IBM® TotalStorage®
SAN Volume Controller Release 3.1 Performance

Bruce McNutt and Vernon Miller
IBM Systems Group
International Business Machines Corporation

October 2005

NOTICES AND DISCLAIMERS

Performance information contained in this document was derived using IBM benchmark measurements. While each item has been reviewed by IBM for accuracy under the given conditions, the results obtained in your operating environment may vary significantly. Accordingly, this performance data does not constitute a performance guarantee or warranty. Please contact your IBM representative for assistance in assessing the performance implications in your specific environment.

Product data is accurate as of initial publication and is subject to change without notice.

No part of this document may be reproduced or transmitted in any form without written permission from IBM Corporation.

References in this document to IBM products, programs, or services do not imply that IBM intends to make such such products, programs or services available in all countries in which they operate. Any reference to an IBM program product in this document is not intended to state or imply that only that program product may be used. Any functionally equivalent program may be used instead.

The information provided in this document is distributed on an ”AS IS” basis without any warranty either express or implied. IBM MAKES NO WARRANTY EXPRESS OR IMPLIED, INCLUDING ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, NON-INFRINGEMENT OR INTEROPERABILITY. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer’s ability to evaluate and integrate them into their operating environment. Customers attempting to adapt these techniques to their own environments do so at their own risk.

IBM, TotalStorage, and AIX are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries or both.

Abstract

From its first introduction, in July 2003, the open-ended, horizontally scalable node architecture of the the SAN Volume Controller (SVC) has offered storage customers a flexible and robust capability to consolide and simplify large storage environments. On October 29, 2004, SVC Release 1.2.1 dramatically increased the potential for storage consolidation, by doubling the maximum number of supported SVC nodes from four to eight. The present white paper documents a further sharp increase in the performance capability of systems managed using the SVC architecture, with the availability of SVC Release 3.1.

1 Introduction

From its first introduction, in July 2003, the open-ended, horizontally scalable node architecture of the the SAN Volume Controller (SVC) has offered storage customers a flexible and robust capability to consolidate and simplify large storage environments.

On October 29, 2004, SVC Release 1.2.1 dramatically increased the potential for storage consolidation, by doubling the maximum number of supported SVC nodes from four to eight.

The present white paper documents a further sharp increase in the performance capability of systems managed using the SVC architecture. As of October 28, 2005, SVC Release 3.1, when running on Storage Engine 336 node hardware, supports an increase of over 50 percent in the maximum I/O throughput of the SVC, as measured using 512 byte read hits (from 805,000 to 1,230,000 I/O's per second). This paper demonstrates the performance and scalability that SVC Release 3.1 now provides, using a variety of random-access and sequential workloads.

SVC Release 3.1 performance was tested using a Storage Area Network (SAN) built with IBM TotalStorage DS4300 Storage Servers, as detailed in Section 2. Although many factors may impact storage system performance, it is anticipated, in the case of a similar SAN built with IBM TotalStorage DS4800 Storage Servers, or with other storage technology that provides comparable or better cache size and I/O handling capability compared with the DS4300, that comparable or better overall performance would be achieved.

Simpler management is the central goal of storage virtualization. Section 3 demonstrates, however, that virtualization using the SVC architecture may also offer a substantial performance advantage for a variety of workloads due to the SVC's caching capability, as well as its ability to "stripe" all host data automatically. Meanwhile, the "latency" incurred by the virtualization layer during system testing (the delay to pass a read miss request through to the physical disks) had so little impact on ordinary system-level measures of performance that it became "lost in the noise".

Sections 4 and 5 explore the ability of the SVC to act as the virtualization layer for systems requiring very high levels of I/O performance. These two sections show the scalability of the SVC both for database as well as sequential storage environments.

2 Test SAN

Figure 1 presents a diagram of the laboratory SAN used for SVC performance tests. For the benefit of any readers who might wish to use Figure 1 as an example, or design "starting point," it is appropriate to point out some of the design choices reflected by this SAN topology:

