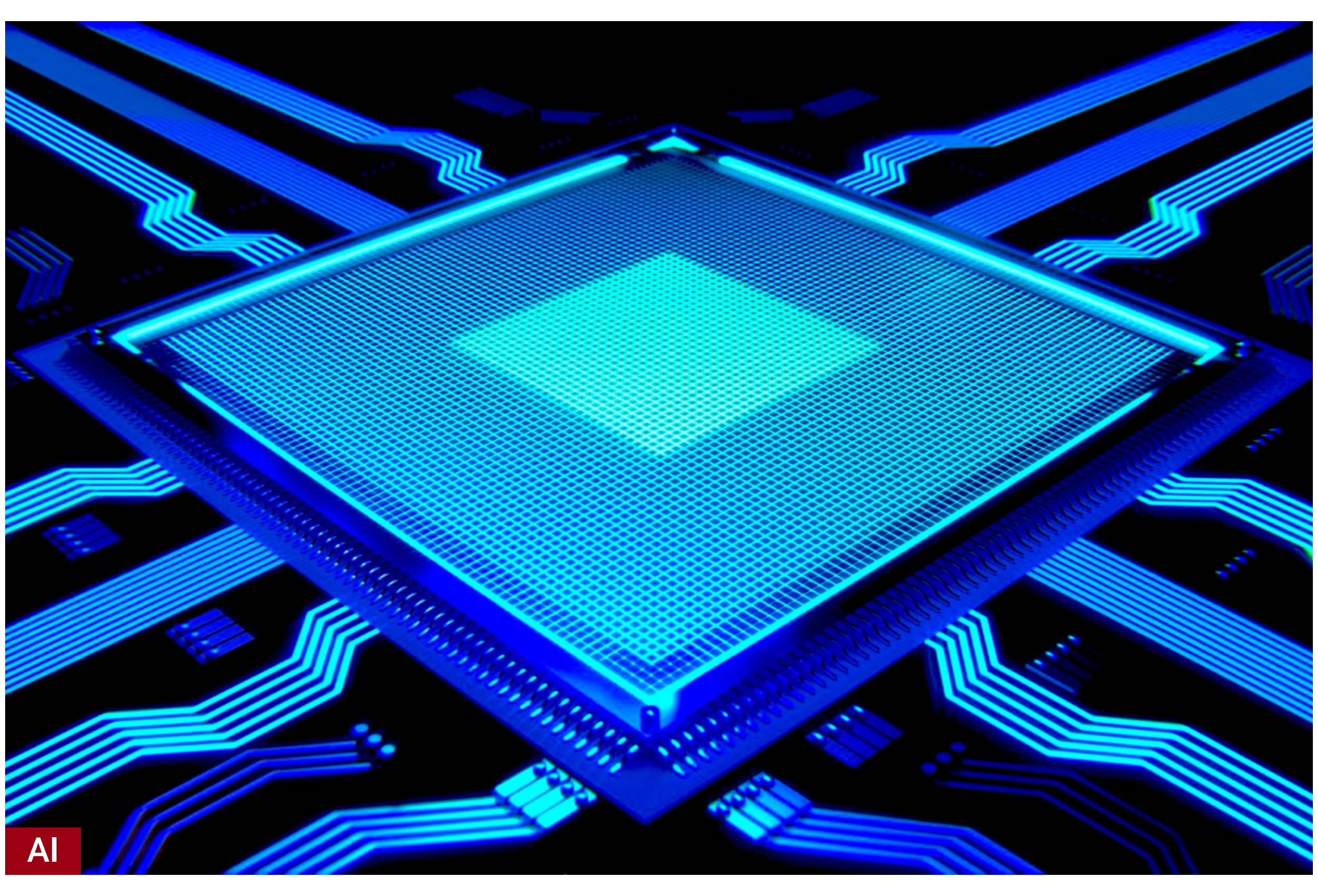
UPDATED 17:22 EDT / FEBRUARY 05 2023



AI

Generative AI drives an explosion in compute: The looming need for sustainable AI

GUEST COLUMN BY SID SHETH

After years of preparation, 2022 highlighted the amazing potential for generative AI as models such as OpenAI LLC's DALL-E and GPT-3 swept across the world. Microsoft Corp., Amazon.com Inc. and Google LLC have been training machine learning models for years, but the introduction of transformer-based large language models or LLMs that could "learn" created a tremendous leap forward in usefulness.

