# DB2 Scales-Up on Wintel Big Iron

Chris Fierros

Microsoft's Windows 2000 Datacenter Server raises the bar for database scalability on the Intel platform. With support for 32 processors and 64 gigabytes of memory the decision to scale-up on this platform has never been easier. This article explores the unique features of Windows 2000 Datacenter Server and demonstrates how DB2 Universal Database leverages these features to scale-up on enterprise ready Intel based servers.

Resources

•DB2 UDB on Windows 2000
www.software.ibm.com/data/db2/udb/udb-nt/windows2000

Microsoft Windows 2000 Datacenter
www.microsoft.com/windows2000/datacenter

Unisys Corporation
www.unisys.com/windows2000/datacenter

## *Windows 2000 Integration*

Although roughly ninety percent of DB2 Universal Database on distributed platforms such as Windows, Linux, and Unix is common code IBM makes great efforts to integrate the product with the features and capabilities of each operating system it ports to and on Windows 2000 this is no exception.

If fact, according to IBM "*DB2 Universal Database makes use of more features in the Windows system than any other database on the market today, delivering record setting price/performance results.*"

DB2 Universal Database integrates tightly into almost all aspect of the Windows operating systems, including but not limited to:

- Active Directory
- Security Services
- Windows Services
- Event Log / Viewer
- Performance Monitor
- Clustering Services
- Terminal Services

## *Windows 2000 Packaging*

The Windows 2000 family of servers today consists of Windows 2000 Server, Windows 2000 Advanced Server, and Windows 2000 Datacenter Server. All Window 2000 server products build upon the already reach features of Windows NT 4.0 Server and Windows NT 4.0 Server Enterprise Edition.

### Windows 2000 Server

Windows 2000 Server provides the same basic support functionality of its predecessor Windows NT 4.0 Server. This version of the Windows operating system can is licensed to support from one (1) to four (4) processors and can address up to four (4) GB of memory.

Although limited to 4 GB of addressable physical memory, Windows 2000 server as well as all other versions of the operating system, include the Address Windowing Extension API. This API allows 32-bit applications to address real physical memory above the 4 GB line using a windowing approach that will be discussed in detail later in this article.

### Windows 2000 Advanced Server

Windows 2000 Advanced Server provides the same basic server functionality of its predecessor, Windows NT 4.0 Server Enterprise Edition. It includes all of the existing features of the base Windows 2000 Server, is licensed to support from one (1) to eight (8) processors, and can address up to eight (8) GB of memory.

Windows 2000 Advanced Server also includes 4 GB Tuning and the Microsoft Clustering Services. 4 GB Tuning allows the Windows 2000 operating system to make an additional 1 GB of memory available to applications. The MSCS can be installed as an optional component. As with Windows NT Server Enterprise Edition it provides support for two (2) node failover clustering.

### Windows 2000 Datacenter Server

Windows 2000 Datacenter Server is what the industry is coming to know as “Wintel Big Iron”. Windows 2000 Datacenter Server includes all of the features of Windows 2000 Advanced Server, can be licensed to support 1-8, 1-16 and 1-32 processors, and can address up to 64 GB of physical memory.

Windows 2000 Datacenter Server extends the MSCS support to include up to 4 nodes. Four (4) node clustering is an extremely attractive feature as it reduces the cost of failover clustering by twenty five percent (25%). It also works well in a partitioned database environment where multiple logical nodes fan out to more than one server.

## Windows 2000 Upgrades

There is no upgrade path from Windows 2000 Server or Windows 2000 Advanced Server to Windows 2000 Datacenter Server. This is because Windows 2000 Datacenter Server is not software that is burned on a disk and shipped in a box, but rather it is as a complete solution including hardware, software, and services.

## Datacenter Server Program

Microsoft designed the Windows Datacenter Program for businesses that require the most scalable, reliable, and available enterprise ready Intel based systems. The Windows Datacenter Program provides these customers with a complete solution that includes rigorously tested hardware, software, and support services. (source: Microsoft Corporation)

*Datacenter Server Hardware*

All OEM hardware must be submitted to the Windows Hardware Quality Labs (WHQL) and pass the Datacenter Server Program's Hardware Compatibility Test (HCT) before being placed on the Datacenter Server Program's Hardware Compatibility List (HCL). And then only the exact configuration submitted is HCL approved.

| **Vendor** | **Hardware** | **Processors** | **Memory** | **Uptime** |
|---|---|---|---|---|
| Compaq | Proliant 8500<br>Proliant DL760 | (8) PIII Xeon<br>700/900 Mhz | 16 GB | 99.9%<br>99.99% |
| Dell | PowerEdge<br>8450 | (8) PIII Xeon<br>700/900 Mhz | 32 GB | 99.9% |
| HP | Lxr 8500dc | (8) PIII Xeon<br>700/900 Mhz | 32 GB | 99.9% |
| IBM | xSeries 370<br>xSeries 440 | (16) PIII Xeon<br>700/900 Mhz | 32 GB | 99.9% |
| Unisys | ES7000 | (32) PIII Xeon<br>700/900 Mhz | 32 GB<br>64 GB | 99.9%<br>99.99% |

(Table 1. OEM Server Hardware)

Microsoft requires a 99.9% guaranteed uptime of all Datacenter Server Program OEM vendors. This equates to no more than 8 hours or less of unplanned downtime in a twelve (12) month period. Some OEM vendors provide an additional optional 99.99% uptime guarantee, but usually require a clustered failover server solution. The Unisys ES7000 Server is currently the only server providing 32-way SMP support. The Unisys ES7000 is capable of supporting 64 GB upon availability of 1 GB memory modules.

