(20)

Spectrum AUTHOR'S COPY

Decision Resources, Inc.

Bay Colony Corporate Center
1100 Winter Street
Waltham, Massachusetts 02154

Telephone 617.487.3700
Telefax 617.487.5750

Information Systems Industry

The Outlook for Scalable Parallel Processing

Gordon Bell
Consultant to Decision Resources, Inc.

Business Implications

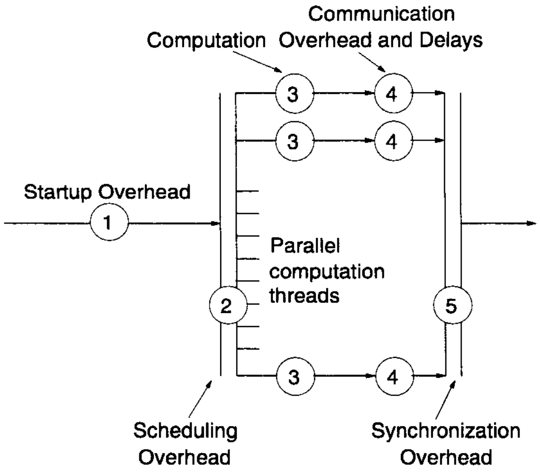
- Scalable, massively parallel processing computers promise to become the most cost-effective approach to computing within the next decade, and the means by which to solve particular, difficult, large-scale commercial and technical problems.
- The commercial and technical markets are fundamentally different. Massively parallel processors may be more useful for commercial applications because of the parallelism implicit in accessing a database through multiple, independent transactions. Ease of programming will be the principal factor that determines how rapidly this class of computer architecture will penetrate the general-purpose computing market.
- Vendors that succeed in developing general-purpose scalable parallel computers have the opportunity, by early in the next decade, to be able to address the computer systems market, including most of the traditional roles of mainframes and supercomputers and today's specialized scalable computers.
- The direction offering the most promise for scalable parallel processing computer development involves the use of standard processing and networking elements and programming environments and ensuring compatibility with traditional multiprocessors, workstations, and PCs.

It is likely that this decade will usher in the beginning of an era in which general-purpose[1] scalable parallel computers assume most of the applications currently run on mainframes, supercomputers, and specialized scalable computers. A scalable computer is a computer designed from a small number of basic components, without a single bottleneck component, so that the computer can be incrementally expanded over its designed scaling range, delivering linear incremental performance for a well-defined set of scalable applications. General-purpose scalable computers provide a wide range of processing, memory size, and I/O resources. Scalability is the degree to which performance increments of a scalable computer are linear. Ideally, an application should be usable at all computer size scales and operate with constant efficiency.

Parallel computers are defined by their ability to share or communicate data among multiple processors. Figure 1 shows the basic structure of a parallel computation. The computation starts with a sequential thread (1) that includes job scheduling and other serial computation. A basic loop starts with supervisory scheduling (2) followed by the computation (3) and intercomputer message (4) phases of a thread. Synchronization (5) occurs prior to returning to scheduling the next unit of parallel work (2). The length of time until a computation thread must synchronize with another parallel thread indicates the granularity of a parallel structure.

[1] Test for general purposeness: Can the computer efficiently process a wide range of jobs (including a workload consisting of sequential to parallel processing, small to large job sizes, short to long runtimes, and interactive to batch response times) requiring a variety of processing, memory, database, and I/O resources?

Figure 1
The Basic Structure of Parallel Computation



Source: Gordon Bell.

The most basic parallelism is using multiprogramming at the workload level, where a common pool of computational resources (processing, primary and secondary memory, and networking) is available to trade off among a large job mix with varying degrees of parallelization (including completely scalar operations). For peak performance of a single job, two forms of parallelism may be required:

- Transparent (or implicit) parallelism in which the computer breaks a job into parallel computational threads without intervention by the user, and
- Explicit multiprocess parallelism in which the user is required to formulate a job in terms of both functional and data parallelism.

Evolvability (i.e., generation or technology scalability) is the ability to implement a follow-on computer of the same family using faster components. Evolvability is an essential property of a scalable parallel computer because of the time and financial investment required to develop parallel programs. It requires that all rate and size metrics (such as processing, memory and I/O bandwidth, memory size, and especially interconnection bandwidth) increase proportionally from generation to generation.

The Software Driver

Computers that are used for a single problem, function, or workload can be built to scale over a range of several thousand processors; they are limited only by systems software and applications. The transition from what currently exists to the scalable parallel computer systems of the future will not be automatic, however, because of the difficulty in establishing standards for parallel processing, which enable applications to run efficiently on a range of parallel machines. Only when standards have been established, standards to which all manufacturers adhere, will software applications for scalable parallel computing truly flourish and drive market growth.

Scalable parallel computers have evolved from two independent and distinct application directions based on two different sets of requirements: technical (i.e., scientific/engineering) and commercial.

Technical Applications

Technical applications are based on floating-point operations used in analysis, simulation, and design. Technical applications focus on achieving the greatest number of floating-point operations per second (FLOPS), although some technical applications, such as genome sequencing, are fundamentally database-oriented. Most of the fundamental understanding about parallelism has been derived from attempts to provide highly parallel technical computers.

Evolvability is an essential property of a scalable parallel computer.

