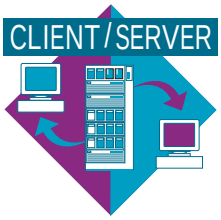
Selecting Nodes for the IBM RS/6000 SP



By Clive Harris

This article describes how to select nodes for the IBM RS/6000 SP. The node types will be described in detail followed by some guidance in selecting SP nodes for particular applications.

Since the initial introduction of the IBM RS/6000 SP™, IBM has continued to announce new, faster node types with better price/performance. The SMP-based High Nodes for the RS/6000 SP are designed for commercial applications and, typically, are not used for scientific/technical applications when high floating-point performance is required. However, the new SMP nodes, based on the PCI bus, are suitable for both commercial and scientific/technical workloads.

The 332 MHz SMP Nodes, based on PCI architecture, are the most recent addition (April 1998) of node types supported by the RS/6000 SP. All previous nodes are based on Micro Channel® Architecture (MCA).

IBM now offers a variety of both uniprocessor nodes and symmetric multiprocessor (SMP) nodes for the SP system.

How to Choose an SP Node

Several node types are available for selection: Thin Node (uniprocessor or SMP), Wide Node (uniprocessor or SMP), and High Node (PowerPC™ 604e SMP). Within these types, there are various configurations, as shown in Figure 1.

With so many node choices, how can you select the most appropriate node in any particular case? There are several factors to consider.

The first step is to review the required capacity for adapters, internal disk, and memory. This can help determine whether a Thin Node (which often provides better performance) has sufficient capacity for the needed requirements. Once that is determined, consider the performance requirements and how the application can exploit the different types of nodes. Even though a Thin Node may have sufficient capacity in terms of adapters, disk, and memory, other requirements may necessitate selecting another node (for example, a need for using Micro Channel adapters). Be sure to include all Micro Channel adapters that will be required, both now and in the future. For high availability solutions, it often is necessary to include additional adapters for redundancy.

Applications have different uses for the internal disk. Commercial applications use the internal disk on each node only for the AIX® operating system, paging space, application executables, and any temporary data. Typically, any critical customer data is stored on external disks so the disks can be “twin-tailed” to provide high availability if the primary node fails. The internal disk cannot be accessed from another node if the node itself has failed. Some scientific/technical applications have a lower volume of data, and high availability is not so important. For these applications, the internal disk is the appropriate solution.

Clive Harris

Node Type	RS/6000 Equivalent	Processor	Current / Non-current[1]
Thin 1	Model 390	POWER2 (66 MHz)	Non-Current
Thin 2	Model 39H	POWER2 (66 MHz)	Non-Current
Thin P2SC	N/A	POWER2SC (120 MHz)	Non-Current
Thin P2SC	Model 397	POWER2SC (160 MHz)	Current
Thin PCI	Model H50	PowerPC 604e (332 MHz)	Current
Wide 1	Model 590	POWER2 (66 MHz)	Non-Current
Wide RPQ	Model 59H	POWER2 (66 MHz)	Non-Current
Wide 2	Model 591	POWER2 (77 MHz)	Non-Current
Wide P2SC	Model 595	POWER2SC (135 MHz)	Current
Wide PCI	Model H50	PowerPC 604e (332 MHz)	Current
High 1 (2-way)	Model R40	PowerPC 604 (112 MHz)	Non-Current
High 1 (4-way)	Model R40	PowerPC 604 (112 MHz)	Non-Current
High 1 (6-way)	Model R40	PowerPC 604 (112 MHz)	Non-Current
High 1 (8-way)	Model R40	PowerPC 604 (112 MHz)	Non-Current
High 2 (2-way)	Model R50	PowerPC 604e (200 MHz)	Current
High 2 (4-way)	Model R50	PowerPC 604e (200 MHz)	Current
High 2 (6-way)	Model R50	PowerPC 604e (200 MHz)	Current
High 2 (8-way)	Model R50	PowerPC 604e (200 MHz)	Current

[1] Current indicates the most recent node available; non-current implies that particular node has been superseded by newer models with superior price/performance

Figure 1. Common RS/6000 SP nodes

Memory is another factor to consider regarding node capacity. The SP nodes have specific memory slots available (these are separate for Micro Channel or PCI slots) but not every combination of memory will be available, particularly if memory cards already are in use in an SP node.

