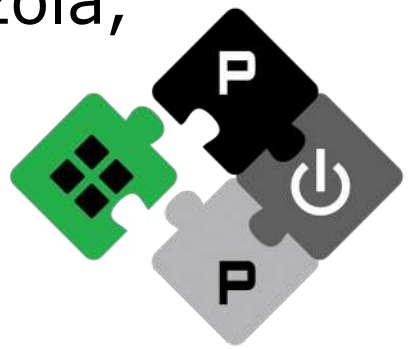
A 410GFLOP/s, 64 RISC-V Cores, 204.8GBps Shared-Memory Cluster in 12nm FinFET with Systolic Execution Support for Efficient B5G/6G AI-Enhanced O-RAN

Yichao Zhang, Marco Bertuletti, Sergio Mazzola, Samuel Riedel and Luca Benini

IIS, ETH Zurich, Switzerland
University of Bologna, Italy
yiczhang@iis.ee.ethz.ch



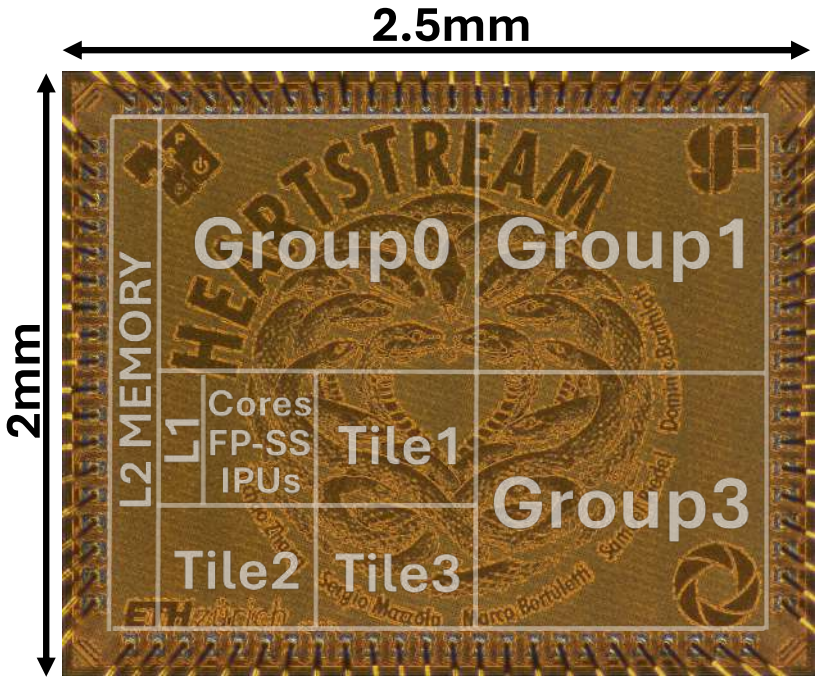
Outline

- “HeartStream” Introduction
- Background and Design Target
- Computing Cluster Design
- Performance and Energy Efficiency
- Conclusion

Outline

- "HeartStream" Introduction
- Background and Design Target
- Computing Cluster Design
- Performance and Energy Efficiency
- Conclusion

”HeartStream” SoC Overview



Technology	12nm FinFet
Die Area	5mm^2
Core Supply*	0.8V
Frequency*	800MHz
On-Chip Mem	512KiB
IO BW*	6.4Gbps
Power*	1.15W

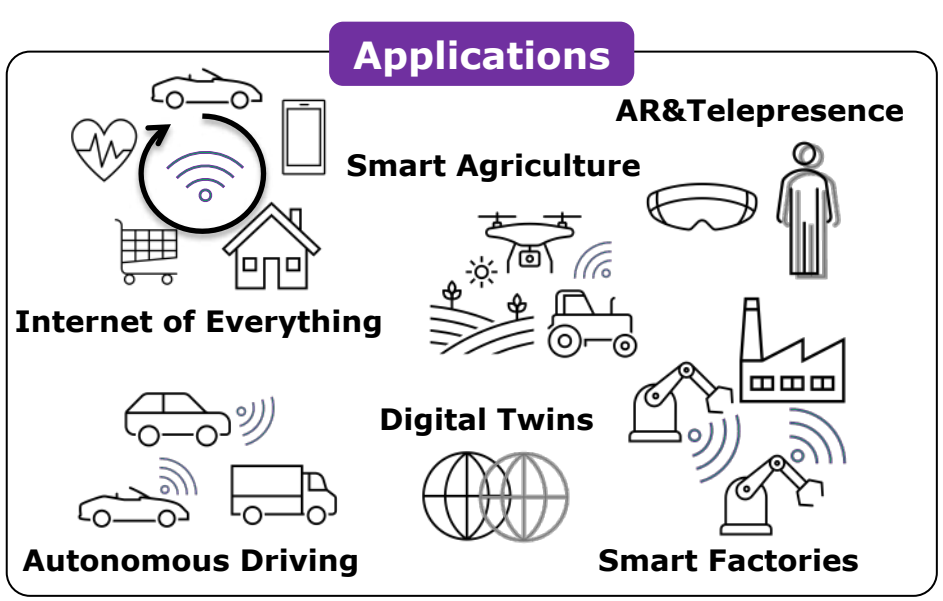
* at 0.8V Core Supply

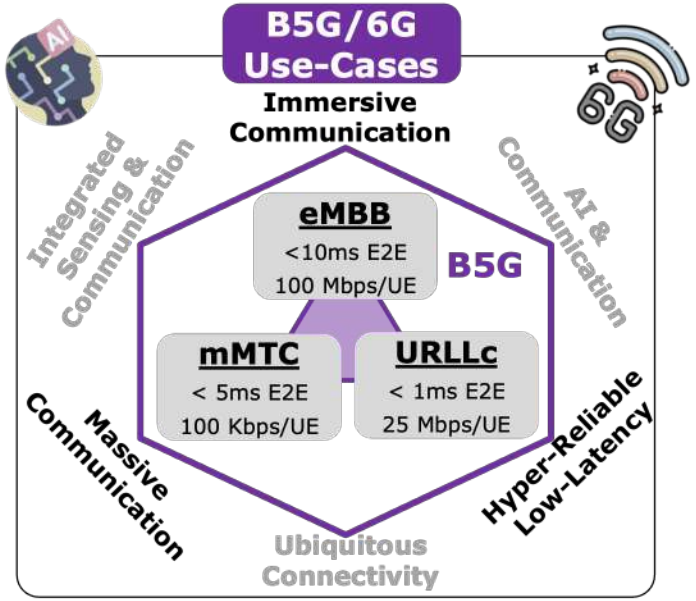
- An efficient **64-core** computing cluster for B5G/6G O-RAN applications:
  - Latency-tolerant PEs, **fully shared L1-memory**
  - **Complex arithmetic** support
  - Efficient **systolic execution**
- Allowing efficient B5G/6G O-RAN applications processing + AI/ML workloads

Outline

- "HeartStream" Introduction
- Background and Design Target
- Computing Cluster Design
- Performance and Energy Efficiency
- Conclusion

6G Open-Radio Access Network





Enhanced **M**obile **B**road**B**and

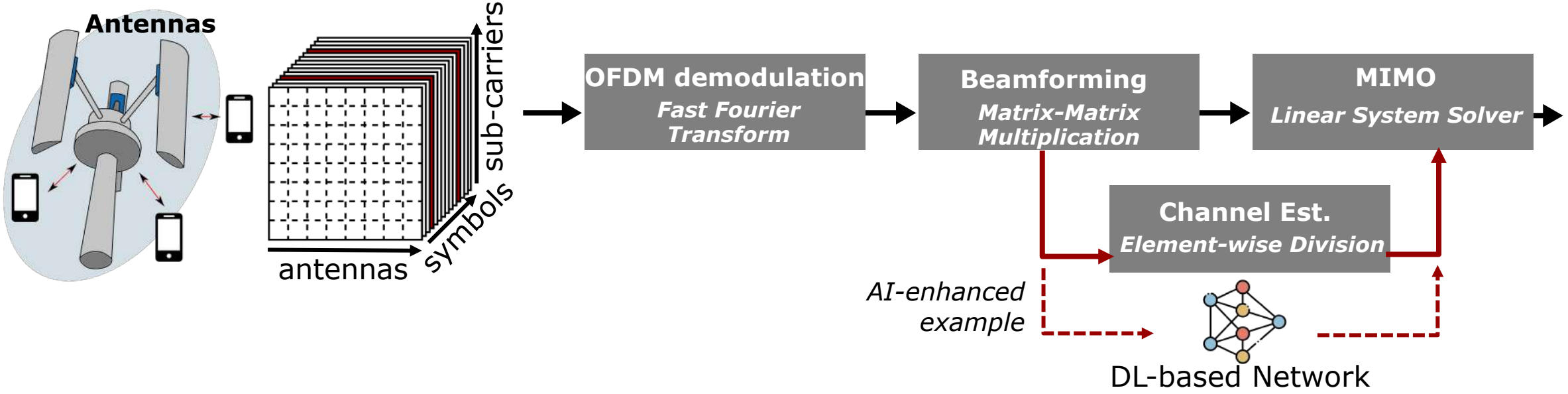
Massive **M**achine **T**ype **C**ommunication

Ultra-**R**eliable **L**ow-**L**atency **C**ommunication

- B5G/6G high-end processing: different scenarios, various requirements
  - < 4ms user-plane latency*
  - 5-20Gbps uplink throughput*
  - AI-enhanced pre/post-processing*

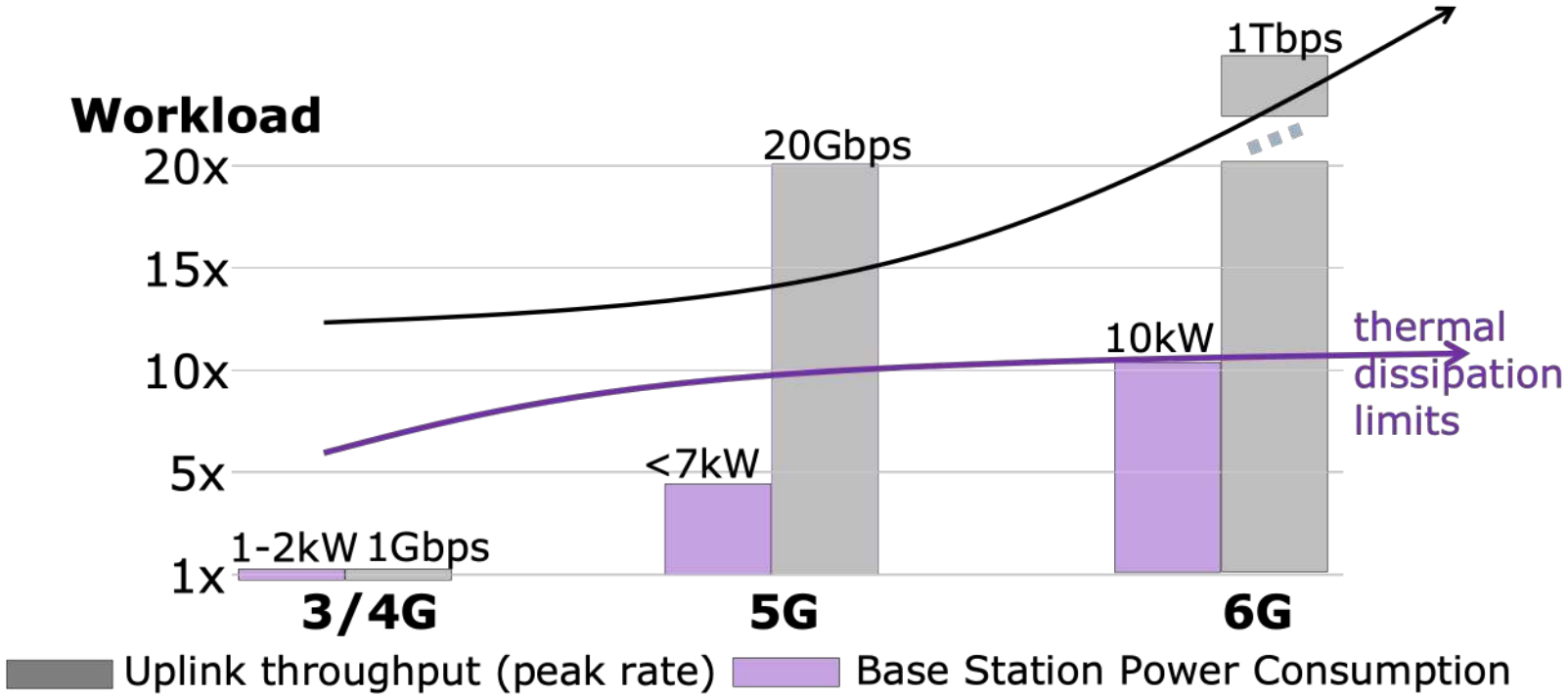
*Md. E. Haque et al., IEEE Access, Apr. 2023, pp. 34372-34396

