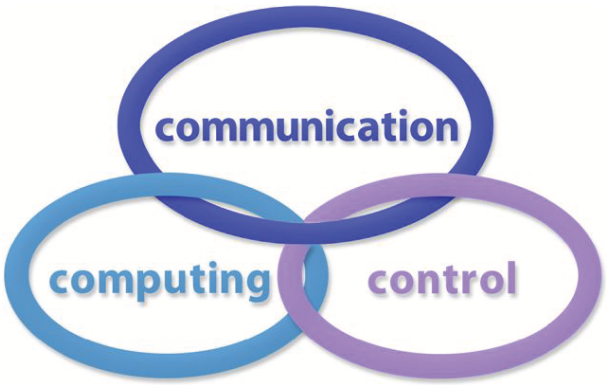
INTERNATIONAL JOURNAL

of

COMPUTERS, COMMUNICATIONS & CONTROL

With Emphasis on the Integration of Three Technologies

IJCCC



Year: 2012 Volume: 7 Number: 2 (June)

Agora University Editing House

CCC Publications

www.journal.univagora.ro

International Journal of Computers, Communications & Control

EDITORIAL ADDRESS:
R&D Agora Ltd. / S.C. Cercetare Dezvoltare Agora S.R.L.
Piaţa Tineretului 8, Oradea, jud. Bihor, Romania, Zip Code 410526
Tel./ Fax: +40 359101032
E-mail: ijccc@univagora.ro, rd.agora@univagora.ro, ccc.journal@gmail.com
Journal website: www.journal.univagora.ro

DATA FOR SUBSCRIBERS
Supplier: Cercetare Dezvoltare Agora Srl (Research & Development Agora Ltd.)
Fiscal code: RO24747462
Headquarter: Oradea, Piata Tineretului Nr.8, Bihor, Romania, Zip code 410526
Bank: MILLENNIUM BANK, Bank address: Piata Unirii, str. Primariei, 2, Oradea, Romania
IBAN Account for EURO: RO73MILB0000000000932235
SWIFT CODE (eq.BIC): MILBROBU

International Journal of Computers, Communications & Control

Mario de J. Pérez Jiménez
Dept. of CS and Artificial Intelligence
University of Seville, Sevilla,
Avda. Reina Mercedes s/n, 41012, Spain
marper@us.es

Dana Petcu
Computer Science Department
Western University of Timisoara
V.Parvan 4, 300223 Timisoara, Romania
petcu@info.uvt.ro

Radu Popescu-Zeletin
Fraunhofer Institute for Open
Communication Systems
Technical University Berlin, Germany
rpz@cs.tu-berlin.de

Imre J. Rudas
Institute of Intelligent Engineering Systems
Budapest Tech
Budapest, Bécsi út 96/B, H-1034, Hungary
rudas@bmf.hu

Yong Shi
Research Center on Fictitious Economy
& Data Science
Chinese Academy of Sciences
Beijing 100190, China
yshi@gucas.ac.cn
and
College of Information Science & Technology
University of Nebraska at Omaha
Omaha, NE 68182, USA
yshi@unomaha.edu

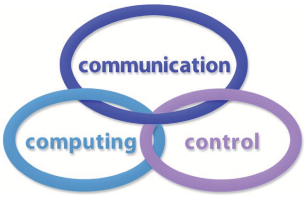
Athanasios D. Styliadis
Alexander Institute of Technology
Agiou Panteleimona 24, 551 33
Thessaloniki, Greece
styl@it.teithe.gr

Gheorghe Tecuci
Learning Agents Center
George Mason University, USA
University Drive 4440, Fairfax VA 22030-4444
tecuci@gmu.edu

Horia-Nicolai Teodorescu
Faculty of Electronics and Telecommunications
Technical University "Gh. Asachi" Iasi
Iasi, Bd. Carol I 11, 700506, Romania
hteodor@etc.tuiasi.ro

Dan Tufiş
Research Institute for Artificial Intelligence
of the Romanian Academy
Bucharest, "13 Septembrie" 13, 050711, Romania
tufis@racai.ro

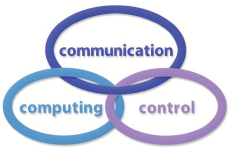
Lotfi A. Zadeh
Professor,
Graduate School,
Director,
Berkeley Initiative in Soft Computing (BISC)
Computer Science Division
Department of Electrical Engineering
& Computer Sciences
University of California Berkeley,
Berkeley, CA 94720-1776, USA
zadeh@eecs.berkeley.edu



International Journal of Computers, Communications & Control

Short Description of IJCCC

Title of journal: International Journal of Computers, Communications & Control
Acronym: IJCCC
Abbreviated Journal Title: INT J COMPUT COMMUN
International Standard Serial Number: ISSN 1841-9836, E-ISSN 1841-9844
Publisher: CCC Publications - Agora University
Starting year of IJCCC: 2006
Founders of IJCCC: Ioan Dzitac, Florin Gheorghe Filip and Mişu-Jan Manolescu
Logo:



Number of issues/year: IJCCC has 4 issues/odd year (March, June, September, December) and 5 issues/even year (March, June, September, November, December). Every even year IJCCC will publish a supplementary issue with selected papers from the International Conference on Computers, Communications and Control.
Coverage:

- Beginning with Vol. 1 (2006), Supplementary issue: S, IJCCC is covered by Thomson Reuters - SCI Expanded and is indexed in ISI Web of Science.
- Journal Citation Reports/Science Edition 2010:
  - Impact factor = 0.650
- Beginning with Vol. 2 (2007), No.1, IJCCC is covered in EBSCO.
- Beginning with Vol. 3 (2008), No.1, IJCCC, is covered in Scopus.

Scope: IJCCC is directed to the international communities of scientific researchers in universities, research units and industry. IJCCC publishes original and recent scientific contributions in the following fields: Computing & Computational Mathematics; Information Technology & Communications; Computer-based Control.
Unique features distinguishing IJCCC: To differentiate from other similar journals, the editorial policy of IJCCC encourages especially the publishing of scientific papers that focus on the convergence of the 3 "C" (Computing, Communication, Control).
Policy: The articles submitted to IJCCC must be original and previously unpublished in other journals. The submissions will be revised independently by at least two reviewers and will be published only after completion of the editorial workflow.

Contents

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 2 (June), pp. 204-217

Application of Chaos Embedded PSO for PID Parameter Tuning

O.T. Altinoz, A.E. Yilmaz, G.-W. Weber

O. Tolga Altinoz
Hacettepe University Bala Vocational School
Electronics Technology Department, Ankara, Turkey
taltinoz@hacettepe.edu.tr

A. Egemen Yilmaz
Ankara University
Electronics Engineering Department, Ankara, Turkey
aeyilmaz@eng.ankara.edu.tr

G. Wilhelm Weber
Middle East Technical University
Institute of Applied Mathematics, Ankara, Turkey
gweber@metu.edu.tr

Abstract: Proportional-Integral-Derivative (PID) control is the most common method applied in the industry due to its simplicity. On the other hand, due to its difficulties, parameter tuning of the PID controllers are usually performed poorly. Generally, the design objectives are obtained by adjusting the controller parameters repetitively until the desired closed-loop system performance is achieved. This allows researchers to use more advanced and even some heuristic methods to achieve the optimal PID parameters. This paper focuses on application of the chaos embedded particle swarm optimization algorithm (CPSO) for PID controller tuning, and demonstrates how to employ the CPSO method to find optimal PID parameters in details. The method is applied to optimal PID parameter tuning for three typical systems with various ordered, and comparisons with the conventional PSO and the Ziegler-Nichols methods are performed. The numerical results from the simulations verify the performance of the proposed scheme.
Keywords: Particle Swarm Optimization (PSO); chaos theory; PID control; multidimensional optimization.

1 Introduction

Numerous control techniques have so far been introduced and applied in industrial problems. On the other hand, since its invention in 1910, Proportional-Integral-Derivative (PID) control has become the best known controller due to its simplcitiy and efficiency in various control problems. Industrial implementation of the PID controller is remarkable that its tuning still presents a challenge in many applications. Primary problem associated with such systems is to obtain the optimal PID controller parameters for a satisfactory control performance, which is a quite difficult task. Therefore, enormous amount of research results have been reported in the literature, among which the well-known tuning method was proposed by Ziegler and Nichols in 1942. Despite the huge number of proposed approaches for PID parameter tuning, most of them (as well as the Ziegler-Nichols method) occasionally yield poor performance in practice.

Generally, conventional tuning methods use the root locus and frequency response for PID parameter tuning. Early presentation of the most common conventional tuning method, the Ziegler-Nichols method, was based on the open-loop step response of the system. With such

definition, the performance was observed to be inadequate and poor. Later, the authors improved the method by making it dependent on the frequency response of the close-loop system. The procedure imposed by this method can be summarized as follows: i) First, the system is assumed to be under proportional control only. The proportion parameter is increased until the system becomes critically stable. The proportion parameter yielding this condition is recorded as well as the corresponding oscillation period; ii) Based on these values, other parameters (i.e. corresponding to the integral and derivative operations) are determined via some look-up-tables which can be found any Control Theory textbook. However, this method yields unsatisfactory phase and gain margins. In addition, due to its off-line unautomatized nature, the Ziegler-Nichols method is not quite applicable to processes in the working systems in practice. For this reason, recently, heuristic methods have been employed for PID parameter tuning in order to improve the controller performance.

In the last decade, heuristic approaches have received increased attention from researchers dealing with engineering problems. Over the years, several heuristic methods have been proposed for PID parameter tuning. In 1995, a novel heuristic approach called the Particle Swarm Optimization (PSO) was introduced by Eberhart and Kennedy [7]. Compared to other optimization algorithms, PSO has a very simple mathematical definition yielding easy-to-implement and short computer programs, which can generate high-quality solutions with reasonable speed. However, like other heuristic methods, conventional PSO suffers from premature convergence especially in higher order complicated problems. In order to improve PSO, so far numerous variants have been introduced. Compared to the conventional PSO, most of these variants show relatively better performance. Much research is still in progress for improvement of the performance of PSO especially in complicated problems. In this study, we present a simple method for PID controller parameter tuning by using the chaos embedded particle swarm optimization algorithm (CPSO).

PID parameter tuning achives in such a way that parameters are adjusted in real-time when the controller is active, which is called on-line tuning or the parameters are measured than the tuned controller is performed on a plant, which is called off-line tuning. Hence, in this study, the CPSO is applied to overall system to obtained the design objectives by adjusting the controller parameters at each iteration repetitively until the desired closed-loop system performance is achieved. This attitude is off-line tuning.

The performance of the closed-loop system can defined in terms of rise time, overshoot, settling time and steady state error. In general, the system with fast rise and settling time under no steady-state error and almost zero overshoot is desired. Hence, in this study to provide a desired performance, mean value of squared error signal, which is different between reference signal and the system output, is minimized by using CPSO and PSO.

These methods (CPSO, PSO and Ziegler-Nichols) are applied to three typical systems of different orders. The first system under investigation is the acceleration model of an object in a frictional environment. The second one is the inverted pendulum control problem, and the last issue is the field-controled of a DC motor problem. When these systems are controlled via PID controller, the closed-loop performance is sensitively dependent to the PID parameters. Hence, determination of these parameters becomes quite critical. Therefore, PID parameter tuning methods should be performed on the closed-loop system.

In Section 2 of the paper, the descriptions and formulations of the controller and problems under investigation are presented. Section 3 contains the definitions of the PSO and CPSO algorithms; and finally, Section 4 presents the numerical results and Section 5 contains the comparisons together with discussions and concluding remarks.

Table 1: Increase in each PID parameter and its impact on transient response

Parameter	Rise Time	Overshoot	Settling Time	Steady State Error
K_P	Decrease	Increase	Increase	Decrease
K_I	Decrease	Increase	Increase	Decrease
K_D	Decrease	Decrease	Decrease	Slight Change

2 Problem Statement

In this study, PID control is discussed and three systems of different orders are investigated for the performance comparison of the proposed CPSO method with the Ziegler-Nichols method and conventional PSO. Each system is summarized via pictorial description as well as its transfer function in the upcoming subsections.

2.1 The Proportional-Integral-Derivative (PID) Control

The basic form of the PID controller composes from sum of the multiplication, integration and differentiation of the error signal, and each operator is multiplied by three control parameters. Time domain representation of PID control equation is presented in Eq. 1, meanwhile the transfer function is given in Eq. 2:

$$u(t) = K_P e(t) + K_I \int e(t)dt + K_D \frac{de(t)}{dt}, \tag{1}$$

$$G(s) = \frac{U(s)}{E(s)} = \frac{K_D s^2 + K_p s + K_I}{s}, \tag{2}$$

where K_P, K_I and K_D are the control parameters, $u(t)$ is the controller output/system input, and $e(t)$ is error signal. In frequency domain, the transfer function of the PID controller has one pole at the origin, and two zeros with locations depending on the tuning strategy.

Table 1 presents the separate and isolated impact of change in each parameter on the system performance. However, in reality, the parameters must be mutually tuned. Therefore, these impacts listed in Table 1 should not be considered as rules-of-thumb, and applied directly for parameter tuning.

2.2 A System of First Order: Acceleration Control of an Object in a Frictional Environment

The first order system in this study results from the acceleration of an object against the friction due to the ground contact, which is one of the fundamental models in vehicular control. The Figure 1 depicts the pictorial description of the relevant system.

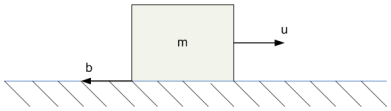


Figure 1: Accelerated object model in a frictional environment.

The transfer function of the system is given in Eq. 3. The model presents the equation of motion for the speed of an object with mass m. Assuming that the input u imposes a force;

Table 2: Setup1a and Setup1b descriptors

Descriptor	Symbol	Setup1a	Setup1b
Friction constant (N.sec/m)	b	50	50
Mass of the object (kg)	m	100	50
Desired velocity of the object (m/sec)	v	1	1

meanwhile the friction force bv restricts the movement of the object, where v is the velocity of the object as well as the output of the overall system. The aim of the controller is to make the object to reach to a desired velocity.

$$\frac{V(s)}{U(s)} = \frac{1/m}{s + \frac{b}{m}} \tag{3}$$

For this particular type of problem, two different setups are constructed and investigated. From now on, these setups will be referred to as Setup1a and Setup1b. The descriptor parameters of these setups are given in Table 2.

2.3 A System of Second Order: Inverted Pendulum

The inverted pendulum is one of the benchmark problems of control theory for illustration and comparison of different control methodologies. The problem comes from the motivation on development of missiles, rockets, robots and other transportation means. In general, the aim in the inverted pendulum problem is to maintain the pendulum at upright position. Therefore, in addition to the PID control, various techniques (such as optimal control [15], linear [9] control, nonlinear control [5], intelligent and adaptive control methods [3]) have also been applied to this problem.

Figure 4 shows the general model of the inverted pendulum. The pendulum is composed of a free moving pendulum with the mass m_p and length l attached to the cart with mass m_t, where a force f is applied to this setup.

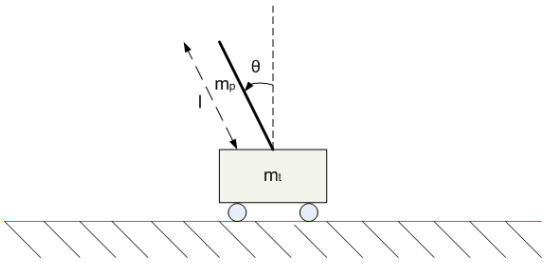


Figure 2: Inverted Pendulum Model.

Eq. 4 shows the transfer function of the inverted pendulum system. The moment of inertia I about the pendulum's mass center is defined as $I = (1/3)m_p l^2$.

$$\frac{\theta(s)}{U(s)} = \frac{m_p l}{((I + m_p l^2) - m_p^2 l^2)s^2 - m_p g l(m_p + m_t)}. \tag{4}$$

For this particular type of problem, four different setups are constructed and investigated. From now on, these setups will be referred to as Setup2a, Setup2b, Setup2c and Setup2d. The descriptor parameters of these setups are given in Table 3.

Table 3: Setup2a - Setup2d descriptors

Descriptor	Symbol	Setup2a	Setup2b	Setup2c	Setup2d
Mass of the cart (kg)	m_t	0.5	0.5	1	1
Mass of the pole (kg)	m_p	1	1	1	1
Length of the pole (m)	l	0.5	0.5	2	2
Initial value of the rod angle (rad)	θ	0.4	0.8	0.4	0.8

Table 4: Setup3a and Setup3b descriptors

Descriptor	Symbol	Setup3a	Setup3b
Motor constant (N.m/A)	k_m	50.10^{-3}	50.10^{-3}
Friction constant (N.sec/m)	b	50.10^{-3}	50.10^{-3}
Resistance (Ω)	R_f	1	10
Field time effect (msec)	τ_f	1	10
Rotor time constant (msec)	τ_l	100	10
Rotor angle (rad)	θ	2	2

2.4 A System of Third Order: Field-Controlled DC Motor

The third problem under investigation is the field-controlled DC motor problem depicted in Figure 1, which has a third order transfer function. The main purpose of this type of systems are to increase speed while reduce the torque. Eq. 5 gives the transfer function of the system.

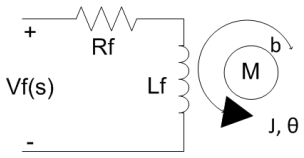


Figure 3: Field-controlled DC motor model.

$$\frac{\theta(s)}{v_f(s)} = \frac{k_m/(bR_f)}{s(\tau_f s+1)(\tau_l s+1)}, \tag{5}$$

where km is the motor constant, b is the friction constant, $\tau_f = L_f/R_f$ is the field time effect, $\tau_l = J/b$ is the rotor time constant, and J is the rotor inertia, v_f is the input and θ is the output of the system.

For this particular type of problem, two different setups are constructed and investigated. From now on, these setups will be referred to as Setup3a and Setup3b. The descriptor parameters of these setups are given in Table 4.

3 Particle Swarm Optimization (PSO)

The PSO algorithm depends on motions of particles (swarm members) searching for the global best in an N-dimensional continuous space. The position of each particle is nothing but a solution candidate, and every time, the fitness of this candidate is re-evaluated. In addition

to its exploration capability (i.e. the tendency for random search throughout the domain), each particle has a cognitive behavior (i.e. remembering its own good memories, and having tendency to return there); as well as a social behavior (i.e. observing the rest of the swarm and having tendency to go where most other particles go).

The original PSO formulation of Kennedy and Eberhart [7] depends on the update of the position $x_i[k]$ and the velocity $v_i[k]$ of the ith particle (swarm member) at the kth iteration as follows:

$$v_i[k+1] = v_i[k] + c_1 \times rand() \times (pbest_i - x_i[k]) + c_2 \times rand() \times (gbest_i - x_i[k]), \quad (6)$$

$$x_i[k+1] = x_i[k] + v_i[k+1] \times \Delta t, \quad (7)$$

where c_1 and c_2 are measures indicating the tendencies of approaching to *pbest* and *gbest*, which are the best positions achieved personally by the ith particle and the whole swarm, respectively. In other words, c_1 and c_2 are the measures of the cognitive and the social behaviors (called cognitive and social parameters), respectively. $rand()$ is a uniformly distributed pseudo-random number generator which produces random numbers between 0.0 and 1.0; and the time step size Δt is taken to be unity for simplicity.

This optimization algorithm demonstrates an outstanding performance under complicated problems. But still, it occasionally faces problems such as getting stuck at local optima and stagnation for multi-dimensional and complex problems. Thus, various improvements have been proposed in order to get rid of these problems. One of the important innovations was introduced by Shi and Eberhart [11], who proposed a term called "inertial weight" in order to improve the performance of the method (used for controlling local and global exploration behavior of the population). By introduction of this term, which puts an additional control on the current velocity of the particles, Eq. 6 is modified as:

$$v_i[k+1] = w[k] \times v_i[k] + c_1 \times rand() \times (pbest_i - x_i[k]) + c_2 \times rand() \times (gbest_i - x_i[k]), \quad (8)$$

which is referred to as the inertial weight PSO, and is currently accepted as the de-facto PSO formulation. Moreover, in a following study [11], Shi and Eberhart showed that the ideal choice for the inertial weight is to decrease it linearly from 0.9 to 0.4. The following pseudocode presents the PSO algorithm.

```
PSO Algorithm
initialize random velocity and position
do
  for i = 1 to swarm size
        1) Calculate fitness function (fit) aimed to be minimized.
        This is the function that the mean squre error of the
        difference between reference signal and the output of the
        system
        if fit < best pattern
             ith particle best position = ith particle position
             best pattern = fit
        end if
  end for
  find min best pattern and corresponding particle
  for i = 1 to swarm size
        Update velocity
```

```
      Update position
  end for
  check the limits of the maximum velocity
while maximum iterations are not exceeded or minimum error is not achieved
```

4 Chaos Embedded Particle Swarm Optimization (CPSO)

After PSO has been presented, various hybrid and innovative methods were introduced in order to get rid of its drawbacks such as premature convergence and stagnation for multi-dimensional and complicated problems. One of the ideas for preventing premature convergence is to embed chaotic maps into PSO. In recent years, chaotic maps have been utilized for generating pseudo-random numbers, and they have been applied to many areas like communication and control theory, since they can produce sufficiently random numbers enhancing the overall performance of the systems. For PSO, the chaotic maps can be helpful for prevention of premature convergence [2]; moreover, it can improve the global/local searching capabilities of the algorithm. Implementation of chaotic maps in PSO can be categorized into two different approaches:

In the first approach, variables and/or random number generators are reinforced with the chaotic maps without any radical changes in the algorithm. In a similar manner, any stochastic algorithm can be improved by using chaotic maps. For PSO, the performance of the algorithm is greatly dependent on the parameters w, c_1, c_2 as well as the random number generators. Thus, any improvement performed on these variables would slightly have influence on the overall performance.

In the second approach, chaos is used in order to interact with the PSO algorithm for searching the solution space. The modification in the PSO algorithm must be fulfilled. In addition to the conventional algorithm, the add-on code is evaluated in order to find a solution candidate with better fitness. At each iteration, the search space boundary is reduced for increasing the searching efficiency. By this way, the particles are kept away from local optimum. This phenomenon is called Chaos Search. In this study, due to its effective performance, the second approach is preferred to the first approach.

The Chaos Embedded Particle Swarm Optimization (CPSO) [8] differs from the conventional PSO with four new operators, which are: chaotic map, variable mapping, inverse variable mapping, and search range operators.

4.1 Chaotic Maps

The discrete-time dynamical system in the iteration form in Eq. 9 is called chaotic mapping or chaotic map.

$$cx_{i+1} = F(cx_i, P), \tag{9}$$

where $F : S \to S$ and $S \in \Re$, $S = [0, 1]$ or $S = [-1, 1]$. P is the chaos parameter, cx is a vector and F is a nonlinear transformation. The equation starting from the initial value cx_0 in the iterative map is obtained. Chaotic maps have sensitivity on initial value and they produce pseudo - random sequences based on cx_0.

One of the most common chaotic maps is the logistic map. The logistic map which is a model of population biology, is frequently used with PSO [8].

$$cx_{i+1}[k] = \mu cx_i[k](1 - cx_i[k]), \tag{10}$$

where μ (which is called bifurcation parameter) is set to 4 for ergodicity as seen in Figure 4. There are two fixed points (If the condition $cx^* = F(cx^*, P)$ holds cx^* is called fixed point) exists

at 3/4 and 0. Moreover, if the chaotic map will be used with CPSO, the fixed points must be checked at the algorithm.

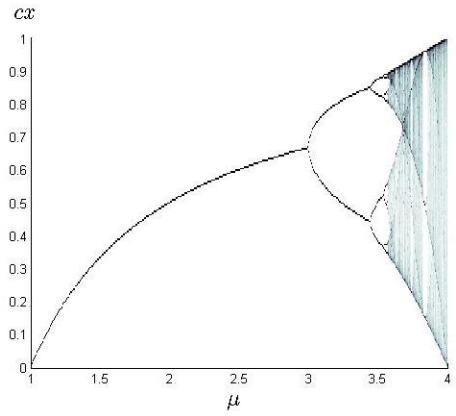


Figure 4: Bifurcation Diagram of Logistic Map.

The chaotic map is a function producing a random number based on the initial value. In CPSO, the initial value is the current position for a particle and chaotic map produces a new position. If the fitness value related to the new position is optimal than the initial position, then the position of a particle is changed as the new position. Thus, it is essential that the chaotic orbit must be ergodic, which means the initial value of the chaotic map must be varied. However, the range of the particle position is not between $[-1, 1]$ or $[0, 1]$. Hence, the particle position should be mapped into chaotic space. Therefore, the variable mapping operator is defined for this purpose.

4.2 Variable Mapping

The position of the particle should be mapped into the chaotic domain by using carrier equation as defined in Eq. 11 [14].

$$cx_i[k] = (x_i[k] - X_{min})/(X_{max} - X_{min}), \tag{11}$$

where cx_i is decision variable, which is the initial value of the chaotic map, x is the position of the particle and X_{min} and X_{max} are the boundaries of the search space. This function maps the range $[X_{min}, X_{max}]$ to the range of chaotic variable $[0, 1]$.

The procedure; variable mapping a particle position and then using this position as an initial value of a chaotic map is called Chaos Search. However, when the optimal point is obtained, it should be mapped to a search space. Hence, the inverse mapping operator is produced.

4.3 Inverse Variable Mapping

After determination of the chaotic variable from a chaotic map with the initial value from variable mapping, the chaotic variable is converted into the particle position by using Eq. 12. This function maps the range of chaotic variable $[0, 1]$ to the range of solution space $[X_{min}, X_{max}]$.

$$x_i[k] = X_{min} + cx_i[k](X_{max} - X_{min}) \tag{12}$$

The position, which is taken from the inverse mapping, is used in order to evaluate the new solutions. If the new solution is better than the non-chaos-search one, the new solution is put pursuant to chaos search.

4.4 Search Range

If the search space extends in a wide area, the searching cannot be completed in the optimal area in a short time. Hence, in order to obtain high performance, the chaotic search is run in a small range. This search area is changed in the current optimal solution neighborhood [4].

$$X_{\min} = \max(X_{\min}, x_g - r(X_{\max} - X_{\min})), \tag{13}$$

$$X_{\max} = \min(X_{\max}, x_g + r(X_{\max} - X_{\min})), \tag{14}$$

where x_g is the global best position so far and r is the variable between 0 and 1. In this study, r is chosen as 0.45, and x_g is chosen as 0 for simplicity. Decreasing the range of the chaotic search will increase the searching efficiency.

The outline of the CPSO used in this study can be listed as follows:

```
Chaos PSO Algorithm
initialize random velocity and position
do
  for i = 1 to swarm size
        1) Calculate the fitness function (fit) aimed to be minimized.
        This function is nothing but the mean square error of the
        difference between reference signal and the output of the
        system
        if fit < best pattern
             ith particle best position = ith particle position
             best pattern = fit
        end if
  end for
  find min best pattern and corresponding particle
  for i = 1 to swarm size
        Update velocity
        Update position
  end for
  check the limits of the maximum velocity
     1) Execute variable mapping that maps the positions into chaos
        variables
     (Section 4.2)
     2) Use Logistic map and find new chaos variable
     (Section 4.1)
     3) Execute Inverse Variable Mapping that converts chaos
     variable into positions
     (Section 4.3)
     if (new position < old position)
         Position = New Position
         break
     end if
     4) Change the search area
     (Section 4.4)
  end for
while maximum iterations are not exceeded or minimum error is not achieved
```

Table 5: Optimization algorithm parameters for the simulation

Property	Value
Number of particles in population	10
Number of iterations ($maxiter$)	50
Social and cognitive parameters (c_1 and c_2)	1.494
Inertia weight (w_{max} and w_{min})	0.9 down to 0.4

Table 6: PID controller parameter values for Setup1a

Parameter	Value Obtained by CPSO	Value Obtained by PSO	Value Obtained by Ziegler-Nichols
K_P	100	22.64	30.52
K_I	38.745	100	109.01
K_D	1	0.1	30.52

5 Simulation Results

In this study, a simulation environment is constructed in order to apply the tuning methods and make performance comparisons. Figure 5 shows the overall setups for all cases; and the selected parameters for the optimization algorithms are given in Table 5.

Figure 5: Overall Setups for all Cases.

The brief summary of the solution strategy to the control problems in this study is as follows: 1) initialize the overall system with random PID parameters; 2) obtain the error, which is the transient response of the system where the reference signal is zero; 3) calculate the fitness function; 4) update the parameters by using Ziegler-Nichols, PSO or CPSO; 5) return to step 2 until the maximum iteration is achieved.

5.1 Results for Setup1

For this type of problem, the PID parameters are tuned by using PSO, CPSO and Zeigler-Nichols methods. Figure 6 shows the transient responses of the closed loop system for Setup1a and Setup1b. Table 6 and 7 present the obtained PID parameters for Setup1a and Setup1b by means of different methods.

It can be observed from the figure that for Setup1a, PSO yields the worst PID parameters. CPSO has no overshoot and better rise time compared to PSO and Ziegler-Nichols. For Setup1b, Ziegler-Nichols yields the worst performance. Even though CPSO seems to outperform to PSO, it can be concluded that CPSO performance is almost the same with PSO.

In summary, for this particular type of problem, CPSO presents better performance compared to PSO and Ziegler-Nichols.

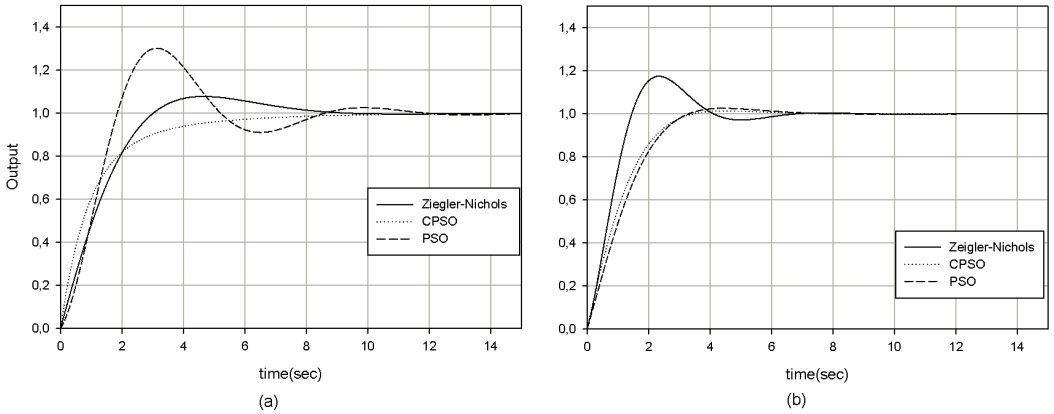


Figure 6: Transient response of the system for a) Setup1a and b) Setup1b.

Table 7: PID controller parameter values Setup1b

Parameter	Value Obtained by CPSO	Value Obtained by PSO	Value Obtained by Ziegler-Nichols
K_P	33.86	24.89	0.85
K_I	49.3	45.57	43.6
K_D	1	1	0.85

5.2 Results for Setup2

For this type of problem, again, the PID parameters are tuned by using PSO, CPSO and Ziegler-Nichols methods. Tables 8 and 9 give the obtained PID parameters, and Figure 3 shows the transient responses of designs.

The results as seen in Figure 3 indicates that CPSO outperforms to PSO and Ziegler-Nichols. However, in Case 2a and Case 2b CPSO performance is similar to that of Ziegler-Nichols and PSO. From the simulations, it is observed that, CPSO is better than PSO; on the other hand, it should be noted that for this particular type of problem, Ziegler-Nichols outperforms to conventional PSO. This result emphasizes the importance of the efforts spent for improvement of the conventional PSO.

5.3 Results for Setup3

The PID parameters obtained for this type of problem are presented in Tables 10 and 11, and illustrated in Figure 8.

Table 8: PID controller parameter values for Setup2a and Setup2b

Parameter	Value Obtained by CPSO	Value Obtained by PSO	Value Obtained by Ziegler-Nichols
K_P	93.6226	60.0136	60.3788
K_I	10.4097	0	4.997
K_D	10.1	3.6553	4.977

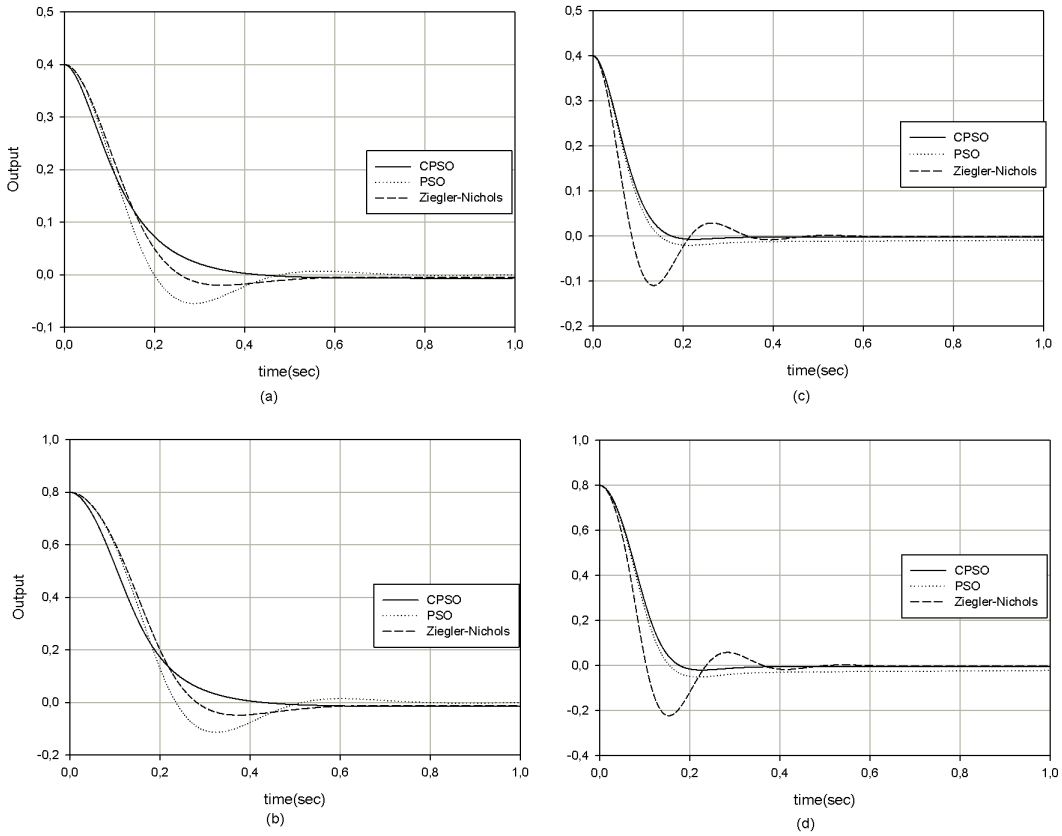


Figure 7: Transient response of the system for a) Setup2a, b) Setup2b, c) Setup2c and d) Setup2d.

Table 9: PID controller parameter values for the Setup2c and Setup2d

Parameter	Value Obtained by CPSO	Value Obtained by PSO	Value Obtained by Ziegler-Nichols
K_P	162.9	169.7	190
K_I	10	79.03	4.9
K_D	10	10	5.12

Table 10: PID controller parameter values for Setup3a

Parameter	Value Obtained by CPSO	Value Obtained by PSO	Value Obtained by Ziegler-Nichols
K_P	395.33	219.20533	1.08
K_I	1.1	55.2488	0.0314
K_D	41.6245	1	6.22

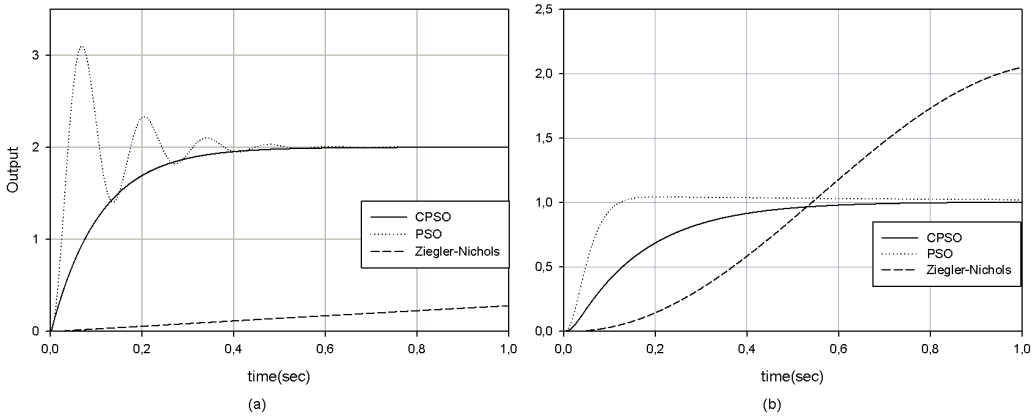


Figure 8: Transient response of the system for a) Setup3a and b) Setup3b.

Table 11: PID controller parameter values for the Setup3b

Parameter	Value Obtained by CPSO	Value Obtained by PSO	Value Obtained by Ziegler-Nichols
K_P	28.73	62.7446	0.2442
K_I	186.95	1.7	87.205
K_D	8.7	1	$1.7\ 10^{-4}$

The results once more indicate the efficiency of the CPSO algorithm. At this point, the following remark shall be made: Even though for Setup3b PSO seems to have a better rise time as well as a settling time compared to CPSO, it yields a steady state error. Therefore, it can still be claimed that CPSO has a better overall performance compared to PSO and Ziegler-Nichols.

6 Conclusion and Future Works

In this study, the chaos embedded particle swarm optimization algorithm (CPSO) is developed, and applied to the some popular engineering problems. The parameter tuning of the PID controller is performed by applying CPSO to three systems. At each iteration, the PID parameters are found off-line; then, the optimized controller is applied to the system. The effectiveness of the proposed method is exposed by comparing it to PSO and Ziegler-Nichols methods. It is observed and concluded that CPSO outperforms to PSO and Ziegler-Nichols, and it is quite efficient for usage in other systems. The scope of the current study is limited to off-line determination of the PID parameters; but simple and parallelizable structure of PSO (and also of CPSO) gives the impression that it is feasible to have real-time implementations of PSO (and also CPSO) in industrial applications for on-line PID controller parameter tuning.

Bibliography

[1] L.S. Coelho, A novel quantum particle swarm optimizer with chaotic mutation operator, *Chaos, Solitons and Fractals*, Vol.37, pp. 1409-1418, 2008

[2] L.S. Coelho, V.C. Mariani, A novel chaotic particle swarm optimization approach using Henon map and implicit filtering local search for economic load dispatch, *Chaos, Solitons and Fractals*, Vol.39, pp. 510-518, 2009

[3] S. Cong and Y. Liang, PID-Like neural network nonlinear adaptive control for uncertain multivariable motion control systems, *IEEE Transactions on Industrial Electronics*, 56(10):3872-3879, 2009

[4] X.Y. Gao, L.Q. Sun, D.S. Sun, An enhanced particle swarm optimization algorithm, *Information Technology Journal*, Vol.8, pp. 1263-1268, 2009

[5] H.N. Iordanou, B.W. Surgenor, Experimental evaluation of the robustness of discrete sliding mode control versus linear quadratic control, *IEEE Transactions on Control Systems Technology*, 5(2):254-260, 1997

[6] C. Jiejin, M. Xiaoqian, L. Lixiang, P. Haipeng, Chaotic particle swarm optimization for economic dispatch considering the generator constraints, *Energy Conversion and Management*, Vol.48, pp. 645-653, 2007

[7] J. Kennedy, R.C. Eberhart, Particle swarm optimization, *IEEE International Conference on Neural Networks*, pp. 1942-1948, 1995

[8] B. Liu, L. Wang, Y.H. Jin, F. Tang, C.X. Huang, Improved particle swarm optimization combined with chaos, *Chaos, Solitons and Fractals*, 25, pp. 1261-1271, 2005

[9] G.A. Medrano-Cersa, Robust computer control of an inverted pendulum, *IEEE Control Systems Magazine*, 19(3):58-67, 1999

[10] Y. Shi, R.C. Eberhart, A modified particle swarm optimizer, *IEEE International Conference on Evolutionary Computation*, pp. 69-73, 1998

[11] Y. Shi, R.C. Eberhart, Empirical study of particle swarm optimization, *Congress of Evolutionary Computing*, pp. 1945-1950, 1999

[12] Y. Song, Z. Chen, Z. Yuan, New chaotic PSO-based neural network predictive control for nonlinear process, *IEEE Transactions on Neural Networks*, 18(2):595-601, 2007

[13] J.M.T. Thompson, H.B. Stewart, *Nonlinear Dynamics and Chaos*, John Wiley and Sons, 2nd Edition, 2002

[14] T. Xiang, X. Liao, K.W. Wang, An improved particle swarm optimization combined with piecewise linear chaotic map, *Applied Mathematics and Computation*, pp. 1637-1645, 2007

[15] G.W. van der Linder, P.F. Lambrechts, H-inf control of an experimental inverted pendulum with dry friction, *IEEE Control Systems Magazine*, 13(4):44-50, 1993

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 2 (June), pp. 218-230

Optimization of Vertical Handoff Decision Algorithm for Wireless Networks

E. Arun, R.S. Moni

Elias Arun, R.S. Moni
Department of Computer Science and Engineering,
Department of Electronics and Communication Engineering,
Noorul Islam University, Tamil Nadu, India.
arunsedly@yahoo.com, r.smoni@yahoo.com

Abstract: To provide mobile users with seamless access anywhere and anytime there is a strong need forinterworking mechanism between cellular networks and wireless local area networks in the next generation wirelessnetworks. Due to the heterogeneous underlying Quality of Service (QoS) support, the admission traffic in these areashas significant impact on overall resource utilization efficiency and QoS satisfaction when multiple services areconsidered. This paper addresses a Call-Level quality of Service (CLS) vertical handoff algorithm between WLANand cellular networks for seamless ubiquitous access. The CLS involves call blocking/dropping probabilities, meandata transfer rate, and number of handoff per call. Based on the above proposed admission strategy, the admissionregion of a cell or WLAN for the traffic can be derived with the function of new call arrival rate, handoff callarrival rate, and the radius of WLANs. The blocking and dropping probabilities are calculated under the guardchanneladmission strategy. The radius of WLAN is determined by using Simulated Annealing (SA) method to minimizethe cost function. Moreover, handoff traffic should be differentiated from new traffic in terms of call admission.When a Mobile Node (MN) moves from an area, with only cellular coverage to an overlaid WLAN area, theongoing call of the MN should be handed over to the WLAN, and handoff from WLAN to cellular network if theyleave the scope. The results based on a detailed performance evaluation study are presented here to demonstrate theefficacy of the proposed algorithm.
Keywords: heterogeneous wireless networks, vertical handoff, seamless mobility, integration handover.

1 Introduction

The past decade has witnessed the fast evolution and successful deployment of a number of wireless access networks. The two most promising ones are cellular networks and wireless local area network (WLAN). Driven by the service anywhere and anytime concept, it is well accepted that Fourth-Generation (4G) wireless networks will be heterogeneous, integrating different networks to provide seamless internet access for mobile users with multi mode access capability. One major challenge in cellular/WLAN internet working is how to take advantage of the wide coverage and almost universal roaming support of cellular networks and the high data rates of WLAN. Many issues should be carefully addressed to achieve seamless interworking,such as mobility management, resource allocation, call admission control, security and billing.This article focuses on how to properly admit incoming traffic to the cell or WLAN and when to take handoff decision. The process of switching connections among networks is called handoff. In each network, an admission control policy either accepts the connection request and accordingly allocates the requested bandwidth or blocks the connection request.Higher priority is usually given

to accept the connection requests from handoff users not the new users. The reason is that from the user's point of view, having a connection abruptly terminated is more annoying than being blocked occasionally on new connection attempts. A service request rejected by its first-choice network can just leave the system or further try to access the other network [1].

There are some related researches on similar problem in two-tier hierarchical cellular networks, in which small size microcells overlay with large macro cells. Many proposed admission strategies [1] [2] are based on user mobility and traffic characteristics. The vertical handoff [3] process involves three stages. The first is the network discovery. In this phase, Mobile Nodes (MN) periodically searches if there are some other different types of wireless networks and take these discovered networks as candidates. The second is the handoff decision phase where MNs compare the state of the current network with candidates, and select one as the handoff target from them according to a certain criterion. The last is the handoff implementation phase where MNs execute the handoff actions and associate with the newly authenticated network. Among these three stages, the handoff decision phase is very important, because it has a direct influence on the network performance and the quality of service of nodes. The objectives of the proposed framework are to maximize network utility through efficient resource allocation, achieve prioritization among different types of connections such as new connections and vertical and horizontal handoff connections, and ensure that the performance of ongoing connections doesn't deteriorate due to accepting too many connections in a service area.

The authors of [4] propose a vertical handoff decision method that simply estimates the service quality for available networks and selects the network with the best quality. In the literature, vertical handoff algorithm is developed in different directions. One of them takes the Received Signal Strength (RSS) and some other factors such as bandwidth, delay and distance into consideration to select the best network through a simple comparison [5] [6]. Another approach utilizes the Artificial Intelligent Techniques such as Neural Network, Fuzzy Logic and machine learning, combining these factors considered to select the best network [7] [8] . The above methods mainly consider the Quality of Service of nodes after handoff, but do not consider the overall system performance such as resource utilization affected by handoff. The service-differentiated admission scheme proposed in [9] applies a different admission strategy for voice service, in which the cellular network is the first choice for voice call and no vertical handoff from the cell to the WLAN is executed for ongoing voice calls. To maximize resource utilization, a complex set of admission parameters need to be determined so that the traffic load is properly distributed to the cells and WLANs. This research is based on the resource management and the handoff execution of a node is used to make the resources be optimally utilized [10]. By properly setting the effective bandwidth of services, packet-level QoS such as packet delay and packet loss can be guaranteed, as long as the allocated bandwidth to traffic is no less than the corresponding effective bandwidth requirements.

In the following, the focus is on call-level QoS in terms of call blocking/dropping probabilities, mean data transfer time and number of handoffs per call. Based on the proposed admission strategy, the admission region of a cell or WLAN for the traffic can be derived with the functions of new call arrival rate, handoff call arrival rate and the radius of WLANs. In view of that in a period, the call arrival rates are stable, so it can be represented all the probabilities as functions of the radius of WLAN. The blocking probabilities and the dropping probabilities are calculated under the guard-channel call admission strategy. The radius of WLANs is determined by using Simulated Annealing method to minimize the cost function.

The rest of the paper is organized as follows. Section 2 provides cellular/WLAN system description. Section 3 proposes the optimal admission control for cellular/WLAN. Section 4 proposes the vertical handoff decision algorithm to minimize the call-level QoS. In Section 5 performance of the proposed algorithm is discussed. Finally the conclusion is stated in Section6.

2 Cellular/ WLAN System Description

Consider an integrated cellular/WLAN system where one or more WLANs may be deployed inside each cell of the cellular system as shown in fig.1. There are two specific coverage areas to be considered: the cellular-only coverage area and the dual cellular/WLAN coverage area. In this context, coverage means service availability. A Mobile Node (MN) can be existing at a given time in the coverage area of a cellular alone. But due to mobility it can move into the regions covered by more than one access networks simultaneously within the coverage area of an UMTS BS and an IEEE802.11 AP. Multiple 802.11 WLAN coverage areas are usually contained within an UMTS coverage area. Horizontal and Vertical Handoffs can occur in different coverage areas. In this section it is described a model to formulate a multi-service integrated UMTS/WLAN system.

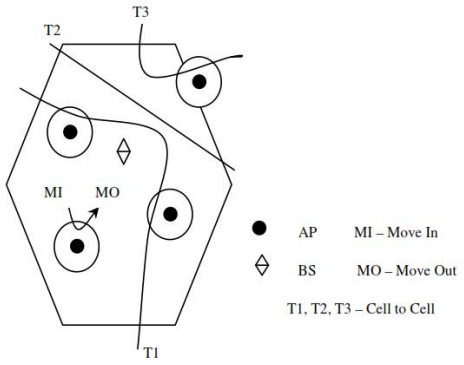


Figure 1: Integrated heterogeneous network

It is assumed that a UMTS network covers k WLANs and all the WLANs have no overlapping areas and are directly adjacent between two UMTS networks. For simplicity, we only draw a single BS and some APs in fig.1, although there are many other cellular networks besides the cellular network. However this has no influence on the design and analysis of our handoff algorithm.

Let the radii of the UMTS network and WLAN is r_u and r_w respectively. The number of channels to be Cu and Cw. WLANs are usually deployed in an indoor environment, where user mobility level is very low and may significantly differ from that of other areas. Hence homogeneous mobility model may not be applicable and it is necessary to differentiate the user mobility characteristics in the double-coverage area from those in the cellular-only area. In the following analysis, a non-uniform model is used to characterize the user mobility within a cell cluster.

Let t_{res}^{co} denote the residence time that a user stays within the cellular-only area before moving to neighboring cells with probability p^{cc} or to the overlaying WLAN with probability p^{cw} and t_{res}^{dc} the user residence time in the double-coverage area. t_{res}^{co} and t_{res}^{dc} are assumed to be exponentially distributed with parameters η^{co} and η^{dc}, respectively. As shown in [11], for the MN with mean velocity V and uniformly distributed movement direction over $[0 - 2\pi]$, the average region boundary cross-over rate ç is given by $\eta = V(L/\pi s)$, where L and S are the boundary length and area of the region respectively.

To analyze the dwell time, we adapt the third model [12] which is defined as the duration that a node stays in a certain region before it moves out of the boundary of the region, of a node

in a region. The third model assumes the nodes are uniformly distributed throughout the whole area and each MN moves in any direction with equal probability. The area of the region in the heterogeneous wireless networks that is in double-coverage defined as S^{dc} and boundary length L^{dc}. Similarly the area of the region that is covered by a WLAN is defined as S^{w} and boundary length L^{w}. The area of the region that is covered by cellular-only area is S^{co} and boundary length L^{co}. Hence the area S^{dc} equals to the sum of S^{w} and S^{co}. The surface area

$$S^{co} = S^{dc} - kS^{w} \tag{1}$$

where k is the k^{th} WLAN.

To model the mobility, it is defined inter-boundary time similar to [13, 14], as the time interval between any two consecutive access network boundary crossings by a mobile user. The wider the coverage area or the more stationary users, longer the inter-boundary times. If the inter-boundary time starts at the moment of entering cell i, then it is denoted by t^{c}_{bi}. If an interboundary time starts at the moment of entering WLAN k then it is t^{w}_{bk}. It is assumed that t^{c}_{bi} and t^{w}_{bk} are exponentially distributed with means $1/\ \eta^{c}_{i}$ and $1/\ \eta^{w}_{i}$ respectively. Hence it is noted that the arrival rates of handoff calls and new calls follow poison distribution, the dwell time in a certain region follows exponential distribution and the call duration time follows exponential distribution with mean value as $1/\eta$. The channel holding time can be defined as the time that a connected mobile user keeps using basic bandwidth resources in each network.

For service s, the channel holding times in cell i and in WLAN k are obtained as $\min(t^{R}_{s}, t^{c}_{bi})$ and $\min(t^{R}_{s}, t^{w}_{bi})$ respectively where t_s is the connection time of a service s. Since t^{R}_{s}, t^{c}_{bi} and $t^{w}_{bi}$are exponentially distributed, the channel holding times are also exponentially distributed with parameters $\mu^{c}_{is} = \vartheta_s + \eta^{c}_{i}$ and$\mu^{w}_{ks} = \vartheta_s + \eta^{w}_{k}$ respectively. With the heterogeneous QoS support of the underlying structure, the incoming traffic in the double-coverage area should be properly admitted to the cell and WLAN. It is assumed that the calls are uniformly distributed in the cellular region, so the call requests in cellular region can be classified as,

(i) New call requests to the WLAN with arrival rate λ^{w}_{n} which is equal to $(\mathrm{S}^{w}/S^{u})\lambda_n$ where λ_n is the new call arrival rate, and S^{u} is the surface area of the cell.

(ii) New call request to the cellular-only area with arrival rate λ^{co}_{n} equaling to $(S^{co}/S^{u})\lambda_n$.

(iii) New call arrival rate in double-coverage area λ^{dc}_{n} which is equal to $(S^{dc}/S^{u})\lambda_n$.

(iv) The arrival rate of handoff calls between neighboring cell λ^{cc}_{h}.

(v) The arrival rate of handoff calls from cellular-only area to overlaying WLAN λ^{cw}_{h}.

(vi) The arrival rate of handoff calls from WLAN to the overlaying cell λ^{wc}_{h}.

With the above parameters the average dwell time in k^{th} WLAN area ,(a_w) is $1/\eta^{w}_{k}$,where the

$$\eta^{w}_{k} = VL^{w}/\pi S^{w} \tag{2}$$

The average dwell time in double-coverage area (a_{dc}) is $1/\eta^{dc}$ where

$$\eta^{dc} = VL^{dc}/\pi S^{dc} \tag{3}$$

The average dwell time in cellular-only area (a_{co}) is $1/\eta^{co}$ where

$$\eta^{co} = VL^{co}/\pi S^{co} \tag{4}$$

Consider an MN is in cellular-only area. When MN is moving out of this region, it may enter into adjacent cell or move into WLAN region. Hence, the average dwell time in cellular-only region before MN moves into another adjacent cell is $1/\eta^{cc}_{h}$ where

$$\eta^{cc}_{h} = VL^{dc}/\pi(S^{dc} - kS^{w}) \tag{5}$$

The average dwell time in cellular-only region before MN moves into WLAN area is $1/\eta_h^{cw}$ where

$$\eta_h^{cw} = VL^w/\pi(S^{dc} - kS^w) \tag{6}$$

3 Optimal Admission Control for Cellular/WLAN Interworking

With Call Admission Control the heterogeneous network blocks some new call requests in order to reduce interference on the network so that the outage probability decreases. Given the cell bandwidth C^c and total offered traffic load, the minimum bandwidth needed to meet the requirements of call blocking and dropping probabilities can be obtained as x_n. N_n^c where x_n is the bandwidth requirement of a new call and N_n^c is the maximum number of new call requests allowed in a cell ($N_n^c \leq C^c/x_n$). It is noted that in cellular-only region only cellular access is available so in this paper randomized guard channel method [15] is applied to give the new and handoff traffic in this area a priority to access the cell bandwidth over the traffic in the doublecoverage area. Because the call blocking and dropping probabilities are very sensitive to the amount of reserved bandwidth , the guard bandwidth for high priority call traffic is randomized instead of an integer number of guard channels.

Each cellular network has C^c channels and each WLAN has C^w channels and $C^c - N_n^c$ channels in cellular network are reserved only for handoff calls. Similarly $C^w - N_n^w$ channels in WLAN are reserved only for handoff calls; when the occupied channels are less than N_n^c in cellular networks, the call admission region of the cell is given in terms of (N_n^c, G_h^c, G_{nh}^c) vectors, in which $G_{nh}^c (\leq N_n^c)$ is a real number to represent a randomized number of guard channels dedicated to new and handoff calls in cellular-only area and G_h^c is the guard bandwidth reserved only for handoff traffic in this region.

When the occupied channels are equal to or more than N_n^c only the handoff calls could be allowed; and the same to the WLANs. Within the two-tier overlaying structure, the vertical handoff from the cell to the overlaying WLAN is not necessary but optional to maintain an ongoing call. Hence the handoff traffic load to the WLAN can be controlled by properly adjusting the admission parameters to the WLAN, by using a simple guard channel method. Due to different quality of service support and resource sharing policy in the underlying networks, the configuration of admission regions of the cell and WLAN can have a significant impact on the overall system performance.

Let $P_{bn}^{req}, P_{dh}^{req}$ and t_d^{req} are the requirements of new call blocking and handoff call dropping probabilities and mean transfer rate respectively. Then the admission control problem can be formulated as

$$max_{N_n^w} \lambda_d \tag{7}$$

Subject to : $P_{bn}^w, P_{bn}^{dc} \leq P_{bn}^{req}$; $P_{bn}^{co} \leq P_{bn}^{req}$; $P_{dh}^c \leq P_{bh}^{req}, E(t_d) \leq t_d^{req}$.

Where P_{bn}^{co} and P_{bn}^{dc} are the blocking probabilities of the cell for new calls in the cellular-only area and double-coverage area respectively, P_{dh}^c is the handoff dropping probability of the cell, $P_{bn}^w$is the probability that a new call is blocked by the WLAN and $E(t_d)$ is the mean transfer rate time. Thus the maximization of λ_d implies a maximization of the total acceptance traffic load and resource utilization.

3.1 New call blocking and dropping probabilities

We use a K+1 dimensional Markov Chain to analyze the guard channel admission algorithm. Let $(k_n^w, k_{nco}^c, k_{ndc}^c)$ denote the state of the new call arrival in a cell cluster, where $k_n^w, k_{nco}^c, k_{ndc}^c$ are the numbers of new calls admitted to the WLAN, to the cell from the cellular-only area and

to the cell from double-coverage area respectively. First, the number of new calls in the WLAN can be described by a birth-death process with respect to k_n^w. Since both new call duration and user residence time in the double-coverage area are exponentially distributed the channel holding time of new calls in WLANs, $min(t_s^R, t_{bi}^c)$ is exponential with mean $1/\mu_n + \eta^{dc}$ where $1/\mu_n$ is the mean new call duration. Then the steady state probability of k new calls in the WLAN is obtained based on m/m/k/k loss system, as

$$\pi_n^w = ([(\lambda_n^{dc} + \lambda_n^{cw})/(\mu_n + \eta^{dc})]^k/k!)/(\Sigma_{i=0}^{N_n^w}([(\lambda_n^{dc} + \lambda_n^{cw})/(\mu_n + \eta^{dc})]^i)/i! \tag{8}$$

Hence the new call blocking probability in the WLAN is

$$P_{bn}^w = \pi_n^w(N_n^w) \tag{9}$$

In the following, we derive the state-dependent transition rates, which are given by:

i) $(K_{nco}^c, K_{ndc}^c) \to (K_{nco}^c + 1, K_{ndc}^c)$ and $K_{nco}^c < N_n^c$.

This happens when there is a new call request in s^{co} or a handoff call request comes from adjacent cellular networks. Since there is no transition at double-coverage area, it is impossible that a handoff call request comes from one of the WLANs, otherwise K_{ndc}^c should be $K_{ndc}^c - 1$. So the transition rate is

$$\lambda_n^{co} + \lambda_h^{cc} \tag{10}$$

ii) $(K_{nco}^c, K_{ndc}^c) \to (K_{nco}^c + 1, K_{ndc}^c)$ and $K_{nco}^c \geq N_n^c$ This will happen when there is a handoff request comes from adjacent cellular network. Because $K_{nco}^c \geq N_n^c$ it is impossible that a new request call is admitted, and because there is no change of state for the MN in K_{ndc}^c it is also impossible that a handoff call request comes from WLAN. So the transition rate is

$$\lambda_h^{cc}. \tag{11}$$

iii) $(K_{nco}^c, K_{ndc}^c) \to (K_{nco}^c, K_{ndc}^c + 1)$ and $K_{ndc}^c < N_n^w$ this will happen when there is a new call request in the k^{th} WLAN. Since there is no channel released in cellular network, it is impossible a handoff call request comes from cellular network. So the transition rate is

$$\lambda_n^w. \tag{12}$$

iv) $(K_{nco}^c, K_{ndc}^c) \to (K_{nco}^c, K_{ndc}^c - 1)$ and $K_{nco}^c = N_c^c$ it shows that the channels in WLAN have been used up, the channel released in cellular networks might be driven by three events: call finishes communication; call leaves for neighbor cellular network; call leaves for the k number of used-up-channel WLANs. Hence the transition rate is

$$K_{nco}^c(\mu_n + \eta_h^{cc}) + (K - k)\eta_n^{co} \tag{13}$$

v) $(K_{nco}^c, K_{ndc}^c) \to (K_{nco}^c, K_{ndc}^c - 1)$ and $K_{nco}^c = C^c$ It means that the channels in cellular network have been used up. The channel released in the kth WLAN may be driven by two events: call finishes its communication and call leaves for the cellular network. So the transition rate is

$$K_{ndc}^c(\mu_n + \eta^w) \tag{14}$$

vi) $(K_{nco}^c, K_{ndc}^c) \to (K_{nco}^c, K_{ndc}^c - 1) and K_{nco}^c < C^c$ This means that the channels in cellular network have not been used up and the channel released in the kth WLAN is only driven by the event that call finishes its communication. It is impossible that a call comes from S^w to S^{co} otherwise (K_{nco}^c should be $K_{nco}^c + 1$. So the transition rate is

$$K^c_{ndc}\mu_n \tag{15}$$

As given in equation [12] and [13] the state departure rates vary with the number of existing new calls in WLAN K^w_n based on which handoff calls from the cell are admitted or blocked by the WLAN. Hence the new calls admitted to the cell from cellular-only region and double-coverage region has different mean channel holding time. Therefore, the cell can be viewed as a multiservice loss system [16].

A product-form state distribution exists and is insensitive to service time distributions, provided that the resource sharing among services is under coordinate convex policies. This requires that transitions between states come in pairs. For loss systems with trunk reservation like guard channel method, the insensitivity property and product-form solutions are destroyed due to the one-way transitions at some states. A recursive method is proposed in [17] to approximate the state distribution, which is shown to be accurate for a wide range of traffic intensities and when the service rates do not greatly differ from each other.

The blocking probabilities are almost insensitive to service time distributions. Hence, we use the recursive approximation in [17] to obtain the steady- state probability of new calls admitted into the cell π^c_n. Thus the blocking probabilities of the cell for new calls in the cellular-only region and double-coverage region are given by

$$P^c_{bnco} = G^c_h - \lfloor G^c_h \rfloor \, \pi^c_n \, \lfloor N^c_{nco} \rfloor + \sum_{i=\lfloor N_{nco} \rfloor + 1}^{N^c_n} \pi^c_n(i) \tag{16}$$

$$P^c_{bndc} = G^c_{nh} - \lfloor G^c_{nh} \rfloor \, \pi^c_n \, \lfloor N^c_{ndc} \rfloor + \sum_{i=\lfloor N_{ndc} \rfloor + 1}^{N^c_n} \pi^c_n(i) \tag{17}$$

and the dropping probabilities of the cell are given by

$$D^c_n = \pi^c_n(N^c_n) \tag{18}$$

3.2 Average arrival rates of Handoff calls

The handoff arrival rates are related to the handoff probabilities. The handoff probability of new calls in the cellular-only area to neighbouring cells is denoted as H^{cc}_{nc} is given by $P^{cc}\varphi(-\mu_n)$ where, φ is the moment generating function. Similarly, the handoff probability of new calls in the cellular-only area to the overlaying WLAN is denoted as H^{cw}_{nc} is given by $H^{cw}_{nc} = P^{cw}\varphi(-\mu_n)$ With an exponentially distributed user residence time in the double-coverage area, the handoff probability of new calls from WLAN to the overlaying cell is

$$H^{wc}_{nc} = \eta^{dc}/(\eta^{dc} + \mu_n) \tag{19}$$

Hence the handoff traffic from the WLAN to the overlaying cell has a mean arrival rate λ^{wc}_h is given by

$$\lambda^{wc}_h = H^{wc}_{nc}(\lambda^{dc}_n + \lambda^{cw}_h)(1 - P^w_{bn}) \tag{20}$$

The mean arrival rates of handoff traffic between neighbouring cells is

$$\lambda^{cc}_h = H^{cc}_{nc}[\lambda^{co}_n(1 - P^{co}_{bn}) + (\lambda^{wc}_h + \lambda^{cc}_h)(1 - P^{co}_{dh}) + \lambda^{dcc}_n(1 - P^{dc}_{bn})H^{wc}_n] \tag{21}$$

And the mean arrival rates of handoff traffic from the cell to the overlaying WLAN is

$$\lambda_h^{cw} = H_{nc}^{cw}[\lambda_n^{co}(1 - P_{bn}^{co}) + (\lambda_h^{wc} + \lambda_h^{cc})(1 - P_{dh}^{co}) + \lambda_n^{dcc}(1 - P_{bn}^{dc})H_n^{wc}] \quad (22)$$

Thus the new call blocking and dropping probabilities can be obtained recursively from (9),(18),(20),(21) and (22).

4 Vertical Handoff Decision Making Algorithm

From the analysis in section 3, it is learnt that if the radius of WLAN is fixed, the drop probability and the block probability will be determined by new call arrival rate and handoff call arrival rate. Usually, the arrival rates in a short period are invariable. Hence the radius of WLAN can be adjusted by regulating the transmission power to optimize these block probabilities and drop probabilities. In order to accomplish an optimization, we formulate a combined cost function

$$G = \pi_n^c + \beta_1 \pi_n^c + \beta_2 \lambda_h^{wc} + \beta_3 \lambda_h^{cc} + \beta_4 \lambda_h^{cw} \quad (23)$$

which is determined by the block probabilities and the drop probabilities in cellular networks and one of the WLAN, and is a function of the radius of WLANs. When the radius of WLAN is determined, all the nodes in the WLAN should communicate with WLAN. Even if the channels in WLAN are used up and there are free channels in cellular networks, the call request in WLAN should not be admitted by cellular network would be blocked or dropped. The communication node shall handoff to WLAN if it enters WLAN and handoff to cellular network if it leaves current WLAN. In the equation (23), β_k , k = 1,2,3,4 denotes the weight of dropping probability in WLAN, handoff probability in WLAN to cellular network, handoff probability in cellular-only area to adjacent cell and handoff probability in cellular network to WLAN respectively.

As to the new call originates in a_w it is covered by cellular network and couldn't have been admitted by cellular network. According to the above algorithm it loses the opportunity and this will not be happened much often, so the weight of β_1 should be bigger than one. As the MN moving from a_{co} to a_w the handoff is not required because it could have been using the original channel to communicate, but handoff is required if MN moves from a_w to a_{co} otherwise, it will lose the channel of WLAN. So the weight of β_4 should be somewhat greater than β_2 and β_3. Because terminating an on-going call is far more annoying than refusing to admit a new call from user's point of view, β_4, β_3 and β_2 should be much bigger than β_1. The objective function and its constraints are,

$$minG = \pi_n^c + \beta_1 \pi_n^c + \beta_2 \lambda_h^{wc} + \beta_3 \lambda_h^{cc} + \beta_4 \lambda_h^{cw} \quad (24)$$

such that $R_{min} \leq R_w \leq R_{max}$, where R_{min} and R_{max} are the minimum radius and the maximum radius of the WLAN.

It is very difficult to determine the optimal radius of WLAN by numerical programming methods. Simulated Annealing (SA) is a stochastic computational technique derived from statistical mechanics for finding near globally minimum-cost solutions to large optimization problems. Finding the global minimum value of an objective function with many degrees of freedom subject to conflict ting constraints is an NP-complete problem.

Therefore, the objective function will tend to have many local minima. A procedure for solving optimization problems of the above nature can be implemented by following the Successive Decent Algorithm by Kirkpatric et al [18] which follows the evolution of a solid thermodynamic equilibrium with a decreasing succession of temperature values. The procedure is as follows:

Step 1: Begin minimization.

Select an initial radius $r_i \in$ E randomly for all WLANs;
Select an initial control parameter T greater than 0;
Select parameter change counter t = 0;
Repeat;
Set repetition counter k=0;
Repeat.
Step 2: Select a new radius r_n in $r_i - 6, r_i + 6$ randomly, r_n is in $[R_{min}, R_{max}]$;
Compute S = $G(r_n) - G(r_0)$.
Step 3: If S less than 0 then $r_0 := r_n$ goto step 5.
Step 4: Else if *rand(0,1)* less than *exp(-S/T)* then $r_i := r_n$.
Step 5: k := k+1.
Until k = R(t);
t:=t+1.
Step 6: If T greater than 30, then output r_0 as the optimal radius.
Otherwise k := 0 and;
goto step 2.

5 Performance Analysis

In this section, we evaluate the performance of the proposed vertical handoff decision algorithm in terms of call blocking/ dropping probabilities. Due to the differentiation of new and handoff traffic in different areas, the analysis is very complex. By applying the call admission control algorithm given in section 3, we can obtain the best configuration for admission parameters to maximize the admissible traffic load with the given cell/ WLAN cluster. Fig.2 shows the relationship between the total acceptance traffic load λ_d and the maximum number of new calls arrived in the WLAN N_n^w under blocking probabilities ≤ 0.01, dropping probabilities ≤ 0.001 and mean data transfer $\leq$ 4s.

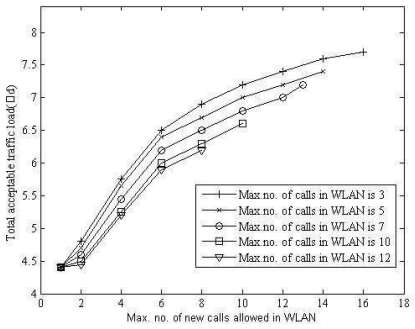


Figure 2: Max. acceptable data traffic load versus max. number of data

It is observed from fig.2 that the total acceptance traffic load increases with N_n^w when N_n^w is relatively small. Fig.3 illustrates the call-level quality of service performance with different N_n^w. It is noted that the simulation results of new call blocking and dropping probabilities are very close to the analytical results. The performance fluctuation of handoff dropping probability is due to the maximum number of calls allowed in the cell and WLAN are both integer variables. From Fig.4 and.5 it is noted that the block probability and drop probability in cellular network

decreases when the radius of WLAN becomes larger. This is because more and more of new call requests are admitted by WLAN.

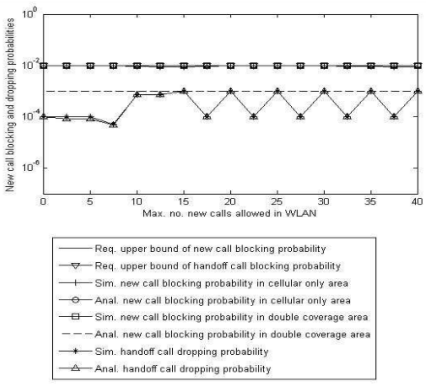


Figure 3: Call level QoS performance with different 5

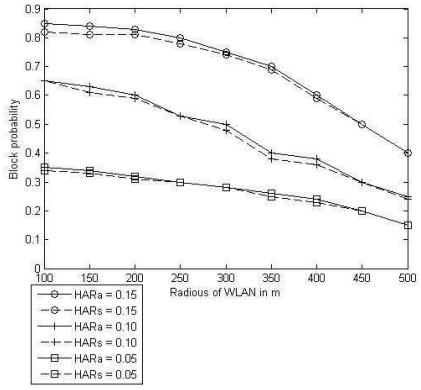


Figure 4: Block probability in cellular networks

Fig.6 and Fig.7 show that the block probability and drop probability in WLAN becomes larger with the radius of WLAN. The reason here is, when the radius becomes expanding, more and more new calls can be admitted to the WLAN, this will result in increase of block probability and drop probability.

