

KT CLOUD พร้อมทั้งจะขยายศักยภาพ AI ด้วย AMD INSTINCT ACCELERATORS

AMD Instinct M1250 accelerators ช่วยให้ KT Cloud ปรับโครงสร้างพื้นฐานให้เหมาะสมเพื่อสร้างสรรคความสามารถใหม่ที่หลากหลายให้กับ AI



ลูกค้า

kt cloud

อุตสาหกรรม

ประเภทบริการ Cloud

ความท้าทาย

ส่งมอบ GPU บน cloud ที่คุ้มค่า และเหมาะสมสำหรับ AI

การแก้ปัญหา

สร้างแพลตฟอร์ม AI ใหม่ที่ขับเคลื่อนด้วย AMD Instinct™ MI250 accelerators

ผลลัพธ์

ประสิทธิภาพเพิ่มขึ้น 1.4 เท่า สำหรับงาน AI และลดค่าบริการ GPU cloud ลงไปได้ 70%

เทคโนโลยี AMD

AMD Instinct MI250 accelerators
2nd Generation AMD CDNA architecture
AMD Infinity Architecture

พันธมิตรด้านเทคโนโลยี

MOREH

กว่าทศวรรษที่ KT Cloud (เดิมชื่อ Korea Telecom) ได้ให้บริการระบบคลาวด์ Cloud ซึ่งเป็นโซลูชันพื้นฐานสำหรับธุรกิจต่างๆ ที่ปลอดภัยและเชื่อถือได้ เทคโนโลยีของ บริษัท ช่วยให้องค์กรต่าง ๆ สามารถใช้ประโยชน์จากความสามารถ การประมวลผลบน Cloud ที่ทรงพลังโดยไม่มีสูญเสียการควบคุมหรือมีความยืดหยุ่นจากความมุ่งมั่นในการให้บริการ และ โซลูชัน ยอดนี้ ทำให้ได้เป็นหนึ่งในบริษัทที่ได้รับการยอมรับมากที่สุดในอุตสาหกรรม ซึ่งให้บริการระบบ Cloud KTCloud มีแผนที่จะมุ่งมั่นเปิดตัวข้อเสนอใหม่หลายอย่าง ที่มีทั้งบริการ AI Cloud สำหรับผู้ใช้ Cloud สาธารณะในรูปแบบของ Infrastructure-as-a-Service (IaaS) นอกจากนี้ KT ยังวางแผนที่จะพัฒนา Infrastructure-as-a-Service (IaaS) เพื่อรองรับงานอื่น ๆ เช่น ศูนย์ให้บริการทางโทรศัพท์อัตโนมัติ นอกจากนี้ KT Cloud ยังตั้งใจที่จะจัดหา Application Programming Interfaces (API) ให้กับบริษัท แอปพลิเคชัน AI เพื่อรองรับแอปพลิเคชัน เช่น Chatbots

GPU (หน่วยประมวลผลกราฟิก) ที่เป็นเลิศในการประมวลผลงาน ทำให้เหมาะสมมากสำหรับงาน AI Rap และ Deep Learning

ยิ่งไปกว่านั้น ยังสามารถมอบพลังการคำนวณที่จำเป็นในการเร่งคำนวณที่ซับซ้อน และอัลกอริธึมที่เกี่ยวข้องกับการเทรนโมเดล AI

การประมวลผลภาษาและแอปพลิเคชันอื่น ๆ ที่เน้นการคำนวณแบบเข้มข้น AM ด้วยเหตุนี้ KT Cloud จึงร่วมมือกับ AMD และ Morech เพื่อสร้างแพลตฟอร์ม AI ใหม่ที่ขับเคลื่อนโดย AMD Instinct MI250 accelerators

ด้วยสถาปัตยกรรม AMD COA รุ่นที่ 2 AMD accelerators ได้รับการออกแบบมาเพื่อ แอปพลิเคชัน Deep Learning ซึ่งมีการทำงานที่มีประสิทธิภาพยอดเยี่ยม รองรับความแม่นยำที่หลากหลาย สำหรับงาน AI และ มีความคุ้มค่าอย่างมีนัยยะสำคัญ

