

UBEP: Re-architecting Expert Parallelism Communication Library for Production Superpods

Yipeng Liu^{1,*}, Chang Liu^{1,*}, Si Shen^{1,*}, Jiaqi Zheng^{1,†}, Mingfan Li², Yuyang Yang¹, Guanhua Li¹, Yuquan Zhang¹, Yimeng Xu¹, Zhongzhe Hu^{2,†}, Zhiyuan Huang², Qihang Duan², Junsong Wang², Wenkai Ling², Baochuan Yang², Xianzhi Yu², Han Bao², Yijie Chen², Guihai Chen¹

¹State Key Laboratory for Novel Software Technology, Nanjing University ²Huawei Technologies Co., Ltd.

ABSTRACT

The deployment of Mixture-of-Experts (MoE) models on production high-bandwidth superpods, such as NVIDIA’s NVL72/576 and Huawei’s CloudMatrix384, introduces critical challenges beyond raw interconnect bandwidth. While these systems provide unified global address spaces and high-bandwidth fabrics, their full potential for sparse MoE communication is hindered by three fundamental bottlenecks: (1) Strict execution serialization imposed by coarse-grained Bulk Synchronous Parallel (BSP) orchestration of interdependent communication phases; (2) Prohibitive synchronization overhead that fails to scale alongside high interconnect bandwidth; and (3) Severe load imbalance resulting from distance-agnostic scheduling of irregular token traffic. To eliminate these bottlenecks, we introduce UBEP (Unified-Bus Expert Parallelism), a production-ready communication library that rethinks MoE’s All-to-All primitives for modern superpod architectures. Through large-scale experiments, UBEP reduces All-to-All latency by up to 52.4% and MoE inference Time Per Output Token (TPOT) by up to 11.1%.

CCS CONCEPTS

• **Networks** → **Data center networks**; • **Computing methodologies** → **Parallel computing methodologies**;

KEYWORDS

Mixture-of-Experts, Expert parallelism, All-to-All communication, Communication library, Superpod, Data center networks

ACM Reference Format:

Yipeng Liu, Chang Liu, Si Shen, Jiaqi Zheng, Mingfan Li, Yuyang Yang, Guanhua Li, Yuquan Zhang, Yimeng Xu, Zhongzhe Hu, Zhiyuan Huang, Qihang Duan, Junsong Wang, Wenkai Ling, Baochuan Yang, Xianzhi Yu, Han Bao, Yijie Chen, and Guihai Chen. 2026. UBEP: Re-architecting Expert Parallelism Communication Library for Production Superpods. In *ACM SIGCOMM 2026 Conference (SIGCOMM ’26)*, August 17–21, 2026, Denver, CO.

*Equal contribution.

†Corresponding author.

USA. ACM, New York, NY, USA, 18 pages. <https://doi.org/10.1145/3789240.3829183>

1 INTRODUCTION

The rapid evolution of Large Language Models (LLMs) has established the Mixture-of-Experts (MoE) architecture as the standard for balancing massive parameter scales with inference efficiency [7]. To support these communication-intensive workloads, datacenter infrastructure is shifting from traditional commodity clusters toward specialized superpod architectures, such as NVIDIA’s NVL72/576 [31] and Huawei’s CloudMatrix384 (CM384) [57]. Unlike traditional clusters that rely on scale-out networks such as InfiniBand (IB)/RoCE with high tail-latency and explicit message passing, superpods integrate hundreds of accelerators via scale-up interconnects such as NVLink/Unified-Bus (UB), forming a unified, high-bandwidth, load/store-accessible domain.

Despite the unprecedented raw capabilities of superpods, efficiently harnessing these capabilities for the irregular and sparse communication patterns inherent to MoE models remains a critical system challenge. Central to this challenge is the Expert Parallelism Communication Library (EPCL), the system layer responsible for orchestrating fine-grained token exchange. Unlike generic Collective Communication Libraries (CCLs) such as NCCL, EPCLs like DeepEP [51] implement the All-to-All primitive through narrow dispatch/combine APIs tailored to modern MoE systems.

As summarized in Table 1, existing EPCLs are not designed for modern multi-tier superpods. DeepEP [51] and UCCL-EP [28] adopt the Bulk Synchronous Parallel (BSP) model for non-superpod architectures with hybrid interconnects (e.g., IB and NVLink) and coarse-grained kernel-level scheduling, where long transmission time dominates and hides software overhead. CANN EP [57] runs on a superpod fabric but retains BSP with intra-only pipelines and global barriers, leaving software overhead exposed on a low-latency fabric. Hybrid-EP [32] adopts Asynchronous Parallel (ASP) model and supports single-tier superpods with finer warp-level scheduling, but it does not account for hierarchical communication constraints. In modern multi-tier superpods, drastically reduced link latency exposes the software overhead of BSP-style serialization and kernel scheduling, resulting in severe underutilization of the high-bandwidth fabric.

In this work, we identify three fundamental bottlenecks in re-architecting EPCL for modern superpods: (1) **BSP-style Serialization**. MoE communication inherently involves interdependent phases like routing and reordering; existing BSP implementations rely on strict global barriers for consistency — a design that serializes execution in low-latency superpods [51, 57]. This previously overlooked “stop-and-wait” behavior prevents overlapping



This work is licensed under a Creative Commons Attribution 4.0 International License. *SIGCOMM ’26, Denver, CO, USA*

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2467-1/26/08

<https://doi.org/10.1145/3789240.3829183>

independent communication phases and leaves high-bandwidth interconnects underutilized during synchronization. (2) **Synchronization Tax.** We reveal that as link bandwidth scales in superpods, the relative cost of synchronization primitives, such as flags, barriers, and kernel launches becomes dominant, a phenomenon we term the “synchronization tax”. Traditional control-data decoupled mechanisms introduce non-negligible overhead, which in ultra-low-latency environments limits scalability by dominating the end-to-end latency budget — an issue not previously quantified in the context of modern EPCLs. (3) **Topology-Agnostic Scheduling.** Although superpods provide a logically unified address space, we demonstrate that physical latency non-uniformity across switching tiers remains a critical yet neglected factor. Current EPCLs treat the fabric as flat and schedule workloads based solely on token counts, ignoring heterogeneous access costs in multi-tier fabrics. We show that this topology-agnostic approach leads to severe stragglers and degraded tail latency — a mismatch between logical abstraction and physical reality that has not been systematically addressed in prior EPCL designs.

To address these bottlenecks, we introduce UBEP, a production-ready EPCL designed specifically for modern superpods. Departing from the traditional BSP model, UBEP adopts a dependency-driven execution model (§3.2) that decomposes the monolithic All-to-All primitive into fine-grained tasks scheduled by data availability instead of global barriers. This enables aggressive overlap of metadata exchange, token dispatch, and reordering. To mitigate the synchronization tax, we propose Data-as-Flag (§3.4), a novel mechanism that embeds synchronization signals directly into data payloads via atomic instructions. This allows implicit, near-zero-overhead coordination at the token level, effectively eliminating control-plane overhead. Finally, to overcome the limitations of topology-agnostic scheduling, UBEP employs a hierarchical token-level scheduler (§3.3) that jointly optimizes token-to-core mapping by considering both load balance and physical fabric distance across switching tiers, thereby minimizing stragglers and tail latency.

We implemented UBEP on the Huawei CANN stack and evaluated it on up to 256 NPU dies allocated from a production CM384 superpod. The evaluation results demonstrate that UBEP reduces All-to-All latency by 52.4% compared to the baseline CANN EP, translating to an 11.1% improvement in end-to-end Time Per Output Token (TPOT) for models at the scale of DeepSeek-R1.

Contributions. Our primary contributions are as follows.

- We characterize bottlenecks in BSP-based EPCLs deployed on superpods, highlighting how synchronization overhead and serialization limit bandwidth utilization.
- We design and implement UBEP, an EPCL natively for modern multi-tier superpods, featuring three core innovations: (1) token-level kernel decomposition to maximize parallelism, (2) a hierarchical scheduler that mitigates tail latency from fabric heterogeneity, and (3) the Data-as-Flag mechanism to minimize synchronization overhead.
- We evaluate UBEP on a production-scale CM384 superpod, demonstrating significant latency reduction and end-to-end performance gains for MoE inference.

¹It supports running on heterogeneous GPUs and NICs.

²UB enables a large-scale superpod execution without inter-node communication.

Table 1: Comparison of MoE communication libraries.

Framework	Net.	Target Arch.	Pipeline	Sync.	Sched.
DeepEP[51]	IB+NV	Non-Superpod	Inter+Intra	BSP	Kernel
UCCL-EP[28]	Hete ¹	Non-Superpod	Inter+Intra	BSP	Kernel
Hybrid-EP[32]	IB+NV	1-Tier-Superpod	Inter+Intra	ASP	Warp
CANN EP[57]	UB	2-Tier (Base)	Intra-only ²	BSP	Kernel
UBEP (Ours)	UB	2-Tier (Opt)	Intra-only	ASP	Core

Ethics. This work does not raise any ethical issues.

2 BACKGROUND

2.1 Model Evolution

From Dense Model to Sparse Model. The MoE paradigm has emerged as a dominant strategy for scaling large models effectively. By activating only a small, input-specific subset of its total parameters (called *experts*), MoE architectures enable a dramatic increase in model size without a proportional increase in computational cost, thereby improving overall capacity and performance. Consequently, many state-of-the-art models have adopted this sparse approach [7, 43, 48]. This architectural shift, however, introduces a fundamental change in communication patterns. Unlike the predictable, structured communication in dense model parallelism, MoE layers generate highly irregular and dynamic inter-GPU communication. During inference or training, tokens are dynamically routed to experts via a learned gating function, with each GPU hosting only a subset of the total expert pool. This necessitates a two-step redistribution: first, token activations must be dispatched across the fabric to reach their assigned experts, and later, the processed outputs must be combined and returned to their original GPUs. The result is a sparse, runtime-dependent all-to-all exchange that forms the core communication bottleneck of MoE models [50], with such communication consuming ~47% of the total execution time on average [24, 49].

2.2 Architectural Evolution

Beyond Traditional Clusters: The Era of Modern Superpods.

Platforms like NVIDIA’s NVL72 [31] and Huawei’s CM384 [57] illustrate the progression from traditional clusters with multi-node 8-accelerator systems to modern superpods featuring 72- or 384-accelerator integrated fabrics. As illustrated in Figure 1(a), traditional clusters typically employ a hybrid interconnect strategy: NVLink for intra-node GPU-to-GPU communication, and IB/RoCE for inter-node networking. While this architecture supports cluster-scale communication through RDMA, its network bandwidth and topology are primarily optimized for data or pipeline parallelism (DP/PP), which generate relatively modest inter-node traffic [57]. In contrast, tensor parallelism (TP) and expert parallelism (EP) demand frequent, fine-grained, and low-latency communication, a requirement that is difficult to satisfy efficiently across traditional cluster nodes [57]. Consequently, many deployments are forced to confine TP/EP groups within a single compute node, constraining the scalability of the model. Modern superpods, such as those realized by NVL72 and CM384, address this bottleneck by integrating hundreds of accelerators into a single, coherent fabric. These systems exhibit three defining characteristics that are crucial for enabling efficient

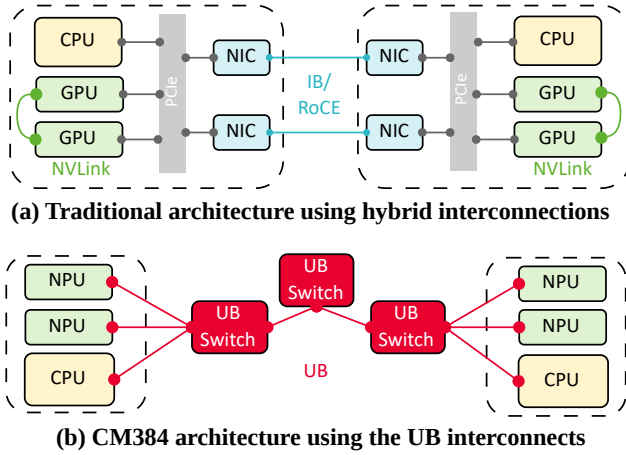


Figure 1: Comparison of traditional cluster using hybrid interconnects (a) with modern superpod utilizing the UB (b).

communication for MoE: (1) Modern superpods adopt advanced interconnect protocols, including NVLink [34], UB [14], Infinity Fabric [3], UALink [44], and SUE [4], which provide high bandwidth and low latency. For example, Huawei’s UB protocol in CM384 removes redundant network layers (Figure 1(b)) and creates a direct connection between the GPU I/O Die and the on-chip Network-on-Chip (NoC), delivering bandwidth nearing 400 GB/s with latencies in the range of hundreds of nanoseconds. Furthermore, these fabrics employ scalable topologies (e.g., Mesh or CLOS) to aggregate multiple ports, ensuring non-blocking, high-bandwidth connectivity across the entire system [25]. (2) Unified Global Address Space (UGAS): superpod provides a unified memory address space with coherent load/store semantics, where all interconnected devices are mapped into a single, globally unique address domain. This architectural feature enables direct, universal memory access, which is essential for the software synchronization required by fine-grained parallel strategies. (3) Multi-Level, Pooled Resource Management: the superpod implements multi-level resource pooling, abstracting distributed compute, memory, and network resources into a cohesive logical pool. This allows for dynamic and flexible scheduling tailored to workload demands.

2.3 Deploying MoE on CM384: Key Insights

Modern superpods offer significant architectural advantages over traditional clusters; however, the direct deployment of MoE models onto these systems often exposes new challenges.

