Supercomputers: Directions in Technology, Architecture and Applications

By Larry Smarr
Director, NCSA and National Computational Science Alliance
University of Illinois at Urbana-Champaign

Abstract

By using the results of the Top 500 listing over the last five years, one can easily trace out the complete transformation of the U.S. supercomputer industry. In 1993, none of the Top500 machines was made by a broadly based market driven company, while today over 3/4 of the Top500 are made by SGI, IBM, HP, or Sun. Similarly, vector architectures have been replaced in market share by microprocessor based SMPs. We now see a strong move to replace many MPPs and SMPs by the new architecture of Distributed Shared Memory (DSM) such as the SGI Origin or HP SPP series. A key trend is the move toward clusters of SMP/DSMs instead of monolithic MPPs. The next major change will be the emergence of Intel processors replacing RISC processors, particularly the Intel Merced processor which should become dominant shortly after 2000. A major battle will then shape up between UNIX and Microsoft's NT operating systems, particularly at the lower end of the Top500. Finally, with each new architecture comes a new set of applications we can now attack. I will discuss how DSM will enable dynamic load balancing needed to support the multi-scale problems that teraflop machines will enable us to tackle.

Applications Continue to Require Exponential Growth in Capacity

The history of supercomputing applications show that the exponential growth in capacity, which has increased by a factor of over one trillion in the last fifty years, is continually absorbed by the user community as they expand the complexity of their computational models. All indications are that this symbiotic relationship will continue into the future. A recent workshop organized by the National Science Foundation produced a chart describing this phenomenon.

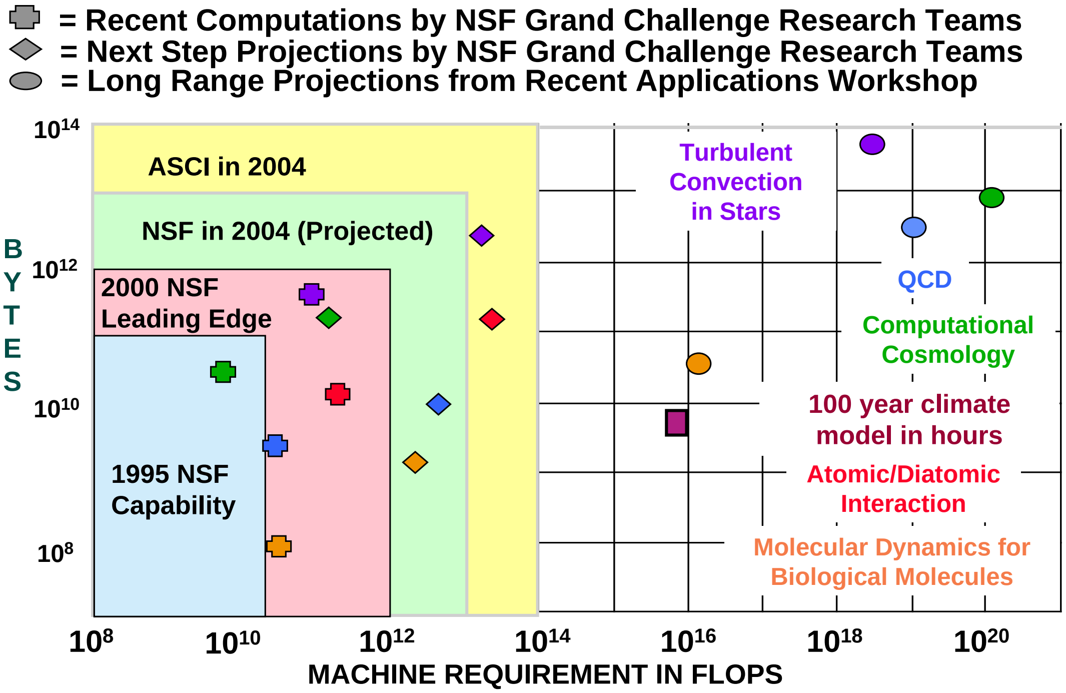


Figure 1: From Bob Voigt, NSF

In the United States, the Department of Energy initiated a major new program several years ago intended to more aggressively pursue high-end computation. The Accelerated Strategic Computing Initiative (ASCI-see www.llnl.gov/asci/) vision is to shift promptly from nuclear test-based methods to computational-based methods for ensuring the safety, reliability, and performance of the United States nuclear weapons stockpile. The ASCI program has installed two generations of new computers so far (ASCI Red with a peak of 1 Teraflop and ASCI Blue with a peak of 3 Teraflops). By 2000 ASCI White will advance the peak performance to 10 Teraflop and a 30 Teraflop peak machine is planned for 2002.