Physical Uplink Shared CHannel



- One of the most **computing-intensive** and **time-critical** channels
- "Many-dimensions" signals, SW-defined O-RAN (**time-to-market** ↓)
- **Heterogeneous** workloads at the edge (base station)
  - FFT, MatMul, MIMO, Channel Estimation
  - AI-enhanced processing → more computing

Energy Efficiency Is the Key



- 90% of operational power for amplifiers, supply, and air-conditioning*
- 10% (100W) budget for analog and digital signal processing*
- Assume 10W per BS component, **>2Gbps/W is need for B5G/6G uplink**

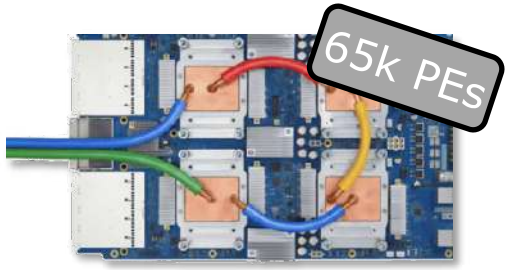
*C.-L et al., Nat. Electron., Apr. 2020, pp. 182-184 *Ericsson, Ericsson Mobility Report, 2021&2024

Goals For HeartStream

- Efficient baseband physical uplink processing:
  - < 4ms user-plane latency, 5-20Gbps throughput, >2Gbps/W

- Cluster of processors architecture:
  - Large workload. → **Scalable Architecture**
  - Low-latency. → **Manycore parallel processing**
  - Limited power budget → **Energy-efficient design**
  - Heterogeneous workloads. → **Efficient ISA extensions**
  - Flexibility, time-to-market. → **Programmable**
  - Community-developed solution. → **Open source**

Our Target: General and Scalable

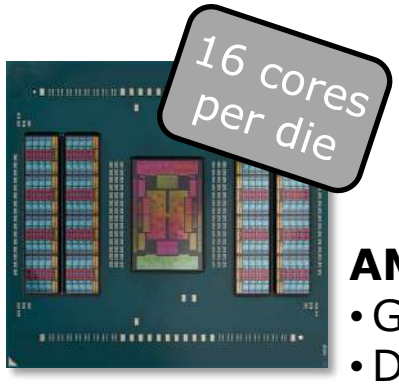


https://cloud.google.com/tpu

Google TPU
- Custom accelerator
- Very specialized

https://developer.nvidia.com/blog/nvidia-hopper-architecture-in-depth/

NVIDIA Hopper GPU
- SIMT
- Complex mem hierarchy
- Flexible



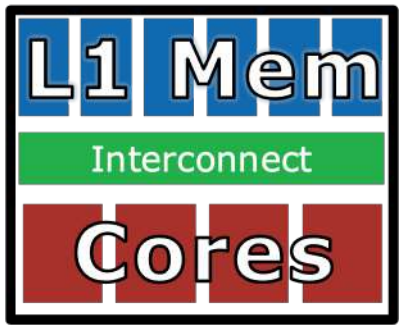
AMD EPYC CPU
- General-purpose
- Doesn't scale

https://www.amd.com/en/products/processors/server/epyc/9005-series.html

scalable

General Processing

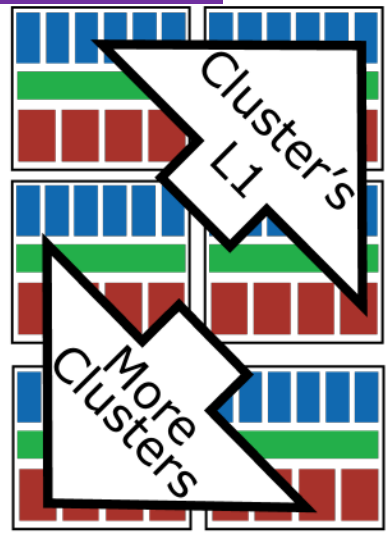
Scaling the computing cluster



Small Compute Cluster

Scale-out

"more"



More Compute Clusters

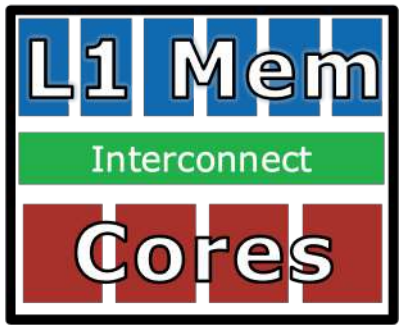
Process Elements

Interconnects

Memories

- "Traditional" way to make more PEs + larger memory footprint

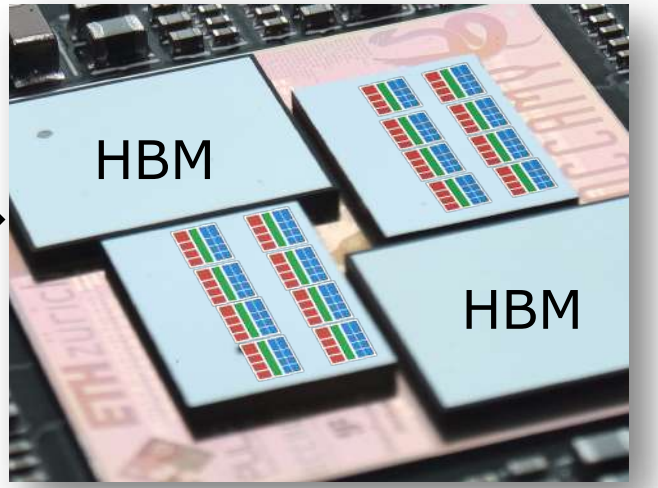
Scaling the computing cluster



Small Compute Cluster

Scale-out

"more"



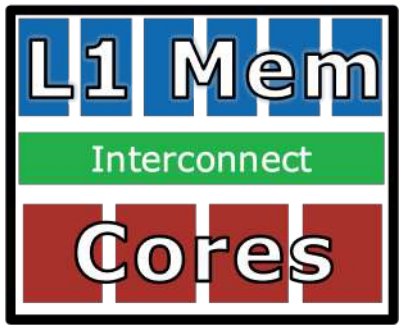
Occamy*: 432-core in 48 clusters

Process Elements
Interconnects
Memories

- "Traditional" way to make more PEs + larger memory footprint
- Scale-out to multi/many loosely coupled clusters:
  - Easy to scale (many small, identical processor clusters)

*P. Scheffler et al., IEEE JSSC, Jan. 2025, pp. 1324-1338

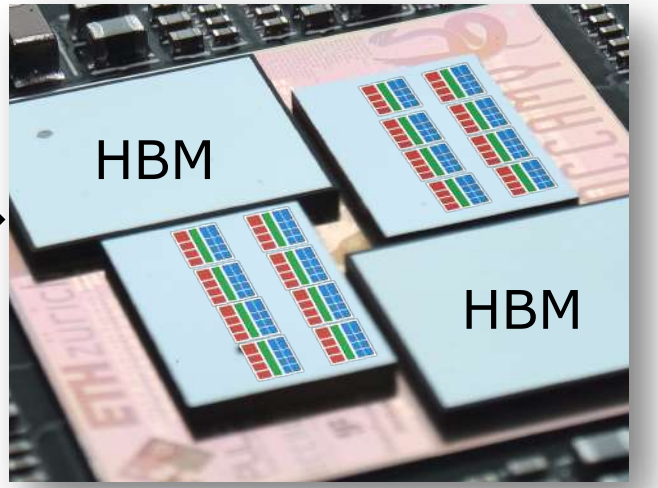
Scaling the computing cluster



Small Compute Cluster

Scale-out

"more"

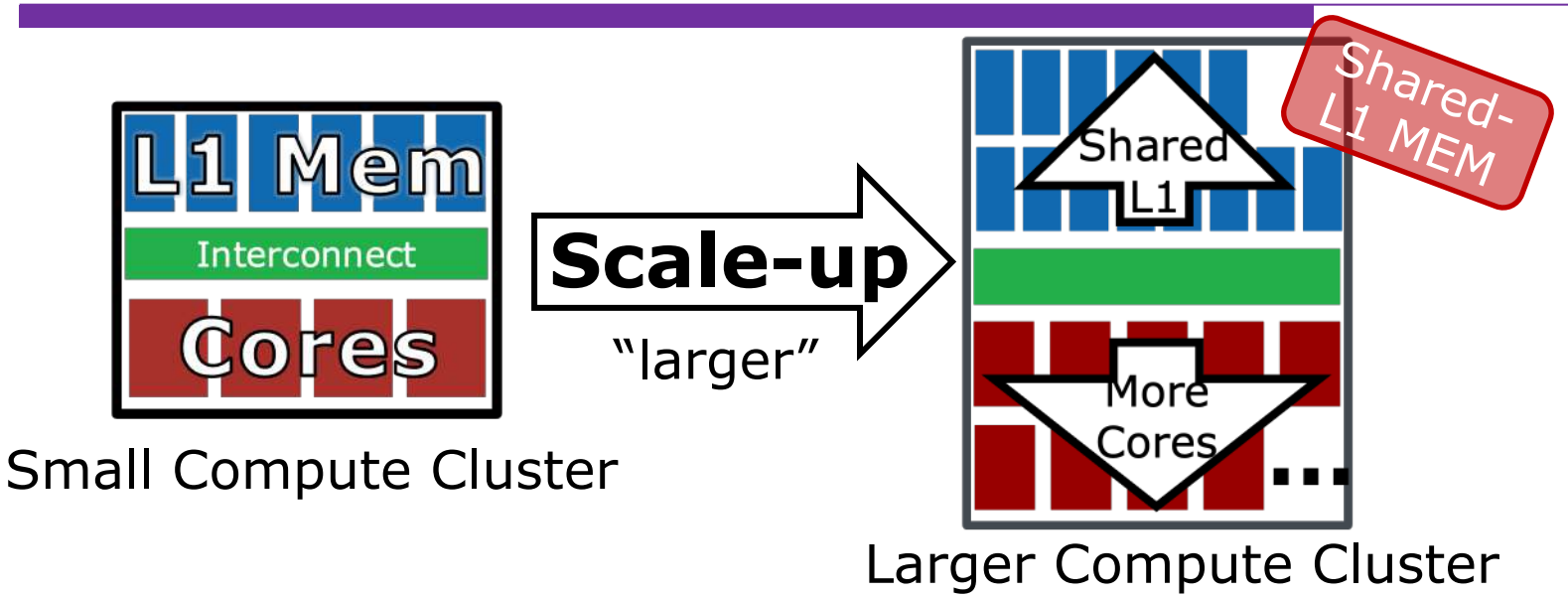


Occamy*: 432-core in 48 clusters

Process Elements
Interconnects
Memories

- "Traditional" way to make more PEs + larger memory footprint
- Scale-out to multi/many loosely coupled clusters:
  - Easy to scale (many small, identical processor clusters)
  - HW&SW overheads:
    - inter-cluster **communication** and **synchronization**
    - Data **allocation-splitting**, workload **distribution**
    - **High-latency** global interconnect to main memory, latency tolerant by limited size of L1

