Tech OnTap

Insights for Simplifying Data Management

NetApp

Tech OnTap Home

January 2006

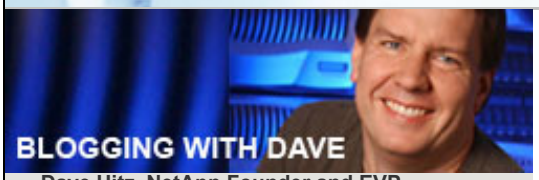
HIGHLIGHTS

- **NetApp Vision of the Grid**
  (Tech OnTap Exclusive Podcast)
- **SQL Server 2005 Benchmark**
- **NFSv4 Benefits and Misconceptions**
- **Free Monitoring Tool: ToasterView**
- **Special Access:**
  Online Launch Event

Six Predictions for 2006

NetApp visionaries predict this year's trends in enterprise security, where IP SANs will appear next, and the expanding role of backups.

More



Dave Hitz, NetApp Founder and EVP

"EMC's strategy is challenging, but I think it may be the best one available to them. Here's why..."

- **Dave's Blog**

DRILL DOWN

- **Most Interesting Technology: iSCSI**
  Learn about iSCSI from the world leader in iSCSI storage and see which technologies your peers are interested in.
- **Exclusive Podcast:**
  NetApp Vision and Roadmap for Grid
  No iPod? Listen on your PC.

SYS ADMIN CORNER

Blake Golliher, Yahoo!

- **Ask the Sys Admin**
  Yahoo!'s Blake Golliher on disk imbalances and volume SnapMirror vs. qtree SnapMirrors
- **ToasterView**
  A free PERL script enabling graphical disk usage monitoring at the volume level

TIPS FROM THE TRENCHES

Evaluating VTL Solutions

Offer Inside

Dianne McAdam, Sr. Analyst and Partner, Data Mobility Group

Analyst advice on choosing the best virtual tape technology for your environment, plus a free 10-page report outlining in-depth questions for prospective vendors.

More

Behind the Scenes: Setting a New TPC-C Record

NetApp, IBM, and Microsoft

Meet the team that benchmarked the fastest TPC-C performance number ever for a 16-way, Xeon-based server:

492,307 tpmC

More

ENGINEERING TALK

Five Things to Know About NFSv4

Mike Eisler, Director of Technology, NetApp

NSFv4 benefits, drawbacks, and misconceptions, plus a Trek-ified history of NFS and the opportunity to blog with the coauthor of O'Reilly's *Managing NFS and NIS*.

More

Six Predictions for 2006

To learn which 2006 events will most impact IT, we turned to a range of NetApp and Decru® visionaries. Here are six trends they think bear watching this year:

- **Enterprises will start taking a militaristic approach to running security.**
  The barrier to full security in most organizations has not been the lack of appropriate technologies but implementation of standard processes. In 2006, enterprises will increasingly recognize the need to secure data in flight and data at rest on both disk and tape. Expect to see a majority of enterprises start to incorporate need-to-know access controls, compartmentalization, and role separation on critical systems. Encryption of tape backup data before sending tapes off site will become the norm.

- **Continued consolidation, simplification will speed IT's move to SOAs.**
  Simplification and on-going consolidation of servers and storage will continue to dominate buying decisions for the rest of the decade. By enabling growth while reducing cost, these projects lay the groundwork for the next major enterprise IT transformation: a renewed focus on Services Oriented Architectures (SOAs).

- **The death of the database feature wars will shift focus to integration.**
  SOA-centric frameworks such as those found in Oracle® Fusion, IBM Websphere, and SAP ESA-based products will facilitate widespread application integration efforts. Over the next 12 months, enterprise database purchasing decisions will shift from core functionality to integration at the application level. Storage vendors with strong integration at the data and application framework level will have a growing value proposition.

- **IP SAN (iSCSI) will become a ubiquitous, multi-OS solution.**
  2005 saw iSCSI become mainstream in Windows® server environments. Second-generation solutions are expanding this "sweet spot" by adding high-availability support, SAN boot support for dense server environments, and performance enhancements. In 2006, iSCSI will expand support to include departmental Linux® and UNIX® environments, particularly for blade servers and small-to-midrange hosts. Expect more stories about iSCSI replacing first-generation FC SANs.

- **Disk-to-disk backup will mature as a true complement to tape libraries.**
  Although disk-to-disk backup will not fully replace tape libraries, expect to see a reduction in reliance on tape in favor of disk-based solutions. The popularity of Virtual Tape Libraries (VTLs) will continue to grow as customers seek seamless ways to transition from tape to disk. In 2006, NetApp forecasts mass adoption of VTL solutions. The year will also bring a strong trend toward the use of replication technologies to automate remote office backups.

- **Backups will be increasingly viewed as strategic reservoirs of information.**
  Expect significant advances in content indexing and the ability to manage data in an application/project-centric approach (as opposed to storage/resource-centric approaches). In 2006, this will largely be leveraged through reporting systems, with storage managers manually moving data "across tiers" for more effective utilization.