Terminology. CM384 is built from a two-tier switching fabric: NPUs on the same baseboard connect through first-tier switches, while multiple baseboards connect through second-tier switches. We classify memory accesses by whether they stay within the local NPU or traverse the switching fabric. Since each NPU contains two compute dies, *Intra-NPU* refers to cross-die HBM access within the local NPU. We do not model same-die HBM access as a separate class, as its latency is negligible compared with accesses that traverse the switching fabric. *One-Hop* and *Two-Hop* denote NPU-to-NPU accesses that traverse one and two switch layers, respectively.

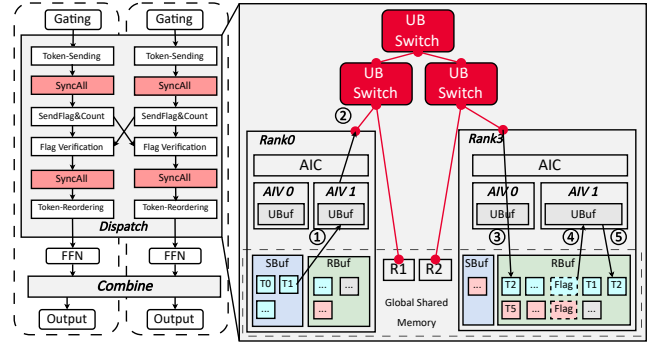
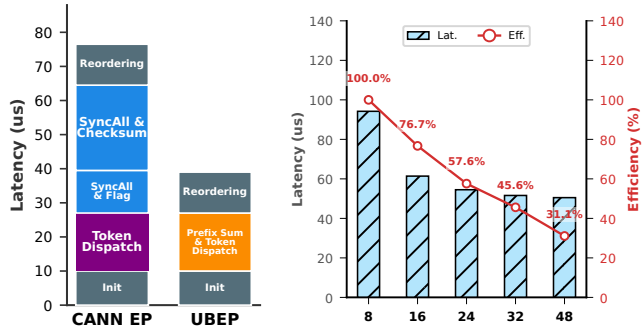


Figure 2: MoE Dispatch workflow. Left: logical view of sequential phases with global barriers (SyncAll). Right: physical view on CM384.

Orthogonal to this topology terminology, we use AIC and AIV to describe the execution hierarchy within each NPU. Each NPU contains multiple AI Cores (AICs), and each AIC follows a decoupled 1-to-N design: one Cube Unit for matrix multiplication and multiple Vector Cores (AIVs) for parallel vector processing are orchestrated by a Scalar Unit that manages instruction dispatch and control flow. **Deploying MoE on CM384.** Typically, MoE layer involves two tightly coupled communication stages: (1) a dispatch stage, which routes tokens to remote GPUs hosting the target experts, and (2) a combine stage, which gathers expert outputs and restores the original token order. Figure 2 details the typical execution flow of the dispatch operation on CM384. Prior to execution, each rank allocates a global shared memory region partitioned by source rank and expert. Leveraging the UGAS, the destination rank determines specific memory ranges based solely on per-expert token counts from each source rank. The specific execution steps for the dispatch operation are as follows: (1) Initialization (Phase 1): Tokens are stored in the local Unified Buffer (UBuf) of each AIV. (2) Transmission (Phase 2): Each AIV sends all locally owned tokens to their designated expert locations. A single token may be dispatched to multiple experts, while each expert receives tokens from multiple AIVs. (3) Global Synchronization and Completion Signaling (Phase 3): A SyncAll operation ensures all token data has been written to global memory. AIVs responsible for signaling transmit per-expert flags and token counts to indicate processing readiness. (4) Flag Verification (Phase 4): Verification is performed in parallel; each AIV polls a subset of flag entries in global memory, followed by another SyncAll to confirm global completion. (5) Offset Calculation & Reordering (Phase 5): Based on the finalized token counts, the system computes reordered token offsets and performs a contiguous reordering of tokens in global memory. This intricate process highlights how the transition from dense to sparse modeling not only changes computational patterns but also demands synchronization-heavy communication protocols to manage the resulting irregular data movement efficiently.

Our Key Insights. In our effort to optimize communication for MoE models on modern superpods, we have re-architected the EPCL. Through this reconstruction, we derived three key insights that guided our architectural decisions and implementation.



(a) Latency breakdown of distinct phases across CANN EP, and UBEP. (b) Scalability of the Token-Sending phase with varying AIV allocations.

Figure 3: Performance analysis of All-to-All communication.

Insight 1: The BSP model’s implicit synchronization incurs explicit barriers, creating a critical bottleneck for superpod. Unlocking their potential requires fine-grained task decomposition to maximize parallelism. The BSP model organizes MoE communication into a sequence of distinct phases, where each phase is guarded by a global synchronization barrier. While this design ensures correctness, it enforces strict ordering constraints and requires every AIV to perform the exact same type of work within a phase. This approach offers little flexibility in scheduling work across AIVs, resulting in a rigid execution path for the current dispatch. As shown in Figure 2, phases such as SetFlagAndCount and Token-Reordering occur in a strict sequence, and a new phase can only begin once every AIV has reached the synchronization point. Since computation is assigned statically, faster AIVs often sit idle to wait for slower ones, leading to underutilized resources.

Figure 3(a) shows that this inefficiency becomes much more visible on modern superpods. In traditional clusters, where scale-out bandwidth was limited (50 GB/s for IB/RoCE v2), the extended time required for data transfer meant that synchronization overhead was simply less pressing compared to the transfer costs. On the CM384 architecture, however, the bandwidth increases by over an order of magnitude to 392 GB/s. As a result, token dispatch reaches bandwidth saturation very quickly, leaving phase-level synchronization as the dominant constraint. We observe that only a subset of AIVs is sufficient to saturate the available bandwidth during dispatch, and adding more AIVs beyond this point provides little latency benefit. As illustrated in Figure 3(b), scaling communication cores from 24 to 48 offers minimal benefit, because the extra cores spend most of their time waiting at barriers rather than doing useful work. This suggests an opportunity to re-evaluate the BSP-style execution. Since we do not need every AIV to participate uniformly in every phase, it is possible to break down these rigid boundaries and assign sub-tasks at a finer granularity. This would relax the strict dependencies and allow for more flexible scheduling.

Insight 2: The explicit synchronization overhead on modern superpod architectures can dominate execution time. Leveraging hardware-guaranteed atomic operations enables

Table 2: CM384 memory-access latency hierarchy under distinct distance categories. *Intra-NPU* denote access HBM in local NPU; *One-Hop* and *Two-Hop* denote NPU-to-NPU accesses that traverse one and two switch layers, respectively.

Memory Access Level	Latency (ns)	Normalized
Intra-NPU	218	1.0×
One-Hop	929	4.3×
Two-Hop	2500	11.5×

implicit synchronization, replacing costly software barriers with memory-level consistency and forming the foundation for fine-grained, low-overhead parallel communication. While decomposing sub-task dependencies improves task parallelism, explicit synchronization primitives limit overall efficiency due to their overhead. On modern superpod architectures, this is exemplified by operations such as SyncAll, whose intrinsic cost can account for approximately 15% of total execution time. To address this, we leverage hardware-guaranteed atomic operations—specifically the 512B atomic Load/Store capability supported by the Ascend NPU within the superpod across scale-up and NoC domains. Crucially, the UB Fabric ensures atomicity of these writes and serializes subsequent reads, enabling *implicit synchronization*. This approach replaces costly explicit software barriers with memory-level consistency, forming the foundation for truly fine-grained, low-overhead parallel communication primitives.

Insight 3: In the hierarchical network of a modern superpod, load-balancing tokens without considering the significant latency gap between one-hop and two-hop memory accesses is counterproductive, resulting in stragglers that degrade overall MoE performance. Modern superpod architectures are rapidly evolving from scales of hundreds of accelerators (e.g., NVL72) to thousands (e.g., CM384, NVL576). To support this massive scale, the interconnect fabric inevitably expands from a single-tier switch to a multi-tier switching architecture. Consequently, while these superpods continue to provide a unified global address space, the underlying physical topology re-introduces distinct Non-Uniform Memory Access (NUMA) characteristics at the cluster level. However, existing implementations remain oblivious to this shift: they are either tailored for single-tier architectures (e.g., NVL72) [32] or focus exclusively on balancing token counts [57], failing to account for the heterogeneous access costs in multi-tier fabrics. This mismatch leads to severe stragglers in Token-Sending phase because, in such hierarchical architectures, NPU-to-NPU memory access falls into three distinct distance categories with markedly different latencies: (1) *Intra-NPU* access to cross-die HBM within the local NPU (e.g., Rank 0 writing to HBM on another die of its local NPU); (2) *One-Hop* access across a single switch layer (e.g., Rank 0 to Rank 1); and (3) *Two-Hop* access traversing two switch layers. As shown in Table 2, the latency gap is profound, with Two-Hop access reaching up to 11.5× that of local access, necessitating a scheduling approach that optimizes for both workload distribution and hierarchical memory access distances.

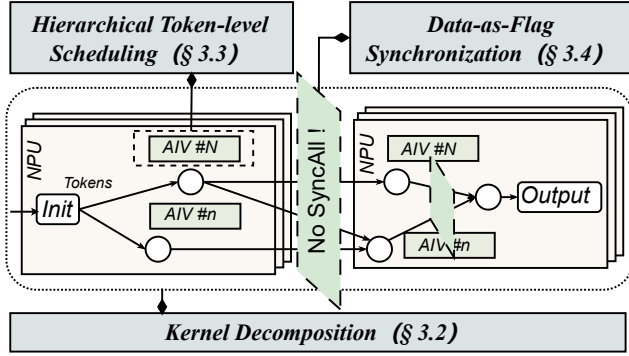


Figure 4: Overview of the UBEP architecture.

3 UBEP

3.1 Overview

UBEP (Unified-Bus Expert Parallelism) is a communication library for low-latency MoE inference. We build UBEP on three key designs (Figure 4): First, we introduce **Kernel Decomposition** (§3.2) to break the sequential constraint of traditional BSP models by partitioning AIVs into distinct groups to execute independent tasks in parallel and replace global barriers with lightweight point-to-point synchronization. Second, to handle the complex latency differences in superpods, we implement a **Hierarchical Token-level Scheduling** (§3.3). This component utilizes a hardware-accelerated mapper to generate optimal allocation schedules within 1 μ s. Finally, to eliminate expensive inter-NPU synchronization, we propose **Data-as-Flag** (§3.4). This mechanism leverages the hardware’s native 512-byte atomic load/store support. By embedding control flags directly within the data payload, UBEP achieves implicit consistency without separate control messages. These three mechanisms are synergistic: kernel decomposition exposes the fine-grained tasks that the scheduler must balance across cores and that Data-as-Flag can synchronize without global barriers; without decomposition there would be no fine-grained work to overlap, without scheduling the workload would become unbalanced across the fabric, and without Data-as-Flag global barriers would reintroduce the synchronization tax.

In principle, the proposed techniques of UBEP’s fine-grained parallelism can be adapted to any BSP-kernel, including both dispatch and combine routines. Since the challenges of global synchronization overhead and load imbalance are predominantly concentrated in the earlier dispatch phase, which fundamentally limits the overall scalability. In this context, the following sections will take dispatch as an illustrative example, without elaborating on trivial yet intricate details in combine.

3.2 Kernel Decomposition

Current MoE dispatch primitive organize communication and data reordering as BSP-style kernel executed. As illustrated in Figure 5(a), the traditional implementation simply divides the communication workflow into task phases executed in parallel across multiple AIVs. This simplifies programming and synchronization across AIVs by employing global barriers (*i.e.*, SyncAll) to satisfy

data dependencies: (1) All AIVs must wait for sending tokens (*i.e.*, Token-Sending) before sending flags for tokens verification and token counts used for address calculation needed by reordering tokens (*i.e.*, Token-Reordering); (2) All AIVs must verify all tokens before performing address calculation (*i.e.*, CalCumSum). However, as each AIV follows a sequential execution order, tasks without dependencies cannot be performed in parallel, thereby restricting the potential for parallelism. Additionally, global barriers cause faster AIVs to wait in idle state during dispatch communication, which reduces effective utilization of NPU resources.

Based on Insight 1 (§2.3), we observe that half of AIVs in each NPU are sufficient to saturate the transmission bandwidth. Therefore, we introduce kernel decomposition involving a redesign of the task decomposition for the entire communication workflow. In this approach, rather than requiring all AIVs to execute the same task sequence, we design a fine-grained partitioning mechanism that assigns independent tasks, such as token transmission and metadata process, to distinct AIVs for execution in parallel. Furthermore, we employ lightweight asynchronous signals to handle the synchronization across different AIVs. Specifically, in Figure 5(b), each NPU decomposes Token-Sending and token counts sending (*i.e.*, TokenCnt-Sending) tasks into two distinct AIV groups: most AIVs send tokens, while other AIVs calculate and send token counts. Once all TokenCnt metadata become available, the latter AIVs compute prefix sums to determine HBM offsets for per-expert tokens used in Token-Reordering. This decomposition allows us to overlap the processing latency of TokenCnt without compromising the bandwidth for token transmission.

Further, maximizing overlap and minimizing communication latency poses a constrained problem: given a fixed total number of AIVs, we must determine the optimal allocation ratio between these two groups. In general, the number of AIVs allocated to process token counts is proportional to the number of experts (calculating token counts for more destinations). Meanwhile, the number of AIVs assigned to token transmission is proportional to the product of the batch size and the Top- k (handling larger communication volume). Detailed cost modeling and the derivation of this optimal allocation appear in Appendix A. We also evaluate and verify the model’s impact in Appendix C.3.