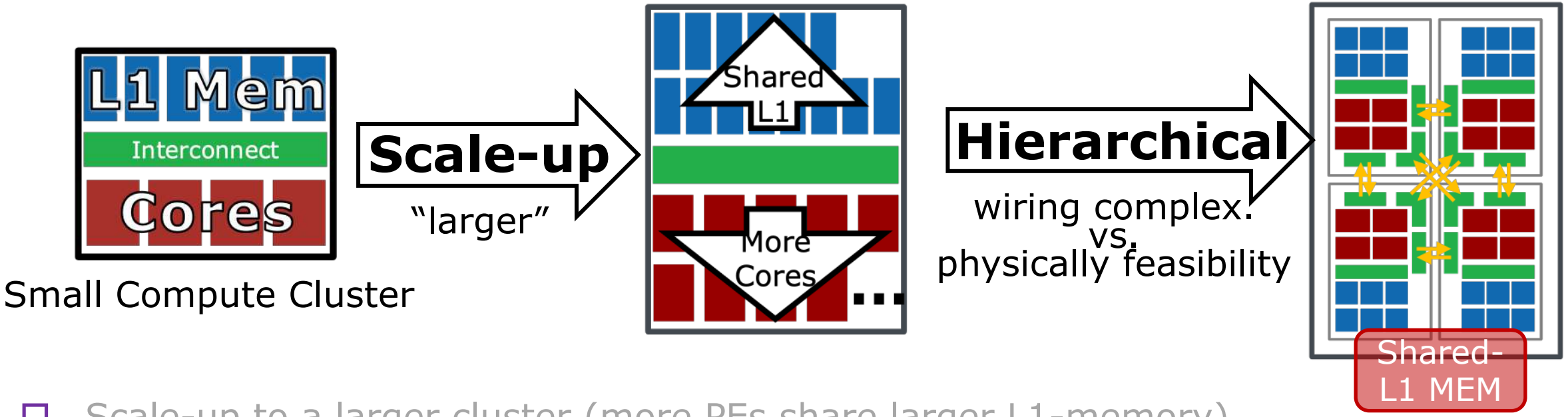
Scaling the computing cluster



- Scale-up to a larger cluster (more PEs share larger L1-memory)
  - reduces data chunks split/transfer/merge* ✓
  - High compute-to-traffic ratio* ✓
  - Easy to program* ✓
  - But… is interconnect implementable? ?
  - Synchronization between cores? ?

*S. Riedel et al., IWASI'25., July 2025, pp. 1-6

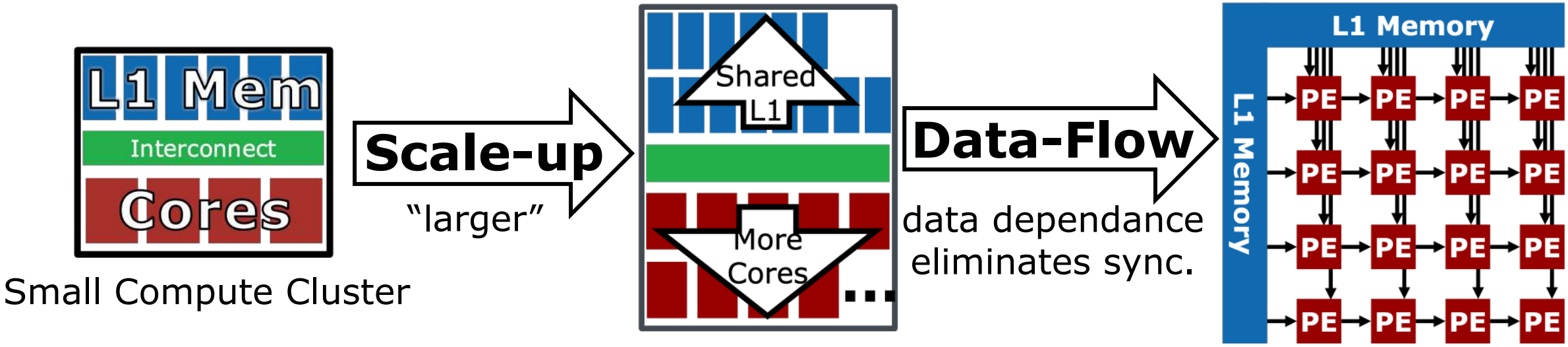
Scaling the computing cluster



- Scale-up to a larger cluster (more PEs share larger L1-memory)
  - reduces data chunks split/transfer/merge*
  - High compute-to-traffic ratio*
  - Easy to program*
  - **But… is interconnect implementable?**
  - Synchronization between cores?

*S. Riedel et al., IWASI'25., July 2025, pp. 1-6

Scaling the computing cluster



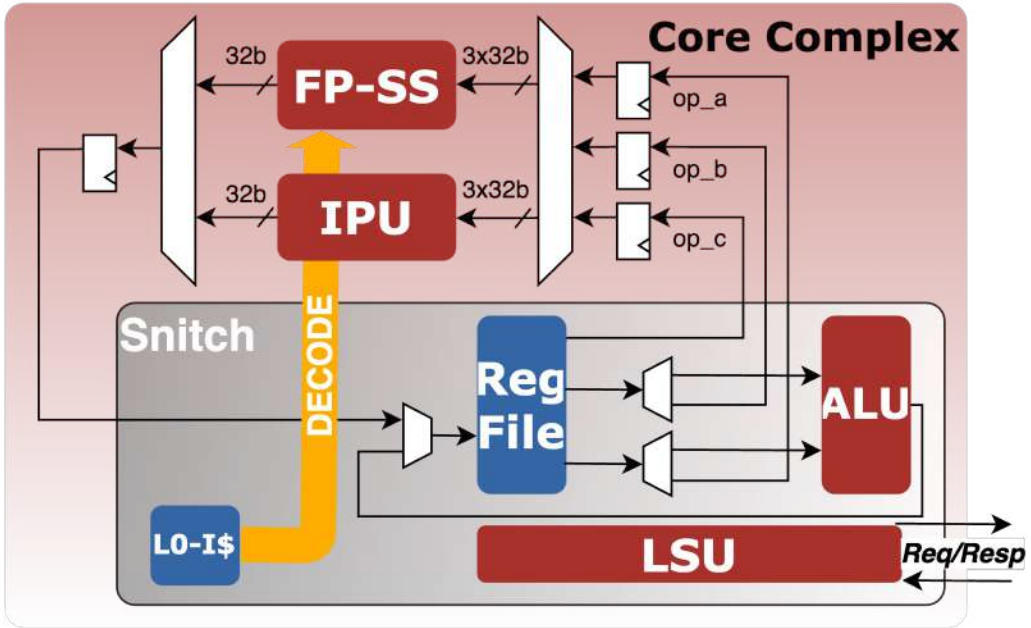
- Scale-up to a larger cluster (more PEs share larger L1-memory)
  - reduces data chunks split/transfer/merge*
  - High compute-to-traffic ratio*
  - Easy to program*
  - But… is it efficiently implementable?
  - **Synchronization between cores?**

*S. Riedel et al., IWASI'25., July 2025, pp. 1-6

Outline

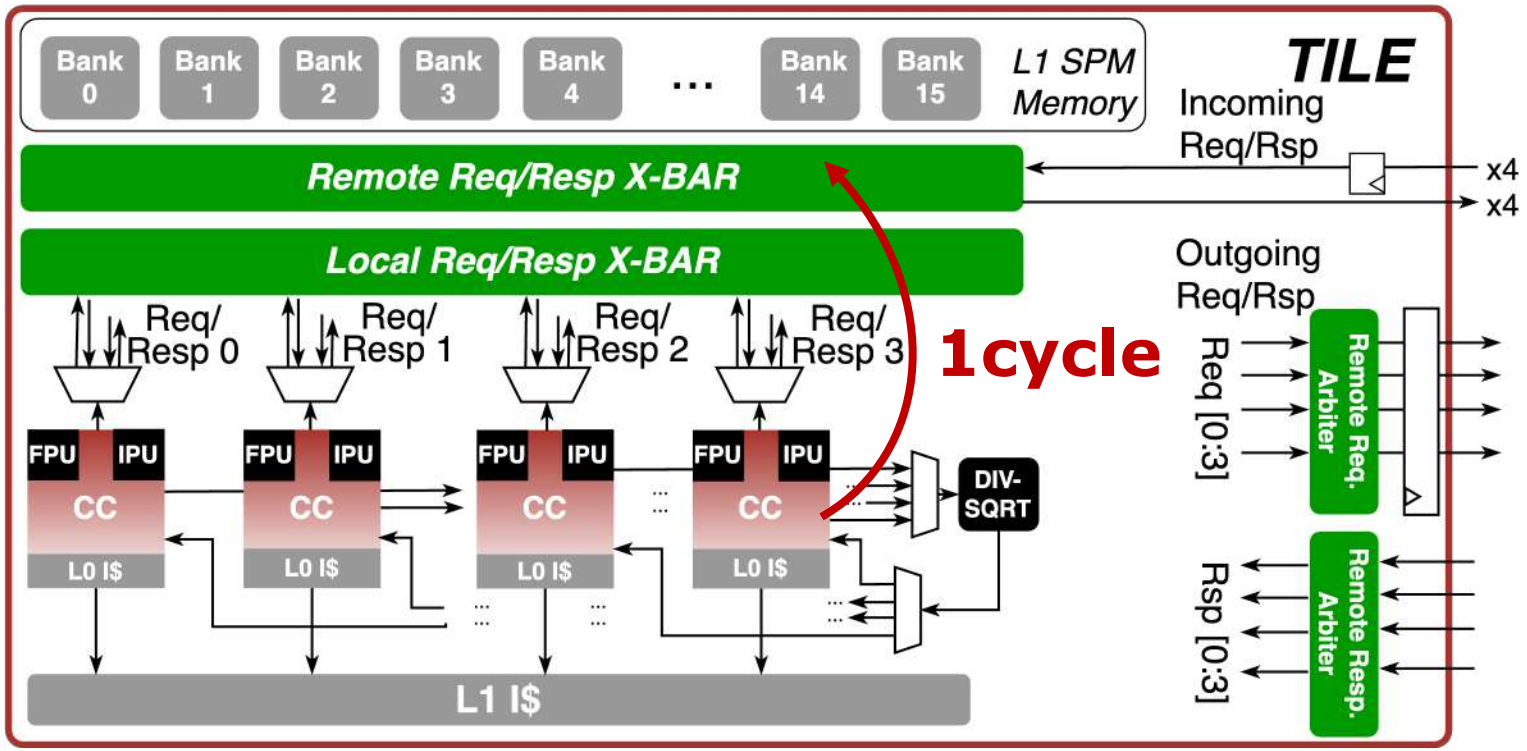
- "HeartStream" Introduction
- Background and Design Target
- Computing Cluster Design
- Performance and Energy Efficiency
- Conclusion

