Open. Together.

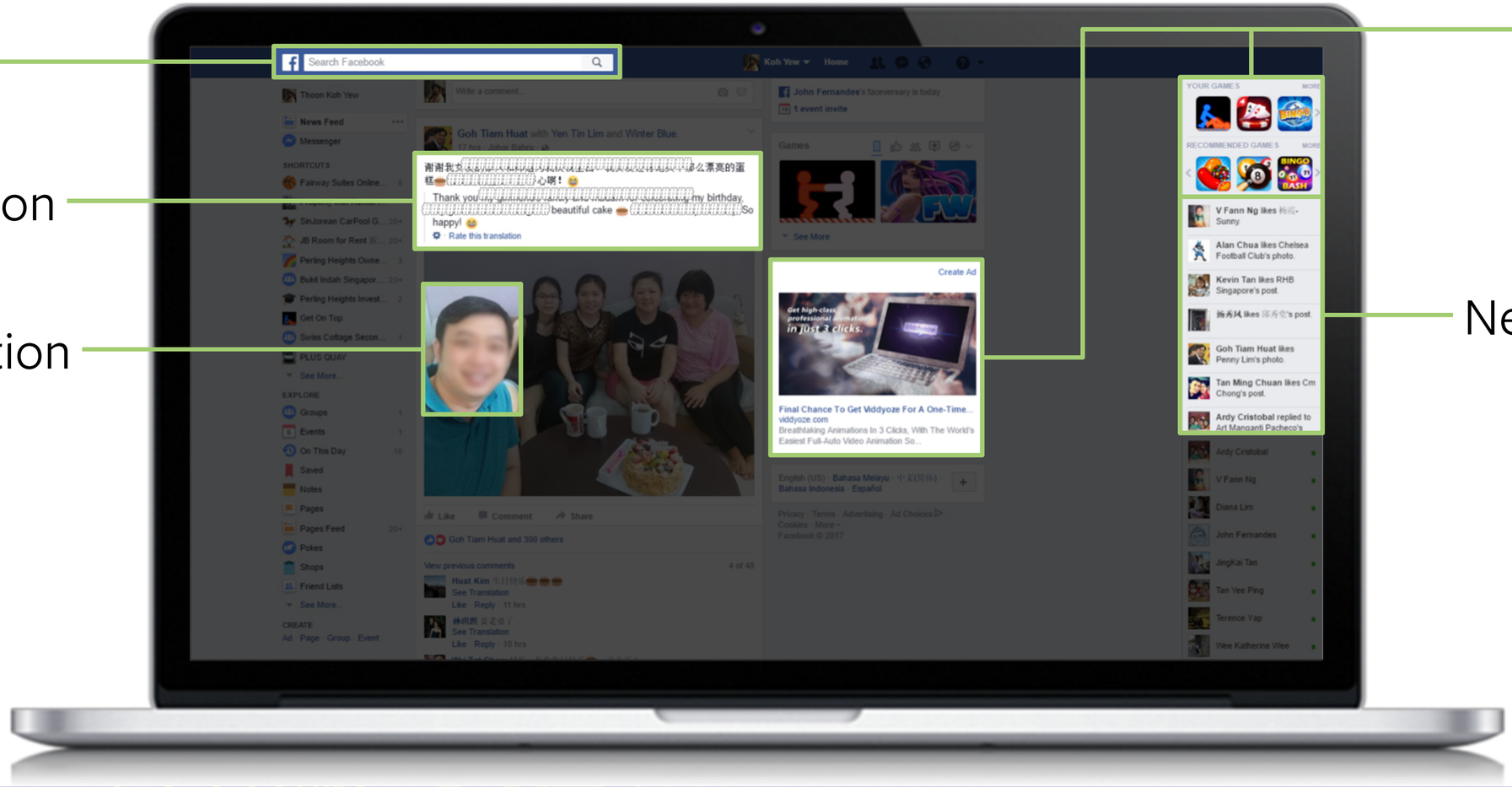
OCP
SUMMIT

OPEN PLATINUM™

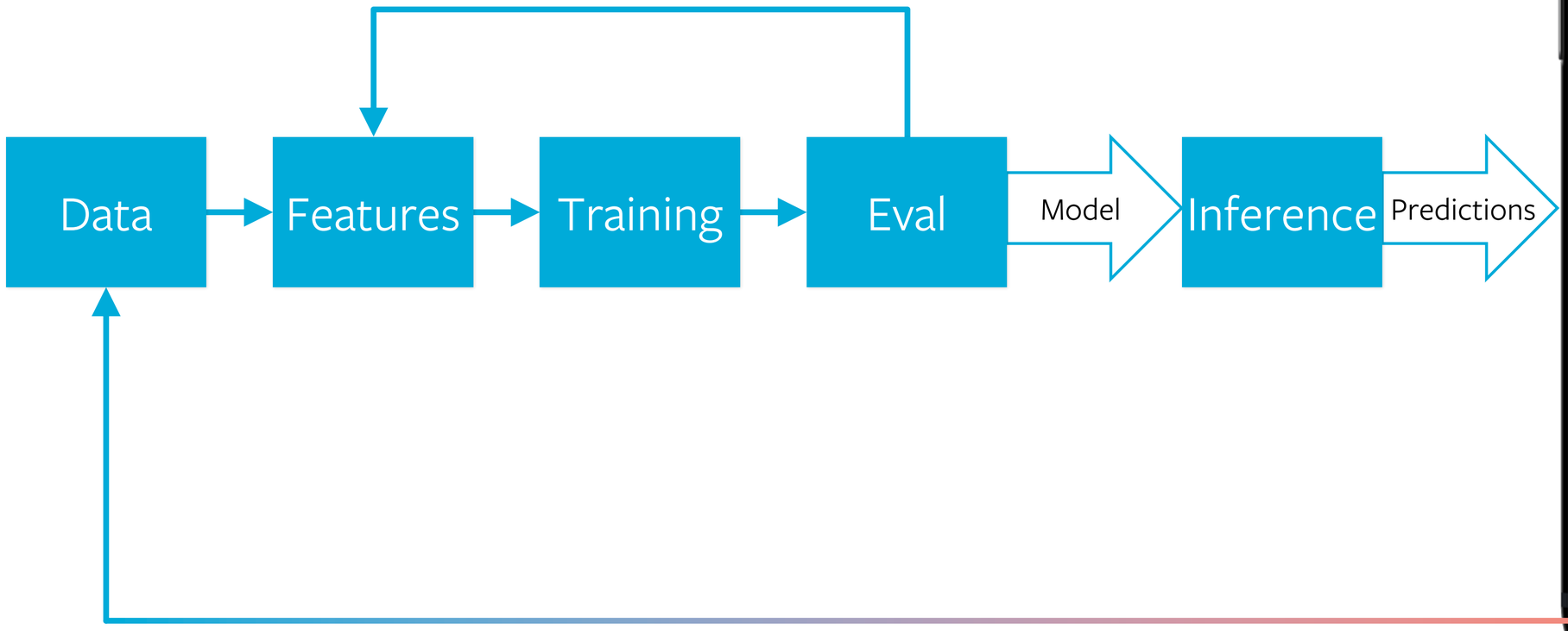
AI at Facebook: An Infrastructure Perspective

