DATA MODELING AND PROCESSING IN DEREGULATED POWER SYSTEM

By

Lin Xu

A dissertation submitted in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

WASHINGTON STATE UNIVERSITY
School of Electrical Engineering and Computer Science

MAY 2005

©Copyright by LIN XU, 2005
All Rights Reserved

ⓒCopyright by LIN XU, 2005
All Rights Reserved

To the Faculty of Washington State University:

The members of the Committee appointed to examine the dissertation of LIN XU find it satisfactory and recommend that it be accepted.

Chair

ACKNOWLEDGMENTS

I would like to begin by thanking Dr. Kevin Tomsovic, my dissertation advisor and mentor for the past 6 years. Dr. Tomsovic has been a wonderful advisor, providing me with invaluable advice, support. His enthusiasm for research, breadth of knowledge inspires me. His encouragement gives me confidence. I thank him for all the time and energy has invested into my research.

I would like to thank Dr. Anjan Bose and Dr. Curtis Dyreson for agreeing to be on my committee. I thank them for reading my dissertation and providing me helpful comments. I was fortunate enough to work with them. Their enthusiasm and optimism always amaze me.

I would like to thank Dr. David Bakken who gave me a lot of help on my research.

I would like to thank the graduate secretary Ms. Ruby Young, for all her help.

I would like to thank National Science Foundation (NSF) for its support on my research.

I would like to thank my family. I am forever indebted to my parents for everything that they have given me. Their unconditional support and encouragement give me strengths to finish this work. I am very lucky to have such a wonderful family. I dedicate this work to them; to all the people who love me and whom I love.

Last, but far from the least, I thank my wife, Jin Ding from bottom of my heart. Her suggestions were very helpful. I thank her for all her patience and her never-ending optimism when I frustrated and stressed. makes my life full of joy and love.

I cannot imagine that I could have completed this work without her support.

DATA MODELING AND PROCESSING IN DEREGULATED POWER SYSTEM

Abstract

by Lin Xu, Ph.D.
Washington State University
May 2005

Chair: Kevin Tomsovic

The introduction of open electricity markets and the fast pace of changes brought by modern information technology bring both opportunities and challenges to the power industry. Vast quantities of data are generated by the underlying physical system and the business operations. Fast and low cost communications allow the data to be more widely accessed. For electric utilities, it is becoming clear that data and information are vital assets. Proper management and modeling of these assets is as essential to the engineering of the power system as is the underlying physical system. This dissertation introduces several new methods to address information modeling and data processing concerns in the new utility environment.

Presently, legacy information systems in the industry do not make adequate use of the data produced. Hence, a new information infrastructure using data warehousing - a data integration technology used for decision support - is proposed for novel management and utilization of data. Detailed examples and discussion are given on the schema building, extract transform and load (ETL) strategies for power system specific data. The benefits of this approach are shown through a new viewpoint of state estimation. Inaccurate grid information, especially topology information, can be a major detriment to energy market traders' ability to make appropriate bids. A two-stage DC state estimation algorithm is presented to provide them with a simpler data viewpoint to make knowledgeable trading decisions. Numerical results show how the results of a DC state estimator can be accurately made available to all concerned.

Additionally, the proposed communication and information infrastructure allow for new formulations and solutions to traditional power problems. In this vein, a new distributed communication model of the power system using publisher/subscriber paradigm is presented and simulated. The simulation results prove its feasibility and show it has adequate performance under today's communication technology. Based on this model, a new state estimation algorithm, which can decentralizes computations and minimizes communication overhead, is derived using a set of overlapping areas to cover the entire network. Numerical experiments show that it is efficient, robust, and has comparable accuracy as the conventional full network state estimation.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

The introduction of open markets and the fast pace of changes brought by modern information technology bring both opportunities and challenges to the electric power industry. Vast quantities of data are generated by the extensive deployment of new recording devices, such as, intelligent electronic devices (IEDs). In addition, the need for new business data, such as, market trading history, bidding information, and so on, has become important. Moreover, fast and low cost communications allow the data to be more widely accessed. Still, the legacy information systems may not make full use of the data produced. Inconsistency, inaccuracy and other problems plague the quality of the data. The large amount of data may decrease the performance of data processing in the control centers or regional transmission operators (RTOs), which are already heavily loaded. After recent serious blackouts both in the US and abroad, people have a renewed focus on reliability and the secure operation of the grid. Users need to be able to benefit from the ever increasing data. The following summarizes the various data and information system changes:

- **Increasing volume of data.** Substation automation systems and IEDs [1] are becoming locally integrated and better interconnected with the control centers. These systems can generate far more data than can be reasonably used at the control center. In addition, the introduction of energy markets produces large amounts of business data every day.

- **Greater complexity of data access requirements.** Utility users and market participants generally want to access data in a fast and timely way with the expectation of high data quality. With the free trading of the electric energy, the grid information needs to be appropriate for each of the market players while at the same time guaranteeing privacy. For example, operators need complex, detailed data in the models used for power system security applications while traders may need simple grid information indicating the main transmission bottlenecks.

- **Improved communications.** The communication network has become more efficient with bandwidth increasing exponentially. The progression of information technology, especially in the network and communication areas, makes the communication network far cheaper, more capable and more robust. This allows information exchange between different entities to be feasible and more reliable. Most utilities are only just beginning the process of moving to modern distributed communication systems but it is inevitable.

- **More decentralized computation and "intelligence". [1–3]** Modern substations have extensive computing capabilities and "intelligence" through new sensors and self-diagnostic systems. Substation automation systems are increasingly common in the grid. At the same time, the ability to process data and extract information locally is tremendously enhanced by new devices with dedicated processors. With this abundant computation power, more tasks can, and should, be done locally.

- **Renewed focus on reliability.** Following recent blackouts and new homeland security concerns, the industry has returned to placing greater emphasis on network security. The desire to make full use of the huge amount of available data to this end is apparent.

At the same, these changes are also creating new problems for the existing data and control center systems widely used throughout the power industry. This includes:

- **Legacy data systems [4]**. The legacy data system in most power utilities cannot handle the massive increase in data efficiently and properly. The traditional approach of allocating the data and related information processing into separate infrastructures within the utility does not make full use of data and information. New users want the information produced as accurately and as timely as possible while avoiding the complexity of data handling issues. The transparency of related infrastructures for collecting, processing, and communicating data and information is the key to a successful application of the data. Making the process of integrating data and extracting information fully automated while creating a major business advantage remains a challenge. To establish the value of data and information as well as to make the process of data integration and the required information extraction as efficient and as transparent as possible remain two of the major concerns in the utility industry today.

- **Inadequacy of existing security analysis tools.** Deregulation has also put more stress on the existing network security analysis tools. A recent trend is to broaden cooperation and attempt to run the power market over an even larger area in a so-called mega-RTO for better market efficiency. The energy management systems (EMS) are already overloaded struggling to cope with increasing large system models while the abundant computation power in local substation automation systems is wasted.

- **Single information consumer.** Currently, the control center is the only information consumer in the grid. The information is primarily hard-wired from the field to the control center. If the communication paths have a problem or the control center becomes overloaded, the effective system operations will cease. Nearly all the calculations are carried out in the control center, while available computational power in the grid is wasted since they lack access to the needed data.

- **Limited information for market participants.** The information that RTO makes available for the market participant is not sufficient or accurate enough for transparent trading. The constraint that the transmission grid poses on the free trading of electrical energy is a constant source of frustration to both power brokers and generation companies. The present system of posting available transfer capacities (ATC) on the OASIS system does not provide enough information for the traders to predict under what levels of transaction the system will face congestion. For such full transparency, each participant should be able to determine this availability independently. One way to achieve this is to make all control center data available to all participants, for example, the state estimator results can be posted in real-time. There are obviously privacy concerns and bandwidth problems in this approach. Still, this will certainly allow the traders to participate or verify all RTO decisions on transmission constraints but it will also require the traders to have the same level of sophistication as the operators. The investment needed in expertise and software for this level of information exchange may be unfair to the smaller brokers or generation owners.

1.2 Contributions of this Research

1.2.1 Data Warehousing

To better integrate the ever increasing amount of data, we propose data warehousing technology. Data warehousing is a technology to address the data integration problem for a large amount of data. A data warehouse is a "subject-oriented, integrated, time-varying, non-volatile collection of data that is used primarily in organizational decision making" [5]. It is a kind of database that contains consolidated data from many sources, augmented with summary data and covering a long time period. They usually are much larger than other kind of databases. Data warehousing is also a process that brings data from heterogeneous sources into an organized format that facilitate data analysis and decision making.

By using data warehousing technology, one can implement:

- efficient management of the huge amount of data available in the modern power system;

- a uniform view of the heterogeneous data, which masks the heterogeneous data source;
- a user-customizable the view appropriate for their own needs;
- data cleansing functions that reduce the amount of the inconsistency in the original data source; and
- new analytical tools for historical analysis and prediction of future trends.

In this dissertation, several examples are developed on how to construct a data warehouse that meets the requirement of the power industries. Methods to populate the data warehouse from several common power system data sources are introduced.

1.2.2 Two-stage DC State Estimation to Improve the Accuracy of the Network Topology Information

One of the main goals in introducing data warehousing was to give market participant simple useful tools to make their own decisions. One example, developed in this section, is to provide all participants with a simple DC power flow data for real-time conditions. Given that many ISOs are adopting methods based on distribution factors to make transmission decisions, the DC power flow may be accurate enough for the traders to anticipate transmission constraints and make informed decisions. On the other hand, DC power flow data can be handled with readily available off-the shelf software or easily integrated into existing trading software. This work addresses how accurate DC power flow results of real-time conditions can be made available. A DC state estimator is proposed and a method is developed to correct for topology errors. In all state estimation, a topology error, unlike an analog measurement error, can make the state estimator results useless and much research is available for topology error detection and correction for the AC state estimator. Here, a novel two-stage DC state estimator that can correct for topology errors is introduced.

The traditional full AC state estimator has many technical advantages in the detection and identification of errors. Unfortunately, it suffers from several disadvantages from a trader's viewpoint. First, it requires a large amount of data, all of which may introduce new errors or observability problems and most of which will not be directly relevant to a given trade. Second, convergence problems that often arise in practice are an unnecessary complication for the purposes of conducting transactions. Third, many of the market rules that are of concern to a trader are based on a simplified DC power flow, such as in the flowgate model, so that the resulting state estimate must be modified to be meaningful for the market.

This dissertation addresses these problems by beginning with a DC state estimator and adjusting the topology error processing for such a system view. The author suggests that traders given access to real-time data could operate such an estimator independently. The primary difficulty is in the inherent errors in the DC model that limit topology error processing. In the proposed approach, state estimation is performed at the bus/branch level. If any errors are detected, the suspect area is expanded into a bus-section/switching-device model. Then the state estimate is repeated over this expanded model. A new method is proposed that more effectively distinguishes between modeling approximation errors and data errors.

There exists extensive literature that addresses the topology error identification problem. All the approaches use full AC state estimation. In the proposed method, the modeling error is estimated in order to compensate for the inherent inaccuracy of the DC method, while maintaining the advantages of robustness and efficiency. Further, these linear computations are more appropriate from a market viewpoint. Calculation proves the feasibility of the proposed method.

1.2.3 Distributed Information Infrastructure and Distributed State Estimation

Since local processors are becoming more and more common, using distributed processing can ease the burdens on the control center and provide better performance and more reliability. Instead of a centralized control and processing model of the existing power system information structure, a distributed information processing model that using the publisher/subscriber paradigm is introduced. In this model, each entity can be an information publisher or subscriber or both and the relations between different power system entities are peer to peer relation instead of master-slave hierarchical relation in the traditional power systems. Each entity, such as a substation, can publish its data, for example bus voltages and breaker status. At the same time, it can be an information subscriber for information published from other entities. A traditional control center can viewed as a large information subscriber that consumes the information published by substations, power plants, and so on, throughout the network.

By using this model, one can achieve:

- greater flexibility, since the system can be easily reconstructed;
- increased reliability, since the information transfer is more due to the possibility of re-routing packets unlike in the existing hard-wired network; and

- more evenly distributed computational load as the control center is not the only entity that can receive the necessary information - any subscriber can have the data it needs to perform computation locally.

In order to validate the distributed information infrastructure, simulations are performed on the IEEE testing systems. The simulations revealed that the new information structure is feasible and has acceptable performance.

State estimation is a major analysis tool that helps operators run the system securely and efficiently. Based on the proposed information structures, a complete distributed state estimation that can make full use of the abundant computation resource in the power grid while easing the burden of the control center is introduced.

The idea of this distributed state estimation is to divide the network into overlapping areas. The whole network is covered by the union of the overlapping areas, that is, any bus or branch in the network is inside at least one of the areas. Each area also has some overlap with the some of the rest of the areas. The connections between areas are assumed to form a connecting graph for the network.

In each area, one local computation center subscribes to all the necessary data inside the area, and carries out the conventional state estimation for that area. The local state estimation result is passed to a control center and the whole network result is assembled in the control center. This approach complete decouples the state estimation calculation inside each area, that is, the result of one area does not depend on the result of another area. This can maximize the parallelism and minimize the communication cost between the areas. This approach is shown feasible based on the simulation of the communication network. Today's network and communication are able to provide an adequate communication infrastructure.

One of the advantages of this approach is that it can avoid the so called large system problem. Since as the network is becoming larger and larger, convergence, numerical stability and error data identification can be a big problem for the state estimation calculation. This approach differs from most of the current approaches to distributed/parallel state estimation. Almost all the current approaches are based on the conventional information structure and formulate the distributed state estimation as an optimization problem for each area and use the equality of the voltage angle and magnitude on border buses from different areas as constraints. Those approaches have one thing in common, each iteration of one area, or more areas, depends on the result of another area. This impairs parallelism, since some calculation may stall while waiting for the result of other iterations. Numerical examples reveal that this approach is feasible and has nearly the same accuracy as traditional methods.

1.3 Organization of the Dissertation

In the following chapters, each individual approach is discussed in great detail. Chapter 2 addresses the data integration problem by using data warehousing technology. In Chapter 3, the two stage DC state estimation is developed to address the topology error identification problem. Chapter 4 presents the distributed information infrastructure followed by Chapter 5 focuses on the proposed distributed state estimation algorithm. The result of simulations on different IEEE testing systems and the test case of the distributed state estimation are presented and analyzed.

Chapter 2

Data Integration using Data Warehousing Technology in Power System

2.1 Introduction

Information in the power system plays a very important role. With the introduction of deregulation and IEDs, data in the power system are changing in both quantity and functionality. Utilities previously focused on the controlling data that reflected the grid status. Today, such data is exploding due to the increasing number of IEDs used in the field. Moreover, energy markets require the utilities to face greater competition and the proper management of vital business data is often the key to the success. In traditional power utilities, data are isolated in different departments and frequently inconsistent. As it becomes clearer that the data and information are one of their most important assets, information integration is becoming a vital area for the utilities to gain a competitive edge, provide better service to customers, gain market share and increase profits.

Data integration and data warehousing are not new to the power system. Gillerman [6] pointed out the problem of data duplication and isolation. His answer to the problem is the Common Information Model (CIM) data model. But the CIM data model itself is not a very good model for historical data. Werner [7] proposed very good requirements for what a data warehouse for power system should do. His point of view is more from how to handle historical data in the Supervisory Control and Data Acquisition (SCADA) system or EMS efficiently. His introductory paper [8] talked about the concept of data warehousing but fell short of implementation and related data model. Sundhararajan [9] reports using data warehousing in ERCOT for business history data.

In this work, we introduce data integration by using data warehousing technology. First, we present the challenge power utilities face today: the isolated parts of the data models/databases and ever increasing volume of historical data. We then introduce the multi-dimensional modeling that is widely used in Online Analytical Processing (OLAP) followed by the methods to construct data warehousing. Possible applications of the data warehousing are discussed.

2.2 Existing Problems

Two main data problems exist in many power companies:

- There is numerous duplicate and inaccurate data residing in different systems isolated across several departments. This arises from different user needs as well as historical reasons.

- There is a rapidly growing amount of data available due to the recent introduction of IEDs in the field, expanding business data from deregulation and increasing data exchange between utilities.

2.2.1 Data Duplication and Inconsistency

There are “sound” historical reasons for separate databases for purchasing, planning, operations, maintenance purposes, and so on. Different data users will have varying needs and views of the data. Further, the fact that there are multiple entries for a single data element across those databases gives rise to the possibility that across the enterprise incorrect data entries exceed the number of correct data entries. There are three main data considerations then:

- Duplication: similar or identical data may exist in different departments.
- Inconsistency: Since different departments place varying degrees of importance on particular data, non-critical data for some department may be updated infrequently or without proper verification, so that such data is unreliable.
- Query construction: Users may have trouble constructing valid queries even if they know where they can find valid data since the databases in different departments may have different schema.

As an illustration, consider the perspectives of several users for transmission line data.

Operation Engineer

The operation engineer is usually a user of the SCADA/EMS system, which requires real-time data from the field. In order to carry out security analysis, the EMS system requires resistance, reactance, short-term and long-term flow limits, the statuses of the breakers/switches at the ends of the line, and so on. Operation schedules of that line may also be stored. Since EMS/SCADA systems have a relatively long history and many different vendors, there are various database models. These include hierarchical, network and some relational models with each vendor developing their own schemas. Thus, these proprietary databases usually have the following characteristics:

- Describe the system with a bus-breaker model. The database has the detailed physical parameters for the power system analysis, but lacks data for other purposes, such as, cost, maintenance cost, physical dimensions, and so on.
- Employ heterogeneous schema depending on the various vendors.
- Emphasize real-time fast access so they typically use hierarchical or network models. Even if a relational database is used, it is usually not normalized.

Planning Engineer

Planning engineers usually select the type of the transmission line and estimate the resistance and reactance for an approximate corridor of the transmission line. They perform calculations to ensure the new line will meet security and other requirements. They also need cost information of the transmission line to complete their analysis. But after the line construction, the data will be duplicated and refined in other department databases and the planning data will be stale.

Maintenance Engineer

Maintenance engineer data requirements include operation schedule of the line, exact location of the towers, tree clearance information, maintenance history, testing data, and so on. Usually, these engineers are more concerned with geographical and related data of the line and operation schedules.

Market Participant

Market participants as business professionals are generally not interested in the detailed physical parameters of the transmission line. Still, they are concerned with reaching correct business decisions, such as, the operation schedule of the transmission line or contribution to Available Transmission Capacity(ATC). Databases used in power markets are usually commercial relational databases, such as Oracle, DB2, and SQL Server. This kind of database differs significantly from those used in real-time SCADA systems. This is simply because the short history of, and the commercial nature of, the power market.

The following are the different characteristics between the power market and the EMS/SCADA system databases. The power market databases:

- mostly build upon the relation model,
- have much higher requirements on error recovery and transaction control,
- reflect more stringent security requirements, and
- access data primarily via canned queries.

Clearly, a single transmission line data object has multiple instances and resides in different databases with multiple views. These problems become more troublesome since there may not be a uniform schema for users to apply even if they know the location of accurate data.

2.2.2 Increasing Volume of Historical Data

As with other industries, power utilities have a need to record their business process as well as the operation of the power system. With the introduction of IEDs, the volume of data is rapidly increasing. Not surprisingly, most EMS' have functions to store historical data in order to create reports and perform subsequent case studies. Still, most recording capabilities have the following limitations:

- Many use proprietary database or files and are not easily integrated with database standards for field devices. As a result, there are no standard tools, e.g., SQL, to access historical data. This situation limits further enhancement to the post event analyses.
- The systems are not flexible. Typically, the data schema is fixed and the user can only select from a set of canned queries.
- Most are dedicated real-time systems so the performance of the built-in historical functions is not good since those functions may slow down the real-time SCADA/EMS system performance.
- There are almost no historical data analysis tools with the system since SCADA/EMS is dedicated for real-time operations.
- It is only accessible by operation engineers.

The problem that arises in this situation is how can, for example, a planning engineer not in operation department, easily access the historical data in order to make reasonable plan for future expansion.

2.3 Information Integration

2.3.1 Two Approaches

The answer to the above problems is by using appropriate data integration technologies. Providing integrated access to multiple, distributed, heterogeneous database and other data sources has become one of the leading issues in database research and industry. There are two methodologies available to address the data integration problem. One is called lazy approach [10]:

1. Accept a query, determine the appropriate set of information sources to answer the query, and generate the appropriate subqueries or commands for each information source.

2. Obtain results from the information sources, perform appropriate translation, filtering, and merging of the information, and return the final answer to the user or application (hereafter called the client).

We refer to this process as a lazy or on-demand approach to data integration, since information is extracted from the sources only when queries are posed. (This process also may be referred to as a mediated approach, since the module that decomposes queries and combines results often is referred to as a mediator [11]).

The natural alternative to a lazy approach is an eager or in advance approach to data integration. In an eager approach:

- Information from each source that may be of interest is extracted in advance, translated and filtered as appropriate, merged with relevant information from other sources, and stored in a (logically) centralized repository.

- When a query is posed, the query is evaluated directly at the repository, without accessing the original information sources.

This approach is commonly referred to as data warehousing, since the repository serves as a warehouse storing the data of interest.

A lazy approach to integration is appropriate for information that changes rapidly, for clients with unpredictable needs, and for queries that operate over vast amounts of data from very large numbers of information sources. However, the lazy approach may incur inefficiency and delay in query processing, especially when queries are issued multiple times, when information sources are slow, expensive, or periodically unavailable, and when significant processing is required for the translation, filtering, and merging steps. In cases where information sources do not permit ad-hoc queries, the lazy approach is simply not feasible.

In the warehousing approach, the integrated information is available for immediate querying and analysis by clients. Thus, the warehousing approach is appropriate for [10]:

- clients requiring specific, predictable portions of the available information.
- clients requiring high query performance (the data is available locally at the warehouse), but not necessarily requiring the most recent state of the information.
- environments in which native applications at the information sources require high performance (large multi-source queries are executed at the warehouse

instead), and

- clients wanting access to private copies of the information so that it can be modified, annotated, summarized, and so on, or clients wanting to save information that is not maintained at the sources (such as historical information).

The lazy and warehousing approaches are each viable solutions to the data integration problem, and each is appropriate for certain scenarios. In this study we address the existing data problems in power companies using a warehousing approach.

2.3.2 Data Warehousing

A data warehouse is a “subject-oriented, integrated, time-varying, non-volatile collection of data that is used primarily in organizational decision making” [9]. It is a database that contains consolidated data from many sources, augmented with summary data and covering a long period of time. They usually are much larger than other databases. Data warehousing is a process that brings data from heterogeneous sources into an organized format that facilitates data analysis and decision making. Such a database promises to provide uniform and user-friendly access to all historical data, flexible enough to provide powerful data mining capability and powerful enough to handle the information pressure caused by the huge amount of inflow data from the power grid and energy market.

Difference between Data Warehouse and the Databases used in EMS/SCADA and Market Transaction

The databases used in EMS and power markets are operational systems that either reflect the change of the power grid in real-time or capture the transaction of the trade business. The main priority of these systems is processing performance and availability. Queries against these systems are usually narrow, one-record-at-a-time. The data warehouse does not directly record the daily transaction/operational data. Instead, it is loaded for analytic usage. The reasons for having a different data base are two-fold:

- Performance. Data warehouse has different characteristics that require quite different implementation and tuning from the production/operational database. For example, special tuning for OLAP and data mining usage.
- Data quality. Data inside the data warehouse is clean and uniform for the whole enterprise, instead of different forms and duplication for different department and applications.

Data warehouse represents the merging of the traditional administrative data warehouse and the historical data from the online EMS. Therefore, the entire company can access the same data warehouse, which greatly reduces the data duplication and maintenance cost.

2.3.3 Data Schema Used in Data Warehouse

Multi-Dimensional Database and Star Schema

The dominant conceptual data model for the data warehouse is a multidimensional model. In the multidimensional data model, the focus is on a collection of numeric measures that are the objects of analysis, such as energy consumption, bills, revenue, cost, etc. Each measure depends on a set of dimensions. Consider as an example the monthly bill for a customer. The numerical measures might be the total consumption (kwH) the customer used and the total energy charge. The

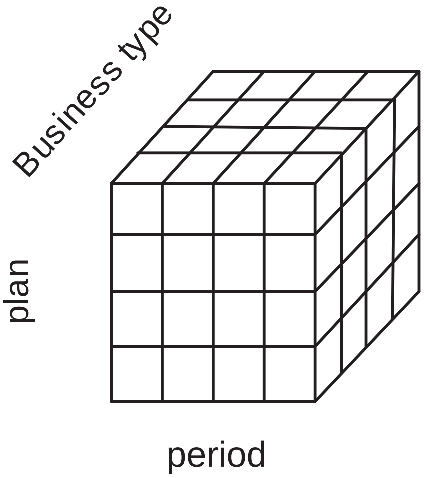


Figure 2.1: Multi-Dimensional Model

dimensions associated with that price are customer name, address, business type, pricing plan, billing period, average temperature, and so on. Together these dimensions are assumed to uniquely determine the measure. Thus, the multidimensional data views a measure as a value in multidimensional space. Each dimension is described by a set of attributes. The attributes of a dimension may be related via a hierarchy of relationships. In the above example, the street address is related to the city and county through such a hierarchical relationship. The business type attribute has a similar hierarchical relationship.

The distinctive feature of the model is the stress on aggregation by one or more dimensions as one of the key operators, for example, calculating the total consumption of residential customers for a certain period. Another comparison could be comparing the two different measures such as revenue and cost.

For multi-dimensional data, other operators include rollup and drill-down, slice and dice, etc. Rollup corresponds to taking the current data object and doing a further grouping on one of the dimensions. Thus, it is possible to roll-up the energy consumption data, perhaps already aggregated on time, additionally by business type. The drill-down operation is the converse of rollup. Slice and dice corresponds to reducing the dimensionality of the data, i.e., taking a projection of the data on a subset of dimensions for selected values of the other dimensions. For example, we can slice and dice sales data for a specific product to create a table that consists of the dimensions plan and the period of bills. The other popular operators include ranking (sorting), selections and defining computed attributes [12].

Entity Relationship diagrams and normalization techniques are popularly used for database design in OLTP environments. However, the database designs recommended by ER diagrams are inappropriate for decision support systems where efficiency in querying and in loading data (including incremental loads) are important. Most data warehouses use a *star schema* to represent the multidimensional data model. The database consists of a single fact table and a single table for each dimension. Each tuple in the fact table consists of a pointer (foreign key, often uses a generated key for efficiency) to each of the dimensions that provide its multidimensional coordinates, and stores the numeric measures for those coordinates. Each dimension table consists of columns that correspond to attributes of the dimension.

Design Examples

Here, we provide some design examples on multidimensional modeling for power companies. The following is a dimensional model for measurement data. For measurement data, the fact table (numerical value of interest) is obviously the value of the meter reading. The dimensional data for that includes: device information on which the meter is measured, the RTU information, the measurement value limits, and temporal information, such as, date and precise measurement time. Figure 2.2 shows the associated star schema.