Occasionally, selecting a node can be a straightforward decision. For example, some applications may require a certain number of Micro Channel or PCI adapters, or a certain amount of memory. The more difficult decisions will be based on performance and, in particular, when to use uniprocessor nodes or SMP nodes in the SP system.

Node Characteristics

Each of the various nodes have certain functions and characteristics that make them well-suited to certain applications.

The 332 MHz SMP Nodes. The design of these PCI-based nodes has high RAS characteristics and excellent attention to detail. They can coexist with all previously supported nodes within any type of SP frame, provided the frame has been upgraded for new power supply requirements.

These PCI nodes are available in two formats: PCI Thin and PCI Wide. (The remainder of this article refers to the 332 MHz SMP Nodes as PCI Thin and PCI Wide Nodes.) These nodes are similar to previous Thin and Wide nodes: they occupy the same space in a frame and the same rules apply with respect to sharing switches between frames. The attraction is better price/performance using the latest technology.

The new PCI nodes are suitable for a wide range of applications, including all commercial and most scientific and technical applications. In some cases, the PCI Wide Nodes demonstrate superior performance to the PCI Thin Nodes, even though they share the same type and number of processors. This is because the Wide Node has a second PCI bridge (or controller) that attaches the additional eight PCI slots to the internal (MX) bus.

The performance differences can be significant in some I/O-related tasks. For example, with Serial Storage Architecture

(SSA) disk performance, the data rate could increase from about 40 MB/sec on a PCI Thin Node to approximately 70 MB/sec on a PCI Wide Node. This alone might provide a good reason to choose the Wide Node rather than the Thin Node.

Micro Channel Thin Nodes. The Micro Channel Thin Nodes have four slots available with a built-in Ethernet adapter that can be used for the RS/6000 SP Ethernet, which is mandatory.

PCI Thin Nodes. The PCI Thin Nodes also have a built-in Ethernet adapter and SCSI adapter for the internal disk. In addition, the optional SP Switch adapter does not use a PCI slot: it plugs into the MX bus and does not require a PCI slot. This leaves the two PCI slots available for normal use.

Micro Channel Wide Nodes. The Micro Channel Wide Nodes do not have a built-in Ethernet adapter. One slot always will be used for this purpose, leaving seven available slots.

PCI Wide Nodes. The PCI Wide Nodes have a built-in Ethernet adapter and SCSI adapter for the internal disk. As with the PCI Thin Node, the optional SP Switch adapter does not use a PCI slot: it plugs into the MX bus and does not require a PCI slot. Ten PCI slots are available for normal use. The PCI Wide Node consists of a PCI Thin Node with an expansion cabinet—fitted at the manufacturing plant—to provide additional expansion for adapters and disk. It is not possible to field upgrade from a PCI Thin Node to a PCI Wide Node.

Micro Channel High Nodes. The High Nodes, which use only a Micro Channel bus, have 16 adapter slots, but only 14 are usable because one is used for an Ethernet adapter and one for a SCSI adapter. In many cases, each node will require an SP Switch adapter, an alternate (user) network adapter, an Ethernet or Token-Ring adapter, plus some kind of disk adapter (SSA or SCSI).

In practice, Thin Nodes can have the available slots filled quickly. When High Availability solutions (HACMP) are required, Thin Nodes almost certainly will not be appropriate.

Figure 2 shows the Micro Channel and PCI adapter capacity.

Internal Disk Capacity

Internal disks contain the AIX operating system, any temporary files, and possibly application executables. For many commercial applications, the real data will be stored on external disks (for example, SSA or RAID) outside the SP nodes to allow access from another node if the primary node fails.

