

Enterprise Application Scaling of Terminal Services on 64-Bit Microsoft Windows Server 2003

Derek Roberts
Senior Technical Consultant
Scapa Technologies

Charles Auger
Independent Consultant
eKnowlogy

May 2006

Version 1.0



Contents

Executive Summary	2
The Test	3
Test Tooling	3
Environment	3
Comparing 32-Bit and 64-Bit systems	4
Server Login Count.....	4
Server Login Response Time.....	4
Task Response Time.....	4
Throughput	5
Analysis of 64-Bit Limits	6
CPU	6
Memory.....	7
System Paged and Non-paged Pool.....	8
Conclusion.....	9
Contact Us.....	10

Executive Summary

Scapa Technologies worked with an independent consultant from eKnowlogy and a broad range of technology partners to prove the scalability of an enterprise application running under Terminal Services on the 64-Bit version of Microsoft® Windows Server 2003®.

Scapa Test and Performance Platform was used to analyse the performance of the entire solution: application publishing, load balancing, security, operating system and application, confirming the viability of a native Terminal Services approach to this high-profile application deployment.

Conventional wisdom suggests that a native Terminal Services solution would struggle in a deployment of this scale. However, a combination of 64-Bit hardware and a 64-Bit operating system was proved to work together to provide a highly scalable and secure solution at a significantly reduced cost over competitive approaches, without any modification to the existing 32-Bit application.

The 64-Bit version of Microsoft Terminal Services was shown to scale to 200 concurrent users on a single server, 3 times as many concurrent users as the 32-Bit version, and to deliver 2.6 times the number of business transactions per second. The corresponding reduction in the number of physical servers significantly reduces the overall complexity and cost of the solution.

Of particular interest is the effectiveness of the 64-Bit Terminal Services memory management architecture in breaking through the long-standing 32-Bit physical limits.

The Test

This white paper contains an extract of the results obtained from a large scale application stress test and sizing engagement which included a sizing exercise to confirm that a significantly higher number of concurrent users could be supported on an x64-based versions of Windows Server 2003, than on a 32-Bit version, while running a 32-Bit enterprise application.

Test Tooling

All stress testing and sizing tests were undertaken using Scapa Test and Performance Platform which is a unique systems performance testing and analysis product from Scapa Technologies. By using Scapa with Microsoft Terminal Services, it is possible to determine how many users each Microsoft® Terminal Server can support for both off-the-shelf and bespoke applications and how the server behaves with large numbers of users.

Scapa is unique in its ability to script at the object level (due to the sophisticated Scapa architecture) and includes with the fullest coverage of objects in the broad range of old and new user interface technologies in use in Enterprise environments. It also has verification – a functional testing technology which compares the state of an application with the expected state. This is particularly useful in Microsoft Terminal Services applications which can suffer from functional failure under load.

The Object-Oriented Recording capability within Scapa identifies objects by their internal object names, not by screen coordinates. If objects change locations or their text changes the script will continue to run.

Environment

- HP® Proliant® DL385:
 - AMD® Opteron® 8000 Series Chipset
 - 2 x AMD Opteron 280 (2.4GHz dual-core) Processor with 1MB L2 Cache per core
 - 16 GB PC3200 400MHz DDR Memory
 - 1 x NC7782 Dual Port PCI-X Gigabit Server Adapter (embedded)
 - Integrated Lights-Out (iLO) Standard Management on system board
 - Integrated Ultra320 Smart Array 6i Controller with 64 MB RAM
 - 1 x 1" Ultra320 15000 rpm SCSI hot plug hard drives (72.8GB)
 - DVD-ROM Slim 8/24 Drive
 - 2 x 575 Watt, 12 Volt hot plug power supply, CE Mark Compliant
 - 5 x hot plug fans
- The application was a customized version of a major two-tier client-server packaged enterprise vertical application
- Microsoft Windows Server® 2003 Enterprise Edition (32-Bit and 64-Bit Versions)

