# 4-Node SV1 Cluster Implementation

*T. Jones, B. Sarnowska,* Logicon Information Systems and Services, Inc.,
Building 1001, RM 101, Stennis Space Center , Mississippi, USA

tljones@navo.hpc.mil sarnowsk@navo.hpc.mil

with

*D. Cole*, NAVOCEANO Major Shared Resource Center
Naval Oceanographic Office Code N7, Stennis Space Center, Mississippi, USA
cole@navo.hpc.mil

**ABSTRACT:** *In September, 1999, the NAVOCEANO MSRC became the first Domestic U.S. site to place a multi-node SV1 cluster into production. In June, 2000, the system was upgraded to a four-node configuration. This paper will present NAVOCEANO MSRC's experiences with a multiple-node SV1 deployed in a cluster configuration. It will discuss the system configuration achieved to support diverse user workloads, integration of the SV1 Cluster product in a large-scale HPC environment, and measured performance.*

## 1 INTRODUCTION

In September, 1999, the NAVOCEANO Major Shared Resource Center (MSRC) became the first site in the United States and the second in the world to install a multiple node Cray SV1 cluster. Consisting of two nodes, the SV1-2 was deployed as a capacity cluster to absorb the workload from a Cray C916. In June, 2000, the SV1-2 cluster was upgraded to an SV1-4 Supercluster, serial 3502, the second of its kind, to replace a T932/20 system. Performance and stability of both SV1 clusters have been excellent. Integration and deployment of the SV1-4 was likewise troublefree. Challenges surrounding the Rev A CPUs and initial field availability of the SV1-2 cluster were relatively minor and overcome in a straight-forward manner.

### *1.1 NAVOCEANO MSRC Mission*

The NAVOCEANO MSRC Mission is to acquire, maintain, and support cutting-edge High Performance Computing (HPC) capabilities for use by DoD scientists and engineers. HPC plays a major role in the ability of the United States to develop and deploy superior weapons, warfighting capabilities, and mission support systems. HPC technology improves the performance of systems analysis, design, development, test, and deployment; helps avoid environmental damage; and improves the integration and effectiveness of complex weapons systems. High-fidelity modeling and simulation techniques are also being used to explore more design options and identify important testing priorities, at a fraction of the cost and time of utilizing traditional, theoretical, experimental, or operational methods.

Production versions of new vector and scalable software by application development teams have further expanded the use of HPC for DoD scientists and engineers. New high-speed, high-bandwidth network connections have enhanced network services, bringing us closer to a true meta-computing environment extending to the remote scientist's desktop.

These enhanced capabilities have increased the demand for NAVOCEANO DoD MSRC computational resources, necessitating the continued close management and prioritization of HPC assets across the user community.

### *1.2 NAVOCEANO MSRC Computational Assets*

The NAVOCEANO MSRC has consistently ranked among the top 10 most powerful computer installations in the world. Installed within are several of the worlds most powerful, individual systems[1], each carefully selected and implemented to achieve the mission statement. The major installed

| SYSTEM | PEAK GFLOPS | CPUs (Nodes) | DISK (GB) | MEMORY (GB) | TOP 500 Rank[1] |
|---|---|---|---|---|---|
| IBM RS6000/SP | 2000.0 | 1336 (334) | 17,750 | 1,336 | 4 |
| CRAY T3E-900 | 980.0 | 1,088 | 1,316 | 404 | 13 |
| SGI ORIGN 2000 | 113.9 | 256 | 900 | 128 | 169 |
| CRAY SV1-4 | 64.0 | 64 | 1,576 | 96 | NR |
| SUN HPC E10K | 51.2 | 64 | 1,440 | 64 | 404-470 |

**TABLE 1-1**
**NAVOCEANO MSRC Computational Resources**

computational assets are listed in Table 1-1, NAVOCEANO MSRC Computational Assets.

### *1.3 Background*

The SV1-4 system was initially installed as a 2-node (SV1-2) cluster in 1999, replacing a C916/16-1024 system delivered in early 1995. This system was one of the first computational assets of the NAVOCEANO MSRC and had reached the end of its life-cycle. Increasing maintenance costs and the vendor's plans to discontinue support for Model E IOS systems provided the impetus to plan for an upgrade. A mission-critical and continuing investment in vector codes by segments of the user community, plus a commitment to provide a full-spectrum, balanced center architecture to the user community indicated the need to provide a compatible system to replace the C916.

In 2000, the maintenance and cost issues surfaced with the T932 system installed in 1998. With the replacement of the C916 by the SV1-2 cluster, the SV1-2 had assumed mission-critical support for NAVOCEANO near-real-time operational programs. Support for these programs was required to continue on the upgraded SV1 vector cluster. Moreover, due to the near-real-time nature of some of these applications, interruptions in service during the upgrade process were required to be minimal and brief.

A workload analysis was conducted which indicated that 2 additional computational nodes, each with 16 SV1 CPUs and 16 GB of memory would meet the existing allocated T90 workload as well as continue to support existing applications, including the near-real-time workload. The combined enhancement within both the Rev C CPUs and Type-N Memory was anticipated to a minor performance improvement (5% to 10%) depending on the application. As with any aspect of system performance, results were expected to vary depending on the nature and type of applications used. Therefore, these anticipated performance improvements were largely discounted in the workload analysis when the projected capacity of the SV1-4 was determined.

The upgrade to a four-node Cray SV1 supercluster replaced the T932, becoming the sole vector computing resource for the NAVOCEANO MSRC. This was accomplished in a paced transition that began on May 22, 2000 with the delivery of the equipment to upgrade the SV1-2 cluster to become the SV1-4 supercluster. On June 29, 2000, the system was accepted and all resources were made available to the user community. Between June 30 and September 30, 2000, T90 users were provided with unallocated (free) access to the SV1-4 cluster to migrate their applications and environments. On October 1, 2000, the T90 system was retired and removed from the NAVOCEANO MSRC.

### *1.4 System Description*

The NAVOCEANO MSRC SV1-4/64-12288 Vector Cluster consists of four nodes, each with 16 SV1 CPUs. Two of the four nodes are configured

with Revision B CPUs and 16 Gigabytes (GB) of 70 ns memory and the remaining two nodes are configured with the Revision C CPUs and 32 GB of 50 ns (Type-N) DRAM memory. The Revision B CPUs were the first generation of full-feature with full-performance (vector cache enabled) SV1 processors. The Revision C CPUs contain a modification to the cache memory circuitry to improve cache performance for certain instruction sequences. The 32 GB memory in the Revision C nodes reflect SV1 product improvements over the earlier Revision B nodes.

## 2 SV1 Architecture and Performance Discussion

Cray vector systems are register-oriented machines. There are no operations which act directly on data stored in memory. All operands, vector and scalar, must be loaded into either the scalar or vector registers before operations can act upon them. The result is also placed into a register, which must then be stored back to memory. Because of this architectural aspect, one area of code optimization focuses on subsequent reuse of the results from previous computations, still resident in the registers. This is done to avoid memory latencies and bottlenecks such as bank conflicts on loads and stores.

In the architecture of the original Cray vector systems (Cray-1, X/MP, Y-MP, C90, T90), all operands were fetched directly from memory into the registers for all operations. The Cray J90 architecture provided Cray functionality and supercomputer performance at a lower price-performance point than the traditional Cray vector systems. System performance was lower than traditional systems; however, the overall cost of the system was several orders of magnitude less. One of the ways this lower cost was achieved was through the use of lower-cost memory technologies (DRAM). As is often the case, with the lower cost memory also came a reduced performance. As memory is the critical resource in a traditional vector supercomputer, this trade-off is a significant limitation in overall system performance. To offset this limitation, a 128-word cache was added to the J90 architecture for scalar data transferred between memory and the scalar registers.

Each J90se CPU produced two floating-point results (one add, one multiply) and one scalar result each 10 nanosecond (ns) cycle (100 MHz). This equates to a peak performance of 200 MFLOPS. Memory bandwidth for the J90se system depended on the chassis. The J916 cabinet used a 4 X 4 memory “backplane” switch, with a peak bandwidth of 25.6 GB/second. The J932 cabinet used an 8 X 8 “backplane” switch, which provided a total memory bandwidth of 51.2 GB/second. By contrast, a Cray C916 system with 8 sections of memory had a total memory bandwidth of 245.8 GB/second and a CPU cycle time of 4.2 ns (1 GFLOP).

The SV-1 was originally named the J90+ and represented the third-stage in the J90 product life-cycle. When Cray Research, Inc. was acquired by SGI, the J90 product line was realigned and brought out as the successor to the discontinued C90 and T90 systems. The new architecture was marketed as a low-cost, vector-cluster system, upward-compatible with the successful J90se but with C90 and T90 performance. The key to reaching T90 performance is through utilization of two new special CPU features and clustering.

The two new features were the Multi-Streaming Processor (MSP) and vector data cache. The MSP feature provided a means of closely coupling four SV1 CPUs to form a single processor capable of processing four simultaneous streams. The performance for the J90 CPU was enhanced by doubling the number of floating-point results from two to four per clock cycle. The clock frequency was increased from 100 MHz (10 ns) to 300 MHz (3.33 ns). This boosted the total theoretical peak performance from 200 MFLOPS to 1.2 GFLOPS; a six-times speedup. This rating placed the SV1 CPU at 200 MFLOPS faster than the C90 CPU (1.0 GFLOPS) and 600 MFLOPS below the T90 CPU (1.8 GFLOPS). Through only a slight increase in parallelism, it appeared possible to replace a far more costly T90 platform with an SV1 Cluster.

### *2.1 Theoretical System Performance Analysis*

Using theoretical peak performance when sizing replacement systems is guaranteed to result in a design which is undersized for assuming the original system’s workload. Developing a reasonable anticipated performance (delivered performance) estimate requires a deeper look into the candidate system’s architecture. Only after this is completed can system sizing and proposed configurations be developed. In

the case of the NAVOCEANO MSRC SV1, the system was sized to assume the workload of the C90 and later, the T90. Therefore, an architectural performance and capacity analysis of three systems was required. Upon examination, it became obvious that the most significant architectural limitation to the SV1's capacity and performance envelope was the retention of the J90se memory subsystem. To begin, the memory bandwidths for these systems are compared in Table 2-1, CRAY Vector Systems Memory Bandwidths.

The memory constraint in the SV1 architecture is immediately apparent. The J932se memory bandwidth is evenly divided among all CPUs in the system. There are four CPUs per module, and each CPU receives 1/4th of the memory bandwidth. The memory bandwidth was 6.4 GB/sec for each module, resulting in each CPU receiving 1.6 GB/sec (0.25 x 6.4 GB/sec). This makes the memory subsystem for the J90se balanced for the architecture.

| System | CPU Cycle Time (ns - MHz) | Floating Point Results/Cycle | CPU Peak Theoretical Performance (MFLOPS) | CPU to Memory Bandwidth (GB/sec) | Memory to CPU Bandwidth (GB/sec) |
|---|---|---|---|---|---|
| C916 | 4.0 ns - 250 MHz | 4 | 1,000 | 8.0 | 11.5 |
| T932 | 2.22 ns - 450 MHz | 4 | 1,800 | 14.4 | 21.6 |
| J932 | 10.0 ns - 100 MHz | 2 | 200 | 1.6 | 1.6 |
| SV1 | 3.33 ns - 300 MHz | 4 | 1,200 | 9.6 | 6.4 |

**TABLE 2-1**
**CRAY Vector Systems Memory Bandwidth**

The column labeled "CPU to Memory Bandwidth" lists the bandwidth required by each type of system CPU to sustain full performance of that processor. The column labeled "Memory to CPU Bandwidth" lists the maximum rate at which the memory subsystem can feed the system's CPU. The memory to CPU bandwidth is the memory subsystem's main performance rating. To be in balance with the processor performance, it must be at least the same speed as the CPU to memory bandwidth rating. Further, total aggregate bandwidth of the memory subsystem must be at least the sum of the number of processors in the chassis times the CPU to Memory bandwidth. If it is less, system performance will be degraded for all processors. The system's theoretical performance rating then falls to the maximum rate at which the CPUs can transfer data between themselves and memory. Table 1-2, CRAY Memory Bandwidth Analysis compares traditional Cray vector systems and presents a memory bandwidth analysis of each.

The SV1 uses the exact same memory subsystem as the J90se. This subsystem was retained for the upgraded J90+ design (i.e. SV1) to improve the price/performance of the new system. However, the enhanced performance of the CPUs, essentially 6 times more performance than the J90se CPUs, requires significantly more aggregate memory bandwidth than the 51.2 GB/second that is available from the J90se-vintage 8-by-8 non-blocking crossbar switch. [2]

In the traditional Cray vector systems, memory bandwidth and CPU performance were kept in balance through the use of high-performance, and high-cost memory technologies. A recent estimate placed the increase in processor speeds at 80% per year while the speed of memory devices has been increasing at a rate of 7% per year. [3] For Cray systems, this became evident in the architecture of the T90, where the original CM02 memory (1995) did not provide sufficient bandwidth or adequate CPU/mem-

ory ratios to support a fully configured T932 at maximum performance. Successive generations of memory were released (CM03 - 1996, and CM04 - 1997) to increase the memory bandwidth and capacity. [4] However, the T90 was high-end technology and the J90 family was designed from the onset to a far lower price/performance point. Lower cost DRAM memory was deployed in these architectures and the SV1 was designed to continue this trend. Therefore, instead of redesigning the memory subsystem, which would have added further to the design cycle and the system cost, a vector cache was added alongside the scalar cache to attempt to match the higher processor performance to the J90 memory subsystem performance.

| System | Number of CPUs | CPU to Memory Bandwidth (GB/sec) | Aggregate Required Memory Bandwidth (GB/Sec) | Memory to CPU Bandwidth (GB/sec) | Aggregate Memory to CPU Bandwidth (GB/Sec) | Memory Subsystem Peak Bandwidth (GB/Sec) |
|---|---|---|---|---|---|---|
| C916 | 16 | 8.0 | 128.0 | 11.5 | 184.0 | 245.8 |
| T932 | 32 | 14.4 | 460.8 | 21.6 | 691.2 | 800+ |
| J932se | 32 | 1.6 | 51.2 | 1.6 | 51.2 | 51.2 |
| SV1 | 32 | 9.6 | 307.2 | 6.4 | 51.2 | 51.2 |

**TABLE 2-2**

**CRAY Memory Bandwidth Analysis**

The cache subsystem on the Cray SV1 is a 256 KB, 4-way set associative, write-through cache implemented using SRAM technology. The data element length is one word (8 bytes), which creates what can be used as a very long vector register, capable of holding 32K elements. The cache interfaces with the CPU at 9.6 GB/second, the speed necessary to keep the processor fully utilized, and transfers between itself and memory at 6.4 GB/second. This allows for performance matching between the faster CPU and the DRAM memory.

The cache size is optimized for the vector applications. The cache size of one word is significantly different from non-vector systems. Non-vector systems have cache lines much longer than a single data element. Non-vector system caches are designed with the principle of locality of reference in mind. Locality of reference simply means that if a data element fetched into cache is used, the probability is high that the data elements adjacent or near to the fetched data element will be used in the near future. However, gather/scatter and strided loads comprise a very important portion of vector operations. Having a one-word cache line length allows for these operations without incurring the overhead of loading unneeded data. The SV1 cache subsystem allows the flexibility to prefetch up to seven surrounding words from memory for certain scalar operations. Further the SV1 cache guarantees memory coherency through its write-through and write-allocate design where data elements are always placed in both the cache and memory. The one word cache line size couples with this feature to aid coherency in that no read/modify/write operations are required. [5]

### *2.1.1 Standard Streaming Benchmarks*

Architectural analysis shows the theoretical limits of the system; however, practical limitations and real-world workloads combine to produce less than theoretical results. One method of gauging the anticipated delivered performance of a new system is through benchmarks. Since memory bandwidth is the constraining factor in the SV1 architecture, we look to results from the standardized STREAM benchmark. STREAM is a synthetic benchmark, designed specifically to determine memory bandwidth performance. It measures the performance of the four fundamental long vector operations listed in Table 2-3, STREAM Benchmark Kernels

In STREAM, the array sizes are defined so that each array is larger than the cache of the system being

tested, and the code is structured so that data reuse is not possible[3]. The standard results for the C90, T90, J90se and SV1 systems are available in [6] and are presented in Table 2-4, Cray Vector Systems Standard STREAM TRIAD Benchmark Results.

Machine balance is defined as the ratio of the number of memory operations per CPU cycle to the number of floating-point operations per cycle for a given processor. To overcome systematic bias, long, uncached vector operands with unit stride are used instead of the theoretical peak memory performance.[7] The resulting ratio indicates a relative measure of memory/CPU performance balance. A machine balance of 1.0 represents the optimal design and function of the memory subsystem.

As shown in Table 2-4, the Cray C916/16 exhibits a machine balance of 1.2, and reaches the highest percentage of efficiency in terms of Peak GFLOPS attained, 56.3%. This is the best of all the system listed. This matches well with the theoretical analysis performed above. The SV1-1/32 system, with its C90+ theoretical CPU performance and its J90se memory subsystem reaches only a little over 2 GW per second and 5.4% of its peak GFLOP rating. Since STREAM is written so that there is no reuse of data and larger than cache memory, it is particularly unfriendly to cache memory machines. Therefore the memory performance for the SV1-1/32 in Table 2-5 shows the worst case example with machine balance of 12.4. Optimized code could be expected to perform better; however, many users do not often make the investment to optimize their codes once they are stable and producing correct results. Further, as requirements

| Operation | Kernel | Bytes/Iteration | FLOPS/Iteration |
|---|---|---|---|
| **COPY** | a(i) = b(i) | 16 | 0 |
| **SCALE** | a(i) = q*b(i) | 16 | 1 |
| **SUM** | a(i) = b(i) + c(i) | 24 | 1 |
| **TRIAD** | a(i) = b(i) + q*c(i) | 24 | 2 |

**TABLE 2-3**
**STREAM Benchmark Kernels**

| System | CPUS | Measured Memory Bandwidth (GW/sec) | Measured GFLOPS | Peak GFLOPS | % Peak GFLOPS | Machine Balance |
|---|---|---|---|---|---|---|
| C916 | 16 | 12.98 | 8.65 | 15.36 | 56.3% | 1.2 |
| T932 | 32 | 44.91 | 29.94 | 57.60 | 51.9% | 1.3 |
| J932se | 32 | 2.36 | 1.25 | 6.40 | 19.5% | 2.7 |
| SV1 | 32 | 3.10 | 2.07 | 38.40 | 5.4% | 12.4 |
| SV1 | 16 | 3.05 | 2.03 | 19.20 | 10.6% | 6.3 |

**TABLE 2-4**
**CRAY Vector Systems Standard STREAM TRIAD Benchmark Results**

change for user codes, optimization previously performed may be lost by software or algorithmic changes necessary to support the new required functionality.

The STREAM results for an SV1-1/16 system (Table 2-4) demonstrates improved performance over the 32 processor SV1 system. The measured memory bandwidth and CPU GFLOPS are nearly identical; however, because the CPU count is one-half that of the 32 CPU system, the performance indices are expected to show a 2-times improvement. This is understandable from Table 2-2, CRAY Memory Bandwidth Analysis which lists the total bandwidth requirements as number of CPUs times the memory bandwidth requirement for a single CPU. A trend is therefore suggested in the STREAM performance data which leads to the optimal sizing of a given node. Table 2-5, SV1 Node STREAM TRIAD Standard Results lists the standard STREAM results for the SV1 across incremental CPU counts

| CPUS | Measured Memory Bandwidth (GW/sec) | Measured GFLOPS | Peak GFLOPS | % Peak GFLOPS | Machine Balance |
|---|---|---|---|---|---|
| 2 | 0.624 | 0.416 | 2.4 | 17.3% | 3.8 |
| 4 | 1.232 | 0.822 | 4.8 | 17.1% | 3.9 |
| 8 | 2.327 | 1.551 | 9.6 | 16.2% | 4.1 |
| 12 | 2.800 | 1.868 | 14.4 | 13.0% | 5.1 |
| 16 | 3.053 | 2.035 | 19.2 | 10.6% | 6.3 |
| 32 | 3.097 | 2.065 | 38.4 | 5.4% | 12.4 |

**TABLE 2-5**
**SV1 Node STREAM TRIAD Standard Results**

The performance indices of% Peak GFLOPS and Machine Balance are plotted in Graph 1-1, SV1 Performance Indices. The performance envelope of the system makes two decrements, one at 8 CPUs and again at 16 CPUs. While these are understandable in terms of system architecture, the important consideration is to maximize the performance of each node while minimizing the system's physical size, cabinet count, CPU count, etc. With this in mind, it is clear from the data presented in Graph 2-1 SV1 Performance Indices that 16 CPUs per Node is the optimal node configuration using the STREAM Triad performance as governing metric.

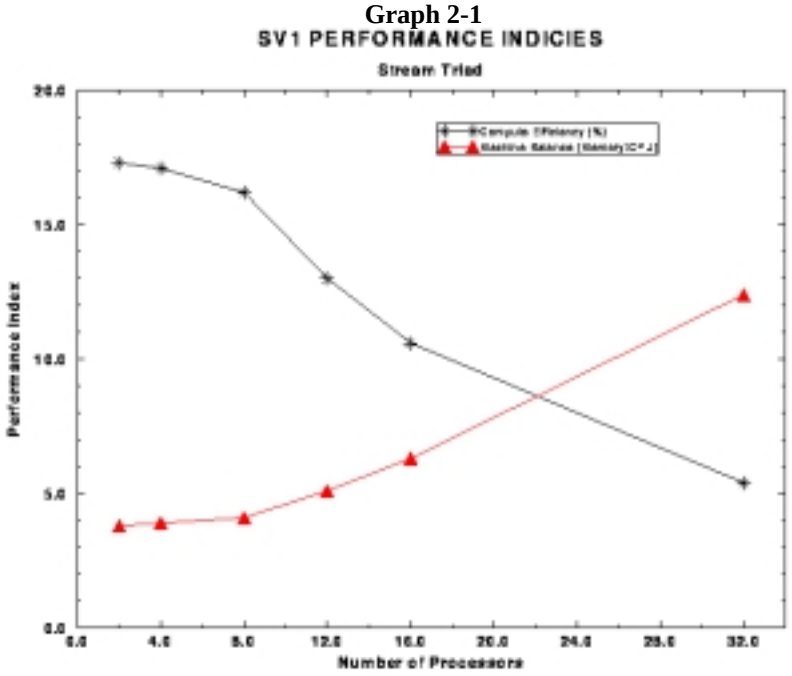
Real-world workloads vary greatly from the analysis in this section. The final test of any configuration planning is the performance of a throughput benchmark which represents the actual workload that the system will support. While theoretical and standard benchmark analysis are sufficient for narrowing down the field of systems and configuration options, the final determiner of the configuration should be a real-world test. NAVOCEANO MSRC conducted a series of real-world benchmark tests using codes which comprised a representative portion of the system workload that the SV1 was to assume. These benchmark tests are described in detail in Section 4.0, Throughput Benchmarks.