A data dependency exists within the entire workflow: Token-Reordering relies on the addresses calculated by CalCumSum. Rather than employing global synchronization via SyncAll, we implement point-to-point synchronization based on asynchronous data signals. Specifically, the AIVs tasked with CalCumSum write the calculated addresses into global shared memory and other AIVs poll for corresponding memory offset and execute Token-Reordering immediately upon detecting the update.

3.3 Hierarchical Token-level Scheduling

As highlighted in Insight 3, the structural disparity across multiple tiers induces latency asymmetry. In practical scheduling, this asymmetry manifests primarily between one-hop access and two-hop access traversing multiple switch layers; consequently, we omit the negligible impact of intra-NPU access. Furthermore, employing naive load balancing strategies in token dispatch often exacerbates

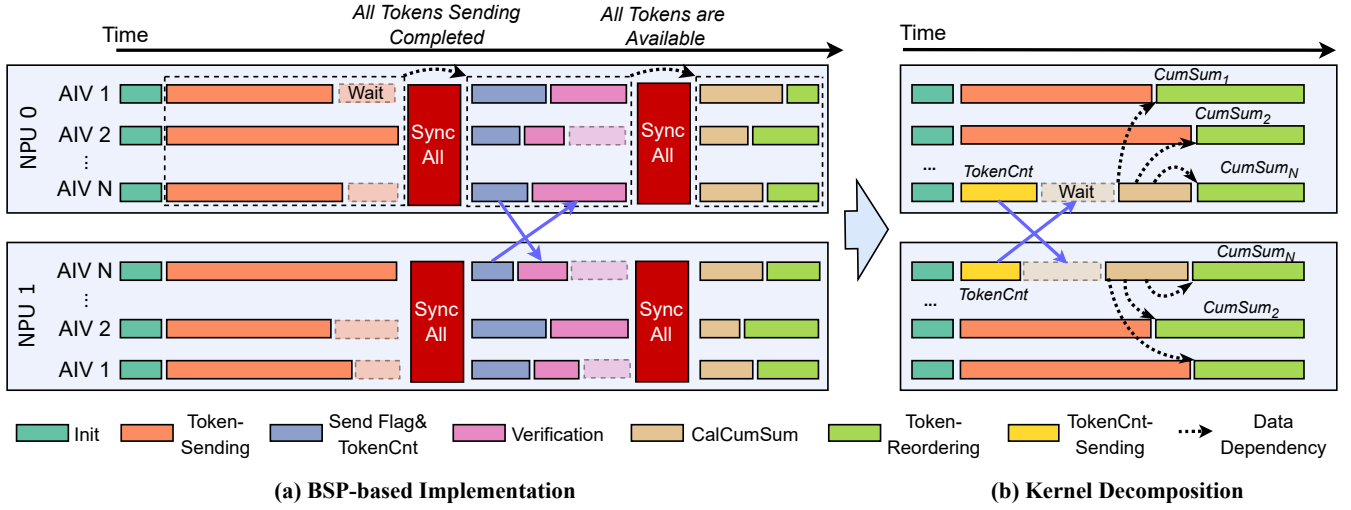


Figure 5: Comparison of BSP-based multi-phase execution and kernel decomposition for token dispatch.

the straggler problem (as detailed in §5.4). Consequently, we must rethink token scheduling by establishing a unified model that considers the impact of load balancing and hierarchy.

Formulation and Hardness Analysis. Consider a specific NPU within a superpod that outputs m tokens, denoted by the set $\mathcal{T} = \{1, \dots, m\}$, which must be dispatched to n AIVs, represented by $\mathcal{C} = \{1, \dots, n\}$. The goal is to determine an optimal assignment of tokens to AIVs such that the overall communication and processing latency is minimized, subject to satisfy load-balancing and assignment constraints. Denote by ℓ_i the transmission latency of the token i and by RTT_i the round-trip time between the token i and its destination (relevant when network delays are considered). Binary variables $\{x_{ij} \mid i \in \mathcal{T}, j \in \mathcal{C}\}$ indicate the assignment between tokens and AIVs; $x_{ij} = 1$ if and only if token i is assigned to AIV j ; and the set of tokens assigned to AIV j is defined as $\mathcal{M}_j = \{i \mid x_{ij} = 1\}$. The objective is to minimize the maximum completion time across all AIVs, which consists of the cumulative transmission latency of assigned tokens and the worst-case network delay within each AIV's token set.

$$\text{minimize} \quad \max_{j \in \{1, \dots, n\}} \left(\sum_{i \in \mathcal{M}_j} \ell_i + \max_{i \in \mathcal{M}_j} \{RTT_i\} \right) \quad (1)$$

$$\text{subject to} \quad \sum_{j=1}^n x_{ij} = 1, \quad \forall i \in \{1, \dots, m\} \quad (1a)$$

$$\sum_{i=1}^m x_{ij} \leq \left\lceil \frac{m}{n} \right\rceil, \quad \forall j \in \{1, \dots, n\} \quad (1b)$$

$$x_{ij} \in \{0, 1\}, \quad \forall i, j \quad (1c)$$

The constraint (1a) characterizes that each token is assigned to exactly one AIV. The binary decision variable x_{ij} indicates whether the token i is assigned to AIV j . Load balancing is enforced by the constraint (1b) so that no AIV receives more than a fair share of tokens. This formulation captures both the communication overhead (via ℓ_i) and the network heterogeneity (via RTT_i), while ensuring that the workload distribution remains balanced across AIVs—a key requirement for scalable and low-latency token dispatch in

MoE-based architectures. We formally prove that the program (1) is NP-hard and the detailed proof is provided in Appendix B.

Latency Homogenization. Directly solving this NP-hard optimization problem to optimality is prohibitively expensive in the context of our sub-microsecond latency budget. Even state-of-the-art integer programming solvers would require much more time to compute a solution, which would itself become the dominant bottleneck, utterly negating the performance gains sought from an optimized schedule. We propose that the objective can be heuristically solved through latency homogenization: making the latency composition of each AIV similar, thereby avoiding bottlenecks caused by some AIVs processing too many high latency tokens. To approximate the NP-hard objective, our heuristic algorithm decomposes the problem based on the two additive terms in the program (1).

First, we consider the transmission time. We note two key observations: (1) in LLMs, tokens generally have uniform sizes across the model, which are much smaller than the bandwidth. Therefore, the transmission time per token (ℓ_i) can be approximately treated as the same; (2) in modern high-bandwidth superpods, the transmission time for a single token (typically at the KB level) is much shorter than the network propagation latency. Based on these observations, we can propose that $\sum_{i \in \mathcal{M}_j} \ell_i \propto \sum_{i=1}^m x_{ij}$. Therefore, this aspect can be satisfied together with constraint (1b). What's more, the second observation justifies why our subsequent optimization can focus primarily on the network propagation term in the objective function, as it becomes the dominant source of latency variance.

Second, we consider the RTT. We categorize tokens by hop count and aim to equalize the hop-count distribution across AIVs. Let v_j be the hop distribution vectors of the AIV j and let Σ be the their covariance matrix. Minimizing the variance of each hop-count component across AIVs is equivalent to minimizing $\text{tr}(\Sigma)$. This reduces the heuristic to the simple per-hop capacity bound $v_{j,h} \leq \lceil V_h/n \rceil$, where V_h is the total number of h -hop tokens.

Hardware-Accelerated Mapper. The theoretical formulation of the token scheduling problem provides a foundation for optimizing token assignments. Practical deployment, however, requires a mechanism that operates efficiently under hardware constraints. In this

Algorithm 1 Quaternary Search

Input: Reordered Expert List E (tensor), Target Global Token Index T_{target} , Total Number of Experts N_{exp}

Output: Target Expert ID Exp_{idx} , Local Token Offset $Offset$

```

1:  $\triangleright$  Initialization
2:  $[L, R] \leftarrow [0, N_{exp}]; \quad C_L \leftarrow 0; \quad Gap \leftarrow \lceil (R - L)/4 \rceil$ 
3: while  $Gap \geq 1$  do
4:    $\triangleright$  Step 1: Define Pivots
5:    $P_i \leftarrow L + i \cdot Gap \quad \forall i \in \{1, 2, 3\}$ 
6:    $\triangleright$  Step 2: SIMD Parallel Counting
7:    $C_i \leftarrow \text{VECTORCOUNT}(E < P_i) \quad \forall i \in \{1, 2, 3\}$ 
8:    $\triangleright$  Step 3: Narrow down (Find target quadrant  $q$ )
9:   Define  $P_0 \leftarrow L, P_4 \leftarrow R, C_0 \leftarrow C_L$ 
10:  Find smallest  $q \in \{1, 2, 3\}$  such that  $T_{target} < C_q$ 
11:  if no such  $q$  exists then  $q \leftarrow 4$ 
12:  Update:  $L \leftarrow P_{q-1}; \quad R \leftarrow P_q; \quad C_L \leftarrow C_{q-1}$ 
13:   $Gap \leftarrow \lceil (R - L)/4 \rceil$ 
14:  if  $R \leq L$  or  $Gap = 0$  then
15:    break
16:  $\triangleright$  Finalization
17:  $Exp_{idx} \leftarrow L$ 
18:  $C_{final} \leftarrow \text{VECTORCOUNT}(E < Exp_{idx})$ 
19:  $Offset \leftarrow T_{target} - C_{final}$ 
20: return  $Exp_{idx}, Offset$ 

```

work, we propose a hardware-accelerated algorithm design that ensures load balancing over all tokens and maintains balance across categories. The design comprises two tightly coupled components: (1) Expert Remapping via Logical Matrix Transposition: We construct a virtual matrix where rows correspond to AIVs and columns to experts. By reading this matrix in column-major order, we generate a remapped expert sequence for each AIV. Using efficient vectorized matrix operations, we change the original token sequence that AIVs were responsible for sending, thereby ensuring that each AIV accesses a balanced mix of one-hop and two-hop experts, homogenizing the expected communication latency. (2) Token-level Load Partition: The global token sequence is partitioned into contiguous slices of equal size, and each AIV is assigned one slice. This guarantees that the maximum load difference between any two AIVs is at most one token. Since the remapped expert sequence produces non-uniform per-expert token counts, this is a prefix-sum search: given a target token index T_{target} , find the expert whose cumulative token range covers it. Each AIV performs this lookup via quaternary search (Algorithm 1), which maintains a search window $[L, R]$ over the expert space and evaluates three equidistant pivots per iteration. The cumulative token counts at all three pivots are computed in parallel via a single SIMD VectorCount instruction. The algorithm then narrows the window to the quadrant containing T_{target} . Upon convergence, the left boundary L gives the expert ID, and a final VectorCount yields the local offset within that expert's buffer. The search converges in 4–5 steps for typical expert counts and stays within 1 μs .

Generalization Beyond Two Tiers. While we evaluate CM384's two-tier topology, the scheduler readily extends to an arbitrary number of hop classes H by applying the same per-hop capacity

constraint, $\lceil V_h/n \rceil$. Provided the topology exhibits distinguishable latency between different hops, the scheduler effectively balances tokens across AIVs. On fabrics like CM384, where per-hop latency variance is large, the benefits of the scheduling strategy are most pronounced.

3.4 Data-as-Flag Synchronization

Unlike strictly ordered systems, data transfer with memory semantic within the modern superpod necessitate explicit memory barriers to enforce visibility ordering between payload and flags, causing hardware pipeline stalls. To alleviate this, we propose Data-as-Flag, a lightweight synchronization scheme that ensures correctness through verification at a smaller scale by leveraging the 512 bytes atomic memory access capability of the CM384.

As illustrated in Figure 6(a), Token-Flag Fusion (TFF) encapsulates the payload and the synchronization flag within a single 512 bytes DataBlock, where the 32 bytes leading serves as the flag field and utilizing only the remaining 480 bytes for payload. Memory semantic enables direct load/store from the local memory of a source NPU to the HBM of a destination NPU. This mechanism necessitates the movement of token data from the HBM to the local memory, where a leading flag is naturally embedded to construct a 512 bytes DataBlock. Subsequently, each DataBlock is written atomically into the HBM of the receiver. Verification of the flag alone is sufficient to ensure entire data transmission has been completed. This approach eliminates the barrier on the sender side, thereby achieving barrier-free communication. However, embedding the flag within the DataBlock introduces an additional transmission overhead, which leads to a reduction in effective bandwidth.

To further reduce the overhead introduced by embedded flags, Data-for-Checksum (DC) completely eliminate explicit flags and instead leverage the token data itself for verification. Specifically, as illustrated in Figure 6(b), the leading 32 bytes of each DataBlock is utilized as a verification marker. Upon completing the token transmission, the sender calculates an accumulated checksum of these data markers and transmits it as a separate packet. The receiver then uses this checksum to verify data integrity and confirm the completion of transmission. However, while this approach improves bandwidth utilization, it necessitates waiting for the checksum of an entire batch, which hinders fine-grained, token-level pipelining.

Sentinel Polling (SP) further mitigates the overhead of flag generation and verification by shifting both operations entirely to the receiver, as illustrated in Figure 6(c). Before communication, the receiver initializes the receive buffer with a specific initial value, while the sender transmits raw payload without any pre-processing. The receiver then determines data arrival by verifying whether the designated bytes (e.g., 32 B) within DataBlocks differ from the initial value. Although this approach eliminates the need for flag transmission and theoretically achieves 100% effective bandwidth utilization, it necessitates active resetting of the receive buffer after data consumption. More critically, a synchronization deadlock occurs if the transmitted data is identical to the initial value. While increasing the number of comparison bytes can mitigate the collision probability, it escalates the computational overhead, and this probability theoretically never reaches zero.

