A publication of ISCA*:
International Society for Computers
and Their Applications

INTERNATIONAL JOURNAL OF COMPUTERS AND THEIR APPLICATIONS

TABLE OF CONTENTS

*** "International Journal of Computers and Their Applications is abstracted and indexed in INSPEC and Scopus."**

Volume 26, No. 3, Sept. 2019 **ISSN 1076-5204**

International Journal of Computers and Their Applications

A publication of the International Society for Computers and Their Applications

ISCA ···•···278 Mankato Ave, #220, Winona, MN 55987 USA···•···Phone: (507) 458-4517
E-mail: isca@ipass.net • URL: http://www.isca@isca-hq.org.

Guest Editorial Preface
Special Issue from ISCA CATA 2019

This Special Issue of IJCA is a collection of five refereed papers, which are extensions of conference papers selected from the *34th International Conference on Computer and Their Applications* (CATA 2019).

Each paper was evaluated by at least two reviewers, judging the originality, technical contribution, significance and quality of the manuscript. Following CATA 2019, a number of high-quality papers were recommended by the Program Co-Chairs to be considered for publication in this Special Issue of IJCA, resulting in the set of papers provided here.

The papers in this special issue cover a wide range of research interests in areas of computers and applications. The topics and main contributions of the papers are briefly summarized below.

Kasi Periyasamy and Saleh Alsyefi, in their paper entitled *Design and Implementation of Nephro Net - a Healthcare Social Network,* address the topic of healthcare and in particular the dissemination and usage of healthcare information through social networks. The authors introduce Nephro Net, the social network for healthcare information exchange and show how such a system can provide valuable communication to all involved. They illustrate the power of the network with a kidney disease example and show how security and privacy can be maintained while using their social network.

Lifeng Liu, Meilin Liu, and Chongjun Wang focus on the topic of general purpose graphics processing units in their paper entitled *Locality-aware CTA mapping for GPUs.* In particular, the authors develop a cooperative thread array mapping method to address the problems of most GPU hardware architectures and, when combined with a programmable warp scheduler, they show performance improvement of over 50% when compared to traditional GPUs.

In their paper *Comparison of Metal-Contamination Removal Rates when using Three Magnets of Different Shapes*, authors Takashi Ohnishi, Takashi Okamoto, and Keiichi Watanuki discuss the problem of removing foreign metal contaminants using magnetic separators. The performance of these devices depends upon the shape of these devices and they investigate the removal rates of small stainless-steel particle contaminants using triangular, pear and circular shaped magnets. Experiments were performed and data analysis were generated to show that the pear-shaped device produces the best performance in metal contaminant removal.

Ana Farhat, Kyle Hagen, Ka C Cheok and Balaji Boominathan consider the research problem of modeling electronic brake systems in their paper *Co-Active Neuro-Fuzzy Inference System Modeling with Clustering Methods.* The EBS provides high performance braking in automobiles, with such desirable features as being lightweight and possessing faster response times than other braking mechanisms. In order to provide full usage of these devices, it is important to adequately model them. The authors develop a novel EBS model identification method, based upon the co-active, neuro-fuzzy inference system and show, through experimentation, that their model provides an accurate model for the EBS which will help automobile manufacturers in designing higher performance control systems for braking systems.

Finally, in their paper entitled *High Level Synthesis Optimization of AES-128/192/256 Encryption Algorithms*, Luka Daoud, Fady Hussein and Nader Rafla address the issue of network security protocols. In particular, they implement the Advanced Encryption Standard processor using Xilinx Vivado HLS on a FPGA for different sizes and investigate pipelining to optimize throughput. Results show a performance enhancement to 26 giga bits per second using the Artix-7 FPGA.

We hope you enjoy this special issue of the IJCA and we look forward to seeing you at a future ISCA conference. More information about the ISCA society can be found at http://www.isca-hq.org.

Guest Editors:

Ying Jin, California State University at Sacramento, USA, CATA 2019 Program Co-Chair
Gordon Lee, San Diego State University, USA, CATA 2019 Program Co-Chair

August 2019

Design and Implementation of Nephro Net - a Healthcare Social Network

Kasi Periyasamy* and Saleh Alsyefi*
University of Wisconsin-La Crosse, La Crosse, WI 54601, USA

Abstract

Technological advancements made it possible to introduce social networks for common people. Through social networks, people can easily communicate with others, quickly exchange vital information and get to know about social events. Healthcare social networks form a branch of social networks where the information exchanged, discussed and revealed all relate to healthcare. Such networks provide valuable information for patients and also establish a quickly accessible medium to communicate with healthcare providers. In addition, healthcare providers may also use this forum to share their service and experience with each other. A recent medical survey indicates that healthcare social networks are very helpful in promoting awareness of health issues, discussing related health problems with other patients and healthcare providers, and finding quick solutions for some of the health problems. This paper describes the design and implementation of a healthcare network focusing on two aspects - security and privacy. These aspects are considered to be more important in the development of a healthcare social network as reported in the literature. The authors chose Nephrology, the study of kidney diseases, for illustration. However, the design and implementation of the network have been made sufficiently generic so that they can be used for other health domains such as Gynecology and Psychiatry.

Key Words: Healthcare, social network, security, privacy, nephrology.

1 Introduction

Social media and social networks are powerful tools that provide a forum to disseminate and exchange information very quickly to the public. In particular, the advent of mobile technology made it easier for anyone to access social information in a few seconds. As reported in Oxford Research Encyclopedia [11], 37% of world's population used social media in 2017. There is a ten-fold increase in the use of social media by Americans during the time 2005-2015. A similar trend is also seen in the United Kingdom in 2016 where more than six in 10 adults use social media. While a social media is a channel to disseminate information, a social network, on the other hand, also provides additional mechanisms for interactions among users to exchange information.

In the healthcare domain, social media is used to promote healthcare education by disseminating vital information such as an outbreak of a disease. In addition, it is also used to educate people regarding healthcare issues such as community service (e.g., blood donation), raising awareness on health issues and promoting public awareness in legal problems involving healthcare [1]. Healthcare networks, on the other hand, are more interactive; they are used to seek, query, comment, share and exchange healthcare information. Some healthcare networks are available to only healthcare providers (Sermo[16], Doximity[4], and QuantiaMD[14], to name a few). These networks provide a forum for providers to discuss specific cases or problems, and exchange diagnostic and medical information. For example, Sermo[16] provides a virtual lounge for physicians to share and solve challenging cases. Sermo, Doximity and QuantiaMD are considered to be general purpose healthcare networks in which physicians from different disciplines join and share information. Networks such as ASN Communities[2] and Kidney Health Australia[8] provide services to physicians in a specific discipline (in this case, Nephrology).

There are some other healthcare networks, available for both groups - patients and healthcare providers (PatientsLikeMe[12], Mercy Health Network[10], and Triad Healthcare Network[17] to name a few). Griffiths and others[7] have discussed the advantages of using healthcare networks and how they support both patients and healthcare providers. In particular, they argue that "social networking has the potential to change patterns of health inequalities and access to healthcare, alter the stability of healthcare provision and lead to a reformulation of the role of health professionals." Moorhead[11] describes a wide range of activities that are supported by healthcare networks which include providing answers to medical questions, facilitating dialogues between patients, and between patients and healthcare professionals. In addition, some healthcare networks such as First Opinion[6] also provide facilities for online consultations. The authors of this paper are highly motivated by the benefits of using a healthcare network as described by Moorhead. These include increased interactions among patients, interactions between patients and healthcare professionals, more available, shared, and tailored information, and increased peer, social and

*Department of Computer Science. Email: kperiyasamy@uwlax.edu, saleh.alsyefi@gmail.com

emotional support.

Two of the most important factors to be addressed in the development of healthcare social networks are *security* and *privacy*[11, 9]. Social networks such as *Facebook* and *Twitter* are used for general purpose in the sense that any topic can be posted or discussed in these networks. As a result, anyone can join these networks and exchange any information. Though there are some security controls available in these networks, they are not sufficiently strict. Li pointed out that openness in virtual communication and the growing trend of using technology for sharing healthcare information create increased opportunities for misuse of healthcare data. Even though the Health Information Portability and Accountability Act (HIPPA) of 1996 in the United States protects people from unauthorized access of healthcare information, HIPPA does not protect the privacy of an individual when the individual himself or herself posts his/her healthcare related information on a social network[9]. Many users of general healthcare media such as *Facebook* and *Twitter* are not aware of this openness in communication and the huge audience behind it. Because of the sensitivity of healthcare information, the designer of healthcare networks should focus on additional mechanisms by which a patient's healthcare information is not leaked accidentally or inadvertently. Security of a healthcare network (in fact, any computer network) mostly depends on the design of such a network. The design involves several factors including the types of users, type or sensitivity of information to be maintained and/or to be shared, the resources available, and potential threats known or published in the literature.

This manuscript describes the design and implementation of a social network called *Nephro Net*. The authors presented security and privacy aspects of Nephro Net at an international conference[13]. Though the current implementation of Nephro Net was created for patients with kidney problems and professionals in Nephrology (a study of kidney diseases), the network can be easily adapted to other domains as well. Section 2 describes an overview of Nephro Net, followed by discussions on some design decisions of Nephro Net in Section 3 and its operations in Section 4. The final section includes concluding remarks, challenges and future work.

2 Nephro Net and its Features

Patient-centric networks are gaining more popularity these days in which patients have more control in entering and releasing personal healthcare information[5]. A recent survey published by the Massachusetts Medical Society [18] indicates that 52% (out of 601 responses) indicated that patient-to-patient support will be very useful; 85% considered that social healthcare networks are useful for chronic disease management; and 75% mentioned that disease specific patient support groups will be the most effective. These statistics along with the notion of patient-centric design motivated the authors to develop a healthcare social network called *Nephro Net*, specifically designed for patients who have kidney problems and healthcare providers in Nephrology. The aim of this network is to provide a forum for patients to search, post and discuss kidney related healthcare problems. Professionals in Nephrology can also join the network to disseminate specific information to patients (such as an optimal dosage of a medication for dialysis patients with diabetes), to answer questions posted by other users (both patients and professionals) in the system and engage in private consultation with a patient. Nephro Net ensures that privacy and security of information passing through the network are strictly enforced. Some major advantages of Nephro Net are listed below:

- Patients can interact with each other discussing their problems, treatments and consequences, and suggestions. As an example, two or more people may have similar problems and take similar treatments but their results might be different. In such cases, the inadequacy in the treatments or the reasons for differences in the results may be identified. Several factors such as age groups, gender and other health conditions (e.g., diabetes) play major roles in this type of discussion. Notice that personal information such as age and gender of a patient will not generally be revealed to other users; it is up to the patient who can disclose personal information in a discussion. For example, a patient may say "I am 65 years old male diagnosed with ... I appreciate if anyone of my age or closer can help me in identifying whether the medication ... will work correctly for me."
- Using Nephro Net, patients will be able to identify and choose better health care facilities and services by exchanging their problems, solutions and ideas. As an example, a patient with Type-2 diabetes undergoing Level-2 dialysis might ask questions related to diabetics. A diabetologist[1] may respond back with more information which is not provided by the nephrologist[2] who treats the patient. Cost of treatment, location of treatment, availability of drugs and a list of healthcare providers are some of the factors generally discussed in this category. The current version of Nephro Net does not store any of these discussion parameters in its database; these parameters will be part of the messages exchanged between patients. Future versions could be enhanced by storing frequently asked questions and information.
- Sometimes a patient may want to have a one-to-one private conversation with a selected physician. For example, a patient may find a nephrologist having more experience than others and has treated a case similar to his/her case. So, the patient may want to discuss his/her case in more detail with this nephrologist. In some other situations, patients may want to have second or third opinions with other specialists individually. A private conversation such as this may involve the patient's personal healthcare information. Nephro Net provides an extra level of security

[1] A physician, usually an internist of endocrinologist, who specializes in the treatment of diabetes mellitus (taken from Dictionary.com)

[2] A physician specialized in Nephorology

Figure 1: Use case diagram for Nephro Net

to ensure that other users of the network cannot access such information.

- Physicians will be able to answer medical questions posted by patients. They will also be able to post educational information for patients. For example, a nephrologist may post a warning message on using a particular drug for patients with diabetes.
- Physicians will also use this forum to discuss multiple cases with other healthcare professionals and educate themselves. It is quite common among healthcare providers discussing common problems, their treatments and results. Medical forums, journals and conferences are all created specifically for this purpose. While these interactions are quite formal, their interactions through Nephro Net are quite informal, short and can be focused on specific cases.
- The role of administrators is very important in Nephro Net. There can be more than one administrator accounts in Nephro Net depending on the number of users and frequency of messages posted. All users' registrations are strictly verified by administrators. In addition, the administrators have the rights to view, block and/or remove most of the information posted on the network. However, administrators will not be able to view passwords, personal profiles of the users if they are set to be private, and will not be able to see any part of a private consultation between a patient and a physician. The authors have specifically designed this feature to support complete privacy of a consultation.
- Security of information is provided by appropriate protection through passwords, encryption/decryption mechanisms and session controls. These will be explained in detail in later sections in the paper.
- Nephro Net provides several features to support privacy of patients' information. However, the patient-centric nature of Nephro Net allows a user to relax some of the constraints enforced on privacy leaving the responsibility to the patients themselves. For example, a patient may reveal his/her complete health records to a physician during a private consultation.

2.1 High Level Requirements of Nephro Net

Users of Nephro Net are classified into three groups - *Patients*, *Physicians* and *Administrators*. Patients and physicians have several functionalities in common such as registration, and searching, viewing and posting messages. Administrators can invoke any functionality that is invoked by a patient or a physician. In addition, there are some additional functionalities which can only be invoked by administrators such as approval or denial of registrations, handling messages, archiving and retrieving messages and so on. Individual functionalities that can be invoked by each type of user are shown in Figure 1. The following assumptions are made while developing Nephro Net:

- Every user must have an account to use Nephro Net which must have been registered before using the system. Administrator accounts are handled separately which is beyond the scope of Nephro Net. Other user accounts must be approved by an administrator to use the system. Accounts may be locked if the user fails to log in within three attempts. Only an administrator can unlock a locked account. An administrator, and only an administrator, can terminate a user account.
- Each user has a short profile describing some basic information about the user (e.g., a physician's short profile provides his/her contact information and specialty). In addition, each user can also create a more detailed profile (called *complete profile*) about him/her. A complete profile of a patient, for example, may include his/her medical problems and the current treatment. More details of these profiles are provided in later sections. While the short profile of a user is used specifically to search for that user, the complete profile of a user can only be viewed with the permission of that user. Consequently, a user can set or revoke the permission to view his/her complete profile at any time.

2.2 Architecture of Nephro Net

A high level architecture of Nephro Net is shown in Figure 2. It uses two different databases. One of them, called *Healthcare Provider database*, is possibly a view created from the database of the healthcare organization that uses Nephro Net. This would consist of all demographic information about the users of Nephro Net. The *Healthcare Provider database* is mainly used at the time of registration of a new user. The registration process and how this database is used at the time of registration are explained in section 3. The second database, called *Nephro Net database*, is created and maintained by the administrators of Nephro Net. This would house all postings and interactions created by the users of Nephro Net. The administrators will take care of archiving messages from this database as and when needed, in order to free up space for new messages.

Each of the three user types has a different user interface portal. This makes it easier to implement and maintain a distinct set of services for each type of user, and at the same time, able to share some common services such as login and setting permissions on complete profiles.

3 Ensuring Security and Privacy

As stated earlier, privacy and security of information were given utmost importance in the design of Nephro Net. This section elaborates the important design decisions considered in the implementation of Nephro Net.

Registration. Every user who wants to join Nephro Net must register with the system by filling up relevant information in a registration form. Appropriate validation checks have been implemented to ensure that all required fields are duly filled in

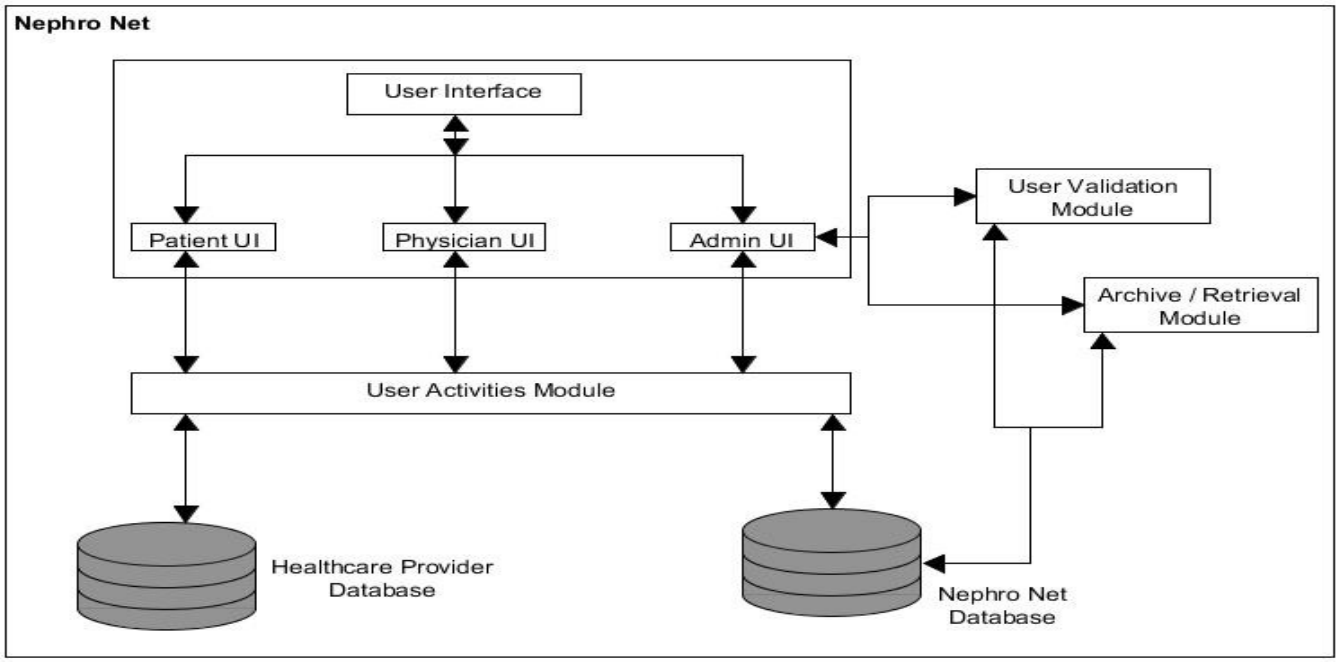


Figure 2: High Level Architecture of Nephro Net

by the registrant. All communication from and to the system will use the email address provided in the registration form. Upon completion, an administrator will evaluate and approve (or deny) the request. If approved, an account will be created for the user and an email is sent to the user with system-created username and an initial password. The user is required to change the password when logging in for the first time.

Nephro Net is expected to be launched within a healthcare organization such as a hospital or clinic. Every user of Nephro Net is expected to be a registered member of the healthcare organization so that when new users register with Nephro Net, the information provided during registration will be verified against the *Healthcare Provider database* of the organization. A new user is required to provide his/her unique ID within the organization which will be used to retrieve information from the *Healthcare Provider database*. This cross verification helps Nephro Net find fraudulent users who may steal IDs of other users and try to register with Nephro Net. In addition, Nephro Net ensures that each user has only one account. For example, if a user tries to get another account with a different name, it will be flagged as an error during the verification process because the user's ID has been already registered with Nephro Net. Once verified, personal information of the new user will be extracted from the *Healthcare Provider database* in order to create a short profile of the user. This short profile can be used to search and/or identify the user by other users of Nephro Net. For example, a patient may want to search a particular physician who has served in the organization for more than five years and is an expert in treating dialysis patients. More details on short profiles are given in a later section.

Authentication. A two-factor authentication mechanism has been implemented to gain access into Nephro Net, even for administrators. That is, in addition to username and password, a user is required to answer to a security question to login into Nephro Net. A user is required to choose three security questions and set up answers for them when logging in for the first time. Thereafter, every time when the user logs in, one security question at a time from the pool of three selected questions by the user will be randomly selected and displayed. If the user answers incorrectly to all the three security questions, the user's account will be blocked. Only an administrator can restore a blocked account. This requires negotiation with the administrator after showing proofs or evidences of valid access requests. This two-factor authentication process is a step forward towards tightening security as expected for a healthcare network. Use of biometric sensors could be another option to consider; however, additional devices and processes are required to implement biometric sensors which will limit the use of Nephro Net by potential users.

Password Encryption. Passwords in Nephro Net are all hashed using a SHA-256 hashing algorithm. Hashing occurs right after a user enters a password for the first time and then the hashed password is stored in the database. By this way, every password is protected even when it is sent to the database.

Sessions. In addition to sessions that are normally used in web-based applications, Nephro Net also uses an additional session's token to increase the level of security. When a user logs in, a session's token is generated and internally stored by

Nephro Net. Thereafter, whenever the user accesses any page in Nephro Net, the session's token is compared to ensure the access rights of the user. The use of session's token prevents a user from logging into Nephro Net more than once at a time.

Health Records. Since Nephro Net provides interactions between patients and healthcare professionals, there will be several messages containing health information of patients. For example, a patient may post a message asking a question about a particular syndrome he/she has. However, Nephro Net does not access the health record of any patient; nor it reveals/displays any health-related information of a patient without the patient himself/herself revealing that information as part of a message. Thus, the patient-centric design of Nephro Net ensures that its users have complete control over their personal information.

User Profiles. Recall that there is a short profile created for each user at the time of registration. In addition, each user can create a complete profile about himself/herself with more details. This profile will be under complete control of the user who created it. By default, no other user in the system has access to another user's profile. To ensure privacy of information, a complete profile may be password protected. That is, the creator of the profile has the option to set a password at the time of creating the profile. Thereafter, every access to the profile, even by the same user, requires authentication using the profile's password. If the user forgets the profile password, the profile can never be accessed. The creator has the option to let other users access his/her profile by giving the profile password. However, such actions are highly discouraged in order to ensure tight security of the system. The complete profile itself is encrypted using a AES-256 encryption algorithm.

Consultation Messages. A patient has the option of exchanging private messages with a physician. These messages are called "consultation messages". Such a private conversation is meant to provide personalized service by a physician to a patient. Consequently, Nephro Net ensures that every message exchanged during consultations is encrypted using a AES-256 encryption algorithm and is not visible or accessible by any other user of the system, including administrators. A patient or physician who had private consultations before can search through the messages that he/she exchanged with at a later time.

4 Operations of Nephro Net

The primary focus of Nephro Net is to provide a forum to exchange information between patients and healthcare professionals in the domain of Nephrology. That is, patients with kidney disease will be able to join Nephro Net, ask questions about kidney related problems, disseminate information regarding their personal experience in dealing with kidney related problems, suggest solutions to other patients who have similar kidney related problems, and direct to appropriate healthcare professionals who provide services to them. Nephrologists will also be able to join Nephro Net, answer to questions posted by other users (both patients and healthcare professionals), post questions on their own, and disseminate information regarding kidney related problems.

Nephro Net has been developed for the domain of Nephrology. As stated earlier, it is preferable to have a healthcare network focusing on one domain. Consequently, it is necessary to ensure that the messages passing through the network are related to the chosen domain, in this case, Nephrology. In the current implementation of Nephro Net, administrators closely watch each message when posted and ensure that it is related to Nephrology. Any message that seems to be irrelevant for the domain will subsequently be removed by the administrators. It may be laborious to watch and ensure that every message is related to the domain, but the process can be automated in the future with some work on natural language parsing. In addition, Nephro Net also provides the option of any user complaining about irrelevancy of a message which will be directed to the administrators. Thus, Nephro Net ensures that only domain-related messages are posted by the users.

Users of Nephro Net can post messages by creating and/or participating in one of the three types of postings - *dissemination*, *discussion* and *consultation*. A posting refers to a structure that contains a type, a topic, the creator's identification and one or more messages (ordered on date/time of creation of these messages). A topic includes a title (user for listing and searching purposes) and a description (explaining the purpose of the topic and what is expected from the messages posted under this topic). Each topic has a unique internal identifier which is not visible to the users. Hence, a user may see the same title for different topics but they are differentiated by the creator, and the date and time of creation of the topic.

A user may be able to browse through the postings and read the messages under a selected topic. The three types of postings are described in detail below:

Dissemination. A dissemination can be created by a patient or a physician. The purpose of a dissemination posting is to share information that is deemed to be useful for other users. It is similar to posting something on a message board in a work place. Every dissemination posting, when created, needs to be approved by an administrator before it is actually posted on Nephro Net. Since no further response is expected for a dissemination posting, there is only one message posted under its topic. A typical example of a dissemination posting is shown in Figure 3.

Discussion. A discussion posting can be created by a patient or a physician. It is meant for exchanging information among several users of Nephro Net, discussing a particular issue or problem. For example, a patient may want to discuss the side effects of one of the medications that he/she takes. So this patient opens up a discussion posting under a topic "Side effects of non-steroidal anti-inflammatory drugs for patients on hemodialysis." Like dissemination posting, a discussion posting must also be approved by an administrator before it is visible to other users. Once approved, any user of Nephro Net can search for the topic of discussion and can also join the discussion by posting messages under the same topic. However, every message in a discussion must be approved by an administrator.

Figure 3: A dissemination posting

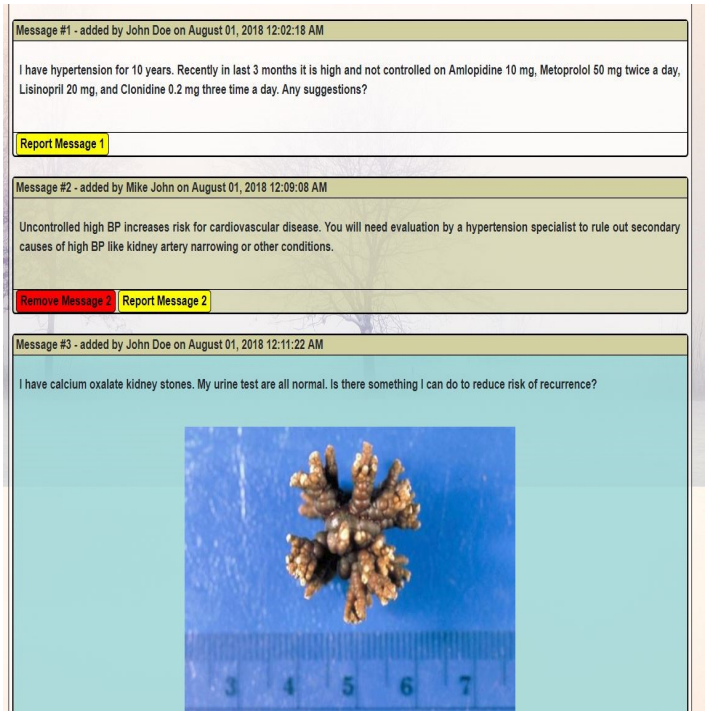


Figure 4: A discussion posting

Figure 4 shows an example of a discussion.

