

Industrial

Electronics

Biomedical

Civil

Aerospace

Computer

Electrical

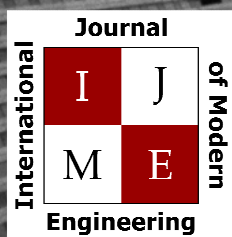
Chemical

Mechanical



ENGINEERING





[www.ijme.us](http://www.ijme.us)

Print ISSN: 2157-8052  
Online ISSN: 1930-6628



[www.iajc.org](http://www.iajc.org)

## INTERNATIONAL JOURNAL OF MODERN ENGINEERING

### ABOUT IJME:

- IJME was established in 2000 and is the first and official flagship journal of the International Association of Journal and Conferences (IAJC).
- IJME is a high-quality, independent journal steered by a distinguished board of directors and supported by an international review board representing many well-known universities, colleges and corporations in the U.S. and abroad.
- IJME has an impact factor of **3.00**, placing it among the top 100 engineering journals worldwide, and is the #1 visited engineering journal website (according to the National Science Digital Library).

### OTHER IAJC JOURNALS:

- The International Journal of Engineering Research and Innovation (IJERI)  
For more information visit [www.ijeri.org](http://www.ijeri.org)
- The Technology Interface International Journal (TIIJ).  
For more information visit [www.tiij.org](http://www.tiij.org)

### IJME SUBMISSIONS:

- Manuscripts should be sent electronically to the manuscript editor, Dr. Philip Weinsier, at [philipw@bgsu.edu](mailto:philipw@bgsu.edu).

For submission guidelines visit  
[www.ijme.us/submissions](http://www.ijme.us/submissions)

### TO JOIN THE REVIEW BOARD:

- Contact the chair of the International Review Board, Dr. Philip Weinsier, at [philipw@bgsu.edu](mailto:philipw@bgsu.edu).

For more information visit  
[www.ijme.us/ijme\\_editorial.htm](http://www.ijme.us/ijme_editorial.htm)

### INDEXING ORGANIZATIONS:

- IJME is currently indexed by 22 agencies.  
For a complete listing, please visit us at [www.ijme.us](http://www.ijme.us).

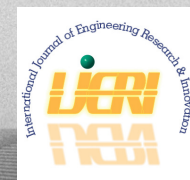
### Contact us:

**Mark Rajai, Ph.D.**

Editor-in-Chief  
California State University-Northridge  
College of Engineering and Computer Science  
Room: JD 4510  
Northridge, CA 91330  
Office: (818) 677-5003  
Email: [mrajai@csun.edu](mailto:mrajai@csun.edu)



[www.tiij.org](http://www.tiij.org)



[www.ijeri.org](http://www.ijeri.org)

---

# INTERNATIONAL JOURNAL OF MODERN ENGINEERING

The INTERNATIONAL JOURNAL OF MODERN ENGINEERING (IJME) is an independent, not-for-profit publication, which aims to provide the engineering community with a resource and forum for scholarly expression and reflection.

IJME is published twice annually (fall and spring issues) and includes peer-reviewed articles, book and software reviews, editorials, and commentary that contribute to our understanding of the issues, problems, and research associated with engineering and related fields. The journal encourages the submission of manuscripts from private, public, and academic sectors. The views expressed are those of the authors and do not necessarily reflect the opinions of IJME or its editors.

## EDITORIAL OFFICE:

Mark Rajai, Ph.D.  
Editor-in-Chief  
Office: (818) 677-2167  
Email: [ijmeeditor@iajc.org](mailto:ijmeeditor@iajc.org)  
Dept. of Manufacturing Systems  
Engineering & Management  
California State University-  
Northridge  
18111 Nordhoff Street  
Northridge, CA 91330-8332

## THE INTERNATIONAL JOURNAL OF MODERN ENGINEERING EDITORS

*Editor-in-Chief:*

**Mark Rajai**

California State University-Northridge

*Associate Editors:*

**Alok Verma**

Old Dominion University

**Li Tan**

Purdue University North Central

*Production Editor:*

**Philip Weinsier**

Bowling Green State University-Firelands

*Subscription Editor:*

**Morteza Sadat-Hossieny**

Northern Kentucky University

*Financial Editor:*

**Li Tan**

Purdue University North Central

*Executive Editor:*

**Paul Wilder**

Vincennes University

*Manuscript Editor:*

**Philip Weinsier**

Bowling Green State University-Firelands

*Copy Editor:*

**Li Tan**

Purdue University North Central

*Publisher:*

**International Association of Journals and Conferences**

*Web Administrator:*

**Saeed Namyar**

Advanced Information Systems

---

# TABLE OF CONTENTS

<i>Editor's Note: Invitation to Attend the 4th IAJC/ISAM Joint International Conference</i> .....	3
<i>Philip Weinsier, IJME Manuscript Editor</i>	
<i>Ultra-compact Graphene-based Electro-optic Modulators on a Silicon-on-insulator Platform</i> .....	5
<i>Chenran Ye, The George Washington University; Sarah Pickus, The George Washington University; Ke Liu, The George Washington University; Volker J. Sorger, The George Washington University</i>	
<i>Optimal Tuning of PI Controllers Subject to Process Constraints: Experimental Evaluation</i> .....	15
<i>Constantine Tzouanas, Clear Lake High School; Vassilios Tzouanas, University of Houston–Downtown</i>	
<i>Altering the Density of Ferrofluids Through the use of Magnetic Fields: A Proof-of-concept Study for the Post-consumer Plastic Recycling Industry</i> .....	22
<i>Matthew Franchetti, University of Toledo; Andrew Moening, University of Toledo; Connor Kress, University of Toledo</i>	
<i>An Autonomous Passive RFID-assisted Mobile Robot System for Indoor Positioning</i> .....	28
<i>Kumar Yelamarthi, Central Michigan University</i>	
<i>Implementation of Robot Joint Trajectory Control: An Applied Method</i> .....	38
<i>Shaopeng (Frank) Cheng, Central Michigan University</i>	
<i>Adaptive Learning Factor of Backpropagation in FeedForward Neural Networks</i> .....	47
<i>Hong Li, The City University of New York</i>	
<i>Fluid Structure Interaction of Traumatic Brain Injury: Effects of Material Properties on SAS Trabeculae</i> .....	54
<i>Mahsa Ghaffari, University of Illinois at Chicago; Mohamad Zoghi, State University of New York; Mostafa Rostami, Amirkabir University of Technology; Nabiollah Abolfathi, Amirkabir University of Technology</i>	
<i>A Laplace-Adomian Solution for Classical Fluid Dynamics Problems</i> .....	62
<i>Amir Mobasher, Alabama A&amp;M University; Mebougna Drabo, Alabama A&amp;M University; Satilmis Budak, Alabama A&amp;M University</i>	
<i>Modeling and Analysis of a Hydrogen Fueling Station Network</i> .....	69
<i>Fang Clara Fang, University of Hartford; Nefthali Torres, University of Texas at Austin</i>	
<i>Reflection and Transmission of Electromagnetic Waves in a Non-stationary Dielectric Medium</i> .....	77
<i>Deepthi Madhavapeddy, University of Connecticut; Ravindra Thamma, Central Connecticut State University</i>	
<i>Computational Aspects of Modeling Flow Boundary Layers</i> .....	84
<i>Sabah Tamimi, Al Ghurair University, UAE</i>	
<i>Testability-guided Abstraction for Scalable Bounded Model Checking of Embedded Software</i> .....	90
<i>Nannan He, Minnesota State University at Mankato; Michael S. Hsiao, Virginia Tech</i>	
<i>Instructions for Authors: Manuscript Submission Guidelines and Requirements</i> .....	99



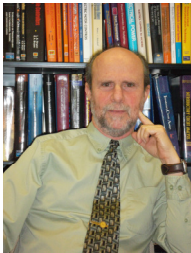
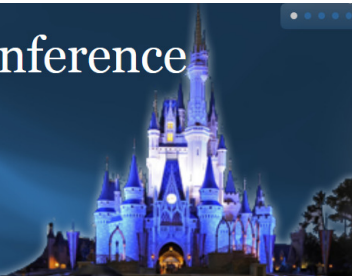


# 4th IAJC/ISAM Joint International Conference

September 25-27, 2014 – Orlando, Florida



The leading indexed high impact factor conference on engineering and related technologies.



## EDITOR'S NOTE

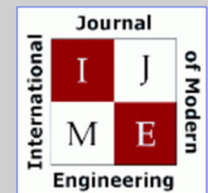
Philip Weinsier, IJME Manuscript Editor

The editors and staff at IAJC would like to thank you, our readers, for your continued support, and we look forward to seeing you at the upcoming IAJC conference. For this fourth IAJC conference, we will be partnering with the International Society of Agile Manufacturing (ISAM). This event will be held at the new Embassy Suites hotel in Orlando, FL, September 25-27, 2014, and is sponsored by IAJC, IEEE, ASEE, and the LEAN Institute. The IAJC/ISAM Executive Board is pleased to invite faculty, students, researchers, engineers, and practitioners to present their latest accomplishments and innovations in all areas of engineering, engineering technology, math, science, and related technologies.

I am also proud to report that, based on the latest impact factor (IF) calculations (Google Scholar method), the International Journal of Modern Engineering (IJME) now has a remarkable IF = 3.0 and continues its march toward the top 20 engineering journals. IJME's sister journal, the International Journal of Engineering Research and Innovation (IJERI) also had a strong showing with an IF = 1.58, which is noteworthy as it is a relatively young journal (only in publication since 2009). Any IF above 1 is considered high, based on the requirements of many top universities, and places the journals among an elite group.

**Conference Statistics:** Of the submissions for this conference, there were 466 abstracts from 132 educational institutions and companies from around the world (multiple submissions from the same authors were not counted). This represented 48 of the 50 U.S. states and 49 countries were represented. After a multi-level review process, only 82 full papers and 19 abstracts were accepted for presentation at the conference and for publication in the conference proceedings. This reflects an acceptance rate of about 22%, which is one of the lowest acceptance rates of any international conference.

### Main Sponsors



### Indexed In



## Editorial Review Board Members

Listed here are the members of the IAJC International Review Board, who devoted countless hours to the review of the many manuscripts that were submitted for publication. Manuscript reviews require insight into the content, technical expertise related to the subject matter, and a professional background in statistical tools and measures. Furthermore, revised manuscripts typically are returned to the same reviewers for a second review, as they already have an intimate knowledge of the work. So I would like to take this opportunity to thank all of the members of the review board.

As we continually strive to improve upon our conferences, we are seeking dedicated individuals to join us on the planning committee for the next conference—tentatively scheduled for 2016. Please watch for updates on our website ([www.IAJC.org](http://www.IAJC.org)) and contact us anytime with comments, concerns, or suggestions.

If you are interested in becoming a member of the IAJC International Review Board, send me (Philip Weinsier, IAJC/IRB Chair, [philipw@bgsu.edu](mailto:philipw@bgsu.edu)) an email to that effect. Review Board members review manuscripts in their areas of expertise for all three of our IAJC journals—IJME (the International Journal of Modern Engineering), IJERI (the International Journal of Engineering Research and Innovation), and TIIJ (the Technology Interface International Journal)—as well as papers submitted to the IAJC conferences.

Kevin Berisso	Ohio University (OH)	Anand Nayyar	Institute of Mgmt & Techn (INDIA)
Pankaj Bhambri	Guru Nanak Dev Eng Coll. (INDIA)	Aurenice Oliveira	Michigan Technological U. (MI)
Shaobiao Cai	Penn State University (PA)	Shahera Patel	Sardar Patel University (INDIA)
Hans Chapman	Moorehead State University (KY)	Karl Perusich	Purdue University (IN)
Michael Coffman	Southern Illinois U-Carbondale (IL)	Thongchai Phairoh	Virginia State University (VA)
Brad Deken	Southeast Missouri State U. (MO)	John Rajadas	Arizona State University (AZ)
Z.T. Deng	Ferris State University (MI)	Desire	
David Domermuth	Appalachian State University (NC)	Rasolomampionona	Warsaw Univ. of Techn (POLAND)
Dominic Fazarro	Sam Houston State University (TX)	Mulchand Rathod	Wayne State University (MI)
Youcef Himri	Safety Eng. Sonelgaz (ALGERIA)	Michael Reynolds	Univ. of Arkansas Fort Smith (AR)
Xiaobing Hou	Central Connecticut State U. (CT)	Anca Sala	Baker College (MI)
Sudershan Jetley	Bowling Green State U. (OH)	Ehsan Sheybani	Virginia State University (VA)
Rex Kanu	Ball State University (IN)	Siles Singh	St. Joseph U. (Tanzania, AFRICA)
Tolga Kaya	Central Michigan University (MI)	Carl Spezia	Southern Illinois University (IL)
Satish Ketkar	Wayne State University (MI)	Vassilios Tzouanas	Univ. of Houston-Downtown (TX)
Ognjen Kuljaca	Brodarski Institute (CROATIA)	Liangmo Wang	Nanjing U. Science&Tech (CHINA)
Chakresh Kumar	Uttar Pradesh Technical U. (INDIA)	Jonathan Williams	Bowling Green S.U. Firelands (OH)
Zaki Kuruppallil	Ohio University (OH)	Faruk Yildiz	Sam Houston State University (TX)
Ronald Land	Penn State University (PA)	Yuqiu You	Morehead State University (KY)
Chao Li	Florida A&M University (FL)	Biao Zhang	US Corporate Research (ABB INC.)
Guoxiang Liu	University of North Dakota (ND)	Jinwen Zhu	Missouri Western State U. (MO)
Louis (Yu) Liu	University of New Orleans (LA)		

# ULTRA-COMPACT GRAPHENE-BASED ELECTRO-OPTIC MODULATORS ON A SILICON-ON-INSULATOR PLATFORM

Chenran Ye, The George Washington University; Sarah Pickus, The George Washington University; Ke Liu, The George Washington University; Volker J. Sorger, The George Washington University

## Abstract

Electro-optic modulators (EOM) convert electronic signals into high-bit-rate photonic data. Its on-chip design plays an important role in the integration of electronic and photonic components for various types of applications, including photonic computing and telecommunication. Graphene is an emerging material allowing the design of ultra-compact EOMs due to its extraordinary electro-optic properties, addressing the trade-offs between high bandwidth and modulation strength. Two Graphene-based EOM designs are presented here for absorption and phase-shifter devices. The high-performance Graphene-based absorption modulator was analyzed numerically, demonstrating an extinction ratio and insertion loss of 7.77 dB/mm and 0.75 dB, respectively. This sub-wavelength compact (0.78 $\lambda$  long) absorption modulator is capable of broadband operation over a 500 nm bandwidth. The second design, a Mach-Zehnder modulator formed by push-pull Graphene-based phase shifters, exhibits an insertion loss of  $\sim$ 2.7 dB/mm with a 5.6 mm arm shifter length operating at telecom wavelength (1.55 $\mu$ m). These EOM performance results have the potential to become essential building blocks for optical interconnects in future integrated optoelectronic systems.

## Introduction

The advancement of modern technology is mainly driven by performance enhancement and cost minimization. As a promising technology, photonics has the potential to make optical computing and next generation Internet data rates a reality. The success and on-going trend of on-chip photonic integration anticipates a photonic road-map [1] leading to compact photonic integrated components and circuits. The electro-optic modulator (EOM) has been identified as one of the key drivers for optical communication [1-5]. The general device function may be perceived as an optical transistor with an optical source/drain and an electrical gate. While silicon-based EOMs have been demonstrated before, their device performance numbers are typically limited, due to weak light-matter-interactions (LMI) and non-linear electro-optic effects that lead to large device footprint requirements,

usually on the order of millimeters if utilizing a Mach-Zehnder interferometer (MZI) design [2], [3], [6].

LMIs are generally used to measure modulation strength and may be enhanced by deploying high-quality (Q) factor cavities, such as a ring resonator [3]. However, within the context of an on-chip design, the narrow spectral bandwidth, high power consumption (i.e., energy per bit), and long photon lifetimes severely limit future development of this technique. In particular, the modulation bandwidth is restricted to a few tens of GHz, due to a high-Q factor in the range of 10-100 thousand. Moreover, the deployment of such resonance-based photonic modulators is limited by their temperature-tuning requirements, due to high power penalties [2], [6]. The footprint and bandwidth constrains can simultaneously be overcome by enhancing an optical mode overlap with an actively index-modulated region and increasing the LMI of an optical mode with an actively index-modulated material by improving the electromagnetic field strength [4], [7-10]. In this study, LMI enhancement was achieved by utilizing a plasmonic Metal-Oxide-Semiconductor (MOS) type of optical hybrid-plasmon-polariton (HPP) mode. This concentrates the part of a propagating mode's field into a nanometer-thin region overlapping the actively index-modulated material, resulting in a deep sub-diffraction limited mode area [11-15]. Even though plasmonic waveguides have a fundamental trade-off between their mode confinement and propagation length, the HPP waveguide was found to be relatively superior [11], [15]. This hybrid approach of combining the dielectric waveguide mode and surface plasmon mode was proposed and demonstrated to achieve both sub-wavelength confinement and long propagation distances. HPP waveguides have an improved propagation length with deep sub-wavelength mode confinement, significantly below the diffraction limit (i.e., 1/50 of the diffraction limit) [11-15]. However, the propagation length of a HPP waveguide is obviously shorter than conventional dielectric waveguides [16], which have been previously proposed [11] and demonstrated [12], [13]. Here, the results suggest that choosing separated waveguide and material systems for passive optical data routing and active data switching can indeed be accomplished by using a silicon-on-insulator (SOI) substrate and a plasmonic hybrid integration architecture [8], [9].



The optically active (i.e., voltage-bias switched) material takes a second-most-important role in the EOM design. Some emerging materials with high modulation capability have been proposed and demonstrated for EO applications [8-10], [17]. Since its discovery in 2004, Graphene has become a highly regarded substance in the area of electronics because of its high mechanical strength, electrical and thermal conductivities, and compatibility with silicon [18-29]. With its unique optical properties, including strong LMI [21], high-speed operation [28], and gate-variable optical conductivity [25], [29], Graphene shows a significant potential for electro-optic applications [10], [17].

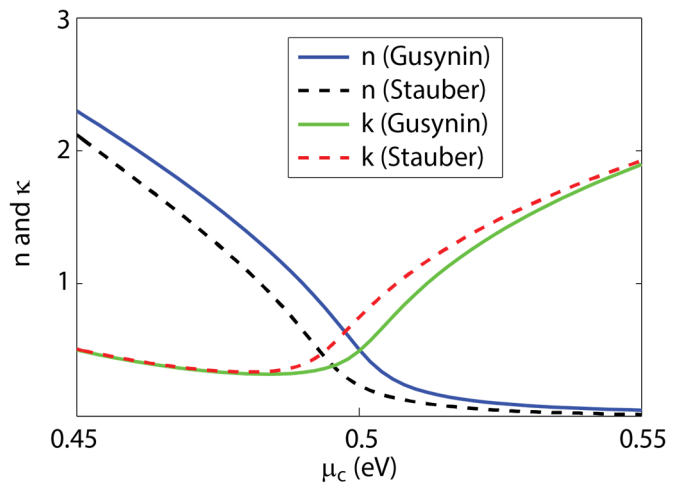
In this study, the authors report the design and performance details of Graphene-based broadband, ultra-compact plasmonic EOMs that allow for seamless integration into a SOI waveguide platform is reported. Two types of EOMs were investigated for absorption and phase-shift purposes. One was a plasmonic MOS-type optical absorption modulator and the other an MZI-based modulator formed by Graphene-oxide-silicon phase shifters. The performances of both EOMs were optimized using a finite element method with the commercial solver, COMSOL. Their performance parameters, including extinction ratio, insertion loss, and bandwidth, were analyzed and are discussed here.

## Graphene Electro-optic Modulator

Graphene, a monolayer material with gate-variable optical conductivity and strong interacting characteristics with light, is considered as a desired material for future EOM designs [25], [29]. Moreover, the interaction is tunable by varying chemical potentials. The extremely high current density and intrinsic mobility (100 times more than in silicon) makes Graphene an attractive material for optoelectronic devices [10], [17], [18], [28].

The chemical potential, or Fermi level, is considered as one of the essential properties that determines the optical conductivity or complex refractive index of a material—in this case, Graphene. The chemical potential may be tuned by an electrical gating voltage resulting in an effective index change of the optical mode, thereby affecting device performance. It is, therefore, important to study Graphene's refractive index as a function of its chemical potential. The optical refractive index of Graphene, which depends on the chemical potential, can be calculated by either of two different approaches developed by Gusynin et al. [19] and Stauber et al. [23], two of the most commonly used methods for the design and analysis of Graphene devices at a 1.55  $\mu\text{m}$  wavelength [5]. Figure 1 depicts the real and imaginary parts of the refractive index of Graphene for  $0.45 < m_c < 0.55$  eV ( $m_c = \text{chemical potential} = f(V_b)$ ). It is worth mentioning

that the real part of the refractive index,  $n$ , slightly decreases, while the imaginary part,  $k$ , increases in the range from 0.49 eV to 0.5 eV.



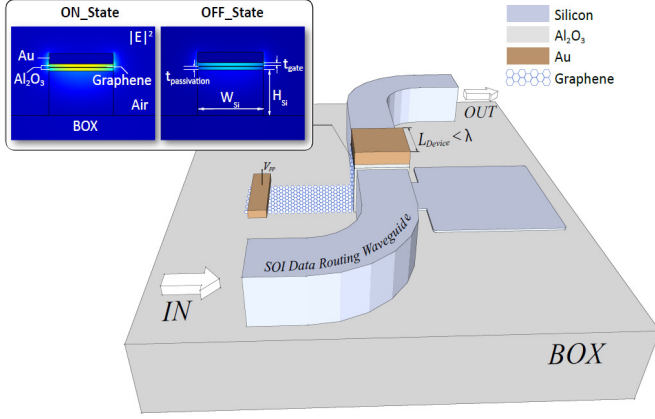
**Figure 1. Complex Refractive Index of Graphene versus Chemical Potential**

By selecting a proper range of chemical potential, graphene can be implemented into two different EOM mechanisms; namely, absorption modulation and phase modulation.

## Device Design and Performance

The ultra-compact, l-size EOM used in this study consisted of an SOI waveguide covered by a passivation oxide layer and a MOS modulation gate stack placed on top, forming a MOS capacitor (see Figure 2). Note that the Graphene film is sandwiched into the MOS stack, which overlaps with the high field density of the HPP plasmonic mode in order to enhance the LMIs. Applying a voltage bias between the Graphene and silicon shifts the effective index of the Graphene and, ultimately, that of the propagating HPP mode, which is the basis of the modulation.

Compared with state-of-the-art Graphene-based modulators [10], [17], the device in this study takes advantage of the merits of a hybrid plasmonic structure. Instead of using a Graphene-oxide-silicon configuration, a metal-oxide-semiconductor stack on the top of silicon waveguide was adopted in the current modulator. Even though the use of metals in optical waveguiding adds the per-length modal loss, the effective mode modulation can be enhanced, leading to the opportunity to design micrometer-sized devices with relatively low insertion loss, as discussed here. As a result, this type of design is highly appropriate for densely integrated photonic on-chip components.



**Figure 2. Schematic of Sub-1 Graphene-based Absorption Modulator**

As mentioned earlier, the top Gold (Au) metal layer: i) provides a strong LMI that results in sub-wavelength plasmonic confinement; ii) provides a convenient electrical contact; and, iii) serves as a heat sink [8], [9], [11-15]. The device dimensions used for the performance optimization included the silicon waveguide core width and thickness ( $W_{Si}$ ,  $H_{Si}$ ), a passivation oxide thickness ( $t_{passivation}$ ), a gate oxide thickness ( $t_{gate}$ ), and a device length ( $L$ ). The effective thickness of Graphene was considered to be 0.7 nm in this study and could be used in future studies to tune the complex conductivity of Graphene [17].

The extinction ratio ( $ER$ ) represents the signal discrimination between the on and off states, and the insertion loss ( $IL$ ) quantifies the undesired loss that occurs when light is propagating through the device with the bias voltage applied (light-ON state). Both parameters were used to define the performance of modulation and can be expressed as:

$$ER = \frac{P_{out}(V_b = V_{OFF})}{P_{out}(V_b = V_{ON})} \quad (1)$$

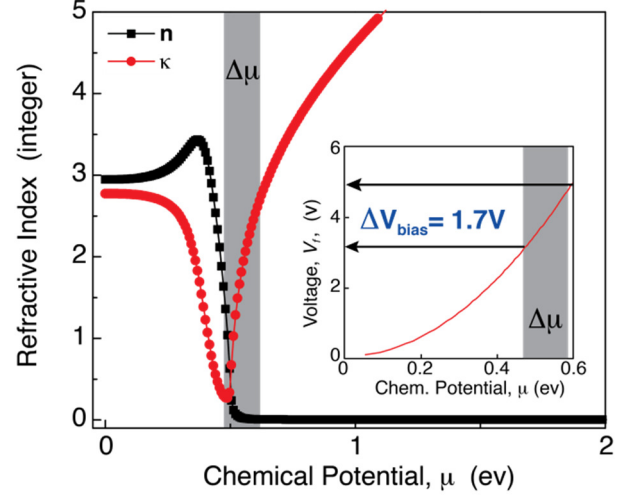
$$IL = \frac{P_{in} - P_{out}(V_b = V_{ON})}{P_{in}} \quad (2)$$

where,  $P_{out}$  ( $P_{in}$ ) is the optical power at the output (input) of the device.

## Absorption Modulator (AM)

The model of absorption modulation was built by tuning the imaginary part of the refractive index,  $k$ , also known as the extinction coefficient, of Graphene via the bias voltage. The bias-shifting chemical potential of Graphene results in

the refractive index change shown in Figure 3. Note that while  $k$  shifts dramatically, the real part of refractive index was only slightly altered. A ‘dielectric Graphene’ at 0.445 eV may be transformed into a ‘metallic Graphene’ at 0.515 eV. The bias voltage required to achieve this transition was expressed by Low [30] in Equation (3):



**Figure 3. Selected Chemical Potential Range for an Absorption Modulator**

$$V = \frac{\left( \frac{\mu_c}{\hbar_{bar} \cdot V_f} \right)^2 \cdot e + \epsilon_{Si} \cdot E_{Si}}{\epsilon_{OX} \cdot \pi} \cdot t_{OX} - \Phi_g + \Phi_{metal} \quad (3)$$

The electric field across the oxide is given by Equation (4):

$$E_{OX} = \frac{1}{t_{OX} (\Phi_g - \Phi_m + V_m)} \quad (4)$$

The electric field in the silicon adjacent to the Graphene is approximated using Equation (5):

$$E_{Si} = \sqrt{\frac{2e \cdot n_{Si} \cdot (V_i - V_{Si})}{\epsilon_{Si}}} \quad (5)$$

where,

- $\hbar_{bar}$  =  $6.59 \times 10^{-16}$  m<sup>2</sup>kg/s Planck constant
- $v_f$  =  $10^6$  m/s Fermi velocity in Graphene
- $e$  =  $1.6 \times 10^{-19}$  constant of elementary charge
- $V_{Si}$  = 0 V bias voltage for silicon
- $V_m$  = 1 V bias voltage for metal
- $n_{Si}$  =  $1.45 \times 10^{10}$  doping level on silicon
- $\epsilon_0$  =  $6.59 \times 10^{-16}$  F/m vacuum permittivity
- $\epsilon_{Si}$  =  $11.68 \epsilon_0$  silicon permittivity

- $\epsilon_{ox}$  = 3.9  $\epsilon_0$  oxide permittivity
- $F_g$  = 4.6 eV work function of Graphene
- $F_m$  = 5.1 eV work function of metal (gold)
- $\mu_c$  = chemical potential
- $t_{ox}$  = thickness of oxide
- $V_i$  = built-in potential in silicon

In this study, adopting a 5 nm oxide thickness in the device design only required a 1.7 V bias alternation in order to enable the modulator to be switched on and off. The detailed performance was analyzed with respect to various geometrical device changes and is summarized in Figure 4. Starting by altering the silicon waveguide core with the fixed passivation and gate oxide thickness ( $t_{passivation} = t_{gate} = 5$  nm), the modulation performance can be improved by choosing a silicon core height and width that results in the smallest modal area and longest propagation length—in this case, captured as the  $ER/IL$  ratio.

Both extinction ratio and insertion loss were used to determine device performance.  $Dk$  was a main contributor to  $ER$  and was based on Equation (6):

$$ER = \frac{T(L, \Delta\alpha)}{T_0} \quad (6)$$

where,  $T$  is the transmitted optical signal;

$$\Delta\alpha = 4\pi \cdot \frac{\Delta k}{\lambda} \quad (7)$$

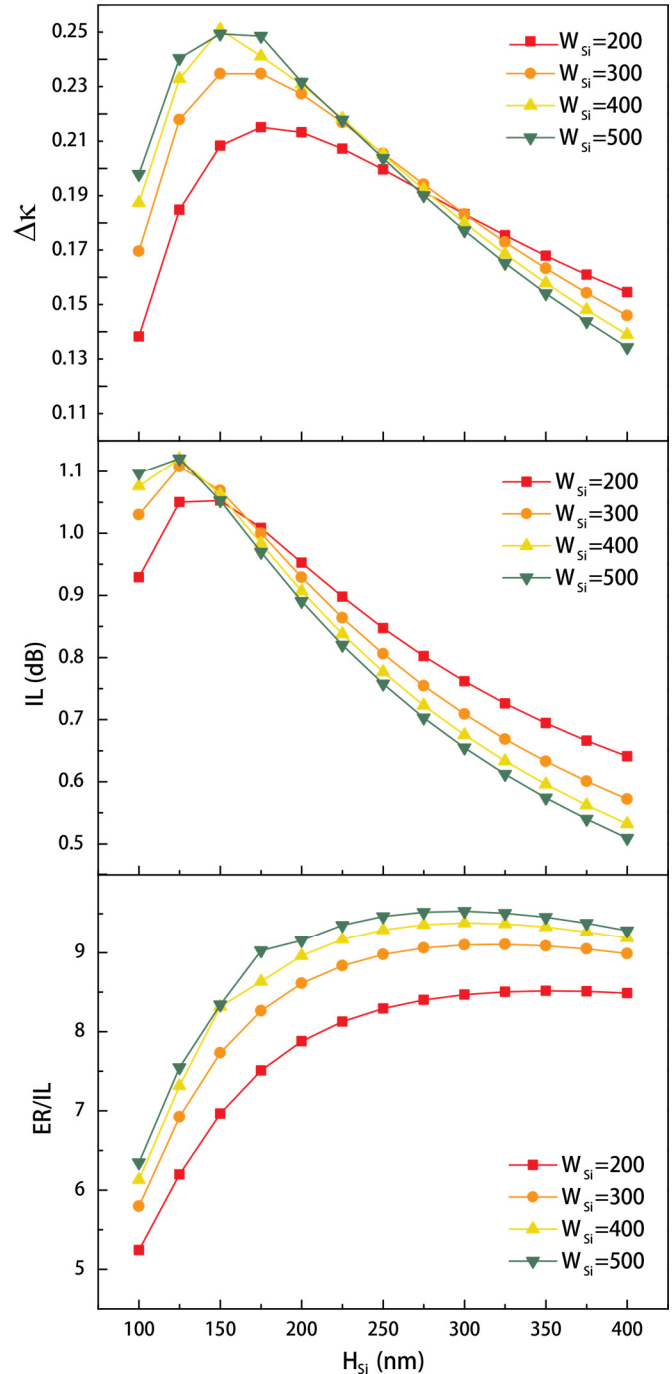
is the incurred loss over the length of the device,  $L$ ;  $\lambda$  is the operation wavelength; and,  $k$  is the imaginary part of the complex effective HPP mode index.  $IL$  is given by Equation (8):

$$IL = 1 - \frac{T(L, \alpha_{ON})}{T_0} \quad (8)$$

The ratio  $ER/IL$  is considered as a figure of merit to track the device's performance, where the main goal is to maximize this ratio—that is, a higher  $ER$  and a lower  $IL$ .

When tuning the thickness of the silicon core,  $Dk$  reaches the maximum value before the SOI cut-off condition ( $l/2n_{eff} = 258$ nm,  $n_{eff} = 3.0$ ,  $l = 1550$  nm) is reached. Here, the field overlaps dominantly with the active material resulting in an undesirably high  $IL$ . Interestingly, as  $H_{Si}$  shrinks beyond the cut-off point, the narrower waveguide leads to a lower  $Dk$ , as more fields reside in the plasmonic gap overlapping the Graphene. However,  $IL$  decreases faster with the narrower

$W_{Si}$  as  $H_{Si} < 150$  nm, which was actually unexpected. Although both  $ER$  and  $IL$  decrease gradually, the  $ER/IL$  ratio exceeded the maximum of 9.53, where  $H_{Si} = 300$  nm and  $W_{Si} = 500$  nm, resulting in an optimized design.



**Figure 4. Device Performance with Various Silicon Core Dimensions**



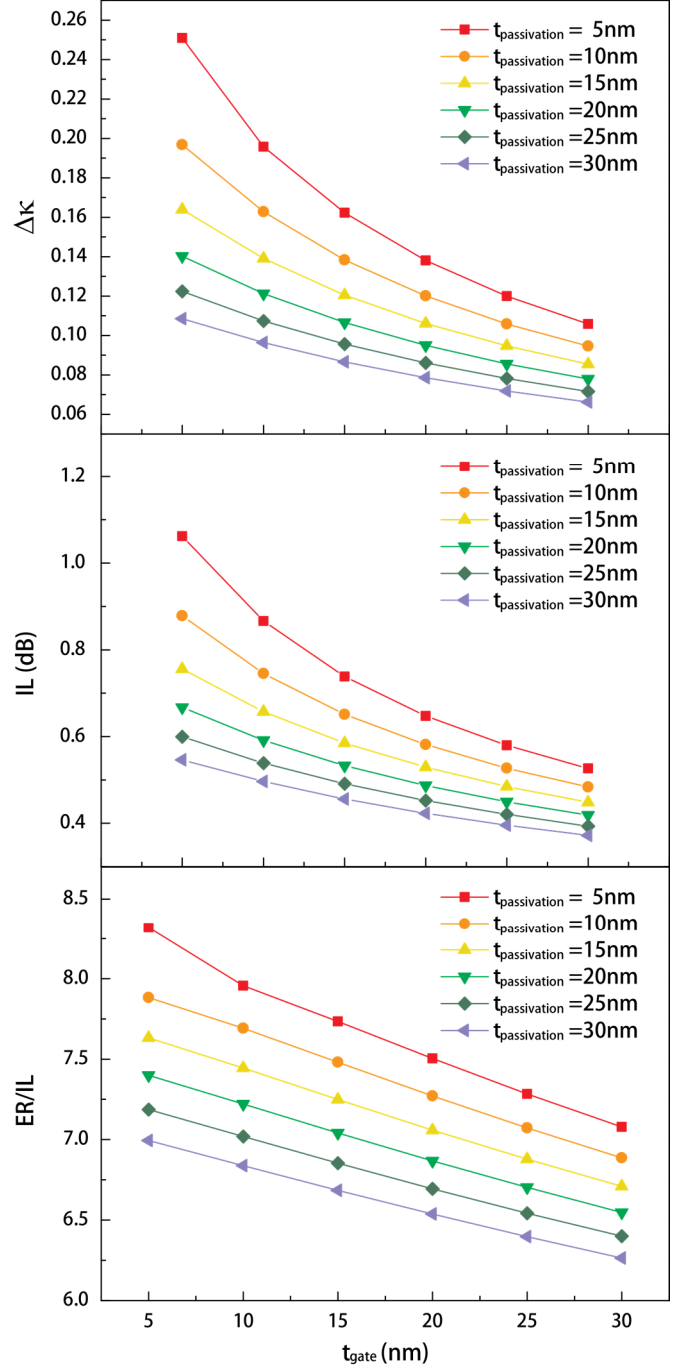
Two other geometric parameters that affect device performance are MOS oxide thickness ( $t_{gate}$ ), which controls the electrostatics of the Graphene, and the passivation layer thickness ( $t_{passivation}$ ). The latter originated in the current design from an experimental consideration in order to protect the silicon waveguide, allowing the authors to essentially avoid damage during the fabrication process. Moreover, it can help tune the mode towards a lower  $IL$ , due to the increased field overlap with the low-loss oxide region located inside the gap.

Keeping  $H_{Si} = 150$  nm and  $W_{Si} = 400$  nm constant in order to produce a maximum  $Dk$ , the effects of passivation and gate oxide were investigated and the authors made the following observations: the thinner the values of  $t_{gate}$  and  $t_{passivation}$ , the stronger the electric field inside the MOS mode overlapping the Graphene, thus leading to high  $ER$  but also large  $IL$  (see Figure 5). According to the study on  $ER/IL$  ratio,  $ER$  increases faster with a decreasing  $t_{gate}$ , leading to better performance for thinner oxides. As a result, the authors found that 5 nm passivation oxide and 5 nm gate oxide levels yielded the best results in this optimization. Also, while smaller oxides result in higher performance, 5 nm was chosen to allow for higher process control during nanofabrication.

## Broadband Potential

In telecommunications, wavelength-division-multiplexing (WDM) technology has demonstrated superior data and signal routing, due to the notion of parallelism of data by using different wavelengths in a single optical fiber. Thus, future on-chip devices and systems should be compatible with broadband operation, due to the multiple wavelengths being used. In order to avoid challenging and failure-prone tuning of individual resonator-based architectures, the Graphene absorption modulator studied here was investigated for its broadband operation performance by scanning a wavelength from 1.30 to 1.85  $\mu\text{m}$ .

Monitoring the performance parameter  $ER/IL$  ratio, the authors found a monotonically increasing value is found for longer wavelengths [31] (see Figure 6), which can be explained as follows: while the silicon core height was kept constant at 300 nm, the cut-off condition of the silicon core increased with wavelength. Therefore, less optical field overlaps with the Graphene in the gap, which reduces  $IL$  by a factor of 2x while  $ER$  only decreases by a factor of one third. This high performance over a 500-nm-wide bandwidth suggests the capability of broadband operation for the device and could be helpful for future systems operating with WDM.



**Figure 5. Passivation and Gate-oxide-layer Effect on Modulation Performance**

## Phase Modulator

A typical configuration for phase-based modulations of an optical signal is the MZI structure which includes two inter-

ferometric waveguide arms of equal length connected between two 3 dB optical couplers (see Figure 7). The output intensity, a function of either constructive or destructive interference, is controlled by the phase shifts. According to the study of the refractive index of Graphene, this material possesses a great capability for implementation in phase modulators. Two chemical potentials, 0.42 eV and 0.51 eV, were chosen for the model in Figure 8. When the extinction coefficient,  $k$ , remains constant, the phase factor,  $n$ , changes from 2.30 to 0.57. This indicates a great opportunity to achieve a phase change within two effective modes. Figure 7 shows the design layout of an MZI implemented with Graphene-based phase-shifter arms. The two arms of the MZI consist of the Graphene-oxide-silicon waveguide that may be applied with voltage to modulate the light transmission.

$IL$  was calculated based on Equation (8), and  $ER$  is determined using Equations (9) and (10):

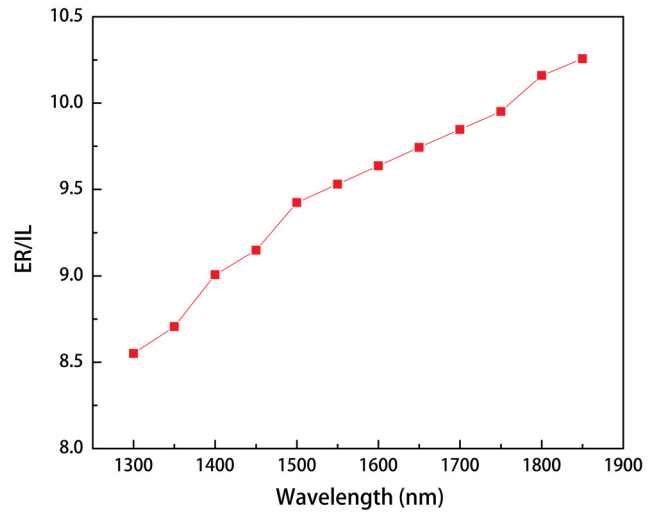


Figure 6. ER/IL Ratio over a Wide Range of Wavelengths for Absorption Modulator

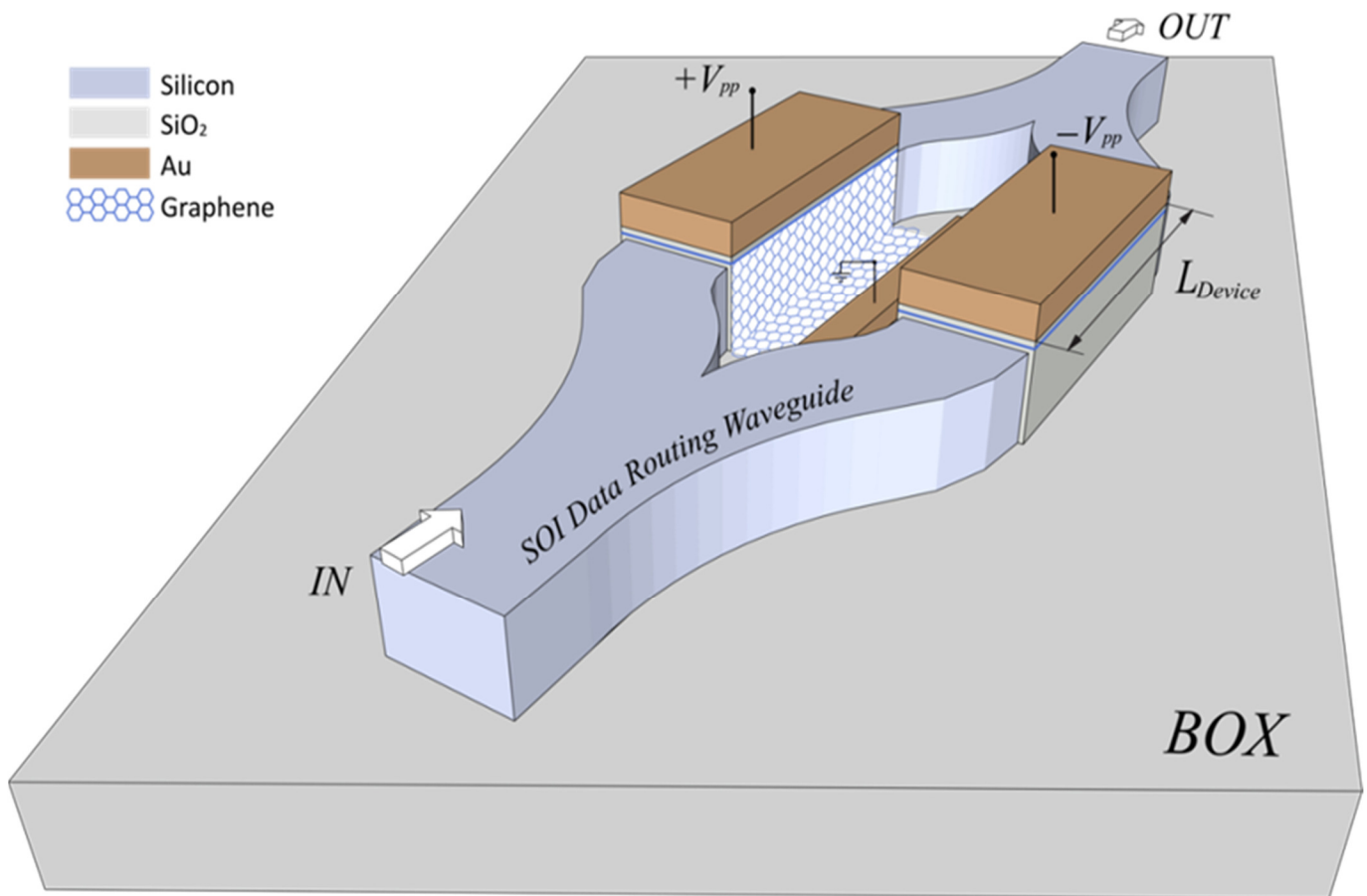


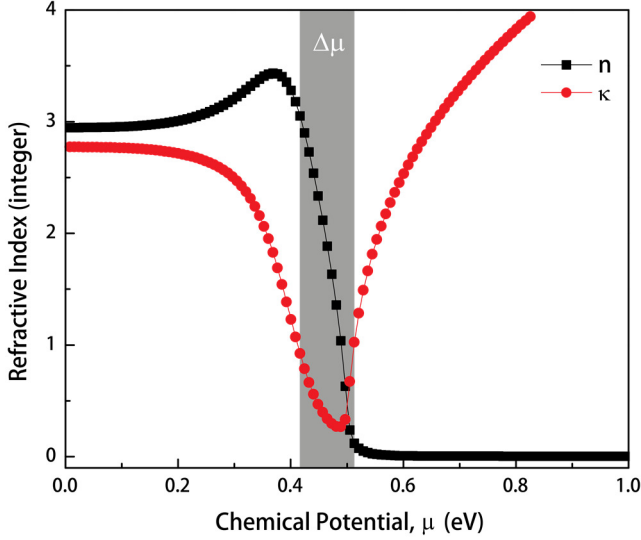
Figure 7. Schematic of Mach-Zehnder Modulator Formed by a Graphene-based Phase Shifter

$$ER = \frac{(1 + \cos \Delta \varphi)}{2} \quad (9)$$

$$\Delta \varphi = \frac{2\pi \cdot L}{\lambda} \cdot \Delta n_{eff} \quad (10)$$

where,  $Dj$  is the phase change between the two states over the length of the device,  $L$ ;  $\lambda$  is the operation wavelength; and,  $\Delta n_{eff}$  is the modulated effective mode index of the plasmonic HPP mode [6]. In order to realize an intensity modulator with the MZI design, a phase change of  $\pi$  between the two arms' states is required. Thus, the device length is given by Equation (11):

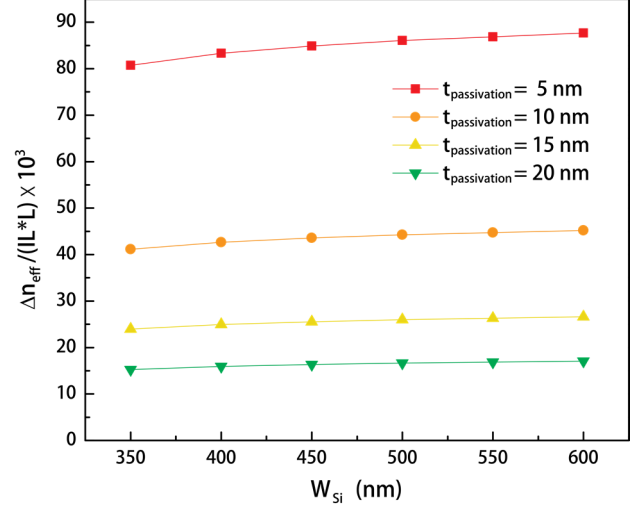
$$L = \frac{\lambda}{2 \cdot \Delta n_{eff}} \quad (11)$$



**Figure 8. Selected Chemical Potential Range for Phase Modulation**

In other words, the  $ER$  is fixed, while the optical signal propagates through the entire length of the device. For a fixed silicon core height of 200 nm, the device performance was analyzed for various  $W_{Si}$  and passivation oxide thicknesses. Here, the thickness of the gate oxide layer was assumed to have been held constant. The main purpose of optimization is to maximize the ratio of  $\Delta n_{eff}/IL \cdot L$ , which leads to a high  $Dn_{eff}$ , while decreasing  $IL$  and  $L$  simultaneously. Figure 9 shows that a thinner oxide layer results in better performance, as is the case for the absorption modulator. In addition, a smaller silicon core height yields a higher merit ratio. This is because the field has been pushed into

the Graphene gap when the silicon core height is comparable to the cut-off thickness of the pure silicon core mode. The best design has an  $IL$  of 2.76 dB for a 5.65-mm-long device, a waveguide width of 600 nm, and a passivation oxide thickness of 5 nm.



**Figure 9. Figure of Merit of Phase Modulator for Various Silicon Core Width and Passivation Oxide Thicknesses**

## Electro-optic Performance

The key performance metrics used to characterize an EOM are the modulation efficiency (energy/bit) and the bandwidth (i.e., speed). Altering the device's geometry as well as the modulator's  $ER$  can optimize both of these along with  $IL$ , as discussed earlier. Tables 1 and 2 summarize the quantitative performance of Graphene-based absorption and phase modulators operating at a wavelength of 1.55  $\mu\text{m}$ . For EOMs, the operation speed is typically limited by RC delay. For the MOS-based device used in this study, the authors estimated the speed by choosing two realistic values between 100 and 1000 Ohms.

For the absorption modulator, there are two optimal choices, depending on considerations of signal integrity and speed, which are denoted by bandwidth. Strong signal switching requires a longer device to achieve a high signal-to-noise ratio, while a low RC delay time for high-speed applications is satisfied by a short device length. According to Table 1, it is obvious that a 2-mm-long device produces strong data switching of 18 dB with a THz modulation speed, while a 0.5-mm-short option offers a bandwidth on the order of tens of THz with a medium signal switching strength. For the phase modulator, RC delay time along with bandwidth was estimated (see Table 2). Hundreds of GHz



of bandwidth could be achieved for high-speed operations with reasonably low energy per bit in the range. Compared with the current study of a silicon-based optical modulator [32], this device gives a  $V_{\pi} \times L = 0.78 \text{ m} \cdot \text{m}$  ( $V_{\pi} = \cdot DV = 4.5 - 3.1 = 1.4 \text{ V}$ ), which is a few orders of magnitude more compact and efficient.

**Table 1. Quantitative Performance Analysis Optimized Signal Integrity and Speed for the Absorption Modulator Case**

Absorption Modulator	Signal	Speed
L (Device Length – mm)	2	0.5
R (Resistance – Ohm)	100 – 1000	
C (Capacitance – fF)	2.6	0.7
BW (Bandwidth – THz)	3.8 – 0.4	15.1 – 1.5
IL (Insertion Loss – dB)	2.1	0.5
ER (Extinction Ratio – dB)	18	4
V (Voltage – Volts)	3.5 – 4.6	
E/bit (Energy per bit – fJ)	16.2 – 28.0	4.1 – 7.0

**Table 2. Quantitative Performance Analysis for the Phase Modulator**

Phase Modulator	
L (Device Length – mm)	5.6
R (Resistance – Ohm)	100 – 1000
C (Capacitance – fF)	4.5
BW (Bandwidth – THz)	2.2 – 0.2
IL (Insertion Loss – dB)	2.7
V (Voltage – Volts)	3.1 – 4.5
E/bit (Energy per bit – fJ)	21.7 – 45.8

## Conclusion

In this study, an ultra-compact plasmonic Graphene-based EOM was proposed and analyzed. Both the absorption and phase modulators were optimized, based on a finite element method in COMSOL. The numerical simulation results show a promising MOS mode for the enhancement of light-matter interaction. For the absorption modulator, the insertion loss can be reduced to 0.8 dB while the ER remains at 7.8 dB/mm, opening up opportunities for sub- $\mu\text{m}$  EOMs. The estimated energy per bit was as low as 4.1 fJ/bit, with a maximum bandwidth of  $\sim 15$  THz when a 3.5 V voltage was applied. For the phase modulator, a 5.6 mm compact device

may burn about 20 fJ per bit, while being modulated at  $\sim 2$  THz. These proposed EOMs have the potential to become essential building blocks for optical interconnects in future integrated optoelectronic systems.

## Acknowledgements

This work was partially supported by funds from the George Washington University; by funds from the National Natural Science Foundation of China under grant 61377059; and from the Beijing Municipal Natural Science Foundation under grant 4142004.

## References

- [1] Kirchain, R., & Kimerling, L. (2007). A roadmap of nanophotonics. *Nature Photonics*, 1, 303–305.
- [2] Reed, G. T., Mashanovich, G., Gardes, F. Y., & Thomson, D. J. (2010). Silicon optical modulators. *Nature Photonics*, 4, 518–526.
- [3] Xu, Q., Schmidt, B., Pradhan, S., & Lipson, M. (2005). Micrometre-scale silicon electro-optic modulator. *Nature*, 435(7040), 325–327.
- [4] Cai, W., White, J. S., & Brongersma, M. L. (2009). Compact, high-speed and power-efficient electro-optic plasmonic modulators. *Nano Letters*, 9, 4403–4411.
- [5] Ye, C., Khan, S., Li, Z. R., Simsek, E., & Sorger, V. J. (2014).  $\lambda$ -Size ITO and Graphene-Based Electro-Optic Modulators on SOI. *IEEE Journal of Select Topic in Quantum Electronics*, 20(4), 3400310.
- [6] Liu, A., Jones, R., Liao, L., Samara-Rubio, D., Rubin, D., Cohen, O., et al. (2004). A high-speed silicon optical modulator based on a metal-oxide-semiconductor capacitor, *Nature*, 427, 615–618.
- [7] Dionne, J. A., Diest, K., Sweatlock, L. A., & Atwater, H. A. (2009). PlasMOSTor: A metal-oxide-Si field effect plasmonic modulator. *Nano Letters*, 9(2), 897–902.
- [8] Huang, C., Lamond, R. J., Pickus, S. K., Li, Z. R., & Sorger, V. J. (2013). A Sub- $\mu\text{m}$  Modulator Beyond the Efficiency-Loss Limit. *IEEE Photonic Journal*, 5 (4), 202411.
- [9] Sorger, V. J., Lanzillotti-Kimura, N. D., Ma, R. M., & Zhang, X. (2012). Ultra-compact silicon nanophotonic modulator with broadband response. *Nanophotonics*, 1(1), 17–22.
- [10] Lu, Z., & Zhao, W. (2012). Nanoscale electro-optic modulators based on Graphene-slot waveguides. *Journal of the Optical Society of America, Optical Physics*, 29(6), 1490–1496.