In addition to the OEM Vendors that provide server solutions for the Datacenter Server Program, **Datacenter Infrastructure Vendors** (DIV) provide infrastructure components such as Storage Area Networks.

*Datacenter Server Software*

All OEM vendor software that runs on the Windows Datacenter Server platform must be certified. IBM Corporation is a Microsoft Gold Certified Partner and DB2 Universal Database was the first database product to be Certified for Windows 2000. It was certified for Windows 2000 Datacenter Server in June of 2001.

*Datacenter Server Services*

In addition to providing hardware and software, OEM vendors that participate in the Datacenter Server program must provide Datacenter Server services. These services include some of the following:

- Installation Configuration Services
- Support Services
    - Joint Support Queue
- Change Control Management Service
- Reliability Measurement Services

The OEM vendor is required to install and configure Windows 2000 Datacenter Server. This configuration must pass an availability assessment test. They are required to provide 24x7 support services with onsite service guarantees.

A Joint Support Queue is staffed by both OEM and Microsoft personnel to ensure tight collaboration between the hardware and operating system vendors and has access to all OEM Datacenter HCL hardware configurations for problem reproduction and isolation. Customers have a choice between the OEM or Microsoft as first contact into the JSQ.

The OEM vendor must control and manage any changes to the Windows 2000 Datacenter Server operating system including Windows service packs, kernel level drivers, and hot fixes. The are also required to provide reliability measurements back to Microsoft in the form of Windows Event Logs, Dr Watson, Blue Screen of Death, and Crash Dumps.

## *Windows 2000 Scalability*

Microsoft designed the Windows Datacenter Program for businesses that require the most scalable, reliable, and available enterprise ready Intel based systems. The Windows Datacenter Program provides these customers with a complete solution that includes rigorously tested hardware, software, and support services. (source: Microsoft Corporation)

### Large SMP Processor Support

The Windows 2000 family of server operating systems today scales from one to 4-way SMP servers with Windows 2000, from four to 8-way SMP servers with Windows 2000 Advanced Server, and all the way up to 32-way SMP servers with W2K Datacenter Server.

The following screen capture of Windows Task Manager was taken while benchmarking DB2 UDB v7.2 on Windows 2000 Datacenter Server running on a Unisys ES7000 with 32-way SMP and 16 GB of physical memory. Note that total CPU utilization is 91% and that over 14 GB of memory has been committed. Another important note is that less than 100 MB of memory is currently used for the Windows system cache.

