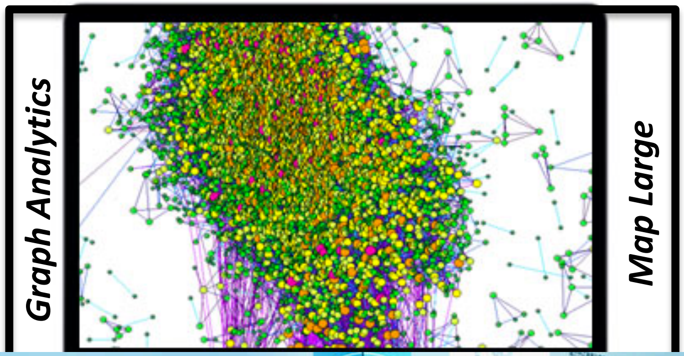
EMC^2 Workshop @ ISCA June 23, 2019

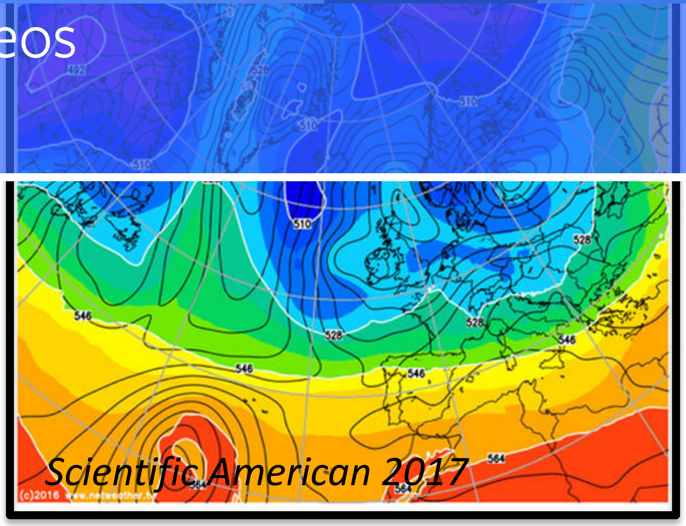
Machine Learning @ Scale

Understanding Inference at the Edge

Carole-Jean Wu
AI INFRA RESEARCH, FACEBOOK

Mobile Industry News





Every minute,

- Weather channel receives 18 million requests
- Google delivers 3.6 million searches
- Wikipedia users publish 600 new edits
- YouTube users watch 4.1 million videos

→ Over 3 million GB image data

Machine Learning at Facebook

Ranking of posts in news feeds

Content understanding

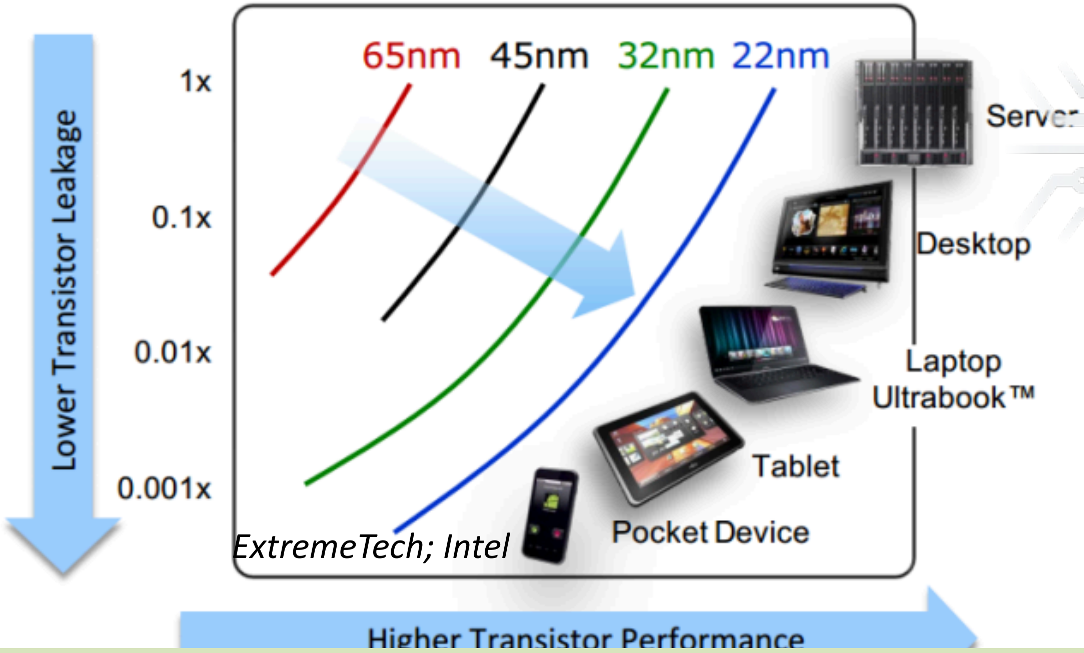
Object detection, segmentation, and tracking

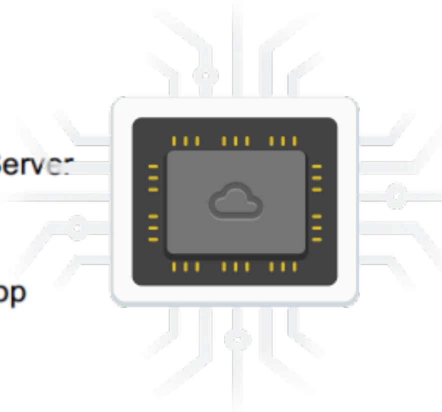
Speech recognition / translation

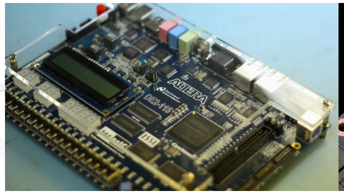
And Many More!

- Objectionable content detection
- Fraudulent account detection
- Content integrity
- Sentiment analysis

Deep Learning is Fueling the Hardware Renaissance







- **Exascale computing**
- **High energy efficiency**

CPU

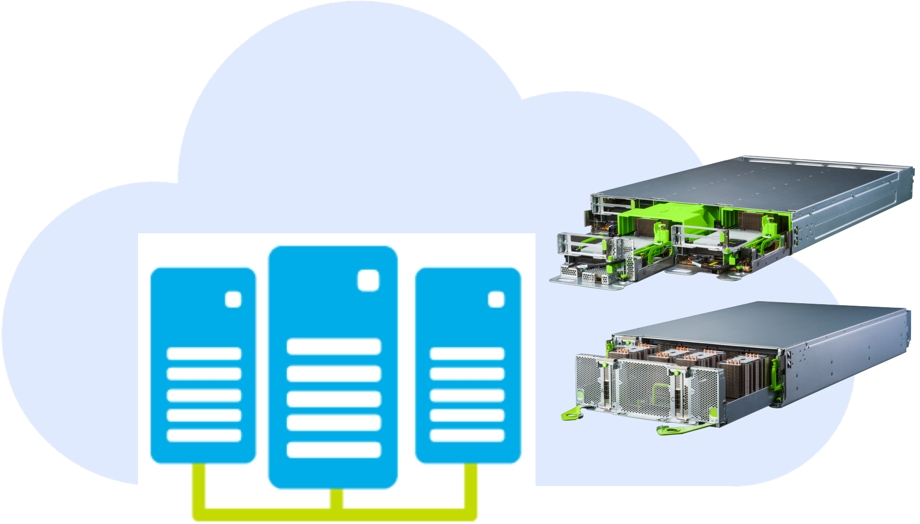
[**MICRO-2011**] C.-J. Wu, A. Jaleel, M. Martonosi, S. Steely Jr., and J. Emer, "PACMan: Prefetch-Aware Cache Management for High Performance Caching."
[**MICRO-2011**] C.-J. Wu, A. Jaleel, W. Hasenplaugh, M. Martonosi, S. Steely Jr., and J. Emer, "SHiP: Signature-Based Hit Predictor for High Performance Caching."

GPU

[**PACT-2014**] S.-Y. Lee and C.-J. Wu, "CAWS: Criticality-Aware Warp Scheduling for GPGPU Workloads."
[**ISCA-2015**] S.-Y. Lee, A. Arunkumar, and C.-J. Wu, "CAWA: Coordinated Warp Scheduling and Cache Prioritization for Critical Warp Acceleration for GPGPU Workloads."
[**ISCA-2017**] A. Arunkumar et al., "MCM-GPU: Multi-Chip-Module GPUs for Continued Performance Scalability."
[**HPCA-2018**] A. Arunkumar, S.-Y. Lee, V. Soundararajan, and C.-J. Wu, "LATTE-CC: Latency Tolerance Aware Adaptive Cache Compression Management for Energy Efficient GPUs."
[**HPCA-2019**] A. Arunkumar, E. Bolotin, D. Nellans, and C.-J. Wu, "Understanding the Future of Energy Efficiency in Multi-Module GPUs."

From Cloud to the Edge

- Minimizing network bandwidth
- Reducing response latency
- Improving user data privacy
- Exploiting features available only at the edge



