Predicting the Number of Fatal Soft Errors in Los Alamos National Laboratory's ASC Q Supercomputer

Sarah E. Michalak, Kevin W. Harris, *Member, IEEE*, Nicolas W. Hengartner, Bruce E. Takala, and Stephen A. Wender

Invited Paper

***Abstract*—Early in the deployment of the Advanced Simulation and Computing (ASC) Q supercomputer, a higher-than-expected number of single-node failures was observed. The elevated rate of single-node failures was hypothesized to be caused primarily by fatal soft errors, i.e., board-level cache (B-cache) tag (BTAG) parity errors caused by cosmic-ray-induced neutrons that led to node crashes. A series of experiments was undertaken at the Los Alamos Neutron Science Center (LANSCE) to ascertain whether fatal soft errors were indeed the primary cause of the elevated rate of single-node failures. Observed failure data from Q are consistent with the results from some of these experiments. Mitigation strategies have been developed, and scientists successfully use Q for large computations in the presence of fatal soft errors and other single-node failures.**

***Index Terms*—Cosmic-ray-induced neutron, life estimation, linear accelerators, memory testing, neutron beam, neutron-induced soft error, neutron radiation effects, semiconductor-device radiation effects, semiconductor-device testing, single-event upset, soft-error rate, static random access memory (SRAM) chips.**

I. Introduction

THE Advanced Simulation and Computing (ASC) Program is a collaboration between the U.S. Department of Energy and Los Alamos, Livermore, and Sandia National Laboratories that supports the development of computing capabilities needed for stewardship of the U.S. nuclear weapons stockpile. As part of this effort, the ASC Q supercomputer was deployed at Los Alamos National Laboratory (LANL) in the 2002–2003 timeframe. In June 2003, it was the second-fastest supercomputer on the Top 500 list, with a Linpack benchmark score of 13.88 TeraOps [1]. Q is composed of 2048 Hewlett-Packard (HP) AlphaServer ES45 nodes [2]. Each node houses four Alpha 21264 1.25-GHz processors [3] that are situated on four separate CPU boards. Further information about Q's architecture is available in the Appendix.

Q is used for scientific computing, jobs that typically require many processors for many hours, and is shared by Los Alamos, Livermore, and Sandia National Laboratories. A large calculation is performed on Q by dividing it into pieces, each of which is calculated on a different processor. Performing a calculation in this manner requires interprocessor communication using message passing interface [4] or a similar parallel programming model so that data may be shared by different pieces of the calculation that are on different processors. The state of the calculation is backed up periodically, a process referred to as checkpointing. When a single node fails, the entire job must be restarted from its last known state, i.e., the output from its most recent checkpoint. Consequently, single-node failures can increase the runtimes of large calculations.

Prior to the deployment and integration of Q, estimates of its hardware reliability had been calculated based on component-level reliability information provided by the vendor. Q's initial single-node-failure rate was greater than these estimates predicted. Investigation of error logs from Q revealed that board-level cache (B-cache) tag (BTAG) parity errors were largely responsible for the elevated rate of single-node failures. A BTAG parity error occurs when the sum of bits in a cache line in a BTAG SRAM does not have the correct parity, indicating that an odd number of bits in the cache line have changed parity. An error that changes a bit in a computer's memory system is referred to as a soft error.

Error correction code (ECC) may be used to detect and correct soft errors, and most of Q's components were provided with ECC so that soft errors in these ECC-protected components typically do not cause failures. However, the use of ECC can increase application runtimes because it increases memory access time. While the ability to detect an odd number of soft errors per cache line (parity-error detection) was supplied for Q's BTAG SRAMs, ECC was not. Thus, a soft error in one of Q's BTAG SRAMs could be detected, but not corrected. When such an error is detected, information about the exact entry of the BTAG with the parity error is not available with the BTAG parity-error message, so the entire contents of the B-cache is untrustworthy. As a result, a BTAG parity error results in a

Manuscript received February 14, 2005; revised May 25, 2005. This work was supported in part by the U.S. Department of Energy.

S. E. Michalak and N. W. Hengartner are with the Statistical Sciences Group at Los Alamos National Laboratory, Los Alamos, NM 87545 USA (e-mail: michalak@lanl.gov; nickh@lanl.gov).

K. W. Harris is with the High Performance Computing Division, Hewlett-Packard Company, Nashua, NH 03062 USA (e-mail: k.harris@hp.com).

B. E. Takala and S. A. Wender are with the Neutron and Nuclear Science Group at Los Alamos National Laboratory, Los Alamos, NM 87545 USA (e-mail: takala@lanl.gov; wender@lanl.gov).

Digital Object Identifier 10.1109/TDMR.2005.855685

node crash. While other types of soft errors that are fatal, i.e., that cause node crashes, are possible, the study described in this paper focuses on BTAG parity errors caused by cosmic-ray-induced neutrons because they are believed to be the most frequent type of fatal soft error that occurs in Q. Throughout, we use the term fatal soft error to refer to such errors.

