Open. Together.

OCP
REGIONAL
SUMMIT

Misha Smelyanskiy

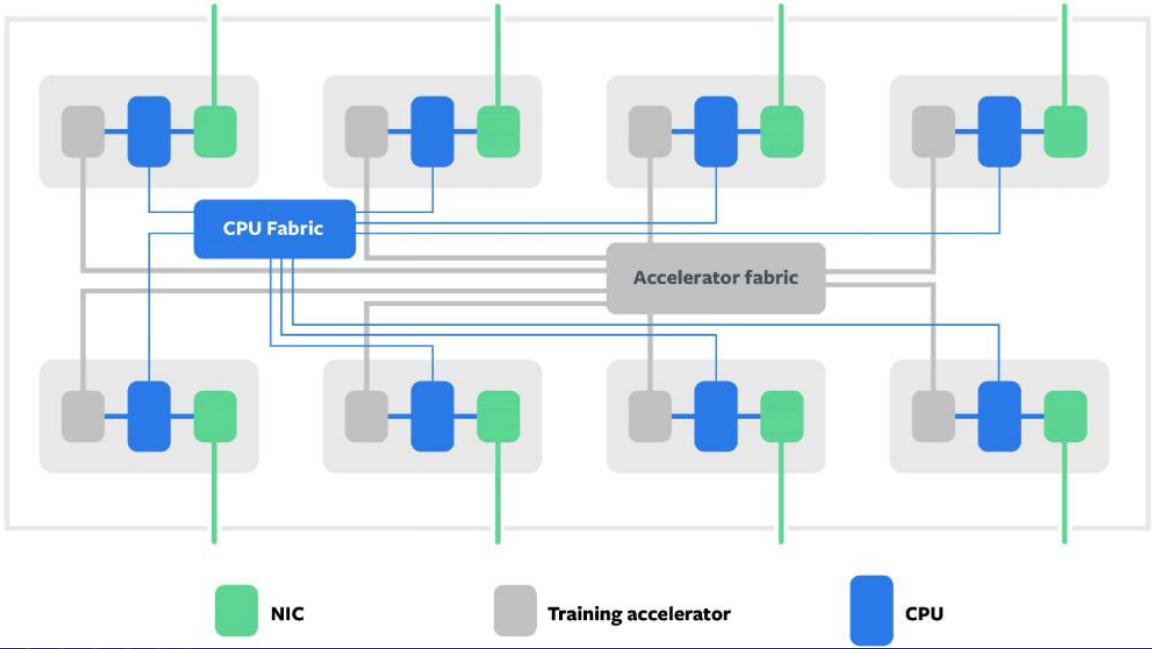
Director, AI System Software/Hardware, Facebook

