Chiplet use cases in AI and ML

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

Jan 24th, 2023

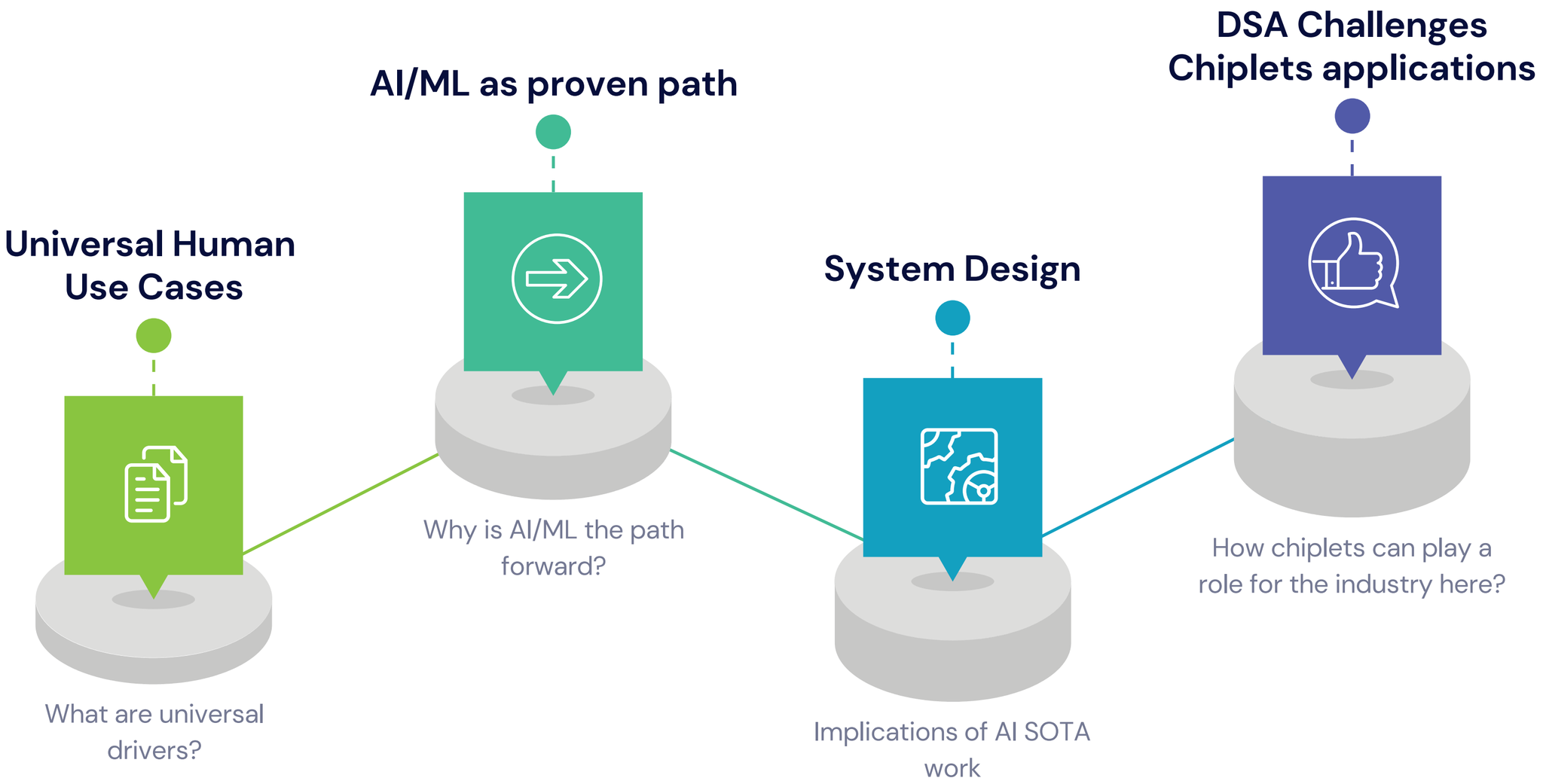
Dharmesh Jani (“DJ”)
Infrastructure Partnerships/Ecosystems Lead @ Meta

CHIPLET SUMMIT

OPEN Compute Project®

Meta

Arc of the talk



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

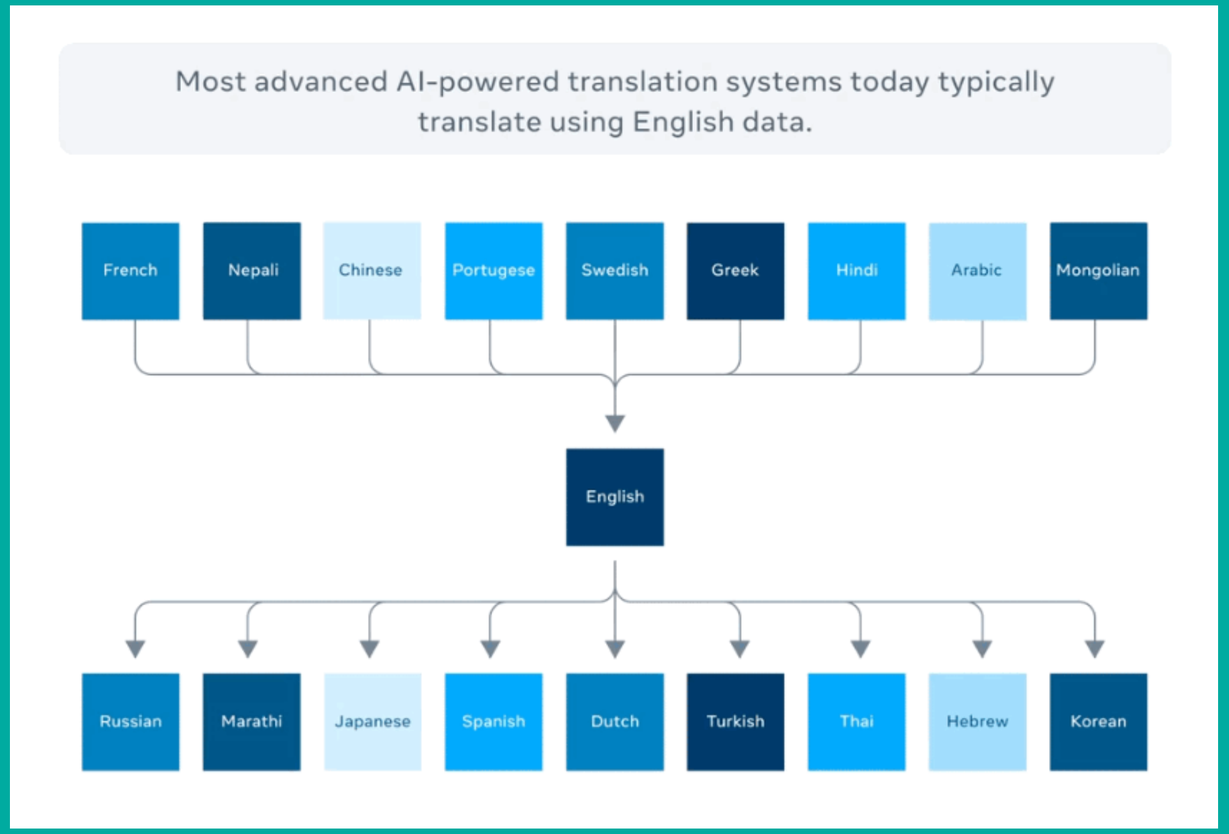
Universal use cases that are drive technology

Recognition

Build identification models by machines of real world

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

Requires training!



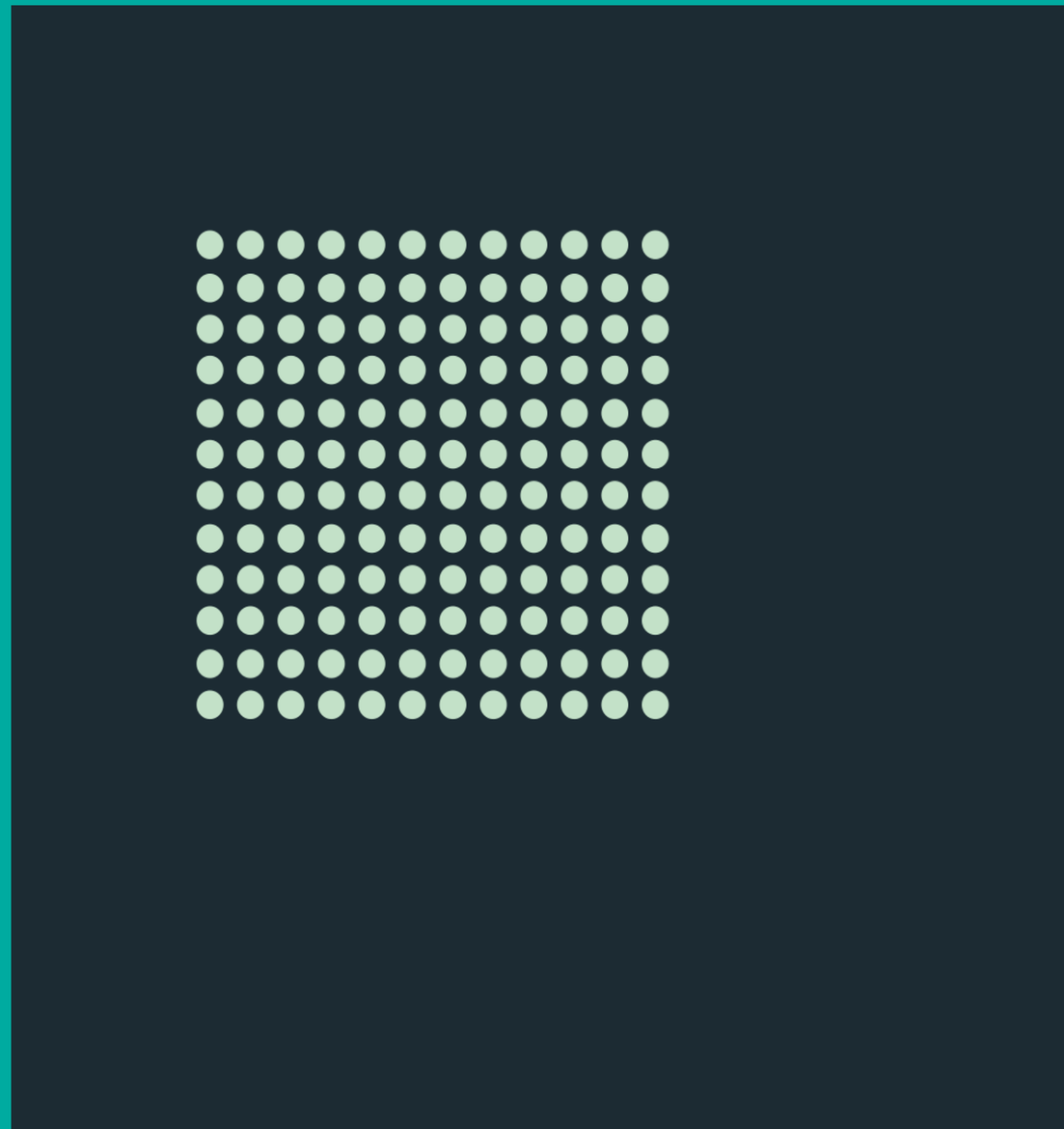
Universal use cases that are drive technology

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!



