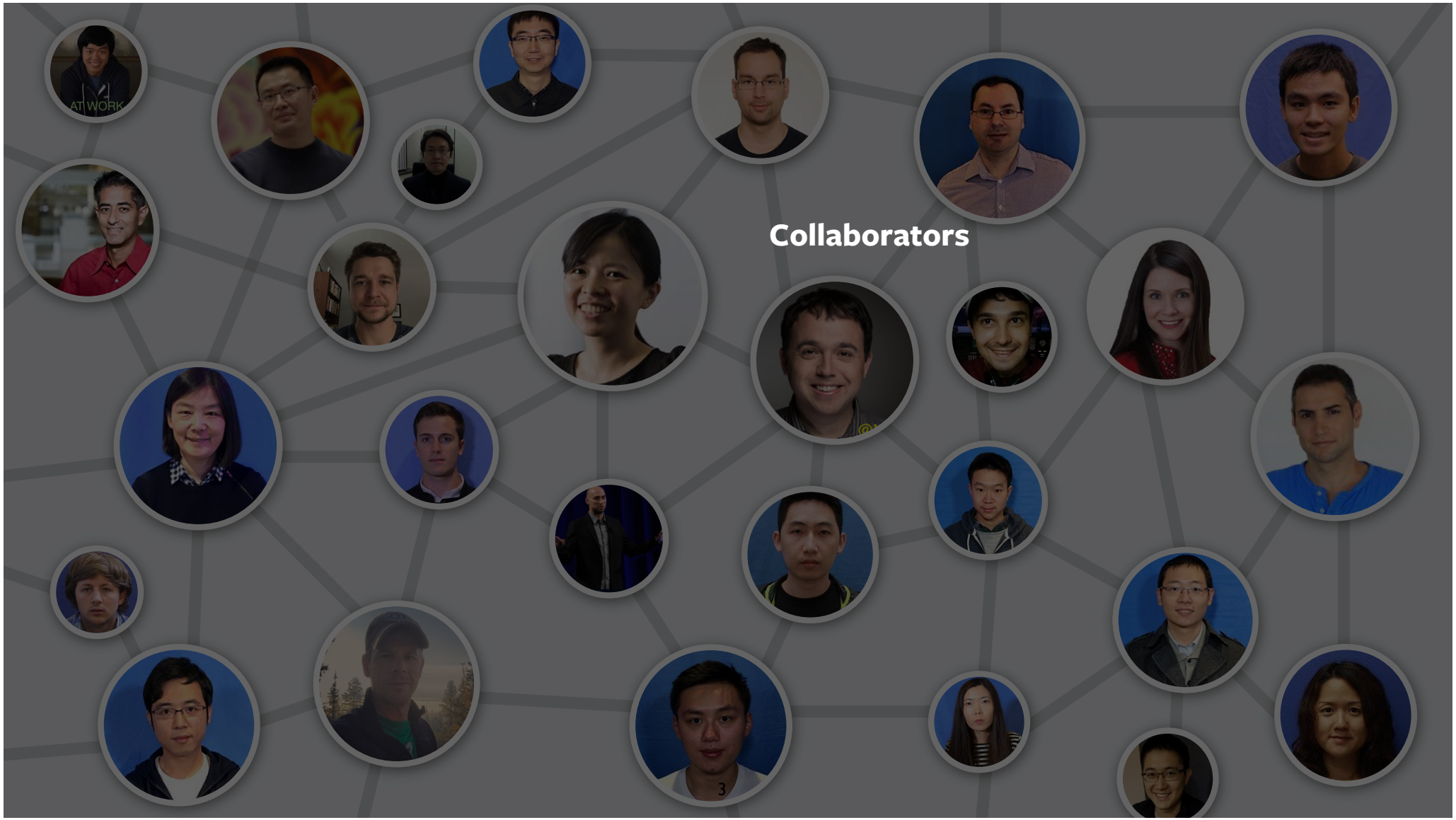
Machine Learning @ Facebook

Understanding Inference at the Edge

IEEE International Symposium on High Performance Computer Architecture (HPCA-2019)

Carole-Jean Wu

Research Scientist @ Facebook
AI Infra Research


Collaborators

Machine Learning @ FB

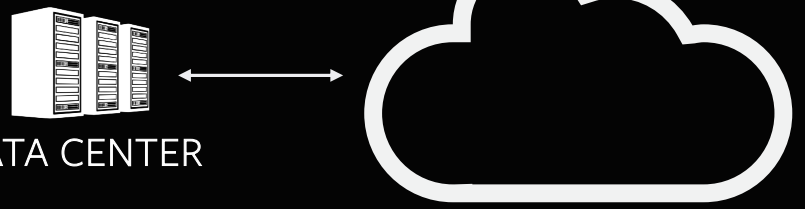
- Ranking of posts in news feeds
- Content understanding
- Object detection, segmentation, and tracking
- Speech recognition / translation

From data centers to the edge

- **Minimizing network bandwidth**
- **Improving response latency**
- **Exploiting features available only at the edge**

Keypoints Segmentation

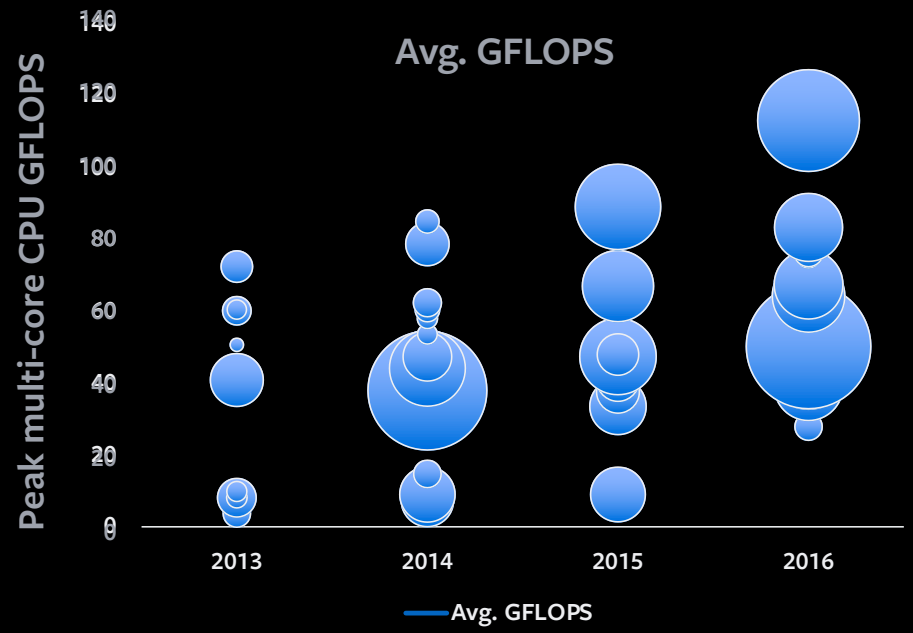
Augmented Reality with Smart Camera



Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective. Hazelwood et al. HPCA-2018.