Processing Element: Core Complex



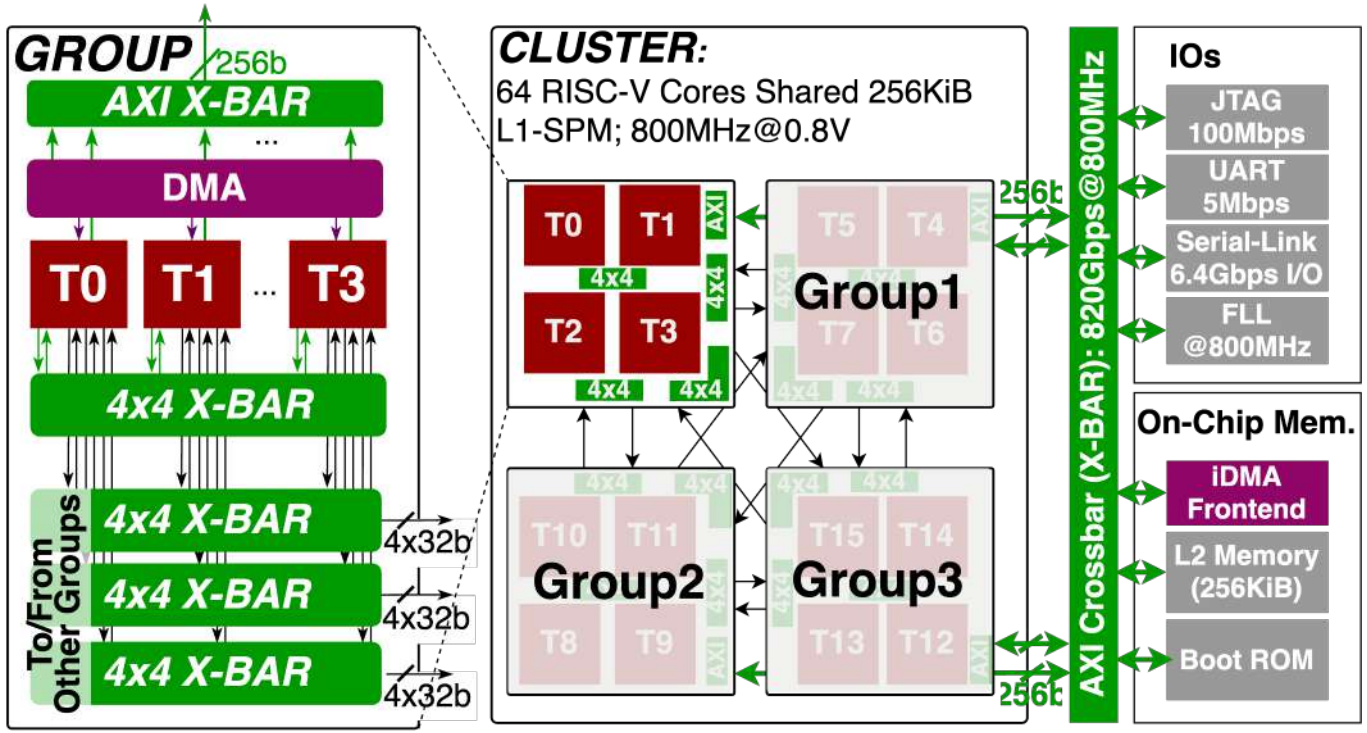
- Snitch core:
  - Single-issue, single pipeline stage
  - Latency-tolerant, many outstanding transactions supported (8 in this instance)
- Core Complex (RV32IMDFA)
  - Core + Integer processing unit (IPU)+FP Sub-System (FP-SS)
  - **Complex Arithmetic** (16b Real & Imaginary)+**Tailored ISA extensions** for baseband & AI

Tile: Building Block



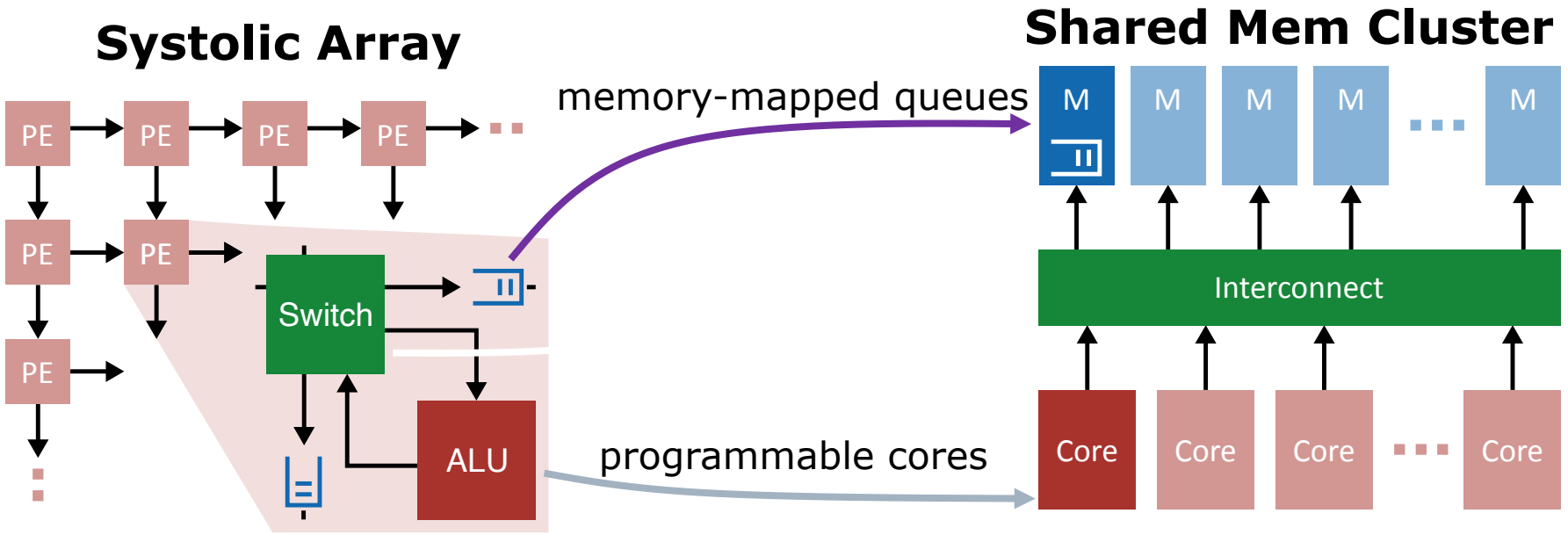
- Tile = Core + interconnect + Memory + remote interface
- Fully combinational log crossbar → **1 cycle access latency** in the Tile
- Registers only at the boundary cut wire delay to the remote Tiles

Cluster: Hierarchical Interconnect



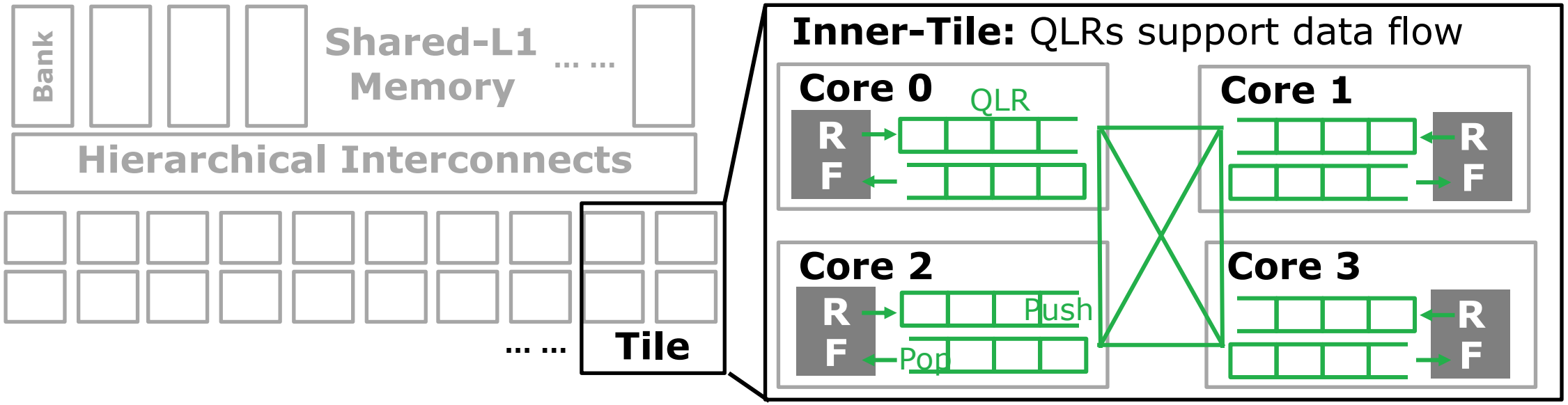
- Hierarchy:
  - 4CCs x 4 Tiles x 4 Groups = 64 CCs **fully shared-L1** (256 Banks) cluster
  - L1 latency: **1 cycle** in Tile, **3 cycles** in Group, **5 cycles** between Groups
  - Hierarchical interconnect → Scalable architecture