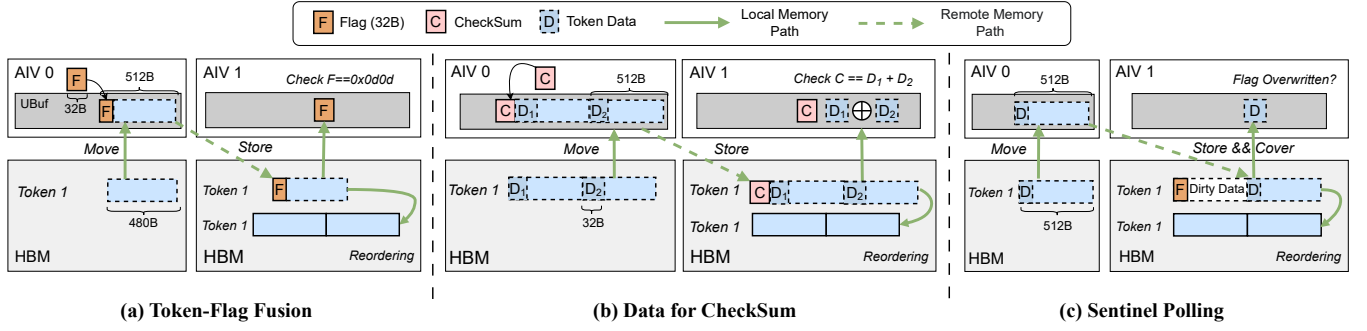


Figure 6: Comparison of three Data-as-Flag-based synchronization mechanisms across NPUs.

Correctness and safety assumptions. Data-as-Flag relies on 512B atomicity (enforced by UB Fabric): A 512B DataBlock is written atomically and observed by remote NPUs as a single transaction. Under this assumption, each Data-as-Flag variant admits a simple happens-before argument. In TFF, the sender writes (*flag, payload*) in one atomic 512B store; the receiver polls the flag and, once it changes, reads the same 512B block. Atomicity guarantees that the flag and payload are observed together, so $Write_{flag} \rightarrow Read_{payload}$. In DC, the sender writes raw payload blocks and, after the batch completes, writes a checksum; the receiver waits for the checksum and then reads the payload. The checksum write acts as a per-batch barrier: $Write_{payload} \rightarrow Write_{checksum} \rightarrow Read_{payload}$. In SP, the receiver pre-initializes the buffer with a sentinel value and the sender’s atomic write of non-sentinel data overwrites it. Because the write is atomic, the receiver never observes a partially written DataBlock, so $Write_{data} \rightarrow Read_{data}$ unless the data itself equals the sentinel.

To eliminate SP’s theoretical deadlock when payload matches the sentinel, we reserve a sentinel encoding that cannot be produced by normal computation (e.g., an unused bit pattern in BF16/FP16) and verify at model initialization that the MoE weights and activations never emit it. With a 32B (256 bits) sentinel, the probability of accidental collision under a uniform value distribution is 2^{-256} ; in practice the bound is far lower because valid activations occupy a narrow subset of the value space.

4 IMPLEMENTATION

We implemented UBEP based on the CANN software stack [15], developing an optimized EPCL in approximately 10K lines of Ascend C. The implementation builds on UB primitives to leverage the available high-bandwidth, low-latency interconnect.

SIMD Vectorized Operation. We leverage the vector instruction set of Ascend NPU throughout UBEP’s communication kernels to avoid scalar per-token loops. For the hierarchical mapper, VectorCount performs parallel prefix-sum lookups in a single instruction (Algorithm 1). For the Token-Sending path, we use CompareScalar for masking and validation, and FindFirstValue to locate token indices. This structure lets the kernel handle small batches efficiently while keeping the control flow straightforward.

Global Load/Store for Point-to-Point Synchronization. To support kernel decomposition, we replace BSP-based barriers with fine-grained point-to-point synchronization (§3.2). On conventional

Table 3: MoE Model configurations. R/S experts means routed and shared experts.

Model Name	Total	Active	Layers	Hidden	Top- <i>k</i>	Experts (R/S)	Precision
Qwen3-30B	30.5B	3.3B	48	2048	8	128/0	BF16
GLM-4.7	358B	33.6B	92	5120	8	160/1	BF16
DeepSeek-R1	671B	37B	61	7168	8	256/1	W8A8
DeepSeek-V3.2	685B	37B	61	7168	8	256/1	W8A8

clusters, this style of synchronization is often limited by coarse kernel scheduling and by DMA overhead for small messages. CM384 provides global addressing that supports cross-NPU notifications via instruction-level load/store. We use this capability both for synchronization and for data movement. During initialization, we pre-partition the receive address space, so senders can write tokens directly into the destination’s global memory. This removes local aggregation and reduces shuffling on the critical path [56].

5 EVALUATION

In this section, we first examine UBEP’s operator-level performance gains by focusing on core communication improvements (§5.2). We then demonstrate AIV-level pipelining enabled by UBEP’s kernel decomposition (§5.3). Next, we analyze the hierarchical token-level scheduling and the synchronization mechanism (§5.4) through ablation studies. Finally, we demonstrate UBEP’s end-to-end performance improvements across different models and verify the gains under various MoE model configurations (§5.5).

5.1 Setup

Testbed. We evaluate UBEP on a production-scale Huawei CM384 superpod. Our deployment consists of 16 Ascend servers allocated from this superpod, interconnected via a high-bandwidth, low-latency unified fabric, totaling 256 NPU Dies.

Models and workload. To demonstrate the versatility and robustness of UBEP across a wide spectrum of model architectures, we evaluate our system using a diverse suite of four state-of-the-art LLMs, ranging from 30B to 685B parameters. Detailed specifications for all evaluated models are summarized in Table 3. We deploy these models on the CM384 cluster, adjusting the parallelism strategies (TP, DP, EP) according to each model’s scale to ensure optimal hardware utilization.

Baseline. We evaluate UBEP against the following baselines:

Table 4: Performance of different EPCLs. DeepEP is included as a protocol-capability reference on H800.

EP	DeepEP[51] (on H800)		CANN EP[57] (on CM384)		UBEP (on CM384)	
	Latency (μ s)	BW (GB/s)	Latency (μ s)	BW (GB/s)	Latency (μ s)	BW (GB/s)
16	118	63	103	71	73	100
32	155	48	120	61	86	85
64	173	43	139	53	101	73
128	192	39	144	51	106	69
256	194	39	151	48	112	66

- CANN EP [57]: a DeepEP-like engineering adaptation to CM384, representing the conventional BSP paradigm used in existing MoE frameworks.
- UBEP w/o mapping: An ablation variant that retains basic load balancing but removes the Tier-2 network-aware scheduling, used to verify the benefits of our topology-aware design.
- Synchronization Variants: To isolate the impact of synchronization strategies, we compare the baseline Stop-and-Wait (SW) mechanism (used in CANN EP) against three variants of our Data-as-Flag design: Token-Flag Fusion (TFF), Data Checksum (DC), and Sentinel Polling (SP).

Metrics. We evaluate UBEP with two key metrics: operator latency and end-to-end latency. We first conduct a fine-grained dispatch operator latency breakdown to analyze sub-process execution time and measure total operator latency to compare with baselines. Then we use Time Per Output Token (TPOT), a critical metric for user experience in LLM serving, to show how our optimizations lead to lower token delivery latency at scale.

5.2 Performance Evaluation

EP Performance Comparison. We first evaluate the communication latency and effective bandwidth of UBEP in large-scale clusters. Table 4 details the results. UBEP improves performance by implementing kernel decomposition to increase parallelism, using Data-as-Flag mechanism to remove global synchronization barriers, and applying fine-grained AIV-level scheduling. Compared to CANN EP on identical hardware and topology, UBEP improves bandwidth by 35.3%–40.8%, isolating the algorithmic gains of our design.

Sensitivity to Workload Parameters. To analyze the performance characteristics of UBEP in more detail, we evaluate it under different batch sizes, cluster scales, and expert counts. Unless otherwise stated, experiments use 128 ranks, a batch size of 64, and 1 expert per NPU. A more comprehensive analysis across different parameters is deferred to Appendix C.1 and C.2 due to space constraints.

Figure 7(a) shows that UBEP demonstrates consistent improvements across BS ranging from 8 to 64. By replacing the baseline mechanism with fine-grained, token-level pipelining, UBEP achieves an average gain of approximately 46.4%. However, at $BS = 128$, the gain decreases to 35.9%. This decline is primarily due to the increased communication volume associated with larger

batch sizes. Although we minimize the kernel resources for the TokenCnt calculation to a single core, the computation time remains shorter than the communication latency. Consequently, the computation phase cannot fully mask the expanding communication overhead, which limits the overall speedup.

Figure 7(b) illustrates how UBEP scales with the number of ranks. In the baseline, overheads for routing and global checksums grow with the rank number, accumulating latency due to repeated global barriers. UBEP mitigates this by pipelining the preprocessing steps, maintaining an average advantage of 43.9% as the cluster scales. As the rank number increases to 256, the improvement drops to 35.1%. Two factors contribute to this. First, the communication volume increases by 13.2%. Second, to handle the more complex routing and global checksums at this scale, we must allocate 12 compute cores for these tasks. This reduces the number of cores available for communication. The reduced parallelism in communication, combined with the larger data volume, prolongs the total transmission time.

Figure 7(c) highlights the benefits of UBEP under complex routing logic. Increasing the number of experts exacerbates load imbalance. In the Baseline, straggler effects cause synchronization overheads to consume around 70% of the total runtime. UBEP restricts this overhead to approximately around 30%, yielding speedups of 38.6–41.9%. As the number of ranks increases, the performance improvement is consistent with our previous analysis. It is worth noting that UBEP exhibits slightly longer raw communication latency compared to the Baseline. This is because we divert specific AIV resources to perform CalCumSum through kernel decomposition, leaving fewer cores for data transmission. However, this decision proves highly beneficial: the minor increase in transmission time is far outweighed by the drastic reduction in synchronization costs, resulting in a net decrease in end-to-end latency.

5.3 Latency Breakdown

To evaluate the effect of Kernel Decomposition, we conducted the tests with a fixed EP scale of 64 NPUs, using a configuration of one expert per NPU. Figure 8 illustrates the detailed latency breakdown of the dispatch operator for three specific batch sizes. In CANN EP, execution proceeds through five stages: Init, Token-Sending, SendFlag & Verification, CalCumSum and Token-Reordering. Global barriers need to be inserted between the two phases where data dependencies exist, causing stragglers to increase the latency of the entire operator. UBEP removes these global barriers by combining Kernel Decomposition with Data-as-Flag synchronization, reducing the idle time of AIVs. Furthermore, fine-grained kernel decomposition enables parallel execution of Token-Sending and CalCumSum. Since the number of AIVs assigned to token transmission scales in proportion to the sending volume, the AIVs available for CalCumSum decrease as the batch size increases. The experiment results indicate that UBEP overlap the processing latency of CalCumSum with the communication latency of Token-Sending across different batch sizes, resulting in a performance improvement from 34.7% to 52.4% over CANN EP. We also evaluated optimal decomposition strategies under different EP scales, and the detailed experimental results are provided in the Appendix C.3.

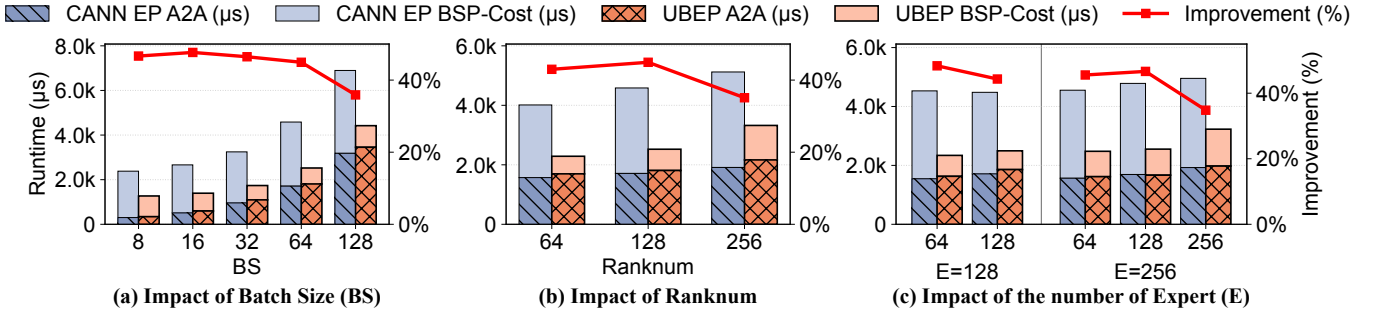


Figure 7: Latency composition of the MoE dispatch operator across representative execution configurations.

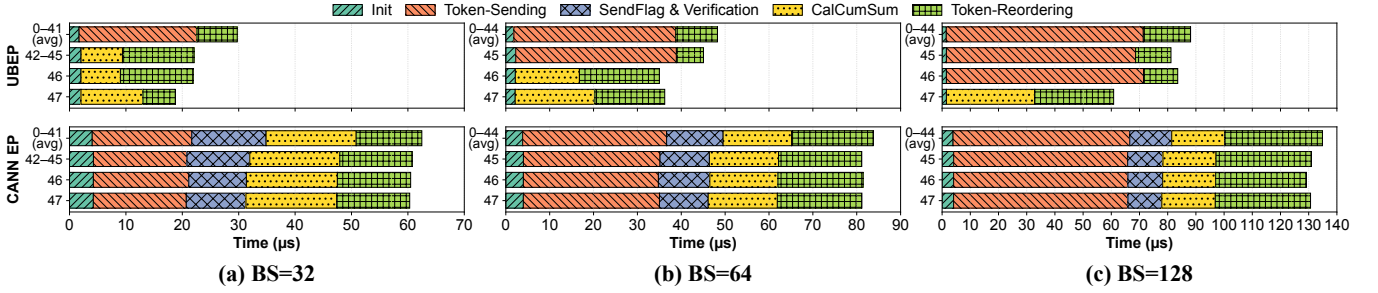


Figure 8: Task-wise latency breakdown of the dispatch operator across varying batch sizes.

