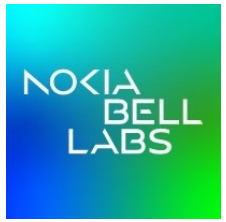
# FaaScale: Unlocking Fast LLM Scaling for Serverless Inference

**Minchen Yu**[1]*, Rui Yang[2]*, Chaobo Jia[1], Zhaoyuan Su[2], Sheng Yao[3], Tingfeng Lan[2], Yuchen Yang[3], Zirui Wang[2], Yue Cheng[2], Wei Wang[3], Ao Wang[4], Ruichuan Chen[5]

[1]CUHK-Shenzhen [2]UVA [3]HKUST [4]Alibaba Group [5]Nokia Bell Labs

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

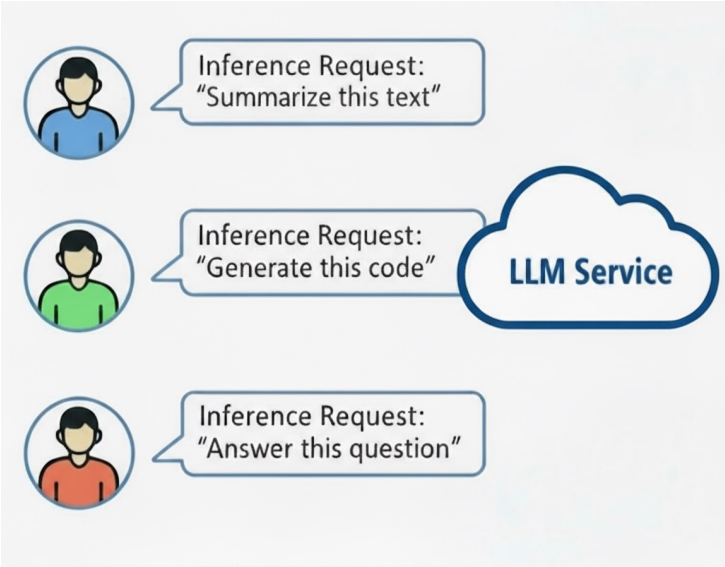


# Serverless LLM inference

LLM inference as an online service

- SLOs under dynamic requests
  - e.g., Time-To-First-Token (TTFT)
- Cost effectiveness



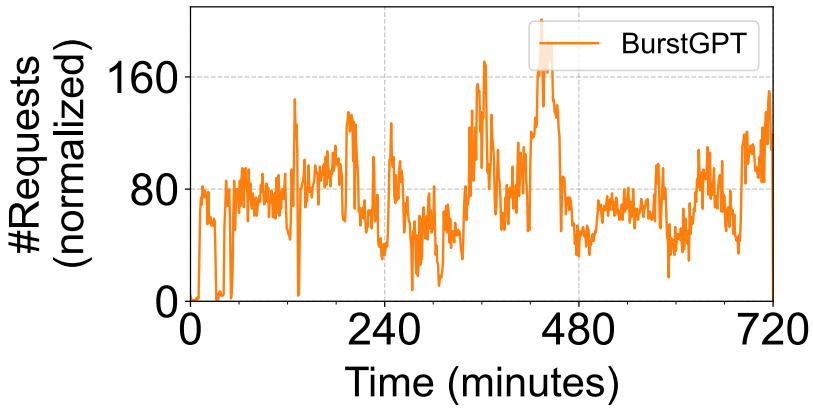

## Serverless LLM inference

- Managed inference services
- Automatic scaling with pay-per-use billing

Leading cloud providers offer serverless LLM inference APIs

# Why scalable LLM inference challenging?

- Request bursts arrive quickly
  - Demand can spike 10x within minutes
- Models are too large to fetch on demand
  - Hundreds of GBs to several TBs of memory
- Too many models to keep warm
  - >2M models on Hugging Face



