Poster ID: **P61954** | Topic / Category: **Accelerated Computing Tools & Techniques - Performance Optimization**

Nawras Alnaasan: alnaasan.1@osu.edu | **Dhabaleswar K (DK) Panda:** panda.2@osu.edu | **Aamir Shafi:** shafi.16@osu.edu
Hari Subramoni: subramoni.1@osu.edu

NVIDIA GTC

HiDL High-Performance Deep Learning http://hidl.cse.ohio-state.edu

MVAPICH http://mvapich.cse.ohio-state.edu

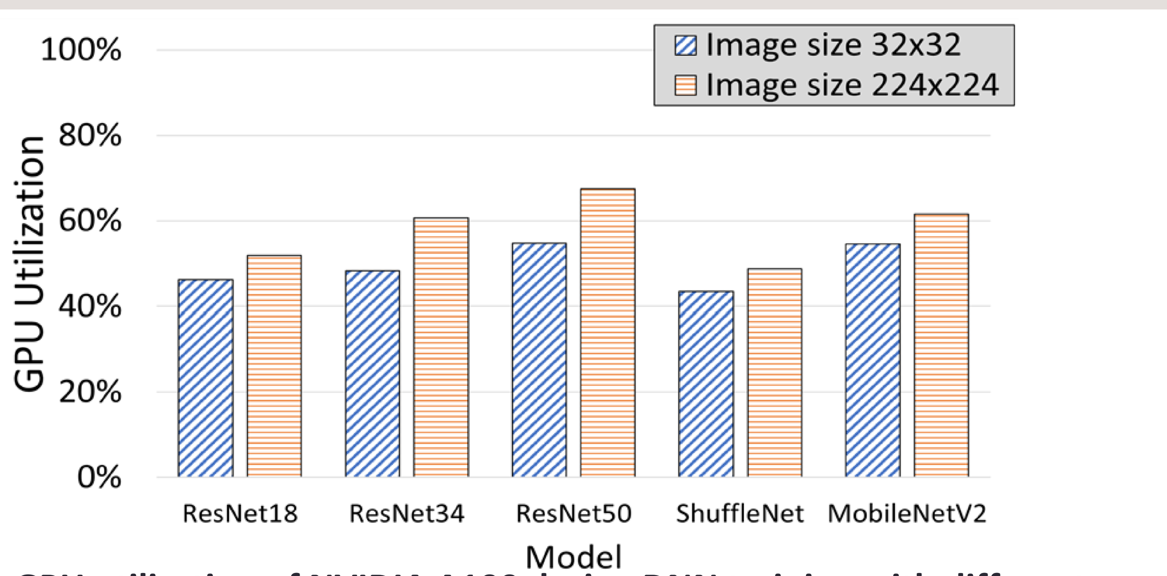
AccDP: Accelerated Data-Parallel Distributed DNN Training for Modern GPU-Based HPC Clusters

Nawras Alnaasan, Aamir Shafi, Hari Subramoni, and Dhabaleswar K Panda
{alnaasan.1, shafi.16, subramoni.1, panda.2}@osu.edu

The Ohio State University

MOTIVATION

- Modern GPUs, like NVIDIA A100, are computational workhorses used in modern HPC systems in parallel to accelerate DNN training. However, such powerful GPUs may not be fully utilized during small-to-medium DNN training and/or input sizes.
- NVIDIA Nsight Systems is used to evaluate GPU utilization.
- Under-utilization of A100 GPUs is observed across different DNNs including ResNet18, ResNet34, ResNet50, ShuffleNet, and MobileNetv2.
- Small models like ResNet18 achieves a 43% utilization, while larger models, like ResNet50, achieve 68% utilization. All evaluations are performed with the batch size that gives the best performance.