When a discussion starts, any user can post a message under the topic of the discussion. Messages may also include images. It is expected that the users participating in a discussion will only post messages that are relevant to the topic and are useful to other users. If any message seems to be irrelevant to the current topic of discussion, any user can send a complaint on that particular message. Several users may complain about the same irrelevant message. An administrator, after reviewing the complaint(s), will decide whether to keep or remove that message.

During a discussion, the creator of a message can remove the message at any time; an administrator also has the freedom to remove a message.

Only administrators have the ability to terminate a discussion. A discussion, after termination, is still accessible by other users when they search for the topic. However, no new messages can be added to that topic. An administrator may decide to remove the discussion at any time after the discussion is terminated. Administrators will archive removed discussions at their own convenience. If a user asks for that discussion, an administrator can reload the discussion again so that the corresponding topic and its associated messages will appear when searching.

Consultation. Nephro net also provides a mechanism for personal consultation between a patient and a healthcare professional. A consultation is similar in structure to a discussion but is restricted to only two participants. To start with, a patient who wishes to consult with a physician must create a consultation posting (just like creating a new discussion, but must indicate the intention for private consultation.) The posting will already include the identification of both the patient and the physician. An administrator should approve the topic and the consultation. Once approved, both participants will post messages through Nephro Net, just like in any other discussion. However, neither the consultation topic nor its associated messages will be visible to any other user of Nephro Net. Administrators may be able to view the topic of a consultation, but will not be able to view its messages. These messages will be encrypted to ensure privacy. The patient who created the consultation is the only user who can terminate the consultation. After termination, either one of the participants can search for the topic and/or the messages of the consultation at any time later but will not be able to post any new messages.

4.1 Short and Complete Profiles

A short profile of a user will contain name(s), demographic information and some additional information (e.g., specialty of a physician) of that user. The information for a short profile will be extracted from the organization's database during registration. Consequently, a new user, while registering with Nephro Net, is expected to provide his/her identification in the *Healthcare Provider database* from which Nephro Net will extract information for a short profile. A user of Nephro Net may be able to search and/or identify another user by reviewing the short profile of that user. For example, a patient may want to check the specialty of a physician before launching a consultation with the physician.

The complete profile of a user contains more details added

by the user himself/herself. It will be different for the different types of users. For a patient, the profile may include personal and health related information; see Figure 5 for a sample complete profile of a patient. The complete profile of a physician (see Figure 6) may include the current cases he/she is working on, places where he/she worked before and interests of the physician. Figures 5 and 6 only show samples of the profiles. Users can add more fields and additional information, if needed. A complete profile is password protected. That is, when a user creates a complete profile, Nephro Net asks the user to set a password which will be stored along with the profile. This password may be different from the user's login password, but it is encrypted and stored in the same way as the login password. If a password is set for a complete profile, then every access to the profile, even for the creator, requires authentication using this password. In order to ensure privacy, it is highly recommended that Nephro Net users handle complete profiles very carefully and give access to other users only when it is absolutely necessary. For example, a patient may give access to his/her complete profile during a consultation with a physician. It is recommended that the patient changes the password for the profile after the consultation is terminated.

4.2 Administrative Tasks

There can be one or more administrators in Nephro Net. When installed for the first time, Nephro Net will come up with a default administrator account. Thereafter, new administrator accounts can be created through the same registration process as any other user (patient or physician). An administrator is also required to use the three-piece authentication process to login into Nephro Net, as described earlier. The following is a list of tasks that an administrator could perform in Nephro Net:

- An administrator may approve/deny a new user registration. This requires validation of information provided by the user against the healthcare organization's database. Currently, the approval process is performed manually. Some of the tasks in the verification could be automated to ease the load on administrators.
- An administrator may approve/deny the creation of a new posting. The administrator is required to check the relevance of the new posting to Nephrology. Sometimes, a user may request to start a new discussion on a topic for which there is already an on-going discussion in Nephro Net. So, the administrator can point the user to simply join that discussion, instead of creating another discussion. As another example, a user may request to post a dissemination message that has already been posted by another user. In this case, the administrator will point the user to the previously posted dissemination.
- An administrator may review complaints against a posted message in a discussion. As shown in Figure 4, every discussion message comes with a tag titled "Report Message". If a user finds that the message is offensive or not relevant to the topic of current discussion, then the user can complain about the message by pressing this tag. An additional dialog box opens up to include additional information about the complaint. These complaints are stored in a queue for administrators to review at the time of their convenience. If an administrator finds the complaint is valid, then he/she can remove the message from the discussion.
- Apart from the complaints received from users about irrelevant or offensive messages, an administrator may also find a message being irrelevant. In this case, the administrator can once again remove the message from the discussion.
- A discussion can be terminated normally by the user who created it. However, an administrator can also abnormally terminate a discussion based on some situations or decisions. Once a discussion is terminated, no one can post any new messages to the discussion. However, they can still see any or all messages in the discussion.
- It is obvious that the number of users and number of messages will grow and eventually fill up a lot of storage space. So, the administrators have the option of archiving the postings, particularly the discussions. The current implementation only supports manual archival of postings and hence the administrators have to choose the time to archive, select the storage space to archive and consequently remove those postings from Nephro Net database.
- Sometimes a user may want to start a new discussion on a particular topic. But the user would like to know whether there was any discussion(s) before on the same topic which could have been archived. In such situations, the user will contact an administrator who will search for discussion(s) on that topic. If found, the administrator will reload the discussion for everyone to view. Since it is a previously terminated discussion, no new messages can be added to it. If there is a need to continue this discussion, the user should request for a new discussion on the same topic and the administrator will archive the reloaded discussion first and then approve the new discussion.

5 Conclusion and Future Work

There are many challenges in the development of a healthcare social network. Two most important factors to be considered in the development of such a network are security and privacy of information passing through the healthcare network. Unlike social networks such as Facebook and Twitter where security and privacy of information are not strictly monitored, healthcare social networks need additional measures to be included to ensure security and privacy. In the United States of America, HIPAA regulations are strictly enforced in maintaining these two aspects. Other countries also enforce such strict regulations.

This manuscript describes the design and implementation of a healthcare network called Nephro Net with more emphasis on its security and privacy aspects. Though Nephro Net has

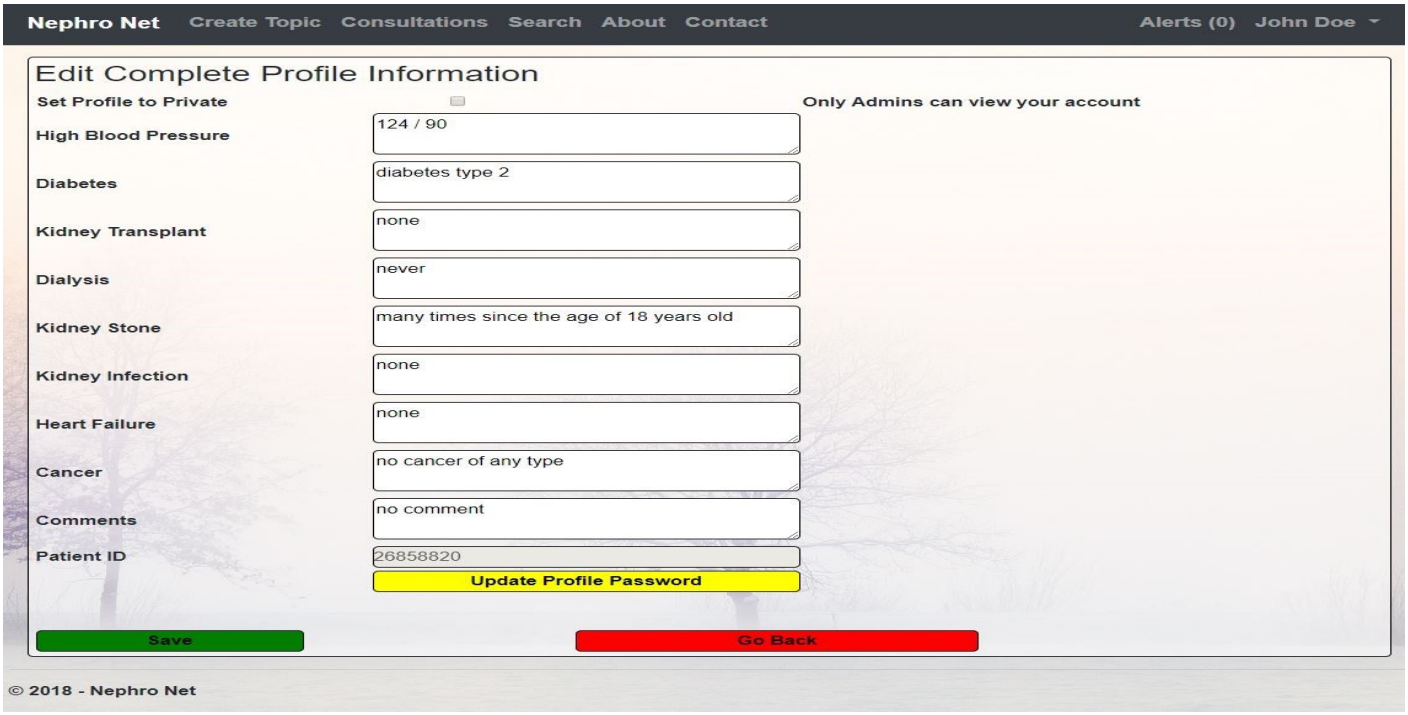


Figure 5: A sample complete profile of a patient

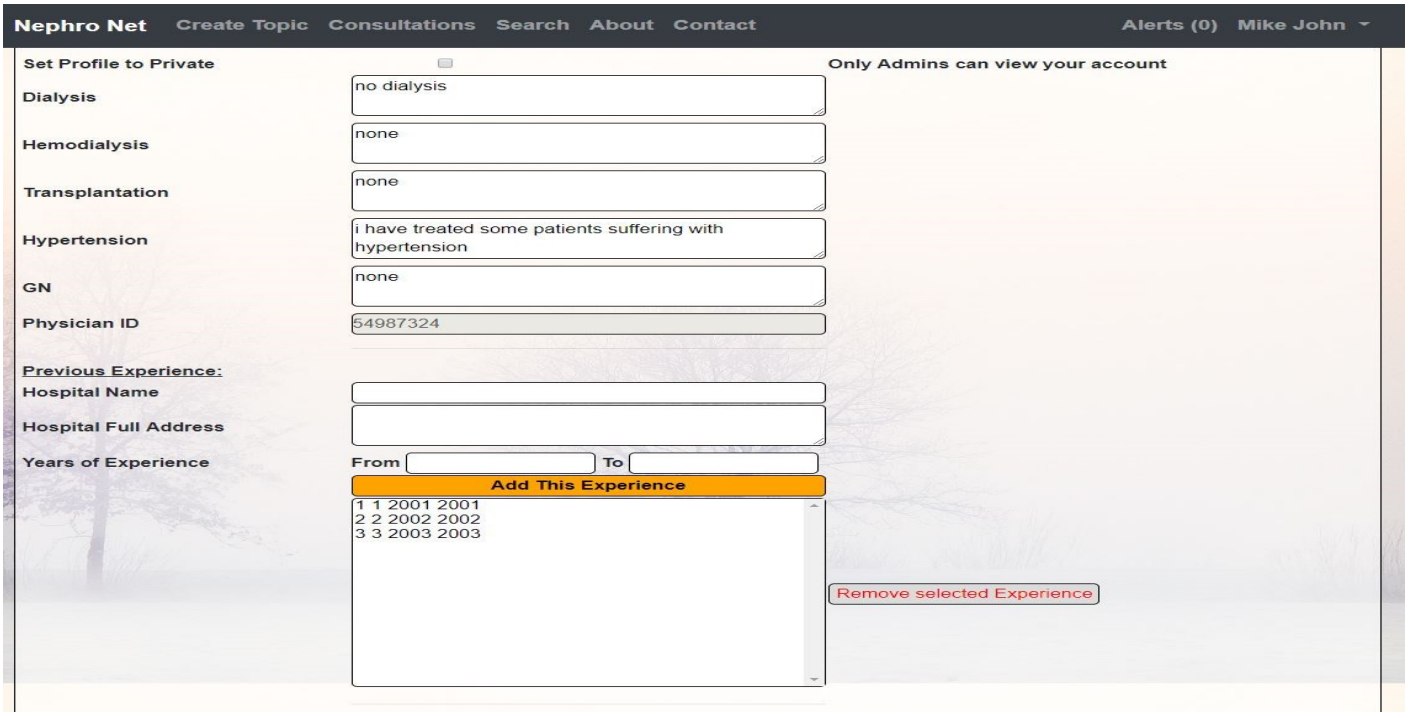


Figure 6: A sample complete profile of a physician

been designed for patients and physicians in the domain of Nephrology, only the case studies and samples have been taken from Nephrology; the design is somewhat generic so that it can be easily tailored to other health domains such as Family Medicine, Psychiatry, Gynecology and so on. It is also evident from a recent survey from Massachusetts Medical Society[18] which shows that healthcare networks for specific domain will be more useful and welcome by the users. Though Nephro Net was tested for its functionalities, it was not yet used in any practical setting and hence was not compared with other healthcare networks.

Below, we enumerate some of the limitations of the current implementation of Nephro Net.

- The current design of Nephro Net enforces strict monitoring of messages by administrators so that every message floating around the network should be approved by an administrator. This may pose performance problems as the number of users and messages increase over time. In order to ease posting of messages and faster responses in discussion postings, this restriction could be relaxed. However, such relaxation can make Nephro Net like any other social network. An alternate solution for this problem is to automate the approval of a message by the administrators. Administrators can still monitor messages and remove them if they are deemed irrelevant. Similarly, at present, verification of credibility of a user is done manually by an administrator and it depends on the data collected during the registration process. Automating this process will also improve the speed of response of the network.
- Social networks expect the users hold ethical responsibilities of proper usage of the network. Raus and others[15] discuss such ethical responsibilities, particularly by patients, when using healthcare networks. Nephro Net is not an exception. As an example, a patient may reveal his/her complete profile to a physician during a consultation by giving the password associated with the profile. The physician is expected not to misuse a patient's personal and/or health information obtained during consultation. Similarly, the patient himself/herself is expected to reset the password after it was viewed by another user. Currently, Nephro Net does not include any additional rules to enforce or monitor ethical responsibilities of its users.
- Nephro Net does not allow commercial users to join and make use of the data available in the network. However, it is possible for someone to steal the data by posing as a patient. Administrators must watch for this kind of abuse while approving registration of a new user.
- The current implementation of Nephro Net supports three types of users - patients, physicians, administrators. It can be easily extended to support more users in the future. For example, the type 'Physician' can be generalized into 'Healthcare providers' to accommodate other types of healthcare service providers such as nurses and physician assistants. However, in practice, the roles of these healthcare providers vary and hence additional features need to be implemented to guarantee adequate security and privacy of healthcare information. To illustrate, consider a patient asking questions about a particular symptom or a disease. A physician, in response, can provide a detailed explanation and relate that information to another case. However, a nurse may not be able to provide such detailed explanation because of his/her knowledge, experience and qualification and hence should not be allowed to respond to such a question from a patient.
- Generally, security aspects involve Confidentiality, Integrity and Availability. It is also known as the CIA triad. This paper focuses more on confidentiality of information. Due to lack of time, the other two security aspects have not been addressed.

Acknowledgments

The authors wish to express their sincere thanks to Dr. Venkateshwaran Iyer, M.D., Nephrology, Mayo Clinic, La Crosse, Wisconsin for his valuable guidance to this project. Almost all information related to Nephrology reported in this manuscript was discussed with him. He also provided help in identifying HIPAA regulations during the implementation.

References

[1] American Health Lawyers Association, `https://www.healthlawyers.org/find-a-resource/alhastore/Pages/default.asps`, 2019, Online; accessed on April 22, 2019.

[2] ASN Communities, `https://community.asn-online.org`, Online; accessed on April 29, 2019.

[3] Bassette Healthcare Network, `https://www.bassett.org/education/medical-education/medical-student-programs/electives/nephorology`, Online; accessed on April 29, 2019.

[4] Doximity - a Healthcare Network, `https://www.doximity.com`, 2018, Online; accessed on October 10, 2018.

[5] L. Fass, "Patient-centric Healthcare", *Third IET International Conference on Medical Electrical Devices and Technology (MEDTECH 2007)*, London, UK, pp 77–109, October 2007.

[6] First Opinion Healthcare Network, `https://firstopinionapp.com`, 2019, Online; accessed on July 01, 2019.

[7] F. Griffiths, J. Cave, F. Boardman, J. Ren, T. Pawlikowska, R. Ball, A. Clarke and A. Cohen, "Social Networks: The Future for Healthcare Delivery", *Social Science & Medicine*, 75:2233–2241, 2012.

[8] Kidney Health Australia, https://kidney.org.au, Online; accessed on April 29, 2019.

[9] J. Li, "Privacy Policies for Healthcare Social Networking Sites", *Journal of the American Medical Informatics Association*, doi: 10.1136/amiajnl-2012-001500, 20:704–707, 2013.

[10] Mercy Health Network, https://www.mercyhealthnetwork.com, 2018, Online; accessed on October 10, 2018.

[11] S. A. Moorhead, "Social Media for Healthcare Communication", *Oxford Research Encyclopedia of Communication*, doi: 10.1093/acrefore/9780190228613.013.335, August 2017.

[12] Patients Like Me - a Healthcare Network, https://www.patientslikeme.com, 2018, Online; accessed on October 01, 2018.

[13] K. Periyasamy and S. Alsyefi, "Implementation of Security and Privacy Aspects in a Healthcare Social Network", *Proceedings of the 34th International Conference on Computers and Their Applications (CATA 2019)*, (ed.) G. Lee and Y. Jin, Honolulu, Hawaii, doi: https://doi.org/10.29007/d1s9, pp 210–218, March 2019.

[14] Quantia MD, https://en.wikipedia.org/wiki/QuantiaMD, 2018, Online; accessed on October 10, 2018.

[15] K. Raus, E. Mortier and K. Eeckloo, "The Patient Perspective in Health Care Networks", *BMC Medical Ethics*, doi: 10.1186/s12910-018-0298-x, 19(52), 2018.

[16] Sermo - Talk Real World Medicine, https://www.sermo.com, 2018, Online; accessed on October 10, 2018.

[17] Triad Healthcare Network, https://www.triadhealthcarenetwork.com, 2018, Online; accessed on October 10, 2018.

[18] K. G. Volpp and N. S. Mohta, "Patient Engagement Survey: Social Networks to Improve Patient Health", *NEJM Catalyst*, Massachusetts Medical Society, http://catalyst.nejm.org, 2017.

Kasi Periyasamy is a Professor and Program Director for the Master of Software Engineering (MSE) program in the Department of Computer Science at University of Wisconsin-La Crosse (UW-L), Wisconsin since 1999. Prior to joining UW-L, he was in the Department of Computer Science at University of Manitoba, Winnipeg, Canada for eight years. Dr. Periyasamy has published numerous papers in software engineering and had been the keynote speaker in some international conferences. His primary research interests include formal specifications, verification and validation, and software project management. He is the co-author of the book *Specification of Software Systems*; the first edition of the book was published in 1998 and the second edition in 2011, both by Springer-Verlag. He also contributed to a chapter in a book on teaching software project management skills; the book title is *Overcoming Challenges in Software Engineering Education*, Liguo Yu (Ed.), published by IGI Global in Spring 2014.

Saleh A. Alsyefi graduated with a Bachelors degree in Computer Information Systems from Wilkes University in 2011, and with a Masters degree in Software Engineering from University of Wisconsin-La Crosse in 2018. He was working as an IT Projects Manager at Royal Saudi Land Forces in Saudi Arabia from 2012 until now. He is one of the two authors of a paper published in the International Conference on Computers and Their Applications (CATA 2019) which was nominated for Best Paper Award.

Locality-Aware CTA Mapping for GPUs

Lifeng Liu, Meilin Liu
Wright State University, Dayton, Ohio, U.S.A.

Chongjun Wang
Nanjing University, Nanjing, China

Abstract

General purpose GPUs (GPGPUs) have been used as the platform for many scientific applications with thread-level parallelism due to their powerful computation ability and massively parallel features. In this paper, we design a locality-aware CTA (cooperative thread array) mapping scheme based on polyhedron model for GPUs to take advantage of the inter-CTA data reuses in the GPU kernels. Using the data reuse analysis based on the polyhedron model, we can detect inter-CTA data reuse patterns in the GPU kernels and control the CTA mapping pattern to improve the data locality on each SM (streaming multiprocessor). The locality-aware CTA mapping scheme based on polyhedron model for GPUs can also be combined with the programmable warp scheduler to further improve the performance. The experimental results show that our CTA mapping algorithm can improve the overall performance of the input GPU programs by 23.3% on average and by 56.7% when combined with the programmable warp scheduler. In addition, the overall L1 cache misses can be reduced by 18.9% by the CTA mapping algorithm and 29.3% when the CTA mapping algorithm is combined with the programmable warp scheduler.

Key Words: GPU, compiler, polyhedron model, data locality, CTA mapping.

1 Introduction

As the power consumption and chip cooling technology are limiting the frequency increase of the single core CPUs, multi-core and many-core processors have become the major trend of computer systems[1, 10, 14, 22, 7]. General purpose GPU (GPGPU) is an effective many-core architecture for scientific computations by putting hundreds or even thousands of stream processor cores into one chip [10, 19, 3, 13, 12]. Compared to the traditional CPUs such as Intel X86 serial CPUs, GPGPUs have significant advantages for many scientific applications with thread-level parallelism [10, 19].

Generally speaking, GPUs provide significant computation speed improvement for parallel applications as the co-processor of traditional CPUs. In addition, the development of GPU programming interfaces such as CUDA and OpenCL makes the programming on GPUs much easier [10, 7]. More and more developers have ported their applications from the traditional CPU based platforms to GPU platforms [5, 11, 17, 6]. However, several challenges still limit further performance improvements and make GPU programming challenging for programmers who lack the knowledge of GPU hardware architecture.

Limited by the current DRAM technologies, accessing off-chip memory is very time consuming. In traditional CPUs, large L1 and L2 caches play an important role in bridging the speed gap between CPU cores and the off-chip DRAM memory. Both L1 and L2 caches are managed automatically by the hardware and are transparent to programmers. Programmers only need to consider a uniform continuous memory space. However, the memory hierarchy on GPU platforms is different. On each SM core of a GPU chip, a small high-speed shared memory is designed to cache a small scale of frequently used data and the data shared among different threads in the same thread block.

When an SM has available CTA slots, a thread block will be selected and mapped to the SM in a round-robin manner [16]. The selection process does not consider the possible data locality or data reuses among the thread blocks mapped to the SM. Figure 1 shows a possible mapping result of a general scenario for 1D applications. Assume the inter-CTA data locality exists among consecutive CTAs, the CTA mapping scheme shown in Figure 1 breaks the data locality, which increases the L1 cache footprint and degrades the overall performance.

The same problem can be observed for the 2D applications too. Assuming we have four SMs and a grid of 5×5 thread blocks. The thread block mapping pattern with the original mapping strategy will break inter-thread block data reuses along the x direction or the y direction as illustrated in Figure 2(a). The inter-thread block data reuses can be preserved if we map the CTAs along the x direction as illustrated in Figure 2(b) because the CTAs having inter-warp data locality are all mapped to the same SM. Similarly, we can preserve the inter-thread block data reuses along the y direction by mapping the CTAs along the y direction as illustrated in Figure 2(c).

Many research works show that warp scheduling techniques improve the resource utilization and hide the long latency memory operations. In Chapter 4 of the dissertation [18], we

analyze the intra-warp and inter-warp data reuses in the L1 data cache. The analysis is based on an assumption that the number of the warps is limited in the same thread block. However, to the best of our knowledge, few research works have studied the impact of CTA mapping on resource utilization on GPUs. In addition, very few research works have investigated the impact of combining the CTA mapping with warp scheduling.

In real GPU systems the threads are grouped and executed as thread blocks or CTAs. The best size of the high priority warp group can be larger than the number of warps in a single thread block. If we prioritize the warps from different thread blocks without considering the data locality among them, the data locality might be broken and the overall performance would be degraded. To construct the high priority warp groups with the warps coming from the thread blocks with inter-CTA data reuses we must control the CTA mapping pattern according to the data reuse pattern among the thread blocks in the same kernel grid.

In [16], Lee et al. proposed a block CTA scheduling algorithm to preserve the inter-CTA locality. The block CTA scheduling algorithm can assign consecutive CTAs along the x direction to the same SM. However, without the inter-CTA reuse pattern detection mechanism, the algorithm cannot handle the inter-CTA reuses along the y direction. In addition, the algorithm cannot handle 2D applications.

In this paper, we present a locality-aware CTA mapping scheme based on polyhedron model for GPUs that can enable the users to set the CTA mapping scheme before a GPU kernel is launched. The locality-aware CTA mapping scheme based on the polyhedron model can detect and control the CTA mapping pattern automatically. Then, we combine the CTA mapping scheme with the compiler-assisted programmable warp scheduler to further improve the performance of the GPU kernels.

The rest of this paper is organized as follows: in Section 3, we present the compiler-assisted locality-aware CTA mapping scheme to detect the CTA mapping pattern. In Section 4, we illustrate how to combine the compiler-assisted CTA mapping scheme with the programmable warp scheduler while Section 5 discusses the issue of balancing CTAs among streaming multiprocessors. In Section 6, we evaluate the locality-aware CTA mapping scheme and the performance of the input benchmarks when they are optimized by the locality-aware CTA mapping scheme combined with the programmable warp scheduler. Related works are provided in Section 7. We conclude this paper in Section 8.

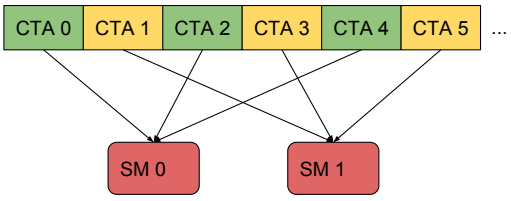


Figure 1: The default CTA mapping for 1D applications

2 Basic Concepts

In this section, we present the basic concepts used in this paper [18].

2.1 The Overview of GPGPUs

As illustrated in Figure 3, GPGPUs [10, 9, 4] consist of several Streaming Multiprocessors (SMs), each of which has multiple streaming processors (SPs). The GPU kernel programs and the data are transferred into the global memory of the GPUs by the host CPU through the PCI-e bus. The global memory usually uses the off-chip Double Data Rate (GDDR) DRAMs. In order to achieve the high global memory access bandwidth, several memory access controllers are introduced, each of which has an on-chip L2 data cache. Those SMs and memory access controllers are connected together by the on-chip interconnection network. The parallel tasks loaded on a GPU system will be managed by the execution manager.

