8

RAS Modeling of a Large InfiniBand Switch System

Dong Tang and Ola Torudbakken
Sun Microsystems, Inc.
USA

1. Introduction

Computer clusters or grids constructed from open and standard commercial off the shelf (COTS) systems now dominate the top 500 supercomputer sites (Top500, 2008), providing an attractive way to rapidly construct high performance computing (HPC) systems of interconnected nodes. The largest of these HPC systems are now driving toward petascale deployments, delivering petaflops of computational capacity and petabytes of storage capacity. However, designing and building these large HPC systems involves significant challenges, including:

- Rapidly building and expanding the computational capacity of HPC clusters to meet growing demands
- Increasing levels of computational density while staying within constrained envelopes of power and cooling
- Reducing complexity and cost for physical infrastructure and management
- Implementing interconnect technology that can connect hundreds or thousands of processors without introducing unacceptable levels of latency

Interconnect technology plays a vital role in addressing all of these issues. InfiniBand has emerged as a compelling interconnect technology, and now provides more scalability and significantly better cost-performance than any other known fabric. In spite of its ability to provide high-speed connectivity and low latency, connecting and cabling thousands of compute nodes with smaller discrete InfiniBand switches remains problematic. With traditional approaches, the largest HPC clusters can require hundreds of switches, as well as thousands of ports and cables for inter-switch connectivity alone. The result can be significant added cost and complexity, not to mention energy and space consumption.

To address these challenges, the Sun Datacenter Switch 3456 (DS3456) system (Sun Microsystems, 2007) provides the world's largest standards-based DDR (dual data rate) InfiniBand switch, with direct capacity to host up to 3,456 server nodes. Only slightly larger than two conventional datacenter racks, the system drastically reduces the cost, power, and footprint of deploying very large-scale standards-based high performance computing fabrics. DS3456 is tightly integrated with the Sun Blade 6048 modular rack system (Sun Microsystems, 2008) which supports InfiniBand leaf switch, facilitating deployment of HPC systems up to 13,824 Nodes. Together these technologies offer low latency, high compute density, reduced cabling and management complexity, and lower power consumption than with other solutions.

Given this new large switch system, an important issue that needs to be addressed is the quantification of the associated RAS features. In this study, we developed a hierarchical Markov availability model (Trivedi, 2001) for DS3456 to assess its reliability, availability, and serviceability (RAS), using RAScad (Tang et al., 2002), a Sun internal RAS modeling tool that supports hierarchical modeling and automatic model generation.

The rest of this chapter is organized as follows: Section 2 gives an overview of Sun DS3456; Section 3 defines RAS metrics; Section 4 describes the model and parameters; Section 5 presents results and analysis; and Section 6 concludes the study.

2. Overview of DS3456

InfiniBand is a technology developed to address low-latency, high-performance, and low overhead communications between servers and I/O devices. It defines an architecture of networking principles – switching and routing – to provide a scalable, high-performance server I/O fabric (Cisco Systems, 2006). InfiniBand is a loss-less interconnect providing ordered packet delivery across the fabric through the use of credit-based flow-control. To ensure data integrity, its end-to-end protocols include fault tolerant features such as link-level and end-to-end CRC, packet re-transmission, multi-path routing, and automatic path migration. Upper-layer protocols, built on top of these provisions, allow seamless fit into existing networking and storage protocols. In addition, QoS (Quality of Service) and congestion control mechanisms are natively included in InfiniBand. All of these provide an excellent, converged fabrics solution for running storage, networking and clustering traffic.

DS3456 is the world's largest InfiniBand switch system, with capacity for connection of up to 3,456 nodes. The basic switch element used in DS3456 is the InfiniScale III (IS3) 24-port InfiniBand switch chip (Mellanox Technologies, 2009). The DDR version of IS3 supports 16 Gbps per 4x port, delivering up to 768 Gbps of aggregate bandwidth. The chip architecture features an intelligent non-blocking packet switch design with an advanced scheduling engine that provides QoS with switching latencies of less than 140 nanoseconds. DS3456 has been deployed in several HPC systems, including Ranger, the world No. 6 HPC system with peak performance of 579.4TFlops (Top500, 2008), located at Texas Advanced Computing Center, University of Texas at Austin.