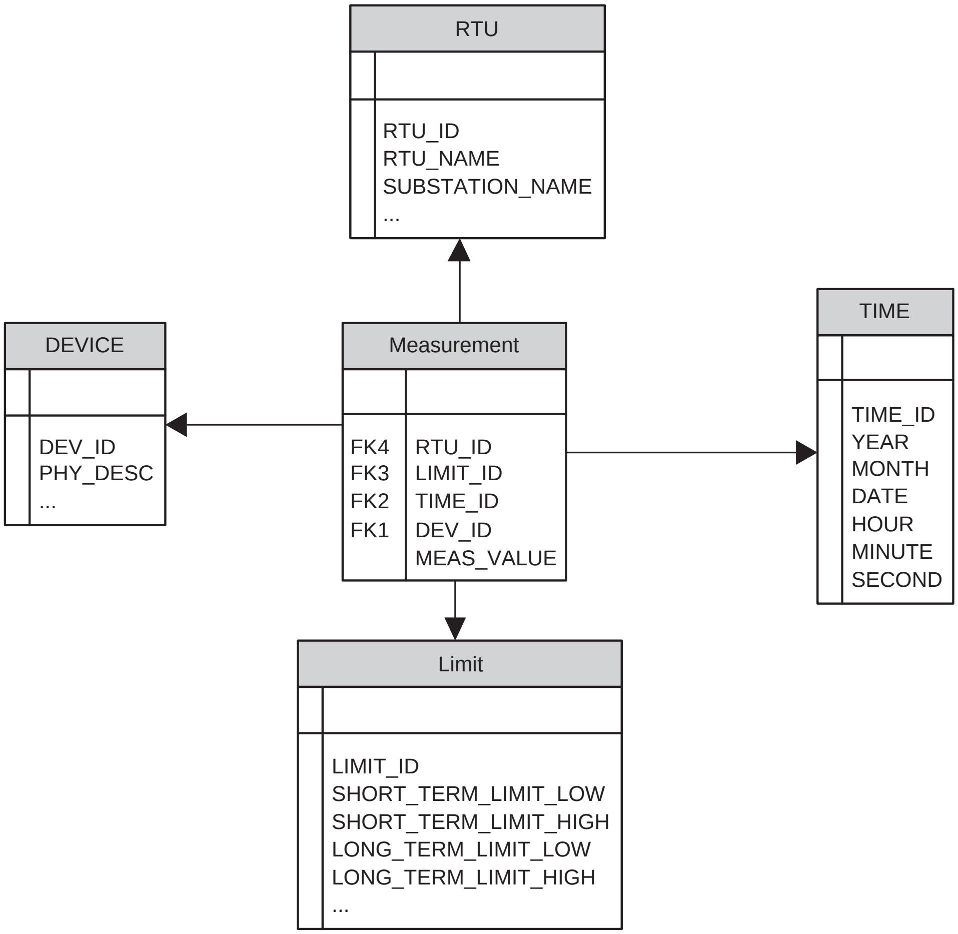


Figure 2.2: Star Schema for Measurement Data

The marketer also needs a database that holds historical data that can help them make proper decision during the bidding process. They can somewhat predicate the trend by carefully examining the historical data. In the power market, one can use the bid price as the fact table. The related dimension could be date and time, fuel price at that time, weather info, and market type. The resulting star schema is shown in Figure 2.3.

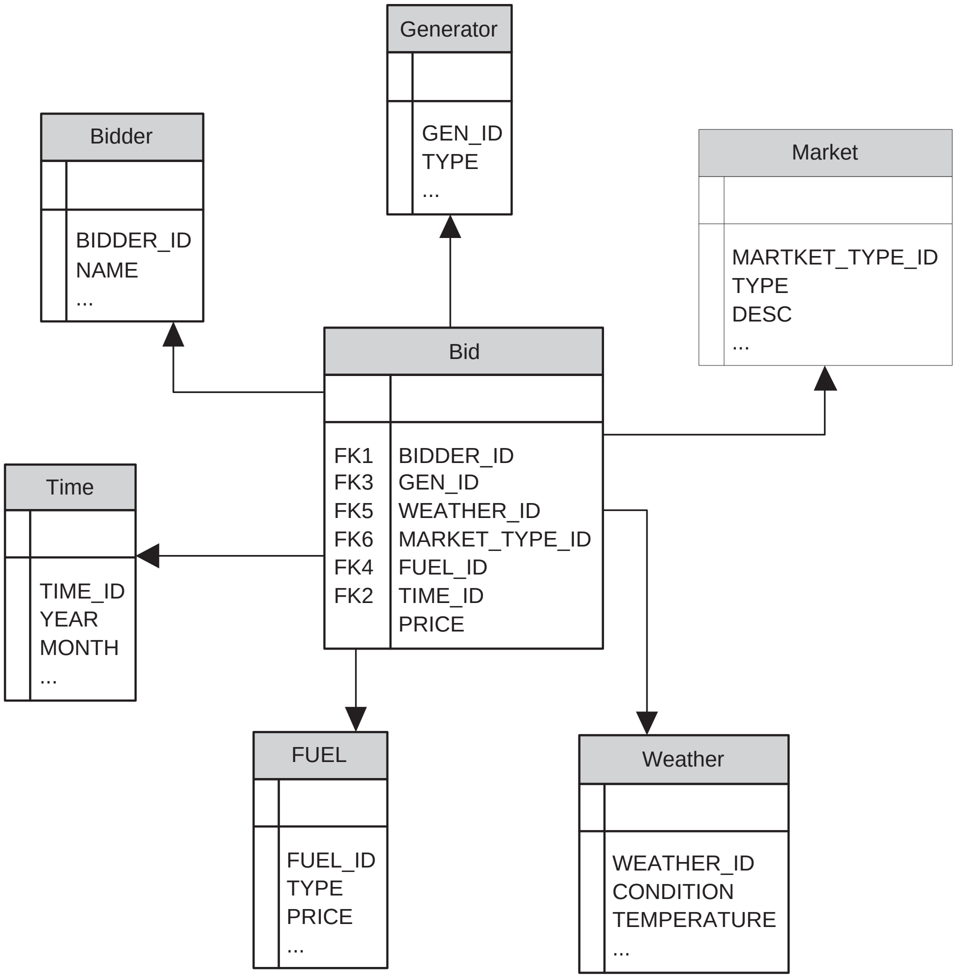


Figure 2.3: Star Schema for Marketer

For the utility, a fact table could be the monthly usage of a customer; the related dimension table could be the name and address of the customer, the account number, the month and year, the average of the temperature of that period and so on.

CIM and Multi-Dimensional Data Model

The CIM is a data model developed by the Control Center Application Program Interface (CCAPI) task force of Electric Power Research Institute (EPRI). It provides a modern view of the power system of the entities in a grid and provides a vehicle for control centers to exchange data. However, most EMS vendors still use a proprietary database in most applications and only use CIM as an information exchange vehicle. The main reason for that is, of course, the vendor does not have any financial motivation to replace their current database with the CIM schema.

The CIM data model is more suitable for the operational database. For data warehousing, users are more interested in aggregate values and the CIM data model itself does not do a good job in recording temporal information. The CIM tends to view the system at particular time instants and lacks historical support. Another reason the CIM is not a good candidate for the data warehousing schema is that it is too normalized and may have bad performance when dealing with large amounts of historical data. Normalization functions to reduce update and insertion abnormality, which are not serious problems for the data warehouse.

2.3.4 Building Data Warehousing

Data conversion for large amounts of data usually involves extract, transform and load (ETL) phases. We discuss the data conversion according to those phases followed by some examples to explain the process in detail.

Extract

The raw data coming from the source systems is read into tables or loaded into memory depending on the size. These tables or memory arrays are called the staging area. It is the working space for ETL. There is minimal restructuring with no significant content transformation.

Cleaning

In some cases, the data quality and format in the data source and in the data target are different. For example, some data has incorrect format or some constraints may need to be enforced in the data or the naming convention is incompatible and so on.

Transform

The transform work is carried out in the staging area. The data is reorganized and recalculated. For example, if the power system data describe the grid on a bus-breaker level and the target is on node-branch model, a topology processor is needed to construct the network and calculate certain parameters.

Loading

The data is loaded into target tables from staging area.

Performance Considerations

Data warehouses may contain extremely large volumes of data. To answer queries efficiently, therefore, requires highly efficient access methods and query processing techniques. Several issues arise. First, data warehouses use redundant structures such as indices and materialized views. Choosing which indices to build and which views to materialize is an important physical design problem. The next challenge is to effectively use the existing indices and materialized views to answer queries. Optimization of complex queries is another challenging problem. Also, while for data-selective queries, efficient index scans may be very effective, data-intensive queries need the use of sequential scans. Thus, improving the efficiency of scans is important. Finally, parallelism needs to be exploited to reduce query response times.

Performance is a critical issue for realtime power system operations and controls. Parallelism plays a significant role in processing massive data. All major vendors of database management systems now offer data partitioning and parallel query processing technology. One technique relevant to the read-only environment of decision support systems is that of piggybacking scans requested by multiple queries. Piggybacking scan is a technology that overlaps scans of multiple concurrent requests. Therefore, it can reduce the total work as well as response time.

The general way to speed up data conversion is to exploit parallelism in the conversion process. There are two common parallelisms: pipelined parallelism and partitioned parallelism [13]. Pipelined parallelism is achieved by streaming the output of one operator into the input of another operator. Therefore, the two operators can work in series. Partitioned parallelism is achieved by partitioning the incoming data among available operators so that each operator can work on a part of the data.

2.3.5 Typical Data Conversion Scenario

Translate from CIM into target data warehousing may be very common. For example in [14], the authors emphasized on the level of modeling (bus-breaker and node-branch), here we will address the process from ETL point of view.

From CIM Data to Data Warehouse

Translating CIM data efficiently into other data models is not a trivial task. Many of the existing data conversion programs require several hours to load the database from CIM format into its proprietary database. Currently, a typical database may have a quarter of million measurement points, several hundred generators, and close to ten thousand lines. The time constraint is tight for the data conversion. We provide some suggestion on each stage of the ETL process:

Extract and Clean

Extract and clean phase in ETL means successfully extracting the data, cleaning the inconsistency and irregularities in the data and arranging them properly in staging area for future transformation. For a small amount of data, we recommend reading data directly into memory and building a proper index. For a large amount of data, it would be best to first load data into a temporary staging database table. The CIM data can exist in many formats, as a flat text file, a database image or Extensible Markup Language (XML) format. With the widespread familiarity of XML, more and more data are encoded in this format. Therefore, there are several scenarios in extracting data from the CIM model, depending on the format of the data. The following sections, we discuss managing each CIM data format respectively.

- **Text file**

  Plain text file is one of the most common ways to exchange data. It can be read directly into memory or load into database, since most of commercial databases have bulk loading utility. Another data cleaning program may be needed to make sure the each data field is correct in both format and semantic meaning.

- **Database Table**

  The CIM data already exists in database tables. It should consider relative clean and we just need extract the table to the staging area for further process. Typically, a ODBC/JDBC is needed in this case, or just plain SQL if the transform database is the same as the source database.

- **XML**

  XML itself is little more than a tagged text file, but has advantages of self-descriptiveness and flexibility. Right now, it is gaining the data exchanging market by swiping speed. XML is good in several ways, but performance wise, it is not that awesome. Its native structure makes it difficult to exploit parallelism. Usually, there are two existing ways to parse XML data, Simple API for XML (SAX) and Document Object Model (DOM). For the size of the CIM data for a typical power grid, it is not recommended to use DOM, since it expands all the elements instantiated in the memory. It will be awkward and inefficient for the system to handle that amount of data. If the target data model is not at the same level as the CIM model (bus-breaker), then a topology processor is needed. In this situation, we need random access XML data. DOM is much better than SAX at random access, but given the size of the data itself, DOM is still not a good choice due to its large memory footprint. We load XML data into database tables first and run the topology processor on top of those tables. Even if the target data model and the CIM are at the same level, it is still better to utilize a database since the schema are very likely to be different and a database is far more superior in the operations, such as, join and lookup.

Data warehouse and CIM data may have different naming convention. And the limit constraint are different or model used (such as transformer, etc) can be different. This needs to be addressed in the data cleaning stage. 2.1 summarizes the different ways to manage different CIM data formats.

Table 2.1: Different ways to deal with different CIM data format

	Data Amount is Small	Data Amount is Large
Plain Text	Build table (array) with index in memory	Load into database and build index
Relation Table	Read them into memory	Use JDBS/ODBC or plain SQL
XML	Build table with index in memory, using either DOM or SAX	Load into database using bulk loader or hand coded using SAX

Data Transform

As mentioned before, CIM describes the power system at the bus-breaker level. However, there is no consensus on the level the target data models describes. EMS/SCADA usually models the power system at the same level as the CIM, while analysis packages, such as PSS/E, view the power system at the node-branch level. In this situation, a topology program is needed to construct the network topology.

There can be two scenarios in the translation,

- CIM data can be extracted, transformed and loaded into a proprietary database by using only relational calculations, if the target data model and the CIM are modeling the power system on the same level (bus-breaker level). In this case, one does not need to construct the network to convert the data. Further, we can leverage the existing database functionality as it not necessary to develop new data extract and transform functions. Currently, commercial relational databases, such as, Oracle, DB2, and Microsoft SQLServer have very good performance in performing relational calculations.
- The conversion cannot be completed by ordinary relational operation. This occurs because either the CIM or target data model are not modeling at the same level. In this case, there is to choice but to write a conversion program, which is similar to a typical topology processor.

Specific Data Transform/Cleanse Applications

Beside business information such as those data from consumer support, energy market, etc, the power system has a large amount of data from the underlying physical system. However, if data warehousing is initially made for business data, it needs to be adapted to work well for the power system.

Topology Processing A typical example is that the granularity of the data stored in the warehouse is at bus-breaker level. This is the finest granularity of data, which conform to the rule that a data warehouse should store the lowest possible grain of data [15]. System operators, of course, need the data in that level. But not all the data warehouse users want data in that detailed level. Planning engineer often want the data in the node-branch level that can easily been analyzed by the load flow, contingency analysis programs. Traders in the power market may want data in even coarser granularity. They may only want the system to be represented in a simplified equivalent version, which has all the important transmission lines but other things as few as possible. All though the data granularity is becoming more and more coarse, the operation to get less fine grain data is a simple aggregation operation mentioned above. In this case, a network topology processor is needed. By using a topology processor, the user can actually roll-up and drill-down the data.

State Estimation Another example of specific operation on the data warehousing for power system is state estimation. Data in the bus-breaker level may contain errors such as measurement errors and communication glitches. It is common that data should be cleaned before they are loaded into the target database. From the data quality point of view, the state estimation acts as a data cleansing program in the ETL process. On one hand, the state estimation cleans the analog data by filter out the measurement noise, on the other hand, state estimation can also be used to identify the topology error existed in the raw data. It is important to have good data since some department in the power utilities such as planning need sensible data to carry out their calculation. Bad data will make their analysis tools had to converge or have wrong conclusion. Author presents a two stage DC that can be used as part of the hand-coded data cleansing program [16].

Data Loading

After data is transformed in the staging area, they can be loaded into target data warehousing tables by using either pure SQL or bulk loading utility. After this, the data is ready for user's query.

2.3.6 Benefits of the Data Warehousing

Since data warehousing is one of the key technologies for data integration, it can provide an accurate and uniformed view of the data throughout the enterprise and provide a feasible way to minimize the data duplication problem without overhauling the entire enterprise's data system. Since the huge amount of data stored in a data warehouse and data inside are organized in a flexible way can suitable for a kinds of analysis. Given the real-time nature of power system data, this prevents users from falling into the habit of viewing the power system at isolated time instants while forgetting the historical data. Data warehousing enables one to explore the history, find patterns and make projections into the future. One of the greatest benefits of the data warehousing is that it can easily access the historical data in the time dimension. Users can take advantage of this and analyze the data over a long time span. Following the earlier example, the benefits for are outlined below.

For generation company

A power generation company can store the actual generated power together with the contract price it was sold. By examining the earning data over years, it can re-evaluate its market strategies to see if it is really successful. Another usage is, by integration the huge amount of power flow data over the transmission lines at a series of very close time instants, one can precisely calculate the energy flow through for market clearing and settlement purpose.

For planning engineer

Data warehousing makes it possible to store and analysis a huge amount of historical operational data without slowing real-time SCADA/EMS operations down. This makes it possible to evaluate, for example, the calculated transmission capacities against the actual power transmitted through the network as well as seasonal load patterns. By studying the pattern and trend, planning engineer can decide the long term plan of the network expansion, optimize the power transmission contracts, and so on. Since the huge amount of data is already loaded into the data warehouse and the planning engineers can easily retrieve a snapshot of the grid at a particular instance. This gives planning engineers much more freedom and more accurate result can be expected. This is difficult to do without the support of the data warehouse, since the data may exist in heterogeneous sources and extracting data without slowing down the SCADA/EMS is difficult.

For maintenance engineer

Historical data provides significant information for maintenance. For example, they can identify the pending meter failures by examining the history of that meter. If they can find a frequent abnormal reading from the meter, there could be a high likelihood of failure.

For market participant

The data warehousing allows market participants to have the ability to forecast future conditions by looking at the history. For example, they can find patterns of fuel prices along with the electricity price, providing a competitive edge in bidding.

For customer service

Customer service is of growing importance with wholesale and retail choices. It is possible to investigate the common characteristics on how customers react to different programs and preferences. This will help utilities to design pricing.

2.4 Conclusion

Due to historical reasons, data in power utilities are isolated in different departments resulting in data duplication, numerous errors and a lack of uniform schema. With deregulation and the rapid introduction of IEDs, utilities must begin handling an ever increasing amount of data effectively. In this work, we propose data warehousing technology to solve the data integration problem. Data warehousing can provide a clean and uniform view of the data. More important, it provides a good platform for possible data analysis and data mining. We propose new system applications made possible by this structure. Power system operations has a tendency to focus only on instantaneous data. The data warehousing approach provides engineers the opportunity to explore the historical data and reach better operational decisions.

Chapter 3

Topology Error Identification using a Two-Stage DC State Estimator

One of the fundamental tenets in deregulation of the power system is to provide fair and open access to transmission facilities. This requires that market participants, both power brokers and generation companies, have complete and timely information as to the transmission availability. The present system of posting available transmission capacities (ATC) is useful but limited because there is no information to predict how these ATCs will change with changing power transfers. It has been proposed that all traders have access to the real-time data of the full transmission model, i.e. state estimator results from the control centers, but this may be too complex and voluminous to be useful to the traders. Instead, making DC power flow data for real-time conditions may provide enough transmission data for traders to make knowledgeable decisions. In this chapter, we show how the results of a DC state estimator can be accurately made available to all concerned.

3.1 Introduction

The Constraint that the transmission grid poses on the free trading of electrical energy is a constant source of frustration to both power brokers and generation companies. The present system of posting available transfer capacities (ATC) on the OASIS system does not provide enough information for the traders to predict under what levels of transaction the system will face congestion. For such full *transparency*, each participant should be able to determine this availability independently. One way to achieve this is to make available to all participants the state estimator results that are available to the ISO/RTO. Although this will certainly allow the traders to participate or verify all ISO decisions on transmission constraints, it will also require the traders to have the same level of sophisticated software tools as the ISOs to do so. The investment needed in expertise and software for this level of information exchange may be unfair to the smaller brokers or generation owners.

An alternative suggestion is to provide all participants with DC power flow data for real-time conditions. Given that many ISOs are adopting methods based on distribution factors to make transmission decisions, the DC power flow may be accurate enough for the traders to anticipate transmission constraints and make informed decisions. On the other hand, DC power flow data can be handled with readily available off-the shelf software or easily integrated into existing trading software. This paper addresses how accurate DC power flow results of real-time conditions can be made available. A DC state estimator is proposed and a method is developed to correct for topology errors. In all state estimation, a topology error, unlike an analog measurement error, can make the state estimator results useless and much research is available for topology error detection and correction for the AC state estimator. We present a novel two-stage DC state estimator that can correct for topology errors.

The traditional full AC state estimator has many technical advantages in the detection and identification of errors. Unfortunately, it suffers from several disadvantages from a trader's viewpoint. First, it requires a large amount of data, all of which may introduce new errors or observability problems and most of which will not be directly relevant to a given trade. Second, convergence problems that often arise in practice are an unnecessary complication for the purposes of conducting transactions. Third, many of the market rules that are of concern to a trader are based on a simplified DC power flow, such as in the flowgate model [17], so that the resulting state estimate must be modified to be meaningful for the market.

This chapter addresses these problems by beginning with a DC state estimator and adjusting the topology error processing for such a system view. The author suggests that traders given access to real-time data could operate such an estimator independently. The primary difficulty is in the inherent errors in the DC model that limit topology error processing. In the proposed approach, state estimation is performed at the bus/branch level. If any errors are detected, the suspect area is expanded into a bus-section/switching-device model. Then the state estimate is repeated over this expanded model. A new method is proposed that more effectively distinguishes between modeling approximation errors and data errors.

There exists extensive literature that addresses the topology error identification problem. Monticelli [18–20] used a physical level model and modeled the zero-impedance branch by its power flow. By assessing this flow, one can tell the status of the zero-impedance branch. However, pinpointing the suspicious area is crucial, otherwise the method suffers computationally. Liu and Wu [21, 22] and Clements [23] both modeled the topology error as a change in the measurement matrix, and subsequently identify the error analytically. Lugtu [24] used residual analysis and empirical judgment to determine the topology error. Abur [25] has proposed a two-stage method similar to our approach. All of the above approaches use full AC state estimation. In the proposed method, the modeling error is estimated in order to compensate for the inherent inaccuracy of the DC method, while maintaining the advantages of robustness and efficiency. Further, these linear computations are more appropriate from a market viewpoint.

3.2 Background

3.2.1 DC State Estimation

In this section, we first review the classical formulation of the state estimation and a liberalized version of it. A model is introduced that includes the topology errors and model (DC) error. The nonlinear equations relating the measurements z and the state vector x are:

$$z = h(x_{true}) + \epsilon \tag{3.1}$$

where ϵ is measurement error vector with zero mean and covariance matrix R. With a linear model of the power system bus angle θ_i as X_{true}, real power injection P_i and line flow P_{ij} as measurement z, 3.1 simplifies to:

$$z = Hx_{true} + \epsilon \tag{3.2}$$

where H is the Jacobian matrix.

The state estimation problem usually is formulated as a least square problem (WLS) to minimize:

$$J(x) = (z - Hx)^T W(z - Hx) \tag{3.3}$$

where W is a weighting matrix (inverse of the covariance matrix) and x is the state vector.

3.2.2 Error Modeling

There are three different types of errors in the DC model:

- **Measurement error** Measurement error can be modeled as a zero mean with non-zero covariance, assuming no gross error exists.
- **Model error** Since the linear model is only an approximation to the real system model, another error is added to the linear model:

  $$z = Hx_{true} + \gamma + \epsilon \tag{3.4}$$

  where γ is the model error vector, i.e., the difference between the accurate model and DC model. It is not a random but a deterministic value that depends on the current state of the system.
- **Topology error** If topology errors exist, the model becomes:

$$z = (H + dH)x_{true} + \gamma + \epsilon \tag{3.5}$$

where, dH is introduced by the incorrect topology. There are several kinds of topology errors of interest here.

Branch Outage and Addition **-** Incorrect information on the breaker of the line/branch will cause false line outage condition. For example, the false line $i - j$ outage will result in errors on:

- Branch flow measurement error on the row regarding the line measurement.
- Injection measurement error on the row of i and j.

These can be modeled as additional or reduced line flow from the corresponding line or node [22, 23]. Assume there is an error in the outage in line $i - j$ as illustrated in Figure 3.1.

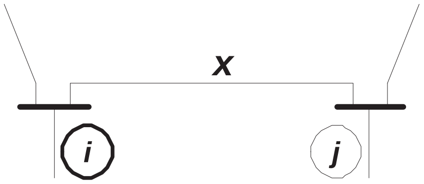


Figure 3.1: Line $i - j$ with impedance x

For a line flow measurement, the corresponding row of H matrix related to the flow measurement on line $i - j$ has the following change:

$$\begin{array}{ccccccc} [& \cdots & -1/x & \cdots & 1/x & \cdots &] \\ & & i & & j & & \end{array}$$

For an injection measurement, the corresponding row of the H matrix related to the injection measurement on node i changes by

$$\begin{array}{ccccccc} [& \cdots & -1/x & \cdots & 1/x & \cdots &] \\ & & i & & j & & \end{array}$$

and similarly for the row of the H matrix related to the injection measurement on node j.

Bus Split - False breaker status can result in a different configuration of the substation. The bus split can also be modeled as multiple line outage [21, 22].

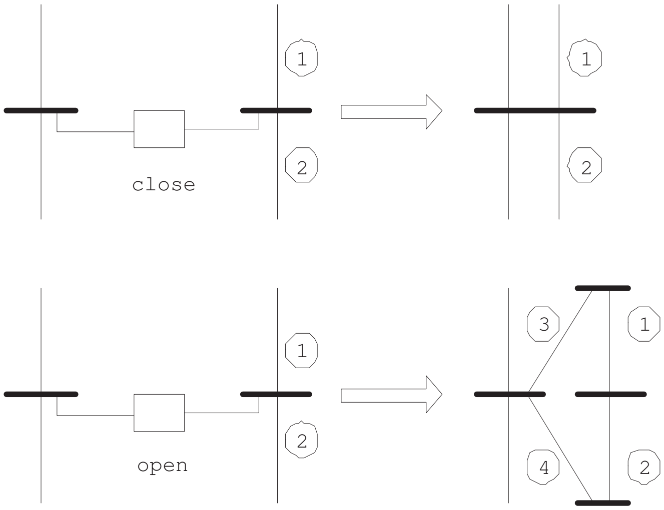


Figure 3.2: Modeling bus split as line outage and addition

3.3 Proposed Approach

The approach to topology error identification is using two-stage state estimation. The first stage is using state estimation on bus/branch model, if suspect area detected, this area is converted to detailed bus section/switching device model, then a second stage generalized state estimation is used on the mixed model to to identify the topology error. DC state estimation has an advantage over AC here since the DC estimator can greatly reduce the calculation burden on the mixed model and only moderate accuracy is needed at this stage. The primary concerns are: one, the relative accuracy of the DC state estimation, and two, indices that can pinpoint the error location. In the following section, various methods to improve DC state estimation are presented. Two indices on topology identification are introduced in section V. Calculations on test cases reveal that this approach is feasible.

3.4 DC State Estimation Solution Methods

3.4.1 Conventional WLS Method

The well-known solution to 3.3 is

$$\hat{x} = G^{-1} H^T W z \tag{3.6}$$

where G is gain matrix, and

$$G = H^T W H \tag{3.7}$$

The residual value for an estimation is defined as:

$$r = z - H\hat{x} \tag{3.8}$$

Assuming only measurement errors e are present, simple algebra shows:

$$r = S\epsilon \tag{3.9}$$

where

$$S = I - H(H^T W H)^{-1} H^T W \tag{3.10}$$

The expected values for $\hat{x}$ and r are then unbiased with covariance:

$$cov(\hat{x}) = G \tag{3.11}$$

$$cov(r) = SW^{-1} \tag{3.12}$$

If model errors are present, 3.9 becomes:

$$r = S(\epsilon + \gamma) \tag{3.13}$$

expected values for $\hat{x}$ and r are no longer unbiased and become:

$$E(\hat{x}) = x_{true} + G^{-1} H^T W \gamma \tag{3.14}$$

$$E(r) = S\gamma \tag{3.15}$$

3.4.2 WLS with Linear Equality Constraints

One way to improve the accuracy of conventional WLS DC state estimation is to treat virtual measurements (zero injections) as linear equality constraints. There are several methods to deal with these constraints. A simple approach used here is to simply weight heavily (i.e., assume some small covariance) any virtual measurement.