2.2 The CTA mapping

Threads belonging to different thread blocks (CTAs) can be executed independently. Multiple CTAs can be assigned and executed concurrently on one SM. During the execution of a GPU kernel, the maximum number of CTAs that can be mapped to an SM is limited by the hardware resources occupied by each CTA and the hardware parameters. The relevant hardware resources are the register usage and the shared memory usage. The maximum number of CTAs mapped to an SM can be calculated as follows [10, 9]:

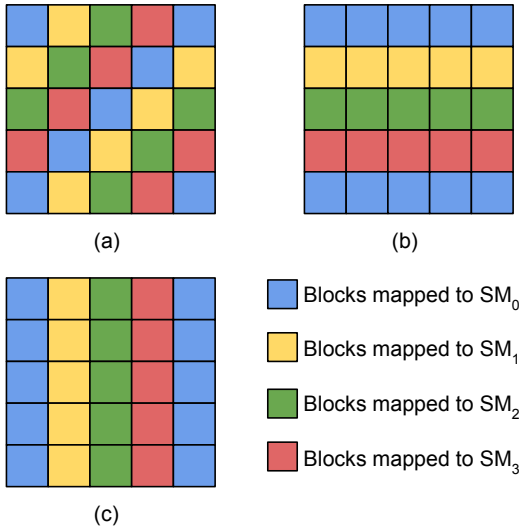


Figure 2: The CTA mapping.(a)Original mapping.(b)Mapping along the x direction (c)Mapping along the y direction

$$\#CTAs = min(\frac{total\#regs}{\#regsPerCTA}, \frac{totalSharedMem}{SharedMemPerCTA}, \#CTASlots) \quad (1)$$

For Nvidia Fermi and Kepler GPUs, the maximum number of hardware CTA slots is eight [8]. New thread blocks will be assigned to an SM when a thread block on this SM terminates. The default thread block is selected in a round-robin manner [16, 10] as illustrated in Figure 1.

3 The CTA Mapping Pattern Detection

We design a CTA mapping control API to modify the CTA mapping strategy before a kernel is launched. The API will modify the value of a special register that controls the CTA mapping unit of each SM core. For performance-complexity trade-off, we only consider the three most common CTA mapping strategies: mapping along the x direction, mapping along the y direction and mapping in the round-robin manner. The default CTA mapping strategy is the round-robin mapping strategy.

Based on the CTA mapping API, we design a compiler-assisted locality-aware CTA mapping scheme to insert the CTA mapping control APIs automatically based on the inter-CTA data reuse analysis. The basic rules of the compiler-assisted locality-aware CTA mapping scheme is:

1. If there are inter-thread block data reuses along the x direction, then the CTAs are mapped along the x direction.
2. If there are inter-thread block data reuses along the y direction, then the CTAs are mapped along the y direction.
3. If there are inter-thread block data reuses along both the x and the y directions, the CTAs are mapped along the direction that has larger reuse distance as the memory blocks can have more reuses along the direction having larger data reuse distance. For example, if the memory blocks accessed by a thread block can be reused by the following N thread blocks, then these memory blocks can be reused $N-1$ times during the execution of the GPU kernel. So the larger the N is, the more times the memory blocks can be reused.
4. Otherwise, use the round-robin CTA mapping strategy.

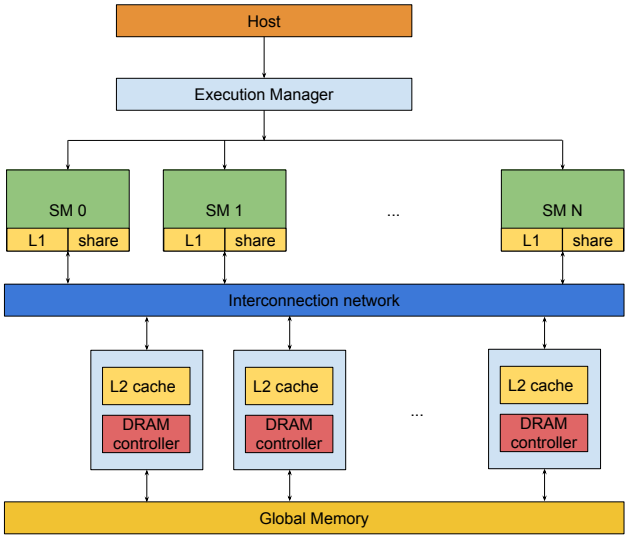


Figure 3: The basic architectures of GPGPUs [10]

The inter-thread block data reuses along the x direction can be formally defined as

Inter-CTA data reuses along the x direction During the execution process of a thread block, if there exists two memory accesses M and M' that meet all of the following conditions, then inter-CTA data reuses exist among the thread blocks along the x direction.

1. M and M' are issued by different thread blocks with the same thread block index in the y direction.
2. M' reuses the data in the L1 cache brought in by M.

The first constraint can be represented as

$$\begin{cases} 0bidy, bidy' < gdimy \\ 0bidx, bidx' < gdimx \\ 0tidy, tidy' < bdimy \\ 0tidx, tidx' < bdimx \\ bidy' = bidy \end{cases} \quad (2)$$

The second constraint can be represented as

$$\begin{cases} \vec{\alpha'}_i = \vec{\alpha}_i, i = 1...m-2 \\ \vec{\alpha'}_{m-1} = \beta' * B + \gamma' \\ \vec{\alpha}_{m-1} = \beta * B + \gamma \\ \beta' = \beta \\ -B + 1\gamma' - \gamma B - 1 \\ 0\gamma' B - 1 \\ 0\gamma B - 1 \\ 0\beta' \\ 0\beta \end{cases} \quad (3)$$

Where B indicates the cache block size; β' and β indicate the cache block indexes of M' and M; γ' and γ indicate the offsets inside each cache block.

The target parameter that we are interested in is the reuse distance along the x direction:

$$\zeta_x = |bidx' - bidx| \quad (4)$$

Then, the problem of detecting inter-thread block data reuses can be transformed into an integer linear programming problem:

$$\begin{array}{ll} max & \zeta_x \\ s.t. & \zeta_x = |bidx' - bidx| \\ & bidx', bidx \in G_x \end{array} \quad (5)$$

Where G_x is the polyhedron defined by (2) and (3). If the problem (5) has no solution or $\zeta_x = 0$, then it indicates that there is no inter-thread block data reuses along the x direction.

Similarly, we can define inter-CTA data reuses along the y direction as **Inter-CTA data reuses along the y direction**. During the execution process of a thread block, if there exists two memory accesses M and M' that meet all of the following conditions, then inter-CTA data reuses exist among the thread blocks along the y direction.

1. M and M' are issued by different thread blocks with the same thread block index in the x direction.
2. M' reuses the data in the L1 cache brought in by M.

The first constraint is represented as

$$\begin{cases} 0 \le bidy, bidy' < gdimy \\ 0 \le bidx, bidx' < gdimx \\ 0 \le tidy, tidy' < bdimy \\ 0 \le tidx, tidx' < bdimx \\ bidx' = bidx \end{cases} \tag{6}$$

The second constraint is the same as (3). We can obtain the maximum inter-thread block reuse distance along the y direction ζ_y in a similar way:

$$\begin{array}{ll} max & \zeta_y \\ s.t. & \zeta_y = |bidy' - bidy| \\ & bidy', bidy \in G_y \end{array} \tag{7}$$

Where G_x is the polyhedron defined by (6) and (3).

Algorithm 1: Detect the CTA mapping direction

Input: GPU kernel G
Output: *direction*, *maxResueDistance*

```
1  max_x = 0;
2  max_y = 0;
3  for each memory access pair M and M' in G do
4  |   Get the reuse distance r_x along the x direction by
   |   solving problem 5;
5  |   Get the reuse distance r_y along the y direction by
   |   solving problem 7;
   |   max_x = max(max_x, r_x);
   |   max_y = max(max_y, r_y);
   end
6  if max_x==0 and max_y==0 then
7  |   direction = round-robin;
   else
8  |   direction = max_y>max_x ? y : x;
9  |   maxResueDistance=max(max_y, max_x);
   end
```

Algorithm 1 is presented to detect the CTA mapping direction. As illustrated in Algorithm 1, if there are multiple inter-CTA data reuses existing in a single GPU kernel, we record the maximum inter-CTA reuse distance along the x direction and the y direction separately (Line $1-5$). When there are inter-thread block data reuses along both the x direction and the y direction, the CTAs are mapped along the direction that has larger reuse distance (Line $6-9$). Otherwise, the round-robin CTA mapping strategy would be used.

4 Combining the Programmable Warp Scheduler and the Locality-Aware CTA Mapping Scheme

To take advantage of the inter-thread block data locality and inter-warp data locality simultaneously, we combine the compiler-assisted programmable warp scheduler and the compiler-assisted locality-aware CTA mapping scheme as illustrated in Algorithm 2.

Algorithm 2: Combine the CTA mapping scheme and the programmable warp scheduler

Input: GPU kernel G
Output: Optimized GPU kernel with both CTA mapping and programmable warp scheduler applied

```
1  Get CTA mapping direction through Algorithm 1;
2  Insert CTA mapping control API before G is launched;
3  for each loop L in G do
4  |   Apply the Programmable Warp scheduling
   |   Algorithm [18] on L with the constrains modified by
   |   (8) and (9) to insert scheduler control instructions for L.
   end
```

The compiler-assisted locality-aware CTA mapping scheme will not affect the intra-warp data reuse detection algorithm. However, we must modify the inter-warp data reuse detection algorithm and the high priority warp group size detection algorithm. The constraints in the Programmable Warp scheduling Algorithm [18] are modified by the extra constraints presented in (8) and (9). Also, we must remove the constraints of $w, w' < ceiling(bdimx * bdimy/32)$ because the warp IDs could exceed the thread block boundary. In addition, the check condition in Line 14 of the Programmable Warp scheduling Algorithm [18] must be modified to "$\min(\xi_{min}, \theta_{min}) < \#warpsInBlock * maxResueDistance$".

$$\begin{cases} bidy' = bidy & \text{If CTAs are mapped along the } x \text{ direction} \\ bidx' = bidx & \text{If CTAs are mapped along the } y \text{ direction} \end{cases} \tag{8}$$

$$\begin{cases} 32wbdimx * tidy + tidx + (gdimx * bidy + bidx) * wpb < 32(w+1) \\ \quad \text{If CTAs are mapped along the } x \text{ direction} \\ \quad \text{or in the round-robin manner} \\ 32wbdimx * tidy + tidx + (gdimy * bidx + bidy) * wpb < 32(w+1) \\ \quad \text{If CTAs are mapped along the } y \text{ direction} \end{cases} \tag{9}$$

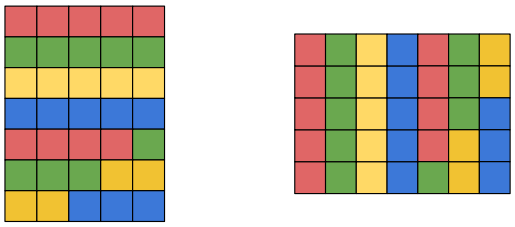


Figure 4: Balance the CTAs among SMs when mapping along the x direction (a) and the y direction (b)

Where $wpb = ceilling(bdimx * bdimy/32)$, which is the number of warps per thread block. The warp constraints for w' is modified in the same way.

5 Balancing the CTAs Among SMs

Algorithm 3: CTA mapping

```
Input: gdimx, gdimy, #sms, direction
Output: the CTAs mapped to each SM
1  if direction == x then
2      a = gdimy / #sms;
3      b = a * gdimx;
4      c = (gdimy - a) * gdimx;
5      d = (c + 1) / #sms + 1;
6      for each B = CTA(x, y) do
7          if y < b then
8              Assign B to SM core y / a;
           end
9          e = (x + y * gdimx) - b;
10         Assign B to the SM core e / d;
       end
   end
11 else if direction == y then
12     a = gdimx / #sms;
13     b = a * gdimy;
14     c = (gdimx - a) * gdimy;
15     d = (c + 1) / #sms + 1;
16     for each B = CTA(x, y) do
17         if x < b then
18             Assign B to the SM core x / a;
           end
19         e = (y + x * gdimy) - b;
20         Assign B to the SM core e / d;
       end
   end
21 else
22     assign the CTAs to the SMs in the round-robin manner;
   end
```

We modify the CTA mapping units in GPGPU-sim to evaluate our compiler-assisted locality-aware CTA mapping algorithm. The CTA mapping unit can be configured to map CTAs along the x direction, along the y direction or using the round-robin manner. The CTA mapping configuration can be set by the CTA mapping control API we developed, which is supported by the GPGPU-sim modified by us. The CTA mapping control API can be executed before the GPU kernel launch. To balance the number of CTAs assigned to an SM, and preserve the inter-CTA data reuses as much as possible when the mapping direction is selected, we use the following rules

1. Assign a whole row or column of CTAs to an SM as long as we can evenly distribute them among the SMs.
2. For the rest of the CTAs, we evenly distribute them along the x direction or the y direction.

The CTA mapping algorithm that follows these two rules is illustrated in Algorithm 3. In Algorithm 3, a, b, c, d and e are intermediate variables, *direction* is the CTA mapping direction that is detected by Algorithm 1. Figure 4 shows the mapping results of Algorithm 3 with $\#sms = 4$ for mapping along both the x direction and the y direction.

6 Evaluation

6.1 Evaluation Platform

We configure the GPGPU-sim developed by Aamodt et al. [2, 3] to simulate the GTX480 GPU as the test platform to evaluate our CTA mapping technique for GPGPUs. The detailed configuration for the baseline GPU system is shown in Table 1. We use NVCC 4.2 to compile the output source code of our source-to-source compiler framework.

The benchmarks we selected to evaluate our CTA mapping algorithm are listed as follows:

(1) micro-benchmark: A simple GPU kernel designed to add neighboring blocks together. We can see that the same block in input will be reused among CTAs along the x direction.
(2) matmul: Block matrix multiplication kernel, whose input size is 4kx4k.
(3) conv: Two dimensional convolution algorithm, whose input size is 4kx4k.
(4) demosaic: Image demosaicing algorithm, whose input size is 8kx8k.
(5) imregionmax:The algorithm to find the regional maximum for an image, whose input size is 4kx4k.
(6) tr: Matrix transpose algorithm, whose input size is 4kx4k.
(7) mum: Parallel local-sequence alignment program [3], whose input size is 50k.
(8) lps: A Laplace discretization on a 3D structured grid [3], whose input size is 4M.

Table 1: The baseline simulator configuration

Module	Description
Number of SMs	15 (as Nvidia GTX480)
Number of integer processing units per SM	2
Number of floating point processing units per SM	1
SIMD width	32
Size of L1 data cache	16KB, 4 way associative
Size of L2 data cache	512KB, 8 way associative
Size of shared memory	48KB

6.2 Experimental Results

To evaluate our compiler-assisted locality-aware CTA mapping algorithm, we measure the performance of the GPU kernels optimized by the compiler-assisted locality-aware CTA mapping algorithm (labeled as "cm" in Figure 5) and the performance of the GPU kernels optimized by the CTA mapping algorithm combined with the programmable warp scheduler discussed in Chapter 4 of the dissertation [18] (labeled as "cm+ps" in Figure 5). We compare the performance of the GPU kernels optimized by our compiler-assisted locality-aware CTA mapping algorithm with the performance of the GPU kernels optimized by the GTO warp scheduler, which is used as the baseline, the performance of the GPU kernels optimized by the CCWS warp scheduling algorithm proposed by Rogers et al. [21] and the performance of the GPU kernels optimized by the compiler-assisted programmable warp scheduler (labeled as "ps" in Figure 5).

The performance results of the GPU kernels are summarized in Figure 5 and the corresponding L1 cache miss rates of the GPU kernels are reported in Figure 6. The higher the speedup and the lower the L1 cache miss rate, the better the overall performance is. In Figure 5, the first column shows the normalized performance of the input benchmarks optimized by the GTO algorithm. The second column shows the normalized speedups of the input benchmarks optimized by the CCWS warp scheduling algorithm. The third column shows the normalized speedups of the input benchmarks optimized by the compiler-assisted programmable warp scheduler. The fourth column shows the normalized speedups of the input benchmarks optimized by the compiler-assisted CTA mapping algorithm. The fifth column shows the normalized speedups of the input benchmarks optimized by the compiler-assisted CTA mapping algorithm combined with the programmable warp scheduler.

In Figure 6, the first column shows the normalized L1 cache miss rates of the input benchmarks optimized by the GTO algorithm. The second column shows the normalized L1 cache miss rates of the input benchmarks optimized by the CCWS warp scheduling algorithm. The third column shows the L1 cache miss rates of the input benchmarks optimized by the compiler-assisted programmable warp scheduler. The fourth column shows the L1 cache miss rates of the input benchmarks optimized by the compiler-assisted CTA mapping algorithm. The fifth column shows the L1 cache miss rates of the input benchmarks optimized by the compiler-assisted CTA mapping algorithm combined with the programmable warp scheduler.

As illustrated in Figure 5, our compiler-assisted locality-aware CTA mapping algorithm can improve the performance of the selected benchmarks by 23.3% on average over the GTO warp scheduling algorithm. The GTO warp scheduler just considered the locality within each CTA, however, the inter-thread block data reuses are ignored due to the round-robin CTA mapping strategy that always assigns adjacent CTAs to different SMs. According to Figure 6, our CTA mapping algorithm can reduce the L1 cache miss rates of the selected benchmarks by 18.9% on average compared to the GTO algorithm.

Compared to the GTO warp scheduling algorithm, the CCWS warp scheduling algorithm and the compiler-assisted programmable warp scheduler can avoid self-evictions in the L1 cache and preserve intra-warp and inter-warp data reuses. The CCWS warp scheduling algorithm and compiler-assisted programmable warp scheduler improve the overall performance of the input benchmarks by 34.1% and 35.0% respectively on average over the GTO warp scheduling algorithm. However, The CCWS warp scheduling algorithm and the compiler-assisted programmable warp scheduler do not consider the data reuses among different CTAs, so they cannot further exploit the data reuses in the L1 cache. In Figure 6, we can see that the L1 cache miss rates of the input benchmarks are reduced by the CCWS warp scheduling algorithm and the compiler-assisted programmable warp scheduler by 12.0% and 12.7% respectively on average, which are smaller than the L1 cache miss rates decreased by the compiler-assisted locality-aware CTA mapping algorithm.

The performance of the input benchmarks is improved by 56.7% on average when our compiler-assisted CTA mapping algorithm is combined with the compiler-assisted programmable warp scheduler, since the self-evictions in the L1 cache are avoided by limiting the total number of active warps and taking advantage of the inter-CTA data reuses. In addition, our compiler-assisted locality-aware CTA mapping algorithm can construct high priority warp groups across the CTA boundaries, which can completely hide long latency operations by feeding the warp scheduler with enough warps and taking advantage of the high data reuses in the L1 cache. This conclusion can also be supported by the reduced L1 cache miss rates of the input benchmarks optimized by the compiler-assisted CTA mapping algorithm combined with the

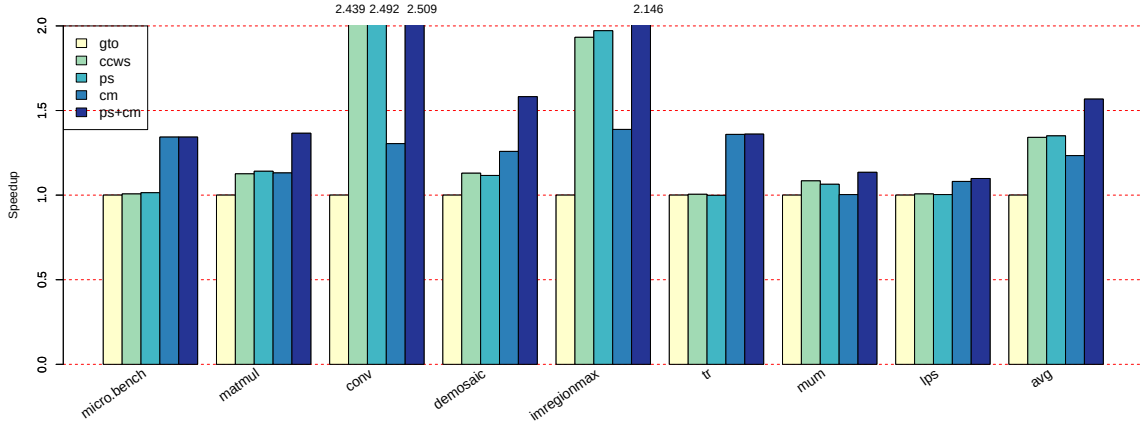


Figure 5: Speedups

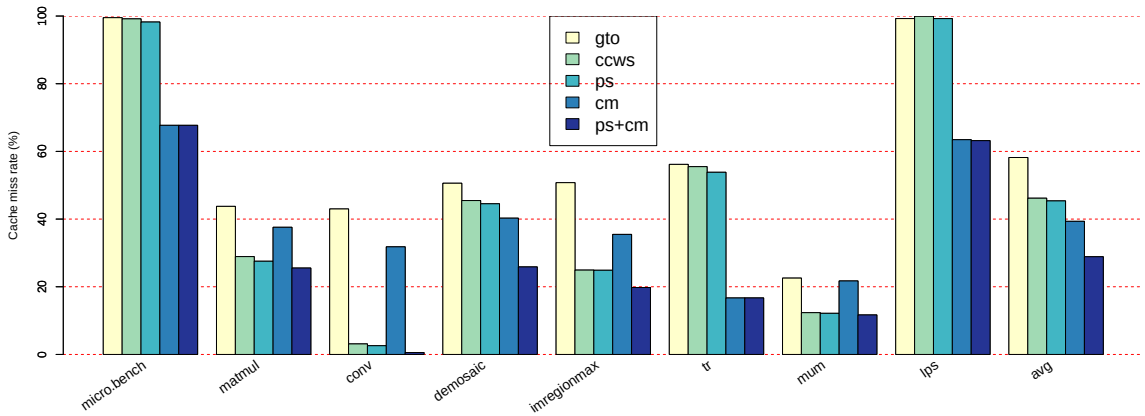


Figure 6: The L1 cache miss rates

programmable warp scheduler.

The benchmarks matmul, conv, demosaic, imregionmax and mum have both good intra-CTA and inter-CTA data locality, so the compiler-assisted programmable warp scheduler contributes most to the performance improvements (64% for matmul, 79% for conv, 55% for demosaic, 66% for imregionmax and 97% for mum) for these input benchmarks. While the rest of the benchmarks mainly obtained performance gain from the inter-CTA data reuses (100% for micro_bench, 99% for tr and 82% for lps). Both the compiler-assisted programmable warp scheduler and the CCWS warp scheduler cannot further improve the performance of these input benchmarks.

From Figure 5 and Figure 6, we can also see that the performance gained from the L1 cache optimization is also different. For example, the combined scheme (the compiler-assisted CTA mapping algorithm combined with the programmable warp scheduler) reduces the L1 cache misses of the input benchmark demosaic by 25%. In contrast, the combined scheme reduces the L1 cache misses of the input benchmark lps by 36%. However, the performance of demosaic is increased by the combined scheme by 50%, higher than lps. The reason is that the ratios between the calculation instructions and the memory access instructions of different benchmarks are different. When the memory accesses of the benchmarks are the bottleneck, such as in the benchmarks conv, demosaic, imregionmax and micro_bench, significant performance improvement can be gained by reducing the L1 cache miss rates. The performance improvement of the input benchmarks can achieve 89.5% on average from the L1 cache miss reduction of 32.2%. However, for the rest of the benchmarks, the performance improvement can only reach 23.9% from the L1 cache miss reduction of 23.9%.

7 Related Work

In [16], Lee et al. proposed two optimized CTA scheduling algorithms to improve the performance of the input GPU kernels. The lazy CTA scheduling method can limit the total number of active CTAs assigned to each SM to improve the L1 cache hit ratio and reduce the resource competition, which is similar to the idea of the warp limiting algorithm introduced by the CCWS warp scheduler [21] when applied on thread blocks. They also proposed the block CTA scheduling strategy to take advantage of the data locality among different thread blocks. Compared to their work, we propose a compiler-assisted locality-aware CTA mapping algorithm which uses

a systematic way to control the CTA mapping pattern more accurately using the data reuse analysis based on the polyhedron model. In addition, the CTA limiting can also be performed in our framework by combining the compiler-assisted locality-aware CTA mapping algorithm with the programmable warp scheduler, which can tune the number of active warps across the CTA boundaries.

Our compiler-assisted CTA mapping algorithm is also combined with the compiler-assisted programmable warp scheduler to exploit inter-CTA data reuses in addition to the intra-warp data reuses and the inter-warp data reuses, which exhibits more data reuses compared to the compiler-assisted programmable warp scheduling algorithms, such as the CCWS warp scheduling algorithm proposed by Rogers et al. [21] and the two-level warp scheduling algorithm proposed by Narasiman et al. [20]. Both of these two warp scheduling algorithms considered the data locality and the resource competition in the L1 cache among different warps on an SM, which can improve the overall performance by limiting the number of concurrent warps. However, just as we analyzed in this paper, random CTA assignment to SMs might break the data locality and increase the L1 cache thrashing. Our proposed compiler-assisted CTA mapping scheme overcomes this drawback and further improves the overall performance.

In [15], the consecutive CTA mapping strategy is used to facilitate the design of optimized GPU memory prefetcher based on the two level warp scheduler proposed in [20]. Their prefetcher takes advantage of the spatial locality among consecutive CTAs assigned to each SM. However, only the benefit obtained from this CTA mapping strategy is analyzed and no inter-CTA data reuse analysis has been performed.

Yang et al. proposed a task scheduling algorithm considering data reuses in the cache, memory footprint and cache coherence for chip-multiprocessor systems [23]. They group tasks that have the maximum amount of data sharing into sharing groups. These tasks will be assigned to the same CPU core to maximize the data reuses in the L1 cache and minimize the cache coherence traffic, which is similar to the idea of our compiler-assisted CTA mapping algorithm that also assigns the thread blocks with data reuses to the same SM. The concept to enable the memory footprint fit in the shared cache is also similar to the warp limiting technique used in our compiler-assisted programmable warp scheduler. However, they did not apply the data reuse analysis and memory footprint estimation at compile time.

8 Summary

In this paper, we design a locality-aware CTA mapping scheme based on polyhedron model for GPUs to take advantage of the inter-CTA data reuses in the GPU kernels. Using the data reuse analysis based on the polyhedron model, we can detect inter-CTA data reuse patterns in the GPU kernels and control the CTA mapping pattern to improve the data locality on each SM. The compiler-assisted locality-aware CTA mapping scheme can also be combined with the programmable warp scheduler to further improve the performance. The experimental results show that our CTA mapping algorithm can improve the overall performance of the input GPU programs by 23.3% on average and by 56.7% when combined with the programmable warp scheduler. In addition, the overall L1 cache misses can be reduced by 18.9% by the CTA mapping algorithm and 29.3% when the CTA mapping algorithm is combined with the programmable warp scheduler.

References

[1] Era of tera. `http://www.intel.com/pressroom/archive/releases/20070204comp.htm`.

[2] T. M. Aamodt and W. W. Fung. Gpgpu-sim 3.x manual. `http://gpgpu-sim.org`.

[3] A. Bakhoda, G. Yuan, W. Fung, H. Wong, and T. Aamodt. Analyzing cuda workloads using a detailed gpu simulator. In *Performance Analysis of Systems and Software, 2009. ISPASS 2009. IEEE International Symposium on*, pages 163–174, April 2009.

[4] D. B.Kirk and W. W. Hwu. *Programming Massively Parallel Processors, second edition*. Morgan Kaufmann Publishers, 2013.

