AI Computing Broker

Fujitsu Limited

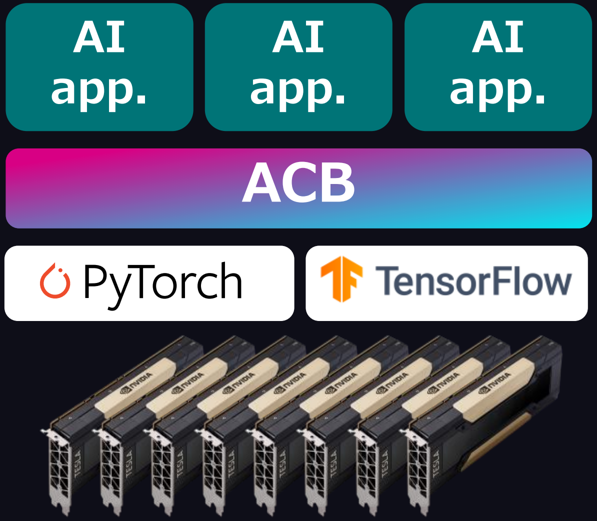
AI Computing Broker

A middleware to share GPUs among AI apps.

- works with a wide range of AI apps. based on **PyTorch** and **TensorFlow**
- just install and use; no code changes are needed

Key Features

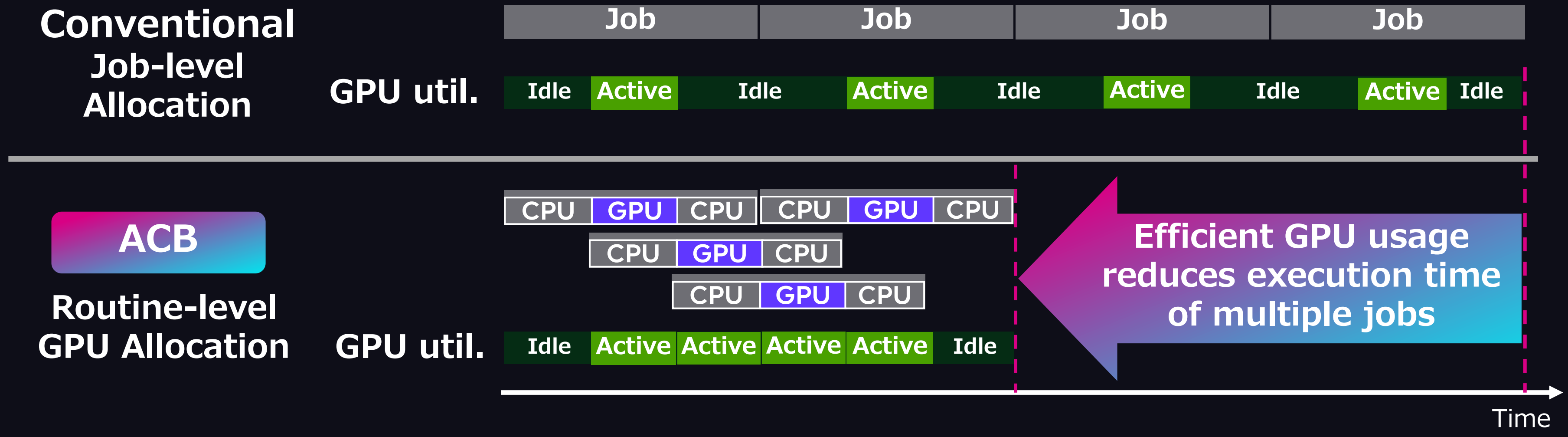
- Best-in-class GPU utilization efficiency
- Enabling full GPU memory for each job



Feature of AI Computing Broker

Best-in-class GPU utilization efficiency

- “Routine-level” allocation that detects actual GPU parts of jobs and dynamically allocates GPU accordingly



