

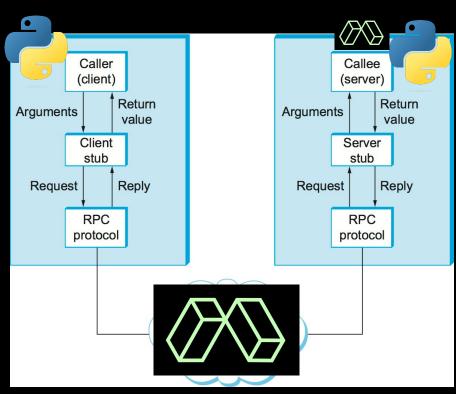
Serverless ML Infra That Does Not Suck

Demo first!



An old idea revisited: RPC.

Your code on your machine



Your code on our machine

book.systemsapproach.org/e2e/rpc.html



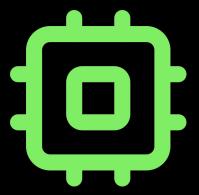
Demo again!







Preserve information



Compute

Manipulate information



I/O

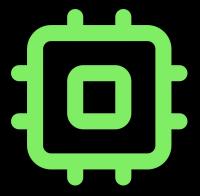
Connect the world to information







Dictionaries, Queues, Volumes, Mounts



Compute

Functions, Crons, GPU acceleration



I/O

Web endpoints, web servers



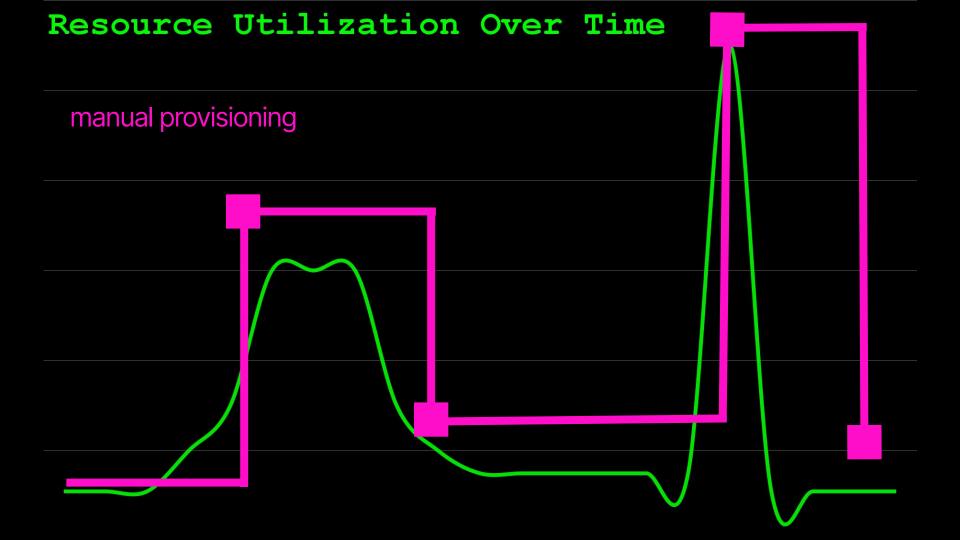
Serverless Computing

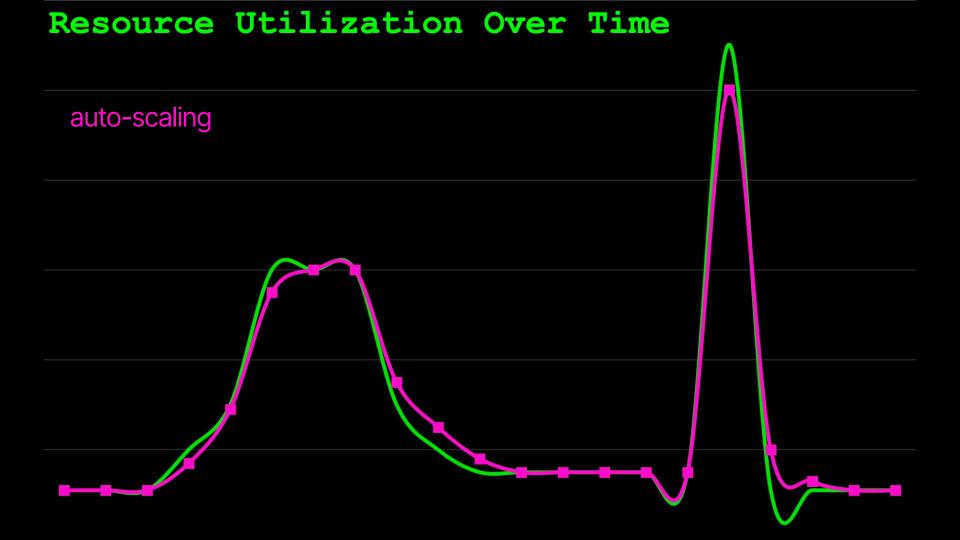
a cloud-computing paradigm, not just a marketing term





Resource Utilization Over Time provision for peaks





Resource Utilization Over Time

demand aggregated by serverless platform

Cloud Programming Simplified: A Berkeley View on

Serverless Computing

Johann Schleier-Smith Vikram Sreekanti Chia-Che Tsai Anurag Khandelwal Qifan Pu Vaishaal Shankar Joan Menezes Carreira Karl Krauth Neeraja Yadwadkar Joseph Gonzalez Raluca Ada Popa Ion Stoica David A. Patterson Electrical Engineering and Computer Sciences Technical Report No. UCB/EECS-2019-3 Technical Report No. UCB/EECS-ZU19-3
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2019/EECS-2019-3.html February 10, 2019

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)

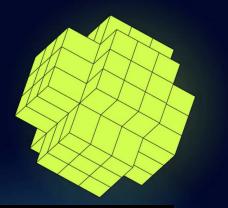




Creating our Own Kubernetes & Docker to Run Our Data Infrastructure

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-htt



More demos!

Flux Fine-Tune of Your Pet?

OpenAl-Compatible vLLM Service?

Blender Render Farm?





https://modal.com