chitecture scalability and serviceability requirements may preclude this approach, except in small systems). The MPS designers will have to package their standard microprocessors in a custom die, just as in the supercomputer, making the two systems very similar. When this eventually occurs, the supercomputer company will either be vertically integrated—that is, owned by a semiconductor manufacturer—or have vertical cooperation agreements to allow manufacturing of the custom dies on a memory processing line. This shared use allows the semiconductor company to recover some of its very large capital investment in a memory processing line for those situations in which some excess capacity is available.

Memory bandwidth is one of the key parameters determining the performance of parallel systems. From the Cray Research Y-MP to the Cray C90 to the Cray C95, there are rather dramatic advances in the memory bandwidth being provided: a factor of 6 from the Y-MP to the C90 and a factor of 4 expected from follow-on systems. Looking at specific cases of MPS suppliers, we find, in the case of Thinking Machines, that as they moved from the CM1 to the CM2 to the CM5, there has been a decrease in the number of processors and an increase in the system's power. In the case of MasPar, there has been an eightfold increase in the power of the processor with the same number of processors.

If the term "massively" refers only to the number of processors (independent of the other complex and far more important system considerations), it is misleading because current trends indicate that the industry is seeking a balance between numbers and power, not an unusual situation in the history of science. It is the simple pendulum effect. The efforts started with extremely powerful single processors that were internally highly parallel, followed by a multiplicity of small, lower power processors. We are now seeking a balance between the two extremes. Nothing could be more natural. This trend influences packaging dramatically and is consonant with the advantages of compactness and low cost in MCMs.

As clock periods have been reduced, there has been a trend toward paying ever more attention to impedance matching at all levels of the interconnect, within the module and between modules, including through the connectors. Although there has not yet been major concern with impedance matching on the chip, there is little question about the need to deal with this in the very near future for the majority of packaging techniques used today.

One current challenge is to build a balanced supercomputer processor consisting of about 10 million gates with a peak performance well over 1 gigaflop. The processor must also possess sufficient bandwidth to supply the functional units with multiple data words to and from memory at every clock period. Key systems decisions involving tradeoffs that must be made include such considerations as the use of custom chips housed in small MCMs, as opposed to gate arrays and sophisticated MCMs and combinations thereof. Custom logic is useful if there are few options per system and a mix of storage and logic is required. One must also assume that there is enough production volume to justify a return on investment from the nonrecurring costs of a custom approach. The success of a custom design requires compatibility between a suite of excellent computer-aided design (CAD) tools and a cooperative semiconductor supplier with advanced processes.

Requirements for CAD systems include the ability to handle thermal analysis, to interconnect designs, and to address mechanical considerations. Today one must consider memory versus smart memory or combined memory and logic on the same die. This is a function of the speed and on-off cycles of the chip, the kind of special functions being supported, the volume, and the level of cooperation from an integrated circuit supplier. In the near future, the industry expects 500 to 1000 input-output pins, diamond conduction cooling, liquid cooling, impedance-controlled MCMs, and impedance-controlled high-density interconnects.

We believe that the current supercomputer companies have the requisite systems-integration technology and the packaging experience as demonstrated by compact physical size, power, cooling, and interconnectability. It will be easier for such companies to move into the MPS market quickly than for the current supplier of MPSs to move to the sophisticated packaging required of the highest performance computers. The MPS supercomputer game, therefore, is for the incumbent companies to win or lose.

Workstation Clusters Rise and Shine

Bill Buzbee

Not very long ago, there was only one option for researchers interested in high-performance computing: the supercomputer. But these powerful machines are extremely expensive—so much so that only large research facilities can afford to buy and maintain them. Researchers at other locations can use these supercomputers by working over high-speed networks, but the number of users usually exceeds the available resources. Recently, a lower cost alternative to single-site supercomputing has become practical, with comparable performance: the workstation cluster.

Workstation clusters consist of an ensemble of workstations or high-performance microprocessor systems that are networked together in some fashion and that often appear to the user as a single resource. The equipment can be all of one type, or a mixture of different workstations and several different networks can be used. Potential benefits of workstation clusters include (i) a cost-effective alternative to mainframe systems, (ii) a cost-effective alternative to providing a workstation to each scientist and engineer in an organization, (iii) an approach to utilizing otherwise unused cycles on personal workstations, and (iv) loosely coupled parallel capability.