There's one problem with this generative AI explosion, though: Every time DALL-E creates an image or GPT-3 predicts the next word, this requires multiple inference calculations that add up to significant electrical demand. Current graphics processing unit and central processing unit architectures can't operate efficiently enough to meet the looming demand, creating a big problem for hyperscalers.

Data centers will become the world's largest energy consumers, rising from 3% of total electricity use in 2017 to 4.5% by 2025. China predicts its data centers will consume more than 400 billion kWh of electricity in 2030 — 4% of the nation's total electricity use.

Cloud providers recognize the massive quantity of electricity they use, and have implemented efficiency measures such as locating data centers in arctic countries to capitalize on natural cooling and renewable energy. It won't be enough for the AI explosion, though: Lawrence Berkeley National Laboratory found that efficiency gains have kept this trend under control for the past 20 years, but "modeled trends indicate efficiency measures of the past may not be enough for the data center demand of the future."

We need a better approach.

Data movement is the killer

The efficiency problem is rooted in how CPUs and GPUs work, especially for running an AI inference model versus training the model. You've heard about "moving beyond Moore's Law" and the physical limitations of packing more transistors onto larger die sizes. Chiplets are helping to address these challenges, but current solutions have a key weakness when it comes to AI inference: Shuttling data in and out of random-access memory leads to significant slowdowns.

Traditionally, it has been cheaper to manufacture processors and memory chips separately, and for many years, processor clock speeds were the key gating factor for performance. Today it's the interconnection between chips that's holding things back. "When memory and processing are separate, the communication link that connects the two domains becomes the primary bottleneck of the system," Jeff Shainline from NIST explains. Professor Jack Dongarra from Oak Ridge National Laboratory said succinctly that "when we look at performance today on our machines, the data movement is the thing that's the killer."

AI inference versus AI training

An AI system uses different types of calculations when training an AI model compared to using it to make predictions. AI training loads a transformer-based model with tens of thousands of images or text samples for reference, then starts crunching away. The thousands of cores in a GPU are very effective at digesting large sets of rich data like images or video, and if you need results faster, you can just rent as many cloud-based GPUs as you can afford.

AI inference requires less power up front to make a calculation — but the massive number of calculations and predictions needed to decide what the next word should be in an autocomplete across hundreds of millions of users takes much more energy than training over the long run. Facebook AI observes trillions of inferences per day across its data centers — and this has more than doubled in the past three years. Facebook AI also found that running inference on an LLM for language translation can use two to three times as much power as initial training.

An explosion of demand

We saw how ChatGPT swept the industry late last year, and GPT-4 will be even more impressive. If we can adopt a more energy-efficient approach, we can expand inference to a wider range of devices and create new ways of doing computing.

Microsoft's Hybrid Loop is designed to build AI experiences that dynamically leverage both cloud and edge devices. This allows developers to make late binding decisions on running inference in the Azure cloud, or the local client computer or mobile device. This maximizes efficiency while users have the same experience regardless of where inference happens. Similarly, Facebook introduced AutoScale to help efficiently decide at runtime where to compute inference.

New approaches to efficiency

If we want to open up these possibilities, we need to get over the barriers slowing down AI today. There are several promising approaches.

Sampling and pipelining can help speed deep learning by trimming the amount of data processed. SALIENT (for SAmpling, sLIcing, and data movemeNT) was developed by researchers at the Massachusetts Institute of Technology and IBM Corp. to address key bottlenecks. This approach can dramatically cut down the requirements for running neural networks on large datasets which can contain 100 million nodes and 1 billion edges. But it also limits accuracy and precision — which can be OK for selecting the next social post to display, but not if trying to identify unsafe conditions on a worksite in near real time.

Apple Inc., Nvidia Corp., Intel Corp. and Advanced Micro Devices Inc. have announced processors with dedicated AI engines incorporated into or sitting next to traditional processors. Amazon Web Services Inc. is even creating the new Inferentia2 processor. But these solutions are still using traditional von Neumann architecture of processors, integrated SRAM and external DRAM memory — which all require electricity to move data in and out of memory.

There's one other approach to break down the "memory wall" that researchers have identified — and that's moving compute closer to the RAM.

In-memory computing improves latency, reduces energy

The memory wall refers to the physical barriers limiting how fast data can be moved in and out of memory. It's a fundamental limitation with traditional architectures. In-memory computing or IMC addresses this challenge by running AI matrix calculations directly in the memory module, avoiding the overhead of sending data across the memory bus.