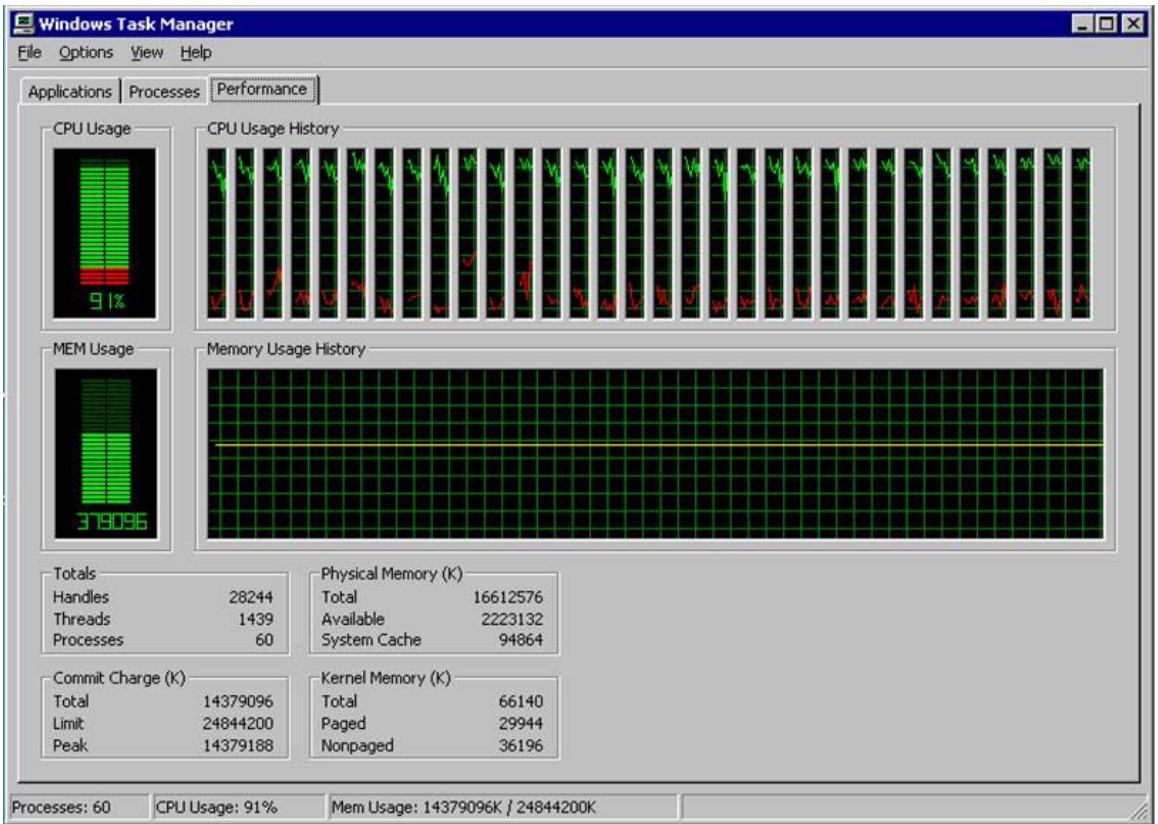

Figure 1. Windows Task Manager.

*Job Object Extension*

Windows 2000 Servers extend the process model to include the concept of a job object. The purpose of the job object is to allow a process or group of processes to be managed. For example, without the Job Object Extensions most processes can modify scheduling priorities and/or processor affinity, a process defined within a Job Object cannot.

*Process Control Tool*

In addition to the Job Object Extensions APIs that are included with all Windows 2000 server products, Windows 2000 Datacenter Server includes both a graphical user interface and command line interface into these APIs called the Process Control Tool.

The Process Control Tool provides a graphical user interface into the Job Object Extension APIs. This system administration tool, located in the Administration Tools folder, can be used to control single processes or groups of processes as a single unit by limiting scheduling priority, processor affinity, processor time, and memory utilization. This tool only provides the system administrator with a graphical user interface into the JOE API. The process of actually managing job objects is performed by the Process Control service.

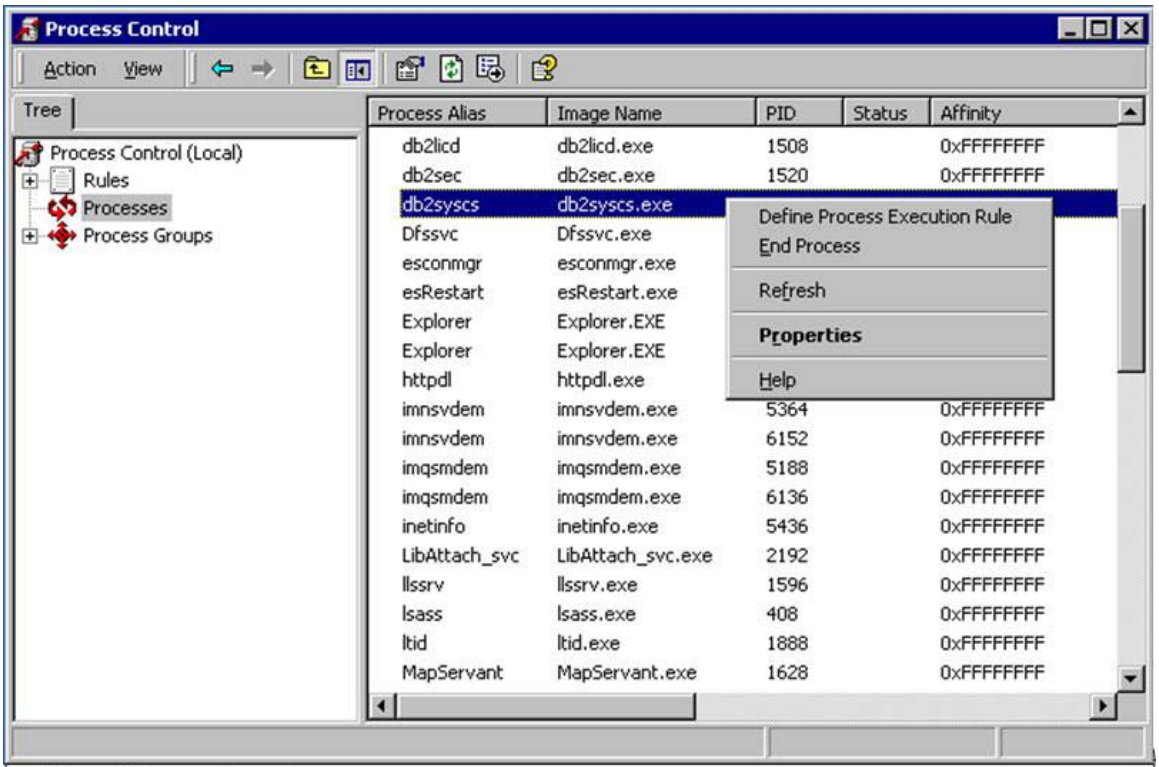

Figure 2 Process Control.

