

# ConCCL: Optimizing ML Concurrent Computation and Communication with GPU DMA Engines

Anirudha Agrawal

Advanced Micro Devices, Inc.  
anirudha.agrawal@amd.com

Shaizeen Aga

Advanced Micro Devices, Inc.  
shaizeen.aga@amd.com

Suchita Pati

Advanced Micro Devices, Inc.  
suchita.pati@amd.com

Mahzabeen Islam

Advanced Micro Devices, Inc.  
mahzabeen.islam@amd.com

**Abstract**—Concurrent computation and communication (C3) is a pervasive paradigm in ML and other domains, making its performance optimization crucial. In this paper, we carefully characterize C3 in ML on GPUs, which are most widely deployed for ML training and inference. We observe that while C3 leads to performance uplifts, the uplifts are far lower than ideal speedups (serial computation and communication versus maximum of computation or communication; all times from isolated executions). That is, C3 on average achieves only 21% of ideal speedup. This is so, due to known challenges of compute and memory interference between concurrent GPU kernels (that is, sharing of GPU's compute units, caches and HBM).

To attain better performance for C3, first, we evaluate dual strategies of schedule prioritization and careful resource partitioning of compute units on GPUs to push performance attained with C3 (on average 42% of ideal speedup). We also provide heuristics that can guide a runtime while employing these strategies. To further enhance C3 performance, we propose to mitigate C3 interference by offloading communication tasks to the GPU's DMA engines. To this end, we build concurrent communication collectives (ConCCL) proof-of-concepts that harness DMA engines for communication. We show how ConCCL considerably closes the gap between realized and ideal speedup for C3 (on average 72% of ideal speedup is realized, up to  $1.67\times$  speedup). Overall, our work makes a strong case for GPU DMA engine advancements to better support C3 on GPUs.

**Index Terms**—Concurrency, DMAs, GPU, ML

## I. INTRODUCTION

Large-scale machine learning (ML) models continue to harness distributed computing over increasingly large clusters of GPUs. A consequence of this is intermingling of computation and communication which, in absence of data dependencies, can be executed concurrently on the GPU. There are several ML algorithmic choices that lead to concurrent computation and communication, termed **C3** in this work, such as data-parallelism [1], fully sharded data parallel (FSDP) [2], nano-batching [3] and more. As such, characterization of and optimization for C3 on GPUs is important.

To achieve this, we first create a taxonomy of C3 and provide a detailed characterization of C3 anchored on this taxonomy. Given our ML focus, we study C3 manifestations wherein the computation kernel is a matrix-matrix multiplication (or GEMM) kernel and the communication kernel is a collective operation such as an all-gather among multiple GPUs. By analyzing C3 scenarios from the training of the LLaMA-70B and LLaMA-405B [4] models, along with some

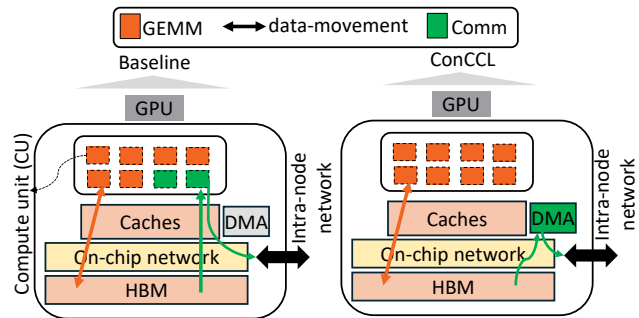


Fig. 1. Baseline C3 (left) and C3 with ConCCL via DMA offloads (right).

synthetic C3 scenarios, we provide broad coverage for our proposed taxonomy.

Next, using detailed experiments, we determine the compute and memory bandwidth requirements of isolated executions of the computation and communication kernels under study. This enables us to get an assessment of ideal performance uplift via concurrency. We observe that, when it comes to compute needs, on a state-of-the-art GPU such as AMD Instinct™ MI300X which has ample compute units (or GPU compute cores), communication kernels need only up to 10-20% of available units, making the rest available to concurrent computation kernels (Figure 1, left). GEMMs, as expected, manifest a spectrum of behaviors, some resilient to losing compute units to concurrent communication while the rest are sensitive. On the memory front, except for memory bandwidth-bound GEMM kernels, we observe that the bandwidth needs of compute-bound GEMMs and communication kernels can be met in tandem by the high memory bandwidth made available by MI300X GPUs. Overall, for C3 scenarios under study, compared to a baseline which serializes computation and communication, if the shorter of the computation or communication kernels is completely hidden, ideal speedups of  $1.6\times$  (average) and  $2\times$  (maximum) are possible.

However, here we observe that of this ideal speedup, on average only 21% ( $1.13\times$ ) is realized. This is not unexpected, as concurrent GPU kernels cause mutual interference with each other as they share compute cores, caches, high bandwidth memory (HBM) bandwidth and more. That is, the available resources for a given GPU kernel are lower in the presence

of other kernels than in isolation. In fact, prior work [5] has observed that due to such interference, C3 can even lead to slowdowns and loss of ML throughput.

To bridge the above performance gap, we first evaluate two strategies: schedule prioritization and careful partitioning of GPU compute units. Specifically, using our isolated execution analysis above, we observe that prioritizing scheduling of kernels with lower resource requirements prevents starvation and leads to overall better performance. Similarly, we also demonstrate that the resource partitioning available on MI300X GPUs, wherein certain compute units on GPUs can be exclusively marked for a given stream of work, can be harnessed to push C3 performance. Overall, using these strategies, we can push the performance achieved with C3 to an average of about 42% of available ideal speedup. Additionally, we also provide heuristics that can guide a runtime in employing these strategies.

Finally, to further enhance C3 performance, we tackle the interference incurred with C3 manifestations today. That is, for compute interference, instead of splitting available compute units among compute and communication kernels (Figure 1, left), we harness existing direct memory access (DMA) engines on the MI300X GPU and offload communication to them making all compute units available for concurrent GEMM computations (Figure 1, right). Given the placement of DMA engines in MI300X GPU (Section II-A), this has the added advantage of lowering interference in subset of caches as well. To this end, we build concurrent communication collectives (**ConCCL**) proof-of-concepts (PoCs) for ML collectives such as all-gather and all-to-all<sup>1</sup>. We demonstrate that our simple POCs are on par with existing communication libraries for bandwidth-bound scenarios while also reducing interference in C3, leading to attainment of on average 72% of ideal speedup, up to a maximum of  $1.67\times$  speedup. Overall, our work makes a strong case for continued investment and betterment of GPU DMA engines, and we conclude with a discussion of further GPU enhancements to support C3 efficiently.

The key contributions of this work are as follows:

- As concurrent computation and communication (**C3**) is an important paradigm in ML, high performance computing (HPC) and other domains, we present a detailed taxonomy and characterization of this paradigm on GPUs.
- Our analysis shows that while C3 can lead to performance uplifts compared to serial execution, not all of the potential speedup is realized (21% of ideal speedup is realized). This is not unexpected, as compute and memory interference among concurrent kernels causes slowdowns.
- To address the above gap, we first observe that C3 performance uplifts can be improved (42% of ideal speedup) by dual strategies of schedule prioritization and careful resource partitioning of compute units on GPUs.
- To attain further performance uplifts, unlike current state-of-art communication libraries which offload commu-

<sup>1</sup>Note that, as GPU DMA engines do not support arithmetic operations, we do not consider offloading all-reduce kernels to DMAs.