With this understanding, investigation of Q's higher than expected rate of single-node failures focused on determining the mechanism(s) causing bits in Q's BTAG SRAMs to change parity and lead to BTAG parity errors. The vendor for Q's BTAG SRAMs provided data on their efforts to reduce radioactive contamination and ^{10}B in these devices. The BTAG SRAMs in Q were from the most recent processes that eliminated a large fraction of ^{10}B. These data, in conjunction with that from other vendors, indicated that the SRAMs in Q were at or near the limit of the available technology for reducing susceptibility to soft errors. In addition, Q was experiencing an elevated rate of fatal soft errors relative to other systems that used the same BTAG SRAMs, further suggesting that a mechanism other than one related to the manufacturing process was responsible for Q's rate of fatal soft errors. Investigation of Q's error logs [5] suggested that the BTAG parity errors might be the result of cosmic-ray-induced neutrons. This hypothesis was tested via device testing at LANL's Los Alamos Neutron Science Center (LANSCE).

Compared to other systems, the impact of fatal soft errors on Q was magnified because of the manner in which it is used and the elevation at which it is located. Specifically, while each BTAG SRAM may experience fatal soft errors at a very low rate, each of Q's nodes contains four processors with one BTAG SRAM included in the B-cache on the board that houses each processor, for a total of four BTAG SRAMs per node. Consequently, a job that requires 500 nodes uses 2000 BTAG SRAMs, thus yielding an overall fatal soft-error rate that may affect application runtimes for lengthy jobs (calculations requiring 500 nodes are not uncommon). In addition, Q is housed at a high altitude (approximately 7500 ft), where the cosmic-ray-induced neutrons that can lead to soft errors are roughly 6.4 times more prevalent compared to at sea level [6], [7]. So, if a single BTAG SRAM experiences fatal soft errors at a rate of roughly 1 every 50 years at sea level, then a job using 500 nodes (2000 BTAG SRAMs) at 7500 ft will experience fatal soft errors at a rate of roughly 1 every 34 h.

Because of the potential impact of single-node failures on application runtimes, there was a strong interest in determining whether cosmic-ray-induced neutrons were the primary cause of Q's BTAG parity errors or if there were additional causes. To address this question, a series of experiments was undertaken using the neutron beam at LANSCE's Irradiation of Chips and Electronics (ICE) House facility. The neutron beam at LANSCE is unique because its neutron spectrum is very similar to that at terrestrial and aircraft altitudes, except that it is many times more intense [8]. It has been used for accelerated testing of semiconductor and other devices that are susceptible to cosmic-ray-induced neutrons and related research [9]–[18]. Some studies compare the results from testing undertaken in the field or other measurements from the field with those from accelerated testing performed at LANSCE and/or other facilities (see, for example, [10], [17], [19], and [20]). An abbreviated report of this work has been presented elsewhere [21].

II. Experimental Protocol

The experimental goal was to predict the average weekly number of fatal soft errors in Q. The testing needed to be completed in a short timeframe, so existing technology was leveraged as much as possible in the development of the experimental protocol.

The BTAG SRAMs in Q are physically the same as the DATA SRAMs in Q's B-caches, so both components should have the same susceptibility to soft errors and both were tested to estimate Q's fatal soft-error rate. Fatal soft errors typically do not occur in Q's DATA SRAMs since ECC was provided for them. Although the soft errors in Q's DATA SRAMs are typically not fatal, the rate at which they occur may be used to estimate the rate at which fatal soft errors occur in Q's BTAG SRAMs since the two types of SRAMs are physically identical and should have the same susceptibility to soft errors deriving from cosmic-ray-induced neutrons. In doing so, one must adjust the rate at which soft errors occur in Q's DATA SRAMs since all of the bits in Q's DATA SRAMs may be used while only a quarter of the bits in Q's BTAG SRAMs are used.

This paper presents experimental results from two test programs, memtest and btagexer. Memtest is a memory diagnostic that exercises a node's main memory, and consequently its B-cache memory, while the node is in a low-level console mode. This program was installed in each of Q's nodes at delivery. Because the node is in a low-level console mode when memtest is running, soft errors in its DATA SRAMs and BTAG SRAMs are logged, but typically do not cause the node to crash. Thus, when using the memtest program, counts of soft errors in the DATA SRAMs that would typically not be fatal under normal operating conditions and counts of soft errors in the BTAG SRAMs that typically would be fatal under normal operating conditions are possible depending on the components at which the neutron beam is aimed. The second test program, btagexer, was written for our testing. It exercises a node's B-cache memory while it is in its normal operating mode. Thus, if LANSCE's neutron beam is aimed at the DATA SRAMs and the btagexer program is running, counts of soft errors that typically would be nonfatal under normal operating conditions result since ECC has been provided for Q's DATA SRAMs. If the beam is aimed at one or more BTAG SRAMs while the btagexer program is running, the node will fail when the first BTAG parity error is encountered and a "time," measured in neutrons per square centimeter, until failure results. Thus, the choice of test program (memtest or btagexer) and beam aim (BTAG SRAM and/or DATA SRAMs) determines the type(s) of data that result: the number of neutrons per square centimeter until failure, a count of soft errors that would typically be fatal, and/or a count of soft errors that would typically be nonfatal.