[5] M. Boyer, D. Tarjan, S. T. Acton, and K. Skadron. Accelerating leukocyte tracking using cuda: A case study in leveraging manycore coprocessors. IPDPS '09, pages 1–12, Washington, DC, USA, 2009. IEEE Computer Society.

[6] S. Che, M. Boyer, J. Meng, D. Tarjan, J. Sheaffer, S.-H. Lee, and K. Skadron. Rodinia: A benchmark suite for heterogeneous computing. In *Workload Characterization, 2009. IISWC 2009.*, pages 44–54, Oct 2009.

[7] N. Corporation. *NVIDIA CUDA Compute Unified Device Architecture Programming Guide*. NVIDIA Corporation, 2007.

[8] N. Corporation. *NVIDIA's Next Generation CUDA Compute Architecture:Fermi*. 2010.

[9] N. Corporation. *PARALLEL THREAD EXECUTION ISA, Version 4.1*. 2014.

[10] N. Corporation. *NVIDIA CUDA (Computer Unified Device Architecture): Programming Guide, Version 7.5.* 2015.

[11] N. Devarajan, S. Navneeth, and S. Mohanavalli. Gpu accelerated relational hash join operation. In *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on*, pages 891–896, Aug 2013.

[12] W. Fung, I. Sham, G. Yuan, and T. Aamodt. Dynamic warp formation and scheduling for efficient gpu control flow. In *Microarchitecture, 2007. MICRO 2007. 40th Annual IEEE/ACM International Symposium on*, pages 407–420, Dec 2007.

[13] M. Gebhart, D. R. Johnson, D. Tarjan, S. W. Keckler, W. J. Dally, E. Lindholm, and K. Skadron. Energy-efficient mechanisms for managing thread context in throughput

processors. *SIGARCH Comput. Archit. News*, 39(3):235–246, June 2011.

[14] H. P. Hofstee. Power efficient processor architecture and the cell processor. In *HPCA '05: Proceedings of the 11th International Symposium on High-Performance Computer Architecture*, pages 258–262, Washington, DC, USA, 2005. ISBN 0-7695-2275-0.

[15] A. Jog, O. Kayiran, A. K. Mishra, M. T. Kandemir, O. Mutlu, R. Iyer, and C. R. Das. Orchestrated scheduling and prefetching for gpgpus. In *Proceedings of the 40th Annual International Symposium on Computer Architecture*, ISCA '13, pages 332–343, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2079-5. doi: 10.1145/2485922.2485951. URL `http://doi.acm.org/10.1145/2485922.2485951`.

[16] M. Lee, S. Song, J. Moon, J. Kim, W. Seo, Y. Cho, and S. Ryu. Improving gpgpu resource utilization through alternative thread block scheduling. In *High Performance Computer Architecture (HPCA), 2014 IEEE 20th International Symposium on*, pages 260–271, Feb 2014.

[17] J. Li, V. Sharma, N. Ganesan, and A. Compagnoni. Simulation and study of large-scale bacteria-materials interactions via bioscape enabled by gpus. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, BCB '12, pages 610–612, New York, NY, USA, 2012. ACM.

[18] L. Liu. *An Optimization Compiler Framework Based on Polyhedron Model for GPGPUs*. PhD thesis, Wright State University, 2017.

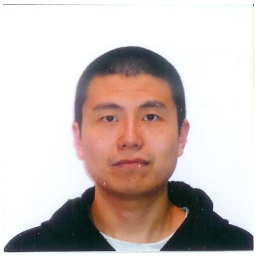
[19] J. Meng, D. Tarjan, and K. Skadron. Dynamic warp subdivision for integrated branch and memory divergence tolerance. *SIGARCH Comput. Archit. News*, 38(3):235–246, June 2010.

[20] V. Narasiman, M. Shebanow, C. J. Lee, R. Miftakhutdinov, O. Mutlu, and Y. N. Patt. Improving gpu performance via large warps and two-level warp scheduling. In *Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO-44, pages 308–317, New York, NY, USA, 2011. ACM.

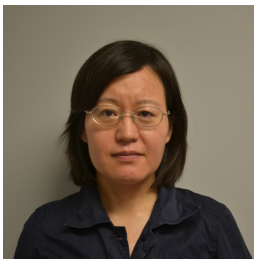
[21] T. G. Rogers, M. O'Connor, and T. M. Aamodt. Cache-conscious wavefront scheduling. MICRO-45, pages 72–83, Washington, DC, USA, 2012. IEEE Computer Society.

[22] S. Williams, J. Shalf, L. Oliker, S. Kamil, P. Husbands, and K. Yelick. Scientific computing kernels on the cell processor. *Int. J. Parallel Program.*, 35(3):263–298, 2007.

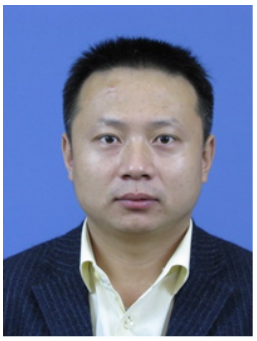
[23] T.-F. Yang, C.-H. Lin, and C.-L. Yang. Cache-aware task scheduling on multi-core architecture. In *Proceedings of 2010 International Symposium on VLSI Design, Automation and Test*, pages 139–142, April 2010. doi: 10.1109/VDAT.2010.5496710.



Lifeng Liu received his Ph.D. degree in Computer Science from the Department of Computer Science and Engineering at the Wright State University, Dayton OH, United States in 2017. He received a B.Sc. and a M.S.E.E. degree in 2008 and 2011 from the department of electrical engineering at the Shanghai Jiaotong University, Shanghai, China. His research interests include computer architecture, parallel computing, compiler optimization, and GPU computing. He is a student member of IEEE.



Meilin Liu is an associate professor at the department of Computer Science Engineering at Wright State University. She received her Ph.D. degree in Computer Science from The University of Texas at Dallas in 2006. Her research interests include optimizing compiler for specific architectures, parallel computing, GPU computing, embedded system, and information security.



Chongjun Wang is a full Professor at the Department of Computer Science and Technology at Nanjing University. He received his Ph.D in Computer Science from Nanjing University, China in 2004. His research interests are Intelligent Agent and Multi-Agent Systems, Complex Network Analysis, Big Data and Intelligent Systems.

Comparison of Metal-Contamination Removal Rates When Using Three Magnets of Different Shapes

Takashi Ohnishi[*,†,§]
Magnetec Japan Ltd., Saitama 359-1164, JAPAN
Saitama University, Saitama 338-8570, JAPAN

Takashi Okamoto[*,†]
Magnetec Japan Ltd., Saitama 359-1164, JAPAN

Keiichi Watanuki[*,§]
Saitama University, Saitama 338-8570, JAPAN

Abstract

In order to create safe and secure products, foreign metal removal is a key process in the quality control of food and pharmaceutical industries. Previously, metal detectors were used to remove foreign metals. However, in the recent years, magnetic separators that can capture small metal particles have been combined with metal detectors to improve the manufacturing yield. Currently, the most common foreign-metal encountered is austenitic stainless steel, because most of the product-processing equipment are manufactured using austenitic stainless steel to make them corrosion proof. SUS304 and SUS316L are the material most commonly encountered in such equipment. Small metal particles adhere to the equipment during sliding and other processes, thus contaminating the equipment. Although austenitic stainless steel cannot be magnetized, weak magnetization is often observed following martensite transformation during sliding and collisions. However, it is not easy to remove small stainless-steel particles during production processes that involve powder flow. In this paper, we investigate the removal rates of small stainless-steel particles when using three magnets of different shapes, under the same specifications and experimental conditions.

Key Words: Magnetic separator, contamination, removal, ANSYS, Finite element method, computer aided engineering, foreign metal, pear-shaped magnet, magnetic flux density, magnetic field, martensite-transformed, austenitic stainless steel.

1 Introduction

Generally speaking, magnetic separators are used to capture and remove foreign-metal particles from powders, grains, and liquids, in the food and pharmaceutical industries. However, magnetic separators can capture and remove magnetized material only. Austenitic stainless steels, which are nonmagnetic materials, are commonly used in production equipment, as shown in Figures 1-2, in order to make them corrosion proof. These particles can be removed by magnetic separators after martensite transformation, as shown in Figure 3. While polishing the equipment, for sanitation purposes, small metal particles can be formed from materials that are already martensite transformed. Thus, it is possible to remove these small austenitic stainless-steel particles by using a magnet under the right conditions.

Reference [5] presents an investigation on a specific magnetic separator, SUS304; however, it does not clarify how the results compare to those of other magnetic separators. This document reports experimental results and clearly indicates the type of metal, size of particle, and processing method used; the effect of magnet shapes on the removal rate is also evaluated. Reference [4] documents an investigation on the removal of iron particles using a plate magnet. However, bar magnets with magnetic flux densities exceeding 1.0 T are commonly used in the food and pharmaceutical industries, and in the recent years, the more commonly removed material is austenitic stainless steel and not iron. Reference [2] documents a liquid filter employing a magnet for the removal of iron particles. This paper shows that magnets are effective in removing austenitic stainless-steel particles. Reference [3] documents a magnetostatic field analysis of the magnetic attachment attractive force and the influence of the magnetic circuit. This paper shows a magnetostatic field analysis for an area of a magnetic circuit used to remove metal particles. Reference [1] documents a magnetic field analysis of the electromagnetic field of ITER in-vessel components, using the same finite element method analysis as that used in ANSYS Maxwell 3D. This paper presents the magnetostatic field analysis of a magnetic separator.

In this paper, we investigate three different magnet shapes circular, triangular, and pear-shaped, and analyze their magnetic flux lines and magnetic fields. Although circular and triangular magnets are used more often, pear-shaped magnet

* Advanced Institute of Innovative Technology, 255 Shimo-okubo, Sakura-ku, Saitama-shi.
† 5-521-1, Mikajima, Tokorozawa-shi,. Email: {ohnishi, okamoto} @magnetec.co.jp
§ Graduate School of Science and Engineering, 255 Shimo-okubo, Sakura-ku, Saitama-shi. Email: watanuki@mech.saitama-u.ac.jp.

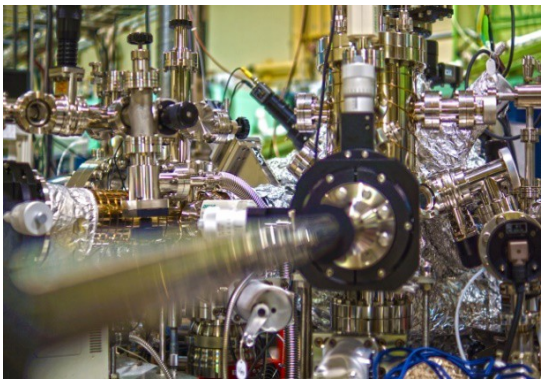


Figure 1: Production equipment

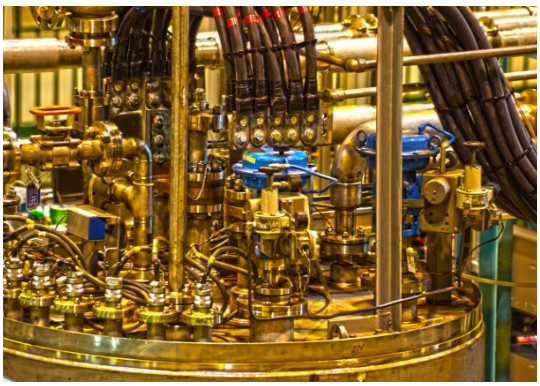


Figure 2: Production equipment

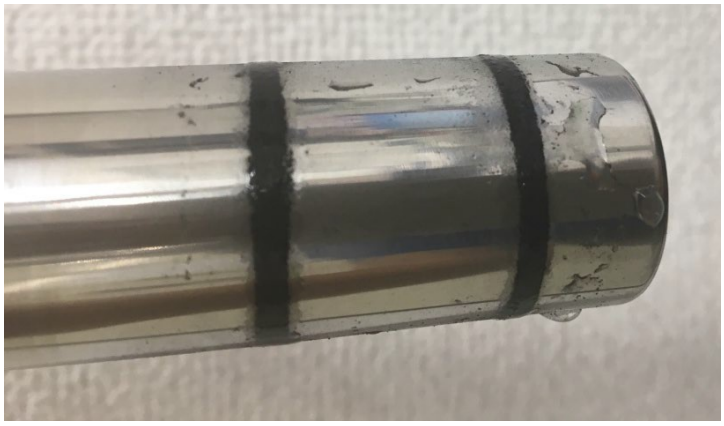


Figure 3: Captured stainless steel particles by magnet

exhibit all the advantages of the other two without any of their limitations.

2 Materials

2.1 Magnetic Separator

Magnetic flux lines are generated evenly around circular magnets, as shown in Figure 4(b). An advantage of circular magnets is that the captured metal particles move to the bottom with the flow of powder; the powder particles are not in contact with the captured metal particles, which stain the bottom of the magnet. However, a disadvantage of the circular magnet is that metal particles may not be attracted by the magnet if powder particles accumulate at the top of the magnet and block the attraction, as shown in Figure 4 (a).

Triangular magnets are shaped in such a way that an acute angle is formed at the top of the magnet; thus, the magnet is always uncovered and can attract metal particles continuously, because powder particles do not accumulate. A disadvantage of the triangular magnet is that the attracted metal particles do not move to the bottom; rather they drop down the sides, owing to the shape of the magnet, as shown in Figure 5(a).

Pear-shaped magnets are shaped in such a way that an acute angle is formed at the top and their bottom is rounded; thus, they can continuously attract metal particles, and these attracted metals stain the bottom, as shown in Figure 6(a).

This study demonstrates the effect of magnet shape on metal removal, by measuring the removal rate of weakly magnetized austenitic stainless steel using three types of magnets, under similar conditions.

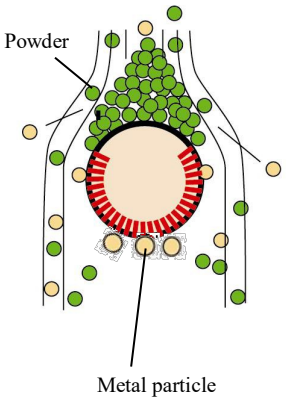
Three configurations were set up with the same number of bar magnets and layers and the same magnetic flux densities, as shown in Figures 7-9, to ensure similar conditions for each magnet shape. The widths of the three magnets are shown in Figure 10.

2.2 Powder Sample

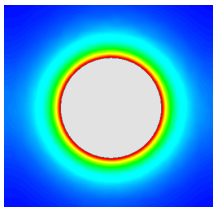
We used medium-strength flour for this experiment, because it had high accumulability, as shown in Figure 11 and Table 1, and could be easily differentiated.

2.3 Metal Particles

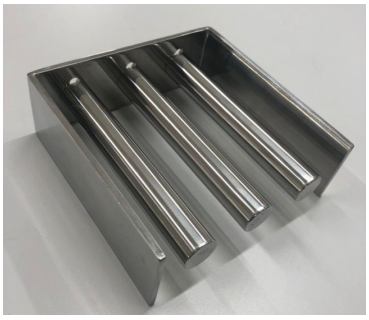
The three types of magnets considered in this work were very effective in the removal of iron and martensite-transformed austenitic stainless-steel particles; thus, it was



(a) Schematic of a magnet

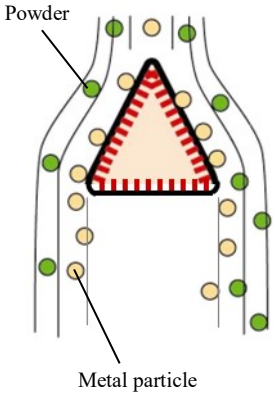


(b) Magnetic flux line

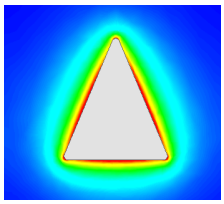


(c) Photo of circular magnet

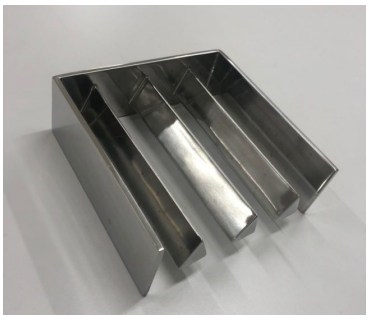
Figure 4: Features of a circular magnet



(a) Schematic of a magnet

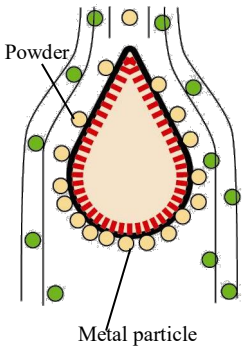


(b) Magnetic flux line

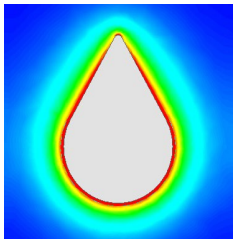


(c) Photo of triangular magnet

Figure 5: Features of a triangular magnet



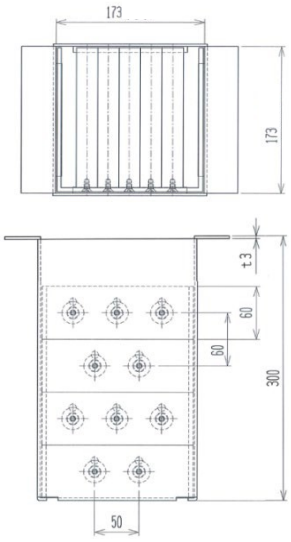
(a) Schematic of a magnet



(b) Magnetic flux line

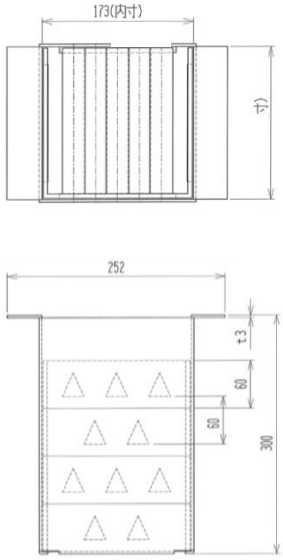
(c) Photo of pear-shaped magnet

Figure 6: Features of a pear-shaped magnet



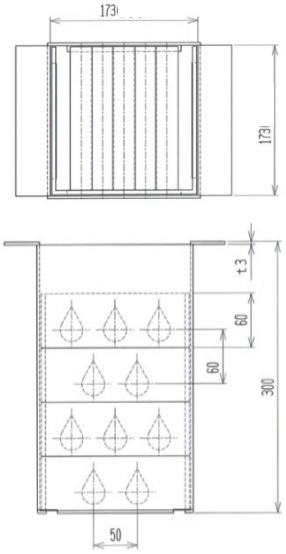
Number of bar magnets: 10
Number of layers: 4
Magnetic flux density: 1.0 T

Figure 7: Grate magnet with four layers of circular magnets



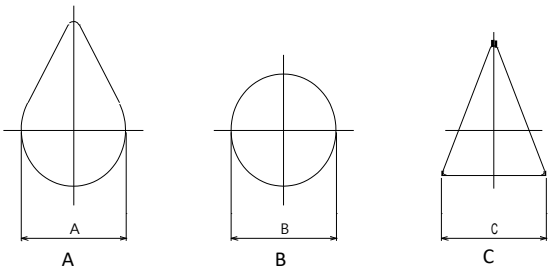
Number of bar magnets: 10
Number of layers: 4
Magnetic flux density: 1.0 T

Figure 8: Grate magnet with four layers of triangular magnets



Number of bar magnets: 10
Number of layers: 4
Magnetic flux density: 1.0 T

Figure 9: Grate magnet with for layers of pear-shaped magnets



No.	Type	Width
A	Pear-shaped	25.2mm
B	Circular	25.2mm
C	Triangular	25.2mm

Figure 10: Comparison of width of magnet

Figure 11: Powder sample

Table 1: Powder specifications

No.	Item	Value	Unit
1	Loose bulk density	0.44	g/cc
2	Tight bulk density	0.68	g/cc
3	Repose angle	88	degree
4	Collapse angle	72	degree
5	Spatula angle	95	degree
6	Spatula collapse angle	75	degree
	Temperature/humidity	15.8 ℃/46%RH	

difficult to compare their results. In order to simplify this, we prepared SUS304 particles with 0.1 mm diameter, by atomization followed by 100 times shot blasting, to create weak martensite-transformed particles, as shown in Figure 12.

Figure 12: Stainless steel particle

2.4 Experimental Environment

From the past experiments, it could be observed that the flowability of the powder was unstable at high humidities; thus, the removal rates would also be unstable. We assume that this is because, as the powder becomes wet, it takes longer for the sieving to break the lumps in the powder during pretreatment. A difference of 73.7-79.3% in the removal rate was observed in an experiment with a pear-shaped magnet for similar conditions when only the temperature and humidity varied. High-magnetic-force austenitic stainless steel subjected to 300 times shot blasting was used in the experiment. As the removal rates of the past experiments were unstable, we considered balancing the temperature and humidity. For relative humidity (RH) 74% and 42%, the removal rates were 73.7% and 79.3%, respectively, for the conditions shown in Figure 13. A large difference in the removal rate was observed with a change in humidity. In other experiments, we balanced the temperature and humidity in a small space with two dryers and a circulator; the progress of the balance is shown in Figure 14. From these experiments, we were able to infer that the temperature and humidity had to be balanced to stable values, according to the amount of powder used, and we could adjust for them.

The experimental environment was set up in a small space with two dehumidifiers and a circulator. The humidity was maintained at 36-44% RH.

The experimental setup is shown in Figures 15(a)-(d) and Figure 16.

3 Methods

3.1 Experimental Method

The experimental steps are as follows.

1. Pass 1 kg of flour through an ultrasonic sieve with an aperture size of 1 mm to break any lumps in the flour and maintain consistency. Add 1 g of metal particles and mix evenly (approximately 100 times).
2. Feed the flour and metal mixture through the electro-magnetic feeder and allow it to flow through the circular, triangular, and pear-shaped magnetic separators for 43-50 s.
3. Measure the amount of metal particles removed by each layer of magnets using an electronic balance.

We arranged the drop position of flour according to the magnetic pole that generated the magnetic field to ensure consistency in conditions and a valid comparison because each magnet had a different magnetic pole position.

Ultrasonic sieve: Artec DGS35-100/200
Electromagnetic feeder: Sinfonia technology WCF-3
Electronic balance: A&D GH-12

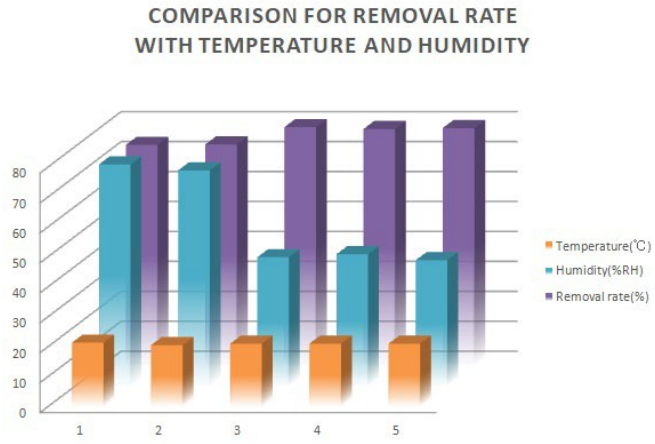


Figure 13: Comparison for removal rate with temperature and humidity

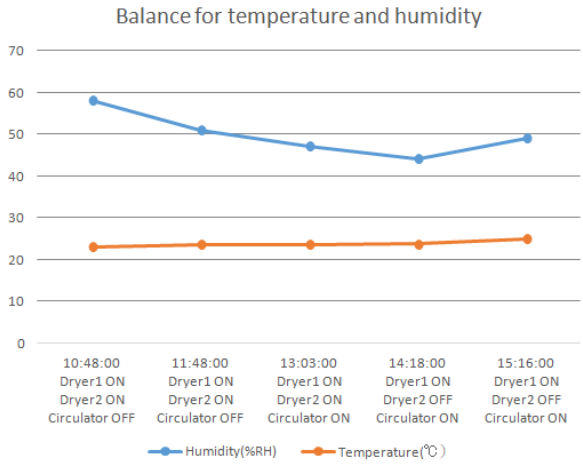
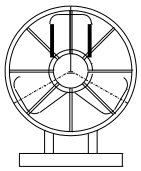
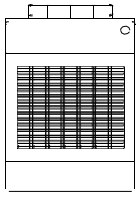


Figure 14: Balance for temperature and humidity



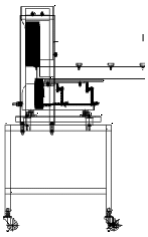
Function: This is used for air circulation in order to ensure a constant value of humidity in a small space.

(a) Circulator



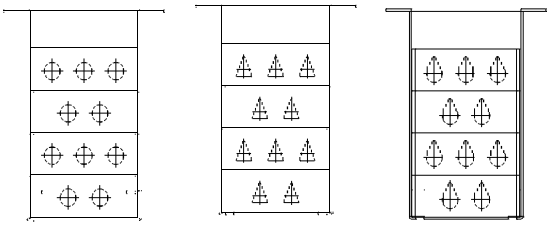
Function: This is used to maintain the humidity at approximately 40% RH to ensure consistent powder characteristics.

(b) Dehumidifier



Function: This is used for feeding the powder and to ensure the same feeding speed and flow rate.

(c) Electromagnetic feeder



Function: Three types of magnets are used for capturing metal particles that are contained in the powder.

(d) Three types of magnet

Figure 15: Devices used in the experiment

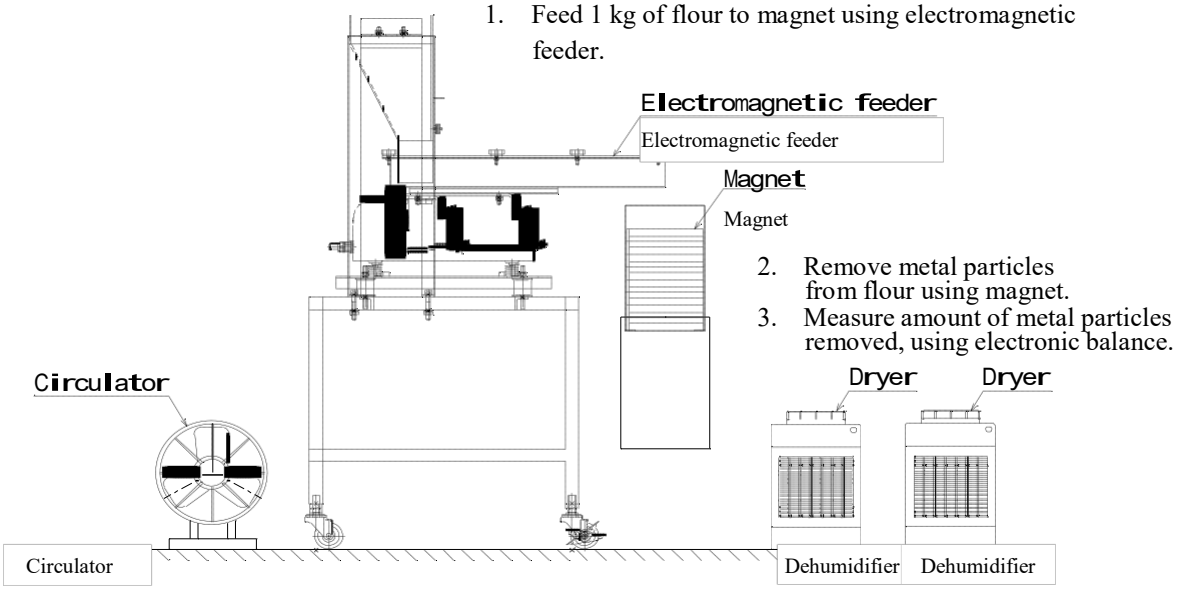


Figure 16: Experimental environment

3.2 Evaluation Method

The SUS304 particles that were separated by the circular, triangular, and pear-shaped magnets of each layer were collected. The removed metal particles were measured using an electronic balance, and the removal rate was calculated from the amount of SUS304 particles removed, as shown in Table 2.