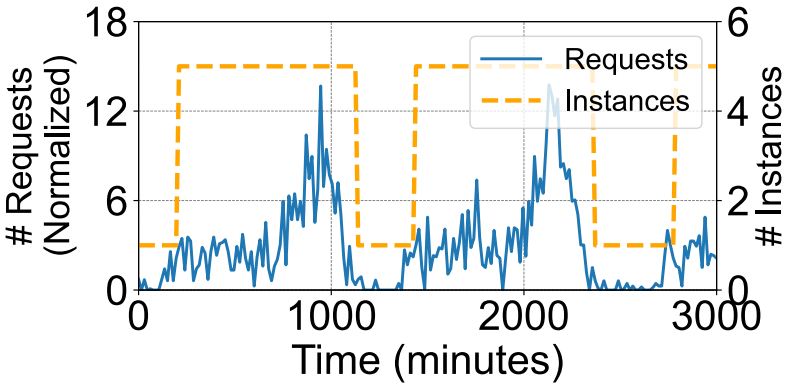

Scalable LLM inference is hard to achieve

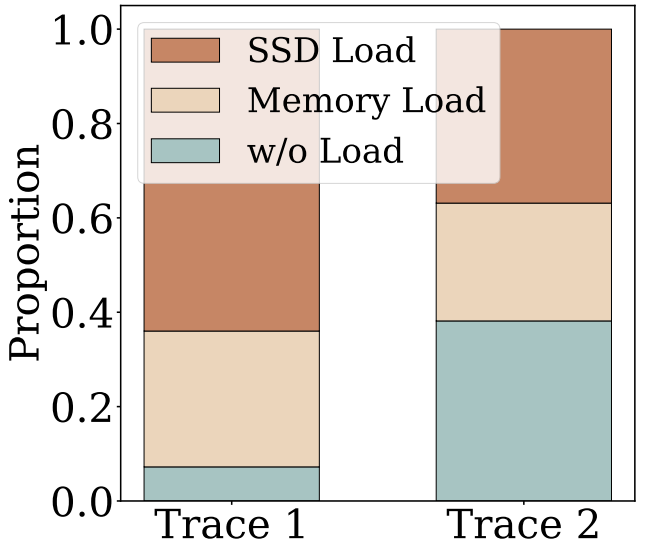
# Existing solutions to serverless inference

- Load models from remote storage
  - Slow startup and bandwidth contention under concurrent fetches

- Overprovision GPUs
  - Fast responses, but costly idle capacity

- Cache models in host memory and SSDs
  - Works only when the requested model remains warm
  - Cache misses fall back to slow SSD or remote loading

