5 เหตุผลที่ AMD INSTINCT™ MI210 ACCELERATORS ช่วยทำให้ระบบการประมวลผลของการวิจัยทางด้านวิทยาศาสตร์ และการค้นคว้าด้านนวัตกรรม รวดเร็วขึ้นกว่าเดิม

ข้อมูลโดยสรุป

ทำอย่างไรถึงจะทำให้นักวิจัยและองค์กรต่างๆ ที่ทำงานบนแอปพลิเคชั่น HPC และ AI ได้ผลลัพธ์ที่รวดเร็ว และมีประสิทธิภาพอย่างน่าพึงพอใจ? AMD Instinct MI210 Accelerators ไม่เพียงแค่ช่วยคุณได้ แต่จะทำให้คุณถึง เพราะเป็นผลิตภัณฑ์ที่ถูกสร้างขึ้นบนเทคโนโลยีการประมวลผลในระดับ Exascale ที่ช่วยยกระดับประสิทธิภาพการทำงานให้กับผู้ใช้งานทางด้าน HPC ไปอีกขั้น ทำให้ลดระยะเวลาในการประมวลผลสำหรับข้อมูลเชิงลึกและข้อมูลใหม่ ๆ ได้เป็นอย่างดี นอกจากนี้ยังมีระบบนิเวศทางเทคโนโลยีที่หลากหลายเพื่อรองรับการใช้งานซอฟต์แวร์ระบบเปิดได้มากกว่าเดิม พร้อมด้วยคุณสมบัติในการรองรับเวิร์คโหลดได้จำนวนมหาศาล ในการใช้งานที่หลากหลาย และช่วยลดขั้นตอนการทำโปรแกรมการทำงานระหว่าง CPU และ GPU ได้เป็นอย่างดี AMD Instinct MI210 Accelerators จะช่วยหลุดพ้นจากทุกพันธนาการและข้อจำกัด เพื่อสร้างคุณค่าที่สูงขึ้นอีกระดับให้กับองค์กรของคุณ

1 มอบประสิทธิภาพการประมวลผลสูงสุดให้กับเวิร์คโหลดของคุณ

สร้างประสบการณ์ที่น่าพึงพอใจยิ่งกว่า ด้วยประสิทธิภาพที่สูงกว่าเดิมและช่วยให้คุณทำงานเสร็จเร็วขึ้นอย่างน่าเหลือเชื่อ

AMD Instinct™ MI210 accelerators ผลิตภัณฑ์ที่ทำให้ความเป็นผู้นำด้านการประมวลผลของ AMD โดดเด่นขึ้นอีกระดับ ด้วยความรวดเร็วแม่นยำกว่าเดิมด้วยประสิทธิภาพการประมวลผลแบบ FP64 สำหรับเวิร์คโหลดด้าน HPC และ AI ในดาต้าเซ็นเตอร์ พร้อมด้วยการสนับสนุนจากสถาปัตยกรรมทางเทคโนโลยีอย่าง 2nd Gen AMD CDNA (compute DNA) และ AMD Instinct MI200 series GPUs ที่ช่วยมอบระบบประมวลผลที่ทรงพลังที่สุดในอุตสาหกรรมให้กับงานด้าน HPC.

2 ประสิทธิภาพที่เหนือขึ้นไปอีกระดับ

ให้การลงทุนด้าน HPC และ AI ของคุณประสบความสำเร็จมากขึ้น

CDNA2 ของ AMD เป็นสถาปัตยกรรมทางเทคโนโลยีรุ่นใหม่ที่ถูกออกแบบมาในขนาด 6 นาโนเมตร (6nm) ที่ติดตั้งบน AMD Instinct MI210 (PCIe, 64GB, 300W) ทำให้ประสิทธิภาพการประมวลผลแบบ FP64 เพิ่มขึ้นถึง 2.3 เท่าต่อวัตต์ เมื่อเทียบกับ Nvidia A100 (PCIe, 80 GB, 300W) จึงเหมาะสำหรับการใช้งานแอปพลิเคชั่นที่ต้องการประสิทธิภาพการประมวลผลที่ดีที่สุด