Two basic programming paradigms are used for technical computing: data parallel and multiprocess. In the data parallel approach, a FORTRAN dialect (such as FORTRAN 90, High Performance FORTRAN [HPF], or just FORTRAN 77) is used with multiple copies of a single program that operate on multiple data items in parallel (called SPMD).

The multiprocess approach, as in FORTRAN M, uses a program that is divided into subproblems and distributed among the nodes that communicate by explicit message passing. Multiprocess applications can be divided by function (i.e., different processes handle different types of tasks) or by data (i.e., different processes handle different data). Ordinary operating system mechanisms such as pipes, sockets, and threads facilitate parallelism by providing communication among and within processes. Programming environments that operate on all computer structures, including networked PCs and workstations, have been developed for multiprocessing. They include Oak

Ridge National Laboratory's Parallel Virtual Machine (PVM), Scientific Computing Associates' Linda, Parasoft's Express, and various programs (for example, IBM's LoadLeveler) that can manage a computer cluster as a single facility.

Commercial Applications

Commercial applications are usually database-centered for transaction processing and database analysis. Transaction processing is implicitly parallel, and many customer-specific applications are easily portable because of the nature of the interface and implicit parallelism. Once a database port has been made, many uses are possible because the database is parallel. Data analysis or "data mining" is organized to utilize the parallel access to a single database. Because data analysis is not typically considered mission-critical, it has been the entry point for parallel applications in commerce.

The first parallel computers for the commercial market were from Tandem and Teradata.[2] In these systems, a transaction-processing monitor operated on a number of independent transactions using a variety of applications, which were distributed within the nodes of a scalable computer cluster. Transaction processors usually access a single database, which is written in such a way that it runs in parallel on the independent computing nodes.

Ironically, commercial applications are more likely to be parallelized than technical applications are because (1) parallelization is implicit once a back-end database (e.g., Informix, Oracle, and Sybase) has been parallelized (i.e., it can access all disks in parallel) and (2) multiple, simultaneous transactions that access the database are parallel. In data analysis or decision support applications, the database is simply mined in multiple ways in parallel to generate data for further analysis and additional reports.

Parallel Programming Environments

Although spectacular increases in performance derived from microprocessors are noteworthy, perhaps the greatest breakthroughs for parallel processing have come from software environments such as Linda, PVM, and Express together with parallelizing compilers. These products permit users to structure and control a collection of processes (using message passing) to operate in parallel on independent computers. Linda, for example, enables a set of computers to view a set of objects stored in a common, virtually shared memory that any processor can symmetrically access. Linda handles only the coordination functions, which include establishing the common memory space, process creation, interprocess communication, and control. All objects can be run in parallel under the right controlling circumstances. The base language, such as C and a FORTRAN dialect, acts in a normal fashion, while Linda adds four functions—*in, out, read,* and *evaluate*—to the language.

User interface software, debuggers, performance monitors, and many other tools are part of these basic parallel environments. New sets of tools that treat a cluster of workstations as a single entity and then allow users to utilize the cluster in parallel for a variety of tasks have been recently introduced by IBM, Platform Computing, and Scalable Technologies.

For multiprocessors, small degrees of parallelism are supported through such mechanisms as multitasking and Unix pipes in an explicit or direct user control fashion. Linda extends this model to manage the creation and distribution of independent processes for parallel execution in a shared address space.

Medium (10-100 processors) and massive (1,000+ processors) degrees of parallelism for a single job can be carried out in either an explicit message passing or implicit fashion. The most straightforward implicit method is the SPMD model for hosting FORTRAN across a number of computers. Recent FORTRAN translators enable multiple workstations to be used in parallel on a single program in an evolutionary fashion. Furthermore, a program written in this fashion can be effectively used across a number of different environments from supercomputers to workstation networks. Alternatively, a new language that has more inherent implicit parallelism, such as dataflow, could evolve; however, no candidate is on the horizon.

Current Scalable Parallel Computers

The current generation of scalable parallel computers is based on four independent lines of architecture development.

Shared-memory multiprocessors, in which two or more processors share a common memory, have evolved over the last 30 years and have become the main line

[2] A former technology partner of NCR, Teradata was acquired by AT&T shortly after its purchase of NCR. It is now part of AT&T Global Information Solutions, the new name for AT&T's computer systems business.

of computing. Product introductions by Convex, Cray Research, and Kendall Square Research have demonstrated that scalable shared-memory multiprocessors with logically centralized but physically distributed memory are feasible. Given this development, shared-memory multiprocessors are likely to continue as an important architecture.

Scalable multicomputers and scalable computer clusters (sometimes referred to as "shared-nothing" systems) are a collection of an arbitrary number of independent computers, each of which runs its own copy of the operating system, and are connected using either a proprietary switch or a network switch such as asynchronous transfer mode (ATM) or Ethernet. Scalable multicomputers and computer clusters supplied by Intel, Meiko, Parsytec, nCube, and Thinking Machines have been the basis for developing technical parallel computing technology, and Teradata's multicomputer has provided the basis for commercial parallel computing development. A multicomputer can simulate shared-memory multiprocessing. As the scalable multicomputer evolves, it will continue to develop characteristics of shared-memory multiprocessors along the lines of computers from Cray Research and Convex. IBM's SP1, using RISC-based RS/6000 headless (no monitor) workstations and running a scalable version of IBM's AIX (Unix) operating system, is likely to be the archetype of this form of scalable parallel computers. However, SP1 will have to significantly reduce latency to compete with scalable multiprocessors. Table 1 shows the basic differences between multiprocessors and multicomputers based on a number of attributes.