## 3 SYSTEM CONFIGURATION

### *3.1 Hardware Configuration*

As described in Section 1.4, System Description, the NAVOCEANO MSRC SV1-4/64-12288 Vector Cluster consists of four nodes, each with 16 SV1 CPUs. Two of the four nodes are configured with Revision B CPUs and 16 Gigabytes (GB) of 70 ns memory and the remaining two nodes are configured with the Revision C CPUs and 32 GB of 50 ns (Type-N) DRAM memory. The four nodes of SV1-4 cluster are named ZEUS, POSEIDON, TRIDENT and ATHENA. The full list of components for the SV1-4 Super-

**Graph 2-1**




| Description | Qty. |
|---|---|
| Cray SV1 4 Node SuperCluster Chassis, Does not include memory or processors | 1 |
| PC-10 Rack Cabinets for GigaRing Peripherals | 4 |
| Cray SV1 Module with four 1.2 GFLOP (Rev C) processors | 8 |
| Cray SV1 Module with four 1.2 GFLOP (Rev B) processors | 8 |
| One 4-Gigabytes main memory module for Cray SV1 Supercomputer Systems. | 16 |
| One 2-Gigabytes main memory module for Cray SV1 Supercomputer Systems. | 16 |
| MPN-1, Multi-Purpose Node. For connection to ATM, Ethernet, FDDI, and SCSI peripherals | 4 |
| HPN-1, HiPPI Channel Node. Contains two 100 Megabyte/second Channels. | 4 |
| NSR-1, I/O Node Subrack. Can hold up to 4 I/O nodes | 4 |
| FDI-10 FDDI Interface, provides single or dual-attach fiber distributed data interchange interface. | 4 |
| ATM-10, ATM Interface for all GigaRing I/O systems. ATM OC-3 performance | 4 |
| ETN-11, 100 Mb/s Ethernet NIC | 8 |
| FCN-1 Fibre Channel Node. Connects to DD/DA-308 and DD/DA 309 disk drives | 8 |
| DA-309 Fibre Channel disk units (80 GB each) | 12 |
| 10 DD-308 Fibre Channel disk units (38 GB each) | 16 |

**TABLE 3-1 SV1-4 Hardware Inventory**

Cluster is listed in Table 3-1, SV1-4 Hardware Inventory

Each node is configured with the resources as listed in Table 3-2, SV1-4 Node Resources

### *3.2 Software Configuration*

The SV1-4 cluster operates under the UNICOS 10.0.0.7 operating system and SWS-ION (Gigaring I/O Software) 6.0. Table 3-3, SV1-4 Software Configuration, lists the software installed on the SV1-4 Cluster

| Resource | Zeus | Poseidon | Trident | Athena |
|---|---|---|---|---|
| **CPUs** | 16 (Rev B) | 16 (Rev C) | 16 (Rev C) | 16 (Rev B) |
| **Memory** | 16 GB (70 ns) | 32 GB (50 ns) | 32 GB (50 ns) | 16 GB (70 ns) |
| **FCN-1** | 2 | 2 | 2 | 2 |
| **DA-308** | 4 (152 GB) | 2 (76 GB) | 2 (76 GB) | 4 (152 GB) |
| **DA-309** | 4 (320 GB) | 4 (320 GB) | 4 (320 GB) | 2 (160 GB) |
| **Total Disk Space** | 472 GB | 396 GB | 396 GB | 312 GB |
| **DSF** | 4 | 5 | 5 | 4 |
| **HPN-1** | 1 | 1 | 1 | 1 |
| **ATM-10** | 1 | 1 | 1 | 1 |
| **FDI-10** | 1 | 1 | 1 | 1 |
| **ETN-11** | 2 | 2 | 2 | 2 |

**TABLE 3-2**
**SV1-4 Node Resources**

#### *3.2.1 Cluster Software Components*

SV1-2 SuperCluster configurations use an OEM-supplied, licensed network software bundle (SuperCluster Bundle). The components of this bundle used for the SV1-4 configuration are:

- **NQE/NQS/NLB:** The Network Queuing Environment with Network Queueing System and Network Load Balancer. Manages the execution of batch jobs and balances the load across all nodes in the cluster. Users submit jobs on one node and the NLB component determines which node will execute the job. The results are returned to the user's current directory
- **NFS (V2):** Network File System Version 2. Bundled with UNICOS Operating System
- **BDSpro:** Bulk Data Services are NFS enhancements for large file transfers.

Software embedded in UNICOS Kernel

- **MPT:** Message Passing Toolkit allows for parallel execution using two or more CPUs located in one or more nodes

#### *3.2.2 Compilers*

The SV1-4 system is operational with Cray Programming Environment (PE) release 3.4 which includes compilers that are specific to the SV1 environment. SV1-specific compilers were issued beginning with the PE 3.4 release. CrayLibs remained

common to all platforms. The decision to create a separate Cray SV1 compiler offered potentially significant benefits for users.

The Cray SV1 hardware and software environments differ from those of other PVP platforms in ways that are significant for compilers. These differences include the presence of a vector cache and the new Multi-Streaming Processor (MSP) configuration option. The compiler code development has been divided into two separate sets of software to better accommodate these differences. This change provides several advantages for SV1 users. For SV1 users, compiler customizations that are system specific can be implemented without restrictions that result from possible degradations that can occur on other platforms. For users of traditional PVP systems, the division allows those compilers to be simpler and more stable via the omission of the SV1-specific components. SV1 users had to re-compile their codes to take advantage of this optimization.

### *3.3 Cluster Configuration*

The NAVOCEANO MSRC SV1-4 system is deployed as a capacity cluster.

A clustered computer system is defined as a collection of multiple machines coupled to each other by networks or other similar interconnects. Each of the machines in a cluster is called a node. A node is comprised of processors and memory and is under the control of a single operating system image. Each individual SV1 mainframe is a single node within an SV1 cluster.

| Description | Version Level |
|---|---|
| UNICOS Operating System | 10.0.0.7 |
| CF90 FORTRAN Programming Environment | 3.4.0.1 |
| C++ Programming Environment | 3.3 |
| Craylibs | 3.4.0.3 |
| Cray Tools | 3.4.0.0 |
| CVT | 3.1 |
| Solaris Operating System for SWS | Solaris 7 |
| SWS I/O Node Software | 6.0 |
| Cluster Bundle for SV1-4 System: | |
| NQE/NQS/NLB Network Queueing Environment with Network Queuing System and Network Load Balancer | 3.3.0.15 |
| MPT Message Passing Toolkit | 1.3.0.2 |
| BDS Bulk Data Services | Incl. in UNICOS S/W |

**TABLE 3-3**
**SV1-4 Software Configuration**

SV1 nodes are coupled within an SV1 cluster by either GigaRings or other high-performance network interconnects such as HIPPI. Clustered computer systems can be software-configured in several ways depending on requirements. SV1 software architecture best supports a capacity cluster, where the management and allocation of resources is performed across the entire cluster but where each application runs entirely within a single node. The OEM's product name for this configuration is the SV1 SuperCluster. The use of parallel programming models such as Message

Passing Interface (MPI) allows the scheduling of application processes in several nodes of a cluster. This is called a capability cluster. [8] The software architecture for the SV1 does not adequately address the tightly coupled resource allocation and management required to provide satisfactory performance on a consistent and automatic basis.