3 เชื่อมั่นได้กับแผนงานการดำเนินการที่ต่อเนื่องของ AMD

เพื่อให้คุณมั่นใจได้ว่าเราพร้อมสำหรับรองรับงานวิจัยทางวิทยาศาสตร์และนวัตกรรมในอนาคตของคุณ

การที่ AMD นำเทคโนโลยีระดับ Exascale มาใช้กับ AMD Instinct MI210 accelerator สำหรับตลาดกระแสหลัก และยังคงมุ่งมั่นในการเป็นผู้บุกเบิกเพื่อสร้างสรรค์สิ่งใหม่ๆให้กับตลาด GPU อย่างสม่ำเสมอ เช่นเดียวกับผลิตภัณฑ์ที่ได้คิดค้นและผลิตมาก่อนหน้านี้ ไม่ว่าจะเป็น AMD Instinct รุ่นต่างๆ หรือจะเป็น AMD EPYC processor นั้น สิ่งเหล่านี้ช่วยสะท้อนถึงแผนงานการดำเนินการระยะยาวของ AMD ที่ต้องการเสนอผลิตภัณฑ์ที่มีประสิทธิภาพดีที่สุด และมีความโดดเด่นอย่างชัดเจน เพื่อความมั่นใจและประโยชน์สูงสุดของผู้ใช้งานได้เป็นอย่างดี

4 ปรับระบบของคุณให้เหมาะสมกับสัดส่วนของงาน วัตถุประสงค์และปริมาณงานที่ต้องการ

ด้วยการใช้แพลตฟอร์มของ CPU และ GPU ที่ออกแบบมาโดยเฉพาะเพื่อการเร่งความเร็วในการประมวลผล

สถาปัตยกรรม AMD Infinity รุ่นที่ 3 ประกอบด้วยระบบการเชื่อมต่อของเทคโนโลยี AMD Infinity Fabric เพื่อการเชื่อมต่อระหว่างโปรเซสเซอร์ AMD EPCY และตัวเร่งความเร็วในการประมวลผลของ AMD Instinct MI210 ซึ่งทำให้ความสามารถในการติดตั้งโปรแกรมง่ายขึ้น เพื่อปลดล็อกให้ความสามารถในการทำงานที่สูงขึ้นและสามารถประยุกต์ใช้กับสถาปัตยกรรมที่มีประสิทธิภาพสูง

5 สิ่งสำคัญที่สุดคือ การมุ่งเน้นไปที่เป้าหมาย

ควรหลีกเลี่ยงการให้ความสำคัญกับส่วนประกอบย่อยมากเกินไป

ซอฟต์แวร์ ROCm 5 ช่วยขยายความสามารถให้กับแพลตฟอร์มซอฟต์แวร์ระบบเปิดของ AMD ซึ่งไม่เพียงเหมาะสมสำหรับแอฟพลิเคชั่นงานประเภท HPC และ AI แต่ยังช่วยเพิ่มขีดความสามารถให้กับอุปกรณ์อื่นๆ ด้วยโค้ดเบสแบบโอเพ่นซอร์ส ที่ช่วยในการพัฒนาระบบและอุปกรณ์ต่าง ๆ

