



<https://hao-ai-lab.github.io/dsc204a-f25/>

DSC 204A: Scalable Data Systems

Fall 2025

Staff

Instructor: Hao Zhang

TAs: Mingjia Huo, Yuxuan Zhang



[@haozhangml](https://twitter.com/haozhangml)



[@haoailab](https://twitter.com/haoailab)



haozhang@ucsd.edu

Where We Are

Machine Learning Systems

Big Data

Cloud

Foundations of Data Systems

2000 - 2016

1980 - 2000



Logistics

- BoQ Survey completion rate: 107%
 - All of you will get 1 bonus point
- Overall Feedback
 - Want more advanced stuff, less basics
 - Want to learn ML/LLM Systems (Many expect $\geq 30\%$ content on it)
 - Want slower pace on content
- Waitlisted Students:
 - Please get in touch with me/TA to get enrolled (there should be room now)

Evolution of Cloud Infrastructure

- Data Center: Physical space from which a cloud is operated
- 3 generations of data centers/clouds:
 - Cloud 1.0 (Past)
 - Cloud 2.0 (Current)
 - Cloud 3.0 (Ongoing Research)

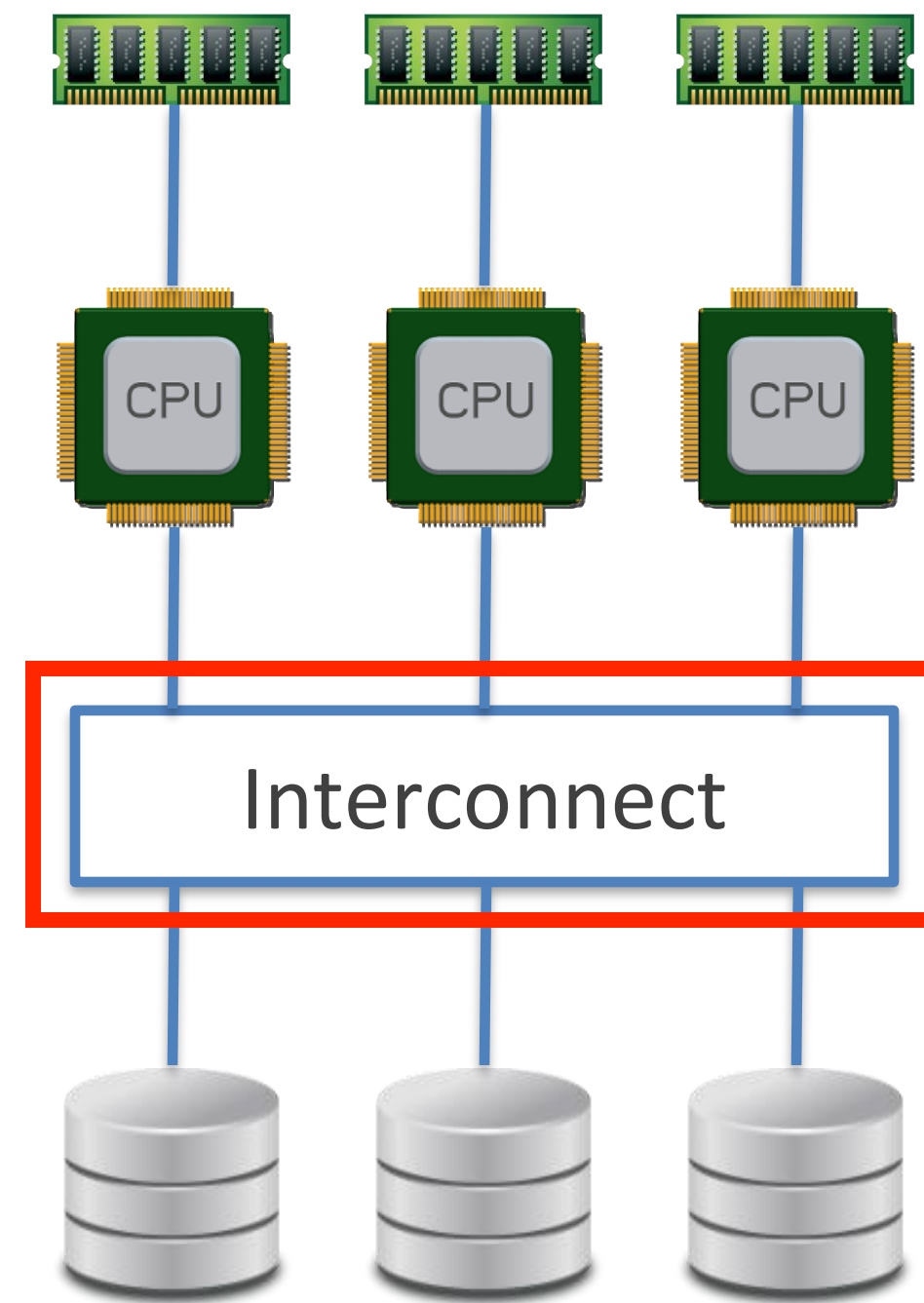
Cloud 1.0 (Past)

- Networked servers;
- User rents servers (time-sliced access) needed for data/software

Cloud 2.0 (Current)

- “Virtualization” of networked servers;
- User rents amount of resource capacity (e.g., memory, disk);
- Cloud provider has a lot more flexibility on provisioning (multi-tenancy, load balancing, more elasticity, etc.)

