Evaluation Metrics for the Paragon XP/S-15

Bernard Traversat[1], David McNab[1], Bill Nitzberg[1]
and Sam Fineberg[1]

Report RND-93-017 December 1993

`travers@nas.nasa.gov`
NAS Systems Development Branch
NAS Systems Division
NASA Ames Research Center
Mail Stop 258-6
Moffett Field, CA 94035-1000

Abstract

On February 17th 1993, the Numerical Aerodynamic Simulation (NAS) facility located at the NASA Ames Research Center installed a 224 node Intel Paragon XP/S-15 system. After its installation, the Paragon was found to be in a very immature state and was unable to support a NAS users' workload, composed of a wide range of development and production activities. As a first step towards addressing this problem, we implemented a set of metrics to objectively monitor the system as operating system and hardware upgrades were installed. The metrics were designed to measure four aspects of the system that we consider essential to support our workload: availability, utilization, functionality, and performance.

This report presents the metrics collected from February 1993 to August 1993. Since its installation, the Paragon availability has improved from a low of 15% uptime to a high of 80%, while its utilization has remained low. Functionality and performance have improved from merely running one of the NAS Parallel Benchmarks to running all of them faster (between 1 and 2 times) than on the iPSC/860. In spite of the progress accomplished, fundamental limitations of the Paragon operating system are restricting the Paragon from supporting the NAS workload. The maximum operating system message passing (NORMA IPC) bandwidth was measured at 11 Mbytes/s, well below the peak hardware bandwidth (175 Mbytes/s), limiting overall virtual memory and Unix services (i.e. Disk and HiPPI I/O) performance. The high NX application message passing latency (184 μs), three times than on the iPSC/860, was found to significantly degrade performance of applications relying on small message sizes. The amount of memory available for an application was found to be approximately 10 Mbytes per node, indicating that the OS is taking more space than anticipated (6 Mbytes per node).

[1] Computer Sciences Corporation, NASA Contract NAS 2-12961, Moffett Field, CA 94035-1000

1.0 Introduction

The Numerical Aerodynamic Simulation (NAS) facility located at the NASA Ames Research Center has a mission to provide to the aerospace community, by the year 2000, an operational computing system capable of sustaining a TeraFlops computing rate to solve large scale Computational Fluid Dynamics (CFD) applications [1]. Highly Parallel Processor systems, composed of thousands of processors, are thought to be one possible architecture for achieving such high performance. In order to replace conventional vector supercomputer systems (i.e. Cray Y-MP, Cray C90) currently in production use at NAS, these systems must be able to support a production workload. This workload is composed of a wide range of development and production activities involving large scale CFD computations. Interactive and batch jobs are mixed, and a system is commonly shared between a large number of simultaneous users (~100). The ability of highly parallel systems to support this workload is essential for their integration into the NAS computing environment.

On February 17th 1993, a 224 node Intel Paragon XP/S-15 was installed at NAS, to complement a 128 node iPSC/860 and a 128 node CM-5 system. Upon installation, the Paragon was found to be in a very immature state, and unable to support a minimal workload. The system was unable to complete any significant tasks without crashing or hanging. A simple "Hello World" program, in which each node printed the words "Hello World", froze the system when run on more than 16 nodes. The Embarrassingly Parallel (EP) benchmark was the only NAS Parallel Benchmark that would sporadically run on the system.

As a first step towards addressing this problem, a set of metrics was implemented to evaluate system performance as operating system and hardware upgrades were installed. The metrics measure four aspects of the system that we consider essential for our workload:

- **Availability:** the amount of time the system is up to service users' requests.
- **Utilization**: the fraction of time the system is busy servicing users' requests while the system is up.
- **Functionality**: the ability of the system to fulfill users' requests.
- **Performance**: the sustained performance obtained on a representative set of CFD applications.

Availability is essential to ensure that a system is reliable and stays up long enough so users can accomplish their work without frequent interruptions. *Utilization* indicates which portion of the system capacity is being used. A high utilization is usually a good measure of the overall usability of a system. *Functionality* verifies that the system is correctly servicing users' requests (i.e. compilation, job submissions), so that users can perform their work. *Performance* measures how much of the peak hardware capability is achievable under a realistic workload condition. Until significant sustained performance is delivered, it is unlikely that users will move to a new system.

This report presents the metrics collected during the period of February 17th 1993 to the end of August 1993. During this period, significant improvements have been observed in availability, functionality and performance, however utilization has remained low. Critical operating system limitations have been identified which may restrict further progress. This report presents the status of our evaluation and concludes with our overall perception of the progress made at the end of August 1993.

