The 2nd High Performance Chiplet and Interconnect Architectures (HiPChips)

HPCA-2023, Montreal, Canada

Enable Polymorphic AI Architecture via Composable Chiplet Technologies

Weifeng Zhang

Chief Architect & VP of Software
Lightelligence

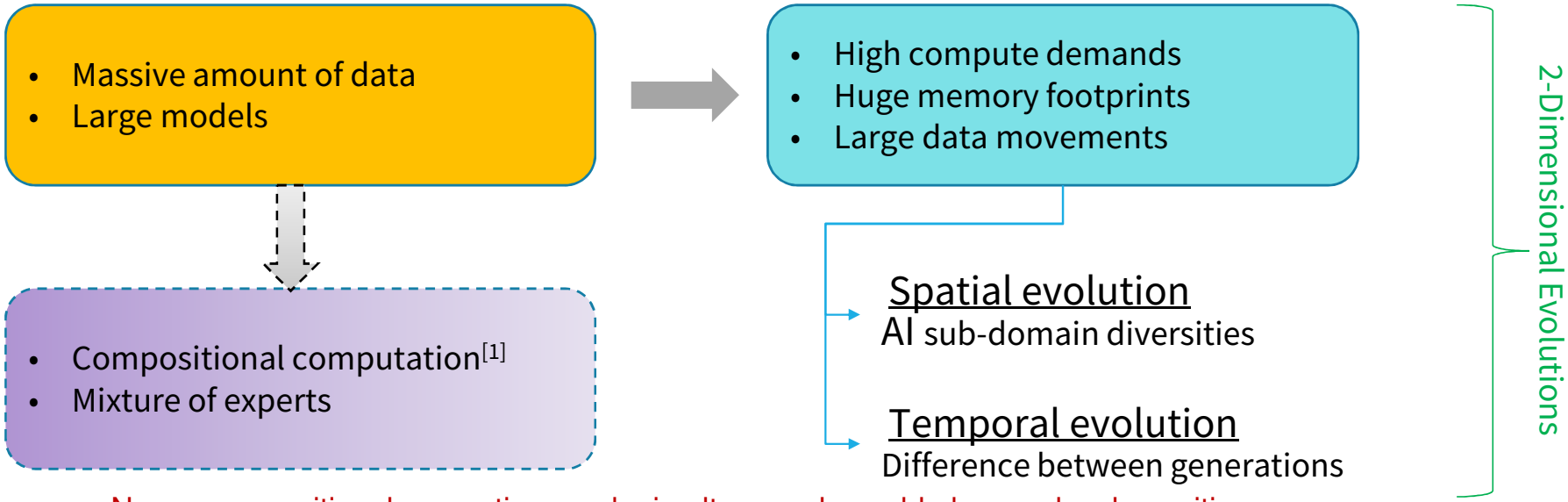
Co-Chair, HiPChips

HiPChips
HPCA-2023

Thanks all speakers for your great presentation
on
Chiplet, Architecture, Interconnect, Standards, and Ecosystem!

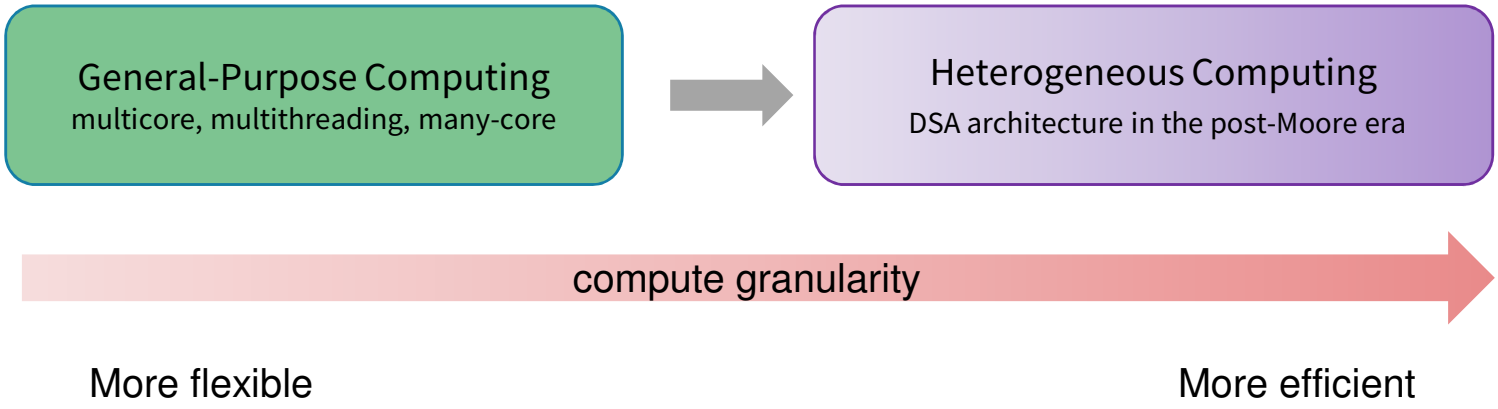
What opportunities do these advancements provide
to
empower AI and HPC applications?