OCP REGIONAL SUMMIT

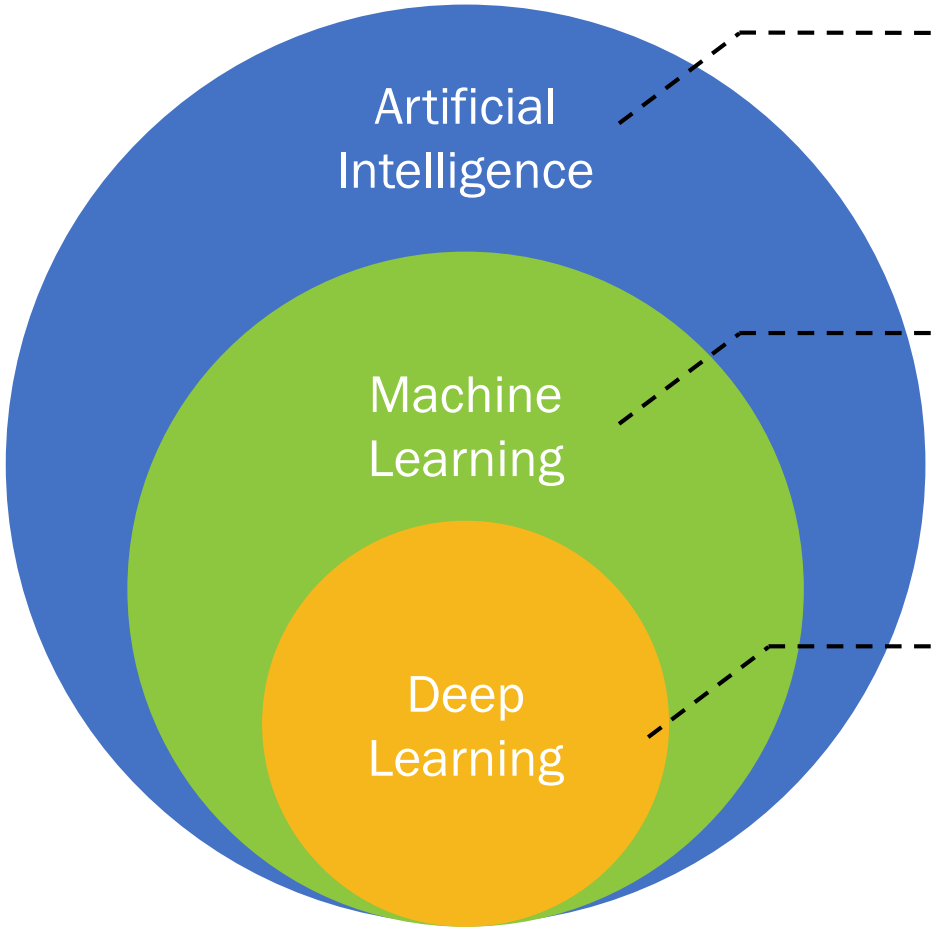
facebook

Challenges and Opportunities of Architecting AI Systems at Datacenter Scale

Misha Smelyanskiy
Director, AI Systems Co-Design — Facebook



Deep Learning



Programs with ability to learn and reason like humans

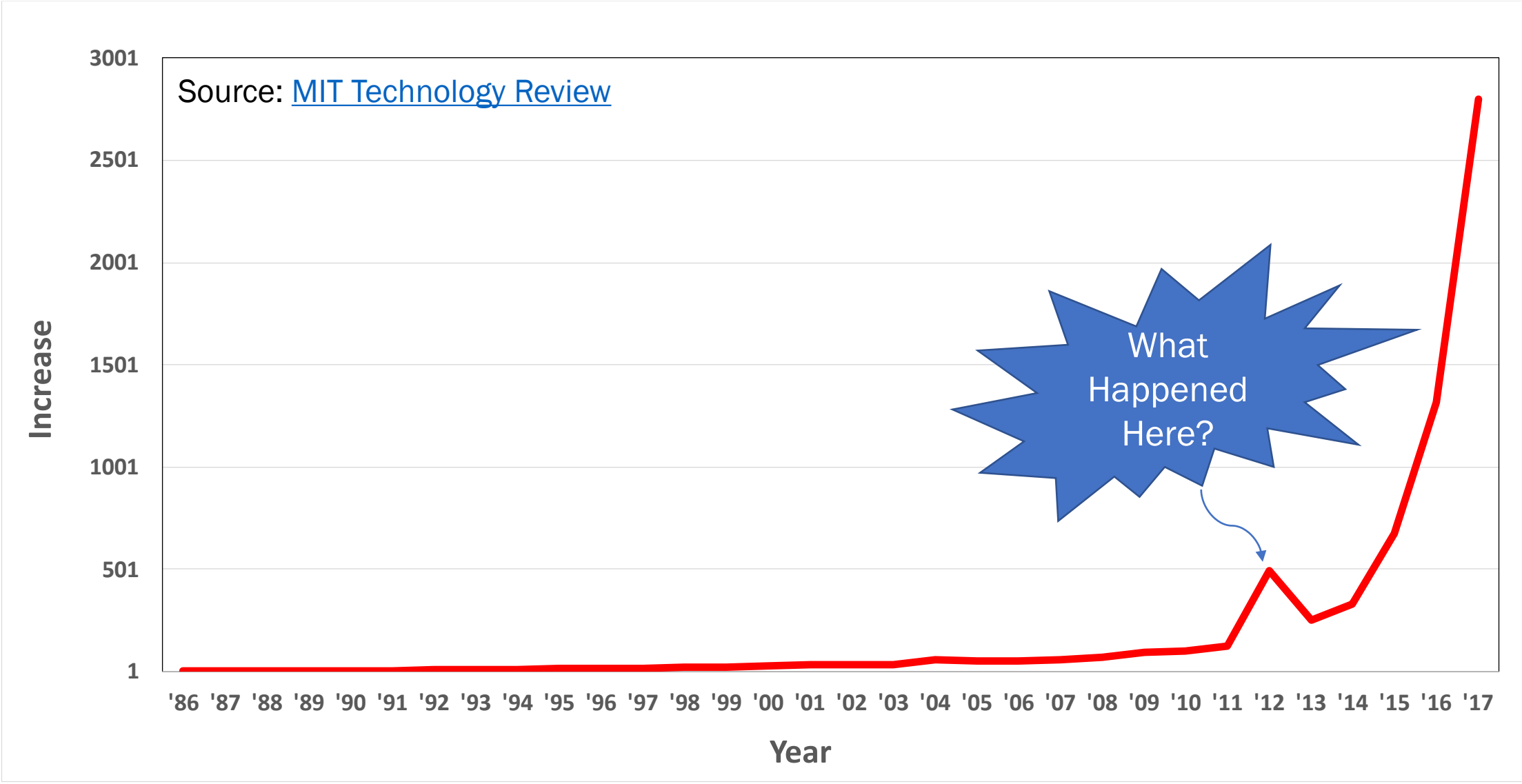
Set of statistical techniques that enable machines to improve with experience

Multi-layer neural networks which adapt and learn from vast amounts of data

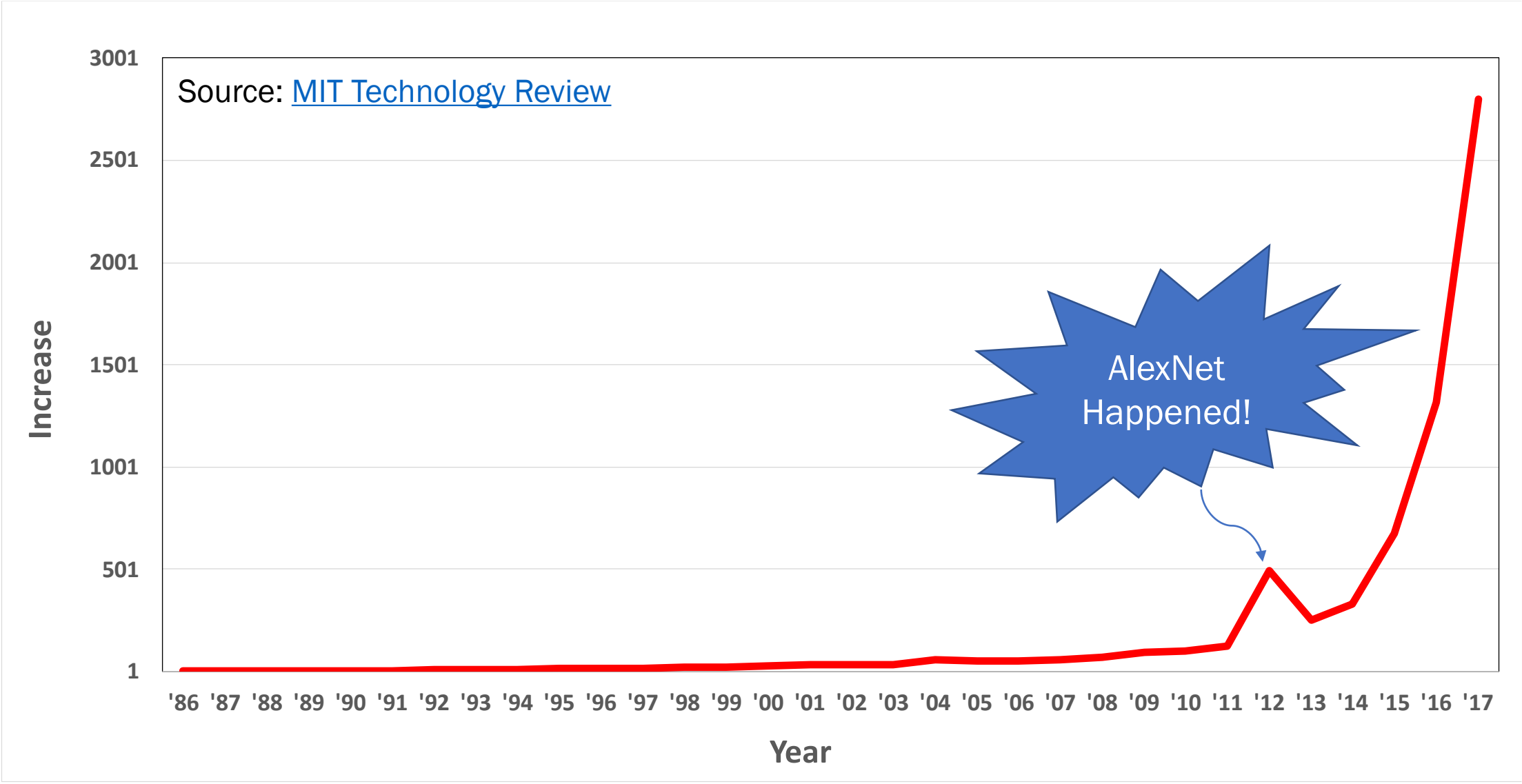
Due to deep learning success some equate it to ML and even AI