- [11] Oulton, R. F., Sorger, V. J., Pile, D. F. B., Genov, D. & Zhang, X. (2008). Nano-photonic confinement and transport in a hybrid semiconductor-surface plasmon waveguide. *Nature Photonics*, 2, 496–500.
- [12] Sorger, V. J., Ye, Z., Oulton, R. F., Bartal, G., Wang, Y., Bartal, G., et al. (2011). Experimental demonstration of low-loss optical waveguiding at deep sub-wavelength scales. *Nature Communication*, 2(5), 331.
- [13] Sorger, V. J., Pholchai, N., Cubukcu, E., Oulton, R.F., Kolchin, P., Borschel, C., et al. (2011). Strongly Enhanced Molecular Fluorescence inside a Nanoscale Waveguide Gap, *Nano Letters*, 11, 4907-4911.
- [14] Alam, M. Z., Meier, J., Aitchison, J. S., & Mojahedi, M. (2007). Super mode propagation in low index medium. *Conference on Lasers and Electro-Optics/Quantum Electronics and Laser Science Conference and Photonic Applications Systems Technologies, OSA Technical Digest Series (CD) (Optical Society of America), paper JThD112*.
- [15] Oulton, R. F., Bartal, G., Pile, D. F. P., & Zhang, X. (2008). Confinement and propagation characteristics of subwavelength plasmonic modes. *New Journal of Physics*, 10(10), 105018.
- [16] Li, G., Yao, J., Thacker, H., Mekis, A., Zheng, X., & Shubin, I., et al. (2012). Ultralow-loss, high-density SOI optical waveguide routing for macrochip interconnects. *Optics Express*, 20(11), 12035-12039.
- [17] Liu, M., Yin, X., Ulin-Avila, E., Geng, B., Zentgraf, T., Long, J., et al. (2011). A Graphene-based broadband optical modulator. *Nature*, 474(7349), 64–67.
- [18] Andersen, D. R. (2010). Graphene-based long-wave infrared TM surface plasmon modulator, *Journal of the Optical Society of America B*, 27, 818–823.
- [19] Gusynin, V. P., Sharapov, S. G., & Carbotte, J. P. (2007). Magneto-optical conductivity in Graphene, *Journal of Physics: Condensed Matter*, 19, 026222.
- [20] Peres, N. M. R., Guinea, F., & Castro Neto, A. H. (2006). Electronic properties of disordered two-dimensional carbon. *Physics Review B*, 73, 125411.
- [21] Wunsch, B., Stauber, T., Sols, F., & Guinea, F. (2006). Dynamical polarization of Graphene at finite doping. *New Journal of Physics*, 8, 318.
- [22] Hwang, E. H., & Das Sarma, S. (2007). Dielectric function, screening, and plasmons in two-dimensional Graphene. *Physics Review B*, 75, 205418.
- [23] Stauber, T., Peres, N. M. R., & Geim, A. K. (2008). Optical conductivity of Graphene in the visible region of the spectrum. *Physics Review B*, 78, 085432.
- [24] Simsek, E. (2013). A Closed-Form Approximate Expression for the Optical Conductivity of Graphene. *Optics Letters*, 38(9), 1437–1439.
- [25] Li, Z. Q., Henriksen, E. A., Jiang, Z., Hao, Z., Martin, M.C., Kim, P., et al. (2008). Dirac charge dynamics in Graphene by infrared spectroscopy. *Nature Physics*, 4, 532.
- [26] Schedin, F., Lidorikis, E., Lombardo, A., Kravets, V. G., Geim, A. K., Alexander, N., et al. (2010). Surface-Enhanced Raman Spectroscopy of Graphene. *ACS Nano*, 4(10), 5617–5626.
- [27] Nair, R. R., Blake, P., Grigorenko, A. N., Novoselov, K. S., Booth, T. J., Stauber, T., et al. (2008). Fine structure constant defines visual transparency of Graphene. *Science*, 320, 1308.
- [28] Xia, F., Mueller, T., Lin, Y. M., Valdes-Garcia, A., & Avouris, P. (2009). Ultrafast Graphene photodetector. *Nature Nanotechnology*, 4, 839–843.
- [29] Wang, F., Zhang, Y., Tian, C., Girit, C., Zettl, A., Krommie, M., et al. (2008). Gate-variable optical transitions in Graphene. *Science*, 320, 206 – 209.
- [30] Low, T. (2013). *Personal discussions with Tony Low*.
- [31] Palik, E. D. (1991). *Handbook of Optical Constants of Solids II*. Academic, San Diego, CA.
- [32] Thomson, D. J., Gardes, F.Y., Fedeli, J. M., Zlatanovic, S., Youfang, H., Kuo, B. P. P., et al. (2012). 50-Gb/s Silicon Optical Modulator. *IEEE Photonics Technology Letters*, 24(4), 234-236.

## Biographies

**CHEN RAN YE** was born in Beijing, China, in 1988. She received the B.S. degree in Automation from the North China University of Technology, China, in 2010; the B.S. degree in Electrical Engineering from Southern Polytechnic State University, Marietta, GA, in 2010; and, the M.S. degree in Electrical Engineering from the University of Florida, Gainesville, FL, in 2012. Currently, she is working as a research assistant. Recently, she entered the field of nanophotonics and joined the Nanophotonic Lab lead by Prof. Sorger at The George Washington University. Her research interests include the investigation of novel electro-optic devices and the design, simulation, and fabrication of gate-level computer architecture via nanophotonics. She is also a member of IEEE. Ms. Ye can be reached at [yechenran@gmail.com](mailto:yechenran@gmail.com).

**SARAH PICKUS** is currently a senior student at the George Washington University pursuing a B.S. in Biomedical Engineering. In the spring of 2013, she joined the Sorger Group to help conduct research in the field of nanophotonics, mainly simulating performance and device parameters for electro-optic modulators. In addition to being an undergraduate research assistant, she has been a member of GW's Varsity Women's Rowing Team for the last four years and

---

was voted captain her senior year. After graduating, Sarah plans to pursue a doctorate in Biomedical Engineering and conduct research within the area of biophotonics. Ms. Pickus can be reached at [spickus@gwmail.gwu.edu](mailto:spickus@gwmail.gwu.edu).

**KE LIU** received the B.S. degree in electronic engineering from the Nanjing University of Science and Technology, Nanjing, China, in 1996; the M.S. degree in circuits and systems from Jilin University, Changchun, China, in 1999; and, the Ph.D. degree in integrated optics from the City University of Hong Kong, Hong Kong, in 2005. He has been a professor of Electronic Information and Control Engineering at the Beijing University of Technology since 2011. Recently, he joined Prof. Sorger's research group in the Department of Electrical and Computer Engineering at The George Washington University, USA. From 2006 to 2011, he worked at the University of North Carolina at Charlotte and the University of Texas at Dallas, respectively, as a Research Associate. He has authored or co-authored more than 40 journal publications and conference papers. His current research interests include integrated photonic devices such as silicon-based plasmonic switches and modulators, and MQW InP-based photonic integrated circuits. He is a member of IEEE and OSA. Dr. Li can be reached at [liukcityu@163.com](mailto:liukcityu@163.com).

**VOLKER J. SORGER** is an assistant professor in the Department of Electrical and Computer Engineering and the director of the Nanophotonics Labs at The George Washington University. He received his Ph.D. and M.S. from the University of California – Berkeley in 2011 and the University of Texas – Austin in 2005, respectively, where he conducted research in the fields of nanoscale opto-electronics and plasmonic devices. Prior to coming to GWU, Dr. Sorger was a leading post-doctoral researcher for the NSF Nanoscale Science and Engineering Center at UC Berkeley. His current research areas include enhanced devices with enhanced light-matter-interactions in plasmonics and nanophotonics, ultrafast switching, and modulation for photonic integrated circuits for aJ/bit data communication. His achievements include his pioneering contributions on nanoscale waveguides, the first demonstration of a semiconductor plasmon laser, and the realization of cavities with highest Q-factors for sub-wavelength optical fields. Recently, he showed the true potential plasmonics by demonstrating the first plasmonic-enhanced electro-optic modulator, which was the second-most downloaded paper of OSA's annual Advanced Photonics Congress in 2012. Dr. Sorger established himself as a leading figure in science and engineering by becoming the editor-in-chief for the journal of Nanophotonics, and organizing various topical meetings and conferences. Throughout his career Dr. Sorger received multiple awards such as the AFOSR YIP award, the MRS

Graduate Gold Award, Intel PhD Fellowship, to name a few. Lastly, Dr. Sorger was and is an active STEM advocate and involved in outreach holding the position as executive chair of the OSA Nanophotonics technical group, and actively involved in the National Photonic Initiative (NPI). He is a member of IEEE, OSA, SPIE, and MRS. Dr. Sorger can be reached at [sorger@gwu.edu](mailto:sorger@gwu.edu).



# OPTIMAL TUNING OF PI CONTROLLERS SUBJECT TO PROCESS CONSTRAINTS: EXPERIMENTAL EVALUATION

Constantine Tzouanas, Clear Lake High School; Vassilios Tzouanas, University of Houston–Downtown

## Abstract

A key contribution of the work from this current study is the experimental evaluation of a new optimal tuning method and the comparison of its performance to that of classical tuning methods. The experimental process was an interacting twin-tank system, where water flows from one tank to the other due to a difference in hydraulic pressure. For this interacting and non-linear process, and considering only a PI controller, the study used experimental results to show that the new tuning method could yield satisfactory performance, while user-specified constraints were being respected.

## Introduction

A method was recently presented for the tuning of proportional-integral-derivative (PID) controllers subject to process constraints [1]. Compared to other optimal tuning methods, this method accounts for a number of constraints such as controlled variable, manipulated variable, and rate of change for the manipulated variable. Since the PID controller is the most widely used control algorithm in the processing industries [2], there has been considerable interest in developing efficient tuning methodologies for such controllers. O'Dwyer [3] offers an extensive list of PI and PID tuning methods. Over the years, methods have been proposed which result in better control performance and improved robustness [4], [5]. Most of these techniques are analytic in nature and require the use of a linear process model. In addition, a number of methods have been proposed on the optimal tuning of PID controllers [6-11]. Almost exclusively, such methods are concerned with optimizing a performance criterion and the efficiency of the optimization methodology. None of them, however, consider process constraints when they determine the optimal tuning parameters of the PID controller.

From a practical point of view, it is important to tune PID controllers subject to operating constraints. Such an approach helps to improve process reliability, reduce maintenance costs, and strike a balance between control performance and system robustness. The method considered in this study accounts for a number of constraints such as controlled variable, manipulated variable, and rate of change

for the manipulated variable [1]. The objective of the current study was to demonstrate the method's applicability using an experimental setup and compare its performance to that of classical tuning methods.

## The Tuning Methodology

Consider a process under feedback control, as shown in Figure 1:

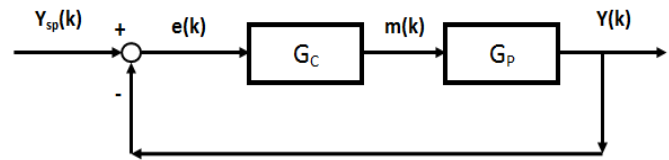


Figure 1. Schematic of a Feedback Control Loop

where,

- $G_p$  is the process model
- $G_c$  is the controller transfer function
- $m(k)$  is the manipulated variable
- $y(k)$  is the controlled variable
- $y_{sp}(k)$  is the controlled variable setpoint
- $e(k)$  is the control error
- $k$  is the discrete time instant

In this study, a PI-only control algorithm was used because it is sufficient to control the experimental setup. However, the same concepts apply to a PID controller, as has been demonstrated with simulation studies [1]. The PI algorithm is given by Equation (1):

$$m(k) = m(k-1) + K_c \cdot [(e(k) - e(k-1)) + \frac{\tau_s}{\tau_i} \cdot e(k)] \quad (1)$$

where,

- $K_c$  is the process gain
- $\tau_i$  is the integral time
- $\tau_s$  is the execution period

According to the tuning methodology, constraints on the controlled variable, manipulated variable, rate of change of manipulated variable, and the expected tuning parameter range were considered, while an optimization criterion was

minimized. The optimization criterion can be the Integral Absolute Error (IAE) or the Integral Square Error (ISE). Specifically, the integral absolute error is:

$$IAE = \sum_{j=0}^k |e(j)| \quad (2)$$

while the integral square error is:

$$ISE = \sum_{j=0}^k e(j)^2 \quad (3)$$

Given a process model, a control algorithm, and a desired setpoint, the objective is to minimize Equations (2) or (3), subject to the constraints noted in Equations (4)-(8):

Manipulated Variable Constraints:

$$m_{LL} \leq m(t) \leq m_{UL} \quad (4)$$

$$\Delta m_{LL} \leq \Delta m(t) \leq \Delta m_{UL} \quad (5)$$

Controlled Variable Constraints:

$$y_{LL} \leq y(t) \leq y_{UL} \quad (6)$$

Tuning Parameter Constraints:

$$K_{c,LL} \leq K_c \leq K_{c,UL} \quad (7)$$

$$\tau_{i,LL} \leq \tau_i \leq \tau_{i,UL} \quad (8)$$

where, the subscripts *LL* and *UL* refer to the lower limit and upper limit of a variable, respectively.

## The Experimental Process

In this section, a short description of the process is presented along with the non-linear, dynamic model.

Figure 2 shows the twin-tank water system. It is based on an experimental system at the University of Houston–Downtown (UHD). In this twin-tank process, water is pumped into the first tank (Tank 1) using a variable-speed pump. From tank 1, water flows into the second tank (Tank 2) due to a difference in hydraulic pressure. From tank 2, water flows back into the reservoir. The control objective is to maintain the level of tank 2 at a desired setpoint using a PI controller. The manipulated variable is the speed of the pump or, more precisely, the DC voltage to the pump. The controller must account for the upper and lower limit con-

straints of the two tank levels, the pump voltage, the rate of change for the pump voltage, and the range of tuning parameters. Figure 3 shows the experimental water tank system along with the LabVIEW-based computer control system.

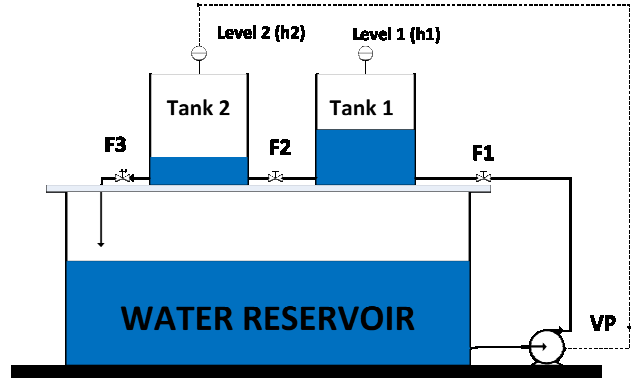


Figure 2. Schematic of the Interconnected Twin-tank Process

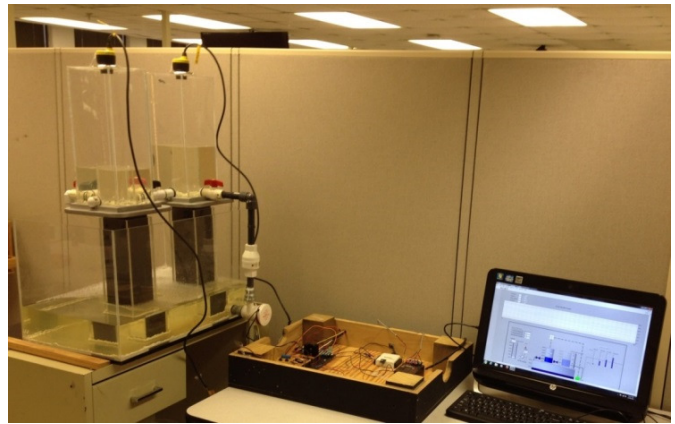


Figure 3. The Experimental Setup

For the twin-tank process, a dynamic non-linear system was developed using material balances. Highlights of this model are presented next.

## Process Model

Tank 1 Model:

$$\frac{dh_1}{dt} = \frac{F(V_p)}{A_1} - \frac{K_1}{A_1} \cdot \sqrt{h_1 - h_2} \quad (9)$$

Tank 2 Model:

$$\frac{dh_2}{dt} = \frac{K_1}{A_2} \cdot \sqrt{h_1 - h_2} - \frac{K_2}{A_2} \cdot \sqrt{h_2} \quad (10)$$

Flowrate Equation (pump curve):

$$F(V_p) = 462.64 \cdot V_p^4 - 7017.5 \cdot V_p^3 + 37481 \cdot V_p^2 - 78333 \cdot V_p + 54356 \quad (11)$$

Hydraulic Pressure Constants:

$$K_1 = \frac{F_{1s}}{\sqrt{h_{1s} - h_{2s}}} \quad (12)$$

$$K_2 = \frac{F_{1s}}{\sqrt{h_{2s}}} \quad (13)$$

Initial Conditions:

At time  $t=0$ ,  $V_p = V_{ps}$ ,  $h_1 = h_{1s}$ ,  $h_2 = h_{2s}$  are known.

In Equations (9)-(13), the variables are defined as follows:

- $h_1$  tank 1 level
- $h_2$  tank 2 level
- $F(V_p)$  flowrate for a pump voltage  $V_p$
- $K_1$  tank 1 hydraulic constant
- $K_2$  tank 2 hydraulic constant
- $A_1$  tank 1 cross section
- $A_2$  tank 2 cross section

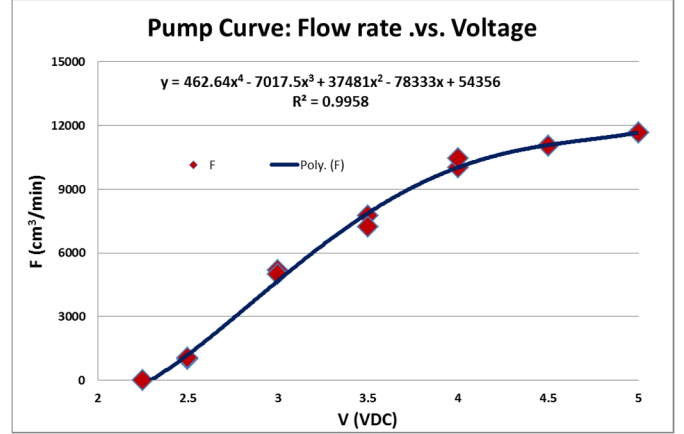
The subscript  $s$  is used to denote the values of the variables at steady-state conditions. Numerical values of these variables are given in Table 1. In all simulation studies, the above model has been solved using Euler's method with an integration step,  $Dt$ , of 3 seconds.

**Table 1. Initial Values of Tank Model Variables**

Variable	Value	Units
$V_{ps}$	2.75	V DC
$F_{1s}$	48.45	$\text{cm}^3/\text{s}$
$h_{1s}$	6.50	cm
$h_{2s}$	3.50	cm
$K_1$	27.98	$(\text{cm}^3/\text{s})/\text{cm}^{0.5}$
$K_2$	25.90	$(\text{cm}^3/\text{s})/\text{cm}^{0.5}$
$A_1$	164	$\text{cm}^2$
$A_2$	164	$\text{cm}^2$

Equation (11) gives the flowrate delivered by the pump for a given voltage applied to the pump. It is a non-linear equation and has been developed experimentally. Its ability to predict experimental data is shown in Figure 4. As

demonstrated in Figure 4, the pump cannot deliver water for voltage values less than about 2.3 VDC.

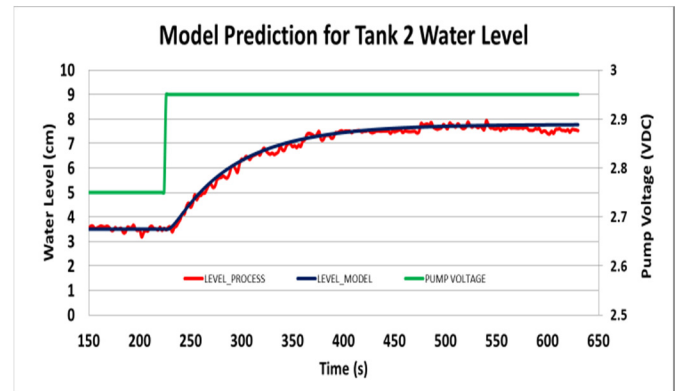


**Figure 4. Pump Curve**

The values of the hydraulic pressure constants,  $K_1$  and  $K_2$ , were calculated using Equations (12) and (13) such that the level derivatives given by Equations (9) and (10) are zero at initial steady-state conditions.

## Model Validation

Since the previously described model will be used to develop tuning parameters, its accuracy was tested against the behavior of the experimental process. In an open-loop fashion (i.e., the controller in manual mode), and with the process at steady state, the pump voltage was increased from 2.75 VDC to 2.95 VDC. The model response for the level of tank 2 (blue line) to this step change in pump voltage is given in Figure 5. Also, Figure 5 shows the experimental process response (red line) under the same conditions. It can be seen that the model prediction and the actual process response match reasonably well.



**Figure 5. Model Prediction and Actual Process Response**

## Empirical Model

The tuning method presented earlier can use the non-linear process model to estimate tuning parameters. However, classical tuning methods such as IMC and SIMC require a linear model. Such an empirical, linear model was derived using the data from Figure 5. The twin-tank response was fitted with a first-order-plus-dead-time model. The model is given by Equation (14) with the units of time in seconds:

$$\frac{h_2(s)}{V_p(s)} = G_p(s) = \frac{21.25 \cdot e^{-10s}}{66 \cdot s + 1} \quad (14)$$

## PI Controller Tuning

Three tuning methods were compared using simulation and experimental studies: the optimization subject to constraints method [1], the IMC method [4], and the SIMC method [5].

## Optimization Method

In this study, the integral absolute error was minimized, subject to a number of equality and inequality constraints.

### Equality Constraints

$$F(V_p) = 462.64 \cdot V_p(k)^4 - 7017.5 \cdot V_p(k)^3 + 37481 \cdot V_p(k)^2 - 78333 \cdot V_p(k) + 54356 \quad (15)$$

$$\frac{\Delta h_1}{\Delta t}(k) = \frac{F(V_p)}{A_1} - \frac{K_1}{A_1} \cdot \sqrt{h_1(k) - h_2(k)} \quad (16)$$

$$\frac{\Delta h_2}{\Delta t}(k) = \frac{K_1}{A_2} \cdot \sqrt{h_1(k) - h_2(k)} - \frac{K_2}{A_2} \cdot \sqrt{h_2(k)} \quad (17)$$

$$h_1(k) = h_1(k-1) + \frac{\Delta h_1}{\Delta t}(k) \cdot \Delta t \quad (18)$$

$$h_2(k) = h_2(k-1) + \frac{\Delta h_2}{\Delta t}(k) \cdot \Delta t \quad (19)$$

$$e(k) = h_{2,sp}(k) - h_2(k) \quad (20)$$

$$V_p(k) = V_p(k-1) + K_c \cdot [\{e(k) - e(k-1)\} + \frac{\tau_s}{\tau_i} \cdot e(k)] \quad (21)$$

$$\Delta V_p(k) = V_p(k) - V_p(k-1) \quad (22)$$

### Inequality Constraints

Manipulated Variable Constraints:

$$V_{p,LL} \leq V_p(k) \leq V_{p,UL} \quad (23)$$

$$\Delta V_{p,LL} \leq \Delta V_p(k) \leq \Delta V_{p,UL} \quad (24)$$

Controlled Variable Constraints:

$$h_{2,LL} \leq h_2(k) \leq h_{2,UL} \quad (25)$$

Other Physical Constraints:

$$h_{1,LL} \leq h_1(k) \leq h_{1,UL} \quad (26)$$

Tuning Parameter Constraints:

$$K_{c,LL} \leq K_c \leq K_{c,UL} \quad (27)$$

$$\tau_{i,LL} \leq \tau_i \leq \tau_{i,UL} \quad (28)$$

where, the subscripts *LL* and *UL* refer to the lower and upper limits of a variable, respectively. Such limits are process dependent and user specified. Table 2 summarizes the upper and lower limits of the different variables.

**Table 2. Upper/Lower Limits for Optimization Method**

Variable	Lower Limit	Upper Limit
Pump Voltage (V DC)	2.3	5.0
Voltage Rate (V DC/s)	-0.17	0.17
Level 2 (cm)	0	9
Level 1 (cm)	0	31
Proportional Gain	0.0	5
Integral Time (s)	1.0	500

The tuning parameters obtained using the IMC and SIMC methods were used to define a reasonable range for the tuning parameters in the optimization method. As suggested by the authors in an earlier study [1], the optimization method was implemented using Microsoft Excel's Solver function. The solving method chosen was the GRG Nonlinear engine.

## IMC and SIMC Tuning Methods

For the sake of brevity, the equations for these two methods are given in the Appendix. Table 3 summarizes the tun-



ing parameters obtained using the three tuning methods. Also, simulation and experimental results for the maximum voltage rate of change, IAE, and ISE are included. It is worth mentioning that the controller execution was set to 3 seconds.

**Table 3. Tuning Parameters**

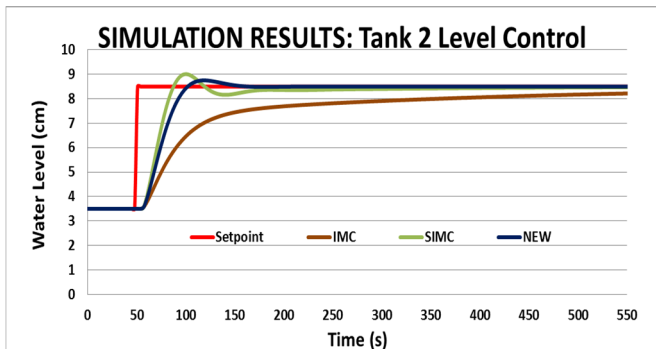
Variable	Optimization Method	IMC	SIMC
Proportional Gain, KC (VDC/cm)	0.099	0.041	0.155
Integral Time, TI (s)	103.41	198.0	198.0
Max Voltage Rate of Change – simulation	0.17	0.07	0.26
IAE-simulation (cm)	46.44	187.71	56.87
ISE-simulation (cm <sup>2</sup> )	163.31	319.20	145.10
Max Voltage Rate of Change – experiment	0.19	0.12	0.39
IAE-experiment (cm)	189	500	204
ISE-experiment (cm <sup>2</sup> )	553	1255	403

The closed-loop system performance under PI control was tested for a setpoint change of 5 cm, using the previously noted sets of tuning parameters. Simulation and experimental results were obtained using the twin-tank model and the experimental process, respectively.

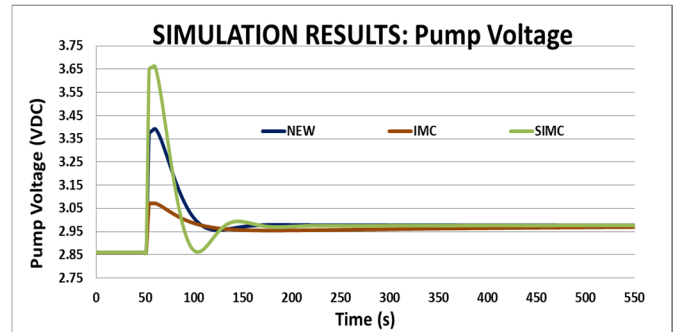
## Closed-loop Tank 2 Level Control

### Simulation Results

Figures 6 and 7 show the simulation results from the evaluation of the different tuning methods. Figure 6 shows the response of the controlled variable, tank 2 level, to a step change in its setpoint. Figure 7 shows the movement of the manipulated variable.



**Figure 6. Tank 2 Level Control (Simulation)**



**Figure 7. Pump Voltage (Simulation)**

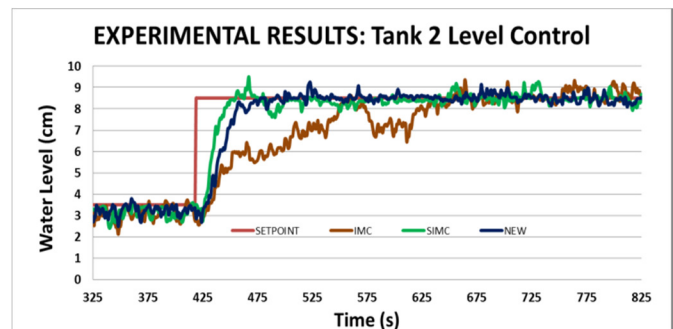
Based on the simulation results and the IAE reported in Table 3:

1. The optimization method (labeled as NEW) yielded the least IAE subject to a maximum voltage rate constraint of 0.17 VDC/s. In the experimental runs, the resulting maximum voltage rate of change was 0.19 VDC/s.
2. The IMC method had the largest IAE and ISE but was the most conservative of the three methods.
3. The SIMC method had an IAE close to that of the proposed method, but was the most aggressive of the three methods.
4. The SIMC method had the least ISE of the three methods. By minimizing the IAE, there is no guarantee that the ISE will also be minimum.

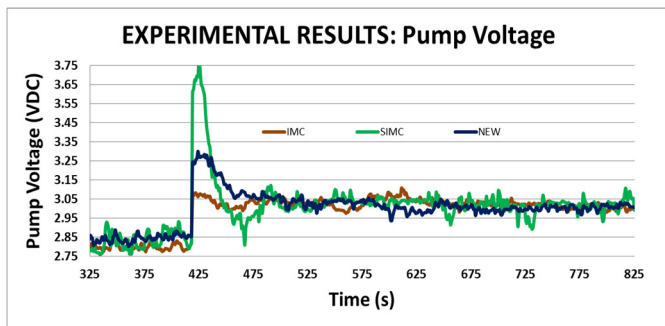
In conclusion, the optimization method performed at least as well as the widely used benchmark IMC and SIMC tuning methods. It balanced speed of response and control performance subject to constraints.

### Experimental Results

In this section, similar results are presented using the experimental process instead of the process model. Figure 8 shows the closed-loop performance of a PI controller for a step change in the setpoint for the level of tank 2. Figure 9 shows the change in the manipulated variable.



**Figure 8. Tank 2 Level Control (Experiment)**



**Figure 9. Pump Voltage (Experiment)**

It is apparent that the experimental results are in agreement with the simulation results and support the observations made earlier. The optimization method performed at least as well as, if not better than, the other tuning methods, while taking into consideration process constraints.

## Conclusions

In this study, the tuning method proposed by the authors in a previous study [1] was evaluated using simulation and experimental studies on a non-linear, interacting twin-tank system. A PI controller was tuned, subject to controlled variable, manipulated variable, rate of change constraints, other important process constraints, and tuning parameter constraints. Simulation and experimental results indicated that the optimization method performed well when compared to the benchmark tuning methods of IMC and SIMC.

Furthermore, this work also dealt with the development and validation of a dynamic, non-linear model of the twin-tank process. The non-linear model was employed by the optimization method in the calculation of tuning parameters.

## References

- [1] Tzouanas, C., & Tzouanas, V. (2013). Optimal Tuning of PID Controllers Subject to Process Constraints. *International Journal of Modern Engineering*, 14(1), 70-80.
- [2] Desbourough, L., & Miller, R. (2002). Increasing customer value of industrial control performance monitoring - Honeywell's experience. *Sixth International Conference on Chemical Process Control, AIChE Symposium Series*, 326(98), 172-192.
- [3] O'Dwyer, A. (2003). *Handbook of Controller Tuning Rules*. Imperial College Press, London.
- [4] Rivera, D. E., Morari, M., & Skogestad, S. (1986). Internal model control 4. PID controller design. *Industrial & Engineering Chemistry Process Design and Development*, 25, 252-265.
- [5] Skogestad, S. (2003). Simple analytic rules for model reduction and PID controller tuning. *Journal of Process Control*, 13(4), 291-309.
- [6] Astrom, K. J., Panagopoulos, H., & Hagglund, T. (1998). Design of PI controllers based on non-convex optimization. *Automatica*, 34(5), 585-601.
- [7] Tavakoli, S., & Tavakoli, M. (2003). Optimal Tuning of PID Controllers for First Order Plus Time Delay Models Using Dimensional Analysis. *The Fourth International Conference on Control and Automation (ICCA'03)*. Montreal, Canada.
- [8] Madhuranthakam, C. R., Elkamel, A., & Budman, H. (2008). Optimal tuning of PID controllers for FOPTD, SOPTD and SOPTD with lead processes. *Chemical Engineering and Processing*, 47, 251-264.
- [9] Martins, F. (2005). Tuning PID Controllers Using the ITAE Criterion. *International Journal of Engineering Education*, 21(5), 867-873.
- [10] Romero, J. A., Sanchis, R., & Balaguer, P. (2011). PI and PID auto-tuning procedure based on simplified single parameter optimization. *Journal of Process Control*, 21, 840-851.
- [11] Sanchis, R., Romero, J. R., & Balaguer, P. (2010). Tuning of PID Controllers based on simplified single parameter optimization. *International Journal of Control*, 83(9), 1785-1798.

## Biographies

**CONSTANTINE TZOUANAS** is currently a junior at Clear Lake High School, Houston, TX. His interests focus on economics and engineering.

**VASSILIOS TZOUANAS** is an Assistant Professor of Control and Instrumentation at the University of Houston – Downtown. Dr. Tzouanas earned a Diploma in Chemical Engineering from Aristotle University, a Master of Science degree in Chemical Engineering/Process Control from the University of Alberta, and a Doctor of Philosophy degree in Chemical Engineering/Process Control from Lehigh University. His professional experience includes technical and management positions with major operating companies. His research interests focus on process control systems, process modeling and simulation. He is a member of AIChE and ASEE. Dr. Tzouanas can be reached at [tzouanasv@uhd.edu](mailto:tzouanasv@uhd.edu).

---

## Appendix

In this Appendix, the equations for the different tuning methods used in this study are summarized. It was assumed that the process model was of the form:

$$G_p(s) = \frac{K_p \cdot e^{-\theta s}}{\tau_p \cdot s + 1} \quad (29)$$

IMC Method [4]:

$$K_c = \frac{1}{K_p} \cdot \frac{\tau_p}{\tau_p + \theta} \quad (30)$$

$$\tau_i = \tau_p \quad (31)$$

Simplified IMC (SIMC) Method [5]:

$$K_c = \frac{0.5}{K_p} \cdot \frac{\tau_p}{\theta} \quad (32)$$

$$\tau_i = \min(\tau_p, 8\theta) \quad (33)$$

# ALTERING THE DENSITY OF FERROFLUIDS THROUGH THE USE OF MAGNETIC FIELDS: A PROOF-OF-CONCEPT STUDY FOR THE POST-CONSUMER PLASTIC RECYCLING INDUSTRY

Matthew Franchetti, University of Toledo; Andrew Moening, University of Toledo; Connor Kress, University of Toledo

## Abstract

In this study, the authors developed a technique that utilizes a novel process to sort plastic particles of varying types by using electromagnetic (EM) waves and ferrofluids. The process involves placing various types of shredded plastic particles into a tank containing a ferrofluid and subjecting the plastic particles and ferrofluid to an EM wave using an EM coil. The EM wave creates a density-changing effect in the ferrofluid and causes the shredded plastic particles to rise and sink at different vertical levels within the ferrofluid tank, based on the plastic particles' respective densities. This method is an efficient, accurate, and low-cost method for sorting plastic particles as compared to conventional technologies. Overviews of the methodology, experimental design, and test results are presented here to demonstrate a proof-of-concept technique. Experimental testing resulted in a density-changing effect of 2.88% within the ferrofluid when subjected to the EM wave.

## Introduction

Since 1960, the U.S. Gross Domestic Product (GDP) has increased from \$520.5 billion to \$15.7 trillion in 2012 [1]. During this same timeframe, waste generation in the U.S. has increased from 88.1 million tons to 249.9 million tons [2]. As the U.S. has grown more productive, and as more goods are consumed, more waste is being generated. In order to address the increased waste-generation problem, improved methods and technologies are required. Recycling is becoming a more viable option, due to space constraints and the use of finite resources for the production of goods. Higher recycling levels can be used as a tool to combat these problems. If the rate and scale of recycling is to be increased, the recycling industry and the related sortation technologies will need to be upgraded and advanced in terms of accuracy, efficiency, and cost.

Of the products that are recycled from the waste stream, plastic is of significant importance. Plastic refuse is problematic for society and has low recycling rates; for example, the U.S. generated 31.8 million tons of plastic waste in 2011

of which only 8.3% was recycled [2]. Compared with other commonly recycled materials, such as glass and aluminum which were recycled at rates of 27.6% and 20.7%, respectively, plastic recycling is significantly lower [2]. With increasing demand for lower-cost plastics and heightened public attention to environmental concerns, the expanding recycling industry has provided an opportunity to lower raw material costs and create jobs. This is expected to be greatly influenced by the U.S. Recycling Works Program. The Recycling Works Program is a recycling initiative that establishes a 75% solid municipal waste recycling goal by the year 2015 [3]. While a recycling rate of 75% may sound unattainable, as the current national recycling rate is approximately 33% [3], several states and cities have positively demonstrated that such results can be achieved. For example, San Francisco has a recycling rate of 70%, and the state of California itself has a recycling rate of 58% [2]. In contrast to these areas with high recycling rates, there are eleven states with recycling rates below 10%, indicating significant room for improvement. Along with possible subsidies from the federal government, the program is expected to generate 1.5 million new jobs in both recycling and post-manufacturing waste recovery [3]. In the arena of waste processing, the costs associated with modern technologies can be difficult to justify, due to the need for large capital investments. If federal and state governments were to subsidize these operations, it could significantly influence their adoption. With this potentially large increase in recycling levels, the recycling industry would benefit from a technology that can quickly and cheaply sort plastic.

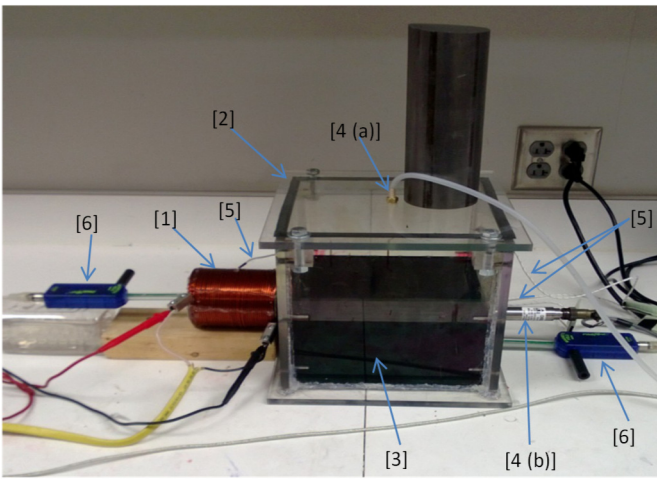
The primary objective of this current study was to provide proof of concept for a new plastic sortation technique that utilizes ferrofluids and electromagnetic (EM) waves as the primary mechanism for sortation. The process involves placing various types of shredded plastic particles into a tank filled with a ferrofluid. The plastic particles and ferrofluid are then subjected to an EM wave generated using an EM coil. A density-changing effect in the ferrofluid can be caused by the EM wave. By altering the density of the ferrofluid, different types of plastic can be made to float to the surface of the fluid where they can then be skimmed off as the means of separation. This is an efficient, accurate,



and low-cost method for sorting plastic particles as compared to conventional technologies. The use of EM waves to sort metal scrap has been applied previously in the recycling industry to sort aluminum scrap [4], [5] and for the elimination of non-ferrous scrap using electromagnetic filtration [6]. This study expands the use of EM waves in the recycling field to a new application and process related to plastic materials.

## Materials and Methods

In Figure 1 is an image of the experimental setup that was used for testing. Every component of the system is numbered, followed by a corresponding description of each item.



**Figure 1. Experimental Setup**

1. Solenoid Design
2. Pressure Vessel
3. Ferrofluid
4. Pressure Data Acquisition
  - a. Manometer
  - b. Pressure Transducer
5. Thermal Data Acquisition
6. Magnetic Field Sensor

The test vessel consisted of a 0.2m x 0.2m x 0.2m anti-static, sealed acrylic box (labeled 2 in Figure 1). The vessel was pressure tested prior to experimentation in order to ensure the accuracy of the measurements. The experimental configuration allowed for the precise measurement of the pressure within the ferrofluid by use of a pressure transducer (labeled “a” in Figure 1), and the air in the test vessel above the ferrofluid using a manometer (labeled “b” in Figure 1).

## Theory and Calculations

The method developed in this study involved submerging plastic particles in a ferrofluid. Because the ferrofluid responds to magnetic fields, it can be manipulated via an externally applied magnetic field gradient; the method is based on the concept that nonmagnetic bodies or magnetic holes in ferrofluids will experience repulsive forces from sources of magnetic field gradients [7]. The interrelationship of a magnetic moment in a medium with a magnetic field gradient results in a body force on the medium. This force is known as the magnetostatic force [8]. The void produced by the nonmagnetic particle (NMP) possesses an effective magnetic moment equal in magnitude and opposite in direction to the displaced fluid [9]. Therefore, an NMP in a ferrofluid with magnetization  $\vec{M}$  is equivalent to a magnetic particle with magnetization  $-\vec{M}$  in a nonmagnetic fluid [10]. Consequently, the force on a nonmagnetic body submerged in a ferrofluid, provided the magnetic field is relatively constant across the body volume, can be written as shown in Equation (1) [10], [11]:

$$\vec{F}_M = -\mu V(\vec{M} \cdot \nabla)\vec{H} \quad (1)$$

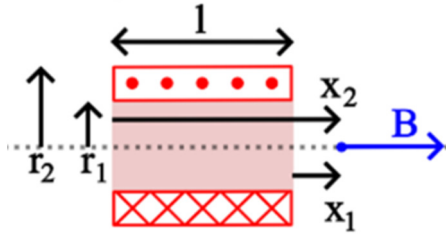
where,  $V$  is the volume,  $\mu V$  is the permeability of free space,  $\vec{M}$  is the magnetization of the ferrofluid, and  $\vec{H}$  is the magnetizing field. The force expression in Equation (1) was found and utilized in different applications involving magnetic holes [8], [12].

The concept of a solenoid is similar to the concept of a magnetic field being created by an electric wire. In the case of a solenoid, the wire is formed into a specific shape to achieve the desired result. A solenoid consists of a wire, which is wound into tightly packed coils. For the solenoid used in this current experiment, the wires were wrapped around an iron core. The use of a solenoid allows for a uniform and controllable magnetic field. The magnetic field of a solenoid is explained by Equation (2) and Figure 2, as derived from previous research studies [13]:

$$B = \frac{\mu_0 \mu_r i}{2d} \left[ \frac{x_2}{\sqrt{x_2^2 + r^2}} - \frac{x_1}{x_1^2 + r^2} \right] \quad (2)$$

where,

$$\begin{aligned} \mu_0 &= 4\pi \times 10^{-7} \text{ H/m} \\ \mu_r &= 2425 \\ i &= 4 \text{ Amperes} \\ d_{12 \text{ AWG wire}} &= 0.002053 \text{ m} \\ x_1 &= 0.025 \text{ m} \\ x_2 &= 0.225 \text{ m} \\ r &= 0.02 \text{ m} \end{aligned}$$



**Figure 2. Single-layer Iron-core Solenoid**

From Equation (2), the calculated strength of the electromagnet was 0.64 tesla or approximately 6400 gauss. This is roughly 10 times the amount of magnetic field intensity needed for the saturation of the ferrofluid, given that the ferrofluid requires 650 gauss to become excited [13]. A strong magnetic field was desirable because of the expected loss of intensity resulting from the rapid change in magnetic field. This loss is referred to as attenuation. Attenuation is commonly found when solenoids use an alternating current to manipulate magnetic fields at the frequency of the current. Although this loss is well documented, attenuation could not be calculated because of the complexity of induction between the core and ferrofluid.

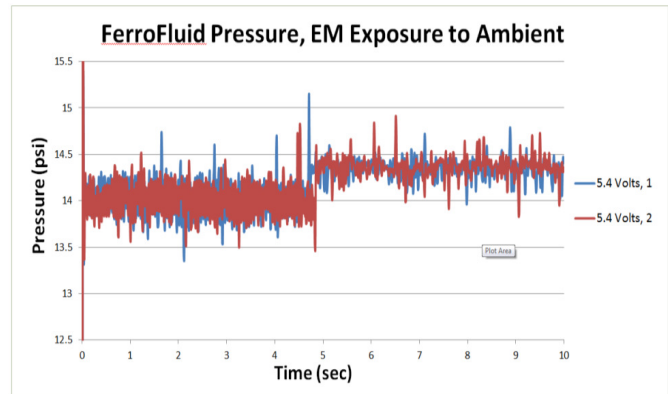
The change in density of the ferrofluid could not be directly measured with the test equipment. The ideal gas law was used to prove a density-changing. The two variables were related by the following function:

$$PV = nRT \quad (3)$$

where,  $P$  is absolute pressure,  $V$  is volume,  $T$  is absolute pressure,  $n$  is the number of moles,  $R$  is the universal gas constant, and  $T$  is the absolute temperature. Assuming that the temperature was constant along with  $R$ , density would change inversely with pressure. Since the system was closed and the test box was rigid and sealed, the volume did not change by any significant amount. It was hypothesized that the excitation of the ferrofluid by the EM waves would cause the density of the fluid to increase. This would in turn decrease the volume of the ferrofluid causing a vacuum to be created in the air. Therefore, a data acquisition system was designed to measure the negative pressure change in the air in order to derive the pressure change in the ferrofluid.

To measure pressure changes within the ferrofluid, tests were conducted at three voltages for the EM coil: 5.4V, 10.3V, and 14.4V. For each voltage level, 15 dynamic tests were conducted within a timeframe of 10 seconds at a sample rate of 100 samples per second. Before each test, the ambient pressure of the ferrofluid was measured. During the first 5 seconds of the test, the measurements were recorded during EM exposure; during the last 5 seconds, the measurements were recorded at ambient conditions (i.e., with no

EM exposure). The average pressure was calculated for each of the 15 tests during the exposure to the EM wave (first five seconds) and with no exposure to the EM wave (last five seconds). Figure 3 depicts the sample test data at an input of 5.4 V over the 10-second test interval for two representative samples.



**Figure 3. Ferrofluid Pressure with an Input Voltage of 5.4V**

In all, 15 ten-second tests were conducted for each of the three voltage levels (5.4 V, 10.3 V, and 14.4 V) for a total of 45 tests. Table 1 displays the test data for the paired observations (with and without EM exposure) for the 15 tests conducted at 5.4V.

**Table 1. Paired Ferrofluid Pressure Observations at 5.4V**

Sample	Ambient Pressure of Ferrofluid (Pa)	Pressure of Ferrofluid with EM Activated (Pa)	$d_i$
1	0.1157	0.1177	-0.00195
2	0.1156	0.1176	-0.00200
3	0.1157	0.1177	-0.00200
4	0.1158	0.1176	-0.00180
5	0.1156	0.1177	-0.00210
6	0.1157	0.1177	-0.00200
7	0.1157	0.1177	-0.00200
8	0.1157	0.1177	-0.00200
9	0.1156	0.1176	-0.00200
10	0.1157	0.1178	-0.00210
11	0.1156	0.1176	-0.00200
12	0.1157	0.1178	-0.00210
13	0.1156	0.1177	-0.00210
14	0.1157	0.1178	-0.00210
15	0.1158	0.1177	-0.00190

Hypothesis tests were conducted on the results at each voltage level to compare the mean pressures of the ferrofluid with and without EM exposure. The pressures were assumed to be normally distributed and a paired t-test was conducted at the 0.05 level of significance to test whether the ferrofluid pressure altered when exposed to the EM wave at 5.4 V, 10.3 V, and 14.4 V. A paired t-test was selected to account for ambient pressure differences of the ferrofluid during the tests. For the hypothesis tests,  $\mu_1$  and  $\mu_2$  represented the mean pressure of the ferrofluid at ambient pressure (no EM exposure) and while exposed to the EM wave, respectively. The null and alternate hypotheses for each voltage at the  $\alpha=0.05$  level of significance were:

$$H_0 : \mu_1 = \mu_2 \quad (4)$$

$$H_0 : \mu_1 \neq \mu_2 \quad (5)$$

The test statistic used is given by Equation (6):

$$t = \frac{\bar{d}}{s_D \sqrt{n}} \quad (6)$$

where,  $\bar{d}$  is the mean pressure difference,  $s_d$  is the standard deviation of the pressure differences for the 15 samples, and  $n$  is the sample size. Considering a sample size of 15 for each test, the critical region was  $t < -2.145$  and  $t > 2.145$ .

## Results and Discussion

Table 2 displays the hypothesis test results for the 45 samples (15 samples for each voltage level of 5.4 V, 10.3 V and 14.4 V).

**Table 2. Experimental Data**

Voltage	Mean Ferrofluid Pressure Difference Ambient versus EM Activated (Pa)	Standard Deviation of Ferrofluid Mean Pressure Difference Ambient versus EM Activated (Pa)	t-test Statistic	Decision
5.4V	-0.00201	8.484E-05	-6.118	Reject $H_0$
10.3V	-0.00307	8.484E-05	-8.155	Reject $H_0$
14.4V	-0.00334	9.854E-05	-8.741	Reject $H_0$

As displayed in Table 2, all null hypotheses were rejected, indicating that a mean pressure difference exists at the 0.05 level of significance when the ferrofluid is exposed to an EM wave at each voltage level. From a safety standpoint,

the temperature of the ferrofluid and EM coil were monitored. The temperature in the coil rose slightly during testing; at the worst case scenario with an input voltage of 14.4 V, the temperature of the coil rose 10 Co. The air and ferrofluid within the chamber remained at a constant temperature throughout all of the testing.

For the proof of concept related to sorting plastic particles in a more efficient and cost-effective system, the change in the density of the ferrofluid is the key parameter to allow for sortation. Table 3 displays the input voltages and the ferrofluid pressure changes at 5.4 V, 10.3 V, and 14.4 V. Table 4 displays the densities of several common plastic types.

**Table 3. Pressure Change Comparison**

Voltage	Mean Ferrofluid Pressure at Ambient (Pa)	Mean Ferrofluid Pressure with EM Activated (Pa)	Difference (Pa)	Difference (%)
5.4V	0.1157	0.1177	-0.00201	-1.74%
10.3V	0.1157	0.1188	-0.00307	-2.65%
14.4V	0.1157	0.1190	-0.00334	-2.88%

**Table 4. Specific Gravities by Plastic Type**

Plastic Type	Specific Gravity (Pa)
Polypropylene	0.0916 - 0.0925
Low-density Polyethylene	0.0936 - 0.0955
High-density Polyethylene	0.0956 - 0.0980
Bulk Polystyrene	0.1050 - 0.1220
Polyvinyl Chloride	0.1304 - 0.1336
Polyethylene Terephthalate	0.1330 - 0.1400

When comparing Tables 3 and 4, a density-changing within the ferrofluid at a given voltage range will allow plastic types to sink or swim within the ranges of the density changes.

It is significant that a change in density within the ferrofluid was observed during experimentation by subjecting the ferrofluid to the EM wave. The experiment indicated a maximum density change in the ferrofluid of 2.88% or 0.00334 Pa, well within the range of non-olefin plastics. Based on experimentation results, higher voltages beyond the test parameters (ranging from 5.4 V to 14.4 V) may result in higher density changes as well. The ability to control the density of the ferrofluid makes it possible to sort plastics of varying densities using similar test configurations

demonstrated for these experiments. For example, if this system were being used to sort two plastics with different densities, the density of the ferrofluid could first be manipulated to be between the two plastic densities. This would result in one of the plastic types rising to the surface, while the other type would sink towards the bottom. The plastic that floated to the surface could then be collected through mechanical methods, such as skimming.

## Conclusions

The experiment provides a proof of concept that it is possible to manipulate the internal pressures of ferrofluids using a predetermined EM pulse. After applying a low-pass filter, the research team was able to determine the resulting net change of pressure and specific gravity (see Table 3). When exposed to the pulse, the ferrofluid creates a magnetostatic force that causes a decrease of specific gravity. This novel method offers several advantages over existing methods for the sortation of plastic particles. This method is able to sort shredded plastic particles, whereas the most commonly used method of near infrared scanning is utilized for whole plastic containers. From a cost standpoint, the ferro-EM method requires significantly less equipment and space versus the most common method of using differential fluids. The differential-fluid method requires several sequenced tanks containing fluids of different densities to cascade the plastic particles through and achieve sortation. This ferrofluid-EM-based method would only require one tank and less fixed equipment to move the particles through the system, thereby achieving cost advantages.

Limitations of this study and possible sources of error are related to the ferrofluid type that was used for the study; for this study, an oil-based fluid was utilized that was purchased from a third-party vendor. Due to the limitations of the research equipment, only a voltage range between 5 V and 15 V could be tested; larger voltages may cause larger density-changing effects based on the test results. Finally, the tests examined short time frames of 10 seconds. Larger test windows may produce different results and will be explored in future studies. Additionally, the different types of plastics have overlapping specific gravities and densities; this may result in different plastic types with the same specific gravity floating at the same magnetic field exposure.

This study serves as a proof of concept of the ferro-EM method of sorting shredded plastic. From this experiment, the research team will progress into stage two of this experiment and re-examine the core design and the frequency of the voltage in order to maximize and refine the magnetic pulse to produce more accurate changes in the specific grav-

ity of the ferrofluid. Future plans also include the design and creation of a larger scale test setup.

