XFaaS

HYPERSCALE AND LOW COST SERVERLESS FUNCTIONS AT META

Alireza Sahraei[1]	Soteris Demetriou[2]	Amirali Sobhgol[1]	Haoran Zhang[3]	Abhigna Nagaraja[1]	Neeraj Pathak[1]	Girish Joshi[1]	Carla Souza[1]	Bo Huang[1]	Wyatt Cook[1]
Andrii Golovei[1]	Pradeep Venkat[1]	Andrew McFague[1]	Dimitrios Skarlatos[4]	Vipul Patel[1]	Ravinder Thind[1]	Ernesto Gonzalez[1]	Yun Jin[1]	Chunqiang Tang[1]	

[1] Meta

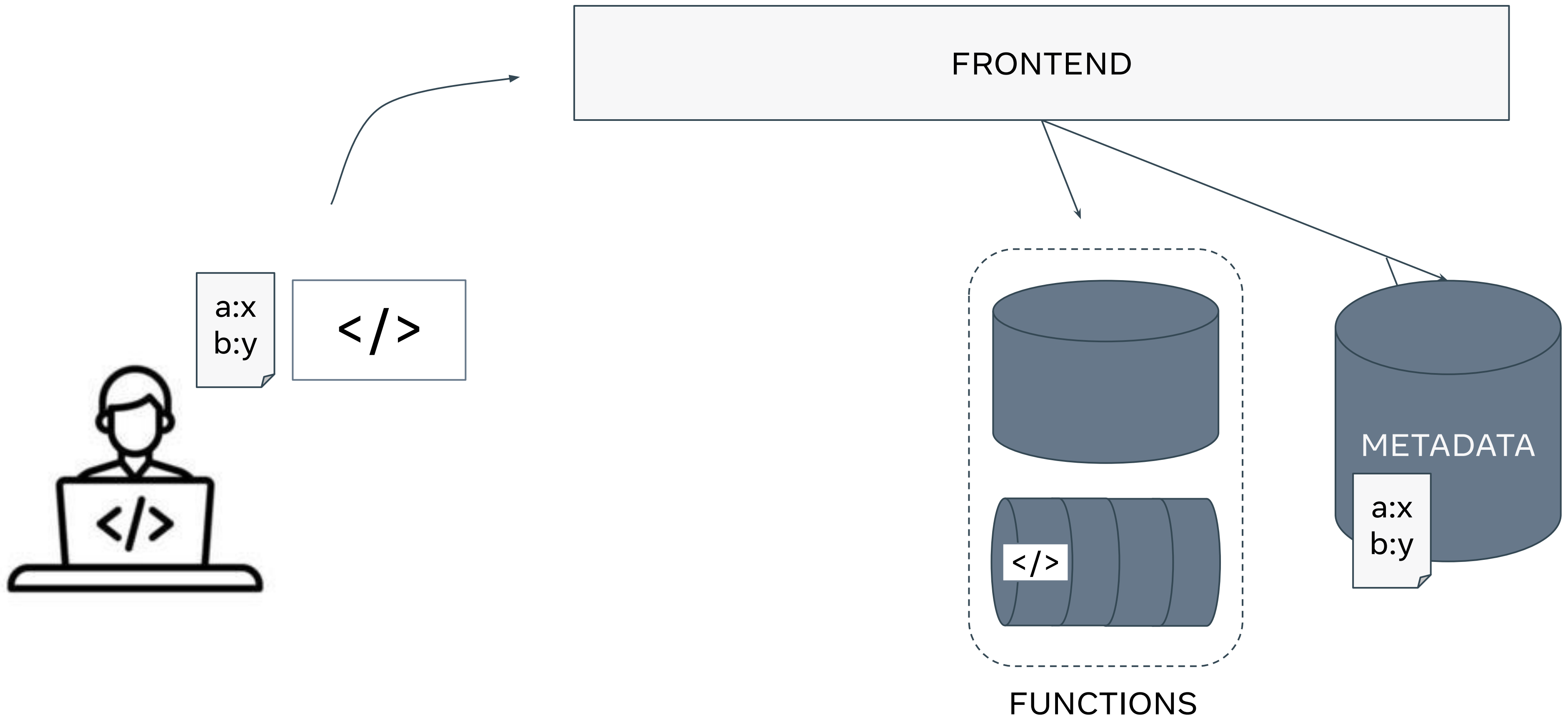
[2] Imperial College London

[3] Penn University of Pennsylvania

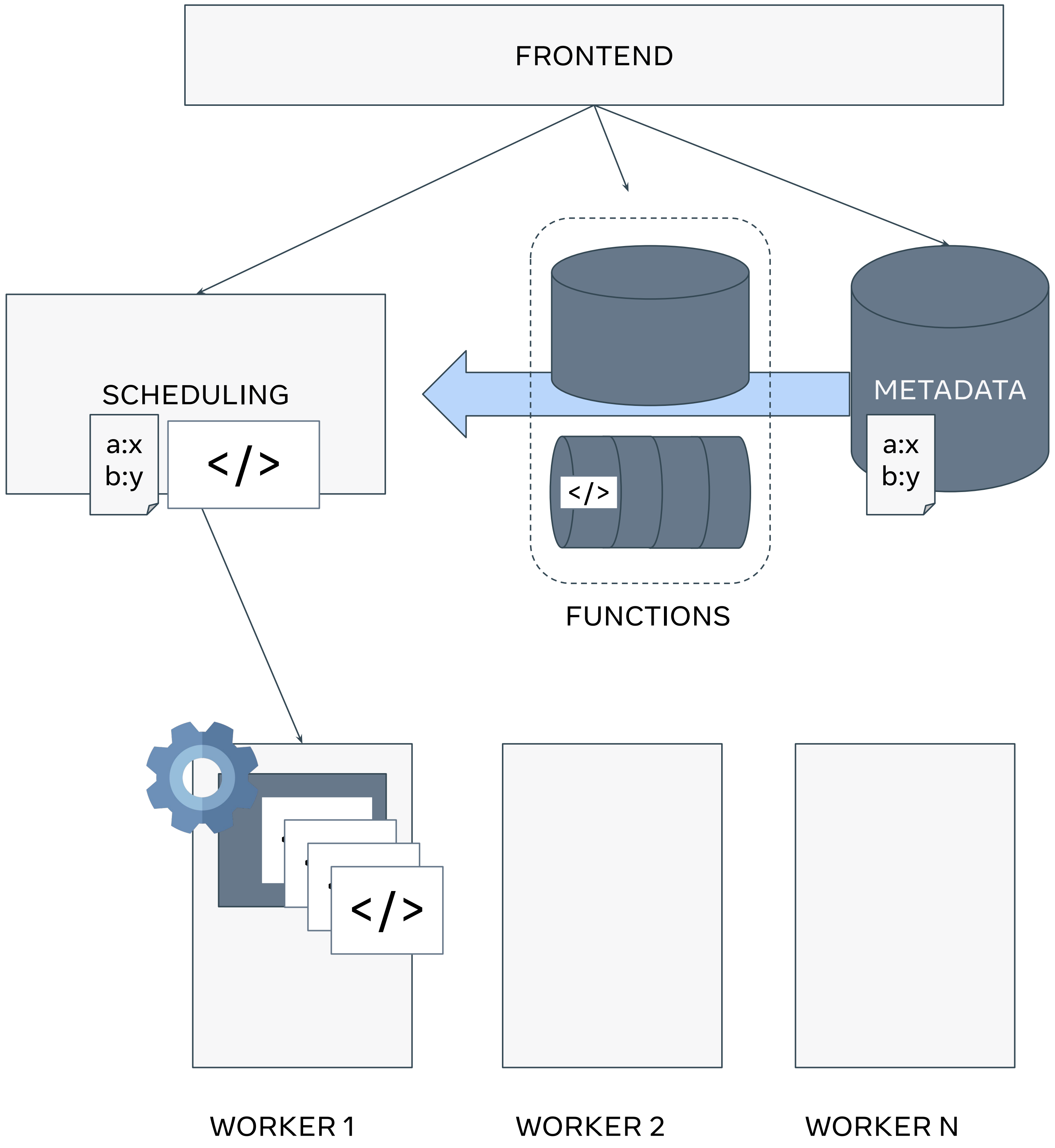
[4] Carnegie Mellon University Computer Science Department