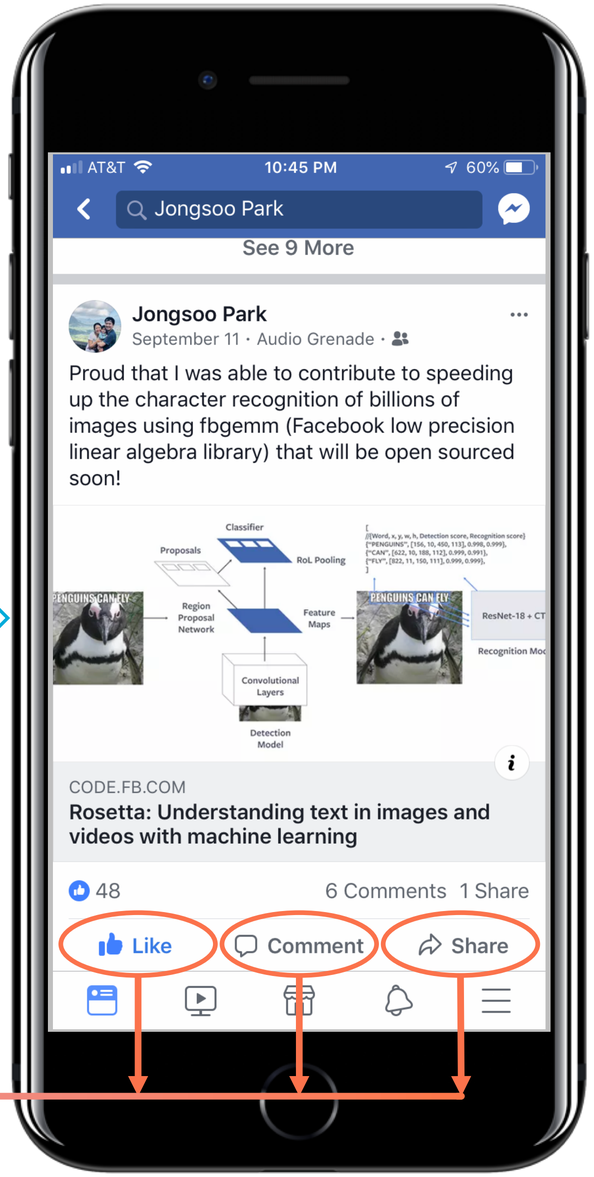
Whitney Zhao, Hardware Engineer, Facebook
Sam Naghshineh, Technical Program Manager, Facebook

How does FB use Machine Learning?

