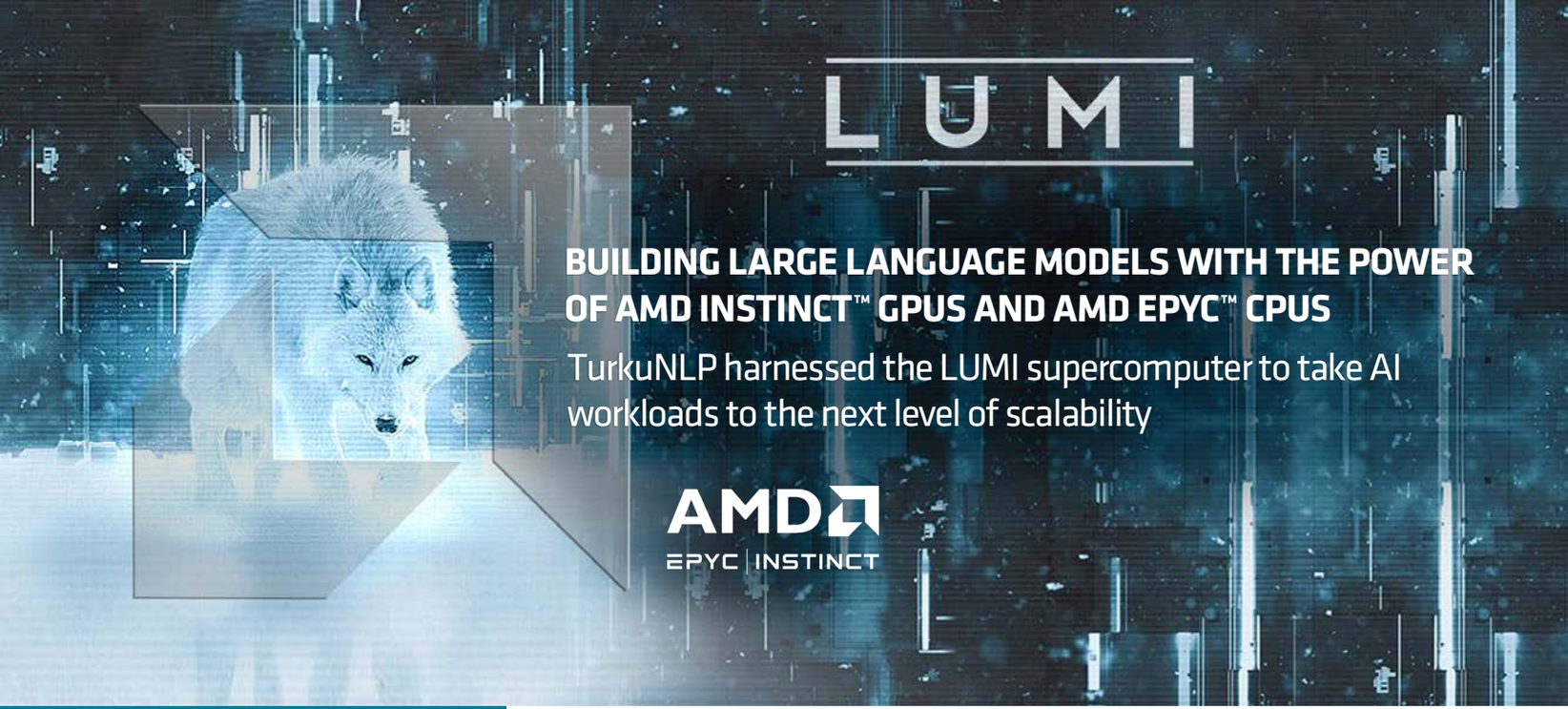


BUILDING LARGE LANGUAGE MODELS WITH THE POWER OF AMD INSTINCT™ GPUS AND AMD EPYC™ CPUS

TurkuNLP harnessed the LUMI supercomputer to take AI workloads to the next level of scalability

CUSTOMER

LUMI

INDUSTRY

Research and education

CHALLENGES

Large Language Models require high-performance computing with massive scalability to ensure sufficiently rapid iteration

SOLUTION

Deploy LUMI supercomputer powered by AMD EPYC™ CPUs and Instinct™ GPUs

RESULTS

Scaling to 192 nodes, taking two weeks to run training on a 176 billion parameter model for 40 billion tokens and several smaller monolingual Finnish models for 300 billion tokens

AMD TECHNOLOGY AT A GLANCE

AMD EPYC™ CPUs
AMD Instinct™ MI250X GPUs

TECHNOLOGY PARTNER

Hewlett Packard Enterprise

There has been a lot of interest in Large Language Models (LLMs), thanks to the high profile of ChatGPT. But training an LLM takes a huge amount of compute power, and models like ChatGPT are usually both proprietary and based on English. When University of Turku Research Fellow Sampo Pyysalo wanted to extend the value of LLMs to wider research applications, he needed performance to train the models in a useful timeframe. The LUMI supercomputer, based on the HPE Cray EX supercomputer architecture and powered by AMD EPYC™ CPUs and AMD Instinct™ GPUs, provided the scale the workloads needed.

