PETER BAER GALVIN

Pete's all things Sun: analytics for the masses (of storage administrators)

Peter Baer Galvin is the Chief Technologist for Corporate Technologies, a premier systems integrator and VAR (www.cptech.com). Before that, Peter was the systems manager for Brown University's Computer Science Department. He has written articles and columns for many publications and is co-author of the *Operating Systems Concepts* and *Applied Operating Systems Concepts* textbooks. As a consultant and trainer, Peter teaches tutorials and gives talks on security and system administration worldwide. Peter blogs at http://pbgalvin.wordpress.com and twitters as "PeterGalvin." His All Things Sun Wiki is http://wiki.sage.org/bin/view/Main/AllThingsSun.

pbg@cptech.com

WHILE WORKING WITH A CLIENT ON deploying a new server that was attached to its SAN, the performance was supposed to improve. Unfortunately, even though we theoretically doubled the CPU and memory being used to run the database, performance was about the same. After digging into the problem (with DTrace, of course), we determined that the SAN was providing disk I/O at very low rates. Unfortunately, the SAN had been having performance problems for quite some time but the vendor had been unable to determine the cause of the problems or, better yet, a solution to the problems. This situation is all too common. Storage is very difficult to performance-analyze and is frequently blamed (sometimes wrongly) for sitewide performance issues. Wouldn't it be great to have the convenience of centralized storage, but with the power of DTrace? Welcome to the Sun Storage 7000.

When Sun introduced DTrace [1] as part of Solaris 10, it provided an unprecedented new tool for system administrators, systems programmers, and developers. DTrace won awards, but more importantly it won the hearts and minds of those trying to analyze the performance of their applications and systems. Almost literally, systems performance analysis moved from a world where performance problems were ignored, had best-guess logic applied, or were tested via trial and error and where problems needed to be recreated in a nonproduction environment for exploration, to a new world where a problem could be dissected astonishingly quickly, in production, with no special development or compilation steps needed. In summary, DTrace moved systems analysis from the Stone Age to the Iron Age.

With the release of the Sun Storage 7000 line of storage appliances, Sun has included an "Analytics" toolkit that once again moves performance analysis forward, this time to the Modern Age. These analytics are based on DTrace but essentially hide the DTrace complexity under a cloak of Ajax-based browser graphics. With a few clicks of the mouse a storage administrator can determine which clients are causing which files on the server to be "hot," or resource-use–intensive. A few more clicks and that

administrator can see the latency of each request to the blocks of that file, or how many requests of each protocol are being processed, or how many cache hits a file had. The list is almost endless. In this episode of PATS, I'll explore the Sun Storage 7000 analytics tool, the Modern Age of storage performance analysis.

OVERVIEW

Because the new analytics are based on DTrace, we should start there. The first public unveiling of DTrace was as a paper at the 2004 USENIX Annual Technical Conference. From that auspicious start, DTrace has been the talk of the town. It has been added to other operating systems, including FreeBSD and Mac OS X, and has led to much discussion in the Linux community about adding similar features there. DTrace, along with ZFS and Containers, provides Solaris with a world-class set of capabilities that is causing even some long-term adversaries, including IBM and Dell, to support Solaris on their hardware.

DTrace is a scalpel of a system analysis tool, given its depth and breadth of abilities, but like a scalpel is best used only by surgeons. DTrace has its own programming language, and it started as command-line only. Since then, DTrace has been included in three GUI programming environments. Sun Studio Express [2], the latest version of Sun's IDE, includes project D-Light for DTrace visualization. Also, the NetBeans DTrace GUI Plugin can be installed into the Sun Studio IDE using the NetBeans Plugin Portal for Java debugging [3]. For those using Mac OS X, the free XCode IDE includes the "instruments" feature, which is a DTrace-based GUI [4]. Given their IDE flavor, these tools are mostly useful for developers, but they can be stretched for general-purpose use. Chime [5] is an open-source project that provides a visualization wrapper around DTrace. Although it is a step in the right direction, it is far from being a general-purpose DTrace GUI for system administration use.