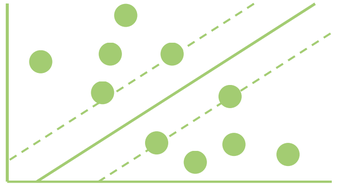


Machine Learning Execution Flow





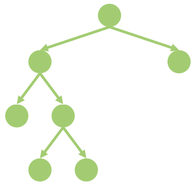
Major AI Services and Algorithms



SVM

Support Vector Machines

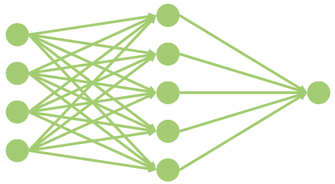
- Facer



GBDT

Gradient-Boosted Decision Trees

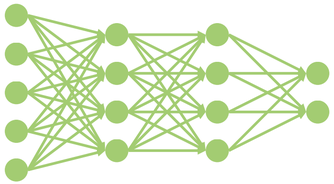
- Sigma



MLP

Multi-Layer Perceptron

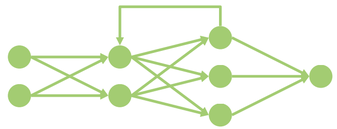
- News Feed
- Ads
- Search
- Sigma



CNN

Convolutional Neural Nets

- Facer
- Lumos

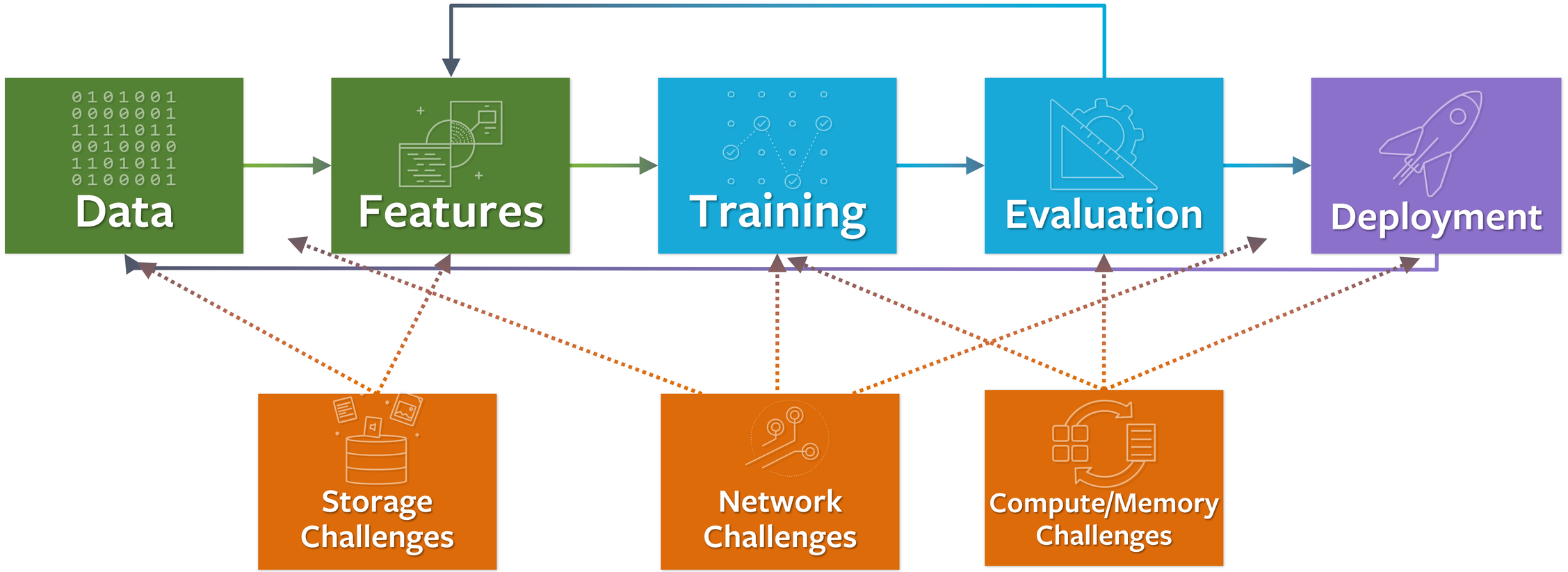


RNN

Recurrent Neural Nets

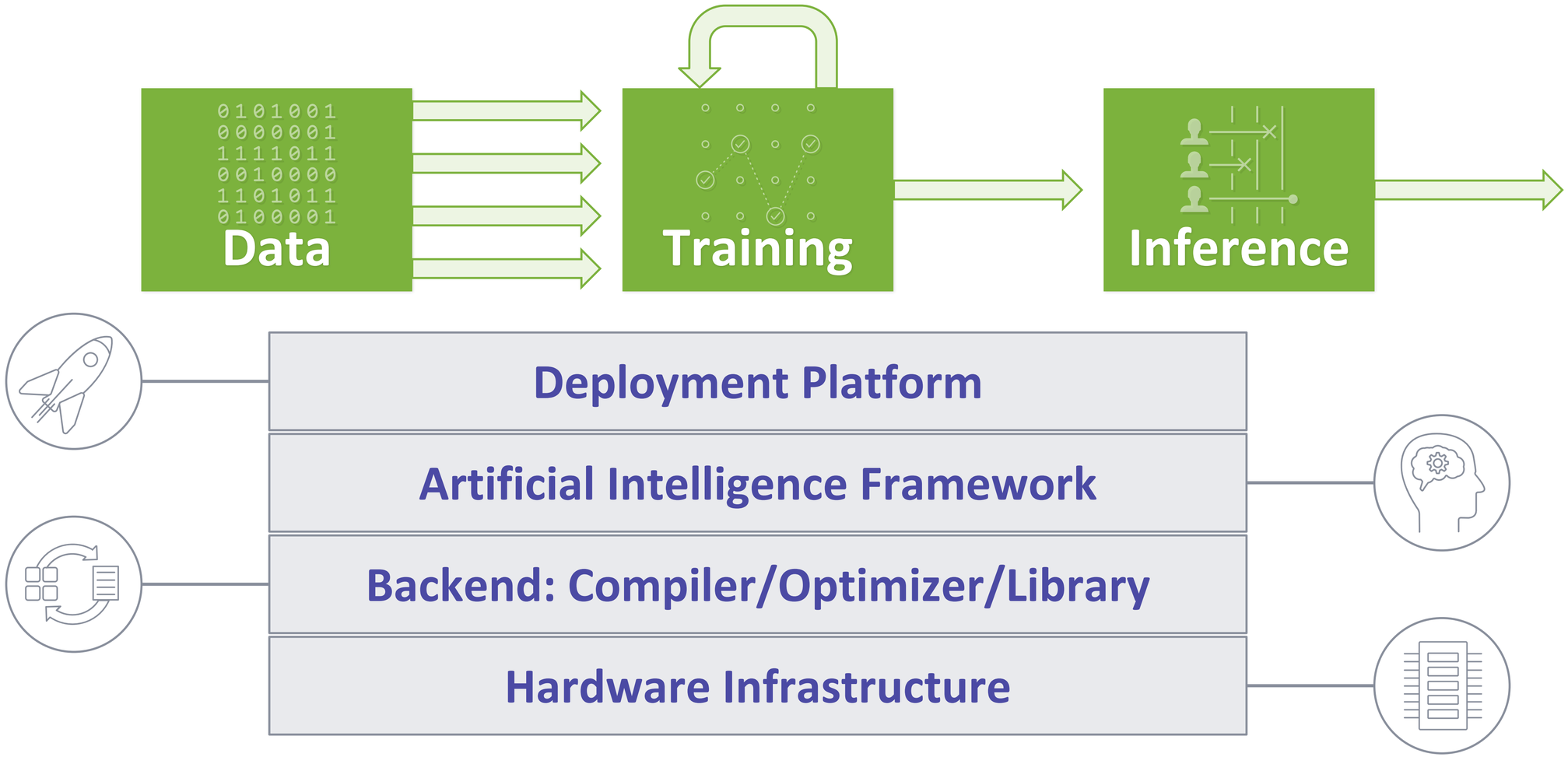
- Language Translation
- Speech Recognition
- Content Understanding

