Modelling Supercomputer Maintenance Interrupts: Maintenance Policy Recommendations

Jagadish Cherukuri

University of Tennessee - Knoxville, jcheruku@vols.utk.edu

Follow this and additional works at: https://trace.tennessee.edu/utk_gradthes

Part of the Industrial Engineering Commons, Other Engineering Commons, Risk Analysis Commons, and the Statistical Models Commons

Recommended Citation
https://trace.tennessee.edu/utk_gradthes/3442

This Thesis is brought to you for free and open access by the Graduate School at TRACE: Tennessee Research and Creative Exchange. It has been accepted for inclusion in Masters Theses by an authorized administrator of TRACE: Tennessee Research and Creative Exchange. For more information, please contact trace@utk.edu.
To the Graduate Council:

I am submitting herewith a thesis written by Jagadish Cherukuri entitled "Modelling Supercomputer Maintenance Interrupts: Maintenance Policy Recommendations." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Reliability and Maintainability Engineering.

Xiaoyan Zhu, Major Professor

We have read this thesis and recommend its acceptance:

Ramon V. Leon, Alberto Garcia, Mingzhou Jin

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)
Modelling Supercomputer Maintenance Interrupts: Maintenance Policy Recommendations

A Thesis Presented for the
Master of Science
Degree
The University of Tennessee, Knoxville

Jagadish Cherukuri
August 2015
ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my graduate advisor, Dr. Xiaoyan Zhu for her support throughout my graduate studies. I am fortunate to have an advisor who gave me the freedom to explore on my own and guided me to overcome any difficulties in my thesis work. She taught me how important it is to put the thoughts and work on paper, which made the documentation part easy. Her advice in planning my course work helped me to complete it in time. I am thankful to her for forwarding my case to the department and help me get the graduate teaching assistantship. Her candid and stoic behavior has a positive influence on me.

I am grateful to Dr. Ramon Leon. His course instruction style and one on one interaction helped me increase my capabilities in statistical data analysis. His valuable insights into the analysis of maintenance record data are very useful for the thesis work. His down-to-earth nature and his ability to provide easy explanation on complex topics are something that I wish to inculcate.

I am grateful to Dr. Mingzhou Jin and Dr. Alberto Garcia for being my committee members. Various seminars organized by Dr. Jin, in the department are instrumental in enhancing my understanding of industrial engineering. Dr. Garcia emphasis on the importance of an engineer in building a healthy society always motivates.

I thank all the professors who have instructed my graduate courses. They are all instrumental in providing me a good understanding of reliability and maintainability concepts. I thank the department administrative staff for all the support that made things easy for me. I thank all my colleagues and friends who enriched my academic experience in the university.

I appreciate the financial assistantship provided by Industrial and Systems engineering department for my graduate studies. I appreciate Dr. Mark Fahey from ORNL for sharing the supercomputer maintenance data that formed the basis of my thesis.

My parents, brother and my friends in India, I cannot thank them enough for their constant love, care and support throughout my life. I am blessed to have them.
ABSTRACT

A supercomputer is a repairable system with large number of compute nodes interconnected to work in harmony to achieve superior computational performance. Reliability of such a complex system depends on an effective maintenance strategy that involves both emergency and preventive maintenance. This thesis analyzes the maintenance records of four supercomputers operational at The National Institute of Computational Science located at Oak Ridge National Laboratory. We propose to use the generalized proportional intensities model (GPIM) to model the maintenance interrupts as it can capture both the reliability parameters and maintenance parameters and allows the inclusion of both emergency and preventive maintenance. We use this model to obtain the reliability parameters indicating the system performance and maintenance parameters indicating the effectiveness of maintenance actions for each of the four supercomputers. System performance measures such as reliability and availability are used to evaluate the effectiveness of the existing maintenance policy and to propose a new maintenance policy that increases the system availability and reduces maintenance cost.
# TABLE OF CONTENTS

1 INTRODUCTION.................................................................................................................................................. 1
  1.1 Background.................................................................................................................................................. 1
  1.2 Motivation.................................................................................................................................................... 2
  1.3 Maintenance Records of the Supercomputers ......................................................................................... 2
  1.4 Methodology.............................................................................................................................................. 3
  1.5 Purpose of the Thesis................................................................................................................................. 4
  1.6 Organization of the Thesis........................................................................................................................ 4

2 LITERATURE REVIEW........................................................................................................................................ 5
  2.1 Failure Analysis on Supercomputers ....................................................................................................... 5
  2.2 Failure Analysis of Repairable Systems................................................................................................ 6
    2.2.1 Component failure analysis .................................................................................................................... 7
    2.2.2 Repairable system analysis .................................................................................................................... 7
    2.2.3 As good as new maintenance models ................................................................................................. 8
    2.2.4 As good as old maintenance models .................................................................................................. 9
    2.2.5 Models considering maintenance effectiveness ................................................................................ 10
    2.2.6 Generalized maintenance models ..................................................................................................... 11
  2.3 Summary...................................................................................................................................................... 12
  2.4 Uniqueness of this Thesis ........................................................................................................................ 12

3 GENERAL PROPORTIONAL INTENSITY MODEL............................................................................................ 13
  3.1 GPIM and Methodology ............................................................................................................................. 13
    3.1.1 Maximum likelihood estimate (MLE) ................................................................................................. 17
    3.1.2 Optimizing the log-likelihood ............................................................................................................ 18
  3.2 Time Trend Test ......................................................................................................................................... 18
  3.3 Results......................................................................................................................................................... 19
    3.3.1 Athena .................................................................................................................................................. 20
    3.3.2 Jaguar .................................................................................................................................................. 21
    3.3.3 Jaguar PF ............................................................................................................................................ 22
    3.3.4 Kraken ................................................................................................................................................ 22
  3.4 Summary...................................................................................................................................................... 24
4 REPAIR TIMES ANALYSIS ................................................................. 26
  4.1 Analysis Approach ................................................................. 26
  4.2 Model Selection Measures ....................................................... 26
  4.3 Tested Distributions ............................................................... 27
  4.3.1 Weibull distribution .............................................................. 28
  4.3.2 Lognormal distribution ......................................................... 28
  4.4 Emergency Maintenance ........................................................ 29
  4.4.1 Athena ............................................................................. 29
  4.4.2 Jaguar .............................................................................. 30
  4.4.3 Jaguar PF .......................................................................... 30
  4.4.4 Kraken ............................................................................. 31
  4.5 Preventive Maintenance ............................................................ 31
  4.5.1 Athena ............................................................................. 32
  4.5.2 Jaguar .............................................................................. 32
  4.5.3 Jaguar PF .......................................................................... 33
  4.5.4 Kraken ............................................................................. 33
  4.6 Emergency and Preventive Maintenance ...................................... 34
  4.6.1 Athena ............................................................................. 34
  4.6.2 Jaguar .............................................................................. 35
  4.6.3 Jaguar PF .......................................................................... 35
  4.6.4 Kraken ............................................................................. 36
  4.7 Summary .............................................................................. 37

5 AVAILABILITY AND MAINTENANCE POLICY ...................................... 39
  5.1 Availability ............................................................................ 39
    5.1.1 Achieved availability ......................................................... 39
    5.1.2 Predicted point availability ................................................. 40
    5.1.3 Predicted average availability .............................................. 41
  5.2 Maintainability .................................................................... 43
    5.2.1 Athena ........................................................................... 44
    5.2.2 Jaguar ............................................................................ 44
    5.2.3 Jaguar PF ....................................................................... 44
5.2.4 Kraken

5.3 Summary

6 CONCLUSION

6.1 Summary of Thesis

6.2 Proposed Maintenance Policy

6.3 Future Scope

REFERENCE

VITA
### LIST OF TABLES

Table 1: Number of maintenance records for each supercomputer ........................................... 3  
Table 2: GPIM model parameter estimates ..................................................................................... 20  
Table 3: Laplace trend test results ................................................................................................. 20  
Table 4: Distributions of emergency maintenance repair times for Athena ......................... 29  
Table 5: Distributions of emergency maintenance repair times for Jaguar ......................... 30  
Table 6: Distributions of emergency maintenance repair times for Jaguar PF ................. 31  
Table 7: Distributions of emergency maintenance repair times for Kraken ....................... 31  
Table 8: Distributions of preventive maintenance repair times for Athena ...................... 32  
Table 9: Distributions of preventive maintenance repair times for Jaguar ....................... 33  
Table 10: Distributions of preventive maintenance repair times for Jaguar PF .............. 33  
Table 11: Distributions of all the maintenance repair times for Athena ......................... 35  
Table 12: Distributions of all the maintenance repair times for Jaguar ......................... 35  
Table 13: Distributions of all the maintenance repair times for Jaguar PF .................. 36  
Table 14: Distributions of all the maintenance repair times for Kraken ......................... 36  
Table 15: Maintenance (repair) time statistics .............................................................................. 37  
Table 16: Achieved availability ..................................................................................................... 40  
Table 17: Failure intensity function after the last recorded maintenance ....................... 41  
Table 18: Mean time to interrupts and predicted average availability for 2160 hours .... 43
LIST OF FIGURES

Figure 1: Geometric interpretation of failures modelled by stochastic point process ........ 8
Figure 2: Failure intensity for AGAN model .......................................................... 8
Figure 3: Failure intensity for AGAO model ......................................................... 9
Figure 4: Failure intensity for age reduction model ............................................. 11
Figure 5: Schematic representation of the supercomputer interrupt-maintenance process ...................................................................................................................................... 13
Figure 6: Schematic representation of the supercomputer interrupt process .......... 14
Figure 7: Geometric interpretation of stochastic point process for the supercomputers .. 15
Figure 8: Failure intensity of Athena vs time – GPIM ........................................... 21
Figure 9: Failure intensity of Jaguar vs time – GPIM .......................................... 22
Figure 10: Failure intensity of Jaguar PF vs time - GPIM ...................................... 23
Figure 11: Failure intensity of Kraken vs time - GPIM .......................................... 24
Figure 12: Schematic representation of the supercomputer maintenance process ....... 26
Figure 13: Point availability vs time ........................................................................ 42
1 INTRODUCTION

First, Sections 1.1 and 1.2 provide the background and motivation of the thesis, respectively. Then, Sections 1.3 and 1.4 discuss the data and methodologies used in this thesis, respectively. Section 1.5 presents the purpose of the work. Finally, Section 1.6 gives the structure of this thesis.

1.1 Background

In modern times, human life is being increasingly dependent on various advanced technologies. There is a growing demand for handling huge data storage, analysis of big data, implementation of complex algorithms, etc. for addressing modern day problems. Modelling and simulating different real life scenarios are becoming increasingly important to scientific predictions. With these growing computational and simulation needs, supercomputers become necessary and play a critical role in various fields such as nuclear, oil and gas explorations, environmental studies, medical research, healthcare, communications, transportation, and so on.

A supercomputer is a system that provides supercomputing capabilities and has the highest performance with largest capability and capacity in a period of time (Xie, Fang, Hu, & Wu, 2010). The latest supercomputers are capable of reaching one thousand million floating point operations per second, usually called as one petaflop.

Supercomputer architecture is an interface between hardware, and software and form the basis for its high computing capability. Supercomputers may run complex algorithms requiring a relatively long, a few months to solve. Any interruption to its normal functioning at an intermediate point will require the entire task to be restarted or to start from the previous checkpoint, if check pointing is possible and implemented. This makes the supercomputer unavailable not only during the emergency maintenance time, needed to fix the problem but also the entire period from which the job is started, considering that it should have successfully completed other smaller jobs during this time. It can be interrupted due to failures in processors, memory storage devices and other physical component(s), termed as hardware failures or due to failures in the input/output system and the interconnect system, termed as software failures. Any of these failures lead to improper functioning of supercomputer and consequently result in loss of data, termination or pause of on-going tasks, and so on.

It is of high interest to design a highly dependable supercomputer to run the jobs without any interruption. With large number of components that can possibly fail and cause interruption, a supercomputer’s dependability is a function of its reliability, availability and maintainability (Stearley, 2005). In this aspect, it becomes necessary for a supercomputer to have an effective maintenance policy that can restore it after a failure and that can effectively identify and address the incipient failures. In order to maintain the availability of such a system at a required level, it is important to understand its failure characteristics and keep a track of the effectiveness of its maintenance policy.
1.2 Motivation

This work is motivated by the demand to analyze the real-time maintenance records of four supercomputers, namely, Athena, Jaguar, Jaguar PF, and Kraken, which are located and operated at The National Institute for Computational Sciences in Oak Ridge National Lab, Tennessee, US.

Athena, Cray XT4 supercomputer is dedicated for climate, weather, and quantum chromodynamics research. With 18,048 cores, 4,512 compute nodes, two AMD Opteron 2.3 GHz quad-core processors at each node and 18 terabytes of memory, it can reach a peak performance of 166 teraflop (Baer, 2010). Jaguar and Jaguar PF are used for the internal differentiation of Cray XT4 and Cray XT5 machines, respectively. Cray XT5 has a peak performance of 1.38 petaflop, while Cray XT4 has 0.26 petaflop. Both these put together have a system memory of 362 TB and 45,208 quad-core Opteron processors. SeaStar2+ 3D torus connects the compute nodes (Bland, Kendall, Kothe, Rogers, & Shipman, 2009). Kraken is Cray XT5 supercomputer with a peak performance of 1.17 petaflop, 147 TB of compute memory and 9,408 compute nodes, it is the most powerful supercomputer used for academic purposes. SeaStar2+ router connects the compute nodes that have two 2.6 GHz hex-core processors (NICS, 2014).