5.4 Ablation Study

We conduct an ablation study to examine the contribution of individual design components in UBEP.

Impact of Token Scheduling. This study isolates the impact of token scheduling on dispatch latency, focusing on expert load imbalance and cross-Tier-2 communication. We compare UBEP with CANN EP and UBEP w/o mapping on a 64-NPU system with 256 experts and Top- $k=8$, organized as four 16-NPU nodes connected via a Tier-2 switch with eight hot experts. Figure 9 shows the token transmission latency of 28 representative AIVs selected from the 48 AIVs on a single node with 16 NPUs. CANN EP exhibits a highly skewed latency distribution due to hotspot experts that concentrate token sending on a small subset of AIVs, leading to a high maximum AIV latency of $62.2 \mu s$. UBEP w/o mapping redistributes token sending load but remains topology-agnostic, causing most transmissions to traverse the Tier-2 switch. As the cross-Tier-2 latency is higher than local HBM access (Table 2), the AIV-level latency remains uneven, with a maximum latency of $48.1 \mu s$. In contrast, UBEP jointly balances token sending load and aligns token scheduling with topology, effectively eliminating expert hotspots and latency imbalance caused by cross-switch traffic, achieving near-uniform token transmission latency across AIVs and reducing the maximum latency to $43.5 \mu s$.

Impact of Data-as-Flag. We evaluate three Data-as-Flag variants (TFF, DC, and SP) against CANN EP (SW) under different synchronization granularities and NPU scales. Figure 10 compares CANN EP with these variants. Across different granularities and rank numbers, all Data-as-Flag methods reduce latency up to 31.0%–57.1% relative to CANN EP, mainly by removing the stop-and-wait execution pattern and reducing cross-AIV synchronization overhead. On

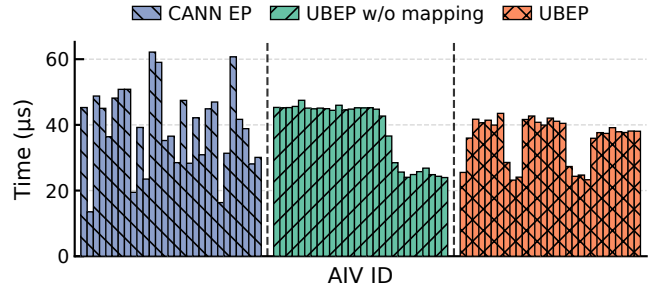


Figure 9: AIV-Level token dispatch latency under different scheduling schemes.

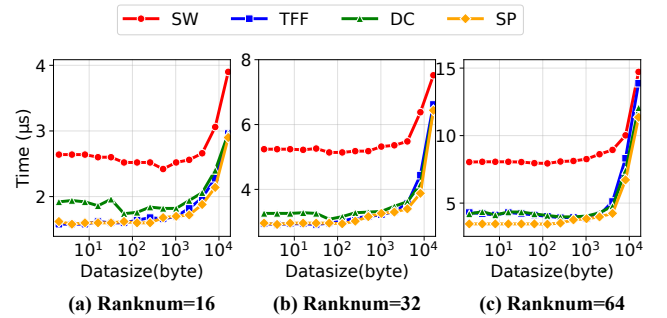


Figure 10: Dispatch latency of Data-as-Flag variants versus the CANN EP across varying synchronization granularity and ranks.

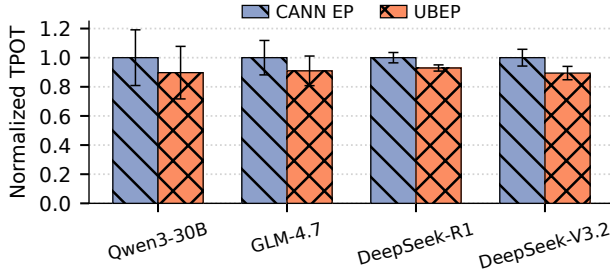


Figure 11: Normalized TPOT performance of CANN EP and UBEP

the other hand, The benefit decreases as synchronization granularity increases. At a typical dispatch granularity of around 10 KB, the gain stays at about 1 μ s. This occurs because as the synchronization granularity increases, the communication time accounts for a larger proportion of the total time, which gradually reduces the relative gain obtained from optimizing synchronization. Furthermore, as the number of ranks increases, the synchronization overhead grows. Consequently, the absolute performance gain of TFF, DC, and SP reaches 3.7–4.6 μ s when the data volume is small. Among the three variants, DC shows higher latency under small granularity or fewer NPUs because it waits for token batches, which weakens token-level pipelining. SP better balances bandwidth utilization and pipelining, and delivers lower latency across most settings, although it may encounter a rare deadlock if the sentinel value matches actual data.

5.5 End-to-End Inference Performance

In this section, based on the vLLM-Ascend v0.14.0 [45] (a hardware plugin for vLLM [21] on Ascend NPU), we evaluate the end-to-end inference performance of UBEP on Mixture-of-Experts (MoE) large language models across varying scales, including Qwen3, GLM and DeepSeek. For Qwen3-30B model, we deploy it on 16 ranks. GLM-4.7 are deployed on 128 and 160 ranks respectively, adopting one-expert-per-rank mapping. DeepSeek-R1 and DeepSeek-V3.2 are evaluated on 128 ranks with a configuration of two experts per rank.

Figure 11 reports the normalized end-to-end TPOT distribution across MoE models, with each per-token latency sample normalized to the corresponding CANN EP baseline. UBEP reduces P99 end-to-end latency by up to 11.1%, shifting the distribution downward, though the gain is smaller than the operator-level speedup because MoE All-to-All communication is only one component of the decoding step. Profiling further shows that MoE communication accounts for roughly 50% of per-token latency (consistent with prior work [49]) but only about 20% of actual hardware execution time, suggesting substantial dependency stalls and runtime overhead. The remaining time is spent on FFN computation, attention, framework scheduling, kernel launch, and memory movement. These findings motivate future work to extend the same fine-grained dependency-driven machinery to other collective communication operations and computation. Because model configurations lead to different operator-level characteristics, the normalized gains vary across models. Each box is collected from at least 100 decoding steps.

6 DISCUSSION

Portability and Hardware Assumptions. To help readers map UBEP onto other superpod fabrics, we explicitly separate the ideas that are fundamentally portable from those that rely on CM384-specific features. The core ideas do not depend on the Ascend Instruction Set Architecture (ISA): (1) token-level kernel decomposition to expose fine-grained dependencies and overlap meta-data preparation with data movement; (2) replacing global barriers with point-to-point data signals in globally-addressable memory; (3) topology-aware token scheduling that homogenizes the per-core hop-distance mix. These ideas apply whenever the fabric provides a unified global address space with remote load/store and enough fine-grained concurrency units to exploit the exposed parallelism.

Three assumptions are currently tuned to CM384, summarized in Table 5: (1) Data-as-Flag uses the 512B atomic write granularity provided by the UB fabric. On fabrics with smaller atomic units, UBEP can use smaller DataBlocks at lower payload efficiency; without atomic writes, it falls back to DC or explicit fences/acks; (2) UBEP uses AIV-level concurrency for fine-grained decomposition and overlap. On GPU-like architectures, this role maps to warp- or thread-block-level specialization, including NVIDIA-style persistent kernels. Unlike fused persistent kernels that often rely on rigid resource allocation and software-managed completion polling, UBEP uses receiver-driven data signals and instruction-level atomics to interleave communication and computation at sub-microsecond granularity; (3) the hierarchical scheduler assumes a multi-tier fabric with uniform bandwidth but hop-dependent latency. On flatter fabrics, it reduces to token-level load balancing with weaker straggler mitigation.

Attention-FFN Disaggregation (AFD). To address the resource dichotomy between memory-intensive Attention layers and compute-heavy FFN experts, recent architectures [42, 55] propose AFD to physically decouple these components onto specialized clusters. This paradigm fundamentally alters the communication topology from symmetric All-to-All to a bipartite Many-to-Many (M2N) pattern, where nodes assume distinct sender or receiver roles. Crucially, our design philosophy remains invariant under this shift. Within the superpod context, UBEP’s fine-grained, dependency-driven orchestration is topology-agnostic; it effectively overlaps the latency of M2N data transfers by decoupling synchronization from payload movement, regardless of the underlying traffic asymmetry.

Next-Generation Fabric Semantics. We demonstrated the efficiency of AIV-level parallelism in UBEP on the CM384 superpod. As inference kernels enter the microsecond regime, breaking the global synchronous barriers of BSP becomes critical. While UBEP leverages UGAS to optimize bandwidth and hierarchical access, software overheads like Data-as-Flag (§3.4) still limit fine-grained scaling. Offloading these control signals to dedicated hardware units represents a more promising mechanism—notable examples include the Scatter-Gather Engine within the SparseCore of TPUs (e.g., TPUv7 [10]) or the CC-Core in AWS Trainium [2]. Meanwhile, enforcing receiver-side ordering capabilities, like *Write-with-Notify* or *Remote-Sync-Write* semantics, can immediately trigger downstream computation pipelines. Representing the key evolutionary direction for UB, these capabilities would reduce programming

Table 5: Portability checklist for UBEP components.

Capability	Used by	Degradation if absent
AIV/warp-level concurrency	Kernel decomposition (§3.2, §3.3)	Coarser scheduling; less overlap
Uniform-bandwidth fabric with hop-dependent latency	Hierarchical scheduler (§3.3)	Falls back to load balancing; stragglers remain
512B atomic write	Data-as-Flag (§3.4)	Smaller atomic blocks reduce efficiency; no atomicity requires DC or fences/acks

complexity and 0.5 RTT overhead, unlocking greater potential for fine-grained computation-communication overlap.

7 RELATED WORK

Communication Optimization. Efficient All-to-All communication is pivotal for MoE scalability [8, 20, 30, 37]. A large body of work has optimized this primitive through specialized EPCLs or general-purpose communication runtimes. EPCLs such as DeepEP [51] and UCCL-EP [28] target traditional IB/NVLink clusters with coarse-grained, BSP-style kernels. Hybrid-EP [32] adopts asynchronous parallelism for single-tier superpods but does not address hierarchical topologies. FUSCO [56] reduces overhead by fusing layout handling, yet it does not target the superpod. More general runtimes such as NVSHMEM [33] and MSCCL++ [16] support flexible data movement but rely on explicit ordering between data movement and notification (e.g., software fences, work-queue entries, or separate flag updates). These designs—whether MoE-specific or generic—target traditional scale-up/scale-out pipelines or flat single-tier superpods; they do not exploit the unified memory semantics and hierarchical topology of modern multi-tier superpods, where traffic patterns and latency non-uniformity render existing adaptation techniques ineffective [18, 41].

Computation-Communication Overlapping. Hiding latency via concurrency is a standard optimization in MoE systems [1, 12, 13, 17, 19, 23, 26, 29, 39, 40, 49, 52, 53]. One line of work schedules decomposed All-to-All operations alongside expert computation, as in FasterMoE [12], PipeMoE [39], and Comet [49]. Another line relies on persistent or fused kernels to blur the boundary between communication and computation. FlashDMoE [1] uses persistent kernels for device-initiated asynchronous communication and fine-grained pipelining, while UniEP [53] integrates dispatch, grouped GEMM, and combine within MoE megakernels. System-level frameworks such as Tutel [17], Lancet [19], and HierMoE [26] optimize scheduling and communication, but target conventional GPU clusters with NVLink+IB interconnects rather than unified-memory superpods. These overlap strategies, whether based on data slicing, persistent kernels, or megakernels, require resolving complex data dependencies and allocating runtime resources precisely. The resulting contention and synchronization overheads can negate the intended speedups [35, 36, 46]. Unlike prior work that overlaps EP communication with expert computation, UBEP hides latency within the EP communication primitive itself by decomposing it into data-dependency-aware sub-tasks, such as overlapping address calculation with token transmission.

Load Balance. Prior works primarily address straggler effects at the *algorithm-level* [5, 6, 9, 11, 17, 22, 24, 27, 38, 47, 54]. Approaches include dynamic routing optimization [5, 6, 27], expert

replication [11, 47], and inverted routing paradigms [54], which ensure balance at the cost of model fidelity or logic modification. For instance, EPLB [6] and LPLB [5] mitigate skew by dynamically altering token assignments based on global statistics, effectively reshaping the traffic pattern at the application layer. In contrast, UBEP targets *system-level* orchestration. We accept the irregular traffic patterns generated by these upper-layer strategies and optimize the underlying dataflow and synchronization mechanisms to maximize effective bandwidth utilization on high-performance superpods without altering model semantics.

8 CONCLUSION

In this paper, we presented UBEP, a production-ready communication library re-architected for the era of superpods. UBEP dismantles the rigid BSP execution model by decomposing the monolithic All-to-All primitive into dependency-driven tasks. To eliminate explicit software barriers, we propose the Data-as-Flag synchronization protocol that leverages hardware-native atomic semantics. Furthermore, our hierarchical token-level scheduling mechanism neutralizes the load imbalance among AIVs and the straggler effects caused by hierarchical network latencies. Deployed on the Huawei’s CM384 superpod, UBEP demonstrates significant performance gains, reducing All-to-All latency by up to 52.4% and improving end-to-end MoE inference TPOT by up to 11.1%. These results validate that to fully unleash the potential of next-generation AI superpods, communication software should evolve from coarse-grained orchestration to fine-grained parallelism.

