

TensorDIMM: A Practical Near-Memory Processing Architecture for Embeddings and Tensor Operations in Deep Learning

Youngeun Kwon
School of Electrical Engineering
KAIST
yekwon@kaist.ac.kr

Yunjae Lee
School of Electrical Engineering
KAIST
yunjae408@kaist.ac.kr

Minsoo Rhu
School of Electrical Engineering
KAIST
mrhu@kaist.ac.kr

ABSTRACT

Recent studies from several hyperscalars pinpoint to embedding layers as the most memory-intensive deep learning (DL) algorithm being deployed in today's datacenters. This paper addresses the memory capacity and bandwidth challenges of embedding layers and the associated tensor operations. We present our vertically integrated hardware/software co-design, which includes a custom DIMM module enhanced with near-memory processing cores tailored for DL tensor operations. These custom DIMMs are populated inside a GPU-centric system interconnect as a remote memory pool, allowing GPUs to utilize for scalable memory bandwidth and capacity expansion. A prototype implementation of our proposal on real DL systems shows an average 6.2–17.6× performance improvement on state-of-the-art DNN-based recommender systems.

CCS CONCEPTS

• Computer systems organization → Parallel architectures.

KEYWORDS

System architecture, memory architecture, near-memory processing, DIMM, machine learning, neural network, graphics processing unit (GPU), neural processing unit (NPU)

ACM Reference Format:

Youngeun Kwon, Yunjae Lee, and Minsoo Rhu. 2019. TensorDIMM: A Practical Near-Memory Processing Architecture for Embeddings and Tensor Operations in Deep Learning. In *MICRO '52: The 52nd Annual IEEE/ACM International Symposium on Microarchitecture, October 12–16, 2019, Columbus, OH, USA*. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3352460.3358284>

1 INTRODUCTION

Machine learning (ML) algorithms based on deep neural networks (DNNs), also known as deep learning (DL), is scaling up rapidly. To satisfy the computation needs of DL practitioners, GPUs or custom-designed accelerators for DNNs, also known as neural processing units (NPUs) are widely being deployed for accelerating DL workloads. Despite prior studies on enhancing the compute throughput of GPUs/NPUs [3, 5, 8, 10, 11, 33, 34, 63], a research theme that has

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MICRO '52, October 12–16, 2019, Columbus, OH, USA

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6938-1/19/10...\$15.00

<https://doi.org/10.1145/3352460.3358284>

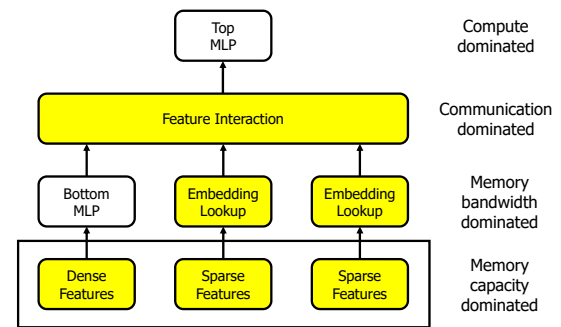


Figure 1: Topological structure of emerging DL applications. The figure is reproduced from Facebook's keynote speech at the Open Compute Project summit 2019 [67], which calls out for architectural solutions providing “High memory bandwidth and capacity for embeddings”. This paper specifically addresses this important memory wall problem in emerging DL applications: i.e., the *non-MLP* portions (yellow) in this figure.

received relatively less attention is how computer system architects should go about tackling the “memory wall” problem in DL: emerging DL algorithms demand both high memory capacity and bandwidth, limiting the types of applications that can be deployed under practical constraints. In particular, recent studies from several hyperscalars [15, 24, 27, 52, 64] pinpoint to *embedding lookups* and *tensor manipulations* (aka *embedding layers*, Section 2.3) as the most memory-intensive algorithm deployed in today's datacenters, already reaching several hundreds of GBs of memory footprint, even for inference. Common wisdom in conventional DL workloads (e.g., convolutional and recurrent neural networks, CNNs and RNNs) was that convolutions or matrix-multiplications account for the majority of inference time [25, 63]. However, emerging DL applications employing embedding layers exhibit drastically different characteristics as the embedding lookups and tensor manipulations (i.e., *feature interaction* in Figure 1) account for a significant fraction of execution time. Facebook for instance states that embedding layers take up 34% of the execution time of all DL workloads deployed in their datacenters [64].

Given this landscape, this paper focuses on addressing the memory capacity and bandwidth challenges of embedding layers (Figure 1). Specifically, we focus our attention on *recommender systems* [1] using embeddings which is one of the most common DL workloads deployed in today's datacenters for numerous application domains like advertisements, movie/music recommendations, news feed, and etc. As detailed in Section 2.3, the model size of embedding layers (typically several hundreds of GBs [27, 64]) far

exceeds the memory capacity of GPUs. As a result, the solution vendors take is to store the entire embedding lookup table inside the capacity-optimized, low-bandwidth CPU memory and deploy the application 1) by only using CPUs for the entire computation of DNN inference, or 2) by employing a hybrid CPU-GPU approach where the embedding lookups are conducted on the CPU but the rest are handled on the GPU. We observe that both these approaches leave significant performance left on the table, experiencing an average 7.3–20.9× slowdown compared to a hypothetical, GPU-only version which assumes the entire embeddings can be stored in the GPU memory (Section 3.2). Through a detailed application characterization study, we root-cause the reason behind such performance loss to the following factors. First, the embedding vectors are read out using the low-bandwidth CPU memory, incurring significant latency overheads compared to when the embeddings are read out using the bandwidth-optimized GPU memory. Second, the low computation throughput of CPUs can significantly slowdown the computation-heavy DNN execution step when solely relying on CPUs, whereas the hybrid CPU-GPU version can suffer from the latencies in copying the embeddings from CPU to GPU memory over the thin PCIe channel.

To tackle these challenges, we present a vertically integrated, hardware/software co-design that fundamentally addresses the memory (capacity and bandwidth) wall problem of embedding layers. Our proposal encompasses multiple levels in the hardware and software stack as detailed below.

(Micro)architecture. We present TensorDIMM, which is based on commodity buffered DIMMs but further enhanced with near-memory processing (NMP) units customized for key DL tensor operations, such as embedding *gathers* and *reductions*. The NMP units in TensorDIMM are designed to conduct embedding gathers and reduction operations “near-memory” which drastically reduces the latency in fetching the embedding vectors and reducing them, providing significant improvements in effective communication bandwidth and performance. Additionally, our TensorDIMM leverages commodity DRAM devices as-is, so another key advantage of our proposal as opposed to prior in-memory architectures [72, 76, 77] is its practicality and ease of implementation.

ISA extension and runtime system. Building on top of our TensorDIMM architecture, we propose a custom tensor ISA and runtime system that provides *scalable* memory bandwidth and capacity expansion for embedding layers. Our proposal entails 1) a carefully designed ISA tailored for DL tensor operations (TensorISA), 2) an efficient address mapping scheme for embeddings, and 3) a runtime system that effectively utilizes TensorDIMMs for tensor operations. The TensorISA has been designed from the ground-up to conduct key DL tensor operations in a memory bandwidth efficient manner. An important challenge with conventional memory systems is that, regardless of how many DIMMs/ranks are physically available per each memory channel, the maximum memory bandwidth provided to the memory controller is fixed. TensorISA has been carefully co-designed with both TensorDIMM and the address mapping scheme so that the aggregate memory bandwidth provided to our NMP units increases proportional to the number of TensorDIMMs. In effect, our proposal offers a platform for *scalable* memory bandwidth expansion for embedding layers. Compared to the baseline system,

our default TensorDIMM configuration offers an average 4× increase in memory bandwidth for key DL tensor operations.

System architecture. Our final proposition is to aggregate a pool of TensorDIMM modules as a *disaggregated memory node* (henceforth referred to as TensorNode) in order to provide scalable memory capacity expansion. A key to our proposal is that TensorNode is interfaced inside the NVLINK-compatible, GPU-side high-bandwidth system interconnect¹. In state-of-the-art DL systems, the GPUs are connected to a high-bandwidth switch such as NVSwitch [62] which allows high-bandwidth, low-latency data transfers between any pair of GPUs. Our proposal utilizes a pooled memory architecture “inside” the high-bandwidth GPU-side interconnect, which is fully populated with capacity-optimized memory DIMMs – but in our case, the TensorDIMMs. The benefits of interfacing TensorNode within the GPU interconnect is clear: by storing the embedding lookup tables inside the TensorNode, GPUs can copy in/out the embeddings much faster than the conventional CPU-GPU based approaches (i.e., approximately 9× faster than PCIe, assuming NVLINK(v2) [61]). Furthermore, coupled with the memory bandwidth amplification effects of TensorISA, our TensorNode offers scalable expansion of “both” memory capacity and bandwidth. Overall, our vertically integrated solution provides an average 6.2–15.0× and 8.9–17.6× speedup in inference time compared to the CPU-only and hybrid CPU-GPU implementation of recommender systems, respectively. To summarize our **key contributions**:

- The computer architecture community has so far primarily focused on accelerating the computationally intensive, “dense” DNN layers (e.g., CNNs/RNNs/MLPs). To the best of our knowledge, this work is the first to identify, analyze, and explore architectural solutions for “sparse” embedding layers, a highly important building block with significant industrial importance for emerging DL applications.
- We propose TensorDIMM, a practical near-memory processing architecture built on top of commodity DRAMs which offers scalable increase in both memory capacity and bandwidth for embedding gathers and tensor operations.
- We present TensorNode, a TensorDIMM-based disaggregated memory system for DL. The efficiency of our solution is demonstrated as a proof-of-concept software prototype on a high-end GPU system, achieving significant performance improvements than conventional approaches.

2 BACKGROUND

2.1 Buffered DRAM Modules

In order to balance memory capacity and bandwidth, commodity DRAM devices that are utilized in unison compose a *rank*. One or more ranks are packaged into a memory module, the most popular form factor being the dual-inline memory module (DIMM) which has 64 data I/O (DQ) pins. Because a memory channel is typically connected to multiple DIMMs, high-end CPU memory controllers often need to drive hundreds of DRAM devices in order to deliver the command/address (C/A) signals through the memory channel.

¹While we demonstrate TensorDIMM’s merits under the context of a high-bandwidth GPU-side system interconnect, the effectiveness of our proposal remains intact for NPUs (e.g., Facebook’s Zion interconnect [19]).

Because modern DRAMs operate in the GHz range, having hundreds of DRAM devices be driven by handful of memory controllers leads to signal integrity issues. Consequently, server-class DIMMs typically employ a *buffer device* per each DIMM (e.g., registered DIMM [71] or load-reduced DIMM [28]) which is used to repeat the C/A signals to reduce the high capacitive load and resolve signal integrity issues. Several prior work from both industry [46] and in academia [4, 6, 21] have explored the possibility of utilizing this buffer device space to add custom logic designs to address specific application needs. IBM’s Centaur DIMM [46] for instance utilizes a buffer device to add a 16 MB eDRAM L4 cache and a custom interface between DDR PHY and IBM’s proprietary memory interface.