1.3 Maintenance Records of the Supercomputers

The main data in the work are the maintenance records of the four supercomputers, Athena, Jaguar, Jaguar PF and, Kraken. The data was collected for the period from October 2009 to December 2010. These records include both emergency maintenance and preventive maintenance. Emergency maintenance is a maintenance activity performed immediately after a system’s failure with an intention to fix the problem and restore it to its normal operating mode. Preventive maintenance is a maintenance activity that is usually performed at regular pre-determined intervals with an intention to identify and reduce incipient failures that can potentially cause a system failure at a later stage, thus reducing the occurrence of actual failures. The maintenance policy for the four supercomputers calls for immediate emergency maintenance after every failure and preventive maintenance at regular interval of two weeks. It can be noted from the maintenance records that the preventive maintenance was not performed at a fixed interval for any of the four supercomputers, probably due to some practical restrictions. It may be due to the fact that the supercomputer is not intentionally interrupted while performing a job.

Each maintenance performed is entered into the event log system manually and includes:

- **Downtime**: it is the date and time at which the system’s normal functioning is interrupted due to either the emergency maintenance or the preventive maintenance. This is also called as *Incident Start Time*.

- **Uptime**: it is the date and time at which the system is brought back to its normal functioning after either the emergency maintenance or the preventive maintenance. This is also called as *Incident End Time*.
- *Type of the interrupt:* it indicates if the interruption is due to emergency maintenance or preventive maintenance.

The raw maintenance records obtained for each supercomputer are refined by eliminating any redundancies. Table 1 shows the number of emergency and preventive maintenance records after refining the raw data for each supercomputer, over the recorded period.

<table>
<thead>
<tr>
<th>Maintenance Type</th>
<th>Supercomputer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Athena</td>
</tr>
<tr>
<td>Emergency</td>
<td>35</td>
</tr>
<tr>
<td>Preventive</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>42</td>
</tr>
</tbody>
</table>

1.4 **Methodology**

A system that can be restored to satisfactory performance after failing to perform one or more of its intended functions satisfactorily by employing a method other than replacing the entire system is called a repairable system (Ascher & Feingold, 1984). This is the most commonly used definition of a repairable system. A supercomputer is a combination of several clusters, with each cluster having several modules. Each module has its own failure characteristic and can fail independently, causing a system failure. Each cluster can be repaired independently and is done by repair or replacement of its module(s). The maintenance records available do not contain any information at a module level but are at a system level. Thus, the maintenance records can be considered as a superimposition of interruptions at module level. Therefore, the analysis of maintenance records is done considering the failure of the system as a whole. Each of the four supercomputers is considered to be a complex repairable system as they can be restored to satisfactory performance after each interruption.

We have a situation where each time there is an interruption of a supercomputer, a maintenance activity returns it to a satisfactory operating condition, creating an interrupt-maintenance-interrupt-maintenance cycle. The interruption can be due to the emergency maintenance or the preventive maintenance. In order to estimate a supercomputer performance measures, it is required to understand the interrupt characteristics and the maintenance characteristics. Interrupt characteristics can be understood by modelling the maintenance records using a statistical model that best fits the time to interrupts of a supercomputer. Maintenance characteristics can be understood by modelling the repair time using a probability life distribution model.

Chapter 2 presents the literature review done on supercomputers in order to gain understanding of analysis approaches used to analyze failure data of supercomputers. It also presents the literature review done on repairable systems in general to understand the
The evolution of statistical models used to model their failure data. It has been established that general proportional intensity model (GPIM) can be used to model the maintenance records of supercomputers. This model has the provision to model both emergency and preventive maintenance together and also considers their effectiveness of system restoration. GPIM is used to model the interruption process of each of the four supercomputers separately. Probability life distributions are used to model the repair times of the maintenance activities of each of the four supercomputers separately. The results of these two models are used to estimate the availability of each of the four supercomputers. All these results are used to review the existing maintenance policy and propose new maintenance policy if there is a scope for improvement.

1.5 Purpose of the Thesis

The purpose of this thesis is to analyze the maintenance records of each of the four supercomputers, Athena, Jaguar, Jaguar PF, and Kraken to understand their dependability. Dependability of a supercomputer is a function of its reliability, availability and maintainability.

The real time maintenance data of the four supercomputers is used to obtain the below outcomes, for each of it.

- Statistically model the process of occurrence of maintenance interrupts
- Estimate the reliability and maintainability parameters
- Model the repair time to fit life probability distributions
- Calculate the achieved availability during the observed period and estimate the predicted availability
- Comment on the existing maintenance policy and propose an improved maintenance policy

1.6 Organization of the Thesis

The remainder of the thesis is organized as follows. Chapter 2 presents the literature review on the basic reliability and availability models and those related to the reliability of high performance computing. Chapter 3 presents GPIM and the approach followed to obtain the reliability and maintainability parameters of each supercomputer, along with the results for each supercomputer. Chapter 4 presents the repair time analysis and the basic statistics of the repair times of each supercomputer. The analysis is conducted for the repair times of only the emergency maintenance, only the preventive maintenance, and both emergency and preventive maintenances. Chapter 5 presents the system availability measures for each supercomputer along with comments on existing maintenance policy and proposes a new maintenance policy. Chapter 6 summarizes the work done in this thesis and discusses future scope.
2 LITERATURE REVIEW

This chapter is divided into four sections. Section 2.1 focuses on the literature review of failure analysis done on high performance computing systems. Section 2.2 focuses on the literature review of repairable systems and the approach used to analyze their failure occurrences. Section 2.4 summarizes the findings of the literature review that are applicable to this thesis. Finally, Section 2.4 explains how the work done in this thesis is different from the other publications.

2.1 Failure Analysis on Supercomputers

Supercomputers are computing systems with the highest capability that can handle the computational needs in a given period of time. The evolution of supercomputers, propelled by the discoveries made in materials and manufacturing over the past 40 years, is noteworthy and they have seen a 10-fold increase in the performance, every 4 years. Their computational performance has been improved with changes in architecture, software, applications, etc., the latest being massive parallel processing and cluster systems that integrate large number of components (Riganati, 1984; Scheneck, 1990; Xie, Fang, Hu, & Wu, 2010; Yang, Liao, & Song, 2011). The failures on any component, be it hardware, software, input/output systems or interconnect system can hinder its normal functioning. A proper repair action has to be performed to get it back to its normal operational mode. It is required to understand the failure pattern to make important organizational decisions such as maintenance policy to be followed, inventory requirements, etc.

The body of literature on supercomputer shows that there is a lot of focus on the supercomputer architecture and ways to improve its computing performance. The primary interest is to study the part of literature that specifically focuses on the work done on failure analysis of supercomputers, aiming at understanding its performance and suggesting ways to improve its maintenance. It is observed that failure analysis has been done on high performance computing systems operational in both laboratory and commercial setups. The work varied depending on the type of data collected and the focus of the study.

Two publications focused on the causes of failure and presented the basic statistical data. One publication studied Tandem supercomputer over a period of 3 years and presented the basic statistical data comparing its failure causes. The failure causes are categorized as environmental, human related, interconnect, software and hardware. It also presented the data showing that the reliability of the hardware component of the system is improved by proper maintenance plan (Gray, 1990). The other publication characterized the causes of machine reboot of Windows NT, the network system in a commercial environment. The observations show that most of the reboots are caused by software. It is also shown that rebooting does not solve the problem but only contributes to the machine downtime (Kalyanakrishnam, Kalbarczyk, & Iyer, 1999). There is one publication that focuses on understanding the correlation of failures with factors external to the system. Failure analysis of a large scale server environment, holding up to 400 servers show that system errors exhibit a time varying behavior with periodic patterns. It has been shown that the
failure rates have a strong correlation with workload and the number of hours of operation in a day. (Sahoo, Squillante, Sivasubramaniam, & Zhang, 2004).

Two publications focused on time between failures, repair times, and rate of failure. These publications used probability life distributions to fit the time between failures, and repair times and used the probability density function to estimate the rate of failure. It is shown that Weibull distribution and exponential distributions are a good fit for the time to failures. Failure analysis of internet service interruptions on a collection of interconnected computers treated as a high performance computing system shows that time between failures is modelled by a Weibull distribution with shape parameter less than 1, showing a decreasing failure rate (Heath, 2002). 23000 failures records of about 20 different large computing clusters are analyzed and life distributions are fitted to the time between failures, and the repair times. Weibull distribution with shape parameters ranging from 0.7 to 0.8 are fitted for the time between failures, and lognormal distributions are fitted for the repair times. It is reported that the failures show decreasing hazard rate and repair times vary a lot across different clusters (Schroeder & Gibson, 2010).

Three publications have focused on understanding the availability of high performance computing clusters. One publication studies two DEC VAX cluster multicomputer systems and established that there are correlated failures caused due to shared resources. It is shown that about 40% of failures occurs in bursts. The failure data is used to model the availability of system based on K out of n systems approach (Tang, Iyer, & Subramani, 1990). The other publication studied the individual server failures to understand failure propagation between 503 servers of Networked Windows NT system, in a commercial environment. It was observed that software and hardware failures are the major contributors and the average availability of individual server is over 99% (Xu, Kalbarcyzyk, & Iyer, 1999). The third publication studied real time event logs from a 512 node clusters in Lawrence Livermore National Laboratory. A single framework that coordinates event monitoring, filtering, data analysis and dynamic availability modeling is presented based on the Markov chain models (Song, Chokchai, Nassar, Gottumukkala, & Scott, 2006).

To the best of our search, we did not find any publication with a focus on analyzing the maintenance records of supercomputers that takes into account both emergency maintenance and preventive maintenance. Also, we did not find any publication that focuses on optimally scheduling the preventive maintenance interval of a supercomputer. In this regard, it is important to review the literature on modelling approaches to model the failure data of a repairable system to see if there exists a model that can be used for the maintenance records of the four supercomputers, Athena, Jaguar, Jaguar PF and Kraken.

2.2 Failure Analysis of Repairable Systems

This section presents different models used in literature to model the failure data of a repairable system and discusses the suitability of them to model the maintenance records of the four supercomputers.
2.2.1 Component failure analysis

In the context of this thesis, a component is anything that once failed, the cost of repair is almost equal to cost of the component and so it is better to replace rather than repair it, after a failure. Thus, for a component, the failure analysis is based on the time to first failure. To model the reliability of a component, one needs to fit an appropriate probability life distribution to the time to first failures of all the components tested. Fitting probability distributions requires selecting a probability model that best fits the failure data, from a list of models that are generally used in reliability analysis (Barlow & Proschan, 1975; ReliaSoft). Proper model selection criteria is to be used to select the best model and one needs a good understanding on various facets of multi-model selection (Akaike, 1974; Schwarz, 1978; Burnham & Anderson, 2004; Quesenberry & Kent, 1982; Ye, Meyer, & Neuman, 2008). Once the probable model is selected, the parameters that best fit the failure data are estimated (Basu, 1964; Scholz, 2004).

The approach is used to fit the life distributions to the time between failures of supercomputers in some publications (Heath, 2002; Schroeder & Gibson, 2010) but it is not appropriate. This approach is only suitable to model the time to first failures as they can be treated as independent and identically distributed (IID). Life distributions are not appropriate to model the time between failures (interrupts) of a supercomputer (a repairable system) as they are neither independent nor identically distributed but can be used to model the repair times of the supercomputers as repair times form a series of random variables that are IID.

2.2.2 Repairable system analysis

The time between failures of a repairable system are not IID and so a repairable system has to be modelled by a process rather than a distribution (ReliaSoft). Literature on repairable systems is primarily focused on modeling failure occurrences using counting theory or point process theory.

Duane model is the earliest model that accounts for the changes in failure rates over the system life (Duane, 1964). AMSAA model proposes a more accurate model to fit the occurrence of failures of a repairable system and is based on stochastic point process (Crow L. H., 1984). It defines the intensity of the failures observed in a repairable system as the rate of change of expected number of failures with respect to time. Figure 1 shows geometric representation of the failures modelled by stochastic point process. \( T_1, T_2, T_3, \ldots \) represent the time of occurrence of failures and \( X_1, X_2, X_3, \ldots \) represent the time between failures. A lot of publications focus their study in this area and various other models have been proposed. The summary of about fifteen distinct growth models is provided in The Military Handbook: Reliability Growth Management (Department of Defense, 1981). NHPP based on power law model is the most popular process in literature used to obtain the failure intensity function of a repairable system as it can model both varying and constant failure intensity (ReliaSoft; Park & Pickering, 1997).
2.2.3 *As good as new maintenance models*

When a maintenance activity restores the repairable system such that it is brought back to as good as new (AGAN) condition, in which it is just as it was first operated, then it is called a perfect maintenance. In this case, the time between failures of the system form random variables that are IID and the failure process is said to follow a renewal process (Taylor & Karlin, 1994). Figure 2 shows the graph of failure intensity vs time for a system modelled by ASAN maintenance model.
The linear baseline intensity is used for illustration purpose. It can be seen that the failure intensity is reset to zero by each maintenance action performed after a failure has occurred. ASAN model is a bad model for complex repairable systems as any maintenance will only fix a part of the system to restore it to a satisfactory performance but does not repair majority of other components. Predominant presence of aged components means that the system is not renewed with respect to its reliability aspects. Thus, this model is ruled out to model the maintenance records of the four supercomputers.

