ISSN: 1813-1786
Volume No. 15
Indexed in:
ULRICH'S P. D.
PASTIC

TECHNICAL JOURNAL

2010

University of Engineering and Technology
Taxila

ISSN: 1813-1786
Volume No. 15
Indexed in:
ULRICH'S P. D.
PASTIC

TECHNICAL JOURNAL

2010

University of Engineering and Technology Taxila

EDITORIAL BOARD

CONTENTS

Design, Architecture, Standards and Protocols for Patient Information Systems

Khurram Zeeshan Haider[1], Farhan Aadil[2], and Rabia Attiqa[3]

Abstract

Electronic Medical Record (EMR) (sometimes also referred to as Electronic Health Record (HER) or Patient Information System (PIS)) systems are emerging as an essential part of the modern healthcare framework. EMR systems have now become an integrated, enterprise-wide providing access to patient healthcare data. EMR systems also serve as a core component of Clinical Decision Support System (CDSS). EMR systems promise to facilitate healthcare professionals with the necessary information to improve patient care efficiently and effectively. Though there has not been a great deal of research to provide a guideline for successful EMR systems implementation. This review paper provides a literature review of EMR from an architectural perspective and also lists some global EMR standards and protocols widely used in the development of EMR systems.

Keywords: Electronic Medical Record (EMR), Clinical Decision Support System (CDSS).

INTRODUCTION

EMRs can be defined as a software suite of integrated functionalities built around a common database. Such functionalities are not limited to, but typically include Electronic Health Records, Diagnostic Tools, Patient Billing, and Electronic Prescribing, Practice Management [1]. These modular functionalities are found at various integration levels. The term “electronic medical record” and terms such as “electronic health record (EHR)”, and “computerized patient record” are sometimes used to describe a person’s medical history in electronic form [1]. EMR software allows the users to create, store, edit, and retrieve patient charts on a computer. A successful EMR project allows a practice to replace its paper charts with electronic charts. This offers tremendous productivity and efficiency benefits to a practice. By storing all the data, an EMR replaces the racks of chart folders with a computer. EMR is not only a database application of patient medical records to retrieve and manipulate patient information but it also acts as a diagnosis tool and useful information can be used to facilitate doctors and physicians to make their job more efficient and easy.

Literature review yields that healthcare are lagging the application of information technology (IT) as compared to other fields. Healthcare is considered as knowledge based enterprise, but knowledge is not taken as the part of the value proposition which is the major cause of minimal academic research into healthcare information, compared to other industries. The EMR systems have emerged recently and a lot of research has been done of late. The information obtained from EMR is also being used in warehousing to develop clinical decision support systems (CDSS). It is evident from the literature that there will be great benefits from the integration of the healthcare and information technology disciplines. The EMR implementation is also critical in the sense that it has to work with as much sensitive data as somebody’s health. This reflects the importance of need for successful implementation of an EMR system.

With rapidly growing importance of information technology in the healthcare area, medical practitioners should have access to sound theoretical and practically relevant research to train them in the adoption, implementation, and use of such systems. EMR rely on real-time access to a common database, on a platform that aims to systematize, integrate, and streamline business processes and workflow. Wang et al. have discussed the effectiveness of Web Technology for their Web based intelligent on-line Monitoring system for Intensive Care Units (IMI) [12]. EMR systems provide accurate and timely information which could sometimes be very helpful in saving somebody’s life. To illustrate; in 2004, the American Medical Association reported 98000 preventable deaths per year due to information errors [1].

Warehousing of EMR data and data mining techniques allow mining for information that will allow healthcare providers to predict risks and measure medical care against benchmarks. The EMR has been augmented by a component that utilizes current technological developments such as internet technology to create a more complete source of healthcare data management [6].

[1, 2] Computer Engineering Dept., University of Engineering & Technology, Taxila, Pakistan, [3] Wah Engineering College, University of Wah

EMRs are the inevitable next step in the continued progress of healthcare. Medicine, perhaps the most information-intensive of all professions, is now ready— after many false starts—to take advantage of the advances in information technology that have transformed our society [2].

An EMR directly impacts patient care therefore making a successful transition to EMR may be the most important project that a medical practice can undertake. EMRs deal with information which involves the lives of the physicians, nurses, administrative staff, and the patients themselves. This makes the impact of an EMR implementation substantial. An unsuccessful project can be frustrating and expensive and even dangerous because as mentioned earlier; EMR involves the lives of the physicians, nurses, administrative staff, and the patients themselves. A successful EMR project has the potential of improving the clinical and administrative efficiency of medical practices, as well as enhancing overall quality of care [2]. Clinical decision support(CDS) can be (and has been) built starting from complementary knowledge representation models, for instance rule-based representations or workflow models to mention just a few [4] and EMR are the primary data source for CDS .

In spite of all the advantages of an EMR system there are some risk factors involved with EMR systems as well. There are also privacy and security fears associated with having people's medical records and history electronically accessible, although the relative newness of the technology means that little data exists regarding actual security breaches in these systems [1]. The EMR is relatively very young, compared to other industries in which IT is being used, with most articles from the last four years. Physicians, healthcare organizations, patients, insurance companies, pharmacies, and all other stakeholders in the healthcare value chain have a vested interest in successful implementation of these enterprise software systems [1]. There is also a need for proper training of medical practitioners to encourage the use of EMR. Physicians, especially those in private practice, are often overbooked with patients and may see the learning curve of an EMR system as too great a hindrance to workflow [1]. The physicians' behavior towards the use of technology will highly be influenced with the ease of use, EMR performance, time and financial constraints. There are some articles in the literature that depict the resistance of physicians towards the use of EMR. Strictly speaking, the EMR should have the capacity to help the physicians to make their job simple and efficient rather then being an overburden for them. The increasing use of integrated EMR systems is a relatively recent phenomenon; theory regarding the implementation process is sparse. Physicians are supposed to enjoy two major sorts of benefits from EMR system. Physicians will be able to a) see more patients in a day, due to time and workflow efficiencies offered by EMR systems or b) spend more quality time with the same number of patients.

Successful Implementation of EMR System Guidelines

William et al. [1] suggest the following guidelines for the successful implementation of EMR.

P1: Accomplishment of EMR or CPR projects depends on a clear business case for the project. For the accomplishment of the project, in pleasing circumstances, the calculated and cost effective explanation of the project is essential; on other hand it is also important to the healthcare organization's ability to assess the success of the project.

P2: Strong support from the practice physician(s) is indispensable for successful execution of EMR projects. In the case of EMR/CPR implementations, physician support can be seen as executive or top management support. There are previously some evidences that physician owned practices are less likely to adopt EMRs than practices owned by a healthcare organization [10], so physician buy-in is critical.

P3: Successful EMR implementation projects will be marked by an internal project "champion". For EMR implementations, champion will not necessarily be the practice doctor(s). While, physician support is essential, due to the time constraints of their practice, many physicians will not be able to play the role of project champion.

P4: In unbeaten EMR or CPR implementation projects there will be a vigilant and purposeful planning phase. The planning phase involves converting the business case into explicit goals and objectives for the implementation process. The same activity is performed while project resources are acquired.

P5: Presence of a project manager with strong project management skill and experience will lead the implementation of EMR project to success. Healthcare organizations, especially smaller practices, may

lack workers with project management experience and will need to look to independent consultants or vendor consultants to fill that need.

P6: Attractive EMR implementation projects will be marked by a willingness to change workflow and procedures on the part of the practice. Most complex and firmly incorporated software systems, such as EMR/ERP systems, are only configurable to a point, and typically require the adopting organization to be conventional their business processes to the software. According to William et al. business process reengineering (BPR) has become an accepted part of the price of implementing an enterprise information system and the implementation of EMR systems is likely no different. The theoretical model is displayed in figure 1 as given in [1].

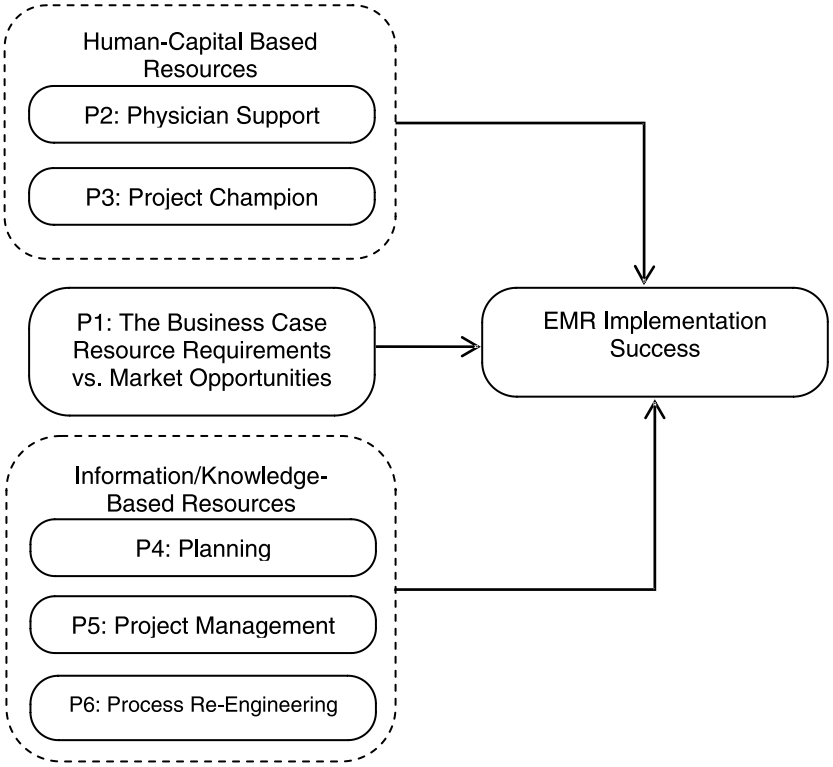


Figure 1: Theoretical model of an EMR System.

Advantages of EMR

An EMR building organization Wolf Medical System lists following major benefits of EMR [2].

Improved Patient Care: An EMR can dramatically improve patient care by sharing information easily, accessing and retrieving information quickly and easily, improving office communication, pro-actively managing patient care by using robust follow-up systems and rules engines. EMR also helps facilitating higher quality documentation (legible, organized, complete), providing built-in protocols and reminders (including health maintenance), and improving medication management.

Revenue Enhancement: An EMR can impact the top line of a practice by allowing physicians to invest time savings into seeing more patients. In addition, the robust follow-up system will ensure that patients are being seen appropriately for follow-up.

Improved Administrative Efficiency: Successful EMR sites are more efficient than traditional offices. These improvements can be attributed to the following :

1. Fewer chart pulls and less filing
2. Universal access to the chart (by more than one person at a time) and less searching for lost charts

3. Reduction in phone tag
4. Improved internal office communication
5. Fewer call-backs from pharmacies
6. Easier compliance with chart requests and chart audits

Improved Effectiveness: Adoption of an EMR allows you to practice in ways that you cannot with a paper chart. One such example is the graphing of electronic lab results versus medications. Creation of powerful rules for reminders – e.g. Diabetic patient that has not had a HbA1c in the past 6 months, automatic graphing of growth curves integrated tools for calculation of expected peak flow, mini-mental status, depression scores etc are also among the evidences of improved effectiveness provided by an EMR. Automatic calculation of cardiac risk and Automatic Drug Interaction checks for Drug to Drug, Drug to Allergy or Drug to Condition reactions are possible with EMRs.

Less Stress for Physicians: With the move to an EMR most physicians notice a definite decrease in stress as they become confident that their EMR will prompt them with appropriate reminders for patient care – they no longer need to hold everything in their heads! Along with this, the ease of access to the system from the home, hospital or office allows physicians to go home after their last patient leaves the office and finish up any remaining work from home as needed. Finally many offices note a real improvement in office morale due to the improvement in communication that results from the adoption of an EMR.

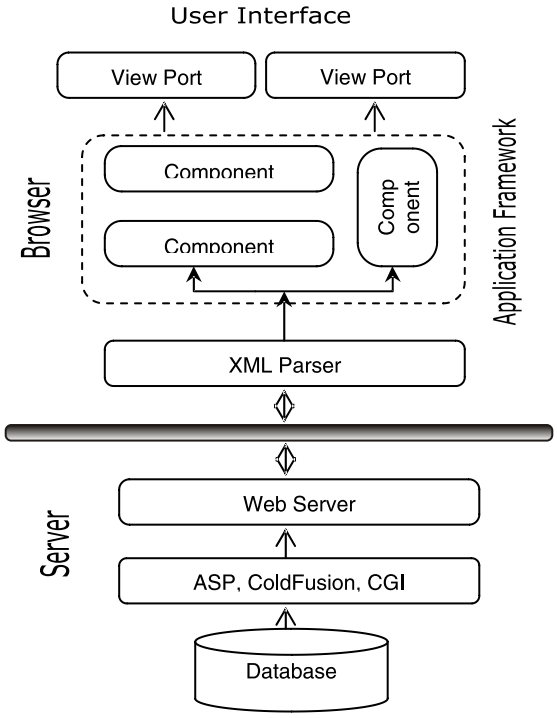


Figure 2: Architecture of EMR.

EMRs may also be implemented as a web portal. There is a wide range of web portals, such as search engines, available for public. These portals provide an interface to present data in an organized manner i.e. a straight forward means to access the data. Public as well as corporate portals provide access to potentially vast amounts of complex, distributed information through a Web browser. These are based on Web technologies and usually are accessed through the Web browser. Medical information itself is quite vast, complex and is distributed in different locations. Medical portals can also be made available via web and can provide the medical community with access to medical information through the Web browser. Appropriate portals and channels within those portals can be defined to provide access from the desk of the physician, the hospital administrator, the insurer or the consumer of health services [3].

Public such as Yahoo! and Google permit searching by query and by browsing hierarchical classifications of the Web-based information with each sub-tree or branch of the hierarchical classification rooted at the portal representing a major subject area such as education, health or entertainment etc. These sub-trees within portals are termed as “channels” of information. Corporate portals or enterprise information portals have been developed over the last few years. A corporate portal permits enterprise information to be made available across the company, normally through the Web [3]. A corporate portal can be thought of as a database view, i.e., the information channels within the portal are analogous to database attributes and portals may be defined differently for different individuals. These information channels permit access to information on the Web, and this Web-based information may be in a Web markup language such as HTML (Hypertext Markup Language) or XML (eXtensible Markup Language), in databases, in spreadsheets, or in word processing documents[3].

A patient sees a doctor who assesses health (and may or may not record it) and suggests the treatment for the patient based upon his/her present condition. At initial encounter the clinician collects information about patient health and today, most often, records the information on paper. Prior information about health status or lab information is often difficult to obtain, and the new information, when hand written on a piece of paper may also be unavailable for subsequent use at other locations [3].

The purpose of EMR is to accelerate the appropriate use of information technology to provide useful support for access to medical information for healthcare, research, teaching, health services administration and patient care. Ideally, a patient should have access to decision support tools to be informed about the need for a medical visit. Some of the patient information is fixed; some varies with the passage of time. In order to provide fixed and varied information, electronic information tools are needed which should be portable, fast, easy to use, connected to both a large valid database of medical knowledge and to the patient record, and will be a servant of patients as well as doctors. Over the last few years, there has been a shift from administrative health systems concerned primarily with billing procedures to clinical information systems that provide support for providers of health care [3]. Such clinical information systems may build on an EMR as a way to unify the data, even though that data may come from many sources and be of many different types. However, vendors at the May, 1999, TEPR Conference (Towards an Electronic Patient Record) indicated that while the use of administrative systems is almost universal for billing, less than five percent of physicians are using an electronic patient record [3]. According to the Computer-based Patient Record Institute (CPRI), A computer-based patient record (CPR) is electronically maintained information about an individual's lifetime health status and health care [3]. CPRs are also referred as Electronic Medical Records (EMRs). The content of CPR or EMR normally consists of medical history, laboratory test results, x-ray and/or CT scan images, current medications, etc. and these databases are not to enormous or huge, there are independent work stations at individual care sites with least connectivity requirements. With the maturity of the internet and World Wide Web (WWW), most system vendors are now presenting a web based clinical and medical systems. Most of these new systems are based, at least in part, on XML, the eXtensible Markup Language, defined by the World Wide Web Consortium (W3C) – a forum for information, commerce, communication, and collective understanding [3]. Most customers of health care desire web-access to the data due to uncomplicated access to appropriate health information. The major benefit of Web-based EMRs or CPRs is that these medical information systems grow to be more accessible to a wider range of users than ever before.

Even though doctors generally regulate with new information technology very gradually, the accessibility of any data or information, at any time, from any place, transforms the dynamics. Access of any data means access to multiple types of data and information including hand-written patient histories, medical images (X-Rays, CT scans etc.), lab reports, prescribing profiles etc. This information is also reachable 24 hours a day, 7 days a week, e.g., medical images (X-Rays, CT scans etc.) are available through the World Wide Web even though the bureau or workplace is closed. The data or information can be accessed straightforwardly without any restriction of particular place for it i.e. data can be accessed through the home computer, or from doctor’s desktop, or from a PDA (Personal Digital Assistant) connected to internet over wireless network. In this case the issues like security and privacy emerges, therefore the essential information should be available to the appropriate personnel(s) only. EMRs/CPRs should also ensure transparency of use, transparency of use is a fundamental attribute.

Medical Portals

Many systems provide portal like access across an intranet, rather than on the Web itself [3]. One such system is InfoClique[11] which is an intranet-based system that defines views by user type; health system administrator, clinical care coordinator, intranet staff, physician, physician office staff, and clergy. These views provide the access to the users based upon the rights of the users' group. In the InfoClique system, all data from the various systems are downloaded several times a day to either SQL servers or text files on one of two computers and these computers provide the intranet access for the users [3]. One major advantage of using EMR as a Web portal it that downloads to centralized servers are not required. This is due to the fact that, each contributing system contains a repository of accessible data. The system that first gets the data makes the data to the Web application, thus making the new data available to the users without downloading the data to the centralized server.

Three-tier architecture is most commonly used in particular for Web applications. Shepherd et al. [3] also presented three-tier architecture of the portal system. The components of the architecture are connected by the Internet. All three tiers are built on the Internet using Internet and Web technologies. The first tier consists of the access devices of the users. In most cases, these will be desktop computers running standard Web browsers, but with the interfaces configured for the appropriate portal. Each user would have access through a portal that would provide access to the required data. The hand held devices would connect to the Internet through a standard wireless communications system. The third tier consists of the data repositories and applications at the appropriate servers to which the user requires access. Second tier is the glue that makes this possible and consists of a proxy Web server and a suite of programs and databases. Data communications between the first and third tiers flow through and are controlled by this middle tier. It provides security and access to the data and applications in the third tier. When record of a patient is requested, that information may also distributed across many other Websites. The middle tier will identify the distributed parts of this virtual record and integrate them via a hypertext link structure, displayable in a Web browser at the first tier. In this system, once the second tier delivers the link to the first tier, the first tier can access the bottom or third tier directly without necessarily going through the middle tier again. This hypertext structure differs from the W3-EMRS architecture in that each tier-three server in the W3-EMRS architecture converts the required data to the Health Level 7 (HL7) message format and sends these messages to the middle tier. This middle tier, called the Agglutinator, integrates the data and converts the result into an HTML page and sends this page to the client in the first tier.

Users' portals can be accessed through a web browser. These portals can be made secure by password authentication. The portal software acts as the user's gateway to the required data and applications in the third tier. Access to JDBC (Java Database Connectivity) and ODBC (Open Database Connectivity) compliant databases in the third tier can be made directly over the Web through the use of appropriate device drivers in the second tier [3]. For non-Web compliant databases and resources, appropriate messaging protocols can be used between the second and third tiers to retrieve this data [3]. There are many messaging standards and protocols for the exchange of medical data. HL7 protocol standards are widely used for the electronic interchange of clinical, financial and administrative information among different health care oriented computer systems. Data Interchange Standard HL7 version 3 Clinical Document Architecture (CDA) is a XML encoded standard that specifies the structure and semantics of clinical documents for data storage. In case of a hand-held device portal, the middle tier will have to specially format the data for display on the device and balance the data flow due to the limited size of the target device and the limited bandwidth of the wireless communications network.

The distributed nature and workings of such a medical portal system should be transparent to the user. User is blind to the middle tier. There are three types of interaction between the first tier and the third tier, through the middle tier [3]. These are pull, push and update.

Pull: Pull is used as a standard Web technology to view a Web document or information in a Web accessible database. The relevant information is displayed to the user in the client browser whenever the user clicks a link or types in a Web server. In case of a medical portal, the user would download a patient record by requesting or pulling that information from the server. The request to pull goes to the middle tier which finds the information at the appropriate third-tier servers and returns that information to the user [3].

Push: When a Web server pushes information to the client browser, push occurs even though the user does not request for any information from the Web server. Push technology has been tested successfully for a real-time patient monitoring system [13]. Shepherd et al. illustrate that there are three types of push; continuous, periodic and triggered. A continuous push is the comparable of monitoring a device that is transferring continuous information. They gave an example of monitoring a patient's heart rate over a period of time. They elucidate the periodic push as if the server is sending information to the client browser at regular intervals. According to Shepherd et al. the example in this scenario might be a server that pushes stock market quotations every five minutes. The triggered push is enlightened by them as, when information is sent from the server to the client on the occurrence of a particular event, i.e., the event triggers the push, and an example of a triggered push would be the notification that a lab report is now available.[3]

Update: Update is the third and increasingly important portal interaction via which the information is updated on the web server. The user must be able to update or add information and this update must be made to data held by the appropriate third-tier server. The update occurs at the third tier via middle tier from the user. Such updates might include the entry of new observations during a physical examination, admission to a hospital, payment of a bill, new guidelines for care, etc. The update of pulled information might take place at the server but will not be visible to the user until the next pull (or download) of that information while the update of pushed information will be seen on the next push of that information from the server to the client [3]. For a triggered or continuous push, the client will receive that information immediately. If, however, the push is periodic, the client will have to wait on the average of half the length of the period before the server pushes the updated information back to the client [3].

2.2 Design Considerations

There are different types of portal interactions which have direct impact on the design of these portals. Handling of cache updates and security are influenced directly.

Cache updates: Most Web browsers keep the most recently viewed Web pages at the client in the cache to reduce network load and to decrease response time at the client. Whenever a client requests a page (pull), the browser first looks in the cache and, if the page is present, simply displays that page rather than go across the network to the server to pull the page again. This may cause problems for all volatile data that might be updated. Pushed data contains new or updated information for the user and will not be present in any document or information cached by the user. Similarly, the information at the server may be updated by one user while being pulled by another user. In this case, the user will not have access to the updated data even if the user does another pull as the browser will return the information resident in the cache, which is now out of date [3].

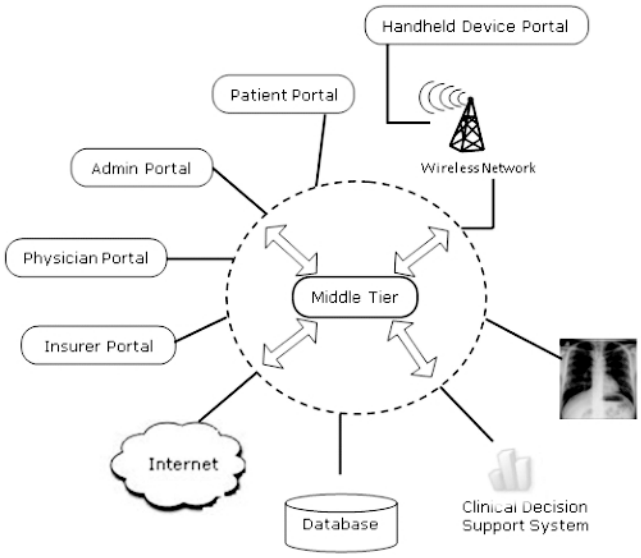


Figure 3: Three tier architecture for a medical portal

This problem can be solved by not caching any data but this will affect the performance. Solutions implemented in distributed databases and operating systems for cache coherency can also tackle this problem.

Automatic link association: In this case when updating occurs appropriate links must be automatically generated among patient's records because, the data is not centralized. Rather, the data is maintained in Web-accessible databases in the third tier. There are two possibilities to do this. First, This entails either that the third tier Web server notify the middle tier so that these links can be generated or that sophisticated dynamic link algorithms be present in the middle tier. Figure 3 depicts the three tier architecture for a medical portal [3].

EMR Standards for Interoperability

Berkowicz et al. [9] suggest that viewing and manipulating clinical data in its original form becomes difficult for the physicians because available data is in heterogeneous form in nature. In 1987, HL7 was introduced standards for the electronic interchange of clinical, financial and administrative data. It defines a communication between two independent applications rather than between closely coupled, client-server type applications. This is basically a messaging format having broad content definitions. It only uses its own flexible syntax for messages. Thus, to validate the structure of a particular HL7 message can be difficult. We have selected XML as syntax for data exchange and document type definition (DTD) has been used to define and validate data structures. XML is a subset of SGML. It has recently been accepted by the W3 consortium as a standard. HL7 has a special interest group (SIG) for SGML/XML representation of health care information. It defines a communication between two independent applications, rather than between closely coupled, client-server type applications.

Although internet IETF standards have led to much success with transport and application development, there standards do not attempt to assign semantic meaning to the data that is sent from application to application across the Internet [5]. Semantic and data encoding standards such as Health Level-7 (HL7) and XML are emerging from other standard development organizations.

Among several other benefits that an EMR can provide, it provides increased legibility and distributed access, to the most basic level. More benefits can be achieved by carefully structuring the data such that knowledge acquisition and data interoperability can be done efficiently. Implemented optimally, an EMR application should improve communication, enhance clinical decision making, improve compliance with documentation and treatment standards, minimize redundancy, enable context specific information presentation, integrate clinical documentation and billing functions, and facilitate quality improvement and clinical research [5]. The use of a well known standard for medical data can actually make feel these benefits. The lack of common standards for representing clinical data results in ambiguities effecting data structure and semantics. Some EMR standards – that can help enjoy the full benefits of an EMR – are discussed below.

Service Oriented Architecture

Both iRevive[1] and BCSEMR[2] are based on a service-oriented architecture providing the necessary agility and flexibility to link with the various internal and external healthcare systems that are essential to an independent, critical care environment. Both iRevive and BCSEMR exchange information via web services, which provides an API that is easy to program, thus promoting data exchange.

The Healthcare Information Technology Standards Panel

The importance of common standards to exchange medical information cannot be over emphasized. This topic was recently summarized in a report from the July 2006 hearing on the Functional Requirements for a Nationwide Health Information Network. The Healthcare Information Technology Standards Panel (HITSP) has defined a "Minimal Data Set" to describe many types of medical information at many levels. This minimal data set includes the semantic meaning of medical information that can be requested, acceptable

[1] iRevive is an out-of-hospital patient documentation application designed for Emergency Medical Care.
[2] BCSEMR is a web based in-hospital patient documentation system developed at Harvard's BWH.

responses, and the structure messages should take when exchanging information between applications. The recommendation from the HITSP committee includes a group of standards that harmonize many heterogeneous standardization efforts into a manageable group, in order to promote interoperability that will improve treatment and reduce costs.

Emerging EMR Standards

The many components of an EMR complicate the interoperability of health care applications. The key modules of a typical in-hospital EMR are administrative systems, clinical documentation, laboratory, radiology, pharmacy, and physician order entry. These areas have overlapping and competing standards, all of which have been developed by different organizations (e.g. HL7, CEN, and ASTM). Examples of overlapping terms include 11 different ways to define and spell "Total cholesterol" [Stanford and Thornton]. Below is a summary of some of the more important standards:

1. International Classification of Disease (ICD) is published by the World Health Organization. ICD is primarily used to identify a disease or problem for billing purposes.

2. Systematized Nomenclature of Medicine (SNOMED) was developed by a division of the College of American Pathologists to provide a "comprehensive, multi-axial, controlled terminology" [Stanford and Thornton] for indexing an entire medical record. SNOMED-CT (Clinical Terms) specifies the core file structure of SNOMED medical terms.

3. Logical Observation Identifiers, Names, and Codes (LOINC) is used to identify individual laboratory results, clinical observations, and diagnostic study observations.

4. Health Level 7 (HL7) is a messaging protocol for exchanging health care information. It includes several vocabularies, such as patient demographics. Unfortunately, there is poor backward compatibility between early and later versions of HL7; thus, there is poor compatibility between vendors who support different versions of the same standard.

5. National EMS Information System (NEMSIS is a standard for pre-hospital care endorsed by the National Highway Traffic and Safety Administration (NHTSA). This is a domain focused standard for out of hospital emergency medical services.

HITSP is reconciling these often overlapping and inconsistent standards to enable a consistent elemental description of the EMR. It is adopting LOINC for assessing a patient's condition. This instrument consists of a set of questions and allowable answers. The questions and answers specified by LOINC use the vocabulary of other standards (e.g. HL7, SNOMED, and ICD-9) to semantically define the meaning of each message. HL7 is then used as the messaging standard for defining the flow of messages being exchanged.