4 Results and Discussion

The results of SUS304 removal from flour using the circular, triangular, and pear-shaped magnets are shown in Table 2. The results show that the performance of the magnets, in terms of the removal rate, is in the order of pear-shaped, circular, and triangular, in Runs 1 through 6. However, the order for Runs 7 through 9 is pear-shaped, triangular, and circular. Little

research has been conducted to show that the humidity of the powders affects their accumulability. This may be the reason why the triangular magnets can remove a greater amount of metal than the circular ones, as the amount of powder accumulated at the top of the magnet increases as the humidity of the powder changes.

As shown in Figures 17-19, there is an increase in powder accumulation at the top of the magnet during the third test.

Table 2: Results of comparison of three types of magnets

Run	Type of magnet	Size of SUS 304	Amount of SUS304(g)	Removed amount for Magnet layer(g)				Total(g)	Temperature /humidity	Drop position	Flow time
				1st	2nd	3rd	4th				
1	Pear-shaped	Diameter 0.1 mm	1.0465	0.194	0.1682	0.1091	0.0403	0.5116	22.0℃	116mm	48.36sec
			100.0%	18.5%	16.1%	10.4%	3.9%	48.9%	37%RH		
2	Triangular		1.0494	0.0587	0.1335	0.1286	0.0415	0.3623	22.4℃	120mm	46.08sec
			100.0%	5.6%	12.7%	12.3%	4.0%	34.5%	36%RH		
3	Circular		1.0462	0.1241	0.1352	0.1051	0.0372	0.4016	22.7℃	117mm	45.93sec
			100.0%	11.9%	12.9%	10.0%	3.6%	38.4%	37%RH		
4	Circular		1.0484	0.1239	0.1313	0.1067	0.0346	0.3965	22.6℃	117mm	48.25sec
			100.0%	11.8%	12.5%	10.2%	3.3%	37.8%	39%RH		
5	Pear-shaped		1.0441	0.1577	0.1461	0.1078	0.0452	0.4568	22.6℃	116mm	43.88sec
			100.0%	15.1%	14.0%	10.3%	4.3%	43.8%	41%RH		
6	Triangular		1.0343	0.0563	0.1400	0.1120	0.0420	0.3503	22.6℃	120mm	43.35sec
			100.0%	5.4%	13.5%	10.8%	4.1%	33.9%	41%RH		
7	Pear-shaped		1.0326	0.1615	0.1480	0.1023	0.0374	0.4492	21.2℃	116mm	48.92sec
			100.0%	15.6%	14.3%	9.9%	3.6%	43.5%	37%RH		
8	Circular		1.0242	0.1150	0.1101	0.1018	0.0315	0.3584	22.5℃	117mm	48.00sec
			100.0%	11.2%	10.7%	9.9%	3.1%	35.0%	44%RH		
9	Triangular		1.0412	0.0593	0.1598	0.1256	0.0390	0.3837	23.0℃	120mm	46.90sec
			100.0%	5.7%	15.3%	12.1%	3.7%	36.9%	43%RH		

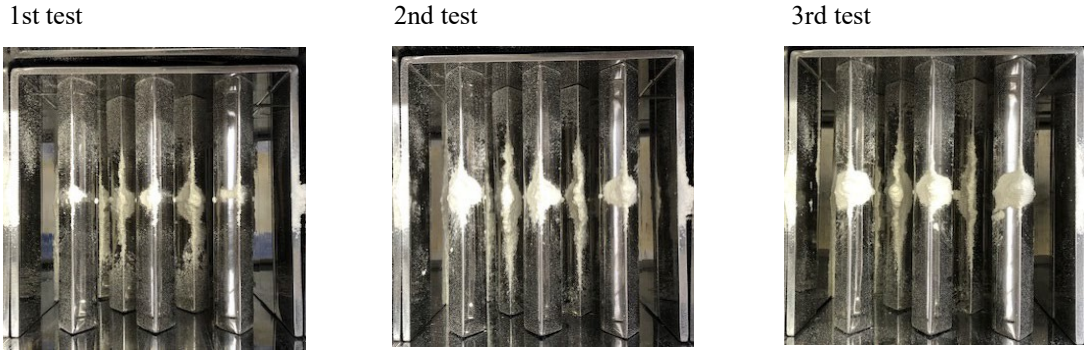


Figure 17: Powder accumulation for pear-shaped magnet

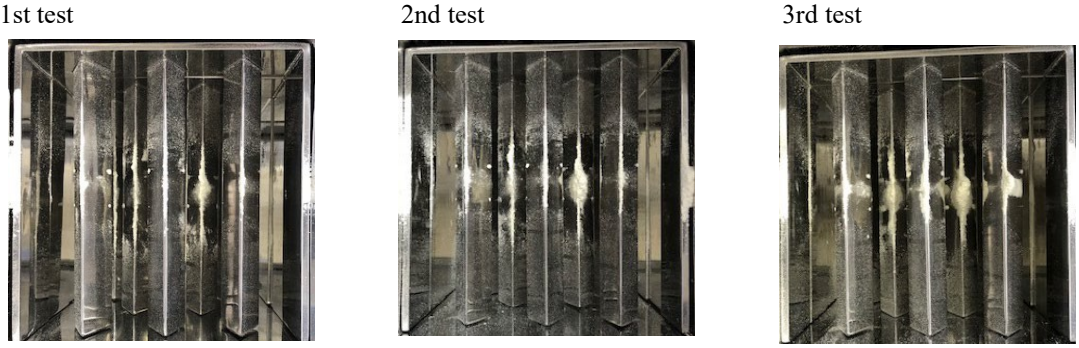


Figure 18: Powder accumulation for triangular magnet

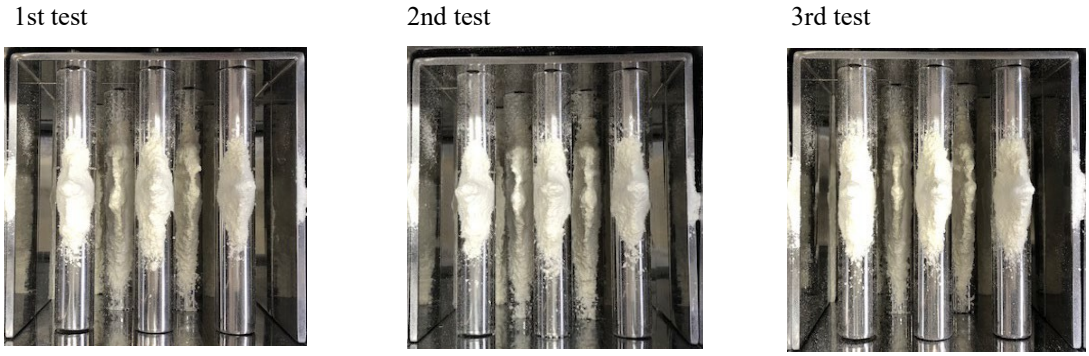


Figure 19: Powder accumulation for circular magnet

The results show that the pear-shaped magnet is the most effective in the removal of austenitic stainless steel.

In order to remove metal particles in a powder flow effectively, we need to consider both the magnet shape and powder flowability. The powder conditions used in this investigation involved passing flour through an ultrasonic sieve with an aperture size of 1 mm to break any lumps, followed by passing this flour through a magnetic feeder to ensure a fixed quantity and reduce the thickness of the strata to facilitate magnetic attraction.

The powder sample had high accumulability, which meant that the pear-shaped and triangular magnets were effective, as the powder did not accumulate on the top of these magnets; their magnetic fields were uncovered and they could attract metal particles.

The widths of the three different magnets were the same (25 mm), and the gap between each bar magnet in a horizontal line was the same as the width (25 mm). Thus, the magnetic separators were designed such that the metal particles in the powder could be attracted by any magnet; however, the accumulated powder blocked this attraction. This meant that the process of effective metal removal called for, essentially, not accumulating powder on top of the magnet. Moreover, weakly magnetized materials could easily fall to the sides with the powder flow, and thus, pear-shaped or circular magnets were found to be effective in staining the metal particles at the bottom of the magnets.

We also simulated the magnetic field for the case before the experiment. The purpose of this simulation was to clarify that the magnetic flux lines were invisible and unstable to measure the magnetic flux density using the Hall element, with the current technology. We performed the magnetic field analysis because it could provide information on the strength and weakness of the magnetic field and the range of capability of the magnetic flux lines. Thus, the effective magnetic field area could be clarified, enabling estimation of performance of the magnet and verification of its effect. The ranges of capability of the magnetic flux lines and magnetic flux densities changed according to the magnetic field applied in a direction and the characteristic of the magnetic material. The placement of the yoke between two magnetic- pieces also changed the capability of the magnetic flux lines. The characteristic values of the model were obtained from the parameters, as shown in Table 3, and our simulation results of the finite element method analysis and measurement were approximated, as shown in Figure 20. The results were obtained by adding the parameters, and not from the beginning. The specifications of the finite element method analysis are shown in Table 4 and the mesh analyses are generated as shown in Figures 21-23.

In addition, the bar magnets repelled each other and were attracted to the metal particles that flowed between them. The magnetic field was simulated as shown in Figures 24, 25, and 26, for the three types of magnets. The pear-shaped and circular magnets exhibited strong attraction to each bar magnet, unlike, the triangular type. Figures 24, 25, and 26, show the simulation results of the magnetic field analysis of the magnet, prepared using computer-aided engineering analysis. The red area is the effective area over which small metal particles are attracted; a wider area tends to attract more particles. Red areas were generated at the bottoms of the bar magnets of the circular and pear-shaped types, but not in the case of the triangular magnet, as shown in Figures 24(a), 25(a), and 26(a). In fact, the pear-shaped magnet was the most effective among the three types of the magnets, as shown in Figure 27, because

Table 3: Parameters

Name		Relative permeability	Magnetic coercivity	Composition
Pear-shaped	Magnet	1.03880530314316	-13.426kOe	Solid
	Yoke	B-H Curve	0A per meter	Solid
Circular	Magnet	1.01050757518664	-97965A per meter	Solid
	Yoke	B-H Curve	0A per meter	Solid
Triangular	Magnet	1.03880530314316	-13.426kOe	Solid
	Yoke	B-H Curve	0A per meter	Solid

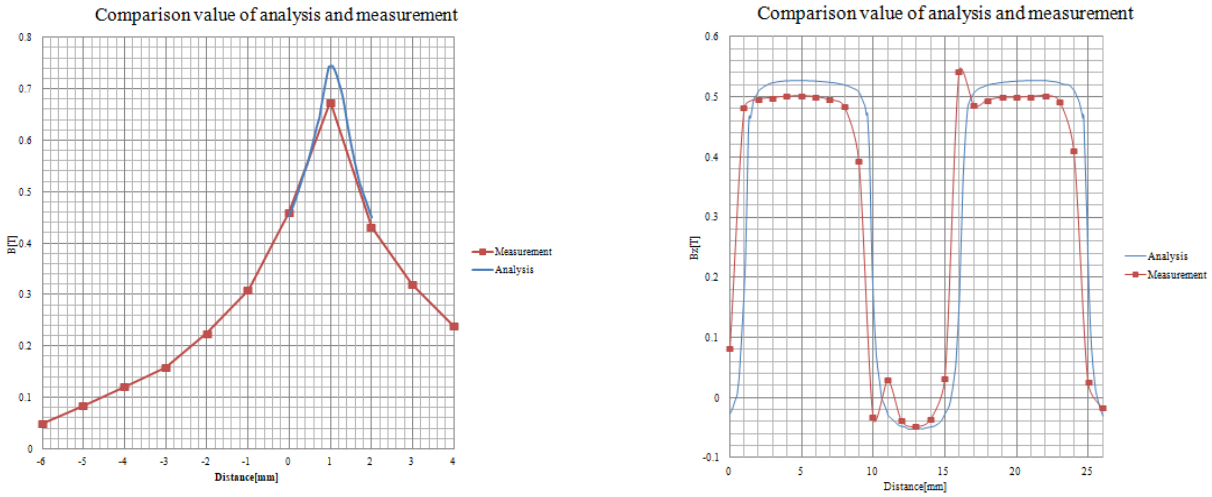


Figure 20: Comparison value of analysis and measurement

Table 4: Specifications of the finite element method analysis

Software of finite element method : ANSYS Electromagnetics Suite 19.1

Analysis device : DELL PRECISION TOWER 7810

	Mesh size	Tetrahedra	Real time	Solver	Scheme
Pear-shaped	2.5 mm	840886	1:12:05	Magnetostatic	Three-dimensional Maxwell's equations
Circular	2.5 mm	931932	1:31:26	Magnetostatic	Three-dimensional Maxwell's equations
Triangular	2.5 mm	756196	1:01:17	Magnetostatic	Three-dimensional Maxwell's equations

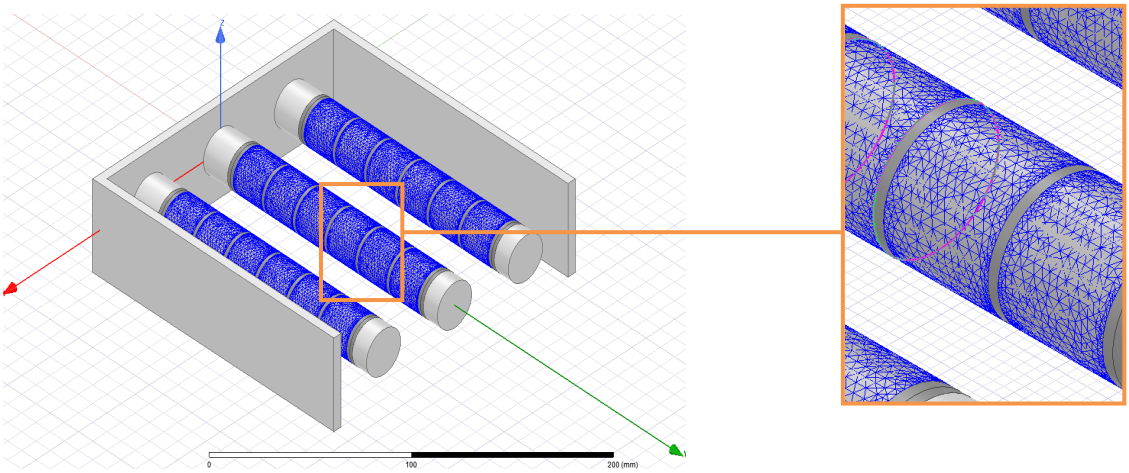


Figure 21: Analysis mesh of circular magnet

it had wide red areas at the bottom of the bar magnet, as simulated. The result was predicted using computer-aided engineering analysis. Although the results of the computer-aided engineering analysis had good accuracy, in the future, the flow analysis will be performed along with the magnetic field analysis to improve the accuracy.

We plotted a graph of the averaged accumulated removal rate, which is shown in Figure 27; the values are listed in Table 5. From the graph and table, it can be seen that the pear-shaped magnet was the most effective for removing metal

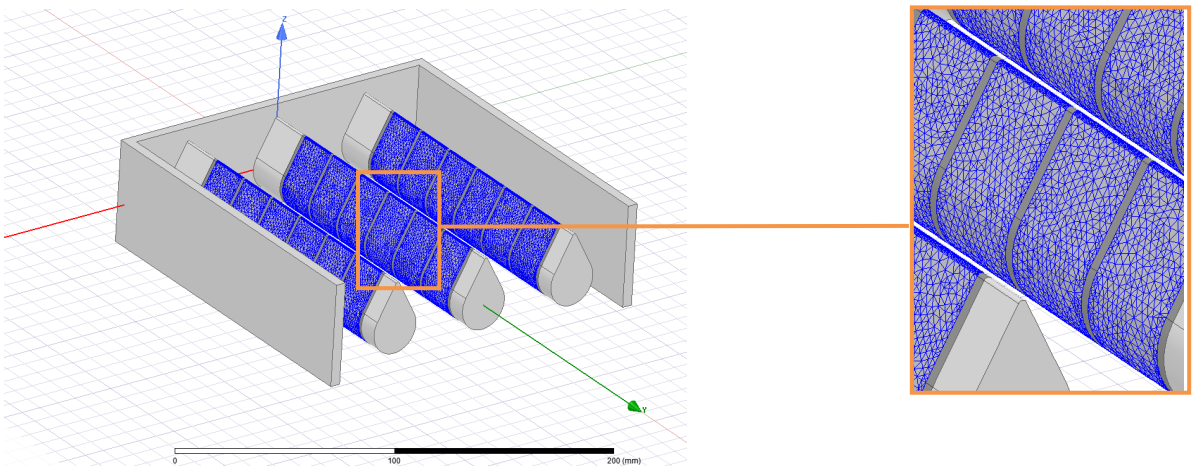


Figure 22: Analysis mesh of pear-shaped magnet

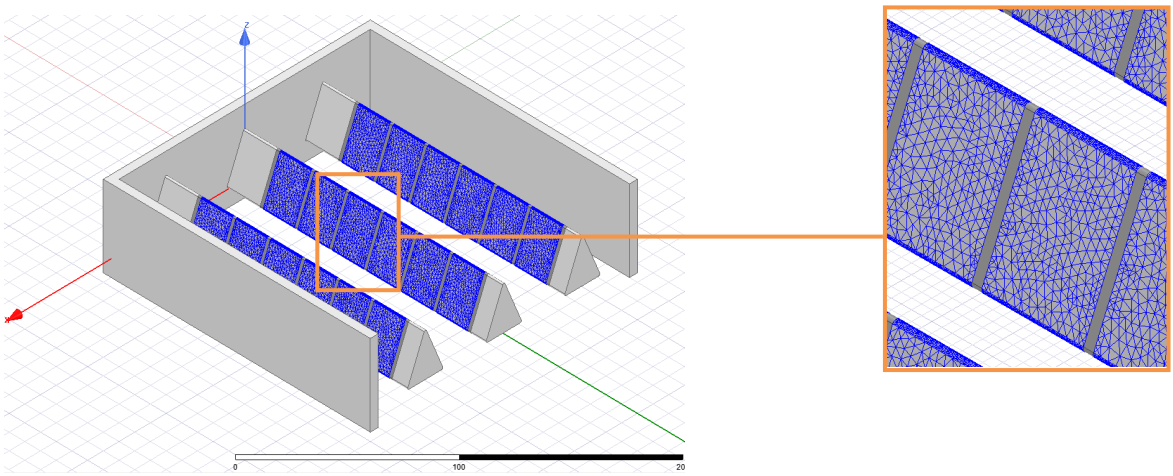


Figure 23: Analysis mesh of the triangular magnet

particles in high-accumulability powder.

5 Conclusions

In this paper, we compared different shapes of magnets (circular, triangular, and pear-shaped) in terms of their ability to remove foreign metal particles from powders used in the food and pharmaceutical industries. We used austenitic stainless steel as the sample, which is the material most commonly used in product-processing equipment. The pear-shaped magnet was found to be the most effective for foreign-metal removal. Owing to its shape, powder did not accumulate at the top of the pear-shaped magnet. This could be attributed to the magnetic field also, which continuously attracted metal particles and prevented them from staining the bottom of the magnet. We found that the difference in the effect of magnet shapes on metal-contamination removal depended on the powder not accumulating on top of the magnets and the metal particles staining the bottom of the magnets. This effect was further increased if the falling metal hit the magnet while falling, instead of being attracted by it, in which case, the magnet would exert its intrinsic performance if the powder does not accumulate on top of the magnet; this could be the reason why the effective removal rate of the circular-type magnet was poor. However, the circular-type magnet could exert its intrinsic performance if the powder flowed easily. The circular-type magnet has the advantages of easy manufacturing and low cost; moreover, magnetic flux lines are generated evenly around circular magnets. Therefore, we need to find an idea and a device to that will not accumulate powder on top of the magnet.

We also found that the removal rate of metal particles varied according to the condition of accumulation of powder on top of the magnet. From the past experiments, the removal rate of metal particles, when using magnetic separators, was found to change according to temperature and humidity changes, especially, when the powder become wet at the beginning and end of the experiment. The temperature and humidity should

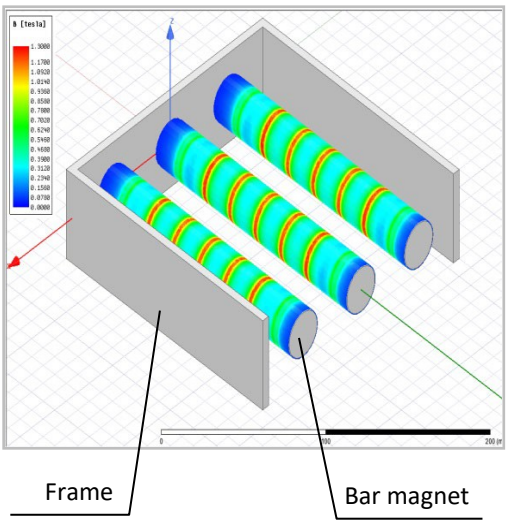


Figure 24: Magnetic field analysis; circular

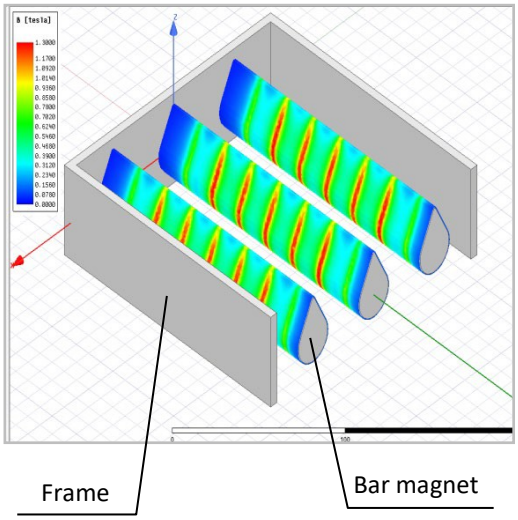


Figure 25: Magnetic field analysis; Pear-shaped

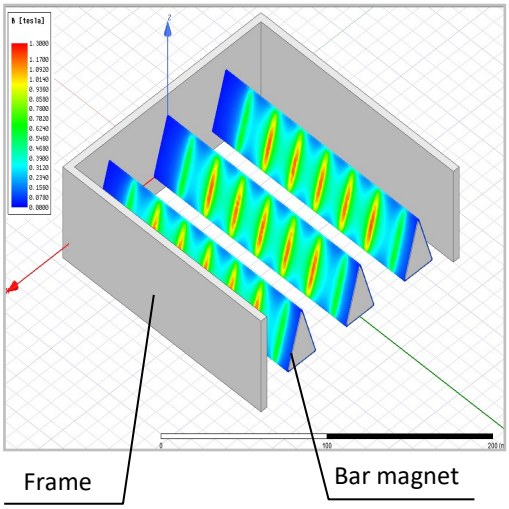


Figure 26: Magnetic field analysis; triangular

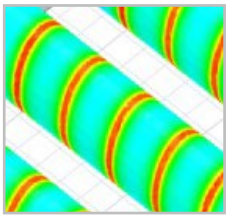


Figure 24(a): Red area

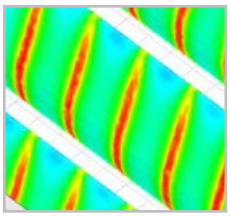


Figure 25(a): Red area

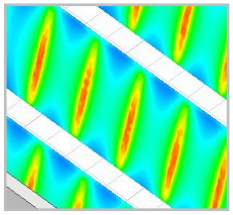


Figure 26(a): Red area

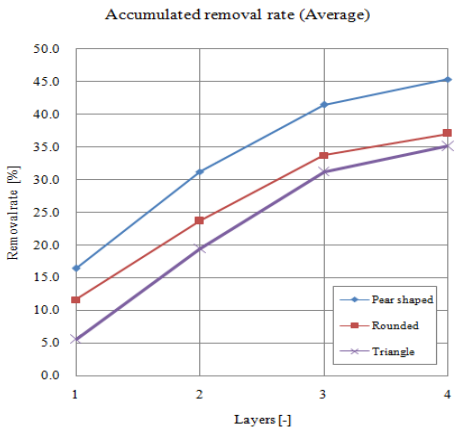


Figure 27: Accumulated removal rate (Average)

Table 5: Accumulated removal rate (Average)

Type of magnet	Accumulated removal rate (Average) [%]			
	1st layer	2nd layer	3rd layer	4th layer
Pear-shaped	16.4	31.2	41.4	45.4
Circular	11.6	23.7	33.7	37.1
Triangular	5.6	19.4	31.2	35.1

be maintained within certain ranges to increase the removal rate of the metal particles in the powder by the magnetic separators.

SUS304 metal particles with diameter 0.1 mm were prepared by the atomization method and 100 times shot blasting, to create weak martensite-transformed particles for comparison and to make the experimental results replicable. There is no global standard for the samples of metal particlesTherefore, for the particles we prepare, we must obtain mill certificate and clarify the process of manufacture. We must also obtain further specifications such as the magnetic moment and magnetic permeability, to ensure to replicability and traceability for further experimentation. We must also adjust the standard for the sample of austenitic stainless steel.

From our experiments, we found that the pear-shaped magnet was the most effective for the removal of SUS304 particles with 0.1 mm diameter from powder with high accumulability. This was because the powder did not accumulate on top of the magnets and the metal particles stained the bottom of the magnets. The removal rate was found to change in accordance with changes in the temperature and humidity, it was not clarified to replicable for process of manufacture for the sample of the metal particle of austenitic stainless steel. To increase the removal rate of metal particles, using the magnet, it is necessary to not only increase the surface flux density of the magnet but also optimize the shape of the magnet, magnetic field, magnetic flux density, magnetic circuit, environmental atmosphere, and flowability of the powder. The manufacturing process and detailed specifications of the sample of metal particles must be clarified to improve the replicability and provide evidence for its suitability. The things needed for the designed factors include flow rate and powder specifications.

Therefore, in the future, we must optimize the magnetic circuit and magnetic field, with respect to the flowability of the powder, targeting foreign-metal particle. Then, we must ensure that the metal particles hit the magnet, in order to facilitate more effective capture. We must also clarify the distance over which the magnet can attract particles, which will be useful in the design of magnetic flux lines.

Our original purpose was the removal of metal particles from a flowing powder, by the attraction of a magnet. We will conduct a study on the flow analysis to improve the removal of finer metal particles, on the basis of this study.

