Machine Learning @ Facebook

Understanding Inference at the Edge

Brandon Reagen

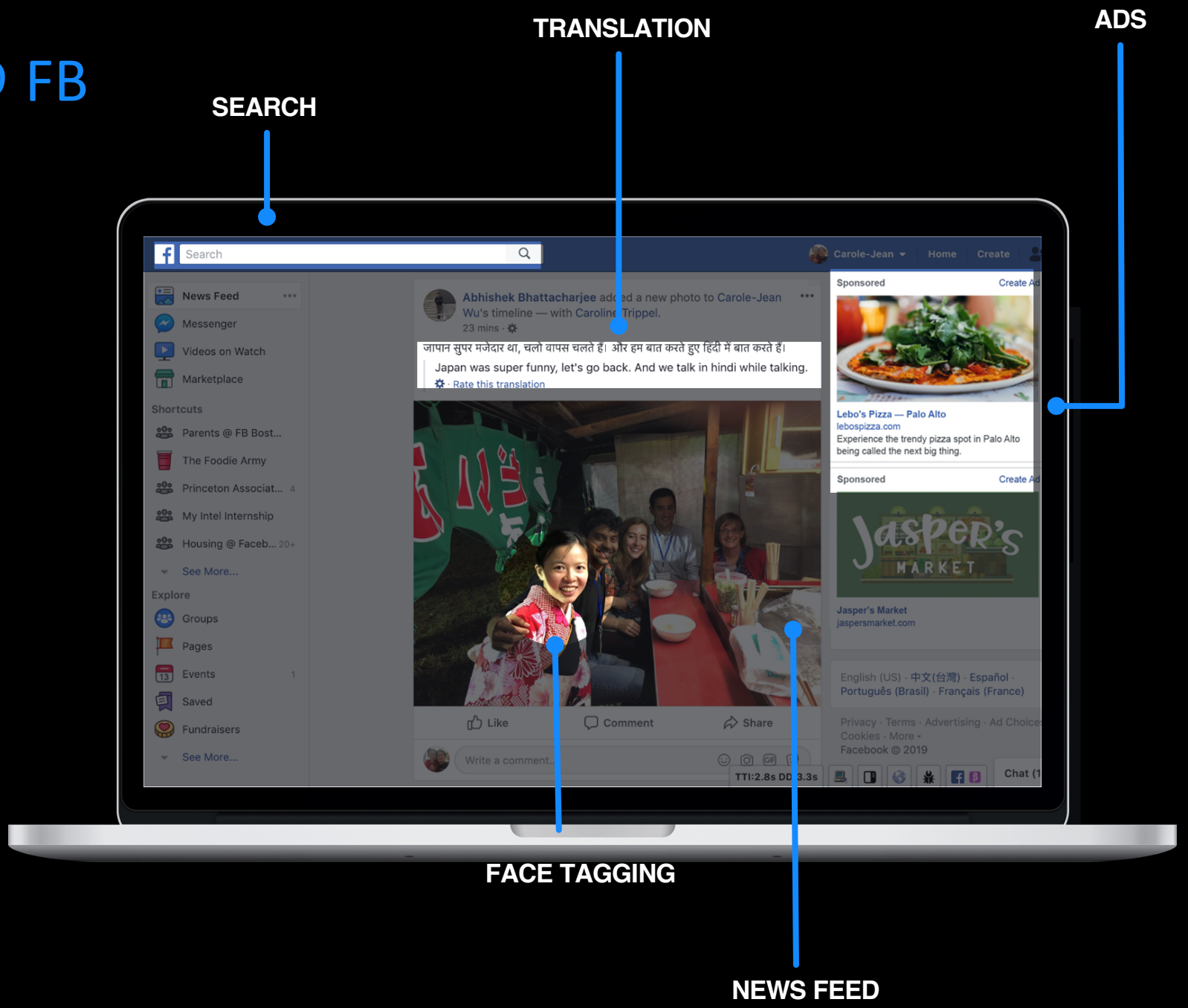
Research Scientist @ FAIR

AMLD, Switzerland 2020

Machine Learning @ FB

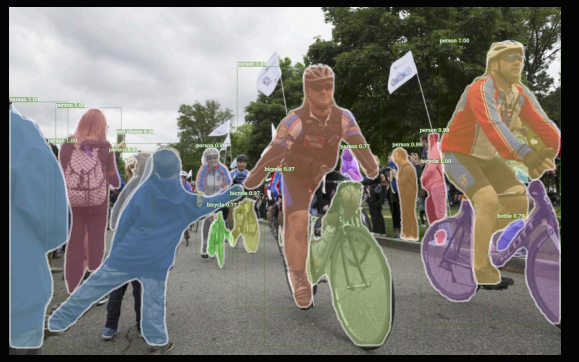
- Ranking of posts in news feeds
- Object detection, segmentation, and classification
- Speech recognition / translation

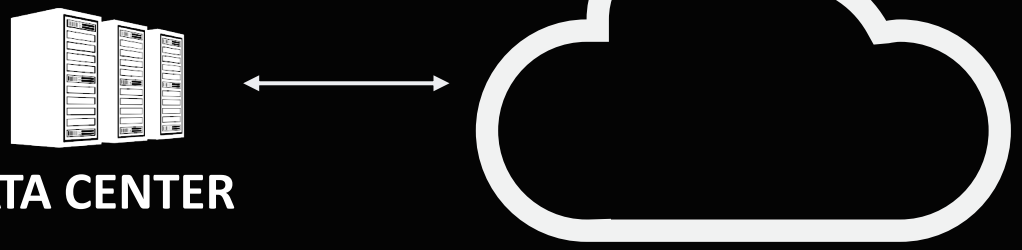
High model diversity
Large request volume



