Paper 292

Taking Advantage of the SAS® System on the Windows Platform

Gary Mehler, SAS Institute Inc., Cary, NC

ABSTRACT

The Microsoft Windows environment has become an ubiquitous desktop application platform, and continues to gain capabilities that may lead to its dominance in the server market as well. This paper presents a discussion of the state of the Windows platform and how the Nashville Release of the SAS System takes advantage and integrates with the operating system. SAS System exploitation of advanced server hardware is also covered.

INTRODUCTION

In the history of SAS software on the Windows platform, SAS Institute has provided various methods of taking advantage of features and other applications available on personal computers. In addition to reflecting the native look and feel of the Windows platform, interoperability has been a hallmark of the SAS System for Windows. This interoperability has taken the form of support for e-mail, Lotus Notes, Windows Explorer integration, and other methods of interacting with client application software.

As the Windows platform has evolved as a capable server platform, this focus has shifted towards server exploitations of high-end hardware capabilities and large-scale processing abilities in Windows NT. These exploitations center around the most effective use of the SAS System on large NT Server systems in applications such as data warehousing and decision support. In one sense, taking advantage of the SAS System on the Windows platform is easy because SAS Institute has included much of the basic support in the software itself, ready for use. In another sense, though, the task is more complex than before, because server-based features can be technically more difficult to understand as they relate to the operation of a specific use of SAS software.

This paper will cover the current state of Windows operating system software and high-end server hardware likely to be applied to large-scale computational tasks. Some of this background will help the reader understand the quickly changing landscape of basic capabilities on personal computers.

The Windows Family

Microsoft Windows NT sales continued growth in 1998, rising to 25 million licenses, up from 11 million a year ago. Among all of the server platforms, NT claims about 36% of the market, according to various sources, and is growing at a rate of about 80% annually. The long-awaited upgrade for Windows NT 4.0 is still in the future, but more information about it is now known. Although we'd been planning to call the next version NT 5, Microsoft seems to be responding to customer confusion about the NT line, and plans to use a unified naming scheme in the future, beginning with Windows 2000. The Windows 2000 family will include some other renamings as well, such as the product hierarchy of: Professional, Server, Advanced Sever, and Datacenter Server replacing the existing products of: Workstation, Server, and Enterprise Server.

The names in the new product line will be significant for their gradual inclusion of features and scalability. For example, the number of processors supported in the product line listed above range through 1 and 16 processors, and memory limits range from 4 through 64 gigabytes (GB). As larger amounts of memory are used by the SAS System, understanding these limits becomes important.

The projected minimum hardware requirements for Windows 2000 have been upgraded as well: Microsoft recommends a minimum of 64MB memory and at least a 300MHz processor. Most newly purchased systems will meet these requirements, but older, smaller systems may struggle more under the increased load. The requirements have been raised because of a wealth of new features added by Microsoft, with estimates of 35 million lines of code comprising the Windows 2000 operating system. These new requirements will be covered as they pertain to use of the SAS System on the Windows platform. In terms of reliability, Windows 2000, like NT, will remain the safest environment for running mission critical and very large applications, as compared with Windows 95, under which an ill-behaved application can still impact another application.

The consumer-oriented Windows platform, formerly known as Windows 95, has moved forward in 1998 with the availability of Windows 98. Although still based on the Windows 95 kernel of a mix of 16-bit and 32-bit code, several

computing publications have concluded that Windows 98 is noticeably more reliable than its predecessor. Windows 98 also adds a collection of interesting features over the initial version of Windows 95, including web browser integration, FAT32 filesystem support, and other performance and manageability enhancements.

Although we'd previously been led to believe that Windows 98 was the last Windows platform based on the mix of 16-bit and 32-bit code, it seems likely that another version will be released before it becomes a truly 32-bit platform like Windows NT and Windows 2000. As of this writing, the next release is expected to be called Windows 2000 Personal Edition. Although it would share the Windows 2000 family name, it is essentially a different platform internally, and further discussions in this paper on Windows 2000 will focus on the operating system formerly known as NT.

The Pentium Family

During 1998, the Intel Pentium family of processors grew at both the high and low ends of the computer market. On the low end, the Intel Celeron processor, initially perceived as an underperformer, has recently become an excellent value. With the addition of a Level 2 cache (allowing the processor to run more effectively at the posted speed) as well as faster processors (currently up to 400MHz), the value end of personal computer lines has advanced very far during the last year. Advances from other vendors such as AMD have also led to more affordable performance and a wealth of consumer choices.

