Optically Connected Multi-Stack HBM Modules for Large Language Model Training and Inference

Yanghui Ou, *Graduate Student Member, IEEE*, Hengrui Zhang, Austin Rovinski, *Member, IEEE*, David Wentzlaff, *Member, IEEE*, and Christopher Batten, *Member, IEEE*

***Abstract*—Large language models (LLMs) have grown exponentially in size, presenting significant challenges to traditional memory architectures. Current high bandwidth memory (HBM) systems are constrained by chiplet I/O bandwidth and the limited number of HBM stacks that can be integrated due to packaging constraints. In this letter, we propose a novel memory system architecture that leverages silicon photonic interconnects to increase memory capacity and bandwidth for compute devices. By introducing optically connected multi-stack HBM modules, we extend the HBM memory system off the compute chip, significantly increasing the number of HBM stacks. Our evaluations show that this architecture can improve training efficiency for a trillion-parameter model by 1.4× compared to a modeled A100 baseline, while also enhancing inference performance by 4.2× if the L2 is modified to provide sufficient bandwidth.**

***Index Terms*—Memory architecture, silicon photonics.**

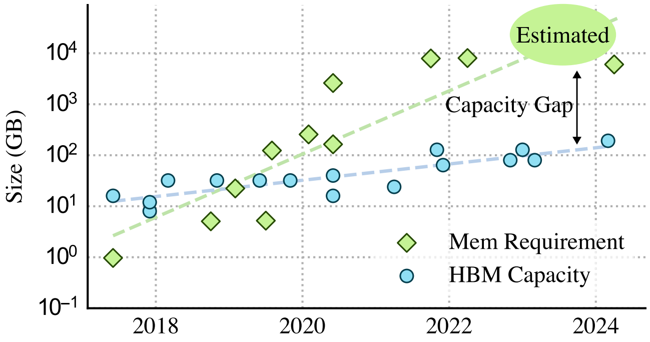


Fig. 1. Memory requirement for training LLMs and HBM capacity per device over time – We assume 16-bit precision for LLM parameters and gradients and 32-bit precision for optimizer state. HBM capacity data is adapted from [15].

I. Introduction

In RECENT years, modern high-performance compute devices such as GPUs and TPUs have largely shifted to using high bandwidth memory (HBM) due to its superior memory bandwidth compared to DDR and GDDR. For instance, in 2017, NVIDIA's V100 GPU introduced HBM2 with 900 GB/s bandwidth and 16 GB capacity. More recently, HBM3e has further pushed these limits, offering 8 TB/s bandwidth and 192 GB capacity in the NVIDIA GB200 GPU. Fig. 1 shows the trend of HBM capacity per device over time. While the capacity per HBM stack has grown from 4 GB to 24 GB over the past seven years, the aggregated HBM capacity remains fundamentally limited by the number of stacks per compute chip. The HBM stacks are connected to the compute chip via short-reach chiplet I/Os, and the number of HBM stacks is ultimately limited by the perimeter of the compute chip. Other memory expansion solutions, such as CXL memory modules [9], [12], come at the cost of reduced bandwidth.

Meanwhile, large language models (LLMs) have demonstrated remarkable capabilities across a broad range of applications [2], [4]. Driven by the need for higher accuracy and the ability to perform more sophisticated tasks, the size of LLMs continues to increase and has recently surged into trillions of parameters [5] (see Fig. 1). This poses significant challenges for the memory capacity of compute devices. For instance, training the GPT-3 model with 175 billion parameters [2], requires approximately 3 TB of memory for storing the model parameters, gradients, and optimizer state. Today's larger models, like Switch Transformer [5], can have 1–2 trillion parameters and require up to 32 TB of memory to train. Early LLMs fit within a single compute device, enabling efficient scaling of training through *data parallelism*. However, today's largest models far exceed single-device memory capacity, and thus require the use of *model parallelism*, including tensor parallelism [13] and pipeline parallelism [10], [11]. Tensor parallelism splits the computation of each layer across multiple devices, while pipeline parallelism segments the model into different stages processed sequentially. However, these parallelism strategies introduce new complications. With tensor parallelism, the frequent all-reduce communication between devices can reduce overall efficiency. Similarly, pipeline parallelism suffers from bubble inefficiencies where stages of the pipeline are underutilized, particularly during the ramp-up and ramp-down phases of the pipeline. This leads to the key observation that motivates this work: *Efficiently exploiting model parallelism for LLM training is largely limited by per-device memory capacity.*