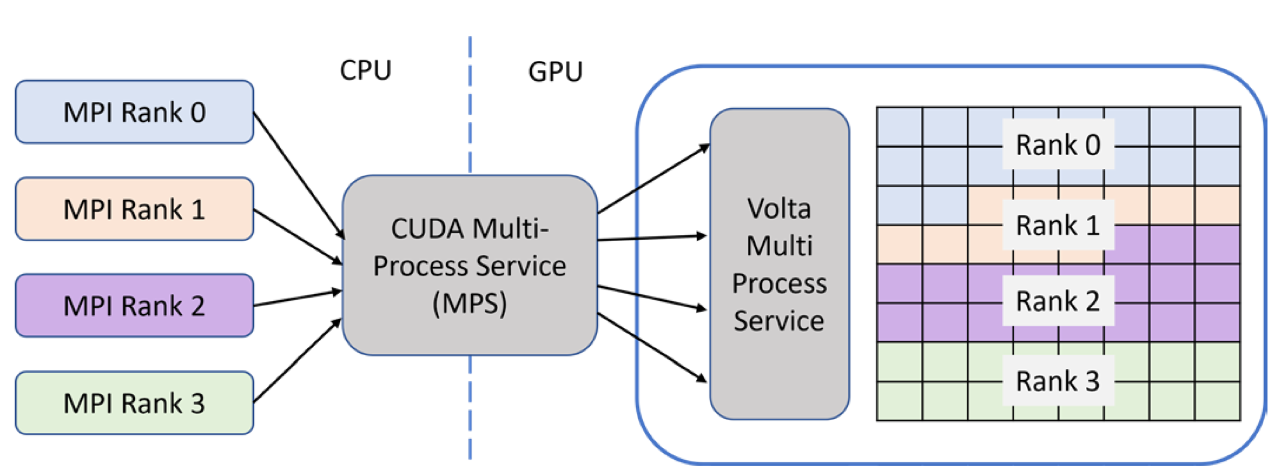


GPU utilization of NVIDIA A100 during DNN training with different models, input sizes, and best batch size.

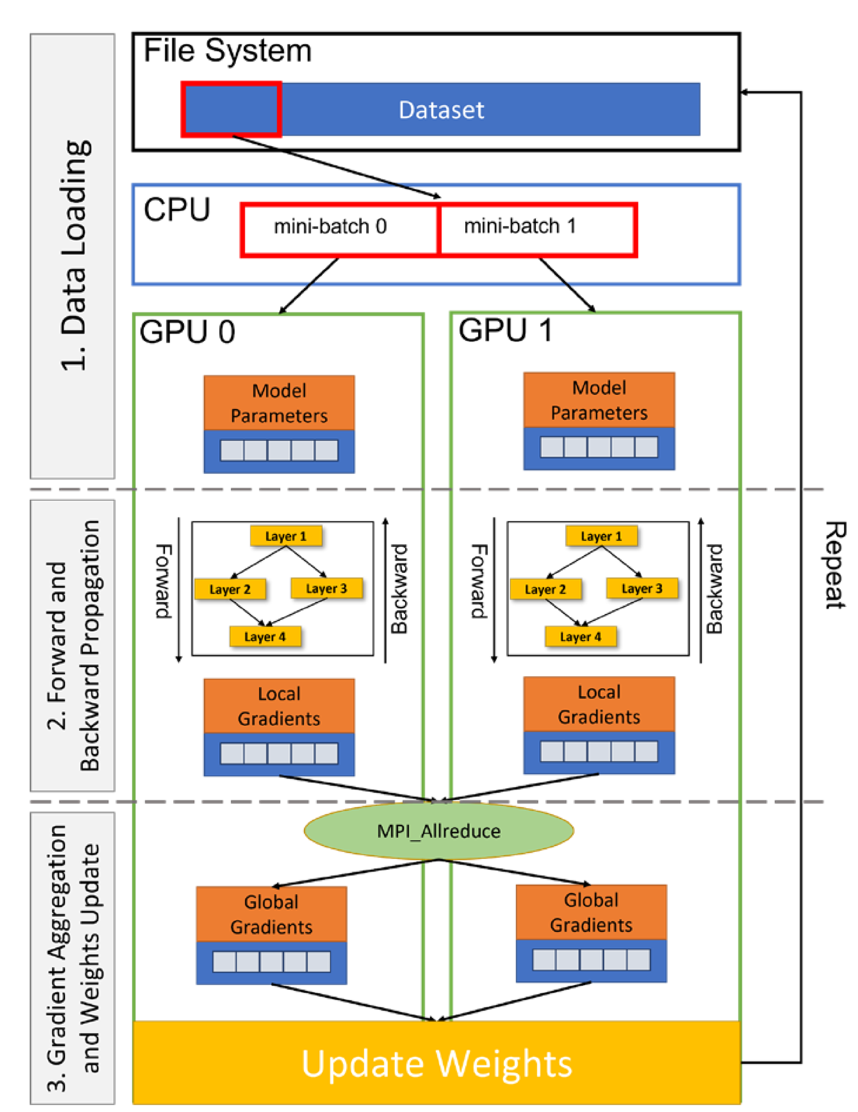
RESEARCH CHALLENGES

- What hardware/software techniques can be used to increase the utilization of processing elements like GPUs for DNN training workloads?
- How can we take advantage of architecture-specific features to increase GPU utilization for DNN training?
- How can we configure and tune DNN training hyperparameters and architecture-specific parameters, to improve hardware utilization during DNN training?
- What is the impact of the number of data loading workers on the GPU utilization and overall performance of DNN training?