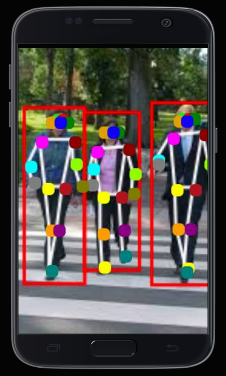
From data centers to the edge

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





Keypoints Segmentation

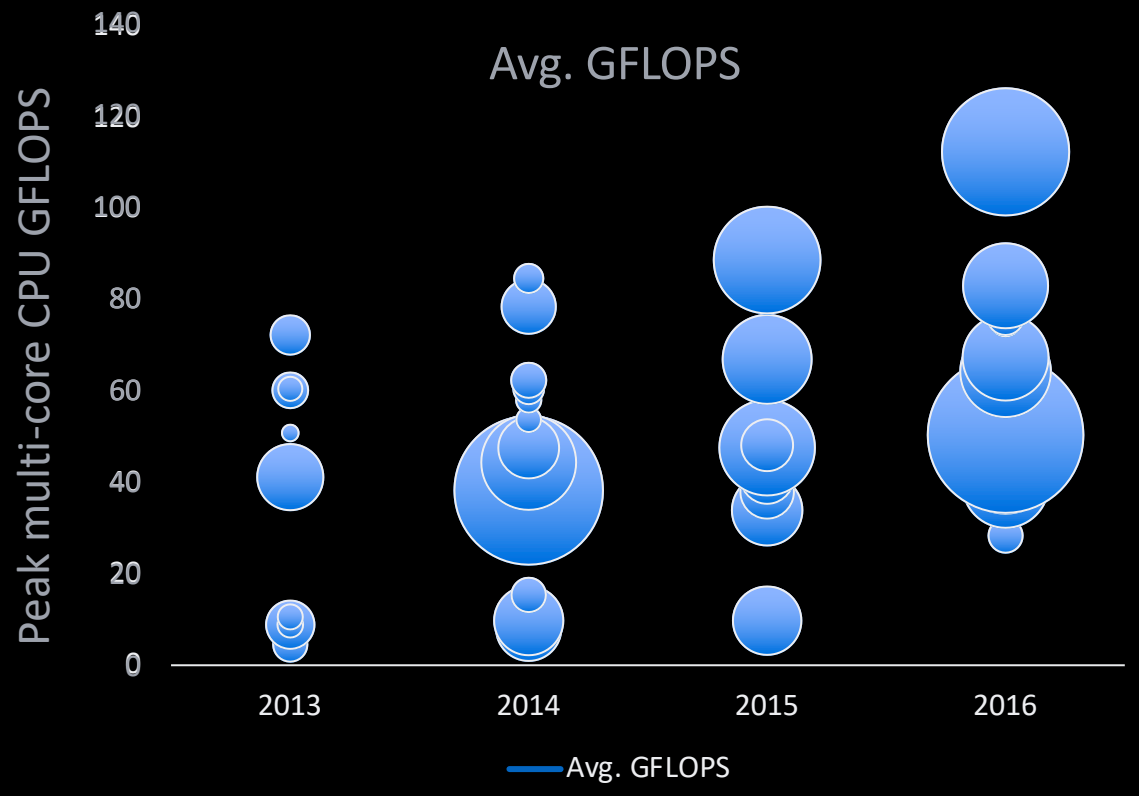


Augmented Reality with Smart Camera

Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective. Hazelwood et al. HPCA '18.
The Architectural Implications of Facebook's DNN-based Personalized Recommendation. Gupta et al. HPCA '20

Challenges of Complex Models on Constrained Edge Devices

- 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 widespread



DELIVERING CONSISTENT INFERENCE PERFORMANCE IS CHALLENGING

Unique Challenges for Edge vs Cloud

Diversity of Mobile Hardware and Software is Not Found in the Cloud

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

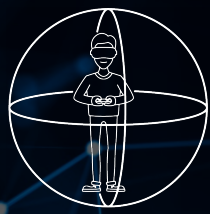
How do we optimize system designs for real-time ML inference?

FRAGMENTED ECOSYSTEM MAKES PER-SYSTEM OPTIMIZATION HARD

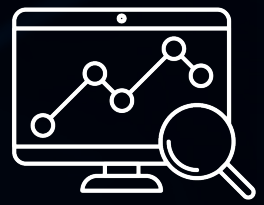
Introduction:
Machine Learning @ FB & Unique Challenges for Edge Inference

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



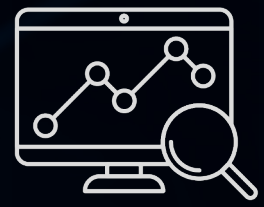
Inference in the Wild:
Performance Variability

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

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

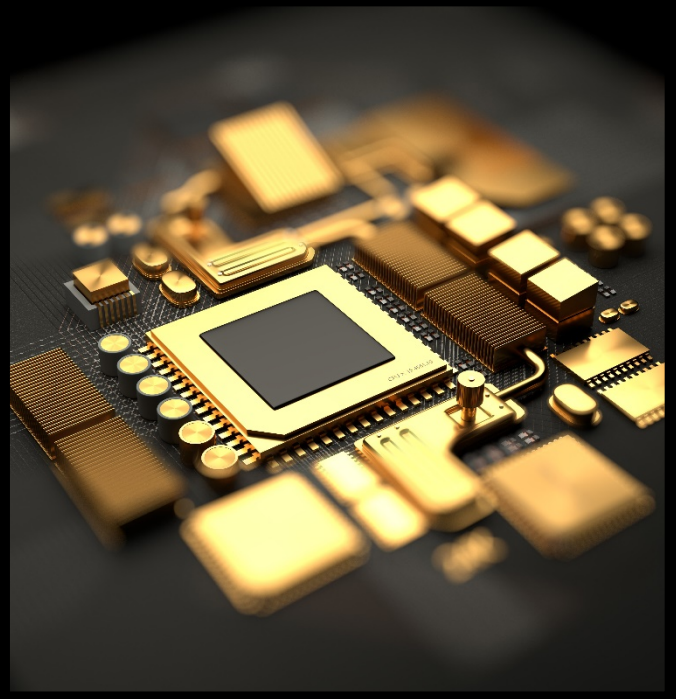
Horizontal Integration:
Making Inference on Smartphones