  As a result of the increased adoption of disk-to-disk backup and the availability of powerful search and indexing solutions, organizations will expand their use of backup copies beyond basic data recovery. Backup data will be used for business needs such as data mining and legal discovery while online backup images will accelerate and simplify application development, testing, and QA functions.

Additional high-profile trends in 2006 will include:

- Adoption of scale-out storage architectures to incrementally and non-disruptively grow capacity and performance and to better balance load across storage resources
- Storage networking enhancements (i.e. intelligent switches)
- Geographically distributed data sharing and collaboration (i.e. WAFS)

Comment on this article.

© 2006 Network Appliance, Inc.

RELATED INFORMATION

- Is NetApp Storage Simpler?
- Choosing between iSCSI and FCP
- Build a Better Backup Strategy
- IP SAN (iSCSI) Storage Center
- Evaluating VTL Solutions
- Uncompromised Security

IP SAN (iSCSI) Developments in 2006

IP SAN will be become the de facto choice for Windows environments.
In addition, second-generation features such as Multi-Connection Sessions and ErrorRecoverLevel>0 will enable iSCSI-based IP SANs to make use of the built-in parallelism and fault tolerance of Ethernet networking and TCP/IP for higher performance and availability with reduced complexity. High performance is simply achieved with sessions that span multiple 1GB paths, while high availability can be accomplished without the expense and complexity of special HBAs and proprietary multipathing software.

IP SAN will gain traction in departmental Linux and UNIX environments.
2006 will see the availability of native iSCSI software initiators from all major UNIX and Linux vendors. In addition, native storage stack support for multipathing and other high-availability options will propel Linux and UNIX IP SAN solutions into the mainstream.

iSCSI boot-from-SAN will revolutionize server provisioning.
Removing disk from low-end servers eliminates a key point of failure and significantly changes the economics of server infrastructures. Once provisioning becomes this simple, it will be possible to almost instantaneously provide more CPU and to enable true "on demand" environments.

DIANNE MCADAM
Senior Analyst and Partner, Data Mobility Group

Leveraging more than three decades of experience working for industry-leading storage and systems vendors, Dianne McAdam provides practical insights on replication technology, business continuance, and networked storage. Dianne directs research and advisory services for industry research firm Data Mobility Group. In addition to reporting on trends in the storage industry, Dianne collaborates with vendors to define their product roadmaps and strategies while helping end users to evaluate alternate technologies.

Evaluating Virtual Tape Systems

Q&A with Dianne McAdam

Q: You recently completed a report on virtual tape technologies. What do you see driving interest in using virtual tape as a backup target?

Dianne: Virtual tape technology essentially involves disk-based storage that interacts with the rest of the IT environment as tape. VTL, or virtual tape library, is the generally accepted term, although some people call it virtual tape, disk-emulating-tape, or disk as tape.

Although a VTL solution offers the performance and rapid recoverability of disk-based backup, to the rest of the backup environment it is indistinguishable from a tape library. This means VTLs can be plugged into existing backup environments without requiring architectural changes.

In contrast, using raw or native disk as a target means that the backup application isn't talking to a tape drive anymore. This approach usually requires changing scripts and retraining. Many backup applications put their own file system on the disk systems so they can write backups to them. Writing backups to a file system, deleting those backups, and then writing new backups cause fragmentation and performance degradation. As a result, the IT staff must monitor the system to avoid outstripping capacity and initiate defragmentation if necessary.

A well-designed VTL solution will provide the performance and reliability of disk-based backup without the added administrative costs associated with backing up to a native disk device.

Q: Your report outlines nine factors that should be considered as people evaluate alternate VTL solutions. What are some of the most important things people should be thinking about?

Dianne: On the surface, a lot of the VTL solutions look alike. Many of them will do the job, but they may not do the job well. Some solutions are more difficult to implement and maintain, while others cannot deliver on their performance promises. A solution that meets today's backup requirements may not scale to support future needs.

When you look under the covers, there are several things to consider.

The first consideration is how efficiently the system writes to the disk drives. You want a solution capable of efficiently writing across all the disks by automating the load-balancing. Otherwise the system will write to one particular spot and create hotspots. In addition to resulting in a higher-maintenance solution, this means that a VTL that runs perfectly today might run in a more degraded mode in the future.

Another factor to consider is compression. Compression can save capacity by shrinking the data as it comes in, but software compression has a terrible performance hit. It can cause up to 50% performance degradation for all of your backup jobs, which is clearly not the result many are seeking for a VTL solution.

A third major factor involves how effectively the solution deals with tape drives. Some standalone VTL solutions don't talk to tape drives at all. In a situation where you need to create a physical tape cartridge to send off-site or to lock up, they will require the backup application to write a second backup directly to tape. This two-step process places an additional burden on the backup server.

Other VTL solutions offer you the option to write more directly to the physical tape drive, in essence sending the communication to the backup application to go clone the tape. Because the VTL uses internal processing power — not that of the backup server — to actually copy to the physical tape, it avoids overburdening the backup server.

