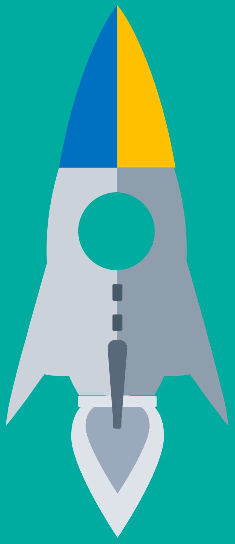
Challenges in AI and ML for Chiplets to address

Why, how and what of chiplets for AI/ML space

HiPChips Workshop @ HPCA 2023

Feb 26, 2023

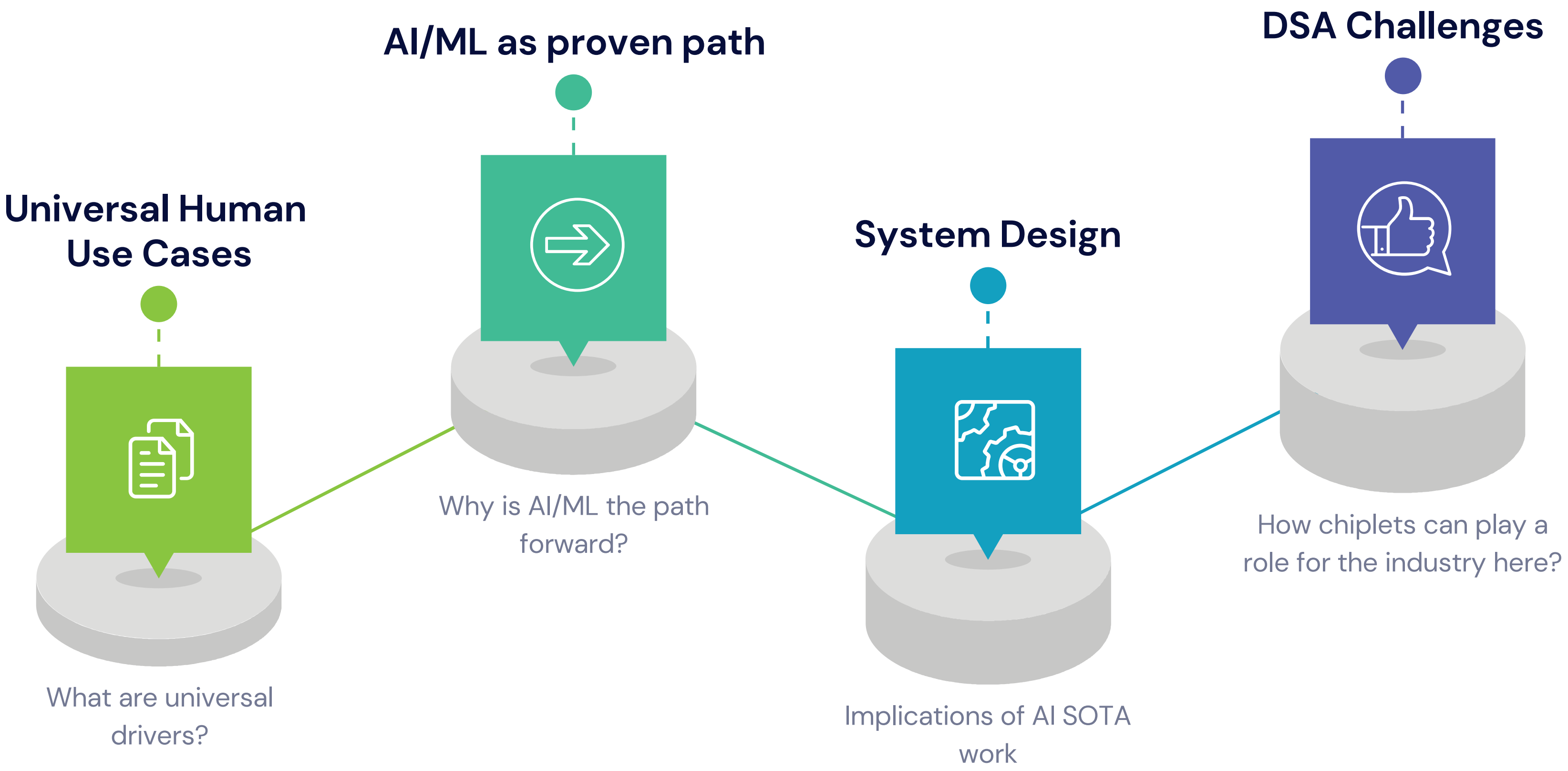


Dharmesh Jani ("DJ")

Infrastructure Partnerships/Ecosystems Lead @ Meta

Meta

Arc of the talk



Universal Human Use Cases

What are universal drivers?

AI/ML as proven path

Why is AI/ML the path forward?

System Design

Implications of AI SOTA work

DSA Challenges

How chiplets can play a role for the industry here?

Arc of the talk

Universal Human Use Cases

What are universal drivers?

Universal use cases that are drive technology

Recognition

Mining

Synthesis

Fundamental use cases have recurring theme of recognition, mining and synthesis for learning and knowledge creation

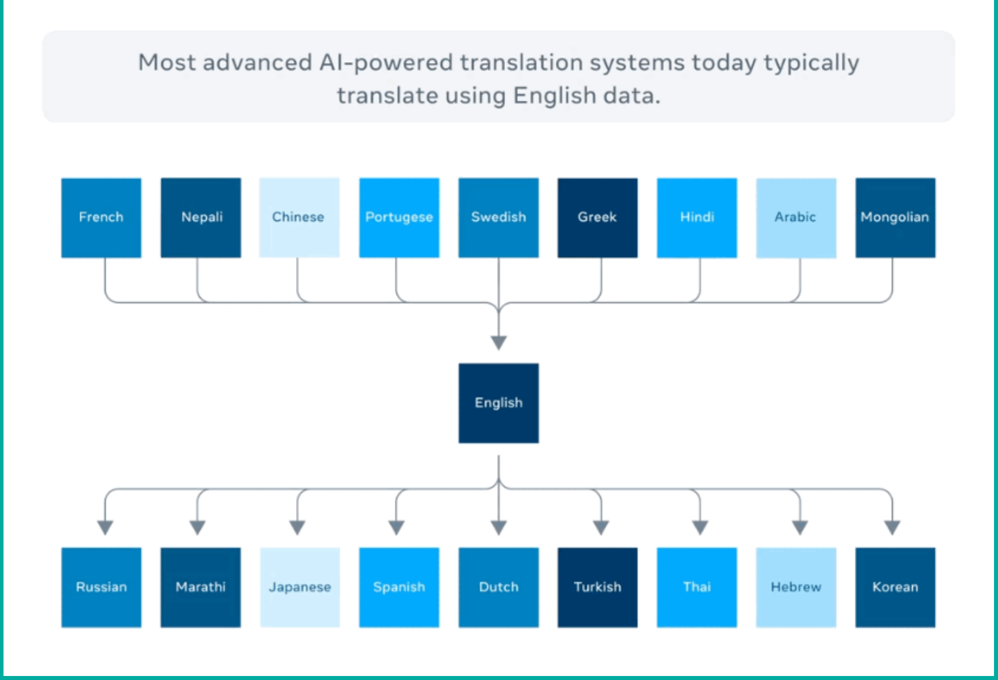
Universal use cases that are drive technology

Recognition

Build identification models by machines of real world

Recognition is the “what is” and create a canonical representative model

Requires training!



Most advanced AI-powered translation systems today typically translate using English data.

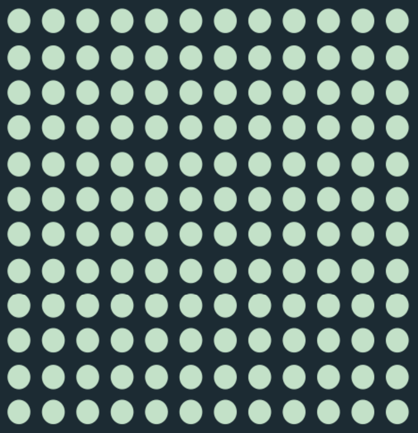
Universal use cases that are drive technology

Mining

Search instances of the model in the sea of data

Mining is searching across all forms of data (e.g., Image, text, video, logs etc.)

Requires inference!



