PARALLEL AND DISTRIBUTED SUPERCOMPUTING AT CALTECH

Paul Messina

Caltech Concurrent Supercomputing Facilities
Caltech, Mail Code 158-79
Pasadena, California 91125 U.S.A.

Abstract

Caltech uses parallel computers for a variety of large-scale scientific applications. Earlier in the decade locally designed and built machines were used. More recently we have acquired commercial parallel computers, some of which have performance that rivals or exceeds that of conventional, vector-oriented supercomputers. A new project has been started that builds on our experience with concurrent computers and attempts to apply our methods to the simultaneous use of parallel and vector supercomputers at four institutions that will be connected by a 800 Mbits/sec. wide-area computer network. Distributed supercomputing experiments will be carried out on this testbed.

Concurrent Supercomputing

Researchers at the California Institute of Technology (Caltech) have been using concurrent computers for a wide variety of scientific and engineering applications for nearly a decade. The hypercube computer architecture developed and first implemented at Caltech in the early 1980's provided the stimulus and opportunity for such early production use of parallel computers. People at Caltech invested their time and effort in learning how to program parallel computers for the usual two reasons: the promise of better price-performance than sequential machines and the ability of parallel computers to scale to much greater performance levels and memory sizes than traditional supercomputers. The first of these has of course been realized by many commercial parallel computers. The second of these advantages, which at Caltech is the more important, has only recently been demonstrated, for example, by the Connection Machine model CM-2 in its largest configuration.

In pursuit of the advantages of parallel computers, in 1988 Caltech established the Caltech Concurrent Supercomputing Facilities (CCSF). CCSF has acquired and operates a variety of concurrent computers, most of which are fast enough that they achieve speeds comparable to one or more CRAY Y-MP processors for real applications. Figure 1 shows the current complement of machines in CCSF. The first-generation NCUBE, the JPL/Caltech Mark IIIfp, the Intel iPSC/860, the Symult S2010, and the CM-2 are all heavily used. Each of these systems has several heavy users at any one point in time. Typically, demand for computational resources is greater than availability, so time on the machines has to be rationed among the users.

Our decision to do large-scale computations on parallel computers has paid off reasonably well. Many groups have used CCSF systems to do computations of a size that would have required prohibitively large time allocations on traditional supercomputers. Perhaps because almost all of our application programs are moderate in size (less than 10,000 lines) and because many of them were developed specifically for a parallel computer, getting efficient parallel implementations has not been the most difficult problem. General system instability, system software immaturity, and bottlenecks in the machine configurations have been major impediments. In other words, it has been the newness (and therefore immaturity) of the computers rather than their parallel architectures that has created the most problems. For example, the Fortran and C compilers for most of the machines produce rather inefficient sequential code, so that it is frequently necessary to do a little assembler programming to get a reasonable fraction of the hardware's peak speed.

Applications are programmed in Fortran or C with message passing extensions. Two message-passing systems are widely used on our Multiple Instruction Multiple Data (MIMD) systems: Express, which is a commercial product based on early work at Caltech and Cosmic Environment/ Reactive Kernel which is the product of Charles Seitz's research group in the Computer Science Department. Both run on a variety of computers, including network-connected workstations, and thus provide portability for parallel programs at the source level.

Caltech is by no means the only institution that is working actively in large-scale computing on parallel computers. Many research laboratories and other universities have also acquired and used parallel computers. In the fall of 1990, thirteen organizations joined Caltech in forming the Concurrent Supercomputing Consortium (CSC). The CSC was formed to acquire computers larger than any single

