Leaders in the storage world
offer valuable advice for making
more effective architecture
and technology decisions.

CTO Roundtable

STORAGE

Featuring seven world-class storage experts, this discussion is the first in a new series of CTO Roundtable forums focusing on the near-term challenges and opportunities facing the commercial computing community. Overseen by the ACM Professions Board, this series has as its goal to provide working IT managers with expert advice so they can make better decisions when investing in new architectures and technologies. This is the first installment of the discussion, with a sec-

ond installment slated for publication in a later issue.

Recognizing that Usenix and ACM serve similar constituencies, Ellie Young, Usenix executive director, graciously invited us to hold our panel during the Usenix Conference on File and Storage Technologies (FAST '08) in San Jose, Feb. 27, 2008. Young and her staff were extremely helpful in supporting us during the conference, and all of us at ACM greatly appreciate their efforts.

—Stephen Bourne

CTO Roundtable STORAGE

MODERATOR

MACHE CREEGER Principal, Emergent Technology Associates

PARTICIPANTS

STEVE KLEIMAN Senior vice president and chief scientist, Network Appliances

ERIC BREWER Professor, Computer Science Division, University of California, Berkeley; Inktomi co-founder (acquired by Yahoo)

ERIK RIEDEL Head of Interfaces and Architecture Department, Seagate Research, Seagate Technology

MARGO SELTZER Herchel Smith Professor of Computer Science, professor in the Division of Engineering and Applied Sciences, Harvard University; Sleepycat Software founder (acquired by Oracle Corporation); architect at Oracle Corporation

GREG GANGER Professor, electrical and computer engineering, School of Computer Science; director, Parallel Data Lab, Carnegie Mellon University

MARY BAKER Research scientist, HP Labs, Hewlett-Packard

KIRK MCKUSICK Past president, Usenix Association; BSD and FreeBSD architect.

CREEGER Welcome to you all. Today we're talking about storage issues that are specific to what people are coming into contact with now and what they can expect in the near term. Why don't we start with energy consumption and see where that takes us?

BREWER Recently I decided to rebuild my Microsoft Windows XP PC from scratch and for the first time tried to use a 32-gigabyte flash card instead of a hard drive. I'm already using network-attached storage for everything important, and information on local disk is easily re-created from the distribution CD. Flash consumes less energy and is much quieter.

Although this seemed like a good idea, it didn't work out that well because XP apparently does a great deal of writing to its C drive during boot. Writing to flash is not a good idea, as the device is limited in the number and bandwidth of writes. Even though the read time for flash is great, I found the boot time on the Windows machine

to be remarkably poor. It was slower than the drive I was replacing, so I'm going to have to go back to a disk in my system. But I still like the idea and feel that the thing that I need to boot my PC should be a low-power flash device with around 32 gigabytes of storage.

RIEDEL This highlights one of the problems with the adoption of new technologies. Until the software is appropriately modified to match the new hardware, you don't get the full benefit. Much of the software we run today is old. It was designed for certain paradigms, certain sets of hardware, and as we move to new hardware the old software doesn't match up.

CREEGER I've had a similar experience. In my house, my family has gotten addicted to MythTV—a free, open source, client-server DVR (digital video recorder) that runs on Linux (<http://www.mythtv.org/>). Mindful of energy consumption, I wanted to get rid of as many disk drives as possible. I first tried to go diskless and do a network boot of my clients off of the server. I found it awfully difficult to get a network-booted Linux client to be configured the way I wanted. Things like NFS (network file system) did not come easily, and you had to do a custom kernel if you wanted to include stuff outside a small standard set.

Since I wanted small-footprint client machines and was concerned about heat and noise, I took a look at flash but quickly noted that it was write-limited. Because I did not have a good handle on my outbound writes, flash didn't seem to be a particularly good candidate for my needs.

I settled on laptop drives, which seemed to be the best compromise. Laptop drives have lots of storage, are relatively cheap, can be shaken, don't generate a lot of heat, and do not require a lot of power to operate. For small audiovisual client computers, laptop drives seem to be the state-of-the-art right solution for me.

RIEDEL Seagate has been selling drives specifically optimized for DVRs. The problem is we don't sell them to the retail channel but to integrators such as TiVo and Comcast. Initially, the optimization was for sound. We slowed down the disk seek times and did other things with the materials to eliminate the clicky-clacky sound.

