

VERITAS Foundation Suite™ 2.0 for Linux

PERFORMANCE COMPARISON BRIEF - FOUNDATION SUITE, EXT3, AND REISERFS

Linux Kernel 2.4.9-e3 enterprise

Executive Summary

The SPECsfs3.0 NFS Server Benchmark running on VERITAS Foundation Suite™ 2.0 for Linux achieved a peak throughput that was faster than two other open source journaling file systems EXT3 and ReiserFS running on the standard Linux Multiple Device driver (MD). The Foundation Suite was 478% greater peak throughput than when running on EXT3 and 236% greater peak throughput than when running on ReiserFS. Foundation Suite provided scalable and significantly faster response time to client requests than either open source configuration. Enterprise environments running NFS will benefit from predictable and scalable performance running on Foundation Suite as opposed to open source EXT3 or ReiserFS.

Introduction

The purpose of this paper is to compare performance using SPECsfs NFS Server benchmark running under these three environments:

- VERITAS Foundation Suite 2.0 for Linux which includes: VERITAS Volume Manager (VxVM) an VERITAS File System (VxFS) on Advanced Server 2.1
- MD running on EXT3 file system on Red Hat Advanced Server 2.1
- MD running on ReiserFS on Red Hat Advanced Server 2.1

EXT3 is Red Hat's journaling file system. ReiserFS is the journaling file system from Namesys. The benchmarking was restricted to NFS version 3 using UDP due to limitations of Linux. To evaluate file server performance, we used the Standard Performance Evaluation Corporation (SPEC) System File Server (SFS) benchmark called SPECsfs97_R1. It is also known as SPECsfs 3.0.

Test Configurations

The SPECsfs3.0 NFS Server Benchmark tests were conducted using the following software, hardware, and setup.

HARDWARE TEST CONFIGURATION

- Dell 6450 with four 700MHz P-III Xeon CPUs and 6GB of memory
- Dell PowerVault 200S disk arrays

A Dell 6450 with four 700MHz P-III Xeon CPUs and 6GB of memory was utilized as the NFS version 3 server. The storage subsystem employed for all tests was four Dell PowerVault 200S disk arrays. Each disk array contained eight 18GB Seagate Cheetah ST318451LC 15K RPM disk drives for a total of 32 drives to be used for the tests. Each PowerVault was attached to the 6450 via one port of an Adaptec 39160 SCSI Ultra-160 host bus adapter. Two 2-port Adaptec 39160's were used in the benchmarking. The 6450 was also configured with two internal 18GB drives utilizing the built-in SCSI interface. These drives were used for various overhead requirements such as the OS and a VxVM single disk rootdg disk group. A maximum of six clients were used to generate the NFS workload, the NFS clients were Sun Microsystems Ultra5 workstations with one 400MHz UltraSparc III CPU and 128 MB of RAM. One Cisco 100/1000BaseT Network switch (Catalyst 3500XL) was employed to create the private client network. The NFS server was configured to use 128 nfsd's.

SOFTWARE TEST CONFIGURATION

The performance tests were run for RAID-0 (striped) volume configurations. The VxFS SPECsfs runs employed VxVM; the EXT3 and ReiserFS runs used MD. In each case the volume layouts were configured to be identical regardless of the volume manager being used. In the tests each volume manager was used to configure sixteen 4GB volumes, two disks per volume, with each of the two disks in a volume from a separate controller (not just a different port on the same controller).

The VxFS volumes were created using the **vxassist** command. The EXT3 and ReiserFS volumes were loaded at boot time with raidtab entries, using partitions sized to match the volumes made with VxVM. For each configuration the same sets of

disks were used. For example, the set of disks used as the first volume in the VxVM configuration was also used to create the first volume in the MD configuration.

SOFTWARE RELEASES

The following Operating System and VERITAS Software releases were used in testing:

- Red Hat Advanced Server 2.1 (kernel: 2.4.9-e3enterprise)
 - EXT3 and ReiserFS that ship with above
- VxFS 3.4 Update 2 (from Linux Foundation Suite 2.0)
- VxVM 3.2 Update 1 (from Linux Foundation Suite 2.0)

SOFTWARE CONFIGURATION DETAILS

The Foundation Suite test used the following command lines for volume and file creation. (The placeholders within the brackets, < >, were replaced with specific values for each volume.)