It’s an Infrastructure Challenge!

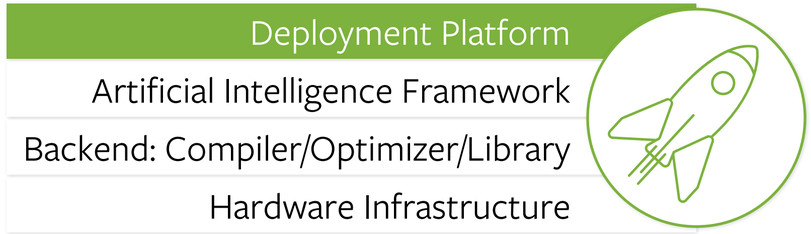


Accelerators could shift the system bottleneck from compute intensity to memory/network

FB AI Development Ecosystem



FBLearner Platform



FBLearner **FEATURE STORE** → FBLearner **FLOW** → FBLearner **PREDICTOR**

- AI workflow for automated model management and deployment
- Programmable, reusable, distributed training pipeline
- Library of reusable ML algorithms
- Provide sharable and reusable history of past experiments

Framework

	PROTOTYPING	DEPLOYMENT
From then	PyTorch	Caffe2
To now	PyTorch	

Backends

Caffe2 | PyTorch | TensorFlow | MXNet | Microsoft CNTK | ... — Frameworks

ONNX — Shared model exchange

Glow compiler

libonnxifi-a.so — Vendor A

libonnxifi-b.so — Vendor B

onnxifi-c.dll — Vendor C

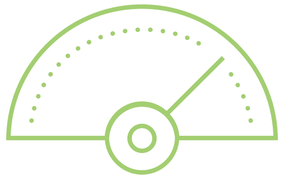
libonnxifi-d.dylib — Vendor D

Inference accelerators

Compilers, Optimizers & Libraries

ML Compiler

Facebook Glow



Vendor Optimizers

Apple CoreML

Nvidia TensorRT

Intel Nervana nGraph

Qualcomm SNPE

…

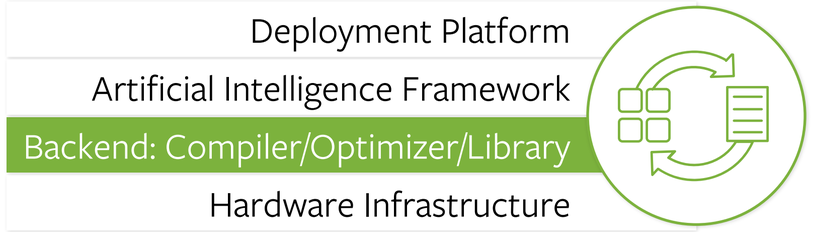
ML Libraries

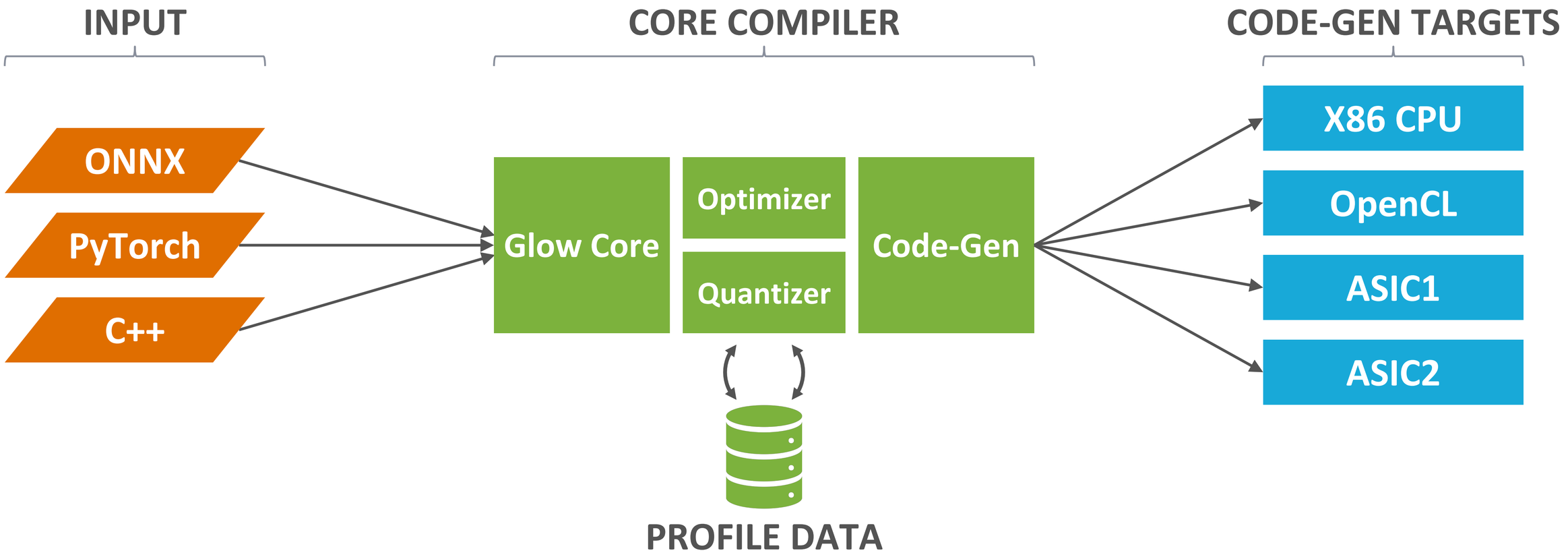
QNNPACK for mobile CPU

FBGEMM, Intel MKL for server CPU

CUDNN for GPU

Compilers, Optimizers & Libraries





Facebook AI Hardware
Today

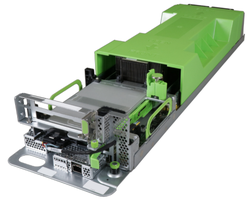
Deployment Platform
Artificial Intelligence Framework
Backend: Compiler/Optimizer/Library
Hardware Infrastructure

Bryce Canyon

- High density storage
- Storage-heavy tasks

Big Basin

- 8 NVIDIA P100s/V100s
- Very compute-heavy tasks



Tioga Pass

- Dual socket, many cores CPU
- Compute & mem-heavy tasks

Twin Lakes

- Single socket, fewer cores CPU
- Compute & low latency tasks

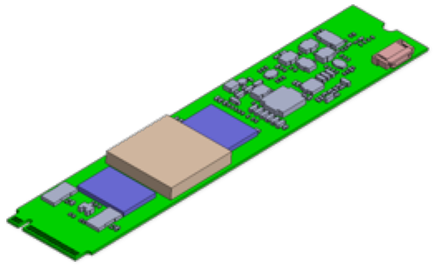
Facebook AI Hardware for Tomorrow

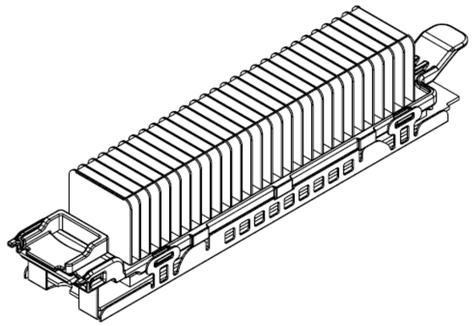
Inference Accelerators

King's Canyon

Deployment Platform
Artificial Intelligence Framework
Backend: Compiler/Optimizer/Library
Hardware Infrastructure