Networked workstations that communicate along a slow local area network (LAN) by passing messages but share little or nothing in terms of memory, I/O, and so on, are scalable. However, they have little to no ability to handle a workload distributed among the nodes or a parallel task because of the long latency, low bandwidth, and high software overhead involved in message passing. Fortunately, these deficiencies can be remedied. As standard, fast switches become more cost-effective and more widely available over the next 3-4 years, then scalable, networked workstation clusters will most likely replace multicomputers that are built from proprietary nodes and switches and use unique software.

Single instruction multiple data (SIMD) computers are considered to be massively parallel because several thousand processing elements operate in parallel (controlled by a single instruction), but their scalability is limited. The Cray-style supercomputer vector processor is a form of SIMD, but with limited parallelism.

SIMDs are limited by sequential problems, but for problems that are highly data parallel (e.g., signal and image processing and certain database operations), a SIMD may perform exceptionally well. MasPar is the leading vendor of SIMD computers. However, many SIMDs are provided as a computer attached to a workstation, a configuration that provides cost-effective technical computation. Adaptive Solutions, Alex Parallel Computers, HNC (SNAP), Mercury Computer Systems, Microway, and Sky Computers all provide an array of attached processors that connect to various workstations and provide exceptional processing power. The HNC SNAP-64 has a peak announced performance (PAP) of 2.56 gigaflops (GFLOPS) at a price of $90,000. Some of the technical applications (e.g., neural simulation, signal processing, and image processing) can be effectively carried out using these workstation-attached processors.

Table 2 gives the general characteristics for a representative sample of each scalable parallel computer architectural type and our view of their strengths and weaknesses. The following section describes these computers in more detail.

Scalable Shared-Memory Multiprocessors

Convex Exemplar. The Convex Exemplar uses a fast switch to interconnect up to 128 Hewlett-Packard (HP) PA-RISC processors. The PAP for a 128-processor system is 25 GFLOPS. Memory is scalable to 32 GB of globally shared physical memory. The Exemplar SPP design has four goals: (1) provide a fast switch so that the nodes appear as a single, shared memory; (2) run FORTRAN 77 supercomputer programs without modification (through automatic parallelization), thus not forcing users to convert programs to HPF (high-performance FORTRAN); (3) offer a scalable system that is no more than 15% more expensive than comparably priced workstations; and (4) support the use of unmodified, binary, single-threaded HP PA-RISC/HP-UX applications.

Cray Research T3D. In September 1993, Cray Research announced its development of the Cray T3D with up to 2,048 150 MFLOPS Alpha-based (from Digital Equipment) computing nodes organized as a shared-memory multiprocessor; that is, any node can directly access the memory of another node using the

Table 1
Attributes of Parallel Multiprocessors and Multicomputers

Attribute	Multiprocessor	Multicomputer
Control of memory consistency	Single, sequential consistent memory supported by hardware	Controlled by overhead software (if at all)
Access to data and programs	Equally accessible to all processors	Allocated among computers; accessible through software
Data communications	Implicit by directly accessing memory	Explicit message passing (may be hidden from user by hardware or compiler)
Resource management	Fungible	Controlled by operating system
Work management	Work queue accessible by any processor	Work is moved as load on computer nodes changes
Exploit memory locality	Automatic mechanism to implicitly control and exploit locality	Nonlocal access requires software for address translation, message passing accesses, and memory management
Function in general-purpose fashion	Inherently general-purpose	Works best in independent, statically determined partitions that run to completion
Handle large jobs	Any node may run any size job	Limited by node's memory size
Achieve parallelism	Provide standard programming environments for rapid porting of applications	Two approaches: 1. New dialects of C and FORTRAN with explicit data management statements 2. Explicit message passing that requires new programs and algorithms

Source: Gordon Bell.

T3D's high-bandwidth, low-latency network. Nodes in the T3D are interconnected via a 3-D torus topology. The computing nodes have substantial hardware to facilitate parallel processing and lower latency, including block transfers, pre-fetch and post-store of data, barrier synchronization, loop scheduling, and so on.

The initial T3D requires a Cray host supercomputer for I/O and management. Each node is controlled by a microkernel that carries out a task or calls the host supercomputer. The initial programming model assumes explicit message passing and includes PVM. Subsequent software will include Cray's MPP FORTRAN.

Kendall Square Research KSR2. In 1993, Kendall Square Research introduced the KSR2 scalable shared-memory multiprocessor. The structure and programming model consists of up to 5,000 or more processor nodes that access a common memory. Each node operates at a PAP of 80 MFLOPS and comprises a 32 MB primary memory and a 64-bit superscalar processor (e.g., IBM RS/6000).