This leads us to the Sun Storage 7000, which was announced and started shipping in November 2008 [6]. The 7000 is a line of NAS storage appliances which includes the 7110, 7210, and 7410 products. It was developed by the Fishworks team at Sun as a fully integrated software and hardware (fish) solution. It is Sun's first OpenStorage product, based on commodity hardware and open-source software (OpenSolaris), but it is far more than the sum of its parts. The 7000 line is a full appliance, with no sign of Solaris (or DTrace, or ZFS, for that matter). I won't review the full feature set here, but it includes clustering, phone home support, snapshots, clones, replication, compression, NFS, CIFS, FTP, HTTP, WebDav and iSCSI protocols, and GUI and CLI management interfaces. The 7000 line uses read-oriented and write-oriented SSDs to increase performance, while using SATA and SAS drives for density, power conservation, and cost-effectiveness. All this comes in addition to the analytics, which will be covered in the remainder of this column.

Try It, You'll Like It

There are two great ways to explore the Sun Storage 7000. Sun has a try-and-buy program for many of its products, including some of the 7000 line [7]. Sun pays to ship the system to you, and it will even pay to ship it back if you decide not to keep it. This is a very hassle-free way to be sure the Sun products meet your needs before committing to their purchase. If you prefer instant gratification and want to try out the 7000 appliance software, simply

download the Sun Unified Storage Simulator, provided by Sun as a VMware image. On Windows or Linux you could use VMware player [8], for free, to run the virtual machine. On Mac OS X there is no free version of VMware Fusion [9], but it does have a trial period. Also, virtualbox, the open-source virtual machine package now owned by Sun, should be able to play VMware images.

For this column I used the simulator, which is a full implementation of the software that runs the 7000 appliance. Although the simulator cannot be used for performance testing (or production storage use), it is as close to the real thing as is needed for evaluation, planning, and experimentation.

Analytics

The analytics component of the Sun Storage 7000 line is feature-rich. Most important, it can provide an astonishing amount of useful information to a storage administrator who is trying to manage and monitor the appliance and the files and blocks stored there. Just like DTrace, the analytics run in real time, and they allow quick progression from hypothesis through data gathering to new hypothesis, data, and conclusions. Unlike DTrace, the analytics component has a very complete and useful graphical interface and visualization engine.

Joerg Moellenkamp has posted a nice blog entry proving a walkthrough of setting up the 7000 software, configuring it to a point that it is ready to be managed by the GUI [10]. After setting up my virtual machine, I configured the virtual disks that are included in that machine to be RAID double parity. NFS service is enabled by default, so nothing was needed there. I then created the userid "pbg" for myself and created a share called "test" owned by "pbg." The share was automatically exported as /export/test. I mounted that share from my Mac and used the analytics to watch the virtual appliance. The GUI is accessed by browsing via https to its IP address at port 215.

For more details on how analytics work, take a look at the presentation put together by members of the Fishworks team that implemented them [11]. All things Fishworks-centric (videos, blogs, white papers, and more) are also available online [12].

Some examples of what Sun Storage 7000 analytics can do should go a long way toward understanding the power and flexibility of the tool. The analytics can show:

- What clients are making CFS requests
- What NFS files are currently being accessed
- How long NFS operations are taking
- What LUNS are being written to
- What files are being accessed from a specific client
- The read/write mix for a specific file by a specific client
- How long writes are taking to a specific file at a specific offset in that file

All of these metrics, and many more, are shown and optionally recorded on a per-second basis. Recorded data can be examined historically for event correlation or trend analysis. Generally, instrumentation is done at a level of abstraction above the implementation, at a level of detail that system administrators care about. The system conveys both the load placed on the appliance and how the appliance is reacting to the load. For example, a problem could be too much load or not enough appliance resources, and the details are available to make that determination. The 7000's analytics allow ad hoc instrumentation, not just precanned or predetermined diagnostics, for site-

specific or problem-specific debugging. The standard UNIX tables-and-numbers diagnostic output is frequently not easy to interpret and slow for the brain to understand, so the GUI manages visualization as a first-class aspect of storage analytics.