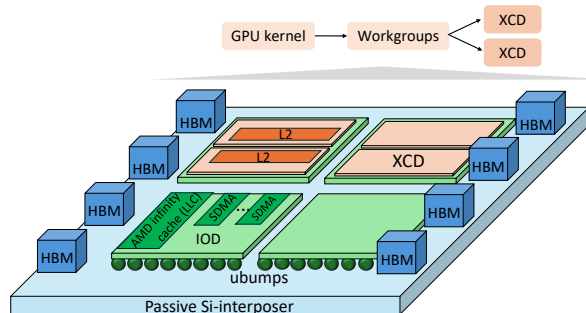


Fig. 2. State-of-art AMD Instinct™ MI300X.

nication to GPU compute-units, we build concurrent communication collectives (**ConCCL**) proof-of-concepts, which offload communication to DMA engines on GPUs.

- Using ConCCL, we demonstrate that C3 interference is lowered, considerably closing the gap between realized and ideal speedup (on average 72% of ideal speedup is realized).
- Overall, we make a strong case to enhance the capabilities of DMA engines on GPUs, as they can play an important role in efficiently supporting C3 in GPUs.

## II. BACKGROUND

### A. AMD Instinct MI300X GPU and Compute Orchestration

In this paper, we study C3 using AMD Instinct MI300X GPU depicted in Figure 2. A single MI300X GPU employs advanced packaging to integrate heterogeneous chiplets. Specifically, each MI300X is comprised of eight accelerator complex dies (XCD) [6] vertically stacked over four I/O dies (IOD), two XCDs per IOD [6], [7].

The IODs are comprised of AMD Infinity Cache™, a shared memory-side last-level cache (LLC) which is 256MB in size. Also, the IODs contain the memory interface to the on-package HBM. Each MI300X has a total of eight HBM stacks, each with 24GB (total of 192GB) for a combined peak memory bandwidth of 5.3TB/s [8]. Additionally, the IODs also contain 14 DMA copy engines<sup>2</sup>, termed SDMA (or system DMA) which are available for intra-node data transfers between GPUs with necessary address mapping support.

Each XCD is comprised of 38 active compute units (CUs), which are highly threaded and parallel GPU processor cores. Each XCD also has shared L2 cache of 4MB shared across all CUs within the XCD. Overall, in an MI300X GPU, there are 304 CUs. Also depicted in Figure 2, is compute orchestration. Computations are offloaded to GPUs as *kernels*, each comprising multiple *workgroups* which are scheduled on available CUs across all XCDs.

Large-scale ML (the focus of this work) often employs multiple GPUs in tandem. In this work, we focus on the AMD

<sup>2</sup>AMD HSA runtime API [9] allows querying of available DMA engines on MI300X.

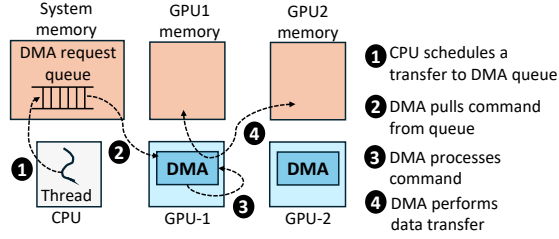


Fig. 3. Offloading a data-transfer to DMA in MI300X.

MI300X Infinity Platform comprised of a 8x MI300X node with a fully-connected topology. Each MI300X connects to seven other MI300X GPUs using AMD Infinity Fabric™ [8] bi-directional links (each with uni-directional bandwidth of 64GB/s/link).

### B. Direct Memory Access (DMA) Engines in GPUs

As discussed above, each MI300X is comprised of 14 SDMA copy engines available for transfers between address-space shared GPUs. We depict the steps involved for users to use these DMA engines in Figure 3. Using either heterogeneous-computing interface for portability (HIP) [10] or heterogeneous system architecture (HSA) [9] runtime API calls, at user-level, a programmer requests a single data transfer to be done using SDMA engines. Under the hood, this causes the runtime on CPU to place a command packet (1) in the DMA queue placed in system memory for a specific GPU (either source or destination GPU for the transfer). The GPU DMA engine gets notified and fetches the command from the queue (2) and processes it (3). Once decoded, the DMA engine issues necessary reads/writes from/to HBM memory of source/destination GPU respectively to complete the transfer (4).

### C. ML Operators and C3 in ML

As large-scale ML training and inference increasingly rely on GPU clusters, GPUs need to support both efficient computation and efficient communication. Furthermore, there are several algorithmic choices in ML that lead to *concurrent* computation and communication. C3 entails a pair of computation kernel and communication kernel that have no data dependencies and as such as be scheduled concurrently on a given GPU.

Examples of such algorithmic choices include data-parallelism [1], FSDP [2], nano-batching [3] and more. In data parallelism, the backpropagation phase of ML training comprises matrix-matrix multiplication gradient kernels (both input and weight gradients), which can be scheduled concurrently with reduction (communication) of weight gradients across GPUs for intra-layer concurrency (input gradient) or inter-layer concurrency (weight gradient of previous layer). Similarly, techniques like FSDP, gather model weights for a given layer on a GPU (communication), while performing computations of previous layers. Finally, optimizations

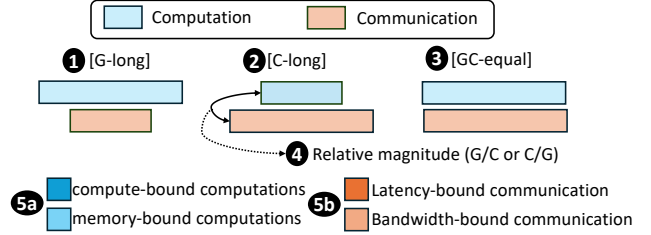


Fig. 4. C3 taxonomy.

like nano-batching [3] break down a single batch of inputs into nano-batches opening up opportunity for compute and communication kernels from different nano-batches to be co-scheduled on a GPU. Overall, given the ample prevalence of C3 in ML, characterization and optimization of C3 on GPUs is important and hence the focus of this work.

## III. C3 TAXONOMY

We begin with a taxonomy for C3 and subsequently anchor our characterizations and optimizations on this taxonomy.

As discussed in Section II-C, the focus of our work is on the manifestations of C3 in ML. More specifically, while there exist a variety of operators/computations in ML, in this work we focus on two primary operators, namely general matrix-matrix multiplications (or GEMM) kernels and communication kernels, as they contribute to the majority of the ML execution time in both training and inference scenarios across a spectrum of models [11]. Based on this, we depict our proposed C3 taxonomy in Figure 4. Specifically, we consider three key types of C3: 1 **G-long**, 2 **C-long** and 3 **GC-equal**.

We use execution times in isolation for our taxonomy. Thus, G-long is a C3 manifestation where GEMM time in isolation is  $>115\%$  of communication time. Similarly, C-long implies communication time in isolation is  $>115\%$  of GEMM time. Finally, GC-equal is a C3 manifestation where both kernels are comparable (within 15% of each other).

Even within this framework, we further observe that the relative magnitude (4 in Figure 4) of computation and communication kernels with respect to each other has an effect on expected interference incurred and as such performance attained. That is, when two kernels run concurrently on a GPU, they interfere with each other in both compute (splitting of compute units) and memory subsystem (caches, HBM). As such, we consider this part of our taxonomy and vary the relative magnitudes of the C3 scenarios under study in our work.