HW-Supported Systolic Data-Flow



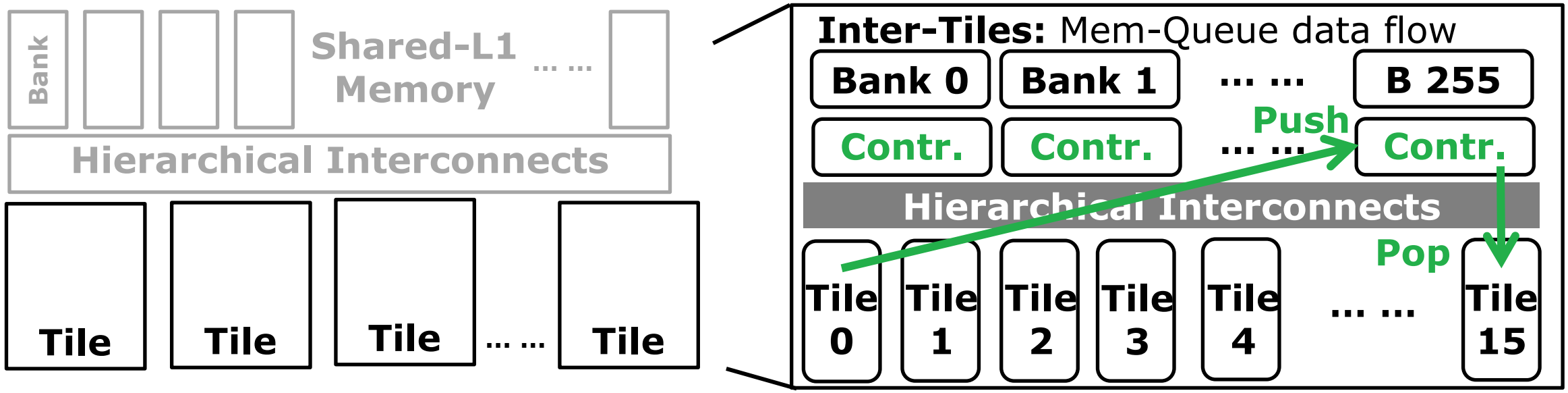
- 64 cores systolic data-flow: programmable, ISA extension
  - reduce memory control & access instructions
  - Data flow eliminates synchronization between cores

HW-Supported Systolic Data-Flow



- 64 cores systolic data-flow: programmable, ISA extension
  - reduce memory control & access instructions
  - Data flow eliminates synchronization between cores
- Queue-Linked Registers (QLRs): implicit inter-core RF R&W

HW-Supported Systolic Data-Flow

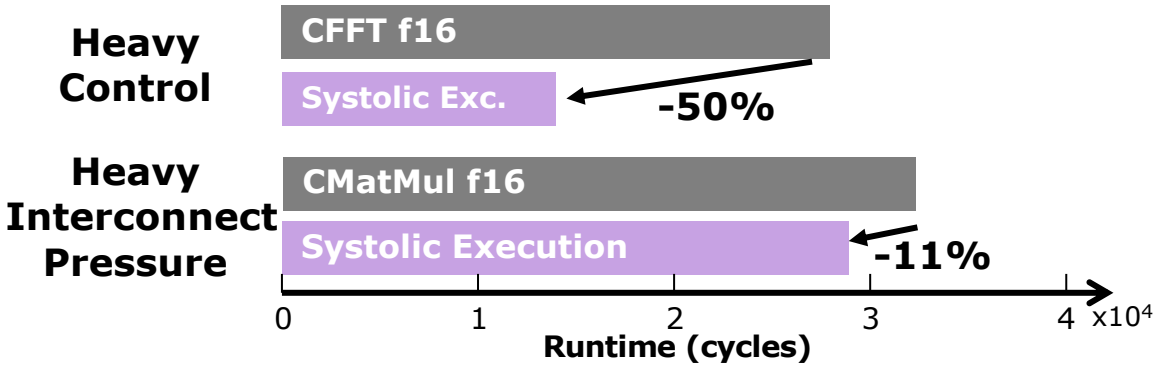


- 64 cores systolic data-flow: programmable, ISA extension
  - reduce memory control & access instructions
  - Data flow eliminates synchronization between cores
- Queue-Linked Registers (QLRs): implicit inter-core RF R&W
- Memory-Mapped Queues: Routed by the cluster interconnects

Outline

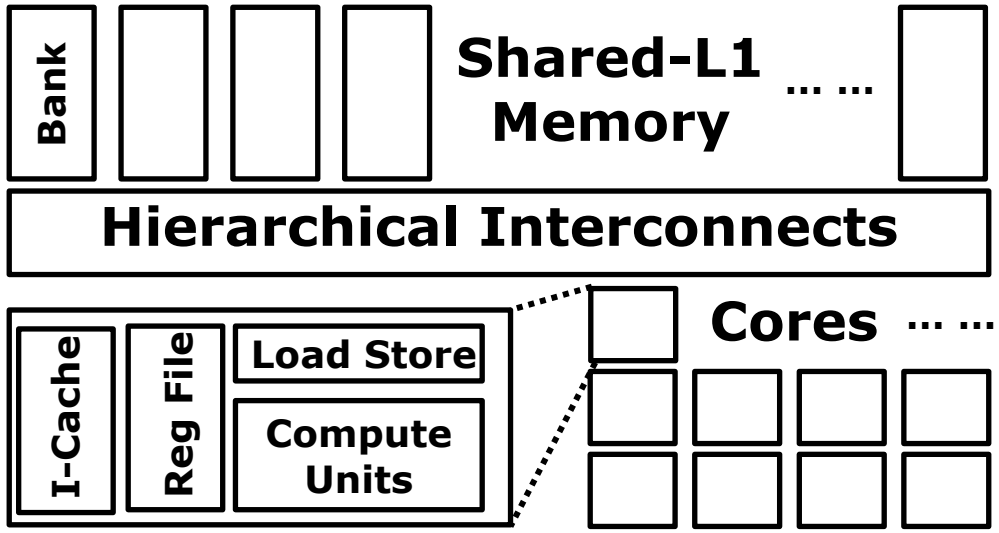
- "HeartStream" Introduction
- Background and Design Target
- Computing Cluster Design
- Performance and Energy Efficiency
- Conclusion

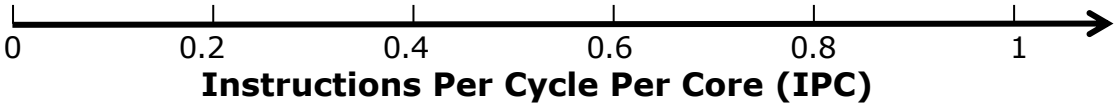
Systolic Boosts Performance



- **Heavy-control kernels** (e.g., Demodulation (Complex FFT))
- **Heavy-interconnect-pressure** kernels(e.g., Beam Forming (Complex MatMul))
- The total instruction count decreases
  - Reducing data movement control
  - Eliminate synchronization between computing stages

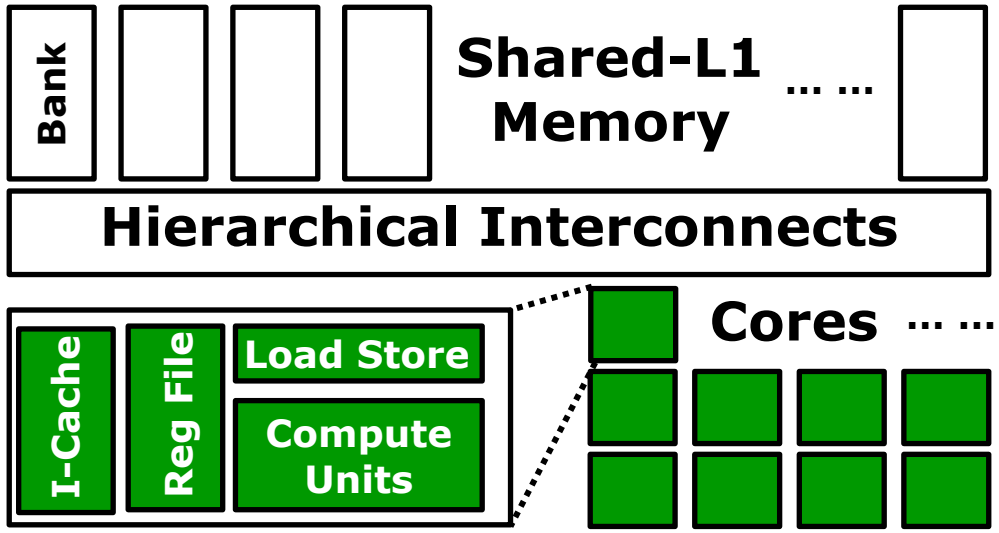
Core Utilization

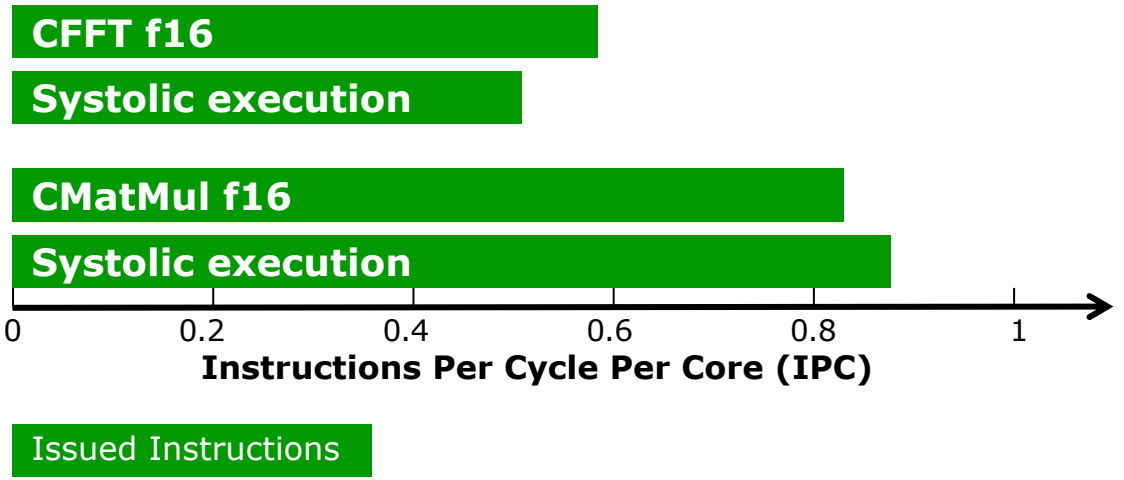




- Analysis from different parts of the design, mainly:
  - Cores: control + compute
  - Interconnects
  - L1 Memory