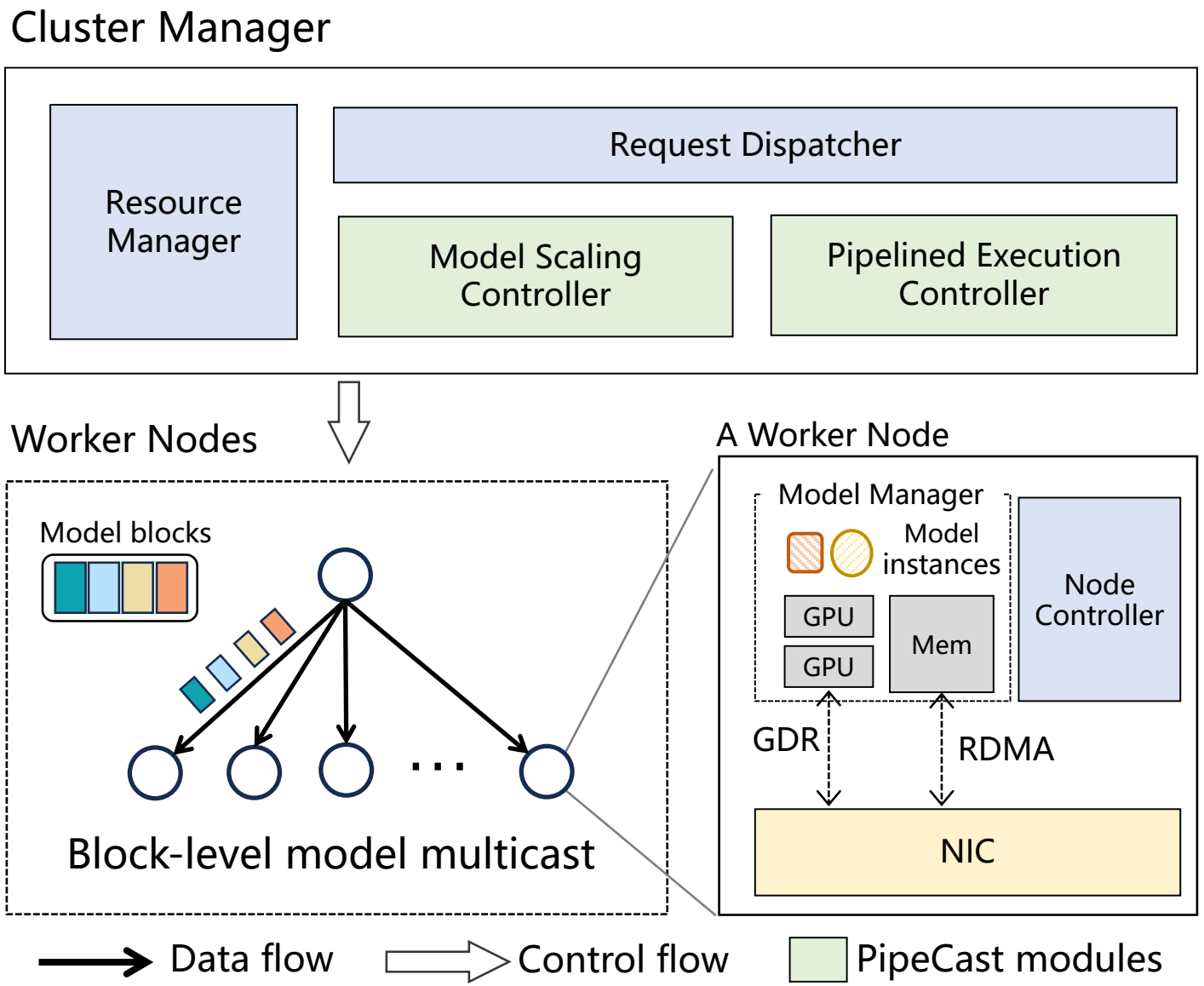


# Key insights

- Move models between GPU nodes over high-speed interconnects (200–400 Gbps with RDMA)
  - Fast model multicast without caching many models in memory

- Inference execution can begin while models are being loaded
  - Execute-while-load: collaborative inference during model multicast




# FaaScale overview

- Execute-while-load: start inference before all blocks arrive
- Block-level model multicast
- GDR-based model transfer from GPU/host memory

# Key challenges and design overview

## Coordinated model multicast and inference execution

- Inference-aware model multicast
- Pipelined, collaborative inference execution

## Cross-storage-tier model management

- Locality-driven model startup
- Efficient GPU and host memory management

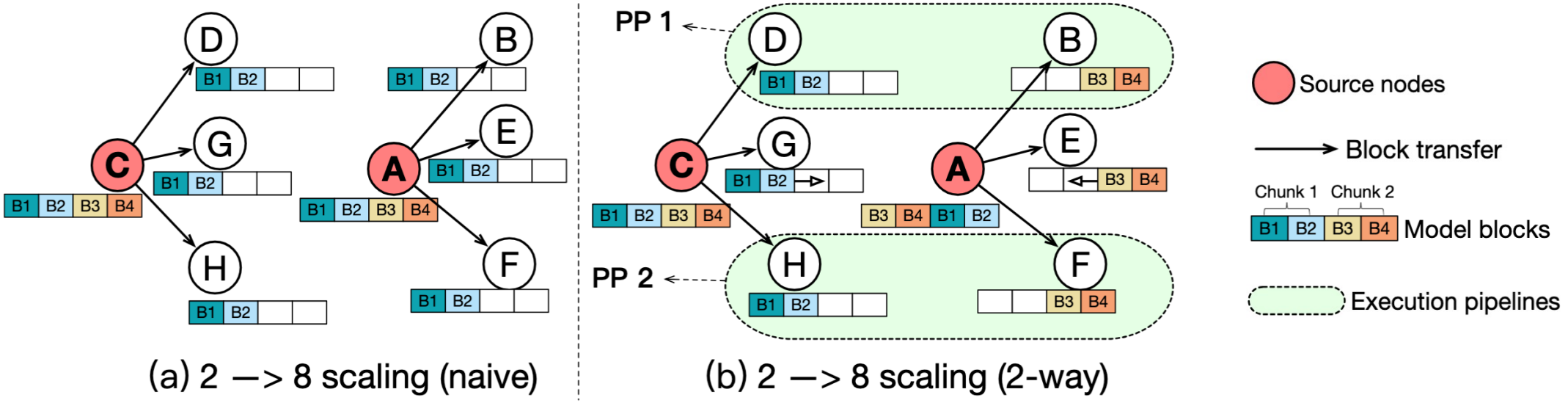
# FaaScale Design

# Inference-aware model multicast

How to perform multicast for minimizing request response times?

Collectively construct complete model copies as early as possible!

