Cyber-bricks and paradigms

David Clark
dwclark@earthlink.net

> It's unacceptable in the day and age of the Internet to have any finite upper limit on how much scalability you can deliver.—*Mitch Shults, Director of Intel Server Platform Marketing, Enterprise Server Division*

After more than 15 years since it was commercially introduced by Digital Equipment Corp., *cluster* is a buzzword again—and *paradigm* has become a fuzzword. But whatever words you choose to use, clusters of NT computers and mass storage devices—connected locally and remotely into a single, highly scalable system—will quickly become the dominant, well, *model* of enterprise computing.

Over 80% of the multiserver systems today—linked as clusters or not—are connected for high availability. The demands of accessibility and reliability in today's Internet, intranet, and enterprise server environments increasingly require *failover* capability—if one server goes down, the other takes over automatically. But the ultimate goal of clustering on the NT platform is to achieve what Microsoft's Jim Gray calls *cyber-brick* scalability—where each cyber-brick you connect to your system of NTs adds most of that brick's processing power to the cluster "wall."

Gray worked on clustering and distributed-database problems at Digital with VMS architect Gordon Bell; they both now work at Microsoft. Their expertise, and technology-sharing partnerships with both Compaq-owned Tandem Computer and Digital, makes analysts such as Dataquest's Jerry Sheridan believe that NT can soon deliver Unix-level cluster capability.

Clustering again

Why has the idea of clustering become new again? Because technology and markets have changed since VMS introduced the ability to cluster workstations in the 1980s. The growth of the Internet, for example, has been dramatic. The Internet is a peer-to-peer architecture, and so is the idea of clustering, says Tandem's Jim Henry. "It's not just that Microsoft announced clustering as part of NT, but it's also because the Internet is driving us toward the same kind of architecture—which is a clustered architecture—even though the clustered units may be many miles apart." Mitch Shults of Intel agrees: "The story isn't really about clustering," he says. "We're moving to a distributed paradigm, and clustering is a natural complement to that."

Internet growth has also created a huge demand for servers—especially NT servers. The relatively low-cost, commodity nature of NT servers has projected them into spaces where only big iron used to fit; the NT server market has become huge. As the price/performance picture gets better, the cycle continues. Internet growth has "built so much value around the server paradigm that the payback for making the [server] investment is obvious," says Microsoft's Mark Wood. Naturally, large markets drive development. "It's a synergy between the massive expansion of the Internet, as well as a massive expenditure to create things for the Internet," according to Shults.

Another enabling difference between clusters of today's proprietary Unix systems and the newer

NT approach has to do with standards. The Internet, by its open nature, has set some standards—in protocols, and in its way of distributing intelligence throughout the network. In addition, the relatively new idea of middleware in software development implies standardization. Developers will use a single (Microsoft) application programming interface to deliver clustering on the NT platform—a standard that Microsoft's Wood points out will nonetheless provide "a far broader choice in vendors than other solutions." Since VMS was introduced, software and hardware vendors have become accustomed to developing products based on industry standards. According to Intel's Shults, a new open-cluster architecture will provide, among other things, standardization in a "multivendor, binary-compatible competition."

Called the Virtual Interface architecture specification, the standards initiative was announced in April by Microsoft, Intel, and Compaq. More than 40 companies are involved in drafting standards for hardware and software interfaces used in cluster communications. The standard will provide "multiple choices of interconnect, multiple choices of server platforms, multiple choices of operating systems on top of that server platform, and [distributed] applications on top of the OS—all of which are able to take advantage of a very high-speed, low-latency, standards-based interconnect," explains Shults. Approval of the VI technology standard is expected in December, and companies including Tandem, Giganet, and Mirrorcom are standing by to "push the button on their silicon implementations to create VI optimized network interface cards" when the specification is ready.

To achieve the cluster holy grail—the ability to grow a cluster organically by adding a cyber-brick at a time as you need more horsepower—the VI architecture has to break the bottleneck of latency. The recent trend in servers has been to add processors to gain power. Four-way serial message-passing NT servers have become common. However, in SMP environments, finite levels of resources are available. Each processor and process competes for available memory and processor bandwidth. The same holds true when you cluster SMP systems. Each added node increases contention geometrically. Because most cluster populations today have fewer than four nodes, contention levels are considered acceptable. However, that won't be the case as the demand for greater scalability increases.