2.0 The Intel Paragon XP/S-15

The Paragon XP/S is a distributed-memory multiprocessor system using a two dimensional mesh interconnection network [2]. Each node consists of two Intel i860 XP microprocessors (one used for computation and one intended for communication) with 16-32 MBytes of local memory. The communication coprocessor was not used during the observed period. The i860 XP runs at 50MHz with a 75 MFLOPS (double precision) peak performance. The mesh routing hardware is capable of delivering a node-to-node peak bandwidth of 175 Mbytes/s (full duplex). The Paragon Operating System (Paragon OS) is based on the Open Software Foundation Advanced Development operating system (OSF/1 AD) [3,4]. OSF/1 AD is a distributed-memory operating system based on the Mach microkernel from Carnegie Mellon University and the OSF/1 Unix implementation. Paragon OS includes the Transparent Network Computing (TNC) extensions from LOCUS Systems providing Unix single system image semantics, the NX message-passing application interface and the Parallel File System (PFS). The Paragon OS architecture is fully and transparently distributed across the system nodes. The *microkernel* resident on each Paragon node implements core operating system functions such as memory management, process management, and inter-process communication (IPC). Higher level Unix services (i.e. file system, networking) run on specific support nodes. Access to the support nodes is transparently provided by the microkernel. In this way, the operating system can potentially scale to support an increased number of nodes or Unix service requirements.

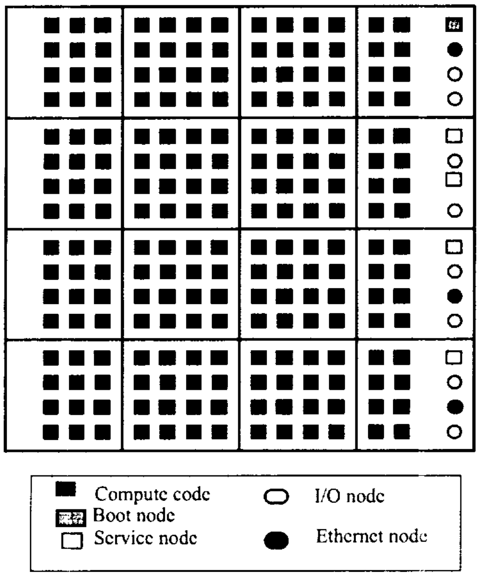


Figure. 1 Paragon XP/S-15 Configuration

The XP/S-15 configuration installed at NAS has two hundred and eight compute nodes (16 Mbytes), eight support nodes (one boot node (32Mbytes), four service nodes (three 32 Mbytes and one 16Mbytes),three Ethernet nodes), and eight additional I/O nodes attached to the RAID disk drives for a capacity of 38 Gbytes of usable space. The service nodes serve as "front-ends" to the system and provide traditional Unix interactive facilities such as editing, compiling and program execution.

3.0 Evaluation Metrics

The metrics implemented provide a limited view, and in some cases, a very rough approximation (e.g. functionality metrics) of the abstract aspects that we were attempting to quantify and measure. However, the metrics were deemed adequate to monitor significant system improvements or degradations as new hardware and software upgrades were installed.

Ideally, each of these aspects should be measured independently. In practice this is impossible. For example, in order to test performance, a certain level of functionality must be present as well as a sufficient level of system availability. The availability and functionality were collected first after the system was installed. As availability and functionality improved, utilization and performance were monitored.

3.1 Availability Metrics

The availability metrics collected *uptime, number of system reboots,* and *mean time to incidents.* An incident was a hardware failure or a system crash or hang.

3.1.1 Uptime

Uptime measured the percent of time the system was available to service user's requests. Dedicated (i.e. maintenance and diagnostic) hours were counted as available. The uptime procedure performed the following steps to determine if the system was available:

1) Check if the system was in dedicated mode.

2) Compile and run "Hello World" on the service node.

3) Compile and run the EP benchmark on 64 compute nodes.

5) Verify correct results.

The procedure ran every hour via cron from February to August. If any step aborted, the test failed and the system was considered down.

Table 1. Uptime Monthly Average

Month	February 93	March 93	April 93	May 93	June 93	July 93	August 93
Uptime (%)	15.53	16.28	54.66	63.33	51.25	73.35	77.44

During the data collection period, the system was under eight hours, five days a week vendor maintenance. Limited vendor support was provided during night hours and weekends

