SYBASE for HACMP/6000: An Architected Approach to Clustered Systems

By Josh Bersin

This article discusses Sybase's product and plans for clustered and parallel systems, including Sybase's current and future support for IBM's HACMP/6000. It also compares using SYBASE for shared-disk clusters like HACMP/6000 to the shared-nothing environment such as the IBM SP2.

As client/server computing becomes a key architecture for enterprise applications, we see the following key trends in server systems:

- **Very large databases**: It is now typical for customers to have databases in the tens of gigabytes, growing to hundreds of gigabytes and beyond.
- **Hundreds to thousands of online users**: As the amount of data grows, so does the number of users who need to access that data. The SYBASE architecture design assumes that these client users are accessing the system through a network.
- **Demand for Online Transaction Processing (OLTP) performance**: Sybase has targeted the OLTP environment from the beginning. These applications demand subsecond response time to the end-user client workstation. It is not enough to allow users to connect—the system must scale up as OLTP workloads increase.
- **Continuous availability of the application**: With hundreds of client users, applications must be continuously available. At the application level, the SYBASE Replication Server addresses this need by allowing a replicated server to run even if the primary server or network fails. At a single-server level, this can mean online backup without major impact on performance, automatic and rapid fallover, workload balancing, and full transaction recovery.
- **Large-scale query and decision support**: As customers put their users and customers online, the demand for query and reporting on the large client/server databases grows as well, often requiring the server to scan databases many gigabytes in size.
- **Mixed workloads on a single server**: In many cases, customers are asking single-server systems to handle both the OLTP and the decision-support workload in a single location.

In short, these requirements resemble the IBM mainframe environment. This article discusses how Sybase is designed to address these six requirements, and in particular, how HACMP/6000 plays an important role in addressing these needs.

SYBASE System 10: Architected for Enterprise Client/Server

When Sybase was founded in 1984, the company brought two key technologies to the Relational Database Management System (RDBMS) market: client/server architecture and OLTP performance. This meant developing client and server Application Programming Interfaces (APIs), as well as developing key new concepts such as a multithreaded server, compiled stored procedures, triggers, and client/server and server/server Remote Procedure Calls (RPCs).

May '93
p.18

Josh Bersin

SYBASE Open Client and Open Server

SYBASE SQL Server 10 is still the only RDBMS product that has integrated networking and both client and server APIs. All Sybase products are built on SYBASE Open Server and Open Client, a network-based, multithreaded interface designed for high-performance networked computing.

Open Server and Open Client, widely used interfaces in the industry, support both SQL and RPC communications between clients and servers and between servers and servers. They run across a large variety of network protocols and vendor network libraries, and on nearly every platform in the industry including AIX and MVS/ESA™.

The Sybase client/server architecture is based on multithreaded connections that allow both synchronous and asynchronous communications. They allow event notification and polling as well as store-and-forward message queuing through the SYBASE Replication Server.

SYBASE SQL Server for Uniprocessors and SMP Systems

The SYBASE SQL Server, which forms the base of the family, is a multithreaded RDBMS engine designed from the ground up for network-based access. By using an optimized threading model, SQL Server can handle thousands of simultaneous users in a single process. A client user accessing SQL Server from Open Client requires less than 50 KB of server memory.

SQL Server's design point has always been optimized for OLTP. For example, in September 1993, Sybase and IBM completed a TPC-A benchmark that attached 2,760 client users to a single SQL Server on a RISC System/6000 Model 990 and achieved average response time of .79 second.

SQL Server pioneered the use of compiled, shared, stored procedures and triggers, which can not only control the local database, but can also send RPCs to other remote servers. This architecture (shown in Figure 1) makes possible a high-performance, guaranteed integrity, client/server database.

SQL Server has many features designed for continuous availability. Online backup has always been available. In System 10, SQL Server includes a separate backup server process that provides online backup with little or no impact on the performance of the SQL Server. Backup server can back up tens of gigabytes per hour to multiple devices and can operate on a separate machine to allow the OLTP server to run unaffected during backup.