RELATED INFORMATION

- A Guide to Evaluating Virtual Tape *(Tech OnTap exclusive!)*
- Build a Better Backup Strategy
- NetApp Open Storage Networking
- NetApp Partner Solutions *(customers only, password required)*
- Data Mobility Group

Seven Questions to Ask a Potential VTL Vendor

1. Can performance and capacity scale to keep pace with my data growth?
2. What is the penalty for using compression?
3. Can the solution deliver maximum performance without manual configuration and tuning?
4. Are physical tapes written in the identical format as the backup application?
5. Is the VTL fully compatible with the leading backup software and tape devices?
6. Can the VTL create tapes directly for optimum performance?
7. Can the VTL fully utilize the speed and media savings provided by tape drive hardware compression?

Read the full Guide to Evaluating Virtual Tape Systems.

True or False: Tape Is Slow

FALSE! Many people perceive tape as slow, but this is incorrect. Here's why:

If an older tape drive is not receiving a constant stream of data, it has to stop and wait for data before it can continue writing on the tape cartridge. Because tape is a flexible medium, an abrupt stop would put excessive tension on the tape and cause tears.

Instead, the tape cartridge must gradually slow down before stopping. No data can be written while the tape cartridge slows, which leaves a blank spot.

Before writing data to the tape again, the cartridge first rewinds to avoid wasting space.

These topics are covered in depth in A Guide to Evaluating Virtual Tape Systems.

Q: Do you have any other advice for people trying to decide between different VTL options?

Dianne: I would boil it down to three key issues:

1. Are you buying from a trusted vendor that will be there to support you if you have any problems?
2. Does the vendor provide a wide range of backup solutions? If you're running a lot of different applications, each one probably has its own requirements.
3. Does the vendor understand that this technology is going to continue to evolve, and does the vendor have a vision and a roadmap?

This back-hitch operation is what degrades performance. Disk avoids the back-hitch process, making it much more efficient.

Tech OnTap Exclusive: Advance access to the new report from Data Mobility Group, **A Guide for Evaluating Virtual Tape Systems**.

© 2006 Network Appliance, Inc.

The TPC-C Benchmark Performance Team

Over 40 people from Network Appliance, IBM, and Microsoft worked to achieve a record-setting tpmC throughput result of 492,307. The project's seven core team members were Steve Daniel (Director of Database Platforms and Performance Technology), Ray Engler (Lead Performance Engineer, IBM), Sanjay Gulabani (Senior Database Performance Engineer, NetApp), Jeff Kimmel (Technical Director, NetApp), Mark Kapoor (Performance Analysis, IBM), Dan Morgan (Senior Manager of Database Performance, NetApp), and Ricky Stout (Lab Manager, NetApp).

Pictured here clockwise from left: Ray Engler, Ricky Stout, Dan Morgan, Chris Lemmons, Steve Daniel, Lee Dorrier, Keith Griffin, Mark Kapoor, Sanjay Gulabani, and Yogesh Manocha.

Behind the Scenes: Achieving a New TPC-C Performance Record

In November of 2005, IBM and NetApp conducted the TPC Benchmark C for Microsoft SQL Server 2005 on the IBM eServer xSeries 460 configured as a client/server system with NetApp FAS3050 storage systems using the Fibre Channel SAN protocol.

This effort resulted in the fastest TPC-C performance number ever for a 16-way, Xeon-based server: 492,307 tpmC.

We proudly invite you to read the 422-page full disclosure report, but will understand if you prefer the 3-page executive summary.

Hardware	Software	Total System Cost	tpmC	$/tpmC	Total Solution Availability Date
IBM eServer xSeries 460	Microsoft SQL Server 2005 Enterprise x64 Edition (SP1) Microsoft Windows Server 2003 Datacenter x64 Edition	$3,138,060 USD	492,307	$6.37 USD	May 20, 2006

Here's why it matters: the TPC-C benchmark is designed to measure maximum sustained system performance and response times for high-transaction-rate online applications.

By demonstrating a 492,307 tpmC benchmark in a 16-CPU Xeon server environment, NetApp has proven its ability to meet the performance requirements of sustained high-transaction database environments.

Net-net: Customers can be confident in their ability to scale out NetApp Fibre Channel solutions for even their most demanding applications.

Pretty impressive, eh?

What's even more impressive is that although most projects of this scale take six months or more, the entire benchmark process was completed in just under four months (114 days).

While the Microsoft choice of NetApp storage to demonstrate SQL Server 2005 performance clearly provides a technological edge, the reality is that an amazing team worked on the project.

This project involved a joint collaboration among the NetApp team at Research Triangle Park (RTP), North Carolina; IBM teams in RTP and Kirkland, Washington; and the Microsoft SQL Server team in Seattle, Washington.

Quick facts:

- Over 40 people were involved in the project, including 7 core team members: Steve Daniel, Ray Engler, Sanjay Gulabani, Jeff Kimmel, Mark Kapoor, Dan Morgan, and Ricky Stout.

RELATED INFORMATION

- The TPC-C Benchmark Result
- SQL Server 2005: Decision Support Scalability Improvements
- IP SAN (iSCSI) Storage Center: Focus on SQL Server
- Best Practices for SQL Server on NetApp Storage
- Double Parity RAID for Enhanced Data Protection with RAID DP
- Ask the Sys Admin Column

TPC-C Benchmark Test Overview

This test was designed to reflect real-world performance results. The NetApp systems used to run the benchmark had very little special tuning. The only major difference between the test environment and a real-world production environment was the lack of a disaster recovery application.