The World According to Deep Learning

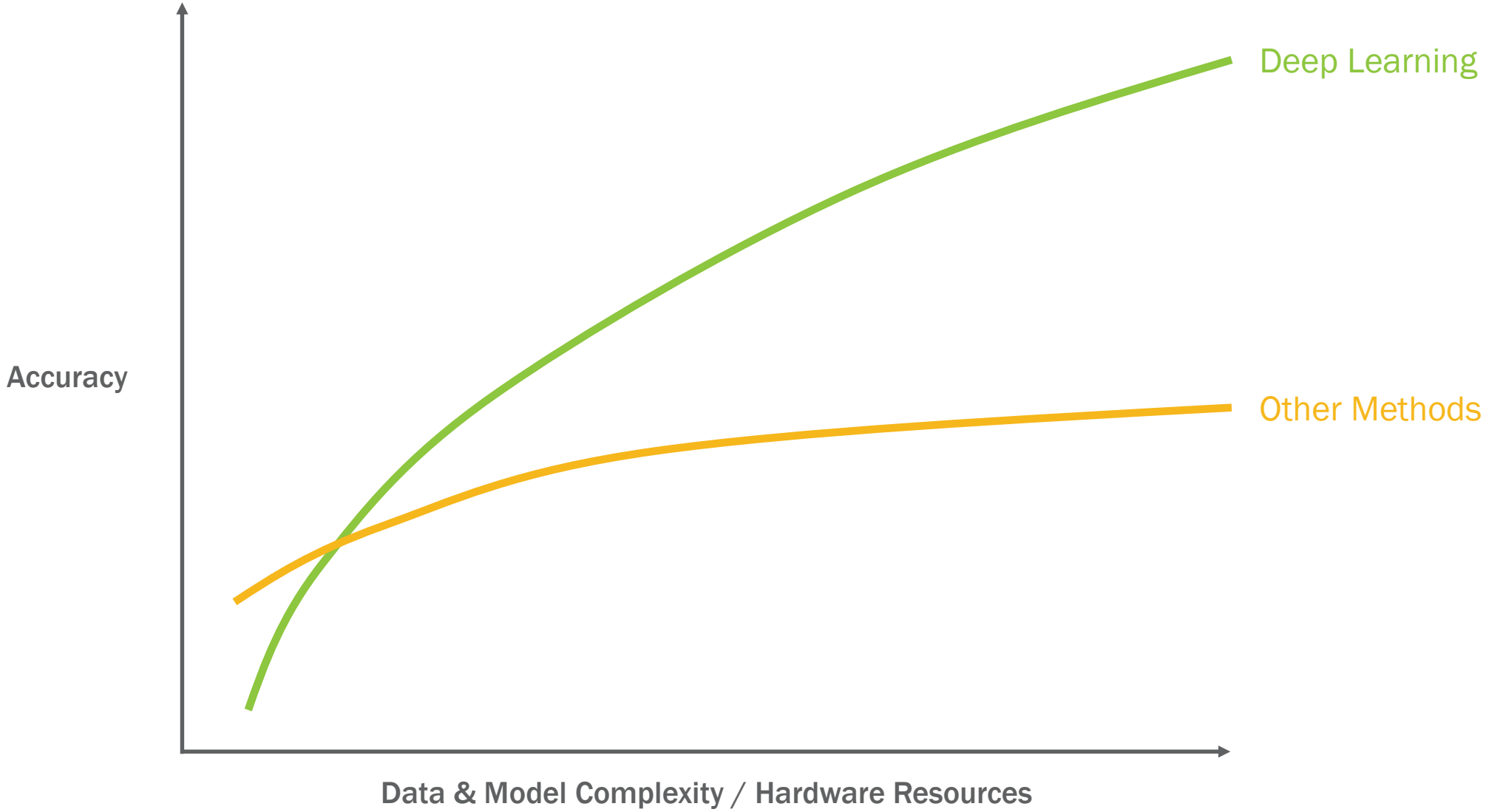
Increase of the term "deep learning" in research