Let's have a look at the appliance management screen. The main "Status" screen gives an overview of the entire appliance, including space used and free, protocols enabled, and basic performance metrics (Figure 1). From there, clicking on a metric brings that metric into a "worksheet" on the Analytics page. An administrator may create many worksheets, store them, and switch among them to quickly look at various custom views of the activity of the appliance. This is done by selecting the "Saved Worksheets" subpage from Analytics. Many performance aspects are constantly sampled and made available for archiving, deletion, or adding to a worksheet from the "Datasets" subpage. Further, the administrator may have many open worksheets and can clone a worksheet to make another copy from the current worksheet. There seems to be no limit to the flexibility for viewing various system performance aspects, both current and past.

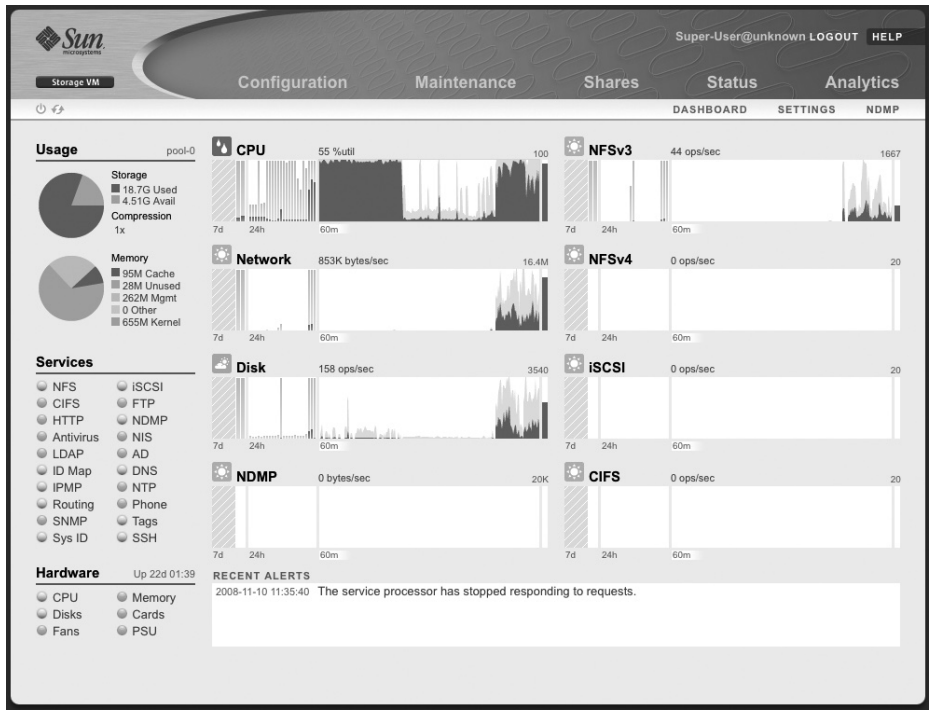

FIGURE 1: SUN STORAGE 7000 MAIN "STATUS" SCREEN

On a worksheet, the "Add Statistic..." button brings up a menu of "statistics," or system metrics, that can be added and manipulated. These statistics can in turn be broken down into constituent elements. Adding a statistic creates a new panel containing a graph of that statistic updated in real time. Averages and other "breakdown" details (for example, when a specific sample or time is clicked on within the graph) are shown to the left of the graph. Likewise, selecting an item (or multiples via shift-click) in the breakdown table highlights the corresponding data in the graph. For files and devices, a "show hierarchy" option creates a subpanel to visualize the details of that item in a pie chart and with file and device names enumerated. Again, selecting any item in any part of the panel highlights that item in the other parts of the panel. Another type of statistic is a quantified breakdown, displayed as a heat map (color-coded histogram) of the data. This is used for latency and size offsets, where scalar representation does not make sense.