Protocols for EMR

Janhke et al. [7] suggest that the CDS system is not able to directly query the native database or data structures underlying the EMR system. Rather, we need to rely on a core patient summary in a standardized format to be sent to the CDS component for the purpose of analysis. One challenge is to determine exactly what data elements should be considered in this core patient summary. If the summary is defined too comprehensive, it may put too many requirements on the EMR system to interface with the CDS components. Moreover, large data volume may decrease the speed of invoking the CDS function to an unacceptable level. On the other hand, if the summary is too small, the usefulness of the CDS system may be decreased, e.g., cholesterol screening may require reviewing up to ten years of lab data). One solution would be to define a multi-step protocol, which allows the CDS component to "ask for more data", if required during the analysis. However, care must be taken to keep the protocol simple in order to lower the cost of integrating the CDS component into EMR systems, i.e., to facilitate adoption as much as possible. Figure 4 depicts a stateless, optional multi-step protocol for EMR and CDS system [7].

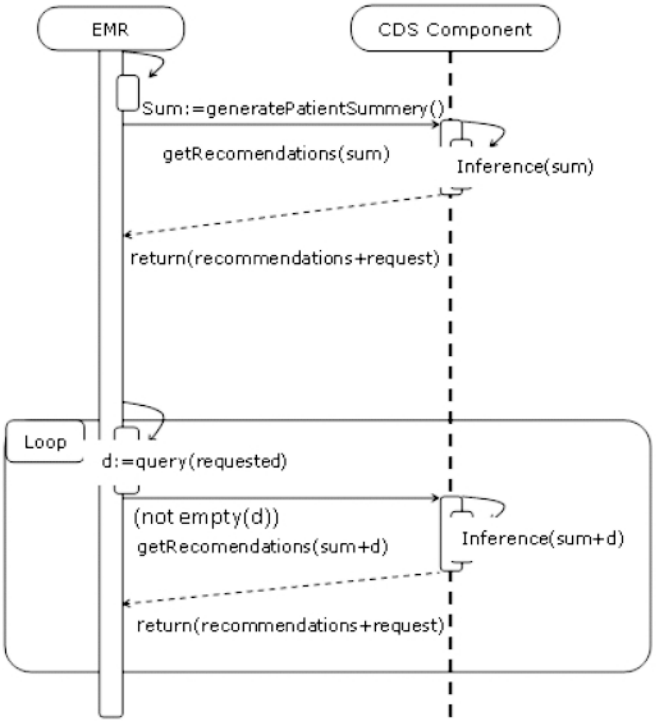


Figure 4: Multi-step Protocol for EMR and CDS

Conclusion

The importance, advantages, architecture, standards, protocols and issues have been discussed in detail. Some guidelines have also been mentioned for a successful implementation of an EMR system. Literature review yields that EMR are becoming an increasingly vital facet of patient safety, quality of care, and efficiency. They reduce medical errors, capture complete patient information, decrease the risk of law suits, improve reimbursement for services, ease compliance with regulations, and clinical decision support. Medical record systems so far have been shown to do the job but significant problems are encountered in the management of many systems. Web-based systems for information management will be the first step in making systems workable. Such systems will eliminate the problems caused by frequent power outages that may affect data storage causing loss and damage to data storage and backup. This will set the stage for more comprehensive development of EMR and then Patient Health Records (PHR).

References

[1] MacKinnon, W. Wasserman, M.: Integrated Electronic Medical Record Systems : Critical Success Factors for Implementation. In: Proceedings of the 42nd Hawaii International Conference on System Sciences, pp. 1–10, 2009

[2] Wolf Medical Systems, A Guide to Electronic Medical Records, http://www.wolfmedical.com (accessed May 12, 2009)

[3] Shepherd, M., Zitner, D., Watters, C.: Medical Portals: Web-Based Access to Medical Information. In: Proceedings of the 33rd Hawaii International Conference on System Sciences, pp. 1–10, 2000

[4] Verlaene, K., Joosen, W., Verbaeten, P.: Arriclides: An Architecture Integrating Clinical Decision Support Models. In: Proceedings of the 40th Hawaii International Conference on System Sciences, pp. 135–144, 2007

[5] Gaynor, M., Myung, D., Gupta, A., Rawn, J., Moulton, S.: Interoperability of Medical Applications and Devices. In: Proceedings of the 41st Hawaii International Conference on System Sciences, pp. 241–250, 2008

[6] Sood, S. P., Nwabueze, S.N., Mbarika, W.A.V., Prakash, N., Chatterjee, S., Ray, P., Mishra, S.: Electronic Medical Records: A Review Comparing the Challenges in Developed and Developing Countries. In: Proceedings of the 41st Hawaii International Conference on System Sciences, pp. 248–257, 2008

[7] Weber-Jahnke, J.H., McCallum, G.: A Light-Weight Component for adding Decision Support to Electronic Medical Records. In: Proceedings of the 41st Hawaii International Conference on System Sciences, pp. 251–260, 2008

[8] Wiggins, C., Pumphery, L., Beachboard, J., Trimmer, K.: Entrepreneurial Governance in a Rural Family Practice Residency Program. In: Proceedings of the 39th Hawaii International Conference on System Sciences, pp. 97a–106a, 2006

[9] Berkowicz, D.A., Barnett, G.O., Chueh, H.C.: Component Architecture for Web Based EMR Applications. In: Proceedings of the AMIA Annual Fall Symposium, pp. 116–120, 1997

[10] Burt, C.W., Sisk, J.E.: Which Physicians and Practices are Using Electronic Medical Records? In: Health Affairs, pp. 1334–1343, 2005

[11] https://www.infoclique.com/ (accessed May 12, 2009)

[12] Wang, K., Kohane, I., Bradshaw, K.L., Fackler, J.F.: A Real Time Patient Monitoring System on the World Wide Web In: Proceedings of the AMIA Annual Fall Symposium, pp. 729–732, 1996

Performance Sensitive Power Aware Multiprocessor Scheduling in Real-Time Systems

Ayaz Ali Khan[1], Muhammad Zakarya[2]

Abstract.

Multiprocessor environment is used for processor intensive real-time applications, where tasks are assigned to processor subject to some pre-defined criteria such as CPU load. Conventionally, real-time systems are paying attention on periodic task models, in which tasks are released at regular time periods. On the other hand, with maturity of multiprocessor structural design, today most real-time systems function in dynamic environment where human activities (aperiodic tasks) are predictable. Aperiodic tasks are to be completed as soon as possible. Consequently the priority assigned to such aperiodic tasks ought to be higher than those of periodic tasks. In distinction to its counter part the field of scheduling hybrid tasks i.e. periodic and aperiodic tasks on multiprocessor systems, still remains relatively unexplored. Similarly, higher power consumption issues arise as a challenge associated with such systems. Power aware scheduling is the cutting edge technique for reducing power constraints of multiprocessor systems. These systems generally remain under utilized thus becomes an ideal candidate for power aware scheduling. Recently a lot of work has been done on minimizing the energy requirement of processors. As a drawback of reducing power consumption of such systems, its response time is increased unreasonably, hence degrades the overall performance of the systems. In the prior work, higher importance is given to energy reduction while minimizing the response time of hybrid tasks is unnoticed. In this work we consider the importance of response time while reducing the power expenditure and utilization of the system. We propose a solution that reduces the power utilization and use of a multiprocessor system while the response time of tasks is kept with in a bound time. We assign aperiodic tasks to the processor that is under utilized for two reasons.

1. The underutilized processor has enough room to complete the aperiodic task with smallest possible time window and
2. There is a potential for running the system with low possible speed in addition to meeting deadline constraints.

The above solutions are obtained through mathematical foundation and experimental result supports our theoretical framework.

Keywords: (PEDS) Priority Exchange Dynamic Server, (RTS) Real Time System, (RM) Rate Monotonic, (DFS) Dynamic Frequency Scaling, (DM) Deadline Monotonic, (EDLS) Earliest Deadline Late Server, (EDF) Earliest Deadline First, (PDA) Personal Digital Assistant, (DVS) Dynamic Voltage Scaling, (WCET) Worst Case Execution Time

Introduction & concepts

Multiprocessor can be defined as a computer system having more than one processor, each one sharing system main memory & peripherals, to concurrently process programs. Hence the scheduling algorithms optimal on uniprocessor machines are not subject to be optimal on multiprocessor machines. So for multiprocessors are concerned, we use different scheduling algorithms. According to the law of Gordon Moore which states that the number of transistors in microprocessors would double every year [1]. But the advances in microprocessors are not face with battery power, and the battery capacity is tripled since 1990's [2, 3]. Therefore there is a need for applying energy efficient scheduling techniques that has become a major design consideration in realtime computing environment. There are many applications where aperiodic tasks are executed, as an outcome of exterior events. The time in which these proceedings take place are not controlled in reality by an application designer. For example, when keyboard buttons are pressed.

[1] COMSATS Institute of Information Technology, Islamabad, Pakistan, [2] Abdul Wali Khan University (AWKU), Mardan, Pakistan

When there are periodic and aperiodic tasks in the given system, our goal is to reduce the response time of aperiodic tasks in such a manner that all the periodic tasks are still optimal & schedulable. The most cut down technique is to give out available time slots, that are left unused by the periodic tasks, to aperiodic tasks. The background scheduling is used to schedule aperiodic tasks which is quite simple; however using background scheduling the desired response is never obtained. To reduce the response time of aperiodic tasks, we use the concept of aperiodic server. The aperiodic server or total bandwidth server services aperiodic requests as soon as possible [4]. The aperiodic server is consisting of a period and fixed execution time called server capacity and is scheduled with the same algorithm that is used for the periodic tasks.

There are two types of scheduling techniques for multiprocessors including

1. Partitioned scheduling and
2. Global scheduling

In partitioned scheduling, each task is assigned to a specific processor and then it is executed on that processor without migration. These processors are then scheduled independently and separately. This reduces the multiprocessor scheduling into a set of uniprocessor scheduling. With this scheduling we can use an optimal uniprocessor scheduling algorithm for multiprocessor systems. In partitioned scheduling the run time overhead is low as compared to global scheduling, because in partitioning technique the task does not migrate to other processors. But there is a shortcoming in schedulability bounds as the deadline to be missed if the total processor utilization exceeds ($\beta m + 1$) ($\beta + 1$), where $\beta = 1/\alpha$ and α is a maximum utilization of individual tasks [5]. Let $\alpha = 1$ and $m \rightarrow \infty$ then the processor utilization is bound by 50%. We use partitioning method in our work because the overhead is low so the context switching and energy consumption would be reduced. The alternative is the global scheduling, in which all the tasks are stored in a single priority queue. The scheduler selects the task having the high priority for execution. In global scheduling the tasks are not fasten to a particular processor and it can be executed on any processor. The optimal uniprocessor scheduling algorithm (EDF, RM etc) gives low utilization on multiprocessors. The global scheduling is best in the worst case schedulability. All the task sets are schedulable, if the processor utilization is less than or equal to 100%. But the number of context switching and migration is a problem. The worst case processor utilization is less than 50%.

In multiprocessors the main issue is heating and energy. Our goal is to minimize the energy consumption so that the cooling cost will be reduced. We are scheduling periodic and aperiodic tasks such that the load is balanced among different processors and the energy consumption is reduced. Runtime power reduction mechanisms can reduce the energy expenditure. For energy reduction we can use the DVS in latest processors. It means that power is a linear function of frequency i.e. f and a quadratic function of the voltage i.e. V given by ($p \propto fV^2$). The voltage adjustment at an instant of time is called DVS, which is an effective way for power saving in current systems [2].

In addition to saving energy, another advantage of having reduced power consumption is lower cooling cost of the multiprocessing environment (web farms, clusters etc).

In recent processors the relationship between frequency f and power p gives foundation to Dynamic Voltage Scaling

$$E = Pt \qquad (1)$$

Where E is energy consumed, t is time taken and P is power consumed. The average power dissipation in processor is:

$$P_{avg} = P_{capt} + P_l + P_{stdby} + P_{sc} \qquad (2)$$

Where P_{capt}, P_l, P_{stdby}, and P_{sc} is capacitance, leakage, standby and short circuit power. The P_l, P_{stdby} and P_{sc} are important but they are least important as compared to P_{capt}. So we will not consider P_l, P_{stdby}, and P_{sc}. So the P_{capt} is equal to:

$$P_{capt} = \alpha C V_d^2 f \qquad (3)$$

Where α is the transition activity dependant parameter, C, V_d and f is switched capacitance, supply voltage, and clock frequency. Equation (3) shows that the supply voltage V_d is quadratic as compared to clock frequency f; furthermore it also shows that lowering the supply voltage would be the most efficient way to reduce the power consumption. But when V_d is reduced then the circuit delay t_d would be increased:

$$t_d = \frac{m V_d}{(V_d - V_{tv})^2} \qquad (4)$$

Where t_{delay} is threshold voltage and m is a constant which will depend on gate size and capacitance. As from equation (4) the f and t_d are inversely proportional, so it would mean that the energy expenditure would be reduced in CMOS devices at the expense of performance delay. The frequency f is:

$$f = \frac{(V_d - V_{tv})^2}{k V_d} \qquad (5)$$

Equation (5) shows that the clock frequency is directly proportional to supply voltage. If we would consider $P = P_{capt}$, then equation (3) can be written as:

$$P = \alpha C V^2 f \qquad (6)$$

Equation (6) shows that when the clock speed f and voltage is changed then it would effect power consumption linearly and quadratically, respectively.

We use partition-based system where the workload is partitioned among processors. We derive results for a single processor and then extended it to multiprocessors. We would schedule mixed tasks (periodic and aperiodic tasks) with EDF scheduling algorithm, and those tasks, which have the earliest deadline, would have the highest priority. Our processor would have the discrete speed and voltage levels where $v_1 < v_2 < ..... < v_n$ and $f_1 < f_2 < .....f_{\max}$. We would consider the overhead of scheduling algorithm and voltage transition insignificant. The power spending by a processor with the speed f is given by $g(f)$ and the energy consumption during the interval $[t_1, t_2]$ is $\int_{t_1}^{t_2} g(f(t))dt$ [3].

Related Work & Existing techniques

In this section we discuss some existing mechanisms and techniques.

Real time systems are those systems that would give us results in a given time period. When a computer, that controls a device, sensor would give the data at regular time period and the computer responds by sending signals to an actuator, there must be a time bound in which the computer must respond to it. The ability of the system to respond with in a given time period depends on its capacity. If the system is unable to fulfill the demands, then we say that the system has inadequate resources. The system with limitless resources can fulfill the demands within the given time period. When the system is unable to respond within a given time period, it has different consequences i.e. there may be no upshot at all, or it may be minor or it may be catastrophic. The real time system is very simple like microcontroller or it may be complex like flight control system. Real time system examples are process control system, flight control system, intelligent highway systems, robotics, and high speed media communication system [6, 7].

Scheduling may be defined as the act of assigning resources to tasks. The scheduling can be applied to the tasks by defining the start time, so that no more than one task can request for a resource at that specific time, this is called static scheduling [xP90, xu93, foh94]. Other type of scheduling would be to assign priorities to tasks and then execute the task with the highest priority when there are more requests than the number of resources then the scheduling decision would be made. The priority scheduling can meet all deadlines where the time table schedule cannot. This situation occurs when the tasks that wants to execute is not known is advance or design, and the tasks execute when event occur.

Global Scheduling Algorithms, Such algorithms store the tasks in one queue that is mutual and common amongst all processing units. All tasks in the queue are maintained and are allocated with their specific priority. Suppose there are more than one task, then the task with the highest priority is selected from the queue and is executed on the specific processor using preemption and migration technique [21, 28]. Each processing unit preserves a status table that designates which tasks have previously committed to run. Additionally, each processing unit has a table that indicates how many processors have spare computational power. The time axis is split into slots, and these slots are of some fixed duration, and each processing unit on a regular basis sends information to its counterparts about the next slot that is free.

When the processor is overloaded it checks its surplus information and selects the processor that is most appropriate to complete the task with in its time period. However, it is possible that the surplus information is out of date and the selected processor can not execute the task. This problem can be solved by sending the task to the selected processor at the same time the originating processor ask from lightly loaded processor that how quickly they can execute the task. Then these responds are sent to the selected processing unit. When the selected processing unit is unable to execute the task with in its time period, then it can review the responds that which other processing unit is able to execute the task within its time period and then transfer the task to that processor.

Partitioning Scheduling Algorithms, Such algorithms divide the tasks in such a way that many task sets are created and then each task set is executed and schedulable on a particular processor. The tasks cannot migrate from one processor to the other, so the multiprocessor scheduling dilemma is altered into many uniprocessor scheduling dilemmas [21, 28]. In partitioning method several task sets are created and each task set is contained in different queues associated with each processor. As the multiprocessor scheduling problem is changed into many uniprocessor scheduling problem so we can use an optimal uniprocessor scheduling algorithm to schedule the tasks. Global scheduling strategies have many shortcomings over partitioning scheduling strategies. For example, Partitioning typically has a little scheduling overhead as put side by side to global scheduling, as tasks do not require to migrate across processing units. In addition, partitioning scheduling strategies divide a multiprocessing unit scheduling problem to a set of uniprocessor scheduling problem and then some optimal uniprocessor scheduling algorithms can be used. On the other hand, partitioning scheduling strategy has two disadvantages over global scheduling. First, to find the optimal handing over of jobs to the processing units is a bin-packing problem that is an NP-complete problem. Therefore, jobs are frequently partitioned using non-optimal heuristics. Secondly [13], there exists tasks that are schedulable if they exist non-partitioned. In spite all this partitioning strategies are extensively used. Additionally the hybrid of partitioning / global scheduling algorithm can be used. For example, at any instant of time the task is allocated to a single processing unit and is allowed to migrate as well.

Existing Problem

In recent times it is realized that there is a need for energy reduction in processors, a lot of work has been done on minimizing the system energy consumption. As a drawback of reducing energy consumption of the system, its response time is increased which degrades the overall performance of the systems. In prior work, higher importance is given to energy reduction and reducing response time of aperiodic tasks remains unnoticed.

We consider the importance of response time while reducing the power consumption of the multiprocessor system. With mathematical foundation we proposed a solution that reduces the power consumption of a multiprocessor system while the response time of tasks is kept within bound.

Proposed solution

Multiprocessor environment is used for processor intensive real-time applications, where tasks are assigned to processor subject to some pre-defined criteria such as CPU load etc. Traditionally, the focus of real-time systems are periodic task model where the release time of tasks are known, however with the advancement of multiprocessor design, the real-time systems are also using aperiodic tasks where the release time are not known in advance. The aperiodic tasks should be completed as quickly as possible; therefore the priority of aperiodic tasks must be greater than periodic tasks. In contrast to its counter part i.e. uniprocessor systems, the field of scheduling mixed tasks (periodic and aperiodic) on multiprocessor system still remains unexplored. Similarly, higher power consumption issues arise as a challenge associated with such systems. Power aware scheduling is the culling edge technique for reducing power constraints of multiprocessor systems. These systems generally remain under utilized thus becomes an ideal candidate for power aware scheduling.

In recent times it is realized there is a need for energy reduction in processors, a lot of work has been done on minimizing the energy reduction. As a drawback of reducing energy consumption of the system, its response time is increased, hence degrades the overall performance of the systems. In prior work, higher importance is given to energy reduction and reducing response time of hybrid tasks is unnoticed.

We consider the importance of response time also while the energy reduction is achieved. In our work we propose a solution that reduces the power consumption of a multiprocessor system while the response time of tasks is kept within bound.

Mixed workload scheduling (Periodic and aperiodic Tasks)

Periodic Tasks

Periodic tasks are those types of tasks that would appear after a fixed interval of time. A Periodic task set $T = \{T_1, T_2, ..., T_n\}$ that arrive at time t = 0, where every task T_i has Two parameters (p_i, c_i), where p_i is the time period and c_i is WCET of the task.

- All tasks are independent and preemptable.
- The task T_i has relative deadline D_i is equal to p_i.
- The released instance $R_{i,j}$ of task T_i is called the $j-th$ job of task T_i. The $R_{i,j}$ release Time would be $p_i \times (j-1)$.
- The task T_i WCET would be known in advance.

Aperiodic Tasks

Aperiodic tasks are those tasks that appear at any time and we doesn't know that when these tasks would reappear. The Aperiodic jobs $\{\sigma_m / m = 1, 2..... \}$ have two parameters (r, e), where r is the release time of job and not known in advance, e is the WCET of σ_m. We considering 'r' as the arrival time and 'e' as the execution time, the aperiodic load would be $\omega = e/r$. In our work we would change the load up to the maximum level. The aperiodic tasks would be run on a special type of server called Total Bandwidth Server. The TBS [4] has the capacity $u_s = c_s / p_s$ where c_s is the execution budget and p_s is period of the server. The $m-th$ aperiodic job σ_m, having the execution time e_m and arrival time r_m, is given the deadline:

$$d_m = \max(r_m, d_{m-1}) + \frac{e_m}{u_s} \qquad (7)$$

Where e_m is WCET. So Equation (7) shows that when we have higher u_s then d_m would be earlier.

SCHEDULING MIXED WORKLOAD WITH M PROCESSORS:

According to EDF, a task set is schedulable iff

$$u_{tot} = \sum_{i=1}^{n} \frac{c_i}{p_i} \leq 1 \qquad (8)$$

Where, u_{tot} is total system utilization, c_i is execution time and p_i is the time period. With periodic tasks we have also aperiodic tasks so we must also consider it with periodic tasks, such that the utilization of periodic and aperiodic tasks must be less than or equal to one.

$$u_p + u_s \leq 1 \qquad (9)$$

In Equation (3) we have $0 \leq u_p, u_s < 1$ and $0 < u_p + u_s \leq 1$, at frequency $f = f_m$ ($f = 1$ in this case), as f =1 so we say that this approach is not a DVS approach because system is running at full speed.

Equation 9 shows that the system is running at full speed and gives us the lowest system utilization. So all the tasks running at their WCET, so the system utilization is far less than 1. It means that the processor is doing nothing for most of the time. It also means that the energy is wasted during the idle time intervals, as the processor is running at their full speed. The energy consumption can be reduced by lowering the speed of the processor, but lowering the speed would take the task longer to complete and the response time of the tasks would be increased. According to [39], the frequency component is added to Equation (9) as:

$$u_p + u_s \leq \frac{f_i}{f_m} \qquad (10)$$

Where f_m is maximum speed of the processor and f_i is the suitable speed so that the task set is feasible schedulable. We represent the initial speed of the processor f_i by f_{static} , and we denote f_i / f_m by α_b .

Response Time Constraint

As Equation (10) gives us the lowest frequency, so that the mixed tasks are feasibly schedulable but the execution time are scaled by a factor of $1/\alpha_b$. When we decrease the frequency then the voltage consumption will be reduced, and according to Equation $p_{cmos} = v^2 f$, when we reduce the speed then the power consumption would be reduced, but the response time of the task would be increased and the system performance would be degrade. When we run an application then the energy requirement at time t would be $E = p.t$. So energy consumption would be $E \propto v^2$.

The tasks are running at lower frequency α_b , so the execution times of the tasks are increased. The deadline of TBS [40] would become:

$$d_m(\alpha_b) = \max(r_m, d_{m-1}) + \frac{e_m}{u_s . \alpha_b} \qquad (11)$$

The deadline for TBS is delayed as

$$d_m(\alpha_b) - d_m = \max(r_m, d_{m-1}) + \frac{e_m}{u_s . \alpha_b} - \max(r_m, d_{m-1}) + \frac{e_m}{u_s} = \frac{e_m}{u_s}\left(\frac{1}{\alpha_b} - 1\right) \qquad (12)$$

It is clear from Equation (6) that the deadline is increased, so the response time of aperiodic tasks would also be increased. The aperiodic tasks are very less in real time applications as compared to periodic tasks such as java based videophone, which runs the garbage collector (aperiodic task) almost every 600ms for every 3.732ms [40]. In those systems where aperiodic tasks came less frequently as compared to periodic tasks, and those aperiodic tasks need quick response then one solution is Equation (6) in which we run the task at full speed. But there is a disadvantage as the curve of energy and voltage is convex in nature [41]. When we increase the voltage then the power consumption would be increased quadratically. Equation (6) shows that we decrease the scheduling priority and the response time of aperiodic tasks is increased. We consider that when the response time of aperiodic task is smaller than p_s (worst case), then there is no need of frequency scaling. To

avoid the performance degradation of aperiodic tasks, we restrict this deadline delay. As it is showed earlier that the TBS has to execute aperiodic jobs for c_i intervals during any interval of length p_s In our work, when we apply DVS then this delay must be less than or equal to p_s i.e. $d_m(\alpha_b) - d_m \leq p_s$. As we have a range of speed levels ($f_1 < f_2 < .....f_m$), authors in [42] gives a technique to find suitable frequency f_k for aperiodic job σ_m.

$$\alpha_b = (1 + \frac{p_s.u_s}{e_m})^{-1} \qquad (13)$$

In case ($d_m(\alpha_b) - d_m \leq p_s$), our algorithm completes aperiodic load before p_s. Equation (13) is for single processor, however, for multiple processors we need to find the α_b on all m processors and allocate the aperiodic task σ_m to the processor having lower α_b. We would use the partitioning scheduling method for the periodic tasks and global scheduling method for aperiodic tasks.

The above formulation is valid for a single processor, while the intended purpose of this work is to accommodate aperiodic tasks on multiprocessor system and make sure aperiodic jobs completed as soon as possible, and no periodic task sever miss the deadline. We are using global scheduling mechanism for the aperiodic jobs.

Initially, periodic tasks are assigned to all available processors i.e., periodic load is uniformly distributed among processors. Also, every processor has a TBS for aperiodic tasks and it's very likely that TBS has different capacity on all processors and the larger capacity it has, the better it would be, because aperiodic job will complete much early as compare to low capacity TBS.

With our approach, let processor 1 has TBS and the required speed for completing aperiodic job is obtained with α_{1b}, where subscript 1 points to processor 1.

$$\alpha_{1b} = (1 + \frac{p_s.u_s}{e_m})^{-1}$$

The required speed for processor 2 is obtained with:

$$\alpha_{2b} = (1 + \frac{p_s.u_s}{e_m})^{-1}$$

Similarly for m-th processor it would be:

$$\alpha_{mb} = (1 + \frac{p_s.u_s}{e_m})^{-1}$$

We get speed for all the processors. Once this step is done, we encounter the solutions:

Run Aperiodic Task on Slowest Possible Speed

We determine the lowest possible speed for aperiodic task on all the processors $\alpha_l = \min(\alpha_{1b}, \alpha_{2b}, ...\alpha_{mb})$ so that it gets completed with the bounded time where α_l is the lowest speed of all processors. This solution result in reducing energy consumption of the over all system, as it runs on lowest speed. However, the response time of aperiodic task gets large due to the lowest system speed

Run Aperiodic Task with Highest Possible Speed

This solution gives the maximum available speed for aperiodic task on all the processors. In other words: $\alpha_h = \max(\alpha_{1b}, \alpha_{2b}, ...\alpha_{mb})$. We find the speed of aperiodic task on all the processors $1,2,3,..... m$ and then the processor with the highest possible speed α_h is selected. This means that the TBS on that particular processor has largest capacity and can only respect the time constraint when executed on the α_h (highest speed) i.e. the processor I needs α_h (highest) speed to completed

aperioid task within period. As discussed earlier, running a processor at maximum speed mean consuming maximum system power, which can not be compromised unnecessarily. In this work, since, we are maintaining a queue of aperiodic tasks, a particular aperiodic task will be consider for execution at individual processors and the one which can execute the task with highest speed is assigned the task. We opt for the first option (Run aperiodic task on slowest possible speed) because it will result in lower energy consumption and the aperiodic task will be completed with assigned time window, which is the main contribution of the work. The proposed technique will make sure to complete the aperiodic task within the time window and reduce the total power consumption of the multiprocessor system. In other words, based on ps of all processors, our technique will find the smallest speed α_l such that aperiodic task gets completed within a permissible time.

Implementation, Simulation and Results

First we produce periodic tasks and calculate their utilization and average values. Similarly the task set generated is then divided and mapped onto multiprocessors. Utilization of all tasks are shown and aperiodic tasks are assigned to processors that are under utilized i.e lesser periodic load is assigned. As a final outcome, response time of aperiodic task, to frequency of the system and corresponding power consumption is drawn at the end.

