Ishfaq Ahmad
Hong Kong University of
Science and Technology
Clear Water Bay, Kowloon
Hong Kong
iahmad@cs.ust.hk

Gigantic clusters: Where are they and what are they doing?

Clusters are like skyscrapers—mammoth structures towering above their smaller counterparts—except that skyscrapers have longer lives. The race to build ever bigger clusters is on, largely driven by emerging scientific and engineering applications. Several research efforts are underway at various universities and US research laboratories. In this article, I examine some of the largest clusters in the world, providing some recent news and opinions from the experts.

A cluster is a collection of complete computers (nodes) interconnected by a high-speed network. Typically, each node is a workstation, PC, or symmetric multiprocessor (SMP). Cluster nodes work collectively as a single computing resource and fill the conventional role of using each node as an independent machine. A cluster computing system is a compromise between a massively parallel processing system and a distributed system. An MPP system node typically cannot serve as a standalone computer; a cluster node usually contains its own disk and a complete operating system, and therefore, also can handle interactive jobs. In a distributed system, nodes can serve only as individual resources while a cluster presents a single system image to the user.

In recent years, the performance of commodity off-the-shelf components, such as processor, memory, disk, and networking technology, has improved tremendously. Free operating systems, such as Linux, are available and well-supported. Several industry-standard parallel programming environments, such as MPI, PVM, and BSP, are also available for, and are well-suited to, building clusters at considerably lower costs than their counterpart MPP systems.

Scalability—a system's ability to scale its resources and sustain its performance accordingly—is fundamental to parallel computing. Although this concept has been realized in MPP systems (both nonshared and shared memory architectures), it has been less certain with clusters, largely because of the software and hardware overheads in the interconnection networks. Integration of all the system resources, such as processors, memory, disks, networking, and I/O subsystems, is an important factor. This situation is now changing, and considerably larger clusters (in excess of 32) are being built. Some questions that arise are

- how big are these clusters?
- where are they?
- how are they being used?

Gigantic clusters and their use

Let's look at some of the more interesting applications of cluster computing.

Networks of Workstations

The Network of Workstations project at the University of California at Berkeley combined elements of workstation and MPP technology into a single system. The system, comprising 100 Sun Ultrasparcs, used fast communication primitives with active messages as well as fast implementations of conventional communication layers, such as Sockets and MPI. NOW served successfully for many applications, including the killer Web search engine application that became "Inktomi," according to project leader David Culler. NOW's successor, the Millennium project, aims to develop and deploy a hierarchical campus-wide system to support advanced applications in scientific computing, simulation, and modeling. Individual nodes are SMPs, making Millennium essentially a cluster of clusters. Millenium's largest cluster is comprised of 208 processors, which soon will increase to 290. It is being used for a variety of large simulation services, including analysis of electron beam lithography for next-generation CAD, earthquake modeling through finite-element methods, and large network simulation.

Beowulf

Beowulf-class computers run operating systems such as Linux and use cost-effective, commodity off-the-shelf technologies. (*Beowulf* is commonly used to describe a cluster of PCs.) They deliver heavy computing power for scientific and engineering applications at the lowest possible price. Although they are similar to NOW, Beowulf computers are referred to as a "Pile-of-PCs" to emphasize their use of mass market commodities; dedicated processors (rather than stealing cycles from idle workstations); and a private communications network. HIVE, a Beowulf-class parallel computer at NASA's Goddard Space Flight Center, consists of four subclusters containing 332 processors (mostly Pentium Pros and Pentium III Xeons). The network includes Foundry Network Switches with fast Ethernet and gigabit ports.

This massive system, which has several gigabytes of cumulative memory and more than a terabyte of disk space, is solving problems associated with large data sets, particularly in Earth and space sciences communities. One problem is photorealistic rendering, which involves the physically-based simulation of light transport in a complex geometric domain. This kind of domain might contain a diversity of materials, each scattering light according to a different bidirectional reflectance distribution function. To generate high-quality images that accurately predict the appearance of a hypothetical scene, a variety of radiometric phenomena must be simulated, including light scattering, interreflections among surfaces, and camera effects such as depth of field and motion blur.

> "Multiple multiple-terascale Linux-based superclusters will be built in the next year with a 10-terascale or better Linux super-cluster highly likely to appear in about a year."

