White Paper

Intel Storage Technology and Products
Intel® Xeon® Scalable Processors

Washington University: Deploying high-speed storage with Intel® technologies

Accelerating SDS performance with Croit, Ceph, DAOS and Intel® Optane™

Contributors

Chip Schweiss
IT team lead, Neuroinformatics Research Group (NRG), Washingtom University School of Medicine

Denis Nuja
Technical Pre-sales Engineer, Croit

Andy Muthmann
Director, Sales, Croit

Chris Feltham
Industry Technical Specialist (Cloud), Intel UK

Table of Contents

Executive summary

At Washington University School of Medicine in St. Louis, Missouri, Chip Schweiss and his team lead the development and management of the IT infrastructure at the university's Neuroinformatics Research Group (NRG).

The NRG is responsible for developing technologies that support neuroimaging and imaging informatics to help scientists better understand the living brain. The group also creates open-source software used around the world to study and treat various neurological conditions, including: Alzheimer's disease, schizophrenia, depression, brain cancer, autism and strokes.

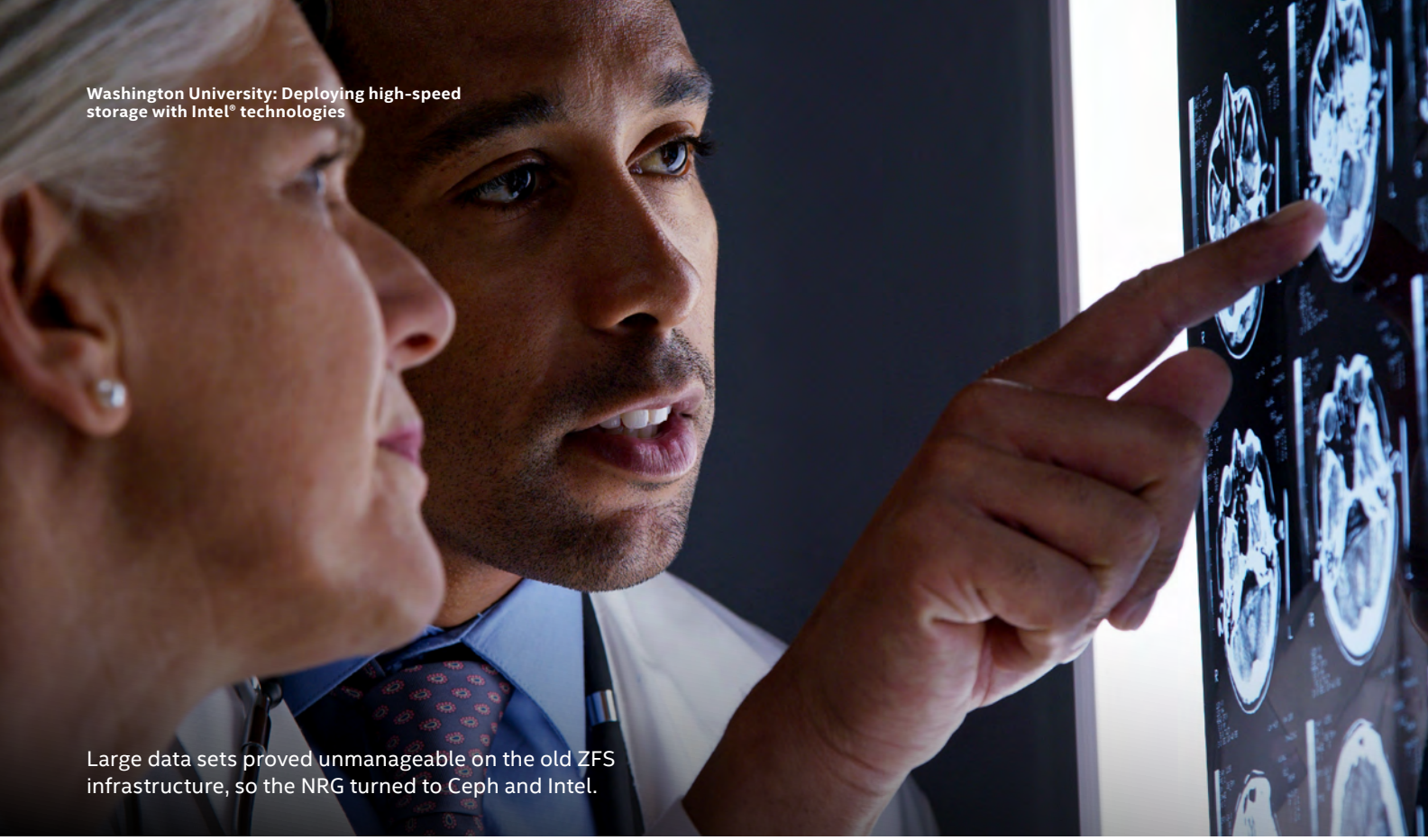
Faced with growing storage requirements, both in capacity and speed of access, Schweiss and the NRG embarked on a project with SDS/Ceph experts Croit to deploy a new storage system powered by Intel technology. Increased transfer speeds, a reduction in operational costs and lower management overheads will see the abilities of the NRG increase. This will give the IT team more time to focus on other areas of the project.