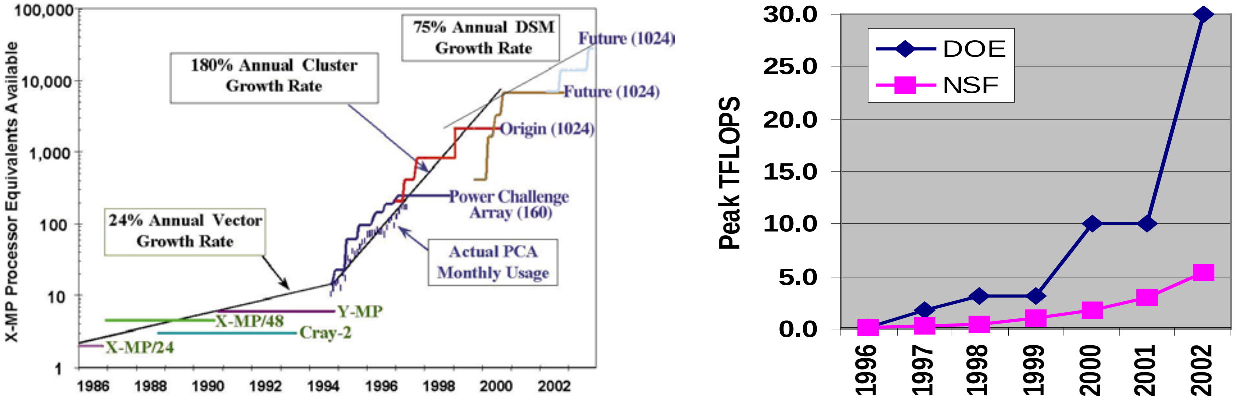


Figure 2: Left-Upgrade Path at NCSA Over Twenty Years
Right-Divergence of Peak Capacity of DOE ASCI and NSF PACI Supercomputers. Shown are the largest single systems in either the ASCI or PACI centers.

The National Science Foundation Partnerships for Advanced Computational Infrastructure (PACI) has two Leading Edge Sites at the National Center for Supercomputing Applications (NCSA) and the San Diego Supercomputer Center (SDSC), supplemented by partnering site facilities. Their supercomputers provide the largest scale systems for the U.S. academic community. The upgrade path for NCSA is shown on the left. Note that NCSA made the switch from shared memory vector supercomputers to shared memory microprocessor systems in early 1995. After that time we experienced the fastest growth rate in computational capacity in our history, roughly 180% compounded annual growth! This was possible because we were able to combine the Moore’s Law speed growth in microprocessors (60% annual rate) with the growth in parallelism from 4-way to 1024-way by moving to scalable systems.

However, as impressive as NCSA’s growth in capacity has been, it is falling behind the leading edge in the United States which is set by the ASCI program described above. As one can see from the graph above, the accelerated timeline for the ASCI program is causing a divergence between the largest single systems available to the DOE ASCI community and to the NSF academic community. This “raising of the bar” by the DOE ASCI program is leading to a broad discussion in the U.S. about how cooperation between the major Federal agencies can assure continued close coupling of the academic community to such teraflop systems.

The Top 500 and Market Transformations

The Top 500 provides a unique window into the complete transformation of the United States high performance market over the last five years. I have found the Top 500 listing invaluable in my role as NCSA Director during this turbulent period.

As can be seen from the following figure, in 1993 essentially all the U.S. Top 500 systems were made by stand-alone supercomputer companies creating either parallel vector processors (PVP-Cray Research and Convex Computer) or massively parallel processors (MPP-Thinking Machines and Intel Supercomputing). None of the Top 500 was built by an U.S. microprocessor based company.

Only five years later, 80% of the Top 500 machines are built by such companies (SGI, IBM, Hewlett-Packard, and Sun). All of the stand-alone companies were absorbed by these market-driven companies or went out of business. Today, more and more high-end computers appear as the upper end of a product line that starts on the desktop. This is a major and permanent change in the market place for high performance computers.