Moving packets quicker

"You can't reduce latency by building faster wires," says Intel's Shults. "The key is to get the software—the operating system and the protocol that drives network connections—out of the way so that packets can move from one server to another." He describes the distributed message-passing (DMP) nature of the VI architecture like this:

> In a standard binary-compatible way, the architecture specifies a register-level interface between applications and the network interface card, which is built to the VI standard. You go directly out of application space, from a data buffer that's controlled by a database or some other communicating application, through the interface card, right onto the wire. Then across the wire through some sort of system-area network (SAN)—there are many choices—up through the other network interface card, and directly into application memory space in the receiving application. Acknowledgment is sent through the same mechanism. The operating system doesn't intervene at all. There's no protocol stack involved, other than anything the application might be inserting for reliability purposes. By doing that we can reduce the amount of software processing overhead and latency by an order of magnitude.

To achieve the cluster holy grail—the ability to grow a cluster organically by adding a cyber-brick at a time as you need more horsepower—the VI architecture has to break the bottleneck of latency.

Intel brings considerable experience with multiprocessor DMP technology to the VI alliance. For example, earlier this year, it delivered a teraflops system to Sandia National Laboratory. That system linked over 9,000 Pentium Pro processors and had a 400 Mbit/s mesh interconnect with an end-to-end switching latency of under 25 microseconds between the system's motherboards. That technology is proprietary and not suitable to the commercial marketplace; it was developed under a contract from the US Department of Energy. However, Intel's experience proves the technology is sound and that they know how to make that kind of switching technology work.

The software community has recognized that DMP is the best solution to the SMP latency problem as well. All the major database vendors have already delivered, or will be delivering, DMP-based distributed database applications. Oracle and Informix have announced commitment to VI; others, including IBM, are expected to announce support shortly.

Distributed message passing will help break another, largely hidden, bottleneck: the enormous number of system messages generated by object orientation in application software and in distributed object models. As systems become more complex and more objects are added, the number of messages grows exponentially. VI's low-latency DMP architecture should provide good performance and scalability in increasingly common distributed-object environments.

Another significant aspect of the VI specification is standardization of a middleware interface, which will lead to development of distributed applications and which will allow existing software to take better advantage of clusters. Middleware programmers will be able to write to a single interface, no matter what OS or network is being used. "The key to distributed applications is middleware," says Tandem's Henry. "If you have cluster-aware middleware, application programs don't change at all. They can run on a hundred nodes and they won't know the difference because the middleware—database, transaction monitor, or messaging system—is aware of the cluster and hides it from the programmer."

Around the corner

What will the future bring? Greater reliability—on the order of 10^{-18} bit errors,

Continued on p. 17

Focus, continued from p. 12

according to Shults—and definitely greater, more efficient scalability. Currently, the cluster server of Microsoft's NT Server Enterprise Edition can link only two nodes, but that will change with Phase II, expected mid-1998. With that will come more and better distributed applications—improved databases and other software that takes advantage of highly scalable clusters. We'll also see new and different uses for NT clusters, including departmental *data-marts*, where companies with huge mainframe data warehouses will extract and download department-specific database tables to each department's server—providing the department's managers with better information and the ability to data mine.

Disk-connection technology will change; today's SCSI (Small Computer Systems Interface) standard isn't the ultimate solution. We'll see an evolution of standards, including fiber connections with SCSI protocols over fiber to get around some of the electrical and distance issues with SCSI. Compaq and others have talked about fiber being a strategic direction. The SCSI protocols themselves will have to evolve. A likely alternative to SCSI will be fiber channel-arbitrated loop. We'll probably see switched fabrics, such as Tandem's ServerNet, or other equivalent SANs being used to hook up servers to disks.

Platforms will also evolve. We'll eventually see a move away from the PCI-based platform, and new microprocessors that take full advantage of some of the VI architecture. Intel's Shults says, "We're not making any announcements right now, and today our direction is PCI-based, but ultimately you can predict that the SAN and the high-speed, low-latency interconnect will become a base capability of the Intel architecture for a standard high-volume server platform."

A future of never-fail, infinitely scalable walls of clusters, built with low-cost cyber-bricks, might not happen tomorrow. However, the promise of the VI alliance, along with independent development by many NT vendors, suggests we might get closer to that ideal—sooner than we thought. If you suffer from slow Web service, you might even consider buying your Internet service provider an extra cyber-brick for Christmas.