# AT-SCALE AI TRAINING ON AMD INSTINCT™ MI350/MI300X SERIES GPUS

WHITEPAPER | 2025

## KEY TAKEAWAYS

- AMD evaluated training performance on a 2T parameter Mixture of Experts (MoE) model on a cluster of AMD Instinct™ MI300X Series GPUs to show near-linear scaling
- We present training parallelism strategies designed specifically for AMD Instinct GPU cluster topologies, aimed at reducing bottlenecks and optimizing Model FLOPs Utilization (MFU)
- The AMD Primus framework offers a unified, modular, and high-performance solution for training, and can be used to reproduce the silicon measurements and estimates from this white paper
- AMD Instinct™ MI350/MI300X Series GPUs are generally well-suited for training large foundation models

## ABSTRACT

Training foundation models efficiently require a strategy beyond brute-force hardware. This paper presents strategies for scaling AI training on AMD Instinct™ MI350X/MI300X Series GPUs, emphasizing the importance of aligning parallelism techniques with hardware topology to maximize Model FLOPs Utilization (MFU). By addressing key challenges in compute, communication, and memory, and leveraging advanced sharding and parallelization strategies, the paper details training benchmark results and extrapolations highlighting near-linear scaling to large cluster sizes on trillion-parameter models.

## THE TRAINING CHALLENGE: MANAGING COMPUTE, COMMUNICATION, AND MEMORY

Training large foundation models is a complex balancing act between three fundamental pressures. Success demands a training strategy that addresses each to achieve high **Model FLOPs Utilization (MFU)**—the fraction of peak math used at scale, while being aware of batch size constraints

- **The Communication Bottleneck:** Scaling requires repetitive collective operations —reduce-scatter/all-reduce (for Data and Tensor Parallelism) and all-to-all (for MoE). Scale-up communication is primarily used for all-to-all in our mappings. Scale -out is used for all-gather, for FSDP, and reduce-scatter for gradient synchronization.
- **The Memory Capacity Constraint:** Model weights, sharded optimizer states, and intermediate activations for backpropagation all need to fit within the available memory of each GPU.
- **The Compute Efficiency Challenge:** The goal is to keep the powerful matrix engines constantly fed and ensure we overlap memory/comms/vector ops with matrix core computations.

Solving these interconnected challenges is not simply a hardware or software problem—it is a system-level optimization task that requires a deep, hardware-software co-design approach.

## TRAINING STRATEGIES FOR AMD INSTINCT™ GPUs

A core training principle is aligning training techniques with hardware topology to minimize communication bottlenecks and maximize Model FLOPs Utilization (MFU).

### Batch Size and Replica Management

Before distributing the model, we established two core objectives to ensure both statistical and system efficiency:

**Optimal Batch Size:** Global Batch Size (GBS) selection is the prerogative of the model researchers and developers. While a large GBS can provide flexibility in the choice of sharding strategy and better compute utilization, a smaller GBS provides better generalization due to flatter loss landscapes. Additionally, a smaller GBS allows more weight updates to be performed using a given dataset. A sharding strategy must optimize data consumption along the batch dimension to scale training to a large number of GPUs at high compute utilization. The AMD sharding plan uses context parallelism for the non-MOE layers of the model, to limit batch size consumption when necessary. This is described in more detail in Section 3.2.

**Replica Size Selection:** For a given model, sequence length, GBS and world size, we explored all model and sequence parallelization strategies to build the largest possible replica without sacrificing efficiency. Multiple replicas are combined into a large training cluster with a distributed data parallelism strategy using Fully Sharded Data Parallelism (FSDP) or ZeRO.

Increasing the replica size implies a reduction in the size of the model shard in each GPU that leads to a proportional reduction in the overhead of gradient synchronization and weight propagation. However, replica size cannot be increased indefinitely. At a certain point, the replica size is capped by the inherent parallelization potential in the model (and data), or by the smallest shard size that results in efficient execution on a GPU.

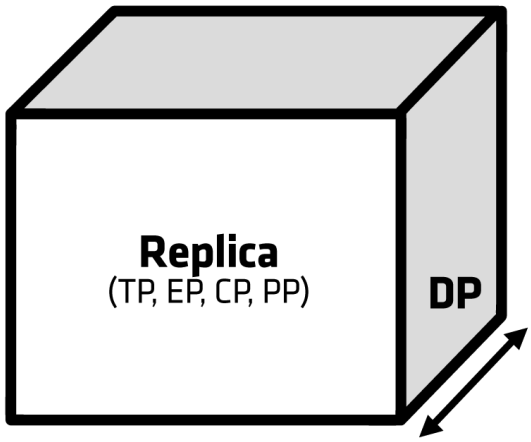


**Figure 1: Parallelization Strategy**

The above figure provides an illustration to describe the replica size optimization problem viewing the world size as a volume with a replica represented by one of the surfaces, and DP (data parallel) represented by depth. The objective of our sharding plan is to maximize the surface area for a given volume without compromising efficiency.