A CPU board in one of Q's nodes houses eight DATA SRAMs and one BTAG SRAM. The number of SRAMs being tested during an experiment at LANSCE depends on the number of CPU boards in the node exposed to the neutron beam and the components at which the beam is aimed. Q's nodes may

house up to four CPU boards, each with the same configuration and orientation relative to the beam. The beam may be aimed at the location of the BTAG SRAM, a pair of DATA SRAMs (pairs of DATA SRAMs "sandwich" Q's CPU boards), the BTAG SRAM and a pair of DATA SRAMs, or two pairs of DATA SRAMs. So, up to four BTAG SRAMs or 16 DATA SRAMs may be tested during a single experiment. Depending on the number and type of SRAMs under test, the number of BTAG SRAM equivalents under test may be determined. In particular, since only a quarter of the bits in a BTAG SRAM are used while all of the bits in a DATA SRAM may be used, one DATA SRAM under test is equivalent to four BTAG SRAMs under test. Thus, if there is one CPU board in the node under test and the neutron beam is aimed at one pair of DATA SRAMs, this is equivalent to testing eight BTAG SRAMs.

HP staff determined the test procedures and performed the testing at the ICE House facility at LANSCE. The general test procedure was as follows. The node under test was determined to contain the appropriate number of components for testing and then aligned in the neutron beam using a laser alignment system. Next, the appropriate test program was launched. Then, the beam shutter was opened, allowing the node to be exposed to the beam of neutrons. During this time, errors were logged in files that were recovered for analysis following the experiment. At the end of the experiment, the beam shutter was closed. See [5, Fig. 1] for a diagram of the experimental setup.

A fission ionization chamber [22] was used to determine the number and energies of neutrons per square centimeter to which the node was exposed during a given experiment. The fission ionization chamber samples neutrons from the beam just before the beam enters the experimental facility. A liquid crystal display (LCD) counter records the number of "fission pulses" from the beam to which the device under test is exposed. This count of fission pulses is proportional to the number of neutrons per square centimeter from the beam to which the device under test has been exposed. LANSCE personnel calculate the average number of neutrons per square centimeter that occur per fission pulse over time periods that may include several experiments and provide this figure to the experimental personnel. The number on the LCD counter is noted at the end of an experiment, and the product of it and the average number of neutrons per square centimeter per fission pulse provides an estimate of the number of neutrons per square centimeter to which the device under test was exposed during the experiment.

For experiments intended to yield count data, HP personnel could control the length of the experiment in wall-clock time by shutting the beam shutter at a chosen time. However, for the failure-time experiments, HP personnel did not control the length of the experiment. Instead, they monitored the node to detect when it had failed. In this latter situation especially, there might be a time lag between the end of the experiment and the time at which the number of fission pulses from the LCD monitor was recorded. In particular, there might be a lag between the time at which the node actually failed and the time at which it was detected to fail and then a lag between the time at which it was detected to fail and the time at which the number of fission pulses was recorded from the LCD monitor. If such a lag occurred, it could lead to overestimation of the number of neutrons per square centimeter to which the device under test was exposed and, consequently, underestimation of the device's susceptibility to neutrons.

Finally, while HP personnel were able to control many factors that could affect the experimental results, such as the number of BTAG SRAM equivalents under test, the test program used, and the beam aim, they were not able to control every relevant factor. Those not under the explicit control of HP personnel include the duration of node exposure to neutrons during experiments that included node crashes, any mid-experiment downtime resulting from node crashes and restarts, the precise intensity of the beam, since it may fluctuate with time, and high-energy particles from sources other than the beam which might result in soft errors in the SRAMs under test.

III. Results and Discussion

A. Data Analysis

Data cleaning was required before the data could be analyzed. This was a nontrivial task as the error logs contained multiple counts of some soft errors and soft errors that occurred in SRAMs that were not being tested. Next, an exploratory data analysis was conducted. We discuss results from three datasets formed following this exploratory data analysis. The first dataset consists of counts of soft errors from experiments in which the memtest program was used and the beam was aimed at the location(s) of the BTAG SRAM, a pair of DATA SRAMs, two pairs of DATA SRAMS, or the BTAG SRAM and a pair of DATA SRAMs. Thus, this dataset combines counts of soft errors that typically would be nonfatal and counts of soft errors that typically would be fatal. We refer to this dataset as "memtest count data." As discussed in Section II, counts of soft errors in the DATA SRAMs that typically would be nonfatal and counts of soft errors in the BTAG SRAM that typically would be fatal may be combined to estimate Q's susceptibility to fatal soft errors since Q's DATA SRAMs and BTAG SRAMs are physically the same, and hence, should have the same susceptibility to soft errors resulting from cosmic-ray-induced neutrons, whether fatal or nonfatal. The second dataset includes counts of soft errors in the DATA SRAMs that were recorded by the btagexer program. These soft errors typically would be nonfatal. We refer to this dataset as "btagexer count data." As with the memtest count data, these counts of soft errors that typically would be nonfatal may be used to estimate Q's fatal soft-error rate since Q's DATA SRAMs and BTAG SRAMs are physically the same. The final dataset includes times until failure, measured in neutrons per square centimeter until failure, recorded when the btagexer program was used and the beam was aimed at the location of the BTAG SRAM. We refer to this dataset as "btagexer failure-time data." We excluded three counts from the 37 soft-error counts in the memtest count and btagexer count datasets from the analyses because they were suspect.