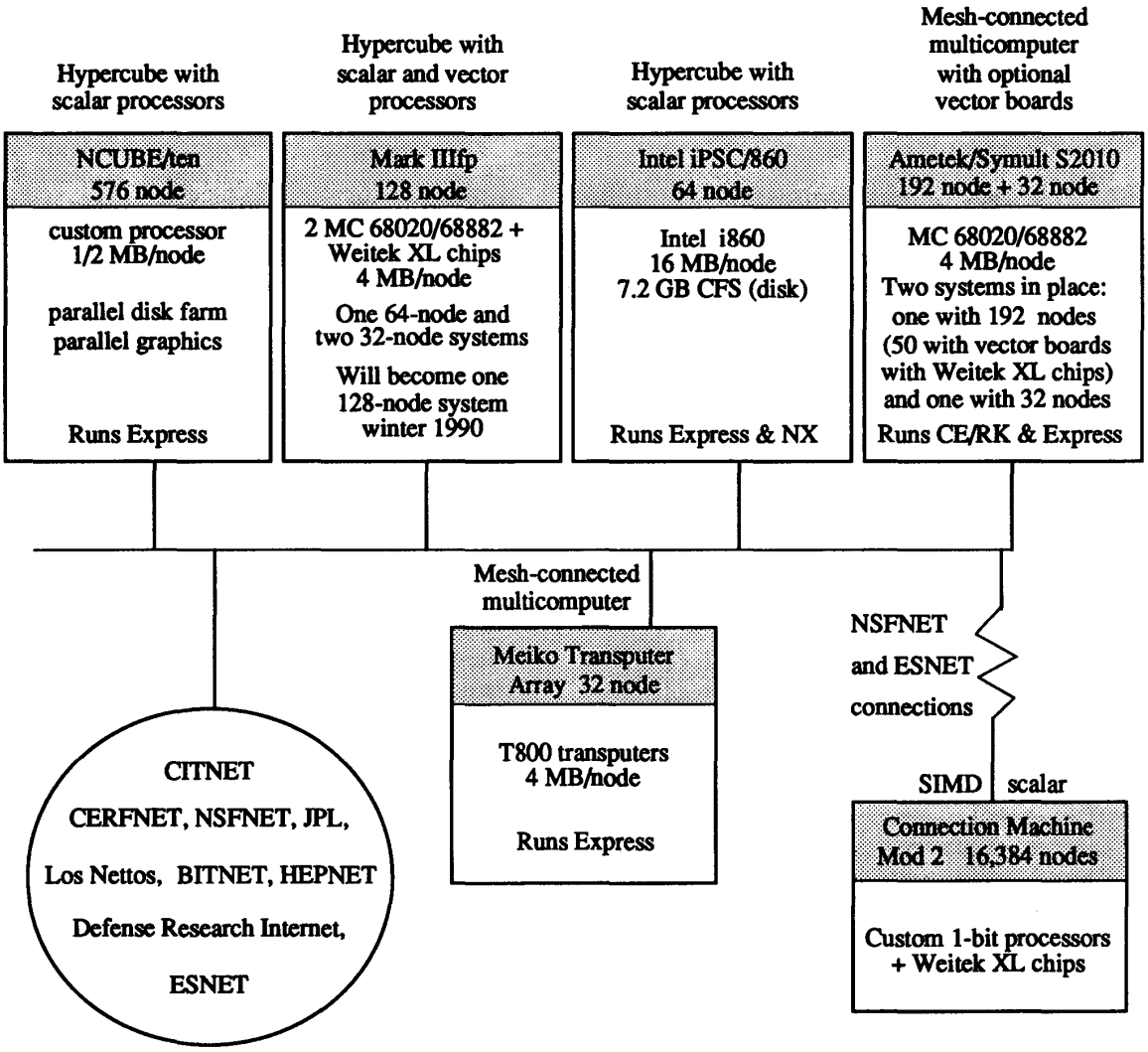

Figure 1. Major Systems in the Caltech Concurrent Supercomputing Facilities.

member can currently afford and to share information and expertise on large-scale scientific computing on massively parallel computers. Caltech will acquire and operate the Intel Touchstone Delta System on behalf of the CSC. The Delta will be delivered in the spring of 1991. It is a distributed memory MIMD system whose nodes are connected in a two-dimensional mesh by mesh-routing chips developed by Seitz at Caltech. With a peak speed of 32 gigaflops and over 8 gigabytes of memory it will provide a powerful new tool to computational scientists at CSC sites.

Distributed Supercomputing

Using concurrent machines to carry out supercomputer-level computations has proved to be feasible and, with systems like the Intel Delta and Thinking Machines CM-2, provides a path to greater computing power than traditional approaches. However, even those systems are not appropriate or adequate for certain tasks. As scientists study ever more complex phenomena through computer simulation, they often require multidisciplinary approaches, information stored in varied databases, and computing resources exceeding any single supercomputer. The necessary resources are geographically dispersed. All supercomputers cannot be put in one place. Scientists with expertise on various aspects of a problem reside at several universities and research laboratories. Huge, frequently-updated databases are maintained at only one site. Furthermore, the design details of the most advanced supercomputers make some better suited for certain computations than others. In a typical large-scale simulation, there are several computational steps; some phases of the computation may be most efficient on a Single Instruction Multiple Data (SIMD) parallel machine, other phases may run better on a MIMD computer.