Paging space can be split between internal disks (the ROOTVG Volume Group) and external disks (SSA disks). It is supported to mirror the ROOTVG Volume Group using standard AIX mirroring, which may lead to additional internal disks. Currently, the RS/6000 SP does not support an

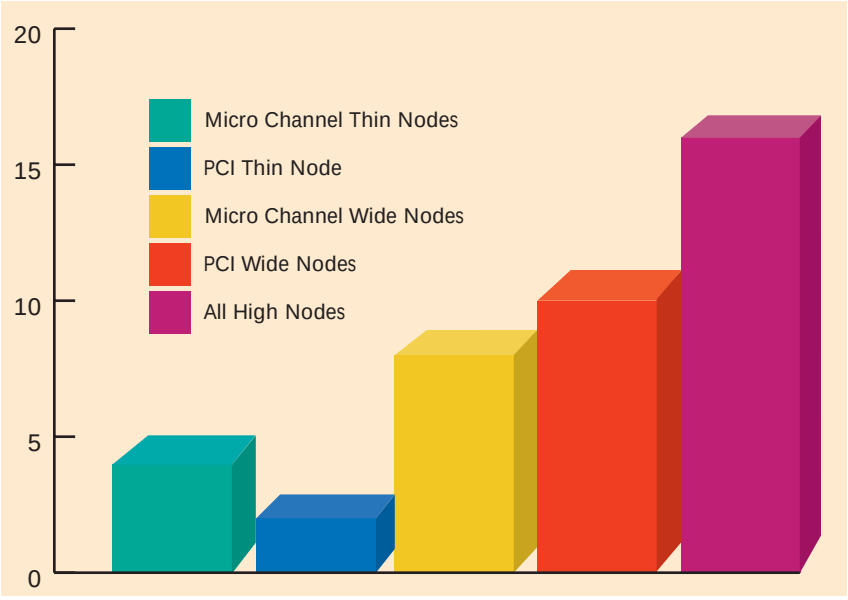


Figure 2. Micro Channel and PCI adapter slots

external ROOTVG Volume Group on SSA disks or booting off external SSA disks.

Maximum Memory Capacity

Memory cards are available for 32 MB, 64 MB, 128 MB, 256 MB, and 512 MB. Different memory cards can be used within the SP system, but certain combinations of cards cannot be used.

Thin Node. In a Thin 1 Node, only 64 MB, 128 MB, and 256 MB memory cards can be used. A Thin 1 Node allows a maximum of two memory cards of any type. The SP Thin 2 (39H) Node uses a combination of both memory cards and memory SIMMs (placed on the CPU board within the SP node). The memory card must match the CPU memory SIMMs card. Any of the 32 MB, 64 MB, 128 MB, and 256 MB memory cards can be used with the corresponding memory SIMMs on the CPU board. One of each always will be used.

Thin POWER2 Super Chip Node. This node has four memory card slots, which allows up to 1 GB of total memory. It doubles the memory capacity of a Thin Node and increases the memory bandwidth by 80%—to 1.9 GB per second. Four memory cards will give optimal memory performance in a Thin POWER2 Super Chip Node.

SP MCA Wide Node. The total number of memory cards must be two, four, or eight. If two memory cards are installed, they must be the same type of card. If more than two memory cards are installed in a Wide Node, they must be installed in matching groups of four identical memory cards. Optimal memory performance is obtained by using four memory cards in an SP Wide MCA Node.

PCI Node. This node uses memory cards specific to PCI nodes or systems. The cards cannot be interchanged with memory from MCA nodes or systems.

It is important to consider upgrade options and performance implications when choosing the initial memory configuration.

Other Non-Application Specific Selection Factors

There are several other factors that can help you decide whether to select a smaller number of more powerful nodes or a larger number of less powerful nodes. Those factors include: high availability, system and application management, application isolation, and system partitioning.

High Availability

High availability within the SP environment can be achieved in three ways:

- Provide a system that has a high degree of reliability and is less prone to error. The reliability of the standard RS/6000 components is proven, and AIX software is one of the best in the UNIX® world. In addition, the SP system, frame, and switch have additional functionality to provide higher reliability and availability.
- Provide ways within the SP to achieve concurrent maintenance when applications are not available due to regular maintenance and housekeeping work.
- Provide a failover functionality, to be instigated in the event of any kind of failure. The SP uses High Availability Clustered Multiprocessing (HACMP) software and component redundancy for this functionality. Redundancy within the SP is provided by RAID or mirrored disks, alternate networks, additional adapters, and spare or additional nodes used in the event of failure.