## References

- [1] The World Bank, Gross Domestic Product. (n.d.). Retrieved June 19, 2013, from <http://data.worldbank.org/indicator>.
- [2] The EPA, Wastes – Resource Conservation: Plastics. (n.d.). Retrieved June 19, 2012, from <http://www.epa.gov/osw/conserve/materials/plastics.html>.
- [3] Recycling Works. Federal Policy Recommendations. Recycling Works! Campaign: A Jobs, Energy and Climate Solution. (n.d.). Retrieved January 5, 2014 from <http://www.recyclingworkscampaign.org/federal-policy-recommendations/>.
- [4] Mesina, M. B., DeJong, T. P. R., & Dalmijn, W. L. (2007). Automatic sorting of scrap metals with a combined electromagnetic and dual energy X-ray transmission sensor. *International Journal of Mineral Processing*, 82, 222-232.
- [5] Krivtsova, G. B., Pimenov, A. N., & Petukhov, V. V. (2009). Electromagnetic separation of nonferrous metals before a metallurgical conversion in a field of high-frequency current. *Metallurgist*, 53, 524-530.
- [6] Zhenming, X., Tianxiao, L., & Yaohe, Z. (2003). Elimination of Fe in Al-Si cast alloy scrap by electromagnetic filtration. *Journal of Materials Science*, 38, 4557-4565.
- [7] Kose, A. R., Fischer, B., & Koser, H. (2008). Towards Ferro-microfluidics for Effective and Rapid Cellular Manipulation and Sorting. *Proceedings of the 3rd IEEE Int. Conf. on Nano/Micro Engineered and Molecular Systems*, 903-906.
- [8] Berkovsky, B. M. (1993). *Magnetic Fluids*. Oxford Science Publications, New York.
- [9] Skjeltorp, A. T., Clausen, S., & Helgesen, G. (2001). Ferrofluids, complex particle dynamics and braid description. *Journal of magnetism and Magnetic Materials*, 226-230(1), 534-539.
- [10] Drenckhan, W., Elias, F., Hutzler, S., Weaire, D., Janiaud, E., & Bacri, J.-C. (2003). Bubble size control and measurement in the generation of ferrofluid foams. *Journal of Applied Physics*, 93(12), 107-118.
- [11] Zahn M., & Greer, D. R. (1995). Ferrohydrodynamic pumping in spatially uniform sinusoidally time varying magnetic fields. *Journal of Magnetism and Magnetic Materials*, 149, 165-173.
- [12] Rosensweig, R. E. (1985). *Ferrohydrodynamics*. Dover Publications, Inc., Mineola New York.
- [13] Tan, H., & Elahinia, M. E. (2008). A nonlinear model for ferromagnetic shape memory alloy actuators.



## Biographies

**MATTHEW FRANCHETTI** is an Assistant Professor of Mechanical, Industrial, and Manufacturing Engineering at The University of Toledo. He earned his B.S., M.S., and Ph.D. degrees from The University of Toledo. Dr. Franchetti is currently teaching at the University of Toledo. His interests include sustainable systems and lean manufacturing. Dr. Franchetti may be reached at [matthew.franchetti@utoledo.edu](mailto:matthew.franchetti@utoledo.edu).

**ANDREW MOENING** is a Graduate Research Assistant at The University of Toledo. He earned his B.S. in Mechanical Engineering in 2013. Mr. Moening may be reached at [andrew.moening@rockets.utoledo.edu](mailto:andrew.moening@rockets.utoledo.edu).

**CONNOR KRESS** is Graduate Research Assistant at The University of Toledo. He earned his B.S. in Mechanical Engineering in 2014. Mr. Kress may be reached at [connor.kress@rockets.utoledo.edu](mailto:connor.kress@rockets.utoledo.edu).

# AN AUTONOMOUS PASSIVE RFID-ASSISTED MOBILE ROBOT SYSTEM FOR INDOOR POSITIONING

Kumar Yelamarthi, Central Michigan University

## Abstract

Indoor positioning systems (IPS) locate objects in closed structures such as warehouses, hospitals, libraries, and office buildings, where Global Positioning System (GPS) typically does not work due to poor satellite reception. Inherently, indoor positioning has been a vital challenge facing industry for a long time. Most of the available IPS operate based on optical tracking, motor encoding, and active RFID tags, often limited by accuracy or high hardware costs. Answering this challenge, the author of this current study developed a new passive radio frequency identification-based (RFID) localization system for indoor positioning using a mobile robot. The concept is based on placing passive RFID tags in a uniform triangular fashion for low tag density, and using an RFID reader on a mobile robot for localization and navigation. The inputs to the proposed systems are location coordinates stored on the RFID tags, and their respective time of arrivals as read by the RFID reader. The proposed system first estimates its location using a centroid method then utilizes time difference of arrival (TDOA) to accurately estimate its position and trigonometric identities to estimate its current orientation. Experimental results demonstrated that the proposed system can effectively perform indoor positioning in order to navigate to its destination with an average accuracy of 0.07 m, while avoiding any obstacles in its path.

## Introduction

Since the industrial revolution, machines have increasingly become more efficient and effective in assisting humans in a variety of tasks. In recent years, advancements in robotics have had a profound impact on the personal and professional lives of people around the world [1-9]. In industry, robots have become commonplace due to the difficult and often time-consuming tasks they can handle with uniformity, speed, and precision. Though the intellectual ability of industrial robots has matured in certain applications, it has not grown proportionally in certain applications. For instance, there has been tremendous interest in applications of navigation assistive robots that operate in unstructured environments, but limited development.

One of the fundamental challenges in mobile robotics and warehouse management is indoor positioning. While Global Positioning System (GPS) is a widely accepted solution for

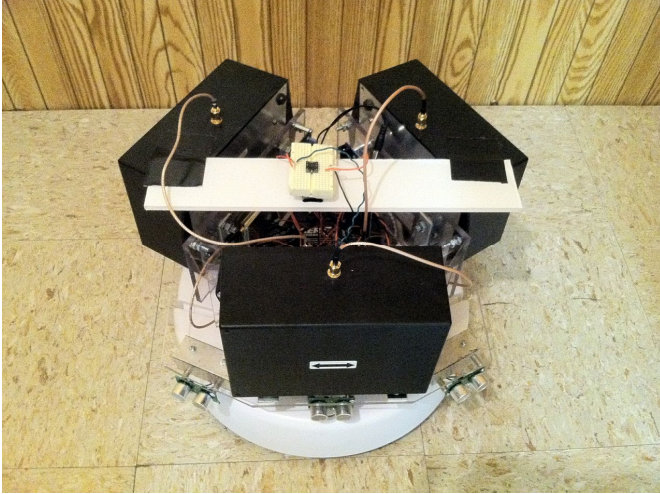
outdoor operation, its accuracy is very limited indoors due to lack of satellite reception. Responding to this challenge of indoor positioning and navigation would be of immense help for several applications such as navigational assistance for the blind, tour guide robots, inventory and asset tracking, and healthcare [1], [10-12]. With the primary application of radio frequency identification (RFID) being object identification, RFID technology has seen significant growth. While it can be used to efficiently identify an object, its scope is challenging in localization. According to Chunag [13], robust implementation of RFID-based positioning for a mobile robot is a challenging task, especially when using passive RFID tags, as communication between tag and reader is sensitive to the environment.

For a mobile robot to reach a certain location with a high degree of accuracy, it has to localize itself with the surrounding, calculate the shortest and safest route to its destination, monitor its path for obstacles, and safely travel the path towards the destination avoiding these obstacles. Such smart robots with indoor localization abilities can serve multiple applications. Identifying this opportunity, the author presents here the research, design, and implementation of an autonomous mobile robot and passive RFID-based positioning and navigation system, as shown in Figure 1. The proposed system utilizes a passive RFID system with three antennas placed in a triangular fashion for increased read range, and an array of ultrasonic (sonar) sensors in front of the robot to allow for maximum visibility of obstacles. Utilizing a combination of methods and an array of systems, the proposed IPS performs localization with an average accuracy of 0.07m.

## Previous Work and Research Objectives

### Robot Navigation

Robot navigation is its ability to find a safe path from a starting position to the end goal, while localizing itself in the environment using sensorial data. Given a map and a goal location, path planning involves finding a geometric path from the actual location of the robot to the goal/target [12]. While obstacles might remain static, the mobile robot holds a different purpose as it must evaluate its continuously changing trajectory in the environment.



**Figure 1. Prototype of an RFID-based Indoor Positioning System**

Many methods have been attempted in order to achieve an ideal path-planning algorithm for an indoor positioning robot. These methods are all different and all have positive and negative effects on the path of the robot and the overall project. A few methods that have been proposed in the past are Rapidly Exploring Random Trees (RRT), Generalized Sampling-based Methods, and Visibility Graphs [12]. RRT operates by accumulating environmental data and creating random probability paths over time based on the relative distance of the object to the robot. It allows the robot to traverse several different areas without exclusivity as it can continue to interact and adapt to new environments by continuously analyzing and computing the probability of an object's existence. The problem with RRT is that it does not allow for direct and optimal time travel to an end goal. Also, while it can analyze the environment efficiently, it fails at minimizing the time to reach the end goal as its computational power is focused on calculating probabilities. Overall, a problem in RRT is that it produces a path with many branches over time, due to its inherent behavior of using a randomized technique [14].

## RFID Localization

Many different methods have been proposed for localization of RFID systems including Angle of Arrival (AOA), Received Signal Strength Indicator (RSSI), and Time of Arrival (TOA). An AOA measurement determines the angle between the transmitter/receiver line and the reference direction [15], with accuracy highly dependent upon the number of receivers or the rotating device, and is more suitable for operating outdoors. RSSI, on the other hand, uses signal strength to determine the distance between sender and receiver [15]. The advantage of this method is its cost effi-

ciency and ease of implementation. The disadvantage, however, is that it requires heavy computation and extensive prior knowledge of the environment.

In addition, several researchers have proposed algorithms to increase the accuracy of RFID positioning. Chunag [13] utilized RFID technology along with object localization in order to locate tag locations. He used AOA and signal strength to determine the location of the RFID tags, but performance was limited to an average accuracy of 0.62 m. Schneegans et al. [16] focused on using RFID vision-based snapshots for mobile robot self-localization and obtained an accuracy of 0.4 m. Park and Hashimoto [17] used trigonometric functions and RFID tags' Cartesian coordinates in a regular grid-like pattern and stated that the RFID tag placement pattern should be adapted through experiments to suit different applications; then were able to obtain an average accuracy of 0.1 m. Kim and Chong [18] validated mobile robot docking with an RFID transponder in the area occupied by obstacles. Their algorithm quantifies the potential error in the Direction of Arrival (DOA) estimate through dual-directional RFID antenna. While each of these methods serves the basic purpose, all have limitations including accuracy, obstacle detection, and cost. Improvement of these projects presents a feasible RFID system to accurately detect a position while indoors.

Time of Arrival (TOA) is another method that operates based on estimating the distance between the sensor and target nodes with the presumption that this distance is directly proportional to the propagation time [19]. While this method can deliver promising results, it has two limitations. First, all sensor nodes and target systems must be precisely synchronized at the microsecond level. Second, all transmitted signals must have a timestamp incorporated into them in order to accurately estimate the distance traveled. These challenges in the TOA method can be overcome using the Time Difference of Arrival (TDOA) method, where the difference in signal arrival time at different sensor nodes is used for target localization. The TDOA measurement defines a hyperbolic area with the possible location of the target with two paired sensors as foci.

## Research Objectives

In this current study, the problems in existing RFID-based localization and navigation algorithms were considered and a potential solution was developed using a passive RFID-based indoor positioning system for a mobile robot. Presented here is a model that addresses the uncertainties in existing RFID-based localization systems by combining centroid and TDOA methods. With the assumption that the environment in which the mobile robot travels is well structured

with uniformly placed RFID tags, and the obstacles along the navigation path are primitive geometrical objects (straight with regular inclination and/or circular), the author developed a hierarchical localization and navigation algorithm with system architecture, as presented in Figure 2.

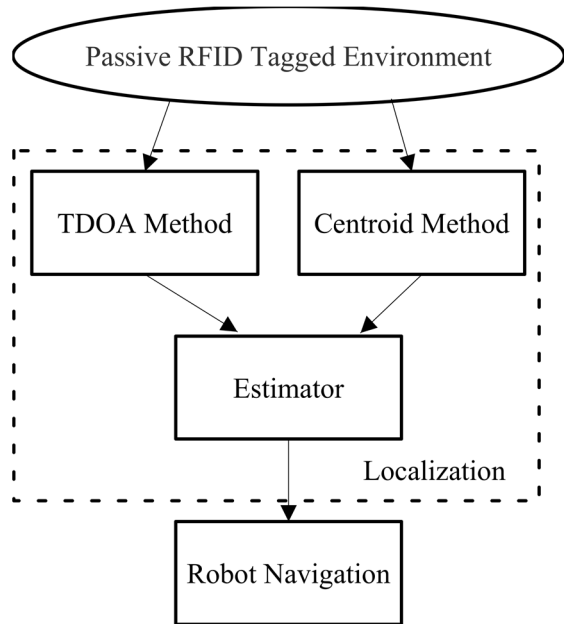


Figure 2. Proposed RFID Positioning System Architecture

## Design Implementation

The RFID Positioning Robot's chassis presented in Figure 1 was based on an iRobot platform [20]. The embedded system of the RFID Positioning Robot was classified into an input module and an output module, where the input module consisted of the localization device, the obstacle detection sensors, and the user input device. The output module consisted of the position display and the robot platform. Both of these modules were interfaced using an embedded microcontroller that serves as the central command module and provides information regarding the current location. Localization is performed using a SkyeTek DKM10 system [21]; obstacle avoidance navigation is performed using an array of ultrasonic range finders [22]. Three RFID reader antennas,  $A_1$ ,  $A_2$ ,  $A_3$ , were mounted on the top of the robot in an equilateral triangle in order to obtain an RFID tag read range of nearly  $360^\circ$ , as presented in Figure 3.

## Localization

Localization of the proposed RFID positioning robot was performed using a mobile robot equipped with an RFID reader and three antennas placed as shown in Figure 1, and

RFID tags placed on the floor in a triangular pattern, as shown in Figure 4. The absolute coordinates of the tag location are stored on the respective tags and sent to the reader upon request. During navigation from start to the destination, the RFID reader continuously monitors a scanning radius,  $r$ , and reads all of the tags in the vicinity. In addition to reading tag data, the system also records time of arrival from each tag. As read range of passive RFID tags is susceptible to factors such as multipath fading, tag orientation, distance, etc., a combination of methods was used in the proposed system.

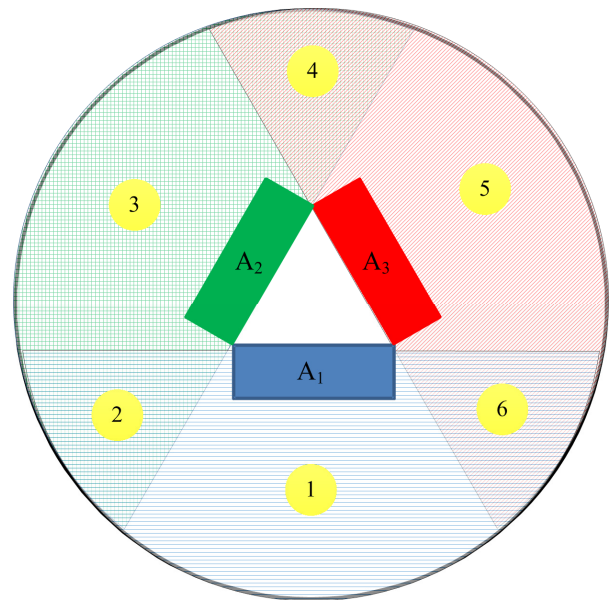


Figure 3. Theoretical Read Range of RFID Reader Antennas

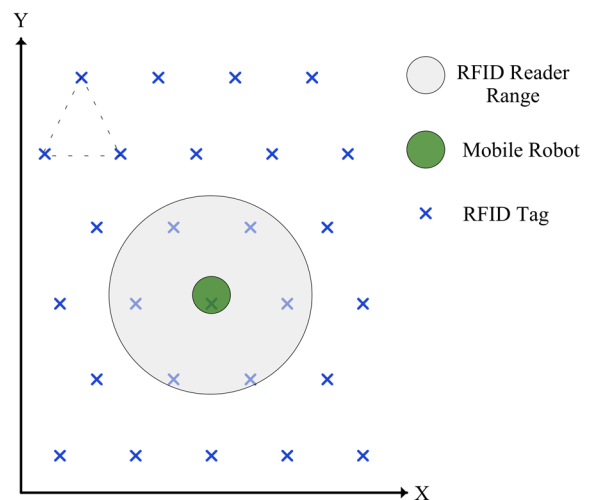


Figure 4. Triangular Placement of Passive RFID Tags and Read Range of RFID Positioning System



First, preliminary localization was performed using a centroid method. With the three RFID antennas placed on the robot, coarse location could be performed by identifying the antenna that would read the respective tags (see Figure 3). For instance, when a tag is read only by antenna  $A_1$ , it can be broadly stated that the tag is potentially located in regions 1, 2, or 6, as marked. Similarly, if it is read by antenna  $A_2$ , the tag is potentially located in regions 2, 3, or 4. With this foundation, passive RFID tags were placed in a triangular fashion for increased accuracy and lower density [23] with their respective locations stored on each tag. When the centroid-based localization [24] starts, the RFID reader scans the area and sequentially gathers information on all tags read and stores it in the memory. When all of this information is obtained, the position of mobile robot  $(x_{pose-1}, y_{pose-1})$  is estimated through Equations (1) and (2), where  $N$  represents the total number of passive RFID tags read, and  $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$  are the locations of the respective tags.

$$x_{pose-1} = \frac{\min\{x_1, x_2, \dots, x_N\} + \max\{x_1, x_2, \dots, x_N\}}{2} \quad (1)$$

$$y_{pose-1} = \frac{\min\{y_1, y_2, \dots, y_N\} + \max\{y_1, y_2, \dots, y_N\}}{2} \quad (2)$$

Second, localization was further performed using TDOA-based hyperbolas. To get a clear understanding of the TDOA method, consider a hyperbola with the transverse axis aligned with the  $y$ -axis, as in Figure 5, with the characteristic equations [see Equations (3)-(5)], where  $a$  is the distance from the center to either vertex;  $b$  is the length of the perpendicular segment between each vertex and the asymptotes; and  $c$  is the distance from the center to either of the focus points. For any point  $P(x, y)$  on either vertex, the absolute difference between the eccentricity  $|d_2 - d_1|$  is always equal to twice the distance between the center and the vertex. The proposed TDOA method works based on the assumption that the RFID reader is located somewhere on the hyperbola and uses the intersection of two or more hyperbolas to identify its respective location.

$$\frac{y^2}{a^2} - \frac{x^2}{b^2} = 1 \quad (3)$$

$$a^2 + b^2 = c^2 \quad (4)$$

$$|d_2 - d_1| = 2a \quad (5)$$

Consider a region where passive RFID tags ( $N_1, N_2, N_3, \dots$ ) are located at random locations  $((x_1, y_1), (x_2, y_2), (x_3, y_3), \dots)$ , as in Figure 6. When the RFID reader initiates its inventory search routine, it marks the TOA of each tag read in its sensing radius as  $T_1, T_2, T_3, \dots$ . Once this information is ob-

tained, the reader designates the tag with the lowest TOA as the reference tag then draws hyperbolas between the reference tag and three other tags. In the example presented in Figure 6,  $N_1$  was identified to be the reference tag and hyperbolas were drawn between  $N_1 - N_2, N_1 - N_3$ , and  $N_1 - N_4$ . Per the hyperbola characteristic equations [see Equations (3)-(5)], if  $T_1$  and  $T_2$  were TOA to  $N_1$  and  $N_2$ , the distance between vertices to the center could be computed using Equation (6), where  $V$  is the speed of target signal. Furthermore, based on the location of  $N_1$  and  $N_2$ , the distance between the tags,  $c$ , and length of the perpendicular segments from each vertex to the asymptotes,  $b$ , can be computed using Equations (7)-(8).

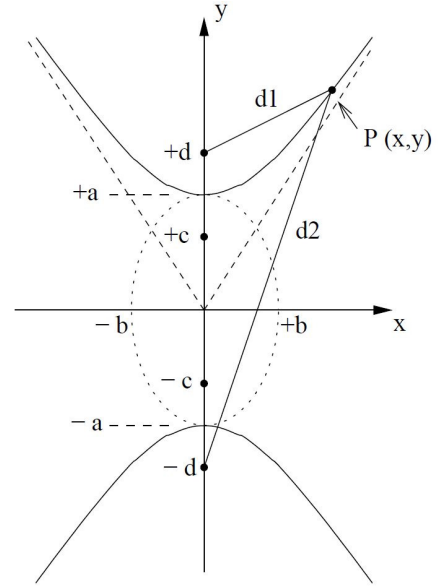


Figure 6. Intersection of Three Hyperbolas to Estimate RFID Reader Location

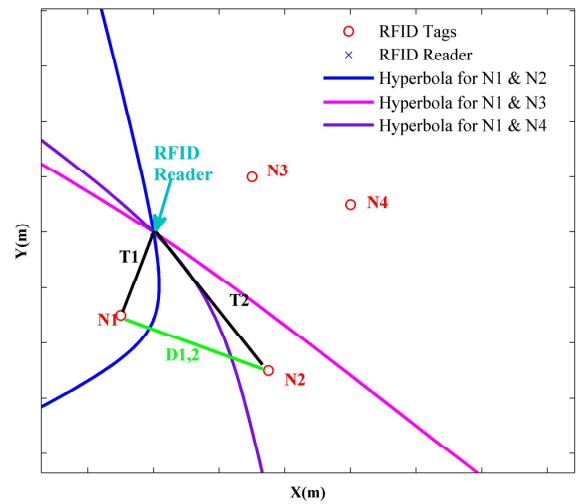


Figure 5. Embedded Geometry of Hyperbola [25]

$$a = \frac{T_2 - T_1}{2} = \frac{V\Delta t}{2} \quad (6)$$

$$c = \frac{D_{12}}{2} = \frac{\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}}{2} \quad (7)$$

$$b = \sqrt{c^2 - a^2} \quad (8)$$

With this foundation, consider a two-dimensional area  $(x, y)$  with four RFID tags  $(N_A, N_B, N_C, N_D)$  and their respective signal times of arrival from the target being  $T_A, T_B, T_C,$  and  $T_D,$  respectively. If  $T_D$  is the lowest then the TDOA between  $N_D$  and the other nodes can be calculated using Equations (9)-(11), and the distance between the nodes can be calculated using Equations (12)-(14). Solving Equations (12) and (13) for target location yields Equations (15)-(17); solving Equations (13) and (14) yields Equations (18)-(20). Furthermore, solving Equations (18) and (20), the intersection hyperbolas and the position of the RFID reader  $(x_{pose-2}, y_{pose-2})$  based on TDOA can be estimated using Equations (21)-(22). Finally, the true position  $(x_{pose}, y_{pose})$  of the RFID robot can be estimated using Equation (23).

$$T_{DA} = T_D - T_A \quad (9)$$

$$T_{DB} = T_D - T_B \quad (10)$$

$$T_{DC} = T_D - T_C \quad (11)$$

$$\sqrt{x_T^2 + y_T^2} - \sqrt{(x_A - x_T)^2 + (y_A - y_T)^2} = V \times T_{DA} \quad (12)$$

$$\sqrt{x_T^2 + y_T^2} - \sqrt{(x_B - x_T)^2 + (y_B - y_T)^2} = V \times T_{DB} \quad (13)$$

$$\sqrt{x_T^2 + y_T^2} - \sqrt{(x_C - x_T)^2 + (y_C - y_T)^2} = V \times T_{DC} \quad (14)$$

$$y = k_1 x + b_1 \quad (15)$$

$$k_1 = (T_{DC}x_A - T_{DA}x_C) / (T_{DA}y_C - T_{DC}y_A) \quad (16)$$

$$b_1 = \frac{T_{DA}(x_C^2 + y_C^2) - T_{DB}(x_A^2 + y_A^2) + V^2 T_{DB} T_{DA} (T_{DA} - T_{DB})}{2T_{DA}y_B - 2T_{DB}y_A} \quad (17)$$

$$y = k_2 x + b_2 \quad (18)$$

$$k_2 = (T_{DB}x_A - T_{DA}x_B) / (T_{DA}y_B - T_{DB}y_A) \quad (19)$$

$$b_2 = \frac{T_{DA}(x_B^2 + y_B^2) - T_{DB}(x_A^2 + y_A^2) + V^2 T_{DB} T_{DA} (T_{DA} - T_{DB})}{2T_{DA}y_B - 2T_{DB}y_A} \quad (20)$$

$$x_{pose-2} = (b_2 - b_1) / (k_1 - k_2) \quad (21)$$

$$y_{pose-2} = k_1 x_T + b_1 \quad (22)$$

$$b_2 = \frac{T_{DA}(x_B^2 + y_B^2) - T_{DB}(x_A^2 + y_A^2) + V^2 T_{DB} T_{DA} (T_{DA} - T_{DB})}{2T_{DA}y_B - 2T_{DB}y_A} \quad (23)$$

## Obstacle Detection and Avoidance

The first step in obstacle avoidance is to accurately identifying the location of the obstacle with respect to the robot. With an array of ultrasonic range finders at the front of the robot, (see Figure 7), a clear 180° obstacle-detection range can be obtained. Based on the preliminary assumption that these sensors are evenly spaced, the angle between any two consecutive sensors can be computed using Equation (24), where  $n$  is the number of sensors in the array. Accordingly, the angle between any two consecutive sensors was found to be 25.71°. In addition, the angle of sensor- $i$  with respect to the front of the robot was calculated using Equation (25). This information would assist in performing a coarse estimation of the location of the obstacle. When a sensor detects an obstacle, it outputs a distance,  $d$ , as marked in Figure 7. With each sensor having a unique obstacle detection angle, the location of the obstacle can be detected with fine accuracy using basic trigonometry. To do this, first find the coordinates of origin  $(x_o, y_o)$  then find the coordinates of each respective sensor using Equation (26), where  $i$  is the number of sensors. When the robot starts navigating, the sonar array continuously looks for obstacles and each sensor computes the distance  $(d_0, d_1, d_2, \dots, d_7)$  to the obstacles and determines its respective position using Equation (27).

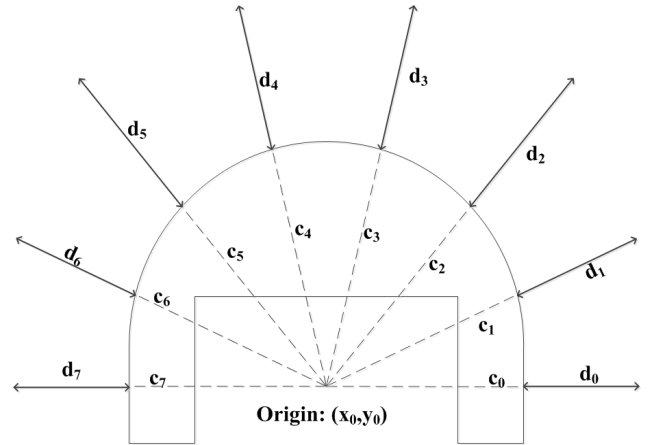


Figure 7. Placement of Ultrasonic Range Finders on the Robot

$$\theta = (180/n - 1) \quad (24)$$

$$\theta_i = 25.71 \times i \quad (25)$$

$$(x_i, y_i) = (c_i \times \cos(\theta_i), c_i \times \sin(\theta_i)) \quad (26)$$

$$\begin{pmatrix} x_{obs-i} \\ y_{obs-i} \end{pmatrix} = \begin{pmatrix} c_i \cos \theta_i + d_i \cos \theta_i \\ c_i \sin \theta_i + d_i \sin \theta_i \end{pmatrix} \quad (27)$$

When the ultrasonic range finder determines the potential location of an obstacle, the robot will determine the rotational angle for avoiding the obstacle. As presented in Figure 8, if the obstacle is determined to be in front of the robot at an angle  $\alpha^\circ$ , the robot will turn by  $(\alpha+90)^\circ$ , move forward until the sonar-0 no longer detects the obstacle within the unsafe zone then track the distance by  $x_d$ . At this point, the robot will turn by  $-90^\circ$ , move forward until sonar-0 no longer detects the obstacle in the unsafe zone then track the distance by  $y_d$ . Later, the robot will turn by  $-90^\circ$ , move forward by  $x_d$ , and again turn by  $90^\circ$  to reach coordinates on the original route. At this point, it will recalculate its position, compute the new rotational angle in order to reach its destination then move accordingly.

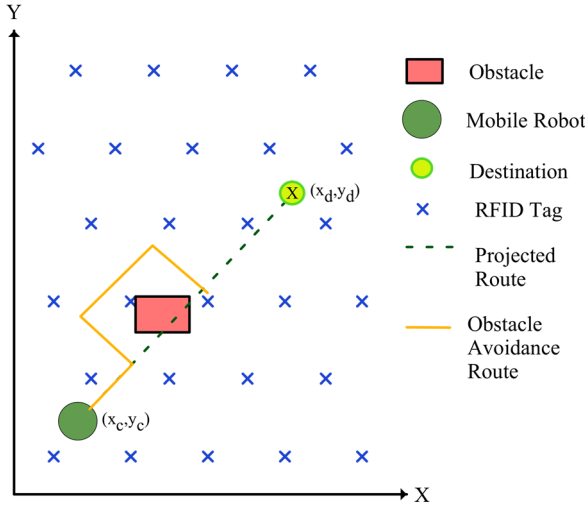


Figure 8. Obstacle Avoidance Subroutine

## Robot Navigation

Two important factors in robot navigation are localization and obstacle avoidance. The proposed mobile robot performs this localization using the TDOA method described earlier, and uses trigonometric identities to calculate its route to the destination. Assuming that the robot is at origin  $O(x_o, y_o)$ , and is provided its destination coordinates  $(x_d, y_d)$ , it will calculate the rotation angle ( $\theta_{origin-dest}$ ) and distance

( $d_{origin-dest}$ ) using Equations (28) and (29) in order to reach the destination. Next, it rotates by  $\theta_{origin-dest}$  and starts moving towards the destination and designates this current angle as  $\theta_{current}$ . Later, and periodically every 100 ms, the onboard RFID system estimates its current location  $(x_c, y_c)$  and recalculates the new angle ( $\theta_{current}$ ) to the destination using Equation (30). If the angle is found to be approximately equal to zero, the robot moves forward towards the destination. If not, the robot rotates accordingly and then moves towards the destination. In this manner, the robot repeats its rotation and forward movements until it reaches the destination. As the robot is moving forward, the onboard ultrasonic range finder constantly scans for potential obstacles. If any obstacle is detected, the obstacle avoidance subroutine (presented in the previous section) is invoked and the robot moves around it.

$$\theta_{origin-dest} = \text{Tan}^{-1} \left( \frac{y_d - y_o}{x_d - x_o} \right) \quad (28)$$

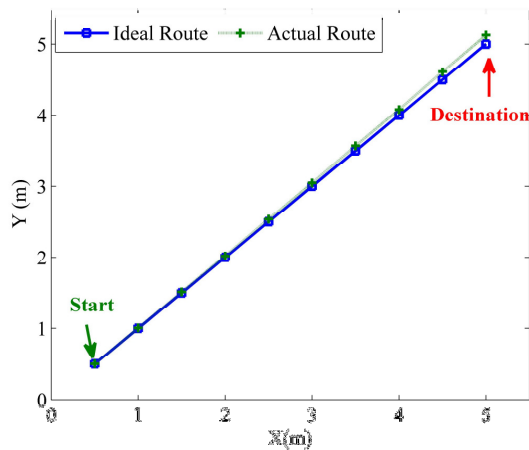
$$d_{origin-dest} = \sqrt{(y_d - y_o)^2 + (x_d - x_o)^2} \quad (29)$$

$$\theta_{current} = \text{Tan}^{-1} \left( \frac{y_d - y_c}{x_d - x_c} \right) - \theta_{current} \quad (30)$$

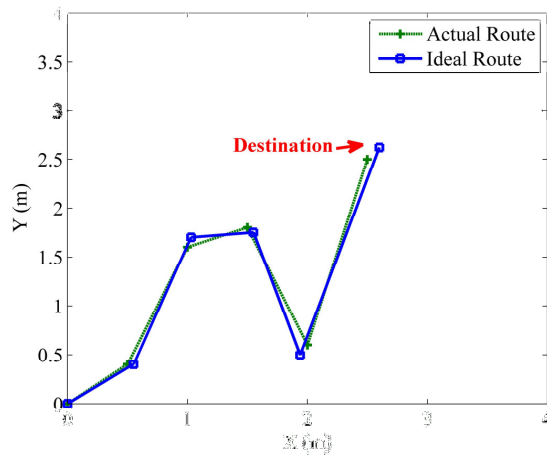
## Experiments and Results

The proposed approach was validated through a proof-of-concept system to perform passive RFID localization, and analyze the average error in doing so. The aim here was to demonstrate the feasibility of a dependable system for indoor positioning; to this end, several experiments were conducted focusing on testing each sub-system including: 1) the robot's movement accuracy; 2) the obstacle detection range; and, 3) RFID localization.

First, the robot's movement accuracy was tested via two different experiments [4]: following a line segment and following a complex path. The first experiment aimed at testing the basic functionality of the robot to track a straight line. The initial position of the robot was set to  $(0.5, 0.5)$ , and the destination was set to  $(5, 5)$ ; the respective movement pattern is shown in Figure 9. After 10 trial runs, the average position error at the destination was calculated to be 0.04 m. The second experiment aimed at testing the robot's movement when turning at acute angles. The initial position of the robot was set to  $(0, 0)$ , and the destination was set to  $(2.5, 2.5)$ , with acute turns at multiple locations, as shown in Figure 10. After 10 similar runs, the average position error at destination was calculated to be approximately 0.07 m.



**Figure 9. Robot Movement Accuracy while Following a Line Segment**

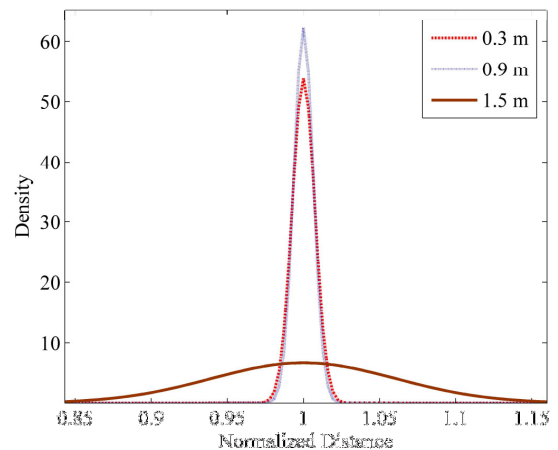


**Figure 10. Robot Movement Accuracy while Turning at Acute Angles**

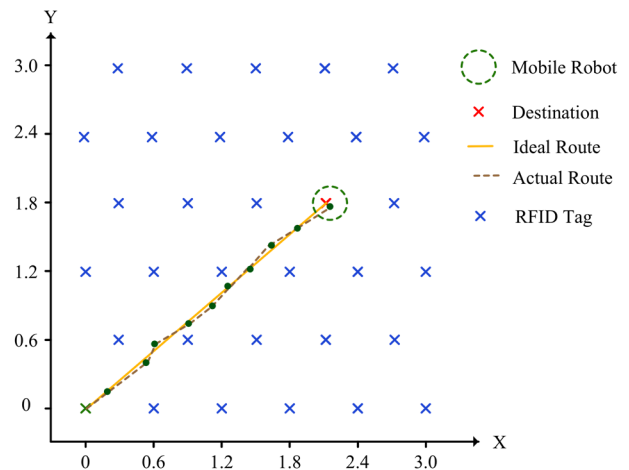
Second, the robot's ability to sense the environment and detect obstacles through the ultrasonic range finders was tested. Obstacles were placed at different distances from the robot and tests were performed while the robot was either stationary or moving. The sensing area was tested up to 1.5 m, due to the scope of this project. While the ultrasonic sensor can read up to 3 m, obstacles within the range of 1 m would be of importance. Figure 11 shows the distribution of the accuracy after 500 readings. From the data, it was found that distances less than 1 m had an error rate under 5%. In the scope of this proposed system scale, obstacles at distances greater than 1 m were not given any priority.

Finally, the efficiency of the proposed passive RFID localization for mobile robots was evaluated through multiple experiments: In the first experiment, the initial position of

the robot was set to (0,0), and the destination was set to (2.1,1.8), as shown in Figure 12. For this test in an indoor environment, 33 passive RFID tags were placed in a triangular pattern over an area measuring 3m x 3m, with spacing of 0.6 m between tags. During this preliminary test, the environment was free of obstacles. At the beginning, the rotation angle was computed using Equation (28) as the robot started moving towards its destination. Periodically, every second, the onboard RFID reader obtained an inventory of all tags in the vicinity for localization. Whenever the calculated position was found to be 0.05 m away from the ideal position on the route (green dots), a new rotational angle was calculated using Equation (30); the robot then rotated towards the destination accordingly. Repeating the experiment through multiple trials provided further validation, and an average positioning accuracy was found to be 0.076 m, as presented in Table 1.



**Figure 11. Accuracy of Ultrasonic Rangefinder at Different Distances**

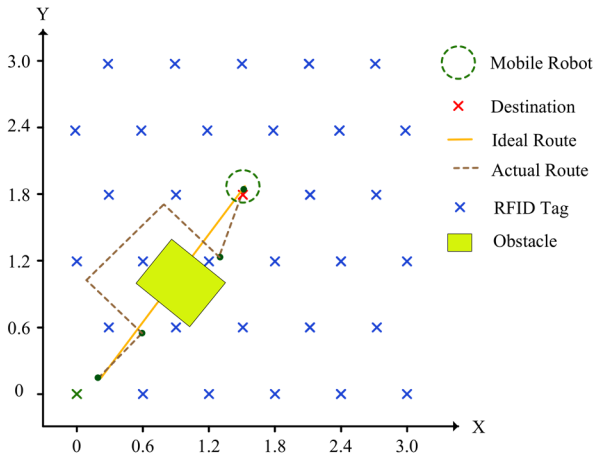


**Figure 12. RFID Positioning under Experiment 1**

**Table 1. RFID Localization Results of Experiment 1 from Multiple Trials**

Trial No.	Measured X (m)	Measured Y (m)	Error (m)
1	2.09	1.85	0.050
2	2.05	1.86	0.078
3	2.17	1.76	0.080
4	2.03	1.78	0.072
5	2.20	1.82	0.100
Average			0.076

To validate the accuracy of the proposed RFID localization in the presence of obstacles in the travel path, a second experiment was conducted with a similar setup (see Figure 13). After several trials, it was found that the robot was able to efficiently avoid obstacle and still reach its destination with an accuracy of 0.07 m—a video demonstration is available on the Internet [26].



**Figure. 13: RFID Positioning under Experiment 2**

Table 2 shows a quantitative comparison of the proposed IPS with published previous work. The object localization strategy proposed by Chunag [13] is clearly limited by its localization accuracy and tag density compared with the system proposed here. In addition, the Chunag system uses active RFID tags that are expensive, larger in physical size due to an onboard battery, and require periodic maintenance, unlike passive tags. Schneegans et al. [16] proposed an RFID snapshot-based method for localization with a very low tag density. This was achieved at a high cost of localization accuracy, thereby limiting the overall effectiveness of the system. On the other hand, Park and Hashimoto [17] proposed a localization strategy similar to the design of this current study. However, their system was limited in all three factors of localization accuracy, tag density, and absence of an obstacle avoidance mechanism. A direction-sensing RFID system was proposed by Kim and Chong [18] to

demonstrate that localization could be performed in the presence of a large number of obstacles. While a good study, it used active RFID tags that require periodic maintenance, and was not tested in the presence of multiple RFID tags to warrant real-time implementation. While the differential-driving robot proposed by Han et al. [23] had localization accuracy close to the system proposed here, its performance was limited by a very high tag density and a lack of an obstacle avoidance system.

**Table 2. Comparison of IPS with Previous Work**

Method	RFID tag type	Localization accuracy (m)	Tag density (no. tags/m <sup>2</sup> )
[13]	Active	0.62	4.33
[16]	Passive	0.40	2.0
[17]	Passive	0.14	9
[18]	Active	n/a	n/a
[23]	Passive	0.09	350
Proposed	Passive	0.07	3.66

To compare cost effectiveness, consider an inventory warehouse on the order of 100 m<sup>2</sup>. While the number of RFID tags required for this warehouse is large, and with the current price of passive and active tags at \$0.15 and \$25, respectively, a simple calculation will show that the RFID tag implementation cost of the proposed system is very low compared to other systems. Furthermore, when the cost of the mobile robot used in this current system (\$150) is compared with similar robotic systems (\$4000 and above), the cost effectiveness of the proposed indoor positioning system is clearly evident.

## Conclusion

In this study, the author developed an efficient passive-RFID-based indoor positioning method. The aim of this proposed approach was to perform accurate localization based on information read from the passive RFID tags in the vicinity. Unlike other methods found in the literature, the proposed method does not require additional sensors or systems to evaluate the navigating environment. In this current system, the proposed IPS was validated through indoor experiments, including but not limited to the robot's movement accuracy, obstacle detection system, and RFID positioning system. The results of this study demonstrated that the proposed system would be adequate in providing mobility for several applications. Future work of the proposed work will include navigating the system through dynamic environments, reducing the tag density, and increasing the overall system robustness to operate in harsh environments.



---

## Acknowledgment

This research was conducted from Central Michigan University's Early Career Investigator grant number C61984. The authors would like to thank the anonymous reviewers for their feedback in improving the quality of this presentation.

## References

- [1] Yelamarthi, K., Sherbrook, S., Beckwith, J., Williams, M., & Liefief, R. (2012). An RFID based Semi-Autonomous Indoor Tour Guide Robot. *Proceedings of the IEEE International Midwest Symposium on Circuits and Systems*, (pp. 562-565). Boise, ID.
- [2] Hehn, M., Ritz, R., & D'Andrea, R. (2012). Performance Benchmarking of Quadrotor Systems Using Time-Optimal Control. *Autonomous Robots*, 33(1), 69-88.
- [3] Yelamarthi, K. (2012). RFID-Based Interdisciplinary Educational Platform to Improve the Engineering and Technology Curriculums. *Journal of STEM Education: Innovations and Research*, 13(5), 46-51.
- [4] Gueaieb, W., & Miah, M. S. (2008). An Intelligent Mobile Robot Navigation Technique using RFID Technology. *IEEE Trans. Instrumentation and Measurement*, 57(9), 1908-1917.
- [5] Das, A., Thakur, D., Keller, J., Kuthirummal, Kira, Z., & Pivtoraiko, M. (2013). R-MASTIF: Robotic mobile autonomous system for threat interrogation and object fetch. *Proceedings of SPIE 8662, Intelligent Robots and Computer Vision XXX: Algorithms and Techniques*, doi: 10.1117/12.2010720.
- [6] Kumar, V., & Michael, N. (2012). Opportunities and challenges with autonomous micro aerial vehicles. *The International Journal of Robotics Research*, 31(11), 1279-1291.
- [7] Heng, I., Zhang, A. S., & Yap, R. (2012). A Novel Design of a Human-Robot Finger Controller. *International Journal of Modern Engineering*, 13(1), 18-26.
- [8] Thamma, R., Kesireddy, L. M., & Wang, H. (2012). Centralized Vision-Based Controller For Unmanned Guided Vehicle Detection. *International Journal of Modern Engineering*, 12(2), 58-65.
- [9] Bryan, C., Grenwalt, M., & Stienecker, A. W. (2011). Embedded Onboard Control of a Quadrotor Aerial Vehicle. *International Journal of Modern Engineering*, 12(1), 5-10.
- [10] Yelamarthi, K., Haas, D., Nielsen, D., & Mothersell, S. (2010). RFID and GPS Integrated Navigation System for the Visually Impaired. *IEEE International Midwest Symposium on Circuits and Systems*, Seattle, WA.
- [11] Dancer, K., Martin, W., Rock, K., Zeleny, C., & Yelamarthi, K. (2009). The Smart Cane: An Electrical Engineering Design Project. *Proceedings of ASEE North Central Section Conference*, Grand Rapids, MI.
- [12] Sezer, V., & Gokasan, M. (2012). A Novel Obstacle avoidance algorithm: Follow the Gap Method. *Robotics and Autonomous Systems*, 60(9), 1123-1134.
- [13] Chunag, L. (2012). Object localization strategy for a mobile robot using RFID. *Master's Thesis, Department of Computing Science, Umeå University, Umeå, Sweden*.
- [14] Ardiyanto, I., & Miura, J. (2012). Real-time Navigation Using Randomized Kinodynamic Planning with Arrival Time Field. *Robotics and Autonomous Systems*, 60(12), 1579-1591.
- [15] Zhang, Z. (2011). I am the Antenna: Accurate Outdoor AP Location using Smart phones. *Proceedings of ACM MobiCom*, Las Vegas, NV.
- [16] Schneegans, S., Vorst, P., & Zell, A. (2007). Using RFID Snapshots for Mobile Robot Self-Localization. *Proceeding of European Conference on Mobile Robots*, (pp. 241-246). Freiburg, Germany.
- [17] Park, S., & Hashimoto, S. (2009). Autonomous Mobile Robot Navigation Using Passive RFID in Indoor Environment. *IEEE Transaction on Industrial Electronics*, 56(7), 2366-2373.
- [18] Kim, M., & Chong, N. (2009). Direction Sensing RFID Reader for Mobile Robot Navigation. *IEEE Transactions on Automation Science and Engineering*, 6(1), 44-54.
- [19] Amundson, I., Sallai, J., Koutsoukos, X., & Ledeczi, A. (2010). Radio interferometric angle of arrival estimation. *Lecture Notes in Computer Science*, 5970, 1-16.
- [20] iRobot Create Programmable Robot, iRobot. Retrieved Aug 12, 2013, from: <http://www.irobot.com/en/us/learn/Educators/Create>
- [21] SkyeModule M10 Developer Kit, Skyetek. Retrieved Aug 12, 2013, from <http://www.skyetek.com>
- [22] SRF-04 Ultrasonic Range Finder, Robot Electronics. Retrieved Aug 12, 2013, from <http://www.robot-electronics.co.uk>
- [23] Han, S., Kim, D., & Lee, J. (2007). An Efficient Localization Scheme for a Differential-Driving Mobile Robot Based on RFID System. *IEEE Transactions on Industrial Electronics*, 54(6), 3362-3369.
- [24] Yun, S., Lee, J., Chung, W., & Kim, E. (2005). Centroid Localization Method in Wireless Sensor Networks Using TSK Fuzzy Modeling. *Proceedings of International Symposium on Advanced Intelligent Systems*, (971-974).

- 
- [25] Peng, P. (2009). A Cooperative Target Location Algorithm Based on Time Difference of Arrival in Wireless Sensor Networks. *Proceedings of IEEE International Conference on Mechatronics and Automation*, (696-700). Changchun, China.
- [26] An Autonomous Passive RFID-Assisted Mobile Robot System for Indoor Positioning. (n.d.). Retrieved Aug 12, 2013, from [http://people.cst.cmich.edu/yelamlk/rfid\\_positioning\\_robot.mov](http://people.cst.cmich.edu/yelamlk/rfid_positioning_robot.mov)

## Biographies

**KUMAR YELAMARTHI** received his Ph.D. degree from Wright State University in 2008. He is currently an Assistant Professor of Electrical Engineering at Central Michigan University, Mt Pleasant, MI. His research interests include RFID, embedded systems, robotics, integrated circuit optimization, and engineering education. He has served as a technical reviewer for several international conferences and journals, and has published over 75 articles in archival journals and conference proceedings. He is a member of the Tau Beta Pi engineering honor society, the Omicron Delta Kappa national leadership honor society, and a senior member of IEEE. Dr. Yelamarthi may be reached at [k.yelamarthi@ieee.org](mailto:k.yelamarthi@ieee.org).

# IMPLEMENTATION OF ROBOT JOINT TRAJECTORY CONTROL: AN APPLIED METHOD

Shaopeng (Frank) Cheng, Central Michigan University

## Abstract

Accurate point-to-point (PTP) robot motion is the most essential of industrial robot applications. It requires the joint controller to control each robot joint according to the joint motion specification. In this paper, the author discusses the key concepts and technologies related to implementing programmable robot joint trajectory control systems. The developed method affords the integrated solutions on the specification, control, and interface of joint motion variables in the microprocessor-based control systems. This includes robot point-to-point (PTP) motion trajectory, joint dynamics and independent servo control, trajectory planning function and parameters, motion data representation in the microcontroller, proportional-integral-derivative (PID) control, pulse-width-modulation (PWM) control interface, and a DC motor drive circuit. As a result, the control principle, requirements, and relevant relationships are introduced to aid the designer in completing the design quickly without mistakes or confusion. A case study shows the application of the method and the performance of a designed single-joint trajectory control system.

## Introduction

The trajectory of an industrial robot is the time history of the motion of robot tool-center-point (TCP) within the robot's working envelope. It has been shown that if all robot joints start and stop their motions exactly at the same time during a given robot point-to-point (PTP) motion, the robot's TCP will move in an unpredictable, but repeatable, trajectory with a uniform trajectory distance [1]. Because of this trajectory characteristic, it is possible for the robot to move the TCP in a safe, reliable, and collision-free manner. In industrial robot systems, such a trajectory has already been implemented as the basic PTP motion control instruction and used in advanced PTP motions like straight-line trajectory control [2].

Robot trajectory control is a large, complex, and important subject. On the one hand, some studies looked at theoretical developments of the subject and provided analytical solutions to some specific problems such as robot trajectory planning and closed-loop robot control systems [3-7]. However, the applied issues of trajectory control solutions ranging from system specification, data representation,

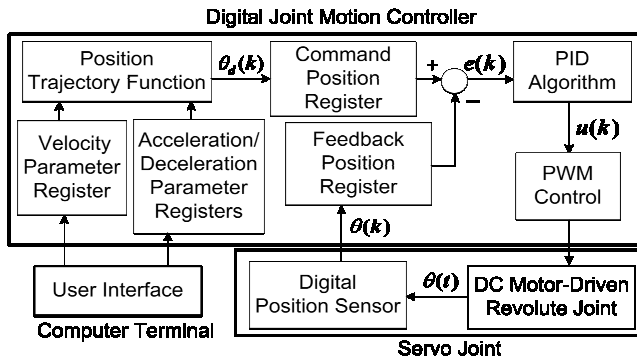
and function implementation in digital microcontrollers have not been fully addressed [8]. On the other hand, commercial industrial robots and motion control products have intensively implemented the robotics and control theories for enhanced product functions, but their documentation usually has very little information about how the relevant theories and technologies are applied in their product designs. Consequently, control and robotics engineers at different levels often find themselves at positions in which they use the functions of the commercial products to do robotic and motion control applications, but with only a limited understanding of the theories and technologies implemented in the design.

In this paper, the author tries to fill in some of the technical gaps in robot trajectory control by presenting an integrated discussion on the specification, control, and interface of joint motion variables in microprocessor-based servo control systems. The objective is to develop an applied method in which the specified motion trajectory of a given joint with unknown joint parameters can be accurately controlled by a closed-loop digital motion control system. It is expected that this method can be used in controlling all joints of a robot arm for basic and advanced PTP motions. To achieve this goal, the author established an applied framework on closely related issues including joint dynamics and independent servo control, joint trajectory planning, motion data representation in a microcontroller, proportional-integral-derivative (PID) control, pulse-width-modulation (PWM) control interface, and DC motor drive circuits. Relevant control principles, requirements, and relationships were logically introduced and precisely formulated to aid the designer in completing the implementation task faster and easier with less confusion. The method was successfully applied in the case study of implementing programmable digital joint servo control systems that are able to control the motion trajectory of DC motor-powered joint systems.

## Joint Dynamics and Independent Servo Control

Controlling all robot joints to follow their desired trajectory functions in real-time is a very complex task. This is because the joint dynamics equation is nonlinear due to the coupled joint inertia, centrifugal force, and gravity of the robot arm [3]. To implement robot trajectory control, a de-

centralized independent joint control scheme was developed and used by partitioning the nonlinear joint dynamics equation into a linear, second-order joint servo system and a nonlinear, model-based computational torque disturbance applied to the servo system. This means that the PTP motion of a robot arm can be controlled by each independent single-input, single-output (SISO) closed-loop joint position control system with the characteristics of high disturbance rejection and accurate trajectory tracking [6], [7]. In this study, this control scheme was implemented as a digital closed-loop joint trajectory control system that controlled a single revolute joint through a trajectory module and a servo control module, as shown in Figure 1.



**Figure 1. Block Diagram of Digital Closed-loop Joint Position Trajectory Control System**

The dynamics of an independent revolute joint can be represented by a linear, second-order differential equation of joint position  $q(t)$  as given in Equation (1):

$$\tau_d = J_{eq} \ddot{\theta}(t) + B_{eq} \dot{\theta}(t) \quad (1)$$

where,  $J_{eq}$  and  $B_{eq}$  are the equivalent inertia and viscous damping of the joint, respectively, and  $t_d$  is the required driving torque transmitted from the motor torque,  $t_m$ , through gear ratio  $N$  [1], [6]. Analytical results showed that joint parameters  $J_{eq}$  and  $B_{eq}$  in the joint dynamic equation naturally produce a large system time constant, which usually causes a long settling time,  $t_s$ , of the joint response to a joint position change and a large position lagging in joint position tracking control [5], [6]. For implementing joint trajectory control in this study, it was assumed that: 1) the joint parameters  $J_{eq}$  and  $B_{eq}$  were unknown, and 2) the DC motor was able to generate the required torque,  $t_d$ , during a joint PTP motion.