Finally, not all GEMM or communication kernels are the same. For GEMMs, we consider two broad categories of *compute-bound* and *memory-bound* (5a in Figure 4) kernels. We define a kernel to be compute-bound if its measured op-to-byte ratio is larger than machine op-to-byte as calculated from peak compute and memory throughput of underlying processor (kernel is memory-bound otherwise). For communication, we consider two types of commonly used multi-GPU ML collective kernels - *all-gather* and *all-to-all*. Similar to GEMM

TABLE I  
COMPUTATIONS (GEMMS) STUDIED, TAGS AND SOURCE.

gemm-tag	gemm-size	source
cb1	8192x8192x8192	LLaMA-70B
cb2	16384x8192x16384	LLaMA-405B
cb3	16384x16384x8192	LLaMA-405B
cb4	18432x8192x16384	LLaMA-405B
cb5	106496x8192x16384	LLaMA-405B
mb1	8192x57344x8192	LLaMA-70B
mb2	16384x106496x8192	LLaMA-405B

kernels, such ML collective kernels, depending on communication size, can be *latency-bound* or *bandwidth-bound* (5b). We classify a communication kernel with its associated size as latency-bound if the kernel latency at/before this size does not increase commensurate to size. We further discuss the specific sizes and C3 manifestations that we consider to cover this taxonomy, the source (ML model) for these manifestations and rationale in Section IV-A.

#### IV. BASELINE C3 CHARACTERIZATION

##### A. Methodology

1) *System Setup*: As discussed in Section II-A, we study C3 in ML using the AMD MI300X Infinity Platform comprised of 8x MI300X node with a fully connected topology.

Further, recall that we focus on two primary ML operators, namely, GEMM kernels and for communication, ML collectives kernels such as *all-gather* and *all-to-all*. For former, we employ AMD ROCm<sup>™</sup> [12] rocBLAS library [13] comprised of high-performance GEMM kernels. For ML collectives, we employ AMD ROCm Communication Collectives Library or RCCL [14], a library of standard collective communication routines for GPUs.

We use multiple GPU streams [15] or independent set of GPU kernels that can be co-scheduled on the GPU to concurrently launch GEMM (computation) and ML collective (communication) kernels by scheduling each type of kernel in its independent stream. Additionally, we leverage the feature available on MI300X GPU to reserve CUs for specific stream to study compute needs of kernels. We use rocpf [16] GPU kernel performance profiling tool to assess kernel execution times for this study. We run total of 15 executions, first 6 are warm ups and then 9 are actual measured.

2) *C3 Manifestations Under Study*: We discuss the GEMM kernels we study in this work, their sizes and their sources in Table I. As listed, we source our GEMM sizes from training of LLaMA-70B and LLaMA-405B [4] models processing 8192 tokens (i.e., product of input length and batch-size) in a given iteration. Further, for ease of reading, we tag these GEMM sizes/kernels as compute-bound (**cb**) or memory-bound (**mb**). Recall that, we define a kernel to be compute-bound if its measured op-to-byte ratio is larger than machine op-to-byte as calculated from peak compute and memory throughput of underlying processor (kernel is memory-bound otherwise).

TABLE II  
C3 COMBINATIONS CONSIDERED AND TAXONOMY.

##### C3-type: G-long

C3	source	C3	source
mb1_896M	LLaMA-70B	mb2_3.25G	LLaMA-405B
mb1_4G	synthetic	mb1_6G	synthetic
cb3_512M	LLaMA-405B	cb4_512M	LLaMA-405B
cb5_1.63G	LLaMA-405B	cb4_1G	synthetic

##### C3-type: C-long

mb1_13G	synthetic	cb2_3.25G	LLaMA-405B
cb4_2.5G	synthetic	cb1_896M	LLaMA-70B
cb5_20G	synthetic		

##### C3-type: GC-equal

mb2_26.5G	synthetic	cb5_13G	synthetic
-----------	-----------	---------	-----------

We list the C3 manifestations we study in this work in Table II. We first provide a tag for C3 being studied using GEMM type followed by the array/data size of the parallel collective operation. This is listed in column **C3** in the table. Further, in all our analysis, we separately present the results for different collective types (all-gather, all-to-all). We also list in the table the **source** for each C3 manifestation. As shown, of 15 unique C3 combinations we have, seven are manifested in training of LLaMA-70B and LLaMA-405B [4] models (we assume 8-way sharding and FSDP [2]) using all-gather collective. To provide good coverage for our proposed taxonomy in Section III, we add additional synthetic C3 manifestations, wherein, we keep the GEMM kernel size as observed in LLaMA models (Table I) but add more communication sizes and repeat all C3 scenarios for all-to-all collective as well.

We also list the taxonomy for each C3 in Table II. Notice that we have more G-long scenarios than C-long scenarios than GC-equal scenarios. This is so, as shown in Table II, majority of C3 manifestations in LLaMA models today are of G-long type and we consciously wanted to limit the synthetic scenarios we added. However, we provide all our results grouped by C3 type to understand the performance for each of these different types.

##### B. Isolated Execution Characterization

We first profile the computation and communication kernels in isolation to ascertain their compute and memory needs and assess potential for performance when they are executed concurrently.

1) *Compute Needs*: Recall that, to execute computation and communication concurrently on GPUs, we launch two independent kernels and as such available compute cores or compute units (CUs) for MI300X (Section II-A) are split between these two kernels. This compute interference can slowdown each of these kernels. We study this in Figure 5 (a) for GEMM kernels and (b)/(c) for subset of all-gather and all-to-all sizes (rest of the sizes show similar behavior).

For GEMM kernel slowdown, we depict two of the available seven GEMMs under study (Table I) in Figure 5(a), as they represent the extremes in terms of slowdown. To plot slowdown, we compare a GEMM execution when the specified

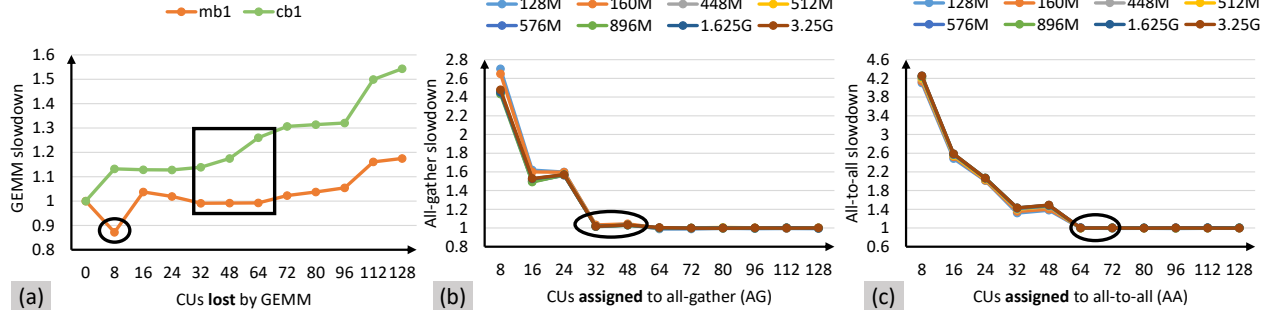


Fig. 5. (a) GEMM kernel slowdown due to loss of compute units (CUs) in GPU. (b) All-gather, (c) All-to-all kernel slowdown with specific # CUs assigned vs. default CUs (All-gather default #CUs=64, All-to-all default #CUs=56). For single partition MI300X with eight XCDs, eight is the minimum number of CUs that can be assigned to a kernel.