Continue reading for more technical detail

ข้อมูลทางเทคนิคโดยละเอียด

#1 มอบประสิทธิภาพการประมวลผลสูงสุดให้กับเวิร์คโหลดของคุณ

- ด้วยสถาปัตยกรรมการออกแบบของ AMD CDNA เจนเนอร์เรชั่นที่ 2 ส่งผลให้การทำงานด้าน HPC ของ AMD Instinct™ MI210 accelerators ได้ผลลัพธ์ที่เหนือกว่า PCIe และ Data Center GPUs และช่วยเพิ่มประสิทธิภาพในการประมวลผลแบบ FP64 เพิ่มขึ้นถึง 2.3 เท่า ซึ่งเป็นข้อได้เปรียบที่เหนือกว่า Nvinia Ampere A100 GPUs โดยเฉพาะเมื่อมีการใช้งานแอปพลิเคชั่น HPC ในจำนวนมหาศาลและหลากหลาย
- ด้วยหน่วยความจำแบบ HBM2e ที่มีความเร็วที่ 64GB และให้แบนด์วิดท์หน่วยความจำประสิทธิภาพสูงขนาด 16 TB/s ทำให้ AMD Instinct™ MI210 accelerators มีแบนด์วิดท์สูงขึ้น 33% และมีความจุของหน่วยความจำ เพิ่มขึ้นถึง 2 เท่า เมื่อเทียบกับการประมวลผลของ AMD Instinct GPU เจนเนอร์เรชั่นก่อนหน้านี้
- สถาปัตยกรรมการออกแบบ AMD CDMA2 และ Matrix Cores ใหม่ ส่งผลให้ประสิทธิภาพการประมวลผลแบบ FP64 Matrix สูงขึ้น 3.9 เท่า เมื่อเทียบกับการประมวลแบบเดียวกันของ AMD Instinct GPU เจนเนอร์เรชั่นก่อนหน้านี้
- AMD Instinct MI210 ที่ถูกสร้างขึ้นมาบนพื้นฐานของเทคโนโลยี AMD Matrix Core ช่วยขยายขอบเขตประสิทธิภาพการคำนวนแบบ mixed-precision เพื่อให้การฝึกอบรมและการเรียนรู้ในเชิงลึกมีความรวดเร็วมากขึ้น มีการประเมิณว่าการคำนวนแบบ FP16 และ BF16 จะมีประสิทธิภาพสูงสุดในทางทฤษฎี ถึง 181 teraflops สำหรับการทำงานที่ผสานกันระหว่างแพลตฟอร์ม HPC และ AI

#2 ประสิทธิภาพที่เหนือขึ้นไปอีกระดับ

- ประสิทธิภาพการทำงานต่อ core สูงขึ้น... เมื่อพบว่า AMD Instinct MI210 ประสบความสำเร็จในการพัฒนาประสิทธิภาพของการประมวลผลแบบ FP64 ด้วยสตรีมโปรเซสเซอร์น้อยกว่า 1,034 ตัว เมื่อเทียบกับการประมวลผลของ AMD Instinct GPU เจนเนอร์เรชั่นก่อนหน้านี้
- AMD Instinct MI210 ปรับขนาดหน่วยความจำและประสิทธิภาพการทำงานอย่างเต็มที่ มีการเพิ่มความจุของหน่วยความจำและประสิทธิภาพการคำนวนแบบ FP64 เป็น 2 เท่าเมื่อเทียบกับเจนเนอร์เรชั่นก่อนหน้านี้ เพื่อให้ได้ประสิทธิภาพสูงสุดเมื่อนำไปปรับใช้กับงานด้าน HPC และ AI

#3 เชื่อมั่นได้กับแผนงานการดำเนินการที่ต่อเนื่องของ AMD

- AMD Instinct MI210 ได้รับการออกแบบโดยใช้สถาปัตยกรรมเดียวกันกับที่ใช้ในซูเปอร์คอมพิวเตอร์รุ่นใหม่ๆ รายละเอียดเพิ่มเติม คลิก AMD.com/Exascale
- นวัตกรรมเชิงเทคโนโลยีของ AMD ในด้านสถาปัตยกรรม บรรจุภัณฑ์ และการทำงานแบบบูรณาการ เป็นองค์ประกอบที่ช่วยผลักดันขอบเขตในการประมวลผลขั้นสูง เพื่อให้การทำงานร่วมกันของ CPU และ GPU มีประสิทธิภาพสูงและรวดเร็วที่สุด
- AMD Instinct MI210 เป็นผลิตภัณฑ์ที่แสดงให้เห็นถึงคุณค่าและการพัฒนาอย่างต่อเนื่องของ Instinct accelerators ซึ่งปัจจุบันเป็นเจนเนอร์เรชั่นที่ 3 แล้ว