01 BACKGROUND & MOTIVATION

FUNCTION AS A SERVICE



FUNCTION AS A SERVICE



FUNCTION AS A SERVICE

PUBLIC

aws

Brooker, Marc, et al. "On-demand Container Loading in {AWS} Lambda." 2023 USENIX Annual Technical Conference (USENIX ATC 23). 2023

Agache, Alexandru, et al. "Firecracker: Lightweight virtualization for serverless applications." 17th USENIX symposium on networked systems design and implementation (NSDI 20). 2020

Azure Functions

Shahrad et al. Serverless in the wild: Characterizing and optimizing the serverless workload at a large cloud provider. In USENIX Annual Technical Conference, 2020.

Alibaba Cloud

Wang, Ao, et al. "{FaaSNet}: Scalable and fast provisioning of custom serverless container runtimes at alibaba cloud function compute." 2021 USENIX Annual Technical Conference (USENIX ATC 21). 2021

Google Cloud Functions

PRIVATE

Meta

This work...

FUNCTION AS A SERVICE AT META

- highly heterogeneous workloads

Workload	Trigger	Calls/ second	CPU (MIPS)	Execution Time (s)	Memory (MB)
Notifications	Data Warehouse	3.4M	65-200	0.55 - 1.1	10 - 90
Morphing Framework	Queue	25K	1.5M - 27M	65 - 155	30 - 230

WHAT ABOUT HARDWARE COSTS?

> *“81% of the applications are invoked once per minute or less on average. This suggests that the cost of keeping these applications warm, relative to their total execution (billable) time, can be prohibitively high”*

Shahrad et al. Serverless in the wild: Characterizing and optimizing the serverless workload at a large cloud provider. In USENIX Annual Technical Conference, 2020.

02 CHALLENGES

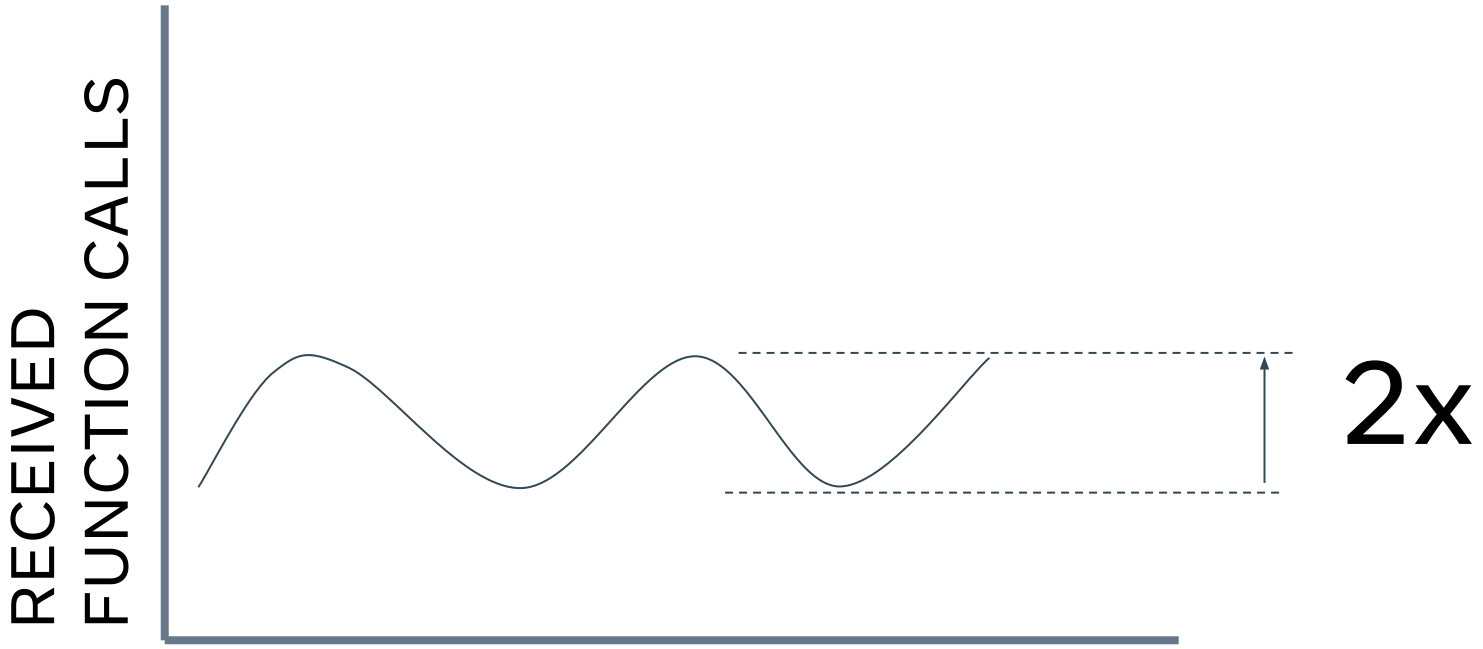
- A Lengthy Cold Start [NOT COVERED IN THIS TALK]
- B High Variance of Load
- C Downstream Overloads