The KSR2 is similar to a multiprocessor mainframe because it is general-purpose, runs a single operating system, and can allocate any of its resources to a common workload. Unlike a mainframe, however, the KSR2 is scalable from 32 to over 5,000 processors in a

Table 2
Scalable Parallel Computers

Vendor/ Model	Type	Use	Scaling Range	Performance per Node in MFLOPS	Processor Architecture	Strengths	Weaknesses
AT&T 3600	Multicomputer	Database	2-1,024	NA	X86	Teradata computers provided experience and customers; focus on evolvability and compatibility; uses Intel micros; moving to use of standard databases (SQL); applications multiprocessors run Unix; one node type (in the future)	Poor ability to deliver products in timely fashion; three culture architecture: AT&T/ NCR/Teradata; proprietary database to support; no benchmark data yet available
Convex Exemplar	Scalable multiprocessor	Technical	4-128	198	PA-RISC	Uses PA-RISC, HP U/X, and many HP workstation applications; understands supers, compilers, and applications; applications on HP workstation farms; shared memory programming model	Convex-unique nodes vs. HP workstations, lack of parallel applications
Cray T3D	Scalable multiprocessor	Technical	32-2,048	150	Alpha	Understands supers, compilers, and applications; shared memory program model; host supercomputer provides full generality; Alpha architecture; becoming a state computer[a] vendor	Alpha architecture: iocompatible with O/S and applications; requires a host supercomputer
Digital Workstation Farm	Networked workstation	General purpose	2-100	Varies based on specific workstations in farm.	Alpha	High-speed Alpha architecture; 64-bit address; supports heterogeneous systems	Lack of volume and scalar applications
Fujitsu VPP 500	Multicomputer	Technical	4-222	1,600	Vector Processor	Fastest vector processing nodes, can be used as independent supercomputers; evolutionary	Not VP compatible; not CMOS-high cost/FLOPS and cost/MB
IBM SP2	Multicomputer cluster	General purpose	4-128	266	POWER architecture and POWER2 microprocessor	IBM salesforce and large customer base; IBM commitment and understanding about parallelism; POWER2 microprocessor and fastest nodes—uses workstation nodes, many compatible vertical market applications	Multicomputer must evolve to shared memory program model; lacks state computer[a] imprimatur

(continued)

Vendor/ Model	Type	Use	Scaling Range	Performance per Node in MFLOPS	Processor Architecture	Strengths	Weaknesses
Intel Paragon	Multi-computer cluster	Technical	2-1,000	75	i860	Large company can sustain market development; early MPP vendor and installed base for upgrades; built distributed OSF; Unisys as a commercial partner; large customer base; switch is upgradable for next generation; a state computer[a] vendor	Dead-end i860 nodes; message passing FORTRAN requires a rewrite of applications; poor RAP/PAP (high software overhead, poor nodes)
KSR2	Scalable multi-processor	General purpose	32-5,000+	80	KSR/Series	KSR processor architecture provides shared memory program model (based on the ALLCACHE memory-management architecture) that all systems may all evolve to; general purpose for technical and commercial	KSR-unique architecture; lacks state computer[a] imprimatur that provides user base with software assistance and applications
MasPar MP-2	SIMD	Technical	1,000-16,000	0.15	MP-2	Simple SIMD programming model, effective for highly parallel jobs	Limited scaling range; not general purpose for jobs or workload; must find point applications
Meiko CS-2	Multi-computer	General purpose	4-1,024	200	SPARC+ vector processor	SPARC and Solaris compatible with Fujitsu vector processing, switch performance; ability to run Sun applications	No performance data; company is very small to attack multiple markets
nCube	Multi-computer	Database and video	8-8,192	4.1	nCube	Early MPP vendor and large installed base; Larry Ellison's ownership ensures Oracle database and applications; poor floating point focuses nCube on commercial market; company working on video server	Proprietary nodes; expensive to maintain proprietary O/S as nodes evolve, few non-Oracle applications
NEC Cenju-3	Multi-computer with multi-processor functions		8-256	50	Mips	Mips architecture and implementations	Mips micros have limited MFLOPS; limited experience
Silicon Graphics Challenge Array	Multi-computer	Technical	n x (2-36) (n=number of nodes)	75	Mips	Large memory and shared-memory program model; independent CPU, memory and I/O scalability; compatible with workstations and their applications	LAN connection with long latency limits types of problems that can be solved effectively

(continued)

Vendor/ Model	Type	Use	Scaling Range	Performance per Node in MFLOPS	Processor Architecture	Strengths	Weaknesses
Power Challenge Array	Multi-computer	Technical	n x (2-18) (n=number of nodes)	300	Mips	Same as ChallengeArray	Same as ChallengeArray
Thinking Machines CM5	Multi-computer	Technical	32-16,000	160	Super SPARC+ vector processor	Early MPP vendor and installed base for upgrades; SPARC front-ends; simple to use SPMD compiler and data programming model; RAID for data mining applications; a state computer[a] vendor	Incompatible SPARC and TM floating point unit = unique nodes; not general for jobs and workload; poor scalar; poor fine grain

[a] "State computer" companies are those that have significant direct government support of their research and development.

Source: Gordon Bell.

3-level hierarchical structure. Each set of 32 processors can support up to 500 GB of disk storage; thus, disk capacity can grow to 160 terabytes. A 1,088-node system provides almost 30 times more processing power, primary memory, I/O bandwidth, and mass storage capacity than a multiprocessor mainframe.

Scalable Multicomputers and Multicomputer Clusters

Fujitsu VPP 500. Fujitsu's VPP 500 supercomputer is a medium to coarse grain, asymmetrical (inhomogeneous) multicomputer with 4-222 1.6 GFLOPS vector supercomputer nodes, each with a 256 MB memory, interconnected via cross-bar switch. Because the nodes are so powerful, a factor of 10-20 fewer nodes can achieve the same level of performance as a computer using CMOS microprocessors. A configuration of 64 nodes achieves 100 GFLOPS. The fast nodes require a lower-latency, lower-overhead switch than is needed for microprocessor-based multicomputers. The 800 MB/sec, low-latency cross-bar switch and interface manage process-to-process data transmission without processor intervention.