A Poisson model was used for the memtest count data and the btagexer count data, while an exponential model was used for the btagexer failure-time data. The Poisson model for the count of unique soft errors from experiment j in dataset i,

where i indexes the two count datasets (memtest count data and btagexer count data) y_{ij}, is as follows

$$y_{ij}|\lambda \sim \mathrm{Po}(\lambda b_{ij} p_{ij} f_{ij})$$

where b_{ij} is the number of BTAG SRAM equivalents under test during experiment j in dataset i, p_{ij} is the number of fission pulses recorded from the LCD counter for experiment j in dataset i, f_{ij} is the estimate of the average number of neutrons per square centimeter per fission pulse that the node was exposed to during experiment j in dataset i, λ is an unknown parameter with units fatal soft errors per BTAG SRAM per neutron per square centimeter, and p_{ij} and f_{ij} are assumed to be known exactly for the analyses presented in this paper. With this model, $\lambda b_{ij} p_{ij} f_{ij}$ is the mean or expected value of y_{ij}.

Our analyses assume that the results from different experiments are independent. In other words, we assume that conditional on the Poisson means $\lambda b_{ij} p_{ij} f_{ij}$ and $\lambda b_{ik} p_{ik} f_{ik}$, the number of unique soft errors counted during experiment j in dataset i provides no information about the number of unique soft errors counted during experiment k in dataset i.

Our model includes a Gamma prior for λ

$$\lambda \sim \mathrm{Gamma}(\alpha, \beta)$$

which is parameterized to have mean α/β and variance α/β^2. We let $\alpha = 18.31$ and $\beta = 11\,672\,626$, values based on Q CPU failure data during an 8-week period beginning in mid-November 2002 and ending in mid-January 2003. Since fatal soft errors are one type of CPU failure that is logged for Q, the data on which these parameter values are based likely overstate the number of fatal soft errors observed during the corresponding time period. However, it is believed that the majority of CPU failures are caused by fatal soft errors, so these values should not overestimate the number of fatal soft errors during this 8-week time period substantially.

Modeling efforts focused on estimation of the parameter λ. Once estimated, λ must be adjusted to reflect the ambient flux of neutrons per square centimeter per second capable of causing a fatal soft error in Q's BTAG SRAMs in the location where Q is housed. This yields a rate of fatal soft errors per BTAG SRAM per second. We cannot use the results from the LANSCE experiment to directly estimate the rate of fatal soft errors per BTAG SRAM per second since the LANSCE beam may contain neutrons that do not have enough energy to cause a fatal soft error in Q.

With the Poisson model, inference about λ is based on its posterior distribution:

$$\lambda|\ \vec{y} \sim \mathrm{Gamma}\left(\sum_{j=1}^{n_i} y_{ij} + \alpha, \sum_{j=1}^{n_i} b_{ij} p_{ij} f_{ij} + \beta\right)$$

where n_i is the number of experiments in dataset i and $\vec{y}_i = (y_{i1}, \ldots, y_{in_i})$ is the vector of soft-error counts from the n_i experiments in dataset i.

We use an exponential model for the btagexer failure-time data in which we assume that the number of neutrons per square centimeter until failure in experiment j is given exactly by the product $p_j f_j$ (we omit the subscript for dataset in this discussion since there is a single failure-time dataset). With this assumption, the exponential model has the following form

$$p_j f_j|\lambda \sim \mathrm{Exponential}(b_j \lambda)$$

where b_j, p_j, f_j, and λ have analogous definitions to those in the Poisson model. We further assume that given $b_j\lambda$ and $b_k\lambda$, the failure time for experiment j is independent of the failure time for experiment k.

We use a Gamma prior distribution for λ that has the same parameter values as in the Poisson model. With this model specification, the posterior distribution of λ in the exponential model is Gamma as follows

$$\lambda|\ \vec{p}, \vec{f} \sim \mathrm{Gamma}\left(n + \alpha, \sum_{j=1}^{n} b_j p_j f_j + \beta\right)$$

where n is the number of btagexer failure-time experiments, $\vec{p} = (p_1, \ldots, p_n)$ is the vector of counts of fission pulses recorded from the LCD counter for the n btagexer failure-time experiments, and $\vec{f} = (f_1, \ldots, f_n)$ is the vector of estimates of the average numbers of neutrons per square centimeter per fission pulse for each of the n btagexer failure-time experiments.

