Alibaba Cloud
Worldwide Cloud Services Partner

GPU Resource Pooling and the benefit of deploying CUPTI

Alibaba Group

Lingling Jin, Lingjie Xu

Alibaba is an AI company

Alibaba Cloud
Worldwide Cloud Services Partner

Alibaba Cloud
aliyun.com

- E-commerce
  - Search
  - Advertisement
- Financing & payment
- Video
- Logistics

All these services are backed up by Alibaba cloud

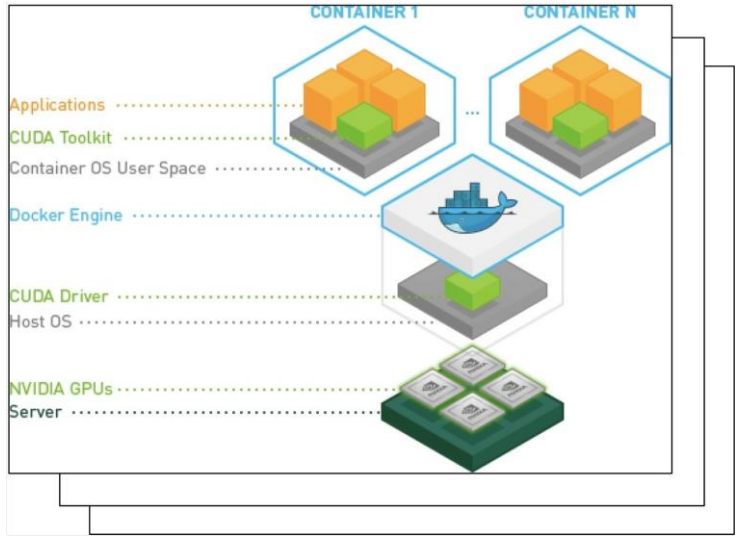
GPU pool

IDC data center

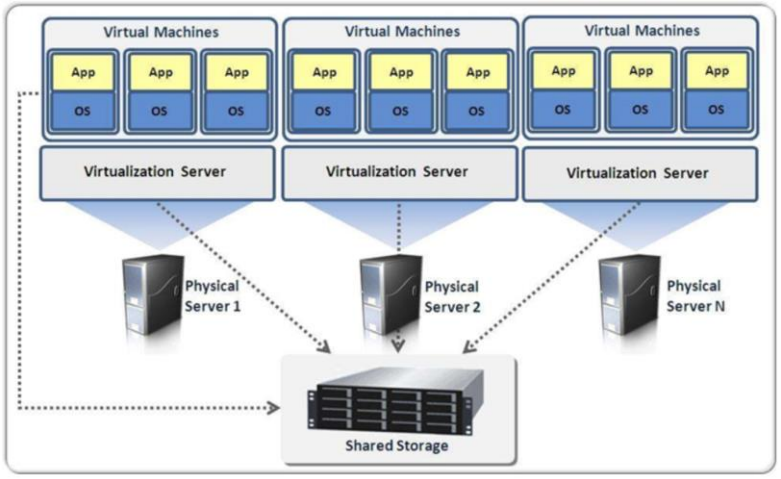
- GPU accelerator powered data center

GPU servers are managed elastically

Alibaba Cloud
Worldwide Cloud Services Partner



Kubernetes



KVM

However GPU utilization is still low

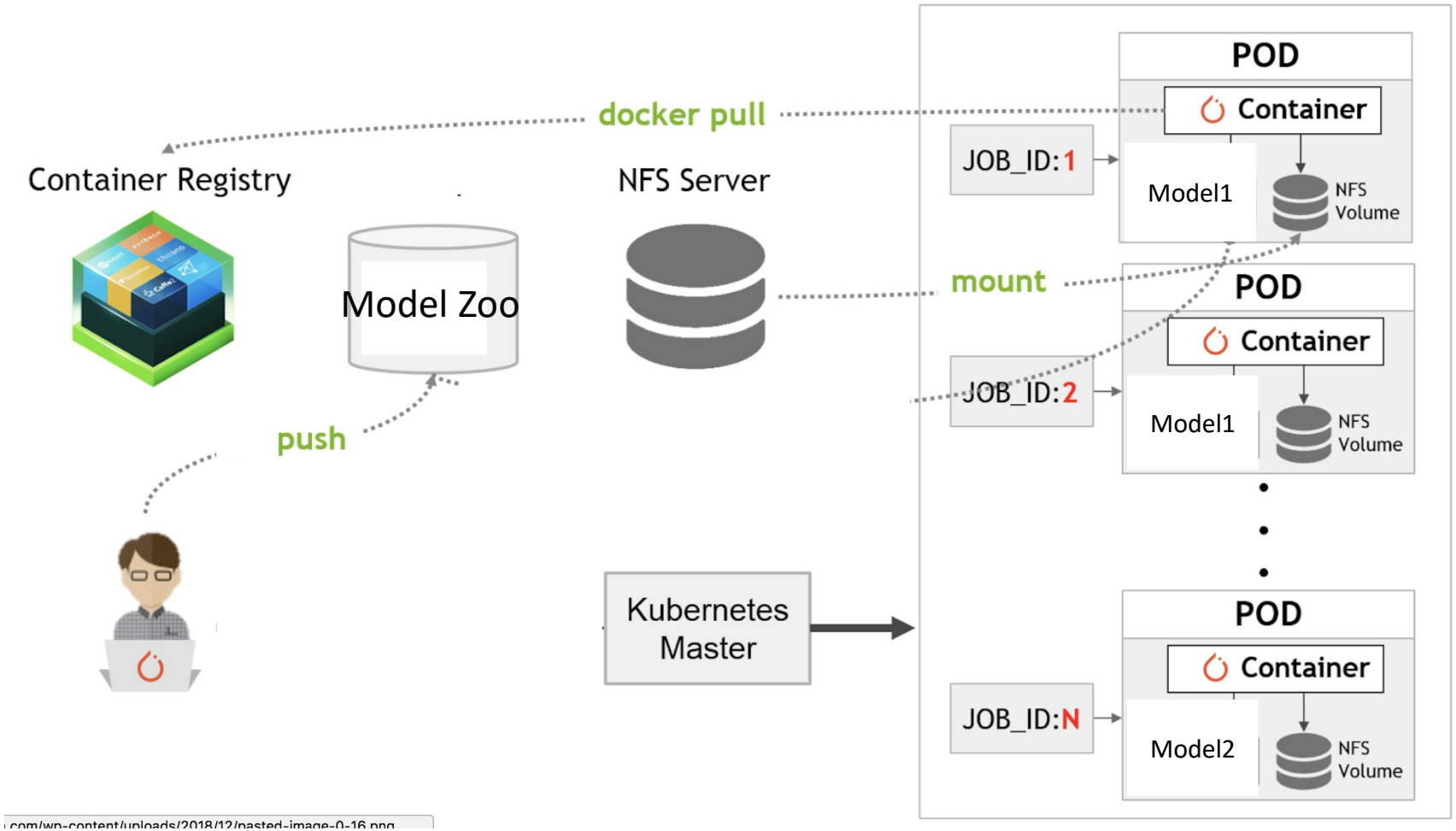
Agenda

Alibaba Cloud
Worldwide Cloud Services Partner

- **GPU accelerated inference service pool: multi-tenant**
- GPU resource pool over ethernet

GPU inference service pool

Alibaba Cloud
Worldwide Cloud Services Partner

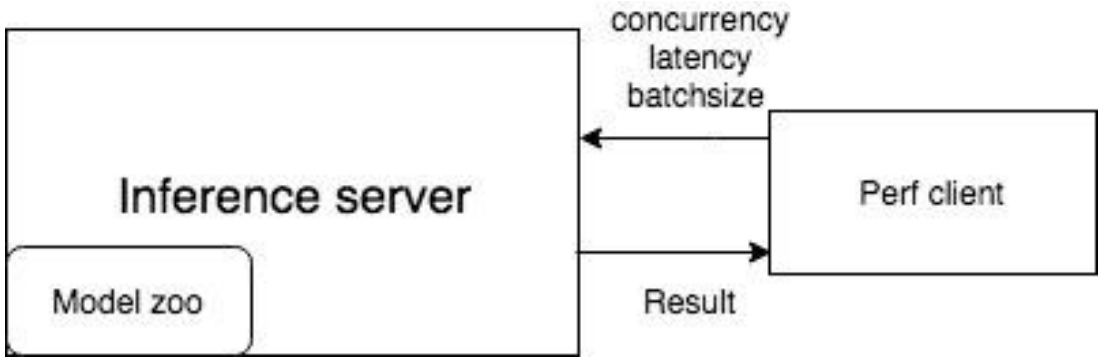


- GPUs are allocated in single card granularity
- Capacity is designed for worst case traffic
- Mixing different inference possible?

