

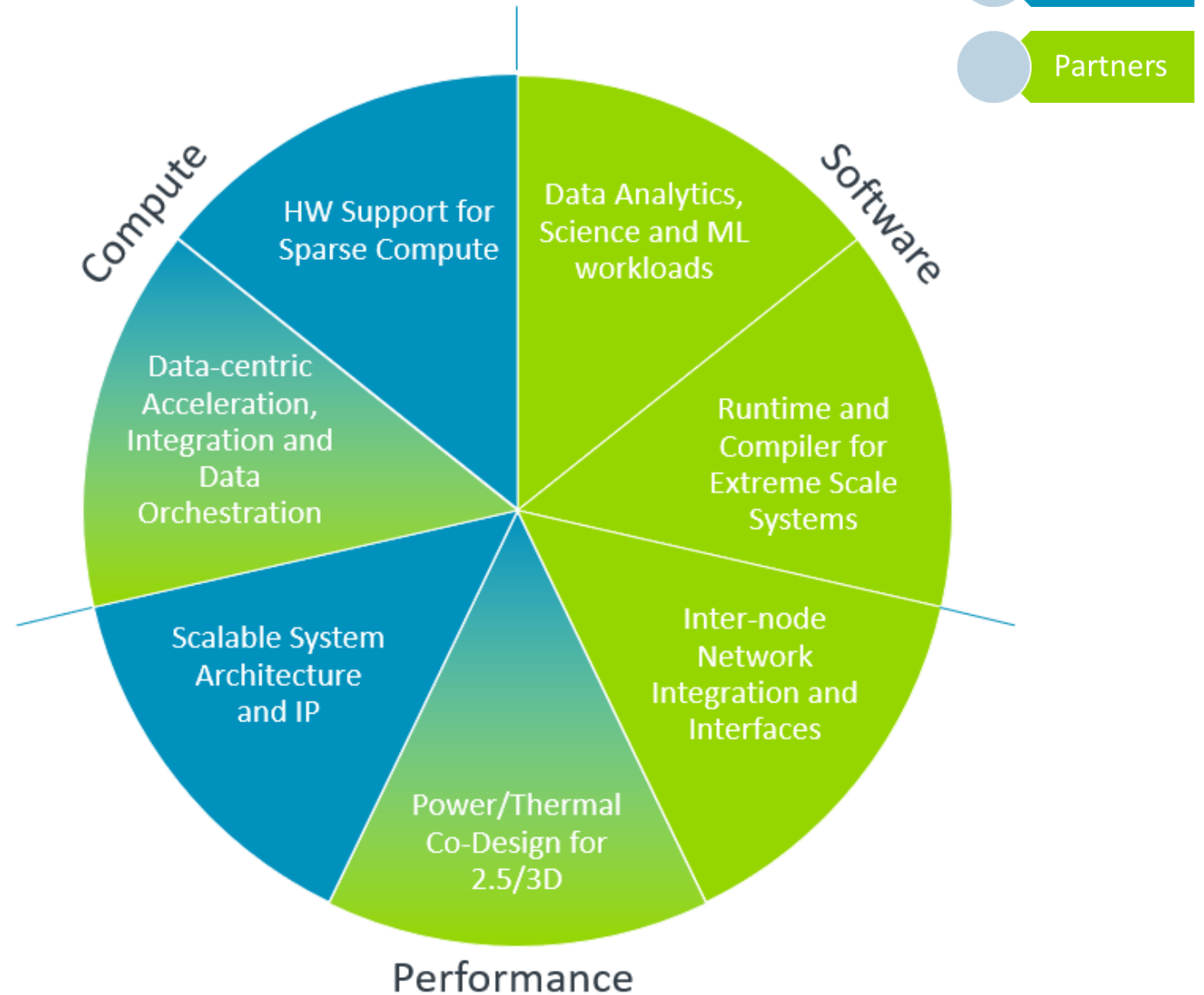


Co-Design of Data Centric HPC Systems Salishan 2022

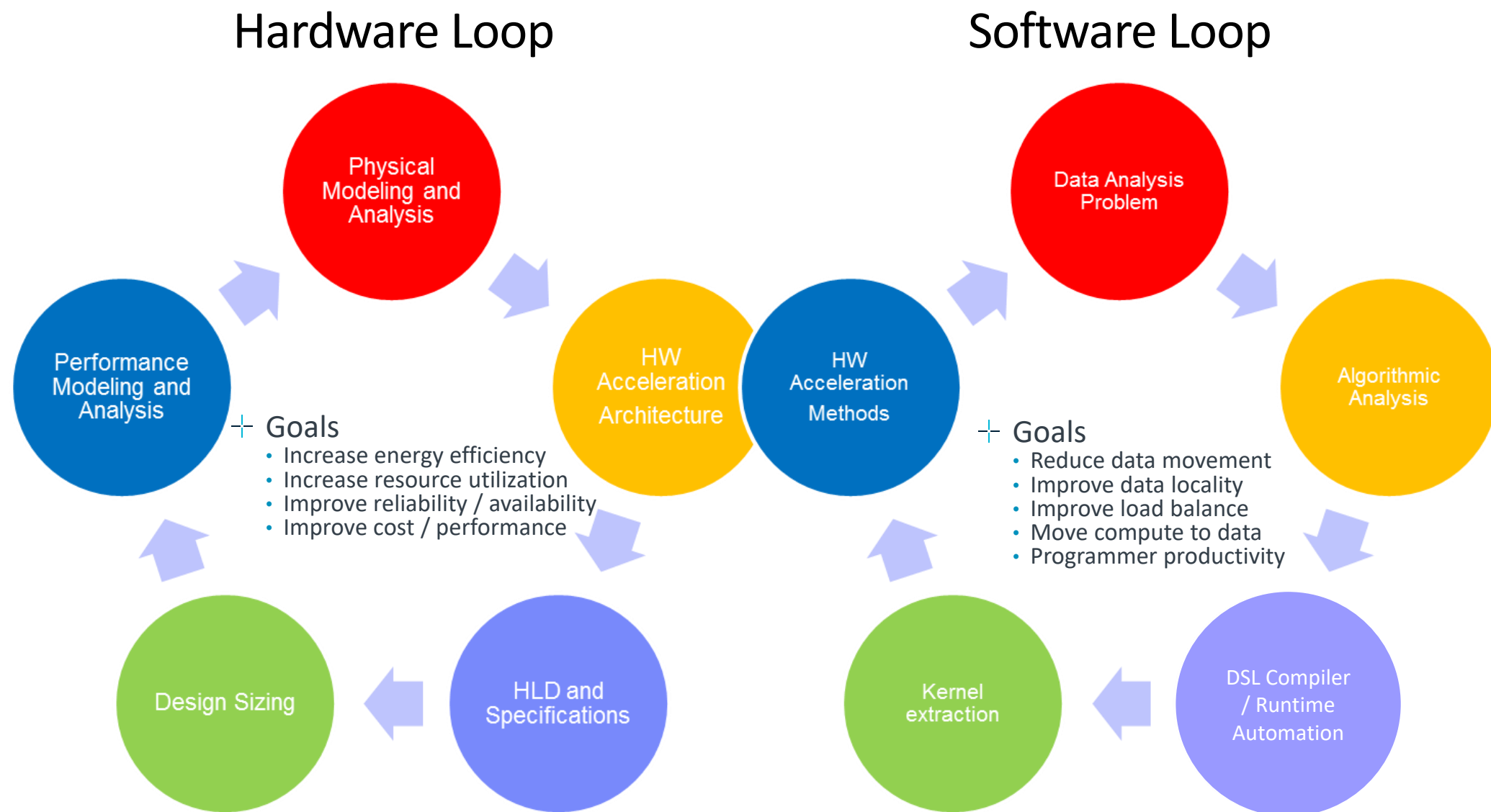
Doug Joseph, ARM Research
ARM Inc

Co-Design Focus Areas

- + Application Areas
 - Data Analytics
 - Machine Learning
 - Science and Engineering
- + Algorithmic Methods
 - Graph Analytics
 - Sparse / Dense Tensor Algebra
 - Statistical Analysis
 - Deep Neural Networks
- + Software Support
 - DSL Compiler and Runtimes
- + Hardware Architecture
 - Network on Chip
 - Memory Hierarchy
 - HW Support for Sparsity
 - Data Centric Accelerators
 - Component Interfaces
- + Enabling Technologies
 - 3D Hybrid Bonding
 - Wafer Fanout Packaging
 - Integrated Power Delivery
 - Advanced Thermal Management
 - 2.5D / 3D Chiplet Interfaces
 - Co-Packaged Si Photonics



Comprehensive Co-Design Flow

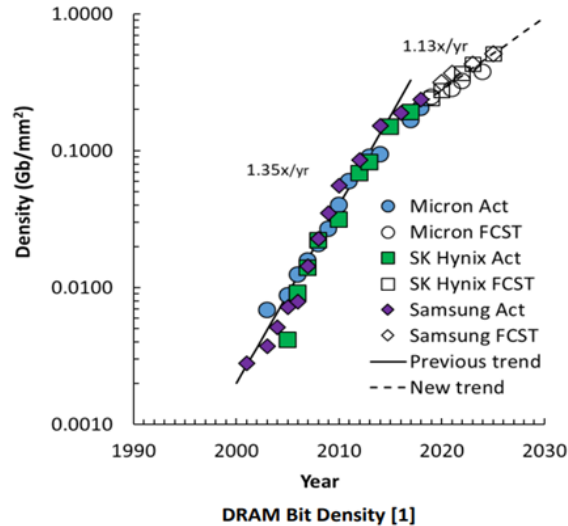


Technology Landscape

Memory Scaling

DRAM Bit Density

- Bit density is die capacity in Gb divided by die size in mm².
- The solid black line is the long term trend based on actual values.
- The dashed black line is the forecasted trend going forward.



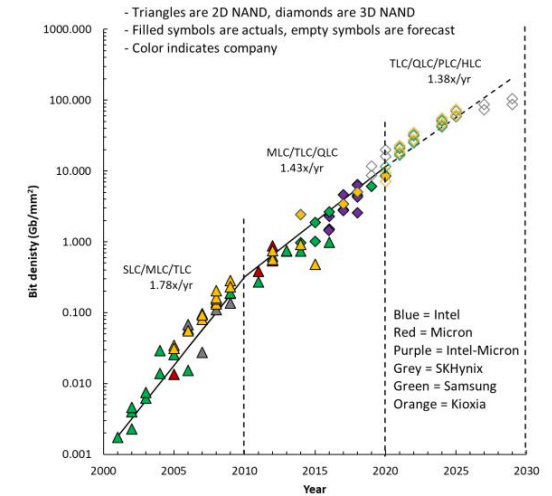
[1] Strategic Cost Model – 2020 – revision 00

ICKNOWLEDGE LLC

1

NAND Bit Density

- The transition from 2D NAND to 3D is enabling the continuation in bit density scaling by using the third dimension.
- Bit density is the number of gigabits of memory on the die divided by the die size.
- Multiple points for the same company in the same year represent MLC/TLC/QLC/PLC/HLC.



[1] Strategic Cost Model – 2020 – revision 00

ICKNOWLEDGE LLC

7

Logic Scaling

N3 PPA (vs. N5 V1.0)

Speed Improvement at Same Power	Power Reduction at Same Speed	Logic Density	SRAM Density	Analog Density
10~15%	25~30%	~1.7x	~1.2X	~1.1x

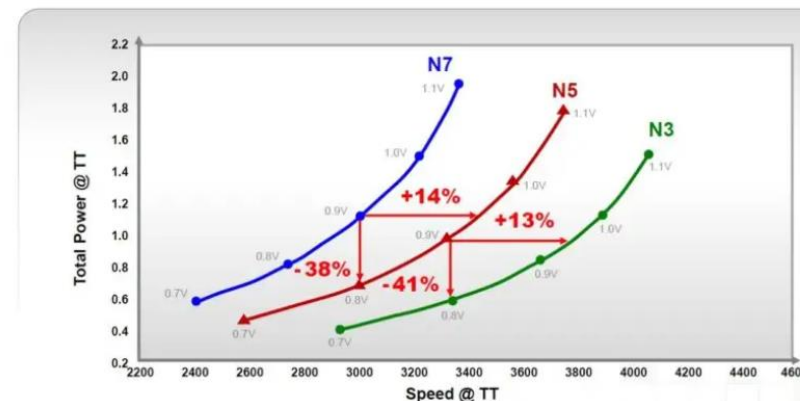
Scaling variance increases each generation.