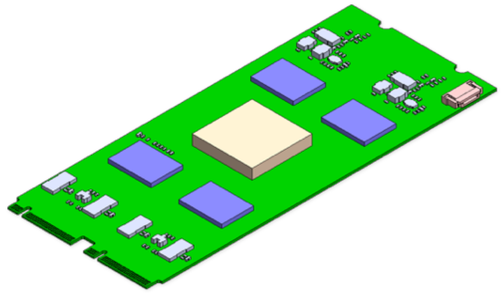
Standard M.2 Form Factor

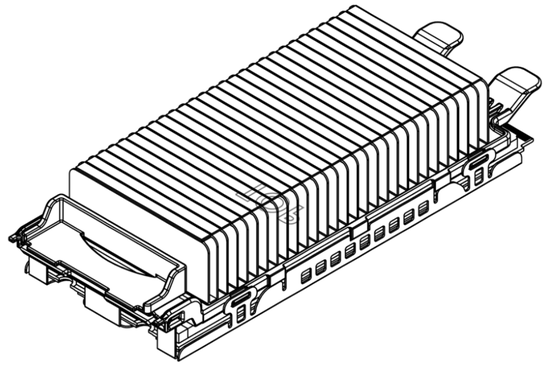




- ASIC
- DRAM
- 12W TDP
- PCIe x4
- UART/JTAG

Dual M.2 Form Factor





- ASIC
- DRAM
- 20W TDP
- PCIe x8
- UART/JTAG

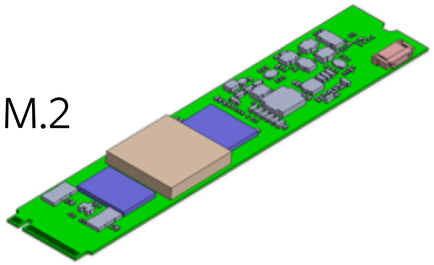
Facebook AI Hardware for Tomorrow

Inference Accelerators

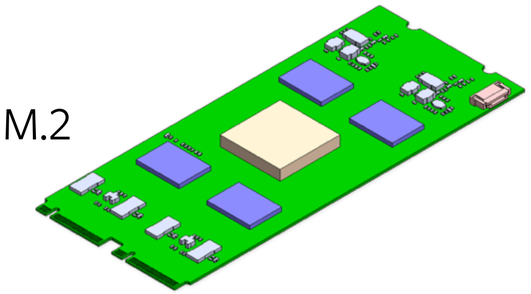
King's Canyon