High Variance of Load

Problem

1. Previous work reported a high peak-to-trough ratio of function calls
2. At Meta, the ratio can be as high as 4.3

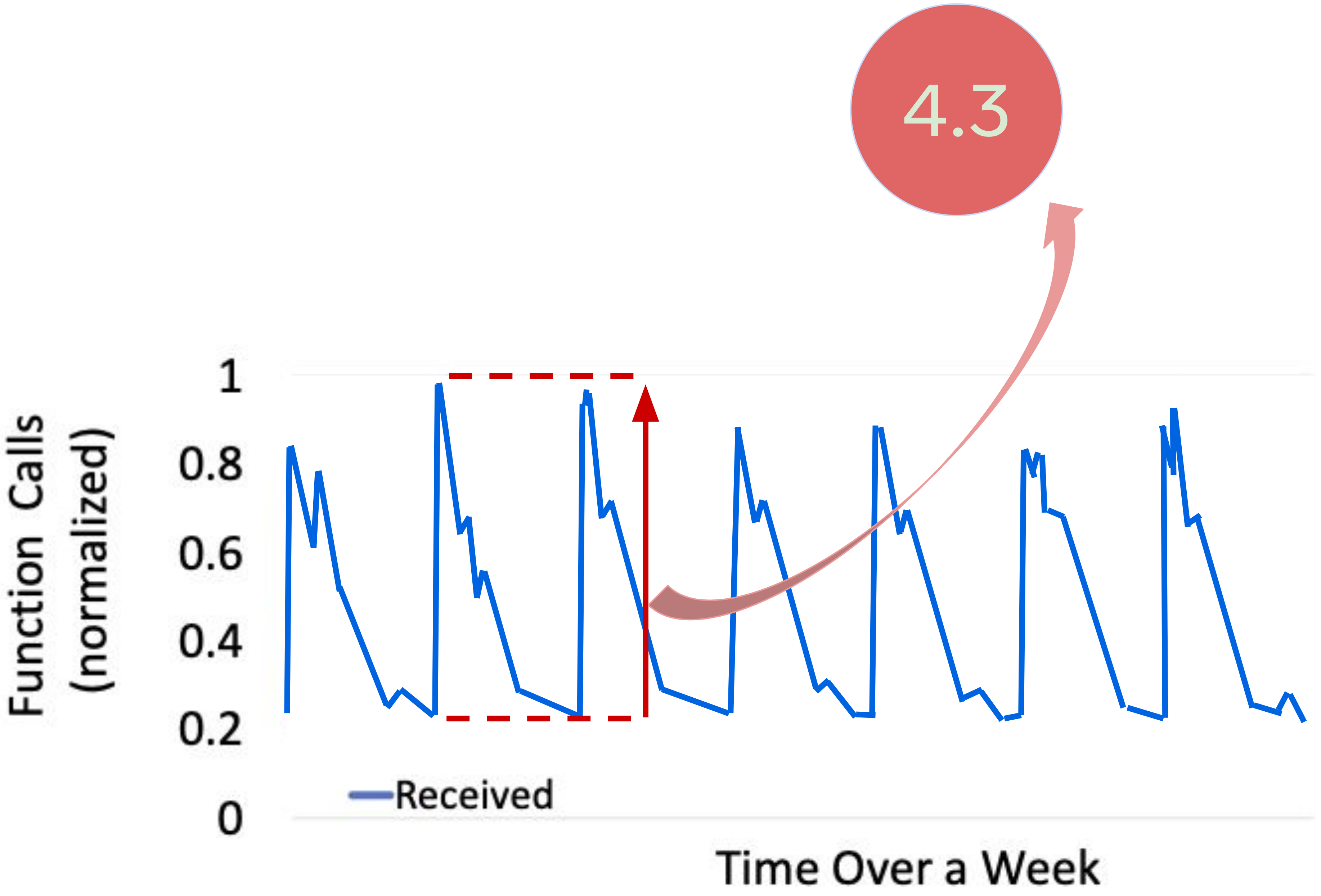
Shahrad et al. Serverless in the wild: Characterizing and optimizing the serverless workload at a large cloud provider. In USENIX Annual Technical Conference (USENIX ATC 20). 2020.



High Variance of Load

Problem

1. Previous work reported a high peak-to-trough ratio of function calls
2. At Meta, the ratio can be as high as 4.3

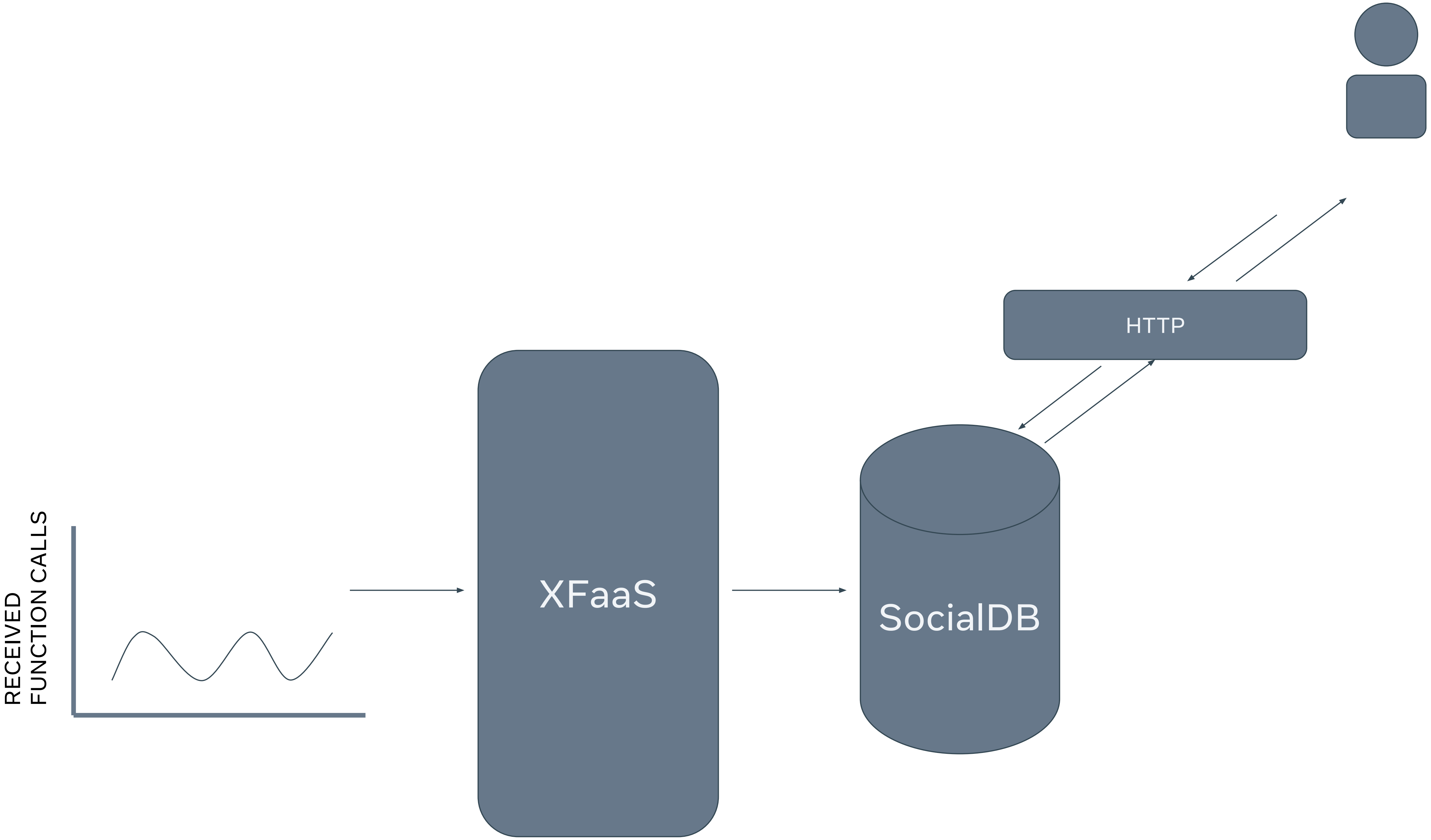


02 CHALLENGES: OVERLOADING DOWNSTREAM SERVICES

DOWNSTREAM OVERLOADS

Problem

SocialDB Outage

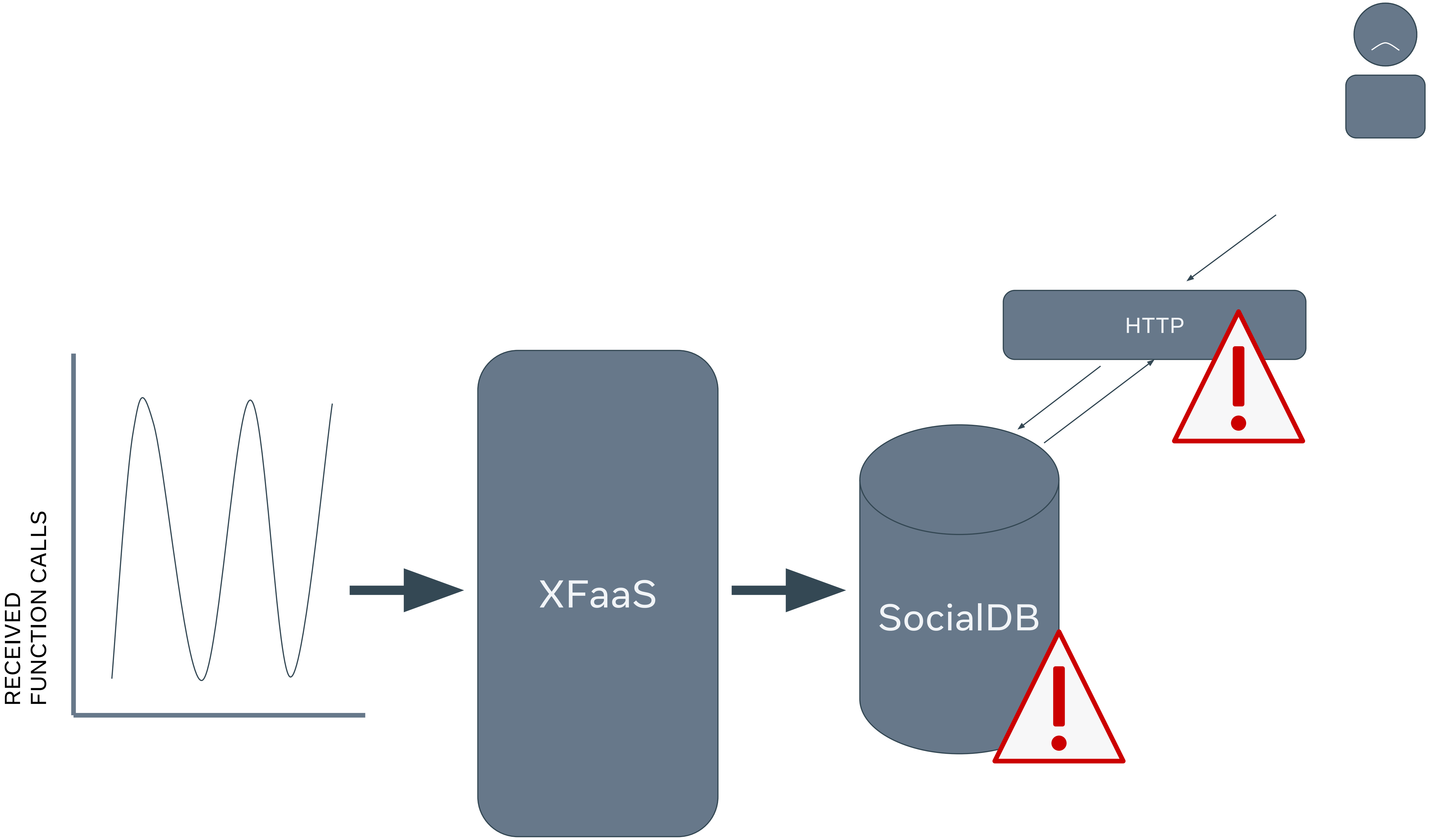


DOWNSTREAM OVERLOADS

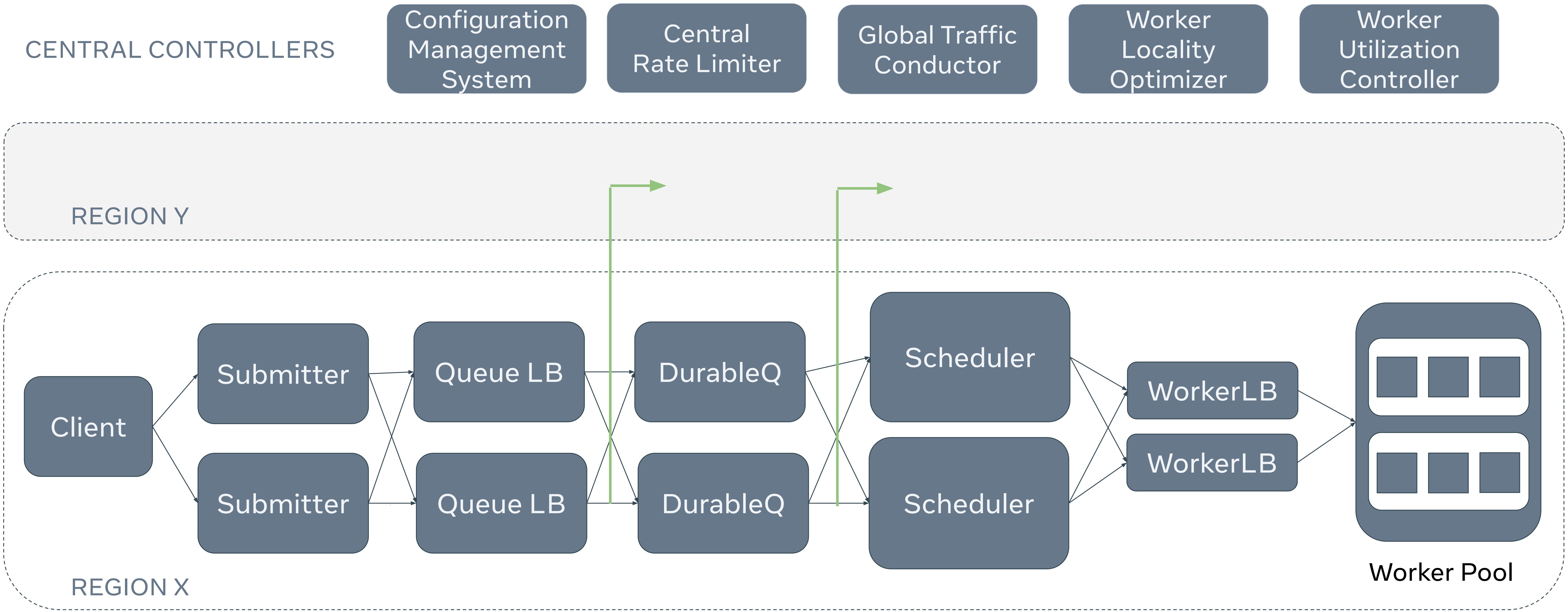
Problem

SocialDB Outage

- manual resolution
- several hours to resolve
- coarse-grained



03 SYSTEM OVERVIEW


CENTRAL CONTROLLERS
Configuration Management System
Central Rate Limiter
Global Traffic Conductor
Worker Locality Optimizer
Worker Utilization Controller
REGION Y
Client
Submitter
Submitter
Queue LB
Queue LB
DurableQ
DurableQ
Scheduler
Scheduler
WorkerLB
WorkerLB
Worker Pool
REGION X