SYBASE Client/Server

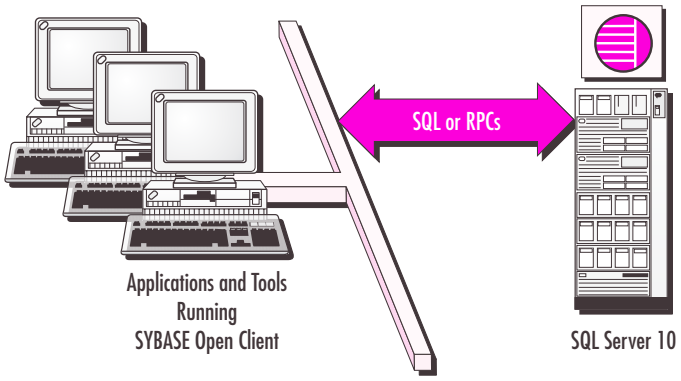
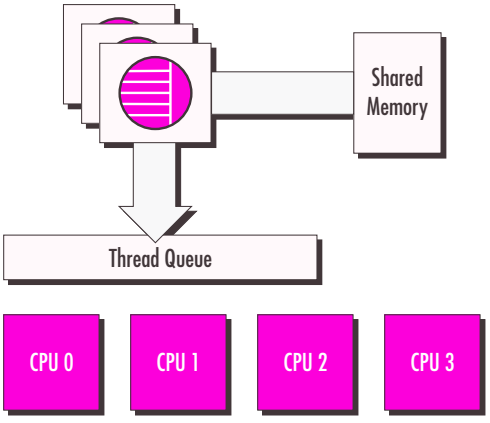


Figure 1. SYBASE client/server architecture

SYBASE SQL Server for SMP Systems

For Symmetric Multiprocessing (SMP) systems, Sybase sells a multiprocess implementation in which multiple SQL Servers are loaded in memory and CPUs are dedicated to the server complex. The SQL Servers communicate through shared memory and can dispatch threads to a run queue that looks for the next available processor. The result is a highly scalable design, leveraging

SQL Server VSA for SMP Systems



- n-1 Servers Loaded
- Efficient Thread Dispatching
- Single Virtual Server
- Highly Scalable to 6–8 CPUs

Figure 2. SQL Server VSA for SMP systems

the strength of the SQL Server uniprocessor architecture along with a single-server image to applications, shown in Figure 2.

Distributed and Parallel Servers from Sybase

Other products that encompass the SYBASE architecture include the following:

Replication Server: Allows customers to create primary and replicate copies of data with automatic, fault-tolerant transaction distribution from primaries to replicates. Replication Server uses a store-and-forward protocol to provide continuous or timed transaction distribution to replicate sites. Replication Server uses a subscription process to allow replicate sites to receive continuous or timed updates from a primary site at a table, row, column, or user-defined level.

A network-based software architecture is fundamental to leveraging future hardware architectures.

OmniSQL Gateway™: Allows any client application to access multiple heterogeneous data sources transparently as if they were located in a single database. OmniSQL Gateway gives an application turnkey access to data in SYBASE SQL Server, DB2, Oracle, Ingres, Informix, and a variety of other data formats.

Control Servers: Consists of a family of client/server performance monitoring and system management tools.

Navigation Server™: Couples together SQL Server engines in a message-based parallel architecture to provide parallelized OLTP, queries, backup, and utilities for applications that demand large databases, large queries, and large numbers of users.

Navigation Server, jointly developed between Sybase and NCR, uses multiple SQL Servers, one per node, in a shared-nothing architecture with fully partitioned data. To a client, Navigation Server looks like a single, large SQL Server, allowing applications to exploit the power of a large parallel system as if it were a single-node machine. Unlike other parallel server offerings, it is fully message-based and optimized for partitioned, shared-nothing machines. Navigation Server is scheduled to be available on IBM's SP family of systems in early 1995.