Simulations Study

Periodic Tasks

Execution Time	Time period	Utilization
20.0000	204.0000	0.0980
30.0000	214.0000	0.1402
14.0000	214.0000	0.0654
17.0000	218.0000	0.0780
27.0000	231.0000	0.1169
18.0000	257.0000	0.0700
29.0000	259.0000	0.1120
27.0000	261.0000	0.1034
25.0000	261.0000	0.0958
21.0000	288.0000	0.0729
4.0000	288.0000	0.0139
30.0000	290.0000	0.1034
5.0000	310.0000	0.0161
13.0000	323.0000	0.0402
1.0000	332.0000	0.0030
3.0000	348.0000	0.0086
27.0000	352.0000	0.0767
24.0000	365.0000	0.0658
30.0000	383.0000	0.0783
2.0000	385.0000	0.0052
14.0000	400.0000	0.0350
29.0000	406.0000	0.0714
9.0000	409.0000	0.0220
10.0000	410.0000	0.0244

22.0000	414.0000	0.0531
19.0000	420.0000	0.0452
2.0000	421.0000	0.0048
18.0000	429.0000	0.0420
9.0000	440.0000	0.0205
19.0000	448.0000	0.0424
29.0000	449.0000	0.0646
4.0000	454.0000	0.0088
27.0000	486.0000	0.0556
6.0000	486.0000	0.0123

Total Utilization = 1.8661
Average = 0.6220

Tasks Executed on Processor 1 =

Execution Time	**Time period**
20	204
30	214
14	214
17	218
27	231
18	257
29	259
27	261

Tasks Executed on Processor 2 =

Execution Time	**Time period**
0	0 (8 Times)
25	261
21	288
4	288
30	290
5	310
13	323
1	332
3	348
27	352
24	365
30	383
2	385
14	400

29	406
9	409
10	410
22	414
19	420
2	421

Tasks Executed on Processor 3 =

Execution Time	Time period
0	0 (27 Times)
18	429
9	440
19	448
29	449
4	454
27	486
6	486

Utilization of Processor 1 = 0.7840
Utilization of Processor 2 = 0.8360
Utilization of Processor 3 = 0.2461
The less utilized Processor is Processor 3 and it can execute Aperiodic load up to 0.7539

Aperiodic Tasks

Execution Time	Time period	Utilization
1.0000	670.0000	0.0015
1.0000	666.0000	0.0015
1.0000	655.0000	0.0015
2.0000	982.0000	0.0020
2.0000	952.0000	0.0021
2.0000	619.0000	0.0032
3.0000	883.0000	0.0034
4.0000	749.0000	0.0053
5.0000	869.0000	0.0058
5.0000	825.0000	0.0061
4.0000	616.0000	0.0065
6.0000	802.0000	0.0075
8.0000	986.0000	0.0081
8.0000	923.0000	0.0087
5.0000	527.0000	0.0095
8.0000	786.0000	0.0102
7.0000	651.0000	0.0108

10.0000	909.0000	0.0110
9.0000	665.0000	0.0135
11.0000	783.0000	0.0140
10.0000	681.0000	0.0147
14.0000	951.0000	0.0147
13.0000	872.0000	0.0149
11.0000	716.0000	0.0154
11.0000	667.0000	0.0165
14.0000	822.0000	0.0170
12.0000	680.0000	0.0176
12.0000	675.0000	0.0178
11.0000	604.0000	0.0182
10.0000	546.0000	0.0183
19.0000	974.0000	0.0195
13.0000	616.0000	0.0211
19.0000	891.0000	0.0213
20.0000	933.0000	0.0214
17.0000	788.0000	0.0216
12.0000	549.0000	0.0219
12.0000	520.0000	0.0231
13.0000	557.0000	0.0233
24.0000	989.0000	0.0243
18.0000	722.0000	0.0249
21.0000	839.0000	0.0250
25.0000	998.0000	0.0251
13.0000	518.0000	0.0251
22.0000	830.0000	0.0265
25.0000	934.0000	0.0268
24.0000	839.0000	0.0286
20.0000	698.0000	0.0287
17.0000	591.0000	0.0288

Aperiodic load =

Columns 1 through 8

0.0015 0.0030 0.0045 0.0066 0.0087 0.0119 0.0153 0.0206

Columns 9 through 16

0.0264 0.0324 0.0389 0.0464 0.0545 0.0632 0.0727 0.0829

Columns 17 through 24

0.0936 0.1046 0.1182 0.1322 0.1469 0.1616 0.1765 0.1919

Columns 25 through 32

0.2084 0.2254 0.2430 0.2608 0.2790 0.2974 0.3169 0.3380

Columns 33 through 40

0.3593 0.3807 0.4023 0.4242 0.4472 0.4706 0.4948 0.5198

Columns 41 through 48

0.5448 0.5698 0.5949 0.6215 0.6482 0.6768 0.7055 0.7342

Response time =

Columns 1 through 8

0.0001 0.0003 0.0009 0.0009 0.0020 0.0158 0.0195 0.0284

Columns 9 through 16

0.0327 0.0386 0.0525 0.0593 0.0593 0.0617 0.0791 0.0867

Columns 17 through 24

0.0994 0.1023 0.1144 0.1221 0.1337 0.1350 0.1393 0.1496

Columns 25 through 32

0.1617 0.1679 0.1795 0.1913 0.2058 0.2224 0.2229 0.2369

Columns 33 through 40

0.2405 0.2425 0.2500 0.2666 0.2843 0.3005 0.3005 0.3106

Columns 41 through 48

0.3163 0.3163 0.3345 0.3409 0.3433 0.3493 0.3607 0.3761

Response Time Frequency =

Columns 1 through 8

0.0005 0.0013 0.0027 0.0029 0.0044 0.0188 0.0233 0.0331

Columns 9 through 16

0.0385 0.0458 0.0612 0.0698 0.0715 0.0759 0.0954 0.1053

Columns 17 through 24

0.1206 0.1262 0.1413 0.1520 0.1669 0.1717 0.1797 0.1938

Columns 25 through 32

0.2099 0.2203 0.2364 0.2528 0.2720 0.2937 0.2993 0.3187

Columns 33 through 40

0.3279 0.3357 0.3491 0.3718 0.3958 0.4186 0.4251 0.4420

Columns 41 through 48

0.4547 0.4615 0.4867 0.5003 0.5101 0.5237 0.5428 0.5662

Response Static Time =

Columns 1 through 8

0.0001 0.0004 0.0011 0.0013 0.0027 0.0168 0.0212 0.0309

Columns 9 through 16

0.0364 0.0440 0.0597 0.0688 0.0714 0.0770 0.0980 0.1096

Columns 17 through 24

0.1269 0.1351 0.1531 0.1674 0.1863 0.1959 0.2094 0.2295

Columns 25 through 32

0.2523 0.2701 0.2942 0.3192 0.3477 0.3791 0.3955 0.4265

Columns 33 through 40

0.4483 0.4698 0.4981 0.5362 0.5762 0.6157 0.6405 0.6767

Columns 41 through 48

0.7098 0.7387 0.7866 0.8242 0.8595 0.9001 0.9475 1.0000

Maximum Energy Consuumed =

Columns 1 through 14

1 1 1 1 1 1 1 1 1 1 1 1 1 1

Columns 15 through 28

1 1 1 1 1 1 1 1 1 1 1 1 1 1

Columns 29 through 42

1 1 1 1 1 1 1 1 1 1 1 1 1 1

Columns 43 through 48

1 1 1 1 1 1

Energy Required =

Columns 1 through 8

0.0212 0.0424 0.0636 0.0675 0.0887 0.1099 0.1311 0.1522

Columns 9 through 16

0.1734 0.1946 0.2158 0.2370 0.2582 0.2794 0.3006 0.3218

Columns 17 through 24

0.3430 0.3642 0.3854 0.4066 0.4278 0.4490 0.4702 0.4913

Columns 25 through 32

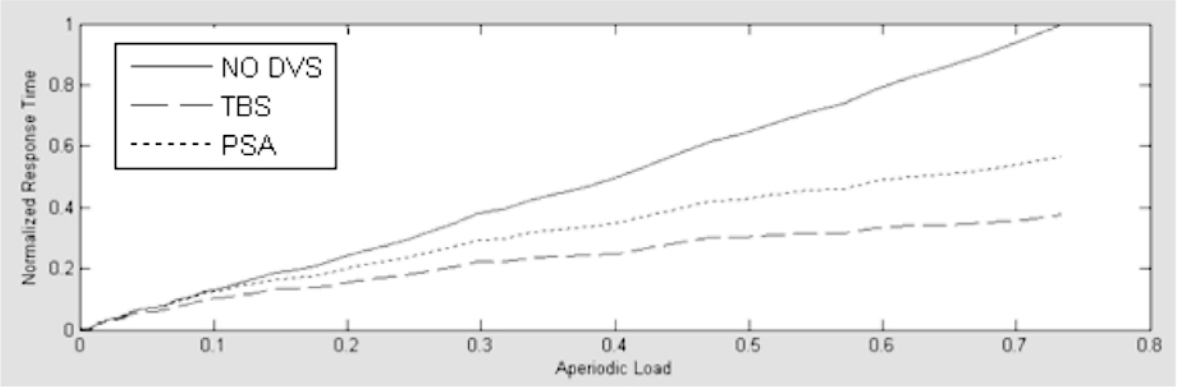
0.5125 0.5337 0.5549 0.5761 0.5973 0.6185 0.6397 0.6609

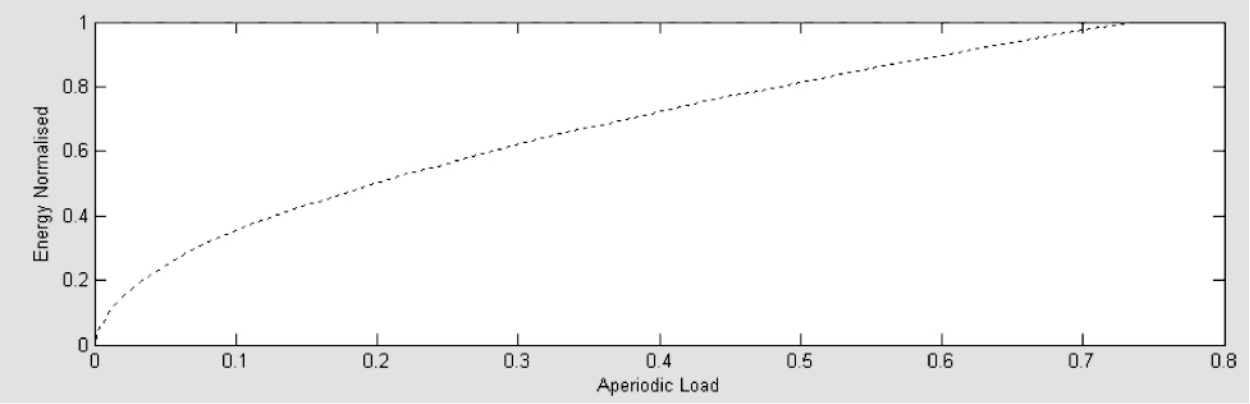
Columns 33 through 40

0.6821 0.7033 0.7245 0.7457 0.7669 0.7881 0.8093 0.8304

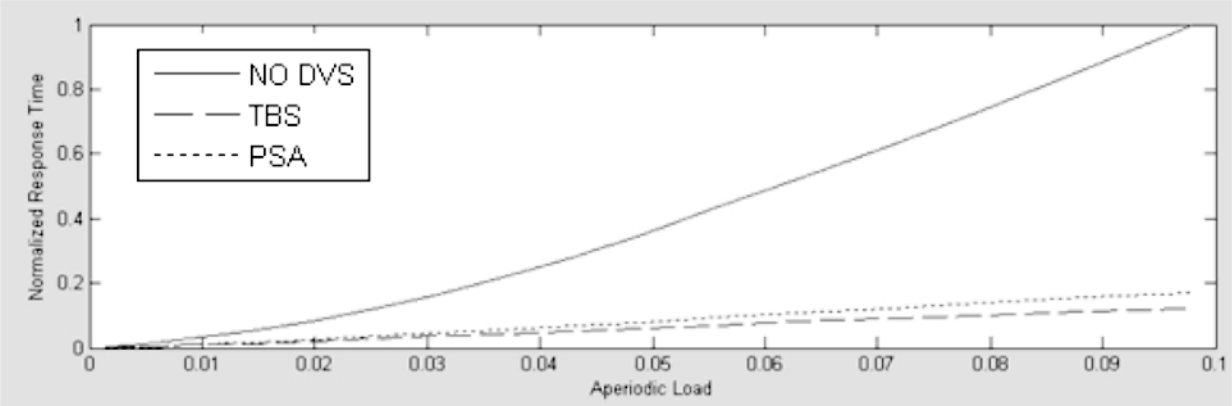
Columns 41 through 48

0.8516 0.8728 0.8940 0.9152 0.9364 0.9576 0.9788 1.0000

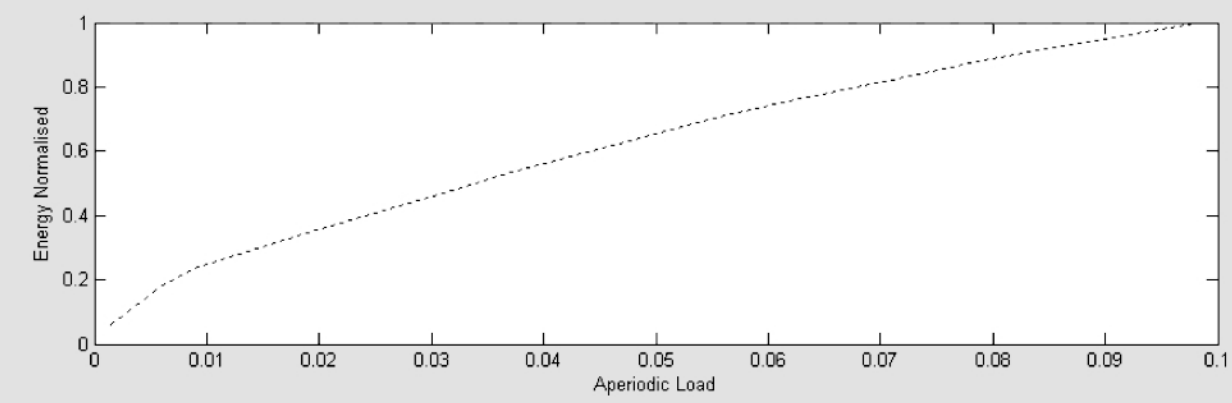




Energy Consumption Aperiodic load (Case 1)



Response Time versus Aperiodic load (Case 2)



Energy Consumption Aperiodic load (Case 2)

Conclusion & Future Work

Recently a lot of efforts are put for energy reduction of processors. As a drawback of reducing energy consumption of the system, its response time is increased, hence degrades the overall performance of the systems. In prior work, higher importance is given to energy reduction and reducing response time of hybrid tasks is unnoticed. We consider the importance of response time also while the energy reduction is achieved. In our work we propose a solution that reduces the power consumption of a multiprocessor system while the response time of tasks is kept within bound. In future the algorithm can be extended to accomplish and schedule large number of jobs over a huge network environment like Grid and Cloud Computing.

References

[1] A. P Chandrakasan, S. Sheng, and R. W. Brodersen. Low Power CMOS Digital Design, IEEE journal of Solid State Circuits, 1992, pp .472-484.

[2] T. D. Burd., T. A. Pering, A. J. Stratakos, and R. W. Rodersen, Adynamic Voltage Scaled Microprocessor system. IEEE Journal of Solid State Circuits, Vol. 35, No. 11,pp. 1571-1580,2000.

[3] D. Shin and J. Kim:, Dynamic voltage scaling of mixed task sets in priority-driven systems. IEEE Transaction on CAD of Integrated Circuits and Systems 25(3): 438-453 (2006).

[4] M. Spuri and G. Buttazzo, Scheduling Aperiodic Tasks in Dynamic Priority Systems, Journal of Real-Time Systems, 10(2):179-210, 1996.

[5] P. Pillai, and K. G. Shin, Real-Time Dynamic Voltage Scaling for Low-Power Embedded Operating Systems, In Proc. of ACM Symp. On Operating Systems Principles, pages 89-102, 2001.

[6] M. Joseph, “Real-time Systems: Specification, Verification and Analysis,” Prentice Hall, 1996.

[7] P. A. Laplante, “Real-time Systems Design and Analysis, An Engineer Handbook,” IEEE Computer Society, IEEE Press, 1993.

[8] C. M. Krishna and K. G. Shin, “Real-Time Systems,” MIT Press and McGraw-Hill Company, 1997.

[9] G. C. Buttazzo, “Hard Real-Time Computing Systems: predictable scheduling algorithms and applications,” Springer company, 2005.

[10] K. Frazer, “Real-time Operating System Scheduling Algorithms,” , 1997.

[11] W. A. Halang and A. D. Stoyenko, “Real Time Computing,” NATO ASI Series, Series F: Computer and Systems Sciences, Volume 127, Springer-Verlag company, 1994.

[12] J. A. Stankovic and K. Ramamritham, , “Tutorial Hard Real-Time Systems,” IEEE Computer Society Press, 1988.

[13] M. Garey, D. Johnson, “Complexity Results for Multiprocessor Scheduling under Resource Constraints,” SICOMP, Volume 4, Number 4, pp. 397-411, 1975.

[14] A. K. Mok, “Fundamental Design Problems of Distributed Systems for the Hard Real-Time Environment,” Ph. D. thesis. Department of Electronic Engineering and Computer Sciences, Mass. Inst. Technol., Cambridge MA, May, 1983.

[15] J. Y.-T. Leung and J. Whitehead, “On the complexity of fixed priority scheduling of periodic real-time tasks,” Performance Evaluation, Volume 2, pp. 237-250, 1982.

[16] S. Baruah, N. Cohen, G. Plaxton, and D. Varvel, “Proportionate progress: A notion of fairness in resource allocation,” Algorithmica , Volume 15, Number 6, pp. 600-625, June, 1996.

[17] C. A. Phillips, C. Stein, E. Torng, and J. Wein, “Optimal time-critical scheduling via resource augmentation,” In Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing, pp. 140-149, El Paso, Texas, 4-6 May, 1997.

[18] M. Moir and S. Ramamurthy, “Pfair scheduling of fixed and migrating tasks on multiple resources,” In Proceedings of the Real-Time Systems Symposium, IEEE Computer Society Press, Phoenix, AZ, December, 1999.

[19] J. Anderson and A. Srinivasan, "Early release fair scheduling," In Proceedings of the EuroMicro Conference on Real-Time Systems, IEEE Computer Society Press, pp. 35-43, Stockholm, Sweden, June 2000.

[20] H. Aydin, P. Mejia-Alvarez, R. Melhem, and D. Mosse, "Optimal reward-based scheduling of periodic real-time tasks," In Proceedings of the Real-Time Systems Symposium, IEEE Computer Society Press, Phoenix, AZ, December, 1999.

[21] S. Funk, J. Goossens, and S. Baruah, "On-line Scheduling on Uniform Multiprocessors," , 22nd IEEE Real-Time Systems Symposium (RTSS'01), pp. 183-192, London, England, December, 2001.

[22] J. M. Bans, A. Arenas, and J. Labarta, "Efficient Scheme to Allocate Soft-Aperiodic Tasks in Multiprocessor Hard Real-Time Systems," PDPTA 2002, pp. 809-815.

[23] Z. Xiangbin and T. Shiliang, "An improved dynamic scheduling algorithm for multiprocessor real-time systems," PDCAT'2003. In Proceedings of the Fourth International Conference on Publication, pp. 710- 714, 27-29 August, 2003.

[24] S. Lauzac and R. Melhem, "An Improved Rate-Monotonic Admission Control and Its Applications," IEEE Transactions on Computers, Volume 52, Number 3, pp. 337-350, March, 2003.

[25] P. Holman and J. H. Anderson, "Using Supertasks to Improve Processor Utilization in Multiprocessor Real-Time Systems," 15th Euromicro Conference on Real-Time Systems (ECRTS'03), Porto, Portugal, 2-4 July, 2003.

[26] D. A. El-Kebbe, "Real-Time Hybrid Task Scheduling Upon Multiprocessor Production Stages," International Parallel and Distributed Processing Symposium (IPDPS'03), Nice, France, 22-26 April, 2003.

[27] B. Andersson and J. Jonsson, "The Utilization Bounds of Partitioned and Pfair Static-Priority Scheduling on Multiprocessors are 50 percent," 15th Euromicro Conference on Real-Time Systems (ECRTS'03), Porto, Portugal, July 02-04, 2003.

[28] C. M. Krishna and K. G. Shin, "Real-Time Systems," MIT Press and McGraw-Hill Company, 1997.

[29] J. Carpenter, S. Funk, P. Holman, A. Srinivasan, J. Anderson, and S. Baruah, "A Categorization of Real-time Multiprocessor Scheduling Problems and Algorithms," Handbook of Scheduling: Algorithms, Models, and Performance Analysis, Edited by J. Y. Leung, Published by CRC Press, Boca Raton, FL, USA, 2004.

[30] B. Sprunt, J. Lehoczky, and L. Sha, "Exploiting Unused Periodic Time For Aperiodic Service Using the Extended Priority Exchange Algorithm," In Proceedings of the 9th Real-Time Systems Symposium, pp. 251-258. IEEE, Huntsville, AL, December, 1988.

[31] J. A. Stankovic, M. Spuri, K. Ramamritham, and G. C. Buttazzo, "Deadline Scheduling for Real-Time Systems, EDF and related algorithms," Kluwer Academia Publishers, 1998.

[32] B. Sprunt, "Aperiodic Task Scheduling for Real-Time Systems," Ph.D. Thesis, Department of Electrical and Computer Engineering Carnegie Mellon University, August, 1990.

[33] M. Spuri and G. C. Buttazzo, "Efficient Aperiodic Service under Earliest Deadline Scheduling," In Proceedings IEEE Real-Time Systems Symposium, pp. 2-11, San Juan, Puerto Rice, 7-9 December, 1994.

[34] M. Spuri and G. Buttazzo, "Scheduling Aperiodic Tasks in Dynamic Priority Systems," The Journal of Real-Time Systems.

[34] Osman S. Unsal and Israel Koren. System-level power-aware design techniques in real-time systems. Proceedings of the IEEE, 91(7):1055{1069, 2003.

[35] Wu chun Feng, Michael S. Warren, and Eric Weigle. The bladed beowulf : A coste _ective alternative to traditional beowulf. In IEEE International Conference on Cluster Computing (CLUSTER 2002), 23-26, Chicago, IL, USA, 2002.

[36] Vivek Tiwari, Deo Singh, Suresh Rajgopal, Gaurav Mehta, Rakesh Patel, and Franklin Baez. Reducing power in high-performance microprocessors. In DAC '98: Proceedings of the 35th annual conference on Design automation, pages 732{737, New York, NY, USA, 1998. ACM Press.

[37] K. Nowka, G. Carpenter, and B. Brock. The design and application of the powerpc 405LP energy-e_cient system on chip. IBM Journal of Research and Development, 47(5/6), September/November 2003.

[38] Kanishka Lahiri, Sujit Dey, Debashis Panigrahi, and Anand Raghunathan. Battery-driven system design: A new frontier in low power design. In ASPDAC '02: Proceedings of the 2002 conference on Asia South Paci_c design automation/ VLSI Design, page 261, Washington, DC, USA, 2002. IEEE Computer Society.

[39] P. Pillai, and K. G. Shin, Real-Time Dynamic Voltage Scaling for Low-Power Embedded Operating Systems, In Proc. of ACM Symp. On Operating Systems Principles, pages 89-102, 2001.

[40] D. Shin and J. Kim:, Dynamic voltage scaling of mixed task sets in priority-driven systems. IEEE Transaction on CAD of Integrated Circuits and Systems 25(3): 438-453 (2006)

[41] T. Ishihara and H. Yasuura, Voltage Scheduling Problem for Dynamically Variable Voltage Processors, In Proceedings of International Symposium On Low Power Electronics and Design, 1998, pp. 197-202.

[42] Nasro Min-Allah, Asad-Raza Kazmi2, Ishtiaq Al3, Xing Jian-Sheng, Wang Yong-Ji “Minimizing Response Time Implication in DVS Scheduling for Low Power Embedded Systems”

Comparison Between MADM Algorithms for Vertical Handoff Decision

Huma Ayub Vine[1]

Abstract

Composite communication platform is the basic requirement of today's Mobile Terminals. The need for this communication platform is to provide seamless roaming to the user without much distortion in their services. So far many different approaches have been adopted, for raking these switching decisions and among all of these approaches, MADM (Multiple Attribute Decision Making) is considered as one of the approach for solving such types of vertical handover problems. The types of few MADM algorithms are AHP, TOPSIS, MEW and SAW. In this paper comparisons between these MADM algorithms have been done. Results of each algorithm and their response against the decision ranking have been analyzed. It is observed that TOPSIS is suffering from non stability behaviour; MEW shows penalizing behaviour towards poor attributes, where as AHP and SAW shows less risk in its decision ranking with minimum standard error mean and statistical variance.

Keywords: Vertical Hand off, Composite Communication Platform, Ranking, Sensor devices, MADM

Introduction

Now a day, researchers are focusing to develop a composite communication platform [1] for the mobility management. The aim behind this composition is to reduce the communication gap between various heterogeneous wireless network environments. From architectural point of a view, a vast set of these wireless technologies such as Code Division Multiple Access (CDMA2000), Satellite Network, Sensor Networks, and Global System for Mobile Communications (GSM), Wireless Local Area Network (WLAN), Universal Mobile Telecommunication System (UMTS), Mobile Ad Hoc Network (MANET), and Home RF network and Blue tooth basically requires a platform which seems to provide seamless roaming to the user without any distortion. The demand for this seamless roaming or handoff becomes more important, when this roaming is vertical not horizontal. Vertical handoff is quite different from horizontal handoff. As in horizontal handoff the mobile user roams from one base station to another within the same network. Where as in vertical handoff the mobile user roams between two different network technologies. Many issues are involved in this vertical handoff, which indirectly affects the decision factors of mobile user. According to the requirement of the user preferences as well as the available network services, parameters such as bandwidth, cost, application type support, bit error rate, coverage area and other QoS attributes, forcefully compel the wireless devices, to take a dynamic decision of net-work selection. Various vertical handover decision algorithms have been proposed and discussed in literature [4], [5], [6], [7] which are using either fuzzy logic or based on policies [9], [10], [11], [12].

In our previous paper [23], we have proposed an idea of Intelligent Intermediate Robust Gateway (IIRG) for handling handover decisions of remote sink sensor applications problems. The Gateway module discussed in that paper is based on fuzzy logic and Analytical Hierarchal Process (AHP), which use fuzzy comparison ratio based criteria for the handling weights of different criteria's by using pair wise judgment. The framework of IIRG is based on three basic modules called Network Monitor, Data Mine, Sensor Application Based Module and FANS [23] as shown in Fig 1. For making analysis of this framework, we extend our work by conducting a comparative study of different MADM based vertical handoff algorithms. Although there is much work done in this vertical handoff but there is lack of comparison between their performance works. The main contribution of our work in this paper is to make comparison among different decision making techniques and check their pros and cons. The algorithms which are used in our comparison work are Technique for Order Preference by Similarity to ideal Solution (TOPSIS) [13], Multiplicative Exponent Weighting (MEW) [13], Analytical Hierarchy Process (AHP) [3], and Simple Additive Weighting (SAW) [13]. A different type of performance attributes such ranking, non stability behavior, Standard Error of Mean (SEM), Minimum Variance Unbiased Estimator (MVUE) and their resultant is discussed in this paper. We have used Matlab for the simulation of our work. Results show that

[1] Departmet of Software Engineering, University of Engineering & Technology, Taxila

more that 80-85% of times the decision made by SAW, MEW, TOPSIS and AHP are same, whereas 20% of time they made different ranking decisions. The reason for selecting different ranking criteria by different algorithms depends on the abnormality behavior of TOPSIS, penalizing behavior of MEW, minimum response value of SEM and MVUE for AHP as well as SAW. For checking the performance of each algorithm, study is made on the bases of quantitative as well as qualitative data. The resultant of this approach shows that TOPSIS is a non stable algorithm in its decision ranking, MEW is suffering from penalizing behavior in case of poor attributes, whereas AHP and SAW show much stable and better decision as compared to other two.

In section II a brief overview of different types of MADM techniques are discussed. Whereas in Section III shows a comparative study of these techniques. Finally in section IV conclusion is drawn from their comparative results for the future perspective.

Decision Making Techniques

As discussed in section I, that Vertical Handoff consists of three major phases described as system discovery, handoff decision, and handoff execution [2]. At system discovery phase the availability of different networks and their provided set of parameters are studied. At handoff decision phase the final selection of network take place from among the list of candidate networks. The final decision is made after considering both user preferences as well as available network parameters. User preferences give weightage to the QoS parameters, which provide much help in the selection of network from the list. Where as in the last phase of handoff execution, the step is taken to change the current network/cell/technology into the newly selected network/cell/technology. Before going in the details of our work, a brief over view of four different MADM techniques are discussed.

- **Analytical Hierarchal Process (AHP)**

  It was developed by Prof T. Saaty (1980) at Wharton School of Business [3]. This process decomposes a complex decision problem into a hierarchical structure. AHP hierarchal structure has at least three major steps.