Hitting the storage limit

In developing technologies to support its medical research, Washington University's Neuroinformatics Research Group constantly faces a need for more storage capacity. Over the last decade, huge demands have been placed on the existing system with storage requirements growing exponentially.

"Washington University is probably the largest radiology research university in the United States," explains Chip Schweiss, who leads the NRG IT team. "We do software development and IT infrastructure for radiology researchers. This means pretty much anything that uses an MRI or a PET scanner, CT, or any of those technologies that you produce imagery from. Through the research, we generate tonnes of data... I've been with [the university] for about 10 and a half years. In that time frame, storage has grown from 300 terabytes to something approaching 12–13 petabytes today."

A decade ago, the university had moved its storage to a ZFS platform. This file system initially suited their requirements. But Schweiss admits that he "had to develop a lot of scripting just to manage replication, snapshot and things that there weren't good add-ons for." So, it was never a perfect solution.



Large data sets proved unmanageable on the old ZFS infrastructure, so the NRG turned to Ceph and Intel.

As the NRG's storage demands have grown, so have the complexities of the ZFS infrastructure they use, adding additional management overhead into the mix. “The biggest thing has been just trying to keep up with growth,” says Chip Schweiss. “We built our first ZFS pool fully to a petabyte. But what we ran into is [that] it becomes unmanageable at that size, and you don't gain scale or performance past about 180–200 disks.

“With that limitation, we started building more ZFS pools. But then you're splitting the data... We have one dataset now that’s split across six different storage pools. It's a good hour or two per week of my time just to keep that data balanced out on those six pools. So, there's been a lot of overhead. Not just because of the sheer size of the datasets, but because of the limitations of the software dealing with those datasets.”

A simpler solution with Croit

Faced with these inefficiencies, it was clear that the NRG needed to upgrade to a cluster file system that didn't have an upper storage limit. This would not only enable them to cope with current storage demands, but allow room for future expansion. As the NRG is grant-funded, turning to an open-source solution made the most sense. So, Schweiss and his team settled on Ceph.

Ceph is an open-source software-defined storage (SDS) platform that uses object storage on a distributed cluster, providing object, block and file-level storage. It’s built to be self-healing and self-managing to reduce admin time. After an initial Ceph project with another supplier fell through, Schweiss sent out what he dubbed a “hail Mary” email to the Ceph community. Within 20 minutes, he’d got a reply. It came from the team at Croit, a Ceph specialist that has developed software to make installation and management of a Ceph system faster and smoother.

After an initial discussion, Schweiss was taken with how Croit’s software made it easier to install CephFS. As Croit's Technical Pre-sales Engineer Denis Nuja points out: "Install manuals [for Ceph] are approximately 60 pages of CLI. So we do it in a nice web UI in 15 minutes."