- SRAM scaling essentially ends at 3nm
- Analog scaling essentially ends at 5nm

5

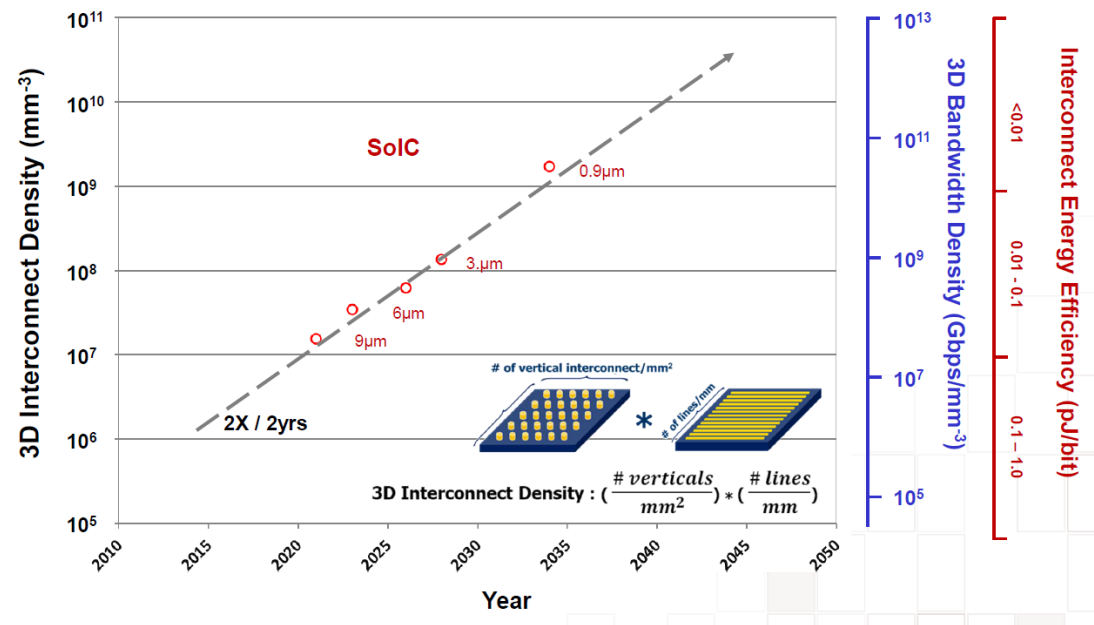
ARM A78 Sub-Block Speed/Power

- with TSMC High Performance Library and Solutions



arm

Inter-chip Interconnect Scaling Roadmap

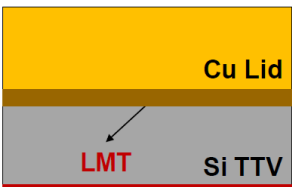


Thermal Management & Power Delivery are Primary Concerns for 3D Integration. (Power Density vs. Power Efficiency Trade-off)

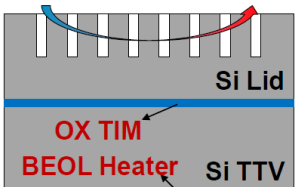
Integrated Si Micro-Cooler (ISMC) for Ultra-HPC

- Thin SiOx bonding interface (OX TIM) by fusion bonding Si lid and Si TTV
- Low interface TR, even though K_{SiOx} at low single digit $\text{W/m}\cdot\text{K}$

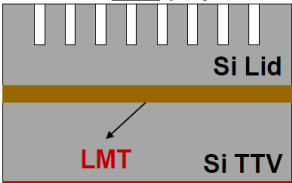
Cu Lid with LMT (Liquid Metal TIM)



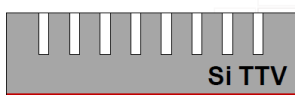
Si Lid with OX TIM



Si Lid with LMT (Liquid Metal TIM)



DWC (Direct Water Cooling)

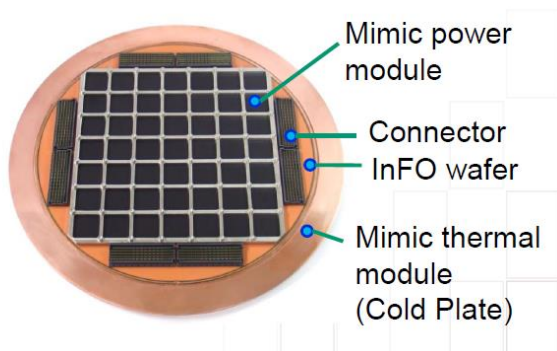


TSMC-SoIC™
Lite-IO (TSMC)
25 Tbps/mm ²
0.02~0.04 pJ/bit
2~4 Gbps

UCle

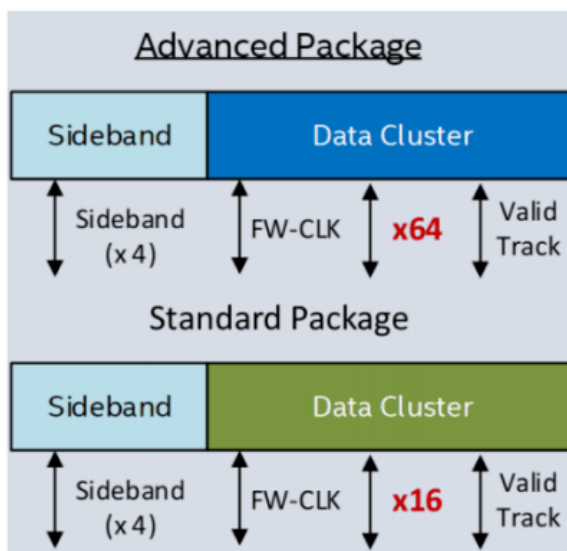
Universal Chiplet Interconnect Express

- InFO_SoW (System-on-Wafer)**



CoWoS®	InFO
HBI (Cadence, Synopsys), LIPIN	
1.15~2 Tbps/mm	1.15~2 Tbps/mm
0.2~0.5 pJ/bit	

Characteristics / KPIs	Standard Package	Advanced Package	Comments
Characteristics			
Data Rate (GT/s)	4, 8, 12, 16, 24, 32		Lower speeds must be supported -interop (e.g., 4, 8, 12 for 12G device)
Width (each cluster)	16	64	Width degradation in Standard, spare lanes in Advanced
Bump Pitch (um)	100 – 130	25 - 55	Interoperate across bump pitches in each package type across nodes
Channel Reach (mm)	<= 25	<=2	
Target for Key Metrics			
B/W Shoreline (GB/s/mm)	28 – 224	165 – 1317	Conservatively estimated: AP: 45u for AP; Standard: 110u; Proportionate to data rate (4G – 32G)
B/W Density (GB/s/mm²)	22-125	188-1350	
Power Efficiency target (pJ/b)	0.5	0.25	
Low-power entry/exit	0.5ns <=16G, 0.5-1ns >=24G		Power savings estimated at >= 85%
Latency (Tx + Rx)	< 2ns		Includes D2D Adapter and PHY (FDI to bump and back)
Reliability (FIT)	0 < FIT (Failure In Time) << 1		FIT: #failures in a billion hours (expecting ~1E-10) w/ CXi Flit Mode



Die - 1		Die - 2			
x16	<-->	x16	CL-0 x16	<-->	CL-0 x16
x32	<-->	x32	CL-0 x16	<-->	CL-0 x16
			CL-1 x16	<-->	CL-1 x16
x64	<-->	x64	CL-0 x16	<-->	CL-0 x16
			CL-1 x16	<-->	CL-1 x16
			CL-2 x16	<-->	CL-2 x16
			CL-3 x16	<-->	CL-3 x16

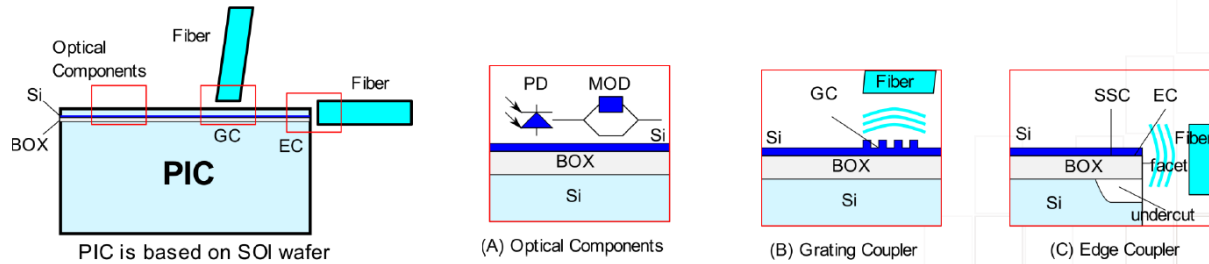
(1, 2, or 4 Clusters can be combined in one UCle Link)

Photonics (TSMC, Hot Chips 33)

Optical Interface (1/2): Overview

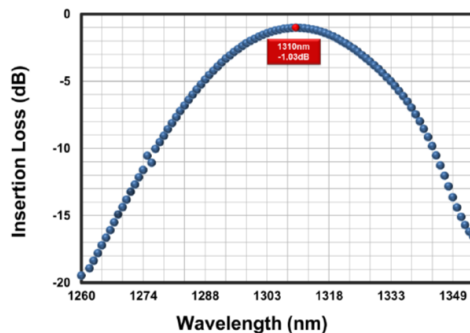
Light can be coupled either vertically (GC) or horizontally (EC):

- GC, as a surface coupler, requires cleanliness and integrity of the optical path from grating surface all the way to the fiber core.
- For EC, care must be taken to prevent the expanded optical mode from overlapping with the bulk silicon underneath SSC.

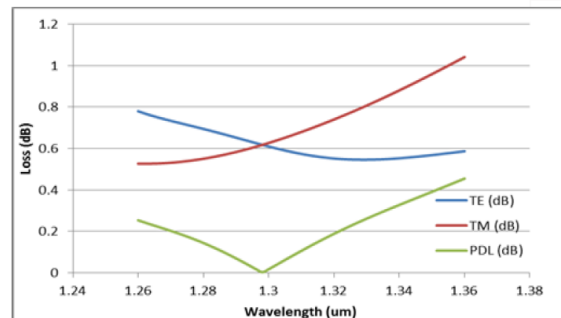


Optical Interface (2/2): GC and EC with COUPE

- GC is designed with optical path intrinsically sealed with dielectrics all the way to the fiber attachment unit, achieving IL (1D apodized GC) -1.03dB @1310nm for TE
- EC avoids optical loss due to beam overlapped with underneath Si, achieving IL -0.6dB @1310nm for TE&TM modes
- With COUPE, GC and EC can built with essentially the same structure.