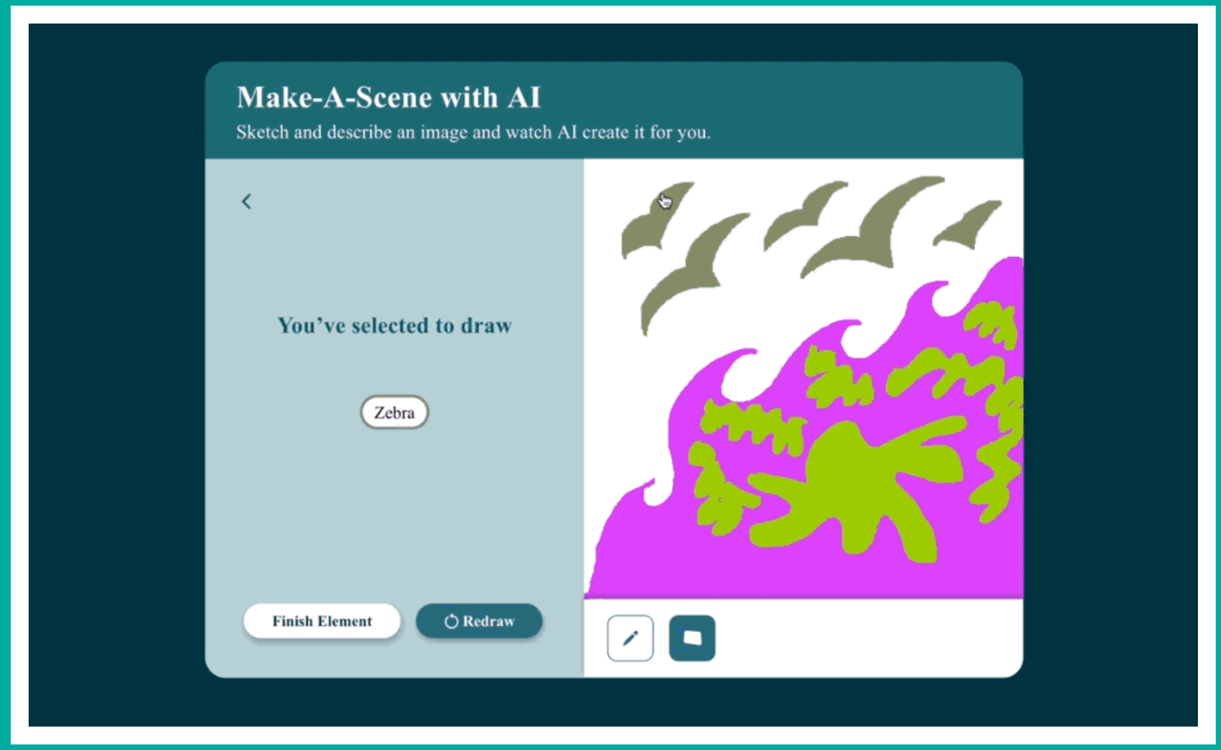
Universal use cases that are drive technology

Synthesis

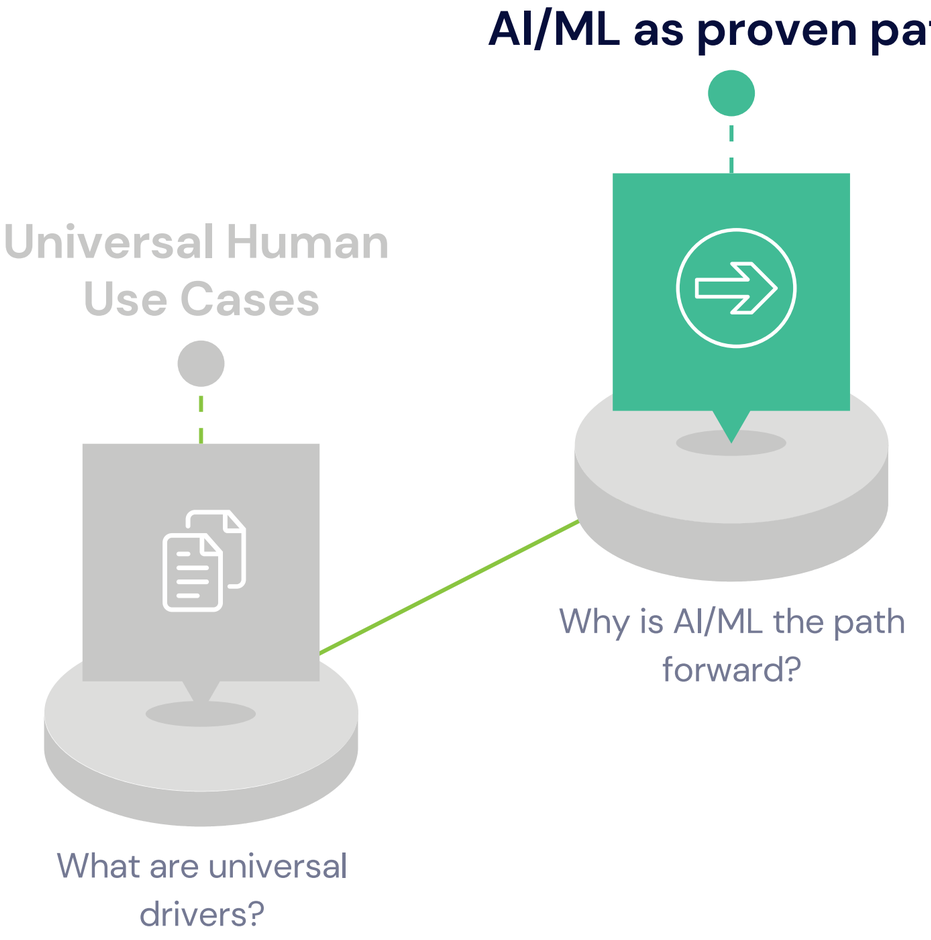
Creating new instance of models where one does not exist

Synthesis is creation by machines of new ideas

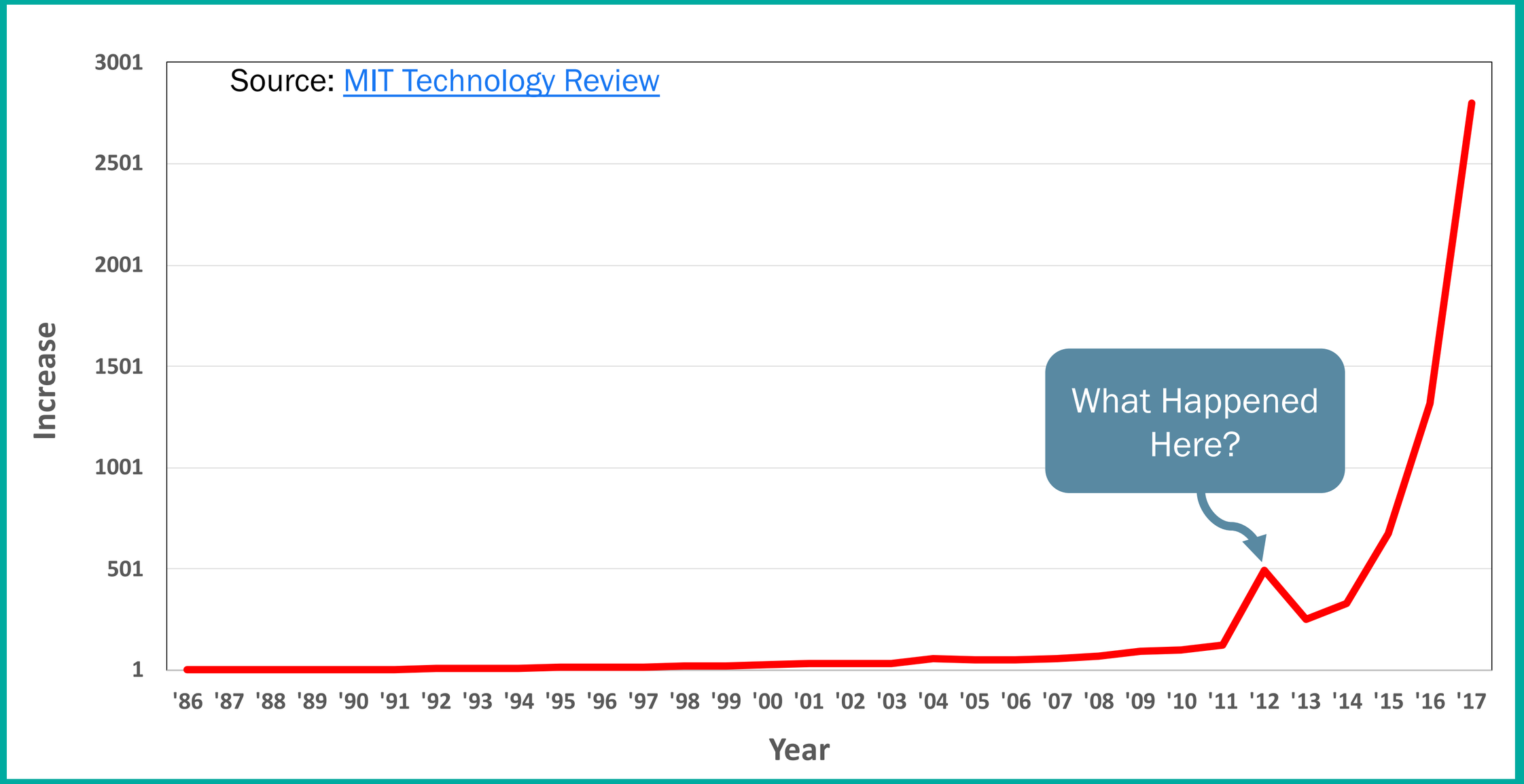
Requires multi-modality, GANs!