หน่วยความจำ 128 GB ที่มี high bandwidth HBM2e ซึ่งพร้อมการยอมรับ ECC และ AMD Instinct MI250 accelerators ที่มีประสิทธิภาพระดับแนวหน้า

ซึ่งสามารถเชื่อมต่อกับตัวเร่งความเร็วอื่นๆ และ โปรเซสเซอร์ AMD EPYC™ ผ่านทางบิตยกรรม AMD Infinity ซึ่งมีแบนด์วิดธ์สูงถึง 800 GB/S KT Cloud ได้กำหนดเป้าหมาย สอง อย่าง สำหรับแพลตฟอร์ม AI ใหม่ ข้อกำหนดอย่างหนึ่งคือ การเพิ่มประสิทธิภาพการใช้ คลัสเตอร์ GPU สำหรับผลิตภัณฑ์และบริการที่ใช้ AI ประเภทที่สอง คือการสร้างแบบจำลอง ภาษาขนาดใหญ่เพื่อรองรับ ข้อกำหนดการใช้งานเชิงพาณิชย์ในการตลาดเกาหลี

พลิกโฉมการประมวลผลแบบ Cloud ด้วย AI Computing ระดับไฮเปอร์สเกลของ KT Cloud และ AMD Instinct MI250

ผู้ให้บริการ cloud เฝ้าจับตามองความท้าทายหลายประการ เมื่อนำเสนอ GPU เพื่อให้บริการกับลูกค้า ผู้ให้บริการจำเป็นต้องเรียกเก็บค่า GPU ไม่ว่าจะมีการใช้งานอยู่หรือไม่ ทำให้บริการนี้มีความคุ้มค่าสำหรับลูกค้าบางรายการน้อยกว่าการใช้เซิร์ฟเวอร์ภายในบริษัท

“ด้วย AMD Instinct accelerators ที่มีการทำงานซึ่งคุ้มค่า กับการลงทุน KT Cloud คาดว่า จะช่วยลดบริการ GPU Cloud ได้อย่างมีประสิทธิภาพ ลงถึง 70%”

JooSung Kim รองประธานของ KT Cloud กล่าว

ผู้ให้บริการยังต้องลงทุนกับ GPU จำนวนมาก เพื่อตอบสนองความต้องการของลูกค้า การไม่ได้นำโซลูชันฮาร์ดแวร์ GPU การจำลองเสมือนมาใช้ ทำให้เกิดความท้าทายมากมาย

ด้วยการตั้งใจที่จะแก้ไขปัญหเหล่านี้ KT Cloud จึงได้เปิดตัว ตัวต Hyperscale AI Computing ซึ่งเป็น IaaS-บริการ Cloud ระดับ AI บนแพลตฟอร์มคลาวด์ของ MoAI MoAI ของ Moreh และ AMD Instinct M1250 accelerators หลายร้อยตัว

“ด้วย AMD Instinct accelerators ที่คุ้มค่า และ รูปแบบการกำหนดราคาแบบจ่ายตามการใช้งาน KT Cloud คาดว่า จะสามารถลดราคาค่าบริการ GPU cloud ลงได้ถึง 70%” JooSung Kim รองประธานของ KT Cloud กล่าว

ความท้าทายหลายประการ

KT Cloud ตระหนัก ถึงความสำคัญของการถ่ายโอนข้อมูลที่ราบรื่นและการเร่งเวลาในการพัฒนาสำหรับนักพัฒนา AI

เพื่อให้เป็นไปตามข้อกำหนด KT Cloud จึงนำ AMD Instinct GPU มาเป็นตัวเลือกที่ไม่ต้องลือคอคอนเทนที่จะใช้ของ NVIDIA CUDA
AMD Instinct GPU เข้ากันได้กับกรอบการทำงานการเขียนโปรแกรมและไลบรารีมาตรฐานอุตสาหกรรมที่ช่วยให้นักพัฒนาสามารถเขียน hardware-agnostic code ได้ด้วยการใช้ ตัวด้ายการใช้ Code และความเชี่ยวชาญที่มีอยู่ นักพัฒนา AI สามารถปรับเปลี่ยนให้ใช้บนแพลตฟอร์ม AMD ได้อย่างราบรื่นซึ่งทำให้มั่นใจได้ว่าจะมีการทำงานที่ราบรื่นด้วย