For the training exercise shown later in this white paper, we demonstrate strong-scaling to a training cluster size 64K GPUs, and in order to do so, we constructed a replica of 384 GPUs with a batch size per replica of nearly 0.75 M tokens to scale to a training cluster of 64K GPUs with a GBS of 128 M tokens (refer to Table 1, Section 4.1).

## AN MoE-OPTIMIZED SHARDING STRATEGY

Our sharding plan for expert-heavy MoE models involves a search for the replica size that provides the highest MFU at the target GBS and world size. Through careful sharding planning, we arrive at the following strategy which implements a hierarchical parallelism scheme that efficiently maps the training workload to the Instinct hardware hierarchy:

- At the innermost level, within the layers in each pipeline stage, we employ Expert Parallelism (EP) for the MOE layers and Context Parallelism (CP) for the attention (non-MOE) layers. If the EP option being considered is wider than the scale-up domain of 8 GPUs, the strategy employs DeepEP-like techniques to overlap all-to-all communication on both the scale-up and scale-out fabrics.
- Within each replica, we shard using Pipeline Parallelism (PP) across layers of the model. Further an optimized PP scheme involving Virtual Pipeline Parallelism (VPP) is used to create more units of work to mitigate pipeline bubbles.
- At the outermost level of the hierarchy, we employ data parallelism with either **FSDP or ZeRO**.

The sharding planning phase also considers strategies that trade-off compute for communication to achieve system objectives. By carefully balancing batchsize management, hierarchical parallelism, and targeted compute–communication tradeoffs to the system characteristics, we can efficiently train large foundational models.

# SCALING TRAINING PERFORMANCE

AMD conducted performance benchmarking of training a 2T MoE model on an AMD Instinct MI300X GPU cluster. Training was performed on 768 GPU cluster (2 replicas) with a Global Batch Size (GBS) of 128M tokens. To optimize throughput and efficiency, a hybrid parallelization strategy was employed, combining Expert Parallelism (EP), Context Parallelism (CP) and Pipeline Parallelism (PP).

Performance was extrapolated to larger cluster sizes, using measurements from the 768 GPU cluster as a baseline. By decreasing the local batch size on the 768 GPU cluster and accounting for the overheads associated with scaling, simulated performance estimates for expanded cluster size configurations were obtained. The scaling efficiency (defined as ***Realized Performance/Ideal Performance***) is shown for each configuration.

| SILICON MEASUREMENT (GBS - # OF GPUS) | EQUIVALENT SCENARIO GBS=128M (# OF GPUS) | ESTIMATED SCALING EFFICIENCY (REALIZED / IDEAL) |
|---|---|---|
| 128M-768[1] | 768 | 1.0 |
| 96M-768 | 1,024 | 1.0 |
| 48M-768 | 2,048 | 1.0 |
| 6M-768 | 16,384 | 1.0 |
| 3M-768 | 32,768 | 1.0 |
| 1.5M-768 | 65,536 | 0.9 |

**Table 1: 2T-MoE Training Scaling Efficiency Summary (2 replicas)**

The extrapolation estimates reveal minimal performance loss, attributed to the relatively low overhead of data parallel *allgather* and *reducescatter* operations, which contribute only a small fraction to overall training time.

# PROJECTING TRAINING PERFORMANCE AT SCALE ON MI355X GPUs

AMD expects training performance to scale across generations of GPUs. The latest AMD Instinct MI355X GPU offers a generational performance uplift over the AMD Instinct MI300X GPU, on key device level metrics, including HBM capacity and BF16/FP8 compute, both critical datatypes for training workloads. AMD plans to evaluate training workloads on AMD Instinct MI355X GPU clusters and expects performance uplift in-line with the improved feature specs.

| | AMD Instinct™ MI300X | AMD Instinct™ MI355X | IMPROVEMENT |
|---|---|---|---|
| **HBM Capacity (GB)** | 192 | 288 | 1.5x |
| **HBM BW (TB/s)** | 5.3 | 8.0 | 1.5x |
| **All-to-All BW (GB/s)** | 448 | 537.6 | 1.2x |
| **Peak BF16 (PF)** | 1.3 | 2.5 | 1.9x |
| **Peak FP8 (PF)** | 2.6 | 5.0 | 1.9x |

We used an internal modeling tool to estimate training performance for even larger cluster sizes on AMD Instinct MI355X GPUs, relying on a sharding planner to identify the best parallelization strategies (as described in 3.2). The results of the internal modeling tool have been correlated with silicon measurements on multi-generations of AMD Instinct GPUs across a diverse set of workloads, from small to large scale.