2.2 System Architectures for DL

As the complexity of DL applications skyrocket, there has been a growing trend towards dense, scaled-up system node design with multiple PCIe-attached co-processor devices (i.e., DL accelerators such as GPUs/NPUs [23, 60]) to address the problem size growth. A multi-accelerator device solution typically works on the same problem in parallel with occasional inter-device communication to share intermediate data [37]. Because such inter-device communication often lies on the critical path of parallelized DL applications, system vendors are employing high-bandwidth interconnection fabrics that utilize custom high-bandwidth signaling links (e.g., NVIDIA’s DGX-2 [58] or Facebook’s Zion system interconnect fabric [19]). NVIDIA’s DGX-2 [58] for instance contains 16 GPUs, all of which are interconnected using a NVLINK-compatible high-radix (cross-bar) switch called NVSwitch [62]. NVLINK provides 25 GB/sec of full-duplex uni-directional bandwidth per link [61], so any given GPU within DGX-2 can communicate with any other GPU at the full uni-directional bandwidth up to 150 GB/sec via NVSwitch. Compared to the thin uni-directional bandwidth of 16 GB/sec (x16) under the CPU-GPU PCIe(v3) bus, such high-bandwidth GPU-side interconnect enables an order of magnitude faster data transfers.

2.3 DL Applications with Embeddings

DNN-based recommender systems. Conventional DL applications for inference (e.g., CNNs/RNNs) generally share a common property where its overall memory footprint fits within the (tens of GBs of) GPU/NPU physical memory. However, recent studies from several hyperscalars [15, 24, 27, 52, 64] call out for imminent system-level challenges in emerging DL workloads that are extremely memory (capacity and bandwidth) limited. Specifically, these hyperscalars pinpoint to *embedding layers* as the most memory-intensive algorithm deployed in their datacenters. One of the most widely deployed DL application using embeddings is the *recommender system* [1], which is used in numerous application domains such as advertisements (Amazon, Google, Ebay), social networking service (Facebook, Instagram), movie/music/image recommendations (YouTube, Spotify, Fox, Pinterest), news feed (LinkedIn), and many others. Recommendation is typically formulated as a problem of predicting the probability of a certain event (e.g., the probability of a Facebook user clicking “like” for a particular post), where a ML model estimates the likelihood of one or more events happening at the same time. Events or items with the highest probability are ranked higher and recommended to the user.

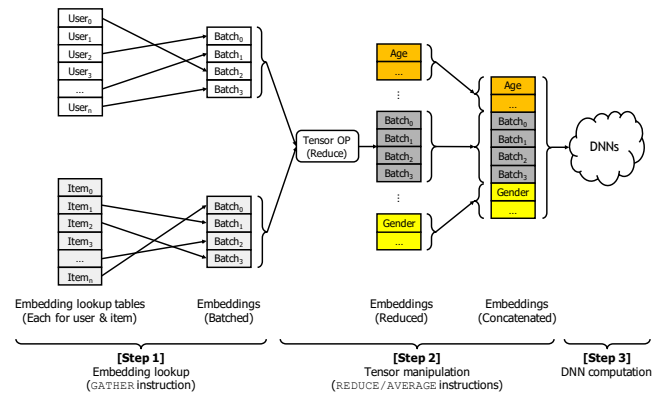


Figure 2: A DNN based recommender system. The embedding layer typically consists of the following two steps. (1) The embedding lookup stage where embeddings are “gathered” from the (potentially multiple, two tables in this example) look-up tables up to its batch size and form (batched) embedding tensors. These tensors go through several tensor manipulation operations to form the final embedding tensor to be fed into the DNNs. Our proposal utilizes custom tensor ISA extensions (GATHER/REDUCE/AVERAGE) to accelerate this process, which we detail in Section 4.4.

Without going into a comprehensive review of numerous prior literature on recommendation models, we emphasize that current state-of-the-art recommender systems have evolved into utilizing (not surprisingly) DNNs. While there exists variations regarding how the DNNs are constructed, a commonly employed topological structure for DNN-based recommender systems is the neural network based collaborative filtering algorithm [26]. Further advances led to the development of more complex models with wider and deeper vector dimensions, successfully being applied and deployed in commercial user-facing products [64].

Embedding lookups and tensor manipulation. Figure 2 illustrates the usage of embeddings in recommender systems that incorporate DNNs [26]. The inputs to the DNNs (which are typically constructed using fully-connected layers or multi-layer perceptrons, FCs or MLPs) are constructed as a combination of dense and sparse features. Dense features are commonly represented as a vector of real numbers whereas sparse features are initially represented as indices of one-hot encoded vectors. These one-hot indices are used to query the *embedding lookup table* to project the sparse indices into a dense vector dimension space. The contents stored in these lookup tables are called *embeddings*, which are trained to extract deep learning features (e.g., Facebook trains embeddings to extract information regarding what pages a particular user liked, which are utilized to recommend relevant contents or posts to the user [64]). The embeddings read out of the lookup table are combined with other dense embeddings (*feature interaction* in Figure 1) from other lookup tables using tensor concatenation or *tensor “reductions”* such as element-wise additions/multiplications/averages/etc to generate an output tensor, which is then forwarded to the DNNs to derive the final event probability (Figure 2).

Memory capacity limits of embedding layers. A key reason why embedding layers consume significant memory capacity is because each user (or each item) requires a unique embedding vector

inside the lookup table. The total number of embeddings therefore scales proportional to the number of users/items, rendering the overall memory footprint to exceed several hundreds of GBs just to keep the model weights themselves, even for inference. Despite its high memory requirements, embedding layers are favored in DL applications because it helps improve the model quality: given enough memory capacity, users seek to further increase the model size of these embedding layers using *larger embedding dimensions* or by using *multiple embedding tables* to combine multiple dense features using tensor reduction operations.

3 MOTIVATION

3.1 Memory (Capacity) Scaling Challenges

High-end GPUs or NPUs commonly employ a bandwidth-optimized, on-package 3D stacked memory (e.g., HBM [32] or HMC [48]) in order to deliver the highest possible memory bandwidth to the on-chip compute units. Compared to capacity-optimized DDRx most commonly adopted in CPU servers, the bandwidth-optimized stacked memory is capacity-limited, only available with several tens of GBs of storage. While one might expect future solutions in this line of product to benefit from higher memory density, there are technological constraints and challenges in increasing the capacity of these 3D stacked memory in a *scalable* manner. First, stacking more DRAM layers vertically is constrained by chip pinout required to drive the added DRAM stacks, its wireability on top of silicon interposers, and thermal constraints. Second, current generation of GPUs/NPUs are already close to the reticle limits of processor die size² so adding more 3D stack modules within a package inevitably leads to sacrificing area budget for compute units.

3.2 Memory Limits in Recommender System

As discussed in Section 2.3, the model size of embedding lookup tables is in the order of several hundreds of GBs, far exceeding the memory capacity limits of GPUs/NPUs. Due to the memory scaling limits of on-package stacked DRAMs, a solution vendors take today is to first store the embedding lookup tables in the CPU and read out the embeddings using the (capacity-optimized but bandwidth-limited) CPU memory. Two possible implementations beyond this step are as follows. The CPU-only version (CPU-only) goes through the rest of the inference process using the CPU without relying upon the GPU. A *hybrid* CPU-GPU approach (CPU-GPU) [59] on the other hand copies the embeddings to the GPU memory over PCIe using `cudaMemcpy`, and once the CPU→GPU data transfer is complete, the GPU initiates various tensor manipulation operations to form the input tensors to the DNN, followed by the actual DNN computation step (Figure 2).

Given this landscape, DL practitioners as well as system designers are faced with a conundrum when trying to deploy recommender systems for inference. From a DL algorithm developer's perspective, you seek to add more embeddings (i.e., more embedding lookup tables to combine various embedding vectors using tensor

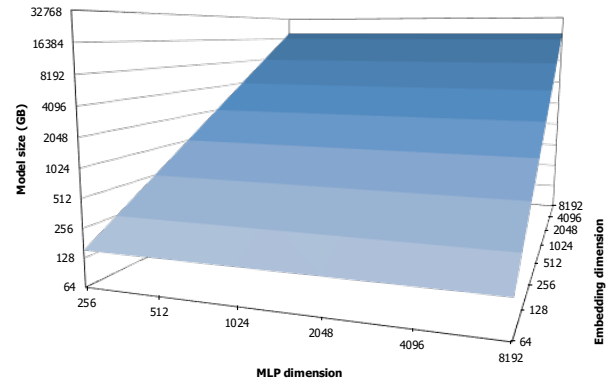


Figure 3: Model size growth of neural collaborative filtering (NCF [26]) based recommender system when the MLP layer dimension size (x-axis) and the embedding vector dimension size (y-axis) is scaled up. Experiment assumes the embedding lookup table contains 5 million users and 5 million items per each lookup table. As shown, larger embeddings rather than larger MLP dimension size cause a much more dramatic increase in model size.

manipulations, such as tensor reduction) and increase embedding dimensions (i.e., larger embeddings) for complex feature interactions as it improves model quality. Unfortunately, fulfilling the needs of these algorithm developers bloats up overall memory usage (Figure 3) and inevitably results in resorting to the capacity-optimized CPU memory to store the embedding lookup tables. Through a detailed characterization study, we root-cause the following three factors as **key limiters** of prior approaches (Figure 4) relying on CPU memory for storing embeddings:

- (1) As embedding lookup tables are stored in the low-bandwidth CPU memory, reading out embeddings (i.e., the embedding *gather* operation) adds significant latency than an unbuildable, oracular GPU-only version (GPU-only) which assumes infinite GPU memory capacity. Under GPU-only, the entire embeddings can be stored locally in the high-bandwidth GPU memory so gathering embeddings can be done much faster than when using CPU memory. This is because gathering embeddings is a memory bandwidth-limited operation.
- (2) CPU-only versions can sidestep the added latency that hybrid CPU-GPU versions experience during the PCIe communication process of transferring embeddings, but the (relatively) low computation throughput of CPUs can significantly lengthen the DNN computation step.
- (3) Hybrid CPU-GPU versions on the other hand can reduce the DNN computation latency, but comes at the cost of additional CPU→GPU communication latency when copying the embeddings over PCIe using `cudaMemcpy`.

3.3 Our Goal: A Scalable Memory System

Overall, the key challenge rises from the fact that the source operands of key tensor manipulations (i.e., the embeddings subject for tensor concatenation or tensor reduction) are initially located inside the embedding lookup tables, all of which are stored inside the capacity-optimized but bandwidth-limited CPU memory. Consequently, prior

²NVIDIA V100 for instance has already reached the reticle limits of 815 mm² die area, forcing researchers to explore alternative options such as multi-chip-module solutions [5] to continue computational scaling.

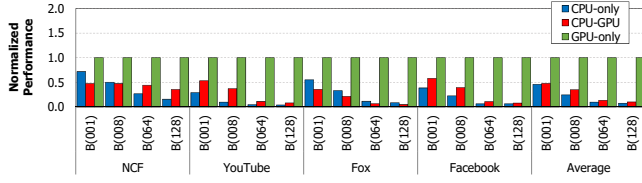


Figure 4: Performance of baseline CPU-only and hybrid CPU-GPU versions of recommender system, normalized to an oracular GPU-only version. B(N) represents an inference with batch size N. CPU-only exhibits some performance advantage than the CPU-GPU version for certain low batch inference scenarios, but both CPU-only and CPU-GPU suffers from significant performance loss compared to oracular GPU-only. Section 5 details our evaluation methodology.

solutions suffer from *low-bandwidth embedding gather operations* over CPU memory which adds significant latency. Furthermore, CPU-only and CPU-GPU versions need to trade-off the *computational bottleneck* of low-throughput CPUs over the *communication bottleneck* of PCIe, which adds additional latency overheads (Figure 1). What is more troubling is the fact future projections of DL applications utilizing embedding layers assume even larger number of embedding lookups and larger embeddings themselves [27, 64], with complex tensor manipulations to combine embedding features to improve algorithmic performance. Overall, both current and future memory requirements of embedding layers point to an urgent need for a system-level solution that provides *scalable* memory capacity and bandwidth expansion. In the next section, we detail our proposed solution that addresses both the computational bottleneck of low-throughput CPUs and the communication bottleneck of low-bandwidth CPU-GPU data transfers.