Vertical Integration:
Processing Inference for Oculus VR



Inference in the Wild:
Performance Variability

Understanding the Mobile Landscape



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

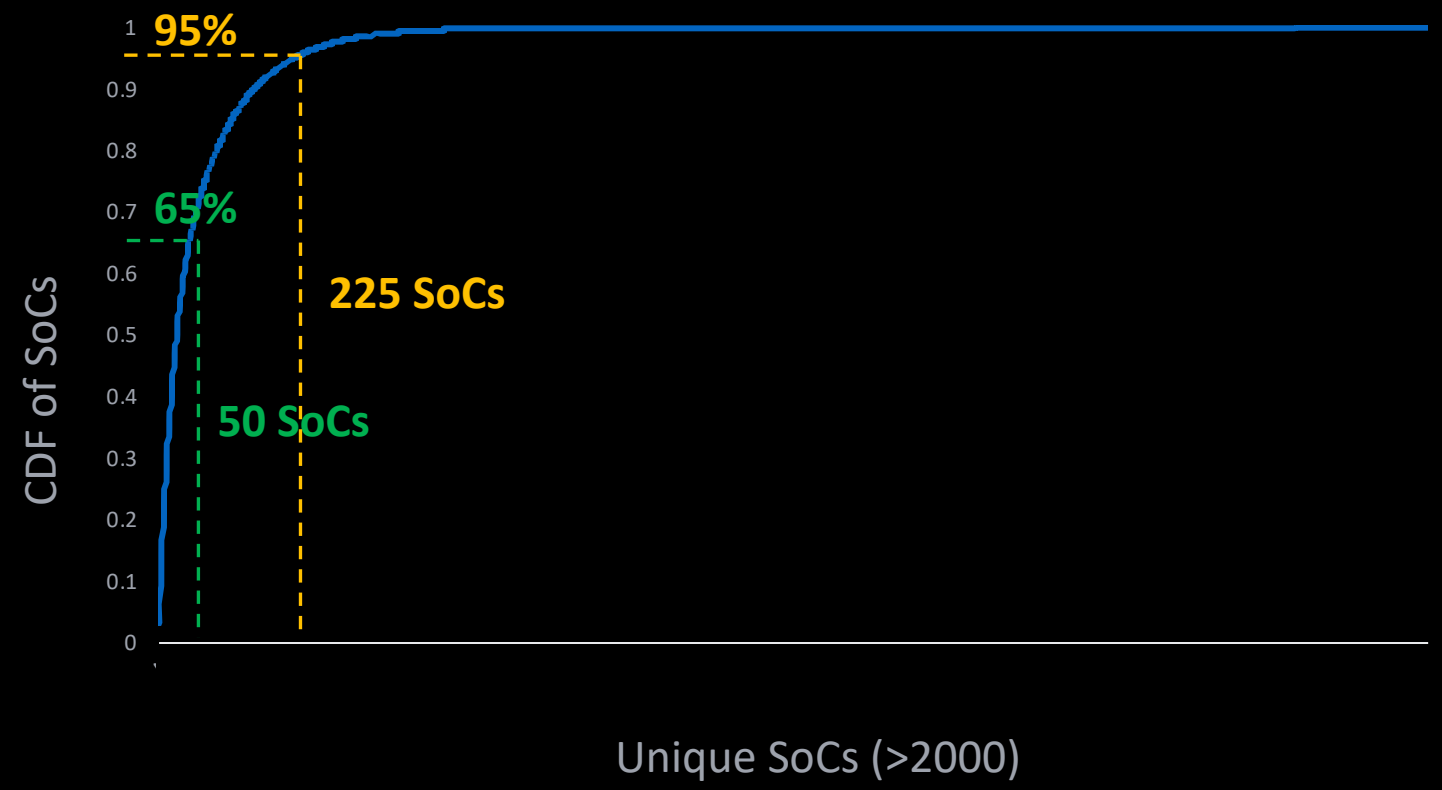
Programmabilit y

Programmability is a Primary Roadblock for Using Mobile Co-

Lay of the Land

FRAGMENTATION

Can we optimize for the common case?



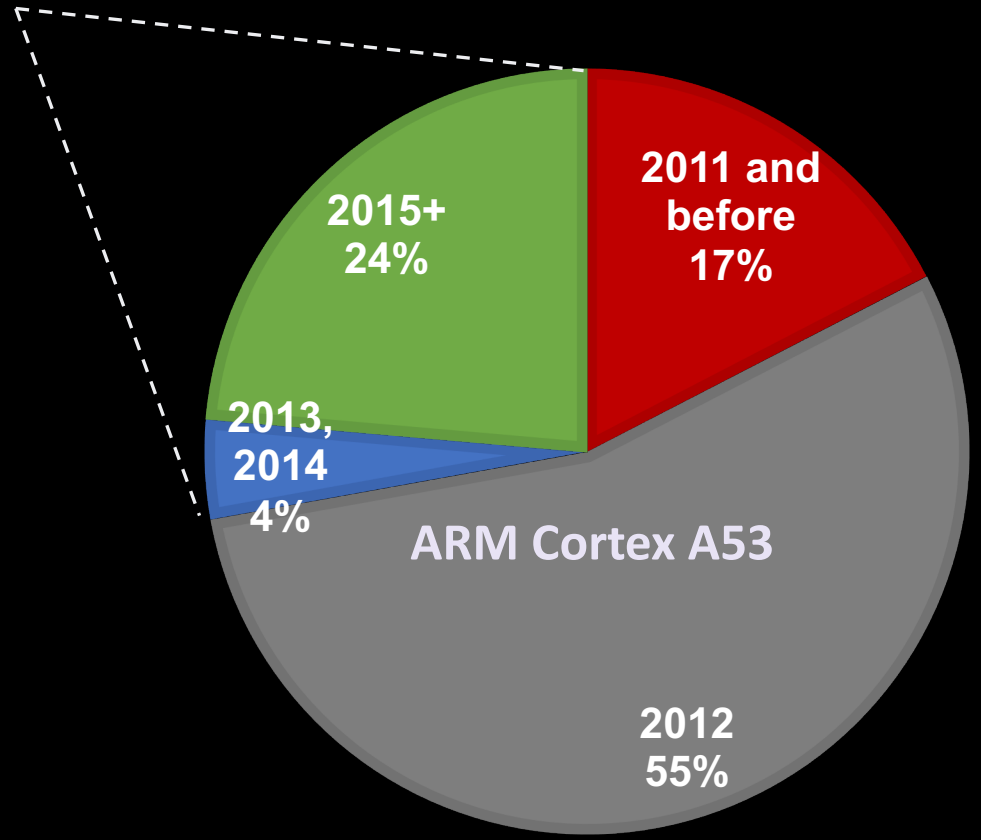
- Qualcomm Snapdragon
- Samsung Exynos
- MediaTek Helio
- HiSilicon Kirin et al.