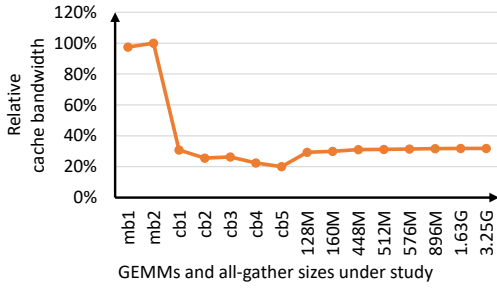


Fig. 6. Relative AMD Infinity Cache™ bandwidth utilization.

number of CUs are taken away from the kernel to an execution where all CUs are available to the kernel (0 on x-axis is when the kernel has all 304 CUs available to it). As depicted in the figure, memory-bound GEMMs are resilient to CU loss, even attaining speedups<sup>3</sup> (highlighted with a circle), while compute-bound GEMMs can incur increasing slowdowns as more CUs are taken away from the GEMMs. As for communication, we do a similar slowdown analysis in Figure 5(b)/(c) and observe that unlike GEMMs, all-gather kernels need 32 CUs, while all-to-all kernels need 64 CUs beyond which there is no benefit to allocating more CUs to the kernel.

Overall, considering the compute needs of these kernels in tandem, it can be expected that memory-bound GEMMs can co-exist with communication kernels, because memory-bound GEMMs are resilient to loss of about 32-64 CUs needed by communication kernels. On the other hand, for the same CU loss, compute-bound kernels suffer up to 17-27% slowdowns.

2) *Memory Needs*: Similar to compute needs analysis above, we next consider memory needs of these kernels. As discussed in Section II-A, MI300X GPU employs a shared last-level AMD Infinity Cache which, being memory-side, subsumes all of HBM traffic. As such we study the memory bandwidth needs of kernels under study by providing their

<sup>3</sup>We observe better cache behavior for this GEMM due to less number of concurrent threads.

relative AMD Infinity Cache bandwidth utilization in Figure 6. Note that, we only show all-to-all kernels and skip all-gather kernels as the latter have about 14% lower bandwidth vs. all-to-all kernels.

As depicted in Figure 6, the bandwidth needs of memory-bound GEMM kernels dwarf that of all other kernels under study. That said, the compute-bound GEMM and communication kernels are in similar ballpark of bandwidth needs. Further, given that the combined bandwidth needs of compute-bound GEMM kernels and communication kernels still leaves headroom in peak bandwidth available, these kernels can share available bandwidth in concurrence.

3) *Ideal speedup projection*: Based on above isolated execution analysis of compute and memory needs, we project in Figure 7 the ideal speedup possible for C3 scenarios under study. For this analysis, we assume that the best speedup possible is when the smaller of the two kernels (computation or communication) happens completely in the shadow of the other. As depicted, wide variety of speedups are possible ( $1.1\times$  to close to  $2\times$ ) and this is largely dictated by relative magnitudes of these kernels (Section III). However, based on compute/memory needs analysis, due to compute/memory interference not all of this ideal performance can be attained.

Further, note that, based on runtime decisions or GPU-GPU execution variation or kernel-size/type, different degrees of overlap can manifest, resulting in different ideal speedups. To precisely tackle this we design C3 taxonomy so that we can consider such varied manifestations and analyze our proposals across these different C3 manifestations.

### C. C3 Characterization

We present characterization of C3 for different scenarios under study in Figure 8. In this figure we refer to baseline C3 performance as **c3\_base**. Note that, with concurrent scheduling, we schedule GEMM kernel first (as scheduling both the kernels at the exact same time is challenging, we minimized the scheduling delay between the two through code optimizations) in **c3\_base** executions.



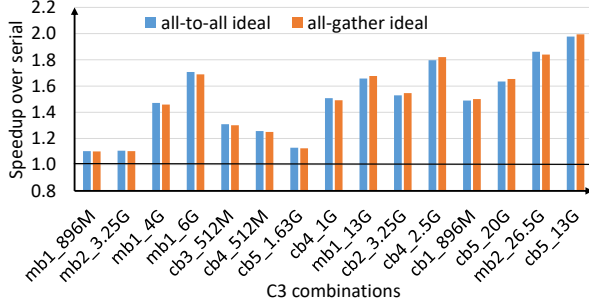


Fig. 7. Ideal speedup possible for C3 scenarios under study.

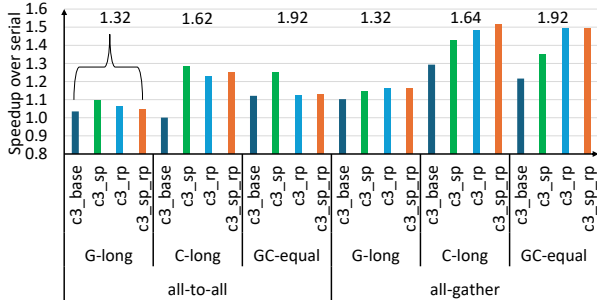


Fig. 8. Speedups for C3 scenarios under study with and without schedule prioritization and resource partitioning.

In Figure 8, we present average speedups for different groupings of C3 scenarios (collectives, and taxonomy). We also show in the graph the ideal speedups possible (marked at the top of the graph). The figure shows that **c3\_base** speedups range from no speedups to up to  $1.3\times$  speedups. On average, **c3\_base** attains  $1.13\times$  speedup which is about 21% of ideal speedup. Note that, this is not unexpected. In fact, prior work [5] observed slowdowns with C3 due to mutual interference of concurrent kernels.

Further, all-to-all attains about 0-13% of ideal speedups while all-gather attains about 24-46%. The slight edge for all-gather is due to the fact the overall memory traffic (hence memory interference) and compute needs are lower for all-gather vs. all-to-all (Section IV-B1). To understand lower memory traffic of all-gather, in simplistic terms, in all-gather, each GPU begins with a single data buffer and the end state is that every GPU holds the complete aggregated data from all other GPUs. In contrast, with all-to-all, each GPU starts with distinct data buffer for every other GPU (lot more data) and concludes with each GPU receiving unique data from all other GPUs (effectively a transpose of data buffers).

## V. OPTIMIZING C3 PERFORMANCE

Next, to improve C3 performance reported above, we evaluate dual strategies of schedule prioritization and careful resource partitioning of compute units on GPUs.

### A. Schedule Prioritization

Note that, thus far, for concurrent scheduling, we chose to schedule GEMM kernel first. But based off our compute needs analysis, we observe that communication kernels have far lower CU needs as compared to GEMM kernels (Section IV-B1). As such, if GEMM kernels are scheduled first, the internal GPU scheduler can in some cases allocate majority of CUs to GEMM kernel leading to potential starvation. In contrast, if we employ, *schedule prioritization*, that is, schedule communication kernel first, later scheduled GEMM kernel will definitely have compute resources to be allocated. Overall, the key insight of schedule prioritization is that prioritizing or providing quality of service for kernel with smaller and complementary resource requirement helps concurrent performance.

To realize schedule prioritization, from CPU-side, we first schedule communication kernel in a stream and then immediately after, schedule GEMM kernel in concurrent stream. We present the results of schedule prioritization, referred in Figure 8 as **c3\_sp** and show the speedups attained. Across all C3 types and collectives, we observe that schedule prioritization improves the speedups attained for C3. Overall, instead of 0-13% of ideal speedup, for all-to-all, **c3\_sp** attains 27-46% instead. Similarly, instead of 24-46% of ideal speedup, for all-gather, **c3\_sp** attains 38-67% of ideal speedup. Overall, on average, **c3\_sp** helps attain 42% of ideal speedup (up from 21% of **c3\_base**).

### B. Resource Partitioning

Note that schedule prioritization is not the only way to prevent starvation of kernel with lower resource needs. We also observe that we can leverage the feature available on MI300X GPU to reserve compute units (CUs) for specific stream to more deterministically allocate CUs to communication kernels and attain similar benefits.

We show this in Figure 8 as **c3\_rp** which adds resource partitioning (rp) to **c3\_base**. We sweep all possible powers-of-two CU allocations for communication kernels (and consequently take CUs away from GEMM kernel) and plot the best performing one as **c3\_rp**. As depicted, **c3\_rp** delivers performance improvements over **c3\_base** which is almost the performance delivered by **c3\_sp** (41% of ideal speedup). Note, that, adding resource partitioning to **c3\_sp**, depicted as **c3\_sp\_rp**, did not improve performance any further.

### C. Runtime Heuristics for Improved C3 Performance

We believe that either schedule prioritization or resource partitioning can help improve C3 performance, while former is indeed simpler to implement. To this end, we provide some heuristics that can guide a runtime while employing them.

**Heuristic for Schedule Prioritization:** As runtimes launch GPU kernels, they can use the information about number of workgroups per kernel (Section II-A) as a proxy for CU requirement of a kernel. Using this information, excepting any other prioritization that supersedes, the runtime can employ

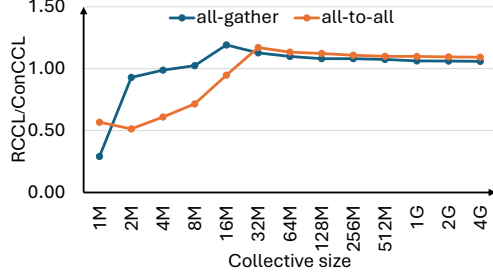


Fig. 9. ConCCL speedup over CU-based collective (RCCL).

scheduling order in the order of resource requirements (number of workgroups), low to high.

**Heuristic for Resource Partitioning:** For resource partitioning across concurrent kernels, we describe a runtime heuristic to allocate CUs across concurrent kernels. To do so, we picked one memory-bound GEMM kernel, one compute-bound GEMM kernel, and two all-gather and all-to-all collective sizes (latency-bound, bandwidth-bound) and used the compute need analysis done in Section IV-B1 to build a lookup table of potential slowdowns when the kernel loses a given number of CUs. Note that, for a given GPU this is to be done once.

Next, for any C3 scenario, we scale roofline GEMM and communication times by these slowdowns for different number of CUs to identify the CU allocation that leads to best performance, that is  $\max(\text{GEMM}, \text{communication})$  time is lowest. For roofline times, we simply focus on peak compute, memory and network throughputs and assume 70% efficiency (taking the average of our/observed [17] peak compute, memory and network efficiencies). We observe that this simple heuristic, predicts CU allocation necessary for 24 of 30 C3 scenarios. For the rest, in comparison to sweeping all possible CU allocations, our heuristic, at best loses 1.5%.

## VI. OPTIMIZING C3 WITH CONCCL

### A. Motivation for DMA Offloads of Communication

For co-scheduled computation and communication, the performance limiters are the compute and memory interference incurred by sharing of compute and memory resources between concurrent kernels. That is, a GEMM kernel in isolation could have been allocated all of 304 compute units (CUs) available on MI300X. In C3, however, number of CUs allocated will be lower causing what we term as compute interference and leading to lower speedups. Similar interference also occurs in the memory sub-system (e.g., caches, HBM).

One way to tackle compute interference completely and memory interference partially, is to offload communication to DMA engines on MI300X. By doing so, we can free up all available compute units for concurrent computation kernel. Further, as DMA engines are placed in IOD beyond L2 caches (Section II-B), this also eliminates any L1/L2 portion of memory interference. These dual benefits motivate us to offload communication to DMA engines on MI300X.

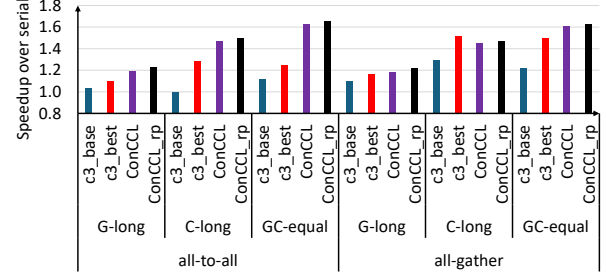


Fig. 10. C3 speedup with ConCCL.

### B. ConCCL: DMA-collectives PoCs

With the above strong motivation, we build **ConCCL** proof-of-concepts (PoCs), which are DMA-based collectives wherein we offload the collectives to DMA engines on MI300X. As MI300X DMA engines do not expose any computational functionality, we only build these PoCs for all-gather and all-to-all (and not all-reduce collective which involves reduction).

We harness the fully-connected topology of MI300X to keep the design of collectives simple. That is, we break down the collective operation into a series of individual transfers (going to different GPUs) and we schedule each such transfer on a specific available DMA engine on MI300X (recall that 14 engines are available). We leverage AMD HSA API call `hsa_amd_memory_async_copy_on_engine` [18] to schedule a single transfer at a given engine. Also, we implement a simple direct algorithm for our collectives. As an example, for all-gather this means that our collective implementation comprises a single step: every GPU reads the data buffer it owns and writes it to every other GPU. Additionally, unlike RCCL communication kernels which are orchestrated by launching GPU kernels, as discussed in Section II-B, DMA engine transfers are orchestrated by CPU.

Finally, we term ConCCL as PoCs for our goal in this work is not to build a high-performant communication collectives library but to use these PoCs to analyze if offloading concurrent communication collectives to DMA engines helps improve speedups possible with C3. We leave building a high-performance DMA-based collectives library to future work.

### C. Isolated ConCCL Characterization

We first provide a comparison of our PoCs to state-of-art RCCL communication library on MI300x. We show this comparison in Figure 9. We follow the same setup details as outlined in Section IV-A1 and model multiple warm-up executions before actual measured executions.

As shown in Figure 9, our simple PoCs are slower than RCCL library for <32MB collective size by as much as 4 $\times$ . This is so, as it can be costly to both launch transfers from CPU to DMA engines on GPU and synchronize with CPU once the transfers are done. This launch/sync cost is not amortized for smaller sizes. That said, for larger sizes, our proposed all-gather PoCs is at par with RCCL library. Finally, recall that, as the smallest communication size we consider

in our C3 scenarios is 128MB (Table II), evaluating C3 with RCCL or ConCCL is a fair evaluation as in this region the performance of both is at-par.

#### D. C3 with ConCCL Characterization

Next, we evaluate the speedups of C3 with ConCCL over sequential execution. We depict this in Figure 10.

We first compare **c3\_base** configuration, which is C3 without schedule prioritization or resource partitioning, to ConCCL without such optimizations. As shown, across the board, ConCCL delivers higher speedups in comparison to **c3\_base**. This is attributed to lowering the compute interference incurred by GEMM kernels in **c3\_base**. Additionally, GEMM kernels also do not incur any memory interference in any higher level of caches such as L1/L2 with DMA offloads.

Overall, while **c3\_base** attains 21% of available ideal speedup, with DMA offloads, ConCCL attains 66% of ideal speedup. Further, ConCCL benefits are even more pronounced for all-to-all (**c3\_base**: 1.05 $\times$ , ConCCL: 1.43 $\times$ ) which has higher compute needs (64 CUs) and as such incurs more compute interference and also has higher memory traffic/interference in higher level caches.

#### E. Schedule prioritization with ConCCL

Unlike **c3\_base**, where both computation and communication are scheduled on GPU CUs making schedule prioritization important, with ConCCL, as communication is scheduled to DMA engines and computation to CUs (two separate entities), schedule prioritization is not necessary for ConCCL.