Experiment setup

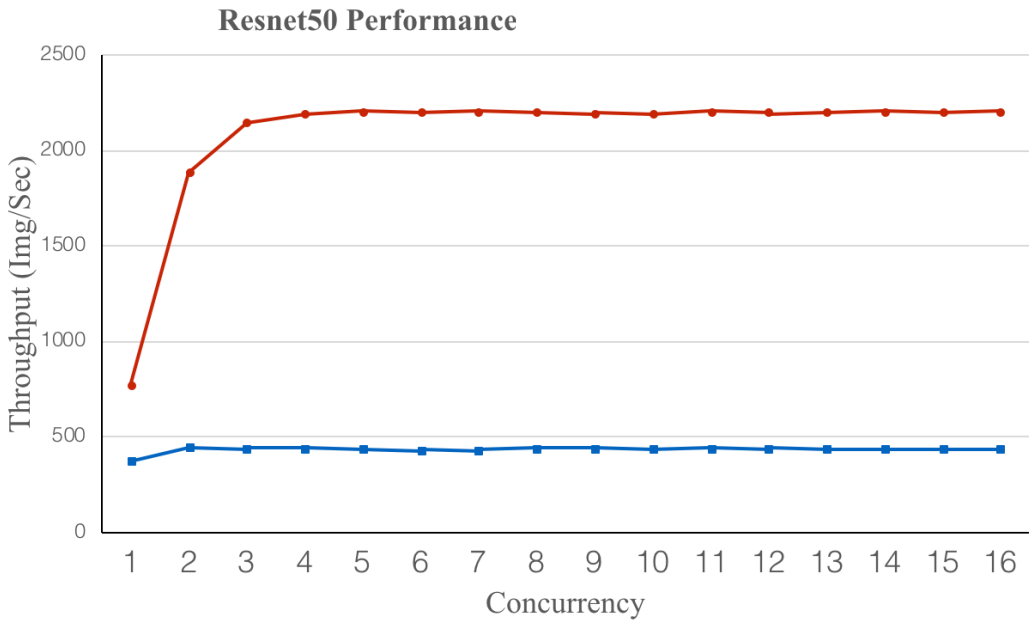
Alibaba Cloud
Worldwide Cloud Services Partner

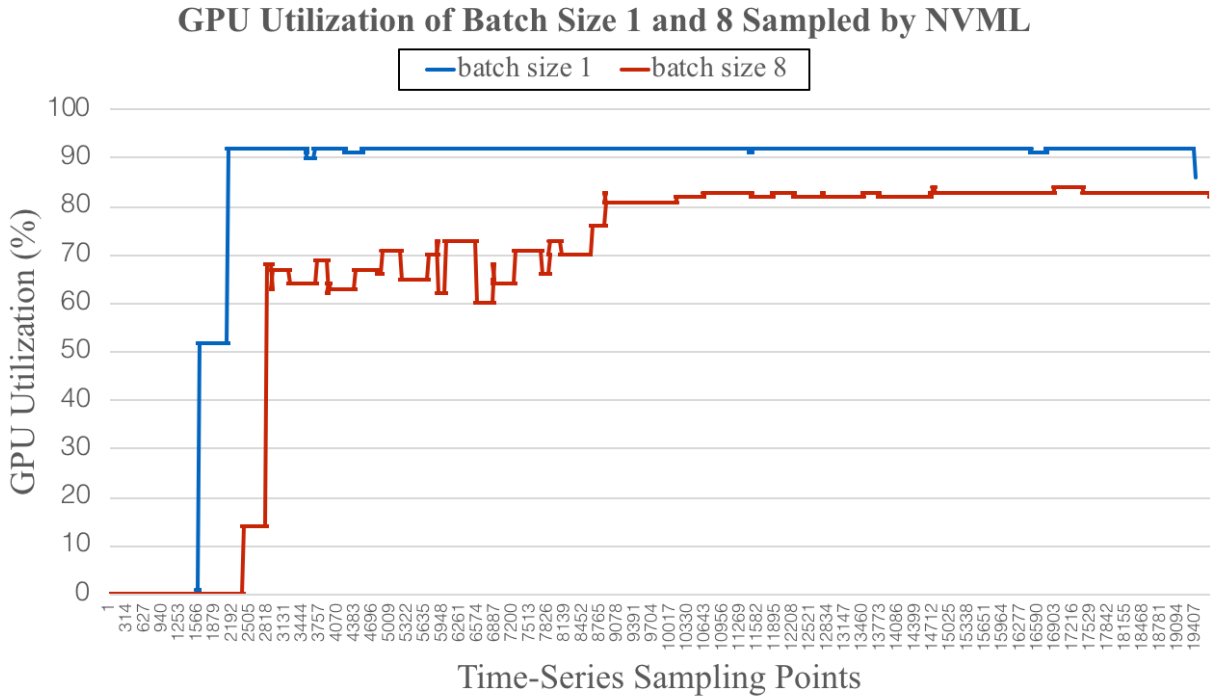
- Single GPU V100 system
- Model: resnet50
- Start inference server, capture NVML GPU utilization at backend
- Client performance test
  - multiple concurrent request,
  - max allowed latency
  - desired batchsize
  - Calculate server throughput



NVML GPU utilization

Alibaba Cloud
Worldwide Cloud Services Partner



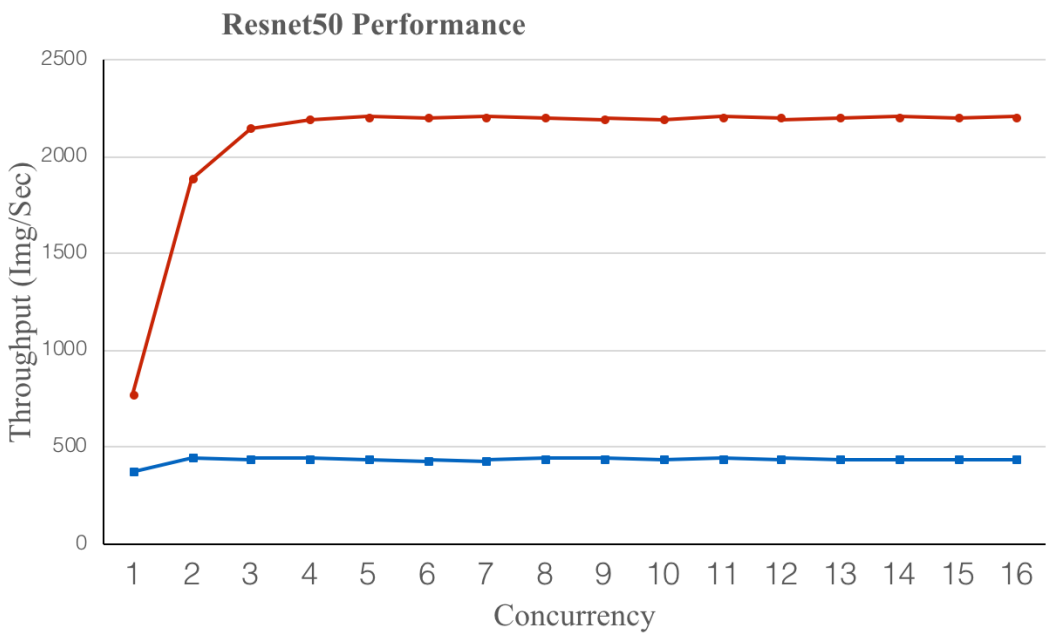


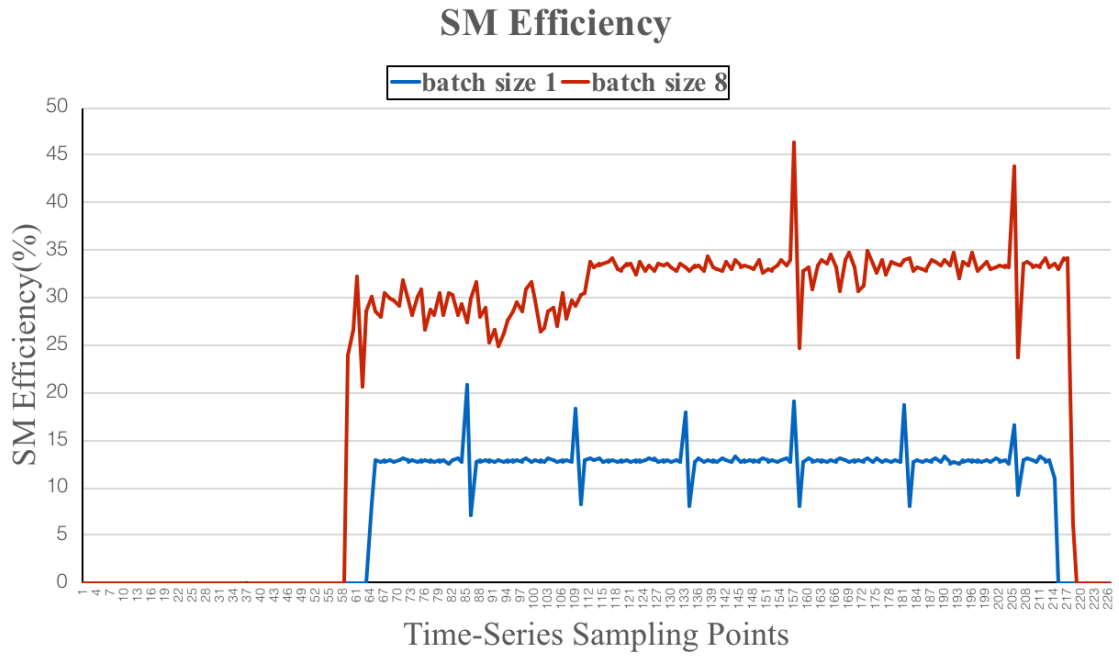
GPU_utilization: Percent of time over the past sample period during which one or more kernels was executing on the GPU