4 TENSORDIMM: AN NMP DIMM DESIGN FOR EMBEDDINGS & TENSOR OPS

4.1 Proposed Approach

Our proposal is based on the following **key observations** that open up opportunities to tackle the system-level bottlenecks in deploying memory-limited recommender systems.

- (1) An important tensor operation used in combining embedding features is the element-wise operation, which is equivalent to a tensor-wide *reduction* among N embeddings. Rather than having the N embeddings individually be gathered and copied over to the GPU memory for the reduction operation to be initiated using the GPU, we can conduct the N embedding gathers and reductions all “near-memory” and copy a single, *reduced* tensor to GPU memory. Such near-memory processing (NMP) approach reduces not only the latency to gather the embeddings but also the data transfer size by a factor of N and alleviates the communication bottleneck of CPU→GPU cudaMemcpy (Figure 5).
- (2) DL system vendors are deploying GPU/NPU-centric interconnection fabrics [19, 58] that are decoupled from the legacy host-device PCIe. Such technology allows vendors to employ custom high-bandwidth links (e.g., NVLINK, providing 9×

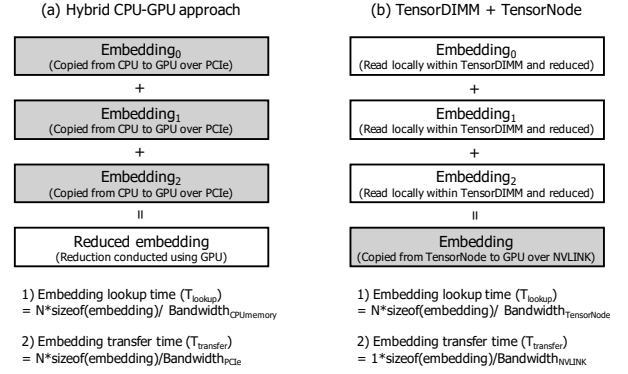


Figure 5: Our proposed approach: compared to the (a) hybrid CPU-GPU version, (b) our solution conducts embedding gathers and tensor-wide reductions locally “near-memory” first, and then transfers the *reduced* tensor over the high-bandwidth NVLINK. Consequently, our proposal significantly reduces both the latency of gathering embeddings from the lookup table (T_{lookup}) and the time taken to transfer the embeddings from the capacity-optimized CPU/TensorNode memory to the bandwidth-optimized GPU memory ($T_{transfer}$). The CPU-only version does not experience $T_{transfer}$ but suffers from significantly longer DNN execution time than CPU-GPU and our proposed design.

higher bandwidth than PCIe) for fast inter-GPU/NPU communication. But more importantly, it is possible to tightly integrate “non”-accelerator components (e.g., a disaggregated memory pool [39, 41, 42]) as separate interconnect endpoints (or *nodes*), allowing accelerators to read/write from/to these non-compute nodes using high-bandwidth links.

Based on these observations, we first propose TensorDIMM which is a custom DIMM design including an NMP core tailored for tensor “gather” and tensor “reduction” operations. We then propose a disaggregated memory system called TensorNode which is fully populated with our TensorDIMMs. Our last proposition is a software architecture that effectively parallelizes the tensor gather/reduction operation across the TensorDIMMs, providing scalable memory bandwidth and capacity expansion. The three most significant performance limiters in conventional CPU-only or hybrid CPU-GPU recommender system are 1) the embedding gather operation over the low-bandwidth CPU memory, 2) the compute-limited DNN execution using CPUs (for CPU-only approaches) and 3) the CPU→GPU embedding copy operation over the PCIe bus (for hybrid CPU-GPU), all of which adds severe latency overheads (Section 3.2). Our vertically integrated solution fundamentally addresses these problems thanks to the following three **key innovations** provided with our approach (Figure 5). First, by storing the entire embedding lookup table inside the TensorDIMMs, the embedding gather operation can be conducted using the ample memory bandwidth available across the TensorDIMMs (4× higher than CPU) rather than the low-bandwidth CPU memory. Second, the TensorDIMM NMP cores conduct the N tensor reduction operation *before* sending them to the GPU, reducing the TensorNode→GPU communication latency by a factor of N , effectively overcoming the communication bottleneck. Furthermore, the high-bandwidth TensorNode↔GPU links (i.e., NVLINK)

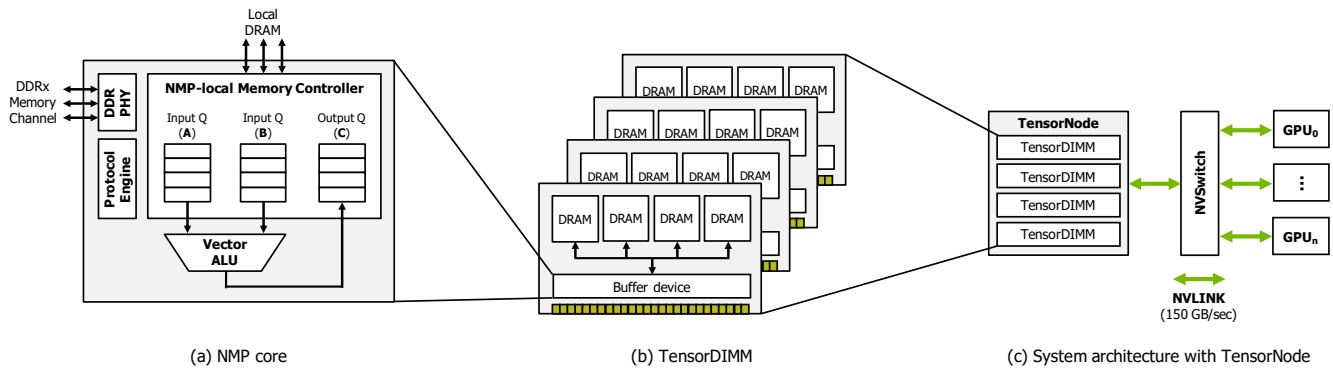


Figure 6: High-level overview of our proposed system. (a) An NMP core is implemented inside the buffer device of each (b) TensorDIMM, multiples of which are employed as a (c) disaggregated memory pool, called TensorNode. The TensorNode is integrated inside the high-bandwidth GPU-side interconnect. The combination of NVLINK and NVSwitch enables GPUs to read/write data to/from TensorNode at a communication bandwidth up to 9 \times higher than PCIe.

enable further latency reduction, proportional to the bandwidth difference between PCIe and NVLINK (approximately 9 \times). Lastly, all DNN computations are conducted using the GPU, overcoming the computation bottleneck of CPU-only implementations. As detailed in Section 6, our proposal achieves an average 6.2–15.0 \times and 8.9–17.6 \times performance improvement over the CPU-only and hybrid CPU-GPU, respectively, reaching 84% of an unbuildable, oracular GPU-only implementation with infinite memory capacity. We detail each components of our proposal below.

4.2 TensorDIMM for Near-Memory Tensor Ops

Our TensorDIMM is architected under three key design objectives in mind. First, TensorDIMM should leverage commodity DRAM chips as-is while be capable of being utilized as a normal buffered DIMM device in the event that it is not used for DL acceleration. Second, tensor reduction operations should be conducted “near-memory” using a lightweight NMP core, incurring minimum power and area overheads to the buffered DIMM device. Third, in addition to memory capacity, the amount of memory bandwidth available to the NMP cores should also scale up proportional to the number of TensorDIMM modules employed in the system.

Architecture. Figure 6(a,b) shows the TensorDIMM architecture, which consists of an NMP core and the associated DRAM chips. As depicted, TensorDIMM does not require any changes to commodity DRAMs since all modifications are limited to a buffer device within a DIMM (Section 2.1). The NMP core includes a DDR interface, a vector ALU, and an NMP-local memory controller which includes input/output SRAM queues to stage-in/out the source/destination operands of tensor operations. The DDR interface is implemented with a conventional DDR PHY and a protocol engine.

TensorDIMM usages. For *non*-DL use-cases, the processor’s memory controller sends/receives DRAM C/A and DQ signals to/from this DDR interface, which directly interacts with the DRAM chips. This allows TensorDIMM to be applicable as a normal buffered DIMM device and be utilized by conventional processor architectures for servicing load/store transactions. Upon receiving a

TensorISA instruction for tensor gather/reduction operations however, the instruction is forwarded to the NMP-local memory controller, which translates the TensorISA instruction into low-level DRAM C/A commands to be sent to the DRAM chips. Specifically, the TensorISA instruction is decoded (detailed in Section 4.4) in order to calculate the physical memory address of the target tensor’s location, and the NMP-local memory controller generates the necessary RAS/CAS/activate/precharge/etc DRAM commands to read/write data from/to the TensorDIMM DRAM chips. The data read out of DRAM chips are temporarily stored inside the input (A and B) SRAM queues until the ALU reads them out for tensor operations. In terms of NMP compute, the minimum data access granularity over the eight x8 DRAM chips is 64 bytes for a burst length of 8, which amounts to sixteen 4-byte scalar elements. As detailed in Section 4.4, the tensor operations accelerated with our NMP cores are element-wise arithmetic operations (e.g., add, subtract, average, ...) which exhibit data-level parallelism across the sixteen scalar elements. We therefore employ a 16-wide vector ALU which conducts the element-wise operation over the data read out of the (A and B) SRAM queues. The vector ALU checks for any newly submitted pair of data (64 bytes each) and if so, pops out a pair for tensor operations, the result of which is stored into the output (C) SRAM queue. The NMP memory controller checks the output queue for newly inserted results and drain them back into DRAM, finalizing the tensor reduction process. For tensor gathers, the NMP core forwards the data read out of the input queues to the output queue to be committed back into DRAM.

Implementation and overhead. TensorDIMM leverages existing DRAM chips and the associated DDR PHY interface as-is, so the additional components introduced with our TensorDIMM design are the NMP-local memory controller and the 16-wide vector ALU. In terms of the memory controller, decoding the TensorISA instruction into a series of DRAM commands is implemented as a FSM control logic so the major area/power overheads comes from the input/output SRAM queues. These buffers must be large enough to hold the bandwidth-delay product of the memory sourcing the data to remove idle periods in the output tensor generation stream.

A conservative estimate of 20 ns latency from the time the NMP-local memory controller requests data to the time it arrives at the SRAM queue is used to size the SRAM queue capacity. Assuming the baseline PC4-25600 DIMM that provide 25.6 GB/sec of memory bandwidth, the NMP core requires $(25.6 \text{ GB/sec} \times 20 \text{ ns}) = 0.5 \text{ KB}$ of SRAM queue size (1.5 KB overall, for both input/output queues). The 16-wide vector ALU is clocked at 150 MHz to provide enough computation throughput to seamlessly conduct the element-wise tensor operations over the data read out of the input (A and B) queues. Note that the size of an IBM Centaur buffer device [46] is approximately (10mm×10mm) with a TDP of 20 W. Compared to such design point, our NMP core adds negligible power and area overheads, which we quantitatively evaluate in Section 6.5.

