Accelerometer: Understanding Acceleration Opportunities for Data Center Overheads at Hyperscale

facebook | UNIVERSITY OF MICHIGAN

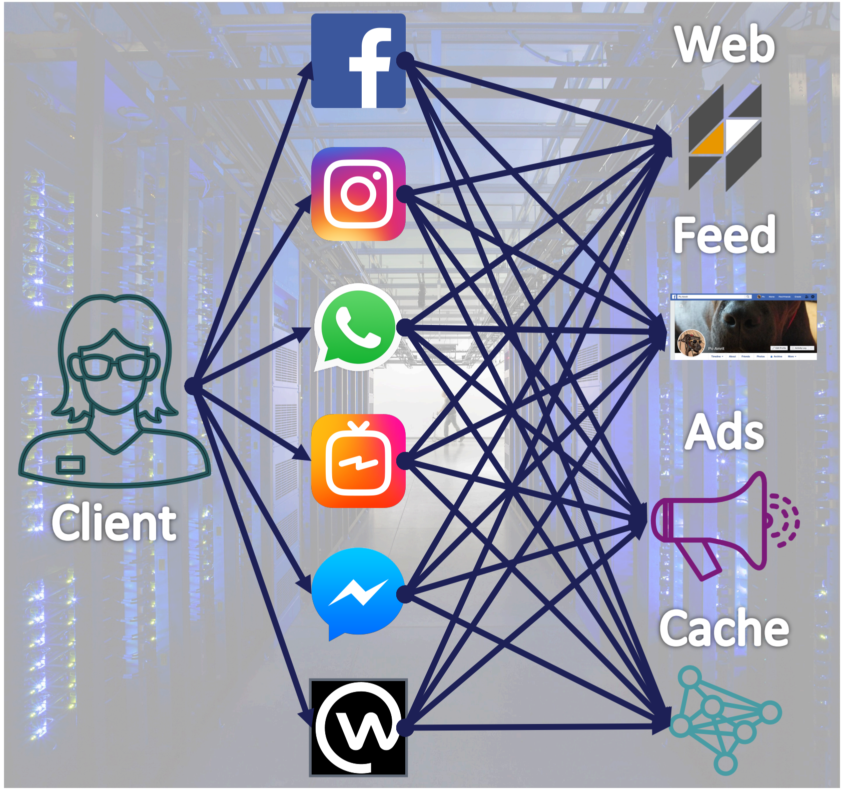
Akshitha Sriraman*

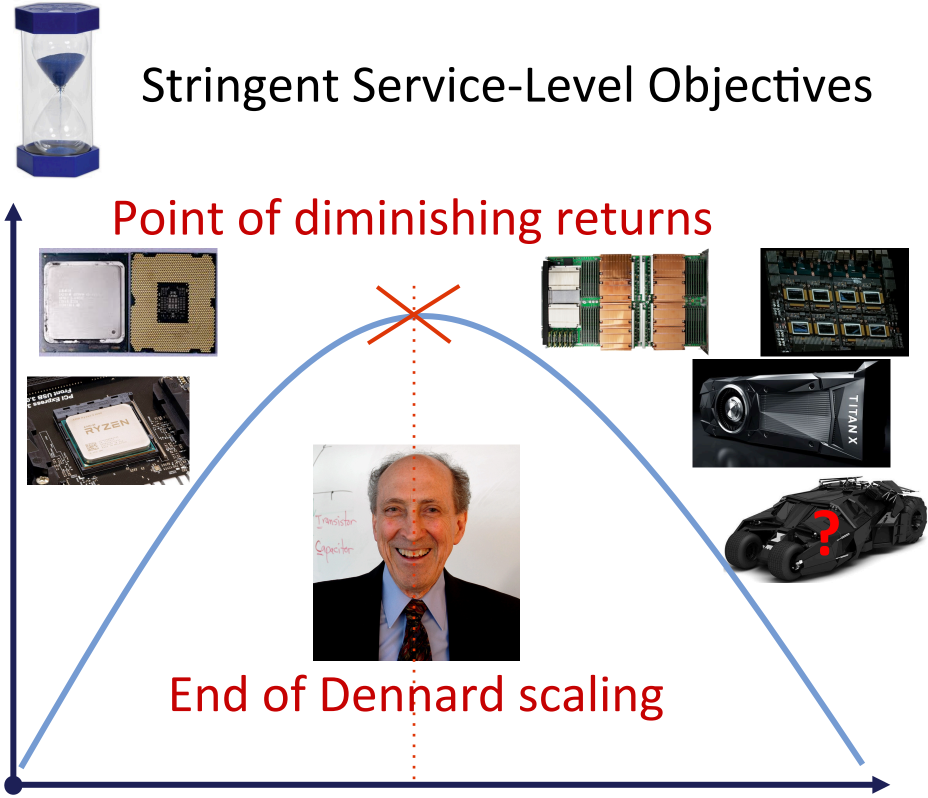
facebook

Abhishek Dhanotia

The 25th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2020)

Rapid Increase in Modern Web Services





µservice SLOs + end of Dennard scaling -> increase in custom HW

But.... What Should we Accelerate?

In-Datacenter Performance Analysis of a Tensor Processing Unit

Minerva: Enabling Low-Power, Highly-Accurate

HARE: Hardware Accelerator for Regular Expressions

TETRIS: A Streaming Accelerator for Physics-Limited 3D Plane-Wave Ultrasound Imaging

Vaib

Brendan L. West
University of Michigan
westbl@umich.edu

Jian Zhou
Arizona State University
jzhou50@asu.edu

Ronald G. Dreslinski
University of Michigan
rdreslin@umich.edu

J. Brian Fowlkes
University of Michigan
fowlkes@umich.edu

Oliver Kripfgans
University of Michigan
greentom@umich.edu

Chaitali Chakrabarti
Arizona State University
chaitali@asu.edu

Thomas F. Wenisch
University of Michigan
twenisch@umich.edu

Abstract—I

Abstract—Deep ... rapidly becoming ... a major component ... tasks, such as speech r...