In this letter, we propose a novel memory architecture using silicon photonic interconnects to expand the memory capacity and bandwidth of compute devices. We introduce optically connected multi-stack HBM modules, a separate chip package with multiple HBM stacks and connected to the compute chip via co-packaged optics. With co-packaged optics, we extend the HBM memory system off the compute interposer, circumventing the chip packaging constraint and allowing more HBM stacks to be connected to the compute chip. In an augmented A100 system, we achieve 576 GB of memory capacity and 12 TB/s of bandwidth using the same HBM technology. Our system improves model FLOPS utilization (MFU) by up to 1.4× for trillion-parameter LLM training. In addition, the increased bandwidth of our system can also benefit LLM inference, improving decoding performance by up to 4.2× with sufficient L2 bandwidth.

Received 14 October 2024; revised 1 February 2025; accepted 3 February 2025. Date of publication 18 February 2025; date of current version 26 February 2025. This work was supported in part by NSF PPoSS under Grant 2118709. This material is based upon research supported by the Air Force Research Laboratory (AFRL) and Defense Advanced Research Projects Agency (DARPA) under Grant FA8650-18-2-7862. *(Corresponding author: Yanghui Ou.)*

Yanghui Ou and Christopher Batten are with Cornell University, Ithaca, NY 14853 USA (e-mail: yo96@cornell.edu; cbatten@cornell.edu).

Hengrui Zhang and David Wentzlaff are with Princeton University, Princeton, NJ 08544 USA (e-mail: hengrui.zhang@princeton.edu; wentzlaf@princeton.edu).

Austin Rovinski is with New York University, New York, NY 10012 USA (e-mail: rovinski@nyu.edu).

Digital Object Identifier 10.1109/LCA.2025.3540058

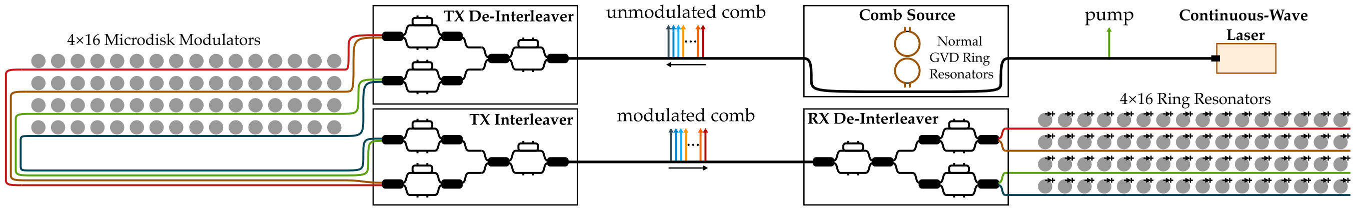


Fig. 2. Optical channel datapath.

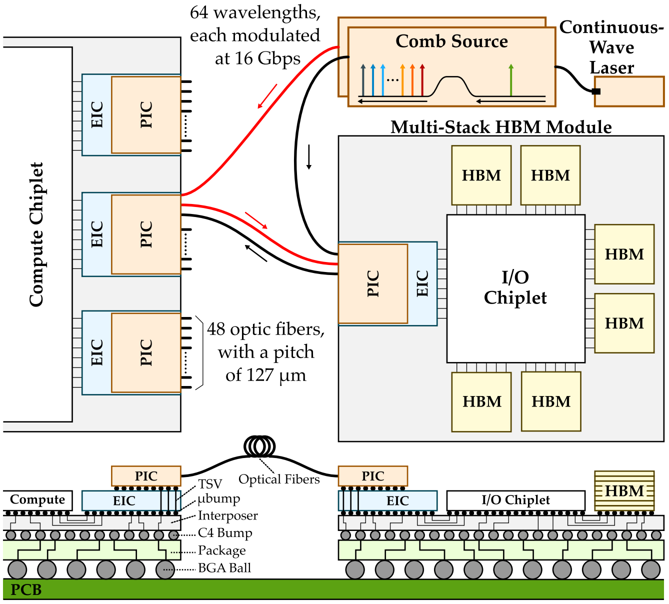
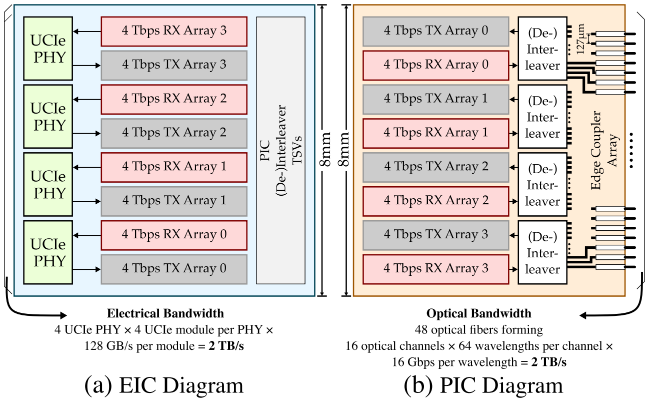