On the high-end, the Pentium II Xeon processor line has recently become available, and is the large boost many have been waiting for in the PC-based server world. With processor speeds up to 450MHz, Level 2 caches ranging to 2MB, bus speeds up to 100MHz, RAM limits to 8GB, and SMP capabilities up to 8 processors, this is a great leap from the previous server plateau of 200MHz Pentium Pros, with L2 caches of 256 or 512KB, bus speed at 66MHz, and SMP limited to 4 processors. The Pentium II line increased the clock speed, but didn't allow SMP past two processors, and kept the L2 cache at the same size and continues to require it to run at only 50% of the main processor speed.

Beyond the other enhancements available with the Xeon processor, one new feature is singularly significant: support for ESMA (Extended Server Memory Architecture). This architecture allows an application to utilize more memory than the operating system is capable of accessing. Since the Pentium family of processors are 32-bit based, the maximum amount of directly accessible memory has remained at 2^{32} bytes, or 4GB. ESMA support in the Xeon processor allows 36-bit access to 2^{36} bytes of RAM, or 64GB. This capability allows high-end applications to provide much faster access to data that would have previously required disk-based storage. The Version 8 of the SAS System takes advantage of this support, and is discussed later in this paper.

As of this writing, the Pentium III line has been announced for availability later in 1999. The future of Intel's 32-bit processor line holds several things in store: higher clock speeds, server-based throughput enhancements, and more performance optimizations available to application writers. SAS Institute is studying these developments closely and plans to take full advantage of any processor enhancements that lead to significant benefits for our products and customers.

Beyond the 32-bit world, news of the Merced processor has been available for some time. This processor is expected to ship sometime in the 2000 timeframe. SAS Institute views the 64-bit family of Intel processors (of which Merced is the first implementation) as very significant because of the vastly increased scalability it will allow for truly high-end applications. SAS Institute plans to support the Merced processor with SAS software available when the processor becomes available.

PLATFORM FEATURES

Beyond taking advantage of basic operating system and processor capabilities, the personal computer platform has evolved towards a first-class system on which to run mission-critical applications. SAS System support for these capabilities is covered in this section.

Interoperability

The SAS System takes advantage of interoperability with other applications available on the Windows platform. This is important because of the popularity of applications on the platform: SQL Server is now regarded as the second most widely-used database, up from fifth place two years ago. Exchange Server use is growing rapidly, with over 14 million seats added in 1998.

SAS System communication to applications such as Lotus Notes Mail or e-mail packages is handled through MAPI support in the SAS System. For Lotus Notes Mail versions through 4.x, VIM is the required protocol, and is also supported by the SAS System. As long as basic mail services are installed, the SAS System will

use the MAPI client interface to allow postings to packages like Microsoft Exchange Server and Lotus Notes Mail (starting with Notes Release 5). Since each of these client packages (again, with Notes support starting with Notes 5) also support connectivity with SMTP (Simple Mail Transfer Protocol), the SAS System can also post mail to SMTP/POP3-style internet mail systems by utilizing these client packages. In addition, if an administrator has set up routing for specific mail recipient identifiers to locations such as public folders, the SAS System can be used to direct mail directly into Microsoft Exchange Server public folders for collaborative data viewing.

The SAS System makes use of a web browser for support of hyperlinking, on-line help, and HTML-generated output. When Version 4 or greater of Microsoft Internet Explorer are present, the SAS System will utilize the browser as an application component within the SAS user experience for a more seamless user experience.

Security

An essential part of an industry strength client workstation or mission critical server platform is security. NT provides for per-file and per-directory security within the NTFS file system, as well as local desktop security. A user id and password is required for login to Windows NT. Across the network, APIs are provided for a single network login, and NT prevents use of resources by unauthorized users. In contrast, operation on Windows 95 or 98 don't enforce any rigorous security at the operating system level.

The SAS System takes advantage of the security features of Windows NT within the file system. In addition, SAS products such as SAS/CONNECT® and SAS/SHARE® only allow data access from authorized remote sessions.