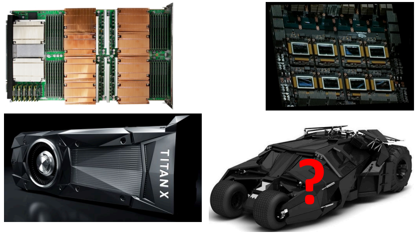
ABSTRACT

High volume acquisition rates are imperative for medical ultrasound ... rates to track high frequency motion. These applications also require large imaging apertures to capture sufficient resolution in a ...

time AI". The growing demand for computationally expensive, ... in live, interactive services, such as web search, advertising,

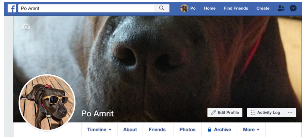
EVERYTHING?!

Key (?) Acceleration Opportunities

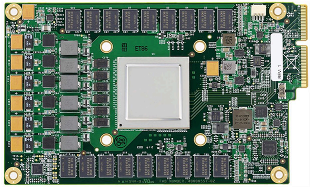


Too many custom HW -> EXPENSIVE

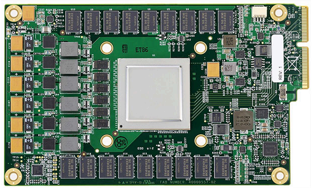
Feed microservice



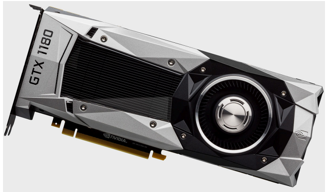
Accelerating ML inference



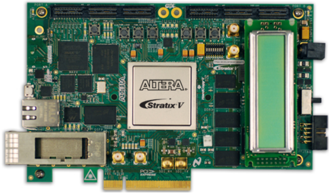
Widely-studied accelerators



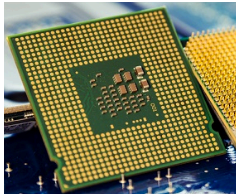
ML Inference



Graphics



Search ranking

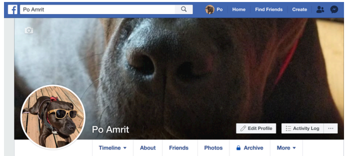


Graph processing

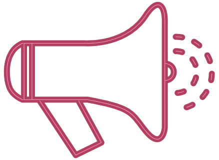
What is end-to-end Feed's throughput increase? < 1.5x!

Acceleration Opportunities at Facebook

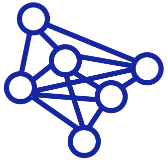
Web



Feed1, Feed2



Ads1, Ads2

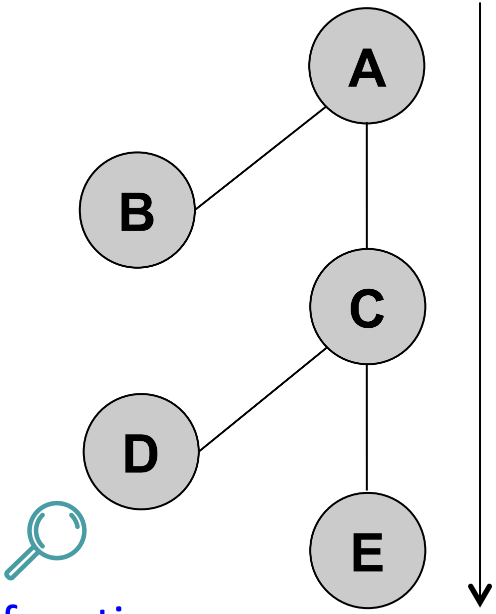


Cache1, Cache2

Which μservice operations consume the most CPU cycles?

Do μservices have common overheads that can inspire future HW designs?

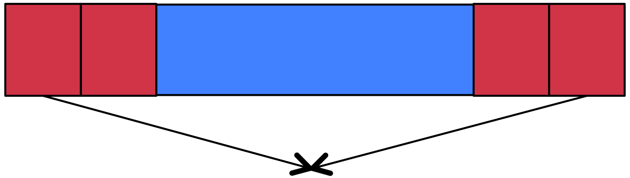
Contribution 1: Where Did My Cycles Go?



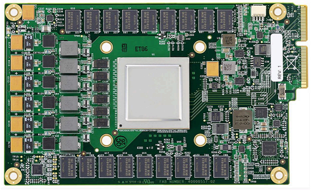
Leaf function:
memcpy()