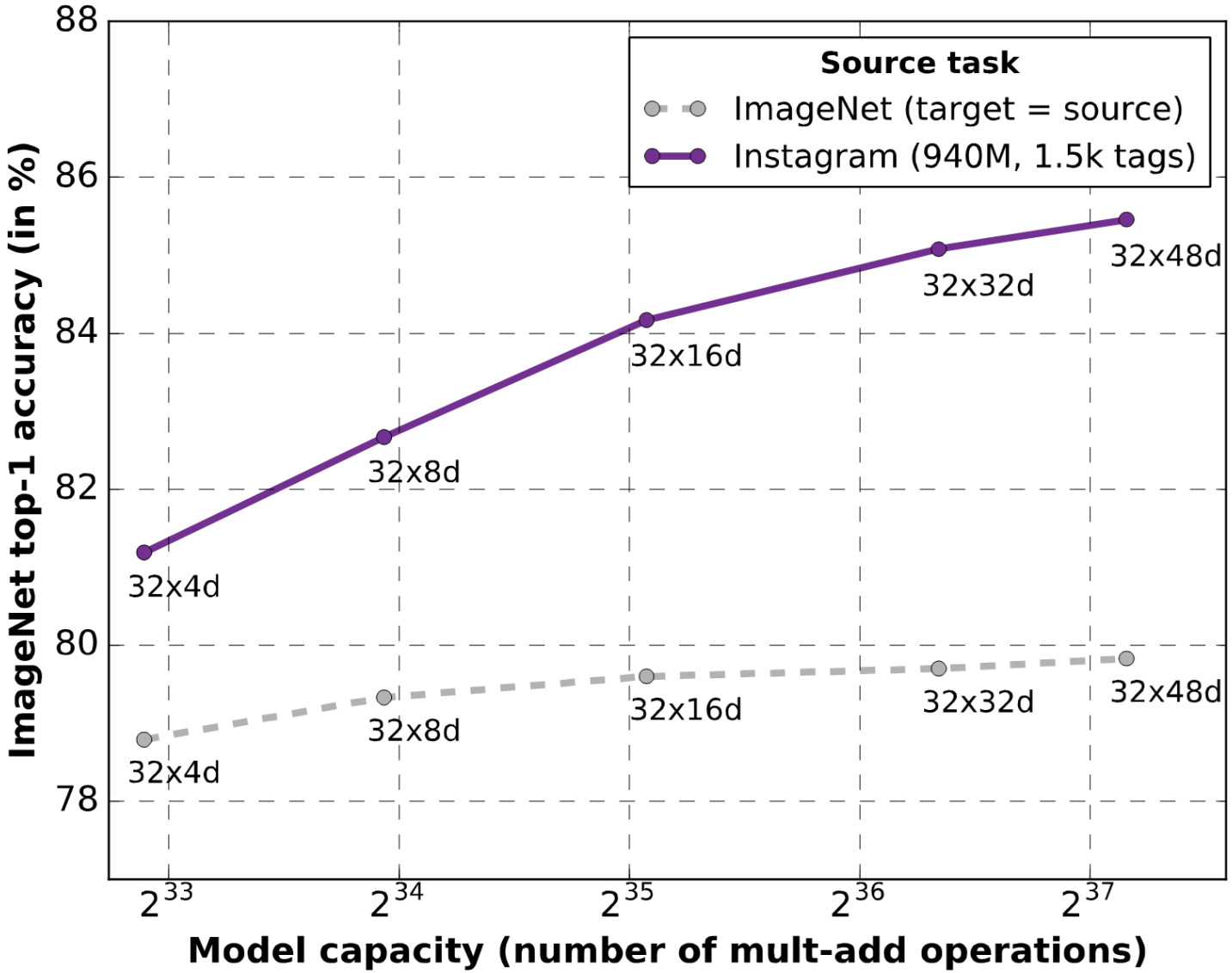
Increase of the term "deep learning" in research