File Systems

Windows NT and 2000 filesystem support includes FAT and NTFS. The FAT file system retains its compatibility with MS-DOS. The NT File System (NTFS) is a compressible, secure, logged file system that provides for 64-bit addressing, supporting files larger than 2 GB. Security can be set at the file or directory level, and logging helps ensure easy recoverability after an abnormal termination (such as a power failure). NTFS also supports fault tolerance, volume sets and striping in software. These mean that multiple physical disks can be arranged in larger configurations for increased reliability, growth, or performance. NTFS 5 will also support encryption at the filesystem level, which will be the best way to ensure data security at the system level.

Windows 98 added support for FAT32, which allows disk volumes to grow larger than 2GB Windows 2000 also can use this format, but if individual files are expected to grow past 2GB, NTFS is required. The SAS System supports the 64-bit addressing of NTFS, allowing SAS Data Sets to be accessed that are larger than 2 GB. Windows NT and the SAS System also support long filenames up to 256 characters long.

Of notable addition in Windows 2000 is operating system support for disk defragmentation. Particularly useful on large disks or when data are frequently modified, defragmentation can lead to significant performance enhancements since well-organized data tends to be quickly accessible data. Previously, a third-party application was required to analyze and correct fragmentation problems. When data change frequently, defragmentation is important and needs to be regularly-scheduled to keep system performance optimal. Stated another way: don't worry about any other issues (other than insufficient memory) until fragmentation has been minimized.

User Experience

Windows 98 and Windows 2000 add support for multiple monitors on a single computer. For knowledge workers, this can lead to increased productivity, since even large monitors eventually fill up with application windows. In a multi-monitor configuration, the user can drag windows from one screen to the other as user tasks evolve. The SAS System functions well in a multi-monitor environment with other applications.

Terminal Server Operation

Windows Terminal Server became a product during 1998, bringing multi-user capabilities to the NT platform. Terminal Server allows multiple users to share a single server system for the computational and other resource requirements (including disk drive use) while displaying a user session on a different PC. This communication occurs across a network, and the speed of user interface updates will depend on the speed of that network link. In this configuration, multiple uses are sharing the resources of a single NT Server system, so care should be taken to ensure sufficient resources are available to share.

General guidelines for behaving well on the Windows Terminal Server platform are outlined below and the SAS System's compliance with them are noted:

- Separate Application Bits and User Data. On a multi-user system, user data that is stored in a central location is subject to inadvertent damage by other users. The Nashville Release of the SAS System suggests default locations for user data that will be unique and secure from other users.

- Don't use hard-coded paths. Since the administrator could configure the system as desired, and Terminal Server systems tend to have higher security, assumptions about application paths need to be avoided. The SAS System doesn't store paths unless specifically requested by the user.

- Use environment variables wherever possible. Instead of using specific paths as in the previous point, use standard path information stored in the system registry. The SAS System looks up default values from the registry at installation time to configure itself appropriately.

- Use the "TEMP" directory to store non-persistent data. On a secure system, most locations won't be user modifiable. In addition, for security, each user has their own temporary space, writable to them, but not even readable by others. The TEMP location is guaranteed to be writable and secure by non-administrative users, and should be used for all temporary information. The SAS System suggests default settings for the user's SASWORK library to work correctly on Terminal Server.

Data Access - ODBC Support

The SAS System provides open data access support. To provide data access to SAS Data Sets from other applications such as Microsoft Excel or Microsoft Access, the SAS ODBC Driver provides support for both local and remote access. Local access is supported by a local SAS ODBC Server provided in BASE SAS software. Remote access to SAS Data Sets is provided through the SAS/SHARE ODBC Server. The SAS ODBC Driver supports data access through TCP/IP sockets to any remote TCP/IP accessible system running SAS software. For remote use applications, the ODBC Driver can be downloaded from the SAS Institute web site (http://www.sas.com), following links to software downloads.

SAS/ACCESS® to ODBC software provides access to any ODBC compliant data source. This allows data from sources external to the SAS System to become available for processing. This is now the preferred method of accessing data from Microsoft SQL Server as well as packages from other major database vendors.

Data Access – SAS Universal ODBC Driver

The Universal ODBC Driver allows other applications to access SAS data without requiring availability of the SAS System. This driver is new since Version 6 and supplements the basic SAS ODBC Driver that provides ODBC data access through the Base SAS System. While the SAS ODBC Driver provided both read and write access to data through the SAS System, the Universal driver provides read-only access.

