

You're listening to the Sun Microsystems Podcast Network.

Welcome to this edition of Innovating@Sun, with your host Hal Stern. Today's topic: the Sun Fire X4500, code named Thumper. And now, here's Hal Stern.

Hal: Hello and welcome to the Innovating@Sun podcast. I'm your host Hal Stern, Vice President of system engineering at Sun, and I'm joined today by three great guests. We have Jeff Bonwick who's a distinguished engineer and the chief technology officer of our storage products group. He's also one of the co-inventors of ZFS, along with Bill Moore who's currently a hardware and software architect in our systems group. And finally Bob Sokol, the chief technologist for the content side of our communications industry group. So three great perspectives to talk about something disruptive in content management and data management, namely our Thumper product. So welcome to the show, guys.

Thanks, Hal.

Thanks, Hal.

So, Jeff you said that you kicked us off a few years ago with some observations on general-purpose systems, so why don't you kind of tell us the genesis of the process today we've nicknamed Thumper?

Jeff: Yeah. The interesting thing actually is that the genesis of Thumper and ZFS, you might think looking at the two of them, given how well they work together, that they were developed in tandem. And the truth is, it was actually a very fortunate lining up of the stars because the reason for developing ZFS was that we were looking at the existing file system products and the trends in storage and saying, you know, none of this stuff can actually cope with the scale that's coming and with the error rates that we're going to have to cope with as we have larger and larger file systems and yet the error rates on the drives are not really improving by all that much. And storage networks get, we're getting evermore complex, you would see more and more sources for error creep in and no way really of detecting that, much less being able to recover from it. So the real reason for setting out to build ZFS was to make something that would be very simple to use and that would run on general-purpose hardware because we could certainly see the trends in the computing industry in general, like for the same reason that nobody makes LIS machines anymore, in five, ten years time, you're not going to have people making rate controllers anymore either. You're getting off of these special-purpose asics and onto general-purpose CPUs, general purpose disk drives. That happened a long time ago, and meanwhile, independent of what we were doing, there was a company called Kealia where Andy Bechtolsheim was, and they were off building this box called Thumper. And at the time, they were targeting specifically the streaming media market. And in that context, it's not terribly important to have a box with great AJ characteristics. You just need a bunch of space. And if for some reason, something goes wrong with the box, well, you've got a copy on tape somewhere and you can restore it. So there wasn't a need, as in most storage boxes, to say we're going to put in NV RAM. We're going to put in rate controllers. We're going to put in all this expensive stuff. They just said, let's make this box as cheap as we can, as dense as we can because that's the market that we're going after. Well, when we bought Kealia, one of the things that we realized is that ZFS and Thumper were made for each other because what ZFS can do is take a collection of unreliable components and make reliable storage. So ZFS was very much in need of a platform to showcase this new capability, and Thumper, in order to go into a more general-purpose market, was very much

in need of a file system that could provide robustness on top of the commodity parts underneath. So the two of them together have turned out to be a great combination.

Hal: You made, I'd say, a subtle point there that I think is understated in its importance. We always tend to think of failures in the system as the catastrophic, as the disk had a head crash or there's smoke coming out of the machine. But you mentioned error rates, and with the explosion of content that we're seeing and the explosion of bandwidth to go serve up that content, the things that we always took as one in a billion or one in a trillion chances now are within the realm of normal operation, aren't they?

Jeff: They are. And in fact, I've got a - and it's not just the media errors from the disk and they'll talk to that in a sec. But I got an email from someone just last night at the CDC and they've got some Thumpers there running ZFS - sorry. This was actually not a Thumper config. They had an EMC DMX 3500 and a SAN switch and then a bunch of stuff in front of it. And it turned out that the - one of the ports in their SAN switch was corrupting data. So from the client side, it wasn't doing anything wrong. And the EMC box that they paid a bunch of money for, well, it wasn't doing anything wrong either, but the SAN switch was corrupting the data such that you send good data over the switch, it garbles it, goes on to the DMX. It then faithfully returns the garbage that was written to it, goes back to the client, and they had several file systems running on this SAN, and all the other ones were either just not noticing that they were getting bad data back or had weird failure modes, and they put ZFS on it and just immediately found out that they were getting corruption in the switch because, you know, ZFS checks everything that goes by.