- k → N: k source nodes distribute the model to N – k nodes
- K-way transmission: split into k sub-group and rotate block transfers via circular shift

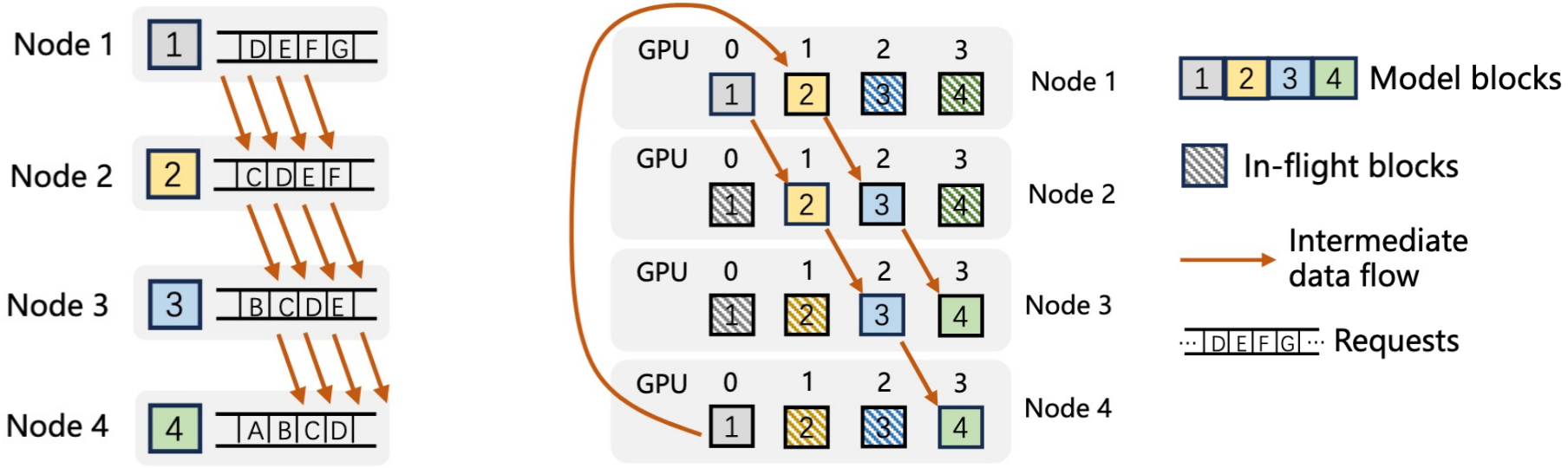


# Pipelined collaborative inference

How to overlap inference with multicast without slowing data transfer?

Keep it simple: transfer only intermediate activations, not KV cache

- 2D execution pipeline: stage model blocks on GPUs and pipeline request batches across nodes



# Cross-tier model management

## Locality-driven model startup

- GPU (hot start): schedule requests first
- Host memory (warm start): pipeline across warm instances
- No cache (cold start): scale quickly from hot/warm instances

## Memory-management optimizations

- Consolidate tensors into contiguous chunks for bulk transfer
- Maintain a GPU memory pool to reduce allocation overhead