As shown in Figure 1, prior to conducting a PTP motion, the trajectory module receives the desired trajectory parameters of a PTP motion from the upper-level robot control system (or a user interface) and stores them in the specified

32-bit registers of the microcontroller. During the PTP execution, the trajectory module uses the chosen position trajectory function to generate the desired joint position value and send it to the servo control module for joint motion control. The servo module provides the joint system with two fundamental abilities: position tracking and disturbance rejection. The good position tracking ability allows the joint to quickly move to and stabilize at a new joint position within a short settling time,  $t_s$ . The strong disturbance rejection ability allows the system to greatly reduce position inaccuracy caused by unexpected system parameter changes and shaft torque disturbances. To implement the digital servo module, the closed-loop control system must sample all analog trajectory and servo information on a fixed, regular interval called sampling time,  $T_s$ . At a particular sampling cycle,  $k$ , the desired joint position,  $q_d(k)$ , generated from the specified joint position trajectory function is compared with the sensor-measured actual joint position,  $q(k)$ , to create the position error,  $e(k)$ . The controller then uses the position error,  $e(k)$ , and the designed PID control algorithm to calculate the required position control command  $u(k)$  and converts it into the corresponding PWM signal for controlling the joint through a DC motor drive circuit.

## Joint Trajectory Planning Functions and Parameters

To coordinate all robot joints for conducting a PTP motion that results in a repeatable robot trajectory, the upper-level robot control system must perform the following steps for determining the common travel time,  $t_f$ , used by all robot joints:

1. Calculate the required travel distance of each joint for the required PTP motion as  $Dq = q_f - q_i$ ;
2. Calculate the required travel time of each joint in the PTP motion by using the allowable joint maximum velocity and the required joint travel distance,  $Dq$ ;
3. Choose the longest joint travel time as the common travel time,  $t_f$ , for all joints in the PTP motion;
4. Calculate the actual joint travel velocity,  $w = (Dq)/t_f$ , for each joint; and,
5. Send the specified values for  $t_f$  and  $w$  to each joint control system.

Based on the required joint travel time,  $t_f$ , and velocity,  $w$ , each joint control system must plan how to physically carry out the joint motion in real-time. In robot trajectory planning, the desired joint position trajectory,  $q_d(t)$ , is generated from the desired velocity trajectory profile,  $\dot{\theta}_d(t)$ . Trapezoid and S-curve velocity profiles are realistic for an actual robot joint to implement because of the finite joint acceleration and deceleration used in these two profiles [1], [3], [4].

To simplify the discussion, the author of this current study used a symmetrical trapezoid velocity profile,  $\dot{\theta}_d(t)$ , as indicated by the blue line in Figure 2. Under this profile, the revolute joint starts a PTP motion from the initial position,  $q_i$ , at zero velocity and accelerates at constant acceleration,  $a$ , until reaching the required joint velocity,  $w$  (i.e., the cruise velocity) at time  $t_b$ . The joint continues to move at cruise velocity,  $w$ , until decelerating at time  $(t_f - t_b)$  with a constant deceleration of  $-a$ . At time  $t_f$ , the motion stops at the destination position,  $q_f$ . The joint position trajectory function,  $q_d(t)$ , in Figure 2 can be mathematically modeled by three blended motion segments, as shown in Equation (2) [1].

$$\left\{ \begin{array}{l} \text{For } 0 \leq t \leq t_b \\ \theta_d(t) = \theta_i + \frac{1}{2}at^2 \\ \text{For } t_b \leq t \leq t_f - t_b \\ \theta_d(t) = \left( \theta_i + \frac{1}{2}\left(\frac{\omega}{t_b}\right)t_b^2 \right) + \omega(t - t_b) \\ \text{For } t_f - t_b \leq t \leq t_f \\ \theta_d(t) = \theta_f - \frac{1}{2}\left(\frac{\omega}{t_b}\right)(t_f - t)^2 \end{array} \right. \quad (2)$$

where,

$$t_b = \frac{\omega}{a} \quad \text{and} \quad t_f = \frac{\omega}{a} + \frac{(\theta_f - \theta_i)}{\omega}$$

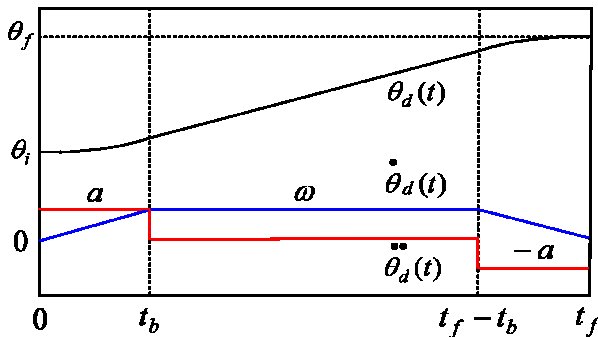


Figure 2. Trapezoid Velocity Trajectory Profile

Since the joint PTP travel time,  $t_f$ , and joint cruise velocity,  $w$ , are usually given by the upper-level robot control system, the joint control system must determine the corresponding acceleration,  $a$  (or deceleration,  $-a$ ). From the relationship  $t_f = \frac{\omega}{a} + \frac{(\theta_f - \theta_i)}{\omega}$  in Equation (2), the acceler-

ation,  $a$  (or deceleration,  $-a$ ) is determined by Equation (3):

$$a = \frac{\omega^2}{\omega t_f - (\theta_f - \theta_i)} \quad (3)$$

## Joint Motion Data Representation in a Microcontroller

The digital microcontroller in the servo module updates all trajectory and servo values at sampling time  $T_s$ . A 32-bit Time Register is available in the microcontroller that holds the current processor time,  $t$ , measured as the total number of sampling cycles,  $k$ , executed since system power-up or reset. The conversion from time  $t$  to sampling cycles of the microcontroller is given by Equation (4):

$$\text{cycles} = \frac{t}{T_s} \quad (4)$$

where,  $t$  and  $T_s$  must be in the same units. Considering that most commercial microcontrollers have sampling time  $T_s$  in microseconds ( $\mu\text{s}$ ), then the number of sampling cycles in one second is given by Equation (5):

$$\text{cycles} = \frac{10^6 (\mu\text{s})}{T_s (\mu\text{s})} \quad (5)$$

During a joint PTP motion, the microcontroller continuously samples the position trajectory function,  $q_d(t)$ , at each sampling time,  $T_s$ . The sampled desired position,  $q_d(k)$ , in degrees at a particular cycle,  $k$ , must be converted to a signed integer,  $c_d(k)$ , in counts and stored in the 32-bit Desired Position Register. The conversion is determined by the sensitivity,  $S_e$ , (in counts/degree) of the digital position sensor used in the joint control system, as given by Equation (6):

$$c(k) = S_e \times \theta(k) \quad (6)$$

For example, if the sampled desired position is  $q_d(k) = 139$  degrees and the digital sensor sensitivity is  $S_e = 11.38$  counts/degree, the corresponding converted value is  $c_d(k) = +1581.82$  counts, which is then saved as the signed integer of  $c_d(k) = +1582$  counts in the Desired Position Register. Similarly, the specified values of cruise velocity  $w$  and acceleration (or deceleration)  $a$  in Equation (2) must be converted into the corresponding unsigned value in counts/cycle and counts/cycle<sup>2</sup>, and stored in the 32-bit Velocity Register and Acceleration (or Deceleration) Register, respectively. The corresponding conversions are given by Equations (7) and (8), respectively:



$$\omega(\text{counts / cycle}) = \frac{\omega(\text{degree/sec}) \times S_e}{\text{cycles}} \quad (7)$$

$$a(\text{counts/cycle}^2) = \frac{a(\text{degree/sec}^2) \times S_e}{\text{cycles}^2} \quad (8)$$

where, the value of *cycles* is defined in Equation (4) or Equation (5). For example, if the sampling time of the microcontroller is  $T_s = 256 \mu\text{s}$ , after the conversion, the trajectory parameters— $w = 120 \text{ degrees/sec}$  and acceleration  $a = 405 \text{ degrees/sec}^2$ —become small floating-point values of 0.3496 counts/cycle and 0.0003 counts/cycle<sup>2</sup>, respectively. If the microcontroller does not support the floating-point unit, the floating-point numbers must be represented as fixed-point numbers in the registers. To do this, the first 16 bits of the Velocity Register and Acceleration (or Deceleration) Register can be used as a scaling factor of  $2^{16} = 65,536$ , which is multiplied by the converted small floating-point values obtained in Equations (7) and (8), as shown in Equations (9) and (10):

$$\omega(\text{counts/cycle})|_{\text{Fixed-Point}} = \omega(\text{counts/cycle}) \times 2^{16} \quad (9)$$

$$a(\text{counts/cycle}^2)|_{\text{Fixed-Point}} = a(\text{counts/cycle}^2) \times 2^{16} \quad (10)$$

For example, with Equations (9) and (10), the velocity  $w = 120 \text{ degrees/sec}$  and acceleration  $a = 405 \text{ degrees/sec}^2$  are finally represented as fixed-point integer values of 22,906 counts/cycle in the Velocity Register and 20 counts/cycle<sup>2</sup> in the Acceleration Register, respectively.

It is important to note that correct conversions of motion data are critical for implementing a digital joint control system. Clearly, the sensitivity,  $S_e$ , of the chosen digital position sensor affects the accuracy of the converted motion data. In general, if a digital rotary position sensor measures the joint position of one revolution by a total of  $n$  digitized counts, the sensitivity,  $S_e$ , of the sensor and the smallest position angle,  $Dq$  (i.e., resolution) to be measured by the sensor are given by Equations (11) and (12):

$$S_e = \frac{n}{360^\circ} \quad (11)$$

$$\Delta\theta = \frac{1}{S_e} \quad (12)$$

## Tuning-up PID Control Function

The PID controller in the servo module is designed to improve the transient and steady responses of the joint-to-joint position changes. It works by calculating the control

command,  $u(t)$ , based on the joint position tracking error,  $e(t)$ , as given by Equation (13):

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau) d\tau + K_d \frac{de(t)}{dt} \quad (13)$$

where, proportional gain  $K_p$  amplifies the error, integral gain  $K_i$  amplifies the accumulation of the errors over time, and derivative gain  $K_d$  amplifies the changing rate of the error. For digital PID implementation, it is necessary to discretize the PID calculation by approximating the integral and derivative terms, as given by Equation (14):

$$\frac{de(t)}{dt} \approx \frac{e(t) - e(t - T_s)}{T_s} \quad \text{and} \quad \int_0^t e(t) dt \approx T_s \sum_{n=0}^k e(nT_s) \quad (14)$$

where,  $T_s$  is the sampling time of the microcontroller and  $k$  is the number of sampling cycles defined in Equation (4). If  $T_s$  is small enough, this approximation is accurate. Thus, the PID calculation becomes:

$$u(kT_s) = K_p e(kT_s) + K_i T_s \sum_{n=0}^k e(nT_s) + K_d \left( \frac{e(kT_s) - e((k-1)T_s)}{T_s} \right) \quad (15)$$

By referencing the iterations (i.e., cycle numbers,  $k$ ) instead of time  $t = kT_s$ , the PID algorithm can be expressed as shown in Equation (16):

$$u(k) = K_p e(k) + K_i T_s \sum_{n=0}^k e(n) + K_d \left( \frac{e(k) - e(k-1)}{T_s} \right) \quad (16)$$

where,  $e(k)$ ,  $e(k-1)$ , ... are samples of position error,  $e(t)$ , at the  $k^{\text{th}}$ ,  $(k-1)^{\text{th}}$ , ... sampling cycles, respectively. The relationships of PID gains between PID algorithm and PID calculation are  $K'_p = K_p$ ,  $K'_i = K_i T_s$  and  $K'_d = K_d/T_s$ , which show how sampling time,  $T_s$ , affects the actual PID gain values used in the PID algorithm. Similarly, the value of  $u(k)$  must be converted to counts and saved in the 16-bit Motor Commander Register with the sign bit. Most commercial motion controllers have the built-in PID algorithm. In this case, the designer only needs to manually tune up three PID gains for a fast and stable joint position step-response without using the mathematical model of the joint system. However, this task can be complicated as a change in one gain can shift the optimal settings of the others.

In general, a large proportional gain,  $K_p$ , gives the joint control system a quicker response and a stronger ability to reject any disturbance appearing in the system. However, if  $K_p$  is too large, it can cause large system overshoot and even instability. The derivative gain,  $K_d$ , acts like a damper to reduce system overshoot. However, if  $K_d$  is too large, the system response may become slow. The integral gain,  $K_i$ , is

used to eliminate the steady-state error and should be as small as possible. This is because the integral term is calculated from the sum of the present and all previous position errors. A small  $K_i$  can still make the integral contribution large enough to eliminate the final error of the system in a long historic windup. If  $K_i$  is too large, the integral contribution to the control output,  $u(k)$ , can be too much within a short period of time, causing the system to overshoot or even become unstable. Commercial motion controllers generally provide a settable integral limit term to cap the total contribution of the integral term so that an unruly amount of power will not be delivered if the joint is to jam or simply find itself against a travel limit.

The best way to complete the tuning task is to change one gain at a time while observing the response of the joint to a commanded position through the use of a position trace facility such as a scope. The procedure is as follows:

1. Initialize  $K_i$  and  $K_p$  to zero, and  $K_d$  to a small value.
2. Start to increase  $K_p$  until the joint position shows an underdamped step-response.
3. Increase  $K_d$  until the joint position is critically damped.
4. Repeat steps 2 and 3 until the system becomes slightly noisy (i.e., the grinding sound). Back off the  $K_p$  and  $K_d$  values for a final, reasonably quiet system.
5. Increase integral term  $K_i$  from zero to a small, non-zero value to eliminate the final position error.
6. Set the integral limit to be a large value to prevent the integral contribution to  $u(t)$  from being saturated.

## The Principle of the PWM Control Mode

In the servo module, the digitized PID control command,  $u(k)$ , needs to control the analog DC motor of the joint. One commonly used interface technology is pulse-width-modulation (PWM). The magnitude, or width, of a PWM pulse signal is defined by its duty cycle,  $t$ , within a PWM period,  $T$ , as  $t/T$ , ranging from 0 to 100%, as shown in Figure 3. Under PWM control mode, the microcontroller converts the PID control command,  $u(k)$ , into signed PWM pulses and sends them out from the PWM sign pin and magnitude pin.

The method of converting the absolute value of  $|u(k)|$  into PWM magnitude is also shown in Figure 3. First,  $|u(k)|$  is divided by the maximum counts,  $A$ , of the 16-bit PWM Command Register without a sign-bit to define the PWM duty cycle, as shown in Equation (17):

$$PWM(\%) = \frac{|u(k)|}{A} \times 100\% = \frac{t}{T} \times 100\% \quad (17)$$

where,  $A = 2^{15} = 32,768$ . Then, the converted PWM duty cycle is digitally scaled by the maximum counts,  $C$ , of the 11-bit Resolution Register, as given by Equation (18):

$$PWM = \frac{|u(k)|}{A} \times C \quad (18)$$

where,  $C = 2^{11} = 2,048$ . In this way, the  $u(k)$ -controlled PWM duty cycle finally becomes a number of “high” signal counts on the PWM magnitude pin within each PWM period,  $T$ . For example, if  $u(k) = +12,345$  at sampling cycle  $k$ , the converted PWM magnitude will be  $(12,345/32,768) \times 2,048 = 771.56 = 772$  counts. This means that the PWM magnitude pin will be “high” for 772 counts and “low” for the remaining 1276 counts within each PWM period,  $T$ . Obviously, the PWM period,  $T$ , should be much shorter than the sampling time,  $T_s$ , of the microcontroller in order for the PWM magnitude to precisely represent the sampled PID output,  $u(k)$ , in the servo module.

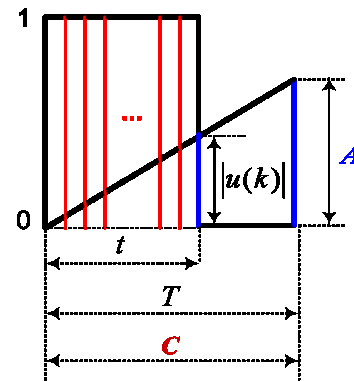


Figure 3. PWM Duty Cycle and Magnitude Generation

## DC Motor Driver Circuit

The sign-magnitude PWM signal generated by the microcontroller must be applied to an H-bridge driver circuit, as shown in Figure 4, in order to control the direction and speed of the joint DC motor. In the case where a “high” signal on the PWM sign pin defines the “forward” joint rotation, then the logic circuit in Figure 4a sets the MOSFET switches SW4 and SW2 constantly closed and open respectively; and, the PWM magnitude, or duty cycle, controls the connection of the DC motor to the source,  $V_s$ , by closing and opening MOSFET switch SW1 with the modulated PWM pulses. The logic circuit also works for the case in which the “low” signal on the PWM sign pin defines the “reverse” joint rotation. In both cases, the ON/OFF switching of  $V_s$  to the DC motor will have little effect on the motor rotating performance as the PWM period,  $T$ , is very small, resulting in the PWM frequency ( $1/T$ ) being significantly

larger than the natural frequency of the motor rotor. This means that the motor does not respond to the ON/OFF cycle of  $V_s$ , except when being driven by an average voltage,  $V_{avg}$ , which is controlled by the PWM magnitude as  $V_{Avg} = V_s \cdot \text{PWM} (\%)$ . For example, if  $V_s = 12 \text{ V}$  and the PWM magnitude is 30%, then the actual average voltage that drives the DC motor is  $12\text{V} \times 0.3 = 3.6 \text{ V}$ .

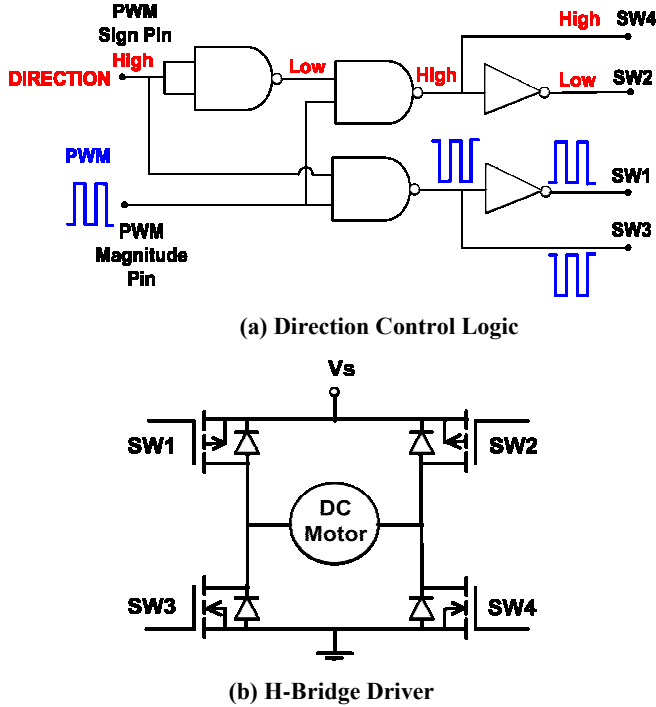


Figure 4. DC Motor Direction Control and Drive Circuit

## Case Study

The method developed in this study was used to implement a joint trajectory control system, as shown in Figure 5.

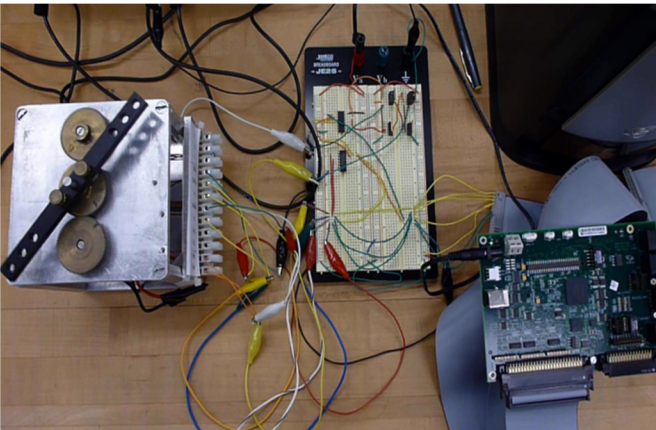


Figure 5. An Implemented Digital Joint Motion Control System

In the system, the motion controller from Performance Motion Devices (PMD), Inc. controls a DC-motor-powered revolute joint to perform the required PTP motion following a user-specified symmetric trapezoid velocity profile (see Figure 2). Although the trajectory module was implemented with the trajectory algorithm provided by the PMD motion controller, the user must program the trajectory module with user-specified trajectory parameters. In this case study, the performance of the system was tested by moving the joint from an initial position of  $q_i = 0$  degrees to a final position of  $q_f = 90$  degrees, with a required travel time of  $t_f = 1.0463$  seconds and a cruise velocity of  $w = 120$  degrees/s. From Equation (3), the corresponding acceleration (or deceleration),  $a$ , and blending time,  $t_b$ , in the position trajectory function were determined to be:

$$a = \frac{\omega^2}{\omega t_f - (\theta_f - \theta_i)} = \frac{120^2}{(120)(1.0463) - (90)} = 405 \text{ deg./s}^2$$

$$t_b = \frac{\omega}{a} = \frac{120}{405} = 0.296 \text{ sec.}$$

Thus, the desired position trajectory function,  $q_d(t)$ , used for the required joint PTP motion, was established with Equation (2) to be:

$$\left. \begin{aligned} &\text{For } 0 \leq t \leq 0.296 \text{ sec.} \\ &\theta_d(t) = \theta_i + \frac{1}{2}at^2 = 202.5t^2 \\ &\text{For } 0.296 \text{ sec.} \leq t \leq 0.75 \text{ sec.} \\ &\theta_d(t) = \left( \theta_i + \frac{1}{2} \left( \frac{\omega}{t_b} \right) t_b^2 \right) + \omega(t - t_b) = 120t - 17.76 \\ &\text{For } 0.75 \text{ sec.} \leq t \leq 1.0463 \text{ sec.} \\ &\theta_d(t) = \theta_f - \frac{1}{2} \left( \frac{\omega}{t_b} \right) (t_f - t)^2 = 90 - 202.5(1.0463 - t)^2 \end{aligned} \right\}$$

Three sample joint position values were directly calculated from the desired joint trajectory position function,  $q_d(t)$ :

$$\theta_d(t = 0.154) = 202.5 \times (0.154)^2 = 4.802^\circ$$

$$\theta_d(t = 0.307) = 120 \times (0.307) - 17.76 = 19.08^\circ$$

$$\theta_d(t = 0.845) = 90 - 202.5(1.0463 - 0.845)^2 = 81.79^\circ$$

In the servo module, a quadrature incremental encoder was attached to the joint shaft. For one revolution of joint rotation, the sensor generated a total of 1024 square-wave pulse signals from channel  $A$  and channel  $B$ , respectively. Because the two signals were offset by 90 degrees, one signal (channel  $A$  or  $B$ ) will always lead by 90 degrees, which

makes it possible for the PMD motion controller to detect the joint rotation direction. Also, the total count of the leading and trailing edges of the two 90-degree offset square pulses per joint revolution was  $n = 4 \times 1024 = 4096/\text{rev}$ . Thus, from Equations (11) and (12), the sensitivity,  $S_e$ , and the smallest position resolution,  $\Delta\theta$ , of the quadrature encoder in the servo module were:

$$S_e = \frac{n}{360^\circ} = \frac{4096}{360} = 11.38 \text{ counts/degrees}$$

$$\Delta\theta = \frac{1}{S_e} = \frac{360}{4096} = 0.0879 \text{ degrees/count}$$

The sampling time of the PMD motion controller was  $T_s = 256 \mu\text{s}$ . From Equation (5), one second is equal to 3906.25 cycles in the PMD motion controller. From Equations (6)-(10), the required joint trajectory parameters of the initial position ( $q_i = 0^\circ$ ), final position ( $q_f = 90^\circ$ ), velocity ( $w = 120^\circ/\text{s}$ ), and acceleration or deceleration ( $a = 405^\circ/\text{s}^2$ ) were converted into the corresponding register values of the PMD motion controller, as shown in Table 1.

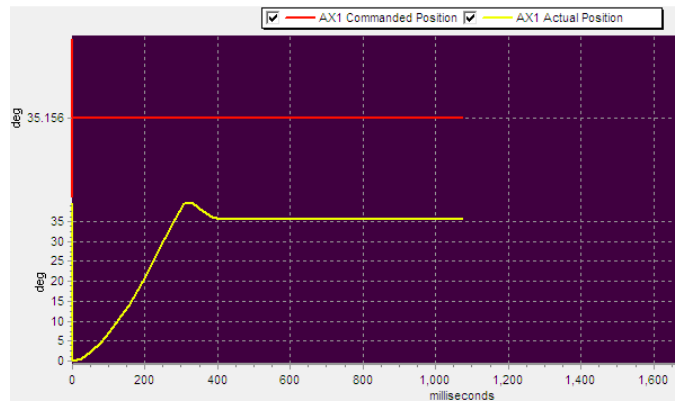
**Table 1. Joint Trajectory Parameters and Conversions**

Trajectory Parameters	Parameter Values	Converted Parameter Values
Initial Position $q_i$	0 degrees	0 count
Final Position $q_f$	+90 degrees	+1024 counts
Cruise Velocity $w$	120 deg./sec.	22906 counts/cycle
Acceleration $a$	405 deg./sec. <sup>2</sup>	20 counts/cycle <sup>2</sup>
Deceleration $a$	405 deg./sec. <sup>2</sup>	20 counts/cycle <sup>2</sup>

The PID control function in the servo module was implemented by the PID algorithm of the PMD motion controller. Using the “Scope” facility provided by the PMD motion controller, the gains of the PID controller were manually tuned for a fast and stable step-response, as shown in Figure 6.

As indicate by Figure 6, the joint was able to reach and stabilize to an instantaneous position change of 35.1 degrees with a settling time of  $t_s = 400 \text{ ms}$ . The tuned PID gains are listed in Table 2. The large  $K'_p$  of 3500 provided the system with a good ability to generate a fast response and reject disturbance, while the  $K'_d$  of 500 reduced the system overshoot. The integral gain  $K'_i$  of 200 was the result of  $K'_i = K_i$

$T_s = 0.78 \times 256$ ; thus, the actual integral gain of  $K_i = 0.78$  used in the PID calculation was small, which is consistent with the PID tune-up method discussed earlier. The integral limit of 40,000 set for the integral term was large enough for  $K'_i$  to generate an integral contribution in order to eliminate the final joint position error.



**Figure 6. Step-Response of Joint Servo Position System**

**Table 2. Tuned PID Gains**

PID Gains	Values
$K'_p$	3500
$K'_i$	200
$K'_d$	500
Integral Limit	40000

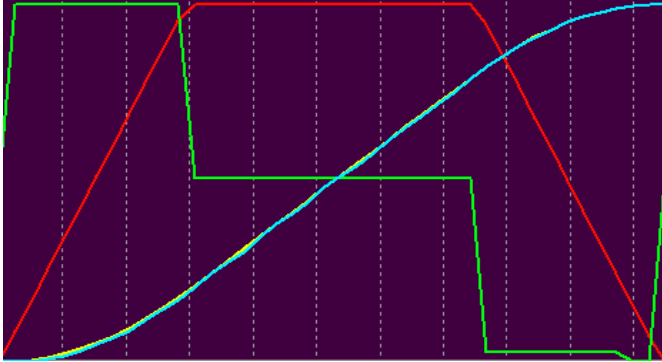
The specified trajectory parameters from Table 1 and PID gains from Table 2 were sent to the PMD motion controller through the user-developed PTP motion command file in Table 3.

**Table 3. Joint PTP Motion Command File**

SetProfileMode 0 ; select trapezoid trajectory mode
SetPosition 1024 ;input destination position (90 deg.)
SetVelocity 22906 ;input cruise velocity (120 deg./sec.)
SetAcceleration 20 ;input acceleration (405 deg./sec <sup>2</sup> )
SetDeceleration 20 ;input deceleration (405 deg./sec <sup>2</sup> )
SetPositionLoop 0, 0, 3500 ;input prop. gain $K_p$ (3500)
SetPositionLoop 0, 1, 200 ;input integral gain $K_i$ (200)
SetPositionLoop 0, 2, 500 ;input derivative gain $K_d$ (500)
Update ;Execute command file

During the PTP motion, the PID control command,  $u(t)$ , was automatically converted into the PWM sign and magnitude that control the DC motor through the H-Bridge driver circuit (see Figure 4). The actual PWM period of  $T = 0.05$

$\mu\text{s}$  (i.e., 20 MHz) was much smaller than the sampling time of  $T_s=256 \mu\text{s}$  for the motion controller, which guarantees the accuracy of the PWM conversion. The desired and actual joint position trajectories were traced by the “Scope” facility of the PMD motion controller, as shown in Figure 7.



**Figure 7. Joint Trajectory Tracking Performance**

In Figure 7, the yellow, red, and green lines are the desired trajectory functions of position,  $q_d(t)$ , velocity,  $\dot{\theta}_d(t)$ , acceleration and deceleration,  $\ddot{\theta}_d(t)$ , for the required PTP joint motion, respectively; the blue line is the actual joint position trajectory,  $\theta(t)$ , measured by the encoder. Table 4 shows the corresponding traced joint position values.

**Table 4. Traced Joint Trajectory Motion Data**

Trace Sample Time (sec.) $t$	Commanded Position (deg.) $\Theta_d(t)$	Actual Position (deg.) $\Theta(t)$
0.077	1.1425	0.70313
0.154	4.74609	5.00977
0.231	10.81055	10.2832
0.307	19.16016	18.89648
0.384	28.38867	28.38867
0.461	37.61719	37.61719
0.538	46.8457	46.75781
0.614	56.07422	56.07422
0.691	65.30273	65.12695
0.768	74.44336	74.35547
0.845	81.91406	81.91406
0.922	86.92383	86.83594
0.998	89.56055	89.47266
1.075	90	89.91211

First, comparing the values of  $q_d(t)$  in Table 4—with three sample position values of  $q_d(0.154)=4.802^\circ$ ,  $q_d(0.307)=19.08^\circ$ , and  $q_d(0.845)=81.79^\circ$  calculated from the user-specified joint position trajectory function modeled by Equation (2)—it can be seen that the user-programmed trajectory parameters in the register values of the PMD motion controller and the PMD trapezoid trajectory algorithm successfully generated the commanded joint position value,  $q_d(t)$ , in the digital joint control system. Second, the actual joint position,  $\theta(t)$ , measured by the joint encoder was almost the same as the commanded joint position value,  $q_d(t)$ , in Table 4.

This means that the joint moves by exactly following the commanded position trajectory function,  $q_d(t)$ , during the joint PTP motion, as shown in Figure 7. As a result, the actual joint travel time of  $t_f = 1.075$  seconds in Table 4 is nearly the same as the user-specified joint travel time of  $t_f = 1.046$  seconds. Finally, the actual joint position error of 0.088 degrees at the destination is caused by the smallest angle,  $\Delta q = 0.088$  degrees/count, that the quadrature encoder is able to measure in the joint control system.

## Conclusion

In this study, the author addressed the basic problems and solutions related to the implementation of digital joint trajectory control systems that are able to control DC motor-driven robot joints with unknown joint parameters. The results showed that the chosen joint trajectory profile was able to provide the joint controller with the desired position values for the required PTP motion control, and the digital closed-loop joint position control system allowed the joint to follow the desired position trajectory function with the required accuracy. It also showed that the successful implementation of the independent joint trajectory control relies on the correct design of the desired joint trajectory functions, PID control function, and DC motor drive in the digital closed-loop system.

The motion data must be accurately represented in digital motion controllers, which involve sampling time  $T_s$ , digital sensor sensitivity  $S_e$ , and data scaling for the registers of the microcontroller. The sensitivity of the digital sensor determines the accuracy of the data conversion. The PID gains can be manually tuned in the closed-loop joint trajectory control system without the use of a mathematical model of the joint. The sign and magnitude of the PWM signal can be generated by the microcontroller to precisely represent the PID control command and control the DC motor through an H-bridge drive circuit. Finally, the generality and scalability of the solutions formulated in this study can be applied to multiple-joint robot control systems.



---

## References

- [1] Niku, S. B. (2011). *Introduction to Robotics: System, Control, and Application*. (2<sup>nd</sup> ed.). John Wiley & Sons.
- [2] FANUC R-30iA LR Mate Robot Handling Tool Software and Operator's Manual (2007). FANUC Robotics, Inc.
- [3] Nguyen, K. D., Ng, T., & Chen, I. (2008). On Algorithms for Planning S-curve Motion Profiles. *International Journal of Advanced Robotic Systems*, 5(1), 42-52.
- [4] Won, J. H., Choi, B. W., & Chung, M. J. (1991). Smooth Joint Trajectory Planning for a PTP Task. *Proceedings IROS '91. IEEE/RSJ International Workshop on Intelligent Robots and Systems*.
- [5] Sciavicco, L., & Siciliano, B. (2000). *Modeling and Control of Robot Manipulator*. Springer.
- [6] Kelly, R., Santibanez, V., & Loria, A. (2005). *Control of Robot Manipulator in Joint Space*. Springer.
- [7] Ohri, J., Dewan, L., & Soni, M. K. (2007). Tracking Control of Robots Using Decentralized Robust PID Control For Friction And Uncertainty Compensation. *Proceedings of the World Congress on Engineering and Computer Science 2007*, San Francisco, USA.
- [8] Hsia, T., Lasky, T. A., & Guo, Z. (1991). Robust independent joint controller design for industrial robot manipulators. *IEEE Transactions on Industrial Electronics*, 38(1), 21-25.

## Biography

**SHAOPENG (FRANK) CHENG** is an associate professor at Central Michigan University. He has taught various courses on robotics, automatic control systems, applied circuits, instrumentation, and fluid power for engineering and technology programs. His research interests include intelligent industrial robotics, mechatronics, integrated systems, and process controls. He is a certified instructor of FANUC Robotics, Inc. and a member of IEEE. Dr. Cheng can be reached at [cheng1fs@cmich.edu](mailto:cheng1fs@cmich.edu).

# ADAPTIVE LEARNING FACTOR OF BACKPROPAGATION IN FEEDFORWARD NEURAL NETWORKS

Hong Li, The City University of New York

## Abstract

In this paper, the author presents a stability analysis of Feed Forward Multilayer Perceptron (FMP) by introducing a Lyapunov function defined as an error function of the backpropagation algorithm. An adaptive learning factor was identified at an iteration of the learning process. A range of possible learning factors was derived, while the Lyapunov condition was satisfied in order to guarantee the convergence. Performance of such network learning with adaptive learning factors is presented here to demonstrate how the adaptive learning factor can enhance the performance of training, while avoiding the oscillation phenomenon.

## Introduction

Artificial Neural Networks (ANNs) have been proven in many real-world applications to be useful in various tasks of modeling nonlinear systems such as signal processing, pattern recognition, optimization, and weather forecasting, to name a few. ANN is a set of processing elements (neurons or perceptrons) with a specific topology of weighted interconnections between these elements and a learning law for updating the weights of interconnections between two neurons. To respond to the increased demand of system identification and forecasting with a large set of data, many different ANN structures and learning rules, supervised or unsupervised, have been proposed to meet the various needs such as robustness and stability. The FMP networks have been shown to obtain successful results in system identification and control [1], [2].

The Lyapunov function approach was used to obtain a stability analysis of the backpropagation training algorithm of such a network [3-9]. However, the training process can be very sensitive to initial conditions such as the number of neurons, the number of layers, the value of weights, and learning factors which are often chosen by trial and error. In this paper, the author also presents a detailed analysis of the FMP architecture and its stability. The backpropagation algorithm was used for learning (i.e., weight adjustment). In this study, the Least Square error function was defined and verified that it can satisfy the Lyapunov condition in order to guarantee the stability of the system. The analysis carried out a method for defining a range of learning factor values at each iteration, which would ensure that the condition for

stability was satisfied. In the simulation, instead of selecting a learning factor by trial and error, the author defined an adaptive learning factor to satisfy the convergence condition and adjust the connection weight accordingly. The simulation results are presented here to demonstrate the performance.

## Basic Principles of an FMP Network

A system identification problem can be outlined as follows: A set of data is collected from the system including input data and corresponding output data observed, or measured as the target output of the identification problem. The set is often called a “training set”. A neural network model with parameters, called weights, is designed to simulate the system. When the output from the neural network is calculated, an error representing the difference between the target output and the calculated output from the system is generated. The learning process of the neural network is to modify the network and minimize the error.

Consider a system with  $n$  inputs  $X = \{X_1, \dots, X_n\}$  and  $M$  output units  $Y = \{Y_1, \dots, Y_M\}$ . A recurrent FMP network combines a number of neurons, called nodes, which are fed forward to the next layer of nodes (see Figure 1). Suppose  $N_j$  is the number of nodes in the  $j^{\text{th}}$  layer; each output from the  $(j-1)^{\text{th}}$  layer will be used as an input for the next layer. A system of a single layer with  $M$  outputs can be expressed as Equation (1):

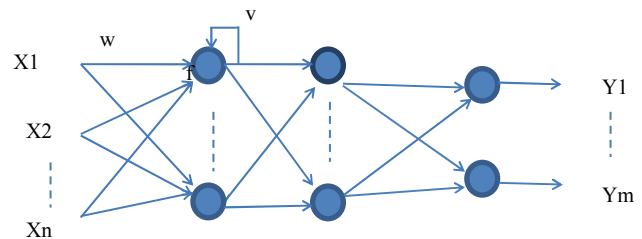


Figure 1. Feed-Forward Multilayer Network

$$Y_{jp} = f(Z) = f\left(\sum_{i=1}^N w_{ij} X_{ip} + \sum_{i=1}^D v_{ij} Y_{i(p-1)}\right) \quad (1)$$

where,  $w_{ij}$  is called the connection weight from input  $X_i$  to output  $Y_j$ ;  $v_{ij}$  is called the connection weight of local feed-

back at the  $j^{\text{th}}$  node with  $i^{\text{th}}$  delay; and,  $f()$  is a nonlinear sigmoid function—see Equation (2):

$$f(z) = \frac{1 - e^{-\theta z}}{1 + e^{-\theta z}} \quad (2)$$

where, the constant coefficient,  $\theta$ , is the slope;  $p = 1, \dots, T$ ;  $T$  is number of patterns;  $N$  and  $M$  are the number of inputs and outputs, respectively; and,  $D$  is the number of delays used in the local feedback.

The backpropagation algorithm has become the standard algorithm for training feed-forward multilayer perceptrons. It is a generalized Least Mean Square algorithm that minimizes the mean squared error between the target output and the network output with respect to the weights. The algorithm looks for the minimum of the error function in the weight space using the method of gradient descent. The combination of weights, which minimizes the error function, is considered to be a solution of the learning problem. A proof of the backpropagation algorithm, presented by Rojas [10], is based on a graphical approach in which the algorithm reduces to a graph labeling problem.

The total error,  $E$ , of the network training set is defined by Equation (30):

$$E = \frac{1}{T} \sum_{k=1}^{N_L} \sum_{p=1}^T e_k^2(p) \quad (3)$$

where,  $e_k^2(p)$  is the error associated with the  $p^{\text{th}}$  pattern at the  $k^{\text{th}}$  node of the output layer—see Equation (4):

$$e_k^2(p) = (d_k(p) - Y_k^L(p))^2 \quad (4)$$

where,  $d_k(p)$  is the target at the  $k^{\text{th}}$  node and  $Y_k^L(p)$  is the output of network at the  $k^{\text{th}}$  node. The learning rule was chosen following the gradient descent method in order to update the network connection weights iteratively—see Equations (5) and (6):

$$\Delta W_j = -\mu \frac{\partial E}{\partial W_j}; j = 1, \dots, M \quad (5)$$

$$\Delta v_j = -\mu \frac{\partial E}{\partial v_j}; j = 1, \dots, D \quad (6)$$

where,  $W_j = (w_{1j}, \dots, w_{Nj})$  and  $v_j = (v_{j1}, \dots, v_{Dj})$  are weight vectors in the  $j^{\text{th}}$  node, and  $\mu$  is a constant called the learning factor.

## Stability Analysis

Stability for nonlinear systems refers to the stability of a particular solution. There may be one solution which is stable and another which is not stable. There are no inclusive general concepts of stability for nonlinear systems. The behavior of a system may depend drastically on the inputs and the disturbances. However, the Lyapunov theory has been used in many studies to examine the stability of nonlinear systems.

The definition of the Lyapunov function and Lyapunov theorem are as follows [11]:

Definition 1 (Lyapunov function): A scalar function  $V(x)$  is a Lyapunov function for the system—see Equation (7):

$$x(t+1) = f(x(t)), f(0) = 0 \quad (7)$$

if the following conditions hold:

1.  $V(0) = 0$  and  $V(0)$  is continuous in  $x$ .
2.  $V(x)$  is positive definite; that is,  $V(x) \geq 0$  with  $V(x) = 0$  only if  $x = 0$ .
3.  $\Delta V(x) = V(f(x(t))) - V(x(t))$  is negative definite; that is,  $V(f(x(t))) - V(x(t)) \leq 0$  with  $\Delta V(x) = 0$  only if  $x = 0$ .

Theorem 1 (Lyapunov Theorem): The solution  $x(t) = 0$  for the system given in Equation (7) is asymptotically stable if there is a Lyapunov function in  $x$ . The stability of the learning process in an identification approach leads to better modeling and a convergent process. According to the Lyapunov theorem, the determination of stability depends on the selection and verification of a positive definite function. For the systems defined by Equations (1) and (2), assume that the backpropagation learning rule is applied and that the error function and weights updating the rule are defined by Equations (5) and (6); the definition is then given by Equation (8):

$$V(t) = \frac{1}{N_L T} \sum_{j=1}^{N_L} \sum_{p=1}^T e_k^2(t) \quad (8)$$

The proof is given in the following theorem that  $V(t)$  satisfies the Lyapunov condition.

Theorem 2: Assuming that the nonlinear sigmoid function  $f()$  defined in Equation (2) is continuous and differentiable, and that the network is defined by Equations (1) and (2) with learning rules from Equations (5) and (6), then the system is stable under the conditions defined by Equation (9):

$$\mu < \frac{TM}{\sum_{j=1}^M \sum_{p=1}^T \left( \left\| \frac{\partial Y_p}{\partial W_j} \right\|^2 + \left\| \frac{\partial Y_p}{\partial v_j} \right\|^2 \right)} \quad (9)$$

Proof: Suppose that  $e_p$  is defined to be the same as in Theorem 1, and the definition function is given by Equation (10):

$$V(k) = \sum_{j=1}^M \sum_{p=1}^T e_{pj}^2 \quad (10)$$

Verification of the function  $V$  as a Lyapunov function for the system, as defined by Equations (1) and (2), is given by Equation (11):

$$\begin{aligned} \Delta V(k) &= V(k+1) - V(k) \\ &= \sum_{j=1}^M \sum_{p=1}^T [e_{pj}^2(k+1) - e_{pj}^2(k)] \\ &= \sum_{j=1}^M \sum_{p=1}^T [(e_{pj}(k) + \Delta e_{pj}(k))^2 - e_{pj}^2(k)] \\ &= \sum_{j=1}^M \sum_{p=1}^T [2e_{pj}(k)\Delta e_{pj}(k) + (\Delta e_{pj}(k))^2] \end{aligned} \quad (11)$$

Applying the first-order Taylor expansion of  $e_{pj}(k+1)$  with respect to weight vectors  $W_j$  and  $v_j$ , yields Equation (12):

$$\begin{aligned} \Delta e_{pj} &= e_{pj}(k+1) - e_{pj}(k) \\ &= \frac{\partial e_{pj}}{\partial W_j} \cdot \Delta W_j + \frac{\partial e_{pj}}{\partial v_j} \cdot \Delta v_j \end{aligned} \quad (12)$$

Substituting Equations (5) and (6) into Equation (12), the calculation yields Equation (13):

$$\begin{aligned} \Delta e_{pj} &= \frac{\partial e_{pj}}{\partial W_j} \cdot \left( -\mu \frac{\partial E}{\partial W_j} \right) + \frac{\partial e_{pj}}{\partial v_j} \cdot \left( -\mu \frac{\partial E}{\partial v_j} \right) \\ &= \frac{\partial e_{pj}}{\partial W_j} \cdot \left( -\frac{\mu}{MT} \sum_{s=1}^T e_{sj} \frac{\partial e_{sj}}{\partial W_j} \right) + \frac{\partial e_{pj}}{\partial v_j} \cdot \left( -\frac{\mu}{MT} \sum_{s=1}^T e_{sj} \frac{\partial e_{sj}}{\partial v_j} \right) \\ \Delta e_{pj} &= -\frac{\mu}{MT} \left[ \frac{\partial e_{pj}}{\partial W_j} \cdot \sum_{s=1}^T e_{sj} \frac{\partial e_{sj}}{\partial W_j} + \frac{\partial e_{pj}}{\partial v_j} \cdot \sum_{s=1}^T e_{sj} \frac{\partial e_{sj}}{\partial v_j} \right] \end{aligned} \quad (13)$$

Substituting Equation (13) into Equation (11) as follows:

$$\begin{aligned} \Delta V(k) &= \frac{-2\mu}{MT} \sum_{j=1}^M \sum_{p=1}^T e_{pj}(k) \left[ \frac{\partial e_{pj}}{\partial W_j} \cdot \sum_{s=1}^T e_{sj} \frac{\partial e_{sj}}{\partial W_j} + \frac{\partial e_{pj}}{\partial v_j} \cdot \sum_{s=1}^T e_{sj} \frac{\partial e_{sj}}{\partial v_j} \right] \\ &+ \frac{\mu^2}{(MT)^2} \sum_{j=1}^M \sum_{p=1}^T \left[ \frac{\partial e_{pj}}{\partial W_j} \cdot \sum_{s=1}^T e_{sj} \frac{\partial e_{sj}}{\partial W_j} + \frac{\partial e_{pj}}{\partial v_j} \cdot \sum_{s=1}^T e_{sj} \frac{\partial e_{sj}}{\partial v_j} \right]^2 \end{aligned}$$

$$\begin{aligned} &\leq \frac{-2\mu}{MT} \sum_{j=1}^M \left[ \left| \sum_{p=1}^T e_{pj} \frac{\partial e_{pj}}{\partial W_j} \right|^2 + \left| \sum_{p=1}^T e_{pj} \frac{\partial e_{pj}}{\partial v_j} \right|^2 \right] \\ &+ \frac{2\mu^2}{(MT)^2} \sum_{j=1}^M \sum_{p=1}^T \left[ \left\| \frac{\partial e_{pj}}{\partial W_j} \right\|^2 \left| \sum_{s=1}^T e_{sj} \frac{\partial e_{sj}}{\partial W_j} \right|^2 + \left\| \frac{\partial e_{pj}}{\partial v_j} \right\|^2 \left| \sum_{s=1}^T e_{sj} \frac{\partial e_{sj}}{\partial v_j} \right|^2 \right] \end{aligned}$$

and using the following notation for simplicity,

$$\alpha = \sum_{j=1}^M \left[ \left| \sum_{s=1}^T e_{sj} \frac{\partial e_{sj}}{\partial W_j} \right|^2 + \left| \sum_{s=1}^T e_{sj} \frac{\partial e_{sj}}{\partial v_j} \right|^2 \right]$$

yields Equation (14):

$$\Delta V(k) \leq \frac{-2\mu}{MT} \alpha + \frac{2\mu^2}{(MT)^2} \alpha \sum_{j=1}^M \sum_{p=1}^T \left[ \left\| \frac{\partial e_{pj}}{\partial W_j} \right\|^2 + \left\| \frac{\partial e_{pj}}{\partial v_j} \right\|^2 \right] \quad (14)$$

To satisfy the Lyapunov condition, let the right-hand side of Equation (14) be less than zero:

$$\frac{-2\mu}{MT} \alpha + \frac{2\mu^2}{(MT)^2} \alpha \sum_{j=1}^M \sum_{p=1}^T \left[ \left\| \frac{\partial e_{pj}}{\partial W_j} \right\|^2 + \left\| \frac{\partial e_{pj}}{\partial v_j} \right\|^2 \right] < 0$$

Further calculations show that the system will be stable when the learning rate satisfies the conditions of Equation (15):

$$\mu < \frac{MT}{\sum_{j=1}^M \sum_{p=1}^T \left[ \left\| \frac{\partial Y_p}{\partial W_j} \right\|^2 + \left\| \frac{\partial Y_p}{\partial v_j} \right\|^2 \right]} \quad (15)$$

For a neural network system with one hidden later, the stability can be similarly proved.

Theorem 3. Assume that a system with one hidden layer can be represented in the form given by Equations (16) and (17):

$$Y_{ip} = f(Z_{ip}^1) = f\left(\sum_{i=1}^N w_{ij}^o X_{ip}^1 + \sum_{i=1}^D v_i^o Y_{j(p-i)}\right) \quad (16)$$

$$X_{ip}^1 = f(Z_{ip}) = f\left(\sum_{i=1}^N w_{ij}^h X_{ip} + \sum_{i=1}^D v_i^h X_{j(p-i)}\right) \quad (17)$$

then the gradient descent rule can be expressed by Equations (18)-(21):

$$\Delta W_o^j = -\mu \frac{\partial E}{\partial W_o^j}, j=1, \dots, M \quad (18)$$

$$\Delta v_o^j = -\mu \frac{\partial E}{\partial v_o^j}, j=1, \dots, D \quad (19)$$

$$\Delta W_h^j = -\mu \frac{\partial E}{\partial W_h^j}, j=1, \dots, M \quad (20)$$

$$\Delta v_h^j = -\mu \frac{\partial E}{\partial v_h^j}, j=1, \dots, D \quad (21)$$

where,

$$W_o^j = (w_{o1}^j, \dots, w_{oN}^j)^T, v_o^j = (v_{o1}^j, \dots, v_{oD}^j)^T$$

$$W_h^j = (w_{h1}^j, \dots, w_{hH}^j)^T, v_h^j = (v_{h1}^j, \dots, v_{hD}^j)^T$$

are weight vectors in the  $j^{\text{th}}$  node in the output and hidden layers, respectively.  $H$  is the number of nodes in the hidden layer. The system will be stable when the learning rate from Equations (18)-(21) satisfies the condition given in Equation (22):

$$\mu < \frac{TM}{2 \sum_{j=1}^M \sum_{p=1}^T \left[ \left\| \frac{\partial Y_{pj}}{\partial W_o^j} \right\|^2 + \left\| \frac{\partial Y_{pj}}{\partial v_o^j} \right\|^2 + \left\| \frac{\partial Y_{pj}}{\partial W_h^j} \right\|^2 + \left\| \frac{\partial Y_{pj}}{\partial v_h^j} \right\|^2 \right]} \quad (22)$$

Proof: Similarly, the Lyapunov function can be defined by Equation (23):

$$V(k) = \sum_{j=1}^M \sum_{p=1}^T e_{pj}^2$$

$$\Delta V(k) = \sum_{j=1}^M \sum_{p=1}^T \left[ 2e_{pj}(k) \Delta e_{pj}(k) + (\Delta e_{pj}(k))^2 \right] \quad (23)$$

Using the first order of the Taylor expansion of  $e_{pj}(k+1)$  yields Equation (24):

$$\Delta e_{pj} = -\frac{\mu}{MT} \left[ \frac{\partial e_{pj}}{\partial W_o^j} \cdot \sum_{s=1}^T e_{sj} \frac{\partial e_{sj}}{\partial W_o^j} + \frac{\partial e_{pj}}{\partial v_o^j} \cdot \sum_{s=1}^T e_{sj} \frac{\partial e_{sj}}{\partial v_o^j} \right.$$

$$\left. + \frac{\partial e_{pj}}{\partial W_h^j} \cdot \sum_{s=1}^T e_{sj} \frac{\partial e_{sj}}{\partial W_h^j} + \frac{\partial e_{pj}}{\partial v_h^j} \cdot \sum_{s=1}^T e_{sj} \frac{\partial e_{sj}}{\partial v_h^j} \right] \quad (24)$$

Substituting Equation (24) into Equation (23):

$$\Delta V(k) \leq \frac{-2\mu}{MT} \sum_{j=1}^M \left[ \left| \sum_{p=1}^T e_{pj} \frac{\partial e_{pj}}{\partial W_o^j} \right|^2 + \left| \sum_{p=1}^T e_{pj} \frac{\partial e_{pj}}{\partial v_o^j} \right|^2 \right.$$

$$\left. + \sum_{i=1}^H \left| \sum_{p=1}^T e_{pj} \frac{\partial e_{pj}}{\partial W_h^j} \right|^2 + \left| \sum_{p=1}^T e_{pj} \frac{\partial e_{pj}}{\partial v_h^j} \right|^2 \right]$$

$$+ \frac{4\mu^2}{(MT)^2} \sum_{j=1}^M \sum_{p=1}^T \left\| \frac{\partial e_{pj}}{\partial W_o^j} \right\|^2 \left\| \sum_{s=1}^T e_{sj} \frac{\partial e_{sj}}{\partial W_o^j} \right\|^2$$

$$+ \left\| \sum_{s=1}^T e_{sj} \frac{\partial e_{sj}}{\partial v_o^j} \right\|^2 + \sum_{i=1}^H \left\| \frac{\partial e_{pj}}{\partial W_h^j} \right\|^2 \left\| \sum_{s=1}^T e_{sj} \frac{\partial e_{sj}}{\partial W_h^j} \right\|^2 + \left\| \frac{\partial e_{pj}}{\partial v_h^j} \right\|^2 \left\| \sum_{s=1}^T e_{sj} \frac{\partial e_{sj}}{\partial v_h^j} \right\|^2$$

and using the following notation for simplicity,

$$\alpha = \sum_{j=1}^M \left[ \left\| \sum_{s=1}^T e_{sj} \frac{\partial e_{sj}}{\partial W_o^j} \right\|^2 \left\| \sum_{s=1}^T e_{sj} \frac{\partial e_{sj}}{\partial v_o^j} \right\|^2 \right.$$

$$\left. + \sum_{i=1}^H \left\| \frac{\partial e_{pj}}{\partial W_h^j} \right\|^2 \left\| \sum_{s=1}^T e_{sj} \frac{\partial e_{sj}}{\partial W_h^j} \right\|^2 + \left\| \frac{\partial e_{pj}}{\partial v_h^j} \right\|^2 \left\| \sum_{s=1}^T e_{sj} \frac{\partial e_{sj}}{\partial v_h^j} \right\|^2 \right]$$

a similar calculation then yields Equation (25):

$$\Delta V(k) \leq \frac{-2\mu}{MT} \alpha + \frac{4\mu^2}{(MT)^2} \alpha \sum_{j=1}^M \sum_{p=1}^T \left[ \left\| \frac{\partial e_{pj}}{\partial W_o^j} \right\|^2 \right.$$

$$\left. + \left\| \frac{\partial e_{pj}}{\partial v_o^j} \right\|^2 + \sum_{i=1}^H \left\| \frac{\partial e_{pj}}{\partial W_h^j} \right\|^2 + \left\| \frac{\partial e_{pj}}{\partial v_h^j} \right\|^2 \right] \quad (25)$$

To satisfy the Lyapunov condition, let the right-hand side of Equation (25) be less than zero, where further calculation leads to:

$$\mu < \frac{TM}{2 \sum_{j=1}^M \sum_{p=1}^T \left[ \left\| \frac{\partial Y_{pj}}{\partial W_o^j} \right\|^2 + \left\| \frac{\partial Y_{pj}}{\partial v_o^j} \right\|^2 + \left\| \frac{\partial Y_{pj}}{\partial W_h^j} \right\|^2 + \left\| \frac{\partial Y_{pj}}{\partial v_h^j} \right\|^2 \right]}$$

In general simulations, the learning factor was predefined as a constant whose value was selected by trial and error. The simulation performance value differs from the value of the learning factor. The learning process may converge or may not reach a satisfactory threshold with a different learning factor. From the results of the theorem above, the convergence is guaranteed if an adaptive learning factor is selected at an iteration of learning which satisfies the stability



condition. For purposes of simplifying the simulation, instead of calculating all  $\partial Y_{pj}/\partial W^l$  and  $\partial Y_{pj}/\partial v^l$  for  $l = 1, \dots, L$ ;  $j = 1, \dots, N_L$  the following corollary will provide a more restrictive but easier calculated condition.