µservice functionalities:
ML inference

ML inference

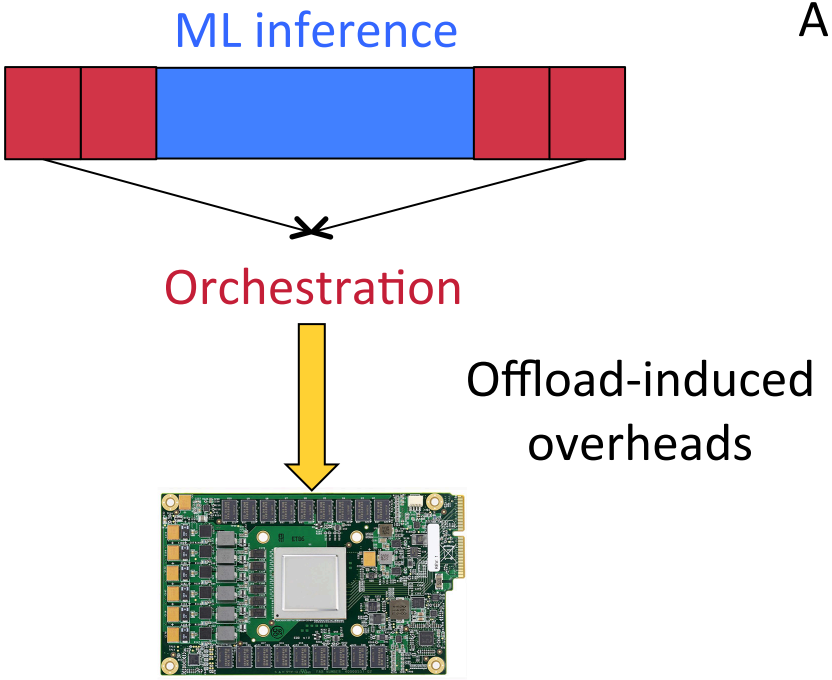


Orchestration

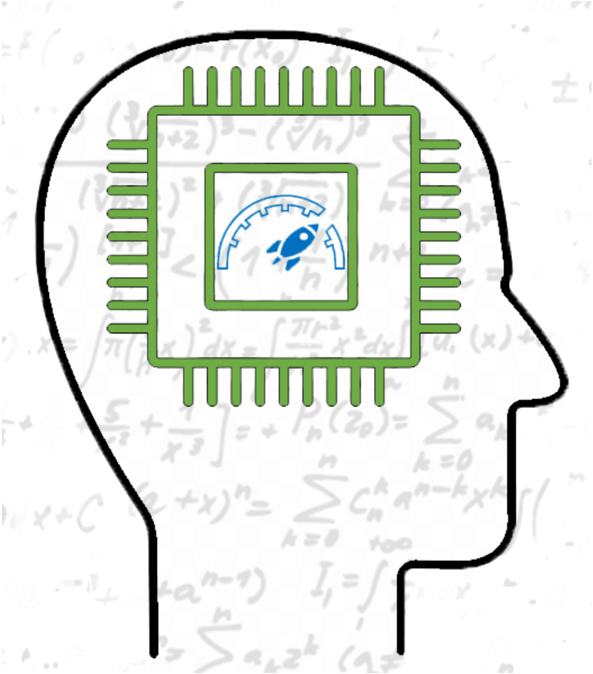


High & common orchestration overheads: Worth accelerating

Contribution 2: Accelerometer

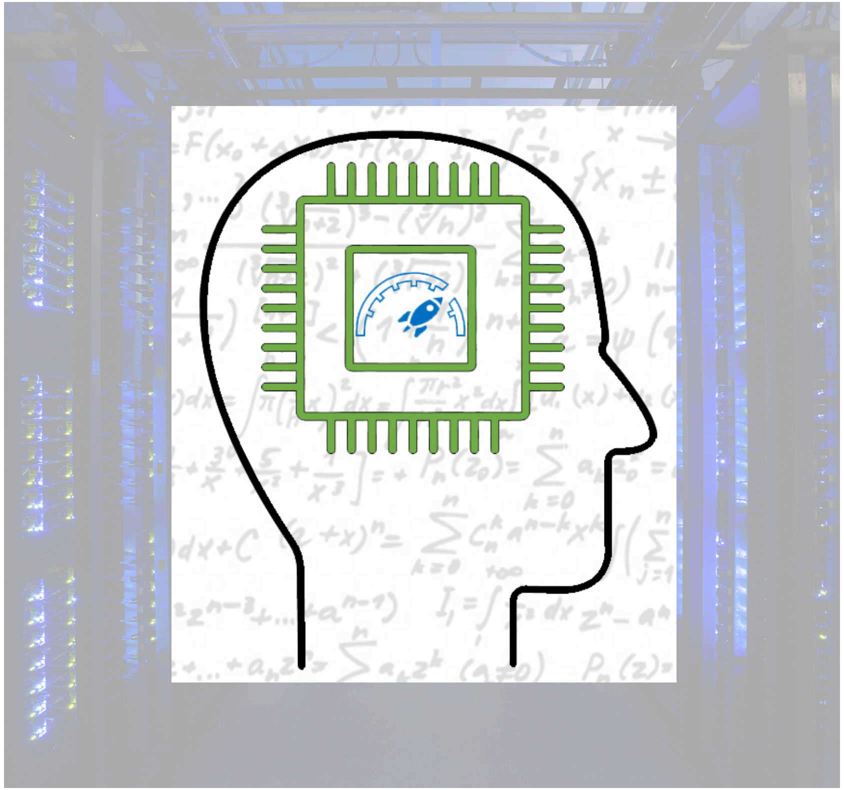


Analytical model for hardware acceleration



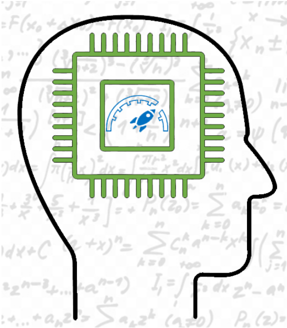
Accelerometer: Projects perf. bounds early in the HW design phase

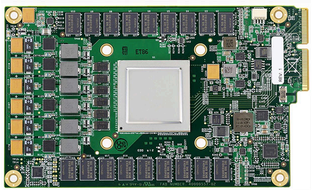
Contribution 3: Validating Accelerometer

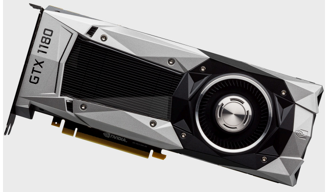


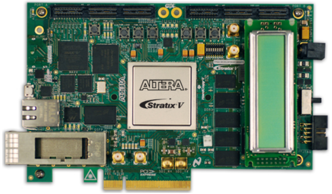
Validating in ***production****: Three retrospective case studies*

Contribution 4: Applying Accelerometer









How can you apply accelerometer to make good HW investments?

Introduction
Post Dennard scaling: What to accelerate?

Characterization
Where do data center cycles go?

Accelerometer
Analytical model for HW acceleration

Validation
Production case studies

Application
How we use Accelerometer

Introduction

Post Dennard scaling:
What to accelerate?

Characterization

Where do data center cycles go?