Recently, power is more of a concern. You have to balance power with storage capacity. When you go to a notebook drive, it's a smaller drive with smaller platter, so there are fewer bits. For most DVRs, you still care about how many HD shows you can put on it (a typical hour of high-definition TV uses more than five times the storage capacity of standard-definition TV).

BAKER Talk about noise! We have three terabytes of storage at home. What used to be my linen closet is now the machine room. While storage appliances are supposed to be happy sitting in a standard home environment, with three of them, I get overheating failures. Our house isn't air conditioned, but the linen closet is. It doesn't matter how quiet the storage is because the air conditioner is really loud.

CREEGER What we're finding in this little microcosm are the trade-offs that people need to consider. The home server is becoming a piece of house infrastructure for which people have to deal with issues of power, heat generation, and noise.

McKUSICK We have seven machines in our house and we wanted to cut our power consumption at 59 cents a kilowatt-hour. We got Soekris boxes that will support either flash or laptop drives (<http://www.soekris.com/>). The box uses six watts plus the power consumption of the attached storage device.

The first machine that we tried was our FreeBSD gateway. We used flash and it worked out great. FreeBSD doesn't write anything until after it's gone multiuser; as a result we were able to configure our gateway to be almost write-free.

Armed with our initial success, we focused on our Web server. We discovered the Web server, Apache, writes stuff all the time, and our first flash device write-failed after 18 months. But flash technology seems to be improving. After we replaced it with a 2X-sized device, it has not been as severely impacted by writes. The replacement has been going strong for almost three years.

SELTZER My guys who are studying flash claim that the write problem is going to be a thing of the past very soon.

KLEIMAN Yes and no. Write limits are going to go down over time. However, as long as capacity increases enough so that at a given write rate you're not using it up too fast, it's OK. It is correct to think of flash as a consumable, and you have to organize your systems that way.

McKUSICK But disks are also consumable; they last only three years.

KLEIMAN Disks are absolutely consumable. They are also obsolete after five years, as you don't want to use the same amount of power to spin something that's a quarter

of the storage space of the current technology.

The implications of flash are profound. I've done the arithmetic. For as long as I can remember it has been about a 100-to-1 ratio between main memory and disk in terms of dollars per gigabyte. Flash sits right in the middle. In fact, if you look at the projections, at least on a raw cost basis, by 2011–2012 flash will overlap high-performance disk drives in terms of dollars per gigabyte.

Yet flash has two orders of magnitude better dollars per random I/O operation than disk drives. Disk drives have a 100-to-1 difference in bandwidth between random and serial access patterns. In flash that's not true. It's probably a 2- or 3-to-1 difference between read and write, but the dynamic range is much less.

GANGER It's much more like RAM in that way.

KLEIMAN Yes. My theory is that whether it's flash, phase-change memory, or something else, there is a new place in the memory hierarchy. There was a big blank space for decades that is now filled, and a lot of things need to be rethought. There are many implications to this, and we're just beginning to see the tip of the iceberg.

BAKER A lot of people agree with you, and it's going to be fun to watch over the next few years. There is the JouleSort benchmark contest (<http://joulesort.stanford.edu/>) in which you attempt to achieve, within certain constraints—performance or size—the lowest power at which you can sort a specific data set. The people who have won so far have been experimenting with flash.

KLEIMAN I went to a Web site that ranked the largest databases in the world. I think the largest OLTP (online transaction processing) databases were between 3 and 10 terabytes. I know from my friends at Oracle that if you cache 3 to 5 percent of an OLTP database, you're getting a lot of the interesting stuff. That means that a few thousand dollars worth of flash can cache the largest OLTP working set known today. You don't need hundreds of thousands of dollars of enterprise hoo-ha if a few thousand dollars will do it.

With companies such as Teradata and Netezza, you have to ask if doing all these things to reorganize the data for DSS (decision support systems) is even necessary anymore.

CREEGER For the poor IT managers out in Des Moines struggling to get more out of their existing IT infrastructures, are you saying that they should really look at existing vendors that supply flash caches?

KLEIMAN No. I actually think that flash caches are a temporary solution. If you think about the problem, caches

CTO Roundtable STORAGE

are great with disks because there is a benefit to aggregation. If I have a lot of disks on the network, I can get a better level of performance than I could from my own single disk dedicated to me because I have more arms working for me.