Equipment and configuration:

- Four IBM x460 servers with 256GB of total RAM in a NUMA configuration
- 16 NetApp FAS3050s with 84 disks (database data files)
- 1 NetApp FAS3050 (1.5TB of SQL Server transaction logs)
- A total of nearly 1400 disks
- Data ONTAP® 7G grid-based storage
- NetApp RAID-DP™ and Snapshot™ technology

TPC-C Benchmark Result

NetApp Best Practice Tip for SQL Server Backups

When in doubt, NetApp generally recommends backing up each database every hour and transaction logs every 30 minutes during production time.

Learn more SQL Server Best Practices.

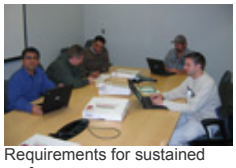
- Members of the core team worked every weekend for four months to achieve the benchmark result.
- This was a 24x7 operation; trips to the office at 2 or 3 a.m. were not unusual.
- The full team participated in 28 conference calls plus hundreds of additional calls on a daily basis.
- Over 5,000 e-mails messages were exchanged with NetApp project coordinator Dan Morgan.

From left to right: Ricky Stout, Keith Griffin, Lee Dorrier, Steve Daniel, Yogesh Manocha, Sanjay Gulabani, Mark Kapoor, Chris Lemmons, Ray Engler, and Dan Morgan. Additional team members not shown: Jeff Kimmel and Phil Larson.

Comments from the team:

"RAID-DP made it possible to run the benchmark without having to restore backups on failed disks. For a project of this magnitude failed disks are inevitable, but all we had to do was to replace them with hot spares and continue."
– Sanjay Gulabani, Senior Database Performance Engineer, NetApp

"Database build time was about 50 hours and we encountered a total of 12 failed disks. If we had to rebuild for every failed disk, the operational build time would have topped 600 hours (25 days)!"
– Dan Morgan, Senior Manager, Database Performance, NetApp


Requirements for sustained performance:

3 boxes x 12 ea. =
36 donuts/5 engineers =
7.2 donuts per engineer

"Various backup restores are necessary operational requirements for running an audited benchmark. We did almost 30 full backup restores, which took only a few minutes using NetApp SnapRestore® and Snapshot technology. A full backup restore of this size could easily take 3.5 hours, so this saved about 105 hours (nine 12-hour days)."
– Sanjay Gulabani, Senior Database Performance Engineer, NetApp

Congratulations to the team — Job well done!!

Send a comment to the team.

Can I go home now?

© 2006 Network Appliance, Inc.

Mike Eisler
Technical Director, Engineering, Network Appliance

Mike Eisler works for Network Appliance on all things related to NFS. Mike is particularly interested in security and global namespaces for NAS; before joining NetApp in 1992 he led the industry effort to standardize NFS security using Kerberos V5. An active participant in the Internet Engineering Task Force (IETF) NFSv4 working group, Mike coauthored the NFSv4 specification RFC3530 and was the primary author of RFC2203, which enhances NFS security. He also coauthored the O'Reilly book Managing NFS and NIS.

Mike posts answers to NFS questions and issues at www.nfsworld.blogspot.com.

The Top Five Things You Need to Know about NFSv4

Sun™ Network File System (NFS) has been the standard distributed file system for UNIX® systems for several decades.

Today, the original NFSv2 protocol appears to be dying, while the wildly successful NFSv3 is being pushed beyond its limits. Although the first client for NFSv4 became available in 2002, UNIX-based clients and servers have been available only since 2004, and the protocol is just starting to gain mainstream acceptance.

At the Large Installation System Administration Conference (LISA '05) in December, Mike outlined the five things that everyone who works with storage should know about NFSv4:

1. How NFSv4 Compares to NFSv3	NFSv3	NFSv4
	A collection of protocols (file, mount, lock, status)	One protocol to a single port (2049)
	Stateless	Lease-based state
	UNIX-centric, but seen in Windows too	Supports UNIX and Windows file semantics
	Deployed with weak authentication	Mandates strong authentication
	32 bit numeric uids/gids	String-based identities
	Ad-hoc caching	Real caching handshake
	UNIX permissions	Windows-like access
	Works over UDP, TCP	Bans UDP
	Needs a-priori agreement on character sets	Uses a universal character set for file names

2. The Benefits of NFSv4	
	Mandates strong security • Every NFSv4 implementation has Kerberos V5 • You can use weak authentication if you want (but why would you?)
	Finer grained access control • Go beyond UNIX owner, group, mode
	Read-only, read-mostly, or single writer workloads can benefit from formal caching extensions
	Multi-protocol (NFS, CIFS) access experience is cleaner
	Byte range locking protocol is much more robust • Recovery algorithms are simpler, hence more reliable

RELATED INFORMATION

- Eisler's NFS Blog
- NFSv4 LISA '05 Presentation
- The Future of NFS Presentation
- Managing NFS and NIS, 2nd Edition
- NFSv4 Information and Resources
- NetApp NOW Customer Site *(password required)*