Consider the infinite norm notation for any vector  $X = \{x_1, x_2, \dots, x_n\}$  that  $\|X\|_\infty = \max_{1 \leq i \leq n} \{x_i\}$ ; for simplicity, the notation  $\|\cdot\|$  is used in this paper and represents  $\|X\|_\infty$ . Applying an infinite norm from Equation (16) and using the following notation:

$$|v_j^i| = \sum_{d=1}^{D_v} v_{jd}^i$$

the calculation of  $\|\partial Y/\partial W^o\|$  and  $\|\partial Y/\partial v^o\|$  and leads to:

$$\left\| \frac{\partial Y}{\partial W^o} \right\| \leq \frac{\theta}{2} \left[ 1 + |v_j^o| \left\| \frac{\partial Y}{\partial v^o} \right\| \right]$$

$$\left\| \frac{\partial Y}{\partial v^o} \right\| \leq \frac{\theta}{2} \left[ 1 + |v_j^o| \left\| \frac{\partial Y}{\partial v^o} \right\| \right]$$

which then becomes Equation (26):

$$\left\| \frac{\partial Y}{\partial v^o} \right\| \leq \frac{\theta}{(2 - \theta |v_j^o|)}, \text{ and } \left\| \frac{\partial Y}{\partial W^o} \right\| \leq \frac{\theta}{(2 - \theta |v_j^o|)} \quad (26)$$

Further calculation of  $\|\partial Y/\partial W^h\|$  and  $\|\partial Y/\partial v^h\|$  leads to Equation (27):

$$\left\| \frac{\partial Y}{\partial W^h} \right\| \leq \frac{\theta}{2} \left( \sum_{j=1}^N w^o \left\| \frac{\partial Y}{\partial W^h} \right\| + |v_j^o| \left\| \frac{\partial Y}{\partial v^h} \right\| \right)$$

$$\leq \frac{\theta}{2} \left( |w^o| \left( \frac{\theta}{(2 - \theta |v_j^o|)} \right) + |v_j^o| \left\| \frac{\partial Y}{\partial v^h} \right\| \right)$$

$$\left\| \frac{\partial Y}{\partial W^h} \right\| \leq \frac{\theta^2 |w^o|}{(2 - \theta |v_j^o|)^2}, \text{ and } \left\| \frac{\partial Y}{\partial v^h} \right\| \leq \frac{\theta^2 |w^o|}{(2 - \theta |v_j^o|)^2} \quad (27)$$

Adding Equations (26) and (27) results in the following Corollary:

Corollary 1: The system defined by Equations (16) and (17) converges if the learning factor from Equations (18)–(21) satisfies the conditions of Equations (28) and (29):

$$2 - \theta |v^o| > 0 \quad (28)$$

$$\mu < \frac{(2 - \theta |v_j^o|)^2}{4\theta(2 + \theta(|w^o| - |v_j^o|))} \quad (29)$$

## Simulation

In this section, an example of a chaotic system, known as a Mackey-Glass chaotic time series system, is used to demonstrate the effectiveness of the methods developed in this study. The Mackey-Glass is a differential equation with a time delay system described by Equation (30):

$$\dot{Y}(t) = -bY(t) + \frac{aY(t-\tau)}{1+Y^{10}(t-\tau)} \quad (30)$$

with the initial condition given by Equation (31):

$$Y(0) = 1.2 \quad (31)$$

and assuming that  $Y(0) = 0$  when  $t < 0$ . The discrete form of the equation can be expressed as Equation (32):

$$Y(t+1) = bY(t) + \frac{aY(t-\tau)}{1+Y^{10}(t-\tau)} \quad (32)$$

Further assuming that  $a = 0.2$ ,  $b = 0.9$ , and  $\tau = 30$ , Figure 2 shows 300 data points generated from Equation (32).

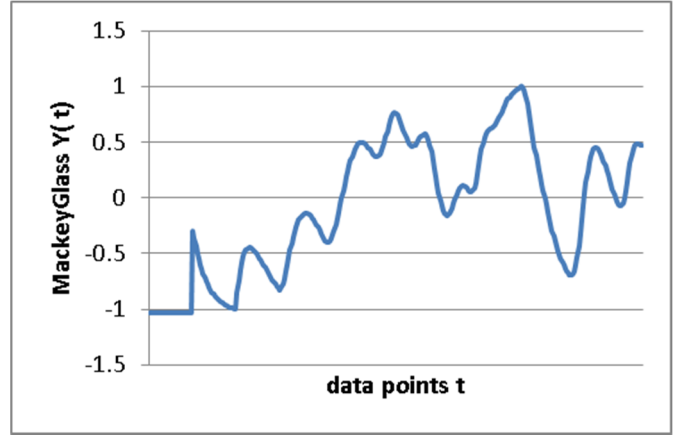


Figure 2. Mackey-Glass Time Delay System

A three-layer neural network structure was selected with five inputs, five nodes in the hidden layer, and one output. For each output,  $Y(t)$ , five inputs were defined:  $Y(t-6)$ ,  $Y(t-12)$ ,  $Y(t-18)$ ,  $Y(t-24)$ , and  $Y(t-30)$ .

With the constant learning factor, several values were used for the learning trials. After a number of attempts—with the slope set to 0.8, the learning factor set to a constant 0.1, and randomly generated initial weights—the system reached an absolute error of 0.05 after 34,666 iterations. Using 0.1 as an initial learning factor and applying the adap-

tive method as defined by Equation (29) at each iteration, learning took 7,139,911 iterations to reach an error of 0.05. Figure 3 shows the first 100 patterns of output of the neural network model. It was also observed that the error decreases steadily when the adaptive learning factor was applied. The oscillation of error was observed while a predefined constant learning factor is applied. Figure 4 shows the error behavior of learning with a constant factor of 0.1; Figure 5 shows the error behavior of learning using an adaptive learning factor with an initial factor of 0.1.

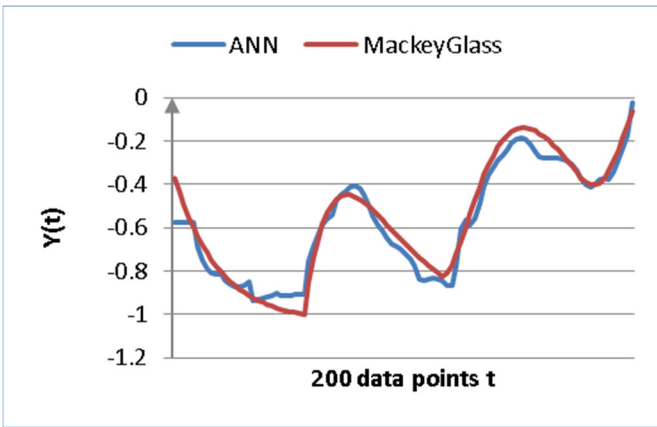


Figure 3. ANN Simulation of Mackey-Glass with a Constant Learning Factor of 0.1

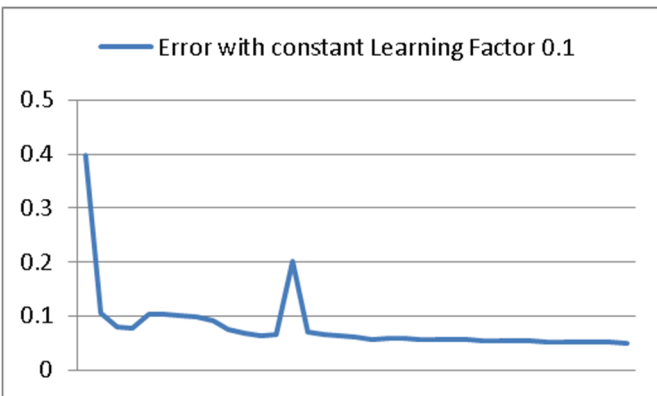


Figure 4. Error Behavior with a Learning Factor of 0.1

The adaptive learning factor guarantees that the errors will steadily decrease. The drawback is that requires many more computations, since an updated learning factor needs to be calculated at every weight update based on the back-propagation algorithm. It also might take more iterations to reach the same threshold. Applying the constant learning factor avoids the calculation, but a proper constant learning factor has to be identified through trial and error. Randomly selecting a learning factor of 0.3, in Figure 6, the plot shows a comparison of errors from the learning process with a constant learning factor and the adaptive learning factor. The

constant learning factor was selected with a value of 0.1. To compare the effectiveness of the adaptive learning method, the value of 0.1 was used as an initial learning factor in the adaptive method. Points for the plot were taken from the errors of every 1000 iterations.

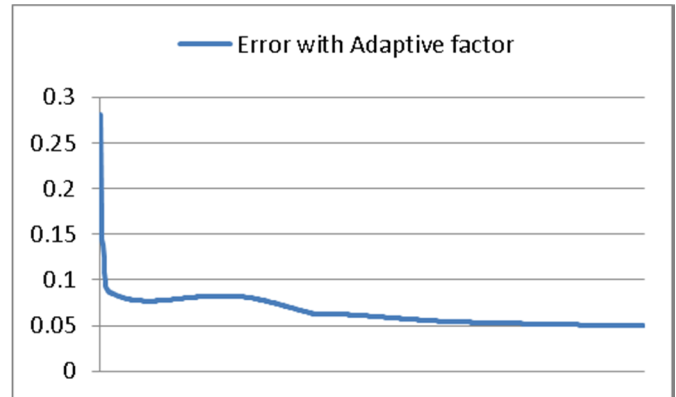


Figure 5. Error Behaviors with an Adaptive Learning Factor

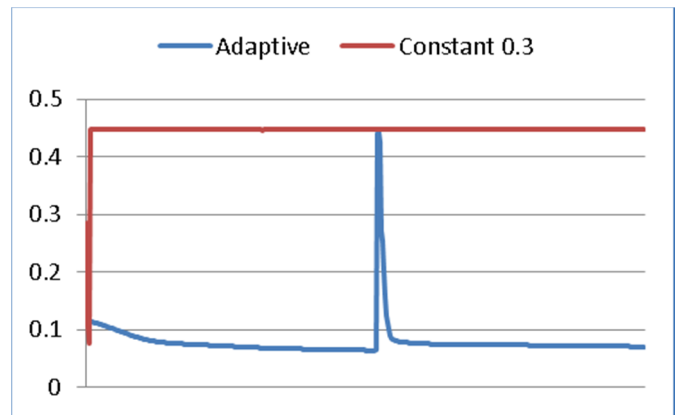


Figure 6. Comparison of Errors from Learning with Adaptive and Constant Learning Factors

## Summary

A Lyapunov function was defined for the learning process of the backpropagation algorithm and the stability of the system was proved when the process satisfied certain conditions. Furthermore, a more simplified condition was used to provide a feasible implementation of the adaptive learning factor. At each iteration of the learning process, an adaptive learning factor was selected satisfying the stability condition in order to avoid unstable phenomena. Simulation results of the Mackey-Glass system demonstrated that a learning factor chosen arbitrarily out of the predefined stability domain leads to an unstable identification of the considered system; however, an adaptive learning factor satisfying the conditions chosen for this study ensured the stability of the identification system.

---

## References

- [1] Raeisi, R., & Kaur, A. (2012). Artificial Neural Networks for Realization and Verification of Digital Logic Circuits. *International Journal of Modern Engineering, Spring/Summer, 12(2)*, 5-9.
- [2] Yu, W., Poznyak, A. S., & Li, X. (2001). Multilayer Dynamic Neural Networks for Non-linear System Identification. *International Journal of Control, 74* (18), 1858-1864.
- [3] Korkobi, T., Djemel, M., & Ctourou, M. (2008). Stability Analysis of Neural Networks-based System Identification. *Modeling and Simulation in Engineering, Volume 2008, Article ID 343940*.
- [4] Zhu, E .(2010). Stability Analysis of Recurrent Neural Networks with Random Delay and Markovian Switching. *Journal of Inequalities and Applications*, doi:10.1155/2010/191546.
- [5] Akea, H., Alassar, R., & Covachev, V. (2005). Stability of Neural Networks with Time Varying Delays in the Presence Impulses. *Advances in Dynamical Systems and Application, 1(1)*, 1-15.
- [6] Kolla, S. R. (2012). Comparison of Time Delay Controllers for A class of Networked Control System. *International Journal of Modern Engineering, Fall/Winter, 13(1)*, 51-56.
- [7] Choon, K. A. H. (2012) Stability Conditions for Fuzzy Neural Networks. *Advances in Fuzzy Systems, Volume 2012, Artical ID 281821*.
- [8] Zhao, H. (2003). Global Stability of Neural Network with Distributed Delays. *Journal of Neural, Parallel & Scientific Computations, 11(3)*, 237-252.
- [9] Liang, G. (2010). Global Asymptotically Stability of Cellular Neural Networks with Time-varying Delay. *2010 8<sup>th</sup> World Congress on Intelligent Control and Automation*, (pp 5031-5036).
- [10] Rojas, R. (1996). *Neural Networks*. Springer-Verlag.
- [11] Sontag, E. D. (1998). *Mathematical Control Theory*. (2<sup>nd</sup> ed.). Springer-Verlag, New York

## Biography

**HONG LI** is an Associate Professor of Computer Systems Technology at The New York City College of Technology of The City University of New York. She earned her Ph.D. in mathematics from the University of Oklahoma. Her interests include mathematical modeling, artificial neural networks, and software application development. Dr. Li may be reached at [hli@citytech.cuny.edu](mailto:hli@citytech.cuny.edu).

# FLUID STRUCTURE INTERACTION OF TRAUMATIC BRAIN INJURY: EFFECTS OF MATERIAL PROPERTIES ON SAS TRABECULAE

Mahsa Ghaffari, University of Illinois at Chicago; Mohamad Zoghi, State University of New York; Mostafa Rostami, Amirkabir University of Technology; Nabiollah Abolfathi, Amirkabir University of Technology

## Abstract

Traumatic Brain Injury (TBI) continues to be a major source of injury worldwide. Understanding the mechanism of injury will help in the construction of preventive solutions for TBI. This current study focused on subarachnoid space (SAS), which contains cerebrospinal fluid (CSF) and randomly oriented trabeculae. The research focus was on the role of the SAS region in traumatic brain injury. A set of three 2D finite element (FE) models was developed. The first model, a solid model, was a 2D mid-sagittal model, which was validated against a cadaver impact test. The solid model was used to determine the relative displacement between brain and skull. The second model was a fluid model to determine the CSF pressure under a blunt head impact. To simulate the fluid-structure interaction, an equivalent fluid was introduced, which represented the SAS trabeculae and the CSF. To study trabeculae locally, the relative displacement between the brain and skull, along with the pressure changes over time, were employed as boundary conditions in a new local model. The results showed that the response of the system is highly dependent upon the material properties of the SAS region. It was also shown that trabeculae are unable to transfer any loads to the brain in the compression mode, but trabeculae can control brain motion by transferring load to brain. Therefore, in tension, a tethering effect was observed, suggesting the load bearing capacity in this mode.

## Introduction

Traumatic Brain Injury (TBI) is a serious public health problem resulting from collisions in vehicular crashes and contact sports or owing to falls from a height. Injuries from transportation-related accidents are the most frequent types of personal injuries. The massive number of injuries sustained in accidents is a growing problem worldwide, especially in developing countries. In the U.S., approximately 1.4 million traumatic brain injuries occur each year [1]. Brain injury not only represents a serious disability for those involved, but also imposes a huge social and economic burden on society.

The development of effective head injury prevention strategies requires a better understanding of the mechanisms of head injury. In the last few decades, many efforts have been made by biomechanical investigators, which can be classified as experimental studies, analytical studies, physical models, and computational studies. Due to the complicated anatomy of the brain, finite element (FE) methods have been widely used to investigate the physical processes producing TBI. Furthermore, local stress and strain, which are the most influential parameters in the determination of the degree of head injury, can be extracted from a FE impact process in any part of the human head. For these reasons, FE modeling has played a growing role in the study of TBI. To date, various FE models have been developed such as the 2D works of Lee et al. [2], Ruan et al. [3], Ueno et al. [4], and Zhang et al. [5]. There are more extensive works performed in 3D models; however, the complicated geometry of the SAS and the trabeculae make it impossible to model all of the details in 3D.

Many experimental studies were taken into account to validate various physical, mathematical, and FE models. In this regard, two extensive cadaver studies were performed by Nahum et al. [6] and Trosseille et al. [7]. The resulting pressure and acceleration time histories from Nahum's research have been widely used as a validation source [8-11]. Anatomically, from exterior to interior, the human head consists mainly of scalp, skull, dura mater, arachnoid mater, subarachnoid space (SAS), pia mater, and brain. The skull is a three-layer structure consisting of trabecular bone sandwiched by two layers of cortical bone. Furthermore, the space between the arachnoid mater and pia mater, known as SAS, is filled with cerebrospinal fluid (CSF), which stabilizes the shape and the position of the brain during head movements.

The SAS that includes CSF and the trabeculae has a complex geometry. This is due to an abundance of trabeculae and their stretching (tree-like) shape from the arachnoid mater to the pia mater [9]. It has been shown that subarachnoid space (SAS) trabeculae are of great importance in damping head impact, thereby reducing traumatic brain injury (TBI) [12], [13]. The computational modeling of CSF in the subarachnoid space (SAS) was developed by Gupta et

al. [14] in order to study the flow dynamics of CSF in this space. In their study, the trabeculae were extended normal to the domain walls with the same configuration and variation of their shape and orientation neglected.

Based on the morphology of the SAS region, three material types—including soft solid, viscous fluid, and porous media—were proposed [14], [15]. A wide range of values of the mechanical properties of trabeculae were also reported [5], [15], [16]. In this regard, Saboori and Sadeqh [9] studied the effect of different material properties of trabeculae on the transfer of the load to the brain, and proposed the optimum material properties of the three material models.

## Material and Methods

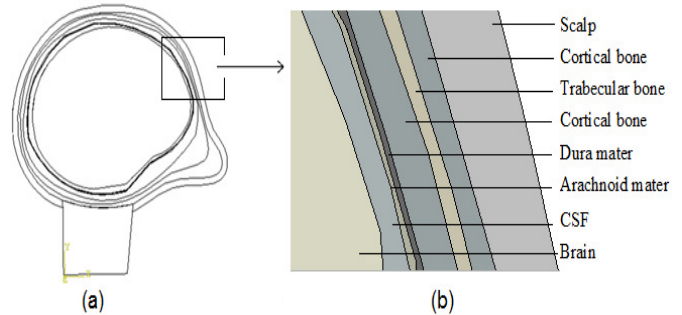
In order to analyze the role of trabeculae in traumatic brain injury, three FE models were developed. The first model was the solid model, which was employed for calculating the stress, strain, and the relative displacement between the brain and the skull. The second model was the fluid model, by which the CSF pressure data was determined when the head is subjected to an impact. The local model was the third model in which the extracted data from the first two models were employed as the boundary condition. In other words, the relative displacement between the brain and skull, extracted from the solid model, and the pressure distribution resulting from the fluid model, were used in the local model in which the stress and strain data in the meningeal layers, trabeculae, and brain could be extracted.

While the need for a detailed FE model of head trauma research may be obvious, the development of such a model has been a challenging task, due to the human head's complex geometry, material composition, as well as insufficient experimental data for model validation.

### The Solid Model

In order to analyze the role of trabeculae in the brain, three FE models were developed. The first, or solid model, was extracted from Horgan and Gilchrist's model [17], which is available on the Internet through the BEL repository managed by the Istituti Ortopedici Rizzoli, Bologna, Italy [18]. The mid-sagittal section of Horgan and Gilchrist's model was used to create the solid model, which consisted of three parts the brain, CSF, and skull. To enhance the solid model, some anatomical layers were added. That is, the solid model consisted of scalp, skull, dura mater, arachnoid mater, brain, and neck, as shown in Figure 1. Abaqus 6.10-1 software was used for pre-processing, post-processing, and

analysis of the model. The plane strain condition was assumed for this model. A summary of all of the materials used can be found in Table 1. Since the neck restraint will not affect the head response in a short duration impact [3], a free boundary condition was assumed in the analysis.



**Figure 1. (a) Solid Model used for Extracting the Relative Displacement; (b) Cutaway View of the Meningeal Layers of the Brain**

**Table 1. Material Properties used for the Solid and Local Models**

Layer	Thick-ness (mm)	Density (kg/m <sup>3</sup> )	Long-term elastic modulus (MPa)	Poisson's ratio
Calp	5	1130	16.7	0.42
Cortical bone	4	2000	13000	0.22
Trabecular bone	2	1300	888	0.3
Dura mater	0.4	1140	11720	0.23
Arachnoid mater	0.35	1130	19.32	0.45
Pia mater	0.15	1130	19.320	0.45
Trabeculae	1.5	1130	0.001	0.48
Brain	—	1040	0.0228	0.49

The skull was comprised of a layer of trabecular bone sandwiched between two layers of cortical bone. The compact bone was simulated as a linear viscoelastic material, and the experimental data of the compact bone of a human tibia, obtained by Lakes et al. [19], were applied in this study. The trabecular bone was also treated as a viscoelastic material, obtained from Yue and Wang [20]. The Prony series approximation for relaxation of the shear modulus of the trabecular bone and the cortical bone are shown in Table 2.



**Table 2. Properties of Materials with Viscoelastic Behavior**

Layers	Viscoelastic response (Prony series approximation)
Cortical bone	$g_R(t) = 1 - 0.0293(1 - e^{-t/9}) - 0.0656(1 - e^{-t/950}) - 0.0278(1 - e^{-t/9500}) - 0.107(1 - e^{-t/90,000})$
Trabecular bone	$g_R(t) = 1 - 0.6224(1 - e^{-t/711.23}) - 0.2143(1 - e^{-t/4267.4})$
Dura mater	$g_R(t) = 1 - 0.1088(1 - e^{-t/40}) - 0.0959(1 - e^{-t/10,000}) - 0.0922(1 - e^{-t/1000,000})$
Arachnoid mater	$g_R(t) = 1 - 0.919(1 - e^{-t/0.002})$
Pia mater	$g_R(t) = 1 - 0.919(1 - e^{-t/0.002})$
Brain	$g_R(t) = 1 - 0.815e^{-t/0.000143}$

In the solid model, according to the results from Bashkatov et al. [21], the thickness of the dura mater was 0.4mm. The viscoelastic model was assumed for the dura mater, the parameters for which were obtained from Yue and Wang [20]. The arachnoid mater was added to the model with a thickness of 0.35mm [22]. Its long-term elastic modulus was defined to be 19.32 MPa with Poisson's ratio of 0.45. The dimensionless relaxation modulus of the dura mater and the arachnoid mater are listed in Table 2 [22]. The thickness of the subarachnoid space was reduced approximately 1.5 mm [21] (varying throughout the model) by adding two layers of the arachnoid mater and dura mater. This space was only occupied by CSF. The CSF was modeled as a solid with an elastic modulus, bulk modulus, and Poisson's ratio of 0.15 MPa, 2.273 GPa, and 0.499989, respectively [23].

The material properties of brain tissue are important to the characteristics of relative brain-skull motion [23]. In order to model the brain, combined material properties of the grey and white matter of the brain was employed [25]. The long-term elastic modulus, bulk modulus, and Poisson's ratio were 22.8 kPa, 2.278 GPa, and 0.499991, respectively. In terms of viscoelasticity, the relaxation shear modulus,  $G_{R(t)}$ , was determined by the dimensionless function,  $g_{r(t)}$ , which is a way of expressing a Prony series in Abaqus:

$$g_r(t) = 1 - 0.8150e^{-t/0.00143} \quad (1)$$

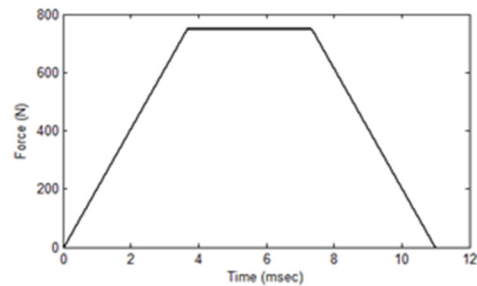
Based on the Abaqus documentation,  $G_0$  is the instantaneous shear modulus, and  $g_{R(t)} = G_{R(t)}/G_0$ . As time  $t$  goes to infinity, the long-term shear modulus can be obtained from the Prony series as:

$$G_\infty = G_{R(\infty)} = g_{R(\infty)} \times G_0 \quad (2)$$

The validated solid model was subjected to the frontal impact reported by Zoghi et al. [25] in the forklift truck accident. An equivalent dynamic force corresponding to the impact velocity of 5 m/s was applied to the solid model. This dynamic force was calculated [26] using Equation (3) [1]:

$$P_{dyn} = W \cdot \sqrt{\frac{v^2}{g \cdot \delta_{st}}} \quad (3)$$

where,  $P_{dyn}$  is the dynamic force;  $W$  is the head weight of 50 N;  $v$  is the impact velocity of 5m/s;  $g$  is the gravitational acceleration of 9.81 m/s<sup>2</sup>; and,  $\delta_{st}$  is the static deflection caused by the load  $W$  being applied on the model and equal to 1.8E-5 m. The equivalent dynamic force of the impact was 750 N, which was modified for the plane strain assumption and was applied in a trapezoid shape, as shown in Figure 2.



**Figure 2. Dynamics Input Load for the Solid Model [6]**

The interface between the skull and the brain has been modeled in many different ways ranging from purely tied (no-slip) to sliding (free slip). It has been indicated that the models using the tied contact definition correlated well with the experimental results in a frontal impact, while the ones with a sliding interface did not match up well [24]. Thus, in this study, all interfaces were tied. To avoid stress concentration, the load was distributed equally on 10 nodes. Due to the nature of the loading, a dynamic explicit mode was used in the solid model simulation. The solid model consisted of 4939 quadratic triangular elements of type CPE6M (10,980 nodes).

## The Fluid Model

One of the main goals of this study was to couple fluid dynamic forces in the SAS region with the FE solid models. For this investigation, based on the solid model, a 2D fluid model was created. The outer wall represents the interface

between skull and dura mater and the inner wall describes the interface between brain and pia mater layer. Gambit 2.3.16 software was used for pre-processing in order to generate mesh. Fluent 14.0 was used for the processing and post-processing of the results. A total of 3318 4-noded elements were used (2079 nodes). An unsteady dynamic analysis was performed and a fluid pressure diagram was determined in the three regions of frontal, superior, and occipital.

The trabeculae play a major role when taking pressure data from a fluid model because they act as obstacles in the pathway of flow. The complex geometry of trabeculae makes it difficult to model them in the fluid model. As long as the model is a 2D model, it will be impossible to develop an anatomy model of trabeculae because they will stop the circulation of CSF. In this part, two types of models were analyzed based on CSF behavior. In the first model, the CSF behaved as an incompressible Newtonian fluid with the density and viscosity the same as that of water at 37° [27]. In the second model, an equivalent fluid with consideration of trabeculae restriction proposed by Zoghi and Sadegh [12] was employed. The aim in that study was to determine the characteristics of an equivalent fluid which could be employed as the representative of the CSF and the trabeculae. The equivalent viscosity was calculated as  $\mu=0.24$  Pa.sec [12] and the density of the equivalent fluid was determined to be  $\rho=1100$  kg/m<sup>3</sup>, which is based on the mixture of CSF (30% with a density of 1000 kg/m<sup>3</sup>) and trabeculae (70% with a density of 1130 kg/m<sup>3</sup>).

### The Local Model

As mentioned above, a local model was developed to further analyze the results of the solid and fluid models. The objective of this model was to locally study the behavior of the trabeculae in the SAS. The emphasis of this model was on the meningeal layers and SAS region. The local model consisted of six layers representing the skull, dura mater, arachnoid, SAS region with randomly oriented trabeculae, pia mater, and brain, as shown in Figure 3. Based on Saboori and Sadegh's [9] experiment, approximately 70% of the trabeculae have a tree-shaped configuration where the branches are attached to the arachnoid and the stems are attached to the pia mater. The architecture and morphology of the local model was extracted from that study. In the current model, the trabeculae were modeled into two forms: pillar-shaped and tree-shaped configurations. The materials used in this model were the same as the solid model (see Tables 1 and 2). For the modeling of SAS trabeculae, the soft solid material model was chosen [9].

Since this model was a local model and cut from the real-size model, the boundary conditions at the planes of cut

were carefully considered. To suppress the effect of boundary conditions, the edges of the model were extended to twice the model size. Due to Saint-Venant's principle, there is no effect of boundary conditions far from the loading site. Abaqus 6.10-1 software was used for pre-processing, analysis, and post-processing. The total number of elements (linear quadrilateral elements of type CPE4R) and nodes was 3096 and 3218, respectively. All of the internal surfaces in the SAS region were subjected to the pressure, which was calculated from the fluid model.

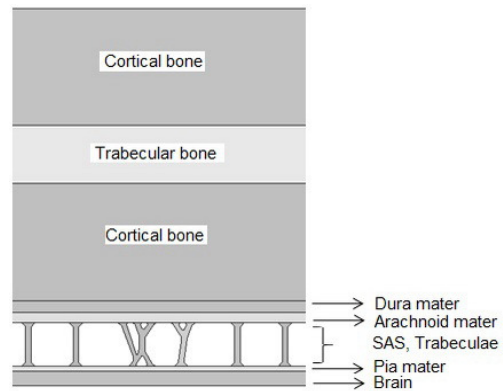


Figure 3. An illustration of the Local Model

## Results and Discussion

The FE model was well validated against the experimental data of the impact experiment by Nahum et al. [6]. In that experiment, the authors performed two series of experiments on cadavers. Series I consisted of eight individual impacts on a single specimen. Experiment number 37 (experiment #2 from series I) was selected for this validation. In this experiment, a seated, stationary cadaver subject was impacted by a rigid mass of 5.59 kg with a constant velocity of 9.94 m/s. The impact was delivered to the frontal bone in the mid-sagittal plane, in an anterior-posterior direction. The dynamic measurement of input force, which is shown in Figure 4, was recorded.

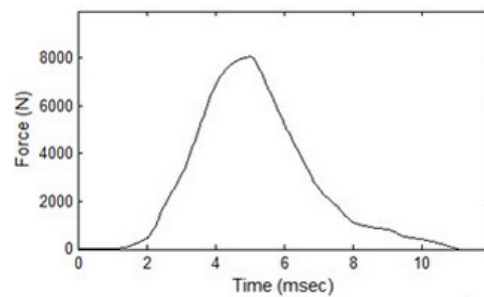
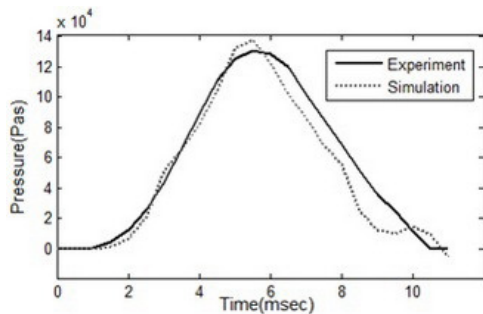


Figure 4. Input Force [6]

In the current model, the impact condition was applied by the input force as a function of time in the frontal direction. Note that the load was converted to the proper load for a plane strain assumption. The pressure response can be plotted during the impact process in the frontal region, and can be used for validation against experimental data presented by Nahum et al. [6]. As shown in Figure 5, the response of the model is in good agreement with the experimental results of Nahum et al. Figure 6 depicts the results of the relative displacement between the skull and brain, determined from the solid model, in the X and Y directions in frontal, occipital, and superior regions. Figure 6 also shows the relative displacement of the superior region in which the maximum tangential movement occurred.

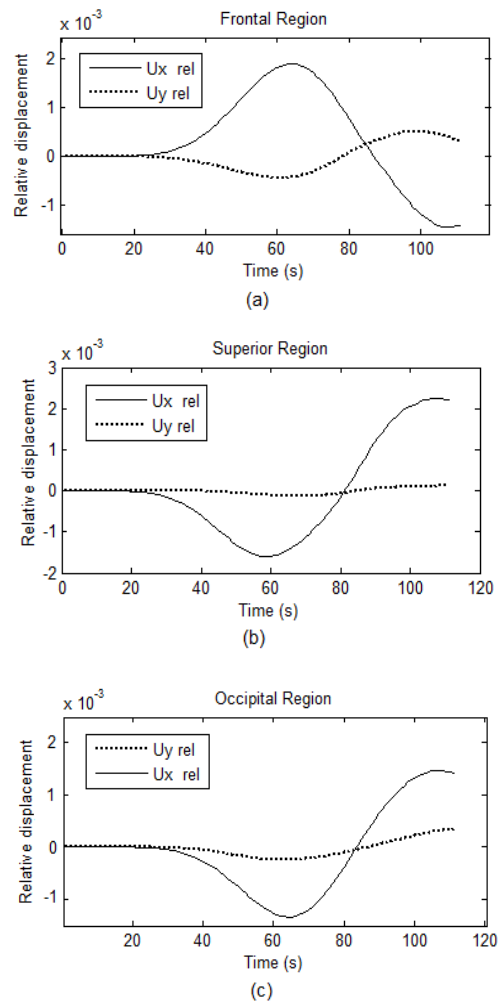


**Figure 5. Comparison and Numerical Analysis of a Cadaver Experiment [6]**

To compare the role of trabeculae and CSF, the fluid model was analyzed with the normal physiological condition of CSF and the equivalent CSF properties. As was previously discussed, the equivalent fluid represents the CSF and SAS trabeculae. In this analysis, the fluid characteristics of CSF were investigated. In particular, the pressure distribution over time was determined. Figure 7 shows that the position of the peak pressure moves faster in the absence of trabeculae over a specific period of time. Also, the pressure variation of CSF over time at the three locations of the head, frontal, superior, and occipital are shown in this figure. As shown, the pressure variation is smoother for the case with equivalent CSF properties. Figure 8 shows the pressure distribution of the CSF during the impact with consideration of trabeculae in different instances: 27, 54, 81, and 108 msec. As shown in this figure, the CSF pressure fluctuates back and forth during the impact, suggesting that all of the regions go under compression/tension states of stress. This variation is more significant in the frontal and occipital regions.

In order to focus on the trabeculae, the CSF effect was indirectly modeled in the local model. That is, pressure boundary conditions for each of the frontal, superior, and occipital regions, as shown in Figure 7, were employed in

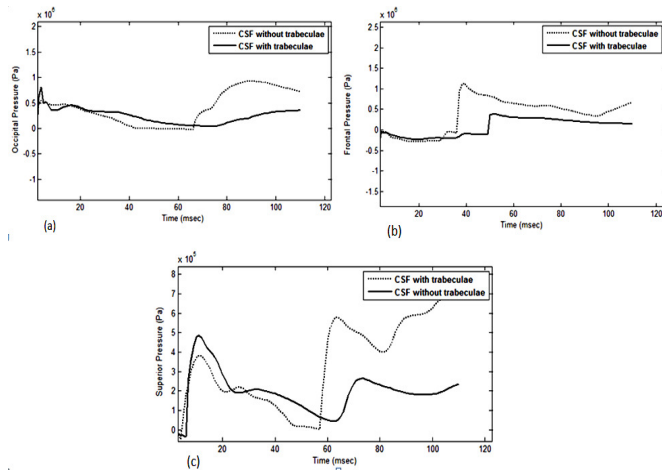
all trabeculae surfaces to simulate the CSF pressure. The model was subjected to the relevant displacement boundary condition in each region.



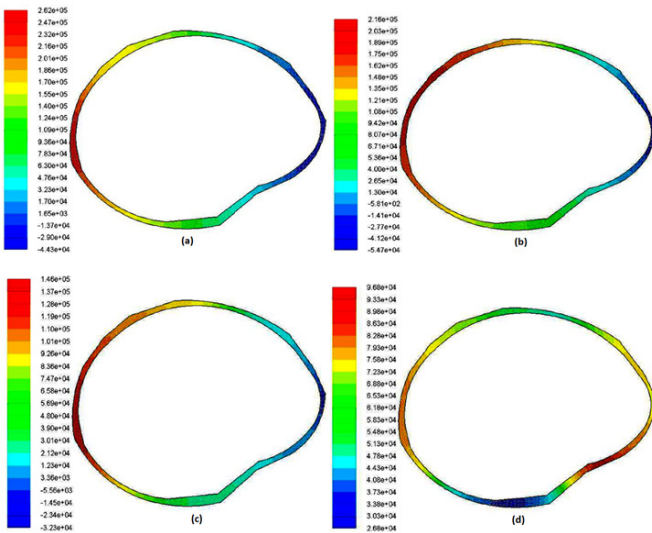
**Figure 6. Results of the Relative Displacement Between the Skull and Brain: Frontal (a), Superior (b), and Occipital (c) Regions**

In practice, buckling is characterized by a sudden failure of a structural member subjected to high compressive stress, where the actual compressive stress at the point of failure is less than the ultimate compressive stress that the material is capable of withstanding. In the local model, the trabeculae sustain no compressive stress before deformation. For this reason, the trabeculae cannot play a damping role in high compressive loads produced by blunt head impact. That is, trabeculae have a passive role in high compressive stress. On the other hand, in tension, the load was transferred to the pia mater and the brain through the trabeculae. In this situation, it can be assumed that the skull is harnessing the brain to control its movement. In other words, the active phase of

trabeculae would be evident. This behavior was seen in both frontal and occipital regions. In the superior region, in which the highest tangential force is applied to the model, it was also observed that the load is transferred to the pia mater through trabeculae. Furthermore, the tethering effect of trabeculae was highlighted.



**Figure 7. Comparison of Pressure Changes over Time—With and Without Trabeculae in the Frontal (a), Occipital (b), and Superior (c) Regions: Pressure Changes of the Trabeculae were Employed as a Boundary Condition for the Local Model**



**Figure 8. CSF Pressure Distribution over Time with Consideration of Trabeculae: (a)  $t = 27$  msec, (b)  $t = 54$  msec, (c)  $t = 81$  msec, (d)  $t = 108$  msec**

As discussed, depending on the loading and position of the trabeculae, they contribute both passively (in compression) and actively (in tension) in damping the impact. This shows that the combination of the trabeculae and CSF play a role in damping the impact.

## Conclusion

A major role for cerebrospinal fluid (CSF) is to provide effective damping against sudden intracranial brain motions during dynamic head impact. The morphology of the SAS with its complex trabeculae structure increases the flow resistance. It was shown that trabeculae yield smoother pressure responses in a head impact simulation. Thus, with the reduction of CSF pressure, trabeculae have an indirect damping effect. The aim of this study was to focus on the role of trabeculae in the SAS region. For this purpose, the local model was developed without the fluid elements of CSF. Instead, the pressure changes of CSF were employed as boundary conditions. It was shown that, in a case of SAS compression, trabeculae are unable to transfer any load to the brain; instead, they develop hydraulic resistance in combination with the CSF pressure. However, in tension, they continue to serve as damping elements both by providing tethering effects and contributing to hydraulic resistance.

In summary, a significant step in modeling the injury mechanism for TBI is to properly address the SAS region. This region comprises both solid state (trabeculae) and a fluid medium (CSF). The fluid-solid interaction along with the viscoelastic behavior of the head components (meningeal layers and brain) provides a damping mechanism for the impact. That is to say that without the fluid-solid interaction in the SAS region, the brain tissue would undergo the constant reciprocating coup-countercoup pressure. This would consequently result in tissue damage. However, due to the damping effect of the SAS region, the impact would rapidly diminish in the range of milli-seconds. This study delineated the significance of properly modeling the SAS region. As mentioned in the Results section, in the absence of trabeculae, the CSF pressure varies rather quickly; hence, the damping effect does not appear. On the contrary, by including the trabeculae (indirectly by using equivalent CSF fluid), the pressure distributions behave more smoothly, thereby providing an efficient damping mechanism. This is a preliminary step in properly defining the FE modeling of the head-brain. The subsequent steps will be in developing a 3D set of models to further analyze the injury mechanism. As stated earlier, validation continues to be a challenge for such models; therefore, more experimental research would be greatly beneficial in constructing a viable solution for TBI.

## References

- [1] Centers for Disease Control and Prevention (CDC), National Center for Injury Prevention and Control. (2005). *Facts about traumatic brain injury*.



- [2] Lee, M. C., Melvin, J. W., & Ueno, K. (1987). Finite Element Analysis of Traumatic Subdural Hematoma. *The 31st Stapp Car Crash Conference, paper number 872201*. USA.
- [3] Ruan, J. S., Khalil, T. B., & King, A. I. (1993). Finite Element Modeling of Direct Head Impact. *The 37th Stapp Car Crash Conference*, San Antonio, CA, USA.
- [4] Ueno, K., Melvin, J. W., Li, L., & Lighthall, J. W. (1995). Development of Tissue Level Brain Injury Criteria by Finite Element Analysis. *Journal of Neurotrauma*, 12(4), 695-706.
- [5] Zhang, L., Bae, J., Hardy, W. N., Monson, K. L., Manley, G. T., Goldsmith, W., et al. (2002). Computational Study of the Contribution of the Vasculature on the Dynamic Response of the brain. *Stapp Car Crash Journal*, 46, 145-64.
- [6] Nahum, A., Smith, R., & Ward, C. (1977). Intracranial Pressure Dynamics During Head Impact. *The 21st Stapp Car Crash Conference*, Warrendale, USA.
- [7] Trosseille, X., Tarriere, C., & Lavaste, F. (1992). Development of a FEM of the Human Head According to a Specific Test Protocol. *The 36th Stapp Car Crash Conference*, (pp. 235-253). Seattle, Washington, USA.
- [8] Horgan, T. (2005). *A Finite Element Model of the Human Head for Use in the Study of Pedestrian Accidents*. Ph.D. thesis at University College Dublin, Volume 9335.
- [9] Saboori, P., & Sadegh, A. (2011). Brain Subarachnoid Space Architecture: Histological Approach. *International Mechanical Engineering Congress and Exposition*, (pp. 89-94). Denver, USA.
- [10] Willinger, R., Kang, H. S., & Diaw, B. (1999). Three-Dimensional Human Head Finite-Element Model Validation against Two Experimental Impacts. *Annals of Biomedical Engineering*, 27(3), 403-410.
- [11] Zhou, C., Khalil, T. B., & King, A. I. (1995). A New Model Comparing Impact Responses of the Homogeneous and Inhomogeneous Human Brain. *The 39th Stapp Car Crash Conference, Society of Automotive Engineers*, Warrendale, PA, USA.
- [12] Zoghi, M., & Sadegh, A. (2010). Equivalent Fluid Model for CSF and SAS Trabeculae Using Head/Brain Damping. *International Journal of Biomed Engineering and Technology*, 4(3), 195-210.
- [13] Zoghi, M., & Sadegh, A., (2009). Global/local Head Models to Analyze Cerebral Blood Vessel Rupture Leading to ASDH and SAH. *International Journal of Computer Methods in Biomechanical and Biomedical Engineering*, 12(1), 1-12.
- [14] Gupta, S., Soellinger, M., Boesiger, P., & Kurtcuoglu, V. (2009). Three-dimensional Computational Modeling of Subject-specific Cerebrospinal Fluid Flow in the Subarachnoid Space. *Journal of Biomechanical Engineering*, 133, 123-133.
- [15] Zhang, L., Yang, K. H., Dwarampudi, R., Omori, K., Li, T., Chang, K., et al. (2001). Recent Advances in Brain Injury Research: A New Human Head Model Development and Validation. *Stapp Car Crash Journal*, 45, 369-394.
- [16] Jin, X., Lee, J., Leung, L., Zhang, L., & King, A. (2006). Biomechanics Response of the Bovine Pia-arachnoid Complex to Tensile Loading at Varying Strain-rates. *Stapp Car Crash Journal*, 50(6) 37-49.
- [17] Horgan, T., & Gilchrist, M. (2003). The Creation of Three-dimensional Finite Element Models for Simulating Head Impact Biomechanics. *International Journal of Crashworthiness*, 8(4), 353-366.
- [18] Istituti Ortopedici Rizzoli, Bologna, Italy [http://www.biomedtown.org/biomed\\_town](http://www.biomedtown.org/biomed_town).
- [19] Lakes, R. S., Katz, J. L., & Sternstein, S. S. (1979). Viscoelastic Properties of Wet Cortical Bone: Part I, Torsional and Biaxial Studies. *Journal of Biomechanics*, 12(9), 657-678.
- [20] Yue, X., & Wang, L. (2008). Amendment on the Strain Measurement of Thin-Walled Human Skull Shell as Intracranial Pressure Changes. *Journal University Science Technology Beijing*, 15(2), 202-208.
- [21] Bashkatov, A. N., Genina, A. E., Sinichkin Y. P., Kochubey, V. I., Lakodina N. A., & Tuchin V. V. (2003). Glucose and Mannitol Diffusion in Human Dura Mater. *Biophysics Journal*, 85(5), 3310-3318.
- [22] Ma, C., Jin, X., Zhang, J., & Huang, S. (2008). Development of the Pia-Arachnoid Complex Finite Element Model. *Bioinformatics and Biomedical Engineering, ICBBE, the 2nd International Conference*, (pp. 1717-1720). Shanghai, China.
- [23] Couper, Z., & Albermani, F. (2007). Infant Brain Subjected to Oscillatory Loading: Material Differentiation, Properties, and Interface Conditions. *Biomechanics and Modeling in Mechanobiology*. 7(2), 105-125.
- [24] Kleiven, S., & Hardy, W. N. (2002). Correlation of an FE Model of the Human Head with Local Brain Motion—Consequences for Injury Prediction. *Stapp Car Crash Journal*, 46, 123-144.
- [25] Zoghi, M., Watkins, C. B., Sadegh, A., & Dunlap, D. (2002). Simulation of Stand-Up Lift Truck Accidents to Evaluate Their Design and Operator Training Implications, Safety Engineering and Risk Analysis Division. *International Mechanical Engineering Congress and Exposition*, New Orleans, USA.
- [26] Ugural, A. C., & Fenster, S. K. (1995). *Advanced Strength and Applied Elasticity*. (3<sup>rd</sup> ed.). Prentice-Hall, Inc., NJ.
- [27] Kurtcuoglu, V., Soellinger, M., Summers, P. K.,



---

Boomsma, K. D., Poulidakos, D., Boesiger, P., et al. (2005). Reconstruction of Cerebrospinal Fluid Flow in the Third Ventricle Based on MRI Data. *Medical Image Computer and Computer-Assisted Intervention, 8th International Conference, Palm Springs, Computer Science*, (pp. 786-793). CA, USA.

## Biographies

**MAHSA GHAFFARI** is a Ph.D. student at the University of Illinois in Chicago, IL. She earned her B.S. degree from Islamic Azad University (Biomedical Engineering, 2010) and M.Sc. (Biomedical Engineering, 2012) from Amirkabir University of Technology. Her interests include computational modeling, fluid structure interaction, cellular and tissue biomechanics, micro-manufacturing, and microfluidics. Ms. Ghaffari may be reached at [ghaffari.mahsa@gmail.com](mailto:ghaffari.mahsa@gmail.com).

**MOHAMAD ZOGHI** is an Assistant Professor of Mechanical Engineering at the State University of New York-Farmingdale State College. He earned his B.S. degree from Sharif University, Tehran (Mechanical Engineering, 1998) and Ph.D. (Mechanical Engineering, 2006) from City University of New York. Dr. Zoghi is currently teaching at State University of New York-Farmingdale. His interests include computational modeling, energy harvesting, and energy storage. Dr. Zoghi may be reached at [mohamad.zoghi@farmingdale.edu](mailto:mohamad.zoghi@farmingdale.edu).

**MOSTAFA ROSTAMI** is an Associate Professor of Biomedical Engineering at Amirkabir University of Technology. He earned his M.Sc. and Ph.D. (Robotics, 1995 and 1999, respectively) from the University of Poitiers, France. Dr. Rostami is currently teaching at the Amirkabir University of Technology. His interests include trauma biomechanics, motion analysis, and robotic surgery. Dr. Rostami may be reached at [rostami@aut.ac.ir](mailto:rostami@aut.ac.ir).

**NABIOLLAH ABOLFATHI** is an Assistant Professor at Amirkabir University of Technology in Tehran. He earned his B.S. degree from Sharif University of Technology, (Mechanical Engineering, 1996) and Ph.D. (Mechanical Engineering, 2009) from the University of North Dakota. Dr. Abolfathi is currently teaching at Amirkabir University of Technology. His interests include traumatic brain injury and bio-MEMS. Dr. Abolfathi may be reached at [nabolfathi@aut.ac.ir](mailto:nabolfathi@aut.ac.ir).

# A LAPLACE-ADOMIAN SOLUTION FOR CLASSICAL FLUID DYNAMICS PROBLEMS

Amir Mobasher, Alabama A&M University; Mebougna Drabo, Alabama A&M University; Satilmis Budak, Alabama A&M University

## Abstract

A number of problems in engineering require solutions of partial differential equations (PDE). Such problems include water hammer problems, Couette flow, Burgers' equation, thermoelectric systems, chemical processing, general manufacturing, bio-heat transfer, electronic cooling, comfort heating and cooling, and a number of natural phenomena from upwelling currents in the oceans to heat transport in stellar atmospheres, to name a few. In this study, a methodology for the solution of classical thermal and fluid dynamic problems, governed by a set of partial differential equations, was developed. The technique employed the Laplace transform for the transformation of partial differential equations to ordinary differential equations and then utilized the Adomian polynomials to solve the resulting equations. Unlike the separation-of-variables method, this technique makes no assumption such as  $U=X(x).Y(y)$ . Two typical classical fluid dynamic problems are presented here in order to demonstrate this methodology. It will also be shown that, in benchmark cases, the Laplace-Adomian method can yield identical solutions to the analytical methods.

## Introduction

Heat transfer continues to be a major field of interest to engineering and scientific researchers. The wealth of applications includes a wide variety of components and systems for energy devices, including general power systems, heat exchangers, high-performance gas turbines, and other power-conversion devices [1]. One of the more important areas concerned with heat flow is in thermoelectric materials. Thermoelectric devices convert heat differences into electricity. Currently, research is being done on different thermoelectric, multilayer thin-film systems for energy harvesting from heat waste [2], [3]. Theoretical modeling is needed to explain heat flow in highly efficient thermoelectric devices, since the efficiency of the fabricated thermoelectric devices is directly related to the heat flow in the devices. Many elaborate numerical schemes, mostly categorized under finite element, finite difference, or finite volume, are available for solving the governing equations with high accuracy. Once a numerical scheme is developed to test its accuracy, one must test it against certain benchmark problems with known analytical solutions.

However, once the problem becomes multidimensional and nonlinear (such as Navier-Stokes), the analytical solutions can only be employed for the solution of limited and simplified forms of those physical problems. Two of the most widely used methods for obtaining analytical solutions to the governing equations of heat transfer and fluid flow problems are the separation of variables (also known as the Fourier method) and Integral transform methods [4], [5]. When solving the typical heat equation using the separation of variables technique, one assumes the solution to be a linear function of the independent variables; i.e.,  $u=X(x)T(t)$ . Substitution of the solution into the original PDE results in two ordinary differential equations which are solved independently. The application of boundary conditions generates a series solution to each of the resulting ODEs with constant coefficients which are determined via the Fourier method.