Unique Challenges for Edge Inference

- Feature-rich edge inference is enabled by the ever increasing mobile performance
- Increasing core counts leads to theoretical peak performance increase. But, when looking at the entire ecosystem, the theoretical peak performance is a widespread.



DELIVERING CONSISTENT INFERENCE PERFORMANCE IS CHALLENGING

Unique Challenges for Edge Inference

The Diversity of Mobile Hardware and Software is Not Found in the Controlled Datacenter Environment.

2	3	20+	20+	10+
MAJOR MOBILE OS	MAJOR GRAPHICS APIs	MAJOR CHIPSET VENDORS	MAJOR CPU UARCH	MAJOR GPU UARCH

How do we optimize system designs for real-time ML inference?

FRAGMENTED SMARTPHONE ECOSYSTEM POSES UNIQUE CHALLENGES FOR EDGE INFERENCE

Introduction:
Machine Learning @ FB & Unique Challenges for Edge Inference

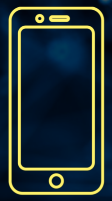
Lay of the Land:
Closer Look at Smartphones that FB Runs on

Horizontal Integration:
Making Inference on Smartphones

Vertical Integration:
Processing Inference for Oculus VR

Inference in the Wild:
Performance Variability

Introduction:
Machine Learning @ FB & Unique Challenges for Edge Inference



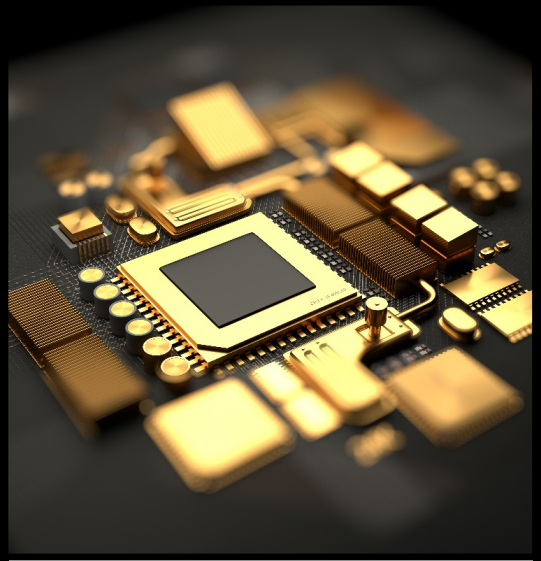
Lay of the Land:
Closer Look at Smartphones that FB Runs on

Horizontal Integration:
Making Inference on Smartphones

Vertical Integration:
Processing Inference for Oculus VR

Inference in the Wild:
Performance Variability

What is Challenging for Mobile Inference?



Fragmentation

There is no standard mobile SoC to optimize for.
Mobile CPUs Show Little Diversity

Performance

The Performance Difference between a Mobile CPU and GPU is Narrow

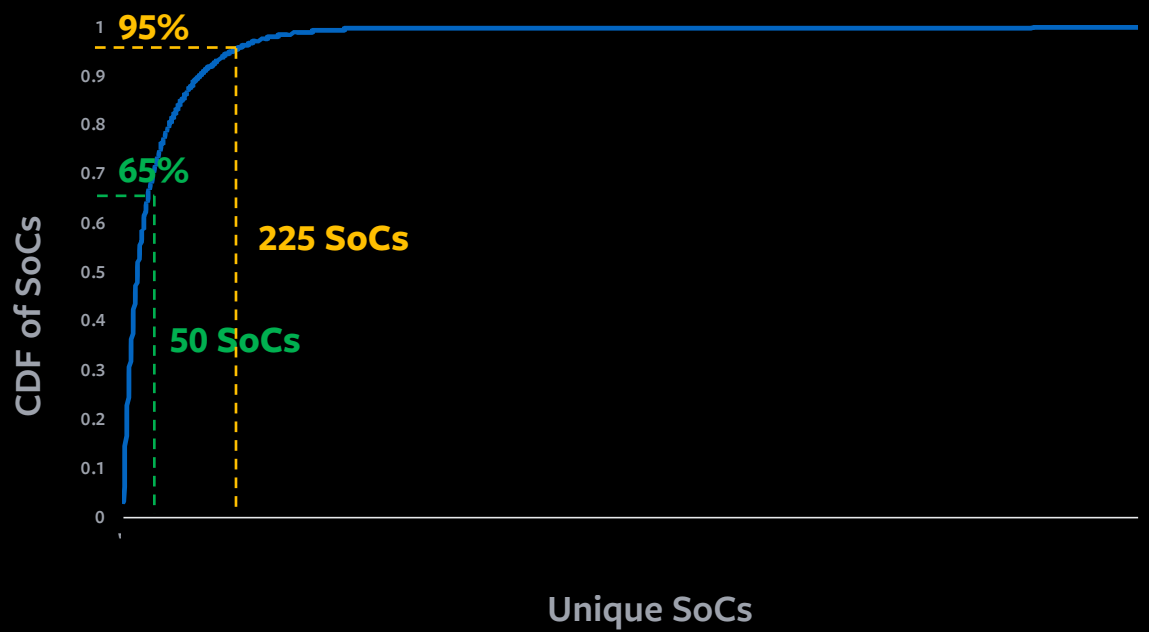
Programmability

Programmability is a Primary Roadblock for Using Mobile Co-processors

Lay of the Land

FRAGMENTATION

Taking a Closer Look at Smartphones Facebook Runs on



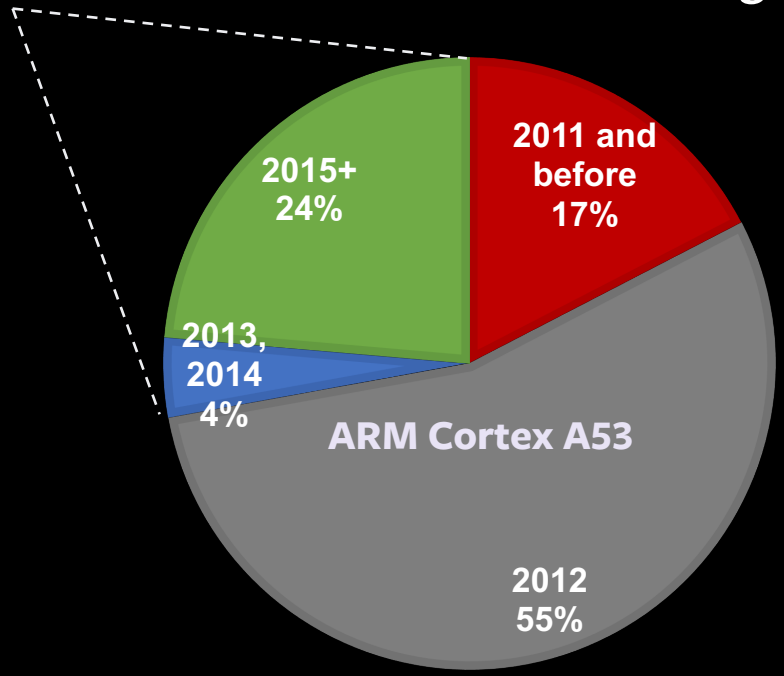
- Qualcomm Snapdragon
- Samsung Exynos
- MediaTek Helio
- HiSilicon Kirin et al.