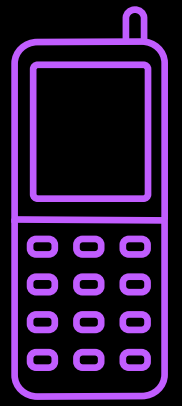
THERE IS NO STANDARD SOC TO OPTIMIZE FOR

Lay of the Land: CPU Cores

FRAGMENTATION

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





72%

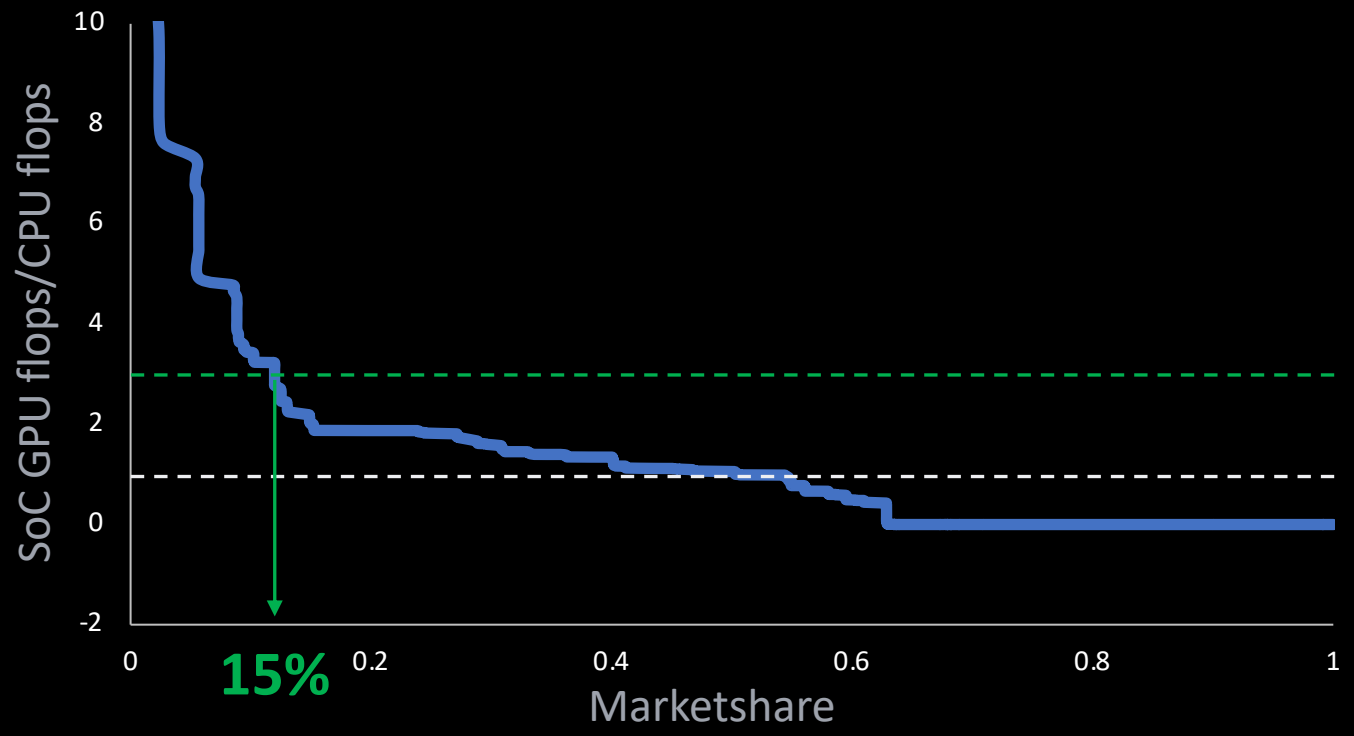
OF THE WORLD'S CELL PHONES ARE MORE THAN 7 YEARS OLD