Consider data where a value of zero must be distinguished from no data, such as the response time of a request.

It is difficult to describe in words the power of analytics, but the use is intuitive and the power really must be tried to be believed. A screenshot of a few graphs on one worksheet with Analytics is shown in Figure 2. There are many controls to manage a panel. Some of these are time controls. These include moving backward and forward in time, pausing the display (but not the data capture), zooming in and out (showing more or less time in the chart), and going to minute, hour, day, week, and month views of the data. Other controls manage viewing specific data, such as going to the smallest recorded sample or the largest, or directly comparing samples by showing them as line graphs rather than a stacked graph. If there is a specific time of interest in one graph, pressing the "synchronize" icon changes all graphs in the worksheet to show that time and the same time scale and to stay synchronized.

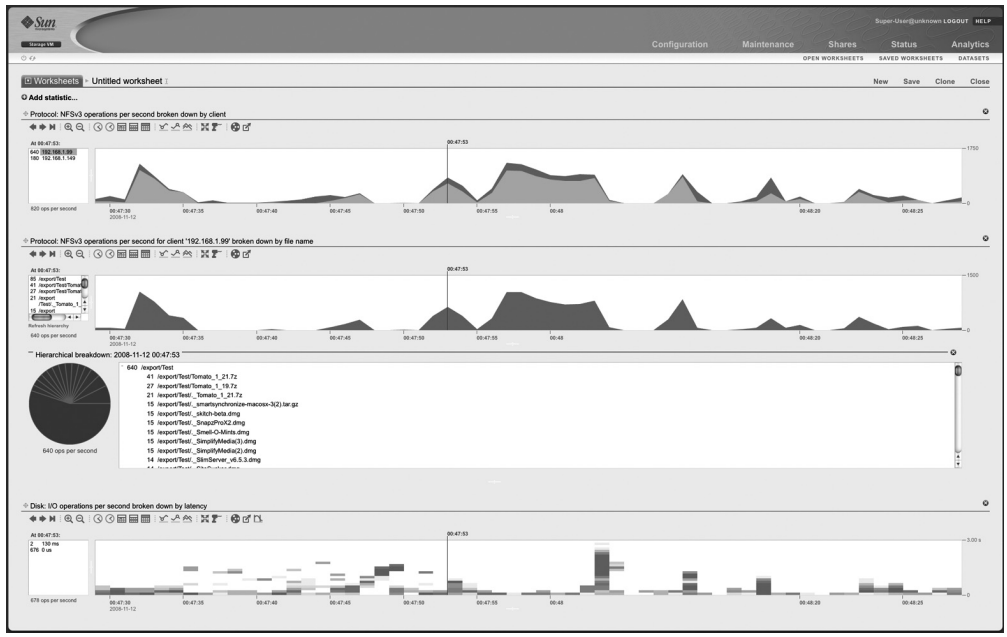

FIGURE 2: SCREENSHOT OF GRAPHS ON A WORKSHEET WITH ANALYTICS

The "drilldown" icon starts from the current graph, allows selection of a specific attribute, and creates a new graph of just that attribute of the current graph. For example, from the "CPU: percent utilization" graph, choosing "drilldown" allows selection of "by user name" or "by application name," among other choices. The drilldown choices are specific to the graph being controlled. For example, in the graph "Disk: I/O operations per second broken down by type of operation," selecting a point in time shows the types and number of each type of operation that occurred at that time. Selecting a type of operation from the list left of the graph and choosing "drilldown" allows creation of a new graph based on that operation type, but further broken down by disk, size, latency, or offset of the operation. Shift-clicking on the drilldown icon highlights every breakdown, creating a "rainbow" differentiating every breakdown on the graph.