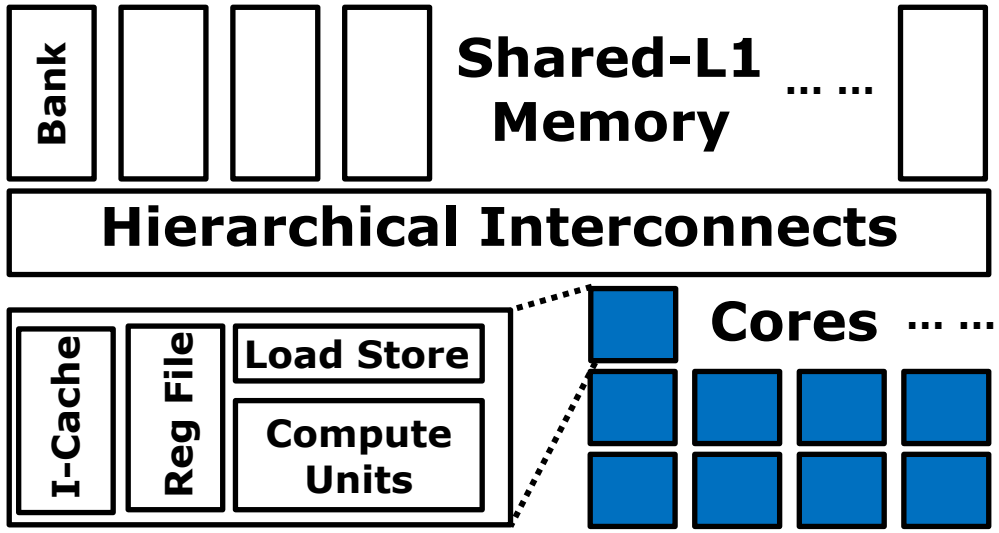
Core Utilization

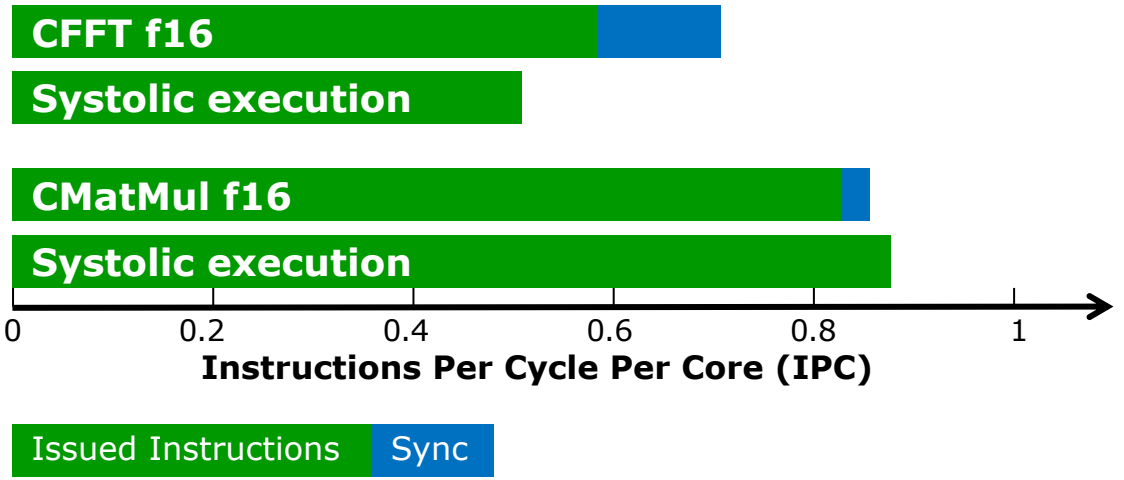




- Heavy-control kernels (e.g. Demodulation(CFFT))
  - 0.6 high IPC with 64-core parallelization
- Heavy-interconnect-pressure kernels (e.g. BeamForming(CMatMul))
  - 0.9 high IPC for compute-intensive execution

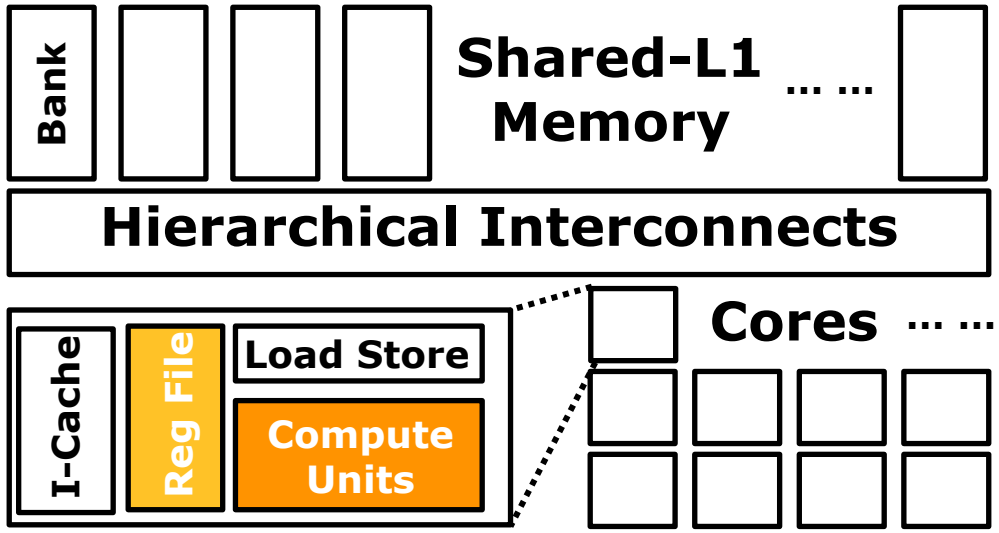
Core Utilization

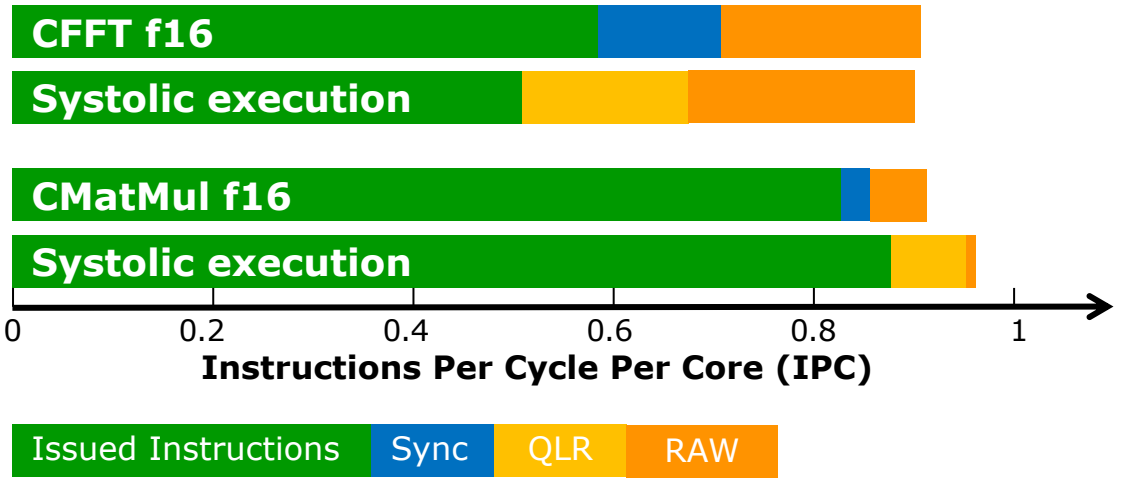




- Heavy-control kernels
  - Synchronization between computing stages
- Heavy-interconnect-pressure kernels
  - Synchronization at the end of computing
- **Systolic eliminates synchronization** by data flow

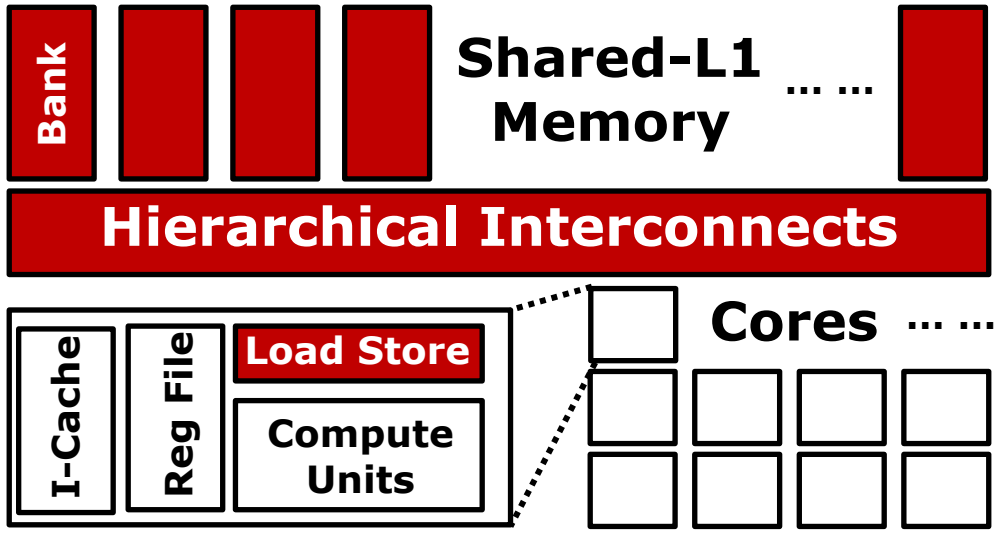
Core Utilization

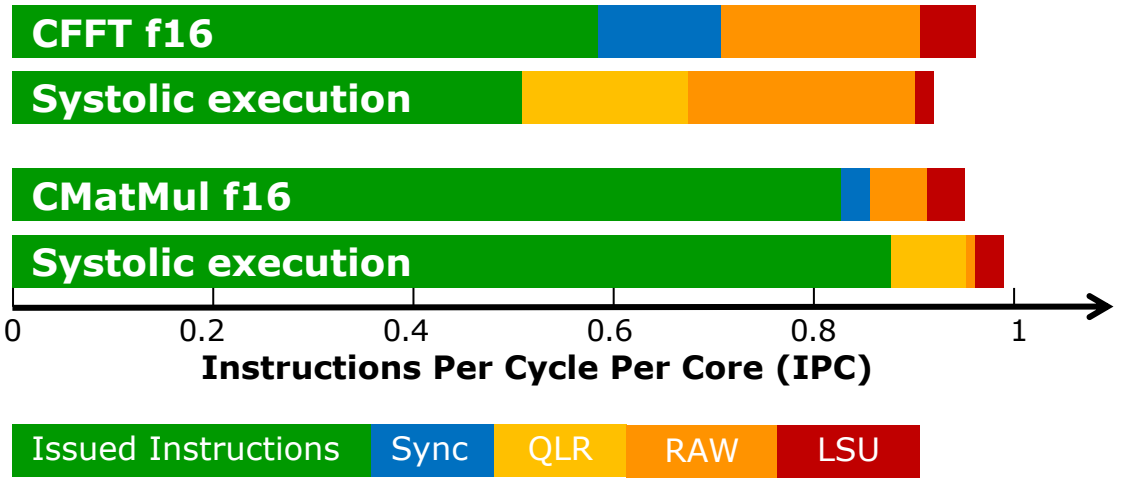




- Data dependency stalls
  - "Queue Linked Registers": wait for operands ready for the next PE execution
  - "Read After Write": wait for operands ready in pipelined compute units
- Latency-hiding is difficult for heavy-control kernels
  - Heavy-control = more register usage
  - Outstanding requests are limited by the number of registers (32) in the ISA