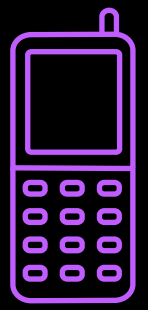
THERE IS NO STANDARD SOC TO OPTIMIZE FOR

Lay of the Land

FRAGMENTATION

In 2018, ~28% of SoCs Use CPUs Designed in 2013 or Later





72%

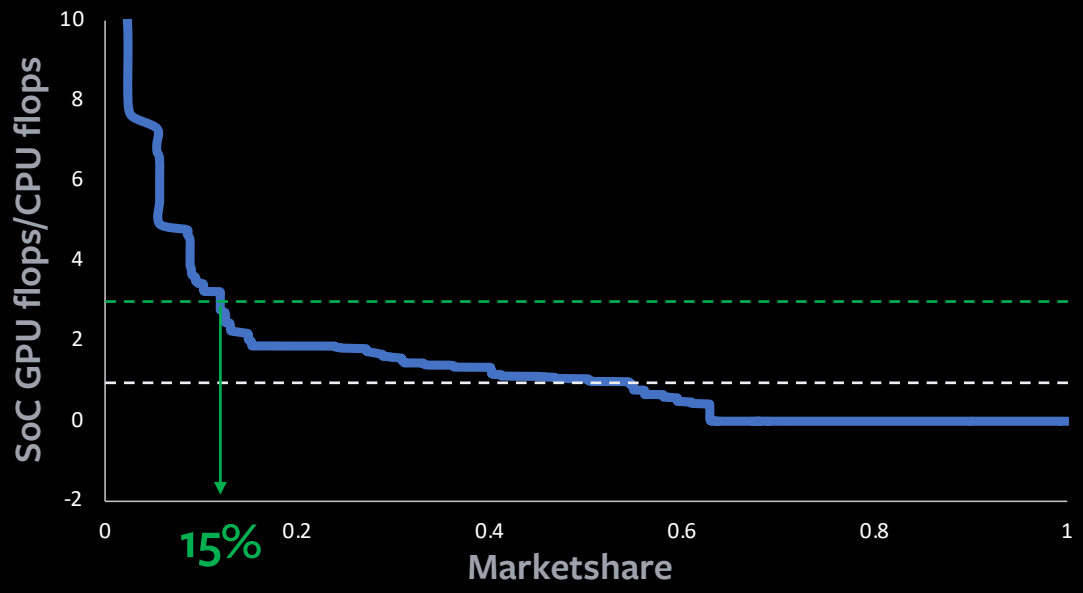
OF THE WORLD'S CELL PHONES ARE MORE THAN 7 YEARS OLD

MOBILE CPUS SHOW LITTLE DIVERSITY

Lay of the Land

PERFORMANCE

The Performance Difference between a Mobile CPU and GPU is Narrow



ON A MEDIAN SMARTPHONE, THE GPU PROVIDES AS MUCH THEORETICAL PEAK PERFORMANCE AS ITS CPU

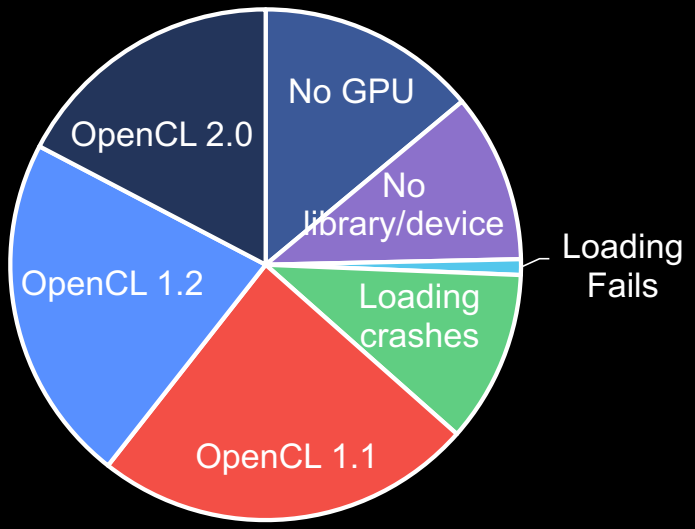
LESS THAN 15% SMARTPHONES HAVE A GPU THAT IS 3 TIMES AS POWERFUL AS ITS CPU

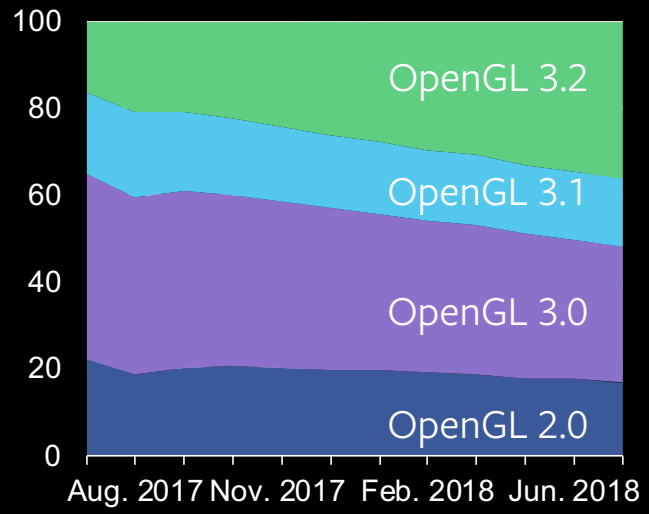
Lay of the Land

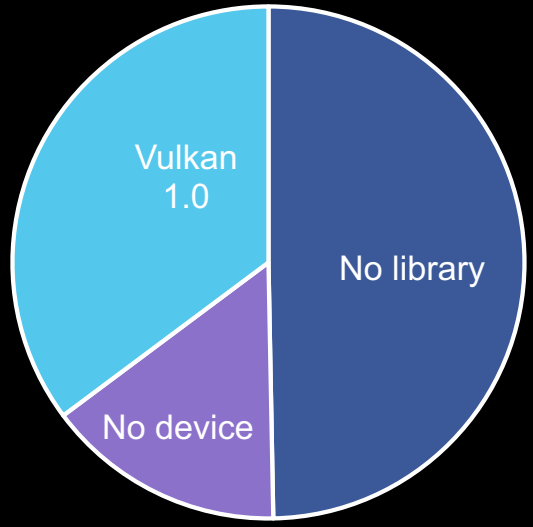
PROGRAMMABILITY

Programmability is a Primary Roadblock for Using Mobile Co-processors

- OpenCL, OpenGL ES, Vulkan for Android GPUs







ANDROID GPUS HAVE FRAGILE USABILITY AND POOR PROGRAMMABILITY WHILE IOS HAS BETTER SUPPORT WITH METAL

Quantitative Approach to Mobile Inference Designs

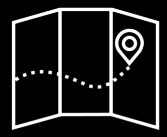
State of the Practice for Mobile Inference is Using CPUs

FRAGMENTATION

