*view*point

Gordon Bell

Photo illustration by Robert Vizzini

1995 Observations on Supercomputing Alternatives: Did the MPP Bandwagon Lead to a Cul-de-Sac?

For over a decade, government and the technical computing community has focused on achieving a teraflop speed supercomputer. In 1989, I predicted this goal would be reached in mid-1995 for a $30 million computer by using interconnected, "killer" complimentary metal oxide semiconductor (CMOS) microprocessors [3–5]. The goal is likely to be reached in 1996 in a much more dramatic fashion than predicted because it is likely to be based on PC technology. Furthermore, by clustering PCs using System Area Nets (SANs), scalable computing can be widely available at low cost.

During 1995, Cray Research, Fujitsu, IBM, Intel, NEC, and Silicon Graphics introduced new technical computers. Intel announced the P6, a PC-compatible chip with a peak advertised performance (PAP) of 133Mflops to be raised to 266Mflops. In September, Sandia ordered a $45.6 million, 9,072 processor, 1.81Gflops computer using the chip scheduled to be installed in November 1996 that will provide 39Kflops/dollar or 1.2Tflops at the $30 million supercomputer price in 1989. Adjusting for inflation allows the 1996 supercomputer price to rise to $40 million and gets 1.6Tflops. Compaq Computer and Tandem Computers announced scalable computer clusters based on P6 for the commercial market. Dongarra's Survey of Technical Computing Sites shows that the world's top 10 have installed peak capacity of about 850Gflops, all of which contain hundreds of computers.

Teracomputer, an ARPA-funded state computer company, went public with an initial public offering to raise money to complete its computer. In the same period, Thinking Machines, a state computer company, and Kendall Square Research, which offered massive parallelism with over 1,000 processors, filed for Chap-

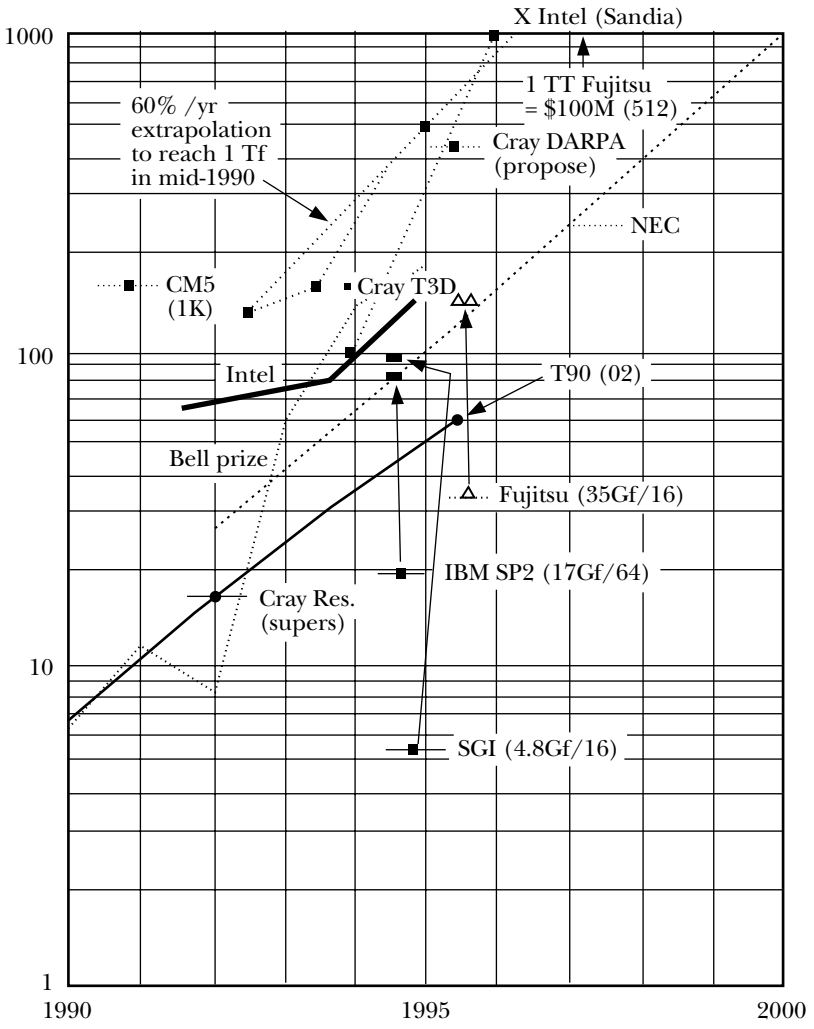
ter 11 but reemerged to offer software and systems based on interconnected workstations. In March, Cray Computer filed for Chapter 11, following the demise of ACRI, aka Stern Computer Company of Lyon, France. Convex, which uses Hewlett-Packard's PA-RISC chips, was bought by Hewlett-Packard. Other small companies making parallel computers are certain to fail, while other companies, such as Digital Equipment are still entering the market.