The Process Control Tool can be used to manage the DB2 Instance a.k.a. the DB2 system controller service (db2syscs.exe) process.

## *Large Memory Support*

The Windows 2000 family of server operating systems today scales physical memory from 4 GB with Windows 2000 Server, to 8 GB with Windows 2000 Advanced Server, all the way up to 64 GB with Windows 2000 Datacenter Server.

The following screen capture of Windows System Monitor was taken while benchmarking DB2 UDB v7.2 on Windows 2000 Datacenter Server running on a Unisys ES7000 with 32-way SMP and 16 GB of physical memory. Note that the DB2 System Controller Service process, db2syscs.exe has over 14 GB committed memory.

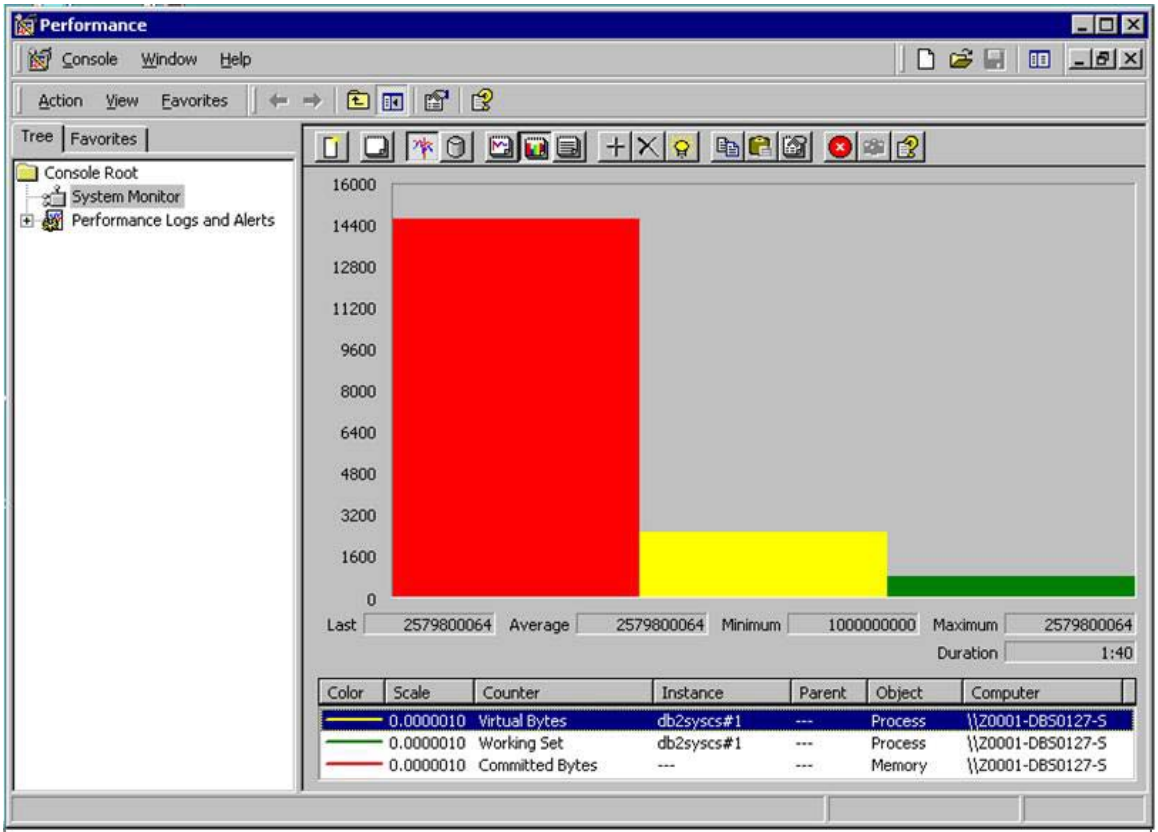


Figure 3. Windows System Monitor a.ka. WinPM.

*4 Gigabyte Tuning*

Windows 2000 Advanced Server and Datacenter Server both support 4 GB Tuning. Microsoft introduced the concept of 4 GB tuning, with Windows NT 4.0 Enterprise Server (Service Pack 3). This memory tuning feature, which can be enabled with the /3GB boot.ini switch, allows 32-bit applications that are aware of the /3GB switch to increase there virtual address space by an additional 1 GB of memory for a total of 3 GB virtual address space.