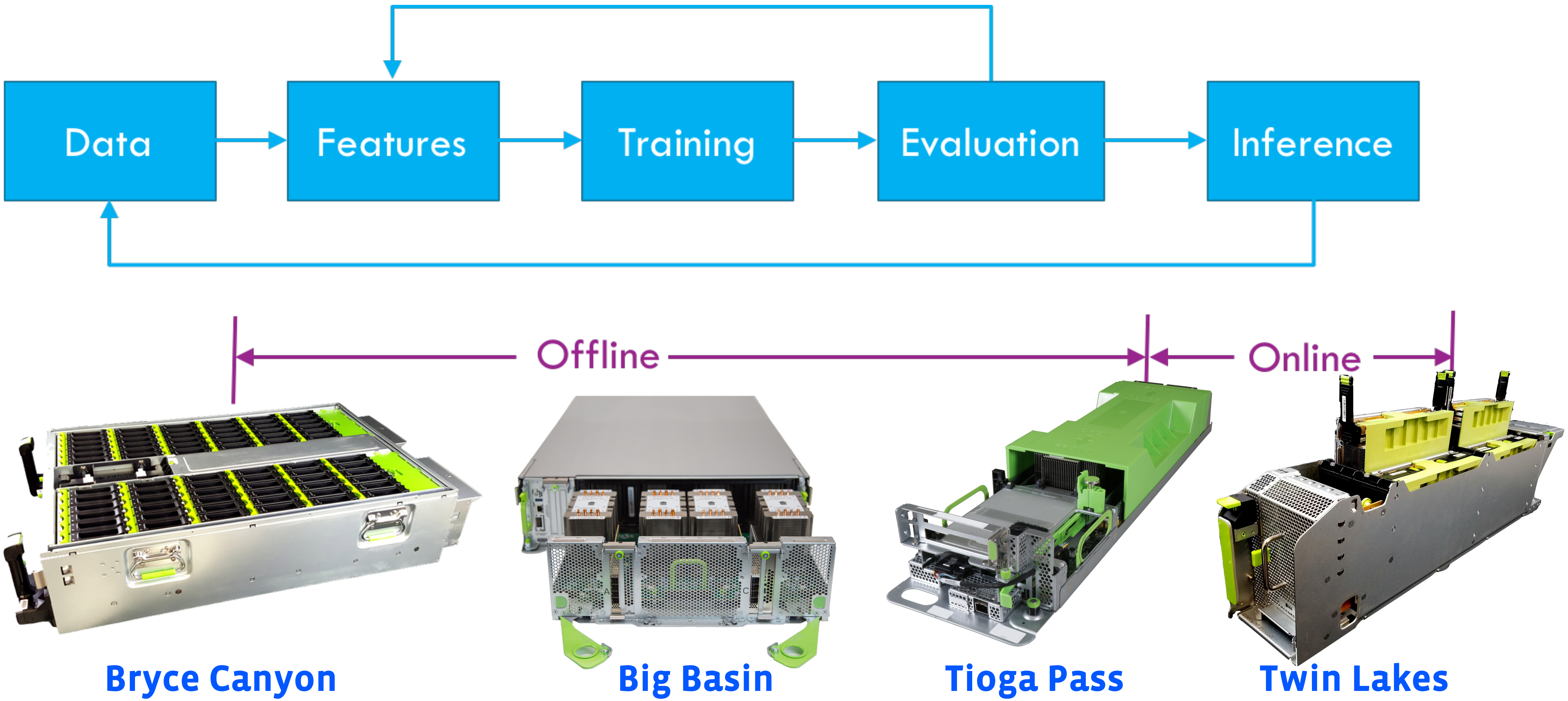
Keypoints Segmentation

Augmented Reality with Smart Camera

K. Hazelwood et al., "**Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective**", HPCA 2018.

C.-J. Wu et al., "**Machine Learning at Facebook: Understanding Inference at the Edge**", HPCA 2019.

Facebook Machine Learning Execution Flow



What We Are Doing at AI Infrastructure Research

Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective

Kim Hazelwood, Sarah Bird, David Brooks, Soumith Chintala, Utku Diril, Dmytro Dzhulgakov, Mohamed Fawzy, Bill Jia, Yangqing Jia, Aditya Kalro, James Law, Kevin Lee, Jason Lu, Pieter Noordhuis, Misha Smelyanskiy, Liang Xiong, Xiaodong Wang

Facebook, Inc.

[Hazelwood, HPCA’18]

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

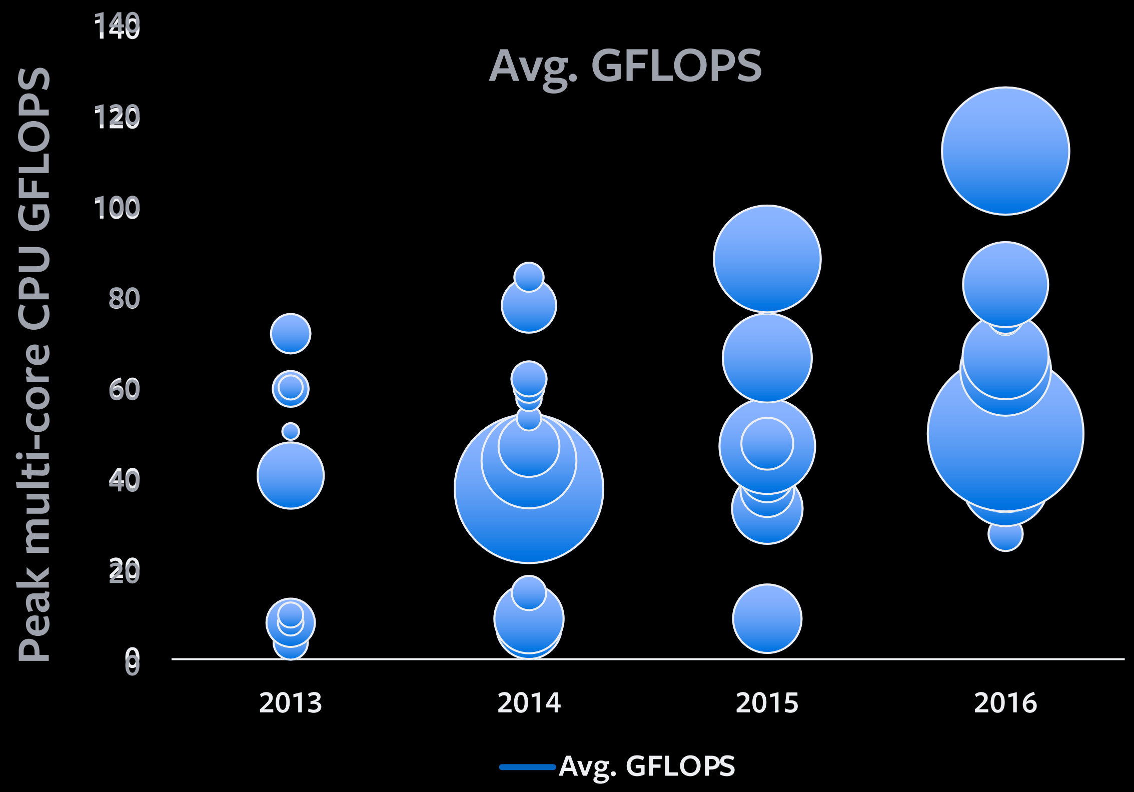
Facebook, Inc.

[Wu, HPCA’19]

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

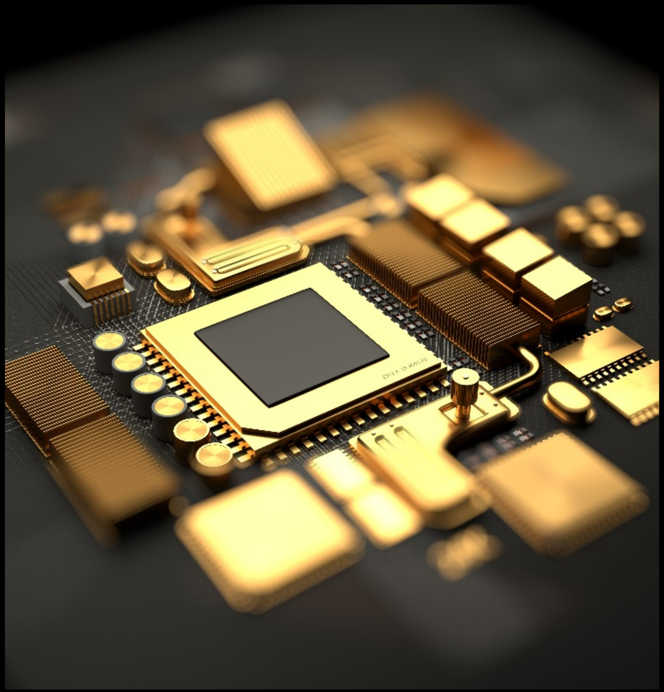
Lay of the Land:
Closer Look at Smartphones that FB Runs on

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?