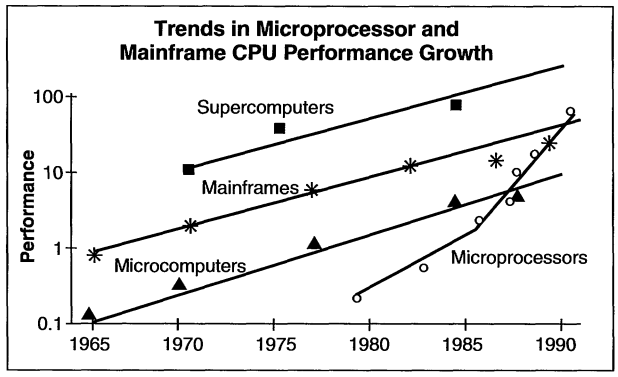
All of these benefits are a consequence of the steady and remarkable progress in very large scale integrated-circuit (VLSI) technology. The cost performance, measured in millions of floating-point operations (flops) per dollar, for top-of-the-line workstations has been growing at a compounded rate of 38% per year, in contrast with 10 to 15% for other systems (see figure) (*1*). It is no surprise that top-of-the-line microprocessors are sometimes referred to as "killer micros," owing to their tendency to devour other systems in the marketplace. Today, a top-of-the-line microprocessor often matches the scalar performance of a single central processing unit (CPU) in a supercomputer, and even in vector mode, a supercomputer CPU seldom outperforms a top-of-the-line micro by more than an order of magnitude. Also, thanks to progress in VLSI technology, microprocessor systems can be cost-effectively equipped with megawords of memory.

These technology trends combined with semiconductor standardization and high-volume production make possible microprocessor systems that cost much less than mainframe and supercomputers. The resultant cost performance advantages are the basis of growing interest in and use of workstation clusters.

A recent acquisition at Lawrence Liver-

The author is director of the Scientific Computing Division, National Center for Atmospheric Research, Boulder, CO 80307.

more National Laboratory (LLNL) provides one of the most dramatic examples of how clusters can provide a cost-effective alternative to mainframes and supercomputers (2). For several years, LLNL has operated an Open Computing Facility (OCF) that was originally built around a Cray X-MP/48. In 1991, the X-MP was directed to other applications. Recognizing the potential cost-effectiveness of workstation clusters, LLNL sought to purchase a system with a minimum level of performance rather than a particular type of machine. Specifically, LLNL requested a minimum throughput of 2.7 times that of an X-MP CPU and provided a suite of benchmark codes that ran on the X-MP. Also, a minimum aggregate memory capacity of 64 megawords (where one word contains 64 bits) was specified.



Going up. Trends in the growth of microprocessor and mainframe CPU performance. Cost performance is measured in millions of flops per dollar for top-of-the-line workstations.

The winner was IBM with a bid of 14 RS/6000 Model 550 workstations interconnected with a fiber-distributed data interface (FDDI). The annual operational cost of the X-MP–based OCF included 12 full-time people and X-MP maintenance charges of about $0.5 million (3). The annual operational cost of the cluster-based OCF includes two full-time people and maintenance charges of about $0.1 million. But equally important is that with suitable software, clusters can provide both interactive and batch computing to a relatively large number of users and do so in a more cost-effective fashion than if each user's office was equipped with a comparable workstation. This is because workstations in individual offices have a low utilization rate on a 24-hour basis. For example, the LLNL cluster has a total user population of over 300 people; typically, fewer than 50 of them are active at any point in time.

Fermi National Accelerator Laboratory (Fermilab) was among the first organizations to recognize the potential of workstation clusters. Fermilab began the development and use of clusters in 1984. Because of its pioneering effort in this area and in order to distribute and manage cooperating processes on clusters, Fermilab developed a software system called Cooperative Process Software (CPS) that is now in use at other sites. Fermilab also developed a capability called Unix Product Support (UPS) to simplify the distribution and maintenance of system and application software.