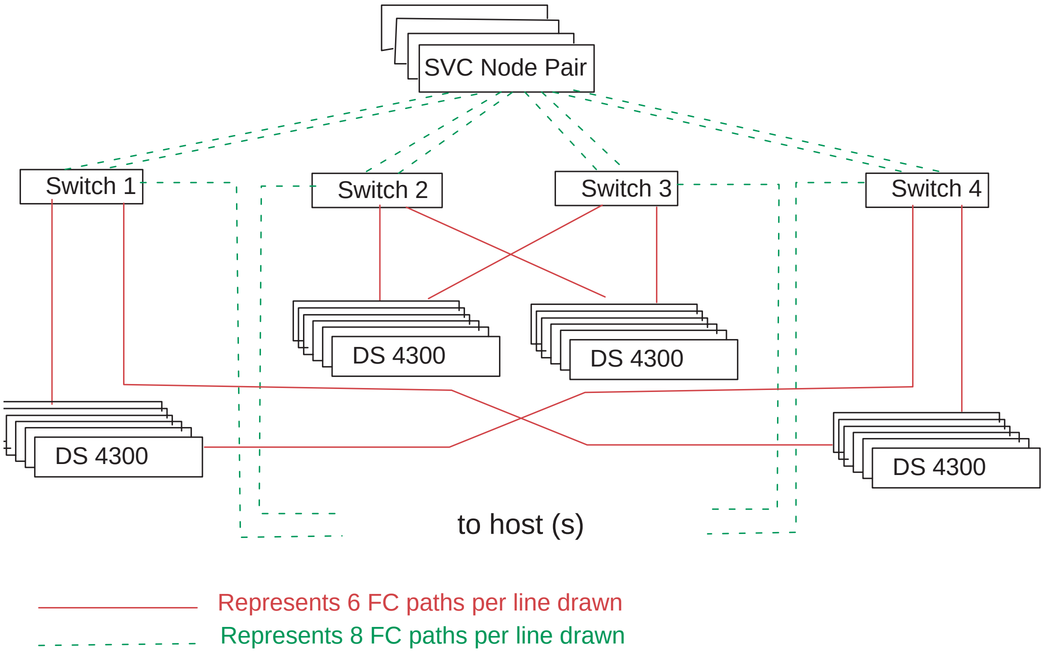

Figure 1 *Test SAN Configuration. Each DS4300 attaches one EXP700 expansion drawer, for a total of 672 15K RPM disks.*

Component	Setting	Value
SVC	Extent Size (MB)	256
SVC	Management Mode	Striped
DS4300	Segment Size (KB)	256
DS4300	Readahead	1
DS4300	Cache Mirroring	Enabled
DS4300	RAID-5 Array Size	12+P+S
DS4300	RAID-10 Array Size	14
AIX	hdisk queue depth	128

Figure 2 *Miscellaneous SVC and DS4300 settings.*

- A *dual fabric* is provided; that is, the fabric divides into two self-sufficient halves (left and right), either of which by itself can provide full connectivity. This type of fabric redundancy allows a range of “quick-and-dirty” repair and maintenence alternatives, since service can be interrupted to multiple communicating components, or even an entire half of the fabric.

- Every node connects to each switch using an identical number of switch ports (the actual number of switch ports, given the dual fabric, turns out to be either one or two).

- Every DS4300 enclosure connects to one switch on each half of the dual fabric.

- To prevent a mix of direct and indirect path alternatives from being found during host LUN discovery, no interswitch links were used.

During host discovery, both the SVC host path definitions and the switch zoning affect the number of occurrences of a given vDisk that the host will identify. It is recommended that the number of occurrences should be kept to four or fewer, because increasing the number to more than four does not improve either performance or reliability, and may cause more time and trouble for the administrator responsible for the host OS.

Tests were done with a variety of strategies for these two related elements of an SVC layout. For a SAN topology with the characteristics just presented, the following strategy seems to suggest itself as a simple and effective approach:

- Pair off the available host ports, so that in each pair there is one member that sees the SVC through each half of the dual fabric.
- Assign each vDisk to one pair of host ports.
- If each SVC connects to a given switch via a single port, define a single host zone in the switch, shared by all SVC node ports and all host ports. Otherwise, define two host zones in the switch, splitting the ports associated with a given SVC node between them; divide the host ports as evenly as possible between the two host zones.
- Define one storage zone in each switch, shared by all SVC node ports and all storage controller ports.

With this design, the number of vDisk occurrences that the host sees during discovery will be equal to four (the vDisk will be seen twice via each side of the dual fabric, once on the preferred node and once on the alternate node).

Figure 2 documents a variety of SVC and DS4300 parameter settings which were used in the lab configuration. No assurance can be offered that this specific combination of settings is “optimal”; however, the use of these settings to obtain the levels of performance documented in the present paper shows that they are at least reasonably effective.

The SVC configuration presented in Figure 1 is designed to demonstrate the performance capability of eight SVC nodes, and for this reason includes a larger number of nodes than would normally be configured to support the disk storage capacity (up to 45 terabytes when configured as RAID-5). A more “balanced” system would support the same storage capacity with four rather than eight nodes.