3.4.3 Singular Value Decomposition and Rank-Deficient Least Square Problems

Singular Value Decomposition(SVD) [26–28]. Any m-by-n matrix A with can be written as

$$A = U\Sigma V^T \tag{3.16}$$

where U is m-by-n and satisfies $U^T U = I$, and $\Sigma = diag(\sigma_1, ...\sigma_n)$ with $\sigma_1 \geq ... \geq \sigma_n \geq 0$. The columns, u_1,..., u_n, of U are called left singular vectors. The columns v_1,..., v_n, of V are called right singular vectors. The σ_1, ... , σ_n, are called singular values.

Rank-Deficient Least Square Problems (RDLSP) [26–28]. When matrix A is rank deficient or "close" to rank deficient, the least square problems to minimize $||Ax - b||_2$ become the so called RDLSP SVD is one of the most commonly used methods to solve this kind of problem. For the rank deficiency least square problem, let A be an m-by-n with $m \geq n$ and $rank(A) = r < n$. There is an $n - r$ dimensional set of vectors x that minimize $||Ax - b||_2$. When A is singular, the SVD of A is:

$$S = [U_1, U_2] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} [V_1, V_2]^T = U_1 \Sigma_1 V_1^T \tag{3.17}$$

where Σ_1 is $r \times r$ and nonsingular, and U_1 and V_1 have r columns. Let $\sigma =$

$\sigma_{min}(\Sigma_1)$, the smallest nonzero singular value of A. Then the solution x that minimizes $||Ax - b||$ can be characterized by:

1. All solutions x can be written as

   $$x = V_1 \Sigma_1^{-1} U_1^T b + V_2 z \tag{3.18}$$

   where z is an arbitrary vector.

2. the solution x has a unique minimum norm $||x||_2$ precisely when $z = 0$, in which case

   $$x = V_1 \Sigma_1^{-1} U_1^T b \tag{3.19}$$

   and

   $$||x||_2 \leq ||b||_2 / \sigma \tag{3.20}$$

   Thus, among the $n - r$ dimensional set of solutions, the minimal norm solution (42) for RDSLP exists and is unique.

3.4.4 Total Least Squares Method

Since the DC model is not an accurate due to model error, another possible improvement is to correct the matrix H instead of estimating model error. This leads to a class of problems called the *Total Least Squares Problems* [10-12]. This is described in the following. The objective is to x to minimize

$$||D[E, \epsilon]T||_F \tag{3.21}$$

subject to

$$(H + E)x = z + \epsilon \tag{3.22}$$

where $H, E \in R^{m \times n}$, $x \in R^{n \times l}$, $z, \epsilon \in R^{m \times l}$, and $D \in R^{m \times m}$, $T \in R^{(n+1)\times(n+1)}$. E and ϵ are unknown; D, T are weighting matrices.

Thus, the model error can be partially eliminated when extimating *x*. The literature [29] and our calculations reveals that in typical applications, gains of 10-15% in accuracy can be obtained by using TLS instead of standard least squares methods.

The condition for the TLS problem to have a unique solution is that the least singular value of H is larger than the least singular value of $[H, z]$. For typical state estimations, these conditions are satisfied. The common method to solve TLS problem is using SVD. If $D[H, z]T = U\Sigma V^T$ and U, Σ, V are partitioned as:

$$U = \begin{array}{c} \left[\begin{array}{cc} U_1 & U_2 \end{array}\right] \\ \begin{array}{cc} n & 1 \end{array} \end{array}, \Sigma = \begin{array}{cc} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} & \begin{array}{c} n \\ 1 \end{array} \\ \begin{array}{cc} n & 1 \end{array} & \end{array}, V = \begin{array}{cc} \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix} & \begin{array}{c} n \\ 1 \end{array} \\ \begin{array}{cc} n & 1 \end{array} & \end{array}$$

then

$$D[H, z]T = -U_2 \sum_2 [V_{12}^T, V_{22}^T] \tag{3.23}$$

and letting $T_1 = diag(t_1, ...t_n)$ and $T_2 = diag(t_{n+1})$

$$x = -T_1 V_{12} V_{22}^{-1} T_2^{-1} \tag{3.24}$$

3.4.5 Multiple Scan Methods

The above derivation is based on a single measurement scan. There are several recursive estimation methods based on a dynamic state estimation model that uses a sequence of scans.

Averaging - A natural and simple extension to the common DC state estimation is to average several consecutive results from the state estimator. Suppose *n* scans is used, in recursive form:

$$\hat{x}_{avg,i} = \hat{x}_i/i + (i-1)\hat{x}_{avg,i-1}/i \tag{3.25}$$

where $\hat{x}_i$ is the estimation from the i^{th} measurement z_i. $i = 1, ..., n$, and $\hat{x}_{avg} = \hat{x}_{avg,n}$. Assuming measurement errors are independent between scans, it can be shown that

$$E(\hat{x}_{avg}) = E(\hat{x}_i) \tag{3.26}$$

and

$$cov(\hat{x}_{avg}) = \frac{1}{n} cov(\hat{x}_i) \tag{3.27}$$

Thus, the averaging method generally provides a more consistent result, but it does not remove any bias that may arise, as would occur with a modeling error.

Kalman Filter [30] - Assuming that the states of the power grid do not change quickly in a short period, the following equations for the power system can be established at snapshot i

$$x_{i+1} = x_i + q_i \tag{3.28}$$

$$z_i = Hx_i + v_i \tag{3.29}$$

where q_i, v_i are Gaussian error vectors, with covariance Q and R, respectively, assuming the covariances are constant in time. The covariance Q is assumed to be small but could be approximated by any number of statistical techniques, including Monte Carlo simulation. Each step of the Kalman filter for the above systems is as follows.

- Start from the prior estimation $\hat{x}_i^-$ and its error covariance matrix P_i^-. Compute the Kalman gain matrix K_i:

$$K_i = P_i^- H_i^T (R + HP_i^- H^T)^{-1} \tag{3.30}$$

- Update the estimate $\hat{x}_i$ with the i^{th} scan measurement z_i:

$$\hat{x}_i = \hat{x}_i + K_i(z_i - H\hat{x}_i^-) \tag{3.31}$$

- Compute the error covariance matrix P_i for the updated estimate:

$$P_i = (1 - K_i H)P_i^- \tag{3.32}$$

- Project ahead to predict the new error covariance matrix P_{i+1} and new estimation $\hat{x}_{i+1}^-$

$$P_{i+1}^{-} = P_i + Q \tag{3.33}$$

$$\hat{x}_{i+1}^{-} = \hat{x}_i \tag{3.34}$$

3.4.6 Generalized State Estimation

Monticelli [18–20] introduced the concept of generalized state estimation. Generalized state estimation is performed on a model in which parts of the network can be represented at the physical level, i.e. bus-section/switch-device level. This allows modeling zero-impedance devices and switching devices. It expands the state variables in the conventional state estimation by including load flow through those zero-impedance branches and switch devices. For the linear DC model, the expanded state variable includes real power flows in those zero-impedance devices. Thus, one can judge the status of, say a breaker, by observing the power flow.

3.5 Toplogy Error Detection and Identification

Correct detection and identification of the suspect topology error is critical for reducing the calculation of the 2-stage estimation. The approach used in this work is using χ^2-test on $J(x)$ and if it fails (i.e., indicates likely errors), to use the proposed indices to identify the suspect nodes/area. Since the DC model is used here, coexisting modeling errors (4) corrupt the result, in both the χ^2-test and residual test. This makes correct error detection and identification more difficult. The proposed method tries to estimate and hence eliminate parts of the modeling error for a more accurate result.

3.5.1 Error Detection

Hypothesis Testing and Residual Test - Hypothesis testing performs a χ^2-test on $J(x)$. Since the state estimation is based on the hypothesis that there is no gross error in the measurement, gross errors should result in $J(x)$ and some normalized residuals r_n to be above some threshold. When using the linearized DC state estimation, the inaccuracies in the model itself may cause such a test to fail. Typically, for a large high voltage network, there may be as much as 5% model error. This error corrupts the residual test in several ways:

- One cannot detect topology errors in lightly loaded areas, since the errors may be less than the model error.
- Numerous “false alarms” may occur because the model error is larger than the residual threshold.

When model error is present, $J(x)$ is

$$\begin{aligned} J_{with_model_error}(x) &= \sum_{i=1}^{m}(z_i - \hat{z}_i)^2/cov(z_i) \qquad (3.35) \\ &= \sum_{i=1}^{m} w_{ii}((S_i\epsilon)^2 + 2(S_i\epsilon)(S_i\gamma) + (S_i\gamma)^2) \\ &= J_{no_model_error}(x) + \sum_{i+1}^{m} w_{ii}((S_i\gamma)^2 + 2(S_i\epsilon)(S_i\gamma)) \end{aligned}$$

where m is the number of measurements; S_i is the i^{th} row of the projection matrix S; w_{ii} is the i^{th} diagonal element of matrix W; $\hat{z} = H\hat{x}$; γ_i, z_i and $\hat{z}$ are the i^{th} component of the γ, z, and $\hat{z}$ vector respectively.

From 3.36, one finds that $J(x)$ changes with W, e, and γ. With the existence of the model error γ, the variance of $J(x)$ will increase. Still for a particular case, the model error γ and measurement error ϵ may cancel each other leaving less residue. From another point of view, the introduction of model error γ changes the distribution of the $J(x)$. So the threshold of the χ^2-test must increase significantly, especially when the covariance of the measurements is small (w_{ii} is large and $|e| < |\gamma|$). This increases the difficulty of detecting topology errors with a DC model.

3.5.2 Estimation of the Model Error

In 3.13, ϵ is a Gaussian random vector while S and γ are deterministic. One can conceivably estimate γ through r. Unfortunately, S is not of full rank but of rank $m - n$ (where m is the number of measurements and n is the number of states). Thus, γ cannot be estimated completely. This problem belongs to a category called Rank-Deficient Least Square Problems.

- Method Using Singular Value Decomposition (SVD)

  One common approach to solving a rank deficiency problem is using SVD. For the model error estimation, rewrite 3.13 as

  $$r = S\gamma + S\epsilon \tag{3.36}$$

  the minimal norm solution is

  $$\gamma = V_1 \sum\nolimits_1^{-1} U_1^T r \tag{3.37}$$

  where

$$S = [U_1, U_2] \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} [V_1, V_2]^T = U_1 \Sigma_1 V_1^T \qquad (3.38)$$

with U_1, U_2, V_1, V_2, and $\sum_i$ as given in section 3.4.3.

Though SVD gives an analytically sound result, it requires more computation than the normal equation. Typically, when $m >> n$, SVD is about twice the cost of the normal equations and when m is small, SVD is about four times of the cost of the normal equations. Since our goal is to build a simple sound method suitable for traders, by introducing some reasonable simplifications, a faster method is proposed.

- A Simplified Method

  Based on some assumptions, portions of the error γ ($m-n$ out of n elements) can be estimated. There are multiple ways to choose the $m-n$ dimension subvector of γ. As [24] illustrates, if sufficient redundancy exists, topology errors tend to affect injection measurements much more than line flow measurements. Selecting those elements of γ corresponding to injections should lead to a better overall result. Assuming, for simplicity, injection measurements are available at every node as well as line flow measurements on every line, the elements in γ corresponding to the injection at each node can be selected as the modeling errors to be estimated.

Doing so is equivalent to assuming that model errors only exist on the node injection measurements with all the modeling errors set to zero. Based on this assumption, 3.13 becomes:

$$r = S_{node}\gamma_{node} + S\epsilon \tag{3.39}$$

where, S_{node} is an m by n submatrix of S obtained by deleting all columns that correspond to flow measurements, and γ_{node} is the model error that arises on the node injection measurements. Using similar WLS methods, one obtains

$$\gamma_{node} = (S_{node}^T S_{node})^{-1} S_{node}^T r \tag{3.40}$$

Thus, the modeling error γ is

$$\gamma = \begin{bmatrix} \gamma_{node} \\ \gamma_{branch} \end{bmatrix} = \begin{bmatrix} \gamma_{node} \\ . \\ . \\ . \\ 0 \end{bmatrix} \tag{3.41}$$

3.5.3 Error Identification

If errors are detected, the next step is identification of the specific errors. Correctly locating the suspicious area is the key to reducing the computational effort in the generalized state estimation. If the model error does not change greatly between measurements, one can use the estimated model error for each subsequent scan. In this way, $J(x)$ and r_n indices are representative of measurement errors. To identify topology errors, two indices are proposed in the following.

1. **Node Index**

Usually, topology error causes larger errors in the vicinity of those related buses. Thus, simply grouping and averaging the normalized residual r_n by nodes will lead to a better index. The proposed index is built as follows:

Step 1: Initialize arrays *nodeIndex* and *nodeCount* to 0.

Step 2: For all r_n,

- Case 1: $r_n(i)$ is flow measurement on line $i - j$, then add $|r_n(i)|$ to $nodeIndex(i)$ and $nodeIndex(j)$, increase both $nodeCount(i)$ and $nodeCount(j)$ by 1.
- Case 2: $r_n(i)$ is an injection measurement node i, then add $|r_n(i)|$ to $nodeIndex(i)$, increase $nodeCount(i)$ by 1.

Step 3: For all nodes,

$nodeIndex(i) = nodeIndex(i)/nodeCount(i)$

2. **Topology Index**

When topology errors are present, the system is represented by 3.4. Using a similar method in estimating modeling error, the residual vector can be written as:

$$r = S(dHx_{true} + \gamma + \epsilon) \tag{3.42}$$

This formulates another estimation problem, if γ has been approximated. Lugtu [25] pointed out that if sufficient redundancy exists, the topology error would cause the largest residual on the node injection measurement. By using similar method to estimating model error, we can estimate the mismatch γ_{node_error} on each node injection.

$$r = S_{node}\gamma_{node_error} + S\epsilon \tag{3.43}$$

The estimated γ_{node_error} is used as a topology error index. One expects that large values will appear on the nodes that have topology errors nearby.

3.6 Test Results

The proposed method is evaluated here on the IEEE 30, 39, 57, 118 bus test systems [15].

3.6.1 Case 1 - Comparison of DC and AC state estimation

First, a comparison of DC and AC state estimation is presented. Tables 3.1, 3.2, 3.3 show the estimation errors and $J(x)$ for each of the systems. Table 3.1 shows the result of the WLS method with and without using linear equality constraints. One finds that using linear equality constraints, improves estimation accuracy. Table 3.2 shows the estimation results using TLS. As expected, the results improve slightly. Next, Table 3.3 shows the estimation result obtained by using AC state estimation. It is listed here for comparison only. Obviously, it is more accurate than the DC methods. Still, the improvement is not tremendous and the DC method appears sufficiently accurate. The results show that model correction reduces J when no error is present, thus decreasing the chance of a “false alarm”.

Note Tables 3.1, 3.2, 3.3, 3.4, 3.5:

Table 3.1: WLS Method with/without linear equality constraints

System	with		without		Degree of freedom
	State error	$J(x)$	State error	$J(x)$	
30	0.0801	38.64	0.1467	33.30	41
39	0.1312	50.46	0.1206	45.57	46
57	0.0988	270.0	0.1278	78.81	80
118	0.1238	284.5	0.1346	200.0	186

Table 3.2: DC State Estimation Using TLS

System	DC(TLS)		
	State error	$J(x)$	Degree of freedom
30	0.0938	34.41	41
39	0.1103	46.28	46
57	0.1349	98.17	80
118	0.1183	346.9	186

Table 3.3: AC State Estimation

System	AC		
	State error	$J(x)$	Degree of freedom
30	0.0113	8.538	164
39	0.0217	118.1	184
57	0.0236	23.46	320
118	0.0312	113.5	744

- All measurements have 5% Gaussian noise
- $StateError = \sqrt{\sum_i (x_i - \hat{x}_i)^2}$

Table 3.4: Estimation with no model error correction

System	IEEE 30 BUS			
	State error	$J(x)$	Degree of freedom	$\gamma^2 > 0.99$
Kalman	0.021	56.82	41	64.95
Average	0.021	80.12	41	64.95
DC	0.0238	47.12	41	64.95
AC	3.62e-0.5	10.75	164	209.0

3.6.2 Case 2 - Benefit of model error correction

The result of three different estimation methods, conventional WLS method, averaging, and Kalman filter are presented on the system with no measurement and

Table 3.5: Estimation with model error correction

System	IEEE 30 BUS			
	State error	$J(x)$	Degree of freedom	$\gamma^2 > 0.99$
Kalman	0.0169	37.08	41	64.95
Average	0.0238	55.09	41	64.95
DC	0.0167	33.95	41	64.95

topology errors. Tables 3.4 and 3.5 show the estimation errors with and without model error correction, respectively, under typical loading conditions.

3.6.3 Case 3 - Error detection and identification

1. **Single topology error**

   A test is carried out on the IEEE 30 bus system [31]. In this case, we simulate a false breaker status on line 15-14 (i.e., false branch outage on branch 15-14).

   - No gross measurement error

     Here, there are 5% Gaussian errors on each measurement but not gross measurement error. With the degrees of freedom $(m - n) = 41$, a confidence level of 0.95, the threshold for the residual test is 56.9. We find $J(x) = 172.18$, which clearly indicates that errors exist. For identification, the node and topology indices are used. The errors at the largest eight nodes, ranked in ascending order, are shown in Table 3.6. For comparison, the r_n index is also listed. From these three indices, one can immediately find that all indices are able to identify the topology error. Among them, the topology index most clearly pinpoints nodes 15 and 14 as the largest. A second step using generalized state estimation shows that the line flow on line 15-14 is 0.65, large enough to indicate that the breaker is closed. The true value of line flow 15-14 is 0.70.

- With gross measurement error

  Besides the 5% Gaussian error, a 70% gross error is included for the measurement on line 15-18. $J(x)$ is 4084, which clearly indicates error. Table 3.7 shows the three indices in this case. One finds that topology error has more impact on the indices than gross measurement error. And once again, these indices can identify the topology error. Applying generalized state estimation, the estimated value on line 15-14 is 0.65, indicating the incorrect breaker status.

2. **Multiple topology errors**

   This test is using the IEEE 30 system [31]. Topology errors are introduced for the breaker status on lines 27-30 and 10-20 i.e., false branch outages on branch 27-30 and 10-20. The three indices are shown in Table 3.8. In this case, the proposed indices especially, the topology index, are better than r_n index, since they provide a clearer view of the errors in the system. This shows that the proposed methods are also valid for identifying multiple topology errors.

Note in tables 3.6, 3.7, 3.8, the prefixes Inj, Ln in the Meas. column represent power injection and line flow measurement respectively. Power values are in p.u.

Table 3.6: Three indices (I)

Meas.	Abs(r_n)	Node	Node Index	Node	Topology Index
Ln.12-14	4.031	14	42.15	15	0.718
Lnj.14	4.010	46	23.89	14	0.624
Ln.22-24	3.501	13	13.31	1	0.271
Inj.24	3.434	47	11.22	8	0.258
Inj.23	3.058	49	7.948	13	0.153
Ln.28-27	3.052	15	7.363	9	0.127
Inj.2	2.661	51	6.073	3	0.0542
Inj.4	2.601	50	6.030	12	0.0518

Table 3.7: Three indices (II)

Meas.	Abs(r_n)	Node	Node Index	Node	Topology Index
Inj.14	59.05	14	42.35	15	0.641
Ln.14-13	54.74	46	24.86	14	0.604
Inj. 46	49.68	13	13.31	7	0.359
Inj. 47	23.18	47	12.72	1	0.226
Inj. 15	16.71	15	7.637	13	0.185
Inj. 48	13.51	49	7.092	26	0.163
Inj. 14-46	13.25	48	7.043	23	0.154
Inj.13	12.69	51	5.720	8	0.147

Table 3.8: Three indices (III)

Meas.	Abs(r_n)	Node	Node Index	Node	Topology Index
Inj. 20	16.00	20	12.84	10	0.0983
Ln. 19-20	9.687	30	9.026	20	0.0917
Inj. 30	9.275	19	7.456	27	0.0802
Ln. 29-30	8.776	27	4.927	30	0.0683
Inj. 10	8.390	29	4.850	4	0.0560
Inj. 19	7.258	10	3.979	1	0.0480
Inj. 12	6.861	16	3.455	2	0.0472

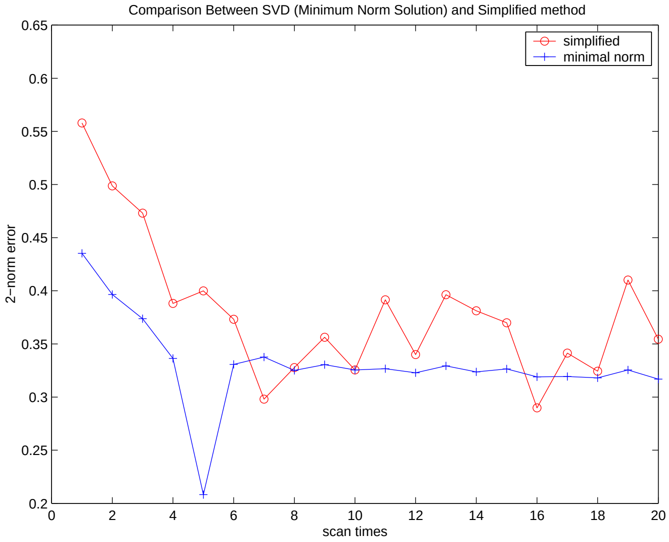


Figure 3.3: Performance comparison between SVD and proposed method

3.6.4 Case 4 - Comparison of SVD and the proposed method

A comparison is made based on the difference between the estimated error $\hat{e}$ and the actual model error e, i.e. $\sqrt{\sum_i (e_i - \hat{e}_i)^2}$, over successive scans. As one might expect the SVD achieves more consistence results, but the performance is quite similar. Since the proposed method requires far less computation, it is suggested here as the preferred method.

3.7 Discussion and Conclusions

A two-stage DC estimation is proposed to detect and identify topology errors. In the first stage, state estimation is performed on the bus/branch level. When errors are detected, the suspicious area is converted to bus-section/switching-device level and the second stage state estimation is performed. Multiple scan DC state estimation methods are introduced. The DC model modeling error is also partly estimated. Results on several IEEE test systems show the validity of the method.

The DC estimator is not proposed here to be a replacement for a full AC estimator, which might be needed by the system operator, but rather as a simplified view of the power system appropriate for certain market participants. An open electricity market has many players with different viewpoints of the system and needs for accuracy. DC state estimation has many advantages and could easily be implemented outside the control center given availability to select measured data and system parameters. Further, the results can be more easily related to typical market rules. The author suggests that where the proposed estimator begins to break down under the burden of modeling errors, it is also likely that the limits of the trading rules will begin to be reached.

Chapter 4

Power System Communication System Based on Distributed Event Service

4.1 Introduction

With the on-going deregulation, power system is facing some fundamental changes. At the same time, the network and information technology are also undergoing tremendous changes. The deregulation introduces more participants beyond control centers, such as individual power traders and small generation companies, and requires fair and transparent sharing of information to all the participants in the power energy market. The existing power communication system is less in flexible and may not fit the needs of the new environment. A new way of handling data and information exchange in power system based on the distributed event service (Fig. 4.1 [32]) is proposed. Different architectures are presented. The performance under certain architectures is also studied.

4.2 Motivation

The existing power communication network is a centralized hardwired structure. All the data generated for the RTU is passed directly to the control center. This model has served power system for decades and almost all the existing EMS/SCADA systems still uses this model. It is hard to understand this model without taking a look at the power system at first. The traditional power industry is a monopoly franchise and owns all the generation, transmission and distribution facilities. It is safe to say the control center and the centralized power communication infrastructure is the byproduct of the monopoly structure. The weakness of this model is lack of flexibility. In the days of centralized power system controls, this is not a problem since all the supervisory and controls are done by the control center. However, power industries are under restructuring since the government believes that the advantages of competition among energy suppliers and wide choice for electricity consumers, outweighs the benefits of the long-established arrangement. They believe deregulation can bring the consumer long term benefits though there might be some new or potential problems. One of the requirements of the deregulated power system and open power energy markets is to make sure every participant has a fare share of the power grid information. At this moment, almost all the information is location inside the control center and all others have to obtain the data from the control center. The structure works but it is awkward. A new structure of power communication system is needed to fit the restructured power system.

The proposed power system communication infrastructure is based on distributed event service, where a group of objects on different location coordinate and pass data through asynchrony events.

The service is implemented as a group of servers that provide access points to clients. Clients using access points to advertise and publish its information. They can also use these access points to subscribe on event of interest and event service can also notify the clients of the event of interest. This is the so called publish-subscribe paradigm. The clients that publish data are called publisher and clients that subscribe to events are called subscribers.

As for the power system communication network, the typical publisher could be a substation, generation plant, or so on, and typical subscriber could be a control center, ISO, or power energy market traders. In fine granularity, the publisher could be the IEDs in the substations, and the subscribers could be the power applications, such as, a state estimator or a topology processor.

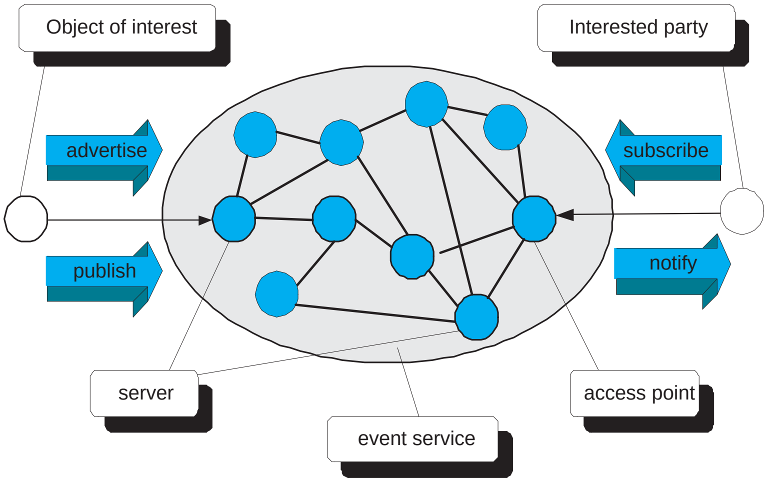


Figure 4.1: Distributed Event Service

4.3 Architectures

It is hard to obtain reasonable simulation result of the communication system without addressing the internal architecture of the system. In order to cooperate with each other and deliver events across a wide-area network, the servers must therefore be arranged into an interconnection topology and make use of some server/server communication protocol. We refer both topology and protocol as the architecture of the event service system. There are several existing different architectures, server topologies and protocols, including hierarchical client/server, acyclic peer-to-peer, general peer-to-peer and some hybrid architectures [32].

The architecture is assumed to be implemented on top of a lower-level network infrastructure. In particular, a topological connection between two servers does not necessarily mean a permanent or direct physical connection between those servers. Moreover, the server/server protocol might make use of any one of a number of existing network protocols.

4.3.1 Hierarchical Client/Server Architecture

A natural way of connecting event servers is according to a hierarchical topology [32], as illustrated in Figure 4.2 [32]. In this topology, pairs of connected servers interact in an asymmetric client/server relationship. Hence, we use a directed graph to represent the topology of this architecture, and we refer to this architecture as hierarchical client/server architecture. A server can have any number of incoming connections from other "client" servers, but only one outgoing connection to its own "master" server. A server that has no "master" server of its own is referred to as a root. The protocol between the "client" server and "master" server is similar to the client/server protocol. The advantage of the architecture is that it is easy to manage and apply policy. It is a straightforward extension of a centralized architecture.