Each node operates under identical versions of all software. The nodes have separate instances of the UNICOS operating system and of the Network Queueing System (NQS) queuing system. They share one /home file system mounted from one SV1 node (SV1-1/16-16). All nodes, however, have separate /tmp file systems.

### *3.4 Workload Categories*

The NAVOCEANO MSRC SV1-4 cluster is required to support several diverse workload categories within its configuration. As is often the case, these workload categories are often diametrically opposed in their requirements. Accommodation of diverse workloads in a single system can become a major challenge; however, in a cluster configuration, significantly greater flexibility is inherent to overcome these challenges.

Specifically, the NAVOCEANO MSRC must support the following workload categories:

- Allocated Batch Usage: These are projects which have been allocated a fixed number of CPU hours within a given year. Allocations must be renewed each year. This category represents the majority of the workload
- Operational Production: NAVOCEANO-specific, mission critical workload that occurs in a daily production cycle. These create production products which have a short shelf-life. Processing must be completed within a specified window of execution.
- Challenge Projects: Emphasized, large-scale projects which stretches the limits of computational technology. Challenge projects are accepted after a peer-review of proposals. Allocations must be renewed each year.
- Non-Allocated Usage: Background usage by allocated projects for which no CPU hours are charged against their allocations. This is available to fill slack periods with codes when system capacity exceeds allocated user demand.

#### *3.4.1 SV1-4 Cluster Deployment*

A great deal of the expertise used in the deployment of the SV1-4 cluster came from experiences with an SGI Power Challenge Array (PCA). Functionally, PCA clusters greatly resembles the SV1 cluster configuration. It consisted of single nodes, no intrinsic global filesystem, no visibility of hardware resources between nodes, and no synchronization of software resources among nodes for multi-node applications. The same techniques were used to create a cluster using SGI PCAs as are used to create an SV1 SuperCluster.

##### *3.4.1.1 Global File System*

Node ZEUS is the primary login node for the SV1-4 cluster and file-server for /u/home within the cluster. User data residing in /u/home on the SV1-4's zeus node is mounted across fast-ethernet (ETN-11) on all other nodes in the cluster using Network Filesystem (NFS) Version 2. This provides a global filesystem space, consistent across all nodes. A common, global filesystem for a cluster is absolutely essential to users. As the nodes do not support hardware resource access by other nodes in the cluster, the three options for providing a global filesystem are:

- NFS V2: Intrinsic to UNICOS O/S. Lower performance than other options, no added cost, simple to implement, very low risk
- NFS V3 (ONC+): Extra cost license fee. Highest performance NFS. Compatibility questions with other platforms in MSRC, simple to implement, moderate risk

- DCE/DFS: Significantly increased complexity. Additional license cost, DCE core services not supported on Cray platforms, requires additional server and license for core services, performance and stability in SV1 environment uncertain, high risk

The SV1 processors initially represented a new architecture to the NAVOCEANO MSRC. Further, since the system was to be the first to be deployed in the USA, and one of the very first in the world, a higher level of risk was associated with the new CPUs which had yet to establish a field pedigree. In consideration of this higher risk, it was deemed prudent to minimize risk by using the low risk NFS V2. Performance was a concern; however, the NFS V2 software was determined to be sufficient to support both the SV1-2 and SV1-4 configurations.

*3.4.1.2 Scratch Work Space*

Scratch workspace is allocated locally on each node. As UNICOS does not support shared disk access across Gigaring, each node is required to have its own scratch space. While this leads to fragmentation and some inefficiencies in the use of disk space, it has not yet presented any performance or capacity problems.

*3.4.1.3 Batch Subsystem/Workload Category Accommodation*

The batch system on the SV1-4 system is the Network Queuing Environment (NQE), which manages the Network Queuing System (NQS) and the Network Load Balancer (NLB) which manages the execution of batch jobs and balances the load across all nodes in the cluster. The SV1-4 Batch Queue structure, has been constructed to provide the MSRC user community with a uniform queue structure across all HPC systems in the NAVOCEANO MSRC. This uniform queue structure consists of four queues, one to match each workload category. The queues are listed in Table 3-4, Batch Queues Description

Interactive access to all four nodes is permitted and NQS controlled batch jobs may be submitted from any of the four nodes. For allocated batch usage (“batch” queue), users submit batch work to NQE using "qsub" alone and their jobs are scheduled by the NLB software (Network Load Balancer) which routes the batch jobs to the least loaded (CPU Utilization) node in the cluster at the time that the job has been submitted. All other workload category jobs are submitted directly to the corresponding queues on the destination node.

Since jobs may be routed to and run on any of the four nodes, users must assure that all needed input data and executable files are transferred to the /tmp filesystem on the node where their job begins execution. To permit NLB to run a job on any of the four nodes, users must have a list of all four nodes in their .rhosts file in their home directory.

To facilitate batch data transfers, users may use the non-kerberized rcp and remsh commands from any SV1-4 node to the mass storage system, MSAS1.

NQE/NLB could not prevent users from submitting their jobs directly to a batch queue on any node. As a result, one node could be overloaded with jobs (saturated) while at the same time the other node would be un-used; the NLB could not do its jobs since not all jobs were directed to queues by NLB. A local wrapper of the qsub command solved this problem. It prevents users from submitting directly to any batch queue.

The SV1-4 Batch Queue structure has been simplified to four queues. The batch and background queues are available on all four nodes. The priority queues, available only on nodes TRIDENT and ATHENA, are restricted for Challenge support. The internal queue is available only on node POSEIDON for the WSC users. Users can run jobs up to 24 hrs and 512 MW of memory. The queue limits are set based on the center workload.

## 4 THROUGPUT BENCHMARKS

### *4.1 Benchmark Description*

The C916 system was required to support production operational oceanographic models which ran daily within specific time windows. The proposed SV1 system had to be able to provide the same throughput for these models on a daily basis. Three codes were identified by NAVOCEANO MSRC management as representative of the C916 required workload. Together, these three codes represented 39.8% of the total C916 workload. Table 4-1, SV1-2 Throughput Demonstration Applications lists these applications.

The SWAFS code can perform ocean wave simulations of a wide range of geographic areas; however, the benchmark developed operates on only one region, the Pacific ocean. The Yellow Sea code is an old version of the SWAFS code, specifically targeted to the Yellow Sea geographic area. The WAM model is a production code which also simulates ocean waves in geographic regions.