Figure 1 is the physical view of DS3456. The major high-level DS3456 components and related RAS features are listed as follows.

- Twenty-four horizontally-installed line cards with each providing 48 12x connectors delivering 144 DDR 4x InfiniBand ports. Each line card connects to pass-through connectors in a passive orthogonal midplane.
- Eighteen vertically-installed fabric cards directly connected to the line cards through the orthogonal midplane. Each fabric card also features eight modular high-

performance fans that provide front-to-back cooling for the chassis. The eight fans are N+1 redundant and hot swappable.

- Two fully-redundant chassis management controller cards (CMCs) monitoring all critical chassis functions including power, cooling, line cards, fabric cards, and fan modules. CMC is hot swappable.
- Sixteen power supply units (PSUs) divided into two banks of eight units, with each bank providing N+1 redundant PSUs to half the line cards and half the fabric cards. PSU is hot swappable.

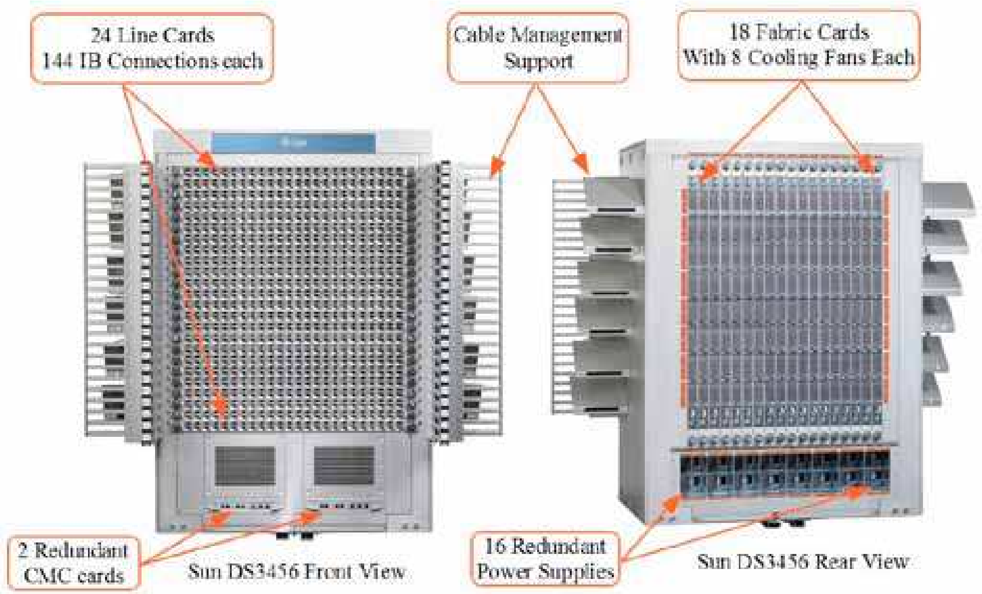

Fig. 1. DS3456 Physical View

Figure 2 shows the connectivity between line cards and fabric cards for DS3456. The passive midplane provides 432 8x8 orthogonal connectors arrayed in an 18x24 grid. Each line card contains 24 IS3 switch chips, 12 interfacing to the midplane, and 12 interfacing to the 12x connectors at the front of line card. A total of 144 4x InfiniBand ports are provided by each line card, expressed as 48 physical 12x connectors. Each fabric card contains eight IS3 switch chips connected to the midplane, providing interconnect between different line cards.

Thus, a communication path starts from an external port connected to an IS3 chip at the bottom row of a line card, goes through an IS3 chip at the top row of the same line card, an IS3 chip on a fabric card, two IS3 chips on the destination line card (one at the top row and one at the bottom row), and ends at another external port connected to the destination IS3 chip. That is, a message packet goes through as many as five stages of switch from the source port to the destination port.