Fig.8 illustrate the optimal radius of WLAN under different new call arrival rate and different handoff call arrival rate. It is observed that when the new call arrival rate is fixed, the radius varies consistent with the handoff call arrival rate. This is because the handoff call requests are only sent to the cellular network, which will increase the drop probability in cellular network, and this leads to maximize the cost function. In order to minimize the cost function, the drop probability must be reduced. This is done by redirecting some of the new call request to WLAN. It is also noted that the handoff call arrival rate is fixed; the radius varies inversely with the new call arrival rate. This is because the number of channels in cellular network is more than that in WLAN.

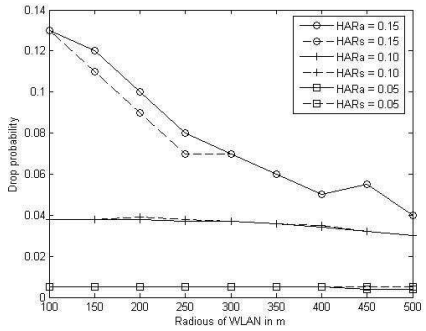


Figure 5: Drop probability in cellular networks

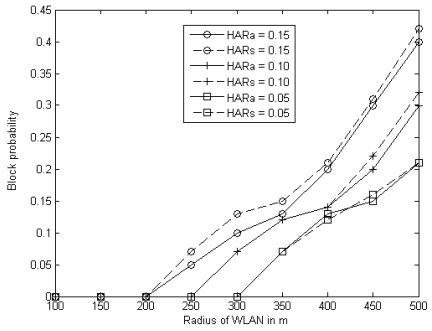


Figure 6: Block probability in WLAN

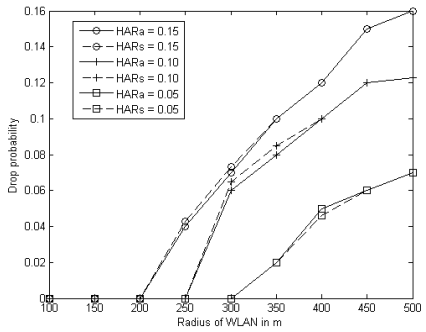


Figure 7: Drop probability in WLAN

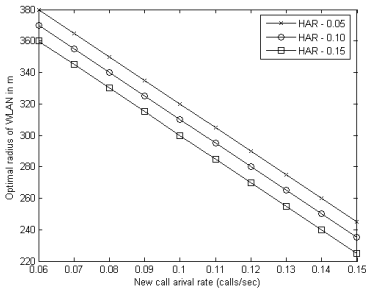


Figure 8: Optimal radius of WLAN

6 Conclusion

When connections need to migrate between heterogeneous networks for performance and high availability reasons, then seamless vertical handoff is necessary the first step. In this paper, we tried to highlight the block probability of new calls and drop probability of handoff calls in heterogeneous networks and their computation and the cost function is proposed which is based on the block probabilities and drop probabilities. The optimal radius of WLAN is determined using the simulated annealing method. All the mobile nodes entering the scope of WLAN should handoff from cellular network to WLAN, and the MN leaving the scope should handoff from WLAN to cellular network. Our performance results based on detailed simulations illustrate that the proposed algorithm could achieve good effects.

Acknowledgement.

This work was supported by AICTE, Govt. of India, File No : 8023/BOR/RID/RPS - 66/2009-10.

Bibliography

[1] K.Maheswary and A.Kumar, Performance analysis of microcellization for supporting two mobility classes in cellular wireless networks, *IEEE Trans.Veh. Technolog*, 49(2):321-333, 2000

[2] T.Klein and S.J.Han,Assignment Strategies for mobile data users in hierarchical overlay networks: Performance of Optimal and adaptive strategies, *IEEE Int J. of select. Areas. Commun.*, 22(5):849-861, June 2004.

[3] J.Ncnair, and F.Zhu, Vertical handoff in 4G multi network environments, *IEEE Wireless Communications*, 11(3):8-15, 2005

[4] N.Nasser, A.Hasswa and H.Hassanein, Handoff in 4G heterogeneous networks, *IEEE communication magazine*, 44(10):96-103, 2006

[5] E.Arun, R.S.Moni, A novel decision scheme for vertical handoff in 4Gwireless networks, *Global Journal of Computer Science and Technology*, 10(5):28-33, 2010

[6] A.H.Zahran, B.Liang, and A.Aaleh,Signal threshold Adaptation for vertical handoff in Heterogeneous wireless networks, *vMobile Network Applications*, 11:625-640, 2006

[7] E.Steven, S.Navarro, V.W.S.Wong, and Y.X.Lin, A vertical handoff decision algorithm for Heterogeneous wireless networks, *Wireless Communication and networking Conferrence*, pp. 3199-3204, 2007

[8] E.Arun, R.S.Moni, Optimization algorithm for a handoff decision in wireless heterogeneous networks, *International Journal of Next-Generation Networks (IJNGN)*, 2(3):99-117, 2010

[9] W.Song and W.Zhuang, QoS provisioning via admission control in cellular/wireless networking, in *Proc. 2nd Intl conf. on Broadband Networks (BROADNETS'05)*, 1:585-593, 2005

[10] P.K.Tang, Y.H.Chew and L.C.O.Michael, Improvement in Grade-of-service in a cooperative overlaying heterogeneous networks, *6th Int conf. on Information Communication and Signal Processing*, pp. 1-5, 2007

[11] B.Jabbari, Teletraffic aspects of evolving and next generation wireless communication networks, *IEEE Pers. Comm.*, 3(6):4-9, 1996

[12] Q.A.zeng and D.P.Agerwal, Modeling and efficient handling of handoff in Integrated Wireless Mobile Networking, *IEEE Trans. on Veh. Tech.*, 51(6):1469-1478, 2002

[13] I.Akyildiz and W.wong, A dynamic location management scheme for next generation multityre PCS systems, *IEEE Trans. on wireless communications*, 1(1):178-189, 2002

[14] D.Hong, and S.S.Rappaport, Traffic model and performance analysis for cellular mobile radio telephone systems with priority and non polarized handoff procedures, *IEEE Trans. on Veh. Tech.*, 35(3):77-92, 1986

[15] N.C.Phuong, S.H.Lee and J.M.Moon, *Priority-based call admission control of multi classes in mobile networks*, ICACT, pp.1471-1474, 2006.

[16] K.W.Ross, *Multiservice loss models for broadband telecommunication networks*, Newyork: Springer-verlag, 1995.

[17] P/Tran-Gia and F.Hubner, An analysis of trunk reservation and gradeoff service balancing mechanism in broadband networks, in *Proc.TC6 Task group/WG6.4. Int'l. workshop*, pp.83-97, 1993

[18] Koza.J.R, *Genetic programming II automatic discovery of reusable programs*, the MIT Press 1994.

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 2 (June), pp. 231-240

The Effect of Heterogeneous Traffic Distributions on Load Balancing in Mobile Communications: An Analytical Model

K.-C. Chu, C.-S. Wang, W.-W. Jiang, N.-C. Hsieh

Kuo-Chung Chu, Wey-Wen Jiang, Nan-Chen Hsieh
Department of Information Management,
National Taipei University of Nursing and Health Sciences
No.365, Mingde Rd., Beitou Dist.,
Taipei City 11219, Taiwan (R.O.C.)
{kcchu, jiang6, nchsie})@ntunhs.edu.tw

Chun-Sheng Wang
Department of Information Management,
Jinwen University of Science and Technology
No.99, Anzhong Rd., Xindian Dist.,
New Taipei City 23154, Taiwan (R.O.C.)
seanwang@just.edu.tw

Abstract: This paper investigates the load balancing problem in an environment of heterogeneous traffic distributions. An analytical model is proposed to determine the effect of heterogeneous traffic distributions on load balancing, in which a generic measure of load balancing level (LBL) that is a function of traffic type coefficient (TYC) and call blocking probability of cells is to analyze the expected level of the load balance. We consider both voice traffic and data traffic to determine which kind of traffic has the greater effect. The performance of cellular systems with sectorization is evaluated; they are normal case (N) of homogeneous distribution and linear case (L) of heterogeneous distribution. The analysis results indicate that the TYC has a significant effect on the accommodation capacity, in which voice calls outperform data calls because the LBL can easily distinguish between normal and linear distributions. Load balancing can be achieved more easily for voice only traffic than for data only traffic. Sectorization is more effective in achieving load balancing in the scenario of the heavier loads than in the lighter loads. The paper results are useful for network planning to optimize the channel allocation for different traffic type's distribution.
Keywords: analytical model, heterogeneous distributions, load balancing, mobile communications, QoS.

1 Introduction

Several studies have evaluated the capacity of mobile cellular systems, but most of them (e.g., [1,2]) assume a homogeneous spatial traffic distribution, as it best fits the system's characteristics to have all signals share all the spectral resources However, homogeneous traffic distribution among base station (BS)/sectors/cells (equal cell loads) is very uncommon in practice. Even though sufficient capacity is planned in a cellular system, heterogeneous traffic distribution may occur in other cells, creating a "hot spot" that exceeds the pre-determined capacity and introduces a large blocking probability, as the quality of service (QoS) of such cells may be degraded, especially below a pre-defined threshold. This is why load balancing is the most important issue to be discussed before network planning in terms of optimal resource allocation.

In [3], Ning et al. discuss load balancing by using a hybrid scheme of channel borrowing scheme and load transfer, it allows borrowing channels from light load cells, and ongoing calls can be transferred from heavy load cells into the overlapping cells they are light load. To improve global resource utilization and reduce regional congestion given heterogeneous arrivals, [4] requires load balancing among multiple cells. However, their works cannot be applied to general system because a lot of issues are different from other systems (WCDMA, CDMA2000, HSPA), e.g. channel definition, interferences, soft handoff. In general, soft handoff enforced by power control has been proposed as a possible solution to local traffic imbalances among cells [5]. Actually, power control is one of the most important processes, as interference is the predominant factor that affects the capacity and signal-to-interference ratio (SIR). To maximize system capacity, power control can be used efficiently to adapt cell sizes for load balancing; however, the trade-off between coverage and capacity should be carefully considered [6, 7]. An adaptive load-shedding scheme combines the power control and the soft handoff function to force some mobile stations (MSs) farthest from the cell to enter forced soft handoff, and transfer their traffic load to neighboring cells that are lightly loaded. In this way, heavily loaded cells dynamically down-size their coverage area in order to handle traffic, while adjacent cells that are less heavily loaded increase their coverage to accommodate the extra traffic. However, in a hot-spot sector, powering up all MSs in the sector results in excessive interference with the MSs in neighboring cells, so they cannot maintain sufficient SIR levels at their sector sites. Previous studies also attempt to achieve constant received mean power from each MS within a sector [8–10].

This paper investigates the load balancing problem in an environment of heterogeneous traffic distributions. An analytical model is proposed to determine the effect of heterogeneous traffic distributions on load balancing, in which a generic measure of load balancing level (LBL) that is a function of traffic type coefficient (TYC) and call blocking probability of cells is to analyze the expected level of the load balance. We consider both voice traffic and data traffic to determine which kind of traffic has the greater effect. The performance of cellular systems with sectorization is evaluated; they are normal case (N) of homogeneous distribution and linear case (L) of heterogeneous distribution.

The remainder of this paper is organized as follows. In Section 2, we discuss the mobile cellular system background. In Section 3, we present an analytical model of load balancing, and define SIR. Section 4 details the numerical results, and Section 5 contains some concluding remarks.

2 Background of Mobile Cellular Systems

2.1 Sectorization

Generally speaking, a BS configuration is uniformly sectorized in one sector (with omni-directional antenna, 360° per sector), in three sectors (120° per sector), and in six sectors (60° per sector). The capacity of each sector is calculated subject to the system's SIR requirements. A sector that is lightly loaded usually experiences more interference than a heavily loaded sector, which leads to a higher blocking probability in the lightly loaded sector. Denote B as a set of BSs, and S as a set of sectors configured in the BS. We further denote the sector s in BS j as sector_{js} ($\forall s \in S$, $j \in B$) and K as the set of sector configurations. In this paper, the following two probable configurations are given for a BS ($|K| = 2$): a single sector configuration with an omni-directional antenna, (360° per sector), and a three-sector configuration (120° per sector); k is assigned as the identification (ID) for each configuration. The sector ID i identifies the sector in the configuration k in an anti-clockwise direction. Table 1 summarizes the sector candidates

Table 1: Sector candidates for the BS

The value of S	Candidate $s_{k,i}$	Configuration I.D. (k)	Sector I.D. (i)
$S=1$	$S(1,1)$	1	1
$S=2$	$S(2,1)$	2	1
$S=3$	$S(2,2)$	2	2
$S=4$	$S(2,3)$	2	3

Table 2: Coverage of candidate sectors

Candidate $S_{k,i}$	Sector I.D.i	Coverage of $S_{k,i}$
1	1	$(\phi, \phi+360°)$
2	1	$(\phi, \phi+120°)$
3	2	$(\phi+120°, \phi+240°)$
4	3	$(\phi+240°, \phi+360°)$

for each BS for a combination of k and i. Let S be the set of sectors; then, each sector $s_{k,i}$ ($\forall s_{k,i} \in S$) is defined by the sector configuration (k) and the sector ID (i). Table 2 details the coverage (in degrees) of each sector, where ϕ is the degree of the baseline. In general, it can be assigned arbitrarily, but in this paper we given $-30°$ in our cellular structure example, as shown in Figure 1.

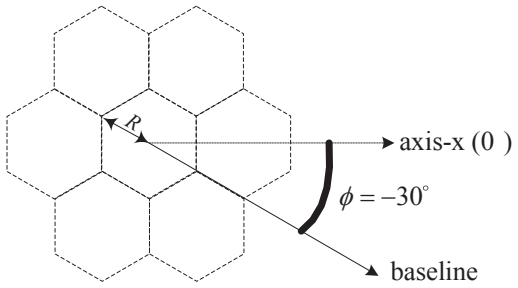


Figure 1: The baseline deployed in a cellular structure

2.2 Interference between sectors

To calculate the interference between sectors, the sector configuration information in Table 1 and the sector coverage information in Table 2 must be given. Without loss of generality, sector $s_{k,i}(s_{k\prime,i\prime})$ is replaced by $s(s\prime)$; and sector s in BS j is denoted by sector$_{js}$, as shown in Figure 2. Because of the baseline degree deployed in all cells is the same using $\phi = -30°$, no matter what the BS is configured, the mutual interference between sectors can be well-known. If we define the interference indicator functions $\Omega^{UL}_{js j\prime s\prime}$ and $\Omega^{DL}_{js j\prime s\prime}$ for the respective uplink (UL) and downlink (DL) connections between sector$_{js}$ and sector$_{j\prime s\prime}$, they can be pre-calculated. To pre-calculate the indicator functions, the sector candidates to be configured in the BS must be defined. Assuming (x_{js}, y_{js}) and $(x_{j\prime s\prime}, y_{j\prime s\prime})$ are the respective locations of BS j and BS $j\prime$, the vectors $\overrightarrow{A_1}(\overrightarrow{A\prime_1})$ and $\overrightarrow{A_2}(\overrightarrow{A\prime_2})$ covering sector$_{js}$(sector$_{j\prime s\prime}$) are defined as follows:

$$\overrightarrow{A_1} = \left[x^1_{js} - x_{js}, y^1_{js} - y_{js}\right], \overrightarrow{A_2} = \left[x^2_{js} - x_{js}, y^2_{js} - y_{js}\right]$$

$$\overrightarrow{A\prime_1} = \left[x^1_{j\prime s\prime} - x_{j\prime s\prime}, y^1_{j\prime s\prime} - y_{j\prime s\prime}\right], \overrightarrow{A\prime_2} = \left[x^2_{j\prime s\prime} - x_{j\prime s\prime}, y^2_{j\prime s\prime} - y_{j\prime s\prime}\right]$$

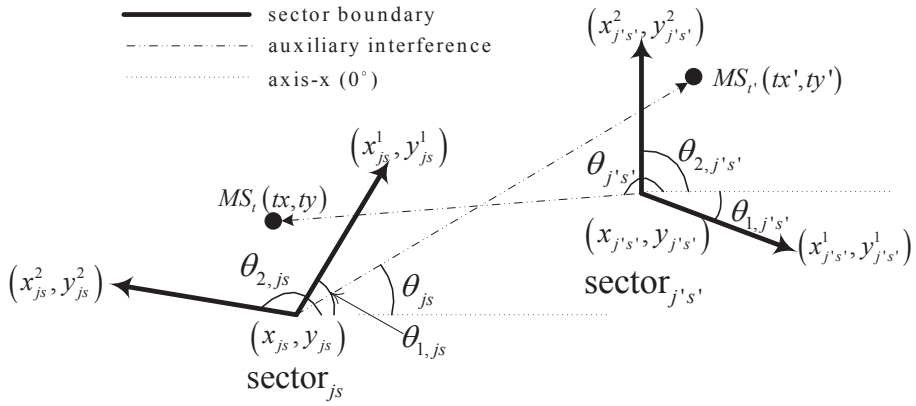


Figure 2: Mutual interference between sectors

where

$$x^1_{js} = R_j \cos(\theta_{1,js}) + x_{js}, y^1_{js} = R_j \sin(\theta_{1,js}) + y_{js},$$
$$x^2_{js} = R_j \cos(\theta_{2,js}) + x_{js}, y^2_{js} = R_j \sin(\theta_{2,js}) + y_{js}$$

$$x^1_{j\prime s\prime} = R_{j\prime} \cos(\theta_{1,j\prime s\prime}) + x_{j\prime s\prime}, y^1_{j\prime s\prime} = R_{j\prime} \sin(\theta_{1,j\prime s\prime}) + y_{j\prime s\prime},$$
$$x^2_{j\prime s\prime} = R_{j\prime} \cos(\theta_{2,j\prime s\prime}) + x_{j\prime s\prime}, y^2_{j\prime s\prime} = R_{j\prime} \sin(\theta_{2,j\prime s\prime}) + y_{j\prime s\prime}$$

Then, the DL interference $\Omega^{DL}_{js j\prime s\prime}$ between sector_{js} and $\text{sector}_{j\prime s\prime}$ can be analyzed by an auxiliary vector $\overrightarrow{A_3} = [tx\prime - x_{js}, ty\prime - y_{js}]$, where $(tx\prime, ty\prime)$ is the arbitrary position of MS $t\prime$ ($\forall t\prime \in T$ and T is the set of MSs) serviced by $\text{sector}_{j\prime s\prime}$. Furthermore, $(\theta_{1,js}, \theta_{2,js})$ and $(\theta_{1,j\prime s\prime}, \theta_{2,j\prime s\prime})$ can also be calculated easily. According to $\overrightarrow{A_3}$, θ_{js} is calculated by $\theta_{js} = \tan^{-1}\frac{ty\prime - y_{js}}{tx\prime - x_{js}}, 0 \le \theta_{js} < 360°$.

After calculating θ_{js}, the Algorithm Cal_InterF[2] is applied to calculate $\Omega^{DL}_{js j\prime s\prime}$ Meanwhile, the UL interference $\Omega^{UL}_{js j\prime s\prime}$ between sector_{js} and $\text{sector}_{j\prime s\prime}$ can be analyzed by an auxiliary vector $\overrightarrow{A\prime_3} = [tx - x_{j\prime s\prime}, ty - y_{j\prime s\prime}]$, where (tx, ty) is the arbitrary position of MS $t(\forall t \in T)$ serviced by sector_{js}; $\theta_{j\prime s\prime}$ is calculated by $\theta_{j\prime s\prime} = \tan^{-1}\frac{ty - y_{j\prime s\prime}}{tx - x_{j\prime s\prime}}, 0 \le \theta_{j\prime s\prime} < 360°$. Again, applying Algorithm Cal_InterF to the calculation of UL interference $\Omega^{UL}_{js j\prime s\prime}$.

3 Analytical Model of Load Balancing

3.1 SIR definition

Denote z_{jst} as a decision variable, which is 1 if MS t is admitted by sector_{js} subject to the SIR requirements and 0 otherwise. Assuming the power of both the UL and DL are perfectly controlled, the received power in sector_{js} from MS t with constant value $P^{UL}_{c(t)}$ will be in the same traffic class$-c(t)$ in the UL, and the received power at MS t from sector_{js} with constant value $P^{DL}_{c(t)}$ will be in same traffic class$-c(t)$ in the DL. If D_{jt} is the distance from MS t to sector_{js}, and given an attenuation factor $\tau = 4$ which is the degree to which a beam of radiation has been attenuated, the intra-sector interference on the UL and the DL is given by (1) and (2) respectively, where both $\alpha^{UL}_{c(t)}$ and $\alpha^{DL}_{c(t)}$ are activity factors of traffic class$-c(t)$. The inter-sector

[2]Detailed algorithm procedure is omitted due to the length limitation of the paper. A complete version of the procedure is available upon request.

interference on the UL and the DL is expressed by (3) and (4) respectively, where both $\Omega^{UL}_{j\prime s\prime js}$ and $\Omega^{DL}_{j\prime s\prime js}$ indicate the interference between sectors.

$$I^{UL}_{jst,intra} = \sum\nolimits_{\substack{t\prime\in T\\ t\prime\neq t}} \alpha^{UL}_{c(t\prime)} P^{UL}_{c(t\prime)} z_{jst\prime} \tag{1}$$

$$I^{DL}_{jst,intra} = \sum\nolimits_{\substack{t\prime\in T\\ t\prime\neq t}} \alpha^{DL}_{c(t\prime)} P^{DL}_{c(t\prime)} \left(\frac{D_{jt\prime}}{D_{jt}}\right)^{\tau} z_{jst\prime} \tag{2}$$

$$I^{UL}_{jst,inter} = \sum_{\substack{j\prime\in B\\ j\prime\neq j}} \sum_{\substack{s\prime\in S\\ s\prime\neq s}} \sum_{\substack{t\prime\in T\\ t\prime\neq t}} \Omega^{UL}_{j\prime s\prime js} \alpha^{UL}_{c(t\prime)} P^{UL}_{c(t\prime)} \left(\frac{D_{j\prime t\prime}}{D_{jt\prime}}\right)^{\tau} z_{j\prime s\prime t\prime} \tag{3}$$

$$I^{DL}_{jst,inter} = \sum_{\substack{j\prime\in B\\ j\prime\neq j}} \sum_{\substack{s\prime\in S\\ s\prime\neq s}} \sum_{\substack{t\prime\in T\\ t\prime\neq t}} \Omega^{DL}_{j\prime s\prime js} \alpha^{DL}_{c(t\prime)} P^{DL}_{c(t\prime)} \left(\frac{D_{j\prime t\prime}}{D_{j\prime t}}\right)^{\tau} z_{j\prime s\prime t\prime} \tag{4}$$

$$SIR^{UL}_{js,c(t)} = \frac{W^{UL}}{d^{UL}_{c(t)}} \cdot \frac{P^{UL}_{c(t)} + (1 - z_{jst})V}{(1-\rho^{UL}) I^{UL}_{jst,intra} + I^{UL}_{jst,inter}} \tag{5}$$

$$SIR^{DL}_{js,c(t)} = \frac{W^{DL}}{d^{DL}_{c(t)}} \cdot \frac{P^{DL}_{c(t)} + (1 - z_{jst})V}{(1-\rho^{DL}) I^{DL}_{jst,intra} + I^{DL}_{jst,inter}} \tag{6}$$

Let $W^{UL}(W^{DL})$ be the spectrum allocated to the UL (DL), and $d^{UL}_{c(t)}$ ($d^{DL}_{c(t)}$) be the information rate in the UL (DL). The SIR values $SIR^{UL}_{js,c(t)}$ and $SIR^{DL}_{js,c(t)}$ in the UL and the DL are defined in (5) and (6) respectively, where ρ^{UL} (ρ^{DL}) is the UL (DL) orthogonality factor. Equations (5) and (6) give a very large artificial constant value V in the numerator in order to satisfy the SIR constraints. This is because the SIR value must be larger than a pre-defined threshold, say the bit energy to noise ratio (BENR), if MS t is to be admitted by sector$_{js}(z_{jst} = 1)$; in other words, the constraint BENR $\leqq$ SIR must be satisfied. For example, in the UL in Equation (5), if MS t is to be admitted by sector$_{js}(z_{jst} = 1)$, the SIR value $SIR^{UL}_{js,c(t)}$ is calculated by $(W^{UL}/d^{UL}_{c(t)}) \cdot P^{UL}_{c(t)}/((1-\rho^{UL})I^{UL}_{jst,intra} + I^{UL}_{jst,inter})$ to determine whether the SIR constraint can be satisfied. In contrast, if MS t ($z_{jst} = 0$) is rejected, the SIR value is always larger than BENR (BENR $\ll$ SIR) because the value V is dominant $P^{UL}_{c(t)}$; thus, $SIR^{UL}_{js,c(t)}$ is calculated as a very large value. This implies that the constraint BENR $\leqq$ SIR can be ignored, as it is always satisfied.

3.2 The Analytical Model

In this paper, we consider traffic with multiple classes, and use the Kaufman model [12] as a performance measure to analyze the blocking probability of each traffic class effectively. Assume that M channels are shared by all traffic requirements. Then, for each traffic class$-c(\forall c \in C)$ with distinct channel requirements, the traffic arrival is a stationary Poisson process with mean rate λ; and the channel requirement b is an arbitrary discrete random variable ($Prob\{b = b_c\} = q_c, \forall c \in C$). A call request with channel requirement b_c has a mean holding time of $1/\mu_c$. Thus, traffic with channel requirement b_c is generated in the Poisson arrival process with mean rate $\lambda_c = \lambda q_c$ and the class$-c$ offered load $a_c = \lambda_c/\mu_c$. The blocking probability of traffic class$-c$ is defined in (7) [12], where the distribution of $q(\cdot)$, which is the probability of the total number of channels occupied by the complete sharing policy, satisfies Equation (8) [11], and $q(x) = 0$ for $x < 0$ and $\sum_{j=0}^{M} q(x) = 1$.

$$B^c(a,b) = \sum\nolimits_{i=0}^{b_c-1} q(|M| - i) \quad \forall c \in C \tag{7}$$

$$\sum\nolimits_{c\in C} a_c b_c q\,(j-b_c) = jq(j) \quad j=0,1,\ldots,M \tag{8}$$

To deal with variations in the traffic load, we can seek load balancing with an average value of resource utilization [12]. In order to evaluate the experiment results, we define a diversity function for system load balancing, from which the standard deviation (SD) of the call blocking probability among sectors can be derived. The smaller the SD, the better the balancing results will be. Let $g_{js}^c = \sum_{t\in T} z_{jst}/\mu_{c(t)}$ be the traffic intensity of class$-c$. Then, $g_{js} = \sum_{c\in C} g_{js}^c$ is the aggregate traffic (in Erlangs) in sector$_{js}$, where g_{js} is equivalent to the traffic load a in (7). We also define $m_{js} = \sum_{t\in T} z_{jst} m^{c(t)}$ as the total number of channels allocated in sector$_{js}$, where m_{js} is equivalent to the required channels b in (7), and $m^{c(t)}$ is the number of channels required for traffic class$-c(t)$. The performance measure B_{js}^c (the call blocking probability of traffic class$-c$ in sector$_{js}$) is expressed by (7), where the sub-script js in B_{js} indicates B^c in sector$_{js}$. If we define L_{LB} as the LBL, the load balancing model can be formulated as (9), where $SD\left(B_{js}^c\right)$ is the SD function of B_{js}^c. The model is calculated subject to SIR Constraints (5) and (6).

$$L_{LB} = \sum\nolimits_{j\in B}\sum\nolimits_{s\in S}\sum\nolimits_{c\in C} K^c SD\left(B_{js}^c\right) \quad \forall j\in B, s\in S \tag{9}$$

To assess the impact of different traffic types on load balancing, we denote K^c as a ratio of traffic class$-c$, where $\sum_{c\in C} K^c = 1$. K^c is a traffic type coefficient (TYC) used to analyze the expected level of the load balance. Given two classes of call requests, e.g., voice and data traffic, if $K^v = 1$ and $K^d = 0$, we only investigate the effect of voice traffic on load balancing; however, if $K^v = 0$ and $K^d = 1$, we only investigate the effect of data traffic on load balancing.

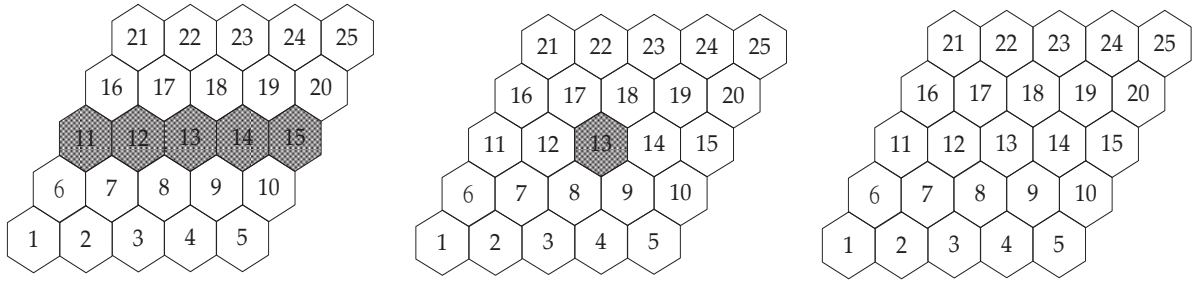
4 Numerical Results

4.1 Parameters

Two heterogeneous traffic distributions are considered in the structure of a 5×5 two-dimensional array with hexagonal cells, and their impact on the system's load balance is compared with that of a homogeneous distribution between sectors, as shown in Figure 3. In the figure, each dark cell has an heterogeneous load that is either heavier or lighter than the load of the normal cells (the light cells). It is assumed that the user density in each cell is homogeneous. Each cell is configured with 3 sectors ($|S| = 3$), and assigned a radius $R_{js} = 5.0$km. The required BENR for voice (v) and data (d) traffic is given by $(E_b/N_{TOTAL})_v^{UL} = (E_b/N_{TOTAL})_v^{DL} =$ 7dB and $(E_b/N_{TOTAL})_v^{DL} = (E_b/N_{TOTAL})_d^{DL} =$ 10dB respectively [13]. The information rates $d_v^{UL} = d_v^{DL} = 9.6$bps, $d_d^{UL} = 19.2$bps, $d_d^{DL} = 38.4$bps [14–16], and the activity factors $\alpha_v^{UL} = \alpha_v^{DL} = \alpha_d^{UL} = \alpha_d^{DL} = 0.5$ [13, 16, 17] are also given. The number of channels required is $m^v = 1, m^d = 4$, and the orthogonality factor is $\rho^{UL} = 0.9, \rho^{DL} = 0.7$ [13]; and the power is perfectly controlled by $P_v^{UL} = 10$dB, $P_v^{DL} = 15$dB, $P_d^{UL} = 15$dB, $P_d^{DL} = 20$dB. The assigned service rate is $\Phi_{js}^v = \Phi_{js}^d = 0.1$ [17, 18].

4.2 Traffic Models

For each sector, call requests for both voice and data calls are generated in the Poisson arrival process with λ_v and λ_d respectively. The mean call holding time is given as $1/\mu_v = 180$(sec), $1/\mu_d = 600$ (sec) [18]. Denote $(E_b/N_{TOTAL})_{c(t)}^{UL} \le SIR_{js,c(t)}^{UL}$ and $(E_b/N_{TOTAL})_{c(t)}^{DL} \le SIR_{js,c(t)}^{DL}$ as the QoS requirements of the UL and DL respectively. All traffic calculated in g_{js} must satisfy the QoS requirements and the condition $z_{jst} D_{jt} \le R_{j_s}\delta_{jst}$, where δ_{jst} is the indicator function if MS t is in the coverage of sector$_{js}$. Power is perfectly controlled in both the UL and the



(a) Heterogeneous linear model (b) Heterogeneous hot spot model (c) Uniform model

Figure 3: Traffic distribution scenarios

DL, and soft handoff is not taken into account. The traffic distributions considered in this work are uniform (U), hot spot (H), and linear (L), as shown in Figure 3. Recall that the cells with heterogeneous loads in Figure 3 (a) and Figure 3 (b) have either heavier or lighter loads than normal cells in a homogeneous distribution. To evaluate the heterogeneous scenario, we introduce two traffic models. We denote heterogeneous cells with heavier loads as M1 and heterogeneous cells with lighter loads as M2. If normal cells are given traffic arrivals λ_c for traffic class$-c$, arrivals in heterogeneous cells are assigned 200% of λ_c in M1, and 50% of λ_c in M2. Thus, the level of load balance for traffic with multiple classes can be evaluated effectively in a near-realistic environment.

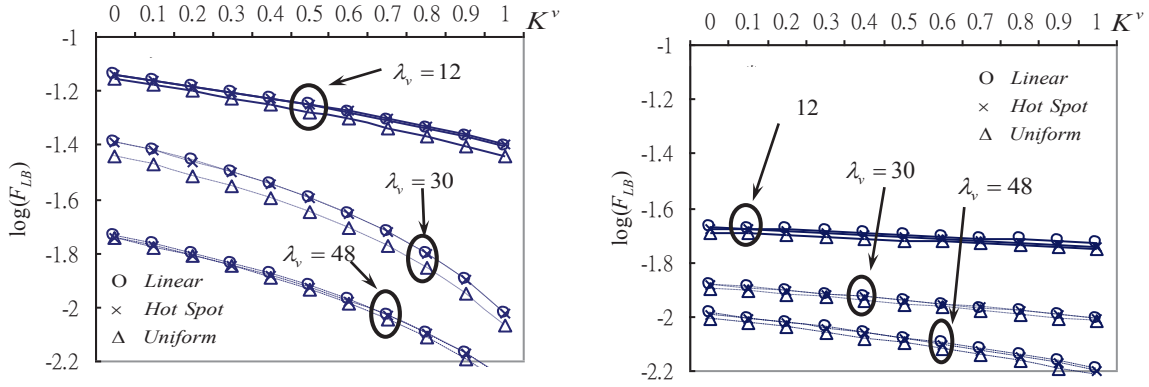
4.3 Analysis

Without loss of generality, the level of load balance is represented in logarithmic form $\log(F_{LB})$. If a smaller value of $\log(F_{LB})$ is calculated, a better level of load balance will be achieved. In Figure 4 (a), no matter what the distribution (linear, hot spot, or uniform) and the offered voice arrivals λ_v are, $\log(F_{LB})$ is a decreasing function of K^v. This implies that load balancing can be achieved more easily for voice only traffic than for data only traffic. If only data traffic is considered, given $K^v = 0$ for all distributions, $\log(F_{LB})$ is nearly -1.75, whereas $\log(F_{LB})$ is nearly -2.4 if only voice traffic is considered ($K^v = 1$). With regard to the effect of traffic intensity, it is easier to achieve load balancing with more offered voice traffic than less offered traffic. For example, in Figure 4 (a), given $K^v = 0.5$ with $\lambda_d = 6$, $\log(F_{LB})$ calculates $(-1.2, -1.6, -1.9)$ for arrivals (λ_v =12, 30, 48). Again, given $\lambda_v = 12$ in Figure 4 (a), $\log(F_{LB})$ is in the range -1.15 to -1.4 in Figure 4 (a), but it is in the range -1.7 to -1.75 in Figure 4(b) with $\lambda_d = 24$.

For the traffic models (M1 vs. M2), there is no significant difference in the load balance with $\lambda_v = 12$ and $\lambda_d = 6$ in both Figure 4 (a) and Figure 5 (a). However, given $\lambda_d = 6$ in Figure 5 (a), the load balance level varies in heavily loaded voice traffic ($\lambda_v = 30, 48$). In another case, given $\lambda_d = 24$ in Figure 5 (b), $\log(F_{LB})$ calculates the same results for variations in the load balance. From the analysis, we conclude that the level of load balancing is more stable in M1 than in M2. A better scheme is needed to handle load balancing in cases of heterogeneous cells with light traffic loads.

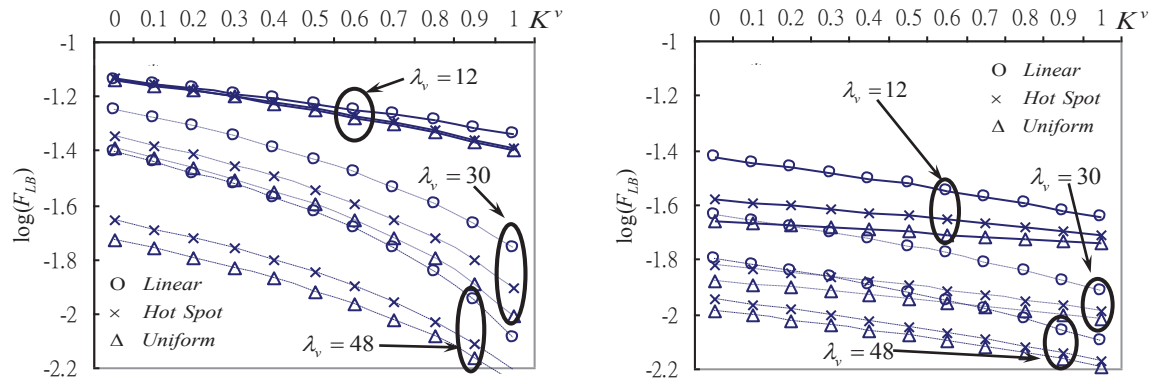
5 Conclusion

In this paper, we propose a load balancing model to deal with the ever-increasing number of heterogeneous distributions in mobile wireless communication systems. We have studied the



(a) $\lambda_d = 6$ (b) $\lambda_d = 24$

Figure 4: BLF as a function of BLC K^v with respect to λ_v, given traffic model M1 and $|S| = 3$



(a) $\lambda_d = 6$ (b) $\lambda_d = 24$

Figure 5: BLF as a function of BLC K^v with respect to λ_v, given traffic model M2 and $|S| = 3$

effect of heterogeneous traffic distributions on load balancing, as well as the effect of sectorization. The numerical results indicate that the level of load balancing is affected by spatial traffic distributions, especially by lighter loads in heterogeneous cells (the M2 model). In the scenario of heterogeneous cells with heavier loads (the M1 model), the level of load balancing has similar values of $\log(F_{LB})$ in three distributions, i.e., the linear, hot spot, and uniform models. Sectorization is more effective in achieving load balancing in the scenario of the heavier loads than in the lighter loads. To achieve load balancing as well as capacity maximization in a system with heterogeneous distributions, a hybrid FMDA/CDMA scheme can be utilized. Usually, available wideband spectrum can be divided into a number of subspectra with smaller bandwidths; each of them is further deployed by CDMA technique. Each subspectrum employs direct sequence spectrum spreading with reduced processing gain, which is transmitted in one and only one subspectrum. The scheme moderately mitigates interference by allocating an appropriate subspectrum in each cell. The results of this work are useful for network planning to optimize the channel allocation for different traffic type's distribution.

Bibliography

[1] J.-S. Wu and J.-K. Chung, Analysis of uplink and downlink capacities for two-tier cellular system, *IEEE Proc.-Communications.*, 144(6):405-411, 1997

[2] J.C. Liberti, T. S. Rappaport, Analytical Results for Capacity Improvements in CDMA, *IEEE Transactions on Vehicular Technology*, vol. 43, pp. 680-690, Sep. 1994

[3] G. Ning, G. Zhu, L. Peng, and X. Lu, Research on hybrid dynamic load balancing algorithm in heterogeneous hierarchical wireless networks, *Journal on Communication*, 28(1):75-81+86, 2007

[4] A. Sang, X. Wang, M. Madihian, and R. D. Gitlin, Coordinated load balancing, handoff/cell-site selection, and scheduling in multi-cell packet data systems, *Wireless Networks*, 14(1):103-120, 2008

[5] X.H. Chen, Adaptive traffic-load shedding and its capacity gain in CDMA cellular systems, *in Proc. IEE Communications*, pp. 186-192, 1995

[6] V.V. Veeravalli and A. Sendonaris, The coverage-capacity tradeoff in cellular CDMA systems, *IEEE Transactions on Vehicular Technology*, 48(5):1443-1450, 1999

[7] G. Hampel, K.L. Clarkson, J.D. Hobby, and P.A. Polakos, The tradeoff between coverage and capacity in dynamic optimization of 3G cellular networks, *in Proc. IEEE VTC-Fall*, vol. 2, pp. 927-932, 2003

[8] W.-M. Tam and F.C.M. Lau, Analysis of power control and its imperfections in CDMA cellular systems, *IEEE Transactions on Vehicular Technology*, 48(5):1706-1717, 1999

[9] X.H. Chen and K.L. Lee, A novel adaptive traffic load shedding scheme for CDMA cellular mobile systems, *in Proc. IEEE ICCS*, vol. 2, pp. 566-570, 1994

[10] X.H. Chen, Adaptive traffic-load shedding and its capacity gain in CDMA cellular mobile systems, *IEE Proceedings-Communications*, 142(3):186-192, 1995

[11] J.S. Kaufman, Blocking in a shared resource environment, *IEEE Transactions on Communications*, vol. 29, pp. 1474-1481, 1981

[12] A. He, Performance comparison of load balancing methods in multiple carrier CDMA systems, *in Proc. IEEE PIMRC*, pp. 113-118, 2000

[13] S.W. Kim, D.G. Jeong, W.S. Jeon, and C.-H. Choi, Forward link performance of combined soft and hard handoff in multimedia CDMA systems, *IEICE Transactions on Communications*, vol. E85-B, no.7, pp. 1276-1282, July 2002

[14] W. Choi and J.Y. Kim, Forward-link capacity of a DS-CDMA system with mixed multirate sources, *IEEE Transactions on Vehicular Technology*, 50(3):737-749, 2001

[15] D. Kim and D.G. Jeong, Capacity unbalance between uplink and downlink in spectrally overlaid narrow-band and wide-band CDMA mobile systems, *IEEE Transactions on Vehicular Technology*, 49(4):1086-1093, 2000

[16] W.S. Jeon and D.G. Jeong, Call admission control for mobile multimedia communications with traffic asymmetry between uplink and downlink, *IEEE Transactions on Vehicular Technology*, 50(1):59-66, 2001

[17] W.S. Jeon and D.G. Jeong, Call admission control for CDMA mobile communications systems supporting multimedia services, *IEEE Transactions on Wireless Communications*, 1(4):649-659, 2002

[18] K. Kim and Y. Han, A call admission control scheme for multi-rate traffic based on total received power, *IEICE Transactions on Communications*, vol. E84-B, no. 3, pp. 457-463, March 2001

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 2 (June), pp. 241-251

Network Element Scheduling for Achieving Energy-Aware Data Center Networks

W. Fang, X. Liang, Y. Sun, A.V. Vasilakos

Weiwei Fang, Xiangmin Liang, Yantao Sun
School of Computer and Information Technology
Beijing Jiaotong University, Beijing 100044, China
{wwfang, 11125116, ytsun}@bjtu.edu.cn

Athanasios V. Vasilakos
Department of Computer and Telecommunications Engineering
University of Western Macedonia
Kozani GR 50100, Greece
vasilako@ath.forthnet.gr

Abstract: The goal of data center network is to interconnect a massive number of servers so as to provide reliable and scalable computing and storage infrastructure for cloud-based Internet services and data-intensive scientific applications. Recent studies reveal that the network elements consume 10∼20% of the overall power in a data center, which has introduced a challenge to reducing network energy cost without adversely affecting network performance. Considering unique features of traffic patterns and network topologies in data centers, this paper proposes a novel Network Element Scheduling Scheme (NESS) to reduce data center energy consumption from the networking perspective. The core idea is to turn on only a minimal subset of network elements to satisfy routing requirements, and put to sleep or shut down the rest unneeded ones for energy saving. In NESS, the logical network architecture formed by the active elements not only achieves the basic purpose for server interconnections in data centers, but also can support multi-path routing between pairs of hot servers for load balancing. Simulation experiments are performed in representative data center network topologies, and the results demonstrate the effectiveness of NESS in energy conserving on network elements in data centers.

Keywords: data center networks, green computing, energy aware, Steiner tree.

1 Introduction

In recent years, many large data centers are built around the world to provide highly reliable and scalable infrastructure for cloud services and scientific computations. As the actual tendency is to exponentially increase the number of demanded servers, a natural consequence is that the power consumption becomes a critical concern for data center operators. For example, it has been reported that data center power usage in U.S. doubled between 2000 and 2006 to nearly 61 billion kilowatt-hours, representing 1.5% of all U.S. electricity consumption [1]. Researchers are now seeking to find effective solutions to make data centers reduce power consumption while keep the desired service performance. Most of the recent research has focused on reducing the two major components of data center power usage: servers and cooling [2]. However, the underlying network infrastructure, namely routers, switches, high-speed links, still lacks effective energy management solutions. The networking part of data center has been found to consume 10∼20% of its total power consumption [2], and thus should not be neglected.

Fortunately, there is an opportunity for substantial reductions in the energy consumption of data center networks due to two factors. On one hand, the high network capacity of data center networks is specially provisioned for worst-case or busy-hour load, and far from being exceeded by traffic load most of time. Moreover, data center traffic varies considerably over time exhibiting temporal patterns (e.g., daily, weekly, monthly and yearly [3]), and over location exhibiting spatial patterns (e.g., hot/non-hot server racks [4]). In the data center with rich link connectivity, a great number of network elements may consequently work in idle state. On the other hand, today's network elements are not energy proportional, since fixed overheads such as fans, switch chips and transceivers waste power at low loads. It has been found that the energy consumption of common networking devices at the idle state still accounts for more than 85% of that at the working state [5]. The implication of these factors is that significant amount of power energy is wasted on idle elements in the data center network.

To address the challenges stated above, this paper proposes a network element scheduling scheme (NESS) that acts as a network-wide energy optimizer for data center networks. It selects a subset of network elements that must stay active to meet traffic routing requirements, and then puts as many unneeded routers, switches and links as possible into dormant mode. Considering unique features of network topologies and traffic patterns in data centers, NESS achieves this goal in two steps. Firstly, NESS models and solves the basic problem of network element scheduling for guaranteeing server interconnection and traffic routing by using the Steiner tree framework. On the basis of the initial selection result, NESS then additionally activates as few unselected network elements as possible to support different degrees of multi-path routing between server pairs. Network elements not involved in the routing service are finally powered off or put into sleep state. We have conducted extensive simulations in typical architecture models of data center network to illuminate the effectiveness and performance of the proposed scheme.

The rest of this paper is organized as follows. Section 2 presents related works. Section 3 describes the NESS scheme in detail. Section 4 explores on experiment results, and finally the paper is concluded in Section 5.

2 Related Works

The issue considered in this paper involves two network-related research directions, data center networking and green networking. The rest of this section presents the state-of-art related to these two directions.

The research on data center networking mainly focuses on how to implement a network infrastructure that achieves the following goals [2] [6]: (1) it must be scalable to an increasing number of servers; (2) it must be fault tolerant against various types of hardware failures; (3) it must be able to provide high network capacity; (4) it must achieve high utilization and be cost efficient. Due to the limitations of the conventional tree-based architecture [7], a number of novel network architectures for data center networks have been proposed recently, which can be roughly divided into two categories. One is the switch-centric architecture, which organizes switches into structures rather than trees and puts interconnection intelligence on switches, such as Fat-Tree [8], VL2 [7] and Portland [9]. The other is the server-centric architecture, which puts interconnection intelligence on servers and uses switches only as cross-bars, such as BCube [10], FiConn [11] and DCell [6]. Accordingly, each of the architecture proposals has its own solution for node addressing and traffic routing [6] [8]. Furthermore, multi-path routing [6] [8] [12] [13] has been exploited for load balancing in data center networks. Other complementary research for data center networking has focused on TCP incast problem [14], traffic-aware virtual machine migration [15], switch design [16] [17] or cost efficiency [2] [3], etc.

Reduction of unnecessary energy consumption, referred to as "green networking", has become

a major concern in wired/wireless networking, because of the potential economical benefits and the expected environmental impact. [18] is a pioneer work on this topic, and the authors of [18] suggested putting network elements to sleep for saving energy in local area network in a later paper [19]. Additionally, link rate adaptation is also employed to reduce energy consumption in Ethernet [20]. Based on these techniques, the IEEE 802.3az Energy Efficient Ethernet task force proposes the Low Power Idle solution to reduce power consumption of Ethernet devices [21]. ElasticTree [3] is a pioneer work that optimizes the energy consumption of data center networks by turning off unnecessary links and switches during off-peak hours. It models the problem based on the multi-commodity flow model, in which some parameters, e.g., server traffic demand, are difficult to be accurately obtained in practice. Besides, ElasticTree focuses on only tree-based topologies such as FatTree. Another similar work [22] models the same problem as a 0-1 Knapsack model, and proposes a heuristic based solution. However, it doesn't support multi-path routing between server pairs for load balancing. VMFlow [23] is a recent work on how to migrate virtual machines among data center servers to minimize the amount of serving network elements while satisfying a large fraction of the network traffic demands.

3 NESS: Network Element Scheduling Scheme

3.1 Research Motivations

A data center network is typically provisioned for peak traffic load, and run well below capacity most of the time. From previous researches, we can discover the two important characteristics (i.e., temporal and spatial) of data center traffics, both of which can help us to define the design goals.

On the one hand, network traffic and its temporal dynamics implicitly reflect the behavioural pattern of end users for whom the data center provide services [3]. For example, traffic may vary daily (e.g., more email exchanging during the day), weekly (e.g., more enterprise data processing on weekdays), monthly (e.g., more multimedia file sharing on holidays), and yearly (e.g., more e-shopping and e-payment in December). Rare events like cable breaks or breaking news may hit the peak capacity, but most of the time data center traffic follows the temporal pattern and actually can be satisfied by a subset of active network elements [3].

On the other hand, physical servers in the data center are commonly organized into racks and richly-interconnected by a number of links, switches and routers. The typical upper-layer applications usually generate a traffic demand with only a few of server racks being hot (i.e., sending or receiving a large volume of traffic) [4]. Moreover, the hot racks generally exchange much of their data with only a few other racks. Such a spatial pattern of data center traffic is determined by the role and tasks of servers in the current application. To avoid network congestion in data centers [4], it is necessary to further provide different degrees of multipath routing for the flows from/to hot servers.

Motivated by the analysis above, we propose NESS, a network element scheduling scheme that acts as a network-wide energy optimizer for data center networks. The following subsections will present the key issues on its design and implementation in detail.

3.2 Design Details

In data center networks, the physical servers are interconnected by a number of high-speed links, switches and even routers. The data center network topology can be modelled as a simple undirected weighted graph *G(V, E)* with equal edge weights [15], where V is the set of vertices and $E \subseteq V \times V$ is the set of edges. There are two types of vertices in V: the servers and the

networking devices (i.e., switches and routers). The sets of them can be denoted by V_s and V_d respectively. Therefore, $V = V_s \cup V_d$. The edge $e \in E$ represents a communication link between a server and a networking device, or between a pair of networking devices. As an illustration, we show four typical data center topologies in Figure 1, namely 2N-Tree [3], VL2 [15], Fat-Tree and BCube. It must be noted that in the first three architectures $V_s \cap V_d = \emptyset$, while in BCube $V_s \subset V_d$.

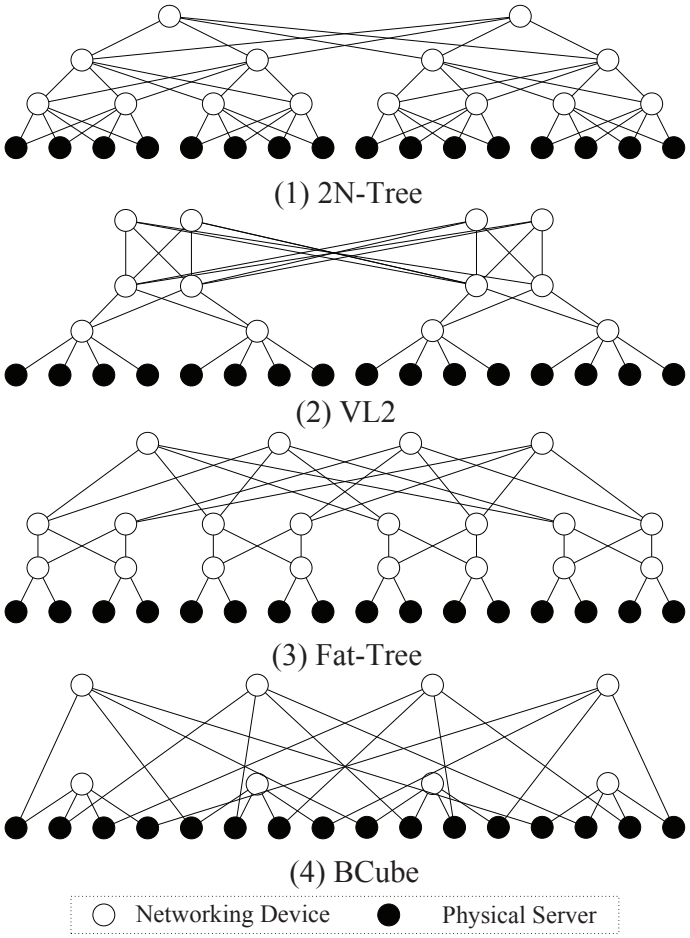


Figure 1: Illustration of state-of-the-art data center topologies.

Before problem formulation, we first define the following notation: $|X|$ denotes the cardinality of set X, $f_{i,j}$ denotes the flow from server i to server j, $p_{i,j}$ denotes the path having i and j as its ends, and $L_{i,j}$ denotes the physically upper bound for total routing paths. Now, the network element scheduling problem can be formally formulated as follows:

Minimize:

$$|V_d'| \tag{1}$$

Subject to:

$$G'(V', E') \subseteq G(V, E) \tag{2}$$

$$V' = V_d' \cup V_s' \tag{3}$$

$$V_d' \subseteq V_d \tag{4}$$

$$V_s' = \{i \in V_s \mid \sum_{\forall j \in V_s} (f_{i,j} + f_{j,i}) > 0\} \tag{5}$$

$$|\{p_{i,j} \mid i,j \in V'_s, \forall k \in p_{i,j}, k \in V'_d\}| \propto \sum (f_{i,j} + f_{j,i}) \tag{6}$$

$$1 \leq |\{p_{i,j} \mid i,j \in V'_s, \forall k \in p_{i,j}, k \in V'_d\}| \leq L_{i,j} \tag{7}$$

In the above formulation, the term of the objective represents the set of active networking devices for interconnections of servers having incoming and/or outgoing traffic. Formula (5)∼(7) guarantee different degrees of multipath routing between pairs of servers can be provided according to traffic requirements. Moreover, Formula (2) implies that unneeded idle links can actually be turned off to further reduce energy consumption [3] [5] since $E' \subseteq E$.

Such a scheduling problem can be easily proved to be NP-complete by restricting that only one path is established between any server pair (i, j). In the equal-edge-weight graph $G(V, E)$, the objective of minimizing $|V'_d|$ is equal to minimizing $|E'|$, and thereby is equal to determining the minimal-weight connected sub-graph $G'(V', E')$ spanning V'_s. This is identical to the Sterner tree problem in the equal-edge-weight graph, which is a NP-complete problem [24]. Therefore, the problem above is NP-complete. Because by now no polynomial-time algorithm is available to obtain the optimal solution of a NP-complete problem, we propose NESS, a scheme that solves the scheduling problem modelled above heuristically in two steps.

In the first step, NESS solves the Steiner tree problem on graph $G(V, E)$ with the terminal node set V'_s. A lot of heuristic algorithms can be used to obtain the results, among which MPH (Minimum Path Heuristic) is one of the best-known solution [24] [25]. The output vertex set V'_{d1} is a subset of V_d, representing the minimal number of networking devices that can guarantee the basic purpose for server interconnection and traffic routing.

In the second step, NESS selects from $V_d \backslash V'_{d1}$ as few vertices as possible to construct multiple path between some vertices in V'_s that represents the hot servers. To achieve this goal, NESS integrates the well-known ECMP (Equal-Cost Multiple-Path) routing mechanism [12] for discovering multiple paths connecting hot servers, and then chooses the paths containing more vertices already in the set V'_{d1}. Denoting the set of newly selected vertices in this step as V'_{d2}, we finally have $V'_d = V'_{d1} \cup V'_{d2}$.

To illustrate the approach stated above, we take an example in the Fat-Tree topology shown in Figure 2. Assume $V'_s = \{v_{21}, v_{25}, v_{27}, v_{28}\}$, and v_{21}, v_{28} are hot servers requiring two routing paths to be established between them. In the first step, NESS executes the MPH algorithm to solve the Steiner tree problem, and obtains $V'_{d1} = \{v_1, v_5, v_7, v_{13}, v_{15}, v_{16}\}$. In the second step, NESS can discover totally four equal-cost shortest paths through a basic version of ECMP [8] as follows:

$$path1 : v_{21} \to v_{13} \to v_5 \to v_1 \to v_7 \to v_{16} \to v_{28}$$
$$path2 : v_{21} \to v_{13} \to v_5 \to v_2 \to v_7 \to v_{16} \to v_{28}$$
$$path3 : v_{21} \to v_{13} \to v_6 \to v_3 \to v_8 \to v_{16} \to v_{28}$$
$$path4 : v_{21} \to v_{13} \to v_6 \to v_4 \to v_8 \to v_{16} \to v_{28}$$

Then we can choose path1 and path2 for v_{21}, v_{28} since path2 requires to additionally activate only v_2, while path3 and path4 requires to activate v_6, v_3, v_8 and v_6, v_4, v_8 respectively. Therefore, $V'_{d2} = \{v_2\}$.

3.3 Implementation Issues

As a scheduling software, NESS consists of four logical modules, i.e., network analyzer, network scheduler, power controller and route controller. The network analyzer obtains network topology and traffic requirement by performing statistics and analysis to the collected running data of the data center network. The role of network scheduler is to find the minimum network subset that can satisfy current traffic according to the approach stated in Section 3.2. With

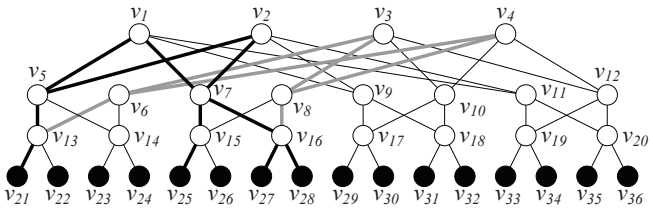


Figure 2: An Fat-Tree example for illustrating network element scheduling in NESS.

the input of topology and traffic conditions from the analyzer, the scheduler outputs the set of active elements to power controller and the set of flow routes to route controller. The power controller toggles the power states of different types of network elements (i.e., links, switches and routers), while the route controller checks routing paths for traffic flows and pushes routes into the network.

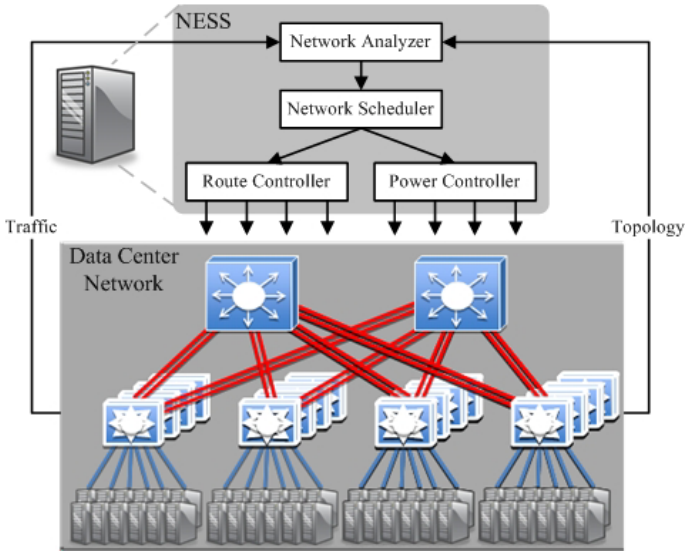


Figure 3: System diagram of NESS.

Similar to ElasticTree [3], NESS can be implemented as a NOX application [26] to run atop a network of OpenFlow switches [27]. OpenFlow is an open standard for commercial switches and routers to enable controlling the forwarding plane by a software running on a separate server, and NOX is an open-source OpenFlow controller that is designed to provide a simplified platform for writing network control software in C++ or Python. The route controller in NESS can be implemented with NOX. Besides, we can leverage the existing mechanisms such as SNMP Set operations and command line interface to support the power control features of NESS. Moreover, the network analyzer can collect the state records of traffic and topology through SNMP Get operations and passive packet tracing [28].

4 Experimental Evaluation

In this section, we evaluate the performance of NESS experimentally, using a custom simulator developed in C++ with the Boost Graph Library [29]. This simulator supports constructing the four types of data center network architectures described in Section 3.2 with a number of 48-port Gigabit Ethernet switches [8] [27]. By using this simulator, we have created data center

communication scenarios with different network sizes (e.g., a network of S=2304/4608/13824 end servers), different traffic conditions (e.g., a network of servers with h=1%/10% hot ones [4] in which each hot server randomly selects about 10% other ones under different racks as its communication counterparts) and different routing requirements (e.g., a pair of hot servers may require 2∼6 equal-cost paths for traffic multiplexing). For a specific scenario, the simulation has been carried out independently for multiple times by network reconstruction, and the results are averaged over these runs. Figure 4 shows the percentage of the dormant network elements in 2N-Tree, VL2, Fat-Tree and BCube respectively.

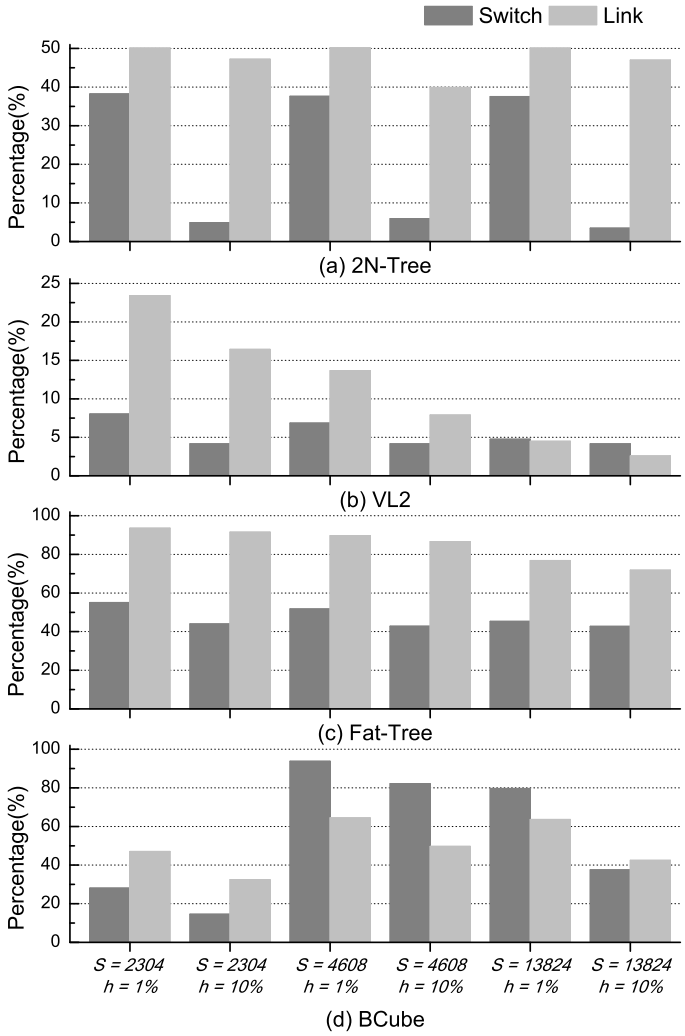


Figure 4: The percentage of dormant network elements in six different networking scenarios.

We have several observations concerning the simulation results:

(1) NESS effectively reduces the total amount of active switches as well as active links to different extents in typical communication scenarios under the four data center network architectures.

(2) All else being equal, the communication scenario with more hot servers generally will have greater demand for network resources, i.e., more network elements have to be kept active by NESS for fulfil routing requirements.

(3) For 2N-Tree, we can find that its dormant percentages of network elements remain essentially flat under the same condition of hot rate h. That's because in 2N-Tree both the total

amount of network elements (i.e., switches or links) and the amount of the ones selected by NESS are approximately in direct proportion to the network size. In general, our simulation results on 2N-Tree (Figure 4 (a)) indicate that we can roughly predict the energy saving for applying NESS to the large-scale data center network built in this architecture through small-scale experiments.

(4) For VL2, we can find that its dormant percentages of both switch and link are the lowest among the four architectures. An examination of simulation traces reveals that the total amount of switches remains the same (i.e., 336) for all the six simulation scenarios, while most of these switches (i.e., 288) and the links from them to the servers have to be kept active. Therefore, the optimization space for NESS is quite limited in VL2 architecture, especially when the hot rate h is relatively high. From the results in Figure 4 (b), the scenario with fewer hot servers has higher dormant percentage of switches, especially when the number of hot servers is less than 230 (i.e., 10% of 2304). Moreover, the dormant percentage of links decreases accordingly along with the increment of network size.

(5) For Fat-Tree, we can find that that it has the relatively highest dormant percentage of links among the four architectures. Besides, the dormant percentage of switches is also relatively high (i.e., about 50%) for each scenario. An examination of simulation traces reveals that Fat-Tree has the most switches (i.e., 2880) and links (i.e., 50000∼70000) among the four architectures. While such redundancy of available network elements makes the fat-tree topology attractive for fault-tolerance, it also brings about considerable energy wastage due to element idle in most cases [3] [22]. Moreover, as that of VL2, all scenarios have the same amount of switches and the one with fewer hot servers will have higher dormant percentage of switches. In all, the simulation results in Figure 4 (c) indicate that there is a tradeoff between fault tolerance and energy consumption for Fat-Tree.

(6) For BCube, the simulation results are significantly different from those of switch-centric architectures, i.e., the dormant percentages increase sharply when S grows lager than 2304. That's because: when S=2304, a 2-level BCube_1 built with totally 96 switches (48 for level 0 and 48 for level 1) is enough. However, a 3-level BCube_2 has to be constructed with 2304 switches for level 2 when S>2304. All these level 2 switches only interconnects two BCube_1 when S=4608 and six BCube_1 when S=13824. Actually, the utilization of these switches are quite low in result of high redundancy, and thus most of them can be put into dormant state. Moreover, it can also be found that the increment of network size under the same condition of BCube level, e.g., from 4608 to 13824, will lead to higher element utilization and lower dormant percentages. In all, the simulation results in Figure 4 (d) indicate that for each BCube level the dormant percentages of network elements will decrease from high to low along with the increment of network size.

Accordingly, Table 1 specifically illustrates to how much extent power energy can be saved by applying NESS to data center networks. The power consumption parameters of network elements are obtained from a previous work [5]. For the three switch-centric architectures above, $P_{VL2}<P_{2N-Tree}<P_{Fat-Tree}$. For the sever-centric architecture BCube, the scenario with a smaller network size at a BCube level (e.g., S=4608 at BCube_2) will have higher reduction percentage of power consumption due to high redundancy of network elements.

5 Conclusion

In this paper, we introduce the energy cost problem in today's data centers. Based on detailed analysis on traffic patterns and network topologies, we propose a novel scheme (NESS) to reduce data center energy consumption from the networking perspective. Extensive simulation results demonstrate that it is possible to switch off idle networking elements, so that the total network power consumption can be reduced without adversely affecting network performance. The work presented in this paper is somewhat preliminary, but shows that energy-aware networking in the

Table 1: Power Consumption Reduction Percentage P.

S	2304		4608		13824	
h	1%	10%	1%	10%	1%	10%
2N-Tree	43.99%	25.44%	43.74%	22.50%	43.68%	24.78%
VL2	10.66%	6.26%	8.59%	5.12%	4.63%	3.43%
Fat-Tree	66.08%	57.59%	62.87%	55.69%	55.56%	52.18%
BCube	37.40%	23.34%	80.87%	78.97%	76.09%	38.69%

data center is promising [30] [31]. In the future, we plan to implement NESS on our prototype system of data center network, which is still under development now.

Acknowledgements

This work was partially supported by the China Postdoctoral Science Foundation (20110490282), the Fundamental Research Funds for the Central Universities of China (2011JBM225) and the Intel China Research Institute - BJTU Collaborative Research Program (K09L00030).