- There are more than **2000+ different SoCs** but mobile CPUs show little diversity with ARM's Cortex A53 dominating the market

PERFORMANCE

- Performance difference between mobile **CPUs** and **GPUs** is narrow



PROGRAMMABILITY

- Programmability is a major road block for **co-processors** (e.g. Android GPUs)

MOBILE INFERENCE OPTIMIZATION IS TARGETED FOR THE COMMON DENOMINATOR OF THE FRAGMENTED SOC ECOSYSTEM

Introduction: Machine Learning @ FB & Unique Challenges for Edge Inference

Lay of the Land: Closer Look at Smartphones that FB Runs on

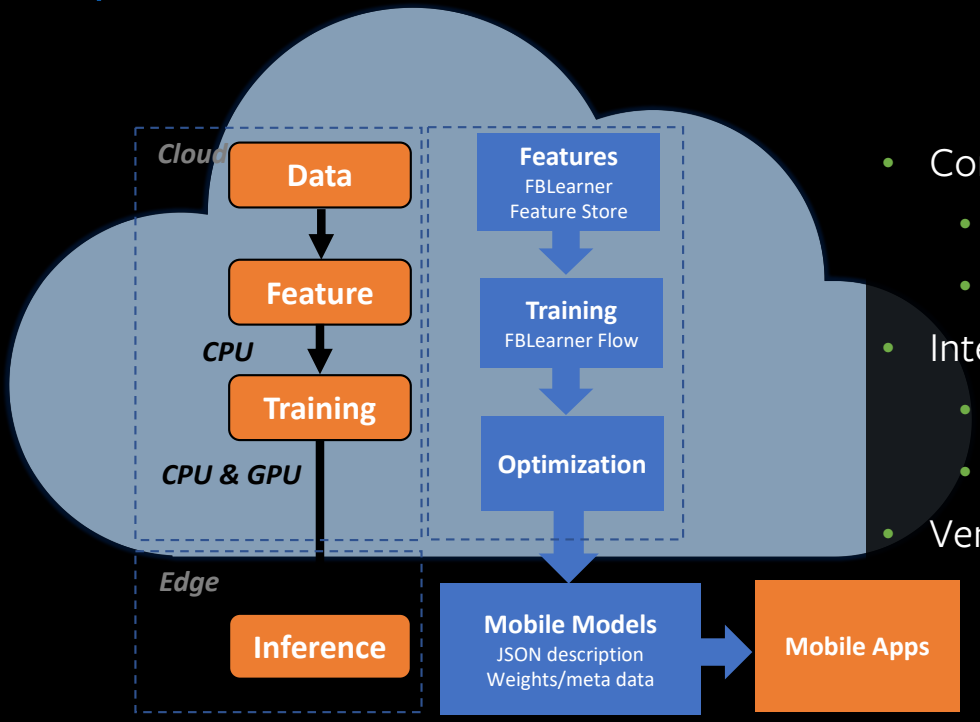
Horizontal Integration: Making Inference on Smartphones

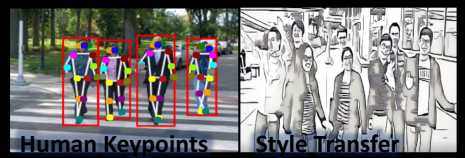
Vertical Integration: Processing Inference for Oculus VR

Inference in the Wild: Performance Variability

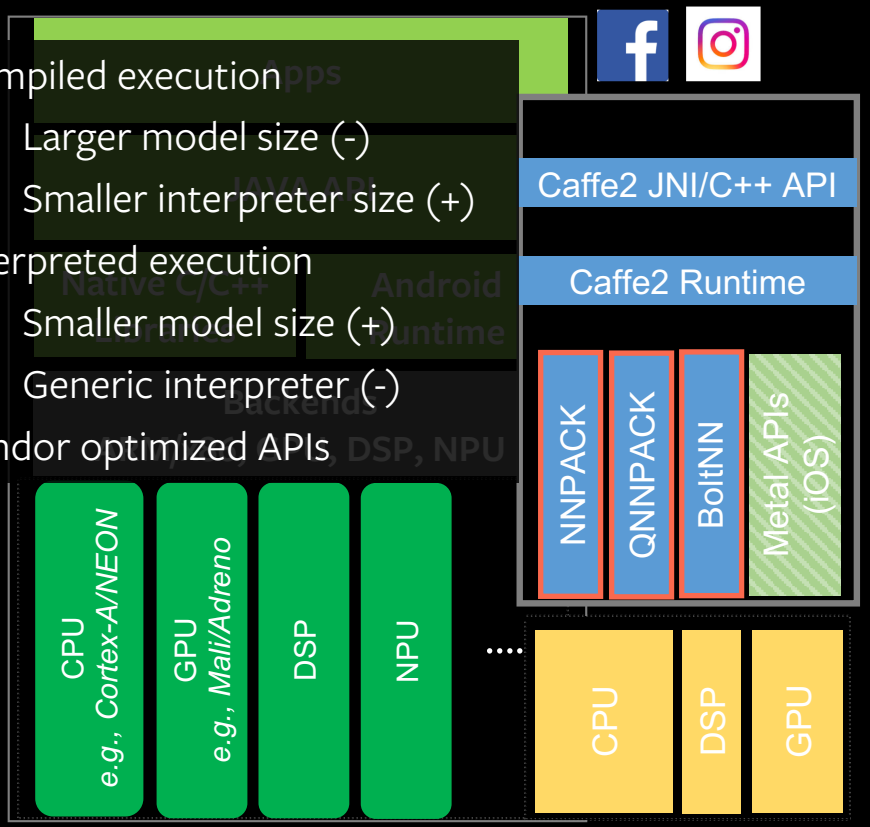
Horizontal Integration

Making Inference on Smartphones in the Wild





- Compiled execution
  - Larger model size (-)
  - Smaller interpreter size (+)
- Interpreted execution
  - Smaller model size (+)
  - Generic interpreter (-)
- Vendor optimized APIs



Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective. Hazelwood et al. HPCA-2018.