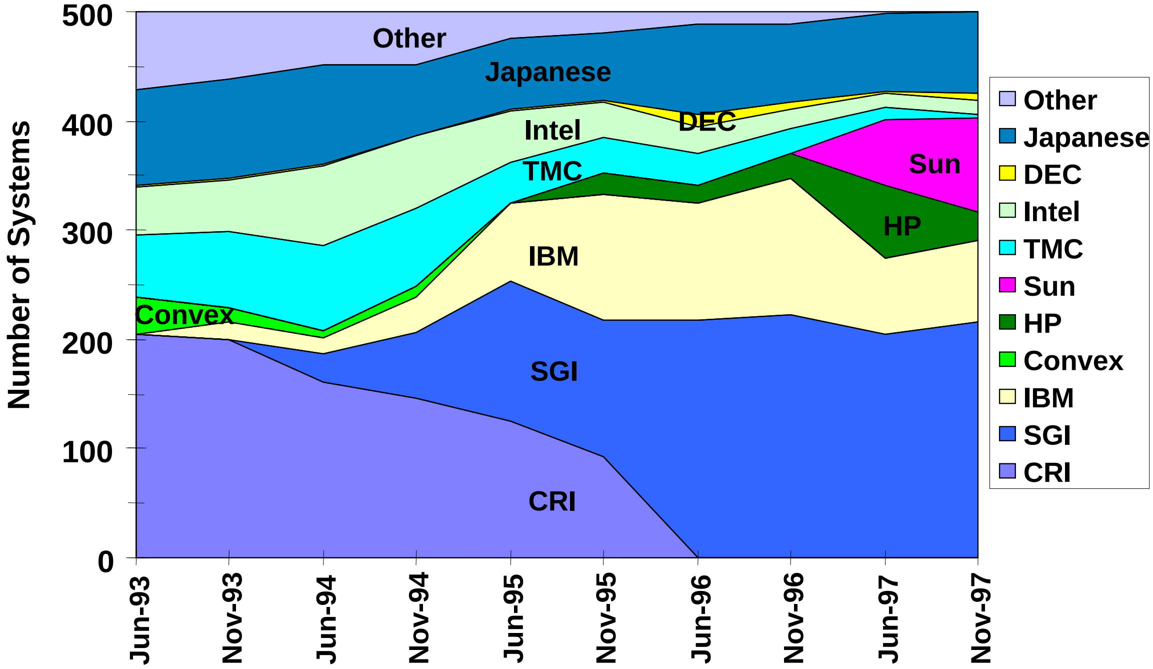


Figure 3: Top 500 Systems by Vendor—Installed Base

Furthermore, during this same period another important change, this one in supercomputer architecture, can be tracked by careful analysis of the Top 500. In the charts below, one can see that in 1993 well over half of the Top 500 machines were PVPs. However, shortly thereafter the market share of shared memory microprocessor-based architectures began to grow rapidly and the PVP market share began to shrink. The shared memory sector includes both monolithic shared memory computers (SGI Power Challenge, Sun E-10000) and Distributed Shared Memory (Kendall Square, HP Exemplar, and SGI Origin). By last year, SMP/DSM machines had grown to become half of the Top 500 and are clearly the largest market share of the three classes of high performance architectures (PVP, MPP, SMP/DSM)!

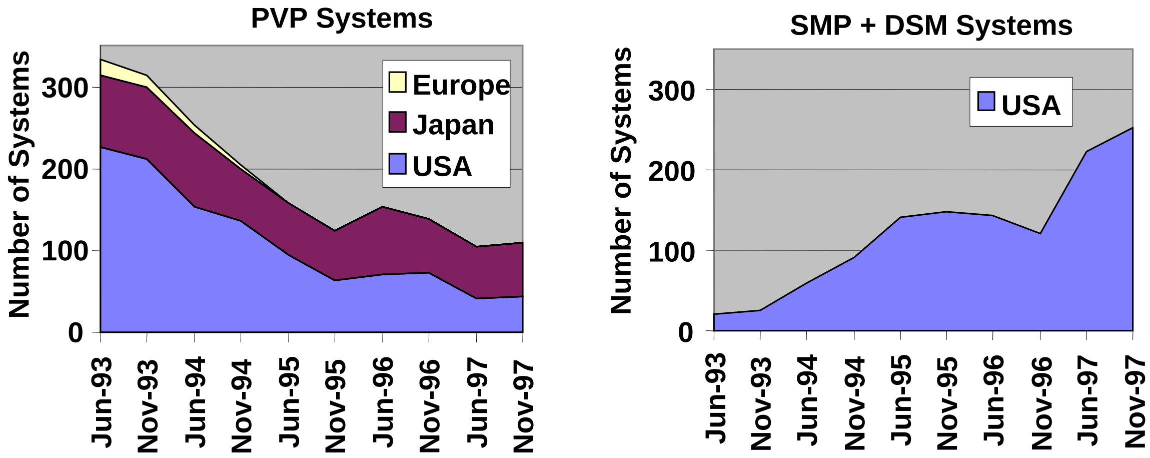


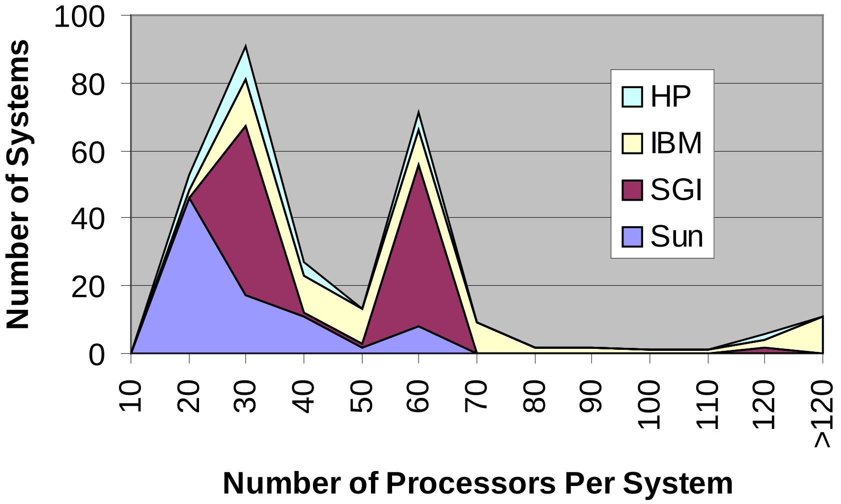
Figure 4: Evolution of Top 500 Shared Memory Systems Installed Base

From the U.S. perspective, another important feature is that while the shrinking PVP market is increasingly dominated by the Japanese manufactured vector supercomputers, the growing microprocessor-based

SMP/DSM market is essentially completely dominated by the U.S. Furthering this microprocessor trend, many of the MPP machines have migrated from specialized processors (such as the Intel i860 CPU in the Intel Paragon and the four vector units per CPU on the CM-5) to standard microprocessors (such as DEC alpha in the Cray Research T3E and RS/6000 in the IBM SP2) as well.