The new PCI nodes are suitable for a wide range of applications, including all commercial and most scientific and technical applications.

If high availability is important, consider using several less powerful nodes in combination, so if any single one fails, the rest will carry on. It is possible to build a highly

available solution in this manner without dramatically increasing the cost of the overall system. This also avoids the weakness of using a single SMP system to run all applications. A single High Node uses two drawers within the SP frame, whereas those same two drawers can house four Thin Nodes. This provides a greater degree of redundancy, and in some environments, a better solution.

These types of decisions should be based on a combination of required performance, price/performance of the different options, and a view of the level of availability required.

A system is less complex if it supports fewer applications. Also, a uniprocessor is less complex than a powerful SMP processor. If a failure occurs, you lose only part of your system. For example, losing one of the four Thin Nodes means that 75% of the total system and performance still is available. Since the possibility of four simultaneous failures causing the loss of the entire system is extremely remote, the four-node system gives several options for concurrent maintenance and housekeeping not afforded by a single node.

Two examples illustrate these considerations. Some applications, such as Lotus Notes®, will see no performance benefits from anything more powerful than a 4-way SMP system because of their design. Therefore, within the same space, you could implement either one 4-way High Node, or four Thin Nodes. The Thin Nodes provide better performance and also provide higher availability in terms of redundancy.

SAP™ is another example in which having more nodes is usually better than putting all of the workload onto one larger node. Running the application servers on separate nodes isolates and guarantees a given level of performance. Running everything on one large node (even if it were a powerful SMP-based node) will not give this level of flexibility or availability.

System and Application Management

As a rule of thumb, fewer nodes will make the SP system easier to manage. In reality, manageability depends on the type of applications running on the RS/6000 SP. For applications in which system management is not a complex task, fewer nodes will not be so important. But when the system is being used to consolidate servers, fewer nodes mean reduced management overhead.

SP system management is much easier compared to a similar number of separate servers. However, organizations often select a smaller number of nodes within the SP system. In fact, customers using their SP system for server consolidation or client/server applications (for example, SAP) have been the earliest adopters of SMP nodes.

Any parallel database environment that has mixtures of nodes is likely to make system and application management more complex. Running a parallel database on eight thin uniprocessor nodes and adding two SMP nodes into that parallel database requires more work to get adequate performance from the additional power afforded by the SMP nodes. In fact, for some parallel databases, it may be necessary to run additional workload on the SMP nodes to enable them to handle a greater workload. This will increase management and administration overhead.

Any parallel database environment that has mixtures of nodes is likely to make system and application management more complex.

The ideal in a parallel database environment is equal power nodes across the board. But in practice, as enterprises move to SMP nodes, the combination of uniprocessor and SMP nodes will occur.

Application Isolation

Isolation is one factor that may lead you to additional, less powerful nodes. Generally, UNIX systems are not good at running multiple applications on the same system. Shared resources often cause contention, and it is impossible to prevent the effect of one application on another.

For example, if applications A, B, and C run on the same UNIX system, and application A is used heavily, its increased workload on the system will have a detrimental effect on applications B and C. It is impossible to prevent this on a single (uniprocessor or SMP) system, which directly contrasts to an MVS® system, which is well-suited to running multiple applications simultaneously.

Different applications can run on different nodes within the SP system, with guaranteed performance and service levels for the applications on each separate node. This is important when running tests, development systems, or different relational database management system (RDBMS) software.

This capability is critical to SAP. Multiple nodes allow separate application servers, database servers, update and batch servers, and separate instances of SAP to run on the same SP system. A cost-effective approach may be to use a larger number of less powerful nodes to allow for application isolation rather than to run everything on a small number of SP nodes.