A Trek-ified History of NFS

1984:
NFS was Kirk's Federation
New, brash, annoying, and unilateral

1992:
NFS was Picard's Federation
Wiser, conciliatory, and multilateral DEC was the Klingon Empire, forcing the change in attitude

2003+:
Divergence from Trek Canon
NFS became the Borg by assimilating other file access cultures to become NFSv4

Read Mike's Presentation about The Future of NFS

NFSv4 Features by Implementation

	AIX	BSD	Hummingbird	Linux	NetApp	Solaris
Access Control Lists	YES			YES	YES	YES
Automated Lock Recovery	YES	YES	YES	YES	YES	YES
Eliminating Adjunct Protocols	YES	YES	YES	YES	YES	YES
Aggressive Caching				YES	YES	YES
Kerberos V5	YES	YES	YES	YES	YES	YES
File Streams					YES	YES
ACLs	YES	YES	YES	YES	YES	YES
Global Namespace						

Read Mike's Presentation about The Future of NFS

3. The Drawbacks of NFSv4	▸ Fewer implementations than NFSv3 • But unlike with NFSv3, Linux is not the laggard ▸ Not all features uniformly implemented right now

4. Common Misconceptions About NFSv4	▸ NFSv4 is a new protocol, so I can use more than 16 supplemental gids • No, the 16 gid limit is a property of the weak authentication flavor of the remote procedure call • Use Kerberos V5, and you can go beyond 16 gids ▸ I need NFSv4 in order to use Kerberos V5 • No, Kerberos V5 works on NFSv[23] and has for years on EMC, Hummingbird, NetApp, Solaris

5. Which Vendors Support NFSv4	▸ IBM (AIX 5.3) ▸ EMC ▸ Hummingbird ▸ Network Appliance (best is 7.0.x) ▸ FreeBSD 5.3 ▸ Linux 2.6 (Fedora Core) ▸ Solaris 10 ▸ More to come ...

Want to know more about NFSv4? Check out Mike Eisler's blog or his recent presentations:

- NFSv4, Large Installation System Administration Conference (LISA), Dec 2005.
- Future of NFS, SNIA Developer Solutions Conference, Aug 2005.

Have a question about NFS?

© 2006 Network Appliance, Inc.

www.cuddletech.com

Ben Rockwood
Systems Administrator, Homestead.com

A self-professed Sun™ zealot, Ben became an advocate of NetApp NAS and iSCSI storage after using heterogeneous NetApp NAS storage to power Homestead.com. In his words ...

"NetApp can serve NFS faster than the people who wrote it. NetApp also nailed iSCSI out of the box. The great thing about Network Appliance™ systems is that they are exactly that, appliances. They just work! And when problems do occur, a NetApp representative contacts me before I know about the problem myself. Of all the storage systems in my data center, the 940s set the standard that I hold other vendors to."

January's Tool of the Month: ToasterView

This month Tech OnTap begins showcasing free tools that just might make your life a little easier. Recommend a tool and get a free NetApp cycling jersey.

Author: Ben Rockwood, systems administrator for Homestead.com, author of *The Sysadmin's Guide to Oracle*, and creator of www.cuddletech.com

What it is: ToasterView's primary focus is to graphically monitor disk usage at the volume level. This simple PERL script uses the NetApp MIB and Net-SNMP to collect basic data about a NetApp storage appliance (or NetCache®) and display it in the simplest way possible. No frills, no chills, no nifty trick or tactics — this is a dead simple little stupid tool that gives you the status of your system as quickly as possible in a page that can be skimmed very quickly.

An RRDtool extension enables trends analysis by creating a table (or graph) displaying the amount of volume change in the last hour, six hours, one day, one week, and one month. ToasterView was inspired by tools such as phpSysInfo.

How it works: The app takes a single argument, the hostname (or IP) of a NetApp system that you wish to generate a report for. An HTML page is output to STDOUT. The intent is to be run from cron in the manner "toasterview my_filer > /opt/htdocs/toasterview/my_filer.html."

Why I like it: It is extremely extensible. The code is fairly loose knit, so modification is pretty simple. I'm not looking to be *l337* or prove my PERL skillz, I just want something easy to hack on to get what I want. Feel free to poke your fingers into it.

Caveats: Support for qtrees and quotas does not currently exist, but only because I don't personally use them. I have a patch for qtree support but do not have a way to test it. Visit the Web site for more information.

- **ToasterView download and full release info**

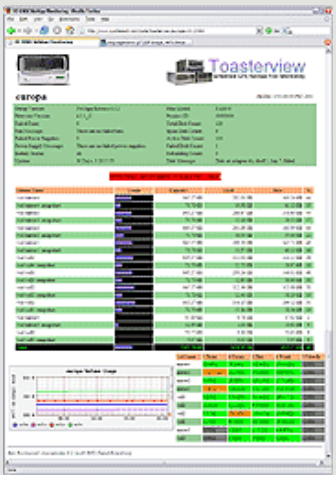
Want to recommend a tool? If we feature it, you get a limited edition NetApp cycling jersey.