Memory bandwidth scaling. An important design objective of TensorDIMM is to provide scalable memory bandwidth expansion for NMP tensor operations. A key challenge with conventional memory systems is that the maximum bandwidth per each memory channel is fixed (i.e., signaling bandwidth per each pin \times number of data pins per channel), regardless of the number of DIMMs (or ranks/DIMM) per channel. For instance, the maximum CPU memory bandwidth available under the baseline CPU system (i.e., NVIDIA DGX [57]) can never exceed $(25.6 \text{ GB/sec} \times 8) = 204.8 \text{ GB/sec}$ across the eight memory channels (4 channels per each socket), irrespective of the number of DIMMs actually utilized (i.e., 8 DIMMs vs. 32 DIMMs). This is because the physical memory channel bandwidth is time-multiplexed across multiple DIMMs/ranks. As detailed in the next subsection, our TensorDIMM is utilized as a basic building block in constructing a disaggregated memory system (i.e., TensorNode). The key innovation of our proposal is that, combined with our TensorISA address mapping function (Section 4.4), the amount of aggregate memory bandwidth provided to all the NMP cores within TensorNode increases proportional to the number of TensorDIMM employed: an aggregate of 819.2 GB/sec memory bandwidth assuming 32 TensorDIMMs, a 4 \times increase over the baseline CPU memory system. Such memory bandwidth scaling is possible because the NMP cores access its TensorDIMM-internal DRAM chips “locally” within its DIMM, not having to share its local memory bandwidth with other TensorDIMMs. Naturally, the more TensorDIMMs employed inside the memory pool, the larger the aggregate memory bandwidth becomes available to the NMP cores conducting embedding gathers and reductions.

4.3 System Architecture

Figure 6 provides a high-level overview of our proposed system architecture. We propose to construct a disaggregated memory pool within the high-bandwidth GPU system interconnect. Our design is referred to as TensorNode because it functions as an interconnect endpoint, or *node*. GPUs can read/write data from/to TensorNode over the NVLINK compliant PHY interface, either using fine-grained CC-NUMA or coarse-grained data transfers using P2P `cudaMemcpy`³. Figure 6(c) depicts our TensorNode design populated with multiple TensorDIMM devices. The key advantage of TensorNode is threefold. First, TensorNode provides a

platform for increasing memory capacity in a *scalable* manner as the disaggregated memory pool can independently be expanded using density-optimized DDRx, irrespective of the GPU’s local, bandwidth-optimized (but capacity-limited) 3D stacked memory. As such, it is possible to store multiples embedding lookup tables entirely inside TensorNode because the multitude of TensorDIMM devices (each equipped with density-optimized LR-DIMM [28]) enable scalable memory capacity increase. Second, the aggregate memory bandwidth available to the TensorDIMM NMP cores has been designed to scale up proportional to the number of DIMMs provisioned within the TensorNode (Section 4.4). This allows our TensorNode and TensorDIMM design to fulfill not only the current but also future memory (capacity and bandwidth) needs of recommender systems which combines *multiple* embeddings, the size of which is expected to become even *larger* moving forward (Section 3.2). Recent projections from several hyperscalars [27, 64] state that the memory capacity requirements of embedding layers will increase by hundreds of times larger than the already hundreds of GBs of memory footprint. TensorNode is a scalable, future-proof system-level solution that addresses the memory bottlenecks of embedding layers. Third, the communication channels to/from the TensorNode is implemented using high-bandwidth NVLINK PHYs so transferring embeddings between a GPU and a TensorNode becomes much faster than when using PCIe (approximately 9 \times).

4.4 Software Architecture

We now discuss the software architecture of our proposal: the address mapping scheme for embeddings, TensorISA for conducting near-memory operations, and the runtime system.

Address mapping architecture. One of the key objectives of our address mapping function is to provide scalable performance improvement whenever additional TensorDIMMs (i.e., number of DIMMs=NMP cores) are added to our TensorNode. To achieve this goal, it is important that all NMP cores within the TensorNode concurrently work on a distinct subset of the embedding vectors for (gather/reduce) tensor operations. Figure 7(a) illustrates our address mapping scheme which utilizes rank-level parallelism to maximally utilize both the TensorDIMM’s NMP computation throughput and memory bandwidth. Because each TensorDIMM has its own NMP core, maximally utilizing aggregate NMP compute throughput requires all TensorDIMMs to work in parallel. Our address mapping function accomplishes this by having consecutive 64 bytes⁴ within each embedding vector be interleaved across different ranks, allowing each TensorDIMM to independently work on its own slice of tensor operation concurrently (Figure 7(b)). As our target algorithm contains abundant data-level parallelism, our address mapping technique effectively partitions and load-balances the tensor operation across all TensorDIMMs within the TensorNode. In Section 6.1, we quantitatively evaluate the efficacy of our address mapping function in terms of maximum DRAM bandwidth utilization.

It is worth pointing out that our address mapping function is designed to address not just the current but also future projections on how DL practitioners seek to use embeddings. As discussed in Section 3.2, given the capability, DL practitioners are willing to

³CC-NUMA access or P2P `cudaMemcpy` among NVLINK-compatible devices is already available in commercial systems (e.g., Power9 [29], GPUs within DGX-2 [58]). Our TensorNode leverages such technology as-is to minimize design complexity.

⁴We assume TensorDIMM is built using a x64 DIMM. With a burst length of 8, the minimum data access granularity becomes 64 bytes.

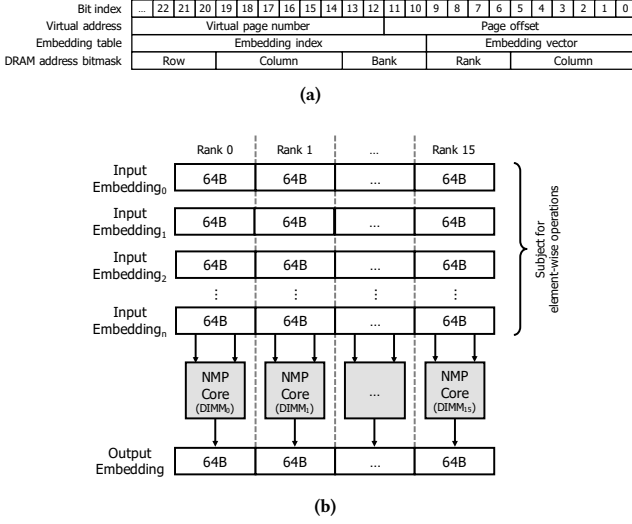


Figure 7: (a) Proposed DRAM address mapping scheme for embeddings, assuming each embedding is 1 KB (256-dimension) and is split evenly across 16 ranks (or DIMMs) (b) Example showing how rank-level parallelism is utilized to interleave and map each embedding across the 16 TensorDIMMs. Such rank-level parallelism centric address mapping scheme enable the memory bandwidth available to the NMP cores to increase proportional to the number of TensorDIMMs employed.

adopt *larger* embedding vector dimensions to improve the algorithmic performance. This translates into a larger number of bits to reference any given embedding vector (e.g., enlarging embedding size from 1 KB to 4 KB increases the embedding's bit-width from 10 bits to 12 bits in Figure 7(a)) leading to both larger memory footprint to store the embeddings and higher computation and memory bandwidth demands to conduct tensor reductions. Our system-level solution has been designed from the ground-up to effectively handle such user needs: that is, system architects can provision more TensorDIMM ranks within TensorNode and increase its memory capacity proportional to the increase in embedding size, which is naturally accompanied by an increase in NMP compute throughput *and* memory bandwidth simultaneously.

TensorISA. The near-memory tensor operations are initiated using our custom ISA extension called TensorISA. There are three key TensorISA primitives supported in TensorDIMM: the GATHER instruction for embedding lookups and the REDUCE and AVERAGE instructions for element-wise operations. Figure 8 and Figure 9 summarize the instruction formats of these three instructions and pseudo codes describing each of its functional behavior. Consider the example in Figure 2 which assumes an embedding layer that uses two embedding lookup tables with a batch size 4 to compose two tensors, followed by an element-wise operation among these two for reduction. As discussed in Section 2.3, an embedding layer starts with an embedding lookup phase that *gathers* multiple embeddings up to the batch size from the embedding lookup table, followed by various tensor manipulations. Under our proposed system, the GPU executes this embedding layer as follows. First,

Instruction format	OpCode	InputBase	AUX	OutputBase	Count
Embedding lookup	GATHER	tableBase	idxBase	outputBase	count
Element-wise reduction	REDUCE	inputBase1	inputBase2	outputBase	count
Element-wise average	AVERAGE	inputBase	averageNum	outputBase	count

Figure 8: Instruction formats for GATHER/REDUCE/AVERAGE.

```
# READ64B(void* src): read 64 bytes from address src
# WRITE64B(void* dst, REG64B v): store 64 bytes value to address dst
# nodeDim: number of TensorDIMMs within a TensorNode
# tid: unique ID assigned to each TensorDIMM
# A, B and C: 64 bytes vector register
# X is a 64 bytes array which can be accessed in 4 bytes granularity (i.e., X[0] to X[15])
```

```
00 #pseudo code of embedding lookup (GATHER)
01 for i in range(count/16):
02   X ← READ64B(idxBase + i)
03   for j in range(16):
04     C ← READ64B(tableBase + (X[j] * nodeDim) + tid)
05     WRITE64B(outputBase + (i * 16 + j) * nodeDim + tid, C)
```

(a)

```
00 #pseudo code of element-wise operations (REDUCE)
01 for i in range(Count):
02   A ← READ64B(inputBase1 + i * nodeDim + tid)
03   B ← READ64B(inputBase2 + i * nodeDim + tid)
04   C ← A <OP> B
05   WRITE64B(outputBase + i * nodeDim + tid, C)
```

(b)

```
00 #pseudo code of element-wise average (AVERAGE)
01 for i in range(count):
02   C ← 256'b0
03   for j in range(averageNum):
04     A ← READ64B(inputBase + (i * averageNum + j) * nodeDim + tid)
05     C ← C + A
06   C ← C <divide> averageNum
07   WRITE64B(outputBase + i * nodeDim + tid, C)
```

(c)

Figure 9: Pseudo code explaining the functional behavior of (a) GATHER, (b) REDUCE, and (c) AVERAGE.

the GPU sends three instructions, two GATHER instructions and one REDUCE, to the TensorNode. The TensorISA instruction is broadcasted to all the TensorDIMMs because each NMP core is responsible for locally conducting its share of the embedding lookups as well as its slice of the tensor operation (Figure 9(a,b)). For instance, assuming the address mapping function in Figure 7(a) and a TensorNode configuration with 16 TensorDIMMs, a single GATHER instruction will have each TensorDIMM gather $(4 \times 64 \text{ bytes}) = 256 \text{ bytes}$ of data under a contiguous physical address space, the process of which is orchestrated by the TensorDIMM NMP-local memory controller using DRAM read/write transactions (Section 4.2). The GATHER process is undertaken twice to prepare for the two tensor slices per each TensorDIMM rank. Once the two tensors are gathered, a REDUCE instruction is executed by the NMP cores (Section 4.2).

Runtime system. DL applications are typically encapsulated as a direct acyclic graph (DAG) data structure in major DL frameworks [9, 66, 75]. Each node within the DAG represents a DNN layer and the DL framework compiles down the DAG into sequence of host-side CUDA kernel launches that the GPU executes one layer at a time. The focus of this paper is on recommender systems which utilizes embedding lookups and various tensor manipulations. Under our proposed system, embedding layers are still executed using normal CUDA kernel launches but the kernel itself is wrapped around specific information for our TensorDIMM runtime system

Table 1: Baseline TensorNode configuration.