#### F. Resource partitioning with ConCCL

Finally, we consider resource partitioning for ConCCL. Since ConCCL does not use any compute units for communication, the allocation decision here is different from **c3\_base**. Here we observe that, *only* for memory-bound GEMM kernels, as depicted in Figure 5, taking CUs away from GEMM kernel can lead to performance improvement due to improved cache behavior. Such improvement can also aid in C3 runs where communication is offloaded to DMA engines.

To study this, only for memory-bound GEMM kernels, we create a variant of ConCCL, **ConCCL\_rp** and depict its performance in Figure 10. This variant performs slightly better than ConCCL. Point to note that, both ConCCL and **ConCCL\_rp** performs significantly better than **c3\_best** variant (best of all **c3** variants in Section IV). We see here that on average, while **c3\_best** attains 48% of ideal speedup, ConCCL attains 66% and **ConCCL\_rp** attains 72% of ideal speedup, considerably closing the gap between attained and ideal speedup.

#### G. Runtime Heuristic for Resource partitioning with ConCCL

For resource partitioning with ConCCL, we recommend a much smaller subset of steps as proposed in Section V-C for baseline C3. Specifically, only the CU-loss slowdown table for a memory-bound GEMM kernel ought to be created. This table helps the runtime identify the necessary number of CUs

to remove from concurrent memory-bound GEMM kernel to maximize speedup. In our analysis, for MI300X, taking away eight CUs lead to speedups for memory-bound GEMM kernels.

## VII. DISCUSSION

### A. System Evolution for Efficient C3

While ConCCL leads to more efficient C3 on GPUs, we believe additional system evolution can further help in this regard. We highlight some potential techniques below.

1) *Addressing Memory Interference in C3*: Our approach mitigates compute interference by offloading communication to the GPU's DMA engines. On the MI300X GPU, DMA transfers inherently do not interact with the L1 and L2 caches, effectively eliminating cache interference at these levels. That said, contention for HBM bandwidth remains, impacting performance. We leave to future work exploration of techniques such as memory-bandwidth partitioning via memory space (channel) partitioning amongst kernels, memory-aware scheduling [19] to explicitly tackle memory interference. Specifically for latter, C3 taxonomy hints to hardware can aid in memory traffic prioritization.

2) *Accelerating C3 with All-Reduce Collective*: Our ConCCL PoCs focus on offloading all-gather and all-to-all collectives to DMA engines. However, all-reduce operations involve both communication and computation (e.g., summing values across GPUs) and since DMA engines do not currently support arithmetic operations we could not offload all-reduce collectives to DMA engines. That said, all-reduce is also used in ML and offloading this to DMA engines can be useful. Investigating addition of arithmetic units to DMA engines is a possibility although the area/power costs of doing so ought to be balanced with possible returns. Alternately, a hybrid approach can be followed. That is, as all-reduce is comprised of a reduce-scatter and all-gather operation, only for latter DMA engines can be harnessed.

Further, we believe that our proposed techniques of schedule prioritization and resource partitioning are also applicable to all-reduce. This is so, as the key insight of the techniques is that prioritizing or providing QoS for kernel with smaller/complimentary resource requirement helps concurrent performance. As all-reduce has this behavior just like all-gather/all-to-all vs. GEMMs (low CU needs, etc.), these techniques can be applicable to all-reduce as well.

### B. Other considerations

1) *Generalizing C3 Heuristics*: We propose in this work heuristics for both schedule prioritization (SP) and resource partitioning (RP) for two concurrent kernels. Our SP heuristic can be extended to more kernels, that is, instead of two kernels, runtime can prioritize scheduling of multiple kernels again in the order of their resource requirements (number of workgroups), low to high. Similarly, our RP heuristic timing analysis can be extended to more kernels. As for two kernels, we can slowdown multiple kernels using the proposed



analytical model and assess CU allocation that leads to best performance. However, this model does not factor in the increasing memory interference with more concurrent kernels and further investigation can be necessary in this regard. We leave evaluating these heuristics for more concurrent kernels to future work.

2) *Caching Considerations*: On MI300X, switching from CU-compute to DMA transfers does not incur any extra cache write-backs. This is so, as L2 is private per XCD (Figure 2), it is already written-back to AMD Infinity Cache at GPU kernel completion. Further, sDMA engines read data from AMD Infinity Cache and write to HBM allowing subsequent GPU kernel to read sDMA transfer's output. That said, in alternate architectures, preceding GPU kernels can proactively write to common cache levels between compute and DMA engines preventing any cache-related overheads (e.g., cache write-back, flush) from happening on the critical path.

3) *Inter-Node Communication Considerations*: Our study primarily addresses intra-node communication within an MI300X Infinity Platform node. While this is most prominent for inference and small-scale training, large-scale distributed ML training involves inter-node communication as well. In such cases, ML algorithms attempt to use both intra and inter node communication in tandem with attempting to maximize former as much as possible for better performance. This is done as intra-node bandwidth is significantly higher than inter-node bandwidth. In such cases, even with a focus on intra-node communication only, ConCCL can be useful end-to-end. Further, large-scale multi-node ML employ hierarchical collectives which break down collectives into intra and inter-node steps [20] and ConCCL can be utilized for intra-node steps.

4) *Global/Local Optimization*: Our proposed heuristics/techniques can in some sense be classified as local optimization strategies. That is, we identify given two or more concurrent kernels, which of schedule prioritization or resource partitioning gets better performance. We understand that for better end-to-end performance for an application, with many kernels, such local decisions have to be combined at global-level to get better performance. Prior works in this regard [21] can be used in tandem with our work to do so.

5) *Power Considerations*: As GPUs get more capable, power constraints are increasingly at the forefront in their design. A power-agnostic scheduler could, by over-employing C3, lower performance by causing GPU power to be stressed leading to power management events. Alternately, a more power-aware scheduler can employ C3 more judiciously by prioritizing concurrency for complementary power kernels. We leave design of such heuristics to future work.

6) *Implications of CPU orchestration*: As discussed in Section VI-C, DMAs are orchestrated using CPU, the launch/sync cost for transfers are not amortized at smaller sizes. A GPU control-path might help solve this problem and we leave investigating software/hardware optimizations to address this as part of future work.

## VIII. RELATED WORK

Given the prevalence of C3 in ML, optimized C3 is key to continued scaling of ML. In our work, we methodically study C3, analyze resulting interference and most importantly, without making any invasive changes in existing system software and hardware, we identify techniques to use available software and hardware resources in optimized way to maximize benefit of C3.

**Computation-Communication Kernel Overlap**: Several GPU communication kernel libraries have been developed which improve collective performance in isolation - RCCL [14], MSCCLang [22], and more [23]–[25]. However, none of these take any special measures to mitigate any interference where collective kernels execute concurrently with computation kernels which are abundant in ML executions. While some prior works study this overlap algorithmically, they do not provide insights into the resource contention therein [11]. Other works study C3 scenarios prevalent in DLRM recommendation [26] and Megatron-LM Transformer [27] model executions using both microbenchmarks as well as end-to-end models executions [28]. Specifically, they study all-reduce slowdowns in concurrence with GEMMs, embedding table lookup and other backprop compute operations. Their study finds that all-reduce can slowdown on an average  $1.4\times$  in the Transformer model when overlapped with backprop operations, and by up to  $6.2\times$  when overlapped with both GEMMs and memory-bound embedding lookup in DLRM. Our work expands this to systematically study execution of collectives (from RCCL [14] which incorporates optimizations from other libraries for an AMD MI300X GPU) in concurrence with compute operations. It studies additional collectives such as all-gather and all-to-all, and furthermore, proposes a taxonomy to incorporate a wide-range of possible C3 scenarios. This taxonomy enables the study of C3 scenarios with different overlap characteristics (e.g., G-long, C-long) and includes communication-computation pairs that are not only prevalent in models today, but can be possible in future models and/or distributed setups. Finally, we not only study these contentions, but also develop two optimizations including, schedule prioritization and compute (CU) partitioning, that improve end-to-end performance of C3 without any additional accelerator and/or hardware changes.. While there have been numerous works on improving concurrent kernel execution on GPUs via appropriate resource partitioning [29], [30], prioritization [21] and concurrency-aware kernels [31]–[33], they primarily focus on multiple compute kernels. Our work applies these techniques to coarse-grain compute-communication kernels. NanoFlow [3] also performs resource partitioning for C3; however, they create very fine-grained kernels by slicing ML input from micro into nano-batches, and find necessary resource allocation from a huge and complex search space unlike our simple yet useful lookup based heuristic.

