System Operations and Development Team

1. Team members

Atsuya Uno (Team Head)
Hitoshi Murai (Research & Development Scientist)
Motoyoshi Kurokawa (Research & Development Scientist)
Keiji Yamamoto (Postdoctoral Researcher)
Fumio Inoue (Research & Development Scientist)
Mitsuo Iwamoto (Technical Staff)
Katsufumi Sugeta (Technical Staff)

2. Research Activities

The K computer is a distributed-memory parallel computer consisting of 82,944 compute nodes. It has played a central role in the High Performance Computing Infrastructure (HPCI) initiative granted by the Ministry of Education, Culture, Sports, Science and Technology. The HPCI has achieved the integrated operation of the K computer and other supercomputer centers in Japan and has enabled seamless access from user machines to a cluster of supercomputers that includes the K computer. Moreover, the HPCI has provided large-scale storage systems that are accessible from all over Japan.

The System Operations and Development Team (SODT) has conducted research and development on advanced management and operations of the K computer. While analyzing operational statistics collected during shared use, the SODT has improved the system configuration, including aspects involving job scheduling, the filesystem, and user environments. For example, achieving higher system utilization is very difficult because the K computer has to process various sizes and types of jobs simultaneously. The SODT has responded flexibly to user requests and analyzed the operational status, thereby realizing a high level of utilization of approximately 75% in FY2014. Moreover, the SODT has developed tools that improve the usability of the K computer.

The K computer's power consumption exceeded the limit several times during FY2013. This is a problem because it forces us to renew the contract-power upper limit, which means expense for the electricity increases. To prevent this increase, the SODT investigated an emergency job-stopping method based on the estimated power consumption of each job using thermal sensors in computer racks.

The SODT also helped users manage the K computer and utilize K computer resources effectively by improving the system software. This support was conducted together with the software development team.

3. Research Results and Achievements

3.1. System software improvements of the K computer

We fixed and improved many aspects of the system software through shared use. Here, we describe the major activities in FY2014.

- Analysis of operation statistics

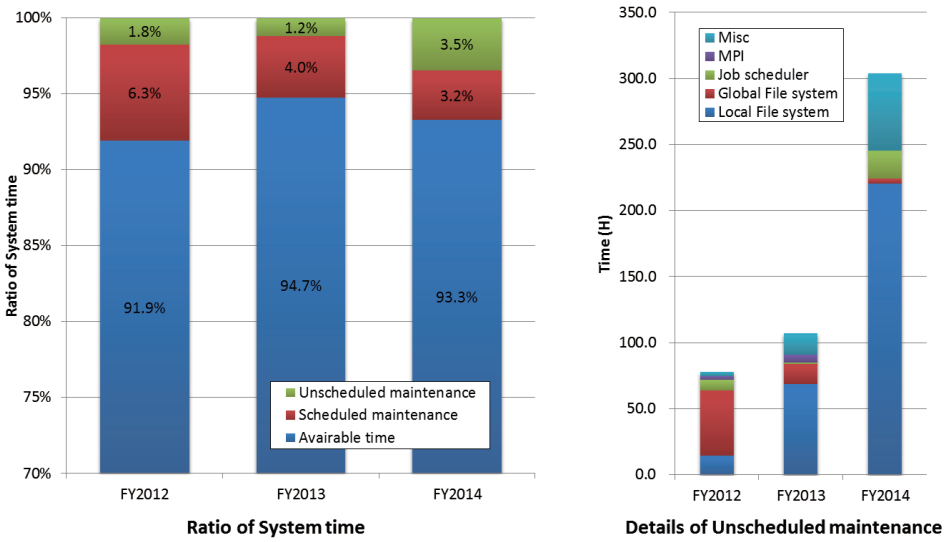

Figure 1 Ratio of system time (left) and details of unscheduled maintenance (right)

Figure 1 shows the ratio of system time (left) and details of unscheduled maintenance (right). Scheduled maintenance primarily consisted of system software updates and maintenance at the end of FY2014, whereas unscheduled maintenance consisted of system unavailability because of unexpected system failures, such as hardware errors, system software bugs, and so on. Overall, we have made the K computer available to users for more than 90% of the time. As the K computer has stabilized, the scheduled maintenance time decreased, but the unscheduled maintenance time in FY2014 was tripled as compared with FY2013. Primarily, the unscheduled maintenance in FY2014 was because of local filesystem failures as the recovery from one local filesystem failure consumes a long time to reconfigure the filesystem. The reduction of such failures is the most important problem that we need to resolve promptly.