The California Institute of Technology and the Jet Propulsion Laboratory maintain a number of Beowulf-class PC clusters. Most of these clusters are comprised of 32 processors and include Compaq Alpha as well as Pentium II and III-based systems and use Gigabit Ethernet or Myrinet as the interconnection network. The largest system is comprised of 140 Pentium Pros and Pentium IIs (a heterogeneous mix) using Fast Ethernet. These clusters are being used for climate modeling, synthetic aperture radar processing, environment modeling, astronomy database, cosmology *n*-body studies of galactic evolution, space-flight mission simulation, image processing, antenna design using genetic algorithms, the Laser Interferometric Gravitational Observatory, and electromagnetic transmission system modeling. Caltech's Thomas Sterling says these systems are highly useful because they are capable of sustaining about 10 Gflops on nontrivial problems at an affordable cost. Because these systems can be dedicated to a single parallel program for many days, they can do the same work as much larger systems that must be shared among a large and diverse workload.

Collosal clusters at Alliance centers

The Albuquerque High Performance Computing Center at the University of New Mexico has long been a proponent of colossal clusters. The AHPCC and the National Computational Science Alliance (the Alliance), comprising more than 50 academic, government, and industry research partners from across the US, have formed a partnership that the National Science Foundation funds. The Alliance, which wants to provide an advanced computational infrastructure, is running a 128-processor Linux SuperCluster with Myrinet (Roadrunner) from Alta Technologies using dual Intel 450-MHz nodes, each with 512 Mbytes of RAM. The AHPCC is acquiring a 512-processor Linux SuperCluster known as Los Lobos, reports David Bader of the University of New Mexico. The Alliance intends to make Los Lobos the largest open-production Linux supercluster geared to the research community.

Los Lobos uses dual Intel 733-MHz IA-32 processor Netfinity 4500R nodes; 1 Gbyte of memory per node; 1 Tbyte of SSA RAID; and 2 Tbytes of tertiary storage (tape robot) to deliver a peak theoretical performance of 375 Gflops. The high-performance interconnect network between the cluster nodes is Myricom's Myrinet, providing speeds exceeding 1 Gbits per second, which is comparable to the fastest interconnects in today's traditional supercomputers. The system maximizes the computing power per square foot on an Intel-based platform. This thin server is designed to deliver the highest computing power per square foot on Intel-based platforms and is one of the industry's most complete rack-optimized product lines for Linux, Windows, and Novell servers, according to IBM. Patricia Kovatch, High Performance Computing Systems Group manager, thinks Linux clusters are compelling for several reasons, especially for the cost-performance ratio. Furthermore, the cost is much less than buying a traditional supercomputer, and the performance rivals one. Another benefit for applications folks is the ease of porting their applications to Linux from other Unix environments. The prospects for the near future look even more momentous: "Multiple multiple-terascale Linux-based superclusters will be built in the next year with a 10-terascale or better Linux supercluster highly likely in about a year," says Frank Gilfeather, Executive Director of High Performance Computing.

Another Alliance partner, the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, is running a 128-processor dual-Pentium-III Xeon, 550-MHz, 1-Gbyte RAM, and Myrinet network technology on an NT operating system.

The National Labs

Los Alamos National Laboratory Center for Nonlinear Studies and Theoretical Division has a 140-processor Alpha Beowulf cluster, named Avalon, which was constructed from commodity personal computer technology. Each Avalon node contains a DEC Alpha workstation in an ATX case, 533-MHz

21164A microprocessor on a DEC Alpha PC 164LX motherboard, 256 Mbytes of memory per node, a 3,079-Mbyte hard disk per node, and a switched network of 144 fast Ethernet ports with 3-Gbit links on each switch. Avalon is running the Linux operating system.

The Tennessee Oak Ridge Cluster (TORC) project, a collaborative effort between the University of Tennessee's Innovative Computer Laboratory and the Computer Science and Mathematics Division of Oak Ridge National Laboratory (ORNL), is running a 128-processor Pentium III Linux cluster called "High TORC." ORNL's Al Geist says, "This Linux cluster, although not the largest of its kind in overall power, has the distinction of being the largest cluster in the nation interconnected with Gigabit Ethernet." Researchers around the nation are using it for gene modeling and for developing a toolkit for managing and monitoring multiple clusters.