DRAM specification	DDR4 (PC4-25600)
Number of TensorDIMMs	32
Memory bandwidth per TensorDIMM	25.6 GB/sec
Memory bandwidth across TensorNode	819.2 GB/sec

to utilize for near-memory tensor operations. Specifically, as part of the embedding layer’s CUDA kernel context, information such as the number of table lookups, the embedding dimension size, tensor reduction type, the input batch size, and etc are encoded per TensorISA instruction’s format (Figure 8) and is sent to the GPU as part of the CUDA kernel launch. When the GPU runtime receives these instructions, they are forwarded to the TensorNode for near-memory processing as discussed in Section 4.2.

As TensorNode is a remote, disaggregated memory pool from the GPU’s perspective, the runtime system should be able to (de)allocate memory inside this remote memory pool. Our proposal builds upon our prior work [39] which proposes several CUDA runtime API extensions for remote memory (de)allocation under a GPU-side disaggregated memory system. We refer the interested readers to [13, 39] for further details on the runtime CUDA APIs required for memory (de)allocations within a disaggregated memory system.

5 EVALUATION METHODOLOGY

Architectural exploration of TensorDIMM and its system-level implication within TensorNode using a cycle-level simulator is challenging for several reasons. First, running a single batch inference for DL applications can take up to several milliseconds even on high-end GPUs, so running cycle-level simulation on several tens to hundreds of batches of inference leads to intractable amount of simulation time. Second, our proposal covers multiple levels in the hardware/software stack, so a cycle-level hardware performance model of TensorDIMM and TensorNode alone will not properly reflect the complex interaction of (micro)architecture, runtime system, and the system software, potentially resulting in misleading conclusions. Interestingly, we note that the key DL operations utilized in embedding layers that are of interest to this study are completely *memory bandwidth limited*. This allows us to utilize the existing DL hardware/software systems to “emulate” the behavior of TensorDIMM and TensorNode on top of state-of-the-art real DL systems. Recall that the embedding lookups (GATHER) and the tensor reduction operations (REDUCE/AVERAGE) have extremely low compute-to-memory ratio (i.e., all three operations are effectively *streaming* applications), rendering the execution of these three operations to be bottlenecked by the available memory bandwidth. Consequently, the effectiveness of our proposal primarily lies in 1) the effectiveness of our address mapping scheme on maximally utilizing the aggregate memory bandwidth for the tensor operations across all the TensorDIMMs, and 2) the impact of “PCIe vs. NVLINK” on the communication latency of copying embeddings across the capacity-optimized “CPU vs. TensorNode” memory and bandwidth-optimized GPU memory (Figure 5). We introduce our novel, hybrid evaluation methodology that utilizes both cycle-level simulation and a proof-of-concept prototype developed on real DL systems to quantitatively demonstrate the benefits of our proposal.

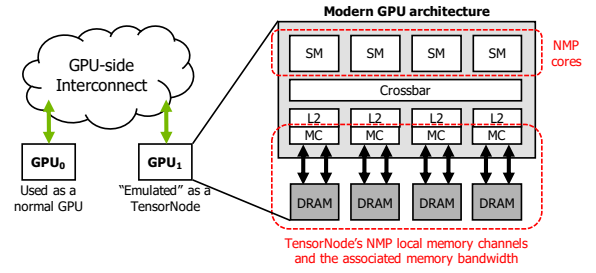


Figure 10: Emulation of TensorDIMM and TensorNode using a real GPU: the GPU cores (aka SMs [56]) and the GPU-local memory channels (and bandwidth) corresponds to the NMP cores within TensorDIMMs and the aggregate memory bandwidth available across the TensorNode, respectively.

Cycle-level simulation. As the performance of TensorDIMM and TensorNode is bounded by how well they utilize the DRAM bandwidth, an evaluation of our proposal on memory bandwidth utilization is in need. We develop a memory tracing function that hooks into the DL frameworks [66, 75] to generate the necessary read/write memory transactions in executing GATHER, REDUCE, and AVERAGE operations for embedding lookups and tensor operations. The traces are fed into Ramulator [36], a cycle-accurate DRAM simulator, which is configured to model 1) the baseline CPU-GPU system configuration with eight CPU-side memory channels, and 2) our proposed address mapping function (Section 4.4) and TensorNode configuration (Table 1), which we utilize to measure the effective memory bandwidth utilization when executing the three tensor operations under baseline and TensorNode.

“Proof-of-concept” prototype. We emulate the behavior of our proposed system using the state-of-the-art NVIDIA DGX [57] machine, which includes eight V100 GPUs [60]. Each V100 GPU contains 900 GB/sec of local memory bandwidth with six NVLINKs for communicating with other GPUs up to 150 GB/sec [61]. To emulate the system-level effects of the high-bandwidth NVLINK communication channels between TensorNode↔GPU, we use a pair of GPUs in DGX and treat one of them as our proposed TensorNode while the other acts as a normal GPU (Figure 10). Because the tensor operations accelerated using our TensorDIMMs are memory bandwidth-limited, streaming workloads, the performance difference between a (hypothetical) real TensorNode populated with N TensorDIMMs and a single V100 that we emulate as our TensorNode will be small, provided the V100 local memory bandwidth matches that of our assumed TensorNode configuration. As such, when evaluating the effective memory bandwidth utilization of TensorNode, we configured the number of ranks (i.e., TensorDIMMs) to be $N=32$ such that the aggregate memory bandwidth available within a single TensorNode approximately matches that of a single V100 (900 GB/sec vs. 819.2 GB/sec, Table 1). After validating the effectiveness of TensorNode in utilizing memory bandwidth, commensurate to that of V100, we implement a software prototype of an end-to-end recommender system (configurable to the four DL applications discussed below) using Intel’s Math Kernel Library (MKL version 2019.0.3 [30]), cuDNN (version 7 [55]), cuBLAS [54], and our in-house CUDA implementation of embedding layers (including

Table 2: Evaluated benchmarks and default configuration.

Network	Lookup tables	Max reduction	FC/MLP layers
NCF	4	2	4
YouTube	2	50	4
Fox	2	50	1
Facebook	8	25	6

GATHER/REDUCE/AVERAGE) as well as other layers that do not come with MKL, cuDNN or cuBLAS. We cross-validated our in-house implementation of these memory bandwidth-limited layers by comparing the measured performance against the upper bound, ideal performance, which exhibited little variation. Under our CUDA implementation of tensor reduction operations, the GPU cores (called SMs in CUDA [56]) effectively function as the NMP cores within TensorNode because the CUDA kernels of REDUCE/AVERAGE stage in/out the tensors between (SM \leftrightarrow GPU local memory) in a streaming fashion as done in TensorDIMM. The TensorNode \leftrightarrow GPU communication is orchestrated using P2P cudaMemcpy over NVLINK when evaluating our proposal. When running the sensitivity of TensorDIMM on (TensorNode \leftrightarrow GPU) communication bandwidth, we artificially increase (decrease) the data transfer size to emulate the behavior of smaller (larger) communication bandwidth and study its implication on system performance (Section 6.4).

Benchmarks. We choose four neural network based recommender system applications using embeddings to evaluate our proposal: (1) the neural collaborative filtering (NCF) [26] based recommender system available in MLPerf [49], (2) the YouTube recommendation system [14] (YouTube), and (3) the Fox movie recommendation system [7] (Fox), and the Facebook recommendation system [52] (Facebook⁵). To choose a realistic inference batch size representative of real-world inference scenarios, we refer to the recent study from Facebook [64] which states that datacenter recommender systems are commonly deployed with a batch size of 1–100. Based on this prior work, we use a batch size of 64 as our default configuration but sweep from batch 1 to 128 when running sensitivity studies. All four workloads are configured with a default embedding vector dimension size of 512 with other key application configurations summarized in Table 2. We specifically note when deviating from these default configurations when running sensitivity studies in Section 6.3.

Area/power. The implementation overheads of TensorDIMM are measured with synthesized implementations using Verilog HDL, targeting a Xilinx Virtex UltraScale+ VCU1525 acceleration dev board. The system-level power overheads of TensorNode are evaluated using Micron’s DDR4 power calculator [47]. We detail these results in Section 6.5.

6 EVALUATION

We explore five design points of recommender systems: the 1) CPU-only version (CPU-only), 2) hybrid CPU-GPU version (CPU-GPU), 3) TensorNode style pooled memory interfaced inside the high-bandwidth GPU interconnect but utilizes regular capacity-optimized DIMMs, rather than the NMP-enabled TensorDIMMs (PMEM), 4) our

⁵Facebook’s Deep Learning Recommendation Model [52] was released *after* this work was submitted for peer-review at MICRO '52. As such, we implement and evaluate this model using our proof-of-concept emulation framework, rather than using Facebook’s open-sourced PyTorch [66] version, for consistency with the other three models.

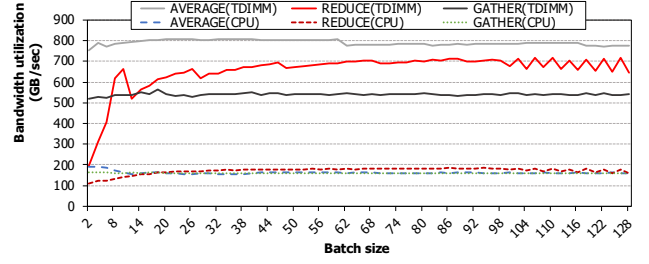


Figure 11: Memory bandwidth utilization for the three tensor operations. TensorNode assumes the default configuration in Table 1 with 32 TensorDIMMs. For CPU-only and CPU-GPU, the tensor operations are conducted over a conventional CPU memory system, so a total of 8 memory channels with 32 DIMMs (4 ranks per each memory channel) are assumed.

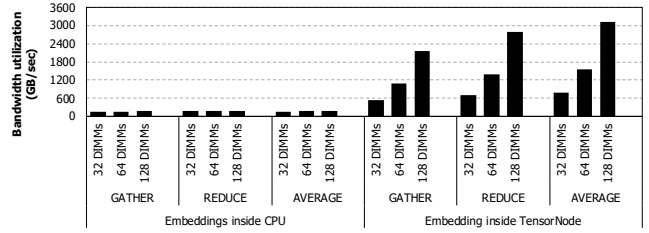


Figure 12: Memory throughput as a function of the number of DIMMs employed within CPU and TensorNode. Evaluation assumes the embedding size is increased from the default value by 2–4 \times , which proportionally increases the embedding lookup table size, requiring a proportional increase in memory capacity (i.e., more DIMMs) to store these lookup tables.

proposed TensorNode with TensorDIMMs (TDIMM), and 5) an unbuildable, oracular GPU-only version (GPU-only) which assumes that the entire embeddings can be stored inside GPU local memory, obviating the need for cudaMemcpy.

6.1 Memory Bandwidth Utilization

To validate the effectiveness of TensorNode in amplifying effective memory bandwidth, we measure the aggregate memory throughput achieved with TensorNode and the baseline CPU-based systems, both of which utilizes a total of 32 DIMMs (Figure 11). TensorNode significantly outperforms the baseline CPU system with an average 4 \times increase in memory bandwidth utilization (max 808 vs. 192 GB/sec). As discussed in Section 4.2, conventional memory systems time-multiplex the memory channel across multiple DIMMs, so larger number of DIMMs only provide enlarged memory capacity not bandwidth. Our TensorDIMM is designed to ensure that the aggregate memory bandwidth scales proportional to the number of TensorDIMMs, achieving significant memory bandwidth scaling. The benefits of our proposal is more pronounced for more future-looking scenarios with enlarged embedding dimension sizes. Figure 12 shows the effective memory bandwidth for tensor operations when the embedding dimension size is increased by up to 4 \times , necessitating larger number of DIMMs to house the proportionally increased embedding lookup table. As depicted, the baseline

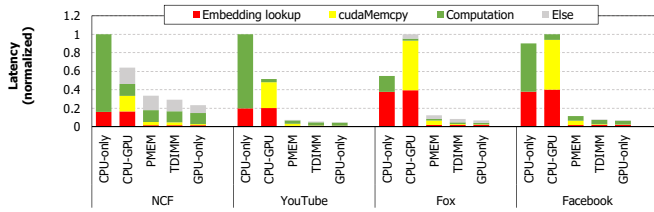


Figure 13: Breakdown of latencies during an inference with batch 64, normalized to the slowest design point (i.e., CPU-only or CPU-GPU).