To create a VxVM RAID-0 volume:

```
vxassist -o ordered -g <disk group> make <volume> 4g layout=stripe < disk names >
```

Below is a typical VxFS file system creation command line used for all tests:

```
/sbin/mkfs.vxfs <volume name>
```

Below is an example of the mount command used with the Foundation Suite stack:

```
/sbin/mount.vxfs -o log <volume name> <mount point>
```

The open source tests used the following for volume and file creation.

The following /etc/raidtab entry is an example of the raidtab used to create the MD RAID-0 volumes:

```
raiddev          /dev/md1
raid-level       0
nr-raid-disks   2
chunk-size      64
nr-spare-disks  0
persistent-superblock 1
device          /dev/sdc1
raid-disk       0
device          /dev/sds1
raid-disk       1
```

Below are the command lines used for EXT3 and ReiserFS file system creation for all tests:

```
/sbin/mkfs.ext2 -j <volume name>
```

```
/sbin/mkreiser -q <volume name>
```

Below is an example of the mount command used for EXT3 and ReiserFS:

```
/bin/mount <volume name> <mount point>
```

Results and Analysis

Table 1 shows the improvement in peak throughput that the Foundation Suite obtained over MD running on EXT3 and MD running on ReiserFS. The Foundation Suite stack provided 478% greater peak throughput than the EXT3 stack, and 236% greater peak throughput than the ReiserFS stack.

Configuration	Foundation Suite stack Improvement over EXT3 & MD	Foundation Suite stack Improvement over ReiserFS & MD
RAID-0 UDPV3	478%	236%

Table 1: Increase in SPECsfs97_R1 Peak Throughput Obtained By Foundation Suite stack.

UDP v.3		
	Max Throughput (NFS Ops/sec)	Overall Response Time
Foundation Suite	4480	3.72
EXT3 & MD	818	10.66
ReiserFS & MD	1335	7.39

Table 2: UDP V3 Max Throughput and Overall Response Time.

The NFS UDPv.3 results in Table 2 above show the Foundation Suite stack outperformed the competing stack's throughput. In the case of the EXT3 stack's Overall Response Time is 187% higher than the Foundation Suite stack and the ReiserFS stack is 99% higher than Foundation Suite stack.

Detailed Results

Table 3 shows detailed results of each benchmark run for the different file systems and Volume Managers. The results show that Foundation Suite performs significantly better with higher throughput than MD running on EXT3 or ReiserFS.

Foundation Suite's peak throughput is 478% greater than EXT3 and MD, and 236% greater than ReiserFS and MD in a RAID-0 volume configuration. Despite enabling a higher load on the server, the Foundation Suite stack provided Overall Response Time (ORT) that is 187% higher than EXT3 and MD, and 99% higher in the case of the ReiserFS and MD. (The ORT provides a measurement of how the system responds, over the entire range of tested loads.)

UDP V3					
Foundation Suite		EXT3		ReiserFS	
Ops/Sec	Response Time (Msec/Op)	Ops/Sec	Response Time (Msec/Op)	Ops/Sec	Response Time (Msec/Op)
448	1.6	80		139	2.3
906	1.7	160	3.3	283	2.5
1372	1.8	240	8.4	425	15.7
1825	1.9	325	6.2	566	3.5
2263	2.4	397	9.7	711	11.4
2728	3.5	506	12.9	839	8.4
3193	4.7	574	14.4	996	4.8
3656	6.4	652	16.5	1139	9.7
4097	8.5	566	25.0	1289	13.0
4480	11.5	818	16.4	1335	12.7

Table 3: SPECsfs97_R1 Statistics for RAID-0 Volume Configuration. In the throughput (Ops/sec) columns, Foundation Suite achieved 478% greater peak throughput than EXT3 and 236% greater peak throughput than ReiserFS for UDP V3 results.

Figure 1 illustrates the throughput and response time limits for the file systems in the RAID-0 volume configuration.