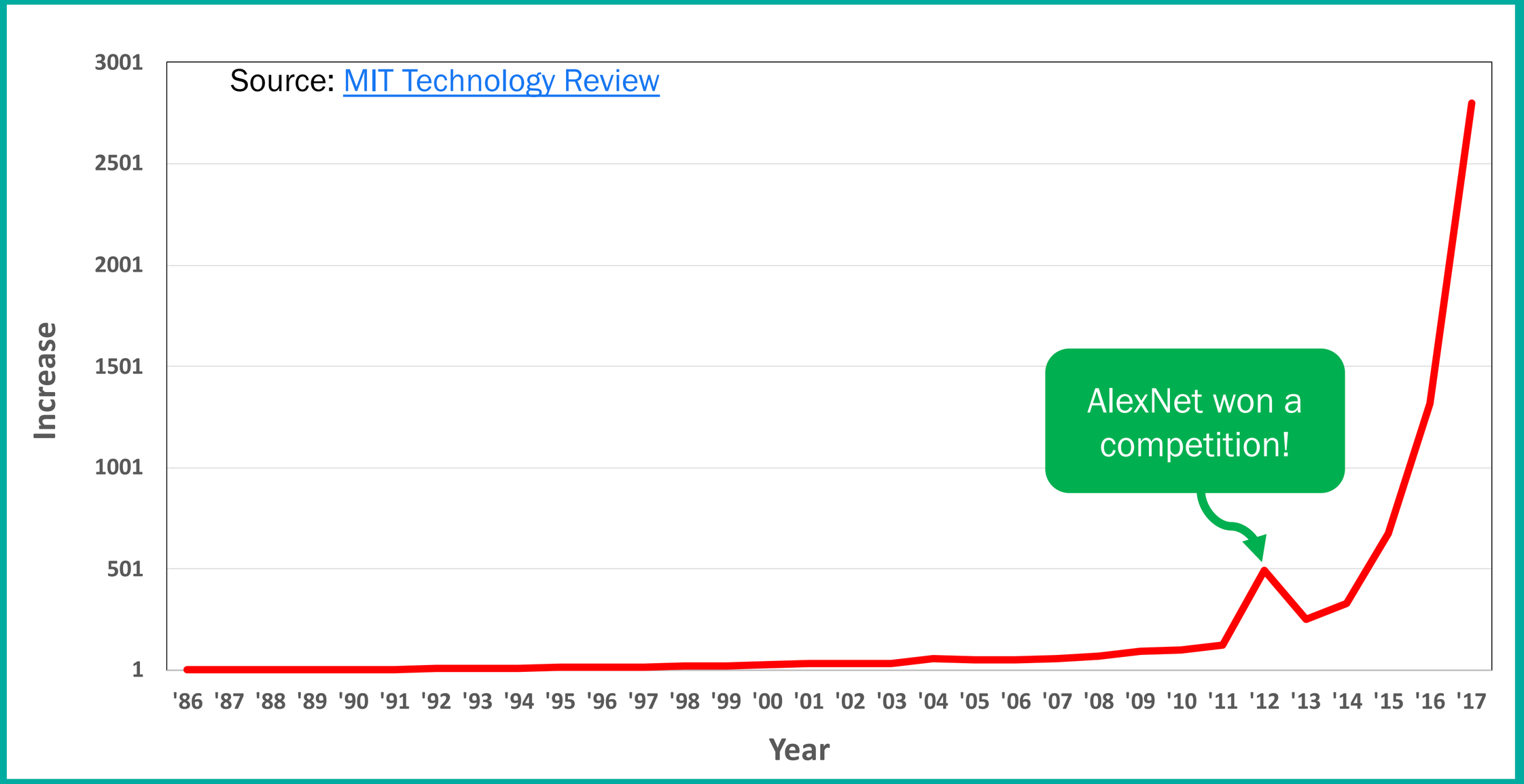
Arch of the talk



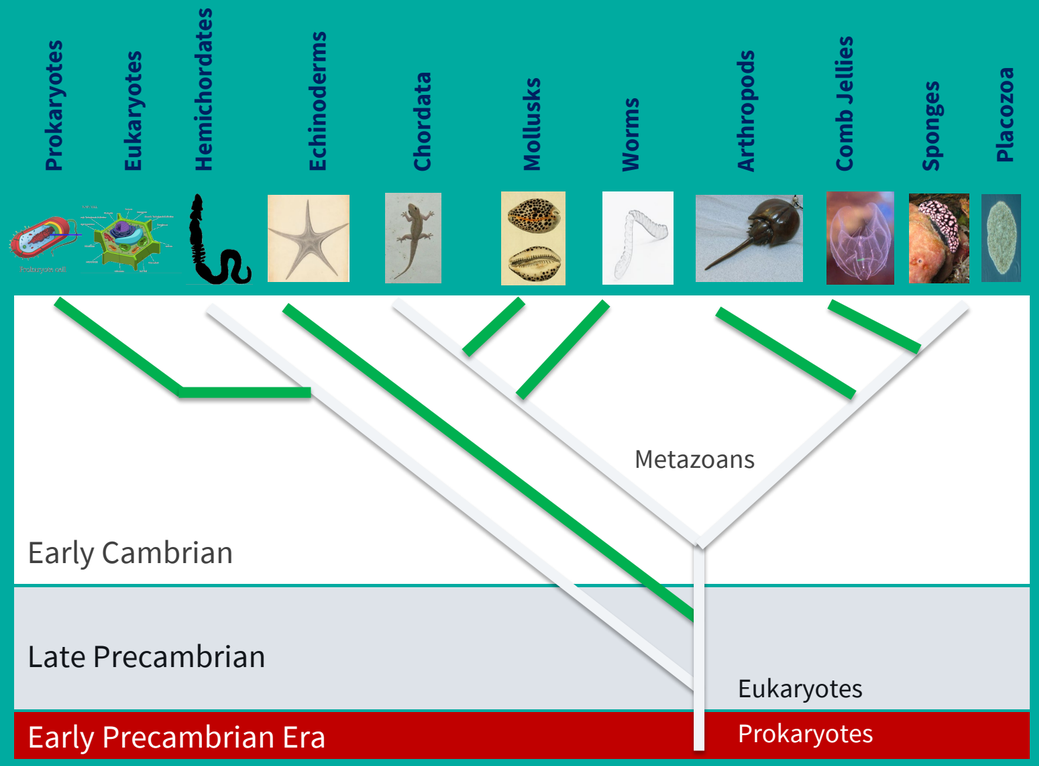
Growth of the term "deep learning" in research



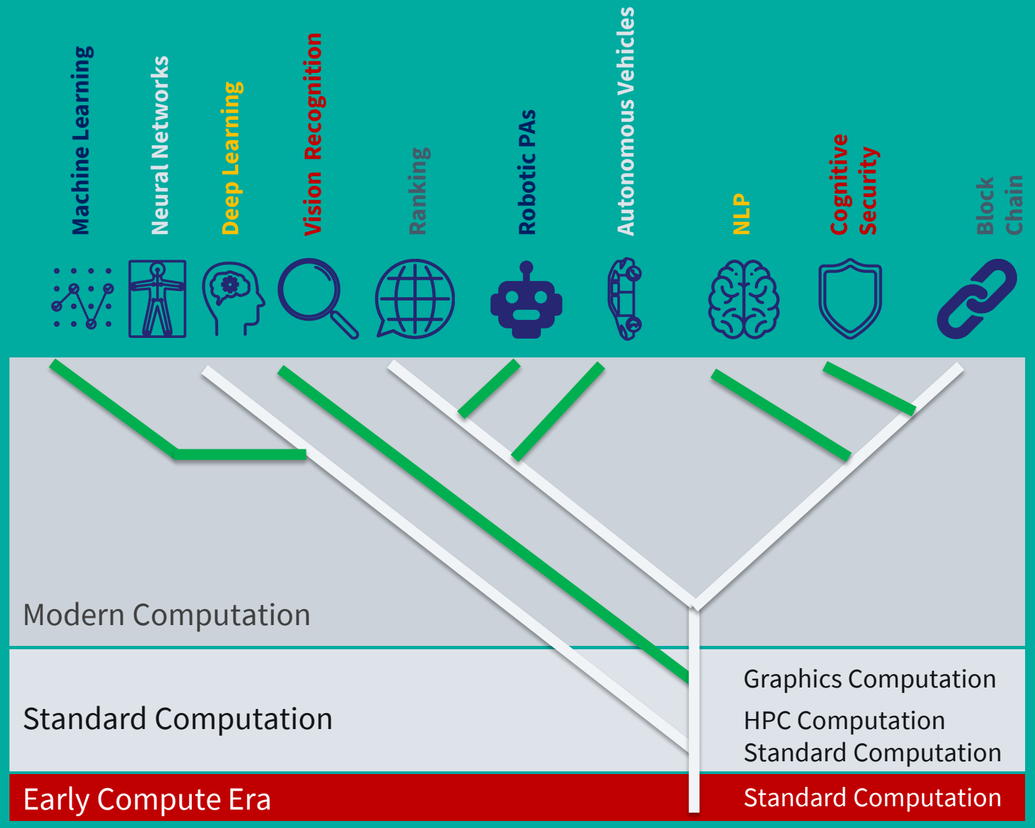
Growth of the term "deep learning" in research



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

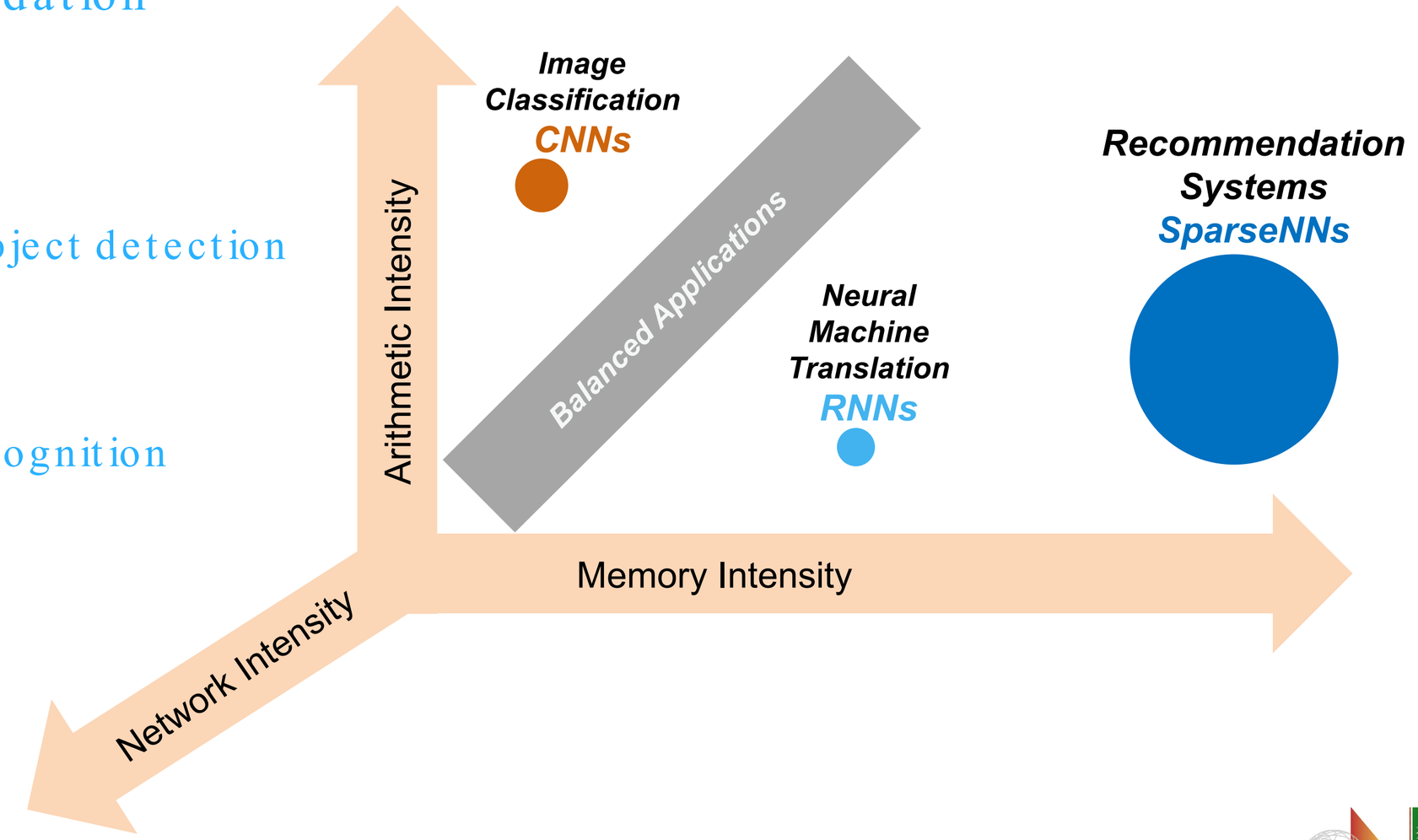


AI and Machine-learning and data-heavy workloads have exploded in 7 years and will diversify as new applications are discovered constantly…

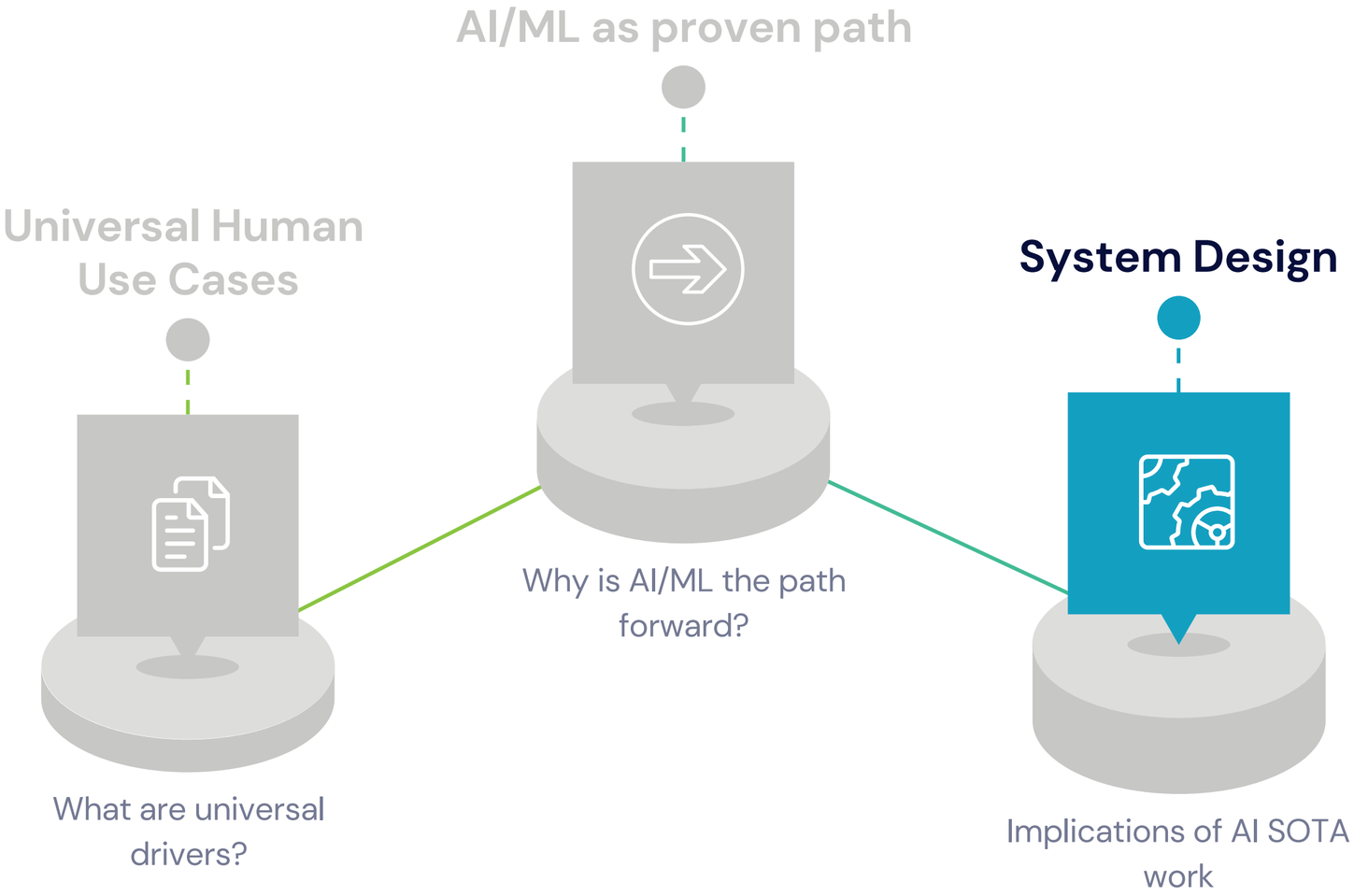
What are the dominant AI serving workloads?