Reduced cost and higher performance

As well as promising simplicity and reduced management overheads, working with Croit helped the NRG spec the right hardware for their next generation storage nodes. Thanks to a close working relationship with Intel, Croit better understood the components that would be required for the new Ceph system to run optimally. It could also make sure that there was a clear pathway to upgrade it.

“We saved roughly $1,500 per storage node on the new system because of the changes,” explains Schweiss. The new configuration (detailed below) used higher performance Intel® Xeon® Processors and more memory.

Importantly, the partnership with Croit isn't static, and as the project for Washington University's NRG has evolved, the requirements have been adapted. “As we progressed the relationship and started installing it, we also started discussing some of their other needs," says Denis Nuja. This led to discussions about other technologies, such as Intel® Optane™ Persistent Memory and Intel DAOS (see page 4).

An MRI image can consist of several thousand JPEGs with detailed metadata, all of which need processing and storing.

Although Ceph is hardware agnostic and can utilise any hardware and storage, Intel specialises in providing the full technology stack, delivering performance, reliability and scale. When pulled together by a company like Croit, Intel's hardware portfolio can be fully utilised. With this in mind, the NRG project is a multi-phase rollout, with storage nodes that will use the latest CPU, storage and memory technologies.

3rd Gen Intel® Xeon® Scalable Processors

The Ceph nodes for phase two (adding to the NRG's original 14-node cluster) use a combination of Intel hardware and high-speed Ethernet networking. For compute, Supermicro Storage SuperServers are equipped with 3rd Gen Intel® Xeon® Scalable Processors. Running on a 10nm process, these processors offer a 1.46x[1] average performance gain over the previous generation of CPUs.

Key for faster storage is PCI Express Gen4, which doubles the bandwidth over the previous generation and enables support for higher performance Intel® 3D NAND SSDs. 3rd Gen Intel® Xeon® Scalable processors also support Intel® Optane™ Persistent Memory, which can operate as affordable larger capacity memory or as extremely fast persistent storage.

For the initial nodes, Washington University used Intel NVMe drives, to balance performance against cost and capacity.

Intel's solid-state memory technology is built to deliver the performance that data-intensive workloads require. So, the NRG's phase two specs include six Intel® SSD DC P4610 Series 1.6TB NVMe drives in a trio of MDS/MON nodes. There are also two Intel® SSD DC P4511 Series 1TB NVMe drives in the admin nodes (one hot, one in cold standby).

Beyond hardware, Intel also provides software to push the limits of what its hardware can achieve, including support for the Intel Intelligent Storage Acceleration Library (ISA-L). A collection of low-level applications, ISA-L provides the tools to minimise disk space use and maximise throughput, security and resilience.

Despite the potential complexity of these systems, thanks to Croit's software and expertise, the initial Ceph installation took just an hour to roll out.

Freedom to explore new projects

The phase two rollout will use a mix of legacy and new Intel-powered Ceph nodes. So, there's an expected performance boost in addition to increased capacity. Where the original hardware provided 2.5 PB of 6+2 erasure-coded CephFS storage, phase two will boost that to around 6 PB. Phase three will double the total to 12 PB.

"With the first expansion, they're getting more metadata servers," says Croit's Denis Nuja. "This includes some more storage servers on the Intel platform. So, I would expect that the system, when it arrives with the first expansion, will be doing something in the line of 40 gigabytes a second."

Chip Schweiss agrees and is encouraged by the early results. "The performance testing we've done with Ceph is an order of magnitude faster than we were getting with ZFS. So out of the gate, it's going to be better."

Crucially, having a scalable solution like Ceph means that all of the NRG's data can be stored in one place. The management of that data becomes much less onerous as a result. This frees up time for the IT team to focus on other tasks and to shift resources to help support the NRG in other areas. For example, one issue that the NRG had noticed was how jobs were handled.

On the ZFS infrastructure, processing pipelines were handled by scripts that were put onto a job scheduler on an HPC to do the processing. One of the big downfalls with this approach is that, if you build a data pipeline for these complex mathematical interpretations of imagery, simply upgrading software might change results. So, you end up with incomparable datasets because they were processed at two different software levels.

Washington University is one of the largest radiology research universities in the United States.