Bibliography

[1] U.S. Enviromental Protection Agency (EPA), Report to Congress on Server and Data Center Energy Efficiency: Public Law 109-431, *EPA ENERGY STAR Program*, August, 2007

[2] A.G. Greenberg, P. Lahiri, D.A. Maltz, P. Patel and S. Sengupta, The Cost of a Cloud: Research Problems in Data Center Networks, *ACM SIGCOMM Computer Communication Review*, 39(1):68-73, 2009

[3] B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee and N. Mckeown, ElasticTree: Saving Energy in Data Center Networks, *Proceedings of the 7th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 249-264, 2010

[4] D. Halperin, S. Kandula, J. Padhye, P. Bahl and D. Wetherall, Augmenting Data Center Networks with Multi-Gigabit Wireless Links, *Proceedings of the 2011 ACM SIGCOMM conference*, 38-49, 2011

[5] A.P. Bianzino, C. Chaudet, F. Larroca, D. Rossi and J.L. Rougier, Energy-Aware Routing: a Reality Check. *Proceedings of the 2010 IEEE GLOBECOM workshops*, 1422-1427, 2010

[6] C.X. Guo, H.T Wu, K. Tan, L. Shi, Y.G. Zhang and S.W. Lu, DCell: A Scalable and Fault-Tolerant Network Structure for Data Centers. *Proceedings of the 2008 ACM SIGCOMM conference*, 75-86, 2008

[7] A. Greenberg, J.R. Hamilton and N. Jain, VL2: A Scalable and Flexible Data Center Network. *Proceedings of the 2009 ACM SIGCOMM conference*, 51-62, 2009

[8] M. Al-Fares, A. Loukissas and A. Vahdat, A Scalable, Commodity Data Center Network Architecture. *Proceedings of the 2008 ACM SIGCOMM conference*, 63-74, 2008

[9] R.N. Mysore, A. Pamboris, N. Farrington, N. Huang, P. Miri, S. Radhakrishnan, V. Subramanya and A. Vahdat, PortLand: A Scalable Fault-Tolerant Layer 2 Data Center Network Fabric. *Proceedings of the 2009 ACM SIGCOMM conference*, 39-50, 2009

[10] C.X. Guo, G.H. Lu, D. Li, H.T. Wu, X. Zhang, Y.F. Shi, C. Tian, Y.G. Zhang and S.W. Lu, BCube: A High Performance, Server-Centric Network Architecture for Modular Data Centers. *Proceedings of the 2009 ACM SIGCOMM conference*, 63-74, 2009

[11] D. Li, C.X. Guo, H.T. Wu, Y.G. Zhang and S.W. Lu, Ficoon: Using Backup Port for Server Interconnection in Data Centers. *Proceedings of the 28th IEEE International Conference on Computer Communications (INFOCOM)*, 2276-2285, 2009

[12] C. Hopps, Analysis of an Equal-Cost Multi-Path Algorithm. *IETF RFC 2992*, November 2000

[13] K. Xi, Y.L. Liu and H.J. Chao, Enabling Flow-based Routing Control in Data Center Networks using Probe and ECMP. *Proceedings of the 30th IEEE International Conference on Computer Communications (INFOCOM) Workshops*, 614-619, 2011

[14] J. Zhang, F.Y. Ren and C. Lin, Modelling and Understanding TCP Incast in Data Center Networks. *Proceedings of the 30th IEEE International Conference on Computer Communications (INFOCOM)*, 1377-1385, 2011

[15] X.Q. Meng, V. Pappas and L. Zhang, Improving the Scalability of Data Center Networks with Traffic-aware Virtual Machine Placement. *Proceedings of the 29th IEEE International Conference on Computer Communications (INFOCOM)*, 1154-1162, 2010

[16] D.A. Joseph, A. Tavakoli and I. Stoica, A Policy-aware Switching Layer for Data Centers. *ACM SIGCOMM Computer Communication Review*, 38(4): 51-62, 2008

[17] N. Farrington, G. Porter, S. Radhakrishnan, H.H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen and A. Vahdat. Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers. *Proceedings of the 2010 ACM SIGCOMM conference*, 339-350, 2010

[18] M. Gupta and S. Singh, Greening of the Internet. *Proceedings of the 2003 ACM SIGCOMM conference*, 19-26, 2003

[19] M. Gupta and S. Singh, Using Low-Power Modes for Energy Conservation in Ethernet LANs. *Proceedings of the 26th IEEE International Conference on Computer Communications (INFOCOM)*, 2451-2455, 2007

[20] C. Gunaratne, K. Christensen, B. Nordman and S. Suen, Reducing the Energy Consumption of Ethernet with Adaptive Link Rate (ALR). *IEEE Transactions on Computers*, 57(4): 448-461, 2008

[21] R. Hays, Active/Idle Toggling with Low-Power Idle. *Presentation for IEEE 802.3az Task Force*, 2008

[22] Y.F. Shang, D. Li and M.W. Xu, Energy-Aware Routing in Data Center Network. *Proceedings of the 1st ACM SIGCOMM Workshop on Green Networking 2010*, 1-8, 2010

[23] V. Mann, A. Kumar, P. Dutta and S. Kalyanaraman, VMFlow: Leveraging VM Mobility to Reduce Network Power Costs in Data Centers. *Lecture Notes in Computer Science Volume 6640 (IFIP Networking 2011)*, 198-211, 2011

[24] M.R. Garey, D.S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness. *W. H. Freeman Publishers Company*, 1979

[25] H. Takahashi, A. Matsuyama, An Approximate Solution for the Steiner Problem in Graphs. *Mathematica Japonicae*, 24: 571-577, 1980

[26] N. Gude, T. Koponen, J. Pettit, B. Pfaff, M. Casado, N. Mckeown, S. Sehnker. NOX: Towards an Operating System for Networks. *ACM SIGCOMM Computer Communication Review*, 38(3): 105-110, 2008

[27] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Perterson, J. Rexford, S. Shenker, J. Turner, OpenFlow: Enabling Innovation in Campus Networks. *ACM SIGCOMM Computer Communication Review*, 38(2): 69-74, 2008

[28] T. Benson, A. Anand, A. Akella, M. Zhang, Understanding Data Center Traffic Characteristics. *ACM SIGCOMM Computer Communication Review*, 40(1), 92-99, 2010.

[29] The Boost Graph Library, *http://www.boost.org/doc/libs/release/libs/graph*.

[30] D. Abts, M.R. Marty, P.M. Welles, P. Klausler, H Liu, Energy Proportional Datacenter Networks. *Proceedings of the 37th ACM/IEEE International Symposium on Computer Architecture (ISCA)*, 338-347, 2010

[31] K. Chen, C.C. Hu, X. Zhang, K. Zheng, Y. Chen, A.V. Vasilakos, Survey on Routing in Data Centers: Insights and Future Directions. *IEEE Network*, 25(4), 6-10, 2011

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 2 (June), pp. 252-263

Prioritization of Traffic for Resource Constrained Delay Tolerant Networks

G. Fathima, R.S.D. Wahidabanu

G. Fathima
Adhiyamaan College of Engg
Hosur. TamilNadu
fathima_ace@yahoo.com

R.S.D. Wahidabanu
Govt. College of Engg
Salem. TamilNadu
drwahidabanu@gmail.com

Abstract: In networks with common shared wireless medium, the available bandwidth is always valuable and often scarce resource. In addition to it, memory available at nodes (eg., sensor nodes) might be limited relative to the amount of information that needs to be stored locally. As Delay Tolerant Networks (DTNs) rely on node mobility for data dissemination, the high node mobility limits the duration of contact. Besides the issue of contact opportunities between nodes, the bandwidth, available storage at peering nodes and contact duration also affect data forwarding. These factors also influence the mechanisms such as buffer replacement and scheduling policies. So there are secondary problems that routing strategies may need to take care of such as to deal with limited resources like buffer, bandwidth and power. Furthermore, despite inherent delay tolerance of most DTN driving applications, there can be situations where some messages may be more important than the others and expected to get delivered earlier. So considering the network limitations and application requirements, the problem of choosing the messages to be transmitted when a contact opportunity arises and the messages to be dropped when buffer full is formulated. A buffer management policy to address these issues is proposed and analysed in this paper. Additionally the buffer utilization of various DTN routing protocols and the impact of buffer size on the performance of DTN are studied.
Keywords: Delay Tolerant Networks, Buffer management, Prioritization of messages, Delivery ratio and Delivery delay.

1 Introduction

Communication networks whether they are wired or wireless operate with the assumption of existence of end-to-end path always between source and destination. However, emerging applications such as earth-quake monitoring, habitat monitoring, Vehicular Ad hoc Networks (VANETs) [17] and military ad hoc networks, will likely to change the typical conditions under which networks operate. In fact in such scenarios, networks pose new challenges like frequent disconnections, limited bandwidth, long delays and high bit error rates. The critical fact is that in such situations, the Transmission Control Protocol (TCP) often does not work [26]. To enable those applications to operate under such challenging conditions, a new network paradigm called Delay Tolerant Networking has been proposed by researchers [16]. Networked environments which operate under such intermittent connectivity are referred as Delay/Disruption-Tolerant

Networks (DTNs).

Delay-Tolerant Networking architecture given by [25] is designed to provide support for message based asynchronous communication over network prone to frequent disconnection, long and variable delays. The details of Delay Tolerant Networking architecture are available in [11]. From the literature survey [9], [13], [15], [22], [24] it is understood that a large amount of research has been performed in developing efficient routing algorithms for DTNs. DTNs operate with the principle of store, carry and forward. According to this principle, a node may store a message in its buffer and carry it along for long period of time until it can forward it further. In addition, to achieve high delivery probability, messages are replicated multiple times. Most of the routing protocols in DTN assume that the buffer available as infinite, which is not the case in reality. Therefore the buffers in the node will run out of capacity at certain point of time due to long term storage and extensive message replication. Further, transmission takes place when nodes come into communication range of each other. The node has to decide which of the messages to be transmitted among the messages those are available in the buffer. It is a challenging problem because of the resource constraints like limited bandwidth and limited period of contact due to node mobility. In this paper, a prioritized buffer management approach is proposed which takes care of both: which messages to be transmitted when a new contact arises and which messages to be dropped when buffer is full.

The opportunistic Network Environment (The ONE) simulator [12], [14] which is designed specifically for DTN environment is used for evaluation of the proposed approach. The result of evaluation is compared with other dropping policies. The remainder of this paper is organized as follows: Section 2 gives overview of various DTN routing protocols and their buffer utilization. Additionally, the impact of buffer size on performance of DTN is analysed. The Buffer Management and related work is discussed in section 3. Section 4 discusses about the proposed policy. The simulation setup and the results are discussed in section 5. Section 6 concludes the paper.

2 Routing Protocols of DTN and their Buffer Utilization

2.1 Overview of Routing Protocols of DTN

Extensive study of different routing mechanism is essential to understand the design of DTNs. The details of various routing protocols of DTN are available in [9], [13], [15]. These protocols differ in the knowledge that they use in making routing decisions and the number of replication they make. Examples of some of the DTN protocols are Direct Delivery, First Contact, Epidemic [4], [5], Spray and Wait [3], [21], PRoPHET [2], and MaxProp [10] routing. Recently, the work in [28] categorized the routing solutions for DTN based on the approach they follow and the information known, as deterministic or scheduled, enforced and opportunistic routing. Deterministic routing approaches are used when knowledge of contact information are known ahead of time. In Enforced routing scheme, special-purpose mobile devices like message ferries proposed in [29] and data mules proposed in [27] are employed. The Opportunistic routing approach relies on arrival of contact opportunities. It is used where there is no knowledge about connectivity or mobility is known a priori and no network infrastructure exists to provide connectivity. In this paper the proposed buffer management policy is used with opportunistic routing approach.

2.2 Buffer utilization of DTN Routing Protocols

In this section, the buffer utilization of various routing protocols at different transmission ranges are observed. Buffer utilization is defined as the ratio of total size of the buffer occupied

and the total size of the buffer. It is represented mathematically as

$$\text{Buffer utilization} = \frac{\text{total size of buffer occupied}}{\text{total size of the buffer}}$$

The ONE simulator is used to perform the experimentation. A new simulation environment is created that combines movement modelling, routing simulation, visualization and reporting in a single framework. They are loaded dynamically based on the given configuration files. The Figure 1 depicts the percentage of buffer utilization by various DTN routing protocols. It is observed from the results as shown in Figure 1 that, irrespective of transmission ranges, the buffer utilization is identical and also less for Direct Delivery routing protocol as expected. This is due to the fact that the messages are delivered directly to the destination. The buffer utilization is at the maximum in the transmission range of 250 m for multi-copy routing protocols like Epidemic and Prophet routing as they do replication based on node encounters and they have comparatively more neighbours in this range. Due to the individuality of Spray and Wait routing, it has less buffer usage compared to that of Epidemic and Prophet routing. The utilization of buffer by MaxProp routing though varies with the transmission ranges still, it is noticeably less. It is due to the fact that it uses acknowledgement and removes the stale data from the buffer. It is also evident from the result in Figure 1. In the transmission range of 50 m and below, the possible contact opportunities are less. Therefore the requirement of buffer and their utilization is also less for all routing strategies.

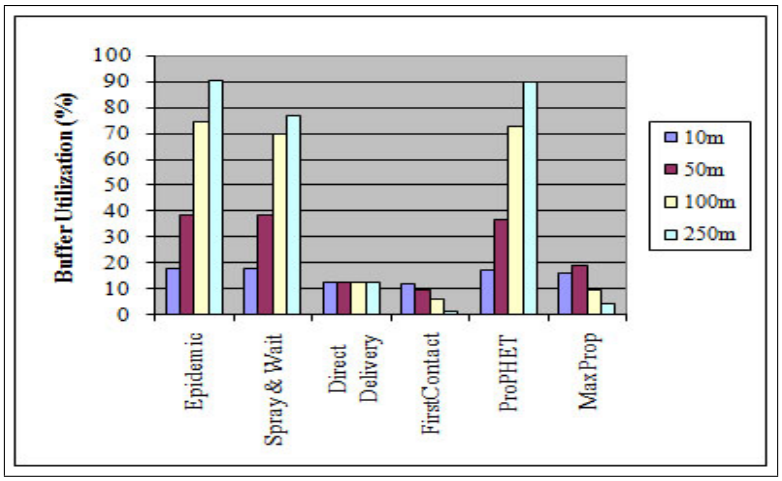


Figure 1: Buffer Utilisation

2.3 Impact of Buffer Size

It is essential to understand the impact of buffer size on performance as this resource is limited in reality (example, sensor nodes). Therefore the impact of buffer size is analyzed in this section. It is examined by varying the buffer size in terms of percentage of total number of messages generated. Both Epidemic and Spray and Wait routing are chosen for evaluation. Both the protocols operate on the assumption of infinite buffer and are respectively of type uncontrolled and controlled flooding which requires huge buffer space. The trade-off between the delivery probability and the buffer size is explored at heavy and light traffic load and the results are shown in Figure 2 and 3 respectively. The traffic load is varied by changing the message generation intervals. It is derived from the results that a buffer size of 5-25 % of the generated messages is sufficient to achieve high delivery ratio with reasonable latency.

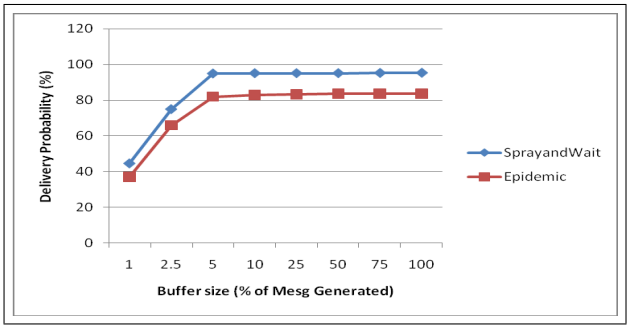


Figure 2: Delivery Probability vs. Buffer size (at Heavy Traffic Load)

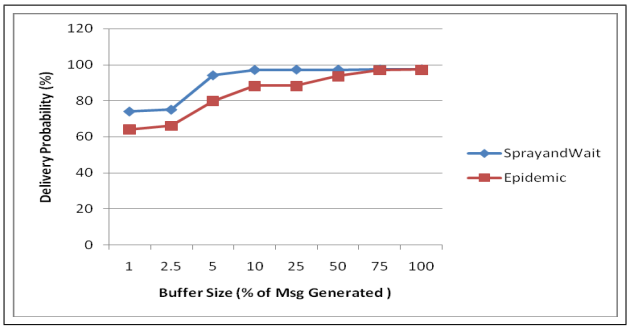


Figure 3: Delivery Probability vs. Buffer size (at Light Traffic Load)

3 Buffer Management and Related Work

This section reviews existing literature on buffer management in DTN. The single-copy protocols like Direct Delivery and First Contact routing and multi-copy protocols like Epidemic and Spray and Wait routing transmit messages in FCFS order, that is the messages are transmitted in the order in which they were stored in the buffer. PRoPHET routing makes forwarding decision based on delivery predictability of the destination of the message. It needs history of past encounters for calculation of delivery predictability. MaxProp routing [10] assigns priority to messages based on hop count and delivery likelihood. It needs to maintain history of data to estimate delivery likelihood.

RAPID protocol [7] explicitly calculates per-packet utility value for administrator-specified routing metric and forwards the message with highest utility value first. In Prioritized Epidemic Routing [19], transmission and dropping is done based on the priority assigned to each bundle. The priority is based on hop count the bundle has traversed thus far. The Optimal policy in [18], [23] deduces the optimal function for the metrics like delivery ratio and delivery delay independently and the message with smallest utility value is dropped when the buffer is full. The authors in [20] had given a framework for devising optimal routing and scheduling algorithm. It is a centralized algorithm. Irrespective of the routing protocols, several dropping and scheduling policies were proposed in the literature [1]. A different combination of queuing and forwarding strategies had been proposed in [6]. But those policies can be used only with PRoPHET routing. It is observed from the study that so far no policies had considered the importance of messages or the requirement of the application.

4 Proposed Prioritized Policy

In addition to network and individual node features and capabilities, application specific requirement must be taken into account when developing DTN routing mechanism. More specifically, some applications that use Delay Tolerant Networking service require preferential delivery of certain messages. For example, field agents wish to communicate their findings, regarding environment hazards to other field agents which are more important than the regular findings. However there is no existing solution which can be applied to variety of DTN applications given their requirements. Under such conditions, a forwarding policy is required to serve different types of traffic differently. Consequently it would be necessary to introduce traffic differentiation mechanism and ensure that they get best possible service according to their requirement. Therefore the goal is to determine the policy which maximizes the delivery probability or equivalently minimizes the delivery latency of high priority messages. A buffer management policy to address these issues is proposed and analyzed in detail in this section. The proposed approach attempts to differentiate traffic based on Class-of-Service (CoS) and schedules according to their class. Further the messages are ordered according to their deadline within the class. The assumptions regarding system model is discussed in the following section.

4.1 System Model

It is assumed that the network is partially connected with low node density and the node meetings are short lived. When two nodes meet, transmission between them succeeds instantaneously. The messages are stored in the buffer until contact opportunity arises or until storage is full. The bundle protocol specified in [8] is used for transfer of messages in DTN. The Bundle Processing Control Flags Bit in the DTN bundle protocol is used to differentiate the traffic through Class-of-Service (CoS) field. The Class-of-Service is identified in this work by the size of the messages. The priority class is based on the concept proposed by DTN architecture. It is assumed that there are three priority classes of traffic: high, medium and low. The messages are transmitted as a whole from node to node. In this proposed approach, the available buffer is logically divided into three queues to hold the incoming messages. Separate queue is maintained for each priority class as shown in Figure 4. The size of the queue is represented in terms of number of messages. Their sizes are defined at the beginning. They can be varied as the correlation of traffic changes. Initially the buffer is divided equally among all queues.
It is assumed that each node has a buffer B of size b = n(B) which is logically divided into three queues: B1, B2, B3 to accommodate high, medium and low priority bundles respectively such that $B = B1 \cup B2 \cup B3$. The size of B1, B2, B3 is b1, b2, b3 respectively such that b = b1+b2+b3. A minimum size q_{min} is specified to reserve space for medium and low priority queues, to avoid the complete negligence of medium and low priority traffic. It is an algorithmic parameter which is set dynamically according to the requirement of the application.

4.2 Proposed Approach

The proposed approach comprises (i)*Bundle classifier* which classifies the bundles based on the traffic class as soon as they arrive and stores them in appropriate queue, (ii)*Bundle scheduler* which schedules the bundles based on the priority from high priority to low priority, (iii) *Bundle dropper* which drops the message according to the policy.

Each bundle S_i in the buffer has a set of information stored with it such as source id, traffic class and Time-To-Live. Several criteria can be used to categorize the bundles such as source id, destination id, size and the TTL of the bundle. In this paper, it is assumed that emergency

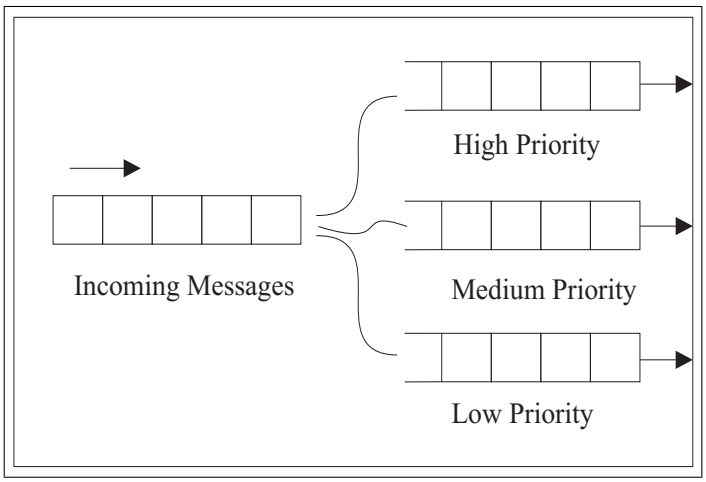


Figure 4: Maintaining Priority Queue

messages are small and have short deadline which is quiet a natural phenomenon. With this assumption, the bundles are classified as high priority when the size of the bundle is between 1 and 10KB. The bundles are classified as medium priority when the size of the bundle is between 11KB and 100KB. The bundles are classified as low priority when the size of the bundle is between 101KB and 1000KB. The Bundle Classifier is a function of newly arrived bundle S_{new} such as

$$f(S_{new}) = \begin{cases} (S_{11}, S_{12}, ...S_{1b1}) \cup (S_{new}) & \text{if } S_{new} \text{ is of high Priority} \\ (S_{21}, S_{22}, ...S_{2b2}) \cup (S_{new}) & \text{if } S_{new} \text{ is of medium Priority} \\ (S_{31}, S_{32}, ...S_{3b3}) \cup (S_{new}) & \text{if } S_{new} \text{ is of low Priority} \end{cases}$$

Where $(S_{11}, S_{12}, ..., S_{1b1})$ are the messages in high priority queue, $(S_{21}, S_{22}, ..., S_{2b2})$ are the messages in medium priority queue, $(S_{31}, S_{32}, ..., S_{3b3})$ are the messages in low priority queue.

The bundle classify procedure is shown in Algorithm 1.

Algorithm.1 Bundle Classify Procedure

```
Bundle_classify()
Receive bundle;
If buffer full
      Call Bundle_drop procedure;
Else
      Check for Class-of-Service of the bundle;
If received bundle is expedited message,
      Store it in high priority queue;
Else if received bundle is normal message,
      store it in medium priority queue;
Else if received bundle is Bulk message,
      Store it in low priority queue;
```

Bundle scheduler is invoked when contact opportunity arises. The number of messages that can be transmitted is limited by the bandwidth and the duration of the contact between the nodes. In such cases only few messages are transmitted and the order in which the messages

are transmitted is significant. Bundle scheduler transmits the bundles from high priority to low priority. The procedure of bundle scheduler is shown in Algorithm 2. The selection of the bundle to be transmitted S_t is represented mathematically as follows:

$$S_t = \begin{cases} S_i | S_i \in B_1 & \text{if high priority queue is not empty} \\ S_i | S_i \in B_2 & \text{if medium priority queue is not empty} \\ S_i | S_i \in B_3 & \text{if low priority queue is not empty} \end{cases}$$

The messages whose destination encountered are the first to be transmitted irrespective of the scheduling policy adopted and the same may be deleted from the buffer. Nodes do not delete messages that are forwarded to other nodes (other than destination) as long as there is sufficient space available in the buffer.

Algorithm.2 Bundle Schedule Procedure

```
Bundle_schedule()
while high priority queue is not empty,
      transmit bundle from high priority queue;
while medium priority queue is not empty,
      transmit bundle from medium priority queue;
while low priority queue is not empty,
      transmit bundle from low priority queue;
```

Once the buffer is full, the Bundle dropper is invoked. As the TTL of the bundle expires, it is dropped automatically. The low and medium priority bundles are dropped to give room for high priority bundles. It is also taken care that a node should not drop its own bundle (source) to give room for newly arrived bundles. The idea of giving priority to source bundles has been proposed in [19], and was shown to improve the average delivery ratio. So the same idea is followed here. The Bundle drop procedure is shown in Algorithm 3. There are three possible classes of bundles that arrive like high priority, medium priority and low priority. Bundle dropping is a function which selects the bundle S_d to be dropped based on the priority of the bundle that arrive. The identification of the bundle to be dropped is mathematically represented as follows:
Case 1: when high priority bundle arrives

$$S_d = \begin{cases} S_i | S_i \in B_3 & if L_{min} > q_{min} \\ S_i | S_i \in B_2 & if M_{min} > q_{min} \\ S_i | S_i \in B_1 & \text{otherwise} \end{cases}$$

Case 2: when medium priority bundle arrives

$$S_d = \begin{cases} S_i | S_i \in B_3 & if L_{min} > q_{min} \\ S_i | S_i \in B_2 & \text{otherwise} \end{cases}$$

Case 3: when low priority bundle arrives

$$S_d = S_i | S_i \in B_3$$

Algorithm.3 Bundle Drop Procedure

Bundle_drop( )
Let L_{min} and M_{min} be the currently occupied space by low priority and medium priority bundles respectively.
Let q_{min} be the threshold value which controls the minimum space for low and medium priority bundles.
Case 1: High priority bundle arrives
  If $L_{min} > q_{min}$
    Drop bundle from low priority and the room is added to high priority queue;
  Else if $M_{min} > q_{min}$
    Drop bundle from medium priority and the room is added to high priority queue;
    Else the remaining time of existing bundles of high priority queue is compared with newly arrived bundle. The bundle with least remaining time is dropped;
Case 2: Medium priority bundle arrives
  If $L_{min} > q_{min}$
    Drop bundle from low priority and the room is added to Medium priority queue;
  Else the remaining time of existing bundles of medium priority queue is compared with newly arrived bundle. The bundle with least remaining time is dropped;
Case 3: Low priority bundle arrives
    The remaining time of existing bundles of low priority queue is compared with newly arrived bundle.The bundle with least remaining time is dropped;

By providing differentiated service based on the class, the best effort service has been enhanced. Moreover the proposed policy takes a state less approach that minimizes the need for nodes in the network to remember anything about flows. It is more practical to implement. The messages are marked in a way that describes the service level that they should receive. These markings are used to provide appropriate service without the need to remember extensive state information for every flow. The metrics used to evaluate the performance of DTN are the delivery ratio and average delivery delay. Both the metrics are defined below.

$$\text{Delivery ratio} = \frac{\text{number of messages delivered}}{\text{total number of messages sent by the sender}}$$

$$\text{Delivery delay} = \text{Average(time taken by all the messages to reach from source to destination)}$$

When compared to other approaches of [18], [19], [20], [23], the proposed approach has the credit of no requirement to maintain state information. Moreover it has less overhead than other approaches as there is no exchange of control traffic before bundle exchange. The proposed approach avoids the complete negligence of medium and low priority messages by reserving minimum space for them.

5 Simulation Results and Analysis

To evaluate the performance of proposed policy, experimentation is done using The ONE simulator. The simulation environment consists of sparsely distributed mobile nodes. They are capable of communicating when they are in the communication range of one another. The parameter of the nodes like buffer size, transmit range, transmit speed, number of nodes are set as mentioned in the Table 1. Further, q_{min} is an algorithm parameter which is set as 10% of the messages generated. The environment where nodes move randomly is considered. So mobility model is set as Random Waypoint Model, in which nodes move independently to a randomly chosen destination. Priority based scheduling is not a basic primitive of DTN routing. Therefore

it has to be incorporated with any of the routing algorithms. Epidemic routing is chosen as baseline for evaluation due to its simplicity and recognition as unbeatable routing protocol from the point of reliable delivery. Furthermore it is suitable for opportunistic environment as it does not rely on previous contact information. However the proposed policy can be incorporated with any of the multi-copy routing protocols.

Table.1 Simulation parameters

Parameters	**Values**
Number of Nodes	50
Transmit Range (m)	250
Transmit speed ($Mbps$)	2
Node Speed (km/hr)	10 - 50
TTL of message (min)	60
Buffer size (MB)	1500
Message size	10 KB - 1 MB
Number of Messages /min	4 - 20
Movement Model	Random Waypoint Model
Simulation Time (min)	160

The performance of Epidemic routing under different buffer management policies are compared in terms of metrics like delivery probability and average delivery latency. Figure 5 and 6 depict the behavior of the proposed policy versus other dropping policies like Drop Front (DF), Drop Old (DO) and Drop Random (DR) with respect to delivery probability and average delivery latency respectively at various traffic loads. The traffic load is increased by increasing the number of messages generated per interval. It is observed from the result that as and when the traffic load increases, the delivery probability decreases. It also shows that incorporating the prioritized policy does not degrade the performance when compared to other policies.

From the literatures [6], [15] it is noticed that DF policy results in highest delivery ratio and DO

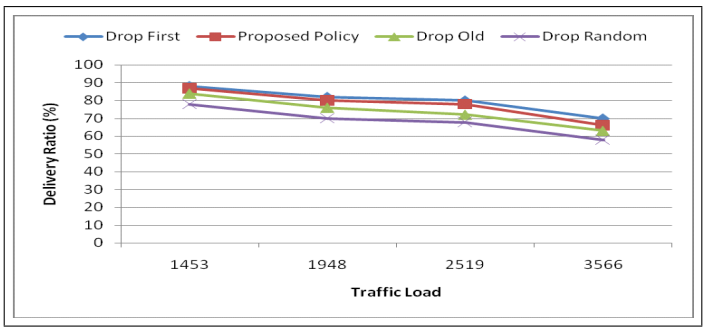


Figure 5: Delivery ratio as a function of Load

results in least average delivery latency. The simulation results shown in Figure 5 and 6 support it. The rationale behind this result is that messages nearing the deadline may get automatically removed from the buffer on the expiry of their deadline. So forcing such messages to drop will not decrease the delivery ratio. Furthermore, high priority messages with earliest deadline are forwarded first. So they have more chances of reaching the destination quickly than other messages before missing their deadline. As the size of high priority message is less compared to other priority class, the number of messages that are transmitted per contact duration is also more. This results in good delivery ratio.

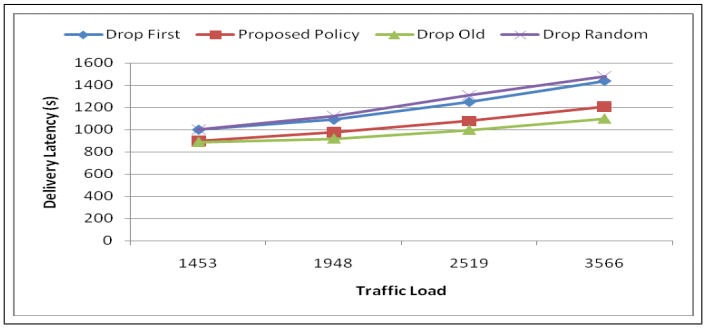


Figure 6: Delivery Latency as a function of load

6 Conclusion

The paper studies the buffer utilization of different routing protocols and the impact of buffer size on performance. The work targets on application that requires preferential delivery in opportunistic DTN environment. The prioritized approach presented in this paper differentiates the traffic based on Class-of-Service and does scheduling and dropping based on their priorities. Thereby it ensures the delivery of high priority messages first with least latency satisfying the application requirement. The proposed policy is more suitable and advantageous in strict resource constrained environment with emergency applications. The service required can be specified by the application. So it can be used in vehicular networks where accident notification messages are more important than other messages. In order to avoid starvation of low priority messages, weighted fair queuing can be used which is carried as future work.

Bibliography

[1] James. A. Davis, Andrew H. Fagg, and Brian N. Levine, N., Wearable Computers as Packet Transport Mechanisms in Highly-Partitioned Ad-Hoc Networks, in *Proceedings of International Symposium on Wearable Computing*, pp. 141-148, 2001

[2] Anders Lindgren, Avri Doria and Olov Schelen, Probabilistic Routing in Intermittently Connected Networks, *Springer LNCS*, Vol. 3126, pp. 239-254, 2004

[3] Spyropoulos, T., Psounis, K. and Raghavendra, C.S., Spray and Wait: an Efficient Routing Scheme for Intermittently Connected Mobile Networks, in *Proceedings of the ACM SIGCOMM Workshop on Delay-Tolerant Networking*, 2005

[4] Alan Demers, Dan Greene, Carl Houser, Wes Irish, John Larson, Scott Shenker, Howard Sturgis, Dan Swinehart and Doug Terry, Epidemic Algorithms for Replicated Database Maintenance", in *Proceedings of ACM Symposium on Principles of Distributed Computing*, pp. 1-12, 1987

[5] Amin Vahdat and David Becker, Epidemic Routing for Partially-Connected Ad Hoc Networks, Technical Report CS-200006, 2000

[6] Anders Lindgren and Kaustubh Phanse, S., Evaluation of Queueing Policies and Forwarding Strategies for Routing in Intermittently Connected Networks, in *Proceedings of International Conference on Communication System softWAre and MiddlewaRE -COMSWARE*, 2006

[7] Aruna Balasubramanian, Brian Levine and Arun Venkataramani, DTN Routing As A Resource Allocation Problem, *ACM SIGCOMM Computer Communication Review*, Vol. 37, No. 4, 2007

[8] Scott, K. and Burleigh, S., Bundle Protocol Specification, RFC 5050, 2007

[9] Evan Jones, P.C. and Paul Ward, A.S., Practical Routing in Delay-Tolerant Networks, *IEEE Transaction on Mobile Computing*, 6(8):943-959, 2007

[10] Burgess, J., Gallagher, B., Jensen, D., and Levine, B.N., MaxProp: Routing for Vehicle-Based Disruption-Tolerant Networks, in *Proceedings of IEEE International Conference on Computer Communications*, pp. 1-11, 2006

[11] Fall, K., A Delay-Tolerant Network Architecture for Challenged Internets, in *Proceedings of SIGCOMM'03*, 2003

[12] Are Keranen, Jorg Ott and Teemu Karkkainen, The ONE simulator for DTN protocol evaluation, in *Proceedings of the 2nd International Conference on Simulation Tools and Techniques*, pp. 1-10, 2009

[13] Sushant Jain, Kevin Fall and Rabin Patra, Routing in Delay Tolerant Networks, *ACM SIGCOMM Computer Communication Review*, Vol. 34, No. 4, 2004

[14] TKK/COMNET. Project page of the ONE simulator. 2008. www.netlab.tkk.fi/tutkimus/dtn/theone/

[15] Zhensheng Zhang, Routing in Intermittently Connected Mobile Ad Hoc Networks and Delay Tolerant Networks: Overview And Challenges, *IEEE Communication Surveys and Tutorials*, 8(1):24-37, 2006

[16] Cerf, V., Burleigh, S., Hooke, A., Torgerson, L., Durst, R., Scott, K., Fall, K. and Weiss, H., Delay Tolerant Networking Architecture, IETF Network working group, RFC 4838, 2007

[17] Basu, P. and Little T.D.C., Networked Parking Spaces: Architecture And Applications, in *Vehicular Technology Conference*, Vol.2, pp. 1153-1157, 2002

[18] Amir Krifa, Chadi Barakat and Thrasyvoulos Spyropolous, Optimal Buffer Management Policies for Delay Tolerant Networks, in *Proceedings of IEEE Conference on SECON*, pp. 260-268, 2008

[19] Ramnathan, R., Hansen, R., Basu, P., Rosales Hain, R. and Krishnan, R., Prioritized Epidemic Routing For Opportunistic Networks, in *Proceedings of the 1st International MobiSys workshop on Mobile Opportunistic Networking*, pp. 62-66, 2007

[20] David Hay and Paola Giaccone, Optimal Routing And Scheduling For Deterministic Delay Tolerant Networks, in *Proceedings of International Conference on Wireless On-Demand Network Systems and Services*, pp. 27-34, 2009

[21] Spyropoulos, T., Psounis, K. and Raghavendra, C.S., Efficient Routing in Intermittently Connected Mobile Networks: The Multi-Copy Case, *IEEE/ACM Transactions on Networking*, 16(1):77-90, 2008

[22] Elizabeth Daly, M. and Mads Haahr, The Challenges of Disconnected Delay-Tolerant MANETs, *Elsevier Ad Hoc Networks Journal*, 8(2):241-250, 2010

[23] Amir Krifa, Chadi Barakat, Thrasyvoulous Spyropolous, An Optimal Joint Scheduling and Drop Policy for Delay Tolerant Networks, *IEEE WoWMoM*, 2008

[24] Delay tolerant networking research group, [Online]. Available: http://www.dtnrg.org

[25] Fall, K. and Farrell, S., DTN: an Architectural Retrospective, *IEEE Journal on Selected Areas in Communications*, 26(5):828-836, 2008

[26] Farrell, S., Cahill, V., Geraghty, D., Humphreys, I. and McDonald, P., When TCP Breaks: Delay- and Disruption- Tolerant Networking, *IEEE Internet Computing*, 10(4):72-78, 2006

[27] Shah, R.C., Roy, S., Jain, S. and Brunette, W., Data MULEs: Modeling a Three-tier Architecture for Sparse Sensor Networks, in *Proceedings of the IEEE International Workshop on Sensor Network Protocols and Applications*, pp. 30-41, 2003

[28] Thrasyvoulos Spyropoulos, Rao Naveed Rais, Thierry Turletti, Katia Obraczka and Athanasios Vasilakos, Routing for Disruption Tolerant Networks: Taxonomy And Design, *Wireless Networks*, Vol. 16, No. 8, 2010

[29] Zhao, W., Ammar, M. and Zegura, E., A Message Ferrying Approach For Data Delivery In Sparse Mobile Ad Hoc Networks, in *Proceedings of the 5th ACM International Symposium on Mobile Ad hoc Networking and Computing, MobiHoc*, pp. 187-198, 2004

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 2 (June), pp. 264-272

A Decision-Making Perspective for Designing and Building Information Systems

F.G. Filip

Florin Gheorghe Filip
The Romanian Academy: BAR and INCE,
and the Nat. Inst. for R&D in Informatics -ICI,
Bucharest, Romania
ffilip@acad.ro

Abstract: This paper aims at highlighting several aspects and associated decision situations which may be met in the process of designing and building modern information systems, such as: choosing the approach and methods to be utilized for building the system and selecting the IT tools, integrating the system into the enterprise and evaluating the project. A particular emphasis is put on evaluation criteria to be utilized in solving the various decision problems.
Keywords: decision criteria, IT&C tools, methodology, prototype, SaaS, standards.

1 Introduction

A large number of models and corresponding solvers have been proposed and reported in the literature with a view to getting optimal solutions for the academic test-problems or real-world management and control applications. They can be found in the technical literature. In many cases, in practical applications, a necessary condition to make the models and the corresponding solvers utilized is to incorporate them into adequate *information systems* (IS).

In the same time, the IT&C (*information technology and communication*) vendors release to the market ever more modern hardware and software products. New trends can be noticed on the software market [21], such as "merger mania" (consisting in mergers, acquisitions, partnerships and strategic alliances between business software vendors), functional expansion, a clear dominance of a few (three) databases (Oracle, Microsoft SQL Server, and IBM DB2), increased usage of Internet, viewing IT/software as a service and so on. Exhaustive services for project management in the software field to serve those managers who want to save time are already available on the market [18].

One can notice, during the process of designing and implementing information systems, sequences of decisions which should be made with respect to the choice of the most adequate alternatives concerning several critical aspects, such as system orientation, composition of the team, method to be adopted, IT&C tools to be utilized, resources to be allocated and so on. In this context, this paper aims at surveying from a decision-making perspective several methodological and practical aspects of designing effective (*usable, useful*, and actually *utilized*) IS. The paper extends, details and up-dates an earlier preliminary version of the paper [9]. The remaining part of this paper is organized as it follows. In the second section, several factors which may influence decisions meant to design and implement information systems in organizations are reviewed. Then, several design and implementation critical aspects are presented, such as the design approach adopted and the selection of the IT&C tools. The paper concludes with a discussion of technical and non-technical integration issues and evaluation aspects. Throughout the paper the evaluation criteria to be used in various decisions are highlighted.

2 Main Influence Factors

There are several factors which can influence the process of designing and implementing an information system, such as: people involved, the orientation and the purpose of the system, the organizational setting, standards utilized and so on [8]. Those influence factors should be taken into consideration by the management of the target enterprise and the designer as well, when a decision on introducing and creating an information system is to be made.

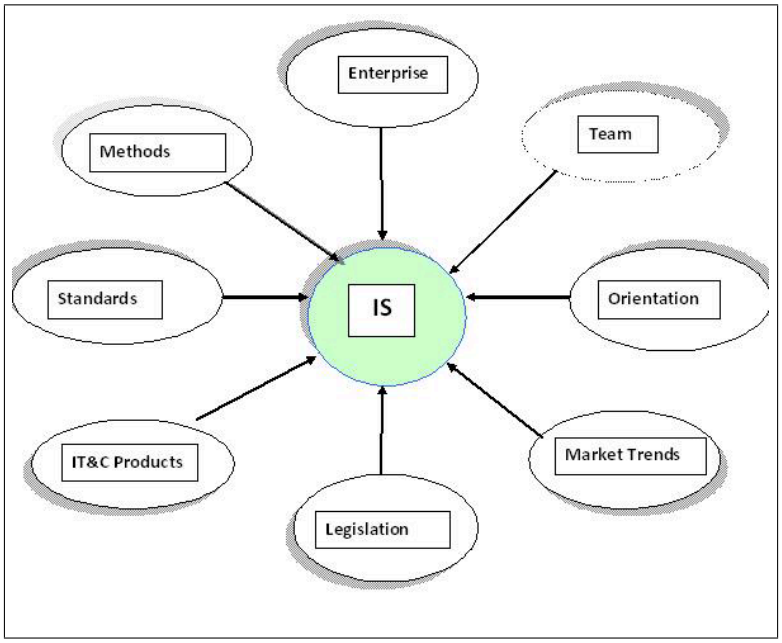


Figure 1: Influence factors (adapted from [9])

The *people involved* in the IS design and implementation should cooperate closely to form a virtual team who aims at obtaining the best solution for the allocated resources (time, manpower, money). There are several classes of people who should take part in various extents of involvement and contributions to the process for the first moment of discussing the idea of IS until its "steady state" operation and impact evaluation. One can identify the following generic classes:"clients", designers, and IT&C vendors. The members of the "client" class include the project "sponsor" (a manager) and the project "champion" who represents the interests of the future direct ('hands-on") or indirect ("beneficiary") actual *users* (who may be also involved in design and implementation of the IS). The "project champion" possesses the necessary knowledge of the application domain. The "project sponsor" possesses the authority since he/she represents the interests of the organization and, consequently, is empowered to accept or reject the project solutions and allocate the necessary resources. The *designers* can be members of a group of people of the organization or/and a team of analysts and IT&C professionals from a consultancy firm who master the design techniques and who are aware of the IT&C products available on the market. The IT&C *manufacturers* and *vendors* can adapt and alter the IT&C products to be utilized.

The information system may *be oriented* to serve a certain generic class of users ("roles") or to help a specific group of persons with names, identities and specific IT skills ("actors").

The system *purpose* might be either to facilitate and make more comfortable the work of

the users, or to promote the change. The *normative model of the change* [14] includes several steps of cooperation between users and designers which can be met in the process of creating an information system, such as:

- Pioneering (to evaluate the needs of the final users and necessary competences of the systems constructors);
- Acceptance (to establish the objectives to be agreed by the users and the designers);
- Diagnosis (to collect the data, define the problem and estimate the necessary resources);
- Planning (to set up the work plan and allocate the corresponding resources);
- Action (to design and implement the information system and train its users);
- Evaluation of the process and project impact.

The target *organization* where the information system is to be implemented may create a context which strongly influences the solution and the process of the system building. There might be constraints caused by a) the insufficient IT&C knowledge and skills of the future users and b) scarce available data or/and limited internal data access rights of the external consultants. Several integration problems may show up caused by the "legacy IT systems" or/and the operating procedures permitted within the organization.

Standards should play a central role in design. The *International Standard Organization* (ISO) is an excellent source of documents to be utilized to set the stage for useful, usable and used solutions. The standards for usable (traditionally called "user-friendly") interfaces, such as those of the series [13]*ISO 9241* ("Ergonomics of Human-System Interaction"), are recommended and can contribute to obtaining a user-centered solution. The quite recent standards, such as *ISO 9241-171.2008* ("Guidance in Software Accessibility") and *ISO 9241-151.2008* ("Guidance WWW User Interface"), are of a particular importance in the context of modern information systems which are ever more oriented to use www technologies. Suduc [23] gives a comprehensive analysis of the design methods to be utilized in the low-cost interface design. Cojocaru [5] proposes the *intelligent interface* concept.

Other aspects, such as previous experience, industry competitors' moves, legislation and, the most serious one, available budgets and intended due dates, may also influence the IS construction process.

3 Design and Implementation Approaches

There are various approaches to designing, building and implementing an information system. They can be grouped in accordance with several *criteria*, such as:

- IT&C tools and platforms which will be utilized (general-purpose products vs integrated suites/generators/shells);
- Buying IT or using *IT as a Service-ITaS* [11], or *Software as a Service -SaaS* [4] and [25];
- Place for construction (within the target organization or at the consultant's site);
- Method utilized (the lifecycle method or the evolving/adaptive design which is based on the use of the prototype).

The *lifecycle*-based method requires several steps, such as system analysis, design, implementation, and operation which are carried out in sequential ("cascade") manner. It also implies that the well defined procedures and checkpoints are strictly observed and the solutions adopted are well documented. It is, consequently, recommended for large-scale applications.

The origins of the *prototype*-based method [20] in the field of IS could be traced back in mid 70's in the empiric remark that 80% of the design ideas in the field are wrong [17]. Consequently, in order to avoid the waste of resources, the prototype permitted to spend 20% of the resources in the early stages of design and construction for identifying the 80% wrong ideas, so that the remaining 80% of resources should be utilized to implement the remaining 20% of ideas which are likely to be correct.

When adopting the prototype-based method, there are a few *basic principles* which are to be observed such as:

• The process starts with approaching the most critical problems of the target organization, so that the user's confidence could be gained as early as possible. In some cases, a "demo script", which includes a critical business scenario [21], can be also utilized to select the most adequate product from a short list (see the next section);

• The early requirements can be formulated in collaboration with the user in a "quick and dirty "simplified manner;

• The information system is developed in several cycles which include operations, such as experiment, evaluation, modification. The cycles should be as short as possible and the cost of the first version must be very low, in order not to lose the user's interest and confidence;

• The evaluation of the effects of the usage of the preliminary version is carried out on a permanent time basis.

Two main types of prototypes have been commonly utilized [22]: a) the "throwaway" prototype, and b) the "evolving one". While the former is only utilized to test the design ideas and then is discarded (the next versions re-designed by using new technologies), the latter consists in a series of refinements of the initial version (Fig. 2). A decision choice should be made on which type to be utilized.

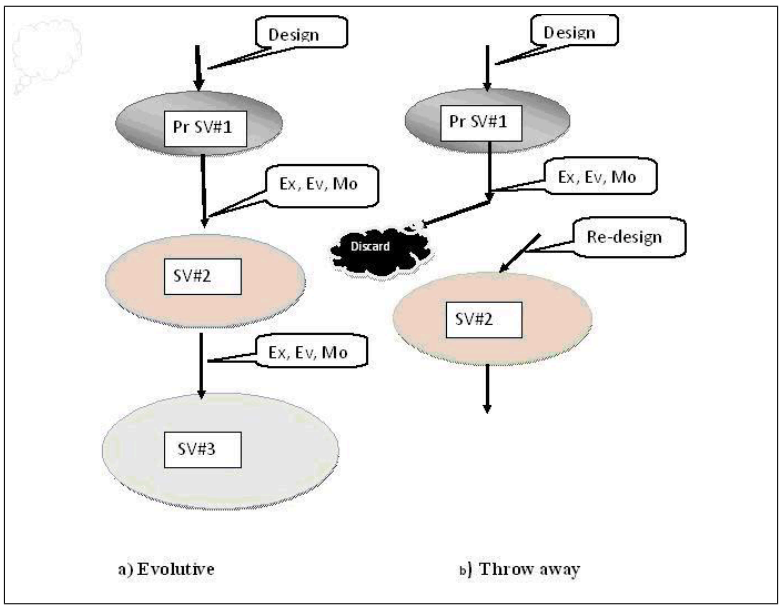


Figure 2: Types of prototypes (Legend: Pr= prototype, SV= System version, Ex= Experiment, Ev= Evaluate, Mo= Modify)

The prototype-based (adaptive/incremental/iterative) methods allow for obtaining a good well customized, early utilizable and helpful solution, even if the information on organization

and its context is scarce and uncertain. On the other hand, methods may stimulate the tendency to continually modify the solution or, on the contrary, to adopt too early a solution which is imperfect or incomplete. In [6] and [8] the story of constructing *DISPATCHER®*, a family of *Decision Support Systems* (DSS) which is meant to assist logistics and production control decisions to be made in the context of the continuous process industries and related fields, is described. The *DISPATCHER®* project started in early 80's as an optimization model and software for production scheduling. Since then, under the influence of several factors such as the users' changing needs and improvement of their IT&C skills, specific characteristic features of target enterprises, and new products and technologies released in the field of IT&C, several application versions were designed and installed in refineries, pulp and paper mills, chemical plants, and water systems. DIPATCHER® has evolved towards a complex solution, a *DSS generator* which could be adapted to new business models (such as the "extended"/networked/"virtual" enterprise), to support new functions and usages. It includes new constituents, such as a three-level modeling scheme of the plant (expressed in terms of final users, analysts and programmers, respectively), AI (*Artificial Intelligence*)-based solvers and model experimentation tools(Fig. 3).

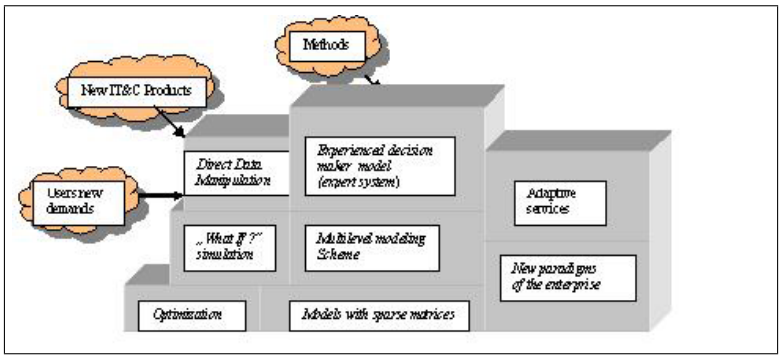


Figure 3: Evolution of Dispatcher system (adapted from [6] and [9])

An important decision consists in making a choice between buying or leasing IT&C products. For example, in recent years, the approach to use SaaS (*Software as a Service*) model has become ever more popular. This new business model means that the software vendors host on their servers the applications to be accessed, via internet, by client organizations only when it is necessary. The pricing scheme is based on paying monthly lease fees, instead of initial license cost and annual maintenance fee. The SaaS scheme can be of particular interest for smaller companies that have limited IT infrastructure and skilled personnel. On the other hand, when a decision is to be made one should take into account long run costs and security issues. Availability of necessary IT infrastructure, ease of usage are the main evaluation criteria when a decision is to be made.

4 Selection of the IT&C Tools

The selection of the IT&C tools should be viewed as a *multi-attribute decision-making* (MADM) problem ([12], [7], [19], [15]). The general criteria to be used in selecting and ranking the possible IT&C products which can be found on the market can be grouped as it follows:

• *Adequacy*: informational transparency, accuracy of expected results, robustness to errors and low quality uncertain input data, response time;

• *Quality of implementation*: scalability, flexibility, easy integration with the "legacy systems", functional transparency, documentation completeness;

• *Delivery quality*: price, delivery time, provider's general reputation, easy adaptation, degree on dependence on the technical assistance from the provider's specialists for implementation and usage).

A particular attention has been received by the software selection. A set of criteria for filtering the software products from the initial long list is recommended by Software Resources [21], such as: a) budget for the new system, b) unique functionality (i.e. multi-currency, multi-company, budgeting, workflow), c) technology preferences, d) reporting, e) scalability, and f) vendor stability. As a subsequent step to make a choice from the short list Software resources recommends the "scripted (or the standard) demo" in order to determine the software vendor or the VAR (*value added reseller*) to modify their product demo to show how the specific business needs of the client can be solved. A systematic methodology for software evaluation and selection through the usage of MADM was proposed by [16] and an experimental expert system was proposed by [27].

An interesting list of pitfalls to be avoided when selecting the software is provided by Software Resources [21]. It includes the following 12 "deadly mistakes": "a) buying the same software as the competitors, b) buying software based on features alone, and overlooking other critical factors (scalability, flexibility, excessity, technology and cultural fit, affordable cost, insufficient technical support and infrastructure), c) neglecting the proper consideration of the vendor reputation, d) buying software without focusing on the implementation partner, e) taking into consideration only the low initial costs and overlooking significantly higher ongoing costs, f) buying software using input from an elite group without getting buy-in from the organization at large, g) choosing the popular software without considering all the posible and affordable options h) buying software that is too complex, i) making a choice without properly defining your requirements, j) buying software that is either at the end or at the beginning of its product lifecycle, k) buying software that is based on dying technology, l) selecting a software only to fix the current business problems" (instead of implementing the change).

Comprehensive on-line independent support for software evaluation and selection can be obtained from specialized consulting firms, such as Technology Evaluation Center [24], Software Resources [21], Project Perfect [18].

5 Integration

In many cases a new information systems should be integrated into the existing or planned IT&C infrastructure of the target organization. Several *principles* are recommended by [26] which are still valid for *technical integration*, such as:

• Adopting an "open system" architecture;

• Neutralizing the information which can be achieved by using standardized data formats;

• Semantic unification which means a symbol has a unique meaning throughout the whole system.

There are, however, several new problems which can show up due to *non-technical causes*, for example:

• *Wrong orientation* of the solution which does not facilitate solving the central problems of the organization; this may be associated with *informational opacity* (the system provides more/less than necessary outputs);

• *Functional opacity* which means that he user is not given the necessary information and incentives to understand how the system works;

• *Frustration* of the "hands-on" user due to a long response time or an un-adequate (insufficient/excessive) number of functions to perform his/her task.

6 Evaluation

Evaluation of IS has been subject of interest for long time [10]. There are several *main principles* to be observed in the process of designing, building and implementing an information system, such as:

• Evaluation is necessary all over in the design and implementation process to support making a decision choice from the set of possible alternatives (project continuation, giving up, allocating some additional resources and so on);

• The objectives and the degree of detail of evaluation depend on various factors, as : a) the project scope, b) technical complexity, c) duration and cost of the project, d) the person who requested the evaluation, e) overall state of the target enterprise;

• Involving the designer into the evaluation team is necessary especially in the case of a large project.

As previously stated, the evaluation is meant to support a decision-making process. Consequently, a set of *evaluation criteria* should be set up, namely:

• Impact on the efficiency of users' professional performance in accomplishing his/her tasks and quality of life (intellectual development, possible additional stress caused, comfort of performing the task);

• Impact on target enterprise general evolution;

• Implementation and further running costs.

A set of more detailed set of criteria which was used in a specific context is given by [1].

There are several methods which can be utilized for evaluation, for example: a) benefits/cost analysis, though the NVP ("net present value") of the investments, b) value analysis, c)"rating and scoring", d) event logging and so on. Agouram [2] gives a useful methodology to assess the IS success.

7 Conclusion

The activities of designing and implementing an information system form, in practical applications, a process which may include many decisions to be made at different stages. There are several critical aspects, both of the technical and non-technical nature, which should be taken into consideration. Among the main aspects which might cause problems are the evolution of the technical constituents associated with the increased requirements for the solution quality set by the users who are ever more informed and skilled and have to face an ever more fierce competition. Multi-attribute decision models could be effectively utilized to solve the decision situations which can be encountered.

Bibliography

[1] Al-adaileh R. M, An evaluation of information systems success: A user perspective - the case of Jordan Telecom group, *European Journal of Scientific Research*, 37(2):226-239, 2009

[2] Agouram H., Defining information system success in Germany, *International Journal of Information Management*, 29:129-137, 2009

[3] Bizoi, M., Sisteme support pentru decizii bazate pe comunicatii (Communication-Based Decision Support Systems), Ph.D. Thesis, (in Romanian), (http://www.racai.ro /Doctorate/Bizoi_Rezumat_teza.pdf, 2010, accessed on 09.07.2011)

[4] Carraro G., Chong F., Software as a Service (SaaS): An Enterprise Perspective, *MSDN Library* (http://msdn.microsoft.com/en-us/library/aa905332.aspx, 2006, accessed on 09.02.2012)

[5] Cojocaru S., Interfete inteligente ("Intelligent Interfaces"), in [9], pp. 213-215, 2007

[6] Filip F. G. , Towards more humanized real-time decision support systems, In *Balanced Automation Systems: Architectures and Design Methods* (L. Camarinha-Matos, H. Afsarmanesh, eds), Chapman & Hall, London, pp. 230-240, 1995

[7] Filip F.G., Decizie asistata de calculator; decizii, decidenti, metode de baza si instrumente informatice associate ("Computer-Aided Decision-Making; Decisions, Decision-Makers, Basic Methods and Software Tools"), 2nd Edition, Editura Tehnica, Bucuresti(in Romanian), 2005

[8] Filip F. G., Sisteme support pentru decizii ("Decision Support Systems"), 2nd Edition, Editura Tehnica, Bucuresti (in Romanian), 2007

[9] Filip F.G., Designing and building modern information systems: a series of decisions to be made, *Computer Science Journal of Moldova*, 119-129, 2011

[10] Hamilton S., Chervany N.L., Evaluating Information System Effectiveness - Part I: Comparing Evaluation Approaches, *MIS Quarterly*, 5(3):55-69, 1981

[11] Hine J., Laliberte B., Enabling IT as a Service, *White Paper. ESG* (http://www.cisco.com/en/US/prod/collateral/netmgtsw/ps6505/ps11869/esg_enabling_it.pdf, 2011

[12] Gaindric C., Luarea deciziilor: metode si tehnologii ("Decision Making: Methods and Technologies"), Editura Stiinta, Chisinau (in Romanian), 1998

[13] ISO 9241 Ergonomics of human-system interaction, 2012 http://www.iso.org/iso/search.htm?qt=9241&sort=rel&type=simple&published=on; accessed on 09.02.2012)

[14] Kolb D.A., Frohman A. L., An organization development approach to consulting, *Sloan Management Review*, 12(4):51-65, 1970

[15] Gang Kou, Yanqun Lu, Yi Peng, and Yong Shi, Evaluation of classification algorithms using CDM and Rank Correlation, *International Journal of Information Technology & Decision Making*, Vol. 11, Issue: 1, DOI: 10.1142/S0219622012004872, 2012.

[16] Moriso M., Tsoukias A., JusWare: a methology for evaluation and selection of software products, *IEE Proc-Softw. Eng.*, 144(2):162-174, 1997

[17] Ness D.N., Decision support systems: theory and design, In *Wharton Office of Naval Research on DSS*, Philadelphia, November, pp. 4-7, 1975

[18] PP: Project Perfect: Project Management Software, 2012 http://www.projectperfect.com.au/, 2012, accessed on 09.02.2012)

[19] Resteanu C., Somodi M., Alexe B., *Multi-Attribute Decision-Making*; E-course, ICI, Bucharest, 2007

[20] Shelly G. B., Cashmanan T. J., Rosenblatt H.J., *Systems Analysis and Design*, 8th Ed., Thomson Course Technology, Boston, Mass, 2010

[21] SR: Soft Resources: Empowering Software Decisions, 2012, (http://www.softresources.com/, accessed on 06.02.2012)

[22] Sprague jr. R. H., Carlson E. D., *Building Effective Decision Support Systems*, Prentice Hall, Englewood Cliffs, N. J., 1982

[23] Suduc A.M., *Interfete avansate pentru sisteme support pentru decizii*("Advanced Interfaces for DSS"), PhD Thesis, (in Romanian), 2010 (http://www.racai.ro/Doctorate/Suduc_ Rezumat_teza.pdf , accessed on 09.07.2011)

[24] TEC: Technology Evaluation Center, 2011, (http://www.technologyevaluation.com/software/, accessed on 09.07.2011)

[25] Trumba , *White Paper: Five Benefits of Software as a Service*, 2007 (http://www.trumba.com/ connect/knowledgecenter/software_as_a_service.aspx, accessed on 09.02.2012)

[26] Vernadat A., *Enterprise Modeling and Integration Principles and Applications*, Chapman & Hall, London, 1996

[27] Vlahavas I., Stamelos Refanidis I., Tsoukias A., ESSE: an expert system for software evaluation, *Knowledge-Based Systems*, 12(4):183-197, 1999

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 2 (June), pp. 273-284

A General Approach for Minimizing the Maximum Interference of a Wireless Ad-Hoc Network in Plane

V. Haghighatdoost, M. Espandar

Vahid Haghighatdoost, Maryam Espandar
Computer and Electrical Engineering Department,
Shahed University, opposite Holy shrine of Imam Khomeini,
Khalij Fars Expressway, Tehran, Iran.
{haghighatdoost, espandar}@shahed.ac.ir

Abstract: The interference reduction is one of the most important problems in the field of wireless sensor networks. Wireless sensor network elements are small mobile receiver and transmitters. The energy of processor and other components of each device is supplied by a small battery with restricted energy. One of the meanings that play an important role in energy consumption is the interference of signals. The interference of messages through a wireless network, results in message failing and transmitter should resend its message, thus the interference directly affect on the energy consumption of transmitter. This paper presents an algorithm which suggests the best subgraph for the input distribution of the nodes in the plane how the maximum interference of the proposed graph has the minimum value. The input of the application is the complete network graph, which means we know the cost of each link in the network graph. Without any lose of generality the Euclidean distance could be used as the weight of each link. The links are arranged and ranked according to their weights, in an iterative process the link which imposition minimum increase on the network interference with some extra conditions which is proposed in future sections, is added to resulting topology and is eliminated from list until all nodes are connected together. Experimental results show the efficiency of proposed algorithm not only for one dimensional known distribution like exponential node chain, but also for two dimensional distributions like two Exponential node chains and α-Spiral node chains.
Keywords: wireless ad-hoc network, sensor network, interference, spanning tree.

1 Introduction

Wireless ad-hoc networks consist of mobile nodes equipped with, among other components, a processor, some memory, a wireless radio, and a power source. Due to physical constraints, nodes are primarily powered by a weak battery, so energy is a scarce resource in wireless ad-hoc networks. In a general way, topology control can be considered as the task of, given a network communication graph, constructing a subgraph with certain desired properties while minimizing energy consumption. The subgraph needs to meet some requirements, the minimum requirement being to maintain connectivity and it should be a spanner of all nodes in original graph; Additionally, symmetric links are desired as they permit simpler higher-layer protocols [1]. One of the foremost approaches to achieve substantial energy conservation is by minimizing interference between network nodes. The concept of topology control restricts interference by reducing the transmission power levels at the network nodes and cutting off long-range connections in a coordinated way. At the same time transmission power reduction has to proceed in such a way

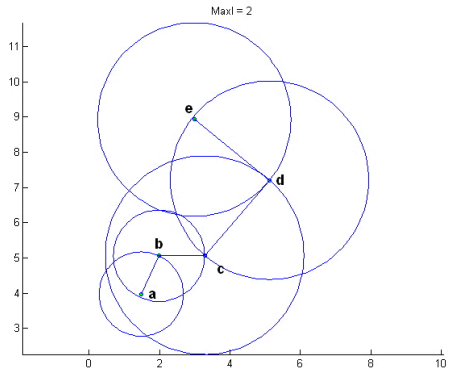


Figure 1: The interference model of a graph with 5 vertexes

that the resulting topology preserves connectivity. Some other works focused on topology control algorithms emphasizing locality while exhibiting more and more desirable properties [2–4], sometimes presenting distributed algorithms that optimize various design goals concurrently. All these approaches have in common, however, that they address interference reduction only implicitly. The intuition was that a low minimizing the maximum degree of nodes of graph would solve the interference issue automatically. As depicted in [1] this intuition was proved wrong in [5], starting a new thread that explicitly studies interference reduction in the context of topology control [6–8]. The general interference model introduced in [9], proposes a natural way to define interference in ad-hoc networks. The general question is: How can one connect the nodes such that as few nodes as possible disturb each other? In the following, we discuss the network and interference model presented in [9].

So far, not many results have been published in the context of explicit interference minimization. For networks restricted to one dimension the authors in [9] present a $\sqrt[4]{n}$-approximation of the optimal connectivity preserving topology that minimizes the maximum interference. For the two dimensional case, the authors in [10] propose an algorithm that bounds the maximum interference to $O(\sqrt{n})$. If average interference of a graph is considered, there is an asymptotically optimal algorithm achieving an approximation ratio of $O(logn)$ [11]. Kevin Buchin in [12] proved that this problem is NP-complete. In the following sections the proposed algorithm is explained briefly. The time complexity of producing the spanning tree with minimum interference is $O(n^5)$.

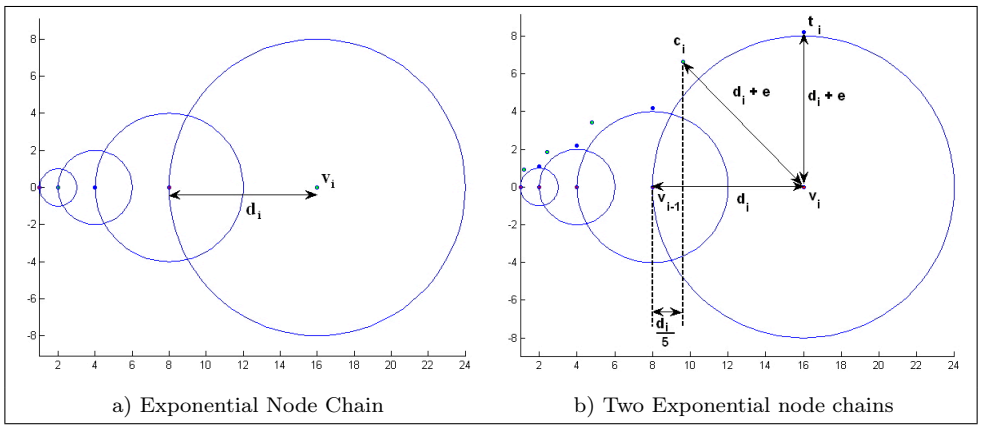
2 Interference Model of Network

The network is modeled as a geometric graph $G = (V, E)$. As mentioned in previous section the links between nodes are symmetric and have not direction, it means that a message sent over a link can be acknowledged by sending a corresponding message over the same link in the opposite direction. So the matrix E is symmetric.

Let N_u denote the set of all neighbors of a node $u \in V$ in the resulting topology. Then, each node u features a value r_u defined as the distance from u to its farthest neighbor. More precisely $r_u = max_{v \in N_u}\{|u - v|\}$, where $|u - v|$ denotes the Euclidean distance between nodes u and v. Since we assume the nodes to use omnidirectional antennas, $D(u, r_u)$ denotes the disk centered at u with radius r_u covering all nodes that are possibly affected by message transmission of u to one of its neighbors. Then the interference of a node v is defined as the number of other nodes that potentially affect message reception at node v.

Definition 1. Given a graph $G = (V, E)$, the interference of a node $v \in V$ is defined as:

$$I(v) = \left| \left\{ u | u \in V \diagdown \{v\}, v \in D(u, r_u) \right\} \right| \tag{1}$$



a) Exponential Node Chain b) Two Exponential node chains

Figure 2: Two Special node distributions

Note that even though each node is also covered by its own disk, we do not consider this kind of self-interference. The graph interference is the maximum interference occurring in a graph:

Definition 2. The interference of a graph $G = (V, E)$ is defined as:

$$I(G) = max_{v \in V} I(v) \tag{2}$$

As shown in Figure 1 the interference of nodes is as follow:

Node:	a	b	c	d	e
Interference:	2	2	2	2	1

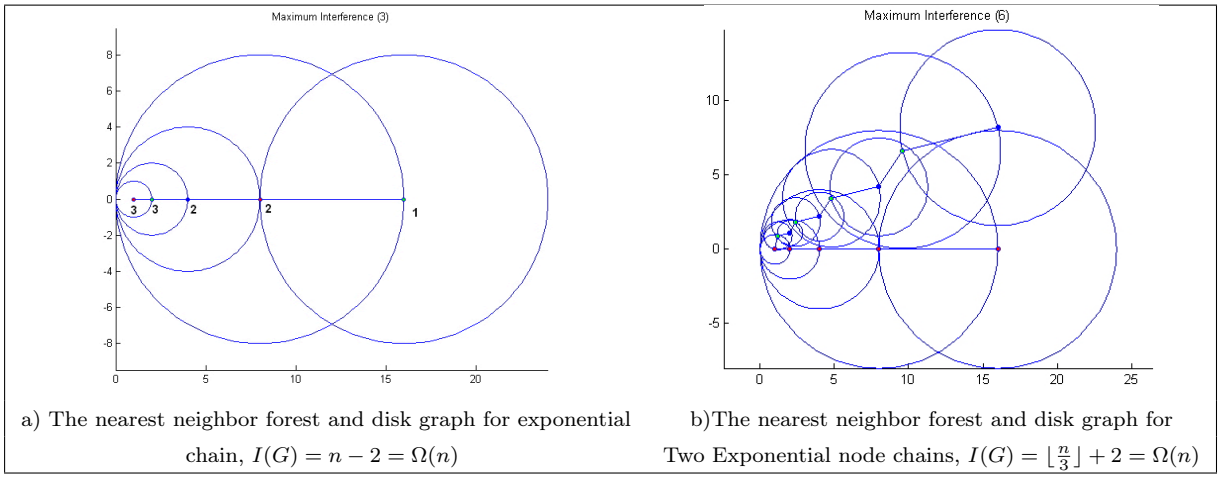
According to Definition 2 the Interference of graph $I(G)$ is 2.

3 The Nearest Neighbor Forest

In the first view of the interference problem, one may say the nearest neighbor forest or minimum spanning tree is the best subgraph which results in minimum interference. In this section, it is shown that this is already a substantial mistake, as thus interference becomes asymptotically incomparable with the interference-minimal topology. For Some special distribution the nearest neighbor forest results in the worst interference. Authors in [13], introduced an instance which seems to yield inherently high interference: the so called *exponential node chain* is a one-dimensional graph $G = (V, E)$ where the distance between two consecutive nodes grows exponentially from left to right as depicted in Figure 2(a).