ACKNOWLEDGMENTS

We sincerely thank all the anonymous reviewers and our shepherd for their helpful comments on drafts of this paper. The work is partly supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM901) and the NSF of China (62422207).

REFERENCES

- [1] Osayamen Jonathan Aimuyo, Byungsoo Oh, and Rachee Singh. 2025. FlashDMoE: Fast Distributed MoE in a Single Kernel. (2025). arXiv:cs.DC/2506.04667 <https://arxiv.org/abs/2506.04667v2>
- [2] Amazon Web Services. 2025. Trainium2 Architecture — AWS Neuron Documentation. (2025). <https://awsdocs-neuron.readthedocs-hosted.com/en/latest/about-neuron/arch/neuron-hardware/trainium2.html> Official AWS documentation. Trainium2 includes 16 CC-Cores for collective communication. Accessed June 2026.
- [3] AMD. 2021. Introducing AMD CDNA 2 Architecture. (2021). <https://www.amd.com/content/dam/amd/en/documents/instinct-business-docs/white-papers/amd-cdna2-white-paper.pdf>
- [4] Broadcom Inc. 2024. *Scale-Up Ethernet Framework (SUE)*. Technical Report. Broadcom Inc. <https://docs.broadcom.com/doc/scale-up-ethernet-framework> White Paper.

- [5] DeepSeek-AI. 2025. An early research stage expert-parallel load balancer for MoE models based on linear programming. <https://github.com/deepseek-ai/LPLB>. (2025).
- [6] DeepSeek-AI. 2025. Expert Parallelism Load Balancer. <https://github.com/deepseek-ai/eplb>. (2025).
- [7] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojuan Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qishui Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shanyan Zhou, Shanhua Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shutong Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanbiao Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhen Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2025. DeepSeek-V3 Technical Report. (2025). arXiv:cs.CL/2412.19437 <https://arxiv.org/abs/2412.19437>
- [8] Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P Bosma, Zongwei Zhou, Tao Wang, Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. 2022. GLaM: Efficient Scaling of Language Models with Mixture-of-Experts. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.), Vol. 162. PMLR, 5547–5569. <https://proceedings.mlr.press/v162/du22c.html>
- [9] William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research* 23, 120 (2022), 1–39. <http://jmlr.org/papers/v23/21-0998.html>
- [10] Google Cloud. 2025. TPU7x (Ironwood). (2025). <https://docs.cloud.google.com/tpu/docs/tpu7x> Google Cloud official documentation. Accessed June 2026.
- [11] Yu Han, Lehan Pan, Jie Peng, Ziyang Tao, Wuyang Zhang, and Yanyong Zhang. 2025. GRACE-MoE: Grouping and Replication with Locality-Aware Routing for Efficient Distributed MoE Inference. *arXiv preprint arXiv:2509.25041* (2025).
- [12] Jiaao He, Jidong Zhai, Tiago Antunes, Haojie Wang, Fuwen Luo, Shangfeng Shi, and Qin Li. 2022. FasterMoE: modeling and optimizing training of large-scale dynamic pre-trained models. In *Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP '22)*. Association for Computing Machinery, New York, NY, USA, 120–134. <https://doi.org/10.1145/3503221.3508418>
- [13] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, and zhifeng Chen. 2019. GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/093f65e080a295f8076b1c572246aa2-Paper.pdf
- [14] Huawei. 2025. unifiedbus 2.0 Specification. (2025). <https://www.unifiedbus.com/>
- [15] Huawei Technologies Co., Ltd. 2025. CANN: Compute Architecture for Neural Networks. (2025). <https://www.hiascend.com/en/software/cann>
- [16] Changho Hwang, Peng Cheng, Roshan Dathathri, Abhinav Jangda, Saeed Maleki, Madan Musuvathi, Olli Saarikivi, Aashaka Shah, Ziyue Yang, Binyang Li, Caio Rocha, Qinghua Zhou, Mahdieh Ghazimirsaeed, Sreevatsa Anantharamu, and Jithin Jose. 2026. MSCCL++: Rethinking GPU Communication Abstractions for AI Inference. In *Proceedings of the 31st ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*. ACM, 1201–1215. <https://doi.org/10.1145/3779212.3790188>
- [17] Changho Hwang, Wei Cui, Yifan Xiong, Ziyue Yang, Ze Liu, Han Hu, Zilong Wang, Rafael Salas, Jithin Jose, Prabhat Ram, HoYuen Chau, Peng Cheng, Fan Yang, Mao Yang, and Yongqiang Xiong. 2023. Tutel: Adaptive Mixture-of-Experts at Scale. In *Proceedings of Machine Learning and Systems*, D. Song, M. Carbin, and T. Chen (Eds.), Vol. 5. Curran, 269–287. https://proceedings.mlsys.org/paper_files/paper/2023/file/5616d34cf8ff73942cfd5aa922842556-Paper-mlsys2023.pdf
- [18] Alexander Ishii and Ryan Wells. 2022. The NVLink-Network Switch: NVIDIA's Switch Chip for High Communication-Bandwidth Superpods. In *2022 IEEE Hot Chips 34 Symposium (HCS)*. <https://doi.org/10.1109/HCS55958.2022.9895480>
- [19] Chenyu Jiang, Ye Tian, Zhen Jia, Shuai Zheng, Chuan Wu, and Yida Wang. 2024. Lancet: Accelerating Mixture-of-Experts Training via Whole Graph Computation-Communication Overlapping. In *Proceedings of Machine Learning and Systems*, P. Gibbons, G. Pekhimenko, and C. De Sa (Eds.), Vol. 6. 74–86. https://proceedings.mlsys.org/paper_files/paper/2024/file/339caf45a6fa281cae8adc6465343464-Paper-Conference.pdf
- [20] Chao Jin, Ziheng Jiang, Zhihao Bai, Zheng Zhong, Juncal Liu, Xiang Li, Ningxin Zheng, Xi Wang, Cong Xie, Qi Huang, Wen Heng, Yiyuan Ma, Wenli Bao, Size Zheng, Yanghua Peng, Haibin Lin, Xuanzhe Liu, Xin Jin, and Xin Liu. 2025. MegaScale-MoE: Large-Scale Communication-Efficient Training of Mixture-of-Experts Models in Production. (2025). arXiv:cs.LG/2505.11432 <https://arxiv.org/abs/2505.11432>
- [21] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*. 611–626.
- [22] Dmitry Lepikhin, HyukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668* (2020).
- [23] Jianmin Li, Yimin Jiang, Yibo Zhu, Cong Wang, and Hong Xu. 2023. Accelerating Distributed MoE Training and Inference with Lina. In *2023 USENIX Annual Technical Conference (USENIX ATC 23)*. USENIX Association, Boston, MA, 945–959. <https://www.usenix.org/conference/atc23/presentation/li-jianmin>
- [24] Yan Li, Pengfei Zheng, Shuang Chen, Zewei Xu, Yuanhao Lai, Yunfei Du, and Zhengang Wang. 2025. Speculative MoE: Communication Efficient Parallel MoE Inference with Speculative Token and Expert Pre-scheduling. (2025). arXiv:cs.LG/2503.04398 <https://arxiv.org/abs/2503.04398>
- [25] Heng Liao, Bingyang Liu, Xianping Chen, Zhigang Guo, Chuanming Cheng, Jianbing Wang, Xiangyu Chen, Peng Dong, Rui Meng, Wenjie Liu, Zhe Zhou, Ziyang Zhang, Yuhang Gai, Cunle Qian, Yi Xiong, Zhongwu Cheng, Jing Xia, Yuli Ma, Xi Chen, Wenhua Du, Shizhong Xiao, Chungang Li, Yong Qin, Liudong Xiong, Zhou Yu, Lv Chen, Lei Chen, Buyun Wang, Pei Wu, Junen Gao, Xiaochu Li, Jian He, Shizhuan Yan, and Bill McColl. 2025. UB-Mesh: A Hierarchically Localized nD-FullMesh Data Center Network Architecture. *IEEE Micro* 45, 5 (2025), 20–29. <https://doi.org/10.1109/MM.2025.3592688>
- [26] Wenxiang Lin, Xinglin Pan, Lin Zhang, Shaohuai Shi, Xuan Wang, and Xiaowen Chu. 2025. HierMoE: Accelerating MoE Training with Hierarchical Token Deduplication and Expert Swap. (2025). arXiv:cs.DC/2508.09591 <https://arxiv.org/abs/2508.09591>
- [27] Xinyi Liu, Yujie Wang, Fangcheng Fu, Xupeng Miao, Shenhan Zhu, Xiaonan Nie, and Bin Cui. 2025. NetMoE: Accelerating MoE Training through Dynamic Sample Placement. In *The Thirteenth International Conference on Learning Representations*. https://proceedings.iclr.cc/paper_files/paper/2025/file/e0c256700465c158de71081b4cf5e8c3-Paper-Conference.pdf
- [28] Ziming Mao, Yihan Zhang, Chihan Cui, Kaichao You, Zhongjie Chen, Zhiying Xu, Scott Shenker, Costin Raiciu, Yang Zhou, and Ion Stoica. 2025. UCCL-EP: Portable Expert-Parallel Communication. (2025). arXiv:cs.DC/2512.19849 <https://arxiv.org/abs/2512.19849>
- [29] Deepak Narayanan, Aaron Harlap, Amar Phanishayee, Vivek Seshadri, Nikhil R. Devanur, Gregory R. Ganger, Phillip B. Gibbons, and Matei Zaharia. 2019. PipeDream: generalized pipeline parallelism for DNN training. In *Proceedings of the 27th ACM Symposium on Operating Systems Principles (SOSP '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. <https://doi.org/10.1145/3341301.3359646>
- [30] Xiaonan Nie, Qibin Liu, Fangcheng Fu, Shenhan Zhu, Xupeng Miao, Xiaoyang Li, Yang Zhang, Shouda Liu, and Bin Cui. 2024. LSH-MoE: Communication-efficient MoE Training via Locality-Sensitive Hashing. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 54161–54182. <https://doi.org/10.52202/079017-1716>
- [31] NVIDIA. 2024. NVIDIA GB200 NVL72. (2024). <https://www.nvidia.com/en-us/data-center/gb200-nvl72/>
- [32] NVIDIA. 2025. Hybrid-EP. <https://github.com/deepseek-ai/DeepEP/tree/hybrid-ep>. (2025).