CUPTI

- The CUDA Profiling Tools Interface (CUPTI) enables the creation of profiling and tracing tools that target CUDA applications.
- all event/metrics available in NVPROF
    - sm_efficiency
    - achieved_occupancy
    - dram_read/write
    - Single/double/half_precision_fu_utilization
- Collect GPU metrics during runtime
- 1%-10% overhead

CUPTI SM Efficiency





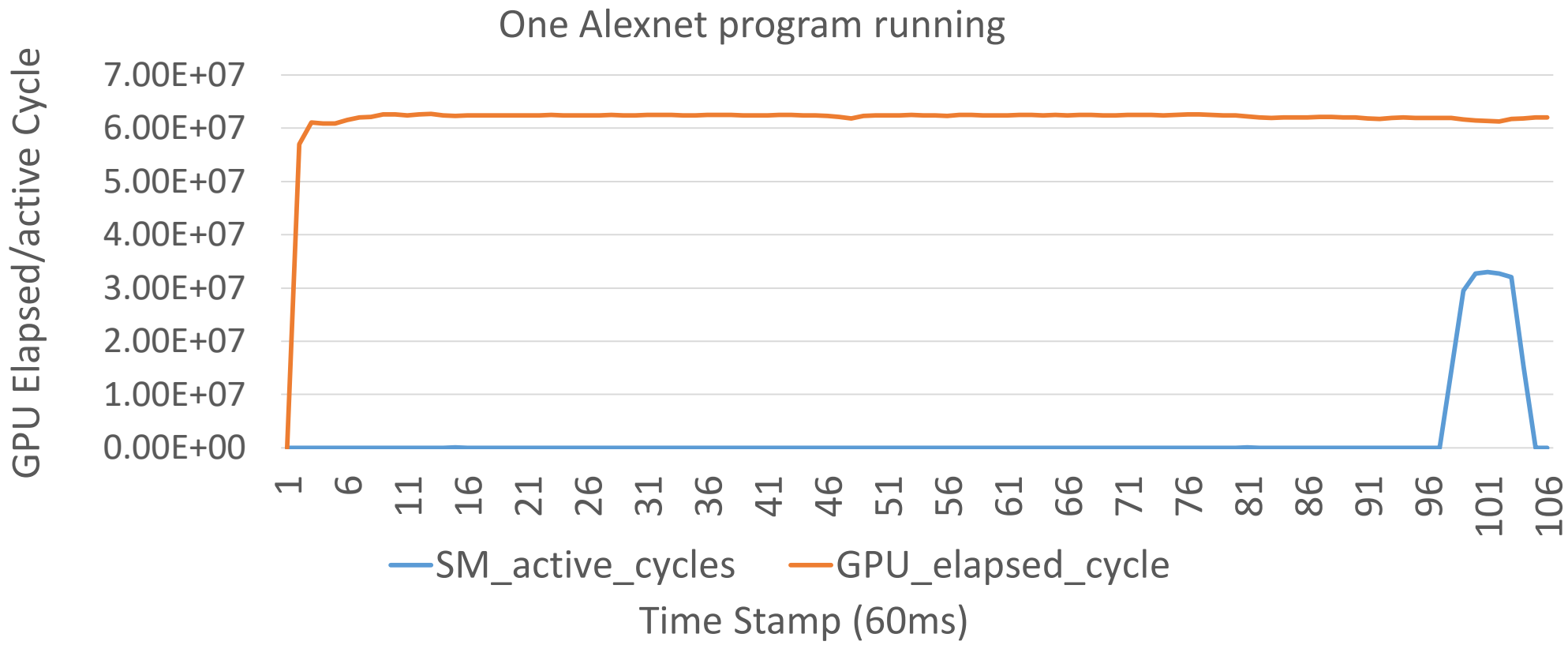
SM efficiency: The percentage of time at least one warp is active on a multiprocessor averaged over all multiprocessors on the GPU.

CUPTI is what we want.

Explore two processes GPU sharing

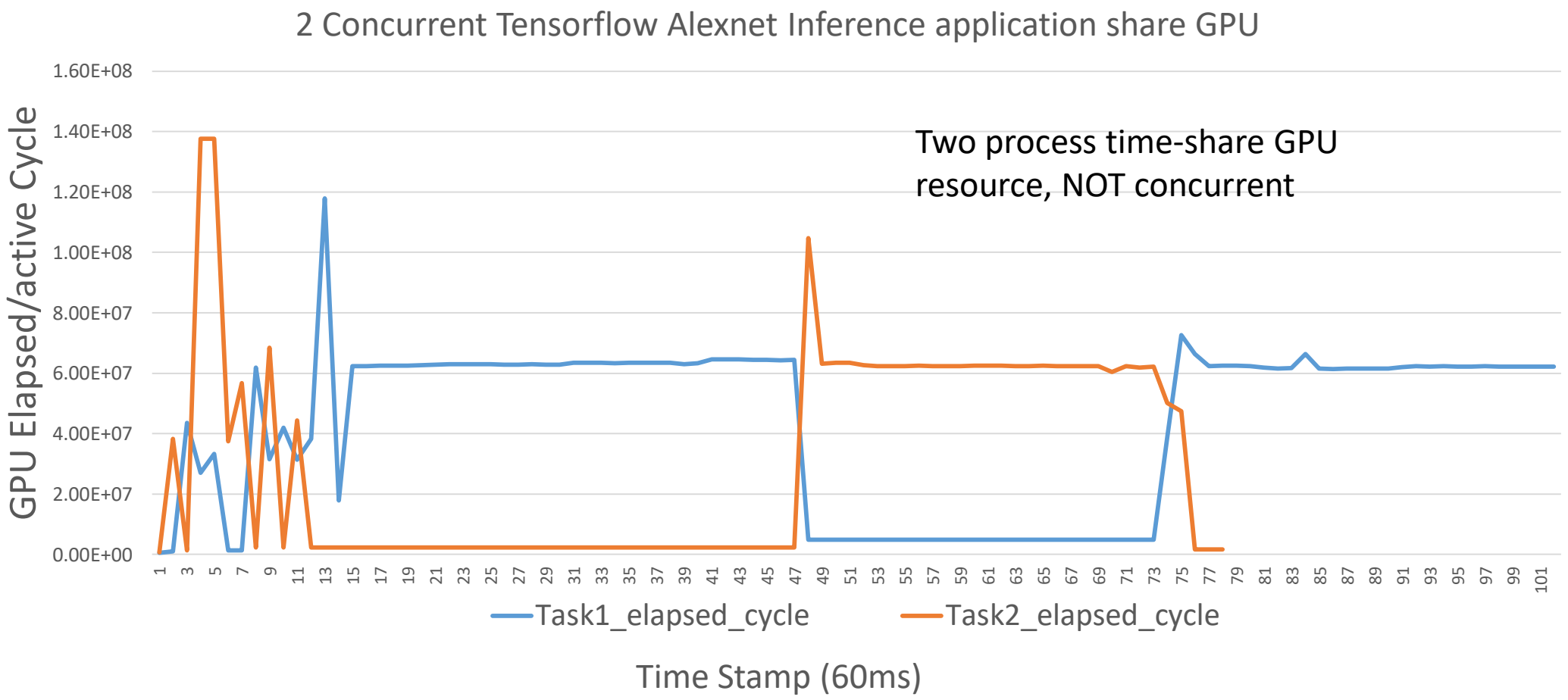
Alibaba Cloud
Worldwide Cloud Services Partner