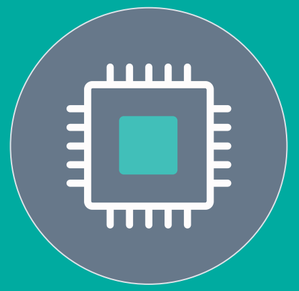
Current and emergent

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



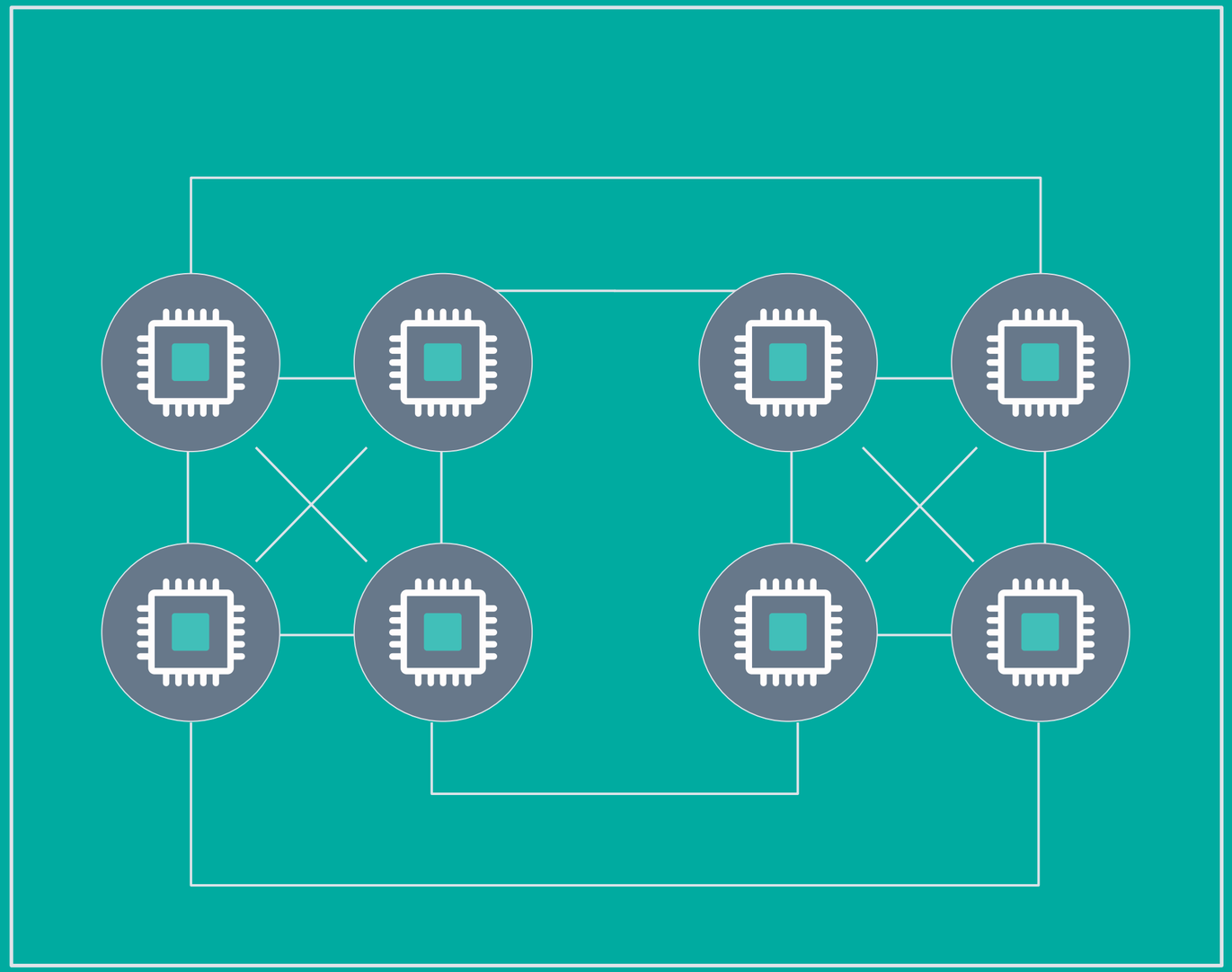
Arc of the talk





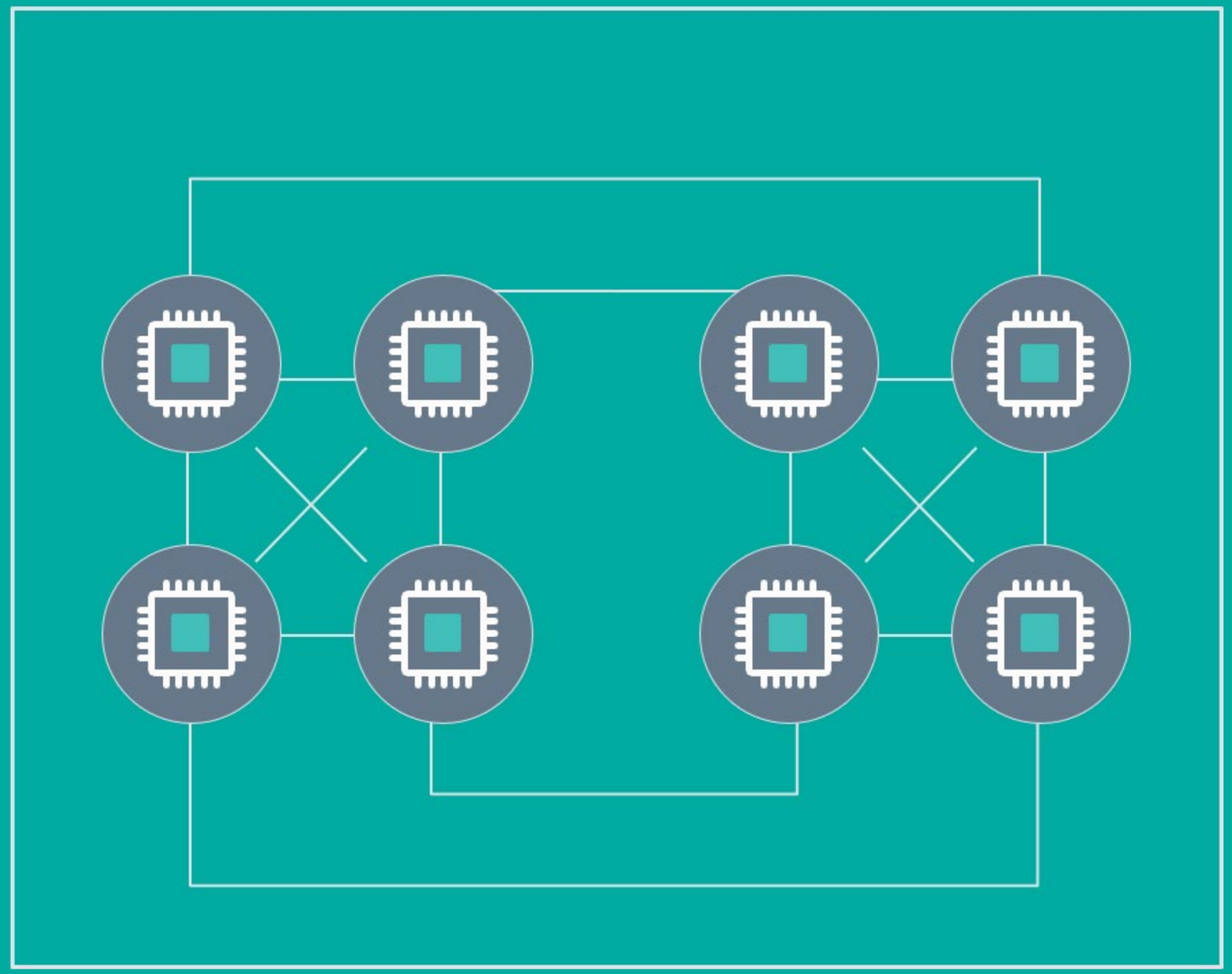
Domain Specific Accelerators

ACCELERATOR WORKLOAD UNIT



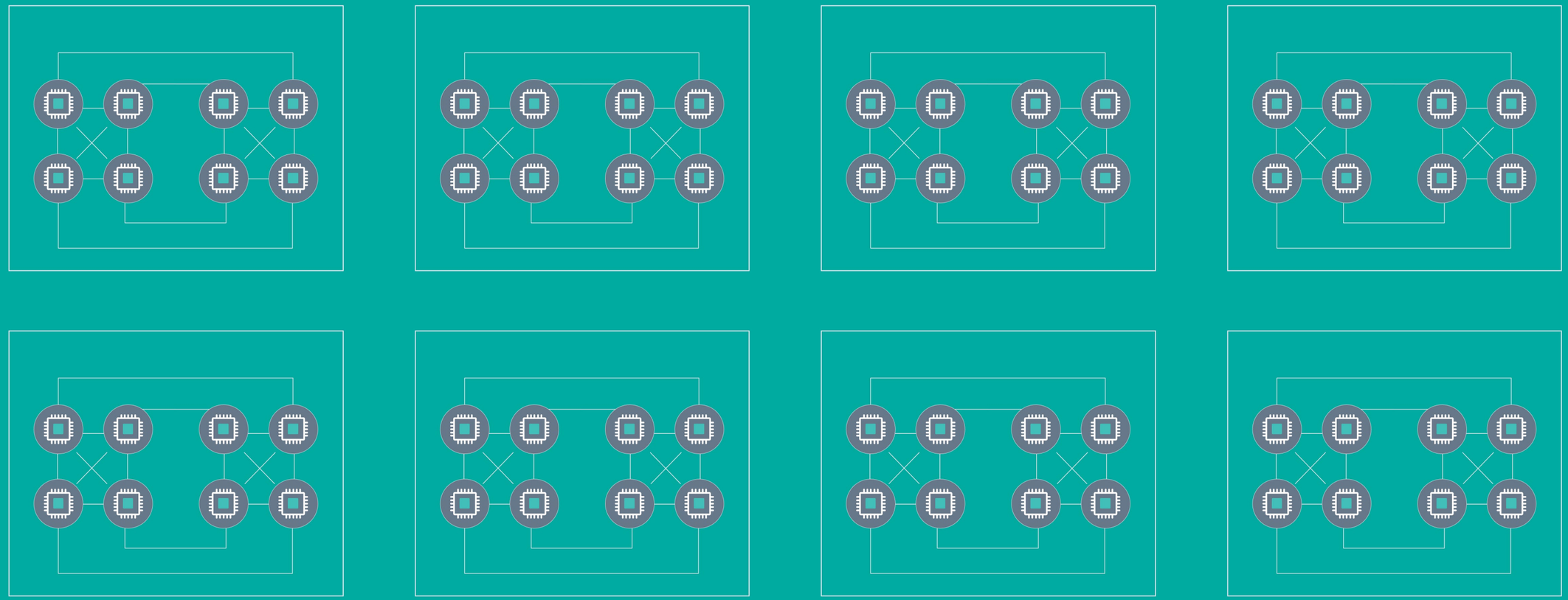
*ignoring the CPU, NICs, SSDs, and everything else...