B. Estimation of the Expected Weekly Number of Fatal Soft Errors in Q

Once the posterior distribution of λ had been calculated based on one of the three datasets described above, the following equation was used to estimate the expected weekly number of fatal soft errors in Q:

$$\lambda \times \varphi \times 8192 \times (3600 \times 24 \times 7)$$

where φ represents the neutron flux, measured in neutrons per square centimeter per second, in the location where Q is housed, the 8192 term scales from a single BTAG SRAM to the 8192 BTAG SRAMs in Q, and the term in parentheses scales the results from seconds to weeks. To calculate a point estimate of this quantity, we replaced λ by its posterior mean based on the results from one of the three experimental datasets described above and φ by $0.025\ \mathrm{neutrons/cm^2/s}$, a value based on [7].

As mentioned in the previous section, the parameter λ has units fatal soft errors per BTAG SRAM per neutron per square centimeter and must be adjusted to reflect the ambient flux of neutrons in Q's location that are capable of causing a soft error in Q's BTAG SRAMs. The value φ provides this adjustment and should include all neutrons capable of producing a fatal soft error in Q. Because the minimum energy required to produce a bit flip in Q's BTAG SRAMs was unknown to us, we used a cutoff value of 10 MeV, a value commonly used in such calculations [9]–[14] and suggested in [6]. Because there are roughly as many neutrons in the ambient neutron spectrum with energy less than 10 MeV as there are with energy greater than 10 MeV [6], this approximation means that our estimates are accurate up to a factor of 2.

TABLE I
EXPERIMENTAL RESULTS

Dataset	$\hat{\lambda}$ Fatal Soft Errors/BTAG SRAM/neutron/cm^2 (95% Posterior Interval)	Estimated Average Weekly Number of Fatal Soft Errors in Q (95% Posterior Interval)
1. Memtest Count Data	1.41e-07 (1.34e-07, 1.48e-07)	17.4 (16.4, 18.5)
2. Btagexer Count Data	1.82e-07 (1.77e-07, 1.87e-07)	22.6 (21.6, 23.6)
3. Btagexer Failure Time Data	7.94e-08 (6.00e-08, 1.02e-07)	9.8 (7.4, 12.7)

Results from three experimental datasets. The results from the memtest count data and the btagexer count data are similar, while those from the btagexer failure-time data are roughly half as large as those from the memtest count data and the btagexer count data. Determination of the cause of this difference would require additional research.

TABLE II
Q HARDWARE FAILURE DATA

Failure Type	Average Weekly Count of Failures (9/5/04–10/23/04)
1. BTAG Parity Errors	24.0
2. CPU Failures	27.7
3. Hardware Failures	35.6
4. Total Failures	47.1

Average weekly counts of four categories of failures that occur in Q. The average weekly number of BTAG parity errors, which may underestimate or overestimate the average weekly number of fatal soft errors in Q, and the average weekly number of CPU failures, which may overestimate the average weekly number of fatal soft errors in Q, are consistent with the predictions in the first two rows of Table I. Fatal soft errors are a substantial proportion of the hardware failures and the total failures experienced by Q.

For each dataset, we calculated 95% posterior probability intervals for λ and the expected weekly number of fatal soft errors in Q. The 95% posterior intervals for the latter quantity incorporate both the uncertainty in the posterior distribution of λ and the variability in the ambient neutron flux where Q is housed. We modeled the ambient neutron flux as normal with mean 0.025 $\mathrm{neutrons/cm^2/s}$ and standard deviation 4.4×10^{-4}. This standard deviation is derived from measurements from a neutron counter deployed in Q's location. Posterior intervals for the expected weekly number of soft errors in Q were calculated via Monte Carlo techniques.

We performed both model checking and sensitivity analyses. Model checking revealed no lack of fit except that the btagexer count data appeared more variable than the model used for them. A sensitivity analysis, which assessed the sensitivity of the results to the assumptions that the number of fission pulses and the mean number of neutrons per square centimeter per fission pulse in each experiment were known without error, was conducted. The results suggested that accounting for this source of variability would not substantially affect point estimates. However, it would likely lead to wider posterior intervals than those presented here. Sensitivity of the results to the prior specification was also assessed. With the exception of the btagexer failure-time data, the results were not sensitive to the prior specification.

C. Results

Table I presents the results of our study: estimates of λ and the estimated average weekly number of fatal soft errors in Q based on the three datasets discussed above along with a 95% posterior interval for each. The first and second rows present results from the memtest count data and the btagexer count data, respectively. The third row presents results from the btagexer failure-time data.

While the results from the memtest count data and the btagexer count data in the first two rows of Table I predict that roughly three fatal soft errors will be observed in Q per day, the results from the btagexer failure-time data in the third row predict roughly half as many fatal soft errors. It is not obvious which of these datasets is more relevant to predicting the average weekly number of fatal soft errors in Q. Although the results from the memtest count data and the btagexer count data are based on more information, the btagexer failure-time data were obtained in a manner that more closely approximates how Q experiences fatal soft errors in a production environment. That is, the node was operated in its normal operating mode until a fatal soft error caused it to crash.