Opening Up Large Language Models

Pyysalo's goal with partners Risto Luukkonen and Ville Komulainen in the, TurkuNLP, was to open up LLMs for academic use. "The big players are large multinational corporations who keep their models closed," he says. "In academia we want practical access to models like these, so we have been creating them ourselves and this requires supercomputer resources." Finnish was the natural starting point for a university based in Finland, such as Turku. "We've created foundation models that must be fine-tuned for specific research requirements. The next steps include training these models so they can follow instructions in a sensible way or work as part of a dialogue like ChatGPT."

"AMD did a great job importing the most important software in this area to their platform."

Sampo Pyysalo, Senior Researcher, TurkuNLP, University of Turku

Building LLMs relies on advanced Artificial Intelligence (AI) and Machine Learning (ML) toolsets. Pyysalo has been working with Hugging Face for this. "We've collaborated with Hugging Face on several projects," he says. "We were part of the BigScience initiatives that created BLOOM, the largest open language model. The biggest model we trained in this effort during the LUMI pilot was to teach BLOOM Finnish. We took the 176 billion-parameter model that Hugging Face had created and combined this with Finnish using 40 billion more words."

Models of this size require immense computing scale, which is where LUMI proved essential. "We found out that this wonderful supercomputer was going to be available," he says. LUMI, owned by the EuroHPC Joint Undertaking, was funded 50/50 by the EuroHPC JU and the LUMI consortium consisting of ten European countries. It is based in Finland at CSC – IT Center for Science's data center – and hosted by the LUMI consortium. The LUMI-G GPU partition dwarfs other GPU partitions hosted by CSC. The organization's Mahti AI consists of 24 GPU nodes, while Puhti offers 80 GPU nodes, both with four GPUs per node. LUMI, in contrast, boasts 2,560 nodes powered by AMD EPYC processors, each with four AMD Instinct MI250x accelerators, for a total of 10,240 GPUs and 20,480 Graphics Compute Dies (GCDs).

AMD provided comprehensive assistance with getting Pyysalo's LLMs to work on LUMI. "AMD did a great job importing the most important software in this area to their platform," he says . "We used the Megatron DeepSpeed language model software, which had been ported. That was the foundation that we built on. We took BigScience, the Hugging Face fork of Megatron DeepSpeed, and then AMD ported the ROCm kernels. AMD technical staff also worked closely with us during the LUMI pilot period helping us get over bottlenecks. For example, a communications overhead issue was resolved using a custom module with libfabric access. That fundamentally changed our ability to continue scaling to several hundred nodes."

Scalable Performance with LUMI

"We need a lot of compute to create a model in a reasonable timeframe," says Pyysalo. "The big challenge at this scale is getting stuff to run at all, but also being able to maintain throughput. We must be able to pull data efficiently from storage, run efficient kernels, and shuffle data back and forth between GPUs and the main memory. Another big scaling challenge is communication. After each of the GPUs computes its parts of the model, everything must be integrated. We need reasonable overall throughput while distributing the computation over hundreds or thousands of devices."

"Large-scale experiments like this are providing really valuable information for us," says Väinö Hatanpää, Machine Learning Specialist at CSC. "The optimization of the libfabric connection, for example, gives valuable information for CSC that we can include in our guides, which then helps others to use our systems more efficiently. The computing capacity and the ability to scale further with LUMI [powered by AMD EPYC™ CPUs and AMD Instinct™ GPUs] enables our customers to push the boundaries of Machine Learning/AI."

"The computing capacity and the ability to scale further with LUMI [powered by AMD EPYC™ CPUs and AMD Instinct™ GPUs] enables our customers to push the boundaries of Machine Learning/AI."

Väinö Hatanpää, Machine Learning Specialist at CSC

The difference in scale LUMI provides cannot be overemphasized. "Around four years ago we trained our first Finnish BERT model on CSC's previous generation supercomputer," says Pyysalo. "That was a pilot project on the CSC, a 110 million-parameter model. But the biggest one that we trained on LUMI was 176 billion, more than a thousand times larger. LUMI is two orders of magnitude bigger than the previous generation machines available in Finland. It would have been inconceivable to do something at this scale on the hardware that was previously available to us."