Grating Coupler Insertion Loss



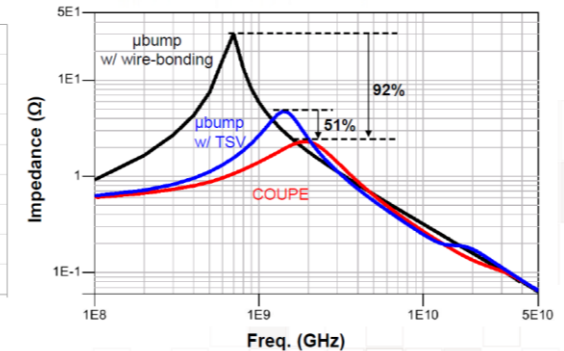
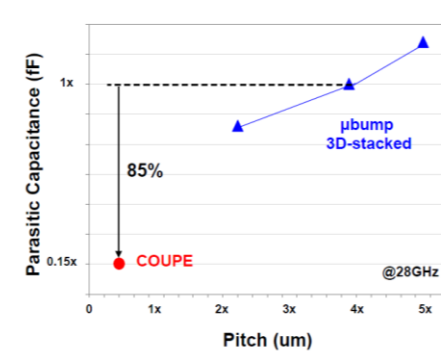
Edge Coupler Insertion Loss

DARPA PIPES (Columbia-AIM)

	Completed Phase 1	Phase 2
Key outcomes	EIC/PIC MCM 1Tbps/link 2 port prototype. Benchtop component demonstration	Integrated link demonstration performance traceable to scaled system
Energy per bit	0.5 pJ/bit	0.2 pJ/bit
Areal bandwidth density	5 Tbps/mm ²	5 Tbps/mm ²
Channel data rate	16 Gbps	
Comb bandwidth >0.5mW	45nm; 80nm >0.1mW	
Aggregate bandwidth	2 Tb/s	10 Tbps
Total port count	2	≥ 1
Power Penalty	16 dB	
Link latency	40 ns + TOF	100 ns + TOF
Link reach (between packages)	1 meters	10 meters
Bit error ratio (BER)	10 ⁻⁹	10 ⁻¹²
Hardware delivered	Benchtop MCM prototype, components demo	2 demo units
Operating temperature range	Room temperature	Room temperature to 80°C

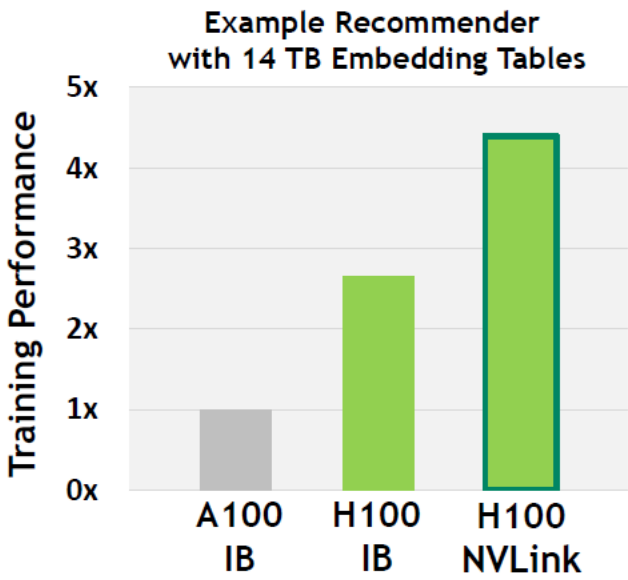
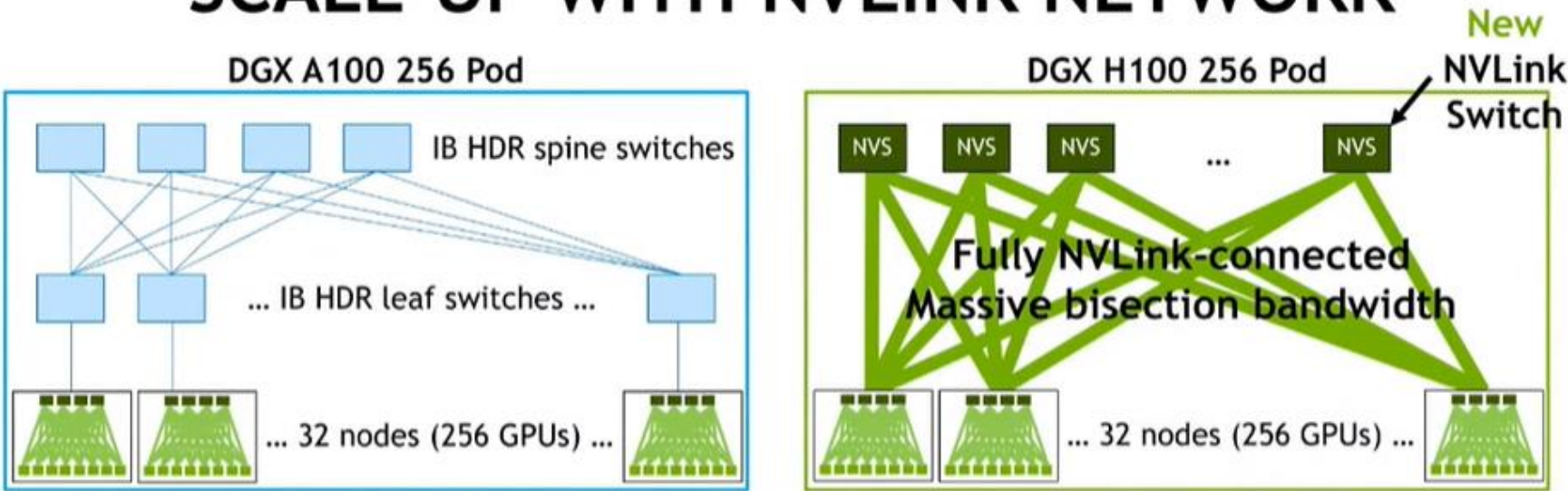
Electrical Interface (1/2)- Parasitics and PDN Impedance

- COUPE has low parasitics at EIC-PIC Electrical Interface, 85% lower capacitance compared with uBump
- 51% reduction in PDN impedance comparing with uBump w/ TSV; and 92% reduction of uBump w/ wire-bonding.



NVidia Grace-Hopper SuperPod

SCALE-UP WITH NVLINK NETWORK

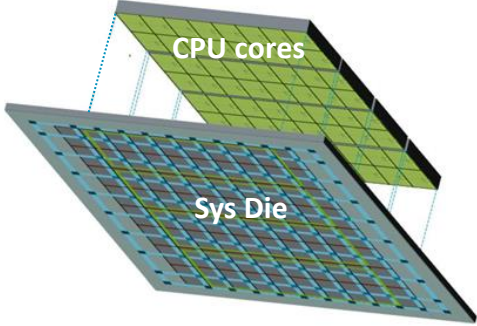
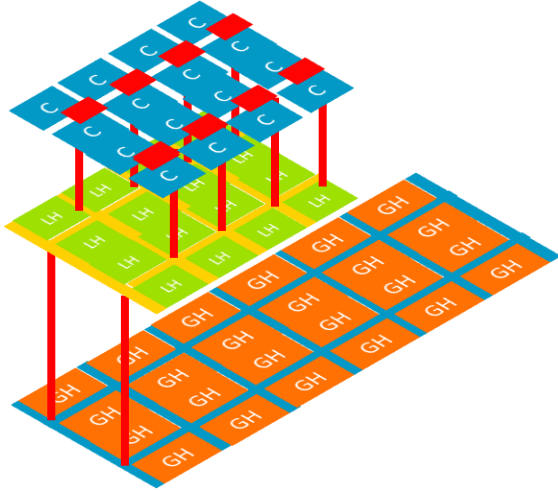
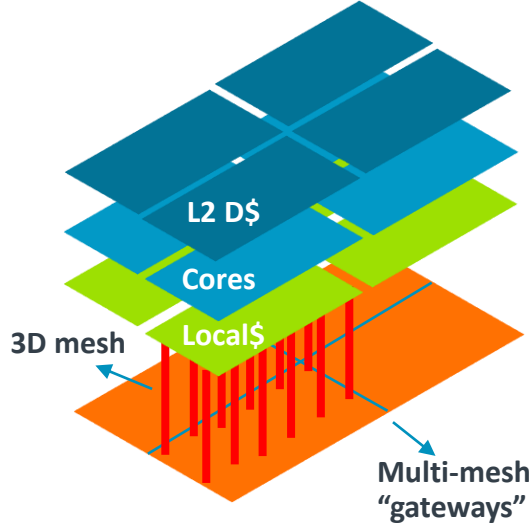
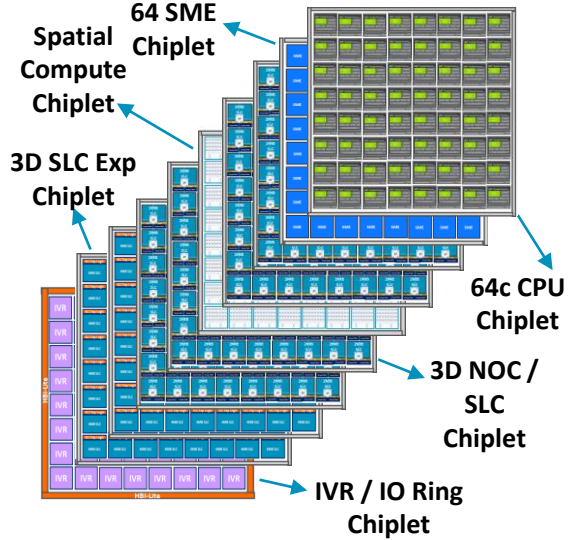


	A100 SuperPod			H100 SuperPod			Speedup	
	Dense PFLOP/s	Bisection [GB/s]	Reduce [GB/s]	Dense PFLOP/s	Bisection [GB/s]	Reduce [GB/s]	Bisection	Reduce
1 DGX / 8 GPUs	2.5	2,400	150	16	3,600	450	1.5x	3x
32 DGXs / 256 GPUs	80	6,400	100	512	57,600	450	9x	4.5x

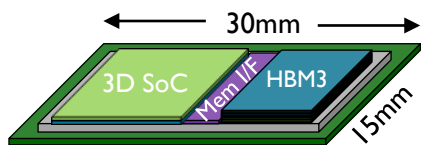
Moving compute to data has big payoff at scale !

The need for integrated Si Photonics is growing !

3D Co-Design Study Roadmap

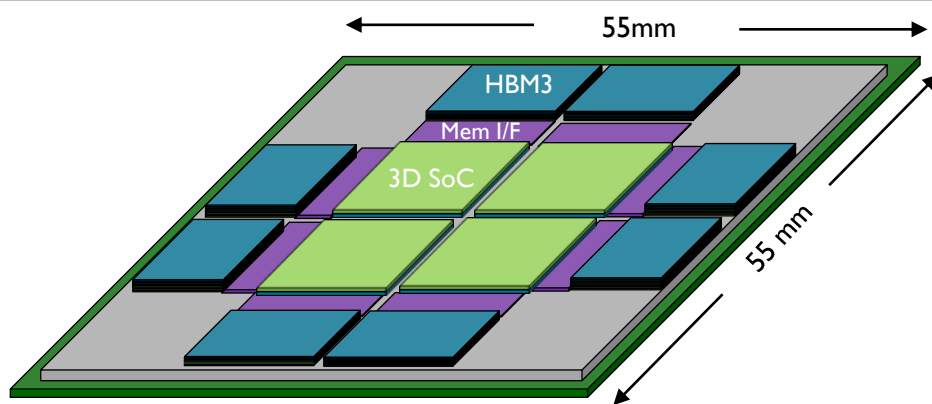
	2023	2024	2026	2030
Features	 <p>CPU cores Sys Die</p>	 <p>CPU Local\$ GH</p>	 <p>L2 D\$ Cores Local\$ 3D mesh Multi-mesh "gateways"</p>	 <p>64 SME Chiplet Spatial Compute Chiplet 3D SLC Exp Chiplet 64c CPU Chiplet 3D NOC / SLC Chiplet IVR / IO Ring Chiplet</p>
	<ul style="list-style-type: none"> CPU layer: higher core density SYS: mesh, SLC and mem/IO 	<ul style="list-style-type: none"> CPU: higher core density Local: cluster cache and mesh SYS: global mesh, SLC and mem/IO 	<ul style="list-style-type: none"> CPU with large 3D L2 Data SRAM CPUs with SMEs Multi 3D mesh system Face-to-face bonding Back-side power delivery 	<ul style="list-style-type: none"> CPU and SME in 3D Spatial data orchestration engines Multi-layer SLC Multi 3D mesh system Integrated IVR for power/thermal mgmt. Cryo thermal solution
Challenges	<ul style="list-style-type: none"> System partitioning and exploration Power delivery thermals 	<ul style="list-style-type: none"> System partitioning and exploration 3D timing for CMN components Power delivery/thermals+ 	<ul style="list-style-type: none"> NoC topology and adaptive routing System partitioning and exploration 3D timing for CPU-L2 interface System challenges: back-side PDN Power delivery/thermals++ 	<ul style="list-style-type: none"> 3D timing for HNF-SLC interface 3D timing for CPU-SME interface Software for general use of acceleration and data orchestration Power delivery/thermals++++

2.5D + 3D Scaling Opportunities



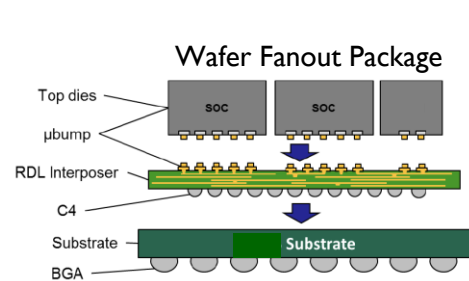
Edge (64 cores ~100W)