Deep Learning is Unique

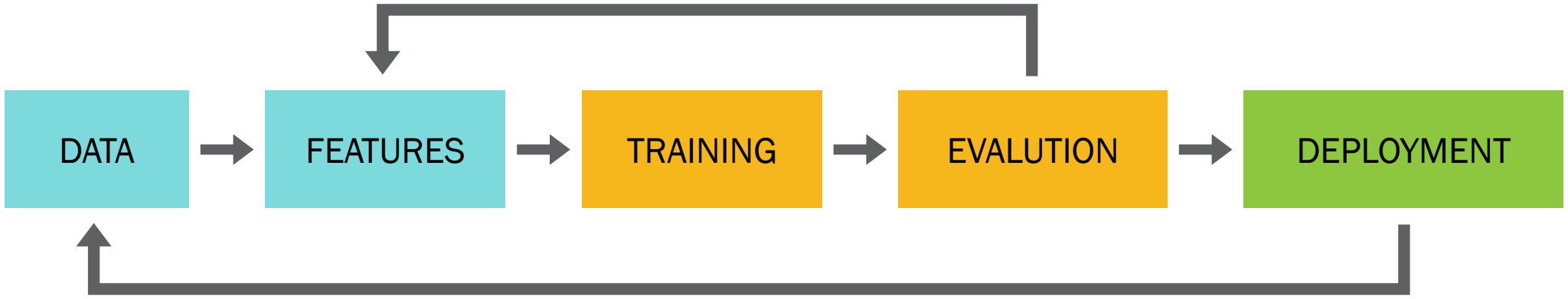


Facebook Example



Source: [Advancing state-of-the-art image recognition with deep learning on

ML Growth and Scale at Facebook



ML data growth

- Usage in **2018**: 30%
- Usage **today**: 50%
- Growth in **one year**: 3X

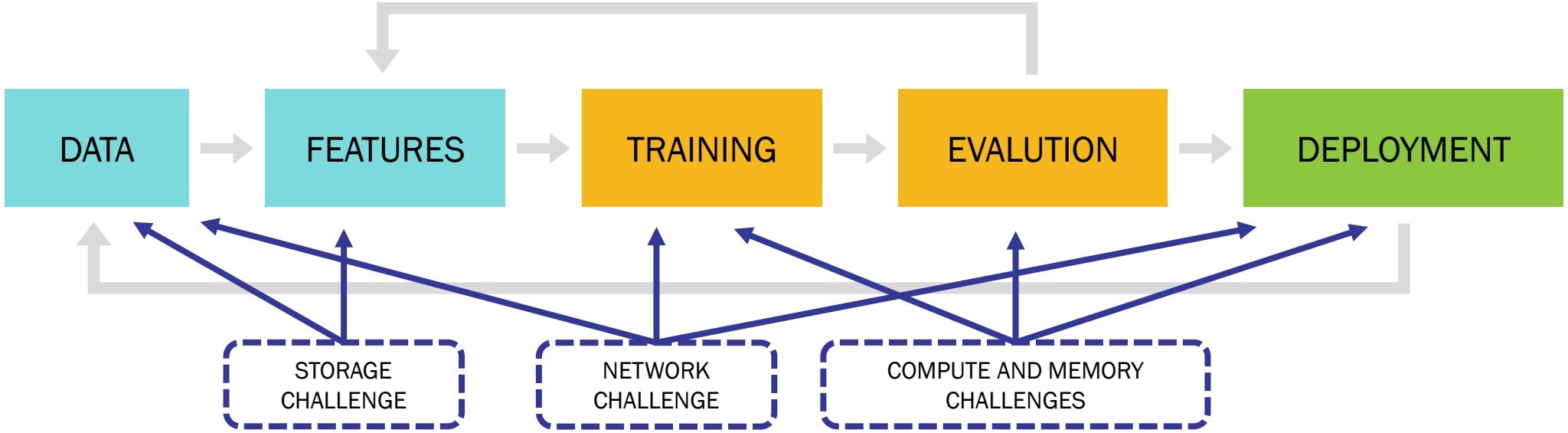
1-year Training growth

- Ranking engineers: 2X
- Workflows trained: 3X
- Compute consumed: 3X

Inference Scale per Day

- # of predictions: 200T
- # of translations: 6.5B
- Fake accounts removed: 99%

Infrastructure Challenges

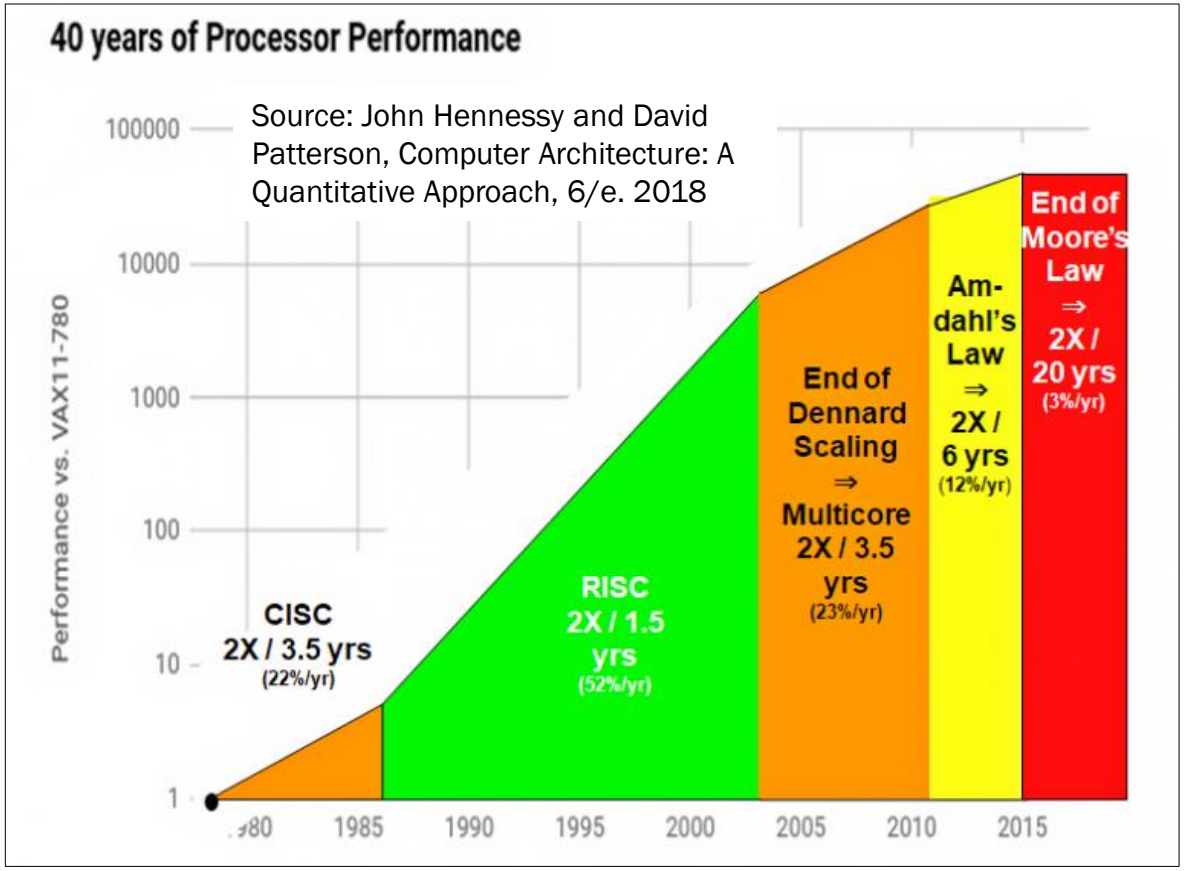


- Strains compute, memory, storage, and network
- Speed of innovation requires high-performance and flexibility

“No Exponential is Forever”

‘The important thing is that Moore's Law is exponential, and no exponential is forever…
But we can delay forever’ – *Gordon Moore*

- Data & Model Complexity ➔ Hardware Resources
- Moore’s Law has declined!
- Solution: Specialization via HW/SW co-design

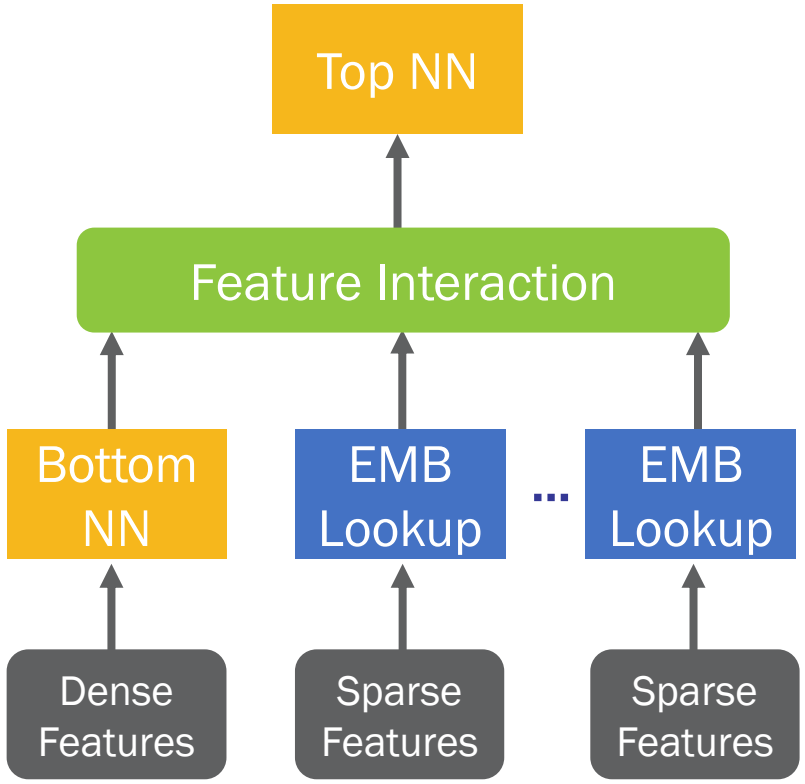


What Are The Workloads?