- Leaf function
- Service functionality

Accelerometer

Analytical model for HW acceleration

Validation

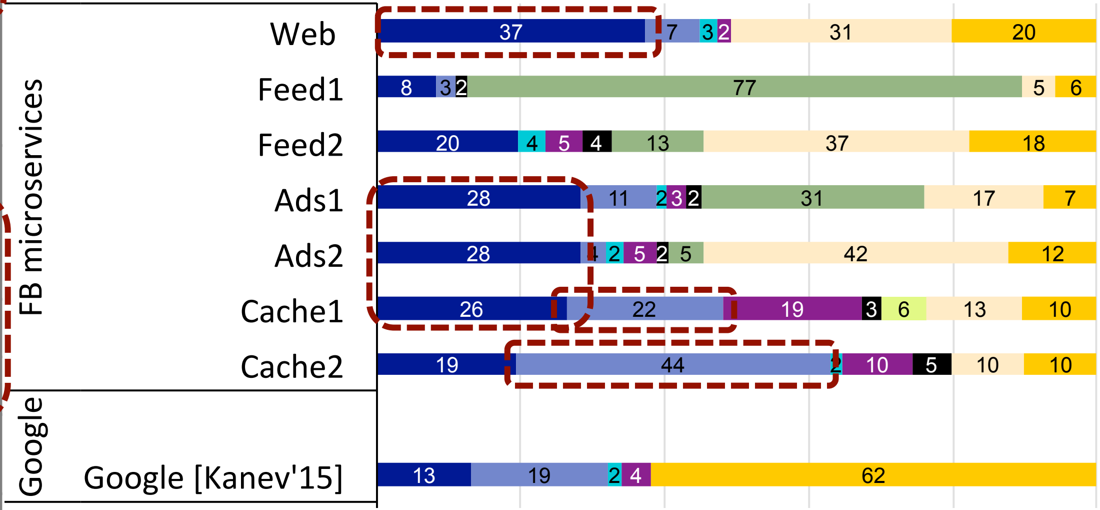
Production case studies

Application

How we use Accelerometer

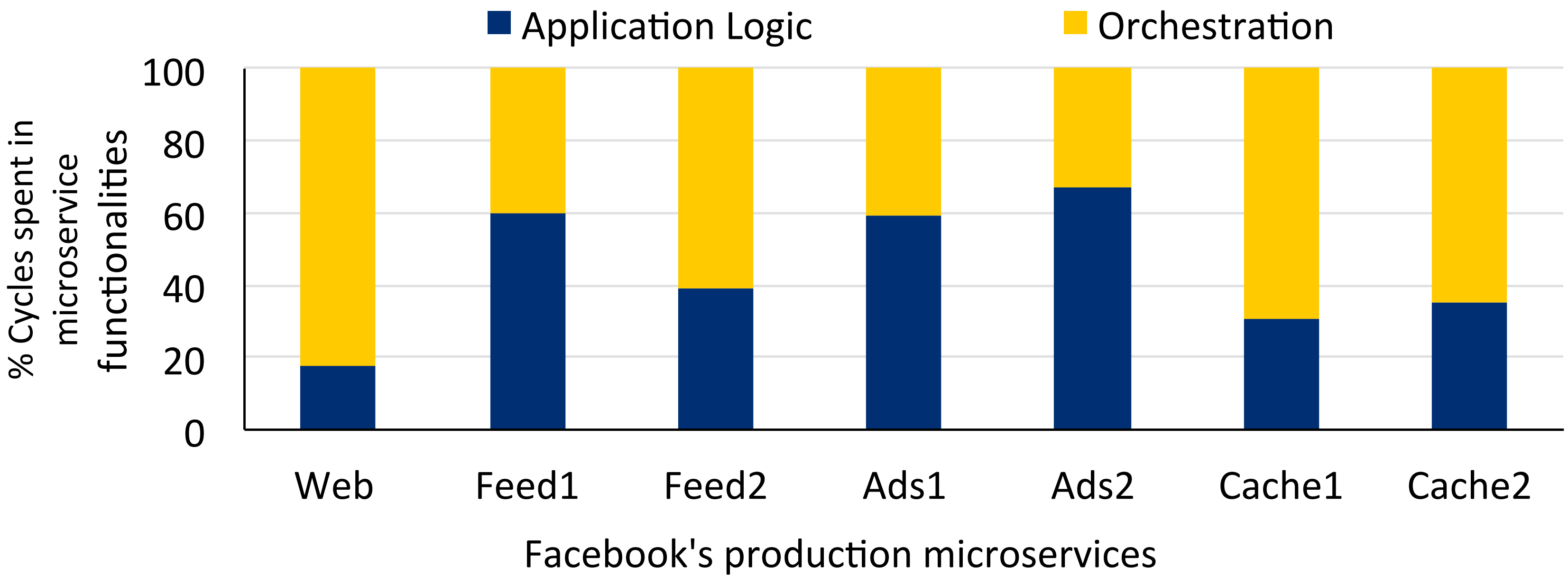
Leaf Function Characterization

Leaf category	Examples
Memory	Copy, allocation, free, compare
Kernel	Scheduling, interrupt handling, network comm., mem. mgmt.
Hashing	SHA
Synchronization	C++ atomics, mutex, spin locks, CAS
ZSTD	(De)compression
Math	Intel's MKL, AVX
SSL	En(de)cryption
C Libraries	Search algorithms, array & string compute
Miscellaneous	Assorted functions