Parallelism in the Cloud

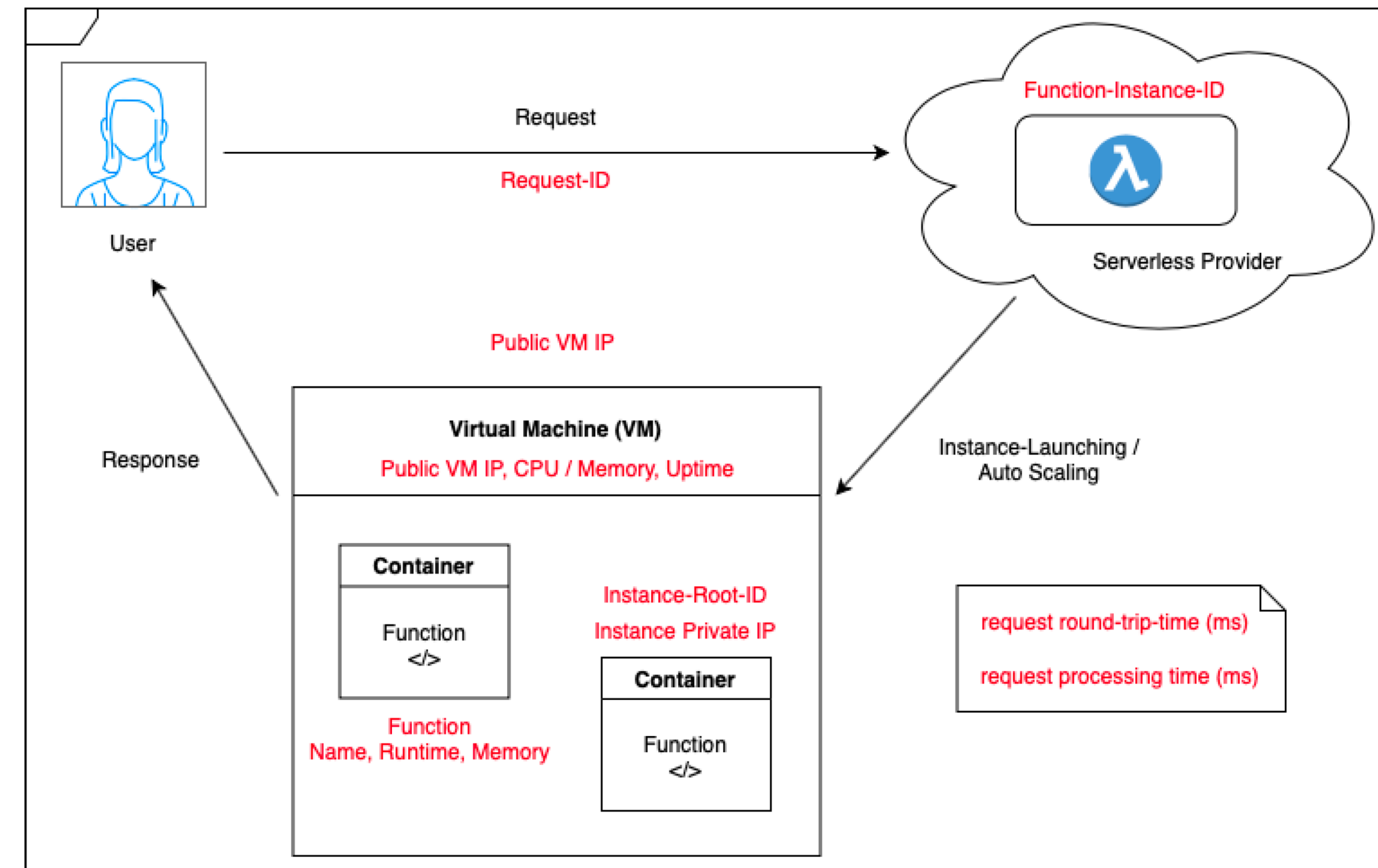


Modern networks in data centers have become much faster: 100GbE to even TbE!