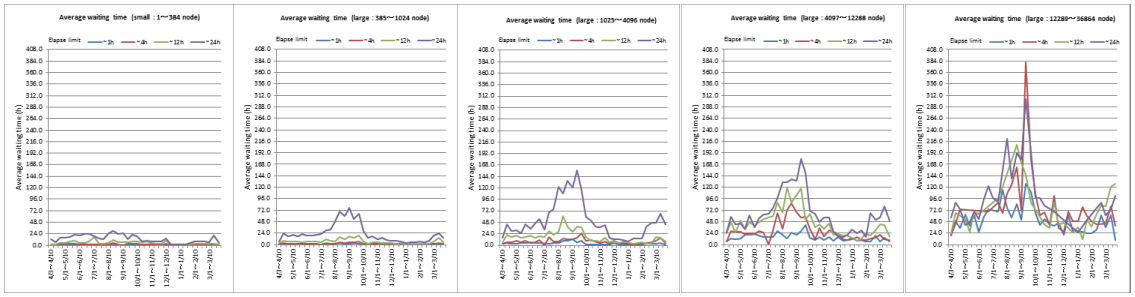

Figure 2 Average waiting times in FY2014

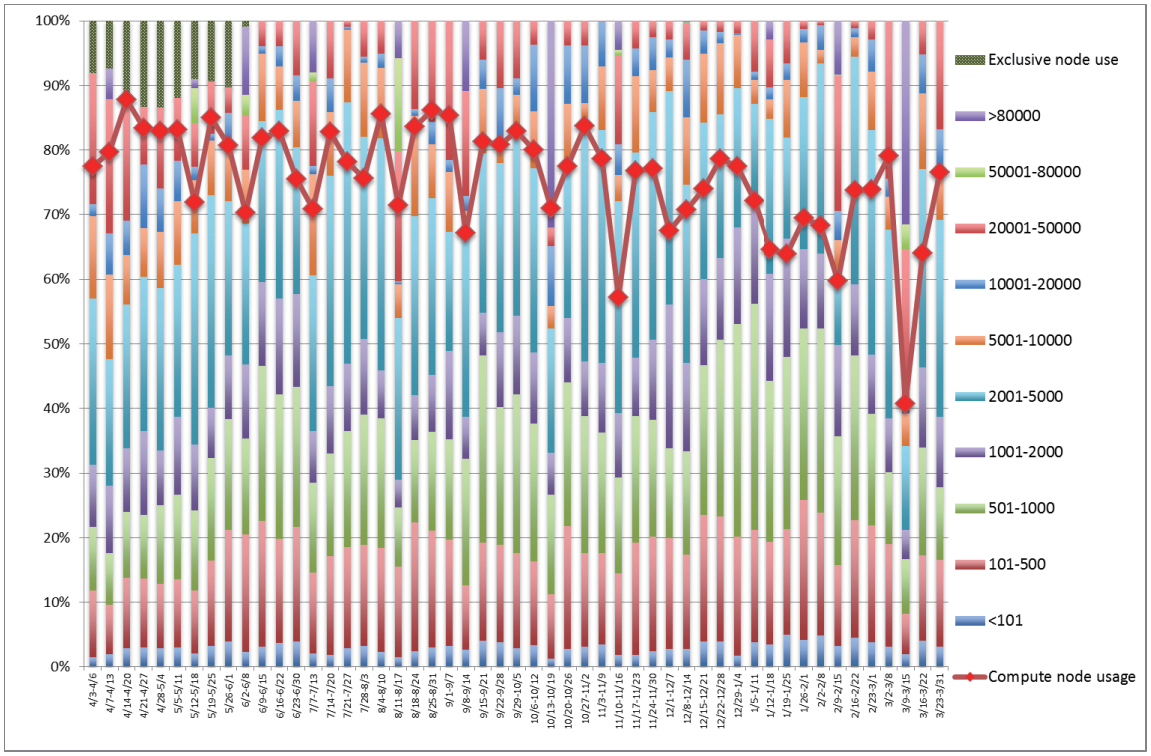

Figure 3 Details of resource usage in FY2014

Figure 2 shows average waiting times with respect to both job sizes and elapsed times in FY2014. The average waiting times were directly proportional to both elapsed time and the number of nodes, indicating that the fairness in job execution has been kept. The average waiting times in September 2014 were substantially longer than those in other months. Each group had appropriate computing resources for a year, and these resources were divided into two terms as follows: (1) from April to September and (2) from October to March. To consume the remaining computing resources before they expire, users tend to submit many jobs at the end of each term. In FY2013, job congestions clearly occurred in August and September 2013 and March 2014, and in FY2014, they occurred during the same months; however, this year's average waiting times were approximately 50%–70% as compared to FY2013. In FY2014, computing resources assigned to users were a few percentage points less than those in FY2013. This decrease led to a significant reduction of the average waiting times in FY2014.

Figure 3 shows resource usage details for FY2014. We achieved approximately 75% node usage in FY2014; more specifically, node usage during the first term of FY2014 was 79%, whereas that during the second term was 72%. If a user uses all the first term's resources within the term, he/she can use the second term's resources also in the first term. Therefore, if too much of the second term's resources are used in the first term, there will be a shortage in the second term's resources. In FY2014, this happened, and the second term's node usage was decreased. To prevent this, for FY2015, we will assign computational resources more than those of FY2014 and less than those of

FY2013; moreover, we will make new rule to suppress the overuse of the second term's resources in the first term.

- Analysis of job scheduling

We analyzed job scheduling on the K computer using a simulator. In FY2014, we evaluated system utilization on the basis of differences in start times of file staging preceding job execution; the results are shown in Figure 4. System utilization increased in the preceding time from 0 to 6 hours, descending from 12 to 24 hours. When the preceding time was long, the compute node assigned to the job were determined earlier; thus, if the compute nodes in which the job could be assigned in an earlier start time were free, the job could not be rescheduled. This means that the preceding time of file staging was one of the important parameters in job scheduling.