With DRAM-based caches, I get a benefit to aggregation because DRAM is so expensive it's hard to dedicate it to any single node. Neither of these is true of network-based flash caches. You can get only a fraction of the performance of flash by sticking it out over the network. I think flash migrates to both sides: to the host and to the storage system. It doesn't exist by itself in the network.

CREEGER Are there products or architectures that people can take advantage of?

KLEIMAN Sure. I think for the next few years, cache will be important. It's an easy way to do things. Put some SSDs (solid-state disks) into some of the caching products, or arrays, that people have, and it's easy. A lot of people will be consuming SSDs. I'm just talking about the long term.

CREEGER This increases performance overall, but what about the other issue: power consumption?

KLEIMAN I'm a power-consumption skeptic. People do all these architectures to power things down, but the lowest-power disk is the one you don't own. Better you should get things into their most compressed form. What we've seen is that if you can remove all the copies that are out in the storage system and make it only one instance, you can eliminate a lot of storage that you would otherwise have to power. When there are hundreds of copies of the same set of executables, that's a lot of savings.

SELTZER You're absolutely right. Getting rid of duplication helps reduce power. But that's not inconsistent; it's a different kind of power management. If you look at the cost of storage, it's not just the initial cost but also the long-term cost, such as management and power. Power is a huge fraction, and de-duplication is one way to cut that down. Any kind of lower-power device, of which flash memory is one example, is going to be increasingly more attractive to people as power becomes increasingly more expensive.

KLEIMAN I agree. Flash can handle a lot of the very expensive, high-power workloads—the heavy random I/Os. But I am working on the assumption that disks still exist. On a dollar-per-gigabyte basis, there's at least a 5-to-1 ratio between flash and disks, long term.

SELTZER If it costs five times more to buy a flash disk than a spinning disk, how long do I have to use a flash disk before I've made up that 5X cost in power savings over a spinning disk?

KLEIMAN It's a fair point. Flash consumes very little power when you are not accessing it. Given the way electricity costs are rising, the cost of power and cooling over a five-year life for even a "fat" drive can approach the raw cost of the drive. That's still not 5X. The disk folks are working on lower-power operating and idle modes that can cut the power by half or more without adding more than a few seconds of latency to access. That improves things to only 50 percent over the raw cost of the drive.

Look at tape-based migration systems. The penalty for making a bad decision is really bad, because you have to go find a tape, stick it in the drive, and wait a minute or two. Spinning up a disk or set of disks is almost the same since it can take longer than 30 seconds. Generally, those tape systems were successful where the expected behavior was that the time to first data access might be a minute. Obviously, the classic example is backup and restore, and that's where we see spin-down mostly used today.

If you want to apply these ideas to general-purpose, so-called "unstructured" data, where it's difficult to let people know that accessing this particular data set might have a significant delay, it's hard to get good results. By the time the required disks have all spun up, the person who tried to access an old project file or follow a search hit is on the phone to IT. With the lower-power operating modes, the time to first access is reasonable and the power savings is significant. By the way, much of the growth in data over the past few years has been in unstructured data.

RIEDEL That's where the key solutions are going to come from. Look at what the EPA is doing with its recent proposals for Energy Star in the data center. It addresses a whole series of areas where you need to think about power. It has

a section about the power-management features you have in your device. The way that it's likely to be written is you can get an Energy Star label if you do two of the following five things—for example, choosing between things like de-duplication, thin provisioning, or spin-down.

If you look at the core part of the spec, however, there's a section that is focused on idle power. This is where we have a big problem in storage. The CPU folks can idle the CPU. If there is nothing to do, then it goes idle. The problem is storage systems still have to store the data and be responsive when a data request comes in. That means time-to-data and time-to-ready are important. In those cases people really do need to know about their data. The best idle power for storage systems is to turn the whole thing off, but that doesn't give people access to their data.

We've never been really careful because we haven't had to be. You could just keep spending the watts and throwing in more equipment. When you start asking

"What data am I actually using and how am I using it?" then you have to do prediction.

KLEIMAN My point is that there is so much low-hanging fruit with de-duplication, compression, and lower-power operating modes before you have to turn the disk off that we can spend the next four or five years just doing those things and save much more energy than spinning it down will do.