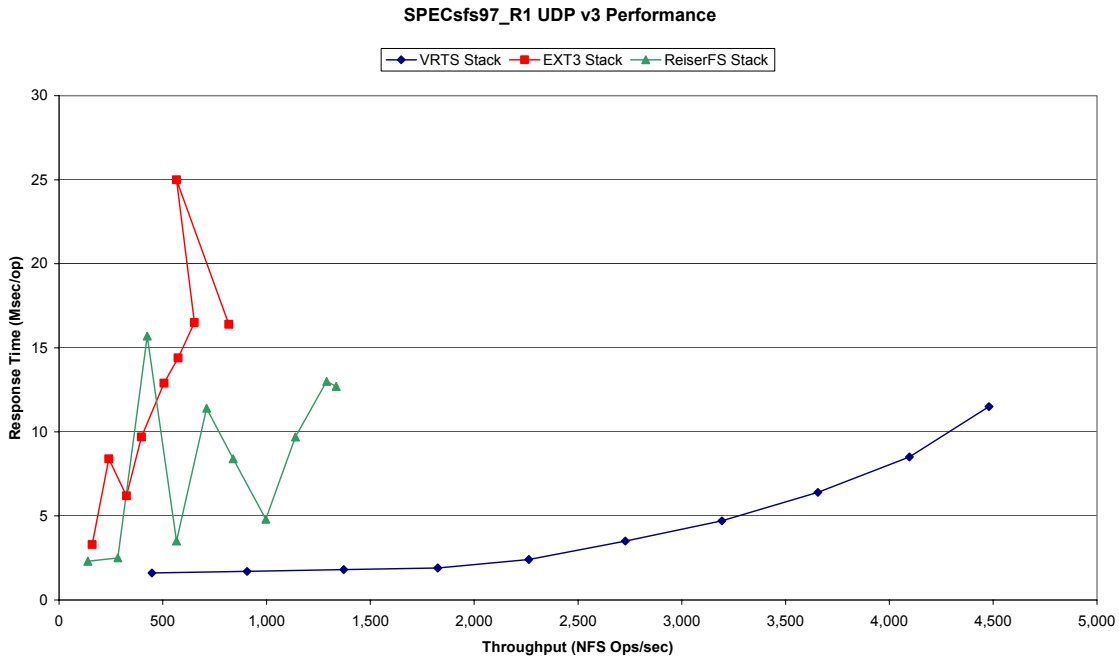


Figure 1: Throughput and Response times for Foundation Suite, EXT3, and ReiserFS stacks. “Goodness” in this graph is down and to the right. In the throughput (Ops/sec) columns, Foundation Suite achieved 478% greater peak throughput than EXT3 and 236% greater peak throughput than ReiserFS for UDP V3 results.

Figure 2 illustrates the throughput and CPU Utilization curves for the file systems in the RAID-0 volume configuration. Figure 2 shows that both the EXT3 and ReiserFS stacks require widely varying CPU utilization during their benchmarking runs, while the Foundation Suite's CPU utilization is quite even.