At first, the replacement of specialized processor supercomputers by microprocessors occurred at the low end of the Top 500. However, with the announcement of the winners of the DOE ASCI procurement, one saw the emergence of complete victory by the microprocessor in the United States. The first system, Sandia National Laboratory's ASCI Red machine was comprised of 9000 Intel Pentium Pros. Each of the Livermore and Los Alamos ASCI Blue systems will consist of roughly 5000 microprocessors. The Los Alamos National Laboratory Blue Mountain system will be organized as a cluster of SGI DSM Origins, while Livermore National Laboratory's ASCI Blue Pacific system will be an IBM SP organized as an MPP with 4-processor SMP nodes. The next generation ASCI White machine due into Livermore in 2000 will have 8096 IBM microprocessors in a tight cluster of 512 16-processor SMPs. Thus, the fastest supercomputers in the United States are going to be clusters of shared memory microprocessor-based machines.

If we take a closer look at the SMP/DSMs statistics in the Top 500, we can observe several other interesting trends. First, because the Top 500 has a lower cutoff based on Linpack performance, the distribution function of SMP/DSMs arranged by number of processors is incomplete at the lower end. Nonetheless, one can easily see that as one goes from 128 to 64 to 32 processors per machine, one has a rapidly rising installed base of such machines. If there were a Top 2000 list, the number of 16 processor machines would dwarf the number of 32 processors.



Includes: HP SPP-2000, IBM SP, SGI Origin, Sun HPC 6000,10000

Figure 5: Competing Servers Top500 November 1997 - Processors Per System

Herein lies the secret to the success of the microprocessor-based systems. Because companies can sell many shared memory machines with few processors for every large parallel machine they sell, there is a rational business model to support the high-end. Essentially the revenue from the large number of smaller machines supports the R&D costs on the larger machines. Furthermore, the large installed base of computers with small (4-16 way) parallelism guarantees software development by third party companies so that the SMP/DSMs will have a comparatively larger application software base than found on stand-alone supercomputer companies.

Clusters of SMPs/DSMs as the New MPP

The new scalable DSMs have developed over a number of years (see Scalable Shared-Memory Multiprocessing by Daniel Lenoski and Wolf-Dietrich Weber). These machines have distributed shared memory, but utilize hardware cache coherency to give the user a single system image, thus bringing scalability to the benefits of shared memory. Emerging from the DARPA funded DASH project, the first market products were the Kendall Square KSR-1 and the Convex Computer (now Hewlett-Packard) Exemplar. Although Kendall Square went out of business, Hewlett-Packard is doing very well with its V-series servers that were derived from the Exemplar and several major university installations have large second generation Exemplars (CalTech, NCSA, NCAR, University of Kentucky, and others).

The latest product utilizing DSM architecture is the Silicon Graphics Origin2000. It is available up to 128-processors in shared memory and can be clustered for larger parallelism. Two of the largest SGI Origin supercomputers are the clusters of Origins at NCSA and Los Alamos National Laboratory. Already, both sites have operational 128-processor Origins with single system image.

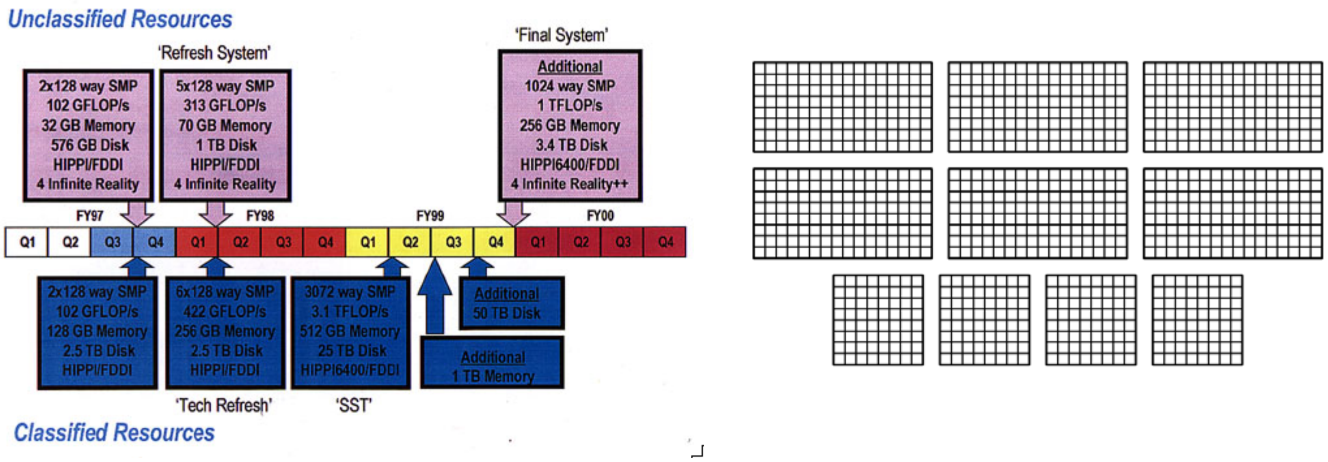


Figure 6: Left--The Los Alamos Cluster of Origins Delivery Schedule (http://www.lanl.gov/projects/asci/bluemtn/Hardware/schedule.html) and
Right-- the proposed FY99 NCSA Cluster of Origins. Each small square represents a MIPS processor and each large rectangle represents a single memory image [6x128 and 4x64=1024 processors]

DSM allows for a shared memory programming model, but with scalability heretofore only found on MPPs. Experiments have been done in running applications codes across four Origins coupled by HIPPI with good results for latency-tolerant codes. Next year, HIPPI-64 should be ready, greatly increasing the bandwidth between separate Origins. Thus, NCSA plans to have reserved time in which appropriately designed codes can run in 4x128=512 or even in 6x128=768 way parallel. Furthermore, the new 250 MHz MIPS R10k microprocessor seems to give a roughly 30% speedup over the current 195 MHz MIPS processor. Thus, on the largest configuration at NCSA, we hope to see sustained application speeds of 75 GFLOPS.