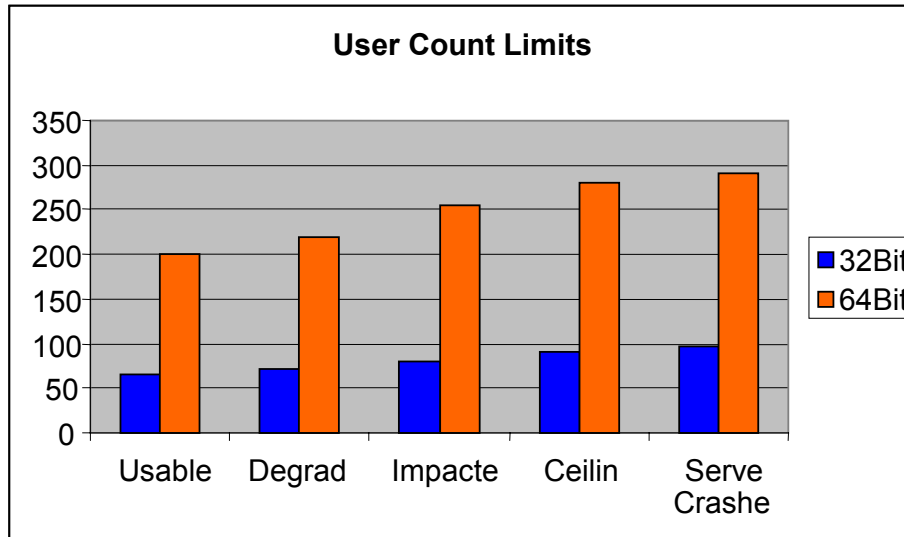
Comparing 32-Bit and 64-Bit systems

Server Login Count

A preliminary set of tests were executed to determine the maximum number of concurrent user sessions that are possible on a 32-Bit and 64-Bit Terminal Server of the DL385 specification mentioned above. The tests involved connecting directly to the application via the Microsoft Terminal Services Client, and continuously executing a simple usage script within each session. Below is a table illustrating the results

	Usable	Degraded	Impacted	Ceiling	Server Crashes
32-Bit	65	72	80	90	98
64-Bit	200	220	255	280	291

There is some subjectivity about where some of these limits were reached, but in general there is a factor of 3 between the numbers.



In subsequent sections of this report, results are presented for 70 users on a 32-Bit operating system and 280 users on a 64-Bit operating system.

Server Login Response Time

Users logged into the 64-bit operating system around 3 times faster than the 32-Bit version.

(Measurement in Seconds)

RDP Login Time	Users	Average	Maximum	Minimum
32-Bit OS	70	3.3954	4.4195	3.074
64-Bit OS	280	1.2836	1.732	1.105

Task Response Time

Scalability tests are presented for a single type of task, which involved retrieval, processing and modification of a customer record. The overall task

time consists of a combination of a fixed amount of time the user spends “thinking” and the time that the system takes to respond to the user. The data is shown for phases of the test where all users had logged on to the application.

The average task response time for 280 active users on the 64-Bit version of the operating system was 40% longer than that for 70 users on the 32-Bit version, but still well within the requirements of the customer.

Task Response Time	Users	Average	Maximum	Minimum
32-Bit	70	88.5	97.0	83.0
64-Bit	280	126.1	164.0	119.0