#4 ปรับระบบของคุณให้เหมาะสมกับสัดส่วนของงาน วัตถุประสงค์และปริมาณงานที่ต้องการ

- AMD Instinct MI210 ให้การสนับสนุนที่สำคัญต่อแพลตฟอร์มการประมวลแบบรวมศูนย์ในดาต้าเซ็นเตอร์ ซึ่งสามารถรองรับปริมาณของงานจากโซลูชั่น เซิร์ฟเวอร์เดี่ยว ไปจนถึงซูเปอร์คอมพิวเตอร์ที่มีขนาดใหญ่ที่สุดในโลก และเพื่อรองรับงานที่มีวัตถุประสงค์และเวิร์คโหลดที่หลากหลาย
- เทคโนโลยี AMD Infinity Fabric เจนเนอร์เรชั่น 3 ยังนำเสนอการเชื่อมต่อและการปรับสเกล เพื่อให้สามารถเชื่อมต่อแบบ peer-to-peer (P2P) ได้ทั้งแบบ dual และ quad ด้วย GPU ที่มีดีไซน์แบบรังผึ้ง ผ่านลิงค์ AMD Infinity Fabric จำนวน 3 ลิงค์ ทำให้การเชื่อมต่อมีอัตราการส่งข้อมูล 300 GB/s (แบบ dual) และ 600 GB/s (แบบ quad) จากยอดรวมทั้งหมดของการเชื่อมต่อแบบ P2P บน I/O แบนด์วิดท์ภายใน GPU
- สามารถทดสอบประสิทธิภาพที่ AMD Accelerator Cloud (AAC) เพื่อให้คุณมั่นใจได้ว่า AMD Instinct MI210 accelerators เหมาะสมกับงานของคุณ

#4 สิ่งสำคัญที่สุดคือ การมุ่งเน้นไปที่เป้าหมาย

- AMD ROCM 5 เป็นแพลตฟอร์มด้านการพัฒนาซอฟต์แวร์แบบโอเพ่นซอร์สแรก สำหรับการประมวลผล GPU ระดับ HPC / ไฮเปอร์สเกล ซึ่งนำปรัชญาทางเลือกของระบบ UNIX พร้อมกับความเรียบง่ายแบบมินิมอล และการพัฒนาซอฟต์แวร์แบบแยกส่วนมาใช้กับการประมวลผล GPU ในระดับอุปกรณ์
- ระบบนิเวศแบบเปิดของ ROCm 5 สามารถรองรับ GPU จากผู้จำหน่ายและสถาปัตยกรรมที่แตกต่างและหลายได้เป็นอย่างดี
- ROCm ยังรองรับเฟรมเวิร์คหลักในตลาด ที่คิดค้นโดยบริษัทชั้นนำ เช่น TensorFlow, PyTorch และ ONNX-RT
- ROCm 5 ช่วยให้นักพัฒนาสามารถเร่งความเร็วในการทำงานได้อย่างต่อเนื่อง จากดาต้าเซ็นเตอร์ถึงเดสก์ท็อปที่รองรับเวิร์กสเตชันสำหรับ GPU AMD Radeon™ Pro W6800
- AMD Infinity Hub ช่วยให้นักวิจัย นักวิทยาศาสตร์ด้านข้อมูล และผู้ใช้งาน สามารถทำงานได้รวดเร็วขึ้น และมีวิธีการที่ตรงไปสู่การค้นหา ดาวน์โหลด รวมถึงการติดตั้ง optimized containers สำหรับแอปพลิเคชัน HPC และเฟรมเวิร์กการเรียนรู้ของเครื่อง ที่รองรับบน AMD Instinct MI200 series accelerators และ ROCm 5