One of the major advantages of shared memory is the ability to handle dynamic load balancing much more efficiently than traditional MPPs. As the complexity of our simulations pushes us toward multiple length and time scales or multi-material computational domains, load balancing will become a requirement for efficient computing. I can illustrate this shift in application approach by considering an example drawn from the National Computational Science Alliance (hereafter “Alliance”) Cosmology Application Technologies Team. An MPP is excellent for computing the evolution of matter in a uniform spatial grid up to the point where the gravitational condensations become subgrid in scale. To follow such an evolution with a uniform grid requires large number of grid zones in each dimension even though at late times the vast majority of grid volumes have little gas in them.

In contrast, the use of modern adaptive grid refinement techniques enable one to compute the same evolution on much coarser main grid and then use nested subgrids as the density increases throughout the volume (see Figure 7). Because the details of the evolution will determine where and when subgrids are introduced, an a priori remapping of the gridzones to memory on an MPP is impractical. A dynamic remapping would cause the communication costs on an MPP to be so great as to make that approach unusable.

On the other hand, a shared memory computer can simply send pointers to map work in subgrids to the next available processor, yielding an efficient computation. Since both Nature and engineered devices are typically hierarchical, realistic simulations will require this sort of dynamic load balancing. Therefore, over the next few years, we expect to see a new generation of applications arise which will require the new features of DSM to run efficiently. To maintain portability of these new codes, we are seeing many of our pioneering users move toward a mix of portable standards, such as using MPI with the new OpenMP application program interface. OpenMP supports multi-platform shared-memory programming on Unix platforms and Microsoft Windows NT architectures (see www.openmp.org/).

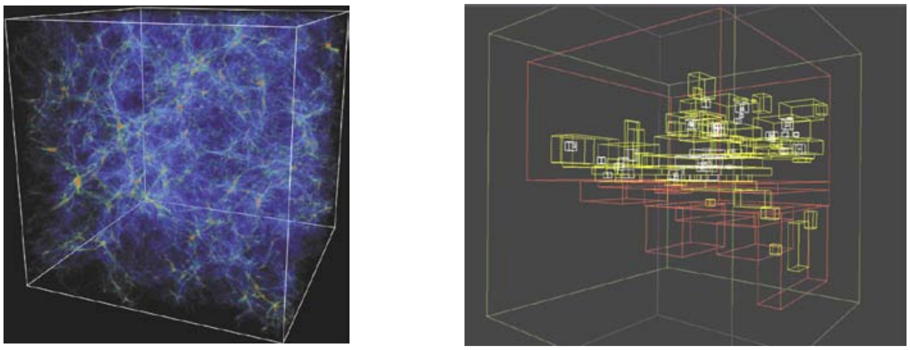


Figure 7: Left—A classic MPP application, homogeneous grids for cosmology with 512x512x512 on the CM-5 (Mike Norman, Greg Bryan, NCSA)
Right—New applications possible on shared memory machines. Adaptive grided cosmology with coarse grid of 64x64x64 and seven level nested grid with equivalent resolution of 8192x8192x8192 on SGI Power Challenge (Norman, Bryan, John Shalf, NCSA).

Linpack versus Baskets of Applications

The Top 500 listing is based on the use of the Linpack Benchmark (http://www.netlib.org/linpack/index.html), a numerically intensive test that has been used for years to measure the floating point performance of computers. Linpack is a collection of Fortran subroutines that analyze and solve linear equations and linear least-squares problems. While the Linpack benchmark has the great advantage of portability, the speeds reported are much higher than those found on real application codes. Therefore, a number of major efforts have been undertaken over the last few years to get a more realistic set of benchmarks.

One of the best known suites are the NAS Parallel Benchmarks (NPB), a set of five kernels and three pseudo-applications (://science.nas.nasa.gov/Software/NPB/), derived from computational fluid dynamics applications. The NPB are based on Fortran 77 and the MPI message passing standard. These implementations, which are intended to be run with little or no tuning, approximate the performance a typical user can expect for a portable parallel program on a distributed memory computer. As useful as the NPB have been, they have the disadvantage of being kernels rather than full applications.

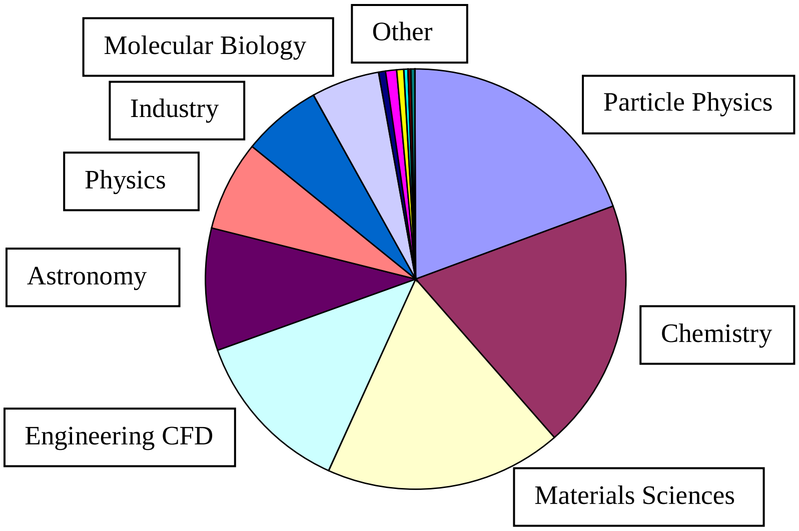


Figure 8: Fraction of CPU-hours run on various application disciplines on the NCSA Origin cluster (384 processors) during March 1998.