- We used cupti to capture GPU elapsed cycle and SM_active_cycle



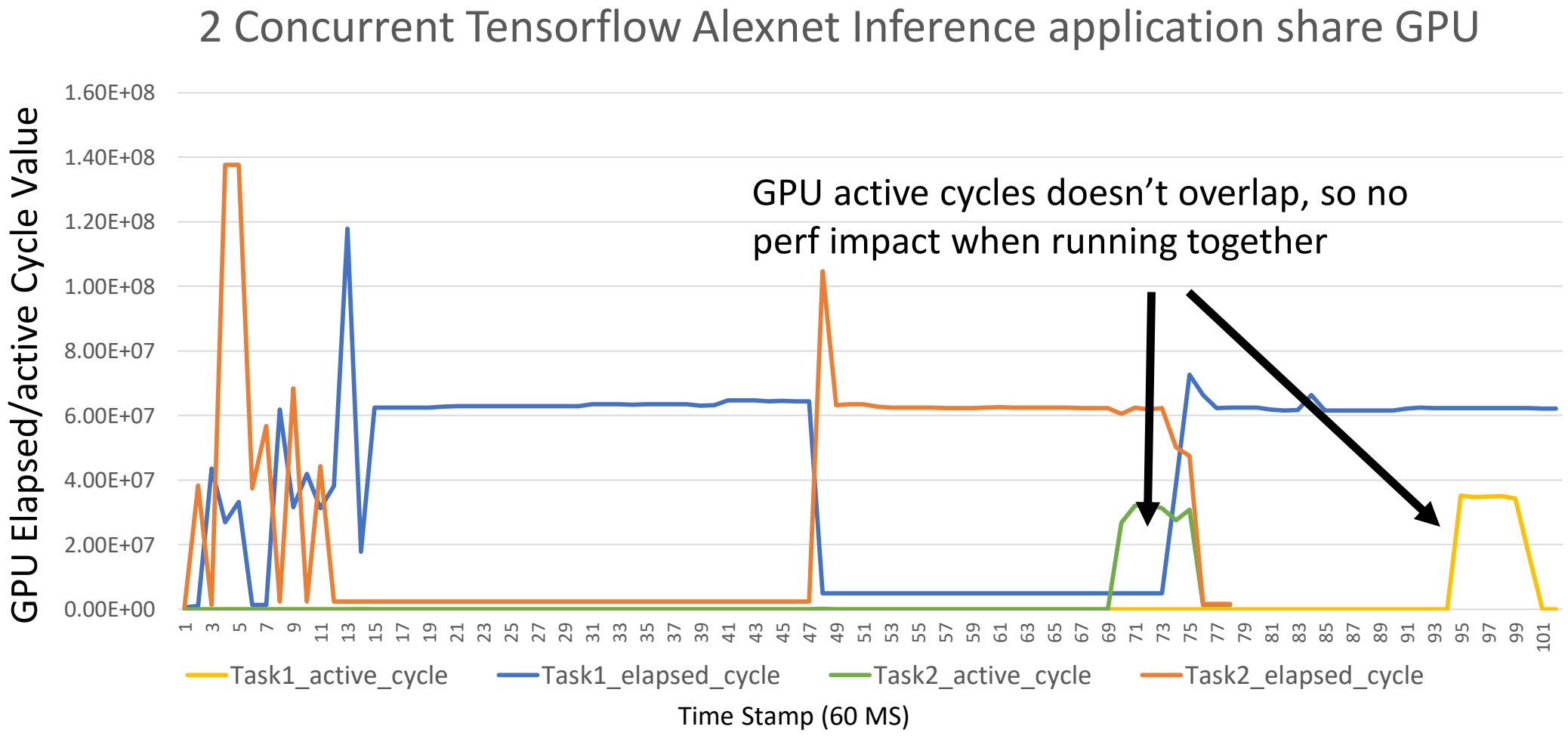
Explore multi-process GPU sharing

Alibaba Cloud
Worldwide Cloud Services Partner



Explore multi-process GPU sharing

Alibaba Cloud
Worldwide Cloud Services Partner



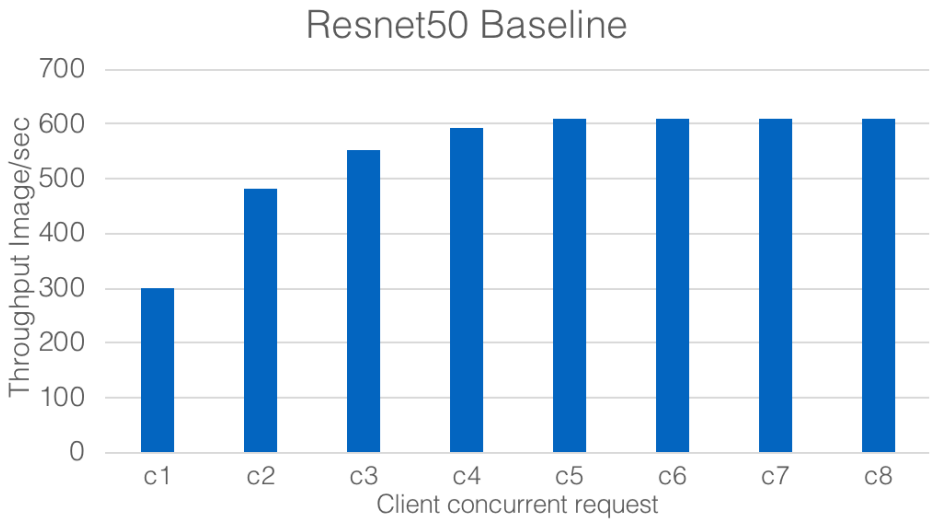
Running two inference servers

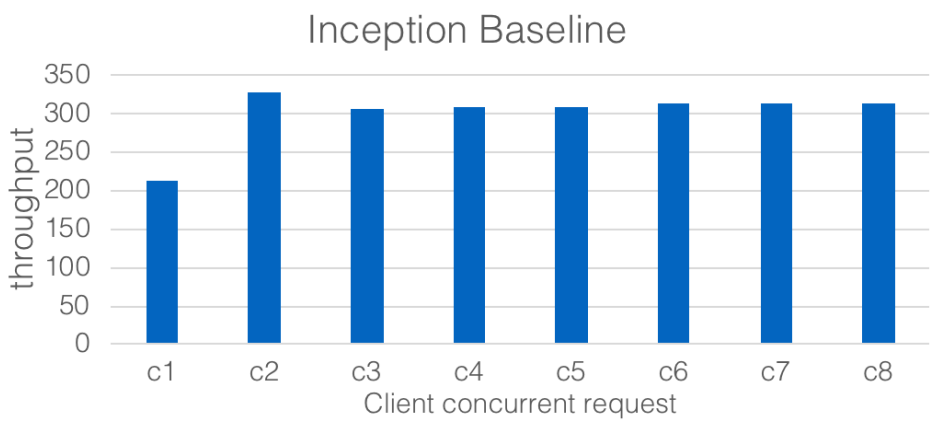
Alibaba Cloud
Worldwide Cloud Services Partner

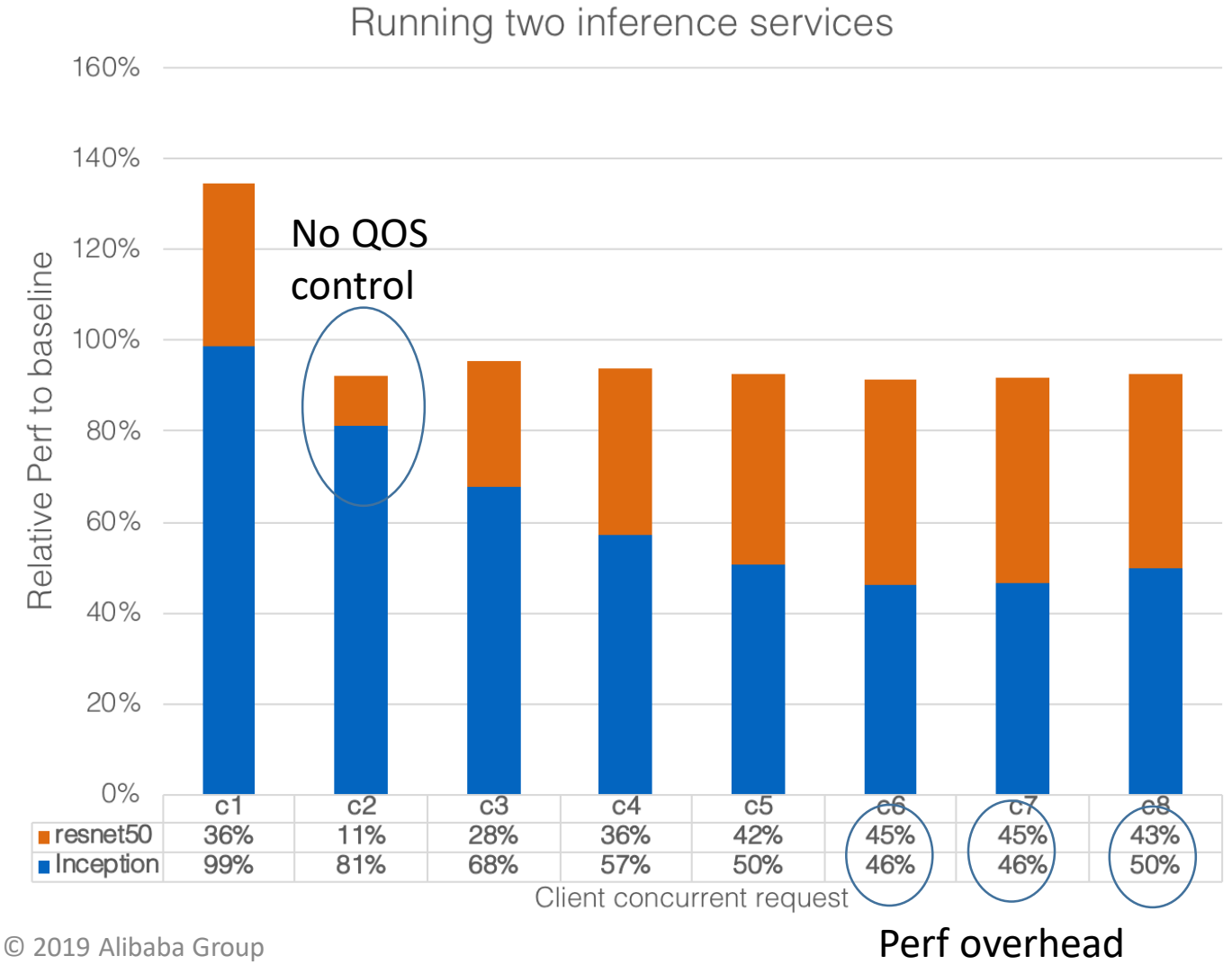
- Single GPU V100 system
- Model
  - resnet50
  - Inception
- Batchsize 4