แพลตฟอร์ม Moreh MoAI ที่มี Abstraction Layer ซึ่งช่วยให้โมเดลการเทรนนิ่งมีความเร็วที่แม่นยำที่เพิ่มขึ้น ขณะนี้ช่วยลดการทำงานแบบ manual ลง Single-device abstraction ช่วยให้ผู้ใช้สามารถเข้าถึงแอปพลิเคชัน AI จากอุปกรณ์เพียงเครื่องเดียวได้อย่างง่ายดาย Just-in-time graph compiler จะปรับปรุงและเพิ่มประสิทธิภาพการทำงานในโมเดลและโมเดลของ AI หลายรายการ Compiler ยังใช้การแปลงโมเดลแบบใหม่ที่เรียกว่า Moreh IR เพื่อปรับให้เหมาะสมกับการดำเนินงานของ Tensor ที่บันทึกไว้ในกราฟคำนวณทำให้ผู้ใช้มีการสร้างแบบจำลองที่มีประสิทธิภาพมากขึ้น เมื่อเทียบกับการใช้ PyTorch/TensorFlow แบบเดิม MOAI ช่วยให้ผู้ใช้สามารถใช้ API ที่จำเป็นในเฟรมเวิร์กการเรียนรู้เชิงลึก เช่น PyTorch และ TensorFlow 2.0 เพื่อให้ใช้งานง่ายและยืดหยุ่นมากยิ่งขึ้น

KT Cloud Hyperscale AI Computing นำเสนอตัวเลือกในการเร่งความเร็วจำลองเสมือนที่หลากหลายแก่ลูกค้าตั้งแต่ GPU ตัวเดียวและ GPU แบบ 64-48 GB และ 24,576 GB เพื่อที่ปรับขนาดจำนวน GPU ในเครื่องจำลองเสมือนได้อย่างง่ายดาย ตัวเลือกที่เลือกจะไม่ส่งผลต่อการกำหนดค่าเครื่องจำลองเสมือน ดังนั้นจึงไม่จำเป็นต้องแก้ไข แอปพลิเคชัน การบริการนี้แนะนำเครื่องจำลองเสมือนที่มีตัวเร่งความเร็วเสมือนตัวเดียว ช่วยให้ผู้ใช้สามารถสร้างโมเดลเหมือนกับว่าพวกเขาใช้ GPU เพียงตัวเดียวในขณะที่ KT Cloud จัดจัดการตั้งค่าปรับปรุงสภาพแวดล้อมคลัสเตอร์ให้

นำการประมวลผล Hyperscale AI ของ KT Cloud มาทดสอบ

KT Cloud และ Moreh เปรียบเทียบประสิทธิภาพประมวลผล Hyperscale AI รุ่นใหม่ซึ่งใช้ตัวเร่งความเร็ว MoAI และ AMD และ AMD Instinct M1250 accelerators กับบริการ GPU รุ่นระดับตำนานที่มี NVIDIA A100 GPU ของ KT Cloud "เราทดสอบแต่ละแพลตฟอร์มโดยใช้ชุดค่าประมาณที่เหมือนกันใน open source model 40 รุ่น " นาย Kim กล่าว "ผลลัพธ์แสดงให้เห็นว่าบริการแบบใหม่ที่ใช้ เซิร์ฟเวอร์ MoAI และ AMD Instinct M1250 accelerator นั้นเร็วกว่า เซิร์ฟเวอร์ที่ใช้ A100 โดยเฉลี่ย 1.4 เท่า"