At NCSA, there is a broad spectrum of application disciplines that are supported on our supercomputers as seen in Figure 8. Working with our users, NCSA is starting a project to assemble a basket of applications drawn from our national user community, for which we have performance data, including single performance data and scaling behavior on a variety of modern supercomputers. The idea is to get a rough sense of how a supercomputer will likely perform on real world applications rather than on kernels. The basket of applications consists of one to two dozen user codes drawn from the disciplines of elementary particles, materials, chemistry, biomolecular dynamics, finite element and turbulent fluid engineering, atmospheric and oceanographic flows, and astrophysical gas dynamics.

While it is two early to get definitive results, one clear point already emerges. Across a wide range of single processors in widely used supercomputers (Cray Research C90, T90, T3E, Origin; IBM SP2; HP Exemplar), this basket of applications sustains speeds at roughly 20% of the machine's corresponding Linpack (N=1000). Our preliminary findings are that the SGI Origin 2000 has the largest fraction of Linpack on the basket of applications, roughly 33%. Whether use of a basket of applications (analogous to the widely used "basket of currencies" in international exchange rate calculations) would change the conclusions drawn from the Linpack ordered Top 500 remains to be seen.

Commodity Supercomputing and the Future of the Top 500

While the market transformation described above does indicate major changes in the HPC companies and architectures, it does not qualitatively change the procedure by which the Top 500 list is compiled. However, there is another emergent trend that clearly does. The Top 500 methodology assumes that the world's fastest computers are manufactured and sold by companies as integrated systems. However, many sites are creating more and more capacity from "build your own" supercomputers.

This phenomenon has been underway for at least ten years in such projects as the University of Wisconsin Condor program (www.cs.wisc.edu/condor/), which developed a middleware that can lay over a distributed network of UNIX workstations to create a high-throughput computer. Today, the active Condor flock at UW has over 300 workstations whose "spare" cycles are being used by many as they would use a supercomputer. Certainly a 300-processor supercomputer would be on the Top 500 list if it were built and sold by a company as an integrated device. The Alliance is spreading Condor flocks across its dozens of universities.

Industry has also been using distributed machines coherently for a long time. Digital Equipment is reputed to have linked thousands of VAX computers throughout the world to allow for fast recompiles of VMS in the 1980s. A number of aerospace corporations in the 1990s have routinely used pools of 500-1000 UNIX workstations to create an "overnight" computational capacity. It really is not known how widespread this practice is.

Such loosely coupled systems are not appropriate for application codes requiring high bandwidth communications. However, there is another technology trend, which may alleviate this shortcoming. A number of research sites have created tightly coupled piles of PCs or workstations, such as Beowulf (://www.biotec.or.th/~supat/beowulf/consortium.html) or the Berkeley NOW project (://now.cs.berkeley.edu/nowOverview.html). These pioneering efforts demonstrated over two years ago that gigaflop speeds with low latency networks could be obtained on these "do-it-yourself" machines and at a major improvement in price/performance. Of course, one pays for this advantage by the disadvantage of no vendor support, a shortage of third party software, and a requirement for a sophisticated local system support effort.

Nonetheless, I believe we are witnessing the birth of a third wave of supercomputing in these experiments. The first wave from the mid-1970s to the mid-1990s created supercomputers from specialized processors and often proprietary operating systems or highly modified UNIX. For instance, when NCSA took delivery of its first Cray Research X-MP it ran the Cray Time Sharing System. The second wave, described earlier and which is only about four years old, has vendors building supercomputers out of standard RISC microprocessors using the same UNIX as is found on desktop workstations. I expect this wave to continue to dominate supercomputing for at least another five years. The third wave is this new phenomenon of user-assembled large-capacity clusters of workstations or PCs. While the individual workstations or PCs are manufactured by computer companies such as Sun, Compaq, or Hewlett-Packard, those companies currently don't actively support the integration of their machines into such large capacity pools.

There are two major market trends, which will accelerate this phenomenon. As shown in the following charts, Intel will next year announce its first 64-bit microprocessor, created jointly with Hewlett-Packard, and code-named Merced. It will be very fast and software backward compatible with the several hundred million desktop PCs which use the Intel 32-bit processors such as Pentium and Pentium II. The combination of high volume and the world's most extensive software base almost guarantees that Merced will sweep all other microprocessors off the desktops over the next five years (for a detailed analysis, see the Microprocessor Newsletter, November 17, 1997). My guess is that it will happen on the server side as well by 2005. This means that either pre-existing loosely coupled systems of desktops or stand-alone racks of PCs in tightly coupled clusters will have awesome computing power. My expectations are that a teraflop system could be purchased for under one million dollars by 2005.