That is, the distance between nodes v_i and v_{i+1} is 2^i for $i = 0, 1, 2, ..., n-1$. So as shown in Figure 3(a) the nearest neighbor forest results in the interference of $\Omega(n)$. Also in [9] they introduced the Two Exponential node chains as shown in Figure 2(b), on the bottom, there is a horizontal chain of nodes v_i with exponentially growing distances, the same as the one dimensional exponential chain, thus distance between v_i and v_{i+1} is 2^i. Each of these nodes v_i has a corresponding node t_i vertically displaced by a little more than v_i's distance to its left neighbor, that is, $|v_i - t_i| > d_i$ where $d_i = |v_i - v_{i-1}| = 2^{i-1}$. Note that the nodes t_i also form a (diagonal) exponential node chain. Finally, between two of these diagonal nodes t_{i-1} and t_i an additional helper node c_i is placed such that $|v_i - c_i| \geq |v_i - t_i|$. The Nearest Neighbor Forest for this node distribution is shown in Figure 3(b). The nearest neighbor forest for the distributions in Figure 2 and their disk graph is shown in Figure 3.

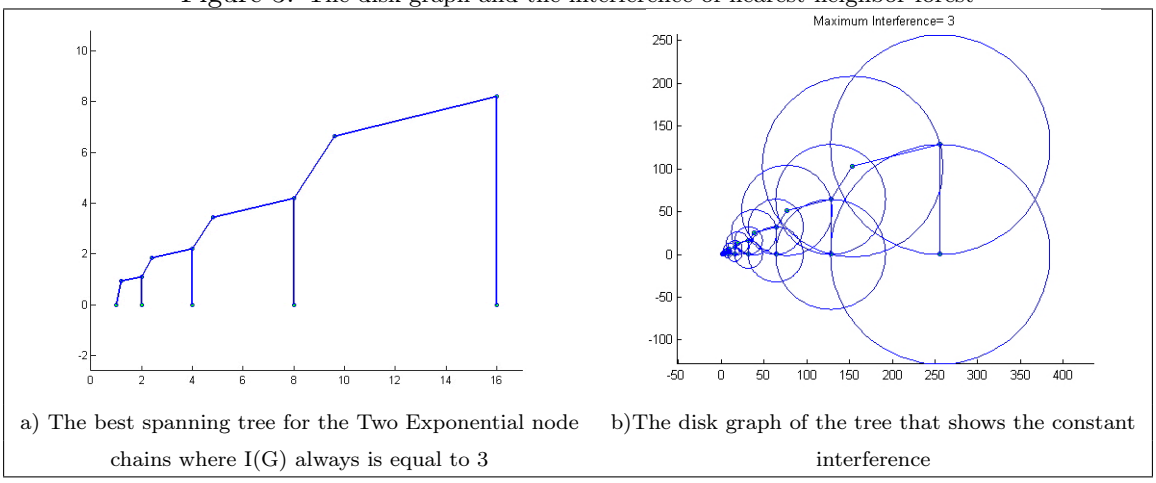
The algorithm proposed in [9] finds a subgraph for the exponential node chain (Figure 2(a)) with $I(G) \in O(\sqrt{n})$. And they proved bellow theorem:



a) The nearest neighbor forest and disk graph for exponential chain, $I(G) = n - 2 = \Omega(n)$

b)The nearest neighbor forest and disk graph for Two Exponential node chains, $I(G) = \lfloor \frac{n}{3} \rfloor + 2 = \Omega(n)$

Figure 3: The disk graph and the interference of nearest neighbor forest



a) The best spanning tree for the Two Exponential node chains where I(G) always is equal to 3

b)The disk graph of the tree that shows the constant interference

Figure 4: Suggested topology for Two Exponential node chains in [9]

Theorem 1: Given an exponential node chain $G = (V, E)$ with $n = |V|$, $\sqrt{n}$ is a lower bound for the interference $I(G)$.

They also proposed a topology with constant interference for the Two Exponential node chains which is depicted in Figure 4 but there is no algorithmic method which generate automatically similar subgraph.

According to the construction of the exponential node chain, only nodes connecting to at least one node to their right increase the v_1's (leftmost node) interference. They called such a node a hub and define it as follows [9]:

Definition 3. Given a connected topology for the exponential node chain $G = (V, E)$. A node $v_i \in V$ is defined to be a hub in G iff there exists an edge (v_i, v_j) with $i < j$.

The A_{exp} algorithm which is proposed in [9] for exponential chain is as follows:

The algorithm starts with a graph $G_{exp} = (V, E_{exp})$, where V is the set of nodes in the exponential node chain and E_{exp} is initially the empty set. Following the scan-line principle, A_{exp} processes all nodes in the order of their occurrence from left to right. Initially, the leftmost node is set to be the current hub h. Then, for each node v_i, A_{exp} inserts an edge $\{h, v_i\}$ into E_{exp}. This is repeated until $I(G_{exp})$ increases due to the addition of such an edge. Now node v_i becomes the current hub and subsequent nodes are connected to v_i as long as $I(G_{exp})$ the overall interference does not increase. Figure 5 shows the resulting topology when A_{exp} algorithm is used for the exponential node chain with 17 nodes and I(G)=6.

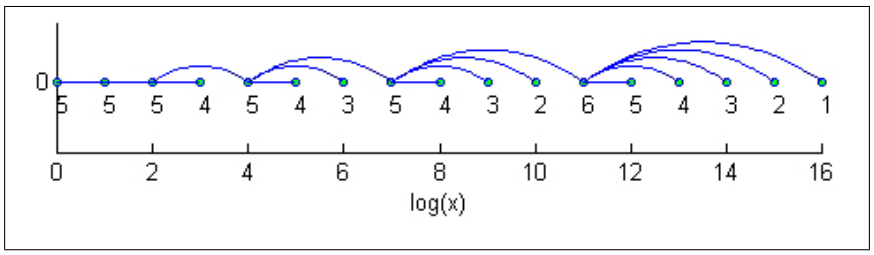


Figure 5: The result topology of A_{exp} algorithm for exponential node chain. The interference of the exponential node is bounded by $O(\sqrt{n})$. For clarity of representation edges are depicted as arcs and x dimension is shown in logarithmic scale. The interference of each node is wrote under the node position

In the next section a new algorithm which is the extension of A_{exp} is proposed for the nodes in the plane. This algorithm finds the best sub graph with minimum interference. The output of proposed algorithm for exponential chain is very similar to A_{exp}'s output with equal complexity. Also it shows the subgraph with constant interference for Two Exponential node chains.

4 Proposed Algorithm (A_{pln})

The following algorithm A_{pln} is the extension of A_{exp} for finding a spanning tree with minimum interference of nodes distributed in the plane. The A_{pln} the same as A_{exp} is an iterative algorithm which generates the resulting graph in n steps. Where n is the number of nodes. In the beginning the result graph has no edge and in each step of algorithm one edge is added to result graph until the connected result graph is generated.

Suppose $G_{in} = (V_{in}, E_{in})$ be the input graph, where $V_{in} = \{v_1, v_2, ..., v_n\}$ is the set of n separate nodes in the plane. (x_i, y_i) is the coordinate of v_i. Also E_{in} is the $n \times n$ adjacent matrix of G_{in} with e_{ij}^{in} elements. e_{ij}^{in} is the Euclidean distance between i'th and j'th nodes.

$$e_{ij}^{in} = |v_i - v_j| = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \qquad \forall i, j = 1, 2, ..., n \tag{3}$$

The steps of generating the spanning tree with minimum interference from G_{in} are as follows:
Step 1) Preparing data: Generate the adjacent list of G_{in}'s edges. The adjacent list of edges of a graph is a list of triples (i, j, e_{ij}); where i points to i'th vertex and j points to j'th vertex of graph and e_{ij}) is the weight of edge (in our problem it is the Euclidean distance). G_{in} is a complete graph and the adjacent matrix E_{in} is symmetric so the adjacent list would have $\frac{n(n-1)}{2}$ elements.

Step 2) Preparing data: Sort the elements of adjacent list according to the weight of each edge and call the new list $SrtE$[1] . Thus the elements of the $SrtE$ are arranged as follow:

$$EdgeLength(SrtE_i) \leq EdgeLength(SrtE_{i+1}) \qquad \forall i = 1, 2, ..., \frac{n(n-1)}{2} \tag{4}$$

Step 3) Find Start Point: Find the smallest edge from $SrtE$ list; Set $E_{min} = SrtE_1$.
Step 4) Initialization: Suppose that $G_{pln} = (V_{pln}, E_{pln})$ is the result graph. Set $V_{pln} = V_{in}$ and E_{pln} contains of only the smallest edge $E_{min} = (h, k, e_{hk})$, in other word all elements of E_{pln} are valued with zero instead of e_{hk}^{pln} and e_{kh}^{pln}. Initialize the maximum interference of result

[1]SrtE: obtained from Sorted Edge list

graph $MaxI_{pln}$ with:

$$\begin{cases} a) & V_{pln} = V_{in} \\ b) & e_{ij}^{pln} = 0 \qquad \forall i,j = 1,2,...,n \\ c) & e_{hk}^{pln} = e_{kh}^{pln} = e_{hk}^{in} \\ d) & MaxI_{pln} = 1 \end{cases} \tag{5}$$

Step 5) Initialize the Active Vertexes List: The nodes which are in sub connected graph are named as **Active Vertexes**. At the beginning sub connected graph consist of only the smallest edge. The Active Vertex list AV is a $1 \times n$ array; iff the node v_i be an Active Vertex the i'th element of $AV(AV_i)$ gets value 1 otherwise its value is 0.

$$AV = \begin{cases} 1 & \text{if } i = h,k \\ 0 & \text{if } otherwise \end{cases} \qquad \forall i = 1,2,...,n. \tag{6}$$

In other word:

	1	2	...	k	...	h	...	N
$AV=$	0	0	0	1	0	1	0	0

(7)

Where k,h are the corresponding nodes of smallest edge.

Step 6) Check the Status: while all nodes of G_{pln} are not connected together repeat steps 7 to 9.

Step 7) Generate Candidate Edges: According to AV list, generate the Active Edge List (AE). AE is a subset of $SrtE$; An edge from $SrtE$ exist in AE iff only one of its endpoints have value 1 in AV.

$$\begin{aligned} &\text{For Each } SrtE_k \in SrtE \text{ if } (AV_i = 1 \oplus AV_j = 1) \quad \rightarrow \quad \text{Add } SrtE_k \text{ to } AE \\ &WhereSrtE_k = (i,j,e_{ij}) \end{aligned} \tag{8}$$

The xor symbol ($\oplus$) in the predicate part of relation (8) ensures the intolerance of multi selecting one edge and recursion in final subgraph. According to (4) and (8) we can write:

$$EdgeLength(AE_i) \leq EdgeLength(AE_{i+1}) \quad \forall i = 1,2,...,size(AE) - 1 \tag{9}$$

Step 8) Find The best Edge: Select the edge $AE_m = (p,q,e_{pq}^{in})$ from AE , which leads to minimum increase on $MaxI_{pln}$ when is added to G_{pln}. After adding the AE_m to G_{pln} , update the AV and $MaxI_r$. The algorithm of finding the AE_m will be explained in next section. The specification of AE_m is as follow:

$$\exists AE_m \in AE | \forall AE_j \in AE \rightarrow I(G_{pln}, E_{pln} \cup AE_m) \leq I(G_{pln}, E_{pln} \cup AE_j) \tag{10}$$

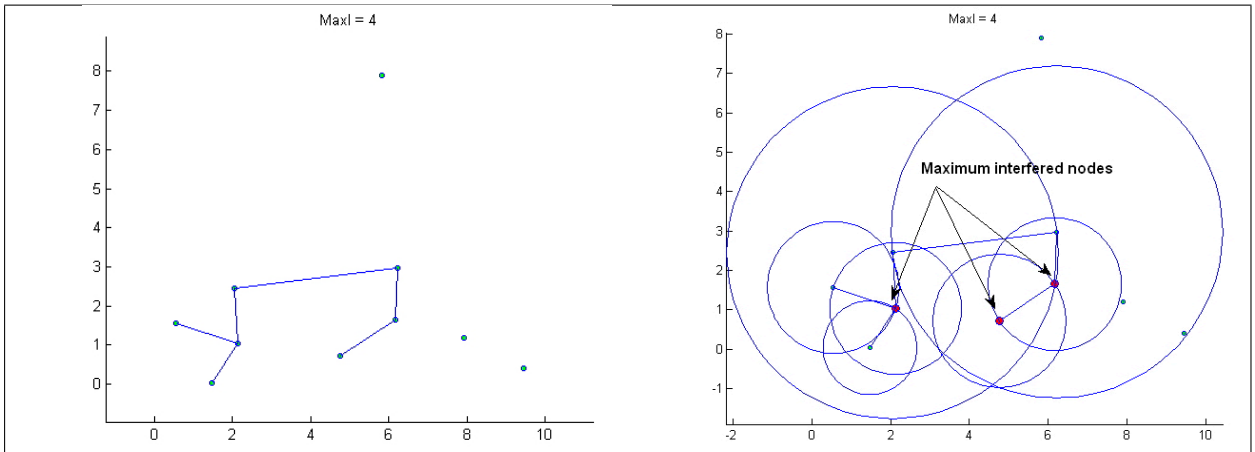
$I(G,E)$ determines the maximum interference of graph G according to adjacent matrix E. updating the variables is done as follow:

$$\begin{aligned} &(a) \quad Suppose \quad AE_m = (p,q,e_{pq}^{in}) \\ &(b) \quad E_{pln} = E_{pln} \cup AE_m \quad \underrightarrow{means} \quad e_{pq}^{pln} = e_{pq}^{in}, e_{qp}^{pln} = e_{qp}^{in} \\ &(c) \quad MaxI_{pln} = I(G_{pln}, E_{pln}) \\ &(d) \quad AV_p = 1, AV_q = 1 \end{aligned} \tag{11}$$

Go to step 6 for checking the status.

Step 9) Finalizing: the obtained $G_{pln} = (V_{pln}, E_{pln})$ is the spanning tree with minimum interference for the input distribution of nodes V_{pln} in the plane.

Step 10) Finish



a) connected component consist of 7 nodes and 3 new vertex with 21 possible edge are candidate to joint to connected component

b)Three vertexes have maximum interference and their interference is 4.

Figure 6: The topology of graph after 6 step with I(G)=4

5 Find best Edge Algorithm

In Step 8 of the A_{pln} algorithm we have addressed an algorithm which has found the best edge from AE list. The brief introduction of finding the best edge from AE list is as follow: Suppose that the current state of algorithm is a connected sub graph with m vertex and $(n-m)$ disconnect nodes which are shown in Figure 6(a) we want to expand the connected sub graph with minimum possible increase of $MaxI$.

Lemma 1) Adding a new vertex to connected sub graph in the worst case result in two unit increase of the current Maximum interference.

Proof) suppose R_i determines the distance from V_i to its farthest neighbour and D_i is a disk with radius R_i and centered by V_i. The disk D_i shows the domain of i'th transmitter and the nodes inside the D_i affect by V_i. When a new vertex V_k is added to the sub graph by connecting to V_j, the new disk D_k is added to current disks and the radius of D_j may increase. Thus in the worst case D_j and D_k will dominate all other vertexes. So the Current max-interfered node is affected by two new transmitter signals and for this reason we will have two unit increase of $MaxI$. And in the best case $MaxI$ not changes.

$$\text{Adding a new Edge in the worst case} \quad \rightarrow \quad MaxI_{new} = MaxI_{old} + 2 \tag{12}$$

Figure 6 shows the increase of interference when a new vertex is added to sub graph.

The algorithm of finding the best edge is as follows:

Step 1) as lemma 1 determined, In the worst case we will have 2 unit increase on Maximum interference of the graph so repeat the bellow Steps for $\triangle I = 0, 1, 2$

Step 2) For Each AE_i from AE list repeat bellow

Step 3) Set $MaxI_{new} = I(G_{pln}, E_{pln} \cup AE_i)$

Step 4) if $MaxI_{new}$ is equal to $(MaxI_{npln} + \triangle I)$ determine the AE_i as the best edge and go to Step 6, else check for next Edge.

Step 5) Set $\triangle I = \triangle I + 1$ and go to Step 2.

Step 6) Finish.

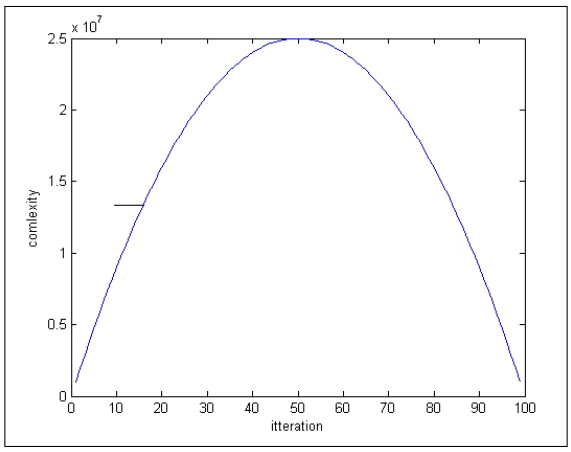


Figure 7: Relation of iteration and complexity of find the best edge for n=100.

The complexity of determining the best edge is as follow:

$$\begin{aligned} \text{Complexity} &= O(3 * length(AE) * O(I(G,E))) \\ &= O(3 * m(n-m) * n^2) \\ &= O(mn^3 - m^2n^2) \end{aligned} \tag{13}$$

Where m is number of active vertexes and n is the number of total vertexes. The relation (13) shows that in the first iteration ($m = 1$) the complexity of finding the best edge is $O(n3)$ and in the final Step ($m = n - 1$) it is also $O(n3)$ but in the middle iteration when $m = n/2$ the complexity is as follow:

$$If(m = n/2) \rightarrow O(mn^3 - m^2n^2) = O(\frac{n^4}{2} - \frac{n^4}{4}) = O(\frac{n^4}{4}) = O(n^4) \tag{14}$$

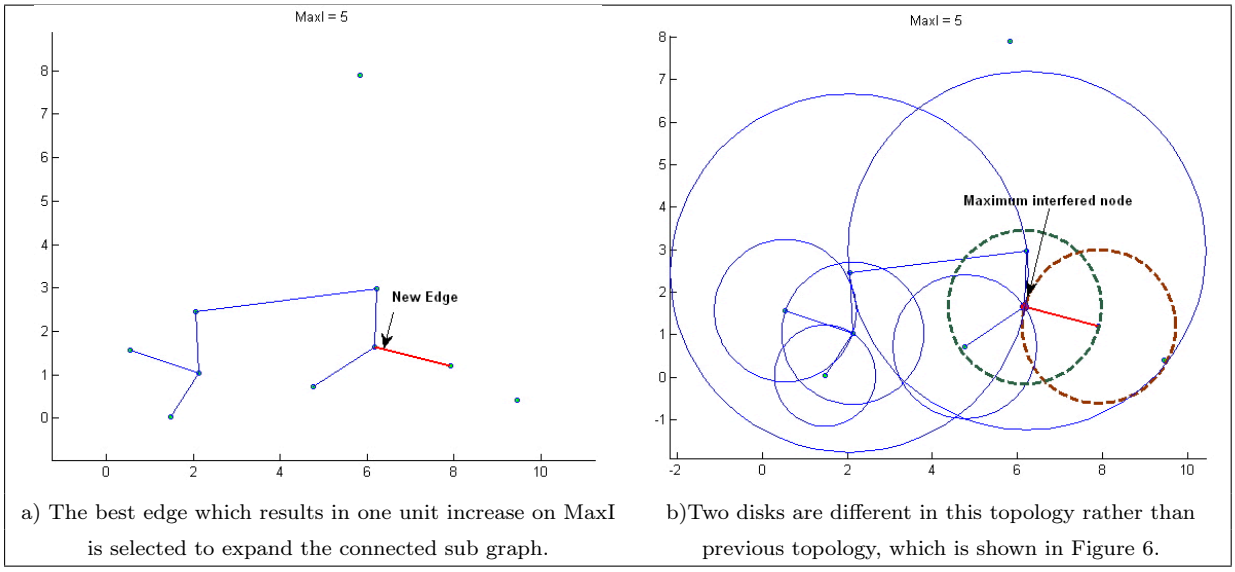
The relation between complexity of determination of best edge and iteration is depicted in Figure 7.

Applying the Best Edge algorithm for the node distribution which was shown in Figure 6 is depicted in Figure 8.

6 Experimental Results

At first the proposed algorithm A_{pln} is compare with A_{exp}. As depicted in Figure 9 the interference of the exponential chain for both algorithms are equal but the topologies are different.

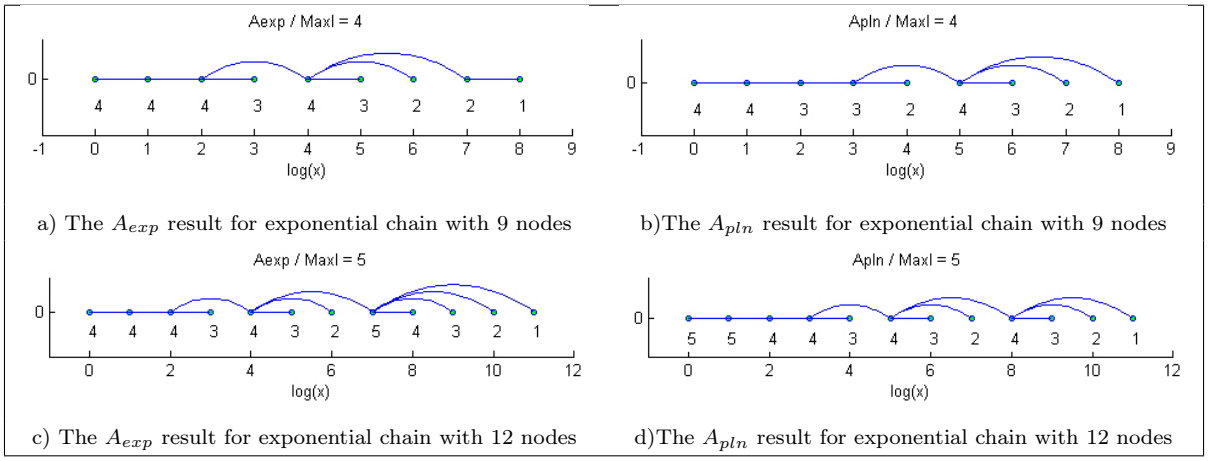
For A_{exp} the order of nodes is important but the A_{pln} does not have any precondition on the distribution. In following the strength of A_{pln} for finding the spanning tree with minimum interference for Two Exponential node chains and some special similar distributions is depicted. Figure 10 depicts the resulting topology if A_{pln} is applied to the Two Exponential node chains.The most important note on this topology is that the A_{pln} algorithm does not know this distribution but the resulting topology is the same as best topology which is created by human brain in [9]. For random distribution of nodes in plane the A_{pln} algorithm always suggest a better topology with equal or smaller interference value rather than nearest neighbor forest. Another test case is α-Spiral Node Chain. In α-Spiral node chain which is shown in Figure 12, the k'th node is placed in $(2^k \cos(ak), 2^k \sin(ak))$.



a) The best edge which results in one unit increase on MaxI is selected to expand the connected sub graph.

b)Two disks are different in this topology rather than previous topology, which is shown in Figure 6.

Figure 8: The topology of graph after 7 step with I(G)=5



a) The A_{exp} result for exponential chain with 9 nodes

b)The A_{pln} result for exponential chain with 9 nodes

c) The A_{exp} result for exponential chain with 12 nodes

d)The A_{pln} result for exponential chain with 12 nodes

Figure 9: The spanning trees obtained by A_{exp},A_{pln}, algorithms for exponential chain. For clarity of representation edges are depicted as arcs and x dimension is shown in logarithmic scale. The interference of each node is wrote under the node position

The results of applying the A_{pln} and nearest neighbor forest (A_{MST}) to some α-Spiral Node Chains are depicted in Figure 13.

7 Conclusion

In this paper a general algorithm named A_{pln} for finding the spanning tree of separate nodes in the plane has been proposed. The A_{pln} algorithm presents an iterative routine for minimizing the maximum interference of the resulting spanning tree.

At the beginning the resulting tree has only one edge which is the smallest edge in the input graph, until all input nodes are not connected together, the algorithm adds a new edge to the resulting tree.

For adding a new edge to sub graph the best edge which imposes minimum increase on the interference of all nodes from all available edges is selected. In section 6 the experimental result of using A_{pln} and A_{exp} for some special distributions are depicted and all of them show the good performance of the proposed algorithm. The A_{pln} is a general algorithm for any two dimensional

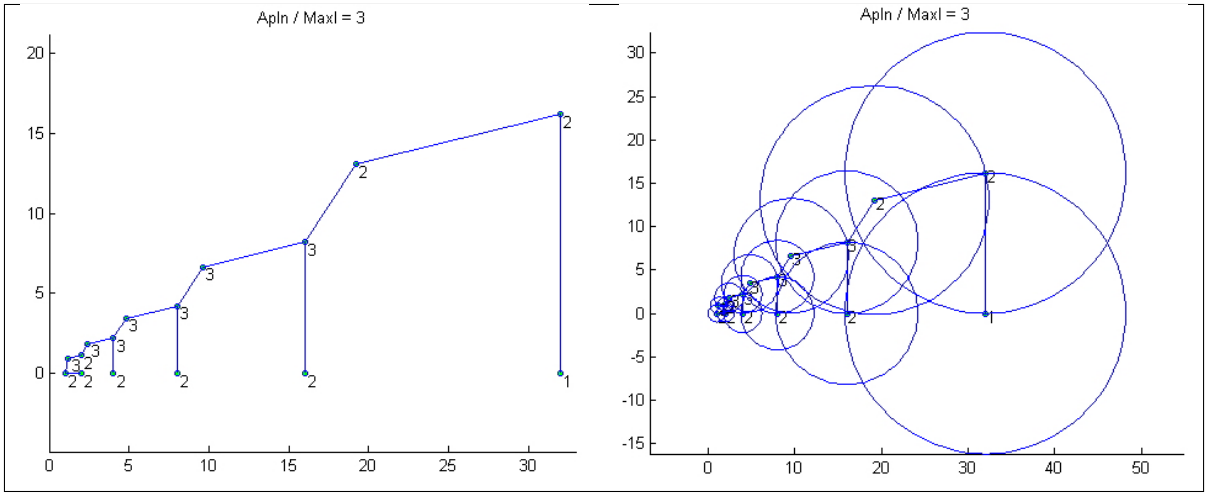
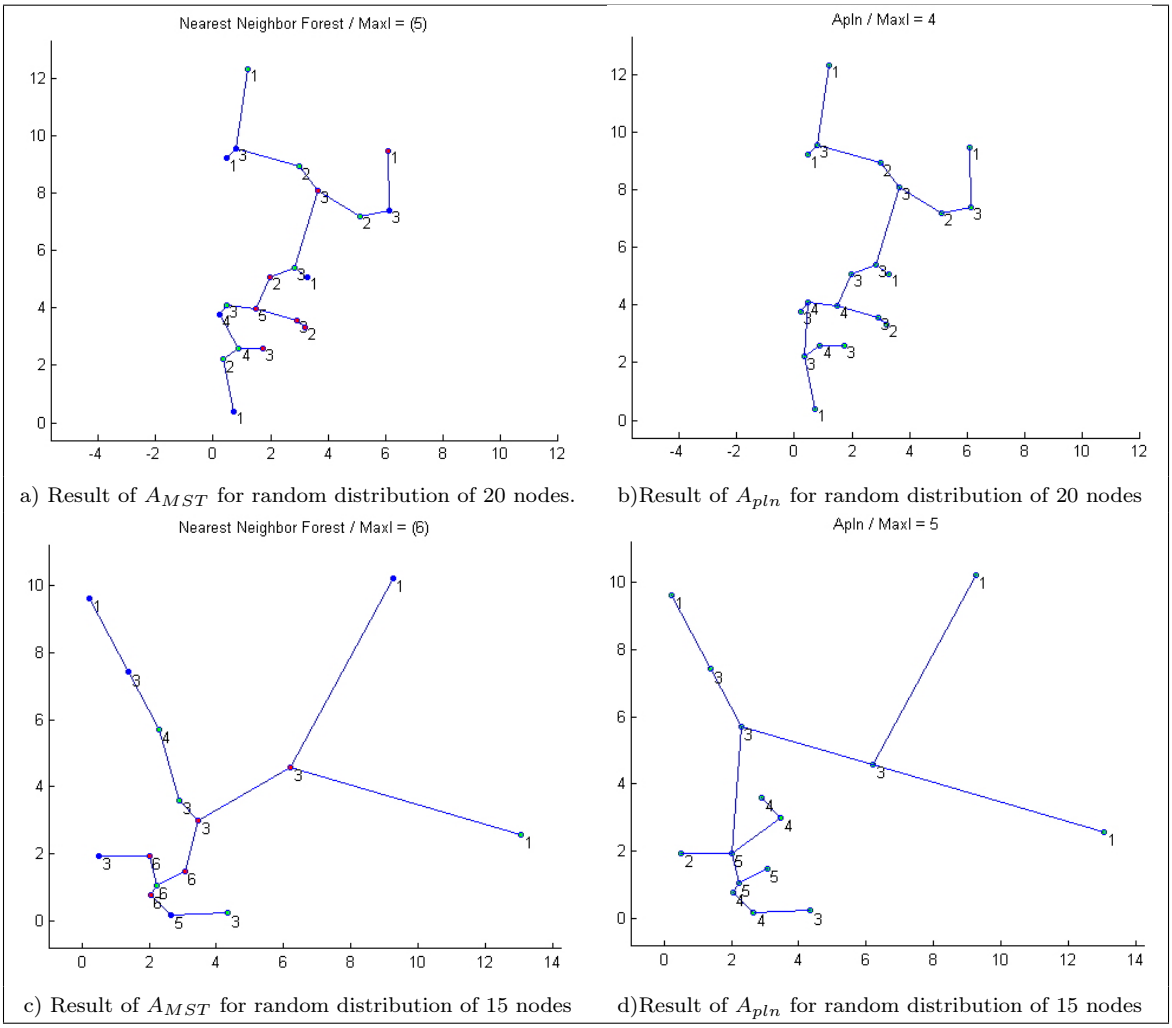


Figure 10: Constant Interference for Two Exponential node chains obtained by A_{pln} algorithm. The interference of each node is wrote beside the node position



a) Result of A_{MST} for random distribution of 20 nodes.

b)Result of A_{pln} for random distribution of 20 nodes

c) Result of A_{MST} for random distribution of 15 nodes

d)Result of A_{pln} for random distribution of 15 nodes

Figure 11: Result of applying nearest neighbor forest and A_{pln} algorithms on random distribution of nodes.

distribution and it has no limit or special conditions for the input distribution and there is no need to inform it about the input distribution, the algorithm itself finds the best spanning

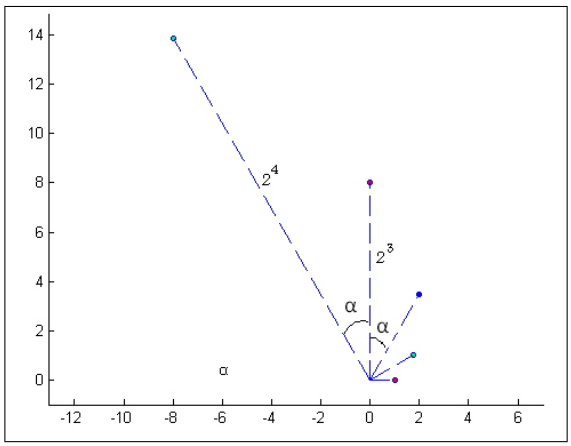
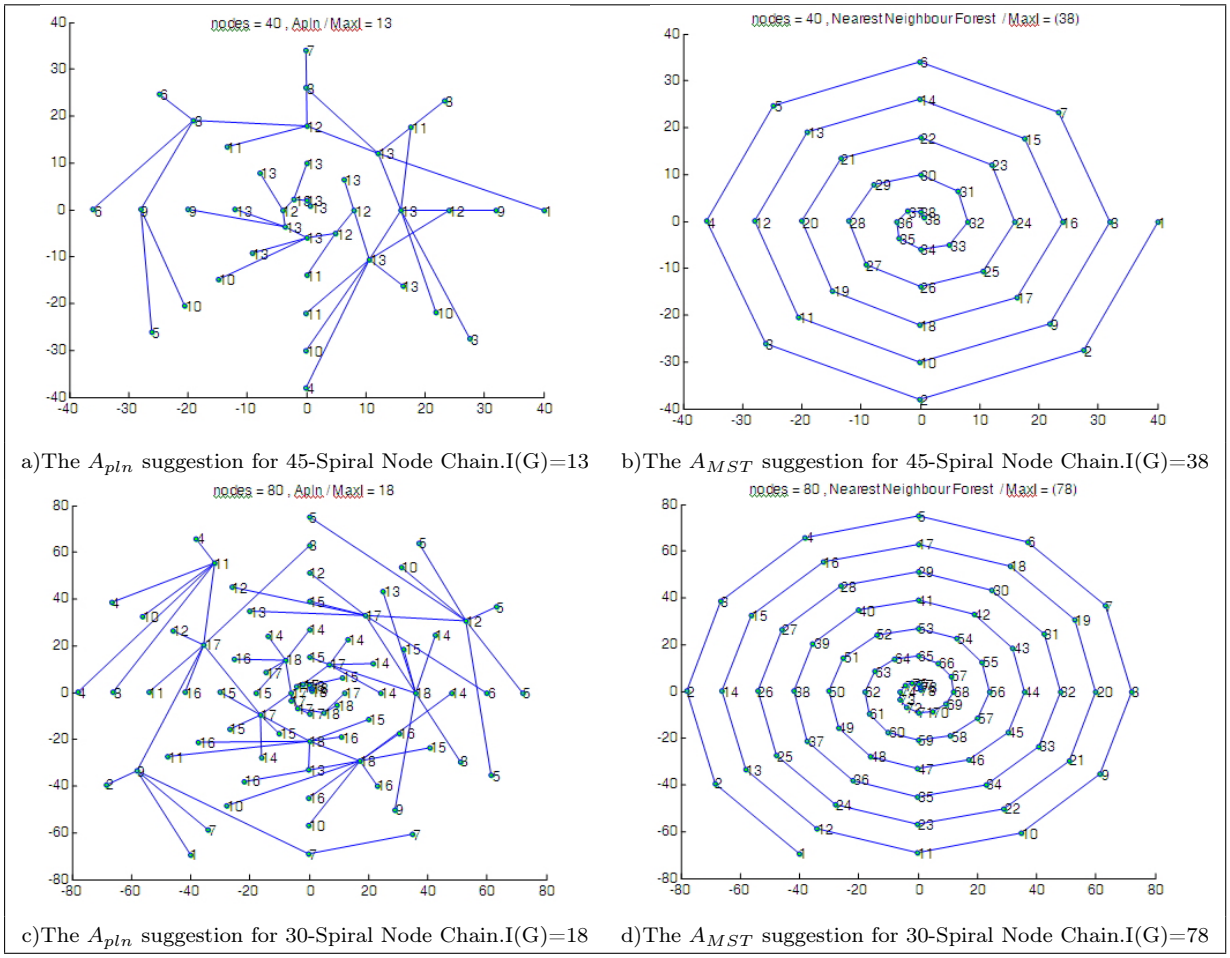


Figure 12: α-Spiral Node Chain.



a)The A_{pln} suggestion for 45-Spiral Node Chain.I(G)=13 b)The A_{MST} suggestion for 45-Spiral Node Chain.I(G)=38

c)The A_{pln} suggestion for 30-Spiral Node Chain.I(G)=18 d)The A_{MST} suggestion for 30-Spiral Node Chain.I(G)=78

Figure 13: 45-Spiral Node Chain with 40 and 30-Spiral Node Chain with 80 nodes and proposed topologies with A_{pln} and A_{MST}. Note that for clarity of representation, the k'th node is positioned in $(kcos(\alpha k), ksin(\alpha k))$ instead of $(2^k cos(\alpha k), 2^k sin(\alpha k))$. The interference of each node is wrote beside the node position.

tree with minimum interference. In this paper we can not compute the final interference order according to the input size but for some special distributions which are criterions for interference problem the resulting topology is generated and all resulting topologies are satisfactory. For the

future work by mathematical relations we will find the order of final topology according to the number of input nodes for the worst case of node distribution.

Bibliography

[1] T. Locher, P. von Rickenbach, and R. Wattenhofer: Sensor networks continue to puzzle: Selected open problems. *In Proc. 9th Internat. Conf. Distributed Computing and Networking (ICDCN)*, 2008

[2] P. Santi: Topology Control in Wireless Ad Hoc and Sensor Networks. *Wiley* , 2005

[3] X.Y. Li , W.Z. Song , W. Wan: A Unified Energy Efficient Topology for Unicast and Broadcast. *In: Proc. of the 11th Int. Conf. on Mobile Computing and Networking (MOBICOM)*, 2005

[4] M. Damian, S. Pandit, S. Pemmaraju: Local Approximation Schemes for Topology Control. *In: Proc. of the 25th Annual ACM Symposium on Principles of Distributed Computing (PODC)*, 2006

[5] M. Burkhart, P. von Rickenbach, R. Wattenhofer, A. Zollinger: Does Topology Control Reduce Interference?. *In: Proc. of the 5th ACM Int. Symposium on Mobile Ad-hoc Networking and Computing (MobiHoc)*, 2004

[6] K. Moaveni-Nejad, X.Y. Li: Low-Interference Topology Control for Wireless Adhoc Networks. *Ad Hoc & Sensor Wireless Networks: An International Journal 1(1-2)*,2005

[7] T. Johansson, L. Carr-Motyčková: Reducing Interference in Ad hoc Networks through Topology Control. *In: Proc of the 3rd ACM Joint Workshop on Foundations of Mobile Computing (DIALM-POMC)*, 2005

[8] M. Benkert, J. Gudmundsson, H. Haverkort, A. Wolff: Constructing Interference- Minimal Networks. *Computational Geometry: Theory and Applications*, 2007

[9] P. von Rickenbach, S. Schmid, R. Wattenhofer, A. Zollinger: A Robust Interference Model for Wireless Ad-Hoc Networks.*In: Proc. of the 5th Int. Workshop on Algorithms for Wireless, Mobile, Ad Hoc and Sensor Networks (WMAN)*, 2005

[10] M.M. Halldórsson, T. Tokuyama: Minimizing Interference of a Wireless Ad-Hoc Network in a Plane. *In: Proc. of the 2nd Int. Workshop on Algorithmic Aspects of Wireless Sensor Networks(ALGOSENSORS)*, 2006.

[11] T. Moscibroda, R. Wattenhofer: Minimizing Interference in Ad Hoc and Sensor Networks. *In: Proc. of the 3rd ACM JointWorkshop on Foundations of Mobile Computing (DIALM-POMC)*, 2005

[12] K. Buchin: Minimizing the Maximum Interference is Hard. *CoRR 2008 and sited in arXiv/0802.2134* , 2008

[13] N. Clark, C. J. Colbourn, D.S. Johnson: Unit Disk Graphs. *Discrete Mathematics, 86:165-177*, 1990

[14] M. Heide, C. Schindelhauer, K. Volbert, M. Gruenewald: Energy: Congestion and Dilation in Radio Networks. *In Proc. of the 14th Annual ACM Symposium on Parallel Algorithms and Architectures (SPAA), pages 230-237*, 2002

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 2 (June), pp. 285-301

Authoring Adaptive Hypermedia using Ontologies

H. Jung, S. Park

Hyosook Jung, Seongbin Park
Korea University,
Anam-dong, Seongbuk-gu, Seoul, Korea
{est0718, hyperspace}@korea.ac.kr

Abstract: Adaptive hypermedia has been developed to overcome the problems of disorientation by providing personalized presentation and link structure. An adaptive hypermedia system consists of an adaptation model, a domain model, and a user model. The user model describes various aspects of a user such as interests, knowledge, preferences, etc. The domain model describes the whole knowledge accessible in adaptive hypermedia. The adaptation model consists of adaptation rules that define both how to generate the personalized presentation and update the user model [12]. Authoring adaptive hypermedia typically starts by designing a domain model so that apprropiate adaptation model and user model can be created based the domain model. While there are authoring tools developed for creating domain model of an adaptive hypermedia, authors need to manually create basic concepts as well as their relationships for a domain of interests. In this paper, we present a system that transforms an ontology into the domain and adaptation model of adaptive hypermedia so that the authors can generate adaptive hypermedia easily. The system transforms classes and relationships between the classes defined in OWL ontology [7] into concepts and relationships between the concepts defined in the domain model of AHA! system which is one of best known open source general-purpose adaptive hypermedia systems [1]. Using our system, authors can utilize well-defined knowledge structure in ontologies for authoring adaptive hypermedia. On top of this, since designing domain model is generally an initial step to author adaptive hypermedia, our system can help authors reduce tasks to create adaptive hypermedia by automatically generating domain model from an ontology.
Keywords: Ontology, Adaptive hypermedia.

1 Introduction

A hypermedia system allows users to freely navigate through a large hyperspace. However, the navigational freedom makes the users experience the problems of disorientation and cognitive overload [11]. Adaptive hypermedia has been developed to overcome these problems by providing personalized content and navigational support based on user model, domain model and adaptation model. The user model describes various aspects of a user such as interests, knowledge, preferences, etc. The domain model describes the whole knowledge accessible in adaptive hypermedia. The adaptation model consists of adaptation rules that define both how to generate the personalized presentation and update the user model [12]. Authoring adaptive hypermedia is a difficult and complex task that involves designing different levels of abstraction and multiple links, defining adaptation on an abstract conceptual level, etc. On top of this, the authored materials are hardly reused because there is no standardized approach to adaptive techniques and behavior [8–10].

In this paper, we propose a system that helps authoring adaptive hypermedia using ontologies. Domain model describes concepts related to a domain and their relationships. An ontology also represents the knowledge of a domain by a set of concepts within the domain and the relationships between the concepts [6]. Based on this similarity, our system converts an ontology into a domain model of an adaptive hypermedia system so that authors can easily create domain models using ontologies. Our system has been implemented using AHA! which is an open source general-purpose adaptive hypermedia [1]. The reason that we selected AHA! system is that it is one of the best open source adaptive hypermedia systems which can be easily used in educational environments. Moreover the web site about the system containts a lot of helpful documents that explain various aspects as well as applications of the system [26].

Figure 1 shows the whole structure of the proposed system. Once an OWL document is read, Parser extracts concepts and their relationships by using Protégé-OWL API [17]. It looks for elements of the OWL vocabulary corresponding to the concepts and their relationships of the domain model. For example, <owl:Class> is a concept and <rdfs:subClassOf> is a concept hierarchy. <owl:ObjectProperty> is a prerequisite relationship between two concepts. Its domain class defined in <rdfs:domain> is a prerequisite for its range class defined in <rdfs:range>. Then, Converter creates a domain model by transforming the concepts and their relationships to the format for AHA! Graph Author that is one of authoring tools used in AHA! system.

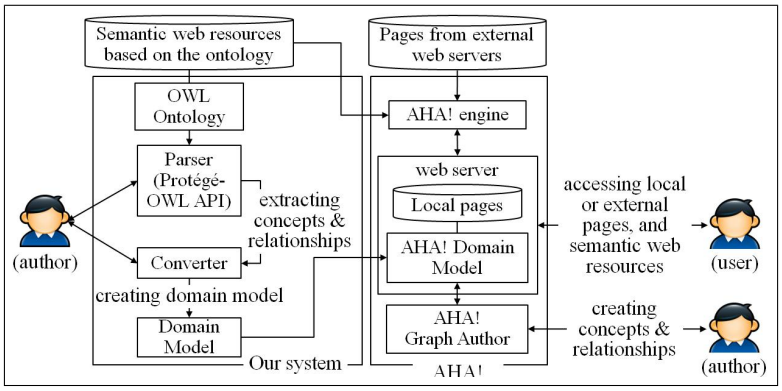


Figure 1: The structure of our system

We have conducted a pilot evaluation for our system with users who had experiences using AHA! system, and their comments were positive in that our system did help easily construct domain models as well as adaptation models using ontologies that could be found on the Web. Since the number of ontologies that are available online is ever increasing [28], it is possible to use well-designed ontologgies for domain model construction even if authors are not experts in the domain of interests.

This paper is structured as follows. Section 2 describes related works. Section 3 describes the structure of the system as well as the transformation steps from an ontology into the domain model and the adaptation model in detail. Section 4 describes illustrative examples that show how the proposed system works. Applications of the system and pilot evaluation about the system are explained in section 5. Section 6 concludes the paper and describes the future works.

2 Related Works

The Semantic Web is an environment where Web contents are represented in a form that is machine processable [14]. There are several languages to represent machine interpretable content on the Web. XML offers a surface syntax for structured documents and XML Schema is a lan-

guage for restricting the structure of XML documents. RDF is a data model for objects and their relations and supports a simple semantics for the data model. RDF Schema is a vocabulary for representing properties and classes of RDF resources. OWL adds more vocabulary for describing properties and classes: among others, relations between classes, cardinality, equality, richer typing of properties, characteristics of properties, and enumerated classes [7]. An ontology is a specification of an abstract data model that is independent of its particular form [6]. Ontologies are used by people, databases, and applications in order to share domain information. They contain computer-usable definitions of basic concepts in the domain and the relationships among them [15]. OWL Web Ontology Language is one of languages to represent machine interpretable content on the Web. An OWL ontology consists of a set of constraints on sets of classes and types of relationships between them. A class contains individuals which are instances of the class, and other subclasses. A property specifies class characteristics. It can be either a datatype or object property. Instances are individuals that belong to the classes defined. OWL supports various operations on classes such as union, intersection and complement [7].

Adaptive hypermedia systems such as AHA! [1], Mag [25], AHyCo [27], etc. have been developed to overcome the problem of disorientation by providing appropriate contents for users with different backgrounds and interests based on the user model, adaptation model, and domain model.

In AHA!, an author needs to design the overall conceptual structure by using AHA! authoring tools. They allow the author to define the domain model for an application along with the adaption model related to it, which means that the author defines requirements for each concept and a set of generate rules that represent connections between concepts [2]. In AHA!, most authors do not have to know about the adaptation model because the rules are generated automatically by Graph Author. When a concept is created in Graph Author, a set of attributes and adaptation rules is generated. It has templates for different types of concept relationships [3].

Mag [25] is a tutoring system that supports learning programming languages in various courses. It provides adaptive learning and personalization of a content delivery based on learner model. It consists of the adaptation model, learner model, application model and domain model. It uses ontologies to present a domain, building learner model and presenting activities in the system in order to achieve knowledge sharing and reuse, learner modeling and extension of a system. The domain model presents storage for all concepts, tutorials and tests. It describes how the information content is structured. Instructors can create the domain model based on domain ontology by using an authoring tool.

AHyCo (Adaptive Hypermedia Courseware) [27] is an adaptive educational hypermedia system to create and reuse adaptive courseware. It focuses on adaptive navigation support and lessons sequencing. It consists of the domain model, the student model, and the adaptation model. It allows authors to design the courseware such as creating concepts, linking concepts by prerequisite relationships, and generating test questions. It offers a graphic editor for concept networks that enables the authors to define the prerequisite relationships with a drag-and drop interface like AHA! Graph Author. Authoring a course contains the development of both the network of lessons and the tests that represents the domain model of the course. For constructing a course, the authors create a set of hypermedia fragments that represent the content of a lesson. A set of lessons are connected by prerequisite relationships and grouped into a module. A course is a set of modules connected by prerequisite relationships. The authors define the prerequisite graphs of concepts and modules and determine the difficulty of lessons and tests. They do not have to consider any other adaptation rules because AHyCo automatically generates the adaptation rules based on the definition of prerequisites and the assignment of the difficulty level for lessons and tests. All information about the domain model is stored in a database.

MOT [5] is an adaptive hypermedia authoring system. It implements domain maps, goal and

constraint maps, user and presentation maps, and adaptation maps. The domain maps structure and organize the resources of the learning environment. They consist of hierarchical domain concept maps. The goal and constraints maps contain all resources and links between them. They add all the necessary pedagogic material and linking for students. The user maps contain all necessary variables and initial values to represent the user. The presentation maps make different presentation according to the physical properties and environment. The adaptation maps describe the dynamic of the adaptation process based on the LAG model [4] which is a three-layer model and classification method for adaptive techniques; direct adaptation rules, adaptation language and adaptation strategies. MOT makes authors create a new concept map of a lesson that consists of concepts and their attributes. There are many tasks to create the concept map such as adding each concept, naming them, selecting their attributes, etc. However, our system helps the authors create the concept map without those manual tasks because it provides a hierarchy of basic concepts with their attributes by converting the ontology related to the lesson. The authors just edit or change the created concept map.

The interoperability between different Adaptive Educational Hypermedia (AEH) systems has been investigated by using conversion systems. Interbook to AHA! compiler translates the source format for Interbook to AHA! with Layout model which presents concepts (pages) to authors [20]. MOT to AHA! conversion engine translates from MOT used as a an authoring system to AHA! used as a delivery system for AES. It uses the common adaptation format (CAF) domain and lesson map descriptions and the adaptive strategies written in LAG to create adaptive presentations in AHA! [21]. MOT to WHURLE converter translates MOT to WHURLE used as a delivery system for reusing the authored content on a different system [22]. However, these are limited to conversions between different AEH systems. The conversion should be based on popular standards for reducing re-designing material each time one AEH system moves to another and encouraging the use of AEH systems. [23] presents a solution for AEH interfacing via web services. It uses WSDL for conversion of data between MOT and WHURLE. [24] describes the integration of the generic AH authoring environment, MOT, into a semantic desktop environment.

Our system enables users to access Semantic Web data represented in RDF or OWL format. It stores the URIs of the classes of the ontology as the resource path of the corresponding concept when converting the ontology. The users can access the information about the classes via their URIs. Our system focuses on creating adaptive hypermedia by using an OWL ontology which is one of the core techniques to build the semantic web.

Figure 2 shows related research areas to our work.

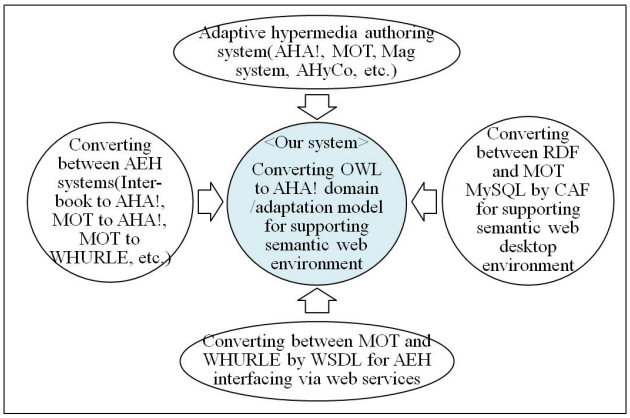


Figure 2: Related works to our research

3 The Structure of the System

In this section, we describe the structure of the proposed system.

Figure 3 shows how OWL ontology are mapped into AHA! domain / adaptation model.

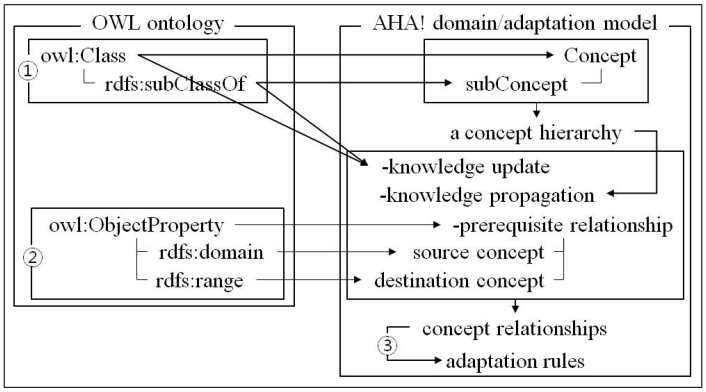


Figure 3: Conversion from an OWL ontology to AHA! domain model and adaptation model

First, our system extracts all named classes in an OWL ontology and transforms the classes into concepts of AHA! domain model in order to generate concepts. A class is transformed into a concept. If the class has subclasses, the system transforms the subclasses into subconcepts of the corresponding concept. It generates a concept hierarchy by using the association between the concepts and their subconcepts. It also defines the information of a concept that is its name and optional description. The system assigns the name of a class to the name of a concept. The URI (Uniform Resource Identifier) of the class is converted to the resource of the concept. It describes a class with subclasses as an abstract concept and the subclasses as page concepts. In AHA! domain model, an abstract concept does not have a resource and a page concept must have resources. However, the optional description can be changed by AHA! Graph Author. It saves the concepts hierarchy and the concept information in AHA! domain model. The Graph Author allows authors to edit concepts and concept relationships of the domain model. The Graph Author also saves the domain model in another authoring format used by Concept Editor (.aha) as translating concept relationships into adaptation rules. Users can edit concepts, attributes, and adaptation rules using Concept Editor which can control the functionality of AHA! such as concepts, attributes and adaptation rules.

Second, our system extracts all OjectProperties in the OWL dontology and their domain and range classes in order to generate concept relationships. It converts an ObjectProperty into a prerequisite relationship of AHA! domain model. In AHA!, when concept A is a prerequisite of concept B, users should read about concept A before continuing with concept B. The concept A is a source concept and the concept B is a destination concept. The system transforms a domain class of a ObjectProperty into a source concept of the prerequisite relationship. It does a range class of the ObjectProperty into a destination concept of the prerequisite relationship. (i.e., the domain class is converted into a prerequisite for the range class.) In addition, the system assigns additional concept relationships to each concept. Each concept with a page has a concept relationship called knowledge update. When a page of a concept is read, its knowledge is updated. The concept hierarchy is used of knowledge propagation. When the knowledge of a concept is changed, the change is propagated to the concepts that are higher than the concept in the concept hierarchy. It saves the concepts relationships in AHA! domain model.

Third, all concept relationships in AHA! are translated into adaptation rules in adaptation model by AHA! Graph Author. The rules are executed conditionally when a page related to a

concept is accessed.

Figure 4 shows how our system fits into AHA! system.

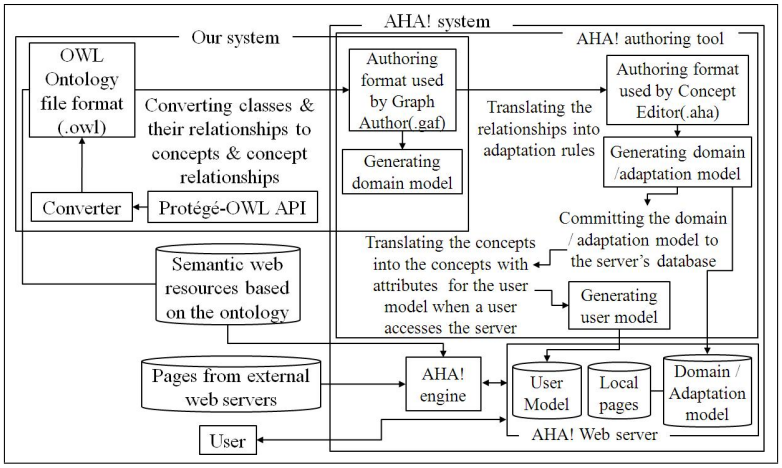


Figure 4: Authoring adaptive hypermedia using our system in connection with AHA! system

When the authors commit the application to the AHA! server's database such as XML or MySQL representation, the AHA! domain and adaptation model are automatically created. If end-users access the application, their user models are created as well. When users access AHA!, it creates their own profiles based on the user model. If a user requests a page, the request triggers the adaptation rules that update the user's profile. The requested page is presented based on the updated profile of the user. The links on the page are also provided based on a ĄŽsuitabilityĄŻ requirement that is part of the domain and adaptation model.

4 Illustrative Examples

In this section, we show two simple examples that illustrate how our system converts OWL ontologies into the domain models and adaptation models of AHA! system. In addition, we show how the user models of AHA! can be created.

4.1 Example 1

Figure 5 shows the structure of an OWL ontology (sweb.owl) about Semantic web that defines 11 classes and 4 object properties. Class Basic has 4 subclasses such as Metadata, Agents, Ontologies, and Logic. Class Technologies has 5 subclasses such as XML, RDF, RDFS, OWL and Rules. ObjectProperty Metadata_is_describedBy has domain class Metadata and range class XML and RDF. ObjectProperty Ontology_is_describedBy has domain class Ontologies and range class XML, RDF, RDFS, and OWL. ObjectProperty Logic_is_describedBy has domain class Logic and range class Rules.

Figure 6 shows the structure of the domain model of AHA! system that is generated from the ontology (sweb.owl) in figure 5 using our system. The name of the OWL ontology (i.e., "sweb" in sweb.owl) becomes the name of an AHA! application or course. Our system extracts all named classes and transforms them into concepts of the domain model such as sweb.Basic, sweb.Technologies, etc. It also generates a concept hierarchy based on the hierarchy between the classes. For example, concept sweb.Basic has subconcepts such as sweb.Metadata, sweb.Agents, sweb.Ontologies, and sweb.Logic. Concept sweb.Technologies also has subconcoept such as sweb.XML, sweb.RDF, sweb.RDFS, sweb.OWL, and sweb.Rules. The system also transforms

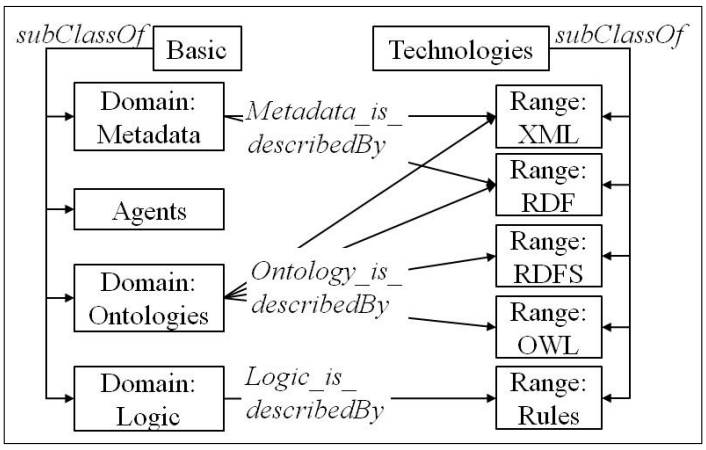


Figure 5: The structure of an OWL ontology about Semantic Web

ObjectPoperties into prerequisite relationships of the domain model. When concept A is a prerequisite of concept B, the concept A is a source concept and the concept B is a destination concept. The system transforms a domain class into a source concept and a range class into a destination concept. For example, ObjectProperty Logic_is_describedBy is transformed into a prerequisite relationship. The domain class Logic of Logic_is_describedBy becomes a source concept and its range class Rules becomes a destination concept.

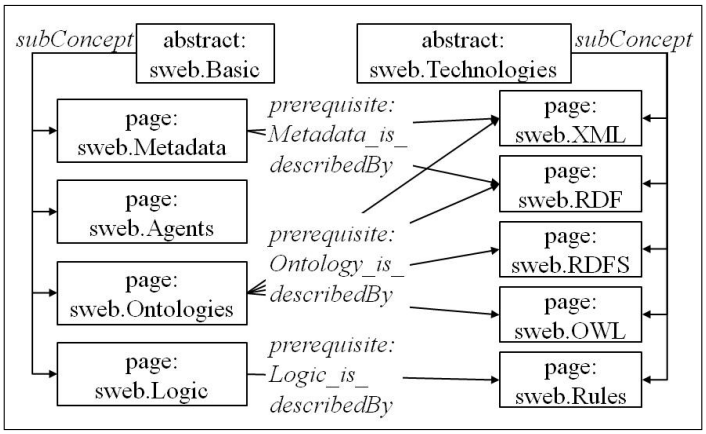


Figure 6: The structure of AHA! domain model transformed from the ontology in figure 5

After the OWL ontology is converted into a domain model, the domain model is saved as a file in an authoring format edited by the Graph Author (sweb.gaf). When an author commits the file to the serverĄŻs database, all concept relationships defined in the domain model are automatically translated into adaptation rules. The adaptation rules and the domain model are saved in a file which is an authoring format edited by the Concept Editor (sweb.aha).

Adaptation rules that are related to concept “sweb.Rules” in “sweb” application which are automatically generated are as follows.

```
<generateListItem isPropagating="true" >
      <requirement>! sweb.Rules.suitability &amp;&amp;
                sweb.Rules.knowledge &lt; 10</requirement>
      <trueActions>
       <action>
```

```
        <conceptName>sweb.Rules</conceptName>
        <attributeName>knowledge</attributeName>
        <expression>35</expression>
       </action>
      </trueActions>
    </generateListItem>
```

The above rule is connected to concept relationship "knowledge update". When the rule is triggered, its condition called "requirement" is checked. It is executed when sweb.Rules.suitability is false and sweb.Rules.knowledge is lower than 10. The action will assign the value 35 to the knowledge of sweb.Rules.

```
<generateListItem isPropagating="true" >
      <requirement>true</requirement>
      <trueActions>
       <action>
        <conceptName>sweb.Technologies</conceptName>
        <attributeName>knowledge</attributeName>
        <expression>sweb.Technologies.knowledge +
               (0.2 * _sweb.Rules.knowledge)</expression>
       </action>
      </trueActions>
    </generateListItem>
```

The above rule is connected to concept relationship "knowledge propagation". When the knowledge of a concept is changed, the change is propagated to the concepts that are higher in the concept hierarchy. Concept sweb.Rules is a child of concept sweb.Technologies in "sweb" application. The action will add the knowledge of sweb.Technologies to 20% of the knowledge of sweb.Rules and assign the value to the knowledge of sweb.Technologies.

```
<attribute name="suitability" type="bool" isPersistent="false"
        isSystem="false" isChangeable="false">
   <description>the suitability of this page</description>
   <default>((sweb.Logic.knowledge &gt; 0))</default>
</attribute>
```

The above rule is connected to concept relationship "prerequisite relationships". sweb.Logic is a prerequisite for sweb.Rules. The suitability of sweb.Rules depends on the knowledge of sweb.Logic. The rule requires the knowledge of sweb.Logic to be higher than 0 in order to access to the page of sweb.Rules.

When a user logs in the AHA! server, the user model of the user is automatically generated. The user model consists of a set of concepts with attributes. It contains an overlay model which means that there is a concept in the user model for every concept in the domain model. After the domain and adaptation model are constructed, the following attributes related to concept sweb.Rules are added to the existing user model when a user accesses the AHA! server.

```
<record>
    <key> sweb.Rules.access</key>
    <type>3</type>
    <persistent>false</persistent>
```

```
    <value>false</value>
  </record>
  <record>
    <key> sweb.Rules.interest</key>
    <type>1</type>
    <persistent>true</persistent>
    <value>0</value>
  </record>
  <record>
    <key> sweb.Rules.suitability</key>
    <type>3</type>
    <persistent>false</persistent>
    <value>false</value>
  </record>
  <record>
    <key> sweb.Rules.visited</key>
    <type>1</type>
    <persistent>true</persistent>
    <value>0</value>
  </record>
  <record>
    <key> sweb.Rules.knowledge</key>
    <type>1</type>
    <persistent>true</persistent>
    <value>0</value>
  </record>
```

4.2 Example 2

Figure 7 shows another the correspondence between an OWL ontology (health.owl) and AHA! domain model. Our system transforms all named classes such as Health, Disease, and Food into concepts of the domain model such as health.Health, health.Disease, and health.Food. It also generates a concept hierarchy based on the hierarchy between the classes. For example, class Health has subclass Disease and Food. Concept health.Health has subconcept health.Disease and health.Food. The system also transforms ObjectPoperties into prerequisite relationships of the domain model. For example, ObjectProperty causedBy is transformed into a prerequisite relationship. Domain class Disease of causeBy becomes a source concept and its range class becomes a destination concept.

After the OWL ontology is converted into a domain model, the domain model is saved as a file in an authoring format edited by the Graph Author (health.gaf). When authors commit the file to the serverĄŻs database, all concept relationships are automatically translated into adaptation rules. A file including the domain model and adaptation rules are saved in an authoring format edited by the Concept Editor (health.aha).

The adaptation rules that are related to concept health.Food in “health” application which are automatically generated are as follows.

```
<generateListItem isPropagating="true" >
   <requirement>! health.Food.suitability &amp;&amp;
         health.Food.knowledge &lt; 10</requirement>
   <trueActions>
```

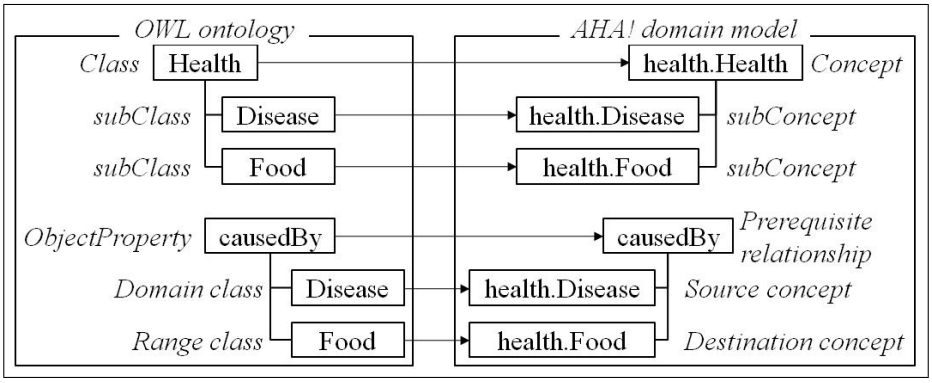


Figure 7: Correspondence between an OWL ontlogy and AHA! domain model

```
      <action>
        <conceptName>health.Food</conceptName>
        <attributeName>knowledge</attributeName>
        <expression>35</expression>
      </action>
   </trueActions>
</generateListItem>
```

The above rule is connected to concept relationship "knowledge update". When the rule is triggered, its condition called "requirement" is checked. It is executed when health.Food.suitability is false and health.Food.knowledge is lower than 10. The action will assign the value 35 to the knowledge of health.Food.

```
<generateListItem isPropagating="true" >
   <requirement>true</requirement>
   <trueActions>
      <action>
        <conceptName>health.Health</conceptName>
        <attributeName>knowledge</attributeName>
        <expression>health.Health.knowledge +
                 (0.5 * _health.Food.knowledge)</expression>
      </action>
   </trueActions>
</generateListItem>
```

The above rule is connected to concept relationship "knowledge propagation". When the knowledge of a concept is changed, the change is propagated to the concepts that are higher in the concept hierarchy. Concept "Food" is a child of concept "Health" in "health" application. The action will add the knowledge of health.Health to 50% of the knowledge of health.Food and assign the value to the knowledge of health.Health.

```
<attribute name="suitability" type="bool" isPersistent="false"
       isSystem="false" isChangeable="false">
   <description>the suitability of this page</description>
   <default>((health.Disease.knowledge &gt; 0 ))</default>
</attribute>
```

The above rule is connected to concept relationship "prerequisite relationships". health.Disease is a prerequisite for health.Food. The suitability of health.Food depends on the knowledge of health.Disease. The rule requires the knowledge of health.Disease to be higher than 0 in order to access to the page of health.Food.

When a user logs in the AHA! server, the user model of the user is automatically generated. The user model consists of a set of concepts with attributes. It contains an overlay model which means that there is a concept in the user model for every concept in the domain model. After the domain and adaptation model are constructed, the following attributes related to concept health.Food are inserted to the existing user model when a user accesses the AHA! server.

```
<record>
    <key>health.Food.access</key>
    <type>3</type>
    <persistent>false</persistent>
    <value>false</value>
  </record>
  <record>
    <key>health.Food.interest</key>
    <type>1</type>
    <persistent>true</persistent>
    <value>0</value>
  </record>
  <record>
    <key>health.Food.suitability</key>
    <type>3</type>
    <persistent>false</persistent>
    <value>false</value>
  </record>
  <record>
    <key>health.Food.visited</key>
    <type>1</type>
    <persistent>true</persistent>
    <value>0</value>
  </record>
  <record>
    <key>health.Food.knowledge</key>
    <type>1</type>
    <persistent>true</persistent>
    <value>0</value>
  </record>
```

5 Application

In this section, we show how the proposed system can be used for authoring adaptive hypermedia. In addition, we describe the pilot evaluation about our system.

Assume that an author wants to create an AHA! application that introduces basic concepts of biology. Although it is possible to design a domain model by using AHA! authoring tools, building a well organized domain model is not a simple task because it involves choosing proper concepts and logically making relationships between them. If the author can use existing ontologies written

by domain experts, creation of a domain model might be easier. So, the author decides to use our system and finds ontologies on the Web using ontology search engines such as Swoogle [18]. Figure 8 shows the structure of an OWL ontology about biology in the tree-view of Protégé-OWL [16]. Our system transforms the OWL ontology document to an XML document that is

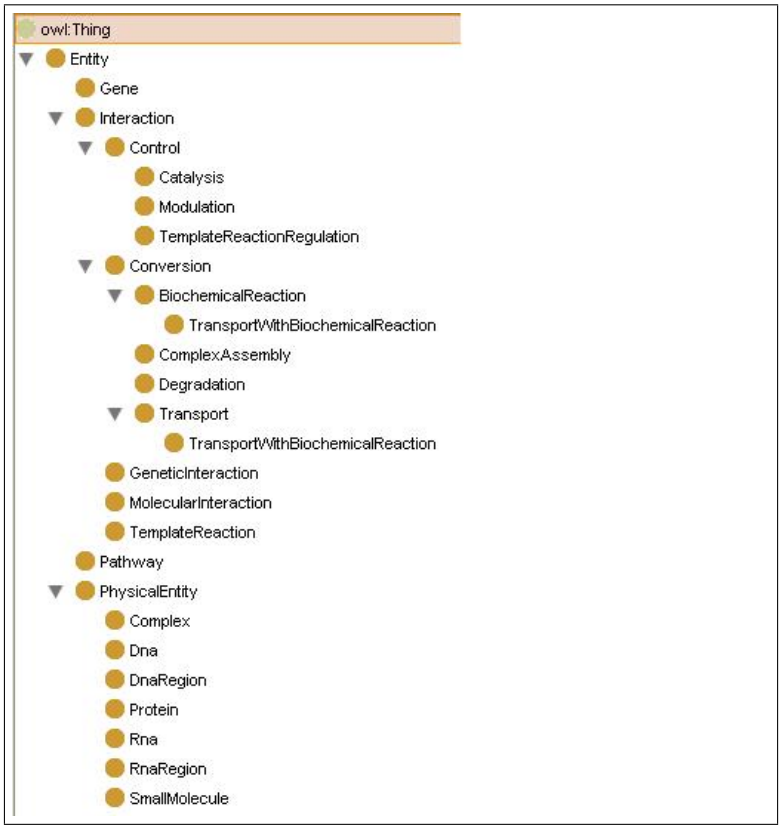


Figure 8: biopax.owl ontology

recognized in AHA! Graph Author. If the author opens an OWL document (biopax.owl) and clicks a button "Converter" on the top, the OWL document is transformed to the XML document (biopax.gaf). (figure 9) The author opens the transformed document in AHA! Graph Author and edits parts of the file appropriately. (figure 10) Similarly, if an author wants to create a domain model about a koala, the author can start with a simple ontology (koala.owl) provided by Protégé site [17]. Figure 11 shows the constructed domain model (koala.gaf) in Graph Author.