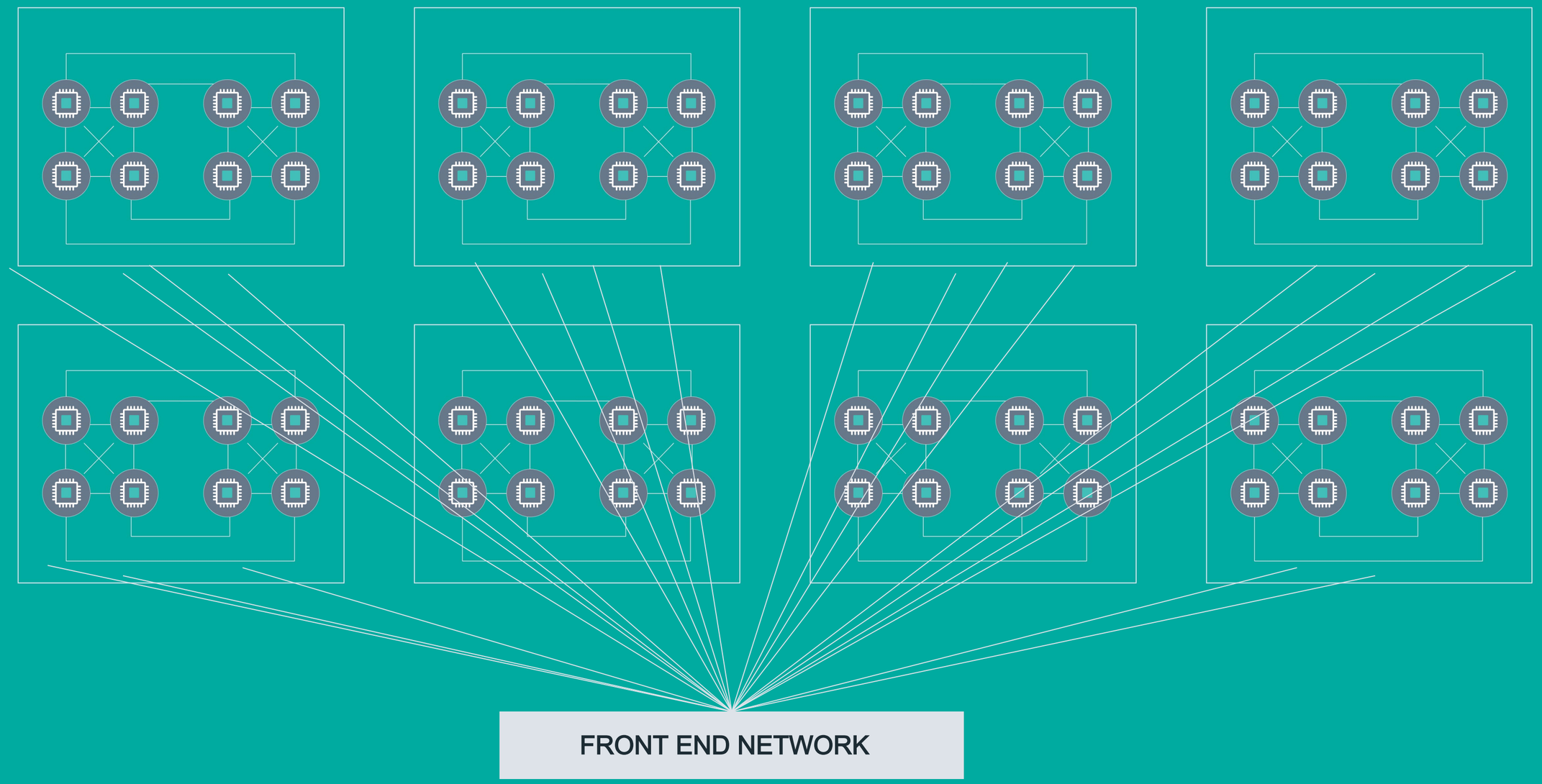
ACCELERATOR WORKLOAD UNIT

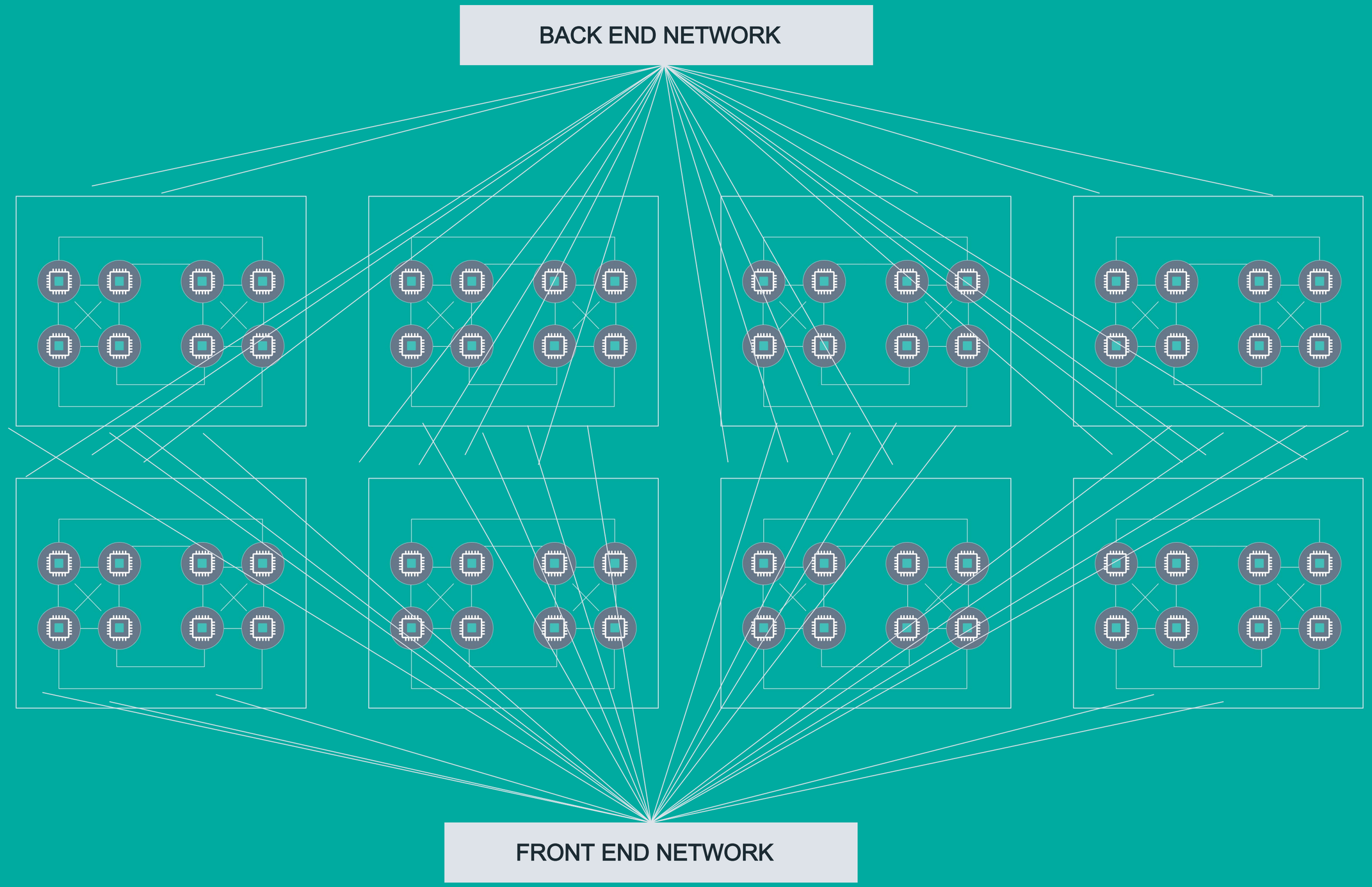


*ignoring the CPU, NICs, SSDs, and everything else...

