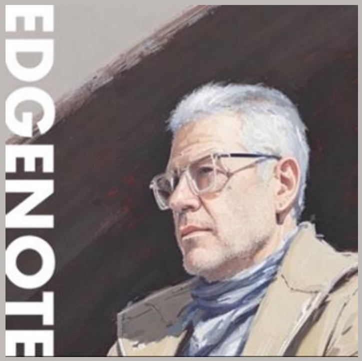

NVIDIA Takes It to The Edge

Pete Bernard – EDGECELSIOR

KEY THEME – The NVIDIA GTC Conference was one for the ages…

The NVIDIA GTC (GPU Technology Conference) for 2024 has ended, the first one in-person in five years, at the epicenter of the semiconductors in San Jose, CA. Since 2009, this conference has been focused on accelerated computing and the use NVIDIA GPU in AI, and in prior years, crypto mining, and gaming.

The Cambrian explosion in generative AI was clearly the focus of this year's event, which reportedly attracted over 20,000 people and was highlighted by CEO **Jensen Huang's keynote** in the SAP center, surrounded by a packed house and as of this writing has over 10m views on YouTube.

MONSTER CHIPS and DEVELOPER LEVERAGE

Key announcements included:

- The latest NVIDIA GPU – Blackwell, and associated upgrades to their NVLINK interconnect capabilities as well as the new BP200 systems will populate NVIDIA designed racks and cooling systems in the biggest data centers from the usual hyper scalars. The Blackwell is a 208 billion transistor beast consuming 1000w of power. You will NOT find this configuration in edge equipment any time soon.

- NVIDIA also announced the NIMS (**N**VIDIA **I**nference **Mi**croservice**s)**. NIMs are cloud-native microservices designed for deploying large language models (LLMs) and other AI models and contain APIs, code for specific functions, and optimized for NVIDIA GPUs. It's a vehicle for NVIDIA to take some of the friction out of developing and deploying services, chatbots and other code that leverages their hardware platform. It's subscription model that will generate recurring revenue for NVIDIA from their 2m+ developers and poses an alternative route to market than using microservices and offerings from AWS, Azure and other platform providers.

Humanoid Robot from Enchanted Tools using twin Jetson Orin platforms

GENERATIVE AI IS ACCELRRATING THE EDGE

The key question is what was unveiled and discussed for edge computing and how these cloud-focused investments will trickle down and accelerate the edge? There were several areas here to dig into:

- **Omniverse** is becoming a rich simulation and a digital twin environment for cloud-to-edge AI solutions. NVIDIA has been investing in Omniverse for years, which is now hosted in Azure, and they have avoided the "Metaverse Curse" by focusing on commercial value propositions for digital twin-based telemetry and now training AI models on "real world" conditions using detailed synthetic Omniverse environments. Although creating Omniverse environments for customers can be extensive undertaking, it will be more and more critical for simulating end to end solutions that impact real world environments. Seemantini Godbole, CDO of Lowe's, discussed several scenarios that they are deploying after first simulating them in their Omniverse instance, including shelf replenishment automation and enhancing plan-o-grams. One of the key aspects of Omniverse is training AI models for robotics platforms, which leads to the next key announcement – Gr00t.

- **Gr00t** is NVIDIA's ambitious effort to radically improve robot functionality by introducing generative AI principles into robot learning and execution. Gr00t is a new foundational AI model that multi-modal and enables training through observation, combined with OSMO, their workflow orchestrator for AI models and learnings. Robotic platforms, including humanoid robots, will be able to be trained by observing humans doing the same task. This kind of training can also be simulated in the Omniverse to provide a rich set of synthetic observable behaviors and scenarios for hours – without tiring out /endangering actual humans in real commercial environments. Companies like Agility

Robotics, Apptronik, Fourier Intelligence, and Unitree Robotics are adopting Gr00t and the show had a strong element of humanoid robotics – which was a full 30 minutes of Jensen’s keynote presentation.

The semiconductor innovations at the **Blackwell** level will eventually trickle down into next gen version of more edge-appropriate platforms, although the Jetson Orin platform was in full force in the Exhibit Hall. Adding NIMs to the extensive CUDA platform, Omniverse, and leveraging millions of developers should unlock innovation for more holistic cloud to edge scenarios, and the talks and exhibits demos focused on generative AI models running on Orin edge platforms, working in concert with TinyML and other low power techniques for edge input analysis and action.

Although NVIDIA has their roots in data center level semiconductors and systems, the GTC show highlighted how generative AI is accelerating the edge as we start to think more holistically about how AI will be applied across virtually every industry vertical.

KEY NVIDIA EDGE ANNOUNCEMENTS:

- *PRESS RELEASE* “NVIDIA Blackwell Platform Arrives to Power a New Era of Computing” (**link**)
- *PRESS RELEASE* “NVIDIA Launches Generative AI Microservices for Developers to Create and Deploy Generative AI Copilots Across NVIDIA CUDA GPU Installed Base“, March 18, 2024 (**link**)
- *PRESS RELEASE* “NVIDIA Announces Omniverse Cloud APIs to Power Wave of Industrial Digital Twin Software Tools“, March 18, 2024 (**link**)
- *PRESS RELEASE* “NVIDIA DRIVE Powers Next Generation of Transportation — From Cars and Trucks to Robotaxis and Autonomous Delivery Vehicles“, March 18, 2024 (**link**)
- *PRESS RELEASE* “NVIDIA Announces Project GR00T Foundation Model for Humanoid Robots and Major Isaac Robotics Platform Update“, March 18, 2024 (**link**)