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
What are universal drivers?

Why is AI/ML the path forward?

Implications of AI SOTA work

How chiplets can play a role for the industry here?
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!
What are universal drivers?

Universal Human Use Cases

AI/ML as proven path

Why is AI/ML the path forward?
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!
Bio-Diversity Exploded from single cells into multi-cell organisms during the Cambrian explosion; all major phyla 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
What are universal drivers?

Universal Human Use Cases

AI/ML as proven path

Why is AI/ML the path forward?

System Design

Implications of AI SOTA work

Arc of the talk
Domain Specific Accelerators
ACCELERATOR WORKLOAD UNIT

*ignoring the CPU, NICs, SSDs, and everything else...
*ignoring the CPU, NICs, SSDs, and everything else...
DEEP LEARNING WORKLOADS - CHARACTERISTICS

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

Chiplets applications

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

- HW FLOPS: 90000x / 20 yrs (3.1x/2yrs)
- DRAM BW: 30x / 20 yrs (1.4x/2yrs)
- Interconnect BW: 30x / 20 yrs (1.4x/2yrs)

Source: AI and Memory Wall, Amir Ghodsi, 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
Arc of the talk

Universal Human Use Cases

AI/ML as proven path

Why is AI/ML the path forward?

System Design

Implications of AI SOTA work

DSA Challenges

Chiplets applications

How chiplets can play a role for the industry here?
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

Diagram showing various components and technologies such as:
- 7nm and 28nm Memory
- Compute 7nm
- ON/OFF Chip Network I/Os 28nm
- FPGA 10nm
- Video Transcoder 7nm
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!