# Evaluation

# Experimental settings

Setup

- H800 worker nodes, each with 400Gbps NIC
- Up to 12 workers

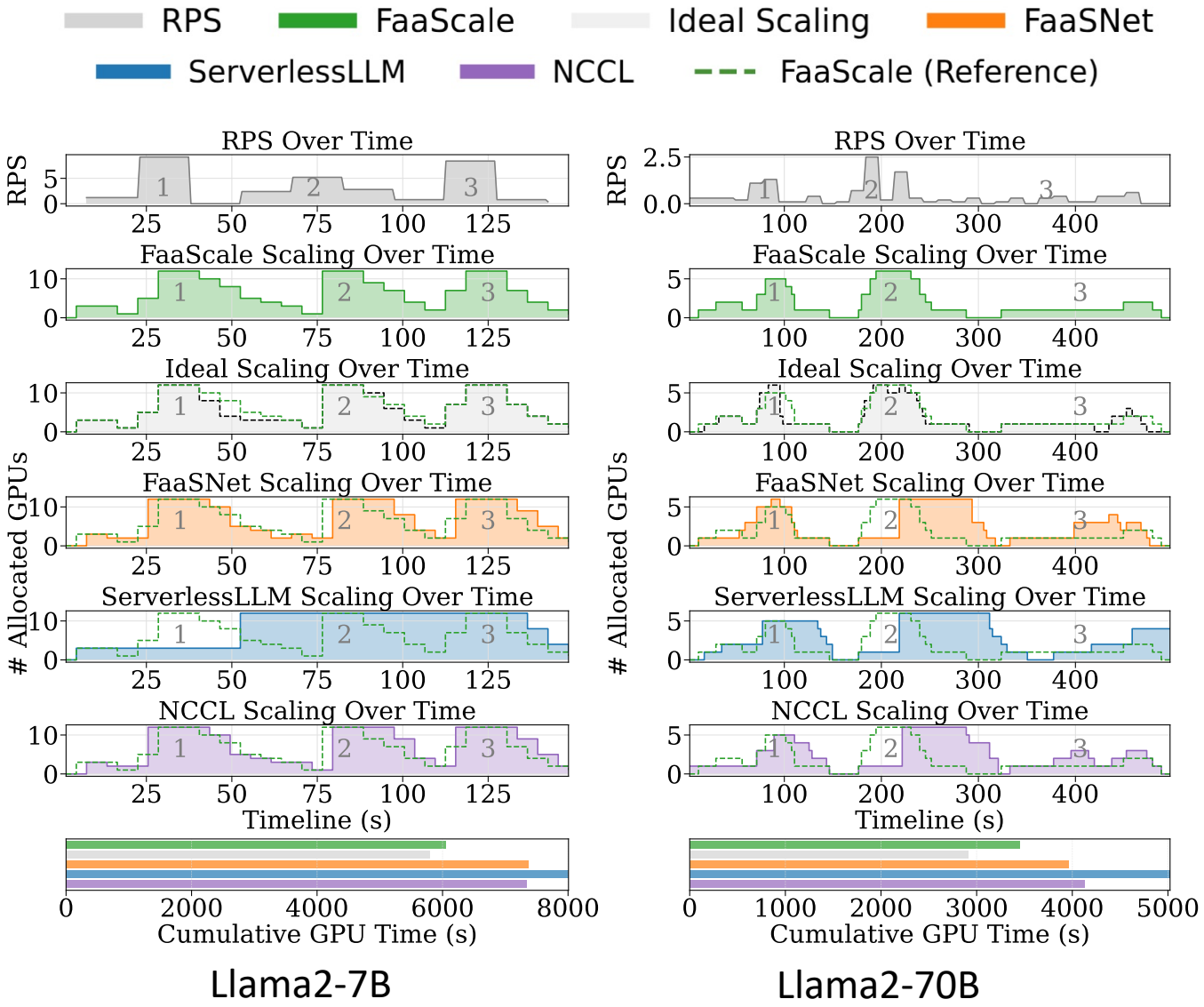
Metrics

- Latency (TTFT)
- Throughput (TPS)
- Resource cost (GPU time)

Baseline

- ServerlessLLM, FaaSNet, NCCL

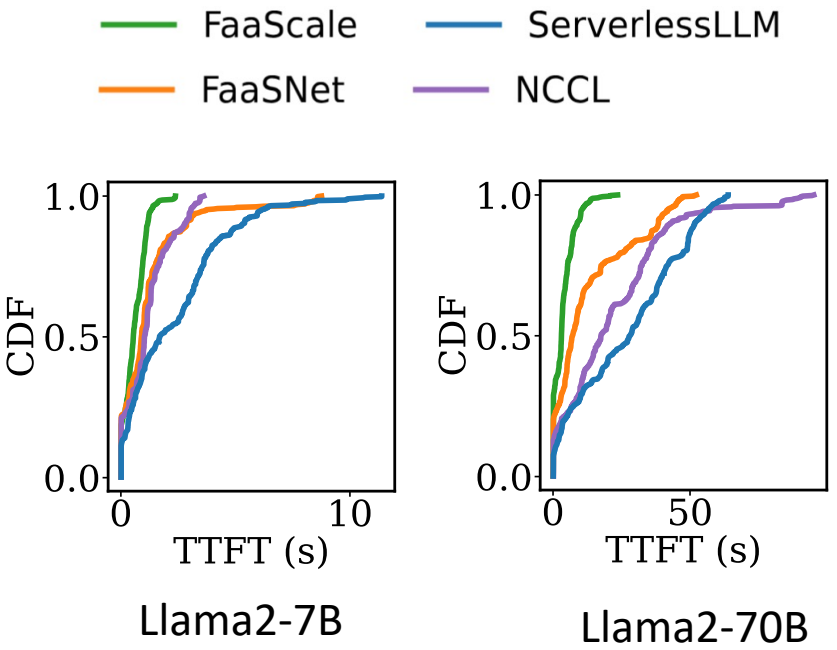
# End-to-end performance




Llama2-7B




Llama2-70B




Llama2-7B

Llama2-70B

- 31% cost reduction
- 5x P90 TTFT latency improvement

# Summary

FaaScale enables fast scale-out for serverless LLM inference

- Leveraging high-speed interconnects for fast model multicast
- Cooperative, distributed inference execution during model loading
- Real-world traces show lower GPU cost and better tail latency

香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

UNIVERSITY of VIRGINIA

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Alibaba Group 阿里巴巴集团

NOKIA BELL LABS

# Questions?

yuminchen@cuhk.edu.cn