Fig. 3. Example system architecture.

II. System Architecture

Fig. 3 depicts an example system architecture with optically connected multi-stack HBM modules. The proposed system consists of a compute multi-chiplet module (MCM) and multiple multi-stack HBM modules. The compute MCM incorporates six electrical interface chiplets (EICs) and six photonic interface chiplets (PICs), which are co-packaged using 3D integration. The multi-stack HBM module includes an EIC-PIC pair and is connected to the compute MCM directly via optical fibers. The system design is based on an A100-sized compute chip. Key design parameters, such as chip dimensions, optical fiber pitch, and optical bandwidth are adapted or derived from recent works on opto-electronic transceivers and MCMs [8], [17]. Given that the width of the EIC-PIC pair is close to that of an HBM stack, the six HBM stacks in the A100 GPU chip can be replaced with six EIC-PIC pairs. Each EIC-PIC pair has a total of 48 optical fibers with a pitch of 127 μm, constituting 16 optical channels. Each optical channel is comprised of three fibers: one for unmodulated comb lines, one for transmitter (TX) signals and one for receiver (RX) signals. Each signal fiber carries 64 wavelengths modulated at 16 Gb/s, resulting in a total unidirectional bandwidth of 1 Tb/s. In total, our design provides 12 TB/s of bandwidth and 576 GB of capacity using the same HBM2e technology. The system shown in Fig. 3 is just one conceptual example, and the proposed architecture can also be adapted to other packaging strategies such as embedded PIC [17], EIC on top of PIC [3], and monolithic EIC-PIC [14].

Fig. 2 shows the detailed structure of the optical datapath highlighted in Fig. 3. The TX array is driven by a comb source that generates hundreds of low-noise frequency channels (comb lines) from a continuous-wave (CW) laser. The comb lines are subdivided by two stages of de-interleavers into four buses, each containing 16 wavelengths. Each wavelength is modulated by a microdisk modulator in the TX array. The modulated wavelengths are combined by two stages of interleavers and transmitted via a single fiber. At the RX side, the modulated wavelengths are de-interleaved into four buses and sent to arrays of cascaded ring resonators that drop each wavelength onto a photodetector to generate electrical signals.



(a) EIC Diagram (b) PIC Diagram

Fig. 4. EIC and PIC architecture – mixed-signal transceivers in the EIC is based on [8]. PIC design is adapted from [17].

Fig. 4(a) shows the EIC structure. The EIC is connected to the compute or I/O chiplet via Universal Chiplet Interconnect express (UCIe) PHYs. We derive the design parameters from the 16 GT/s UCIe PHY from the UCIe 1.0 specification [16]. Each UCIe PHY contains four 128 GB/s UCIe modules. The EIC includes four UCIe PHYs to provide a total bandwidth of 2 TB/s, matching the optical bandwidth. The EIC also includes mixed-signal transceiver arrays to communicate with the PIC via high density microbumps and through-silicon vias (TSVs). Fig. 4(b) illustrates the PIC structure, adapted from [17]. It includes an edge coupler array, de-interleavers and interleavers, and transceiver arrays. The edge coupler array is used for attaching optical fibers. Each TX array contains 256 microdisk modulators and each RX array contains 256 ring resonators, corresponding to four optical channels. The transceivers in both EIC and PIC include include integrated heaters and closed loop control to ensure the devices stay at their target temperature.