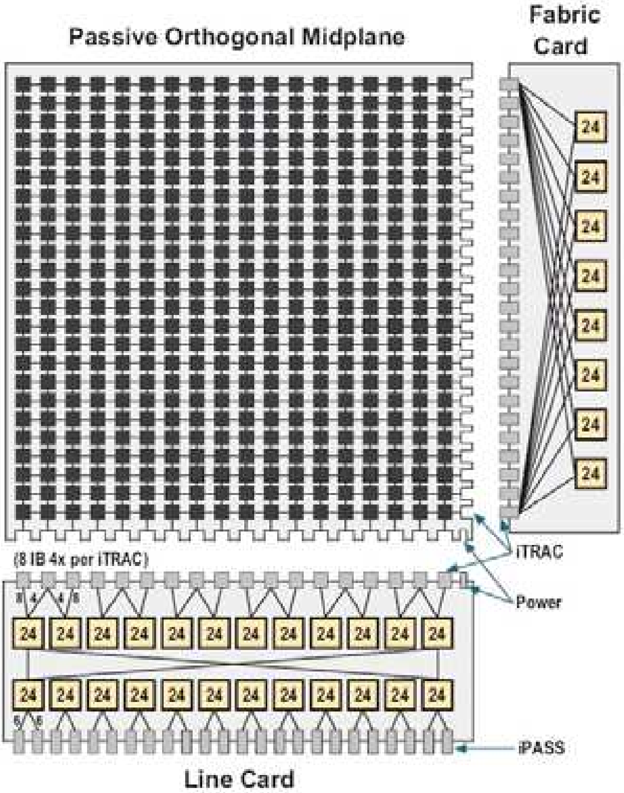

Fig. 2. DS3456 Internal Connectivity

3. RAS metrics defined

To quantify RAS for the target system, we first define RAS metrics and related concepts in this section. For simplicity, the capacity of the target system is assumed to be fully used, i.e., all 3,456 ports of the switch are utilized to connect server nodes.

3.1 Reliability

Connectivity between the server nodes using the switch for communication is a reliability measure for the switch system. A connectivity failure is defined as the loss of communication between a server node physically connected to the switch and another server node physically connected to the same switch due to hardware problems in the switch. We use *Mean Time Between Connectivity Failures* (MTBCF) to quantify reliability for the switch system.

A line card or fabric card failure would cause partial communication paths in the switch to be unavailable. Unavailability of partial paths caused by a fabric card failure does not affect connectivity as paths that were routed across the faulty fabric card can be re-routed to the operational fabric cards. Unavailability of partial paths caused by a line card failure may or may not translate to connectivity failures, depending on redundancy in the interconnect topology between the switch and server nodes.

- Non-redundancy case. If each server node connects to only one port on the switch, a line card failure would result in connectivity failures for some of the server nodes connected to the switch.
- Redundancy case. Typically, each server node connects to two or four ports on different line cards in the switch. In this case, unavailability of partial paths caused by the failure of one line card does not generate any connectivity failures.

3.2 Availability

The traditional availability definition is the proportion of time that the system is operational and delivering required services. At any time point, the system is in either an up or a down state. However, for a degradable system, the system can be in a partially available state. For the non-redundancy case of DS3456, unavailability of partial paths does not disable the function of the entire switch, but degrades system capacity. Thus, the system can be in partially available states, in addition to the fully available and failure states. For instance, when a line card fails, paths related to the faulty line card (out of 24 line cards) are unavailable and the system capacity is reduced by 1/24. Therefore, we defined availability for this state as 23/24. The RAScad performability (Trivedi, 2001) evaluation capability is used to generate this performance-oriented availability.

3.3 Service cost

In traditional service strategies, every component failure in the system translates to a service call. For such a large system as DS3456, replacing a line card or fabric card is particularly time consuming, because it may take several hours for the system to complete the restart process after a power-off repair. It is thus desirable to reduce service frequency as much as possible.

Previous studies showed that adoption of deferred repair service strategies for redundant components can greatly reduce unscheduled service events and associated system downtime (Sun, 2005). In this study, we once again analyzed the effect of deferred repair on system availability and service cost for the redundancy case. We use *Unscheduled Mean Time Between Services* (U_MTBS) to quantify service cost for the switch system.

3.4 Failure rate estimation

These metrics are calculated from a system-level RAS model built by utilizing information on the system configuration and its RAS characteristics (redundancy, hot or cold swap, etc.), applying a failure rate to each component, and then integrating them into the model. These failure rates are estimated from previous field data using the field-based *Mean Time Between Failures* (MTBF) prediction method described below where MTBF = 1/failure rtae.