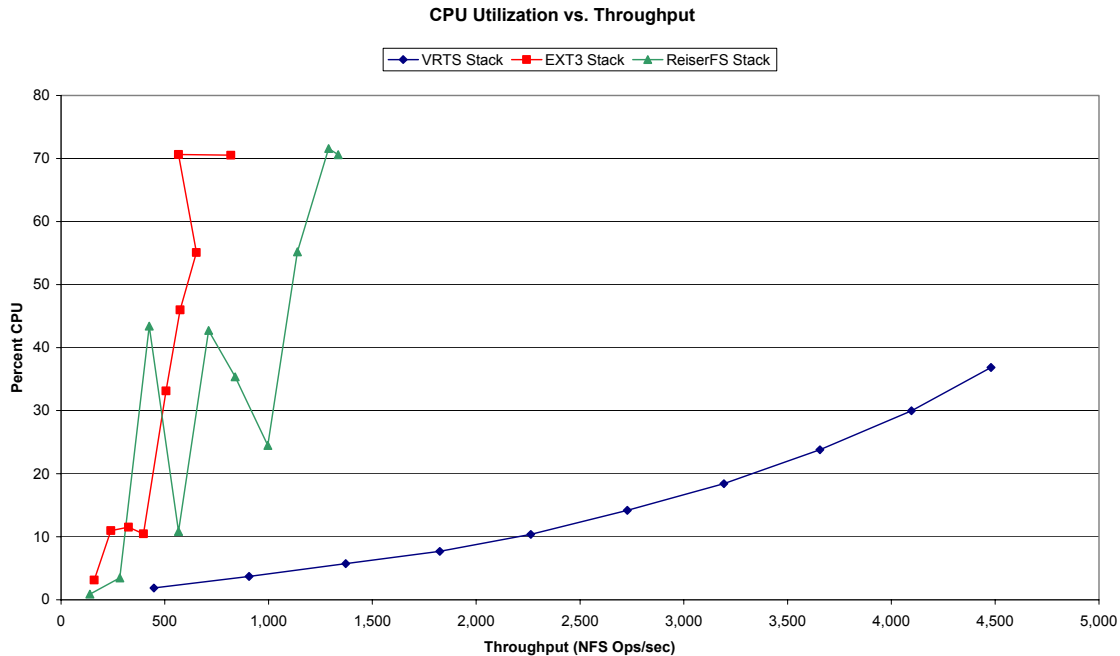


Figure 2: Throughput and Percent CPU for Foundation Suite, EXT3, and ReiserFS stacks. “Goodness” in this graph is down and to the right. Notice the irregular CPU utilization of the EXT3 and ReiserFS stacks over the period of their benchmark runs.

Disk I/O utilization is very difficult to measure on a Linux system due to the lack of a reliable *iostat*. It was possible to use *vxstat* to get disk I/O utilization for the benchmark runs that used VxVM. From the limited disk I/O utilization measurements we were able to conclude that the Foundation Suite stack was not bottlenecked on disk I/O.

A sidebar to the benchmark results for this paper, and only partially visible through the numbers, is the great difficulty that was encountered in benchmarking EXT3 and ReiserFS. As can be seen from the results, the benchmark of EXT3 and ReiserFS are very uneven. This is not the normal results from a SPECsfs benchmark run and a great many runs and testing on the configuration was necessary before it became apparent that this unevenness was endemic to EXT3 and ReiserFS when being benchmarked with SPECsfs. The second problem in benchmarking EXT3 and ReiserFS was the number of SPECsfs benchmark runs that failed before the necessary data points were achieved, roughly half the benchmarking runs attempted on EXT3 and ReiserFS failed before the run finished. This contrasts with the Foundation Suite stack where there were no failed attempts in over 30 benchmark runs.

Of further note is that during the early stages of this project Red Hat 7.2 (kernel 2.4.9-13) was used as a test OS. While this OS was in use, EXT3 using VxVM was benchmarked and a peak throughput of about 2,600 Ops/sec was achieved. When the same hardware was used with Advanced Server 2.1, EXT3 using VxVM only achieved about 800 Ops/sec. Red Hat had no readily available answer to our questions of why this should be.

Conclusion

The benchmarks discussed in this paper show the Foundation Suite achieves much greater throughput than both EXT3 and ReiserFS stacks. Using RAID-0 and utilizing the UDP protocol, a peak throughput of about 4,500 NFS Ops/sec is achieved with the Foundation Suite. The EXT3 stack produces about 800 NFS Ops/sec and ReiserFS and MD produces about 1,300 NFS Ops/sec. These results show that Foundation Suite has significantly higher throughput than ReiserFS and overwhelmingly better than EXT3.

Enterprise customers running NFS for file services benefit from faster and more predictable performance when using VERITAS Foundation Suite than EXT3 or ReiserFS. With Foundation Suite, CPU utilization is quite even while CPU utilization varies widely with increasing workload when using EXT3 or ReiserFS.

Appendix A

Table 4 shows the workload distribution of SPECsfs 3.0 using NFS version 3.