1. At the top of the hierarchy, problem Statement and objectives are placed.
2. In the middle of the hierarchy a list of criteria's are mentioned that are required to the alternatives.
3. Where as at the lowest levels the sets of alternatives is defined as shown in Fig 1.
4. A pair wise comparison matrix is established which shows the relationship between upper level elements with respect to the level immediately below it.
5. Two questions are answered during this comparison procedure i.e.
   1. Which criteria are more important?
   2. How much extend it is important as compared to other criteria

Table 1: Comparison Saaty, 1980)
Comparison Saaty, 1980) [3]

INTENSITY OF IMPORTANCE	DEFINATION
1	Equal Importance
2	Equal to Moderately Importance
3	Moderate Importance
4	Moderate to strong Importance
5	Strong Importance
6	Strong to very strong Importance
7	Very strong Importance
8	Very to Extremely Strong Importance
9	Extreme Importance

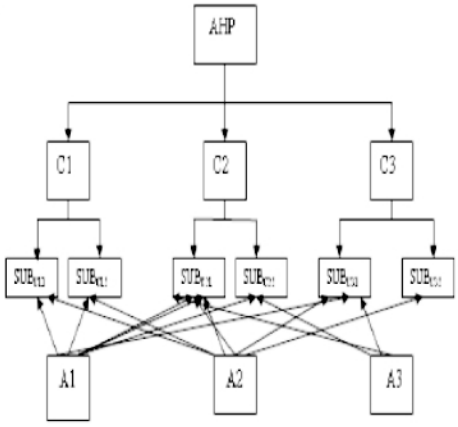


Figure 2. Hierarchy Decision Making Process

Buckley's in 1985 [15], [16] suggested fuzzy AHP by allowing fuzzy numbers for pair wise comparison and calculate fuzzy weights and fuzzy performance/results. Linguistic variables are used to express fuzzy numbers. These linguistic variables use linguistic terms like Low, Medium, and High to classify different perception of the related subject. For example the bandwidth can be low medium or high and similarly the distance covered by the network can be Low, Medium or High. Saaty uses scale table as shown in Table I for the judgment and comparison of criteria's.

6 At the final step of Saaty AHP method, eign vector of comparison matrix is calculated, which is used to find the relative weights among the criteria's (sub criteria's and one level above of the hierarchal system). Saaty [3] gives different methods for the calculation of the eign vectors. These methods are summarized below:-

Normalization of Row Average

$$\mathbf{w_i} = \sum_{j=1}^{n}(b_{ij}) / \sum_{i,j=1}^{n}(b_{ij}) \quad \textbf{(1)} \qquad w_i = \prod_{j=1}^{n}(a_{ij})^{1/n} / \sum_{k=1}^{n}\prod_{j=1}^{n}(a_{kj})^{1/n} \quad \textbf{(2)}$$

where in eq 1 nominator part is showing sum of each row and denominator is showing sum of over all rows. Hence there division gives a normalized output.

Geometric mean of Rows

It takes nth root of multiplication of all elements in each row and divides them with the sum of product of all elements in each and every row.

Average of Normalized Column

Converts fraction pair wise comparisons to decimal equivalent and calculate normalized valueby dividing each element by its column total. Row wise total is then taken for this normalized matrix. Average normalized column is obtained by dividing row sum by the number of elements in the row.

$$w_i = 1/n(\sum_{j=1}^{n}(a_{ij}) / \sum_{k=1}^{n} a_{kj}) \quad \textbf{(3)} \qquad CI = (\lambda_m axn)/n - 1 \quad \textbf{(4)}$$

Consistency of Matrix

Consistency checking is considered as one of the major key point for AHP decision analysis. It is used to verify the reliability of our judgment. The judgment is considered to be consistent if the maximum eign value _max of the reciprocal matrix is equal to the order of the matrix. Whereas it is considered as in consistent if the value of maximum eign value is greater than the order of the matrix. For measuring the consistency of the matrix we use the Consistency Index (CI) as shown in the

Eq 3. CR =CI/RI (5)

For consistent pair wise matrix CI would be zero. Consistency of judgment can be further calculated by using the Consistency Ratio i. e. Random Index (RI) for different order of matrix are calculated by Saaty [3] and shown in Table II. If the calculated CR is less than 0.10, it means less inconsistency exists in our assumed matrix, where as if CR¿0.10 than calculated matrix is not consistent and whole matrix should be revised or re-examined.

Table 2: Random Index Table (Saaty, 1980)

MATRIX SIZE	RANDOM INDEX	MATRIX SIZE	RANDON INDEX
1	0	9	1.43
2	0	10	1.49
3	0.58	11	1.51
4	0.9	12	1.48
5	1.12	13	1.56
6	1.24	14	1.57
7	1.32	15	1.59
8	1.41		

Technique for Order Preference by Similarity to Ideal Solution (TOPSIS)

In this method two artificial alternatives are hypothesized: i.e.

1) Ideal alternative (Positive alternative)

2) Negative alternative

The basic rule behind these hypotheses is that the chosen alternative should have shortest distance form positive ideal solution and longest distance from negative ideal solution [17].

Best/Ideal alternative: considered to be one which shows the best value for all attributes. Negative ideal alternative: the one which has the worst attribute value. TOPSIS makes the decision of alternative that is very near to the best result and very far from Pessimistic option alternative. In this method the problem is classified on the bases of m alternatives (options) and n attributes/criteria. Each option or alternative is assigned a score against each criterion. For the explanation of the mathematical process if xij is the score of option i with respect to criterion j and matrix X = xij is a m x n matrix then it is consider that J be the set of benefit attributes or criteria (more is better and J' be the set of negative attributes or criteria (less is better). The following steps are followed for the calculation of this method.

Step 1: Make a decision matrix using normalized values. Through this step, different measuring units of attributes are transformed into uniform unit attributes, which allows comparisons across criteria. The normalize scores or data are as given in Eq 6:

$$r_{ij} = x_{ij}/\sqrt{\sum (x_{ij})^2} \qquad (6) \qquad\qquad v_{ij} = w_j * r_{ij} \qquad (7)$$

Step 2: Make a decision matrix having weighted normalized values. Assume we have a set of weights for each criteria wj for j = 1…n. Multiplications have to be done between columns of the normalized decision matrix and its corresponding weights. An element of the new matrix is given as in Eq 7: for i = 1, ..., m; j = 1,..., n.

Step 3: Determine the ideal and negative ideal solutions. The positive or ideal solution can be obtained as in Eq 8

$$A* = v_1*, ..., v_n* \qquad (8) \qquad\qquad A' = v'_1, ..., v'_n \qquad (9)$$

$$\text{where } v_{j*} = max v_{ij} if j \in J; \text{ min } v_{ij} \text{ if } j \in J' \qquad \text{where } v' = \text{min } v_{ij} if j \in J; \text{ max } v_{ij} if j \in J'$$

Whereas the negative ideal solution is given as in Eq 9

Step 4: Calculate the separation measures for each alternative. The separation from theideal alternative is given in Eq 10

$$S_i^* = [\sum (v_j^* - v_{ij})^2]^{1/2} \qquad (10) \qquad\qquad S_i' = [\sum (v_j' - v_{ij})^2]^{1/2} \qquad (11)$$

where i = 1....., m And similarly, the separation from the negative ideal alternative is as given in Eq 11: where i = 1,...., m

Step 5: Calculate the relative closeness to the ideal solution Ci*. Select the Alternative with Ci*closest to 1.

$$C_i^* = S_i'/(S_i^* + S_i') \qquad (12) \qquad\qquad S_i = \sum_{j=1}^{M} (w_j * r_{ij}) \qquad (13)$$

where $0 < C_i < 1$

Simple Additive Weighting (SAW)

Here the normalized value of the criteria for each alternative is multiplied with the importance of the criteria i.e. weights assigned to the attributes/criteria. After that the total score against each alternative is calculated by summing these products over all these attributes. At the end the alternative having highest score is selected finally. Equation 13 gives the mathematical way for calculating the sum these products over all of these attributes. Where i = 1... n, j=1,.....,m, Si represents the Overall score of the ith alternatine,rij is the normalized rating of the ith alternative of the jth criterion,xij is the value of the jth criterion of ith alternation, wj is the weight of the jth Criterion, M is the number of criteria and N is the number alternative.

Multiplicative Exponent Weighting (MEW)

MEW is another way of solving multi attributes Problem. Here the attribute problem consists of a matrix having N number of alternative and M number of criteria's against them. The score for each network i can be calculated as

$$S_i = \prod_{j=1}^{m} (x_{ij})^{w_{ij}} \qquad (14)$$

Here x_{ij} is the element or value of jth attribute and wj is the weight assigned to each attribute. The value of wj will be considered as positive for benefit matrix x_i^{wij} j and its value will be negative for cost factor i.e. $xi-^{wij}$. The selected network is the best value of each matrix. The highest value in benefit matrix is considered as preferred one; where as the lower value in the cost matrix is selected as final option.

Comparison of Techniques

For making a comparison between these four handover methods i.e. AHP[3], TOPSIS[13], MEW[13]and SAW[13] we have consider an example of traffic, coming from some remotely placed sensor devices. According to the authors in [18] sensor devices provide support for many different types of applications such as Event, Continues, Hybrid and Query. The classification of these data types are done on the bases of data driven module. Each application type has a requirement of diverse QoS parameters. Event type applications of sensor devices always activate in the presence of some triggers, Query type applications responded when some question is generated from the sink, continues type response continually by sending data at some fixed rate, whereas hybrid applications are combination of all these three above applications. In the calculation of network selection for this application, different types of QoS parameters are required both from user as well as from network. The parameters which we used for the evaluation of methods are bandwidth, delay tolerance, security, application type, coverage area, cost and priority. Here the values of these parameters, either received from user or from network, are first converted into fuzzy variables by making use of Gaussian membership function [19]. The advantage of using these fuzzy variables is that one can easily express both quantitative and qualitative data. Later on these fuzzy values are normalized between [0, 1] for developing symmetry of measurement among the different parameters having different units. For the simplicity of our work, the weights for each criterion are calculated by using AHP Eigen value method based on geometric mean. Our simulation, which is carried out in MATLAB2.

Table 3: Event Based AHP Matrice

QoS	App_Pri	Critical	Delay	Bandwith	Coverage
App_Pri	1	1	4	4	7
Critical	1	1	4	4	7
Delay	1/4	1/4	1	1/5	3
Bandwidth	1/4	1/4	5	1	5
Coverage	1/7	1/7	1/3	1/5	1

Table 4: Continue Based AHP Matrice

QoS	App_Pri	Critical	Delay	Bandwith	Coverage
App_Pri	1	1	1/7	1/9	1/3
Critical	1	1	1/7	1/9	1/3
Delay	7	7	1	1/5	5
Bandwidth	9	9	5	1	7
Coverage	3	3	1/5	1/7	1

Table 5: Query Based AHP Matrice

QoS	App_Pri	Critical	Delay	Bandwith	Coverage
App_Pri	1	1/2	1/5	1/9	1/4
Critical	2	1	1/5	1/9	1/4
Delay	5	5	1	1/5	3
Bandwidth	9	9	5	1	7
Coverage	4	4	1/3	1/7	1

Table 6: Hybrid Based AHP Matrice

QoS	App_Pri	Critical	Delay	Bandwith	Coverage
App_Pri	1	1/2	1/7	1/9	1/4
Critical	2	1	1/7	1/9	1/3
Delay	7	7	1	1/5	2
Bandwidth	9	9	5	1	2
Coverage	8	3	1/2	1/2	1

consider performance parameters of five different networks i.e. 802.11a, 802.11b, Satellite, GSM, UMTS as well as User Provided QoS parameters. The parameters provided by each network are generated randomly. Selection of the final network depends on the values of the above mentioned attributes. The AHP Matrices regarding each traffic class is shown in Table III, IV, V and VI and resulting weights calculated against each traffic class are shown in Table VII.

Ranking Approach

For the simulation of our work, the calculations are done against Event based Application. First the parameters provided by four different networks against each QoS criteria's are converted into fuzzy set, as shown in Table VIII. These fuzzed parameters are then multiplied by the Event based calculated weights provided in Table VII. The final decision is based on highest ranking factor. The ranking distribution for four different methods i.e. AHP SAW, MEW and TOPSIS are shown Table IX. The result shows that Network 4 is the final selection due to highest ranking and the ranking order of AHP, SAW andTOPSIS are same where as in case of MEW, there is a slight variation in decision factor of Network 1 and Network 2. This variation factor in ranking is due to penalizing behavior of MEW algorithm. MEW always give lower rank to the alternative which has poor attributes as compared to other one. In this scenario, Network 1 has poor attributes as compared to Network 2, which can be verified by taking the average and variance of attributes as shown in Table VIII. Although this decision of MEW seems to be good but MEW has considered only the over all average of parameters for its decision ranking and did not considered the weights assigned to these parameters. Therefore Network1 which is providing high bandwidth and less delay has not been assigned high rank. This means that MEW lacks behind in assigning a good ranks.

Stability Approach

In this simulation, we concentrate on the non stability ranking problems between algorithms. Focusing on the ranking done by the algorithms in Table X, we start removing the lowest ranking decision alternative from the table. Results shows that AHP, SAW and MEW remain stable in their ranking decision; where as the ranking order of TOPSIS has been changed. This abnormal behavior of TOPSIS is due to the large variation in its assigned values of ranks. Table X and Table XI, is showing abnormal behavior of TOPSIS before and after removal of lowest Rank alternative

Table 7: Consistncy Ratio (CRatio) and Weights W.R.T Application Type

Sensor App	App_Pri	Critical	Delay	Bandwith	Coverage	CRatio
EVENT	0.14	0.14	0.03	0.06	0.15	0.09
CONTINUE	0.0024	0.0024	0.015	0.034	0.0053	0.079
QUERY	0.006	0.008	0.033	0.094	0.018	0.082
HYBRID	0.0071	0.011	0.052	0.109	0.041	0.09

Table 8: Network Provided Parameters

	Sensor App	App_Pri	Critical	Delay	Bandwith	Coverage	Mean	Variance
802.11a	0.787	0.656	0.00004	0.1312	0.4949	0.0383	0.35122	0.1146
802.11b	0.2274	0.3279	0.8995	0.3137	0.2517	0.433	0.40887	0.06291
Satellite	0.8424	0.1845	0.5082	0.4522	0.3256	0.3801	0.44883	0.04964
GSM	0.8865	0.7613	0.8838	0.4574	0.7992	0.134	0.65372	0.08968
UMTS	0.0653	0.3751	0.3735	0.484	0.9695	0.3421	0.43492	0.08815

Table 9: Network Ranking

	AHP	SAW	MEW	TOPSIS
802.11a	0.36250(3)	0.36250(3)	0.4489(4)	0.79960(3)
802.11b	0.17694(4)	0.17694(4)	0.0.49208(3)	0.24901(4)
Satellite	0.38814(2)	0.38814(2)	0.77376(2)	0.84174(2)
GSM	0.47431(1)	0.47431(1)	0.90538(1)	0.98577(1)
UMTS	0.10452(5)	0.10452(5)	0.30011(5)	0.09807(5)

Table 10: Ranking Analysis for Topsis: Case I

	AHP	SAW	MEW	TOPSIS
802.11a	0.305290(5)	0.305290(5)	0.667105(5)	0.1314(5)
802.11b	0.457316(2)	0.457316(2)	0.877490(1)	0.6090(1)
Satellite	0.383(3)	0.383(3)	0.735(3)	0.4702(3)
GSM	0.4577(1)	0.4577(1)	0.8297(2)	0.608(2)
UMTS	0.3818(4)	0.3818(4)	0.6984(4)	0.378(4)

Table 11: Ranking Analysis for Topsis: Case II (Abnormality)

	AHP	SAW	MEW	TOPSIS
802.11a	-	-	-	-
802.11b	0.46(2)	0.46(2)	0.88(1)	0.83(2)
Satellite	0.38(3)	0.38(3)	0.73(3)	0.74(3)
GSM	0.46(1)	0.46(1)	0.83(2)	0.83(1)
UMTS	0.38(4)	0.382(4)	0.698(4)	0.73(4)

Table 12: Standard Error of Mean

N	AHP	SAW	MEW	TOPSIS
5	0.0063	0.0062	0.0024	0.0021
10	0.0068	0.0068	0.0118	0.0089
20	0.0036	0.0036	0.0060	0.0059
50	0.0021	0.0021	0.0038	0.0032
100	0.0014	0.0014	0.0027	0.0023
250	0.0009	0.0009	0.0018	0.0014
400	0.0007	0.0007	0.0014	0.0011
700	0.0006	0.0006	0.0010	0.0009
1000	0.0005	0.0005	0.0009	0.0007

Table 13: Variation: For Comparison of MVUE

AHP	SAW	MEW	TOPSIS
0.03	0.03	0.10	0.16
0.01	0.01	0.02	0.07
0.02	0.02	0.02	0.11
0.01	0.01	0.03	0.10
0.01	0.01	0.01	0.09
0.02	0.02	0.06	0.09
0.02	0.02	0.07	0.11
0.01	0.01	0.02	0.10
0.01	0.01	0.01	0.07
0.01	0.01	0.04	0.11

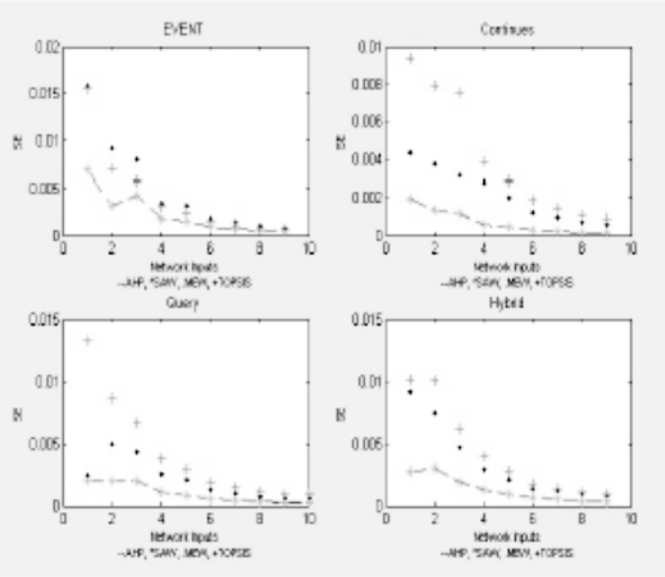


Figure 3: Standard Error of Mean for AHP, SAW, MEW and TOPSIS

Estimator Analysis

Estimator is one of the Statistical branches where estimated value is calculated through given empirical data or measured data. The purpose of estimator is to find the desired results which are embedded as noisy signal in the available data. Although the aim is to find estimator that exhibits optimal behavior. However which is not possible [21]. Among the list of these estimators, we have used SED and MVUE for the analysis of our above algorithms.

Mean squared error

MSE of an estimator also called second moment of error [22] taking both the variance of the estimator and its bias in account. The difference between estimated and true value is actually the error, which comes due to randomness or malfunctioning of estimator. The MSE for unbiased estimator is a variance. On the other hand the under root of MSE is called the root mean squared error or RMSE and at some places also known as Standard Error (SEM). RMSD or standard error is considered as a good measure of accuracy.

For this the term standard error is often used to calculate the unknown values. The value of this standard error is very helpful in providing the indication of the amount uncertainty in the decision based on mean value. SEM is usually calculated as sample estimation of the population standard deviation divided by the square root of the sample size as shown in eq

Stability Approach

In this simulation, we concentrate on the non stability ranking problems between algorithms. Focusing on the ranking done by the algorithms in Table X, we start removing the lowest ranking decision alternative from the table. Results shows that AHP, SAW and MEW remain stable in their ranking decision; where as the ranking order of TOPSIS has been changed. This abnormal behavior of TOPSIS is due to the large variation in its assigned values of ranks. Table X and Table XI, is showing abnormal behavior of TOPSIS before and after removal of lowest Rank alternative

$$S.E = \frac{s}{\sqrt{n}} (15)$$

where s is the sample standard deviation and n is the size of the sample. While considering the case of our AHP, SAW, MEW and TOPSIS method, the analysis has been drawn against SEM. The results of SEM are basically showing the efficiency of a methodology in relation to its unbiased estimator. The minimum the value of SEM estimator, the more efficient will be the methodology. It is because according to statistical rule in SEM as the value of sample size increases the resultant should move in decreasing order. The lower the SEM estimator value, the lesser will be the error or deviation of the quantity from the true value. Table XII and Fig 1 is showing the SEM resultant against different sample sizes and the behavior of this result show that AHP and SAW have the least biased value as compared to other two algorithms. So therefore we can say that AHP and SAW is showing better efficiency as compared to other two algorithms.

Minimum-Variance Unbiased Estimator A Uniformly Minimum-Variance Unbiased Estimator or Minimum-Variance Unbiased Estimator (UMVU or MVUE) is an unbiased estimator that computes low value of variance as compared with any other unbiased estimator. For finding the MVUE usually the comparison is done in terms of ratio between unbiased estimator variances. This comparison ratio is basically the efficiency of estimator. The efficiency for this estimator is normally stated in a relative terms. If we state _1 as unbiased estimator of sample 1 and _2 as unbiased estimator of sample 2, then the ratio between their variance states that VAR (_1) and VAR (_2) represents the measure of relative efficiency of _1 with respect to _1. If VAR (_1) is less than VAR (_2) then _1 is considered as more efficient than VAR _2. For checking the efficiency factor of our algorithms i.e. AHP, SAW, MEW and TOPSIS, we again calculate their variances against each ranking decision. These results are shown in XIII.For MVUE purpose six different ratio factor related variance have been calculated. These six different pairs are

1. VAR AHP vs. VAR SAW
2. VAR AHP vs. VAR MEW
3. VAR AHP vs. VAR TOPSIS
4. VAR SAW vs. VAR MEW
5. VAR SAW vs. VAR TOPSIS
6. VAR MEW vs. VAR TOPSIS

Comparison between their variances ratio again shows that AHP and SAW both have minimum variance unbiased estimator as compared to MEW and TOPSIS. Where as this estimator has ranked MEW at the second position and TOPSIS at last or third position. Thus the non stability and inconsistency in TOPSIS is due to having largest variability ratio as compared to other algorithms. That is why any small change in it rank creates major changes in it decision stability and consistency.

Conclusions

In this paper the selection of network for sensor based applications has been done by making use of AHP, SAW, MEW and TOPSIS algorithms. Parameters included for decision making are application type, bandwidth, delay, coverage area, security and priority. A number of limitations have been identified during

the comparison study of these algorithms. For the simulation of this decision making process, Matlab is used. Results show that more that 80-85% of times the decision made by SAW, MEW, TOPSIS and AHP is same, whereas 20% times they made different ranking decisions. Comparison shows that TOPSIS is suffering from abnormality behavior of ranking, MEW is penalizing poor attributes and makes judgment on the mean and average value of attributes, where as AHP and SAW as compared to other two algorithms shows more stable, less risk proven and penalizing judgment behavior.

References

[1] K. Murray, D. Pesch, State of the Art: Admission Controland Mobility Management in Heterogeneous Wireless Networks," M-Zones State of the Art Paper, soa paper 05/03, Ireland, May, 2003.

[2] E. Stevens-Navarro and V. W. S. Wong, Comparison between Vertical Handoff Decision Algorithms for Heterogeneous Wireless Networks," in Proc. of IEEE Vehicular Technology Conference (VTC- Spring'06), Melbourne, Australia, May 2006.

[3] T. L. Saaty, Book: The Analytical Hierarchy Process, New York: McGraw-Hill, 1980.

[4] P. M. L. Chan, An intelligent Handover Strategy for a Multi- Segment Broadband Network", PIMRC 2001, Sep 2001, San Diego, California.

[5] P. Chan, et. al., Mobility Management Incorporating Fuzzy Logic for a Heterogeneous IP Environment, IEEE Communication Magazine, Dec.2001.

[6] A. Majlesi and B. H. Khalaj, An Adaptive Fuzzy Logic Based Handoff Algorithm for Interworking between WLANs and Mobile Networks, IEEE INT'L Symp. on Personal, Indoor and Mobile Radio Comm, vol 5, 2002, pp.2446-2451.

[7] W. Zhang, Handover Decision Using Fuzzy MADM in Heterogeneous Networks, in Proc. Of IEEE WCNC'04, (Atlanta, GA), March 2004.

[8] Q. Song and A. Jamalipour, Network Selection Mechanism for Next Gen. Networks, in Proc. of IEEE ICC'05,(Seoul, Korea), May 2005."

[9] H.J. Wang, R. H. Katz, and J. Giese, Policy-Enabled Handoffs across Heterogeneous Wireless Networks, Proc. of ACM WMCSA, 1999.

[10] Qingyang Song, Abbas Jamalipour, A Time-Adaptive Vertical Handoff Decision Scheme in Wireless Overlay Networks, IEEE 17th International Symposium on Personal, Indoor and Mobile Radio Communications, 2006.

[11] J. McNair, F. Zhu, Vertical Handoffs in Fourth-Generation Multinetwork environments, IEEE Wireless Communications,pp.8-15, June 2004.

[12] J. Ylitalo, T. Jokikyyny, T. Kauppinen, A. J. Tuominen,J. Laine, Dynamic Network Interface Selection in Multihomed Mobile Hosts", 36th Hawaii International Conference on System Sciences, Hawaii, USA, January 2003.

[13] W. Zhang, Handover Decision Using Fuzzy MADM in Heterogeneous Networks, in Proc. IEEE WCNC'04, Atlanta, GA, March 2004.

[14] Q. Song and A. Jamalipour," "A Network Selection Mechanism for Next Generation Networks, in Proc IEEE ICC'05, Seoul, Korea, May, 2005.

[15] Buckley, Fuzzy hierarchical analysis. Fuzzy Sets Systems, v17 i1.233-247,1985.

[16] Buckley, Ranking alternatives using fuzzy numbers. Fuzzy Sets Systems, v15 i1. 21-31,1985.

[17] Yoon K. P., Hwang C. L., Multiple Attribute Decision Making, An Introduction, Sage University Papers (Series: Quantitative Applications in the Social Sciences), 1995.

[18] S. Tilak, N. Abu-Ghazaleh, and W. Hein Zelman, A Taxonomy of Wireless Micro-Sensor Network Models, Mobile Computing and Communications Review (MC2R) of ACM, vol. 6, No.2, Apr 2002

[19] The Math Works accelerating the pace of engineering and science, http://www. mathworks. com, Nov 2008.

[20] Interpreting results : Variance and coefficient of variation, http://www.graphpad.com /help/ Prism5 prism5help.html/coefficient of variation(cv)htm.

[21] Estimation theory From Wikipedia, the free encyclopedia, http://en.wikipedia.org/wiki/Estimation theory.

[22] Estimation theory From Wikipedia Mean squarederror, From Wikipedia, the free encyclopedia,' http://en.wikipedia.org/wiki/Mean squared error.

[23] H. Qayyum, S. Sohail,G. A. Shah and A. Khanum. A Fuzzy Decision Maker for Wireless Sensor Network in Ubiquitous Network environment,"ieeecomputersociety.org/10.1109/WAINA.2009.85.

Effect of in plane Shearing Stiffness of Infill Walls on Response of Moment Resisting Frames of High Rise Buildings under Seismic Loads

Prof. Dr Saeed Ahmad[1] and Talha Afzal[2]

Abstract

The majority of human occupation is catered by residential and office buildings for our living and working purposes. Because of available space restrictions, there is general trend of high rise construction in urban areas which are prone to earthquake forces. The major load resisting structural elements consists of multilevel beam column frames with or without shear walls.

Non structural infill walls constructed of burnt brick masonry are used in almost all concrete frame buildings in Pakistan to serve architectural functions. (A country where seismicity is one of the prime factor to be considered in design) Such masonry infill walls are considered as non-structural elements as these are constructed after completion of concrete frames. Although they are designed to perform architectural functions, masonry infill walls do resist lateral forces with substantial structural action. In addition to this, infill walls have a considerable strength and stiffness and they have significant effect on the seismic response of the structural system. Most of the past researchers concluded that in filled masonry wall frames have the capability to cater greater design forces as compared to frames without it, Since they contribute to shear stiffness of the structure (the property having prime importance in seismic response of a structure) like shear walls.

Keywords: Masonry infill, Shear walls, Lateral stiffness, Seismic response

Introduction

The effects of non structural infill walls are not included generally during analysis and design of reinforced concrete structures. However, these infill walls have considerable effect on the structural response as studied by many researchers all over the globe. The brick masonry is used as partitions in most of reinforced concrete construction in Pakistan because of its low cost and locally available skilled labor. Therefore, it is necessary to study the effects of these nonstructural infill walls to response parameters of these structures.

The addition of these nonstructural walls increase the lateral stiffness of the frame structure thus reducing its time period of vibration and shifting the period in the elastic response spectra in such direction that yields higher base and storey shears. Moreover, the increased stiffness of the structure may increase the column shears which creates the plastic hinges at the top of columns that are in contact with infill walls since the structure is designed for a ductile response to design level earthquake.

Furthermore, potentially negative effects also occur such as torsional effects induced by in plan-irregularities, soft-storey effects induced by stiffness irregularities and short-column effects.

In recent past, many researchers did the study of effect of stiffness of infill nonstructural walls on the behavior of the moment resisting frame by using different analysis techniques. Almost all the engineers performed a comparative study between the structures having modeled infill walls and another without it.