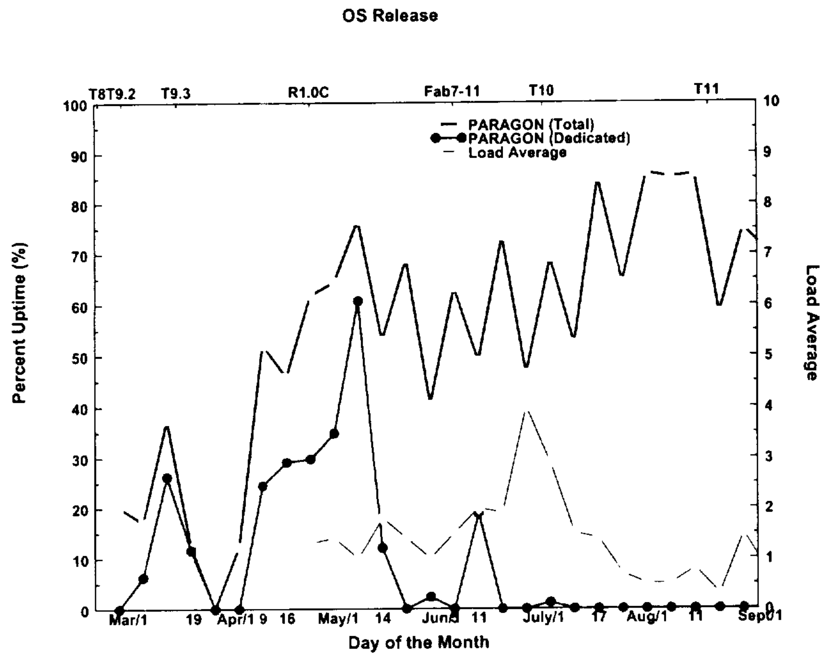


Figure 2. Uptime Weekly Average

Figure 2 shows a significant increase in uptime since the R1.0C OS release was installed on April 23rd. Since then, uptime had fluctuated between a low 50% to a high 86%.

3.1.2 Number of reboots

The number of Paragon's reboots was recorded. Reboots occurred after either a crash, or a hang that prevented useful work from being done. Not surprisingly, the number of reboots was found to be strongly related to system utilization. A day with few reboots did not necessarily imply increased reliability, since often the machine was used minimally (for example during weekend periods).

Table 2. Monthly Number of Reboots

Month	February 93	March 93	April 93	May 93	June 93	July 93	August 93
Number Reboots	22	155	114	73	67	153	90

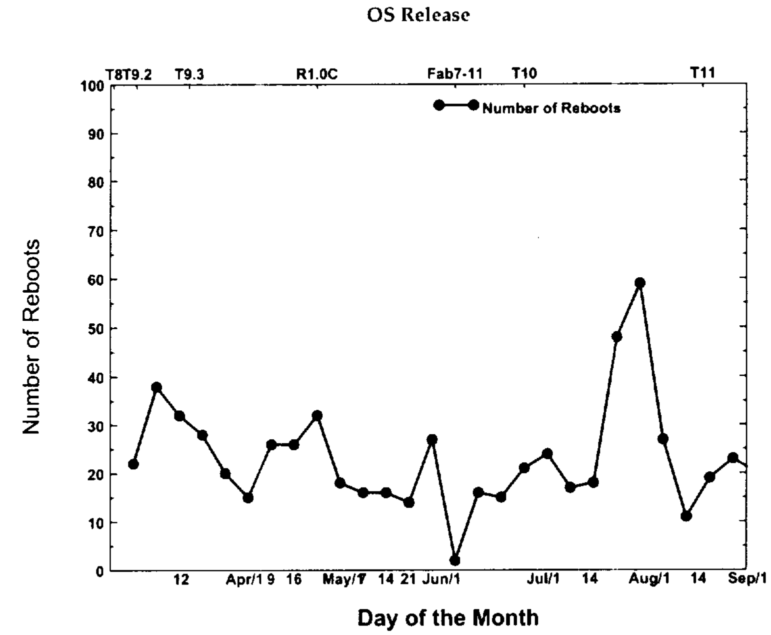


Figure 3. Weekly Number of Reboots

Figure 3 shows a small decrease in the number of reboots as new OS releases were installed. The abnormal increase at the end of July was due to a stress test procedure that was run on the system. This procedure consisted of two copies continuously resubmitted of the NAS Parallel Benchmarks. At the end of August, we observed an average of 3 reboots per day.

3.1.3 Mean Time To Incidents (MTTI)

Mean Time To Incidents (MTTI) was often a better indicator of availability. For example, a machine with 99% uptime may be unusable if it crashes every five minutes (even if it only takes one second to reboot). MTTI was measured using the "uptime" command. The MTTI remained almost constant during the collection period. During the same period, both the uptime and utilization increased. While the system was up more often and the load increased, the frequency of failures did not increase significantly.