IMC works well for AI inference because it involves a relatively static (but large) data set of weights that is accessed over and over again. There's always a need to transfer some data in and out, but IMC eliminates the majority of the energy transfer expense and latency of data movement by keeping data in the same physical unit where it can efficiently be used and reused for multiple calculations.

This approach promotes scalability because it works well with chiplet designs. With chiplets, AI inference technology can scale from a developer's desktop for testing, before deploying to production at the data center. A data center can use an array of cards or a large device with many chiplet processors to efficiently run enterprise-grade AI models.

Over time, we predict IMC will become the dominant architecture for AI inference use cases. It just makes so much sense when you have massive data sets and trillions of calculations. You don't have to waste energy shuttling data across the memory wall, and the approach easily scales up to meet long-term demands.

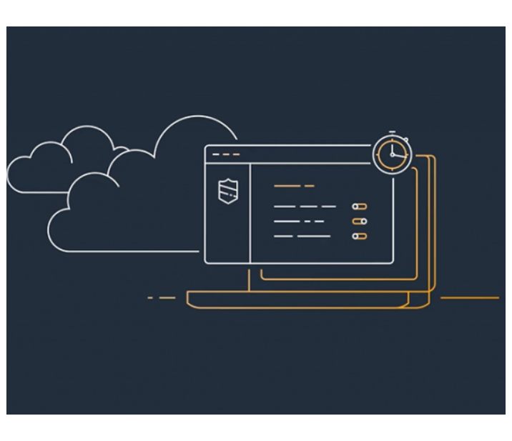
We're at such an exciting inflection point with advancements in generative AI, image recognition and data analytics all coming together to uncover unique new connections and uses for machine learning. But first we need to build a technological solution that can meet this need — because right now, unless we can create more sustainable options, Gartner predicts that by 2025, "AI will consume more energy than the human workforce."

Let's figure out a better approach before this happens.

Sid Sheth is chief executive of d-Matrix. He wrote this article for SiliconANGLE. Disclosure: The company offers its own digital in-memory computing architecture.

Image: Colin Behrens/Pixabay

LATEST STORIES



WS launches AWS Verified ccess into general availability

Innovation targets hard problems at RSA Conference

Spending on cloud infrastructure services tops $63B in the first quarter

CUBE EVENT COVERAGE

◀ Prev Clip | Next Clip ▶

Day 3 Keynote Analysis | RSAC 2023

VIEW FULL VIDEO

LATEST FROM THECUBE

Delivering cloud security inside-out using a purpose-built graph database

The cybersecurity conundrum: integrated platforms vs. best-of-breed solutions

From chaos to cohesion: how integrated platforms streamline cybersecurity

Solving the problem of inefficient enterprise penetration testing

The data defender's mindset: generative AI, LLMs and data protection

DIVE INTO DAVE VELLANTE'S BREAKING ANALYSIS SERIES

Don't be fooled by slowing cloud growth: Cost optimization is a feature, not a bug

Dave Vellante's Breaking Analysis: The complete collection

RSA 2023 and the security identity crisis, part two

RSA 2023 highlights a cybersecurity identity crisis in the age of AI

Hidden gems from HPE's GreenLake Storage Day

UPCOMING CUBE EVENTS

Optimize your Media Supply Chain and Increase Velocity with AWS
May 04-04

ISC High Performance 2023
May 22-24

Dell Technologies World 2023
May 23-25

Red Hat Summit + Ansiblefest 2023
May 23-25

HPE Discover 2023
Jun 20-22

RECENT CUBE EVENTS

RSA Conference 2023

KubeCon + CloudNativeCon Europe 2023

QlikWorld 2023

MWC 2023

AWS Startup Showcase: Top Startups Building Generative AI on AWS 2023

Special Program Series: Women of the Cloud 2023

CloudNativeSecurityCon 2023

Supercloud2 2023

theCUBE Live at Ignite 2022

AWS re:Invent 2022

AWS re:Invent 2022 - Accenture Executive Summit

AWS re:Invent 2022 - Global Startup Program

theCUBE Live at Supercomputing 2022

KubeCon + CloudNativeCon NA 2022

AnsibleFest 2022

IFS Unleashed 2022

UiPath Forward 2022

Fal.Con 2022

AWS Marketplace Seller Conference 2022

AWS Startup Showcase: Cybersecurity - Detect and Protect Against Threats 2022

VMware Explore 2022

Supercloud 2022

Monaco Crypto Summit

AWS re:Inforce 2022

AWS Summit New York 2022

Amazon re:MARS 2022

Snowflake Summit 2022