Several explanations for the discrepancy between the count data and the failure-time data were investigated. The following did not appear to explain the difference: major recording or transcription errors in the data, errors in or misrepresentations of the experimental conditions, the lag in recording the number of fission pulses from the LCD counter, or differences between the two test programs. Additional sources of error include fatal soft errors in Q that may have causes other than cosmic-ray-induced neutrons and differences between the neutron spectrum of the LANSCE beam and the ambient neutron spectrum in Q's location. A full understanding of the discrepancy would require further research.

Table II presents hardware failure data from Q for comparison purposes. This first row in Table II contains the average weekly number of BTAG parity errors observed in Q over a 7-week period from early September 2004 through late October 2004. This value may be lower than the actual number of BTAG parity errors during this time period due to recording omissions. In addition, BTAG parity errors may result from mechanisms other than cosmic-ray-induced neutrons. As a result, it is not clear whether this value overestimates or underestimates the number of fatal soft errors in Q that derives from cosmic-ray-induced neutrons. The second row presents the average weekly number of CPU failures in Q over the same time period. This value likely overestimates the number of fatal soft errors observed during this time period since CPU failure counts include both BTAG parity errors resulting from cosmic-ray-induced neutrons and failures resulting from other mechanisms. The final two rows present the average weekly number of hardware failures and the average weekly number of total failures in

Q over the same time period. These latter two values include CPU failures.

The estimates of the average weekly number of fatal soft errors in Q in the first two rows of Table I are 17.4 and 22.6, respectively. These results are consistent with the failure data from Q in Table II (an average of 24.0 BTAG parity errors and 27.7 CPU failures per week) given that the test programs may use memory in a manner that is different from how it is typically used in Q under normal operating conditions, the unknown minimum energy required to cause a fatal soft error in Q, the uncertainties in the observed Q failure data as an estimate of the average weekly number of fatal soft errors it experiences, and the sources of error listed previously. In particular, the unknown minimum energy required to cause a soft error in Q's BTAG SRAMs means that our results could be as little as half of what they would be if very low energy neutrons are capable of causing soft errors in Q's BTAG SRAMs.

Over the same 7-week time period, the observed average weekly number of hardware failures in Q was 35.6, and the observed average total weekly number of failures was 47.1. The two types of failures that may be used to estimate the number of fatal soft errors occurring in Q (CPU failures and BTAG parity errors) comprise the bulk of the hardware failures and total failures, indicating that the remaining components and subsystems in Q are reliable. These data also indicate that soft errors are an important consideration for large systems.

Since Q is used in the presence of many types of single-node failures, various mitigation strategies have been developed. To mitigate the impact of fatal soft errors in Q, preexecutables that remove preexisting BTAG parity errors in Q's BTAG SRAMs are run before a job is launched on Q. Data to evaluate the efficacy of this strategy are not immediately available.

Other strategies that address node failures including those deriving from fatal soft errors involve the use of checkpointing and reserve nodes. For example, a user submitting a job may optimize the interval between his checkpoints according to system and job characteristics such as the required number of processors for the job, the mean time to system interrupt for the segment of the machine on which the job is running, and the amount of time needed to perform a checkpoint. In this manner, the interval at which checkpoints are performed may be chosen to minimize expected total application runtime [23]. In addition, a user may run on fewer nodes than are in his allocation. In this case, when a single node fails, the application may be immediately restarted on the remaining nodes without waiting for the failed node to be fixed or restarted [24]. While these strategies do not reduce the number of node failures or BTAG parity errors that Q experiences, they do mitigate the impact of such failures on application runtime. For example, based on simulations in [25], nonoptimal choice of the checkpoint restart interval may add as many as 15 or more hours to the length of time required to complete a 500-h calculation on a machine with a mean time to failure of 6 h, a restart time of 10 min, and a checkpoint file write time of 5 min. Simulations in [26] suggest that when using the optimal checkpoint restart interval, running a 500-h calculation with insufficient reserve nodes can increase the total wall-clock time to completion by 20 or more hours for a calculation performed on a system with a mean time to failure of 7.8 h, a restart time of 15 min, and a checkpoint file write time of 6 min. Despite single-node failures of different types, Q's resources are sought after and successfully used by scientists to perform scientific calculations and meet milestones.

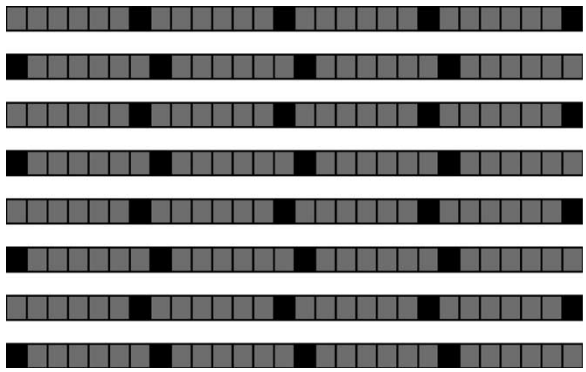

Fig. 1. Basic architecture of QA or QB. The gray blocks denote cabinets that contain server nodes, while the black blocks denote cabinets that contain domain anchor nodes.