**Communication Offload**: Works which improve C3 performance usually offload communication to dedicated customized accelerators on GPUs. For example, ACE [28] is a

custom accelerator for offloading communication which can free up compute units for concurrent compute. It further buffers intermediate, partially reduced data to avoid their reads and writes to memory which reduces memory interference. Similarly, other works require extensive hardware support such as compute-capable switches to offload reduction operations while reducing volume of data moved over network links and memory sub-system for all-reduce [34]. Furthermore, switch offloads still require GPU kernels to orchestrate data movement and in-switch commands, leading to interference with concurrent computation. We consider offload but by leveraging *existing* data-movers (DMAs), requiring no hardware changes and show C3 benefits with two key ML collectives, all-gather and all-to-all.

**Communication Offload to DMAs:** There are also work which offload communication to DMA engines. For example, MSCCL++ [25] offloads larger size collective operations to DMA engines, however it initiates DMA call from GPU involving CU (through proxy channel in CPU), which in our work we prefer to reserve only for compute operations. ARK [5] suggests communication offload to DMA but emphasizes on DMA call overhead time and involves GPU threads to reduce such overhead either by significant SW changes or by addition of new DMA prototype; we, in contrast rely on existing software stack and HW and still show benefit of sDMA offload. Some of the works [35] leverage DMAs for fine-grained C3. However, unlike this work, they provide limited insights into challenges with such overlap and therefore provide limited insights into the potential for DMAs alone to help overcome C3 interference like this work highlights with ConCCL. Other works require hardware modifications to initiate DMA transfers in a fine-grained manner [19].

**Fine-grained Computation-Communication Overlap:** Other interesting works discuss techniques for fine-grain overlap of compute/communication to improve concurrency [3], [19], [35]–[38]. While such fine-graining techniques are promising, they necessitate pervasive changes in SW/HW to realize overall performance benefit. For example, Fused Computation-Collective [36] fuses communication operation directly in GEMM kernel requiring exponential implementations in GEMM library.

## IX. CONCLUSION

In this work, we carefully characterize the performance of concurrent computation and communication (C3), an important ML paradigm, on state-of-art MI300X GPU and observe that speedups attained are, not unexpectedly, on average only 21% of ideal speedup possible due to interference between concurrent GPU kernels. To improve performance of this important primitive, we first evaluate dual strategies of schedule prioritization and careful resource partitioning of compute units on GPUs to push performance attained with C3 (on average 42% of ideal speedup). Additionally, we provide heuristics that can guide a GPU runtime to harness these strategies. To further improve speedups for C3, we offload

communication to DMA engines via building **ConCCL** (Concurrent Communication Collectives) proof-of-concepts. We demonstrate how ConCCL closes the gap between realized and ideal speedup for C3 (on average about 72% of ideal speedup is realized, up to  $1.67\times$  speedup). We believe DMA engines have an important role to play for C3 performance and argue for their continued betterment.

## ACKNOWLEDGMENT

We thank our colleagues Wenkai Du, Gilbert Lee, Alexander Kaganov, Leo Dong, Anthony Asaro, Padmini Nujetti for many helpful discussions. We also thank Nicholas Malaya, Alex Habeger, Kevin Cheng, Gowri Shankar Guttikonda for help with equipment. Finally, we also thank Gabriel Loh and the anonymous ISPASS reviewers for helping improve the paper. AMD, the AMD Arrow logo, AMD ROCm, AMD Instinct, AMD Infinity Cache, AMD Infinity Fabric, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies.

## REFERENCES

- [1] L. G. Valiant, "A bridging model for parallel computation," *Commun. ACM*, vol. 33, no. 8, p. 103–111, Aug. 1990. [Online]. Available: <https://doi.org/10.1145/79173.79181>
- [2] Y. Zhao, A. Gu, R. Varma, L. Luo, C.-C. Huang, M. Xu, L. Wright, H. Shojanazeri, M. Ott, S. Shleifer, A. Desmaison, C. Balioglu, P. Damania, B. Nguyen, G. Chauhan, Y. Hao, A. Mathews, and S. Li, "Pytorch fsdp: Experiences on scaling fully sharded data parallel," 2023. [Online]. Available: <https://arxiv.org/abs/2304.11277>
- [3] K. Zhu, Y. Zhao, L. Zhao, G. Zuo, Y. Gu, D. Xie, Y. Gao, Q. Xu, T. Tang, Z. Ye, K. Kamahori, C.-Y. Lin, S. Wang, A. Krishnamurthy, and B. Kasicki, "Nanoflow: Towards optimal large language model serving throughput," 2024. [Online]. Available: <https://arxiv.org/abs/2408.12757>
- [4] A. M. Llama Team, "The llama 3 herd of models," 2024. [Online]. Available: <https://arxiv.org/abs/2407.21783>
- [5] C. Hwang, K. Park, R. Shu, X. Qu, P. Cheng, and Y. Xiong, "{ARK}:{GPU-driven} code execution for distributed deep learning," in *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*, 2023, pp. 87–101.
- [6] A. Smith, E. Chapman, C. Patel, R. Swaminathan, J. Wu, T. Huang, W. Jung, A. Kaganov, H. McIntyre, and R. Mangaser, "11.1 amd instinctmi300 series modular chiplet package – hpc and ai accelerator for exa-class systems," in *2024 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 67, 2024, pp. 490–492.
- [7] A. Smith, G. H. Loh, J. Wu, S. Naffziger, T. Huang, H. McIntyre, R. Mangaser, W. Jung, and R. Swaminathan, "AMD Instinct™MI300X Accelerator: Packaging and Architecture Co-Optimization," in *2024 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, 2024, pp. 1–8.
- [8] AMD, "The AMD CDNA™ 3 architecture," <https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/white-papers/amd-cdna-3-white-paper.pdf>, 2024.
- [9] AMD, "HSA Runtime API and runtime for ROCm," <https://rocm.docs.amd.com/projects/ROCR-Runtime/en/latest/>, 2024.
- [10] —, "HIP: C++ Heterogeneous-Compute Interface for Portability," <https://github.com/ROCm/HIP>, 2024.
- [11] S. Pati, S. Aga, M. Islam, N. Jayasena, and M. D. Sinclair, "Tale of two cs: Computation vs. communication scaling for future transformers on future hardware," in *2023 IEEE International Symposium on Workload Characterization (IISWC)*, 2023, pp. 140–153.
- [12] AMD, "AMD ROCm™ Software," <https://www.amd.com/en/products/software/rocm.html>, 2024.
- [13] AMD, "ROCm/roBLAS: Next generation BLAS implementation for ROCm platform," <https://github.com/ROCm/roBLAS>, 2024.