Horizontal Integration

Backend Neural Network Libraries in Caffe2 Runtime

NNPACK

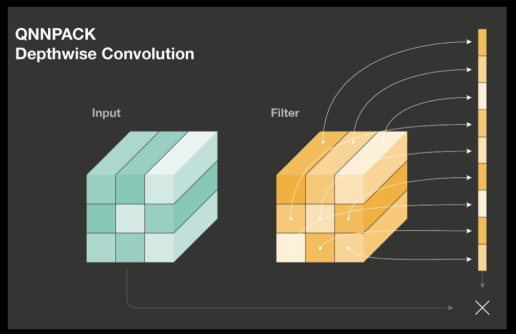
(32-BIT FLOATING POINT)

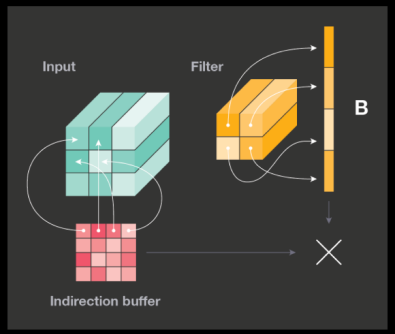
- Optimized convolution implementation using **Winograd** and **FFT**
- Best for NN with 3x3, 5x5 or larger convolutions

QNNPACK/QUANTIZED NNPACK

(8-BIT FIXED POINT)

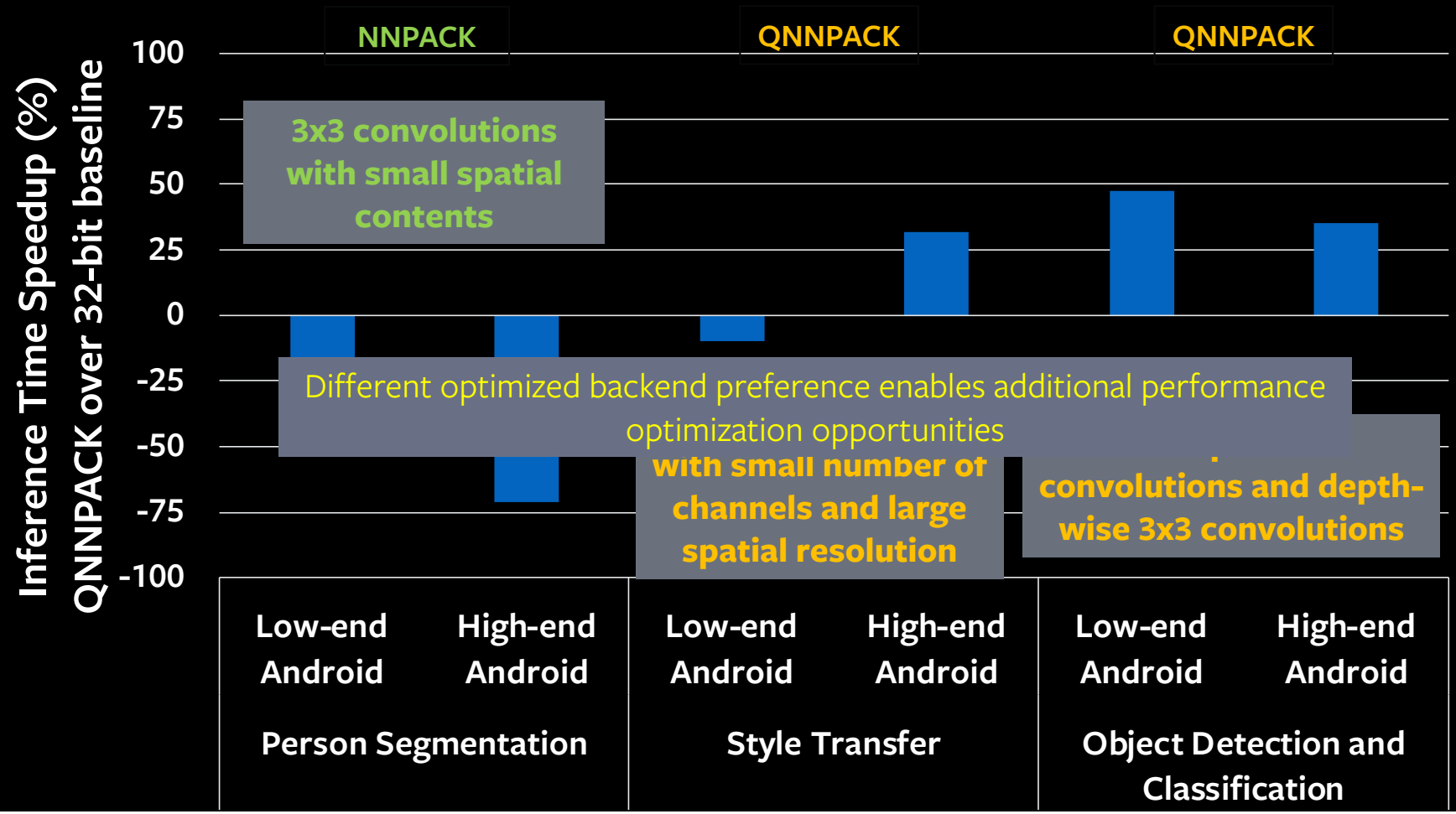
- Optimized direct convolution implementation
- Best for low-intensity convolutions
- Grouped, depth-wise, dilated convolutions
- Eliminate the overhead of im2col and other memory layout transformation