The present paper includes tests of the configuration of Figure 1, in which either two, four, six, or all eight of the nodes are actually used. For ease of comparison with tests performed previously, the present white paper also includes some tests on a storage configuration of two nodes and 168 disks. The configuration of these tests corresponds to the inner two switches of Figure 1, with half of the DS4300’s excluded in order to reduce the number of disks.

3 Virtualization Net Gain

In considering the potential performance effects of providing a “virtualization layer,” many knowledgeable observers make the mistake of focusing on the delays that might occur within such a layer during I/O processing. As measured in the lab, however, such delays were very short. For example, the delay introduced in servicing a 4K read miss was measured to be approximately 60 microseconds, or 0.06 milliseconds. Such a short delay, in the context of read miss processing, would tend to be invisible under production conditions. On the other hand, the SVC offers the opportunity to significantly improve performance through striping, caching, and other virtualization services.

In the author’s view, the capability to stripe across disk arrays is the single most important performance advantage offered by the SVC. In the past, the use of striping across disk arrays has been possible through various host software offerings; however, this has required a degree of pre-planning which many or most system administrators have found to be impractical. By contrast, the SVC can automatically stripe all vDisk images provided to the host; and it can do so across the entire set of supported physical disk resources.

The use of striping can have an important impact, for example, in database environments that resemble the Storage Performance Council benchmark, SPC-1. This benchmark, designed to reflect typical production conditions in a server running OLTP or mail server applications, features “hot spots” with a realistically high concentration of demand.

The SVC’s large cache and advanced cache management algorithms also allow it to improve upon the performance of many types of underlying disk technologies. The SVC’s capability to manage, in the background, the destage operations incurred by writes (while still supporting full data integrity), can be particularly important in achieving good database and/or SPC-1 benchmark performance.

As should be expected from the discussion just presented in the previous paragraphs, the SVC cannot increase the throughput potential of the underlying disks in all cases. Its ability to do so depends upon both the underlying storage technology, as well as the degree to which the workload exhibits “hot spots” and/or sensitivity to cache size or cache algorithms.

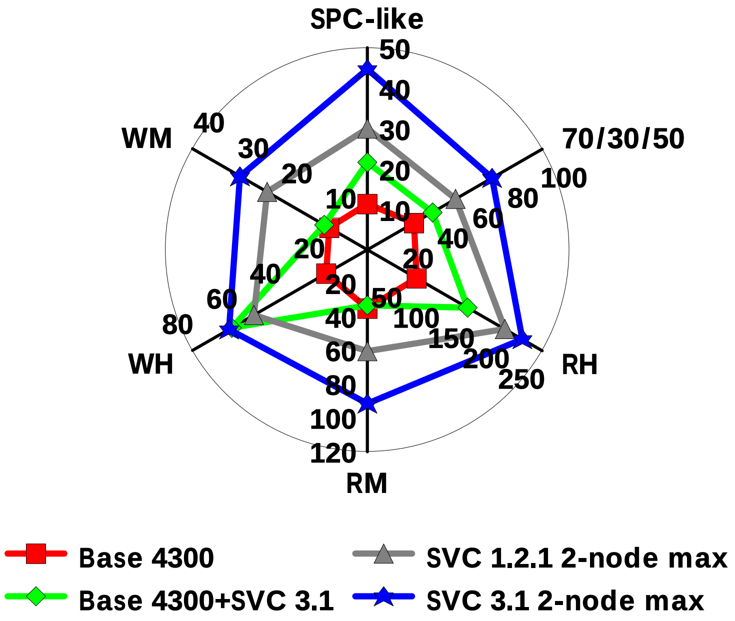

Figure 3 *Performance of the SVC “virtualization layer”. The two innermost curves used 168 RAID-5 disks and 4 fibre channel host attachments.*

Figure 3 provides an overview of the performance impacts for a variety of benchmark I/O workloads, when putting SVC in front of a disk storage technology with a small cache and less advanced cache algorithms. For the purpose of Figure 3, the base model of the DS4300 was used to illustrate this type of older disk storage. In addition to an SPC-1-like test[1], the figure includes results for:

- 70 percent reads/30 percent writes/50 percent read hit, with 4K transfer size (70/30/50).
- 4K read misses (RM)
- 4K read hits (RH)
- 4K write misses (WM)
- 4K write hits (WH)

[1] The SPC-1-like tests reported in the present paper used the mix of I/O workload components as defined in the SPC-1 specification, but were not performed under the conditions required for SPC-1 audit certification.

