INTERVIEW BY MARGO SELTZER

the present and future of SAN/NAS

INTERVIEW WITH DAVE HITZ AND BRIAN PAWLOWSKY OF NETAPP

Dave Hitz is one of the founders of NetApp. At NetApp, he has been a programmer, an evangelist, and the VP of Engineering. Now he focuses on strategy and culture as the Chief Philosophy Officer, asking the timeless questions: Who are we? Where did we come from? Where are we going? (See blogs.netapp.com/dave.)

Dave.Hitz@netapp.com

Brian Pawlowski is Senior Vice President and Chief Technology Officer at NetApp. Since joining NetApp in 1994, he has been involved in the design of high-performance, highly reliable storage systems.

Brian.Pawlowski@netapp.com

Margo I. Seltzer is a Herchel Smith Professor of Computer Science and a Harvard College Professor in the Harvard School of Engineering and Applied Sciences. Her research interests include file systems, databases, and transaction processing systems. She is the author of several widely used software packages. Dr. Seltzer is also a founder and CTO of Sleepycat Software, the makers of Berkeley DB.

margo@usenix.org

DAVE HITZ

MARGO SELTZER INTERVIEWING BRIAN PAWLOWSKY

MARGO: MY FIRST QUESTION IS THAT people use this term "network storage" but I think different people use it to mean different things, so in order to lay some context I'd like you guys to tell me what you think network storage is all about.

Dave: I think we should start with the technical answer.

Brian: Storage that's on a network?

Dave: There's a whole bunch of different dimensions when you look at network storage. Brian gave the answer: it's storage over a network. Yes, but does a Fibre Channel network count as a network or does network storage only include Ethernet? Sometimes people say network-attached storage, which almost always means Ethernet, but is that only file-based protocols or would iSCSI be a form of network-attached storage? And so you can get into really funny kinds of technical semantic arguments about whether a particular type of storage like iSCSI is a form of network-attached storage or not, so I'm not that interested in the vocabulary of it, but I think that there's two dimensions that matter. The first dimension is, "Are you using Ethernet, or are you using some other form of networking like Fibre Channel?" That's important dimension number one; and then the other interesting dimension is, "Is it block-based storage like Fibre Channel or iSCSI (basically, read a block, write a block, talk directly to the disk drive), or is it file-based storage like NFS or CIFS?"

Margo: So let's look at each of those dimensions. Why does it matter whether you're talking over an Ethernet or something else?

Dave: From a technical perspective, technical people tend to look at the difference between Fibre Channel and Ethernet and they say it's not that big of a difference. What really matters is where I plug into the operating system, and plugging in at the block device layer is an important distinction as opposed to plugging into the file system layer.

Margo: So that goes back to your other dimension, and I guess the question is, "Are those dimensions really separable, then?"

Dave: The block file really is where you plug into the OS, and technical people almost always argue that that's the much more important distinction. Business people tend to focus on Fibre Channel versus Ethernet, and the reason business people tend to focus on that is because they worry about things like capital expenditure. If they've spent millions of dollars on a Fibre Channel infrastructure and they're about to buy more storage, they care a whole lot whether that new storage is going

to plug into the millions of dollars' worth of Fibre Channel infrastructure they already bought or whether they're going to plug it into their corporate Ethernet infrastructure, in which case they may need to beef that up.

Brian: So there is a historical artifact here that I think was interpreted as a technical truism: that essentially the evolution of block storage went into the Fibre Channel network and Fibre Channel SANS, which were much better than using run-of-the-mill Ethernet and TCP networking, which was used for low-grade file sharing along the lines of NFS or things that you see in the Microsoft Windows network. And there was this line between the two that was more an artifact of the evolution of the two technologies than a technical requirement. Where we are today is just a total blur, first with iSCSI going over TCP/IP, Fibre Channel protocols being put over Ethernet, and block protocols being tunneled through Fibre Channel networks, and InfiniBand just playing merrily between the two camps.

Dave: This has been something that evolved over time. Ten years ago it was pretty clear where Fibre Channel would make sense and where Ethernet would make sense. If you were looking at heavy-duty database business kind of apps you definitely wanted a Fibre Channel. If you were looking at more distributed users' home directories, you definitely wanted NAS, and it was pretty distinct.

Margo: Why? Is it again just—

Dave: Because Ethernet reliability was not as strong and because the applications had not yet been modified to support NAS. If you went and talked to Oracle they would give you a list of reasons why NFS was not a good solution for running your databases. So 10 or 15 years ago there really was a strong distinction.

Brian: And even if it ran over NFS, Oracle would say they wouldn't support Oracle over NFS, which was a deal breaker for a lot of customers even if they said, "But we just ran the application over NFS and it works fine."