VPP's principal advantage is that it can achieve incredibly high throughput by using a single node; thus, it can be used effectively as a workload computer that requires little or no parallelization beyond vectorization. Because the computer is built using relatively expensive circuit and packaging technology (including gallium arsenide), it is very compact. The small node memory may prove to be a serious limitation, however.

Intel Paragon. The Intel Paragon is a symmetrical (homogeneous) multicomputer with up to 1,000 nodes interconnected by a fast 2-D mesh. Compute nodes consist of an i860 microprocessor,[3] which achieves a PAP of 75 MFLOPS, and a separate i860 microprocessor to handle communication or additional computation. (Older software does not utilize the second processor.) Compute nodes can each support up to 32 MB of memory. Larger service processor nodes handle I/O and user interaction. Paragon is controlled by the micro kernel-based OSF/1 (Mach) operating system. Software parallelization is left to the user by employing explicit message passing.

Although it was introduced in 1991, few benchmarks, applications, and performance data are available for the Paragon. In May 1994, a Paragon XP/S 140 achieved 143.4 double-precision GFLOPS on the Massively Parallel LINPACK benchmark—the highest number ever achieved. However, the relatively small amount of node memory defines a limited computer that requires significant evolution to be useful. Paragon's PAP does not imply significant real application performance (RAP) as shown by NAS benchmark data. A poor RAP/PAP ratio is a result of the i860 architecture, nodes that have insufficient memory, and internode communications overhead.

[3] The Intel i860 was introduced as a desktop supercomputer for graphics processing and highly tuned applications that could be carried out with a small cache and could tolerate long context switching times.

Intel will likely be using X86-based chips in subsequent Paragon systems, giving them a commercial orientation. As a result, we expect that the i860 product line will be discontinued. Hence, evolvability using a compatible architecture for the technical marketplace is yet to be determined.

Intel has an agreement with Unisys to provide "system building blocks" with which Unisys will develop a scalable parallel processor based on the mesh interconnect subsystem using Pentium processors. Unisys is porting Unix and related software for the system for the commercial market, but initial shipments are not scheduled until 1995. Intel also announced an agreement with Microsoft in May 1994 in which Microsoft will offer its Tiger videoserver software on a Pentium-based system with the Paragon interconnect.

Meiko CS-2. The Meiko CS-2 is a symmetrical multicomputer for both the technical and commercial marketplaces. It supports up to 1,024 processing elements (a large printed circuit board) in four expandable configurations of 16, 64, 256, or 1,024 elements. An element can be one of three types: a SPARC processor and two 100 MFLOPS double precision (200 MFLOPS single precision) vector processors, a SPARC processor and I/O channels, or four SPARC processors. A SPARC processor operating at 50 MHz provides a PAP of 150 MIPS, 50 MFLOPS, or 80 SPECmarks. Four elements are interconnected to form a module (a small cabinet) and modules are interconnected to the backplane network switch.

The network and node-to-node interface is a significant feature because it provides fast task-to-task bandwidth (100 MB), low latency (1.4 microseconds), low processor overhead (1 microsecond/message), and the ability to directly load/store data at remote nodes. The architecture provides n+1 redundancy and fault tolerance. Each node runs Sun's Solaris operating system, enabling compatibility with Solaris applications, thus ensuring the CS-2 a large applications base lacking in most scalable computers.

Meiko is one of oldest parallel computing companies. Founded in 1985 in Bristol, England, Meiko's relatively large base of small installations is a result of nearly a decade of operation. In 1993 Meiko won a contract to supply a large computer to Lawrence Livermore Laboratory, however, it is difficult to see how such a small company can support R&D for its specialized nodes and software for both the technical and commercial markets.

AT&T Global Information Solutions (AGIS). In 1983, Teradata introduced its first multicomputer; nine years later, AGIS (then known as NCR) acquired Teradata. It now has an installed base of more than 200 organizations and 400 systems running commercial database applications, mostly on AT&T DBC (Teradata) computers. Over time, AGIS will transition from a Teradata architecture with a proprietary DBC/1012 database to a more general architecture, the AT&T (NCR) 3600, which supports commercial databases and Unix V.4 applications.

The AT&T 3600, based on AGIS multiprocessors and Teradata's multicomputer architecture, was introduced in May 1991 and shipments began in April 1993. Scalability extends to 1,024 Intel X86 processors.

The AT&T 3600 consists of three types of computers linked together by YNET, Teradata's dual tree structured message passing network. Each dual YNET operates at 6 MB/second. The YNETs operate in tandem at an aggregate bandwidth of 10MB/second. The three computer types are the following:

- Up to 32 dyadics (i.e., pairs) of 1-8 Pentium processors (called applications processors [APs]) that have a disk system for traditional applications.
- Up to 1,024 uniprocessor access module processors (AMPs) that control and access database disks.
- Parsing engines that allocate database requests to AMPs.

Both the AMPs and APs have disks that are accessed via redundant paths. User applications are run in the APs that are controlled by Unix. AGIS has announced that Oracle Parallel Server and Sybase Navigation Server will operate in the AP. By 1995, AGIS intends to have only a single multiprocessor node type that will be used for both the AP and AMP, as well as a faster YNET switch.