Throughput

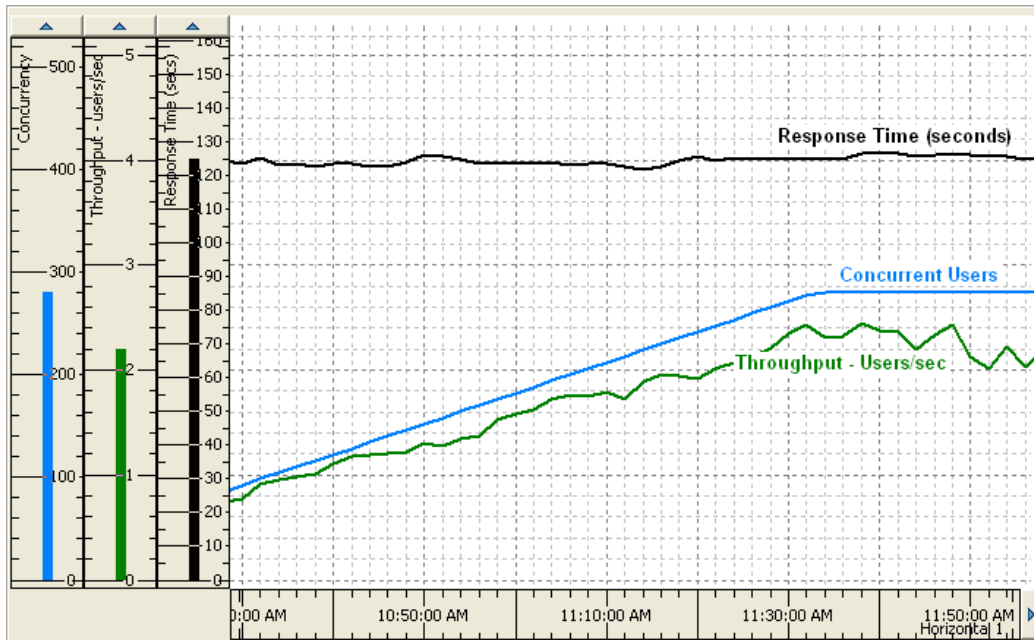
The throughput is the number of tasks that are being completed per second by the population of users. If the number of active users increases with a constant response time, one would expect the throughput to increase in proportion. Values were measured at steady-state.

Tasks per Second	Users	Sustained Steady-State
32-Bit	70	0.9
64-Bit	280	2.3

Analysis of 64-Bit Limits

We illustrate the process with screenshots from Scapa Test and Performance Platform. Our screenshots show a set of scales to the left, and a set of line graphs on a chart on the right. The scale against which the value of the line graph is plotted is indicated by a colored bar in the scale. The horizontal scale on the chart shows the time at which the metric was measured.

Various metrics are used in the diagrams. Color-coding is used consistently, and labels are given on the graphs for those reading in black and white. There are three core metrics derived from the running of the test itself. User Count (Blue), Throughput (Green) and Task Response Time (Black). Additional system-derived metrics are shown in other colours.



As the number of active users increases, the response time hardly changes, and the throughput increases more or less linearly, as one would expect.

CPU

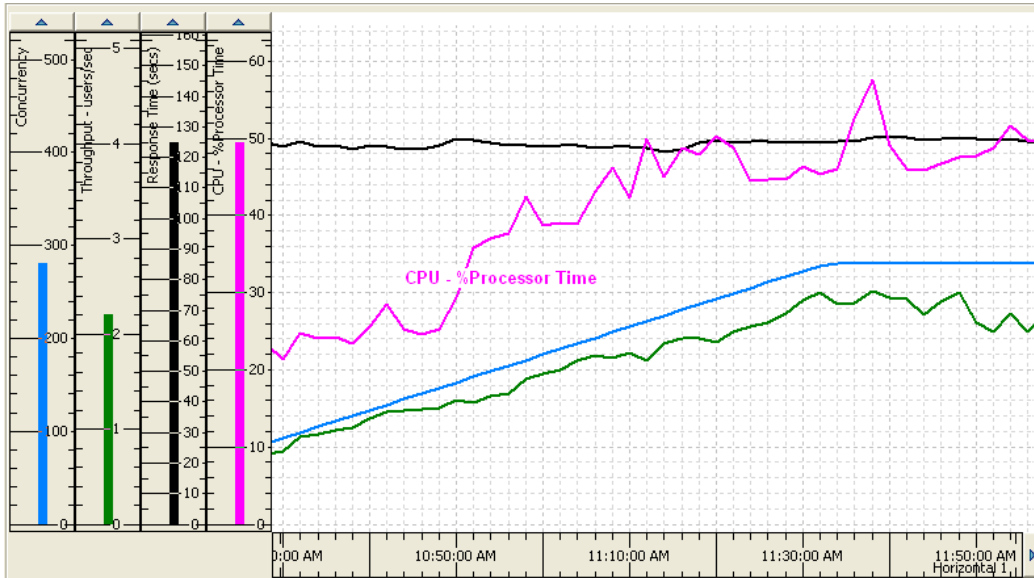
The most fundamental change in 64-Bit environments is the change in CPU architecture. It is important to remember, however, *that although the operating system and processor were both 64-Bit, the application was not.*