Table 3. MTTI Monthly Average

Month	April 93	May 93	June 93	July 93	August 93
MTTI (h:mm)	7:51	6:59	7:20	5:59	8:40

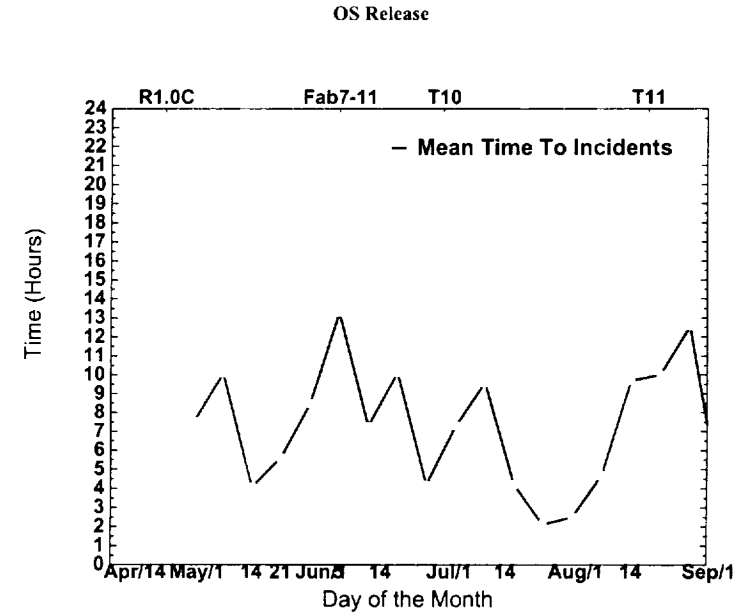


Figure 4. MTTI Weekly Average

The abnormal decrease of MTTI at the end of July (see Figure 4) was due to the previously described stress test.

3.2 Utilization Metrics

The utilization metrics collected user usage of the system. Due to the lack of functional system accounting software, an approximate utilization procedure was implemented. The procedure collected the following information every 15 minutes:

- Number of users logged on the system.
- Load average on the entire compute and service partition (224) nodes. The load average measured the average number of jobs in the run queue over the last 60 seconds on the 224 nodes. A load average of 1 means that one job is running on all 224 nodes.

Table 4. Utilization Monthly Average

Month	April 93	May 93	June 93	July 93	August 93
Average Number of Users	0.73	1.70	3.03	3.09	4.85
Load Average	0.13	0.14	0.29	0.17	0.08

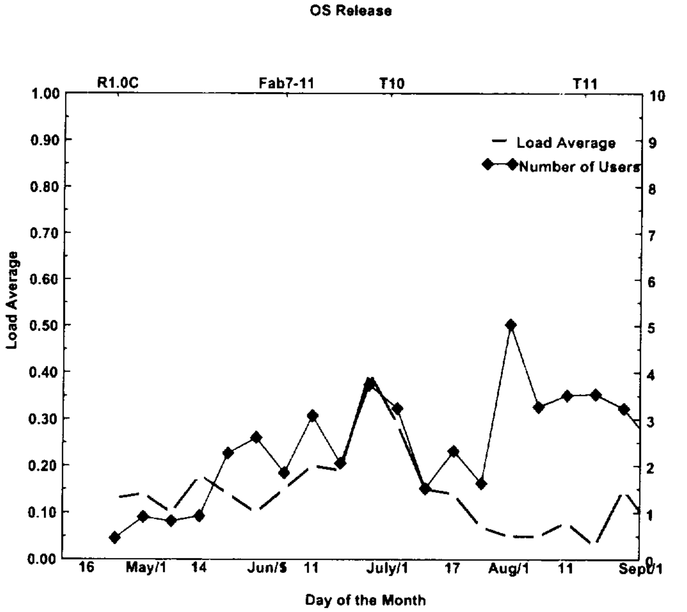


Figure 5. Utilization Weekly Average

From April to the beginning of July, the utilization metric showed a small increase in system usage. Usage of the system remained low (less than 5 users). In Figure 5, the low peak at the end of July was due to the unavailability of the system for the stress test. Since the completion of the stress test at the beginning of August, the utilization of the system has slightly increased as new users were added.

3.3 Functionality Metrics

The functionality metrics measured the ability of the system to fulfill a user request by compiling and running a suite of representative CFD application benchmarks through complete and correct execution. The NAS Parallel Benchmarks [5] and the NHT-1 I/O benchmarks [6] were used to implement the functionality metrics. The NAS Parallel benchmarks are a set of five kernels and three application benchmarks which represent computational parts of important NAS application codes. The NHT-1 I/O benchmarks test peak I/O performance, and I/O performance while executing a typical CFD application.

3.3.1 NAS Parallel Benchmarks

All of the NAS parallel benchmarks compiled and ran correctly after the OS release T10 was installed on June 30th.