NFS Op	Distribution
getattr	11%
setattr	1%
lookup	27%
readlink	7%
read	18%
write	9%
create	1%
remove	1%
readdir	2%
fsstat	1%
access	7%
commit	5%
readdirplus	9%

Table 4: SPECsfs NFS v3 Workload Distribution

The following is a combined SPECsfs full disclosure like report for the three stacks benchmarked in this report.

SPECsfs97_R1.v3 Result

Foundation Suite RAID-0

SPECsfs97_R1.v3 = 4480 Ops/Sec (Overall Response Time = 3.72)

EXT3/MD RAID-0

SPECsfs97_R1.v3 = 818 Ops/Sec (Overall Response Time = 10.66)

ReiserFS/MD RAID-0

SPECsfs97_R1.v3 = 1335 Ops/Sec (Overall Response Time = 7.39)

RAID-0

Foundation Suite		EXT3/MD		ReiserFS/MD	
Ops/Sec	Response Time (Msec/Op)	Ops/Sec	Response Time (Msec/Op)	Ops/Sec	Response Time (Msec/Op)
448	1.6	80		139	2.3
906	1.7	160	3.3	283	2.5
1372	1.8	240	8.4	425	15.7
1825	1.9	325	6.2	566	3.5
2263	2.4	397	9.7	711	11.4
2728	3.5	506	12.9	839	8.4
3193	4.7	574	14.4	996	4.8
3656	6.4	652	16.5	1139	9.7
4097	8.5	566	25.0	1289	13.0
4480	11.5	818	16.4	1335	12.7

CPU, Memory and Power

Model Name	Dell 6450
Processor	700 MHz P-III Xeon
# of Processors	4
Primary Cache	16KBI+16KBD on chip
Secondary Cache	512KB(I+D) on chip
Other Cache	N/A
UPS	N/A
Other Hardware	N/A
Memory Size	6 GB
NVRAM Size	N/A
NVRAM Type	N/A
NVRAM Description	N/A

Server Software

OS Name and Version	Red Hat Advanced Server 2.1 (kernel 2.4.9-e3enterprise)
Other Software	VxFS 3.4 Update 2, VxVM 3.2 Update 1
File System	VxFS, EXT3, ReiserFS
NFS version	3

Server Tuning

Buffer Cache Size	default
# NFS Processes	128
Fileset Size	43.5 GB (Foundation Suite) 7.7 GB (EXT3/MD) 13.5 GB (ReiserFS/MD)

Network Subsystem

Network Type	Gigabit Ethernet
Network Controller Desc.	Intel PPro 1000
Number Networks	1
Number Network Controllers	1
Protocol Type	UDP
Switch Type	Cisco 3500XL Switch
Bridge Type	N/A
Hub Type	N/A
Other Network Hardware	N/A

Disk Subsystem and Filesystem

Number Disk Controllers	3
Number of Disks	34
Number of Filesystems	16 (F1-F16)
File System Creation Ops	default
File System Config	default
Disk Controller	Integrated SCSI Controller
# of Controller Type	1
Number of Disks	2
Disk Type	Seagate Cheetah 18GB 10K RPM & Quantum Atlas 18GB 10K RPM
File Systems on Disks	OS, root diskgroup
Special Config Notes	
Disk Controller	Adaptec 39160
# of Controller Type	2
Number of Disks	32

Disk Type Seagate Cheetah 18GB 15K RPM
 File Systems on Disks F1-F16
 Special Config Notes

Load Generator (LG) Configuration

Number of Load Generators 2 (EXT3, ReiserFS), 6 (Foundation Suite)
 Number of Processes per LG 8
 Biod Max Read Setting 2
 Biod Max Write Setting 2

LG Type LG1
 LG Model Sun Microsystems Ultra Enterprise 5
 Number and Type Processors 1 400 MHz UltraSPARC
 Memory Size 128 MB
 Operating System Solaris 2.8
 Compiler SPEC supplied precompiled binaries
 Compiler Options N/A
 Network Type On-board 100baseT

Testbed Configuration

LG #	LG Type	Network	Target File System	Notes
1-6	LG1	N1	16 (F1-F16)	Foundation Suite
1-2	LG1	N1	16 (F1-F16)	EXT3 & ReiserFS

Notes and Tuning

None.