System Partitioning

System partitioning within the SP enables "logically" separate SP systems. SP environments can be isolated completely from each other, as in two production environments or a production and test environment.

Although SP system partitioning can assist with migration from AIX Version 3 to AIX Version 4 in an SP environment, this is not its primary purpose. The SP system no longer supports AIX Version 3. Combinations of AIX Versions 4.1, 4.2, and 4.3 on the same SP system or within a system partition are easy to run using the coexistence support in Parallel System Support Program (PSSP). For this reason, it is not necessary to partition the SP.

Today, only a few environments require system partitioning. Generally partitioning is done according to defined rules based on Switch Chip boundaries. SP Switch adapters are cabled into the SP Switch at the bottom of the SP frame and connected to various Switch Chips according to their position in the frame. The result is that certain nodes are located within a Switch Chip boundary.

The PCI Nodes follow the same rules of system partitioning and sharing a switch between frames as the existing Thin Nodes and Wide Nodes.

Frame Limitations and Node Upgrade Options

Both the old and the new tall frames support all types of nodes, including the PCI Nodes. The entry-level short frame also supports the PCI Nodes. Since the PCI Nodes require an upgraded power supply, the new tall frames include the power supply.

> ***The CPU upgrade path is popular for organizations that have High Nodes or PCI Nodes within their SP system and require more computing power.***

Both short, entry-level frames and the full-height SP frames support High Nodes (both PowerPC 604 based and PowerPC 604e based). The older entry-level High Performance Switch (LC-8 Switch) does not support High Nodes or PCI nodes. All of the other switches, such as the older, full 16-port version High Performance Switch (HiPS), and newer SP switches (both the full 16-port version and the entry-level SP Switch-8) support the High Nodes.

The PCI Nodes can be used only with the SP Switch (either the 8-port or the 16-port version) and not with the older High Performance Switch. Currently a single SP system cannot support more than 64 High Nodes.

Upgrading High Nodes (two to four or four to eight processors) is straightforward and cost-effective. Similarly, PCI Nodes can be upgraded from two to four CPUs. A PCI node begins with two processors. Thin Nodes are sold only in pairs (a full drawer).

The CPU upgrade path is popular for organizations that have High Nodes or PCI Nodes within their SP system and require more computing power.

The most difficult considerations involve organizations that are running parallel database applications, but would like to switch

to SMP nodes rather than a combination of node types. These upgrades must be carefully planned. It may not be possible to simply remove a Wide Node and replace it with a High Node, because the new High Node will take up two drawers in the SP.

Selecting Nodes Based on Performance

It is important to choose the correct nodes for the applications and have the ability to exploit the most appropriate architecture—serial or parallel, uniprocessor or SMP processor.

The performance measures shown in Figure 3 are estimates of OnLine Transaction Processing (OLTP). The numbers represent performance relative to an IBM RS/6000 Model 250, and simply show a comparison in performance between nodes.

Each type of application requires a different set of factors to consider in terms of performance of the various nodes.

Performance Considerations

It is important to differentiate between serial and parallel applications when evaluating the performance of SP nodes as well as other factors.

Serial and Parallel Applications. A single serial application will use only one node within the SP. It is common to run numerous serial applications on the same SP system, either as server consolidation or client/server applications. Server consolidation applications generally run several different applications on discrete nodes within the SP to reduce management and support costs. These applications often can be good contenders for running on an SMP node, particularly if the application has been designed and written to exploit an SMP architecture. This is not always the case with typical commercial database applications.

Each application running on an SMP node must exploit the SMP architecture, otherwise only one processor out of eight (in an 8-way SMP node, for example) is utilized. For example, Lotus Notes will typically scale only to a 4-way SMP system and will not take advantage of a 6-or 8-way system.