Horizontal Integration

QNNPACK Performance Evaluation



Introduction: Machine Learning @ FB & Unique Challenges for Edge Inference

Lay of the Land: Closer Look at Smartphones that FB Runs on

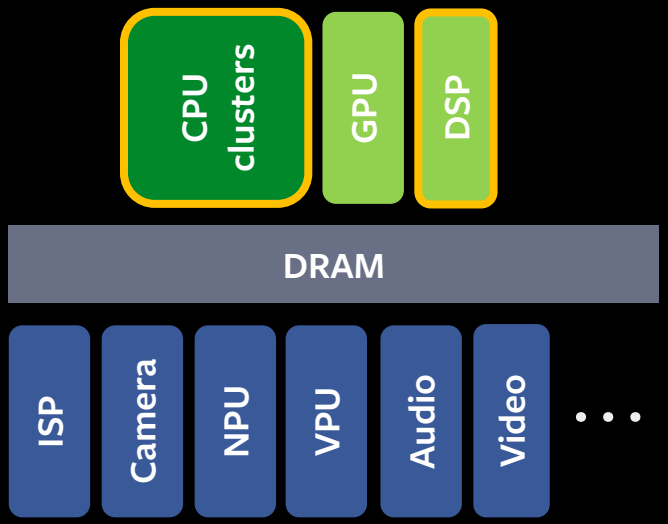
Horizontal Integration: Making Inference on Smartphones

Vertical Integration: Processing Inference for Oculus VR

Inference in the Wild: Performance Variability

Vertical Integrated Systems

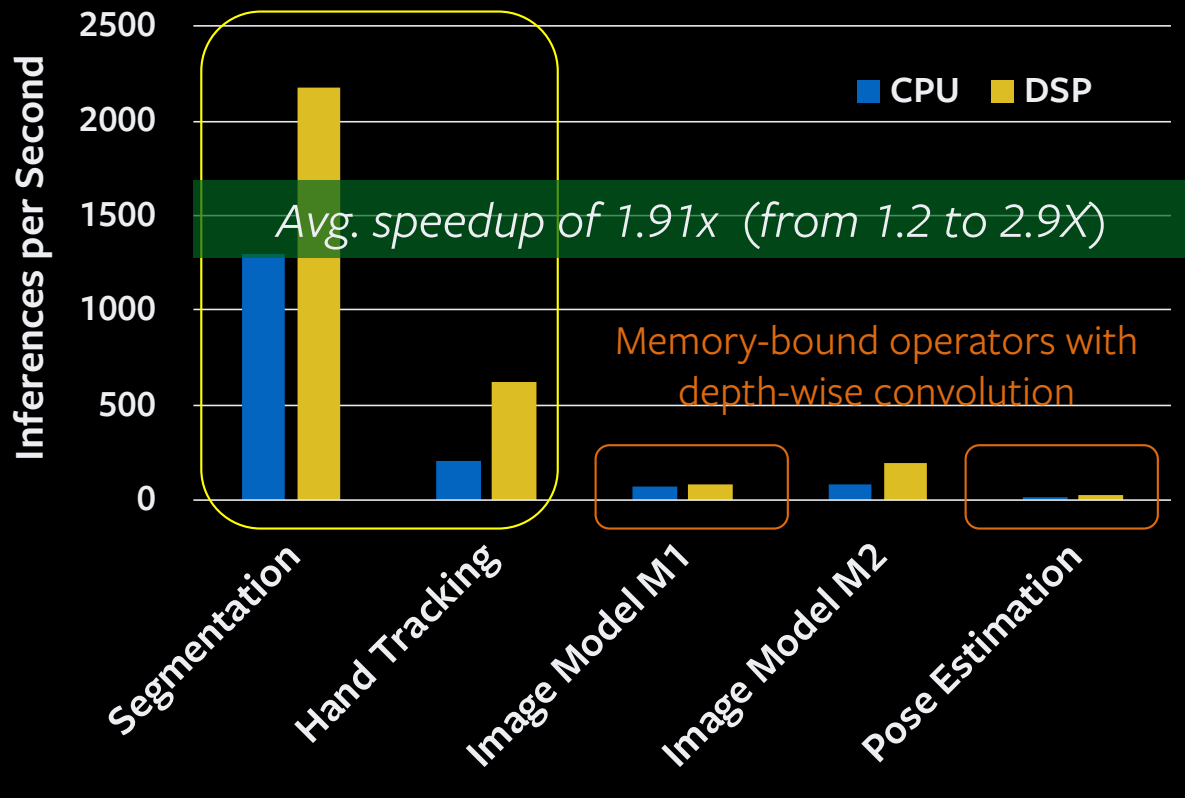
Processing Inference for Oculus VR



Vertical Integrated Systems

Performance Acceleration with Co-processors

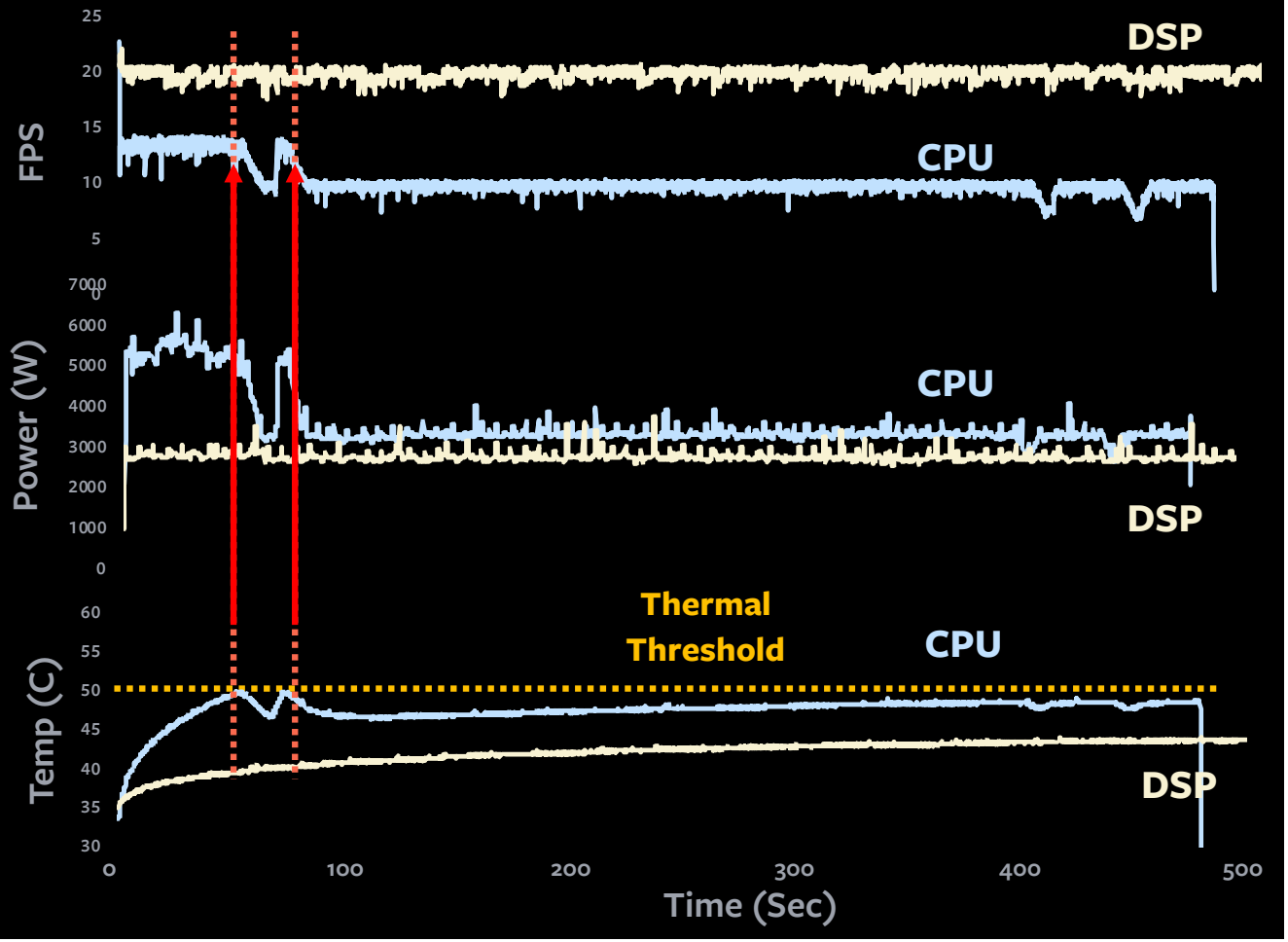
DNN Features	MACs	Weights
Segmentation	1X	1.5X
Hand Tracking	10X	1X
Image Model 1	10X	2X
Image Model 2	100X	1X
Pose Estimation	100X	4X



