Efficient Datacenter Scale AI Acceleration Using 3D Optics

Phillip Burr
Lumai
Oxford, UK
phillip.burr@lumai.co.uk

Xianxin Guo
Lumai
Oxford, UK
xianxin.guo@lumai.co.uk

James Spall
Lumai
Oxford, UK
james.spall@lumai.co.uk

***Abstract*—In this paper we describe Lumai's AI datacenter accelerator which uses 3D optics to perform ultra-fast and energy efficient AI inference. The paper outlines the need for a new approach, the underlying optical computing scheme, the advantages, how it can be easily deployed in datacenters, and how it can help address the sustainability and TCO challenges that the industry faces over the coming years.**

Keywords—AI acceleration, optical compute, datacenter sustainability, total cost of ownership

I. Introduction

Today's answer to the demand for increased AI performance in datacenters is to add more silicon area, more cost and more power. Solutions are chasing diminishing returns – each step in performance requires more and more technology and cost, whether it is the latest silicon process node, advanced packaging, or the latest HBM. Squeezing performance out of current AI accelerator solutions is an expensive business, with one large GPU supplier reported to have spent "about $10bn" on developing their latest generation of device [1]. High manufacturing and material costs, high infrastructure costs (for cooling etc) along with high development costs means that the capital costs of AI hardware are huge. This is even before the environmental/sustainability impact is considered.

A new approach is needed. AI acceleration using optics offers a new and disruptive approach. This paper describes how 3D (or "free-space") optics can be used to solve the challenge of delivering a leap in AI inference performance whilst lowering total cost of ownership (TCO) (i.e. both capital and operational costs). The paper outlines the approach, the roadmap of increasing performance and how this technology is being brought to market.

II. AI Acceleration Using Optics

A. The Case for Optical Computing

The cost, power and sustainability challenges of increased AI performance has prompted the industry to look at new approaches to AI acceleration. Significant research has been undertaken in the field of optical compute as it offers the potential to solve these challenges. The maths underpinning AI lends itself perfectly to optical compute and when done correctly, the performance and efficiency gains can be incredible.

There are two types of optics solutions:

1) Integrated Photonics

Although there have been several research groups and companies looking at using integrated photonics for AI processing, these companies are now heavily focused on interconnect or switching. Developing higher-bandwidth interconnects is a key issue for future AI deployments, and integrated photonics is an exciting technology to help overcome the constraints of electronic-only interconnect solutions. However, for processing, and specifically matrix multiplication (matmul), it is extremely challenging to get the performance needed using integrated photonics, due to the low component density, poor scalability and low compute precision.

2) A better approach – 3D Optics

Lumai's approach is to use free-space optics to perform highly parallel computing which overcomes all of the challenges with integrated photonics, resulting in a leap in performance at approximately 10% of the power compared to a GPU.

B. Lumai's Optical AI Accelerator's Quadratic Scaling Advantage

Lumai's optical AI accelerator encodes vector data in the intensity of laser beams and performs matrix–vector multiplication (MVM) by precisely controlling the beams and harnessing the laws of physics.

Lumai's accelerator can leverage the multitude of fundamental benefits of optics already found in optical communications: wavelength multiplexing, fast clock speeds and negligible energy consumption. Significantly, Lumai's accelerator is fundamentally more scalable than all other '2D' integrated chip architectures, whether digital or analog, electronic or photonic, due to the quadratic scaling advantage of using all three spatial dimensions to perform computations.

The Lumai processor performs the three elements of matrix multiplication (copying, multiplying, and adding) by precisely manipulating millions of individual beams of light. Vectors are encoded with fast and energy-efficient optical emitters; copied and summed by spreading and converging the beams with passive lenses and multiplied by controlling the transmission with liquid crystal or MEMS display devices.

Once the information is encoded in light, the matmul calculation (copying, multiplying and addition) is performed almost 'for free', meaning the entire MVM, with millions of parallel operations, is performed in a single clock cycle with extremely low energy consumption.