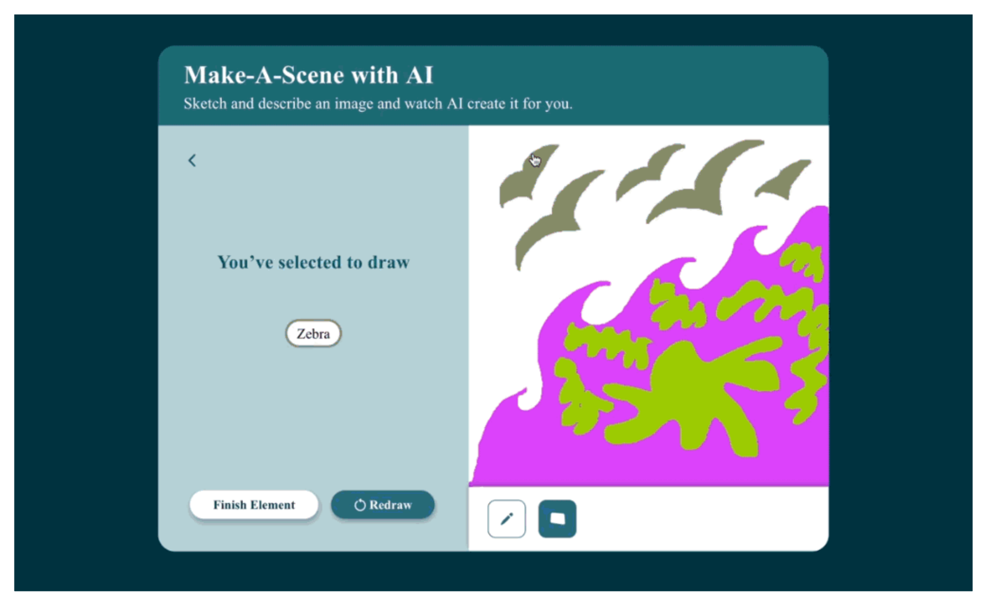
Universal use cases that are drive technology

Synthesis

Creating new instance of models where one does not exist

Synthesis is creation by machines of new ideas

Requires multi-modality, GANs!



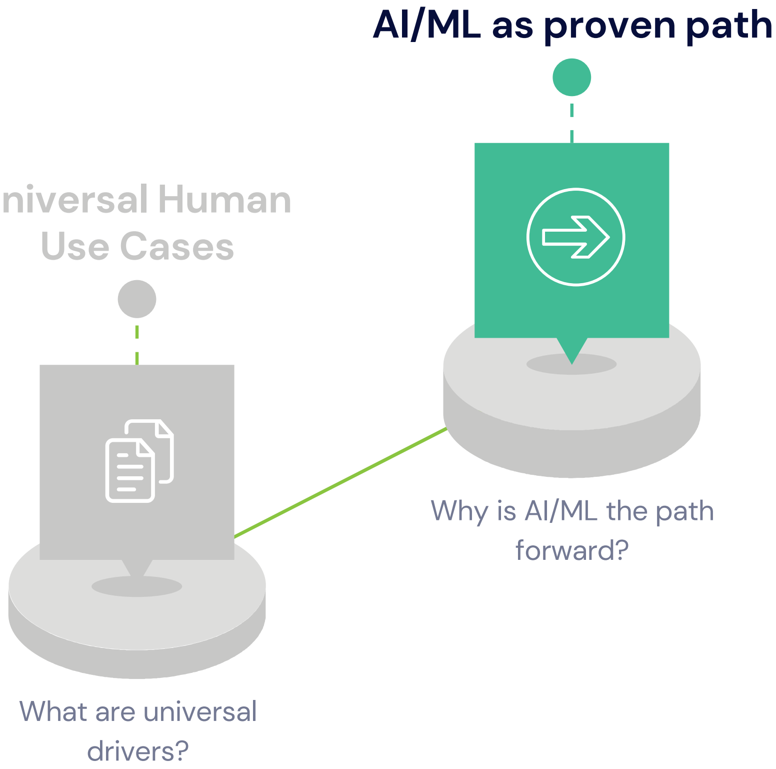
Make-A-Scene with AI

Sketch and describe an image and watch AI create it for you.

You've selected to draw

Zebra

Arch of the talk



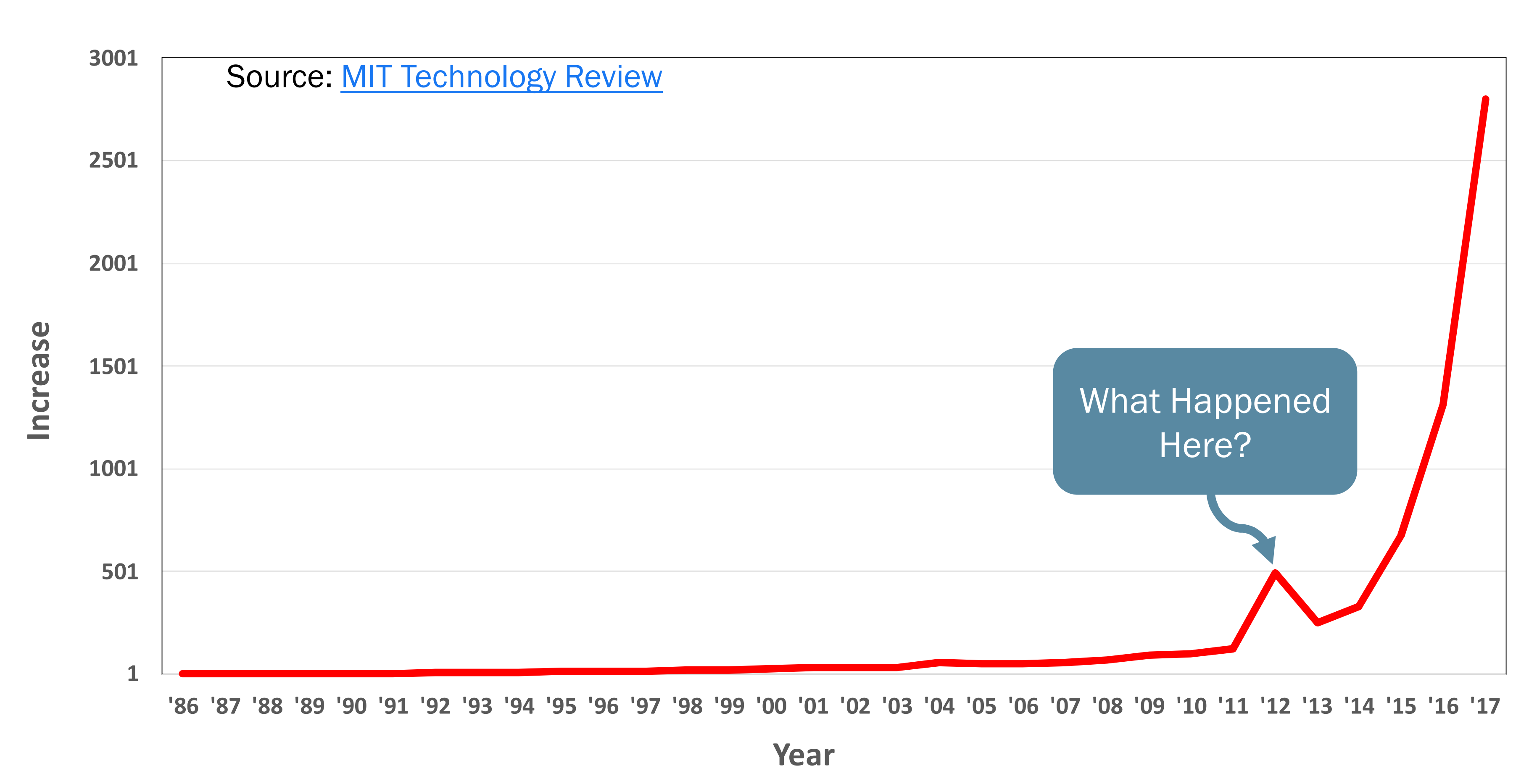
AI/ML as proven path

Why is AI/ML the path forward?

Universal Human Use Cases

What are universal drivers?