Table 5. NAS Parallel Benchmarks Functionality Metrics

Benchmark	Problem Size	Run Correctly on 128 nodes (T10)
EP	2^{28}	yes
MG	256^3	yes
FFT	256^2x128	yes
CG	2.0x10^6	yes
IS	2^{23}	yes
APPSP	64^3	yes
APPBT	64^3	yes
APPLU	64^3	yes

3.3.2 NHT-1 I/O Benchmarks

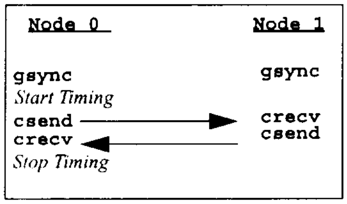
During the period covered in this report, the NHT-1 I/O benchmarks did not run on the Paragon. The I/O application benchmark could not run due to the inability of the Paragon APPBT benchmark implementation to handle the larger problem size (102^3). The peak I/O benchmarks did not run due to the limited functionality of the Paragon Parallel File System (PFS). Parallel I/O operations occurring simultaneously on more than 16 nodes hung the system.

3.4 Performance Metrics

We used the NAS Parallel Benchmarks to evaluate and compare the Paragon performance with the iPSC/860. Specific performance metrics were also implemented to measure performance of key system features. These included: NORMA (No Remote Memory Access) IPC, NX message-passing and virtual memory. NORMA IPC is an extension of Mach IPC used for inter-node communication by the operating system. NX is the message-passing protocol used by users' applications to implement inter-node communication. Both of these communication layers are critical as users' applications are layered on top of them. The purpose of evaluating these system features was also to improve system tuning and identify potential performance bottlenecks.

3.4.1 NX Message Passing Performance

The NX message-passing metrics measured the time necessary for sending a message between two neighbor nodes. The two communicating nodes were synchronized before starting. This attempted to ensure that the receive operation was not posted before the send. The one-way message time was measured by dividing the time for a round trip message by two.



The *csend* and *crecv* synchronous NX message-passing primitives were used to implement this test. The peak bandwidth performance was measured for various message sizes from (0-2 MBytes). The latency was approximated from a linear regression for a message size of zero bytes.

Table 6. NX Peak Bandwidth and Latency

OS Releases	T8	T9.2	T9.3	T9.5 with assertion	T9.5 without Assertion	T9.5 small kernel	R1.0C	T10
Bandwidth (MB/s)	6.5	9.5	9.5	9.5	13.5	16.5	26.5	28
Latency (µs)	250	260	250	250	170	170	185	184

Figure 6 plots message size versus transfer rate for the T10 OS release. Two different communication protocols were used depending on the message size. If the message size was less than 300 Kbytes, bandwidth performance fluctuated between a low (10 Mbytes/s) and a high (20 Mbytes/s) measurement. The bandwidth fluctuation is thought to be due to the extra overhead that occurs when checking for available space in the receiving communication buffer.

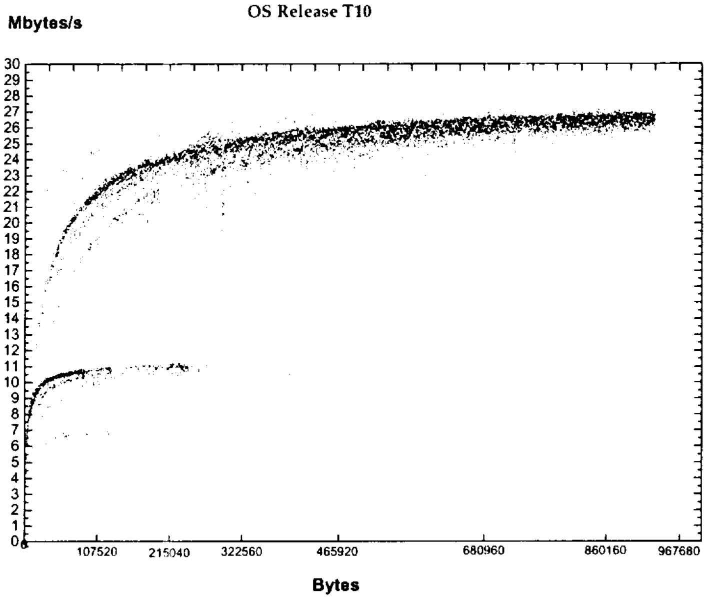


Figure 6. NX Message Passing Bandwidth

During the period covered in this report a steady improvement in communication bandwidth was measured. However, communication latency remained high, at three times the iPSC/860 latency (~50 μs)

3.4.2 NORMA IPC Performance

A test program was implemented to measure NORMA IPC performance. NORMA IPC is an extension of the Mach Inter-Process Communication (IPC) service used for inter-node communication on distributed-memory systems. NORMA IPC performance is critical to the OS and users' applications because it limits the performance of inter-node process communications needed to support OS distributed services such as filesystem, network I/O and virtual memory paging. The lack of performance in the NORMA IPC data structures and its communication protocol can seriously degrade the functionality of a distributed-memory operating system.