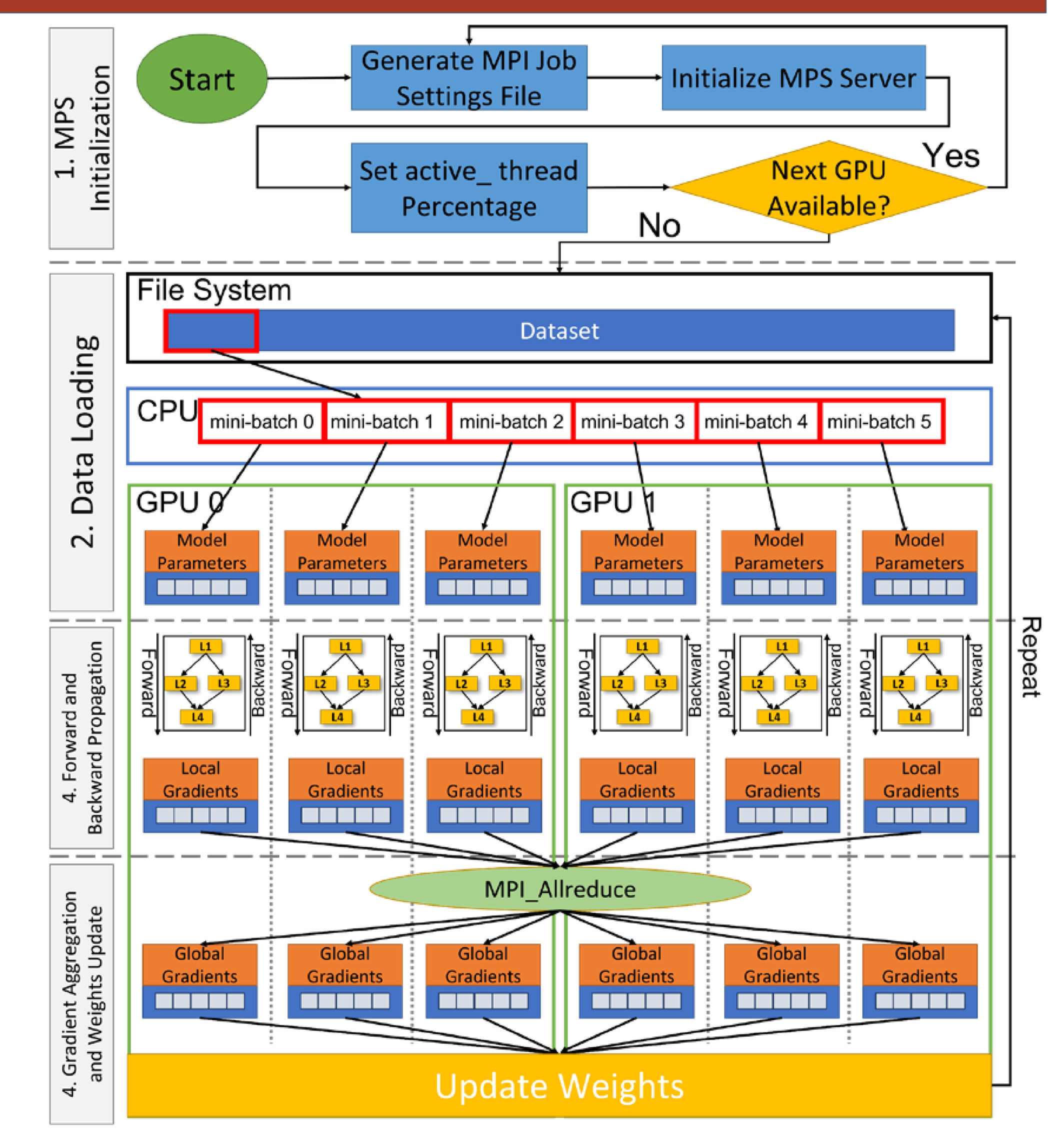
CUDA MULTI-PROCESS SERVICE (MPS)