Integral methods offer an alternative approach to the solution of heat-equation-type problems in that the PDE is transformed to an ODE via Laplace transformation and the resulting ODE is solved by the application of the residue theory. It is interesting to note that both methods lead to identical solutions of the same problem, although they represent completely different approaches.

One over-looked approach for the solution of partial differential equations is the method proposed by Adomian. By using Adomian Polynomials, Adomian obtained a series solution with a fast convergence rate to a number of different partial and ordinary differential equations with and without linearity [6-10]. In some cases, similarity transform methods are used to convert partial differential equations into a set of linear or non-linear ordinary differential equations and the resulting equation (usually nonlinear) is solved by the Adomian Method [11]. However, one drawback of the Adomian method is that in certain cases the solution is only partially convergent. In other words, the solution converges for a part of the domain after which the solution starts to diverge.

For complete convergence it is necessary to increase the number of terms to more than 60. Mobasher et al. [12] used a piecewise technique to address the convergence issue and used this technique to solve problems of a highly stiff nature. In that technique, the domain was divided into sub-domains and a solution was obtained for each sub-domain.

The solution was advanced by substituting the end condition of the previous sub-domain as the initial conditions for the new sub-domain. It was shown that with as few as two terms, complete convergence with high accuracy was achieved. Another drawback of the Adomian method is that the initial conditions for the problem must be well defined, integrable functions. However, in many physical engineering problems, the initial conditions to the problem are constants. For example, in the case of the Couette flow problem, the initial condition is defined as 0. This poses a challenge when calculating the Adomian polynomials, since the subsequent solutions,  $u$ , are functions of the anti-derivatives of the initial condition.

In this paper, the authors present a methodology for solving linear one-dimensional, time-dependent problems such as the heat equation. This technique employs the Laplace transform to reduce the governing partial differential equation to an ordinary differential equation (ODE). Then it utilizes the Adomian polynomials to expand the resulting ODE into a series solution. The resulting solution is then presented as a closed-form solution. Unlike separation of variables and Integral methods, this remarkably simple technique makes no assumptions about the solution that it has separate functions of the independent variables and eliminates the tedious evaluation of Fourier integrals. This method produces identical solutions to those obtained by the separation of variables technique and Integral transform methods. To demonstrate this technique, three examples with known analytical solutions are presented here.

## Methodology

Consider the general partial differential equation of the form given in Equation (1):

$$\frac{\partial^n u}{\partial t^n} - \frac{\partial^m u}{\partial x^m} = 0 \quad (1)$$

Taking the Laplace transform of Equation (1) yields Equation (2):

$$s^n U(s) - \sum_{k=1}^n s^{n-k} u^{(k-1)}(0) - \frac{d^m U}{dx^m} = 0 \quad (2)$$

By reducing the PDE to an ODE, one can express the solution,  $U$ , as the sum of solutions  $U_0, U_1$ , etc., as:

$$U = \sum_{i=0}^{\infty} U_i(x, s) \quad (3)$$

where,

$$U_0 = A_0 + A_1 x + \dots + A_{m-1} x^{m-1} + L^{-m} \left( \sum_{k=1}^n s^{n-k} u^{(k-1)}(0) \right) \quad (4)$$

and where,  $L^{-m}$  is the  $m^{\text{th}}$  anti-derivative and  $A_0, A_1$ , etc., are constant coefficients determined by the specified boundary conditions. Subsequently,

$$\begin{aligned} U_1 &= L^{-m} (s^n U_0(s, x)) \\ U_2 &= L^{-m} (s^n U_1(s, x)) \\ &\vdots \end{aligned} \quad (5)$$

By taking the inverse Laplace transformation of the function  $U(x, s)$ , the solution,  $u(x, t)$ , in the  $(x, t)$  space is obtained and is accomplished either by analytical evaluation of  $U$ , using the residue theory, or numerical evaluation.

## Results and Discussions

The Laplace-Adomian method was demonstrated through three different examples for which analytical solutions exist. In the first case, the methodology was demonstrated via a simple example for which the solution was known to be  $u=xt$ .

Example 1. Consider the partial differential in Equation (6) with the specified boundary conditions (a-c):

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = x \quad (6)$$

$$u(0, t) = 0 \quad (a)$$

$$u(1, t) = t \quad (b)$$

$$u(x, 0) = 0 \quad (c)$$

Taking the Laplace transform of Equation (6) with its associated boundary conditions yields the ordinary differential equation given as Equation (7):

$$-\frac{d^2 U}{dx^2} + sU - u(x, 0) = \frac{x}{s} \quad (7)$$

$$U(0, s) = 0 \quad (a)$$

$$U(1, s) = \frac{1}{s^2} \quad (b)$$

$$U(x, 0) = 0 \quad (c)$$

From Equation (4),

$$U_0 = A_0 + A_1 x - L^{-2} \left( \frac{x}{s} \right) = A_0 + A_1 x - \frac{1}{s} \frac{x^3}{3!} \quad (8)$$

where,  $L^{-2}(\cdot) = \int dx \int (\cdot) dx$ . Therefore,

$$U_1 = L^{-2} U_0 = \left( A_0 \frac{x^2}{2!} + A_1 \frac{x^3}{3!} - \frac{1}{s} \frac{x^5}{5!} \right) s \quad (9)$$

$$U_2 = L^{-2} U_1 = \left( A_0 \frac{x^4}{4!} + A_1 \frac{x^5}{5!} - \frac{1}{s} \frac{x^7}{7!} \right) s^2$$

$$\vdots$$

Substituting Equations (8) and (9) into Equation (3), rearranging terms and applying the boundary conditions (7-a),  $A_1=0$ . The solution,  $U$ , becomes:

$$U = A_0 \left( s^0 + s \frac{x^3}{3!} + s^2 \frac{x^5}{5!} + \dots \right) \frac{\sqrt{s}}{\sqrt{s}}$$

$$- \frac{1}{s} \left( \frac{x^3}{3!} + s \frac{x^5}{5!} + s^2 \frac{x^7}{7!} + \dots \right) \frac{s}{s} - \frac{1}{s} x + \frac{1}{s} x \quad (10)$$

$$= \frac{A_0 \sinh \sqrt{s} x}{\sqrt{s}} - \frac{1}{s^2} \sinh \sqrt{s} x + \frac{x}{s^2}$$

Applying boundary condition of Equation (7b) results in  $A_0 = 1/s^{3/2}$ . Substitution of  $A_0$  into Equation (10) eliminates the first two terms in the series, resulting in  $U = x/s^2$ . Finally, taking the inverse Laplace transform, with  $\mathcal{L}^{-1}\{1/s^2\} = t$ , yields  $u(x,t) = xt$ . A table of inverse Laplace transforms for numerous functions is supplied by Poularikas [5].

Example 2. Solution of Couette flow with  $v = 1$ :

$$\frac{\partial u}{\partial t} - v \frac{\partial^2 u}{\partial x^2} = 0 \quad (11)$$

$$u(0, t) = 0 \quad (a)$$

$$u(1, t) = 1 \quad (b)$$

$$u(x, 0) = 0 \quad (c)$$

Taking the Laplace transform of Equation (11) yields:

$$sU - \frac{d^2 U}{dx^2} = 0 \quad (12)$$

$$U(0, s) = 0 \quad (a)$$

$$U(1, s) = \frac{1}{s} \quad (b)$$

$$U(x, 0) = 0 \quad (c)$$

Here, the initial condition of Equation (12c) is implemented:

$$U_i = L^{-1}(sU_{i-1}) \quad (13)$$

The subsequent solutions are:

$$U_0 = A_0 + A_1 x$$

$$U_1 = s \left( A_0 \frac{x^2}{2!} + A_1 \frac{x^3}{3!} \right) \quad (14)$$

$$U_2 = s^2 \left( A_0 \frac{x^4}{4!} + A_1 \frac{x^5}{5!} \right)$$

$$\vdots$$

And the solution,  $U$ , is:

$$U = \sum_{i=0}^{\infty} U_i(x, s) \quad (15)$$

Regrouping and simplifying terms yields:

$$U = A_0 \left( s^0 + s \frac{x^2}{2!} + s^2 \frac{x^4}{4!} + \dots \right)$$

$$+ A_1 \left( x + s \frac{x^3}{3!} + s^2 \frac{x^5}{5!} + \dots \right) \frac{\sqrt{s}}{\sqrt{s}} \quad (16)$$

or

$$U = A_0 \cosh(\sqrt{s} x) + \frac{A_1}{\sqrt{s}} \sinh(\sqrt{s} x) \quad (17)$$

Applying boundary conditions (12-a) and (12-b) yields  $A_0=0$  and  $A_1 = \frac{1}{\sqrt{s} \sinh(\sqrt{s})}$ , respectively.

Substituting  $A_0$  and  $A_1$  into Equation (17) yields:

$$U(x, s) = \frac{1}{s} \frac{\sinh(\sqrt{s} x)}{\sinh(\sqrt{s})} \quad (18)$$

The solution  $u(x,t)$  is obtained by application of the residue theory:

$$u(x,t) = x + \sum_{n=1}^{\infty} \frac{2}{n\pi} (-1)^n \exp(-n^2 \pi^2 t) \sin(n\pi x) \quad (19)$$

Example 3. One-dimensional Burgers' equation: the one dimensional Burgers' equation developed by Burgers [13] is widely used in benchmarking problems for a variety of schemes [19]. Adomian solved this equation directly to approximate the solution using Adomian polynomials [14], [15]. Although this equation is nonlinear, in certain cases exact solutions exist for this problem and numerical solutions have been tested for this problem [16]. In this current study, the approach taken to solve Burgers' equation was to employ the Cole-Hopf transformation in order to convert the non-linear PDE into a linear, homogeneous PDE in the form of Equation (1) and then apply the Laplace-Adomian methodology to solve the resulting equation. Consider Burgers' equation and the prescribed boundary conditions represented in Equation (20):

$$\frac{\partial \phi}{\partial t} + \phi \frac{\partial \phi}{\partial x} = \nu \frac{\partial^2 \phi}{\partial x^2} \quad (20)$$

$$\phi(x, t) = \sin \pi x \quad (a)$$

$$\phi(0, t) = 0 \quad (b)$$

$$\phi(1, t) = 0 \quad (c)$$

Using the Cole-Hopf transformation, this non-linear equation is reduced to a second-order linear PDE:

$$\frac{\partial u}{\partial t} = \nu \frac{\partial^2 u}{\partial x^2} \quad (21)$$

$$u(x,0) = \exp\left\{-\left(\frac{1-\cos \pi x}{2\nu}\right)\right\} \quad (a)$$

$$u'(0, t) = 0 \quad (b)$$

$$u'(1, t) = 0 \quad (c)$$

Taking the Laplace transform in order to convert the PDE into its corresponding ODE yields Equation (22):

$$\frac{d^2 U}{dx^2} = -u(x,0) + sU \quad (22)$$

The solution, then, is  $U_i = L^{-2}(sU_{i-1})$  with  $U_0 = A_0 + A_1 x - L^{-2}(u(x,0))$ . Since an analytical solution to the integral of the initial condition of Equation (21a) is not available,  $u(x,0)$  was represented by the Taylor series and term-by-term integration was applied to the subsequent series representation. Here, the expansion was carried out to four terms:

$$\begin{aligned} \exp\left(\frac{1-\cos \pi x}{2\nu}\right) = & \\ 1 - a(1-\cos \pi x) + \frac{a^2}{2!}(1-\cos \pi x)^2 & \quad (23) \\ + \frac{a^3}{3!}(1-\cos \pi x)^3 + \dots & \end{aligned}$$

$$\text{where, } a = \frac{1}{2\nu}$$

After simplification:

$$u(x,0) = K'_0 + K'_1 \cos \pi x + K'_2 \cos^2 \pi x + K'_3 \cos^3 \pi x + \dots \quad (24)$$

where,

$$\begin{aligned} K'_0 &= \left(1 - a + \frac{a^2}{2!} - \frac{a^3}{3!}\right) \\ K'_1 &= \left(a - \frac{2a^2}{2!} + \frac{3a^3}{3!}\right) \\ K'_2 &= \left(\frac{2a^2}{2!} - \frac{3a^3}{3!}\right) \\ K'_3 &= \frac{a^3}{3!} \end{aligned} \quad (25)$$

Applying the operator  $L^{-2}$  to the initial condition yields:

$$\begin{aligned} L^{-2}(u(x,0)) &= K_0 \frac{x^2}{2!} - K_1 \frac{\cos(\pi x)}{\pi^2} \\ &- K_2 \frac{\cos(2\pi x)}{(2\pi)^2} \\ &- K_3 \frac{\cos(3\pi x)}{(3\pi)^2} \end{aligned} \quad (26)$$



$$\begin{aligned}
K_0 &= \left( K'_0 + \frac{K'_2}{2} \right) \\
K_1 &= \left( K'_1 + \frac{3}{4} K'_3 \right) \\
K_2 &= \frac{K'_2}{2} \\
K_3 &= \frac{K'_3}{4}
\end{aligned}
\tag{27}$$

Substituting this into Equation (26) yields Equation (28):

$$\begin{aligned}
U_0 &= A_0 + A_1 x + K_0 \frac{x^2}{2!} + K_1 \frac{\cos(\pi x)}{\pi^2} \\
&+ K_2 \frac{\cos(2\pi x)}{(2\pi)^2} + K_3 \frac{\cos(3\pi x)}{(3\pi)^2}
\end{aligned}
\tag{28}$$

$$U_1 = L^{-1}(sU_0) \tag{29}$$

$$\begin{aligned}
U_1 &= \left\{ A_0 \frac{x^2}{2!} + A_1 \frac{x^3}{3!} - K_0 \frac{x^4}{4!} - K_1 \frac{\cos(\pi x)}{(\pi)^4} \right. \\
&\left. - K_2 \frac{\cos(2\pi x)}{(2\pi)^4} - K_3 \frac{\cos(3\pi x)}{(3\pi)^4} \right\}_s
\end{aligned}
\tag{30}$$

After grouping terms and manipulation, the general expression for U is obtained as follows:

$$\begin{aligned}
U &= \sum U_i \\
&= A_0 \left\{ 1 + s \frac{x^2}{2!} + \frac{x^4}{4!} s^2 + \dots \right\} \\
&+ A_1 \left\{ x + s \frac{x^3}{3!} + \frac{x^5}{5!} s^2 + \dots \right\} \frac{\sqrt{s}}{\sqrt{s}} \\
&- \frac{K_0}{s} \left\{ 1 + \frac{sx^2}{2!} + \frac{s^2 x^4}{4!} + \frac{s^3 x^6}{6!} + \dots \right\} + \frac{K_0}{s} \\
&+ \frac{K_1}{\pi^2} \cos \pi x \left\{ 1 - \frac{s}{\pi^2} + \frac{s^2}{\pi^4} - \dots \right\}
\end{aligned}
\tag{31}$$

Further simplification results in Equation (32):

$$\begin{aligned}
U &= A_0 \cosh \sqrt{s} x + \frac{A_1}{\sqrt{s}} \sinh \sqrt{s} x \\
&- \frac{K_0}{s} \cosh \sqrt{s} x + \frac{K_0}{s} \\
&+ K_1 \cos \pi x \left( \frac{\pi^2}{s + \pi^2} \right) \\
&+ K_2 \cos 2\pi x \left( \frac{(2\pi)^2}{s + 4\pi^2} \right) \\
&+ K_3 \cos 3\pi x \left( \frac{(3\pi)^2}{s + 9\pi^2} \right)
\end{aligned}
\tag{32}$$

Differentiating to get U', then, yields Equation (33):

$$\begin{aligned}
U' &= A_0 \sqrt{s} \sinh \sqrt{s} x + A_1 \sqrt{s} \cosh \sqrt{s} x \\
&- \frac{K_0 \sqrt{s}}{s} \sinh \sqrt{s} x + \frac{K_0}{s} \\
&- \frac{K_1 \pi}{s + \pi^2} \sin \pi x - \frac{2K_2 \pi}{s + 4\pi^2} \sin(2\pi x) \\
&- \frac{3K_3 \pi}{s + 9\pi^2} \sin(3\pi x)
\end{aligned}
\tag{33}$$

Applying boundary conditions from Equations (21b) and (21c) gives  $A_0 = K_0/s$  and  $A_1 = 0$ , respectively. Substitution into Equations (32) and (33), solution U and its derivative, U', are obtained, respectively, as Equations (34) and (35):

$$\begin{aligned}
U &= \frac{K_0}{s} + \frac{K_1}{s + \pi^2} \cos \pi x \\
&+ \left( \frac{2K_3}{s + 4\pi^2} \right) \cos(2\pi x) \\
&+ \left( \frac{3K_3}{s + 9\pi^2} \right) \cos(3\pi x)
\end{aligned}
\tag{34}$$

$$\begin{aligned}
U' &= - \frac{K_1 \pi}{s + \pi^2} \sin(\pi x) \\
&- \frac{4K_2 \pi}{s + 4\pi^2} \sin(2\pi x) \\
&- \frac{9K_3 \pi}{s + 9\pi^2} \sin(3\pi x)
\end{aligned}
\tag{35}$$

Taking inverse Laplace transforms of  $U$  and  $U'$ , and substitution into Equation (38), the solution in the  $(x,t)$  domain is obtained.

$$u(x,t) = K_0 + K_1 \exp(-\pi^2 t) \cos \pi x + 2K_2 \exp(-4\pi^2 t) \cos(2\pi x) + 3K_3 \exp(-9\pi^2 t) \cos(3\pi x) \quad (36)$$

$$u'(x,t) = -K_1 \pi \exp(-\pi^2 t) \sin \pi x - 4K_2 \pi \exp(-4\pi^2 t) \sin(2\pi x) - 9K_3 \pi \exp(-9\pi^2 t) \sin(3\pi x) \quad (37)$$

$$\phi(x,t) = -2 \frac{u'(x,t)}{u(x,t)} \quad (38)$$

The exact solution, then, is:

$$\phi_{ex}(x,t) = 2\pi \frac{\sum_{n=1}^{\infty} a_n \exp(-n^2 \pi^2 t) n \sin(n\pi x)}{a_0 + \sum_{n=1}^{\infty} a_n \exp(-n^2 \pi^2 t) \cos(n\pi x)}$$

The solutions for  $\nu=1$  and times 0.005, 0.05, and 0.2 seconds were compared with the Adomian method and exact solutions are shown in Table 1. Figure 1 is a comparison of the two methods at  $t = 0.005, 0.05, 0.1,$  and  $0.2$  seconds. From Table 1, it is apparent that the exact solution compares well with the current method for the case of  $\nu=1$ . For smaller values of  $\nu$ , additional terms of the Taylor series expansion are required to produce higher accuracy.

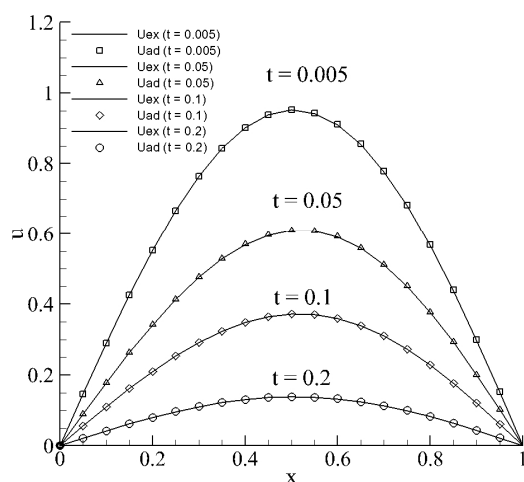


Figure 1. Comparison of Exact Solutions with the Laplace-Adomian Method for  $\nu=1$

Table 1. Comparison of Exact Solutions with the Laplace-Adomian Method for  $\nu=1$

	t=0.005	0.05	0.2			
x	Exact	Adom	Exact	Adom	Exact	Adom
0	0	0	0	0	0	0
0.1	0.2902	0.2902	0.1780	0.1782	0.0419	0.0420
0.2	0.5531	0.5531	0.3414	0.3417	0.0800	0.0801
0.3	0.7636	0.7637	0.4759	0.4763	0.1106	0.1108
0.4	0.9012	0.9016	0.5686	0.5693	0.1308	0.1310
0.5	0.9518	0.9527	0.6091	0.6099	0.1385	0.1387
0.6	0.9092	0.9109	0.5904	0.5914	0.1326	0.1328
0.7	0.7765	0.7790	0.5111	0.5121	0.1135	0.1136
0.8	0.5660	0.5685	0.3767	0.3775	0.0828	0.0830
0.9	0.2982	0.2998	0.1999	0.2003	0.0437	0.0438
1	0	0	0	0	0	0

## Conclusions

The Laplace-Adomian method was used to obtain exact and approximate solutions to the heat equation and non-linear Burgers' equation, respectively. In cases of homogeneous, time-dependent equations, exact solutions were obtained and shown to be identical to solutions found by other standard methods such as separation of variables and integral transform methods. The method is simple to use and, unlike the separation of variable techniques, makes no assumptions. With a four-term Taylor series expansion of the initial conditions, the method was able to provide a very accurate solution to Burgers' equation for  $\nu=1$ . For small values of  $\nu$ , additional terms of the series are needed to produce higher accuracy. In future studies, focus will be given to the solution of these problems directly in experimental heat flow studies for energy harvesting from heat waste in thermoelectric materials along with a comparison of theoretical modeling and experimental results.

## References

- [1] Goldstein, R. J., Eckert, E. R. G., Ibele, W. E., Patankar, S. V., Simon, T.W., Kuehn, T. H., et al. (2002). Heat transfer – a review of 2000 literature. *International Journal of Heat and Mass Transfer*, 45 (14), 2853–2957.
- [2] Budak, S., Parker, R., Smith, C., Muntele, C., Heidary, K., Johnson, R. B., et al. (2013). Super-lattice

- Multi-nano-layered Thin Films of SiO<sub>2</sub>/SiO<sub>2</sub>+Ge for Thermoelectric Device Applications. *Journal of Intelligent Material Systems and Structures*, 24(11), 1357-1364.
- [3] Budak, S., Smith, C., Muntele, C., Chhay, B., Heidary, K., Johnson, R. B., et al. (2013). Thermoelectric Properties of SiO<sub>2</sub>/SiO<sub>2</sub>+CoSb Multi-nano-layered Thin Films Modified by MeV Si Ions. *Journal of Intelligent Material Systems and Structures*, 24(11), 1350-1356.
- [4] Tasos, C. P., Georgios, C. G., & Andreas, N. A. (2000). *Viscous Fluid Flow*. CRC Press LLC.
- [5] Poularikas, A. D., & Seely, S. (2000). *Laplace Transforms*. CRC Press LLC.
- [6] Adomian, G. (1989). *Nonlinear stochastic Systems Theory and Applications to Physics*. Kluwer Academic.
- [7] Adomian, G. (1983). *Stochastic systems*. Academic Press, London.
- [8] Adomian, G. (1984). A new approach to non-linear partial differential equations. *Journal of Mathematical Analysis and Applications*, 102, 420-434.
- [9] Adomian, G. (1991). A Review of the decomposition method and some recent results for nonlinear equations. *Computers & Mathematics with Applications*, 21, 101-127.
- [10] Adomian, G. (1994). *Solving Frontier Problems of Physics: The decomposition Method*. Kulvar Academic, Boston.
- [11] Mobasher, A., Oviedo, R. R., & Vu, B. (2001). Application of Method of Decomposition to Classical Fluid Dynamics Problems. *Fifth Mississippi State Conference on Differential Equations & Computational Simulations*. Starkville, MS.
- [12] Mobasher, A. M., Biazar, A. P., & Zeng, Z. T. (2012). Piecewise n<sup>th</sup> order Adomian polynomial stiff differential equation solver. *International Journal of Modern Engineering*, 13, 13-17.
- [13] Burgers, J. M. (1948). *A Mathematical Model Illustrating the Theory of Turbulence*. *Advances in Applied Mechanics I*. Academic Press.
- [14] Adomian, G. (1987). Analysis of model equations of gas dynamics. *AIAA Journal*, 26, 242-244.
- [15] Adomian, G. (1998). Solutions of nonlinear PDE. *Applied Mathematics Letters*, 11, 121-123.
- [16] Kutluay, S., Bahadir, A. R., & Ozdes, A. (1999). Numerical solutions of one-dimensional Burgers equation: explicit and exact explicit finite difference methods. *Journal of computational and Applied Mathematics*, 103, 251-261.
- [17] Hopf, E. (1950). The partial differential equation  $u_t + u u_x = \mu u_{xx}$ , *Comm. Pure Applied Mathematics*, 3, 201-230.
- [18] Cole, J. D. (1951). On a quasi-linear parabolic equations occurring in aerodynamics. *Quart. Applied Mathematics*, 9, 225-236.
- [19] Benton, E., & Platzman, G. W. (1972). A table of solutions of one-dimensional Burgers equations. *Quart. Applied Mathematics*, 30, 195-212.

## Biographies

**AMIR A. MOBASHER** is an Associate Professor of Mechanical Engineering at Alabama A&M University. Dr. Mobasher may be reached at [amir.mobasher@aamu.edu](mailto:amir.mobasher@aamu.edu).

**MEBOUGNA DRABO** is an Assistant Professor of Mechanical Engineering at Alabama A&M University. Dr. Drabo may be reached at [mebougna.drabo@aamu.edu](mailto:mebougna.drabo@aamu.edu).

**SATILMIS BUDAK** is an Associate Professor of Electrical Engineering at Alabama A&M University. Dr. Budak may be reached at [satilmis.budak@aamu.edu](mailto:satilmis.budak@aamu.edu).

# MODELING AND ANALYSIS OF A HYDROGEN FUELING STATION NETWORK

Fang Clara Fang, University of Hartford; Neftali Torres, University of Texas at Austin

## Abstract

A geo-spatial analysis was applied in order to model the number of hydrogen stations and their placement on a regional scale in the state of Connecticut. In the study, the authors considered a number of unique geographic and socio-demographic factors of the region and associated fueling behaviors in order to predict the number of stations required. Those factors included traffic network, population, existing alternative fueling stations, and transit routes. There were approximately 248 stations identified in Connecticut, based on these high-density factors. The preliminary station arrangement was evaluated, based on the average vehicle travel times for refueling. The study found that the weighted average travel time for a user to reach a hydrogen station was 4.91 minutes. Vehicle travel times by county and population regions were also compiled. Lastly, station arrangement optimization was explored by attempting to minimize vehicle travel times for Connecticut. The study's results showed that the method for locating hydrogen stations decreased the driving time to a station in metropolitan/high-traffic flow regions, while preserving similar travel times within rural regions.

## Introduction

In our modern age, where energy and fuel are primary concerns, the idea of deploying alternative-fuel vehicles is slowly turning into a reality. When making a change to any alternative-fuel vehicle, a major obstacle is deciding on the distribution of the fuel. In the case of hydrogen, the infrastructure for refueling vehicles is almost non-existent. Several studies have explored building a hydrogen infrastructure with various degrees of success. Milbrandt [1] determined the location and number of hydrogen stations nationwide that would make hydrogen fueling available at regular intervals along commonly traveled interstate roads. An infrastructure design and cost analysis was also conducted. However, the search did not examine refueling travel time to ensure that customer convenience was maintained. In addition, the experiment was performed on a nationwide basis, so a focused study on individual states was still required. Melaina [2] succeeded in developing focused criteria for determining the required number of stations based on high-density factors including traffic flow, populations, etc. However, the methodology for station placement and analy-

sis of the network were not evaluated. Nicholas [3] developed a hydrogen station network for Sacramento county, based on the percentage of existing gasoline stations highlighted in earlier studies. Hydrogen stations were placed and arranged utilizing Geographic Information Systems (GIS) based on travel time. Though this experiment provided a solid foundation for station placement and analysis for optimization, factors such as population, traffic flow, and existing alternative refueling stations were not factored into the placement of hydrogen stations.

This current study attempted to combine the favorable aspects of the previous studies and enhance these aspects by appropriately modifying the criteria for the state of Connecticut. This project explored the method of placing refueling stations based upon spatial factors such as population density, traffic flow, existing alternative fueling stations, and transit routes. GIS analyses were used to develop a model that relates the location of stations to the number of hydrogen stations by considering these spatial factors and determining potential sites. The optimal arrangement of the stations was derived by minimizing a person's drive time to the nearest station for a given region.

## Methodology

Figure 1 presents a brief overview of the project methodology. This study was initiated with data collection and assembly from various public resources. Once the data were assembled, the proposed hydrogen stations were placed based on spatial factors such as traffic volumes, population, existing alternative refueling stations, and local transit routes. After a preliminary proposed refueling network was established, the station layout analysis was performed. The output of the analysis was the weighted average travel time from an origin to the proposed hydrogen station for the state of Connecticut. The weighted travel time was then minimized through an iterative process by rearranging the proposed station layout. Lastly, a final proposed station layout was determined.

For this project, GIS was utilized to develop a visual representation of the preliminary hydrogen fueling station network based on factors such as population densities, traffic volumes, existing alternative fueling stations, and transit routes. In addition, Network Analyst and StreetMap, two

functions/software programs in ArcGIS 9.3 [4], aided in the analysis of the preliminary hydrogen network for determining a final station arrangement that minimized the weighted travel time to each station.

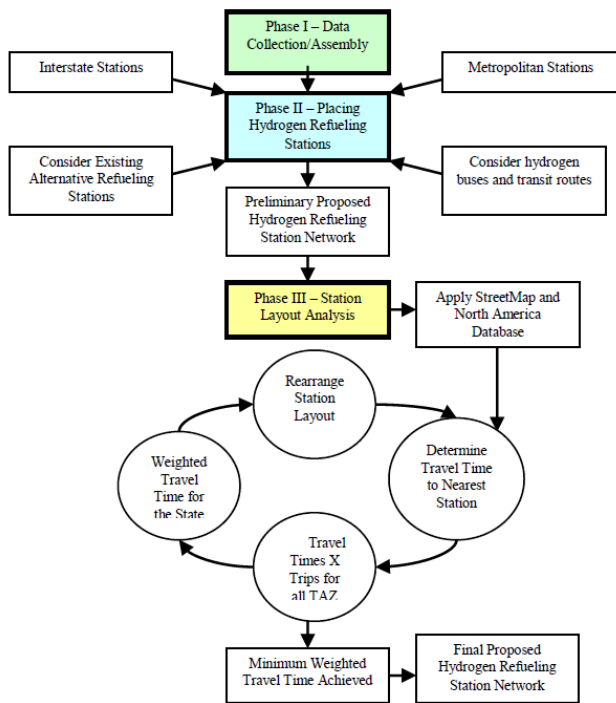


Figure 1. Overview of Project Methodology

## Phase I—Data Collection and Assembly

As a first step in developing a hydrogen network, the data—including existing alternative fueling stations, population densities, traffic volumes, and tract zones—were gathered and spatially assembled using GIS. Following is a brief discussion of the data collection process.

## Existing Hydrogen Production Facilities and Alternative Fueling Stations

Data on existing hydrogen production facilities and alternative fueling stations—including compressed natural gas (CNG), bio-diesel, electric, ethanol 85, and hydrogen refueling stations—were gathered from online websites [5], [6]. The datasets were manually processed in ArcGIS. The exact street addresses of these facilities and stations were identified, allowing for a more accurate placement, as presented in Figure 2. As a result, it was found that the state of Connecticut contained nine CNG, four ethanol 85, three electric, two hydrogen, one bio-diesel fueling stations, and three

hydrogen production facilities.

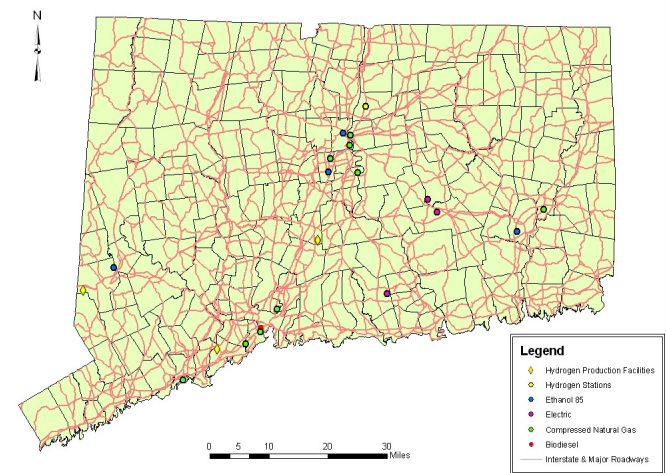


Figure 2. Hydrogen Production Facilities and Existing Alternative Fueling Stations in Connecticut

## Interstate Traffic Volumes

As shown in Figure 3, more than fifty percent of the interstate network experienced traffic volumes that were less than or equal to 10,000 vehicles per day (vehs./d), while the rest had an even distribution of traffic greater than 10,000 vehs./d.

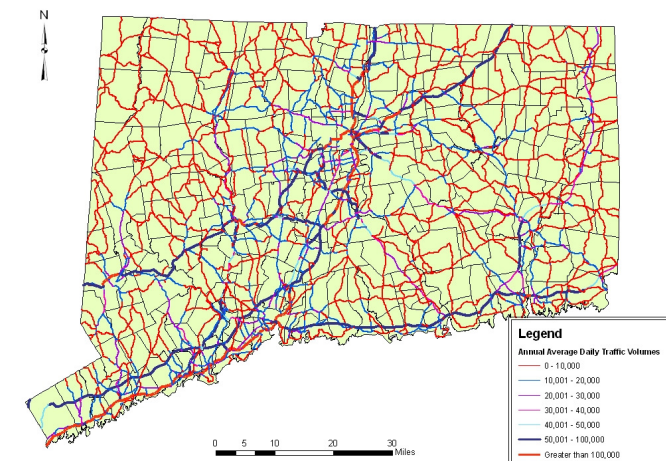


Figure 3. Annual Average Daily Traffic for Connecticut

## Population Densities

Population data were gathered from the UCCON Map and State GIS Geographic Center [7], [8]. The tabular population database was retrieved on population counts (from 1990) and land area (in square miles), sorted by town.



Based on these results, population densities were evaluated utilizing the normalization function in ArcGIS. Consequently, a map of the population densities was created (see Figure 4). The darker colored towns represent higher population densities, while lighter colored ones represent lower population densities. It should be noted that although the population counts were taken from 1990, it was assumed that towns with high population densities would remain densely populated.

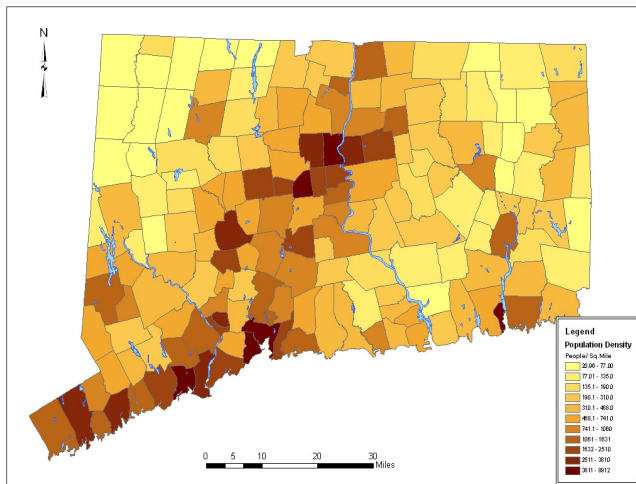


Figure 4. Population Densities for Connecticut

## TRACT Zones

Census 2000 tract data were gathered from the U.S. Census Bureau database [9]. The tract database provided tabular information on population, vehicles, and mean income within a tract zone. Each tract zone was spatially identified as a shape file in GIS. For this study, tract data were selected over traffic analysis zones (TAZs), due to the fact that there were 2,000 zones, which would be computationally challenging for Network Analyst. In addition, tract zones contained valuable information such as vehicle counts and average mean income that was otherwise unavailable in the TAZ data. A map of the census 2000 tract zones was developed based on eight counties in Connecticut, as illustrated in Figure 5. In total, there were 819 tract zones.

## Phase II—Developing a Preliminary Hydrogen Fueling Station Network

With a variety of data assembled, phase II of this project developed the initial hydrogen fueling station network based on factors such as population densities, traffic volumes, existing alternative fueling stations, and transit routes. After

extensively reviewing existing research and reports, this study followed some general guidelines in the development of a model for effective station placement, as highlighted in an earlier study by Melaina [2]. For example, stations were located in close proximity to high traffic volume, high profile areas with potential consumers, and along the interstate. In this study, the authors used these requirements and expanded the analysis by providing a procedure to geographically locate these stations in GIS to consider unique social and demographic characteristics of Connecticut.

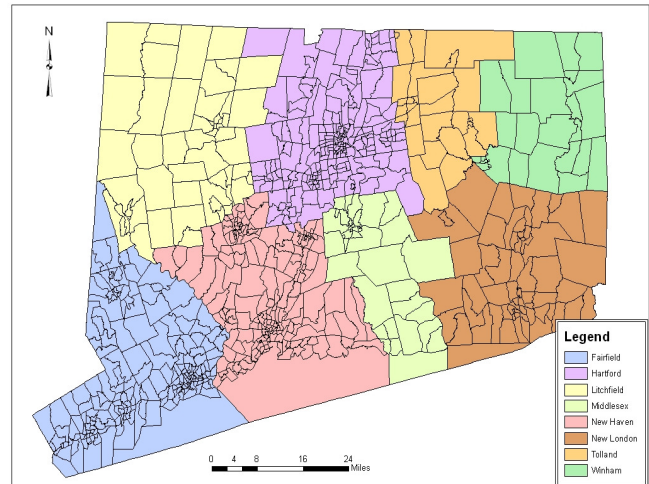


Figure 5. Tract Zones Grouped by County (2000)

## Placing Interstate Stations

Refueling stations must be placed along interstate roads to accommodate high traffic volumes, increase fuel accessibility to potential first fuel cell vehicle (FCV) buyers, and to provide fuel to vehicles during long trips. For this project, the interstate network considered both urban interstates and rural interstates. Urban interstates were identified as road networks that experienced volumes greater than or equal to 50,000 vehs./d—here, stations were placed at a maximum distance of 5 miles apart. Rural interstates exhibited volumes less than 50,000 vehs./d—in this case, stations were placed at a maximum distance of 10 miles apart.

Based on these specific guidelines and an average daily traffic volume distribution map (see Figure 3), a new map that located potential stations was generated by using simple selection queries and the divide function in GIS, as presented in Figure 6. The interstate segments colored in red are networks that experienced traffic volumes equal to or greater than 50,000 vehs./d. A total of 168 stations were placed along the interstates in Connecticut.

## Placing Metropolitan Stations

For the metropolitan approach, stations must be placed exclusively within metropolitan areas so that they are accessible to the first adopters of hydrogen vehicles and to accommodate for the large populations within the areas. This study classified metropolitan populations greater than or equal to 1,000 people per square mile (see Figure 7).

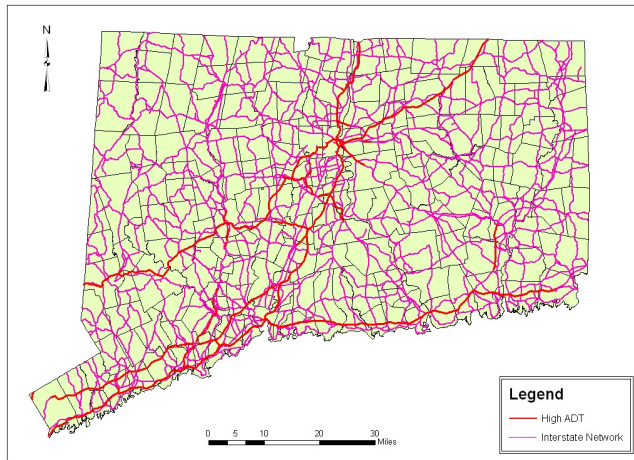


Figure 6. Interstate Traffic of Over 50,000 vehs./d

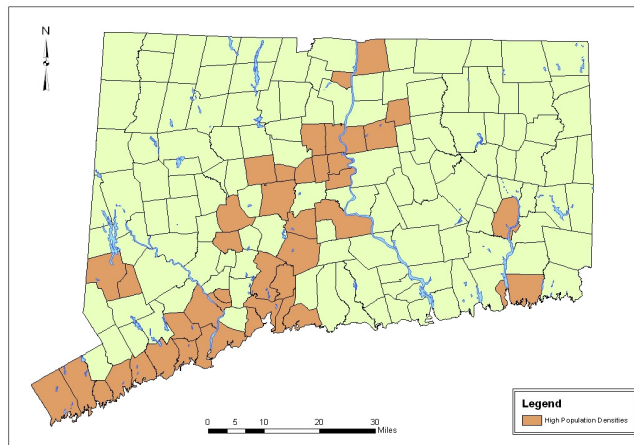


Figure 7. Towns with Population Densities Greater than 1000 People per Square Mile

Stations were placed so that residents are a short distance from their homes. The minimum number of stations required within any given metropolitan region is based on the region's area and the average maximum distance from any station (i.e., 3 miles), assuming the layout was being established for an early hydrogen network. A total of 48 stations were positioned within metropolitan regions.

## Considering Existing Alternative Fueling Stations

Existing alternative fueling stations are ideal for hydrogen station adoption due to the fact that vendors are already experienced in dealing with zoning and relevant regulations. In addition, these stations already serve customers on a regular basis. And, adding new hydrogen refueling stations may increase the probability of hydrogen energy conversion of existing customers. Thus, a total of 22 hydrogen stations were manually placed based on the existing station map (see Figure 2).

## Considering Hydrogen Buses and Transit Routes

Once standard hydrogen vehicles are widely commercialized, hydrogen buses are likely to follow. Currently, the Connecticut transit system has one fuel-cell bus that services the Hartford downtown region; there are future plans to adopt an entire fleet of hydrogen buses for the region. Based on possible hydrogen transit routes and schedules, 10 stations were assigned.

## Preliminary Hydrogen Station Network

Figure 8 illustrates the preliminary hydrogen fueling station distribution after evaluating all of the approaches described above. Figure 9 presents the preliminary hydrogen fueling station distribution by county.

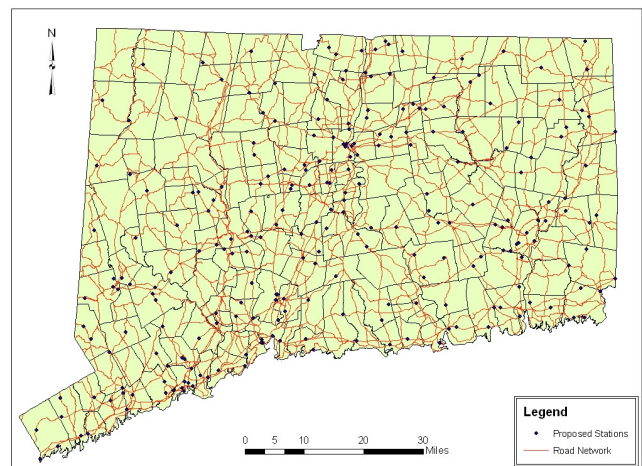


Figure 8. Preliminary Hydrogen Station Layout

A summary of station counts and related statistics for the preliminary hydrogen station network is provided in Table 1. There were 248 proposed stations in Connecticut. Hart-

ford acquired the largest number of stations (58), while Windham had the fewest stations (14). It was found that station placement in urban areas such as Hartford and New Haven generated more stations than in less-populated counties. This was expected based on the fact that metropolitan areas experience greater traffic flows and, therefore, require more refueling stations.

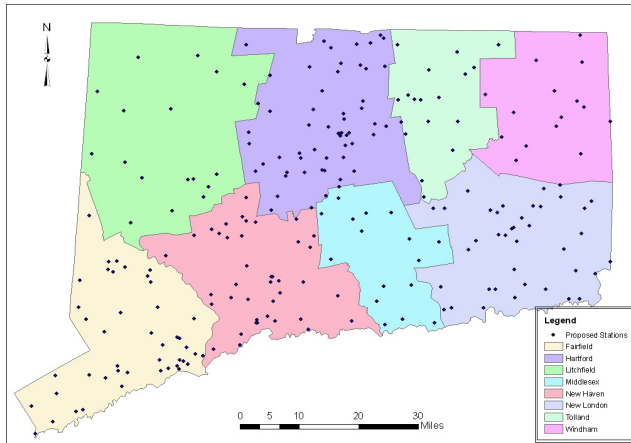


Figure 9. Preliminary Hydrogen Station Layout by County

Table 1. Preliminary Hydrogen Station Statistics by County

County	# of Stations	# of Towns	Stations / Town	Stations by Town	
				Min #	Max #
Fairfield	44	23	1.91	0	6
Hartford	58	29	2.00	0	7
Litchfield	17	26	0.65	0	3
Middlesex	17	15	1.13	0	3
New Haven	42	27	1.56	0	4
New London	38	21	1.81	0	5
Tolland	18	13	1.38	1	3
Windham	14	15	0.93	0	2
Connecticut	248	169	1.47	-	-

### Phase III–Station Layout Analysis

With the proposed hydrogen refueling network established, the optimal arrangement of stations was performed. Previous studies [3] found that a station’s proximity to a customer’s origin or destination was a significant influencing factor in station selection and usage. Therefore, the travel time from a user’s origin or destination to a refueling station was applied in this current study. Two ArcGIS 9.3

extensions were used: Network Analyst and StreetMap. Network Analyst allows the user to build, edit, and analyze network datasets, while StreetMap provides domestic and international routing and address geo-coding capabilities. These two extensions, in combination, provided the means for evaluating travel time from an origin to a proposed station and determining whether the station(s) could be arranged to lower average travel time.

### Developing Traffic Analysis Zones

This study manually combined the tract data (as shown in Figure 3) into groups of three or four within the same county using the GIS Dissolve function. The tract data were reduced to 275 TAZs, appropriate for computations during the station layout analysis. The new population data and vehicle counts within each TAZ were the result of a combined sum of the initial corresponding tract zones. The median income data were preserved by computing the average median incomes for the combined tract zones. Figure 10 displays the TAZs grouped by county.

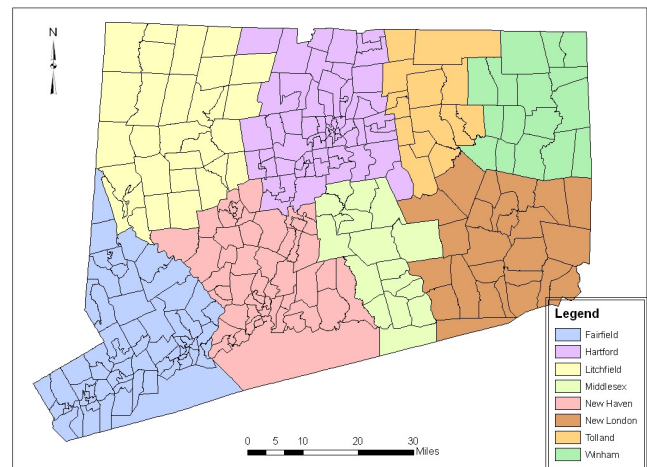


Figure 10. Traffic Analysis Zones Grouped by County

### Analysis of the Station Layout

The analysis assumed that all users within a TAZ would begin at the approximate center of the TAZ and were assigned to the station that output the lowest travel time. This process was used due to the fact that it would be computationally exhaustive if individual resident homes were considered as starting points. All residents within a TAZ were assumed to leave at the same time. As shown in Figure 11, the StreetMap function was then utilized to calculate travel time along the street to the nearest station. The application was utilized to determine the most efficient combination of minor roads and interstates that output the lowest travel



time. If there were multiple stations near the TAZ, then the same procedure was applied to each station until the optimal solution was achieved. If stations outputted the same travel time, the stations with the shortest distance were selected.

All data were manually recorded in tabular format using MS Excel, as presented in Figure 11. In this table, the “Origin TAZ” column represents the TAZ being examined, while the “Station ID” column displays the ID number of the selected station. The “Trips” column corresponds to the vehicle counts that were extracted from the original tract shape files for each TAZ. This column was used to calculate the final weighted travel time for the state. It was presumed that the TAZs generating higher trips would have a greater influence than the TAZs with fewer trips.

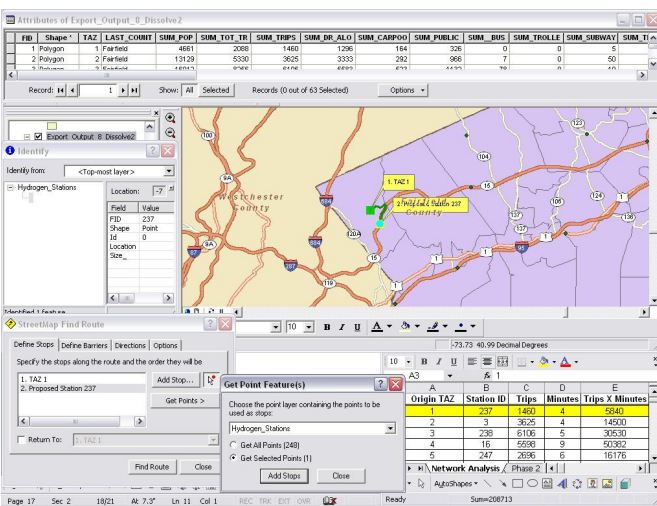


Figure 11. Station Layout Analysis

## Analysis of Results

Table 2 presents the results from the station layout analysis with respect to individual counties in Connecticut. They include the trips generated and the vehicle travel time statistics for each county and the entire state. Vehicle travel times represent the travel time from the center of a TAZ to the station that output the lowest travel time including the minimum, maximum, averaged, and weighted average values. As shown in Table 2, New Haven county drivers would experience the shortest vehicle travel time (4.44 minutes), while Litchfield adopters would drive an average of 8.24 minutes to a hydrogen station. This phenomenon was expected since a greater number of stations were placed in the metropolitan areas that met the high-density requirements, resulting in a reduced vehicle travel time. A similar pattern was documented throughout the rest of the counties. The weighted average travel time was 4.91 minutes for the entire

state. It was also found that the average and weighted vehicle travels times displayed a small difference. The lack of any difference between the average and weighted travel times was due to the fact that high-density factors were already considered during the initial placement of the hydrogen stations.

Table 2. Results of Network Analysis

County/ Region	Trips	Vehicle Travel Times (min)			
		Min	Max	Ave	Weighted
Fairfield	353318	1	11	4.87	4.78
Hartford	368587	1	14	4.52	4.49
Litchfield	86471	1	16	7.93	8.24
Middlesex	74334	1	10	4.88	4.68
New Haven	350766	1	11	4.39	4.44
New London	117972	1	11	5.29	5.31
Tolland	65037	1	12	5.21	4.91
Windham	50615	1	10	5.94	6.00
Connecticut	1467100	1	16	5.15	4.91

Although there was minimal difference between the weighted travel times across each county, the impact can be amplified through simple arithmetic. For instance, there was a 0.29 minute and 15,269 trip difference between the Fairfield and Hartford counties, which would save 4,428 minutes or 3 days’ worth of travel time. A noteworthy impact such as 3 days would save a significant amount of time, money, and reduced environmental impacts when studied on a yearly basis.

Table 3 summarizes the minimum vehicle travel times for representative TAZs when grouped into urban, sub-urban, and rural classifications. Thirty TAZs were randomly selected and arranged to fit one of the three sub-regions. The urban classification was identified as a TAZ with populations greater than 20,000 people, while a TAZ with less than 10,000 people was classified as a rural region. The TAZs that did meet the above requirements were known as sub-urban regions. As presented in Table 3, urban, sub-urban, and rural zones experienced weighted vehicle travel times of 4.50, 6.45, and 7.38 minutes, respectively. This slight difference in weighted travel times was expected, due to the greater number of stations placed within metropolitan and high-traffic-flow regions. Ultimately, shorter travel times in urban areas validated the criteria and assumptions established at the beginning of the experiment.

**Table 3. Refueling Travel Time (in minutes) for Urban, Sub-urban, and Rural Regions**

Region	TAZ ID	County	Population	Trips	Travel Time (min.)	Trips x Mins.	W'ght'd Travel Time (min.)
Urban (Pop > 20,000)	40	Fairfield	25765	6536	1	6536	4.50
	44	Fairfield	21305	9347	5	46735	
	49	Fairfield	20801	10616	3	31848	
	69	Hartford	22533	7972	3	23916	
	110	Hartford	23609	5395	2	10790	
	119	Hartford	23075	10655	4	42620	
	177	New Haven	20325	9958	3	29874	
	189	New Haven	23035	10987	3	32961	
	206	New Haven	28463	12356	10	123560	
	229	New London	24688	11475	7	80325	
Sub-urban (20,000 ≥ Pop. ≥ 10,000)	81	Hartford	14342	6217	13	80821	6.45
	82	Hartford	10347	5066	13	65858	
	84	Hartford	10346	4135	5	20675	
	159	Middlesex	10830	5425	5	27125	
	163	Middlesex	13094	6556	3	19668	
	167	Middlesex	10367	4523	2	9046	
	221	New London	12980	5887	7	41209	
	224	New London	11147	5280	4	21120	
	230	New London	11759	5213	2	10426	
253	Tolland	11056	2837	12	34044		
Rural (Pop < 10,000)	128	Litchfield	3977	1407	15	21105	7.38
	129	Litchfield	2968	1157	4	4628	
	130	Litchfield	1081	510	5	2550	
	232	New London	4991	2515	11	27665	
	243	Tolland	3036	1643	5	8215	
	250	Tolland	5959	3249	6	19494	
	252	Tolland	4971	2502	6	15012	
	264	Windham	1618	749	9	6741	
	265	Windham	3798	1652	1	1652	
	269	Windham	4692	2468	10	24680	

## Conclusions and Recommendations

This study expanded on earlier work to produce guidelines for placing and analyzing preliminary hydrogen refueling station networks. A more detailed and comprehensive dataset for hydrogen station placement was then developed exclusively for the state of Connecticut. A GIS-based station layout analysis was applied in order to minimize user travel time in a traffic analysis zone. The results showed that the hydrogen station setup provided reduced travel times for urban and high-density regions, while preserving similar vehicle travel times for all other regions. Ultimately, high-density factors considered at the beginning of the experiment validated the experimental results. The findings will establish a convenience for highly populated areas, which would better promote early hydrogen vehicle adoption. Furthermore, the outcome of this research may be improved if existing gasoline stations, average household income, and station capacity were examined. A cost analysis may also provide a method for optimizing station placement.