- One 3D SoC
- 32GB HBM3 stack

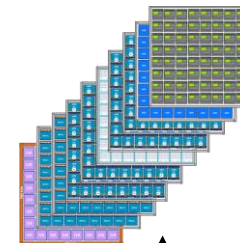


Data Center (256 cores ~500W)

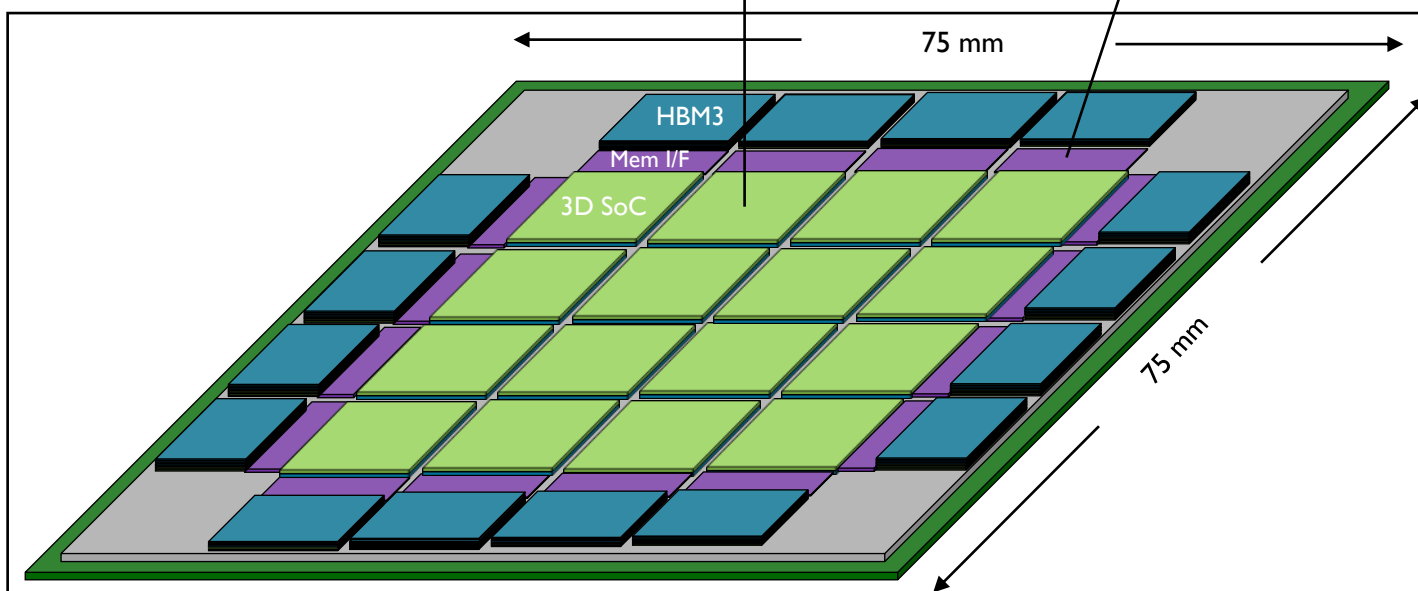
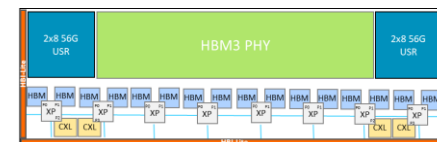
- 2x2 3D SoC
- 256GB in 8x HBM3 stacks



Heterogenous 3D SoC



Key Technology:
Separate Memory & IO Chiplet

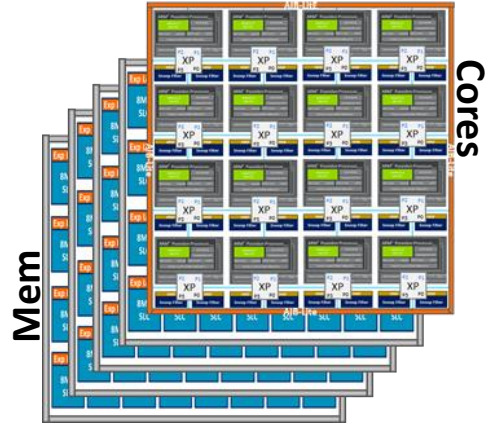


HPC (1024 cores ~2KW)

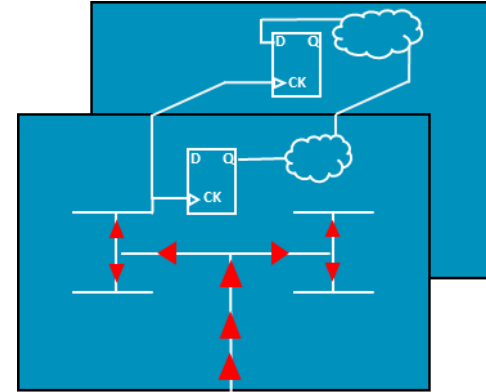
- 4x4 3D SoC
- 512GB in 16x HBM3 stacks

Co-Design: 3D Physical Design

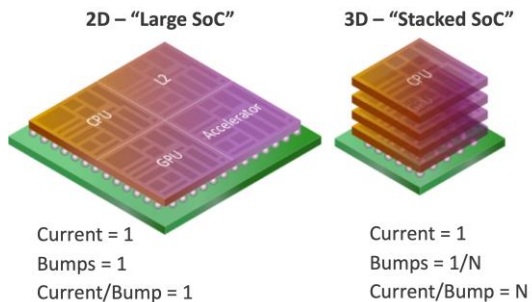
3D design challenges



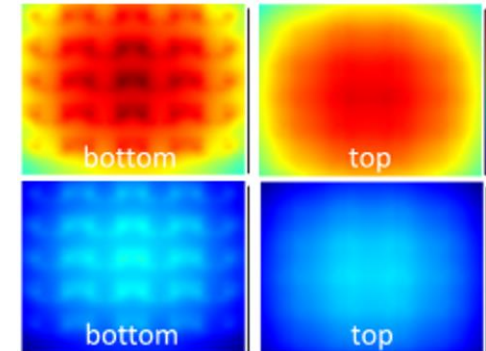
- System Partitioning
 - Node, tier assignment, partitioning and 3D floorplanning



- Timing for synchronous 3D
 - Inter-tier skew and clock design strategies for 3D



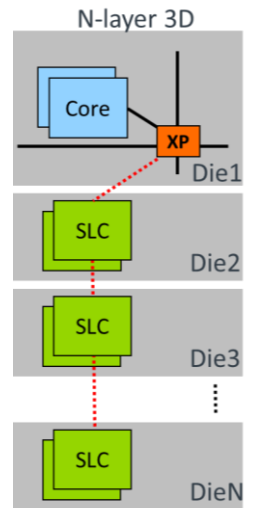
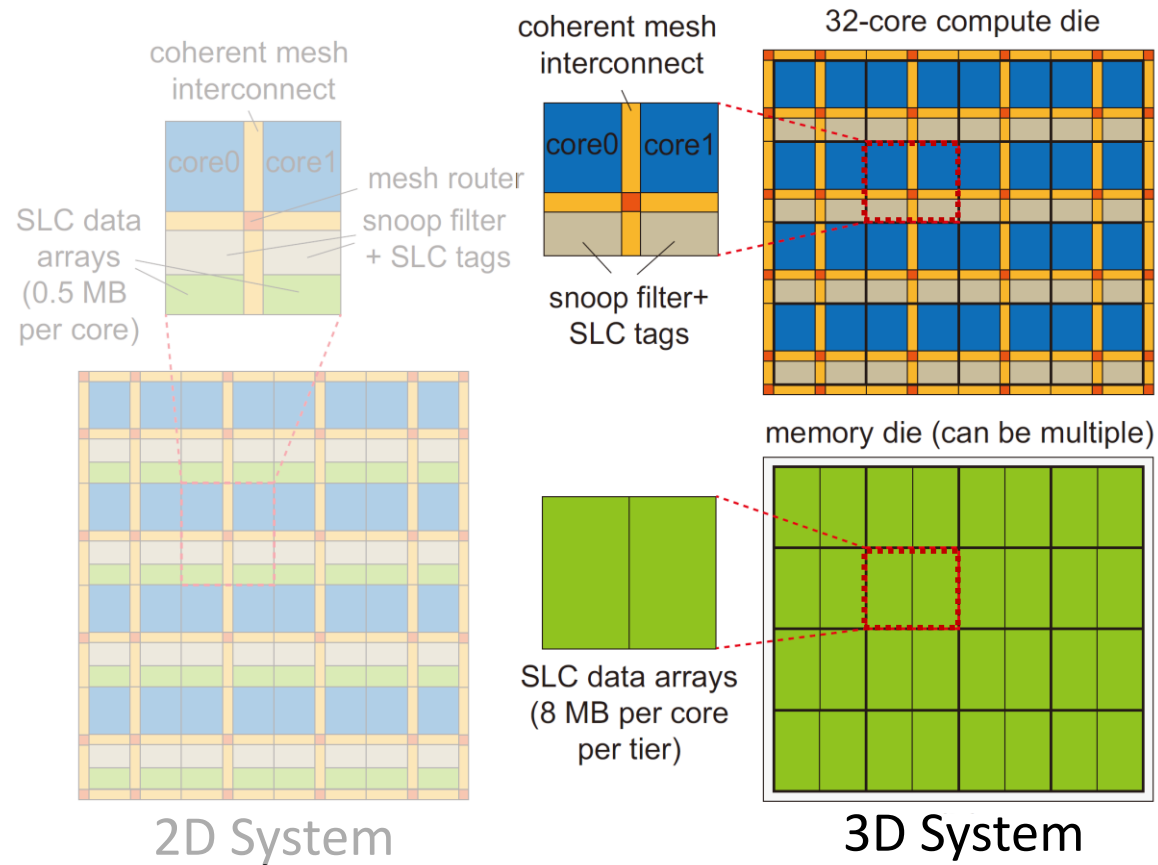
- Power Delivery & Management
 - Power allocation and distribution, voltage droop management



- Thermal Management
 - Thermal sensing capability, and tier placement

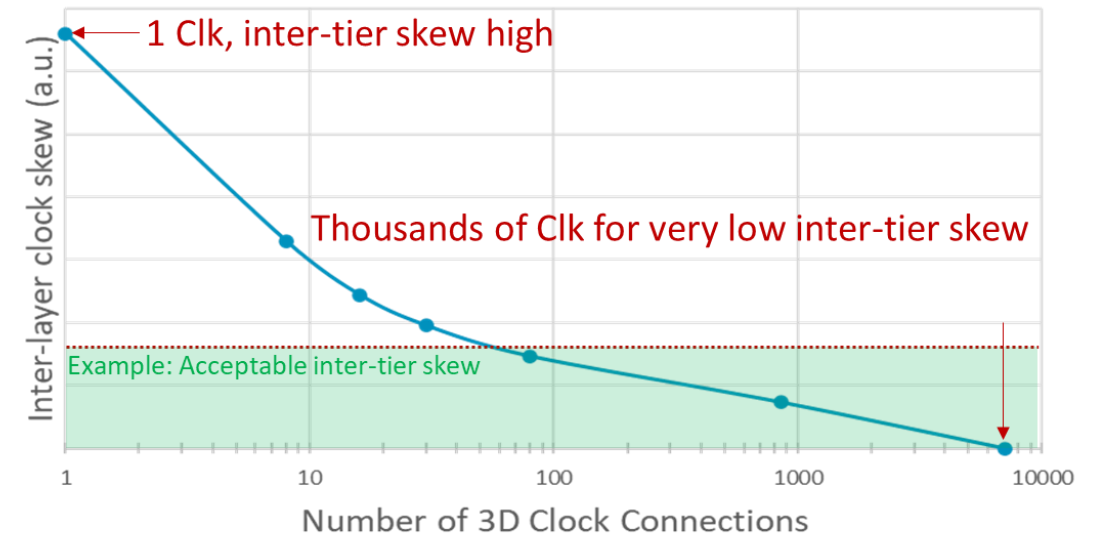
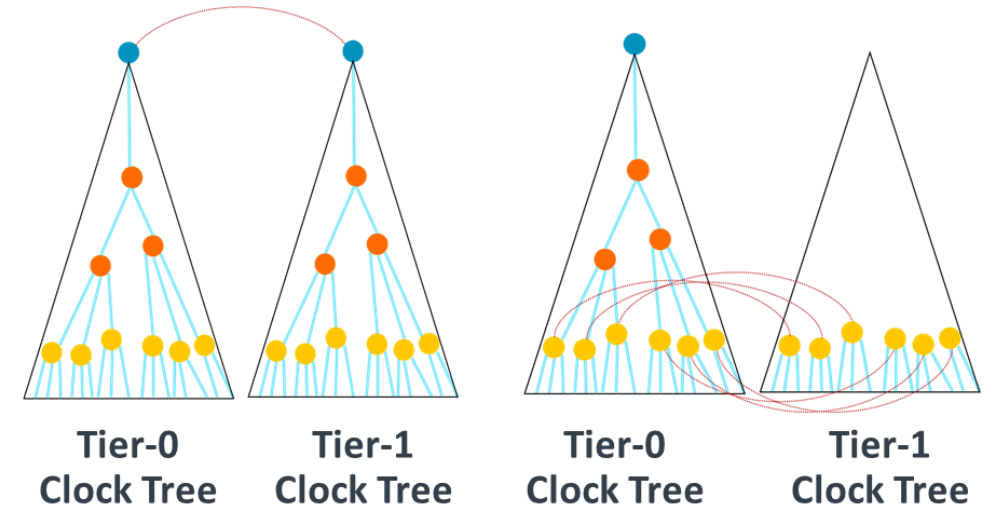
Partitioning: 3D system design case-study