PROPOSED DESIGN



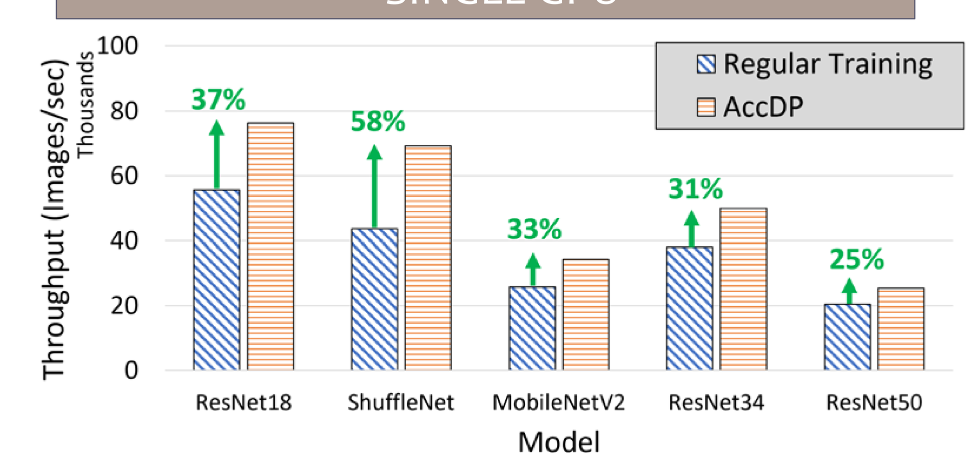
Workflow of traditional data parallelism using two GPU devices.



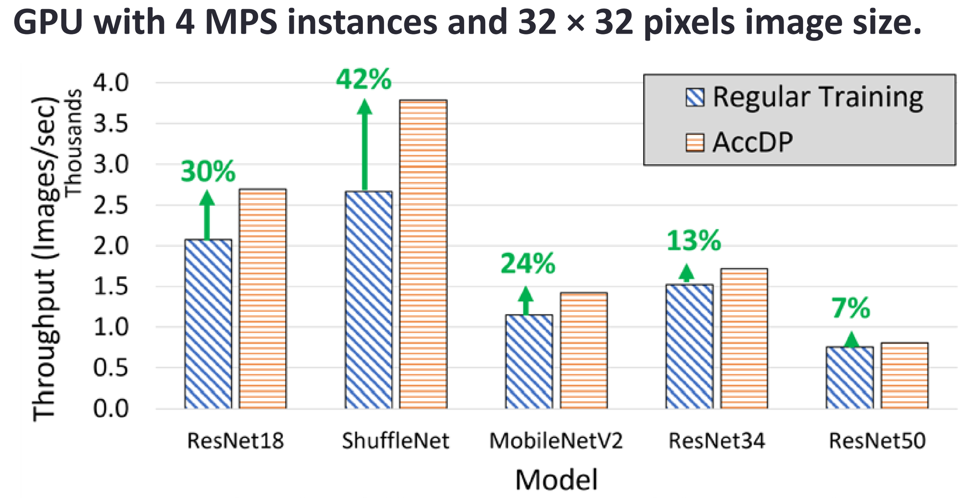
AccDP (proposed design) using data parallelism on two GPU devices and 3 MPS instances per GPU.

PERFORMANCE BENEFITS

SINGLE GPU

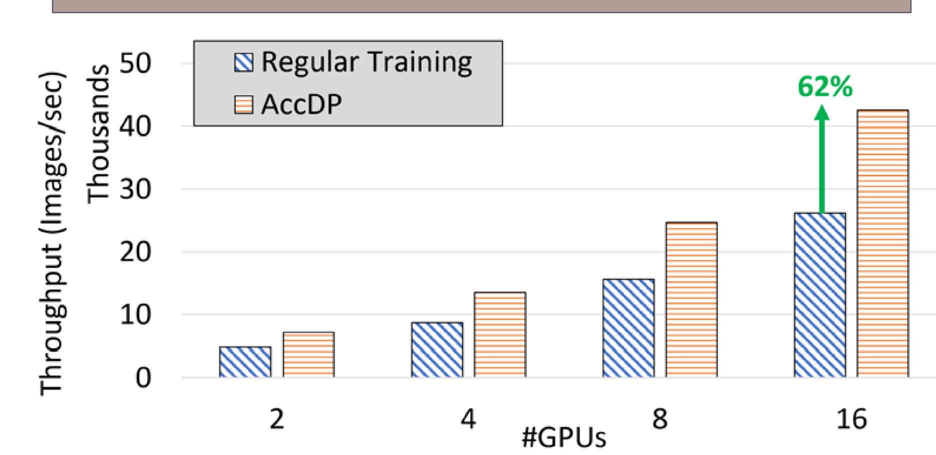


Training speedup for different DNN models on single A100 GPU with 4 MPS instances and 32 × 32 pixels image size.

