AMPERE.

TheNogalesLab | CASE STUDY

Electron Microscopy runs 60% more efficiently at Berkeley's Nogales Labs on Ampere

SNAPSHOT

Organization: Nogales Lab is a cryo-EM lab on the UC Berkeley Campus. Cryo-EM (cryogenic electron microscopy) allows scientists to make detailed 2D and 3D images of single particle and microtubule biological structures to study macromolecular assemblies as whole units of molecular function by direct visualization of their architecture, functional states, and regulatory interactions.

Challenge: The CPU-intensive compute of the EM lab was outgrowing the Intel x86 processor-based servers installed when the lab was first built. The challenges of a small data center footprint, the need for more efficient processing capacity, and an overheating problem that caused thermal throttle sent the lab's HPC Systems Administrator, Kurt Stine, to seek an entirely new CPU architecture to meet their growing needs.

Solution: According to Stine, deploying the Ampere instance in a Supermicro server was surprisingly easy. "The integration was just as easy as dropping in another Intel server. Everything just worked—which was great and what we were really looking for," he said.

Results: While the team was originally only looking to replace their legacy Intel servers, they were surprised to discover that the Ampere® Altra® Cloud Native Processors consumed 60% less system power than their newer AMD EPYC system. In addition, the much lower power consumption and heat generation of the Ampere-powered servers gave researchers 40% more overall compute capacity.

In recent years, cryo-EM has gained popularity as a method for studying biological structures. Using cryo-EM, scientists are creating 3D images of ribosomes, proteins, and viruses and then stitching together thousands of these images to study how structures within a molecule move and interact.

At Nogales Lab, these studies were often halted when the data center reached thermal capacity which required time for servers to be cooled before studies could be continued. A limited data center footprint in a building never designed for a data center and a budget dependent on frequent fluctuations in funding made expanding the compute capacity challenging.

"We found ourselves needing more and more CPU specific servers, and with the increase in the Intel and NVIDIA tax, we began looking for options that could deliver similar performance while lowering power consumption and price. That's where we decided on Ampere," said Nogales Labs HPC Systems Administrator, Kurt Stine.

INCREASED POWER CONSUMPTION PROMPTS AN ARCHITECTURE CHANGE

Even though Relion workloads—typically used for cryo-EM—had never been run on Arm-based processors in the past, the need for more cores, optimal processing speeds, less power consumption, and less strain on cooling systems led Stine to reach out to the high-performance computing community to find a new solution on behalf of Nogales Labs.

It was there that Stine was made aware of the innovative architecture of Arm-based processors and Ampere's Cloud Native solutions. Through further research, Stine was impressed by the performance and efficiency Ampere had achieved on other compute intensive workloads and wanted to see if he could get similar results on the Relion workload.

"Because our x86 servers continued to require more and more power, I began conversations with the high-performance computing community and conducted my own research into other alternatives," Stine recalled. "What I discovered was that an entire architecture change was needed to lower the power consumption, lower the heat generation, and still be performant."

LOWER POWER RESULTS IN 40% MORE COMPUTE CAPACITY

The combination of high-performance computing, a small data center footprint, and power-hungry Intel x86 servers created cooling challenges that often resulted in temporary data center shutdowns. In some cases, thermal throttle was observed. The team at Nogales knew that they didn't have the cooling capacity to handle increasing their CPU capacity but needed a way to squeeze more compute from their small data center.

"One of the issues has always been cooling in our environment. Since our Intel CPU servers run a lot hotter, we can only use about 80% of the CPU cores. On a 56-core server from Intel, we're only able to use about 50 cores. If we use all 56 it will start running overly hot and we've seen it start to thermal throttle," noted Stine.

By switching just one Intel server in the data center to an Ampere Cloud Native server, Mega DC ARS-110M-NR from Supermicro, the researchers at Nogales Lab were able to run more tests and get results faster. They were also surprised by benchmark studies showing that they could get as much compute from the Ampere Altra Max as from the AMD EPYC VMs at 60% less power consumption—prompting them to reconsider their former decision to switch to AMD from Intel.

"With Ampere, we can use all 128 cores and it just won't thermal throttle at all. We can get full performance with all 128 cores being used, and since we're using all 128 cores, our researchers can complete jobs faster," according to Stine. "One of the main reasons we adopted Ampere was the lower power consumption and resulting lower cooling capacity requirements. It was the perfect choice."

40% BETTER PERFORMANCE PER WATT THAN AMD EPYC

From the beginning, the Nogales Labs team saw improvements in what they could accomplish with Cloud Native Processors from Ampere. Although Relion is not optimized for Arm-based Cloud Native Processors currently, the results were impressive.

Performing his own benchmarking studies, Stine discovered that not only does the Ampere Altra Max deliver 40% better performance per watt, but he also saw equivalent and sometimes better performance than AMD EPYC.

"Our initial results with Ampere looked really good—even without Relion being optimized—we're getting similar performance numbers to EPYC. For the vast difference in price, it's perfect for our compute requirements and our limited budget," stated Stine.

The surprising results have caused the team at Nogales to look toward a future where they can continue to increase their compute density without impacting power consumption and cooling systems. In the future, Stine would like to switch over all the CPU servers to Ampere—depending on the funding he receives and when those funds become available for the data center.

"Our next purchase cycle is in four months, and I plan on replacing one or two of our Intel CPU servers with Ampere, just because they're crushing. In the future, I'd like to see nearly all of them switched over," Stine said.

About Ampere

Built for sustainable cloud computing, Ampere Computing's Cloud Native Processors feature a single-threaded, multiple core design that's scalable, powerful, and efficient.

Learn more

See our solutions for a variety of demanding workloads: https://amperecomputing.com/solutions

Visit our Developer Center: https://amperecomputing.com/developers

	Altra Max M128-30	2 AMD Epyc 7702	2 Intel Xeon 6238R	Power10 48-core, 384 thread
Benchmark	128 cores / CPU	64 cores / CPU	28 cores / CPU	48 cores / CPU
3D Classification	58 hours, 22 min (128 core)	42 hours, 12 min (256 threads)	46 hours, 12 min (112 threads)	26 hours, 13 min
AutoPicking	1 hours, 17 min (64 core)	1 hours, 2 min (64 core	1 hours, 3 min	29 min, 28 sec
2D Classification	24 min, 12 sec (128 core)	21 min, 4 sec (256 threads)	52 min, 24 sec (112 threads)	15 min, 23 sec
Notes	Power draw 515W	Power draw 1300W		Does not support GPUs, cost 12X vs Amper

All tests performed with Relion 4.0.1 and CPU acceleration 0.603846154

Disclaimer

All data and information contained in or disclosed by this document are for informational purposes only and are subject to change.

This document is not to be used, copied, or reproduced in its entirety, or presented to others without the express written permission of Ampere®.