

Technical Computing Alternatives: Supercomputers to Ordinary Computers

According to Congress and the press, technical computing is in trouble. But in fact, the future looks brighter than ever. New classes of computers and new software are being created by new and existing companies. Only the growth rate for the traditional supercomputer might be slow.

The reason is straightforward. A user can often do the same computation on a powerful PC, on a workstation, on a minicomputer or microcomputer, on a superminicomputer, a graphics supercomputer or 3-D workstation, a minisupercomputer, a special-purpose supercomputer, or even a traditional supercomputer.

Users can trade execution time for cost because "Flops is Flops." The computer power necessary for technical computation can be substituted across this entire spectrum.

Technical computing is moving away from highly centralized, time-shared supercomputers. The same forces that operated in traditional computing will provide distributed, interactive computers to technical users. These computers offer adequate capacity and peak power for demanding jobs, are cheaper and easier to purchase and use, and offer superlative price/performance ratios.

A *general-purpose supercomputer* is a machine that, at the time of its announcement, costs more than other computers (perhaps \$5 to \$20 million), runs faster *in general*, has greater primary and secondary memory, and is suitable for all scientific, numeric problems.

Supercomputers have evolved along one architectural path. They all employ vectors and powerful multiple processors to gain speed. Fortran is by far the most important programming language; dusty deck programs are expected to port easily.

Automatic compiler tools (vectorizers and such) help users exploit the new machines. The more adventurous eventually reprogram using explicitly parallel techniques. Within a decade, Fortran will undoubtedly have parallel constructs.

By Gordon Bell
Corporate Vice President
and Chief Scientist
Stardent Computer Inc.



For both marketing and technical reasons, several new computer classes have appropriated the desirable tag *super*. They include:

- Minisupercomputers (originated c.1983) \$0.1 to \$2 million.
- Graphics supercomputers (c.1988) \$50 to \$200 thousand with significant 3-D graphics.
- Super workstations (c.1990) \$25 to \$50 thousand with significant 3-D graphics.

To deserve the *super* tag, a machine should be substitutable for a supercomputer. Substitutability for a user or group of users requires three things:

- The machine must provide in one day the same computational performance as could be expected from the largest time-shared supercomputer economically available to that user.
- The machine must have the capacity to sustain high throughput on a wide range of jobs.
- The machine must be the best performer in its price class.

For most people, a machine with 5 to 10 percent of the power of a CRAY and sufficient memory and disk capacity for the normal run of jobs would constitute a *super-substitute*. Such a machine delivers the equivalent of between one and three hours of CRAY service a day — more than all but a handful of the nation's supercomputer users receive.

We can now examine possible *super contenders* to see how well they fit the requirements. The second requirement has a hidden catch in it as well. The range of jobs includes both scalar and vector/parallel programs. A contender must do well on both.

A mainframe with vector facility might have the computational style and capacity necessary for the super tag, but typically such a machine cannot provide the peak performance required. These mainframes are never the best performers in their price class (at least for numeric computing).

Jack Worlton, in an article in the September issue of *Supercomputing Review*, says *wimp supercomputers* provide plenty of computational

power but miss the peaks of best machines available. Such machines are often the supercomputers of an older generation; a CRAY-1 is a wimp because it is two generations old.

Users stuck on wimps might be waiting 24 to 48 hours to get one hour of compute time. They might be paying charges per megaflop much higher than they would for newer machines because the center is still writing off the wimp on a five-year basis.

Special-purpose or monoprogrammed supercomputers utilize a large number of low performance processors (100 to 1,000) or processing elements (10,000 to 100,000) and provide very high peak power on selected applications. There are two main varieties. Multi-computers connect processor and memory pairs to work on one program. Transputers, hypercubes and circuit-switch connected machines fall in this class.

SIMD machines pass a single instruction stream to thousands of processing elements; the Connection Machine is in this class.

Table 1: Power of 1989 Technical Computers in Megaflops/Second

Type	#Proc. Max	LFK per Proc.	LFK per Machine	Linpack 100x100	Linpack 1000x1000	Peak
PC	1		0.1 to 0.5	0.1 to 0.5	0.1 to 1.0	
Workstation	1		0.2 to 1.5	0.5 to 3.0	6	8
Micro/Mini	1		0.1 to 0.5	0.1 to 0.5	0.1 to 0.5	2
Supermini	6	1	4	1	6	24
Graphics Super	4	1.5 to 5	10	6 to 12	80	128
Minisuper	8	2 to 4.3	10	6 to 16	166	200
Main/Vectors	6	7.2	43	13	518	798
Supercomputer	8	19	150	84	2,144	2667

Applications must be reprogrammed to exploit these machines fully, and such programs can run faster than on a super. But the need for reprogramming and slow scalar speeds causes these machines to fail the workload test for serial job streams. They are not directly substitutable for current, general-purpose supercomputers.

Super multiprocessors (developed by BBN and Evans & Sutherland) have 100 or more processors, a common memory and a single job pool controlled by one operating system. The sheer number of processors guarantees good throughput performance, but the relatively slow individual processors will find it difficult to run scalar programs at supercomputer speeds. Still, these machines might be the highest capacity, general-purpose, cost-effective systems that can be built today.

Ordinary computers provide the most technical computing power today. This includes PCs, which are evolving toward the power of 1-D, 2-D and 2½-D workstations; 3-D high-perform-

ance workstations; microcomputers; and superminicomputers. Many of these have impressive scalar performance, but they have no way to hit performance peaks for those programs that make good use of the vector or parallel capabilities of supercomputers.