Throughout this wide variety of tests, the SVC allowed the full capability of the underlying disks to "show through." For some workloads (e.g., read miss), performance is essentially the same with or without SVC virtualization. But for a number of workloads, particularly the SPC-1-like workload that most closely resembles a typical production environment, the SVC can significantly augment the performance capability of the native storage.

4 Database Scalability

We turn now to a discussion of the SVC's capability to scale up to very high levels of I/O demand. The present section focuses on database I/O demands, as reflected by the SPC-1-like workload; the following section then examines SVC scalability for sequential demands.

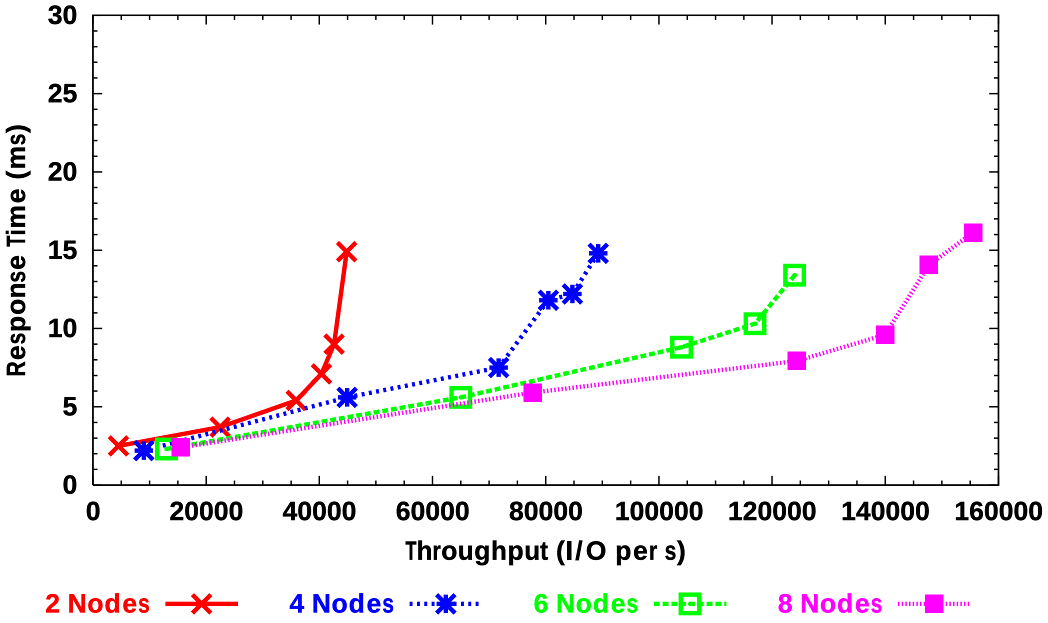

Figure 4 *SPC-1-like workload with 672 15 RPM disks configured as RAID-10. Host connectivity was via 32 fibre channels.*

Figure 4 shows the SPC-1 like performance delivered by two, four, six, or eight SVC nodes, when configured as shown in Figure 1. Figure 5 presents the database scalability results at a higher level, by pulling together the maximum throughputs (observed at a response time of 30 milliseconds or less) for each configuration. As Figure 5 shows, the tested SVC configuration is capable of delivering over 150,000 I/O's per second for the SPC-1-like workload. The reader is encouraged to compare this result against any other disk storage product currently posted on the SPC web site (`www.storageperformance.org`).