"The speed improvement gained from the ability to extend scaling is important if you need to iterate quickly," says Hatanpää. "I have pre-trained a 1 billion-parameter language model on my own computer, but it took me half a year. It's rarely realistic for a research group to spend half a year on training."

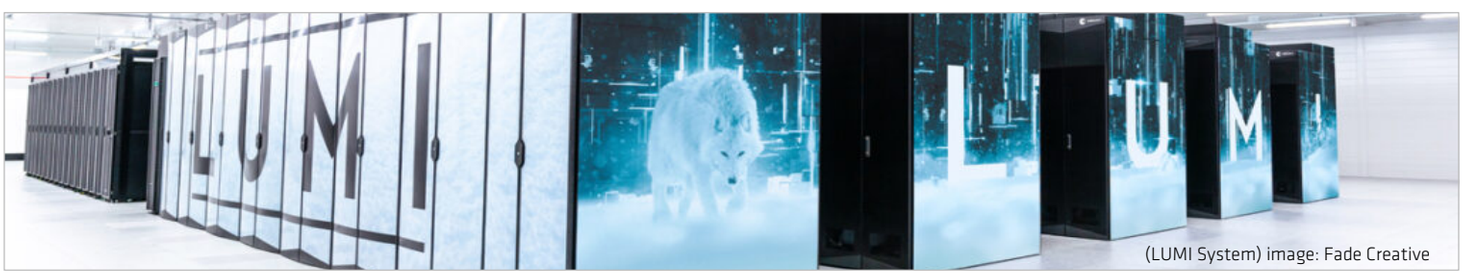
"The time taken to put BLOOM through about 40 billion tokens, which could be characters, syllables, or words, was about two weeks on LUMI," says Pyysalo. "It's theoretically possible to run a small cluster for a couple of years and get the same result, but it will be largely irrelevant by the time you publish it. We scaled to 192 nodes and 1,536 GCDs for the 176 billion-parameter model with 40 billion tokens. We're currently up to 512 nodes on LUMI, so that's 4,096 GCDs."

Towards LLMs for All European Languages

Pyysalo is now looking to leverage this scalability for the future of his LLM program. "TurkuNLP is one of the 10 cooperating university research labs," he says. "We're part of the EU-funded High Performance Language Technologies project, a three-year endeavor now just past its first six months. What we did for Finnish was a test run towards creating foundation models for at least all official EU languages and hopefully quite a few others as well. We'll be building on the technology that we put together to start generating those language models. We'll be releasing them over the next two years and some initial ones later this year."

This will require even greater scaling, but Pyysalo expects LUMI to meet the challenge. "The ambition is to create the largest open model with comprehensive support for European languages," he says. "We will go beyond 10 million GPU hours. Around 1.5 million went into our previous models, so this would be an order of magnitude more ambitious. LUMI is becoming a mature platform for very large-scale AI work. In the future, we will be training for a much larger number of tokens. It's likely that it will be going into a trillion words plus."

"We hope that the models we've now built for Finnish will serve as the foundation for the next generation of Finnish artificial intelligence technology," concludes Pyysalo. With TurkuNLP's future multilingual programs, Pyysalo hopes to extend this vision to every European language, and beyond.



(LUMI System) image: Fade Creative

About LUMI

LUMI (Large Unified Modern Infrastructure), one of the EuroHPC world-class supercomputers and leading platforms for artificial intelligence, is located at CSC's data center in Kajaani, Finland. The supercomputer is hosted by the LUMI consortium including ten European countries. To learn more about LUMI visit lumi-supercomputer.eu.

About TurkuNLP

The TurkuNLP Group is a group of researchers at the University of Turku as well as the UTU graduate school (UTUGS). The focus of their research is natural language processing, language technology and digital linguistics, ranging from corpus annotation and analysis to machine learning theory and applications. For more information visit turkunlp.org.

About AMD

For more than 50 years AMD has driven innovation in high-performance computing, graphics, and visualization technologies. Billions of people, leading Fortune 500 businesses, and cutting-edge scientific research institutions around the world rely on AMD technology daily to improve how they live, work and play. AMD employees are focused on building leadership high-performance and adaptive products that push the boundaries of what is possible. For more information about how AMD is enabling today and inspiring tomorrow, visit the AMD (NASDAQ: AMD) website, blog, LinkedIn, and Twitter pages.