- **Nominate a tool**

LINKS

- Cuddletech
- NetApp Technical Report Library
- Ask the Sys Admin Column
- Solving Five Exchange Issues
- Five Things About NFSv4
- NetApp NOW Customer Site *(password required)*

Sample Output:



© 2006 Network Appliance, Inc.

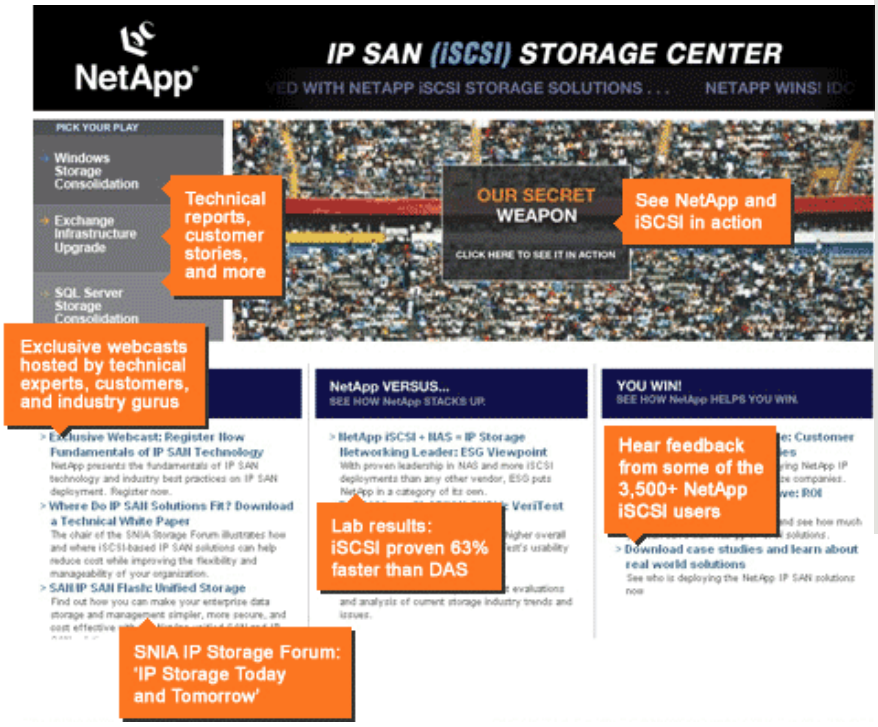
Who Cares about iSCSI?

Clearly, **you're** at least a little curious or you wouldn't have clicked this link.

You're not alone. Half the people who have joined the Tech OnTap program think that iSCSI is something worth learning more about.

Since NetApp happens to be the world leader in iSCSI storage and has over 3,500 deployments in production today, you've come to the right place. To learn about iSCSI from users, analysts, and other experts, check out the iSCSI Webcast series:

Looking for instant gratification on these or other topics related to iSCSI? Peruse the NetApp IP SAN (iSCSI) Storage Center for all things iSCSI.

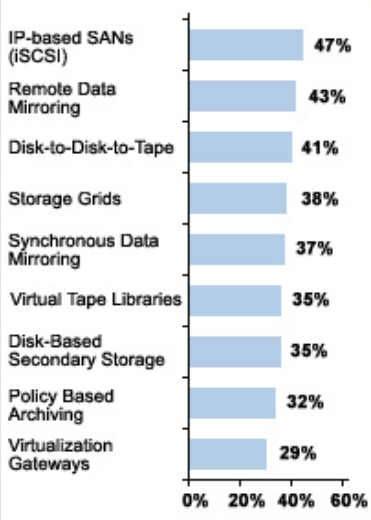


RELATED INFORMATION

- IP SAN (iSCSI) Storage Center
- Choosing between iSCSI and FCP
- IP Storage: State-of-the-Market
- SNIA IP Storage Forum: IP Storage Today and Tomorrow
- iSCSI Webcast Series
- iSCSI Customer Stories

Tech OnTap Members Vote iSCSI "Most Interesting Technology"

When asked, "Which of these technologies interest you most?" Tech OnTap members chose ...



Help us keep Tech OnTap relevant ... what topics interest you?

© 2006 Network Appliance, Inc.

BLAKE GOLLIHER
Storage Administrator and Filer Jedi, Yahoo!

As a sys admin at Yahoo!, Blake Golliher works with NetApp storage almost every single day. After searching unsuccessfully for a book on managing NetApp systems, Blake decided to leverage his more than five years of experience on the subject and write one himself. As part of his background research, Blake has offered to share his tips, tricks, and advice for managing NetApp storage with Tech OnTap members.

We're thrilled to feature Blake in Tech OnTap because he's a super smart guy, but our legal team requires this caveat. Blake has extensive tribal knowledge based on years of experience and working directly with engineers who develop NetApp products. His advice may include workarounds and suggestions not documented in NetApp best practices materials and therefore not supported by NetApp.

Ask the Sys Admin

By Blake Golliher

Last month several people asked about performance analysis. There's no "one size fits all" answer—it depends on your environment, applications, workload, and other variables. I previously wrote about the SIO tool, so this month I decided to talk about a key cause of performance issues: disk imbalances. (Special thanks to Darrin Chapman in NetApp tech marketing and Darren Sawyer in the NetApp performance group for their advice on this month's topics.)