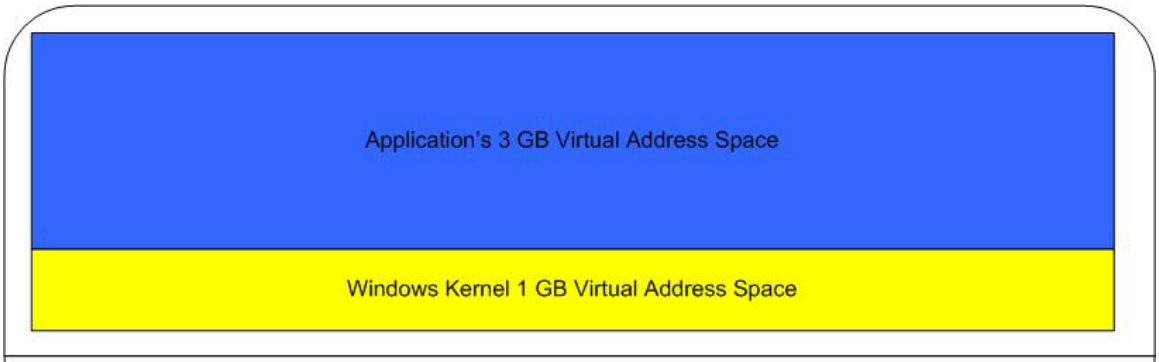


Figure 4. Windows 4GB Tuning with /3GB boot.ini switch.

Note that although Windows 2000 Datacenter Server supports 4 GB Tuning, but only on systems with up to 16 GB of physical memory. This is because the Windows kernel requires a full 2 GB of virtual address space to support physical memory larger than 16 GB.

DB2 UDB WE, EE, and EEE are all /3GB aware applications. This means that even on machines with as little as 4 GB of memory DB2 can allocate large buffer pools, up to 3 GB less DB2's working set memory.

This example shows how to create a buffer pool of 500000 (4k) pages or 2 Gigabytes.

```
--
-- Create Buffer Pool Example
--
CREATE BUFFERPOOL MyBigBad2GbBp SIZE 500000;
```

It is important to note that without the /3GB boot.ini switch after creating this buffer pool of 2 GB, any attempts to activate the database would be successful, however DB2 would return SQL1478W indicating that the defined buffer pools could not be started. Instead one small buffer pool for each page size supported by DB2 has been started.

*Address Windowing Extensions*

The Microsoft Address Windowing Extensions (AWE) API set provides 32-bit windows applications the ability to address up to 64 GB of physical non-paged memory. Although the AWE API is available on all versions of Windows 2000, support is limited to 4 GB of memory with Windows 2000 Server. Its primary purpose on Windows 2000 Server is for development and testing of applications that use the AWE API.

The Address Windowing Extension (AWE) API is enabled using the /PAE boot.ini switch. PAE, which is a Intel acronym for Physical Address Extension as it is the Intel IA-32 architecture that provides the memory addressing capability, Microsoft simply offers the AWE API via a 32-bit PAE enabled kernel.

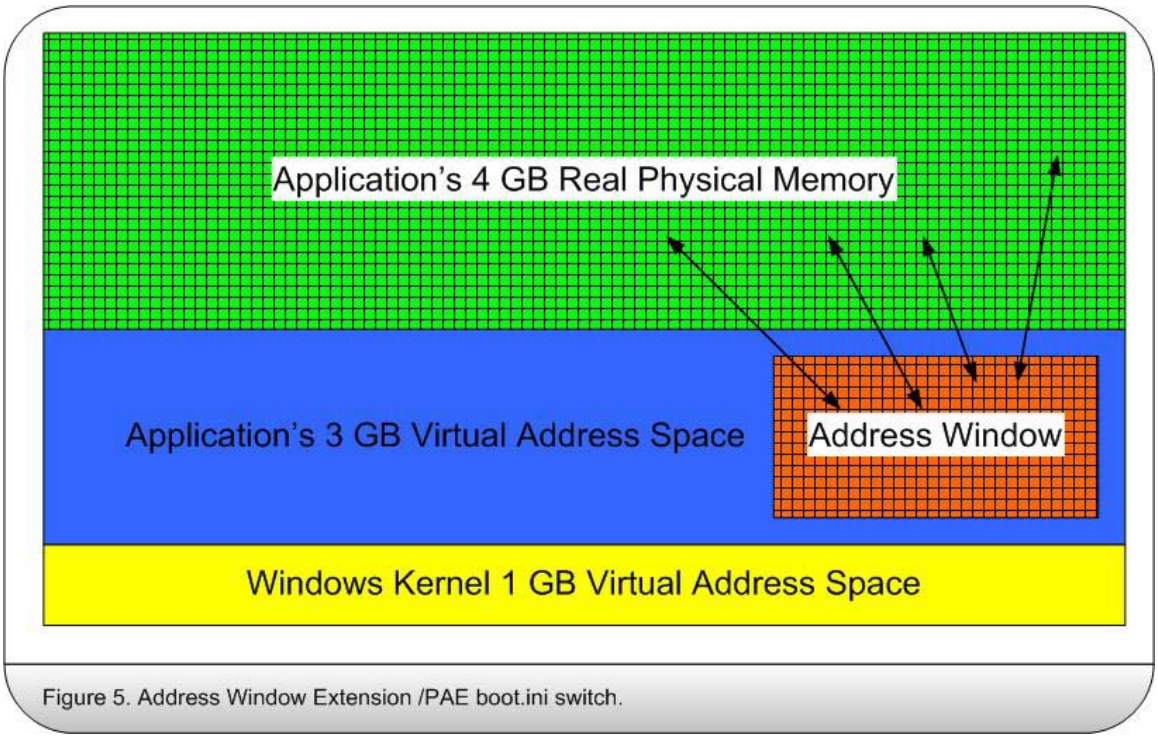


Figure 5. Address Window Extension /PAE boot.ini switch.