Our proposed architecture does not integrate optical interconnect technology directly into the HBM stack. Instead, it maintains the conventional HBM structure while leveraging co-packaged optics to extend connectivity beyond the compute interposer. However, this design choice does introduce added packaging cost and design complexity. Future work will explore the cost effectiveness of our design.

Our design addresses two key constraints limiting the capacity and bandwidth of current HBM-based memory systems.

Constraint #1: The number of HBM stacks is restricted by the perimeter of the compute chiplet. Our proposed design overcomes this by extending the HBM stacks into separate chip packages using co-packaged optics, with slight overhead in latency and energy. Instead of direct microbump connections to HBM, the compute chiplet connects optically to multiple I/O chiplets, and each I/O chiplet is connected to multiple HBM stacks. As a result, without increasing the chip perimeter, significantly more HBM stacks can be connected to the compute chiplet.

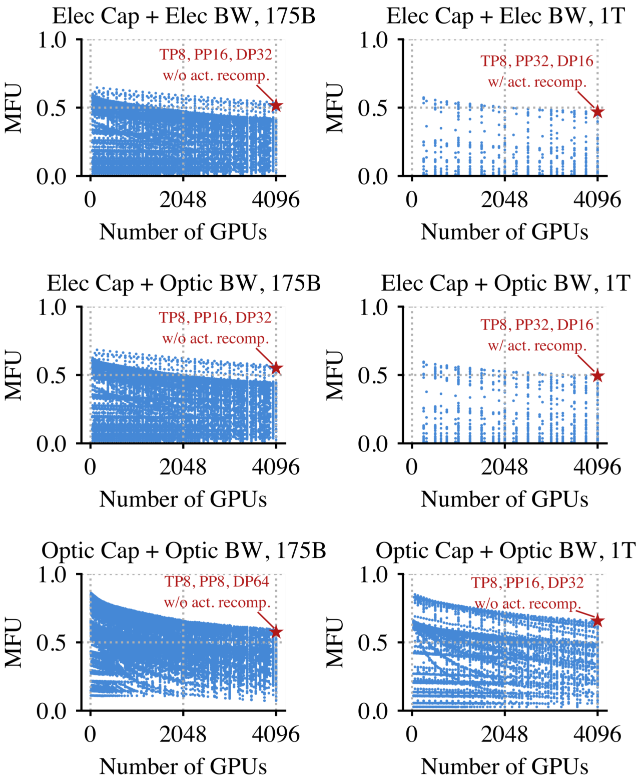
Constraint #2: The HBM interface operates below the maximum possible data rate of chiplet I/O. Current compute chiplets use HBM PHYs to communicate with the HBM stacks at up to 9.6 Gb/s per microbump (HBM3e). However, state-of-the-art chiplet interconnects like UCIe can reach 16 Gb/s or even 32 Gb/s under similar bump pitch. The bandwidth of HBM PHYs are constrained by the DRAM speed rather than the data rate of microbumps. By replacing the HBM stacks with EIC-PIC pairs, our design leverages faster UCIe PHYs to achieve a higher bandwidth. In the multi-stack HBM module, multiple HBM stacks can be accessed in parallel to match the optical bandwidth.

III. Evaluation

We evaluate our proposed system using LLMCompass [18], a performance modeling framework for LLM workloads. We model an 8-GPU A100 compute node as the baseline system. We create a model of our proposed memory system in LLMCompass and integrate it into the A100 model. The memory system, as is detailed in Section II, offers 12 TB/s of memory bandwidth and 576 GB of memory capacity. Both training and inference performance are evaluated.

Fig. 5 shows the evaluation results for modeling training of a 175-billion and a 1-trillion parameter LLM. We use the kernel-level performance model to simulate the execution time of the forward and backward pass. We generate different pipeline schedules based on the memory capacity constraints and create an InfiniBand network model in LLMCompass to simulate the communication time for data parallelism. We compare the achieved MFU of three different systems: the baseline A100 model, the A100 model with only the bandwidth enhancement of our design, and the A100 model with optically connected multi-stack HBM modules. We sweep the number of GPUs from 1 to 4096, and each point in the plot represents a possible mapping with certain degree of tensor, pipeline, and data parallelism. We can see that the improvements in MFU mainly benefit from the capacity enhancement as opposed to bandwidth. This is because the main operations in the training process, matrix-matrix multiplications, are compute-bound. Our design particularly benefited the training of the 1-trillion parameter LLM by offering higher memory capacity. At 4096 GPUs, the best mapping with our design achieves a 1.4× improvement in MFU. The baseline system, constrained by the memory capacity, has to use more model parallelism to partition the model parameters, gradients, and optimizer state. It also requires activation recomputation during the backward pass since it does not have sufficient memory to hold the activations, which leads to reduced MFU.

