ELIYAN TECHNOLOGY MAY REWRITE HOW CHIPLETS COME TOGETHER

COMPANY'S INTERCONNECT APPROACH ELIMINATES THE NEED FOR EXPENSIVE INTERPOSERS ACCELERATING AI PROCESSING; COULD DOUBLE THE AMOUNT OF MEMORY FOR CHATGTP-LIKE AI, SAVING MILLIONS.

Artificial Intelligence is finally having its iPhone moment. The launch of ChatGPT led to waves of industry-wide excitement, with massive focus on large pretrained generative AI models like GPT-3, GPT-4 etc. Humanity has rushed to the edge of significant technology disruption, with new tools and capabilities seemingly limited only by our imagination.

But there is no free lunch. To be truly valuable – and to build a future of technology parity enjoyed around the world - the cost of running these models must be reduced by perhaps an order of magnitude over the next few years to meet the ambitious goals of companies like Microsoft and Google and make these new capabilities accessible to all.

The cost of training large language models (LLMs) and foundation models is quite high, reportedly more than $10M spent on compute hardware and energy to train a single model. Usage of the models – called inference – is significantly more costly than other compute workloads we currently rely on. For comparison, the cost per inference of ChatGPT is estimated to be anywhere from 4 to 70X more than a Google search!

Considerable attention and capital are now concentrating on companies that can increase the compute efficiency needed to handle these massive new workloads. Santa-Clara-based Eliyan is a chiplet startup with a potential game changer, as its interconnect technology enables more memory and lower costs than is currently possible. Let's take a closer look.

WHAT PROBLEM IS ELIYAN SOLVING?

ChatGPT, Bard, and similar AI rely on large language, or foundation, models which are trained on massive GPU clusters and have hundreds of billions of parameters. Training these models demand far more memory and compute that are available on a single chip, so the models must be cascaded onto large clusters of GPUs, or ASICs like Google TPU. While most inference processing can run on CPUs, LLM apps like ChatGPT require 8 NVIDIA A100 GPUs just to hold the model and process every ChatGPT query, and the accelerator memory size is constrained by how many High Bandwidth Memory (HBM) chips can be connected to each GPU / ASIC. That's where Eliyan comes in.

WHAT DOES ELIYAN DO? TECH AND PRODUCT

Today's chip-to-chip interconnects are expensive and add significant engineering time to a chip development schedule. Eliyan's high performance PHY (physical layer chip componentry) technology unlocks design flexibility without impacting communication performance, connecting chips and chiplets more efficiently and to the needs of the target workload.

NuLink PHY, a chiplet interconnect technology based on a superset of industry standards UCIe and BoW, provides similar bandwidth, power, and latency to those interconnects on a silicon-based interposer but on standard organic substrates. NuLink reduces system costs by simplifying the system design. More importantly, for generative AI, NuLink increases memory capacity and thus the performance of HBM-equipped GPUs and ASICs for memory-dense applications.

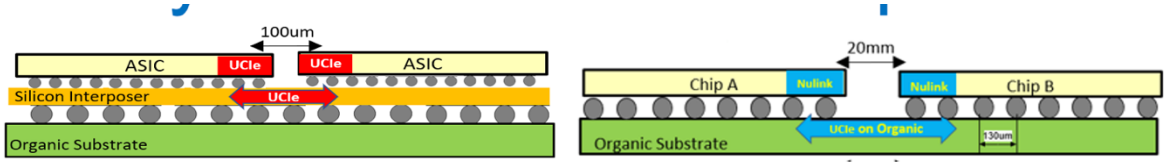

Figure 1: On the left is today's approach to Chiplet interconncts, on the right is Eliyan's NuLink, which can implement UCIe at superior bandwidth to interconnect chiplets without requiring an interposer. Consequently,, Eliyan can increase the distance between chips, allowing more HBM memory per ASIC. This could significantly lower the cost of inference processing for LLM's like ChatGPT.

Eliyan has also created a chiplet called **NuGear**, converting an HBM PHY interface to the NuLink PHY. The NuGear chiplet allowing standard off-the-shelf HBM parts to be packaged with GPUs and ASICs in standard organic packaging without the need of any interposer.

Extending beyond chiplet communications, **NuLinkX** extends the reach of NuLink by 10x to at least 20cm, supporting chip-to-chip external routing over a Printed Circuit Board (PCB). NuLinkX increases the design flexibility for high performance systems by providing unmatched bandwidth efficiency externally, helping system designers optimize for high performance workloads by enabling efficient processor clustering and memory expansion .

NULINK FOR HBM INTERCONNECTS: AN EXAMPLE

NVIDIA, Google, AMD, Intel: all leverage system designs connecting an ASIC, e.g., GPU, to HBM for AI workloads. Today's chip designers use advanced packaging to integrate HBMs with other ASICs, effectively a well-defined set of high performance, expensive interconnections enabling fast communication between logic and memory. It

works, but its rigid – with silicon interposers, given size limitations of processing technology, we're limited to 6 HBM3 blocks per SOC today.

Eliyan's NuLink eliminates the need for such advanced interposers, directly attaching the HBM dies to the ASIC through the organic substrate.