Vertical Integrated Systems

Making Inference on DSPs Leads to Consistent Performance

- CPU thermal throttling causes sudden **FPS drop**
- The primary reason for using co-processors and accelerators are for **lower power** and **more stable performance**



Introduction: Machine Learning @ FB & Unique challenges for Edge Inference

Lay of the Land: Closer look at smartphones that FB runs on

Horizontal Integration: Making Inference on Smartphones

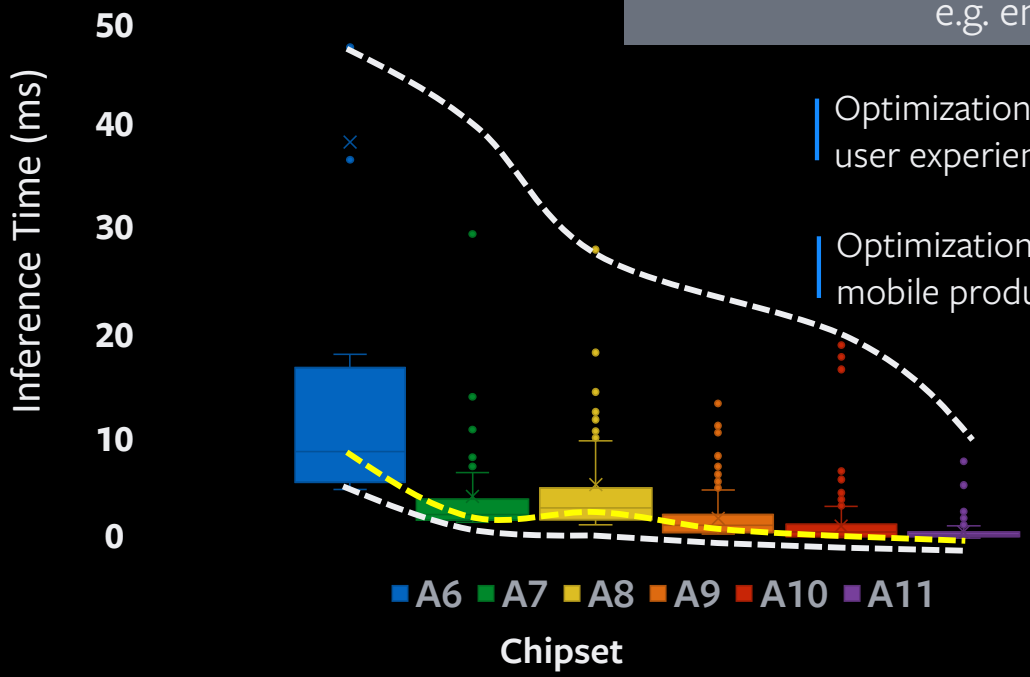
Vertical Integration: Processing Inference for Oculus VR

Inference in the Wild: Performance Variability

Inference in the Wild

Making "Efficient" Inference in the Wild Requires Developers to Deal with Performance Variability

Performance variability makes mobile design challenging for e.g. energy-aware scheduling



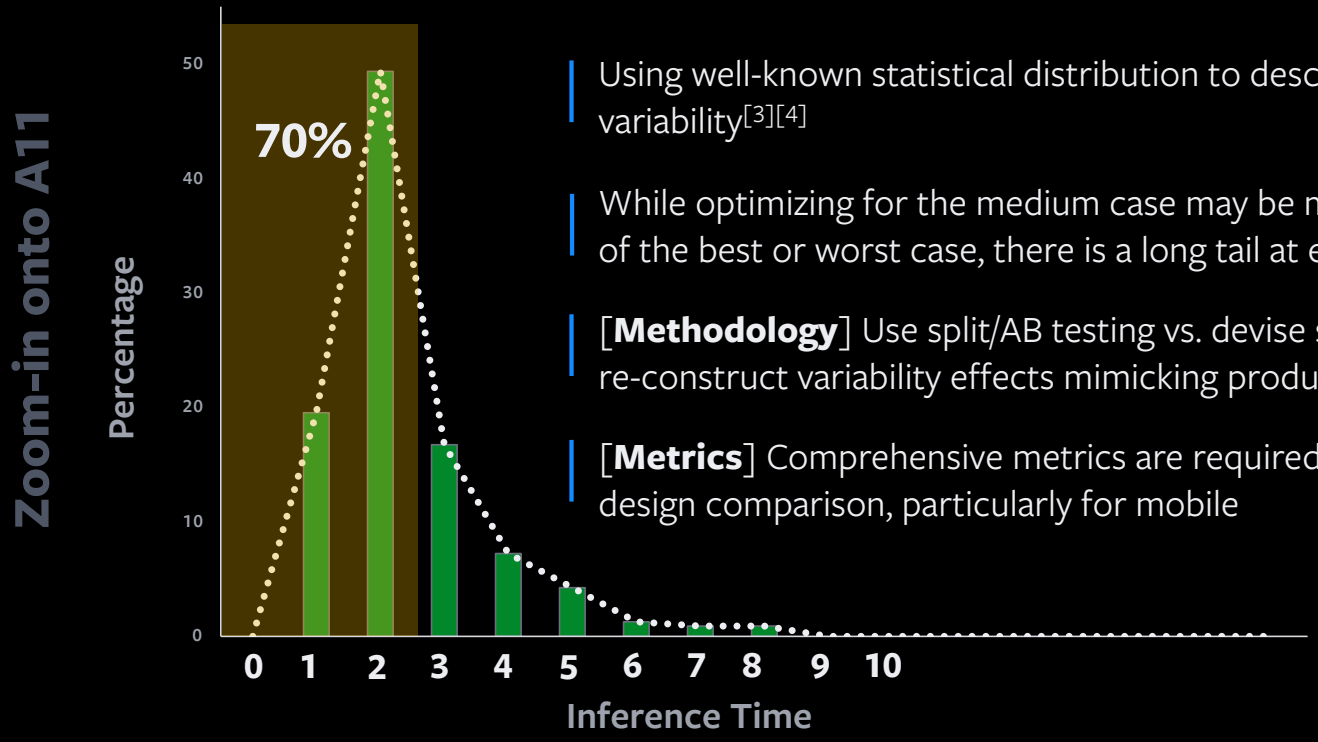
- Optimizations targeting the average or medium case sacrifices user experience
- Optimization must endure the high degree of variation in the mobile production environment

