NAND RESEARCH

RESEARCH NOTE

AMAZON GRAVITON4 & TRAINIUM2

STEVE MCDOWELL, CHIEF ANALYST
DECEMBER 7, 2023

CONTEXT

Amazon introduced generational updates to its Graviton and Tranium custom silicon solutions at its recent re:Invent conference in Las Vegas.

This Research Note delves into the intricacies and potential impacts of these latest innovations from AWS.

GRAVITON4

Graviton is a series of custom-designed processors based on the Arm architecture developed by AWS specifically for its cloud computing services.

The key aspects of the Graviton processors include:

- **ARM Architecture:** Graviton processors use the ARM architecture, which differs from the x86 architecture used by most traditional server CPUs. Arm-based processors often deliver a better performance/watt/dollar story than competing x86 processors from AMD and Intel; Arm-based parts are ideal for general-purpose workloads and are beginning to scale to high-performance ones.

- **Optimized for Cloud Performance:** Graviton processors are designed to maximize the performance of cloud-based workloads. They are tailored to run various applications hosted on AWS, including those that require high computational efficiency.

- **Energy Efficiency:** A key advantage of Graviton processors is their energy efficiency, which is crucial in large data centers where power consumption can be a significant operational cost.

The new Graviton4 is a significant step forward in AWS's chip design. The new processor offers various advancements that enhance the performance and efficiency of computing tasks on the Amazon Elastic Compute Cloud (EC2).

Some of the key attributes and features of Graviton4 include:

1. **Improved Compute Performance:** Amazon tells us that the Graviton4 provides up to 30% better compute performance than its predecessor, the Graviton3.

2. **Increased Core Count:** The new CPU delivers 50% more cores than Graviton3. More cores allow it to handle more tasks simultaneously, improving its ability to manage multi-threaded applications and high-performance computing tasks.

3. **Enhanced Memory Bandwidth:** Graviton4 features 75% more memory bandwidth, allowing faster data transfer rates to and from memory.

4. **Energy Efficiency:** One of the hallmarks of Graviton4 is its focus on energy efficiency.

5. **Broad Application Range:** Graviton4 suits various applications, including databases, analytics, web servers, batch processing, ad serving, application servers, and microservices. Its versatility makes it a compelling choice for diverse computing needs.

6. **Security Enhancements:** The chip includes improvements in security, such as full encryption of all high-speed physical hardware interfaces. This feature adds an extra layer of security against physical attacks and unauthorized data access.

7. **Support for AWS Services:** Graviton4 is supported by many AWS-managed services like Amazon Aurora, Amazon ElastiCache, Amazon EMR, Amazon MemoryDB, Amazon OpenSearch, Amazon RDS, AWS Fargate, and AWS Lambda. This support extends the chip's benefits to a broader range of AWS cloud services.

8. **Use in EC2 Instances:** Graviton4 will be available in memory-optimized Amazon EC2 R8g instances. These instances offer larger sizes with more virtual CPUs and memory, catering to demanding workloads that require high performance and scalability.

The new Graviton processor is based on the Arm “Demeter” Neoverse V2 core, part of the Armv9 architecture. This is similar to NVIDIA’s “Grace” CPU. The V2 core in Graviton4 shows a 13% improvement in instructions per clock over the “Zeus” V1 core used in Graviton3 and Graviton3E.

While the company didn’t say explicitly, AWS likely shifted from a 5-nanometer process to a denser 4-nanometer process for Graviton4, aligning with Nvidia's Grace CPU and Hopper GH100 GPUs production.

Graviton4 features 96 V2 cores, a 50% increase over its predecessors, and includes twelve DDR5 memory controllers with a speed boost to 5.6 GHz.

This configuration results in a 75% higher memory bandwidth per socket than Graviton3 and Graviton3E.

AWS is positioning Graviton4 as a powerful, efficient, and versatile processor that caters to a range of cloud computing needs, highlighting AWS's commitment to continuous innovation in cloud infrastructure technology.

TRAINIUM2

Trainium is an AI accelerator designed to enhance the performance and efficiency of training machine learning models, particularly deep learning models. These accelerators are used for workloads involving large-scale neural networks foundational to many modern AI applications.

Some of the key aspects of Trainium include:

- **Purpose-Built for ML Training:** Trainium chips are tailored to enhance the performance and efficiency of training machine learning models, particularly deep learning models.
- **High Performance:** Trainium chips offer high-performance computing capabilities to handle the demanding requirements of ML training tasks, which often involve processing large datasets and complex algorithms.
- **Efficiency:** Alongside performance, Trainium chips also emphasize energy efficiency, which is crucial given the high-power demands of ML training operations.

The new Trainium2 is specifically engineered to accelerate the training of machine learning models, mainly focusing on large-scale models such as foundation models (FMs) and large language models (LLMs).

The key features and characteristics of Amazon’s new Trainium2 include:

1. **Enhanced Training Performance:** Trainium2 delivers up to four times faster training performance than first-generation Trainium chips.
2. **Large-Scale Deployment Capability:** Trainium2 can be deployed in Amazon EC2 UltraClusters, which can scale up to 100,000 Trainium2 chips. This massive scaling capability enables the training of exceptionally large ML models.
3. **Energy Efficiency:** A significant focus of Trainium2 is on improving energy efficiency. The new part offers up to twice the energy efficiency (performance per watt) compared to the first-generation Trainium chips. This improvement is significant given the high energy demands of large-scale ML model training.
4. **Support for Foundation Models and Large Language Models:** Trainium2 is especially suited for training FMs and LLMs, which are foundational to today's emerging generative AI applications. These applications include creating diverse content such as text, audio, images, video, and software code.

5. **High Memory Capacity:** The chip is designed with a high memory capacity, three times more than the first-generation Trainium chips. This increased memory allows for handling larger datasets and more complex ML models.
6. **Integration with AWS Infrastructure:** Trainium2 chips will be available in Amazon EC2 Trn2 instances, each containing 16 Trainium chips. The integration with AWS infrastructure, including AWS Elastic Fabric Adapter (EFA) and petabit-scale networking, ensures high performance and reliability.
7. **Supercomputer-Class Performance:** The scale and performance of EC2 UltraClusters equipped with Trainium2 chips are likened to supercomputer-class performance, enabling the rapid training of extremely large models.
8. **Cost-Effective Training:** By providing higher performance and efficiency, Trainium2 aims to lower the cost of training large-scale ML models, making it more accessible to a broader range of AWS customers.

AWS Trainium2 represents a significant leap forward in ML training technology for AWS, offering increased speed, scalability, and efficiency for training the most advanced and large-scale machine learning models in the cloud.

ANALYSIS

Amazon's introduction of Graviton4 and Trainium2 continues the on-going shift in how cloud-based computing and AI model training are approached. The new chips are poised to transform various applications, from basic cloud services to the most advanced AI-driven tasks.

Graviton and Trainium allow Amazon to deliver a scalable range of compute options to its customers with better cost and efficiency than with more traditional options from Intel, AMD, and NVIDIA. This is a nice choice for IT buyers, allowing Amazon to reclaim some of the margin that would otherwise flow to other semiconductor suppliers.

By offering increased speed, efficiency, and scalability, these chips demonstrate AWS's commitment to pushing the boundaries of what's possible in cloud computing and AI.

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 warrant 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@nandresearch.com or visit our website at nandresearch.com.