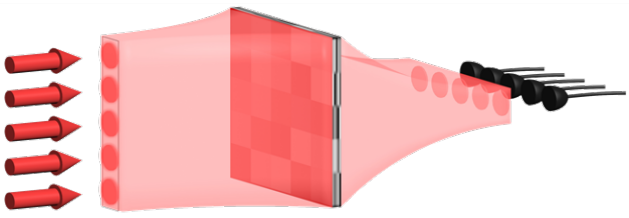

Fig. 1 Conceptual diagram of free-space optical vector-matrix multiplication

The advantage of Lumai's approach is that the system becomes more power efficient as performance increases. This is because the power consumed increases at most linearly as the width of the input vector increases, but the number of mathematical operations performed scales quadratically.

A large matrix size therefore yields a significant increase in throughput without the associated energy cost.

The optical core is interfaced with an otherwise digital electronics hardware stack to maximize the compatibility, programmability and flexibility of the AI accelerator. This hardware stack also performs non-linear accelerator functions.

C. *Memory Advantage*

If not designed carefully, accelerator performance will be limited by memory bandwidth. The Lumai processor has the advantage that the memory can be distributed across the full width of the vector, increasing the memory bandwidth available without needing to use expensive HBM.

In addition, as the Lumai accelerator is able to execute the entire MVM in a single clock cycle, it does not need to divide up the matrix and store intermediate results. This significantly reduces the amount of data that needs to be stored and retrieved either in local cache or HBM memory.

III. Designed for Datacenters

A. *Datacenter Form Factor*

Lumai's accelerator has been designed for use in datacenters. The optical module and accompanying electronic control hardware each use PCIe form-factor cards. These in turn sit within off-the-shelf datacenter trays alongside standard host CPUs. Multiple accelerators can be interconnected for scale-out, and standard interfaces are supported.

Using free-space optical components in datacenters is not new - Google have deployed their Optical Circuit Switch (OCS) across their estate for many years [2], helping demonstrate the reliability and effectiveness of using similar technology within the exacting datacenter requirements.

Fig. 2 Lumai AI Accelerator – Two Accelerators Per 4U Tray

B. *Software*

Because the optical core performs matmul in an extremely efficient way with minimal resource overhead, the compute architecture is simplified, and the software stack can support a wide range of AI models across standard frameworks (Pytorch, Tensorflow etc). By using standard stacks and tools, models can be compiled seamlessly.

The optical part of the design sits at the hardware layer and it is abstracted from software developers who may be unaware that the processing is using photons instead of electrons (other than the cost of processing is significantly lower!).

IV. Sustainablity

As an industry we should be proud that datacenter power demand remained relatively flat between 2015 and 2019, despite workloads nearly tripling [3]. The drive to efficiency has been impressive. We now need to find new, innovative ways to enable an AI revolution, without consuming evermore energy. It has been estimated that by 2028, AI will represent almost one fifth of datacenter power, and that overall datacenter power demand will grow 160% by 2030 [3].

Of course the environmental impact isn't just about energy generation – for every additional Watt of power used, more power and cooling infrastructure is needed and more carbon is produced in creating this. In addition, if we can extend the life of existing datacenters by reducing the power needed for AI accelerators, fewer new datacenters will be needed and the corresponding carbon emissions reduced.

These are sobering figures and a stark reminder that as an industry we now need to do better, including embracing new computational techniques and technologies.

V. Total Cost of Ownership

No discussion of an AI accelerator would be complete without discussing TCO. Lumai's accelerator uses standard optical and electronic components already used within datacenters, re-engineered for the requirements of the optical processor. It does not rely on expensive HBM memory and it does not need the latest silicon technology or packaging. As a result, development, component and manufacturing costs are all a small fraction of those needed for semiconductor-only accelerators. This results in a significantly reduced capital cost for operators. When combined with the reduced power and cooling infrastructure requirements (e.g. no need for liquid cooling), the resulting TCO is a fraction of a GPU.

References

[1] T. Kim "Nvidia CEO Says Blackwell GPU Will Cost $30,000 to $40,000" Barrons, March 2024. https://tinyurl.com/zjvx57tn

[2] R Urata et al "Mission Apollo: Landing Optical Circuit Switching at Datacenter Scale", 2022. https://arxiv.org/pdf/2208.10041

[3] Goldman Sachs "AI is poised to drive 160% increase in data center power demand", May 2024. https://tinyurl.com/476ff4z9