Execution of this benchmark is controlled with job priorities, process priorities, and queue run limits. Directly submitted to individual node's batch subsystems allows an even distribution of the jobs between nodes. Job and process priorities, and the number of CPUs allocated to each job, are presented in Table 4-2, SV1 Benchmark Job Priorities and CPUs Utilized.

| Workload Category | NQS Queue | Comments |
|---|---|---|
| Allocated Batch Usage | batch | Load Balanced across all nodes, unrestricted access |
| Operational Production | internal | restricted access, primary and backup nodes, not load balanced |
| Challenge Projects | priority | Restricted access, primary and backup nodes, not load balanced |
| Non-Allocated Usage | background | All nodes by primary internal queue node |

**TABLE 3-4**
**Batch Queues Description**

| Model | Description |
|---|---|
| WAM | Operational Wave Model |
| SWAFS | Operational Princeton Wave Model |
| Yellow Sea POM | Princeton Operational Model for Yellow Sea Region |

**TABLE 4-1**
**SV1-2 Throughput Demonstration Applications**

| Model | Batch Priority | Process Priority | Number of CPUs |
|---|---|---|---|
| WAM | 30 | 22 | 6 |
| SWAFS | 20 | 27 | 2 |
| Yellow Sea POM | 25 | 29 | 4 |

**TABLE 4-2**
**SV1 Benchmark Job Priorities and CPUs Utilized**

This benchmark can be run on (conceivably) any number of SV1 nodes. It is assumed that the nodes share a common home directory, but that they have separate tmp directories. These codes run multiple times (per node) during the benchmark.

The C916/SV1 Throughput Benchmark was performed on an SV1-2 (2 nodes, 16 CPUs per node, and

16 Gigabytes (GB) of 70 nanosecond (ns) cycle-time central memory per node comprised of Revision B processors. For brevity, benchmark Old RevB will hereafter be referred to as C916/SV1 Throughput Benchmark within this document. The T932/SV1 Throughput Benchmark was performed on the SV1-4 with the configuration of the respective nodes given in the Table 4-3, SV1-4 Node Configuration.

| NODE NAME | CPUs | MEMORY (GB) | DISK (GB) |
|---|---|---|---|
| ZEUS | 16 (Rev B) | 16 (70 ns) | 472 |
| POSEIDON | 16 (Rev C) | 32 (50 ns) | 396 |
| TRIDENT | 16 (Rev C) | 32 (50 ns) | 396 |
| ATHENA | 16 (Rev B) | 16 (70 ns) | 312 |

**TABLE 4-3**
**SV1-4 Node Configuration**

The SV1 nodes utilized in the T932/SV1 benchmark differ significantly from one another. Two are comprised of sixteen Revision B processors each, and 16 Gbytes of 70 nanosecond (ns) cycle-time central memory. The other two nodes ares comprised of sixteen Revision C processors each, and 32 Gbytes of 50 ns cycle-time central memory. For brevity, benchmarks RevB T932/SV1 and RevC T932/SV1 will hereafter be referred to as New Re vB and Rev C Throughput Benchmark within this document.

The Revision C processors have improved data cache enhancements in the CPU and Type-N memory (faster DRAM memory parts). The Cray, Inc. benchmarking group stated that the SV1 Revision C processors improvements are application dependent; the combined enhancement within both the CPU and Memory can provide a minor improvement (5% to 10%) for some programs.

In addition, the new Revision C and Revision B nodes in the SV1-4 Cluster have upgraded compilers and libraries, Cray Programming Environment (PE) release 3.4. Cray PE 3.4 release includes compilers that are specific to the SV1 environment.

Each node operates under identical versions of all software, shares a common /home file system mounted from one SV1 node and has its own /tmp file system. With the exception of several large input files, all files required to execute the benchmark are located in the shared /home directory. All benchmark jobs are executed on each node's /tmp file system and all resulting output files from the jobs are left intact on these /tmp file systems. This model of job execution provides a one-way flow of data from the shared /home file system to the two /tmp file systems and is consistent with operational environment employed by the user community.

### *4.2 C916/16 to SV1-2 Throughput Benchmark Results*

The benchmark mix prescribed, by way of multitasking and simultaneous execution of multiple copies, 30 processes to busy the C916's sixteen processors. Run-time experiments demonstrated that with I/O, synchronization, SDS transfers, and job start-up and shut-downs, the mix used 15.5 of the C916's sixteen processors on average through the benchmarks's duration. Aggregated, the benchmark mix represents 39.8% of the total C916 workload heavily exercising the I/O subsystem in a manner which incurs idle time. The prescribed mix provided a valid representation not only of the C916's capabilities, but also of the machine's daily workload.

### *4.3 T932/20 to SV1-4 Throughput Benchmark Results*

The T932/SV1 benchmark describes the simultaneous execution of the same benchmark on one NAVO CRAY SV1-1/16-16 (1 node, 16 CPUs, 16 GB memory) node with Revision B processors and on one NAVO CRAY SV1-1/16-32 (1 node, 16 CPU, 32 GB memory) node with Revision C processors. It also gave a comparisons of Revision C processor performance to Revision B processor performance.

### *4.4 Comparison of Benchmarks, Old RevB, New RevB. and RevC*

The following Table 3-4, Benchmark Performance Data, presents condensed performance summaries of September 22, 1999 Old RevB benchmark, and June 2000 New RevB and RevC benchmarks, respectively. The multiple runs of a single code are reduced to averages. For example, nine executions of the Yellow Sea code are reduced to a single set of average

| Code | Memory Mwords | Number CPUs | Executions per node | C916/SV1(b) Benchmark Elapsed Seconds | T932/SV1(b) Benchmark Elapsed Seconds | T932/SV1(c) Benchmark Elapsed Seconds |
|---|---|---|---|---|---|---|
| SWAFS | 372 | 6 | 1 | 5,523 | 5,147 | 4,529 |
| Yellow Sea | 38 | 4 | 9 | 1,638 | 1,738 | 1,525 |
| WAM average | 31 | 2 | 33 | 1,366 | 1,479 | 1,316 |

**TABLE 4-4**
**Benchmark Performance Data**

The three instances of the SV1 benchmark executed in the following elapsed times:

- Old Rev B 5,675 seconds
- New Rev B 5,572 seconds
- Rev C 5,003 seconds

This shows good consistency between past and present Revision B SV1 nodes, and a 10% elapsed time reduction gained with the Revision C nodes. Software upgrades since the Old Rev B benchmark account for some of the performance improvements realized in the New Rev B and Rev C benchmarks. The UNICOS operating system was upgraded from 10.0.0.6 to 10.0.0.7 and the compilers and libraries were upgraded from version 3.3 to 3.4. SV1 specific enhancements were contained in these upgrades which contributed to the performance improvements in the New RevB benchmark.

The impact of the benchmark design and other system software changes resulted in a 1.85% performance improvement, as derived by comparing the Old Rev B benchmark elapsed time with New Rev B benchmark elapsed time.

### *4.5 Percent Changes in Benchmark Performance*

The raw performance numbers, presented in Table 4-5, Percent Changes in Benchmark Performance, illustrate the changes in performance realized between three executions of the benchmarks. They compare the Old RevB results to the New Rev B and Rev C results.

To illustrate the performance changes, the data from Table 4-4, Benchmark Performance Data, repeated in Tables 4-5, Percent Changes in Benchmark Performance, but with all performance numbers recast as relative percent changes. For example, comparing the RevC benchmark with respect to the New RevB benchmark, the relative percent change is defined as: Percent Change = [ Rev C – Rev B ] / Rev C * 100. Negative numbers indicate a performance improvement and are enclosed in parentheses.