Memory and kernel leaf functions dominate

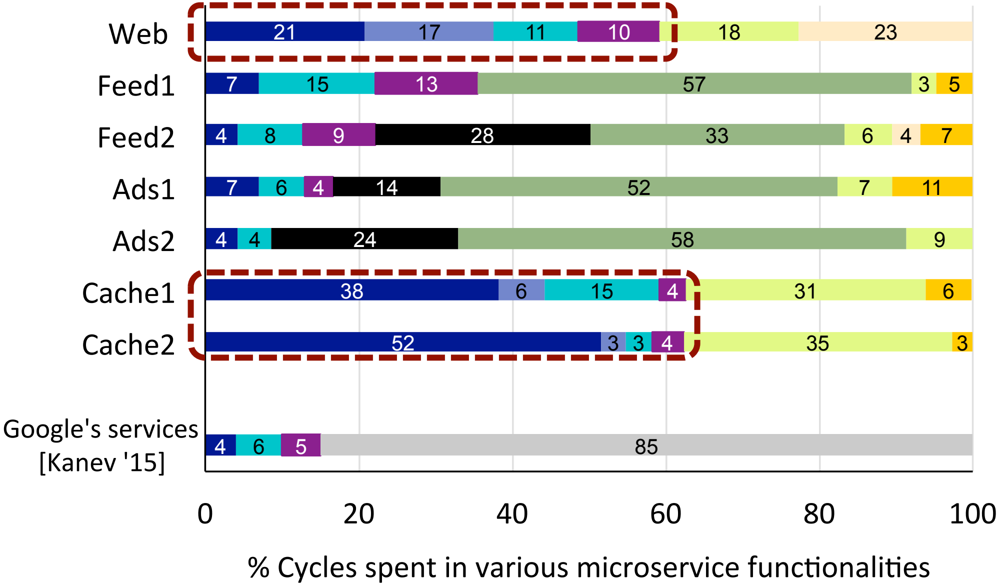
µService Functionality Characterization



Orchestration overheads dominate across microservices

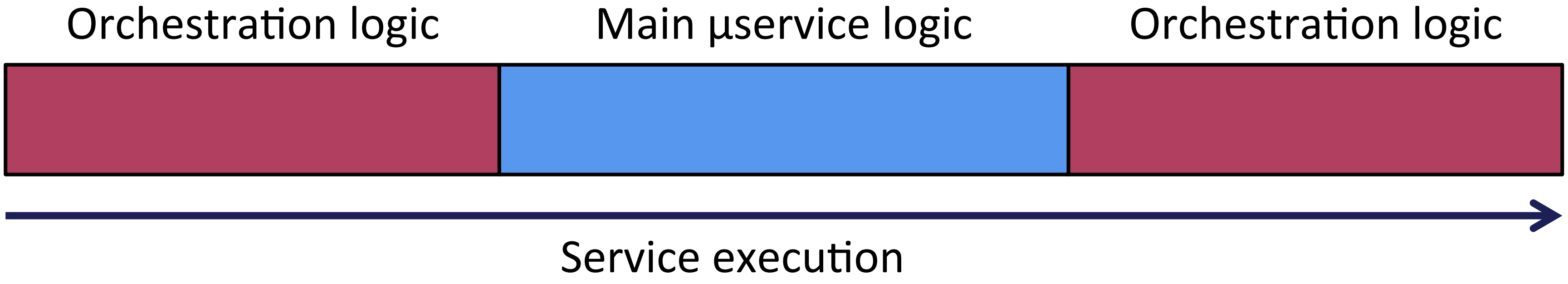
µService Functionality Breakdown

Functionality	Examples
(In)secure IO	IO send/receive
IO pre/post processing	Copy, alloc., etc before/ after IO
(De)compression	(De)compression logic
(De)serialization	RPC De(serialization)
Feature extraction	Feature vector creation in ML services
Prediction/Ranking	ML inference
Core app. logic	Core business logic
Logging	Creating, reading, updating logs
Thread pool mgmt.	Creating, deleting, synchronizing threads