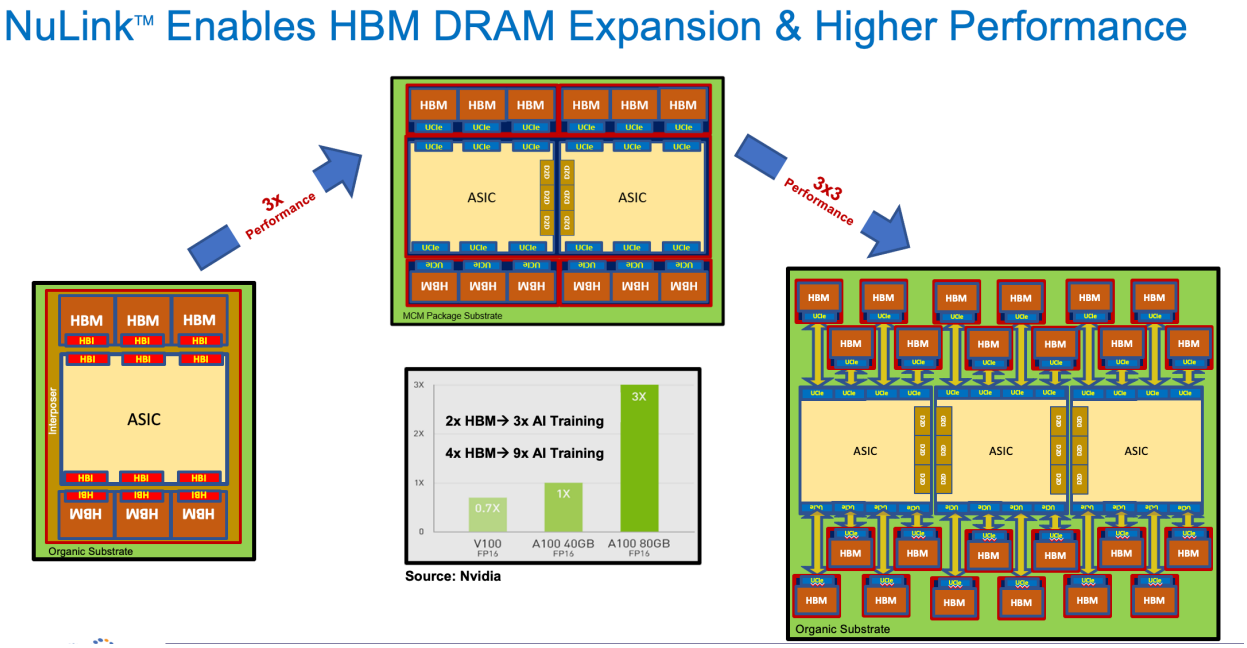

Figure 2: Using NuLink to attach HBM chiplets to an ASIC such as a GPU can increase performance by 3 to 9 fold. More importantly, NuLink can increase HBM count and capacity, which can decrease the number of ASICs or GPUs needed for processing Large Language Modes.

NVIDIA offers two models of the A100 GPU, with 40 and 80GB of HBM, and indicates a 3X performance advantage afforded by the larger memory.

Leveraging NuLink, one could increase the number of HBMs by a factor of two to 160 GB. Assuming linear scaling of the memory benefit in AI training, adopting NuLink triples performance yet again.

Because NuLink connects chiplets without an interposer, the size of the package could increase beyond that of the reticle. One could imagine a three ASIC package with 24 HBM stacks, or 384 GB, as shown in Figure 2. If one assumed the same performance scaling as NVIDIA enjoys going from 40 to 80 GB, then you could potentially realize a 9X performance boost, assuming the 3 ASICs can process the math without becoming compute bound.

While having more memory to train a large language model could be impacted by as much as a 10X speedup, inference processing could benefit as well by reducing the

accelerator footprint needed to hold the model as one could process large models with fewer ASICs.

At 175 billion parameters, GPT-3 is a huge model requiring upwards of 700 GB of high-performance memory to run. At 80 GB per GPU, that means at least 8 GPUs are needed to run ChatGPT. If the GPU/ASIC is poorly utilized, then a smaller number of chips each with more memory could run the inference query with fewer GPUs, saving millions of dollars at scale. The reduction or simplification of the compute cluster would also translate to significantly more sustainable infrastructure. One larger concoction of an Eliyan based system replaces up to 10 individual A100s. Less aggregate material, energy reduction (both into the POD and dissipation) as well as space are secondary but likely important components to consider.

CONCLUSIONS

Eliyan eliminates the need for advanced packaging such as silicon interposers in chiplet designs and all its associated limitations and complexities; therefore, may win client deployments based on their PHY technologies that lower costs, improve yields and improve chip time to market. Additionally, companies such as NVIDIA, Intel, AMD, and Google could license the NuLink IP, or buy NuGear chiplets from Eliyan, to eliminate the performance bottlenecks imposed by limitations of silicon interposers size and enable them to achieve higher-performance AI and HPC SoCs.

We believe that Eliyan has found a niche in the chiplet world that could turn into a bonanza.

\

IMPORTANT INFORMATION ABOUT THIS PAPER

CONTRIBUTORS AND PUBLISHER

Karl Freund, Founder and Principal Analyst, Cambrian-AI Research LLC

INQUIRIES

Contact us if you would like to discuss this report, and Cambrian-AI Research will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Cambrian-AI Research". Non-press and non-analysts must receive prior written permission by Cambrian-AI Research for any citations.

LICENSING

DISCLOSURES

This document was developed with Eliyan Inc. funding and support. Although the document may utilize publicly available material from various vendors, it does not necessarily reflect the positions of such vendors on the issues addressed in this document.

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Cambrian-AI Research disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Cambrian-AI Research and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.