It must be emphasized that these driving times are estimates and may be greater if traffic congestion during rush hour and other relevant factors are considered. In addition, hydrogen station placement may be modified to correspond to the appropriate zoning restrictions and land availability. Even if station locations were slightly modified, it is believed that the adjustment will have little effect on the weighted travel times within a county or region. In addition, this study has also explored station layout rearrangement and optimization. Once a baseline travel time is established, the proposed station layout can be rearranged in order to minimize the weighted average travel time through an iterative process, as described in the Methodology section. However, the experiment showed that if a proposed station were merely placed at the center of the TAZ, it would generate a “flawed” analysis with very low travel time, making the optimization analysis useless. To eliminate this problem, two options can be pursued. One is to partition residential regions within the TAZ and acquire an average vehicle travel time for the TAZ. The other is to have the location of each TAZ. Although the analysis would remain very similar, guidelines on how to rearrange the stations would be required. It is also possible to assign stations to nearby existing gasoline stations and then rearrange the proposed stations. A C++ program or K means clustering is recommended to perform the analysis in a systematic manner in future projects.

In conclusion, the study provides direction on hydrogen station placement and serves as a tool for evaluating station placement decisions by achieving a good arrangement of them. It is anticipated that the results of this research will enhance investigations into a cleaner, more environmentally-friendly transportation system.

## Acknowledgement

This work was supported in part by the NASA Space Grant’s student research fellowship.

## References

- [1] Milbrandt, M. (2005). *Analysis of the Hydrogen Infrastructure Needed to Enable Commercial Introduction of Hydrogen-Fueled Vehicles*. National Renewal Energy Laboratory. Washington, DC.
- [2] Melaina, M. W. (2003). Initiating Hydrogen Infrastructures: Preliminary Analysis of a Sufficient Number of Initial Hydrogen Stations in the US. *International Journal of Hydrogen Energy*, 28, 743-755.



- 
- [3] Nicholas, M. A. (2004). *Hydrogen Station Siting and Refueling Analysis Using Geographic Information Systems: A Case Study of Sacramento County*. University of California Davis, Davis, CA.
  - [4] Environmental Systems Research Institute. (2009). CA
  - [5] Fuel Cells 2000. (2009). Breakthrough Technologies Institute, Washington, D.C. Retrieved from <http://Fuelcells.org>.
  - [6] Alternative Fueling Station Locator. (2009). U.S. Department of Energy. Retrieved from <http://www.afdc.energy.gov/afdc/locator/stations>.
  - [7] Connecticut Department of Transportation GIS Team. (2009). ADT 2003 Database. CTDOT, Newington, CT.
  - [8] UCCONN MAGIC. (2009). Retrieved from [http://magic.lib.uconn.edu/connecticut\\_data.htm](http://magic.lib.uconn.edu/connecticut_data.htm).
  - [9] Tract Data, U.S Census Bureau. (2009). Washington, D.C. Retrieved from <http://www.census.gov/geo/www/cob/tr2000.htm>.

## Biographies

**FANG CLARA FANG** is an Associate Professor at the University of Hartford where she teaches both undergraduate and graduate courses and conducts research in transport engineering. Fang holds a Ph.D. degree in transportation engineering from Pennsylvania State University, two Master degrees in civil engineering from the University of British Columbia, Canada, and the City University of Hong Kong, respectively, and a Bachelor of Science degree in civil engineering from Sichuan University, China. She has authored or co-authored more than forty peer-reviewed journal and conference articles and reports. Dr. Fang can be reached at [fang@hartford.edu](mailto:fang@hartford.edu).

**NEFTALI TORRES** received his Bachelor of Science degree in Civil Engineering from the University of Hartford in May, 2010. He was honored with the John G. Lee Medal, which is awarded to a graduating senior from the Greater Hartford area who has excelled academically and demonstrated a deep commitment to the community. Neftali received his Master of Engineering degree from the University of Texas at Austin in May, 2014, where he worked as a graduate research assistant for the Construction Industry Institute. Mr. Torres can be reached at [torres-nef@gmail.com](mailto:torres-nef@gmail.com).

# REFLECTION AND TRANSMISSION OF ELECTROMAGNETIC WAVES IN A NON-STATIONARY DIELECTRIC MEDIUM

Deepthi Madhavapeddy, University of Connecticut; Ravindra Thamma, Central Connecticut State University

## Abstract

In this study, the authors looked at the reflection and transmission of electromagnetic waves (EM) in a non-stationary dielectric medium, moving with a uniform velocity  $v$  and independent of time and position. Initially, the problem was considered for two cases: when the medium was moving parallel to the interface and then perpendicular. The focal point of the study was to examine the behavior of electromagnetic wave propagation—which is normally incident on a dielectric interface that is moving with constant velocity,  $v$ , perpendicular to the interface—and to study the changes in various physical properties like the reflection and transmission coefficients, angles, and frequencies compared to a stationary medium. Moving Target Indicator (MTI) radar is one of the applications that was used in this study.

## Introduction

The reflection and transmission of electromagnetic waves (EM) from a non-stationary dielectric medium has been a fundamental problem of interest for several years. Numerous authors have investigated this problem as early as in 1958. Pauli [1] and Sommerfeld [2] discussed in detail the frequency shift of the reflected wave by a moving mirror. In 1965, Tai [3] and Yeh [4] gave the first solutions for the problem of reflection and transmission of plane waves by a dielectric half-space in a vacuum moving parallel and perpendicular to the interface, respectively.

Huang [5] discussed the problem of reflection and transmission of electromagnetic waves when the dielectric is moving in any arbitrary direction. The solution for this problem has applications in the fields of optics, radio sciences, and astrophysics. It is also of considerable practical interest in many monitoring and control applications such as: radio communication with moving spaceships; identifying echoes from moving targets from stationary targets; Moving Target Indicator radar systems; plasma diagnostics; power amplification; and, laser Doppler anemometry for fluid flow studies.

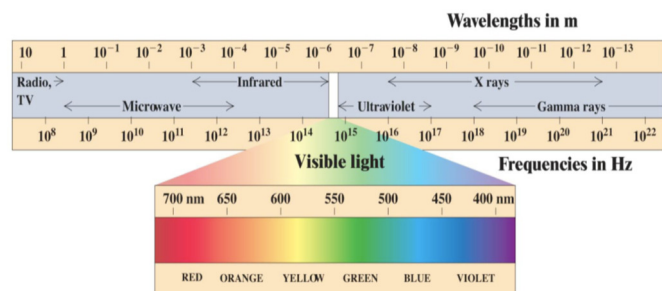
## Theory

Plane waves can be defined as EM waves in free space, which means that the electric and magnetic fields are confined to a plane and are uniform within the plane at all time. EM waves travel at the speed of light:

$$c = \frac{1}{\sqrt{\mu_0 \epsilon_0}} = 3 \times 10^8 \text{ m/s}$$

EM waves can be generated in different frequency bands like radio, microwave, infrared, visible, ultraviolet, x-rays, and gamma rays.

Figure 1 shows the electromagnetic spectrum, where the visible portion of the spectrum is relatively narrow. The boundaries between various bands of the spectrum are not sharp, but instead are somewhat arbitrary ( $1 \text{ nm} = 1 \times 10^{-9} \text{ m}$ ). When a light ray travels from one medium to another, part of the incident light is reflected and part is transmitted at the boundary between the two media. The transmitted part is said to be refracted in the second medium. Polarization plays an important role in wave propagation. The polarization of a uniform plane wave describes the shape and locus of the tip of the electric field vector,  $E$ , at a given point in space as a function of time. A wave is linearly polarized if  $E_x(z, t)$  and  $E_y(z, t)$  are in phase ( $d=0$ ) or out of phase ( $d=p$ ). For circular polarization, the magnitudes of the x- and y-components of  $E(z)$  are equal and the phase difference is  $d=\pm p/2$ . From the point of view of the target looking at the transmitter, left-hand circular polarization has a phase difference of  $d=p/2$  and right-hand circular polarization has a phase difference of  $d=-p/2$ .



Copyright © 2007 Pearson Education, Inc. publishing as Addison Wesley

**Figure 1. Electromagnetic Spectrum [6]**

Figure 2 shows the three wave polarizations for Linear, Circular, and Elliptical. If the vector that describes the electric field at a point in space as a function of time is always directed along a line, which is normal to the direction of propagation, the field is said to be linearly polarized. In general, however, the figure that an electrical field traces in an ellipse is said to be elliptically polarized. Linear and circular polarizations are special cases of elliptical and can be obtained when the ellipse becomes a straight line or a circle, respectively. The figure of the electric field is traced either clockwise (CW) or counterclockwise (CCW). Clockwise rotation of the electric field vector is also designated as right-hand polarization and counterclockwise as left-hand polarization.

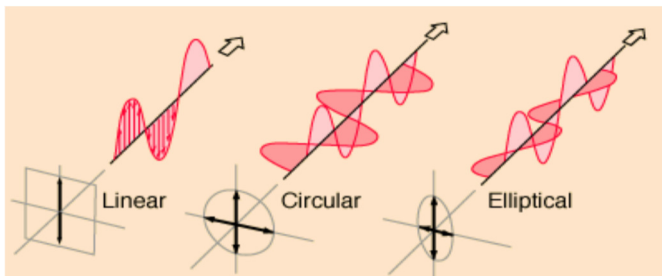


Figure 2. Variations of Electromagnetic Wave Polarizations [7]

In real-world problems, the fields encounter boundaries, scatterers, and other objects and, as a result, the fields must be found by taking into account these discontinuities. In general, the reflection and transmission coefficients are complex quantities and their amplitudes and phases can be varied by controlling the direction of the wave travel (angle of incidence). In face for one of the wave polarizations (parallel polarization), the reflection coefficient can be made equal to zero. When this occurs, the angle of incidence is known as the Brewster angle. This principle is used in the design of many instruments such as binoculars, etc. The magnitude of the reflection coefficient can also be forced to unity by selecting the wave incidence angle. This angle is known as the critical angle and is independent of wave polarization; however, in order for this angle to occur, the incident wave must exist in the denser medium. The critical angle concept plays an important role in the design of transmission lines, such as optical fiber, slab waveguides, and coated conductors—the micro strip is one example.

## Reflection and Transmission Coefficients for Stationery Medium Normal Incidence (Lossless Medium)

As shown in Figure 3, the wave is assumed to be traveling perpendicular (normal incidence) to the planar interface formed by the two semi-infinite lossless media [8]. When

the incident wave encounters the interface, a fraction of the wave intensity is reflected into medium 1, with another part being transmitted into medium 2.

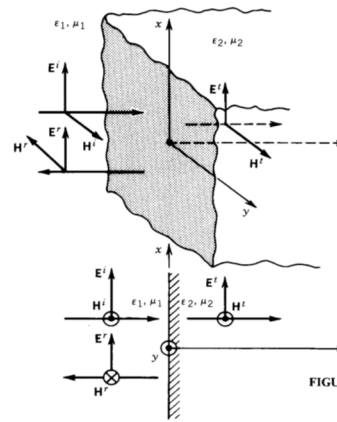


FIGURE 5-1 Wave reflection and transmission at normal incidence by a planar interface.

Figure 3. Wave Reflection and Transmission at Normal Incidence by a Planar Interface [8]

Assuming that the electric field is polarized in the  $x$  direction, expressions can be written for its incident, reflected, and transmitted electric field components, respectively, as:

$$E^i = \hat{a}_x E_0 e^{-j\beta_1 z} \quad (1)$$

$$E^r = \hat{a}_x \Gamma^b E_0 e^{+j\beta_1 z} \quad (2)$$

$$E^t = \hat{a}_x T^b E_0 e^{-j\beta_2 z} \quad (3)$$

where,  $E^i$  is the incident electric field,  $E^r$  is the reflected electric field,  $E^t$  is the transmitted electric field, and  $E_0$  is the incident electric field amplitude.

The terms  $\Gamma^b$  and  $T^b$  are used to represent, respectively, the reflection and transmission coefficients at the interface which are unknown and will be determined by applying the boundary conditions on the fields along the interface. Since the incident fields are linearly polarized and the reflecting surface is planar, the reflected and transmitted fields will also be linearly polarized. Using the right-hand rule, the magnetic field components can be written as:

$$H^i = \hat{a}_y \frac{E_0}{\eta_1} e^{-j\beta_1 z} \quad (4)$$

$$H^r = -\hat{a}_y \frac{\Gamma^b E_0}{\eta_1} e^{+j\beta_1 z} \quad (5)$$

$$H^t = \hat{a}_y \frac{T^b E_0}{\eta_2} e^{-j\beta_2 z} \quad (6)$$

where,  $H^i$  is the incident magnetic field,  $H^r$  is the reflected magnetic field,  $H^t$  is the transmitted magnetic field, and  $\eta_1, \eta_2$  are the wave impedances in medium 1 and medium 2, respectively. The reflection and transmission coefficients will now be determined by enforcing continuity of the tangential components of the electric and magnetic fields across the interface. The plane wave reflection and transmission coefficients of a planar interface for normal incidence are given by Equations (7) and (8):

$$\Gamma^b = \frac{\eta_2 - \eta_1}{\eta_2 + \eta_1} = \frac{E^r}{E^i} = -\frac{H^r}{H^i} \quad (7)$$

$$T^b = \frac{2\eta_2}{\eta_2 + \eta_1} = 1 + \Gamma^b = \frac{E^t}{E^i} = -\frac{\eta_2}{\eta_1} \quad (8)$$

The corresponding average power densities can be written as:

$$S_{av}^i = \frac{1}{2} \text{Re} (E^i X H^{i*}) = \hat{a}_z \frac{|E_0|^2}{2\eta_1} \quad (9)$$

$$\begin{aligned} S_{av}^r &= \frac{1}{2} \text{Re} (E^r X H^{r*}) \\ &= -\hat{a}_z |\Gamma^b|^2 \frac{|E_0|^2}{2\eta_1} \\ &= -\hat{a}_z |\Gamma^b|^2 S_{av}^i \end{aligned} \quad (10)$$

$$\begin{aligned} S_{av}^t &= \frac{1}{2} \text{Re} (E^t X H^{t*}) = -\hat{a}_z |T^b|^2 \frac{|E_0|^2}{2\eta_2} \\ &= -\hat{a}_z |T^b|^2 \frac{\eta_2}{\eta_1} \frac{|E_0|^2}{2\eta_1} = -\hat{a}_z |T^b|^2 \frac{\eta_2}{\eta_1} S_{av}^i \\ &= -\hat{a}_z (1 - |\Gamma^b|^2) S_{av}^i \end{aligned} \quad (11)$$

where,  $S_{av}^i$  is the incident average power,  $S_{av}^r$  is the average reflected power, and  $S_{av}^t$  is the average transmitted power densities. It is apparent that the ratio of reflected to incident power densities is equal to the square of the magnitude of the reflection coefficient. The reflection and transmission coefficients from the equations relate the reflected and transmitted field intensities to the incident field intensity. Since the total tangential components of these field intensities on either side must be continuous across the boundary,

it can be expected that the transmitted field can be greater than unity.

## Reflection and Transmission Coefficients for Stationery Medium Normal Incidence (Lossy Medium)

When a uniform plane wave is normally incident upon a planar interface formed by two lossy media, the incident, reflected, and transmitted fields, the reflection and transmission coefficients, and the average power densities are identical to the normal incidence for a lossless medium except that: a) attenuation constants must be included in each field, and b) the intrinsic impedances and attenuation and phase constants must be modified to include the conductivities of the media.

$$E^i = \hat{a}_x E_0 e^{-\alpha_1 z} e^{-j\beta_1 z} \quad (12)$$

$$E^r = \hat{a}_x \Gamma^b E_0 e^{-\alpha_1 z} e^{+j\beta_1 z} \quad (13)$$

$$E^t = \hat{a}_x T^b E_0 e^{-\alpha_2 z} e^{-j\beta_2 z} \quad (14)$$

$$H^i = \hat{a}_y \frac{E_0}{\eta_1} e^{-\alpha_1 z} e^{-j\beta_1 z} \quad (15)$$

$$H^r = -\hat{a}_y \frac{\Gamma^b E_0}{\eta_1} e^{+\alpha_1 z} e^{+j\beta_1 z} \quad (16)$$

$$H^t = \hat{a}_y \frac{T^b E_0}{\eta_2} e^{-\alpha_2 z} e^{-j\beta_2 z} \quad (17)$$

$$\Gamma^b = \frac{\eta_2 - \eta_1}{\eta_2 + \eta_1}; T^b = \frac{2\eta_2}{\eta_2 + \eta_1} \quad (18)$$

$$S_{av}^i = \hat{a}_z \frac{|E_0|^2}{2\eta_1} e^{-2\alpha_1 z} \text{Re} \left( \frac{1}{\eta_1^*} \right) \quad (19)$$

$$S_{av}^r = -\hat{a}_z |\Gamma^b|^2 \frac{|E_0|^2}{2\eta_1} e^{+2\alpha_1 z} \text{Re} \left( \frac{1}{\eta_1^*} \right) \quad (20)$$

$$S_{av}^t = -\hat{a}_z |T^b|^2 \frac{|E_0|^2}{2\eta_2} e^{-2\alpha_2 z} \text{Re} \left( \frac{1}{\eta_2^*} \right) \quad (21)$$

The total electric and the magnetic fields in medium 1 can be written as:

$$\begin{aligned} E^1 &= E^i + E^r \\ &= \hat{a}_x E_0 e^{-\alpha_1 z} e^{-j\beta_1 z} \left( 1 + \Gamma^b e^{+2\alpha_1 z} e^{+j2\beta_1 z} \right) \end{aligned} \quad (22)$$

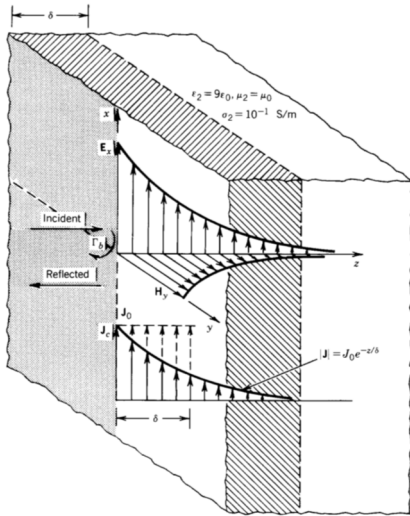


FIGURE 5-9 Electric and magnetic field intensities, and electric current density distributions in a lossy earth.

Figure 4. Electric and Magnetic Field Intensities and Electric Current Density Distribution in a Lossy Earth [9]

$$\begin{aligned} H^1 &= H^i + H^r \\ &= \hat{a}_y \frac{E_0}{\eta_1} e^{-\alpha_1 z} e^{-j\beta_1 z} \left( 1 - \Gamma^b e^{+2\alpha_1 z} e^{+j2\beta_1 z} \right) \end{aligned} \quad (23)$$

## Reflection and Transmission by a Medium Moving Perpendicular to the Interface (Normal Incidence)

Reflection and transmission by moving boundaries, such as reflection from a moving mirror, introduce Doppler shifts in the frequencies of the reflected and transmitted waves. Here, once can see the problem of normal incidence on a dielectric interface that is moving with constant velocity,  $v$ , perpendicularly to the interface; that is, along the  $z$  direction, as shown in Figure 5. The dielectric is assumed to be non-magnetic and lossless with permittivity. The left medium is free space,  $\epsilon_0$ . The electric field is assumed to be in the  $x$ -direction and, thus, the magnetic field will be in the  $y$ -direction. Two coordinate frames can be considered: the fixed frame,  $S$ , with coordinates  $\{t, x, y, z\}$ , and the moving frame,  $S'$  with  $\{t', x', y', z'\}$ . Of interest in this case is the determination of the Doppler-shifted frequencies of the re-

flected and transmitted waves, as well as the reflection and transmission coefficients, as measured in the fixed frame,  $S$ .

The procedure for solving this type of problem—originally suggested by Einstein in his 1905 special relativity paper—is to solve the reflection and transmission problem in the moving frame,  $S'$ , with respect to the boundary at rest, and then transform the results back to the fixed frame,  $S$ , using the Lorentz transformation properties of the fields. In the fixed frame,  $S$ , the fields to the left and right of the interface will have the forms:

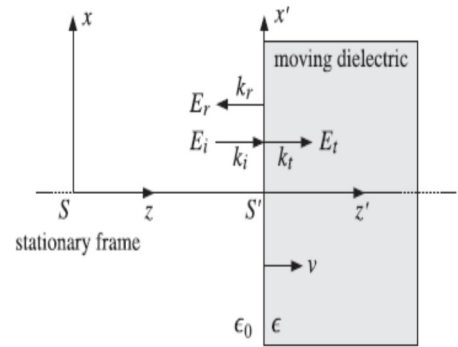


Figure 5. Reflection and Transmission at the Moving Boundary [10]

Left side:

Right side:

$$E_x = E_i e^{j(\omega t - k_z z)} + E_r e^{j(\omega_r t - k_r z)}$$

$$E_x = E_t e^{j(\omega t - k_z z)}$$

$$H_y = H_i e^{j(\omega t - k_z z)} - H_r e^{j(\omega_r t - k_r z)}$$

$$H_y = H_t e^{j(\omega t - k_z z)}$$

where,  $\omega$ ,  $\omega_r$ ,  $\omega_t$ , and  $k_i$ ,  $k_r$ ,  $k_t$  are the frequencies and wave numbers of the incident, reflected, and transmitted waves measured in  $S$ . In the frame,  $S'$ , where the dielectric is at rest, all three frequencies are the same and set equal to  $\omega'$ . This is a consequence of the usual tangential boundary conditions applied to the interface at rest. Note that  $\phi_r$  can be written as  $\phi_r = \omega_r t - (-k_r)z$ , implying that the reflected wave is propagating in the negative  $z$ -direction. In the rest frame,  $S'$ , of the boundary, the wave numbers are:

$$k'_i = \frac{\omega'}{c}, k'_r = \frac{\omega'}{c},$$

$$k'_t = \sqrt{\epsilon \mu_0} = n \frac{\omega'}{c}$$



where,  $c$  is the speed of light in a vacuum and  $n = \sqrt{\epsilon / \epsilon_0}$  is the refractive index of the dielectric at rest. The phase velocities of the incident, reflected, and transmitted waves are:

$$v_i = \frac{\omega}{k_i} = c, v_r = \frac{\omega_r}{k_r} = c,$$

$$v_t = \frac{\omega_t}{k_t} = c \frac{1 + \beta_n}{n + \beta}$$

In the rest-frame,  $S'$ , of the dielectric, the fields have the forms:

$$E'_x = E'_i (e^{j\phi'_t} + \rho e^{j\phi'_r}) \quad (24)$$

$$H'_y = \frac{1}{\eta_0} E'_i (e^{j\phi'_t} + \rho e^{j\phi'_r}) \quad (25)$$

$$E'_x = \tau E'_i e^{j\phi'_r} \quad (26)$$

$$H'_y = \frac{1}{\rho} \tau E'_i e^{j\phi'_r} \quad (27)$$

where,

$$\eta = \frac{\eta_0}{n}, \rho = \frac{\eta - \eta_0}{\eta + \eta_0} = \frac{1 - n}{1 + n},$$

$$\tau = 1 + \rho = \frac{2}{1 + n}$$

The corresponding reflection and transmission in the fixed frame,  $S$ , are given by Equation (28):

$$\frac{E_r}{E_i} = \rho \frac{1 - \beta}{1 + \beta},$$

$$\frac{E_t}{E_i} = \tau \frac{1 + \beta_n}{1 + \beta} \quad (28)$$

## Application

One of the major applications of EM waves is in radar systems. The word radar is derived from an acronym for radio detection and ranging. Radar is an electronic and electromagnetic system that uses radio waves to detect and locate objects. Radar operates by transmitting a particular kind of radio frequency waveform and detecting the nature of the reflected echo. When radio waves strike an object,

some portion is reflected, and some of this reflected energy is returned to the radar set, where it is detected. The location and other information about these reflective objects, or targets, can be determined by the reflected energy. Radio waves have many properties including speed, frequency, and power; they also display many phenomena such as refraction, attenuation, and reflection. The known speed and reradiation properties of radio waves are fundamental to the theory of radar operation, since radar systems are, at their core, transmitters and receivers of radio waves.

Radio waves occupy a portion of the electromagnetic spectrum from frequencies of a few kilohertz to a few gigahertz, which represents less than one-billionth of the total electromagnetic spectrum. In a vacuum, radio waves travel in straight lines. The speed of radio wave propagation in a vacuum ( $3 \times 10^8$  m/s) is a universal constant,  $c$ . The speed of wave propagation differs from  $c$  if the medium of propagation is matter. When electromagnetic waves travel in non-conducting materials such as air, the wave speed,  $v$ , is slower than  $c$  and is given by Equation (29):

$$v = \frac{1}{\sqrt{\epsilon \mu}} \quad (29)$$

where,  $\epsilon$  and  $\mu$  are the permittivity and permeability of the material, respectively. A measure derived from this wave speed change is the index of refraction,  $n$ , given by Equation (30):

$$n = \frac{c}{v_{\text{material}}} \quad (30)$$

A typical value of the index of refraction for air near the surface of the earth is 1.0003. A phenomenon called refraction occurs when radar waves pass through media with different indices of refraction. When a ray, the path of propagation of an EM wave, passes from a material having a smaller index of refraction to a material having a larger index of refraction, the ray is bent upwards. If the ray goes from a material having a larger index of refraction to a material with a smaller index of refraction, the ray is bent downwards. The index of refraction of the atmosphere is not constant and depends on temperature, air pressure, and humidity.

Since the velocity of propagation is inversely proportional to the index of refraction, radio waves move slightly more rapidly in the upper atmosphere than they do near the surface of the earth. From trigonometry, one can determine the radar horizon range, the maximum distance that radar can detect targets. The range is given by Equation (31):

$$r = \sqrt{2h_a k r_e} + \sqrt{2h_t k r_e} \quad (31)$$

where,  $k r_e$  is the effective radius and  $h_a$  and  $h_t$  are the heights of the transmitter and target, respectively. Anomalous propagation, or ducting of electromagnetic waves, occurs when the effective radius,  $k r_e$ , is greater than 2. This occurs when the gradient of the index of refraction of the atmosphere,  $dn/dh$ , is very large. EM waves are traveling waves that transport energy from one region to another. The energy transfer can be described in terms of power transferred per unit area, for an area perpendicular to the direction of wave travel. In a vacuum, the energy flow per unit time per unit area,  $S$ , is given by Equation (32):

$$S = \frac{EB}{\mu_0} \quad (32)$$

where,  $E$  and  $B$  are the electric and magnetic field magnitudes, respectively, and  $\mu_0$  is the permeability of free space. This can also be defined as the Poynting vector, and given by Equation (33):

$$\vec{S} = \frac{1}{\mu_0} \vec{E} \times \vec{B} \quad (33)$$

The total power out of any closed surface is then given by Equation (34):

$$P = \oint \vec{S} \cdot d\vec{A} \quad (34)$$

The atmosphere can cause power to attenuate, depending on range and frequency. Attenuation peaks at certain frequencies due to absorption by atmospheric gases, such as water vapor at 22.24 GHz and oxygen at 60 GHz. Heavy rainfall also attenuates radar waves increasingly at higher frequencies.

## Doppler Effect and Moving Target Indication (MTI) Filtering

There are two types of basic radar: Pulse transmission and Continuous wave. The former transmits a series of pulses separated by non-transmission intervals, during which the radar “listens” for a return and the later is constantly emitting radar. Relative motion of either the radar or the target is required to indicate target position or frequency shift. The frequency and wavelength of EM waves are affected by relative motion; this is known as the Doppler effect. Only the radial (approaching or receding) component of motion produces this phenomenon. If the source of an EM wave is approaching an observer, the frequency increases and the wavelength decrease. If the source is receding, the frequency decreases and the wavelength increase. An object's mo-

tion causes a wavelength shift,  $\Delta\lambda$ , that depends on the speed and direction the object is moving. The amount of the shift depends on the source's speed, and is given by Equation (35):

$$\Delta\lambda = \frac{\lambda_{rest} v_{radial}}{c} \quad (35)$$

where,  $c$  is the speed of light (or wave propagation speed, if in a material);  $\lambda_{rest}$  is the wavelength that would be measured if the source was at rest; and,  $v_{radial}$  is the speed of the source moving along the line of sight. The  $v_{radial}$  term only refers to the component of velocity along the line of sight. If the target moves at an angle with respect to the line of sight, then the Doppler shift ( $\Delta\lambda$ ) only informs about the part of motion along the line of sight.

Radars use the Doppler effect to determine a target's velocity. Many radars use Doppler information only for determining target motion direction, whereas others determine speed of motion more precisely. Radars transmit radio waves of a set wavelength (or frequency), which then reflects off the target. In this case, the target acts as a source, and its speed can be determined from the difference in the wavelength (or frequency) of the transmitted and reflected beams. Unwanted radio wave reflections in radar systems are called clutter. Since clutter is unwanted, radars try to eliminate it using signal and data processing techniques.

Using the Doppler effect's result, that stationary objects will not have frequency shifts, the transmitted waveform frequency is filtered out of the received signal. As a result of this frequency domain filtering, only moving targets will remain. This process is called moving target indication filtering or MTI.

The purpose of MTI radar is to reject returns from fixed or slow-moving, unwanted targets such as buildings, hills, trees, sea, and rain, and retain for detection or display signals from moving targets such as aircraft. MTI radar utilizes the Doppler shift imparted on the reflected signal by a moving target to distinguish moving targets from fixed targets. A Doppler shift allows one to distinguish between target and transmitter leakage. The amount of Doppler shift is determined by the radial velocity of the target, since the radial velocity is the apparent speed that the target is closing on or going away from the radar. A target can move in any direction and across a wide range of speeds; therefore, the radial velocity can change considerably.

## Radar Range Equation

The radar range equation relates the range of a radar system to the characteristics of its transmitter, receiver, anten-

na, target, and environment. In the simplest form of the equation, propagation effects discussed earlier are ignored. If the power of the radar transmitter,  $P_t$ , is transmitted through an isotropic antenna, which radiates uniformly in all directions, the power density at a distance  $R$  from the radar will be given by Equation (36):

$$\text{power density} = \frac{P_t}{4\pi R^2} \quad (36)$$

Radars generally use directional antennas, which channel the transmitter power in a particular direction. The gain,  $G$ , of an antenna is the increased power in the direction of the target, as compared to an isotropic antenna. When a directional antenna is used, the power density at distance  $R$  from the radar becomes:

$$\text{density power of signal transmitted at target} = \frac{P_t G}{4\pi R^2} \quad (37)$$

Radar Cross Section (RCS),  $A_o$ , is a measure of the electromagnetic energy intercepted and reradiated at the same frequency by an object. To determine the RCS of an object, the reradiation properties are compared to an idealized object that is large, perfectly conducting, and reradiates isotropically. An example of this is a large copper sphere, whose RCS is given by Equation (38):

$$\text{RCS of copper sphere} = \pi r^2 \quad (38)$$

where,  $r$  is the radius of the sphere. When the transmitted pulse reaches the target, it is reradiated in all directions, so the power density of the echo signal at the radar is:

$$\text{power density of echo at radar} = \frac{P_t G A_o}{4\pi R^2 4\pi R^2} \quad (39)$$

The radar antenna only captures a portion of the echo power. If the effective area of the radar antenna is denoted  $A_e$ , the power received by the radar,  $P_r$ , is given by Equation (40):

$$P_r = \frac{P_t G A_o A_e}{4\pi R^2 4\pi R^2} = \frac{P_t G A_o A_e}{(4\pi)^2 R^4} \quad (40)$$

The maximum radar range,  $R_{max}$ , beyond which the radar cannot detect targets occurs when the received echo signal is the minimum detectable signal,  $P_{min}$ . Hence,

$$R_{max} = \sqrt[4]{\frac{P_t G A_o A_e}{(4\pi)^2 P_{min}}} \quad (41)$$

## Conclusion

In this paper, the authors summarized the operation of radar systems and the how electromagnetic waves are used to transmit and receive the desired information from various

targets. Also described was how the Doppler effect plays a vital role in practical radar systems determining the target velocity in Moving Target Indicator (MTI) radar systems. It was also shown that the Earth's atmosphere plays a central role in radar operation, as it is the medium of propagation for the radio waveforms. The focus of this study was to understand how EM waves travel when the medium is moving with a velocity perpendicular to the interface and compare physical properties like the reflection and transmission coefficients, angles, and frequencies of both the moving and the stationary mediums.

## References

- [1] Pauli, W. (1958). *Theory of Relativity*. Pergamon, New York.
- [2] Sommerfeld, A. (2<sup>nd</sup> ed.). Akademische, Leipzig, Germany.
- [3] Tai, C. T. (1965). URSI Spring Meeting. Washington, DC.
- [4] Yeh, C. (1965). *Journal of Applied Physics*, 36, 3513.
- [5] Huang, Y.-X. (1994). Reflection and transmission of electromagnetic waves by a dielectric medium moving in an arbitrary direction. *Journal of Applied Physics*, 76, 2575.
- [6] Darling, D. The Worlds of David Darling. Retrieved April 5, 2011, from <http://www.daviddarling.info/encyclopedia/E/emspec.html>.
- [7] HyperPhysics. Retrieved April 5, 2011, from <http://hyperphysics.phyastr.gsu.edu/hbase/phyopt/polclas.html>.
- [8] Orfanidis, S. J. (2004-2010). *Electromagnetic Waves and Antennas*. Online book.
- [9] Introduction to Naval Weapons Engineering. Retrieved April 5, 2011, from <http://www.fas.org/man/dod-101/navy/docs/es310/radarsys/radarsys.htm>.
- [10] The Radar Handbook. (1990). (M. I. Skolnik, Ed.). McGraw-Hill.
- [11] Balanis, C. A. (2012). *Advance Engineering Electromagnetics*. (2<sup>nd</sup> ed.). Wiley.

## Biographies

**DEEPTHI MADHAVAPEDDY** is a graduate student at University of Connecticut. Mr. Madhavapeddy can be reached at dem09011@engr.uconn.edu.

**RAVINDRA THAMMA** is with the Department of Manufacturing and Construction Management, Central Connecticut State University. Dr. Thamma can be reached at thamarav@ccsu.edu.

# COMPUTATIONAL ASPECTS OF MODELING FLOW BOUNDARY LAYERS

Sabah Tamimi, Al Ghurair University, UAE

## Abstract

A wall-element-technique-based finite element method (FEM) was adopted in the near wall zone (N.W.Z.) for predicting the distribution of the pertinent variables in a smooth straight channel. The validation of the imposed technique, which used one-dimensional elements normal to the wall, was tested with two different types of flows: pressure flow and the combination of pressure and Couette flow. Both types were examined for turbulent, fully developed flow, when confined turbulent flow with a one-equation model was used to depict the turbulent viscosity. Also, the validity of the imposed technique was examined for the case in which the N.W.Z. was extended away from the solid wall. The imposed technique was tested and compared with other techniques with respect to turbulent flow.

## Introduction

One of the important applications of Engineering and applied Computer Science disciplines is computational fluid dynamics (CFD), which has become the focus of many studies. It is well known that when a fluid enters a prismatic duct, the values of the pertinent variables change from some initial profile to a fully developed form, which is thereafter invariant in the downstream direction. Numerous theoretical and experimental studies are available on laminar flow [1], [2], but few on turbulent flow. Since it has not been possible to obtain exact analytical solutions to such flows, an accurate numerical approach is needed. Therefore, an effective technique is required to model the variation of the pertinent variables near the solid boundary, where the variation in velocity and kinetic energy, in particular, is extremely large since the transfer of shear forms the boundary into the main domain and the nature of the flow changes rapidly. Consequently, if a conventional finite element is used to model the N.W.Z., a significant grid refinement would be required. Several solution techniques have been suggested in order to avoid such excessive refinement [3-5].

A more common approach is to terminate the main domain subject to discretization at some small distance away from the wall, where the gradients of the independent variables are relatively small, and then use another technique to model the flow behavior in the near wall zone (N.W.Z.). In previous work, the validity of the adopted wall element

technique, based on the use of the finite element in one direction normal to the solid wall, was tested and applied successfully to fully developed flow and heat transfer [6].

In this current study, three stages were considered. In the first stage, the validity of the adopted technique was tested under a combination of pressure and Couette flow for fully developed turbulent flow. In the second stage, the validity of the imposed technique was examined when the N.W.Z. was extended away from the fixed solid wall. In the third and final stage, the development of turbulent flow using parabolic elements was examined in one direction normal to the wall in a straight channel with fixed walls.

## Mathematical Model

The Navier-Stokes (N-S) equations associated with steady—state incompressible two-dimensional turbulent flow of a Newtonian viscous fluid with no body forces acting are:

$$\rho u_j \frac{\partial u_i}{\partial u_j} = -\frac{\partial p}{\partial x_i} + \frac{\partial}{\partial x_j} \left[ \mu_e \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) \right] \quad (1)$$

where,  $i, j = 1, 2$ ;  $u_i$  and  $p$  are the time—averaged velocities and pressure, respectively;  $\rho$  is the fluid density;  $\mu_e$  is the effective viscosity, which is given by  $\mu_e = \mu + \mu_t$ ; and,  $\mu$  and  $\mu_t$  are the molecular viscosity and turbulent viscosity, respectively. The flow field must satisfy the continuity equation, which may be written as:

$$\frac{\partial u_i}{\partial x_i} = 0 \quad (2)$$

Equations (1) and (2) cannot be solved unless a turbulence closure model can be provided to evaluate the turbulent contribution to  $\mu_e$ . In this present work, a one-equation model was adopted:

$$\mu_t = C_\mu \rho k^{1/2} l_\mu \quad (3)$$

where,  $l_\mu$  is the length scale of 0.4 times the normal distance from the nearest wall surface and  $C_\mu$  is a constant. The distribution of the turbulence kinetic energy,  $k$ , [7], [8] can be evaluated by a transport equation given by:

$$\rho \mu_j \frac{\partial k}{\partial x_j} = \frac{\partial}{\partial x_j} \left[ \left( \mu + \frac{\mu_t}{\sigma_k} \right) \frac{\partial k}{\partial x_j} \right] + \mu_t \frac{\partial u_i}{\partial x_j} \left[ \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right] - E \quad (4)$$

where,  $E = C_D \rho k^{3/2} / l_\mu$ ;  $\mu_t / \sigma_k$  is the turbulent diffusion coefficient;  $\sigma_k$  is the turbulent prandtl or Schmidt number; and,  $C_D$  is a constant.

The turbulence model, based on Equations (1), (2), and (4), is called the one-equation (k-1) model. These governing equations have been discretized by using the standard finite element method (FEM) [9-11] and the Galerking weight-residual approach was adopted in order to solve the discretized equations governing fluid motion. For purposes of this study, the main domain was divided into quadratic, 8-noded elements in order to define the variations in velocities and turbulent kinetic energy, while the pressure was predicted using the mixed-interpolation technique developed by Taylor and Hood [12]. This means that linear 4-noded elements should be used for pressure. Green's theorem is used, then, to reduce the order of the equations to unity, resulting in a 'weak formulation' or non-linear equation matrix, which is solved by either a coupled or an uncoupled method. Within the N.W.Z., the universal laws concept [13], or one-dimensional parabolic elements normal to the wall, was used.

## Boundary Conditions

Two types of turbulent flow were considered: pressure and pressure plus Couette flow. At the upstream, constant values were imposed for developing turbulent flow and compatible fully developed values were imposed for fully developed turbulent flow which were updated until a converged condition was satisfied. No slip conditions were imposed on the solid boundaries and tractions were updated downstream. Tractions are given by Equations (5) and (6):

$$\tau_{x_1} = -p + \frac{\mu_e}{\rho} \left( \frac{\partial u_1}{\partial x_1} \right) \quad x_1 \text{---parallel to walls} \quad (5)$$

$$\tau_{x_2} = \frac{\mu_e}{\rho} \left( \frac{\partial u_2}{\partial x_1} + \frac{\partial u_1}{\partial x_2} \right) \quad x_2 \text{---normal to walls} \quad (6)$$

Within the main domain, conventional two-dimensional isoperimetric elements were used to discretize the flow domain using a symmetric two-dimensional mesh consisting of 56 elements and total of 199 nodes, or 63 elements and 222 nodes when the mesh was extended up to the wall. Within the N.W.Z., either the universal laws concept or conventional finite elements (i.e., 2-D elements up to the

wall) were used (see Figure 1).

In this current study, a wall element technique based on an FEM was adopted using one-dimensional, 3-noded elements normal to the wall, as shown in Figure 2.

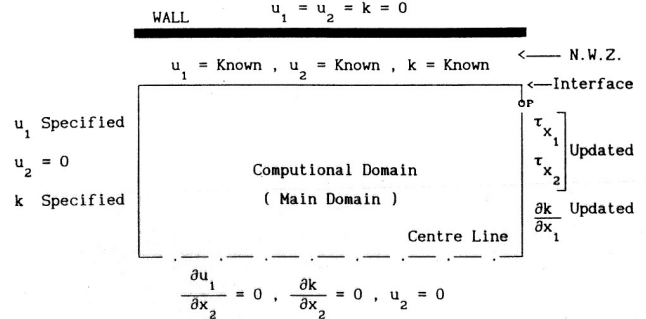


Figure 1. Boundary Conditions When the Mesh is Terminated at Short Distance from the Wall

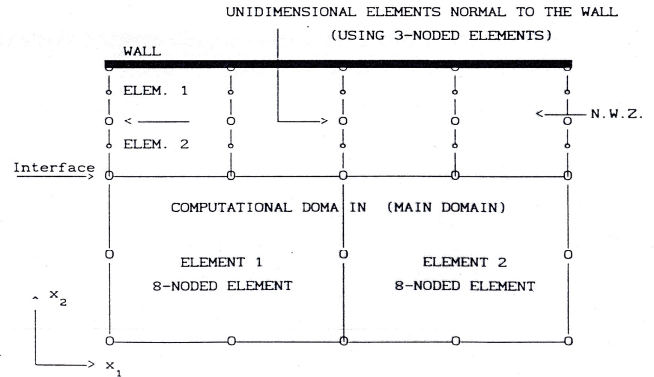


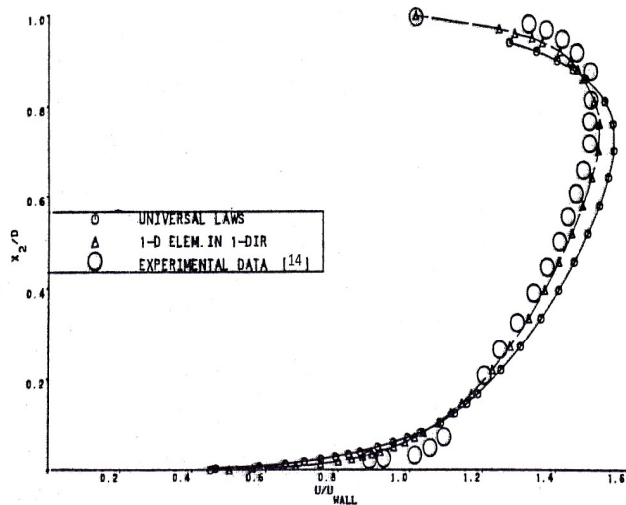
Figure 2. One-dimensional Elements in One Direction Normal to the Wall used in the N.W.Z.

## Results and Discussion

Turbulent flow was considered in a parallel-sided duct of width  $D$ , which is taken as 1.0 and length  $L$ . Different Reynolds numbers based on the width of the channel of 12,000, 50,000, and 70,000 were considered. In studies [6], [14], the validity of the adopted wall element technique (1-D elements in one direction normal to the wall) was tested for fully developed flow and comparisons were made with other accepted techniques and experimental results. It is generally agreed that the use of the universal laws is not acceptable and the general use of 2-D elements up to the wall is not economically viable since it needs excessive refinement which is very costly in computer time and memory size. Therefore, a wall element technique based on the use of FEM was adopted successfully; since each 1-D string of elements was analyzed individually, this saved

computer memory and time required and can replace other methods. For example, converged solutions were obtained in 71 sec when 1-D elements in one direction were employed, and 193 sec when universal laws were employed. However, a considerable saving in both can result if a comparison is made with the analysis when 2-D elements are used up to wall. In this study, three stages were considered. In the first stage, further calculations were concerned with an analysis of combining pressure and Couette flow, while the lower surface was stationary and the upper surface was moving at a constant speed when 1-D elements in one direction normal to the wall were used in the N.W.Z.

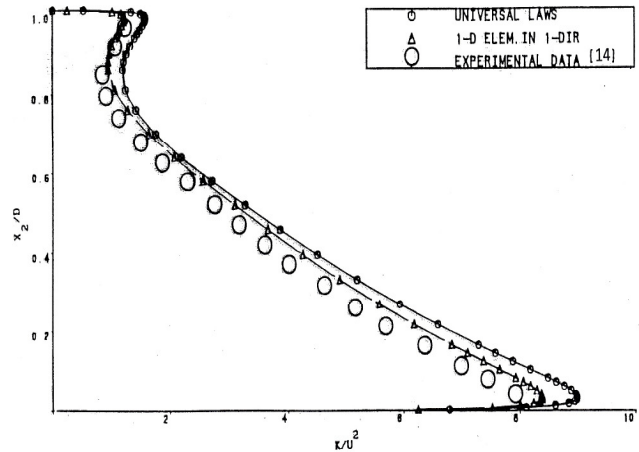
Fully developed turbulent velocity profiles and turbulent kinetic energy distributions with a fixed lower surface and a moving upper surface for a Reynolds number of 70,000 were obtained and are presented in Figures 3 and 4, respectively. Again, the results obtained from the adopted technique using 1-D elements normal to the wall were found to be superior to those obtained by using universal laws, since these laws have been derived in conjunction with empirical data and are applicable within a very small zone near the wall and are only considered valid for certain one-dimensional flow regimes. The results further showed excellent agreement between the imposed technique and experimental results [14] when turbulent fully developed flow was considered.



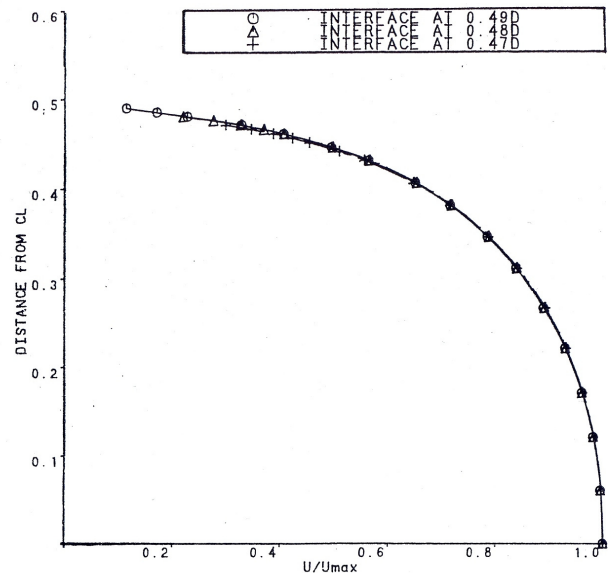
**Figure 3. Turbulent Fully Developed Velocity Profiles for Fixed Lower Surface and Moving Upper Surface: Re=70,000**

The second stage was concerned with the validation of the wall element technique in an extended N.W.Z. with the interface located at 0.48D and 0.47D from the symmetric line in a straight channel with fixed walls. Figures 5-7 show the downstream velocity, kinetic energy, and shear stress when

the N.W.Z. is extended up to 0.47D. Obviously, the results obtained from the adoption of 1-D elements in one direction are still the most advantageous, owing to the number of elements used in the N.W.Z. Also, the location of the N.W.Z. limit does not seem to affect the values of the pertinent variables. This is a distinct advantage over the universal law approach where strict limits must be placed on the location of the interface. Figure 8 indicates excellent agreement between the imposed technique and experimental results [14] for fully developed flow.

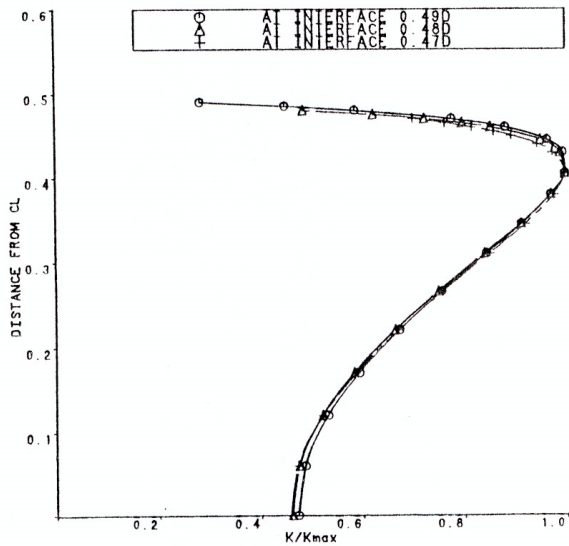


**Figure 4. Turbulent Fully Developed Kinetic Energy Profiles for Fixed Lower Surface and Moving Upper Surface: Re=70,000**

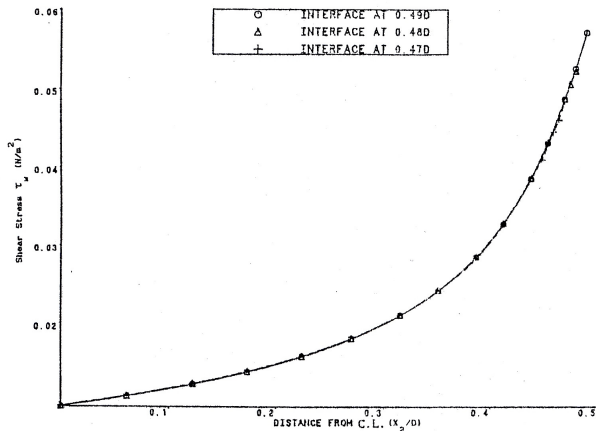


**Figure 5. Fully Developed Velocity Profiles for Turbulent Flow when the N.W.Z. is Extended up to 0.47D**





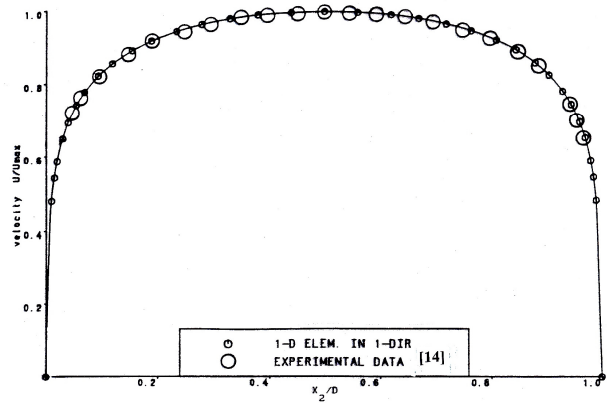
**Figure 6. Downstream Fully Developed Kinetic Energy Profiles when the N.W.Z. is Extended up 0.47D**



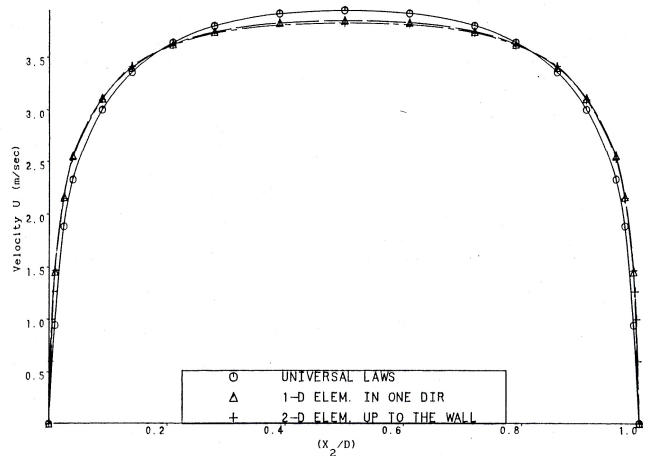
**Figure 7. Downstream Shear Stress Distributions for Fully Developed Flow when the N.W.Z. is Extended up 0.47D**

The final stage was concerned with developing turbulent flow with the wall element technique in one direction normal to the wall used in the N.W.Z. in a straight channel with fixed walls. Convergent velocity profiles at the outlet are shown in Figure 9. These clearly demonstrate that the results obtained from the adopted wall element technique (1-D elements in one direction) and complete mapping (2-D up to the wall) is not identical and different from those obtained when using the universal laws.

