THIS WEEK IN HPC: ALTAIR RELEASES PBS PRO 13

Addison Snell **Michael Feldman**

December 2014

PODCAST

The following is a transcript from the weekly Intersect360 Research podcast, "This Week in HPC," available on iTunes, Stitcher, and through our media partnership with top500.org. The full podcast can be found at http://www.intersect360.com/industry/podcasts.php and is hosted at http://www.top500.org/blog/altair-announces-pbs-pro-13-a-conversation-with-bill-nitzberg.

In this special podcast episode, originally published on December 1, 2014, host analysts Addison Snell and Michael Feldman interview Bill Nitzberg, CTO of the PBS Works division at Altair, concerning the release of PBS Pro 13.

Addison Snell: Michael, we've had a lot to follow up on from [the] Supercomputing [conference], and we're still doing it.

Michael Feldman: Yes, we are. Indeed we are.

AS: And in this sponsored episode of This Week in HPC we've got a special conversation with Bill Nitzberg, who is the CTO of the PBS Works division at Altair. Bill, thanks for joining us.

Bill Nitzberg: Thanks for having me on.

AS: Bill, the reason we wanted to have you on here is because Altair launched a new PBS Pro 13. Why don't you tell us a little bit about the announcement that you had at SC14.

> ***"PBS Pro 13 is architected for Exascale, focused on million-core scalability, end-to-end resilience, and much more."***
> ***– Bill Nitzberg***

BN: We announced and rolled out PBS Professional 13 at Supercomputing. 13 is in a series of releases that we're doing focused on scaling out PBS. We think that 13 is really architected for Exascale. We completely changed some of the internal structures—the underlying framework, I should say—for PBS, and that's what the main feature of 13 is. It's architected for Exascale, focused on million-core scalability, end-to-end resilience, and really much more—power management support, and a bunch of other features.

MF: So, Bill, PBS 12.0 has been around for a couple of years now. This is obviously a major release you're doing. Can you talk about some of the major new capabilities when you compare it to the version that's currently being used.

BN: Sure. Let me, since I used the word Exascale, focus on three main things that I think of when I think of Exascale: scale/speed, reliability, and power management.

Let me do a little history of PBS. Twenty years ago, when we created PBS, we built the internal communication framework of PBS based on UDP, one of the underlying Internet technologies. And we did that because TCP 20 years ago was really not so good and had a lot of problems. Well, a lot's changed in twenty years. Web servers can now handle 60,000 incoming connections with no problem, and that just wasn't the case 20 years ago.

So, we took advantage of the change in the TCP stack that's happened over the last 20 years and re-factored all of our internal communications based on that change. We now actually have a fully multi-threaded, hierarchical, persistent connection, a fault-tolerant, non-blocking infrastructure that connects the components of PBS. And with that, we're getting about a 15-times faster job dispatch rate to, say, 100 jobs per second, supporting 10-times larger systems. We're actually testing up to 100,000 nodes. If anybody listening has a system with 100,000 nodes, we'd love to try it out for real; right now we're doing some virtualization stuff.

AS/MF: [Laughing]

BN: And I don't want to get a call from Amazon telling me I can just simply hand them my credit card, because my credit card wouldn't handle that. But also looking at throughput, so millions of jobs today.

On the resilience side, we've been building up this plug-in framework, really for extending PBS itself. We went back into the plug-in framework, and we've made it very comprehensive with respect to health checking, so we really have a comprehensive health-check framework available now, which is fully hardened as of 13. So it's, we hope, impossible to get around the health checks that we're running, so effectively we now have health-check checks to make sure they run as well.

With 13 we're also introducing C groups, and a new way of handling the launch of really wide, hero-style jobs, so that faults that are found when a large, say, 20,000-way MPI job starts, don't make that job wait in the queue again for another 20 hours before it gets a chance to run. Actually, you effectively reserve those nodes and try again right away.

"We're testing PBS Pro 13 up to 100,000 nodes. And if anybody listening has a system with 100,000 nodes, we'd love to try it out."
– Bill Nitzberg

AS: You know, this has been an interesting conversation that we've talked about throughout Supercomputing and even beforehand. What is the definition of Exascale to begin with? What's the difference between Exascale and Exaflop? Chris Willard, our Chief Research Officer, has the perspective that what Exascale really means is it's an Exaflop, plus the software environment to effectively run at an Exaflop, and that includes the entire middleware stack as well as the applications that can take advantage of it. So you're right in the sweet spot for that.