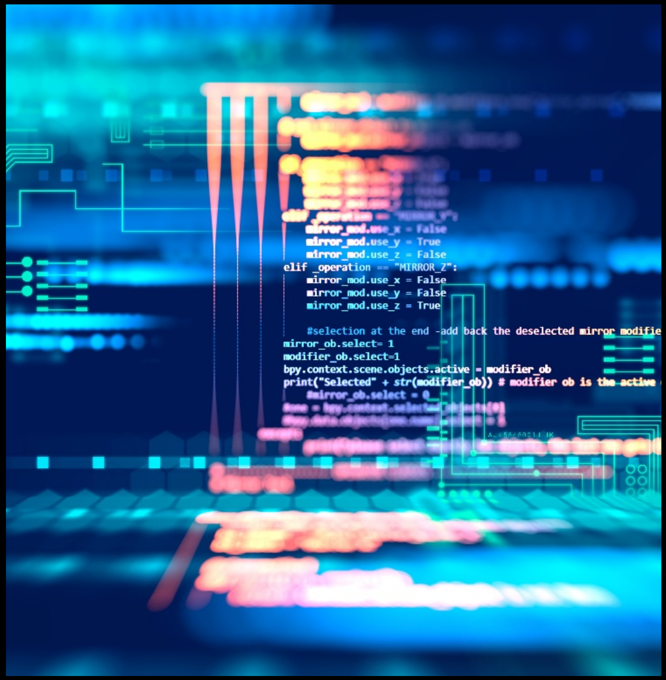


Fragmentation

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



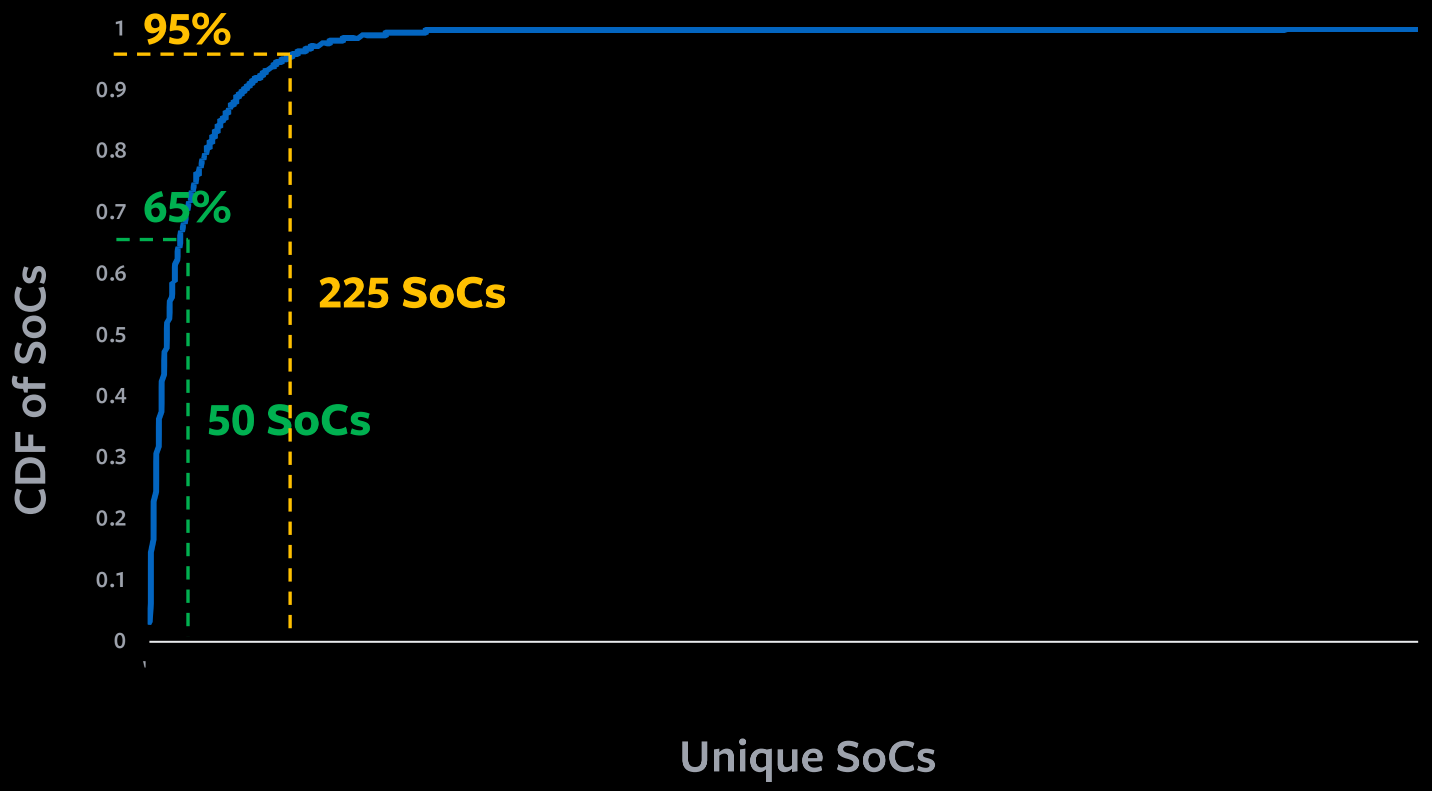
Programmability

Programmability is a Primary Roadblock for Using Mobile Co-processors

Lay of the Land

FRAGMENTATION

Taking a Closer Look at Smartphones Facebook Runs on



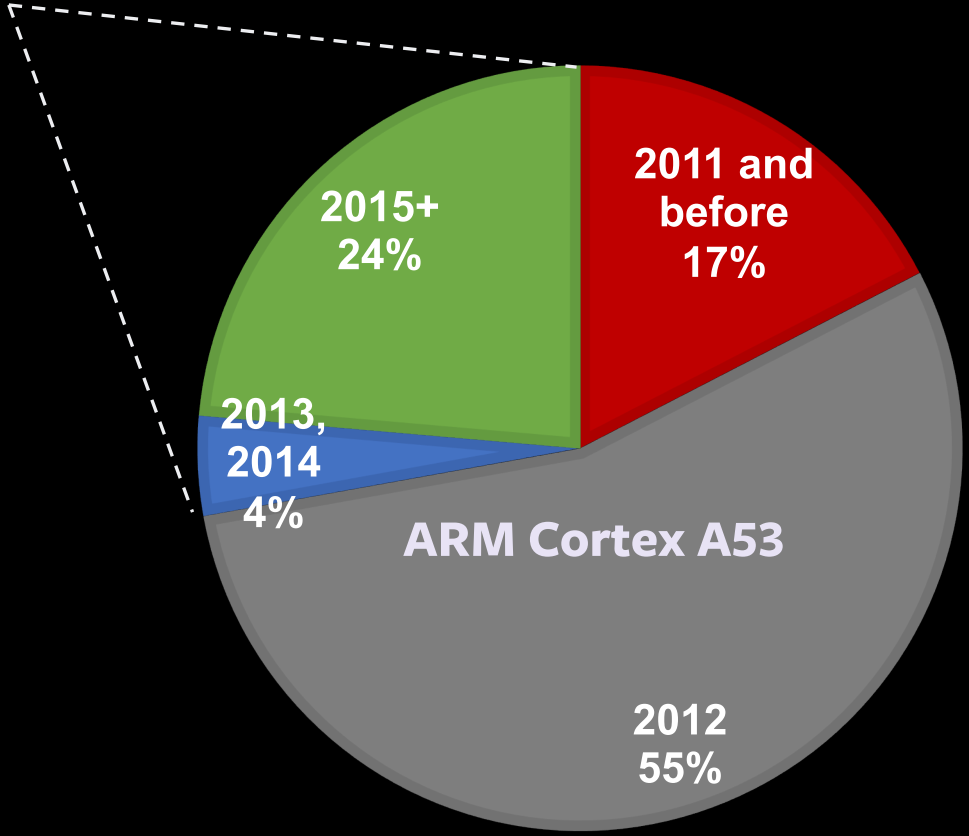
- Qualcomm Snapdragon
- Samsung Exynos
- MediaTek Helio
- HiSilicon Kirin et al.

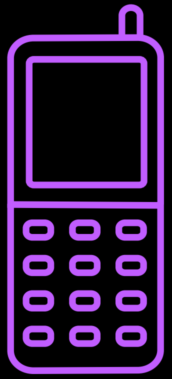
THERE IS NO STANDARD SOC TO OPTIMIZE FOR

Lay of the Land

FRAGMENTATION

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





72%

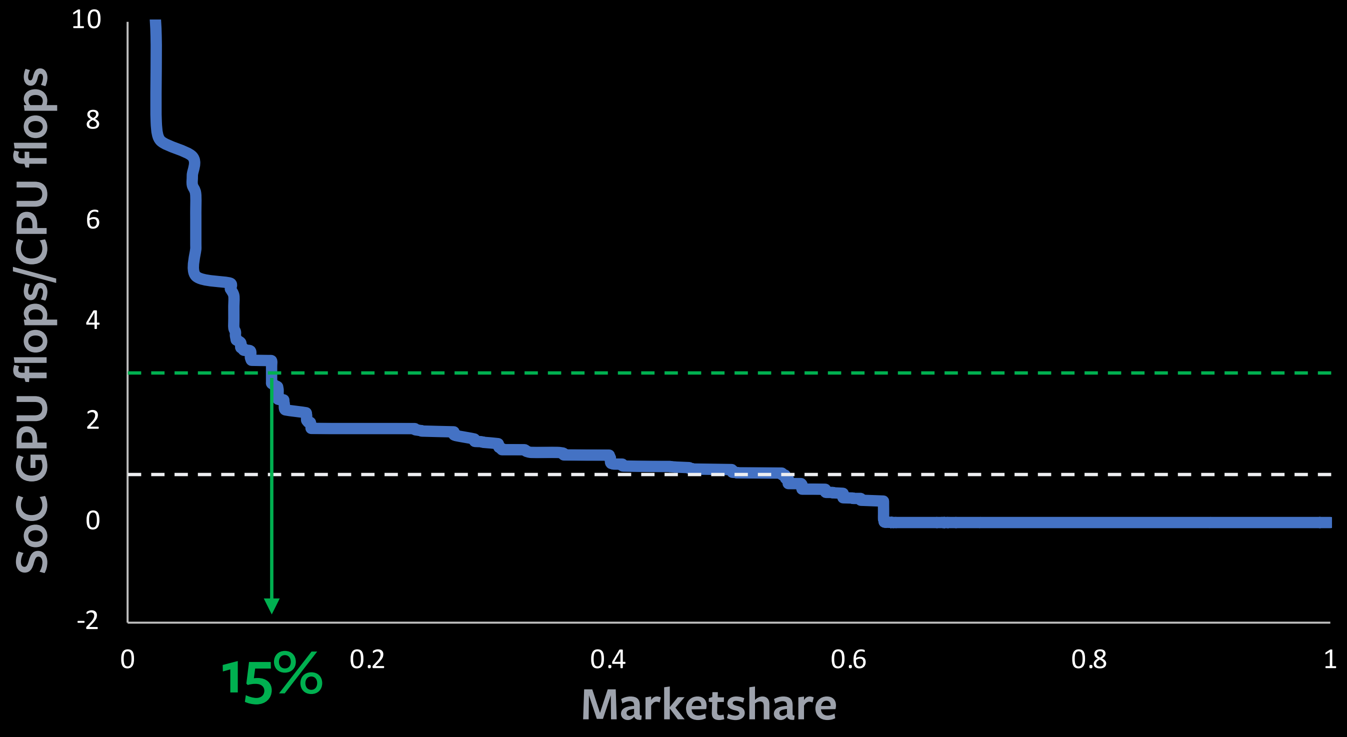
OF THE WORLD'S CELL PHONES ARE MORE THAN 7 YEARS OLD