The main problem exhibited by the hierarchical architecture is the potential overloading of servers high in the hierarchy. Moreover, every server acts as a critical point of failure for the whole network.

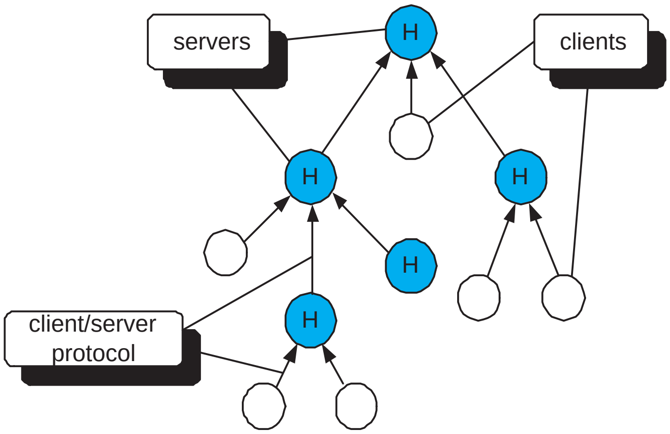


Figure 4.2: Hierarchical Client / Server Architecture

4.3.2 Acyclic Peer-to-Peer Architecture

Acyclic peer-to-peer architecture [32] (figure 4.3 [32]) gets its name because its topology forms an acyclic undirected graph. Servers communicate with each other symmetrically as peers, using a protocol that allows a bi-directional flow of subscriptions, advertisements, and notifications. It is important that the connections among servers has the property of acyclicity, since routing algorithms might rely on the property to assume, for instance, that any two servers are connected with at most one path. However, ensuring this can be difficult and/or costly in a wide-area service in which administration is decentralized and autonomous.

As in the hierarchical architecture, the lack of redundancy in the topology make it difficult to ensure connectivity, since a failure in one server S isolates all the subnets reachable from those servers directly connected to S.

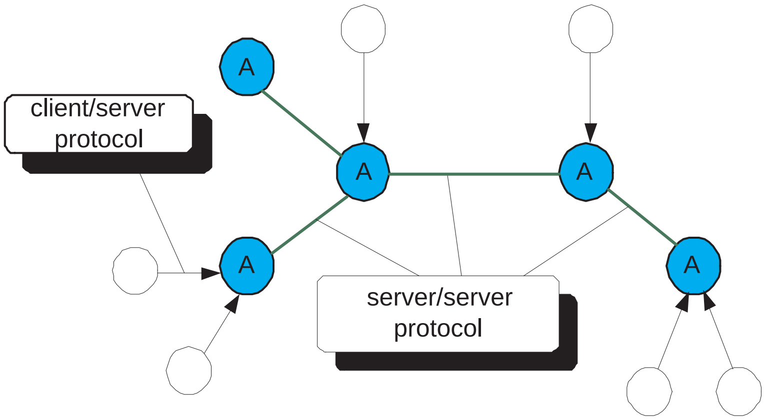


Figure 4.3: Acyclic peer-to-peer architecture

4.3.3 General Peer-to-Peer Architecture

General peer-to-peer architecture [32] (figure 4.4 [32]) is peer-to-peer architecture without the constraint of acyclicity. As in the acyclic peer-to-peer architecture, this architecture allows bi-directional communication between two servers, but the topology can form a general undirected graph, possibly having multiple paths between servers. The advantage of the general peer-to-peer architecture over the previous two architectures is that it requires less coordination and offers more flexibility in the configuration of connections among servers. Moreover, allowing redundant connections makes it more robust with respect to failures of single servers. The drawback of having redundant connections is that special algorithms must be implemented to avoid cycles and to choose the best paths. Consequently, the server/server protocol using in the general peer-to-peer architecture may require additional information.

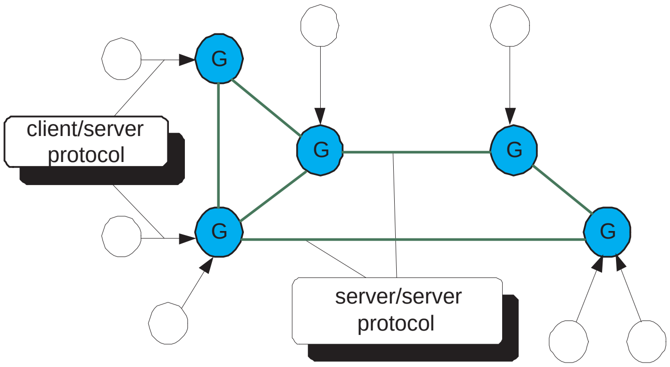


Figure 4.4: General peer-to-peer architecture

4.3.4 Hybrid Architecture

A wide-area, large-scale, decentralized service such as power system communication system poses different requirements at different levels of administration. We can potentially take advantage of these intermediate levels to gain some efficiency by considering the use of different architectures at different levels of network granularity [32]. One possible architecture using hierarchical architecture is to build a event service between different power utilities. Within a certain utility intranet, a high degree of control and coordination is required. It is better to apply hierarchical architecture here. However, between utilities, there is no high level central control, so applying general peer-to-peer architecture is more appropriate(figure 4.5 [32]).

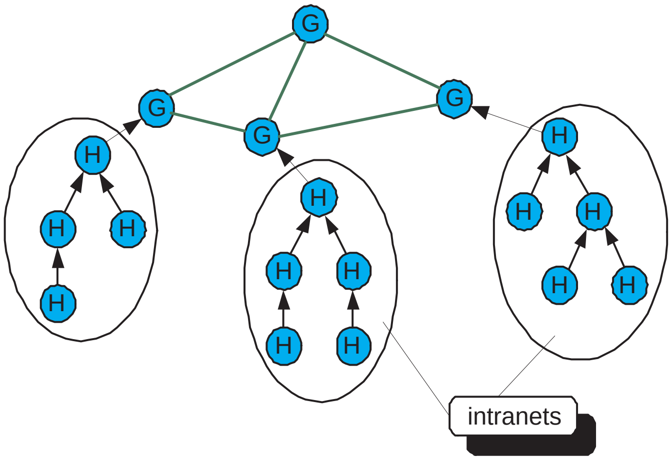


Figure 4.5: Hybrid architecture

4.3.5 Proposed Architecture

The proposed structure (figure 4.6) is similar to the general peer-to-peer server architecture. The simulations here are based on this architecture. The publisher is push out the data to the servers using multicast. The subscriber subscribes the event of interest and uses receiver select protocol to select the server(s) that has/have the minimal delay(s). The subscriber also uses a slide window like mechanism to eliminate duplicates, which is usually introduced by the redundancy. The servers maintain a table that includes the information of all the subscription and periodically exchange data and information. For a certain subscription, the processing is mainly localized in one or more particular server(s).

Design Considerations

The most important consideration is to meet real-time constraints. Power system operations require real-time decisions. Most of the applications have very tight

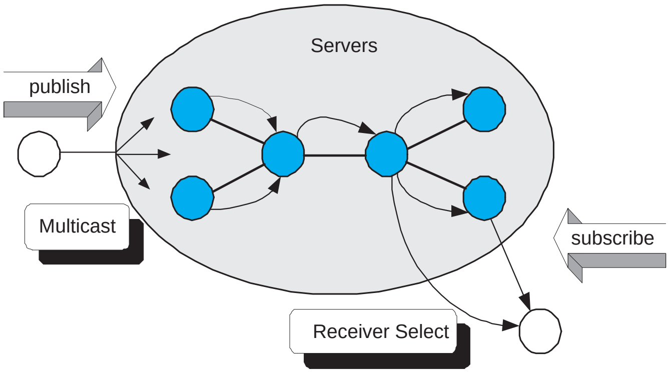


Figure 4.6: Proposed architecture

time constraints. Thus, the proposed event service is implemented using UDP protocol to minimize the transmission overhead to meet the real-time requirements. For similar reasons, acknowledge (ACK) is minimal to the subscription request. Other reasons for minimizing the use of acknowledge packet is to reduce retransmission, since historical data is only of very limited use. It is unwise to compromise the delay of new data to retransmit old data. Thus, ACK only applies to those important control messages that require very high reliability.

The second goal is to achieve a high level of reliability. It is very important for certain of messages to arrive in time and intact to ensure the safe operation of the system. A good design should be flexible and meet the requirements of a wide spectrum of applications. One should be able to specify different quality of service (QoS) requirements.

The State Transition of a Receiver/Subscriber

The implementation here adopts a receiver-select mechanism to select the best service access point. Initially, subscriber send subscribe request to the servers and wait for reply. Then, the messages might come from different servers that have different delay. After a certain amount of time, the subscriber/receiver is able to tell which one has the least average delay and advice other servers to stop sending message to it. The state diagram is show below in figure 4.7.

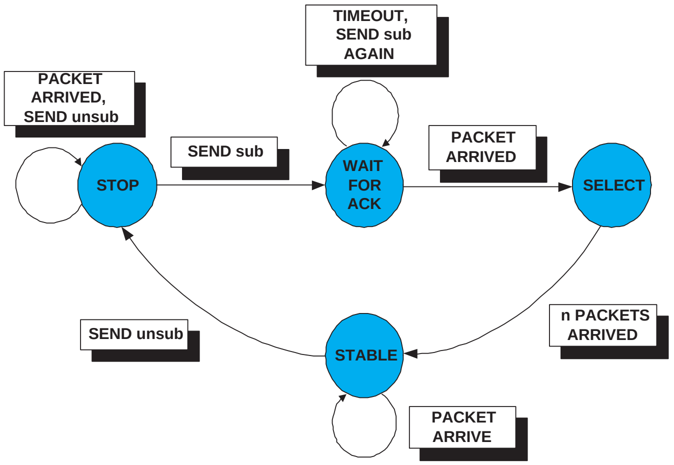


Figure 4.7: State diagram of a subscriber/receiver

Mechanism to Eliminate the Duplicates

When redundancy is present, we need a mechanism to eliminate duplicate messages. The method used is a slide window method. Since we do not have ACK in many cases here, the protocol is different from the traditional slide window protocol on how to advance the window. The receiver maintain two variables, the window size, denoted WS, given the upper bound of the messages that the client agent can hold; NME denotes the next message expected, shown in 4.8.

The receive window advances only in two cases, that is, only two ways can increase NME.

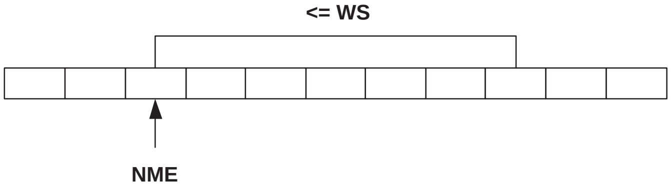


Figure 4.8: Receiver window

- **Case 1:**

  The message just arrived has the *SeqNum* equal to *NME*. The window advances until to the first message expected to come. *NME* is equal to the *SeqNum* of that message.

- **Case 2:**

  The *SeqNum* of the last received message is larger than the *NME + WS*. Then, the window advances to at least *SeqNum - WS - 1*, so that the receive window can hold the latest arrived message. Increase *NME* accordingly. All the messages have not come between the old *NME* and the new *NME* are marked as lost. All the messages that come with *SeqNum* already received or with *SeqNum* less than *NME* are discarded silently.

  The reason we try to give the newest message higher priority is that, in the power system, users are more interested in the most recent data.

  The sequence number *SeqNum* may be modified with a timestamp and the mechanism is still effective.

4.4 Test Cases and Assumptions

Simulation is carried out on four IEEE testing systems, IEEE 30, 57, 118, 300 bus systems, which have 30, 57, 118, 300 nodes respectively. Since actually system topology structure and parameters of the communication network are not available. I make the following assumptions:

- Each node (except nodes that are connected to transformers) in the system represents a substation that is a publisher in the system.
- Each substation publishes two types of data in different interval
  - Regular data, published periodically, every 4 seconds.
  - Contingency data, sent as a Poisson Process
  - Contingency data sent with higher priority than the regular data type.
- The size of the data published by each substation in each cycle is equal to or less than 10k bytes.
- The communication line follows the power transmission line. And according experience, the distance can be estimated and is proportion to the resistance of power transmission line. (approximately 0.15 Ohm per mile).
- Each server node has a fixed its own processing delay (software and hardware) that is related to the arriving packet size. The delay is the result of message pack/unpack and marshalling/unmarshalling. The formula used in the simulation is

$$delay = (1e - 7) * total\_packet\_size\_in\_server + 0.01 \tag{4.1}$$

- The capacity of the communication channel varies from 500kb to 10Mb.
- The client/server ratio is kept constant in all testing system. The ratio is 10.

The packet queue in each node implemented as Random Early-Detection (RED). The minimum threshold is 50, maximum threshold is 90 and its drop mechanism is random drop, which means that when the number of arriving packet exceeds 50, it begins to drop packets at random and when the number of packets exceeds 90, it begins to drop packets by force. The simulation uses centralized multicast and the simulation is implemented in ns-2.

4.5 Results and Analysis

4.5.1 Varying Bandwidth

Packet Delay

Figures 4.9, 4.10, 4.11 and 4.12 are the delays of regular messages that are generated periodically and the delays of urgent messages that have Poisson arrival rate. They are the simulated result by setting all the transmission channels to the specified bandwidth. One can find that the delay decreases with increasing bandwidth. However after certain point, the decrease in delay becomes minimal since server delay dominates. From figures 4.9, 4.10, 4.11 and 4.12, one can also find that the delay grows when system size grows, particularly in the low bandwidth situation. Although the growth is linear, one can find that in IEEE 300 Bus system, the mean delay is close to 1 second. This means that there are some bottlenecks in the network. This can be easily solved by assigning more bandwidth to those channels.

Closer study can reveal that the delay is a little bit less than for regular message. It is because that the urgent message has smaller packet size and RED queue favors small size packet. Also, the smaller size packet has less transmission delay. As above, we need give some “busy” channels more bandwidth to decrease the message delay.

Packet Arrival Percentage

Figures 4.13, 4.14, 4.15 and 4.16 are the arrival percentage of the regular and urgent messages giving packet error rate of 0.0001. As expected, packet arrival percentage rises with increasing bandwidth. IEEE 30, 57, 118 bus systems all have satisfactory mean percentage when bandwidth is large. However, the IEEE 300 bus system still suffers from packet loss. Increasing bandwidth for certain bottlenecks should solve the problem. Since the RED queue favors to small packets. It is not surprising that urgent message has lower packet loss rate.

4.5.2 Varying Message Size

Packet Delay

Figures 4.17, 4.18, 4.19 and 4.20 are the packet delay of regular and urgent message when varying the packet size. The bandwidth for all channels used here

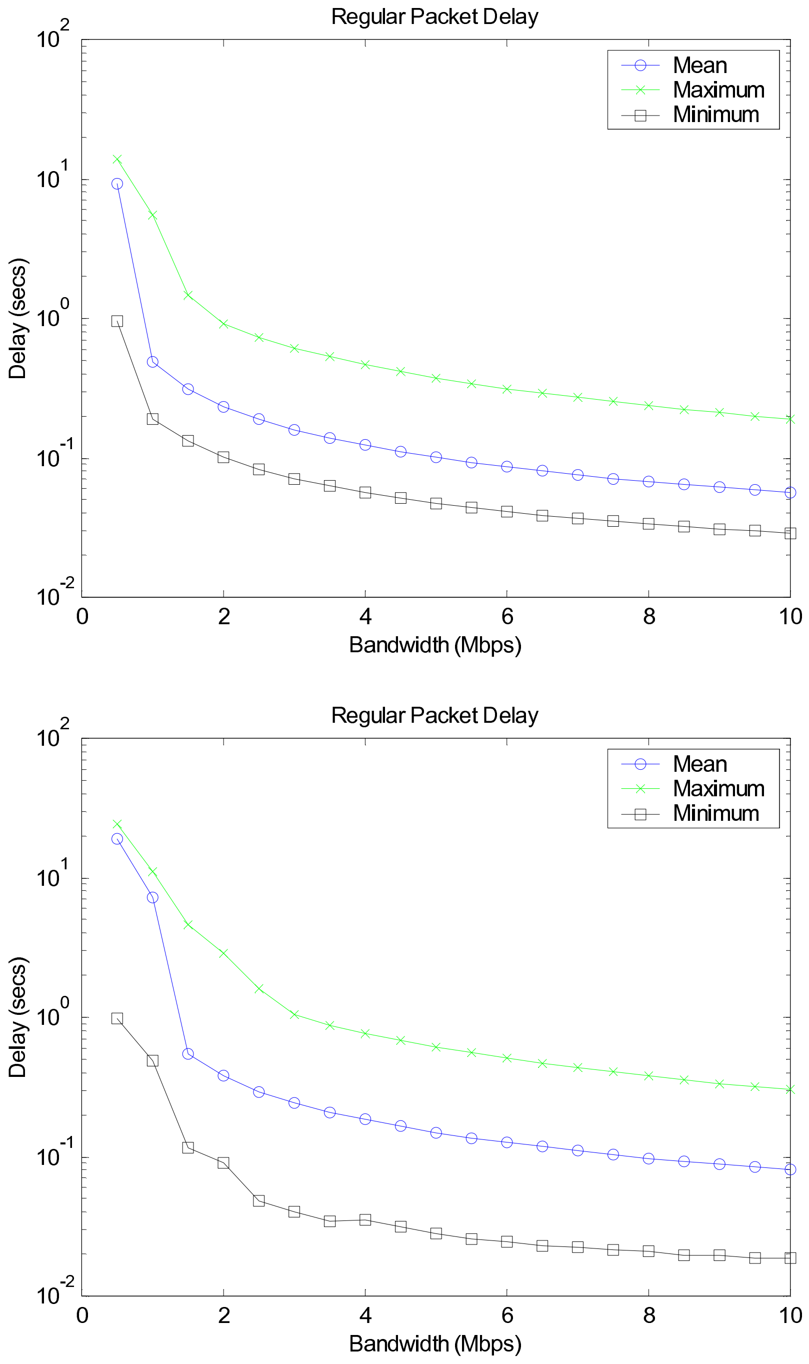


Figure 4.9: Packet delay varying bandwidth for regular message (a) IEEEE 30 bus system (b) IEEEE 57 bus system

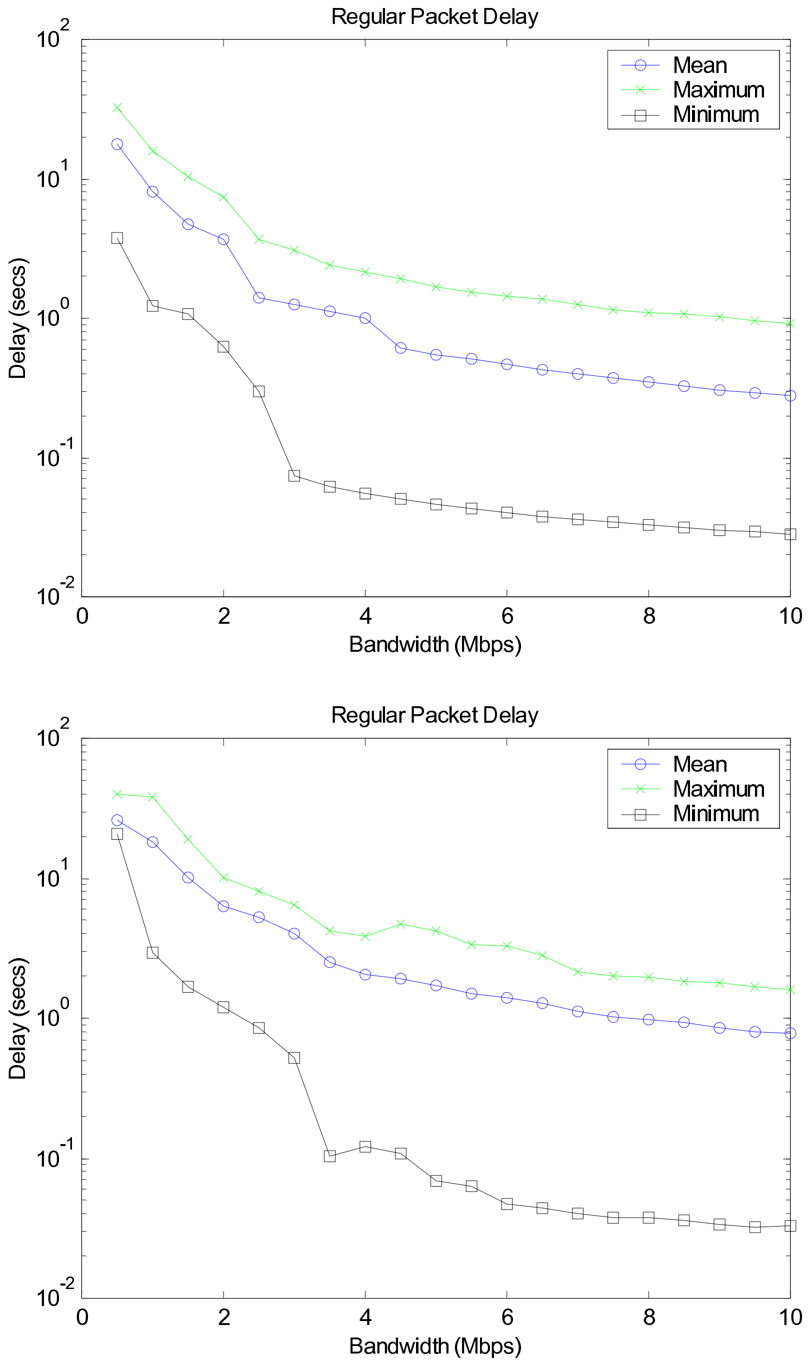


Figure 4.10: Packet delay varying bandwidth for regular message (a) IEEEE 118 bus system (b) IEEEE 300 bus system

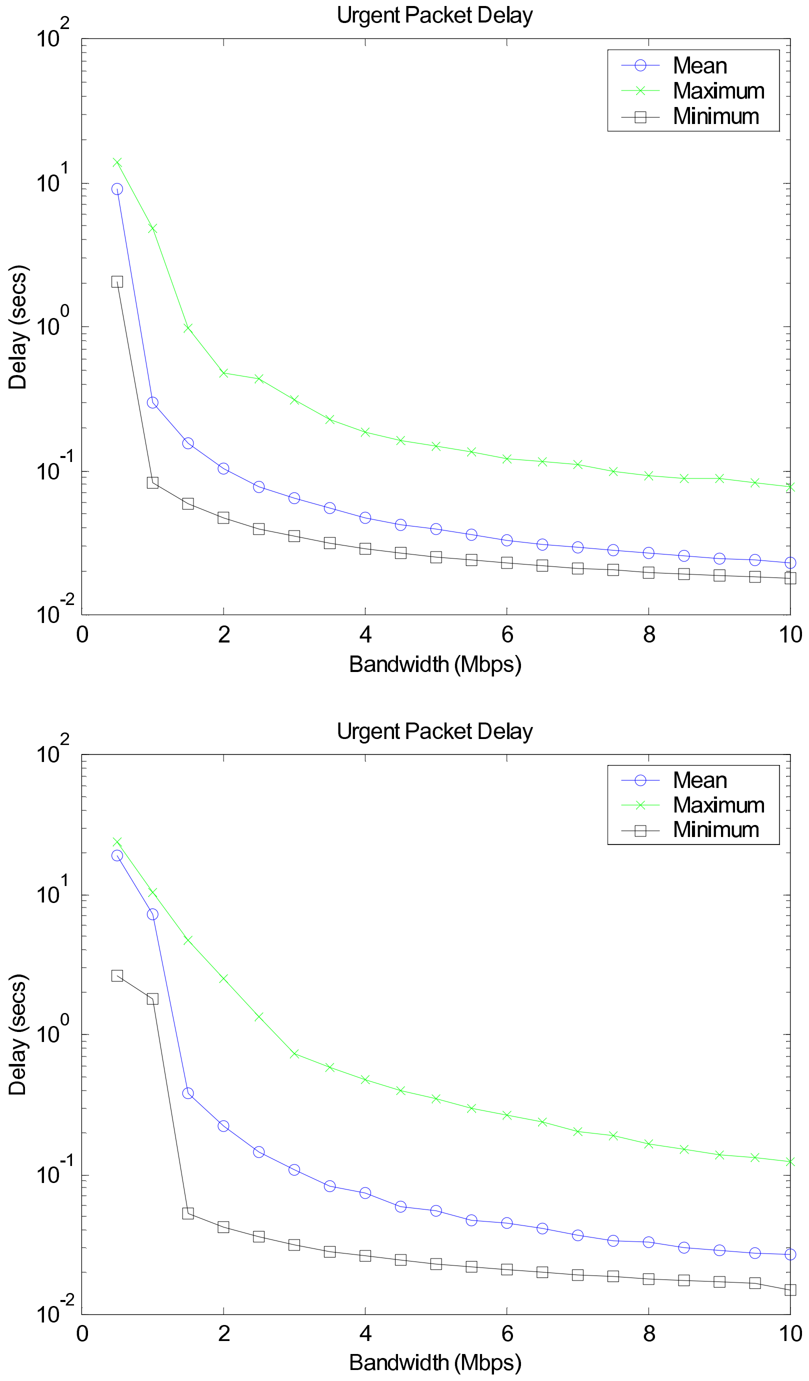


Figure 4.11: Packet delay varying bandwidth for urgent message (a) IEEEE 30 bus system (b) IEEEE 57 bus system

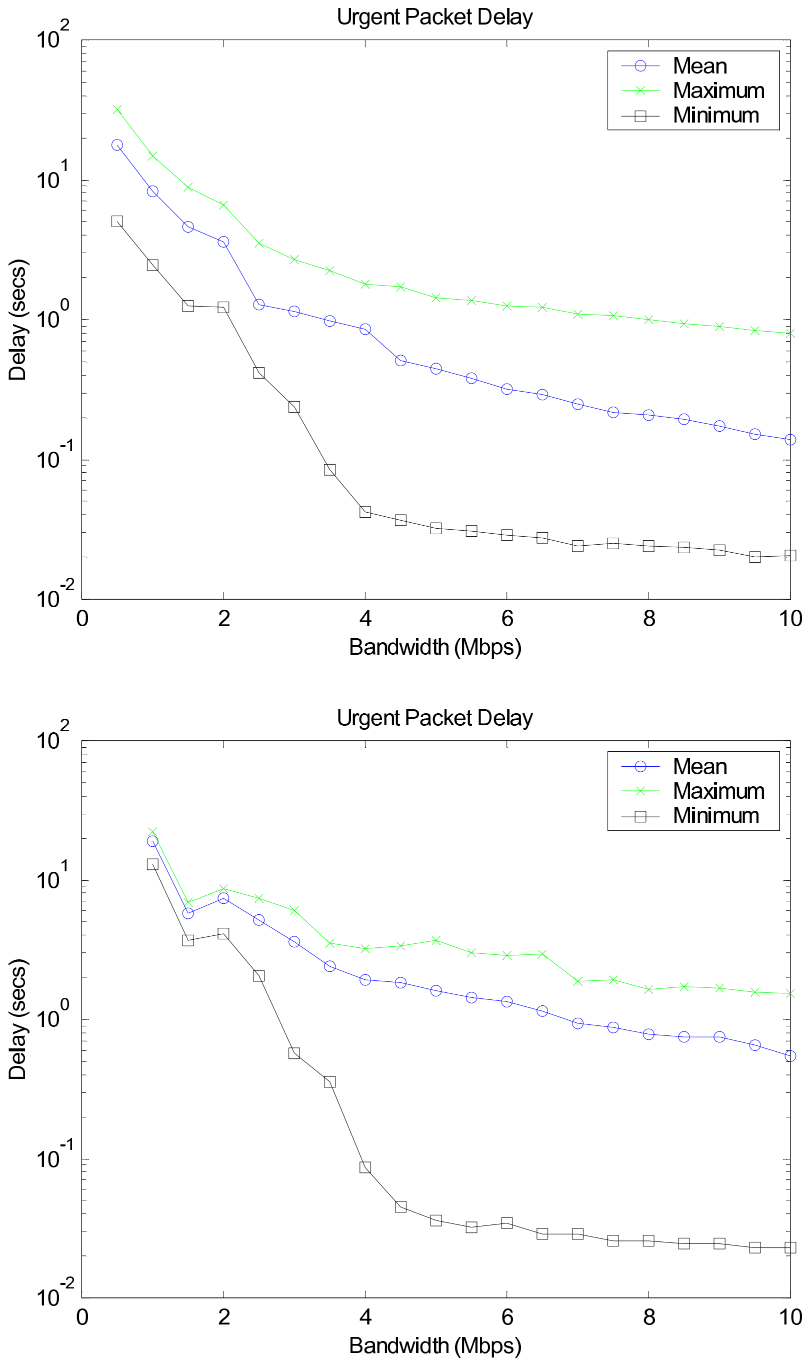


Figure 4.12: Packet delay varying bandwidth for urgent message (a) IEEEE 118 bus system (b) IEEEE 300 bus system