2.2.4 As good as old maintenance models

When a maintenance activity restores the system such that it is brought back to as good as old (AGAO) condition, in which it is just as it was immediately before the occurrence of failure, then it is called minimal maintenance. In this case, the time between failures of the system form random variables that are not IID and the failure process is modelled by Non homogeneous Poisson process (NHPP). If the rate of occurrence of interrupts is constant over time, i.e., constant failure intensity, it is a special case of NHPP that follows homogeneous Poisson process (HPP) (Crow, 1975). Reliability trend tests are used to statistically verify if the failure intensity is constant, increasing, or decreasing (Coit, 2005; Kvaloy & Lindqvist, 1998). Figure 3 shows the graph of failure intensity vs time for a system modelled by AGAO model with linear baseline intensity used for illustration purpose.

![Figure 3: Failure intensity for AGAO model](image)

It can be seen that there is no change in the failure intensity after each maintenance followed by occurrence of a failure. AGAO model, also called as NHPP model is a poor
model for systems with very few components for which the intensity function usually changes following maintenance but it can be considered for systems with large number of components. Thus, for the supercomputers this model cannot be ruled out but it may not be a practical case always as it considers only one possible case of restoration. Also, one has to treat both the emergency and preventive maintenance in the same way to use this model.

2.2.5 Models considering maintenance effectiveness

AGAN and AGAO are only two possibilities that a repairable system can be restored to. AGAN is an extreme case and is the best possible state of restoration. A more practical case may be that after the maintenance, the system is restored to a condition which is worse than new and better than the condition at which it has failed. Brown and Proschan proposed an imperfect repair model in which it is considered that a maintenance restores the system to AGAN with probability $p$ and AGAO with probability $(1 - p)$ (Brown & Proschan, 1983). Chan and Shaw (Chan & Shaw, 1993) model and the quasi renewal model (Jack, 1998) are two other models which are on the similar lines. However, these models still consider that all the maintenance activities restore a system to either AGAN or AGAO, thus creating a need for other models that can consider the general effect of maintenance on the system’s performance. A maintenance can affect a system in a way that it reduces the system’s failure intensity where it actually fixes the problem, or in a way that it increases the system’s failure intensity, cases where it may induce new defects, such as in photocopiers.

The literature contains many models that can accommodate different effects of maintenance on system’s performance. This section covers the prominent models. The age reduction Kijima Type I and Type II models (Kijima, 1989) and proportional intensity model (PIM) (Cox, 1972; Percy, Kobbacy, & Ascher, 1998) are the most commonly used models to model a repairable system under imperfect maintenance. Both these models contain NHPP as the baseline model. Age reduction model modifies the intensity function considering a virtual age to which a system is reset to after maintenance. The virtual age is a fraction of the actual age whose magnitude is decided by the effectiveness of maintenance. This does not modify the baseline intensity. The PIM modifies the baseline intensity function after each maintenance by a multiplicative or an additive factor whose magnitude depends on the effectiveness of maintenance. Doyen and Gaudoin (Doyen & Gaudoin, 2004) proposed models with failure intensity improvement factor after each repair and is based on two approaches, arithmetic reduction, and geometric reduction. Multiplicative scaling of the intensity function after maintenance is a more recent proposed model and better fits the physical situation of a repairable system that is improving, deteriorating or constant with time (David F.P. & Babakalli, 2006). Figure 4 shows the graph of failure intensity for age reduction model with linear baseline intensity used for illustration purpose.
The limitation to these models is that they are designed to model only emergency maintenance (followed by an actual failures) but the supercomputers have emergency as well as preventive maintenance. Thus, other models that can accommodate preventive maintenance are considered.

2.2.6 Generalized maintenance models

Generalized age reduction model (GARM) and Generalized proportional intensity model (GPIM) consider effect of maintenance on system’s performance and have the provision for modelling both the emergency and preventive maintenance together. These models are introduced during the past five years and indicate the increase in awareness of inclusion of preventive maintenance in the maintenance protocol of industries.

GARM (Arwa, Soufiane, & Mounir, 2013) is an extension of the age reduction model and GPIM (Percy & Babakalli, 2006) is an extension of the PIM model. Though GARM is a good statistical model and fits the maintenance data well, it is more of a theoretical model than a model that considers the true nature of the repairable system. GPIM provides a more practical physical model and has a higher potential for maintenance decision making than GARM with the possibility to accommodate future extensions to this study, such as including predictor variables and covariates. GARM is generally used for systems following a block replacement policy and can be a good statistical model but has a limitation that it does not provide a practical description of the failure process. Replacing a failed timing belt of a car does not reduce the age of the car as all the other components are not any less likely to fail. GPIM is finding an increased attention in the literature and has been applied to maintenance modelling in oil and gas industry (David & Babakalli, 2007), gas turbines (Babakalli, 2012) and power transmission sector (Amin, Mahmood, & Mohsen, 2014). This provides a better description of the actual physical situation of the
system as it modifies the failure intensity function after every maintenance activity based on the effectiveness of the maintenance.

2.3 Summary

To the best of our knowledge, the literature review shows that there is no publication that analyzes the maintenance records of supercomputers that contain both emergency maintenance and preventive maintenance.

The literature review shows that the below points can be considered to model the maintenance records of each of the four supercomputers.

- The process of occurrence of interrupts follows stochastic point process.
- GPIM is the best model that fits the maintenance records as it has the provision to model the emergency maintenance and preventive maintenance together and it also has factors representing the effectiveness of these two maintenance actions.
- NHPP with power law intensity can be used as the baseline failure intensity for the GPIM.

Thus, GPIM forms the main model for the failure analysis of the four supercomputers. Chapter 3 explains the approach taken to estimate the reliability and maintainability parameters using the GPIM and presents the results for the four supercomputers.

2.4 Uniqueness of this Thesis

To the best of our knowledge, an exhaustive literature search on failure analysis of supercomputers has shown that this thesis is different from the others publications in below ways:

- It considers the real time maintenance records of four supercomputers in a laboratory environment, spanning over a period of about one year, rather than a single system analysis done in most other publications.
- The four supercomputers are complex repairable systems involving large number of hardware and software components and are maintained by emergency as well as preventive maintenance, thus requiring a model with higher complexity. It is not a case of simple component replacement that can be modelled by AGAN or a simple repairable system that can be modelled by AGAO model.
- It adopts the GPIM to model the maintenance records of the supercomputers as it can capture the reliability parameters and effectiveness of maintenance for both emergency and preventive maintenance. GPIM is used for the first time to model maintenance data of high performance computing systems.
- It models the maintenance times using life distributions and uses the results in combination with the GPIM results to estimate the achieved and predicted availability of each of the four supercomputers. The results are also used to comment on the effectiveness of the existing maintenance policy and to propose an improved maintenance policy.
3 GENERAL PROPORTIONAL INTENSITY MODEL

This chapter fits the GPIM to the maintenance records of each of the four supercomputers. Section 3.1 presents the GPIM and the analysis approach followed to estimate its reliability and maintainability parameters. Section 3.2 presents a statistical test to determine the trend of interrupts. Section 3.3 presents the GPIM parameter values and the trend test results. It also comments on the system’s performance and effectiveness of maintenance actions for each supercomputer. Section 3.4 summarizes the observations made in this chapter.

3.1 GPIM and Methodology

The interrupt-maintenance-interrupt-maintenance cycle observed on the supercomputer cannot be modelled by a probability life distribution as it is usually applicable to model the time to first failure. Moreover, the time between successive interrupts of a supercomputer may not be independent and identically distributed. Thus, a stochastic point process that models the occurrence of interrupts is appropriate for modelling this situation (Ascher & Feingold, 1984). This model considers time to interrupts rather than time between interrupts. Figure 6 presents a schematic representation of the interrupt-maintenance process of the supercomputers. At any given point of time the supercomputer either of the two states, normal operating mode or under maintenance. The maintenance records of each of the four supercomputers start and end with the occurrence of an interrupt, either emergency or preventive.

Figure 5: Schematic representation of the supercomputer interrupt-maintenance process
This chapter is focused on the failure intensity of the supercomputer which is dependent on the time to interrupts. It is assumed that the interruption process is independent of the maintenance time at each interrupt and thus maintenance times are negligible compared to the time to interrupts. Thus, this chapter deals only with the normal operating state of the supercomputer along with the occurrence of the interrupts. The supercomputer maintenance state and the maintenance times analysis is dealt with in Chapter 4. Figure 6 presents a schematic representation of the interrupt-maintenance process of the supercomputers considered in this chapter.

Let $T_1, T_2, T_3, \ldots$ represent the time to successive interrupts of the supercomputer. These times are the supercomputer operating times. Let $T_k$ be the random variable representing the total operating time of the system since it initially started until the $k$th interrupt. For this analysis, it is assumed that the initial start point for each supercomputer is the point at which it resumes normal functioning, after fixing the first recorded interrupt. Thus, the operating time at the initial start point is zero. $T_n$ represents the operating time of the supercomputer at the last recorded interrupt. Let $X_1, X_2, X_3, \ldots$ represent the successive time between interrupts of the supercomputer. Let $X_k$ be the random variable representing the system operating time between interrupt $(k - 1)$ and interrupt $k$. $T_k$ and $X_k$ are related by the below two equations.

\[
X_k = T_k - T_{k-1} \tag{1}
\]

\[
T_k = \sum_{i=0}^{k} X_i \tag{2}
\]

Let $Z(t)$ be the counting function and represents the number of interrupts (both emergency and preventive) that occur during the interval $(0, t]$, for all $t > 0$ (Leemis, 1995). It is of primary interest to understand the behavior of the interruption process and is described by

![Figure 6: Schematic representation of the supercomputer interrupt process](image)
the failure intensity function, $\rho(t)$ which is the rate of change of expected number of interrupts with respect to time given by,

$$\rho(t) = \frac{d}{dt} E[Z(t)]$$  \hspace{1cm} (3)$$

and is also called as rate of occurrence of interrupts. Figure 7 shows the notation of interrupt pattern and geometric representation of the counting and the intensity function for the supercomputers.

![Figure 7: Geometric interpretation of stochastic point process for the supercomputers](image)

The failure intensity function of a GPIM is (Percy & Babakalli, 2006),

$$\rho(t) = \rho_0(t) \left( \prod_{i=1}^{N(t)} s_i \right) \left( \prod_{j=1}^{M(t)} r_j \right)$$  \hspace{1cm} (4)$$

where $s_i$ ($r_j$) is the intensity scaling factor for each emergency (preventive) maintenance and $N(t)$ ($M(t)$) is the total number of emergency (preventive) maintenance activities performed on the supercomputer during the interval $(0,t]$, $\rho_0(t)$ is the baseline intensity function of the supercomputer and $s_i$ and $r_j$ are positive real numbers for $i = 1, 2, 3,\ldots$ $N(t)$ and for $j = 1, 2, 3,\ldots$ $M(t)$, respectively.

The baseline intensity function can take many forms but the appropriate function is the one that corresponds to the hazard rate function of a familiar life distribution. This goes with the reason that for the time to first interrupt of a supercomputer, the intensity function and the hazard rate function are identical. Thus, the intensity function can be a constant intensity, log linear intensity or a power-law intensity corresponding to an exponential hazard, Gumbel hazard and, Weibull hazard, respectively. The most commonly used model
in the literature for a complex repairable systems is the power-law intensity (Crow L. H., 1975; Huairui & Wenbio, 2006) as it is flexible and can model increasing, decreasing and, constant failure rates. Thus, the time to first interrupt of each of the four supercomputers is assumed to have a Weibull distribution and the corresponding failure intensity function defined by power-law is

$$\rho_0(t) = \lambda \beta t^{\beta-1}$$

(5)

where $\lambda$ is the scale parameter and $\beta$ is the shape parameter corresponding to the Weibull hazard.

$\lambda$ and $\beta$ indicate the reliability parameters of the supercomputer and take positive real values. Particularly,

- $\beta$ lies in the interval $(0,1)$ if the rate of occurrence of interrupts is decreasing, indicating that the supercomputer is improving such as experiencing infant mortality phase of the reliability bathtub curve;
- $\beta$ is greater than one, if the rate of occurrence of interrupts is increasing, indicating that the supercomputer is deteriorating such as experiencing wear out phase of the reliability bathtub curve; and
- $\beta$ is equal to one, if the rate of occurrence of interrupts remains constant, indicating that the supercomputer is stable.

The intensity scaling factors can take the form of random variables varying with interrupts or varying with operation time or simply positive constants. For the purpose of this thesis, a reasonable assumption for the preliminary analysis can be that the scaling factors take the form of constants, i.e., $s_i = s$ for $i = 1, 2, 3, \ldots N(t)$ and $r_j = r$ for $j = 1, 2, 3, \ldots M(t)$. Substituting these values and the base line intensity function given by Equation (5) in Equation (4), the intensity function modelled by GPIM is

$$\rho(t) = \lambda \beta t^{\beta-1} s^{N(t)} r^{M(t)}$$

(6)

where $s$ and $r$ are the intensity scaling factor of emergency and preventive maintenance, respectively. They indicate the maintainability parameters of the supercomputer and take positive real values. Particularly,

- $s$ and $r$ lie in the interval $(0,1)$ if the failure intensity reduces after each maintenance action indicating that the maintenance has a positive effect on the performance of the supercomputer;
- $s$ and $r$ are greater than one, if the failure intensity increases after each maintenance action indicating that the maintenance has a negative effect on the performance of the supercomputer; and
- $s$ and $r$ are equal to one, if the failure intensity does not change after a maintenance action indicating that the maintenance does not change the performance of the supercomputer. This is a special case of GPIM where the maintenance restores the supercomputer to AGAO situation and is modelled by NHPP.
GPIM is used to fit the maintenance records for each of the four supercomputers, namely, Athena, Jaguar, Jaguar PF and Kraken. The next two sections elaborate on the method used to estimate the parameters of the GPIM model.

3.1.1 Maximum likelihood estimate (MLE)

Now that the GPIM model is used and the failure intensity function is defined, the next step is to estimate the reliability and maintainability parameters. The values of the parameters have to be such that the intensity function with these parameter values should be the closest statistical model that can replicate the interruption process of each of the four supercomputers. The two general methods used for parameter estimation are least-square estimation and MLE. The latter is the most common method used for wide range of applications. MLE of a parameter is that value of the unknown parameter that results in the highest probability of obtaining the observed data.

\[ T_1, T_2, T_3, \ldots, T_n \] are the successive time to interrupts and form a random sample obtained from a distribution that depends on four unknown parameters \( \lambda, \beta, s, \) and \( r \) with a probability density function \( f(T_k | \lambda, \beta, s, r) \). The joint probability density function of \( T_1, T_2, T_3, \ldots, T_n \) is called the likelihood function,

\[
L(\lambda, \beta, s, r | Q(t)) = f(T_1, T_2, T_3, \ldots, T_n | \lambda, \beta, s, r)
\]  

(7)

where \( Q(t) \) represents the observed maintenance records. Equation (7) is read as likelihood of the parameters \( \lambda, \beta, s, \) and \( r \) given the maintenance records of the supercomputer is equal to the likelihood or probability of observing the given data as a function of these parameters. The MLE of the four parameters are the values of these parameters that maximize the likelihood function. The likelihood function considering that preventive maintenance as a right censored data is

\[
L(\lambda, \beta, s, r | Q(t)) = \exp\left( -\int_0^{T_n} \rho(t) dt \right) \prod_{k=1}^{Z(t)} \rho(T_k | \lambda, \beta, s, r)^{c_k}
\]

(8)

where \( Z(t) = M(T) + N(T) \) and \( c_k \) is the censor indicator. \( c_k = 1 \) for emergency maintenance and \( c_k = 0 \) for preventive maintenance.

Maximizing the product can get quite tedious and hence we maximize the log likelihood, an equivalent of likelihood due to the fact that logarithm is an increasing function. From Equation (8), the log likelihood function is given by

\[
l(\lambda, \beta, s, r | Q(t)) = \sum_{i=1}^{Z(t)} c_k \{ \ln \lambda + \ln \beta + (\beta - 1)\ln t_k + N(t)\ln s + M(t)\ln r \}
\]

\[
- \sum_{k=1}^{Z(t)} s^{N(t_k)} r^{M(t_k)} \lambda^{t_k} \beta^{t_k - t_{k-1}}
\]

(9)
3.1.2 Optimizing the log-likelihood

GPIM can get complex, especially if it is required to fit about 50 to 100 maintenance records. Thus, MATLAB is used to find the parameters that optimize the log-likelihood function of GPIM. MATLAB is a short form of matrix laboratory which is an interactive numerical computing environment using fourth-generation programming language. It is developed by Math Works and allows various matrix computations, plotting of functions, algorithm implementation and can interface with programs written in different languages.

We used \textit{fmincon}, a predefined function in the optimization toolbox of MATLAB to find the reliability and maintainability parameters that optimize the nonlinear log-likelihood function given by Equation (9), to fit the maintenance records of each of the four supercomputers. \textit{fmincon} minimizes the intensity function and takes into consideration the constraints on the variables or parameters and is generally called constrained nonlinear optimization function. Thus to maximize the log-likelihood function the negative log-likelihood function is taken. It optimizes based on the four optimization algorithms, trust-region-reflective algorithm, active set algorithm, interior-point algorithm, and sequential quadratic programming algorithm. The constraints considered while maximizing the log-likelihood of the GPIM are that each of four parameters, \( \lambda, \beta, s, \) and \( r \) are positive.

3.2 Time Trend Test

A time trend test is can statistically determine the type of trend exhibited by the interrupt times. The trend can be decreasing, indicating that the supercomputer is exhibiting an improving behavior, or increasing, indicating that the supercomputer is exhibiting a deteriorating performance. If there is no trend then the system is stable and the process becomes HPP, a special case of NHPP. In this thesis we use Laplace trend test to check whether the interrupts of each of the four supercomputers exhibit a trend.

The hypothesis of the Laplace trend test is given as:
Null hypothesis: \( H_0 \) is that the supercomputer does not exhibit any time trend that is the supercomputer is stable.
Alternate hypothesis: \( H_a \) is that the supercomputer exhibits a time trend that is the supercomputer is either improving or deteriorating.

Laplace trend test is defined for both Type I and Type II data. If the data recording is terminated at a predetermined time, then the maintenance record of that supercomputer is said to be Type I censored. If the data recording is terminated at a point of occurrence of an interrupt, then the supercomputer is said to be Type II censored. The maintenance records of the four supercomputers are terminated at the occurrence of an interrupt, rather than at a predetermined time, thus all the four supercomputers are considered to be Type II censored.

The Laplace test statistic for Type II censored data is
where $n$ is the total number of interrupts considering both emergency and preventive maintenance. The null hypothesis, $H_0$ is rejected if $U$ is out of the range $[Z_{\alpha/2}, Z_{1-\alpha/2}]$ where $\alpha$ is the significance level of the hypothesis test and $Z$ is the value at specified $\alpha$ taken from standard normal distribution table. $\alpha$ is related to confidence limit of the hypothesis test by the equation, $\alpha = 1 - (confidence\ limit/100)$. At 5% level of significance, $\alpha = 0.05$ and $[Z_{0.025},Z_{0.975}] = [-1.96,1.96]$.

If the null hypothesis is rejected and there is enough evidence to accept that there is a trend in the interrupt process, it is of interest to know if the supercomputer has an increasing trend or a decreasing trend. Specifically,

- $U < 0$ if there is a decreasing trend that is the interrupts are becoming less likely and the interval between interrupts are getting larger indicating that the supercomputer is improving.
- $U > 0$ if there is an increasing trend that is the interrupts are becoming more likely and the interval between interrupts are getting smaller indicating that the supercomputer is deteriorating.

3.3 Results

The log-likelihood function presented by Equation (9) is used to fit the maintenance records of each of the four supercomputers. The MLE of the reliability parameters, scale parameter, $\lambda$, and shape parameter $\beta$, corresponding to the Weibull distribution corresponding of the baseline intensity function are estimated. The maintainability parameters, $s$ and, $r$ representing the scaling factors of the emergency and preventive maintenance, respectively are estimated.

Table 2 presents the estimates of the four parameters along with the maximum value of the log-likelihood function for each of the four supercomputers, Athena, Jaguar, Jaguar PF, and Kraken. Table 3 presents the Laplace trend test results for each of the four supercomputers. The rest of this section presents the observations on each supercomputer based on the reliability and maintenance parameters shown in Table 2 and the trend test results shown in Table 3.
Table 2: GPIM model parameter estimates

<table>
<thead>
<tr>
<th>Supercomputer</th>
<th>$\hat{\lambda}$</th>
<th>$\hat{\beta}$</th>
<th>$\hat{s}$</th>
<th>$\hat{r}$</th>
<th>log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Athena</td>
<td>0.0026</td>
<td>1.0415</td>
<td>0.9467</td>
<td>1.6231</td>
<td>-228.0823</td>
</tr>
<tr>
<td>Jaguar</td>
<td>0.0350</td>
<td>0.7500</td>
<td>1.0882</td>
<td>0.8097</td>
<td>-356.3360</td>
</tr>
<tr>
<td>Jaguar PF</td>
<td>0.0025</td>
<td>1.2063</td>
<td>1.0134</td>
<td>0.9965</td>
<td>-686.8591</td>
</tr>
<tr>
<td>Kraken</td>
<td>0.0048</td>
<td>1.0814</td>
<td>1.0251</td>
<td>0.9585</td>
<td>-333.3181</td>
</tr>
</tbody>
</table>

Table 3: Laplace trend test results

<table>
<thead>
<tr>
<th>Supercomputer</th>
<th>$U$</th>
<th>5% significance [-1.96, 1.96]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Athena</td>
<td>-0.0821</td>
<td>Stationary</td>
</tr>
<tr>
<td>Jaguar</td>
<td>-2.0388</td>
<td>Improving</td>
</tr>
<tr>
<td>Jaguar PF</td>
<td>2.5000</td>
<td>Deteriorating</td>
</tr>
<tr>
<td>Kraken</td>
<td>-0.2131</td>
<td>Stationary</td>
</tr>
</tbody>
</table>

3.3.1 Athena

The GPIM model fitted to the maintenance records of Athena show that
- the shape parameter, $\hat{\beta}$ is almost equal to one and thus, Athena exhibits a constant failure intensity and is stationary. This is also confirmed by the Laplace trend test;
- the emergency maintenance scaling factor, $\hat{s}$ is 0.9467, which is less than one. Thus, each emergency maintenance slightly decreases the failure intensity function and has a positive effect on Athena’s performance; and
- the preventive maintenance scaling factor, $\hat{r}$ is 1.6231, which is higher than one. Thus, each preventive maintenance prominently increases the failure intensity function and has a negative effect of Athena’s performance.

Figure 8 shows the change in failure intensity with respect to time for Athena, as modelled by GPIM. It can be seen that the failure intensity between any two consecutive interrupts is almost constant but has a sudden change in magnitude at each interrupt, indicating the effectiveness of the maintenance action. We can see steep increase in failure intensity after each preventive maintenance action representing its scaling factor of 1.6231 and slight decrease in failure intensity after each emergency maintenance action representing its scaling factor of 0.9467. Considering the first ten interrupts,
- Athena has preventive maintenance interrupts at 471, 1109, and 2093 and these points are marked by steep rise in intensity function; and
- Athena has emergency maintenance interrupts at 626, 632, 771, 881, 1329, 2093, 2226, and 2617 and these points are marked by slight fall in intensity function.
3.3.2  Jaguar

The GPIM model fitted to the maintenance records of Jaguar show that

- the shape parameter, \( \hat{\beta} \) is 0.75, which is lesser than one. Thus, Jaguar exhibits a decreasing failure intensity and is improving as the time progresses. This is also confirmed by the Laplace trend test;
- the emergency maintenance scaling factor, \( \hat{s} \) is 1.0882, which is greater than one. Thus, each preventive maintenance slightly increases the failures intensity function and has a negative effect of Jaguar’s performance; and
- the preventive scaling factor, \( \hat{r} \) is 0.8097, which is lesser than one. Thus, each preventive maintenance prominently decreases the failure intensity function and has a positive effect on Jaguar’s performance.

Figure 9 shows the change in failure intensity of Jaguar with respect to time, as modelled by GPIM. It can be seen that the average failure intensity is decreasing with time. Considering the first ten interrupts,

- Jaguar has preventive maintenance interrupts at 34, 152, 191, and 680 and these points are marked by steep fall in intensity function due to the scaling factor of 0.8097; and
- Jaguar has emergency maintenance interrupts at 170, 173, 390, 411, 577, and 732 and these points are marked by slight rise in intensity function due to the scaling factor of 1.0882.
3.3.3 Jaguar PF

The GPIM model fitted to the maintenance records of Jaguar PF show that
- the shape parameter, $\hat{\beta}$ is 1.2063, which is greater than one. Thus, Jaguar PF exhibits an increasing failure intensity and is deteriorating as the time progresses. This is also confirmed by the Laplace trend test; and
- the emergency maintenance scaling factor, $\hat{s}$ is 1.0134 and preventive scaling factor, $\hat{r}$ is 0.9965. Thus, the maintenance actions do not bring in any considerable change in the failure intensity (performance) of Jaguar PF. This indicates that each maintenance on Jaguar PF restores it to AGAO.

Figure 10 shows the change in failure intensity with respect to time for Jaguar PF, as modelled by GPIM. It can be seen that the average failure intensity of Jaguar PF is increasing with time. It can also be seen that there is almost no sudden change in the failure intensity at each interrupt, indicating no change in failure intensity after each maintenance. There seems to be a change in the slope of the curve at about 7200 hours indicating some major change in the system that caused a faster deterioration of the system. The reason for this has to be further explored and is not available with the existing maintenance records.

3.3.4 Kraken

The GPIM model fitted to the maintenance records of Kraken show that
- the shape parameter, $\hat{\beta}$ is almost equal to one. Thus, Kraken exhibits a constant failure intensity and is stationary. This is also confirmed by the Laplace trend test;
Figure 10: Failure intensity of Jaguar PF vs time - GPIM

- the emergency maintenance scaling factor, $\hat{s}$ is 1.0251, which is greater than one. Thus, each preventive maintenance slightly increases the failures intensity function and has a negative effect on Kraken’s performance; and
- the preventive maintenance scaling factor, $\hat{r}$ is 0.9585, which is lesser than one. Thus, each preventive maintenance slightly increases the failure intensity function and has a positive effect on Kraken’s performance.

Figure 11 shows the change in failure intensity with respect to time for Kraken, as modelled by GPIM. Considering the first ten interrupts,
- Kraken has preventive maintenance interrupts at 862, 1018, and 1540 and these points are marked by slight fall in intensity function due to scaling factor of 0.9585; and
- Kraken has emergency maintenance interrupts at 64, 121, 320, 466, 958, 1054, and 1728 and these points are marked by slight rise in intensity function due to scaling factor of 1.0251.

It has to be noted that at about 7200 operating hours, the rate of change of failure intensity gets steeper indicating that Jaguar PF performance decreases at an increased rate. This may be explained by the possibility of addition of a hardware component that is not compatible with the main system.
3.4 Summary

The maintenance records of each of the four supercomputers are modelled by GPIM and the trend in occurrence of interrupts is statistically tested by the Laplace trend test. It has been observed that Athena and Kraken experienced a stationary performance while Jaguar experienced an improvement in performance and Jaguar PF experiences a deterioration in performance during the period October 2009 to December 2010.

The change in failure intensity for the four supercomputers after each emergency maintenance is very marginal and can be considered to be negligible for the purpose of this thesis. Thus, it can be concluded that each emergency maintenance restores the supercomputers to AGAO state. The results are in acceptance with the fact that emergency maintenance is only intended to bring back the supercomputer to normal operating condition by fixing a problem that caused its failure.

It can be observed that the preventive maintenance has varied effects on each of the four supercomputers showing that either the same maintenance protocol for all the supercomputers is not an effective solution or that the maintenance protocol followed for each supercomputer is different. Preventive maintenance has a negative impact on Athena’s performance as the failure intensity of Athena is prominently increased after each preventive maintenance. It may be due to the fact that the maintenance personnel are conducting some performance determination tests during a preventive maintenance that are inducing additional failures. The failure intensity of Jaguar is prominently decreased after each preventive maintenance, creating a positive impact on its performance. This may be
explained by the fact that maintenance personnel could identify the incipient failures and successfully address them. The change in failure intensity for Jaguar PF and Kraken after each preventive maintenance is very marginal and can be considered to be negligible for the purpose of this thesis. This indicates that the preventive maintenance on Jaguar PF and Kraken could not effectively identify the incipient failures and is serving only as an inspection done for formality.
4 REPAIR TIMES ANALYSIS

This chapter focuses on the analysis of repair times of the four supercomputers. Section 4.1 describes the analysis approach followed in this chapter followed by Section 4.2 providing the criteria for selection of a model from the probability models. Sections 4.3 presents the probability models considered. Sections 4.4 - 4.6 analyze the repair times of emergency maintenance, preventive maintenance, and both emergency and preventive maintenance together, respectively, for the four supercomputers. Section 4.7 summarizes the results obtained in this chapter and presents the observations.

4.1 Analysis Approach

The chapter deals with the maintenance state of the supercomputer and analyzes the maintenance times. Emergency maintenance is a repair done when there is an interruption to the normal functioning of the supercomputer due to failure of a component(s), it involves diagnosing the problem and then fixing it to restore the system to its normal operating condition. In this case, the repair time can vary a lot depending on the inventory required to tackle the problem at hand. Preventive maintenance is a pre scheduled maintenance activity with a fixed protocol and most often than not, the inventory required is procured before starting the maintenance. In this case, there is a high possibility that the variation of maintenance time is low. Therefore, it is proposed to model the repair times of emergency and preventive maintenance separately. JMP Pro software is used to model the probability life distributions of the repair times for each supercomputer. Figure 12 shows the schematic representation of the maintenance process of the supercomputer.

![Figure 12: Schematic representation of the supercomputer maintenance process](image)

4.2 Model Selection Measures

This section elaborates the measures used for selecting the best probability distribution that fits the repair times, among the six common distributions considered, exponential, logistic, log-logistic, lognormal, normal, and, Weibull distributions. Traditionally, log-likelihood
values are calculated for all the distributions considered and the one with highest log-likelihood value is chosen to be the best distribution that fits the data. But, this method has a limitation in that the log-likelihood can be increased by adding parameters to a model i.e., a model with higher number of parameters always has a higher log-likelihood value. Higher number of parameters in a model may lead to over fitting which is not always good. There should be a proper balance between fitting the best model, while avoiding the over fitting.

In order to address the limitation of the log-likelihood measure in selecting the best model among the considered models we used AICc (Akaike, 1974) and BIC (Schwarz, 1978) measures introduced by Akaike Hirotugu and Gideon Schwarz, respectively. Both these measures are based on the log-likelihood value and take the form, \(2[-\text{log} \text{likelihood} + pc]\) where \(p\) is the number of parameters in the model and \(c\) is \([n/(n - p - 1)]\) for AICc and \(\ln n\) for BIC, \(n\) being the total number of maintenance activities. AICc is a measure of relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model and a model with lower AICc is closer to the truth. BIC is a measure of the posterior probability of a model being true under a certain Bayesian setup and a model with lower BIC is more likely to be the true model.

BIC penalizes a complex model more heavily than AICc and the only way they may disagree is when BIC chooses a smaller model (model with lesser number of parameters) than AICc. AICc always tends to choose a bigger model (model with higher number of parameters), regardless of sample size whereas BIC has very little chance of choosing a bigger model if sample size is just sufficient to fit a model otherwise it usually tends to select a smaller model. AICc is a better measure in situations where a false negative selection would be more misleading than a false positive and BIC is a better measure otherwise (Burnham & Anderson, 2004).

Without loss of generality \(-2\text{loglikelihood}\) is used instead of \(\text{loglikelihood}\) whenever it is used against AICc and BIC to maintain the same sign and make it comparable to AICc and BIC in terms of magnitude.

4.3 Tested Distributions

Six probability distributions which are widely used to model the reliability data are tested for their fitness to the repair time data of each supercomputer. The six probability distributions are: exponential, logistic, log-logistic, lognormal, normal, and, Weibull distributions. It is observed from our results in this chapter that lognormal and Weibull distributions always fit the repair time data better than other four distributions. Therefore, the rest of this section introduces the lognormal and Weibull distributions.

Let \(T\) be a continuous random variable representing the time to maintenance, be it emergency or preventive maintenance with a probability density function of \(h(t)\), then the cumulative distribution function, \(H(t)\) is the probability that the maintenance time is less
than time $t$. Let $MMT$ and $t_{med}$ denote the mean and median of the maintenance times of the supercomputer.

### 4.3.1 Weibull distribution

The cumulative density function, $H(t)$ for Weibull distribution with scale parameter, $\lambda$ and shape parameter, $\beta$ is

$$H(t) = \exp\left(-\lambda t^\beta\right)$$

The mean, median, standard deviation of time to maintenance of supercomputer, defined by the Weibull distribution are

$$MMT = \int_0^\infty (1 - H(t))dt = \int_0^\infty \exp\left(-\lambda t^\beta\right)dt$$

$$t_{med} = \left(\frac{\ln(2)}{\lambda}\right)^{\frac{1}{\beta}}$$

$$S_{dev} = \sqrt{\frac{\Gamma\left(1 + \frac{2}{\beta}\right)}{\Gamma\left(1 + \frac{1}{\beta}\right)^2}}$$

### 4.3.2 Lognormal distribution

The cumulative density function of failure, $H(t)$ for lognormal distribution with location parameter, $\mu$ and shape parameter, $\sigma$ is

$$H(t) = \Phi_{nor}\left(\frac{\log(t) - \mu}{\sigma}\right)$$

$$\Phi_{nor}(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$

where $\Phi_{nor}(z)$ represent the cumulative density function of a normal distribution. The mean, median, standard deviation of time to maintenance of supercomputer, defined by the lognormal distribution are

$$MMT = t_{med} \exp\left(\frac{\sigma^2}{2}\right)$$

$$t_{med} = e^\mu$$

$$S_{dev} = \sqrt{(e^{\sigma^2} - 1)e^{2\mu+\sigma^2}}$$

Once the mean and standard deviation are calculated, the 95% confidence interval $CI_{0.95}$ (this is equivalent to 5% $\alpha$) for the mean of maintenance times is
\[ CI_{0.95} = MMT \pm Z_{\frac{\alpha}{2}} \left( \frac{\sigma}{\sqrt{n}} \right) \]

where \( Z_{\frac{\alpha}{2}} \) is the value corresponding to \( \frac{\alpha}{2} \) in the standard normal distribution table and \( n \) is the total number of maintenance activities.

### 4.4 Emergency Maintenance

This section shows the analysis done on the maintenance time distributions of the emergency maintenance activities. For each of the four supercomputers, the emergency maintenance records are fitted by the six probability distributions (in Section 4.3) and are compared by three measures, \((-2\text{loglikelihood})\), AICc and BIC (in Section 4.2). Sections 4.4.1 - 4.4.4 presents the results for the four supercomputers, Athena, Jaguar, Jaguar PF, and Kraken, respectively and Section 4.7 summarizes the results.

#### 4.4.1 Athena

For Athena, there are 36 emergency maintenance actions. The \(-2\text{loglikelihood}\), AICc and BIC values for the six tested distributions are calculated. Table 4 shows the results for Athena. Lognormal distribution best fits the emergency maintenance times of Athena with the smallest \(-2\text{loglikelihood}\), AICc and BIC values. The fitted lognormal distribution with \( \mu = 0.42 \) and \( \sigma = 1.07 \) is given by the below equation.

\[ \hat{H}(t) = \Phi_{\text{nor}} \left( \frac{\log(t) - 0.42}{1.07} \right) \]

The estimate of mean emergency maintenance time \((MEMT)\) for Athena is 3.15 hours, standard deviation is 5.66 and the median is 1.53 hours. This shows that the emergency maintenance time distribution is heavily skewed towards positive side. The 95% confidence interval of the estimated mean repair time is (1.61, 6.84). On an average, it will take 3.15 hours to perform an emergency maintenance on Athena and there is a huge variation in the repair times.

| Table 4: Distributions of emergency maintenance repair times for Athena |
|-----------------|-----------------|-------|-------|
| **Distribution** | **(-2Loglikelihood)** | **AICc** | **BIC** |
| Lognormal       | 143.37           | 147.75 | 150.48 |
| Log-logistic    | 145.67           | 150.05 | 152.78 |
| Exponential     | 150.59           | 152.71 | 154.15 |
| Weibull         | 149.02           | 153.39 | 156.13 |
| Logistic        | 189.05           | 193.42 | 196.16 |
| Normal          | 198.73           | 203.10 | 205.84 |
4.4.2 Jaguar

For Jaguar, there are 58 emergency maintenance actions. The \(-2\log\text{likelihood}, \text{AICc}\) and BIC values for the six tested distributions are calculated. Table 5 shows the results for Jaguar. Lognormal distribution best fits the emergency maintenance times of Jaguar with the smallest \(-2\log\text{likelihood}, \text{AICc}\) and BIC values.

The fitted lognormal distribution with \(\mu = 1.14\) and \(\sigma = 0.76\), is
\[
\hat{H}(t) = \Phi_{nor}\left(\frac{\log(t) - 1.14}{0.76}\right)
\] (22)

The estimate of \(\text{MEMT}\) of Jaguar is 4.17 hours, standard deviation is 3.69 and the median is 3.12 hours. On an average, it takes 4.17 hours to perform an emergency maintenance on Jaguar. The 95% confidence interval of the estimated mean repair time is (2.90, 5.02).

Table 5: Distributions of emergency maintenance repair times for Jaguar

<table>
<thead>
<tr>
<th>Distribution</th>
<th>(-2Loglikelihood)</th>
<th>AICc</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lognormal</td>
<td>231.25</td>
<td>235.48</td>
<td>239.22</td>
</tr>
<tr>
<td>Log-logistic</td>
<td>232.53</td>
<td>236.77</td>
<td>240.51</td>
</tr>
<tr>
<td>Weibull</td>
<td>237.46</td>
<td>241.69</td>
<td>245.43</td>
</tr>
<tr>
<td>Exponential</td>
<td>250.89</td>
<td>252.97</td>
<td>254.88</td>
</tr>
<tr>
<td>Logistic</td>
<td>250.71</td>
<td>254.94</td>
<td>258.69</td>
</tr>
<tr>
<td>Normal</td>
<td>264.03</td>
<td>268.26</td>
<td>272.00</td>
</tr>
</tbody>
</table>

4.4.3 Jaguar PF

For Jaguar PF, there are 133 emergency maintenance actions. The \(-2\log\text{likelihood}, \text{AICc}\) and BIC values for the six tested distributions are calculated. Table 6 shows the results for Jaguar PF. Lognormal distribution best fits the emergency maintenance times of Jaguar PF with the smallest \(-2\log\text{likelihood}, \text{AICc}\) and BIC values. The fitted lognormal distribution with \(\mu = 1.12\) and \(\sigma = 0.58\) is
\[
\hat{H}(t) = \Phi_{nor}\left(\frac{\log(t) - 1.12}{0.58}\right)
\] (23)

The estimate of \(\text{MEMT}\) of Jaguar PF is 3.62 hours, standard deviation is 2.29 and the median is 3.06 hours. Thus, it will take 3.62 hours on an average to perform an emergency maintenance on Jaguar PF. The 95% confidence interval of the estimated mean repair time is (3.25, 4.04).
<table>
<thead>
<tr>
<th>Distribution</th>
<th>(-2Loglikelihood)</th>
<th>AICc</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lognormal</td>
<td>491.15</td>
<td>495.25</td>
<td>500.86</td>
</tr>
<tr>
<td>Log-logistic</td>
<td>492.89</td>
<td>496.99</td>
<td>502.60</td>
</tr>
<tr>
<td>Weibull</td>
<td>522.02</td>
<td>526.12</td>
<td>531.73</td>
</tr>
<tr>
<td>Logistic</td>
<td>543.45</td>
<td>547.55</td>
<td>553.15</td>
</tr>
<tr>
<td>Normal</td>
<td>568.53</td>
<td>572.63</td>
<td>578.23</td>
</tr>
<tr>
<td>Exponential</td>
<td>578.62</td>
<td>580.65</td>
<td>583.47</td>
</tr>
</tbody>
</table>

4.4.4 Kraken

For Kraken, there are 133 emergency maintenance actions. The $-2\log\text{likelihood}$, AICc and BIC values for the six tested distributions are calculated. Table 7 shows the results for Kraken. Lognormal distribution best fits the emergency maintenance times of Kraken with the smallest $-2\log\text{likelihood}$, AICc and BIC values. The fitted lognormal distribution with $\mu = 0.52$ and $\sigma = 0.96$ is

$$\hat{H}(t) = \Phi_{\text{nor}}\left(\frac{\log(t) - 0.52}{0.96}\right)$$  \hspace{1cm} (24)

The estimate of $MEMT$ is 2.67 hours, standard deviation is 3.29 and the median is 1.68 hours. On an average, it takes 2.67 hours on an average to perform an emergency maintenance on Kraken but there is a huge variation in the repair times. The 95% confidence interval of the estimated mean repair time is (1.76, 4.19).

<table>
<thead>
<tr>
<th>Distribution</th>
<th>(-2Loglikelihood)</th>
<th>AICc</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lognormal</td>
<td>201.29</td>
<td>205.53</td>
<td>209.24</td>
</tr>
<tr>
<td>Log-logistic</td>
<td>204.36</td>
<td>208.60</td>
<td>212.30</td>
</tr>
<tr>
<td>Exponential</td>
<td>214.98</td>
<td>217.05</td>
<td>218.95</td>
</tr>
<tr>
<td>Weibull</td>
<td>214.97</td>
<td>219.21</td>
<td>222.92</td>
</tr>
<tr>
<td>Logistic</td>
<td>265.25</td>
<td>269.49</td>
<td>273.19</td>
</tr>
<tr>
<td>Normal</td>
<td>271.33</td>
<td>275.57</td>
<td>279.27</td>
</tr>
</tbody>
</table>

4.5 Preventive Maintenance

This section shows the analysis done on the maintenance time distributions of the preventive maintenance activities. For the maintenance records of each of the four supercomputers, the six probability distributions (in Section 4.3) are tested and compared by three measures $-2\log\text{likelihood}$, AICc and BIC (in Section 4.2). Sections 4.5.1 - 4.5.4
presents the results for the four supercomputers, Athena, Jaguar, Jaguar PF, and Kraken, respectively and Section 4.7 summarizes the results.

### 4.5.1 Athena

For Kraken, there are 12 preventive maintenance actions. The $-2\log\text{likelihood}$, AICc and BIC values for the six tested distributions are calculated. Table 7 shows the results for Kraken. Lognormal distribution best fits the preventive maintenance times of Kraken with the smallest $-2\log\text{likelihood}$, AICc and BIC values. The fitted lognormal distribution with $\mu = 0.68$ and $\sigma = 0.46$ is

$$\hat{H}(t) = \Phi_{nor} \left( \frac{\log(t) - 0.68}{0.46} \right)$$  \hspace{1cm} (25)

The estimate of mean preventive maintenance time ($MPMT$) is 2.2 hours, standard deviation is 1.07 and the median is 1.97 hours. On average, it takes 2.2 hours to perform a preventive maintenance activity on Athena. The 95% confidence interval of the estimated mean repair time is (1.56, 3.21).

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$-2\log\text{likelihood}$</th>
<th>AICc</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lognormal</td>
<td>29.19</td>
<td>34.69</td>
<td>33.99</td>
</tr>
<tr>
<td>Weibull</td>
<td>30.23</td>
<td>35.73</td>
<td>35.03</td>
</tr>
<tr>
<td>Log-logistic</td>
<td>30.32</td>
<td>35.82</td>
<td>35.12</td>
</tr>
<tr>
<td>Normal</td>
<td>31.56</td>
<td>37.06</td>
<td>36.36</td>
</tr>
<tr>
<td>Logistic</td>
<td>32.52</td>
<td>38.02</td>
<td>37.32</td>
</tr>
<tr>
<td>Exponential</td>
<td>39.30</td>
<td>41.75</td>
<td>41.70</td>
</tr>
</tbody>
</table>

### 4.5.2 Jaguar

For Kraken, there are 31 preventive maintenance actions. The $-2\log\text{likelihood}$, AICc and BIC values for the six tested distributions are calculated. Table 7 shows the results for Kraken. Lognormal distribution best fits the preventive maintenance repair times of Kraken with the smallest $-2\log\text{likelihood}$, AICc and BIC values. The fitted lognormal distribution with $\mu = 1.59$ and $\sigma = 0.47$ is

$$\hat{H}(t) = \Phi_{nor} \left( \frac{\log(t) - 1.59}{0.47} \right)$$ \hspace{1cm} (26)

The estimate of $MPMT$ is 5.45 hours, standard deviation 2.69 and the median is 4.89 hours. On average, it takes 5.45 hours to perform a preventive maintenance activity on Jaguar. The 95% confidence interval of the estimated mean repair time is (4.42, 6.83).
Table 9: Distributions of preventive maintenance repair times for Jaguar

<table>
<thead>
<tr>
<th>Distribution</th>
<th>(-2Loglikelihood)</th>
<th>AICc</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lognormal</td>
<td>139.06</td>
<td>143.49</td>
<td>145.93</td>
</tr>
<tr>
<td>Log-logistic</td>
<td>139.90</td>
<td>144.33</td>
<td>146.77</td>
</tr>
<tr>
<td>Weibull</td>
<td>143.79</td>
<td>148.22</td>
<td>150.66</td>
</tr>
<tr>
<td>Logistic</td>
<td>145.38</td>
<td>149.81</td>
<td>152.25</td>
</tr>
<tr>
<td>Normal</td>
<td>148.70</td>
<td>153.12</td>
<td>155.56</td>
</tr>
<tr>
<td>Exponential</td>
<td>167.16</td>
<td>169.30</td>
<td>170.60</td>
</tr>
</tbody>
</table>

4.5.3 Jaguar PF

For Jaguar PF, there are 42 preventive maintenance actions. The $-2\log\text{likelihood}$, AICc and BIC values for the six tested distributions are calculated. Table 7 shows the results for Jaguar PF. Lognormal distribution best fits the preventive maintenance repair times of Jaguar PF with the smallest $-2\log\text{likelihood}$, AICc and BIC values.

The fitted Weibull distribution with $\lambda = 0.0555$ and $\beta = 1.1800$, is given by

$$\hat{H}(t) = 1 - \exp\left(-0.0555(t)^{1.1800}\right)$$

(27)

Table 10: Distributions of preventive maintenance repair times for Jaguar PF

<table>
<thead>
<tr>
<th>Distribution</th>
<th>(-2Loglikelihood)</th>
<th>AICc</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weibull</td>
<td>260.87</td>
<td>265.18</td>
<td>268.35</td>
</tr>
<tr>
<td>Logistic</td>
<td>265.19</td>
<td>269.50</td>
<td>272.67</td>
</tr>
<tr>
<td>Normal</td>
<td>266.77</td>
<td>271.08</td>
<td>274.25</td>
</tr>
<tr>
<td>Lognormal</td>
<td>267.01</td>
<td>271.32</td>
<td>274.49</td>
</tr>
<tr>
<td>Log-logistic</td>
<td>268.55</td>
<td>272.86</td>
<td>276.02</td>
</tr>
<tr>
<td>Exponential</td>
<td>279.87</td>
<td>281.97</td>
<td>283.61</td>
</tr>
</tbody>
</table>

The estimate of $MPMT$ is 10.3 hours, standard deviation is 2.4 and the median is 9.85 hours. On an average, it will take 10.3 hours to perform a preventive maintenance activity on Jaguar PF. The 95% confidence interval of the estimated mean repair time is (8.81, 12.21).

4.5.4 Kraken

For Kraken, there are 23 preventive maintenance actions. The $-2\log\text{likelihood}$, AICc and BIC values for the six tested distributions are calculated. Table 7 shows the results for Kraken. Lognormal distribution best fits the preventive maintenance repair times of Kraken.
with the smallest \(-2\text{loglikelihood}\), AICc and BIC values. The fitted lognormal distribution with \(\mu = 1.50\) and \(\sigma = 0.51\) is

\[
\hat{H}(t) = \Phi_{\text{nor}}\left(\frac{\log(t) - 1.5}{0.51}\right)
\]  

(28)

Table 11: Distributions of preventive maintenance repair times for Kraken

<table>
<thead>
<tr>
<th>Distribution</th>
<th>(-2Loglikelihood)</th>
<th>AICc</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lognormal</td>
<td>98.64</td>
<td>103.27</td>
<td>104.82</td>
</tr>
<tr>
<td>Log-logistic</td>
<td>99.25</td>
<td>103.88</td>
<td>105.43</td>
</tr>
<tr>
<td>Weibull</td>
<td>103.07</td>
<td>107.70</td>
<td>109.25</td>
</tr>
<tr>
<td>Logistic</td>
<td>106.63</td>
<td>111.26</td>
<td>112.81</td>
</tr>
<tr>
<td>Normal</td>
<td>108.42</td>
<td>113.05</td>
<td>114.60</td>
</tr>
<tr>
<td>Exponential</td>
<td>115.98</td>
<td>118.18</td>
<td>119.07</td>
</tr>
</tbody>
</table>

The estimate of \(MPMT\) is 5.11 hours, standard deviation 2.77 and the median is 4.50 hours. On an average, it will take 5.11 hours to perform a preventive maintenance activity on Kraken. The 95% confidence interval of the estimated mean repair time is (3.88, 6.89).

4.6 Emergency and Preventive Maintenance

This section shows the analysis done on the repair time distributions of the both emergency and preventive maintenance activities considered together. For each data set of a supercomputer, the six probability distributions (in Section 4.3) are tested and compared by three measures \(-2\text{loglikelihood}\), AICc and BIC (in Section 4.2). Sections 4.4.1 - 4.4.4 presents the results for the four supercomputers, Athena, Jaguar, Jaguar PF, and Kraken, respectively and Section 4.7 summarizes the results.

4.6.1 Athena

For Athena, there are 48 maintenance actions. The \(-2\text{loglikelihood}\), AICc and BIC values for the six tested distributions are calculated. Table 4 shows the results for Athena. Lognormal distribution best fits the maintenance repair times of Athena with the smallest \(-2\text{loglikelihood}\), AICc and BIC values. The fitted lognormal distribution with \(\mu = 0.61\) and \(\sigma = 1.20\) is

\[
\hat{H}(t) = \Phi_{\text{nor}}\left(\frac{\log(t) - 0.61}{1.20}\right)
\]  

(29)

The estimate of mean maintenance time \(MMT\) is 3.80 hours, standard deviation is 6.86 and the median is 1.84 hours. On an average, it will take 3.80 hours to perform maintenance on Athena and there is a huge variation in the repair times.
Table 12: Distributions of all the maintenance repair times for Athena

<table>
<thead>
<tr>
<th>Distribution</th>
<th>-2Loglikelihod</th>
<th>AICc</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lognormal</td>
<td>207.92</td>
<td>212.19</td>
<td>215.62</td>
</tr>
<tr>
<td>Log-logistic</td>
<td>209.01</td>
<td>213.28</td>
<td>216.71</td>
</tr>
<tr>
<td>Weibull</td>
<td>216.55</td>
<td>220.82</td>
<td>224.25</td>
</tr>
<tr>
<td>Exponential</td>
<td>224.40</td>
<td>226.49</td>
<td>228.25</td>
</tr>
<tr>
<td>Logistic</td>
<td>287.05</td>
<td>291.32</td>
<td>294.75</td>
</tr>
<tr>
<td>Normal</td>
<td>309.79</td>
<td>314.07</td>
<td>317.49</td>
</tr>
</tbody>
</table>

4.6.2 Jaguar

For Jaguar, there are 89 maintenance actions. The $-2\text{loglikelihood}$, AICc and BIC values for the six tested distributions are calculated. Table 5 shows the results for Jaguar. Weibull distribution best fits the maintenance repair times of Jaguar with the smallest $-2\text{loglikelihood}$, AICc and BIC values. The fitted Weibull distribution with $\lambda = 0.1471$ and $\nu = 1.2400$ is

$$\hat{H}(t) = 1 - \exp\left(-0.1471(t^{1.2400})\right)$$

Table 13: Distributions of all the maintenance repair times for Jaguar

<table>
<thead>
<tr>
<th>Distribution</th>
<th>-2Loglikelihod</th>
<th>AICc</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weibull</td>
<td>434.39</td>
<td>438.52</td>
<td>443.37</td>
</tr>
<tr>
<td>Exponential</td>
<td>440.93</td>
<td>442.97</td>
<td>445.42</td>
</tr>
<tr>
<td>Log-logistic</td>
<td>441.45</td>
<td>445.59</td>
<td>450.43</td>
</tr>
<tr>
<td>Logistic</td>
<td>455.17</td>
<td>459.31</td>
<td>464.14</td>
</tr>
<tr>
<td>Lognormal</td>
<td>472.68</td>
<td>476.81</td>
<td>481.65</td>
</tr>
<tr>
<td>Normal</td>
<td>493.93</td>
<td>498.07</td>
<td>502.91</td>
</tr>
</tbody>
</table>

The estimate of $\text{MMT}$ is 4.69 hours and the standard deviation is 3.55. On an average, it takes 4.69 hours to perform maintenance on Jaguar.

4.6.3 Jaguar PF

For Jaguar PF, there are 175 maintenance actions. The $-2\text{loglikelihood}$, AICc and BIC values for the six tested distributions are calculated. Table 6 shows the results for Jaguar PF. Weibull distribution best fits the emergency maintenance repair times of Jaguar PF with the smallest $-2\text{loglikelihood}$, AICc and BIC values. The fitted Weibull distribution with $\alpha = 0.1204$ and $\beta = 1.2200$ is
\[ \hat{H}(t) = 1 - \exp\left(-0.1204(t)^{1.2200}\right) \]  

(31)

The estimate of \( MMT \) is 5.31 hours and the standard deviation is 4.38. Thus, it will take 3.62 hours on an average to perform an emergency maintenance on Jaguar PF.

Table 14: Distributions of all the maintenance repair times for Jaguar PF

<table>
<thead>
<tr>
<th>Distribution</th>
<th>(-2Loglikelihood)</th>
<th>AICc</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weibull</td>
<td>869.56</td>
<td>873.63</td>
<td>879.83</td>
</tr>
<tr>
<td>Log-logistic</td>
<td>894.30</td>
<td>898.37</td>
<td>904.57</td>
</tr>
<tr>
<td>Lognormal</td>
<td>897.71</td>
<td>901.78</td>
<td>907.98</td>
</tr>
<tr>
<td>Exponential</td>
<td>905.59</td>
<td>907.61</td>
<td>910.73</td>
</tr>
<tr>
<td>Logistic</td>
<td>985.37</td>
<td>989.44</td>
<td>995.64</td>
</tr>
<tr>
<td>Normal</td>
<td>1024.61</td>
<td>1028.68</td>
<td>1034.88</td>
</tr>
</tbody>
</table>

4.6.4  Kraken

For Kraken, there are 76 maintenance actions. The \(-2\loglikelihood\), AICc and BIC values for the six tested distributions are calculated. Table 7 shows the results for Kraken. Lognormal distribution best fits the maintenance repair times of Kraken with the smallest \(-2\loglikelihood\), AICc and BIC values. The fitted lognormal distribution with \( \mu = 1.02 \) and \( \sigma = 0.85 \) is

\[ \hat{H}(t) = \Phi_{nor}\left(\frac{\log(t) - 1.02}{0.85}\right) \]  

(32)

The estimate of \( MMT \) is 3.96 hours, standard deviation is 4.06 and the median is 4.50 hours. One an average, it takes 3.96 hours on an average to perform a maintenance on Kraken but there is a huge variation in the repair times.

Table 15: Distributions of all the maintenance repair times for Kraken

<table>
<thead>
<tr>
<th>Distribution</th>
<th>(-2Loglikelihood)</th>
<th>AICc</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lognormal</td>
<td>347.17</td>
<td>351.33</td>
<td>355.83</td>
</tr>
<tr>
<td>Log-logistic</td>
<td>352.37</td>
<td>356.53</td>
<td>361.03</td>
</tr>
<tr>
<td>Exponential</td>
<td>364.28</td>
<td>366.33</td>
<td>368.61</td>
</tr>
<tr>
<td>Weibull</td>
<td>363.83</td>
<td>368.00</td>
<td>372.50</td>
</tr>
<tr>
<td>Logistic</td>
<td>431.11</td>
<td>435.27</td>
<td>439.77</td>
</tr>
<tr>
<td>Normal</td>
<td>483.12</td>
<td>487.28</td>
<td>491.78</td>
</tr>
</tbody>
</table>
4.7 Summary

It is observed that lognormal distribution is the best fit for the emergency maintenance times, for the four supercomputers. The preventive maintenance times of each of Athena, Jaguar and Kraken are fitted the best by lognormal distribution, whereas that of Jaguar PF is fitted by a Weibull distribution. Considering both the emergency and preventive maintenance actions together, lognormal distribution is the best fit for the maintenance times of Athena, and Kraken and Weibull distribution is the best fit for the maintenance times of Jaguar, and Jaguar PF. Table 16 presents the maintenance time distributions and the estimated mean maintenance time for the four supercomputers.

<table>
<thead>
<tr>
<th>Supercomputer</th>
<th>Emergency only Distribution</th>
<th>( MEMT )</th>
<th>Preventive only Distribution</th>
<th>( MPMT )</th>
<th>Emergency &amp; Preventive Distribution</th>
<th>( MMT )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Athena</td>
<td>lognormal</td>
<td>3.15</td>
<td>lognormal</td>
<td>2.20</td>
<td>lognormal</td>
<td>3.80</td>
</tr>
<tr>
<td>Jaguar</td>
<td>lognormal</td>
<td>4.17</td>
<td>lognormal</td>
<td>5.45</td>
<td>Weibull</td>
<td>4.38</td>
</tr>
<tr>
<td>Jaguar PF</td>
<td>lognormal</td>
<td>3.62</td>
<td>Weibull</td>
<td>10.30</td>
<td>Weibull</td>
<td>5.31</td>
</tr>
<tr>
<td>Kraken</td>
<td>lognormal</td>
<td>2.67</td>
<td>lognormal</td>
<td>5.11</td>
<td>lognormal</td>
<td>3.96</td>
</tr>
</tbody>
</table>

The observations that can be inferred from the table and the basic statistics presented in Section 4.4 - 4.6 are:

- The mean emergency maintenance times of each of the four supercomputers are comparable to each other, whereas the mean preventive maintenance times are very dissimilar. This may be attributed to the difference in size of each supercomputer leading to the difference in the time taken to complete a preventive maintenance action.
- The variation in the preventive maintenance times for each supercomputer is lesser compared to the variation in the emergency maintenance times. This is as expected, the variation in a pre-scheduled activity will be less than the variation in an unplanned activity. The inventory requirements for the preventive maintenance would be standard and can be arranged for before the activity but in case of emergency maintenance, the inventory is comparatively unpredictable and may need time for procurement.
- The variation in emergency maintenance times of Athena is the highest but the least variation in preventive maintenance times, compared to Jaguar, Jaguar PF, and Kraken. Thus for Athena, giving an estimate of maintenance time once a failure has occurred is challenging and is the least predictable but its maintenance time for a preventive maintenance is the most predictable of all the four supercomputers.
- \( MPMT \) of Jaguar PF is almost twice as that of Athena or Jaguar or Kraken. This can be attributed to the fact that it is the biggest of the four supercomputers in terms of architecture and has the highest speed. Nevertheless, more understanding on the
actual activities performed during a preventive maintenance of Jaguar PF will help root-cause the reasons and come up with a plan to reduce the preventive maintenance time.

- $M\text{PMT}$ of Athena is the least among the four supercomputers and is equal to 2.20 hours, very less compared to that of the other three supercomputers. This may be the reason for the deteriorating effect on Athena after each preventive maintenance action.
5  AVAILABILITY AND MAINTENANCE POLICY

This chapter discusses the achieved and predicted performance measures of the four supercomputers. Section 5.1 presents the availability functions and compares the availability of each supercomputer. Section 5.2 discusses the effectiveness of existing maintenance policy and proposes new maintenance policy. Section 5.3 summarises the work done and the observations to be noted in this chapter.

5.1  Availability

Availability is the probability that a system or component is performing its required function at a specified point of time or over a specified period of time when operated under stated conditions. The terms ‘required function’ and ‘stated conditions’ are predefined and agreed upon by the stakeholders. It is the fraction of a time period that an item is in a condition to perform its intended function upon demand (SEMI, 1986, 2004).

5.1.1  Achieved availability

Achieved availability is the average availability of the supercomputer over the period of time during which the maintenance records are available. It takes both emergency maintenance and preventive maintenance into account. It is the fraction of time that the supercomputer is available, of the total time the supercomputer was observed. The achieved availability of the supercomputer is given by (Availability, 2010),

\[ A_a = \frac{T_n}{T_n + q(MMT)} \]  \hspace{1cm} (33)

where \( T_n \) is the total operating time of the supercomputer and \( q(MMT) \) is the total maintenance time of the supercomputer, during the period for which the maintenance records are available. Figure 5 shows the schematic representation of the total operating time and the total maintenance time. \( MMT \) is used as the maintenance can be emergency or preventive and \( q \) is the total number of expected interrupts given by,

\[ q = \int_0^T \rho(t)dt = \sum_{k=1}^{Z(t)} S^{N(t_k)} r^{M(t_k)} \lambda (t_k^\beta - t_{k-1}^\beta) \]  \hspace{1cm} (34)

The values of GPIM parameters, \( r, s, \lambda, \) and \( \beta \) are taken from Table 2, and the value of \( MMT \) are taken from Table 16, for each of the four supercomputers.

Downtime per year, \( D \) is a frequently used availability metric. The four supercomputers run continuously throughout the year, 365 days that is 8760 hours. Thus the downtime per year in days will be (Vargas, 2000),

\[ D = 365(1 - A_a) \]  \hspace{1cm} (35)

Table 17 presents the achieved availability and downtime per year for each of the four supercomputers. These numbers quantify the availability of each supercomputer during the time period for which the maintenance records are available. It can be observed that Jaguar
PF has the least availability and is explained by the fact that it has 175 interrupts, much higher than the interrupts recorded on each of the other three supercomputers. Also Jaguar PF has the highest MMT contributed by MPMT of 10.3 hours which is almost twice compared to MPMT of the other three supercomputers. This means that Jaguar PF has a higher percentage of time spend on maintenance compared to that of Athena, Jaguar and Kraken. Jaguar has the highest availability followed by Athena and Kraken in the decreasing order.

<table>
<thead>
<tr>
<th>Supercomputer</th>
<th>$\overline{A}\hat{a}$ (days/year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Athena</td>
<td>0.9701</td>
</tr>
<tr>
<td>Jaguar</td>
<td>0.9759</td>
</tr>
<tr>
<td>Jaguar PF</td>
<td>0.9150</td>
</tr>
<tr>
<td>Kraken</td>
<td>0.9374</td>
</tr>
</tbody>
</table>

5.1.2 Predicted point availability

Point availability is the availability of the supercomputer to perform a job at a specified time point, $t$. The point availability is an equivalent term for a repairable system as is reliability to a component. Reliability of a supercomputer is the probability that the supercomputer with its hardware and software components is capable of performing its intended function under normal operating conditions over the specified period of time (Jane, 1996). Thus, the predicted point availability of the supercomputer at a particular time, $t$ is the probability that it will function satisfactorily without interruption up to and until time, $t$ after the system is restored from the last recorded interrupt.

Let $f(t)$ be the probability density function of the interrupt distribution, $F(t)$ be its cumulative density function and $R(t)$ be the reliability function. The relation between these three functions and the failure intensity function is

$$ f(t) = \frac{dF(t)}{dt} = -\frac{dR(t)}{dt} = R(t)\rho(t) $$

(36)

Let $m(t_1, t_2)$ be the expected number of interrupts in the interval $(t_1, t_2)$ and is related to the failure intensity function by

$$ m(t_1, t_2) = \int_{t_1}^{t_2} \rho(t) dt $$

(37)

and is related to reliability function by the below equation.

$$ R(t) = \exp(-m(t_1, t_2)) $$

(38)
The point availability of the supercomputer at time, $t$ after the last recorded maintenance is

$$A_p(t) = \exp(-m_{GPIM}(0,t))$$  \hspace{1cm} (39)

where, $m_{GPIM}(0,t)$ is the number of expected number of interrupts in the interval $(0,t)$ with the failure intensity function modelled by GPIM. The failure intensity function changes after every maintenance action when it is modelled by GPIM. Thus, to understand the point availability of the supercomputer after the last recorded interrupt, it is important to calculate its failure intensity function at the end of its last recorded maintenance. The number of maintenance records shown in Table 1 and reliability and maintainability parameters shown in Table 2 are used to compute the failure intensity function of each supercomputer at the end of the last maintenance action. It has to be noted that the maintenance records are event terminated and thus the system operating time at the point of last interrupt is different for each supercomputer. Table 18 presents the failure intensity function at the end of the last recorded maintenance for each of the four supercomputers.

<table>
<thead>
<tr>
<th>Supercomputer</th>
<th>$\rho(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Athena</td>
<td>$0.1260t^{0.0415}$</td>
</tr>
<tr>
<td>Jaguar</td>
<td>$0.0051t^{-0.2500}$</td>
</tr>
<tr>
<td>Jaguar PF</td>
<td>$0.0153t^{0.2063}$</td>
</tr>
<tr>
<td>Kraken</td>
<td>$0.0073t^{0.0814}$</td>
</tr>
</tbody>
</table>

Figure 13 shows the variation of point availability of each of the four supercomputers with the increase in time (from the point of last recorded maintenance). It can be clearly seen that Jaguar is the most reliable system and Kraken, Jaguar PF, and Athena follow in decreasing order. The probability that Athena is available to perform a job after 50 hours of the last recorded maintenance is almost zero. This is because of its high failure intensity contributed by the scaling factor of 1.6231 after each preventive maintenance. There is 50% probability that Athena, Jaguar, Jaguar PF and Kraken are available to take up a job after 5 hours, 470 hours, 28 hours, and 73 hours respectively, after the last recorded maintenance.

5.1.3 Predicted average availability

This section attempts to predict the average availability of the supercomputer for a required operating time period after the last recorded maintenance observed at an operating time of $T_n$. In order to predict the average availability for a period of time beyond the last recorded maintenance, it is required to estimate the time points at which emergency interrupts and preventive interrupts occur. For preliminary analysis it is assumed that
emergency (preventive) interrupts occur at regular intervals with time period equal to the mean time to emergency (preventive) interrupts, $MTTEI (MTTPI)$ which are estimated from the available maintenance records.

Let $n_e$ be the number of emergency interrupts and $n_p$ be the number of preventive interrupts during the supercomputer operating time interval $(T_n, t + T_n)$. The predicted average availability during this interval, assuming that the supercomputer are maintained as per the existing maintenance policy is

$$A_{p_{\text{ps}}} = \frac{t}{t + MMT\sum_{k=n}^{n+n_e+n_p} s^{N(t_k)} r^{M(t_k)} \lambda(t_{k+1} - t_k^\theta)}$$  \hspace{1cm} (40)

and

$$n_e = \left\lfloor \frac{t}{MTTEI} \right\rfloor, n_p = \left\lfloor \frac{t}{MTTPI} \right\rfloor $$  \hspace{1cm} (41)

The GPIM results presented in Section 3.3 indicate that the preventive maintenance has positive impact only on Jaguar, thus it is of interest to understand the predicted average availability of the supercomputer assuming that the preventive maintenance is discontinued after the last recorded maintenance. In this case, the interruption is only caused due to emergency maintenance and the predicted average availability is

$$A_{p_{\text{ps}\text{emergency}}} = \frac{t}{t + MEMT\sum_{k=n}^{n+n_e+n_p} s^{N(t_k)} r^{M(t_k)} \lambda(t_{k+1} - t_k^\theta)}$$  \hspace{1cm} (42)
The values of GPIM parameters, $r$, $s$, $\lambda$, and $\beta$ are taken from Table 2, and the values of $MMT$ and $MEMT$ are taken from Table 16, for each of the four supercomputers. Table 19 presents the $MTTEI$ and $MTTP1$ for each of the four supercomputers. It also presents the number of emergency and preventive interrupts along with the predicted average availability for an operating period of 2160 hours, for each of the four supercomputers.

<table>
<thead>
<tr>
<th>Supercomputer</th>
<th>$MTTEI$</th>
<th>$MTTP1$</th>
<th>$n_p$</th>
<th>$n_p$</th>
<th>$\hat{A}_{pa;2160}$</th>
<th>$\hat{A}_{pa;emergency;2160}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Athena</td>
<td>263.08</td>
<td>1315.38</td>
<td>8</td>
<td>1</td>
<td>0.7212</td>
<td>0.7573</td>
</tr>
<tr>
<td>Jaguar</td>
<td>183.86</td>
<td>355.46</td>
<td>11</td>
<td>6</td>
<td>0.9783</td>
<td>0.9793</td>
</tr>
<tr>
<td>Jaguar PF</td>
<td>77.57</td>
<td>359.25</td>
<td>27</td>
<td>6</td>
<td>0.2406</td>
<td>0.3173</td>
</tr>
<tr>
<td>Kraken</td>
<td>204.31</td>
<td>470.78</td>
<td>10</td>
<td>4</td>
<td>0.7160</td>
<td>0.7890</td>
</tr>
</tbody>
</table>

It can be observed that the predicted average availability of Jaguar PF is the least and Jaguar is the highest of all the four supercomputers. The predicted average availability considering that the preventive maintenance is stopped after the last recorded maintenance is higher than the predicted average availability considering that the existing maintenance policy is followed, for all the four supercomputers.

5.2 Maintainability

Chapter 4 focuses on the analysis of maintenance times of the four supercomputers and provides insights into the maintenance in terms of the number of man hours spend on the maintenance activity. This section reviews the maintenance policy of each supercomputer based on the reliability and maintenance parameters presented in Table 2 and the different measures of availability presented in Section 5.1.

Emergency maintenance is performed to restore the supercomputer to normal operating condition after an occurrence of a failure and thus it cannot be avoided. Preventive maintenance is performed to effectively identify incipient failures and reduce the probability of them being actual failures. It is observed that for the existing preventive maintenance protocol on the four supercomputers is that it is performed at varying intervals, ranging from 15 to 54 days. Preventive maintenance is intended to reduce the failure intensity of the supercomputer but contributes to its downtime. Thus an effective preventive maintenance policy should be scheduled based on:

- the improvement it shows on performance of the supercomputer
- the increase in downtime of the supercomputer
- the consequences of a failure on the supercomputer
Section 5.2.1 - 5.2.4 comments on the existing maintenance policy and suggests a new preventive maintenance policy to be considered moving further for Athena, Jaguar, Jaguar PF and Kraken, respectively.

5.2.1 Athena

Athena is a stable supercomputer with constant failure intensity. With the existing maintenance policy, the failure intensity of Athena is marginally decreased after every emergency maintenance action and prominently increased after every preventive maintenance action. Thus, it is proposed to

- Stop performing preventive maintenance actions as they are deteriorating the system’s performance. This will increase the system availability;
- Root-cause the reasons for the preventive maintenance protocol not being effective; and
- Continue to monitor the system and watch out for the system phase transition to ‘wear out’ where the system starts deteriorating.

5.2.2 Jaguar

Jaguar is a supercomputer with improving performance. Its failure intensity is not effected after each emergency maintenance action but is prominently decreased after each preventive maintenance action. The improvement in the predicted average availability of the system considering that there is no preventive maintenance is very low as compared to the predicted average availability considering the existing maintenance policy with both emergency and preventive maintenance. The increase is only 0.1% considering the predicted average availability for an operating period of 3 months. Thus, it is proposed that the existing maintenance policy be continued on Jaguar.

5.2.3 Jaguar PF

Jaguar PF is a supercomputer with deteriorating performance. Its failure intensity is not effected by each maintenance action, be it emergency or preventive maintenance. The improvement in the predicted average availability of the system considering that there is no preventive maintenance is prominent as compared to the predicted average availability considering the existing maintenance policy with both emergency and preventive maintenance. The increase is about 32% considering the predicted average availability for an operating period of 3 months. This shows that though the preventive maintenance is not altering the failure intensity of Jaguar PF, it is contributing to the system downtime. Thus, it is proposes to

- Stop performing preventive maintenance actions;
- Root-cause the reasons for the preventive maintenance protocol not being effective and come up with an improved protocol. As Jaguar PF is inherently deteriorating, an effective preventive maintenance protocol can help to improve its performance; and
• Continue to monitor the system. There may be a point where the rate of deterioration is so high that it is required to revamp the system.

5.2.4  Kraken

Kraken is a stable supercomputer with constant failure intensity. Its failure intensity is not effected after each emergency maintenance action and is marginally decreased after every preventive maintenance action. The improvement in the predicted average availability of the system considering that there is no preventive maintenance is prominent as compared to the predicted average availability considering the existing maintenance policy with both emergency and preventive maintenance. The increase is about 10% considering the predicted average availability for an operating period of 3 months. This shows that the decrease in failure intensity of Kraken after each preventive maintenance does not justify its increase in system downtime. Thus, it is proposes to

• Stop performing preventive maintenance actions as the existing protocol does not help in improving the system performance;
• Continue to monitor the system and watch out for the system phase transition to ‘wear out’ where the system starts deteriorating.

5.3  Summary

The achieved availability, predicted point availability and predict average availability for an operating period of 3 month after the last recorded maintenance are estimated and presented for each of the four supercomputers. The maintenance policy of each supercomputer is reviewed and it is proposed that the preventive maintenance should not be performed on Athena, Jaguar PF and Kraken as it does not improve their performance. It is concluded that the existing maintenance policy of Jaguar is good and should be continued.
6 CONCLUSION

First, Section 6.1 summarizes the work done in this thesis along with a brief overview of the analysis approach. Section 6.2 presents the proposed maintenance policy for each of the four supercomputers. Finally, Section 6.3 discusses the scope for future study.

6.1 Summary of Thesis

The demand to analyze the maintenance records of four supercomputers, Athena, Jaguar, Jaguar PF and Kraken, collected over the period from October 2009 to December 2010 forms the motivation to this thesis. These supercomputers are located and operated at The National Institute for Computational Sciences in Oak Ridge National Lab, Tennessee, US.

Each supercomputer is considered as a complex repairable system with the occurrence of events causing interruption to normal system operation being modelled by stochastic point process. In order to understand the system performance characteristics such as reliability, availability, and maintainability it is important to fit a mathematical/statistical model to the available maintenance record data. Different maintenance models to analyze the repairable systems available in literature are broadly classified into maximal maintenance model, minimal maintenance model, and partial maintenance model. These models are reviewed and compared. Partial maintenance model that provides the scope to include the effectiveness of maintenance actions has been considered to be more appropriate to analyze the four supercomputers.

Different approaches to model the partial maintenance models proposed in literature have been reviewed. Among these, GPIM and GARM are the appropriate models to maintenance data that contain both emergency maintenance and preventive maintenance, as is the case with the four supercomputers and allows inclusion of both reliability and maintainability parameters in a single model. Finally, GPIM is chosen to be the main model for this thesis as it provides more practical physical model and higher potential for maintenance decision making than GARM and has the possibility to accommodate future extensions to this study, such as including predictor variables and covariates.

The reliability and maintenance parameters of GPIM that fits the maintenance records are obtained based on MLE method, using MATLAB, for each supercomputer. Based on the reliability parameter estimates it has been observed that Athena and Kraken are stable, while Jaguar is improving and Jaguar PF is deteriorating. The maintenance parameter estimates show that each emergency maintenance marginally decreases the failure intensity of Athena and marginally increases the failure intensity of Jaguar, Jaguar PF, and Kraken. Each preventive maintenance marginally decreases the failure intensity of Jaguar PF, and Kraken, prominently increases the failure intensity of Athena and prominently decreased the failure intensity of Jaguar.

While the occurrence of events interrupting supercomputer’s normal operation are modelled using GPIM, the maintenance times are modelled by life distributions. The best
life distributions fitting the maintenance times, considering only the emergency repairs, only the preventive maintenance, and both emergency and preventive maintenance times are selected using AICc and BIC model selection measures. The parameters of the chosen distributions are estimated using MLE methods. It has been observed that the variation in emergency maintenance times are higher that on the preventive maintenance, for all the four supercomputers. The $MPMT$ of Jaguar PF is distinctly higher than the $MPMT$ of the other three supercomputers. The $MPMT$ values of the four supercomputers are distinctly different and can be attributed to the difference in size of the supercomputers.

Achieved availability and predicted availability for each supercomputer are estimated using maintenance time distributions along with GPIM results. Jaguar PF has the least achieved availability and the least predicted average availability. Athena has the least point availability. Jaguar has the highest achieved availability, point availability and predicted average availability showing that it is the most reliable system of all the four supercomputers.

6.2 Proposed Maintenance Policy

The supercomputers operate continuously and are only interrupted by emergency maintenance or preventive maintenance. The maintenance team plan is to perform preventive maintenance on each of the four supercomputers, once every two weeks. But, it is observed from the maintenance records that the preventive maintenance has not been done at regular intervals on any of the four supercomputers. It may be due to some practical difficulties and may depend on the existing work load on the supercomputers, etc. So the existing maintenance policy is the one modelled by GPIM based on the maintenance records. Based on the study of the existing maintenance records and observation of the system performance and maintenance effectiveness, a new maintenance policy that improves the availability of the supercomputers and reduces the maintenance costs is proposed.

The proposed maintenance policy for each supercomputer is given below.

- Continue the existing emergency maintenance protocol and stop performing preventive maintenance for Athena, Jaguar PF, and Kraken,
- Continue the existing emergency maintenance and preventive maintenance protocols for Jaguar.
- Jaguar PF is a deteriorating system and the predicted average availability for an operating period of 3 months considered after the last recorded maintenance is less than 32%, which is very low for a supercomputer. Thus, Jaguar PF may need a revamp.
- Continue to monitor Athena and Kraken to detect any phase transition from stationary to wear out, where they may show an increasing failure rate.
6.3 Future Scope

The analysis of maintenance records studied in this thesis provides good insights into the system performance and failure characteristics of the four supercomputers. Future extension of this thesis can be in the direction:

- It has to be noted that the proposed maintenance policy is based on the maintenance records available during the period from October 2009 to December 2010. The system performance characteristics might have changed since then and it is important to use the GPIM to fit the latest maintenance records and form the base for a new maintenance policy proposal.

- A computerized maintenance management system can be used to dynamically model the interrupts and monitor the supercomputer based on the GPIM and the corresponding availability analysis.

- Linking the statistical model governing the maintenance records to the inventory is one other practical use. It is required to understand the failures and their causes at a cluster or module level to explore solutions of improving supercomputer availability and suggest better ways to handle the spare parts.

- It is observed there is a drastic difference in the way a preventive maintenance action affects each supercomputer. Also, there is a lot of variation observed in the preventive maintenance times of each supercomputer. This can be due to the fact that the preventive maintenance protocol is different for each supercomputer or the same preventive maintenance protocol for all the supercomputers is not an effective strategy. Root-cause analysis performed on this should provide direction for a more effective preventive maintenance.

- Supercomputers take up huge space and have stringent cooling requirements; in this context further work can be done to understand the spatial locations of the failures to figure out any location correlations. Workload and failure correlations, failure clustering, time span of works run and failure correlations, failure statistics with respect to categories like human error, environmental effects, etc. are some important areas to consider for better understanding of failures.
REFERENCE
50


Jagadish Cherukuri was born in India and his entire schooling is done in Loyola Public School, India during the period 1990-2000. He completed his intermediate education in Yadava Junior College, India and then took up undergraduate studies in R.V.R & J.C College of Engineering under Acharya Nagarjuna University in Mechanical Engineering. He graduated in 2007 with a distinction.

Jagadish took up a role as product engineer and worked for about 5 years in an engineering service providing company, Tech Mahindra. He worked for multiple clients to provide mechanical engineering services and his major area of work is product engineering for an appliance manufacturer. He later decided to complement his work experience with an academic degree and moved to USA in 2013 to pursue Master of Science in Reliability and Maintainability engineering in The University of Tennessee, Knoxville. He choose to graduate from the Industrial and System engineering department of the university and accepted graduate teaching assistantship for the academic year 2013-14.