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

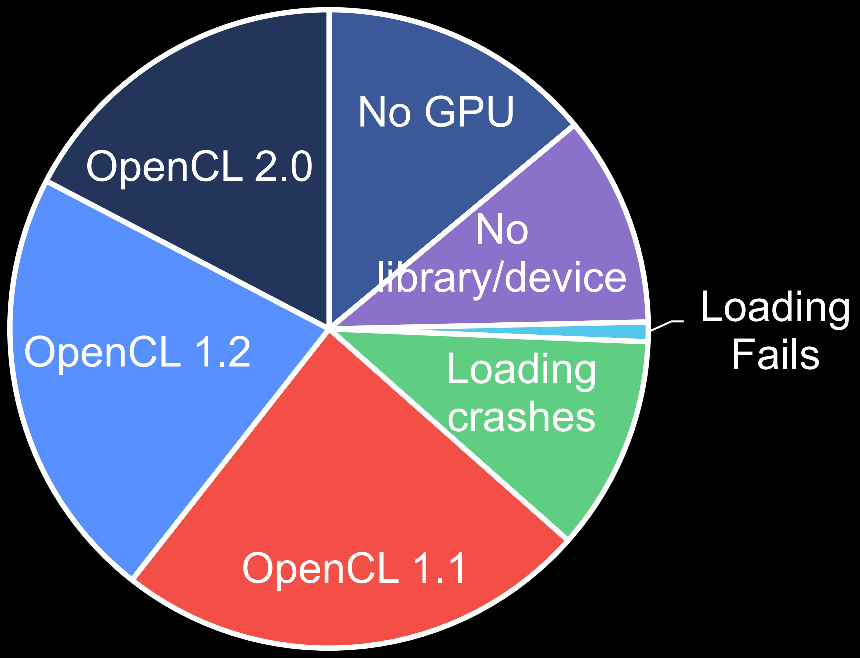
LESS THAN 15% SMARTPHONES HAVE A GPU THAT IS 3 TIMES AS POWERFUL AS ITS CPU

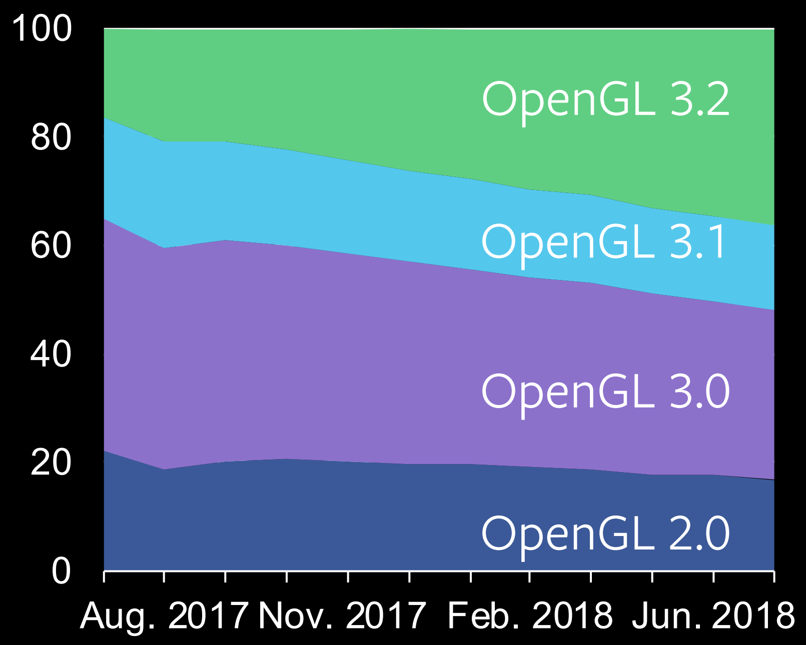
Lay of the Land

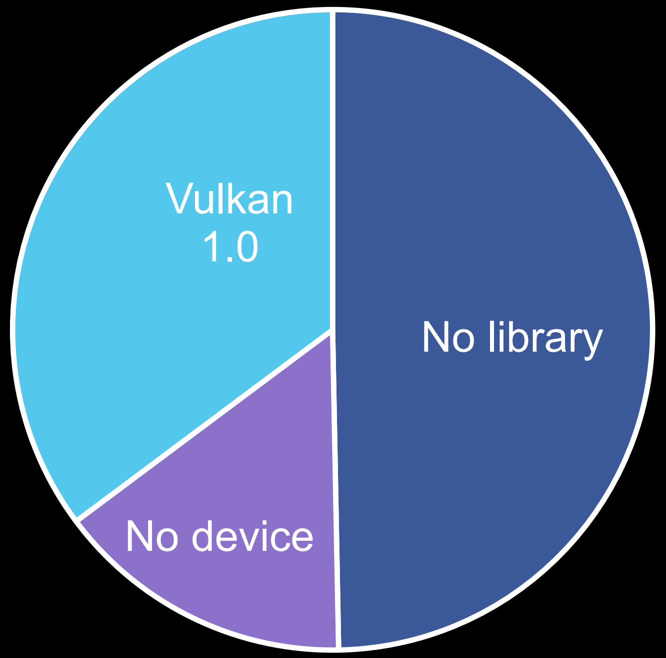
PROGRAMMABILITY

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

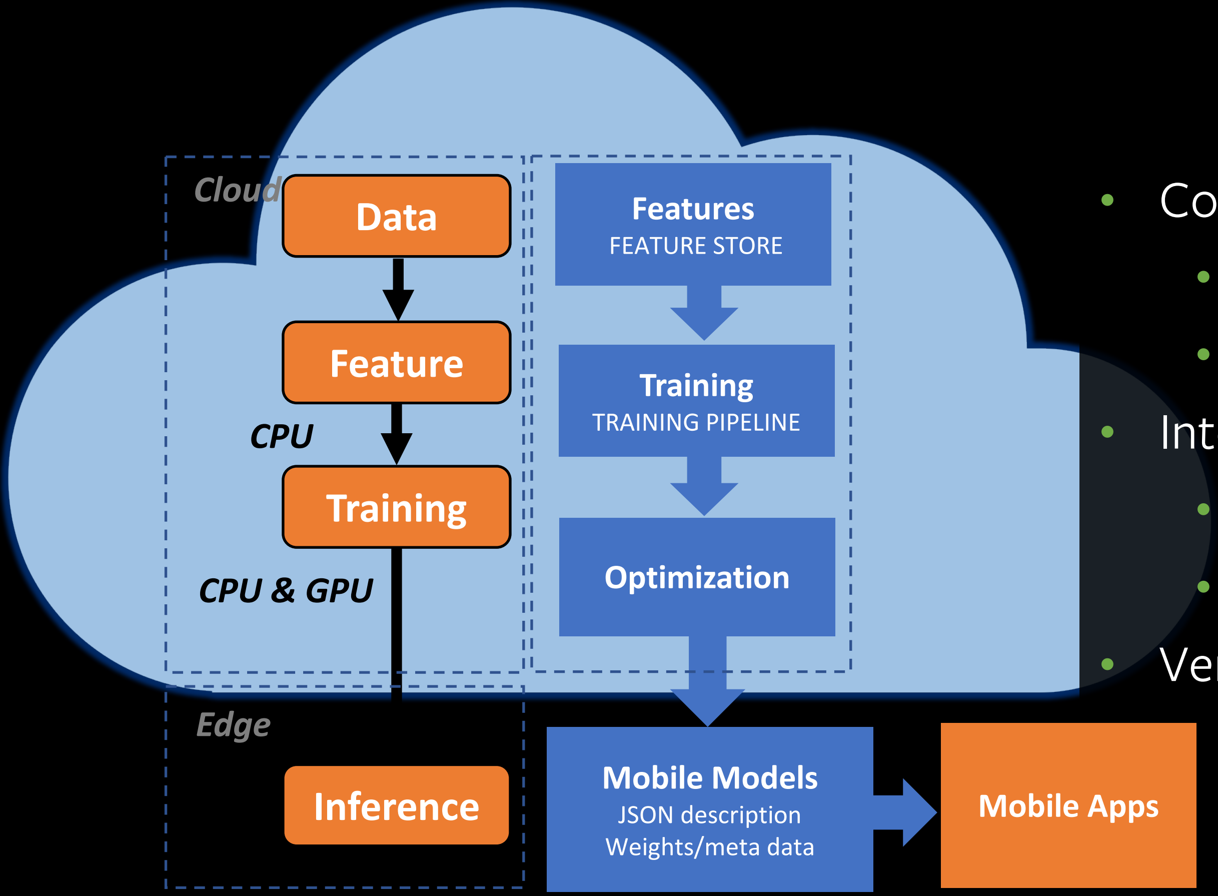
Horizontal Integration:
Making Inference on
Smartphones

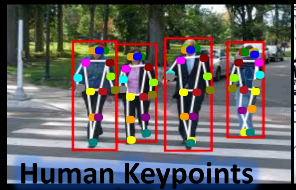
Vertical Integration:
Processing Inference for
Oculus VR

Inference in the Wild:
Performance
Variability

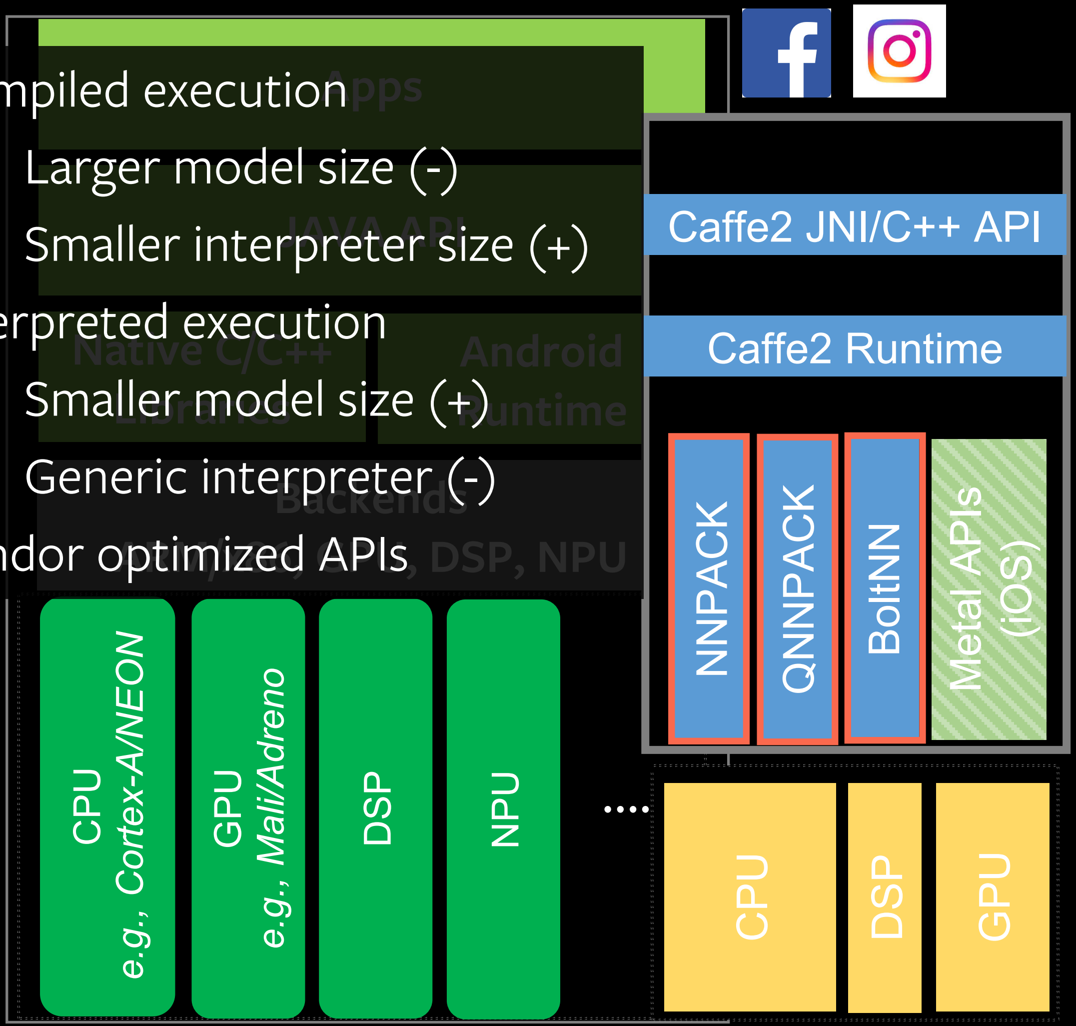
Horizontal Integration

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

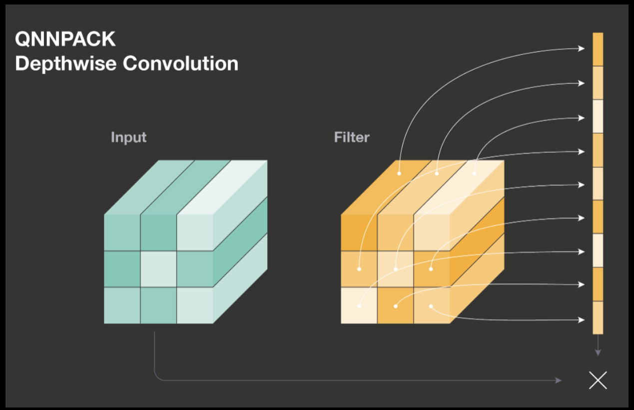
(32-BIT FLOATING POINT)