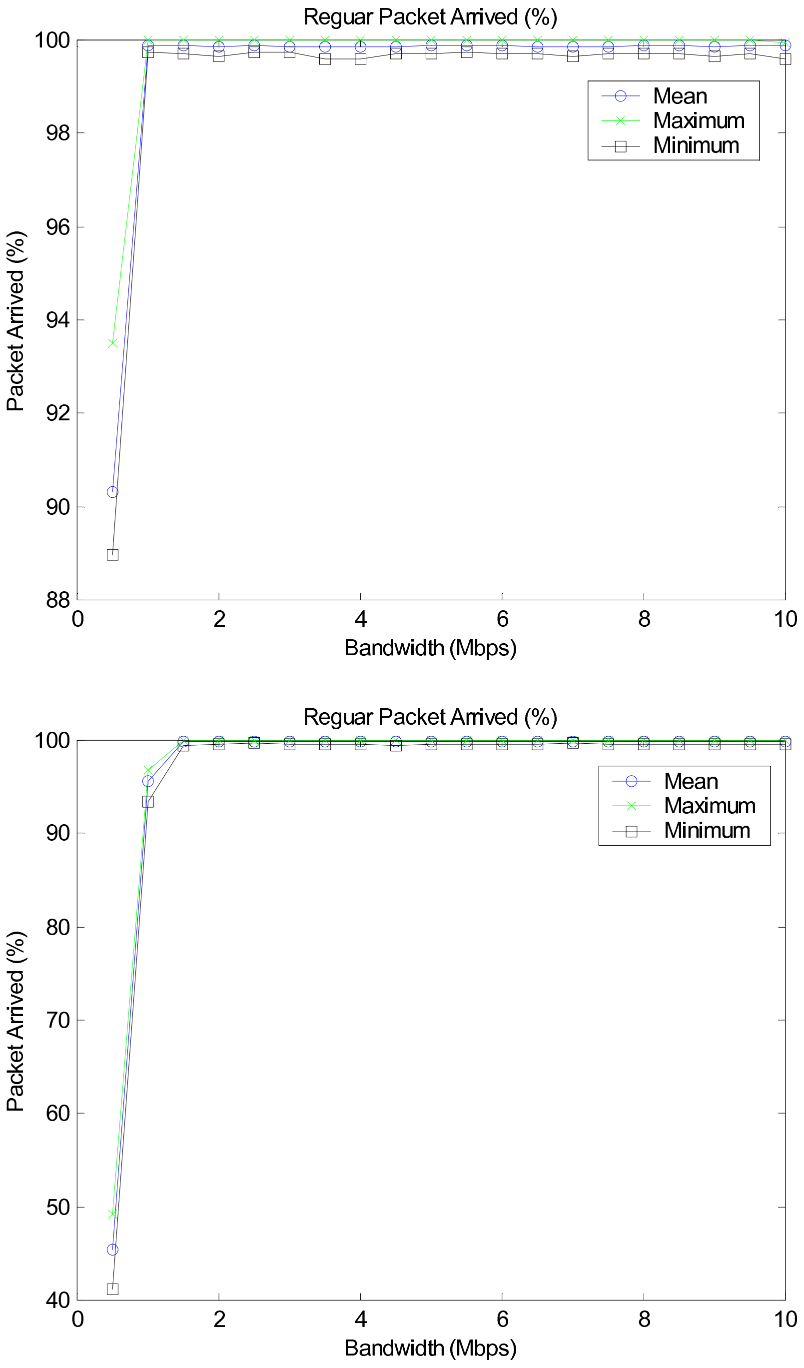


Figure 4.13: Packet arrival percentage varying bandwidth for regular message (a) IEEEE 30 bus system (b) IEEEE 57 bus system

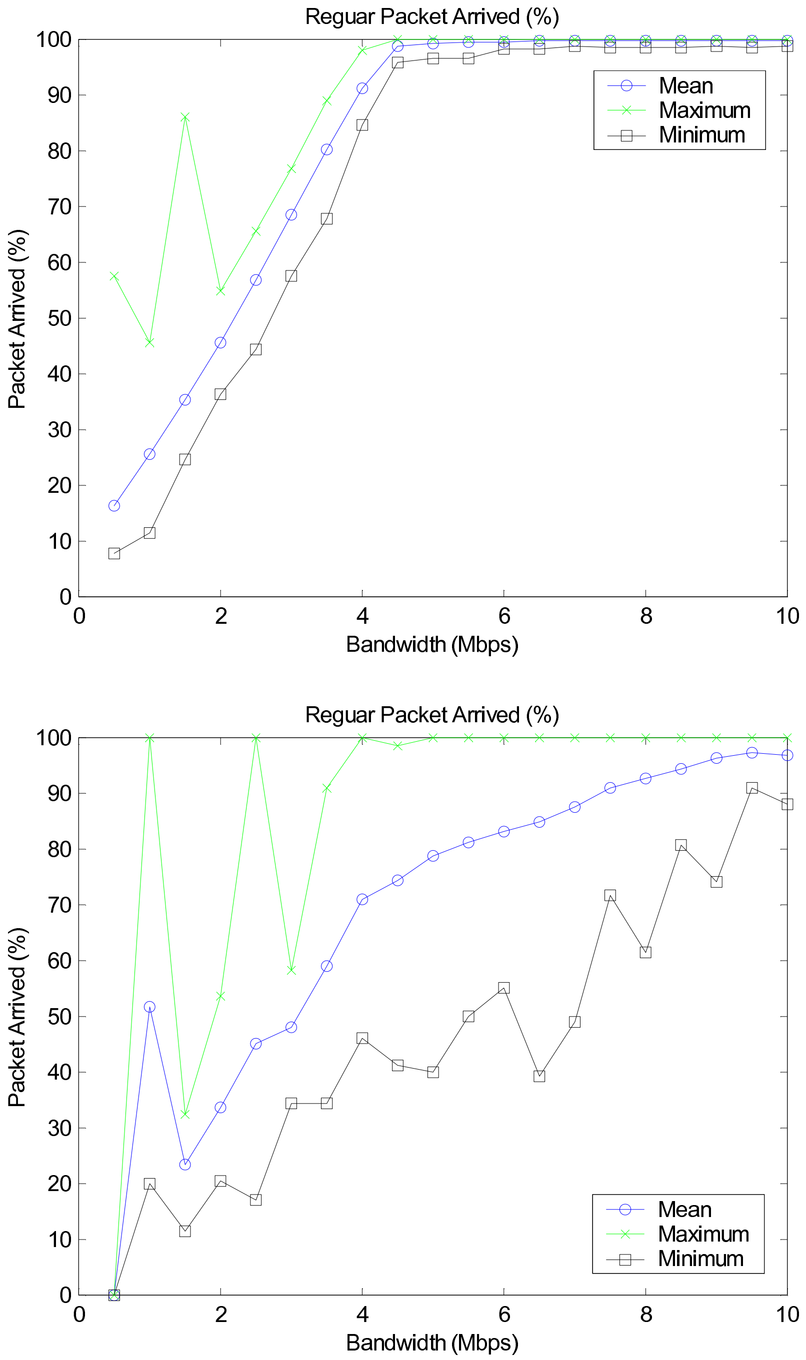


Figure 4.14: Packet arrival percentage varying bandwidth for regular message (a) IEEEE 118 bus system (b) IEEEE 300 bus system

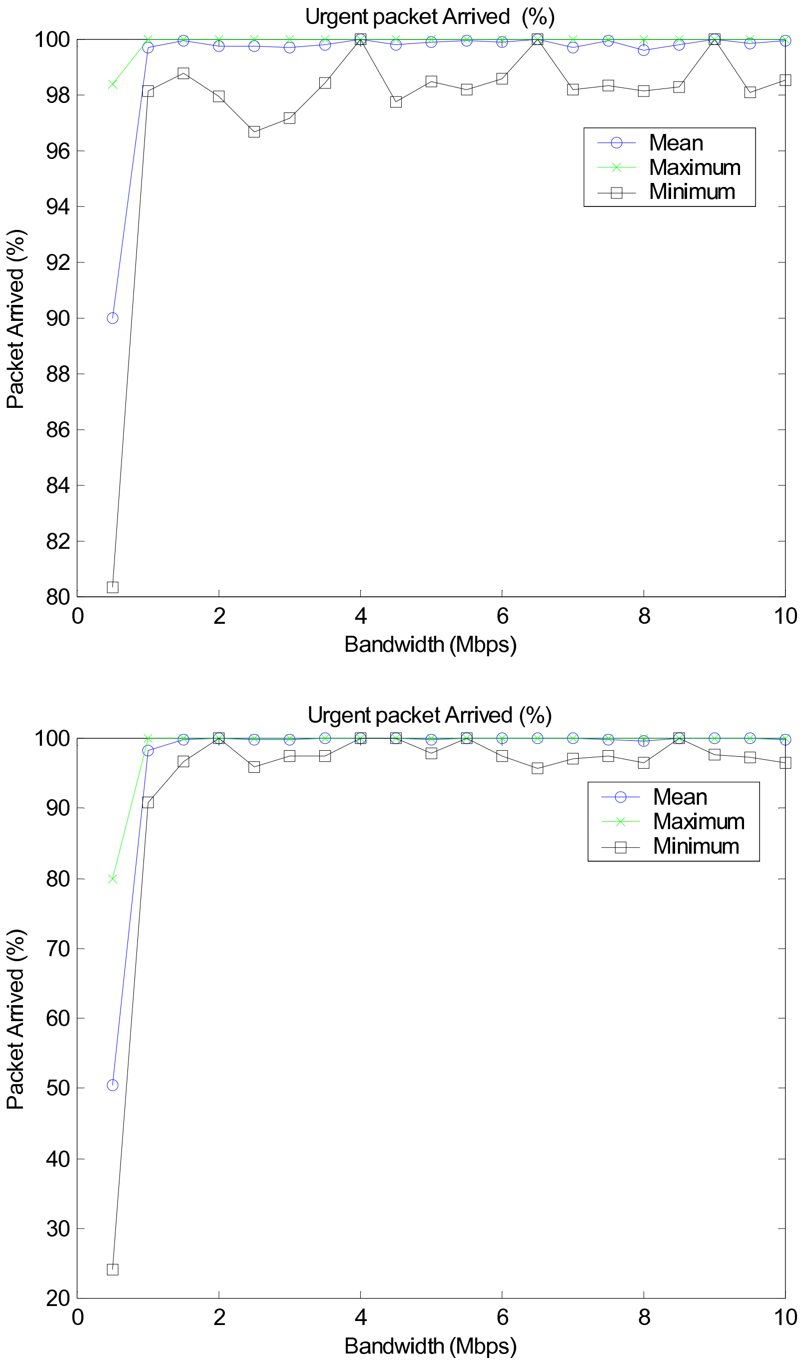


Figure 4.15: Packet arrival percentage varying bandwidth for urgent message (a) IEEEE 30 bus system (b) IEEEE 57 bus system

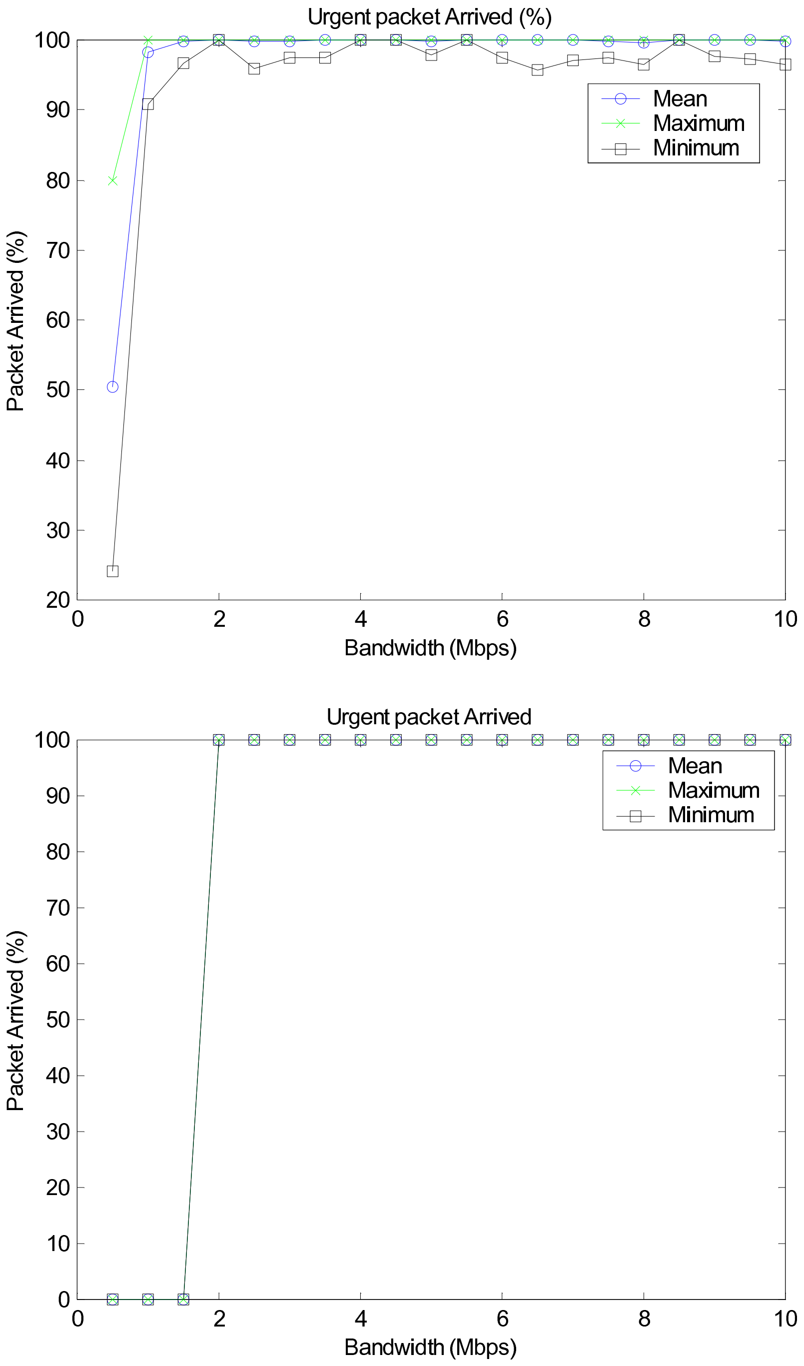


Figure 4.16: Packet arrival percentage varying bandwidth for urgent message (a) IEEEE 118 bus system (b) IEEEE 300 bus system

is 5 Mb. One can find that the delay grows linearly with packet size. And the delay grows with the system size for fixed bandwidth. This is because in the larger system, the channel is easily overloaded and becomes the bottleneck.

Packet Arrival Percentage

Figures 4.21, 4.22, 4.23 and 4.24 are for packet arrival percentage of regular and urgent message when varying the packet size. The bandwidth for all channels used here is 5 Mb. One finds that the percentage remains unchanged with the packet size changes. The only exception is for the IEEE 300 Bus system. The reason for that is since 5 Mb is enough for small or medium size system. However, for a larger system, one needs higher bandwidth and better topology to avoid bottlenecks.

4.5.3 Varying Packet Error Rate and Redundancy

Packet Delay

Figures 4.25, 4.26, 4.27 and 4.28 are the packet delay of regular and urgent message when varying the packet loss rate and redundancy. Redundancy equals to 1 means that each subscriber has one service access point and no redundancy

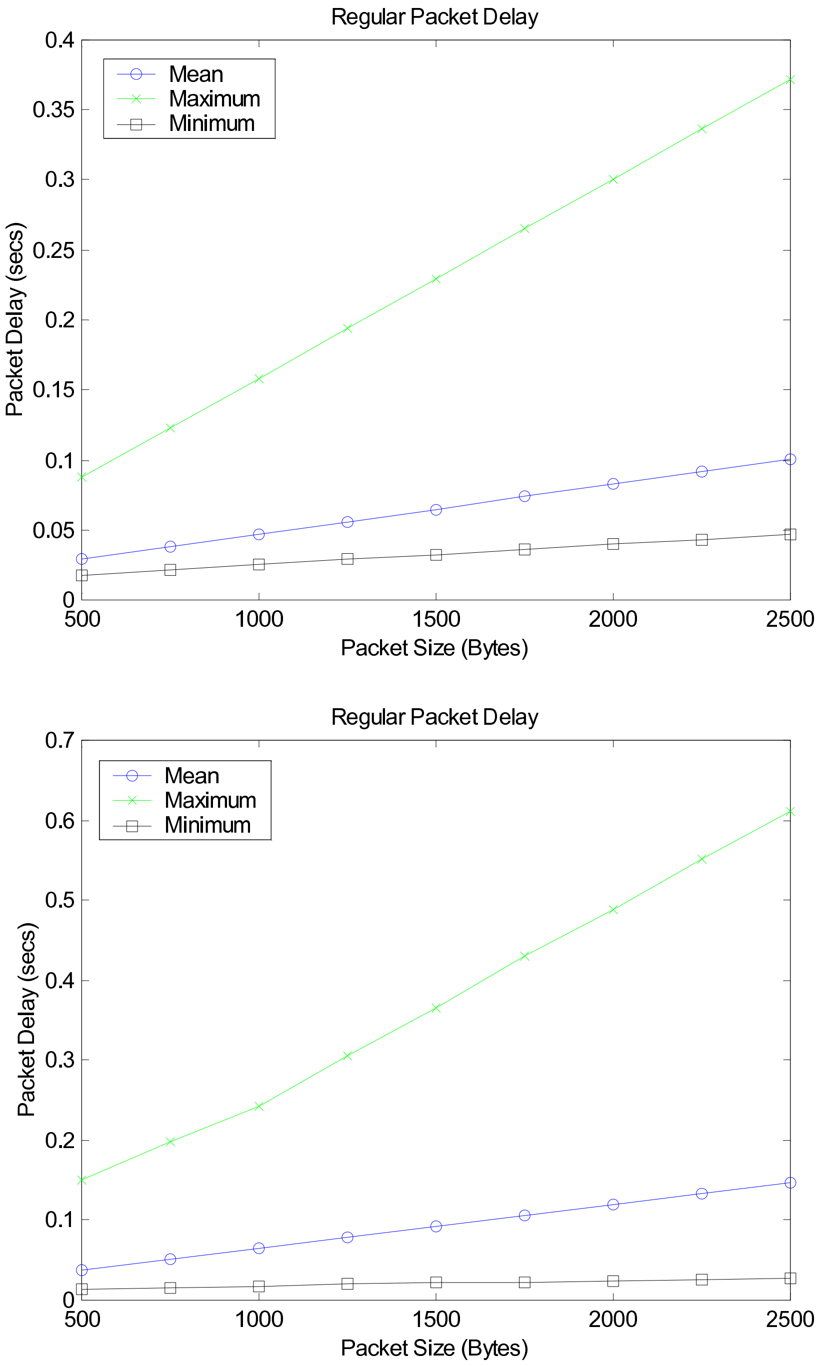


Figure 4.17: Packet delay varying packet size for regular message (a) IEEEE 30 bus system (b) IEEEE 57 bus system

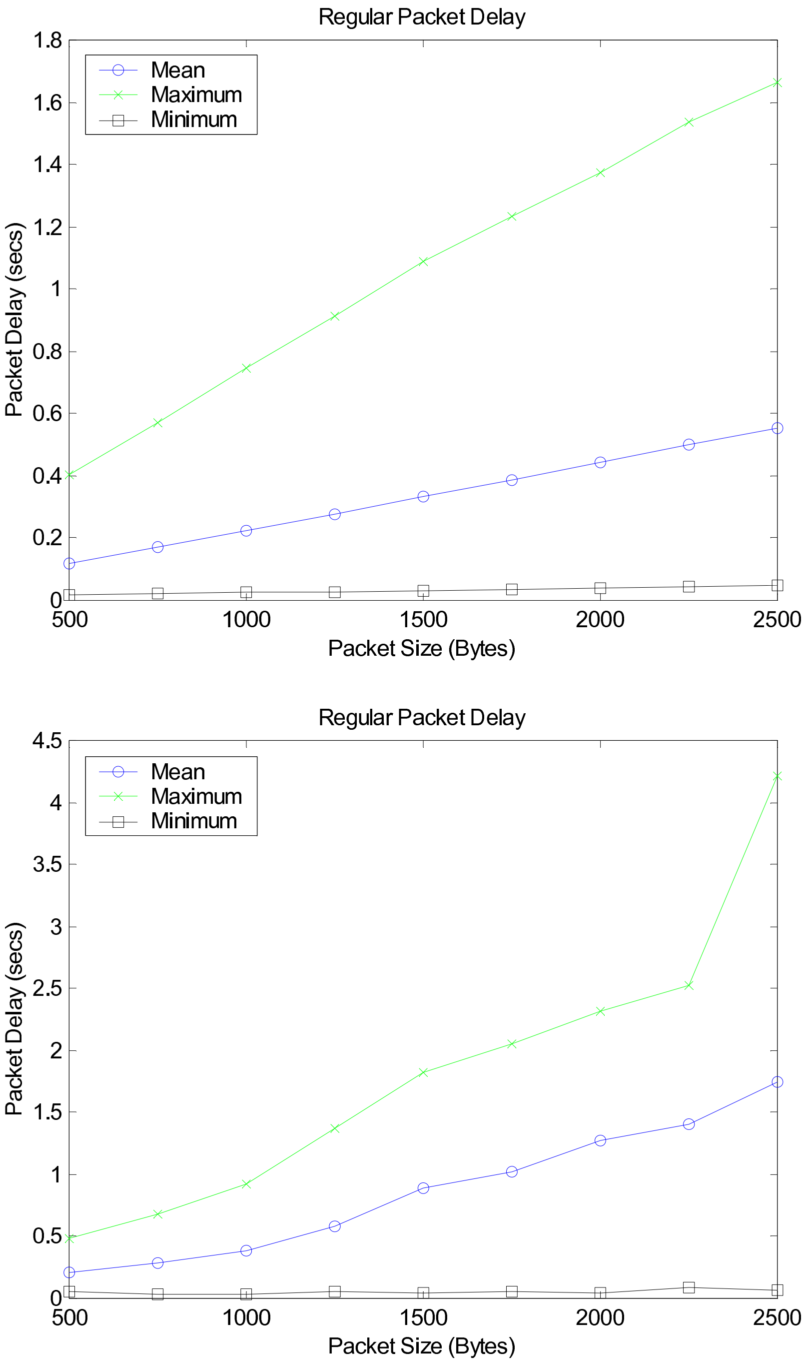


Figure 4.18: Packet delay varying packet size for regular message (a) IEEEE 118 bus system (b) IEEEE 300 bus system

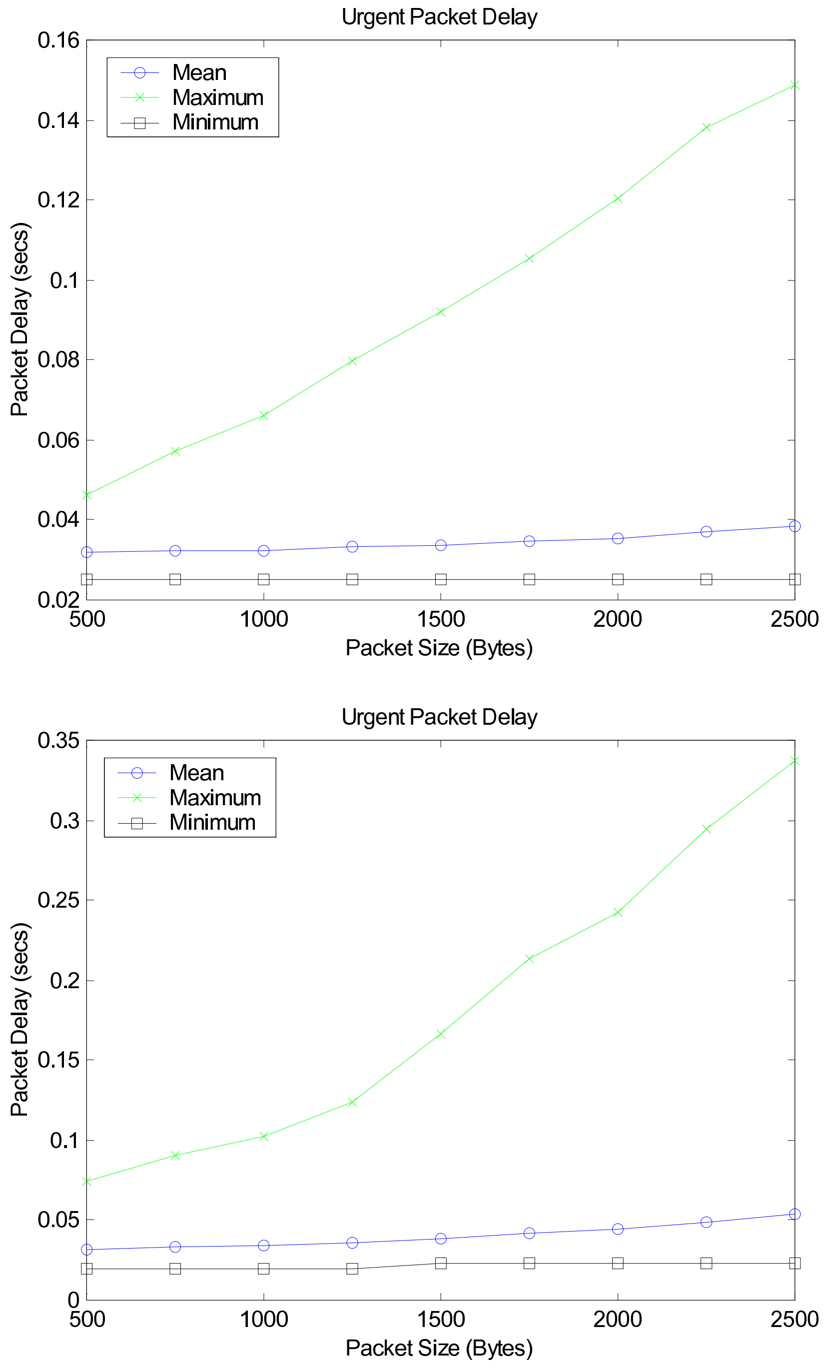


Figure 4.19: Packet delay varying packet size for urgent message (a) IEEEE 30 bus system (b) IEEEE 57 bus system