Field-Based MTBF Prediction Method — The Field Replaceable Unit (FRU) MTBFs are calculated using methods described in Telcordia TR-NWT-000332 (Telcordia Technologies, 2001) with lower component-level (ICs, resistors, capacitors, etc.) failure rates adjusted based on field data, or directly estimated from field data, or provided by the OEM vendors.

The field data used to calibrate component failure rates were collected from tens of thousands of Sun field systems with billions of cumulative operating hours. This approach is called the Sun field-based MTBF prediction method.

4. RAS model and parameters

Similar to many studies of this type, we assumed independent failures on different components and constant failure rates. The target system is modeled as a hierarchy of Markov chains. The top level model is shown in Figure 3. In a RAScad Markov model, the user can define three reward vectors for each state, as displayed in the circles representing states (Tang & Trivedi, 2004): (1) Availability (0 or 1), (2) Performance (≥ 0), and (3) Service Cost (≥ 0).

The first reward vector is used to calculate system availability. The second reward vector is used to calculate system performability. The third reward vector is used to calculate annual service cost or service call rate. In the DS3456 model, up to two failures of line card and fabric card, which have impact on system performability (for the non-redundancy case), were modeled in detail. The notation used in the models is explained as follow:

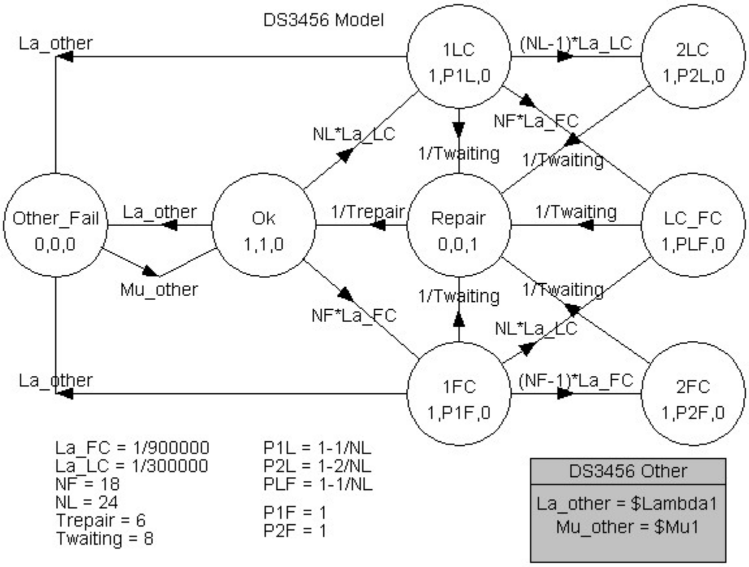

Fig. 3. Top level Markov model

- Ok: state in which the system is functioning properly (no faults)
- 1LC: state in which one line card has failed
- 2LC: state in which two line cards have failed
- 1FC: state in which one fabric card has failed
- 2FC: state in which two fabric cards have failed
- LC_FC: state in which one line card and one fabric card have failed
- Repair: state in which the system is shutdown to replace faulty line card or fabric card

- Other_Fail: state in which the system is down due to other hardware component failures
- NL: number of line cards in the system (18)
- NF: number of fabric cards in the system (24)
- Twaiting: service waiting time – waiting for off-peak hours to repair the system (8 hours)
- Trepair: repair time including restart time (6 hours)
- La_LC: failure rate for line card (1/900K hours)
- La_FC: failure rate for fabric card (1/300K hours)
- La_other: system failure rate due to other hardware faults (calculated from submodel)
- Mu_other: system repair rate for other hardware faults (calculated from submodel)

When one or more line/fabric cards have failed (states 1LC, 1FC, 2LC, 2FC, and LC_FC), the system is scheduled to be shutdown for repair after a service waiting time. For the non-redundancy case, these states may be degraded states, as shown by the performance reward vector in these states (P1L, P2L, etc.). For the redundancy case, these states are still fully functioning states, as shown by the availability reward vector in the states (all values are 1).