We compare the per-layer latency of the prefill and decode stages for the 175-billion and 1-trillion parameter LLM. Since our design exhibits bandwidth inversion, offering 12 TB/s bandwidth which is well above the 7 TB/s L2 bandwidth of the baseline A100, we also evaluate a system with an aggressive 24 TB/s L2 bandwidth to fully utilize the optical bandwidth. As is shown in Fig. 6, for prefill, our design with aggressive L2 bandwidth achieves an average 1.14× speedup across different sequence lengths for both the 175B and 1 T model. For decoding, our design with aggressive L2 achieves 3.17× speedup for the 175B model and 4.23× speedup for the 1 T model on average across different context lengths. Without the L2 modification, the speedup is 1.53× and 1.67×



Elec Cap =80 GB, Elec BW = 2 TB/s, Optic Cap = 576 GB, Optic BW = 12 TB/s.
175B = 96 layers, 96 attention heads, and an embedding size of 12288.
1T = 128 layers, 160 attention heads, and an embedding size of 25600.
TP = tensor parallelism, PP = pipeline parallelism, DP = data parallelism.
act. recomp. = activation recomputation. All experiments are performed using a sequence length of 2048 and a batch size of 4096.

Fig. 5. Evaluation results for training – each dot represents a valid training configuration for a given number of GPUs. Optimal configurations at 4096 GPUs are highlighted.

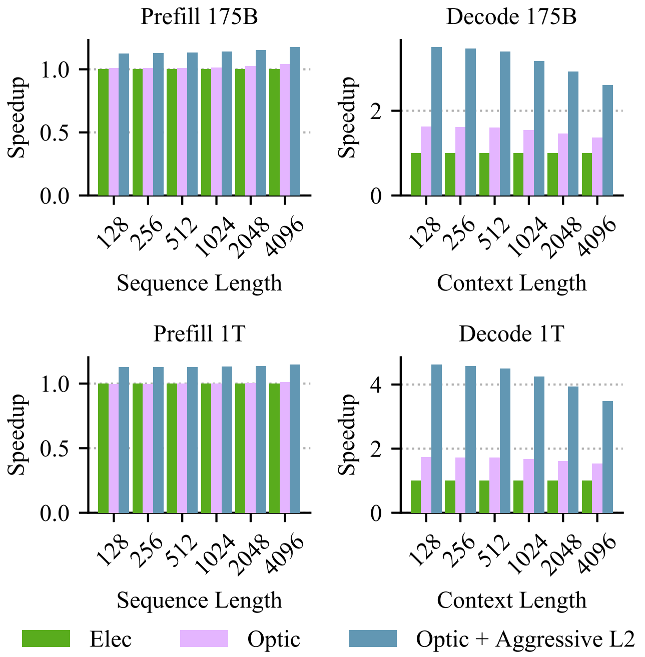


Fig. 6. Per layer speedup for inference – elec = A100 GPU, Optic = A100 GPU with our design, Optic + Aggressive L2 = A100 GPU with our design and 24 TB/s L2 bandwidth. Both prefill and decode is simulated with batch size = 8.

respectively, indicating that our design requires rethinking the memory hierarchy design to fully utilize the massive memory bandwidth offered by the optically connected multi-stack HBM modules. The capacity enhancement of our design can also potentially benefit the latency and throughput of LLM inference by allowing larger sequence length, batch size, KV cache size, as well as enabling more optimal parallelization strategies. Future work will evaluate the impact of our design on end-to-end LLM inference systems.

IV. Related Work

Beamer et al. propose PIDRAM, a photonically interconnected DRAM architecture that uses monolithically integrated silicon photonics to address the bandwidth and power limitations of electrical DDR-based memory systems [1]. Our work is distinct from PIDRAM in that we use 3D integrated silicon photonics to augment the HBM-based memory system.