Inference Module

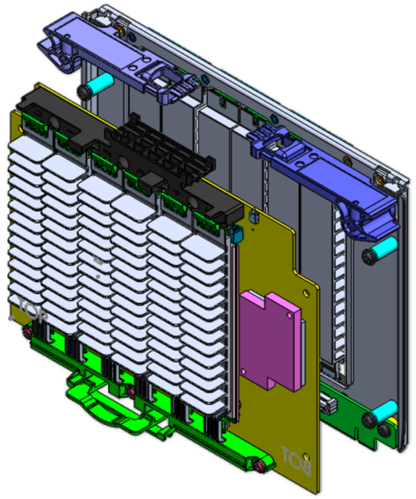
Single M.2



Dual M.2

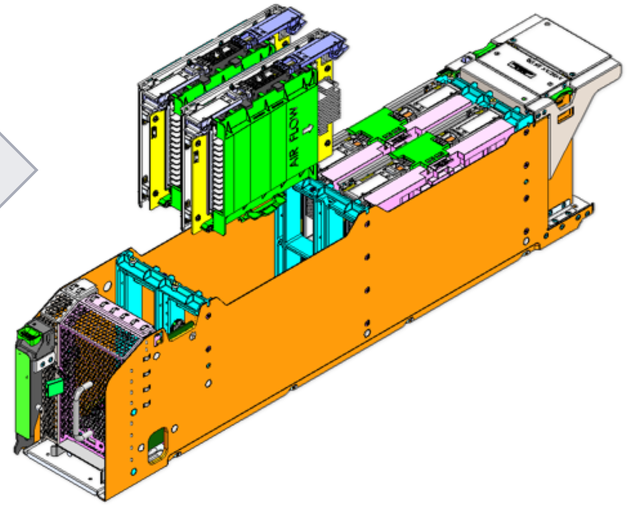


Carrier Card



Up to 12x single /
6x dual M.2

Yosemite V2



2x Twin Lakes with 2x carrier cards

Facebook AI Hardware for Tomorrow

Inference Accelerators

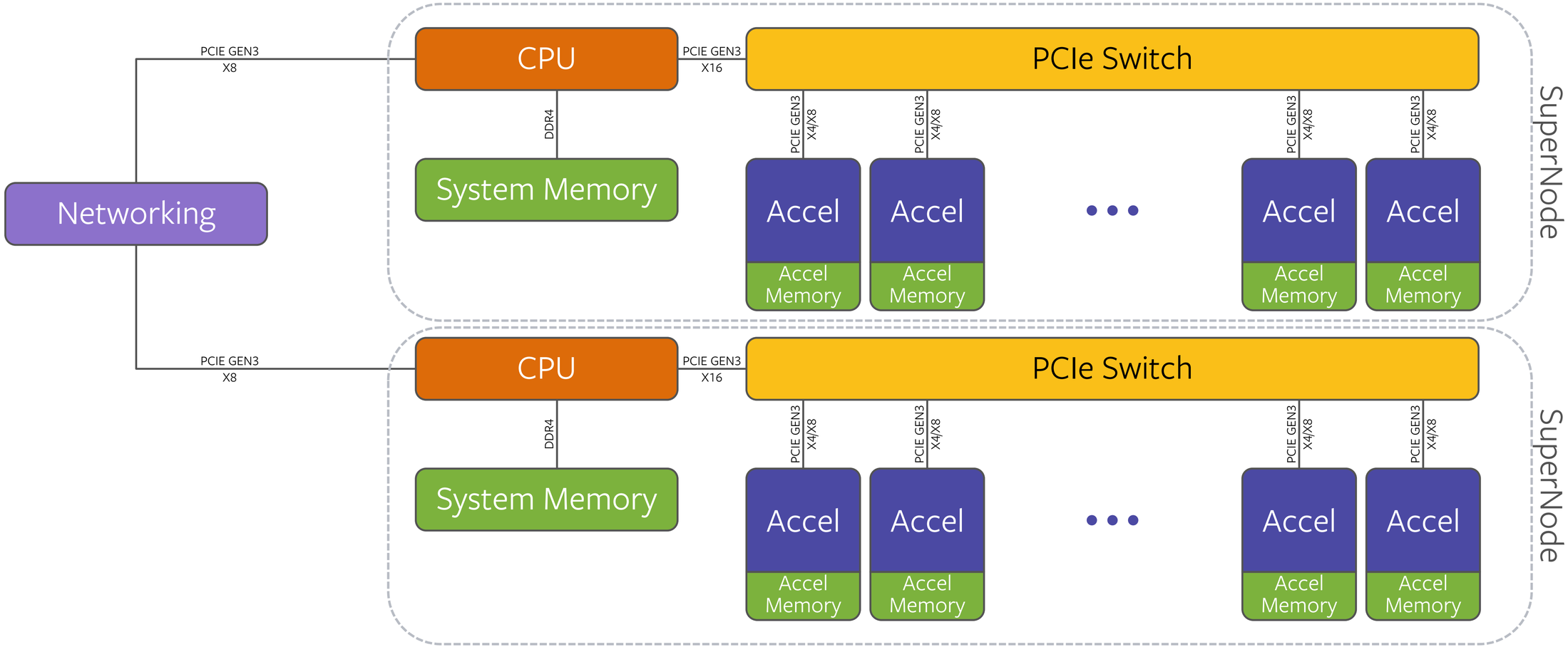
King's Canyon

Deployment Platform

Artificial Intelligence Framework