Performance with two inference servers

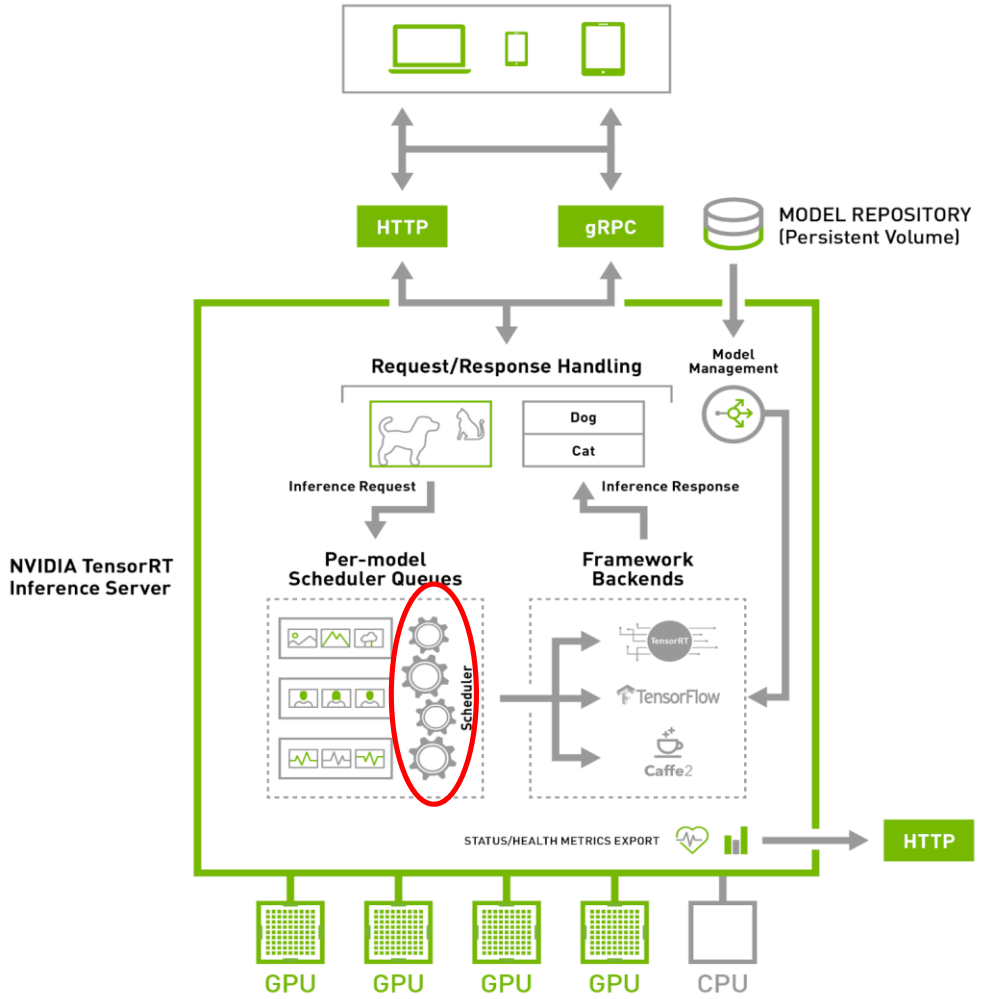
Alibaba Cloud
Worldwide Cloud Services Partner





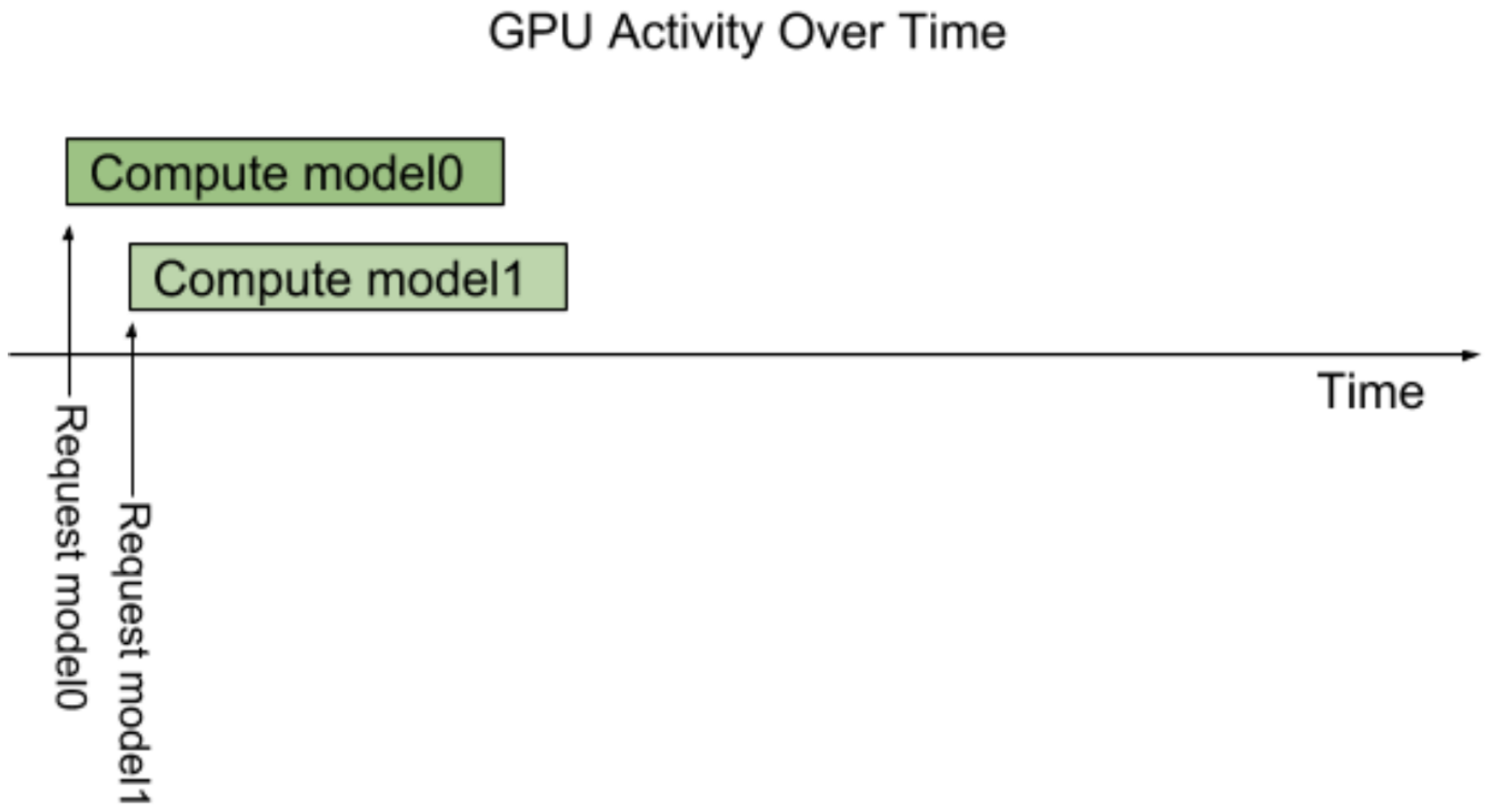


Nvidia inference server



- Each model is implemented as GPU stream, can be run concurrently on GPU
- better control with scheduler

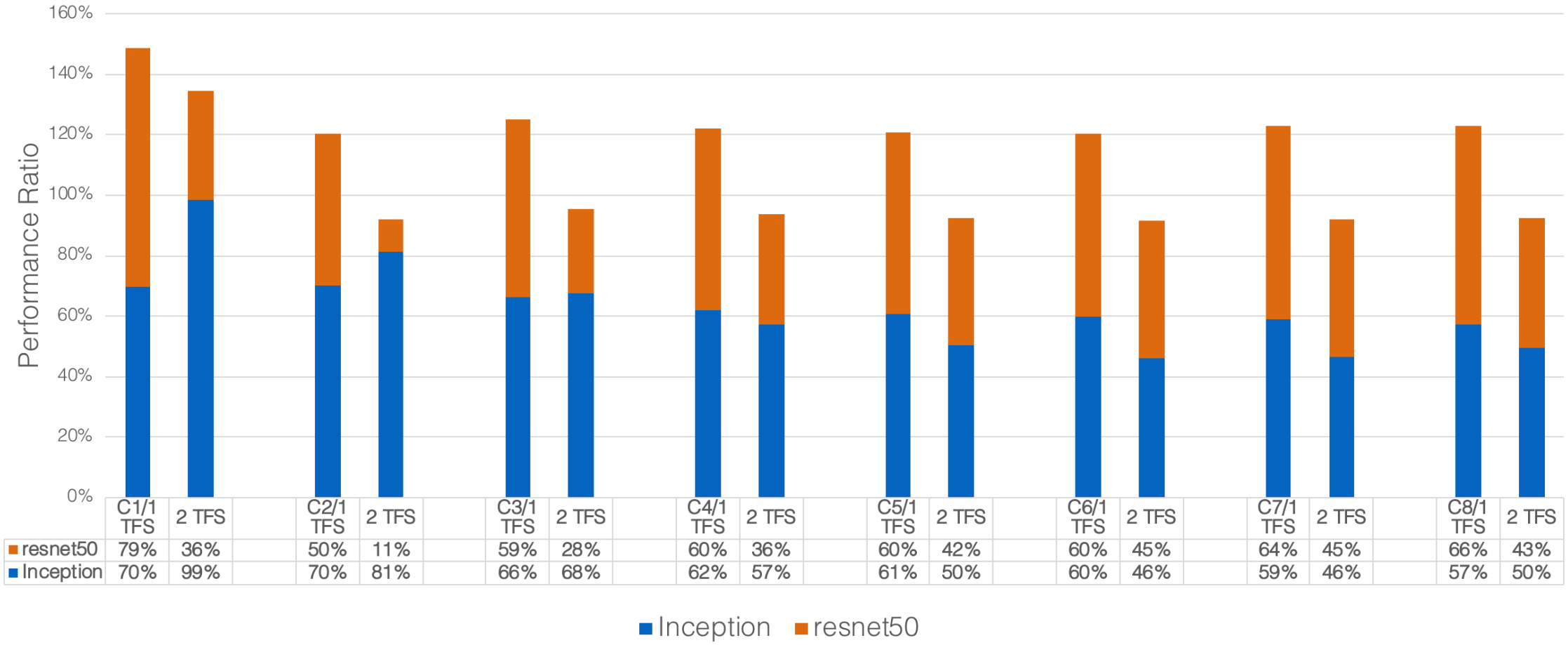
Nvidia inference server



1 server vs 2 server performance comparison

Alibaba Cloud
Worldwide Cloud Services Partner

Inception + resnet50, batchsize8, one instance



	C1/1 TFS	2 TFS	C2/1 TFS	2 TFS	C3/1 TFS	2 TFS	C4/1 TFS	2 TFS	C5/1 TFS	2 TFS	C6/1 TFS	2 TFS	C7/1 TFS	2 TFS	C8/1 TFS	2 TFS
resnet50	79%	36%	50%	11%	59%	28%	60%	36%	60%	42%	60%	45%	64%	45%	66%	43%
Inception	70%	99%	70%	81%	66%	68%	62%	57%	61%	50%	60%	46%	59%	46%	57%	50%

Agenda

Alibaba Cloud
Worldwide Cloud Services Partner

- GPU accelerated inference service pool: multi-tenant
- **GPU resource pool over ethernet**