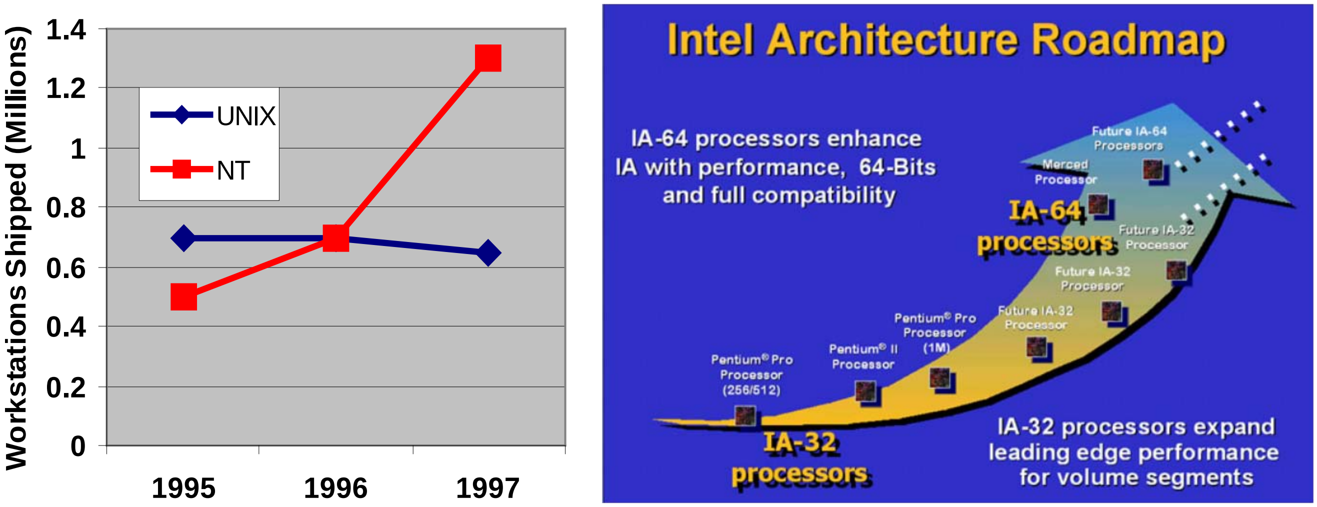


Figure 9: Left-NT Workstation Shipments Rapidly Surpassing UNIX
(Source IDC and Wall Street Journal, March 6, 1998)
Right-Intel 32-bit and 64-bit roadmap (://developer.intel.com/solutions/archive/issue5/focus.htm#FOUR)

The other major market trend is the rapid eclipsing of UNIX market share by Microsoft's NT operating system. As one can see, while the number of UNIX workstations shipped annually has slightly declined over the last three years, NT workstations are on a doubling each year curve and now greatly outnumber the UNIX workstations shipped. As a result, many third party application software vendors are developing their codes on NT first and then porting to UNIX. While UNIX still has many capabilities that NT wont have for years, the advantages of seamlessly integrating an NT/Intel workstation or server into a LAN containing a large number of NT/Intel desktops seems to be winning the market struggle.

Anticipating these strong market trends, NCSA is working to create access to such NT/Intel clusters for the academic research community. We have two projects in this area: NCSA Symbio for loosely coupled NT/Intel clusters and High Performance Virtual Machine for tightly coupled systems.

NCSA Symbio (symbio.ncsa.uiuc.edu/) is a distributed object system that is built upon the Microsoft Distributed Component Object Model (DCOM). Structurally, it consists of two lines of development. First, it is an object management system for a cluster of NT workstations which allocates resources, schedules processes/jobs, implements fault tolerance/object migration and provides user interfaces/API's for controlling and observing distributed processes. Second, Symbio is a set of object libraries that enables developers to create objects which inherently support the interfaces that are required to interact with the management system. Symbio can be used to develop and run parallel, distributed applications exploiting the high availability and low cost of NT workstations.

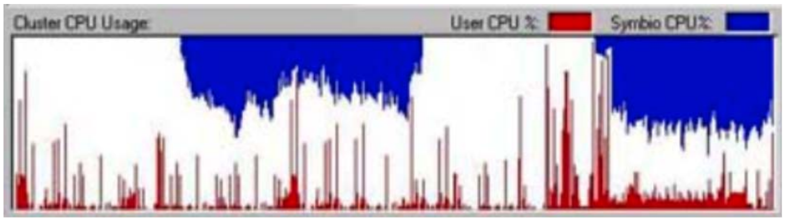


Figure 10: Symbio using Cycles from a LAN of NT/Intel Computers (Briand Sanderson, NCSA)

NCSA's second path is similar to the previous projects mentioned above for tightly coupling PCs, except that it uses the commercial Microsoft NT operating system instead of UNIX or LINUX, which most of the previous Intel clusters have used. For a compilation of performance measurements on a variety of PC clusters see the NASA web page (http://science.nas.nasa.gov/Software/NPB/NPB2Results/cluster.html). The NCSA/UIUC NT supercluster is an easily cloned system built from a mass market operating system (Microsoft Windows NT) and off-the-shelf PCs (HP and Compaq dual Pentium II's currently) using high-bandwidth low-latency networks.

While the supercluster does not have the more sophisticated Distributed Shared Memory architecture of either NCSA's SGI/Cray Origin2000 or Hewlett-Packard SPP-2000, it should run many applications that currently use those systems at NCSA. We view the NT supercluster as a first generation experimental MPP that may be useful to supercomputing applications that don't require shared memory or a 64-bit environment. By providing this alternative platform, we should free up the Origin cluster for leading edge computations.