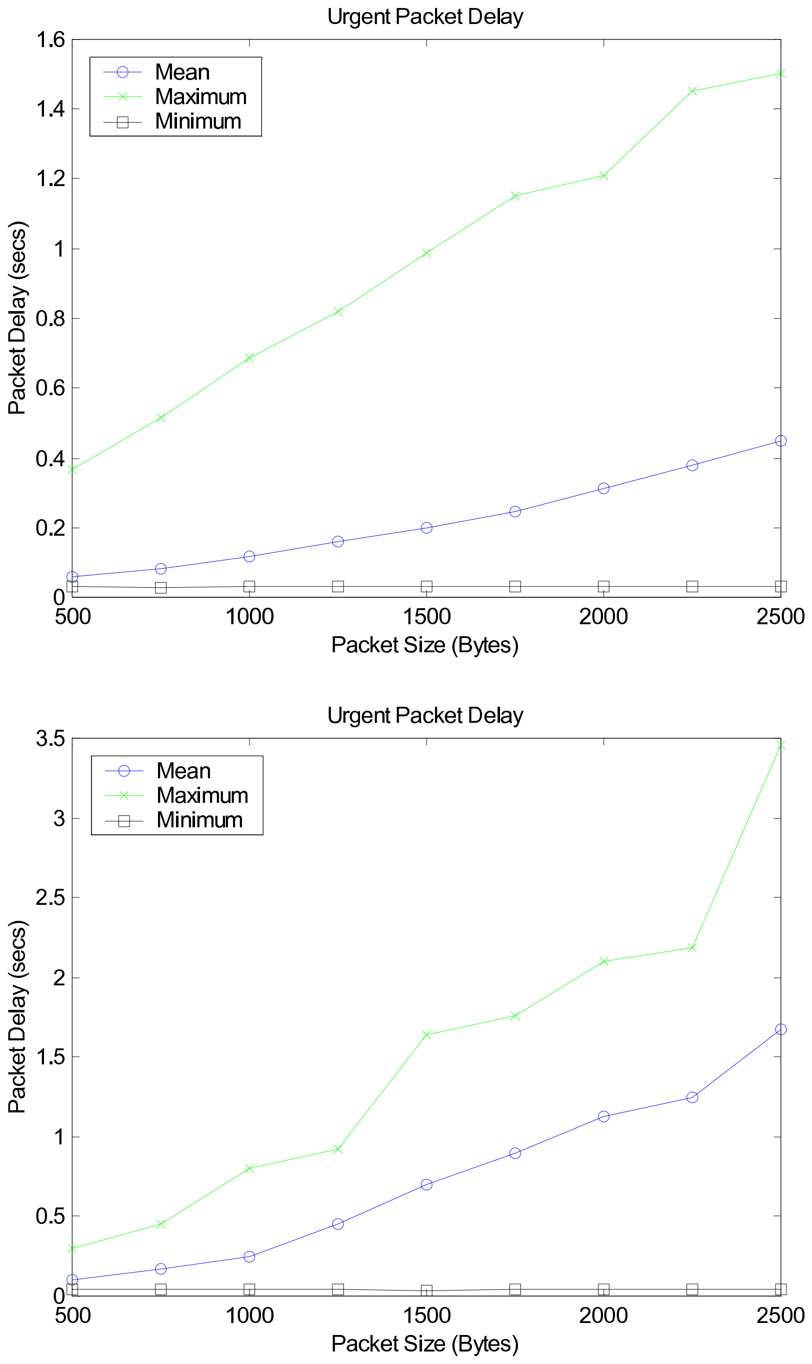


Figure 4.20: Packet delay varying packet size for urgent message (a) IEEEE 118 bus system (b) IEEEE 300 bus system

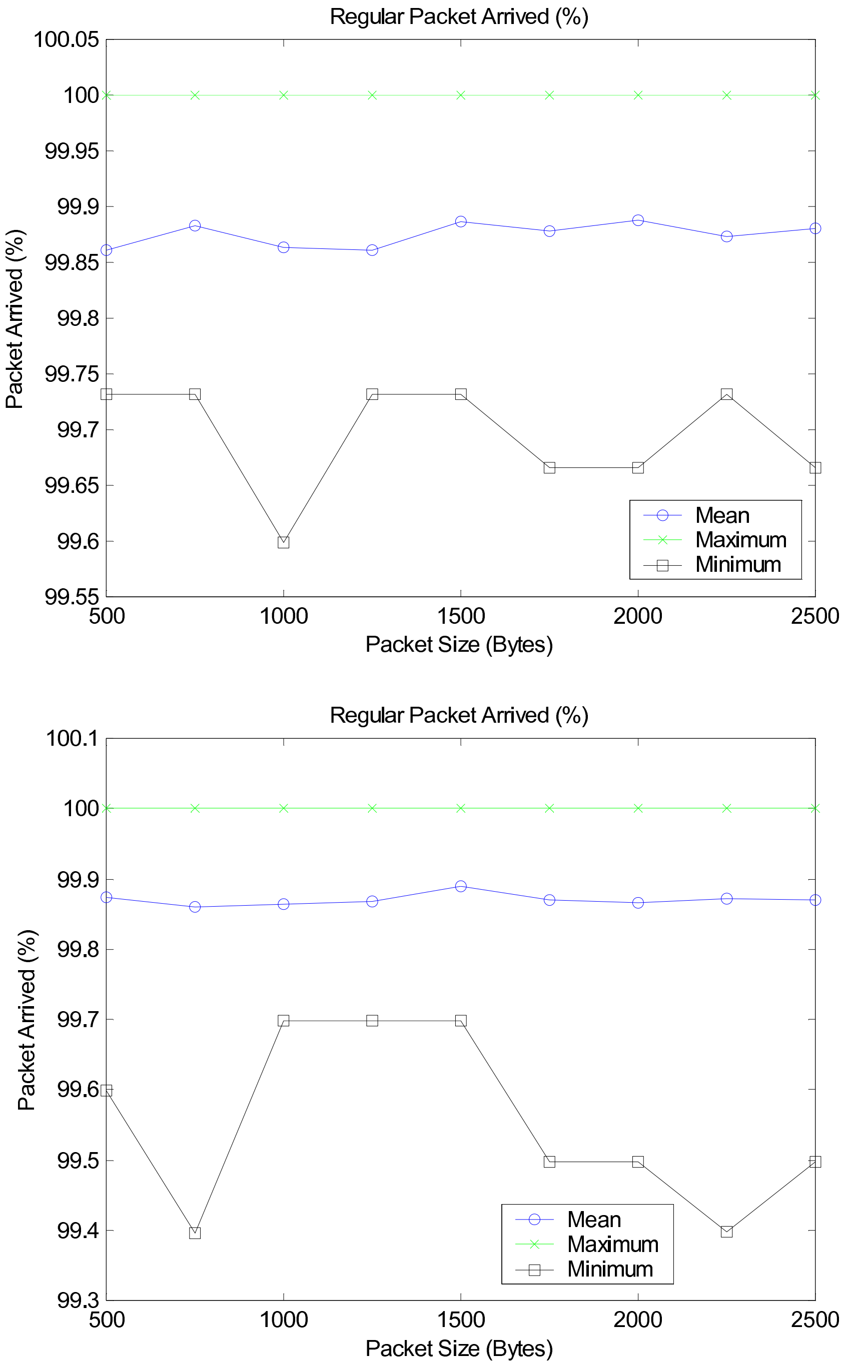


Figure 4.21: Packet arrival percentage varying packet size for regular message (a) IEEEE 30 bus system (b) IEEEE 57 bus system

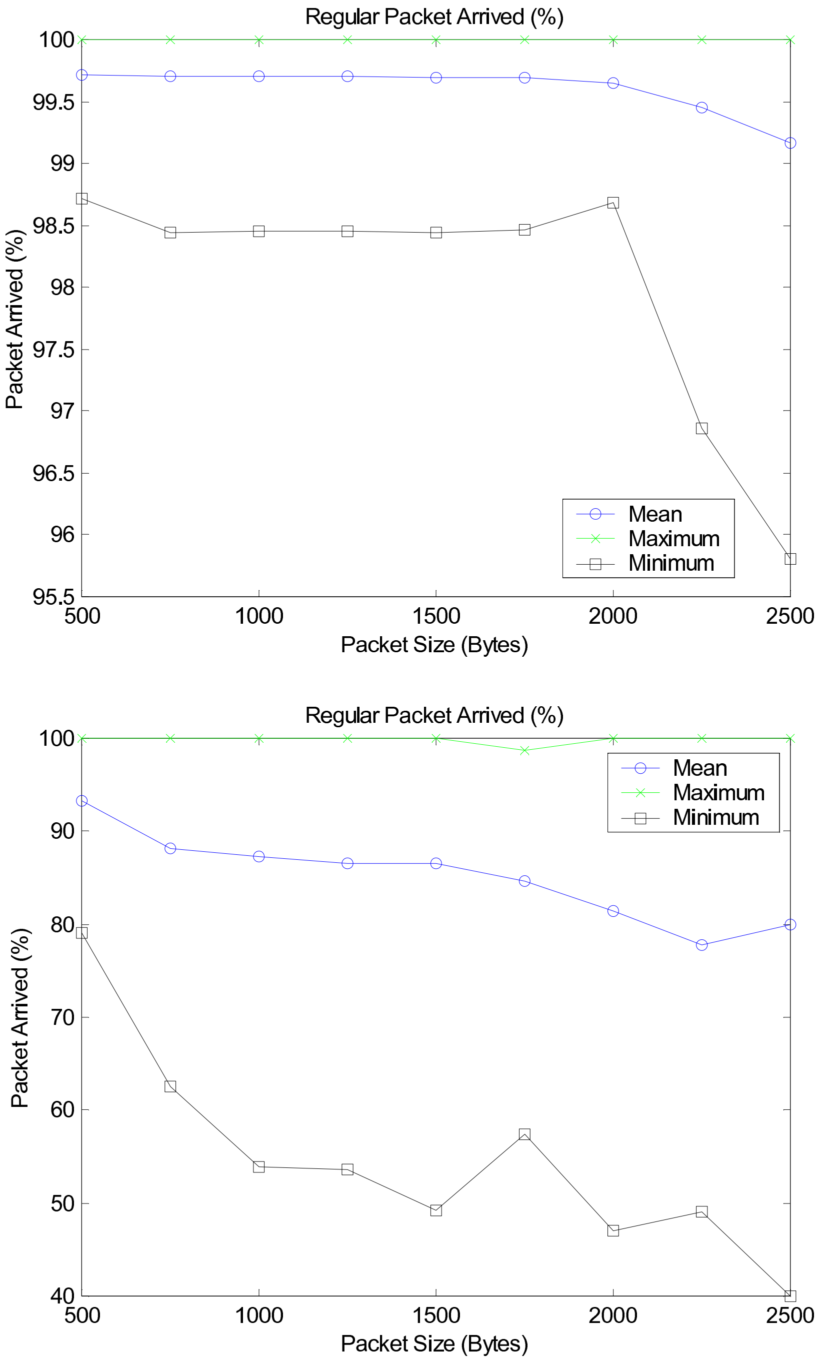


Figure 4.22: Packet arrival percentage varying packet size for regular message (a) IEEEE 118 bus system (b) IEEEE 300 bus system

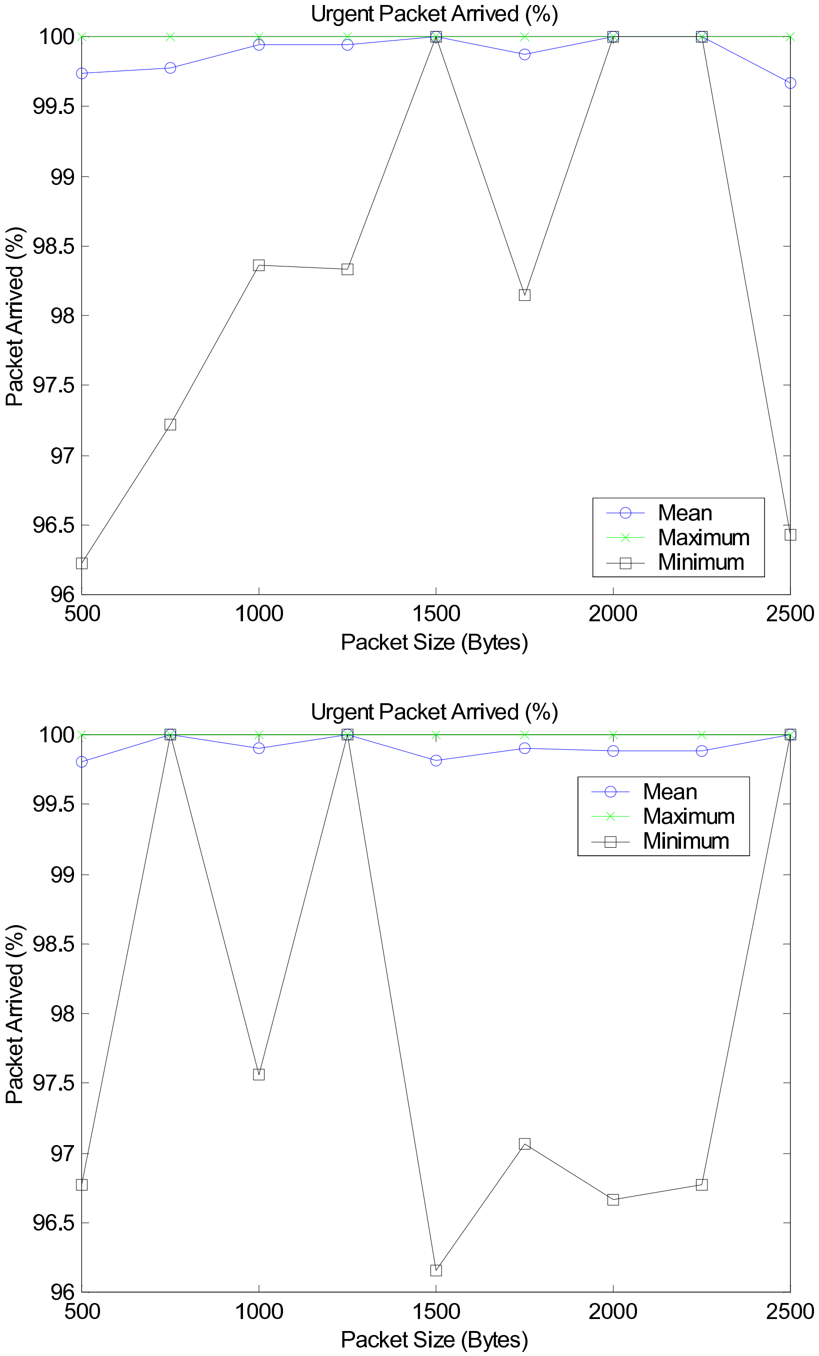


Figure 4.23: Packet arrival percentage varying packet size for urgent message (a) IEEEE 30 bus system (b) IEEEE 57 bus system

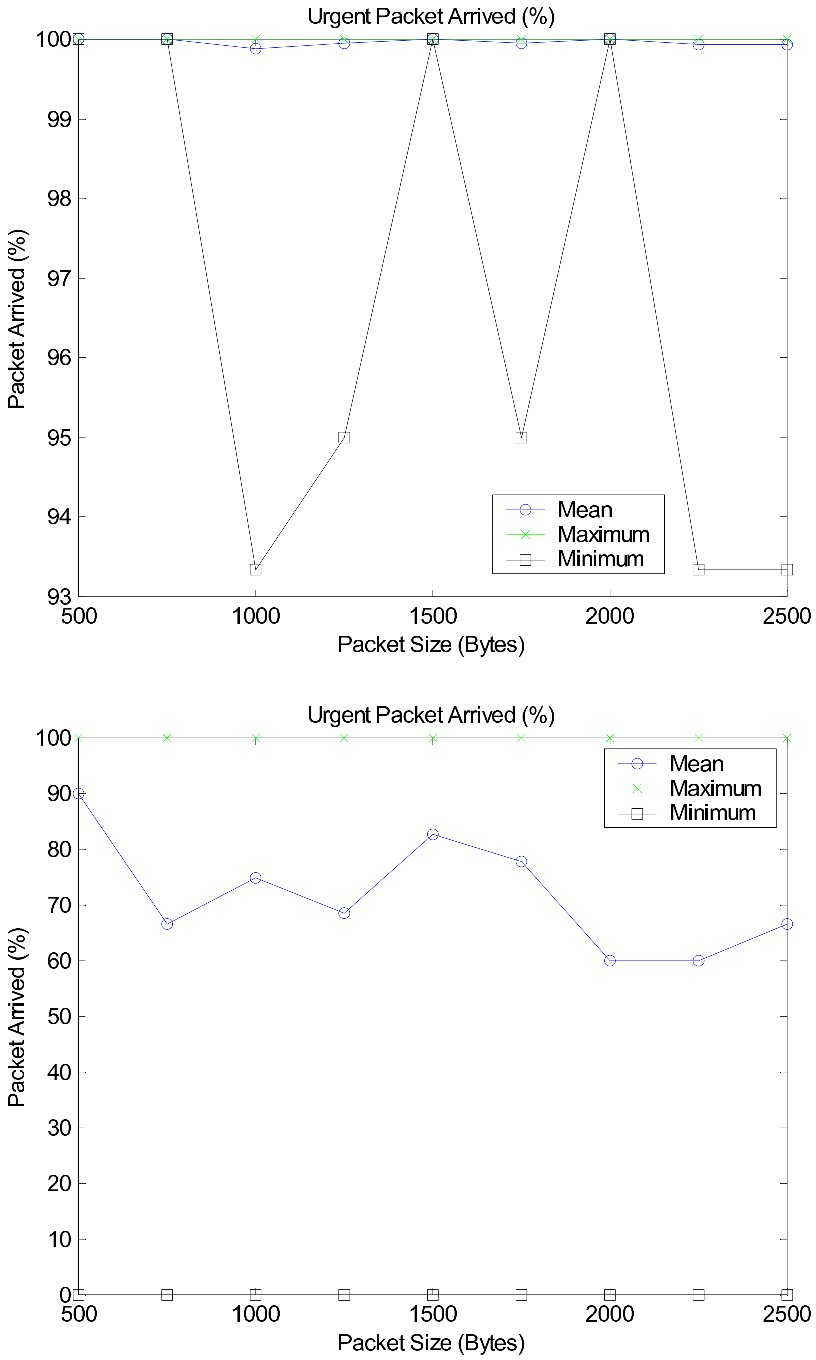


Figure 4.24: Packet arrival percentage varying packet size for urgent message (a) IEEEE 118 bus system (b) IEEEE 300 bus system

exists. Redundancy equals to 2 means that each subscriber has two different service access points and has redundancy. The bandwidth for all channels used here is 10 Mb. One can find that the packet delay is decreasing when increasing the error rate. When we have a higher loss rate, since we build our system on UDP and only have ACK in some cases, there are actually less packets flowing though the channel. The result is that we have less packet delay. Introducing redundancy will increase packet delay since it increases the number of packets flowing in the channel.

Packet Arrival Percentage

Figures 4.29, 4.30, 4.31 and 4.32 are the packet arrival percentage of the regular and urgent messages when varying the packet loss rate and redundancy. The bandwidth for all channels used here is 10 Mb. As expected, one can find that the packet arrival rate drops as the error rate increases. The packet loss becomes significantly when the error rate exceeds 1e-3. By introducing redundancy, the arrival packet percentage rises. With redundancy, we have higher packet arrival percentage, less packet lost. This is particularly true when the error rate is high.

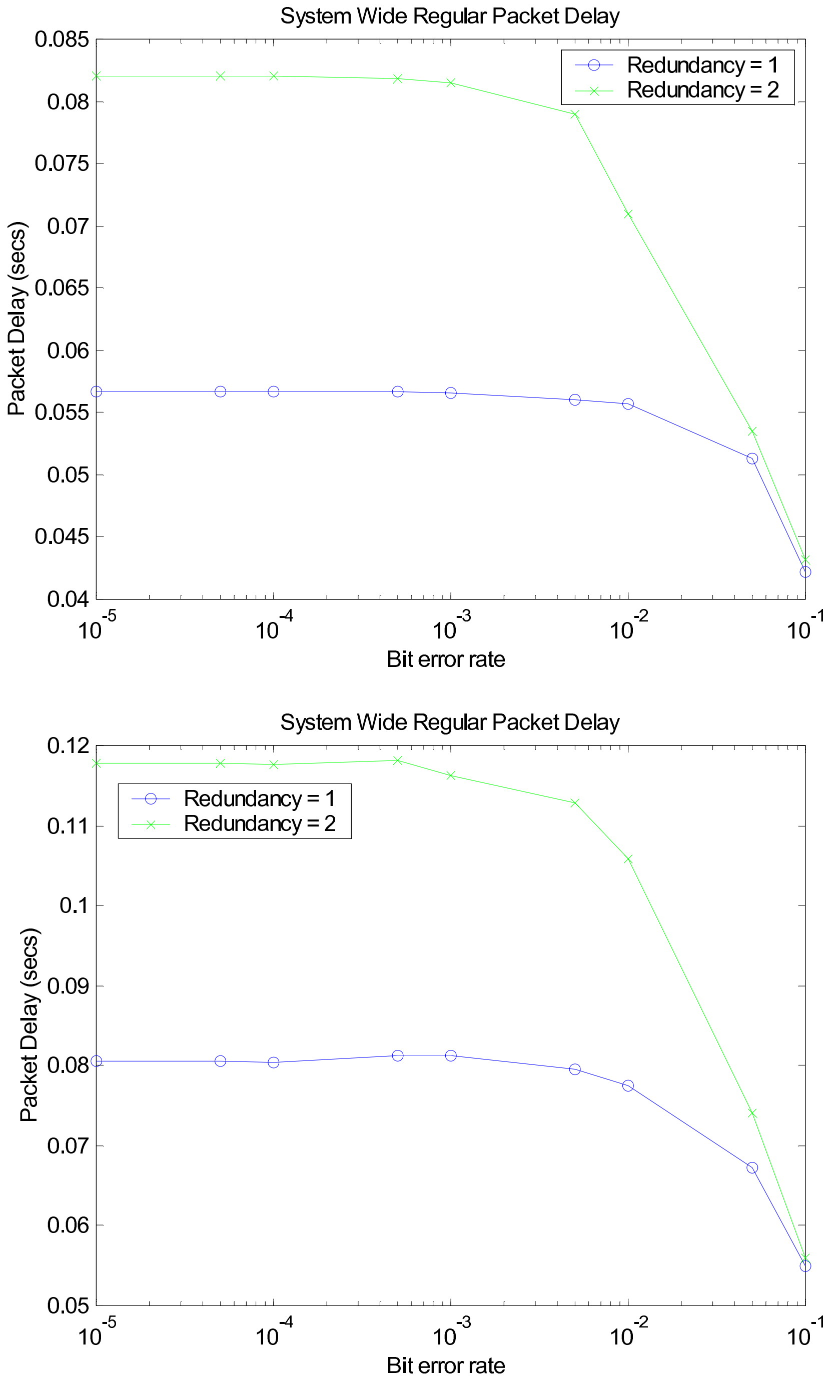
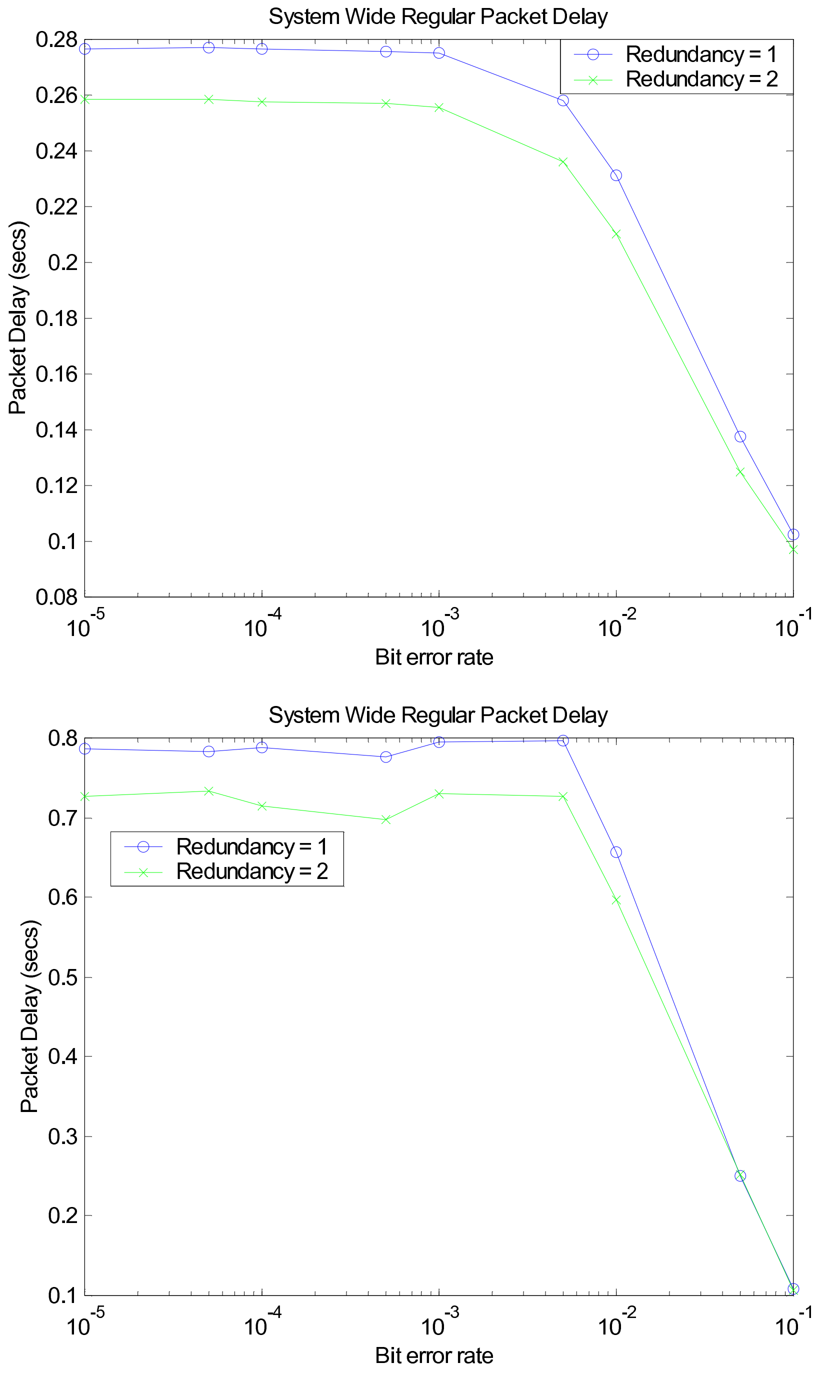


Figure 4.25: Packet delay varying packet error rate and redundancy for regular message (a) IEEEE 30 bus system (b) IEEEE 57 bus system