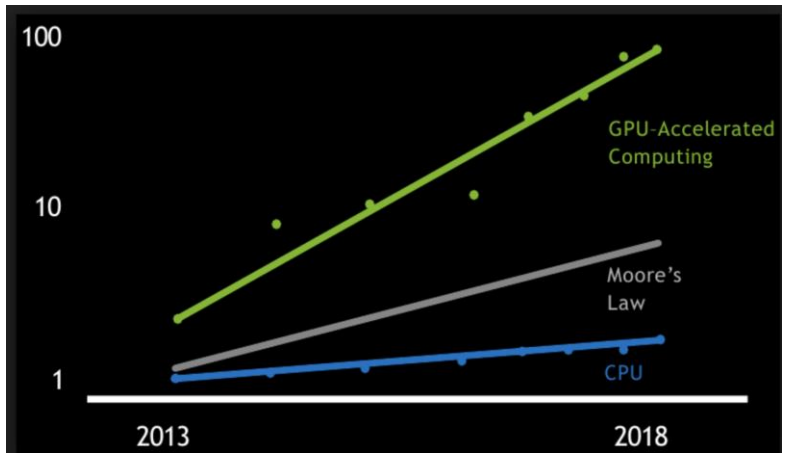
Why

DGX2

DGX1

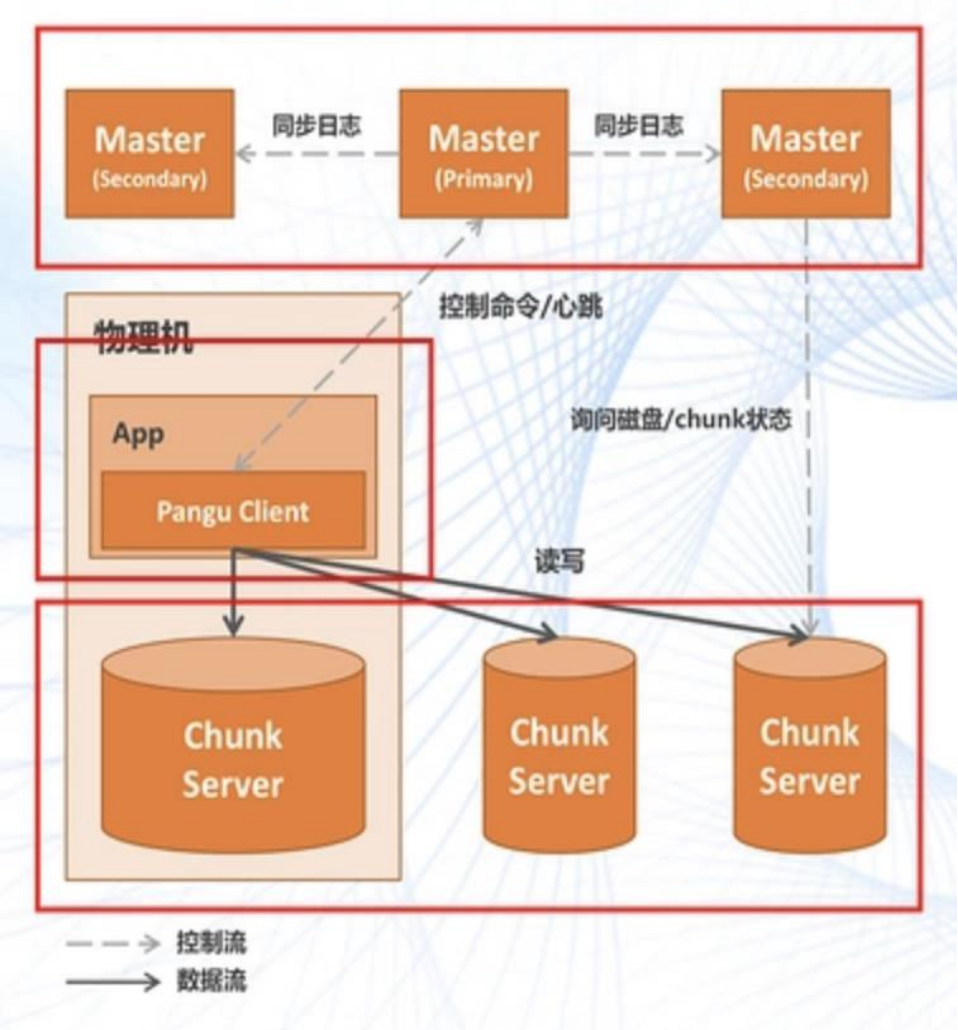
1U GPU server

- Each AI application has different CPU/GPU ratio
- many types GPU servers



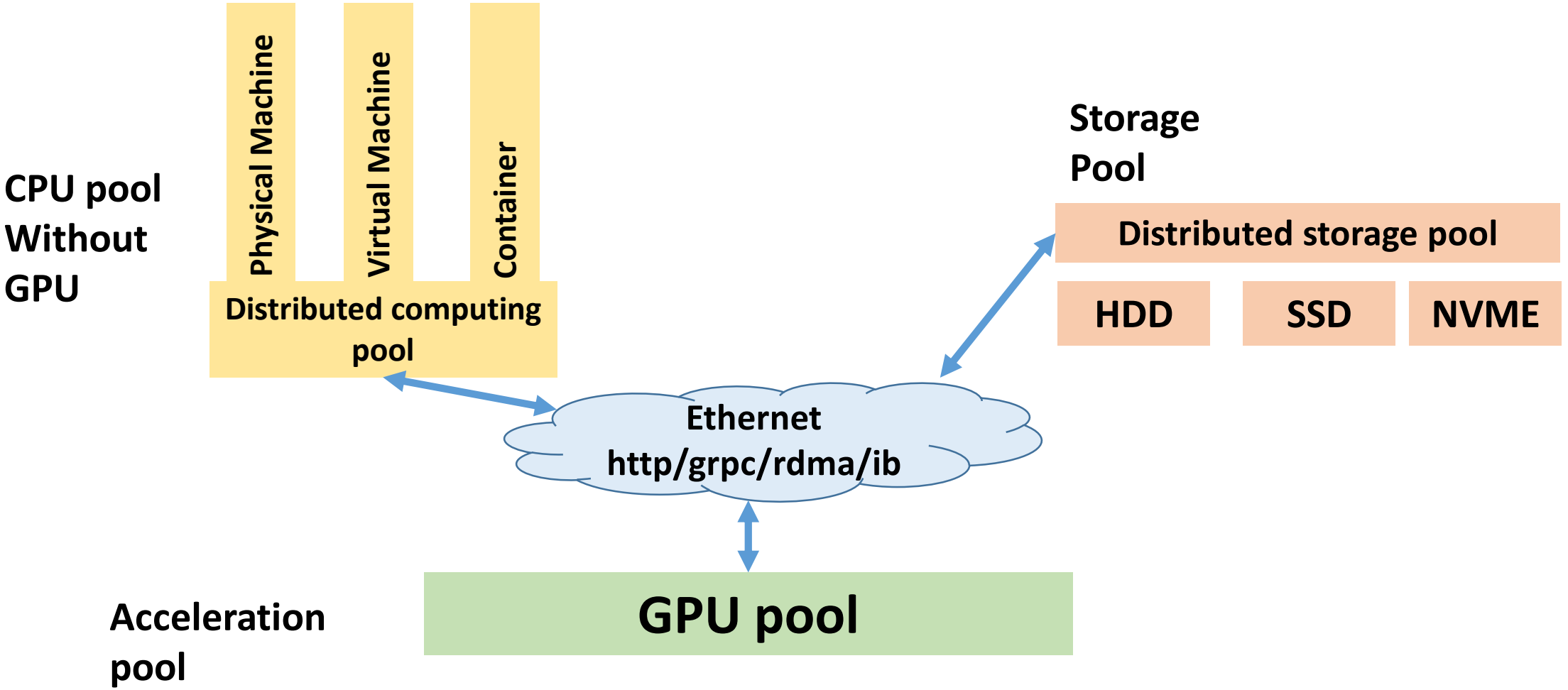
- Single GPU card and GPU server is becoming so powerful, cannot be fully utilized

Distributed storage system-big success



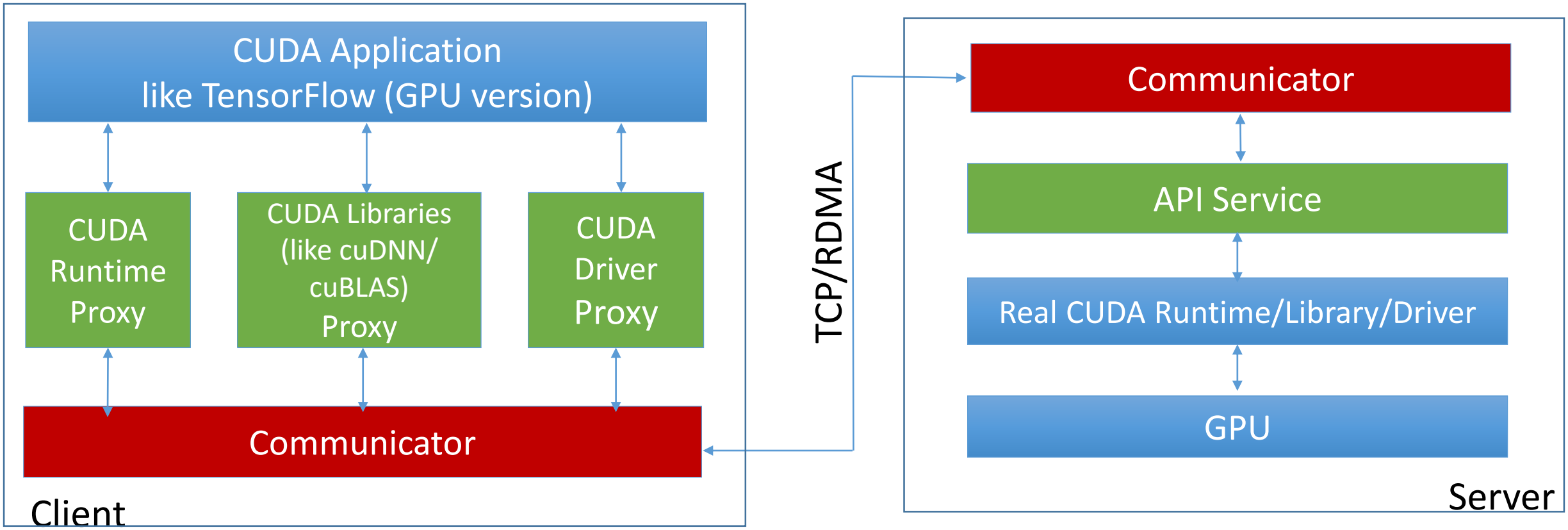
- seperated CPU(client) and storage(chunk server)
- Stable
- High performance
- Easy to maintain
- Low cost

distributed acceleration



GPU pool over ethernet

Alibaba Cloud
Worldwide Cloud Services Partner



Components can be generated automatically with minimum manual work

Components can be shared by GPU/FPGA/etc.

Challenge 1-A lot of CUDA API

Alibaba Cloud
Worldwide Cloud Services Partner

- A intermediate and auto-generated YAML file to define the APIs that need to be forwarded to remote server.
- Need to add some information on pointer arguments manually

```
874   -
875     return_type: "cudaError_t"
876     name: cudaGetDeviceProperties
877     args:
878       -
879         type: "struct cudaDeviceProp *"
880         name: prop
881         is_usr_buf: 1
882         usr_buf_io: out
883         usr_buf_size: sizeof(struct cudaDeviceProp)
884       -
885         type: "int"
886         name: device
```

Challenge 2-Host memory

Alibaba Cloud
Worldwide Cloud Services Partner

- With remote GPU, we need to detect such memories in CUDA kernel parameters to make sure the function correctness
  - cudaMallocHost

Challenge 3 - Hide API forwarding Latency

Alibaba Cloud
Worldwide Cloud Services Partner

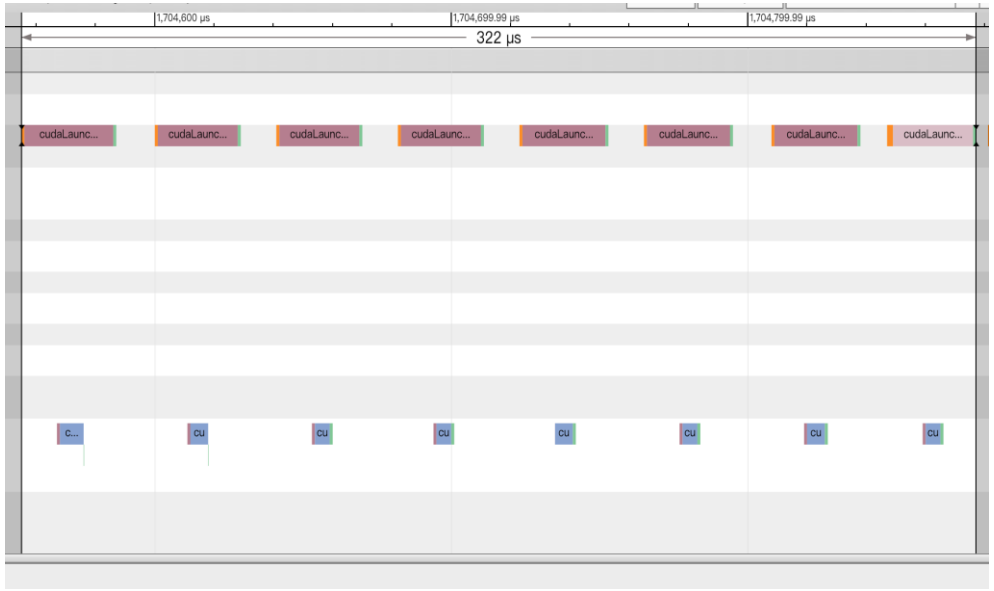
- Changing blocking API calls to asynchronous calls
- Benefit: 2x speedup