The NORMA IPC test sent 1000 messages from a service node process to a compute node process. The message size varied from 32 bytes to 32 Kbytes, in 32 byte increments. Both Mach's in-line and out-of-line transmission modes were used. Out-of-line transmission is more efficient for large messages because it requires two fewer memory-to-memory copies at the user/OS interface. NORMA IPC latency was measured at approximately 800 µs for messages of minimum size. The maximum bandwidth using out-of-line transfers and large packets was approximately 11.0 Mbytes/s. Performance discontinuities at page boundaries, most notable with in-line transfers because of the additional copy in/outs, indicate significant overhead in interactions with the VM system. The change from 4 Kbytes VM pages in R1.0C, to 8K pages in T10, almost doubled NORMA IPC bandwidth, but this still is only 7% of the peak hardware bandwidth.

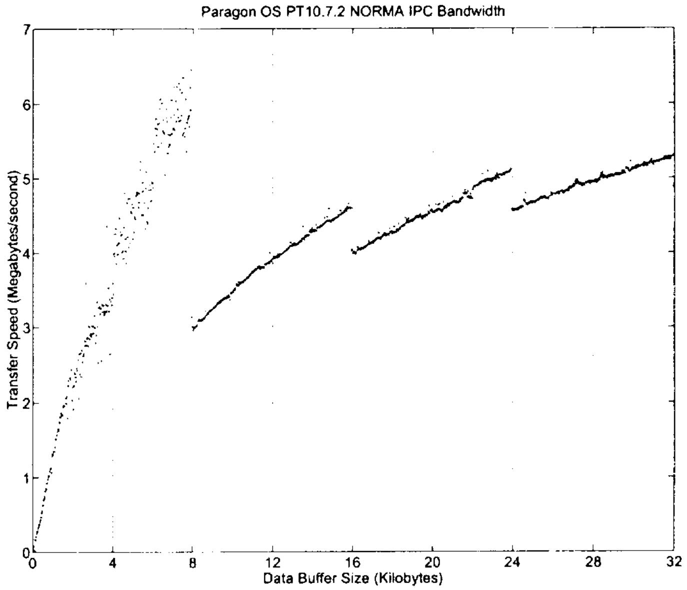


Figure 7. NORMA IPC Bandwidth

Table 7. NORMA IPC Peak Bandwidth and Latency

OS Releases	R1.0C	T10	T10
	in-line	in-line	out-of-line
Bandwidth (MB/s)	4.0	7.5	11
Latency (µs)	800	800	800

3.4.3 Virtual Memory Performance

A test program was implemented to measure and evaluate the Paragon virtual memory (VM) system. The test allocated a chunk of memory and iterated through it twice, first writing one byte to each page, then reading it back. The test was run on a 16 node partition with memory sizes ranging from 1 to 2800 pages. A significant increase in execution time occurs when the memory allocated requires the unused part of the server to be paged out. This first increase occurred at about 8 Mbytes for R1.0C and 7 Mbytes for T10. A second increase in execution time occurs when the application has paged out all unused portions of the server and began paging its unused pages. When the application started page itself, it would often hang as pages needed to continue execution were likely to have been paged out. This second increase occurred around 11 Mbytes for R1.0C and 10 Mbytes for T10. The increased size of the server on T10 was due to added functionality (i.e. PFS, HiPPI).