1 World's fastest data center GPU is the AMD Instinct™ MI250X. Calculations conducted by AMD Performance Labs as of Sep 15, 2021, for the AMD Instinct MI250X (128GB HBM2e OAM module) accelerator at 1,700 MHz peak boost engine clock resulted in 95.7 TFLOPS peak theoretical double precision (FP64 matrix), 47.9 TFLOPS peak theoretical double precision (FP64), 95.7 TFLOPS peak theoretical single precision matrix (FP32 matrix), 47.9 TFLOPS peak theoretical single precision (FP32), 383.0 TFLOPS peak theoretical half precision (FP16), and 83.0 TFLOPS peak theoretical Bfloat16 format precision (BF16) floating-point performance. Calculations conducted by AMD Performance Labs as of Sep 18, 2020 for the AMD Instinct MI100 (32GB HBM2 PCIe® card) accelerator at 1,502 MHz peak boost engine clock resulted in 11.54 TFLOPS peak theoretical double precision (FP64), 46.1 TFLOPS peak theoretical single precision matrix (FP32), 23.1 TFLOPS peak theoretical single precision (FP32), 184.6 TFLOPS peak theoretical half precision (FP16) floating-point performance. Published results on the Nvidia™ Ampere A100 (80GB) GPU accelerator, boost engine clock of 1410 MHz, resulted in 19.5 TFLOPS peak double precision tensor cores (FP64 Tensor Core), 9.7 TFLOPS peak double precision (FP64). 19.5 TFLOPS peak single precision (FP32), 78 TFLOPS peak half precision (FP16), 312 TFLOPS peak half precision (FP16 Tensor Flow), 39 TFLOPS peak Bfloat 16 (BF16), 312 TFLOPS peak Bfloat16 format precision (BF16 Tensor Flow), theoretical floating-point performance. The TF32 data format is not IEEE compliant and not included in this comparison. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf, page 15, Table 1. MI200-01

2 Calculations conducted by AMD Performance Labs as of Jan 14, 2022, for the AMD Instinct MI210 (64GB HBM2e PCIe card) accelerator at 1,700 MHz peak boost engine clock resulted in 45.3 TFLOPS peak theoretical double precision (FP64 Matrix), 22.6 TFLOPS peak theoretical double precision (FP64), and 181.0 TFLOPS peak theoretical Bfloat16 format precision (BF16), floating-point performance. Calculations conducted by AMD Performance Labs as of Sep 18, 2020 for the AMD Instinct MI100 (32GB HBM2 PCIe card) accelerator at 1,502 MHz peak boost engine clock resulted in 11.54 TFLOPS peak theoretical double precision (FP64), and 184.6 TFLOPS peak theoretical half precision (FP16), floating-point performance. Published results on the Nvidia Ampere A100 (80GB) GPU accelerator, boost engine clock of 1410 MHz, resulted in 19.5 TFLOPS peak double precision tensor cores (FP64 Tensor Core), 9.7 TFLOPS peak double precision (FP64) and 39 TFLOPS peak Bfloat16 format precision (BF16), theoretical floating-point performance. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf. page 15, Table 1. MI200-41

3 Calculations conducted by AMD Performance Labs as of Feb 15, 2022, for the AMD Instinct MI210 (64GB HBM2e PCIe card) 300-watt accelerator at 1,700 MHz peak boost engine clock resulted in 45.3 TFLOPS peak theoretical double precision (FP64 Matrix), 22.6 TFLOPS peak theoretical double precision (FP64), 22.6 TFLOPS peak theoretical single precision (FP32), and 181.0 TFLOPS peak theoretical Bfloat16 format precision (BF16), floating-point performance. Calculations conducted by AMD Performance Labs as of Sep 18, 2020 for the AMD Instinct MI100 (32GB HBM2 PCIe card) accelerator at 1,502 MHz peak boost engine clock resulted in 11.54 TFLOPS peak theoretical double precision (FP64), floating-point performance. Published results on the Nvidia Ampere A100 (80GB) 300 Watt PCIe GPU accelerator, boost engine clock of 1410 MHz, resulted in 19.5 TFLOPS peak double precision tensor cores (FP64 Tensor Core), 9.7 TFLOPS peak double precision (FP64). 19.5 TFLOPS peak single precision (FP32), 78 TFLOPS peak half precision (FP16), 39 TFLOPS peak Bfloat16 format precision (BF16), theoretical floating-point performance. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/nvidia-ampere-architecture-whitepaper.pdf, page 15, Table 1. MI200-44

