NAND RESEARCH

RESEARCH NOTE

NVIDIA & AWS'S BROAD AI-FOCUSED COLLABORATION

STEVE MCDOWELL, CHIEF ANALYST

MARCH 20, 2024

CONTEXT

At the Nvidia GTC event in San Jose, Nvidia and Amazon Web Services made a series of wide-ranging announcements that showed a broad and strategic collaboration to accelerate global AI innovation and infrastructure capabilities.

The joint announcements included the introduction of NVIDIA Grace Blackwell GPU-based Amazon EC2 instances, NVIDIA DGX Cloud integration, and, most critically, a pivotal collaboration called Project Ceiba.

PROJECT CEIBA

Project Ceiba is a groundbreaking initiative between Amazon Web Services (AWS) and NVIDIA aimed at constructing one of the world's fastest AI supercomputers. This ambitious project leverages AWS's advanced computing capabilities and NVIDIA's Grace Blackwell GPU platform's cutting-edge technology to push the boundaries of AI research and development.

OBJECTIVES

- **Advance AI Research**: Project Ceiba's primary goal is to significantly propel the field of artificial intelligence forward. By providing unprecedented computational power, it aims to enable breakthroughs in large language models, digital biology, robotics, autonomous vehicles, climate prediction, and more.
- **Support NVIDIA's AI R&D**: While serving as a powerful global resource for AI research, Project Ceiba focuses on supporting NVIDIA's own R&D efforts in generative AI and other domains. This collaboration underscores the strategic partnership between AWS and NVIDIA in driving AI innovation.

INFRASTRUCTURE AND TECHNOLOGY

- **GB200 Grace Blackwell Superchips**: At the core of Project Ceiba are the NVIDIA GB200 Grace Blackwell Superchips. Each system will feature these Superchips, which combine NVIDIA's advanced GPUs and CPUs interconnected by the fifth-generation NVLink technology.
- **Massive Scale**: Project Ceiba is being built using 20,736 GB200 Superchips, capable of delivering 414 exaflops of AI computational power. This scale is designed to handle the most demanding AI workloads and datasets, providing an infrastructure to support the development of multi-trillion parameter models.
- **EFA Networking**: The project utilizes AWS's Elastic Fabric Adapter (EFA) networking to provide low-latency, high-bandwidth connectivity between the Superchips. This networking capability is crucial for distributed training and inference tasks, allowing for efficient scaling of AI models across the entire supercomputer.
- **AWS Infrastructure**: Hosted exclusively on AWS, Project Ceiba benefits from the cloud provider's robust and secure infrastructure, including its advanced virtualization with the Nitro System, hyper-scale clustering with Amazon EC2 UltraClusters, and comprehensive security features.

IMPACT AND APPLICATIONS

Project Ceiba is expected to have a profound impact on a variety of fields. By enabling more complex and larger-scale AI models than ever before, it opens new possibilities in:

- **Healthcare and Life Sciences**: Accelerating drug discovery, understanding complex biological processes, and improving patient care through advanced diagnostics and personalized medicine.
- **Environmental Science**: Enhancing climate models and predictions to better understand and mitigate the effects of climate change.
- **Autonomous Systems**: Advancing the development of self-driving vehicles and robotics, making them safer and more efficient.
- **AI-Driven Innovation**: Fueling a wide range of AI innovations across industries, from entertainment and content creation to finance and cybersecurity.

Project Ceiba exemplifies the deep strategic partnership between AWS and NVIDIA, combining AWS's cloud infrastructure prowess with NVIDIA's AI and GPU technology leadership. It showcases how collaboration between tech giants can lead to advancements that benefit the broader research community.

NVIDIA BLACKWELL ON AWS EC2

The new accelerated instances on AWS, powered by NVIDIA's Grace Blackwell GPU platform, offer a significant leap in cloud computing capabilities, particularly for developing and deploying LLMs and generative AI applications.

Here's an overview of the new instances:

1. **Hardware Configuration**: The instances are equipped with NVIDIA GB200 Grace Blackwell Superchips, which combine NVIDIA's latest Blackwell GPUs and Grace CPUs. Each GB200 NVL72 instance includes 72 Blackwell GPUs and 36 Grace CPUs, interconnected by the fifth-generation NVIDIA NVLink technology. This advanced interconnect allows for high-speed, efficient data transfer between the CPUs and GPUs, enabling unprecedented computational performance.

2. **Performance Capabilities**: The GB200 Grace Blackwell Superchips deliver massive computational power, making them ideal for training and running inference on multi-trillion parameter LLMs. The instances can scale to thousands of GB200 Superchips, providing customers with the ability to execute complex AI workloads at an extraordinary scale. This setup will significantly speed up inference workloads for resource-intensive AI models.