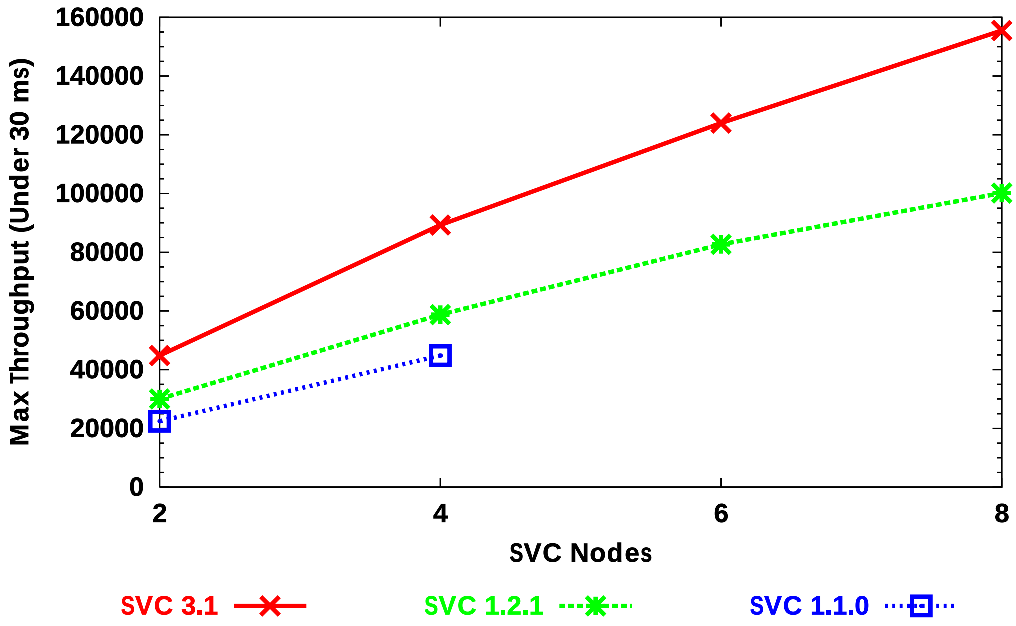

Figure 5 *SPC-1-like workload scalability.*

Figure 5 also shows that SVC 3.1 improves significantly upon the previous level of the product, for every node configuration examined. For example, the maximum SPC-1 like throughput on a single node pair improves by nearly 50 percent.

5 Sequential Scalability

Due to its ability to buffer sequential transfers in its cache, combined with its inherent scalability, the SVC architecture makes possible exceptional levels of sequential performance. Figure 6 presents the sequential throughputs achieved with two, four, six or eight SVC nodes. Using eight nodes, a read sequential throughput of approximately 4.5 gigabytes per second was achieved. The reader is encouraged to compare this data rate to any other demonstrated benchmark result for a single-image storage system.

Figure 6 demonstrates a significant improvement in SVC Release 3.1 compared with the preceding release. For example, the data rate for a single write stream has improved by more than 50 percent; the data rate for sequential reads on two nodes has improved by approximately 25 percent.

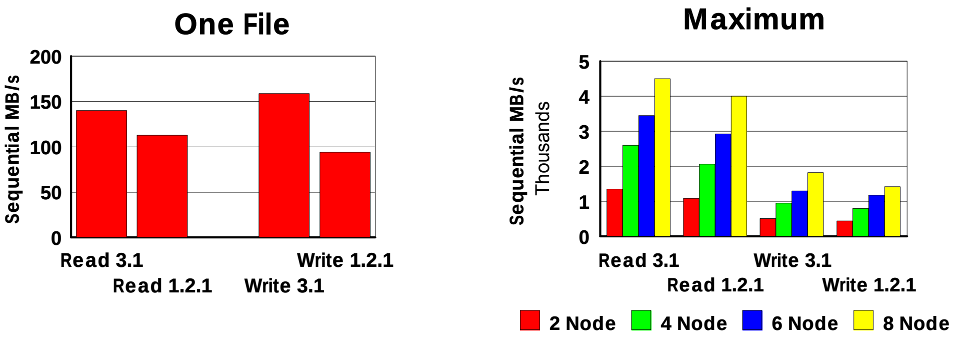

Figure 6 *Sequential throughput results. Storage layout: RAID-5. Transfer size per I/O: 512K bytes. Connectivity was via 32 fibre channel paths. Due to disk limitations, disk-level cache write mirroring was disabled for tests of write sequential throughput. This procedure was needed to demonstrate SVC sequential write throughput capability at the highest levels, since otherwise the SVC configuration would have been constrained by the underlying disk capability.*

6 Summary

The IBM TotalStorage SAN Volume Controller continues to represent an important new step in the evolution of storage technology. The SVC architecture has succeeded in bringing together, for the first time, common administration and management of heterogeneous storage; support for levels of reliability traditionally associated with high-end storage control technology; and an open-ended node architecture with no inherent limit to its scalability.

This paper has demonstrated the high levels of scalability inherent in SVC's open-ended node architecture. With SVC Release 3.1, a virtualized storage configuration can now support over 150,000 I/O's per second in the typical OLTP and/or mail server environments reflected by the SPC-1 like workload.

The Contributors

Many thanks to Howard Rankin, Phil Bryan, Tim Graham, and John Taylor, the members of the SVC team in Hursley without whose active support the tests reported in this white paper would not have been possible. Important contributions to building the needed lab configuration were also made by Bud Muenckler, Frank Gappmayer, Bill Hengen, Matt Geiser, and Beth Reygers. Finally, thanks also to all members of the SVC product development team.