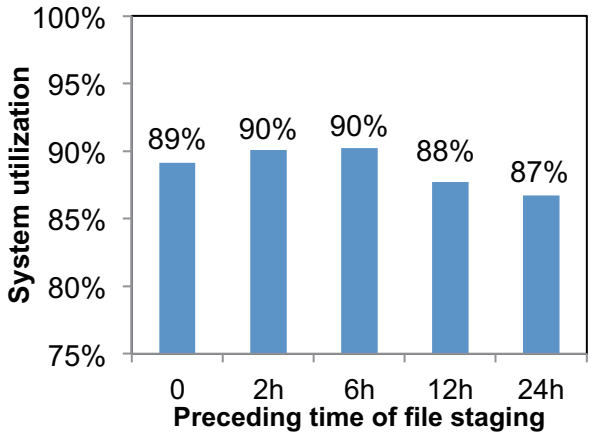

Figure 4 System utilization by time preceding file staging

3.2. Approach for the power consumption problem

The K computer's power consumption exceeded the given limit several times during FY2013; this is an important problem because it forces us to increase the contract-power upper limit, thereby increasing the cost, which cannot be ignored. To prevent this problem, we performed a preliminary review that estimated the power consumption of each job, thereby enabling us to control the overall K computer's power consumption. Moreover, we investigated an emergency job-stopping method based on the estimated power consumption of each job in case power consumption again exceeds the given limit.

- Preliminary review of large-scale jobs

  To prevent power consumption from exceeding the given limit, we performed a preliminary review of large-scale jobs. In this process, we estimated the power consumption of jobs using all compute nodes versus the power consumption of a small-scale job with the same characteristics, thereby permitting users to submit jobs if the estimated power consumption is below the limit.

- Emergency job-stopping method

When power consumption exceeds the given limit, we have to stop the jobs to reduce it. To select the appropriate jobs to be stopped, we proposed a method to estimate the power consumption of each job using thermal sensors in computing racks. By this method, we estimated the job's power consumption using CPU and exhaust air temperature variations. Figure 5 shows the relation between the assigned number of nodes and estimated power consumption of each job during a two-day period from June 11 to June 12, 2014. The upper figure shows the number of compute nodes assigned to each job, indicated by the blue line. The lower figure shows the estimated power consumption of each job, with the red line indicating the observed K computer's power consumption. In both the upper and lower figures, areas with the same color during the same period correspond to the same job. The width of an area denotes the elapsed time from start to termination of the job. The white areas under the blue or red lines indicate all jobs using less than 1,000 compute nodes. From these figures, we note that the estimated power consumption of each job is not proportional to the size of assigned compute nodes, and our proposed method is useful in selecting the appropriate jobs to be stopped.