In 2007, international journal of science and technology published a paper presented by" Kasım, Armagan, Korkmaz, Fuat Demir and Mustafa Sivri" students of " Suleyman Demirel University, Civil Engineering Department, TURKIYE. They perform the nonlinear analysis on different models of various configurations of infill walls as shown. A three (03) story reinforced frame was considered with different patterns of infill walls. They concluded that the presence of infill masonry walls significantly and positively alters the seismic performance of the structure. However its irregular vertical distribution causes some negative effects including soft story phenomenon.

In 2003, another research work had been carried out by CVR MURTY and SUDHIR K JAIN. The purpose of their research work is to highlight the beneficial effects of non structural masonry walls on the bare

[1] Civil Engg. Department UET Taxila Pakistan, [2] Structural Engineer, Nespak Pakistan.

So in this study the response of two essentially same structures, with and without consideration of stiffness of in filled walls were evaluated and compared on the basis of different structural parameters listed next.

Research Objectives

In this comparative evaluation study, the seismic performance of two multistory reinforced concrete building, one with modeled in filled walls to account for its stiffness and other without it, shall be investigated by Elastic Response Spectrum Analysis using UBC-97.

The objectives of the study are summarized in following:

1. To study effect of in filled walls on the performance of the high rise ductile moment resisting frames under seismic loads and to get a quantitative idea of this effect based on different structural parameters like but not limited to:
   a) Time Period
   b) Base Shear
   c) Story Drifts
   d) Relative Story Displacements
   e) Support Reactions
   f) Member End Forces (Shears, Moments, Torsion etc)

   by comparing it with another same structure but without modeling the infill walls.
2. To understand the need to account the stiffness of infill walls.
3. To ensure and understand the need of proper designing of in filled Non structural walls against lateral seismic forces in order to ensure its serviceable performance level without extensive cracking so that its stiffness contribution should remains there.

Case Study Buildings

To study the effect of in plane stiffness of infill walls as prescribed earlier in methodology of research, we will analyze a high rise moment resisting frame building, with and without modeling of infill walls. So for this purpose we model a sample structure having 90ft x 90ft plan dimensions dividing in 6 bays each of 15ft in both principles direction. The total height of building is taken as 109ft having 10 numbers of stories. The height of plinth level is taken as 10ft and a typical story height of 11ft is taken above plinth level comprising of 9 stories.

In first model as shown, loads of masonry in fill walls 9" thick of full story height are applied as member load on all external beams and only on those internal beams situated at grid 'D' and '4'. In second model, infill walls are modeled at prescribed locations using material properties of masonry walls as shown. The brief description of these models and graphical views are given below.

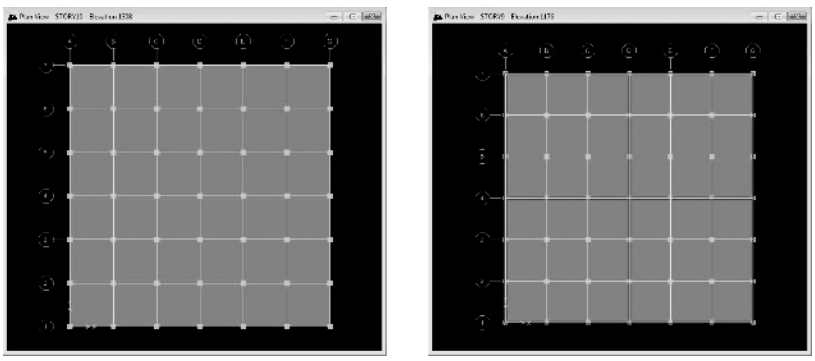


Figure 1: Plan views

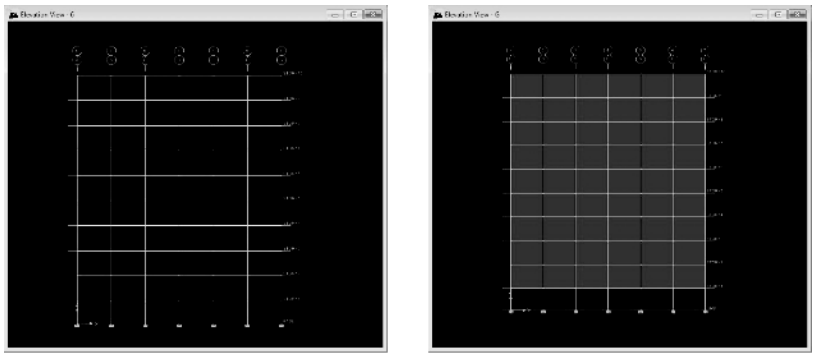


Figure 2: Elevation views

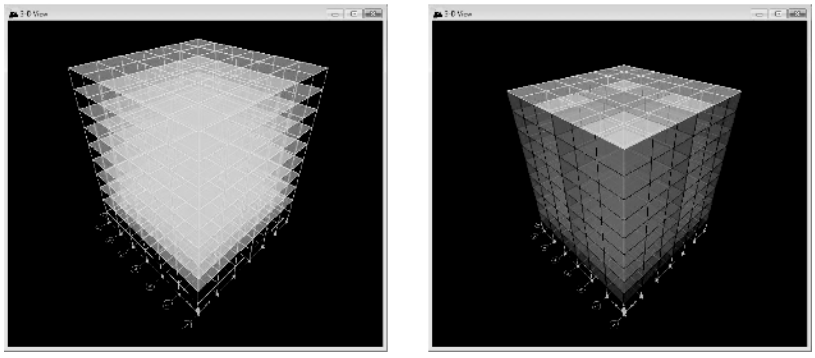


Figure 3: 3D views

Method of Analysis

Dynamic analysis is performed using Response Spectrum method of analysis by using standard UBC Response Spectra whose peak values corresponds to Ca and Cv values used for Zone 2b as per UBC. Eigen vectors analysis type is used to generate different possible number of modes.

In modal analysis, SRSS (square root of sum of squares) technique is used for modal combinations in which 8 no of modes are taken for combination since mass participation appears to be 99% for 8th mode in both principle directions.

SRSS (square root of sum of squares) technique is also used for Directional combinations.

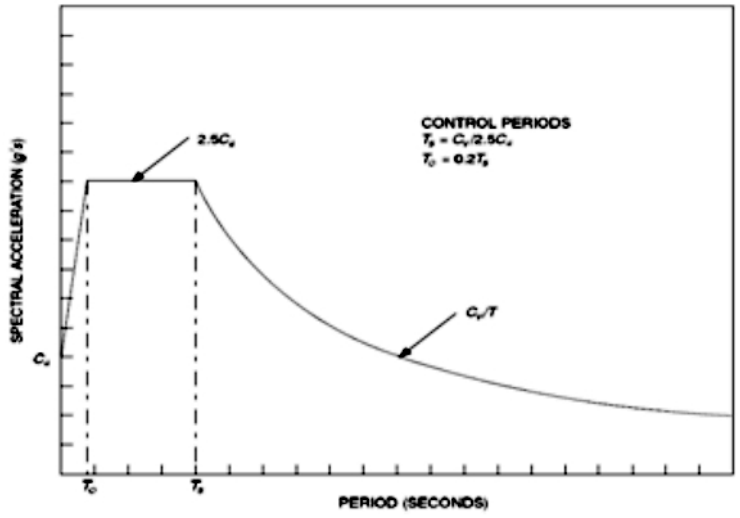


Figure 4: UBC response spectra

Results and Discussion

After performing the dynamic analysis of the two structures, their behavior will be analyzed and compared in terms of the following parameters.

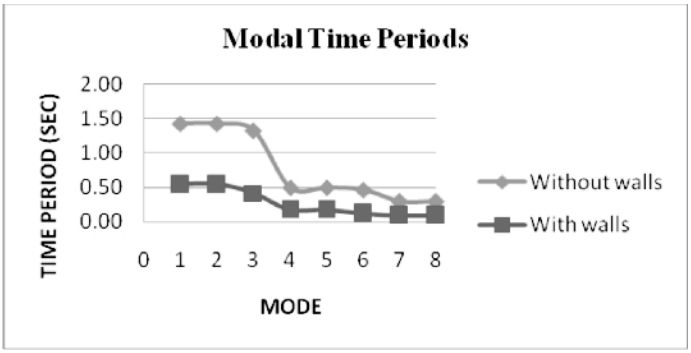
1. Modal Time Periods
2. Storey Shears
3. Storey Displacements
4. Support Reactions
5. Reinforcement %age of Columns

The comparison of results on the basis of above parameters will be made in terms of Tables and Graphs in the proceeding sections.

Modal Time Periods

Table 1: Model time periods

Modal Time Periods (Sec)		
Mode	**T sec (Without walls)**	**T sec (With walls)**
1	1.43	0.54
2	1.43	0.54
3	1.33	0.42
4	0.50	0.18
5	0.50	0.18
6	0.47	0.13
7	0.30	0.09
8	0.30	0.09



Storey Shears

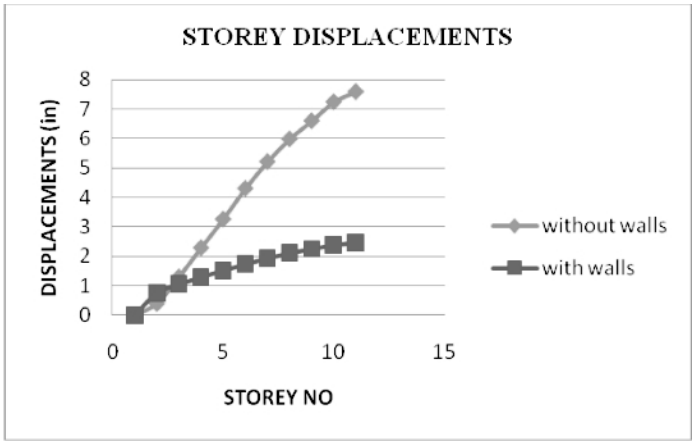
Table 2: Storey shear

Storey Shear		
Storey Id	**Vx (Kip) without walls**	**Vx (Kip) with walls**
Storey1	7435.8	23565.83
Storey2	7361.45	23085.51
Storey3	6993.22	21437.16
Storey4	6452.62	19451.49
Storey5	5864.94	17262.38
Storey6	5233.98	14868.86
Storey7	4511.8	12193.17
Storey8	3673.43	9239
Storey9	2710.85	6130.33
Storey10	1376.65	2885.54

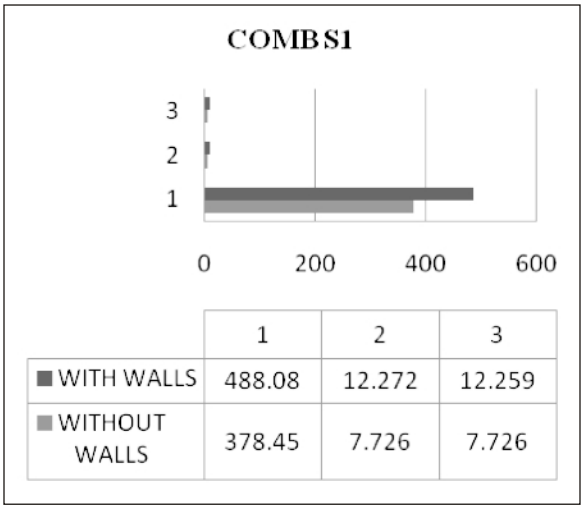
Storey Displacements

Table 3: Storey displacement

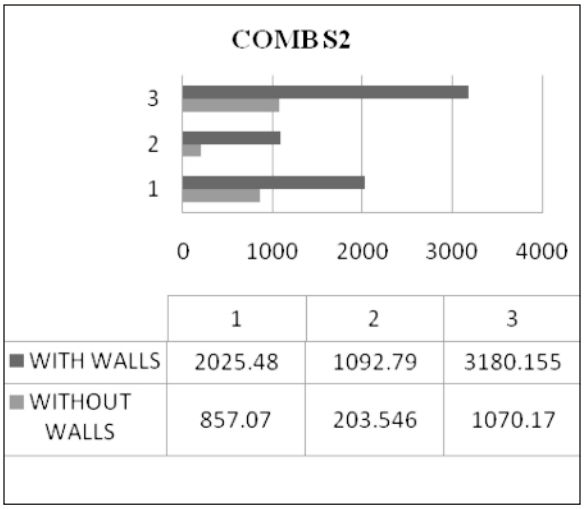
Storey Displacements		
Storey ID	**Ux (inch) without walls**	**Ux (inch) with walls**
Base	0	0
Storey1	0.41	0.77
Storey2	1.31	1.06
Storey3	2.30	1.31
Storey4	3.28	1.53
Storey5	4.32	1.75
Storey6	5.23	1.95
Storey7	6.00	2.12
Storey8	6.61	2.27
Storey9	7.26	2.39
Storey10	7.60	2.48



Support Reactions

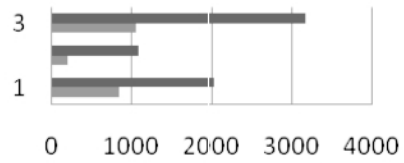


	1	2	3
WITH WALLS	488.08	12.272	12.259
WITHOUT WALLS	378.45	7.726	7.726



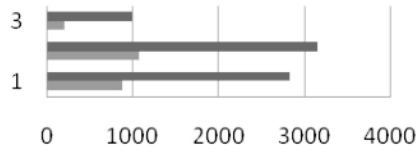
	1	2	3
WITH WALLS	2025.48	1092.79	3180.155
WITHOUT WALLS	857.07	203.546	1070.17

COMB S3



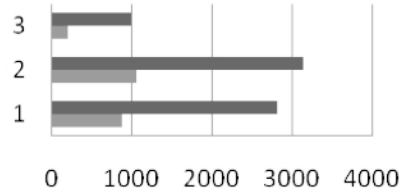
	1	2	3
WITH WALLS	2025.48	1092.79	3180.155
WITHOUT WALLS	857.07	203.546	1070.17

COMB S4



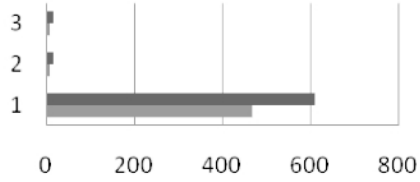
	1	2	3
WITH WALLS	2822.72	3148.794	997.005
WITHOUT WALLS	882.62	1070.227	203.853

COMB S5



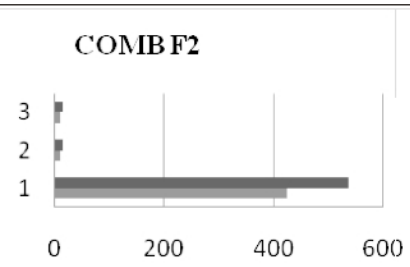
	1	2	3
WITH WALLS	2822.72	3148.794	997.005
WITHOUT WALLS	882.62	1070.227	203.853

COMB F1



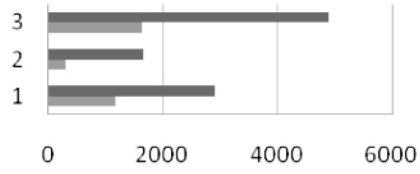
	1	2	3
WITH WALLS	611.75	15.64	15.634
WITHOUT WALLS	469.38	9.528	9.528

COMB F2

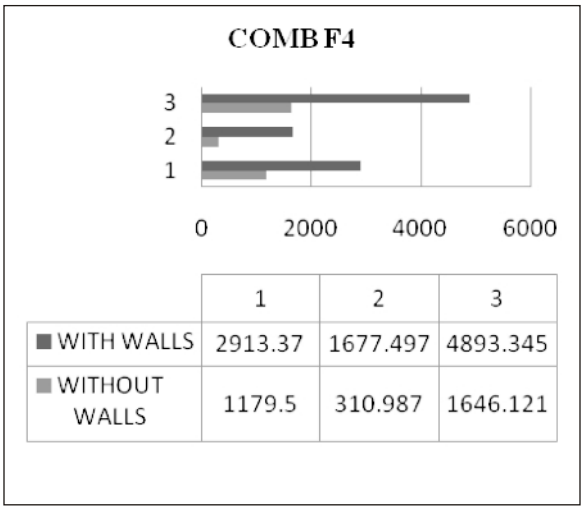


	1	2	3
WIITH WALLS	537.06	13.048	13.027
WITHOUT WALLS	425.84	8.811	8.811

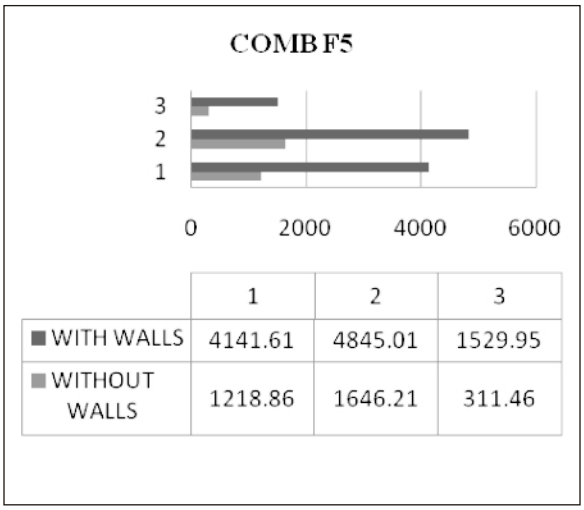
COMB F3



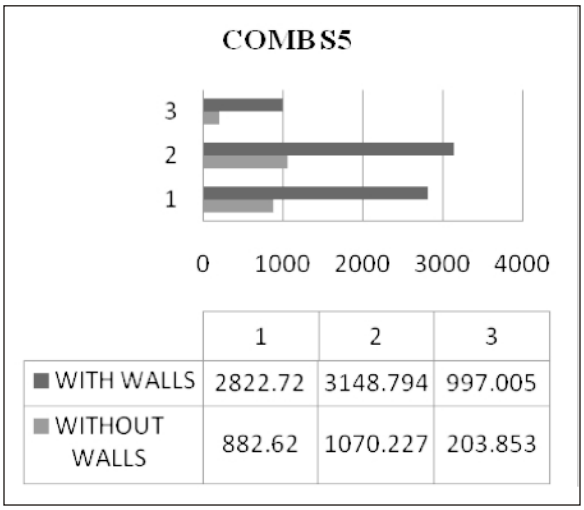
	1	2	3
WITH WALLS	2913.37	1677.497	4893.345
WITHOUT WALLS	1179.5	310.987	1646.121



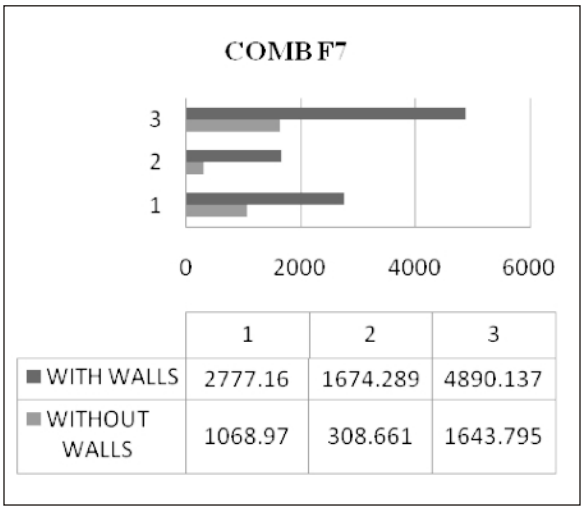
	1	2	3
WITH WALLS	2913.37	1677.497	4893.345
WITHOUT WALLS	1179.5	310.987	1646.121



	1	2	3
WITH WALLS	4141.61	4845.01	1529.95
WITHOUT WALLS	1218.86	1646.21	311.46

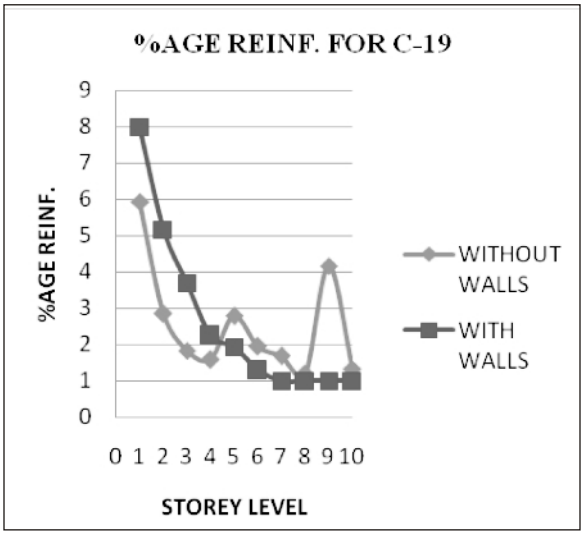


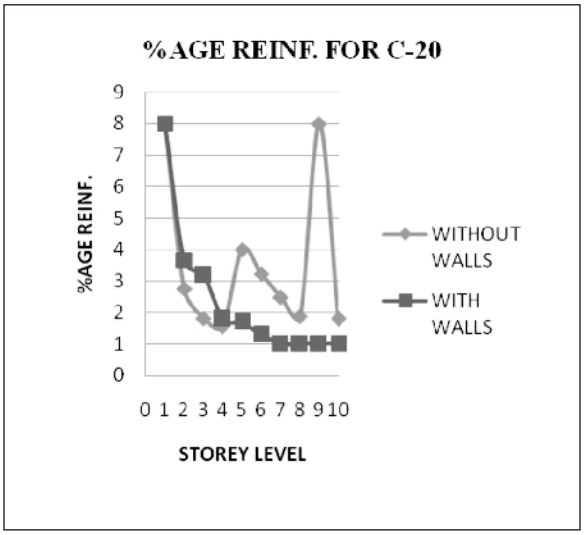
	1	2	3
WITH WALLS	2822.72	3148.794	997.005
WITHOUT WALLS	882.62	1070.227	203.853

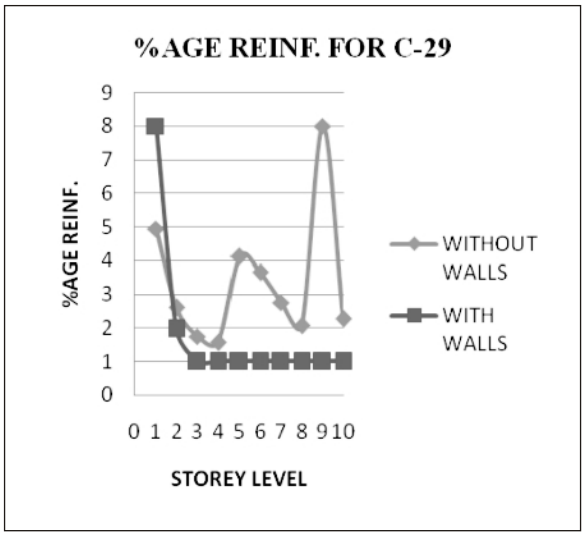


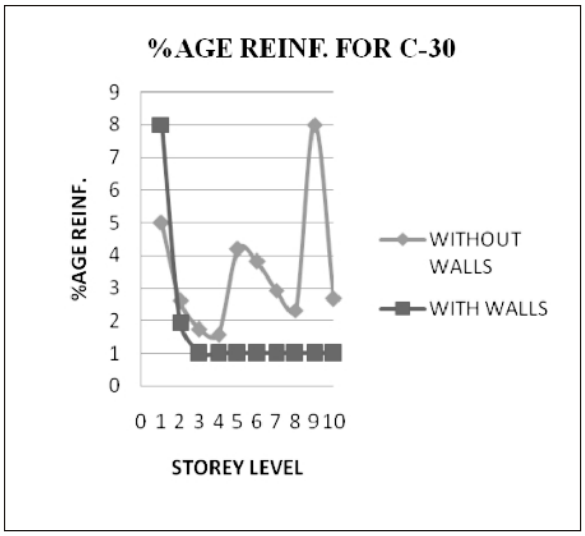
	1	2	3
WITH WALLS	2777.16	1674.289	4890.137
WITHOUT WALLS	1068.97	308.661	1643.795

Reinforcement %Age of Columns









Conclusions

1. Time periods of the model with modeled infill walls are lower than the model without infill walls. It happened because the infill walls increase the overall stiffness of the structure and hence tries to lower the time periods.
2. Storey shear of the structure with modeled infill walls are quite larger than the structure without modeled in fill walls since Time periods of the model with modeled infill walls are lower than the model without infill walls as explained. Hence keeping in view of the above results, it can be suggested that stiffness of infill walls should be considered for design since it represents a true picture of the behavior of structure.
3. Storey displacements in the structure with modeled infill walls are much lower than the structure without infill walls except at 1rst story level. It might be due to reason that the infill walls adds the significant stiffness to the structure which results in the lowering of story displacements and hence moments.
4. However at 1rst storey level the displacement behavior of the structure is reversed since it creates the well known effect of “Soft Storey” as is called by UBC. A soft storey effect appears to occur in the structure at storey level where the difference of storey stiffness of the two adjacent stories are more than 20% as prescribed by UBC. Since the infill non structural walls are not present beneath the plinth level, hence it creates the difference between the storey stiffness of plinth and 1rst level which in turns give rise to the soft storey phenomenon and producing large deflections and hence moments at that level as will be seen in later stages. Keeping in view, it can be suggested that the true behavior of the structure will be presented only when we consider the stiffness of infill walls or proper consideration should be given to soft storey phenomenon.
5. The column reinforcement of the structure in which stiffness of infill walls are considered are less comprehensively showing the fact that stiffness of walls gives beneficial in terms of economy to the structure.
6. The reinforcement in the plinth column is much higher i.e. 8% in the same structure showing the negative effects of soft story phenomenon and hence its importance to consider in design.

References

[1] ACI 318-02, Building Code Requirements for Structural Concrete.

[2] Arshad K. Hashmi and Alok Madan, Department of Civil Engineering, Indian Institute of Technology Delhi, New Delhi 110 016, Damage forecast for masonry infilled reinforced concrete framed buildings subjected to earthquakes in India.

[3] CVR Murty1 And Sudhir K Jain, Beneficial Influence Of Masonry Infill Walls On Seismic Performance of RC Frame Buildings.

[4] Fuat Demir ,Kasım Armagan Korkmaz, and Mustafa Sivri Gueguen, Suleyman Demirel University, Civil Engineering Department, Cunur, Isparta, TURKIYE, Earthquake Assessment of RC Structures With Infill Walls, International journal of Science and Technology 2007.

5] P. G. Asteris, M.ASCE, Lateral Stiffness of Brick Masonry Infilled Plane Frames.

[6] Uniform Building Code-1997.

Water Resources Management and Crop-water Requirement in Arid and Semi Arid Regions of Pakistan

Ossama Algahtani[1], Khaled[2], A. R. Ghumman[3] and N. Saqib[4]

Abstract

The study has focused on the effective and rational utilization of water resources in arid and semi arid areas of Pakistan. The crop water requirements with respect to water sources availability have been analyzed. A case specific study approach was adopted which was kept limited to district boundaries so that the local government and other organizational structure may implement the recommendations and results of this study. Data regarding agriculture statistics, rainfall, evaporation, cropping patteren and other relevant parameters for the area under study was collected from Meteorological Department of Pakistan, Pakistan Council for Research in Water Resources, National Agriculture Research Council Islamabad, Pakistan, Arid Agriculture University Rawalpindi, Pakistan, Water Gate Way Pakistan Website, Federal Bureau of Statistics and Ministry of Food Agriculture and Live Stock. CropWat software was used to estimate the crop water requirements. It is observed that crop water requirements of most of the crops are not met with the average rain and need an additional source of water both in Gawadar and Zhob areas.

Introduction

Sustainable water resources development is becoming highly important world wide. More than one third of the earth's land surface is arid where the process of land degradation has intensified in recent decades causing desertification. About 14 percent of the total area of Pakistan is under main deserts namely Thar, Cholistan, Thal, Kharan and Chagai. On the whole more than half the country receives less than 205 mm of annual rainfall. The mean annual precipitation ranges from less than 100 mm in parts of the Lower Indus Plain to over 750 mm near the foothills in the Upper Indus Plain. The other main source is groundwater. However groundwater in most part of the area under study is saline. When there is no rain in these deserts for long period it causes drought and people are compelled for migration along with their livestock. As a result of drought, grazing lands are reduced or abolished which cause increase in livestock mortality and add untold miseries to human beings (Kahlown 2002 [1])

Twenty three districts in Balochistan, Thar, Dadu and Thatta in Sindh, and Cholistan in Punjab are usually hit by the drought. Its impacts are reflected all over the country. Famine-like situation is faced in the severely affected areas resulting in migration of millions of families to 'safe areas', hundreds of people have lost their lives and countless livestock have died due to lack of water and grass. The cycle of drought does not seem to be ending. The amount of rainfall has been consistently decreasing over the last few years. It is feared that its effect may prolong causing more devastating situation in the country.