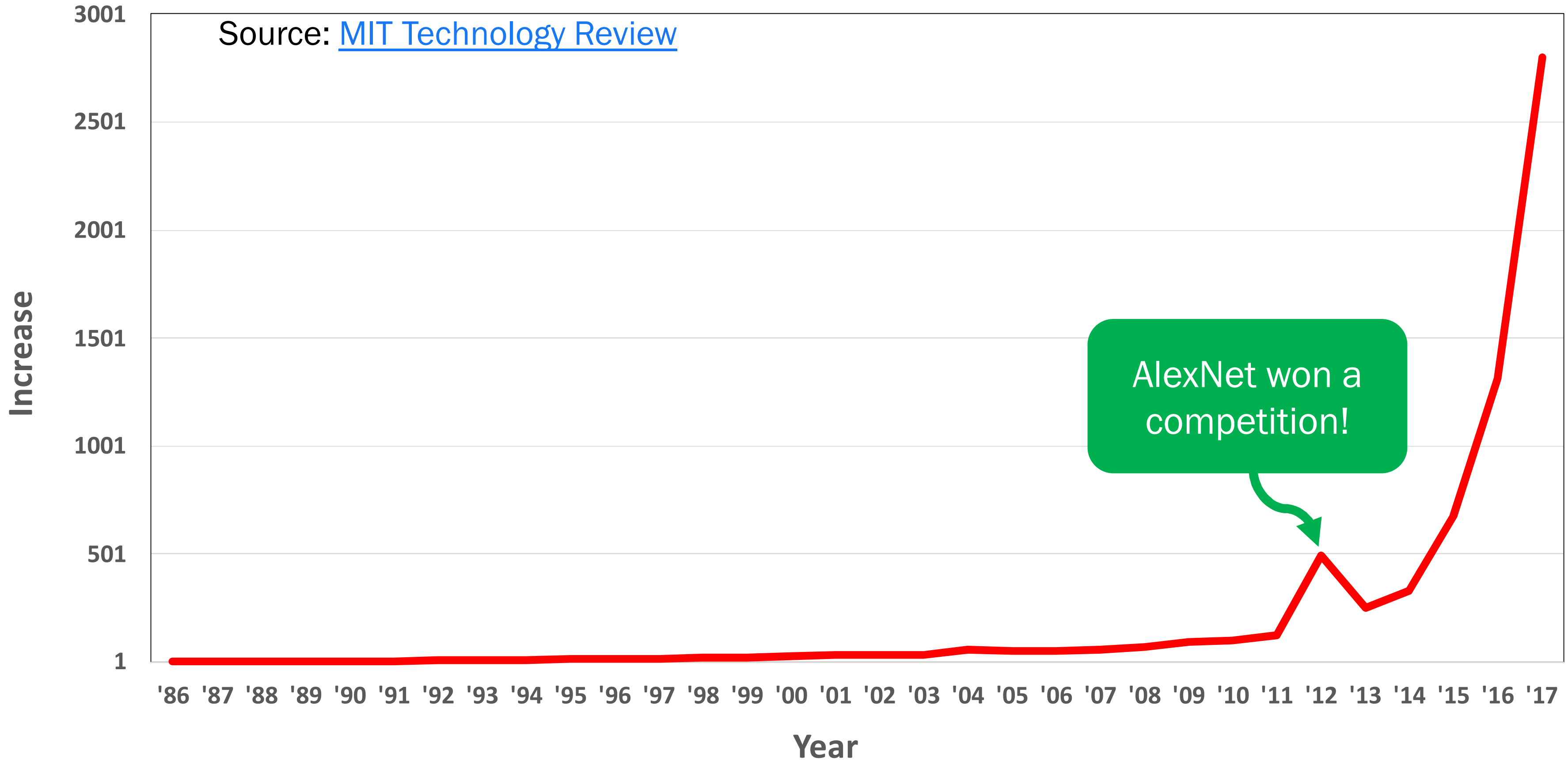
Growth of the term "deep learning" in research



Source: MIT Technology Review

What Happened Here?

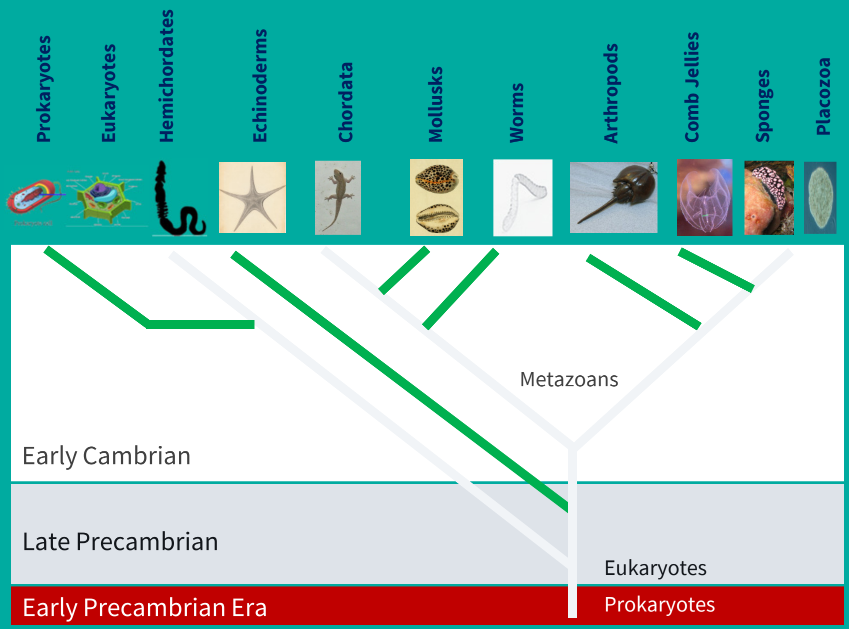
Growth of the term "deep learning" in research



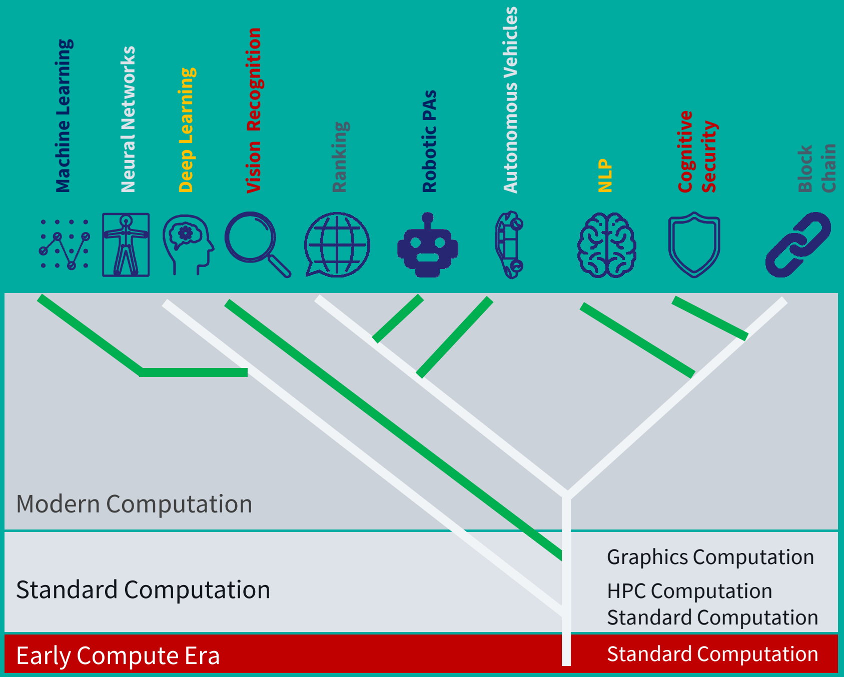
Source: MIT Technology Review

AlexNet won a competition!