While Teradata was first to use a large number of processors to access databases in parallel, nearly all scalable parallel computers described in this report that run a traditional database will provide significant competition and will supply a significant amount of commercial computing aimed at reducing AGIS's market share.

nCube. Founded in 1983, nCube was an early pioneer multicomputer vendor, and now has an installed base of approximately 400 systems. Until Larry Ellison, CEO of Oracle, purchased a controlling interest, nCube concentrated mainly on the technical marketplace using its proprietary node and switch architecture. Given its demonstrated I/O bandwidth and high reliability, the nCube system is particularly suited to two major applications: a parallel database server for the Oracle Version 7.0 environment and a videoserver. Both of these applications are being driven by Oracle. Given nCube's negligible floating-point performance per node, it is no longer targeting the technical market. Ellison has announced his intention to use nCube computers for video-on-demand applications.

The nCube nodes have memories of 4-64 MB and operate at 15 MIPS with a PAP of 4.1 MFLOPS. The nodes are interconnected to one another using a hypercube network (i.e., each node has "n" links to other nodes in a computer with 2^n nodes). The two basic models in the 2S series scale over the following ranges: Model M 5, 8-128 nodes, and Model M 10, 128-1,024 nodes. Three larger 2S models extend the range to 8,192 nodes for a maximum of 123,000 MIPS or 34 GFLOPS.

NEC Cenju-3. The Cenju-3 is a multicomputer with up to 256 50 MFLOPS processing elements (PEs) equipped with a VR4400SC RISC processor (based on Mips R4400 chip). Each PE can accommodate 64 MB of local memory with a maximum total capacity of 16 GB. A 256-PE system provides a PAP of 12.8 GFLOPS. Each PE is connected through a multistaged interconnection network, similar to that of IBM SP1, Meiko CS-2, and ATM switches. A PE can load/store data with other PEs on a word-at-a-time or message-block basis. In addition, barrier synchronization and remote procedure call functions support parallel processing.

Thinking Machines CM5. The CM5 is an asymmetrical multicomputer with 1-32 Sun Microsystem server control computers that "host" user programs and control an array of 32-1,024 computational computers, each of which has four 40 MFLOPS floating-point arithmetic units and 32 or 128 MB of memory. The system has SPARC-based I/O server nodes and a tree-structured switch to interconnect nodes. The system is divided into independent partitions with at least 32 computational nodes managed by each control computer. The CM5 is an evolution of a SIMD architecture with a single instruction multiple data program residing in each computation node and a main control program in the control computer. It can now operate in SIMD or MIMD mode. Because the CM5 is asymmetrical, independent jobs cannot run in the computational computers; thus, a CM5 perpetuates the limitations of SIMD by being unable to process scalar, moderately parallel workloads effectively.

The CM5 consists of three separate networks: control, data message passing, and diagnosis and reconfiguration. Control network messages include broadcasting (e.g., sending a scalar or vector) to all selected nodes, recombining results (carrying out arithmetic and logical operations on data from each node), and global signaling and synchronization for controlling parallel programs. The data network operates at 5-10 MB/second with latencies at the applications level of 7-150 microseconds, depending on the library and O/S. While subsequent computational nodes can evolve to higher performance with greater memory size, a next-generation CM5 requires a proportional increase in the communication network. It is unclear whether CM5's networks can evolve as rapidly as its microprocessor-based nodes to provide generation scalability.

We expect IBM to become the leading supplier of scalable computers.

In March 1994, Thinking Machines announced the availability of Oracle 7, which has demonstrated linear speed-ups. Users have observed a performance that is 50 times better than a comparably priced mainframe. Thinking Machines has described its 1996 architecture as being able to be used in a massively parallel fashion or as independent workstations that are fully ABI compatible with Sun's Solaris operating system.

IBM Scalable POWERparallel (SP) Systems. In 1993, IBM introduced the SP1, which supports 8-64 125 MFLOPS (70 SPECint92 and 121 SPECfp92) processor nodes (headless IBM RS/6000 workstations) with 64-256 MB of memory per node. In April 1994, it introduced the SP2, which supports 4-128 266 MFLOPS processor nodes. These nodes are interconnected via a high-performance switch (HPS) and HPS adapters that have demonstrated 40 MB/second point-to-point data transfer rate and 0.5 microsecond hardware latency. Demonstrated application-to-application latency is less than 40 microseconds. Various nodes can be assigned as compute servers, file servers, mass storage servers, and interfaces to an S/390. The SP2 supports up to 256 GB of internal memory and 1,024 GB of internal disk storage.

The SP2 is controlled by various parallel application interfaces including the IBM AIX Parallel Environment, Express, Forge 90, Linda, and PVM. The cluster is also managed by the IBM LoadLeveler that balances node use, including managing batch operation. Early benchmark performance is impressive; for example, the floating point SPECrate92 efficiency for 16 nodes is 95%.

IBM began delivering the SP1 in February 1993. By the end of 1993, approximately 70 were installed, giving IBM a large installed base and strong customer/ market position. The SP2 is scheduled for general release in July 1994. Within the next year, we expect IBM to become the leading supplier of scalable computers[4] when measured in terms of units, installations, and revenue.

Silicon Graphics Challenge Array. While Silicon Graphics is omitted from most reports on scalable computing, it has demonstrated a 16-node array (20 processors each) of its Challenge server in a 3-D torus similar to the Cray T3D. This array interconnected 320 processors (using 100Mb/second FDDI rings), 28 GB of memory, and 192 GB of disk storage to achieve a peak performance of 16 GFLOPS and a sustained performance of 4.9 GFLOPS.