CPU memory system’s memory bandwidth saturates at around 200 GB/sec because of the fundamental limits of conventional memory systems. TensorNode on the other hand reaches up to 3.1 TB/sec of bandwidth, more than 15× increase than baseline.

6.2 System-level Performance

TensorDIMM significantly reduces the latency to execute memory-limited embedding layers, thanks to its high memory throughput and the communication bandwidth amplification effects of both near-memory tensor operations *and* the high-bandwidth NVLINK. Figure 13 shows a latency breakdown of our studied workloads assuming a batch size of 64. With our proposed solution, all four applications enjoy significant reduction in both embedding lookup latency and the embedding copy latency. This is because of the high-bandwidth TensorDIMM tensor operations and the fast communication links utilized for moving the embeddings to GPU memory. Figure 14 summarizes the normalized performance of the five design points across different batch sizes. While CPU-only occasionally achieves better performance than CPU-GPU for low batch inference scenarios, the oracular GPU-only consistently performs best on average, highlighting the advantages of DNN acceleration using GPUs (and NPUs). TensorDIMM achieves an average 84% (no less than 75%) of the performance of such unbuildable oracular GPU, demonstrating its performance merits and robustness compared to CPU-only and CPU-GPU (an average 6.2× and 8.9× speedup).

6.3 TensorDIMM with Large Embeddings

A key motivation of our work is to provide a scalable memory system for embeddings and its tensor operations. So far, we have assumed the default embedding size of each workloads as summarized in Table 2. With the availability of our pooled memory architecture, DL practitioners can provision much larger embeddings to develop recommender systems with superior model quality. Because our evaluation is conducted over an emulated version of TensorDIMMs, we are not able to demonstrate the model quality improvements larger embedding will bring about as we cannot train these scaled-up algorithms (i.e., the memory capacity is still constrained by the GPU memory size). Nonetheless, we conduct a sensitivity study of the TensorDIMMs for scaled-up embedding sizes which is shown in Figure 15. With larger embeddings, the embedding layers cause a much more serious performance bottleneck. TensorNode shows even higher performance benefits under these settings, achieving an average 6.2–15.0× and 8.9–17.6× (maximum 35×) performance improvement than CPU-only and CPU-GPU, respectively.

6.4 TensorDIMM with Low-bandwidth System Interconnects

To highlight the maximum potential of TensorDIMM, we discussed its merits under the context of a high-bandwidth GPU-side interconnect. Nonetheless, it is still possible for TensorDIMM to be utilized under conventional, CPU-centric disaggregated memory systems. Concretely, one can envision a system that contains a pooled memory like TensorNode interfaced over a low-bandwidth system interconnect (e.g., PCIe). Such design point looks similar to the hybrid CPU-GPU except for one key distinction: the tensor operations are done using the NMP cores but the reduced tensor must be copied to the GPU over a slower communication channel. Figure 16 summarizes the sensitivity of our proposal on the TensorNode↔GPU communication bandwidth. We study both PMEM (i.e., disaggregated memory “without” TensorDIMMs) and TDIMM under low-bandwidth system interconnects to highlight the robustness our TensorDIMM brings about. Overall, PMEM is much more sensitive to the communication bandwidth than TDIMM (maximum 68% performance loss) because the benefits of NMP reduction is lost with PMEM. Our TensorDIMM design (TDIMM) on the other hand only experiences up to 15% performance loss (average 10%) even with a 6× lower communication bandwidth. These results highlight the robustness and the wide applicability of TensorDIMM.

6.5 Design Overheads

TensorDIMM requires no changes to the DRAM chip itself but adds a lightweight NMP core inside the buffer device (Section 4.2). We implement and synthesized the major components of our NMP core on a Xilinx Virtex UltraScale+ FPGA board using Verilog HDL. We confirm that the added area/power overheads of our NMP core is mostly negligible as it is dominated by the small (1.5 KB) SRAM queues and the 16-wide vector ALU (Table 3). From a system-level perspective, our work utilizes GPUs as-is, so the major overhead comes from the disaggregated TensorNode design. Assuming a single TensorDIMM uses 128 GB load-reduced DIMM [28], its power consumption becomes 13W when estimated using Micron’s DDR4 system power calculator [47]. For TensorNode with 32 TensorDIMMs, this amounts to a power overhead of $(13 \times 32) = 416W$. Recent specifications for accelerator interconnection fabric endpoints (e.g., Open Compute Project [20]’s open accelerator module [19]) employ a TDP of 350–700W so the power overheads of TensorNode is expected to be acceptable.

7 RELATED WORK

Disaggregated memory [41, 42] is typically deployed as a remote memory pool connected over PCIe, which helps increase CPU accessible memory capacity. Prior work [13, 38, 39] proposed system-level solutions that embrace the idea of memory disaggregation within the high-bandwidth GPU interconnect, which bears similarity to our proposition on TensorNode. However, the focus of these two prior work is on DL training whereas our study primarily focuses on DL inference. Similar to our TensorDIMM, several prior work [4, 6, 21, 46] explored the possibility of utilizing the DIMM buffer device space to add custom acceleration logics. Alian et al. [4] for instance adds a lightweight CPU core inside the buffer device to construct a memory channel network (MCN), seeking to

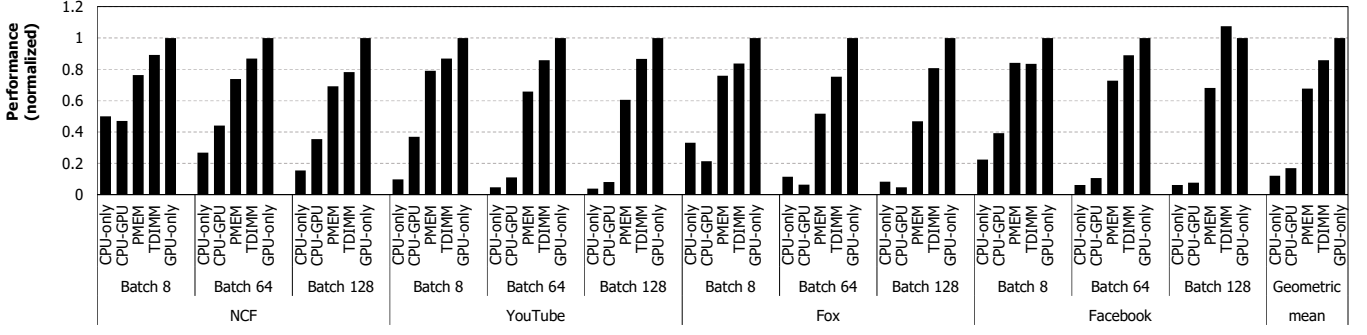


Figure 14: Performance of the five design points of recommender systems, normalized to the oracular GPU (GPU-only).

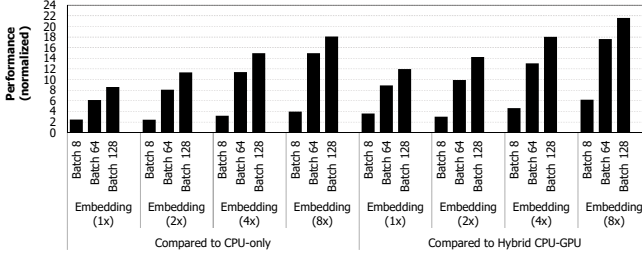


Figure 15: TensorDIMM performance with larger embedding size (up to 8x). Results are averaged across the four studied benchmarks.

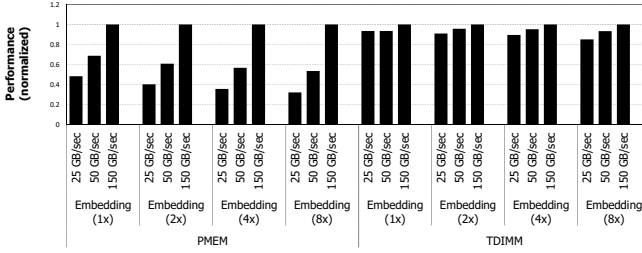


Figure 16: Performance sensitivity of PMEM (i.e., pooled memory without NMP acceleration) and TensorDIMM to the communication bandwidth. Results are averaged across the four studied benchmarks, and are normalized to the default configuration of 150 GB/sec data point.

amplify effective memory bandwidth exposed to the overall system. While TensorDIMM’s near-memory processing approach bears some similarity with MCN, direct application of this design for embedding layer’s tensor operations will likely lead to sub-optimal performance because of MCN’s inability to maximally utilize DRAM bandwidth for embeddings. Concretely, Gupta et al. [24] showed the performance of embedding lookup operations with CPUs, reporting that the irregular, sparse memory access nature of embedding operations render the CPU cache hit-rate to become extremely low. As such, the latency to traverse the cache hierarchy leads to less than 5% of the maximum DRAM bandwidth being utilized, as opposed to the TensorDIMM NMP cores being able to reach close to maximum DRAM bandwidth on average. In general, the scope of all these prior studies is significantly different from what our work focuses on. To the best of our knowledge, our work is the first in

Table 3: FPGA utilization of a single NMP core (i.e., SRAM queues and vector ALU with single-precision floating point (FPU) and fixed point ALU) on Xilinx Virtex UltraScale+ VCU1525 acceleration development board.

	LUT [%]	FF [%]	DSP [%]	BRAM [%]
SRAM queues	0.00	0.00	0.00	0.01
FPU	0.19	0.01	0.20	0.00
ALU	0.09	0.01	0.01	0.00

both industry and academia to identify and address the memory capacity and bandwidth challenges of embedding layers, which several hyperscalars [15, 27, 52, 64, 67] deem as one of the most crucial challenge in emerging DL workloads.

Other than these closely related studies, a large body of prior work has explored the design of a single GPU/NPU architecture for DL [8, 10–12, 17, 18, 22, 31, 35, 43–45, 50, 51, 68, 69, 73, 74] with recent interest on leveraging sparsity for further energy-efficiency improvements [2, 3, 16, 25, 34, 53, 63, 70, 78–81]. A scale-out acceleration platform for training DL algorithms was proposed by Park et al. [65] and a network-centric DL training platform has been proposed by Li et al. [40]. These prior studies are orthogonal to our proposal and can be adopted further for additional enhancements.

8 CONCLUSION

In this paper, we propose a vertically integrated, hardware/software co-design that addresses the memory (capacity and bandwidth) wall problem of embedding layers, an important building block for emerging DL applications. Our TensorDIMM architecture synergistically combines NMP cores with commodity DRAM devices to accelerate DL tensor operations. Built on top of a disaggregated memory pool, TensorDIMM provides memory capacity and bandwidth scaling for embeddings, achieving an average 6.2–15.0x and 8.9–17.6x performance improvement than conventional CPU-only and hybrid CPU-GPU implementations of recommender systems. To the best of our knowledge, TensorDIMM is the first that quantitatively explores architectural solutions tailored for embeddings and tensor operations.