IV. Conclusion and Future Challenges

In this work, we used relatively simple techniques that drew on existing technology to perform experiments that provided useful predictions of the rate at which fatal soft errors occur in Q. These predictions corroborated the hypothesis that cosmic-ray-induced neutrons are the primary cause of BTAG parity errors in Q. Through a process of application hardening and other mitigation strategies, scientists have been successfully using Q for over two years in the face of fatal soft errors and other failures.

Understanding the mechanism underlying the difference between the predictions from the memtest count data and the btagexer count data and the predictions from the btagexer failure-time data would require further investigation. Statistical models that more fully reflect all of the sources of variability in the data are a subject of ongoing research.

The impact of technological advances on device susceptibility to soft errors is unknown, and the issue of soft errors in large systems will need to be considered regardless of the altitude at which they are housed unless such systems are well shielded. It would be beneficial to have rates of soft errors deriving from cosmic-ray-induced neutrons for different components published yearly or on another periodic basis so that customers can make informed decisions that reflect technological advances.

Appendix

This appendix details Q's architecture. For programmatic purposes, Q is typically used as two segments with identical architectures, QA and QB. Fig. 1 contains a schematic of QA or QB. QA (or QB) is composed of eight rows, each of which contains 128 ES45 nodes in 28 cabinets. These 128 nodes are divided into four domains of 32 nodes. Each domain contains two domain anchor nodes in a single cabinet and 30 server nodes in six cabinets, each of which contains five nodes. A Quadrics interconnect connects all 1024 nodes in QA or QB.

ACKNOWLEDGMENT

The authors would like to thank E. Buenafe, J. Daly, M. Devlin, R. Klamann, H. Kutac, R. Miller, J. O'Donnell, S. Shaw, M. Vernon, and M. Vigil for their contributions to this work. The authors would also like to thank two anonymous referees for their insightful comments which improved the presentation of this work.

REFERENCES

[1] TOP500 List. (Jun. 2003). [Online]. Available: http://top500.org/list/2003/06/?page

[2] Compaq Computer Corporation. (2002). *AlphaServer ES45 Owners Guide* [Online]. Available: http://h18002.www1.hp.com/alphaserver/download/ekes450-ug-b01.pdf

[3] R. Kessler, E. McLellan, and D. Webb, *The Alpha 21264 Microprocessor Architecture* [Online]. Available: http://h18002.www1.hp.com/alphaserver/download/ev6chip.pdf

[4] *The Message Passing Interface (MPI) Standard*. [Online]. Available: http://wwwunix.mcs.anl.gov/mpi/

[5] K. W. Harris, "Asymmetries in soft-error rates in a large cluster system," *IEEE Trans. Device Mater. Rel.*, vol. 5, no. 3, pp. 336–342, Sep. 2005.

[6] *Measurement and Reporting of Alpha Particles and Terrestrial Cosmic Ray-Induced Soft Errors in Semiconductor Devices,* JEDC standard, JESD89, 2001.

[7] M. S. Gordon, P. Goldhagen, K. P. Rodbell *et al.*, "Measurement of the flux and energy spectrum of cosmic-ray induced neutrons on the ground," *IEEE Trans. Nucl. Sci.*, vol. 51, no. 6, pp. 3427–3434, Dec. 2004.

[8] S. A. Wender, "Neutron single event effects testing at LANSCE," presented at the IEEE Int. Reliability Physics Symp., Dallas, TX, Mar. 30, 2003.

[9] C. A. Gossett, B. W. Hughlock, M. Katoozi, G. S. LaRue, and S. A. Wender, "Single event phenomena in atmospheric neutron environments," *IEEE Trans. Nucl. Sci.*, vol. 40, no. 6, pp. 1845–1852, Dec. 1993.

[10] E. Normand, D. L. Oberg, J. L. Wert, J. D. Ness, P. P. Majewski *et al.*, "Single event upset and charge collection measurements using high energy protons and neutrons," *IEEE Trans. Nucl. Sci.*, vol. 41, no. 6, pp. 2203–2209, Dec. 1994.

[11] E. Normand, "Single event upsets at ground level," *IEEE Trans. Nucl. Sci.*, vol. 43, no. 6, pp. 2742–2750, Dec. 1996.

[12] Y. Tosaka, S. Satoh, K. Suzuki, T. Sugii, N. Nakayama *et al.*, "Measurement and analysis of neutron-reaction-induced charges in a silicon surface region," *IEEE Trans. Nucl. Sci.*, vol. 44, no. 2, pp. 173–178, Apr. 1997.

[13] Y. Tosaka, S. Satoh, T. Itakura, H. Ehara, T. Ueda *et al.*, "Measurement and analysis of neutron-induced soft errors in sub-half-micron CMOS circuits," *IEEE Trans. Electron Devices*, vol. 45, no. 7, pp. 1453–1458, Jul. 1998.

[14] P. Hazucha, C. Svensson, and S. A. Wender, "Cosmic-ray soft error rate characterization of a standard 0.6-μm CMOS process," *IEEE J. Solid-State Circuits*, vol. 35, no. 10, pp. 1422–1429, Oct. 2000.