Silicon Graphics is the principal supplier of workstations for both visualization and computation because virtually every significant technical application runs on its platforms. It has been delivering both multiprocessors that operate at 75 MFLOPS PAP per processor and parallelizing compilers for 5 years, with over 1,000 installed. The company fundamentally understands and has expertise[5] in building both scalable multicomputers (i.e., workstations) and multiprocessors. We estimate that there are currently over 700 installed Challenge multiprocessors with an average performance of 0.6 GFLOPS. Combined, they provide a PAP of 420 GFLOPS—roughly equivalent to the installed base of the largest supercomputer manufacturer.

Silicon Graphics recently announced its Power Challenge multiprocessor for the commercial market; it set a record of 1,700 transactions per second, a rate that is 1.5 times that of large mainframes. In mid 1994, a Power Challenge multiprocessor is slated to be delivered with 300 MFLOPS processors providing a PAP of 5.4 GFLOPS (18 x 300 MFLOPS)—roughly the same PAP and incremental price per FLOPS as a 32-node CM5. Given the multiprocessor structure, fine-grain applications (including traditional supercomputer codes) will run efficiently through both vectorization and parallelization.

Networked Workstations

DEC Alpha AXP Farm. A Digital Equipment Corporation (DEC) workstation farm (a collection of workstation and/or server nodes) is composed of up to 120 nodes connected via FDDI, Ethernet, or ATM, and is controlled by LSF (load sharing facility) cluster compute and PVM software. LSF provides the ability to move work to the appropriate node with monitoring, Unix's Make command done in parallel, load sharing, and batch operation. LSF also supports heterogeneous farms consisting of workstation nodes from DEC, Sun, IBM, Silicon Graphics, and HP.

Ease of programming will define how fast the scalable parallel computer market grows.

DEC recently introduced packaged, pre-configured workstation farms (AdvantageClusters) based on its GIGAswitch, which connects up to 22 FDDI ports to a cross-bar switch. The GIGAswitch enables 6.25 million connections per second at an aggregate data rate of 3.6 gigabits per second. AdvantageCluster compute and file servers support up to 32 processor nodes. A high availability AdvantageCluster supports 2-3 processor nodes and offers redundancy, volume sharing, automatic recovery and failover, and high availaibility NFS.

SIMD Computers

MasPar MP-2. The MasPar MP-2 is a cost-effective computer that uses the massive SIMD paradigm and 1K-16K processing elements. In order to achieve parallelism, processing elements controlled by a single instruction are placed with distributed memory. The MP-2 is hosted by a VAX computer. (Digital Equipment is a distributor of MP-2 systems.) Data are moved among the processing elements through a nearest neighbor communication or a high-speed switching network. MP-2 has a high-bandwidth memory that can access 2.5 words of memory for each floating-point operation.

[4] This projection does not include Silicon Graphics' multiprocessor servers sold for technical computation.

[5] Professor John Hennessy of Stanford University, a Mips Computer founder, is researching scalable multiprocessors using Silicon Graphics platforms and is a consultant to the company.

The MP-2 has several advantages over its multiple instruction multiple data (MIMD) counterparts: (1) because only one instruction is executed at a time, it is inherently fine grain and synchronized, permitting vector processing style programming; (2) the fast, low-latency network interconnecting the processing nodes means that internode communication delays are small, so that memory can almost be treated as centralized; and (3) it has a fast I/O system for disks and real-time data, such as video or radar data.

Achieving Viability

The speed at which the scalable parallel computer market will grow will be defined and limited by one factor: ease of programming. Because of the difficulty in developing new algorithms and new code to run effectively on scalables, scalable parallel computers must run existing supercomputer applications competitively to achieve at least minimal viability.[6] We believe that by 2000, virtually all computers will be scalable. But the exact way in which they are scalable will depend upon a number of variables, including development of processors, memory, mass storage, switches/networks, operating systems, and applications. The most likely form will be simple computers connected to a high-speed, low-latency, ubiquitous network (e.g., ATM).

Processor performance and memory size are key determinants of speed and both have proven to be generation scalable. Mass storage is also generation scalable—disk capacity has doubled every 18 months at a constant price. Switches and networks are less likely to scale as easily as other components because their bandwidth and latency are not as easily generation scalable. However, switches may be irrelevant, provided that they are fast enough[7] and can scale adequately to support the 100-fold parallelism that most commercial and technical applications can use. Decoupling switch and node designs will allow each to evolve more rapidly, interoperate, and provide intergeneration evolution.

It is time for vendors of scalable parallel computers that utilize unique nodes and networks to reexamine their product strategy. An ideal scalable must not only be size, spatial, and generation scalable, it must also, for survival, be viable. This viability can be accomplished by being compatible with, complementary to, and competitive with other computer structures. All scalable structures have inherent overhead including packaging, power, a switch (either processor to memory or processor to processor), and operating system copies. Thus, today's scalables are not price/performance competitive with multiprocessors on the low end of the market, nor are they competitive with networked workstations that scale at essentially no extra cost.

To succeed in a niche, a scalable computer must be fully compatible with other computer structures by building on their components.

Successful designs reduce the burden of overhead through *elegance,* whereby one component carries out multiple functions. For example, multiprocessors are elegant because the bus/backplane carries processor-memory-I/O communication, packaging, cooling, and power. The shared memory provides memory and infinite communication among processes. Networked workstations are also elegant because the network carries out many communication functions, including support for parallel processing. Scalables that utilize unique nodes and networks have little elegance and must bear the full burden of the inherent overhead.