Technology Trends

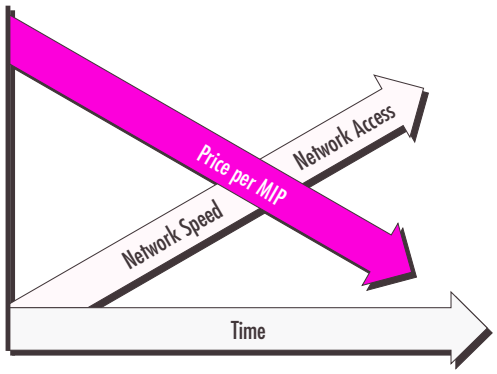


Figure 3. Trends in computing technology

Hardware Technology Evolution: Clustered Architectures versus Others

Hardware technology advances at a dramatic rate. Sybase sees two fundamental shifts occurring in hardware technology (shown graphically in Figure 3). First, CPU performance is increasing at a dramatic rate—doubling price/performance per year. Second, high-speed, low-latency networks are becoming inexpensive and ubiquitous. For the future, we believe that a network-based software architecture is fundamental to leveraging future hardware architectures.

Today's Range of Computer Architectures

To exploit these technology trends, IBM and other vendors are developing new computing architectures that are critical to the design of our software architecture, as shown in Figure 4.

Clusters versus Parallel versus SMP Systems

For high-performance OLTP today, the most scalable systems are uniprocessors and SMP. OLTP workloads place heavy demands on updating the database—therefore when multiple CPUs are involved, they must be able to communicate with each other rapidly to serialize access to data. SMP

System	Uniprocessor	SMP	Cluster	Parallel
Architecture	Shared Everything	Shared Memory	Shared Disk	Shared Nothing
# of nodes	1 CPU	2–16 CPUs	4–32 CPUs	8–64+ CPUs
IBM model	Model 990	PowerPC™ SMP	HACMP/6000	SPx family

Figure 4. Range of hardware architectures for client/server computing

achieves this through shared memory, and can often scale in OLTP workloads at over 90% efficiency for four to six nodes. Design implementations, however, often limit the number of nodes to six or eight.

Parallel systems, on the other hand, or shared-nothing systems such as IBM's SP2 are excellent for large-scale decision support and reporting. By using partitioned parallel technology such as Navigation Server, they can access a large database in parallel, providing a near linear speedup in query performance as additional nodes are added. Navigation Server is also designed for parallel OLTP, allowing the system to increase the number of users as additional nodes are added.

Clusters like HACMP/6000 offer different advantages. The first key advantage is high availability. If one node in a cluster fails, another node takes over. The second key advantage is that each node in a cluster can be a large SMP itself, allowing a cluster to theoretically scale far beyond a large SMP system. This allows thousands of clients to be connected to a single database.

However, there are limitations. Clusters, because they communicate via networks and shared disk, do not scale well in update-intensive workloads. Even light OLTP workloads do not scale well. This is because multiple nodes that want to update a single database record must communicate through a Distributed Lock Manager (DLM) to serialize access to data. The DLM, as implemented today, does not scale well for update-intensive work like OLTP. The best clusters available today get slightly more than two times performance going from one to four nodes.

Clusters are also not optimal for large database queries. Query performance is limited by the time it takes to scan large tables—and this cannot be speeded up significantly without partitioning data. Sybase sees the Navigation Server and SP2 as an excellent solution for this workload.

HACMP/6000 and Sybase Today

For fallover and hot-standby modes, SYBASE SQL Server supports high-speed switchover and a variety of configuration options.