This driver is significant because it allows other applications, such as Microsoft Excel, to directly access data stored in SAS formats without requiring the SAS application software. For organizations storing large amounts of SAS data on PC networks, this driver opens up the server data to more applications than ever before.

The Universal ODBC Driver is sold separately and operates only on Windows platforms. The driver provides access to the standard data sources, including: SAS data sets, Version 6 CFO Vision Financial Database (FDB) files, Version 6 SAS multi-dimensional database (MDDB) files, and JMP files. New for Version 8, the Universal ODBC Driver also provides access to SAS Transport Files, so that data from any SAS platform can be accessed by remote applications. For direct, native access, the Universal ODBC Driver also supports access to Version 6 SAS data generated on other platforms, such as UNIX or OpenVMS directly.

You can download an evaluation copy of the Universal ODBC Driver from the SAS Institute web site (http://www.sas.com), following the links to Demos/Downloads, ODBC Drivers, and SAS Universal ODBC Driver. This driver is also available for purchase at the same web location. Note that this driver is a single-time purchase item that will continue to be useful without renewal or other fees in the future, but that upgrades are not included in the base purchase price.

Windows NT Support

In addition to meeting the requirements for the BackOffice logo, additions to the system have been made in the following areas:

- Performance Monitor support. Operation of the SAS System can be monitored in the Windows NT Performance Monitor.

Performance Monitor is a powerful utility shipped with Windows NT that allows various performance metrics and counters to be displayed graphically or captured in log files for later analysis. The SAS System includes support for specific counters within the SAS System that can help administrators and site engineers determine whether optimal performance is being delivered by the SAS System during various operations. Performance Monitor allows knowledgeable users to do real-world performance analyses with their actual data and application workload. Performance Monitor can access data about the local system as well as from other NT systems on the network.

Performance Monitor operation doesn't compromise the operation of the SAS System. During normal operation, the SAS System will update internal counters during various types of operations. These counters are stored in a separate memory location also accessible to Performance Monitor. When monitoring SAS System operation, these counters are interpreted by SAS System code that runs within Performance Monitor to decode and analyze the values.

Using Performance Monitor with the SAS System is straightforward. The first step is to select counters to monitor. These counters are located under the SAS object in Performance Monitor. Then, use the SAS System normally. While Performance Monitor is running, it will capture information about the operation of the SAS System.

- NT Event Log support. Various diagnostic messages, such as Setup program configuration changes, remote user connections, and runtime errors can be logged to the Windows NT Event Log. Event Log is another administrative utility that is shipped with Windows NT that acts as a central repository for system and application errors or configuration updates. The SAS System supports writing diagnostic and configuration changes messages to the event log. Event log messages can be reviewed locally or for other systems on the NT network.

- Microsoft Systems Management Server Support. To the administrator on an NT Server, SMS is a good way to distribute and inventory software packages on the systems in a large organization. The SAS System functions as such an SMS package.

- Support for ESMA on Xeon Processors With the availability of the Pentium II Xeon processor, Intel-based servers gained the ability to access memory beyond the four gigabyte boundary. Using an architecture called ESMA (Extended Server Memory Architecture), an application can utilize this memory for very large-scale tasks. ESMA requires a driver available from Intel Corporation, and provides access to 8 to 32 gigabytes of RAM.

  The SAS System Version 8 takes advantage of ESMA as either a very large disk cache or in-memory library. When ESMA is utilized effectively, the speed of input/output operations can be greatly improved since the SAS System doesn't have to wait for a relatively slow (compared with RAM) disk drive system to respond to data requests. When in use, Windows NT memory and disk caching algorithms are bypassed, leaving them able to respond to smaller requests more effectively. In the ideal case of utilization of ESMA, the amount of real, elapsed time required to accomplish a processing task approaches the CPU time, which is typically far smaller when large amounts of data are involved. ESMA is covered in more detail in the next section.

- Scatter/Gather I/O For large data processing requirements, this type of operation has been added to Windows NT (starting with NT 4 Service Pack 3). The purpose of scatter/gather I/O is to allow data from non-contiguous areas of memory to be transferred to or from the disk in a single call to the operating system. In short, data can be *gathered* from various parts of the application and written to disk more quickly in some situations. In the reverse situation, data is read from the disk and *scattered* as needed in the application's memory.