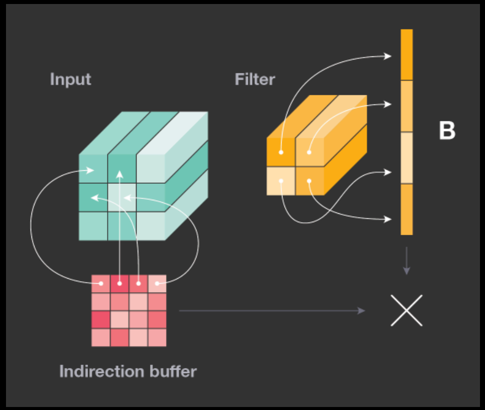
- Optimized convolution implementation using **Winograd** and **FFT**
- Best for NN with 3x3, 5x5 or larger convolutions

QNNPACK/QUANTIZED NNPACK

(8-BIT FIXED POINT)

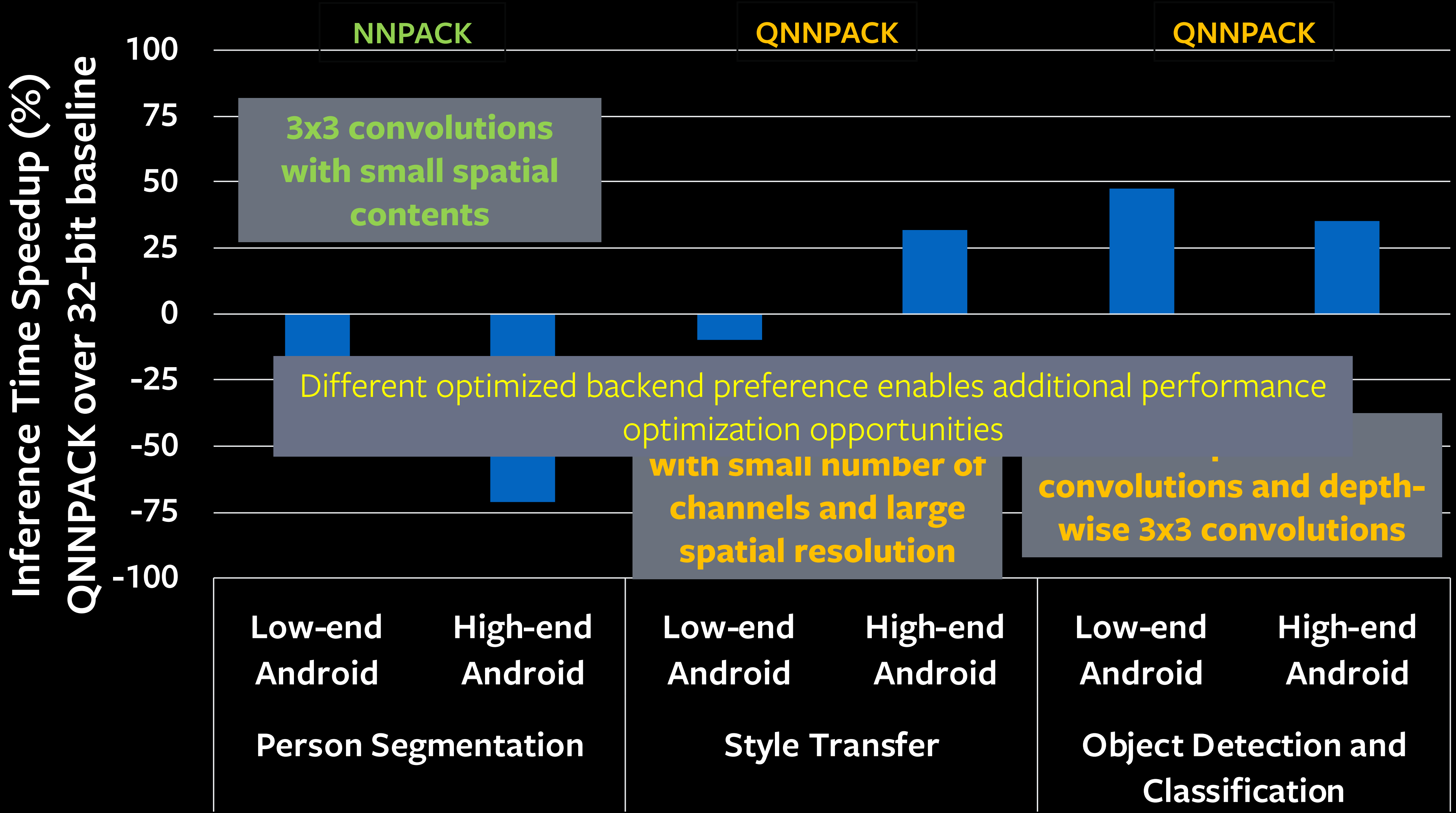
- Optimized direct convolution implementation
- Best for low-intensity convolutions
- Grouped, depth-wise, dilated convolutions
- Eliminate the overhead of im2col and other memory layout transformation





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

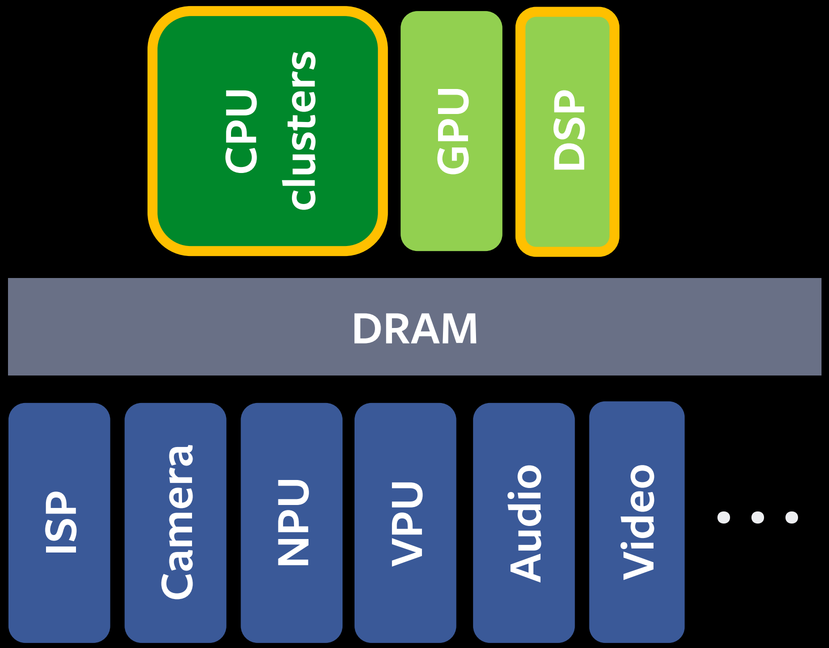
Horizontal Integration:
Making Inference on
Smartphones