These events call for a look at how technical computing is now likely to evolve.

Five distinct computer structures are now vying for survival:

- Cray vector-style architecture supercomputers consisting of multiple, vector processors that access a common memory and build from the fastest ECL (emitter coupled logic) and GaA (gallium arsenide) circuit technology (Cray Research and NEC); Fujitsu and Hitachi have switched to CMOS but remain on this path.
- A computer cluster or multicomputer formed from single, fast vector processor computers connected via a fast, high-capacity switch (Fujitsu). The vector processor is implemented in CMOS technology. NEC has also announced a CMOS vector processor operating at 2Gflops per node that can scale to 512 processors.
- Headless workstation clusters, or multicomputers, formed from workstation "killer" CMOS microprocessor computers connected via SANs that are proprietary, high-bandwidth, low-latency switches; the IBM SP2 uses stacks of workstations. UC/Berkeley is building clusters using off-the-shelf Sun Microsystems workstations interconnected via Myrinet's high-bandwidth switch. Intel's Paragon is formed from specially packaged, CMOS microprocessor computers connected via its high-bandwidth, low-latency switch. Tandem and Compaq have introduced clusters for the commercial market using Tandem's ServerNet to interconnect Compaq 4 processor computers.
- "Multis," or multiple, CMOS microprocessors connected to large caches that access a common memory via a common bus (Cray Superserver using Sun SPARC micros, Silicon Graphics Power Challenge using MIPS micros) that I predicted to be computing's "mainline" structure [2] and have limited scalability of about 10, although Cray's Superserver uses 64 SPARC processors.
- Distributed shared-memory multiprocessors formed from workstation CMOS microprocessor or multimicroprocessor computers that communicate with one another via a proprietary, high-bandwidth, low-latency switch. Processors can access both local and remote memories as a multiprocessor (Convex, Cray). Silicon Graphics is fol-

Figure 1.
PAP* Gflops(t) for supers and MPP's for $30M (unless noted). Peak and # Proc. (in parenthesis)
*Peak Advertised Performance



lowing this path for scalability. Other companies are using the IEEE Scalable Coherent Interface (SCI), to build scalable multis with a single memory to simplify the operating system and apps porting.

Figure 1 shows performance measured in PAP for a $30 million expenditure, or roughly the cost of a supercomputer in the mid-1990s. Technical computing has evolved. Since 1990, ARPA's High Performance Computing and Communication Initiative (HPCCI) has stimulated the market by developing, purchasing and using highly parallel computers for scientific and technical computing. It is especially interesting to observe the effects of this effort as the teraflop quest continues.

From the details of the announcements and figure, I draw 13 major conclusions:

1. There is more diversity in computing alternatives than I predicted. While competition makes for lower hardware cost, it inhibits the attraction of apps software by independent software vendors. Cray (T90), Fujitsu, and NEC are continuing to evolve the supercomputer, utilizing existing apps. Fujitsu's multicomputer is a cost-effective hybrid of the traditional super that enables existing apps to run effectively and be evolved. Silicon Graphics is evolving the workstation and compatible multi with a wide range of apps. Convex, Cray, IBM, Intel, and nCUBE are all trying to establish massively parallel processing (MPP) as a viable computer structure. IBM is likely to be successful based on its ability to fund commercial apps. Intel's P6 microprocessor makes the PC the most likely candidate for the most cost-effective nodes in both the commercial and technical markets.

2. The computing industry has made impressive progress in developing parallel computers. More impressive is the fact that technical users have made progress in realizing the PAP for various apps as shown by the Bell Prize. The growth in apps performance by this measure has roughly doubled yearly, with the 1995 winner operating at 0.5Tflops using a specialized computer. The winning MPP operated at 179Gflops.