Next…

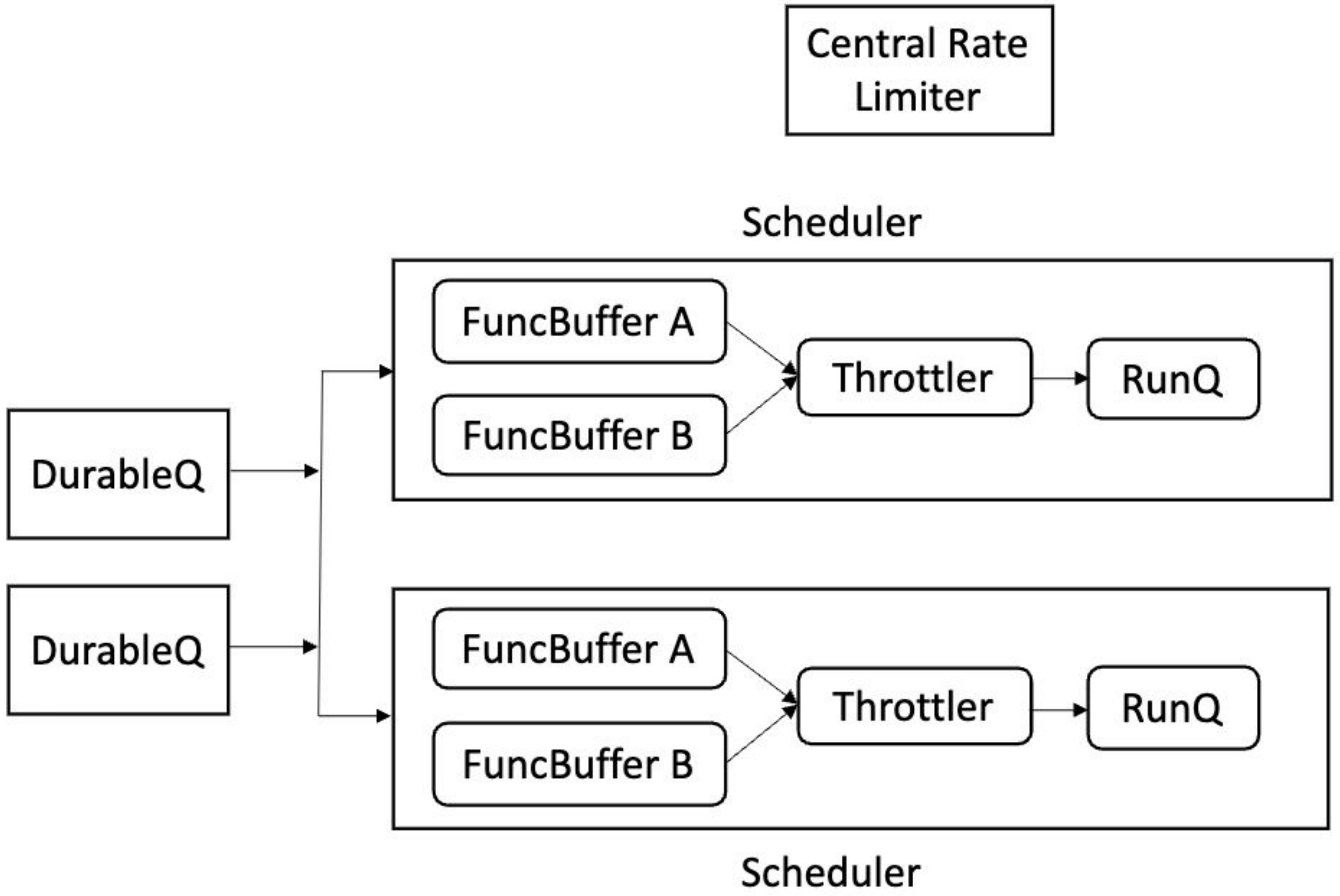
04 DEFERRED COMPUTE – DESIGN & EVALUATION

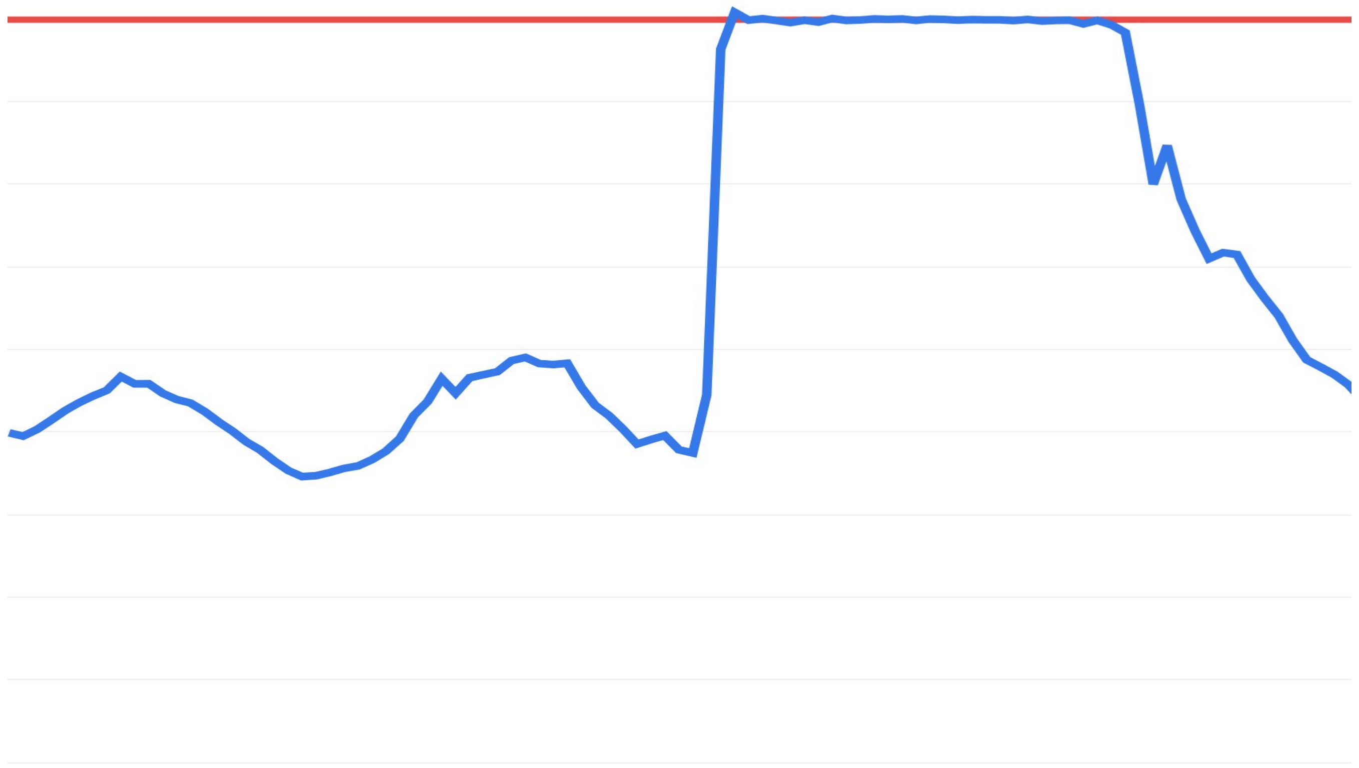
05 DOWNSTREAM PROTECTION – DESIGN & EVALUATION