Cambrian Explosion of Workloads



Bio-Diversity Exploded from single cells into multi-cell organisms during the Cambrian explosion; all major phylla were established in this transition

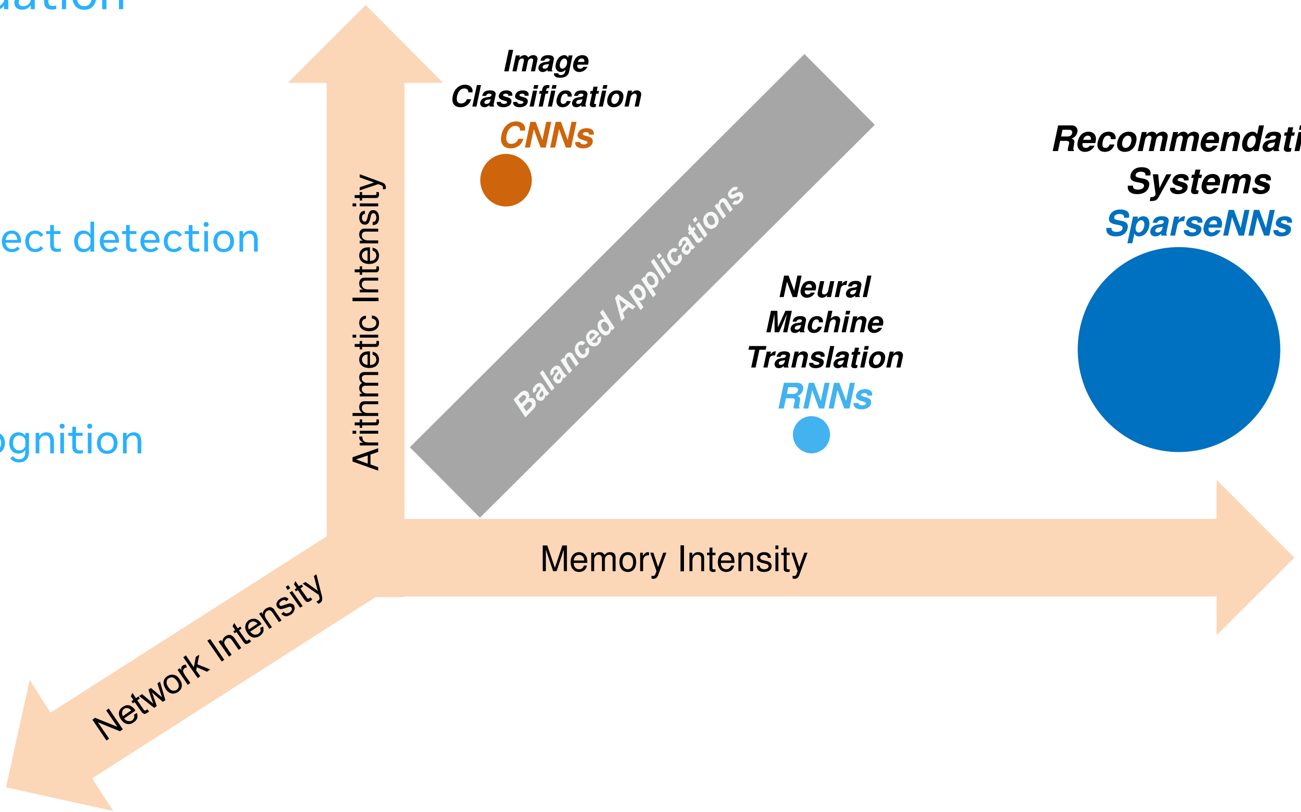


AI and Machine-learning and data-heavy workloads have exploded in 7 years and will diversify as new applications are discovered constantly…

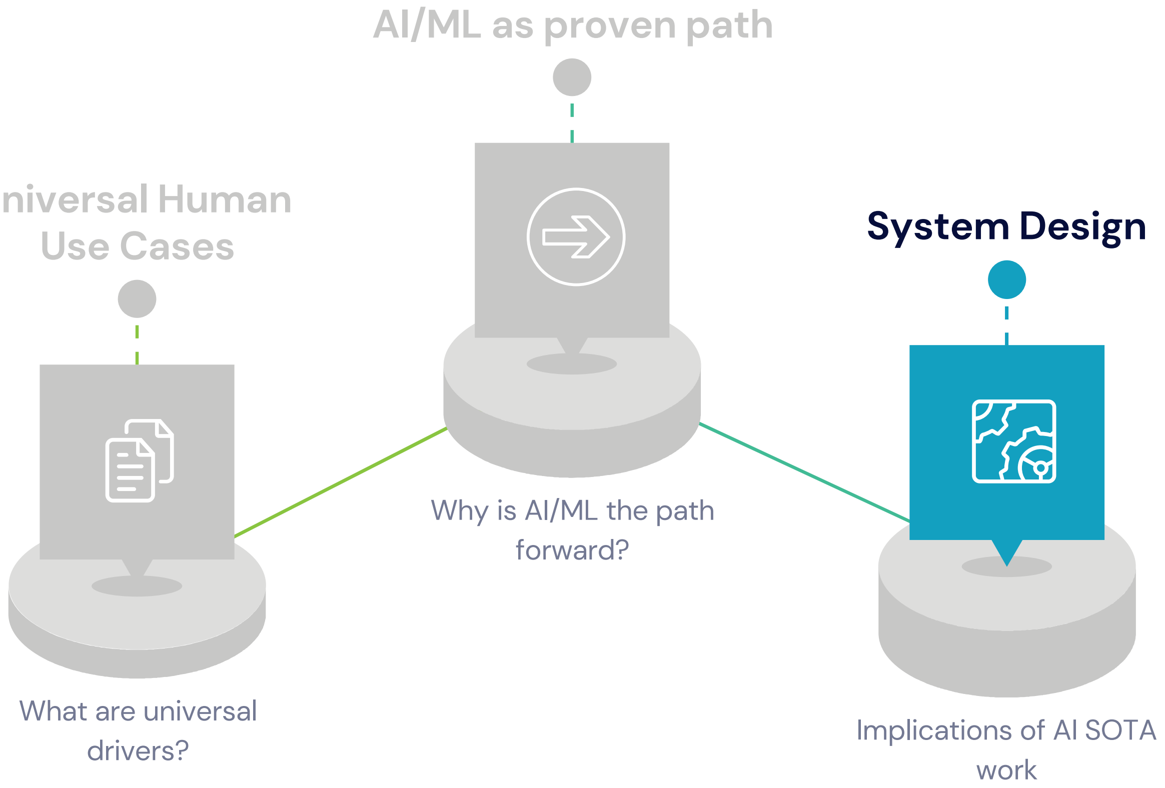
What are the dominant AI serving workloads?

Current and emergent

- Ranking and recommendation
  - News feed and Search
- Computer Vision
  - Image classification, object detection
- Language
  - Translation, speech recognition
- Multi-modal
  - Metaverse synthesis



Arc of the talk



Universal Human Use Cases

What are universal drivers?

AI/ML as proven path

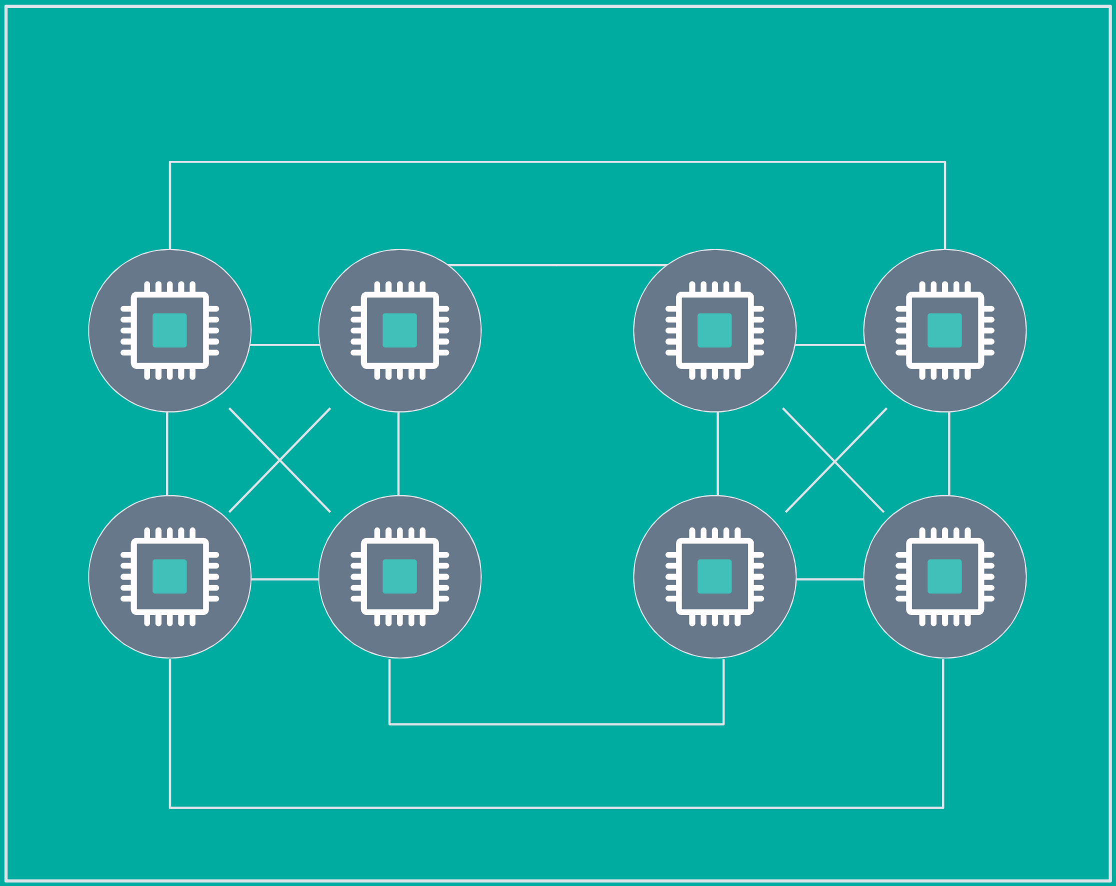
Why is AI/ML the path forward?