We conducted a pilot evaluation with a few PhD students who had experiences using AHA! system. They also had general understanding about OWL ontlogies. They were asked to create domain models about data structure and heals sciences. While they could use Graph Author, they found it difficult to model concept hierarchies for the domain areas. On the other hand, they felt it easy to use existing ontologies found on the Web to create domain models since these ontogies are often times defined by experts in the fields and contain basic concepts and their relationships. Here are some of their comments.

1. Authors can easily find the essential concepts about a domain and their relationships using ontologies defined by domain experts if the authors are not familiar with the domain of interests.

2. Authors can save time and efforts authoring the domain model and adaptation model because they are generated quickly with our system once an OWL ontology is given.

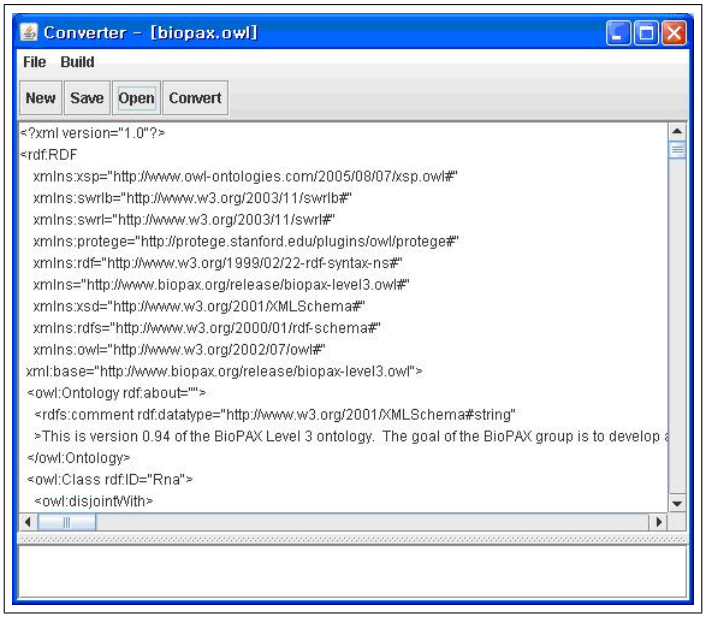


Figure 9: biopax.gaf that was transformed from biopax.owl ontolgy using our system

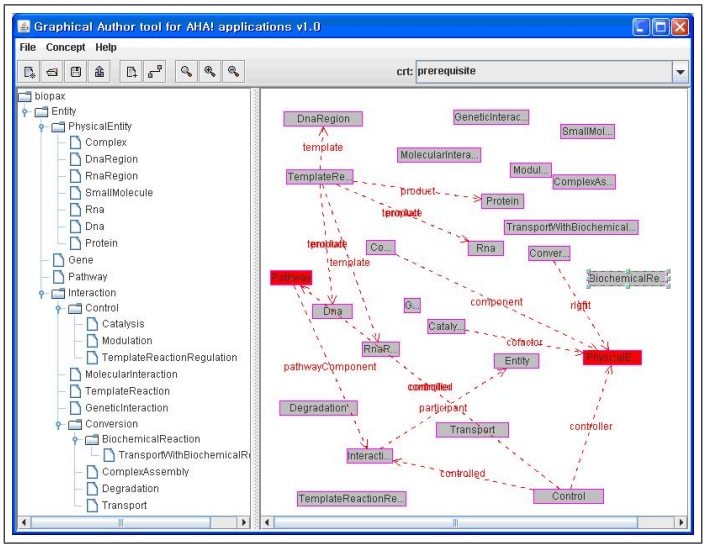


Figure 10: biopax.gaf in Graph Author

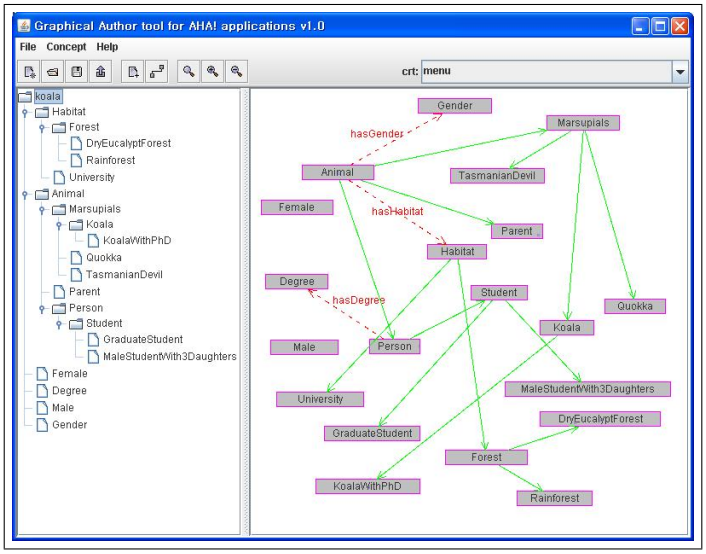


Figure 11: koala.gaf in Graph Author

3. It is helpful when authors have it difficult in conceptualizing a particular domain. In addition, it is easy to reorganize and expand the concepts and concept relationships.

6 Conclusions and Future Works

In this paper, we propose a system that can be used to author adaptive hypermedia using ontologies. Generating the domain model is a basic task for authoring adaptive hypermedia because it influences on generating the adaptation model and the user model. Our system makes it easy to create the domain model using an OWL ontology. The concepts and concept relationships are created automatically by our system and authors do not have to manually create them. When an author designs a domain model, it may not be easy to define concepts of a domain and their relationships. In ontologies, concepts about a domain and their relationships are already defined by domain experts. So, an ontology can serve as a starting point to define a domain model and the author can edit the domain model appropriately.

We plan to extend our system so that it can transform AHA! domain models into OWL ontologies so that it supports authoring a linked data [19]. An AHA! author creates a structure of the domain model consisting of concepts and concepts relationships and writes local resources consisting of a set of xhtml pages. Every page must have a corresponding concept. Both local and remote pages can be associated with a concept. The resources in a domain model can be considered as a linked data. Using our system, authors can gather lots of OWL ontologies transformed from the domain models and create linked data if our system can enable authors to search resources and link one to others. The linked data can be also reused in AHA! system.

We also plan to combine our system with other adaptive hypermedia systems such as Mag system and AHyCo system.

In the case for Mag system, our system helps the instructors define the concepts of the domain model of Mag system by using the domain ontology. It extracts all concepts defined in the ontology and it converts them into the concepts of the domain model. Then, it generates the domain model of Mag system that consists of the converted concepts. If the instructors can use the concepts already defined by domain experts or other instructors, they can easily design the domain model. (Figure 12)

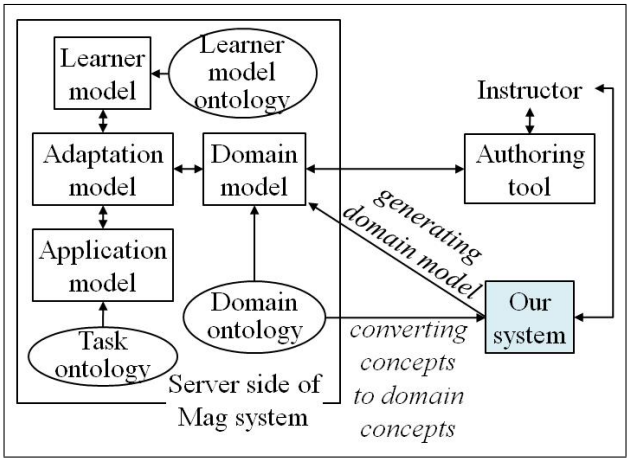


Figure 12: Combination with Mag system

In the case of AHyCo system, our system can help authors define the lessons, group them into modules, and connect between them by the prerequisite relationships. Even if our system cannot support the creation of the real content of each lesson such as text or image, it can offer the structure of the domain model by converting an ontology. It converts a class to a module, and an instance of the class to a lesson. The hierarchy of classes is transformed into the prerequisite relationships between modules. For example, a class is prerequisite for its subclasses. It means that the class is learned before its subclasses. The object property linking between instances of two classes is transformed into the prerequisite relationships between lessons. For example, an instance of its domain class is prerequisite for an instance of its range class. It means that the instance of the domain class is learned before the instance of the range class. Instead of the lessons created by Word, Excel or PowerPoint objects, it stores the URIs of classes or their instances as resources. Our system can provide the authors the predesigned domain model based on the ontology. (figure 13)

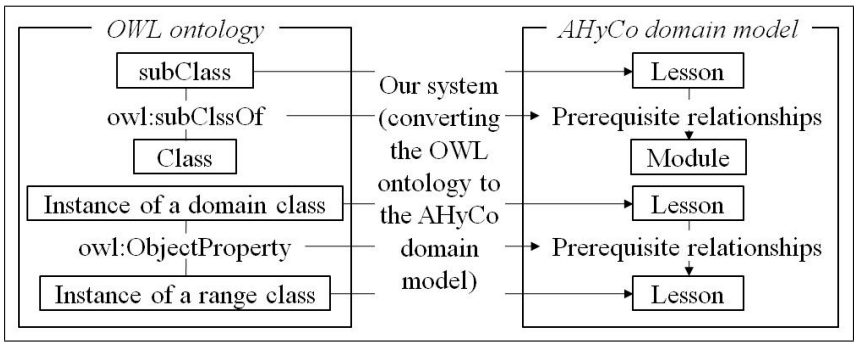


Figure 13: Combination with AHyCo system

Acknowledgement

Seongbin Park is the corresponding author. This research was supported by a College of Education of Korea University Grant in 2010.

Bibliography

[1] De Bra, P., Stash, N., Smits, D., Creating Adaptive Applications with AHA!, Tutorial for AHA! version 3.0, Tutorial at the AH 2004 Conference, Eindhoven, pp.4-20, 2004

[2] Stash, N., De Bra, P., Building Adaptive Presentations with AHA! 2.0, *Proceedings of the PEG Conference*, Saint Petersburg, 2003

[3] De Bra, P., Aerts, A., Berden, B., De Lange, B., Rousseau, B., Santic, T., Smits, D., Stash, N., AHA! The Adaptive Hypermedia Architecture, *Proceedings of the ACM Hypertext Conference on Hypertext and Hypermedia*, ACM, Nottingham, pp. 81-84, 2003

[4] Cristea, A., Calvi, L., The Three Layers of Adaptation Granularity, UM'03, LNCS 2702, Springer, 2003

[5] Cristea, A., Smits D., De Bra, P., Towards a Generic Adaptive Hypermedia Platform, a conversion case study, *Journal of Digital Information*, Vol.8, No.3., 2007

[6] Gruber, T.,Ontology, http://tomgruber.org/writing/ontology-definition-2007.htm

[7] McGuinness, D. L., van Harmelen, F., OWL Web Ontology Language Overview, W3C Recommendation, 2004, http://www.w3.org/TR/owl-features

[8] Wu, H., Houben,G. J., De Bra, P., Authoring Support for Adaptive Hypermedia Applications, *Proceedings ED-MEDIA'99*, Seattle, pp. 364-369, 1999

[9] Cristea, A. I., de Mooij, A., Designer Adaptation in Adaptive Hypermedia Authoring, ITCC'03, IEEE Computer Science, pp. 444-448, 2003

[10] Cristea, A., Evaluating Adaptive Hypermedia Authoring while Teaching Adaptive Systems, SAC'04, ACM, pp. 929-932, 2004

[11] De Bra, P., Houben, G.J., Wu, H., AHAM: A Dexter-based Reference Model for Adaptive Hypermedia, *Proceedings of the ACM Conference on Hypertext and Hypermedia*, Darmstadt, pp. 147-156, 1999

[12] Wu, H., De Kort, E., De Bra, P., Design Issues for General-Purpose Adaptive Hypermedia Systems, *Proceedings of the ACM Conference on Hypertext and Hypermedia*, Aarhus, pp. 141-150, 2001

[13] http://protege.stanford.edu/plugins/owl/ontologies.html

[14] Berners-Lee, T., Hendler, J., Lassila, O., The Semantic Web, *Scientific American Special online Issue*, 2001

[15] Heflin, J., Volz, R., Dale, J., Web Ontology Requirements, Proposed W3C Working Draft, 2002, http://km.aifb.uni-karlsruhe.de/projects/owl/index.html

[16] http://protege.stanford.edu/download/protege/3.4/installanywhere (Protégé editor download site)

[17] http://protege.stanford.edu/download/registered.html (Protégé-OWL source code download site)

[18] http://swoogle.umbc.edu

[19] Bizer, C., Heath, T., Idehen, K., Berners-Lee, T., Linked Data on the Web (LDOW 2008), *Proceedings WWW2008*, Beijing, China, 2008

[20] De Bra, P., Santic, T., Brusilovsky, P., AHA! meets Interbook, and more..., *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2003*, pp. 57-64, 2003

[21] Cristea, A., Smits, D., De Bra, P., Writing MOT, Reading AHA! converting between an authoring and a delivery system for adaptive educational hypermedia, *A3EH Workshop at AIED'05*, Amsterdam, the Netherlands, pp. 36-45, 2005

[22] Stewart, C., Cristea, R., Brailsford, T., Ashman, H., Authoring once, Delivering many: Creating reusable Adaptive Courseware, *Proceedings of the 4th IASTED International Conference on Web-Based Education*, WBE 2005, pp.21-23, 2005

[23] Meccawy, M., Celik, I., Cristea, A., Stewart, C., Ashman, H., Interoperable Adaptive Educational Hypermedia: A Web Service Definition, *Proceedings of Sixth IEEE International Conference on Advanced Learning Technologies (ICALT'06)*, pp. 639-641, 2006

[24] Hendrix, M., Cristea, A., and Nejdl, W., Authoring adaptive educational hypermedia on the semantic desktop, *International Journal of Learning Technology*, 3(3):230-251, 2007

[25] Klasnja-Milicevic, A., Vesin, B., Ivanovic, M., Budimac, Z., Integraion of Recommendations and Adaptive Hypermedia into Java Tutoring System, *Computer Science and Information Systems 2010 OnLine-First*, 2010

[26] http://aha.win.tue.nl/publications.html

[27] Hoic-Bozic, N., Mornar, V., AHyCo: a Web-Based Adaptive Hypermedia Courseware System, *Journal of Computing and Information Technology*, CIT 13, 3, p. 165-176, 2005

[28] Thomas, E., Alani, H., Sleeman, D., Brewster, C., Searching and Ranking Ontologies on the Semantic Web, *Workshop on Ontology Management: Searching, Selection, Ranking, and Segmentation, 3rd K-CAP* Banff, Canada. pp. 57-60, 2005

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 2 (June), pp. 302-311

Noise Characterization in Web Cameras using Independent Component Analysis

M.A.U. Khan, T.M. Khan, R.B. Khan, A. Kiyani, M.A. Khan

Mohammad Asmat Ullah Khan
Department of Electrical and Computer Engineering
Effat University Jeddah, Saudi Arabia.
mohammad_a_khan@yahoo.com

Tariq Mahmood Khan,
Muhammad Aurangzeb Khan
Department of Electrical Engineering
COMSATS Institute of Information Technology,
Islamabad, Pakistan.
tariq_mehmood@comsats.edu.pk,rabiya@ciit.net.pk

Rabia Bahadar Khan, Atiqa Kiyani
Department of Electrical Engineering
COMSATS Institute of Information Technology,
Abbotabad, Pakistan.
rabiya@ciit.net.pk,atiqa@ciit.net.pk

Abstract: An image captured by a web camera contains stationary and non-stationary noise patterns. These noise patterns are of three types i.e. Fixed Pattern Noise (FPN), Interactive Nose (IN) and Temporal Noise (TN). TN is an independent noise pattern and needs an algorithm that does exploit its higher-order dependencies. Previously, these noise patterns have been characterized using Principal Component Analysis (PCA). PCA is restricted to second order dependencies. In this paper Independent Component Analysis (ICA) has been investigated for actual TN noise. The experimental results demonstrates the effectiveness of the proposed method.
Keywords: fixed pattern noise, interaction noise, temporal noise, independent component analysis, principle component analysis.

1 Introduction

Web camera is a real-time device. It describes low-resolution digital video which is used for instant messaging or a PC video calling application. Just like any other electronic device, webcam's acquired images contain noise patterns. To be more specific, it can be claimed that a webcam image contains stationary and non-stationary noise patterns. Generally, the noise patterns are classified as Fixed Pattern noise (FPN), Temporal noise (TN), and Interaction noise (IN) [1].

The first noise pattern FPN is due to a combination of variations in image pixel geometry. It is observed that Fixed pattern noise emerge in very long exposures and is worsen by higher temperatures. Whereas, the second pattern Temporal noise (TN) fluctuates randomly from frame to frame and is characterized by intensity and color fluctuations near the actual image intensity. TN is present at any exposure length. The third noise pattern is referred to as Interaction Noise (IN), which is highly camera-dependent. IN is normally introduced by the camera when it reads data from the digital sensor. This type of noise is more visible in the shadows, or when an image has been excessively brightened.

There has been a considerable research literature available to discuss ways and means for identification of these noise patterns. One common and popular method recently introduced for identification of these noise patterns is based on Principle Component Analysis (PCA) [2]. PCA in general, is statistical method to find useful image representations. Any given image can be decomposed into its combination of the standard basis images. The main objective in PCA is to find an optimum set of base images to uncorrelate the image coordinates into PCA coefficients which cannot be linearly found from each other. It has been an established fact that PCA can only extract and exploit pair wise linear dependencies between pixels. In the joint distribution of PCA coefficients the high-order dependencies are still there. It has already been confirmed that the higher order relationship among the pixels contains most of the important information of an image [3]. Later on, a generalization of PCA, known as Independent Component Analysis (ICA) has indeed provided better recognition rate as compared to PCA when used on Face Recognition data [4]. Motivated by the ICA success we started investigating web camera images. PCA representation of web camera noise characteristics may not be able to capture adequately the high-order structure present in the image. Specifically the object independent noise pattern TN identification is in need of an algorithm that does exploit higher-order dependencies. ICA is one such generalization of PCA that exactly fits the bill.

ICA seeks a set of independent components instead of a set of orthogonal components. Two components are considered as an independent components if we have knowledge about one and don't know any thing about the other. This is a very strong condition than uncorrelated components. Normally, we call ICA a blind source separation. Since, we have to find out a set of original sources from an observed mixture in blind source separation problem.

This paper presents the conjecture that ICA, a generalization of PCA, provides a better identification of noise patterns. For the purpose of practical implementation of ICA, numerous methods are available. The algorithm employed in this research word is developed by [5]. This algorithm has proven its worth by separating randomly mixed auditory signals (the cocktail party problem), and for separating electroencephalogram (EEG) signals [6] and functional magnetic resonance imaging (FMRI) signals [7].

This paper is organized as follows. In Section II, different types of webcam noises has been classified. In Section III, the conventional methods used for characterization of noises in webcams have been introduced. In Section IV, proposed work will be explained and the claim that ICA performs better than PCA will be elaborated further. In section V, results will be provided with discussion. Conclusion will be presented in the last section.

2 Noise Characterization

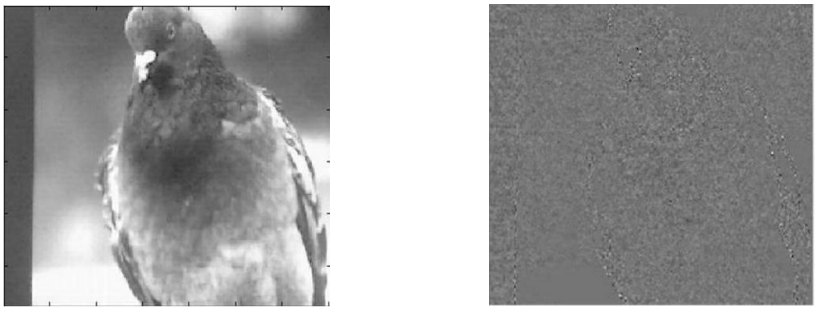
Electronic devices have some degree of noise when they receive or transmit data as a 'signal'. For television this signal is broadcast data; for webcams this signal is light which hits the camera sensors. The most common type of noises in webcams are:

1. Fixed pattern noise (FPN) is used to represent a noise pattern which is observed during longer exposure shots. It is a fixed pixel-to-pixel offset which is formed because of the combination of variations in image substrate material, pixel geometry and dark current as shown in Fig. 1.

2. Temporal noise (TN) sets the fundamental limit on image sensor performance, it is usually much more difficult to remove without degrading the image. Computers have a difficult time discerning TN from fine texture patterns such as those occurring in dirt or foliage, so removal of TN results in the loss of these textures as well, which is shown in Fig. 2.

(a) (b)

Figure 1: a) A sample image b) FPN extracted from a).



(a) (b)

Figure 2: a) A sample image b) TN extracted from a).

3. Interaction noise (IN) is highly camera-dependent, and is introduced by the camera when it reads data from the digital sensor. It is present at those locations in the image where the signal varies the most as shown in Fig. 3.

(a) (b)

Figure 3: a). A sample image b) IN extracted from a).

Separating these from each other helps to improve image quality as it either educe or eliminate some noise components.

2.1 Working with PCA

In this section, the salient features of PCA are elaborated. PCA uses Gaussian source models. PCA is a linear combination of basis vector. Let X be the original $n \times n$ matrix, T be the transformation matrix, Y would be the projection of the original matrix. As described in the following equation:

$$Y = TX$$

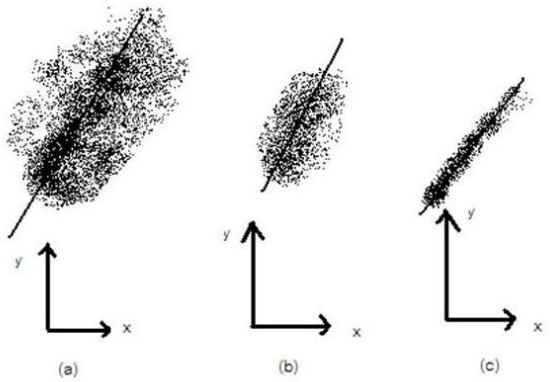


Figure 4: In a),b) and c) the best curve fitting line Y=TX is shown by dark line.

The probability of the data depends only on first and second-order statistics if we have Gaussian source. In PCA, the rows (M) represents the eigenvectors of the covariance matrix of the data [1].

2.2 Working with ICA

It is inadequate when the actual sources are non-Gaussian and we assume a Gaussian sources in PCA. Signals can be described well as linear combinations of sources with long tailed distributions [5]. In such cases, ICA has the following advantages over PCA:

1. It uniquely identifies the mixing matrix M.
2. It provides a better probabilistic model of the data.
3. It finds non-orthogonal basis which are able to reconstruct the data better than PCA even in the presence of noise.
4. It is sensitive to high-order statistics in the data, and not restricted to second order statistics only.

ICA can be implemented by taking X transpose (n-dimensional data vector) and organize the given data so that images are in the columns of X. This approach is illustrated in Fig. 4. In this approach, pixels are random variables and images are realizations. Here, the emphasis is on the independence of pixels or functions of pixels. For example, we have to consider pixel i and j independent if we are moving across the entire set of images. As, we can not predict the value taken by a specific pixel based on the corresponding value taken by pixel on the same image. This approach was inspired by [6] work on the ICs of natural images.

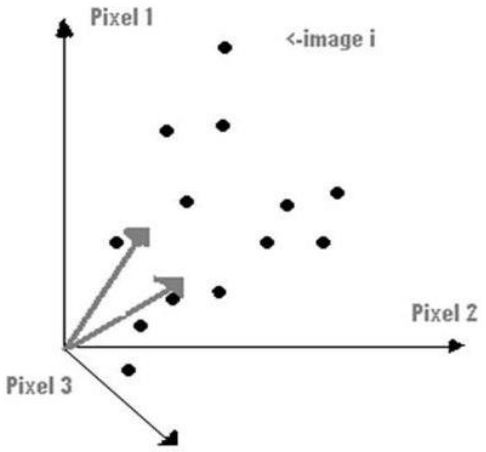


Figure 5: Finds weight vectors in the direction of statistical dependencies among the different images

3 Conventional method

In [1] the PCA method is applied to characterize the artifacts in fully digital image-acquisition systems. For practical application a video sequences was taken with USB cameras. Normally PCA method provides three different sets of spatial-temporal structures Fig. 6. Then a series of quasi sinusoidal patterns with different spatial frequencies were presented to the camera. With this approach, it is possible to characterize the noise using eigen-vector analysis.

In [2], Digital image noise has been characterized on the basis of RAW data. Noise characterization is done by shooting several frames both in complete darkness with several exposure times and also on several exposure levels with fixed field target. Based on these results, a number of parameters describing the camera's noise characteristics and sensitivity are measured.

4 Proposed Method

In digital cameras, there are three types of noises as discussed earlier. IN and TN are co-related to each other. PCA can be used efficiently to characterize these noises, but it is not able to provide adequate representation for the IN noise. One possible explanation that can be described for this failure is the inherent characteristic associated with IN noise. This unique characteristic is independence that PCA ignores but ICA does take into account.

For the ICA implementation for noise characterization of web camera, the setup has been laid out as following. A commercial web camera was connected directly to the computer via USB port. The camera provides frames of 288×352 pixels at a frame rate of 50images/s. The output values from the actual pixels in the field circuitry of the web camera are different from the values of the pixels in the analyzed set of frames. As artifacts are added by the compression algorithms and readout electronics in the analyzed pixels [1].

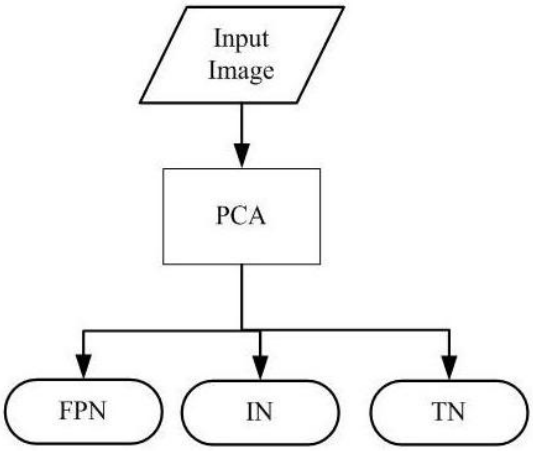


Figure 6: Block diagram of Conventional Method

The object is presented to the camera by using a CRT monitor. The web camera stares at the monitor and a movie is recorded to identify and classify the spatial-temporal artifacts embedded in the image. Each sequence contains 50 frames. The patterns spatial frequency varies slightly from the top to the bottom of the frame in order to include a narrow variation of the spatial frequency when moving along the vertical direction. Its basic aim is to preclude aliasing artifacts along the whole image. The experiment has been performed on three different frequencies to show how noise pattern changes with increasing frequencies. These three chosen frequencies are labeled as low, medium, and high frequencies are shown in Fig. 7.

The images are given as an input signal to PCA. It has already been observed that PCA is not suitable for separating independent noise processes. Therefore, PCA is used as helping tool to obtain two types of noise patterns i.e; FPN and non-FPN noise patterns. The non FPN noise patterns contain IN and TN, both are completely independent of each other. ICA is used to separate these two noise patterns.

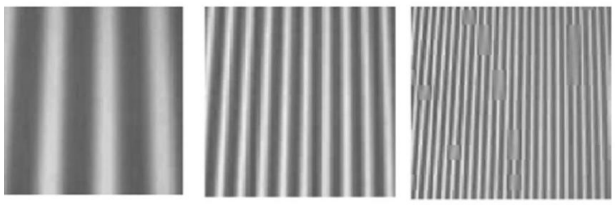


Figure 7: a) Low frequency frame, b) Medium frequency frame, c) high frequency frame.

ICA computes independent components and reconstruction is done by projecting these independent components. The process is demonstrated in Fig. 8.

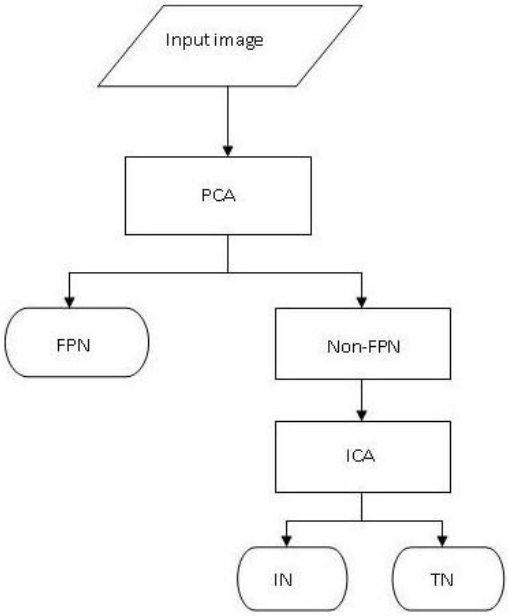


Figure 8: Block diagram of Proposed Method.

5 Results and Discussions

In the Fig. 9 it can be seen that FPN carries the maximum of variation of data. IN expresses less variation compare to TN. As PCA focuses on covariance structure thus gives good results for IN but it fails while discussing about TN noises, because TN is statistically independent noise.

As a solution to above problem, PCA has been applied first to the webcam images from which FPN and non FPN noises have been extracted. The non FPN (IN and TN) noise images are than processed further by the ICA. As they are statistically independent of each other therefore, ICA is performed on these images. The outputs of IN and TN noise images are shown in Fig. 10.

A pure TN is a diagonal covariance-wise for noises as it does not depend on the spatial frequency structure of the object. Comparing Fiq. 11 it can be observed that ICA outperforms PCA giving a pure diagonal. The diagonals using ICA from the three different frequencies are closer to the delta function than those retrieved using PCA.

The structure of the covariance of TN filtered set is strongly dependent on the spatial frequency. Therefore, it is better revealed when plotting the correlation function as an image. TN is a Gaussian noise pattern that is completely independent, thus when ICA is applied for different frequencies for extracting TN the results are more explicit than that attained by PCA as described in Fig. 12.

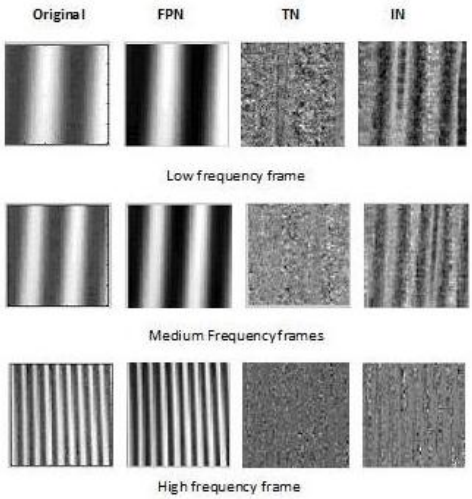


Figure 9: Results using PCA showing the spatial distribution of noises.

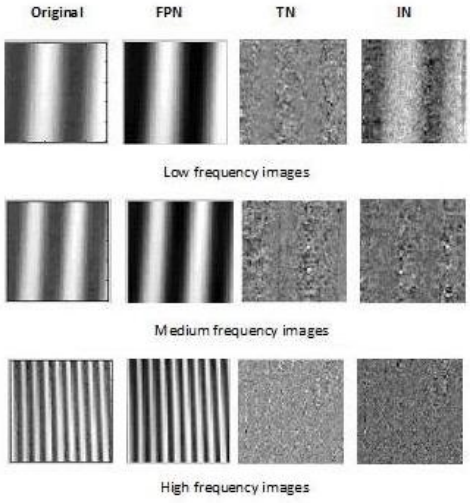


Figure 10: FPN are the results of PCA while IN and TN are extracted from non-FPN images using ICA technique.

6 Conclusion

In this paper ICA method has been applied for the characterization of artifacts in fully digital image-acquisition systems. The practical application was performed using video sequences taken

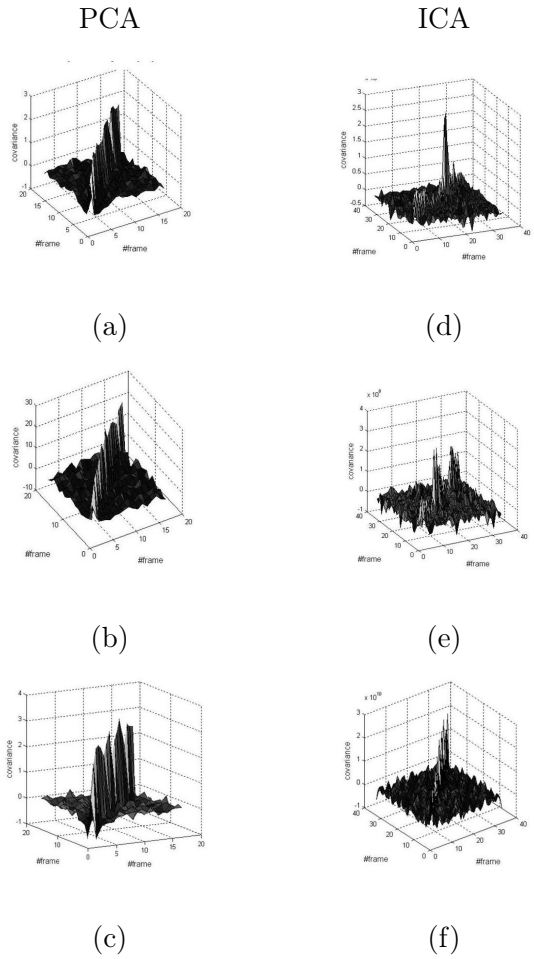


Figure 11: a) Covariance of TN using Low Frequency. b) Covariance of TN using Medium Frequency. c) Covariance of TN using High Frequency. d) Covariance of TN using Low Frequency. e) Covariance of TN using Medium Frequency. f) Covariance of TN using High Frequency

with USB cameras. PCA was applied as an integral part of ICA. By PCA method two different noise patterns have been attained, the FPN and Non-FPN. The Non-FPN was a mixture of IN and TN noise patterns which are completely independent of each other. Therefore, ICA has been applied to it. The resultant IN and TN from ICA methods are much closer in shape to the actual noise characterizations. One of the major reasons for this is the ability of ICA to extract independent components. This work can be further extended for bad pixel analysis.

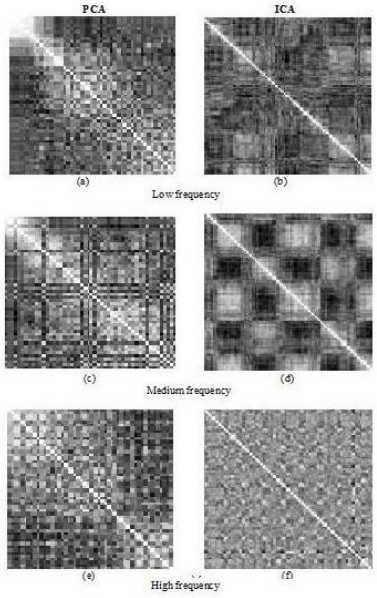


Figure 12: Correlation for IN using PCA in (a), (c), (e) and ICA in (b), (d) and (f).

Bibliography

[1] Jose Manuel, Lopez-Alonso, and Javier Alda, "Characterization of artfacts in fully digital image-acquisition systems: Application to web cameras„" in *Society of Photo-Optical Instrumentation Engineers(SPIE)*, 2004, vol. 43, pp. 257–265.

[2] Heli T. Hytti, "Characterization of digital image noise properties based on raw data„" in *Image Quality and System Performance III*, 2006, vol. 6059, 60590A.

[3] Tetsuya Takiguchi and Yasuo Ariki, "Pca-based speech enhancement for distorted speech recognition„" in *Journal of multimedia*, 2005, vol. 2, pp. 13–18.

[4] M. Stewart Bartlett, Javier R. Movellan, and Terrence J. Sejnowski, "Face recognition by independent component analysis„" in *IEEE Transactions on Neural Networks,*, 2002, pp. 1450–1464.

[5] Bell A. J., T. J. Sejnowski, and Vision Res, "The independent components of natural scenes are edge filters," in *In Neural Comput*, 1997, vol. 37, pp. 3327–3338.

[6] S. Makeig, A. J. Bell, T.-P. Jung, and T. J. Sejnowski, "What is the goal of sensory coding?„" in *In Neural Comput*, 1994, vol. 6, pp. 559–601.

[7] M. J. McKeown, S. Makeig, G. G. Brown, T.-P. Jung, S. S. Kindermann, A. J. Bell, and T. J. Sejnowski, "Analysis of fmri by decomposition into independent spatial components„" in *Human Brain Mapping*, 1998, vol. 6, pp. 160–188.

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 2 (June), pp. 312-324

An Intelligent and Pervasive Surveillance System for Home Security

A. Longheu , V. Carchiolo, M. Malgeri, G. Mangioni

Alessandro Longheu, Vincenza Carchiolo,
Michele Malgeri, Giuseppe Mangioni
Dipartimento di Ingegneria Elettrica, Elettronica ed Informatica
Facoltà di Ingegneria - Università degli Studi di Catania
Viale A. Doria 6, I 95125 - Catania - Italy
{alessandro.longheu, vincenza.carchiolo, michele.malgeri,
giuseppe.mangioni}@dieei.unict.it

Abstract: Domotics is a promising area for intelligent and pervasive applications that aims at achieving a better quality of life. Harnessing modern technologies is valuable in many contexts, in particular in home surveillance scenario, where people safety or security might be threatened.
Modern home security systems endorse monitoring as well as control functions in a remote fashion, e.g. via devices as a laptops, PDAs, or cell phones, thus implementing the pervasive computing paradigma; moreover, the intelligence is now often embedded into modern applications, e.g. surveillance systems could adapt to the environment through a self-learning algorithm.
This work presents an intelligent and pervasive surveillance system for home and corporate security based on the ZigBee protocol which detects and classifies intrusions discarding false positives, also providing remote control and cameras live streaming. Results of tests in different environments show the effectiveness of the proposed system.
Keywords: Home surveillance systems; pervasive systems; ZigBee protocol.

1 Introduction

Home automation [1], [2] exploits the latest technologies to provide an intelligent control of lighting, air conditioning, plumbing systems, home appliances and security systems, achieving comfort, safety, efficiency, costs/energy savings and, in summary, a better quality of life [3], [4].

One of the first scenarios where the penetration of domotics occurred is home surveillance, where sensors, actuators, alarms, controllers or even robots [5] increase safety through continuously environment scanning, sending alarms and recording incidents upon detecting any abnormal event, based on the consolidated principles of corporate security [6].

The amazing part of modern home security systems is that monitoring and control can be performed remotely via devices as a laptops, PDAs, or cell phones. According to this criterion, home security systems can be classified into four categories [7]:

- hardware-based, the simplest systems where both monitoring and control are implemented in hardware,
- passive systems, where only the monitoring is remote (the control is manual),
- phone based systems, with monitoring and control performed through the phone (wired and/or cellular) network
- web-based systems, identical to phone-based but using the Internet as the communication infrastructure.

Hardware-based systems in general offer high performances but with higher costs than other solutions, and they might be subjected to proprietary solutions thus with potentially drawbacks as less interoperability and less flexibility; conversely, lower performances but greatest flexibility at limited costs is achieved with phone- and web-based systems, which are now more and more adopted for home surveillance.

Moreover, the use of these systems together with the cutting-edge hardware and software technologies of mobile devices allows a more effective implementation of the intelligent [8] and pervasive (or ubiquitous) computing paradigma [9], [10].

The intelligence is now often embedded into modern applications, for instance the surveillance system could adapt to the environment through self-learning [11] or it can automatically start some countermeasure as some critical event occurs [12].

The concepts of ubiquitous and pervasive computing, sometimes considered as similar [13] or even overlapped with mobile or embedded computing [14], are implemented on one hand thanks to the advanced pc-based hardware (sensors, actuators etc.) available for home security, in accordance with the ubiquitous computing paradigm where computers "vanish into the background" [16], [15], on the other hand the use of smatphones or PDAs as monitor/controller devices gives users the possibility of accessing information and services anytime from anywhere, as promoted by the pervasive computing approach [17].

Another relevant issue concerning home security systems is the underlying transmission network, indeed a consolidated standardization is still missing [18], whilst the need for a quick deployed and cost-effective wireless solution to support easy remote control and affordable data transmission is emerging rapidly [19].

Currently, several home wireless networks are available as infrared technology (IrDA), Bluetooth and ZigBee protocol. IrDA operates over short distances and is subjected to high error rate, whereas the Bluetooth technology is limited by network capacity and performances.

The most promising standard for wireless home and personal area networks is then represented by the ZigBee technology [20], [21], which comes with features as low complexity, low data error rate, low power and low-cost [22], [23].

The work presented in this paper falls into the scenario outlined above, in particular we propose an intelligent surveillance system for home security that receives intrusion attempts detected by sensors and cameras connected through a ZigBee-based network and classifies such intrusions with a customizable algorithm in order to exclude false positive cases (e.g. leaves moved by the wind); this paper is an extended version of [24].

Potential alarms can be managed via an iPhone® application that receives alarms, and allows users to use the iPhone as a system remote controller and as a monitoring console to view real-time camera images.

The paper is organized as follows: section 2 introduces the architecture of the system, while in section 3 we describe in detail the application that manages devices, processes alarms detection and implements the push notification service. Section 4 shows the calibration that helps false positive intrusion detection as well as the system at work; finally, section 5 presents our conclusions and further works.

2 The Physical Layer

The home surveillance system we propose is represented in Figure 1; its components come from several goals to be addressed:

- First, we want to detect of unauthorized accesses to the perimeter, i.e. whenever a potential intrusion occurs it should be detected using infrared radars and ip cameras)

- the system must be able to classify the events detected, distinguishing between real intrusions and false positives e.g. due to leaves moved by the wind or to animals; this goal is achieved through a customizable algorithm [25] that can be tailored to both indoor and outdoor environments
- Finally, once real intrusions are recognized, the system provides a notification via an iPhone® application so the user is immediately warned; he should also be able to view real-time images from cameras, so real-time countermeasures as siren activation or police action request can be taken.

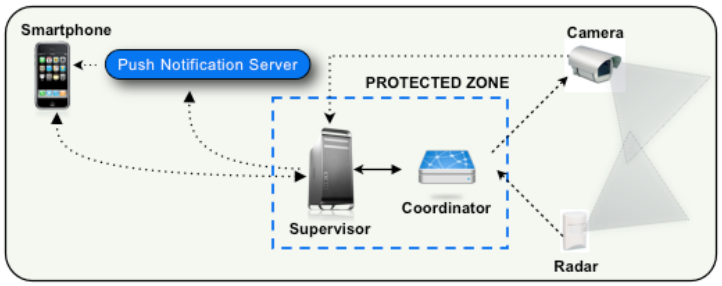


Figure 1: Logical schema of the proposed suirveillance system

Referring to Figure 1, the system key component is the supervisor which manages the network, the data coming from devices (cameras and radars) as well as remote iPhone connections and finally processes video images for intrusions detection.

The supervisor used in our experiments was a medium-performance notebook, including a Pentium® dual-core CPU @ 2.30GHz and 2GB RAM, however the system does not require significant CPU/bandwidth resources, so it can be easily implemented on a low-cost 32-bit microcontroller provided with a serial interface (to allow communications with the coordinator module).

The coordinator acts as the interface between the supervisor and the ZigBee network; it is implemented using a custom ZigBee module, it manages the low-level network and can be configured and managed via standard serial interface.

Moreover, we used only devices that provide standard TTL output, in order to easily achieve interoperability with any programmable digital circuit. To manage the communication between the devices and the ZigBee module we used a microcontroller that listens to events from the device and controls the ZigBee module.

For instance, the PIR (passive infrared) based component, i.e. a typical residential PIR motion detector with multi-Fresnel lens cover, needs to be interfaced with the ZigBee network; to this purpose, we used the PIC16F628A microcontroller (see Fig.2), identical to the well known PIC16F84 except for its additional USART module, which allows a high level management of serial communication. In addition to standard pins connection, as clock and reset, also the RB0/INT (pin 6) and RB2/TX/CK (pin 8) were connected, in particular the former was used to receive asynchronous interrupt signals coming from the PIR radar, whilst the latter was set up as an output for the PIC16F628A and used to connect to the ZigBee network.

The microcontroller run at a 16MHz frequency clock in order to ensure the compatibility with the ZigBee module The microcontroller waits for signals coming from sensors TTL output; as soon as it receives a signal - i.e. a potential alarm - it alerts the coordinator (via the ZigBee module) that controls the camera using the RS232 protocol.

Note that each module acts as a ZigBee End Device, so they spend most of the time in a sleep state, thus saving energy. This however also requires to ”wake up” the device in order to

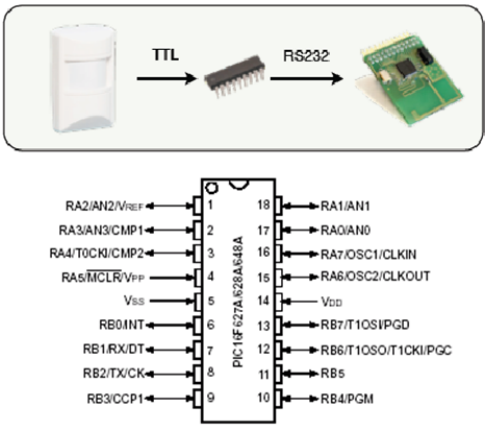


Figure 2: Interfacing a PIR motion detector with the ZigBee network using the PIC16F628A

communicate with it, so the PIC16F628A was programmed so that the device changes from the sleep to the ready state before sending a command, and the same controller allows the device going into the sleep state again after the command has been executed. Finally, note that for the PIR motion detector we also detect potential tampering attempts by connecting the device TAMP pin to the PIC16F628A so that an alarm is sent even in this case.

3 The Application Layer

The architecture described in the previous section operates thanks to the supervisor application, a software that manages devices and evaluates alarms, helping to prune false positive ones. Such application is arranged into the following modules:

- Control module: it controls the physical connection with devices and is based on the QextSerialPort project [27]; this module automatically detects new devices (for instance, a new radar) and allows the supervisor user to add it to the ZigBee network
- ZigBee management module: it is an higher layer manager of the ZigBee devices as cameras and radars;
- Alarm management module: it processes potential alarms to remove false positives and manages the remote alarm notifications as well as control.

All these modules are implemented in C++ within the Qt Framework [26], whereas the OpenCV libraries [28] were used to process images coming from IP camera.

In particular, the available libraries to communicate with the coordinator were developed as an OCX control, so just Windows®-based systems can interface with the system. To overcome this limitation, we develop a C++ version of such libraries; this not only provides more compatibility, but also can be easily interfaced with computer vision (CV) softwares, which are generally written in C language.

The reason for choosing the Qt Framework is that it ensures platform portability and operating system independence; its only drawback is the Phonon framework used by Qt to manage the multimedia layer. Phonon is an abstract framework whose implementation depends on the operating systems, in particular it leverages the Quicktime® libraries for Mac OS-X, DirectShow® libraries within the Windows® platform and GStreamer for Linux OS. Phonon does not allow

an easy single frame management of a video stream, so we exploited the QTKit and QuickTime frameworks to write a class that provided these functionalities.

A complete UML diagram of the supervisor application is depicted in Fig.3, in particular:

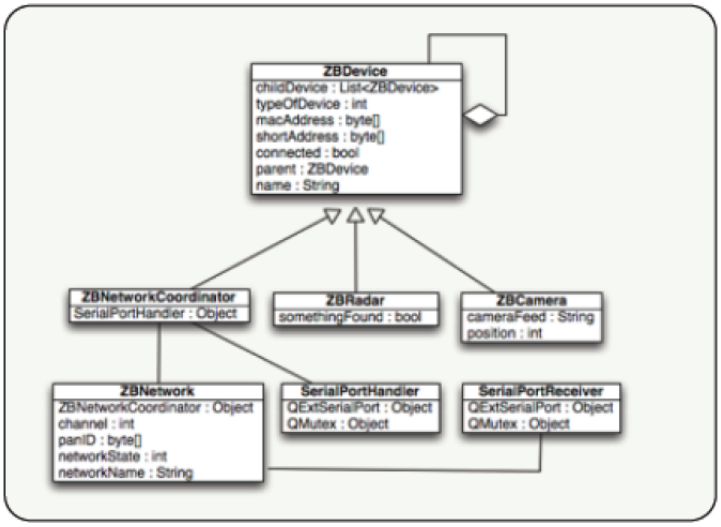


Figure 3: UML diagram of the proposed system

- the ZBDevice class represents a generic ZigBee device with its standards parameters, as the MAC address, the type of device and so on
- three classes are derived from the ZBDevice, i.e. the ZBNetworkCoordinator, the ZBRadar and the ZBCamera, whose roles are clearly understandable; in particular, the coordinator is the only device that can access the physical network via the SerialPortHandler class
- the ZBNetwork class describes the underlying ZigBee network, whereas asynchronous signals in the network are managed by the SerialPortReceiver, that forwards such signals to the right device, for instance as soon as the coordinator is turned on, it sends a message to its supervisor with its mac address in order to be registered following a process mediated by the SerialPortReceiver which also exploits the SerialPortHandler to provide message acknowledgements

A snapshot of the supervisor application is presented in fig.4.

In the following we discuss how the alarms are detected, whereas the subsection 3.2 shows how the remote notification occurs whenever a relevant alarm is detected.

3.1 Alarms detection

The MP4 video stream generated by the IP camera is provided to the supervisor's alarm processing module through a Real Time Streaming Protocol (RTSP) server connected via the wireless connection; this allows the evaluation of whether an intrusion occurred and an alarm should be generated or not.

The received frames are evaluated through the algorithm (details can be found in [25]) in order to assign a precision (i.e. a numeric value) thus establishing whether a false positive has been detected.

To do this, the first (trivial) solution we adopted was to evaluate the difference between the current frame and the previous one, applying a threshold to detect a binary pattern; this worked

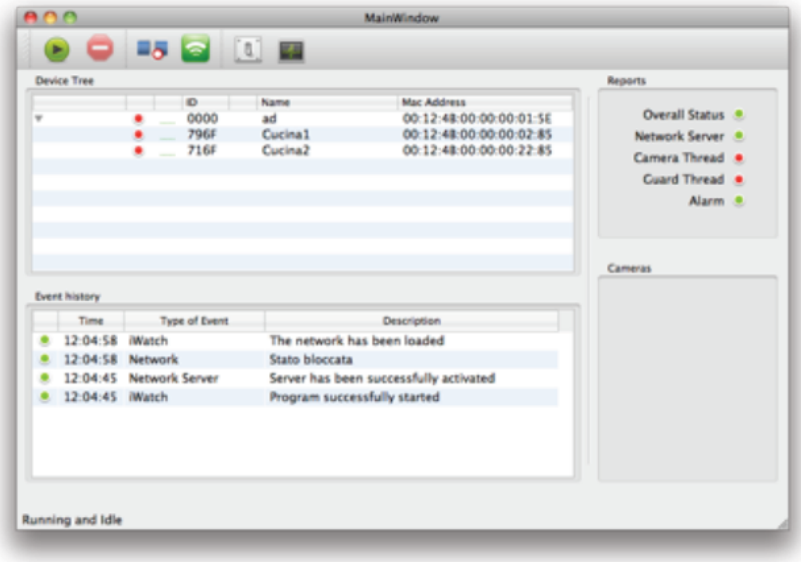


Figure 4: Snapshot of the supervisor application

in almost static background scenarios, as for indoors environments (e.g. a room) with constant lighting, but it is not suitable when frames background is not static, for instance when tree leaves are moved by the wind.

The next step was to apply the Gaussian Background Model [29], citeLee2:2005, in order to effectively remove the (even dynamic) background; the algorithm implemented in the supervisor application works following the steps illustrated in fig. 5: it first converts colour images into a black and white format for binary processing (1), then it detects the foreground (2) and extracts the corresponding bounding rectangle (3), i.e. the area where the gaussian model detected relevant information.

Figure 5: Alarms sources extraction process of the supervisor algorithm

Bounding rectangles across subsequent frames are compared for a given time interval and if something is detected with a specified precision, an alarm is generated; the system stores in XML format any relevant event, should it determine an alarm or not, in accordance with a set of severity levels.

Note that alarm detection is completely customizable by setting algorithm parameters, so a proper configuration allows to correctly distinguish between false positives/negatives and real alarms; for instance, the threshold can be reduced for indoor environments (where the background is almost static) so even a little movement will be detected; similarly, the minimum bounding rectangle can be increased, in order to discard cats/dogs movement detection in outdoor environments and so on.

Parameters should be tailored to the actual scenario the system will be installed into (section 4 shows a typical calibration session); in fig. 6 the C struct implementing such parameters is shown, together with the screenshot of the supervisor application used to set up their values. Details about the meaning of parameters and how they affect the CV recognition process can be found in the OpenCV documentation [28].

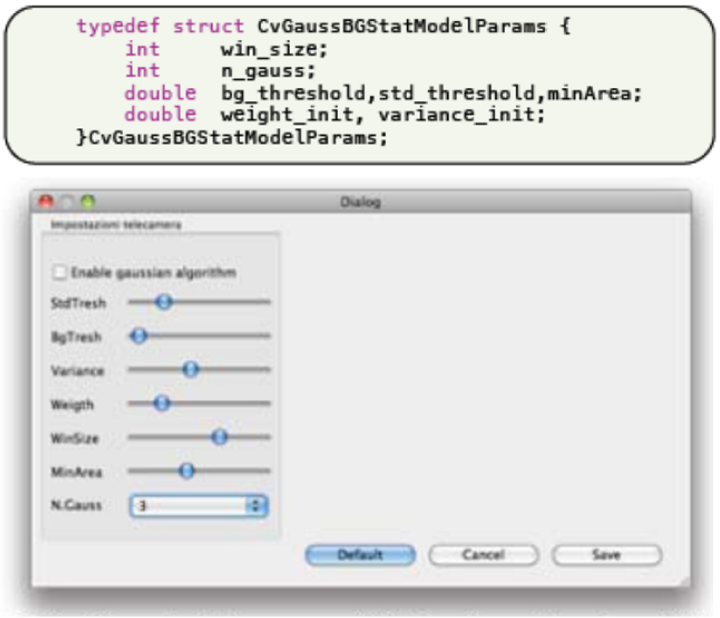


Figure 6: CV Parameters set up

3.2 Alarm notifications and remote control

As soon as a (possibly real) alarm is detected, the system sends notification to the user, also providing a remote control of overall systems functionalities. These are implemented making use of the well-known push technology [31], citePohja:2009, in particular we used the Apple Push Notification Service (APNS in the following) [33].

The supervisor application is first registered at the server providing APNS; then, whenever the supervisor needs to send a notification, it establishes a secure connection to the APNS using the certificate obtained by the APNS during the registration, and finally data are transmitted. The format of the packet being sent is illustrated in figure7, where the deviceToken is a 32 byte device identifier and the Payload represents the notification, created according to the JSON format, a lightweight data-interchange format based on a subset of the JavaScript language [34].

During our tests notifications were received within about 4 seconds since the request, that can be considered a good response time for this kind of application.

To manage notifications and provide remote control functionalities, a mobile application operates in conjunction with the supervisor's software counterpart.

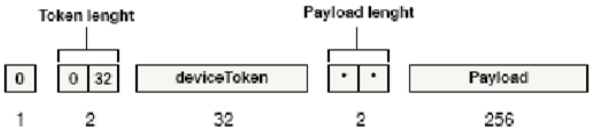


Figure 7: The format of notifications sent by the supervisior

This application was developed in Objective-C within the Cocoa Touch framework [35], and it displays notifications received from the push server, also allowing the management of the ZigBee network (e.g. enabling or disabling devices) and the possibility of remotely examining current camera video stream from the mobile phone, checking whether it deals with a real intrusion on not and applying proper countermeasures, as an alarm activation or a police call for on-site actions.

Finally, note that the system also performs a complete log of all events, each stored with a severity level, a timestamp and a description, so that it will be possible to analyze who, why and when a given alarm notification has been triggered.

4 Experiments and Results

In order to evaluate the performance of the intrusion detection system presented in previous sections, we performed two types of test, the former to assess how the calibration affects alarms classification and the latter is about the behaviour of the ZigBee network, in particular for what concerns the largest area that can be effectively kept under surveillance.

To assess false positive alarms pruning, we used a Panasonic^{TM} standard VGA resolution camera in two different scenarios, an indoor environment (a room) with scarce illumination and the an outdoor environment (a garden in front of a house) with strong lighting.

For each scenario we performed 50 test sessions, 25 with an intruder walking in front of the camera and the remaining 25 without intrusion. Finally, results from each scenario have been evaluated first without any configuration for the detection algorithm parameters (i.e. with default values) and then providing calibration for such values (details are here omitted [25]) In the following we present the test results using diagrams that show the number of tests according to the percentage of accuracy reached (ranging from 0 to 100%).

The measures for the indoor scenario with no calibration (see figure 8 a) show unsatisfactory results. In the set of 25 tests with an intruder, 20% of them shown negative results, i.e. the system did not detect the intruder. The same scenario with no intruder shows incorrect results for 44% of tests (a non-existent intruder is detected); this was due to missing calibration and also by changes in the light coming from the windows (which lead to false positive detection).

After the calibration, results significantly improved as figure 8 b) shows. The algorithm indeed detected the intruder with high accuracy (no false negative occurred), and similarly in the case of absence of the intruder (no false positive have been detected).

The second scenario concerns an outdoor environment with strong lighting (a garden in front of a house in the morning); results are displayed in figure 9. The significantly increasing in illumination allow better results even with no calibration (see fig. 9 a); the algorithm indeed shown high precision in detecting an intruder (no false negatives), whilst we have just 16% of false positives in the case of no intrusion. After the calibration results were further improved (false positives were completely prevented).

The second set of tests was about the ZigBee network, in particular we want to assess the largest area that can be effectively kept under survelliance. To quantify this effectiveness, we

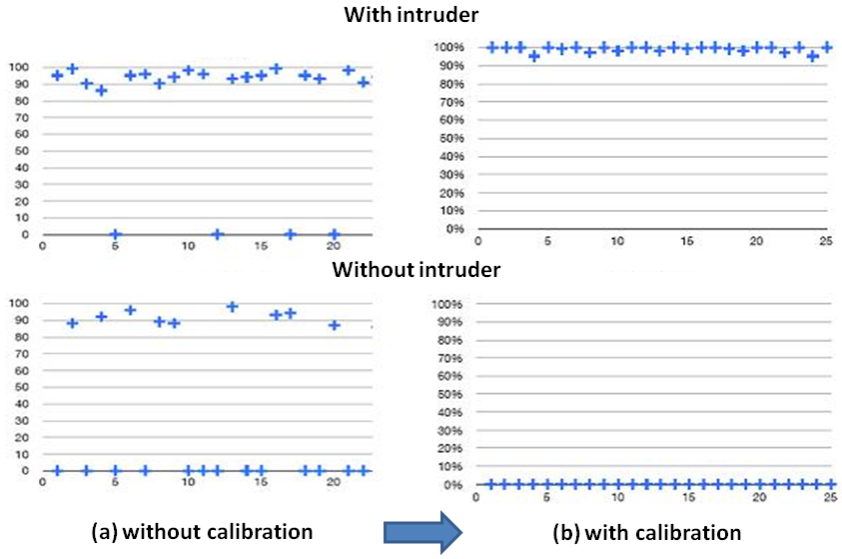


Figure 8: Indoor scenario

used the Link Quality Indicator (LQI), an 8 bit sequence representing a figure of the quality of the link between two ZigBee nodes [36].

To perfom such tests we used a system made up of a coordinator module and a router module which was positioned at different locations (increasing the distance from the coordination) along a given perimeter. At each position we performed 25 measures; results are shown in fig. 10, where the value of LQI for distance from 0 to 12m is plotted, in particular the solid line represents the theoretical LQI value whereas the other line represents measured LQI values. Each point is the average of 25 measures; in the same figure we report some LQI measurements together with corresponding variance values; note that variance remains limited until a distance of 4 meters, whereas it increases significantly over this distance, this is due to the presence of a wall at 5 mt, which hinders the communication between ZigBee modules.

Tests revealed worse performances than those declared on the component datasheets, indeed the system worked well up to a distance of 20 meters even if the LQI decreased significantly since a distance of 5 meters. However, we noted that even with a low LQI the communications between ZigBee modules was acceptable, in particular some connection losses were detected but the system was able to restore the connection in few seconds.

5 Conclusions and Future Works

Several intelligent and pervasive applications are being developed within the domotics context. in particular, an intelligent surveillance system for home security was presented in this paper. It exploits the ZigBee protocol and it can detect and classify intrusions to discard false positive and negative alarms, also providing remote control functions and cameras live streaming to allow users to analyze who and why is triggering alarms. We also presented results about both the detection algorithm effectiveness and the ZigBee network performances.

Some future works include the following issues:

- the application currently was tested on Windows or MacOS X operating systems; to increase portability, we are planning to test it in Linux-based platforms;

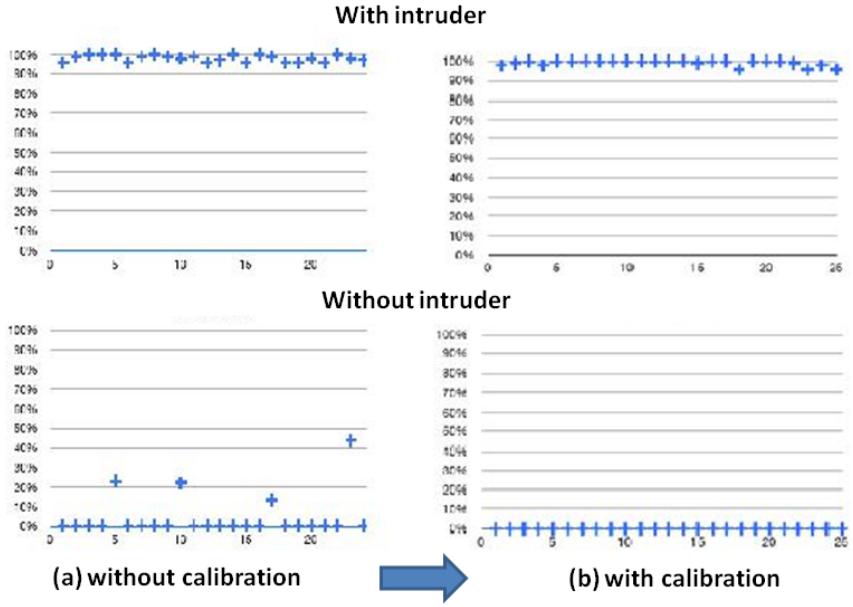


Figure 9: Outdoor scenario

Distance	Average LQI	Absolute variance	Variance (%)
0,8 m	140	0,0	0,000%
1,2 m	132	8,1	6,121%
1,5 m	140	0,9	0,644%
2 m	133	2,4	1,805%
2,5 m	123	0,0	0,000%
3 m	123	0,0	0,000%
4 m	120	0,0	0,000%
6 m	79	17,6	22,373%
9 m	75	12,2	16,267%
12 m	64	22,4	34,879%

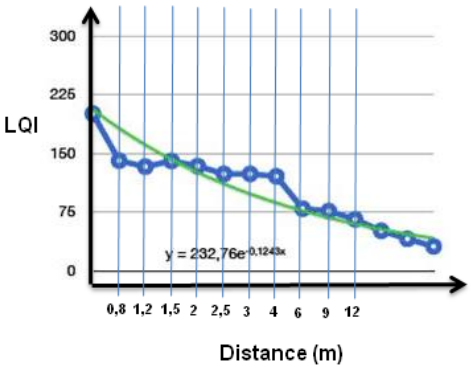


Figure 10: LQI versus distance between coordinator and router ZigBee modules

- the system includes just a coordinator and a router with a camera; adding other components, e.g. smoke and gas sensors, household appliances etc., allows to build a system that guarantees security and safety in a global fashion as electronic keys or biometric sensors to improve the system's capability.

- the algorithm we used to evaluate intrusions was quite simple, and it should be intended as the best choice for first experiments (it indeed did not affect on the overall system performance); better yet slower algorithm should be tested.

6 Acknowledgement

This paper was inspired by a work developed in cooperation with ValueTeam IT consulting and solutions (http://www.valueteam.com). Particularly, we thank Francesco Consoli as Senior Manager at Security Division, for his guidance and support to this work. We also thank Danilo Torrisi for the implementation and test of the system.

Bibliography

[1] Jacobson, J., Understanding Home Automation, *Electronic House*, 14(6):18-21, 2001

[2] Rodden, Tom and Benford, Steve, The evolution of buildings and implications for the design of ubiquitous domestic environments, *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, New York, USA, pp. 9-16,2003

[3] Lorente, S., Key issues regarding Domotic applications, *International Conference on Information and Communication Technologies: From Theory to Applications*, pp. 121 - 122, 2004

[4] Park, Sang Hyun and Won, So Hee and Lee, Jong Bong and Kim, Sung Woo, Smart home-digitally engineered domestic life, *Personal and Ubiquitous Computing*, Springer London, 7(3):189-196, 2003

[5] Luo, R.C. and Hsu, T.Y. and Lin, T.Y. and Su, K.L., The development of intelligent home security robot, *Mechatronics, 2005. ICM '05. IEEE International Conference on*, pp. 422 -427, 2005

[6] European Institute for Corporate Security, *http://www.eicsm.org/index.html*

[7] Chunduru, V. and Subramanian, N., Perimeter-Based High Performance Home Security System, *Consumer Electronics, 2007. ISCE 2007. IEEE International Symposium on*, pp. 1 -7, 2007

[8] Russell, Stuart J. and Norvig, Peter, Artificial intelligence: a modern approach, Prentice-Hall, Inc., 2010

[9] Nieuwdorp, Eva, The pervasive discourse: an analysis, *Comput. Entertain.*, ACM, New York, NY, USA, vol. 5(2):13-13, 2007

[10] Bell, Genevieve and Dourish, Paul, Yesterday's tomorrows: notes on ubiquitous computing's dominant vision, *Personal Ubiquitous Comput.*, Springer-Verlag, London, UK, 11(2):133-143, 2007

[11] Jun Hou and Chengdong Wu and Zhongjia Yuan and Jiyuan Tan and Qiaoqiao Wang and Yun Zhou, Research of Intelligent Home Security Surveillance System Based on ZigBee, *Intelligent Information Technology Application Workshops, International Symposium on*, IEEE Computer Society, Los Alamitos, CA, USA, pp. 554-557, 2008

[12] Krausz, Barbara and Herpers, Rainer, Event detection for video surveillance using an expert system, *AREA '08: Proceeding of the 1st ACM workshop on Analysis and retrieval of events/actions and workflows in video streams*, Vancouver, British Columbia, Canada, ACM, New York, NY, USA, pp. 49-56,2008

[13] Want, Roy and Pering, Trevor, System challenges for ubiquitous & pervasive computing, *ICSE '05: Proceedings of the 27th international conference on Software engineering*, St. Louis, MO, USA, ACM, New York, NY, USA, pp. 9-14,2005

[14] McCullough, Malcolm, Digital Ground: Architecture, Pervasive Computing, and Environmental Knowing, MIT Press, Cambridge, MA, USA, 2004

[15] Weiser, M. and Gold, R. and Brown, J. S., The origins of ubiquitous computing research at PARC in the late 1980s, *IBM Syst. Journal*,IBM Corp., Riverton, NJ, USA, 38(4):693-696, 1999

[16] Weiser, Mark, The computer for the 21st century, *SIGMOBILE Mob. Comput. Commun. Rev.*, ACM, New York, NY, USA, 3(3):3-11, 1999

[17] Hansmann, Uwe and Nicklous, Martin S. and Stober, Thomas, Pervasive computing handbook, Springer-Verlag New York, Inc., New York, NY, USA, 2011

[18] Miori, Vittorio and Russo, Dario and Aliberti, Massimo, Domotic technologies incompatibility becomes user transparent, *Commun. ACM*, New York, NY, USA, 53(1):153-157, 2010

[19] Egan, D., The emergence of ZigBee in building automation and industrial control, *Computing & Control Engineering Journal*, 16(2):14-19, 2005

[20] Fukui Kiyoshi and Tanimoto Akira and Fukunaga Shigeru, ZigBee Technology for Low-Cost and Low-Power Radio Communication Systems, *Journal of the Institute of Electronics, Information and Communication Engineers*, vol. 88(1):40-45, 2005

[21] Wang Dong and Zhang Jin-rong and Wei Yan, Building Wireless Sensor Networks (WSNs) by Zigbee Technology, *Journal of Chongqing University (Natural Science Edition)*, 29(8):95-98, 2006

[22] Li Cai and Nina Dai, The Home Security System Based on ZigBee Technology, *Wireless Communications Networking and Mobile Computing (WiCOM), 2010 6th International Conference on*, pp. 1-2, 2010

[23] The ZigBee Protocol OB/EL, http://www.digi.com/technology/rf-articles/wireless-zigbee.jsp, *http://www.digi.com/technology/rf-articles/wireless-zigbee.jsp*, 2009

[24] Carchiolo V. et al, Pervasive home security: an intelligent domotics application, *Intelligent Distributed Computing IV (IDC 2010) conference*, pp. 145-154, 2010

[25] Danilo Torrisi, Sistema di sicurezza perimetrale con allarmistica e controllo a distanza, 2009, *Tech. Report - Dipartimento di Ingegneria Informatica e delle Telecomunicazioni - Facolta' di Ingegneria - Universita' di Catania*

[26] Qt Nokia Framework, *http://qt.nokia.com/about/news/nokia-releases-qt-4.6.2*

[27] QextSerialPort, *http://qextserialport.sourceforge.net/*

[28] OpenCV libraries, *http://sourceforge.net/projects/opencvlibrary/*

[29] Brian R. Williams and Ming Zhang, Multiple Dimension Chrominance Model for Background Subtraction, *Computational Intelligence*, pp. 438-443, 2005

[30] Lee, Dar-Shyang, Effective Gaussian Mixture Learning for Video Background Subtraction, *IEEE Trans. Pattern Anal. Mach. Intell.*, IEEE Computer Society, Washington, DC, USA, 27(5):827-832, 2005

[31] Eugster, Patrick Th. and Felber, Pascal A. and Guerraoui, Rachid and Kermarrec, Anne-Marie, The many faces of publish/subscribe, *ACM Computing Survey*, ACM, New York, NY, USA, 35(2):114-131, 2003

[32] Pohja, Mikko, Server push with instant messaging, *SAC '09: Proceedings of the 2009 ACM symposium on Applied Computing*, Honolulu, Hawaii, ACM, New York, NY, USA, pp. 653-658, 2009

[33] Apple Push Notification Service, *http://developer.apple.com/ IPhone/ library/ documentation/ NetworkingInternet/ Conceptual/ RemoteNotificationsPG/ ApplePushService/ ApplePushService.html*

[34] Introducing JSON, 2011, *http://www.json.org*

[35] Cocoa Touch Framework, *http://developer.apple.com/iphone*

[36] Philip Orlik and Jinyun Zhang and Bharat Bhargava and Gang Ding and Gang Ding and Zafer Sahinoglu and Zafer Sahinoglu, Reliable Broadcast in ZigBee Networks, *In Proceedings of SECON IEEE Conference*, 2005

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 2 (June), pp. 325-340

Small Signal Monitoring of Power System using Subspace System Identification

A. Mohammadi, H. Khaloozadeh, R. Amjadifard

Aliakbar Mohammadi
Department of Electrical Engineering
Science and Research Branch
Islamic Azad University
Tehran, Iran.
aa.mohammadi@srbiau.ac.ir

Hamid Khaloozadeh
K. N. Toosi University of Technology
Tehran, Iran.
h_khaloozadeh@kntu.ac.ir

Roya Amjadifard
Tarbiat Moallem University
Tehran, Iran.
amjadifard@tmu.ac.ir

Abstract: In this paper, small signal analysis of power systems is investigated using Subspace System Identification (SSI) methods. Classical small signal analysis methods for power systems are based on mathematical modeling and linearized model of power system in an especial operating point. There are some difficulties when such a classical method is applied, specially, in the case of large power systems. In this paper, such difficulties and their bases are investigated and in order to avoid them, it is suggested to use SSI algorithms for small signal analysis of power systems. The paper discusses extracting of small signal properties of power systems and presents some new suggestions for application of subspace system identification methods. Different types of subspace system identification algorithms were applied to different power system case studies using the presented propositions. The benefits and drawbacks of subspace system identification methods and the presented suggestions are studied for small signal analysis of power systems and power system monitoring. Several comparisons were investigated using computer simulations. The results express the usefulness and easiness of proposed methods.
Keywords: small signal, subspace identification, power system, monitoring.