January topics:

- Identifying and solving disk imbalances
- Comparing volume SnapMirror® and qtree SnapMirror

Previous columns:

- December — Adding a disk to the first RAID group in a volume, monitoring client workloads, enabling SSH with public keys, and using the SIO tool to test filer performance

Submit a question

Q: How can I recognize and solve disk imbalances?

In a previous column I explained how to add a single disk to a particular RAID group. NetApp performance engineering recommends against this and suggests you should always add at least an entire RAID group at a time. I'd agree—that's the best way to avoid performance issues caused by disk imbalances.

For example, my brother works for a fairly large photo-hosting service that recently upgraded from a FAS270 to a FAS3050 with SATA storage. As part of the upgrade they did a bulk migration using volume SnapMirror. After the transfer, they broke the mirror and started adding drives. Unfortunately, they waited until each volume was full and then added disks one at a time.

They began experiencing slow NFS performance, so we decided to run two diagnostics tools: statit and nfs_hist. nfs_hist indicates how slow things are running and which op is slow. statit is available in priv set advanced mode and provides a great deal of information regarding filer head utilization and the storage subsystem. Specifically, I was interested in the disk utilization section in statit.

In the sample statit output below, I've cut out most everything except for the disk stats. The first volume is the root volume, which we can ignore. We can also ignore the first two disks for this volume, which are parity disks.

The information highlighted in yellow is what I find most useful.

```
disk  ut% xfers ureads--chain-usecs writes--chain-usecs cpreads-chain-usecs
/aggr0/plex0/rg0:
0a.124  1  0.49  0.01 10.00 41200  0.44  3.81 7161  0.05 16.00 1984
0a.112  1  0.59  0.01 10.00 79200  0.53  3.48 7175  0.05 16.00 1984
0a.123  1  0.61  0.12  1.60 19479  0.44  3.81 7289  0.05 16.00 1948
/vol1/plex0/rg0:
0a.122  4  9.84  0.10  1.00 32042  3.90 13.90  967  5.84 6.05 1025
0a.19   5  9.95  0.10  1.00 78692  4.01 13.58 1050  5.84 6.04 1173
0a.20  42 49.14  42.33  2.55 7990  1.84  5.01 6928  4.97 3.72 3383
0a.21  44 51.60  45.18  2.55 7820  1.49  6.05 5732  4.94 4.10 3602
0a.22  43 49.71  42.73  2.54 8058  2.09  4.66 7805  4.89 3.76 3675
0a.23  43 50.82  44.06  2.54 7779  1.75  4.47 8383  5.01 3.81 3557
0a.24  43 51.31  44.51  2.52 7926  1.82  4.36 8749  4.98 3.82 3656
0a.25  44 51.05  44.59  2.51 8216  1.68  6.22 5944  4.77 3.84 3846
0a.26  44 51.83  44.84  2.55 8007  2.00  5.08 6727  4.99 3.81 3755
0a.27  46 52.56  46.05  2.53 8029  1.63  5.57 5920  4.88 3.82 3649
0a.28  68 68.76  60.00  2.18 13974  3.50 10.69 4258  5.27 4.84 4481
0a.29  66 69.38  60.13  2.16 12808  3.76  9.90 4248  5.48 4.42 4415
0a.113  68 69.80  60.39  2.14 13922  3.85  9.94 4505  5.55 4.49 4558
0a.114  66 69.36  60.68  2.16 12793  3.49 10.50 3935  5.18 4.47 4567
```

The 'ut%' column of the statit output shows the utilization of each disk. It is clear that the last four disks are much more utilized than the others in the volume. This can likely be explained by the 'xfers' column, which shows the number of I/O operations being served by each disk. The four busy disks are doing about 70 operations per second, while the rest of the disks in the volume are only handling around 50 operations per second. Later columns that break down the operations into reads and writes show that the four busy disks are being asked to both read and write more data than the less busy disks. These disks were presumably added after the rest of the volume was full, and all new data has been written to these disks. More reads are coming for the recently written data, and when that data is read, it's read from only the last four disks instead of all the disks in the RAID group.

Based on the statit information, we saw an imbalance between old disks and newly added disks because one disk was added at a time. In the future, adding an entire RAID group at the same time would have given the newly added data enough resources to be read effectively.

This problem will likely solve itself over time. As data is read and written again, WAFL® will find new blocks and do a very good job of balancing out workload across all the disks. In this example, the workload is fairly static (80% reads to 20% writes), and data is rarely if ever rewritten, which is typical for a Web back end.

However, since the application was already experiencing delays beyond an acceptable threshold, we decided to try to fix it quickly instead of waiting for WAFL to resolve the issue.

Data ONTAP 7.0 has a new 'reallocate' feature specifically targeted at solving this problem. It does a very good job. In this case you need to use the -f flag and a path to the volume on which you want to run reallocate. The main page for 'reallocate on the NOW™ site has more details.