RIEDEL We are going to have to know more about the data and the applications. Look at the history of an earlier technology we all know about: RAID. There are multiple reasons to do RAID. You do it for availability, to protect the data, and for performance benefits. There are also areas where RAID does not provide any benefits. When

we ask our customers why they are doing RAID, nobody knows which of the benefits are more important to them.

We've spent all this time sending them to training classes, teaching them about the various RAID levels and how you calculate the XORs. What they know is if they want to protect their data, they've got to turn it up to RAID5, and if they've got money lying around, they want to turn it up to RAID10. They don't know why they're doing that, they're just saying, "This is what I'm supposed to do, so I'll do it." There isn't a deeper understanding of how the data and applications are being used. The model is not there.

SELTZER I don't think that's going to change. We're going to have to figure out the RAID equivalent for power management because I don't think people are going to figure out their data that way. It's not something that people know or understand.

McKUSICK Or they're going to put flash in front of the disk, so you can have the disk power down. You can dump it into flash and then update the disk when it becomes available.

BREWER Many disks have some NVRAM (nonvolatile RAM) in them anyway, so I feel like one could absorb the write burst while the drive wakes up. We should be able to hide that. At least in my consumer case, I know that one disk can handle my read load. Enterprise is more complicated, but that's a lot of disks we can shut down.

KLEIMAN I disagree. Flash caches can help with a lot of applications being consumed in the enterprise. Because there is a 10-to-1 cost factor, however, there are areas where flash adds no benefit. You have to let the disk show through so that cache misses are addressed. That is very hard to predict.

We've long passed the point where you can delete something. Typically, you don't know what is important and what is not, and you can't spend the time and money to figure it out. So you end up keeping everything, which means in some sense everything is equally valued. The problem is that you need a certain level of minimum reliability or redundancy in all the data because it's hard to distinguish what is important and what is not. It's not just RAID. People are going to want to have a disaster-recovery strategy. They're not going to have just one copy of this thing, RAID or no RAID.

RIEDEL At a recent event in my department to discuss storage power, we had a vendor presentation about a CPU scaling system. When system administrators feel they are getting close to peak power, they can access a master console and turn back all the processors by 20 percent. That's a system that they have live running today, and they do

CTO Roundtable STORAGE

it without fear. They figure that applications are balanced and somehow all the applications—the Web servers, the database servers—will adjust to everything running 20 percent slower.

When our group saw that, it became clear that we are going to have to figure out what the equivalent of that is for storage. We need to be able to architect storage systems so that an administrator has the option of saying, “I need it to consume 20 or 30 percent less power for the next couple of hours.”

CREEGER A mantra that I learned early on is that in databases more spindles are better. More spindles allow you to have more parallelism and a wider data path. What you’re all saying now is that we have to challenge that. More spindles are better, but at what cost? Yes, I can run a database on one spindle, but it’s not going to be a particularly responsive one. It won’t have all the performance of a 10-spindle database, but it’s going to be cheaper to run.

KLEIMAN If you think about the database example, I don’t know about that. You can put most of the working set on flash. You don’t have to worry about spinning it.

SELTZER That’s the key insight here. Flash has two attractive properties: it handles random I/O load really well, and it’s very power efficient. I think you have to look at how that’s going to play into the storage hierarchy and how it’s going to help.

In some cases you may be using flash as a performance enhancer, as a power enhancer, or both. This gets back to Erik Riedel’s point, which is that today people don’t know why they’re using RAID. It may very well be the same with flash.

GANGER The general model of search engines is you want to have a certain cluster that handles a given load. When you want to increase the load you can handle, you essentially replicate that entire cluster. It’s the unit of replication that makes management easier.

When it’s Christmas Eve and the service load is low, you could actually power down many of the replicas. While I do not believe this has been done yet, it seems like the thing to do as power costs continue to be a larger issue. In these systems there is already a great degree of replication to provide more spindles during high-load periods.

CREEGER You all said that there is low-hanging fruit to take advantage of. Are there things you can do today as profound as server virtualization?

KLEIMAN The companion to server virtualization is storage virtualization. Things like snapshots and clones take whole golden images of what you’re going to run and instantaneously make a copy so that only the parts that have changed are additional. You might have 100 virtual servers out there with what they think are 100 images, but it’s only one golden image and the differences. That’s an amazing savings. It’s the same thing that’s going on with server virtualization; it’s almost the mirror image.