Hal: But I mean, that's - more and more, I think we need to be sensitive to all these different failure modes, not just the big filter switch but the end-to-end path of the data. So -

Exactly. The EMC box has all kinds of stuff inside it to ensure data integrity, and yet it doesn't do anything to ensure the integrity of the path between the box and the application that ends up using it. And that can be just as fatal as something going wrong with one of the disk drives.

Hal: It's what the user sees that matters the most because that's where the data goes eventually. So, Bill, tell us more about your role as, I would say, the systems architect here.

Bill: So I was involved with Thumper before the Kealia acquisition even closed and, when we first saw it, as Jeff said, since I was working on the ZFS project at the time, we were looking for a box that could really, really showcase the capabilities of ZFS. And with Thumper having both compute and disk in the same box interconnected in a way such that there's no bandwidth, no bottleneck anywhere to be seen, really allowed us to showcase the technology we were working on and, as Jeff said again, this was a match made in heaven. So one of the things that we've done with Thumper since its current conception is tried to make it a little more stable, a little more robust and a little more applicable to general purpose environments rather than just the streaming environment that it was originally designed for before the Kealia guys knew about ZFS. So really, there were only a few minor tweaks made to it instead of, you know, keeping it as it was. And some of those tweaks went from a four-socket design down to a two-socket design because we were going with dual cores. We were using higher

capacity enterprise-class disks. We had, you know, vibrational issues to work out and just, you know, the things you do to take something from a prototype to a real product that you can ship to a large number of customers.

Hal: Great and if we all follow up a little bit on this notion of being a general-purpose system, Jeff, I think you've described a vision of the storage world in the future really thinking about, you know, how bits get to disk and how they live there, not so much what their format is, but really how we maintain them over long periods of time.

Jeff: Yeah. And in fact, I don't even know if I'd call it so much a vision as a necessary observation that - in the same way that the fundamental economics of the marketplace dictated over the course of the last decade or two that special purpose compute machines like, you know, LIS machines, I gave you an example before, those just - there was a time when making them made sense because they were enough faster at the particular application at hand that it was cost-effective to build such a machine if what you had was a bunch of LIS applications you wanted to run. That was compelling. But as general-purpose processors like Intel and AMD and SPARC got faster and faster, the special-purpose things that were made in smaller volume couldn't keep up with the technology curve. And what you're seeing now is the same thing that's starting to happen - in fact, already has started to happen. I mean, if you look inside the EMC boxes, for example, what you'll actually find inside is a PC running Windows as it turns out, not a bunch of special fancy hardware.

Great. So, Bob, I'd say in your perspective, talking to media companies, content companies, you know, Sun customers both old and new, what's the kind of reaction you're getting to this? Do people understand the implications of data management in a very large?

Bob: People are really excited about Thumper and the throughput capabilities and the density of the storage. I think when they start to look at it and say, all right, I've got 48 disks stuck in four rack units, and you look at the density of that and the capabilities of a general-purpose system, and they have to understand how to manage it. And ZFS is a terrific way of addressing of how do you manage the systems, the files. How do you leverage the capabilities of having the dual-core CPUs in there? How do you take advantage of the throughput? So if you look at the - something like transcoding, right? So we have customers who have hundreds or thousands of hours of content that they want to make available in different formats, so if they're using something like DV 25 as an archive format, and that's about somewhere around 11 gigabytes an hour for that format of storage. They can get tremendous amount of content on a single Thumper and leverage the throughput to push off to a transcode form and get multiple formats that they can then distribute to different content devices. And this gives them the ability to have flexibility with both how they distribute the content. So they can leverage it as a very dense content cache throughout the network, or they can use it as centralized content store and then have multiple copies that they then distribute those formats out.

Bill: Yeah. And if could elaborate on that, another thing that people are really finding with Thumper is not only is it very dense and space efficient, but it's also cost effective and power efficient. We recently had a customer do an analysis, and the computing platforms out there, the ratio of CPUs and chassis to disk is much different than our competitors' boxes. And because of that, you have to amortize the cost of separate power supplies, separate motherboards, separate processors over a much smaller number of disks. So the cost

overhead per disk and the power overhead per disk are much greater on other boxes. So with Thumper, because the ratio of CPU to disk is so much lower, that it really allows you to get a lot of power efficiency and a lot of cost effectiveness that you just couldn't get if you had to pick the different ratio.