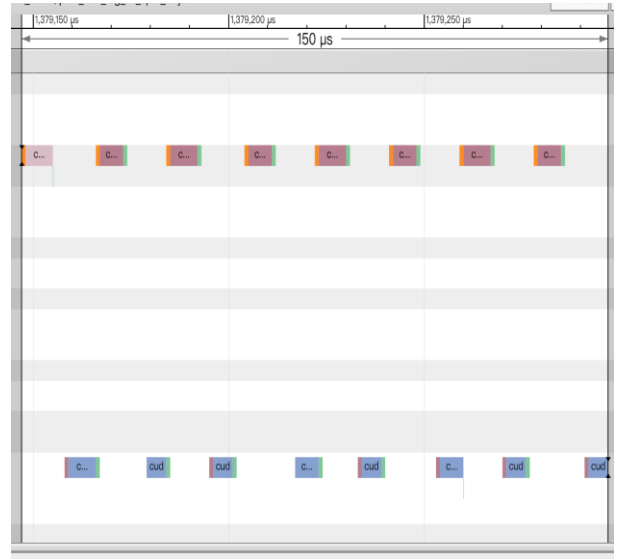
Before

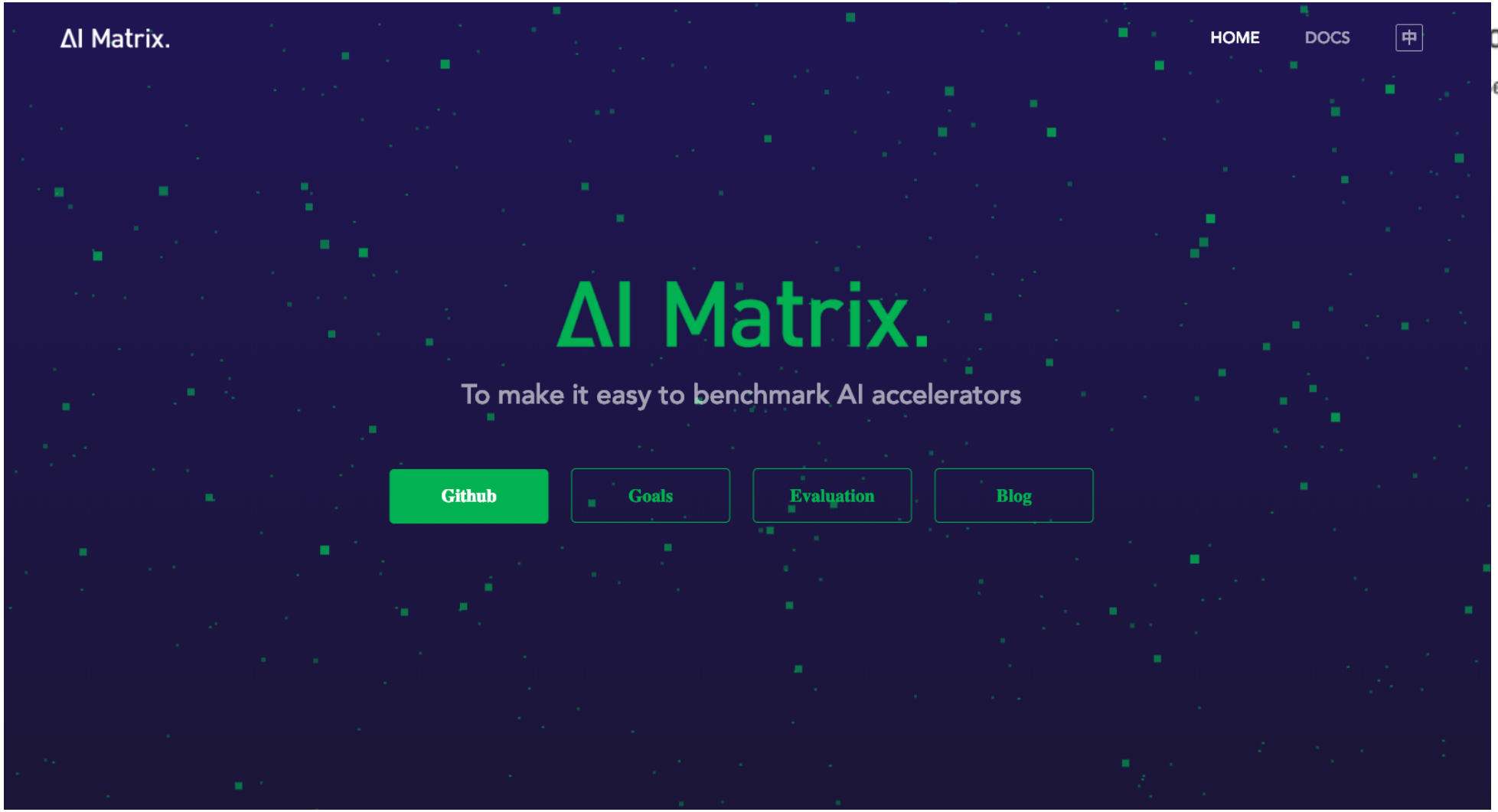
After

API run time in client

API run time in server



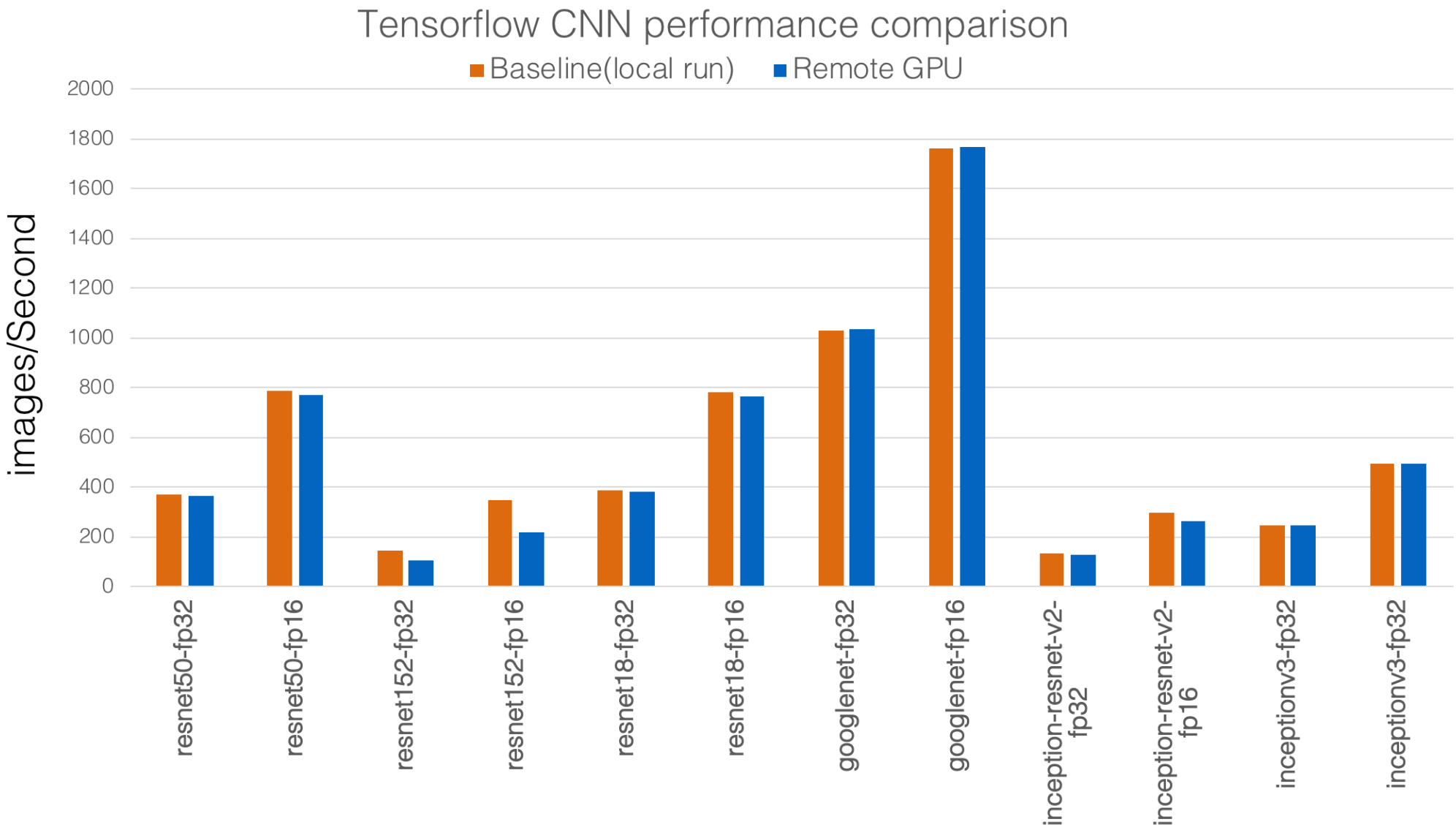




- website：https://aimatrix.ai
- Code repo：https://github.com/alibaba/ai-matrix

CNN Training Performance

Alibaba Cloud
Worldwide Cloud Services Partner



Summary

Alibaba Cloud
Worldwide Cloud Services Partner

- GPU accelerated inference service pool: multi-tenant
- GPU resource pool over ethernet
- All results are preliminary and we are actively working on it
- Discussions and collaborations are welcome
- Email: l.jin@Alibaba-inc.com

Alibaba Cloud
Worldwide Cloud Services Partner

Thanks!
Q&A