Enable Polymorphic AI Architecture via Composable Chiplet Technologies

Weifeng Zhang
Chief Architect & VP of Software
Lightelligence
Co-Chair, HiPChips
Thanks all speakers for your great presentation on Chiplet, Architecture, Interconnect, Standards, and Ecosystem!

What opportunities do these advancements provide to empower AI and HPC applications?
AI Application Characteristics

- Massive amount of data
- Large models

- Compositional computation\[^1\]
- Mixture of experts

- High compute demands
- Huge memory footprints
- Large data movements

Spatial evolution
AI sub-domain diversities

Temporal evolution
Difference between generations

Neuro-compositional computing mush simultaneously enable human-level cognition:
the principles of Compositionality and Continuity
Architecture Advancement

General-Purpose Computing
multicore, multithreading, many-core

Heterogeneous Computing
DSA architecture in the post-Moore era

compute granularity

More flexible

More efficient

2nd High Performance Chiplet and Interconnect Architectures (HiPChips), HPCA-2023, Montreal
Compositional computing
Mixture of expert

- Spatial evolution
- Temporal evolution
- High compute demands
- Huge memory footprints
- Large data movements
- Algorithms outpacing hardware

Diversified domain-specific granularities

Domain Specific Architecture with coarse instruction granularities

2nd High Performance Chiplet and Interconnect Architectures (HiPChip), HPCA-2023, Montreal
Polymorphic Architecture

Future-Oriented Computing

Focus on composability and adaptability
- Spatial and temporal scalability
- Composability with diversified granularities
- Re-configurability and transformability

Goal
Achieve the same level of performance as the purpose-built DSA accelerator with spatial and temporal scalabilities
Critical Components in AI Architecture

- Compute
- Memory

- Interconnect

- Capability: heterogeneity, efficiency
- Scalability: scale up, scale out
- Bandwidth, latency
- Topology, coherency
- Hierarchical, tiered, disaggregated
- Near, far, in-memory
Rationales for Polymorphism

Architectural Composability for neuro-compositional computing

1) DSA based scale-out of computation aligns well with the compositionally structured nature of future AI
2) Steady advancement of composable hardware (e.g. CGRA, HPCM, Pathway)
3) Ultralow interconnect latencies (e.g. via optical links)
4) Coherent interconnect protocols (e.g. CXL, UCIe)
5) Fast reconfiguration of programmable hardware (in the order of tens of cycles)
Hierarchical Composability for Polymorphism

Polymorphism: heterogeneity + composability + transformability

- High-performance chiplet-based heterogeneous PEs
- Hierarchical interconnect for intra-chiplet, inter-chiplet, accelerators, ...
- Composability at hierarchies: from chiplet, compute unit, compute node, ...
- Just-in-time configurable compute funclet, memory, and interconnect
- Compute partitioning and mapping with dark compute funclets
Application-Driven Transformability

AI application execution flow on the polymorphic architecture
Acknowledgement

OCP AI Co-Design Workgroup

&

Alan Cantle, Kevin Cameron, Winston Liu, Brian Hirano, Greg Compton, Danny Moore, Robert Feng, Ritu Gupta, Zhibin (David) Xiao, Igor Muskatblit, Huihuo Zheng, Yang Ki, Keith McKay, Matt Bergeon, Marian Verhelst, Jonathan Zhang, Manoj Wadekar, Michael Choi
Reference