Note that without the /PAE boot.ini switch Windows will not load the PAE enabled kernel and the total system memory reported by Task Manager will be about 4 GB regardless of how much physical memory is actually installed. This is because the Windows kernel itself uses the AWE API for memory management.

DB2 UDB WE, EE, and EEE are all AWE enabled. This means that even on machines with 64 GB of memory DB2 can allocate larger buffer pools, up to 64 GB less DB2's and Windows combined working set memory on a dedicated database server.

It is important to note that not all software and/or drivers will function properly with the Windows PAE kernel on Windows 2000 Server or Windows 2000 Advanced Server. This is because AWE support is not required for certification on these versions of Windows, only on Windows 2000 Datacenter Server.

The DB2 registry variable, DB2_AWE, must be set in order to AWE enable a buffer pool. Given the granularity of the DB2 registry variables at the instance level, we must define only one database in the instance.

**Syntax:**

**DB2_AWE=bp_id, bp_size, aw_size**

**Where:**

**bp_id:** This is the id of the buffer pool to AWE enable. You can find the buffer pool id with a simple query on SYSCAT.BUFFERPOOLS.

**bp_size:** This is the size of the buffer pool in 4k pages. The entire buffer pool resides above the 4GB memory line and is backed by non-swappable physical memory.

**aw_size:** This is the size of the addressing window in 4k pages. The entire address window resides within the DB2 instances virtual address space.

This example shows how to AWE enable buffer pool #2 to use 16 GB (4,000,000 4K pages) of AWE memory with a addressing window of 1 GB (500,000 4K pages). We will also need to tune db2heap so that we have enough memory to manage the large buffer pool.

The bp_size must be at least equal to or greater than aw_size. If bp_size is smaller than aw_size your instance will start and your database will activate without warning or errors, but your buffer pool will not be AWE enabled. You will get a warning in the db2diag.log file, even if diaglevel = 1.

*DB2 Memory Management*

The DB2 registry variable, db2memmaxfree, controls how memory is managed by both the DB2 system controller process (db2syscs.exe) and engine dispatchable units (threads). Originally introduced as an AIX only registry variable, db2memmaxfree was first implemented on the Windows platform in DB2 UDB v7.1 FP3 a.k.a. DB2 UDB v7.2 with only the first parameter, max_keep. It was further enhanced in DB2 UDB v7.1 FP4 with the second parameter, min_keep after testing on large SMP servers indicated that memory latch contention was detrimental to performance.

**Syntax:**

**db2memmaxfree=max_keep,min_keep**

**Where:**

**Max_keep:** This parameter controls how much memory in bytes the DB2 system controller service (db2syscs.exe), a.k.a. the DB2 Instance, releases back to the operating system. On a dedicated DB2 server this parameter should be set as high as possible to reduce the overhead associated allocating and freeing memory (VirtualAlloc and VirtualFree).

**Min_keep:** This parameter controls how much completely unused memory segments each engine dispatchable unit (EDU) a.k.a. threads, keep in their own private memory pool. This parameter should be tuned to minimize latch contention on the DB2 system controller service's memory manager.

The DB2 registry variable, db2ntnochace, enables or disables Windows file caching on files opened by DB2. If enabled (set to 1), file system caching is not performed and the files are open with the NOCACHE option thus reducing the likelihood of double buffering.

**Syntax:**

**db2ntnocache=1**

In addition to DB2NTNOCACHE, it is also a good idea to change the Windows 2000 Datacenter Server memory optimization setting to minimize the operating systems use of memory for file caching. This is accomplished by changing the value of the "Server Optimization" setting on the "File and Printer Sharing for Microsoft Networks Properties" dialog to "Minimize memory used.