Table 1 summarizes the power of various technical computers in 1989. The machines in the bottom half are capable of providing shared supercomputer power. The column labelled #Proc. Max describes the parallelism available. The columns headed LFK give performance on the Livermore Fortran Kernels, a good synthetic mixed workload for scientific machines. These numbers measure the throughput one might obtain in a typical scientific environment.

The columns headed Linpack measure performance on the Linpack linear equation solving test. The 100x100 test shows the rates that might be achieved by normal users of reasonable supercomputer programs, and the 1000x1000 test shows rates that can be achieved by real

programs after considerable tuning.

Notice that peak rates (sometimes listed at about 20 gigaflops) are not shown for the very large machines because no real programs come anywhere near the peaks. By significant tuning to run in parallel, programs operate at over one-half the peak.

WHICH COMPUTER IS THE BEST FOR THE APPLICATION?

Market substitution occurs across all computers. Users have a fixed budget to trade off across the complete range. Computer choice depends on many factors besides purchase price and peak performance: software availability, ease of purchase, installation and use; apparent lifetime; rate of technological change; past and future compatibility; control in the allocation and management of resources; programming knowledge needed; even the machine appearance or the prestige

Table 2: Installed Capacity for Technical Computing (Dataquest)

Type	Dataquest Installed	Ships	'89 LFK Capacity	Companies		
				Selling	Building	Dead
PC	3.4M	1M	1341	100s	?	?
Workstation	0.4M	290K	580	7	?	~50
Micro/Mini	0.9M	51K	30	~20	?	~100
Supermini	0.3M	7.5K	100	7	?	~10
Graphics Super	10.5K	13.6K	182	2	2	2
Minisuper	1.6K	600	32	5	>2	8
Parallel Proc.	365	250	4	24	>9	8
Main/Vectors	8.3K	1600	46	3	?	3
Supercomputer	450	130	100	4	>3	3

of owning a particular computer.

Table 2 shows the installed capacity for technical computing. The most important column is '89 LFK Capacity — the power available to run general-purpose technical workloads measured in units of CRAY Y-MP/8s. Most of the power is provided by machines that are not super anything; even much of the supercomputer power is provided by wimp machines at least one generation old.

Computation on the wrong system is costly and inefficient. There is a supercomputer at one national laboratory making a relatively trivial calculation for contractors throughout the United States. The supercomputer produces a picture, compresses the data, sends it over a slow but expensive network, and graphics workstations recompute the image for static display. The computation and display could be done on a powerful workstation in roughly the same elapsed time without the network or supercomputer. The computer exists to support a super bureaucracy.

Let us compare an Ardent Titan III/2 (two processors) with a CRAY Y-MP processor. Titan III was introduced one year after the CRAY Y-MP and delivers about 9 megaflops LFK throughput. The Titan's throughput is about 1/2 that of a Y-MP processor for LFK, its 100x100 Linpack rate is about 1/6 that of a Y-MP processor; and its peak performance is about 1/50 that of an eight-processor Y-MP.

A Titan III/2 costs about 1/20 what a one-processor CRAY Y-MP does. Furthermore, the Titan III's speed doubled (and its price dropped) from its predecessor in only 18 months. The shortest conceivable gestation for a supercomputer is three years, and five to seven years is more likely.

WILL SUPERS' GROWTH RATE DECLINE LIKE THE MINI AND MAINFRAME?

Supercomputers will continue to evolve more rapidly than minicomputers and mainframes and the performance gap between the latest

generation of supercomputers and the last generation will continue to widen. However, alternative computers that are *fast enough* will continue the trend to distributed computing even for applications that were previously *super-applications*.

Architecturally, the reasons are easy to see. Supercomputers require expensive, high-speed components, elaborate processor-memory connections, very fast, large disks, processing circuits that do relatively few operations per chip and per watt, extensive installation and high operating costs; worse, they have little architectural scalability.

Supercomputer buyers must have great needs, great dedication to the support of the machine and great budgets. For almost all users, economics inevitably dictate the purchase of smaller systems connected in networks.

In addition to the pure economics, current supercomputers lack the visualization capability and interactivity found in distributed computing. Networks coupled with workstations are not adequate to provide the same capability. For example, the use of spread sheets, drawing programs and even word processing is qualitatively different using terminals connected to time-shared computers through LANs, in comparison with the use of personal computers.

BASIC SHIFT TO INTERACTIVE AND DISTRIBUTED COMPUTING

A significant change in computing styles is occurring in providing truly interactive design and analysis using high-performance, super workstations and graphics supercomputers. These provide more than 10 percent of a CRAY Y-MP processor capacity and peak power, are often three to seven times cheaper per delivered megaflop, are purchased and installed by a single user or small group, and overcome many of the disadvantages of supercomputers.

Mechanical and petroleum engi-

neering, chemistry and biochemistry, fluid dynamics and medical image processing are being transformed by the newly available distributed power. A similar change occurred about a decade ago when workstations were introduced into the design of digital systems and chips.

All the supercomputers and alternates are highly parallel — using vector processors, parallel CPUs or multiple processing elements. Computer scientists and applications specialists must cooperate to understand and make use of parallel computing; institutions must respect and encourage research and teaching in these areas.

The arts and sciences of visualization must be propagated into the technical computing community. The cost in new software will be more than repaid by the value of new insights. If the technical computing community can meet these needs, our future is bright. **SR**

How To Make The Dollar Stronger Overseas.

Give to CARE. 95% of every dollar we receive goes to help impoverished people in the third world.

Small wonder that we're the best run, best managed charity in America.

CARE

We're Helping People
Learn To Live Without Us
1-800-242-GIVE