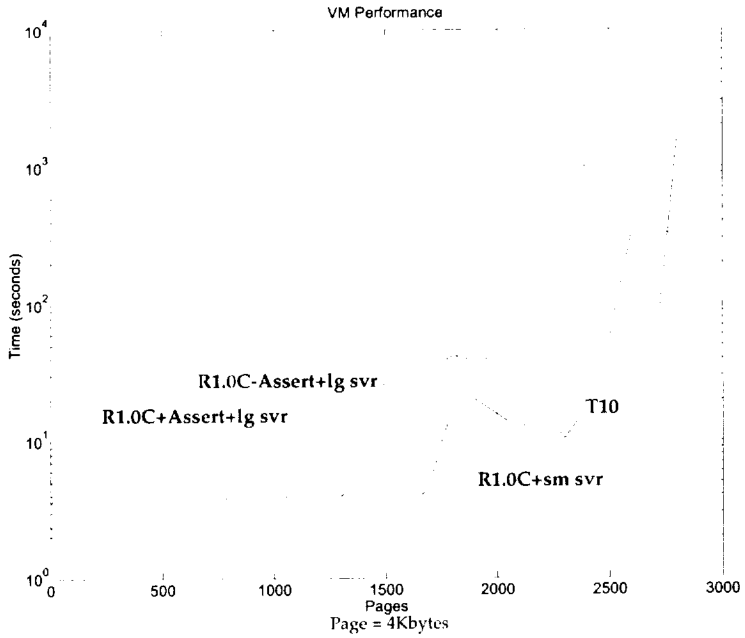


Figure 8. VM Performance

3.4.4 NAS Parallel Benchmarks

EP was the only NAS Parallel Benchmark that was initially able to run on the delivered system. The EP benchmark was run on successive OS releases and system upgrades to monitor performance improvement. A record was kept of the timing for the second run of EP to monitor virtual memory paging-in. The second run-time measured the execution time without paging activities as the full EP application fitted in core memory. A significant difference was observed in the OS releases (T8, T9.2 and T9.3) between the first and second timings. The difference between the first and second run timing approximated the time to page-in the application. Improvements in the page-in overhead was observed after the OS release T8 due to improvements in the NORMA IPC bandwidth and a binary-tree paging algorithm. On T10, the overhead to page-in an application was negligible.

Table 8. EP Timing

Size:	2**28				2**30			
Node	128 (Secs)	128[a] (Secs)	208 (Secs)	208[a] (Secs)	128 (Secs)	128[a] (Secs)	208 (Secs)	208[a] (Secs)
iPSC/860								
NX 3.3.1	31.1				124.3			
Paragon								
T8	268.1	23.3	473.7	13.8	262.9	93.0	423.4	57.0
T9.2	30.0	23.3	20.7	13.8	99.8	93.2	63.6	56.8
T9.3	28.5	23.3	19.2	13.8	95.5	93.2	62.1	56.8
T9.5	20.35		12.87					
R1.0C	20.31	20.21	12.74	12.63	80.93	80.82	50.63	50.52
T10	19.72	19.71	12.33	12.31				

[a] second run of timed DO loop (no paging activity).

All NAS parallel benchmarks ran between 1 and 2 times faster than on the iPSC/860 after the OS release T10. Comparison with published iPSC/860 [7] and CRAY Y-MP/1 [7] timings are provided in the following table.

Table 9. NAS Parallel Benchmarks

Benchmarks	Problem Size	64 Nodes (Secs)	128 Nodes (Secs)	208 Nodes (Secs)	128 Nodes Ratio to iPSC/i80	128 Nodes Ratio to Y-MP/1
EP	2^{28}	39.45	19.72	12.33	1.30	6.40
MG	256^3		6.52		1.32	3.41
FFT	256^2x128	15.29	6.36		1.52	3.88
CG	2.0x10^6		6.75		1.27	1.76
IS	2^{23}	15.43	13.50		1.01	0.85
APPSP	64^3	445.12	281.27		1.59	1.67
APPBT	64^3	475.96	266.58		1.56	2.97
APPLU	64^3	682.00	442.00	326.00	1.01	0.73