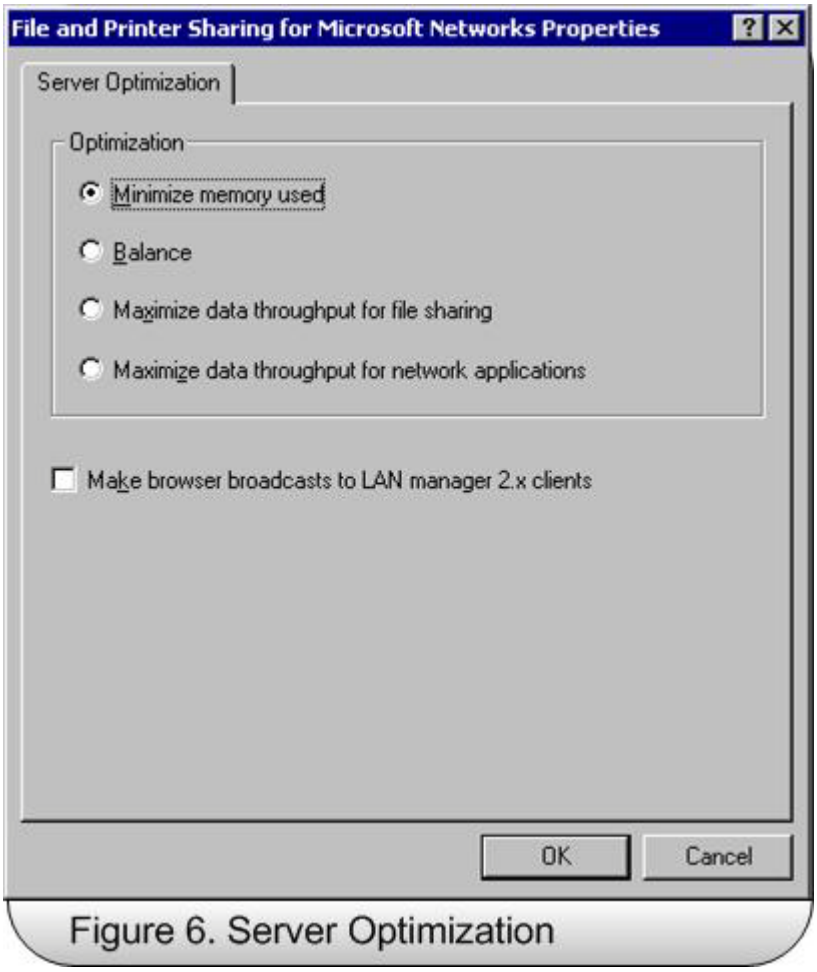

Figure 6. Server Optimization

### *Large Storage Support*

The DB2 registry variable, DB2_STRIPED_CONTAINERS, specifies that DMS table-space containers reside on stripped storage devices such as RAID5 or RAID10 arrays. DMS tablespaces that are created after this registry variable is set will have the first page in the container extended a complete extent size so that the container's extents line up with the RAID stripes.

**Syntax:**

**db2_striped_containers=1**

The DB2 registry variable, DB2_PARALLEL_IO**,** registry variable specifies that DB2 should attempt parallel I/O for all table-spaces on the specified device, even if they only have one table-space container. DB2 can perform parallel I/O on single containers that span multiple hard drives such as on RAID5 subsystems. A value of "*" can be used to indicate that all containers on all devices support parallel I/O. Prior to this registry

variable DB2 would only attempt parallel I/O if the table space was defined with two (2) or more containers.

**Syntax:**

**db2_parallel_io=***

## *High Availability Support*

Although not unique to Windows 2000 Datacenter Server, the Services properties Recovery tab can be used to configure automated recovery actions upon first, second, and subsequent failures of the DB2 Instance. Here we can see that the Recovery action on the first failure is to run a DB2 script that will modify the db2 diagnostic level (diaglevel) to the highest setting and restart the DB2 Instance.

*Fail Over Cluster Support*

Windows 2000 Advanced Server supports 2 Node fail over clustering, Windows 2000 Datacenter Server support 4 node fail over clustering. The db2mscs.exe productivity tool is used in conjunction with a configuration file to enable a DB2 Instance to support MSCS. During the configuration of the DB2 Instance, the Instance directory that is usually found in the \SQLLIB\ directory is moved to the shared storage device in \INSTPROF\ directory.

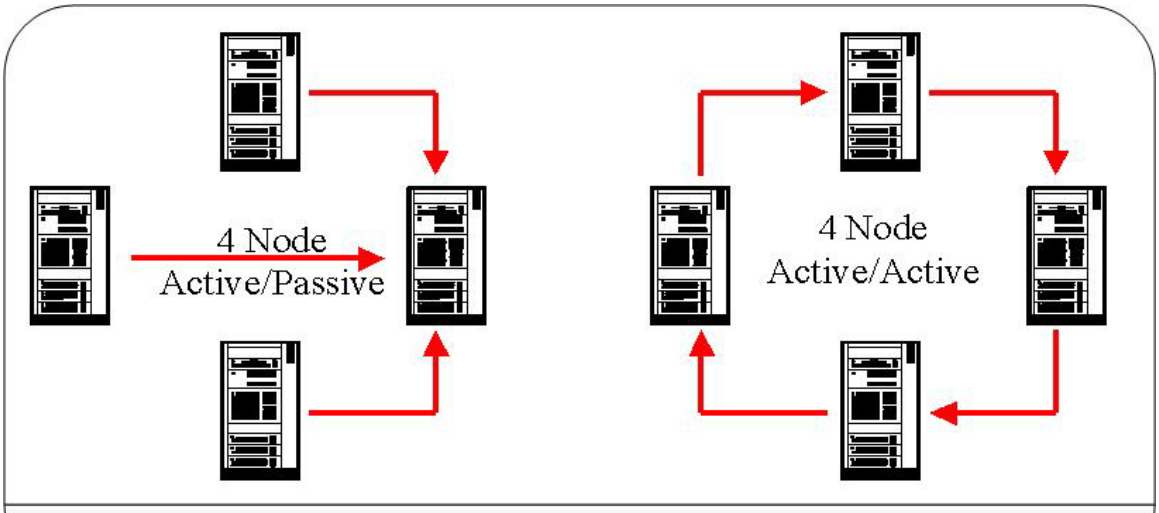


Figure 7. Four (4) Node Failover Clustering

The DB2 registry variable, DB2_FALLBACK, must be enabled if you want your DB2 Instance to fall back with MSCS to its preferred server immediately, otherwise the DB2 Instance will continue to run on the failed over node until the database is shut down.