Conventional
Job-level
Allocation

Job | Job | Job | Job

GPU util. Idle Active Idle Active Idle Active Idle Active Idle

ACB

Routine-level
GPU Allocation

CPU GPU CPU | CPU GPU CPU

CPU GPU CPU

CPU GPU CPU

GPU util. Idle Active Active Active Active Idle

Efficient GPU usage reduces execution time of multiple jobs

Time

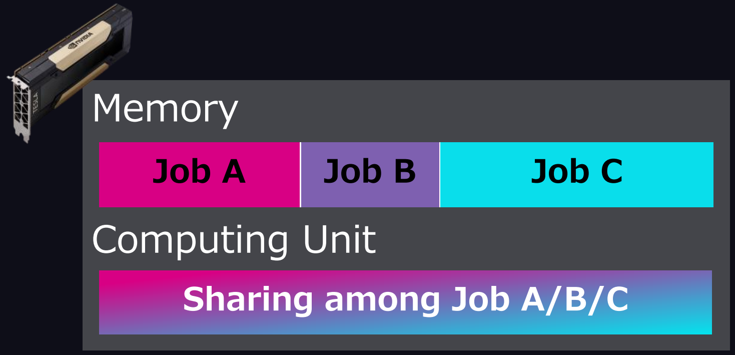
Feature of AI Computing Broker

Enabling full GPU memory for each job

- Allocate GPU to only one job at a time (Temporal-sharing)
- Data of other jobs on GPU is automatically swapped to CPU

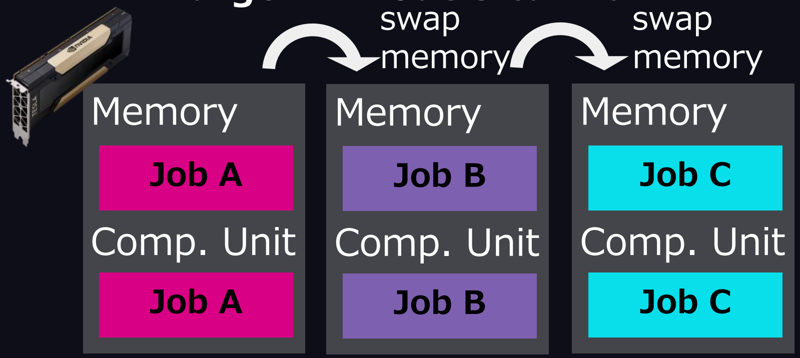
Conventional: Spatial-sharing

Memory is divided among jobs
Limited to small AI models

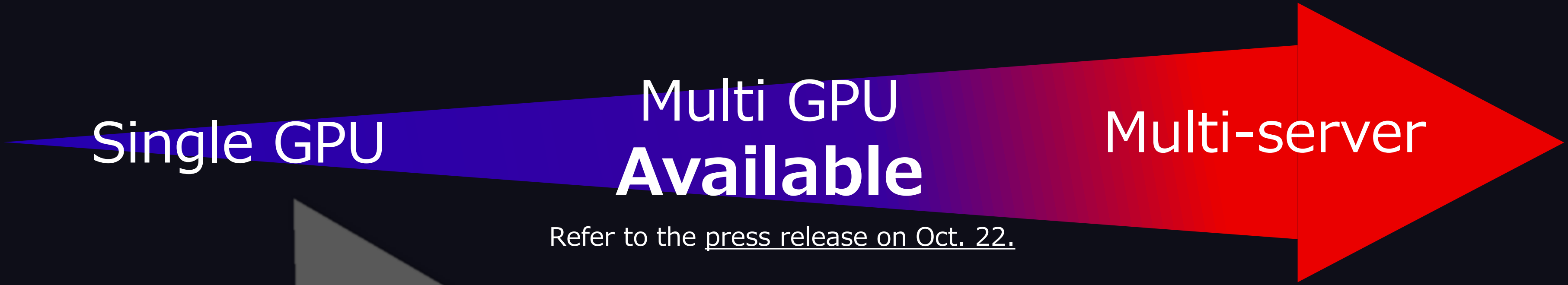


ACB Temporal-sharing

Memory is occupied by each job
Large AI models can run



Development status



Single GPU

Multi GPU

Available

Multi-server

Refer to the press release on Oct. 22.

- Small-scale AI tasks
  (e.g., Image recognition)

- Using multiple GPU in a server
- LLM inference, fine-tuning

- Using multiple servers
- Large-scale LLM training

Currency prediction service

TRADOM Inc.

Doubled the
model training throughput

Data Center Service

Sakura Internet Inc.

Deploying more AI tasks
beyond hardware limits

FUJITSU

Thank you

For more details:
https://www.fujitsu.com/global/products/computing/servers/supercomputer/topics/sc24/