[3] Improving Smartphone User Experience by Balancing Performance and Energy with Probabilistic Guarantee. Gaudette, Wu, and Vrudhula, HPCA-2016.

IS THE PERFORMANCE VARIABILITY PATTERN PREDICTABLE?

Inference in the Wild

Does the Performance Variability Follow Certain Statistical Distributions?



- Using well-known statistical distribution to describe performance variability[3][4]
- While optimizing for the medium case may be more representative than that of the best or worst case, there is a long tail at each direction
- [**Methodology**] Use split/AB testing vs. devise systematic benchmarking to re-construct variability effects mimicking production environment is needed
- [**Metrics**] Comprehensive metrics are required for fair, representative design comparison, particularly for mobile

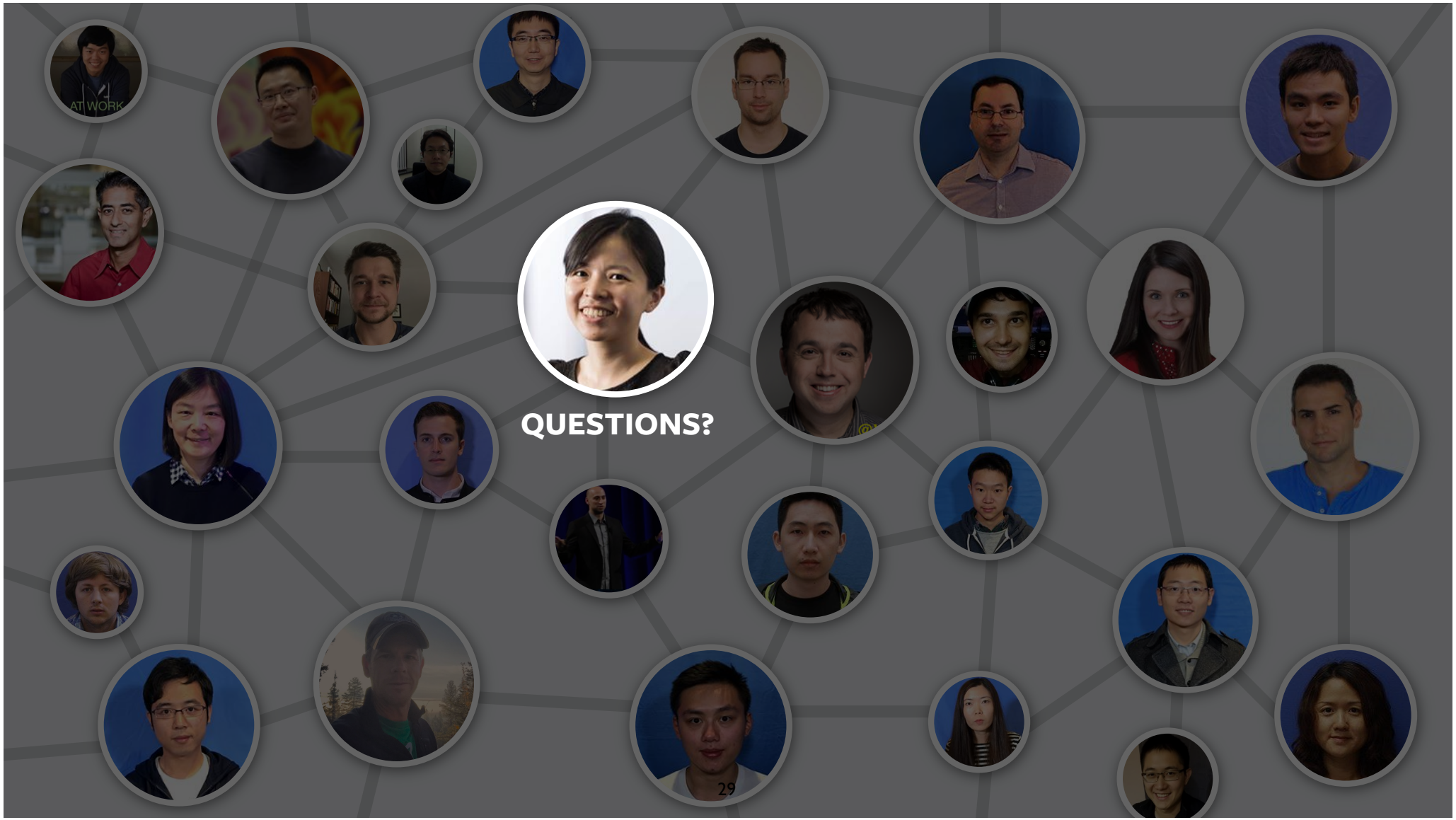
[3] Improving Smartphone User Experience by Balancing Performance and Energy with Probabilistic Guarantee. Gaudette et al. HPCA-2016.
[4] Optimizing User Satisfaction of Mobile Workloads Subject to Various Sources of Uncertainties. Gaudette et al. TMC-2018.

What did we learn?

It is important to consider full-picture and system effects for efficient, practical edge inference designs.

Data-driven approach to summarize the state of the industry practice:

- Lay of the land for mobile SoCs is extremely heterogeneous.
- Majority of mobile inference run on CPUs, that are the current generation.
- Performance difference between a mobile CPU and GPU/DSP is not 100×.
- Inference performance varies much more widely in the field.
- Co-processors and accelerators are used for energy efficiency and stable performance; speedup is often secondary.


QUESTIONS?

Thank you

facebook