System Design

Implications of AI SOTA work

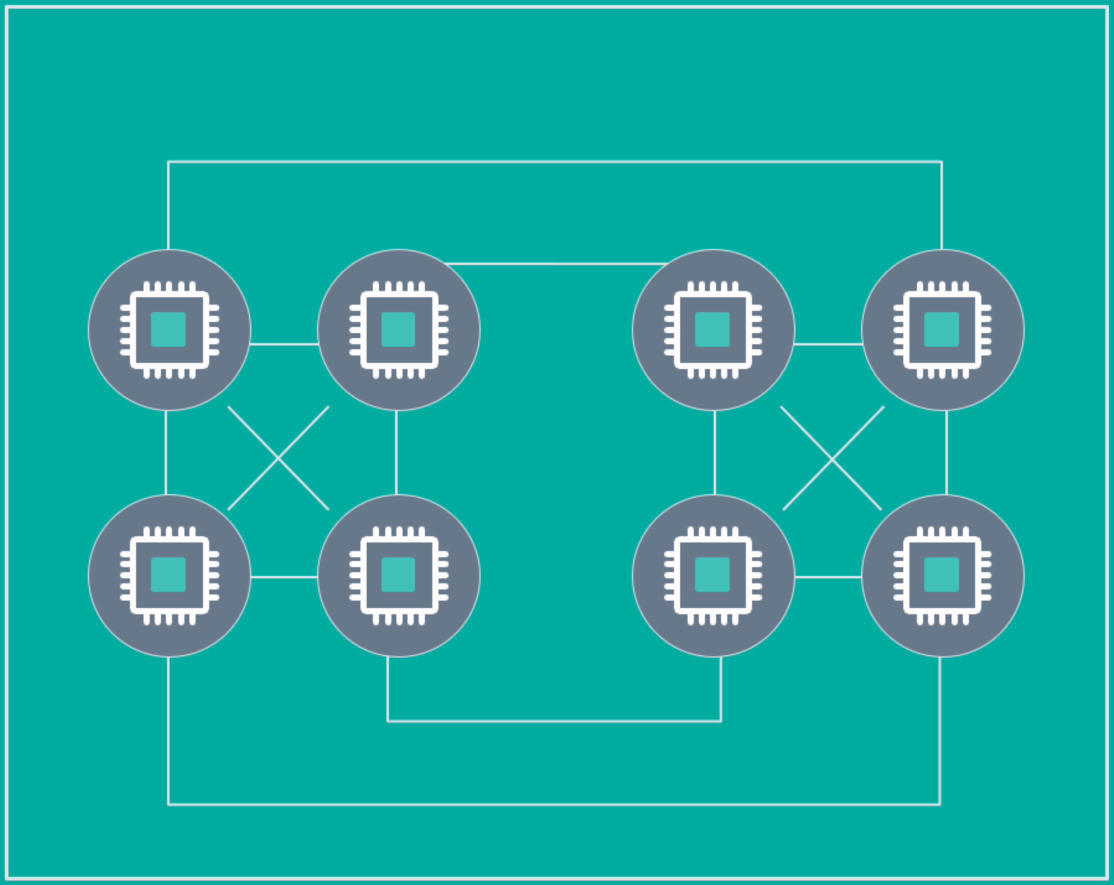
Domain Specific Accelerators

ACCELERATOR WORKLOAD UNIT



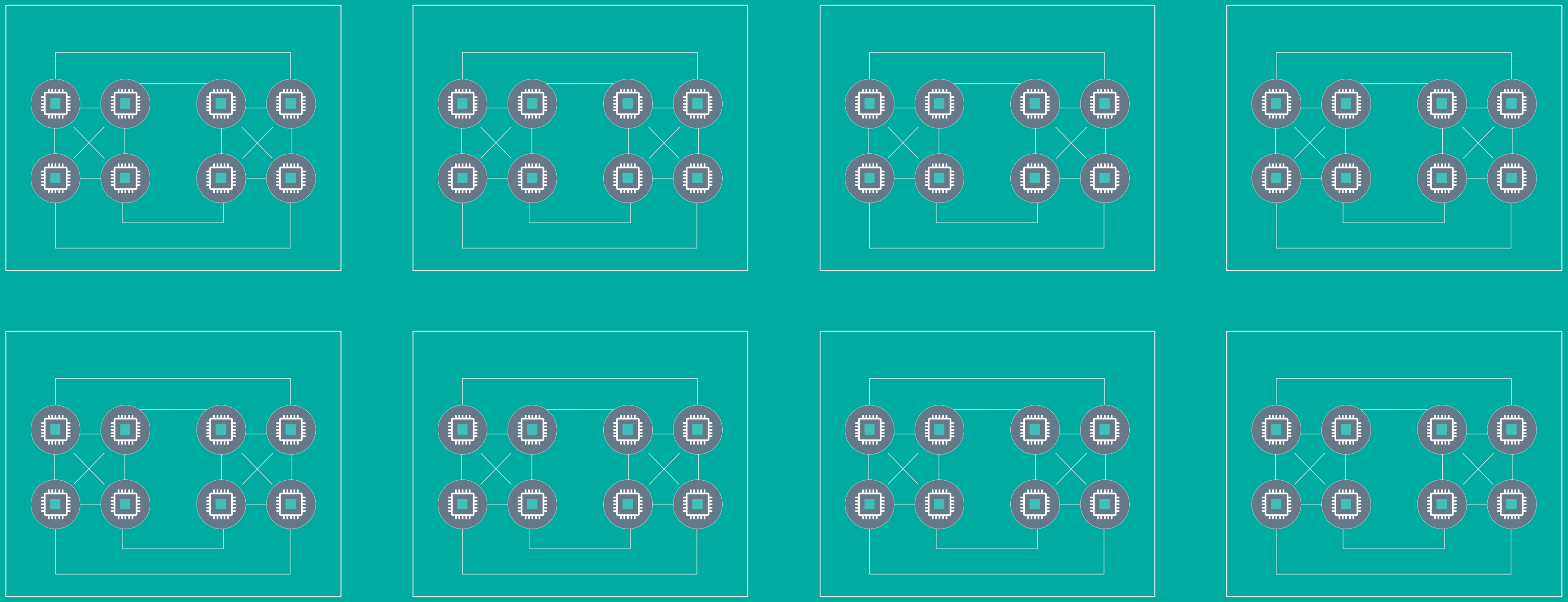
*ignoring the CPU, NICs, SSDs, and everything else...