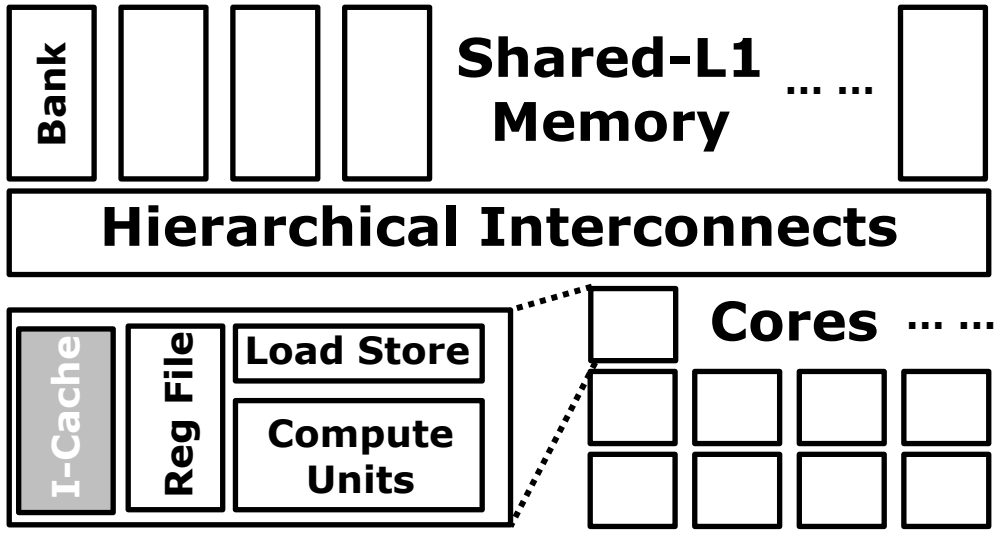
Core Utilization

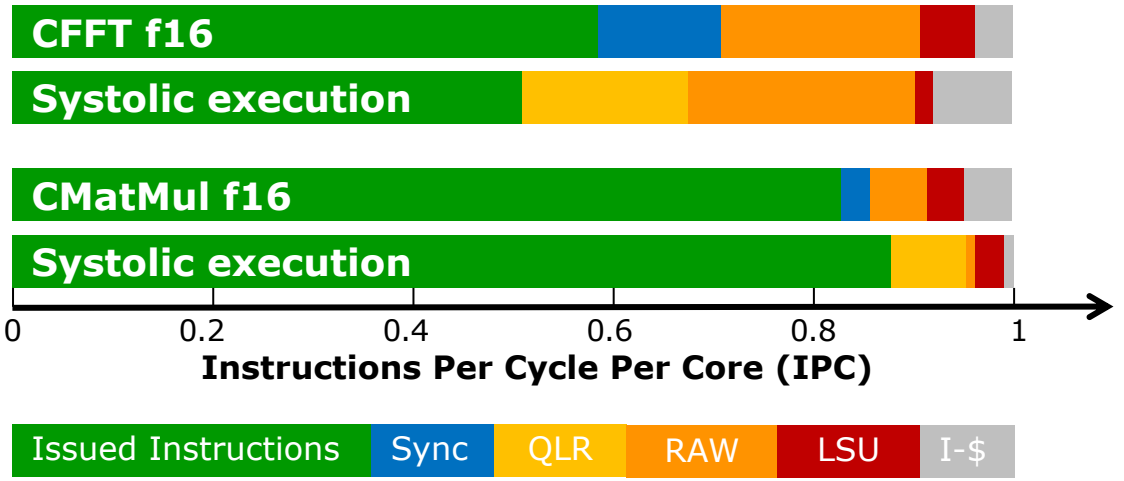




- Our cluster has 64 cores fully shared 256 L1 banks
- Super low PE-to-L1-bank interconnects stalls **(<5%)**
  - **204.8GBps** PE-to-L1 interconnect, **max. 5 cycles** latency.
  - Fully interleaved data reduce bank conflict

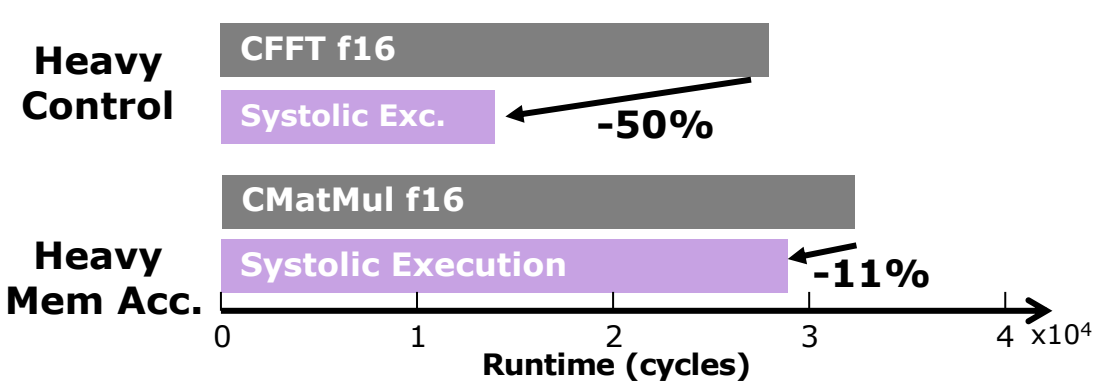
Core Utilization

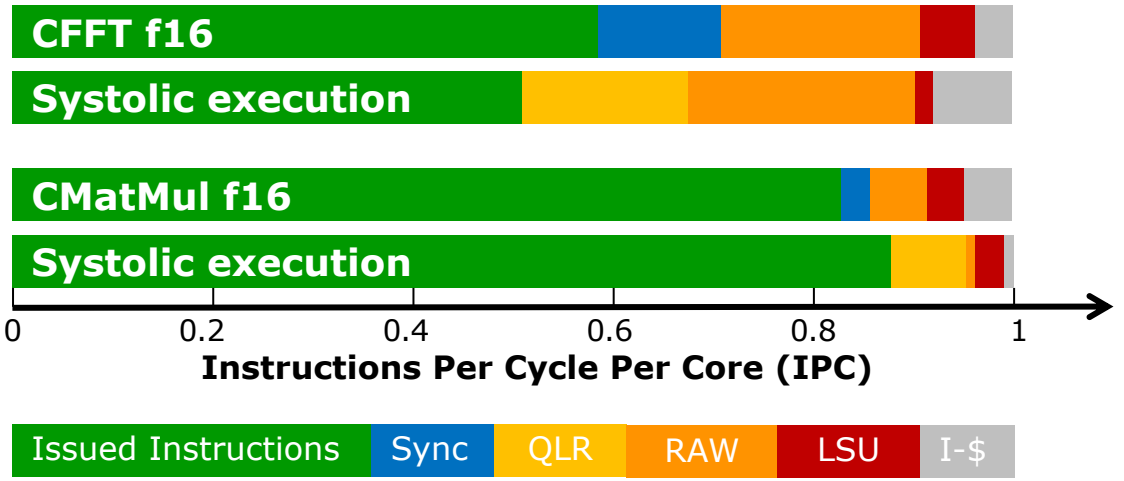




- Instruction refilling
  - 2KiB instruction cache per Tile, shared by 4 cores
  - Refilling from L2 memory
  - Through 820GBps system interconnect (256b width AXI/Tile)

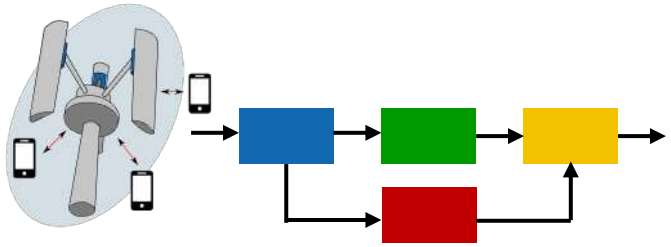
Performance Summary



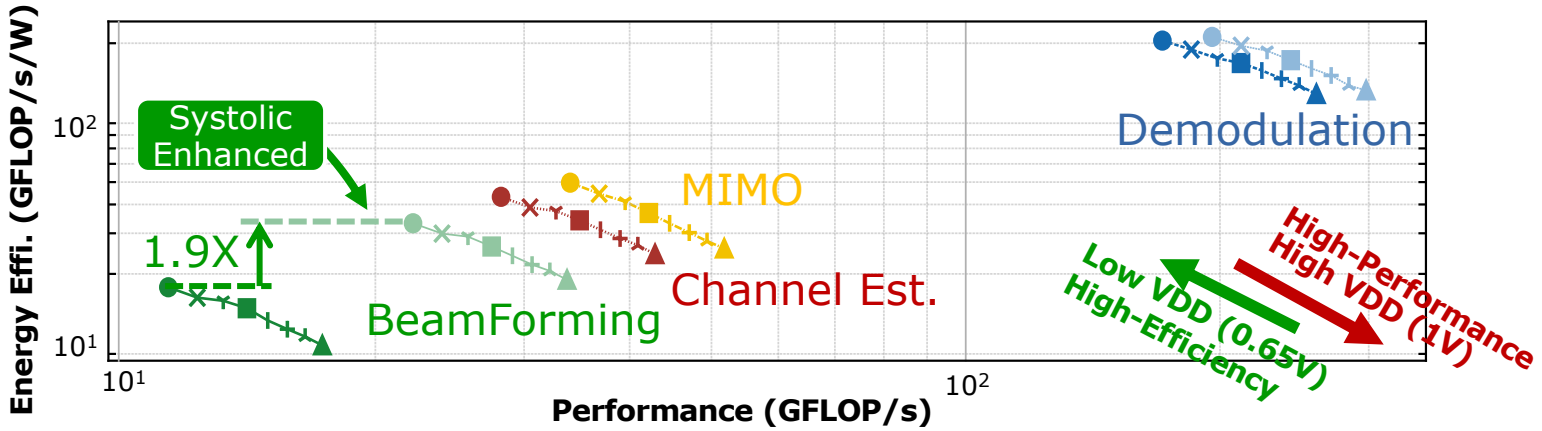


- **Heavy-control kernels** (e.g., CFFT):
  - Systolic boost 50% runtime (1.95x compute utilization)
  - Eliminate memory access control and inter-stage synchronizations.
- **Heavy-interconnect-pressure kernels** (e.g., CMatMul):
  - Negligible interconnect stalls
  - 204.8GBps PE-to-L1 interconnect, max. 5 cycles latency, fully interleaved data

Energy Efficiency of Key Kernels

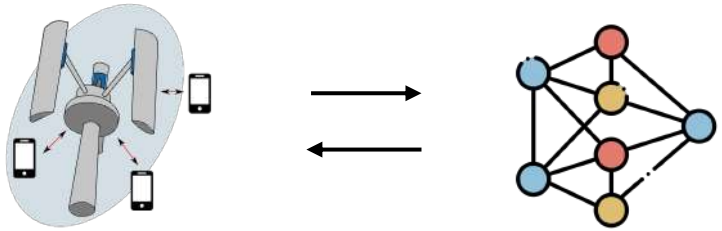


Physical Uplink Kernels:
Complex-f16, data parallel

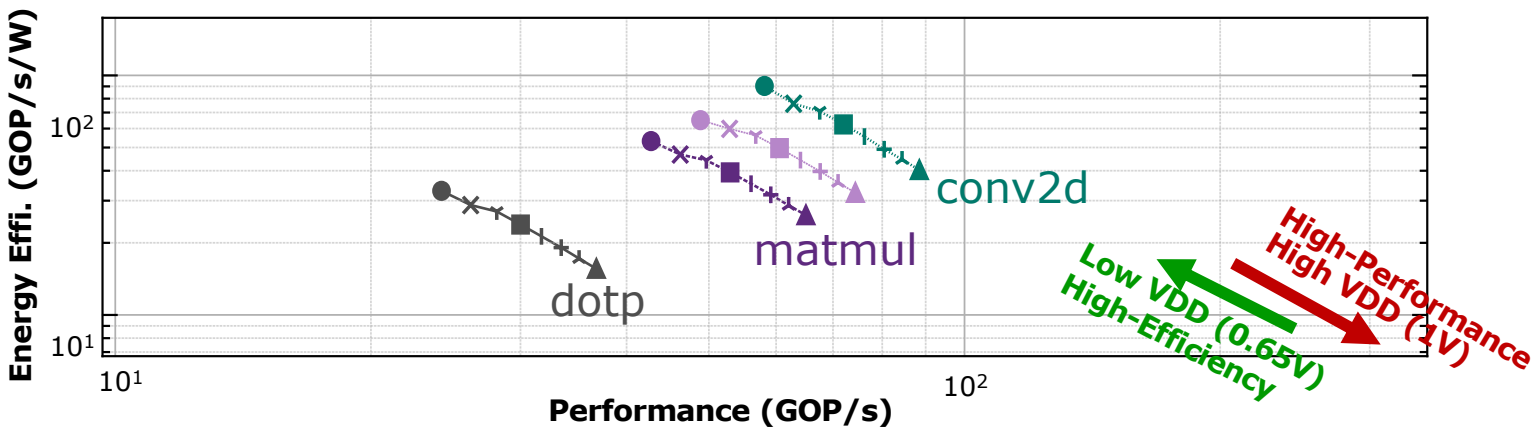


- Low-V(0.65V) for light loads =high energy efficiency
- High-V (1V) for heavy loads = high performance
- Systolic Ext.: improve up to **1.9X energy efficiency**
- Wireless kernels: OFDM:33.2GFLOP/s/W; Beamforming: 213GFLOP/s/W