04 DEFERRED COMPUTE - DESIGN & EVALUATION

1. Reserved Quota
   - CPU cycles a function can consume
   - Transformed to RPS for enforcement

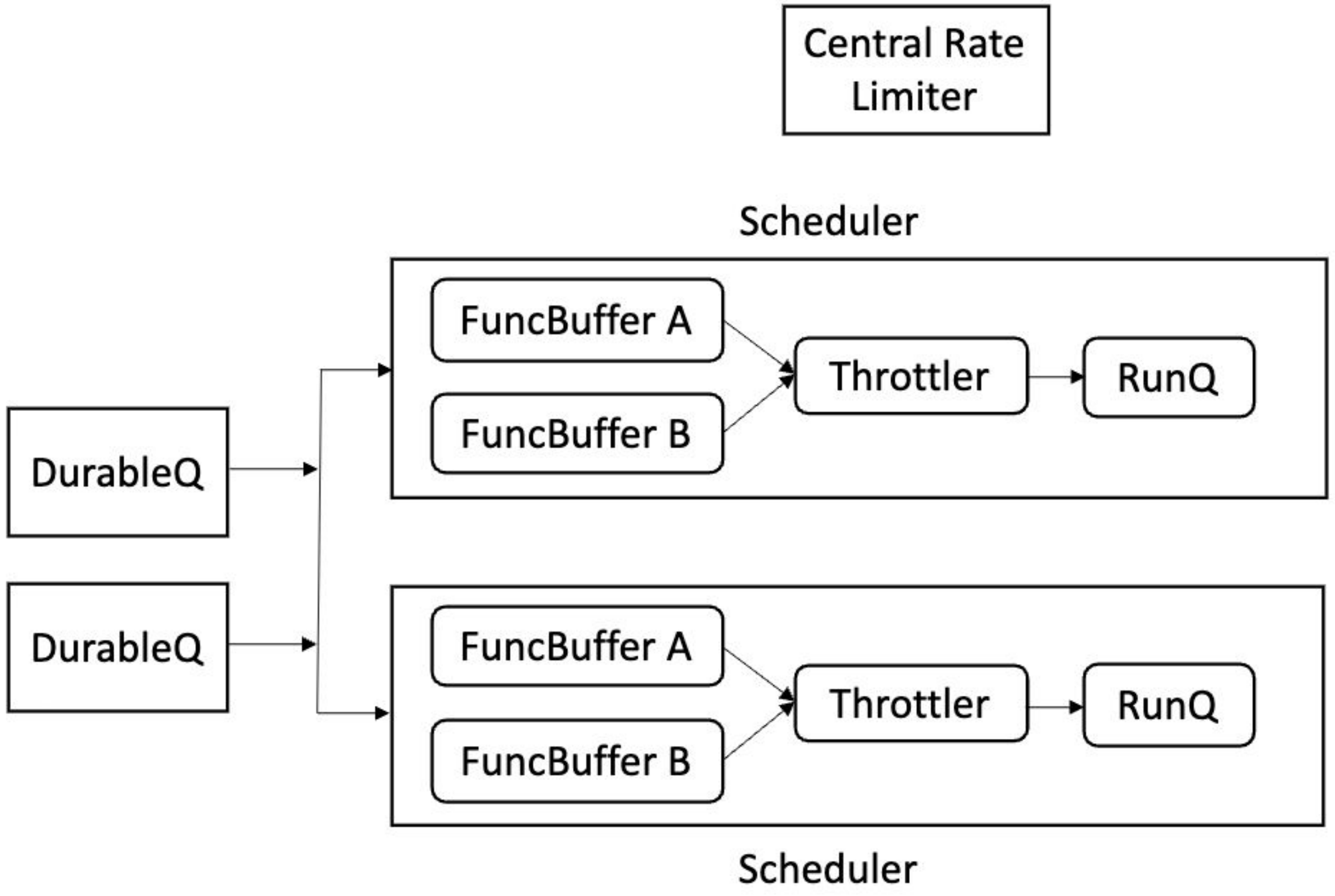
2. Opportunistic Quota





1. Reserved Quota

2. Opportunistic Quota
   - Dynamically adjusted based on worker utilization
   - Deferred to off-peak hours
   - SLA of 24 hrs



throttling_rate = base_rate_from_quota * S

High Worker Utilization: S ↘
Low Worker Utilization: S ↗

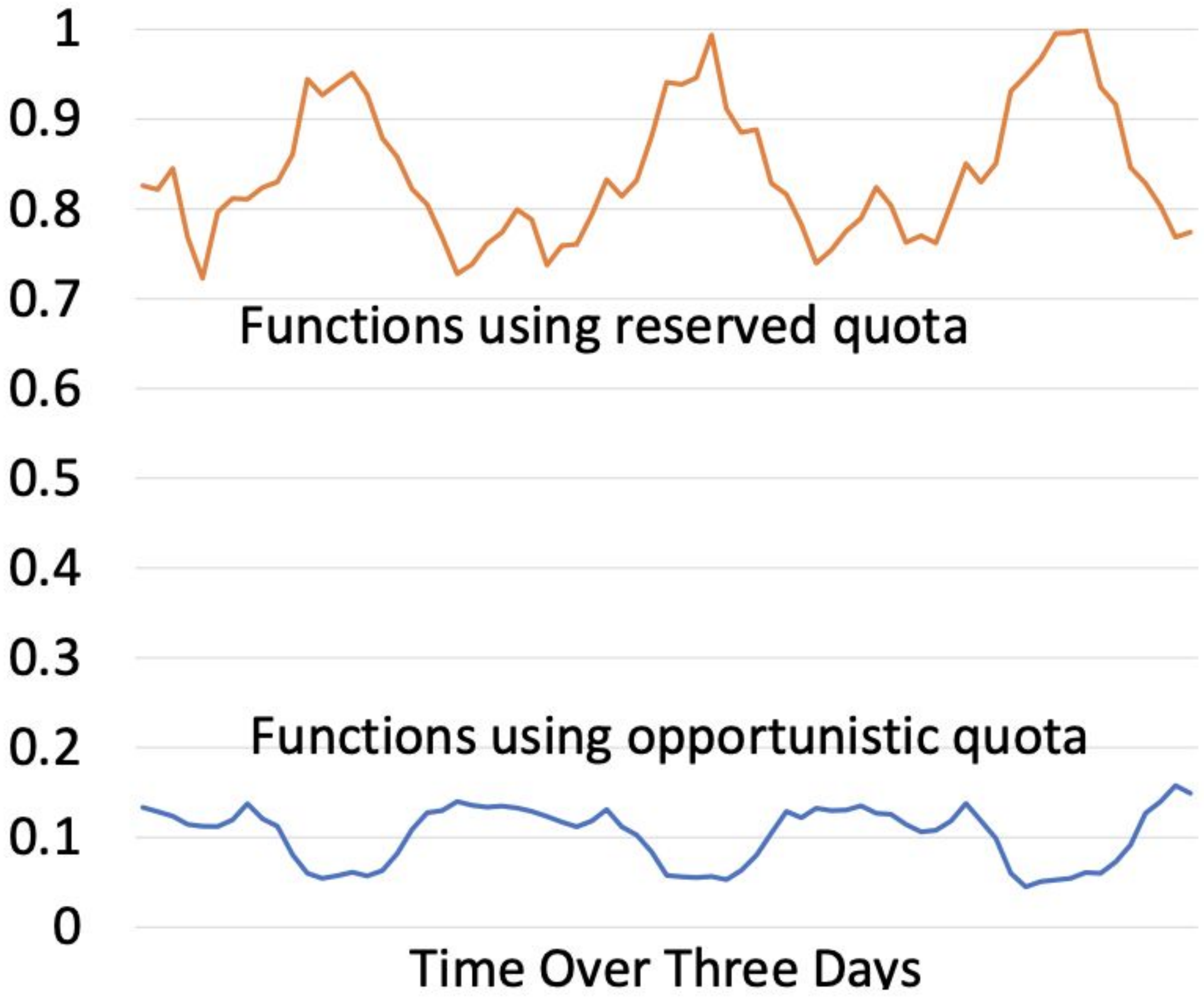
1. Reserved Quota
2. Opportunistic Quota

3. Per function criticality level
4. Explicit future execution start time

[NOT COVERED IN THIS TALK]