System Wide Regular Packet Delay

System Wide Regular Packet Delay

Figure 4.26: Packet delay varying packet error rate and redundancy for regular message (a) IEEEE 118 bus system (b) IEEEE 300 bus system

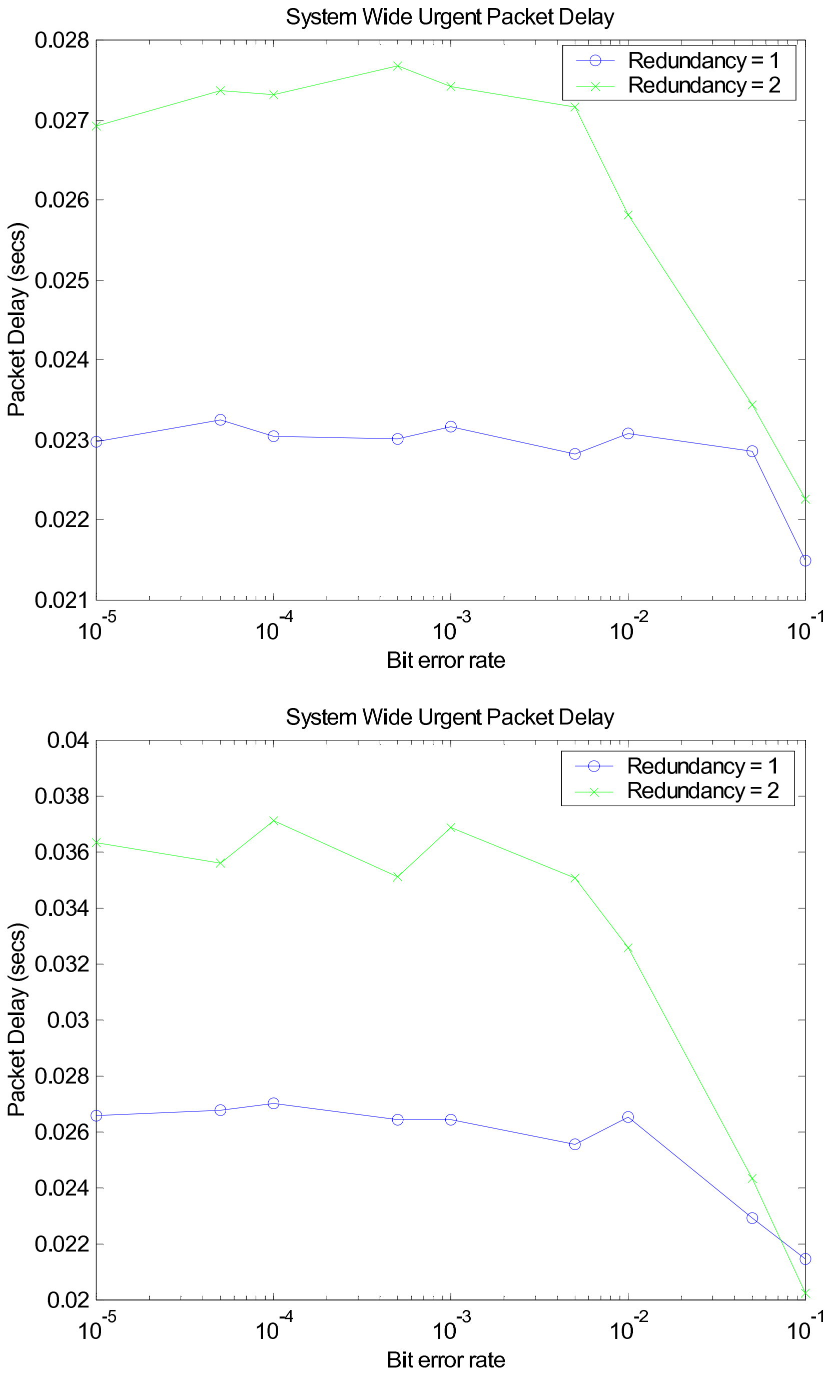


Figure 4.27: Packet delay varying packet error rate and redundancy for urgent message (a) IEEEE 30 bus system (b) IEEEE 57 bus system

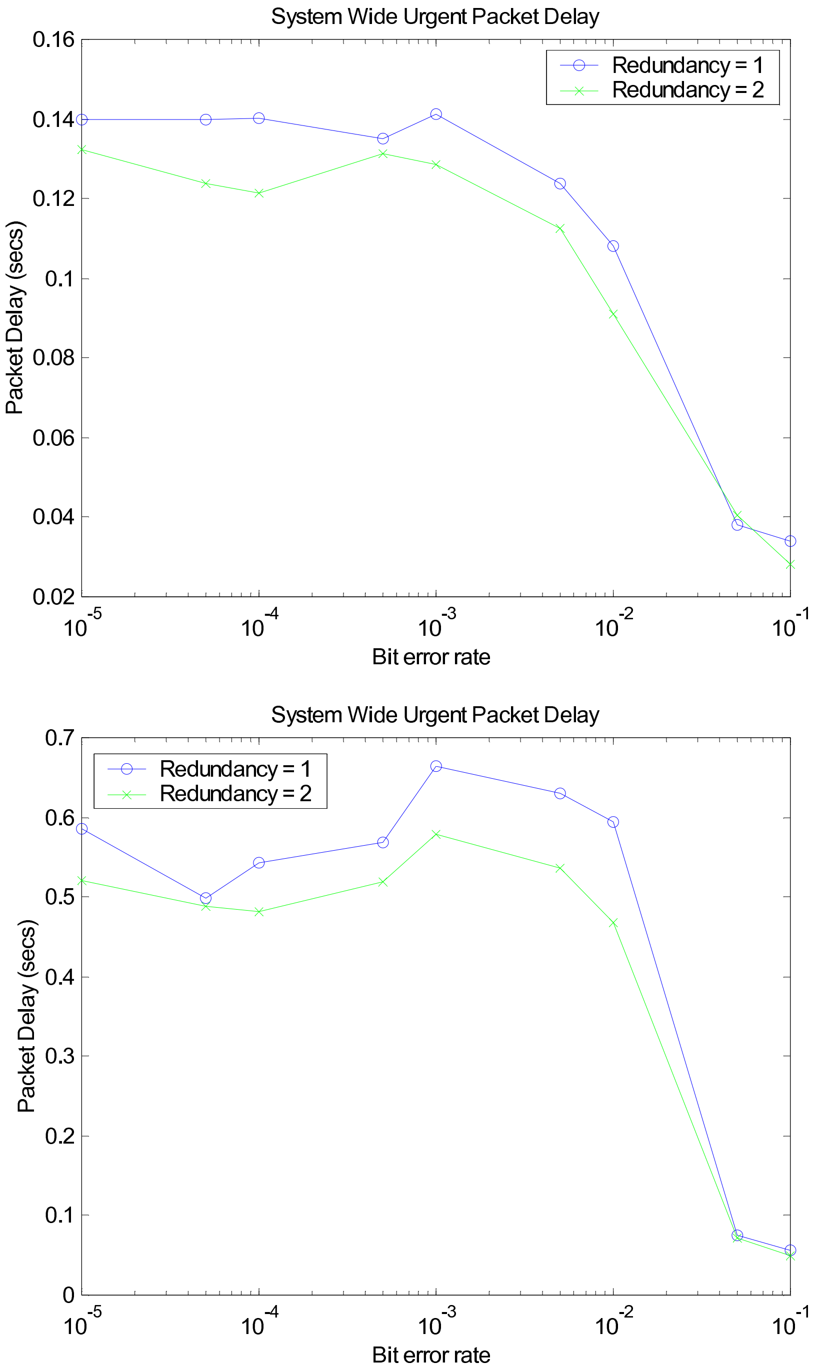


Figure 4.28: Packet delay varying packet error rate and redundancy for urgent message (a) IEEEE 118 bus system (b) IEEEE 300 bus system

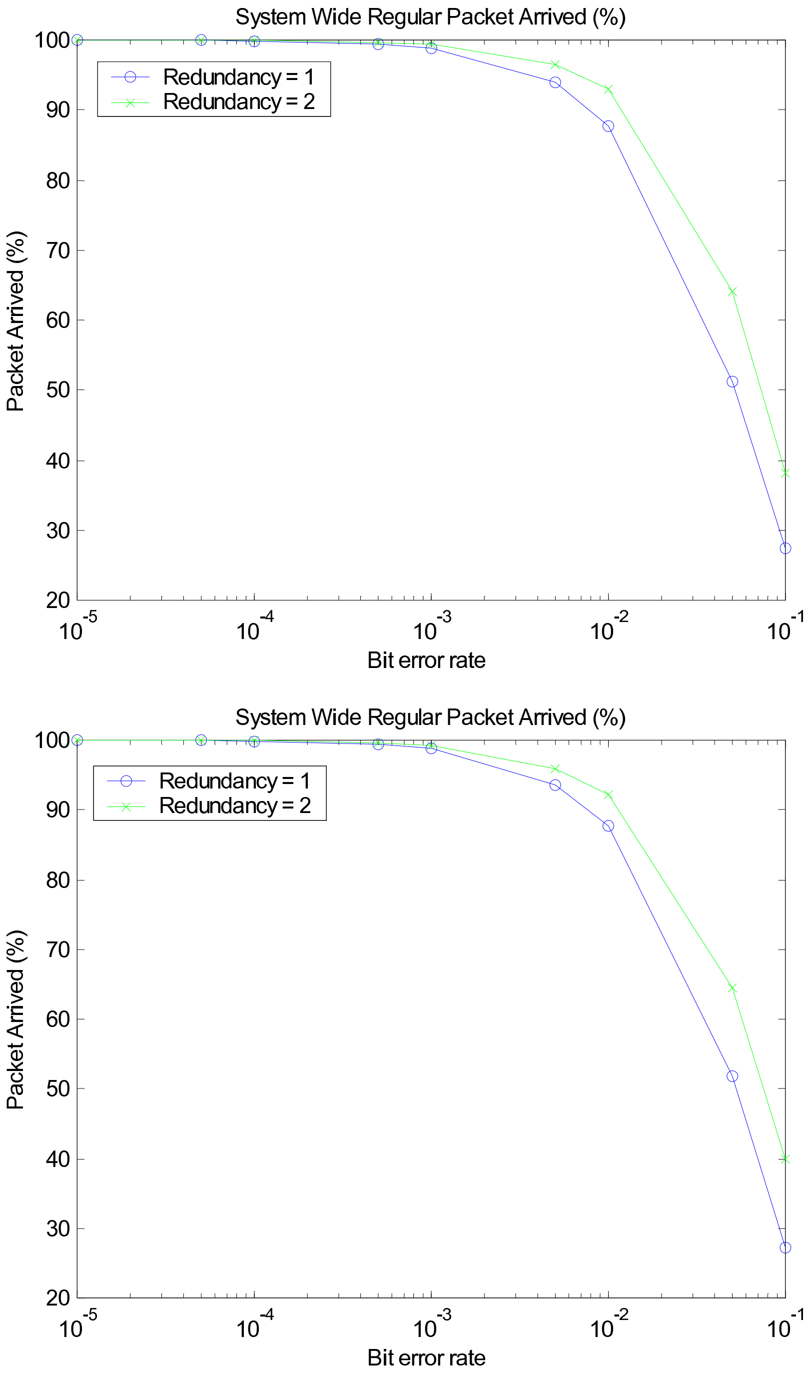


Figure 4.29: Packet arrival percentage varying packet error rate and redundancy for regular message (a) IEEEE 30 bus system (b) IEEEE 57 bus system

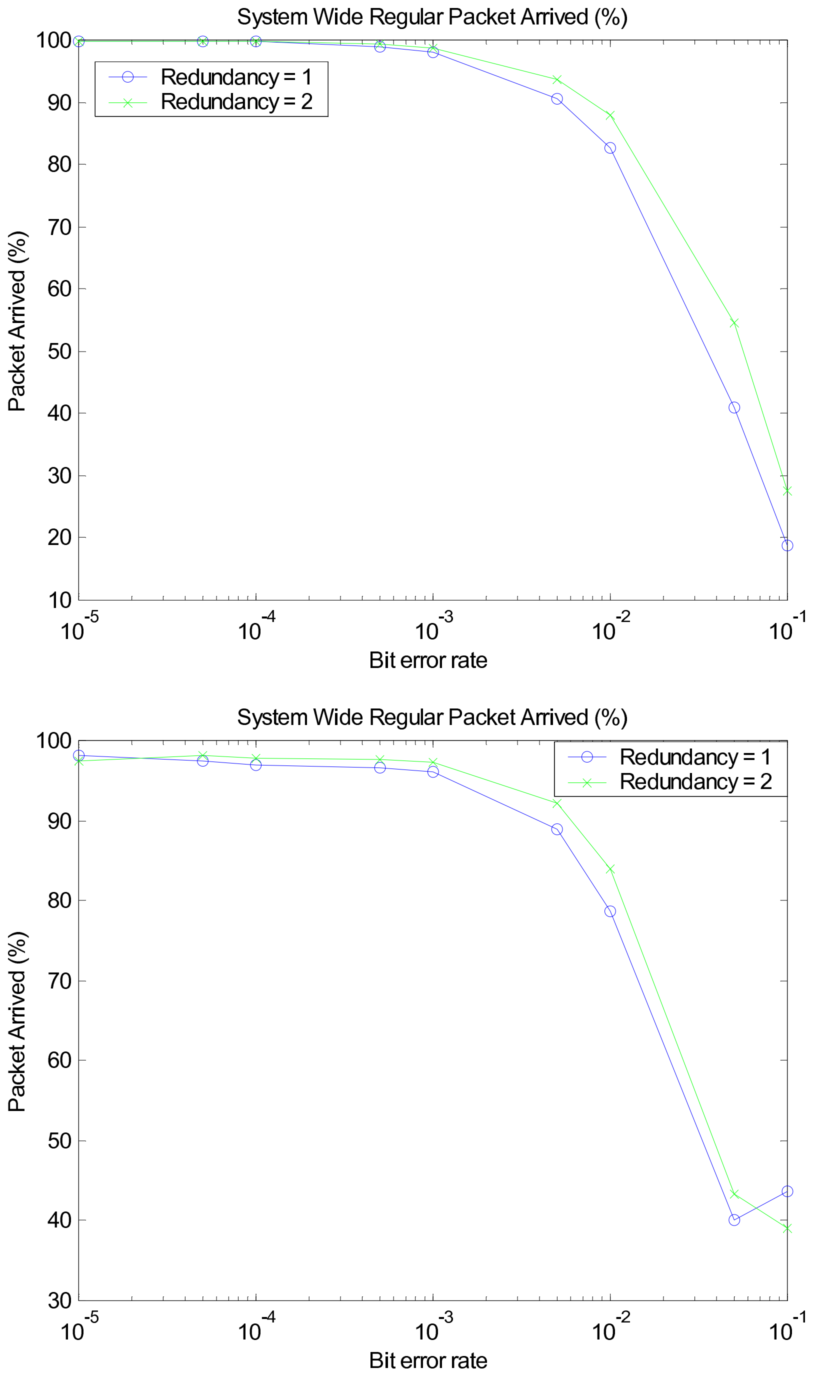


Figure 4.30: Packet arrival percentage varying packet size for regular message (a) IEEEE 118 bus system (b) IEEEE 300 bus system

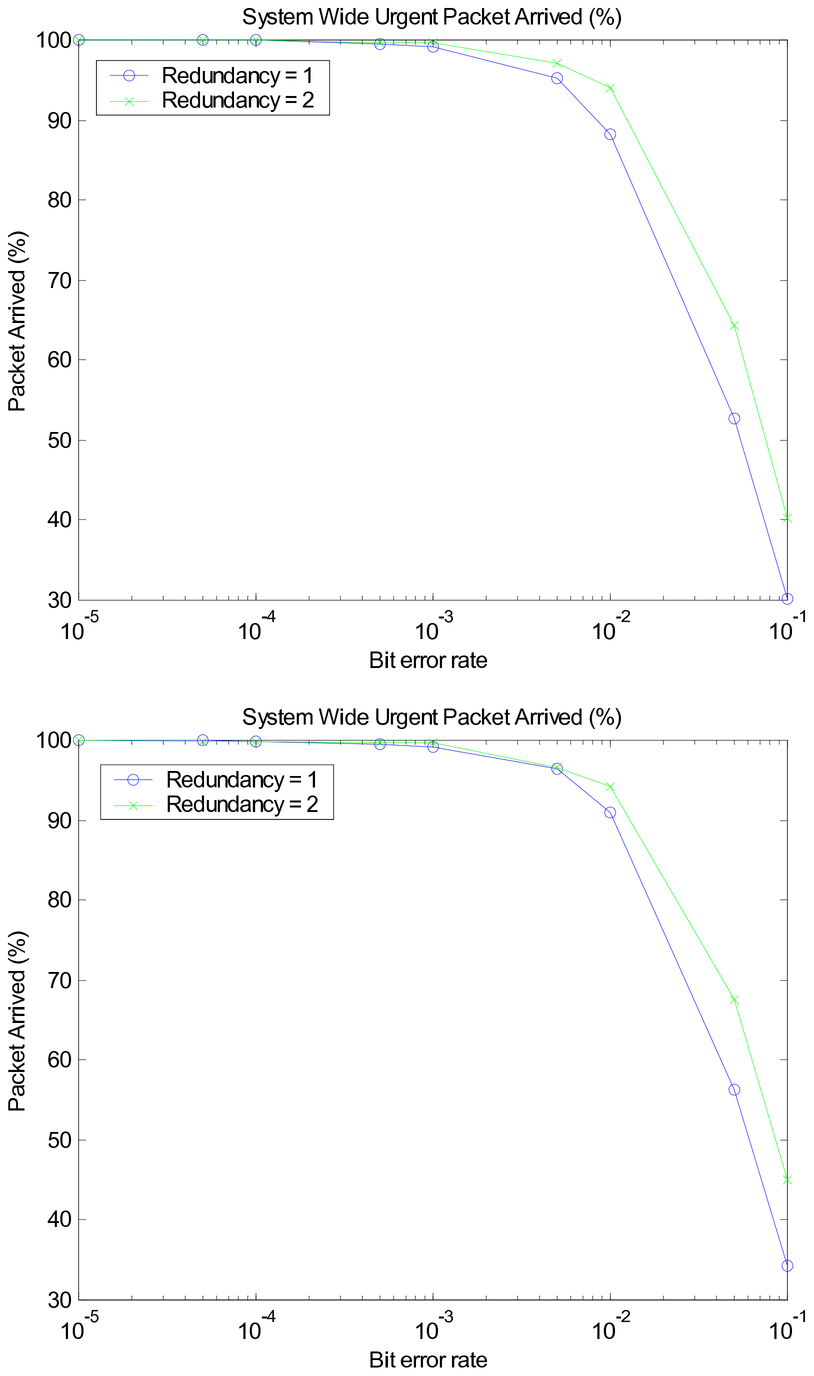


Figure 4.31: Packet arrival percentage varying packet error rate and redundancy for urgent message (a) IEEEE 30 bus system (b) IEEEE 57 bus system

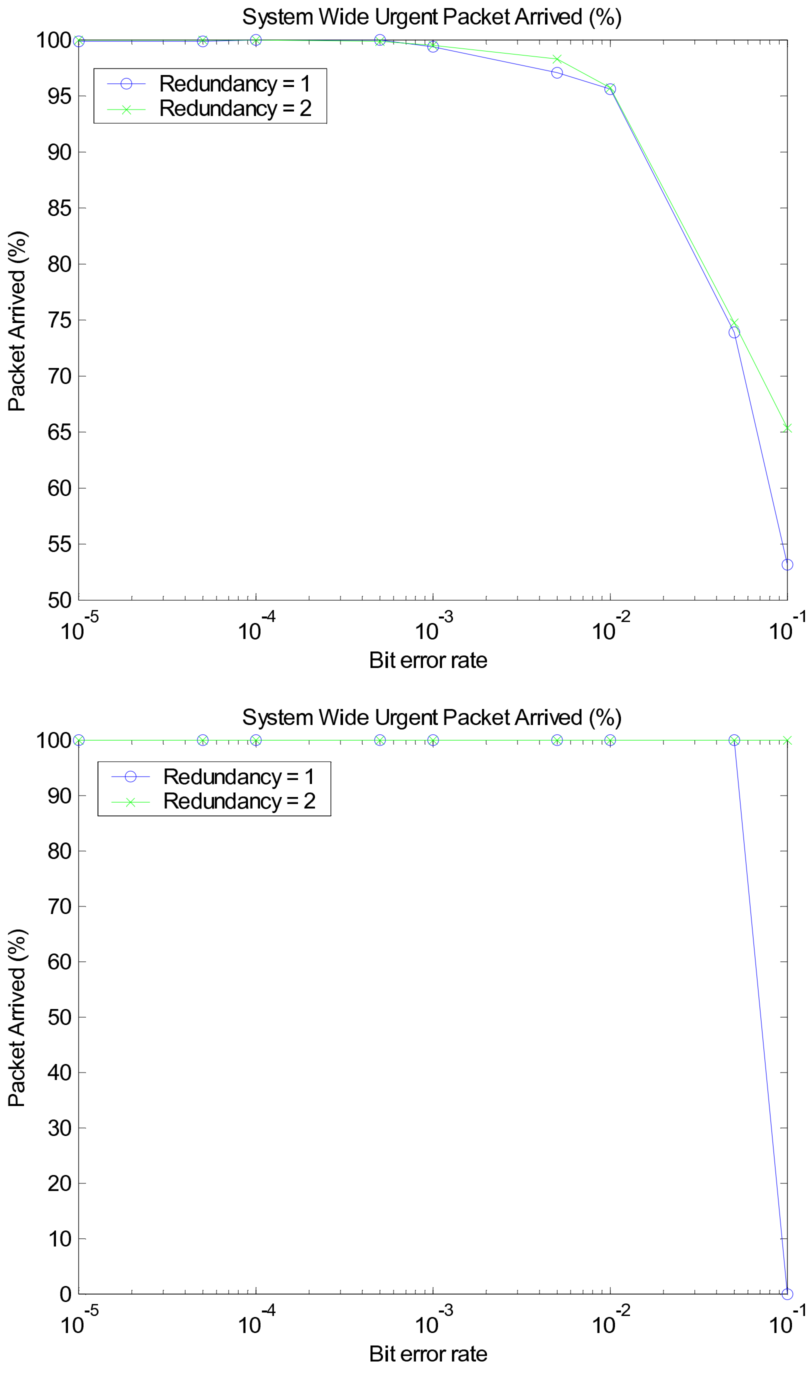


Figure 4.32: Packet arrival percentage varying packet error rate and redundancy for urgent message (a) IEEEE 118 bus system (b) IEEEE 300 bus system

4.6 Conclusion

From the simulation, we find that the system performance is satisfactory if given adequate bandwidth. The delay ranges from 0.05 to 1.5 seconds, which is sufficient for monitoring and supervisory control. We also find that the delay grows when the system size becomes larger. This may be caused by some bottleneck in the system. Thus, carefully plan the network topology and bandwidth is important, especially for large system. In addition, the assumption that all nodes subscribe to all information is not realistic or necessary for very large systems. Other technologies, such as data caching, merging or fusion, can reduce the traffic on the network thus improve the system performance.

Chapter 5

Distributed State Estimation

5.1 Introduction

State estimation is a fundamental component in power system network security analysis. In modern Energy Management Systems (EMS), the state estimation is executed as snapshots, typically once every several minutes. The desire to improve market efficiency and prevent cascading failures is driving two recent trends in state estimation: first, increasing the frequency of execution in order to more closely track system state; and second, covering a much wider area network with the estimates. Both these trends pose challenges on the existing state estimation calculation. At the same time, the growing availability of low cost sensors with communications and computational capability offers promise for improved estimates.

Higher frequency of execution requires fast state estimation algorithms while large system calls for more numerically stable algorithms, particularly with regard to erroneous data. Traditional centralized state estimation is quite mature and relatively stable at this moment and significant improvements in the algorithm are unlikely. Still, with the increase of network size, numerical stability poses a serious problem, since the system matrix can become more easily ill-conditioned and both computational speed and convergence suffer.

A natural way to increase speed and avoid the “curse of dimensionality” is parallelism and distributed processing. The natural geographic distribution of power system measurements can benefit from a similarly decentralized information architecture, where remote processors perform local state estimation and the result is sent back to control center to refine the calculation. The local estimate can be continuously updated and used for local control purposes.

With the recent quantum leap of information technology, especially in the communication area, the distributed processing is more feasible. Traditional power systems use a centralized information model with all information passed to the control center. In the distributed power system information model, one uses a publish/subscribe paradigm. Each power device entity can be a publisher of its own data while users, such as a traditional control center and substation can subscribe to desired portions of the data. This makes it possible for a local processor to carry out the state estimation for a specific area, since the local processor can access the data required to perform a localized state estimation calculation.

In this dissertation, based on the newly proposed distributed publish/subscribe communication model, we present an asynchronous distributed state estimation that takes the advantage of the conventional state estimation and the flexibility of the communication network. The entire network is covered by a set of overlapping areas, in which conventional state estimation is performed on each area independently. That is, instead of the passing the result of each iteration back to the central processor, the result is held locally until the convergence is reached. The central processor consolidates the results to ensure the consistency on the bus states in the overlapping area. Numerical experiments show that the proposed method is sufficiently accurate for typical applications, computationally efficient and more robust with respect to data errors.

5.2 Related Work

5.2.1 Distributed State Estimation

There are many existing parallel and distributed state estimation algorithms. In most approaches, the problem is formulated by first dividing the network into several small areas. For each of these areas, a local optimization problem is solved and constraints are placed on the boundary buses to ensure the consistency of the bus states. This approach can be termed a synchronous method, since each iteration of a local state estimation must be coordinated with the other areas. Among these methods are the earliest attempts using so-called hierarchical methods [33], which depend on a star or master-slave functional style and communication network. Separated processors calculate local iterations and the result of each iteration is passed up to a central computer for further processing. The synchronized result is passed back to each local processor for the next iteration. This approach suffers from an inherent reliability and performance problem. The central master processor becomes a bottleneck and parallelism cannot be further exploited since each processor must wait for all others to complete their computations. Recent approaches are based on parallel processing for the linear system equations associated with nonlinear state estimation algorithms or optimization decomposition techniques. Lin [34] introduced dual recursive quadratic programming but this method suffers from poor computational performance. Falcao, et al [35] proposed a method based on the conventional state estimation algorithm using a coupling constraints optimization technique. Baldick, et al [36] proposed an approach using an optimization technology based on the "Auxiliary Problem Principle" [37]. Both of these approaches require data exchange between different local state estimators during iterations. The object in this work is to minimize such communication requirements. Huang, et al [38] used a similar approach as developed here that relies on overlapping areas. Still, their emphasis is on the measurement exchange between utilities and lacks a detailed analysis of the communication network needed to support their approach.

5.2.2 Time Skew Problem

Time skew is potentially a problem in a distributed approach since there are more stages in the communication. Several papers address time skew problems in traditional state estimation, e.g., [39, 40]. Su and Lu [40] used a stochastic Extended Kalman Filter to reduce errors. Dabbaghchi [39] studied the time skew problem on practical state estimators. Based on several experiments on the American Electric Power system, he shows that the delay within 15 minutes is tolerable and concludes that the impact of time skew is minimal if the delay is within several minutes. Data exchange between different utilities is feasible at this rate. Such analysis does not preclude the difficulty in reconstructing events following outages where the timing of certain events may be important. Still, based on this study, we assume that given the communication requirements in the proposed formulation, time skew is not a problem.