Distributing large computations among several supercomputers provides the opportunity both to bring to bear greater computing power than is available in any single machine and to use the most suitable machine for each step of the task. By correctly decomposing an application, nonlinear speedups might be achieved. The execution time can be decreased by a factor greater than the sum of the effective speeds of each CPU. In addition, by creating a high-performance distributed environment, a further paradigm shift in science methodology can occur. Computational models can be extended into real-time interactive simulations that integrate data from experiments, satellites, and databases. The field of interactive simulation promises to be an extremely powerful tool for science.

Computer networks can provide the needed connection between the people, the machines, and the data, but today's networks are far too slow to support most large-scale applications across wide areas. This mismatch of speeds precludes effective distribution of work among dispersed supercomputers; the exchange of data among the collaborating computers would be very slow compared to the calculation rate. The computers would spend much of the time waiting for intermediate results to be communicated by their fellow workers so that the next phase of the computation can be tackled. Networks with much greater bandwidth can be built, but the greater speed is not sufficient to solve the problem. Because thousands of miles must be traversed, the time to deliver a message from one location to another will still be significant due to propagation delays.

To address these new areas, Caltech has joined with the Jet Propulsion Laboratory, Los Alamos National Laboratory, and the San Diego Supercomputer Center in the CASA project. CASA has support from the Corporation for National Research Initiatives (CNRI), which was awarded a grant by the National Science Foundation and the Defense Advanced Research Projects Agency for such activities. CASA will create a network testbed that will connect all four sites and operate at gigabit/second speeds. The goal of the CASA testbed is to demonstrate that high-speed networks can be used to provide the necessary computational resources for leading-edge scientific problems, regardless of the geographical location of these resources. Three important scientific problems will be studied by harnessing the computer and data resources at the four institutions.

A key challenge is to devise ways to use multiple supercomputers and high bandwidth channels with large latencies to solve important problems. While high latencies can be minimized by the correct protocol and amortized by transmitting large amounts of data, choice of algorithms and application decomposition methods will also be investigated for optimally driving the "meta-computer" that will be formed by linking high-performance computers with the network.

The three applications that the CASA testbed will adapt to run in a distributed fashion are from the areas of chemistry, geophysics, and climate modeling. Chemical reaction dynamics computations will be carried out to study the reaction of fluorine and hydrogen, which is relevant to powerful chemical lasers. These computations involve operations on very large matrices and require frequent communication of large blocks of data between the computers that participate in the calculation. The second application will develop an interactive visualization program for geological applications that takes input from Landsat, seismic, and topographic databases. Among the benefits of such analysis will be much clearer identification of fault zones, plate thrusts, surface erosion effects, and an improved ability to predict earthquake magnitude. The climate modeling application will combine ocean and atmospheric models simultaneously running in separate computers and continually exchanging data across the

CASA network. The resulting concurrent dynamic model will be much more realistic than existing models that use static data for either the ocean or the atmosphere boundary conditions and will be used to predict global change in long-term climate simulations.

The primary hosts on the CASA network will be several CRAY computers, the massively parallel CM-2 from Thinking Machines, and the Intel Touchstone Delta system. To use these systems (and others) in a distributed environment, our approach is to use algorithms, programming systems, and tools that have been developed for parallel computing and adapt them to a wide-area distributed computing environment with high communications latency. The Express programming environment mentioned earlier is being generalized to operate across networks and on new systems such as the CRAY Y-MP at SDSC, the Connection Machine at Los Alamos, and the Delta at Caltech.

Acknowledgments

This work was supported by the U.S. Department of Energy: Applied Mathematical Sciences (Grant DE-FG03-85ER25009); Program Manager of the Joint Tactical Fusion Program Office; and U.S. National Science Foundation: Center for Research on Parallel Computation (Grant CCR-8809615). Tina Mihaly provided editorial assistance and designed Figure 1.