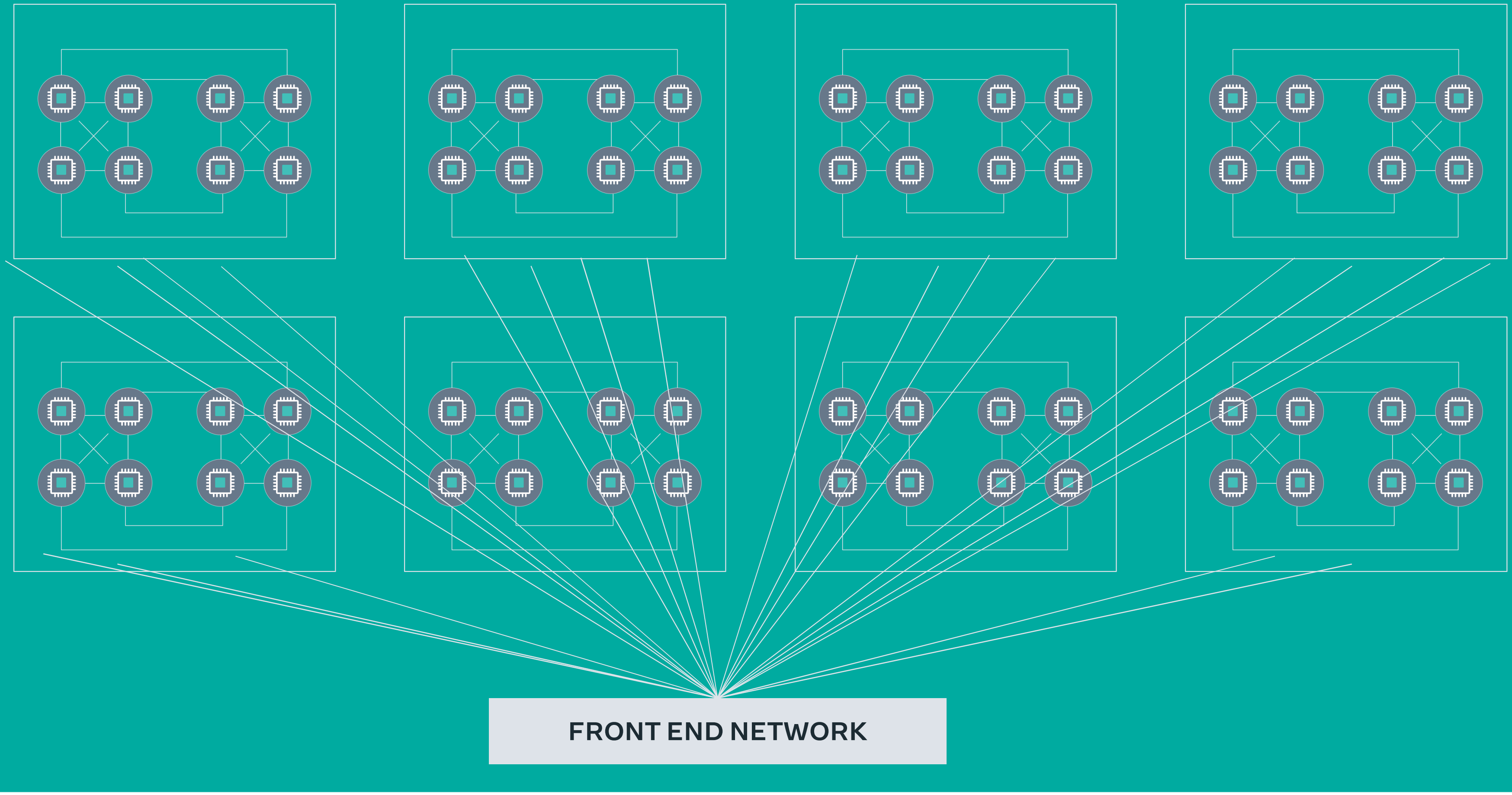
ACCELERATOR WORKLOAD UNIT

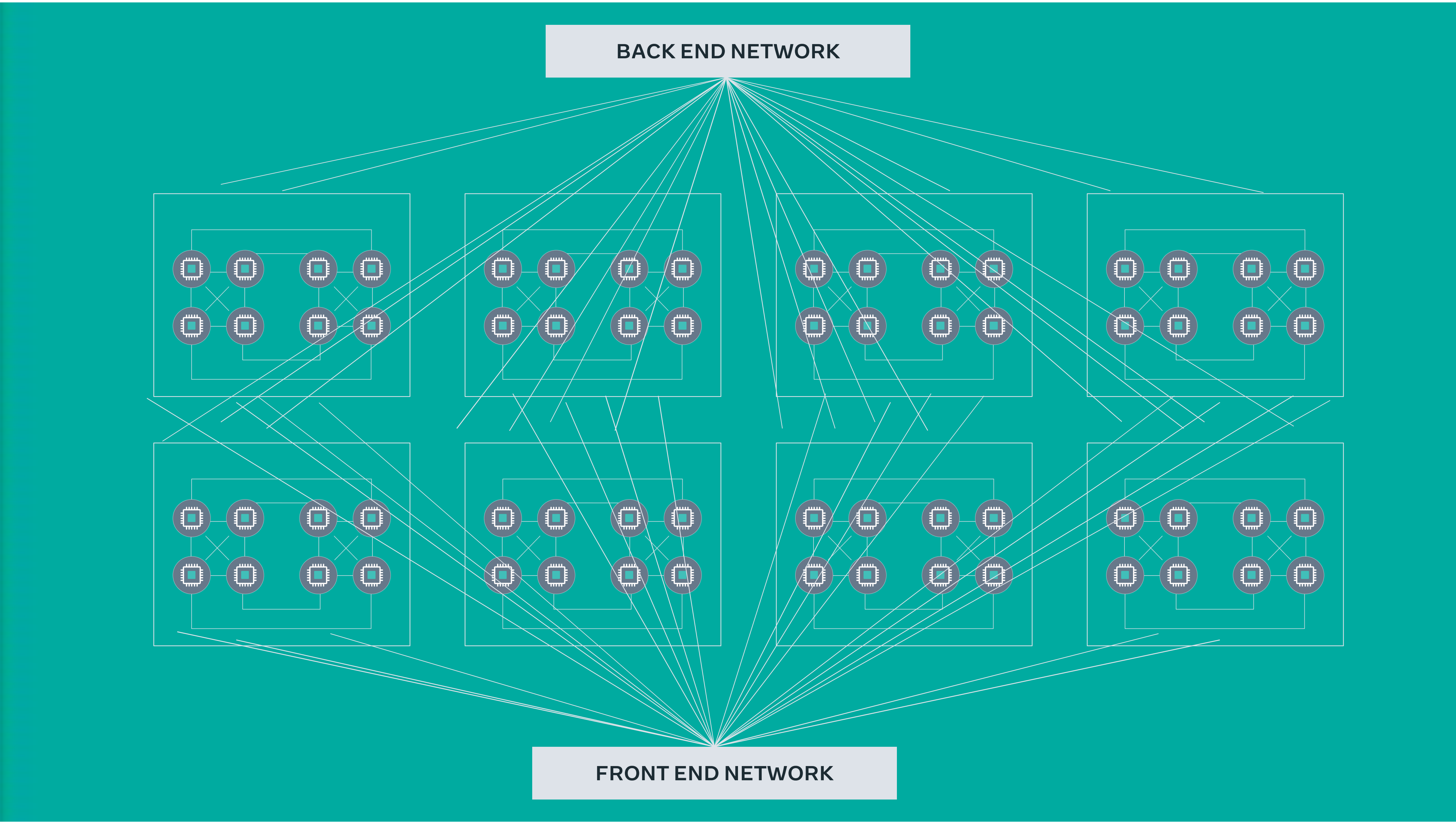


*ignoring the CPU, NICs, SSDs, and everything else...