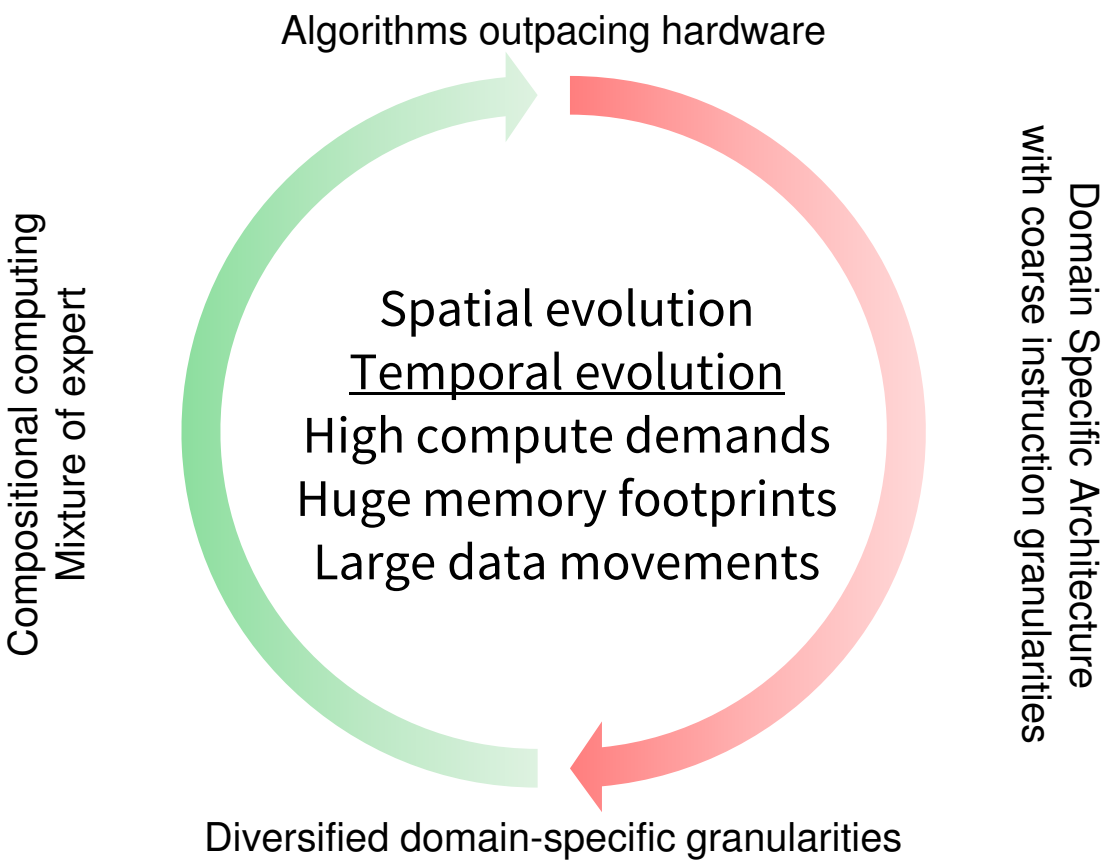
AI Application Characteristics



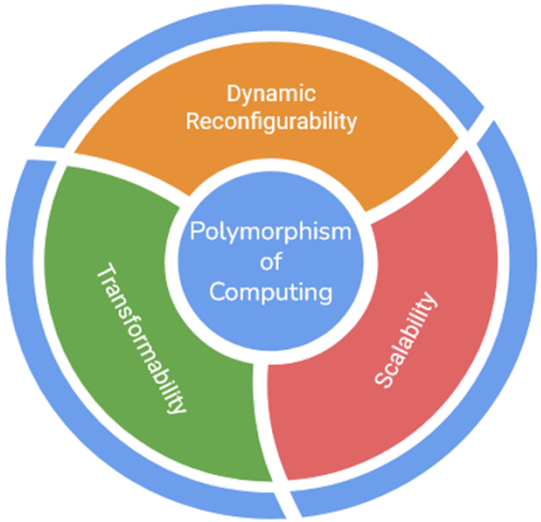
Neuro-compositional computing mush simultaneously enable human-level cognition:
the principles of Compositionality and Continuity

Architecture Advancement




Algorithms outpacing hardware
Domain Specific Architecture
with coarse instruction granularities
Diversified domain-specific granularities
Compositional computing
Mixture of expert
Spatial evolution
Temporal evolution
High compute demands