- Daily Peak Pattern
- Opportunistic Functions are Throttled during Peak

Total CPU Cycles Consumed by Functions



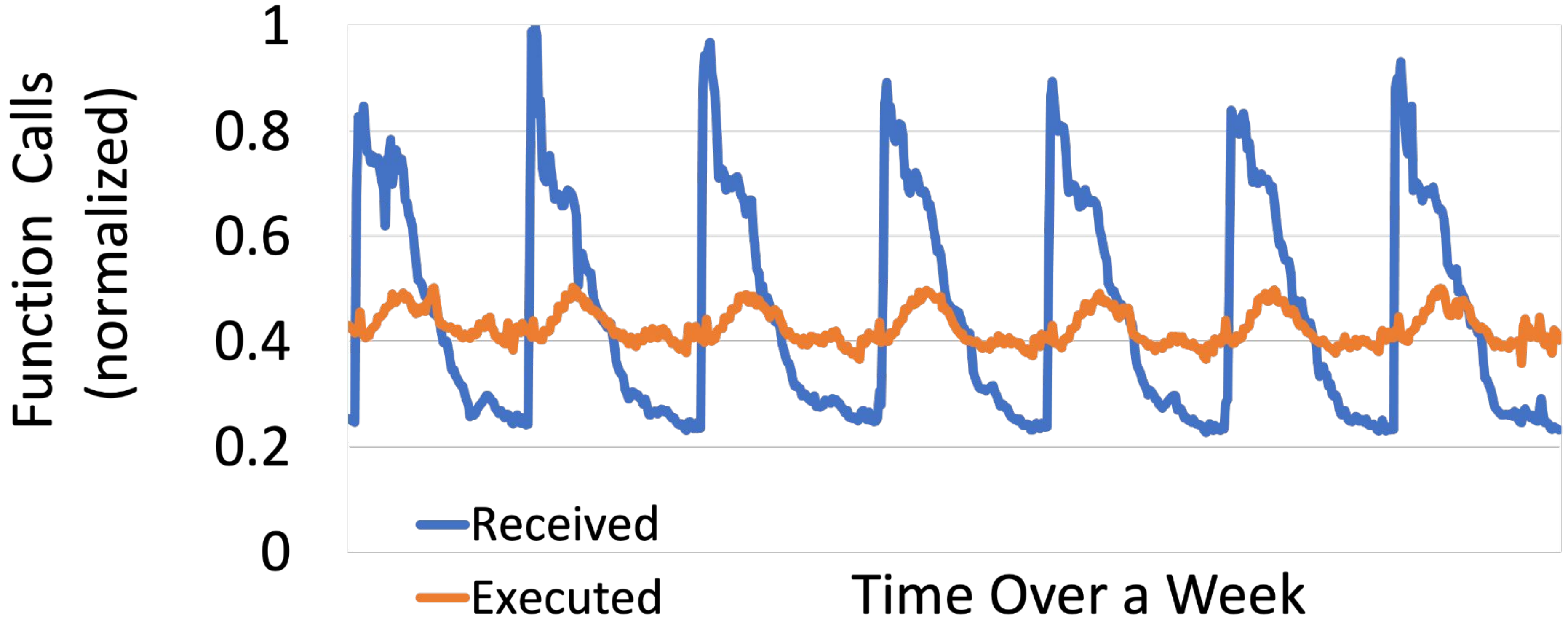
All Deferred Compute Features at Work

- Reserved Quota
- Opportunistic Quota
- Per Function Criticality
- Explicit future execution time

Cross Regional Load Balancing

Results:

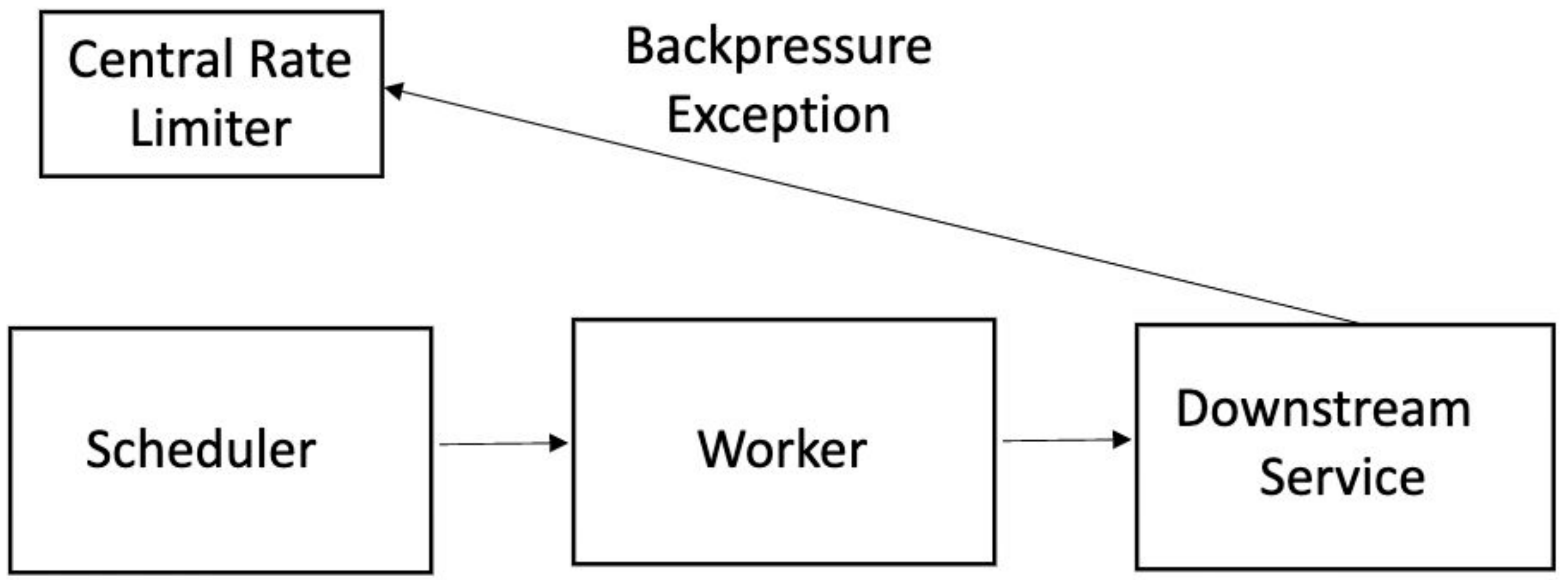
- PeaktoTrough reduced from 4.3x to 1.4x
- 66% Daily Average CPU Utilization