1 Introduction

It is not more than 20 years that a new horizon has been opened in system identification. Subspace System Identification (SSI) has been one of the most attractive methods for system identification since 1990s. There is a large amount of literatures devoted to the algorithms based on subspace system identification [1–3]. Factually, we can present a list of renowned type of SSI algorithms as following; Numerical Subspace State Space System IDentification (N4SID) [4], Multivariable Output Error State Space (MOESP) [5, 6], Past Output (PO-MOESP) [7, 8], Canonical Correlation Analysis (CCA) [9], Orthogonal Decomposition (ORT) [2]. Although they are similar in some general aspects, but there are several differences which may distinguish them. Actually, they don't perform quite the same as each other since they practice various

mathematical tools in different ways. It is not too far to expect fairly different advantages or disadvantages when using SSI algorithms. A useful review of subspace system identification algorithms is available in [10]. Authors in [1] present a unifying theory which may be helpful to understand SSI methods.

During the past years, application of SSI methods is being developed increasingly to different areas of industry and engineering sciences. Maybe the main reason for such a development hides behind the capability and easiness of application of SSI methods for multi-input/multi-output (MIMO) systems. In addition, state space structure of SSI methods is usually considered as an advantage. SSI methods also use robust, fast and consistent mathematical tools for calculations, which provide them with some considerable advantages [2, 3].

Numerous investigators have worked on SSI methods and they have used different SSI algorithms for different applications. Moreover, MIMO specifications and also state space based analysis motivated many investigators to apply SSI methods for different analysis of power systems. However, there are not so many SSI applications in power systems, yet. The first steps of SSI applications in power systems may be seen in [11]. The paper provides low order model of large scale power system using N4SID algorithm. Results of paper express that SSI based model is in lower orders, more optimized and more suitable for controller design in comparison with classical system identification and modeling.

An application of power transformer identification was developed in [12] using frequency response data and SSI methods. Authors in [13] and [14] introduce an algorithm based on SSI and some applications of such algorithm for transformer debugging and parameter estimation.

A Heffron-Phillips model of synchronous generator was identified in [15] using subspace identification algorithms and online measurements. In [16] the parameters of a Heffron-Phillips model of synchronous generator were extracted from closed loop data using SSI algorithms. It divides identification problem of a closed loop system to two open loop identification and then it uses SSI algorithms to identify each open loop transfer function. Using some mathematical processing of the provided transfer function, it provides a transfer function as a generator model.

In [17], authors discuss a model predictive controller design for multi-machine power system using SSI algorithms. The design uses a recursive subspace system identification algorithm in order to provide a MIMO self-tuning adaptive controller; therefore it can be used for online applications. [18] mentions use of different types of power system signals which are applicable to SSI algorithms. It uses such signals to provide identification data. Modal analysis of power system was developed using subspace system identification methods and sampled data. [19] also discusses modal analysis and oscillatory stability study of power systems based on SSI methods. It provides a voltage stability measure using identified critical modes of power system. [20] introduces a power system stabilizer (PSS) using N4SID algorithm of SSI methods. It provides a power system model using SSI and then it designs a MIMO power system stabilizer using identified model. In [21], authors also discuss a PSS using stochastic subspace system identification approaches. They also mention small signal analysis of power systems.

This paper aims at small signal analysis of power systems using SSI. It proposes some useful notes on how to extract small signal properties of a power system using SSI algorithms. The paper clarifies the process of small signal monitoring of power system using subspace system identification algorithms. It also compares several SSI based analysis of power systems with the classical methods to distinguish the advantages.

The paper structure is as following; In order to highlight drawbacks of classical methods, there would be an introduction to classical small signal analysis of power systems in the next section. Subspace system identification algorithms are discussed in section three to high-light the advantages of such methods. Application of SSI methods for power system small signal analysis is introduced in section four. There are some new ideas in section four, which are used to extract

small signal properties of test case power systems in section five. On-line monitoring of power system is discussed in section six. In section seven, the application of different SSI algorithms for several power systems is discussed. Finally, the conclusions are presented.

2 Drawbacks of Analytically Small Signal Analysis of Power Systems

We are usually interested in extraction of modes, participation factors of modes, damping ratios and oscillatory frequencies of power system which all are called small signal properties of power systems. Since a power system is naturally a nonlinear system, one should follow the following stages to achieve small signal properties of a power system:

1. Finding the details of all included elements (Generator constants, Transformer and line parameters,)
2. Finding nonlinear model of power system using constant, parameters and theoretical relations of variables for different power system elements.
3. Solving a load flow problem in order to provide an operating point.
4. Linearizing of nonlinear model using the provided operating point.
5. Application of modern small signal methods to provide small signal properties.

Providing an operating point, a nonlinear modeling and linearizing the model are all tough works in application, especially when the system is large. We also know that the parameters of system may change during the normal operation of system. Therefore, it is obvious that some of the above stages are not applicable to a real system or the result of such an analysis is not reliable. Moreover, There is always a big gap between the analysis done on a piece of paper and the system behavior. Such a method is not applicable for monitoring of power system, neither. This is a considerable drawback for a scientific method. It is always a question at the end of theoretical analysis of power systems; To what extent are the results useful and applicable? it is not easy to response to the question unless we make a bridge between real world phenomena and theoretical analysis.

What can be suggested at this point is usually application of a system identification method, at least, for estimating a linear model. Classical identification methods are useful in many applications. While using a classical system identification method, the biggest difficulty origins from single-input/single-output (SISO) structure of such methods. Classical system identification methods may fall into whirlpool of over parameterization. Coping with such problems is itself a new problem.

Our suggestion for overcoming such problems is to use subspace system identification (SSI) methods. SSI methods are good solution for multi-input/multi-output (MIMO) systems. They can be considered as the bridge for passing over the gap between real world system and theoretical analysis. The next section investigates SSI methods to glorify their useful advantages for small signal analysis of power systems.

3 Subspace System Identification

Generally, we can arrange SSI methods into two categories from the measurement view; stochastic and deterministic SSI algorithms. If the SSI algorithm uses exogenous input measurements in its raw identification data set, it is called as deterministic SSI algorithm (DSSI).

Otherwise, it is called a Stochastic Subspace System Identification (SSSI) method [2, 22]. If it is supposed to provide SSSI methods in an algorithmic way for further understanding, we can arrange the following notations. Further investigations of the following notes are also useful for understanding of DSSI algorithms;

Model: The considered model of system for a typical SSSI algorithm is

$$\left\{ \begin{array}{l} x(t+1) = Ax(t) + w(t) \\ y(t) = Cx(t) + v(t) \end{array} \right. , E\left\{ \left[\begin{array}{c} w(t) \\ v(t) \end{array} \right] \left[\begin{array}{cc} w^T(s) & v^T(s) \end{array} \right] \right\} = \left[\begin{array}{cc} Q & S \\ S^T & R \end{array} \right] \delta_{ts} \tag{1}$$

where $y \in \mathbb{R}^{n_y}, x \in \mathbb{R}^{\mathrm{n}}$are samples of output and state vectors. $v_{\mathrm{t}} \in \mathbb{R}^{n_y}, w_{\mathrm{t}} \in \mathbb{R}^{\mathrm{n}}$ are stationary, zero average state and output noise vectors, consequently.

Identification Data: SSSI methods use only samples of system outputs. However, identification data should usually be provided in the following structure:

$$\begin{array}{l} f(t) \triangleq \left[\begin{array}{cccc} y(t) & y(t+1) & \cdots & y(t+k-1) \end{array} \right]^T \\ p(t) \triangleq \left[\begin{array}{cccc} y(t-1) & y(t-2) & \cdots & y(t-k) \end{array} \right]^T \end{array} \tag{2}$$

where $f(t)$ and $p(t)$ are future and past data set. k should be strictly bigger than n. It can be a guess. Therefore, this is not a restrictive condition. Now, we can formulate SSSI problem as below [1, 2, 5, 10, 22]:

"There are N samples of output vectors, $Y(t) \triangleq \left[\begin{array}{cccc} y(0) & y(1) & \cdots & y(N-1) \end{array} \right]$, from a system of order n. Find matrices *A, C, Q, R, S* and n for the structure defined in (1)."

Block Hankel Matrices: SSSI algorithms begin data processing by forming the following block Hankel Matrices:

$$H_{k,k} = E\{f(t)p^T(t)\} = \left[\begin{array}{cccc} \Lambda(1) & \Lambda(2) & \cdots & \Lambda(k) \\ \Lambda(2) & \Lambda(3) & \cdots & \Lambda(k+1) \\ \vdots & \vdots & \ddots & \vdots \\ \Lambda(k) & \Lambda(k+1) & \cdots & \Lambda(2k-1) \end{array} \right] \tag{3}$$

where $\Lambda(l) = E\{y(t+l)y^T(l)\}, \quad l = 0, 1, \cdots, L \ , \ \ 2k-1 \leq L \ , \ \ k > n$ is the correlation of future and past data. Actually, SSSI uses statistical properties of output samples for further processing. The word Stochastic in the expression (Stochastic Subspace System Identification) may arise from this point.

System Order: SSSI uses the following Singular Value Decomposition (SVD) of Hankel matrix in order to provide system order (n):

$$H_{k,k} = \left[\begin{array}{cc} U_{sys} & U_{noise} \end{array} \right] \left[\begin{array}{cc} \sum_{sys} & 0 \\ 0 & \sum_{noise} \end{array} \right] \left[\begin{array}{c} V_{sys}^T \\ V_{noise}^T \end{array} \right] \simeq U_{sys}\Sigma_{sys}V_{sys}^T \tag{4}$$

We can detect noise singular values by detecting a big gap among the singular values of Hankel matrix. Thus, noise singular values can be neglected since they are very smaller than system singular values. Therefore, system order, n can be defined as

$$n \triangleq \dim(\Sigma_{sys}) \tag{5}$$

Estimation of State Space Matrices: *SVD* of Hankel matrix can also provide us with an extended observability matrix;

$$O_k \triangleq U_{sys}\Sigma_{sys}^{1/2} \tag{6}$$

It can also present controllability matrix:

$$C_k \triangleq \Sigma_{sys}^{1/2} V_{sys}^T \tag{7}$$

We can easily obtain system matrix using extended observability matrices:

$$\begin{aligned} A &= O_{k-1}^{\dagger} O_k(p+1:kn_y, 1:n) \\ C &= O_k(1:n_y, 1:n) \end{aligned} \tag{8}$$

Estimation of Variance Matrices: If we define

$$\bar{C}^T = C_k(1:n, 1:n_y) \tag{9}$$

Variance matrices can be evaluated as below;

$$\begin{aligned} Q &= \Pi_* - A\Pi A^T, \\ S &= \bar{C}^T - A\Pi_* C^T, \\ R &= \Lambda(0) - C\Pi_* C^T \end{aligned} \tag{10}$$

where the essentials of the above formulation is provided by solving the following algebraic Riccati equation;

$$\begin{aligned} \Pi_{k+1} &= A\Pi_k A^T + (\bar{C}^T - A\Pi_k C^T)(\Lambda(0) - C\Pi_k C^T)^{-1}(\bar{C} - C\Pi_k C^T) \\ \Pi_* &= \lim_{k\to\infty} \Pi_k \end{aligned} \tag{11}$$

Innovation Model for state estimation: If we are interested in innovation model for state estimation, we should have an estimation of Kalman gain matrix:

$$K = (\bar{C}^T - A\Pi_* C^T)(\Lambda(0) - C\Pi_* C^T)^{-1} \tag{12}$$

Therefore, states can be estimated using the following dynamic equations,

$$\begin{aligned} x(t+1) &= Ax(t) + Ke(t) \\ y(t) &= Cx(t) + e(t) \\ cov\{e(t)\} &= R \end{aligned} \tag{13}$$

Considering above mentioned notations, we can investigate the deterministic version of SSI algorithms using the following model:

$$\left\{ \begin{array}{l} x(t+1) = Ax(t) + Bu(t) + w(t) \\ y(t) = Cx(t) + Du(t) + v(t) \end{array} \right. , E\left\{ \left[\begin{array}{c} w(t) \\ v(t) \end{array} \right] \left[\begin{array}{cc} w^T(s) & v^T(s) \end{array} \right] \right\} = \left[\begin{array}{cc} Q & S \\ S^T & R \end{array} \right] \delta_{ts} \tag{14}$$

where $u_{\mathrm{t}} \in \mathbb{R}^{n_u}, y_{\mathrm{t}} \in \mathbb{R}^{n_y}, x_{\mathrm{t}} \in \mathbb{R}^{\mathrm{n}}$are samples of input, output, state vectors and $v_{\mathrm{t}} \in \mathbb{R}^{n_y}, w_{\mathrm{t}} \in \mathbb{R}^{\mathrm{n}}$ are stationary, zero mean state noise and output noise vectors. In the case of deterministic identification, subspace system identification problem can be formulated as below [2, 4, 22]: "There are N samples of input vectors $u = \left[\begin{array}{cccc} u_0 & u_1 & u_2 & \ldots & u_{N-1} \end{array} \right]$ and output vectors $y = \left[\begin{array}{cccc} y_0 & y_1 & y_2 & \ldots & y_{N-1} \end{array} \right]$ from a system of order n. Find A, B, C, D, Q, R, S matrices and n for the structure defined in (14)."

There are several different algorithms available for SSI. They usually use some consistence mathematical tools that provide them with pretty useful benefits. There are two well-known subspace system identification algorithms expressed in Table 1. They use the same measurements, the same block Hankel matrices, different types of projections, SVD of different matrices,

the same method for extraction of system order and different extended observability matrices. MOESP does not need to estimate future states of system, but N4SID provides future state vectors by using a weighting matrix. MOESP uses extended observability matrix to extract system matrices but N4SID uses future states and through a least square problem estimates the system matrices. Investigating Table 1 expresses the following advantages for subspace system identification algorithms:

1. SSI Algorithms are the only system identification methods that can easily and extensively be applied to all MIMO and SISO systems.

2. Estimation of system order is one of the steps of SSI algorithms. This advantage reduces amount of time, cost and calculations.

3. SSI methods can handle big packages of data.

4. On-line operations of SSI methods are easier and can easily be applied to MIMO systems.

5. SSI methods use robust mathematical tools such as SVD, LQ decomposition, least square and QR decomposition. They also don't need nonlinear optimization.

6. Some SSI algorithms only use output data to identify a model. This is a considerable advantage.

4 Application of SSI Methods for Small Signal Analysis of Power Systems

The SSI advantages expressed in previous section can be used to overcome the difficulties with classical small signal analysis of power systems. Using SSI methods reduces above five steps to the following three steps:

1. Measuring input/output signals of power system.

2. Identification of a linear model for power system using SSI algorithms.

3. Application of modern small signal methods to provide small signal properties.

As it can be seen, the four first steps vanished and two other steps replaced them. The fifth step left with no change. Therefore, one can provide small signal analysis of power systems in an easier and faster way. Since application of SSI algorithms are very easy, power system small signal analysis will be provided in very low levels of cost and time.

We are usually interested in identification of most oscillatory and least damped modes of power system. Such modes are usually related to electro-mechanical parts of power system. Therefore, in order to identify most critical modes, the angle and speed of electrical machines should be measured. Signal measuring is the starting point of system identification. Since measured signals should have enough persistency of excitation, we should use most effective inputs. In attention to differential equations of a single machine power system [23], mechanical torque and field voltage are proper inputs.

Suppose that input vector u and output vector y of a power system have been measured. The goal is to find small signal properties of power system (Modes, Damping Ratios, Oscillation Frequencies, Participation Factors) using several samples of u and y. It is announced that if

step	Operation	MOESP Algorithm	N4SID Algorithm
1	Model	$\begin{cases} x_{t+1} = Ax_t + Bu_t + w_t \\ y_t = Cx_t + Du_t + v_t \end{cases}$ $E\left\{ \begin{bmatrix} w_t \\ v_t \end{bmatrix} \begin{bmatrix} w_s^T & v_s^T \end{bmatrix} \right\}$ $= \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \delta_{ts}$	$\begin{cases} x_{t+1} = Ax_t + Bu_t + w_t \\ y_t = Cx_t + Du_t + v_t \end{cases}$ $E\left\{ \begin{bmatrix} w_t \\ v_t \end{bmatrix} \begin{bmatrix} w_s^T & v_s^T \end{bmatrix} \right\}$ $= \begin{bmatrix} Q & S \\ S^T & R \end{bmatrix} \delta_{ts}$
2	Measured Data	$u = \begin{bmatrix} u_0 & u_1 & u_2 & ... & u_{N-1} \end{bmatrix}$ $y = \begin{bmatrix} y_0 & y_1 & y_2 & ... & y_{N-1} \end{bmatrix}$	$u = \begin{bmatrix} u_0 & u_1 & u_2 & ... & u_{N-1} \end{bmatrix}$ $y = \begin{bmatrix} y_0 & y_1 & y_2 & ... & y_{N-1} \end{bmatrix}$
3	Block Hankel Matrices	$U_{0,k-1} = \begin{bmatrix} u_0 & u_1 & \cdots & u_{N-1} \\ u_1 & u_2 & \cdots & u_N \\ \vdots & \vdots & \vdots & \vdots \\ u_{k-1} & u_k & \cdots & u_{k+N-2} \end{bmatrix}$ $\in \mathbb{R}^{kn_u \times N}$ $Y_{0,k-1} = \begin{bmatrix} y_0 & y_1 & \cdots & y_{N-1} \\ y_1 & y_2 & \cdots & y_N \\ \vdots & \vdots & \vdots & \vdots \\ y_{k-1} & y_k & \cdots & y_{k+N-2} \end{bmatrix}$ $\in \mathbb{R}^{kn_y \times N}$	$U_{0,k-1} = \begin{bmatrix} u_0 & u_1 & \cdots & u_{N-1} \\ u_1 & u_2 & \cdots & u_N \\ \vdots & \vdots & \vdots & \vdots \\ u_{k-1} & u_k & \cdots & u_{k+N-2} \end{bmatrix}$ $\in \mathbb{R}^{kn_u \times N}$ $Y_{0,k-1} = \begin{bmatrix} y_0 & y_1 & \cdots & y_{N-1} \\ y_1 & y_2 & \cdots & y_N \\ \vdots & \vdots & \vdots & \vdots \\ y_{k-1} & y_k & \cdots & y_{k+N-2} \end{bmatrix}$ $\in \mathbb{R}^{kn_y \times N}$
4	Extra Predefined Matrices	$U_p \triangleq U_{0,k-1}$ $Y_p \triangleq Y_{0,k-1}$	$U_p \triangleq U_{0,k-1}$, $Y_p \triangleq Y_{0,k-1}$, $U_f \triangleq U_{k,2k-1}$, $Y_f \triangleq Y_{k,2k-1}$ $W_p \triangleq \begin{bmatrix} U_p \\ Y_p \end{bmatrix}$, $W_f \triangleq \begin{bmatrix} U_f \\ Y_f \end{bmatrix}$
5	LQ Decomposition	$\begin{bmatrix} U_p \\ Y_p \end{bmatrix} = \begin{bmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_2^T \end{bmatrix}$	$\begin{bmatrix} U_f \\ W_p \\ Y_f \end{bmatrix} = \begin{bmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \\ L_{31} & L_{32} & L_{33} \end{bmatrix} \begin{bmatrix} Q_1^T \\ Q_2^T \\ Q_3^T \end{bmatrix}$
6	Projection	$Y_p / U_p^{\perp} = L_{22} Q_2^T$	$Y_f / {}_{U_f} W_p = L_{32} L_{22}^{\dagger} W_p$
7	Singular Value Decomposition (SVD)	$L_{22} = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \sum_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix}$	$L_{32} L_{22}^{\dagger} W_p = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \sum_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix}$
8	System Order	$n \triangleq \dim(\Sigma_1)$	$n \triangleq \dim(\Sigma_1)$
9	Extended Observability Matrix	$O_k \triangleq U_1 \sum_1^{1/2}$	$O_k \triangleq U_1 \sum_1^{1/2} T \quad , \quad \lvert T \rvert \neq 0$
10	Future State Estimation	———-	$X_f = T^{-1} \Sigma_1^{1/2} V_1^T \in \mathbb{R}^{n \times N}$ $X_k \triangleq X_f = \begin{bmatrix} x_k & x_{k+1} & \cdots & x_{k+N-1} \end{bmatrix} \in \mathbb{R}^{n \times N}$
11	Estimation of State Space Matrixes	$C = O_k(1 : n_y, 1 : n)$ $A = O_k^{\dagger}(1 : n_y(k-1), 1 : n)$ $\cdot O_k(n_y + 1 : kn_y, 1 : n)$ Solving a Least Equare Problem to Estimate B and D	$\begin{bmatrix} \hat{A} & \hat{B} \\ \hat{C} & \hat{D} \end{bmatrix} = \left(\begin{bmatrix} X_{k+1} \\ Y_{k,k} \end{bmatrix} \begin{bmatrix} X_k \\ U_{k,k} \end{bmatrix}^T \right) \cdot \left(\begin{bmatrix} X_k \\ U_{k,k} \end{bmatrix} \begin{bmatrix} X_k \\ U_{k,k} \end{bmatrix}^T \right)^{-1}$

N samples of input/output vectors are availabe, then one can identify the following state space linear model by utilizing a subspace system system identification algorithm:

$$x(t+1) = Ax(t) + Bu(t), \quad y(t) = Cx(t) + Du(t) \tag{15}$$

One can find the system modes and as a result damping factors and damping frequencies by digging the matrix$A$. However, we may encounter some difficulties when calculating the participation factors of modes in states. The problem arises since the state vector x of the identified model is not correspondent to that of the real power system which we may obtain by analytical methods. Therefore, mode in state participation factors can't be utilized using the identified A.

In order to cope with such a problem, it is proposed to use modal canonical realization of (15). Using T as a similarity transformation matrix, one can provide the following modal canonical realization:

$$\begin{aligned} &\dot{z} = \Lambda z + \bar{B}u, \qquad y = \bar{C}z + \bar{D}u \\ &x = Tz,\ \Lambda = T^{-1}AT,\ \bar{B} = T^{-1}B, \bar{C} = CT,\ \ \bar{D} = D \end{aligned} \tag{16}$$

Generally, Λ is in Jordan and block diagonal structure. Thus, mode in state participation factor (p_{ki}) is defined as

$$p_{ki} \triangleq \frac{\partial \lambda_i}{\partial a_{kk}} \tag{17}$$

where a_{kk} is a diagonal element of system matrix. Since in (16), the system matrix is diagonal with modes as its diagonal elements, we can write:

$$p_{ki} \triangleq \frac{\partial \lambda_i}{\partial a_{kk}} = \frac{\partial \lambda_i}{\partial \lambda_i} = 1 \tag{18}$$

Therefore, modal canonical realization can maximize (100 %) mode in state participation factor of model. In order to clarify the point, suppose that u is zero, Λ is diagonal and z^0 is the initial condition vector of modal canonical realization. Thus, we can write:

$$\dot{z} = \Lambda z \ \ \Rightarrow \ \ z_i = e^{-\lambda_i t} z_i^0 \ \ , \ i = 1, 2, \cdots, n \tag{19}$$

Therefore, the only participating mode in state z_i is λ_i, so the participation factor of mode λ_i in the state z_i is 100% and each mode is mapped to a state. Considering above point and output equation of (15), one can write:

$$y = \bar{C}z \ \ \Rightarrow \ \ y_k = \textstyle\sum_{i=1}^{n} \bar{c}_{ki} z_i = \sum_{i=1}^{n} \bar{c}_{ki} z_i^0 e^{-\lambda_i t}, \quad k = 1, \cdots, n_y \tag{20}$$

Therefore, output y_k is affected by mode λ_i and mode in output participation factor (p_{ki}) is proposed as:

$$p_{ki} \triangleq \bar{c}_{ki} z_i^0 \ \ , k = 1, \cdots, n_y \ \ and \ i = 1, \cdots, n \tag{21}$$

In order to provide participation factors, one may need z^0 which can be provided through following relation:

$$z^0 = T^{-1} x^0 \tag{22}$$

x^0 is the initial condition vector of the identified state space model. It is also provided by SSI algorithms.

Some investigators [24] discuss another kind of participation called *State in Mode* participation factor. In most of literatures *state in mode* and *mode in state* participation factors are considered the same as each other and they have been used interchangeably. However,there is a discussion on some differences in [24]. The following proposition challenges the definition of *state in mode* participation factor.

Proposition 1. *"State in Mode Participation" is a meaning-less expression.*

Proof: A system includes some physical elements (such as resistors, capacitors and inductors in electrical systems or dampers, springs and masses in mechanical systems). Each element has a value and a role in topology and configuration of system. Configuration and topology of system provides the system with unique set of differential equations which we call it mathematical model.

A presentation of system mathematical model is State Space Structure which has a system matrix called A which is affected by basic elements and configuration of system. Systems modes are one of the mathematical properties of A, therefore system modes only rely on system elements and system configuration. System states don't affect its modes so there is no state in mode participation and system modes are not affected by its states. □

Example 2. *For instance, one can provide the series RLC circuit with the following state space model:*

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} v_i, \qquad v_c = \begin{bmatrix} \frac{\lambda_1\lambda_2}{(\lambda_1-\lambda_2)} & \frac{-\lambda_1\lambda_2}{(\lambda_1-\lambda_2)} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{23}$$

where

$$\lambda_{1,2} = -\xi\omega_n \pm j\omega_n\sqrt{1-\xi^2}, \;\; \omega_n = 1/\sqrt{LC}, \quad \xi = (R/2)/\sqrt{L/C} \tag{24}$$

Therefore, $\lambda_{1,2}$ rely only on L, C and R and the topology of circuit and the states depend on system modes$\lambda_{1,2}$. Thus, there is no participation of states in modes, and the expression state in mode participation is meaningless.

5 Test Cases and Simulations

Identification process should be provided with sampled input/output signals. Therefore, Computer simulations of the following test case power systems have been conducted using MATLAB/SIMULINK installed on a computer with 2.4 GHz CPU and 4G RAM.

We study the systems shown in Figure 1; A single machine connected to an infinite bus, and a four-machine two-area system. Since we may extract the small signal properties of generator angle and speed, it is recommended to use mechanical torque as input and rotor speed and/or its angle as output signals. Torque and power are the same in per-unit system. Therefore, we use mechanical power as input signal. In order to have enough persistency of excitation in input signals, we added a white noise to input signals. To provide more realistic operating conditions, we added a white noise to output signals, as well. Effect of noises will be investigated later.

5.1 Single Machine Three-Bus System

Power system shown in the left side of Figure 1 is a three bus single machine power system with no control and exciter. The parameters of the system are those used in [23]. We are supposed to extract all small signal properties of system using SSI algorithms and the methods illustrated in previous sections.

We acquired 300 samples of input/output data through a 30 second simulation. Using the SSI algorithms presented in Table 1 , some linear models were identified and the results are presented in Table 2. It is clear that for investigating performance of noises, we cannot manipulate their averages since the operating point may vary. This is not applicable in this study. Therefore, each noise variance was altered separately in order to see its effect. In Table 2, It is expressed that an increase in input noise variance may lead to a better model from the view of FPE (Final Prediction Error) measure. However, we should be conservative when the estimation of

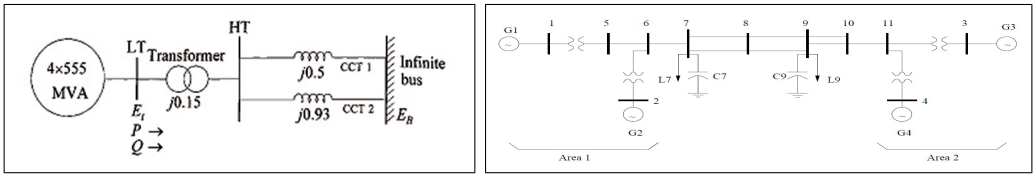


Figure 1: Left; Single machine three-bus power system. Right; Two-area four-machine power system.

	σ_u	σ_y	λ	ξ	$\omega_n\,(Hz)$	$\lvert P\rvert$	**FPE**
CM	-	-	$-0.714 \pm j6.35$	0.112	1.0165	$\begin{bmatrix} 0.503 & 0.503 \\ 0.503 & 0.503 \end{bmatrix}$	-
SSIM01	0.0001	0	$-0.7975 \pm j6.2443$	0.1276	0.9988	$\begin{bmatrix} 0.5037 & 0.5037 \\ 0.5037 & 0.5037 \end{bmatrix}$	7.4e-5
SSIM02	0.001	0	$-0.7192 \pm j6.3203$	0.1139	1.0105	$\begin{bmatrix} 0.5028 & 0.5028 \\ 0.5028 & 0.5028 \end{bmatrix}$	8.5e-5
SSIM03	0.01	0	$-0.7727 \pm j6.1300$	0.1253	0.9825	$\begin{bmatrix} 0.5033 & 0.5033 \\ 0.5033 & 0.5033 \end{bmatrix}$	11e-5
SSIM04	0.001	0.01	$-0.7192 \pm j6.3203$	0.1139	1.0105	$\begin{bmatrix} 0.5028 & 0.5028 \\ 0.5028 & 0.5028 \end{bmatrix}$	0.59
SSIM05	0.001	0.01	$-0.7192 \pm j6.3203$	0.1139	1.0105	$\begin{bmatrix} 0.5028 & 0.5028 \\ 0.5028 & 0.5028 \end{bmatrix}$	0.60

Table 2: Small Signal Analysis of Single Machine Power System Using SSI Algorithms; abriviations are: Classical Model (CM), Subspace System Identified Model (SSIM), Participation Matrix (P), variance of input noise (σ_u), variance of output noise (σ_y), eigenvalue (λ), damping factor (ξ), natural frequency (ω_n), and Final Prediction Error (FPE).

small signal properties is under consideration. Actually, a large increase in input noise variance may alter the operating point or its absorption area. This may lead to instability. In Table 2, if we compare SSIM4 and SSIM5 with SSIM2, we can see that output noise has no effect on subspace system identification. Actually, additive output noise does not have considerable effects on the SSI identification, since SSI algorithms use robust linear algebra tools. It is noticeable that the Left eigenvector of a wide matrix are not considerably sensitive to additive white noise, [2]. Therefore, the identification is not sensitive to additive output noise. What is more, the identification process has no effect on normal operating conditions of power system, since the applied input noise is too weak.

5.2 Two-Area Four-Machine System

The power system introduced in [25] is used as the second case study (see the right side of Figure 1). This power system has four generators and two fully symmetrical areas linked together by a weak line. It was specifically designed [23] to study low frequency electromechanical oscillations in large interconnected power systems. Despite its small size, it can thoroughly mimic the behavior of typical systems in actual operation.

In the case of no PSS (Power System Stabilizer), when the measured signals are differential speeds and differential angles of all generators, the CVA algorithm detected 10 modes shown in Figure 2. As illustrated in Figure 2, there are 8 oscillatory modes, and two non-oscillatory modes. There are three zero modes, since there is no infinite bus as a reference for rotor angles.

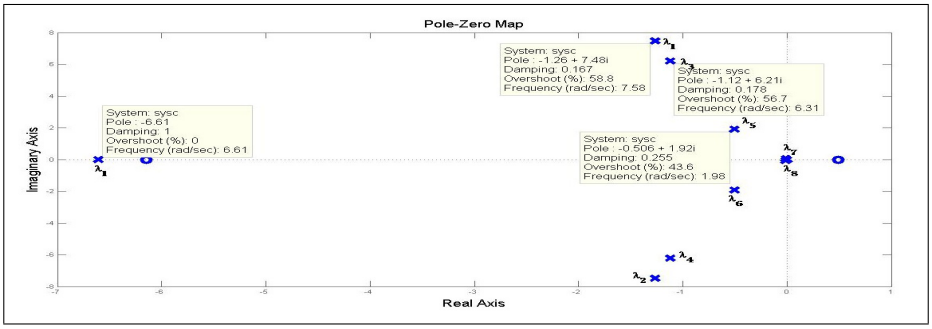


Figure 2: Pole-zero map of the identified model foe two-area system (No PSS) using CVA.

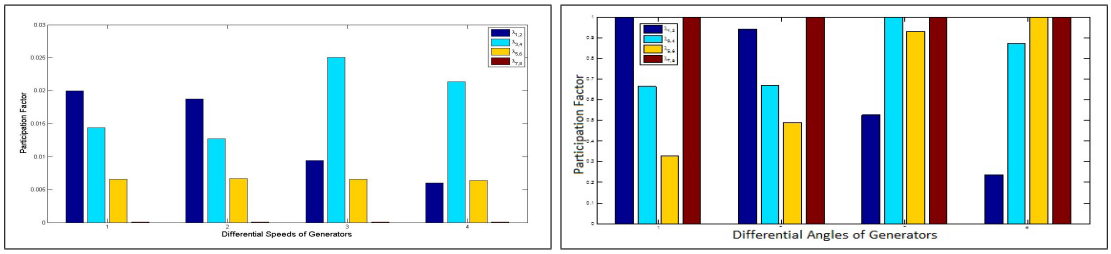


Figure 3: Left: Normalized participation factors of modes in differential speeds of generators. Right: Normalized participation factors of modes in differential Angles of generators. (No PSS, measurements are differential Angles and differential speeds).

Moreover, the speed governors have not been modeled [23], ($\lambda_{7,8}$ and a real pole in the origin).

In order to distinguish between local and inter-area modes, the participation factors of modes are illustrated in Figure 1. Participation factors were calculated using the approach presented in previous section. Table 3 illustrates the small signal properties of two-area system extracted from Figure 2 and Figure 1.

In order to damp oscillatory modes, different power system stabilizers have been designed and applied to two-area power system. Performances of PSS have been investigated using SSI algorithms and the method mentioned in previous section. Table 4 summarizes the damping performance of Multi-Band (MB) PSS, $\Delta\omega$ PSS and ΔPa PSS [26] when applied to two-area power system. All PSSs perform very well. They stabilize the naturally unstable system. However, it is clear that the MB-PSS is superior to the other two PSSs, since it provides significantly more damping for all modes. The results presented here using identification method, are identical to those analytically achieved in [26].

Mode	$\omega_d(Hz)$	$\xi(\%)$	**Participated States**
$\lambda_{1,2}$	1.1904	16.70	δ_1, δ_2 (Local, Area 1)
$\lambda_{3,4}$	0.9871	17.84	$\delta_3, \delta_4, \omega_3, \omega_4$ (Local, Area 2)
$\lambda_{5,6}$	0.3061	24.35	$\delta_3, \delta_4, \omega_1, \omega_2, \omega_3, \omega_4$ (Inter-area)
$\lambda_{7,8}$	0.0095	20.78	$\delta_1, \delta_2, \delta_3, \delta_4, \omega_1, \omega_2, \omega_3, \omega_4$ (Inter-area)

Table 3: Small Signal Properties of Two Area Test System (No PSS).

PSS	$\omega_d(Hz)$	$\xi(\%)$	Mode Type
No PSS	1.1904	16.70	Local
	0.9871	17.84	Local
	0.3061	24.35	Inter-area
MB PSS	0.3942	51.91	Inter-area
$\Delta\omega$ PSS	0.7115	47.71	Inter-area
	0.1113	67.82	Controller
Pa PSS	0.4655	43.02	Inter-area

Table 4: Investigating the effects of PSS on Oscillatory modes of two-area system using SSI.

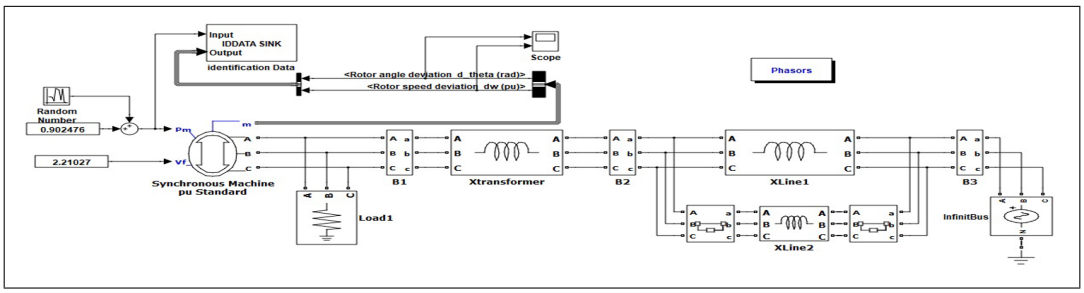


Figure 4: Monitoring of single machine three-bus system

6 On-line Monitoring of Power System Using SSI

On-line monitoring of power system is an open area which attracts tremendous consideration of investigators. Using analytic method for on-line monitoring may lead in calculation problems and the results are not also trustfully. Since the identification process uses sampled signals of power system, there would not be any gap between identification results and power system behavior. Therefore, the results are trustworthy in the case of power system monitoring. SSI algorithms can easily be applied to on-line power system monitoring. They can provide a state space model which is most suitable for on-line monitoring.

A single machine three-bus power system with two transmission lines is used to investigate on-line monitoring of power systems (Figure 2). Some computer simulations were conducted for 60 second. In the first 30 seconds, both lines are in operating condition. As a result of a fault, for the second 30 seconds, line 2 is out of work. Some state space models were identified in every 20 seconds using SSI algorithms. There would be three models and the operating point will change for the second model, so one can call it as an operating point transient model.

Three identified model and their properties are available in Table 5. SSIM6 represents the system performance in the initial operating point before the fault. SSIM8 represents system performance after breaking out of second line. SSIM7 is a middle model between SSIM6 and SSIM8.

As it can be seen from the column of modes column in Table 5 , modes has moved slightly to the right and as a result stability of system has reduced. It implies that the operating point should be re-tuned in order to recover more stable operating conditions.

Damping frequency and also damping factor has reduced. It implies the need for some control in order to provide power system with more suitable synchronizing and damping factors.

It is considerable that the whole monitoring process expresses a little movement of eigenvalues toward the right half plane. Moreover, there is a decrease in damping frequency. Therefore, a control should be applied to mechanical torque and field voltage, because it is predictable that a variation in operating conditions of power system may lead to instability and oscillatory behavior.

Model	λ	$\omega_d\,(Hz)$	ξ	$\lvert P\rvert$
SSIM6	$-0.6010 \pm j8.0143$	1.2755	0.0748	$\begin{bmatrix} 0.5016 & 0.5016 \\ 0.5016 & 0.5016 \end{bmatrix}$
SSIM7	$-0.5269 \pm j7.0677$	1.1249	0.0743	$\begin{bmatrix} 0.5012 & 0.5012 \\ 0.5012 & 0.5012 \end{bmatrix}$
SSIM8	$-0.4810 \pm j6.9452$	1.1054	0.0691	$\begin{bmatrix} 0.5018 & 0.5018 \\ 0.4775 & 0.4775 \end{bmatrix}$

Table 5: Subspace System Identification Models (SSIM) identified during Online Monitoring.

7 Effect of Different SSI Algorithms on Small Signal Analysis of Power Systems

There are several kinds of subspace system identification algorithms. They may all use the same mathematical tools. But they are different in applying tools to measurement data. Therefore, SSI algorithms may show slightly different results for the same input/output set of data. There are three main algorithms which are usually used in the papers discussing power system analysis; MOESP (Multi-variable Output-Error State Space), N4SID (Numerical State Space Subspace System IDentification) and CVA (Canonical Variate Analysis). We are supposed to compare performance of such algorithms when they are used for small signal analysis of power systems in this section.

Figure 1 shows the system under study. Small signal analysis of mentioned power system was conducted using four models and different SSI algorithms. The models were investigated; 1) Classical model of 3-bus single machine power system which has no control and the effect of field is ignored. 2) Single machine power system model containing effects of field voltage. 3) Auto-voltage regulated (AVR) power system. 4) Single machine power system with AVR and power system stabilizer (PSS). In each case, K_D were calculated using the system parameters available in [23].

Table 6 contains results of small signal analysis for several models of single machine power system using two different SSI algorithms. Identification results are almost the same for two SSI algorithms. They are slightly different in estimation of non-oscillatory modes. Table 6 also expresses that SSI algorithms are not capable of identifying non-dominant modes. That is not a draw-back, since non-dominant modes are not so important for system analysis. As a whole, one can use subspace system identification algorithms for sure, when it is needed to estimate most effective modes of power system. Therefore, Subspace system identification algorithms are convenient solution for extracting small signal stability properties of a power system.

8 Conclusions and Future Works

The paper discusses pitfalls of small signal analysis of power systems when using analytic methods. Analytic methods use basic theories and the provided parameters in order to extract operating point and linearized model of power system. In this paper, it is expressed that we can avoid such boring, time consuming and cost effective stages, using system identification methods. The most proper method for power system small signal analysis is then, application of subspace system identification (SSI) algorithms. SSI methods have many other advantages which were discussed in the paper. Extraction of modes and their participation factors are easy when using SSI methods but it needs some modification in SSI algorithms. In order to fulfill such requirements, some suggestions presented in this paper. Using SSI algorithms, some ideas

No.		Classical	MOESP	CVA
1	λ	$0 \pm j6.39$	$0 \pm j6.4158$	$0 \pm j6.4158$
	ξ	0.0000	0.0000	0.0000
	ω_d	1.0200 Hz	1.0211 Hz	1.0211 Hz
2	λ	$\lambda_{1,2} = -0.11 \pm j6.41$ $\lambda_3 = -0.2040$	$\lambda_{1,2} = -0.1095 \pm j6.4114$ $\lambda_3 = -0.2038$	$\lambda_{1,2} = -0.1095 \pm j6.4114$ $\lambda_3 = -0.2038$
	ξ	0.0170	0.0171	0.0171
	ω_d	1.0700 Hz	1.0204 Hz	1.0204 Hz
3	λ	$\lambda_{1,2} = +0.5040 \pm j7.2300$ $\lambda_3 = -20.202$ $\lambda_4 = -31.230$	$\lambda_{1,2} = +0.5045 \pm j7.2321$ $\lambda_3 = -0.0015$ $\lambda_4 = -15.1521$	$\lambda_{1,2} = +0.5045 \pm j7.2321$ $\lambda_3 = -0.0310$ $\lambda_4 = -15.1174$
	ξ	-0.0700	-0.0696	-0.0696
	ω_d	1.1500 Hz	1.1510 Hz	1.1510 Hz
4	λ	$\lambda_1 = -0.7390$ $\lambda_{2,3} = -1.005 \pm j6.607$ $\lambda_{4,5} = -19.797 \pm j12.822$ $\lambda_6 = -39.097$	$\lambda_1 = 0.0000$ $\lambda_2 = -4.5770$ $\lambda_{3,4} = -1.0357 \pm j6.6150$ $\lambda_{5,6} = -16.857 \pm j16.288$	$\lambda_1 = 0.0000$ $\lambda_2 = -4.6266$ $\lambda_{3,4} = -0.9946 \pm j6.5951$ $\lambda_{5,6} = -16.906 \pm j16.265$
	ξ	$\xi_{2,3} = 0.15$ $\xi_{4,5} = 0.84$	$\xi_{3,4} = 0.1547$ $\xi_{5,6} = 0.7191$	$\xi_{3,4} = 0.1491$ $\xi_{5,6} = 0.7204$
	ω_d	$\omega_{d2,3} = 1.05Hz$ $\omega_d 4,5 = 2.04Hz$	$\omega_{d2,3} = 1.0528Hz$ $\omega_{d4,5} = 2.5924Hz$	$\omega_{d2,3} = 1.0496Hz$ $\omega_{d4,5} = 2.5886Hz$

Table 6: Small signal analysis of single machine power system using different types of subspace system identification algorithms. 1) Classical model (No field control, K_D=0) . 2) System model containing effects of field voltage, (K_D=1.53). 3) Auto-voltage regulated (AVR) system, ($K_D = -7.06$). 4) System model with AVR and power system stabilizer (PSS), ($K_D = 14.08$)

such as power system monitoring can be conducted in a more convenient and more realistic way. This point is discussed in several parts of paper while applying SSI algorithms to different power systems.

Future works may aim at application of subspace system identification methods to monitor modes, stability measures and estimation of damping and synchronous factors of power system.

Bibliography

[1] P. van Overschee and B. de Moor, A unifying theorem for three subspace system identification algorithms, *Automatica*, vol. 31, pp. 1853-1864, 1995

[2] T. Katayama, Subspace Methods for System Identification: Springer, 2005

[3] S. J. Qin, An Overview of Subspace Identification, Elsevier, *Computer & Chemical Engineering*, vol. 30, pp. 1502-1513, 2006

[4]] P. Van Overschee and B. De Moor, "N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems, *Automatica*, vol. 30, pp. 75-93, 1994

[5] M. Verhaegen and P. Dewilde, Subspace model identification Part 1. The output-error state-space model identification class of algorithms, *International Journal of Control*, vol. 56, pp. 1187-1210, 1992/10/01 1992

[6] M. Verhaegen and P. Dewilde, Subspace model identification Part 2. Analysis of the elementary output-error state-space model identification algorithm, *International Journal of Control*, vol. 56, pp. 1211-1241, 1992/10/01 1992

[7] M. Viberg, Subspace-based methods for the identification of linear time-invariant systems, *Automatica*, vol. 31, pp. 1835-1851, 1995

[8] M. Verhaegen, Identification of the deterministic part of MIMO state space models given in innovations form from input-output data, *Automatica*, vol. 30, pp. 61-74, 1994

[9] W. E. Larimore, "Canonical variate analysis in identification, filtering, and adaptive control, in *Decision and Control, Proceedings of the 29th IEEE Conference on*, 1990, pp. 596-604

[10] S. J. Qin, An overview of subspace identification, *Computers & Chemical Engineering*, vol. 30, pp. 1502-1513, 2006

[11] G. T. I. Kamwa, L. Gerin-Lajoe, Low Order Black-Box Models for Control System Design Large Power Systems, *IEEE Transaction on Power Systems*, Vol. 11, No. 1, February 1996

[12] S. M. I. H. Akcay, B. Ninness, Subspace-Based Identification of Power Transformer Models from Frequency Responce Data, *IEEE Transaction on Instrumentation and Measurement*, Vol. 48, No. 3, June 1999

[13] H. A. T. McKelvey, Lennart Ljung, Subspace-Based Multivariable System Identification from Frequency Responce Data, *IEEE Transaction on Automatic Control*, Vol. 41, No. 7, July 1996

[14] S. M. I. H. Akcay, B. Ninness, Identification of Power Transformer Models from Frequency Responce data: A Case Study. Elsevier, *Signal Processing*, 68, pp 307-315, 1998

[15] O. P. M. M. Karrari, Identification of Heffron-Phillips Model Parameters for Synchronous Gemerators Using Online Measurment, *IEEE Proceeding of Generator, Transmition, Distribution*, Vol. 151, No. 3, May 2004.

[16] D. W. O. P. M. M. Soliman, Identification of Heffron-Phillips Model Parameters for Synchronous Generators Operating in Closed Loop, *IEEE Generation, Transmition and Distribution*, 2(4):530-541, 2008

[17] O. P. M. Bin Wu. Multivariable Adaptive Control of Synchronous Machines in a Multimachine Power System. IEEE Transaction on Power Systems, Vol. 21, No. 4, November 2006

[18] J. W. P. Ning Zhou, J. F. Hauer, Initial Results in Power System identification From Injected Probing Signals Using a Subspace Method, *IEEE Transaction on Power Systems*, Vol. 21, No. 3, August 2006

[19] C. C. H. Ghasemi, A. Moshref, Oscillatory Stability Limit Prediction Using Stochastic Subspace Identification, *IEEE Transaction on Power Systems*, 21(2):736-745, 2006

[20] L. G.-L. I. Kamwa, G. Trudel, Multi-Loop Power System Stablizers Using Wide-Area Synchronous Phasor Measurements, *Proceedings of the American Control Conference*, Philadelphia, Pennsylvania, June 1998

[21] D. Y. G Cai, Y. Jiao, Ch. Shao, Power System Oscillation Mode Analysis and Parameter Determination of PSS Based on stochastic Subspace Identification, *Asia-Pacific Power and Energy Engineering Conference (APPEEC)*, pp. 1-6, March 2009

[22] P. Overschee and B. L. R. Moor, Subspace identification for linear systems: theory, implementation, applications: Kluwer Academic Publishers, 1996

[23] PP. Kundur and N. J. Balu, Power System Stability and Control: IEEE, 1998

[24] M. A. H. Wael A. Hashlamoun, and Eyad H. Abed, New Results on Modal Participation Factors: Revealing a Previously Unknown Dichotomy. *IEEE Transactions of Automatic Control*, Vol. 54, No. 7, July 2009

[25] J. J. Sanchez-Gasca, V. Vittal, M. J. Gibbard, A. R. Messina, D. J. Vowles, S. Liu, and U. D. Annakkage, Inclusion of higher order terms for small-signal (modal) analysis: committee report-task force on assessing the need to include higher order terms for small-signal (modal) analysis, *Power Systems, IEEE Transactions on*, 20(4):1886 - 1904, 2005

[26] R. Klein, Moorty and Kundur. Analytical investigation of factors influencing PSS performance, *IEEE Trans. on EC*, 7(3):382-390, 1992

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 2 (June), pp. 341-352

Robust Adaptive Neural-Fuzzy Network Tracking Control for Robot Manipulator

T. Ngo, Y. Wang, T.L. Mai, M.H. Nguyen, J. Chen

ThanhQuyen Ngo, YaoNan Wang, T. Long Mai,
M. Hung Nguyen, Jun Chen
College of Electrical and Information Engineering
Hunan University, Changsha, Hunan Province 410082, P.R.China
Faculty of Electrical Engineering
HCM City University of Industry, HCM City, Vietnam
thanhquyenngo2000@yahoo.com, yaonan@hnu.cn,
mailongtk@gmail.com, manhhung@yahoo.com, 2011junchen@gmail.com

Abstract: This paper presents a robust adaptive neural-fuzzy network control (RANFNC) system for an n-link robot manipulator to achieve the high-precision position tracking. Initially, the model dynamic of an n-link robot manipulator is introduced. However, it is difficult to design a conformable model-based control scheme, for instance, external disturbances, friction forces and parameter variations. In order to deal with this problem, the RANFNC system is investigated to the joint position control of an n-link robot manipulator. In this control scheme, a four-layer neural-fuzzy-network (NFN) is used for the main role, and the adaptive tuning laws of network parameters are derived in the sense of a projection algorithm and the Lyapunov stability theorem to ensure network convergence as well as stable control performance. The merits of this model-free control scheme are that not only the stable position tracking performance can be guaranteed but also unknown system information and auxiliary control design are required in the control process. The simulation results are provided to verify the effectiveness of the proposed RANFNC methodology.
Keywords: Adaptive control, Neural-fuzzy network, robot manipulator.

1 Introduction

In the past decade, the applications of intelligent control techniques (fuzzy control or neural control) to the motion control of robotic manipulator have received considerable attention [1]-[3]. In general, robotic manipulators have to face various uncertainties in their dynamics, such as friction, and external disturbance. It is difficult to establish exactly mathematical model for the design of a model-based control system. Thus, the general claim of these intelligent control approaches is that they can reduce the effects of structured parametric uncertainty and the unstructured disturbance by using their powerful learning ability without a detailed knowledge of the controlled plant in the design processes.

Recently, the concept of incorporating fuzzy logic into a neural network has grown into a popular research topic [5]-[8]. The integrated neural-fuzzy-network system possesses the merits of both fuzzy system [9] (e.g., humanlike IF-THEN rule thinking and ease of incorporating expert knowledge) and neural network [10] (e.g., learning and optimization abilities and connectionist structure). In this way, one can bring the low-level learning and computational power of neural network into fuzzy systems as well as the high-level humanlike IF-THEN rule thinking and reasoning of fuzzy systems into neural network.

The main of this paper is to design an intelligent control system scheme for the position control of an n-link robot manipulator by using neural-fuzzy-network controller to compensated uncertainness dynamic model and external disturbance via capability self-learning of neural network and human intuitive.

This paper is organized as follows: Section 2 described a dynamic model of an n-link robot manipulator briefly [11]. Section 3 presents a structure of neural-fuzzy-network. The design process of the RANFNC system is investigated to control an n-link robot manipulator for periodic motion in section 4. The design procedures of the proposed RANFNC system are described in detail. The adaptive learning laws in the RANFNC system are designed in the sense of the Lyapunov stability theorem [12], [13] so that the network convergence and system tracking stability can be guaranteed in the closed-loop control system. Numerical simulation results of a two-link robot manipulator under the possible occurrence of uncertainties are provided to demonstrate the tracking control performance of the proposed RANFNC system in section 5. Conclusions are drawn in section 6.

2 System Description

2.1 Robotic Dynamic Model

In general, the dynamic of an n-link robot manipulator may be expressed in [11] as:

$$M(q)\ddot{q} + C(q,\dot{q})\dot{q} + G(q) + F(\dot{q}) + \tau_d = \tau \tag{1}$$

Where $q, \dot{q}, \ddot{q} \in \Re^n$ are the joint position, velocity and acceleration vectors, $M(q) \in \Re^{n\times n}$ denotes the inertia matrix, $C(q,\dot{q}) \in \Re^{n\times n}$ expresses the matrix of centripetal and Coriolis forces, $G(q) \in \Re^n$ is the gravity vector, and $F(\dot{q}) \in \Re^n$ is the friction. Bounded unknown disturbances are denoted by τ_d and the control input torque is $\tau(t) \in \Re^n$. In this paper a robot manipulator is shown in Fig.1 which is utilized to verify dynamic properties are given in section 5.

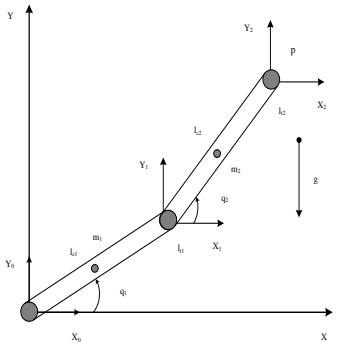


Figure 1: Architecture of two-link robot manipulator.

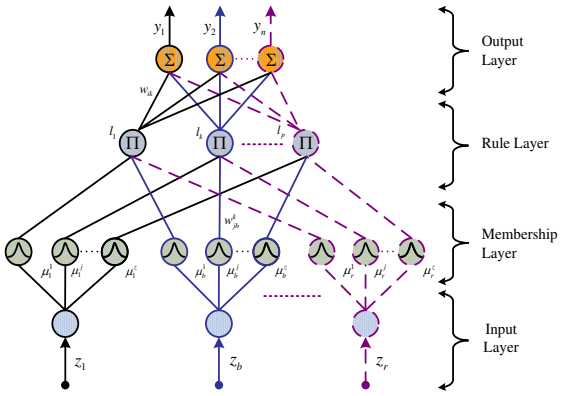


Figure 2: Structure of four-layer NFN

Given a desired arm trajectory $q_d(t) \in \Re^n$ the tracking error is:

$$e(t) = q_d(t) - q(t) \tag{2}$$

And the filtered tracking error (in standard use in robotics) is:

$$r(t) = \dot{e} + \lambda e \tag{3}$$

Where $\lambda = \lambda^T > 0$, differentiating $r(t)$ and using (1), the arm dynamics may be written in terms of the filtered tracking error as:

$$M(q)\dot{r} = -C(q,\dot{q})r - \tau + M(q)(\dot{q}_d + \lambda\dot{e}) + C(q,\dot{q})(\dot{q} + \lambda e) + F(\dot{q}) + G(q) + \tau_d \tag{4}$$

Where the nonlinear robot function is:

$$f(x) = M(q)(\ddot{q}_d + \lambda\dot{e}) + C(q,\dot{q})(\dot{q} + \lambda e) + G(q) + F(\dot{q}) \tag{5}$$

Substituting (5) equation into (4) we have:

$$M(q)\dot{r} = -C(q,\dot{q})r - \tau + f(x) + \tau_d \tag{6}$$

And, for instance $x = \begin{bmatrix} e^T & \dot{e}^T & q^T & \dot{q}^T & \ddot{q}^T \end{bmatrix}^T$

2.2 Defined Control Law

Now, we define a control input torque as:

$$\tau_0 = \hat{f}(x) + K_v r \tag{7}$$

With $\hat{f}(x)$ is an estimate of $f(x)$ and a gain matrix $K_v = K_v^T$ the closed-loop system becomes:

$$M(q)\dot{r} = -(K_v + C(q,\dot{q}))r + \tilde{f}(x) + \tau_d \equiv -(K_v + C(q,\dot{q}))r + \zeta_0, \quad \zeta_0 = \tilde{f}(x) + \tau_d \tag{8}$$

Where functional estimate error is given by:

$$\tilde{f}(x) = f(x) - \hat{f}(x) \tag{9}$$

This is a system error wherein filtered tracking error is driven functional estimate error.

The control τ_0 incorporates a proportional plus derivative (PD) term in $K_v r = K_v(\dot{e} + \lambda e)$.

In the remainder of the paper we shall use (8) to focus on selecting NFN weight tuning algorithms that guarantee the stability of the filtered tracking error $r(t)$. Then, since (3), with the input considered as $r(t)$ and the output as $e(t)$ describes a stable system, standard techniques [13], [15] guarantee that $e(t)$ exhibits stable behaviour. In fact $\|e\|_2 \leq \|r\|_2/\sigma_{min}$, $\|\dot{e}\|_2 \leq \|r\|_2$ with $\sigma_{min}(\lambda)$ the minimum singular value of λ. Generally λ is diagonal, so that $\sigma_{min}(\lambda)$ is the smallest element of λ.

The following properties of the robot dynamics are required [15]. They hold for all revolute rigid-link manipulator.

The inertial matrix $M(q)$ is symmetric and positive definite. It is also bounded as a function of $q : m_1 I \leq m_2 I$. $\dot{M}(q) - 2C(q,\dot{q})$ is a skew symmetric matrix, that is $y^T[\dot{M}(q) - 2C(q,\dot{q})]y$, where y is a $n \times 1$ nonzero vector. The gravity vector $G(q)$ is bounded as a function of $q : G(q) \leq g_d$. The unknown disturbance satisfies $\|\tau_d\| \leq d_b$ and this property is new, that is, the dynamic (8) from $\zeta_0(t)$ to $r(t)$ are a state strict passive system.

3 Structure of NFN

In the RANFNC scheme, as shown in Fig.3, an NFN estimator is designed to tune the nonlinear dynamic function vector, and then, the estimative vector is used to indirectly develop a stable RANFNC law. An NFN controller is directly designed to imitate a predetermined model-based stabilizing control law, and then, the stable control performance can be achieved by using joint position and filtered error vector and information. In this paper, a four-layer NFN

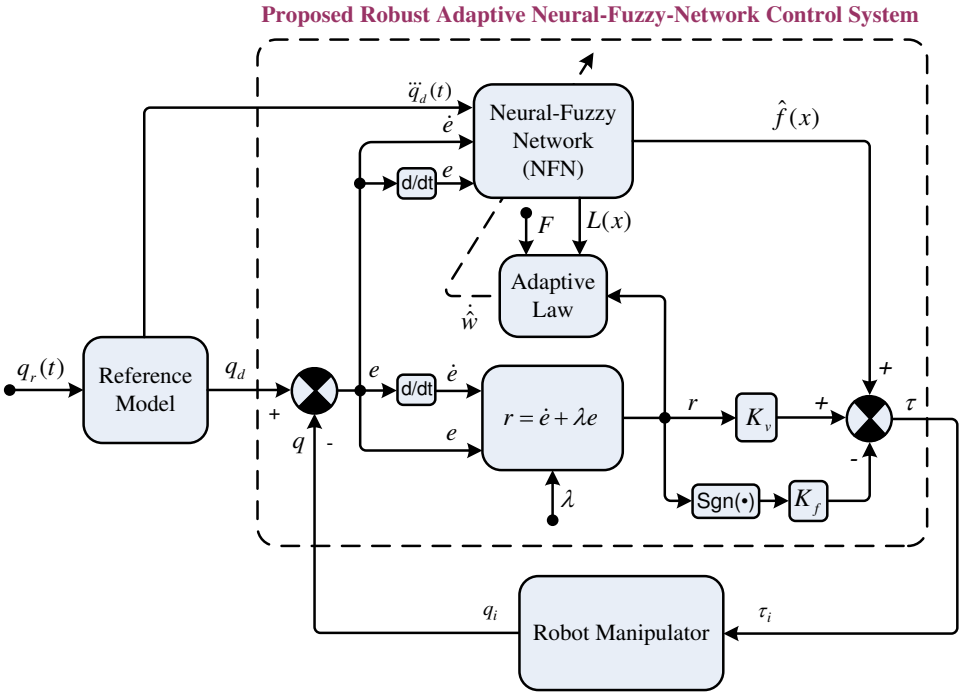


Figure 3: Block diagram of RANFNC scheme

structure showns in Fig.2, which is composed of input, membership, rule, and output layers, is adopted to implement the NFN estimate in RANFNC. The signal propagation and the basic function in each layer of the NFN are introduced as follows.

1. Input layer transmits the input linguistic variables $z_b|_{1,2,\cdots,r}$ to the next layer.

2. Membership layer represents the input values with the following Gaussian membership functions:

$$\mu_b^j(z_b) = \exp\left[-\frac{(z_b - m_b^j)^2}{(t_b^j)^2}\right] \tag{10}$$

Where exp[·] is the exponential function, m_b^j and $t_b^j (b = 1, 2, \cdots, r; j = 1, 2, \cdots, z)$ are the mean and the standard deviation of the Gaussian function in the jth term of the bth input variable z_b to the node of membership layer, respectively. It can be referred as the fuzzification procedure.

3. The output of each node in the rule layer is determined by fuzzy *And* operation. Each node in this rule layer is denoted by $\prod$, which multiplies the input signals and output the result of the product. The product operation is utilized to determine the firing strength. It can be referred as the fuzzy inference mechanism. The output of this layer is given as:

$$l_k = \prod_{b=1}^{r} w_{jb}^k \mu_b^j(z_b) \tag{11}$$

Where $l_k|_k = 1, 2, \cdots, p$ represents the kth output of the rule layer, w_{jb}^k which represents the weights between the membership layer and the rule layer, is assumed to be unity, and p is the number of rules.

4. Final layer is the output layer, and nodes in this layer represent the output linguistic variables. Each node in the output layer $y_i(1, 2, \cdots, n)$ is labelled as $\sum$, which computes the

overall output as the summation of all input signals, and be represented as:

$$y_i = \sum_{k=1}^{p} w_{ik} l_k \tag{12}$$

The output node, together with the links connected to it, acts as a defuzzifier. It can be referred as the normal defuzzification procedure. Moreover, it can be rewritten in the following vector form:

$$y = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}^T = WL \tag{13}$$

Where

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1p} \\ w_{21} & w_{22} & \cdots & w_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{np} \end{bmatrix} = \begin{bmatrix} w_1 & w_2 & \cdots & w_n \end{bmatrix} \in \Re^{n \times p},$$

$$L = \begin{bmatrix} l_1 & l_2 & \cdots & l_n \end{bmatrix}^T \in \Re^{p \times 1}$$

In the RANFNC scheme, the NFN is used to estimate unmodeled nonlinear function, Moreover, the RANFNC law and adaptive tuning algorithms for NFN are introduced from the stability analyses of the closed-loop system by using Lyapunov method. The input of the NFN estimator are the elements in the filtered error vector and joint positions signal, the output of the NFN estimator are the nonlinear dynamic function vectors in the local models.

Based on the powerful approximation ability [4], there exists an optimal NFN estimator to approximate the nonlinear dynamic function in (5) such that

$$f(x) = W^* L(x) + \varepsilon(x) \tag{14}$$

With W^* the ideal weight matrix and the estimative error vector $\varepsilon(x) \in \Re^{n \times 1}$ are assumed to be given by

$$\tilde{W} = \arg \min_{\hat{W} \in M_x} \left[\sup_{x \in M_x} \| f(x) - \hat{W} L(x) \| \right], \quad \| \varepsilon \| \leq \varepsilon_N \tag{15}$$

In which $\| \cdot \|$ is the Euclidean norm, M_x and M_w are the predefined compact sets of x and $\hat{W}$, and the positive constant ε_N can be reduced arbitrarily by increasing the number of rules.