Here are some key features of Sybase for HACMP/6000 today:

- Fast fallover through a tunable transaction log ("recovery interval") to create a cluster with either fast fallover and more frequent checkpointing, or less frequent checkpointing with longer fallover
- Use of Logical Volume Manager (LVM) raw I/O for highest performance fallover and recoverability
- Open Client architecture that allows client applications on the network to transparently switch or retry a failed server during high-availability switchover and takeover of a failed system's IP address
- Replication Server services to allow one node to function as an OLTP server and another as the query and reporting server for maximum performance and high availability

For high-performance OLTP today, the most scalable systems are uniprocessors and SMP.

SYBASE for HACMP/6000 in Mixed OLTP and DSS Workloads

Due to the large and unpredictable amount of CPU and disk I/O required for large queries and reports, mixing OLTP and decision support in a single machine is difficult to manage. One excellent solution to this problem is to partition the workload onto two nodes of a cluster. Many customers are doing this today using SYBASE Replication Server.

With SYBASE Replication Server, customers can implement high-performance OLTP applications on one HACMP/6000 node and replicate transactions to a second or third node for query or reporting applications. This allows client applications to receive predictable, guaranteed response time on the OLTP node without contention by periodic reports or queries. Decision-support and query functions are automatically routed to the second node by the Open Client application and do not impact the mission-critical applications.

Because the nodes are operating independently, there is no overhead from the HACMP/6000 DLM. Both systems run unconstrained at full individual CPU speeds. All systems in the cluster can run efficiently and provide full performance.

During a failure of one of the nodes, HACMP/6000 can automatically bring up the OLTP application on the surviving node and move the decision-support application to the background. This allows mission-critical applications to operate continuously with highly scalable function for the entire suite of applications.

Figure 5 illustrates transaction replication using SYBASE Replication Server.

A Partitioned, Single-Image Cluster

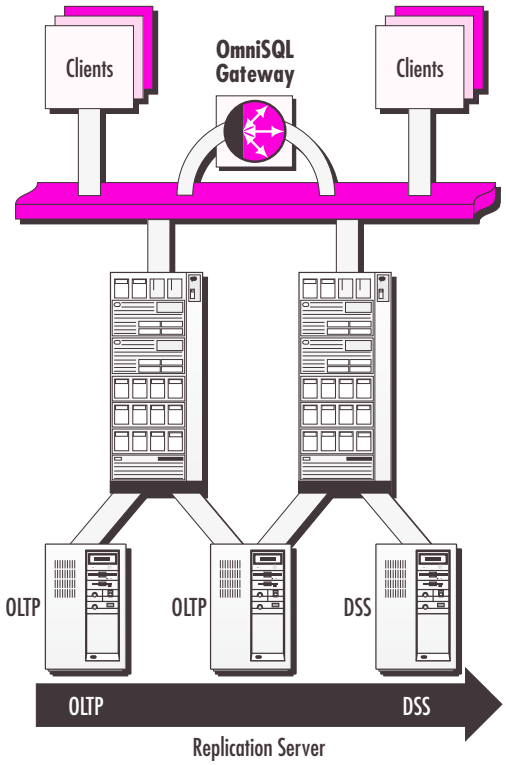


Figure 5. A partitioned, single-image cluster for OLTP and decision support

Message-Based Cluster Versus DLM

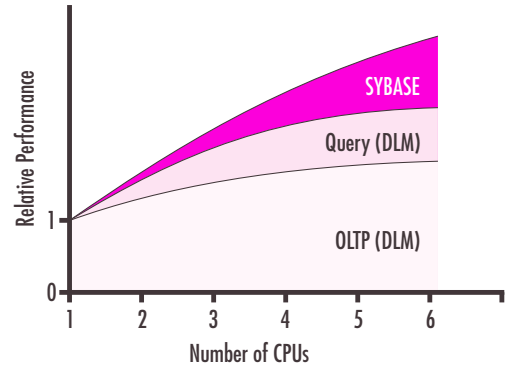


Figure 6. SYBASE message-based cluster performance versus traditional DLM

SYBASE for Clustered Systems: A Message-Based Architecture

The Sybase approach to concurrent disk sharing uses short, low-latency messages between the nodes to serialize access to data. As network speeds improve, the cluster performance improves.