- 32-core system
 - High-performance Arm cores
 - System-level caches (SLC)
 - Cache-coherent mesh interconnect
- Limited space in 2D
 - More compute or more memory?
- 3D integration
 - Decouples increasing number of cores from cache capacity
 - Allows adding SLC expansion tiers

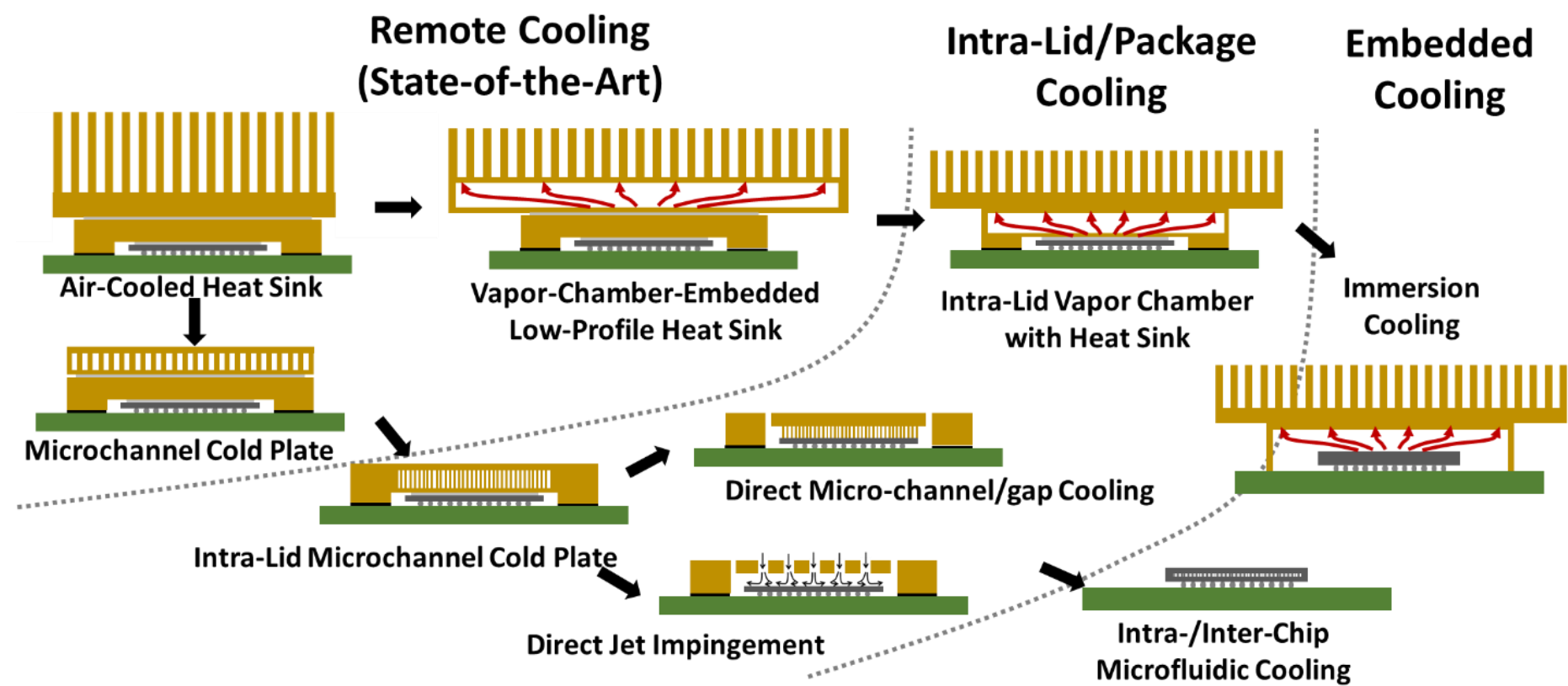


3D timing: Inter-tier skew

- Process variation across tier
 - Leads to inter-tier skew on uncommon clock tree path
- Connect at root
 - Small #3D connections but large uncommon path => Large inter-tier skew
- Connect near leaf
 - Large #3D connections but small uncommon path => Small inter-tier skew

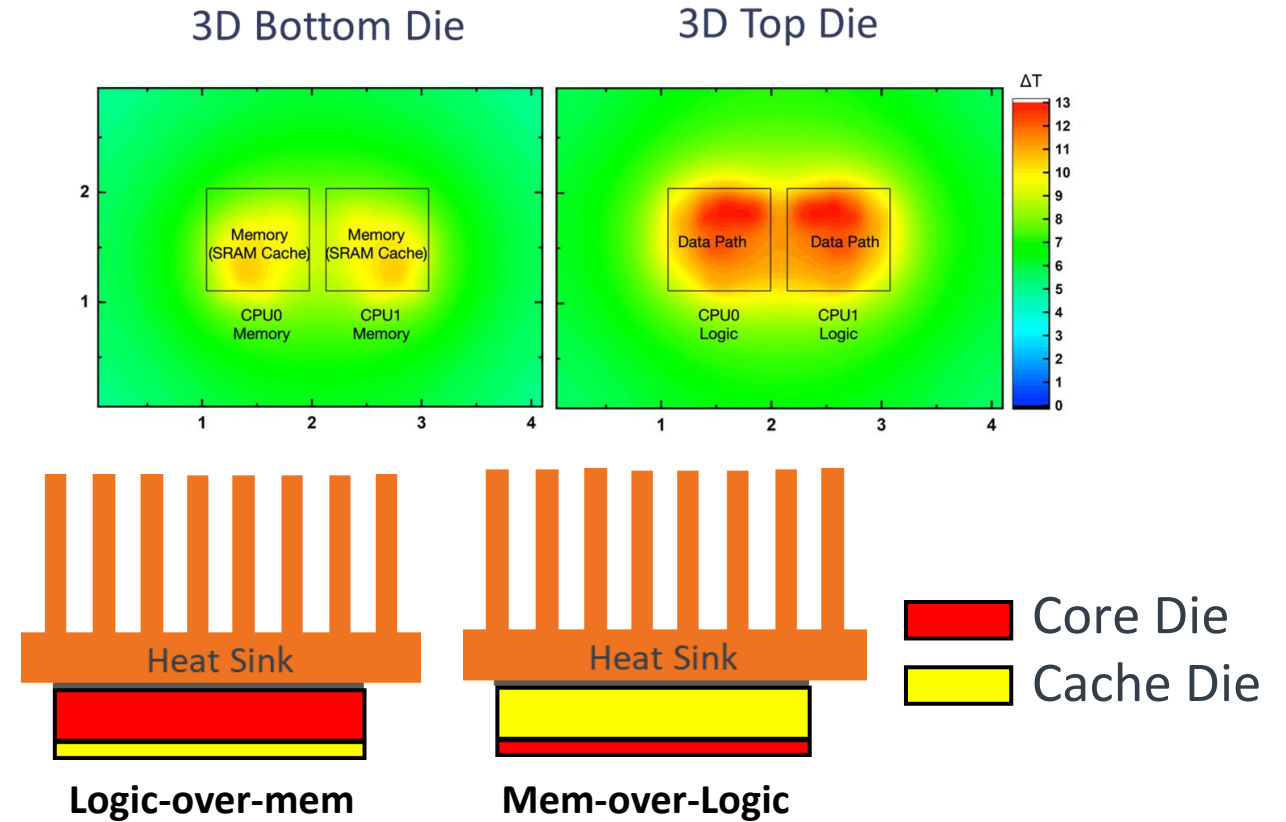
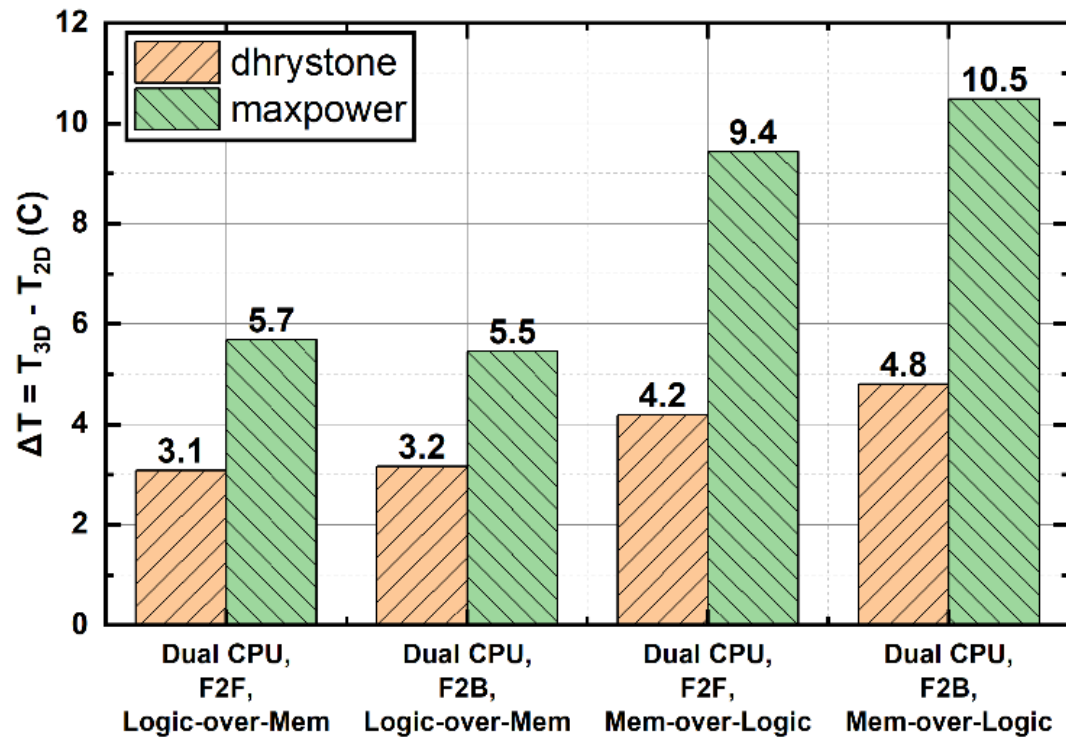