[15] P. Hazucha and C. Svensson, "Cosmic ray neutrons multiple-upset measurements in a 0.6-μm CMOS process," *IEEE Trans. Nucl. Sci.*, vol. 47, no. 6, pp. 2595–2602, Dec. 2000.

[16] T. Granlund, B. Granbom, and N. Olsson, "Soft error rate increase for new generations of SRAMs," *IEEE Trans. Nucl. Sci.*, vol. 50, no. 6, pp. 2065–2068, Dec. 2003.

[17] H. Kobayashi, H. Usuki, K. Shiraishi, H. Tsuchiya, N. Kawamoto *et al.*, "Comparison between neutron-induced system-SER and accelerated-SER in SRAMs," in *Proc. 42nd Int. Reliability Physics Symp.*, Phoenix, AZ, 2004, pp. 288–293.

[18] P. Hazucha, T. Karnik, S. Walstra, B. A. Bloechel, J. W. Tschanz *et al.*, "Measurements and analysis of a SER-tolerant latch in a 90-nm dual-V_T CMOS process," *IEEE J. Solid-State Circuits*, vol. 39, no. 9, pp. 1536–1543, Sep. 2004.

[19] J. F. Ziegler, H. P. Muhlfeld, C. J. Montrose, H. W. Curtis, T. J. O'Gorman *et al.*, "Accelerated testing for cosmic soft-error rate," *IBM J. Res. Develop.*, vol. 40, no. 1, pp. 51–72, Jan. 1996.

[20] T. J. O'Gorman, J. M. Ross, A. H. Taber, J. F. Ziegler, H. P. Muhlfeld *et al.*, "Field testing for cosmic ray soft errors in semiconductor memories," *IBM J. Res. Develop.*, vol. 40, no. 1, pp. 41–50, Jan. 1996.

[21] S. E. Michalak, K. W. Harris, N. W. Hengartner, B. E. Takala, and S. A. Wender, "Using the LANSCE irradiation facility to predict the number of fatal soft errors in one of the world's fastest supercomputers," in *Proc. 18th Int. Conf. Application Accelerators Research and Industry*, Fort Worth, TX, 2004, to be published.

[22] S. A. Wender, S. Balestrini, A. Brown, R. C. Haight, C. M. Laymon *et al.*, "A fission ionization detector for neutron flux measurements at a spallation source," *Nucl. Instrum. Methods Phys. Res., A*, vol. 336, no. 1-2, pp. 226–231, 1993.

[23] J. T. Daly, "A strategy for running large scale applications based on a model that optimizes the checkpoint interval for restart dumps," in *Proc. 1st Int. Workshop Software Engineering High Performance Computing System Applications*, Edinburgh, Scotland, 2004, pp. 70–74.

[24] ——, "Milestone Performances on the Q Machine," Los Alamos Nat. Lab., Los Alamos, NM, Tech. Rep. LA-CP-03-0278, 2003.

[25] ——, "A higher order estimate of the optimum checkpoint interval for restart dumps," in *Future Generation Computing Systems*. Amsterdam, The Netherlands: Elsevier, 2004.

[26] ——, "Evaluating the performance of a checkpointing application given the number and types of interrupts," in *Proc. Workshop High Performance Computing Reliability Issues*, San Francisco, CA, 2005.

Sarah E. Michalak received the B.A. degree in mathematics from Yale University, New Haven, CT, in 1992, and the M.A. and Ph.D. degrees in statistics from Harvard University, Cambridge, MA, in 1996 and 2001, respectively.

She is a Technical Staff Member with the Statistical Sciences Group at Los Alamos National Laboratory (LANL), Los Alamos, NM. Her research interests include methods for predicting soft-error rates based on neutron beam testing, the performance and reliability of supercomputers, statistics and public health, chemometrics, and hierarchical models including prior specification and conditions under which the posterior distribution is proper.

Dr. Michalak is a Member of the American Statistical Association and the International Society for Bayesian Analysis. She received two team awards for her work on the Advanced Simulation and Computing (ASC) Q project.

Kevin W. Harris (M'04) received the B.S. degree in mathematics from the University of Maryland, College Park, in 1974, and the M.S. degree in computer science from Pennsylvania State University, University Park, in 1976.

He is a Research Engineer with the High Performance Computing Division at Hewlett-Packard (HP), Nashua, NH. His early employment history included 3 years in a cosmic-ray detection group under Dr. Robert Hartman in the High Energy Astrophysics Division of NASA at Goddard Space Flight Center. Most of his career was spent at Digital Equipment Corporation working on optimizing compilers and parallel processing technologies. His interests are in pushing the scalability limits of all aspects of computing and bringing the results to market. He currently works on scalability issues for HP's compute cluster products.

Nicolas W. Hengartner, photograph and biography not available at the time of publication.

Bruce E. Takala, photograph and biography not available at the time of publication.

Stephen A. Wender, photograph and biography not available at the time of publication.