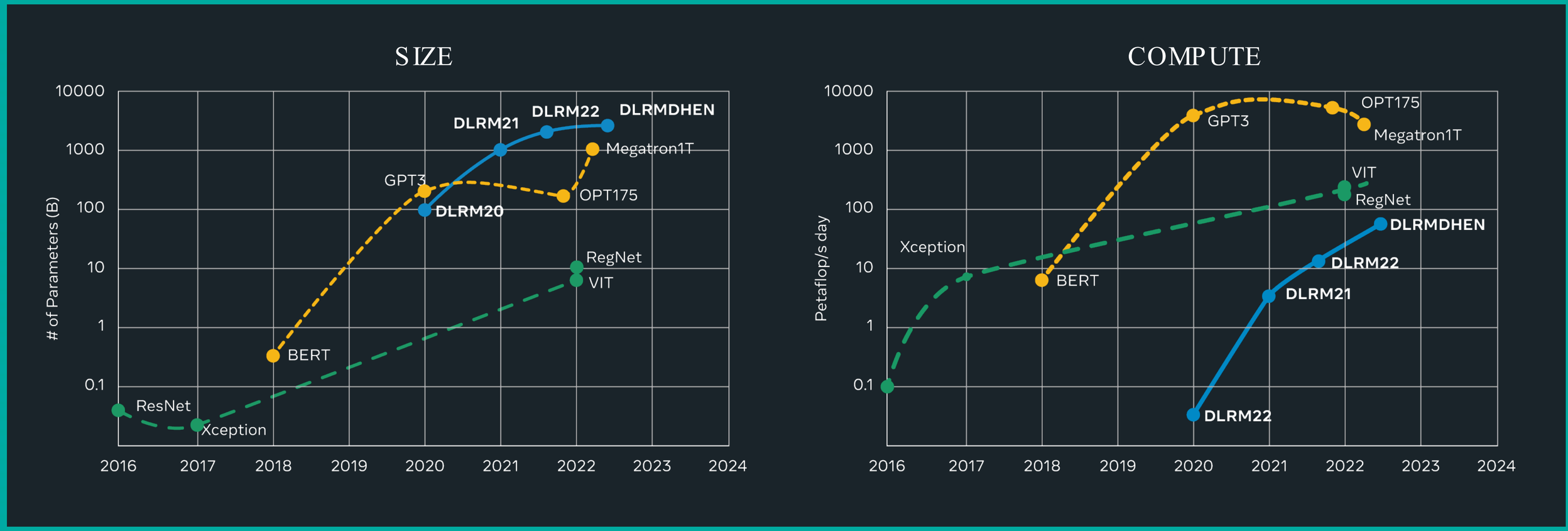
ACCELERATOR WORKLOAD CLUSTER




FRONT END NETWORK

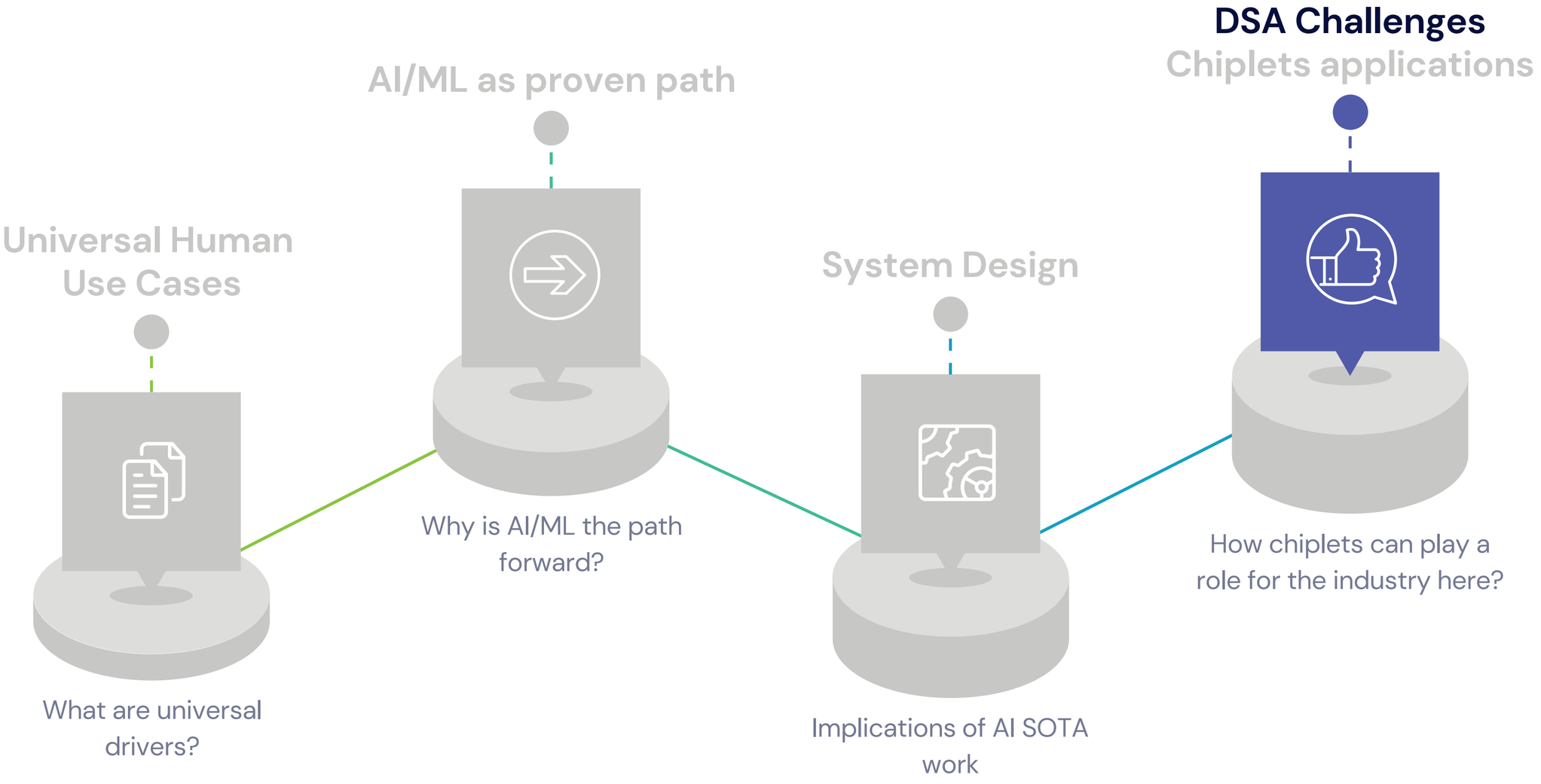

BACK END NETWORK
FRONT END NETWORK

DEEP LEARNING WORKLOADS - CHARACTERISTICS



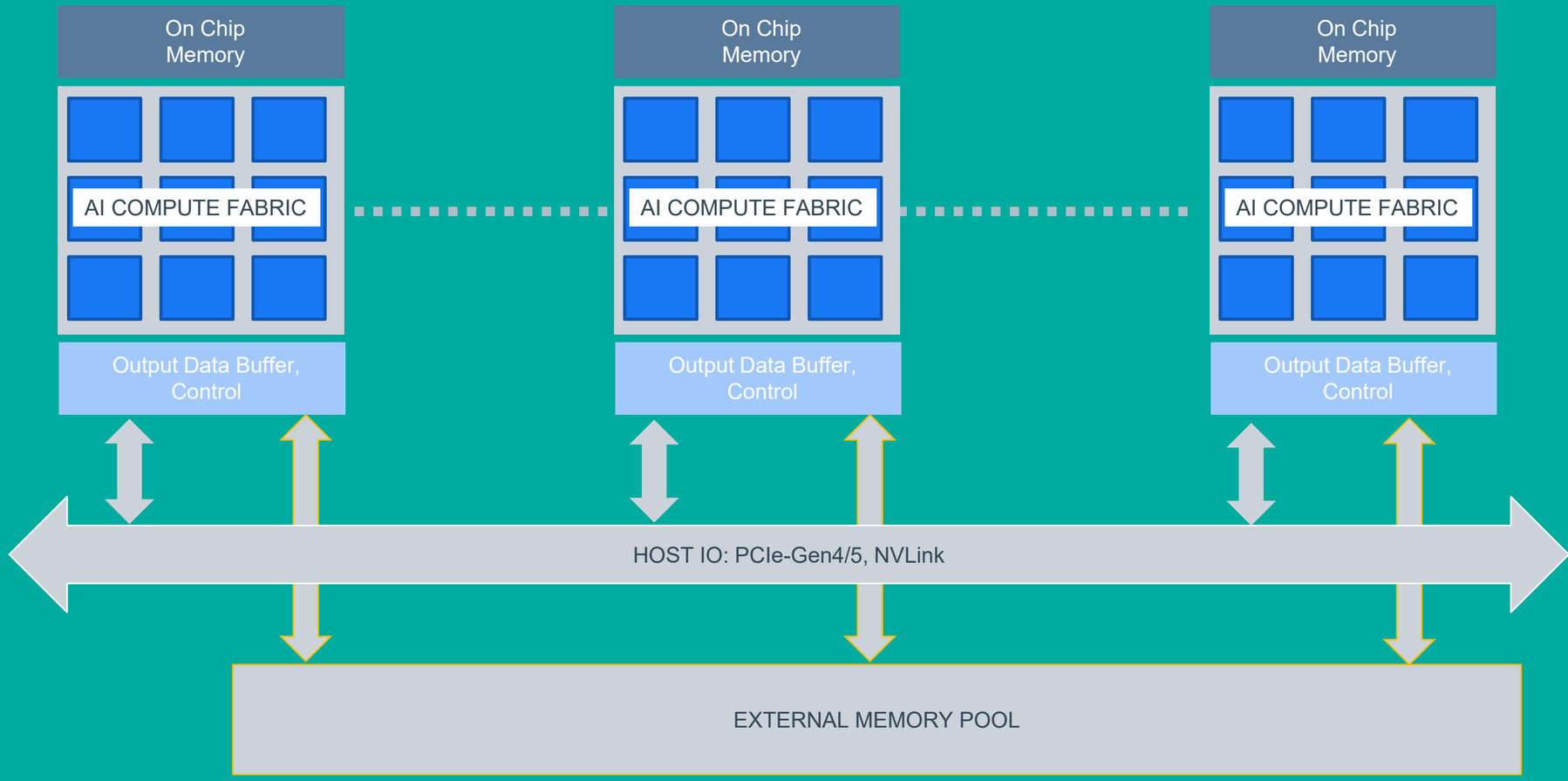
SOURCE: Meta Keynote at OCP Global Summit Oct 2022

