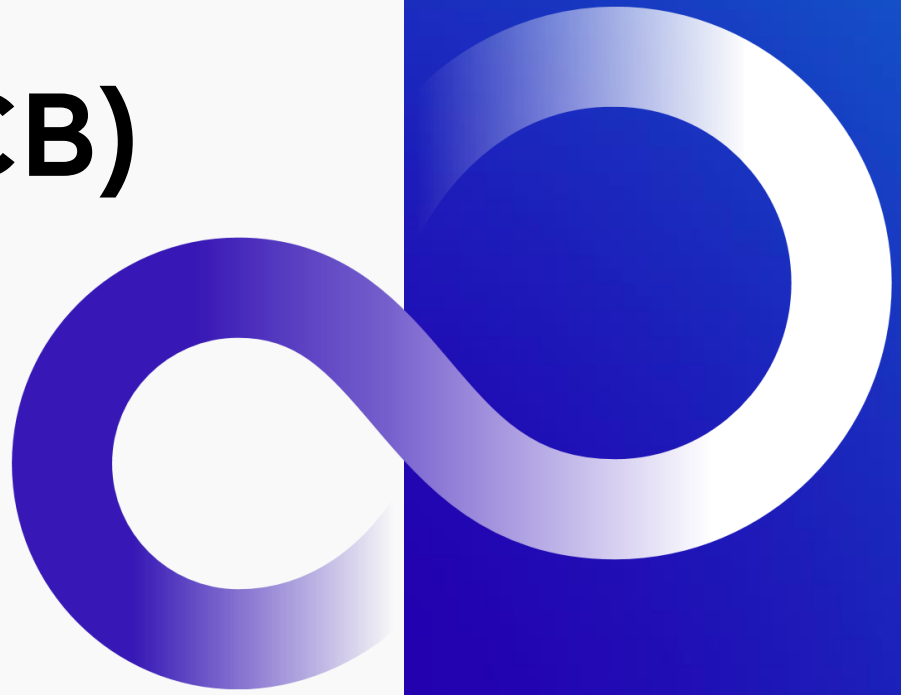
ISC 2024

FUJITSU

AI Computing Broker (ACB)



Computing Laboratory

Fujitsu Ltd.

AI Computing Broker (ACB)

FUJITSU

Significantly Reducing AI Calculation Costs and Winning the AI Development Race

AI race heating up → Global GPU shortage → Price surge & Long lead time

AI power ≒ GPU count

Technology to achieve fast AI with fewer GPUs

Watch the demo !

Technical Points of ACB

- Applicable to a wide range of AI using PyTorch or TensorFlow
- End users can utilize the ACB software by installing it
- Supports from on-premises to cloud environments

Technical Overview

FUJITSU

Realize more efficient use of GPU in AI learning process

Challenges in using GPU

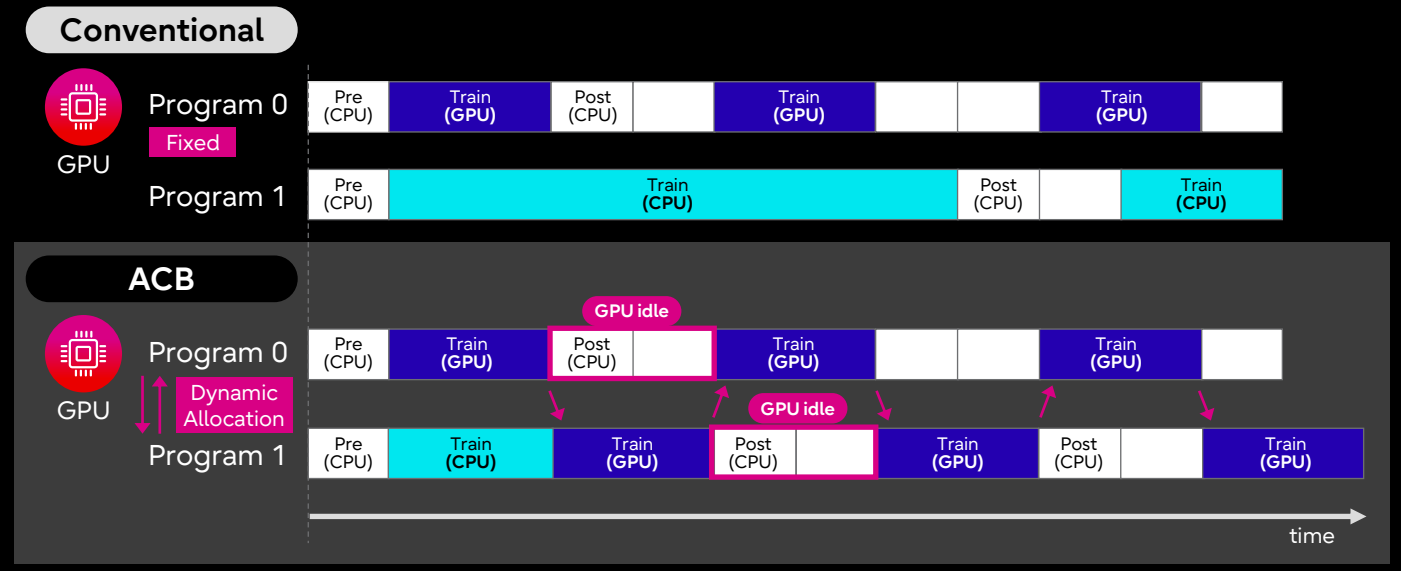
- Since the GPU is allocated on a per-program basis, even if the GPU idle time is large, other programs cannot use the GPU until the program ends.
- When virtualizing the GPU to share it among multiple programs, the GPU memory is divided for each program, reducing the capacity per program.

Features of AI Computing Broker

- GPU scheduling specialized for deep learning
- Control of GPU allocation at the processing unit level within the program
- Improves GPU utilization rate, reducing overall processing time
- More processes can be executed with the same GPU resources
- Unlike sharing of GPUs through virtualization technology, the program can fully utilize the GPU-equipped memory

Technology of AI Computing Broker

Analyze AI processing content, dynamically allocate GPU to processes that require GPU



Conclusion

- We have developed **AI Computing Broker (ACB)** which improves efficiency on HPC systems

- ACB achieves:
  - Adaptive Accelerator allocation
    - Dynamically switches among GPU accelerators and allocates them to processes requiring GPUs
  - Efficient GPU Utilization
    - Less GPUs required for running AI workloads

Thank you