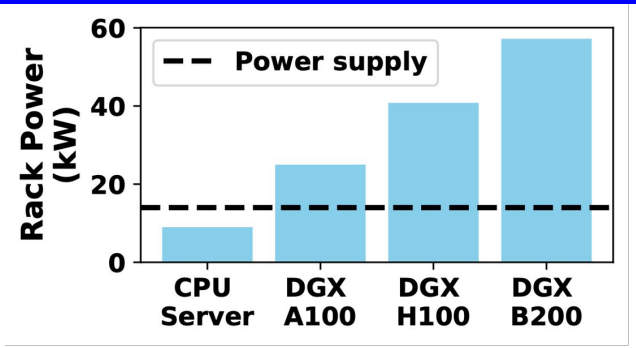
Mange the Workloads, not the Cluster: Designing a Control Plane for Large-Scale AI Cluster

LLM online services are booming

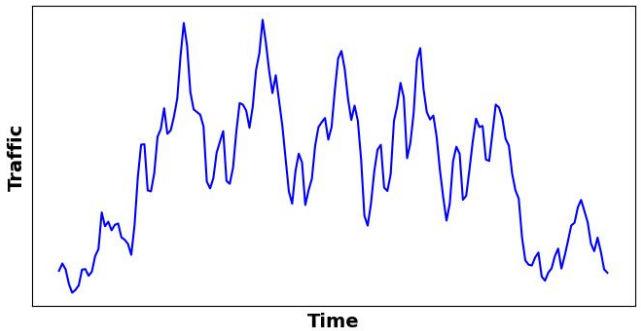
- Meta, OpenAI, Microsoft are building clusters with 10,000 GPUs
- Inference traffic accounts > 90% portion of GPU usage

Power usage is increasing



Highly dynamic traffic pattern

- The request arrival rate fluctuates



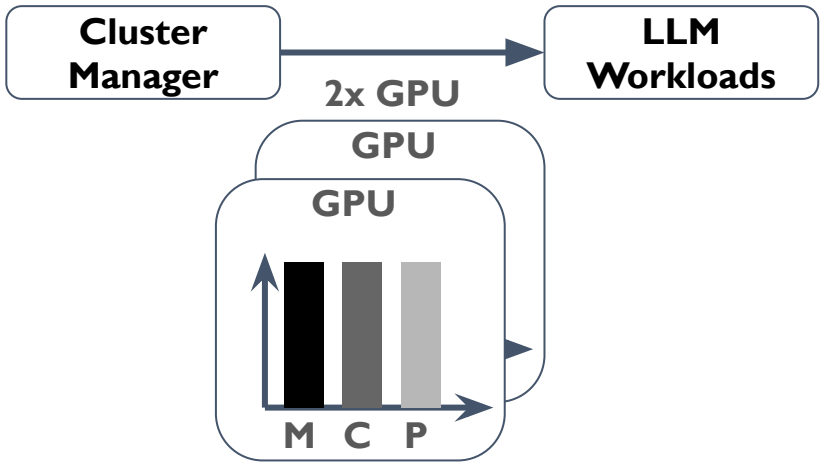
We need auto-scaling capable cluster manager for LLM

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Mange the Workloads, not the Cluster: Designing a Control Plane for Large-Scale AI Cluster

Existing cluster managers

- Allocate resources as an entire GPU
- Resources are tightly coupled with each other



Existing GPU cluster managers provision resources statically

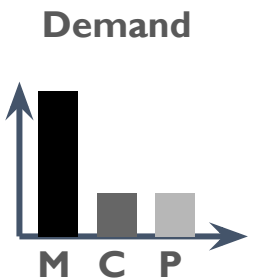
Problem

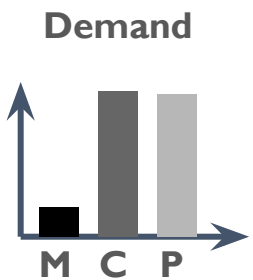
Long Requests

"Please write me a 10,000 words novel"

Short Requests

"What is the capital of Netherlands?"
"What is the capital of USA?"
"What is the capital of …?"





Workload resource demands vary over time

Learn more in our poster!

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE