

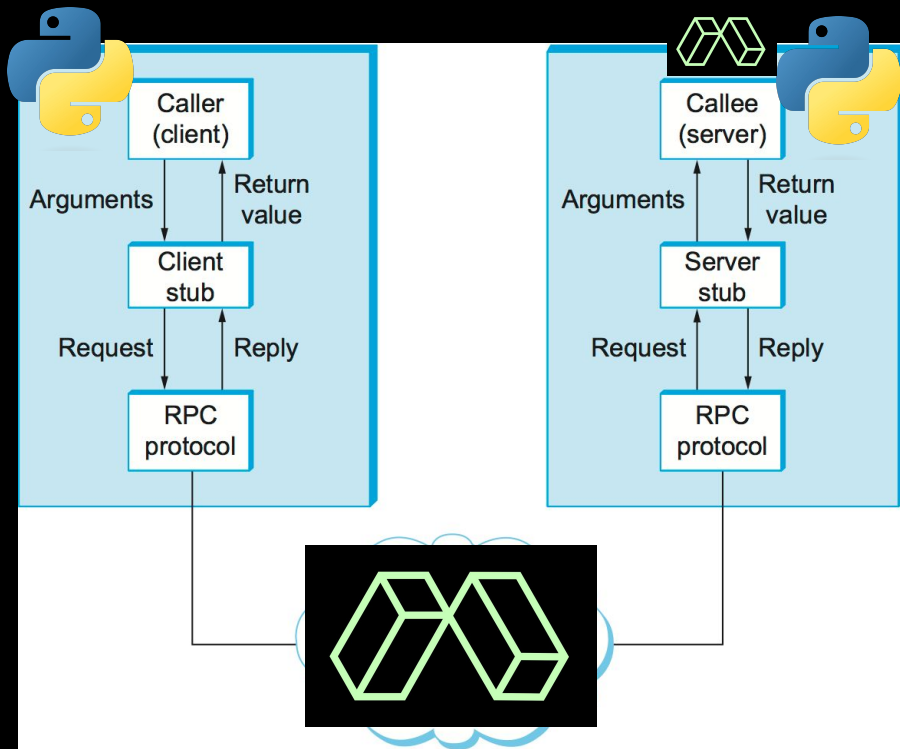


Serverless  
ML Infra  
That Does  
Not Suck

# Demo first!

# An old idea revisited: RPC.

Your code  
on your machine



Your code  
on our machine

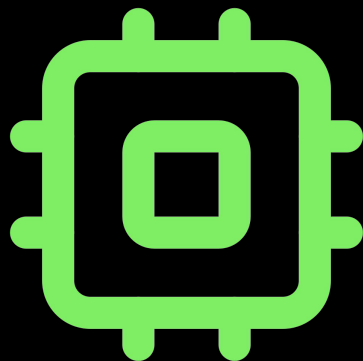
[book.systemsapproach.org/e2e/rpc.html](http://book.systemsapproach.org/e2e/rpc.html)

Demo again!