ACKNOWLEDGMENTS

This research is supported by Samsung Research Funding Center of Samsung Electronics (SRFC-TB1703-03).

REFERENCES

- [1] ACM. 2019. The ACM Conference Series on Recommendation Systems. <https://recsys.acm.org/>.
- [2] Jorge Albericio, Alberto Delmas, Patrick Judd, Sayeh Sharify, Gerard O'Leary, Roman Genov, and Andreas Moshovos. 2017. Bit-pragmatic Deep Neural Network Computing. In *Proceedings of the International Symposium on Microarchitecture (MICRO)*.
- [3] Jorge Albericio, Patrick Judd, Tayler Hetherington, Tor Aamodt, Natalie Enright Jerger, and Andreas Moshovos. 2016. Cnvlutin: Ineffectual-Neuron-Free Deep Convolutional Neural Network Computing. In *Proceedings of the International Symposium on Computer Architecture (ISCA)*.
- [4] Mohammad Alian, Seung Won Min, Hadi Asgharimoghaddam, Ashutosh Dhar, Dong Kai Wang, Thomas Roewer, Adam McPadden, Oliver O'Halloran, Deming Chen, Jinjun Xiong, Daehoon Kim, Wenmei Hwu, and Nam Sung Kim. 2018. Application-Transparent Near-Memory Processing Architecture with Memory Channel Network. In *Proceedings of the International Symposium on Microarchitecture (MICRO)*.
- [5] Akhil Arunkumar, Evgeny Bolotin, Benjamin Cho, Ugljesa Milic, Eiman Ebrahimi, Oreste Villa, Aamer Jaleel, Carole-Jean Wu, and David Nellans. 2017. MCM-GPU: Multi-chip-module GPUs for Continued Performance Scalability. In *Proceedings of the International Symposium on Computer Architecture (ISCA)*.
- [6] Hadi Asghari-Moghaddam, Young Hoon Son, Jung Ho Ahn, and Nam Sung Kim. 2016. Chameleon: Versatile and Practical Near-DRAM Acceleration Architecture for Large Memory Systems. In *Proceedings of the International Symposium on Microarchitecture (MICRO)*.
- [7] Miguel Campo, Cheng-Kang Hsieh, Matt Nickens, JJ Espinoza, Abhinav Taliyan, Julie Rieger, Jean Ho, and Bettina Sherick. 2018. Competitive Analysis System for Theatrical Movie Releases Based on Movie Trailer Deep Video Representation. <https://arxiv.org/abs/1807.04465>.
- [8] Tianshi Chen, Zidong Du, Ninghui Sun, Jia Wang, Chengyong Wu, Yunji Chen, and Olivier Temam. 2014. DianNao: A Small-footprint High-throughput Accelerator for Ubiquitous Machine-learning. In *Proceedings of the International Conference on Architectural Support for Programming Languages and Operation Systems (ASPLOS)*.
- [9] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. 2015. MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems. In *Proceedings of the Workshop on Machine Learning Systems*.
- [10] Yunji Chen, Tao Luo, Shaoli Liu, Shijin Zhang, Liqiang He, Jia Wang, Ling Li, Tianshi Chen, Zhiwei Xu, Ninghui Sun, and Olivier Temam. 2014. DaDianNao: A Machine-Learning Supercomputer. In *Proceedings of the International Symposium on Microarchitecture (MICRO)*.
- [11] Yu-Hsin Chen, Tushar Krishna, Joel S. Emer, and Vivienne Sze. 2016. Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks. In *Proceedings of the International Solid State Circuits Conference (ISSCC)*.
- [12] Ping Chi, Shuangchen Li, Cong Xu, Tao Zhang, Jishen Zhao, Yongpan Liu, Yu Wang, and Yuan Xie. 2016. Prime: A Novel Processing-in-memory Architecture for Neural Network Computation in ReRAM-based Main Memory. In *Proceedings of the International Symposium on Computer Architecture (ISCA)*.
- [13] Esha Choukse, Michael Sullivan, Mike O'Connor, Mattan Erez, Jeff Pool, David Nellans, and Stephen W. Keckler. 2019. Buddy Compression: Enabling Larger Memory for Deep Learning and HPC Workloads on GPUs. <https://arxiv.org/abs/1903.02596>.
- [14] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for Youtube Recommendations. In *Proceedings of the ACM Conference on Recommender Systems (RECSYS)*.
- [15] Jeff Dean, David Patterson, and Cliff Young. 2018. A New Golden Age in Computer Architecture: Empowering the Machine-Learning Revolution. In *IEEE Micro*.
- [16] Alberto Delmas, Patrick Judd, Dylan Stuart, Zissis Poulos, Mostafa Mahmoud, Sayeh Sharify, Milos Nikolic, and Andreas Moshovos. 2018. Bit-Tactical: Exploiting Ineffectual Computations in Convolutional Neural Networks: Which, Why, and How. <https://arxiv.org/abs/1803.03688>.
- [17] Zidong Du, Robert Fasthuber, Tianshi Chen, Paolo Ienne, Ling Li, Tao Luo, Xiaobing Feng, Yunji Chen, and Olivier Temam. 2015. ShiDianNao: Shifting Vision Processing Closer to the Sensor. In *Proceedings of the International Symposium on Computer Architecture (ISCA)*.
- [18] Zidong Du, Daniel D Ben-Dayana Rubin, Yunji Chen, Liqiang He, Tianshi Chen, Lei Zhang, Chengyong Wu, and Olivier Temam. 2015. Neuromorphic Accelerators: A Comparison Between Neuroscience and Machine-Learning Approaches. In *Proceedings of the International Symposium on Microarchitecture (MICRO)*.
- [19] Facebook. 2019. Accelerating Facebook's infrastructure with Application-Specific Hardware. <https://code.fb.com/data-center-engineering/accelerating-infrastructure/>.
- [20] Facebook. 2019. Open Compute Project. <https://www.opencompute.org/>.
- [21] Amin Farmahini-Farahani, Jung Ho Ahn, Katherine Morrow, and Nam Sung Kim. 2015. NDA: Near-DRAM Acceleration Architecture Leveraging Commodity DRAM Devices and Standard Memory Modules. In *Proceedings of the International Symposium on High-Performance Computer Architecture (HPCA)*.
- [22] Mingyu Gao, Jing Pu, Xuan Yang, Mark Horowitz, and Christos Kozyrakis. 2017. TETRIS: Scalable and Efficient Neural Network Acceleration with 3D Memory. In *Proceedings of the International Conference on Architectural Support for Programming Languages and Operation Systems (ASPLOS)*.
- [23] Google. 2017. Cloud TPUs: ML accelerators for TensorFlow.
- [24] Udit Gupta, Xiaodong Wang, Maxim Naumov, Carole-Jean Wu, Brandon Reagen, David Brooks, Bradford Cotel, Kim Hazelwood, Bill Jia, Hsien-Hsin S. Lee, Andrey Malevich, Dheevatsa Mudigere, Mikhail Smelyanskiy, Liang Xiong, and Xuan Zhang. 2019. The Architectural Implications of Facebook's DNN-based Personalized Recommendation. In *arxiv.org*.
- [25] Song Han, Xingyu Liu, Huizi Mao, Jing Pu, Ardavan Pedram, Mark Horowitz, and William J. Dally. 2016. EIE: Efficient Inference Engine on Compressed Deep Neural Network. In *Proceedings of the International Symposium on Computer Architecture (ISCA)*.
- [26] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the International Conference on World Wide Web (WWW)*.
- [27] Joel Hestness, Newsha Ardalani, and Gregory Diamos. 2019. Beyond Human-Level Accuracy: Computational Challenges in Deep Learning. In *Proceedings of the Symposium on Principles and Practice of Parallel Programming (PPOPP)*.
- [28] Hynix. 2017. 128 GB 3DS LRDIMM: The World's First Developed 3DS LRDIMM. (2017).
- [29] IBM. 2017. IBM Power9 Microprocessor.
- [30] Intel. 2019. Intel Math Kernel Library. <https://software.intel.com/en-us/mkl>.
- [31] Hanhui Jang, Joonsung Kim, Jae-Eon Jo, Jaewon Lee, and Jangwoo Kim. 2019. MnFast: A Fast and Scalable System Architecture for Memory-augmented Neural Networks. In *Proceedings of the International Symposium on Computer Architecture (ISCA)*.
- [32] JEDEC. 2018. High Bandwidth Memory (HBM2) DRAM. (2018).
- [33] Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Gaurav Agrawal, Raminder Bajwa, Sarah Bates, Suresh Bhatia, Nan Boden, Al Borchers, Rick Boyle, Pierre-luc Cantin, Clifford Chao, Chris Clark, Jeremy Coriell, Mike Daley, Matt Dau, Jeffrey Dean, Ben Gelb, Tara Vazir Ghaemmaghami, Rajendra Gottipati, William Gulland, Robert Hagmann, C. Richard Ho, Doug Hogberg, John Hu, Robert Hundt, Dan Hurt, Julian Ibarz, Aaron Jaffey, Alek Jaworski, Alexander Kaplan, Harshit Khaitan, Daniel Killebrew, Andy Koch, Naveen Kumar, Steve Lacy, James Laudon, James Law, Diemthu Le, Chris Leary, Zhuyuan Liu, Kyle Lucke, Alan Lundin, Gordon MacKean, Adriana Maggiore, Maire Mahony, Kieran Miller, Rahul Nagarajan, Ravi Narayanaswami, Ray Ni, Kathy Nix, Thomas Norrie, Mark Omernick, Narayana Penukonda, Andy Phelps, Jonathan Ross, Matt Ross, Amir Salek, Emad Samadiani, Chris Severn, Gregory Sizikov, Matthew Snelham, Jed Souter, Dan Steinberg, Andy Swing, Mercedes Tan, Gregory Thorson, Bo Tian, Horia Toma, Erick Tuttle, Vijay Vasudevan, Richard Walter, Walter Wang, Eric Wilcox, and Doe Hyun Yoon. 2017. In-datacenter Performance Analysis of a Tensor Processing Unit. In *Proceedings of the International Symposium on Computer Architecture (ISCA)*.
- [34] Patrick Judd, Jorge Albericio, Tayler Hetherington, Tor Aamodt, and Andreas Moshovos. 2016. Stripes: Bit-serial Deep Neural Network Computing. In *Proceedings of the International Symposium on Microarchitecture (MICRO)*.
- [35] Duckhwan Kim, Jaeha Kung, Sek Chai, Sudhakar Yalamanchili, and Saibal Mukhopadhyay. 2016. Neurocube: A Programmable Digital Neuromorphic Architecture with High-Density 3D Memory. In *Proceedings of the International Symposium on Computer Architecture (ISCA)*.
- [36] Yoongu Kim, Weikun Yang, and Onur Mutlu. 2015. Ramulator: A Fast and Extensible DRAM Simulator.
- [37] Alex Krizhevsky. 2014. One Weird Trick For Parallelizing Convolutional Neural Networks. <https://arxiv.org/abs/1404.5997>.
- [38] Youngeun Kwon and Minsoo Rhu. 2018. A Case for Memory-Centric HPC System Architecture for Training Deep Neural Networks. In *IEEE Computer Architecture Letters*.
- [39] Youngeun Kwon and Minsoo Rhu. 2018. Beyond the Memory Wall: A Case for Memory-Centric HPC System for Deep Learning. In *Proceedings of the International Symposium on Microarchitecture (MICRO)*.
- [40] Youjie Li, Jongse Park, Mohammad Alian, Yifan Yuan, Zheng Qu, Peitian Pan, Ren Wang, Alexander Schwing, Hadi Esmaeilzadeh, and Nam Sung Kim. 2018. A Network-Centric Hardware/Algorithm Co-Design to Accelerate Distributed Training of Deep Neural Networks. In *Proceedings of the International Symposium on Microarchitecture (MICRO)*.
- [41] Kevin Lim, Jichuan Chang, Trevor Mudge, Parthasarathy Ranganathan, Steven K. Reinhardt, and Thomas F. Wenisch. 2009. Disaggregated Memory for Expansion and Sharing in Blade Servers. In *Proceedings of the International Symposium on Computer Architecture (ISCA)*.
- [42] Kevin Lim, Yoshio Turner, Jose Renato Santos, Alvin AuYoung, Jichuan Chang, Parthasarathy Ranganathan, and Thomas F. Wenisch. 2012. System-level Implications of Disaggregated Memory. In *Proceedings of the International Symposium on High-Performance Computer Architecture (HPCA)*.