What has come about over the past few years is the ability to share the infrastructure. You may have one infrastructure, but it’s still a hundred different images; you’re actually not sharing the data. That’s changed in the past five years since we have had cloning technology. This allows you to get the tremendous so-called thin-provisioning savings.

BREWER I disagree with something said earlier, which is that it’s becoming hard to delete stuff. I feel that deletion is a fundamental human right because it gets to the core of what is private and what rights you have over data about you. I want to be able to delete my own stuff, but I also want to be able to delete from groups that I no longer trust that have data about me. A lot of this is a legal issue, but I hate to feel like the technical things are going to push us away from the ability to delete.

KLEIMAN That’s a good point. Although it’s hard to expend the intellectual effort to decide what you want to delete, once you’ve expended that effort, you should be able to delete. The truth is that it’s incredibly hard to delete something. Not only do you have to deal with the disks themselves, but also the bits that are resident on the disk after you “delete” them, and the copies, and the backups on tape.

One thing that is part of our product right now, and that we continue to work on, is the ability to fine-grain encrypt information and then throw away the key. That deletes the information itself, the copies of the information, and the copies of the information on tape.

SELTZER There are two sides to this. I agree that’s a nice

solution to the deletion problem, but it concerns me because you may have an unintended consequence, which is now you've got a key-management problem. Given my own ability to keep track of my passwords, the thought of putting stuff I care about on an encrypted device where if I lose the key, I've lost my data forever, is a little scary.

KLEIMAN We have a technology that does exactly that. It turns into a hierarchical key-management system. Margo is right. When you care about doing stuff like that, you

have to get serious about it. Once you lose or delete that key, it's really, really, truly, gone.

SELTZER Given that my greatest love of snapshots comes from that time that I inadvertently deleted the thing that I didn't want to, inadvertent key deletion really scares me.

KLEIMAN That's why people won't do it, right? I think it will be done for very specific reasons with pre-thought intent that says, "Look, for legal reasons, because I don't want to be sued, I don't want this document to exist after five years."

Today data ownership has a very real burden. For example, you have an obligation to protect things such as your customers' credit card numbers or Social Security numbers, and this obligation has a real cost. This gives you a way of relieving yourself of that burden when you want to.

SELTZER I hear you and I believe it at one level, but at another level, I can't help but think of the dialog boxes that pop up and say, "Do you really mean to do this?" We're all trained to click on them and say "Yes." I'm concerned about how seriously humans will take an absolute delete.

RIEDEL Margo, you've pointed out a much bigger problem. Today, one of the key problems within all security technology is that the usability is essentially zero. In regard to Web-page security, it's amazing what people are willing to click and ignore. As long as there's a lock icon somewhere on the page, it's fine.

BREWER If we made deletion a right, this would get sorted out. I could expect business relationships of mine to delete all records about me after our relationship ceased. The industry would figure it out. If you project out 30 years, the amount you can infer given what's out there is much worse than what's known about you today.

BAKER It's overwhelming, and there's no way to pull it back in. Once it's out there, there's no control.

CREEGER Now that we all agree that there should be a way to make information have some sort of time-to-live or be able to disappear at some future direction, what recommendations can we make?

SELTZER There's a fundamental conflict here. We know how to do real deletion using encryption, but for every benefit, there's a cost. As an industry, we have already demonstrated that the cost for security is too high. Why are our systems insecure? No one is willing to pay the cost in either usability or performance to have true security.

In terms of deletion, there's a similar cost-benefit relationship. There is a way to provide the benefit, but the cost in terms of risk of losing data forever is so high that there's a tension. This fundamental tension is never going to be fully resolved unless we come up with a different technology.

BREWER If what you want is time to change your mind, we could just wait a while to throw away the key.

SELTZER The best approach I've heard is that you throw away bits of the key over time. Throwing away one bit of the key allows recovery with a little bit of effort. Throw away the second bit and it becomes harder, and so on.

BREWER But ultimately either you're going to be able to make it go away or you're not. You have to be willing to live with what it means to delete. Experience always tells us that there's regret when you delete something you would rather keep. **Q**

LOVE IT, HATE IT? LET US KNOW

feedback@acmqueue.com or www.acmqueue.com/forums

© 2008 ACM 1542-7730/08/1000 \$5.00

This article appeared in print in the August 2008 issue of Communications of the ACM.