ใช้ประโยชน์จากศักยภาพของ Machine Learning ด้วย KT Cloud

โครงการแบบจำลองภาษาเกาหลีพารามิเตอร์ 11 พันล้านพารามิเตอร์ของ KT Cloud เพื่อพัฒนาแบบจำลองภาษาขนาดใหญ่สำหรับภาษาเกาหลี ต้องใช้การประมวลผลจำนวนมากซึ่งเป็นการลงทุนด้าน CAPEX ที่สำคัญ ซึ่ง KT Cloud ประสบผลสำเร็จอย่างโดดเด่น ใช้ประโยชน์จากทุนพลังของ AMD Instinct GPUs มากกว่า 1,000 ตัวซึ่งมอบ "แรงม้า" ประมวลผลอันมหาศาลที่จำเป็นในการขับเคลื่อน KT Cloud ได้อย่างมีประสิทธิภาพ



เกี่ยวกับ KT Cloud

KT Cloud ให้บริการการประมวลผลบน Cloud แก่ธุรกิจ รวมทั้งโครงสร้างพื้นฐาน -as-a-service(IaaS) แพลตฟอร์ม -as-a-service(PaaS) และโซลูชัน software-as-a-service(SaaS) บริการของเราทั้งเก็บข้อมูลบน Cloud การประมวลผลบน Cloud การวิเคราะห์ข้อมูลขนาดใหญ่ ปัญญาประดิษฐ์และอื่นๆ KT Cloud เป็นหนึ่งในผู้ให้บริการ Cloud ชั้นนำในเกาหลีใต้ และได้รับการยอมรับ ในด้านความน่าเชื่อถือและปลอดภัย สำหรับข้อมูลเพิ่มเติมโปรดเยี่ยมชม cloud.kt.com

การอ้างสิทธิ์ประสิทธิภาพและการประหยัดต้นทุนทั้งหมดจัดทำโดย KT Cloud และยังไม่ได้รับการตรวจสอบโดย AMD ผลประโยชน์ด้านประสิทธิภาพและต้นทุนได้รับผลกระทบจากตัวแปรต่างๆ ผลลัพธ์ในที่มีไว้สำหรับ KT Cloud โดยเฉพาะและอาจไม่ใช่เรื่องปกติ GD-181 ©2023 Advanced Micro Devices, Inc. สงวนลิขสิทธิ์ AMD, โลโก้ AMD Arow, EPYC และชื่อที่รวมกันเป็นเครื่องหมายการค้าของ Advanced Micro Devices, Inc. ชื่อผลิตภัณฑ์อื่นๆที่ใช้นี้ อาจเป็นเครื่องหมายการค้าของบริษัทอื่นและอาจมีเครื่องหมายการค้าของบริษัทที่เกี่ยวข้อง

Model ตัวเข้ารหัส-ตัวถอดรหัสที่ใช้ Transformer ขนาดใหญ่และเทรนโดยพารามิเตอร์นับพันล้านตัวเป้าหมายคือการนำเสนอบริการที่ใช้ API พร้อมแอปพลิเคชัน เช่น พยานิชย์ที่มีศักยภาพมาก ตัวอย่างเช่น KT วางแผนที่จะสนับสนุน AI chatbot ที่ให้คำปรึกษาด้านจิตวิทยาโดยอ้างอิงจากที่ปรึกษาชื่อดังชาวเกาหลี Model ภาษาเกาหลีตัวแรกของ KT Cloud ใช้ Training parameter 11 พันล้านparameters และได้รับการประเมินโดยวิธีการเทรน สอง วิธีที่แตกต่างกัน

วิธีแรกใช้คลัสเตอร์ Nvidia DGX A100 รุ่นเก่าซึ่งมี 40 NODES (โดยใช้ 320 GPU) เชื่อมต่อกันสำหรับ bandwidth สูง 1.6 Tb/s ต่อ node วิธีที่สองใช้คลัสเตอร์ AMD ที่มีตัวเร่งความเร็ว AMD Instinct M1250 จำนวน 160 ตัว และซอฟต์แวร์แพลตฟอร์ม moai พร้อมด้วยเครือข่ายการเชื่อมต่อที่สอดคล้องกันมากขึ้นโดยมีการเชื่อมต่อ InfiniBand 2 การเชื่อมต่อ node ที่มีความเร็ว 400Gb/s พร้อมด้วยซอฟต์แวร์ที่จัดการการสื่อสารได้อย่างมีประสิทธิภาพพร้อม Application ของผู้ใช้งาน KT ผลประสบความสำเร็จจากการเทรนทั้ง 2 คลัสเตอร์ "ระบบที่ใช้ตัวเร่งความเร็ว AMD Instinct ใช้งาน network switch และสายเคเบิลเพียง 25% เมื่อเทียบกับของคู่แข่ง "ใน Kim อธิบาย