Backend: Compiler/Optimizer/Library

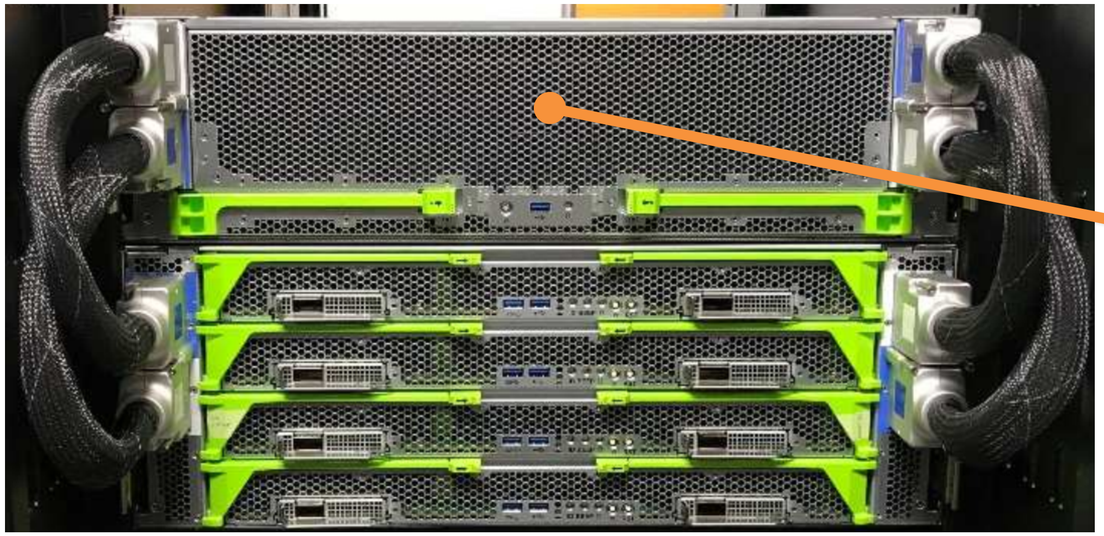
Hardware Infrastructure



Facebook AI Hardware for Tomorrow

Training

Large Memory Unified Training Platform Zion



8* Socket system with 8* Accelerators

8* OCP Accelerator Modules

Facebook AI Hardware for Tomorrow

Training

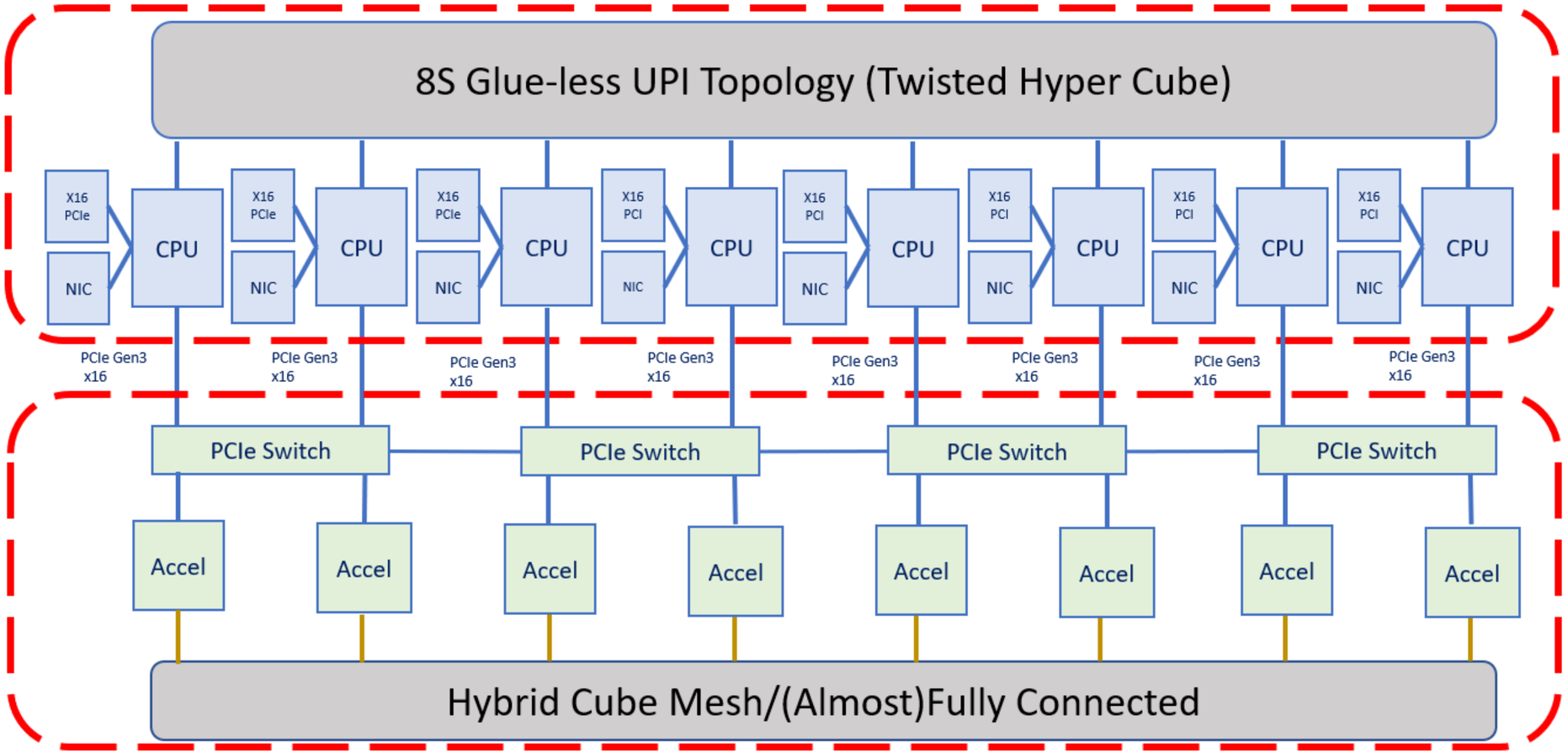
Large Memory Unified Training Platform Zion