Andrew Chien, a professor in the University of Illinois Department of Computer Science and a member of the Alliance Parallel Computing Team, and his research group working with staff from NCSA, recently completed construction of a 256-processor Windows NT supercluster for high performance computing research. The supercluster consists of 32 Compaq and 96 Hewlett Packard Windows NT PC workstations. These 128 workstations, connected by a Myricom Myrinet network, create a 256-processor supercluster. NCSA has plans to upgrade the cluster to 512-processors next year, perhaps eventually to thousands of processors.

The cluster uses Chien's High Performance Virtual Machine (HPVM) software that synthesizes clusters of Windows NT processors into a high-performance environment. HPVM relies on Fast Messages (FM) as the core communication layer which is designed to deliver underlying network's hardware performance to the application, even for small messages, without requiring changes in applications (or protocols) to increase message size. FM is also designed to enable convenient and high performance layering of other API's and protocols atop it. As such it provides key guarantees: reliable, in order delivery, and host-network decoupling as well as a composable interface: efficient gather/scatter, receiver rate control, and per-packet multithreading which make it easy to build higher level interfaces based on FM. Some of the interfaces which have been built include MPI, Shmem Put/Get, and Global Arrays. A version of High Performance Fortran (HPF) that takes advantage of HPVM's Fast Messages Application Programming Interface (API) is expected in several months from the Portland Group. Currently, NCSA and Chien’s group are studying various software techniques to introduce DSM-like capabilities into the NT supercluster.

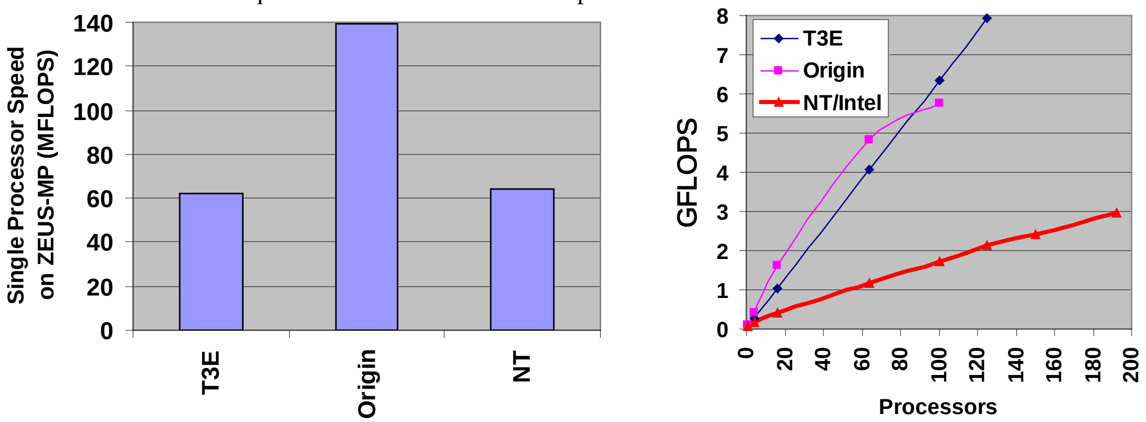


Figure 11: Comparison of Single Processor Performance and Scaling of MPP, DSM, and NT Supercluster

At Alliance’98, held at NCSA May 27-30, 1998, the first demonstration of a supercomputing code running on the NT supercluster was made. It ran the MPI ZEUS-MP cosmology application code used in Figure 7 above, on 192 processors of the supercluster working together as one HPVM machine. This virtual machine had 50 gigabytes of memory, 400 gigabytes of disk space and almost 4 gigabytes per second of bisection bandwidth across the Myrinet network. One can see that the widely used Intel Pentium II (300 MHz) microprocessor runs codes at similar single processor speeds as traditional supercomputers (see graph), but the scaling on these first tests was a factor of 3-4 worse than MPPs or DSMs. On the other hand, the scaling is remarkably linear and this is only the first test.

In addition to ZEUS-MP, ISIS++, a portable object-oriented framework for solving sparse systems of linear equations such as found in large-scale finite element analysis codes, also ran on 192 processors. The code was developed at Sandia National Laboratory. Other supercomputer codes that will run on the NT supercluster during the next month include applications in elementary particles, materials, and fluid turbulence. While it is still too early to predict how many types of applications will parallelize efficiently on the NT supercluster or what the precise price-performance obtained will be, we believe it is encouraging that these supercomputer codes have so easily been ported to this new environment.

The NT supercluster should not be thought of as a substitute for a commercial supercomputer today. A great deal of research remains to be done on several generations of the superclusters that NCSA will build over the next few years. It will be several years before NT has the 64-bit characteristics of modern UNIX operating systems. Indeed, such NT superclusters may never replace commercial supercomputers. However, they seem destined to create a great deal of capacity for users who have applications that can take advantage of the excellent price-performance.

Acknowledgements

I would like to thank NCSA staff members Charlie Catlett, Lex Lane, Rob Pennington, John Towns and Alliance Partner Andrew Chien for help with data analysis and for stimulating conversations. I also appreciate all the NCSA users who sent me their performance data as well as Forest Baskett, John Brown, and John McCalpin of Silicon Graphics for discussions. Thanks to Joseph Smarr for his skillful help with the final manuscript.