Khani et al. introduce SiP-ML, which uses silicon photonics to create a flat network topology for inter-GPU communication [7]. Wu et al. propose SiPAC, co-designing an inter-GPU silicon photonic interconnect and a collective communication algorithm to accelerate distributed deep learning. Our work is complementary to SiP-ML and SiPAC in that our work explores optical interconnects between the compute chip and the HBM-based main memory.

Gonzalez et al. propose an optically connected memory architecture for disaggregated data centers, using silicon photonics to create high-bandwidth, low-latency optical links between different resource pools [6]. The GPU memory system is not modified. In contrast, our work leverages co-packaged optics to augment the HBM-based memory for compute devices.

V. Conclusion

We propose optically connected multi-stack HBM modules to enhance the capacity and bandwidth of HBM-based memory systems. Utilizing co-packaged optics, our design connects compute devices to multiple off-chip HBM stacks. Our evaluations show significant improvements in memory capacity and bandwidth, enhancing the training and inference efficiency for large-scale LLMs. The results also suggest that current memory hierarchy designs need to be reconsidered to fully exploit the advantages of optical interconnects.

Acknowledgments

The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Air Force Research Laboratory (AFRL) and Defense Advanced Research Projects Agency (DARPA) or the U.S. Government.

References

[1] S. Beamer et al., “Re-architecting DRAM memory systems with monolithically integrated silicon photonics,” in *Proc. Int. Symp. Comput. Archit.*, 2010, pp. 129–140.

[2] T. B. Brown et al., “Language models are few-shot learners,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.

[3] S. Daudlin et al., “3D photonics for ultra-low energy, high bandwidth-density chip data links,” *Comput. Res. Repository*, 2023, *arXiv:2310.01615*.

[4] J. Devlin et al., “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *North American Chapter of the Association for Computational Linguistics*. Cambridge, MA, USA: MIT Press, 2019.

[5] W. Fedus, B. Zoph, and N. Shazeer, “Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity,” *ACM J. Mach. Learn. Res.*, vol. 23, pp. 5232–5270, 2022.

[6] J. Gonzalez et al., “Optically connected memory for disaggregated data centers,” *J. Parallel Distrib. Comput.*, vol. 163, pp. 300–312, 2022.

[7] M. Khani et al., “SiP-ML: High-bandwidth optical network interconnects for machine learning training,” in *Proc. ACM SIGCOMM Conf.*, 2021, pp. 657–675.

[8] D. Khilwani et al., “3D-integrated, low power, high bandwidth density opto-electronic transceiver,” in *Proc. Int. Conf. Circuits Syst.*, 2024, pp. 1–5.

[9] H. Li et al., “Pond: CXL-based memory pooling systems for cloud platforms,” in *Proc. Int. Conf. Architectural Support Program. Lang. Operating Syst.*, 2024, pp. 574–587.

[10] D. Narayanan et al., “PipeDream: Generalized Pipeline Parallelism for DNN training,” in *Proc. 27th ACM Symp. Operating Syst. Princ.*, 2019, pp. 1–15.

[11] D. Narayanan et al., “Memory-efficient pipeline-parallel DNN training,” in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 7937–7947.

[12] S.-S. Park et al., “An LPDDR-based CXL-PNM platform for TCO-efficient inference of transformer-based large language models,” in *Proc. Int. Symp. High-Perform. Comput. Architecture*, 2024, pp. 970–982.

[13] M. Shoeybi et al., “Megatron-LM: Training multi-billion parameter language models using model parallelism,” *Comput. Res. Repository*, 2019, *arXiv:1909.08053*.

[14] C. Sun et al., “TeraPHY: An O-band WDM electro-optic platform for low power, terabit/s optical I/O,” in *Proc. IEEE Symp. VLSI Technol.*, 2020, pp. 1–2.

[15] TechPowerUpGPU Database, “Online database,” 2024. [Online]. Available: https://www.techpowerup.com/gpu-specs

[16] UCIe 1.0 specification, 2024. [Online]. Available: https://www.uciexpress.org/specifications

[17] Y. Wang et al., “Silicon photonics chip I/O for ultra high-bandwidth and energy- efficient die-to-die connectivity,” in *Proc. IEEE Custom Integr. Circuits Conf.*, 2024, pp. 1–8.

[18] H. Zhang et al., “LLMCompass: Enabling efficient hardware design for large language model inference,” in *Proc. Int. Symp. Comput. Archit.*, 2024, pp. 1080–1096.