  The SAS System can take advantage of this advanced form of I/O for greater throughput performance for certain types of operations. In this mode of operation, data bypasses the main NT filesystem cache, which can continue to be used for other operations. In addition, when used with sequential access patterns, the SAS System can direct the operating system to read-ahead data that will next be needed and hopefully return it before the SAS System actually needs that data. In the opposite case, data being written to disk can be processed by NT while the SAS System continues its main processing tasks.

  When large amounts of data are being processed, scatter/gather I/O can lead to performance improvements over standard operation. Unfortunately, it is difficult to state in this paper which types of

applications will see improvement over the standard operation of Windows NT. However, if multiple applications are being run on the same server system, scatter/gather I/O has a good change of effecting an improvement, since NT resources can continue to be used for other applications while the SAS System continues to optimally meet its data throughput requirements.

HARDWARE CONSIDERATIONS

In general, Windows NT Workstation (or Windows 2000 Professional) is the preferred environment for SAS data analysis and end user application deployment. It is highly recommended for advanced development with the SAS System, to have at least 16-32 additional megabytes of memory available for use by only the SAS System, above what is required for the operating system. Typically, this would suggest a minimum of 80-96 Megabytes memory on Windows 2000 Professional. Given the expected requirements for Windows 2000 in processor speed, the simple axiom of faster is better should be remembered. As user interfaces become more capable and larger amount of data are processed, machine speed becomes more important. Particularly as new computers become relatively inexpensive, there shouldn't be much keeping an important application running on an out-of-date, slow computer. If this is the case, Windows Terminal Server should be considered, so that application load can be directed towards a more capable server system.

The disk space required by the SAS System varies widely, but can range from under 100 Megabytes to 500 Megabytes for the entire system. The product can be custom installed for less disk utilization. On-line books and geographic map data are also available on separate CD-ROMs, and will require more space if installed to disk.

What you need to know about ESMA

As more enterprises realize the benefits of data warehousing and decision support systems, the size of the largest data requirements can grow beyond the capability of many computer systems to process in an efficient manner. On a typical computer, the ability to process data effectively and efficiently can be limited by the throughput limit of the hardware I/O system, which is responsible from moving information between main storage (disk drives) and the processing units (CPUs). Although large amounts of main memory (RAM) can be used to speed access to this information, the limit of Windows NT memory capacity (four gigabytes of RAM) can still be an impeding factor when hundreds of gigabytes of data are being used. In addition, new processors such as the Intel Pentium II Xeon greatly improve CPU throughput over its predecessors, but disk throughput remains relatively unchanged.

Thankfully, in addition to the performance improvement available in the Pentium II Xeon processors, Intel has also added support for much larger amounts of memory through an architecture called ESMA (Extended Server Memory Architecture). This architecture allows applications to take advantage of main memory beyond the Windows-NT limit of four gigabytes. In currently-available server hardware, this means availability of main memory of eight gigabytes to 32 gigabytes.

Beginning with Version 8 of The SAS System for Windows NT, ESMA capabilities are available to users of the SAS System's suite of data warehousing and decision support software. Through SAS Institute's support for this architecture, SAS applications can make direct use of memory beyond the four gigabyte limit to greatly improve access for various types of data processing operations. Initial tests have shown significant performance improvements for some types of operations that make use of ESMA, particularly those in which access to large amounts of disk-based data are required. In these cases, when much or all of the data can reside in very fast memory, as compared with relatively extremely slow disk drives, the central processing power can be more effectively utilized.

In addition to lowering or eliminating the overhead of waiting for disk systems to handle large data requests, ESMA has several other benefits when it is used effectively:

- when information is retrieved through ESMA, it can dramatically decrease disk channel traffic, resulting in greater performance of other operations that may be occurring on a server system.

- since ESMA can in some cases be immune to disk bottlenecks, main SAS System processing throughput can remain largely unaffected by other server activities, as long as central processing power is available.

- SAS System ESMA operations bypass the Windows NT disk cache, leaving it available to handle other application requirements on a server system.

- the amount of memory available to SAS System processing is limited only by

hardware availability, up to the theoretical ESMA limit of 64 gigabytes of main memory. This can be far greater than typical applications can access through standard access in Windows NT, which is limited to three gigabytes per application. When these very large amounts of memory are available, very impressive performance improvements should result.