Shifting existing pipeline processes into Docker containers gives consistency across the entire stack. But it's a matter of time to implement everything. As Chip Schweiss explains: "Docker containers are a really good solution to getting that done, but we have not had the IT bandwidth to facilitate their needs. I haven't been able to support them that much because I 've spent so much time dealing with storage."

Plans for future expansion

As previously mentioned, beyond the two current phases that are already provisioned, Washington University is looking to further expand its Ceph system. As it replaces the old ZFS system, each iterative phase is designed to expand capacity and performance, all without adding additional overheads for system management.

"We have a third phase in plan that we just received grant funding for. We're still in the research phase of how much compute and how much storage we are going to be purchasing with this," continues Schweiss. "We are potentially looking at doubling our Ceph cluster size again with that grant, to have potentially about 10–15 petabytes of Ceph online, both primary and DR."

Ceph and Intel® DAOS

Of course, while the Ceph storage solution solves a lot of problems that Washington University's NRG has faced, it also has its limitations. According to Schweiss, the Ceph system doesn't quite look ready to be the primary storage system for HPC machine learning tasks. "We may still have a staging storage for those datasets as we try and process them on that," he says.

Another issue that the university faces is what to do with metadata. Although the Ceph system is built to accelerate the processing of this, metadata is still, as Schweiss points out "the bottleneck" in their system.

"For example, an MRI image is a slice of anywhere from a few hundred to a few thousand JPEG images that you're trying to consume and process in relation to each other and possibly additional MRIs, especially when we're going to machine learning. So, now we're dealing with slice sets from hundreds or thousands of people that we're consuming at a high rate. This drills down to all these relatively small files in the hundreds of kilobytes range. So, metadata becomes a thorn in our side," says Schweiss.

Intel is set to help with this problem too. Intel DAOS is a next-generation open-source storage platform, built to take advantage of new solid state technologies. Traditional storage stacks are designed for spinning media. DAOS, however, is being built from the ground-up to use new NVM components, including Intel® Optane™ Persistent Memory.

Fitting into a traditional DRAM socket, Intel Optane Persistent Memory comes in higher capacities than traditional memory (up to 512GB per module). It also retains data, even after a server is powered off. Persistent Memory is also far more cost-effective than DRAM. So, putting such high-speed storage close to the processor opens up a new world of possibilities for the most demanding workloads.

"Going forward, Croit is working on integrating DAOS into the stack with Ceph," says Schweiss. "My understanding is they're

actually looking at making DAOS a possible metadata layer for Ceph, which is [Ceph's] biggest performance downfall currently. And that will be very significant for us if we can do that in the future."

It will be the integration offered by Croit that will make DAOS more easily accessible. Just as the company did with Ceph, Croit can work with Intel to take all of its technologies that deliver leading performance. These include Intel Optane Persistent Memory, Intel Xeon Scalable Processors and Intel® Solid State Drives. It can then combine them with leading software that takes away the complexities. As a result, customers can enjoy a new storage dynamic with ease.