- [33] NVIDIA. 2026. NVSHMEM Memory Ordering. <https://docs.nvidia.com/nvshmem/api/gen/api/ordering.html>. (2026).
- [34] NVIDIA Corporation. 2024. NVIDIA NVLink. (2024). <https://www.nvidia.com/en-us/products/workstations/nvlink-bridges/>
- [35] Yanghua Peng, Yibo Zhu, Yangrui Chen, Yixin Bao, Bairen Yi, Chang Lan, Chuan Wu, and Chuanxiong Guo. 2019. A generic communication scheduler for distributed DNN training acceleration. In *Proceedings of the 27th ACM Symposium on Operating Systems Principles (SOSP '19)*. Association for Computing Machinery, New York, NY, USA, 16–29. <https://doi.org/10.1145/3341301.3359642>
- [36] Kishore Punniyamurthy, Khaled Hamidouche, and Bradford M. Beckmann. 2024. Optimizing Distributed ML Communication with Fused Computation-Collective Operations. In *SC24: International Conference for High Performance Computing, Networking, Storage and Analysis*. 1–17. <https://doi.org/10.1109/SC41406.2024.00094>
- [37] Samyam Rajbhandari, Conglong Li, Zhewei Yao, Minjia Zhang, {Reza Yazdani} Aminabadi, {Ammar Ahmad} Awan, Jeff Rasley, and Yuxiong He. 2022. DeepSpeed-MoE: Advancing Mixture-of-Experts Inference and Training to Power Next-Generation AI Scale. *Proceedings of Machine Learning Research* 162 (2022), 18332–18346. Publisher Copyright: Copyright © 2022 by the author(s); 39th International Conference on Machine Learning, ICML 2022 ; Conference date: 17-07-2022 Through 23-07-2022.
- [38] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarczyk, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *International Conference on Learning Representations*. <https://arxiv.org/abs/1701.06538>
- [39] Shaohuai Shi, Xinglin Pan, Xiaowen Chu, and Bo Li. 2023. PipeMoE: Accelerating Mixture-of-Experts through Adaptive Pipelining. In *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications*. 1–10. <https://doi.org/10.1109/INFOCOM53939.2023.10228874>
- [40] Shaohuai Shi, Xinglin Pan, Qiang Wang, Chengjian Liu, Xiaozhe Ren, Zhongzhe Hu, Yu Yang, Bo Li, and Xiaowen Chu. 2024. ScheMoE: An Extensible Mixture-of-Experts Distributed Training System with Tasks Scheduling. In *Proceedings of the Nineteenth European Conference on Computer Systems (EuroSys '24)*. Association for Computing Machinery, New York, NY, USA, 236–249. <https://doi.org/10.1145/3627703.3650083>
- [41] Xiaoyong Song, Danyuan Zhou, Kai Li, Jiayuan Chen, Hao Zhang, Xiaoguang Zhang, and Xuxia Zhong. 2025. Survey of Intra-Node GPU Interconnection in Scale-Up Network: Challenges, Status, Insights, and Future Directions. *Future Internet* 17, 12 (2025). <https://doi.org/10.3390/fi17120537>
- [42] StepFun, Bin Wang, Bojun Wang, Changyi Wan, Guanzhe Huang, Hanpeng Hu, Haonan Jia, Hao Nie, Mingliang Li, Nuo Chen, Siyu Chen, Song Yuan, Wuxun Xie, Xiaoni Song, Xing Chen, Xingping Yang, Xuelin Zhang, Yanbo Yu, Yaoyu Wang, Yibo Zhu, Yimin Jiang, Yu Zhou, Yuanwei Lu, Houyi Li, Jingcheng Hu, Ka Man Lo, Ailin Huang, Binxing Jiao, Bo Li, Boyu Chen, Changxin Miao, Chang Lou, Chen Hu, Chen Xu, Chenfeng Yu, Chengyuan Yao, Daokuan Lv, Dapeng Shi, Deshan Sun, Ding Huang, Dingyuan Hu, Dongqing Pang, Enle Liu, Fajie Zhang, Fanqi Wan, Gulian Yan, Han Zhang, Han Zhou, Hanghao Wu, Hangyu Guo, Hanqi Chen, Hanshan Zhang, Hao Wu, Haocheng Zhang, Haolong Yan, Haoran Lu, Haoran Wei, Hebin Zhou, Heng Wang, Heng Wang, Hongxin Li, Hongyu Zhou, Hongyuan Wang, Huiyong Guo, Jia Wang, Jiahao Gong, Jialing Xie, Jian Zhou, Jianjian Sun, Jiaoren Wu, Jiaran Zhang, Jiayu Liu, Jie Cheng, Jie Luo, Jie Yan, Jie Yang, Jieyi Hou, Jinguang Zhang, Jinlan Cao, Jisheng Yin, Junfeng Liu, Junhao Huang, Junzhe Lin, Kaijun Tan, Kaixiang Li, Kang An, Kangheng Lin, Kenkun Liu, Lei Yang, Liang Zhao, Liangyu Chen, Lieyu Shi, Liguang Tan, Lin Lin, Lin Zhang, Lina Chen, Liwen Huang, Liying Shi, Longlong Gu, Mei Chen, Mengqiang Ren, Ming Li, Mingzhe Chen, Na Wang, Nan Wu, Qi Han, Qian Zhao, Qiang Zhang, Qianni Liu, Qiaohui Chen, Qiling Wu, Qinglin He, Qinyuan Tan, Qifeng Wang, Qiuping Wu, Qiuyan Liang, Quan Sun, Rui Li, Ruihang Miao, Ruosi Wan, Ruyan Guo, Shangwu Zhong, Shaoqiang Pang, Shengjie Fan, Shijie Shang, Shilei Jiang, Shiliang Yang, Shiming Hao, Shuli Gao, Siming Huang, Siqi Liu, Tiancheng Cao, Tianhao Cheng, Tianhao Peng, Wang You, Wei Ji, Wen Sun, Wenjin Deng, Wenqing He, Wenzhen Zheng, Xi Chen, Xiangwen Kong, Xianzhen Luo, Xiaobo Yang, Xiaojia Liu, Xiaoxiao Ren, Xin Han, Xin Li, Xin Wu, Xu Zhao, Yanan Wei, Yang Li, Yangguang Li, Yangshijie Xu, Yanming Xu, Yaqiang Shi, Yeqing Shen, Yi Yang, Yifei Yang, Yifeng Gong, Yihan Chen, Yijing Yang, Yinmin Zhang, Yizhuang Zhou, Yuanhao Ding, Yuanhao Fan, Yuanzhen Yang, Yuchao Luo, Yue Peng, Yufan Lu, Yuhang Deng, Yuhe Yin, Yujie Liu, Yukun Chen, Yuling Zhao, Yun Mou, Yunlong Li, Yunzhou Ju, Yusheng Li, Yuxiang Yang, Yuxiang Zhang, Yuyang Chen, Zejia Weng, Zhe Xie, Zheng Ge, Zheng Gong, Zhenyi Lu, Zhewei Huang, Zhichao Chang, Zhiguo Huang, Zhirui Wang, Zidong Yang, Zili Wang, Ziqi Wang, Zixin Zhang, Binxing Jiao, Daxin Jiang, Heung-Yeung Shum, and Xiangyu Zhang. 2025. Step-3 is Large yet Affordable: Model-system Co-design for Cost-effective Decoding. (2025). [arXiv:cs.LG/2507.19427](https://arxiv.org/abs/2507.19427) <https://arxiv.org/abs/2507.19427>
- [43] GLM Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, Yean Cheng, Yifan An, Yilin Niu, Yuanhao Wen, Yushi Bai, Zhengxiao Du, Zihan Wang, Zilin Zhu, Bohan Zhang, Bosi Wen, Bowen Wu, Bowen Xu, Can Huang, Casey Zhao, Changpeng Cai, Chao Yu, Chen Li, Chendi Ge, Chenghua Huang, Chenhui Zhang, Chenxi Xu, Chenzheng Zhu, Chuang Li, Congfeng Yin, Daoyan Lin, Dayong Yang, Dazhi Jiang, Ding Ai, Erle Zhu, Fei Wang, Gengzheng Pan, Guo Wang, Hailong Sun, Haitao Li, Haiyang Li, Haiyi Hu, Hanyu Zhang, Hao Peng, Hao Tai, Haoke Zhang, Haoran Wang, Haoyu Yang, He Liu, He Zhao, Hongwei Liu, Hongxi Yan, Huan Liu, Huilong Chen, Ji Li, Jiajing Zhao, Jiamin Ren, Jian Jiao, Jiani Zhao, Jianyang Yan, Jiaqi Wang, Jiayi Gui, Jiayue Zhao, Jie Liu, Jijie Li, Jing Li, Jing Lu, Jingsen Wang, Jingwei Yuan, Jingxuan Li, Jingzhao Du, Jinhua Du, Jinxin Liu, Junkai Zhi, Junli Gao, Ke Wang, Lekang Yang, Liang Xu, Lin Fan, Lindong Wu, Lintao Ding, Lu Wang, Man Zhang, Minghao Li, Minghuan Xu, Mingming Zhao, Mingshu Zhai, Pengfan Du, Qian Dong, Shangde Lei, Shangqing Tu, Shangtong Yang, Shaoyou Lu, Shijie Li, Shuang Li, Shuang-Li, Shuxun Yang, Sibo Yi, Tianshu Yu, Wei Tian, Weihang Wang, Wenbo Yu, Weng Lam Tam, Wenjie Liang, Wentao Liu, Xiao Wang, Xiaohan Jia, Xiaotao Gu, Xiaoying Lin, Xin Wang, Xing Fan, Xingru Pan, Xinyuan Zhang, Xinze Zhang, Xiuqing Fu, Xunkai Zhang, Yabo Xu, Yandong Wu, Yida Lu, Yidong Zhang, Yilin Zhou, Yiming Pan, Ying Zhang, Yingli Wang, Yingru Li, Yinpei Su, Yipeng Geng, Yitong Zhu, Yongkun Yang, Yuhang Li, Yuhao Wu, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yuxuan Zhang, Zezhen Liu, Zhen Yang, Zhengda Zhou, Zhongpei Qiao, Zhuoer Feng, Zhuorui Liu, Zichen Zhang, Zihan Wang, Zijun Yao, Zikang Wang, Ziqiang Liu, Ziwei Chai, Zixuan Li, Zuodong Zhao, Wenguan Chen, Jidong Zhai, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. 2025. GLM-4.5: Agentic, Reasoning, and Coding (ARC) Foundation Models. (2025). [arXiv:cs.CL/2508.06471](https://arxiv.org/abs/2508.06471) <https://arxiv.org/abs/2508.06471>
- [44] ULTRA ACCELERATOR LINK CONSORTIUM. 2025. *Introducing UALink 200G 1.0 Specification*. Technical Report. ULTRA ACCELERATOR LINK CONSORTIUM. https://ualinkconsortium.org/wp-content/uploads/2025/04/UALink-1.0-White_Paper_FINAL.pdf White Paper.
- [45] vllm-project. 2025. Community maintained hardware plugin for vLLM on Ascend. <https://github.com/vllm-project/vllm-ascend>. (2025).
- [46] Hulin Wang, Yaqi Xia, Donglin Yang, Xiaobo Zhou, and Dazhao Cheng. 2025. Harnessing Inter-GPU Shared Memory for Seamless MoE Communication-Computation Fusion. In *Proceedings of the 30th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming, PPoPP 2025, Las Vegas, NV, USA, March 1-5, 2025*. ACM, 170–182. <https://doi.org/10.1145/3710848.3710868>
- [47] Tian Wu, Liming Wang, Zijian Wen, Xiaoxi Zhang, Jingpu Duan, Xianwei Zhang, and Jinhao Zuo. 2025. Accelerating Edge Inference for Distributed MoE Models with Latency-Optimized Expert Placement. *arXiv preprint arXiv:2508.12851* (2025).
- [48] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zhang, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. (2025). [arXiv:cs.CL/2505.09388](https://arxiv.org/abs/2505.09388) <https://arxiv.org/abs/2505.09388>
- [49] Shulai Zhang, Ningxin Zheng, Haibin Lin, Ziheng Jiang, Wenlei Bao, Chengquan Jiang, Qi Hou, Weihao Cui, Size Zheng, Li-Wen Chang, Quan Chen, and Xin Liu. 2025. Comet: Fine-grained Computation-communication Overlapping for Mixture-of-Experts. (2025). [arXiv:cs.DC/2502.19811](https://arxiv.org/abs/2502.19811) <https://arxiv.org/abs/2502.19811>
- [50] Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Huazuo Gao, Jiashi Li, Liyue Zhang, Panpan Huang, Shangyan Zhou, Shirong Ma, Wenfeng Liang, Ying He, Yuqing Wang, Yuxuan Liu, and Y.X. Wei. 2025. Insights into DeepSeek-V3: Scaling Challenges and Reflections on Hardware for AI Architectures. In *Proceedings of the 52nd Annual International Symposium on Computer Architecture (ISCA '25)*. Association for Computing Machinery, New York, NY, USA, 1731–1745. <https://doi.org/10.1145/3695053.3731412>
- [51] Chenggang Zhao, Shangyan Zhou, Liyue Zhang, Chengqi Deng, Zhexuan Xu, Yuxuan Liu, Kuai Yu, Jiashi Li, and Liang Zhao. 2025. DeepEP: an efficient expert-parallel communication library. <https://github.com/deepseek-ai/DeepEP>. (2025).
- [52] Lianmin Zheng, Zhuohan Li, Hao Zhang, Yonghao Zhuang, Zhifeng Chen, Yanping Huang, Yida Wang, Yuanzhong Xu, Danyang Zhuo, Eric P. Xing, Joseph E. Gonzalez, and Ion Stoica. 2022. Alpa: Automating Inter- and Intra-Operator Parallelism for Distributed Deep Learning. In *16th USENIX Symposium on Operating Systems Design and Implementation (OSDI 22)*. USENIX Association, Carlsbad, CA, 559–578. <https://www.usenix.org/conference/osdi22/presentation/zheng-lianmin>
- [53] Size Zheng, Xuegui Zheng, Li wen Chang, and Jidong Zhai. 2026. UniEP: Unified Expert-Parallel MoE MegaKernel for LLM Training. (2026). [arXiv:cs.DC/2604.19241](https://arxiv.org/abs/2604.19241) <https://arxiv.org/abs/2604.19241>
- [54] Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Y. Zhao, Andrew Dai, Zhifeng Chen, Quoc Le, and James Laudon. 2022. Mixture-of-experts

- with expert choice routing. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 515, 12 pages.
- [55] Ruidong Zhu, Ziheng Jiang, Chao Jin, Peng Wu, Cesar A. Stuardo, Dongyang Wang, Xinlei Zhang, Huaping Zhou, Haoran Wei, Yang Cheng, Jianzhe Xiao, Xinyi Zhang, Lingjun Liu, Haibin Lin, Li-Wen Chang, Jianxi Ye, Xiao Yu, Xuanzhe Liu, Xin Jin, and Xin Liu. 2025. MegaScale-Infer: Serving Mixture-of-Experts at Scale with Disaggregated Expert Parallelism. (2025). arXiv:cs.DC/2504.02263 <https://arxiv.org/abs/2504.02263>
- [56] Zhuoran Zhu, Chunyang Zhu, Hao Lin, Xu Fu, Yiming Zhou, Quanlu Zhang, Zhenhua Li, Feng Qian, Chao Yu, Boxun Li, Guohao Dai, and Yu Wang. 2025. FUSCO: High-Performance Distributed Data Shuffling via Transformation-Communication Fusion. *arXiv preprint arXiv:2512.22036* (2025).
- [57] Pengfei Zuo, Huimin Lin, Junbo Deng, Nan Zou, Xingkun Yang, Yingyu Diao, Weifeng Gao, Ke Xu, Zhangyu Chen, Shirui Lu, Zhao Qiu, Peiyang Li, Xianyu Chang, Zhengzhong Yu, Fangzheng Miao, Jia Zheng, Ying Li, Yuan Feng, Bei Wang, Zaijian Zong, Mosong Zhou, Wenli Zhou, Houjiang Chen, Xingyu Liao, Yipeng Li, Wenxiao Zhang, Ping Zhu, Yinggang Wang, Chuanjie Xiao, Depeng Liang, Dong Cao, Juncheng Liu, Yongqiang Yang, Xiaolong Bai, Yi Li, Huaguo Xie, Huatao Wu, Zhibin Yu, Lv Chen, Hu Liu, Yujun Ding, Haipei Zhu, Jing Xia, Yi Xiong, Zhou Yu, and Heng Liao. 2025. Serving Large Language Models on Huawei CloudMatrix384. (2025). arXiv:cs.DC/2506.12708 <https://arxiv.org/abs/2506.12708>

APPENDICES

Appendices are supporting material that has not been peer-reviewed.