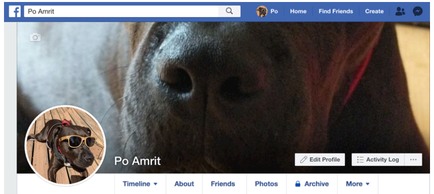


Common orchestration overheads dominate across microservices

Characterization Takeaways



Web



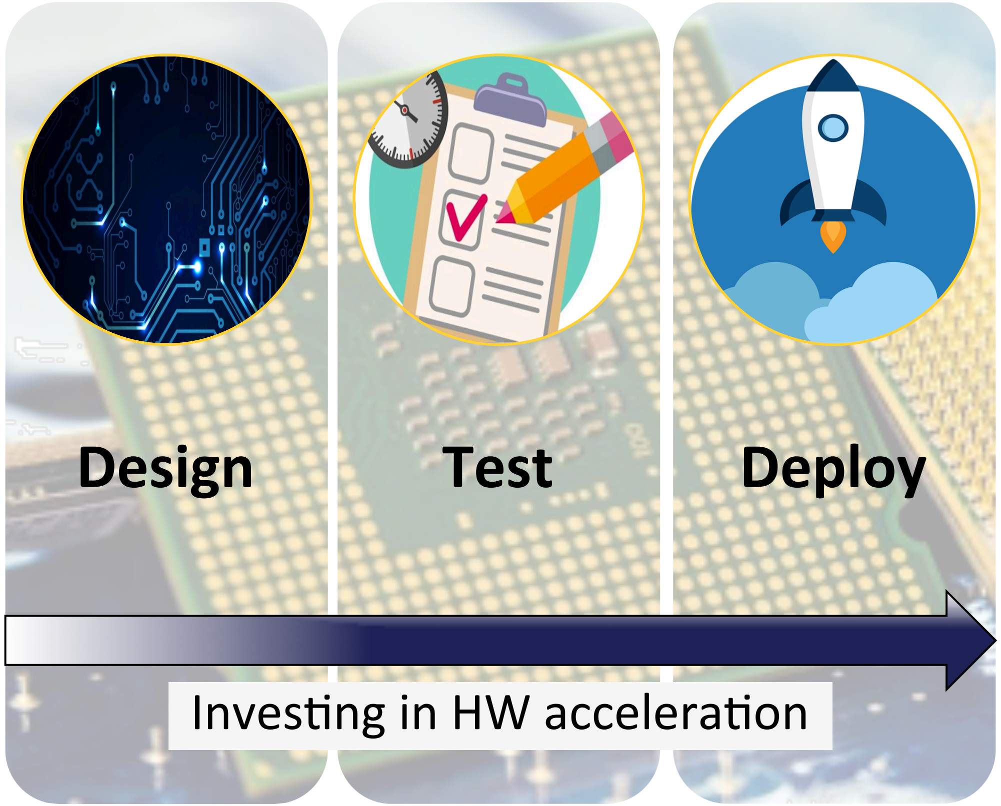
Feed1, Feed2

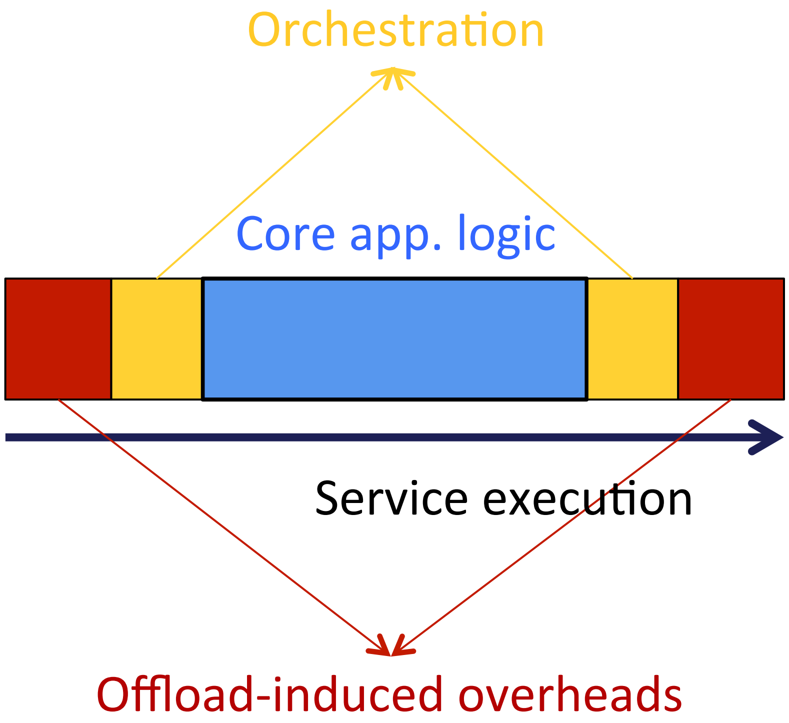
Ads1, Ads2

Cache1, Cache2

Accelerating orchestration can improve perf. across global fleet

Investing in HW Acceleration





Risky @scale due to perf. bounds from offload overheads

Introduction
Post Dennard scaling: What to accelerate?

Characterization
Where do data center cycles go?