Vertical Integration:
Processing Inference for
Oculus VR

Inference in the Wild:
Performance
Variability

Vertical Integrated Systems

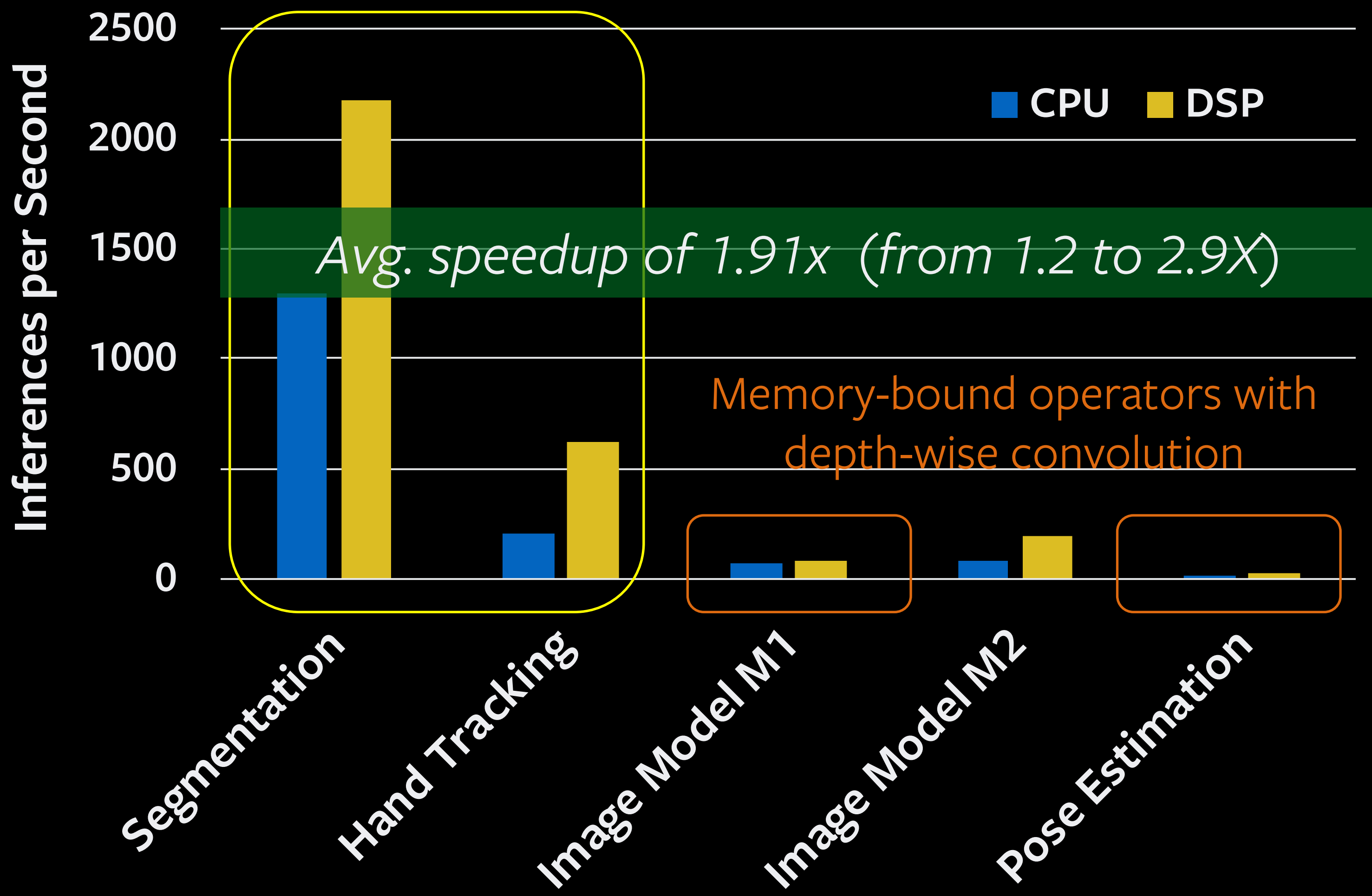
Processing Inference for Oculus VR



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

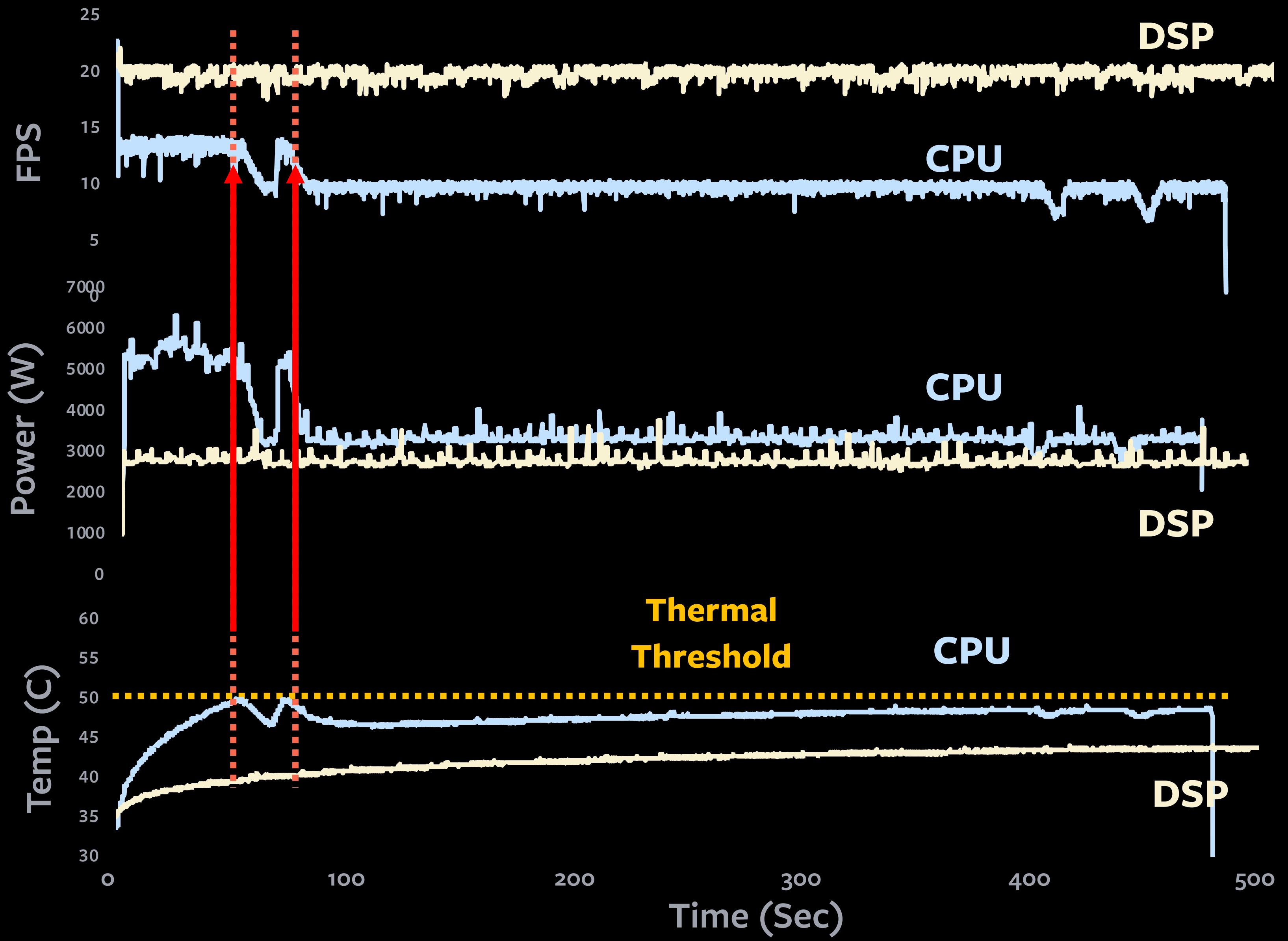


Vertical Integrated Systems

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**



Computing Platforms at the Edge

Workload Characterization

- MobileBench [IISWC-2013]
- Joule/Instruction [IISWC-2014]
- TLP for Mobile [ISPASS-2015]
- Multitasking for Mobile [IISWC-2015]

We use the rigorous workload characterization results to guide designs tailored for mobile

Energy Efficiency Optimization

- STEAM [TECS-2014]
- Statistical PPW Optimization [HPCA-2016] [TMC-2018]
- DORA [ISPASS-2018]

We propose a family of algorithms that maximize smartphone energy efficiency subject to various dynamic execution scenarios

Temperature Management

- Thermal Modeling [IISWC-2017] [ITHERM-2018*]
- Hybrid Cooling Technologies
- Near Sensor Processing

We design a collection of temperature-aware optimization:
Floor-planning;
Advanced cooling technologies for mobile (TEC/PCM);
Near sensor processing

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

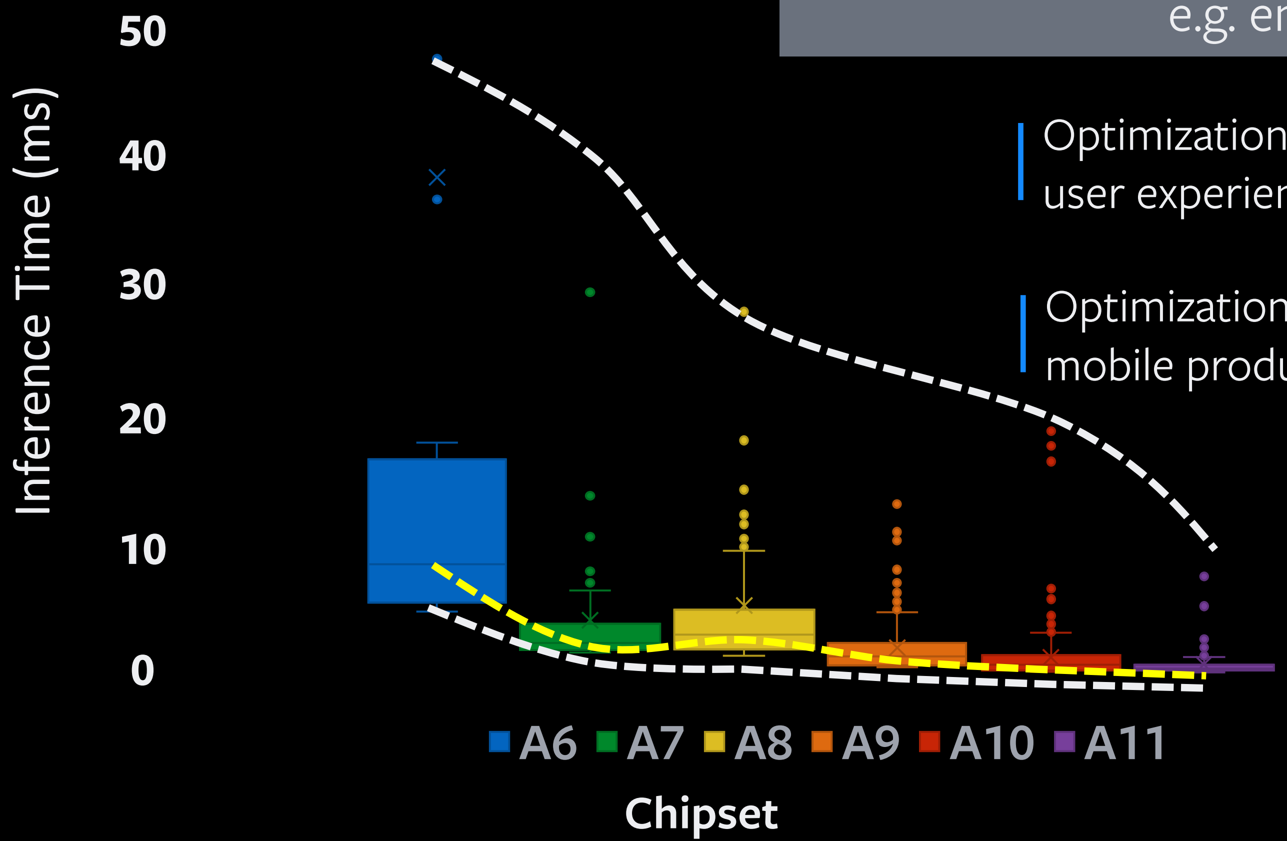
Vertical Integration:
Processing Inference for
Oculus VR