BN: When I think about Exascale, I'm thinking about the problems that Exascale brings. I know that most people think about the opportunities, but where we're sitting is: Protecting people like you from people like us.

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

MF: And it's not just about scalability. I mean, we talk about node counts, and actually people don't even know how many nodes are going to be in some of these first Exascale systems, but there are other aspects that you have to think about as far as the workload manager. What sort of things are you building into 13.0 that addresses some of the other aspects of these bigger systems?

BN: In addition to just the scale I talked about, and some of the resilience features, we're looking at making scheduling more tunable, more fine-grained. As part of 13 we're implementing additional scheduling features. Actually these are almost "bonus" things that for such a major release we were really just focusing on the scalability and the resilience, but we had time to add some extra stuff, and that is around scheduling. We have a new extension to our priority formula, we have a new Fairshare usage formula that we're adding, and a whole fine-grained preemption system that we're rolling out with PBS 13.

AS: You know, Michael asks a really relevant question with regard to the number of nodes, and I think when we look ahead to not just the most scalable systems but at HPC systems at the entry level up through supercomputer going forward, that it's not just the number of nodes we're looking at but the diversity in the nodes. One thing we've been talking about a lot recently is this evolution of parallel computing, where we've left the Beowulf era of industry standards behind, where every node looked pretty much the same. Now you're coping with all different types of parallel programming: x86, accelerated x86, GPU computing, APUs, ARM, FPGAs, DSP, POWER, what have you. What role do you see for PBS Pro as a workload manager as we go into this era of more diversity and specialization?

BN: PBS, and middleware in general, is really about trying to take a really complex and really fragile system—probably made with a bunch of commodity parts, which is what makes it fragile—and turning into a really simple to use and really hardened system. And so the greater the complexity, the more need there is for something that adds an intelligence layer into the system. We built technology into 13 primarily with the enhanced plug-in interfaces that we built, to manage power, to easily configure in Xeon Phis, GPGPUs, FPGAs, and then let you schedule them like they're real resources.

> ***"The greater the complexity, the more need there is for something that adds an intelligence layer into the system." – Bill Nitzberg***

So with 13—although with limited availability, we actually did some great demos with SGI at Supercomputing—you can treat power [energy] like a resource. You can ask for a status, how much has by job used so far? And it comes back. Ah, you're at 26.2 Kilowatt-hours. Okay, how about now? You're at 26.8 Kilowatt-hours. So you can really treat power as a resource. But you can also treat all of these other complicated things as a resource.

I should say that I'm excited by the change in High Performance Computing. When I got into High Performance Computing, (mumble, mumble) many years ago, it was a really, really diverse space, and new stuff was coming out that was crazy new! And it sort of commoditized and became very generic and bland, and as a computer scientist, somewhat uninteresting. And it's becoming way more interesting for me, way more complicated for customers, and way more valuable to have some piece of middleware in between that has a layer of intelligence.

AS: That's the trend that we see as well, that the importance of the middleware stack continues to increase as the scalability questions become more complex. You can link it to the "Solve" report that we did for the

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com

United States Council on Competitiveness[1], that showed that software scalability is the most critical limiting factor for organizations, especially commercial organizations, as they look forward to new levels of scalability, whether it's 10x what they have now or going all the way to Exascale. We're going to see an increased focus on this middleware space, and Altair[2], with PBS Pro being one of the top middleware packages we find in our surveys, you're going to be right in the middle of it. So it's going to be fun, Bill, to see how this continues to evolve. This is a well-timed new version for you.

BN: It is, and I should say that we're in beta right now, and if somebody's interested in trying out our beta, we'd love to hear from you. And our target is to launch the GA [general availability] in the first quarter of next year [2015].

AS: We've been speaking with Bill Nitzberg, CTO of the PBS Works division at Altair. Bill, thanks again for joining us.

BN: Thanks for having me.

AS: And thanks to everybody for listening. You've been listening to This Week in HPC.

[1] U.S. Council on Competitiveness, "Solve. The Exascale Effect: the Benefits of Supercomputing Investment for U.S. Industry," October 2014, free download available from http://www.compete.org/publications/detail/2695/solve/.
[2] Intersect360 Research, "HPC User Site Census: Middleware," April 2014, http://www.intersect360.com/industry/reports.php?id=106

Intersect360 Research
P.O. Box 60296 | Sunnyvale, CA 94088 | Tel. [888] 256-0124
www.Intersect360.com | info@Intersect360.com