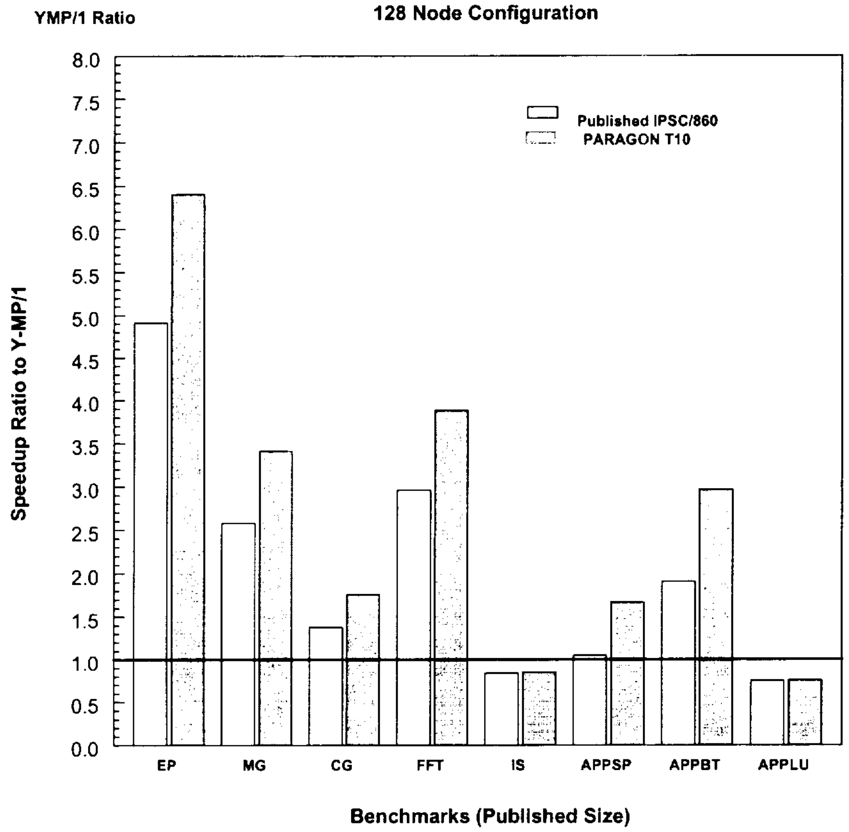


Figure 9. NAS Parallel Benchmarks iPSC/860 vs. Paragon

In Figure 9, the relatively low performance (1.01 times faster than the iPSC/860) for the APPLU and IS benchmarks was due to the high NX message-passing communication latency measured on the Paragon.

4.0 Conclusions

Over the period covered in this report (February 1993 to August 1993), the Paragon average uptime was about fifty percent. Reboots averaged four per day. Single-user availability was found satisfactory, but multi-user availability was precarious. Partition scheduling was unreliable and processes were occasionally left hanging, deadlocking critical resources.

On average, there were three users logged on the Paragon. These users, were predominantly developers and testers; not scientists. Functionality improved from running one NAS Parallel Benchmark (EP), to running all the NAS Parallel Benchmarks on a 128 node partition. At the end of August, all of the NAS benchmarks ran faster (between 1 and 2 times) than on the iPSC/860 on the latest installed T10 OS release.

Although progress was observed, it is our concern that fundamental limitations of the Paragon operating system may impair or restrict further progress. The NORMA IPC bandwidth was measured at 11 Mbytes/s (7% of the peak hardware performance), limiting overall virtual memory, Unix services (i.e. Disk and HiPPI I/O) and the Parallel File System performance. The NX application message passing bandwidth was measured at 28 Mbytes/s, and latency at 184 µs. The high NX latency (three times than on the iPSC/860) was found to significantly degrade performance of applications relying on small message sizes. The amount of memory available for user applications was found to be approximately 10 Mbytes per node, indicating that the Paragon OS is taking more space than anticipated (6 Mbytes per 16 Mbytes node).

While significant improvements were observed in the availability and performance metrics, the previously mentioned Paragon OS limitations were limiting the number of users and the Paragon was unable to support the NAS workload. It is our belief that until considerable redesign and re-implementation of fundamental parts of the Paragon OS (e.g. NORMA IPC) are done, overall functionality and performance will be limited.

5.0 Acknowledgments

The authors wish to thank Russel Carter, Toby Harness and Leigh Ann Tanner of NAS and Ed Kushner and Thanh Phung of Intel SSD for their assistance and comments.

6.0 References

[1] Numerical Aerodynamic Simulation Program Plan. NAS Systems Division, Ames Research Center, October 1988.

[2] Intel Paragon XP/S User Guide, *Intel Supercomputer Systems Division,* April 1993, Order Number: 312489-001.

[3] "An OSF/1 Unix for Massively Parallel Multicomputers", R. Zajcew et al., *in Procedings of the 1993 Winter USENIX Conference,* January 1993, pp. 37-55.

[4] "OSF Mach: Kernel Principles", K. Loepere, Open Software Foundation and Carnegie Mellon University, February 1993.

[5] "The NAS Parallel Benchmarks", D. Bailey et al., NASA Technical Memorandum 103863, *NASA Ames Research Center,* July 1993.

[6] "NHT-1 I/O Benchmarks", R. Carter et al., NAS Report RND-92-016, *NASA Ames Research Center,* November 1992.

[7] "NAS Parallel Benchmark Results", D. Bailey et al., NAS Report RNR-92-002, *NASA Ames Research Center,* February 1993.

[8] "The Art of Computer Systems Performance Analysis", Raj Jain, *Wiley Professional Computing,*1991.

	RND TECHNICAL REPORT
	Title: "Evaluation Metrics for the Intel Paragon XP/S-15"
	Author(s): Bernard Traversat, David M^{c}Nab, Bill Nitzberg and Sam Fineberg
	Reviewers: "I have carefully and thoroughly reviewed this technical report. I have worked with the author(s) to ensure clarity of presentation and technical accuracy. I take personal responsibility for the quality of this document."
Two reviewers must sign.	Signed: ______________________________ Name: Russell Carter Signed: ______________________________ Name: Jeff Becker
After approval, assign RND TR number.	Branch Chief: Approved: ______________________________
Date:	TR Number: