Wide Horizons: NVIDIA Keynote Points Way to Further AI Advances

Chief Scientist Bill Dally described research poised to take machine learning to the next level.

Author: Rick Merritt

Dramatic gains in hardware performance have spawned generative AI , and a rich pipeline of ideas for future speedups will drive machine learning to new heights, Bill Dally, NVIDIA's chief scientist and senior vice president of research, said today in a keynote.

Dally described a basket of techniques in the works — some already showing impressive results — in a talk at Hot Chips, an annual event for processor and systems architects.

"The progress in AI has been enormous, it's been enabled by hardware and it's still gated by deep learning hardware," said Dally, one of the world's foremost computer scientists and former chair of Stanford University's computer science department.

He showed, for example, how ChatGPT, the large language model (LLM) used by millions, could suggest an outline for his talk. Such capabilities owe their prescience in large part to gains from GPUs in AI inference performance over the last decade, he said.

Researchers are readying the next wave of advances. Dally described a test chip that demonstrated nearly 100 tera operations per watt on an LLM.

The experiment showed an energy-efficient way to further accelerate the transformer models used in generative AI. It applied four-bit arithmetic, one of several simplified numeric approaches that promise future gains.

Looking further out, Dally discussed ways to speed calculations and save energy using logarithmic math, an approach NVIDIA detailed in a 2021 patent.

He explored a half dozen other techniques for tailoring hardware to specific AI tasks, often by defining new data types or operations.

Dally described ways to simplify neural networks, pruning synapses and neurons in an approach called structural sparsity, first adopted in NVIDIA A100 Tensor Core GPUs .

"We're not done with sparsity," he said. "We need to do something with activations and can have greater sparsity in weights as well."

Researchers need to design hardware and software in tandem, making careful decisions on where to spend precious energy, he said. Memory and communications circuits, for instance, need to minimize data movements.

"It's a fun time to be a computer engineer because we're enabling this huge revolution in AI, and we haven't even fully realized yet how big a revolution it will be," Dally said.

In a separate talk, Kevin Deierling, NVIDIA's vice president of networking, described the unique flexibility of NVIDIA BlueField DPUs and NVIDIA Spectrum networking switches for allocating resources based on changing network traffic or user rules.

The chips' ability to dynamically shift hardware acceleration pipelines in seconds enables load balancing with maximum throughput and gives core networks a new level of adaptability. That's especially useful for defending against cybersecurity threats.

“Today with generative AI workloads and cybersecurity, everything is dynamic, things are changing constantly,” Deierling said. “So we’re moving to runtime programmability and resources we can change on the fly,”

In addition, NVIDIA and Rice University researchers are developing ways users can take advantage of the runtime flexibility using the popular P4 programming language.

A talk by Arm on its Neoverse V2 cores included an update on the performance of the NVIDIA Grace CPU Superchip , the first processor implementing them.

Tests show that, at the same power, Grace systems deliver up to 2x more throughput than current x86 servers across a variety of CPU workloads. In addition, Arm’s SystemReady Program certifies that Grace systems will run existing Arm operating systems, containers and applications with no modification.

Grace uses an ultra-fast fabric to connect 72 Arm Neoverse V2 cores in a single die, then a version of NVLink connects two of those dies in a package, delivering 900 GB/s of bandwidth. It’s the first data center CPU to use server-class LPDDR5X memory, delivering 50% more memory bandwidth at similar cost but one-eighth the power of typical server memory.

Hot Chips kicked off Aug. 27 with a full day of tutorials, including talks from NVIDIA experts on AI inference and protocols for chip-to-chip interconnects, and runs through today.

Original URL: https://blogs.nvidia.com/blog/2023/08/29/hot-chips-dally-research/