To succeed in a niche, a scalable computer must be fully compatible with other computer structures by building on their components. In terms of hardware, compatibility means utilizing "main line" microprocessors that are adopted by multiprocessors and LAN-based workstations, not special-purpose computers. In terms of software, compatibility means adopting operating systems, tools, libraries, and applications compatible with other computer structures. Furthermore, with high-speed ubiquitous networking, a scalable must build on standard hardware and software network structures that enable spatial scalability. With spatial scalability, massively parallel computers can exist across any environment at "zero" cost by utilizing existing workstations, servers, and standard networking; hence, spatial scalability is a requirement for viability because it is the key to attracting applications.

We believe that the winning approach to scalability is complete compatibility with workstations or PCs. With

[6] Viability is a computer's ability to develop software compatibility among a variety of platforms over a long period of time and to handle a variety of job sizes, application types, and mix of computational resources.

[7] We anticipate line speed for ATM switches to increase from 655 Mbits to several Gbits by 2000.

compatibility, a user will not see a difference in simple applications whether run on the desktop or on a multiprocessor server, such as Silicon Graphics' Challenge XL, or as a collection of headless workstations operating together as a scalable server. The principal difference among the three alternatives is the degree of parallelism that can be achieved based on the interprocessor communication characteristics.

Users should not view massive parallelism as a panacea, providing untold returns using a particular application that no other organization has. Rather, it should be viewed as a technique that can provide both more cost-effective computing in the long term and, in a few cases, solve particular, elusive large-scale commercial and technical problems. Applications fitting this profile include scientific and engineering simulation and analysis and very large-scale commercial systems for database, transaction processing, and data analysis that cannot be solved by other means. Most forecasters predict the commercial market will grow rapidly, eclipsing the technical market. This scenario is feasible provided that running in parallel is transparent to users.

The main barrier to using computers in parallel was, is, and will continue to be developing the right programming languages and environments that will enable training, development of programming tools, and support of standard, third-party applications. The best scenario is that users will not see any differences in computers from various vendors (in terms of the availability of and user environment for applications) other than performance and price/performance differences. The greatest inhibitor of (or competitor to) parallelism is faster sequential processing. The evolution of limited-scalability multiprocessors takes a substantial part of the market that specialized scalable computers might otherwise address. All of these factors suggest that the path to massive, parallel processing will be via standard, mostly uniprocessor computers such as workstations and PCs that are interconnected via emerging high-speed networks—not specialized scalable computers.

We recommend that all vendors consider using standard nodes, networks, and programming environments to reduce development and product costs (building from a single learning curve) and improve time to market, thus allowing them to concentrate their considerable skills on value-added components of parallel processing. Also, all companies that build traditional workstations with compatible multiprocessor servers (including Apple, AGIS, Compaq, DEC, HP, IBM, Intel X86-based system companies, Silicon Graphics, and Sun, as well as all members of their microprocessor *keiretsus*)[8] should offer high-speed, standard networked environments at zero (or minimal) incremental cost. Only then will standardization finally stimulate parallelism.

About the Author

__Gordon Bell__ is a computer industry consultant at large. He spent 23 years at Digital Equipment Corporation as vice president of research and development, where he was the architect of various minicomputers and time-sharing computers and led the development of Digital's VAX and the VAX environment. Mr. Bell has been involved in, or responsible for, the design of many products at Digital, Encore, Ardent, and a score of other companies. He is on boards at Adaptive Solutions, Chronologic Simulation, Cirrus Logic, Kendall Square Research, Microsoft, Visix Software, University Video Communications, Sun Microsystems, and other firms.

Mr. Bell is a former professor of computer science and electrical engineering at Carnegie-Mellon University. His awards include the IEEE Von Neumann Medal, the AEA Inventor Award, and the 1991 National Medal of Technology for his "continuing intellectual and industrial achievements in the field of computer design." He has authored numerous books and papers, including High Tech Ventures: The Guide to Entrepreneurial Success, *published in 1991 by Addison-Wesley. Mr. Bell is a founder and director of The Computer Museum in Boston, Massachusetts, and a member of many professional organizations, including AAAS (Fellow), ACM, IEEE (Fellow), and the National Academy of Engineering.*

Eric P. Blum, Research Program Manager

94-11-59

[8] *Keiretsu* is a Japanese word that describes a group of affiliated companies. For more details on microprocessor *keiretsus*, see "Microprocessor Standards and Markets, Part II: Six Architectural Affiliations," *Spectrum, Information Systems Industry*, Issue 53, 1993.

About Decision Resources, Inc.

Decision Resources is an international publishing and consulting firm that evaluates worldwide markets, emerging technologies, and competitive forces in the information technology, life sciences, and process industries. Decision Resources links client companies with an extensive network of technology and business experts through consulting, subscription services, and reports. For additional information, please contact Marcia Falzone by phone at (617) 487-3749 or by fax at (617) 487-5750.

This material, prepared specifically for clients of Decision Resources, Inc., is furnished in confidence and is not to be duplicated outside of subscriber organizations in any form without our prior permission in writing. The opinions stated represent our interpretation and analysis of information generally available to the public or released by responsible individuals in the subject companies. We believe that the sources of information on which our material is based are reliable and we have applied our best professional judgment to the data obtained. We do not assume any liability for the accuracy, comprehensiveness, or use of the information presented.