Today, Fermilab has over 150 IBM RS/6000 workstations and 200 Silicon Graphics, Inc., systems that are organized into about a dozen "second-generation" clusters (4). In aggregate, these clusters provide approximately 2.5 gigaflops of sustained performance in support of Fermilab research. Fermilab funds its cluster program on a constant budget of about $900,000 per year. Providing comparable sustained performance with supercomputers would require a substantially greater investment.

For some time, loosely coupled parallel processing has been performed on local area networks of workstations with software packages such as Parallel Virtual Machine (PVM) (5). Problems that can be solved in this fashion typically require relatively small amounts of interprocess communication. Clusters offer the possibility of providing high-speed communication links between workstations, and that in turn could significantly enlarge the set of problems that can be parallel-processed, namely "moderately coupled" applications. For example, the National Center for Atmospheric Research (NCAR) is interconnecting a cluster of four IBM RS/6000 Model 550s with serial optical links and a Network Systems PB290 router (6). Preliminary estimates are that this cluster, by means of parallel processing, can support real-time weather forecasting at a level comparable with one processor of a CRAY Y-MP.

The Institute of Electrical and Electronics Engineers (IEEE) Computer Society Technical Committee on Supercomputing Applications (TCSA) sponsors a Scientific Supercomputing Subcommittee (the "IEEE SSS"). Members of the SSS are drawn from organizations with leading edge computing capability. One of the objectives of the IEEE SSS is to track and assess developments in high-performance computing technology. Because of the rapid emergence of workstation clusters as a viable option for high-performance scientific computing, during 1992 the SSS made clusters of workstations one of its focal points. In the IEEE SSS discussions with laboratories using clusters, it was apparent that areas in which major technological improvements are needed include: (i) workstation input/output (I/O) reliability, (ii) I/O capability, and (iii) filing systems.

As an example of the need for greater I/O reliability, Nash (4) discussed an incident in which a small-computer system interface (SCSI) bus on a workstation would intermittently reset itself. This sort of unreliability is not appropriate in a production system used by many people. In particular, unreliable I/O causes users to be suspect of all computations performed. Also, file access and, thus, filing systems become a fundamental issue in the presence of unreliable I/O in a cluster (in fairness to workstation manufacturers, the current I/O bus system was developed for "stand-alone use," and it is sufficient for that task).

Organizations with applications that involve large amounts of data, such as NCAR, see a pressing need for high-bandwidth I/O channels that are compatible with mainframe channels. This reflects the potential cost-effectiveness of using leading edge workstations to process and manage large volumes of data. A number of manufacturers are planning to offer clusters as supported products, so these improvements should soon be available.

Because of the advantages discussed above, members of the IEEE SSS view workstation clusters as an extremely important development. These systems provide remarkable computing capacity at reasonable cost, they can provide a large community of users with both interactive and batch computing, and they can be used to support loosely to moderately coupled parallel processing. Consequently, members of the IEEE SSS expect that many scientific and engineering organizations will quickly adopt this technology for state-of-the-art applications. However, clusters cannot match the highest levels of performance afforded by supercomputers operating in multitask mode or of Massively Parallel Processors (MPPs). Thus, we expect that clusters will augment supercomputers and MPPs in that clusters can be used to handle moderately sized calculations whereas supercomputers and MPPs can be dedicated to large calculations that often characterize leading edge research. In fact, this trend is already evident at a number of government laboratories.

References

1. J. L. Hennessy and N. P. Jouppi, *IEEE Trans. Comp.* **1991**, 18 (September 1991).
2. C. D. Marsan, *Fed. Comput. Week*, 16 December 1991, p. 10.
3. E. Brooks III, "Cluster Computing at LLNL," presentation to the IEEE SSS, 12 August 1992.
4. T. Nash, "Cluster Computing at Fermilab," presentation to the IEEE SSS, 12 August 1992.
5. A. Beguelin, J. Dongarra, G. A. Geist, R. Manchek, V. Sunderam, "A users' guide to PVM: Parallel Virtual Machine," *Tech. Rep. ORNL/TM-11826* (Oak Ridge National Laboratory, Knoxville, TN, 1991).
6. D. Anderson, personal communication.