Arc of the talk

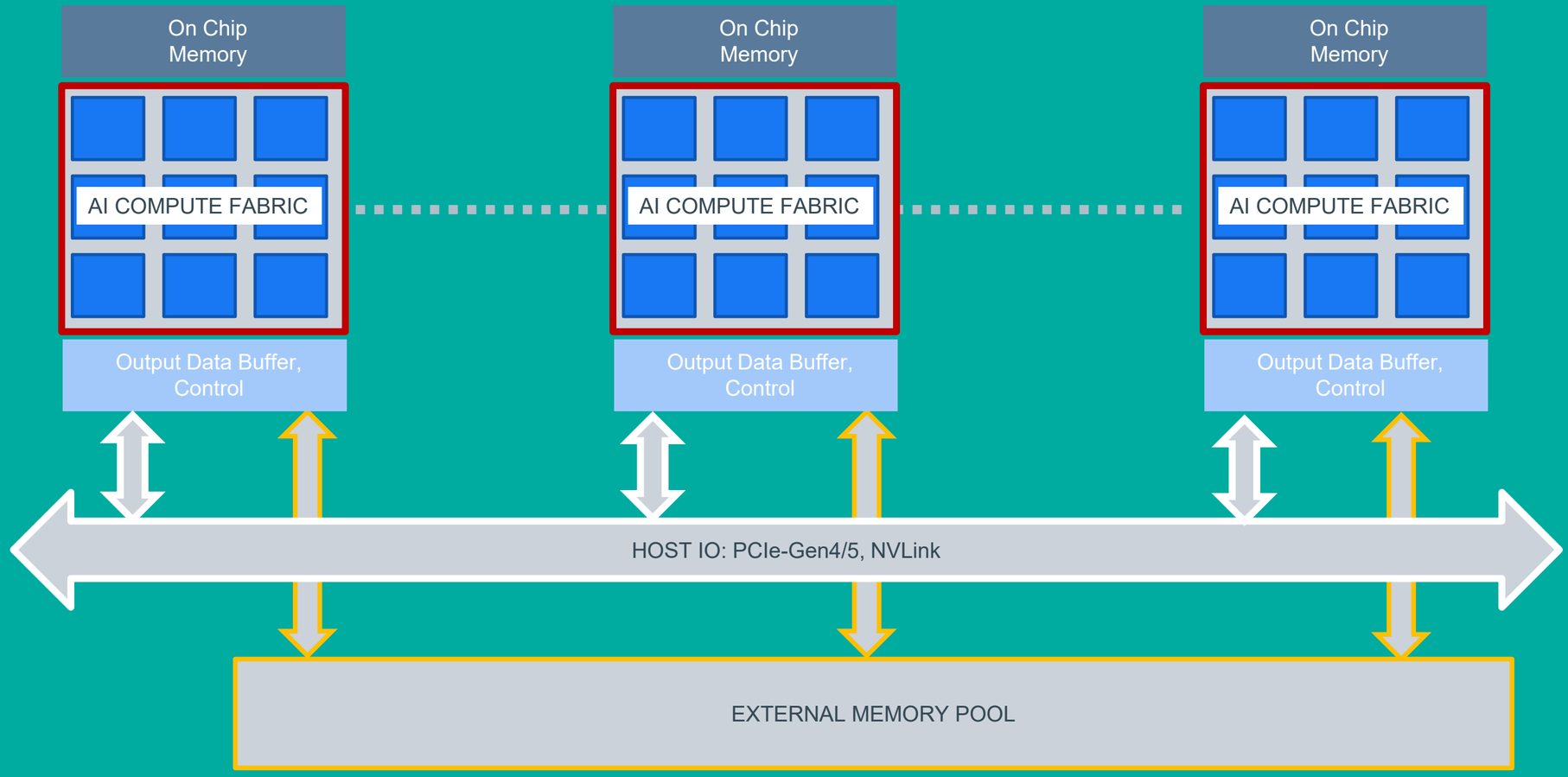


Key challenges for DSAs to address

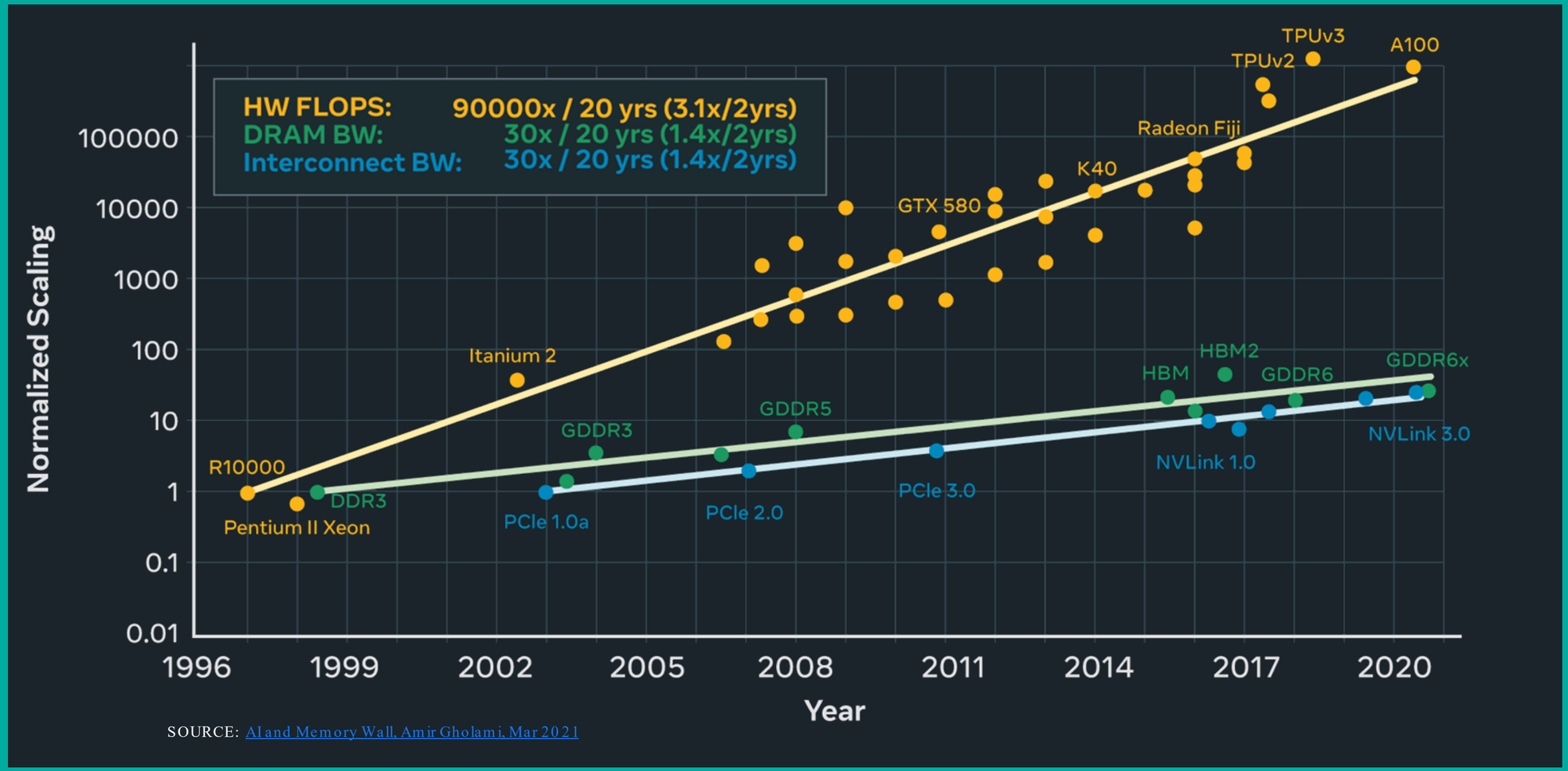
Training based on DSA



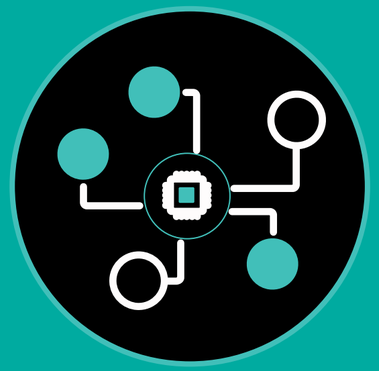
Training based on DSA



Memory and Network Lagging Compute

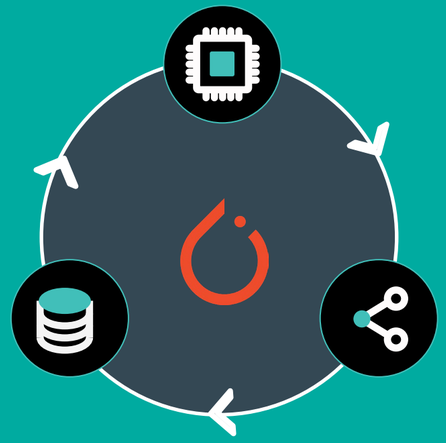


Challenges for AI System to address



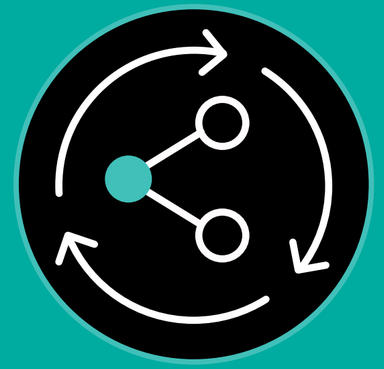
DSA Performance

Accelerator-Memory gap



Model Flexibility

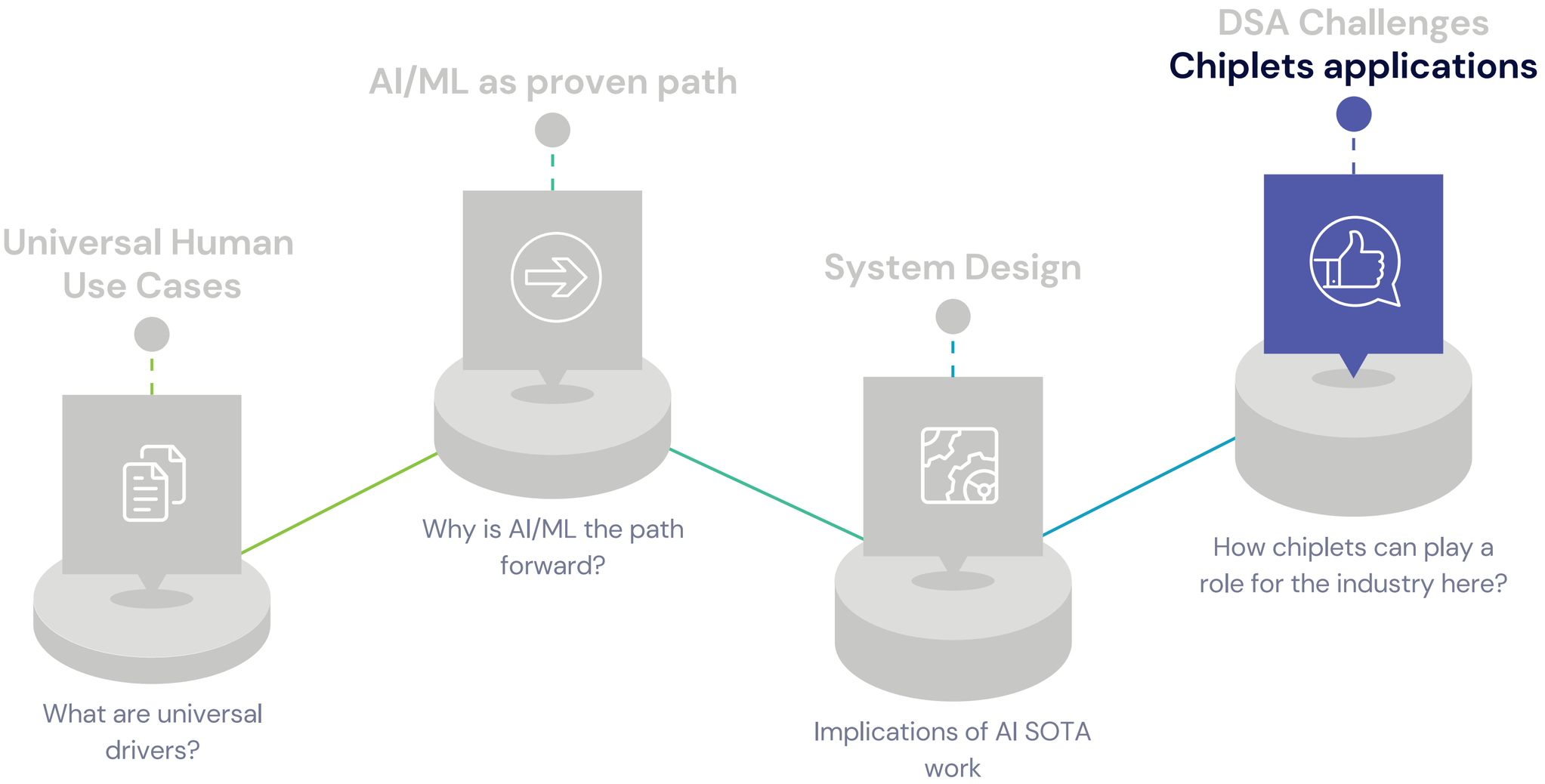
HW/SW co-design



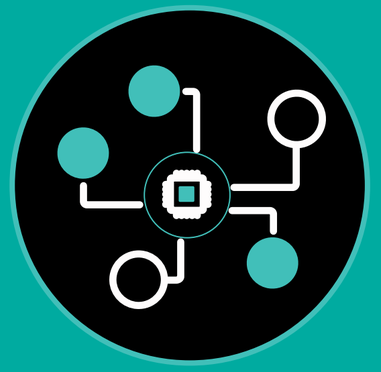
Networking BW

Switching cross sectional BW

Arc of the talk

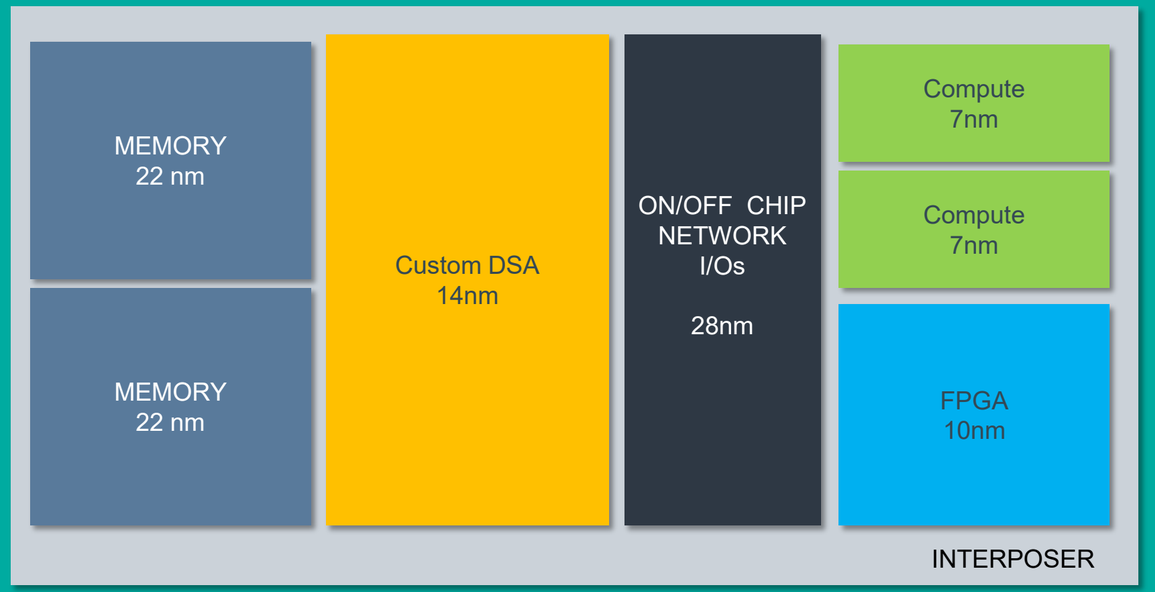


Chiplets for AI Systems: Challenge 1

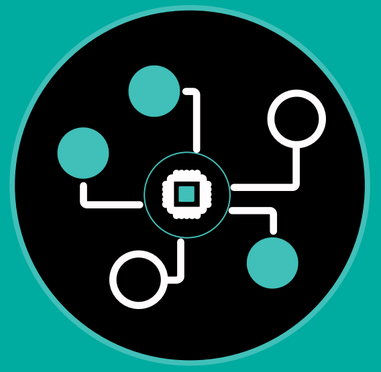


DSA Performance

Accelerator-Memory gap

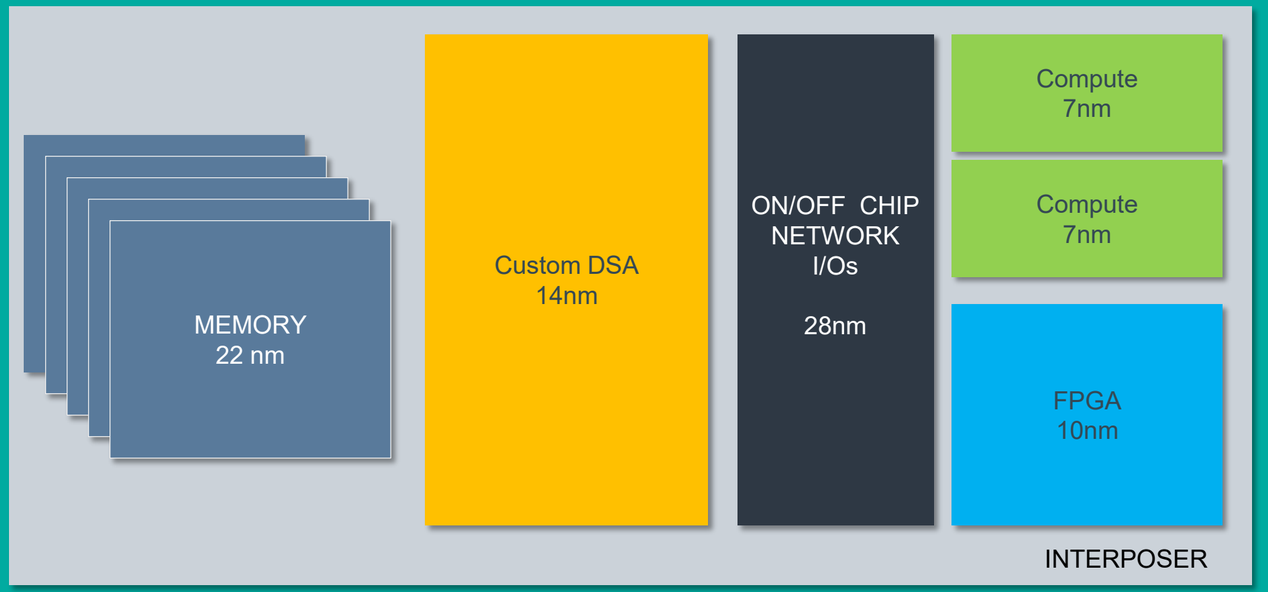