Huge memory footprints
Large data movements

Polymorphic Architecture



Polymorphism of Computing[2]
Address 2-dimensional evolution challenges

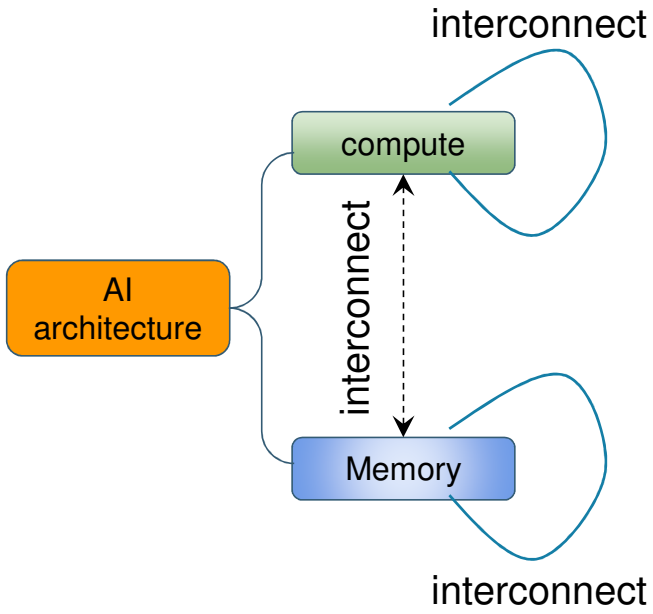
Future-Oriented Computing

Focus on composability and adaptability

- Spatial and temporal scalability
- Composability with diversified granularities
- Re-configurability and transformability

<u>Goal</u>
Achieve the same level of performance as
the purpose-built DSA accelerator
with spatial and temporal scalabilities

