ComputerWeekly.com

Home > AI and storage

KTSDESIGN - STOCK.ADOBE.COM

FEATURE

What are tensor processing units and what is their role in AI?

We look at Google’s TPUs – tensor processing units – and ask what makes them different to CPUs, GPUs and DPUs, as well as how you can take advantage of them in AI processing

By **Antony Adshead,** Storage Editor

Published: **17 Oct 2024**

There are central processing units (CPUs), graphics processing units (GPUs) and even data processing units (DPUs) – all of which are well-known and commonplace now. GPUs in particular have seen a recent rise to prominence with the emergence of artificial intelligence (AI).

You may have also heard of tensor processing units (TPUs), which are a Google creation and only available via their cloud services.

But what are TPUs, and why might you need them?

In short, they are processing units customised for use with the high-dimensional data found in AI processing operations. But, before we set out their characteristics in detail, let’s see how they compare with other forms of processor.

CPUs, as we all probably know, are the core of all computing. Core being the operative word nowadays, as multiple cores – in other words, individual processing units within the CPU – handle multiple compute functions.

CPUs are in fact the general purpose processors of the computing world. They handle incoming instructions and outgoing messaging to other parts of the system and the overall orchestration of events. That’s because they can handle numerous sets of functionality simultaneously and they’re good at it.

But, ever since there have been CPUs, there has also been the need to offload processing-heavy operations to more specialised chips.

What use are TPUs when GPUs exist?

And so, GPUs emerged. They were originally, as the name suggests, built to handle graphics processing in gaming, but later began to find a use in AI. That’s because graphics processing is all about matrix operations. In other words, in calculations that involve matrices – for example, in multiple dimensions – and these were found to be suited to AI operations also.

But, while GPUs can handle matrix operations, they are not quite as customised to this task as TPUs, as we shall see.

Nvidia has somewhat become synonymous with the market in GPUs for AI use cases lately, but they are also available from other vendors.

What do DPUs do?

Then there are also DPUs, deployed in servers to handle data transfer, data reduction, security and analytics. Here, these tasks are offloaded from CPUs with DPUs to bring more specialisation to the tasks involved, freeing up the CPU for more general and orchestration duties.

DPUs are available from Intel, Nvidia, Marvell and AMD, as well as in the cloud from, for example, Amazon Web Services, with its Nitro cards.

What's special about TPUs?

TPUs were first used in about 2016. They are, as mentioned above, tensor processing units. A tensor is a form of matrix, or multi-dimensional number, a key plank of AI processing that assigns and uses high-dimensional numbers to objects to process them.

Whether there are specifically mathematical tensors in use or not in Google's TPUs, the key thing that characterises them is that they are built around ASIC chips customised for calculations involving high-dimensional numbers. ASIC stands for application-specific integrated circuit. That means chip hardware that's designed especially for specific operations. Specifically, the ASICs on a TPU are called matrix-multiply units (MXUs).

That's what makes TPUs different from CPUs, which are general purpose processors. DPUs and GPUs can be built around ASICs, or field programmable gate arrays (FPGAs), that are not designed around one task but can be configured as needed for a wide range of uses.

Google's tensor ASIC in its TPUs was designed for use with its open source TensorFlow AI software framework, which helps run advanced AI analytics models in which data is arrayed via high-dimensional patterns.

TPUs are available from Google, delivered as a service. The latest version, the TPU v5p, allows for peak compute of 459 floating point operations per second (teraflops). Nowadays, in addition to TensorFlow, Google TPUs support other AI frameworks such as PyTorch and Jax, with image classification and generation, and large language model operations possible.

Having said all that about the specialisation of TPUs, anyone thinking of building their own AI systems in-house won't be able to buy them anyway. GPUs will do the job just as well, and the jury is out over whether there's a performance advantage for TPUs. It's just that if you want to work in the (Google) cloud, TPUs will play well with the AI software stack available there.

Read more on AI and storage

- Storage technology explained: Vector databases at the core of AI. We look at the use of vector data in AI and how vector databases work, plus vector embedding, the challenges for storage of vector data and the key suppliers of vector database products.

- Storage technology explained: AI and data storage. In this guide, we examine the data storage needs of artificial intelligence, the demands it places on data storage, the suitability of cloud and object storage for AI, and key AI storage products.

Read more on AI and storage

Google launches Parallelstore file storage at cloud AI training

By: Yann Serra

What is an AI accelerator?

By: Chris Tozzi

GPUs vs. TPUs vs. NPUs: Comparing AI hardware options

By: Stephen Bigelow

Securing data in GCP: A Computer Weekly Downtime Upload podcast

By: Cliff Saran

poin
or U

ent-ele
strong
ina, co

DATA MANAGEMENT

Experts urge DOGE to prioritize tech over agency cuts

Elon Musk and Vivek Ramaswamy want to delete federal agencies. However, some hope the duo will turn their focus to improving ...