The modeling results on AMD Instinct MI355X GPU clusters (refer to Figure 2) show that training performance is projected to track closely with ideal theoretical speedup. From a baseline training configuration with 8192 GPUs taking 22.8 days, scaling the cluster to nearly 100K GPUs brings training time down to 2.6 days. This represents up to an 8.8x speedup, achieving up to 73% of the ideal linear scaling.[2]

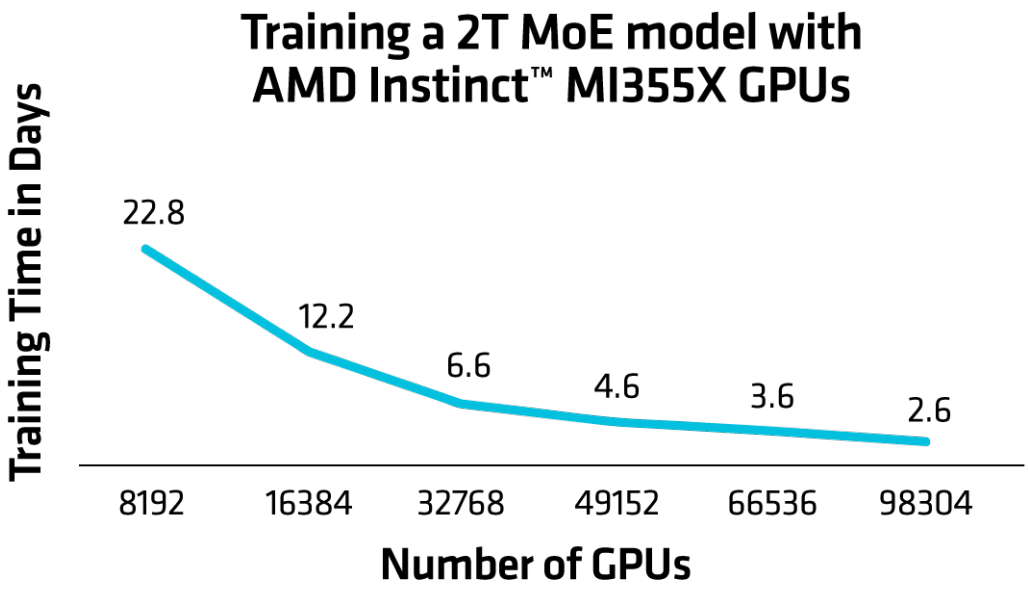


**Figure 2: Projected Scaling across AMD Instinct™ MI355X GPUs**

## GETTING STARTED WITH THE AMD PRIMUS FRAMEWORK

AMD Primus offers a unified and modular training framework which supports Megatron-LM and TorchTitan backends, making it easy to run training workloads on AMD Instinct GPUs. Primus (**github** code base) is a flexible and high-performance training framework designed for large-scale foundation model training and inference. It is designed to support pretraining, post-training, and reinforcement learning workflows, and is compatible with **Megatron** and TorchTitan backends combined with AMD ROCm™ software. The training evaluations conducted in this white paper can be reproduced using the Primus framework.

## CONCLUSION: AMD INSTINCT GPUS ARE READY FOR LARGE SCALE AI TRAINING

AMD Instinct™ MI350/MI300X Series GPUs deliver scalable and efficient solutions for training massive foundation models, including those with trillions of parameters. By leveraging advanced parallelism and optimized sharding strategies, these systems achieve near-linear scaling. AMD Instinct GPUs are well-positioned to support the next generation of large-scale AI training workloads, making them a compelling choice for organizations seeking high-performance, scalable AI infrastructure.

## AUTHORS


**Aditya Nandakumar**
Fellow Software Development Eng. AIG-SHARKS
**Aditya.Nandakumar@amd.com**

**Ashish Panday**
SMTS Software Development Eng. DCGPU MI Perf Eng
**Ashish.Panday@amd.com**

**Matt Ouellette**
Director Product Development Eng. AIG-AI Product Mgt
**matt.ouellette@amd.com**

**Ram Sivaramakrishnan**
Fellow Systems Design Eng. DCGPU AI Perf Eng
**Ram.Sivaramakrishnan@amd.com**

**Shobha Vissapragada**
MTS Silicon Design Eng. DCGPU MI Perf Eng
**Shobha.Vissapragada@amd.com**

**Wen Xie**
PMTS Software Development Eng. AIG-MODELS
**Wen.Xie@amd.com**

**Zhenyu Gu**
Sr. Director Software Development. AIG-MODELS
**Zhenyu.Gu@amd.com**


### FOOTNOTES

1. Based on AMD internal measurements as of September 23, 2025 on training a 2T parameter MoE model on 96 nodes of 8-way MI300X GPUs, Azure Cluster with 400Gbps InfiniBand. Software Configuration: rocm/megatron-lm:v25.8_py310, Primus v0.4.0, Primus-Turbo v0.1.1, Ubunutu 22.04.4, ROCm 7.0. Results may vary based on GPU memory configuration, LLM size, and any potential variances in GPU memory access or the server operating environment. MI300-092.
2. Based on AMD internal calculations as of September 23, 2025. Using AMD Instinct™ MI355X GPU based systems, the estimated total training time for a 2 trillion parameter mixture of experts (MoE) model compared to MI355X GPU based systems. Results may vary based on GPU memory configuration, LLM size, and any potential variances in GPU memory access or the server operating environment. MI350-061.

### DISCLAIMERS



### COPYRIGHT NOTICE