The penultimate per-graph control saves the current graph to the Datasets collection. On the Datasets page, each graph has its data being constantly collected. Uninteresting datasets can be deleted to save space and increase performance a bit. Interesting new graphs can be saved as a dataset and the

pertinent data will then be continually collected for later examination. The last control exports the data shown in the graph to a comma-separated .csv file for importation into a spreadsheet, for example. Even more analytic options are available, and these can be enabled by selecting the "Make available advanced analytics statistics" checkbox on the Configuration->Preferences page.

Overall, the use of the analytics component of the Sun Storage 7000 NAS appliance is amazing, allowing exploration of every aspect of the performance, load, and status of the system. Sometimes the response of the GUI lagged, but the appliance's virtual machine was limited to 1 CPU and 1 GB of memory within VMware Fusion, and I was putting quite a load on the virtual machine to test out the analytics, so I expect that was an aberration. On the whole I'm glad to leave behind the Middle Ages of performance analysis and look forward to the new tools of the modern age (except for the violence inherent in the system).

Random Tidbits

This month I'm adding a new section to my column, called "Random Tidbits." This is a perfect spot for that important command, technique, or news bit that, well, doesn't really fit with the column but is important enough to talk about.

Jeff Victor has written a free and very useful new stat tool, zonestat. Jeff is a Sun employee who spends a lot of time in and around zones, even teaching tutorials about how to build and use them. One aspect of zones that was difficult to grapple with was the performance of the zones. Read Jeff's blog for more details and to download information about this very useful tool [13].

Solaris 10 11/08 (a.k.a. Update 6) shipped in November. The biggest news is a feature that I've been waiting for, for a very long time: ZFS boot / root. ZFS can now be the root file system on both x86 and SPARC systems. With this change comes the power of ZFS (snapshots, clones, replication, checksumming, and practically infinite everything) for the root disk and all system files. It is also very nice to be able to mirror the root disk with one command: `zpool attach rpool <rootdiskname> <newmirror diskname>`.

My company has started blogging about all things IT strategy. The topics there [14] will run the gamut from servers through storage, technologies, and trends. For example, my colleague Jesse St. Laurent has posted recent entries about HSM without headaches and the role of SSDs in storage. I've posted the slides I use when I teach my Solaris 10 tutorials. We hope you enjoy the blog and look forward to your comments.

A new edition, *Operating System Concepts,* 8th edition, the textbook I coauthor with Avi Silberschatz and Greg Gagne, was published in the fall but I failed to mention it here. It is the best-selling undergraduate textbook on the topic. Have a look if operating systems theory and implementation interest you.

For an interesting real-world example of the use of analytics you should check out a blog entry by Brendan Gregg in which he shows that yelling at disk drives decreases their performance: http://blogs.sun.com/brendan/entry/unusual_disk_latency. This includes the YouTube video http://www.youtube.com/watch?v=tDacjrSCeq4.

RESOURCES

[1] http://en.wikipedia.org/wiki/DTrace.

[2] http://developers.sun.com/sunstudio/index.jsp.

[3] http://docs.sun.com/source/820-4221/index.html.

[4] http://developer.apple.com/technology/tools.html.

[5] http://opensolaris.org/os/project/dtrace-chime/.

[6] http://www.sun.com/storage/disk_systems/unified_storage/.

[7] http://www.sun.com/tryandbuy/.

[8] https://www.vmware.com/products/player/.

[9] https://www.vmware.com/products/fusion.

[10] http://www.c0t0d0s0.eu/permalink/A-walkthrough-to-the-Sun-Storage-Simulator-Part-1-Initial-Config.html.

[11] http://blogs.sun.com/bmc/resource/cec_analytics.pdf.

[12] http://blogs.sun.com/fishworks/.

[13] http://blogs.sun.com/JeffV/.

[14] http://ctistrategy.com/.