Store

Preserve  
information



Compute

Manipulate  
information



I/O

Connect the world  
to information



## Store

Dictionaries, Queues,  
Volumes, Mounts



## Compute

Functions, Crons,  
GPU acceleration



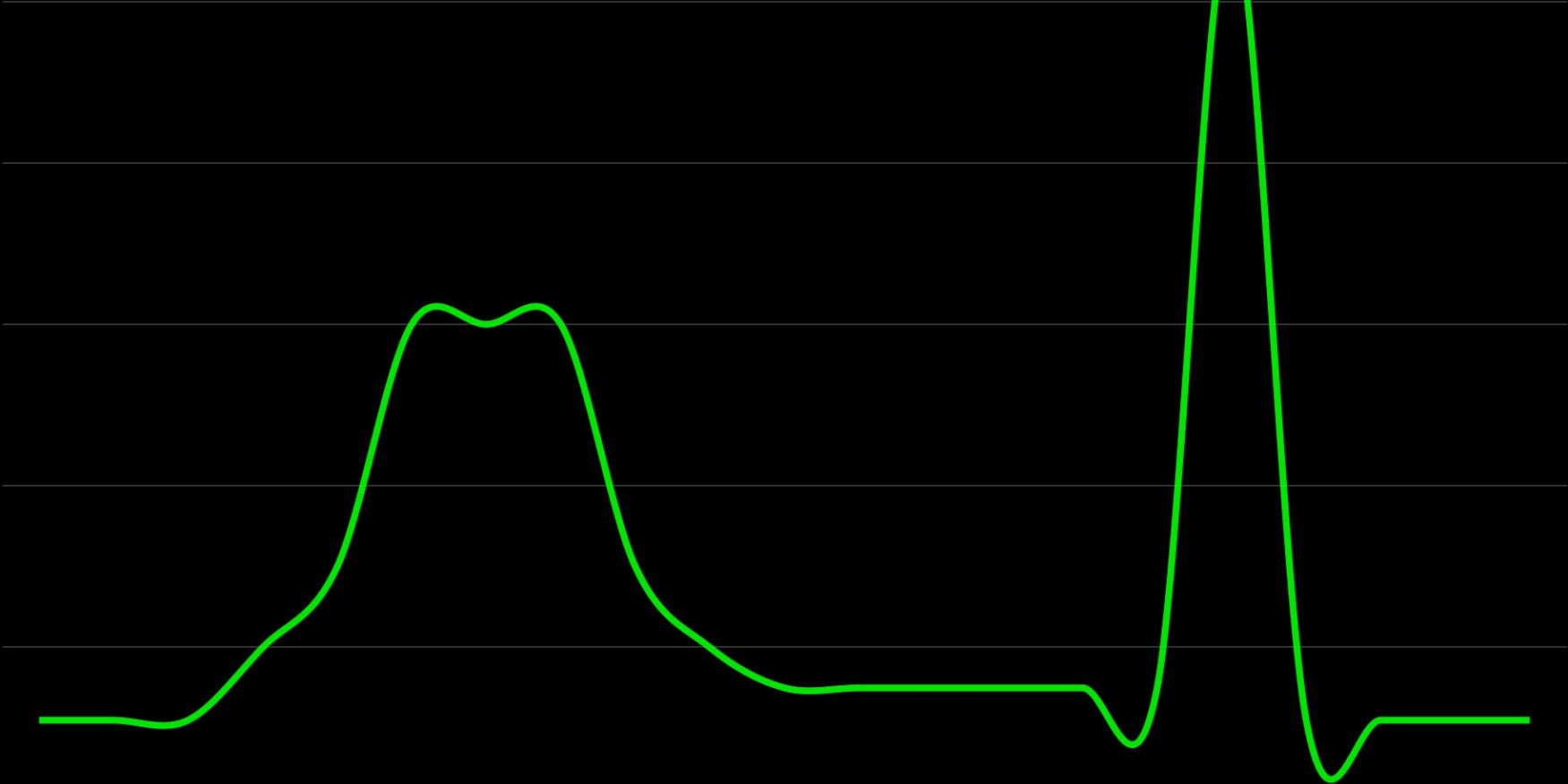
## I/O

Web endpoints,  
web servers

# Serverless Computing

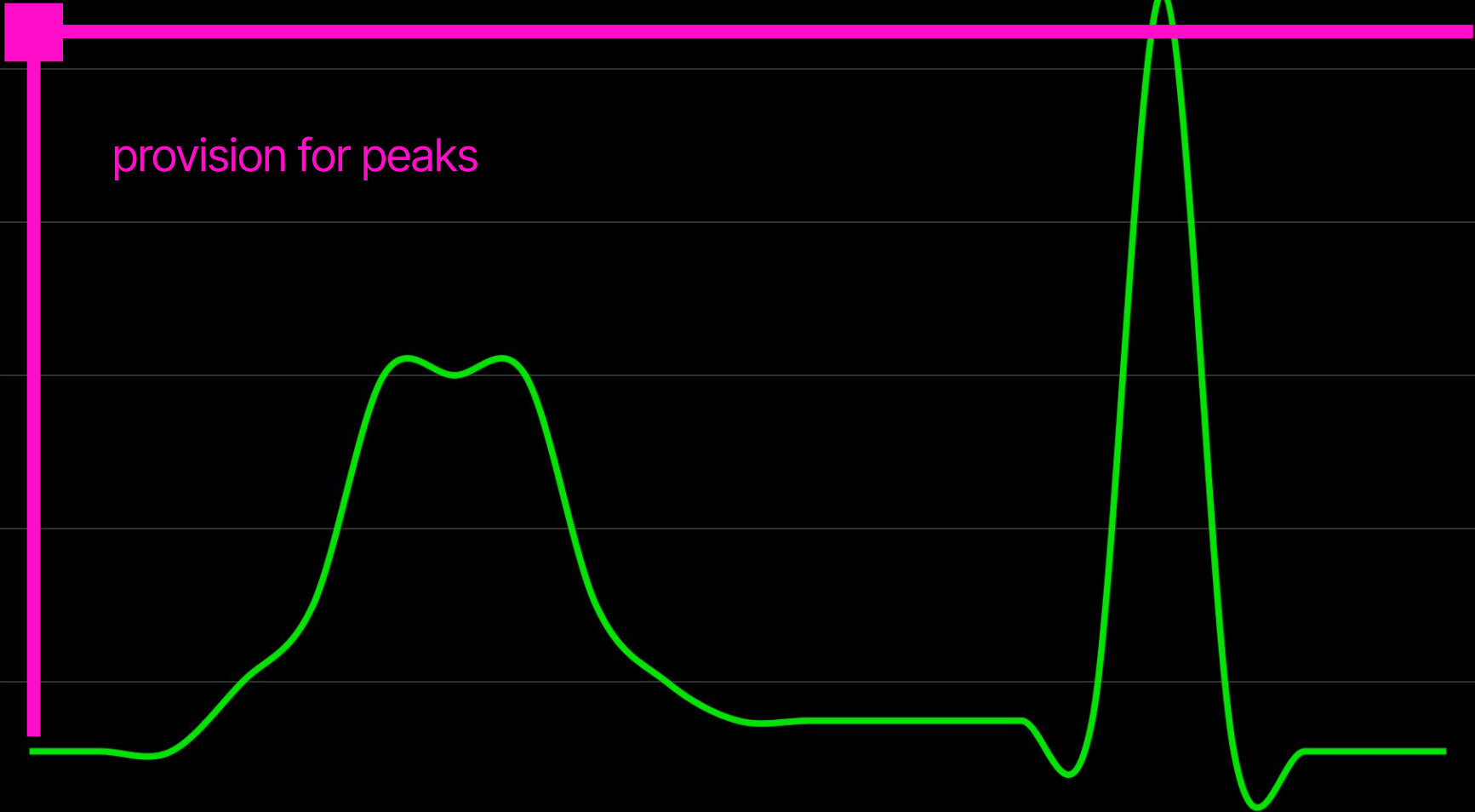
a cloud-computing paradigm,  
not just a marketing term