MOBILE CPUS SHOW LITTLE DIVERSITY, BUT ARE DATED AND FEW HAVE LATEST DEVICES

Lay of the Land: Achieving High Perf. With Accelerators

PERFORMANCE

The Performance Difference between a Mobile CPU and GPU is Narrow



ON MOST SMARTPHONES, THE GPU PROVIDES AS MUCH THEORETICAL PEAK PERFORMANCE AS ITS CPU

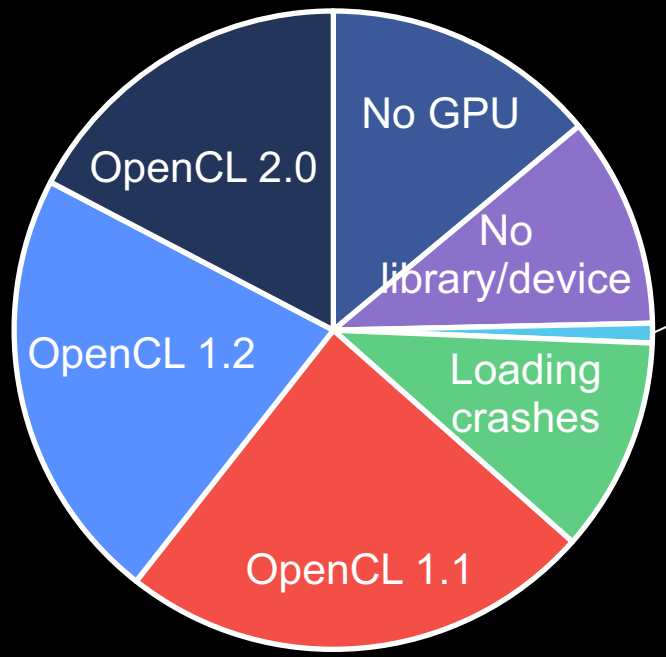
LESS THAN 15% SMARTPHONES HAVE A GPU THAT IS 3 TIMES AS POWERFUL AS ITS CPU

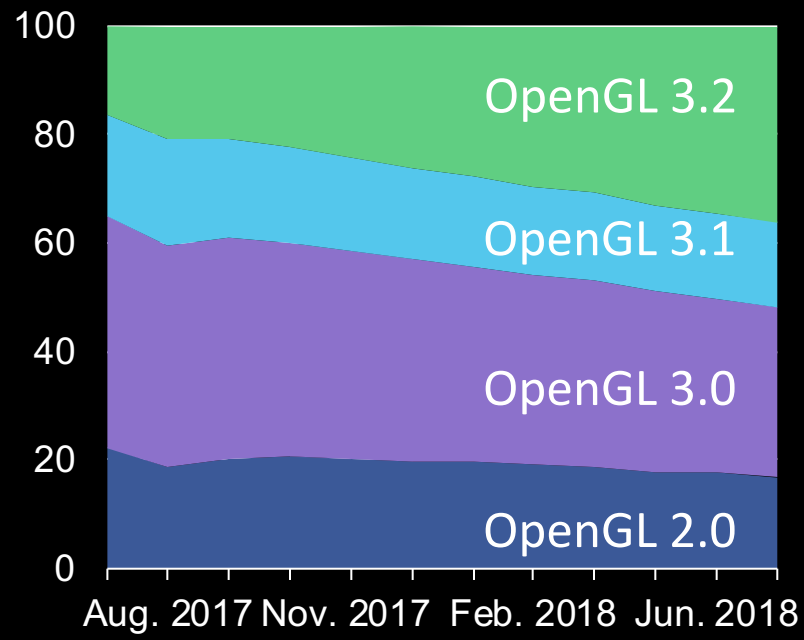
Lay of the Land: Programmability

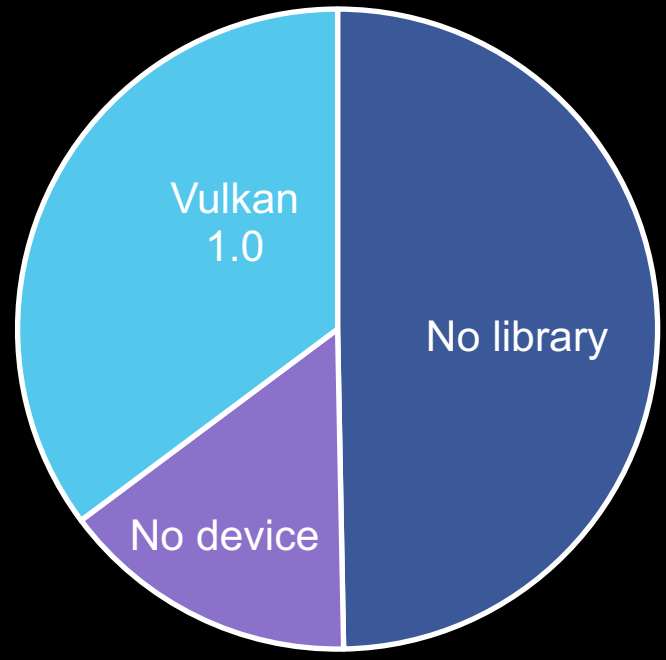
PROGRAMMABILITY

Programmability is a Primary Roadblock for Using Mobile Co-processors

- OpenCL, OpenGL ES, Vulkan for Android GPUs



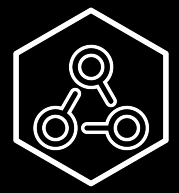




ANDROID GPUS HAVE FRAGILE USABILITY AND POOR PROGRAMMABILITY
WHILE IOS HAS BETTER SUPPORT WITH METAL

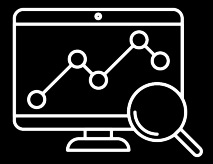
Quantitative Approach to Mobile Inference Designs

State of the Practice for Mobile Inference is CPUs



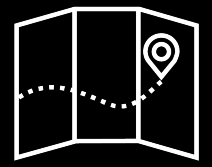
FRAGMENTATION

Over **2000+ different SoCs**
Mobile CPUs show little diversity
ARM's A53 dominates market



PERFORMANCE

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



PROGRAMMABILITY

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

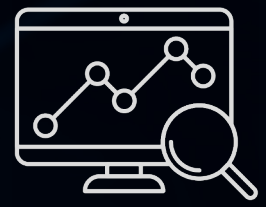
MOBILE INFERENCE OPTIMIZATION IS TARGETED FOR THE COMMON DENOMINATOR OF THE FRAGMENTED SOC ECOSYSTEM

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

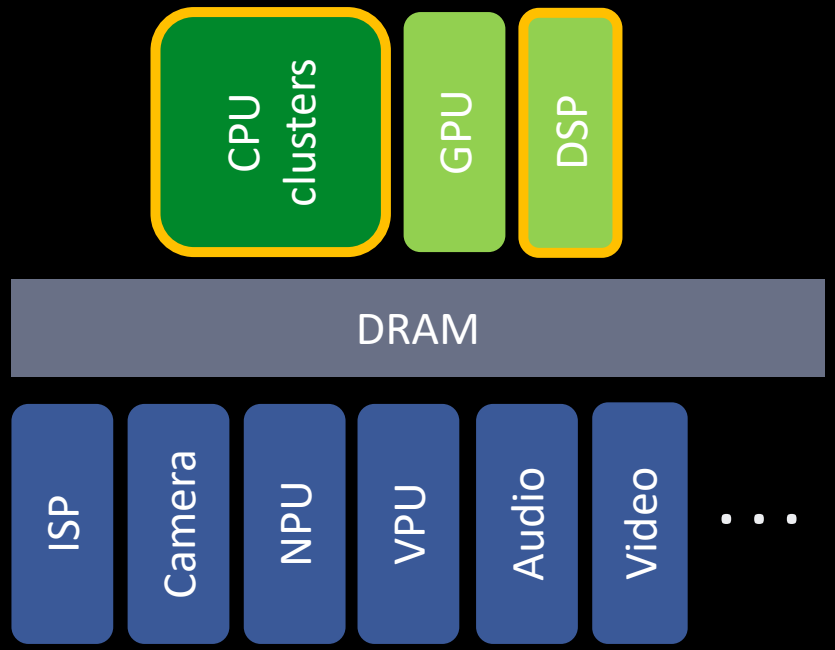
Vertical Integration: Processing Inference for Oculus VR