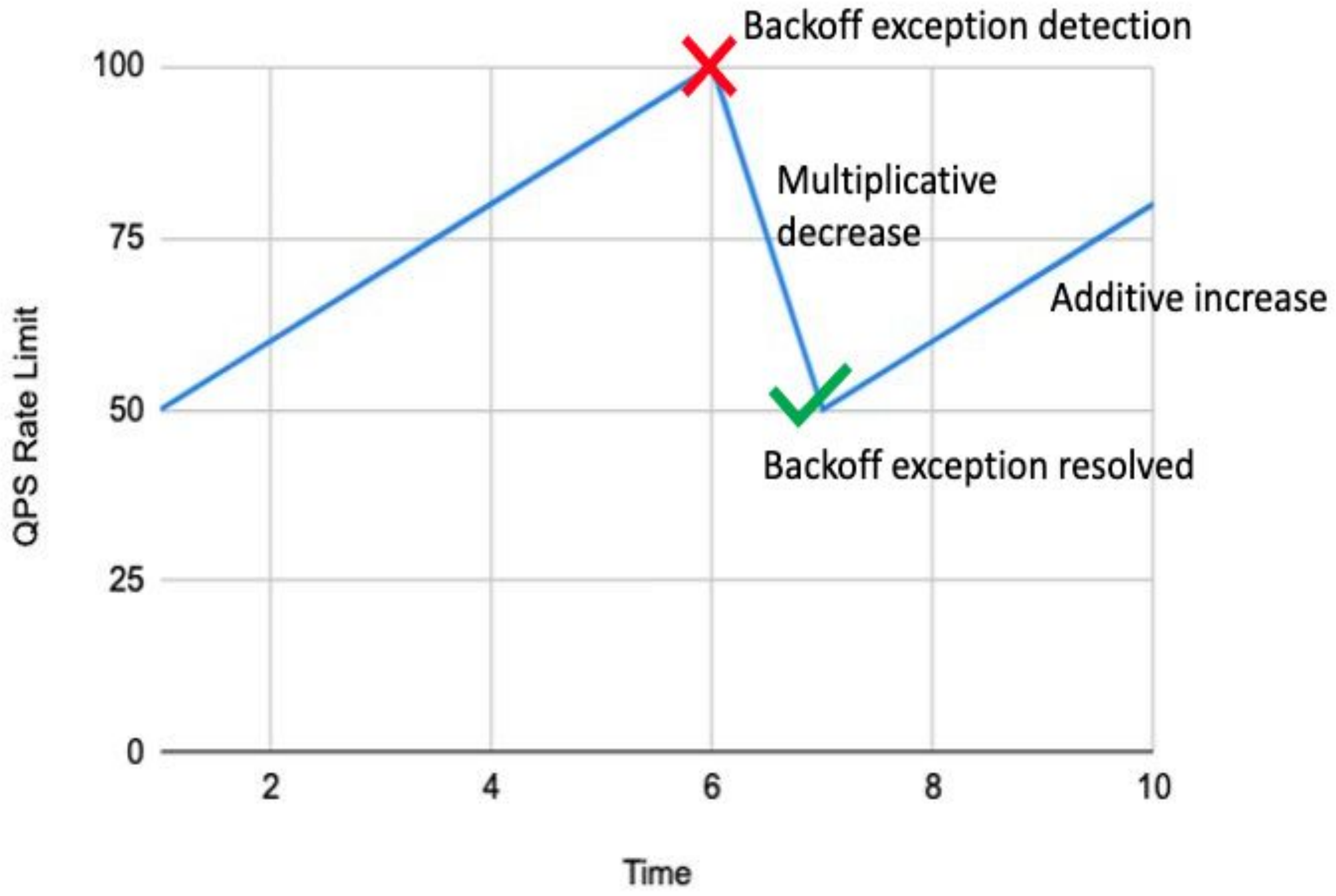


05 DOWNSTREAM PROTECTION - DESIGN & EVALUATION

Backpressure Handling

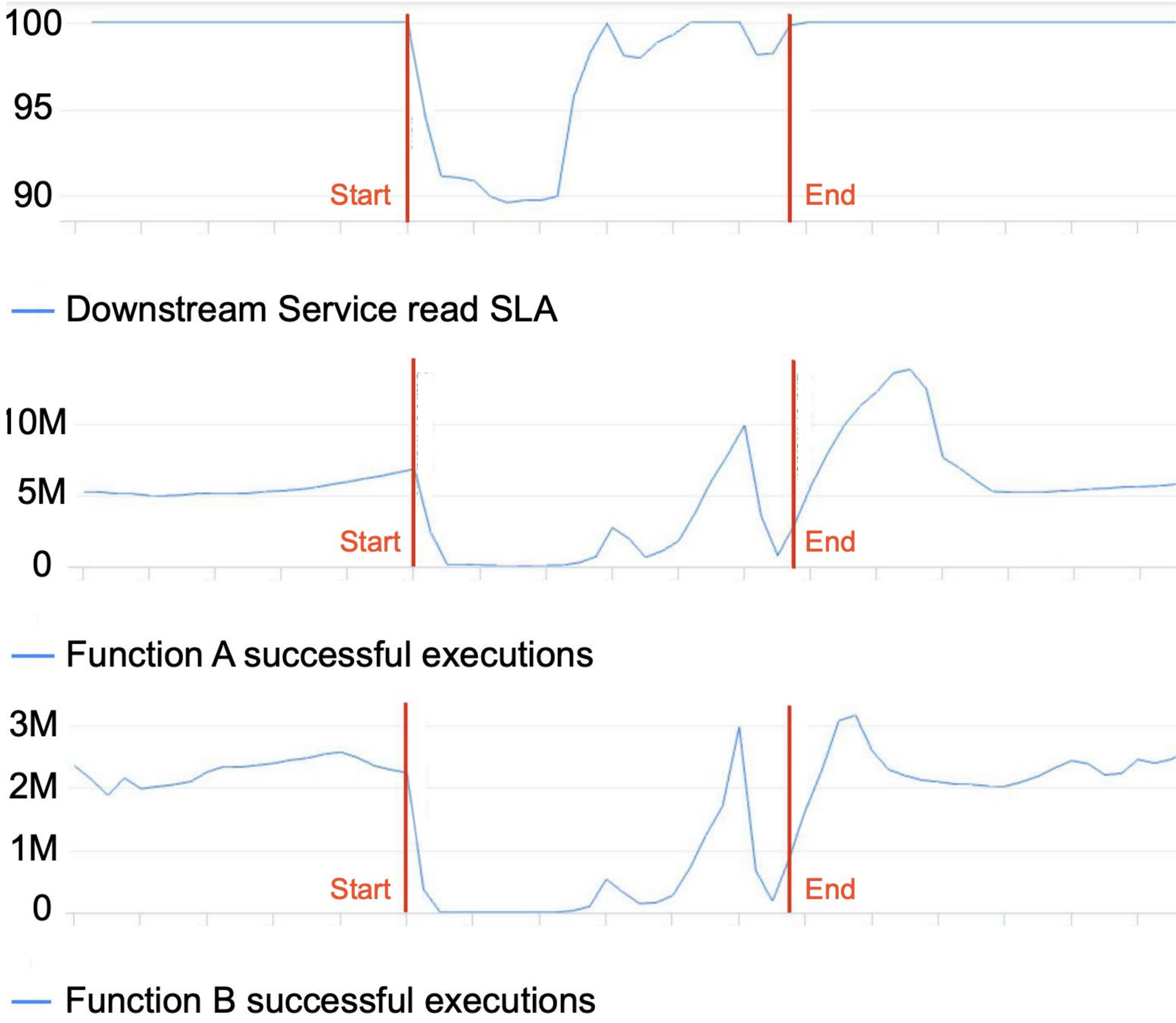
- Responds to Downstream Backpressure Exceptions
- Throttling rate is set by AIMD algorithm





05 DOWNSTREAM PROTECTION – DESIGN & EVALUATION

- Real incident during overload of WTCache in front of Social DB (TAO[1])
- Recovery was complete in two hours without any engineering intervention



[1] Nathan Bronson, et al. "TAO: Facebook’s Distributed Data Store for the Social Graph." In Proceedings of the 2013 USENIX Annual Technical Conference, 2013

XFaaS

HYPERSCALE AND LOW COST SERVELESS FUNCTIONS AT META

06 Summary

$O(10^{12})$

Function Calls/ Day

$O(10^{5})$

Servers

>10

Regions

Alireza Sahraei | asahraei@meta.com
Soteris Demetriou | s.demetriou@imperial.ac.uk

- XFaaS utilizes the concept of universal workers to eliminate cold start [NOT COVERED IN THIS TALK]
- Even if we eliminate cold start, we will still be often underutilized with need to autoscale almost instantaneously by 4x
- XFaaS embodies several methods to smoothen out the function execution curve => **daily avg CPU utilization at 66%**
- Ensures protection of downstream services

Meta Imperial College London Penn University of Pennsylvania Carnegie Mellon University

Meta