# Resource Utilization Over Time



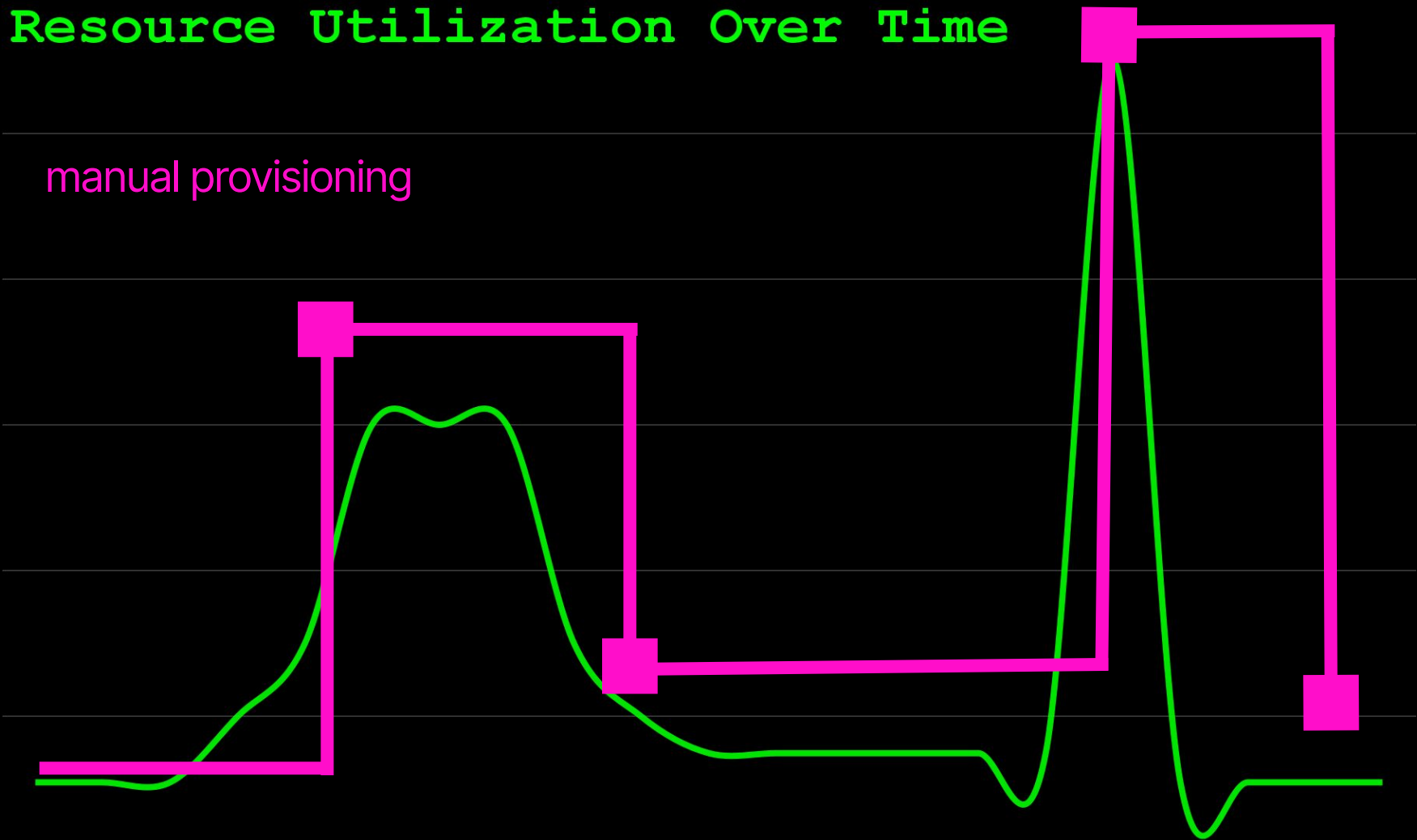


# Resource Utilization Over Time



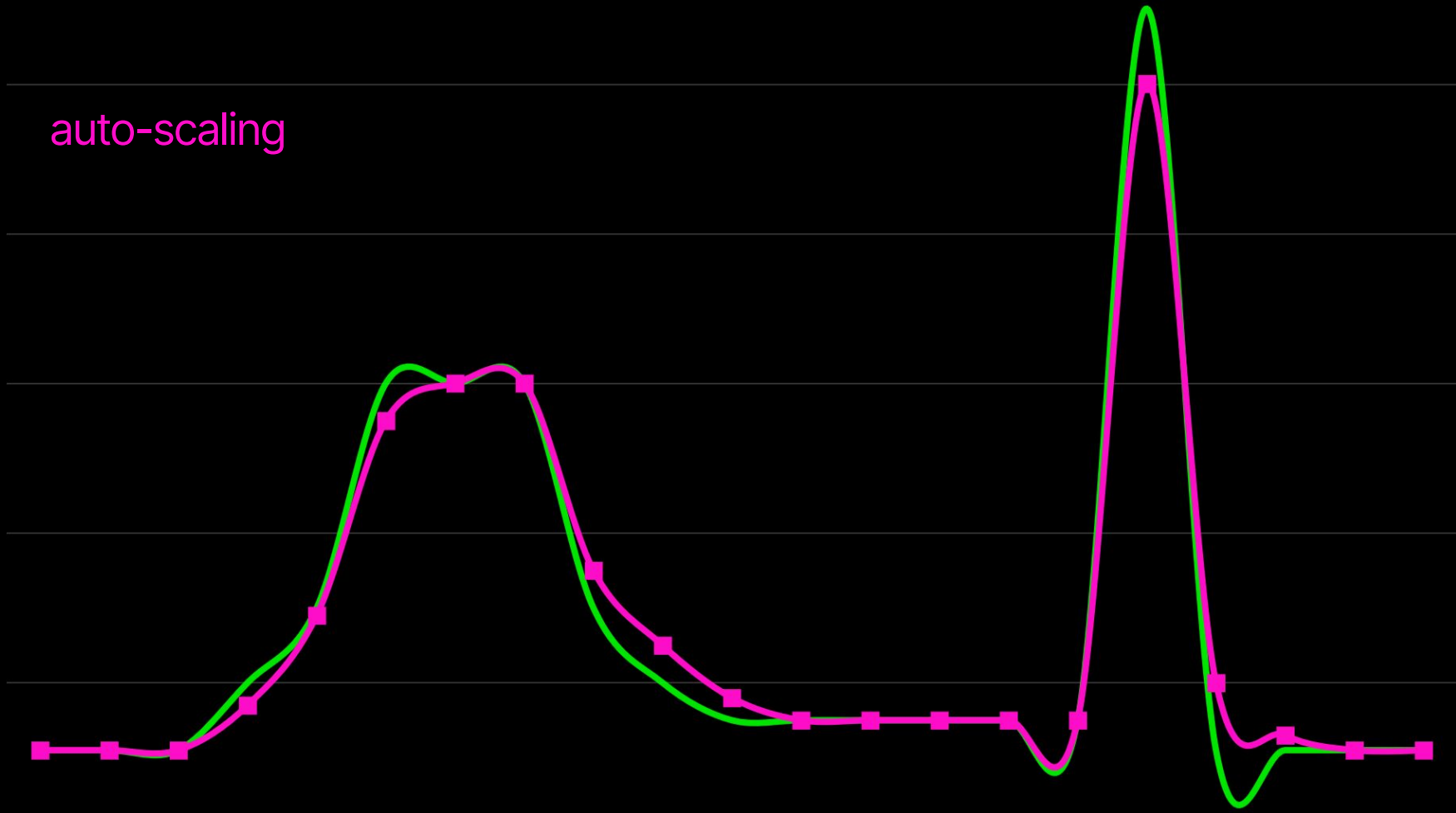
# Resource Utilization Over Time

manual provisioning



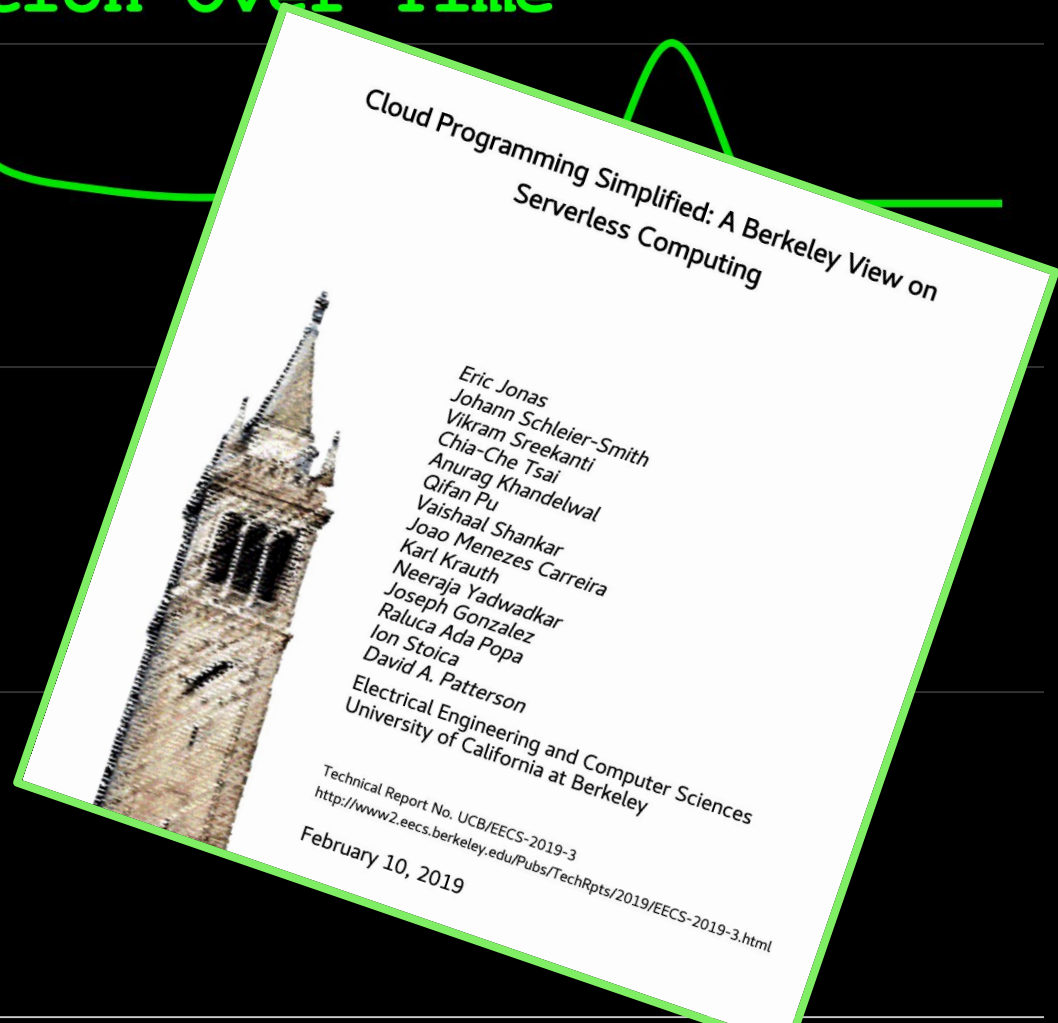
# Resource Utilization Over Time

auto-scaling



# Resource Utilization Over Time

demand aggregated by  
serverless platform



# Why not just Google/AWS/Azure/Oracle?

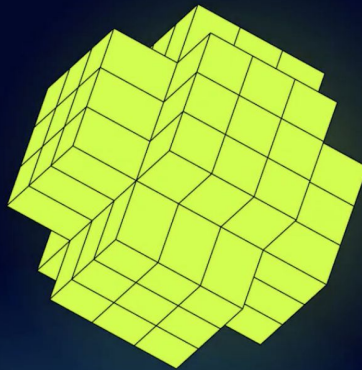
Serverless for web/mobile applications has very different technical requirements than serverless for data/model applications

- Image sizes are larger (many GiB, not shared between tenants)
- Input and/or output sizes are larger (many MiB to GiB)
- Compute requirements are higher (many cores, GPUs)
- Runtimes are longer (worry about HTTP timeouts)

## Creating our Own Kubernetes & Docker to Run Our Data Infrastructure

Erik Bernhardsson  
Founder, Modal Labs

<https://www.youtube.com/watch?v=3jJ1GhGkLY0>



<https://www.latent.space/p/modal>

## Lambda on hard mode: Inside Modal's web infrastructure




**Eric Zhang** @ekzhang1

Founding Engineer

<https://modal.com/blog/serverless-http>

# More demos!

 Flux Fine-Tune of Your Pet?

 OpenAI-Compatible vLLM Service?

 Blender Render Farm?



<https://modal.com>