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



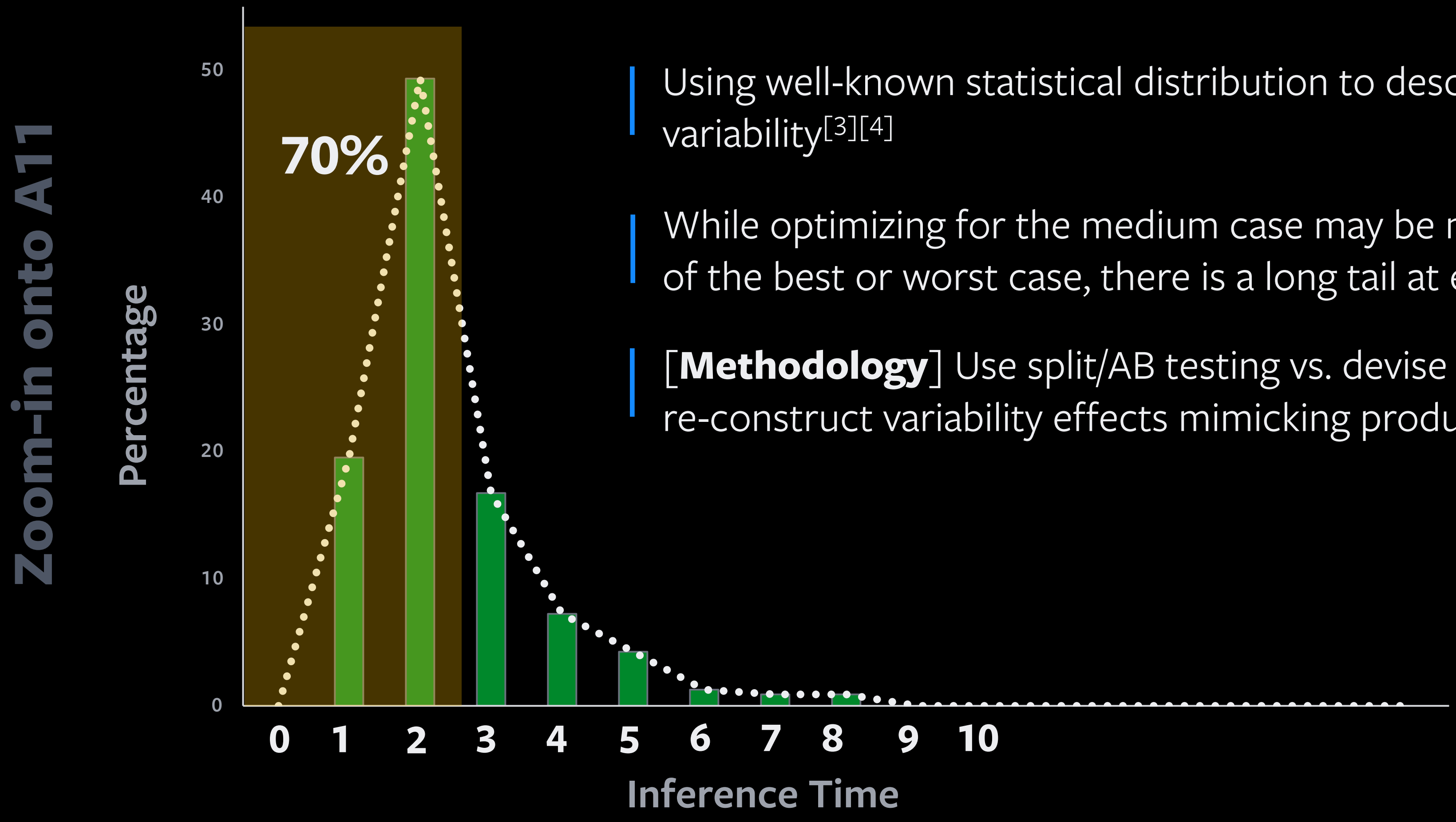
- Optimizations targeting the average or medium case sacrifices user experience
- Optimization must endure the high degree of variation in the mobile production environment

[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

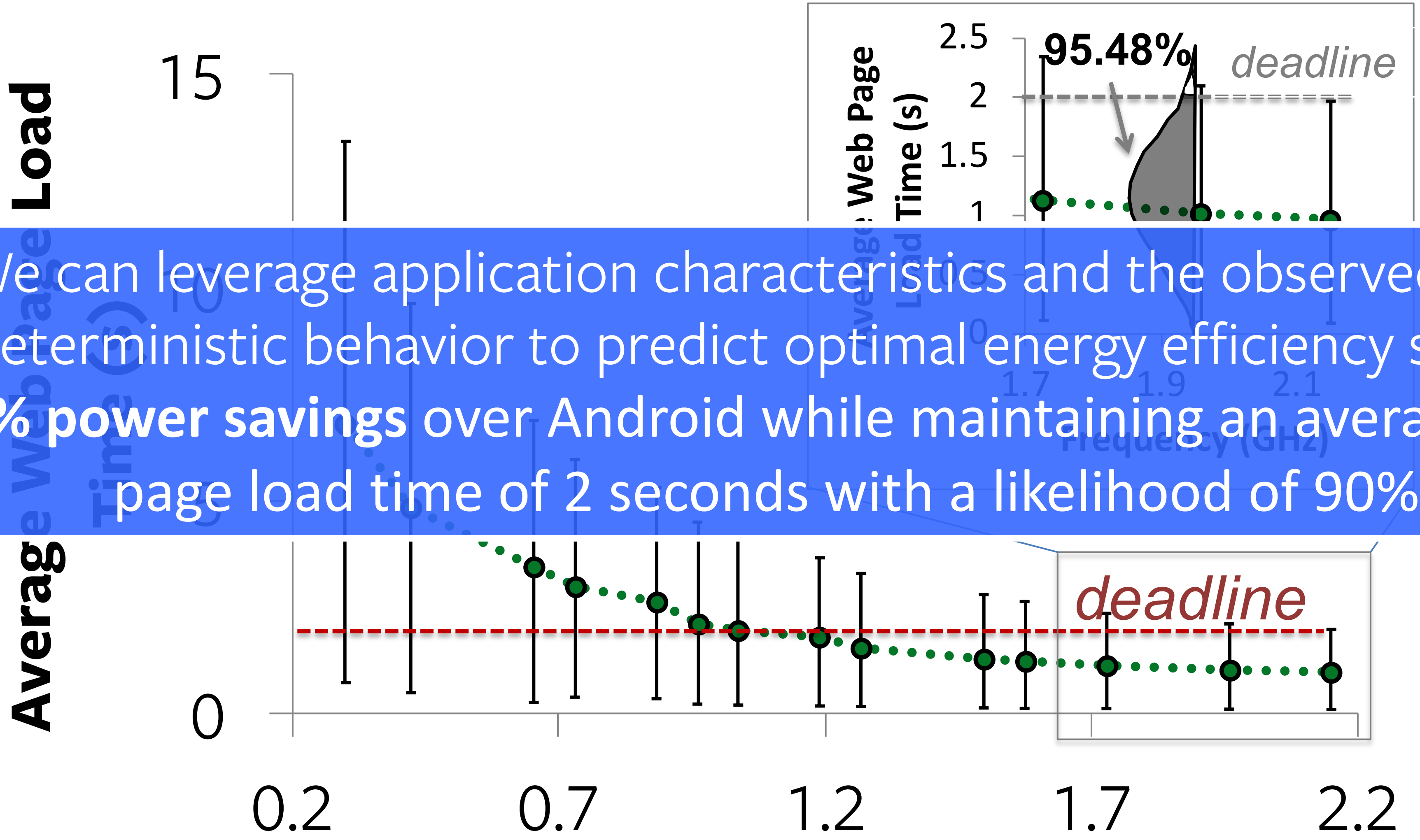
Does the Performance Variability Follow Certain Statistical Distributions?



- 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

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

Energy Efficiency Optimization with Stochastic Assumption

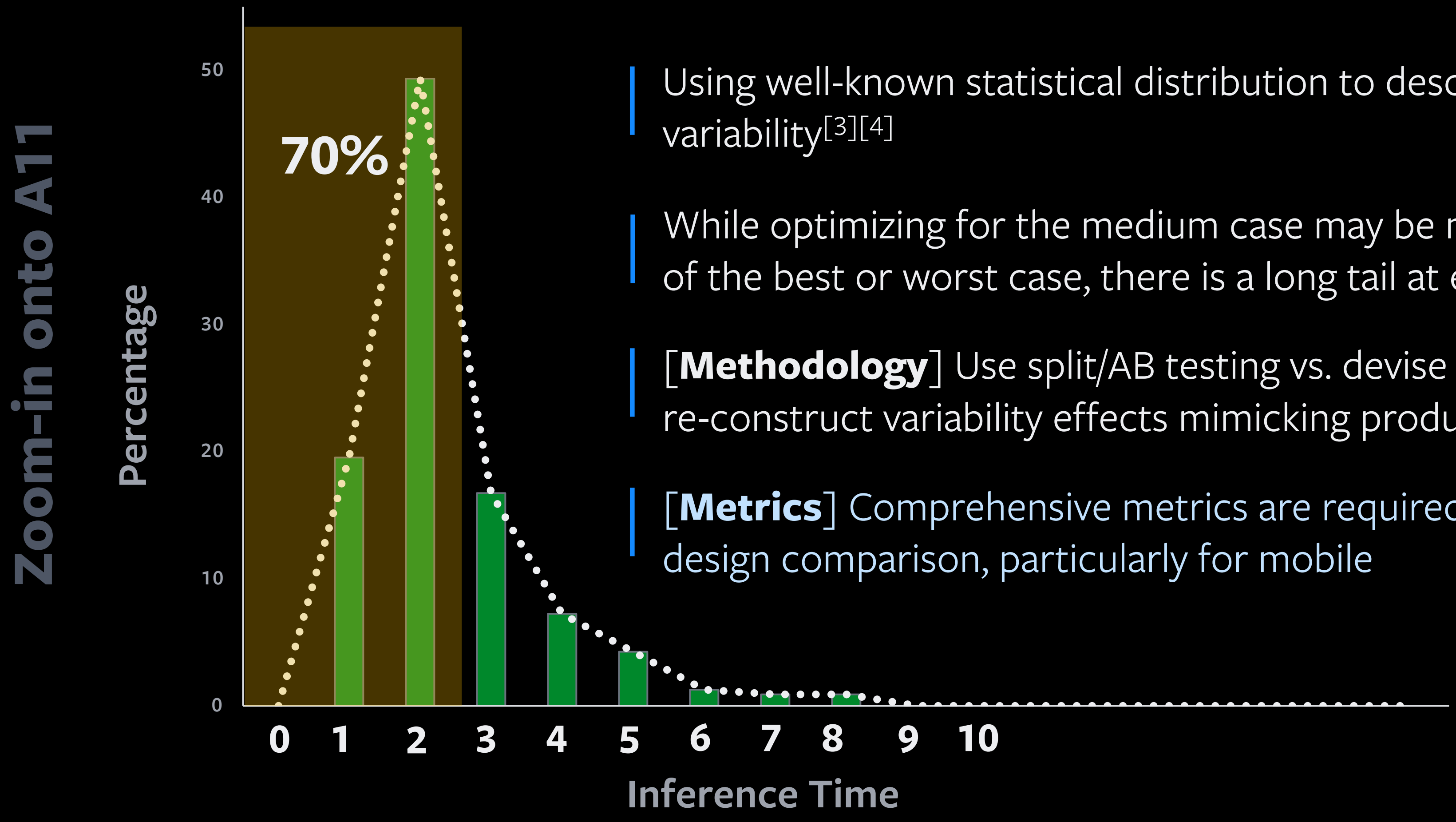


We can leverage application characteristics and the observed non-deterministic behavior to predict optimal energy efficiency states: **29% power savings** over Android while maintaining an average web page load time of 2 seconds with a likelihood of 90%

Gaudette, Wu, and Vrudhula. HPCA-2016; TMC-2018.