ผลลัพธ์แสดงให้เห็นว่าบริการแบบใหม่ที่ใช้ซอฟต์แวร์ MoAI และ AMD Instinct M1250 accelerator นั้นเร็วกว่า เซิร์ฟเวอร์ที่ใช้ A100 โดยเฉลี่ย 1.4 เท่า Joosung Kim รองประธาน KT Cloud

"การติดตั้ง switch แต่ละครั้งมีค่าใช้จ่ายประมาณ 20,000 ดอลลาร์ ดังนั้นเมื่อคลัสเตอร์ของ KT เติบโตขึ้นประโยชน์ก็จะเกิดขึ้นได้อย่างชัดเจน"

"ในแง่ของความคุ้มค่าคลัสเตอร์ที่ใช้ AMD Instinct ซึ่งใช้ซอฟต์แวร์ Moreh ที่มีปริมาณงานที่สูงขึ้น 1.9 เท่า ต่อดอลลาร์เมื่อเทียบกับคลัสเตอร์ของ NVIDIA และปรับปรุงผลลัพธ์ ได้สูงสุดถึง 117%" นาย Kim กล่าว

ประสิทธิภาพของ AI ก้าวกระโดดด้วยคลัสเตอร์ซูเปอร์คอมพิวเตอร์ 1200 GPU ของ KT Cloud

KT Cloud ประกาศผลถึงความสำเร็จในการสร้างคลัสเตอร์ซูเปอร์คอมพิวเตอร์ มี GPU AMD Instinct MI250 ถึง 1,200 ตัวเพื่อการเทรน Model ภาษาเกาหลี ในเวอร์ชันถัดไป

ด้วยพารามิเตอร์ 2 แสนล้านพารามิเตอร์นี้จะมีจุดสูงสุดทางทฤษฎีที่ 434.5 PFLOPS สำหรับปฏิบัติการเมทริกซ์ fp16/bf16, 108.6 PFLOPS สำหรับปฏิบัติการเมทริกซ์ fp32/fp64 และ 54.4 PFLOPS และสำหรับปฏิบัติการเวกเตอร์ fp32/fp64 ซึ่งอาจจะทำให้เป็นหนึ่งในซูเปอร์คอมพิวเตอร์ GPU อันดับต้น ๆ ของโลก

หากต้องการทราบว่า โปรเซสเซอร์ AMD EPYC จะเป็นประโยชน์ต่อการทำงานของของคุณได้อย่างไร

โปรดลงทะเบียนเพื่อรับข้อมูล จากศูนย์ข้อมูลของเรา amd.com/epycsignup



เกี่ยวกับ AMD

เป็นเวลากว่า 50 ปีแล้วที่ AMD ได้ขับเคลื่อนนวัตกรรมด้านเทคโนโลยีระดับสูง เทคโนโลยีประมวลผลประสิทธิภาพสูงกราฟฟิกและเทคโนโลยีจำลองเสมือนผู้บริโภคหลายร้อยล้านคน ซึ่งเป็นผู้นำในธุรกิจ fortune 500 และศูนย์วิจัยวิทยาศาสตร์ ที่ทันสมัยทั่วโลกต่างไว้วางใจ ในเทคโนโลยีของ AMD เพื่อปรับปรุงการใช้ชีวิต ทำงาน และการเล่นเกม พนักงานของ AMD มุ่งเน้นไปที่การสร้างผลิตภัณฑ์ชั้นนำที่มีประสิทธิภาพสูงก้าวข้ามความเป็นไปไม่ได้ หากต้องการข้อมูลเพิ่มเติมเกี่ยวกับวิธีที่ AMD ขับเคลื่อนในวันนี้ และสร้างแรงบันดาลใจในอนาคตโปรดไปที่ AMD (NASDAQ:AMD) Website, Blog, LinkedIn และ Twitter