Instead of using the existing HACMP/6000 DLM, Sybase is developing a clustered SQL server for concurrent disk sharing with HACMP/6000. It is designed to scale beyond what is available with the DLM approach today. We expect the product to be available in 1995. Our approach is to allow each node in the cluster to "own" a portion of the database. When one node needs to update data owned by another node, it sends a message to the owner. The owning node either performs the transaction or releases the lock so that the first node can update the record.

Because this approach is built largely around the network, we believe that it will scale beyond the DLM approach available today. As shown in Figure 6, our design goals are to generate a product that provides good OLTP scaling as additional CPUs are added to the cluster.

This approach also allows for continuous availability. If a node fails, another node takes over its ownership, and the workload is transferred to the surviving nodes. Any transactions in process are rolled back or rolled forward. Client connections can be automatically routed to a surviving node. Sybase's design goals for clusters are to deliver breakthrough performance in two areas: continuous availability with workload balancing and OLTP performance.

Hardware Target: Clusters of Uniprocessors or SMPs

As shown in Figure 7, the SYBASE cluster architecture is designed to work across clusters of uniprocessors, SMPs, or combinations. SYBASE will use shared, concurrent access disk ("concurrent access for HACMP/6000") and will not require the use of the HACMP/6000 DLM. The product is designed to scale to four nodes and beyond, with excellent OLTP performance characteristics. In each of the uniprocessor or SMP nodes, the product will use existing SQL Server uniprocessor or SMP technology.

Benefits of the Sybase Cluster Approach

The following are some benefits we see with the cluster approach.