4 RANFNC Design

Define the NFN functional estimate by

$$\hat{f}(x) = \hat{W} L(x) \tag{16}$$

With $\hat{W}$ the current values of the NFN weight as provided by the tuning algorithm. With the ideal weights required in (14) define the weight deviations or weight estimation errors as

$$\tilde{W} = W^* - \hat{W} \tag{17}$$

With τ_0 defined to be (7), select the control input

$$\tau = \tau_0 - v = \hat{W} L(x) + K_v r - v \tag{18}$$

$$v = (\varepsilon_N + d_b)sgn(r) = K_f sgn(r) \tag{19}$$

With $v(t)$ a function to be determined to provide robustness in the face of the net reconstruction error ε. Then, the closed-loop filtered error dynamics become

$$M\dot{r} = -(K_v + C(q,\dot{q}))r + \tilde{W}L(x) + (\varepsilon + \tau_d) + v = -(K_v + C(q,\dot{q}))r + \zeta_1 \tag{20}$$

Theorem 1: Consider an n-link robot manipulator represented (1). If the RANFNC law is designed as (18), (19) and the weight update law is designed as (21), then the stability of the proposed RANFNC system can be ensured

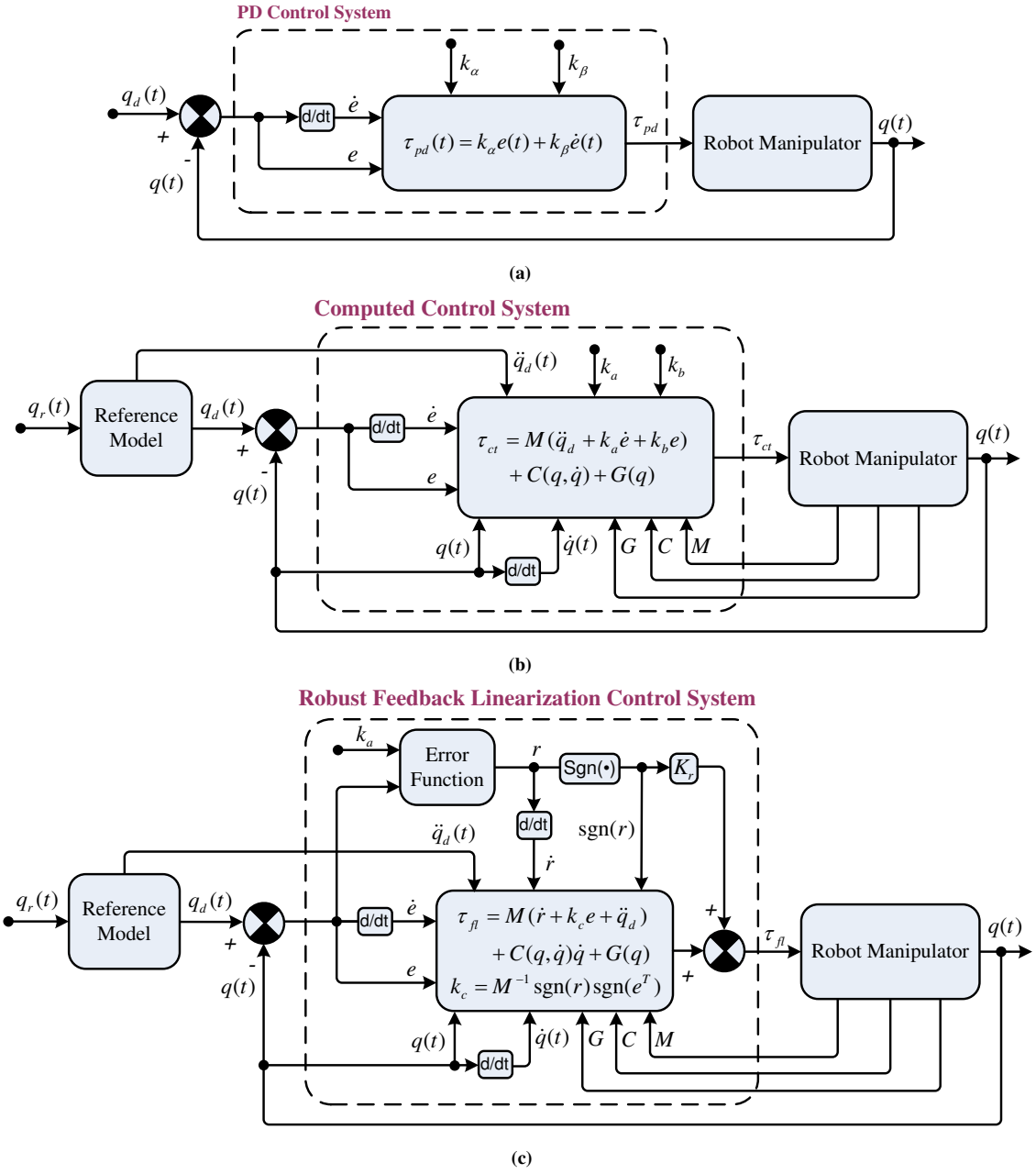


Figure 4: (a) PD control system, (b) computed torque control (CTC) system, (c) robust feedback linearization control (RFLC) system

$$\dot{\hat{W}} = FL(x)r^T \tag{21}$$

Proof: Define a Lyapunov function candidate as

$$V(r(t), \tilde{W}) = \frac{1}{2}r^T Mr + \frac{1}{2}tr(\tilde{W}^T F^{-1}\tilde{W}) \tag{22}$$

Where $tr(\cdot)$ is a trace operator. By differentiating (22) with respect to time and using (19), (20), (21), and using properties of the robot dynamics are introduced in section 2, one can obtain.

$$\begin{aligned}
\dot{V} &= \frac{1}{2}r^T M\dot{r} + \frac{1}{2}r^T \dot{M}r + tr(\tilde{W}^T F^{-1}\tilde{W}) \\
&= -r^T K_v r + \frac{1}{2}r^T(\dot{M} - 2C)r + tr\tilde{W}(F^{-1}\dot{\tilde{W}} + Lr^T) \\
&+ r^T(\varepsilon + \tau_d) - r^T(\varepsilon_N + d_b)sgn(r) \\
&= -r^T K_v r + r^T(\varepsilon + \tau_d) - r^T(\varepsilon_N + d_b)sgn(r) \\
&= -r^T K_v r + r^T(\varepsilon + \tau_d) - \|r\|(\varepsilon_N + d_b) \\
&\leq -r^T K_v r \leq 0
\end{aligned} \tag{23}$$

Since $\dot{V}(r(t), \tilde{W}) \leq 0$, $\dot{V}(r(t), \tilde{W})$ is a negative semidefinite function, i.e. $V(r(t), \tilde{W}) \leq V(r(0), \tilde{W})$. It implies that $r(t)$ and $\tilde{W}$ is bounded functions. Let function $h(t) \equiv r^T K_v r \leq -\dot{V}$ and integrate function $h(t)$ with respect to time

$$\int_0^\tau h(t)d\tau \leq V(r(0), \tilde{W}) - V(r(t), \tilde{W}) \tag{24}$$

Because $V(r(0), \tilde{W})$ is a bounded function, and $V(r(t), \tilde{W})$ is a nonincreasing and bounded function, the following result can be concluded:

$$\lim_{t\to\infty} \int_0^\tau h(t)d\tau < \infty \tag{25}$$

In addition, $\dot{h}(t)$ is bounded; thus, by Barbalats lemma can be shown that $\lim_{t\to\infty} h(t) = 0$. It can imply that $r(t)$ will be converging to zero as time tends to infinite.

5 Numerical Simulation

A two-link robot manipulator as shown in Fig.1 is utilized in this paper to verify the effectiveness of the proposed control scheme. The detailed system parameters of this robot manipulator are given as: link mass $m_1, m_2(kg)$, lengths $l_1, l_2(m)$, angular positions $q_1, q_2(rad)$.

The parameters for the equation of motion (1) are adopted in [11].

$$M(q) = \begin{bmatrix} (m_1 + m_2 l_1^2) & m_2 l_1 l_2(s_1 s_2 + c_1 c_2) \\ m_2 l_1 l_2(s_1 s_2 + c_1 c_2) & m_2 l_2^2 \end{bmatrix}$$

$$C(q, \dot{q}) = m_2 l_1 l_2(c_1 s_2 + s_1 c_2)\begin{bmatrix} 0 & -\dot{q_2} \\ -\dot{q_1} & 0 \end{bmatrix}, \quad G(q) = \begin{bmatrix} -(m_1 + m_2)l_1 g s_1 \\ -m_2 l_2 g s_2 \end{bmatrix}$$

Where $q \in \Re^2$ and the shorthand notations $c_1 = \cos(q_1)$, $c_2 = \cos(q_2)$, $s_1 = \sin(q_1)$ and $s_2 = \sin(q_2)$ are used.

For the convenience of the simulation, the nominal parameters of the robotic system are given as $m_1 = 4.6(kg)$, $m_2 = 2.3(kg)$, $l_1 = 0.5(m)$, $l_2 = 0.2(m)$, $g = 9.8(m/s^2)$ and the initial conditions $q_1(0) = 0.5$, $q_2(0) = 0.5$, $\dot{q}_1(0) = 0$, $\dot{q}_2(0) = 0$. The desired reference trajectories are $q_{d1}(t) = \sin(2t)$, $q_{d2}(t) = \cos(2t)$ respectively.

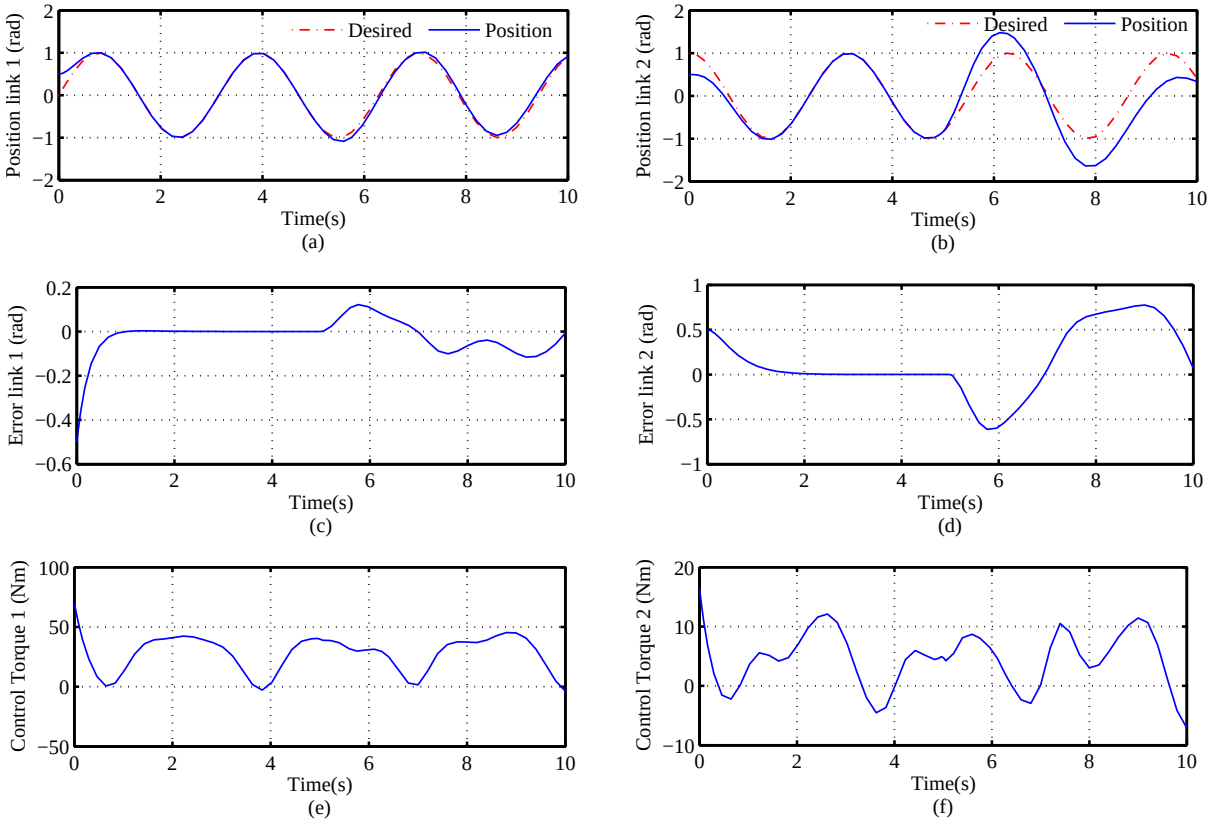


Figure 5: Simulated position responses, tracking errors, and control torques of the CTC control system at joints 1 and 2

The most important parameters that effect the control performance of the robotic system are the external disturbance and the friction term which are denoted $t_l(t)$ and $f(\dot{q})$, in simulation, parameter variation situation and disturbance situation occurring at 5s are considered. The parameter variation situation is that $1(kg)$ weight is added to the mass of link 2, i.e. $m_2 = 3.3(kg)$. The disturbance situation is that external forces are injected into the robotic system, and their shapes are expressed as follows: $t_l(t) = \begin{bmatrix} 5\sin(5t) & 5\cos(5t) \end{bmatrix}^T$. In addition, friction forces are also considered in this simulation and given as: $f(\dot{q}) = \begin{bmatrix} 2\dot{q}_1 + 0.8sgn(\dot{q}_1) & 4\dot{q}_2 + 0.1sgn(\dot{q}_2) \end{bmatrix}^T$.

To this end, the simulation situations are adopted to demonstrate the robust property of the proposed control scheme. In order to exhibit the superior control performance of the proposed RANFNC scheme, three extra control systems including an RFLC system shows in Fig. 4(c), a conventional computed torque control (CTC) and a proportional differential (PD) control are examined in the mean time [13]. Moreover the conventional CTC system as shown in Fig.4(b)can be expressed as

$$\tau_{cd} = M(\ddot{q}_d + k_a\dot{e}(t) + k_b e(t)) + C(q, \dot{q}) + G(q) \tag{26}$$

The PD control system as shown in Fig. 4(a) can be expressed as

$$\tau_{pd} = k_\alpha e(t) + k_\beta \dot{e}(t) \tag{27}$$

The gain in these control system are given as

$$k_\alpha = \begin{bmatrix} 2500 & 0 \\ 0 & 1000 \end{bmatrix}, \quad k_\beta = \begin{bmatrix} 20 & 0 \\ 0 & 25 \end{bmatrix}, \quad k_a = \begin{bmatrix} 6 & 0 \\ 0 & 6 \end{bmatrix}, \quad k_b = \begin{bmatrix} 9 & 0 \\ 0 & 9 \end{bmatrix}, \quad K_v = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix},$$

$$\eta = 20, \quad K_r = d_b = 5. \tag{28}$$

The gain matrices of k_a, and k_b, are determined so that the roots of the characteristic polynomial of $k_a\dot{e} + k_b e$ lie strictly in the open left haft of the complex plane, i.e. $\lim_{t\to\infty} e(t) = 0$. It means that the CTC system shown in (26) is globally asymptotically stable as the root dynamics (1) without the consideration of system uncertainties. However, the stability of the closed-loop control system may be destroyed if the system dynamics are perturbed by external disturbance. However gains k_α and k_β in the PD control are selected according to the Ziegler-Nichols tuning rule. In addition, the selection of learning rates η is dependent on the significance of tuning objects.

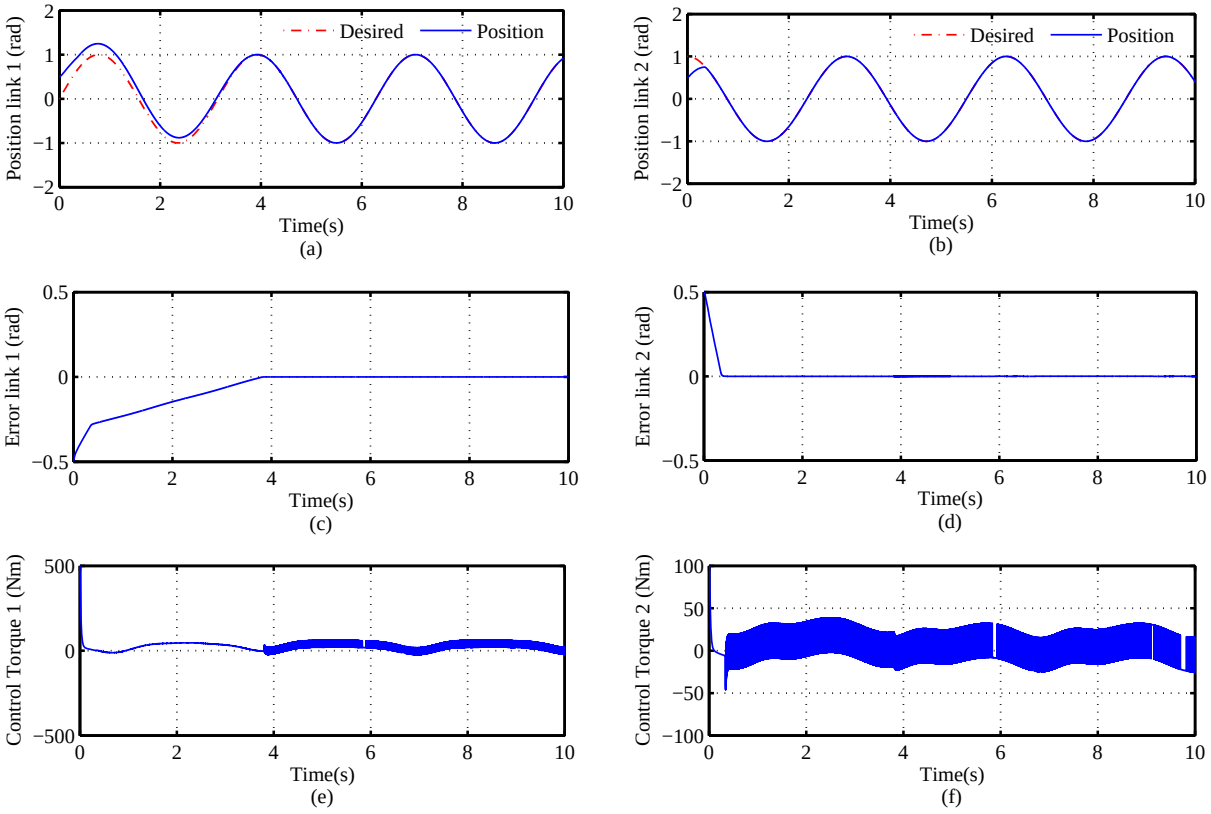


Figure 6: Simulated position responses, tracking errors, and control torques of the RFLC control system at joints 1 and 2

The simulated results of CTC system, the responses of joint position, tracking error and control torque are depicted Fig. 5(a-d), and (e-f), respectively. From the simulated results, in interval from beginning to 5s, favourable tracking responses can only be obtained for the nominal situation. However, since the control gains in (28) are determined without considering the joint

friction and external disturbance. So, poor tracking responses after 5s are resulted due to the occurrence of joint friction and external disturbance. In the RFLC system are depicted in Fig. 6. The joint position responses, tracking error and control torque are depicted in Fig. 6(a-d) and (e-f), respectively. The robust control performance of the RFLC system is obvious under the occurrence of system uncertainties. However, the undesirable chattering phenomenon in the control torque.

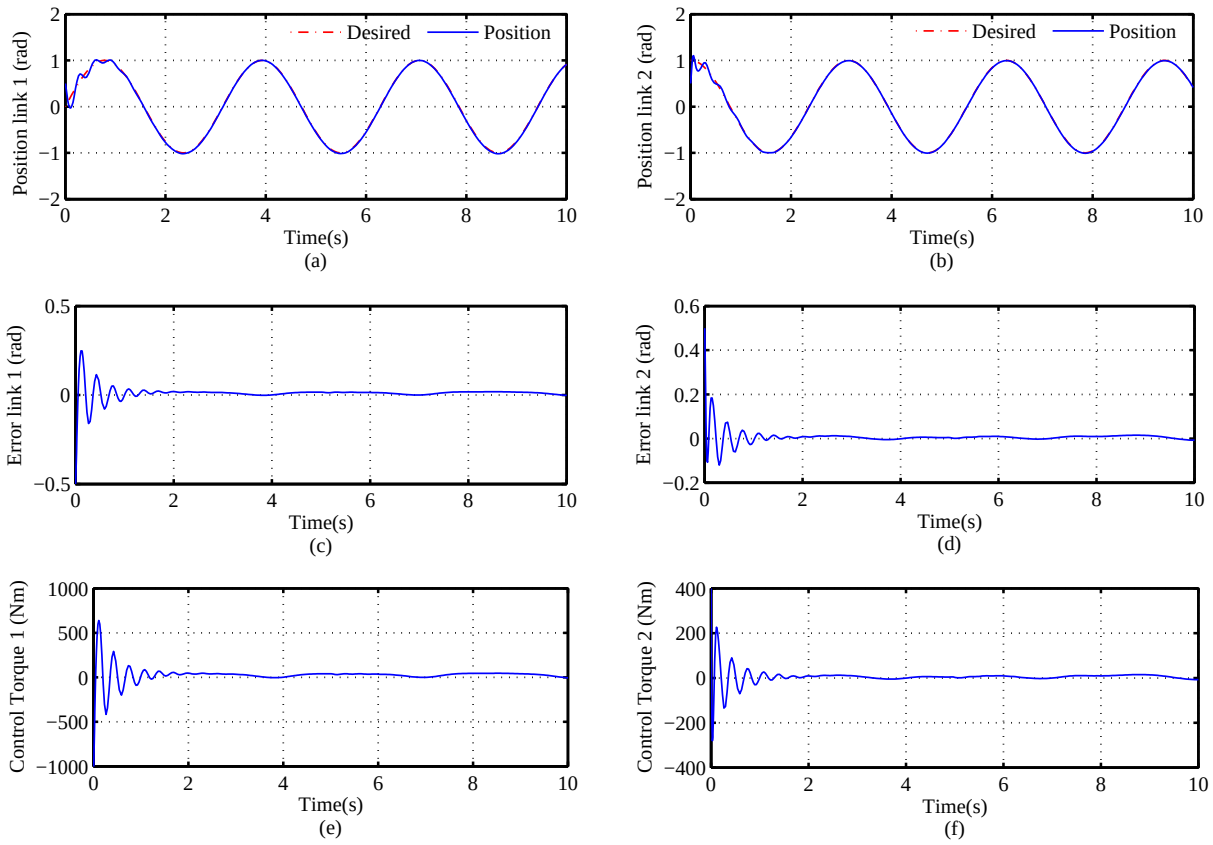


Figure 7: Simulated position responses, tracking errors, and control torques of the PD control system at joints 1 and 2

The PD control system based on model-free design is offered to apply comparable responses to manifest the performance of the RANFNC system. The simulated responses of joint position, tracking error, and control torque, are represented in Fig. 7(a-d) and (e-f). From simulated results see that, the tracking responses are greatly improved and the chattering phenomenon are much reduced.

Now, the proposed RANFNC system depicted in Fig. 3 is applied to control the robot manipulator for comparison. The simulated results of joint position responses, tracking error and control torque are depicted in Fig. 8(a-d) and (e-f), respectively. Because of, all parameters in the NFN are roughly initialized, the tracking errors are gradually decreased though online training process whether the uncertainties exist or not. Furthermore, robust control performance of the RANFNC system, both in the condition of joint friction, parameter variation, and external disturbance are obvious. Compared these results with the CTC, RFLC and PD control systems, the control torque of proposed RANFNC system is not chattering phenomenon.

6 Conclusions

This paper has successfully implemented an RANFNC system to control the joint position of a two-link robot manipulator for achieving desired position control. All the system dynamic could be unknown and no strict constraints. The NFN is used to compensate the uncertainty of the system. All adaptive learning laws in the RANFNC system were derived in the sense of a projection algorithm and Lyapunov theorem so that the network convergence and system-tracking stability of the closed-loop control system can be ensured whether or not the uncertainties occur. Simulated results of a two link robot manipulator via various existing control frameworks including CTC RFLC and PD control were also applied in this paper to compare and display the manipulative performance of the proposed control system. According to the result as depict in Figs. 5-8, the desired position tracking response of the RANFNC system can be controlled closely follow specific reference trajectories under wide range of disturbance. The main of the paper is to construct a simpler and more efficient intelligent control system without dynamic knowledge of plant. While ensuring the convergence and tracking stability of the closed-loop system. The proposed RANFNC system can also be applied to other systems, such as mobile robotic, AC servo system and so on.

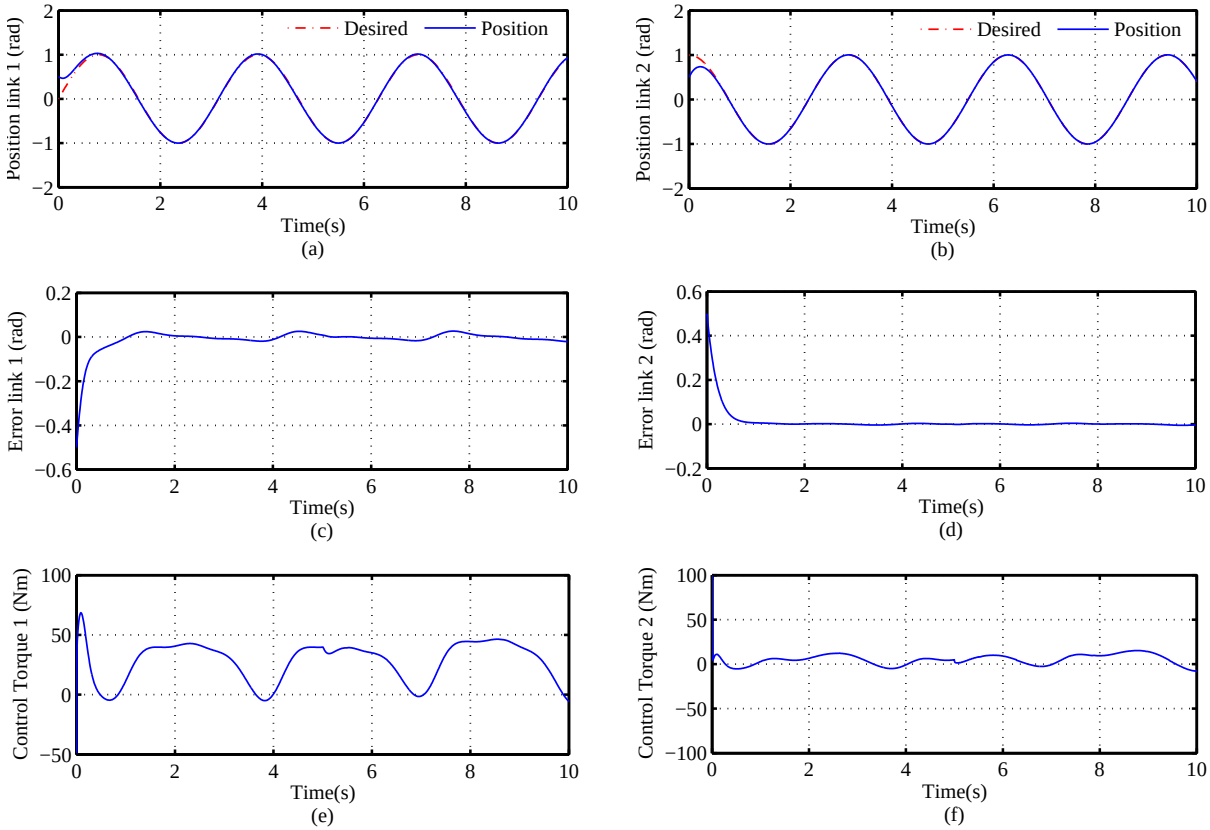


Figure 8: Simulated position responses, tracking errors, and control torques of the RANFNC control system at joints 1 and 2

7 Acknowledgment

This work was supported by the National Natural Science Foundation of China (60775047; 60835004), the National High Technology Research and Development Program of China (863

Program) (2007AA04Z244; 2008AA04Z214). The authors would like to thank the associate editor and the reviewers for their valuable comments.

Bibliography

[1] Jinzhu Peng, Yaonan Wang, Wei Sun, Yan Liu, A neural network sliding mode controller with application to robotic manipulator, *IEEE Conf. Int. Control*, 1:2011-2015, 2000

[2] B.K. Yoo and W.C Ham, Adaptive control of robot manipulator using fuzzy compensator, *IEEE Trans. Ind. Electron*, 8(2):186-199, 2000

[3] Shuzhi S. Ge, Adaptive neural network control of robot manipulator in task space, *IEEE Trans. Ind. Electron*, 44(6):746-752, 1997

[4] C.T. Lin and C.S. George Lee, Neural Fuzzy Systems, *Englewood Cliffs*, Prentice-Hall, 1996

[5] Y.Q. Zhang and A. Kandel, Compensatory neural-fuzzy systems with fast learning algorithms, *IEEE Trans. Neural Newt*, 9(1):83-105, 1998

[6] Vesselenyi T., Dzitac S., Dzitac I., Manolescu M.-J., Fuzzy and Neural Controllers for a Pneumatic Actuator, NT J COMPUT COMMUN, ISSN 1841-9836. 2(4): 375-387, 2007

[7] Alavandar S., Nigam M.J., Neuro-Fuzzy based Approach for Inverse Kinematics Solution of Industrial Robot Manipulators, INT J COMPUT COMMUN, ISSN 1841-9836, 3(3):224-234, 2008

[8] AlavandarS., Nigam M.J., Inverse Kinematics Solution of 3DOF Planar Robot using ANFIS, NT J COMPUT COMMUN, ISSN 1841-9836, 3(S):150-155, 2008

[9] L.X. Wang, A course in Fuzzy Systems and Control, *Englewood Cliffs*, NJ:Prentice Hall, 1997

[10] O. Omidvar and D.L. Elliott, Neural Systems for Control, *Englewood Cliffs*, NJ: Prentice-Hall, 1997

[11] B.S. Chen, H.J. Uang, and C.S. Tseng, Robust tracking enhancement of robot systems including motor dynamics: A fuzzy-based dynamic game approach, *IEEE Trans. Fuzzy syst*, 11(4):538-553, 1998

[12] R.J. Schilling, Fundamentals of Robotics, *Analysis and control. Hoboken*, NJ: Prentice-Hall, 1998

[13] J.J.E. Slotime and W. Li, Applied Nonlinear Control, *Hoboken*, NJ: Prentice-Hall, 1991

[14] H. K. Khalil, Nonlinear Systems. Englewood Cliffs, *Englewood Cliffs*, NJ: Prentice-Hall, 1996

[15] K. Liu and F.L. Lewis, Robust control techniques for general dynamic system, *J. intell. Robotic syst*, 6:33-49, 1992

[16] F.L. Lewis, C.T. Abdallah, and D.M Dawson, Control of Robot Manipulators, *New York: Macmillan*, 1993

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 2 (June), pp. 353-364

Advance and Immediate Request Admission: A Preemptable Service Definition for Bandwidth Brokers

I.T. Okumus, F.U. Dizdar

Ibrahim Taner Okumus
K.Maras Sutcu Imam University, Computer Engineering Department
Avsar Kampusu, 46000, K.Maras, Turkey
okumus@mu.edu.tr

Ferhat Umut Dizdar
Mugla University, Graduate School of Science
Kotekli Kampusu, 48000, Mugla, Turkey

Abstract: Differentiated Services Architecture lacks control level functionalities and Bandwidth Brokers are proposed to fill that gap. In order to provide proper control level functionalities, Bandwidth Brokers need to provide services for both advance requests and immediate requests. There is a tradeoff between preemption of immediate flows and utilization of links. It is important for a resource manager to provide the promised QoS level to a flow without any preemption. In this study, we solve the preemption and the experienced QoS problem by defining a preemptable service and explain how this service works and also show the performance and scalability characteristics of resource manager with the addition of a preemptable service.
Keywords: admission control, IR Flows, AR flows,preemptable forwarding service, quality of services.

1 Introduction

Differentiated Services (Diffserv) Architecture [1] is the de facto standard to provide QoS in IP networks. However Diffserv only specifies the data level functionalities. Control level functionalities are left out from the architectural definition.There are various options to provide control level functionalities of Diffserv. One of the solutions is named Bandwidth Broker (BB) [2], which is essentially an intra-domain resource manager (IDRM) [3] in a Diffserv domain. This manager has admission control, authentication, authorization, inter-BB communication and resource management duties in a Diffserv network.

In the Internet, different applications generate traffic flows. Depending on the application, different flows have different QoS needs. Diffserv architecture defines QoS classes to address the needs of different flows [4] [5]. Again depending on the application, actual traffic generation time will be different. Some applications generate traffic right away and ask for resources to be used immediately. Some applications make advance requests for a traffic that will be generated in the future.

In order to provide QoS for all kinds of applications, IDRM needs to provide service for both advance requests (AR) and immediate requests (IR). In this study, to provide proper QoS for both type of requests we propose a new service called Preemptable Forwarding (PF). We present the details of the architecture with the new PF service and analyze the performance and scalability characteristics of the approach.

Rest of the paper is organized as follows. In the following section, we summarize the previous work in this subject. Section 3 gives the details of the IDRM and also the definition of a PF service. Section 4 provides the analysis results. Comparison of the results with earlier work is given in section 5. We give the concluding remarks in Section 6.

2 Related Work

One of the earlier works in the area of resource planning in advance request agents is provided by Schelen and Pink [6]. In this study, authors tried to reduce the preemption ratio of the IR flows. Authors experimented with different look-ahead times to analyze its effect on preemption rate of IR flows. The look-ahead time is kept constant in time (CLAT). Authors show that when look-ahead time increases, the preemption rate decreases but utilization also decreases.

Lin et.al. [7] proposed to change the look-ahead time depending on the application. In this scheme, authors use prediction methods to calculate the holding time of an IR flow and use this time as the look-ahead time. This scheme requires sensitive prediction techniques which brings high cost to the computing entity.

Ahmad et.al. [8] proposes to use dynamic look-ahead time to solve the problem. Authors argue that look-ahead time is dependent on resource scarcity, IR arrival and release rate, and bandwidth release per IR call. Authors calculate the LAT dynamically (DLAT) by considering these parameters at the time of the calculation. When there is a need for preemption, IR calls that are accepted most recent time are preempted to reduce wasted throughput. However size of these flows is also important. Preempting a recently accepted giant flow will cost more to the network than preempting an old small flow. Also when the AR limit is low and the LAT is high, DLAT model exhibits comparable results with CLAT model.

Degermark et.al. [9] used a model where all flows declare their duration, and admission is based on measurements and future predictions of the traffic load. In this scheme none of the flows are preempted. AR flows are expected to provide the duration of the flow. However assuming all IR calls to provide their duration is not realistic in today's networks.

Greenberg, Srikant, and Whitt [10] proposed a probabilistic approach for AR admission control. In this method, an AR call is admitted based on the call interruption probability. If the probability is below a threshold, request is admitted. This method assumes all AR calls are made for far ahead in time. This scheme allows occasional service disruptions to achieve higher utilization.

Srikant and Whitt [11] used CLT approximation to calculate interrupt probability for resource sharing between AR and IR flows. This method allows multi-class resource scheduling.

Karsten et.al. [12] proposed a policy-based service specification for advance resource reservations. In the study, an IR request is assumed to be nonpreemptable for certain amount of time and preemptable after that duration. During the admission process, IR requests provide the nonpreemptable duration. Also it is possible for a flow to request nonpreemptable for the whole lifetime. However, it is not realistic for a flow to require certain QoS for a limited time and not need that QoS after that. Applications require certain QoS for the whole lifetime. If all IR requests are nonpreemptable, then the network utilization is low. This study does not provide any evaluation result for the suggested scheme.

In order to provide an acceptable service for both AR and IR calls it is important to have both flows accepted in the network and also to have a sensitive preemption scheme to reduce preemption rate of IR flows. When providing QoS, if an IDRM accepts a flow into the network, whether it is AR or IR, flow expects to get the promised QoS through the duration of the flow. Unaware preemption should not be an option in any case. None of the proposed solutions consider user satisfaction and experienced-QoS in their study. In the next section we provide the solution we propose for this problem.

3 Proposed Model: Preemptable Forwarding

In our work, we used IDRM architecture as a resource manager [3]. IDRM is mainly a specialized bandwidth broker. IDRM is responsible for admission control and intra-domain resource management along with other tasks. These two tasks are the important ones for our study.

In order to better manage intra-domain resources, IDRM knows intra-AS topology. IDRM keeps track of the available capacity of individual links in the network. IDRM also keeps track of the currently reserved resources. In this context, to support Advanced Requests (AR), IDRM needs to keep records of start and end times and required capacity of advanced reservation flows. In bandwidth database, IDRM keeps current load of the links and differentiates immediate and advance reserved shares of the bandwidth (Table 1).

Table 1: A sample database state

LinkID	UtilizedBW	ReservedBW
40	5	7
41	5	5

Table 2: Time slot table for reservation states

LinkID	40								
Time Slot	1	2	3	4	5	6	7	8	9
Reserved BW	1	2	4	2	0	1	5	4	0

IDRM also keeps reservation database for AR requests. For the admission of advance requests, IDRM needs to know the available capacity and the reserved capacity of a future time interval. Since time is an analog entity, it is impossible to keep track of the link states in every point in time. Usual way to make time more manageable is to quantify it. We call these quantified time intervals *time slots*. IDRM keeps track of reserved capacity of every time slot through the future. To prevent the scalability problems, the number of time slots into the future needs to be finite. The actual amount of a time slot can be determined according to the network's need. It can be seconds, minutes, hours. A sample state of reservation database is shown in Table 2.

3.1 Admission Control

Admission control methods for IR, AR and PF requests are different. The goal is to reduce the preemption rate of IR flows and to increase the overall throughput as much as possible. In order to achieve this goal, we need to keep a balance between IR and AR flows. If we only accept IR flows, admission control decision is simply based on current available capacity on the path of the flow. Determination of this capacity can be parameter-based or measurement-based which is out of scope of this paper. If we only accept AR flows, then the admission decision is based on the available capacity from the start time to the end time of the request. When we have both IR and AR together, we need to use a mixture of these two admission control methods.

We define LC as the total link capacity, IRLC as the link capacity share of IR flows, CIR_i as the capacity of the i_{th} active IR flow,CAR_i as the capacity of the i_{th} active AR flow, $CIR(t_0)$ as the total IR capacity used by IR flows at time t_0, $CAR(t_0)$ as the total AR capacity used by AR flows at time t_0. $CIR(t_0)$, and $CAR(t_0)$ is calculated as follows:

$$CIR(t_0) = \sum_{i=0}^{n_1} CIR_i, i = 0, 1, ..., n_1 \tag{1}$$

$$CAR(t_0) = \sum_{i=0}^{n_2} CAR_i, i = 0, 1, ..., n_2 \tag{2}$$

Residual capacity at time $t_0$for AR and IR share is calculated as follows:

$$IRLC_r(t_0) = IRLC - CIR(t_0) \tag{3}$$

$$ARLC_r(t_0) = ARLC - CAR(t_0) \tag{4}$$

Admission Control of IR Flows

Considering the preemption and throughput tradeoff, there can be two different admission control approaches for IR admission. In first case, to increase the throughput, IR flows can be allowed to overflow into the AR share of the link capacity. However some IR flows can be preempted if necessary. In the second case, to prevent preemption, IR flows can be limited into their own share of the capacity and not allowed to use the excess AR capacity.

Admission of the first case is based on checking the available capacity on the whole link. If there is enough available capacity at the time of the request, then the IR flow is accepted:

$$c < IRLC_r(t_0) + ARLC_r(t_0) \tag{5}$$

Admission of the second case is based on checking available capacity only on the IR share. If there is enough capacity IR flow is accepted:

$$c < IRLC_r(t_0) \tag{6}$$

Admission Control of AR Flows

Admission control criterion for AR flows is different. AR flows always have priority over IR flows. Admission control decision does not take active IR flow capacity into account. Decision is based on the AR reservations in the duration of the requested flow. If AR flow starts at t_1 and ends at t_2, condition to accept the flow is:

$$ARLC > c + max[C_{AR}(\tau)], \tau = [t_1, t_2] \tag{7}$$

Where ARLC stands for AR Link Capacity share, $max[C_{AR}(\tau)]$ shows the maximum reserved AR capacity in the interval τ. If this condition is satisfied, AR flow is accepted. At time t_1 AR flow will start. It is possible that in between the time of accepting the AR flow and the actual AR flow start time, some IR flows could be accepted. During interval τ, some of the IR flows can be preempted to provide capacity to the pre-accepted AR flows.

Preemptable Forwarding Service

This requirement led us to define a new service type. As we will show in our analysis results, in order to get the most benefit from the network while having both IR and AR flows together, there will be a need to preempt IR flows. Instead of selecting a flow without the consent of the owner, we define a new service called Preemptable Forwarding (PF). In this service type, some IR calls accept the probability of preemption beforehand during the admission control process. From hereon, we refer this flow type as PFIR. For PFIR flows, there is a chance that accepted flow will never be preempted. In this case the experienced QoS will be the same as other IR and AR flows. However if there is a need for preemption, network will select one of the PFIR flows to preempt. None of the regular IR flows will be preempted. Since the user is expecting this preemption and has consented to preemption beforehand, there will not be any user dissatisfaction in terms of the experienced QoS.

This service type is most suitable in cases where there is a limit on both IR and AR flows. Bandwidth capacity is divided between IR & AR flows and none of these flows can overflow

to other flow's share. If there is no capacity available in IR share for a new IR request, PF admission is used.PF flows will be using the excess capacity in the AR share of the link capacity. The amount that can be used by PF flows can be determined beforehand and can be static or dynamic.

Admission Control of PFIR Flows

The admission control decision for PFIR depends on the IR capacity, AR capacity and PF share. If available capacity on the IR share is not enough to accept an IR request, user is prompted to retry for PFIR. PFIR request is accepted only if the residual IR capacity and residual PF share are in total bigger than the requested capacity. Residual PF capacity $PFLC_r$ is dependent on the look-ahead time (LAT), AR reservations in that time interval and the active PFIR flows. Residual AR capacity ($ARLC_r$) is defined as the capacity available on the ARLC and is defined as:

$$ARLC_r = ARLC - max[C_{AR}(LAT)] \tag{8}$$

At time t_0, total PFIR capacity in use is calculated as:

$$PFIR(t_0) = \sum_{i=0}^{n_3} PFIR_i, i = 0, 1, ..., n_3 \tag{9}$$

Residual PF capacity at time t_0 is then defined as:

$$PFLC_r(t_0) = \begin{cases} PFLC - PFIR(t_0) & ARLC_r \geq PFLC \\ ARLC_r - PFIR(t_0) & ARLC_r < PFLC \end{cases} \tag{10}$$

PF admission is based on the residual IR and residual PF capacities:

```
if (c < IRLCr + PFLCr)
 then accept
else reject
```

In order to clarify the admission of PFIR flows, lets take a look at the admission process. An IR request will arrive at the BB via a signaling protocol such as SIBBS [14]. BB will determine the type of the request (IR or AR). If the request is IR, BB performs IR admission procedure. If IR capacity is not enough to accept the request, BB sends back a negative message with PF admission option. This message suggests a PF request and also contains an indicator about the preemption possibility. We define $PFRate(t_0)$ as the PF share usage rate at time t_0:

$$PFRate(t_0) = 1 - \frac{PFLC_r(t_0)}{PFLC}, 0 \leq PFRate(t_0) \leq 1 \tag{11}$$

Average PF rate is calculated using moving average to take into account the previous trends on the PF usage rate. Average PF rate takes values between 0 and 1. 0 indicates that no PF flow is active and PF capacity is available. 1 indicates PF capacity is fully used. Values of variables a and b can be set based on the network needs. In our study, we used values a=0.2 and b=0.8. Average PF rate is calculated as follows:

$$AvgPFRate(t_0) = a * PFRate(t_0 - 1) + b * PFRate(t_0), 0 \leq a, b \leq 1, a + b = 1 \tag{12}$$

Average PF rate shows the possibility of preemption of a PFIR flow in LAT time window. Based on this indicator, user can predict preemption possibility and decide whether or not to use the

PF service. After receiving the negative response and average PF rate, request can be submitted as a PF request.

After PFIR flow is admitted there is a possibility that some AR flows can request a reservation and reclaim the bandwidth currently used by PFIR flows. Then some PFIR flows will be preempted. In this case we propose to preempt latest PFIR flow. A detailed loss-benefit analysis needs to be made on the preemption choices to develop a method to choose the flow to be preempted that will have least harm on the network benefit. However, this study is out of scope of this paper.

4 Simulation Results

Our analysis mainly covers the effects of the admission control methods on throughput and preemption rate. Along with these, to measure the scalability we define benefit and processing load as other measures. 1 Unit benefit is defined as the amount of income network gets from a 1Mbps of flow in 1 sec. If a 100Mbps link is fully utilized for 20 seconds, benefit is 100*20=2000 units. Processing load is the amount of work required to process an incoming request. In order to test this, we calculate the number of accesses to databases by IDRM to produce a response to a request. If an IDRM accesses bandwidth table 2 times and time slot table 4 times that requests' processing load is 6 units.

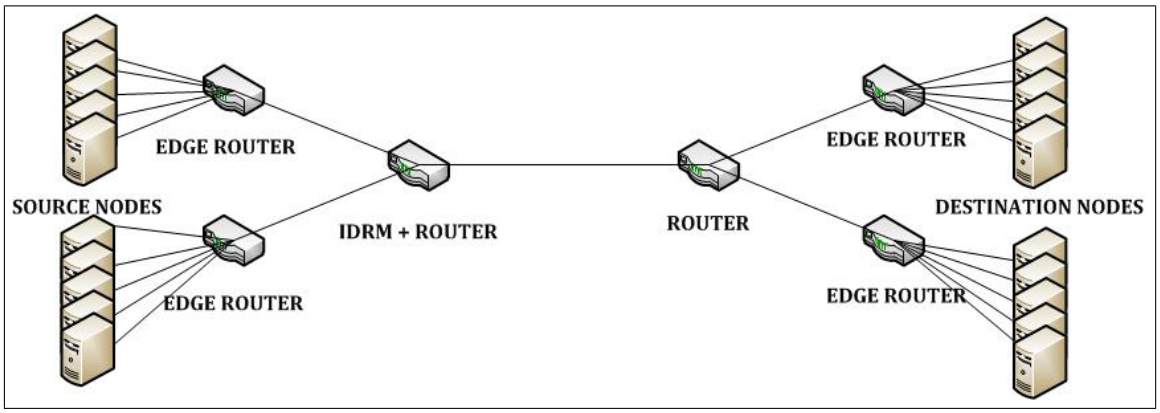


Figure 1: Simulation Topology

IDRM is implemented on ns2 simulation software [15]. We used the topology shown in Figure 1 in our simulation. In this topology, there are 10 source nodes and 10 destination nodes. All the link capacities are 100Mbps. The link between *IDRM+router* and *router* is the main bottleneck link. We assume that all flows are unidirectional from sources to destinations. Flows and reservation requests are entered into the domain from edge routers. Edge routers forward reservation requests to IDRM. IDRM applies admission control methods and sends back an accept/reject answer to the source.

The incoming reservation requests follow Poisson distribution with a mean arrival of 10 requests per time slot. Time slots are used as the time measure and on average 10 request is produced in a single time slot by all the traffic sources. IR & AR limit of these requests are varied depending on the scenario. Every request asks for 1Mbps capacity from the network and flows last for 20 time slots. In all the simulation scenarios, there is a warmup time. Because of the warmup time, first 100 time slots is excluded from evaluation. All results indicate the steady state case.

In order to determine the network behavior and use it as a benchmark for other results, we first tested the network with only IR and with only AR flows.

4.1 IR-only and AR-only flows

In IR-only scenario, sources only make IR requests from the network. There is no limit on the IR flows. These flows can consume the whole link capacity. During the simulation, links are saturated and the network behavior under full capacity is observed. In this case, 29.24% of IR are rejected because of unavailable capacity. Maximum benefit is 20000 units and achieved benefit for this scenario is 19627 units. Average utilization is calculated as 98.13%. Average process-time is 34.17 units and total process time is 70577 units.

In the AR-only case, sources make only AR requests from the network. No limit is imposed on AR flows in terms of the bandwidth use. In this scenario, 38.51% of AR requests are rejected. The difference between IR-only and AR-only case is that, AR-only case works with time slots. That means flows can start only at the beginning of a time slot. However, IR flows can start anytime. This is one of the reasons for high rejection ratio of AR flows. Total network utilization in this scenario is on average 90.73%. Network gained 18146 units of benefit out of 20000 maximum benefit. Total process-time for AR-only case is 166327 units.

Comparing the two cases, AR flows achieve lower network utilization because of the scheduling problems and the slot timing. AR flows achieve lower benefits and also higher process times compared to IR flows.

4.2 IR & AR together, No limit on AR

In this scenario, IR and AR flows appear together in the network. 70% of the requests are IR requests and 30% of the requests are AR requests. Since the holding time for flows is 20 time slots and each flow consumes 1Mbps of bandwidth, resources are saturated in 10 time slots.

Results show that network utilization is 96.62% on average. In case of congestion, IR flows are preempted. 44.2% of the total accepted flows are IR flows and 55.8% are AR flows. 67.9% of IR flows and 0.35% of AR flows are rejected. During the simulation 4.66% of accepted IR flows are preempted. In this IR&AR-no-limit case, when both flows are active in the network, utilization is lower compared to the IR-only case, but it is higher than the AR-only case.

4.3 IR & AR, AR Limited

In this scenario, our goal is to analyze the effect of limiting the AR flows on network utilization. To determine this effect we set the limit of AR flows to 50% in the first simulation and reduced it by 10% for each successive runs.

AR flows can use the limited capacity reserved for them. If there is not enough capacity, then AR flows are rejected. However, IR flows do not have any limit. If there is available capacity in the IR part, flow is accepted immediately. IR flows are preempted in the future when an AR flow asks for the capacity that is being overused by IR flows. Figure 2 shows AR share and accept/reject ratios for both IR and AR flows. Figure 3 shows IR and AR throughput and total throughput for AR-limited scenario. As it can be seen from the figure, as the AR percentage drops, total throughput increases. Network utilization changes from 97.23% for 50% AR share to 98.26% for 10% AR share.

Figure 4 shows the IR drop rate. IR flows are preempted as AR flows move in for their share of the capacity. Figure clearly shows that IR drop rate increases as the AR share increases in the network.

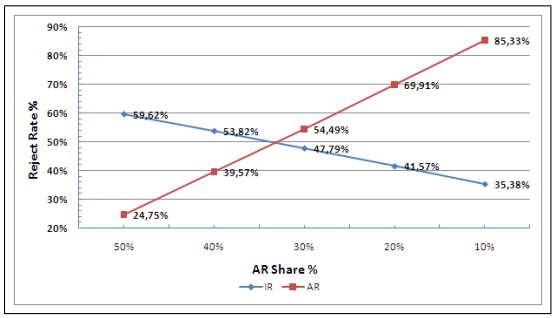


Figure 2: Accept/Reject Rates of IR & AR flows for different limits on AR flows

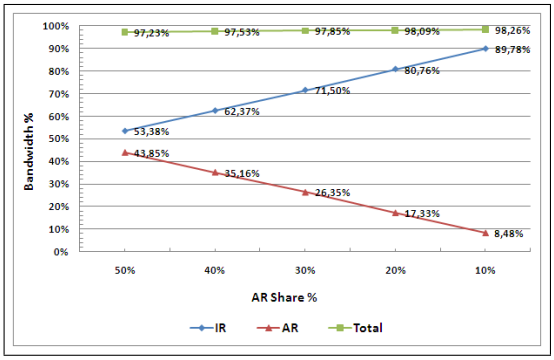


Figure 3: IR, AR bandwidth shares and total throughput for AR limited case.

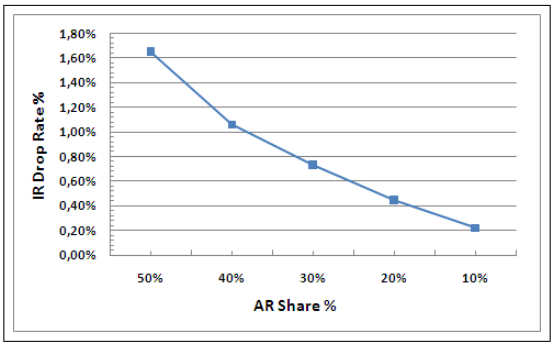


Figure 4: IR Drop Rate for AR limited case

Table 3 shows processing time shares of IR and AR flows for different limits on AR flows. Results show that as the AR limit decreases, total processing time also decreases. This is in accordance with the fact that admittance of AR flows is more costly than admittance of IR flows.

Table 3: Process times for AR limited case

Limit	IRAvg	IRTotal	ARAvg	ARTotal	TOTAL
50%	23	159756	99	279883	439639
40%	25	174704	83	236285	410989
30%	27	190244	66	187944	378188
20%	29	206265	49	140279	346544
10%	32	222212	28	80077	302289

Table 4: Process times for both limited case

Limit(%)	IRAvg	IRTotal	ARAvg	ARTotal	TOTAL
50-50%	21	146917	99	279883	426800
60-40%	23	165047	83	236285	401332
70-30%	26	183177	66	187944	371121
80-20%	29	201307	49	140279	341586
90-10%	31	219347	28	80077	299424

4.4 IR & AR, Both Limited

As we can see from the AR-limited scenario, some IR flows are preempted, which is not desired in a QoS environment. To prevent that, one option is to limit both flows. In this scenario we divide the total capacity between these two flows. We do not allow any of the flow types to use the capacity from the other flow types' share. As in previous case, highest capacity reserved for AR flow is 50%. This limit is reduced by 10% for each simulation run and IR limit is increased by 10% to analyze the effect of the amount of share each flow type gets from the network. We start with 50-50 (IR-AR), then change to 60-40, 70-30 and so on for each successive runs.

Figure 5 shows accept/reject ratios for both limited case. Compared to previous scenarios, IR accept-rate is reduced. The reason is that after using the capacity reserved for IR flows, all IR requests are rejected.

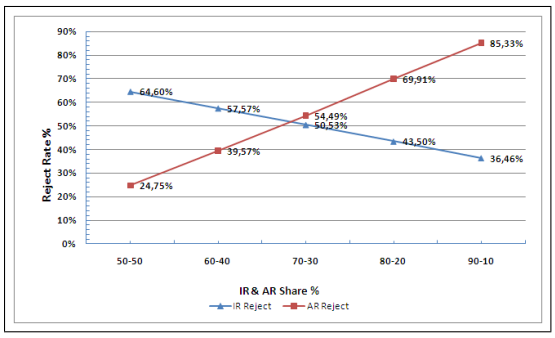


Figure 5: Accept/Reject Rates of IR & AR flows for different limits on both flows

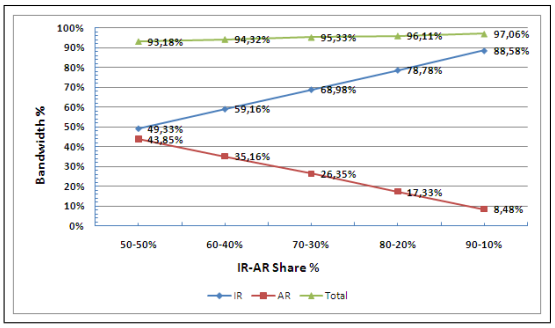


Figure 6: IR, AR shares and Throughput for both limited case

Figure 6 shows IR-AR shares and total throughput for both-limited case. In this scenario, throughput is decreased significantly compared to the base cases and to the AR-limited case. For 50-50 case, throughput is 93.18% and for 90-10 case throughput is 97.06%.

Table 4 shows process times for both limited case. Compared to the AR-limited case, AR process times are same; however IR process time is decreased because less IR flows are accepted into the network.

In this scenario, no IR flows are dropped. Since both flows have their own share and they are bound to it, preemption is not necessary. However, this is achieved by sacrificing from total throughput. This result also confirms that there is a clear tradeoff between throughput and the preemption of IR flows.

Since the main goal of the network provider is to get the maximum benefit from the network, which can be achieved by maximum utilization of the network capacity, we need to find a balance between the preemption and the total throughput. If preemption is inevitable, then the question is which flows should be preempted. In our study we prefer to define a new service called preemptable forwarding (PF). IR flows that are accepted as PF flows will be preempted in case of an overuse.

4.5 Preemptable Forwarding Service with Lookahead Time

In our scheme, in order to balance the throughput and preemption we select to limit the capacities of both IR and AR flows. As we showed in section 4.4, this scheme has low throughput but no IR drops. In order to increase the throughput we allow certain percentage of AR capacity to be used by PF flows if necessary.

We analyzed this scenario on the same topology with the same parameters. In the analysis, we have two parameters to consider. First one is the LAT. How LAT affects the throughput? Second parameter is the PF percentage in the AR region. We analyzed the effect of PF percentage on the throughput. In the simulations we set the IR-AR ratio to 50-50 and changed the PF share from 1% to 15% of the total link capacity and also changed LAT from 1 time slot to 15 time slots.

Figure 7 shows IR lost benefits because of preemption for different PF shares. For low LAT values, IR drop rate is higher for high PF shares. As LAT value increases, lost benefit for all PF shares converge to zero.

Figure 8 shows the total throughput of the network for different LAT values and PF shares. This graph shows that total achieved throughput is higher for lower LAT values and higher PF shares. Again these results show the clear tradeoff between IR drops and the total throughput.

Table 5 gives the utilization values for different PF share and LAT values. Table 6 shows

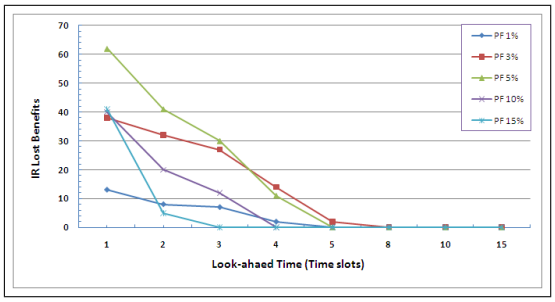


Figure 7: IR Lost benefits

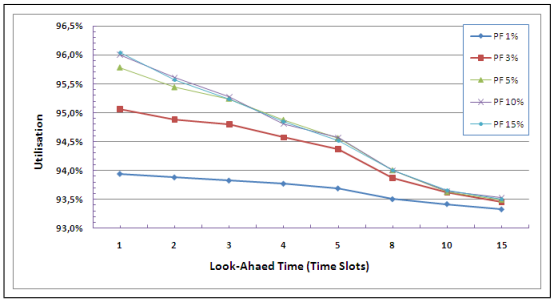


Figure 8: Total throughput

Lost IR benefits for different LAT values and PF shares.

Table 5: Throughput values for different LAT values and PF shares

PF(%)	1	3	5	10	15
LAT	Utilization (%)				
1	93.94	95.07	95.78	96	96.04
2	93.88	94.88	95.44	95.61	95.57
3	93.825	94.80	95.24	95.27	95.23
4	93.77	94.57	94.88	94.81	94.85
5	93.69	94.37	94.56	94.57	94.52
8	93.50	93.88	94.01	94	94.01
10	93.42	93.62	93.63	93.64	93.66
15	93.33	93.46	93.50	93.53	93.5

Table 6: Lost IR benefits for different LAT values and PF shares

PF(%)	1	3	5	10	15
LAT	Lost IR Benefits				
1	13	38	62	40	41
2	8	32	41	20	5
3	7	27	30	12	0
4	2	14	11	0	0
5	0	2	0	0	0
8	0	0	0	0	0
10	0	0	0	0	0
15	0	0	0	0	0

Results on Table 5 and 6 suggest that for 50-50 IR-AR share, best option is to set PF share at 15% and LAT value at 3 time slots. These values give the highest throughput on the network without any preemption. Highest utilization that can be achieved in this scheme is 96.035%. This value is considerably higher than the 50-50 IR-AR both limited case, which has 93.18% utilization.

Processing time for a PF request consist of two components. A PF request is originally an IR request. When there is lack of resources on IR share, request is rejected for resubmission as a PFIR request. PFIR admission is similar to an AR admission. Residual AR capacity is calculated and decision is made depending on that value. So the process is a two pass process. First pass results in rejection of an IR request. Second pass is PF admission. In terms of the processing time, admission of a PFIR request is more costly than both admission of an IR request and also admission of an AR request for the same LAT value. Regular AR admission has a constant LAT value. PFIR processing time increases with the increase of LAT. Table 7 shows average processing times of PFIR requests for different LAT values.

Table 7: Average Processing times for different LAT values

LAT	1	2	3	4	5	8	10	15
Avg. Process Time	82	84	85	87	90	97	103	110

5 Comparison of the Results

Ahmad et.al. [8] proposed a dynamic lookahead time (DLAT) solution for reduce preemption rate at the cost of lower throughput. This study provides results for DLAT and compared the results with the study conducted by Schelen and Pink [6]. Schelen and Pink used constant lookahead time (CLAT) in their study. We will also compare our findings with DLAT and CLAT models.

In our PF model, we used 50%IR and 50%AR share. We compare these results with the same AR limit case in Schelen and Ahmad study. When AR limit is set to 50%, highest normalized throughput achieved by DLAT model is 0.84 with DLAT c=1.0 and highest throughput achieved by CLAT is 0.85 with CLAT 30. These LAT values corresponds to LAT value 2 in our study. With LAT=2, highest utilization in our case is 95.6% (PF %10) and lowest utilization is 93.8% (PF %1). This shows that our scheme performs better than both schemes in terms of the network utilization.

In terms of the preemption rate, our results show the number of preempted flows, while DLAT and CLAT models show preemption probability. We will compare the preemption trends with those studies. Our results indicates that it is possible to prevent preemption by selecting appropriate LAT and PF share values. Also while longer LAT values result in lower preemption, higher LAT values result in higher preemption. DLAT and CLAT models also show the similar trends. Longer DLAT and CLAT values result in lower preemption probability.

As a summary, DLAT model achieves lowest throughput than CLAT and our PF model. PF model performs best in terms of the total throughput. Preventing preemption is necessary in order not to disrupt the QoS of the accepted flows. DLAT model provides this with the cost of throughput. Our model suggest a PF scheme where user pre-agrees a preemption when and if necessary. Also PF scheme can provide low preemption ratios by selecting LAT and PF% values properly. Our proposed scheme results in higher utilization and also in higher user satisfaction in terms of the perceived QoS.

6 Conclusion

In this study we analyzed the admission control method of an IDRM that supports both IR and AR flow types.The main tradeoff in this environment is between preemption of IR flows and the total throughput of the network. As the throughput increases, IR preemption also increases. If we decrease IR drop rate, throughput decreases. Another issue is user satisfaction due to IR drops and also the scheme to select which IR flows to drop in case of a capacity problem.

In order to increase the perceived QoS on the user side, instead of selecting an IR flows among the ones that the network accepted and promised to provide a certain QoS without any interruption, we propose to employ a new QoS class called Preemptable Forwarding (PF). This flow type will be accepted to the network with the condition that the flow will be preempted and not get the promised QoS in case of congestion. Users will be accepting the service with the possibility of a preemption or not get any QoS at all.

We analyzed the effects of the PF share from the total capacity on the IR preemption and the total throughput. As the PF rate increases, total throughput also increases. However, increased PF also causes high IR drop rate.

When we employ lookahead time (LAT) before accepting the PF requests, behavior changes. With high LAT values, IR drop rate is reduced. However, this causes the network utilization to decrease.

Results show that employing a resource manager that uses PF service in admission control of the flows can increase the total throughput and also the user satisfaction in a QoS network.

Bibliography

[1] Blake, S. et.al., An Architecture for Differentiated Services, RFC 2475, 1998

[2] Nichols K. ,Jacobson V. ,Zhang L. , A Two-bit Differentiated Services Architecture for the Internet, RFC 2638, 1999

[3] Mantar H.A., Okumus Ý.T., Hwang J., Chapin S.J., A Scalable Intra-Domain Resource Management Architecture for Diffserv Networks, *Journal of High Speed Networks*, 15, 185-205, 2006

[4] Jacobson V., Nichols K., Poduri K., An Expedited Forwarding PHB, RFC 2598, 1999

[5] Heinanen et.al., Assured Forwarding PHB Group, RFC 2598, 1999

[6] Schelen O., Pink S., Resource Sharing in Advance Request Agents, *Journal of High Speed Networks: Special issue on Multimedia Networking*, 7(3-4):213-228, 1998

[7] Lin Y., Chang C., Hsu Y., Bandwidth Brokers of Instantenaous and Book-ahead Requests for Differentiated Services Networks, *ICICE Transactions on Communication*, E85-B, No.1,278-283, 2002

[8] Ahmad I., Kamruzzaman J., Aswathanarayaniah S., A Dynamic Approach to Reduce Preemption, in *Book-ahead Reservation in QoS-Enabled Networks*, *Computer Communications*, 29(9):1443-1457, 2006

[9] Degermark M. Et.al., Advance reservationss for Predictive Service in the Internet, *Multimedia Systems*, 5(3):177-186, 1997

[10] Greenberg A.G., Srikant R., Whitt W., Resource Sharing for Book -Ahead and Instantaneous -Request Calls, *IEEE/ACM Transactions on Networking*, 7(1):10-22, 1999

[11] Srikant R., Whitt W., Resource Sharing for Book-Ahead and Instantaneous-Request Calls Using a CLT Approximation, *Telecommunication Systems*, 16(3-4):233-253, 2001

[12] Karsten M., Beries N., Wolf L., steinmetz R., A Policy-Based Service Specification for Resource Reservation in Advance, *Proceedings of the International Conference on Computer Communications (ICCC'99)*, Tokyo, Japan, 82-88, Sept 1999

[13] Ahmad I., Kamruzzaman J., Preemption Policy in QoS-Enabled Networks: A Customer Centric Approach, *Journal of Research and Practice in Information technology*, 39(1):61-79, 2007

[14] Adamson A. et.al., QBone Signaling Design Team final Report", Internet2 QBone Sinaling Workgroup, http://qos.internet2.edu/wg/documents-informational/20020709-chimento-etal-qbone-signaling/, Jul 2002

[15] The Network Simulator - ns-2, http://www.isi.edu/nsnam/ns/

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 2 (June), pp. 365-376

Fuzzy Control Design for a Class of Nonlinear Network Control System: Helicopter Case Study

P.Q. Reyes, J.O. Arjona, E.M. Monroy, H.B. Pérez, A.D. Chavesti

P. Quiñones-Reyes
Instituto Tecnológico de Jiquilpan
Carr. Nac., S/N Km. 202, CP 59510,
Jiquilpan, Michoacán, México.

J. Ortega-Arjona, E. Méndez-Monroy,
H. Benítez-Pérez, A. Durán-Chavesti
Universidad Nacional Autónoma de México
Apdo. Postal 20-726, Admón. 20,
Del. A. Obregón, México D. F., CP. 01000.
polbond@hotmail.com, jloa@ciencias.unam.mx,
{hector,chavesti}@uxdea4.iimas.unam.mx

Abstract: This paper presents a fuzzy control approach to a helicopter MIMO nonlinear system, implemented on a Networked Control System, as case study. For this, a hardware-in-the-Loop implementation is developed using several multi-channel A/D Cards, integrated to a computer network system. Variant time delays are considered over Ethernet and CANBUS networks. Fuzzy logic is used to deal with the complexity of the integrated computer network as well as with the dynamics of the system. Two fuzzy logic control systems are coupled for both signals of the helicopter case study: yaw and pitch. Both these tend to concentrate around desired references, considering variant time delays.
Keywords: fuzzy control, networked control systems.

1 Introduction

Reconfiguration is a transition that modifies the structure of a system so it changes its representation of states. Here, it is used as a feasible approach for fault isolation, and also, it is a response to time delay modification.

In control systems, several modelling strategies for managing time delay within control laws have been studied by different research groups. Nilsson [16] proposes the use of a time delay scheme integrated to a reconfigurable control strategy, based on a stochastic methodology. Jiang [12] describe how time delays are used as uncertainties, which modify pole placement of a robust control law. Izadi present an interesting case of fault tolerant control approach related to time delay coupling. Blanke [7] study reconfigurable control from the point of view of structural modification, establishing a logical relation between dynamic variables and the respective faults. Thompson [17] and Benítez-Pérez [4] consider that reconfigurable control strategies perform a combined modification of system structure and dynamic response, and thus, this approach has the advantage of bounded modifications over system response. Recent approximations are presented by Dai, which allow a reasonable but static approximation for time variable strategy. Also, Kim [20] have followed a Maximum Allowable Time Delay (MADB), where complex task behaviour is permitted as long as MADB is preserved [22] [23].

The approach here makes use of a case study that takes time delays due to communication as deterministic measured variables. For this, a Fuzzy Control law [1] is used, where time delays result from the deterministic reconfiguration of communications due to a scheduling algorithm. Fuzzy Control is used for managing extended horizons from system inputs and outputs, to determine several scenarios modified by time delays. Recent results encourage this approximation, as shown in Benítez-Perez [5] [6].

For experimental purposes, the following considerations are taken:

1. Time delays are bounded.

2. The combination tends to be globally stable.

The objective of this paper is to present the design of a fuzzy control strategy developed from the time delay knowledge, as well as computer network behaviour considering communication amongst nodes for a helicopter case study. The novelty is to propose a Fuzzy Control [24] [25] for Network Control System (NCS) based on the defined communication network and variant time delays.

Present approach makes use of time delays due to communication as deterministic measured variables, defined by previous knowledge of computer network. In here, control law views time delays as a result of bounded communications based upon scheduling algorithm.

A basic consideration for this approach is that time delays are bounded to MADB. The main reason for these is the behavior of computer network as well as node processing time. However, in order to integrate this variable as global time delay (Δt_*), it is necessary to consider its nominal value. This is presented as a percentage: 0% refers to the current time delay, and 100% represents extreme time delays. Hence, this value is any value produced as time delay less than or equal to total sampling period.

Having defined global time delays as its nominal value, the fuzzy control structure is proposed as shown in Figure 1.

Fuzzy control is chosen here for implementing a gain-scheduler controller, on contrary of Smith's predictor [18], since it has a smooth transition between scenarios. Furthermore, the chosen operating points are the reference elements of proposed fuzzy control. Thus, any degradation from time delays would degrade the control law, however keeping a stable response from the plant. Time delay degradation is bounded from communication protocol, as explained by [13].

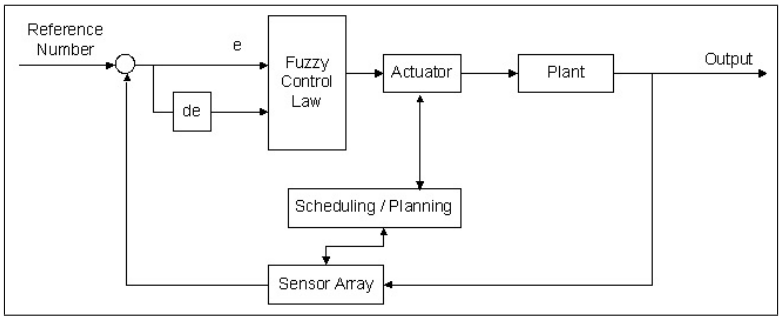


Figure 1: Fuzzy Control Structure.

Notice that the current approach follows a Mamdani strategy [18] rather than a Takagi Sugeno (TKS) [1] for the implementation of the fuzzy control. The development of a TKS strategy is considered as a future work, aiming for integrating time delays into a subsequent part of the fuzzy control rules.

The fuzzy control law for a fault free scenario is presented in Figure 2, based on [18]. Membership functions are Gaussian distributions, where e has six membership functions (PB, PM, PS, NS, NM, NB), and de has 6 membership functions (PB, PM, PS, NS, NM, NB). The output variable has eight membership functions (PB, PM, PS, PZ, NZ, NS, NM, NB). An additional variable, named Current Nominal Time Delay (CNTD), has three membership functions (N, Z, P).

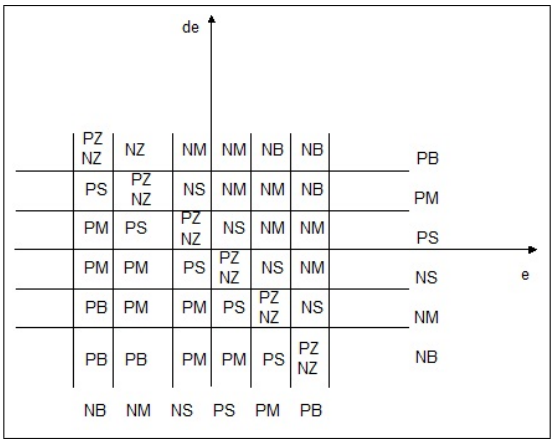


Figure 2: Classical Structure for Fuzzy Control Law

Notice that this implementation is common for fuzzy control design, however, here fuzzy control design focuses on the strategy of the control law due to the communication in the computer network, as well as the consequent time delays. These time delays are measurable and bounded, according to previous knowledge of computer network, based on a finite state machine that selects when valid strategies take place. Here, the scheduling approximation is the inherent EDF algorithm [14] with certain lost data. The aim is to study how this transition is carried out when using a Fuzzy Mamdani approach [1], based on a NCS. Particularly, for the actual NCS, the communication network strongly affects the dynamics of the system, expressed as a time variance that exposes a nonlinear behaviour. Such nonlinearity is addressed by incorporating time delays. From real-time system theory, it is known that time delays are bounded even in the case of causal modifications due to external effects [15]. Using this representation, time delays are counted using simple addition, as described as follows.

2 Case Study and Experimental Setup

The case study here is a Helicopter MIMO system, integrated to a computer network as shown in Figure 3 [21]. It is integrated by three A/D Cards: an AD512 card is the interface of joystick, acting as moving reference; an AD612 card is the interface of actuators yaw and pitch; and a Q4 card is used for sampling the information from two encoders which sense yaw and pitch information from the movement. Two networks are used for this case study: an Ethernet network at 10/100Mhz, and a CAN network at 1Mbit/s. For experimental purposes, the controller node works as well as gateway for both databuses.