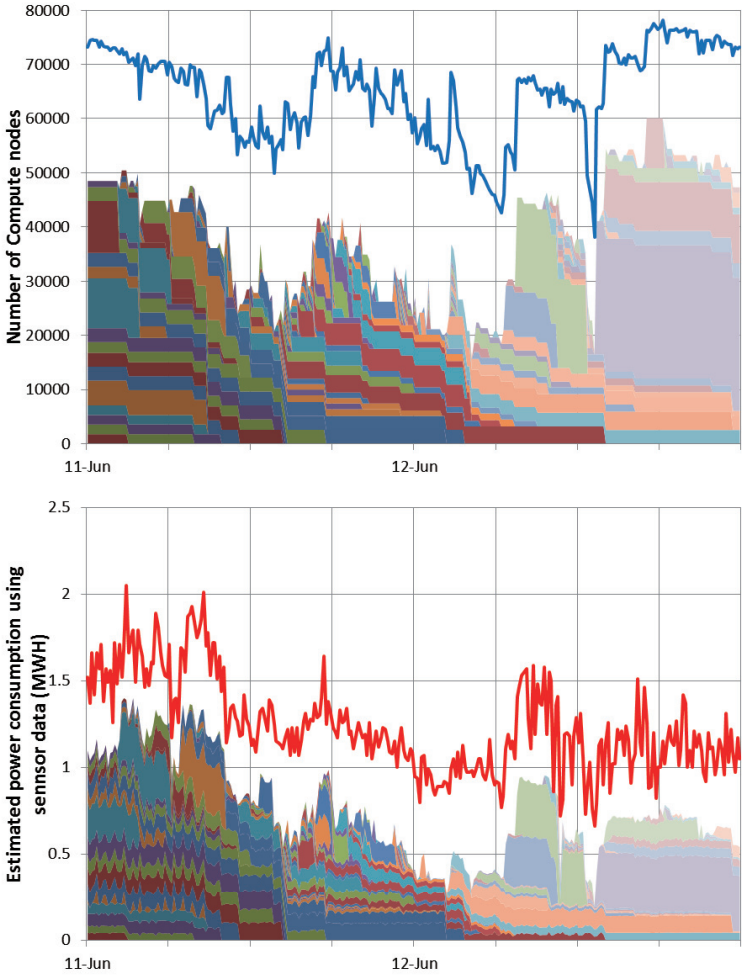

Figure 5 Assigned number of nodes (upper) and estimated power consumption (lower)

3.3. User support

We conducted user support such as consulting services.

- Number of users

  The K computer has over 130 groups and over 1,500 users as of the end of March 2015. The total number of HPCI users and AICS researchers is approximately 1,250 and 250, respectively. The number of daily active users is approximately 110.

- Consulting services

  We supported users through the K support desk, providing users technical information on the K computer, including system environment, system tools, software libraries ,etc. Our consulting services were offered together with the software development team. Figure 6 indicates the number of issues in FY2014, showing the number of new issues in FY2014 to be approximately 170. The number of new issues in FY2013 is approximately 230. Because users have learned how to use the K computer and various approaches and workarounds for problems, we consider the number of new issues to be decreased.