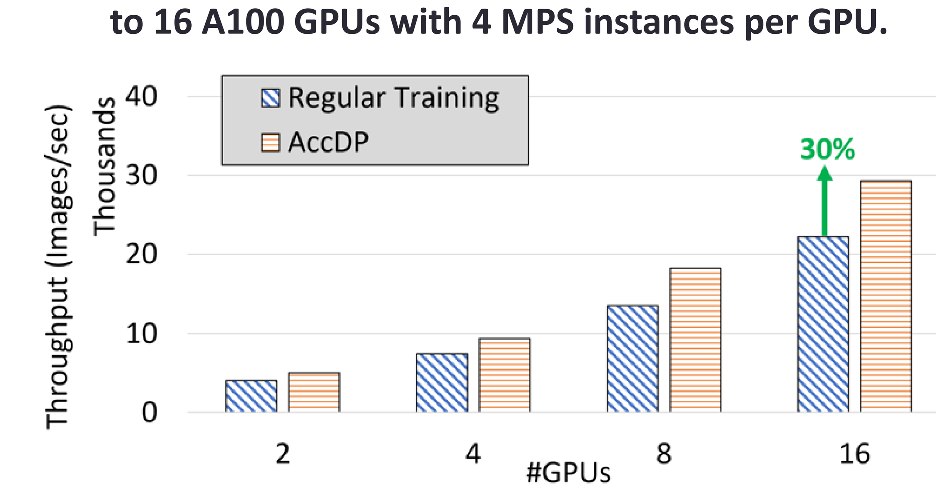


Training speedup for different DNN models on single A100 GPU with 4 MPS instances and 224 × 224 pixels image size.

MULTI GPU



Training throughput and speedup for ShuffleNet on up to 16 A100 GPUs with 4 MPS instances per GPU.



Training throughput and speedup for ResNet18 on up to 16 A100 GPUs with 4 MPS instances per GPU.

Alnaasan et al., "AccDP: Accelerated Data-Parallel Distributed DNN Training for Modern GPU-Based HPC Clusters", HiPC '22

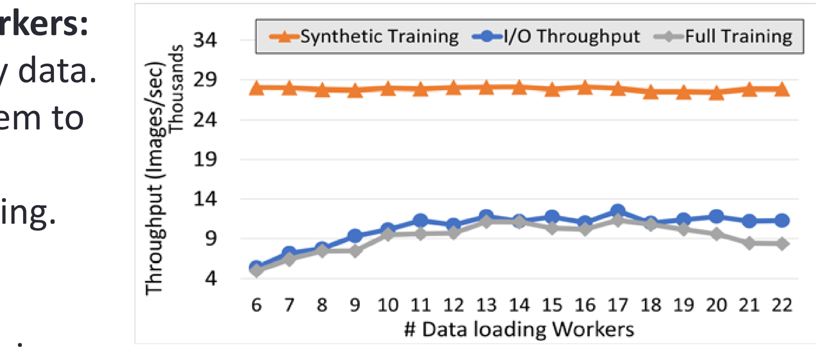
ANALYZING DATA LOADING IMPACT ON PERFORMANCE

Analysis metrics to pick the optimal number of data loading workers:

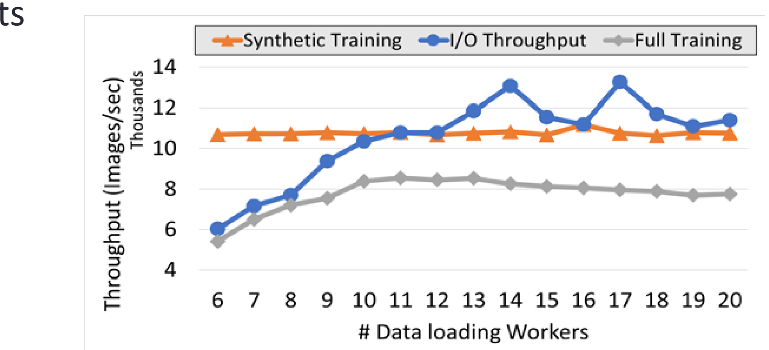
1. Synthetic training throughput: training throughput on dummy data.
2. I/O throughput: throughput of loading data from the file system to the host memory to the GPU memory.
3. Full training throughput: throughput of the regular DNN training.