Thermal Solution Landscape



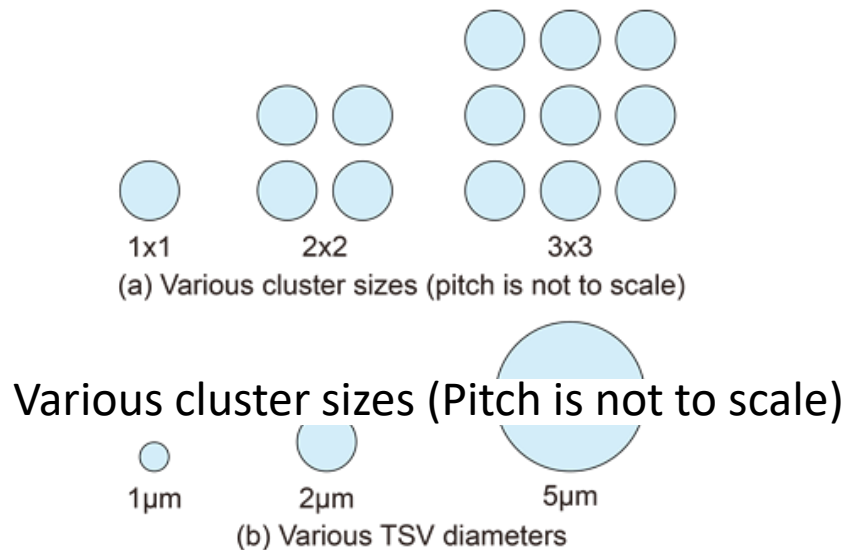
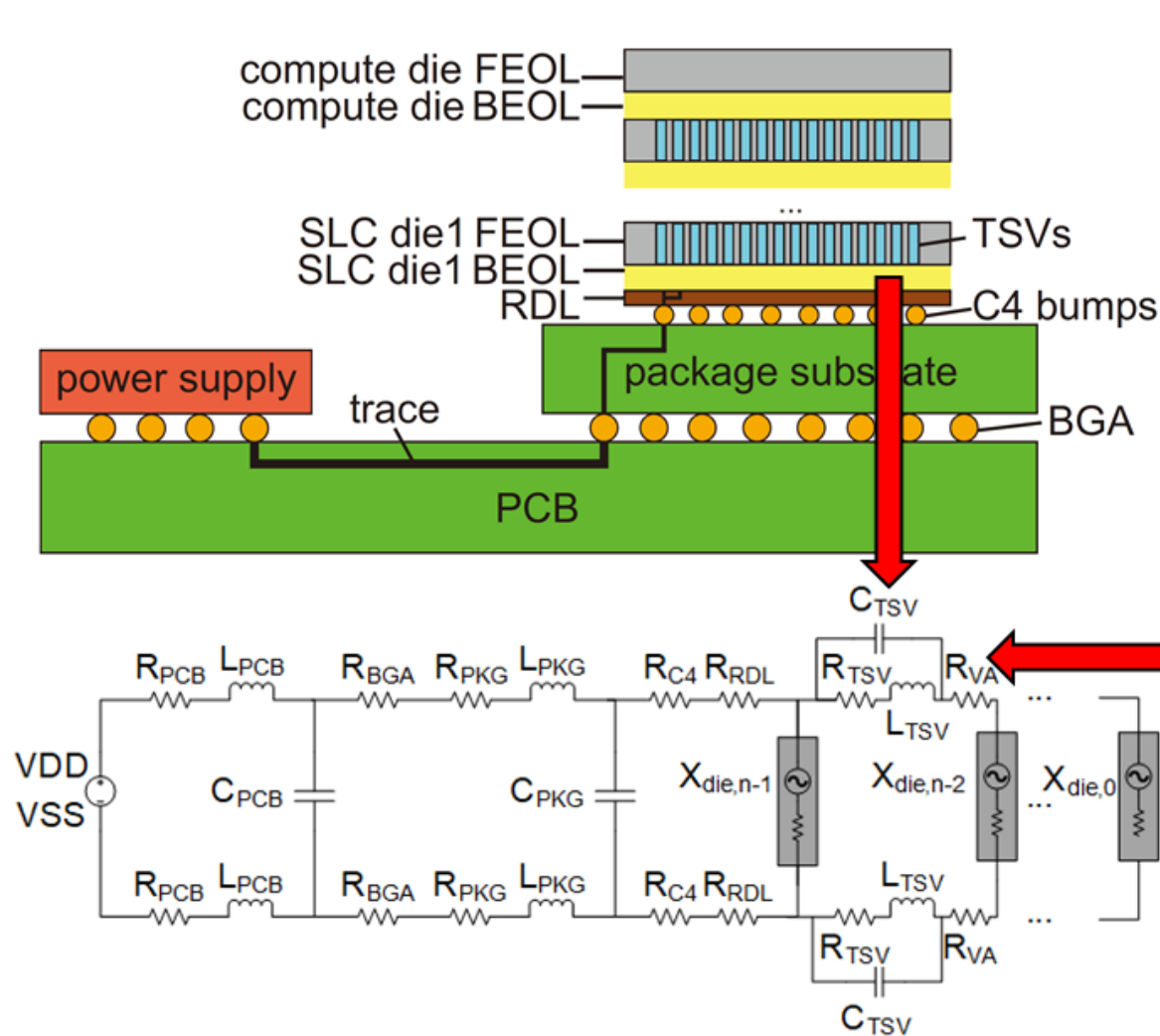
	Remote Cooling	Intra-lid / Package Cooling	Embedded Cooling
Cooling efficiency	Low	Medium	High
Cost	Low	Medium	High

3D thermal design



- Power density increasing as area continues to scale down with newer technology
- Temperature rise is proportional to the power density of the design
- Higher power die near the heat sink is preferred for lower temperature rise

3D power delivery and management

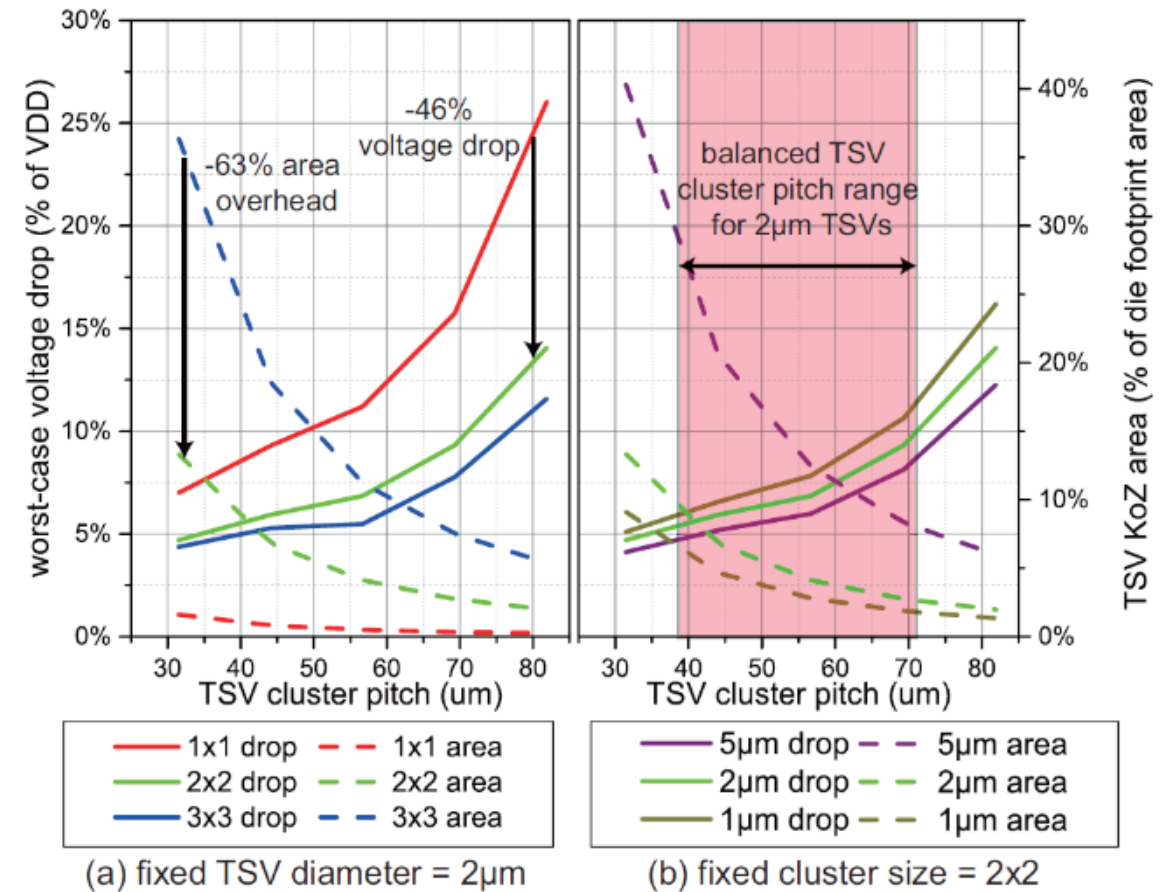


TSV configurations			
Various TSV Diameters 10:1			
TSV cluster size options	1x1, 2x2, or 3x3		
TSV diameter (µm)	5	2	1
TSV length (µm ²)	50	20	10
TSV min. pitch (µm)	10	7.5	2
2x2 cluster KoZ (µm ²)	400	132.25	16
normalized TSV R	1.00	2.50	5.00
normalized TSV L	1.00	0.40	0.20
normalized TSV C	1.00	0.40	0.20

L. Zhu et al., ISLPED'21

3D power delivery and management

- TSV pitch and parasitics have significant impact on voltage drop
- Decreasing power TSV pitch
 - Decreases voltage drop
 - Increases area overhead
- Trade off the voltage drop and area overhead for power delivery TSVs

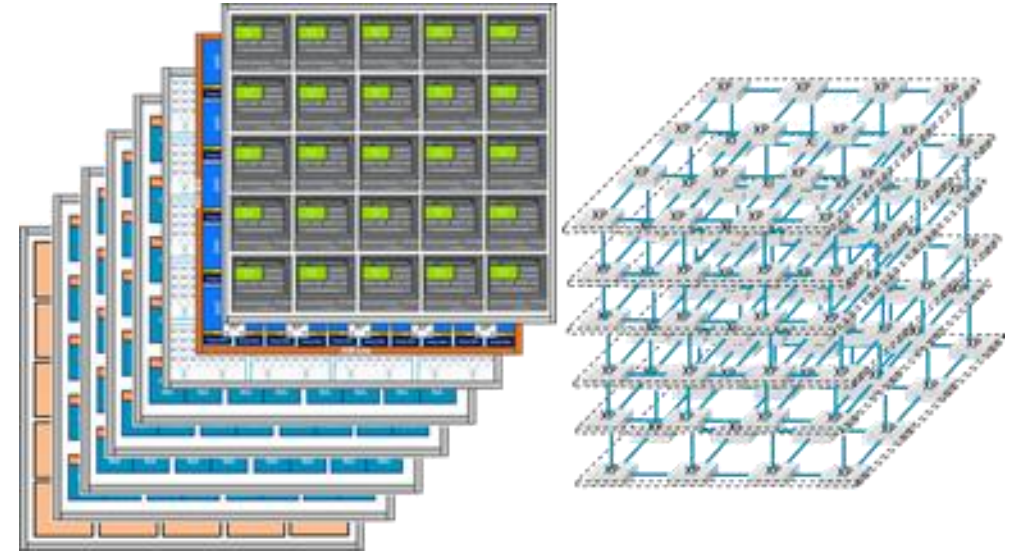


L. Zhu et al., ISLPED'21

Co-Design: 3D Network on Chip

Expanding NoC to 3D layers

- + Lower Manhattan distance between endpoints -> Lower data access latency
- + Higher bi-section bandwidth
- + Research topics
 - 1) Topology and system partitioning exploration
 - 2) Explore adaptive routing algorithms
 - 3) QoS management
 - 4) Cache Coherence Scaling
 - 5) SLC optimizations
 - 6) Support for Multicast and Collectives