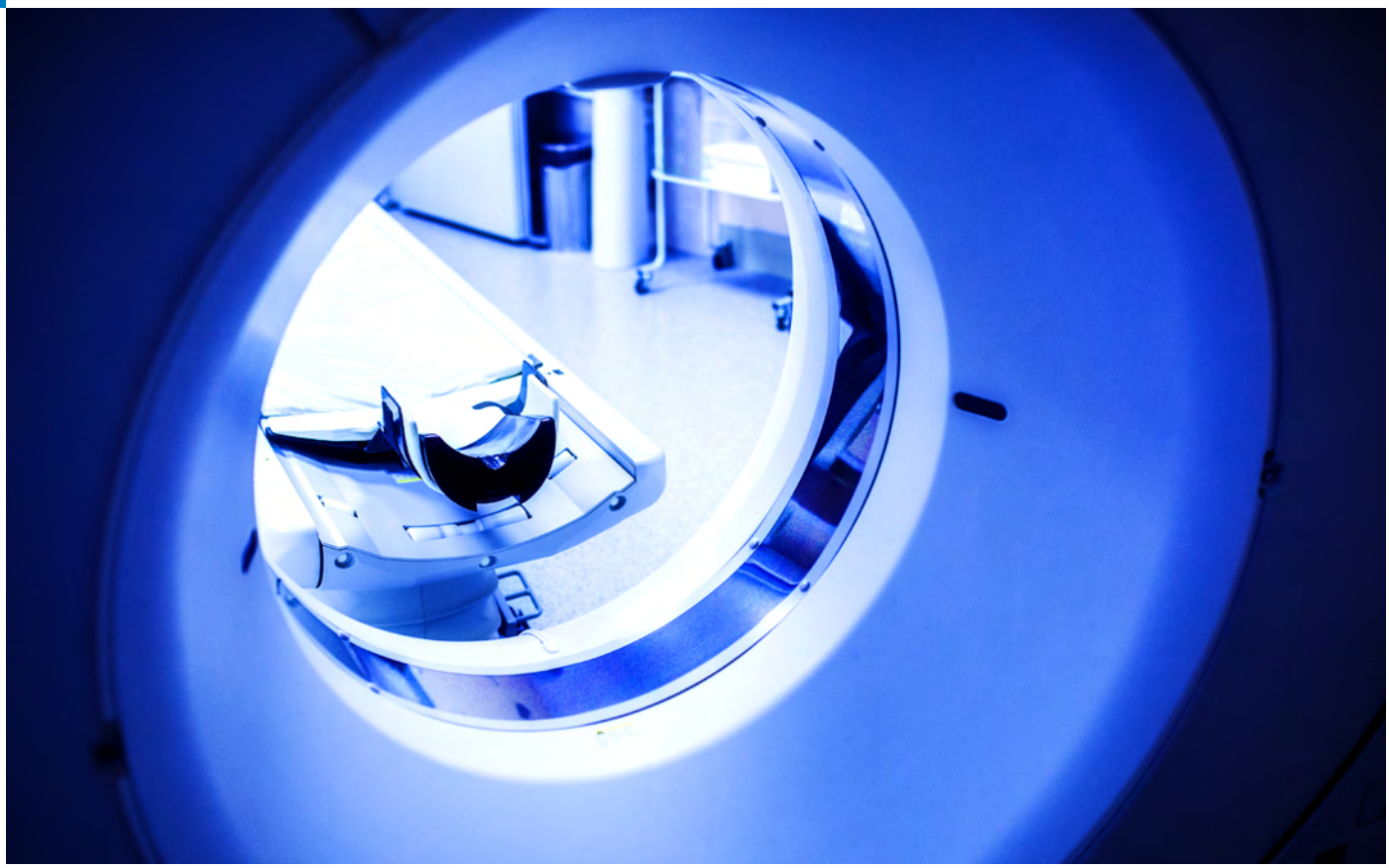
By redesigning the storage hierarchy, the NRG can leverage Intel Xeon Scalable Processors and Intel NVMe technology on Ceph to deliver more performance. Not only that, but with Croit's expertise and its experience as an Intel partner, it can ensure that the NRG is getting the right storage tiers to suit their compute requirements.

3TB to 13PB
The growth in storage requirements over 10 years

20PB
The planned total amount of storage for the new Ceph system

$1,500
Approximate cost saving per storage node Washington University made

50GB/s
The expected throughput of the new storage system



"Intel is actually allowing us to cover the whole spectrum of storage needs," says Croit's Denis Nuja. "With Intel solutions, we can cover everything from backups to active second tier storage to tier zero storage, all with very high performance. The Intel technology and our software work hand in hand. So, we can create what is effectively a full spectrum solution for any storage that customers might need these days."

Learn More

You may find the following resources useful:

- Intel® Xeon® Scalable Processors
- Intel® Optane™ Technology
- Intel Distributed Asynchronous Object Storage (DAOS)
- Croit

Solution provided by

intel. croit

[1] https://www.intel.co.uk/content/www/uk/en/products/docs/processors/xeon/3rd-gen-xeon-scalable-processors-brief.html

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.