In Figure 3, the gray color rectangle box represents the interface between the current model and the submodel called DS3456 Other. All hardware components other than line cards and fabric cards are included in the submodel (details are not discussed in this chapter). If a system failure occurs due to hardware problems other than line card and fabric card faults, the system goes from the Ok state to the Other_Fail state. The associated failure rate (La_other) and repair rate (Mu_other) are bound to the submodel output Lambda1 and Mu1 which are the equivalent failure rate and repair rate (Lanus et al., 2003) of the submodel.

The model parameters, as listed above, were estimated using the Sun field-based MTBF prediction method discussed in Section 3.4 or based on engineering judgements. The repair time was estimated to be 6 hours because the system restart time is long.

5. Analysis of results

In this section, we present RAS results for the target system, including basic results, interval results (assuming deferred repair), and uncertainty analysis on key parameters.

5.1 Basic results

Table 1 shows the steady-state system level results evaluated from the DS3456 model by RAScad. The results show that for the redundancy case, MTBCF is much longer than that for the non-redundancy case. That is, with two or four redundant ports on different line cards, the system reliability is high in terms of connectivity. But this is not the case for system availability due to the large number of line/fabric cards and the long duration of power-off repair time of these cards. The system availability is similar for both redundancy and non-redundancy cases. This is because the system unavailability is dominated by power-off repair events, which are common for both cases. In other words, the system unavailability is not significantly affected by the degraded states for the non-redundancy case.

Configuration	U_MTBS (hours)	MTBCF (hours)	Availability
Non-redundancy	5,937	9,679	0.999372
Redundancy	5,937	3.23E6	0.999398

Table 1. Steady-state results for DS3456

A high availability DS3456 configuration typically implements interconnect between a server node and two (2-way redundancy) or four (4-way redundancy) different line cards, utilizing standard 4x InfiniBand ports. In the following, our discussion is focused on the 4-way redundancy configuration. To investigate which components in the system contribute most to the system unavailability (or downtime) and service events, we did a breakdown analysis as shown in Figures 4 and 5.