- [14] —, “ROCm Communication Collectives Library (RCCL).” [Online]. Available: <https://github.com/ROCm/rccl>
- [15] —, “ROCm: HIPStream,” [https://rocm.docs.amd.com/projects/HIP/en/latest/reference/hip\\_runtime\\_api/modules/stream\\_management.html](https://rocm.docs.amd.com/projects/HIP/en/latest/reference/hip_runtime_api/modules/stream_management.html), 2024.
- [16] AMD, “rocprow — ROC Profiler Documentation,” <https://rocm.docs.amd.com/projects/rocprow/en/docs-5.5.1/rocprow.html>, 2024.
- [17] —, “AMD Instinct™ MI300X Accelerator Performance Validation Guide,” <https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/product-briefs/amd-instinct-mi300x-performance-validation-guide.pdf>, 2024.
- [18] AMD, “ROCm: ROCr-Runtime,” [https://rocm.docs.amd.com/projects/HIP/en/develop/doxygen/html/hsa\\_ext\\_amd\\_8h.html](https://rocm.docs.amd.com/projects/HIP/en/develop/doxygen/html/hsa_ext_amd_8h.html), 2024.
- [19] S. Pati, S. Aga, M. Islam, N. Jayasena, and M. D. Sinclair, “T3: Transparent tracking & triggering for fine-grained overlap of compute & collectives,” in *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, 2024, pp. 1146–1164.
- [20] S. Rajbhandari, C. Li, Z. Yao, M. Zhang, R. Y. Aminabadi, A. A. Awan, J. Rasley, and Y. He, “Deepspeed-moe: Advancing mixture-of-experts inference and training to power next-generation ai scale,” 2022. [Online]. Available: <https://arxiv.org/abs/2201.05596>
- [21] S. Kato, K. Lakshmanan, R. Rajkumar, and Y. Ishikawa, “TimeGraph: GPU Scheduling for Real-Time Multi-Tasking Environments,” in *Proceedings of the 2011 USENIX Conference on USENIX Annual Technical Conference*. Portland, OR: USENIX Association, Jun 2011.
- [22] M. Cowan, S. Maleki, M. Musuvathi, O. Saarikivi, and Y. Xiong, “MSC-CLang: Microsoft Collective Communication Language,” in *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, 2023, pp. 502–514.
- [23] NVIDIA, “NCCL.” [Online]. Available: <https://github.com/NVIDIA/nccl>
- [24] A. Shah, V. Chidambaram, M. Cowan, S. Maleki, M. Musuvathi, T. Mytkowicz, J. Nelson, O. Saarikivi, and R. Singh, “TACCL: Guiding Collective Algorithm Synthesis using Communication Sketches,” in *20th USENIX Symposium on Networked Systems Design and Implementation*, ser. NSDI. Boston, MA: USENIX Association, apr 2023, pp. 593–612. [Online]. Available: <https://www.usenix.org/conference/nsdi23/presentation/shah>
- [25] P. Cheng, R. Dathathri, C. Hwang, A. Jangda, S. Kalivardhan, B. Li, S. Liu, S. Maleki, M. Musuvathi, C. Rocha, O. Saarikivi, A. Shah, W. Tsui, and Z. Yang, “MSCCL++: A GPU-driven communication stack for scalable AI applications.” [Online]. Available: <https://github.com/microsoft/mscclpp>
- [26] M. Naumov, D. Mudigere, H.-J. M. Shi, J. Huang, N. Sundaraman, J. Park, X. Wang, U. Gupta, C.-J. Wu, A. G. Azzolini *et al.*, “Deep learning recommendation model for personalization and recommendation systems,” *arXiv preprint arXiv:1906.00091*, 2019.
- [27] M. Shoenybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, “Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism,” 2019.
- [28] S. Rashidi, M. Denton, S. Sridharan, S. Srinivasan, A. Suresh, J. Nie, and T. Krishna, “Enabling Compute-Communication Overlap in Distributed Deep Learning Training Platforms,” in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture*, ser. ISCA. IEEE. Piscataway, NJ, USA: IEEE Press, 2021, pp. 540–553. [Online]. Available: <https://doi.org/10.1109/ISCA52012.2021.00049>
- [29] J. T. Adriaens, K. Compton, N. S. Kim, and M. J. Schulte, “The Case for GPGPU Spatial Multitasking,” in *IEEE International Symposium on High-Performance Comp Architecture*, ser. HPCA, IEEE. Washington, DC, USA: IEEE Computer Society, 2012, pp. 1–12.
- [30] Q. Jiao, M. Lu, H. P. Huynh, and T. Mitra, “Improving GPGPU energy-efficiency through concurrent kernel execution and DVFS,” in *2015 IEEE/ACM International Symposium on Code Generation and Optimization*, ser. CGO. IEEE, 2015, pp. 1–11.
- [31] S. Pai, M. J. Thazhuthaveetil, and R. Govindarajan, “Improving GPGPU Concurrency with Elastic Kernels,” in *Proceedings of the Eighteenth International Conference on Architectural Support for Programming Languages and Operating Systems*, 2013, p. 407–418.
- [32] L. Ma, Z. Xie, Z. Yang, J. Xue, Y. Miao, W. Cui, W. Hu, F. Yang, L. Zhang, and L. Zhou, “Rammer: Enabling Holistic Deep Learning Compiler Optimizations with rTasks,” in *14th USENIX Symposium on Operating Systems Design and Implementation*, ser. OSDI. Renton, WA: USENIX Association, Nov 2020, pp. 881–897. [Online]. Available: <https://www.usenix.org/conference/osdi20/presentation/ma>
- [33] S. Pati, S. Aga, N. Jayasena, and M. D. Sinclair, “Global optimizations & lightweight dynamic logic for concurrency,” *arXiv preprint arXiv:2409.02227*, 2024.
- [34] B. Klenk, N. Jiang, G. Thorson, and L. Dennison, “An In-Network Architecture for Accelerating Shared-Memory Multiprocessor Collectives,” in *ACM/IEEE 47th Annual International Symposium on Computer Architecture*, ser. ISCA, IEEE. Washington, DC, USA: IEEE Computer Society, 2020, pp. 996–1009.
- [35] Horace He, Less Wright, Luca Wehrstedt, Tianyu Liu, Wanchao Liang, “Introducing Async Tensor Parallelism in PyTorch,” <https://discuss.pytorch.org/t/distributed-w-torch-titan-introducing-async-tensor-parallelism-in-pytorch/209487>, 2024.
- [36] K. Punniyamurthy, K. Hamidouche, and B. M. Beckmann, “Optimizing distributed ml communication with fused computation-collective operations,” *arXiv preprint arXiv:2305.06942*, 2023.
- [37] A. Jangda, J. Huang, G. Liu, A. H. N. Sabet, S. Maleki, Y. Miao, M. Musuvathi, T. Mytkowicz, and O. Saarikivi, “Breaking the Computation and Communication Abstraction Barrier in Distributed Machine Learning Workloads,” in *Proceedings of the 27th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, ser. ASPLOS. New York, NY, USA: Association for Computing Machinery, 2022, pp. 402–416. [Online]. Available: <https://doi.org/10.1145/3503222.3507778>
- [38] S. Wang, J. Wei, A. Sabne, A. Davis, B. Ilbeyi, B. Hechtman, D. Chen, K. S. Murthy, M. Maggioni, Q. Zhang, S. Kumar, T. Guo, Y. Xu, and Z. Zhou, “Overlap Communication with Dependent Computation via Decomposition in Large Deep Learning Models,” in *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 1*, ser. ASPLOS. New York, NY, USA: Association for Computing Machinery, 2022, pp. 93–106. [Online]. Available: <https://doi.org/10.1145/3567955.3567959>