5.3 Algorithms

5.3.1 Conventional State Estimation

The nonlinear equation relating the measurements z and the state vector x is:

$$z = h(x_{true}) + \epsilon \tag{5.1}$$

where z is a ($m \times 1$) measurement vector, x_{true} is a ($n \times 1$) true state vector, $h(.)$ is a ($m \times 1$) vector of nonlinear function, ϵ is ($m \times 1$) measurement error vector with zero mean and covariance matrix R. Conventional least square (WLS) state estimation is to find $\hat{x}$ that minimizes:

$$J(x) = ||z - h(\hat{x})||_2 \tag{5.2}$$

If the system is observable, the Gauss-Newton iteration scheme can be used to solve the nonlinear optimization problem:

$$\Delta x^i = [H^T R^{-1} H]^{-1} H^T R^{-1} (z - h(x^i)) \tag{5.3}$$

with H is the Jacobian matrix of $h(x)$, $H = \partial h / \partial x$, $x^{i+1} = x^i + \Delta x^i$ and x^i is the result of i^{th} iteration. More details on this formulation are widely available, e.g., [41].

5.3.2 Asynchronous Distributed State Estimation

The object of the proposed asynchronous distributed algorithm is to combine the advantage of the existing conventional state estimation and at same time introduce distributed processing to take advantage of today's communication capacity. The approach assumes a distributed communication scheme based on the publisher/subscriber model described elsewhere in this thesis.

In this algorithm, the network is partitioned into several overlapping areas. Each area has its own local processor and subscribes to the measurements published in that area and possibly neighboring regions. The areas overlap not just at the boundary buses as in most distributed state estimation but possibly over significant areas. This means more than one local processor will subscribe to the information published by some measurements. The overlapping areas serve two purposes. First, the results from two different local estimators on the overlapping area can be used in the final stage to reduce the discrepancies. Second, bad data detection and identification near or on the boundary buses can be employed as in traditional distributed state estimations.

Each area performs state estimation individually until obtaining convergence within a specified tolerance. The result is passed to a central processor, which consolidates the results from all areas into a complete network result. If an individual estimator fails to converge, it can be ignored in the overall network estimate. It is also feasible for the estimator to expand or shrink coverage in such a case by changing its subscription data. This introduces new problems with adequate coverage and is not addressed here.

The algorithm steps are:

1. Overlapping areas are partitioned from the full network. This can be performed centrally or by the local estimators and can be dynamic.

2. Multiple local estimators distributed in the different areas are executed simultaneously and asynchronously until they converge individually to the

desired tolerance. There is assumed to be sufficient computational power available at some substation within a region to perform the computation.

3. Each local estimator runs bad data analysis individually.
4. Based on the results of local estimators, determine the state of the full system according to the different accuracy and reliability of the individual estimators.

Partition the Network

The distributed state estimation divides the network into several overlapping areas. The partition must be:

- Complete. Each bus is included in at least one area. This means that every measurement in the network is included in the estimate.
- Connected. Each area has some overlap with neighboring area(s). Consider the Area Connectivity Graph (ACG). Each area is a node in the figure 5.1 and an edge between two nodes in the graph is present if those two areas have overlap. This criterion states the partition should be a connected ACG. The primary need for this condition is to ensure a common angle reference between different areas.

An important consideration in the partition is bad data processing. In traditional distributed state estimation, bad data near/on the boundary buses is difficulty to detect and identify. By enlarging the overlapping area, this difficulty can be alleviated to the degree that redundancy exists within an area. Still, bad data in the overlapping areas will plague the estimation result of all areas that include that bad

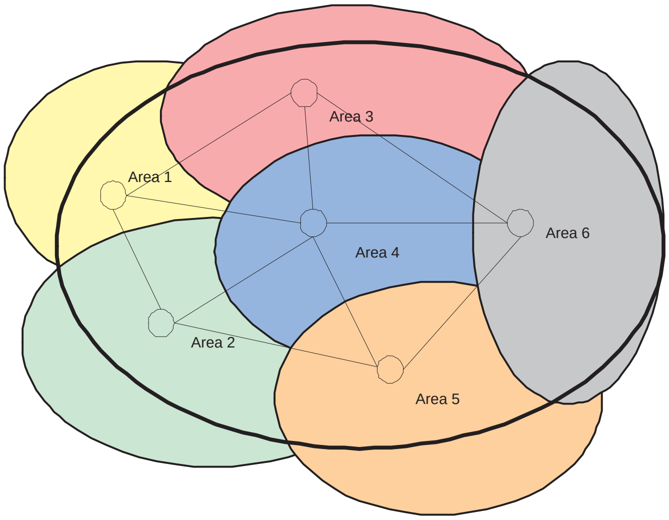


Figure 5.1: Partition the Whole Network and Area Connectivity Graph

measurement. Greater local measurement redundancy will reduce the impact of bad data.

Asynchronous Local State Estimation and Bad Data Processing

Existing conventional centralized state estimation algorithm can be used. The only difference that arises is in needing a reference bus for each area. There are numerous mature and numerically stable algorithms available. Here, a fast decoupled Newton algorithm with orthogonal transformation to improve the numerical stability is employed.

Complete Network State

Since the state estimations performed in different areas use different reference buses, there are discrepancies in the bus states on the overlapping areas. The major task for the central processor is to determine the bus angle difference between different areas. There will, of course, be differences in voltage magnitudes but these differences are related to a common reference and can be managed statistically. The following describes the methodology to consolidate the estimates:

1. Since any two areas may use different reference buses, the angle difference needs to be calculated between these arbitrary references. The angle difference is found by using the two different state estimation results on the overlapping boundary buses. A simple statistical weighting scheme based on the estimation accuracy of different local estimators can be formulated as:

   $$\Delta\theta_{AB} = \sum_{i \in O} (\theta_i^A - \theta_i^B)(g_{ii}^A + g_{ii}^B) / \sum_{i \in O} (g_{ii}^A + g_{ii}^B) \tag{5.4}$$

   where

   - $\Delta\theta_{AB}$ is the angle difference of reference buses between local estimator A and B,
   - O is the set of all the overlapping buses of estimator A and B,
   - θ_i^A is the estimated angle on bus i in estimator A,
   - g_{ii}^A is the i^{th} diagonal element of gain matrix G of area A , which is the inverse of the covariance matrix of the state estimation accuracy.

2. Select a reference bus of one estimator as the global reference bus for the grid.

3. Determine the angle difference between this global reference bus and reference bus in every local estimator. This can be implemented by traversing the area connection graph. The depth-first graph traverse algorithm serves such purposes well. Since angle difference has the transitivity:

   $$\Delta\theta_{AC} = \Delta\theta_{AB} + \Delta\theta_{BC} \tag{5.5}$$

   By using (5.5), the areas that do not have overlap with the global reference area can be calculated.

4. The estimated angle of each local estimator will be adjusted by the angle difference between the global reference bus and the local difference bus.

5. For non-overlapping buses, the state variables of bus voltage magnitude are determined by the current estimation result in local estimators.

6. For overlapping bus i belonging to multiple local estimators j, $j \in S$, the state variables x_i are found as:

   $$x_i = \sum_{j\in S} x_i^j g_{ii}^j / \sum_{j\in S} g_{ii}^j \tag{5.6}$$

   where

   x_i^j is the state variable of bus i in local estimator j.

   g_{ii}^j is the i^{th} diagonal element of gain matrix G of area j, which is the inverse of the covariance matrix that measures the result of the state estimation accuracy.

Bad Data Detection and Identification

Each local area performs bad data processing independently. Traditional distributed state estimation has difficulty with detection and identification of the bad data on the boundary buses. If the overlapping areas have sufficient redundancy, the problem is alleviated. Generally, speaking the local estimator should subscribe to a region of data sufficiently large to assure reasonable robustness with respect to bad data. Too small of an area will leave the system more vulnerable to bad data while too large of an area will make the system more vulnerable to numerical instabilities as well as taxing the communication network. No attempt has been made to optimize this tradeoff here.

5.4 Numerical Experiments

The IEEE 30-bus system (Figure 5.2 [31]) described elsewhere in this thesis is used here as an explanatory example to verify the feasibility and accuracy of the proposed algorithm. Assume there are 5% Gaussian errors relative to the true value on all measurements. In this example, the network is partitioned into 4 areas. Areas A, B, C, and D include buses (1-7), (5-11), (6, 9-22), and (15, 21-30), respectively. The areas were chosen empirically and one can easily verify completeness and connectivity. Fast decoupled state estimation is performed on each area. In each area, the lowest numbered bus serves as the reference bus for that area. Results are shown in Table 5.1.

Following the local convergence, the results of each area are assembled together, starting from area A, then B, C, D - the ordering is not important. For comparison, a traditional fast decoupled state estimation is performed on the full

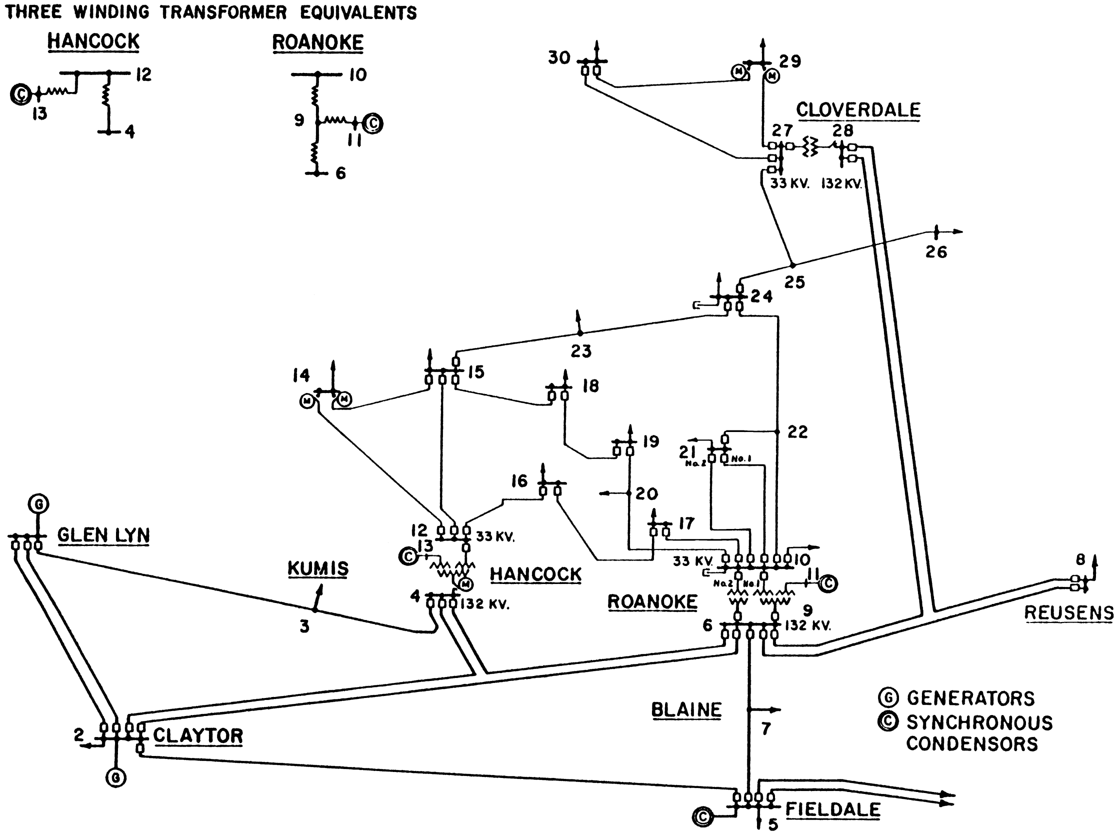


Figure 5.2: IEEE 30-Bus Testing System

Table 5.1: Result from Local Estimators

Area A (1-7)	Angle	(0, -0.0929, -0.1310, -0.1615, -0.2461, -0.1927, -0.2240)
	Magnitude	(1.0600, 1.0431, 1.0208, 1.0118, 1.0102, 1.0103, 1.0025)
Area B (5-11)	Angle	(0, 0.0555, 0.0229, 0.0425, 0.0024, -0.0253, 0.0024)
	Magnitude	(1.0092, 1.0103, 1.0019, 1.0101, 1.0509, 1.0451, 1.0819)
Area C (6, 9-22)	Angle	(0, -0.0512, -0.0780, -0.0512, -0.0645, -0.0645, -0.0801, -0.0815, -0.0749, -0.0810, -0.0925, -0.0956, -0.0921, -0.0859, -0.0856)
	Magnitude	(1.0102, 1.0510, 1.0450, 1.0823, 1.0570, 1.0710, 1.0421, 1.0376, 1.0444, 1.0400, 1.0282, 1.0257, 1.0298, 1.0327, 1.0332)
Area D (15, 21-30)	Angle	(0, -0.0033, -0.0031, -0.0063, -0.0093, -0.0015, -0.0083, 0.0077, 0.0745, -0.0135, -0.0291)
	Magnitude	(1.0377, 1.0328, 1.0333, 1.0272, 1.0216, 1.0171, 0.9992, 1.0230, 1.0065, 1.0033, 0.9918)

network. The comparison is shown in table 5.2, where estimation error is defined as $\sqrt{\sum_i (x_i - \hat{x}_i)^2}$. From table 5.2, one finds that the accuracy of the proposed algorithm is comparable to the traditional full network state estimation.

Table 5.2: Comparison of Estimation Error

Algorithm	Distributed State Estimation	Traditional State Estimation
Estimation Error	0.0145	0.0099

5.5 Conclusion

State estimation is a fundamental component in power system network security analysis. In modern Energy Management Systems (EMS), state estimation is executed as snapshots, typically once every several minutes. This chapter has reported on a new approach to state estimation based on the proposed publisher/subscriber communication infrastructure. This infrastructure affords an opportunity to greatly improve the existing approaches to system security.

The proposed algorithm is based on this new distributed information structure. The algorithm is asynchronous and avoids a central controlling node during the iteration, thus, improving system wide robustness. Each local area performs the state estimation independently and the final result is assembled in a central processor. It is not expected that the local estimators, when combined in this fashion, can find the global maximum likelihood estimate. Still numerical tests verify that the new algorithm is sufficiently accurate compared with existing conventional full network state estimation algorithms. Moreover, the real challenge for modern state estimators is robustness with respect to missing and bad data. Such a distributed approach is far superior in this regard.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

The introduction of open markets and the fast pace of changes brought by modern information technology bring both opportunities and challenges to the electric power industry. Vast amounts of data are being generated by the extensive deployment of new recording devices, as well as the need for new business data, such as, market trading history, bidding information, and so on. In addition, fast and low cost communications allow the data to be more widely accessed. Still, the legacy information systems may not make full use of the data produced. Inconsistency, inaccuracy and other problems plague the quality of the data. The large amount of data may decrease the performance of data processing in the control centers or RTOs, which are already heavily loaded. In addition, requirements for system security are receiving renewed focus following recent blackouts. Users need to be able to benefit from the ever increasing data.

The challenges discussed in the introduction:

Changes

- Increasing volume of data from both physical systems and business side such as power market.
- Greater complexity of data access requirements.
- Improved communications due to readily available of much higher bandwidth, reliable and low cost communication networks and equipments.
- More decentralized computation and intelligence. Substation automation is a typical example of this trend
- Renewed focus on reliability after several serious blackouts.

Problems

- Legacy data systems can not handle the huge amount of data properly and efficiently.
- Inadequacy of existing security analysis tools.
- Hard-wired communication networks make ISO/RTO control centers the single information consumer, while available computational power in the grid is wasted since they lack access to the needed data.
- Limited information for market participants.

To better integrate the ever increasing amount of data, data warehousing technology are introduced. At the same time, effort is made on providing the user higher quality data. A two stage DC state estimation is introduced to address the inaccurate topology data problem that hampers the electrical energy market traders. To take advantage of today's communication and information technologies, a distributed communication framework using publisher/subscriber paradigm and a distributed state estimation based on that framework is presented in the dissertation.

6.1.1 Data Warehousing

Data warehousing is a technology that can integrate huge amount of data. By using data warehousing technology, one can implement:

- efficient management of the huge amount of data available in the modern power system;
- a uniform view of the heterogeneous data, which masks the heterogeneous data source;
- a user-customizable the view appropriate for their own needs;
- data cleansing functions that reduce the amount of the inconsistency in the original data source; and
- new analytical tools for historical analysis and prediction of future trends.

Different from most of the data warehousing used in the business area, the data warehouse has to deal with a lot of data that comes from the physical system in the power system. In this dissertation, several examples are developed on how to construct a data warehouse that meets the requirement of the power industries. Methods to populate the data warehouse from several common power system data sources are discussed.

6.1.2 Two-stage DC State Estimation to Improve the Accuracy of the Network Topology Information

One of the main goals in introducing data warehousing was to give market participant simple useful tools to make their own decisions. Given that many ISOs are adopting methods based on distribution factors to make transmission decisions, the DC power flow may be accurate enough for the traders to anticipate transmission constraints and make informed decisions. A two-stage DC estimation is proposed to detect and identify topology errors. In the first stage, state estimation is performed on the bus/branch level. When errors are detected, the suspicious area is converted to bus-section/switching-device level and the second stage state estimation is performed. Multiple scan DC state estimation methods are introduced. The DC model modeling error is also partly estimated. Results on several IEEE test systems show the validity of the method.

The DC estimator is not proposed here to be a replacement for a full AC estimator, which might be needed by the system operator, but rather as a simplified view of the power system appropriate for certain market participants. An open electricity market has many players with different viewpoints of the system and needs for accuracy. DC state estimation has many advantages and could easily be implemented outside the control center given availability to select measured data and system parameters. Further, the results can be more easily related to typical market rules. The authors suggest that where the proposed estimator begins to break down under the burden of modeling errors, it is also likely that the limits of the trading rules will begin to be reached.

6.1.3 Distributed Information Infrastructure and Distributed State Estimation

Instead of a centralized control and processing model of the existing power system information structure, a distributed information processing model that using the publisher/subscriber paradigm is introduced.

By using this model, one can achieve:

- greater flexibility, since the system can be easily reconstructed;
- increased reliability, since the information transfer is more due to the possibility of re-routing packets unlike in the existing hard-wired network; and
- more evenly distribute computational load as the control center is not the only entity that can receive the necessary information - any subscriber can have the data it needs to perform computation locally.

The simulations on IEEE testing systems revealed that the new information structure is feasible and has acceptable performance.

State estimation is a fundamental component in power system network security analysis. In modern EMS, the state estimation is executed as snapshots, typically once every several minutes. This work has reported on a new approach to state estimation based on the proposed publisher/subscriber communication infrastructure. This infrastructure affords an opportunity to greatly improve the existing approaches to system security.

The proposed algorithm is based on this new distributed information structure. The algorithm is asynchronous and avoids a central controlling node during the iteration, thus, improving system wide robustness. Each local area performs the state estimation independently and the final result is assembled in a central processor. It is not expected that the local estimators, when combined in this fashion, can find the global maximum likelihood estimate. Still numerical tests verify that the new algorithm is sufficiently accurate enough compared with existing conventional full network state estimation algorithms. Moreover, the real challenge for modern state estimators is robustness with respect to missing and bad data. Such a distributed approach is far superior in this regard.

6.2 Future Work

Several technologies and algorithms are introduced in the dissertation to address the problem brought by today's deregulated power systems. There are still many open problems related to the communication and information systems.

1. Data warehousing is introduced to deal with the data from the business world. But in power industries, there is also a huge amount of data coming from the underlying physical system. Common data warehousing products have efficient programs to deal with the business computation such as, aggregation, but they have no tools to deal with data specific to the power system. Effort still needs to be made to have efficient applications to manage the huge amount of physical system data.

2. Truly distribute power system operations are difficult. The simulation done in the dissertation proves it is feasible from a communication viewpoint. Still, it is not clear the degree to which new communications can be integrated with the existing infrastructure without sacrificing performance.

3. With the introduction of the new communication network, one can expect that a lot of existing power system analysis tools need to be revised to take advantage of the distributed computation. The distributed state estimation proposed in the dissertation is one example, but many applications need to be adapted to the new information infrastructure.

Bibliography

[1] J. McDonald, “Substation automation. ied integration and availability of information,” *IEEE Power and Energy Magazine*, vol. 1, pp. 22–31, Mar-Apr 2003.

[2] M. Kezunovic, A. Abur, A. Edris, and D. Sobajic, “data integration/exchange part 1: existing technical and business opportunities,” *IEEE Power and Energy Magazine*, vol. 2, pp. 14–19, Mar-Apr 2004.

[3] M. Kezunovic, A. Abur, A. Edris, and D. Sobajic, “data integration/exchange part 2: future technical and business opportunities,” *IEEE Power and Energy Magazine*, vol. 2, pp. 24–29, May-June 2004.

[4] J. Gillerman, S. Mauser, and D. Nordell, “Utility information integration vision, benefits, strategies, and status,” in *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, Jan. 2000.

[5] W. H. Inmon, *Build the Data Warehouse*. John Wiley, 1992.

[6] J. Gillerman, S. Mauser, and D.Nordell, “Utility information integration - vision, benefits, strategies, and status.” Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, Jan 2000.

[7] M. Werner and U. Hermansson, “Integrated utility data warehousing - a prerequisite to keep up with competition on electricity market,” in *Power System Management and Control, Conference Publication IEE*, pp. 130–135, April 2002.

[8] D. Shi, Y. Lee, X. Duan, and Q. Wu, “Power system data warehouses,” *IEEE Computer Applications in Power*, pp. 49–55, July 2001.

[9] S. Sundhararajan, A. Boecker, J. Dondeti, R. Howard, J. Tamby, S. Grendel, and A. Jayantilal, “Experience with ercot system’s it development and operation,” *IEEE Transactions on Power Systems*, vol. 18, May 2003.

[10] J. Widom, “Research problems in data warehousing,” in *4th International Conference on Information and Knowledge Management*, (Baltimore, Maryland), pp. 25–30, 1995.

[11] G. Wiederhold, “Mediators in the architecture of future information systems,” *IEEE Computer*, vol. 25, no. 3, pp. 38–49, 1992.

[12] S. Chaudhuri and U. Dayal, "An overview of data warehousing and olap technology," *ACM SIGMOD Record*, vol. 26, pp. 65–74, March 1997.

[13] D. DeWitt and J. Gray, "Parallel database systems: the future of high performance database systems," *Communications of the ACM*, vol. 35, no. 6, pp. 85–98, 1992.

[14] A. W. McMorran, G. W. Ault, I. M. Elders, C. E. T. Foote, G. M. Burt, and J. R. McDonald, "Translating cim xml power system data to a proprietary format for system simulation," *IEEE TRANSACTIONS ON POWER SYSTEMS*, vol. 19, pp. 229–235, february 2004.

[15] R. Kimball and M. Ross, *The Data Warehouse Toolkit, 2nd Ed*. John Wiley & Sons, Inc, 2002.

[16] K. T. L. Xu and A. Bose, "Topology error identification using a two-state dc state estimator," *Journal of Electric Power Systems Research*, In Press.

[17] H. Chao and S. Peck, "A market mechanism for electric power transmission," *Journal of Regulatory Economics*, vol. 10, pp. 25–59, 1996.

[18] A. Monticelli and A. Garcia, "Modeling zero-impedance branches in power system state estimation," *IEEE Trans. on Power Systems*, vol. 6, pp. 1561–1570, 1991.

[19] S. E. A. Monticelli, *State Estimation in Electric Power Systems, a Generalized Approach*. Kluwer Academic Publishers, 1999.

[20] O. Alsac, N. Vempati, B. Scott, and A. Monticelli, "Generalized state estimation," *IEEE Trans. on Power Systems*, vol. 13, pp. 1069–1075, Augest 1998.

[21] F. Wu and W. Liu, "Detection of topology errors by state estimation," *IEEE Trans. on Power Systems*, vol. 4, pp. 176–183, Feburary 1989.

[22] W. Liu, F. Wu, and S. Lun, "Estimation of parameter errors from measurement residuals in state estimation," *IEEE Trans. on Power Systems*, vol. 7, pp. 81–89, Feburary 1992.

[23] K. Clements and P. Davis, "Detection and indentification of topology errors in electric power systems," *IEEE Trans. on Power Systems*, vol. 3, pp. 1748–1753, November 1988.

[24] R.L.Lugtu, K. D.F.Hackett, and D.D.Might, "Power system state estimation: Detection of topology errors," *IEEE Trans. on Power Systems*, pp. 2406–2412, Nov./Dec. 1980.

[25] A. Abur, H. Kim, and M. Celik, "Identifying the unknown circuit breaker statuses in power networks," *IEEE Trans. on Power Systems*, pp. 2029–2037, November 1995.

[26] A. Bjorck, *Numberical Methods for Linear Square Problem*. SIAM, 1996.

[27] G. H. Golub and C. F. V. Loan, *Matrix Computation, 2nd Edition*. The Johns Hopkins University Press, 1989.

[28] J. Demmel, "Applied numerical linear analysis," *SIAM*, 1997.

[29] S. V. Huffel and J. Vandewalle, “The total least squares problem: Computational aspects and analysis,” *Frontiers in Applied Mathematics, SIAM*, 1991.

[30] R. G. Brown and P. Y. C. Hwang, *Introduction to Random Signals and Applied Kalman Filtering*. John Wiley & Sons, Inc, 1992.

[31] “Power systems test case archive.” Department of Electrical Engineering, University of Washington, http://www.ee.washington.edu/research/pstca/.

[32] A. Carzaniga, D. S. Rosenblum, and A. L.Wolf, “Design and evaluation of a wide-area event notification service,” *ACM Transaction on Computer System*, vol. 19, pp. 332–383, August 2001.

[33] T. V. Cutsem and M. Ribbens-Pavella, “Critical survey of hierarchical methods for state estimation of electrical power systems,” *IEEE Transaction on Power Apparatus and Systems*, vol. PAS-102, pp. 3415–3424, October 1983.

[34] S.Y.Lin, “A distributed state estimator for electric power systems,” *IEEE Transaction on Power Systems*, vol. 7, pp. 551–557, May 1992.

[35] D. Falcao, F. F. Wu, and L. Murphy, “Parallel and distributed state estimation,” *IEEE Transaction on Power Systems*, vol. 10, pp. 724–730, May 1995.

[36] R. Ebrahimian and R. Baldick, “State estimation distributed processing,” *IEEE Transactions on Power System*, vol. 15, pp. 1240–1246, November 2000.

[37] G. Cohen, “Auxiliary problem principle and decomposition of optimization problems,” *Journal of Optimization Theory and Applications*, vol. 32, pp. 277–305, November 1980.

[38] G. M. Huang and J. Lei, “A concurrent non-recursive textured algorithm for distributed multi-utility state estimation,” in *IEEE PES Summer Meeting*, vol. 3, pp. 1570–1575, July 2002.

[39] I. Dabbaghchi and L. S. VanSlyck, “Inter-utility data exchange for sate estimation,” *IEEE Transactions on Power System*, vol. 3, pp. 1254–1262, August 1988.

[40] C.-L. Su and C.-N. Lu, “Interconnected network state estimation using randomly delayed measurements,” *IEEE Transactions on Power System*, vol. 16, pp. 870–878, November 2001.

[41] A. Monticelli, *State Estimation in Electric Power Systems, A Generalized Approach*. Kluwer Academic Publishers, 1999.