Bob: Yeah. It's really a new price performance indicator because if you start to look at, you know, just - in the past, we may have looked in how many streams you could get for the CPU or, you know, how much you could fit in a rack in terms of CPU power and streaming throughput. You look at Thumper and you're saying, okay, this is really a different price performance because it's not just the price per CPU power, but it's the price for storage per rack unit.

Bill: And the watts per gigabyte, if you will.

Hal: So we're at a point now where we can start thinking about developers in different realms. They have application developers and you have website developers. Now we're trying to talk about data developers, people who focus on how you get the bits onto disk and how you manage that and how you add metadata and tagging and provide for index single searching, and all these other dense storage facilities, the Thumper has shaped the market here a little bit.

Bill: Yeah. One of the things that Thumper allows is because the disks have a direct connection, no cables or anything involved, but it's directly connected over the back plane, is you can have a very wide, very fast connection between the CPUs and the disk, and this allows you to get streaming capabilities that you couldn't get just simply because of the laws of physics if you had all the disks external. For example, inside Thumper, each disk has its own dedicated link between itself and the CPUs, and this goes over six full bandwidth PCI busses and six eight-port SAN controllers, and there's no bottleneck anywhere in the system. Streaming raw off the disks into server memory, assuming you don't process the data, you can get over 3 gigabytes a second sustained off the disks into memory. Now, of course, once you start processing it, that'll slow things down a little bit to, you know, two gigabytes a second or so. But still, that's the kind of bandwidth that if you were going with conventional 2 gigabyte or even 4 gigabyte fiber channel, it would require, you know, ten ports to, you know, talk to disks that quickly. And at a thousand dollars a port for fiber channel, that's just not economical.

Hal: So the disruption here comes from the packaging density, the elimination of the networking component of the storage. And just the ability to - in some cases, compress other tier of the CPU to memory to storage hierarchy, you're getting bits from disk right into usable memory around the top of the CPU.

Bill: Yeah. And with a cost of adding internal to the box, a channel from CPU to disk being so low because, you know, it traces over a back plane, it really, you know, allows you to create this wide bridge between your processing your data that you just wouldn't have in a split configuration box.

Jeff: Yeah. I was talking to a customer the other day and they were asking, in effect, can you guys make me a really, really big Thumper because if you can give me this kind of architecture, then I don't want to have my SAN anymore because all it is, is a headache. It's a source of maddening problems. It's a source of enormous cost. It's not very fast. And

there's a tremendous amount of, I think, pent up "SAN anger" in the market that Thumper can help to alleviate.

Hal: Well, I'll give you credit for being the first person to use the phrase "SAN anger" in a sentence Jeff. So any final thoughts in terms of what we're seeing in the market and what we're seeing happening now with this new emphasis on storage in a box?

Jeff: Yeah. I would just add to what Bill said that the interesting thing that I see happening here is that the thing that makes Thumper unique is that you've put not just a tremendous amount of density, but you've also put, as Bill said, compute with a high bandwidth interconnect right next to the storage. And what that means is that you can go from a model where your storage box does nothing but stream bytes to you, to more of a model where you have a storage appliance that answers questions so that you can - if you think about the way you use Google, that's exactly what it is, right? You send some small query over the net. Google doesn't, in response, stream however many petabytes of data it has to your client and then you run grep on your laptop, right? You send a question to Google. It has thousands of machines that chew on that for a moment and then sends you back an answer. So there's very little data transfer involved, and that model where you have all the compute near the storage and not near the client is going to enable a whole new class of storage applications in the years ahead.

Bob: I think from my perspective, it changes some of the discussions with customers because as you start to talk about infrastructure and workflow and how they process their content, it gives tremendous flexibility. And in terms of performance and price, if you want to have a small content cache as part of a workflow for creation or if you want to have it as part of the distribution, it really doesn't matter. So whereas before you may have depended on large storage arrays or even, you know, some tape to help augment the storage, now you can have fast, online, flexible storage that you can put into that infrastructure.

Hal: Well, great. I want to thank Jeff, Bill and Bob for joining us today, talking about the Thumper product and the future of very dense and [inaudible] storage management. Okay. You've been listening to Innovating at Sun, and I'm your host Hal Stern.

You've been listening to Innovating@Sun. Join us next time for the latest in innovation from Sun Microsystems. Only on the Sun Microsystems Podcast Network.