- [43] Daofu Liu, Tianshi Chen, Shaoli Liu, Jinhong Zhou, Shengyuan Zhou, Olivier Temam, Xiaobing Feng, Xuehai Zhou, and Yunji Chen. 2015. PuDianNao: A Polyvalent Machine Learning Accelerator. In *Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*.
- [44] Shaoli Liu, Zidong Du, Jinhua Tao, Dong Han, Tao Luo, Yuan Xie, Yunji Chen, and Tianshi Chen. 2016. Cambricon: An Instruction Set Architecture for Neural Networks. In *Proceedings of the International Symposium on Computer Architecture (ISCA)*.
- [45] Divya Mahajan, Jongse Park, Emmanuel Amaro, Hardik Sharma, Amir Yazdambakhsh, Joon Kyung Kim, and Hadi Esmaeilzadeh. 2016. TABLA: A unified Template-based Framework for Accelerating Statistical Machine Learning. In *Proceedings of the International Symposium on High-Performance Computer Architecture (HPCA)*.
- [46] Patrick J. Meaney, Lawrence D. Curley, Glenn D. Gilda, Mark R. Hodges, Daniel J. Buerkle, Robert D. Siegl, and Roger K. Dong. 2015. The IBM z13 Memory Subsystem for Big Data. *IBM Journal of Research and Development* (2015).
- [47] Micron. 2017. Micron: System Power Calculator (DDR4). (2017).
- [48] Micron. 2018. Hybrid Memory Cube (HMC). (2018).
- [49] mlperf.org. 2018. MLPerf. <https://mlperf.org/>.
- [50] Duncan J.M. Moss, Srivatsan Krishnan, Eriko Nurvitadhi, Piotr Ratuszniak, Chris Johnson, Jaewoong Sim, Asit Mishra, Debbie Marr, Suchit Subhaschandra, and Philip H.W. Leong. 2018. A Customizable Matrix Multiplication Framework for the Intel HARPv2 Xeon+FPGA Platform: A Deep Learning Case Study. In *Proceedings of the ACM International Symposium on Field-Programmable Gate Arrays (FPGA)*.
- [51] Duncan J.M. Moss, Eriko Nurvitadhi, Jaewoong Sim, Asit Mishra, Debbie Marr, Suchit Subhaschandra, and Philip H.W. Leong. 2017. High Performance Binary Neural Networks on the Xeon+FPGA Platform. In *Proceedings of the International Conference on Field Programmable Logic and Applications (FPL)*.
- [52] Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G. Azzolini, Dmytro Dzhulgakov, Andrey Mallevich, Ilia Cherniavskii, Yinghai Lu, Raghuraman Krishnamoorthi, Anshu Yu, Volodymyr Kondratenko, Stephanie Pereira, Xianjie Chen, Wenlin Chen, Vijay Rao, Bill Jia, Liang Xiong, and Misha Smelyanskiy. 2019. Deep Learning Recommendation Model for Personalization and Recommendation Systems. In *arxiv.org*.
- [53] Eriko Nurvitadhi, Ganesh Venkatesh, Jaewoong Sim, Debbie Marr, Randy Huang, Jason Ong Gee Hock, Yeong Tat Liew, Krishnan Srivatsan, Duncan Moss, Suchit Subhaschandra, and Guy Boudoukh. 2017. Can FPGAs Beat GPUs in Accelerating Next-Generation Deep Neural Networks?. In *Proceedings of the ACM International Symposium on Field-Programmable Gate Arrays (FPGA)*.
- [54] NVIDIA. 2008. cuBLAS Library. (2008).
- [55] NVIDIA. 2016. cuDNN: GPU Accelerated Deep Learning.
- [56] NVIDIA. 2016. NVIDIA CUDA Programming Guide.
- [57] NVIDIA. 2017. The NVIDIA DGX-1V Deep Learning System.
- [58] NVIDIA. 2017. The NVIDIA DGX-2 Deep Learning System.
- [59] NVIDIA. 2018. Accelerating Recommendation System Inference Performance with TensorRT. <https://devblogs.nvidia.com/accelerating-recommendation-system-inference-performance-with-tensorrt/>.
- [60] NVIDIA. 2018. NVIDIA Tesla V100.
- [61] NVIDIA. 2018. NVLink High-Speed Interconnect.
- [62] NVIDIA. 2018. NVSwitch: Leveraging NVLink to Maximum Effect.
- [63] Angshuman Parashar, Minsoo Rhu, Anurag Mukkara, Antonio Puglielli, Rangharajan Venkatesan, Bruce Khailany, Joel Emer, Stephen W. Keckler, and William J. Dally. 2017. SCNN: An Accelerator for Compressed-sparse Convolutional Neural Networks. In *Proceedings of the International Symposium on Computer Architecture (ISCA)*.
- [64] Jongsoo Park, Maxim Naumov, Protonu Basu, Summer Deng, Aravind Kalaiah, Daya Khudia, James Law, Parth Malani, Andrey Malevich, Satish Nadathur, Juan Pino, Martin Schatz, Alexander Sidorov, Viswanath Sivakumar, Andrew Tulloch, Xiaodong Wang, Yiming Wu, Hector Yuen, Utku Diril, Dmytro Dzhulgakov, Kim Hazelwood, and Bill Jia, Yangqing Jia, Lin Qiao, Vijay Rao, Nadav Rotem, Sungjoo Yoo, and Mikhail Smelyanskiy. 2018. Deep Learning Inference in Facebook Data Centers: Characterization, Performance Optimizations and Hardware Implications. In *arxiv.org*.
- [65] Jongse Park, Hardik Sharma, Divya Mahajan, Joon Kyung Kim, Preston Olds, and Hadi Esmaeilzadeh. 2017. Scale-Out Acceleration for Machine Learning. In *Proceedings of the International Symposium on Microarchitecture (MICRO)*.
- [66] PyTorch. 2019. <http://pytorch.org>.
- [67] Vijay Rao. 2019. Accelerating Infrastructure - together. <https://2019ocpglobalsummit.sched.com/event/jiis/accelerating-infrastructure-together-presented-by-facebook>.
- [68] Brandon Reagen, Paul Whatmough, Robert Adolf, Saketh Rama, Hyunkwang Lee, Sae Lee, Jose Miguel, Hernandez-Lobato, Gu-Yeon Wei, and David Brooks. 2016. Minerva: Enabling Low-Power, High-Accuracy Deep Neural Network Accelerators. In *Proceedings of the International Symposium on Computer Architecture (ISCA)*.
- [69] Minsoo Rhu, Natalia Gimelshein, Jason Clemons, Arslan Zulfiqar, and Stephen W. Keckler. 2016. vDNN: Virtualized Deep Neural Networks for Scalable, Memory-Efficient Neural Network Design. In *Proceedings of the International Symposium on Microarchitecture (MICRO)*.
- [70] Minsoo Rhu, Mike O'Connor, Niladrish Chatterjee, Jeff Pool, Youngeun Kwon, and Stephen W. Keckler. 2018. Compressing DMA Engine: Leveraging Activation Sparsity for Training Deep Neural Networks. In *Proceedings of the International Symposium on High-Performance Computer Architecture (HPCA)*.
- [71] Samsung. 2016. (8GB, 1Gx72 Module) 288pin Registered DIMM based on 4Gb E-die. (2016).
- [72] Ali Shafiee, Anirban Nag, Naveen Muralimanohar, Rajeev Balasubramanian, John Paul Strachan, Miao Hu, R Stanley Williams, and Vivek Srikumar. 2016. ISAAC: A Convolutional Neural Network Accelerator with In-situ Analog Arithmetic in Crossbars. *ACM SIGARCH Computer Architecture News* 44, 3 (2016), 14–26.
- [73] Ali Shafiee, Anirban Nag, Naveen Muralimanohar, Rajeev Balasubramanian, John Paul Strachan, Miao Hu, R Stanley Williams, and Vivek Srikumar. 2016. ISAAC: A Convolutional Neural Network Accelerator with In-situ Analog Arithmetic in Crossbars. In *Proceedings of the International Symposium on Computer Architecture (ISCA)*.
- [74] Hardik Sharma, Jongse Park, Divya Mahajan, Emmanuel Amaro, Joon Kyung Kim, Chenkai Shao, Asit Misra, and Hadi Esmaeilzadeh. 2016. From High-level Deep Neural Models to FPGAs. In *Proceedings of the International Symposium on Microarchitecture (MICRO)*.
- [75] Tensorflow. 2016. <https://www.tensorflow.org>.
- [76] Fengbin Tu, Weiwei Wu, Shouyi Yin, Leibo Liu, and Shaojun Wei. 2018. RANA: Towards Efficient Neural Acceleration with Refresh-Optimized Embedded DRAM. In *Proceedings of the International Symposium on Computer Architecture (ISCA)*.
- [77] Hossein Valavi, Peter J. Ramadge, Eric Nestler, and Naveen Verma. 2019. A 64-Tile 2.4-Mb In-Memory-Computing CNN Accelerator Employing Charge-Domain Compute. *IEEE Journal of Solid-State Circuits* PP (03 2019), 1–11. <https://doi.org/10.1109/JSSC.2019.2899730>
- [78] Ganesh Venkatesh, Eriko Nurvitadhi, and Debbie Marr. 2017. Accelerating Deep Convolutional Networks using Low-precision and Sparsity. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- [79] Paul N. Whatmough, Sae Kyu Lee, Hyunkwang Lee, Saketh Rama, David Brooks, and Gu-Yeon Wei. 2017. A 28nm SoC with a 1.2 GHz 568nJ/Prediction Sparse Deep-Neural-Network Engine with >0.1 Timing Error Rate Tolerance for IoT Applications. In *Proceedings of the International Solid State Circuits Conference (ISSCC)*.
- [80] Paul N. Whatmough, Sae Kyu Lee, Niamh Mulholland, Patrick Hansen, Sreela Kodali, David C. Brooks, and Gu-Yeon Wei. 2017. DNN ENGINE: A 16nm Sub- μ J Deep Neural Network Inference Accelerator for the Embedded Masses. In *Hot Chips: A Symposium on High Performance Chips*.
- [81] Shijin Zhang, Zidong Du, Lei Zhang, Huiying Lan, Shaoli Liu, Ling Li, Qi Guo, Tianshi Chen, and Yunji Chen. 2016. Cambricon-X: An Accelerator for Sparse Neural Networks. In *Proceedings of the International Symposium on Microarchitecture (MICRO)*.