Dave: Oracle hadn't chosen to train their problem-solving people on those technologies and so they couldn't really help you. What happened is that Ethernet got to be much, much faster and better. Then Oracle said, "You know, this NFS stuff can save people money." So if you look at it like a Venn diagram of what are all the problems you could solve with NFS and what are all the problems you could solve with SAN, 10 years ago they were disjoint sets. There was not really any overlap. Today for the vast majority of things you might consider using storage for, you could use either one. It's gone to a Venn diagram with 90% overlap.

So from a business perspective, what I tend to believe is whatever you're already doing is probably the cheapest thing to keep doing. From a technical perspective, if you've got the opportunity to come in and redesign a bunch of stuff from scratch—not always but 80 or 90 percent of the time Ethernet storage, either NAS or iSCSI, is almost always going to be easier to manage and lead to lower cost.

Margo: So we can take away from this that in some sense these decisions are no longer important. It used to be that when I wanted storage I went and I bought the best price-performing disk I could. And it's no longer a pure price-performance choice in storage, so what are those other characteristics that you started alluding to and what are the value adds that storage manufacturers are really going to have to compete on?

Dave: Let me start top down. It's humbling as a storage vendor to recognize that CIOs do not care about storage. CIOs have some list of business prob-

lems that they want to solve and in general each of those business problems links to a particular application (e.g., all the employees need to be able to send each other email). Okay, we've chosen Exchange and so the CIO's top-level concern then is, how do I run Exchange—or, if we are going to balance the books, how do I run Oracle's financials? The more that a storage vendor can talk to the CIO about how its storage makes some kind of difference for running that application, the better off you are as a storage vendor. So—your face is all scrunched up.

Margo: That makes it sound like your value-adds are all application-specific and I'm going to claim that there's got to be a set of common value addons that you can argue will help your Exchange server and will help your financial apps and will help something else and that you can't possibly run a business having to argue each individual application independently.

Dave: There are common technologies that can help a lot of different apps, but I'll tell you that when you get into actually working with someone doing an Exchange deployment versus an Oracle deployment, they care about fundamentally different things. Let me use Exchange as just a really specific example. One of the things that people have noticed in Exchange deployments is that the Exchange database tends to get corrupted. So in an Exchange world, Exchange administrators care a lot about, "How do I get back to the earlier version of the stuff I had that used to be good?" And snapshots are a beautiful tool for doing that and so you can get back to that earlier version—and the more automated the better, right?

Think about the challenge in the real data center: You've got an Exchange administrator who typically doesn't own his own server, and there's a server administrator who typically doesn't own the network to the storage, and there's a Fibre Channel or an Ethernet administrator; and then down the line somewhere further on there's a storage administrator. Often each of those people reports to a different director and sometimes a different VP. The poor Exchange guy is just trying to get his database back the way it used to be, right? If you can somehow work with that Exchange guy and say, "Look, here's a tool that lets you do all this stuff," and now your Exchange environment is back up and running again without having to have even talked to the storage guy—that's a whole different model.

In Oracle, on the other hand, one of the big challenges is that people are always running test and development environments. They're not so worried about whether the database is corrupt, but they say, "I've got this giant production database that I'm not allowed to touch but I'm doing some little tweak in the customization that I have for SAP, say, or Oracle financials and I wish I had a playpen I could work in." Snapshots, writeable snapshots, or clones are a great tool for that. You really do have to look—there's a bazillion apps but you look at the combination of the major apps—the Microsoft Suite, the Oracle, the SAP, and VMware as an emerging one that has common characteristics: test and development environment, sort of the typical UNIX home directory. You look at that set of apps and optimize for them based on a common set of underlying capabilities. How do you virtualize your storage more? How do you create snapshots? How do you do thin provisioning? How do you do clones? You've got lots of data here, so how do you get it to there? De-duplication—those are the kinds of building blocks.

Brian: I want to make a comment, because Dave just kind of glossed over a large part of our history. It wasn't that there weren't a lot of NFS servers out there; one of the key differentiators was basically the instantaneous copying of an entire file system at essentially zero cost. And that shattered the Exchange deployment preconceptions about the time required for backup

and the number of recovery points you could have in your Exchange environment: when it divoted on you, what you hope to recover, and how fast the recovery was. Snapshots just blew away the traditional methods of doing backup to tape or any other means. Fast-forwarding, we come to that experience from the late nineties when we started seeing vast incursions into Exchange deployments for our product: a lot of times our customers were coming to us kicking and screaming about many different applications before we were giving them the tools around it.

I think there was a recognition that it's not the primary copy of data that's what is most important and of most concern to people in an organization. It's the secondary copies of data—the recovery points, the archives, etc.—and the ability to leverage and reuse data that has to be managed, because of the cost of making those copies for different purposes but also because of their usefulness in terms of business continuity. The primary copy is what everybody was designing around and everything was optimizing for. But what came circling back to everybody was the cost and value of the secondary copies, around which our fundamental technology enables interesting processes and techniques, regardless of how you access the data. How do we do data management with snapshots? How do we do disaster recoveries? Secondary copy management applies to Fibre Channel SANS and to NAS and file access.