Accelerometer
Analytical model for HW acceleration

Validation
Production case studies

Application
How we use Accelerometer

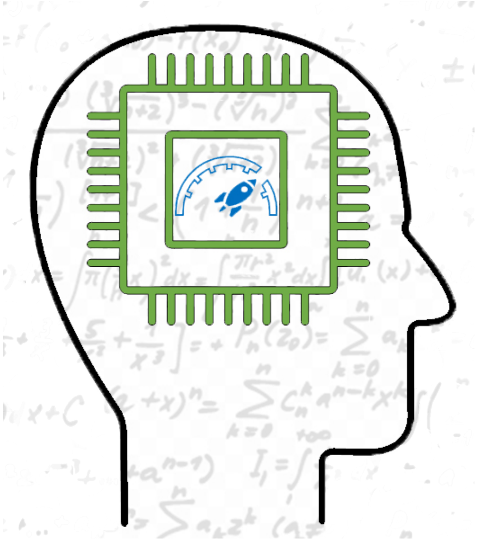
Accelerometer: Analytical Model

Acceleration strategy

On-chip

Off-chip

Remote



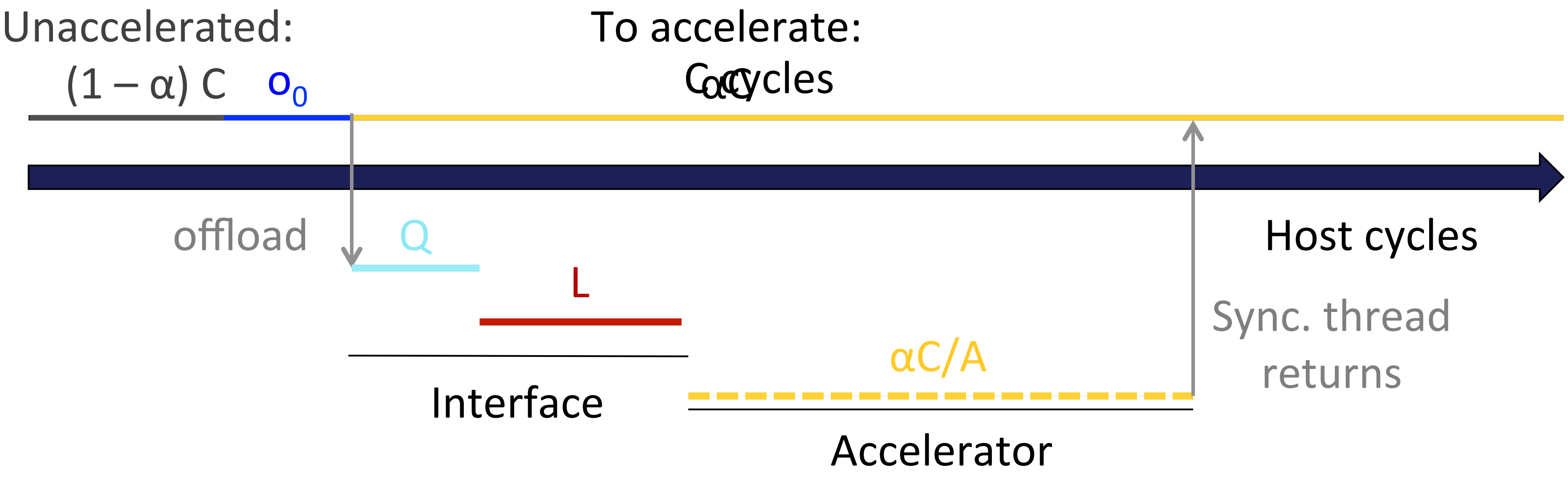
Threading design for offload

Synchronous

Asynchronous

Models throughput speedup and per-request latency reduction

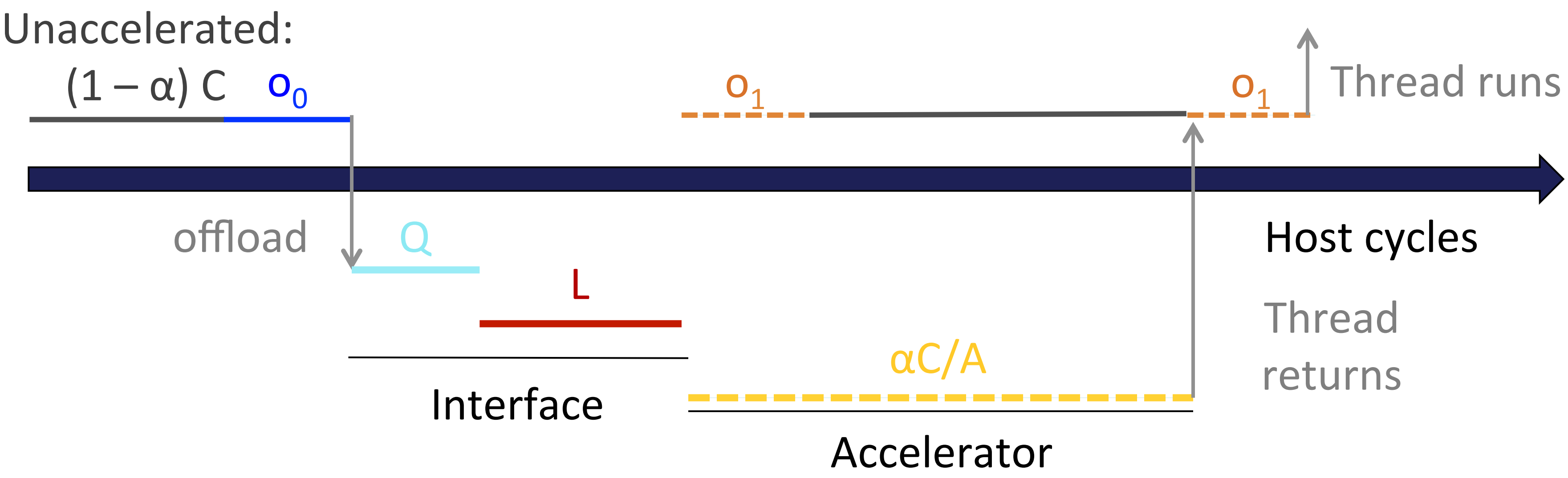
Synchronous Offload



$$\text{Speedup} = \frac{C}{(1-\alpha)C + \alpha C/A + (o_0 + Q + L)\,n}$$

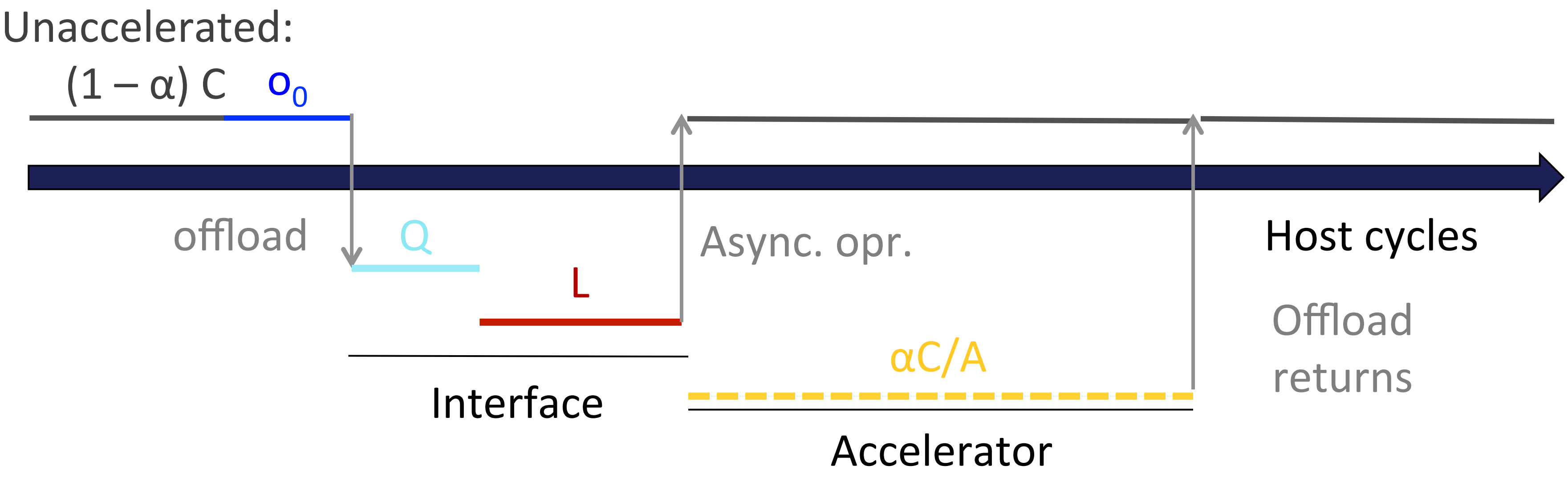
Accelerator cycles critically affect speedup & latency reduction

Sync. Offload with Thread Oversubscription



Context switch penalties impact speedup & latency reduction

Asynchronous Offload



*Accelerator cycles **do not** critically affect speedup*

Introduction
Post Dennard scaling: What to accelerate?

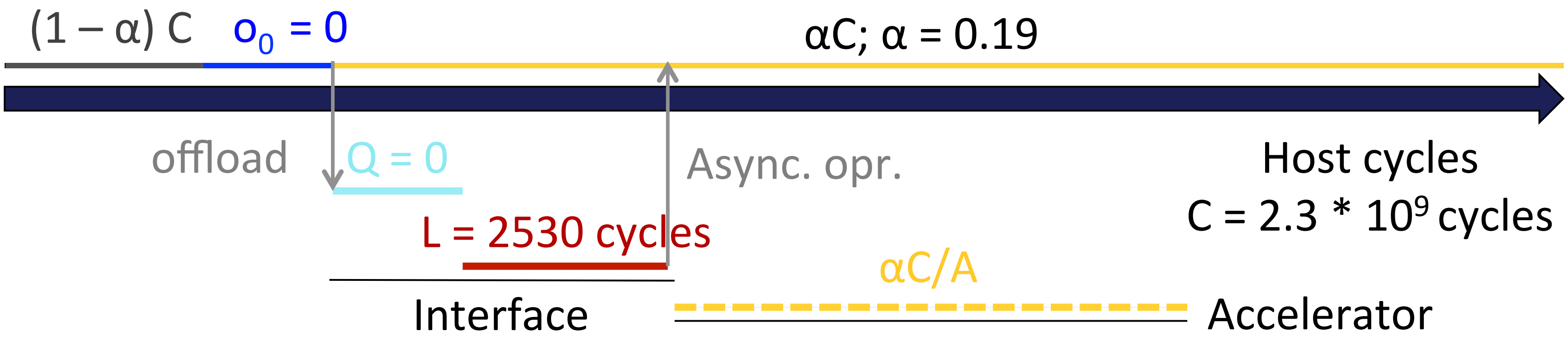
Characterization
Where do data center cycles go?