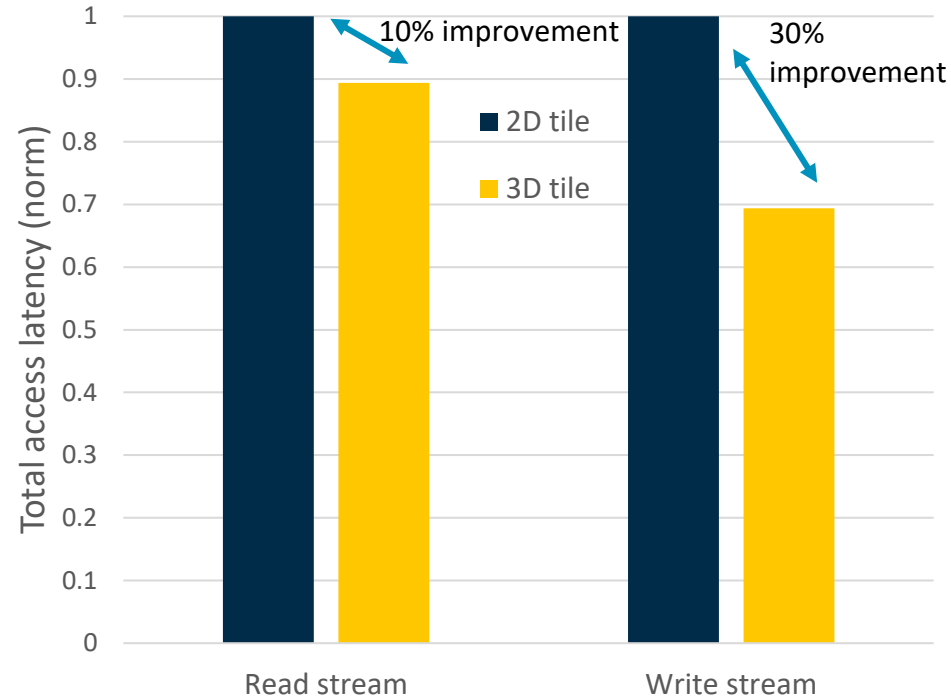


Multi-layer 3D mesh with 4x4x4 XPs

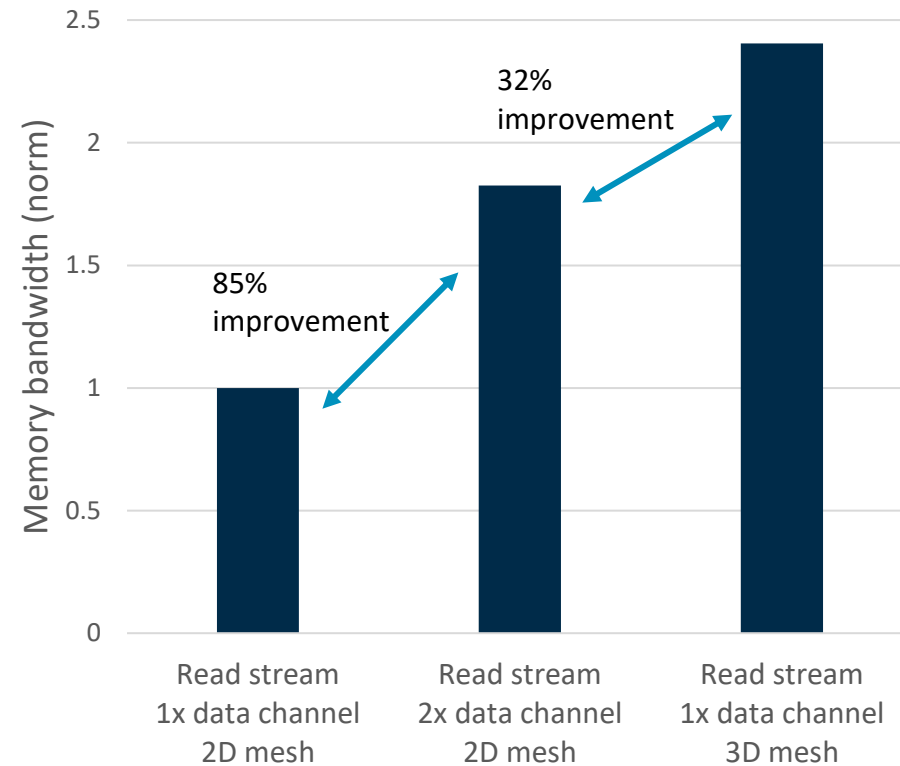
2D vs 3D tiles : latency and bandwidth

10x10 mesh / 128 cores / 4xHBM2 stack

+ 10% – 30% faster accesses with 3D tiles



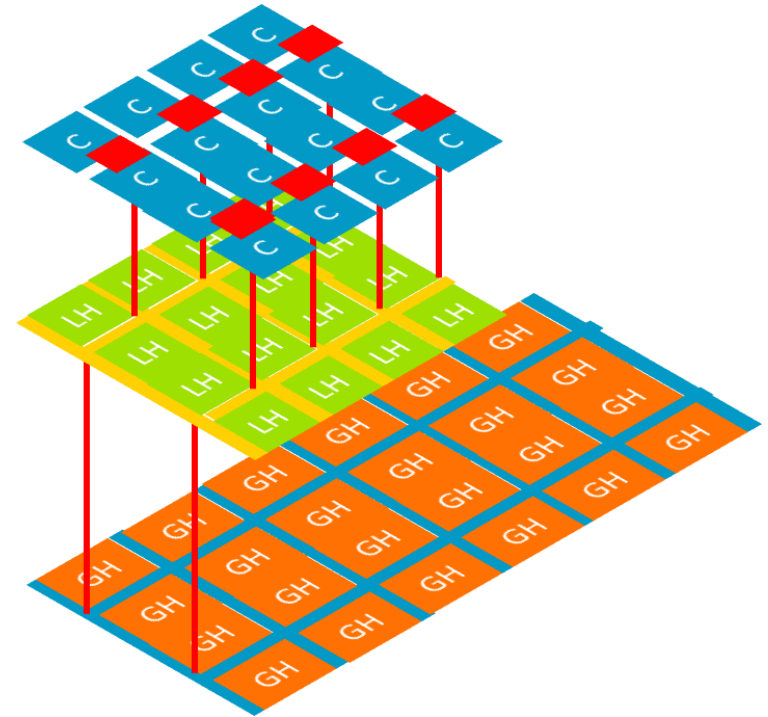
• Bandwidth



- Future-proof NoC need to provision for even more BW (e.g. >4TB/s for HBM3)
- Bandwidth improves by adding more data channels and bisection BW ?
 - Not scalable with a 2D mesh
 - 3D mesh naturally increase channel availability (see notes for this slide for details)

Topology and system partitioning exploration

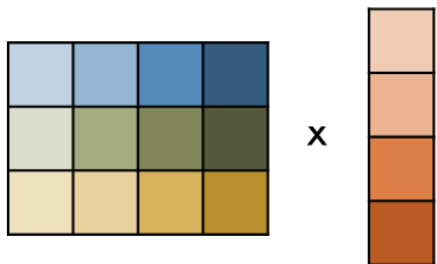
- Explore tradeoffs of different endpoint distribution for cores, SLC, HBM, accelerators and IO
- Explore mesh topologies (e.g. regular/irregular meshes, hypercubes, fat trees)
- Potential for novel cache hierarchy options with 3D integration and 3D NoC
- DSE example: 3 Layers
 - Top: Cores
 - Middle: Local HNFs (LH) and mesh
 - Bottom: Global HNFs (GH) + mesh
- Some more DSE points
 - Mesh on core layer ?
 - CHI channels per layer
 - Num of Z-dim connections vs TSV placement constraints



Co-Design:
Data Centric Accelerators
&
DSL Compiler / Runtimes

Current systems are optimized for regular computations

Regular / Dense (e.g., dense linear algebra)



Vector processing, GPUs
Prefetchers

Memory optimized for bulk transfers
(Lack of) HW synchronization

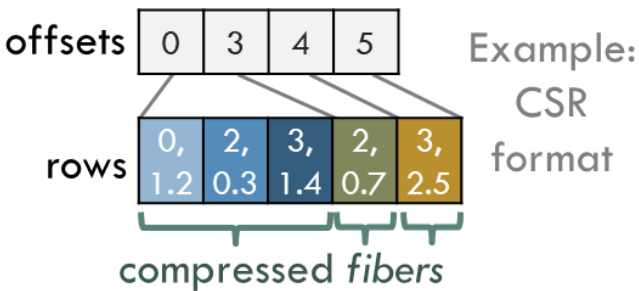
Tiling, polyhedral model
Compiler-supported parallelization

Language-level support
(code and data abstractions)



Irregular / Sparse (e.g., sparse linear algebra)

	0	1	2	3
0	1.2	0	0.3	1.4
1	0	0	0.7	0
2	0	0	0	2.5



Result: Percentage of peak utilization in supercomputers

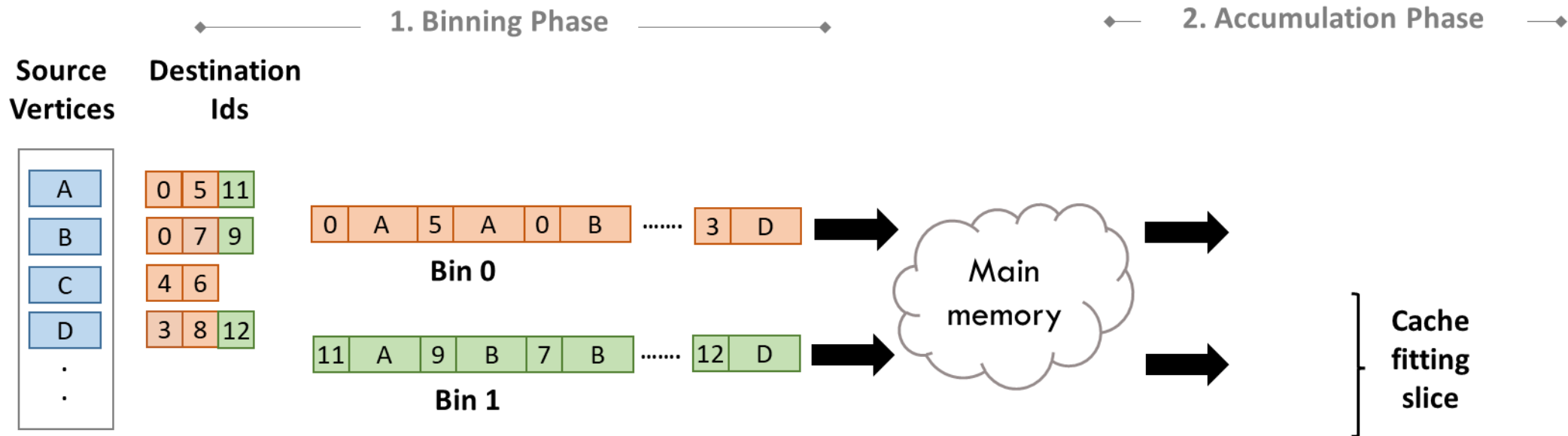
Dense linear algebra	50-80%
Sparse linear algebra	1-3%
Graph analytics	<<1%

Similar inefficiencies in accelerators (e.g., no/limited support for sparse deep learning)

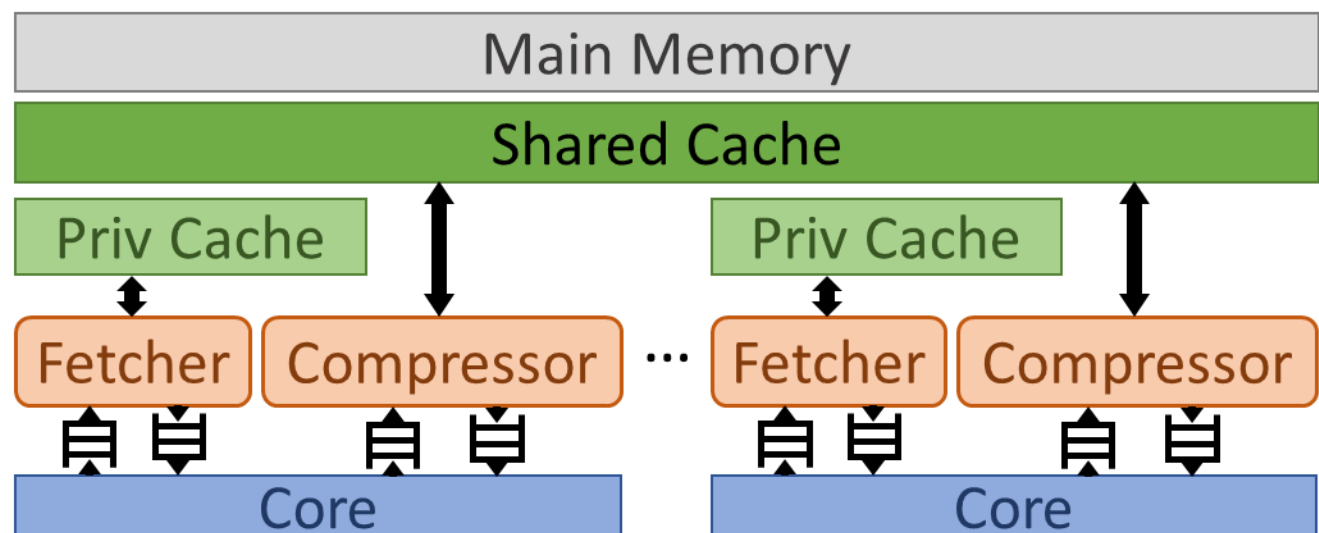
Must rethink full system stack to support irregular computations efficiently