4 Calculations conducted by AMD Performance Labs as of Jan 27, 2022, for the AMD Instinct MI210 (64GB HBM2e) accelerator (PCIe) designed with AMD CDNA 2 architecture 6nm FinFet process technology at 1,600 MHz peak memory clock resulted in 64 GB HBM2e memory capacity and 1.6384 TFLOPS peak theoretical memory bandwidth performance. MI210 memory bus interface is 4,096 bits and memory data rate is 3.20 Gbps for total memory bandwidth of 1.6384 TB/s ((3.20 Gbps*(4,096 bits))/8). Calculations conducted by AMD Performance Labs as of Sep 18, 2020, for the AMD Instinct™ MI100 (32GB HBM2) accelerator (PCIe) designed with AMD CDNA architecture 7nm FinFet process technology at 1,502 MHz peak clock resulted in 32 GB HBM2 memory capacity and 1.2288 TFLOPS peak theoretical memory bandwidth performance. MI210 memory bus interface is 4,096 bits and memory data rate is 2.40 Gbps for total memory bandwidth of 1.2288 TB/s ((2.40 Gbps*(4,096 bits))/8). MI200-42

5 The AMD Instinct™ MI210 PCIe accelerator has 104 compute units (CUs) and 6,656 stream cores. The AMD Instinct™ MI100 accelerator has 120 compute units (CUs) and 7,680 stream cores. Calculations conducted by AMD Performance Labs as of Feb 15, 2022, for the AMD Instinct™ MI210 (64GB HBM2e PCIe card) 300 Watt accelerator at 1,700 MHz peak boost engine clock resulted in 45.3 TFLOPS peak theoretical double precision (FP64 Matrix), 22.6 TFLOPS peak theoretical double precision (FP64), floating-point performance. Calculations conducted by AMD Performance Labs as of Sep 18, 2020 for the AMD Instinct™ MI100 (32GB HBM2 PCIe card) accelerator at 1,502 MHz peak boost engine clock resulted in 11.54 TFLOPS peak theoretical double precision (FP64), floating-point performance. MI200-45

6 Calculations as of Jan 27, 2022. AMD Instinct MI210 built on AMD CDNA2 technology accelerators support PCIe Gen4 providing up to 64 GB/s peak theoretical data bandwidth from CPU to GPU per card. AMD Instinct MI210 CDNA2 technology-based accelerators include three Infinity Fabric™ links providing up to 300 GB/s peak theoretical GPU to GPU or peer-to-peer (P2P) bandwidth performance per GPU card. Combined with PCIe Gen4 support, this provides an aggregate GPU card I/O peak bandwidth of up to 364 GB/s. Dual-GPU hives: One dual-GPU hive provides up to 300 GB/s peak theoretical P2P performance.
Four-GPU hives: One four-GPU hive provide up to 600 GB/s peak theoretical P2P performance. Dual four GPU hives in a server provide up to 1.2 TB/s total peak theoretical direct P2P performance per server. AMD Infinity Fabric link technology not enabled: One four-GPU hive provide up to 256 GB/s peak theoretical P2P performance with PCIe 4.0. AMD Instinct MI100 built on AMD CDNA technology accelerators support PCIe Gen4 providing up to 64 GB/s peak theoretical transport data bandwidth from CPU to GPU per card. AMD Instinct MI100 CDNA technology-based accelerators include three Infinity Fabric links providing up to 276 GB/s peak theoretical GPU to GPU or Peer-to-Peer (P2P) bandwidth performance per GPU card. Combined with PCIe Gen4 support, this provides an aggregate GPU card I/O peak bandwidth of up to 340 GB/s. One four-GPU hive provides up to 552 GB/s peak theoretical P2P performance. Dual four-GPU hives in a server provide up to 1.1 TB/s total peak theoretical direct P2P performance per server. AMD Infinity Fabric link technology not enabled: One four-GPU hive provides up to 256 GB/s peak theoretical P2P performance with PCIe 4.0. Server manufacturers may vary configuration offerings yielding different results. MI200-43