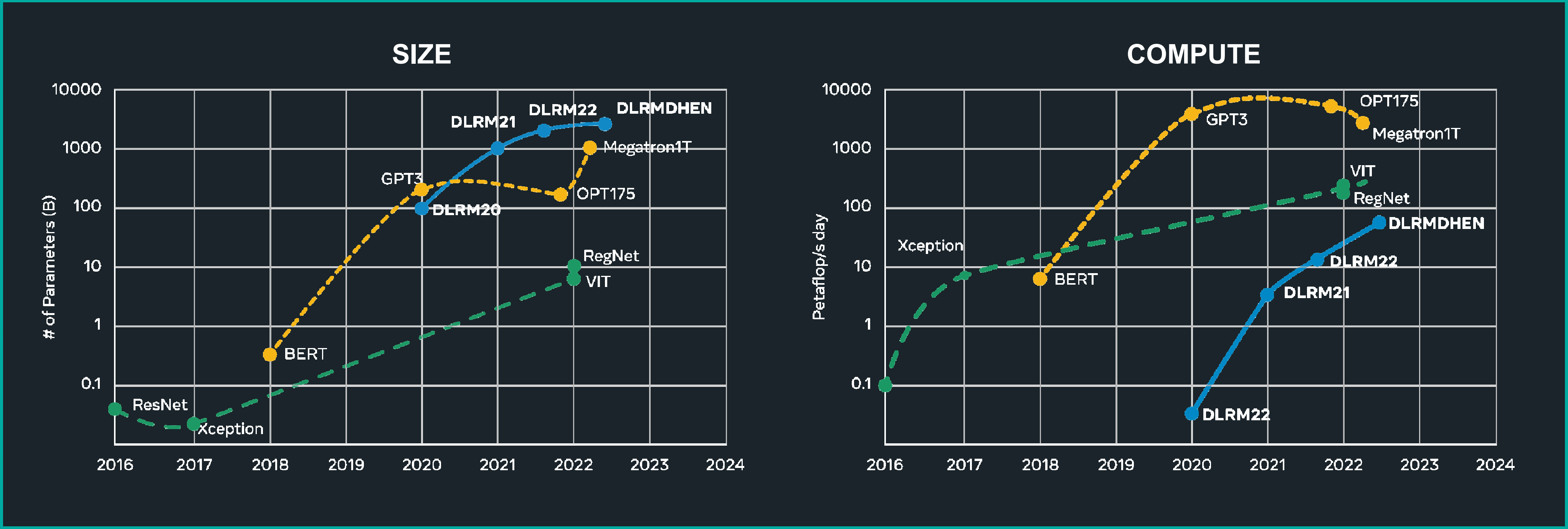
ACCELERATOR WORKLOAD CLUSTER




FRONT END NETWORK


BACK END NETWORK
FRONT END NETWORK

DEEP LEARNING WORKLOADS - CHARACTERISTICS

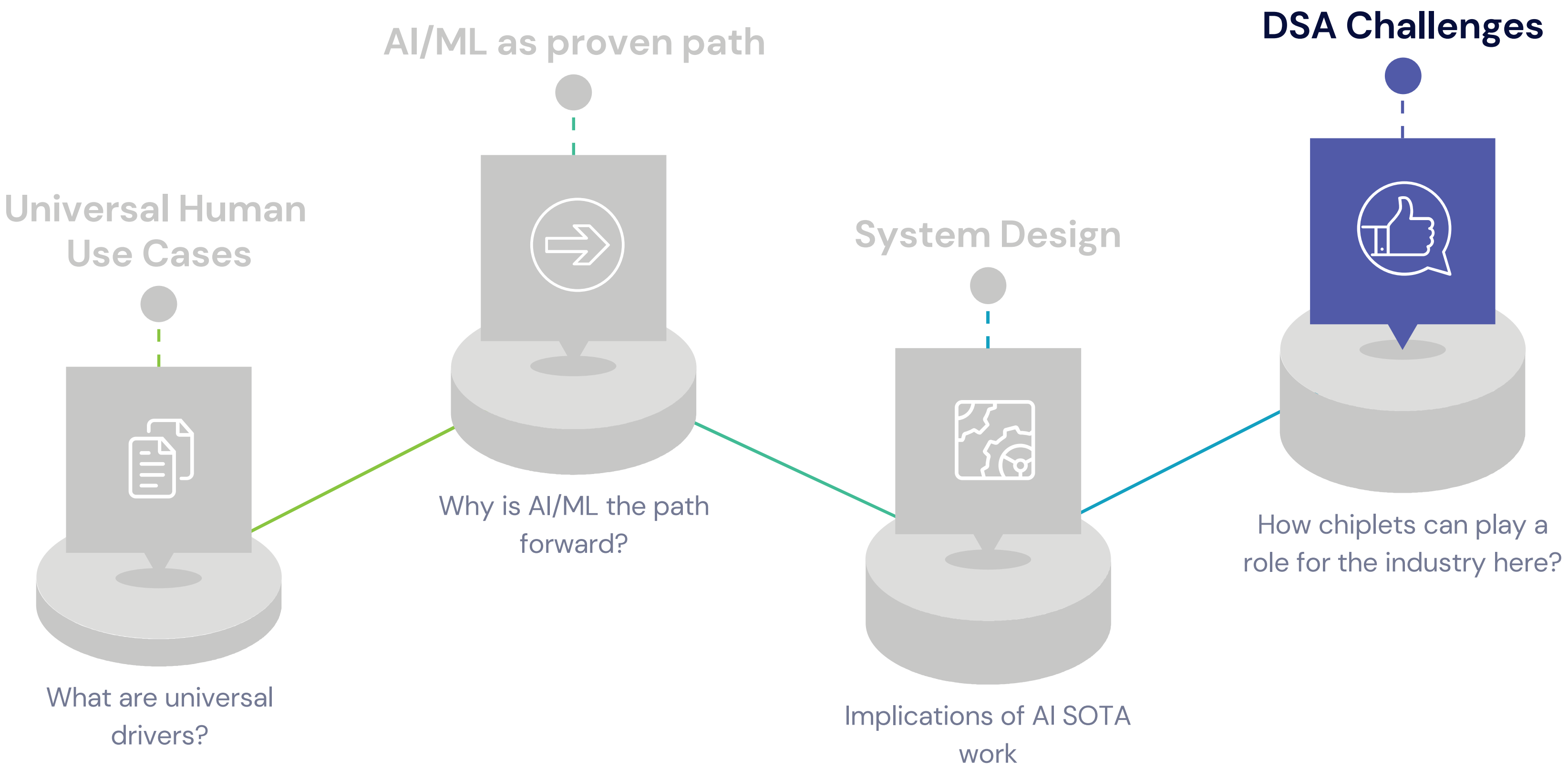


SIZE

COMPUTE

SOURCE: Meta Keynote at OCP Global Summit Oct 2022

Arc of the talk



Universal Human Use Cases

What are universal drivers?

AI/ML as proven path

Why is AI/ML the path forward?