PERFORMANCE IMPROVEMENTS

Every release of the SAS System since 6.12 has been performance tuned by SAS Institute, with the assistance of Intel Corporation. The Institute uses tools that help pinpoint potential performance issues in the application code so they can be addressed. In addition, the Institute has endeavored to utilize any reasonable capability of the operating system or machine hardware that results in more optimal performance on the Windows platform. This covers areas such as application code as well as the internal compiler used to build the system. Recent improvements have capitalized on exploitations of processor optimizations and parallelization.

FUTURE DIRECTIONS – REDUCING TCO

The future on the Windows platform means Windows 2000, ever-faster 32-bit processors, and 64-bit initial availability. Another important consideration is reducing the total cost of ownership (TCO), which can be high on personal computers due to their relatively higher quantity than centralized systems. In addition, the wide range of applications that may be installed on a PC may have varying degrees of correctness in their installation procedures. Finally, the requirement that some installation processes require administrative privileges is worrisome to administrators and end-users alike.

A proposed solution for these problems is use of the Microsoft Software Installer (MSI) service. This service will be part of the operating system in Windows 2000, and can be added to earlier members of the Windows family. On a secure system, MSI provides the only way for an application to make system modifications, such as updating system files and components, making registration entries, or copying application files. So, even if a user doesn't have administrative privileges on their PC, if the administrator gave them permission to install a package, that installation itself would be able to use administrative privileges in the operating system in a controlled fashion. While use of MSI isn't a requirement, it should become the safest and best way to install software in the future.

MSI also ensures that any installation process would be completely monitored, logged, and reversible. In addition, inter-application interactions should be minimized since unsupported actions shouldn't occur.

Beyond facilitating safe installations, MSI standardizes four basic installation types for general use:

- Run from local machine. All software is installed on the local computer, for best performance and least network load.

- Run from Source. The software is run from the source CD-ROM or network drive, minimizing local computer requirements, but maximizing network load.

- Install on Demand. In the first mode, components can be installed later, such as when an unusual feature is first used. In the second mode, an administrator can "advertise" an application by placing the icon on the Windows taskbar, without actually placing the software on the local computer. In this scenario, the icon appears "grayed-out" to indicate that it isn't exactly runnable without some intervention. When the user activates the icon, parts of the application are installed when they are first used.

- Don't Install. This allows an administrator to "hide" software components from inappropriate or unlicensed users.

Further benefits of MSI are automatic recovery, ensuring that the system will be left unaltered if the installation process fails for some reason, and automatic repair. Of these, automatic repair promises to be the most useful. If the user accidentally deletes part of the application, MSI will re-install the missing pieces and otherwise verify that all of the original installation steps have been retained. Automatic repair uses the same functionality as "Install on Demand", and checks out basic functions of the system.

Automatic "Detect and Repair" capabilities check every setting and file to ensure that all are present and the correct version. For example, if the user accidentally deleted a portion of the application, or if another installation changed a vital system setting, the repair process would set the system back to a known good state with respect to operation of the SAS System.

SAS Institute is actively researching MSI for full inclusion as it becomes available. MSI should help organizations deploy and manage their use

of the SAS System more effectively than ever before.

CONCLUSION

The Nashville Release of the SAS System for Windows operates on Windows 95, Windows 98, Windows NT and soon Windows 2000, but targets at least Windows NT for optimal reliability, performance, scalability, and maintainability. SAS Institute will continue investing in research and development for the Windows family and will embrace new technologies as they become available.

Trademarks

REFERENCES

Cates, Mark (1997), “Taking Advantage of the SAS System on Windows NT”, *Proceedings of the Twenty-second Annual SAS Users Group International Conference*

Clegg, Jennifer and Mehler, Gary (1999), “The Nashville Release of SAS System for Windows”, *Proceedings of the Twenty-fourth Annual SAS Users Group International Conference*

Mehler, Gary (1998), “Integrating Windows Clients and the SAS System into the Enterprise”, *Proceedings of the Twenty-third Annual SAS Users Group International Conference*

Author Information

Gary Mehler
SAS Institute Inc.
100 SAS Campus Drive
Cary, NC 27513
(919) 677-8000
gjm@sas.com