Node Type	Relative OLTP Performance
Thin	13.0
Thin	23.3
Thin POWER2 SuperChip (120 MHz)	5.8
Thin POWER2 SuperChip (160 MHz)	6.7
Thin PCI (2-way)	17.9
Thin PCI (4-way)	32.8
Wide	13.9
Wide	24.5
Wide POWER2 SuperChip	5.8
Wide PCI (2-way)	17.9
Wide PCI (4-way)	32.8
High 1 (2-way)	5.8
High 1 (4-way)	10.0
High 1 (6-way)	14.5
High 1 (8-way)	19.2
High 2 (2-way)	9.3
High 2 (4-way)	17.0
High 2 (6-way)	23.8
High 2 (8-way)	30.6

Figure 3. Relative OLTP performance

The High Nodes do not support scientific/technical applications. Since the floating-point performance and price/performance of the POWER2 Super Chip is superior to the PowerPC 604e processor used in the High Node, the POWER2 nodes are the obvious choice for these applications. A scientific or technical environment may need to run particular "commercial" workloads, such as file servers or communications gateways; these commercial workloads (non-floating-point workloads) can run on the SP.

The newer PCI Nodes (using the 332 MHz PowerPC processor) have good commercial (integer) and floating-point performance for various application types.

Several considerations for commercial parallel applications are written to exploit more than one node within the SP system. First, determine if the application supports an SMP (each individual node within the parallel system supports an SMP architecture). Second, make decisions about how to mix nodes within a parallel application. Selecting nodes of equal power and similar configuration may be a good

solution. However, if several POWER2 Thin Nodes already are running an application, and more power is needed by adding SMP nodes, these different types of nodes may lead to increased management/administration overhead.

Another scenario might be starting a serial application, such as a database server for an SAP solution. This serial database might run on one SMP node as it grows from two to eight processors within one High Node. For additional performance, you might implement a parallel database for an additional growth path. However, what would you add to the existing configuration of one large High Node? Initially, you may consider another 2-way or 4-way High Node, but again, this would create a performance imbalance between the two nodes—which you want to avoid.

A good solution would be an additional 8-way SMP node, but the cost may be too high. An alternative to upgrading the High Node would be to introduce a second node earlier, while it was still a 4-way or 6-way node. This would lead to two 4-processor nodes.

Scalability. The scalability of any application clearly depends on many factors. One is the ratio of CPU computing work compared to the communications workload. Nodes must be distributed evenly; and how well this is achieved depends on the compute/communications ratio.

Nodes of Varying Power. Another factor is how parallel databases handle nodes of varying power within the total configuration. Each database product that runs on the SP system continues to add more function. How a database handles different types of nodes will be a deciding factor in evaluating the options for SP system growth.

Price/Performance. The price/performance of the current nodes tends to be similar at any given time. You can select the most appropriate nodes for a particular application based on the best choice for the enterprise, rather than choosing particular nodes based purely on price/performance.

Server Consolidation (Serial Applications)

The most important consideration in a server consolidation environment is whether the specific applications support SMP processors and take advantage of such nodes. Some applications may exploit an SMP system only partially, possibly taking advantage of a 4-way or 6-way SMP system, but not an 8-way.

Server consolidation applications may be among the best for running on SMP nodes within the SP. Adopting SMP nodes can minimize the number of required nodes. But to isolate applications, you still have enough nodes within the SP for some level of potential redundancy from an availability point of view.

Server consolidation applications use only one node, and as the performance of the application grows, the maximum performance available on any individual node can be a gating factor. SMP nodes allow an application to grow from the smallest uniprocessor nodes to a powerful SMP node with greatly increased performance. Since AIX is binary compatible with each node, the same application can run unchanged from one node to another.

> ***SMP nodes allow an application to grow from the smallest uniprocessor nodes to a powerful SMP node with greatly increased performance.***

SP Switch Performance

Different nodes have varying performance characteristics with respect to the SP Switch or the High Performance Switch. This performance can impact node selection, either for nodes that act as a gateway or for scientific/technical applications in which SP Switch performance may be critical.

Performance in real life can vary dramatically, depending on tuning parameters and whether TCP/IP or User Space protocols are used across the SP Switch. Packet sizes have dramatic results on actual performance: a large packet size may give increased overall performance.

Sample Scenarios

Parallel Databases: Shared Disk Systems (I/O Shipping) Oracle Parallel Server.