- Ranking and recommendation
  - news feed, and search
- Computer vision
  - image classification, object detection, and video understanding
- Language
  - translation, speech recognition, content understanding

- Recommendation models are among most important models

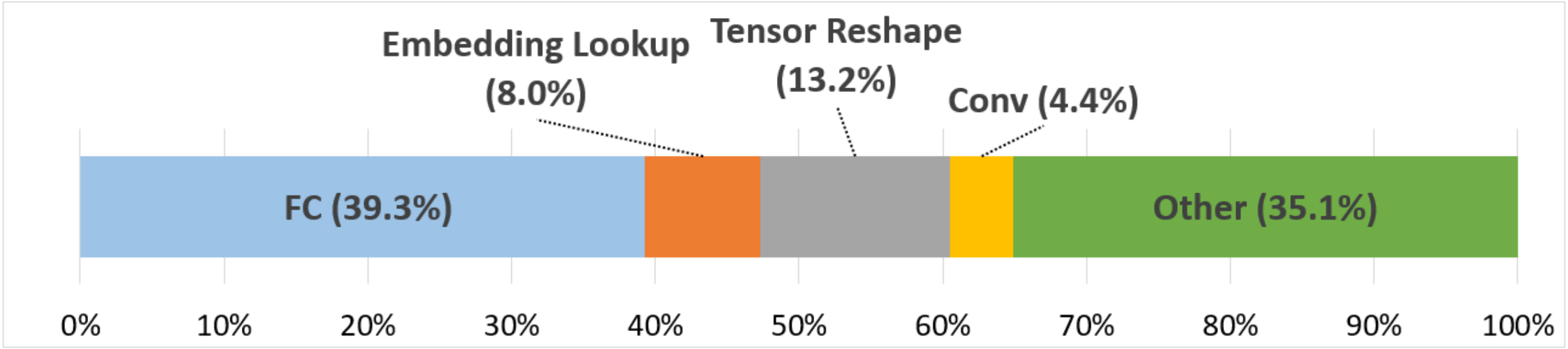
Deep Learning Recommendation Models



- DL recommendation models help user choose small set of items out of many
- Embedding look-ups result in sparse irregular accesses

Compute

It’s not all about Matrix Multiplications (MM)



Only ~40% is spent in MM in FB production

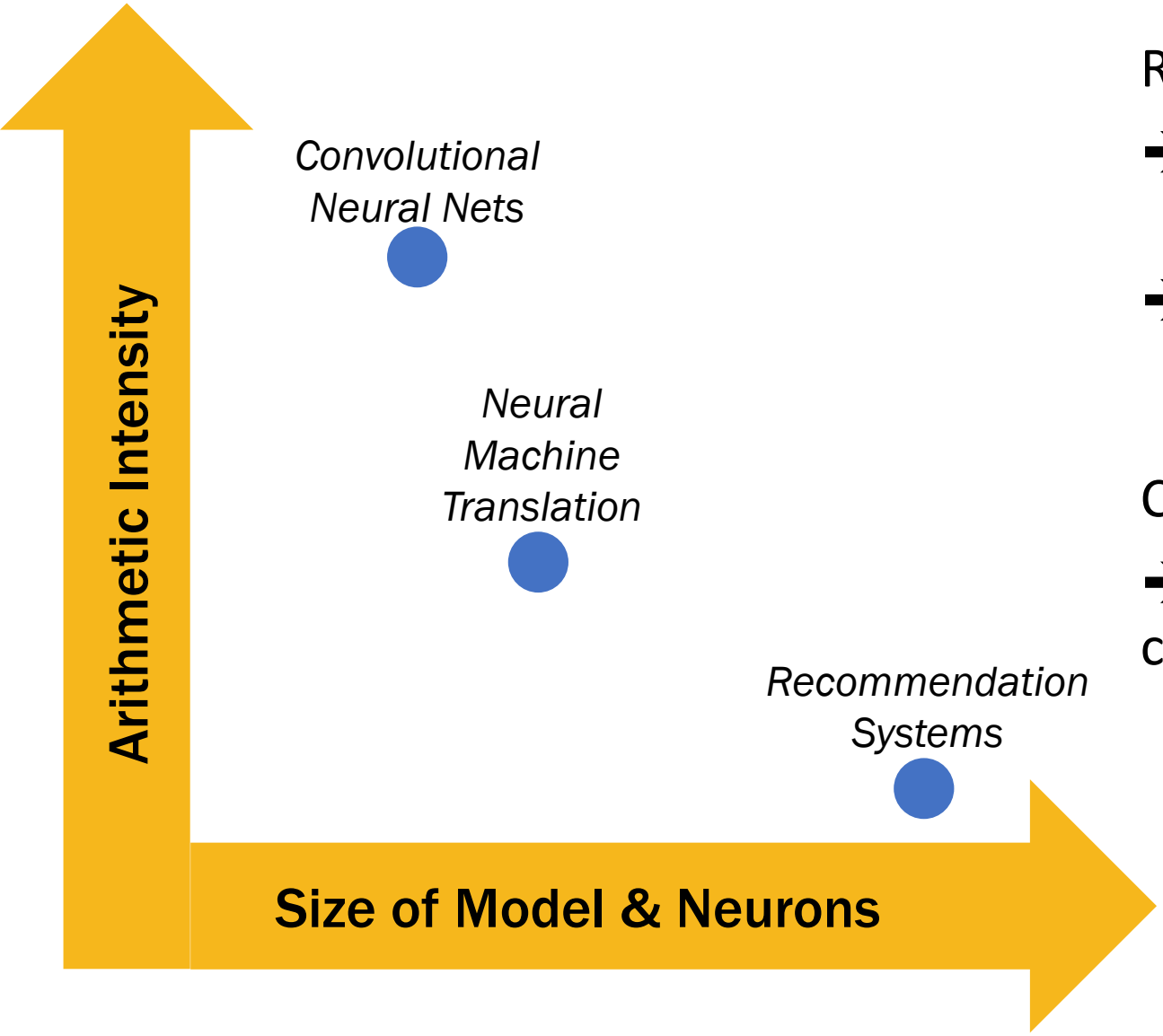
➔ Should not over-design hardware for MM and convolutions