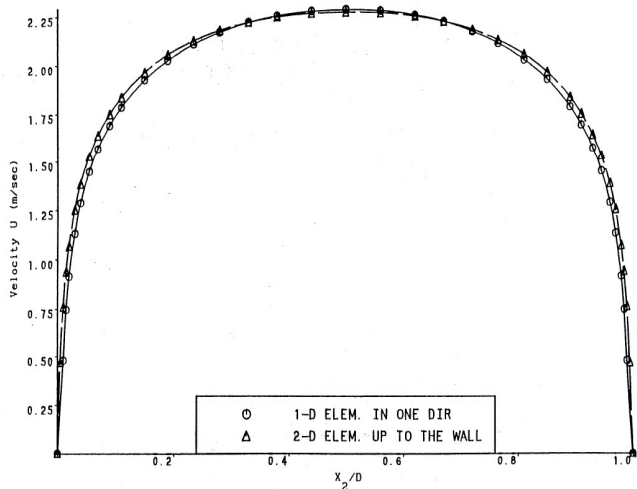
The velocity and the kinetic energy profiles presented in Figures 10 and 11 indicate that for this case the correspondence between the results obtained—when complete mapping and 1-D elements in one direction are used—are not as good as that experienced for fully developed flow. These figures clearly show that the use of 1-D in one direction is not good when compared to use of the complete mapping and 1-D elements in one direction normal to the wall, and clearly not valid for developing flow. The results obtained in Table 1 clearly show that the adopted technique is excellent and can be used with confidence for fully developed turbulent flow but not for developing flow, and the universal laws are not valid and should not be used.



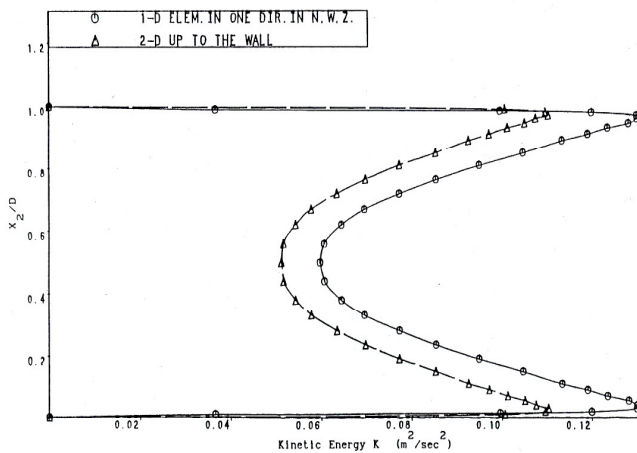
**Figure 8. Turbulent Velocity Profiles for Fully Developed Flow at 10D Downstream:  $L=10D$ , along with Experimental Results,  $Re=50,000$**



**Figure 9. Turbulent Velocity Profiles for Developing Flow at 10D Downstream:  $L=10D$ ,  $Re=12,000$**



**Figure 10. Turbulent Velocity Profiles for Developing Flow at 10D Downstream: L=10D, Re=12,000**



**Figure 11. Developing Kinetic Energy Profiles for Turbulent Flow at 10D downstream: L=10D, Re=12,000**

**Table 1. Results of the Simulation Model for Turbulent Flow**

Technique/ Method used in N.W.Z.	Fully developed Flow	Developing Flow
Universal Laws	1	1
Full mesh 2-D up to the wall	2	2
Wall element technique (using 1-D element in 1-dir. normal to the wall)	3	1

1: Poor (Not valid to be used); 2: Average (Valid to be used but not preferable); 3: Excellent (The best to be used).

## Conclusion

First, the utilization of empirical universal laws is not valid for developing and fully developed flow, since these laws are only really applicable for certain one-dimensional flow regimes. Second, the general use of 2-D elements up to the wall is not economically viable. Therefore, in order to avoid such an excessive refinement, these methods were replaced by introducing a wall element technique, based on the use of the FEM.

Third, the use of the wall element technique 1-D in one direction normal to the wall was applied successfully to both types of flow pressure and combination of pressure and Couette, and proved to be superior to other techniques. Finally, the validation of the wall element technique in an extended N.W.Z. was shown to have more advantages when compared to the use of universal laws. Therefore, this technique can be used to replace other techniques and can be used with confidence for fully developed turbulent flow, but not for developing flow.

## References

- [1] Hawken, D. M., Tamaddon-Jahromi, H. R., Townsend, P., & Webster, M. F. (1990). A Taylor-Galerkin Based Algorithm for Viscous Incompressible Flow. *Proceedings of the Int. Journal Num. Meth. Fluids*.
- [2] Mehrotra, M. K., & Patience, G. S. (1990). Unified Entry Length for Newtonian and Power Law Fluids in Laminar Pipe Flow. *Journal of Chemical Engineering*, 68, 529-533.
- [3] Launder, B. E., & Shima, N. (1989). Second Moment Closure for Near Wall Sublayer: Development and Application. *AIAA Journal*, 27, 1319-1325.
- [4] Haroutunian, & Engelman, S. (1991). On Modeling Wall-bound Turbulent Flows Using Specialized Near-wall Finite Elements and the Standard k-e Turbulent Model, *Advances in Num. simulation of Turbulent flows. ASME*, 117, 97-105.
- [5] Graft, T., Gerasimov, A., Iacovides, H., & Launder, B. (2002). Progress in the Generalization of Wall-Function Treatments. *International Journal for Heat and Fluid Flow*, 148-160.
- [6] Tamimi, S. (2012). Validation Models of Turbulent Flow and Heat Transfer. *Journal of Communication and Computer*, 9(9), 981-989.
- [7] Taylor, C., & Hughes, T. G. (1981) *Finite element programming of the Navier-Stokes equation*. Pineridge Press.
- [8] Wolfshtein, M. (1970). Some solutions of the plant

- 
- turbulent impinging jet trains, *ASME, Journal of Basic Engineering*, 92, 915-922.
- [9] Taylor, C., Hughes, T. G., & Morgan, K. (1980). A finite element model of one and two equation models of turbulent flow. *Recent Advances in Numerical Methods in Fluid*, 1, 311-344.
- [10] Zienkiewicz, O.C. (1977). *The Finite element method*. (3<sup>rd</sup>ed.). McGraw Hill.
- [11] Griffiths, D. V., & Smith, I. M. (2006). *Numerical Methods for Engineers*. (2<sup>nd</sup> ed.) Chapman and Hall/CRC.
- [12] Taylor, C., & Hood, P. (1972). A numerical solution of the Navier-Stokes equations using the finite element technique. *International Journal of Computers and Fluid*, 1-28.
- [13] Davies, T. J. (1972). *Turbulent phenomena*. Academic Press.
- [14] Nayak, U. S. L., & Stevens, S. J. (1973). An experimental study of the flow in the annular gap between a long vehicle and a low close-fitting tunnel, *Report: Dept. of Technology, Loughborough University of Technology*.

## Biographies

**SABAH TAMIMI** is an Associate Professor in the College of Engineering and Computing, Al Ghurair University, Dubai, UAE. He earned his M.Sc. (Computer Science, 1988) and Ph.D. (Applied Computer Science, 1992) degrees from the University of Wales, UK. He has been a member of the university faculty for over 22 years. In addition, for the last 11 years, he has been involved in the administrative sector working as Deputy Dean and Dean, in addition to serving as a member of the teaching faculty. His interests include computer modeling and simulation, computational flow dynamics, software testing techniques, computer graphics, and databases. He has many publications in international journals and conferences, and has reviewed many papers for international journals and conferences. Also, he is an editorial board member of different international journals. Dr. Tamimi can be reached at [sabah@agu.ac.ae](mailto:sabah@agu.ac.ae).

# TESTABILITY-GUIDED ABSTRACTION FOR SCALABLE BOUNDED MODEL CHECKING OF EMBEDDED SOFTWARE

Nannan He, Minnesota State University at Mankato; Michael S. Hsiao, Virginia Tech

## Abstract

In many reliability-critical applications, embedded software becomes increasingly important. Since an error may potentially cause disastrous results or severe economic consequences in these applications, the correctness of such software must be rigorously verified before it is put into operation. Although testing is still the mainstream approach in software verification practices to detect bugs and demonstrate the degree of confidence in the software quality, it cannot provide 100% coverage to rigorously verify safety requirements of the embedded software alone.

On the other hand, model checking, as an automated formal verification technique, is capable of proving or refuting the absence of bugs, with respect to specified properties. In recent years, with the successful application of model checking to formally verify hardware, more researchers are interested in applying this technique to verify embedded software. One major challenge in the scalable application of model checking embedded software in practice is that the state space traversed for formally checking is intractably large. A property-based abstraction-refinement method becomes necessary in scalable software formal verification, due to the fact the information required to verify a given property in a program is typically small. The basic idea is to automatically construct a small abstract model, which extracts property-relevant software segments and rules out irrelevant ones.

In this paper, the authors present an efficient abstraction-refinement method using a controllability metric (CM) and an observability metric (OM), which are two basic testing metrics for the high-level structural guidance for scalable software model checking. This method is based on an under- and over-approximate abstraction framework. Two basic testability metrics were applied to efficiently guide the proposed abstraction procedure. As a result, if the software fails to satisfy the property, this method can quickly find an error trace in the under-approximate model to demonstrate the property violation; otherwise, if the property holds, it can quickly compute a small and accurate over-approximate model by automatically abstracting away irrelevant information to prove the property.

## Introduction

Testing and automated formal methods are two basic software verification approaches. Testing is the mainstream approach in many software verification practices for detecting bugs and demonstrating confidence in the software quality before the product is delivered. Many techniques have been well established for software testing [1]. One of the main advantages of testing techniques is ease of use. A test practice simply applies input stimuli to initiate the program execution, and then compares whether or not actual outputs equal expected ones. More importantly, testing is scalable as it can be performed independent of complex implementation details. However, software testing is inadequate for guaranteeing program correctness in reliability-critical embedded applications, even if a program passes all available tests.

On the other hand, the formal method conducts an exploration of all possible behaviors, based on formal models of the program and the formal specification of intended requirements. The main advantage of the formal method is the completeness it offers in terms of specific properties, which can eliminate the limitation of inadequate coverage that conventional testing faces. This is an attractive feature in reliability-critical embedded applications. Furthermore, some automated formal methods like model checking can be used to detect hard corner-case bugs, which are hard to be detected by testing alone. With significant advances in automatic reasoning and the computing capability of modern computers, software verification using the formal method is no longer only of academic interest.

Model checking is an automated formal method that verifies specified properties of the finite state model (FSM) defined by a software program or hardware design [2]. It aims to prove the absence of bugs with respect to specified properties. Model checking can be fully automatic without requiring much expertise, which is not the case with all other formal verification methods. Furthermore, when the design under verification fails to satisfy a given specification, counterexamples can be generated automatically, which show erroneous behaviors of the design. Three researchers [3] shared the 2007 Turing Award for their work in devel-

oping model checking into a highly effective verification technology in both hardware and software industries.

Recently, one kind of model checking technique, called SAT- or SMT-based Bounded Model Checking (BMC), has become increasingly popular [4]. It unrolls the FSM within a bounded  $K$ , step transitions from the initial state(s), and checks whether a property violation can occur in the bounded  $K$ . When  $K$  reaches the length of the longest transitions in FSM, BMC can completely prove the property. The main advantage of BMC is that it can quickly find the counterexample (i.e., error trace) with the shortest path. Furthermore, with BMC, each non-Boolean variable is modeled as a bit-vector with a certain width (i.e., a vector of 32 bits for representing integers) [5]. Such data modeling is capable of precisely capturing the semantics of verification instances constrained by the physical word size on a computer. In particular, it allows for the detection of arithmetic overflow and underflow bugs, which are hard to find by conventional testing and other formal techniques. Thus, a number of software model-checking tools such as CBMC [6], SATABS [7], Saturn [8], F-SOFT [9], and hardware design validation techniques [10], [11] have employed the bit-vector modeling of program variables.

However, limited scalability is still the major problem with the formal method. In particular, one of the most serious problems with model checking is the state-space explosion problem, where the space increases exponentially with an increasing number of state variables. This problem is typically much worse in software verification models than that in hardware designs. Recently, many research efforts have been invested in alleviating the state explosion problem. The most popular approach is the property-based automated abstraction-refinement method [12], [13]. The basic idea is to create a small abstract model with information that is relevant to the property under verification and to remove irrelevant portions. An initial abstract is typically coarse and cannot include all relevant information. To automate the abstraction, an iterative refinement process usually accompanies the model-checking method. Refinement is necessary to learn and restore any relevant information that has not been identified in a previous abstraction step, until the abstraction is precise enough for verification. According to the verification capability that the abstract model can render, two important abstraction approaches for BMC are distinguished:

- Over-approximation-based abstraction renders the correct verification of proving the property. It may produce abstract error traces, which are infeasible in the original model.
- Under-approximation based abstraction renders the correct verification of refuting the property. It may

not be able to detect all error traces in the original model.

The ultimate aim of this current study was to explore the synergistic effects of the testing and formal methods in order to achieve both scalable and high-assurance verification for the reliability-critical embedded software. Presented here is an approach for utilizing testability metrics in order to guide the under- and over-approximation-based abstraction refinement processes for the scalable SAT- or SMT-based BMC of embedded software. Specifically, CM and OM have been used to direct the creation and refinement of the under-approximate abstraction model in two main ways: 1) the selection of a subset of execution paths in software BMC verification instances; and, 2) the discovery of the non-uniform bit-vector encoding widths of program variables on different execution paths. The testability metrics, especially OM, were used to guide the refinement of verification-relevant software constraints with low computational costs.

## Software Bounded Model Checking

Presented first are two assumptions that were needed for this study. Programs under verification are assumed to be written in ANSI-C, which is a popular programming language in embedded software. The properties being verified were assumed to be safety properties. It specifies in temporal logic that an error will never happen at any execution time instances with respect to all possible input combinations. Given an ANSI-C program,  $P$ , and safety properties,  $A$ , under verification, a typical SAT- or SMT-based software BMC has four basic steps, as shown in Figure 1.

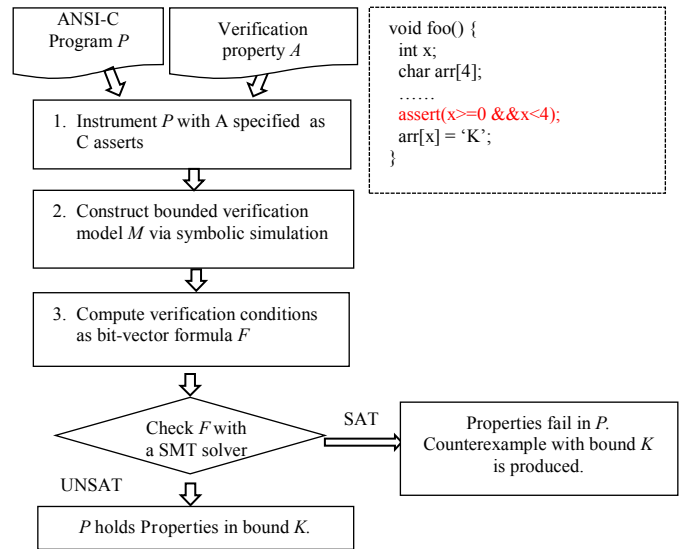


Figure 1. Basic Steps of Software BMC

1. First, implement program  $P$  with verification properties  $A$ , specified as C assertions. The C assertion is a special program statement where the asserted expression must always be true; otherwise, the execution aborts. In Figure 1, an example assertion is given, which claims that the value of the array index,  $x$ , must fall within the size of the array variable,  $arr$ , before each array access operation.
2. Then, construct the bounded verification model,  $M$ , by modeling both data and execution trace in bound for BMC via symbolic simulation in two ways:
  - All program variables are modeled as size-bounded bit-vectors.
  - All loop structures and recursions are unrolled to a bounded depth,  $K$ .

An example of modeling the execution sequence of loop structures and program variables is shown in Figure 2. The code block,  $A$ , includes all statements inside the loop. The maximum number of iterations of the *for* loop is three. On the right side of Figure 2, code block  $A$  is repeated three times and every integer variable is modeled as a vector of 32 bits. Any constant is shown in its binary representation; for example, the constant 0 is modeled as  $[0\dots0000]$  with 32 bits.

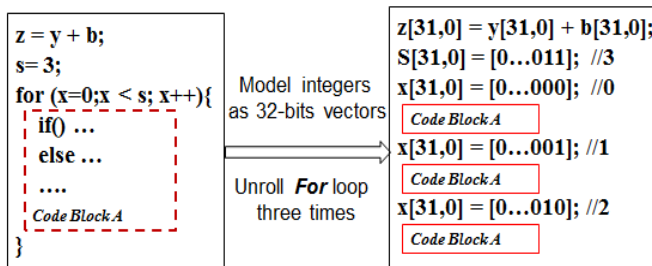


Figure 2. An Example Software Bounded Model

The verification model,  $M$ , is usually represented as the directed data dependency graph, where each node represents a program operation and arrows represent the data flow. Figure 3(b) shows model  $M$  in the Single Static Assignment (SSA) form for the C code example in Figure 3(a). All variables in  $M$  are in the SSA form. This format is an intermediate representation in compiler design, where each variable is defined only once. An original program variable is split into versions, indicated by the original name with a subscript. For example, the variable  $a$  has versions  $a_0, a_1, a_2$  in  $M$ . Source nodes in the graph are program input variables; for example, nodes  $a_0, b_0, d_0, z_0$  in Figure 3(b). The value computed by the operation of the output node represents the property to be verified; for instance,  $p$  in Figure 3(b). The

operator *ite* is the if-then-else operator, similar to the ternary conditional assignment operator  $?:$  in ANSI-C programming language. This step involves several compiler optimization techniques. More details could be found in the study by He and Hsiao [14].

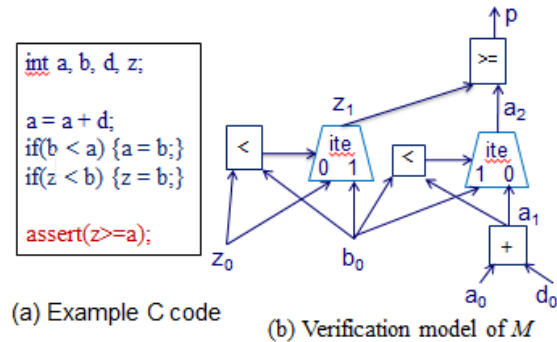


Figure 3. An Example Verification Model

3. Next, encode verification conditions in  $M$  including all bounded program constraints and the asserted property, into the first-order bit-vector arithmetic formula  $F$ . Figure 4 shows the bit-vector formula generated from the verification instance of the C code in Figure 3. All verification constraints in  $M$  are assembled using the AND logic operator  $\cap$ .  $C_0$  and  $C_1$  represent conditional predictors for the two *if* conditional expressions  $b_0 < a_1$  and  $z_0 < b_1$ , respectively. The last term of  $F$  is the negation of the property under verification,  $p$ . In order to prove that  $P$  is always true for any inputs, the negation logic is applied; that is, the complement of  $p$  is never true. Please note, though, that each variable in this formula is a bit-vector from the model.

$$\begin{aligned}
 F = & (c_0 = (b_0 < a_1)) \\
 & \cap (c_1 = (z_0 < b_1)) \\
 & \cap (a_1 = a_0 + d_0) \\
 & \cap (a_2 \text{ ite } c_0 \ b_0 \ a_1) \\
 & \cap (z_1 \text{ ite } c_1 \ b_0 \ z_0) \\
 & \cap (p = (z_1 \geq a_2)) \\
 & \cap \neg p
 \end{aligned}$$

Figure 4. An Example Bit-vector Formula of Figure 3

4. In the final step, the satisfiability of Formula  $F$ , which corresponds to the discovery of an error trace that violates the asserted property, is then decided by an SMT solver. An SMT solver is a software tool that determines the satisfiability of first-order formulas with respect to decidable theories, such as theories of arrays, integers, bit-vectors, lists, etc. [15]. It uses a



combination of SAT-solving and theory-specific decision procedures. The SAT problems derived from this BMC software are only related to the theory of fixed-width bit vectors.

If  $F$  is UNSAT (unsatisfiable), it is concluded that the property holds within a bound of  $K$ . If the depth of unrolling reaches the finite upper bound of loop iterations, the program property can be proved completely. On the other hand, if  $F$  is SAT (satisfiable), a satisfying solution is produced that can be translated to an error trace in the program.

Existing software tools could automate the BMC steps. A bounded model checker was used for checking ANSI-C, called CBMC, from Oxford University. By exhaustively exploring program behaviors up to a given loop bound, it can formally verify safety properties such as adhere-to-array bounds (no buffer overflows), user-specified asserted properties, etc. If a property fails, CBMC automatically returns an error trace. This trace can often be used to greatly simplify debugging of a faulty program. Many embedded applications have strict real-time requirements and limited memory resources. As a result, loop constructs in ANSI-C programs often have small execution iteration bounds. CBMC is capable of formally validating such bounds with the assist of unwinding assertions. Once the bound is established, it can prove the program correctness. Moreover, CBMC can model machine integers and IEEE floating-point arithmetic accurately in order to evaluate machine-level code such as bitwise operators and arithmetic overflows [16]. It is, therefore, able to detect software bugs that may be missed by many other formal verification tools. The source code of CBMC is available for free, and intermediate outputs from each BMC step are also accessible. This is the main reason that this tool was employed in this current study.

## Abstraction and Refinement

Multiple property-based automatic abstraction techniques have been proposed to enhance the scalability of software BMC. Their common objectives are to reduce the search space of the verification model and to eventually reduce the verification time. There are two main categories of abstraction and refinement techniques for model checking safety properties: over-approximation-based abstraction and under-approximation-based abstraction. Comparatively speaking, the over-approximation model,  $Mo$ , has a superset of behaviors of the original model,  $M$ , by definition; whereas, the under-approximation model,  $Mu$ , is defined to have a subset of the behaviors of  $M$ . From these definitions, the following correlations of  $Mo$  and  $Mu$  with  $M$  can be derived.

- All possible execution behaviors in  $M$  are also feasible in  $Mo$ , but some behavior in  $Mo$  may not be feasible in  $M$ . This feature implies that no false positive errors can occur in  $Mo$ . It can be concluded that: if the formula translated from  $Mo$  is UNSAT, the original model,  $M$ , must hold the property under verification.
- All possible execution behaviors in  $Mu$  are also feasible in  $M$ , but some behavior in  $M$  may not exist in  $Mu$ . This feature implies that no false negative errors can occur in  $Mu$ . It can be concluded: if the formula translated from  $Mu$  is SAT, the original model,  $M$ , must violate the property under verification.

## Over-approximation-based Abstraction

Predicate abstraction [17] and localization abstraction [18] are two examples of over-approximation-based abstraction techniques. In this current study, the localization abstraction was investigated, which has been widely used in scalable hardware verification. The idea of such abstraction is to completely remove input constraints of a set of non-input variables in the program and consider these variables as pseudo primary inputs. As a result, the abstract model,  $Mo$ , contains only a subset of computation constraints from the original verification model, which are represented by visible variables and data dependencies.

As shown in Figure 5, removed variables (nodes not filled with any color), including their data dependencies, are removed and become invisible in the abstract model. The variables (nodes filled with the color red) at the boundary between visible and invisible variables become pseudo primary inputs. As a result, the abstract model,  $Mo$ , contains only a subset of computation constraints from the original verification model, which are represented by visible variables and data dependencies.

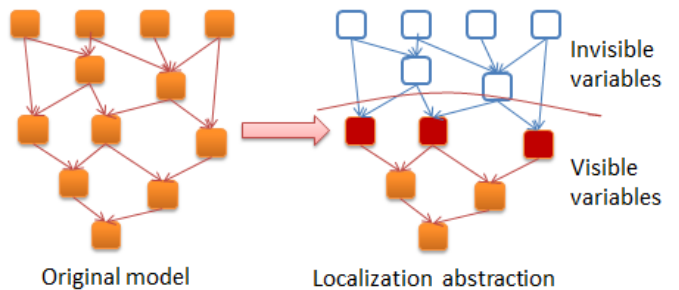


Figure 5. Over-approximate Abstraction

Figure 6 gives a simple localization abstraction example on a small C code segment. Figure 6(b) shows the initial abstraction from which program constraints in lines 1 and 2 are removed. The asserted property “assert( $z \geq y$ )” fails in this model since it removes program dependencies between the variables  $y$  and  $z$  and causes a false negative error. After

restoring the constraints in Line 2, a refined abstraction was obtained, as in Figure 6(c). Since the property holds in this model, so must it hold in the original program.

<pre> 1. y = z * x; 2. z = y; 3. if(x &gt; y) { 4.   z = x;    } 5. assert(z &gt;= y); </pre>	<pre> 1. if(x &gt; y) { 2.   z = x;    } 3. assert(z &gt;= y); </pre>	<pre> 1. z = y; 2. if(x &gt; y) { 3.   z = x;    } 4. assert(z &gt;= y); </pre>
---	---	---

(b) Initial Abstraction

(a) Original Program

(c) After Refinement

Figure 6. An Example Localization Abstraction

## Under-approximation-based Abstraction

Two under-approximation-based abstraction techniques which were used in this current study for software verification are shown in Figure 7.

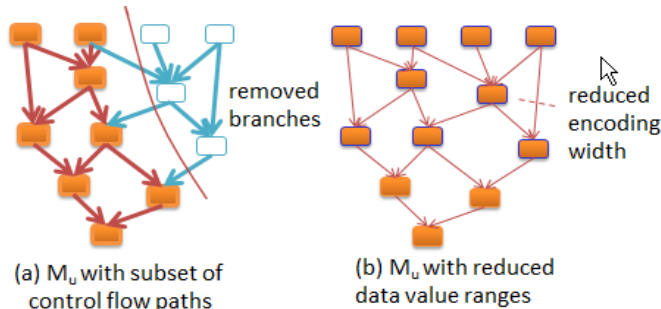


Figure 7. Two Under-approximate Abstractions

First, an abstract model,  $M_u$ , can be built by including a subset of execution paths in the verification model,  $M$ . This is done by forcing condition expressions in the program to a constant value of either 0 or 1 so that only one branch is taken for each enforced condition expression. This is shown in Figure 7(a) for the conditional *if* expression *if(expr1) {...} else {...}*, assuming the value of *expr1* is enforced to 1 only when the *if* branch is included in the abstract model,  $M_u$ , but all statements within the *else* branch are ruled out of  $M_u$ .

Second, another kind of abstract model,  $M_u$ , can be built by constraining the value of variables to a smaller range compared with the original variable value ranges in the original model,  $M$ . This can be realized by setting the encoding width of bit-vector variables smaller than their individual modeling width, as illustrated in Figure 7(b)—represented by thinner arrows compared with ones in Figure 7 (a). For example, assume that the modeling width of a variable  $x$  is 32 bits and whose value range is  $[-2^{31}, 2^{31}-1]$ ; the encoding width of  $x$  is then set to 4 and the higher 28 bits are set to a

constant 0—here, the value range is restricted to  $[0, 2^4-1]$ . This way, the search space of  $M_u$  is greatly reduced.

## Abstraction Refinement Methods

Counterexample-guided abstraction refinement (CEGAR) and proof-based refinement are the two most widely used abstraction refinement methods. CEGAR was initially proposed to automate localization abstraction and has been extended with several variations. The integration of CEGAR with predicate abstraction has been applied in several software model-checking tools such as SLAM from Microsoft Research for model checking device drivers [19]. More details about CESAR can be found in the study by Ball and Rajamani [20].

Proof-based refinement is an alternative abstraction refinement method, which was proposed based on the BMC and SAT/SMT solver. This current study also used the refinement approach together with under-approximate and over-approximate models in alternation.

1. Given an initial bound,  $K$ , it first verifies the satisfiability of the bit-vector arithmetic formula derived from the initial under-approximate abstract model,  $M_u$ . If the formula is SAT, a SAT solution with the bound  $K$  can be obtained to show how the property is violated and the verification is done. Otherwise, no counter-example (i.e., error trace) can be found, within the bound  $K$ , but it is inconclusive about whether the property holds in the original model,  $M$ .
2. It then builds an over-approximate abstraction,  $M_o$ , only with the set of program constraints used in the UNSAT proof of  $M_u$  produced from a SAT- or SMT-solver.
3. It verifies  $M_o$  without any bound constraints. If the property holds in  $M_o$ , it means that the property holds in the original model as well and the verification is done. Otherwise, an abstract counterexample  $\lambda$  that violates the property in abstract model  $M_o$  can be obtained, but it is inconclusive about whether  $\lambda$  is feasible in the original model,  $M$ .
4. Finally, the bound of  $\lambda$ , instead of values assigned in  $\lambda$ , is used to determine a larger bound,  $K'$ , and to refine the under-approximate model,  $M_u$ , and a new refinement iteration starts.

The basic workflow of this method with four stages is shown in Figure 8. The actual refinement stages may vary depending on the construction of the under-approximate and over-approximate models. In comparison, with the proof-based abstraction refinement approach, all counterexamples within the bound,  $K$ , are ruled out at once; whereas, CEGAR may need multiple refinement iterations to rule out

these counterexamples. However, the proof-based approach has one main drawback: it is more computationally intensive to extract the refutation proof that there are no counterexamples of bound  $K$ , than to refute a single abstract counterexample. Thus, the refinement step becomes a bottleneck in the process.

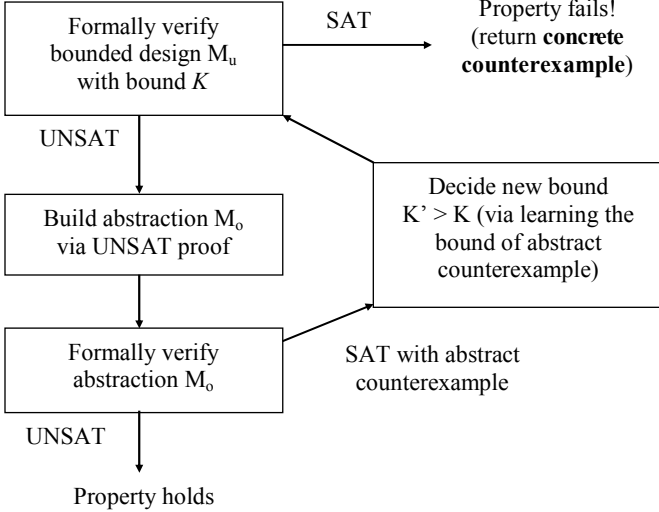


Figure 8. Basic Flow of Proof-based Refinement

## Controllability/Observability Metrics

Testing metrics are used to measure the testability of a hardware design or software product. Controllability and observability are two basic testing metrics. They have been used to identify hard-to-test regions in a design, where faults can hide and not be easily reached by testing [21], and analyze the testability of software components [22].

In software BMC, the CM of a variable,  $x$ , is expected to approximate the difficulty of setting a value of  $x$  along all paths from primary inputs to the variable  $x$ . The difficulty is defined by two main factors: the maximum length of a path and the computation complexity along the path. The OM of a variable,  $x$ , is defined to estimate the difficulty of observing the value-change of the variable,  $x$ , at the output. Formal verification aims to check the correctness of a design with complete input coverage. Thus, CM can provide high-level structural guidance for the search of a SAT solution, according to different expectations. To be specific, it has the potential to either search easy-to-control paths first, if the goal is to test common cases, or search hard-to-control paths first to verify corner cases, where random/directed testing cannot. On the other hand, hard-to-observe variables are less likely to affect the verification property, as variations of their values have little probability for changing the property.

Thus, OM provides guidance for estimating whether a variable should be included in the over-approximate abstraction.

Table 1 shows the Controllability and Observability Coefficient (COC) values of basic program operators, as defined by Bryant et al. [23]. These COCs approximate different amounts of influence that basic operators have on the CM and OM. The CM and OM of each individual variable computed by an operator were proposed to be derived by using its COC.

Table 1. COC of Basic Program Operators

OPs	COC	Ops	COC
&,	2	+	10
XOR	2.5	-	11
==, !=	3	≤, ≥	12
>, <	9	ITE(if-then-else)	2.5

## CM and OM Calculations

CMs of all primary inputs (PIs) of the verification model are initialized to zero because values of these input variables are fully controllable. The controllability of an internal variable is affected by the controllability of its inputs and the type of operator applying on this variable. Thus, the CM of a variable,  $x$ , can be computed using the COC of the operator  $op$  and the CMs of all of its inputs  $IN(i)$ , as shown in Equation (1). The symbol  $N$  is the number of inputs of variable  $x$ .

$$CM(x) = \text{Max}\{CM(IN_i)\} + COC(op) \quad (1)$$

where,  $0 \leq i < N$ .

After the CM values of all variables have been calculated, the OM value of each variable can be computed backwards starting from the primary outputs (POs). The OMs of one or more POs are initialized to zero. Then, the OM of an input variable,  $IN(i)$ , is computed as the sum of the COC value of  $x$ 's operator  $op$ , and the maximum value among the CM, of all side inputs of  $x$  and the OM of  $x$  via Equation (2):

$$OM(IN_i) = COC(op) + \text{Max}\{OM(x), \text{Max}\{CM(IN_{j+i})\}\} \quad (2)$$

The CM/OM computed using Equation (2) has two important properties related to this current study:

- The larger the value of the CM, the harder the controllability.
- The larger the value of the OM, the weaker the observability.
- On any computation path from PIs to a PO,  $CM(x_1) < CM(x_2)$  and  $OM(x_1) > OM(x_2)$  always hold, if  $x_1$  is the predecessor of  $x_2$ .

# Testability-guided Abstraction for Software BMC

An efficient path-oriented bit-vector encoding width computation approach was proposed for constructing the abstract model. First, the key concept of the encoding width,  $We$ , of a variable was defined as follows.

Starting from the least significant bit, the encoding width,  $We(x)$ , for a bit-vector variable  $x$  is defined as the number of consecutive bits in the vector whose values have not been assigned:  $0 \leq We(x) \leq W(x)$ . For each of the remaining  $W(x) - We(x)$  bits, the value is set to be a constant 0 or 1.

Similar to the work done by Bryant et al. [24], this current approach embedded the dynamic computation of encoding width  $We$  in the abstraction-refinement framework. However, the work in this study is distinguished by its path-oriented analysis with the guidance of the static controllability metric and the observability metric in three major ways. First, it computes the initial non-uniform  $We$  of variables on different paths. By setting a bigger initial  $We$  for the variables on the easy-to-control paths, while setting a smaller  $We$  for other paths, this approach can greatly increase the chance of finding a SAT solution in the restricted search space directly, without the need to adjust  $We$  iteratively. Second, in  $We$  adjustment steps (if necessary), this approach gives priority to enlarging the  $We$  of the easily controllable variables by manipulating the abstract counterexample generation guided by CM and OM. This helps to systematically search for the true counterexample or error trace with less effort. Third, it sets  $We$  to zero for some single-bit variables that determine path(s) selection, thereby enforcing constant values on them to restrict choosing only a subset of paths. This can avoid searching those partitions that have been checked in previous refinement steps, especially the ones on which the variables'  $We$  experienced no increase, thus simplifying the problem.

Figure 9 illustrates the basic idea of refining the non-uniform encoding width. Figure 9(a) shows the original verification model,  $M$ , without abstraction, where the encoding width,  $We$ , equals the modeling width,  $W$ , for each variable. Figure 9(b) shows the initial under-approximate abstraction model  $M_u$ , where different variables are restricted with different encoding widths,  $We$ . The variables on some paths are assigned encoding widths different from the ones for the variables on other paths. If an error trace is found in the initial  $M_u$ , it is concluded that the property fails in  $M$ . If no error traces could be found in  $M_u$ , this approach refines  $M_u$  using the proof-based refinement method guided by CM and OM to obtain a new set of encoding widths assigned, as shown in Figure 9(c).

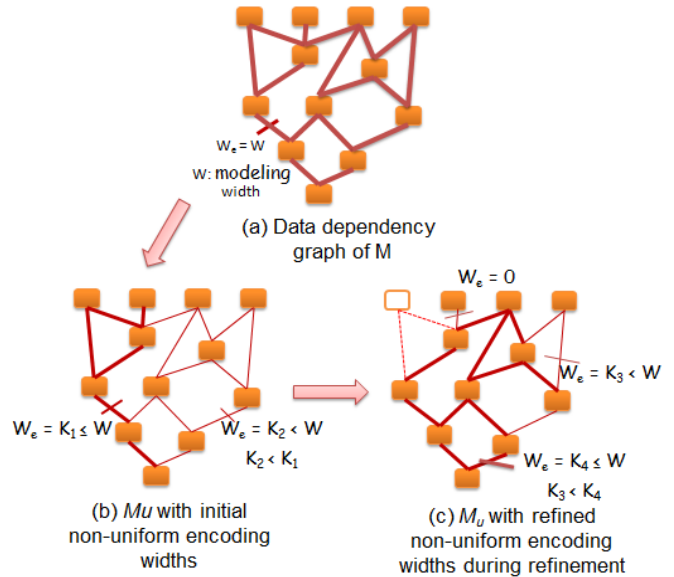


Figure 9. Overview of  $We$  Variance During Refinement

Figure 10 gives the basic flow of the guided abstraction refinement approach. Compared with the basic proof-based abstract refinement workflow in Figure 8, three main steps are enhanced by testability metrics guidance, which are distinguished with dashed lines.

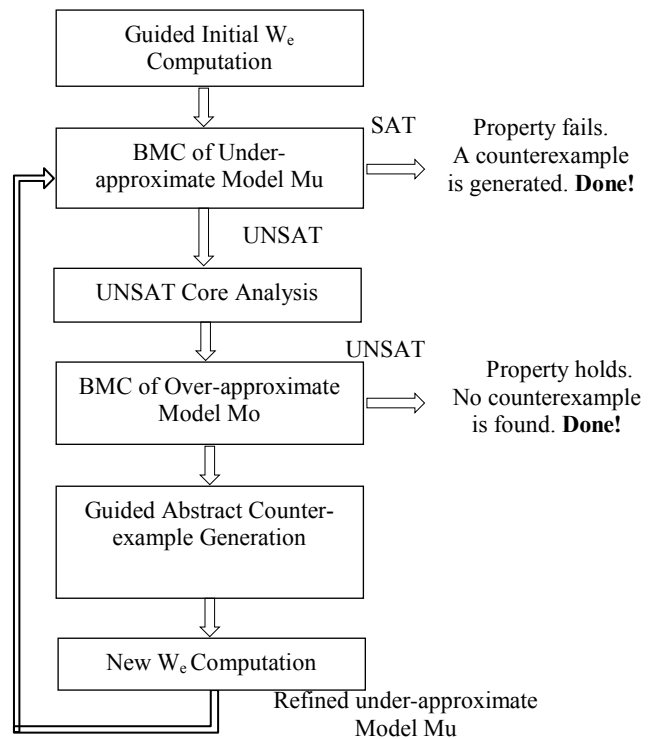


Figure 10. Basic Workflow

---

## Guided Encoding Width Computation

The computation of encoding width  $We$  involves two stages. First, the initial  $We$  computation guided by testability before abstraction starts. Second,  $We$  is adjusted during each abstraction refinement iteration. In the first stage, the initial  $We$  is computed and the constant value to be assigned to the bits beyond the encoding width is determined. For example, an integer variable is modeled with a bit-vector of 32 bits—assuming that  $We$  is 4, the bits being encoded are [0, 3] and the bits beyond the encoding width are [3, 31], each of which is assigned to be 1 or 0. The basic idea is to give preference (a larger initial  $We$  value) to variables on easy-to-control paths, which are identified by the controllability metric in order to increase chances of quickly finding an error trace at the lowest computational costs. Moreover, this keeps the encoding width of a variable consistent with its inputs or relevant output variables and the given operator type.

In the second stage, given the abstract counterexample  $\tau$ , the approach enlarges  $We$  of some variables in  $Mu$  so that the new encoded value range can adequately cover their values assigned in  $\tau$ . It also updates the  $We$  of all other data dependent variables in  $M$  and applies the new  $We$  on them in order to build the new  $Mu$ . To avoid repeatedly searching the same space among refinement iterations, model slicing is applied in order to remove variables whose  $We$  do not increase from the new  $Mu$ . As the search space of  $Mu$  or  $Mo$  is finite, this approach is guaranteed to terminate. In the worst-case scenario, the  $We$  of variables may need to be enlarged in order to be equal to their modeling width  $W$ . In most applications, the verification iteration stops rather quickly.

## Conclusion

In this paper, the authors present a scalable formal verification approach of embedded software written in C, which makes use of bounded model checking, an under-/over-approximate-based abstraction framework, and two testability metrics. In order to enhance the abstraction accuracy and reduce the computation cost of automatic abstraction, the controllability metric and the observability metric are proposed to be utilized as program structural guidance. As a result, if the software fails to satisfy the property, this method can quickly find an error trace in the under-approximate model to demonstrate the property violation; conversely, it can quickly compute a small and accurate over-approximate model by automatically abstracting away irrelevant information to prove the property. The next phase of this research project is underway to apply this proposed approach

to safety-critical embedded software written in C and to formally verify safety properties specified as C assertions.

## References

- [1] Biezer, B. (1990). *Software Testing Techniques*. International Thomson Computer Press.
- [2] Clarke, E. M., Grumberg, O., & Peled, D. A. (2000). *Model Checking*. MIT Press.
- [3] Clarke, E. M., Emerson, E. A., & Sifakis, J. (2009). Model checking: algorithmic verification and debugging. *Communications of the ACM* 52(11), 74-84.
- [4] Biere, A., Clarke, E. M., Fujita, M., & Zhu, Y. (1999). Symbolic model checking using SAT procedures instead of BDDs. *Proceedings of ACM/IEEE DAC*, (pp. 317-320). San Diego, CA.
- [5] Barrett, C. W., Dill, D. L., & Levitt, J. R. (1998). A decision procedure for bit-vector arithmetic. *Proceedings of ACM/IEEE Design Automation Conference*. (pp. 519-531). San Diego, CA.
- [6] Clarke, E. M., & Kroening, D. (2004). A tool for checking ANSI-C programs. *Proceedings of TACAS*, (pp. 168-176). Barcelona, Spain.
- [7] Clarke, E. M., Kroening, D., Sharygina, N., & Yorav, K. (2005). SATABS: SAT-based predicate abstraction for ANSI-C. *Proceedings of Tools and Algorithms for Construction and Analysis of Systems*, (pp. 570-574). Edinburgh, UK.
- [8] Xie, Y., & Aiken, A. (2005). Scalable error detection using Boolean satisfiability. *Proceedings of ACM Symposium on POPL* (pp. 351-363). Long Beach, CA.
- [9] Ivancic, F., Shlyakhter, I., Ganai, M., & Gupta, A. (2005). Model checking C programs using F-Soft. *Proceedings of the International IEEE ICCD* (pp. 297-308). Las Vegas, NV.
- [10] Huang, C. Y., & Cheng, K. T. (2000). Assertion checking by combined word-level ATPG and modular arithmetic constraint-solving techniques. *Proceedings of ACM/IEEE DAC* (pp. 118-123). San Diego, CA.
- [11] Manolios, P., Srinivasan, S. K., & Vroon, D. (2006). Automatic memory reductions for RTL-level verification. *Proceedings of the International IEEE IC-CAD* (pp. 786-793). San Jose, CA.
- [12] Amla N., & Kenneth L. M. (2004). A hybrid of counterexample-based and proof-based abstraction. *Proceedings of IEEE FMCAD* (pp. 260-274). Austin, Texas.
- [13] Amla, N., & McMillan, K L. (2007). Combining Abstraction Refinement and SAT-Based Model Checking. *Proceedings of IEEE TACAS* (pp. 405-419). Braga, Portugal.



- 
- [14] He, N., & Hsiao, M. (2007). Bounded model checking of embedded software in wireless cognitive radio systems. *Proceedings of the International IEEE conference on Computer Design*, (pp. 19-24). Lake Tahoe, CA.
- [15] Barrett, C., Stump, A., & Tinelli, C. (2010). The SMT-LIB Standard: Version 2.0, 2014. *Proceedings of the 8<sup>th</sup> International Workshop on Satisfiability Modulo Theories*. Edinburgh, UK.
- [16] Brillout, A., Kroening, D., & Wahl, T. (2009). Mixed abstractions for floating-point arithmetic. *Proceedings of IEEE FMCAD* (pp. 69-76). Austin, TX.
- [17] Das, S. (2003). *Predicate Abstraction*. Unpublished doctoral dissertation. Stanford University.
- [18] Kurshan, R. P. (1994). *Computer Aided Verification of Coordinating Processes*. Princeton University Press.
- [19] Mang, F. Y. C., & Ho, P. H. (2004). Abstraction refinement by controllability and cooperativeness analysis. *Proceedings of the ACM/IEEE Design Automation Conference*, (pp. 224-229). San Diego, CA.
- [20] Ball, T., & Rajamani, S. (2000). Behop: A symbolic model checker for B programs. *Proceedings of SPIN*, (pp. 113-130). Stanford, CA.
- [21] Hsu, F., Rudnick, E. M., & Patel, J. H. (1996). Enhancing High-level Control-Flow for Improved Testability. *Proceedings of IEEE ICCAD*, (pp. 322-328). San Jose, CA.
- [22] Nguyen, T., Delaunay, M., & Robach, C. (2002). Testability analysis for software components. *Proceedings of ICSM*, (pp. 422-429). Montreal, Canada.
- [23] He, N., & Hsiao, M. (2008). A New Testability Guided Abstraction to Solving Bit-vector Formula. *Proceedings of Joint Workshops of the 6th International Workshop on SMT and 1st International Workshop on Bit-Precise Reasoning*, (pp. 39-45). Princeton, NJ.
- [24] Bryant, R. E., Kroening, D., Ouaknine, J., Seshia, S. A., Strichman, O., & Brady, B. (2007). Deciding Bit-Vector Arithmetic with Abstraction. *Proceedings of TACAS*, (pp. 358-372). Braga, Portugal.
- cal Engineering from the University of Illinois at Urbana-Champaign. His research interests include architectural-level and gate-level automatic test pattern generation (ATPG), design verification and diagnosis, power estimation and management in VLSI, embedded software, etc. He is a fellow of IEEE. Dr. Hsiao may be reached at [Hsiao@vt.edu](mailto:Hsiao@vt.edu).

## Biographies

**NANNAN HE** is currently an Assistant Professor in the ECET Department at Minnesota State University, Mankato. She received a Ph.D. in computer engineering from Virginia Tech. Her teaching and research interests include formal verification and testing of safety-critical embedded software and design of real-time embedded systems. Dr. He may be reached at [Nannan.he@mnsu.edu](mailto:Nannan.he@mnsu.edu).

**MICHAEL S. HSIAO** is a Professor in the ECE Department at Virginia Tech. He received the Ph.D. in Electric



# INSTRUCTIONS FOR AUTHORS: MANUSCRIPT REQUIREMENTS

The INTERNATIONAL JOURNAL OF MODERN ENGINEERING is an online/print publication, designed for Engineering, Engineering Technology, and Industrial Technology professionals. All submissions to this journal, submission of manuscripts, peer-reviews of submitted documents, requested editing changes, notification of acceptance or rejection, and final publication of accepted manuscripts will be handled electronically. The only exception is the submission of separate high-quality image files that are too large to send electronically.

All manuscript submissions must be prepared in Microsoft Word (.doc or .docx) and contain all figures, images and/or pictures embedded where you want them and appropriately captioned. Also, for all accepted manuscripts, each figure, image or picture that was imported into your Word document must be saved individually as a **300dpi or higher JPEG (.jpg)** file and submitted separately; your manuscript and figure numbers must be used in the title of the file (e.g., **M13-F-18 Figure 4**); that means one additional file for each image imported into your manuscript. These 300dpi images do NOT need to be embedded in your manuscript. For tables or graphs created directly in Word, you do not need to submit them as separate files.

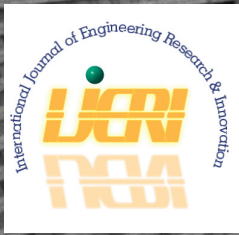
Included below is a summary of the formatting instructions. You should, however, review the [sample Word document](http://www.ijme.us/formatting_guidelines) on our website ([www.ijme.us/formatting\\_guidelines](http://www.ijme.us/formatting_guidelines)) for details on how to correctly format your manuscript. The editorial staff reserves the right to edit and reformat any submitted document in order to meet publication standards of the journal.

The references included in the References section of your manuscript must follow APA-formatting guidelines. In order to help you, the sample Word document also includes numerous examples of how to format a variety of scenarios. Keep in mind that an incorrectly formatted manuscript will be returned to you, a delay that may cause it (if accepted) to be moved to a subsequent issue of the journal.

1. **Word Document Page Setup:** Two columns with 1/4" spacing between columns; Top of page = 3/4"; Bottom of page = 1" (from the top of the footer to bottom of page); Left margin = 3/4"; Right margin = 3/4".
2. **Paper Title:** Centered at the top of the first page with a 22-point Times New Roman (Bold), Small-Caps font.
3. **Page Breaks:** Do not use page breaks.
4. **Body Fonts:** Use 10-point Times New Roman (TNR) for body text throughout (1/8" paragraph indentation); 9-point TNR for author names/affiliations under the paper title; 16-point TNR for major section titles; 14-point TNR for minor section titles; 9-point TNR BOLD for caption titles for tables and figures; other font sizes as noted in the sample document.
5. **In-text Referencing:** List and number each reference when referring to them in the body of your document (e.g., [1]). The first entry must be [1] followed by [2], [3], etc., continuing in numerical order to the final entry in your References section. Again, see the sample Word document for specifics. Do not use the End-Page Reference utility in Microsoft Word. You must manually place references in the body of the text.
6. **Tables and Figures:** Center all tables and figures. Captions for tables must be above the table, while captions for figures are below; all captions are left-justified.
7. **Page Limit:** Manuscripts should not be more than 15 pages (single-spaced, 2-column format).
8. **Page Numbering:** Do not use page numbers.
9. **Publication Charges:** Manuscripts accepted for publication are subject to mandatory publication charges.
10. **Copyright Agreement:** A Copyright Transfer Form must be signed and submitted by all authors on a given paper before that paper will be published.
11. **Submissions:** All manuscripts and associated files must be submitted electronically.

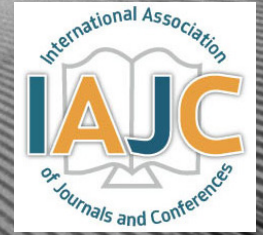
MANUSCRIPTS should be submitted to Dr. Philip D. Weinsier, manuscript editor, at [philipw@bgsu.edu](mailto:philipw@bgsu.edu) along with a copy to [editor@ijme.us](mailto:editor@ijme.us).

FILES containing your high-quality images should **ONLY** be submitted to [philipw@bgsu.edu](mailto:philipw@bgsu.edu).



[www.ijeri.org](http://www.ijeri.org)

Print ISSN: 2152-4157  
Online ISSN: 2152-4165



[www.iajc.org](http://www.iajc.org)

## INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH AND INNOVATION

### **ABOUT IJERI:**

- IJERI is the second official journal of the International Association of Journals and Conferences (IAJC).
- IJERI is a high-quality, independent journal steered by a distinguished board of directors and supported by an international review board representing many well-known universities, colleges, and corporations in the U.S. and abroad.
- IJERI has an impact factor of **1.58**, placing it among an elite group of most-cited engineering journals worldwide.

### **OTHER IAJC JOURNALS:**

- The International Journal of Modern Engineering (IJME)  
For more information visit [www.ijme.us](http://www.ijme.us)
- The Technology Interface International Journal (TIIJ)  
For more information visit [www.tiij.org](http://www.tiij.org)

### **IJERI SUBMISSIONS:**

- Manuscripts should be sent electronically to the manuscript editor, Dr. Philip Weinsier, at [philipw@bgsu.edu](mailto:philipw@bgsu.edu).

For submission guidelines visit  
[www.ijeri.org/submissions](http://www.ijeri.org/submissions)

### **TO JOIN THE REVIEW BOARD:**

- Contact the chair of the International Review Board, Dr. Philip Weinsier, at [philipw@bgsu.edu](mailto:philipw@bgsu.edu).

For more information visit  
[www.ijeri.org/editorial](http://www.ijeri.org/editorial)

### **INDEXING ORGANIZATIONS:**

- IJERI is currently indexed by 16 agencies. For a complete listing, please visit us at [www.ijeri.org](http://www.ijeri.org).

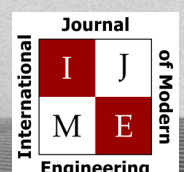
### **Contact us:**

**Mark Rajai, Ph.D.**

Editor-in-Chief  
California State University-Northridge  
College of Engineering and Computer Science  
Room: JD 4510  
Northridge, CA 91330  
Office: (818) 677-5003  
Email: [mrajai@csun.edu](mailto:mrajai@csun.edu)



[www.tiij.org](http://www.tiij.org)



[www.ijme.us](http://www.ijme.us)

# IAJC UPCOMING 2014 CONFERENCE

## 2014.IAJC.ORG

**The latest impact factor (IF) calculation (Google Scholar method) for IJME of 3.0 moves it even higher in its march towards the top 10 engineering journals.**

**IJME IS THE OFFICAL AND FLAGSHIP JOURNAL OF THE  
INTERNATIONAL ASSOCIATION OF JOURNALS AND CONFERENCE (IAJC)**

[www.iajc.org](http://www.iajc.org)



The International Journal of Modern Engineering (IJME) is a highly-selective, peer-reviewed journal covering topics that appeal to a broad readership of various branches of engineering and related technologies. IJME is steered by the IAJC distinguished board of directors and is supported by an international review board consisting of prominent individuals representing many well-known universities, colleges, and corporations in the United States and abroad.

### **IJME Contact Information**

**General questions or inquiries about sponsorship of the journal should be directed to:**

**Mark Rajai, Ph.D.**

**Editor-in-Chief**

**Office: (818) 677-5003**

**Email: [editor@ijme.us](mailto:editor@ijme.us)**

**Department of Manufacturing Systems Engineering & Management**

**California State University-Northridge**

**18111 Nordhoff St.**

**Northridge, CA 91330**