For serial applications, the latest versions of the Oracle® RDBMS run on the RS/6000 SMP systems and SMP nodes within the SP. The SMP node is a good solution for serial applications.

A parallel database comprised of several SMP nodes running Oracle also may be a good solution if SMP nodes are required. A parallel database provides a highly scalable solution, supporting a multi-terabyte database. The level of scalability far exceeds a single SMP system or server.

A parallel database within an SP system can be built using uniprocessor or SMP nodes as the building blocks. The preferred option is running a parallel database with four 4-way SMP nodes rather than 16 POWER2 Wide Nodes, because this option has fewer nodes to maintain, manage, and administer.

The situation is not so clear when mixing nodes—uniprocessor and SMP nodes—within an Oracle parallel database solution. The most recent Oracle software has an intelligent optimizer that will indicate when certain nodes complete their work faster than others. It will give those nodes additional work the next time around. Over time, the SMP should accept more of the workload when compared with other less powerful nodes.

Parallel Databases: Distributed Disk Systems (Function Shipping)

DB2® Universal DataBase (UDB). In a serial database, the IBM RDBMS software (DB2/6000™) can run on SMP systems and exploit an SMP node within the SP. SMP nodes may be a good option for serial applications. UDB works differently in a parallel database, such as the Oracle Parallel Server. DB2 UDB partitions the data and spreads it across the nodes. As with all parallel databases, this is best suited to using equal power and similarly configured nodes across the entire environment.

DB2 UDB also works well with a parallel database running on SMP nodes. However, each database product achieves its objectives differently. The DB2 software maintains a catalogue node that tracks where the required data resides, and ships the SQL request to the node in question. With these types of parallel "function shipping" databases, the SQL sends a request to the remote node within the SP, as opposed to sending just the disk I/O request (I/O Shipping).

Informix XPS. PSSP 2.3 or later must be used to run a parallel Informix XPS database on SMP nodes within the SP system. The Informix parallel database uses the higher performing User Space protocol for communication across the SP Switch. The SMP nodes can be used when running a serial database.

OnLine Transaction Processing (OLTP)

Since OLTP usually involves many users, it is well-suited to an SMP environment. In effect, parallelism is built in to OLTP workloads because they typically process a large number of relatively light transactions. Response times are critical and expected to be fast. Applications such as CICS™, SAP, and other client/server implementations run well in the SP system, where individual components can run on uniprocessor or SMP nodes.

> ***A parallel database within an SP system can be built using uniprocessor or SMP nodes as the building blocks.***

When the system reaches a transaction rate of approximately 10 transactions per second, a TP Monitor can improve throughput and overall performance significantly. Good performance and scalability for OLTP applications on the SP system depend on careful system design and tuning.

The amount of I/O performed on a local disk versus I/O performed on a remote disk will affect scaling performance across nodes in a parallel database. The more disk activity that is local, the better the scaling. TPC-C transactions, for example, perform less than 15% of their I/O remotely. In contrast,