3. **Security Features**: Enhanced security is a critical component of the Blackwell-based instances. They integrate AWS's Nitro System, EFA encryption, and AWS Key Management Service (KMS) with Blackwell's encryption capabilities, offering end-to-end control and security for training data and model weights. This ensures a secure environment for AI applications, protecting against unauthorized access and tampering with model integrity.

4. **Networking and Virtualization**: The instances leverage AWS's Elastic Fabric Adapter (EFA) for petabit scale networking and the Nitro System for advanced virtualization capabilities. EFA provides low-latency, high-throughput networking, essential for distributed training and inference across multiple instances. The Nitro System enhances performance and security by offloading certain functions from the host CPU/GPU to specialized AWS hardware.

5. **Integration with AWS Services**: The Blackwell-based instances are designed to work seamlessly with AWS services such as Amazon EC2 UltraClusters for hyper-scale clustering, allowing for the efficient deployment and management of large-scale AI workloads. Additionally, the integration with Amazon SageMaker and NVIDIA NIM inference microservices enables customers to optimize the price-performance of their foundation models running on GPUs.

6. **Use Cases**: These instances are particularly suited for cutting-edge AI research and development, including the training of LLMs with over a trillion parameters, generative AI model development, digital biology, robotics, autonomous vehicles, and more. They are engineered to meet the needs of enterprises and research institutions requiring the highest levels of computational power and security for their AI projects.

7. **Project Ceiba**: A notable application of the Blackwell-based instances is their deployment in Project Ceiba, a collaboration between NVIDIA and AWS to build one of the world's fastest AI supercomputers. This project highlights the instances' capabilities in handling extraordinarily large and complex AI computations, further advancing AI research and development.

INTEGRATION WITH BLACKWELL ENCRYPTION

The integration of AWS Nitro System, Elastic Fabric Adapter (EFA) encryption, and AWS Key Management Service (KMS) with Blackwell encryption in the new NVIDIA Grace Blackwell GPU-based instances on AWS brings a comprehensive, multi-layered security approach to AI applications.

This integration protects data and model integrity across the entire AI workflow, from data ingestion and model training to inference and deployment.

Here's how each component contributes to the overall security architecture:

1. **AWS Nitro System**: The Nitro System is a collection of building blocks that provide the underlying infrastructure for AWS instances. It includes a suite of dedicated hardware and software innovations that eliminate the need for traditional hypervisors, thereby reducing the attack surface. The Nitro System enhances security by offloading I/O, networking, and storage tasks to dedicated hardware, freeing the CPU to run workloads with higher performance and efficiency. For AI workloads, sensitive operations and data processing can be isolated, reducing the risk of external interference or data leakage.
2. **Elastic Fabric Adapter (EFA) Encryption**: EFA is AWS's high-performance network interface optimized for High-Performance Computing (HPC) and machine learning applications. It supports low-latency, high-bandwidth communication between instances, which is crucial for distributed training of large models. The integration of encryption within EFA ensures that data in transit across the network is secured, preventing unauthorized interception and ensuring the confidentiality and integrity of the data as it moves between instances in a cluster.
3. **AWS Key Management Service (KMS)**: KMS is a managed service that makes it easy for customers to create and control the encryption keys used to encrypt their data. It provides strong security and auditing capabilities, allowing users to securely manage keys and perform cryptographic operations. By integrating KMS with Blackwell-based instances, customers can manage the keys that encrypt their training data and model weights, ensuring that only authorized users and processes can access or use them.
4. **Blackwell Encryption**: The NVIDIA Grace Blackwell Superchips have built-in encryption capabilities that secure the data directly on the chip. This includes physical encryption of the NVLink connections between GPUs, ensuring secure data transfer between components within the same instance.

Additionally, the connection from the Grace CPU to the Blackwell GPU is encrypted, safeguarding the data as it moves from the CPU to the GPU for processing. This layer of encryption at the chip level complements the overall security posture by protecting the data at rest, in use, and in transit.

Together, these technologies create a secure environment for AI workloads on AWS, providing end-to-end control and security for customers' data and models. The Nitro System ensures secure and efficient operation of the instances themselves. EFA encryption secures data as it moves across the network, KMS provides secure key management for data encryption, and Blackwell encryption secures the data at the hardware level.

This comprehensive security approach enables customers to confidently deploy AI applications, knowing their data and intellectual property are protected throughout the entire AI lifecycle.

AMAZON SAGEMAKER INTEGRATION WITH NVIDIA NIM

The integration of Amazon SageMaker with NVIDIA NIM (NVIDIA Inference Server, previously known as Triton Inference Server) represents a significant advancement in the deployment and management of AI models, particularly in optimizing the performance and cost-effectiveness of running foundation models on GPUs.