The problem of water scarcity in Pakistan does not solely stem from a shortage of resources. Its roots also lie in the realm of awareness and willingness to find a participatory solution that is feasible and sustainable. In this regard one of the key parameter which should be known before hand is the crop water requirements. The present study has addressed evaluation of this parameter. There is extensive work on methods to estimate the crop water requirements (Hargreaves and Zohrab (1985) [2], Ullah et al., (2001) [3], Bastiaanssen and Chandrapala (2003) [4], Bastiaanssen et al. (2005) [5], Allen et al., (2005) [6], Kuo et al., (2006) [7], Laghari et al., (2008) [8] and Shakir et al (2010) [9] . Penman–Monteith equation has commonly been used by many researchers. FAO favors a standardized grass of 12 cm height and ASCE has recommended one short crop (grass) and a tall crop (alfalfa) as the reference crops. Although use of the above two methods in various parts of the world is cited (Allen et al., 2005 [6]), only a couple of studies (Ullah et al., 2001 [3]; Laghari et al., 2008) [8]) are found for the Pakistan region employing this scientific method for estimating evapotranspiration and thus crop water requirement. Laghari et al.'s (2008) [8] study does not cover a whole canal system but rather is limited to a few wheat fields, while Ullah et al. (2001) [3] estimated crop water requirements for the whole of the Indus Basin irrigation system. Ullah et al. (2001) [3], however, fell short of estimating the irrigation demands and comparing the estimated irrigation demands

[1, 2] Salman Ibn Abdelaziz University, Saudi Arabia, [3, 4] Civil Engineering Department, University of Engineering and Technology, Taxila.

with the actual supplies. The present study will fill this gap of estimating the crop water requirement based on the ASCE standardized Penman–Monteith (2000) [10] equation, converting it to irrigation demands and then comparing it with the actual water supply to obtain an insight into possible improvement of a canal irrigation system in Pakistan.

Evaluation of Crop Water Requirements (CWR) Using CropWAT for Windows

CropWat for Windows Software has been used to assess crop sustainability through evaluation of CWR for the case study areas. It is very useful software for irrigation planning and management which can be used to evaluate rain fed production and drought effects as well as the efficiency of the irrigation practices.

Methodology

The methodology developed in this study followed the case study approach to assess impacts of drought and to find out sustainable water resource management techniques. Climatic and agriculture data have been used to run models to analyze water resource management and irrigation practices in vogue with an emphasis on need of adopting such crop patterns as are more sustainable for a certain agro-climatic zone.

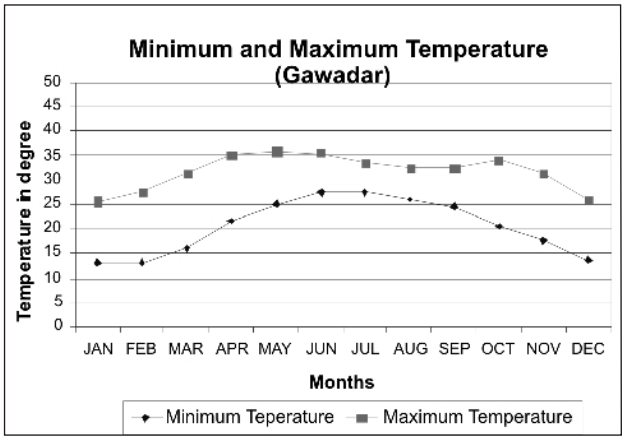
A water a balance approach has been used to determine the effect of the drought on irrigation water budget. This approach first quantifies the variations in water supply (with prime consideration of rainfall for rain fed areas). Secondly the responses of the consumptive use are quantified using a consumptive use model which utilizes climate and plant parameters to estimate the crop consumptive use as well as irrigation water requirement for different crops for the study area. The major considerations for the selection of the area for case study are the availability of, accuracy and adequacy of the data for the modeling and validation processes. Such considerations eventually lead to selection of two case studies from Balochistan Districts of Gawadar and Zhob (worst hit by drought of 1999-2002). Data for the study area was collected from Ministry of Agriculture, Food and Livestock, (MINFAL) Islamabad, Pakistan (MINFAL 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003) [11], Nazir (1993) [12], Kahlown, (2002) [1] and Federal Bureau of Statistics (FBS), 2001) [13]

Case Study for Gawadar Area

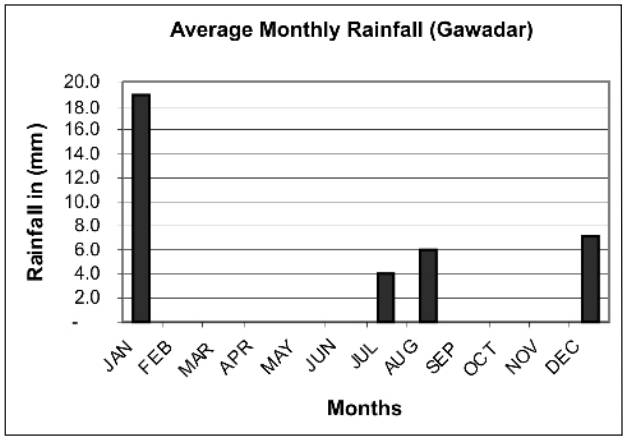
Geography and Climate of the Area

The Gawadar District, lies at 0-300 meters above sea level, and its climate is dry arid hot. It is placed in "warm summer and mild winter" temperature region. The oceanic influence keeps the temperature lower than that in the interior in summer and higher in winter. Monthly temperature and rain is given in figure 1(a, b). The mean temperature in the hottest month (June) remains between 310 C and 320 C. The mean temperature in the coolest month (January) varies from 180 C to 190 C. The uniformity of temperature is a unique characteristic of the coastal region in Balochistan. Occasionally, winds moving down the Balochistan plateau bring brief cold spells, otherwise the winter is pleasant. In Gawadar, winter is shorter than summer. It stays only from December through February (3 months) while summer starts in March and prolongs up to November (9 months). Mean monthly temperature in summer remains between 210 C and 320 C. In the coldest month, January, the mean monthly temperature remains above 100 C. Freezing temperature has been recorded at Pasni but nowhere else in the district.

Aridity prevails all over the district because average annual rainfall is below 250 mm and in some years annual rainfall was even below 100 mm. Both the monsoons and the Western Depressions result in scanty rainfall but overall precipitation level remains low. According to the Pakistan Meteorological Department, total annual precipitation in 1994 was 159.1 mm at Pasni and 110.6 mm at Jiwani. (Pak MET)The extent of precipitation affects the supply of drinking water in Gawadar district as most of it is provided from reservoirs which are rain-dependent.



a) Monthly average of the minimum and maximum daily temperatures °C



b) Monthly average rainfall (mm)

Figure 1: Average Monthly Temperature and Rainfall at Gawadar

Classification of Land and Agriculture

In Gawadar district, agricultural land can be classified into irrigated and un-irrigated. Irrigated land has permanent sources of water like open surface wells or springs. There is not a single karez or kaurjo in the district. Un-irrigated land in Kulanch and Dasht valleys is rain-fed, locally called khushkaba, or flood irrigated (sailaba). Irrigated land is predominantly used for production of fruits however some crops are also cultivated in orchards. Crops like wheat, barley, and jowar are cultivated in un-irrigated land.

The Agriculture Department has sub-divided the net potential area available for cultivation into current fallow, net sown, area sown more than once and culturable waste. In Gawadar, about 97 percent of the area is either not yet reported or not available for cultivation. The arable land constitutes just one percent of the total area. The major reasons are scarcity of irrigation water and unavailability of agricultural labor.

Major Crops and Cropping Pattern

Major crops of the area are given in table 1. In Gawadar agricultural crops are categorized in two types, Rabi and Kharif, according to their cultivation seasons. Rabi crops are sown in winter and harvested in late winter or during early summer while Kharif crops are sown in summer and harvested in late summer or early winter. The major Rabi crops of the district include wheat, barley, muttar pulse, and various vegetables, but their current volume of production is comparatively negligible. Kharif crops in Gwadar district include mainly fruits and water melons, various vegetables and some jowar and mash. Fodder is cultivated throughout the year. Fruits are produced in Kharif season.

Dates, mangoes, fodder, water melons, citrus, tomatoes and wheat are the major agricultural produce of the area. Wheat is cultivated mostly in un-irrigated areas, rain-fed for the most part, while dates, other fruits, water melons, and vegetables are cultivated in irrigated lands. Although average per hectare yield of wheat in Gwadar (1,143 kg/Ha.) is not at par with the average yield of wheat for the province of Balochistan (2,320 kg/Ha.), it fulfils the local needs to some extent. Barley is another crop cultivated mostly in khushkaba lands.

Different vegetables and fodder are cultivated throughout the year. Date, the major produce among fruits, is most frequently planted on irrigated land. Dates require continuous irrigation and more care.

Table 1: Average Annual Production of Major Crops in Gawadar

Crops	Area (Ha.)	Production (tones)	Yield (kg/Ha.)
Fruits	2,496	20,997	8,412
Fodder	128	2,140	16,719
Water Melons	100	1,880	18,800
Vegetables	82	830	10,121
Wheat	70	80	1,143
Pulses	70	35	500
Barley	40	35	875
Jowar	30	25	833
Coriander	10	5	500
Bajra	9	5	556
Guar Seed	3	2	667
Total	3,038	26,034	

Irrigation Practices

The major source of irrigation in Gwadar is streams and springs etc. Other significant sources are tube wells, operated by diesel, and open surface wells from where water is taken out for irrigation with the help of diesel pumps. The remaining is khushkaba or sailaba land dependent upon rainfall. The total number of tube wells has decreased in khushkaba lands over the past years. It may be due to increase in the diesel price.

In flood irrigation, rain-water is harvested into the fields by embankment of fields. This provides enough water for cultivation of crops like wheat and barley. All the tube wells are diesel powered. These are used for irrigation in case diesel, illegally imported from Iran, is cheap. Otherwise tube wells are not operated and farmers wait for rainfall. Here it is necessary to clarify a misunderstanding, that the open surface wells with diesel pumps are often also called tube wells. In Pasni and Shadikaur area, there are 22 such open surface wells being used for irrigation.

Case Study for Zhob

Geography and Climate

The district lies between 30 ° 30¢ to 32 ° 05¢ north latitudes and 67 ° 26¢ to 70 ° 00¢ east longitudes. It is bounded on the north by Afghanistan and South Wazirestan agency of FATA, on the east by the tribal area adjoining Dera- Ismail -Khan district of NWFP and Musakhel district, on the south and south-west by Loralai and Killa Saifullah districts. Total area of district is 20297 square kilometers. Topographically, the district is covered with mountains and hills intersected by the broad valley of Zhob and its tributaries. The Toba-Kakar range covers the western half of the district extending from the boundary of Afghanistan up to the Zhob River. The general elevation of the district is 1500 to 3000 meters.

On the south of Zhob valley, a succession of parallel ridges running from north-east to south-west divide the drainage of the Zhob from that of the Bori valley in the Loralai district.

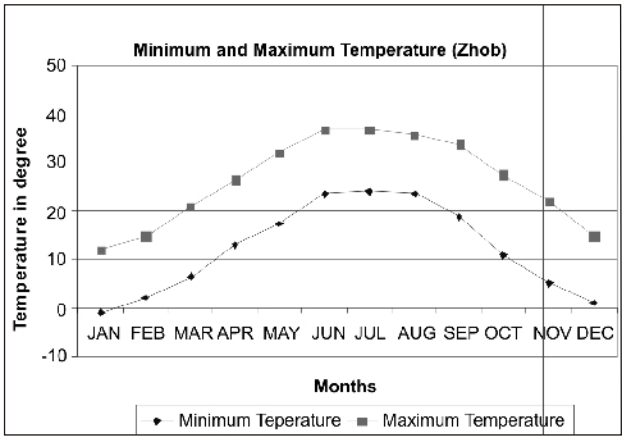
Monthly temperature and rain are shown in figure 2 (a and b). The climate of the district is hot and dry in summer and cold in winter. June is the hottest month with mean maximum and minimum temperature of about 37°C and 23°C respectively. January is the coldest month with mean maximum and minimum temperature of about 13°C and -1°C degree respectively. The dust- storms occur in summer from July to September accompanied by thunderstorms. In winters the wind blows from the west and is very cold. The winds from the Southwest and east are also common, the later invariably brings rain. The wind from the north occasionally blows during September to April bring drought and damage standing crops. Rainfall is scanty and varies with the altitude. Most of the rainfall is received during winter seasons. The district is one

of the biggest districts in Balochistan. The land use table 2 below shows that only 4% of the total geographical area is as yet reported.

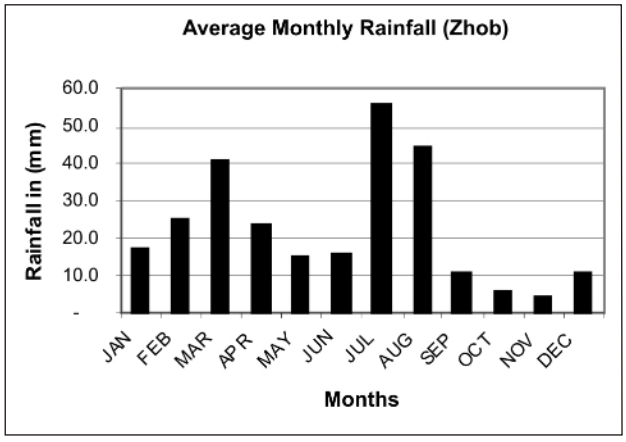
Table 2: Land use statistics for Zhob

Land use	Area (Ha)	%age of total district area
Total Geographical Area	1,651,787	100
Area not Reported	1,590,857	96
Area Reported	60,980	4
- Not available for cultivation	10,853	0.7
- Area under Forest	13,010	0.8
- Culturable waste	13,387	0.8
- Area under water logging / Salinity		
- Arable land	23,730	1.4
Potential area available for cultivation	37,117	2.2

The soil of Zhob district is rocky and shingle gravel. Vegetation and forest are spread over considerable parts of the district. Wheat, Cotton, pulses almond , apricot, cherry and pomegranate are the major agricultural produce of the area.



a) Monthly average of the minimum and maximum daily temperatures °C



b) Monthly average rainfall (mm)

Figure 2: Average Monthly Temperature, Precipitation and Sunshine at Zhob

Irrigation Practices

The two principal drainage channels of the district are the Zhob and the Kundar Rivers, flow into the Gomal River. The Zhob River rises at Tsari Mehtarazai pass, the watershed a distance of about 400 kilometers. The broad plain of the Zhob River is occupied by the alluvial formation. The Kundar River rises from the central and highest point of the Toba Kakar range, a few kilometers northeast of the Sakir. It constitutes boundary between Pakistan and Afghanistan territory for a considerable length. The other subsidiary rivers or streams are the Baskan, Chukhan, Sri Toi, Sawar, Surab, etc. Only 16,206 acres of land is irrigated throughout the district. Majority of the area in the district is irrigated by springs, perennial irrigation, flood irrigation and delay action dam/storage dam irrigation schemes.

Results and Discussion

The analysis of the major crops was done to analyze the sustainability of the Current Crop Pattern under drought conditions and the management of water resources in general. A model was developed using CROPWAT, Software. The results are shown in figures 3 and 4 for Gawadar and Zhob respectively.

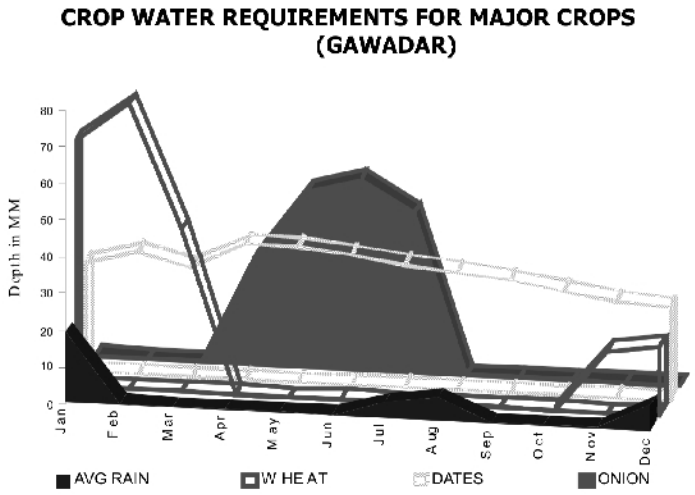


Figure 3: CWR for Major Crops and Average Rainfall Gawadar

The model developed for Gawadar (Figure 1) shows that the average rain is not sufficient to meet High CWR (Crop Water Requirement) for wheat as wheat is grown mostly in un-irrigated areas of the district. On the other hand dates are ever green and can extract their CWR from sub surface water also they are mostly grown in irrigated lands. Onion also has CWR requirement during May to July and is not fit for rain fed areas except in heavy pre monsoon rainfall conditions. In case of Zhob (Figure 2) the crop water requirement for wheat in the study area can be met from the average rainfall. However storage is required to supplement the rain fall if the rain is below average. Sunflower, apple, cherry, potato and pomegranate have comparatively higher water requirement and need storage water in addition to the rain for proper growth. Potential for growth of pulses during September to December is high.

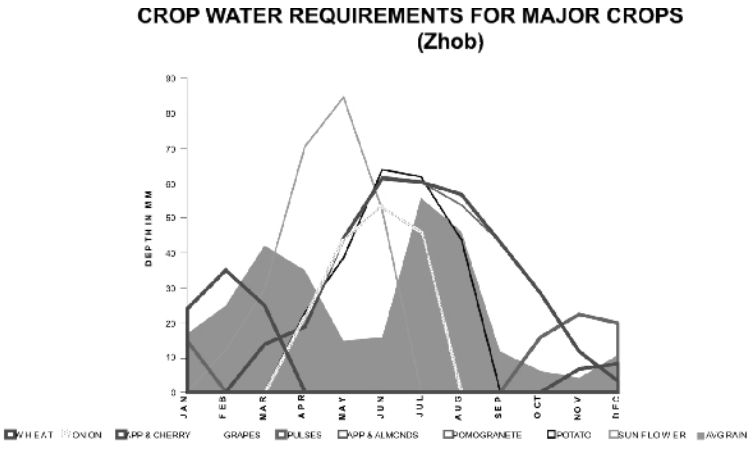


Figure 4: Analysis of Crop Pattern for CWR (Zhob)

Conclusions

There is no easy solution to problem of water conservation in semi-arid regions as it varies from place to place and depends upon the local climate, soils, and vegetation and human requirements. It is observed that the average rainfall trend in Zhob suits the crop water requirements for wheat in study area but below average rainfall period storage is required for supplementary requirement to maintain the rate of

production. Sunflower, apple, cherry, potato and pomegranate need storage water to sustain. Pulses has high growth rate during September to December is high.

In case of Gawadar the average rain is not sufficient to meet crop water requirement for wheat during peak season. Dates in Gawadar are ever green and can extract their water requirement from sub surface water also they are mostly grown in irrigated lands. Onion also has high water requirement during May to July and is not fit for rain fed areas of Gawadar.

From the discussion on the case studies regarding the cropping pattern the following recommendations can be made for better management of water resources under drought conditions.

1. First of all a better communication for drought preparedness and forecasting needs to be established.
2. An important aspect of drought prone areas-Soil degradation should also be controlled.
3. Efficient Irrigation Techniques and new crop patterns should be tested through pilot projects at Local Government Levels for easy implementation and better participation. Crop Patterns with: Low Delta Crops mixed with High Delta Crops. Some examples are listed as: Oil seeds, Pulses, Local vegetables, Tree varieties, Grasses, Medicinal plants, Castor, Guar (cluster bean), Mung bean.
4. Water harvesting and delay action dams are already being implemented but a lot more can be achieved through better participation of locals.

References

[1] Federal Bureau of Statistics (FBS), 2001), Statistics Division, Govt of Pakistan, “Pakistan Statistical Yearbook 2001”, April 2001.

[2] Kahlown, M. A., Majeed, A. and Tahir, M. A. (2002). Water Quality Status in Pakistan, Report 2001-02, Pakistan Council of Research in Water resources, report No. 121-2002, Islamabad, Pakistan, October 2002.

[3] MINFAL Ministry of Agriculture, Food and Livestock, Islamabad, Pakistan (1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003) Agricultural Statistics of Pakistan, Government Press Islamabad.

[4] Hargreaves G. H. and Zohrab A. S., 1985, Reference Crop Evaporation from Temperature,, Applied Engineering in agriculture, Vol 1, issue 2, pp: 96, 99.

[5] Shakir A. S., Khan N. M. And Qureshi M. M., 2010, Canal Water Management: Case Study Of Upper Chenab Canal, In Pakistan, Irrigation & Drainage, Vol. 59, pp : 76-91

[6] Bastiaanssen W. G. M. and Chandrapala, L. (2003). “Water balance variability across Sri Lanka for assessing agricultural and environmental water use.” Agricultural Water Management, 58(2), 171-192.

[7] Bastiaanssen W. G. M, Noordman E. J. M., Pelgrum H, Davids G., Thoreson B. P., Allen R. G., 2005, SEBAL Model with Remotely Sensed Data to Improve Water-Resources Management under Actual Field Conditions. J. of Errigation and drainage Engineering ASCE, Vol. 131, No 1 pp: 85-93.

[8] Kuo, S., Ho s., Liu C., 2006, Estimation irrigation water requirements with derived crop coefficient, for upland and paddy crops in ChiaNan Irrigation Association, Taiwan, Agricultural Water Management, 82(3), 433-451.

[9] Ullah MK, Habib Z, Muhammad S. 2001. Spatial distribution of reference and potential evapotranspiration across the Indus Basin irrigation system. IWMI Working Paper 24, International Water Management Institute, Lahore, Pakistan.

[10] Laghari KQ, Lashari BK, Memon HM. 2008. Perceptive research on wheat evapotranspiration in Pakistan. Irrigation and Drainage 57: 571–584.

[11] Nazir A. 1993, “Water Resources of Pakistan”, Miraj uddin Press, Lahore September 1993.

[12] Allen R. G, Pruitt W. O., Raes D., Smith M., Pereira L. S., 2005, Estimating Evaporation from Bare Soil and the Crop Coefficient for the Initial period using common soil Information, ASCE, J. Irrigation and Drainage Engineering, Vol 131 Issue 14 (10 pages)

A New Technique to Improve Comprehensibility in Inductive Learning

K. Shehzad[1] and S. Khushnood[2]

Abstract

Inductive learning algorithms are well-known for their ability to produce models that are not only more predictive as compared with other techniques but also comprehensible for easy interpretation. RULES-6 is one such algorithm among the RULES family of simple inductive learning algorithms. It employs user-specified search space pruning techniques and can handle large noisy datasets efficiently as compared to its predecessor RULES¬3 Plus. However, because of the issue of overlapping native to the RULES family, the number of rules induced by RULES-6 is still quite large, which adversely affects its comprehensibility. This paper presents a new technique to handle the problem of overlapping native to the RULES family and tests the proposed approach on the RULES-6 algorithm. Experimental results clearly demonstrate that the proposed technique improves the induction capability of RULES-6. The result is a more compact rule set that is not only representative of the regularities in the dataset but also more comprehensible from the user's point of view.

Keywords: Pruning, Data Mining, Knowledge Discovery, Machine Learning, Inductive Learning, Supervised Learning, Rule Induction, Classification.

Introduction

Data mining and machine learning have been the focus of scientific research during the last two decades, and there has been much progress in both the fields. Data mining is the process of discovering patterns in large repositories of data. Machine learning is a tool that can be used for developing data mining applications. It makes use of simple and practical techniques to extract useful information from raw data [1].

Inductive learning techniques are very popular in the field of data mining because of their ability to generate hypotheses that are not only accurate but also easy to understand. The learned hypothesis can then be used for predicting the class of unseen data, which may either be a nominal label or a continuous value. The unseen data must be from the same domain in order to be used for classification or prediction. The reasons for the popularity of inductive learning algorithms for data mining applications are threefold, namely accuracy, efficiency and comprehensibility.

Inductive learning algorithms can be divided into two basic categories, those that learn rules on a class-per-class basis such as CN2 [2, 3], RIPPER [4] etc and those that use a seed example methodology to generates rules such as AQ [5-7], RULES family [8-14] etc. The former class of algorithms generates rules for each class in turn, whereas the latter selects a seed example and the class of the seed is considered to be the target class while all other classes are considered as negative. The problem of overlapping rules is common to both these categories of algorithms.

RULES (RULe Extraction System) is a family of inductive learning algorithms which uses ideas from both AQ and CN2. RULES-1 [8], RULES-2 [9] and RULES-3 [11] developed by Pham and Aksoy were the first three algorithms and the simplest in the family. Later, Pham and Dimov [12] incorporated beam search and a specialization measure called the H measure [15] into RULES-3 to develop RULES-3 Plus [12]. RULES-3 Plus has been employed for the extraction of classification rules for solving different engineering problems, e.g. the recognition of design form features in CAD models for computer aided process planning [16], the mapping of manufacturing information to design features [16] and the classification of defects in automated visual inspection [17]. An incremental version of RULES-3 Plus was later proposed in [13]. RULES-5 [14] was the next algorithm in the family, the main strength of which was its ability to handle continuous attributes.

Because it does not employ pruning and emphasizes completeness and consistency, RULES-3 Plus fails to cope with large and noisy datasets. To overcome these deficiencies, RULES-6 [10] was proposed, which can handle large noisy datasets using search space pruning techniques. Although RULES-6 gave far fewer rules and was also more accurate than RULES-3 Plus, the number of rules induced is still high because of

[1&2] MED, University of Engineering and Technology, Taxila.

the overlapping problem native to the RULES family. This means that because of rules covering common areas within the instance space, multiple rules cover a single instance which increases the total number of rules needed to capture the concept. And the problem is exacerbated by the fact that the RULES family only marks the covered examples instead of removing them.

This paper presents a new technique to handle the problem of overlapping native to the RULES family and confirms the efficacy of the proposed approach by testing it on the RULES-6 algorithm. The proposed technique adopts a simple criterion in order to address the problem of overlapping, which results in a more compact rule set with higher classification accuracy. The criterion is based on the number of new instances covered by a rule, i.e. instances that have not been covered by the rule set created so far. This results in a reduced degree of overlapping between the induced rules which increases the compactness of the rule set and at the same time enhances its accuracy on the test data.

The paper is organized as follows. Section 2 delineates RULES-6, the algorithm that was proposed to improve the scalability and noise handling capabilities of RULES-3 Plus. Section 3 outlines the proposed technique for handling the problem of overlapping with RULES-6. Experimental evaluation of the proposed technique is presented in section 4. Section 5 concludes the paper and suggests directions for future research.

The RULES-6 Algorithm

RULES-6 works in a fashion similar to its predecessor RULES-3 Plus, by performing a general-to-specific beam search to induce a rule from a seed example. A pseudo¬code description of RULES-6 is given in Fig. 1, while the Induce_One_Rule procedure is outlined in Fig. 2. The default values of w, MinNegatives, and MinPositives for RULES-6 are 4, 1, and 2 respectively [10].

```
Procedure Induce_Rules (TrainingSet, BeamWidth)
RuleSet = ∅                                                    (step 1)
While all examples in the TrainingSet are not covered Do       (step 2)
  Take a seed example s that has not yet been covered.         (step 3)
  Rule = Induce_One_Rule (s, TrainingSet, BeamWidth)           (step 4)
  Mark the examples covered by Rule as covered.                (step 5)
  RuleSet = RuleSet ∪ {Rule}                                   (step 6)
End While
Return RuleSet                                                 (step 7)
End                                                            (step 8)
```

Figure 1: A pseudo-code description of RULES-6

The seed is selected at step 4 of the Induce_Rules procedure, after which the control is transferred to the Induce_One_Rule procedure. Subsequently, both PartialRules and NewPartialRules are initialized to empty, and a rule is formed containing the attribute values of the seed example only. All conditions in the antecedent of the rule are then marked as “Not Existing” and the most general rule so formed is declared to be the BestRule, which is then added to PartialRules for specialization. At this point, the While loop at step 2 is activated and continues to iterate until PartialRules becomes empty again.