The CPU usage is shown in the following diagram in which we include an additional metric.

- CPU - %processor time (Pink). The activity level of the CPU (averaged across the two cores).

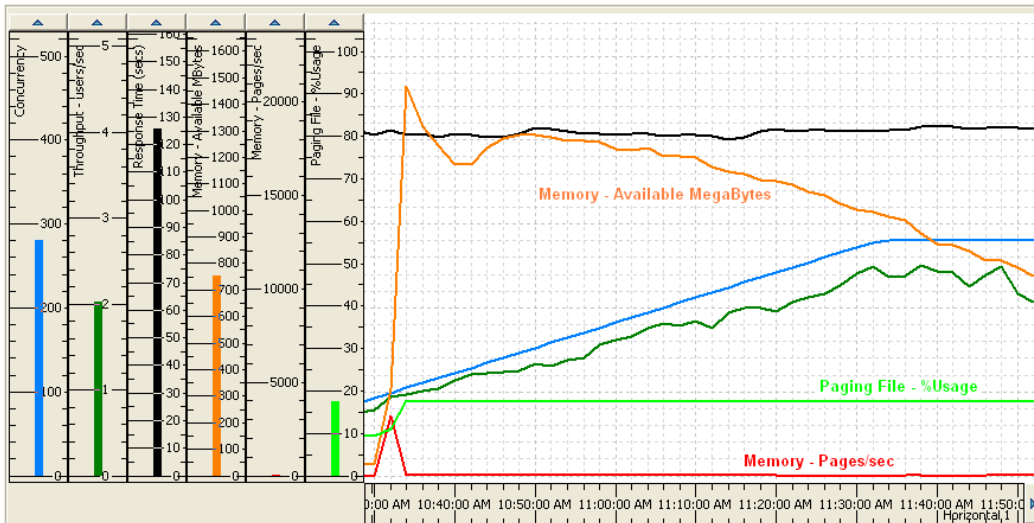
The system stabilizes at around 50% CPU usage with 280 active users. The CPU is a dual-core Opteron, capable of simultaneous execution of 2 parallel instruction streams. 50% usage is slightly suspicious and there is a possibility that the application is serializing and only able to use one of the CPUs at a

time. In practice, however, the 50% activity level is likely to be a coincidence as the CPU usage grows more or less in proportion to the number of users, and there is no increase in response time at higher numbers of users.



Memory

The second most fundamental change in 64-Bit environments is the increase in addressable memory, from 4G in the 32-Bit version of Windows 2003 to 1TB (in Windows Server 2003 64-Bit Enterprise Editions). The instability seen in 32-Bit environments when too many users are run on a machine tends to be associated with hitting this limit.



In the diagram three additional system-derived metrics are shown while the test is running

- Available MegaBytes - (Orange). This is the amount of physical memory available to processes running on the computer.

- Pages/sec - (Red). This is the rate at which pages are read from or written to disk to resolve hard page faults. This counter is a primary indicator of the kinds of faults that cause system-wide delays.
- Paging file %age usage - (Bright Green). The amount of the Page File instance in use in percent.