Skinny MMs due to depth- and group-wise convolutions, small batch, beam search

➔ Fewer smaller tensor units is better than few big ones

Memory and Storage

Workload Characteristics



Rec systems are huge; low arithmetic intensity

➔Need high capacity, high bandwidth memory

➔Unstructured accesses benefit from caches

CV and language models are smaller

➔ Larger on-chip memory helps and gives compiler more flexibility

Network

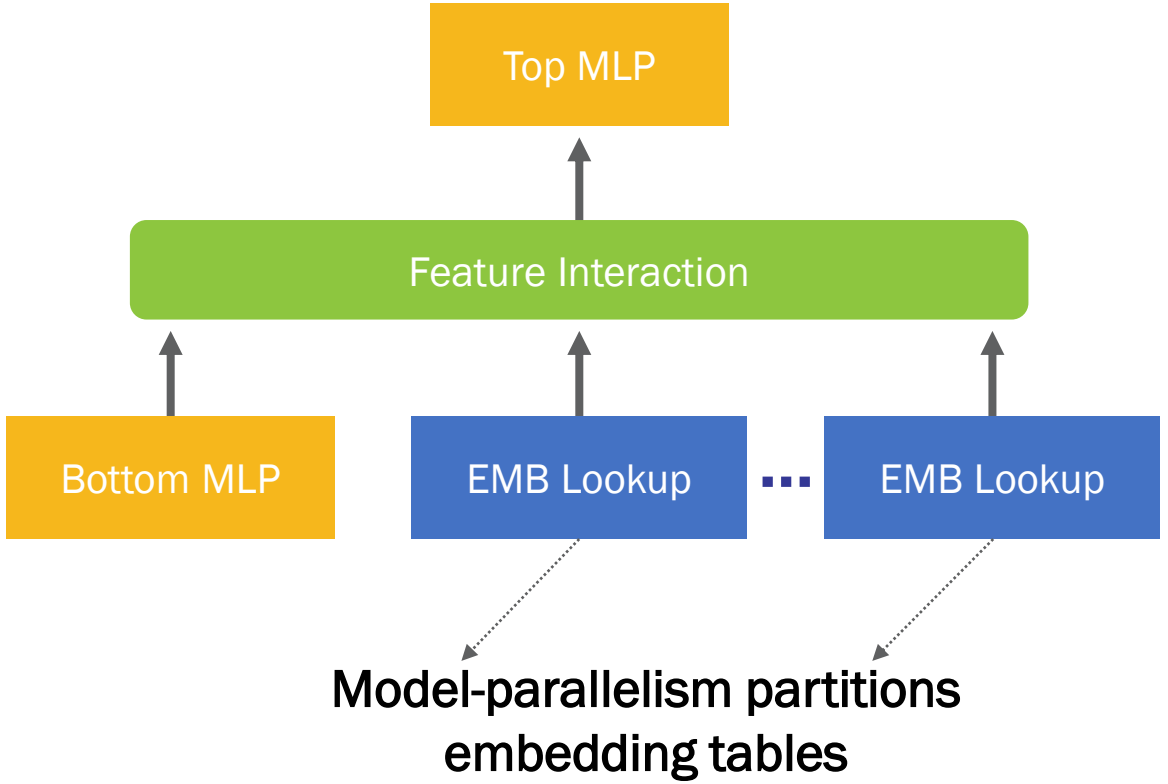
Interconnect Matters!