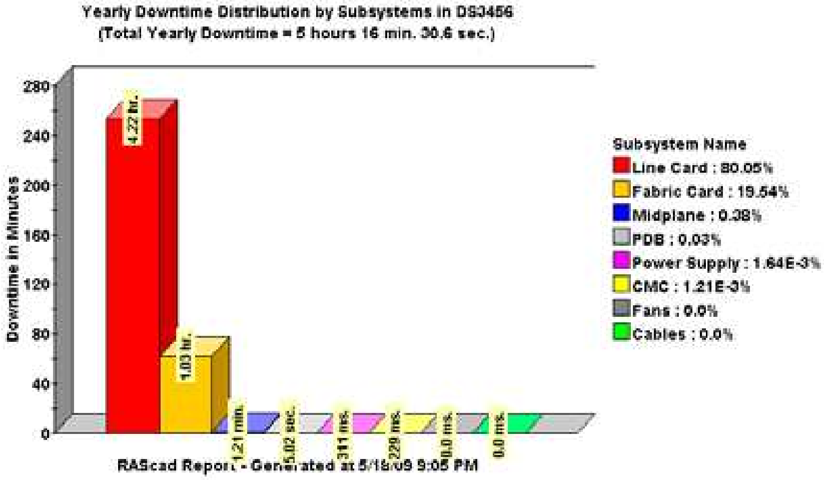

Fig. 4. Distribution of system downtime

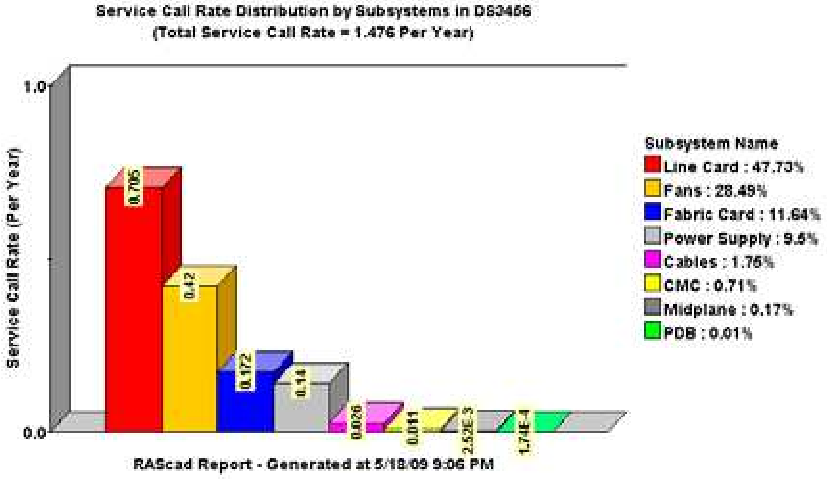

Fig. 5. Distribution of service events

Figure 4 shows that the system unavailability is dominated by shutdown repairs for faulty line cards and fabric cards. Figure 5 shows that the service events are mostly due to the following components: line cards, fans, fabric cards, and power supply units. Deferred repair of these components, if possible, could significantly reduce unscheduled service events and system downtimes. For the 4-way redundancy configuration, we can tolerate at least two line card or fabric card failures without losing any connectivity. Since every eight fans (N+1 redundant) are associated with a fabric card, we can also tolerate the failure of two fans associated with a line card (equivalent to a line card failure) or three fans otherwise.

5.2 Deferred repair

Given these thresholds of component failures that can be tolerated without degrading system performance, the following deferred repair service strategy is proposed for the target system. The system is serviced periodically, referenced as scheduled service, according to a predefined maintenance schedule, to repair all the components that have failed since the last service event. During the time window between two scheduled services, an unscheduled service is triggered upon any of the following events:

- Two line cards have failed.
- Two fabric cards have failed.
- One line card and one fabric card have failed.
- Two fans associated with a fabric card or any three fans have failed.
- Any other hardware component failures that stop the functioning of system (e.g., failure of two PSUs in a power bank).

The Markov model in Figure 3 can be easily modified to model this deferred repair service strategy by removing the transition from state 1LC to state Repair and the transition from state 1FC to state Repair. That is, no repair action is taken upon a failure of line card or fabric card. In addition, one of the submodels in the hierarchy, the fan model, also needs to be modified, as shown in Figure 6. In the diagram, La_fan is the fan failure rate and N is the total number of fans in the system. The failure of two fans associated with a fabric card is modeled by the transition from state 1Fan_Down to state Repair. The failure of any three fans is modeled by the transition from state 2Fan_Down to state Repair.

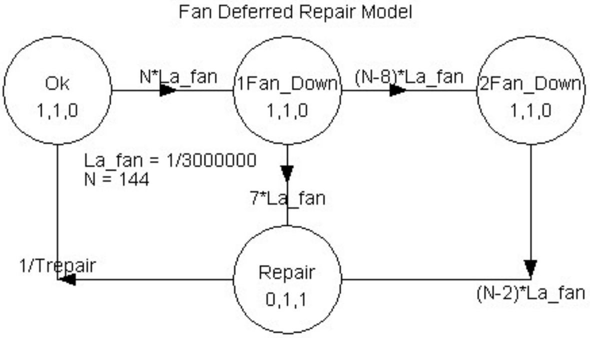

Fig. 6. Deferred repair model for fans

Table 2 shows the interval system-level results for different service strategies generated by RAScad. Our previous study (Tang & Trivedi, 2004) showed that the interval availability (average availability for a time interval from 0 to T) and associated measures, such as interval failure rate and interval service call rate, instead of steady-state measures, should be used for systems with deferred repair. In the table, "Deferred, 3 months" means a deferred repair service strategy with a periodic maintenance schedule of three months.

Service Strategy	U_MTBS (hours)	Unscheduled Yearly Downtime
No deferred repair	5,937 hr.	5 hr. 16.25 min.
Deferred, 1 month	41,992 hr.	11.44 min.
Deferred, 3 months	33,527 hr.	29.98 min.
Deferred, 6 months	26,941 hr.	52.04 min.

Table 2. Interval results for DS3456 under different service strategies

The table indicates that adoption of the proposed deferred repair service strategies can significantly reduce unscheduled service events as well as system downtime. With a quarterly maintenance schedule, the unscheduled MTBS is five times longer and the unscheduled system downtime is reduced by 90%. With a monthly maintenance schedule, the unscheduled MTBS is seven times longer and the unscheduled system downtime is reduced by 96%.