ORNL recently acquired two massive clusters. The first is Eagle, an IBM SP3 cluster that has 184 nodes, each with four 1.5-Gflops IBM Power3 II processors, 372 Gbytes of memory, and 9.2 Tbytes of local storage. The Eagle system has a total theoretical speed of 1.08 Tflops. "At the present time, ORNL has the distinction of having the most powerful open nonclassified computer cluster in the world at 1.08 Tflops," Giest says. It is dedicated to nonclassified scientific computing and is used for a variety of national computational science research projects, including global climate prediction, human genome studies, nanotechnology, materials design, and molecular modeling. The second system, Falcon, is a Compaq Alpha cluster dedicated to large-scale computer science, and particularly to developing better tools and more scalable algorithms for computational science researchers involved with early systems' evaluation. Falcon has 80 nodes, each with four 1.3-Gflops Alpha EV67 processors, 160 Gbytes of memory, and 5.5 Tbytes of local storage. Falcon is an SMP cluster and can deliver 427 Gflops of theoretical speed.

Cornell Theory Center's Advanced Cluster Computing Consortium (AC3) maintains a cluster named Velocity, which consists of 256 processors (64 nodes), a Dell Power Edge 6350 four-way Pentium III Xeon SMP, 500-MHz CPU with 4 Gbytes of RAM per machine, 54 Gbytes of disk storage per machine, and Giganet Gbit technology on an NT operating system. Velocity recently expanded to 344 processors. The system is used for a complex large-scale materials simulation code as part of a project to develop multiscale simulations to let engineers design and predict materials performance from the atomic level all the way up to the product level.

The AC3 group, led by Anthony Ingraffea, is also modeling crack propagation through materials faster than its ever been done before. The AC3's goal is to select an object, generate a mesh for it, and produce 1,000 time steps of crack propagation with one million degrees of freedom in less than one hour. Hemoglobin protein modeling is another major application of the cluster. Hemoglobin carries oxygen molecules from the lungs to the muscles, switching between an "on" state in the lungs to pick up the oxygen and an "off" state in the muscles and other tissues to release the oxygen. To switch states, the protein also must change shape. To understand biochemical processes, protein dynamics must be simulated at the atomic level. The hemoglobin switch requires tens of microseconds to change state that is 10,000 times longer, but the simulations are limited to a few nanoseconds using current supercomputers. A Cornell research group has developed a model to capture the hemoglobin switching process dynamics using STO on up to 96 processors on the AC3 Velocity cluster with near linear speedup. Simulations that previously took up to three years to run on a serial machine using standard methods and software can now run in less than one day.

The Computational Plant Project at Sandia National Laboratories Project started with a cluster of 128 digital personal workstations, each of which contained a 433-MHz 21164 microprocessor, 192 Mbytes of ECC SDRAM, 2 Mbytes of L3 cache, a 10/100Base-Tx-integrated Ethernet, and 2.1 Gbytes of HDD Myrinet Gbit networking hardware on a Unix operating system. This system was upgraded to 400 digital personal workstations, each containing a 500-MHz 21164 microprocessor, 192 Mbytes of ECC SDRAM, 2 Mbytes of L3 cache, and a 10/100Base-Tx-integrated Ethernet. After another upgrade, the system now consists of a compute partition and a system support partition. The compute partition is comprised of 592 Compaq XP1000 workstations, each containing a 500-MHz 21264 microprocessor, 256 Mbytes of RAM, and a 10/100Base-Tx-integrated Ethernet. The system support network is comprised of 36 Compaq XP1000 workstations, each with a hard disk drive.

The Chiba City system at Argonne National Laboratory has a scalable cluster architecture with multiple partitions. The main partition consists of 256 compute nodes, each with dual Pentium III 500-MHz processors, 512 Mbyes of RAM, and 9 Gbytes of local disk. Thus there are a total of 512 computer nodes. In addition, the system contains four dual-processor systems for login nodes. A visualization partition is comprised of 32 IBM Intellistation M Pros with Matrox Millineum G400 graphics cards, 512 Mbytes of RAM, and 9 Gbytes of local disk. A storage partition consists of eight IBM Netfinity 7000s with 500-MHz Xeons, 512 Mbytes of RAM, and 300 Gbytes of disk. Finally, a management partition consists of 12 "mayor" nodes IBM Netfinity 500s with 500-MHz Pentium IIIs, 512 Mbytes of RAM, and 200 Gbytes of disk. The system is interconnected with a high-performance communications 64-bit Myrinet. The cluster, which was built in partnership with IBM and VA Linux Systems, supports scalable parallel computing with major applications that include message passing, job startup, parallel file systems, and systems management.

Now that cluster computing is an accepted form of supercomputing, the challenge is to continue exploring new methods for efficient network, I/O, and system software, while finding niche applications with commercial potential. It will be great if these gigantic clusters could affect the economy directly.