Model parallelism

- Different communication patterns
- Needs high bisection bandwidth

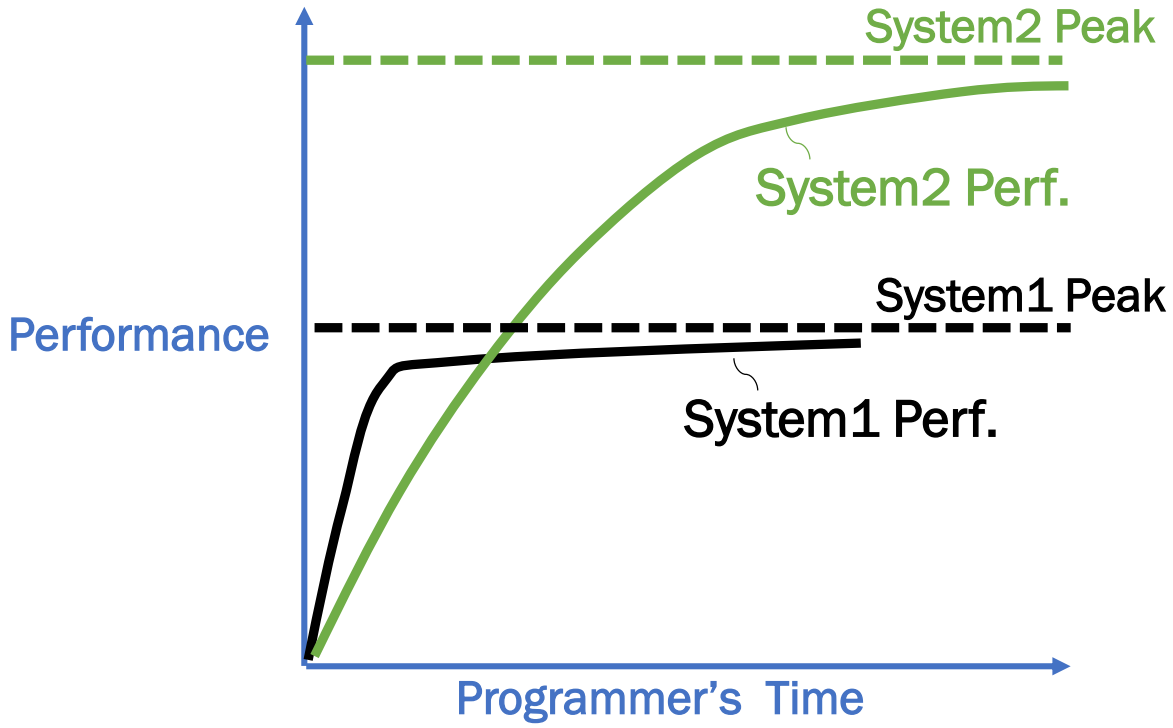
Graph learning

- New emerging application
- Need low latency, low diameter



Programmability

It is Not All About Peak Flops



Those of us who build ML HW need to think about SW at scale

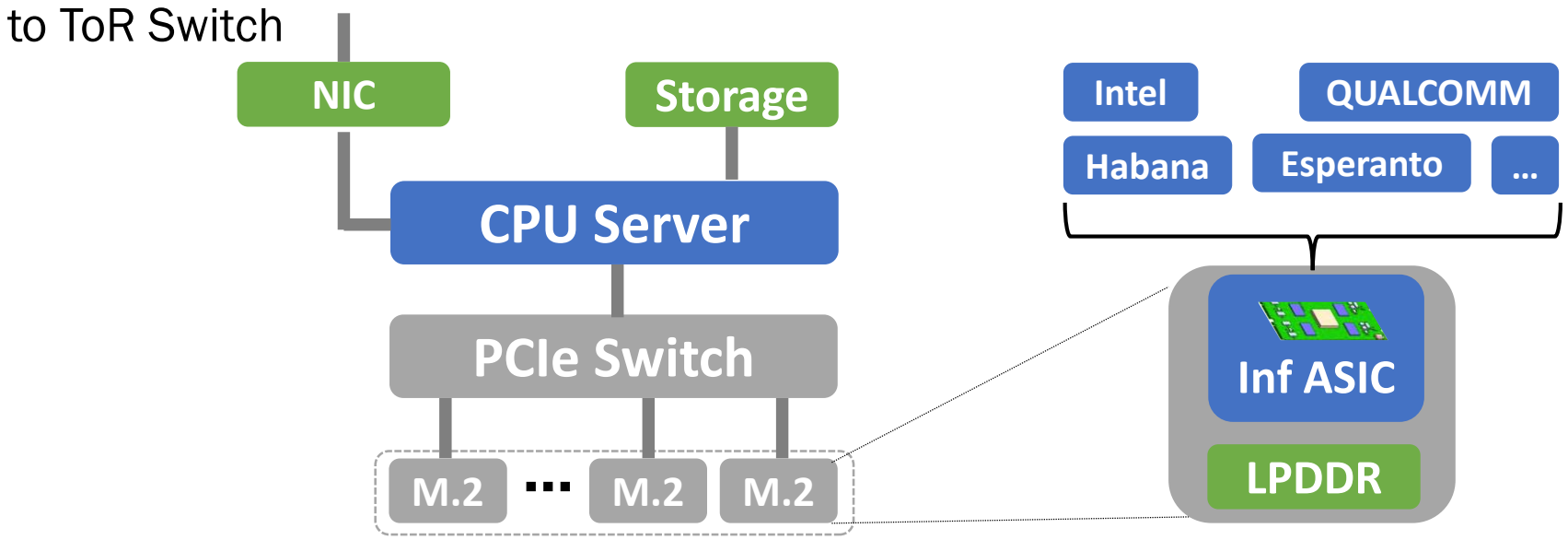
What Makes Programmability Easy (Hard)

Programmability Features	Easier To Program	Harder to Program
Concurrency & control	Few cores	Many cores
Computation	Scalar, SIMD	Tensor units
Data Reuse	Caches	SW-controlled SRAM
Communication	Cache coherence	Explicit
Latency Hiding	HW prefetcher	SW prefetch

- Specialization improves energy efficiency but limits programmability
- How do we get the best of both worlds?

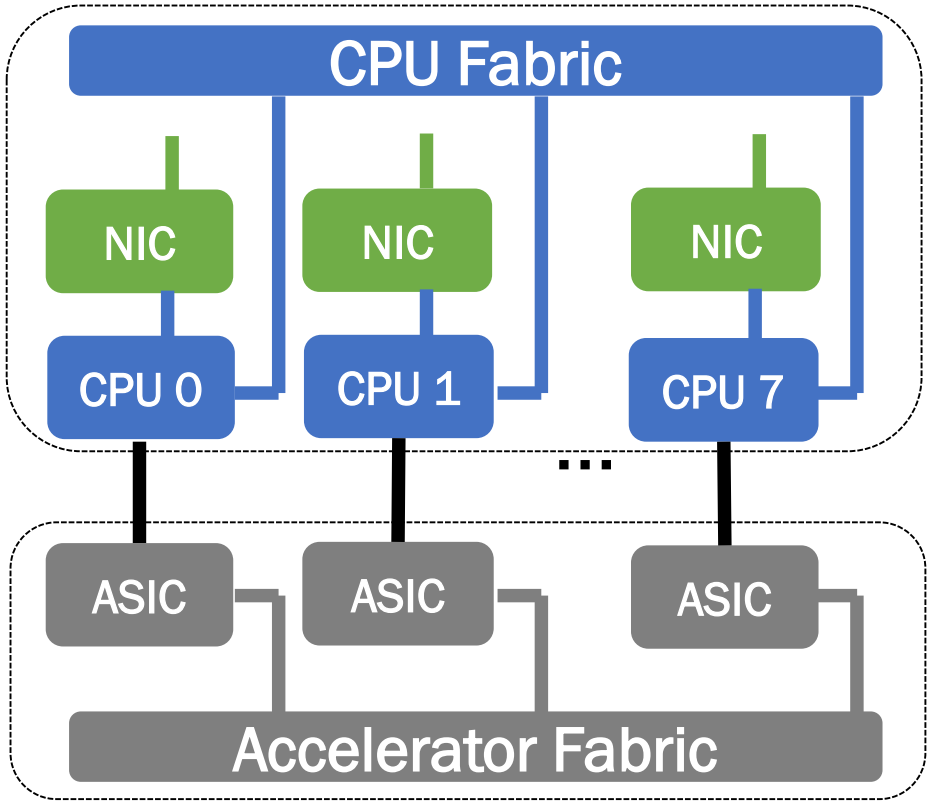
Facebook Approach

Yosemite V2 Inference Platform



- Scale-up compute, mem/SRAM capacity & BW: tightly couple via PCIe switch
- Common M.2 module, common compute and memory requirements for vendors
- Community-driven approach to programmability via GLOW compiler

Zion Training Platform



System

- Unified 100s TFLOPs of BFLOAT16
- High capacity DDR, high bandwidth HBM
- High bandwidth disaggregated fabric

Where does flexibility come from?

- OCP Accelerator Module(OAM)
- Incremental SW enablement

See Zion talk on Friday @ 9:30am by Whitney Zhao and Dheevatsa Mudigere

Call to Action

- Moore's Law slow-down requires specialization and co-design
- Need to tackle problems holistically: memory, compute, network, storage
- *The Only Constant Is Change*: exciting new developments in sparsity, graph learning, unsupervised learning, architecture search, backprop free training, …
- Performance is a start of the conversation; programmability will keep it alive!
- Our journey is only 1% finished
- Let us work together!

Open. Together.

OCP
REGIONAL
SUMMIT