Here's a sample syntax:

```
nfsfiler01*> reallocate start -f /vol/vol05
Mon Jan 2 01:44:06 PST [wafl.scan.start:info]: Starting file reallocating on volume vol05.
Reallocation scan will be started on '/vol/vol05'.
Monitor the system log for results.
nfs37201-mud*> reallocate status -v
Reallocation scans are on
/vol/vol05:
    State: Reallocating: Inode 1722, block 0 of 1
    Flags: doing_force,whole_vol
  Threshold: 4
    Schedule: n/a
    Interval: n/a
  Optimization: n/a
```

When my brother ran 'reallocate' it took a little over a day to complete and had no impact on the application. Once the command was complete, there were no more performance delays, and the application chugged along serving data again. (Ideally I'd share a new statit sample here, but once the issue was resolved my brother moved on to the next fire drill.)

If you have not yet upgraded to Data ONTAP 7.0 and have large file sizes, you can use the priv set advanced command 'wafl scan reallocate'. This is older command that is similar to 'reallocate' but is more limited and somewhat less effective.

Otherwise, you'll need to find a way to rewrite the data to a fresh volume. The process usually involves creating a new volume of the appropriate size to store the data you need rewritten and using a file-based migration tool such as NDMP copy to rewrite the data on that new volume. Incidentally, it's important that you don't try to use a block-based migration tool to rewrite the data. Volume SnapMirror and Vol Copy will attempt to transfer as much as possible a mirror of the original data, including block layouts. In essence, they replicate the problem.

This can be complex, so if you find yourself in this situation, NetApp has service offerings that can help.

Reviewer comment: Using NDMP copy to resolve a disk imbalance is an old NetApp hacker's trick that predates flexible volumes. It so happens that copying data from a tradvol to a tradvol using NDMP copy or QSM will cause the destination volume to be relaid out in the same way that 'reallocate' does. This should be viewed as the hacker's last resort. If you are not ready to move to flexible volumes, upgrading to Data ONTAP 7.0G tradvols and using 'reallocate' is generally the preferred approach.

Q: Which method is faster for replicating data in Data ONTAP 7G: volume SnapMirror or qtree SnapMirror?

For beginners, SnapMirror software provides the ability to replicate individual qtrees as well as whole volumes. The two are literally physical vs. logical. There are tradeoffs, including performance, manageability, configuration, and infrastructure resources.

The answer to this question is usually entirely dependent on your data set and your volume type. Without going into exhaustive detail, I'd sum it up this way:

- Fewer larger files work best with qtree SnapMirror (QSM)
- Lots of small files work best with volume SnapMirror (VSM)

That's a generalization, and here's why. QSM is a process that is treated like any application doing file workload may be treated. It has a finite amount of work to do for each file to be transferred during a SnapMirror update. This generally works well for a small number of large files. However, if your qtree has a small number of files in it and if the volume contains a large number of files, QSM will still have to visit each file to be transferred and therefore might perform slower than VSM. SnapMirror is dependent on snapshot copies, and all snapshots operate at the volume level. Since there aren't any qtree snapshots, QSM must look at the whole snapshot and any other data in it to determine the blocks it must transfer. If there are many qtrees and nonqtrees, there is more work to be done to determine deltas.

VSM has been around for a long time and is very efficient. In general, VSM transfers differences at the block level and hence is not affected by the kind of data that resides in the volume. The time taken to find out the data that has changed primarily depends on the size of the volume. It maps each block on the source storage system to its equivalent block on the destination storage system and will transfer any modified block on a block-by-block level. The drawback is that you are initially transferring the data of the entire volume and don't get the greater granularity of control that you get with QSM. So if you have very large volumes, this may not be a practical solution either.

Another key architecture decision, of course, is that VSM means the secondary data set looks exactly like the primary. If you have 24 hourly, seven nightly, and four weekly snapshots on the primary, that is what you will see on the secondary. With QSM Snapshot copies aren't mirrored by default; you have the ability to maintain entirely different snapshot policies on primary and secondary. Thus you might keep the secondary data versions around for longer.

For those people not using FlexVol™ technology (available with Data ONTAP 7.0), VSM also has the drawback of geometry mismatches impacting performance. VSM attempts to read and write to disks in ways that give the highest performance; therefore, SnapMirror tries to read from all the source disks and write to all the destination disks at the same time. If the disks on the source are different sizes than those on the destination, it is difficult to efficiently read and write to both sets of disks. If the destination is a volume made up from larger size disks, then VSM will not write to as many disks as on the primary volume, and hence performance will be suboptimal. If the sizes of the source disks are not the same as the sizes of the destination disks, problems can occur that result in some spindles not getting data properly distributed across them. For example, data cleanly striped across three 5GB drives on the source that is replicated to a destination system with larger disks, 15GB, would result in the data being laid out on one of the destination system's spindles instead of across all three. Note, however, that in this situation SnapMirror still works without problems; it just works a little more slowly. For optimal performance it is recommended that the source and destination disks be the same size.

FlexVol technology solves a lot of these problems. There's no geometric mismatch anymore due to the abstraction of the aggregate, and the flexibility of FlexVol allows VSM to be used in many places where you used to use qtrees. Essentially, FlexVol gives you closer to the speed of VSM without the inflexibility of traditional volumes.

More information is available in the NetApp SnapMirror FAQ on the NOW site (customers only, password required).

Submit questions or comments

© 2006 Network Appliance, Inc.