Continuous availability—concurrent resource access: SYBASE will provide continu-

SYBASE Cluster Implementation

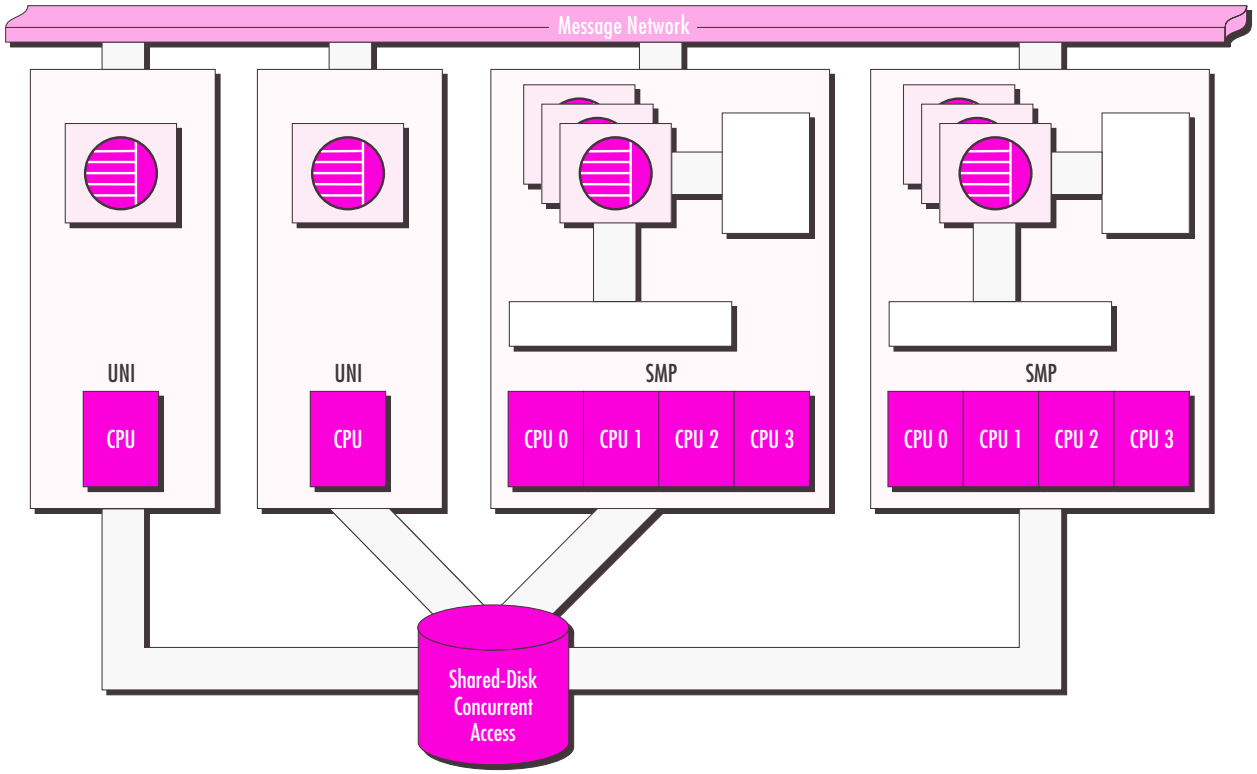


Figure 7. SYBASE cluster implementation (HACMP/6000 Concurrent Access mode)

ous availability at a transaction level, with a client having the ability to start a transaction on any node and having the transaction completed or rolled back if any system fails. If a node fails, the connections are transferred to a surviving node, and the transactions are completed or rolled back.

Workload balancing: If a node fails or is added in this architecture, work is automatically rerouted to the surviving nodes, allowing the cluster to reconfigure itself. This provides dynamic reconfiguration of the workload if hardware changes or unplanned outages occur.

Scalable high performance: Each node in the cluster consists of an SQL Server or SQL Server/VSA system, offering the same level of performance available in a non-clustered environment. When multiple nodes attempt to write to the same record, conflicts are resolved rapidly, providing high scalability. Unlike the current DLM approach, the system is designed for update-intensive applications like OLTP.

Large numbers of clients: The efficiency of the SQL Server and VSA implementations on each node allows the Sybase cluster to support thousands of users accessing a single database. With less than 50 KB of memory required per client, a single-node uniprocessor or SMP will support hundreds of users. The overhead is minimal as additional nodes are added, allowing thousands of users to be connected to the database.

Data integrity: Sybase pioneered using stored procedures and triggers to guarantee data integrity in a networked environment. The Sybase message-based cluster will use this technology to guarantee data and transactional integrity across the network.

Application Environment	Sybase	IBM	Performance
High-performance OLTP and queries	SQL Server/VSA	RS/6000 Uni RS/6000 SMP	300+ TPC-A 400+ TPC-A Hundreds of users
Mixed workload OTLP and DSS	SQL Server/VSA Replication Server	Uni and SMP HACMP/6000	300–400 TPC-A per node Partitioned workload Hundreds of OLTP users Hundreds of DSS users
Scalable OLTP and DSS with high availability	Clustered SQL Server (available in 1995)	HACMP/6000	300–400 TPC-A per node 1000+ TPC-A in cluster Scalable shared disk Workload balancing Thousands of clients
Large-scale queries Very large database (terabytes) Very large OLTP	Navigation Server (available in 1995)	SP2	Thousands of TPC-A Terabytes of data Thousands of clients

Figure 8. Sybase architectures for IBM RISC System/6000-based systems

Conclusions

Figure 8 summarizes Sybase architectures for RISC System/6000-based systems. Sybase believes that with a scalable OLTP solution, a new generation of cluster applications will become possible with HACMP/6000, allowing applications to fully exploit the power of new processing and network technology as it becomes available.

***Josh Bersin**, Sybase, Inc., 6475 Christie Avenue, Emeryville, CA 94608. Internet: joshua.bersin@sybase.com. Mr. Bersin manages Sybase's product and marketing strategy for the IBM product line. He has spent over 10 years in the industry in a variety of technical, marketing, and management positions at IBM and Sybase. He holds a BS from Cornell University, an MS from Stanford University, and an MBA from the University of California at Berkeley.*