Microsoft Cluster Service only supports basic volumes created on basic disks. Dynamic volumes created on dynamic disk that support simple, spanned, mirrored, or raid volumes (a.k.a. software RAID) cannot be clustered.

*Mirrored Log Support*

The DB2 registry variable, db2_newlogpath2, was introduced with DB2 UDB v7.2 and it is used to enable dual logging. The registry variable currently does not support a given path, but rather is simply enabled (set to 1) or disabled (set to 0). Enabling the registry variable creates a second set of mirrored logs in a directory located in the same path as the original logs but with the number two "2" appended to the log path thus requiring the use of volume mounting to place the mirrored logs on a secondary physical device. DB2 UDB v8 is rumored to support a database configuration parameter "mirrorlogpath" that will overcome this limitation.

The Windows 2000 mount volume utility should be used to mount a separate physical drive into the new log path subdirectory.

The release notes for DB2 UDB v7.1 Fixpak 3 indicate that… "Because Windows NT and OS/2 do not allow "mounting" a device under an arbitrary path name, it is not possible (on these platforms) to specify a secondary path on a separate device."… note that this does not apply to Windows 2000.

*Mirrored Database Support*

**DB2 Suspend I/O:** This new high availability feature, introduced DB2 v7.2, provides support for suspending database I/O so that mirrored storage devices can be broken. The split copy of the database can then be used for a query only reporting, offloading backups, and a hot standby server.

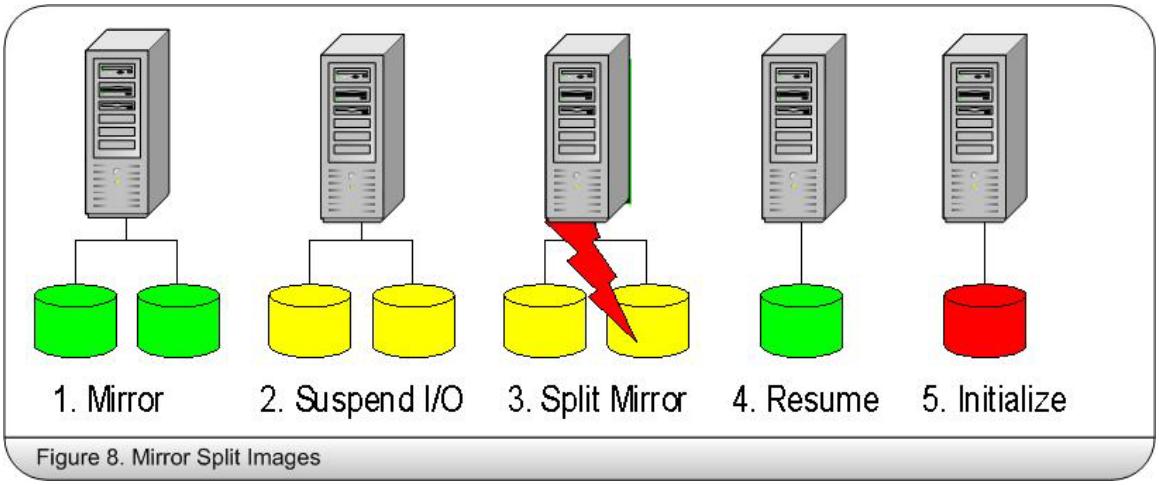

Figure 8. Mirror Split Images

**DB2 Initialize Database:** The initialize database utility is used to restart the split image of the database. The database can be initialized as a snapshot, standby, or mirror

database. Snapshot is used to initialize the database with new database transaction logs. Standby mode will initialize the database in a roll forward pending state. Database transaction logs can then be shipped over from the primary database via user exit routines and the database continuously rolled forward as the logs arrive. Finally, if the primary database is restored from a backup image created from the secondary it can be initialized in mirror mode allowing it to be rolled forward with the existing database logs.

**About the Author:** Chris Fierros is a DB2 Solutions Expert and principle consultant at Ten Digit Consulting, Inc. a DB2 Universal Database Company specializing in services, training, and support for DB2 UDB on Windows. He is an IBM Certified Solutions Expert in DB2 UDB Database Administrator and Application Development as well as an IBM Certified Advanced Technical Expert in Clustering, Data Replication, and DRDA. Chris has worked with DB2 on Intel platforms for over a decade and has spent the last year working with DB2 UDB on Windows 2000 Datacenter Server. He can be reached at chris@tendigit.com.