Update Batching (UB)

- Maximizes spatial locality of memory transfers using two-phase execution
- **Binning phase:** Logs updates to memory, dividing them into cache-fitting slices (bins) of vertices
- **Accumulation phase:** Reads and applies logged updates bin-by-bin

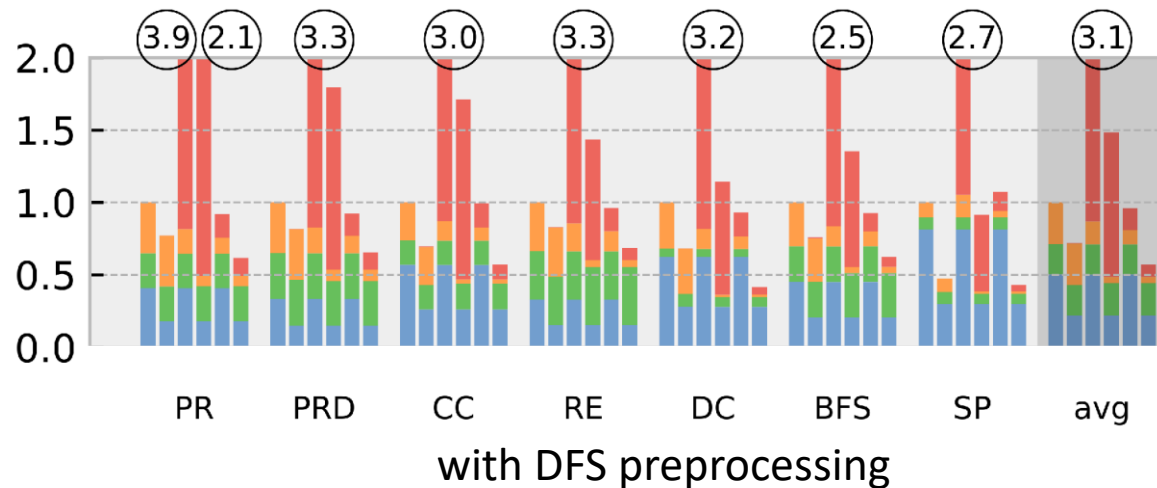
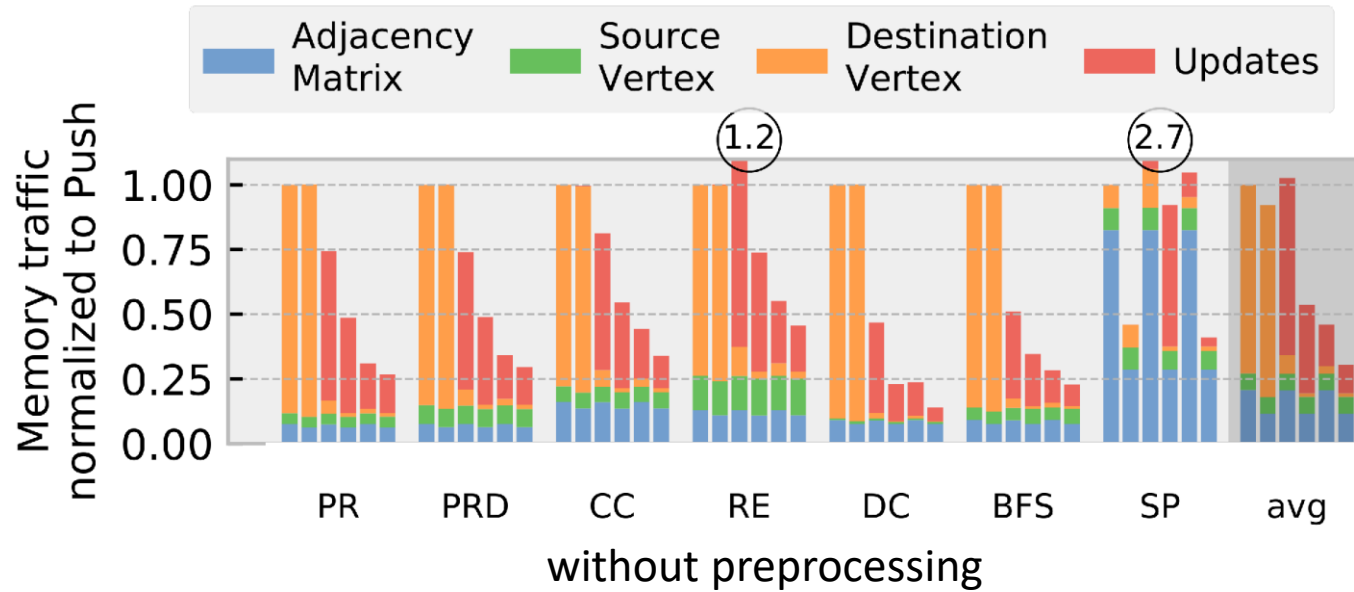


SpZip



- + *SpZip fetcher* accelerates data structure traversal and decompression
- + *SpZip compressor* compresses newly generated data before storing it off-chip
- + Fetcher and compressor execute a configurable dataflow graph of logical operators
 - Handle multiple/complex data structures by composing simple operators
 - Provide general support for graphs and sparse tensors (but trees, hash tables would require more operators)
 - Can be used in the context of a CPU or a specialized architecture

Memory Traffic Reduction



- + UB+SpZip reduces memory traffic
 - 3.3x without preprocessing
 - 1.8x with preprocessing

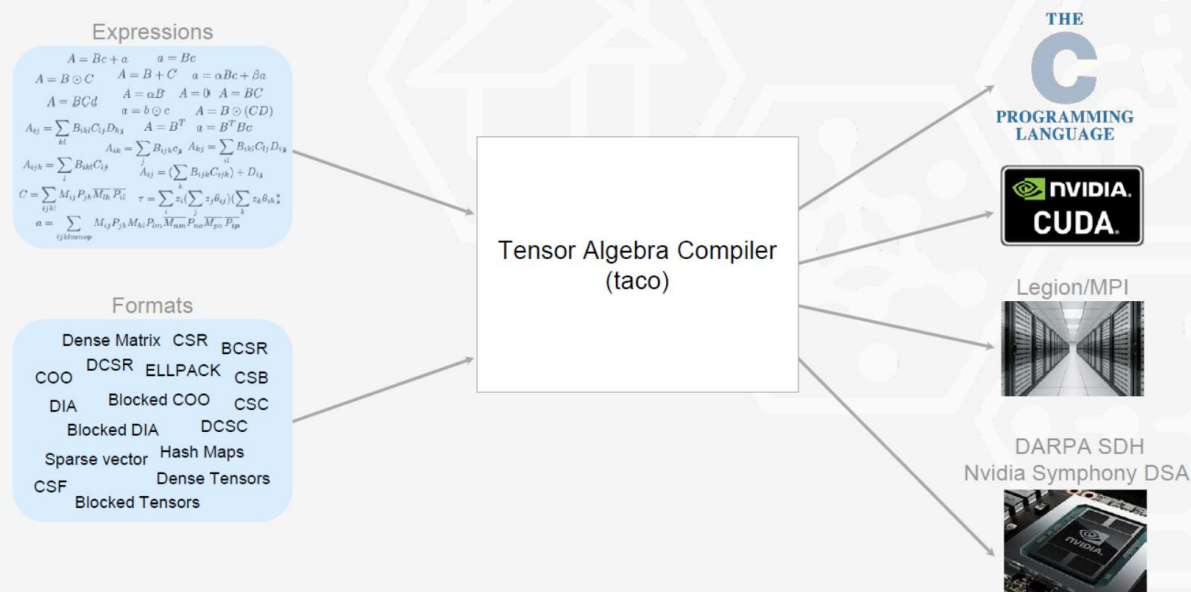
Tensor Algebra Compiler

(<http://tensor-compiler.org/>)

A Domain Specific Language, Compiler and Runtime

- Raising the level of abstraction to enhance programmer productivity
- Generates optimized parallel distributed sparse tensor linear algebra code
- Sparse tensors are the dominant form of tensor
- Other Prominent DSL's: MLIR, Halide, GraphIT, TVM

THE TENSOR ALGEBRA COMPILER (TACO)



TENSORS ARE EVERYWHERE

Data Analytics

Movies

Social Networks

Product Reviews

Machine Learning

Sparse Networks

Sparse Convolutional Networks

TensorFlow

Graph Convolutional Network

PyTorch

Science and Engineering

Robotics

Simulations

Computational Biology

Amazon Product Reviews

Customers

Words

Monitor Swallow Laptop Candidate Jacket The filled Kindle

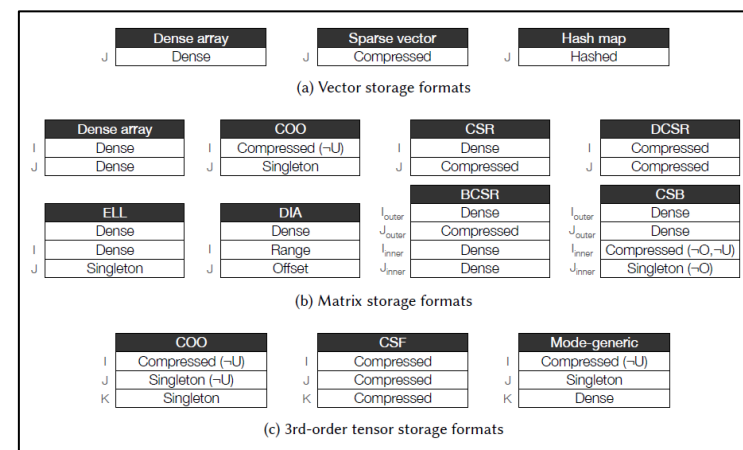
Extremely sparse

Dense storage: 107 Exabytes

Sparse storage: 13 Gigabytes

Images from the

Supports all widely used sparse tensor formats



Initial Characterization of ECP Applications

App	Kernel Time	Kernel Type	Parallelism	Sparse/Dense	Primary API	Limiting factor
AMG2013	72%	CSR SpMV	High	sparse	TACO	Memory Bandwidth
ExaMiniMD	76%	Euclid Distance	High	sparse/graph	GraphIt	Not vectorized, poor branch prediction
Laghos	50%	Tensor contractions	Medium	sparse/graph	TACO	CPU vector unit, MPI comm
miniAMR	88%	7-pt stencil	High	dense with multigrid	CoLa	L3 latency bound, unnecessary indirection
miniQMC	78%	Spline interpolation, DGEMM	High	dense	Halide/Tiramisu	DRAM bandwidth bound, FMA bound
miniVite	??	??	Low	sparse/graph	GraphIt	Memory latency, serial sections, mallocs
nekbone	70%	DGEMM, daxpy	High	dense	Halide/Tiramisu	Memory Bandwidth
PICSARLite	??	dense stencil	Low	dense	Halide/Tiramisu	Thread spawninig
SW4lite	90%	dense stencil	High	dense	Halide/Tiramisu	CPU bound needs better vectorization!
SWFFT	90%	copying/MPI	High	dense	Halide/Tiramisu	Memory Bandwidth, Network Bandwidth
FFTW	95%	butterflies	High	regular but sparse	Halide/Tiramisu	Memory Bandwidth
XS Bench	95%	particle lookup/update	High	sparse/hash	GraphIt	Memory Latency

Common Kernels mapped to APIs:

Sparse Tensor

→ TACO

Graphs

→ GraphIt

Dense Stencils/Tensors

→ Halide/Tiramisu

Multigrid

→ CoLa

Key Limiting Software Factors:

- serial sections, thread overhead, poor vectorization

Key Limiting Hardware Factors:

- CPU vector unit
- Branch Prediction
- Memory Bandwidth & Memory Latency
- Network Communication

arm

Backup