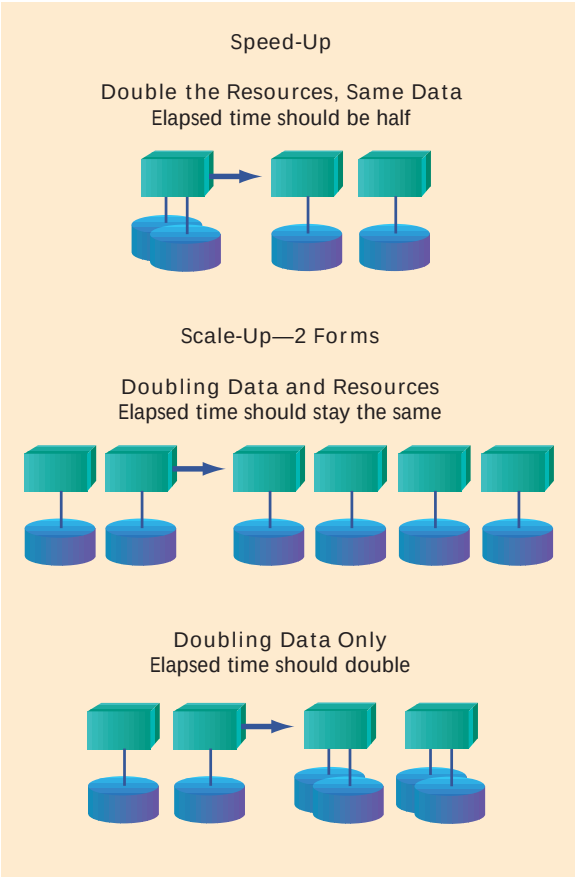


Figure 4. Speed-up and scale up

an application performing 50% of its I/O to remote disks would not scale well.

Decision Support

Decision Support System (DSS) applications, or business intelligence applications, use data warehouses or data marts for centrally storing data. Various applications access the data to gain business intelligence.

DSS applications require larger, more complex queries with fewer users. The database will run on a single node or in parallel if the most powerful SMP node is not enough. Database volumes with hundreds of gigabytes require a parallel database. Combining processors within an SMP system results in a diminishing return on investment as processors are added. However, the scaling factor for DSS workloads probably is better than for OLTP (TPC-C).

In a TPC-C workload, kernel activity requires a large percentage of time, and the operating system sometimes locks. In contrast, a DSS workload spends more time in the application, which results in better scaling, assuming that the application does not introduce an additional level of locking. Therefore, an SMP node running a TPC-D workload generally displays better scaling growth—from two to four or eight processors—when compared with a typical TPC-C type of workload. Such factors will depend on the actual application or database in use.

Figure 4 shows the processors required for scalability and performance.

SAP R/3. SAP R/3 is a well-suited and popular OLTP application. It can be advantageous to isolate some SAP components from others. For example, isolating the application servers and running them on separate SP nodes can guarantee service levels and provide redundancy. At the database server level, SAP (with Oracle, for example) can exploit a uniprocessor node, an SMP node (if more power is required), or a parallel Oracle database. High availability also can be provided for the database server; otherwise, this will be a single point of failure.

Lotus Notes/Domino™. Since Lotus Notes needs to replicate its databases between Notes® systems, it is well-suited to the performance characteristics of the SP Switch. The SP system management advantages also suit this type of application where customers often need to utilize multiple servers. Although Lotus Notes supports SMP processors, it rarely uses more than four or six processors within a single SMP system. Therefore, only an SMP node with four CPUs would be selected.

Using a single High Node with four processors for Lotus Notes may not be the only option. Within the same two drawers, you also could choose four Thin Nodes, which would provide more overall CPU performance and additional high-availability options.

The Lotus Notes Advanced Server can run more than one Lotus Notes “instance” on the same SMP system and also utilize clusters more effectively, as in the SP system. The Lotus Notes Partitioning

function allows multiple Notes servers to run on one SMP system, although each server uses its own memory section. Future requirements include the ability to "throttle back" certain Notes servers and provide dedicated resources to minimize the effect of one Notes server on another.

The Notes clustering functionality allows automatic load balancing and automatic rerouting of clients if a failure occurs. The SP system allows Lotus Notes to be implemented by using the HACMP software.

Scientific/Technical

Typical scientific/technical applications will not exploit the High Nodes; they need the higher floating-point performance of the POWER2 nodes or the PCI Nodes. The individual PowerPC 604 processors in High Nodes do not have high floating-point performance and would not be used for scientific/technical applications. The performance of the processors in the latest PCI Nodes for floating-point applications are enhanced greatly and better suited to single-precision floating-point workloads.

These applications take advantage of a parallel system such as the RS/6000 SP. In general, combinations of POWER2 uniprocessors are likely to be the best solution.

Summary

This article described the various SP nodes, how applications exploit them, and factors that you should consider as you select nodes for the SP system.

Clive Harris, IBM Corporation, Europe, Middle East, Africa (EMEA) SP Business Manager, EMEA Mid-Range Server Business, IBM Basingstoke, U.K.
E-mail: clive_harris@uk.ibm.com.