Accelerometer
Analytical model for HW acceleration

Validation
Production case studies

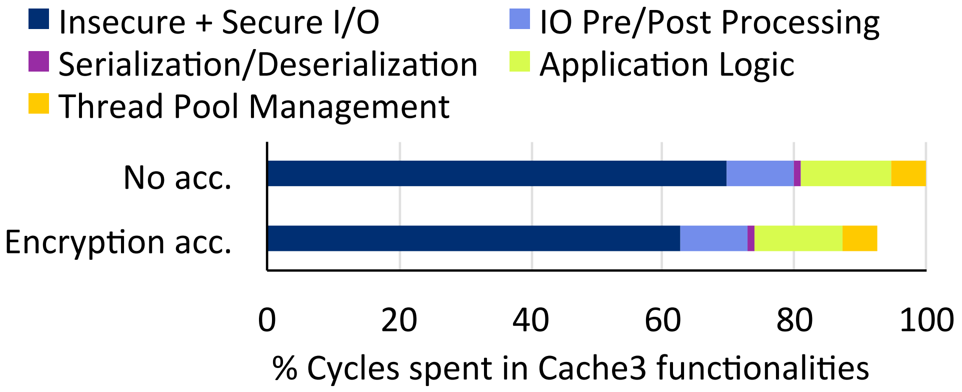
Application
How we use Accelerometer

Validating Accelerometer: Encryption



Estimated speedup = 8.6%

Real speedup = 7.5%



Accelerometer estimates speedup with <= 3.7% error

Introduction

Post Dennard scaling: What to accelerate?

Characterization

Where do data center cycles go?

Accelerometer

Analytical model for HW acceleration

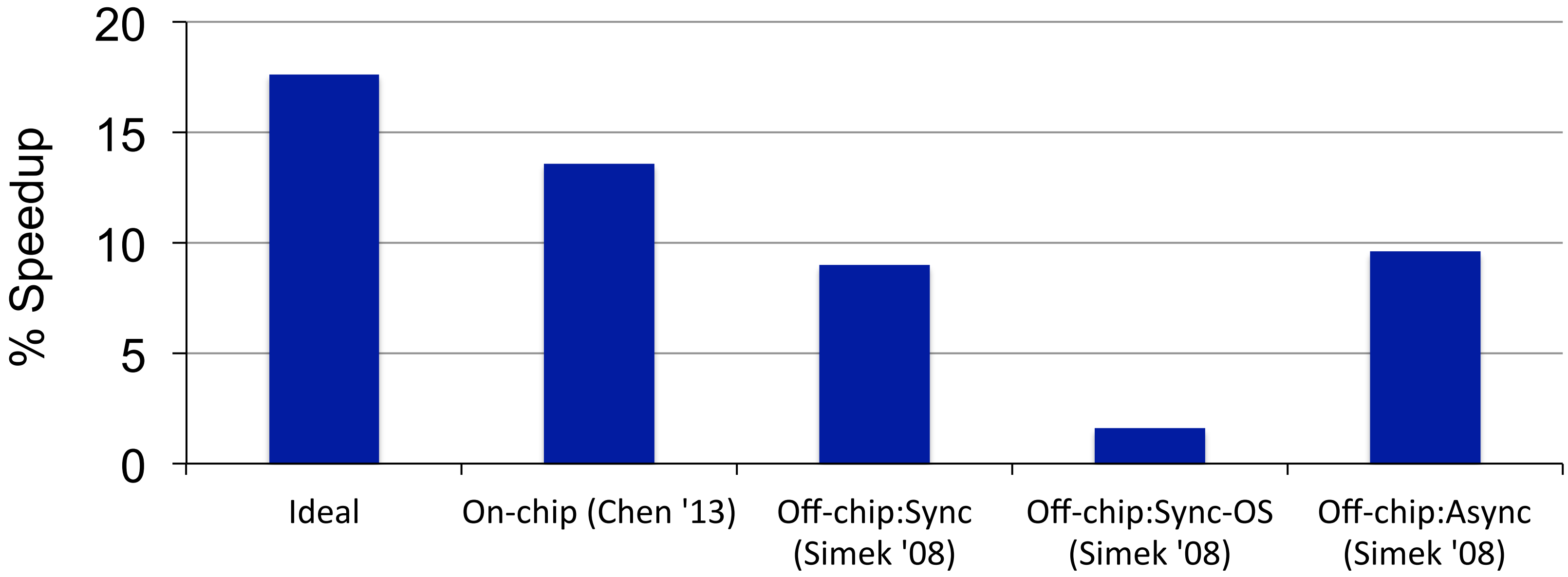
Validation

Production case studies

Application

How we use Accelerometer

Applying Accelerometer: Compression



Accelerometer can identify good accelerator investments

Introduction
Post Dennard scaling: What to accelerate?

Characterization
Where do data center cycles go?

Accelerometer
Analytical model for HW acceleration

Validation
Production case studies

Application
How we use Accelerometer

Accelerometer: Understanding Acceleration Opportunities for Data Center Overheads at Hyperscale

facebook
UNIVERSITY OF MICHIGAN

Akshitha Sriraman*

facebook

Abhishek Dhanotia

The 25th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2020)