Chip Manufacturing ‘Ideal Application’ for AI, NVIDIA CEO Says

NVIDIA CEO Jensen Huang outlines role of accelerated computing and AI in address to semiconductor industry leaders at ITF World 2023.

Author: Brian Caulfield

Chip manufacturing is an “ideal application” for NVIDIA accelerated and AI computing, NVIDIA founder and CEO Jensen Huang said Tuesday.

Detailing how the latest advancements in computing are accelerating “the world’s most important industry,” Huang spoke at ITF World 2023 semiconductor conference in Antwerp, Belgium.

Huang delivered his remarks via video to a gathering of leaders from across the semiconductor, technology and communications industries.

“I am thrilled to see NVIDIA accelerated computing and AI in service of the world’s chipmaking industry,” Huang said as he detailed how advancements in accelerated computing, AI and semiconductor manufacturing intersect.

The exponential performance increase of the CPU has been the governing dynamic of the technology industry for nearly four decades, Huang said.

But over the past few years CPU design has matured, he said. The rate at which semiconductors become more powerful and efficient is slowing, even as demand for computing capability soars.

“As a result, global demand for cloud computing is causing data center power consumption to skyrocket,” Huang said.

Huang said that striving for net zero while supporting the “invaluable benefits” of more computing power requires a new approach.

The challenge is a natural fit for NVIDIA, which pioneered accelerated computing, coupling the parallel processing capabilities of GPUs with CPUs.

This acceleration, in turn, sparked the AI revolution. A decade ago, deep learning researchers such as Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton discovered that GPUs could be cost-effective supercomputers.

Since then, NVIDIA reinvented its computing stack for deep learning, opening up “multi trillion-dollar opportunities in robotics, autonomous vehicles and manufacturing,” Huang said.

By offloading and accelerating compute-intensive algorithms, NVIDIA routinely speeds up applications by 10-100x while reducing power and cost by an order of magnitude, Huang explained.

Together, AI and accelerated computing are transforming the technology industry. “We are experiencing two simultaneous platform transitions — accelerated computing and generative AI,” Huang said.

Huang explained that advanced chip manufacturing requires over 1,000 steps, producing features the size of a biomolecule. Each step must be nearly perfect to yield functional output.

“Sophisticated computational sciences are performed at every stage to compute the features to be patterned and to do defect detection for in-line process control,” Huang said. “Chip manufacturing is an ideal application for NVIDIA accelerated and AI computing.”

Huang outlined several examples of how NVIDIA GPUs are becoming increasingly integral to chip manufacturing.

Companies like IMS Nanofabrication and NuFlare build mask writers — machines that create photomasks, stencils that transfer patterns onto wafers — using electron beams. D2S builds multi-rack computing appliances for mask writers. NVIDIA GPUs accelerate the computationally demanding tasks of pattern rendering and mask process correction for these mask writers.

Semiconductor manufacturer TSMC and equipment providers KLA and Lasertech use extreme ultraviolet light, known as EUV, and deep ultraviolet light, or DUV, for mask inspection. NVIDIA GPUs play a crucial role here, too, in processing classical physics modeling and deep learning to generate synthetic reference images and detect defects.

KLA, Applied Materials, and Hitachi High-Tech use NVIDIA GPUs in their e-beam and optical wafer inspection and review systems.

And in March, NVIDIA announced that it is working with TSMC, ASML and Synopsys to accelerate computational lithography.

Computational lithography simulates Maxwell’s equations of light behavior passing through optics and interacting with photoresists, Huang explained.

Computational lithography is the largest computational workload in chip design and manufacturing, consuming tens of billions of CPU hours annually. Massive data centers run 24/7 to create reticles for new chips.

Introduced in March, NVIDIA cuLitho is a software library with optimized tools and algorithms for GPU-accelerated computational lithography.

“We have already accelerated the processing by 50 times,” Huang said. “Tens of thousands of CPU servers can be replaced by a few hundred NVIDIA DGX systems, reducing power and cost by an order of magnitude.”

The savings will reduce carbon emissions or enable new algorithms to push beyond 2 nanometers, Huang said.

What’s the next wave of AI? Huang described a new kind of AI — “embodied AI,” or intelligent systems that can understand, reason about and interact with the physical world.

He said examples include robotics, autonomous vehicles and even chatbots that are smarter because they understand the physical world.

Huang offered his audience a look at NVIDIA VIMA, a multimodal embodied AI. VIMA, Huang said, can perform tasks from visual text prompts, such as “rearranging objects to match this scene.”

It can learn concepts and act accordingly, such as “This is a widget,” “That’s a thing” and then “Put this widget in that thing.” It can also learn from demonstrations and stay within specified boundaries, Huang said.

VIMA runs on NVIDIA AI, and its digital twin runs in NVIDIA Omniverse , a 3D development and simulation platform. Huang said that physics-informed AI could learn to emulate physics and make predictions that obey physical laws.

Researchers are building systems that mesh information from real and virtual worlds on a vast scale.

NVIDIA is building a digital twin of our planet, called Earth-2 , which will first predict the weather, then long-range weather, and eventually climate. NVIDIA’s Earth-2 team has created FourCastNet, a physics-AI model that emulates global weather patterns 50-100,000x faster.

FourCastNet runs on NVIDIA AI, and the Earth-2 digital twin is built in NVIDIA Omniverse.

Such systems promise to address the greatest challenge of our time, such as the need for cheap, clean energy.

For example, researchers at the U.K.’s Atomic Energy Authority and the University of Manchester are creating a digital twin of their fusion reactor, using physics-AI to emulate plasma physics and robotics to control the reactions and sustain the burning plasma.

Huang said scientists could explore hypotheses by testing them in the digital twin before activating the physical reactor, improving energy yield, predictive maintenance and reducing downtime. “The reactor plasma physics-AI runs on NVIDIA AI, and its digital twin runs in NVIDIA Omniverse,“ Huang said.

Such systems hold promise for further advancements in the semiconductor industry. “I look forward to physics-AI, robotics and Omniverse-based digital twins helping to advance the future of chip manufacturing,” Huang said.

Original URL: https://blogs.nvidia.com/blog/2023/05/16/itf-world-2023/