Critical Components in AI Architecture



Capability: heterogeneity, efficiency
Scalability: scale up, scale out

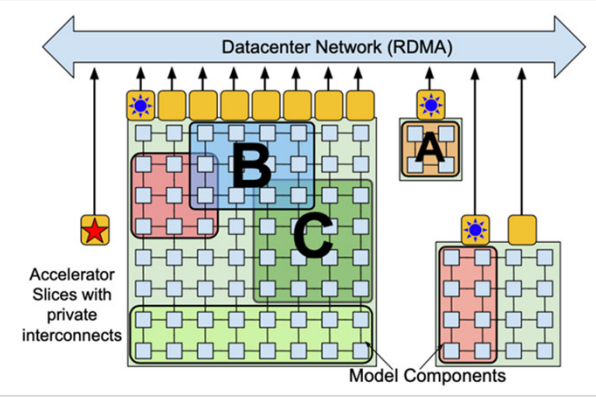
Bandwidth, latency
Topology, coherency

Hierarchical, tiered, disaggregated
Near, far, in-memory

Rationales for Polymorphism

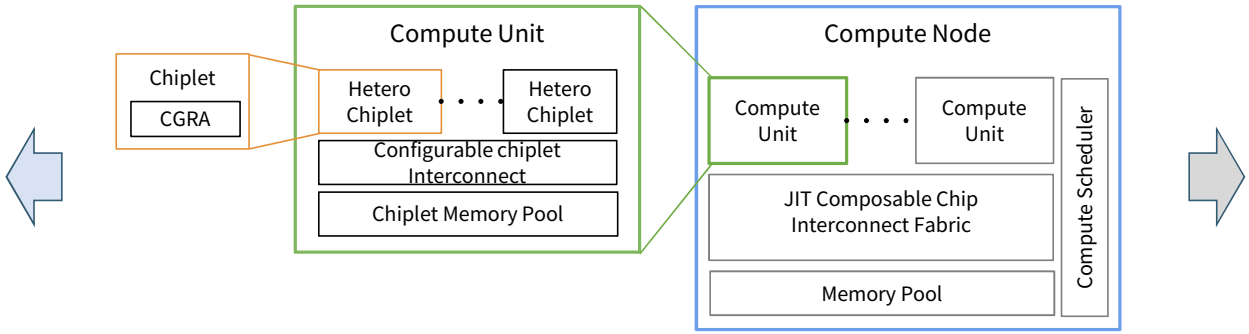
Architectural Composability for neuro-compositional computing

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)



Google Pathway Architecture[3]

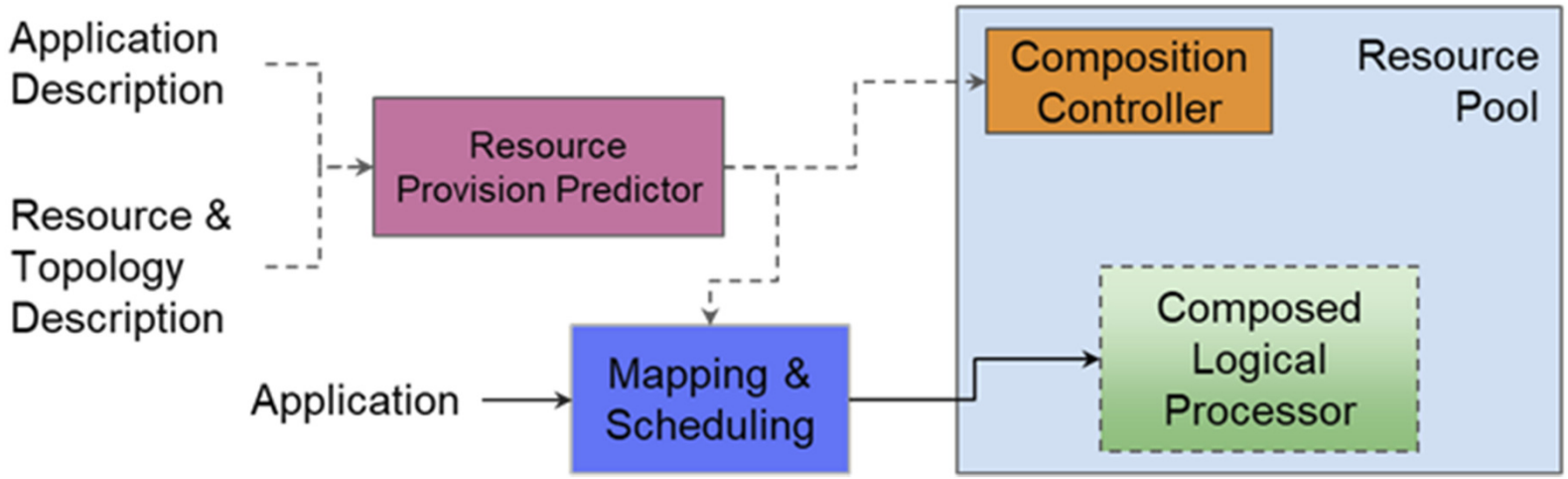
Hierarchical Composability for Polymorphism



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

1. P. Smolensky, et al, "*Neurocompositional Computing: from the Central Paradox of Cognition to a new generation of AI systems*", AI Magazine Vol43, Issue 3, pages 217-340, 2022
2. W. Zhang, "*Polymorphic Architecture for Future AI/ML Applications*", OCP Future Technology Symposium, 2022
3. P. Barham, A. Chowdhery, et al, "*Pathways: Asynchronous Distributed Dataflow for ML*", arXiv, 2022