AMAZON SAGEMAKER OVERVIEW

Amazon SageMaker is a fully managed service that allows every developer and data scientist to build, train, and deploy machine learning (ML) models quickly. SageMaker takes away the heavy lifting of ML model development by providing tools and features that accelerate the entire process from idea to production.

NVIDIA NIM OVERVIEW

NVIDIA NIM (formerly Triton Inference Server) is an open-source platform designed to deploy AI models at scale. It supports multiple frameworks (such as TensorFlow, PyTorch, ONNX Runtime, and more) for running inference operations. NIM optimizes the utilization of NVIDIA GPUs and CPUs in production environments, ensuring efficient resource use and high throughput for inference tasks.

INTEGRATION BENEFITS AND FEATURES

- **Optimized Performance**: The integration allows customers to leverage the computational power of NVIDIA GPUs on Amazon SageMaker, enabling high-performance inference for complex AI models. NIM's support for multi-framework model deployment ensures developers can deploy models optimized for specific hardware, reducing latency and increasing throughput.

- **Cost Efficiency**: By utilizing NIM's advanced model serving capabilities, such as dynamic batching and model versioning, within SageMaker's scalable environment, organizations can achieve more efficient inference operations. This translates to lower operational costs by maximizing resource utilization and minimizing idle times.
- **Seamless Deployment**: SageMaker's integration with NIM simplifies the deployment process of AI models. Thanks to NIM's multi-framework support, developers can quickly deploy models trained in various frameworks without worrying about compatibility issues. This streamlined process helps reduce the time to market for AI solutions.
- **Scalability and Flexibility**: SageMaker provides the infrastructure and tools to scale model inference to meet demand, while NIM offers the flexibility to serve models optimized for different hardware. This combination allows for scalable, flexible deployment options that adapt to varying workload requirements.
- **Enhanced Security and Management**: SageMaker offers built-in security features that protect data and models at rest and in transit. When combined with NIM's capability to manage multiple models and versions efficiently, organizations benefit from a secure and manageable environment for their AI applications.

USE CASES

The integration is particularly beneficial for deploying foundation models and generative AI applications that require GPUs' processing power. It caters to a wide range of industries, including, but not limited to, healthcare, financial services, automotive, and entertainment, where the demand for efficient and effective AI inference is high.

The Amazon SageMaker and NVIDIA NIM integration provides a powerful, flexible, and cost-efficient solution for deploying AI models at scale. It harnesses the best of AWS's cloud infrastructure and NVIDIA's GPU optimization technologies. This collaboration enables customers to accelerate the deployment of AI applications while maintaining high performance and reducing operational costs.

ANALYSIS

The scarcity of Nvidia’s flagship Ampere and Hopper generation GPUs has impacted the cloud landscape, with customers moving AI workloads to where GPUs are most available. Nvidia is not a neutral party in this, favoring partners willing to embrace its platform-focused strategy. Amazon was the last major public cloud provider, in November 2023, to embrace DGX Cloud.

The introduction of Nvidia Grace Blackwell Superchips is a game-changer for the industry, one that keeps Nvidia at the forefront of high-performance AI training. Bringing the parts to AWS is a natural move, and Nvidia made similar announcements with Oracle, Microsoft, and Google.

Cloud-based AI democratizes access to unprecedented computational power, enabling businesses and researchers to tackle more complex problems and innovate faster. Having Nvidia's latest-generation technology available across every major DSP is good for the entire industry.

Project Ceiba is where AWS steps away from the pack, showing a surprisingly deep relationship with Nvidia. The new collaboration is a bold step towards building one of the world's fastest AI supercomputers, exclusively on AWS infrastructure.

The project is not just about raw computational power; it shows a strategic vision shared by AWS and Nvidia in pushing the boundaries of AI research and development. The potential applications of Project Ceiba in areas such as healthcare, autonomous vehicles, and climate modeling are vast and could lead to breakthroughs that significantly impact society.

Overall, the joint AWS and Nvidia announcements show a deep commitment by both organizations to democratize AI technology by making it accessible and secure for a wide range of users. The collaboration is likely to set new standards in the industry, driving innovation and opening up new possibilities for AI applications across various sectors, benefiting enterprises across nearly every industry.

The information contained within this document was believed by NAND Research to be reliable and is provided for informational purposes only. The content may contain technical inaccuracies, omissions, or typographical errors. This document reflects the opinions of NAND Research, which is subject to change. NAND Research does not warranty or otherwise guarantee the accuracy of the information contained within.

NAND Research is a technology-focused industry analyst firm providing research, customer content, market and competitive intelligence, and custom deliverables to technology vendors, investors, and end-customer IT organizations.

Contact NAND Research via email at info@nand-research.com or visit our website at nand-research.com.