Margo: So what I can take away is that snapshots were a truly fundamental value add that helped you differentiate early and that continued to be leveraged to solve a bunch of different business problems.

Brian: Yes.

Margo: What have you done for me lately? So snapshots were a great idea but it's 2008 and what's the next piece of core technology?

Dave: There are a handful of different ones that we can work our way through. One that I think is interesting and people don't understand the ramifications of as much as they might is RAID 6 or RAID DP—the ability to allow any two disks to fail instead of just one. When I say RAID group, I mean one parity drive with however many different disk drives; and as the number of disks gets bigger, each disk drive itself gets bigger. Say you put 10, then your overhead's 10%. You put 20: your overhead is 5%. The more disks you put in there, the more data you have to read to reconstruct a bad one. In fact, if you look at the math, disks are getting so big these days that just looking at the bit failure rate (with the current size of disks, you build a standard RAID group of 7 disks), you would expect to see failures 1% of the time on your RAID reconstructions just as a result of the raw bit failure rate of the underlying disk drives.

So imagine that doubles again because, remember, if one drive fails you have to read all of the other drives. So it was already getting bad for regular disk drives; the real challenge is what about those cheap ATA drives? Wouldn't it be nice to be able to use SATA drives? These are both bigger and slower, so it's going to take longer and be less reliable.

Margo: It's the next generation; now we're going to replace arrays of moderately inexpensive disks with arrays of really super cheap disks that are unreliable and so therefore we're going to have to go to even bigger parity.

Dave: Absolutely. Look at EMC, the way that EMC enabled the transition from the DASD style of drives to the cheaper emerging, more commoditized drives of that era was through the invention of RAID 4. I do think that ATA or SATA drives enable the next generation of this transition.

Brian: I want to connect the two topics of snapshots and RAID DP. The strength of snapshots was basically the commoditization and making snapshots available for everyone at no cost essentially compared to all other solutions. The clever part about RAID DP—having double disk protection—was not a new invention. RAID 6 was certainly okay, but no one could ever enable it because they would regret that decision forevermore because of the performance. The really clever part about RAID DP is that it was enabled with no more than a 1% to 3% performance drop versus single-parity disk protection on all our storage systems, to the point where we ship it out by default in all our systems, from our low-end SMB product, to the S500, and up to our high-end systems.

Dave: Another zone of technology is the data-replication technology that NetApp has—it turns out that snapshots are a beautiful starting point for replicating data to a remote location. The reason for that is one of the biggest challenges of replicating data: If the data that you're copying is changing underneath you and you're moving the blocks, depending on what order the data changes and depending on what order the blocks move in, you may get corruption on the copy that's of a form that even FSCK can't fix. You've got to be really careful, so in a lot of situations when you do bulk copies to a remote location, you may even have to quiesce the system. If you have snapshots as a foundation you can just copy all of the blocks in a snapshot and you know that a snapshot is static and those blocks are locked down so that changes don't go back on top of those blocks.

Margo: You said "just the same," but for using snapshots there's a time delay.

Dave: Sorry, "just the same as it looked at some point in the past; just the same but with a delay."

Brian: And with a well-defined consistency point.

Dave: Yes, with a well-defined consistency model. What a lot of customers started saying as they moved away from tapes for backup was, "I like that model, but on the remote machine it's probably made with cheap boatloads of ATA and probably a lot more drives per head, so the performance wouldn't be there necessarily for running an app but just kind of for reference." They want to keep the snapshots for a lot longer: On your primary systems you only keep snapshots for a day or two or maybe a week, but on your remote system, what if you could keep snapshots for literally a year or multiple years? We started getting banks looking at this and saying tape just isn't scaling; disks are getting bigger and faster—faster than tapes are getting bigger and faster, especially the faster part. A lot of banks have a regulatory requirement to keep data around for seven years and so they started saying "Can we have seven years' worth of snapshots, please?"

Margo: And it seems that when you get into that model of, "Okay, I need my snapshots for seven years," and they're going to be spinning, then I also have a disk lifetime problem and the disks don't necessarily last seven years, and so I also have a problem of refreshing my disk farm as it's running with these unreliable disks.

Dave: Sure. The capital lifetime of this equipment for most customers is three or four years. Some people keep it for an extra year, but really three to five years is typically the replacement cycle. So keeping a snapshot for seven years, that snapshot may not be living on the same system or the same disks, but that snapshot has the same bytes in the same file system organized structure. The snapshot may live for much, much longer than any of the physical components live.

Margo: As a customer, when I'm refreshing my vault, I assume I want to keep my vault spinning, so am I doing sort of a real-time online migration to my new vault and then sort of replacing incrementally, or am I really doing, "Okay, time to copy the vault"?

Read the complete interview transcript online at www.usenix.org/publications/login/2008-06/netappinterview.pdf.