The BestRule is now specialized at step 3 by changing the status of the conditions in its antecedent to “Existing”. This results in n NewRules initially, each with a single condition in its antecedent, where n is the number of attribute-values in the seed example. At step 4, as opposed to its predecessor RULES-3 Plus which prefers consistency over generality, the RULES-6 algorithm replaces the *BestRule* with the *NewRule* based on only the specialization measure of the *New Rule*.

```
Procedure Induce_One_Rule (s: Seed example, Instances: Training set,
w: Beam width)
PartialRules = NewPartialRules = ∅
BestRule = most general rule (the rule with no conditions)          (step 1)
PartialRules = PartialRules ∪ {BestRule}
While PartialRules ≠ ∅ Do                                           (step 2)
  For each Rule ∈ PartialRules Do
  {First, generate all specialisations of the current rule, save useful ones
   and determine all the InvalidValues according to one of the
   conditional tests in steps (5), (6) or (7)}
    For each nominal attribute A_i that does not appear in Rule Do
      If v_is ∈ Rule.ValidValues, where v_is is the value of A_i in s
      Then
        NewRule = Rule ∧ [A_i = v_is]                               (step 3)
        If NewRule.Score > BestRule.Score Then                      (step 4)
          BestRule = NewRule
        If Covered_Positives (NewRule) ≤ MinPositives OR            (step 5)
          Covered_Negatives (Rule) − Covered_Negatives (NewRule)
          ≤ MinNegatives OR                                         (step 6)
          Consistency (NewRule) = 100% Then                         (step 7)
          Parent (NewRule).InvalidValues =
          Parent (NewRule).InvalidValues + {v_is}                   (step 8)
        Else
          NewPartialRules = NewPartialRules ∪ {NewRule}             (step 9)
    End For
  End For
  Empty PartialRules
  For each Rule ∈ NewPartialRules Do
  {Next, delete partial rules that cannot lead to an improved rules and
   determine all the InvalidValues according to the conditional test in
   step (10)}
    If Rule.OptimisticScore ≤ BestRule.Score Then                   (step 10)
      NewPartialRules = NewPartialRules − {Rule}                    (step 11)
      Parent (Rule).InvalidValues = Parent (Rule).InvalidValues +
      Last_Value_Added (Rule)                                       (step 12)
  End For
  For each Rule ∈ NewPartialRules Do
  {Finally, remove from the ValidValues set of each rule all the values
   that will lead to unnecessary construction of useless specialisations at
   subsequent specialisation steps}
    Rule.ValidValues = Rule.ValidValues − Parent (Rule).InvalidValues
                                                                    (step 13)
  End For
  If w > 1 Then
    Remove from NewPartialRules all duplicate rules
    Select w best rules from NewPartialRules and insert into
    PartialRules                                                    (step 14)
    Remove all rules from NewPartialRules
End While
Return BestRule
End
```

Figure 2: A pseudo-code description of the Induce One Rule procedure of RULES-6

If any *NewRule* satisfies any one of the three search space pruning steps 5, 6 and 7, it is pruned and the last value used to specialize it is added to the *InvalidValues* set of its *ParentRule*. Step 10 further attempts to prune rules from the *NewPartialRules* based on whether the optimistic score of the rule is less than or equal to the score of the *BestRule,* in which case the last value used for specialization is added to the *InvalidValues* set of its *ParentRule*. Step 13 then removes from the *ValidValues* set of each *NewRule* the *InvalidValues* set of its *ParentRule.*
Finally, based on whether the beam width w is greater than 1, all duplicate rules are removed from *NewPartialRules* at step 14, and the w *BestRules* are copied from *NewPartialRules* into *PartialRules*. Subsequent to this, the *NewPartialRules* set is initialized to null and the control is transferred to the While loop at step 2, which repeats the specialization iteration. The rules in the *PartialRules* with n condition(s) in their antecedent now act as *ParentRules* to produce *NewRules* with n + 1 conditions in their antecedent.

When there are no rules left in *NewPartialRules* to be copied into *PartialRules*, the While loop at step 2 terminates and the *BestRule* is returned to the *Induce_Rules* procedure. The instances covered by the Rule are marked at step 5, the Rule is added to the *RuleSet*, and the control is transferred to the While loop at step 2 in order to select another seed example and repeat the iteration. When no uncovered instances are left in the training dataset, the final RuleSet is returned.

The Proposed Technique

As already mentioned in section 1, what makes the RULES family different from other covering algorithms is that the examples covered by a rule are only marked instead of being deleted. This is necessary since if it is not done, the same rule may be induced again. Although all covering algorithms suffer from the problem of overlapping which increases the total number of rules needed to capture the concept, the issue is more pronounced in the case of algorithms that only mark the examples instead of removing them. This is because all the examples continue to be used for the purpose of calculating both the accuracy as well as the score of each newly formed rule, resulting in an increased level of overlapping. However, the advantage of taking into account the complete set of examples each time a new rule is formed is that both the fragmentation problem (reduction in the amount of data during the later stages of induction) as well as the small disjuncts problem (low coverage rules with a high error rate) can be avoided [18].

The RULES-6 algorithm (and the entire RULES family for that matter) works in the following fashion. The user-defined data type “rule” has the following data members in it: Rule.Covered, Rule.Classified, Rule.NewClassified etc. This information is calculated for each rule by the procedure Induce_One_Rule. The variable Rule.Classified tells how many instances in the target class have been covered by the newly induced rule. By contrast, Rule.NewClassified gives the total number of instances of the target class that have been covered by the rule from among the instances of that class that have not yet been covered by the rule set created so far. This information is very crucial and can be capitalized to minimize the overlapping between the rules.

Using the NewClassified information of the newly induced rule and the idea of invalid values used in the RULES-6 algorithm, the proposed technique has been designed as follows. When a rule is returned by the procedure *Induce_One_Rule*, it is immediately checked to see how many new instances it has covered. If the new instances covered by the rule are less than a user-specified threshold, then all attribute values marked as “Existing” in the rule are marked as “Invalid” for the seed example. Subsequently, the same seed example is used to induce the rule again. However, in view of the fact that the attribute values marked as “Invalid” in the seed are also transferred to all partial rules created during the search, an alternative rule is generated this time. It is for this reason that the proposed technique has been named OMAR which stands for “Overlapping Minimization using Alternative Rules”. A pseudo-code description of RULES-6 with OMAR is outlined in Fig. 3.

Since the attribute values used in the last rule were not used the next time for specialization, the alternative rule is quite different from the one induced previously. If the new classified instances of the alternative rule are still less than the user-specified threshold, a new rule may be induced yet another time and this process continues until the new classified instances exceed the threshold.

```
Procedure Induce_Rules (TrainingSet, BeamWidth, Threshold)
RuleSet = ∅                                                    (step 1)
While all examples in the TrainingSet are not covered Do       (step 2)
  Take a seed example s that has not yet been covered.         (step 3)
  Rule = Induce_One_Rule (s, TrainingSet, BeamWidth)           (step 4)
  While Rule.NewClassified < Threshold Do                      (step 5)
    Mark the "Existing" attribute values in the Rule as
    "Invalid" for the seed example s.                          (step 6)
    Rule = Induce_One_Rule (s, TrainingSet, BeamWidth)         (step 7)
  Mark the instances covered by the Rule as covered.           (step 8)
  Add Rule to the RuleSet                                      (step 9)
End While
Return RuleSet                                                 (step 10)
End                                                            (step 11)
```

Figure 3: A pseudo-code description of RULES-6 with OMAR

Empirical Evaluation

The technique suggested in this paper has been tested thoroughly on a population of 40 datasets, all of which have been downloaded from the repository of machine learning databases, University of California at Irvine (UCI) [19], with the exception of the dataset Depression from Williams College [20]. These datasets come from a variety of domains and have been summarized in Table 1. Each dataset name is followed by a block capital letter in parenthesis which indicates whether the dataset is nominal, continuous or mixed type. Occasionally, there might also be a suffix L to denote that the dataset is large[1].

All evaluation was carried out on the Windows 7 operating system using an Intel Pentium 2.0 GHz Dual-Core computer with 2 GB of RAM. In order to handle continuous and mixed type datasets, RULES-6 uses its own online discretization procedure [21], which discovers the cut points for each attribute during the learning phase. To ensure the accuracy of the estimate, the evaluation approach used in this study is the standard stratified 10¬fold cross-validation [22]. Two commonly used criteria Rules_Reduction and Accuracy_Increase as well as an Overall_Improvement criterion have been used for testing the efficacy of the proposed technique.

Table 2 compares the results obtained by activating OMAR against those of the default RULES-6 algorithm, while Table 3 summarizes the results in terms of efficacy on the total number of employed datasets. It can be seen from the second column in Table 2 that the threshold value for the majority of the tested datasets (29) is 1%. For 9 datasets, the threshold is 5% while there are only two datasets for which the threshold value is 10%. Based on this, the default value of the threshold for RULES-6@OMAR can be taken to be 1%. The last row in Table 2 shows that the average classification accuracy for RULES-6@OMAR increases by 1.59% while there is a 21.29% reduction in the average number of rules. The top 10 datasets in terms of accuracy increase are Car(N), Cover-Type, Heart-Cleveland, Flags, Promoters, Nursery, Ecoli, Arrhythmia, Depression, and Heart-Hungarian. The percent increase in accuracy for these datasets is 7.28, 5.42, 5.22, 4.17, 3.85, 3.77, 3.51, 3.11, 3.10, and 2.94 respectively. In terms of the average number of rules, the top 10 datasets are Horse-Colic, Flags, Credit-Approval, Depression, Hypothyroid, Heart-Hungarian, Pima-Indians, Nursery, Splice, and Heart-Cleveland. The percent reduction in the number of rules for these datasets is 55.71, 43.86, 40.30, 38.66, 35.38, 32.61, 32.48, 30.17, 29.92, and 29.90 respectively.

[1] Any dataset for which the product of number of instances and number of attributes is greater than 50k is regarded as large in this study.

Table 1: Summary of Datesets

No.	Dataset	# Instances	# Attributes			# Classes
			Total	Nominal	Continuous	
1	Adult(M)L	48,842	14	8	6	2
2	Anneal(M)	798	38	32	6	6
3	Arrhythmia(M)L	452	279	73	206	13
4	Breast-Cancer(C)	699	10	0	10	2
5	Breast-Cancer(N)	286	9	9	0	2
6	Car(N)	1,728	6	6	0	4
7	Chess(N)L	3,196	36	36	0	2
8	*Connect-4(N)L	3,376	42	42	0	2
9	*Cover-Type(M)L	3,100	54	47	7	7
10	Credit-Approval(M)	690	15	9	6	2
11	Depression(M)	428	17	11	6	2
12	Dermatology(M)	366	34	33	1	6
13	Ecoli(C)	336	8	0	8	8
14	Flags(M)	194	29	29	0	8
15	German-Credit(M)	1,000	20	13	7	2
16	Hayes-Roth(N)	160	5	5	0	2
17	Heart-Cleveland(M)	303	13	8	5	5
18	Heart-Hungarian(M)	294	13	8	5	2
19	Hepatitis(M)	155	19	13	6	2
20	Horse-Colic(M)	368	27	18	9	2
21	Hyperthyroid(M)L	3,711	29	22	7	4
22	Hypothyroid(M)L	3,772	29	22	7	2
23	Image(C)	210	19	0	19	7
24	Ionosphere(C)	351	34	0	34	2
25	Iris(C)	150	4	0	4	3
26	Lymphography(N)	148	18	18	0	4
27	Mushroom(N)L	8,124	22	22	0	2
28	Nursery(N)L	12,960	8	8	0	2
29	Parkinsons(C)	195	22	22	0	2
30	Pendigits(C)L	10,992	16	0	16	10
31	Pima-Indians(C)	768	8	8	0	2
32	Post-Operative-	90	8	8	0	3
33	Promoters(N)	106	58	58	0	2
34	Soybean-Large(N)	683	35	35	0	19
35	Spect(C)	267	44	0	44	2
36	SPECT-Heart(N)	267	22	22	0	2
37	Splice(N)L	3,190	61	61	0	3
38	Tic-Tac-Toe(N)	958	9	9	0	2
39	Vehicle(C)	846	18	0	18	4
40	Wine(C)	178	13	0	13	3

* The Cover-Type(M)L had 581,012 instances originally but was sampled to approx. 1/187 of its full size.

It also becomes clear from the summary in Table 3 (row 1) that there is only one dataset for which there is an increase in the number of rules namely the Arrhythmia data. The number of rules remains the same for five datasets while it reduces for 34 datasets. From the accuracy point of view (row 2), there is an increase in accuracy for 31 datasets which equates to 78% of the total number of datasets employed. The accuracy reduces for only 7 datasets while it remains the same for 2 datasets. Furthermore, for 75% of the datasets, there is an overall improvement in the model produced by RULES-6@OMAR.

Table 2: Rules-6@OMAR vs RULES-6

Dataset	Th. {1, 5, 10}	No. of Rules		Redc. (%)	Accuracy (%)		Incr. (%)
		R6	R6@ OMAR		R6	R6@ OMAR	
Adult(M)L	1	191	135	**29.32**	77.90	77.57	0.42
Anneal(M)	5	40	40	0.00	97.92	97.92	0.00
Arrhythmia(M)L	1	103	120	16.50	56.25	58.00	**3.11**
Breast Cancer(C)	1	27	25	**7.41**	94.64	94.06	0.61
Breast Cancer(N)	1	71	50	**29.58**	68.21	69.59	**2.02**
Car(N)	5	92	65	**29.35**	61.40	65.87	**7.28**
Chess(N)L	1	47	35	**25.53**	94.06	94.50	**0.47**
Connect 4(N)L	1	401	305	**23.94**	66.38	67.55	**1.76**
Cover Type(M)L	1	446	327	**26.68**	56.68	59.75	**5.42**
Credit	1	67	40	**40.30**	80.74	82.23	**1.85**
Depression(M)	5	119	73	**38.66**	67.38	69.47	**3.10**
Dermatology(M)	5	35	31	**11.43**	88.29	90.00	**1.94**
Ecoli(C)	1	50	47	**6.00**	78.39	81.14	**3.51**
Flags(M)	5	57	32	**43.86**	60.00	62.50	**4.17**
German Credit(M)	1	239	187	**21.76**	71.90	72.15	**0.35**
Hayes Roth(N)	1	30	25	**16.67**	70.00	71.86	**2.66**
Heart	5	97	68	**29.90**	55.36	58.25	**5.22**
Heart	1	46	31	**32.61**	77.86	80.15	**2.94**
Hepatitis(M)	1	30	25	**16.67**	82.67	82.75	**0.10**
Horse Colic(M)	1	70	31	**55.71**	81.11	82.33	**1.50**
Hyperthyroid(M)L	1	72	69	**4.17**	97.99	97.67	0.33
Hypothyroid(M)L	10	65	42	**35.38**	90.79	92.40	**1.77**
Image(C)	1	37	37	0.00	76.67	78.10	**1.87**
Ionosphere(C)	1	49	35	**28.57**	88.24	89.05	**0.92**
Iris(C)	10	11	10	**9.09**	95.33	94.00	1.40
Lymphography(N)	1	32	30	**6.25**	82.86	84.20	**1.62**
Mushroom(N)L	1	26	25	**3.85**	97.58	99.10	**1.56**
Nursery(N)L	1	358	250	**30.17**	67.99	70.55	**3.77**
Parkinsons(C)	1	30	29	**3.33**	90.56	92.68	**2.34**
Pendigits(C)L	1	500	475	**5.00**	93.19	93.30	**0.12**
Pima Indians(C)	5	117	79	**32.48**	69.34	70.00	**0.95**
PO Patient(M)	1	25	18	**28.00**	65.00	66.70	**2.62**
Promoters(N)	5	20	16	**20.00**	78.00	81.00	**3.85**
Soybean Large(N)	1	53	52	**1.89**	90.63	90.16	0.52
Spect(C)	1	54	53	**1.85**	78.85	78.55	0.38
SPECT Heart(N)	1	34	34	0.00	82.69	82.69	0.00
Splice(N)L	1	254	178	**29.92**	90.35	92.90	**2.82**
Tic Tac Toe(N)	1	28	28	0.00	92.74	92.50	0.26
Vehicle(C)	5	189	157	**16.93**	68.05	69.29	**1.82**
Wine(C)	1	29	29	0.00	87.50	89.55	**2.34**
Average:		106	83	**21.29**	79.29	80.55	**1.59**

Table 3: Summary of Rules-6@OMAR

No.	Criterion	No. of Datasets for which					
		Redc.		Equal		Incr.	
		Total	%age	Total	%age	Total	%age
1	No. of Rules	34	**85%**	5	**13%**	1	**3%**
2	Accuracy	7	**18%**	2	**5%**	31	**78%**
3	Overall Impr.	30					**75%**

Conclusions and Future Work

This paper has presented a new technique for handling the issue of overlapping native to the RULES family. Because they cover common areas within the instance space, overlapping rules tend to increase the total number of rules needed to capture the concept which adversely affects the comprehensibility of the learned rule set. The proposed technique was tested on RULES-6 and was found to be effective in terms of both the number of rules and classification accuracy.

It may be possible to improve the work carried out in this research in several ways. Firstly, additional pre-pruning criteria may be combined with the OMAR technique in order to further reduce the number of rules. Secondly, RULES-6@OMAR may also be combined with certain stopping criteria with the objective of informing the algorithm as to when to halt learning.

References

[1] Witten, I. H. and Frank, E. Data Mining - Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers, San Francisco, USA, 2005.

[2] Clark, P. and Boswell, R. Rule Induction with CN2: Some Recent Improvements. Proc. of the 5th European Working Session on Learning, Porto, Portugal, 1991, 151-163.

[3] Clark, P. and Niblett, T. The CN2 Induction Algorithm. Machine Learning, 1989, 3, 261-283.

[4] Cohen, W. W. Fast Effective Rule Induction. Proc. of the 12th Int. Conf. on Machine Learning, 1995, 115-123.

[5] Hong, J., Mozetic, I., and Michalski, R. S. AQ15 - Incremental Learning of Attribute-Based Descriptions from Examples, the Method and User's Guide, Editor^Editors. 1986, Department of Computer Science, University of Illinois: Urbana. 1041-1045.

[6] Michalski, R. S. On the Quasi-Minimal Solution of the General Covering Problem. Proc. of the 5th Int. Symposium on Information Processing, Bled, Yugoslavia, 1969, 125-128.

[7] Wojtusiak, J., Michalski, R. S., Kaufman, K. A., and Pietrzykowski, J. The AQ21 Natural Induction Program for Pattern Discovery: Initial Version and Its Novel Features. 18th IEEE International Conference on Tools with Artificial Intelligence, 2006, 523-526,

[8] Pham, D.T. and Aksoy, M. S. An Algorithm for Automatic Rule Induction. Artificial Intelligence in Engineering, 1993, 8(4), 227-282.

[9] Pham, D.T. and Aksoy, M. S. RULES: A Simple Rule Extraction System. Expert Systems with Applications, 1995, 8(1), 59-65.

[10] Pham, D.T. and Afify, A. A. RULES-6: A Simple Rule Induction Algorithm for Supporting Decision Making. Industrial Electronics Society, IECON 31st Annual Conf. of IEEE, Raleigh, North Carolina, USA, 2005, 2184-2189.

[11] Pham, D.T. and Aksoy, M. S. A New Algorithm for Inductive Learning. Journal of Systems Engineering, 1995, 5, 115-122.

[12] Pham, D. T. and Dimov, S. S. An Efficient Algorithm for Automatic Knowledge Acquisition. Pattern Recognition, 1997, 30(7), 1137-1143.

[13] Pham, D. T. and Dimov, S. S. An Algorithm for Incremental Inductive Learning. Proc. of the Inst. of

Mechanical Engineers, Part B: Journal of Engineering Manufacture, 1997, 211(3), 239-249.

[14] Pham, D. T., Dimov, S. S., and Bigot, S. RULES-5: A Rule Induction Algorithm for Problems Involving Continuous Attributes. Proc. of the Inst. of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2003, 217, 1273¬1286.

[15] Lee, C. Generating Classification Rules from Databases. Proc. of the 9th Conf. on Application of Artificial Intelligence in Engineering, PA, USA, 1994, 205-212.

[16] Pham, D. T. and Dimov, S. S. An Approach to Concurrent Engineering. Proc. of the Inst. of Mechanical Engineers, Part B: Journal of Engineering Manufacture, 1998, 212, 13-27.

[17] Jennings, N. R. Automated Visual Inspection of Engine Valve Stem Seals. Internal Report, University of Wales Cardiff, Cardiff, UK, 1996.

[18] Afify, A. A. Design and Analysis of Scalable Rule Induction Systems. Ph. D. Thesis, Systems Engineering Division, University of Wales, Cardiff, UK, 2004.

[19] Blake, C. L. and Merz, C. J. UCI Repository of Machine Learning Databases. University of California, Department of Information and Computer Science, Irvine, CA, 1998. Available from: http://archive.ics.uci.edu/ml/ [Accessed: 24 April 2009].

[20] Veaux, R. D. Datasets for Use in the Data Mining Course. Williams College, Williamstown, Department of Mathematics and Statistics, Bronfman Science Center, MA, 01267, USA, 2007. Available from: http://www.williams.edu/Mathematics/rdeveaux/data. html [Accessed: 24 April 2009].

[21] Pham, D. T. and Afify, A. A. Online Discretization of Continuous-Valued Attributes in Rule Induction. Proc. of the Inst. of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2005, 219(8), 829-842.

[22] Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Proc. of the 14th Int. Joint Conf. on Artificial Intelligence, Montreal, Canada, 1995, 1137¬1143, (Morgan Kaufmann).

Grid High Availability & Service Security Issues with Solutions

Muhammad Zakarya[1], Ayaz Ali Khan[2] and Hameed Hussain[3]

Abstract

DDoS attacks are launched through sending a large quantity of packets to a target machine, using instantaneous teamwork of multiple hosts which are distributed throughout the Grid computing environment. Today DDoS attacks on the Internet in general and especially in Grid Computing environment has become a visible issue in computer networks. DDoS attacks are easy to generate but their detection is a very difficult task and therefore, an attractive weapon for hackers. DDoS streams do not have familiar characteristics, therefore currently available IDS cannot detect these attacks perfectly. Similarly, there implementation is a challenging task. In practice, Gossip based DDoS attacks detection mechanism is used to detect such types of attacks in network, by exchanging traffic over line. Gossip based techniques results in network congestion and have overhead of extra packets.

Keeping the above drawbacks in mind, we are going to propose a DDoS detection and prevention mechanism, that has the beauty of being easy to adapt and more reliable than existing counterparts. We are going to introduce entropy based detection mechanism for DDoS attack detection. Our proposed solution has no overhead of extra packets, hence resulting in good QoS. Once DDoS is detected, any prevention technique can be used to prevent DDoS in Grid environment. The rest of paper is organized as follows. In section I we give some introduction, II is about related work. Section III, IV and V is about existing problem and proposed solution. IV describes statistical and simulation results. VII is about performance evaluation. We conclude in section VIII with challenges and future directions.

Keywords: Normalized Entropy (NE), Denial of Service (DoS), Grid Simulator (GridSim)

Introduction & Concepts

Grid Computing, more specifically computational grids is the application of several systems to a single huge problem at the same time, usually to a scientific or technical problem that needs a large number of CPU processing cycles i.e. more CPU power or access to huge and large amounts of data. One of the main Grid Computing strategies is to use different softwares to divide and apportion different pieces of a single program among several individual systems, may be up to many thousands [25]. These systems, taking part in Grid System are called nodes. Grids are called super computers for economically poor organizations. The GS consists of GN and a GNM. When multiple GS are combined in such a way, that at least one of them registers its available services to a Broker as shown in Fig 1. And others Grid Sites (GS) requests for such registered services from the Broker. The Environment is called Grid Computing Environment.

A. HA in Grid Systems

Any system which is always available to its customers is HA. High availability of grid system can be achieved, through implementing a lot of architectures. For example reduce congestion. It is difficult to achieve HA in today's global village because more services are required to customers. The more congested the network, more systems are offline to its customers. Considering TCP congestion scenario, where TCP drops all extra packets resulting in increased queuing delays. Therefore using traditional TCP congestion detection, avoidance mechanisms are not to achieve HA.

B. QoS in Grid environment

We are trying to study different service level security issues in Grid computing especially in wireless grids, and will try to propose new solutions to their security improvements. As service level security issues like DoS Attacks & Network Congestion, are most important. Solving these issues results in High Availability as well as. In high available systems, QoS services are expected from service providers.

[1] Abdul Wali Khan University (AWKU), Mardan, Pakistan, [2, 3] CIIT, Islamabad, Pakistan.

C. Security Issues

As networks are coming common to layperson in computer technology, the need to provide good services to its customers at any time is essential. Grid computing provides its services to its customers on need basis, means whenever, what is required must be provided. Therefore managing QoS and making the systems available, each and every time, to provide its services to Grid users and customers, is a must.

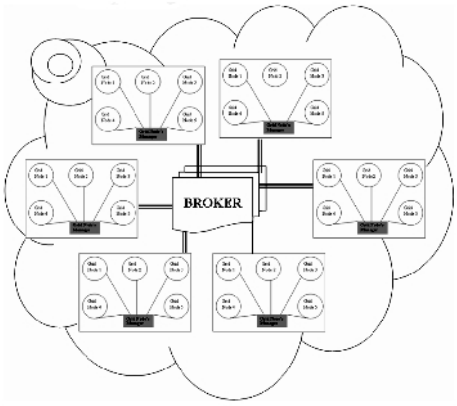


Figure 1: The Grid Computing Environment

D. Distributed DoS Attack

DDoS attacks are launched by sending a large volume of packets to a target machine, using simultaneous cooperation of multiple hosts which are distributed throughout the Grid computing environment. Mostly DDoS attacks are considered as congestion control problem. DDoS attacks are two phases attack. In first phase the attacker finds some vulnerable systems in the network. The attacker install some DDoS tools on these systems, also called zombies or agents. In second phase all zombies create the actual attack on the victim, as shown in figure 3 below [2].

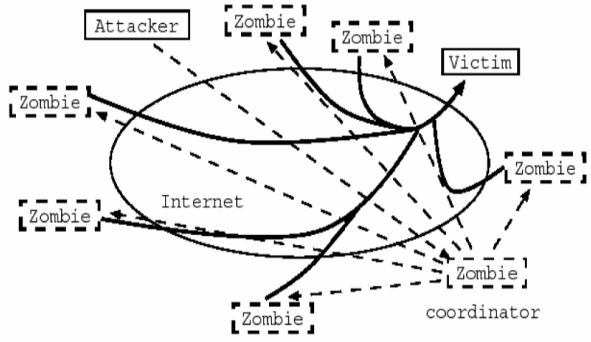


Figure 2: Attacker, Zombies and Victims [2]

E. IP Spoofing

Change of source address in the header of an IP packet is called IP Spoofing. It requires privileged access to network stack (raw socket access). A partial solution to IP Spoofing is to associate a fixed MAC address with each IP address in a subnet to detect spoofing.

Related Work & Existing techniques

In this section we discuss some existing mechanisms and techniques.

A. Mutually Guarded Approach

In wireless communication medium, if a node-A (attacker) (masquerade itself as node-B), sends packets to node-C, where nodes A & B are in the same coverage area, then that packet will also be received by node-B. Therefore node-B will easily catch the attack. But if nodes B & C are in different coverage area, or both nodes B & C are out of range to each other, in that scenario the attacker will successfully launch its attack, as shown in Fig 4.

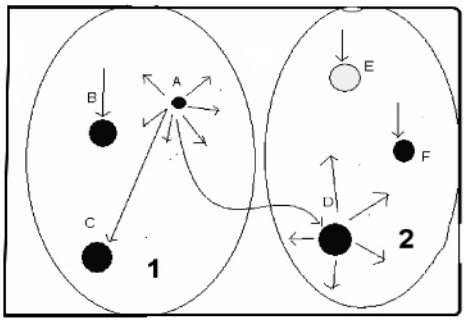


Figure 4: Mutually guarded approach

B. Ingress & Egress Filtering

Ingress & Egress filtering mechanism is shown diagrammatically in Fig 5 [10].

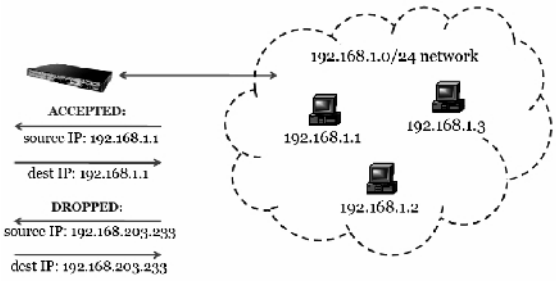


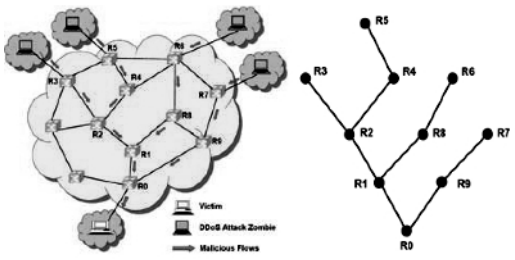
Figure 5: Ingress & egress filtering [10]

C. IP trace-back mechanism

In this technique the attacker is traced, by location. Actually without any mobility, it is some what easy, but when mobility is involved, the attacker cannot be traced easily.

D. Distributed Change point Detection (DCD)

In [6] the authors have proposed a new detection mechanism for DDoS. A CAT is constructed. Nodes in a CAT are ATRs that participate in forwarding the malicious flows. The links in the CAT indicate the path along which attacking traffic goes towards the victim. Once a CAT is constructed, a DDoS attack is detected and ATRs are identified. The next task is to filter out malicious flows.



Figur 6: IP Trace-Back mechanism [6]

E. Moving Target Defense

A Band-Aid solution to a DDoS attack is to change the IP address of the victim computer, thereby invalidating the old address. The technique may work in some cases but administrators must make a series of changes to DNS entries, routing table entries etc.

F. Rate Limiting

Rate-limiting mechanisms compel a rate limit on a set of packets that have been characterized as nasty by the detection mechanism. It is a moderate response technique that is usually deployed when the detection mechanism has many false positives or cannot accurately illustrate the attack flow.