References

[1] D. Giorla, R. Roccella, R. Lo Frano, and G. Sannazzaro "EM Zooming Procedure in ANSYS Maxwell 3D," *Fusion Engineering and Design*, 132:67-72, 2018.

[2] W. Kennedy, *Liquid Filter Employing a Magnet*, United States Patent Office, 1957.

[3] H. Kumano, Y. Nakamura, R. Kanbara, Y. Takada, K. T Ochiai, and Y. Tanaka, "A Three-Dimensional Finite Element Evaluation of Magnetic Attachment Attractive Force and the Influence of the Magnetic Circuit," *Dental Materials Journal*, 33(5):669-673, 2014.

[4] Y. Nagaoka, "Iron Removal from Table Salt," *Bulletin of the Society of Sea Water Science*, 21(1):23-29, 1967.

[5] M. Sada, "Reliability for Food Safety and Safety System," *The Journal of Reliability Engineering Association of Japan*, 30(6):503-509, 2008.

Takashi Ohnishi was born in 1969 in Hyogo prefecture in Japan. He graduated from Rio Hond College. He is currently in the PhD program at Saitama University, working in the Advanced Institute of Innovative Technology in Saitama University and the R&D Section in Magnetec Japan Ltd.

Takashi Okamoto was born in 1990 in Tokyo prefecture in Japan. He graduated from Shibaura Institute of Technology. He is working in the Advanced Institute of Innovative Technology in Saitama University as a researcher and the design department in Magnetec Japan Ltd.

Keiichi Watanuki received his Doctor of Engineering in the Department of Precision Machinery Engineering, Tokyo Institute of Technology, Japan, in 1991. From 1991, he was a member of the Faculty of Mechanical Engineering at Saitama University, and from 2005, he has been a Professor of Mechanical Engineering at Saitama University. He also serves as Director of Human-Machine Interaction Systems Engineering; Deputy Director-General, Research Management Bureau; Director, Advanced Institute of Innovative Technology. Dr. Watanuki's research interests include systems designs that support various human activities in the field of design and manufacturing. Specifically, he develops intelligent computer-aided design and manufacturing (CAD/CAM) systems, design systems for environment, knowledge management and technology transfer systems, virtual reality/augmented reality (VR/AR), cognitive neuroscience, emotional engineering, human-machine interface (HMI), brain-machine interface (BMI), artificial intelligence (AI), Internet of Things (IoT), healthcare and medical technology, intelligent assistive technology, human gait analysis, ambient mobility interface, and intelligent robots.

Co-Active Neuro-Fuzzy Inference System Modeling with Clustering Methods

Ana Farhat*
Oakland University, Rochester, MI USA

Kyle Hagen†
Continental Automotive Systems Inc. Auburn Hills, MI USA

Ka C Cheok*
Oakland University, Rochester, MI USA

Balaji Boominathan†
Continental Automotive Systems Inc. Auburn Hills, MI USA

Abstract

This paper presents a novel approach to Electronic Brake System (EBS) model identification based on co-active neuro-fuzzy inference systems (CANFIS). Data of subcomponents of brake was module obtained from Continental Automotive for modeling purposes. The optimizers having been used to estimate the unknown parameters of the model are Levenberg-Marquardt Algorithm (LMA) and Least Square Error (LSE) for calculating the parameters in antecedent and consequent part of CANFIS. To find the best initial values of membership functions for input data, two clustering methods have been selected which are Fuzzy C-Mean (FCM) and Subtractive Clustering Method (SCM). Finally, the performance of identified model has been evaluated.

Key Words*:* CANFIS, LMA, LSE, clustering, FCM, SCM.

1 Introduction

The introduction of highly efficient and exceedingly capable electrohydraulic actuators has led to a transformation of the Electronic Brake System (EBS) field. These new actuators enable EBS to adapt next generation technologies like brake by-wire and automated driving. These actuators are lighter, deliver braking pressure faster than current conventional hydromechanical brake systems, and provide greater controllability in safety critical applications. Continental brake system, MKC1, using these advanced actuators. Figure 1 shows an overview of the architecture and major components contained within the MKC1 system [9]. Simulation models can be derived from this architecture which enable the analysis of Automotive has developed a next generation electrohydraulic dynamic brake event. Here we explain how the system operates in sufficient detail.

An internal hydraulic actuator is driven through the application of force by a driver on a brake pedal. Determination of the amount of brake fluid output of an electrohydraulic actuator is preformed through travel and pressure measurements of the driver-fed hydraulic actuator. The electrohydraulic actuator provides boosted braking energy through the command of a boost controller unit. The electrohydraulic actuator delivers brake pressure and fluid volume to a controller which distributes fluid to individual wheels. Foundation brake wheel actuators provide torque on the wheels which ultimately translates to deceleration and braking of the vehicle.

Due to the importance of this system and its control applications, this paper will detail the brake model identification method. There are different identification approaches that can be selected according to the system's specifications. In the identification and modeling process, the first step is choosing the right model structure, and the next step will be selecting appropriate optimization algorithm. But before that we need to be sure the data we are using to estimate the model is not corrupted. In practice, data collected from different sources is inherently uncertain because of noise, inconsistency, and incompleteness, and certainly this uncertainty will have a negative impact on the accuracy of the results [8]. Generally, the integrity of data is a very important issue, and it becomes more significant when it comes to cyber-physical systems (CPS) where attackers can remotely attack the sensors and corrupt the measurement data [13, 14]. By being sure about having clean and trustable data set, we can select the model structure. One of the model structures that can identify nonlinear dynamic systems with high accuracy is adaptive network-based fuzzy inference system (ANFIS). This model is used to identify single/multi-input single-output (SISO/MISO) systems.

As is shown in Figure 1, the studied model has two inputs and two outputs. The inputs (U1, U2) are taken as the output of the hydraulic actuator block, and the outputs (R1, R2) are the outputs of electrohydraulic actuator. Therefore, this model is a multi-input multi-output (MIMO) model. To be able to model a MIMO system with ANFIS structure, ANFIS should be modified to coactive neuro-fuzzy inference system

* Email: anafarhat@oakland.edu and cheok@oakland.edu
† kyle.hagen@continental-corporation.com and Balaji.boominathan@continental-corporation.com

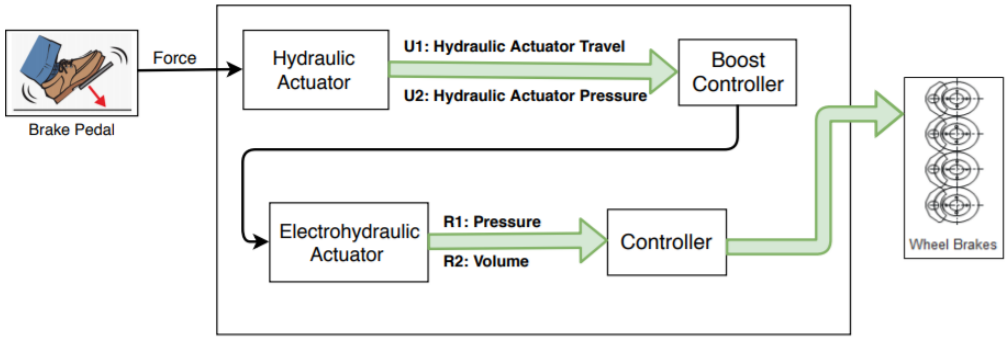


Figure 1: Architecture of brake system

(CANFIS) [6].

As mentioned before, the second step in identification is choosing the optimization method. Different optimization algorithms can be selected to optimize the parameters of the model. Chatterjee and Watanabe [2] use particle swarm optimization (PSO) to update whole parameters of the ANFIS. In [1], a hybrid optimization method which presents PSO for antecedent part and gradient descent (GD) for consequence part is used. Reference [3] compares different training algorithms; the algorithms which are compared are GD, Resilient Propagation, Quick prop, and LMA. The LMA was originally designed [11, 12] to serve as an intermediate optimization algorithm between the Gauss–Newton (GN) method and gradient descent algorithm.

In this paper CANFIS model has been selected to identify the two-input two-output model. As the optimization algorithm, Levenberg-Marquardt algorithm (LMA) is employed to update the parameters in the antecedent part of CANFIS. To estimate the best initial values for these parameters, input data has been clustered by some clustering algorithms. These clusters will provide the initial values of membership functions for the input data in the antecedent part of CANFIS. The consequence parameters are identified by least square estimation (LSE). The article is organized as follows. In Section 2, CANFIS structure and applied optimization algorithms, along with different clustering methods are reviewed. In Section 3, the performance of trained model is verified. Section 4 presents conclusion, the references are presented at the end.

2 CANFIS Model as an Identifier

To identify above mentioned model which is a MIMO system, CANFIS model has been selected as an identifier. CANFIS has the capability of modeling the systems which have more than one output. The structure of CANFIS is similar to ANFIS. Detailed information regarding ANFIS can be found in [7]. CANFIS rules comprise a set of IF-antecedent and THEN-consequent, which are aggregated to produce an output.

For illustration, consider a CANFIS model with two inputs (U1, U2) and two outputs ($O_{1_{delayed}}$, $O_{2_{delayed}}$) as presented in Figure 2. And the number of membership functions (MFs) for each input is set to two. The equations corresponding to Figure 2 is presented in (1).

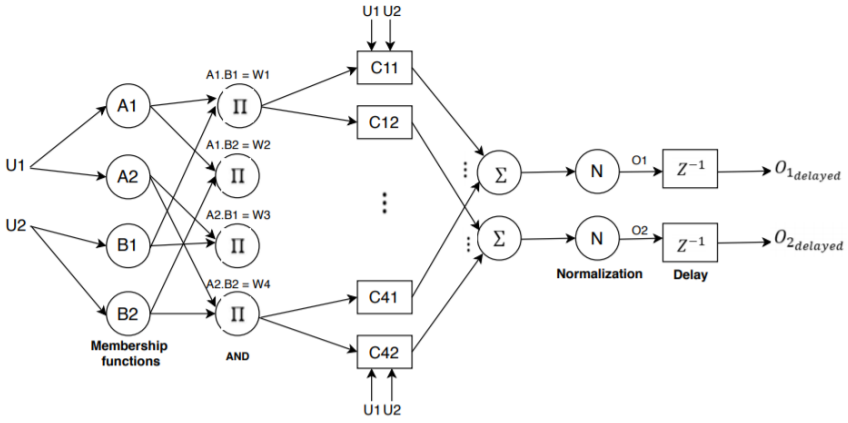


Figure 2: CANFIS **n**etwork

$$If\ U_1\ is\ A_1\ and\ U_2\ is\ B_1, then\ \begin{cases} C_{11} = p_{11}U_1 + q_{11}U_2 + r_{11} \\ C_{12} = p_{12}U_1 + q_{12}U_2 + r_{12} \end{cases}$$

$$If\ U_1\ is\ A_1\ and\ U_2\ is\ B_2, then\ \begin{cases} C_{21} = p_{21}U_1 + q_{21}U_2 + r_{21} \\ C_{22} = p_{22}U_1 + q_{22}U_2 + r_{22} \end{cases}$$

$$If\ U_1\ is\ A_2\ and\ U_2\ is\ B_1, then\ \begin{cases} C_{31} = p_{31}U_1 + q_{31}U_2 + r_{31} \\ C_{32} = p_{32}U_1 + q_{32}U_2 + r_{32} \end{cases}$$

$$If\ U_1\ is\ A_2\ and\ U_2\ is\ B_2, then\ \begin{cases} C_{41} = p_{41}U_1 + q_{41}U_2 + r_{41} \\ C_{42} = p_{42}U_1 + q_{42}U_2 + r_{42} \end{cases}$$

and the outputs of the CANFIS model are

$$O_1 = \frac{C_{11}W_1 + C_{21}W_2 + C_{31}W_3 + C_{41}W_4}{W_1 + W_2 + W_3 + W_4}$$
$$O_2 = \frac{C_{12}W_1 + C_{22}W_2 + C_{32}W_3 + C_{42}W_4}{W_1 + W_2 + W_3 + W_4} \quad (1)$$

where $W_1 = \mu_{A_1}(U_1).\mu_{B_1}(U_2)$, $W_2 = \mu_{A_1}(U_1).\mu_{B_2}(U_2)$, $W_3 = \mu_{A_2}(U_1).\mu_{B_1}(U_2)$, and $W_4 = \mu_{A_2}(U_1).\mu_{B_2}(U_2)$. By considering the number of MFs for each input equal to n, the generalized form of CANFIS rules for i^{th} rule will be:

If U_1 is A and U_2 is B, then

$$C_{ij} = p_{ij}U_1 + q_{ij}U_2 + r_{ij} \quad (2)$$

where $i = 1,2, \dots, n^2$, n is number of membership functions for each input and n^2 is total number of rules. And j = 1, …, k, k is number of outputs. A and B, membership functions of input1 and 2, are Gaussian functions as defined in (3).

$$\mu_{A_f}(U_1) = exp^{-0.5(\frac{U_1 - c_f}{s_f})^2}$$
$$\mu_{B_d}(U_2) = exp^{-0.5(\frac{U_2 - c_d}{s_d})^2} \quad (3)$$

where f = 1, …, n and d = n+1, …, 2n. According to (2) and (3), CANFIS model has some unknown parameters which should be estimated.

2.1 Least Square Error (LSE)

Unknown parameters in consequence part of (2) for output one (O_1) are defined in matrix below:

$$\theta_1 = [p_{11}\ q_{11}\ r_{11} \dots\ p_{n^2 1}\ q_{n^2 1}\ r_{n^2 1}]^T.$$

According to (1), generalized form of first output of CANFIS is

$$O_1 = \frac{C_{11}W_1 + \dots + C_{n^2 1}W_{n^2}}{W_1 + \dots + W_{n^2}} \quad (4)$$

According to (4) we have

$$\begin{bmatrix} W_1U_{11} & W_1U_{21} & W_1 & \cdots & W_{n^2}U_{11} & W_{n^2}U_{21} & W_{n^2} \\ \vdots & & & \ddots & & \vdots & \\ W_1U_{1N} & W_1U_{2N} & W_1 & \cdots & W_{n^2}U_{1N} & W_{n^2}U_{2N} & W_{n^2} \end{bmatrix} \begin{bmatrix} p_{11} \\ q_{11} \\ r_{11} \\ \vdots \\ p_{n^2 1} \\ q_{n^2 1} \\ r_{n^2 1} \end{bmatrix}$$
$$Matrix\ M \qquad Matrix\ \theta_1$$

$$\approx \begin{bmatrix} (W_1 + \dots + W_{n^2})R_{11} \\ \vdots \\ (W_1 + \dots + W_{n^2})R_{1N} \end{bmatrix} \quad (5)$$
$$Matrix\ K$$

$$\Rightarrow\ M\theta_1 \approx K \quad (6)$$

which N is the number of samples, and $R_{1,k}$ with $k = 1, \dots, N$ is the desired output one with N sample points. By LSE

$$\hat{\theta}_1 = \begin{bmatrix} \hat{p}_{11} \\ \hat{q}_{11} \\ \hat{r}_{11} \\ \vdots \\ \hat{p}_{n^2 1} \\ \hat{q}_{n^2 1} \\ \hat{r}_{n^2 1} \end{bmatrix} = (M^T M)^{-1} M^T K$$

The unknown parameters of second output of model (O_2) can also be estimated in a similar way.

2.2 Levenberg-Marquardt Algorithm (LMA)

In this paper, the unknown parameters of antecedent part of CANFIS which are parameters of Gaussian membership functions are estimated using LMA as the learning algorithm. Initial values for these parameters are computed by clustering algorithms presented in the next section. Letting $\boldsymbol{e_1} = R_1 - O_{1_{delayed}}$ and $\boldsymbol{e_2} = R_2 - O_{2_{delayed}}$, cost function can be defined as

$$T = 0.5(\boldsymbol{e_1}^T \boldsymbol{e_1} + \boldsymbol{e_2}^T \boldsymbol{e_2})$$

where $\boldsymbol{e_1}$ and $\boldsymbol{e_2}$ are the error vectors corresponding to first and second output respectively. By considering N sample data and one sample time delay in the output, error vectors are represented as $\boldsymbol{e_1} = [e_{12} \ \dots\ e_{1N}]^T$ and $\boldsymbol{e_2} = [e_{22} \ \dots\ e_{2N}]^T$. Therefore, total error vector will be

$$E = [\boldsymbol{e_1} \quad \boldsymbol{e_2}]^T$$

All unknown parameters of antecedent part, which are centers and standard deviations of MFs, are defined in matrix V as

$$V = [c_1 \ s_1 \dots c_{2n} \ s_{2n}]^T = [v_1 \ v_2 \dots v_{4n}]^T$$

LMA uses Jacobian matrix J which is a gradient matrix representing the partial derivatives of error vector E with respect to V [15]. Here we have the Jacobian matrix equation [15]

$$J = \frac{\partial E}{\partial V} = \frac{\partial \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}}{\partial \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{4n} \end{bmatrix}} = \begin{bmatrix} \frac{\partial e_{12}}{\partial v_1} & \cdots & \frac{\partial e_{12}}{\partial v_{4n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial e_{1N}}{\partial v_1} & \cdots & \frac{\partial e_{1N}}{\partial v_{4n}} \\ \frac{\partial e_{22}}{\partial v_1} & \cdots & \frac{\partial e_{22}}{\partial v_{4n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial e_{2N}}{\partial v_1} & \cdots & \frac{\partial e_{2N}}{\partial v_{4n}} \end{bmatrix}$$

By using Jacobian matrix, LMA update is expressed as

$$V_{q+1} = V_q - ((J_q{}^T J_q) + \mu I)^{-1} J_q{}^T E \qquad (7)$$

which q shows the number of iterations (epochs) of training and factor μ adjusts its value according to the rule depicted in the LMA flowchart in Figure 3. The flowchart shows how the LMA formula adjusts μ to cleverly switch between the coarser delta rule update and finer Gauss-Newton algorithm update [7]. To be able to update matrix V by (7), we need to define the initial value of V which have been defined in the Section 2.3.

2.3 Finding the Initial Values of Membership Functions in Antecedent Part by Clustering Methods

Clustering is a process for organizing patterns into groups according to their proximities to certain characteristics or classes. The patterns belonging to any one of the clusters are similar. In classical clustering methods, the boundary of different clusters is crisp such that one pattern is assigned to exactly one cluster [4]. However, in practice a data point can belong to multiple clusters with different degrees of membership [4]. In this paper, a fuzzy clustering approach, Fuzzy C-Mean (FCM), and Subtractive Clustering Method (SCM) has been used for clustering the input data. FCM is a clustering method that allows each data point to belong to multiple clusters with varying degrees of membership [10]. If there is no clear idea how many clusters there should be for a given set of data, SCM is a fast, one-pass algorithm for estimating the number of clusters and the cluster centers for a set of data [5].

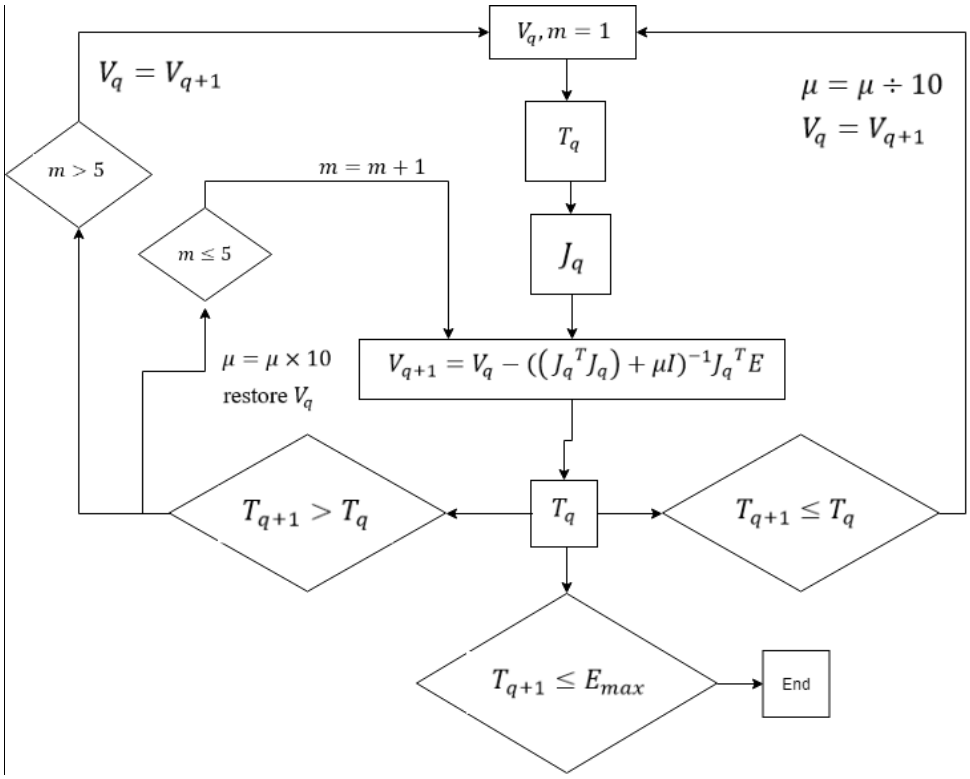


Figure 3: Flowchart for LMA

2.3.1 Fuzzy C-Mean Method, FCM. FCM is based on minimization of the following objective function by determining the values of f_i and $\mu_{i,k}$.

$$J_m(\mu, f) = \sum_{k=1}^{N}\sum_{i=1}^{C}(\mu_{i,k})^m \left\|U_{l,k} - f_i\right\|^2 \quad (8)$$

where N is the number of data points, C is the number of clusters (which is assumed as the number of MFs), $m \in [2,3,\dots\infty)$ is weighting constant, $U_{l,k}$ is the k^{th} data point of l^{th} input, f_i is the center of i^{th} cluster, $\mu_{i,k}$ is the degree of membership of $U_{l,k}$ in the i^{th} cluster. For a given data point, $U_{l,k}$, the sum of membership values for all clusters is one.

$$\mu_{1,k} + \mu_{2,k} + \cdots + \mu_{C,k} = 1$$

FCM performs the following steps during clustering:

Step 1. Randomly initialize the cluster membership values, $\mu_{i,k}$.

$$\mu^{(0)} = \left\{\mu_{i,k}^{(0)}\right\} = \begin{bmatrix} \mu_{1,1}^{(0)} & \cdots & \mu_{1,N}^{(0)} \\ \vdots & \ddots & \vdots \\ \mu_{C,1}^{(0)} & \cdots & \mu_{C,N}^{(0)} \end{bmatrix},\ \mu_{i,k}^{(0)} \in [0,1]$$

Subject to the constraint that $\mu_{1,k}^{(0)} + \mu_{2,k}^{(0)} + \cdots + \mu_{C,k}^{(0)} = 1$. (0) is used to denote the 0^{th} iteration and (t) denotes the t^{th} iteration.

Step 2. Calculate the cluster centers:

$$f_i^{(t)} = \frac{\sum_{k=1}^{N}(\mu_{i,k}^{(t)})^m U_{l,k}}{\sum_{k=1}^{N}(\mu_{i,k}^{(t)})^m}$$

where $i = 1, 2, \dots, C$ and $t = 0, 1, 2, \dots$.

Step 3. Update $\mu_{i,k}$ according to the following:

$$\mu_{i,k}^{(t+1)} = \frac{1}{\sum_{j=1}^{C}\left(\frac{\left\|U_{l,k} - f_i^{(t)}\right\|}{\left\|U_{l,k} - f_j^{(t)}\right\|}\right)^{\frac{2}{(m-1)}}}$$

where $i = 1, 2, \dots, C$ and $k = 1, 2, \dots, N$.

Step4. Calculate the objective function, J_m.

Step5. Repeat steps 2-4 until J_m improves by less than a specified minimum threshold or until after a specified maximum number of iterations [10].

By going through steps 1-5, centers and membership values will be defined by FCM algorithm. In this method, the number of clusters are defined by the user. The purpose of using this clustering method is finding the best initial values of matrix V which are centers and standard deviations of MFs. Clusters defined by FCM are considered as MFs in CANFIS, and their calculated centers are considered as centers of MFs, and standard deviations of MFs are derived by using the values of $\mu_{i,k}$.

2.3.2 Subtractive Clustering Method, SCM. Subtractive clustering assumes that each data point is a potential cluster center. The algorithm does the following:

Step1. Calculate the likelihood that each data point would define a cluster center. Given a collection of N data points, the SCM [5] considers each data point as the candidate for cluster centers. The potential value at data point $U_{l,k}$ is defined as

$$P_k = \sum_{j=1}^{N} e^{-\alpha\left\|U_{l,k} - U_{l,j}\right\|^2} \quad (9)$$

where $k = 1,2, \dots, N$ and $l = 1, 2$. According to (9), a data point which has many neighboring points will have high potential value. After calculating the potential values for each data point, the data point with highest potential value will be selected as the first cluster center.

Step2. Remove all the data points near the first cluster center. If z_1 be the point selected for the first cluster center and P_1 is its potential value, the potential value for each data point will be updated by [4]:

$$P_k = P_k - P_1 e^{-\beta\left\|U_{l,k} - z_1\right\|^2} \quad (10)$$

where $k = 1,2, \dots, N$ and $l = 1, 2$. Therefore, the potential value for the data points near the first cluster center will have significantly reduced potential value which make them unlikely to be selected as the next cluster center.

Step3. Choose the remaining point with the highest potential as the next center.

Step4. Repeat steps 2 and 3 until all the data is within the influence range of a cluster center.

In addition, corresponding standard deviation, σ_i, for each cluster center is estimated by:

$$\sigma_i = \rho\frac{U_{l,max} - U_{l,min}}{\gamma} \quad (11)$$

where l = 1, 2 and $i = 1,2, \dots, M$, M is the total number of clusters in each data dimension, $U_{l,min}$ and $U_{l,max}$ are the minimum and maximum values in each data dimension, and ρ determines the range of influence for each data dimension. The values for parameters presented in (9), (10) and (11) which are $\{\alpha, \beta, \gamma, \rho\}$, throughout the experiments are $\{16, 10.24, \sqrt{8}, 0.5\}$ [4].

By going through the steps 1-4, centers and standard

deviations will be defined by SCM. In this method, number of clusters are specified by the algorithm. As explained in section 2.3.1, these values will be used as initial values of MFs in CANFIS.

2.3.3 Initial Membership Functions Estimated by FCM and SCM. In this section, we are applying FCM and SCM methods of clustering to the input data mentioned earlier to estimate the initial MFs used in CANFIS network. These initial MFs are used as initial values in the training process of CANFIS. Our mentioned model is a two-input two-output model. Input1, U1, is hydraulic actuator travel with the range of [0,13.13] and input2, U2, is the hydraulic actuator pressure with the range of [-1.3,12.7]. The input data is clustered by FCM and SCM. Here we have different sets of MFs which are used as initial values of MFs in antecedent part of CANFIS. First, we estimated these values randomly just by trial and error. Then, clustering methods, FCM and SCM, have been used to cluster the data and estimate the MFs for input data. Estimated MFs by random estimation, FCM, and SCM are shown in Figures 4, 5, and 6, respectively.