Energy Efficiency of Key Kernels

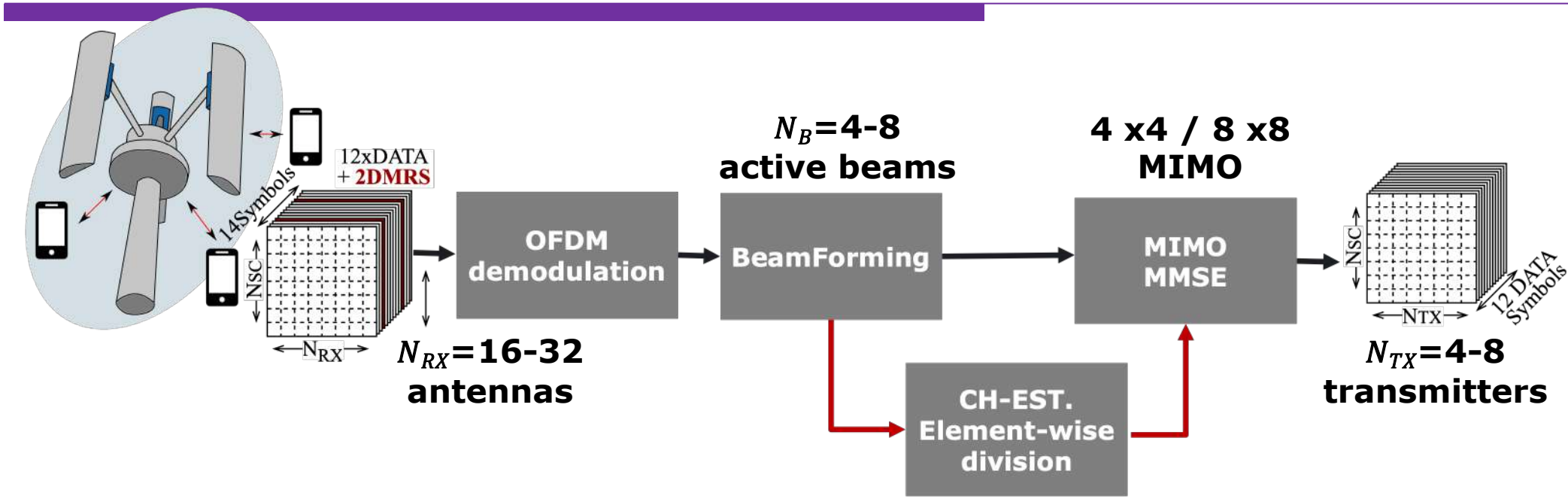


Generic DL Kernels:
Integer 32b, data-parallel



- Low-V(0.65V) for light loads =high energy efficiency
- High-V (1V) for heavy loads = high performance
- Systolic Ext.: improve up to **1.9X energy efficiency**
- Wireless kernels: OFDM:33.2GFLOP/s/W; Beamforming: 213GFLOP/s/W
- High performance for Generic Deep Learning workloads: 37-89GOP/s

Physical Uplink Workload Evaluation



	Our work	Target
Throughput (Raw Antenna*)	9Gbps@0.8V	5-20Gbps
User Plane Latency	3.2ms@0.65V	<4ms
Energy Efficiency*	8.3Gbps/W (49.6 GFLOP/s/W)	>2Gbps/W

*Y. Zhang et al., ISSCC'24., Feb 2024, pp. 48-49

Baseband Processing SoA Design

	This Work	ESSERC'24	CoolChip'22	ISSCC'14	ESSCIRC'22	ISSCC'24	VLSI'20
Technology	12nm, 800MHz@0.8V	22nm 400MHz@0.8V	28nm 800MHz@0.9V	65nm 445MHz@1.2V	22nm 293MHz@0.8V	40nm 200MHz@1.1V	40nm 290MHz@1.1V
Processing Element	64 x RISC-V Cores	4 PEs (core, vector, accel.)	9 PEs (core, ASIPs, accel.)	20 PEs (core, vector, ASIPs)	16 PEs (ASIPs)	ASIC Accel.	ASIC Accel.
Data Precision	Int 32/16 FP 32/16/8 Cmplx support	-	Int32/-, -	Int 16/- FP 32/-	FP 32/16/8 bfloat 16	-	-
GP Program.	Fully ✓	Partial	Partial	Partial	Partial	No	No
Baseband Processing	OFDM Beamforming Channel Est. Equalization ✓ … …	Equalization	Equalization	Equalization, WiMAX	Equalization Decoding	Equalization Decoding Channel Est.	Equalization
Baseband Gbps / Gbps/W	9/8.3 Deep-Learn: 45.2 GOP/s/W	5.6/160	7.3/7.2	1.7/40.4	1.8/33.3	32.8/453	26.1/1893

- *In-phase and quadrature antenna throughput divided by runtime, 800MHz@0.8V.*
- *MIMO processed by accelerator or specialized datapaths ASIP;*
- *Technology normalized to 12nm and 0.8V core supply.*
- *The energy efficiency of deep learning workload (Conv2D example).*

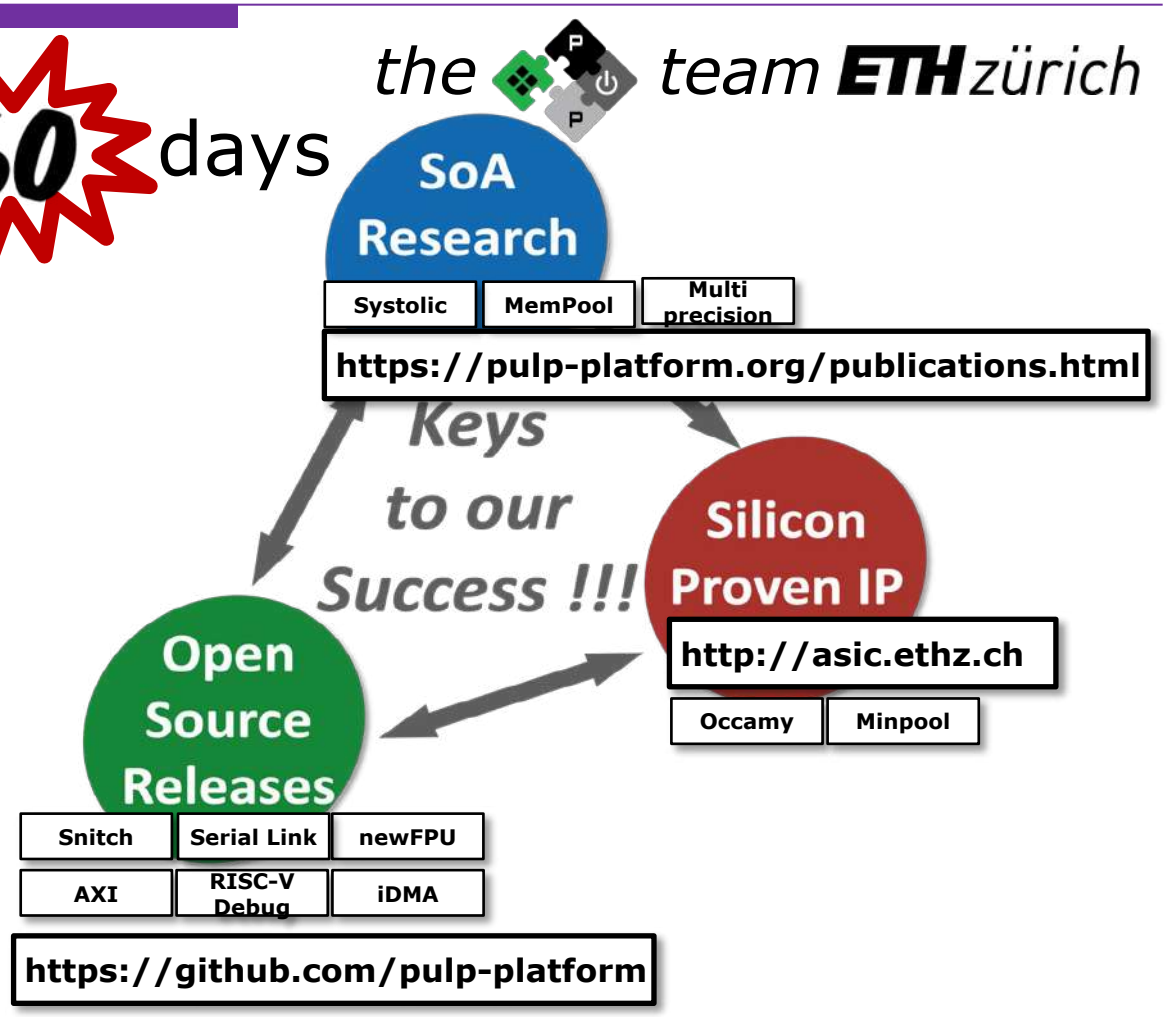
Open-Source Design Strategy

HeartStream tape-out in 60 days

the team ETH zürich

60 Days, 4 PhD Students :

- → Design idea
- → Team setup
- → RTL design
- → Physical Design
- → Tape-out



*https://pulp-platform.org/docs/riscvmunich2024/RISCV-Summit-EU_2024_kgf_60days.pdf

Outline

- "HeartStream" Introduction
- Background and Design Target
- Computing Cluster Design
- Performance and Energy Efficiency
- Conclusion

Conclusion

- HeartStream:
  - 64 Cores shared-L1-SPM computing cluster. → **Large workload, low overhead**
  - Complex Arith., FPU, Systolic-ext., DIVSQRT. → **Efficient AI-enhanced O-RAN**
  - Fully programmable PEs. → **SW-defined, flex baseband, time-to-market**
- Efficient PUSCH Processing:
  - Typical **9Gbps@0.8V** by a single cluster
  - High energy-efficiency: **49.6GFLOP/s/W**, **< 4ms** (3.2ms) latency (low-V usage)
  - Systolic extension improves up to **1.9x energy efficiency**
- **Open-source design** enables a faster design cycle for complex SoCs

https://github.com/pulp-platform/mempool