Inference in the Wild: Performance Variability

Vertical Integrated Systems

Processing Inference for Oculus VR



ML for image and tracking

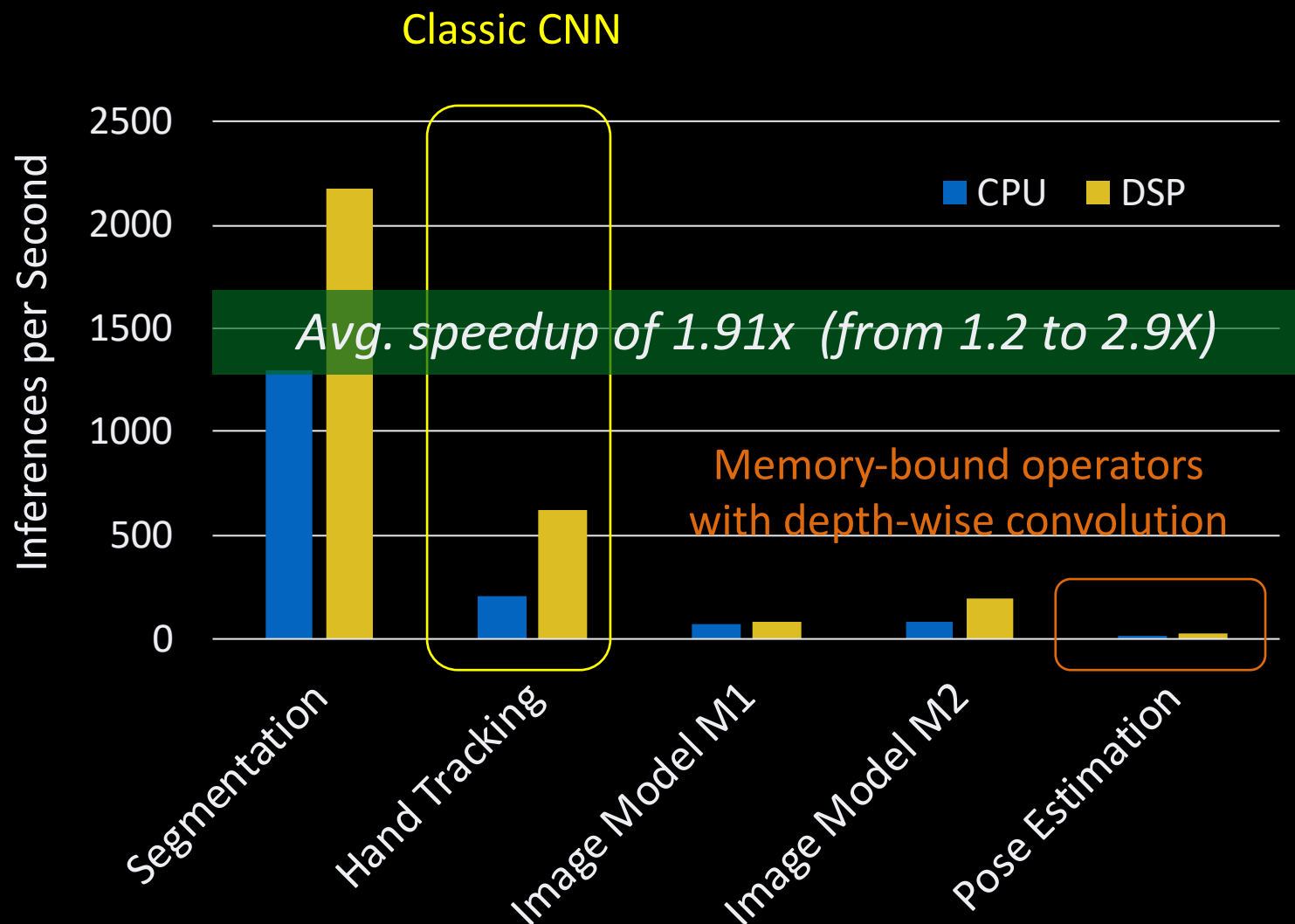
VR 30-60 FPS

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

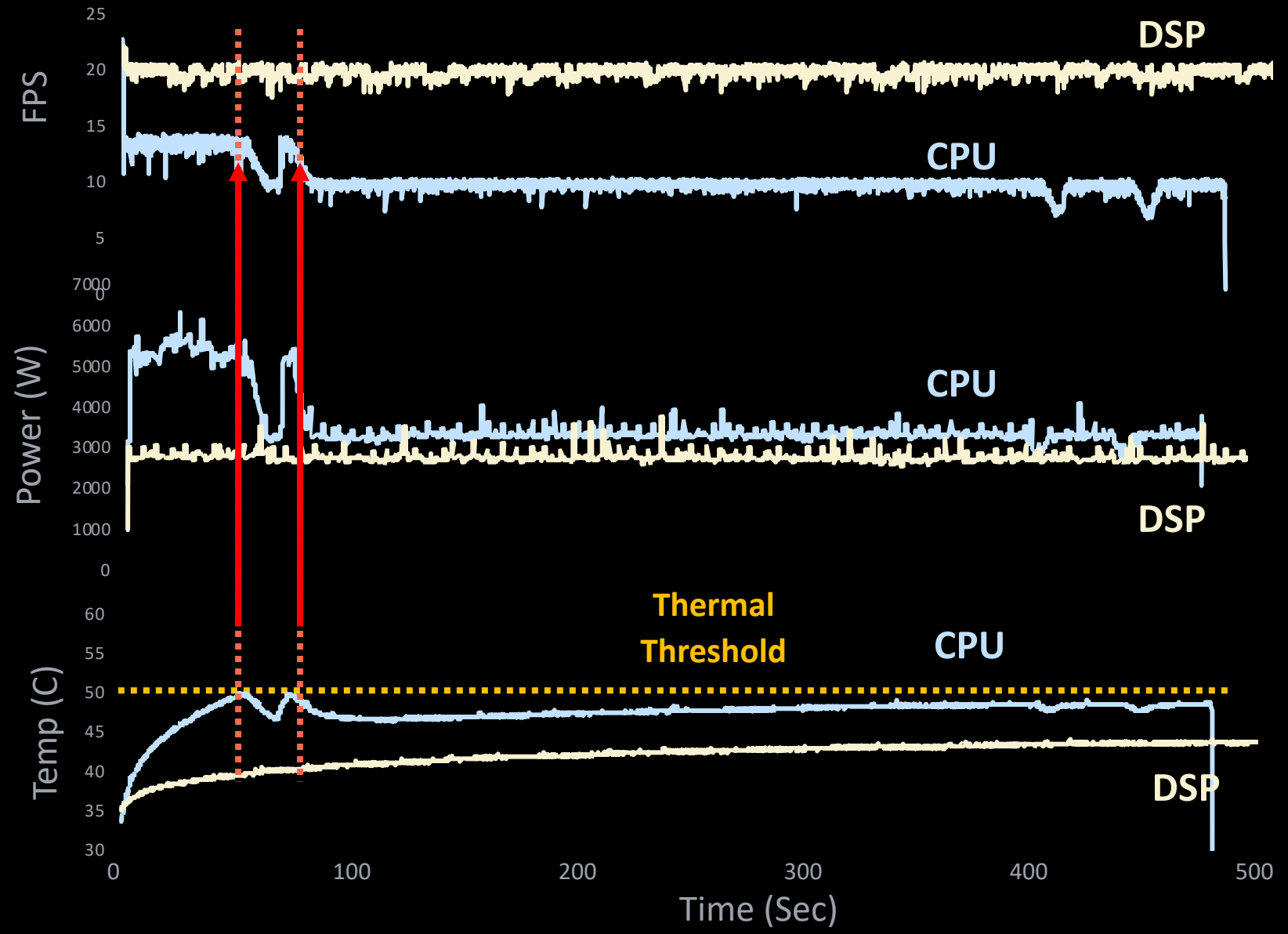
Co-processor speedup
2x not 10x (or 100x)



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

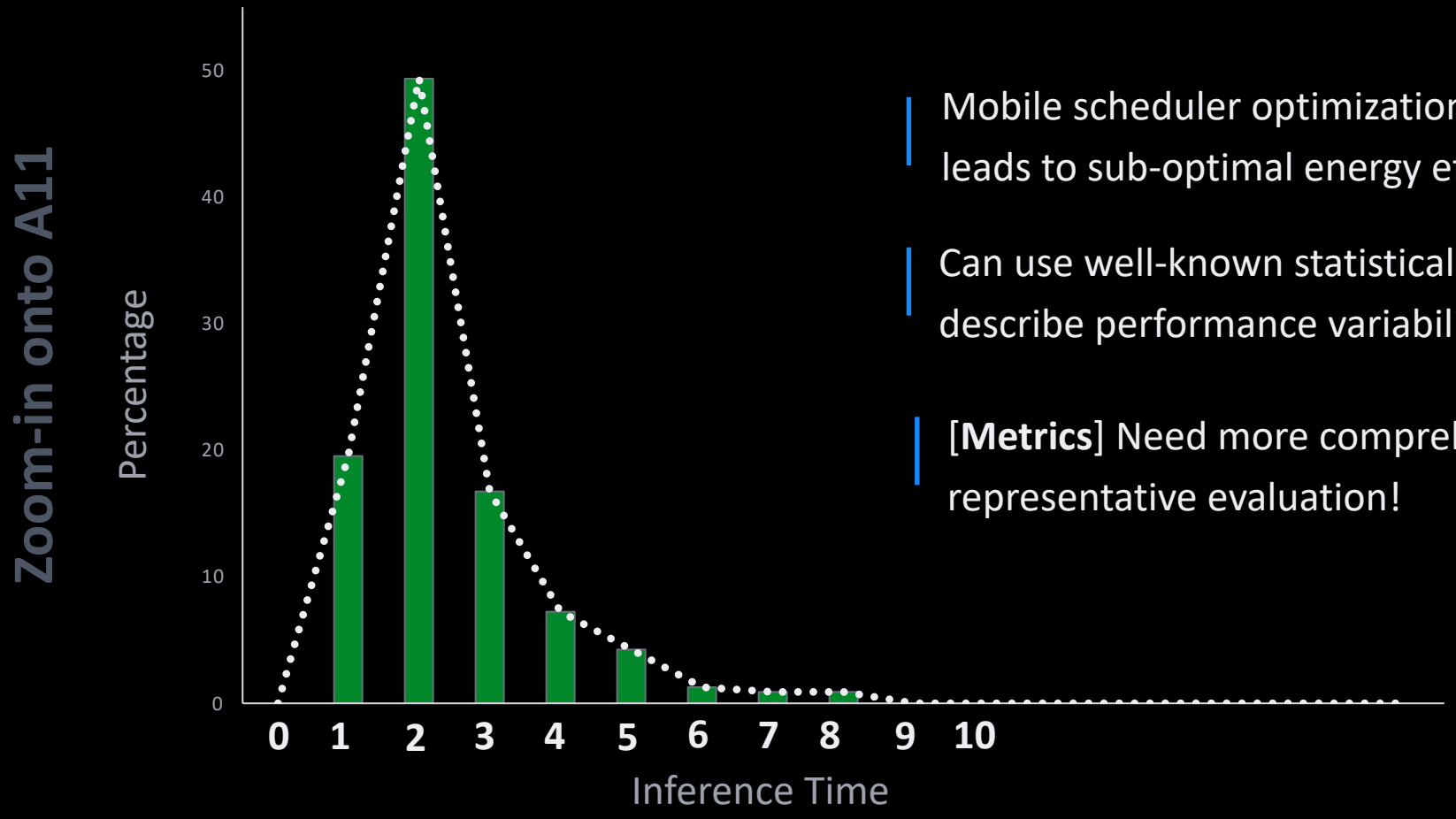
Horizontal Integration: Making Inference on Smartphones

Vertical Integration: Processing Inference for Oculus VR

Inference in the Wild: Performance Variability

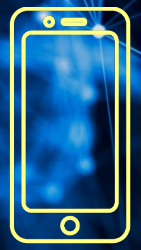
Inference in the Wild

Find Performance Variability in Same Layer and Device



- Mobile scheduler optimizations for the best case leads to sub-optimal energy efficiency [3]
- Can use well-known statistical distribution to describe performance variability[3][4]
- [**Metrics**] Need more comprehensive metrics for fair, representative evaluation!