At the very start of the test, available memory is low, but page file usage is also low. Page file usage is at around 10% (or approximately 2.4GB), and available memory is around 100MBytes (indicating 15.9GB in use), corresponding to roughly 56MB in use per user, which is typical in Terminal Services environments deploying large vertical applications.

However early in the test, something happened which freed up significant amounts of available memory by paging to disk. The paging occurred once, over a short period of time, and didn't affect the response time of the system. Around an additional 7% of the 16GB page file was used (i.e. roughly 1.1GB) which seems to have freed up around 1.2GB of physical RAM.

A possible explanation of this behaviour appears on Microsoft Knowledge base article 889654 <http://support.microsoft.com/kb/889654>. It appears that in the 64-Bit version of Windows, additional paging activity can occur, relating to the loading and unloading of memory-mapped files in scenarios where page file usage is low.

As large numbers of users become active there is again a significant reduction in available memory (although not to the level seen at the start), which continues to reduce even after the number of active users stabilizes. This warrants some additional analysis as there may be a memory leak somewhere in the application which would lead to long-term instability, or more likely there may be some "lazy" memory management in the operating system which would eventually be triggered to free up memory.

Overall, there is nothing in this test that indicates that memory is actually a bottleneck. Once the initial page file activity is over there is no paging at all.

System Paged and Non-paged Pool

For completeness it is worth considering the two memory pools in Windows that have static limits. Paged pool is a region of virtual memory in system space that can be paged in and out of the working set of the system process. Non-paged pool is a memory pool that consists of ranges of system virtual addresses that are guaranteed to be resident in physical memory at all times and thus can be accessed from any address space without incurring paging input/output (I/O). Non-paged pool is created during system initialization and is used by Kernel-mode components (including hardware drivers) to allocate system memory.

Both the 32-Bit and 64-Bit versions of the windows operating system have static limits on the size of the paged and non-paged pool. In 32-bit systems, the exhaustion of non-paged pool can be a cause of instability and catastrophic system failure, particularly where drivers leak pool memory over

extended periods. The 64-Bit limits are sufficiently large to make it unlikely that a problem will occur.

This particular application did not actually approach the 32-Bit limits, even when running at 280 users, and Pool usage was constant once users were logged in. The limits and measured values are shown in the following table.

Architectural component	64-Bit Windows Limit	32-Bit Windows Limit	Measured at 280 users
Paged pool	128GB	470MB	190MB
Non-paged pool	128GB	256MB	140MB

Conclusion

There is nothing in this analysis that suggests there is a problem with running 200 users of this enterprise application on a single 64-Bit Terminal Server, and even at 280 users the system is reasonably stable.

The increased number of users per server, and thus the reduction in servers required is extremely high, significantly reducing the lifetime cost of maintaining and managing the server farm in an enterprise data center environment, and reducing the requirement for complex load balancing and management software. For some applications it makes a simple Terminal Services implementation a viable option without additional third-party O/S software.

Contact Us

For more information about Scapa Test and Performance Platform or to be put in touch with your local Scapa Technologies accredited Reseller, please visit our web site at www.scapatech.com or contact us at the addresses overleaf.

www.eKnowlogy.com

5 Jupiter House, Calleva Park
Aldermaston
Reading
Berkshire RG7 8NN
01344206134

Tel: +44 (0) 1344 206134
Tel US: +1 321 284 3833



Scapa
Technologies

Scapa Technologies Inc.
245 Park Avenue
NYC, NY 10167
USA

Tel: +1 212 792 4032
Fax: +1 212 372 8798

Scapa Technologies Limited
125 McDonald Road,
Edinburgh
EH7 4NW, Scotland

Tel: +44 131 652 3939
Fax: +44 131 652 3299

www.scapatech.com

Scapa is a registered trademark of Scapa Technologies Limited. All other company, brand or product names are either trademarks or registered trademarks of their respective companies.