System Design

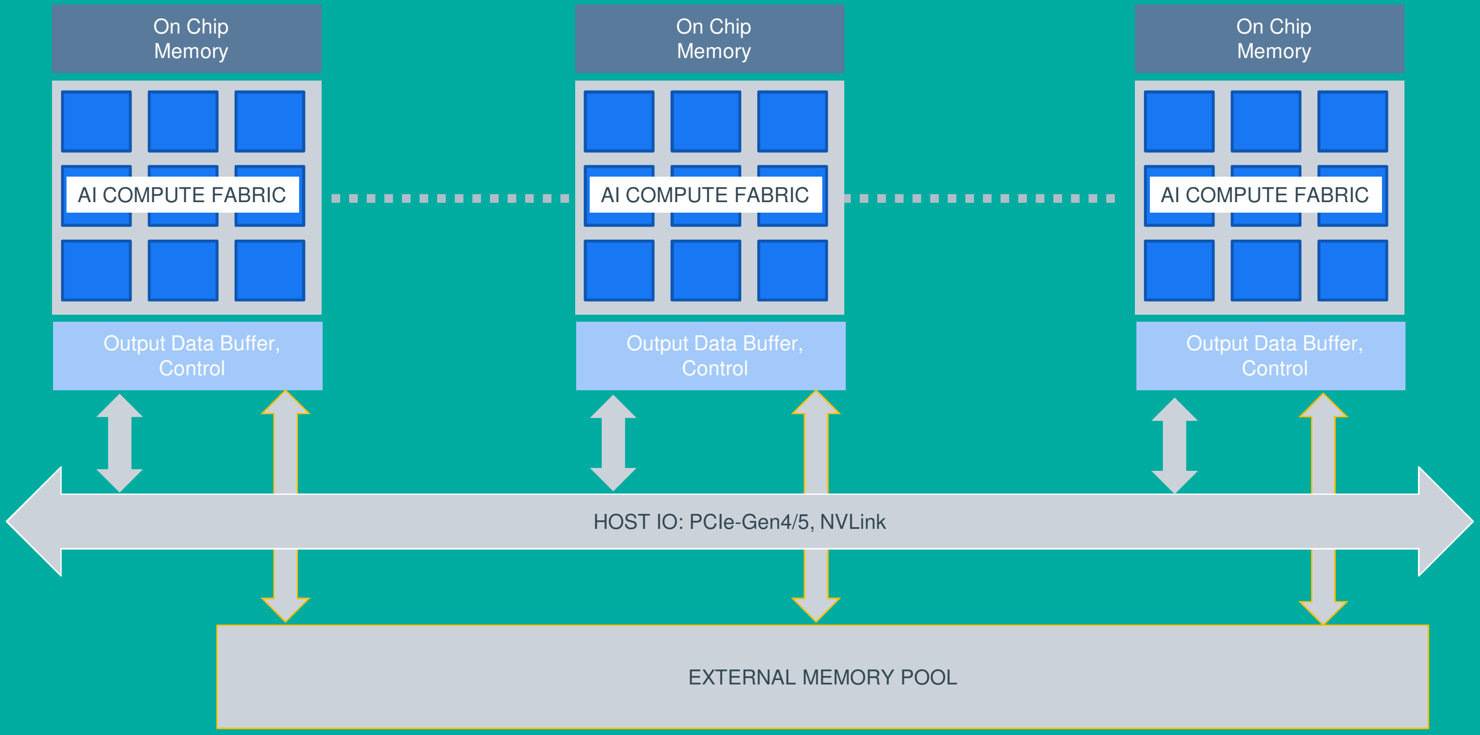
Implications of AI SOTA work

DSA Challenges

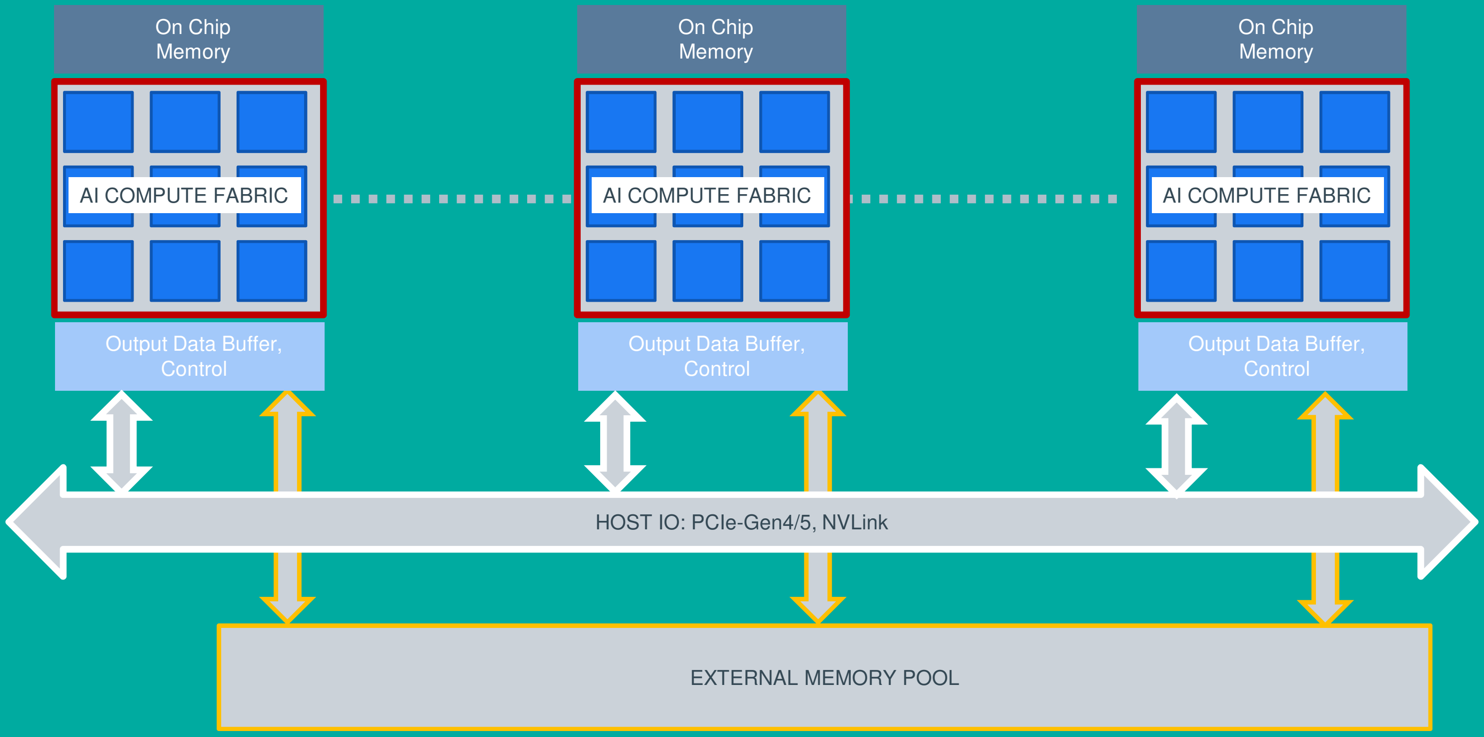
How chiplets can play a role for the industry here?

Key challenges for DSAs to address

Training based on DSA

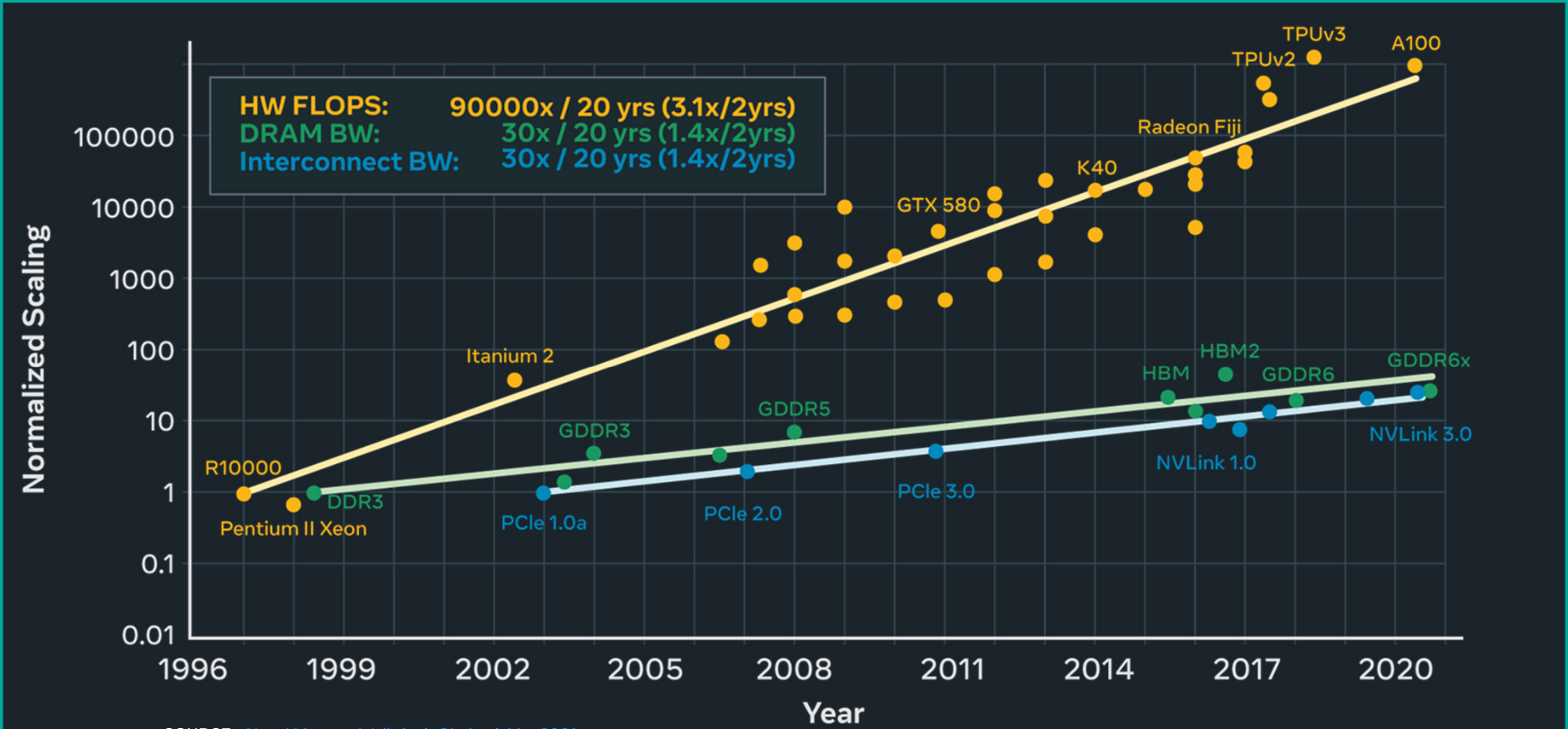


Training based on DSA



Memory and Network Lagging Compute

SCALING OF PEAK HARDWARE FLOPS, AND MEMORY/INTERCONNECT BANDWIDTH



SOURCE: AI and Memory Wall, Amir Gholami, Mar 2021

Challenges for AI System to address

DSA Performance

Accelerator-Memory gap

Model Flexibility

HW/SW co-design

Networking BW

Switching cross sectional BW

Holy Grail it is not…

Apologies to Monty Python!

Meta