A ANALYTICAL MODELING OF DISPATCH LATENCY

We model the MoE dispatch process as two parallel groups executing on partitioned hardware resources: **token count processing** and **token dispatching**. To minimize the total latency, we aim to find the optimal resource partitioning that balances the execution time of these two groups.

Cost Models. Let n denote the total number of available AIVs. We partition these resources into n_{count} AIVs for processing token counts and n_{token} cores for dispatching token data, such that $n_{token} = n - n_{count}$. We define the time cost functions for both groups as follows:

- **Token Count Processing** (T_{count}): The latency is proportional to the number of experts (N_{exp}) and inversely proportional to the allocated compute resources (n_{count}).

$$T_{count} = G\left(\frac{N_{exp}}{n_{count}}\right) + b_{count} \quad (2)$$

- **Token Dispatching** (T_{token}): The latency is proportional to the total token load, defined by the product of the Top- k value (K_{top}) and the batch size (B_{sz}), and inversely proportional to the allocated communication resources ($n - n_{count}$).

$$T_{token} = F\left(\frac{K_{top} \times B_{sz}}{n - n_{count}}\right) + b_{token} \quad (3)$$

Here, $F(\cdot)$ and $G(\cdot)$ represent the linear scaling functions of communication and computation, respectively. b_{token} and b_{count} denote the constant hardware overheads (e.g., kernel launch latency).

Optimization Objective. Since the two AIV groups execute in parallel, the total dispatch latency $T_{dispatch}$ is determined by the slower group (the bottleneck). Our objective is to solve the Min-Max problem:

$$\min_{n_{count}} \left(\max(T_{count}, T_{token}) \right) \quad (4)$$

According to the pipeline parallelism principle, the optimal latency is achieved when the two groups are perfectly overlapped, i.e., $T_{count} = T_{token}$. Substituting the cost models, we get:

$$G\left(\frac{N_{exp}}{n_{count}}\right) + b_{count} = F\left(\frac{K_{top} \times B_{sz}}{n - n_{count}}\right) + b_{token} \quad (5)$$

Approximation and Solution. To derive an analytical relationship for optimal resource allocation, we introduce two practical assumptions based on the workload characteristics of large-scale MoE models:

- (1) **Communication Dominance:** The token dispatching group is significantly more resource-intensive than token count processing. Consequently, the majority of cores are allocated to communication ($n_{token} \gg n_{count}$), implying:

$$n - n_{count} \approx n \quad (6)$$

- (2) **Negligible Constant Overhead:** Under high-load scenarios (large batch size and Top- k), the variable processing time

dominates the constant overheads. Thus, we approximate $b_{count} \approx 0$ and $b_{token} \approx 0$.

Assuming $F(x) = \alpha x$ and $G(x) = \beta x$ are linear functions, Eq. 5 simplifies to:

$$\beta \cdot \frac{N_{exp}}{n_{count}} \approx \alpha \cdot \frac{K_{top} \times B_{sz}}{n} \quad (7)$$

Solving for n_{count} , we obtain:

$$n_{count} \approx \frac{\beta}{\alpha} \cdot n \cdot \frac{N_{exp}}{K_{top} \times B_{sz}} \quad (8)$$

The derivation leads to the following proportionality:

$$n_{count} \propto \frac{N_{exp}}{K_{top} \times B_{sz}} \quad (9)$$

This result indicates that to maintain optimal pipeline overlap, the number of cores allocated to token count processing (n_{count}) should be **directly proportional to the number of experts** (N_{exp}) and **inversely proportional to the token load** ($K_{top} \times B_{sz}$). As the communication load increases, resources must be shifted from calculation to communication to prevent the dispatch group from becoming the bottleneck.

B HARDNESS ANALYSIS

Theorem 1. Program (1) is NP-Hard.

PROOF. We analyze the hardness by reducing from the **Subset Sum Problem (SSP)**. Consider a Subset Sum Problem Instance $\mathcal{I}_2 = \{a_0, a_1, \dots, a_{k-1}\}$, where $\sum_{i=0}^{k-1} a_i = 2B$.

We construct an instance $\mathcal{I}_n$ of the original scheduling problem consisting of a set of $k \cdot n$ tokens:

$$\mathcal{I}_n = \left\{ \underbrace{(a_0, 0), \dots, (a_{k-1}, 0)}_{(n-2) \times (B, 0)\text{'s}}, \underbrace{(0, 0), \dots, (0, 0)}_{((k-1)(n-1)+1) \times (0, 0)\text{'s}} \right\} \quad (10)$$

Consider the following decision problem: Does there exist an assignment $\{x_{ij}\}$ such that:

$$\min_{\{x_{ij}\}} \max_j \left(\sum_{i \in M_j} \ell_i + \max_{i \in M_j} \{\text{RTT}_i\} \right) \leq B \quad (11)$$

If, for $\mathcal{I}_n$, there exists an n -partition $M_1, M_2, \dots, M_n$ that satisfies the requirements of the original problem: Because the total load is:

$$(n-2) \cdot B + \sum_{i=0}^{k-1} a_i = (n-2) \cdot B + 2B = n \cdot B \quad (12)$$

And due to the cardinality constraint:

$$|M_1| = |M_2| = \dots = |M_n| = k \quad (13)$$

Therefore, among $M_1, M_2, \dots, M_n$, there must be exactly $(n-2)$ partitions that take the form:

$$\left\{ (B, 0), \underbrace{(0, 0), (0, 0), \dots, (0, 0)}_{(k-1) \times (0, 0)\text{'s}} \right\} \quad (14)$$

Without loss of generality, let these be $M_3, M_4, \dots, M_n$ as shown above.

Consequently, M_1 and M_2 must strictly contain the terms from $\{(a_0, 0), \dots, (a_{k-1}, 0)\}$ and the remaining several $(0, 0)$ tokens. Let $N_1 = \{a_i \mid (a_i, 0) \in M_1\}$ and $N_2 = \{a_j \mid (a_j, 0) \in M_2\}$. We then have:

$$\sum_{a_i \in N_1} a_i = B = \sum_{a_j \in N_2} a_j, \quad N_1 \cap N_2 = \emptyset \quad (15)$$

This implies that N_1, N_2 constitute a valid solution to the requirements of the Subset Sum Problem instance $\mathcal{I}_2$.

In summary, to solve the Subset Sum Problem, one only needs to construct the corresponding original problem instance and solve it. Since the Subset Sum Problem is NP-Complete, the original decision problem is NP-Hard. $\square$

C SUPPLEMENTARY EXPERIMENT

C.1 Extended Latency Breakdown

We extend the latency composition analysis in §5.4 to additional system configurations. These results examine whether UBEP exhibits consistent behavior across different batch sizes and cluster scales.

Varying Batch Size. Figure 12(a) and 12(b) show latency breakdowns for batch sizes of 32 and 128. UBEP reduces the dominant latency components by 35%–54% over the baseline for both smaller and larger batch sizes, confirming its efficiency across batch sizes.

Varying Cluster Scale. Figure 12(c) and 12(d) report results for 64 and 256 ranks. Despite increased All-to-All communication and synchronization overhead at larger scales, UBEP reduces latency by 30%–42% compared to the baseline, demonstrating robust scalability.

C.2 Robustness to Routing Sparsity (Top- k)

To better test how UBEP’s performance changes with parameters, we have added one more experiments in addition to §5. Figure 13 presents the sensitivity analysis for the Top- k parameter. Although increasing Top- k results in a proportional multiplication of routing traffic, UBEP’s performance improvement rate remains robust, maintaining a high range of 29%–34%. This stability under throughput pressure validates the robustness of our vectorized routing kernel in handling intensive memory accesses.

C.3 Tuning Validation Resource Allocation

We sweep the CalCumSum AIV allocation group size across 64, 128, and 256 ranks to find the best setting for different combinations of rank count and batch size, as shown in Figure 14. For each grid point, we normalize the latency of the best group size to 0 and report the extra latency of the other group sizes relative to the optimum; larger values indicate a larger deviation from the optimum.

Two trends are consistent. First, the best group size decreases as batch size grows. With larger batches, dispatch spends more time on communication, so CalCumSum verification is more likely to be overlapped by communication. Adding more CalCumSum AIVs then brings less benefit and can introduce extra scheduling and contention overhead, making fewer AIVs preferable. Second, the best group size increases with rank count. We fix one expert per NPU die, so more ranks mean more experts and more validation work. Allocating more CalCumSum cores increases parallelism and reduces stragglers on the verification path.

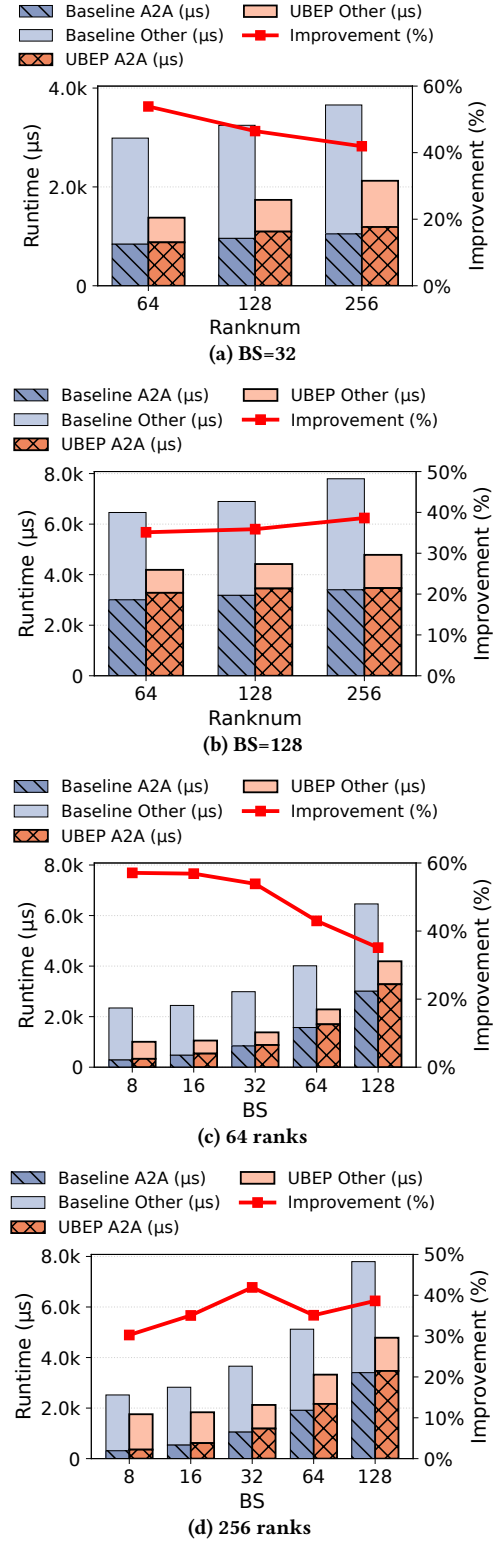


Figure 12: Supplementary latency analysis of the MoE Dispatch primitive

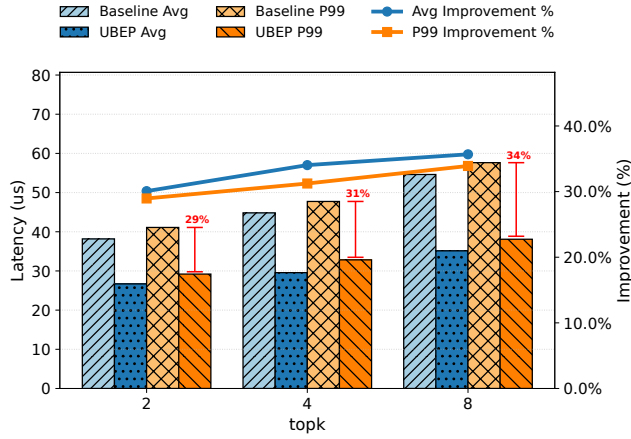
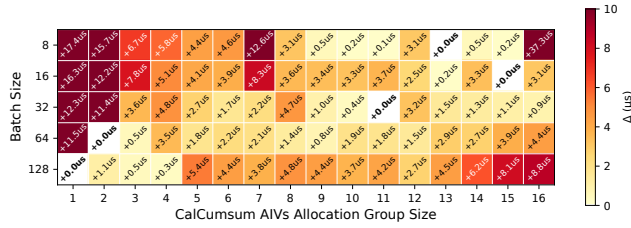
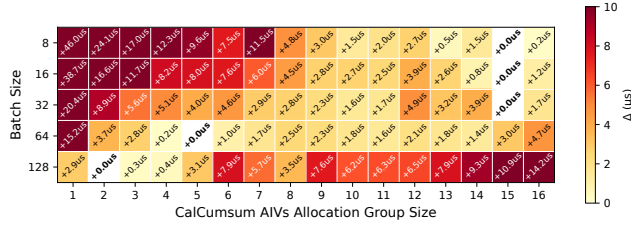


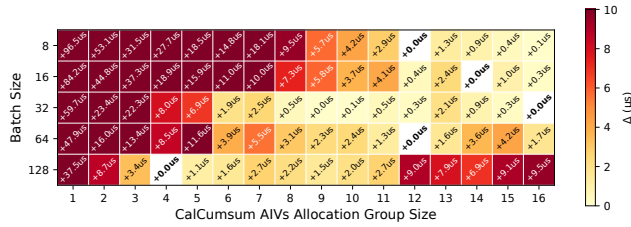
Figure 13: Performance scalability of different Top-k.



(a) 64 ranks



(b) 128 ranks



(c) 256 ranks

Figure 14: Heatmaps illustrating the relationship between the CalCumSum AIVs allocation group size, batch size, and varying rank numbers