Two types of data loading related bottlenecks are identified:

1. **I/O bottleneck**: Observed when the full training performance is bottlenecked by the I/O throughput. Synthetic training represents the potential capability of the if not capped by the data loading performance. ResNet18 is an example of this bottleneck.
2. **Computational bottleneck**: Observed when synthetic training throughput is lower than the I/O throughput. The optimal performance is achieved when I/O throughput matches the synthetic training throughput. ResNet50 is an example of this.



ResNet18 (I/O Bottleneck)

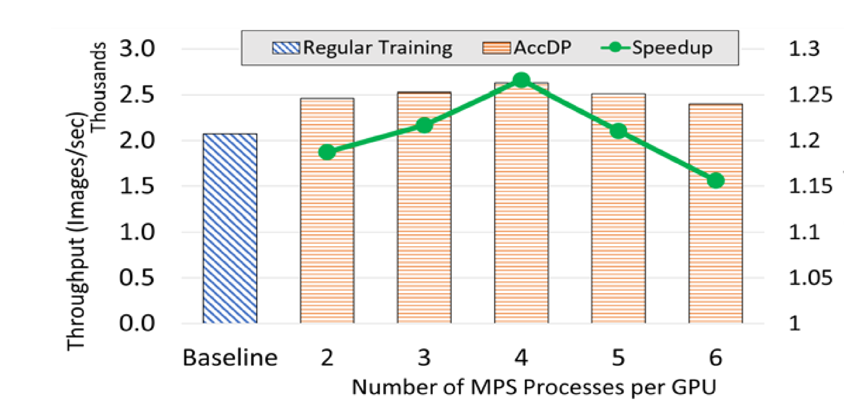


ResNet50 (Computational Bottleneck)

OPTIMIZING MPS PARAMETERS

NUMBER OF MPS INSTANCES

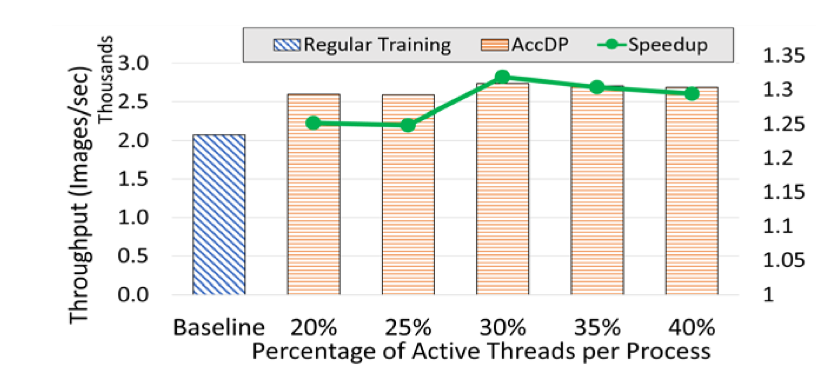
- 4 MPS processes per GPU yields the highest throughput of around 2,700 images per second.
- Increasing the number of MPS processes beyond 4 decreases the performance.
- This trend is consistent for all tested models.



Impact of number of MPS instances on DNN training throughput

THREADS % PER INSTANCE

- Best performance achieved by oversubscribing the GPU resources at 30% per instances.
- Each MPS instance has the flexibility to use additional or fewer resources with a 5% margin of the overall available GPU cores
- This trend is consistent for all tested models.



Impact of varying the percentage of active GPU threads per MPS instance (4 instances in total)

SUMMARY OF CONTRIBUTIONS

- Propose a novel data-parallelism-based training approach using MPS and MPI to improve the utilization of GPU in distributed DNN training.
- Conduct a comprehensive evaluation with different DNN models and report improvements in training throughput of up to 58% on a single A100 GPU and 62% on 16 GPUs.
- Provide in-depth analysis of the impact of different DNN training and MPS parameters.
- Identify data loading bottlenecks and impact of the number of data loading workers on DNN training performance.

FUTURE WORK

- Explore other GPU partitioning mechanisms such as Hyper-Q and Multi-GPU Instance (MIG).
- Explore solutions for other DL workloads such as hyperparameter optimization and LLM training.