Inference in the Wild

Does the Performance Variability Follow Certain Statistical Distributions?



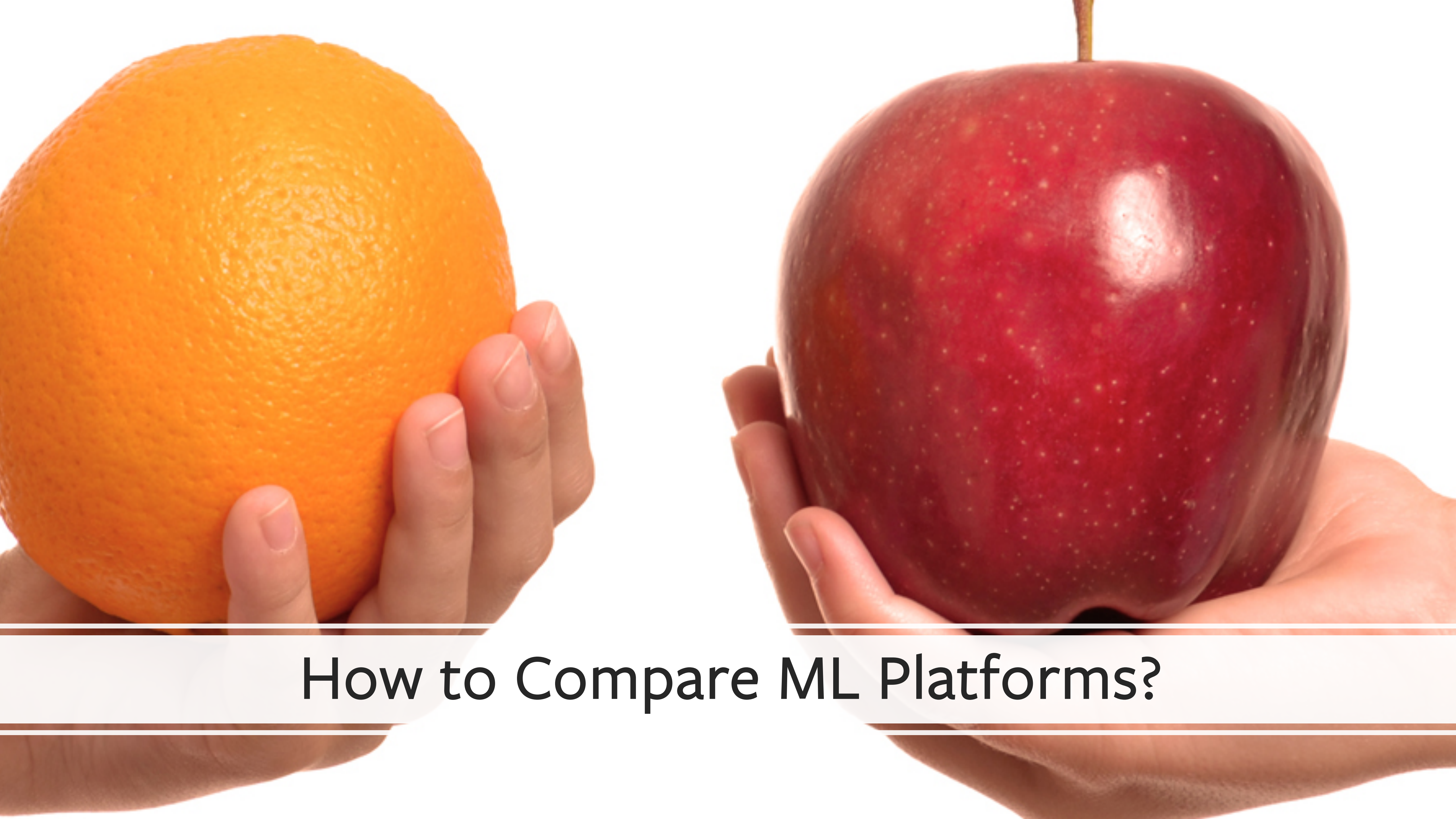
Using well-known statistical distribution to describe performance variability[3][4]

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

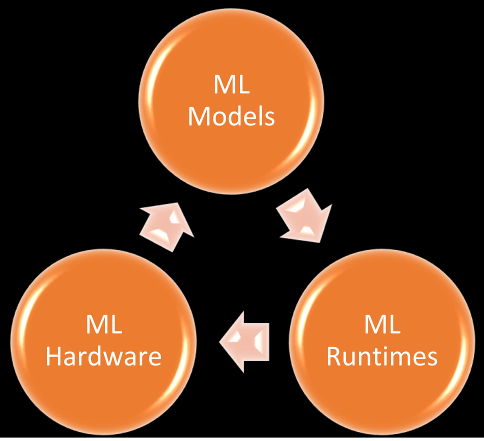
[**Metrics**] Comprehensive metrics are required for fair, representative design comparison, particularly for mobile

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


How to Compare ML Platforms?

MLPerf

www.mlperf.org



- Accelerate progress in ML via **fair and useful measurement**
- Serve both the **commercial and research community**
- **Encourage innovation** to improve the state-of-the-art of ML
- **Enforce replicability** to ensure reliable results
- Use **representative workloads**, reflecting production use cases
- Keep **benchmarking affordable**

MLPerf Inference Benchmark v0.5

Open Challenges & Issues

- Large and high-quality data sets
- Diversity in machine learning models/use cases

- Metrics
  - Performance: how fast is a model for inference ?
  - Quality: prediction accuracy ?

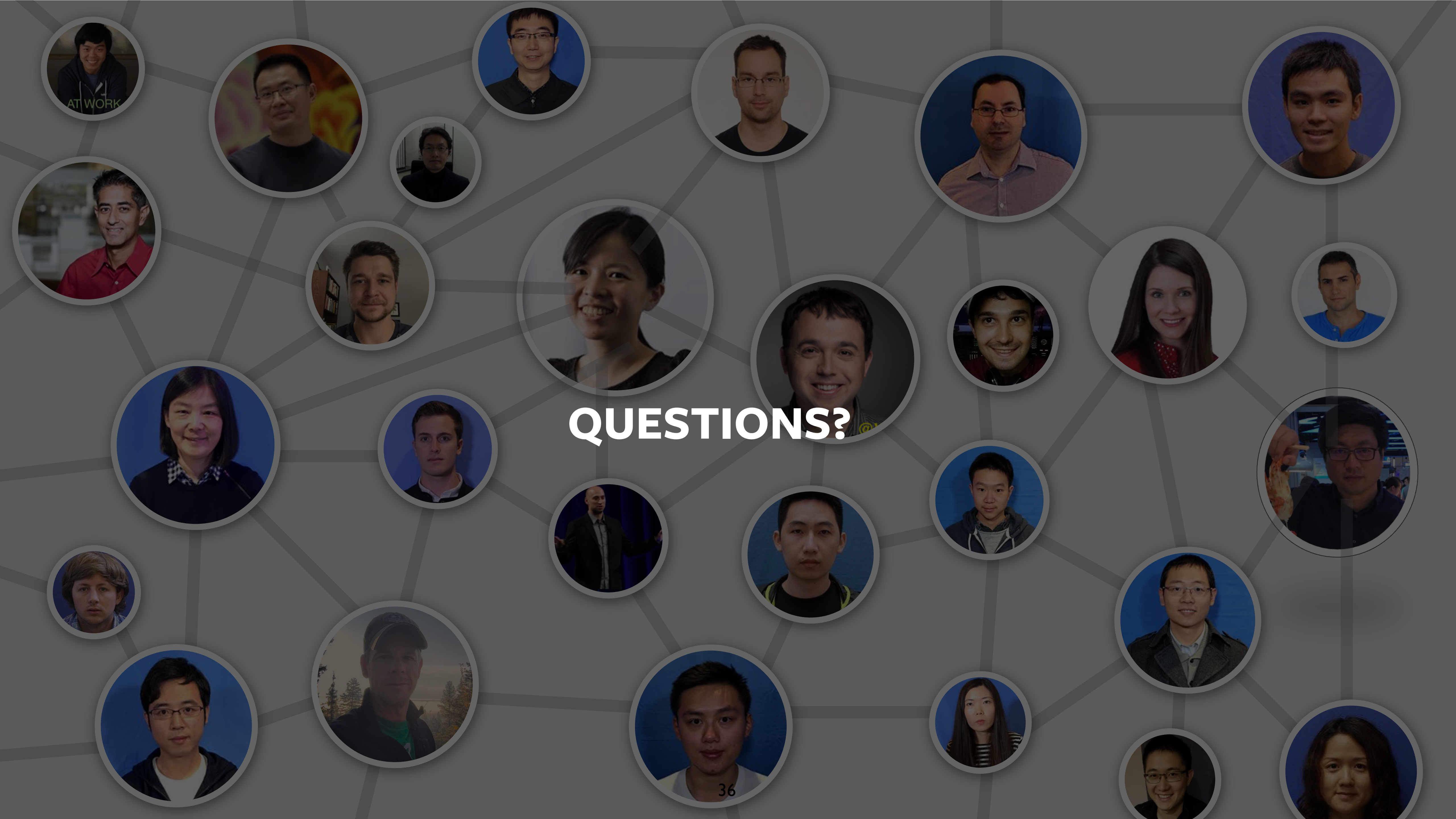
Area	Benchmark	Dataset	Model
Vision	Image classification	ImageNet	MobileNet v1
			ResNet-50
	Object detection	MS-COCO 2017	SSD-MobileNet v1
			SSD-ResNet-34
Language	Translation	Google NMT	WMT Eng-Germ

How to bridge from node to scale?

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

K. Hazelwood et al., "**Applied Machine Learning at Facebook: A Datacenter Infrastructure Perspective**", HPCA 2018.

C.-J. Wu et al., "**Machine Learning at Facebook: Understanding Inference at the Edge**", HPCA 2019.



QUESTIONS?

facebook