Deployment Platform

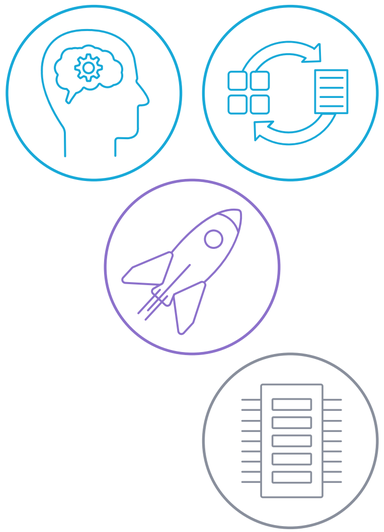
Artificial Intelligence Framework

Backend: Compiler/Optimizer/Library

Hardware Infrastructure



Putting it All Together



FBLearner Feature Store | FBLearner Flow | FBL Predictor

Data APIs | Caffe2, PyTorch, ONNX | C2 Predictor

Back-End: Compiler/Optimizer/Library

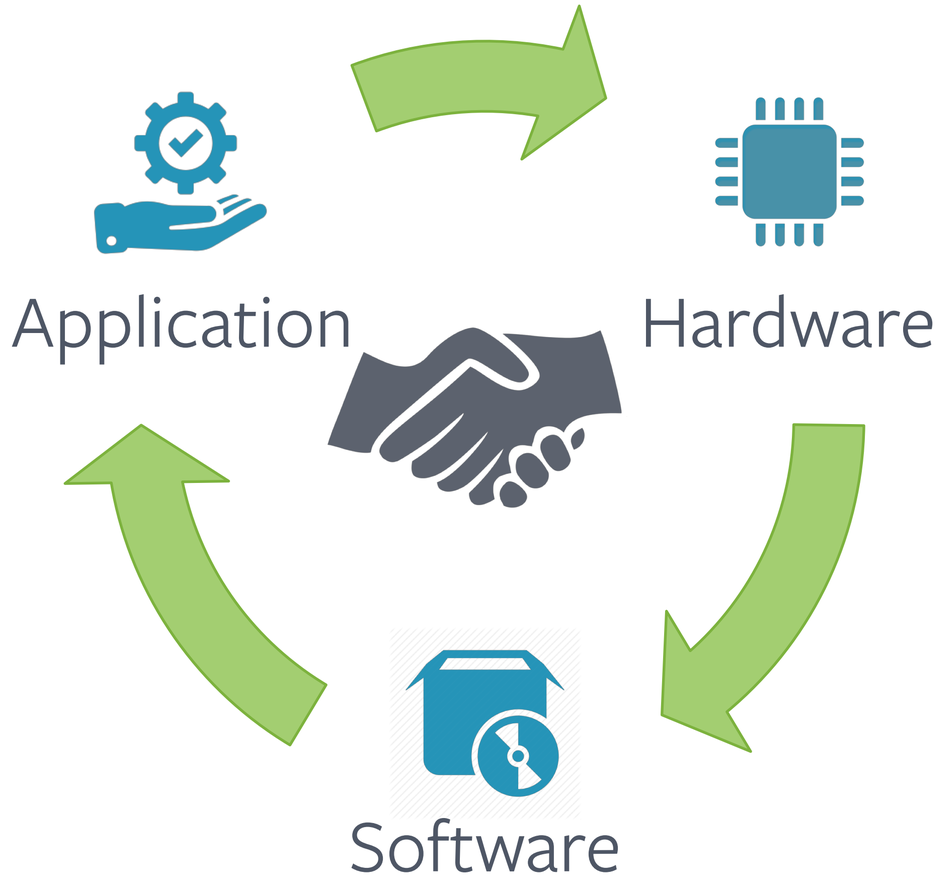
Bryce Canyon | Big Basin, Tioga Pass | Tioga Pass Yosemite V2


What changes when you scale to over
2 Billion People?

Scaling Challenges / Opportunities

- Lots of Data
- Lots of Compute
- Global scale
- Wide variety of models
- Full stack challenges

HW/SW Co-Design



Please join us for Facebook AI-related workshops on 3/15:

2:00 PM OCP Accelerator Module (OAM) System: An Open Accelerator Infrastructure Project

3:30 PM OCP Accelerator Module (OAM)


Open. Together.
OCP Global Summit | March 14–15, 2019
OCP
SUMMIT