[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.


Introduction: Machine Learning @ FB & Unique challenges

Lay of the Land: Closer look at smartphones that FB

Horizontal Integration: Making Inference on Smartphones

Vertical Integration: Processing Inference for Oculus VR

Inference in the Wild: Performance Variability

Machine Learning at Facebook: Understanding Inference at the Edge

Carole-Jean Wu, David Brooks, Kevin Chen, Douglas Chen, Sy Choudhury, Marat Dukhan, Kim Hazelwood, Eldad Isaac, Yangqing Jia, Bill Jia, Tommer Leyvand, Hao Lu, Yang Lu, Lin Qiao, Brandon Reagen, Joe Spisak, Fei Sun, Andrew Tulloch, Peter Vajda, Xiaodong Wang, Yanghan Wang, Bram Wasti, Yiming Wu, Ran Xian, Sungjoo Yoo*, Peizhao Zhang

Facebook, Inc.

Thank you

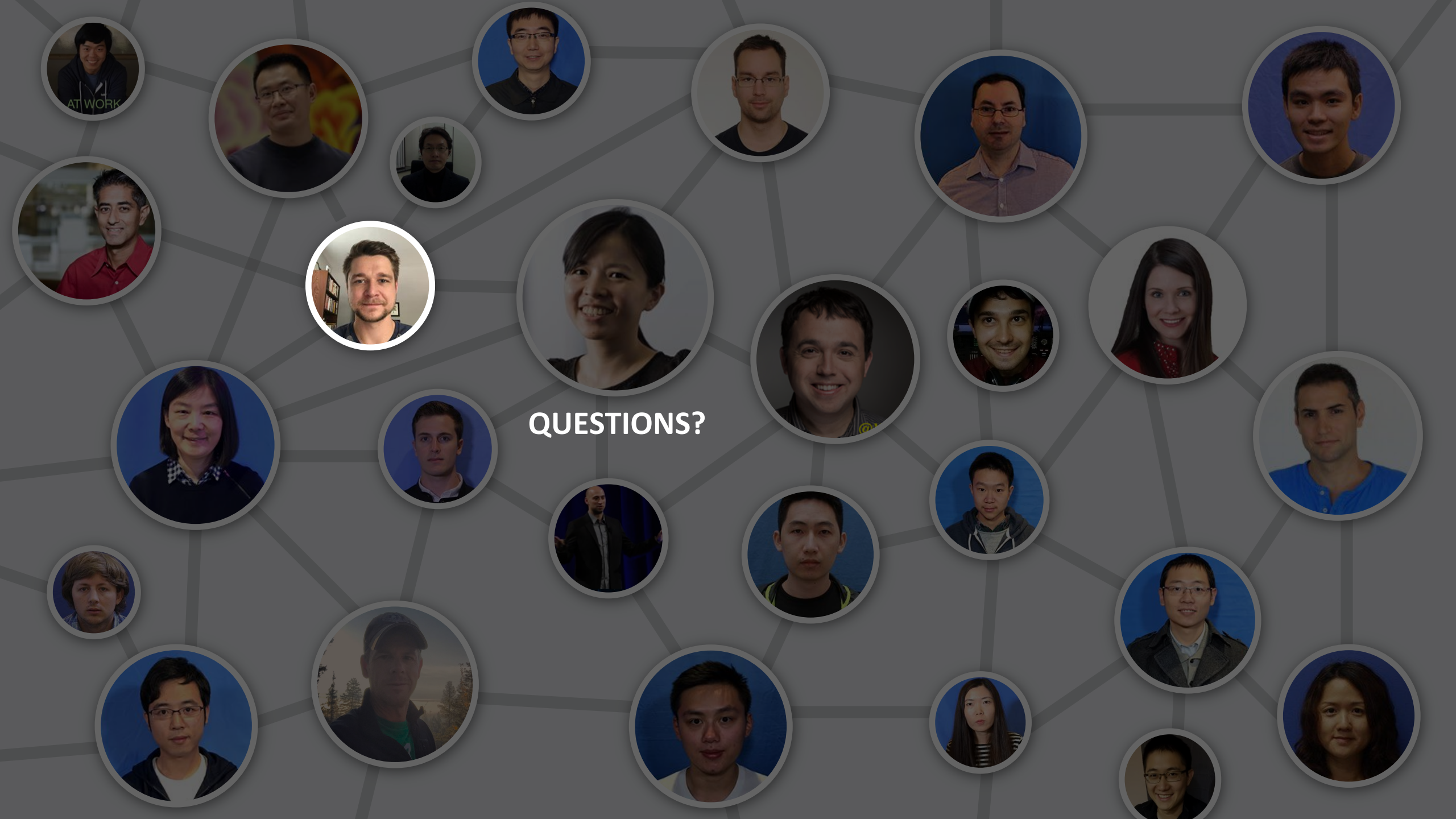
facebook

Question?

It is important to consider full-picture and system effects for efficient edge inference

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

- Lay of the land for mobile SoCs is extremely heterogeneous
- Majority of mobile inference run on CPUs
- Performance difference between a mobile CPU and GPU/DSP is not 100×
- Inference performance varies in the field.
- Co-processors are used for power and stable performance; speedup can be secondary.


QUESTIONS?