The second column of Table 4-5 shows that the New Revision B and Old Revision B nodes performed consistently. The third column shows a significant performance improvement in the Revision C node with respect to the Old Revision B node. The CPU seconds decrease from 8% to 15.5% for each benchmark code. The forth column presents a direct comparison of the Revision B and Revision C nodes. All performance numbers improve upon the Revision C node.

| Code | New Rev B with respect to Old RevB Benchmark Elapsed Seconds | Rev C with respect to Old RevB Benchmark Elapsed Seconds | Rev C with respect to New RevB Benchmark Elapsed Seconds |
|---|---|---|---|
| SWAFS | (7.31) | (21.95) | (13.65) |
| Yellow Sea | 5.75 | (7.45) | (14.00) |
| WAM average | 7.67 | (3.75) | (12.38) |

**TABLE 4-5**
**Percent Changes in Benchmark Performance**

| Benchmark | Elapsed Seconds | % Change from C916/SV1(b) |
|---|---|---|
| Old Rev B | 5,675 | 0.00 |
| New Rev B | 5,572 | 1.85 |
| Rev C | 5,003 | 13.43 |

**TABLE 4-6**
**Percent Changes in Benchmark Performances**

Lastly, Table 4-6, Benchmark Elapsed Time, compares the benchmark elapsed times

For this benchmark suite, the new Revision C node outperforms the old Revision B nodes by 13.43%. The new Revision B node outperforms the old Revision B nodes by nearly 2%.

The T932/SV1 benchmark elapsed time for the Revision B node is 5,572 seconds and 5,003 seconds for the Revision C node. The Revision C node executed the benchmark 11.3% faster than the Revision B node. The two nodes performed precisely the same amount of computational work and I/O during the benchmark. Additionally, as the benchmark codes are well representative of codes employed by NAVO's general user community, the Revision C nodes are capable of performing 11.3% more work on a daily basis than the Revision B nodes.

### *4.6 Benchmark Summary*

The Old Rev B benchmarks required 5,675 elapsed time seconds to complete. The New Rev B benchmarks required 5572 seconds and the Rev C required 5003 seconds to complete. The Revision C node completed the benchmark mix 11.3% faster than the Revision B node.

The New Rev B benchmark results illustrate consistent performance between the Revision B CPU nodes from September 1999 and June 2000. The benchmark mix models were selected and structured to closely model the daily workload performed by the SV1 nodes. Continued improvements in SV1 hardware and software releases will likely exhibit continued reductions in benchmark elapsed time requirements.

These real-world benchmark results, coupled with the workload analysis described on Section 5.1, T932 Workload Analysis, strongly supported the proposed SV1-4 SuperCluster configuration capacity as the replacement system to the T932 and sole vector compute platform at the NAVOCEANO MSRC for current and projected requirements.

## 5 T932 Workload Analysis

As part of the system configuration planning, an analysis of the existing T90 workload was performed

and projected on to a proposed expanded SV1 Cluster to determine if it provided sufficient capacity to support all workload executing on the existing SV1-2 and the T90 systems. Further, the mission-critical support for NAVOCEANO operational programs provided by the existing SV1-2 cluster was required to be continued on the SV1-4, regardless of the workload transferred from the T90.

This analysis indicated that a Cray SV1-4 cluster would support all allocated vector workload at the NAVOCEANO MSRC allowing for the decommissioning of the T932/20 system and a corresponding reduction in operations and maintenance resources. Once the workload was transferred from the T90 to the SV1, the SV1-4 cluster would become the sole vector computing resource for the NAVOCEANO MSRC.

### *5.1 Historical Usage Assessment*

The basis of this assessment was an analysis of the historical usage of the T90 system. Accounting records and the HPCMO monthly utilization reports were used as the sources of data. The data was analyzed for total system utilization. FY1999 and FY2000 were similar with regard to T90 usage. T90 monthly usage across the 17 months is characterized in Table 5-1, T90 Workload Characteristics

All usage of the NAVOCEANO MSRC HPC systems, including the T90, require an allocation of CPU hours assigned by the High Performance Computing Modernization Office (HPCMO) through a yearly renewal process. The T90 and all other MSRC HPC systems employ a scheme wherein users may execute jobs at a significantly reduced priority without charge to their allocations. This provides a means of capturing otherwise unused capacity of the HPC platforms for projects which have exceeded their allocated hours for the year. The non-billable utilization (9%) is included in the total CPU utilization.

| Parameter | Monthly | Weekly |
|---|---|---|
| Average User CPU Hours | 7796 | 1,949 |
| Standard Deviation | 1,287 | 322 |
| % CPU Utilization | 72.5 | 72.5 |

**TABLE 5-1**

**T90 Work load Characteristics**

### *5.2 System Resources*

The SV1 and the T90 system resources are listed in Table 5-2, T90/SV1 System Resources:

The total usable GFLOPS (Giga-Floating Point Operations Per Second) is an estimate based on application performance. The T90 is well balanced in its architecture; a well-tuned application might obtain 75% of peak GFLOPS. Experience gained with optimized codes on the SV1 during the C916/SV1(b) throughput benchmark showed that an application might expect to utilize at most 50% of the peak GFLOPS due to the memory subsystem bandwidth constraints. These are conservative estimates of the efficiency of each architecture.

Because the SV1's has significantly larger central memory resources than the T90, central memory utilization was not a factor in this workload assessment. It was anticipated that memory requirements of the T90 workload can be accommodated by the proposed SV1-4 cluster.

### *5.3 SV1/T90 Equivalency*

Using the T90 workload characteristics, the number of SV1 CPUs required to accommodate the T90 workload, assuming applications that have been reasonably parallelized and optimized for the SV1 architecture is derived from Table 5-3, T90 Utilization Derivation

A utilization of 63.5% on a 16-CPU T90 system translates to 10.16 CPUs being used to perform work charged against project allocations (billable utilization). To approximate the number of SV1 CPUs re-

quired to sustain this workload, the usable GFLOPS ratings for the T90 and SV1 are obtained from Table 4-2, T90/SV1 System Resources. The T90 has a usable GFLOP rating of 24 GFLOPS and the SV1-4 has a usable GFLOP rating of 32 GFLOPS. This translates to 1.5 GFLOPS per T90 CPU and 0.5 GFLOPS per SV1 CPU. This is a 1:3 ratio, meaning that for every 1 T90 CPU, 3 SV1 CPUs would be required to process the same workload. Replacing 10.16 T90 CPUs would require 30.48 SV1 CPUs.

| Resource | T90 | SV1-4 |
|---|---|---|
| **Memory** | 8GB/16GB SSD | 2 Nodes; 16 GB<br>2 Nodes: 32 GB |
| **Disk** | 617 GB | 540 GB/Node |
| **Processor Floating Point Format** | IEEE | Cray Floating Point |
| **CPUs** | 16 (contractual) | 64 (4 nodes, 16 each) |
| **MAX GFLOP/CPU** | 2.0 | 1.0 |
| **Total Max GFLOPS** | 32.0 | 64 (4 nodes, 16 GFLOPS ea.) |
| **Total Usable GFLOPS** | 24 (75% of Peak) | 32 (50% of Peak)<br>(4 nodes, 8 GFLOPS ea.) |

**TABLE 5-2**
**T90/SV1 System Resources**

| | |
|---|---|
| T90 Average User CPU Hours/Week | 1,949 |
| Average Total Utilization of 16-CPU T90 | 72.5% |
| T90 Average Non-billable CPU Hours/Week | 242 |
| Average Billable CPU Hours/Week | 1,707 |
| Average Billable Utilization of 16-CPU T90 | 63.5% |

**TABLE 5-3**
**T90 Utilization Derivation**

### *5.4 Required System Capacity Increment*

From this analysis, The SV1-2/32 cluster would require an additional 32 CPUs to accommodate the T90 billable workload. This conclusion assumed that the additional CPUs would be 100% utilized. T90 Applications must be optimized to be “cache-friendly” and a higher degree of parallelism must be exploited to meet this projection. Single CPU applications will see a significant decrease in throughput (as high as a factor of 3 to 4), which will impact overall SV1-4 cluster throughput.