3. Price differences among the alternatives are often explained by differences in memory size and bandwidth. With computers, you get what you pay for. This rarely shows up in PAP, but appears downstream in RAP (real application performance) and occasionally on benchmarks. However, in 1995, most computers operated well on the Linpack benchmark, provided there was sufficient memory to scale the problem size and cover communication overhead.

4. CMOS has effectively replaced ECL and GaAs as the technology for building the highest-performance computers. Fujitsu's CMOS vector processor has a higher PAP than Cray Research's computers.

5. The Cray vector-style architecture is *not* dead to be replaced by multiple, slow CMOS workstation-style processors. The common wisdom within the U.S. academic community, which is the dominant receptor of research funding and sets the research and funding agenda, appears to have been wrong. The MPP bandwagon ran over vectors, replacing them with many interconnected "killer" micros used for workstations. These workstation micros are low cost and may be tuned for the benchmark de jour to provide high hype. MPP machines often perform poorly for problems where high bandwidth between processor and memory is required. It takes 8 to 10 of the fastest CMOS micros to equal a supercomputer vector processor in peak power. When used in parallel, power can be significantly reduced, depending on the computer (its memory and interconnectability) and problem granularity.

Most vector apps are unlikely to run on multicomputers for a long time. Silicon Graphics' multi is more likely to provide parallelism for fine granularity even though its scalability and memory bandwidth are limited. Silicon Graphics has the largest market share for technical computing, even though it is not the fastest. Convex, Cray, Fujitsu, and NEC are supporting traditional supers and MPPs. Since it is unlikely that MPPs based on CMOS micros can take over supercomputer workloads, the transition, if it happens at all, is certain to be costly. It is more likely CMOS micros will approach the speed of supers because supers trade off vector speed for scalar speed.

6. The prediction by NEC and me [4, 5] that a 1Tflop, classical multiprocessor supercomputer would not be available until 2000 still seems possible, even though the T90 supercomputer isn't quite on this trajectory. The difficulty is building a high-bandwidth, low-latency switch to connect processors and memories, since latency increases with bandwidth. A 1Tflop multiprocessor would require a switch of at least 16Tbytes per second to feed the vector units using the Cray formula.

7. No teraflop before its time. I predicted that a $30 million, 1flop computer would be available in 1995 [3–5], or by mid-1996 at the latest. The price of computation, using Thinking Machines' CM5 PAP as a reference, is only increased by 50% with Cray's T3D MPP. In 1992, I suggested waiting to purchase a $200 million 1Tflop ultra-com-

puter from Thinking Machines [4, 5]. Based on its characteristics and the inevitable progression of technology, I argued that we should wait until the system could be available at a price of only $30 million.

Intel's 1.8Tflops computer more than satisfies the wait. Intel provides a new high watermark in performance and performance/price. P6 offers the power of the fastest workstation micros at a "commodity" PC price level of less than $10,000. In this fashion, future MPPs are likely to be more heavily based on the X86 architecture.

8. Thinking Machines and other competitors vanished. Government subsidies affected the ability to function in a competitive, public marketplace. Larger companies have since entered the market, and only recently have significant apps appeared.

Government should stop the direct subsidy of computer design and associated targeted purchases. The best and perhaps only way I know of to develop an industry is through university research prototypes that go to start-ups or existing companies, and by the competitive purchase of new systems by leading-edge, government-funded users.

9. The price of supercomputers and MPPs has converged more than predicted. In 1992 the two differed by a factor of 10 and in 1996 the prediction is just three. More precisely, low-priced MPPs haven't materialized since Thinking Machines left the market. Better supercomputers may be due to competition and to better fabrication techniques.

10. Cray Research has placed three bets, including its mainline vector multiprocessor (T90), SPARC-based multi (Cray Superserver), and Digital Equipment's Alpha-based multicomputer (T3D,T3E). Cray has stated that it needs to converge its approach to parallelism and a common architecture. Convex is in a similar dilemma.

11. My prediction [4, 5] that MPPs will be built using a shared-memory multiprocessor architecture was optimistic. The multicomputer with multiple, independent computers interconnecting via a switch is the hardware structure for the foreseeable future to obtain the maximum peak power because it uses unmodified workstations. Software often manages and presents the structure as a single memory. Kendall Square Research provided the first scalable multiprocessors. Researchers are focused on the multiprocessor and have made progress. Other efforts are aimed at using SCI for building distributed shared-memory computers. The Convex and Cray (T3D,T3E) provide a shared memory but utilize it as a multicomputer. Silicon Graphics has a physically central memory for its multiprocessor.

12. In 1995, the world's fastest installation was a multicomputer. The Japanese threat continues to materialize with Fujitsu's VPP 300, which is is significant for a number of reasons:

- It is an engineering compromise between a classical Cray multiple, vector processor supercomputer and an MPP;
- It is cost-effective measured by peak performance and several real apps;
- As the fastest vector processor, it is likely to outperform other supercomputers for single processor tasks;
- As a multicomputer, it can function as n-independent computers to compete with supercomputers for workload;
- Because of a high-bandwidth switch and fastest nodes, it can outperform any of the MPPs, is more cost-effective than any of them, and has inherently lower overhead because fewer are needed;
- Having the vector architecture allows it to capitalize on the plethora of supercomputer apps developed over the last 20 years;
- It is aggressively priced (it is CMOS, and uses synchronous DRAMs) scales from a cost of less than $500,000 to a projected cost of $100 million for teraflops by 2000;
- The low entry cost and scalability increases its market size so that it will compete across the technical marketplace from workstations to servers to minisupercomputers and traditional supercomputers and the range of MPPs.

13. Berkeley's NOW (Network of Workstations) [1] project connects workstations through either an ATM or Myrinet switch [7]. PCs and workstations with 1 to 4 processors, no overhead backplane but limited PAP, are the most cost-effective to manufacture. IBM's SP2, based on uniprocessor workstations and its proprietary switch, belies this fact because its price of almost $100,000 per workstation is well above workstation-level prices. In contrast, Intel's system is only $10,000 per dual-processor node. Significant opportunities exist based on the PC.

NOW is important for many reasons, including having independent manufacturers for the network (switches) and platforms that permit multivendor environments. Over time, we expect low-cost SANs to emerge. Myrinet's and Tandem's ServerNet [8] are candidates for standard switches. Jim Gray and I are predicating the future of computing based on a small number of standards for the SNAP (Scalable Network and Platforms) architecture [6].

I believe funding university

purchases of NOW environments that either live in a single room or are distributed with users will prove to be a wise investment. The NOW structure will provide computing power and encourage the adoption of this paradigm. It would be desirable to have more standardized switches that are computer vendor independent and host multiple vendors. With a plethora of NOW environments, standards can form that will attract apps.

Conclusions

It is hard to be completely optimistic about U.S. supercomputing. It appears to be a small, vanishing market niched away by all kinds of computers. I see several options in addition to maintaining a "buy U.S." policy. Cray, IBM, Intel, and Silicon Graphics have large, loyal customer bases, many apps, and inertia. Intel offers the real bright spot by providing a powerful PC that can challenge any workstation. Suppliers have time to validate or rethink future product strategies. MPP apps are still difficult and will only get easier with fewer platform environments.

Government funders should ponder their role and question whether they helped or possibly misled companies, such as Cray Research, through funding and other pressures. The government's role going forward is still crucial.

The myriad of options should continue to keep the technical market vibrant (shaken up and alive) for a long time. The PC is likely to be the greatest change agent. It's a great time to be a user.

References

1. Anderson, T.A., Culler, D.E., and Patterson, D. A. A case for NOW (network of workstations. *IEEE Micro. 15*, 1, (Feb. 1995), 54–64.
2. Bell, C.G. Multis: A new class of multiprocessor computers. *Science 228*, (April 26, 1985), 462–467.
3. Bell, G. The future of high performance computers in science and engineering. *Commun. ACM 32*, 9 (Sept. 1989), 1091–1101.
4. Bell, G. Ultracomputers: A teraflop before its time. *Science 256*, (Apr. 3, 1992), 64.
5. Bell, G., Ultracomputers: A teraflop before its time. *Commun. ACM, 35*, 8 (Aug. 1992), 27–45.
6. Bell, G. and Gray, J. The SNAP (scalable network and platforms) architecture. Report, Mar. 1995.
7. Boden, N.J., et. al. Myrinet: A gigabit-per-second local area network. *IEEE Micro 15*, 1 (Feb. 1995), 29–36.
8. Horst, R.W. TNet: A reliable system area network. *IEEE Micro, 15*, 1 (Feb. 1995), 37–45.

Gordon Bell *is a senior researcher at Microsoft Corporation and a computer industry consultant-at-large.*