We used these MFs as the initial MFs of antecedent part to be trained and updated by LMA. Between these methods, the best results come from initial values estimated by FCM. Therefore,

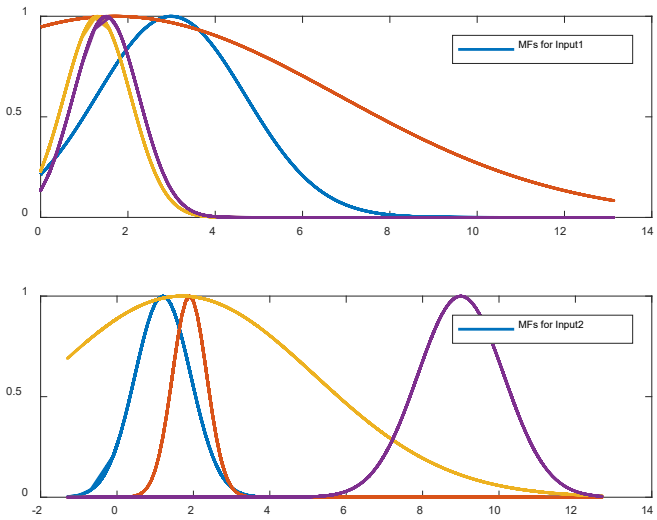


Figure 4: MFs selected randomly by trial and error

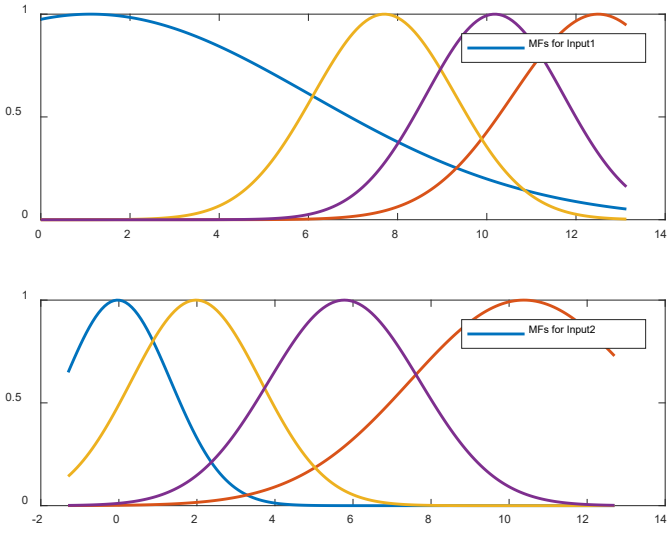


Figure 5: MFs selected by FCM method

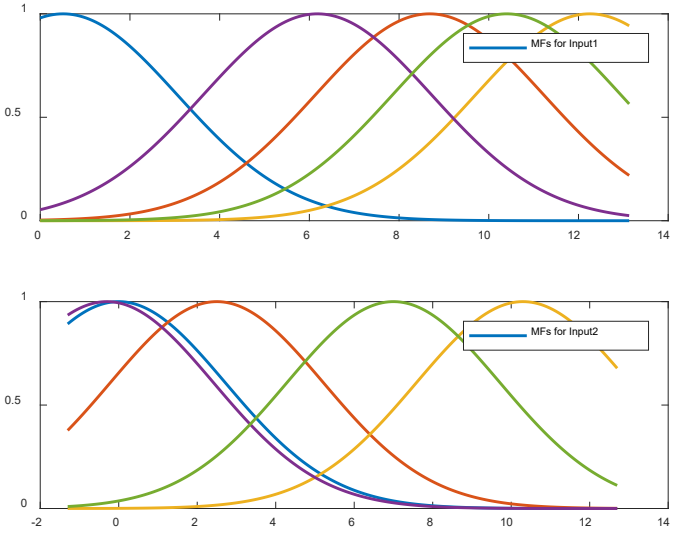


Figure 6: MFs selected by SCM method

the MFs shown in Figure 5 are selected to be considered as the initial MFs of input data set.

The reason that SCM is not as powerful as FCM in estimating the clusters is because SCM is not an iterative optimization-based clustering method; that is, clusters' centers are not being updated based on error evaluation. Another reason can be due to the fixed standard deviation for all the clusters in each set of data. According to (11), σ_i is fixed for each data dimension

3 Training Results

To evaluate the accuracy of the proposed model, input/output data has been taken from the brake module. As mentioned before, this model is designed with two inputs and two outputs. Inputs are pedal travel and pressure change measurements caused by driver, and the outputs are volume and pressure change in the brake actuator. Before training the model we are going to evaluate the effects of number of membership functions on the accuracy of the model. As mentioned earlier, the initial values of MFs are calculated by FCM, and in this method the number of clusters will be defined by the user. Therefore, in the

next section the effect of having different numbers of MFs have been investigated.

3.1 Investigating the Effects of the Number of MFs on the Accuracy of the Model

The number of membership functions in CANFIS model directly affects the accuracy of the model. In this paper different numbers of membership functions (n) have been considered for CANFIS model. We started the modeling with n = 2, and by checking the accuracy of the model, eventually n = 4 has been selected as the number of membership functions for this model. Figures 7 to 9 represent the results. In Figure 7, 2 membership functions are considered for each input in the model. In this figure desired outputs (R1, R2) are compared with outputs of the model (O1, O2). As is shown, the second output of the model is not mimicking the data. The same comparison has been implemented in Figure 8 with 4 membership functions for each input. Corresponding error is available in Figure 9. According to these results, 4 membership functions have been selected to model the data.

3.2 Training the model

In this section, with the selected number of membership functions, the model will be trained. Figure 10 shows the training dataset having been used to train the model. As is shown in this Figure, 3 sets of data are used to train the model and 1 set is kept to be used as test data.

The accuracy of trained model is shown in Figure 11. In this figure the desired outputs (R1, R2) are compared to the estimated outputs (O1, O2).

A magnification of Figure 11 is shown in Figure 12. And error plots corresponding to Figure 11 show the difference between training data and output of the model presented in Figure 13

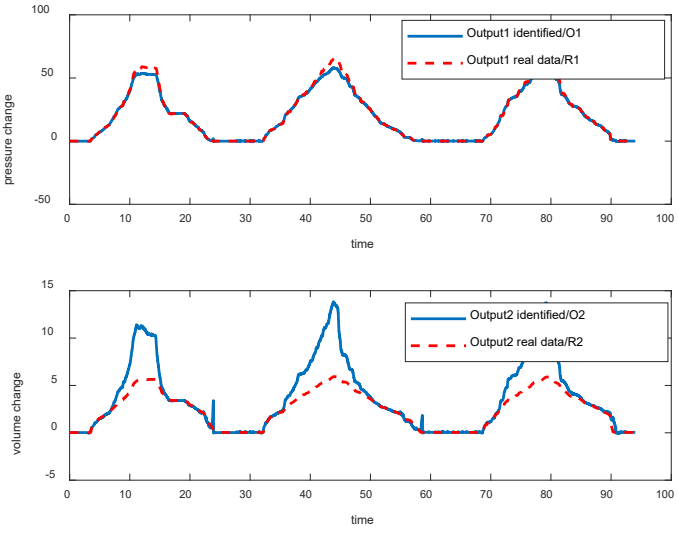


Figure 7: Performance of the model with n = 2

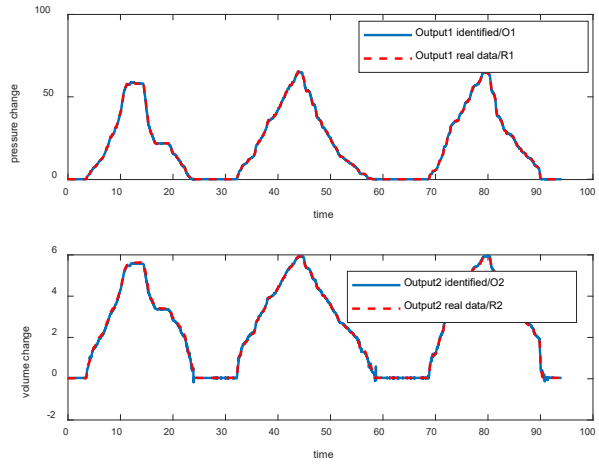


Figure 8: Performance of the model with n = 4

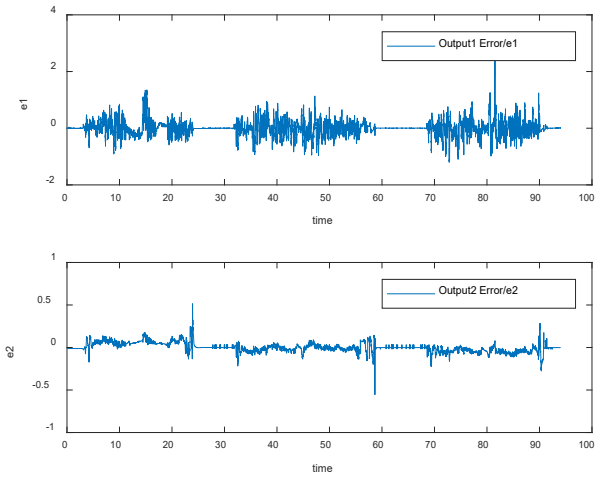


Figure 9: Corresponding error with n = 4

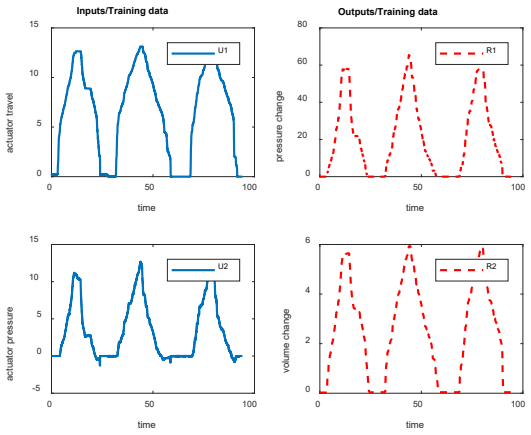


Figure 10: Input/Output training data

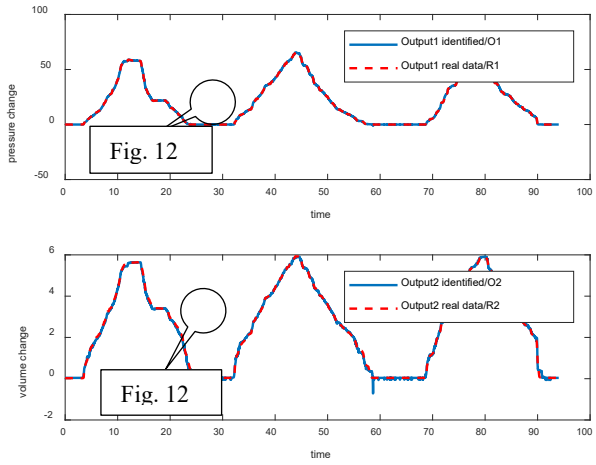


Figure 11: Identified outputs by CANFIS

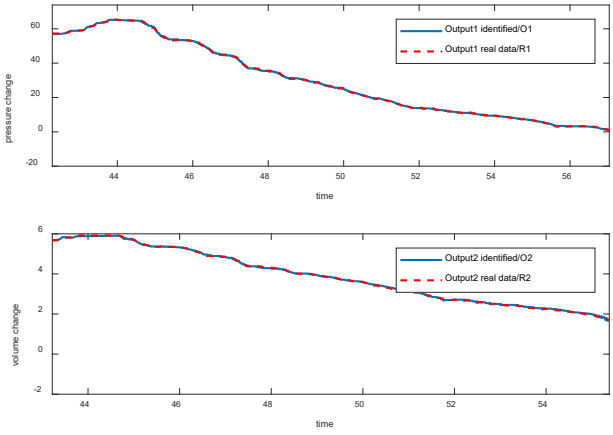


Figure 12: Magnification of part of Figure 10

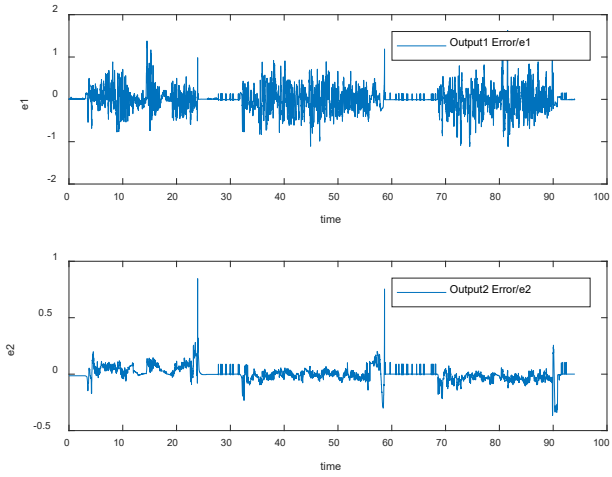


Figure 13: Error plots

3.3 Validating the Model

The trained model has been validated by using three different datasets. As an example, one set of validation data is shown in Figure 14.

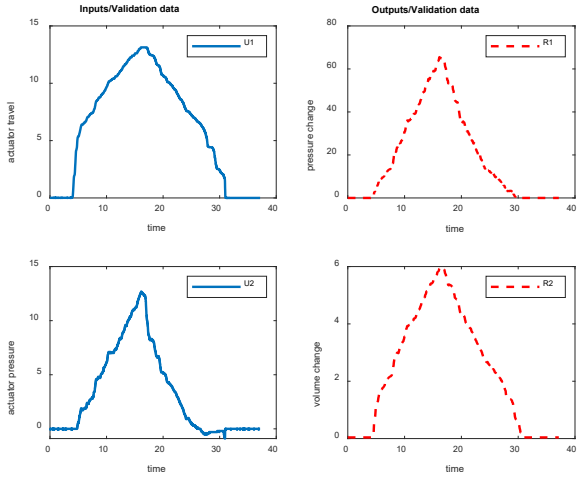


Figure 14: Validation data

The outputs of trained model using validation input data are presented in Figure 15. As is shown in this figure, outputs of trained model (O1/O2) is fitting the real data (R1/R2).

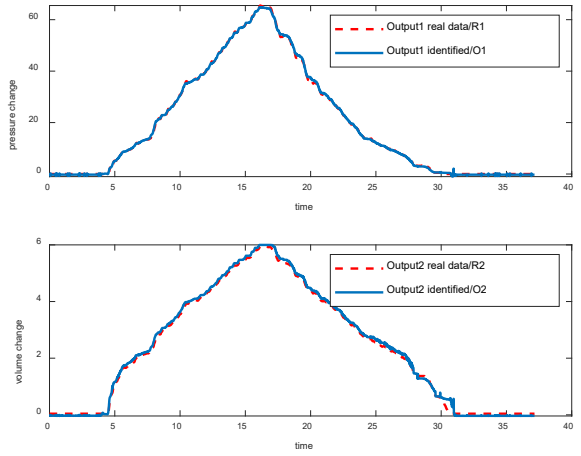


Figure 15: Outputs comparison using validation data

4 Conclusion

We presented a novel approach based on CANFIS modeling for identifying the model of a subcomponent of brake module. To estimate the unknown parameters of the model Levenberg-Marquardt Algorithm (LMA) and Least Square Error (LSE) algorithms have been applied as the optimizers. Best initial values of membership functions for input data are estimated using clustering methods. The effectiveness of the model is

evaluated by comparing the model output and actual data. As is shown neuro-fuzzy network modeling strategies determines system behavior with better accuracy since the correlation of real-world data and the model is more direct.

References

[1] M. Sh Aliyari, M. Teshnehlab, and A. K. Sedigh, "A Novel Training Algorithm in ANFIS Structure," *Proceedings of the American Control Conferences*, **6:5059-5064**, 2006.

[2] A. Chatterjee and K. Watanabe, "An Optimized TakagiSugeno Type Neuro-Fuzzy System for Modeling Robot Manipulators," *Neural Comput Appl.* 15(1):55–61, 2006.

[3] M. S. Chen, "A Comparative Study of Learning Methods in Tuning Parameters of Fuzzy Membership Functions," *IEEE Conf. on Syst. Man. Cybern.*, 3:40–44, 1999.

[4] Mu-Song Chen and Shinn-Wen Wang, "Fuzzy Clustering Analysis for Optimizing Fuzzy Membership Functions," *Fuzzy Sets and Systems*, 103(2):239-254, 1999.

[5] S. Chiu, "Fuzzy Model Identification Based on Cluster Estimation,"*, Journal of Intelligent & Fuzzy Systems*, 2(3), Sept. 1994.

[6] A. Farhat, K. Hagen, K. C. Cheok, ans B. Boominathan, "Neuro-Fuzzy-based Electronic Brake System Modeling using Real Time Vehicle Data," *Proceedings of 34th International Confer.,* 58:444-453, 2019.

[7] Ana Farhat and Ka C. Cheok, "Improving Adaptive Network Fuzzy Inference System with Levenberg-Marquardt Algorithm," 2017 Annual IEEE International Systems Conference, pp 1-6, April 2017.

[8] Reihaneh H. Hariri, Erik M. Fredericks, and Kate M. Bowers, "Uncertainty in Big Data Analytics: Survey, Opportunities, and Challenges," *Journal of Big Data*, 6.1:44, 2019.

[9] https://www.continental-automotive.com/en-gl/Passenger - Cars/Chassis-Safety/Brakes/Electronic- Brakes/MK-C1/ MK-C1, 2018.

[10] https://www.mathworks.com/help/fuzzy/fuzzy-cluster-ing.html, 2017.

[11] Levenberg–Marquardt Training by Hao Yu and Bogdan M. Wilamowski, 2011.

[12] D. Marquardt, "An Algorithm for Least Squares Estimation of Non-Linear Parameters," *J. Ind. Appl. Math*., pp 431–441, 1963.

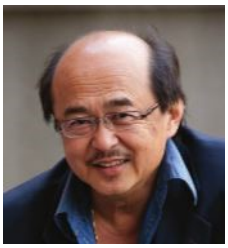
[13] S. Nateghi, Y. Shtessel, J. P. Barbot, and C. Edwards, "Cyber Attack Reconstruction of Nonlinear Systems via Higher-Order Sliding-Mode Observer and Sparse Recovery Algorithm," 2018 IEEE Conference on Decision and Control (CDC), pp 5963-5968, 2018.

[14] S. Nateghi, Y. Shtessel, J. P. Barbot, G. Zheng, and L. Yu, "Cyber-Attack Reconstruction via Sliding Mode Differentiation and Sparse Recovery Algorithm," Electrical Power Networks Application, 2018 15th International Workshop on Variable Structure Systems (VSS), pp 285-290, 2018.

[15] Hao Yu and Bogdan M. Wilamowski, Levenberg-Marquardt Training, Auburn University, 2011.

Ana Farhat is a PhD student in Electrical Engineering at Oakland University, Rochester, MI, supervised by Prof. Ka C. Cheok. Her research interests are Neuro-fuzzy modeling, system identification, and optimization algorithms.

Kyle Hagen is a Senior Systems Engineer at Continental Automotive working in the field of systems simulation and modeling of electronic brake systems. He received a B.S.E. degree in Aerospace Engineering and a M.E. in Space Systems Engineering, each from the University of Michigan Ann Arbor. His current research interests include mechanical, dynamic, and data-driven simulation of vehicular systems.



Ka C. Cheok is a Professor of Engineering and John Dodge Chair at the Department of Electrical & Computer Engineering, Oakland University, Rochester, MI. He has led several successful R&D collaborations in intelligent systems and autonomous robotics for over the years including automated breast cancer diagnostic tester that integrates IR thermography and AI; fuzzy logic based highway and city street lane centering systems; ultra-wideband tracking of omnidirectional mobile robots & assets in GPS denied areas; mine detection robot that sweeps for anti-personnel ordinance.

Balaji Boominathan is working in Continental Automotive Systems Inc., as a Systems Engineering Manager responsible for Platforms, System Design and Mechatronics development for advanced electronic brake system technologies. He received a B.E. degree in Electrical & Electronics Engineering from Madurai Kamaraj University and M.E. degree in Electrical Engineering from Anna University, India. His research areas include autonomous brake system concepts, optimized system architecture modelling and system simulation techniques.

High-Level Synthesis Optimization of AES-128/192/256 Encryption Algorithms

Luka Daoud*, Fady Hussein*, and Nader Rafla*
Boise State University, Boise, ID 83725 USA

Abstract

Advanced Encryption Standard (AES) is applied in many worldwide systems including several private and public sectors to ensure data confidentiality in various applications ranging from data servers to low-power hardware embedded systems. AES encryption algorithm is composed of several functions and serial rounds which makes it challenge to be efficiently implemented and optimized in hardware. As a solution, High-Level Synthesis (HLS) offers flexibility in designing and rapid optimization of dedicated hardware to meet the design constraints. In this paper, we designed a fully pipelined AES encryption engine for key sizes, 128, 192, and 256 using Xilinx Vivado HLS tool chain. The AES architecture was designed and pipelined such that an encryption process is executed every clock cycle after filling the pipeline latency. We show that the proposed HLS-based design achieves 28 Gbps (Gigabit per second) throughput on Artix-7 FPGA occupying 2794, 3306, and 3750 FPGA-slices for AES-128, AES-192, and AES-256, respectively.

Key words: Advanced encryption standard, AES, High-Level Synthesis, HLS, security, optimization, high throughput, low-resources utilization, FPGA.

1 Introduction

Advanced Encryption Standard (AES) is a standardized algorithm approved by the National Institute of Standards and Technology (NIST) [16]. It represents a fundamental building block of many network security protocols to ensure data confidentiality in various applications ranging from data servers to low-power embedded systems. However, AES-based block cipher is computationally intensive. The cipher algorithm consists of multiple rounds of encryption where each round has three main layers of ciphering techniques to apply data confusion through nonlinear transformation and data diffusion by mixing the data state. At each round, the algorithm takes the state array and, after applying a round encryption, returns an updated state array. Performing AES encryption algorithm in software on general purpose processors is time demanding which may require a dedicated computing module. For example, embedded systems often rely on dedicated hardware accelerators for data encryption and decryption. There is a need to accelerate the AES encryption algorithm on application-specific integrated circuit (ASIC) or reconfigurable hard-ware devices such as field programmable gate arrays (FPGAs).

The implementation and optimization of such complex functions in Hardware Description Languages (HDL) is time consuming and not trivial to optimize. In order to achieve an efficient design with minimal effort, High-Level Synthesis (HLS) procedures are applied. HLS is an automated process that accepts a system design created in a high-level language, such as C or C++, and then generates a Register Transfer Level (RTL) design describing the behavior of the system. HLS plays an important role in the design process by reducing the effort of designing and debugging HDL thus providing flexibility in the final hardware implementation to meet design constraints set by the developer.

In this paper, we leverage Vivado High-Level Synthesis [20], to design a fully pipelined AES encryption algorithms and evaluate their performances. The proposed architectures support different key sizes including 128, 192, and 256 bits. The designed AES encryption mode is based on Electronic Code Book (ECB), where each possible block of plaintext has a defined corresponding ciphertext value and vice versa. The design is implemented on the Xilinx Zynq-7000 SoC FPGA chip of the ZedBoard prototyping board [21]. Our proposed AES design was implemented by using only look-up tables (LUTs) and flip flops (FFs) without including any block RAM (BRAM) or DSP slices of the FPGA. Therefore, our design may be appealing to low-cost and high-throughput applications. The AES architecture was designed and fully pipelined such that an encryption process is executed every clock cycle after filling the pipeline latency. Additionally, we compare our proposed HLS implementation of the standard AES algorithm to previous implementations on FPGAs. The rest of this paper is organized as follows. We present the background and related work in Section 2 and provide an overview of HLS and relevant techniques in Section 3. We then present our proposed AES block cipher algorithm and its optimization in Section 4. Finally, Section 5 concludes the paper and provides future research directions.

2 Background and Related Work

AES is a symmetric block cipher [16] and its fundamental operations are performed on byte-level field over the Galois

* Electrical and Computer Engineering. Email: LukaDaoud, FadyHussein]@u.boisestate.edu, nrafla@boisestate.edu.

Field GF(2^8) [17], where each input block is divided into a set of 8-bit vectors. The algorithm encrypts and decrypts a 128-bit block of data by repeatedly applying the same round transformation using a secret key. The key size can be either 128, 192, or 256 bits and is chosen based on the preferred security level. The different versions of AES are known as AES-128, AES-192, or AES-256, where the number represents the key length, and the number of rounds for each version area 10, 12, and 14, respectively. Different versions of AES are supported by different hardware platforms depending on the systems and applications. Therefore, in this paper, we present the implementations of AES-128, AES-192, and AES-256 on FPGA using HLS.

2.1 AES Structure

The standard AES encryption algorithm includes different phases as shown in Figure 1. The algorithm begins by applying an initial round, followed by a number of standard rounds on the output, and ends with a final round. Each standard round, intermediate cipher, has four different operations to scramble and non-linearly transform the data. The intermediate result is called a state and is represented as a 2-D matrix notation of 4×4 bytes.

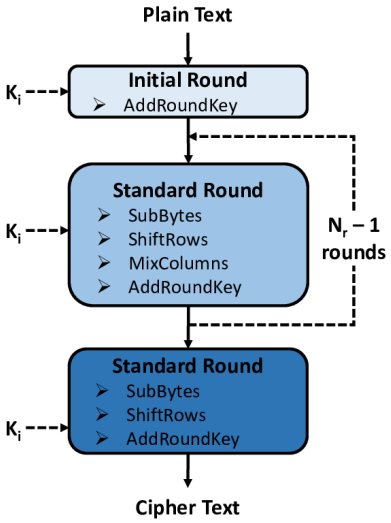


Figure 1: AES design block

The four main operations of the algorithm are summarized as follows:

- SubBytes: This is an invertible non-linear transformation. It is a process of substituting each byte of the input state by another byte from a predefined table (S-box) that contains 256 different elements for each byte. The design criteria of the S-box values is to be resistant against the known differential and linear crypto-analysis. Each possible element of the S-box is generated by computing the multiplicative inverse in GF(2^8) and then applying an affine transformation.
- ShiftRows: This operation of cyclically shifting (rotating) each row with a different offset.
- MixColumns: This operation cyclically shifts (rotates) each row with a different offset.
- AddRoundKey: This operation computes a bit-wise XOR between the round key (sub-key) and the current state. Each round key is derived from the previous sub-key. This requires the encryption algorithm to schedule the key for each round.

For example, in the standard AES-128, the initial round is done by applying AddRoundKey, i.e., XORing the key with the input data. Then, it is followed by 10 repeated rounds, where each round performs SubBytes, ShiftRows, MixColumns and AddRoundKey operations. The final round is slightly different from the other rounds, as it does not include the MixColumns function.

The key schedule takes the original input key and derives a sub-key for each round. The sub-key derivation is different for each version of the AES algorithm. The sub-key derivation is computed recursively, and the first sub-key is the original input key. The AES key schedule is word-oriented with word size = 32 bits. Table 1 presents the number of rounds and key sizes for AES-128, AES-192, and AES-256. For example, for AES-128, the number of rounds is 10, and 11 sub-keys are derived, each of 128 bits, that is the total size of the sub-keys is 176 bytes. Similarly, for AES-192 and AES-256, the total size of sub-keys are 208 and 240 bytes, respectively. Key expansion for each round for each version of the AES is described in [16-17].