- Decoupling of compute+memory from storage is common in cloud
 - *Hybrids* of shared-disk parallelism + shared-nothing parallelism
 - E.g, store datasets on S3 and read as needed to local EBS

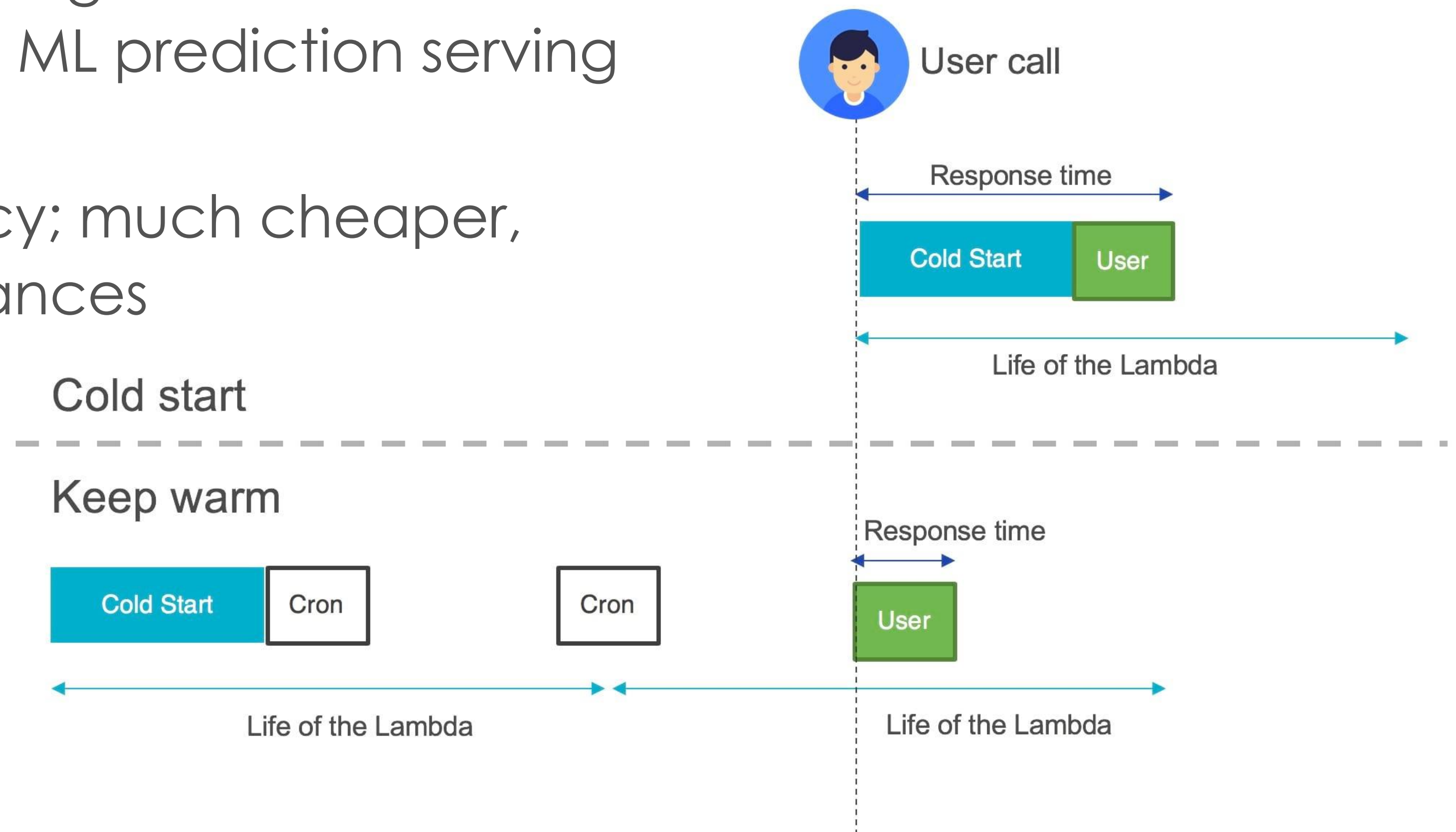
Cloud 3.0 (Ongoing Research)

- Full resource disaggregation! That is, compute, memory, storage, etc. are all network-attached and elastically added/removed
- User gives a program (function) to run and specifies CPU and DRAM needed
- Cloud provider abstracts away all resource provisioning entirely
- Aka *Function-as-a-Service* (FaaS)



Cloud 3.0 (Ongoing Research)

- “Serverless” and disaggregated resources all connected to fast networks
- Serverless paradigm gaining traction for some applications, e.g., online ML prediction serving on websites
- Higher resource efficiency; much cheaper, often by 10x vs Spot instances



Advantage and disadvantage

- Cloud 1.0:
 - +: Simple, Perfect isolation,
 - - : Expensive.
- Cloud 2.0:
 - +: Cheaper than Cloud 1.0.
 - - : Some resource waste
- Cloud 3.0:
 - +: Cheapest
 - - : Cold-start issues, Security & Privacy, Hard to manage.

New Trends in the Deep Learning Era

- New trends:
 - Reservation takes precedence than on-demand/spot
 - GPU vertical cloud
 - Community cloud

LAST UPDATED: 1 WEEK AGO

Provider	A100 40GB	A100 80GB	H100
Lambda Labs	Unavailable	Unavailable	Unavailable
FluidStack	Unavailable	Unavailable	Unavailable
Runpod	Unavailable	\$1.59/hour	\$3.39/hour