## 6 Application Conversion/Transition Issues

### *6.1 C90 to SV1-2 Transition*

Transition from the C90 to the SV1-2 was very smooth. A performance reduction of 3 to 1 was noticed for most codes simply recompiled on the SV1. Increased parallelism was discovered to be the most important key to achieving C90-like throughput. Issues associated with the initially-delivered Revision A CPUs were multiple. These were resolved when the processors were replaced by the vendor with Revision B CPUs.

### *6.2 T90 to SV1-4 Transition*

Likewise, the transition from T90 to SV1-4 was very smooth. There were no significant issues related

to application conversion. Users were mostly already aware of the internal floating point format differences between the SV1 and the T90 CPUs. The SV1 series of CPUs uses the Cray Floating Point format for internal representation of floating point numbers. This is identical to the J90se and C916 systems at the NAVO MSRC. It is not directly compatible to the T932, T3E or Origin 2000 systems which use the IEEE 704 Floating Point standard. However, options available on the UNICOS assign statement permit the input and output of floating point numbers in IEEE formats for data interchange between the SV1-2 and these machines. The differences in internal floating point representation used during computations must be considered if users are porting an application to or from one of these machines and the SV1. Simple recompilation of users' codes solved that issue.

### *6.3 SV1-Specific Compilers*

Cray PE 3.4 release included compilers specific to the SV1 environment. Users had to recompile their codes to take advantage of the SV1-optimized opcode sequences generated by these compilers. As shown in Section 3.5, Percent Changes in Benchmark Performance, the new compiler resulted in a performance improvement of about 2%. While this improvement might be discounted, it should be mentioned that it was for the specific codes in the benchmark suite. Further, this is only the initial release of the targeted compiler. Additional improvements are anticipated.

### *6.4 SSD*

C90 and T90 users could use the fast SSD on these systems. The SV1 systems do not support an SSD. However, the architecture provides for very large central memory. For the users of the NAVOCEANO SV1-2 and SV1-4 clusters, simply changing the assign statement from type = ssd to type = mr (memory resident) resolved the issue.

### *6.5 Code Optimization*

The ability to achieve C90 and near-T90 throughput performance rests in the ability to increase the parallelism in applications. Users initially expressed disappointment with this; however, increased parallelism and improved parallel algorithms are consistent with HPCMO program goals.

User codes transitioning from the traditional vector platforms such as the C90 and T90 were optimized in a manner consistent with the vector-register architecture. Many of these codes no longer perform well in the cache-based SV1 vector environment. Users must re-optimize their codes to take advantage of the vector and scalar cache to stream data through the CPUs. Without this optimization, codes will be severely constrained by the raw memory bandwidth. Additional differences in code optimization for the SV1-4 compared to the T90 and C90 systems is that unrolling a DO Loop actually worsens performance of the code.

### *6.6 Multi-Streaming Processors (MSPs)*

As part of the initial testing of the SV1-2 Cluster, Multi-Streaming Processors were evaluated with several kernel codes. These returned impressive results, consistent with the anticipated performance; however, it is not practical to configure MSPs for the NAVOCEANO MSRC environment. MSPs reduce the granularity a 16-CPU node and are manually allocated. This is anticipated to lead to scheduling inefficiencies (excessive idle time). Automatic allocation of MSPs such that one is created when a process asks for an MSP would go a long way to alleviate this issue.

## 7 REVIEW OF ISSUES

The SV1 cluster software presently lacks certain features which would enhance its usability. The most important feature would be greater cluster-awareness on the part of NQE/NQS and the portions of UNICOS which schedule/manage resources. For example, it is impossible to place a limit on the number of jobs a user can run on the entire cluster. The NAVOCEANO MSRC is currently developing a submit-wrapper script for internal use to set and enforce cluster-wide job limits by queue, user, and group, as well as global cluster limits. Further, distributed shell commands, much like dsh under IBM AIX for SP Clusters would be a great aid to system management and user interface.

NLB routes jobs to whichever node is the least loaded at the time that the user submits the job. THis leaves users guessing where their job is executing, forcing them to look on all nodes to find it. A distributed shell command would allow users to issue a "qstat" command to all nodes in one command.

Additionally, the SV1 cluster would benefit highly from a global, parallel I/O filesystem. This would allow for disk resources to be consolidated for common filesystems visible across the entire cluster. This filesystem should not be limited to a single controlling node or element of a node, but be resiliently supported by multiple, independent node resources. Filesystems for user home, scratch, opt, local applications, security, etc. could be more efficiently allocated if such a capability was available.

Finally, the most important architectural issue that must be solved is the memory bandwidth and machine balance.

## 8 SUMMARY

In summary, the SV1-2 and SV1-4 clusters smoothly replaced the C90 and T90 systems at the NAVOCEANO MSRC. Vector users have adjusted well to the new architecture. The memory bandwidth needs improvement, and the software's "cluster-awareness" should be increased. However, the system has successfully met its goals as the replacement system for the higher-cost C90 and T90 systems.

## 9 ACKNOWLEDGEMENTS

The authors would like to recognize the work of Mr. Michael Patterson, Logicon Information Systems and Services (LISS), Inc. Mr. Patterson performed and analyzed the real-world benchmarks described in Section 3, Throughput Benchmarks. The authors also wish to thank Mr. Steven Adamec, Director, NAVOCEANO MSRC, and Dr. Walter Shackelford, Program Manager, LISS.

## 10 REFERENCES

1) H-W Meur, J. Dongarra, E. Strohmaier, *Top 500 Supercomputer Sites*, Univ. of Mannheim, Mannheim, Germany, Univ. of Tennessee, Knoxville, TN, USA, June 2000, http://www.top500.org/

2) *Cray SV1 Boosts Parallel Vector Systems*, D.H. Brown Associates, Inc., Port. Chester, NY, June 16, 1998. http://www.dhbrown.com

3) J. McCalpin, *Sustainable Memory Bandwidth in Current High Performance Computers*, Advanced Systems Division, SGI, Inc., October 12, 1995, //http://home.austin.rr.com/mccalpin/papers/bandwidth/bandwidth.html

4) Silicon Graphics/Cray offers Enhanced Data Storage Capabilities for Cray T90 Vector Supercomputing Series, SGI Australia, Sydney, Australia, November 7, 1997, 1997 Press Release Archive, http://www.sgi.com.au/news/427nov1.html

5) M. Stewart, *Overview of SV1 Vector Cache Architecture,* Cray Inc., National Energy Research Scientific Computing Center (NERSC), Berkeley, CA, September 21, 2000. http://hpcf.nersc.gov/computers/J90/sv1opt1.html

6) Streams Benchmark Web Site: http://www.cs.virginia.edu/stream/, Results: September 15, 2000.

7) J. McCalpin, *A Survey of Memory Bandwidth and Machine Balance in Current High Performance Computers*, Advanced Systems Division, SGI, Inc.,http://home.austin.rr.com/mccalpin/papers/balance/index.html

8) P; Langer, M. Danisch, Reviewers, *CRAY SV1 SuperCluster Administrator's Guide, Draft*, SG-2253 10.0.0.5, Silicon Graphics, Inc. April 1999, Document Number 004-2253-001