Although the system reliability and unscheduled downtime can be further improved by increasing maintenance frequency, the scheduled downtime (6 hours for each maintenance event) will also increase, leading to lower overall system availability. Given a tradeoff between system reliability and availability, we recommend deferred repair service strategies with a maintenance time window of 1 to 3 months.

5.3 Uncertainty analysis

Two key parameters in the model are the line card and fabric card failure rates. How sensitive are the results to the variance of these parameters? To answer this question, we performed an uncertainty analysis using RAScad. In each experiment of the analysis, the two parameters were randomly selected from the ±50% range of the estimated mean value, respectively, to generate a point of result. The sample size is 1,000. Figure 7 and Figure 8 plot the results for the "Deferred, 3 months" service strategy.

Figure 7 shows that for unscheduled MTBS, the 90% confidence interval is (26856, 40979), with the mean of 33,490 hours. That is, U_MTBS is likely to vary about ±20% around the mean when the uncertainty of the two key parameters is ±50%. Figure 8 shows that for unscheduled yearly downtime, the 90% confidence interval is (12.8, 52.7), with the mean of 30.5 minutes. That is, the system availability is most likely to stay at the 0.9999 level (equivalent to 5.3 to 53 minutes of yearly downtime), given the ±50% uncertainty of the two key parameters. Notice the slight difference between these means estimated from the simulations and those calculated numerically in Table 2. This is due to the nature of random sampling in simulations.

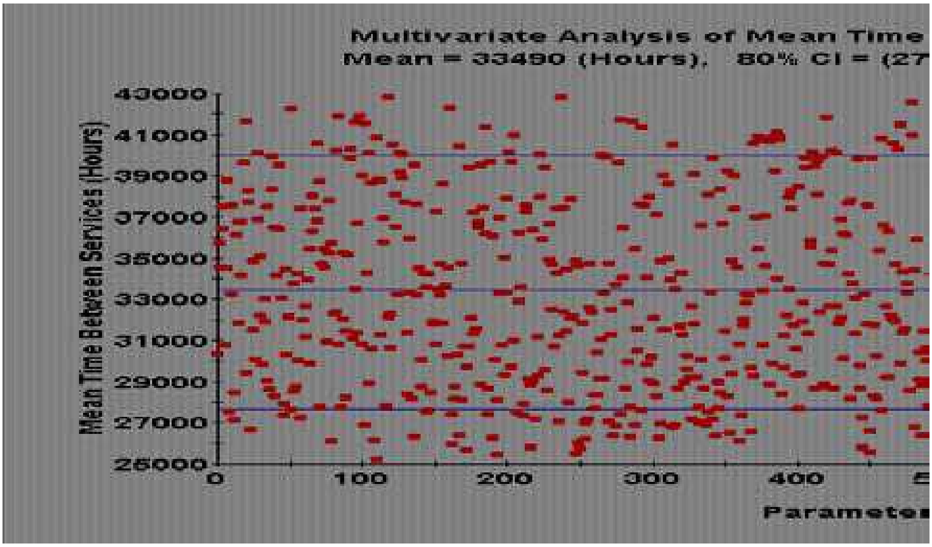

Fig. 7. Uncertainty analysis plot for U_MTBS

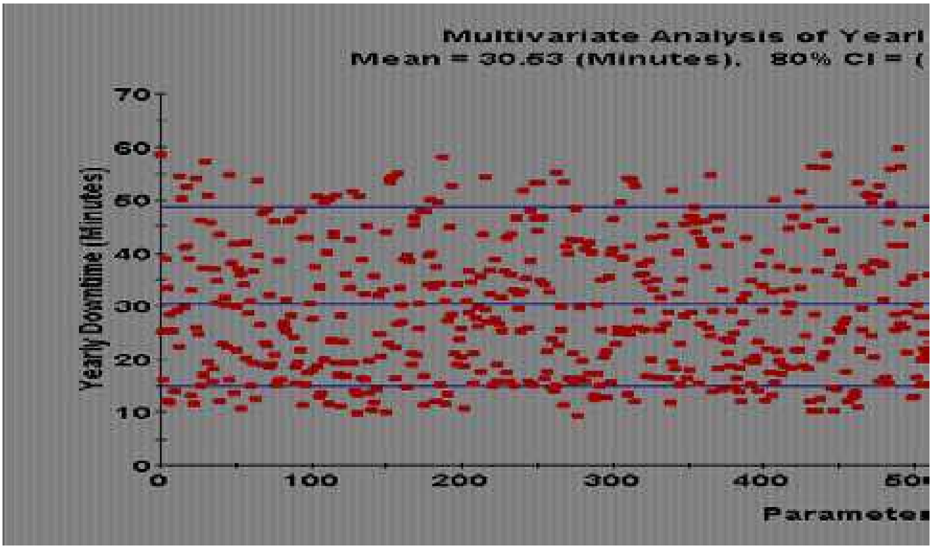

Fig. 8. Uncertainty analysis plot for yearly downtime

6. Conclusion

In this chapter, we presented a reliability, availability, and serviceability modeling and analysis, against hardware faults, for the Sun Datacenter Switch 3456 system, the world's largest InfiniBand switch system. To our knowledge, this is the first effort in RAS modeling of such a large switch system. The study demonstrated how the hierarchical Markov modeling approach can be used on large switch systems to reduce model complexity and therefore the feasibility of RAS quantification for large switch systems.

The results show that the system reliability, in terms of connectivity between the server nodes physically connected to the switch, is high for configurations with redundant ports (MTBCF > 3 million hours). The study also investigated the RAS benefits of practicing deferred repair service strategies and identified optimal maintenance time windows. Adoption of our recommended service strategies can significantly reduce unscheduled service events (five to seven times longer U_MTBS) and system downtime (by 90% to 96%). Finally, an uncertainty analysis was performed to study the sensitivity of results to the variance of key parameters. The analysis generated 90% confidence intervals of system RAS measures for the ±50% uncertainty of two key parameters.

7. References

Cisco Systems (2006). *Understanding Infiniband*, White Paper, http://www.cisco.com/

Lanus M.; Ying L. & Trivedi K. S. (2003). Hierarchical Composition and Aggregation of State-based Availability and Performability Models, *IEEE Transactions on Reliability*, Vol. 52, No. 1, March 2003, pp. 44-52

Mellanox Technologies (2009). *InfiniScale III Product Brief*, http://www.mellanox.com/

Sun H.; Tang D. & Wood R. (2005). Optimizing Service Strategy for Systems with Deferred Repair, *Proceedings of the 11th Pacific Rim International Symposium on Dependable Computing* (PRDC'05), ISBN 0-7695-2492-3, Changsha, China, Dec. 2005, IEEE, Los Alamitos, California

Sun Microsystems (2007). *Sun Datacenter 3456 Switch System Architecture*, White Paper, Nov. 2007.

Sun Microsystems (2008). *Pathways to Petascale Computing: The Sun Constellation System – Designed for Performance*, White Paper, Feb. 2008.

Tang D.; Zhu J. & Andrada R. (2002). Automatic Generation of Availability Models in RAScad, *Proceedings of International Conference on Dependable Systems and Networks*, (DSN 2002), pp. 488-492, ISBN 0-7695-1597-5, Washington DC, USA, June 2002, IEEE, Los Alamitos, California

Tang D. & Trivedi K. S. (2004). Hierarchical Computation of Interval Availability and Related Metrics, *Proceedings of International Conference on Dependable Systems and Networks* (DSN 2004), pp. 693-698, ISBN 0-7695-2052-9, Florence, Italy, June 2004, IEEE, Los Alamitos, California

Telcordia Technologies (2001). *SR332 - Reliability Prediction Procedure of Electronic Equipment*, Issue 1, May 2001.

Top500 Supercomputer Sites (2008). Top 10 Systems – 11/2008, http://www.top500.org

Trivedi K. S. (2001). *Probability and Statistics with Reliability, Queuing, and Computer Science Applications*, ISBN 0-471-33341-7, John Wiley and Sons, New York

Switched Systems
Edited by Janusz Kleban

ISBN 978-953-307-018-6
Hard cover, 174 pages
Publisher InTech
Published online 01, December, 2009
Published in print edition December, 2009

This book presents selected issues related to switched systems, including practical examples of such systems. This book is intended for people interested in switched systems, especially researchers and engineers. Graduate and undergraduate students in the area of switched systems can find this book useful to broaden their knowledge concerning control and switching systems.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821