G. Mitigating DDoS Attacks via Attestation (Assayer)

In [9] the authors have proposed a new hardware based attestation mechanism to detect and prevent DDoS attacks. On a per-packet basis, they proposed to provide the network with the dominant ability to identify, the code on the end host that generated or permitted the packet. The story is shown in Fig 7 below.

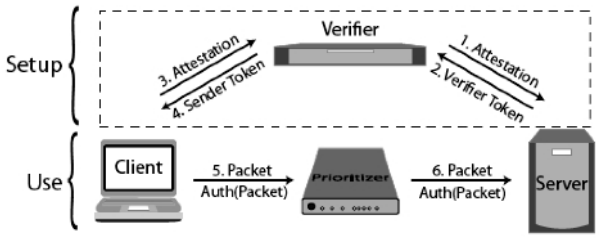


Figure 7: ASSAYER [9]

H. Traffic Shaping

A number of routers available in the bazaar today have features that permit you to limit the amount of bandwidth that some specific type of traffic can consume. This is occasionally referred to as "traffic shaping" technique [10].

I. Internet Protocol Ver 6 (IPv6)

IPv.4 does not have any check or methods to authenticate whether the IP address i.e. source address, that the sender puts into an IPv.4 packet header field, is justifiable or not. As a result, the authentication of source IP address is to be anticipated to enhance and improve an Internet Security against current DoS attacks as shown in Fig 8 [10].

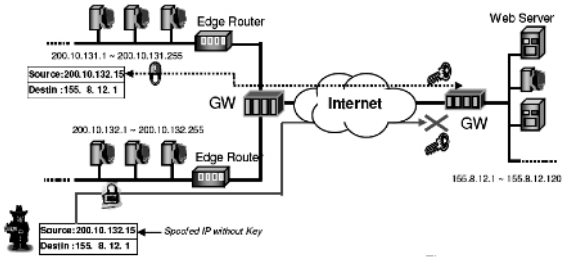


Figure 8: IP Version 6

Existing Problem

We are going to propose a DDoS detection and prevention mechanism, that has the beauty of being easy to adapt and more reliable than existing counterparts. As, in service level security issues DoS Attacks, DDoS & Network Congestion, are most important. Solving the issue

of DDoS also results in High Availability as well as good QoS.

Proposed solution

After a deep study of available techniques, we are going to introduce a new IDS, which can be implemented on our own proposed architecture, resulting in DDoS detection and prevention mechanism.

A. Proposed Architecture

In our proposed architecture, we have divided the whole Grid System into regional areas i.e. GS, where each GS is protected by an AS / GL. Our developed ADS is installed on two places i.e. every Grid Node & AS or on their respective routers. A packet which is detected as cruel once at AS, is marked out, so that Client node can be informed. In our proposed architecture (for future direction), DDoS source is detected for future prevention. A tree is maintained at every router, by marking every packet with path modification strategy, so that the victim is able to trace the sender of the packet. Any packet which was detected as malicious flow, can be confirmed in a second try i.e. confirmation process at GN i.e. victim node. In phase 1 we detect malicious flow, while in phase 2 we have a confirmation algorithm so either to drop the attack flow, or to pass it otherwise. In the given scenario, we consider that AS is configured properly for policed address i.e. the attacker node address or victim IP address.

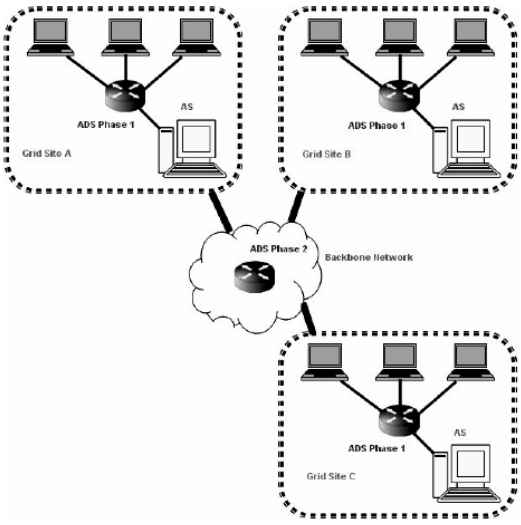


Figure 9: Proposed Grid Architecture

- Authentication Server (AS) or Geographical Authentication & Authorization Server(GAS) is responsible for controlling the geographical area where defined.
- Locally phase 1 is executed & at the core router phase 2 takes place.

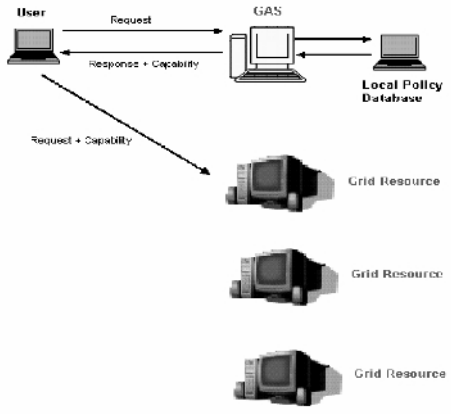


Figure 10: Working diagram of Proposed Grid Architecture

PROS & CONS

- Local Security Policy
- Little computation as compared to Global security policy
- Near the source detection
- No overhead of extra packet
- User accesses GAS, authenticated & authorization check
- Performance Scalability + load balancing + QoS
- No need for resources to check the user identity
- Local & Quick allocation of resources by GAS
- No Single point of failure, affect some part of the Grid
- GAS are required to inform all corresponding GAS in case of new node to any geographical community
- GAS is attacked by DDoS, not possible

B. Intrusion Detection System

IDS may be in software form and/or in hardware form, that will monitor the network for disbelieving activity and alerts the network administrator to take a particular action accordingly. Signature based IDS will observe packets on the network and judge against them to a database maintained with well-known threats. On the other hand, using an ADS, if deviation of user activity is exterior a certain threshold value, it is marked as nasty and a reaction is triggered. After a deep survey of DDoS detection & prevention mechanism we reach to the point that Entropy may be used as DDoS detection metric.

C. Information Theory &Entropy based ADS

According to [14], any statements that have some surprise and meaning are called information. Some consider that information theory is to be a subset of communication theory, but we consider it much more. The word entropy is rented from physics, in which entropy is a measure of the chaos of a group of particles i.e. 2nd law of thermodynamics. If there are a number of possible messages, then each one can be expected to occur after certain fraction of time. This fraction is called the probability of the message. In [23], [24] Shannon proved that information content of a message is inversely related to its probability of occurrence. To summarize, the more unlikely a message is, the more information it contains. In [15], Entropy H(X) is given by

$$H(X) = -\sum_{x \in \mathcal{X}}^{n} p(x) \log p(x)$$

The log is to the base 2 and entropy is expressed in bits. To say randomness is directly proportional to entropy i.e. more random they are, more entropy is there. The value of sample entropy lies between 0 and log(n). The entropy value is smaller when the class distribution belongs to only one & same class while entropy value is larger when the class distribution is more even. Therefore, comparing entropy values of some traffic feature to that of another traffic feature provides a mechanism for detecting changes in the randomness. We use traffic distribution like IP Address & application Port Number i.e. (IP address, Port). If we wants to calculate entropy of packets at a single or unique source i.e. destination, then maximum value of n must be 2^{32} for IPV4 address. Similarly if we want to gauge entropy at multiple application ports then value of n is the total number of ports [16]. In similar way, p(x) where x ϵ X, is the probability that X takes the value x. We randomly examine X for a fix time window (w), then p(x) = mi/m Where, mi is the total number we examine that X takes value x i.e

$$\text{m} = \sum_{i=1}^{n} mi$$

Putting these values in entropy equation 1, we get

$$H(X) = -\sum_{i=1}^{n} (m_i/m) \log (m_i/m)$$

Similarly, if we want to calculate the probability p(x), then m is the entire number of packets, but mi is the number of packets with value x at destination as source [26]. Mathematically given as

$$P(x) = \frac{\text{Number of packets with } x_i \text{ as source (destination) address}}{\text{Total number of packets}}$$

Again if we want to calculate probability p(x) for each destination port, then

$$P(x) = \frac{\text{Number of packets with x as source (destination) port}}{\text{Total number of packets}}$$

Remember that total number of packets is the number of packets observed in a specific time slot (w). When this calculation finishes, normalized entropy is calculated to get the overall probability of the captured flow in a specific time window (w). Normalized Entropy is given by

$$\text{Normalized entropy} = (H / \log n_0)$$

Where no is the number of dissimilar values of x, in a specific time slot (w). During the attack, the attack flow dominates the whole traffic, resulting in decreased normalized entropy. To confirm our attack detection, again we have to calculate the entropy rate i.e. growth of entropy values for random variables, provided that the limit exists, and is given by

$$H(\chi) = \lim_{n \to \infty} \frac{1}{n} H(x_1, x_2 \ldots x_n)$$

Proposed Algorithms

For Detection of Ddos Attack

- A. Mathematical Proof
- Decide a threshold value δ_1
- On edge routers collect traffic flows for a specific time window (w)
- Find probability P(X) for each node packets
- Calculate link entropy of all active nodes separately
- Calculate H(X) for routers using Equation (1)
- Find normalized entropy using Equation (3)
- If normalized entropy < δ_1, identify malicious attack flow

For Confirmation of Attack Flows

- Decide a threshold value δ_2
- Calculate entropy rate on edge router using Equation (4)
- Compare entropy rates on that router, if =< δ_2, DDoS confirmed
- Drop the attack flow

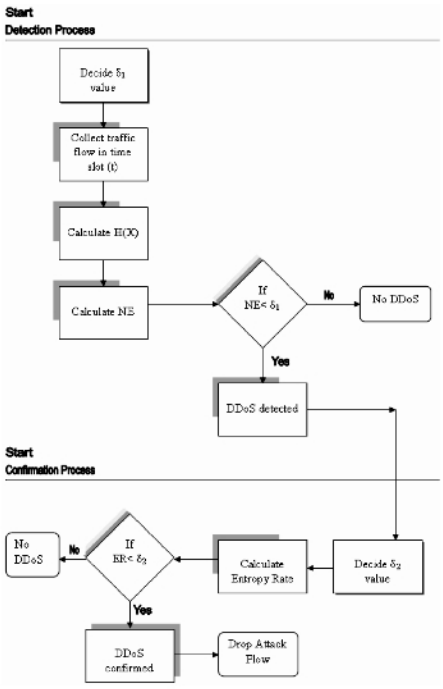


Figure 11: Flow / Transition Diagram

Implementation, Simulation & Results

In this section we describe that how to mathematically or statically implement our proposed scheme, while in section coming after that we have shown our simulation results along with charts form with a practical environment.

A. Mathematical Proof

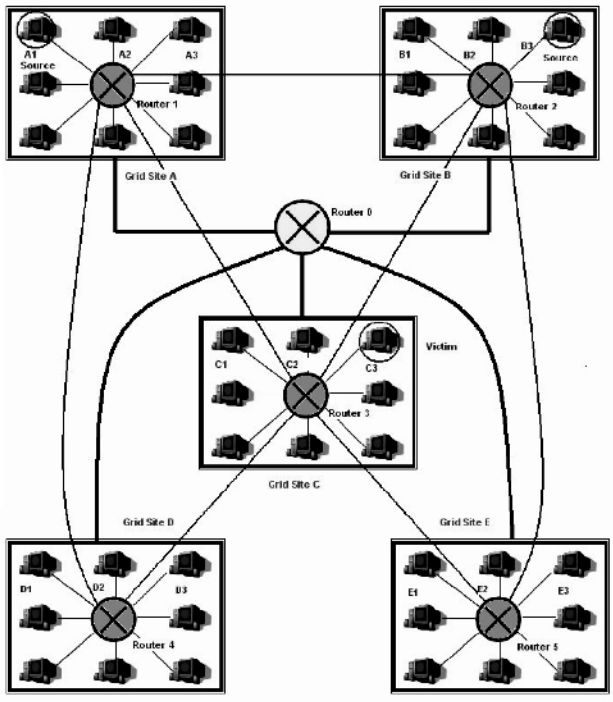


Figure 12: Environment for statistical study

Consider Fig 12, A1 and B3 are attack sources at different Grid Sites, while C3 is the target victim machine. Router 1 will capture traffic flow coming from A1 and Router 2 will capture attack flow thrown by B3, for a specified time window (w). Suppose that we capture the following traffic flow at Router 1 and Router 2, shown in table 1 and table 2, table 3 and table 4 respectively.

Table 1: Traffic at Router 1

Source node	Destination node	No of packets	Entropy
A1	C3	7	0.50
A2	B1	2	0.40
A3	B3	3	0.47
A4	E1	2	0.40

Therefore Router Entropy for Router 1 is 0.50 + 0.40 + 0.47 + 0.40 = 1.77 & as $\log_2 4 = \log 4/\log 2 = 2$ Hence NE is $1.77/\log_2 4 = 0.88$

Table 2: Traffic at Router 2

Source node	Destination node	No of packets	Entropy
B1	D1	2	0.44
B2	A3	1	0.31
B3	C3	6	0.47
B4	E2	2	0.44

Therefore Router Entropy for Router 2 is 0.44 + 0.31 + 0.47 + 0.44 = 1.66 & as $\log_2 4 = \log 4/\log 2 = 2$ Hence NE is $1.66/\log_2 4 = 0.83$

Table 3: Traffic At Router 4

Source node	Destination node	No of packets	Entropy
D1	A1	2	0.46
D2	A3	2	0.46
D3	E3	3	0.52
D4	C2	3	0.52

Therefore Router Entropy for Router 1 is 0.46 + 0.46 + 0.52 + 0.52 = 1.96 & as $\log_2 4 = \log 4/\log 2 = 2$ Hence NE is $1.96/\log_2 4 = 0.98$

Table 4: Traffic at Router 5

Source node	Destination node	No of packets	Entropy
D1	C3	2	0.52
D2	C1	1	0.43
D3	D1	2	0.52
D4	A4	1	0.43

Therefore Router Entropy for Router 2 is 0.52 + 0.43 + 0.52 + 0.43 = 1.90 & as $\log_2 4 = \log 4/\log 2 = 2$ Hence NE is $1.90/\log_2 4 = 0.95$

We can see that as at both routers i.e. Router 1 and Router 2, routers entropy is lesser as only one flow conquered the whole bandwidth. As an outcome NE decreases. If we have a perfect threshold value **δ**, suppose 0.94 then our proposed ADS will consider flows coming from A1 (GS A) and B3 (GS B) as malicious flows, while Grid Site D & Grid Site E have entropy value greater than our considered threshold value 0.95, no attack is detected at these sites.

B. Simulations Study

Simulation Environment

GridSim was used as a simulation environment, for testing the results of our proposed Idea. To simulate our proposed idea we have 3 users with 2 posers of DDoS attack, 2 routers and 3 resources containing any single victim node on the same time, as shown in Fig 13.

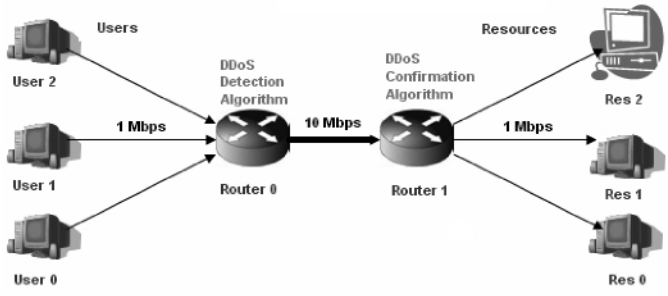


Figure 13: Environment for simulation study

Both routers are connected to each other over a 10 Mbps link, while all other connections are made at 1 Mbps link. Detection algorithm is implemented on router 0, while confirmation is supposed to be implemented on router 1.

Simulation Results

In this section we consider only DDoS detection algorithm on router 0, not to confirm attack.

Case 1:

Table 5: Traffic at Router for User_0

Destination node	Total No of packets	Probability	Entropy
Res_0	5	0.5	0.5
Res_1	2	0.2	0.46
Res_2	3	0.3	0.52

Therefore Router Entropy for Router 2 is 0.5 + 0.46 + 0.52 = 1.48 & as $\log_2 3$ = log3/log2 = 1.58 Hence Normalized Entropy is 1.48/ $\log_2 3$ = 0.93

Table 6: Traffic at Router for User_1

Source node	Total No of packets	Probability	Entropy
Res_0	4	0.4	0.52
Res_1	3	0.3	0.52
Res_2	3	0.3	0.52

Therefore Router Entropy for Router 2 is 0.52 + 0.52 + 0.52 = 1.57 & as $\log_2 3$ = log3/log2 = 1.58 Hence Normalized Entropy is 1.57/ $\log_2 3$ = 0.99.

Table 7: Traffic at Router for User_2

Source node	Total No of packets	Probability	Entropy
Res_0	0	0.0	0.0
Res_1	3	0.3	0.52
Res_2	7	0.7	0.36

Therefore Router Entropy for Router 2 is 0.0 + 0.52 + 0.36 = 0.88 & as $\log_2 2$ = log2/log2 = 1 Hence Normalized Entropy is 0.88/ $\log_2 2$ = 0.88

Performance Evaluation

Our ADS can detect 100% DDoS attack only in case of good threshold value, which is one of the most challenging tasks in developing any ADS. We conclude our story that a threshold value of 0.95 results in good detection rate. A value greater than 0.95, results in good detection rate i.e. 100 % DDoS detection but generate more false positive alarms, as the value is increased from 0.95 to 1.0. The reports are shown in figure 14 and figure 15, are self explanatory

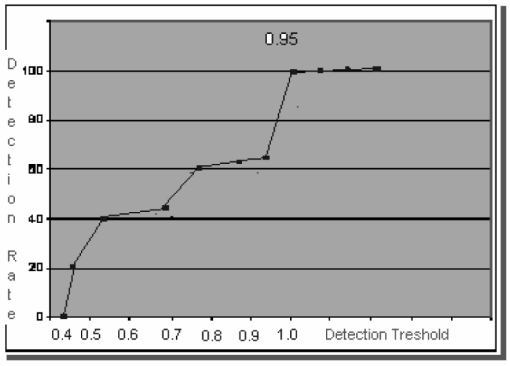


Figure 14: DDoS detection rate

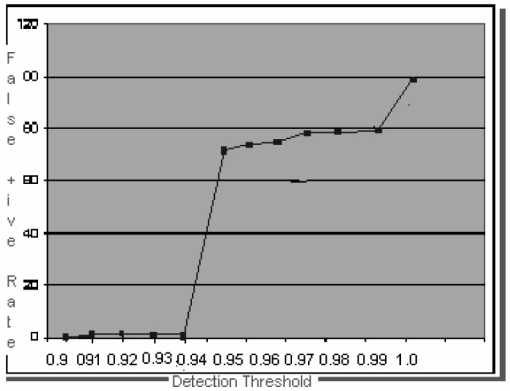


Figure 15: DDoS false positive rate

Conclusion & Future Work

In this paper, we have proposed a new architecture for Grid Computing platform. We have also developed ADS for detection & early prevention of DDoS attacks. In future the proposed idea may be actually implemented over Grid environment to accurately detect DDoS attacks. The idea may also be extended for recovery mechanism for DDoS attacks. Following are some challenges which might be addressed for further enhancement by researchers and scholars.

- Setting perfect threshold values δ1, δ2, some time it must be dynamic in nature to detect DDoS accurately.
- what about different mathematical functions when used for creating attack packets.
- In case of Huge network access separating legitimate flows from attack flows is a challenging task.

References

[1] Bart Jacob, Michael Brown, Kentaro Fukui, Nihar Trivedi, Introduction to Grid Computing, 2005

[2] Kashan Samad, Ejaz Ahmed, Riaz a. Shaikh, Ahmad Ali Iqbal, analysis of ddos attacks and defense mechanisms, 2005

[3] Hang Chau, Network Security – Mydoom, Doomjuice, Win32/Doomjuice Worms and DoS/DDoS Attacks, usa

[4] Puneet Zaroo, a Survey of DDoS attacks and some DDoS defence mechanisms, Advanced Information Assurance (cs 626).

[5] Stephen M. Specht, Ruby B. Lee, Distributed Denial of Service : Taxonomies of Attacks, Tools and Countermeasures, September 2004

[6] Yu Chen, Kai Hwang, Wei-Shinn Ku, Distributed Change point Detection of DDoS Attacks: Experimental Results on DETER Testbed, 2007

[7] Preeti, Yogesh Chaba, Yudhvir Singh, Review of Detection and Prevention Policies for Distributed Denial of Service Attack in MANET, March 2008

[8] S.Meenakshi, Dr.S.K.Srivatsa, A Comprehensive Mechanism to reduce the detection time of SYN Flooding Attack, 2009

[9] Bryan Parno, Zongwei Zhou, Adrian Perrig, Don't Talk to Zombies: Mitigating DDoS Attacks via Attestation, June 2009

[10] Konstantinos Meintanis, Brian Bedingfield, Hyoseon Kim, The Detection & Defense of DDoS Attack, University of Texas

[11] A. Lakhina, M. Crovella, and C. Diot., Diagnosing Network-Wide Traffic Anomalies, ACM SIGCOMM Computer Communication Review, Portland, 2004

[12] L. Feinstein, D. Schnackenberg, R. Balupari, and D. Kindred, Statistical approaches to DDoS attack detection and response, 2003

[13] W. Lee, D. Xiang, Information-theoretic measures for anomaly Detection, IEEE, 2001

[14] DAVID APPLEBAUM, PROBABILITY AND INFORMATION (An Integrated Approach), CAMBRIDGE UNIVERSITY PRESS, 2008

[15] THOMAS M. COVER, JOY A. THOMAS, ELEMENTS OF INFORMATION THEORY, Second Edition, 2006

[16] Dennis Arturo Ludeña Romaña, Yasuo Musashi, Entropy Based Analysis of DNS Query Traffic in the Campus Network, Japan

[17] Rajkumar Buyya, Manzur Murshed, GridSim: a toolkit for the modeling and simulation of distributed resource management and scheduling for Grid computing, 2002

[18] Anthony Sulistio, Gokul Poduval, Rajkumar Buyya, Chen-Kong Tham, Constructing A Grid Simulation with Differentiated Network Service using GridSim, University of Melbourne, Australia

[19] Manzur Murshed, Rajkumar Buyya, Using the GridSim Toolkit for Enabling Grid Computing Education, Monash University, Australia

[20] Anthony Sulistio, Uros Cibej, Srikumar Venugopal, Borut Robic, Rajkumar Buyya, A toolkit for modelling and simulating data Grids: an extension to GridSim, March 2008

[21] Anthony Sulistio, Chee Shin Yeo, Rajkumar Buyya, Visual Modeler for Grid Modeling and Simulation (GridSim) Toolkit, 2003

[22] Microsoft Encarta Encyclopedia, 2009

[23] Claude E. Shannon, A Mathematical Theory of Communication, 1948

[24] Claude E. Shannon, Communication Theory of Secrecy Systems, 1949

[25] Yi-Chi Wu, Wuu Yang, Rong-Horg Jan, DDoS Detection and Trace-back with Decision Tree and Gray Relational Analysis, National Chiao Tung University, Taiwan.

[26] George Nychis, An Empirical Evaluation of Entropy-based Anomaly Detection, May 2007.

Guidelines And Information For Authors

General

Papers may be submitted any time throughout the year. After having received a paper it is sent to three referees, at least one from a technology advanced countries. Papers reviewed and declared fit for publication before 31 December are published next year before 31 March every year. Papers must be submitted on a CD with FOUR Hard copies to the editor Technical journal, University of Engineering and Technology Taxila. Soft copy by e-mail to the following address is preferred.

technical.journal@uettaxila.edu.pk

Authors are required to read the following carefully for writing a paper.

Text

Text should be type-written with M.S word, Arial Font size 10.at single space and with margins as 1.5 inch top, 1 inch right, 1 inch left and 1 inch bottom, on an A-4 size, paper. The title page should include; the title; the name/names of the authors and their addresses, an abstract of about 200 words and keywords followed by the introduction. The text of the paper may be divided into introduction, methodology/Analysis results and discussion, conclusion, references and acknowledgment (if any). All pages should consist of single columns text.

Length

Research paper should not exceed 15 pages as per specifications given above.

Elements of Paper

The basic elements of paper are listed below in the order in which they appear: Title, names of the author and affiliations, Abstract, Body of paper, Acknowledgments, Nomenclature, references, Appendices.

Title

The title of the paper should be concise and definitive.

Names of Authors and Affiliations

Names of authors should consist of first name (or initial), middle initial and last name. The author affiliation should consist of his full address.

Abstract

An abstract up to a maximum of 200 words should open the paper. The abstract should give a clear indication of the objectives, scope and results, the abstract text may be organized to include the background, methods, results and conclusions.

Keywords

Keywords should be included on a separate line at the end of the abstract.

Body of the Paper

Body of the paper may include introduction and literature review, materials and methods, modeling/experimentation, results-discussions and conclusions.

Originality

Only original contributions to engineering and Science literature should be submitted for publication. It should incorporate substantial information not previously published.

Accuracy

All the technical, scientific and mathematical information contained in the paper should be checked with great care.

Use of SI Units

Preferably SI units of Measurements be included.

Mathematics

Equations should be numbered consecutively beginning with (1) to the end of the paper. The number should be enclosed in parentheses (as shown above) and set flush right in the column on the same line as the equation. This number then should be used for referring the equation within the text. Equation may be referenced within the text as “E q. (x)”.”When the reference to an equation begins a sentence, it should be spelled out fully, as “Equation (x).in all mathematical expressions and analyses, symbols (and the units in which they are measured) not previously defined in nomenclature should be explained.

Figures

All figures (graphs, line drawings, photographs, etc.) should be numbered consecutively and have a caption consisting of the figure number and a brief title or description of the figure. This number should be used when referring to the figure in the text. Figure references should be included within the text in numerical order according to their order of appearance. Figure may be referenced within the text as “Fig.-x”. When the reference to a figure begins a sentence, the abbreviation “Fig,” should be spelled out e.g.,” Figure-x”

Tables

All tables should be numbered consecutively. Tables should have a caption consisting of the table number and brief title. This number should be used when referring to the table in text. Table references should be included within the text in numerical order according to their order of appearance. Table should be inserted as part of the text as close as possible to its first reference.

Acknowledgments

All individuals or institutions not mentioned elsewhere in the work who have made an important contribution should be acknowledged.

References

Within the text, references should be cited with name of the author and year in parenthesis. The reference list will be arranged alphabetically.

Example

Chamber (1959) has described a method and Wormleaton (2006) used this method. In case of two authors, name of both the authors will appear with year. For example Khan and Ghumman (2008) studied hydrodynamic modeling for water-saving strategies in irrigation canals. In case of three or more authors it will be cited as: Ghumman et al. investigated use of numerical modeling for management of canal irrigation water in case of continuous references, the references may be separated by comma”,” See the list of sample references.

List of References

References to original source of the cited material as given above as sample reference should be listed together at the end of the paper, footnotes should not be used for this purpose. References should be arranged in alphabetic order. Each reference should include the last name of each author followed by his initials.

1. Reference to journal articles and paper in serial publication include :Last name of each author followed by their initial, Year of publication, Full title of the cited article, Full name of the publication in which it appears, Volume number (if any) in boldface, Issue number (if any) in parentheses, Inclusive page number of the cited article.
2. Reference to the text books and monographs should include: Last name of each author followed by their initial, Year of publication, Full title of the publication, Publisher, City of publication, Inclusive page number of the work being cited, Chapter number(if any).
3. Reference to original conference papers, papers in compiled conference proceedings or any other collection of woks by numerous authors should include: Last name of each author followed by their initial, Year of publication, Full title of the cited paper in quotes, individual paper number (if any) ,Full title of the publication , Initial followed by the name of the editor (if any),followed by the abbreviation, “eds” ,Publisher, City of publication, Volume number (if any),Inclusive Page number of the work being cited.
4. Reference to thesis and technical reports should include: Last name of each author followed by their initial, Year of publication, Full title in quotes, title capitalization, Report number(if any) Publisher or the institution name, City.

Sample References

[1] Konstantinos Meintanis, Brian Bedingfield, Hyoseon Kim, The Detection & Defense of DDoS Attack, University of Texas

[2] A. Lakhina, M. Crovella, and C. Diot., Diagnosing Network-Wide Traffic Anomalies, ACM SIGCOMM Computer Communication Review, Portland, 2004

[3] Strelkof T.1969.One-dimentional equations of open-channel flow. J. Hydraulics Div., ASCE, Vol. 95 (HY3): 861-876

[4] M. Z. Khan and A. R. Ghumman, Hydrodynamic modeling for water-saving strategies in Irrigation canals. Irrigation and Drainage, Wiley InterScience (www.interscience.wiley.com) DOI: 10.1002/ird.375, 2008.

[5] A. R. Ghumman, M. Z. Khan, and M. J. Khan. Use of numerical Modeling for Management of Canal Irrigation Water. Irrigation and Drainage. Wilay Intersciencez : 445-458, 2006.