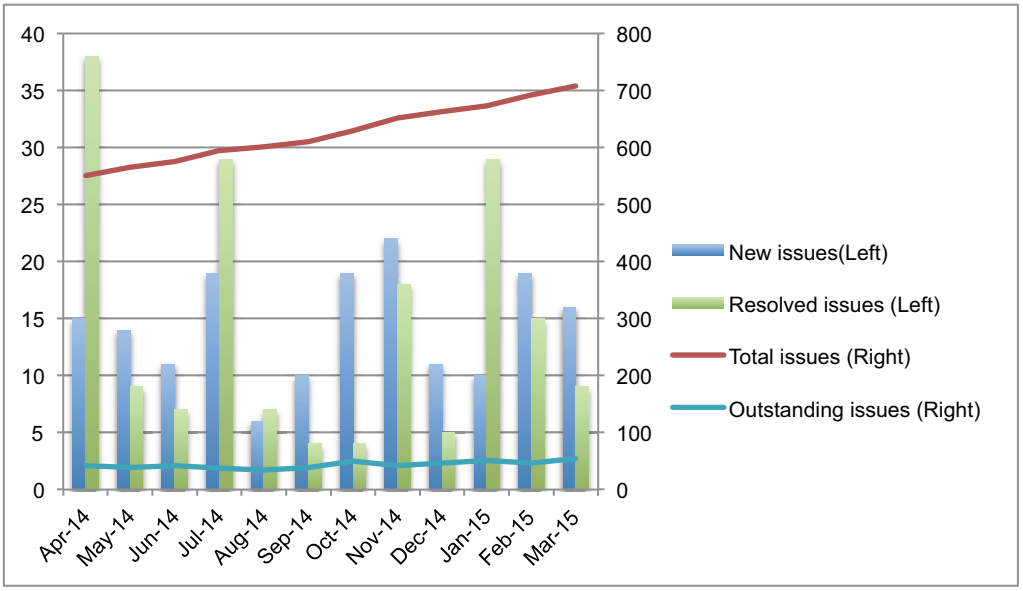

Figure 6 Number of issues in FY2014

4. Schedule and Future Plan

We continue to improve the system software of the K computer and provide user support. Improved system software that improves the usability of the K computer will be released in FY2015. Moreover, we have investigated an approach for the K computer's power consumption problem in terms of job scheduling.

5. Publication, Presentation, and Deliverables

(1) Journal Papers

2. Keiji Yamamoto, Atsuya Uno, Katsufumi Sugeta, Toshiyuki Tsukamoto, Fumiyoshi Shoji, “スーパーコンピュータ「京」の運用状況,” IPSJ Magazine Vol.55 No.8 pp.786-793, 2014. (In Japanese)

(2) Conference Papers

4. Keiji Yamamoto, Atsuya Uno, Hitoshi Murai, Toshiyuki Tsukamoto, Fumiyoshi Shoji, Shuji, Matsui, Ryuichi Sekizawa, Fumichika Sueyasu, Hiroshi Uchiyama, Mitsuo Okamoto, Nobuo Ohgushi, Katsutoshi Takashina, Daisuke Wakabayashi, Yuki Taguchi, Mitsuo Yokokawa, “The K computer Operations: Experiences and Statistics,” International Conference on Computational Science (ICCS), 2014, Australia.
5. Atsuya Uno, Hajime Hida, Fumio Inoue, Naoki Ikeda,Toshiyuki Tsukamoto, Fumichika Sueyasu, Satoshi Matsushita, Fumiyoshi Shoji, “Operation of the K computer Focusing on System Power Consumption,” HPCS2015, 2015, Japan. (In Japanese) (to appear)

(3) Invited Talks

- None

(4) Posters and Presentations

3. Keiji Yamamoto, Atsuya Uno, Ryuichi Sekizawa, Daisuke Wakabayashi, Fumiyoshi Shoji, “区間スケジューリングを用いたジョブスケジューリングの性能評価,” IPSJ-SIGHPC 2014-HPC-146 No.3, 2014. (In Japanese)
4. Fumio Inoue, Atsuya Uno, Toshiyuki Tsukamoto, Satoshi Matsushita, Fumichika Sueyasu, Naoki Ikeda, Hajime Hida, Fumiyoshi Shoji, “電力消費量の上限を考慮した「京」の運用,” IPSJ-SIGHPC 2014-HPC-146 No.4, 2014. (In Japanese)
5. Atsuya Uno, Hajime Hida, Naoki Ikeda, Fumio Inoue, Toshiyuki Tsukamoto, Fumichika Sueyasu, Fumiyoshi Shoji, “「京」におけるジョブ単位の消費電力推定の検討,” IPSJ-SIGHPC 2014-HPC-147 No.20, 2014. (In Japanese)
6. Keiji Yamamoto, Atsuya Uno, Ryuichi Sekizawa, Daisuke Wakabayashi, Fumiyoshi Shoji, “区間ジョブスケジューリング法へのファイルステージング導入に伴う性能評価,” IPSJ-SIGHPC 2015-HPC-148 No.29, 2015. (In Japanese)

(5) Patents and Deliverables

- None