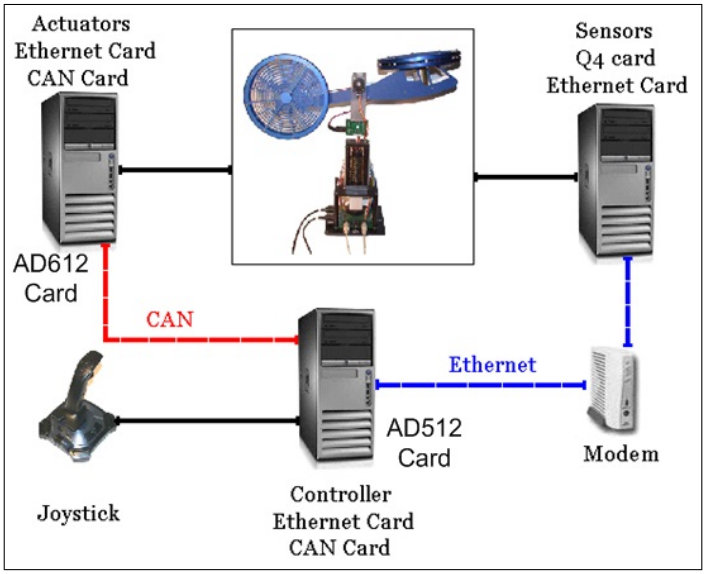


Figure 3: Helicopter MIMO system for the case study

While the description of the helicopter is beyond the scope of this paper, a brief description is provided here. However, further information can be found in the original Quanser [21] documentation. The Quanser 2 DOF Helicopter consists of a helicopter model mounted on a fixed base with two propellers that are driven by DC motors. The front propeller controls the elevation of the helicopter nose about the pitch axis and the back propeller controls the side-to-side motions of the helicopter about the yaw axis. The pitch and yaw angles are measured using high-resolution encoders. The two degrees of freedom helicopter pivots about the pitch axis by angle θ and about the yaw axis by angle ψ.The pitch is defined positive when the nose of the helicopter goes up and the yaw is defined positive for a clockwise rotation. Table 1 lists the various lengths, masses, and moment of inertias associated with this helicopter model.

Variable	**Description**	**Value**	**Unit**
Beq,p	Equivalent viscous damping about pitch axis.	0.800	N/V
Beq,y	Equivalent viscous damping about yaw axis.	0.318	N/V
Jeq,p	Total moment of inertia about pitch pivot.	0.0384	$kg.m^2$
Jeq,y	Total moment of inertia about yaw pivot.	0.0384	$kg.m^2$
Kpp	Thrust torque constant acting on pitch axis from pitch motor/propeller.	0.204	N.m/V
Kpy	Thrust torque constant acting on pitch axis from yaw motor/propeller.	0.0068	N.m/V
Kyp	Thrust torque constant acting on yaw axis from pitch motor/propeller.	0.0219	N.m/V
Kyy	Thrust torque constant acting on yaw axis from yaw motor/propeller.	0.072	N.m/V
lcm	Center-of-mass length along helicopter body from pitch axis.	0.186	Cm
$mheli$	Total moving mass of the helicopter.	1.3872	Kg

Table 1: Helicopter specifications and model parameters.

The linear state-space model of the helicopter is given by Equation 1. Notice that the nonlinear equations of motion are considered linear about the quiescent point (θ_0=0, ψ_0=0, $\dot{\theta}_0$=0, $\dot{\psi}_0$=0). Substituting the state x = [θ, ψ, $\dot{\theta}$, $\dot{\psi}$] and solving for $\dot{x}$:

$$\dot{x} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -\frac{B_p}{J_{eq,p}+m_{heli}l_{cm}^2} & 0 & 0 & 0 \\ 0 & 0 & 0 & -\frac{B_y}{J_{eq,y}+m_{heli}l_{cm}^2} & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} x + \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ \frac{K_{pp}}{J_{eq,p}+m_{heli}l_{cm}^2} & \frac{K_{py}}{J_{eq,p}+m_{heli}l_{cm}^2} \\ \frac{K_{yp}}{J_{eq,y}+m_{heli}l_{cm}^2} & \frac{K_{yy}}{J_{eq,y}+m_{heli}l_{cm}^2} \\ 0 & 0 \\ 0 & 0 \end{bmatrix} u \tag{1}$$

$$y = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} x$$

In order to establish an experimental setup, several conditions nave to be taken into account. For example, the communication network frequencies for Ethernet and CANBUS are set 10/100Mhz and 1Mbit/s, respectively. Moreover, the derivative boundary for pitch is -0.3 to 0.3. The rest of the boundaries are shown Table 2.

	Maximum Value	**Minimum Value**	**Standard Deviations**
Derivative Value	-0.3	0.3	
Error Value	-0.25	0.25	
Control Value	0	25	
Derivative Center	-0.3 -0.2 0 0.2 0.3		0.12
Error Centers	-0.25 -0.05 0 0.05 0.25		0.1
Names of current membership functions	CMP CP CM CG CMG		

Table 2: Signal Characteristics of pitch.

Table 3 shows the related consequent matrix for pitch signal. This is coupled following control law for yaw signal, where time delays are affecting both control laws.

Error-Derv	**MP**	**P**	**M**	**G**	**MG**
mp	CP	CP	CP	CM	CM
p	CP	CP	CM	CM	CM
m	CP	CM	CM	CM	CG
g	CM	CM	CM	CG	CG
mg	CM	CM	CG	CG	CG

Table 3: Membership relationships amongst consequent parts for pitch signal.

Table 4 shows the main characteristics of control approach for the yaw signal. Notice that the numerical values tend to be different to those related to pitch control law. This is basically due to dynamic conditions of yaw signal.

	Maximum Value	**Minimum Value**	**Standard Deviations**
Derivative Value	-0.2	0.2	
Error Value	-0.5	0.5	
Control Value	-20	15	
Derivative Center	-0.2 -0.1 0 0.1 0.2		0.08
Error Centers	-0.25 -0.125 0 0.125 0.25		0.25
Names of current membership functions	PMP PP PM PG PMG		

Table 4: Signal Characteristics of yaw.

Table 5 shows the matrix for yaw signal. This is considerable different from pitch due to dynamic response of the plant.

Error-Derv	**MP**	**P**	**M**	**G**	**MG**
Mp	PMG	PMG	PMG	PMG	PMG
p	PMG	PMG	PG	PG	PG
m	PM	PM	PM	PM	PM
g	PP	PP	PP	PP	PP
mg	PP	PP	PP	PP	PMP

Table 5: Membership relationships amongst consequent parts for yaw signal.

Now for both signals, the Mamdani integration is defined in Eqn 2:

$$\begin{gathered} u = \frac{\sum_{j=1}^{r} \alpha_j \beta_j}{\sum_{j=1}^{r} \alpha_j} \\ r = mxn \\ \alpha_j = \gamma_m \gamma_n \\ j = 1...mxn \\ m = 1...M \\ n = 1...N \\ \gamma_i = exp\left(-\left(\frac{x-c_i}{\sigma_i}\right)^2\right) \end{gathered} \tag{2}$$

where:

- c_i are the centers of the Gaussians
- σ_i are the standard deviations of the Gaussians.
- m is the total number of Gaussians per error
- n is the total number of Gaussians per error derivatives, and

- β_j is the group control centers

Figure 4 presents the control surface for pitch signal. This surface tends to have smooth transitions between rules and consequent parts. It has a local valley due to delay interaction and error. This response is reflected on error and derivative conversion.

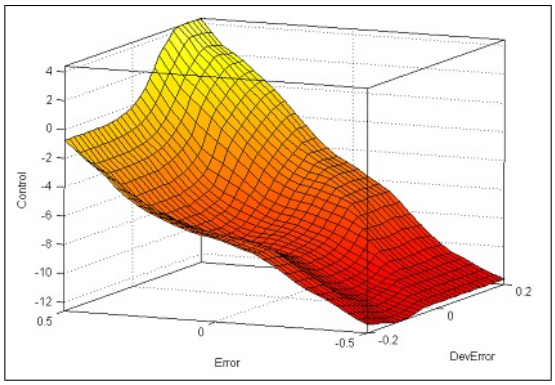


Figure 4: Control surface for pitch signal.

Figure 5 shows the control surface for the yaw signal. This tends to be more accurate in terms of formal fuzzy transitions, where only one small local valley exists.

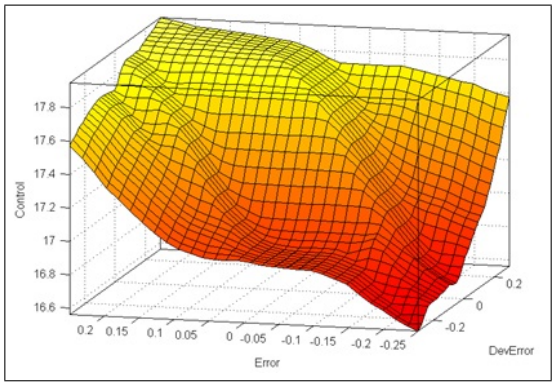


Figure 5: Control signal for yaw signal.

In both resulting surfaces, these tend to be smaller at negative error and derivative error values. On the contrary biggest values are related to both positive values (error and derivative respectively).

3 Experimental Results

Current time delays are bounded to normal representation as shown in Fig. 6. Where the delays are mainly considered as a result of communication media following Fig. 3.

The current experimental results show how the Helicopter MIMO case study adequately performs during time variant conditions. Considering an experimental execution of 50 seconds, several results show that, for instance, current error of yaw signal (Figure 7.a) presents a bounded response from 5 seconds onwards, while its derivative (Figure 7.b) tends to be almost zero, even in the case of change perturbation around 5 seconds.

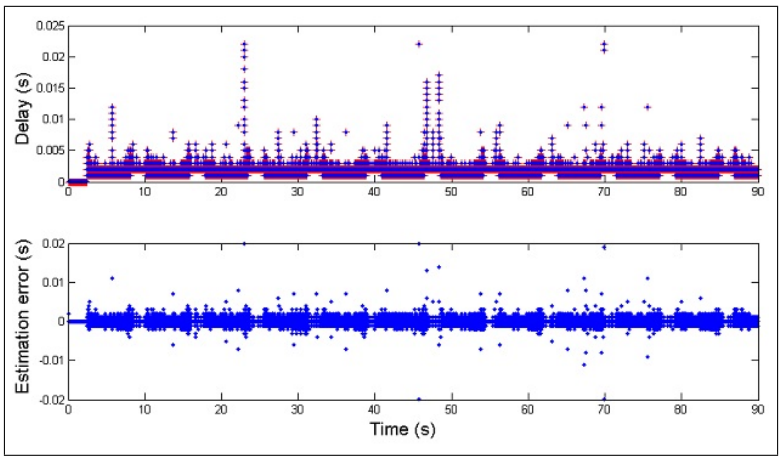


Figure 6: Communication Time Delays and the related Estimation Error.

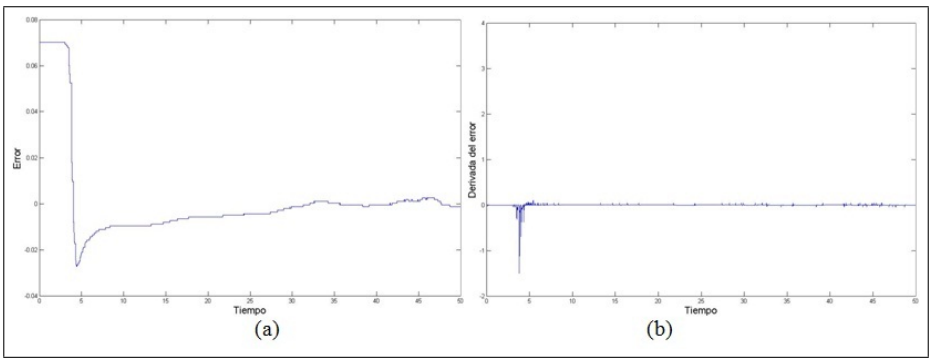


Figure 7: Error response for yaw signal.

Figure 8 shows the control response for the yaw signal. It is a bounded response, where several glitches are spotted. These result from loosing data from the Ethernet network. Such a behaviour is not the actual delay from computer network, but it is an inherent feature of the Ethernet computer network.

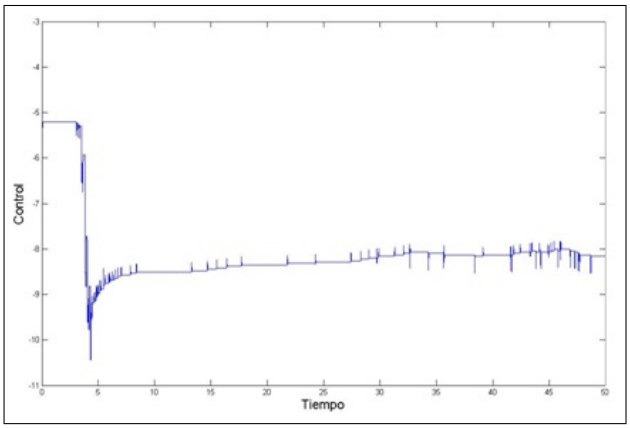


Figure 8: Control response for yaw signal.

Figure 9 presents a not-smooth transition between error and its derivative, due to local valley shown at Figure 4. Although system response is feasible for the current combination, it tends to have drastic, undesirable jumps.

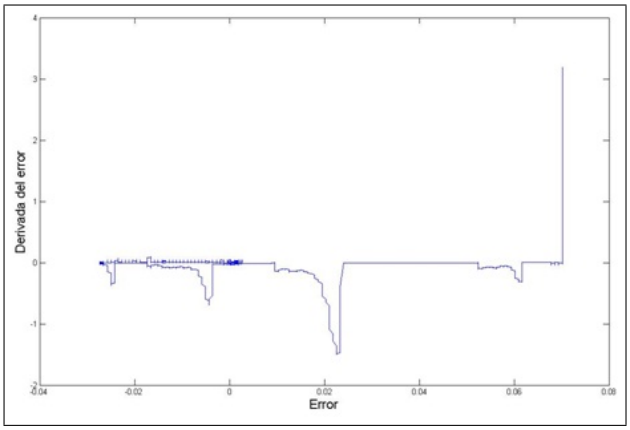


Figure 9: Error vs derivative of error response.

Regarding the pitch signal error, the response tends to be smooth and bounded. There is an inherent oscillation at error response (Figure 10.a), which results of slight movements from the joystick. The related derivative (Figure 10.b) reflects the bounded responses with some glitching.

Figure 11 shows the control response, presenting a similar behaviour, where glitches are present. These glitches result from loosing data in the communication network.

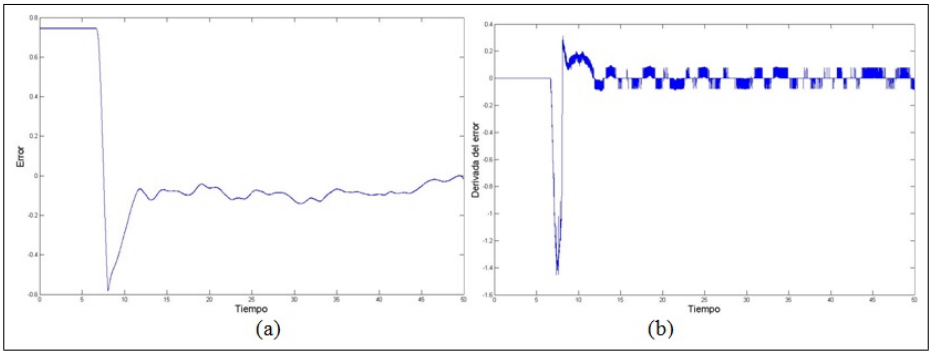


Figure 10: Error response for pitch signal.

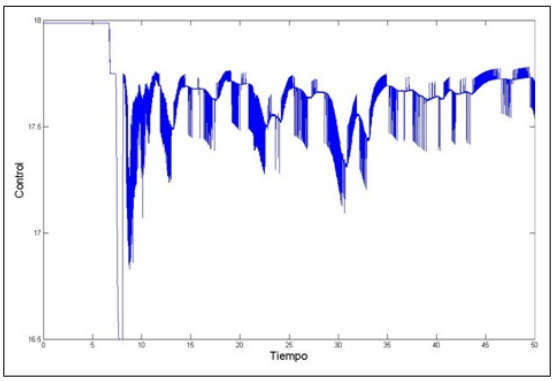


Figure 11: Control response for pitch signal.

Figure 12 shows a common response between error and derivative of error as convergence signal.

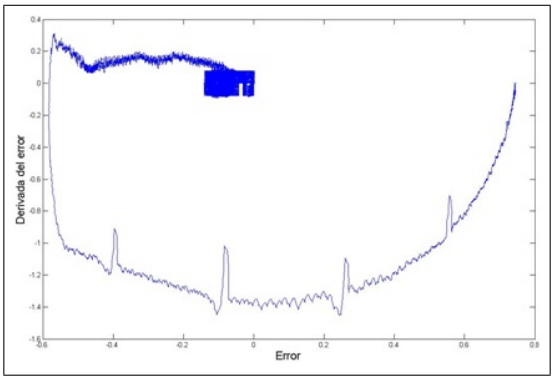


Figure 12: Error vs derivative of error response for pitch signal.

4 Concluding Remarks

This paper presents an approach for the integration of two fuzzy logic controllers, in order to perform control under time variant conditions. These two techniques are applied in parallel. Although there is no formal verification for this sequence, it has been adopted since Mamdami approximation provides stable conditions for control reconfiguration. Moreover, the use of a real-time testbed to approve or disapprove changes on the behaviour of a computer network allows bounding time delays during a specific time frame. This local time delay allows the design of a control law, capable to cope with new conditions. Hardware-in-the-loop implementation is feasible, since complex integration through computer network as well as real-time Operating systems are integrated through bounded time delays

Acknowledgements

The authors would like to thank the financial support of DISCA-IIMAS-UNAM, and UNAM-PAPIIT (IN103310-3) , PICCO 10-53, Mexico in connection with this work.

Bibliography

[1] J. Abonyi, Fuzzy Model Identification for Control, *Birkhďž¨user*, 2003

[2] L. Almeida, P. Pedreiras and J. A.-Fonseca, The FTT-CAN Protocol: Why and How, *IEEE Transactions on Industrial Electronics*, 49(6):1189-1201, 2002

[3] H. Benítez-Pérez and F. García-Nocetti, Switching Fuzzy Logic Control for a Reconfigurable System Considering Communication Time Delays, *Proceedings, CDROM, European Control Conference*, ECC 03 September, 2003

[4] H. Benítez-Pérez and F. García-Nocetti, Reconfigurable Distributed Control, *Springer Verlag*, 2005

[5] H. Benítez-Pérez, Real-Time Distributed Control A Diverse Approach for Nonlinear Problem, *Nonlinear Analysis: Hybrid Systems and Applications*, doi:10.1016/j.nahs.2006.06.004, Vol 2/2 pp 474-490, Junio 2008

[6] H. Benítez-Pérez, J. S.-Gonzalez, F. C.-Flores and F. García-Nocetti, Fault Classification for a Class of Time Variable Systems by using a group of three ART2 Networks, *International Journal Control and Intelligent Systems*, DOI: 10.2316/Journal.201.2008.1.201-1820, Vol 36, No. 1, 2008

[7] M. Blanke, M. Kinnaert, J. Lunze and M. Staroswiecki, Diagnosis and Fault Tolerant Control, *Springer*, 2003

[8] A. Cervin, D. Henriksson, B. Lincoln, J. Eker and K. Arzén, How Does Control Timing Affect Performance?, *IEEE Control Systems Magazine*, Vol. 23, pp. 16-30, 2003

[9] T. Frank, K. F.-Kraiss and T. Kuhlen, Comparative Analysis of Fuzzy ART and ART-2A Network Clustering Performance, *IEEE Transactions on Neural Networks*, Vol. 9, No. 3, May 1998

[10] D. Hanselman and B. littlefield, *Mastering MATLAB*, Prentice Hall, 2002

[11] R. I.-Zamanabadi and M. Blanke, A Ship Propulsion System as a Benchmark for Fault-Tolerant Control, *Control Engineering Practice*, Vol. 7, pp. 227-239, 1999

[12] J. Jiang, and Q. Zhao, Reconfigurable Control Based on Imprecise Fault Identification, *Proocedings of the American Control Conference, IEEE*, pp. 114-118, San Diego, June, 1999

[13] F. Lian, J. Moyne and D. Tilbury, Network Design Consideration for Distributed Control Systems, *IEEE Transactions on Control Systems Technology*, Vol. 10, No. 2, pp. 297-307, March 2002

[14] L. Liu, Real-time Systems, *Wiley*, 2002

[15] Menendez L. de C. A. and H. Benítez-Pérez, Node Availability for Distributed Systems considering processor and RAM utilization Based upon a Local Optimization Procedure, INT J COMPUT COMMUN, ISSN 1841-9836, 5(3):336-350, 2010

[16] J. Nilsson, Real-Time Control with Delays, *PhD. Thesis, Department of Automatic Control, Lund Institute of Technology*, Sweden, 1998.

[17] H. Thompson, Wireless and Internet Communications Technologies for monitoring and Control, *Control Engineering Practice*, vol. 12, pp. 781-791, 2004

[18] D. Driankov, H. Hellendoorn and M. Reinfrank, An Introduction to Fuzzy Logic Control, *Springer-Verlag*, 1994

[19] L. Zhang, Y. Shi, T. Chen and B. Huang, A New Method for Stabilization of Networked Control Systems with Random Delays, *American Control Conference*, pp. 633-637, 2005

[20] D. Kim, D. Choi and P. Mohapatra, Real-Time Scheduling method for Networked Discrete Control Systems, *Control Engineering Practice*, Vol 17, pp: 564-570, 2009

[21] http://www.quanser.com/english/html/solutions/ fs_soln_software_wincon.html

[22] Zmaranda D., Gabor G., Popescu D.E., Vancea C., Vancea F., Using Fixed Priority Preemptive Scheduling in Real-Time Systems, INT J COMPUT COMMUN, ISSN 1841-9836, 6(1):187-195, 2011

[23] Dai L., Chang Y., Shen Z.,An Optimal Task Scheduling Algorithm in Wireless Sensor Networks, INT J COMPUT COMMUN, ISSN 1841-9836, 6(1):101-112, 2011

[24] Negoita C.V., Remembering the Beginnings, INT J COMPUT COMMUN, ISSN 1841-9836, 6(3):458-461, 2011

[25] Nyirenda C.N., Dong F., Hirota K., Distance Based Triggering and Dynamic Sampling Rate Estimation for Fuzzy Systems in Communication Networks, INT J COMPUT COMMUN, ISSN 1841-9836, 6(3):462-472, 2011

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 2 (June), pp. 377-387

A 2-level Metaheuristic for the Set Covering Problem

C. Valenzuela, B. Crawford, R. Soto, E. Monfroy, F. Paredes

Claudio Valenzuela, Broderick Crawford
Pontificia Universidad Católica de Valparaíso
Valparaíso, Chile
{claudio.valenzuela, broderick.crawford}@ucv.cl

Ricardo Soto
1. Pontificia Universidad Católica de Valparaíso
Valparaíso, Chile, and
2. Universidad Autónoma de Chile
ricardo.soto@ucv.cl

Eric Monfroy
Universidad Técnica Federico Santa María
Valparaíso, Chile
eric.monfroy@inf.utfsm.cl

Fernando Paredes
Escuela de Ingeniería Industrial
Universidad Diego Portales
Santiago, Chile
fernando.paredes@udp.cl

Abstract: Metaheuristics are solution methods which combine local improvement procedures and higher level strategies for solving combinatorial and nonlinear optimization problems. In general, metaheuristics require an important amount of effort focused on parameter setting to improve its performance. In this work a 2-level metaheuristic approach is proposed so that Scatter Search and Ant Colony Optimization act as "low level" metaheuristics, whose parameters are set by a "higher level" Genetic Algorithm during execution, seeking to improve the performance and to reduce the maintenance. The Set Covering Problem is taken as reference since is one of the most important optimization problems, serving as basis for facility location problems, airline crew scheduling, nurse scheduling, and resource allocation.
Keywords: metaheuristics, genetic algorithm, scatter search, ant colony optimization, set covering problem.

1 Introduction

The Set Covering Problem (SCP) is a classical problem in computer science and one of the most important discrete optimization problems since it can model conveniently real world problems. Some of these problems — which can be modeled as a set covering problem — include facility location, airline crew scheduling, nurse scheduling, resource allocation, assembly line balancing, vehicle routing, among others [6]. There are several studies which have implemented a SCP solution using metaheuristics [1,2,12]. Depending on the algorithm that has been used, the quality of the solution wanted and the complexity of the SCP chosen, it is defined the amount of customization efforts required. Conveniently, this work proposes transferring part of this customization effort to another metaheuristic (a "high level" metaheuristic) which can handle

the task of parameters adjustment for a low level metaheuristic. This approach is considered as a multilevel metaheuristic since there are two metaheuristics covering tasks of parameter setting, for the former, and problem solving, for the latter [9].

The main design of the implementation proposed considers a Genetic Algorithm (GA) [10] at online (Control) and offline (Tuning) parameter setting for a low level metaheuristic (Ant Colony Optimization (ACO) or Scatter Search (SS)) using a Reactive Search approach and an Automatic Parameter Tuning approach. In Reactive Search, feedback mechanisms are able to modify the search parameters according to the efficiency of the search process, i.e. the balance between intensification and diversification can be automated by exploiting the recent past of the search process through dedicated learning techniques [13]. The Automatic Parameter Tuning is carried by an external algorithm which searches for the best parameters in the parameter space in order to tune the solver automatically. Ant Colony Optimization and Scatter Search techniques [11] have shown interesting results at solving SCP [6] and similar problems [5]. For the purpose of this work, the former is selected by its constructional approach for generating solutions, plus its stochastic-based operators. The latter is considered as an evolutionary (population-based) algorithm which uses, essentially, deterministic operators. Both of them provide good reference metaheuristics in terms of their foundations, their problem solving approaches, their design maturity, and in terms of how different one is from the other, making them highly suitable to the development of this work.

2 Set Covering Problem

A general mathematical model of the Set Covering Problem can be formulated as follows:

$$(1)\ \ Minimize\, Z = \sum_{j=1}^{n} c_j x_j \qquad j = \{1, 2, 3, ..., n\}$$

Subject to:

$$(2)\ \ \sum_{j=1}^{n} a_{ij} x_j \geq 1 \qquad i = \{1, 2, 3, ..., m\}$$

$$(3)\ \ x_j = \{0, 1\}$$

Equation (1) is the objective function of set covering problem, where c_j is the cost of j-column, and x_j is decision variable. Equation (2) is a constraint to ensure that each row is covered by at least one column where a_{ij} is a constraint coefficient matrix of size m x n whose elements comprise of either "1" or "0". Finally, equation (3) is the integrality constraint in which the value x_j can be "1" if column j is activated (selected) or "0" otherwise. Different solving methods have been proposed in the literature for the SCP. There exist examples using exact methods [8], linear programing and heuristic methods [7], and metaheuristic methods [2]. Has being pointed out, that one of the most relevant applications of SCP is given by crew scheduling problems in mass transportation companies where a given set of trips has to be covered by a minimum-cost set of pairings, a pairing being a sequence of trips that can be performed by a single crew.

3 Multilevel Metaheuristics and Parameter Setting

Metaheuristics, in their original definition, are solution methods that orchestrate an interaction between local improvement procedures and higher level strategies to create a process capable

of escaping from local optima and performing a robust search of a solution space [3]. Over time, these methods have also come to include any procedures that employ strategies for overcoming the trap of local optimality in complex solution spaces, especially those procedures that utilize one or more neighborhood structures as a means of defining admissible moves to transition from one solution to another, or to build or destroy solutions in constructive and destructive processes. A number of the tools and mechanisms that have emerged from the creation of metaheuristic methods have proved to be remarkably effective, so much that metaheuristics have moved into the spotlight in recent years as the preferred line of attack for solving many types of complex problems, particularly those of a combinatorial nature.

Multilevel Metaheuristics can be considered as two or more metaheuristics where a higher level metaheuristic controls de parameters of a lower level one, which is at charge of dealing more directly to the problem. Therefore, parameter Setting is a key factor in the design of Metaheuristics and Multilevel Metaheuristics [4], since they improve their solving performance by modifying and adjusting themselves to the problem at hand, either by self-adaptation or supervised adaptation.

3.1 Parameter Setting

Many parameters have to be set for any metaheuristic. Parameter setting may allow a larger flexibility and robustness, but requires a careful initialization. Those parameters may have a great influence on the efficiency and effectiveness of the search. It is not obvious to define *a priori* which parameter setting should be used. The optimal values for the parameters depend mainly on the problem and even the instance to deal with and on the search time that the user wants to spend in solving the problem. A universally optimal parameter values set for a given metaheuristic does not exist.

3.2 Tuning Before Solving

Also know as "Offline Parameter Initialization". As previously mentioned, metaheuristics have a major drawback; they need some parameter tuning that is not easy to perform in a thorough manner. Those parameters are not only numerical values but may also involve the use of search components [15]. Usually, metaheuristic designers tune one parameter at a time, and its optimal value is determined empirically. In this case, no interaction between parameters is studied. This sequential optimization strategy (i.e., one-by-one parameter) do not guarantee to find the optimal setting even if an exact optimization setting is performed. Main flavors of "tuning before solving" techniques include: Parameter Tuning on Preliminary Experiments, Empirical Manual Tuning and Automatic Parameter Tuning by an External Algorithm.

3.3 Control During Solving

Also know as "Online Parameter Initialization". The drawback of the offline parameter setting approaches is their high computational cost, particularly if this approach is used for each input instance of the problem. Indeed, the optimal values of the parameters depend on the problem at hand and even on the various instances to solve. Then, to improve the effectiveness and the robustness of offline approaches, they must be applied to any instance (or class of instances) of a given problem. The control of the solverÕs behavior during the run can be achieved by either modifying its components and/or its parameters. This corresponds, for instance, to an online adjustment of the parameters or heuristics. Such control can be achieved by means of supervised control schemes or by self adaptive rules. Of course, such approaches often rely on a learning process that tries to benefit from previously encountered problems along the search or even

during the solving of other problems. At this section are considered the approaches that change the parameters during the search with respect to the current state and other parameters. Of course, these parameters have a direct influence on the heuristics functions, but also these latter functions stay the same during the solving process. Some remarkable approaches are introduced as follows: Hyperheuristics and Reactive Search [16].

4 Implementation Details

4.1 GA-SS Design

The design proposed for the multilevel implementation is based on documented standards proposed for each metaheuristic. Both metaheuristics are looking to be as close as they can to its origins. Obviously, a multilevel implementation with adaptive parameter control approach forces some changes to the "high levelÓ metaheuristic, in this case, GA. This case will be similar to both combinations: GA-SS and GA-ACO. Hence, the design of GA seems particularly faithful to its basic design.

[*Size of P , BestSet , DiverseSet , EnhanceTrials , MaxSol*]

Figure 1: GA Chromosome Representation for SS.

The chromosome representation shows that the first gene, for this case the "*Size of P*" is the number of initial solutions for SS, will take values between 1 and $n/2$ where n number of variables. Same way for the second gene "*BestSet*", which is the size of the Best Solutions Reference Set. Similarly, "*DiverseSet*" is the size of the Most Diverse Solutions Set. "*EnhanceTrials*", represents the number e of trials for the Improvement Method to try to enhance a solution, where $e = \{1...n\}$. Finally, "*MaxSol*" is the limit of solutions generated by a call to the Scatter Search algorithm.

4.2 GA components

The main components of GA components are:

Initialization function: Initializes the values of genes within the variables bounds. It also initializes (to zero) all fitness values for each member of the population. It takes upper and lower bounds of each variable from user defined parameters. It randomly generates values between these bounds for each gene of each genotype in the population.

Evaluation function: The upper level metaheuristic uses two criterions; (1) the same evaluation function of the lower level metaheuristic, this is the corresponding *objective function* of SCP, and (2) a *Objective Function* of SCP penalized by the processing effort.

"Keep the best" function: This function keeps track of the best member of the population.

Elitist function: The best member of the previous generation is stored.

Selection function: Standard proportional selection for maximization/minimization problems incorporating elitist model — makes sure that the best member survives.

Crossover selection: selects two parents that take part in the crossover. This work implements a single point crossover.

Mutation: Random uniform mutation. A variable selected for mutation is replaced by a random value between lower and upper bounds of this variable.

4.3 SS components

A description of SS components included in GA-SS design:

Diversification Generation Method: This method generates diverse binary initial solutions, it starts with an all-zero arbitrary solution, and then begins adding "1" in every position with a jump of k bits, where $k = \{1...n\}$ and $n < (number\_of\_vars) - 1$. In example, for $k = 1$ it will generate a solution vector $[1, 0, 1, 0, 1, 0, 1, 0...]$. Should be noticed that this method always starts placing a bit "1" at the first position. Together with the previously generated solution, the method generates the complement of that solution, which will be $[0, 1, 0, 1, 0, 1, 0, 1, 0, 1...]$. The quantity of solutions generated by this method is a parameter controlled by the GA.

Improvement Method: transforms a trial solution into a enhanced trial solution. If the trial solution is not feasible, should be fixed until it turns feasible. The input solutions are not required to be feasible. If the input trial solution is not improved as a result of the application of this method, the "enhanced" solution is considered to be the same as the input solution. The limit of enhancement trials (not the fix trials) is controlled by the GA. The duties of fixing and enhancing are improved since a vector of ratios is calculated (the length of the vector is the number of variables, or *columns*). This vector is then sorted from min to max ratio. Each element represents a variable at the objective function. Then, when trying to fix a solution should be followed in-order, so the first item will represent the column which covers more rows of the incidence matrix. If this position is "0" at the solution, should be turned to "1", because the method is wanting to *add* "1" for covering rows and turn the solution to feasibility. This process continues until the solution is feasible. When trying to enhance a solution, the vector of ratios should be accessed in post-order, trying to turn as many "1" to "0" as it can while the solution keeps feasible. The method will try to enhanced until the limit l is reached, where l is a parameter controlled by the GA.

Reference Set Update Method: builds and maintain two reference sets, consisting of the b "best" solutions and the d "*most* diverse" solutions found (where the value of b and d is set by the GA), organized to provide efficient accessing by other parts of the solution procedure. The criteria for adding the "best" solutions to its set is the cost at the objective function, also, the criteria for adding solutions at the "most diverse" set is the Hamming distance to all the solutions at the "best set" and the "diverse set".

Subset Generation Method: operates on the reference set, to produce a subset of its solutions as a basis for creating combined solutions. The most common subset generation method will be used, which generates all pairs of reference solutions (i.e., all subsets of size 2) from the "best set" and the "diverse set" together.

Solution Combination Method: transforms a given subset of solutions produced by the Subset Generation Method into one combined solution. The combination method is based in the *cost at objective function/ rows covered* ratio, which is calculated per column. When combining two solutions, the following procedure is used:

```
if (Sol_1[x] = Sol_2[x]) then
     newSol[x] := Sol_1[x]
elseif (Ratio[x] < median) then
     newSol[x] := 1
else newSol[x] := 0
```

where x is the position of the bit being evaluated with $x = \{1...maxColumns\}$, and the *median* is the median of the vector of ratios for a given SCP instance.

4.4 GA-ACO Design

For this implementation, the same structure of GA components presented at 4.2 is used. Added to this, the implementation of ACO for SCP is very straightforward. The components used by ACO are:

Pheromone. Denoted by $t = \tau_i$, the matrix of pheromones will be used to consolidate information obtained by each ant, i.e. the amount of pheromone stored in each column i. $\tau_i(t)$ specifies the intensity of the pheromone at column i in time t, and updates in a local way (according to the path of the ants), and in a global way (the pheromone evaporates in every matrix column). The matrix is initialized with a value t_0 which will be 10^{-6} for this implementation.

Transition. For apply a transition between two columns, this should be done according to the list of candidates. The list of candidates of a column contains the c more attracting columns, and they are sorted in a sequential manner. The transition is carried according to:

If exists at least one column $j \in$ candidate list, then, choose the next column $j \in J_i^k$, between the c columns in the candidate list according:

$$j = \begin{cases} max_u \in J_i^k[\tau_u(t)][\eta_u]^\beta & \text{if } q \leq q_0 \\ J & \text{if } q > q_0 \end{cases}$$

where η_u represents the heuristic information, and J is chosen with the probability:

$$p_j^k(t) = \frac{[\tau_j(t)][\eta_j]^\beta}{\sum_{l \in N^k}[\tau_l(t)][\eta_l]^\beta}$$

where N^k is a possible neighbor of ant k.

After each selection of a column j, occurs a local modification of the level of pheromone of that column, given by the equation:

$$\tau_j(t) = (1 - \rho)\tau_j(t) + \rho\tau_0$$

This evaporation is done so the visited column will not be interesting for the following ants, stimulating in this way the exploration of solutions. At the equation above, $\rho\tau_0$ is a stabilization factor of the pheromone modification, used with the intention of not having less attractive columns so quickly, allowing the exploration of a number even bigger of new solutions. When every ant ended an iteration (a solution has been found), the pheromone level of the best solution found is updated globally, loading pheromone to each column of the solution according to:

$$\tau_j(t) = (1 - \rho)\tau_j(t) + \rho\triangle\tau_j(t)$$

where $\triangle\tau_j(t)$ is the variation of pheromone left on column j by the best ant. This variation is calculated as the frequency of column j in the routing of each ant, i.e. the number of times where column j is in the solutions found by the ants. To operate, the following parameters controlled by the GA will be defined: *Number of Ants*, ρ (the evaporation factor), β (the importance in the choice of the next column, *Length of the List* (which will be used to limit the number of columns that can be visited next), q_0 (parameter which indicates if the exploration is supported at the moment of next column election) and *MaxIter* (max number of iterations).

4.5 Penalizing function

An aid to the Parameter setting function is needed so the Genetic Algorithm can handle a trade-off between keep improving a solution — ergo more resources applied — or to quit a solution and try another search. A penalizing function is a very good solution to this problem. The direct comparison is between a Parameter Tuning version and a Parameter Control one. As a penalizing function we propose:

$$fitness(sol) = ObjFuncValue(sol) + TimeTaken(sol) * FCT$$

where *sol* is a solution and $ObjFuncValue$ is the value — or cost — of the solution generated evaluated at the SCP Objective Function of the corresponding benchmark. Also, $TimeTaken$ is the amount of time taken to generate that solution. FCT is a correction factor which makes it possible to compare the time with the cost.

4.6 Parameter Control Considerations

Designs introduced in sections 4.1 and 4.4 were directly coded into the Parameter Tuning versions. To obtain the counterparts to the Parameter Control versions, some changes were introduced between transitions of the metaheuristics. GA was modified to enabling an intermediate results memory taking feedback of the performance of the lower level metaheuristic.

5 Experimental Results

5.1 Performance Measurement

The performance measures evaluated will are:

- Best Solution Found (value at SCP's Objective Function)
- Time Taken to find the Best Solution (best represented by "Calls to Objective Function made before best solution is found")
- Average Fitness (penalized) and its Standard Deviation

The quality of the solutions is taken measuring its closeness to the Optimal value of a certain instance (OR-Library SCP file). The average fitness is also an aid to evaluating quality in combination with the Standard Deviation. The computational effort is measured by: the number of calls to objective function needed to generate the best solution.

5.2 Reference Benchmarks

Each implementation — i.e. GA-SS and GA-ACO — are tested using the benchmarks provided by [17] which are widely used by Operation and Optimization Researchers. The files used are Set Covering Problem instances denoted by series SCP4x, SCP5x and SCP6x. To obtain more experimental results — and to compare the performance of each implementation — such benchmarks will be tested against both versions of parameter setting configurations. The benchmarks presented are minimization problems, where the idea is to find the lowest cost at the Objective Function. Each benchmark is based on a matrix of 200 rows and 1000 columns, which makes them a fairly large set of problems.

5.3 Environment

All the algorithms required for testing will be coded under standard C language and compiled with GNU GCC. The Xcode 3.2 IDE was selected for coding and depuration task, and as reference computer an Intel Core 2 Duo CPU with 4Gb of DDR3 RAM was used. To notice, testing implementation code does not allow the use of multiples cores as is the case of current CPU.

5.4 Robustness based on GA's parameters

The first table of this section shows the results for GA-SS with Parameter Tuning. These results are important since accomplish two objectives: to allow to evaluate the measures at this offline parameter setting configuration and to represent the execution of several single SS metaheuristic executions. As was explained before, a Parameter Tuning configuration runs a whole instance of SS metaheuristic with parameters defined — by GA — before its execution. When the SS execution finish, the best result is given as feedback to GA, which might continue its process of experimenting with other SS parameters. Hence, the results obtained consider a neutral parameter configuration — a random one — representing a human trial-and-error testing, where best results might be as good as the ones with GA-SS, but surely requiring more Human work in terms of defining those parameters and to experiment with them.

PXOVER	PMUT	BestFound	CallsOF	Average	Fit-StDev
0.25	0.25	1009	5778578	3017	433
0.5	0.5	1009	4698886	2915	675
0.75	0.75	1007	5750974	3104	571
0.25	0.75	1008	4545205	2803	458
0.75	0.25	1009	4698886	2644	578

Table 1: Different GA-SS parameters configurations using Parameter Tuning. Benchmark: SCP41.

First results (Table 1) show a considerable variability. The overall look represent what a human can expect when experimenting — no good luck considerations. As will be seen in the next tables, *PXOVER* and *PMUT* tend to lose its impact, being this instance the most unexpected.

PXOVER	PMUT	BestFound	CallsOF	Average	Fit-StDev
0.25	0.25	509	3446206	2235	1075
0.5	0.5	509	3462702	2329	1078
0.75	0.75	509	3481599	2002	954
0.25	0.75	509	3502315	1735	797
0.75	0.25	509	3308283	1980	890

Table 2: Different GA-SS parameters configurations using Parameter Control. Benchmark: SCP41.

At the second table (Table 2) best results found show no variation, mainly due to SS Improvement Function. The other measures show a low standard deviation specially in the overall calls to Objective Function. As the average solution results are the ones with more variation, there is logic in thinking that there was a good level of exploration between solution spaces.

In the next table (Table 3) the performance with ACO is evaluated, using Parameter Tuning. Aside of the better results obtained with ACO — and its best resources management — the same fashion as in GA-SS is seen: very low variation at each measure.

PXOVER	PMUT	BestFound	CallsOF	Average	Fit-StDev
0.25	0.25	449	1477894	781	236
0.5	0.5	434	1095816	768	240
0.75	0.75	449	1659306	896	274
0.25	0.75	449	1599972	862	260
0.75	0.25	449	1435961	762	220

Table 3: Different GA-ACO parameters configurations using Parameter Tuning. Benchmark: SCP41.

Using an online parameter setting criteria (Table 4) shows no more variation at the results; actually, a very low variation is obtained. Is clearer that the results under this configuration are the best of all. This will be reviewed in following experiments show down.

PXOVER	PMUT	BestFound	CallsOF	Average	Fit-StDev
0.25	0.25	436	673804	442	7
0.5	0.5	436	691756	442	8
0.75	0.75	436	743647	444	8
0.25	0.75	436	671342	444	8
0.75	0.25	436	703318	441	7

Table 4: Different GA-ACO parameters configurations using Parameter Control. Benchmark: SCP41.

5.5 Convergence to the best solution

This experiments were selected between the most interesting results: which presented a noticeable convergence through time. At first look, the temptation is to compare the obvious; GA-ACO performed better than its counterparts. After a careful review, the comparison is unfair fundamentally because ACO uses a more "intelligent" way of building solutions whereby SS is more blind. A preliminary conclusion is that the constructive process of ACO is better than the evolutive of SS. Anyway, the purpose of this graphics is to allow to observe how each version converges through time to a better solution.

The plot is based on benchmark SCP48 and the clearest idea is that Parameter Control allows to get better results in lesser time, which is mainly due to the intensification of the harvest of certain solution space where "best looking" results are found, and to the saved time when quitting to explore "bad looking" results. Also, a common item in each plot is that GA-SS (Tuning) tent to obtain a better result than GA-SS (Control) when a great quantity of solutions are searched.

5.6 Wide comparison of best results found

Following table introduces the best results found through the various instances executed. Faster convergence of Parameter Control versions has been considered in previous section — which is not represented in this table. Same results obtained by both versions of ACO hide evidently the efficiency factor. Every instance executed to obtain this results was using $PXOVER =$

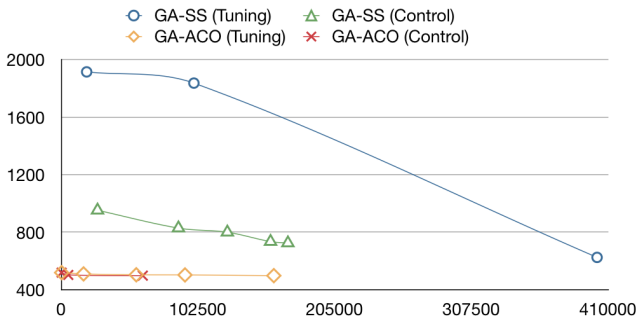


Figure 2: Convergence analysis to a better solution. Benchmark: SCP48

0.5 and $PMUT = 0.5$ which seems a fair tradeoff between crossover and mutation at GA. Under this circumstances, another criteria of robustness is handled: to perform well in a variety of instances. The worse performer through all experiments was GA-SS with Parameter Tuning, but, even with this issue, it was not an extremely bad perform.

	SCP41	SCP42	SCP48	SCP61	SCP62	SCP63
Optimum	429	512	492	138	146	145
GA-SS-Offline	1007	981	561	154	155	166
GA-SS-Online	509	603	642	164	165	176
GA-ACS-Offline	434	529	497	142	154	148
GA-ACS-Online	434	529	497	142	154	148

Table 5: Table of best results.

6 Conclusions

A 2-level metaheuristic has been tested on different SCP benchmarks showing to be very effective. One of the main goals of a multilevel approach is to provide an unattended solving method, for quickly producing solutions of a good quality for different instances. The overall work was extended on several relevant issues and quality measures: quality of solutions, robustness in terms of the instances and insensitivity to small deviation at parameters (GA parameters) solving large-scale problems, easiness of implementation, easiness to combine with other algorithms, automatic setting of parameters. All of them provided a fair approach to the core problem, and many of them have already shown lights of been covered, producing a renovated energy for keep working and admiring the synergy produced. Benchmarks have shown interesting results in terms of robustness of every approach being the Parameter Control the most robust in terms of a good performance in several instances using same parameters and in terms of stable results when small deviations are made to parameters. Also, Parameter Control shown to converge faster to better results than Parameter Tuning, but, when long running times are taken, both seem to obtain equal results. Considering the particular performance of a Constructive approach and an Evolutive one for the SCP, it seems to be more effective to "construct" intelligently a solution rather than "blindly" get one and then evolve it. An interesting fact is that SS with Parameter Control might get a performance closer to ACO. The overall implementation of a multilevel metaheuristic is pretty straightforward, no big obstacles were found. Also, the implementation of the GA using real numbers was a direct representation of what parameters a human user might

choose to operate, therefore achieving an implementation which accomplished our requirements.

Bibliography

[1] B. Crawford, C. Lagos, C. Castro, F. Paredes, A Evolutionary Approach to Solve Set Covering, *ICEIS 2007 - Proceedings of the Ninth International Conference on Enterprise Information Systems, Volume AIDSS, Funchal, Madeira, Portugal, June 12-16, 2007 (2)*, pp.356-363, 2007

[2] U. Aickelin, An Indirect Genetic Algorithm for Set Covering Problems, *Journal of the Operational Research Society*, Vol.53, pp.1118-1126, 2002

[3] F. Tangour, P. Borne, Presentation of Some Metaheuristics for the Optimization of Complex Systems, *Studies in Informatics and Control*, Vol.17, No.2, pp.169-180, 2008

[4] C-M. Pintea, D. Dumitrescu, The importance of parameters in Ant Systems, INT J COMPUT COMMUN, ISSN 1841-9836, 1(S):376-380, 2006

[5] R. Martí, M. Laguna, Scatter Search: Diseño Básico y Estrategias, *Revista Iberoamericana de Inteligencia*, Vol.19, pp.123-130, 2003

[6] D. Gouwanda, S. G. Ponnambalam, Evolutionary Search Techniques to Solve Set Covering Problems, *World Academy of Science, Engineering and Technology*, Vol.39, pp.20-25, 2008

[7] A. Caprara, M. Fischetti, P. Toth, Algorithms for the Set Covering Problem, *Annals of Operations Research*, Vol.98, 1998

[8] J. E. Beasley, K. Jornsten, Enhancing an algorithm for set covering problems, *European Journal of Operational Research*, Vol.58, pp.293-300, 1992

[9] C. Cotta, M. Sevaux, K. Sörensen, *Adaptive and Multilevel Metaheuristics*, Springer, 2008

[10] Z. Michalewicz, *Genetic algorithms + data structures = evolution programs*, Springer, 1996.

[11] F. Glover, G. A. Kochenberger, *Handbook of metaheuristics*, Springer, 2003

[12] B. Crawford, C. Castro, Integrating Lookahead and Post Processing Procedures with ACO for Solving Set Partitioning and Covering Problems, *Proceedings of ICAISC*, pp.1082-1090, 2006

[13] Y. Hamadi, E. Monfroy, F. Saubion, What is Autonomous Search?, *Technical Report MSR-TR-2008-80*, 2008

[14] L. Lessing, I. Dumitrescu, T. Stützle, A Comparison Between ACO Algorithms for the Set Covering Problem, it Proceedings of ANTS, pp.1-12, 2004

[15] E. Talbi, *Metaheuristics: From Design to Implementation*, Wiley Publishing, 2009

[16] R. Battiti, M. Brunato, F. Mascia, *Reactive Search and Intelligent Optimization*, Springer Verlag, 2008

[17] J. E. Beasley, *OR Library, http://people.brunel.ac.uk/ mastjjb/jeb/info.html*

INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 2 (June), pp. 388-395

A Reliability Level List based SDD Algorithm for Binary Cyclic Block Codes

B. Yamuna, T.R. Padmanabhan

B.Yamuna
Assistant Professor,Dept of ECE
Amrita Vishwa Vidyapeetham,
Amrita School of Engineering
Amrita Nagar, Coimbatore. 641 112
Tamil Nadu, India
E-mail: b_yamuna@cb.amrita.edu

Dr. T.R.Padmanabhan
Professor Emeritus, Dept of IT
Amrita Vishwa Vidyapeetham,
Amrita School of Engineering
E-mail: trp@amrita.edu

Abstract: Soft decision decoding (SDD) provides a better coding gain by making use of the unquantized channel output. In this paper we introduce the concept of a Reliability Level List (RLL); based on the RLL a new SDD algorithm for Binary Phase Shift Keying (BPSK) based binary cyclic block codes is proposed. The algorithm guarantees to extract the most reliable codeword in an iterative manner. The formation of the RLL involves a search for the next possible entry into the RLL based on the error probability which is a reflection of the reliability values of the bits of the received word obtained from the channel. The procedure for the formation of RLL which is the central idea of the paper is given as a structured algorithm.
Keywords: cyclic block codes, reliability based decoding, soft decision decoding, probability of error.

1 Introduction

Soft decision decoding (SDD) algorithms in general focus on the extraction of the codeword from the received word by making use of the reliability information available at the channel output which is otherwise discarded in hard decision decoding. SDD algorithms have increased complexity compared to hard decision decoding but they trade off complexity for better error performance.

The idea of exploiting the real channel output to the maximum possible extent instead of quantizing it to 0's and 1's as with hard decision decoding has been the essence of SDD vis-â-vis binary cyclic block codes. Different SDD algorithms have been proposed in the recent years-each with its own trade off between complexity and error performance[1- 3].

The reliability based SDD algorithms aim at decoding a received word which is more reliable even under low signal-to-noise ratio (SNR) conditions. Researchers are focusing on making the best use of the channel output and improving decoding reliability. In these algorithms the channel output is used to quantify reliability values of the bits of the received word. The bits are processed with respect to their reliability values. They can be based on processing the bits from the least reliable end as with Chase decoding algorithms and Generalized-minimum distance (GMD) decoding algorithms [4], [5] or they can be based on processing from the most reliable

end as with Ordered Statistics Decoding algorithms[6].

A SDD algorithm is evolved in this paper. The algorithm uses the channel output reliability in the true sense. The significance of the work is in the fact that the target codeword identified through the algorithm here is the best that reliability based SDD algorithm can lead to.

The method involves searching in a space of 2^n words for the most reliable codeword. But the extent of actual search is confined to a much shorter range as brought out in the paper. The focus of the paper is two-fold: (1) Proving that the codeword extracted is the best possible one from a soft decision point of view and (2) giving a structured algorithm for the extraction of the codeword.

2 Preliminaries of the RLL based SDD Algorithm

The problem of SDD is one of starting with the received word and identifying a codeword which will be the most reliable one; this codeword is called the 'target codeword' here. The process essentially amounts to arranging all the possible 2^n binary bit sequences in the order of increasing reliability and picking the first codeword from it. The sequence so arranged is called the Reliability Level List (RLL) here. Hence RLL represents the search sequence to be followed to find the most reliable codeword.

For an n bit sequence let the pair $\{(b_i,\ m_i)\}$, $0 \leq i \leq n$ represent the sets of the bit values and the corresponding reliability magnitudes. Each bit of hard decision has an associated probability of error given as $\int_{m_i}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$.

Let $Q(\frac{m_i}{\sigma})$=$\int_{m_i}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$, σ being the noise variance. Arrange the set $\{Q(\frac{m_i}{\sigma})\}$ in descending order of magnitudes and assign integer values k to them such that k =1 for the bit with the largest value of $Q(\frac{m_i}{\sigma})$; let $q[1]$ represent this $Q(\frac{m}{\sigma})$ value. Similarly $k = 2$ for the next one with $q[2]$ representing the corresponding $Q(\frac{m}{\sigma})$ value and so on until $k = n$ for the bit with the smallest value of $Q(\frac{m_i}{\sigma})$. Let $\{q[k]\}$ be the rearranged set of numbers and $\{p[k]\} = \{1 - q[k]\}$.

With an (n,k) code all the 2^k possible codewords are included in the total space of 2^n words. All these codewords appear in the RLL each with its own associated probability of being the target codeword; the very first codeword to appear in the list being the most reliable of the codewords is the target codeword; those further down the RLL being less reliable need not be examined. The first few entries in RLL as defined above can be seen to be as given below:

- The topmost entry is the term $\prod_{k=1}^{15} p[k]$ having the largest magnitude. This represents the received word itself. i.e., none of the bits is in error.
- The next entry in RLL is the product $q[1]\prod_{k=2}^{15} p[k]$ having the next higher magnitude. This represents the received word with the least reliable bit with k =1 being in error. Hence the received word with the least reliable bit flipped is examined and if found to be a codeword it is the target codeword and the search stops. If it is not a codeword then the subsequent entry in RLL has to be made and the process of examining the entry for identification of target codeword is to be continued. The process is thus repeated until target codeword is identified. The third and fourth entries are given below for clarity.
- The third entry in RLL is the term $q[2]p[1]\prod_{k=3}^{15} p[k]$
- The fourth entry in RLL is either $q[3]p[2]p[1]\prod_{k=4}^{15} p[k]$ or $p[3]q[2]q[1]\prod_{k=4}^{15} p[k]$ based on whichever candidate has the higher magnitude. In general the position of a word in RLL is decided by the magnitude of the corresponding product

$$\prod_{k_i} p[k_i] \prod_{k_j} q[k_j] \qquad (1)$$

where k_i and k_j represent the set of k values for which the $p[k_i]$ and $q[k_j]$ values respectively appear in the product. This is the crux of selecting the next entry in to the RLL at every stage. With$s[k] = \frac{q[k]}{p[k]}$ define

$$f = \frac{\prod_{k_j} q[k_j]}{\prod_{k_i} p[k_i]} \tag{2}$$

Using f the product $\prod_{\tau=1}^{n} P[\tau] \times f$ can be used in place of the product in (1) above. This directly leads to the following lemma.

Lemma 1:

Let v represent the rank of an entry in RLL and f_v the corresponding f value as defined in equation (2). Then the v^{th} entry in RLL is such that $f_{v-1} \geq f_v \geq f_{v+1}$ for $0 \leq v \leq n$

The above inequality means that the problem of identifying the rank of an entry in the RLL is the same as computing the f_v's and arranging them in descending order of magnitudes. Since log f_v is a monotonically increasing function of f, the inequality implies log $f_{v-1} \geq$ logf_v $\geq$ log f_{v+1} for $0 \leq v \leq n$.

If an entry in RLL is known its log f_v is known. The problem of deciding the very next entry can be looked upon as the problem of identifying a set of log $s[k]$ such that the corresponding Σ log $s[k]$ satisfies the following two conditions

1. It should be smaller than the Σlog $s[k]$ for the present entry.
2. It should be the largest in magnitude amongst the set yet to be entered in the RLL.

With $W = \Sigma$ log $s[k]$, an entry with rank v in RLL is characterized by its W_v. With this the above result can be formalized as a theorem:

Theorem 1:

For any $v \in \{0, 2^{n-1}\}$ the corresponding entry in RLL has a W_v such that $w_{v-1} \geq w_v \geq w_{v+1}$

Log $s[k]$ being a negative quantity, $W = \Sigma$log $s[k]$ is also a negative quantity. It is more convenient to work with log $(\frac{1}{s[k]})$ = (-log $s[k]$) and the corresponding sum M = -Σ log $s[k]$. With this, Theorem 1 implies Lemma 2 below which is more convenient to work with.

Lemma 2:

For any $v \in \{0, 2^{n-1}\}$ the corresponding entry in RLL has a M_v such that $M_{v-1} \leq M_v \leq M_{v+1}$

3 Procedure for Forming the RLL

The procedure involves forming sequences of n bit binary numbers which would be the entries in the RLL. Let v denote the rank of an entry in the list with v ranging from 0 to 2^n. The step by step procedure for making the entries in the RLL is follows.

1. M_0, M_1, M_2, where $M_0 = 0, M_1$ = - log $s[1]$, M_2 = - log $s[2]$ are formed as explained earlier in section 2 and the corresponding binary sequences designated N_0, N_1, N_2 are entered in the RLL.
2. Define a 'pending list' - P_L as the collection of all contenders for the next entry to the RLL.
3. The entries in the P_L are characterized by the following.
   (a) Each entry is a collection of indices.
   (b) Each entry has an associated characteristic magnitude - M.

4. Once the P_L is fully known, the next entry to the RLL can be decided by comparing the M values for each entry in the P_L. It also implies that the left over elements in the P_L will also be members for the P_L for the subsequent entry in the RLL.

5. Once an entry to the RLL is decided, the P_L for the next entry to the RLL is formed through the following steps.

   (a) All the entries left over after the previous entry to the RLL are also to be in the new P_L.

   (b) Additional sets are to be added to the new P_L depending on the present entry to the RLL. This essentially amounts to scanning the indices in the set for the present entry and altering them. A perusal of the set of indices shows that the additional element sets can be formed as follows:

      i. If the least index K_1 for the set is not 1 add 1 to the index set.

      ii. For every index X_i in the set if $X_i+1 < X_{i+1}$ replace X_i with X_i+1. This amounts to adding a set to the P_L where the contribution to M by X_i is increased by the least possible amount in the neighborhood of X_i.

The procedure to form the RLL as a sequence of 2^n binary numbers explained above can be cast as structured algorithm. The list of notations used in the algorithm is given below:

- K: reliability index assigned to the bits of the received word.
- N_v : binary number in RLL with rank v.
- $s[k] = \frac{q[k]}{1-q[k]}$
- n : block length of the codeword
- P_L : Pending list after updating
- P_u : pending list before updating
- T_i : i^{th} set in P_L
- J_{max} :total number of T_i's in P_L
- β_i: number of elements in T_i
- F : flag
- U_i, μ, Δ: Temporary symbols used for T_i's, U_i , and J_{max} respectively

4 RLL based SDD Algorithm

1. **Input**: {-log $s[k]$}, n.
2. **Output**: Sequential entries in RLL
3. **Initial Condition**: $N_0 = \{0\}$, $N_1 = \{1\}$, $N_2 = \{2\}$, $P_L = \{T_1, T_2\}$ where $T_1 = \{3\}$, $T_2 = \{1, 2\}$,$J_{max} = 2$, $F = 0$.
4. Do For v=3 to 2^n

5. Do For i=1 to J_{max}
   Compute $M_{v_i} = -\sum_{\alpha=1}^{\beta_i} log(s[X_\alpha])$
   EndDo i

6. M_{min} = min{ M_{v_i} for i = 1 to J_{max}}

7. Let i_m = the i for which M_{v_i} is minimum

8. Form a binary sequence N_v whose 1 bits are decided by the indices in the set T_{im}

9. $\beta_v = \beta_{im}$

10. Do for i=1 to J_{max}
    If($i \neq i_m$) $U_i = T_i$
    EndDo i
    (Removing T_{im} from the pending list and reassigning the T_i's to the U_i's.)

11. $P_u = \{0\}$

12. Do for $i = i$ to i_{m-1}
    $P_u = \{ P_u : U_i\}$
    EndDo i

13. Do for $i = i_m$ to J_{max}

    (a) $U_i = U_{i+1}$

    (b) $P_u = \{ P_u : U_i\}$

    (c) $\beta_i = \beta_{i+1}$

    EndDo i

14. In N_v (Updating the pending list by forming candidate entries; processing done with present entry)

15. Do for d = 1 to n

    (a) If $K_1 \neq 1$, (If LSB of N_v is 0 add 1 to the present entry)
    {
    i. $\mu = \{ N_v : 1\}$
    ii. if μ is not in P_u ,
    iii. $U_i = \mu$ for $i = J_{max}$
    iv. $\Delta = J_{max}$
    v. $\beta_{J_{max}} = \beta_{J_{max}} + 1$
    vi. $F = 1$
    }

    (b) If $K_d + 1 < K_{d+1}$
    {
    If F =1
    {
    i. Form μ with K_d replaced by K_d +1 in N_v
    ii. If μ is not in P_u ,

iii. $U_{\Delta+1} = \mu$
iv. $\beta_{\Delta+1} = \beta_v$
}
Else
{

15.2.5 Form μ with K_d replaced by K_d +1 in N_v
15.2.6 If μ is not in P_u
15.2.7 $U_\Delta = \mu$
15.2.8 $\beta_\Delta = \beta_{J_{max}}$

}

16. Increment Δ
}
17. $J_{max} = J_{max}$ +1
18. EndDo d
19. $P_L = P_u$ for i =1 to Δ (Pending list updated)
20. $F = 0$.
21. EndDo v

5 Example

Extensive simulation with a variety of binary cyclic block codes like (15, 7), (31, 16), and (127, 64) have been carried out; in each case a few thousand transmissions over an AWGN channel was done for different values of σ and the erroneous received words decoded using the proposed algorithm. In all the cases the resultant target codeword was found to be the most reliable one. Details of a representative set are discussed briefly here.

A (15, 7) binary cyclic block code with $d_{min} = 5$ and t $= \lfloor d_{min}/2 \rfloor = 2$ is considered. The message 030h has been encoded as 304eh and transmitted over an AWGN channel with σ= 0.8. The received word is 78cch. There are 4 errors at bit positions 1, 7, 11, 14. With the input as the set {-log $s[k]$}, $n = 15$ and the Initial Condition as:$N_0 = \{0\}$ $N_1 = \{1\}$, $N_2 = \{2\}$, $P_L = \{T_1, T_2\}$ where $T_1 = \{3\}$, T_2= {1, 2},$J_{max} = 2$, $F = 0$, following the steps of the algorithm the RLL was formed. A representative segment of the RLL is given in Table 1. The table gives the rank v, k value, M_{min} , and the contents of T_i with the corresponding M_{v_i} (both separated by a hyphen in the table).

As seen in Table 1 in each row the specific T_i in italics has the minimum value of M_{v_i} and forms the candidate identified for the next entry. Further in each row the entries in bold represent the updated candidates that are formed by applying the rules in row 15 in the algorithm given above in section 4. For example considering the row with rank 82, the entry T_{36}= {*2,3,6,7 - 0.977*} - with k=2, k = 3, k = 6, k = 7 as 1ι's and Σ log $s[k]$ as 0.977 - is the next entry with rank as 83 ; and this is the one with minimum of M_{v_i}. The entries in the same row with rank 82, namely T_{38} = {**1,4,6,7 - 1.198**} and T_{39} = {**1,3,6,8 - 1.4992**}are the ones formed on applying the rules in rows 15.1 and 15.2.1 of the algorithm given in section 4 respectively.

Table 1: Partial RLL

Rank v	K	M_{min}	Details of T_i's in the Updated pending list P_L
1	1	0.0832	
2	2	0.0957	T_1= {*3 - 0.1223*}, T_2= {1,2 - 0.1789}
3	3	0.1223	P_L= { T_1, T_2, T_3}; T_1= {*1,2 -0.1789*}, T_2= {**1,3 -0.2055**}, T_3= {**4,-0.3558**}
4	1,2	0.1789	T_1= {*1,3 - 0.2055*}, T_2= {4 - 0.3558}
5	1,3	0.2055	T_1= {4 - 0.3558}, T_2= {**2,3 - 0.218**}, T_3= {**1,4 - 0.439**}
.	.	.	.
.	.	.	.
82	1,3,6,7	0.9645	T_1= {1,8 - 0.9978}, T_2= {2,8 - 1.0103}, T_3= {3,8 - 1.0369}, T_4= {1,2,8 - 1.0935}, T_5= {1,3,8 - 1.1201}, T_6= {2,3,8 - 1.1326}, T_7= {1,2,3,8 - 1.2158}, T_8= {4,8 - 1.2704}, T_9= {5,8 - 1.2727}, T_{10}= {6,8 - 1.2937}, T_{11}= {1,4,8 - 1.3536}, T_{12}= {1,5,8 - 1.3559}, T_{13}= {2,4,8 - 1.3661}, T_{14}= {2,5,8 - 1.3684}, T_{15}= {1,6,8 - 1.3769}, T_{16}= {2,6,8 - 1.3894}, T_{17}= {3,4,8 - 1.3927}, T_{18}= {4,5,6 - 1.093}, T_{19}= {4,5,7 - 1.0938},T_{20}= {3,5,8 -1.395}, T_{21}= {4,6,7 - 1.1148}, T_{22}= {3,6,8 - 1.416}, T_{23}= {1,2,4,8 - 1.4493}, T_{24}= {1,2,5,8 - 1.4516}, T_{25}= {1,2,3,4,5 - 1.0151}, T_{26}= {1,2,6,8 - 1.4726}, T_{27}= {1,9 - 1.0262}, T_{28}= {10 - 1.0421}, T_{29}= {1,2,3,4,6 - 1.0361}, T_{30}= {1,2,3,4,7 - 1.0369}, T_{31}= {2,3,4,8 - 1.4884}, T_{32}= {1,2,3,5,6 - 1.0384}, T_{33}= {2,4,5,6 - 1.1887}, T_{34}= {1,2,3,5,7 - 1.0392}, T_{35}= {2,4,5,7 - 1.1895}, T_{36}= *{2,3,6,7 - 0.977}*, T_{37}= {2,3,5,8 - 1.4907}, **T_{38}= {1,4,6,7 - 1.198 }, T_{39}= {1,3,6,8 - 1.4992 }**
83	2,3,6,7	0.977	T_1= {1,8 - 0.9978}, T_2= {2,8 - 1.0103}, ...T_{37}= {1,4,6,7 - 1.198 }, T_{38}= {1,3,6,8 - 1.4992 },**T_{39}={1,2,3,6,7 - 1.0602}, T_{40}={2,4,6,7 - 1.2105}, T_{41}={2,3,6,8 - 1.5117}.**

With each such entry to the RLL the corresponding candidate word is formed by complementing the bits indicated by the entry. The candidate word is examined to check if it is a codeword; if yes it is the desired target codeword. Once the target codeword is identified the process of filling the RLL can be discontinued since any codeword found below this in the RLL is less reliable. For the specific case here the target code word is obtained at the rank of 83 with errors at bits 7, 1, 11, and 14 with the corresponding K values as 2, 3, 6, and 7 respectively.

6 Conclusion

A structured, iterative, and reliability based Soft Decision Decoding algorithm is proposed. The algorithm uses the reliability value of the received word as a soft metric and outputs the desired target codeword through an ordered search in the 2^n word space. The algorithm identifies an entry in the RLL such that the corresponding word is the most reliable yet. For each entry in the RLL the corresponding word formed has to be checked to see if it is a code word and if so it is the desired target codeword. The proposed algorithm guarantees to return *the best* in terms of reliability because of the very fact that any other codeword found below the identified one will be less reliable. The structured algorithm can be easily coded and used for decoding any (n, k) binary cyclic block code.

Bibliography

[1] Wenyi Jin and Marc.P.C.Fossorier,Fellow,IEEE, Reliability-Based Soft-Decision Decoding with Multiple Biases , *IEEE Trans. Inform. Theory*, Vol. 53, pp. 105-120, Jan. 2007

[2] Ye Liu, Member, IEEE, Shu Lin, Fellow, IEEE, and Marc.P.C.Fossorier,Fellow,IEEE, MAP algorithms for Decoding linear block codes based on sectionalized Trellis diagrams, *IEEE Trans. Communications*, Vol. 48, pp. 577-587, Apr. 2000

[3] Yuansheng Tang, Member, IEEE, San Ling and Fang - WeiFu, On the Reliability - Based Soft - Decision Decoding Algorithms for Binary Linear Block Codes, *IEEE Trans. Inform. Theory*, Vol. 52, pp. 328-335, Jan. 2006

[4] David Chase, Member, IEEE, A Class of Algorithms for Decoding Block Codes With Channel Measurement Information, *IEEE Trans. Inform. Theory*, Vol. IT-18, pp.170-182, Jan 1972

[5] G. D. Forney, Jr., Generalized minimum distance decoding, *IEEE Trans. Inform. Theory*, Vol. IT-12, pp. 125-131, Apr. 1966

[6] Marc P. C. Fossorier, Member, IEEE, and Shu Lin, Fellow, IEEE, Soft-Decision Decoding of Linear Block Codes based on Ordered Statistics,*IEEE Trans. Inform. Theory*, Vol. 41, pp. 1379-96, Sep 1995

Author index