Chiplets for AI Systems: Challenge 1

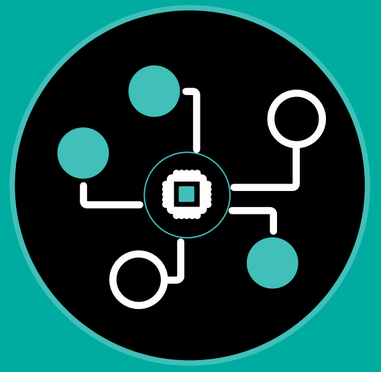


DSA Performance

Accelerator-Memory gap

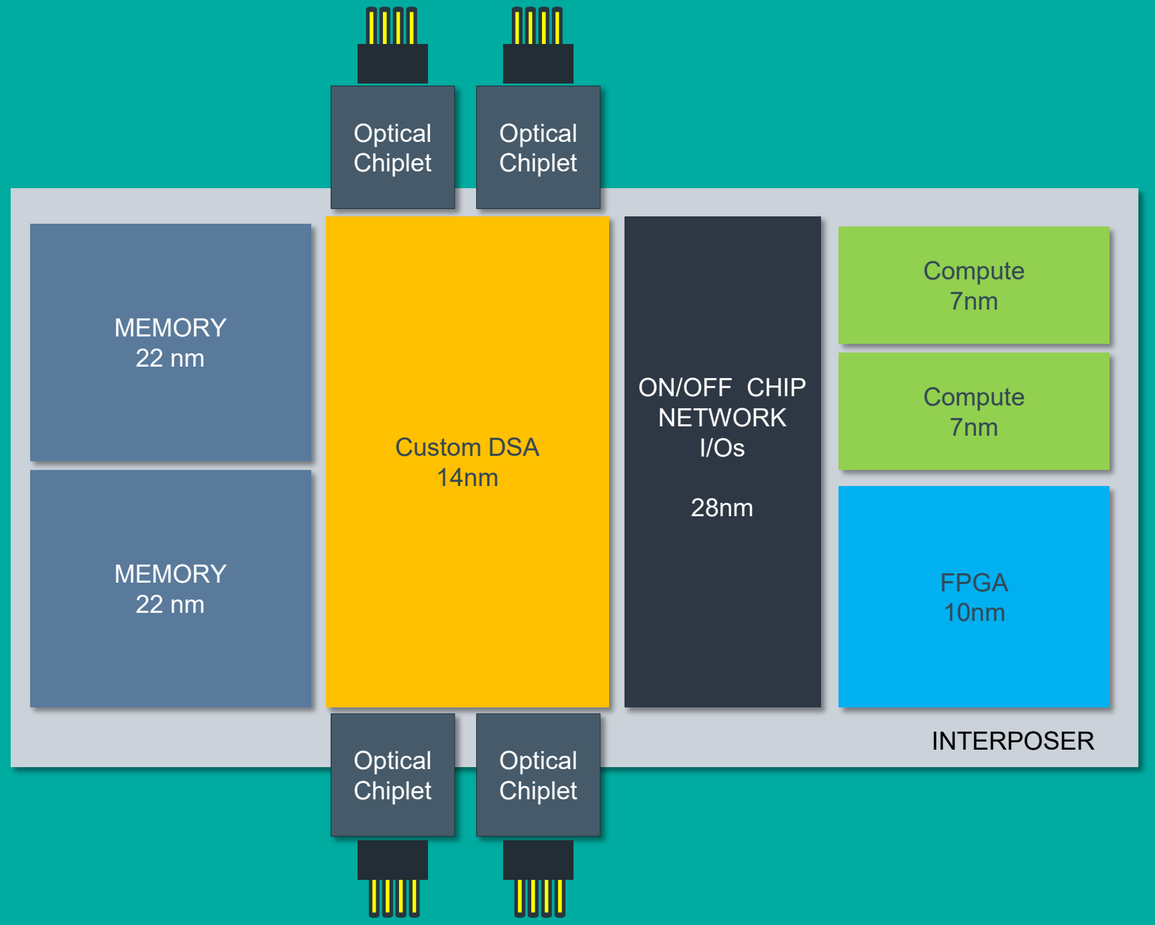


Chiplets for AI Systems: Challenge 1

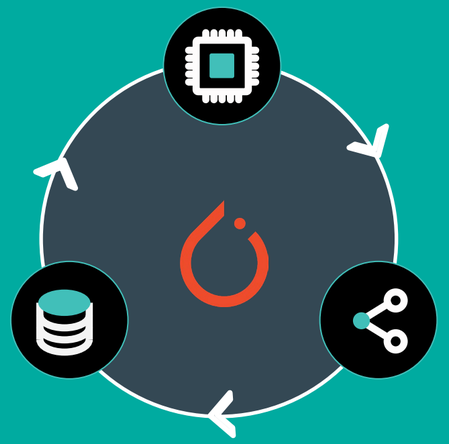


DSA Performance

Accelerator-Memory gap

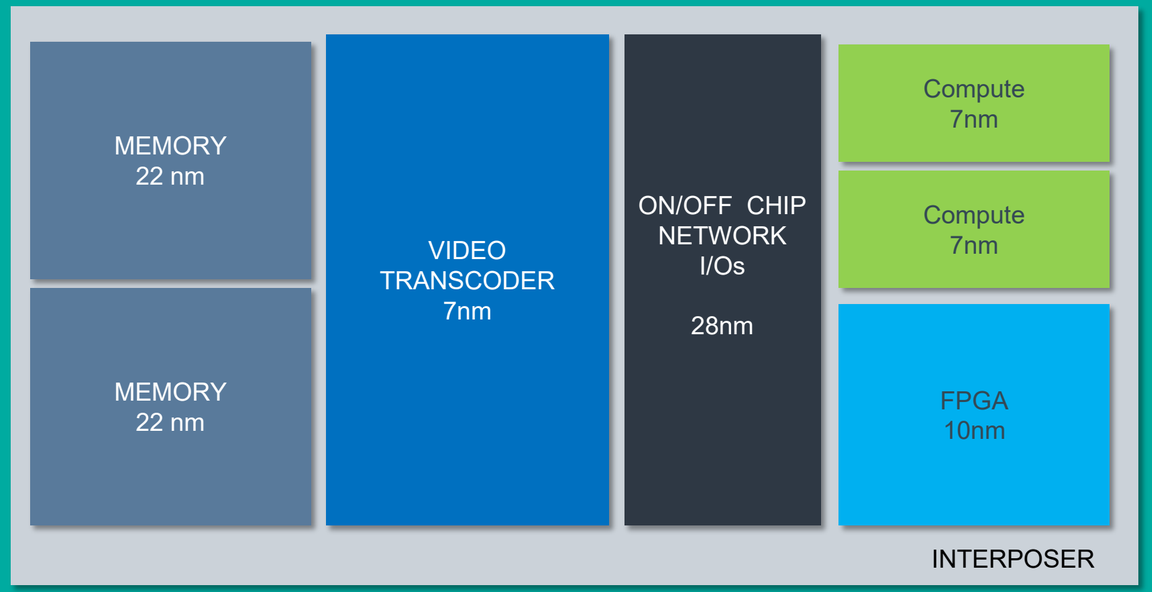


Chiplets for AI Systems: Challenge 2

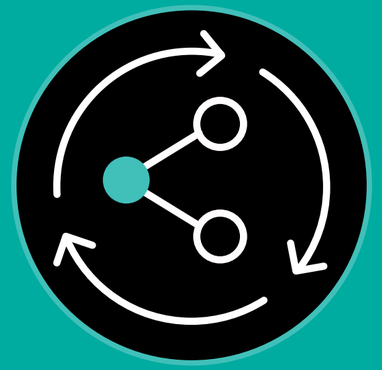


Model Flexibility

HW/SW co-design

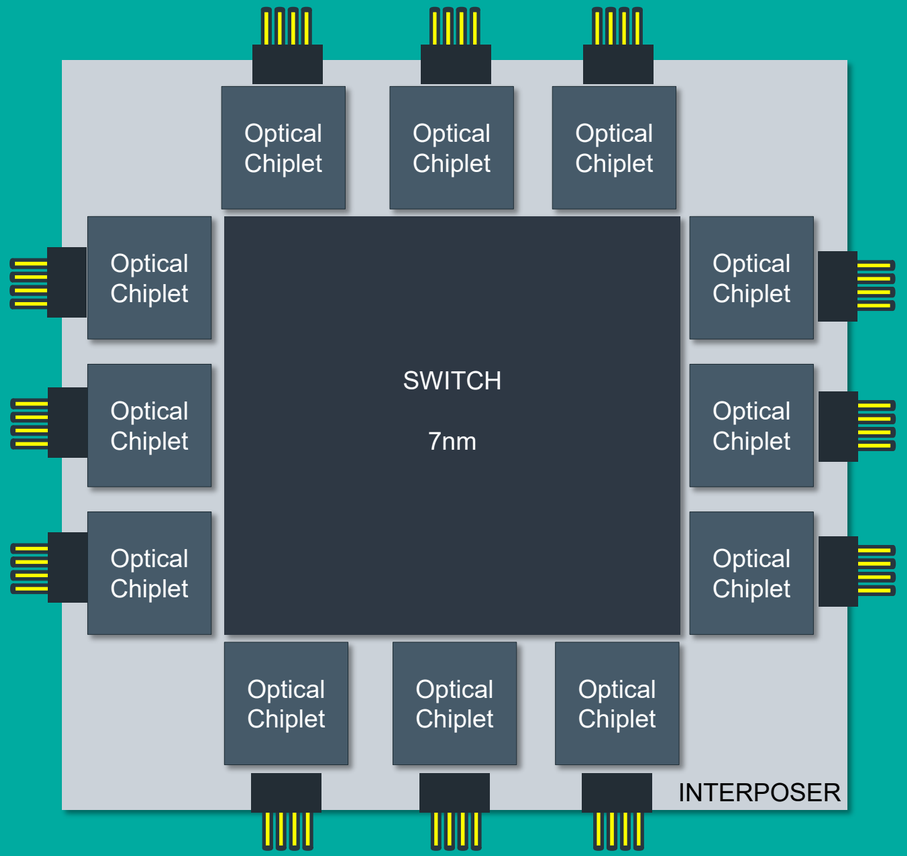


Chiplets for AI Systems: Challenge 3



Networking BW

Switching cross sectional BW

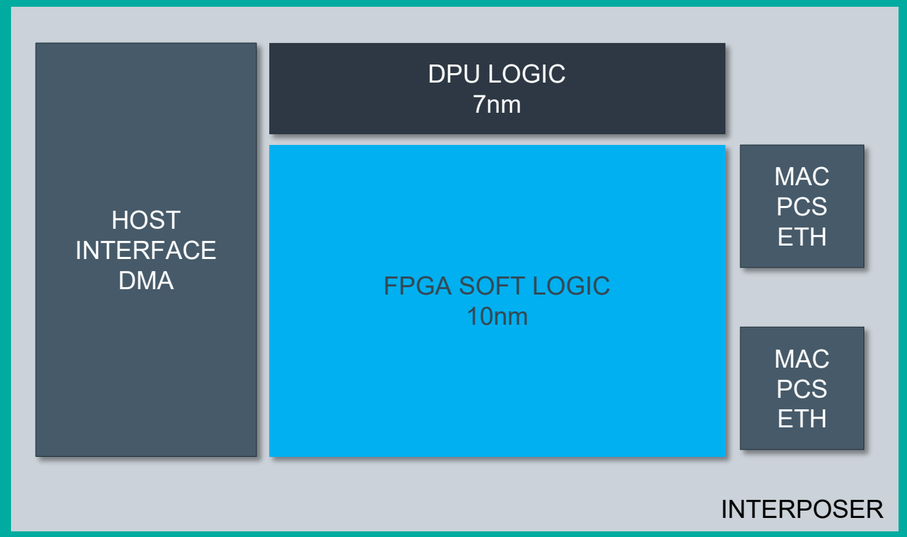


LOW POWER HIGH RADIX
OPTO ELECTRICAL SWITCH

Chiplets for AI Systems: Challenge 3

Networking BW

Switching cross sectional BW



SMART NIC
TRAFFIC OFF-LOAD ACCELERATOR

Holy Grail it is not…

Apologies to Monty Python!


Meta