Table 1: Number of rounds, key size and sub-keys size for AES-128/192/256

AES configuration	Rounds	Key size (bits)	Sub-key size (bytes)
AES-128	10	128	176
AES-192	12	192	208
AES-256	14	256	240

2.2 Related Work

In the literature, several methods have been proposed to implement AES algorithm on FPGA [2, 4, 7, 9] and ASIC [14-15, 18]. Most of the implementations feature high speeds and high costs suitable for high-end applications only. Early AES designs were mostly straightforward implementations of various loop unrolled and pipelined architectures. Recent implementations of the AES on FPGA demonstrate better utilization of FPGA resources by using dedicated on-chip memories that implement S-boxes and DSP slices [8].

In [4], the authors studied and compared the performances of implementations of different candidate AES encryption techniques on FPGA and evaluated their suitability for FPGA-based implementations. They focused on time performance and the encryption throughput. Biasizzo, et al. [2] presented a hardware implementation of the AES algorithm developed for an external data storage unit in a dependable application and

optimized the encryption algorithm to meet the needs of the target application. In [9], the authors focused on optimizing the AES algorithm to suit small embedded applications or low power consumption devices. They achieved a throughput of 121 Mbps at a maximum frequency of 153 MHz targeting small area design and lower-energy consumption per processed block.

HLS has been used to optimize the implementations of cryptography protocols in hardware [5, 10, 12]. There has been some work on using HLS to implement the AES algorithm on FPGAs [1, 11, 13, 19]. In [1], the authors investigated various optimizations of the C-based AES implementation into hardware using C2R [3] methodology for co-processor synthesis. These implementations included baseline hardware design, BRAM-based architecture, a pipelined scheme, and an optimized architecture for improving the performance and area utilization. In [13], the authors explored different hardware implementations of AES using HLS directives and memory partitioning optimizations. In [19], the authors explored four different implementation methods of AES using Vivado HLS for different types of substitution tables. In our previous work [5], we optimized the standard AES-128 encryption algorithm using Vivado HLS. The key idea of the optimization was to calculate the sub-key of each round on the fly during the encryption process. The optimized architecture achieved an encryption with a rate of 1.29 Gbps (Gigabits per second) on the FPGA.

In this paper, we optimized the standard encryption algorithms for AES-128, AES-192, and AES-256. The AES architectures were fully pipelined such that an encryption process is executed every clock cycle after filling the pipeline latency.

3 Vivado High-Level Synthesis

Hardware description languages are mainly used to design hardware architectures and systems, where the design behavior is modeled in RTL primitives. On the other hand, high-level languages can efficiently model systems and applications with less effort and better ease of configurability. Therefore, in order to reduce the amount of design time of hardware-based systems, high-level synthesis tools are used for transforming the systems that are modeled in a high-level language into RTL implementation that can be synthesized into FPGAs [6]. HLS not only reduces the effort of HDL design and debugging but also provides flexibility in the final hardware implementation to meet design constraints. HLS also offers the potential to allow for hardware benchmarking in early design stages and for in-depth analysis of a design's resource usage versus high-level code placement.

The Xilinx Vivado HLS tool [20] transforms a C specification design, such as C, C++, and System-C, into an RTL design. The tool goes through several phases to achieve an optimized design. The HLS tool schedules logic operations for each clock cycle and assigns hardware resources for each scheduled operation. From the flow of the design, the control logic is extracted, creating a finite state machine which sequences the operations in the RTL design.

The Vivado HLS tool synthesizes the C functions into blocks within the RTL hierarchy. The top-level function arguments are synthesized, as appropriate, into I/O ports accompanied by appropriate hand-shaking signals. The HLS tool allows the hardware developers to analyze and optimize the design. Based on the default behavior, constraints, and any optimization directives; the tool creates an optimal RTL implementation of the high-level design. The tool then generates a set of synthesis reports which are used to analyze the implementation and, hence, several optimizations can be applied to meet the design constraints.

In this work, we implemented AES-128, AES-192, and AES-256 encryption algorithms in C-language. The algorithms were synthesized and co-simulated by the Vivado HLS tool to check the functionality of the RTL design. Then, the design was analyzed and optimized to achieve higher throughput. The throughput, T_p, in this paper is calculated as: $T_p = \frac{Block\_Size}{N_C * T_{CLK}}$, and the throughput to area ratio is calculated as: $K = \frac{T_p}{A}$, where *Block_Size* is the size of a block in bits, i.e., 128 bits, N_C is the number of clock cycles necessary to encrypt a single block, T_{CLK} is the maximum delay path, and area, *A,* is the number of slices from the Vivado utilization report. The Xilinx HLS tool enabled us to quickly realize our design and make optimizations which greatly increased the throughput of the AES algorithms.

4 AES Implementation Optimization

In this section, we will explain our implementation and optimization of the AES encryption algorithms for AES-128, AES-192, and AES-256 using Vivado HLS. We will explore two different implementations of the AES: 1) SW-based baseline implementation and 2) Pipelined-based optimization.

AES encryption algorithm is composed of *Nr* rounds of ciphering process, where *Nr* is 10, 12, or 14 for AES-128, AES-192, or AES-256, respectively. The ciphering round is a set of functions as described in Section 2. In HLS optimization, AES top function is synthesized into an RTL block and each sub-function into a sub-block that is instantiated into the top-level design. Additionally, the top-function arguments are synthesized into input/output (I/O) ports. The HLS tool allows us to choose the handshaking protocol to be implemented onto the I/O ports of the designed block(s). I/O ports can be implemented as streaming data from/to a FIFO, or as reading/writing data to/from a memory. There are other handshaking protocols which can be implemented, as necessitated by the design.

4.1 SW-Based Implementation

This scheme is the baseline AES algorithm that is designed for software implementation, where the key expansion is executed first before starting the encryption process. The purpose of the key expansion module is to generate the *Nr* + 1 different extended keys, each of size 128-bits. In this version, all loops are rolled and no optimization is applied. This implementation led to an encryption of one block in 2556, 2785,

and 2803 clock cycles using 154, 197, and 221 slices of the FPGA for AES-128, AES-192, and AES-256, respectively. The pseudocode used for the AES-128 algorithm is presented below.

4.2 Pipelined-Based Optimization

We applied our optimization directives to the SW-based design to achieve a high-performance encryption and a fully pipelined architecture. The aim of the design is to obtain maximum throughput by unrolling loops and applying pipeline with initiation interval = 1. For achieving high throughput, some arrays were fully partitioned and dependence directives were applied to overcome loop-carry dependencies. For comparison purposes, we constrained the HLS tool to synthesize the embedded memory blocks into FPGA slices. In this proposed architecture, the key is first expanded to generate the sub-keys for each round of the ciphering process. The sub-keys are registered for later use in the encryption operation. After expanding the key, the design performs a loop of four main functions, substituting bytes, shifting rows, mixing columns, and adding the sub-key. The trip count of the loop is fixed and based on the encryption version. These functions are inlined, pipelined, and optimized to achieve a highperformance design. The following code shows some of the applied Pragmas to achieve high throughput.

The proposed HLS-based optimization increases the throughput of the encryption process but at the expense of the FPGA's resources. For example, one block is encrypted in 19, 23, and 27 clock cycles for AES-128, AES-192, and AES-256, respectively, and occupying 2794, 3306, and 3750 slices of the FPGA and the maximum delay path is 4.566 ns. However, after executing one encryption process, each block of a plaintext is encrypted every clock cycle. Therefore, our proposed HLS-based pipelined architecture achieves a throughput of 28 Gbps.

As evident from these results, the throughput difference between the pipelined-based implementation and the SW-based implementation is significant. This result indicates that by simply implementing high-level code into a HLS tool and then optimizing in a way that is beneficial for design constraints, we can speed-up the development time and achieve a high-performance design. Therefore, by using HLS, rapid optimization can be accomplished with results similar to dedicated HDL designs.

The proposed optimization was compared to previous HLS implementations in the literature. The area utilization, in slices, maximum achieved frequency, throughput, and throughput to area ratio for the Pipelined-based implementation and the previous works are shown in Table 2. Our proposed optimization in HLS achieved a higher throughput of 28 Gbps. We achieved 10, 8.47, and 7.47 Mbps per slice for AES-128, AES-192, and AES-256 proposed architectures, respectively. Although our proposed AES-128 consumes 4-6 X more slices than [13] and [5], it is fully pipelined and an encryption process is executed every clock cycle.

The synthesized RTL design of the Pipelined-based optimization was exported to Vivado and the design implementation was completed for further analysis. The experiments were conducted using Xilinx Zynq-7000 SoC, Zedboard, XC7Z020-1CLG484C [21] along with the Xilinx Vivado Design suite and SDK 18.2. The FPGA fabric runs

```
// SW-based Encryption Function

#define Nr 10   // Number of rounds (10, 12, 14)
#define KS 16   // Key size in Bytes (16, 24, 32)


void AES_Encrypt(unsigned char state[16], unsigned char msg[16], unsigned char key[KS]){

for (int i = 0; i < 16; i++)
    state[i] = msg[i];

// Rounded key Generation
unsigned char expanded_key[KS*(Nr+1)];
key_expansion(key, expanded_key);

add_round_key(state, key);

for (int j = 0; j < Nr-1; j++) {
    sub_bytes(state);
    shift_rows(state);
    mix_columns(state);
    add_round_key(state, (expanded_key + (16 * (j + 1))));
  }

// Final round
sub_bytes(state);
shift_rows(state);
add_round_key(state, (expanded_key + 16*Nr));
}
```

```
// Pipelined-based Encryption Function

// Key Expansion function
static void KeyExpansion(unsigned char* key, unsigned char*  RoundKey){
#pragma HLS INLINE
#pragma HLS PIPELINE
#pragma HLS RESOURCE variable=s_box core=ROM_nP_LUTRAM
:
:
}

// Substitute each state value with another byte in the Rijndael S-Box
void sub_bytes(unsigned char* state) {
#pragma HLS INLINE
#pragma HLS PIPELINE
    for (int i = 0; i < 16; i++){
    #pragma HLS UNROLL
    state[i] = s_box[state[i]];
    }
}

// Encryptian Function
void AES_Encrypt(unsigned char state[16], unsigned char msg[16], unsigned char key[1ks]){
#pragma HLS ARRAY_RESHAPE variable=state complete dim=1
#pragma HLS ARRAY_RESHAPE variable=msg complete dim=1
#pragma HLS ARRAY_RESHAPE variable=key complete dim=1

unsigned char ExtendedKey[16];
#pragma HLS ARRAY_PARTITION variable=ExtendedKey complete dim=1
:
:
Round_Loop:for (uint4 round = 1; round < Nr; ++round)
    {
    #pragma HLS PIPELINE
    sub_bytes(tempstate);
      shift_rows(tempstate);
      mix_columns(tempstate);
      add_round_key(tempstate, &RoundKey[16*round]);
    }

:
:
}
```

Table 2: AES hardware implementation comparison

Implementation	Area				Max. Freq. Mhz	Throughput Gbps	Mbps/slice
	LUT	FF(SRL)	BRAM	SLICE			
AES-128 [1]	14588	5328	80	7670	144	1.53	0.2
AES-128 [13]	-	-	-	646	-	1.39	2.2
AES-128 [5]	1417	830	0	431	192	1.29	3
Our AES-128	9481	6787(32)	0	2794	219	28	10
Our AES-192	10925	8551(272)	0	3306	219	28	8.47
Our AES-256	13824	8827(832)	0	3750	219	28	7.47

normally on 100 MHz (used in this paper), but it can be configured up to the maximum frequency of the encryption module.

Figures 2, 3, and 4 show the execution time for encrypting one data block (i.e., 128bit) for AES-128, AES-192, and AES-256. After 19, 23, and 27 clock cycles, `ap_done` and `state_ap_vld` signals turn to high indicating that the encryption process is done. In these Figures also, the `ap_ready` signal is always high. This indicates that the encryption module is ready to receive a new input block to be

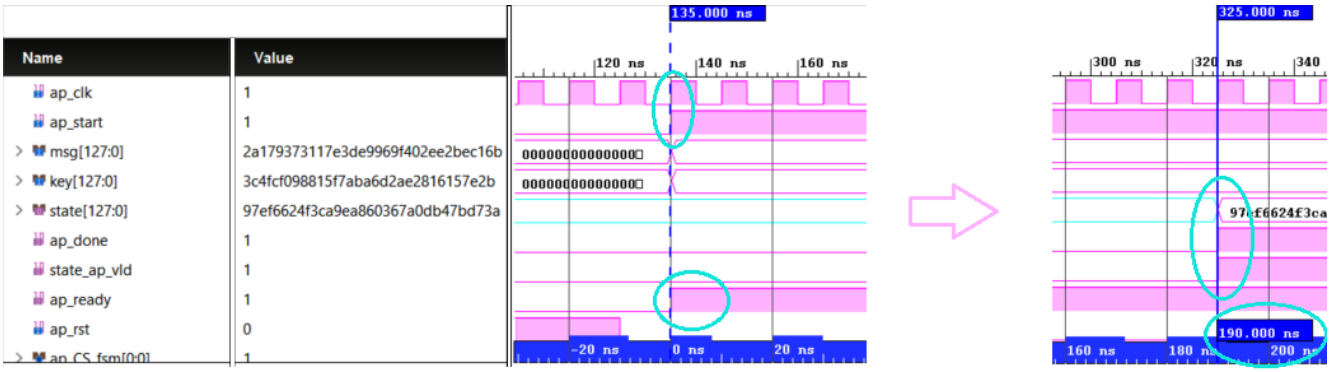


Figure 2: AES-128 simulation showing the start and end of 128-bit data encryption (190nsec, 19 cycles).

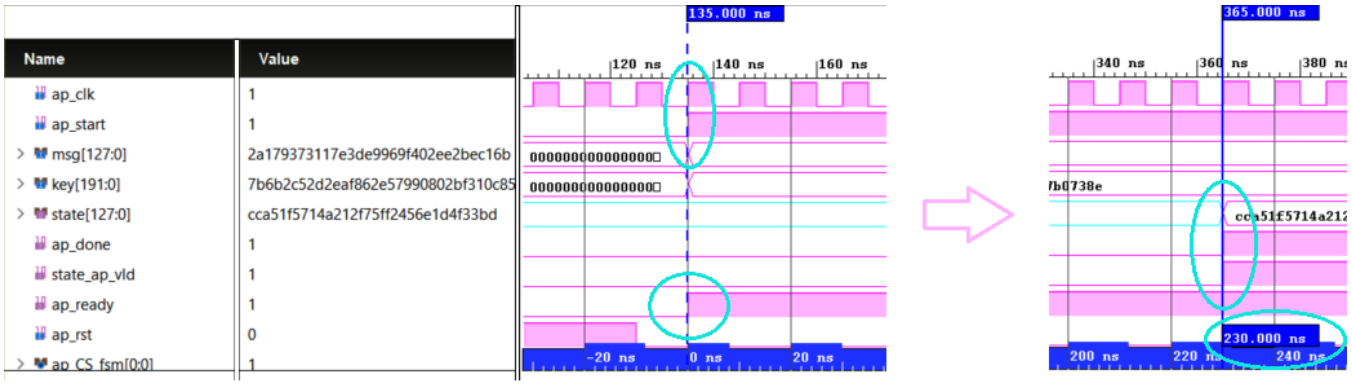


Figure 3: AES-192 simulation showing the start and end of 128-bit data encryption (230nsec, 23 cycles)

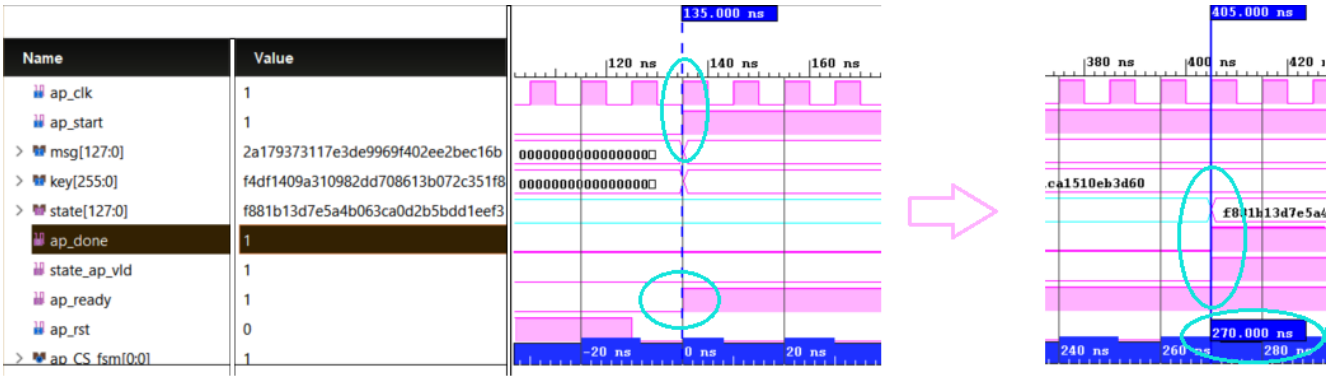


Figure 4: AES-256 simulation showing the start and end of 128-bit data encryption (270nsec, 27 cycles)

encrypted, i.e., an encrypted block can be received every clock cycle after filling the pipeline latency.

5 Conclusion

In this work, we explored the standard AES encryption and its implementation into a Xilinx ZedBoard with the Zynq-7000 APSoC. We focused on the encryption aspect of AES-128, AES-192, and AES-256, but the decryption part could easily be implemented and tested as well. The AES was initially coded in a high-level language and was then implemented with Xilinx Vivado High-Level Synthesis (HLS). The main goal of the optimization of AES is to reduce the initiation interval and allow concurrent execution of the operations in the loops and functions to achieve a pipelined architecture for the encryption process.

The encryption throughput of the proposed architectures in HLS was observed to be 28 Gbps. This rapid development and optimization of HLS-ready code showed that HLS can be used to increase a designer's productivity by applying directives such as pipelining, array shaping, and port mapping to their new and existing designs. A designer is thus able to see a moderate improvement without the need to design RTL with traditional, and time consuming, HDL languages.

In the future work, different modes of encryption for AES to encrypt successive blocks of data including counter (CTR), cipher block chaining (CBC), cipher feedback (CFB), and output feedback (OFB) can be implemented and optimized using HLS and compared to their counterpart RTL implementations on FPGA.

References

[1] S. Ahuja, S. T. Gurumani, C. Spackman, and S. K. Shukla, "Hardware Coprocessor Synthesis from an Ansi c Specification," *IEEE Design & Test of Computers*, 26:58-67, 2009.

[2] M. M. N. Biasizzo, "Hardware Implementation of AES Algorithm," *Journal of Electrical Engineering,* 56:265-269, 2005.

[3] C. Compiler, C2R Compiler, 2018.

[4] A. Dandalis, V. K. Prasanna, and J. D. P. Rolim, "A Comparative Study of Performance of AES Final Candidates using FPGAs," International Workshop on Cryptographic Hardware and Embedded Systems, 2000.

[5] L. Daoud, F. Hussein, and N. Rafla, "Optimization of Advanced Encryption Standard (AES) Using Vivado High Level Synthesis (HLS)," *Proceedings of 34th International Conference on Computers and Their Applications*, 58:36-44, 2019.

[6] L. Daoud and H. Selvaraj, *A Survey of High Level Synthesis Languages, Tools, and Compilers for Reconfigurable High Performance Computing, Advances in Systems Science*, Springer, pp 483-492, 2014.

[7] A. M. Deshpande, M. S. Deshpande, and D. N. Kayatanavar, "FPGA Implementation of AES Encryption and Decryption," 2009 International Conference on Control, Automation, Communication and Energy Conservation, INCACEC 2009, 2009.

[8] S. Drimer, T. Güneysu, and C. Paar, "DSPs, BRAMs and a Pinch of Logic: New Recipes for AES on FPGAs," 16th International Symposium on Field-Programmable Custom Computing Machines, 2008. FCCM'08., 2008.

[9] P. Hamalainen, T. Alho, M. Hannikainen, and T. D. Hamalainen, "Design and Implementation of Low-Area and Low-Power AES Encryption Hardware Core," 9th EUROMICRO Conference on Digital System Design: Architectures, Methods and Tools, 2006.

[10] H. S. Jacinto, L. Daoud, and N. Rafla, "High Level Synthesis Using Vivado HLS for Optimizations of SHA-3," IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), 2017.

[11] M. Kotegawa, K. Iwai, H. Tanaka, and T. Kurokawa, "Optimization of Hardware Implementations with High-Level Synthesis of Authenticated Encryption," *Bulletin of Networking, Computing, Systems, and Software*, 5:26-33, 2016.

[12] M. Latif, H. S. Jacinto, L. Daoud, and N. Rafla, "Optimization of a Quantum-Secure Sponge-Based Hash Message Authentication Protocol," IEEE 61st International Midwest Symposium on Circuits and Systems (MWSCAS), 2018.

[13] R. S. Meurer, T. R. Muck, and A. A. Frohlich, "An Implementation of the AES Cipher using HLS," 2013 III Brazilian Symposium on Computing Systems Engineering (SBESC), 2013.

[14] S. Morioka and A. Satoh, "An Optimized S-Box Circuit Architecture for Low Power AES Design," International Workshop on Cryptographic Hardware and Embedded Systems, 2002.

[15] S. Morioka and A. Satoh, "A 10-Gbps Full-AES Crypto Design with a Twisted BDD S-Box Architecture," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 12:686-691, 2004.

[16] N. I. S. T. F. I. P. S. Pub, "197: Advanced Encryption Standard (AES)," *Federal Information Processing Standards Publication,* 197:0311, 2001.

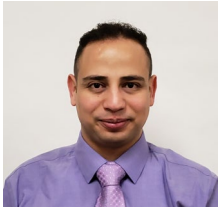
[17] C. Paar and J. Pelzl, *Understanding Cryptography: A Textbook for Students and Practitioners*, Springer Science & Business Media, 2009.

[18] I. Verbauwhede, P. Schaumont, and H. Kuo, "Design and Performance Testing of a 2.29-GB/s Rijndael Processor," *IEEE Journal of Solid-State Circuits*, 38:569-572, 2003.

[19] M. Watanabe, K. Iwai, H. Tanaka, and T. Kurokawa, "High-Speed Implementation of Encryption Circuit using a High-Level Synthesis Tool," *Bulletin of Networking, Computing, Systems, and Software*, 3:63-66, 2014.

[20] X. Inc., "Vivado Design Suite: High-Level Synthesis," July, 2018.

[21] X. Inc., "ZC702 Evaluation Board for the Zynq-7000 XC7Z020 User Guide," June, 2018.



Luka Daoud received the B.S. degree in Electrical Engineering from Fayoum University, Egypt in 2007, and M.S. degree in Electronics and Communications Engineering from Egypt-Japan University of Science and Technology (E-JUST), Alexandria, Egypt in 2012. Luka is currently a Ph.D. candidate in Electrical and Computer Engineering at Boise State University, Boise, Idaho, USA. His main research focuses on hardware security, Network-on-Chip (NoC), hardware acceleration, high performance reconfigurable computing, partial reconfiguration, and High-Level Synthesis (HLS) design.

Fady Hussein received a B.S. degree in electrical engineering from Birzeit University (Ramallah, Palestine) in 2005, an M.S. degree in electrical and computer engineering from Louisiana State University (Baton Rouge, Louisiana USA) in 2011, and a Ph.D. degree in electrical and computer engineering at Boise State University (Boise, Idaho USA) in 2017.

He is a Senior SIG Product Engineer at Micron Technology, Inc. (Boise, Idaho, USA). He worked as a Test Engineer at Micron for 6 years. Dr. Hussein is an IEEE Member and adjunct professor at Boise State University. His main research focuses on security, evolutionary computation, adaptive systems, hardware acceleration, and high-performance reconfigurable computing.

Nader I. Rafla, PhD., P.E., received his MS and PhD in electrical engineering from Case Western Reserve University, Cleveland, Ohio, USA, in 1984 and 1991 respectively. From 1991 to 1996, he was an Associate Professor in the department of manufacturing engineering at Central State University. In 1997, he joined the electrical and computer engineering program at Boise State University where he is currently an associate professor. His main research focuses on reconfigurable computing and neuromorphic architectures; evolvable hardware and genetic algorithms; cryptography and cybersecurity; Implementation and design of secure hardware architectures for the next-generation embedded systems and their applications that have immediate impact on real-world problems and related technical challenges such as data fusion, hardware security, and hardware acceleration.

Instructions for Authors

The International Journal of Computers and Their Applications is published multiple times a year with the purpose of providing a forum for state-of-the-art developments and research in the theory and design of computers, as well as current innovative activities in the applications of computers. In contrast to other journals, this journal focuses on emerging computer technologies with emphasis on the applicability to real world problems. Current areas of particular interest include, but are not limited to: architecture, networks, intelligent systems, parallel and distributed computing, software and information engineering, and computer applications (e.g., engineering, medicine, business, education, etc.). All papers are subject to peer review before selection.

A. Procedure for Submission of a Technical Paper for Consideration:

1. Email your manuscript to the Editor-in-Chief, Dr. Fred Harris, Jr. Fred.Harris@sce.unr.edu.
2. Illustrations should be high quality (originals unnecessary).
3. Enclose a separate page for (or include in the email message) the preferred author and address for correspondence. Also, please include email, telephone, and fax information should further contact be needed.

B. Manuscript Style:

1. The text should be, **double-spaced** (12 point or larger), **single column** and **single-sided** on 8.5 X 11 inch pages.
2. An informative abstract of 100-250 words should be provided.
3. At least 5 keywords following the abstract describing the paper topics.
4. References (alphabetized by first author) should appear at the end of the paper, as follows: author(s), first initials followed by last name, title in quotation marks, periodical, volume, inclusive page numbers, month and year.
5. Figures should be captioned and referenced.

C. Submission of Accepted Manuscripts:

1. The final complete paper (with abstract, figures, tables, and keywords) satisfying Section B above in **MS Word format** should be submitted to the Editor-in-chief.
2. The submission may be on a CD/DVD, or as an email attachment(s). **The following electronic files should be included:**
   - Paper text (required)
   - Bios (required for each author). Integrate at the end of the paper.
   - Author Photos (jpeg files are required by the printer)
   - Figures, Tables, Illustrations. These may be integrated into the paper text file or provided separately (jpeg, MS Word, PowerPoint, eps). title of the paper.
3. Specify on the CD/DVD label or in the email the word processor and version used, along with the title of the paper.
4. Authors are asked to sign an ISCA copyright form (http://www.isca-hq.org/j-copyright.htm), indicating that they are transferring the copyright to ISCA or declaring the work to be government-sponsored work in the public domain. Also, letters of permission for inclusion of non-original materials are required.

Publication Charges:

After a manuscript has been accepted for publication, the author will be invoiced for publication charges of $50 **USD** per page (in the final IJCA two-column format) to cover part of the cost of publication. For ISCA members, $100 of publication charges will be waived if requested.

January 2014