IEEE High Performance Extreme Computing Conference (HPEC), 2017

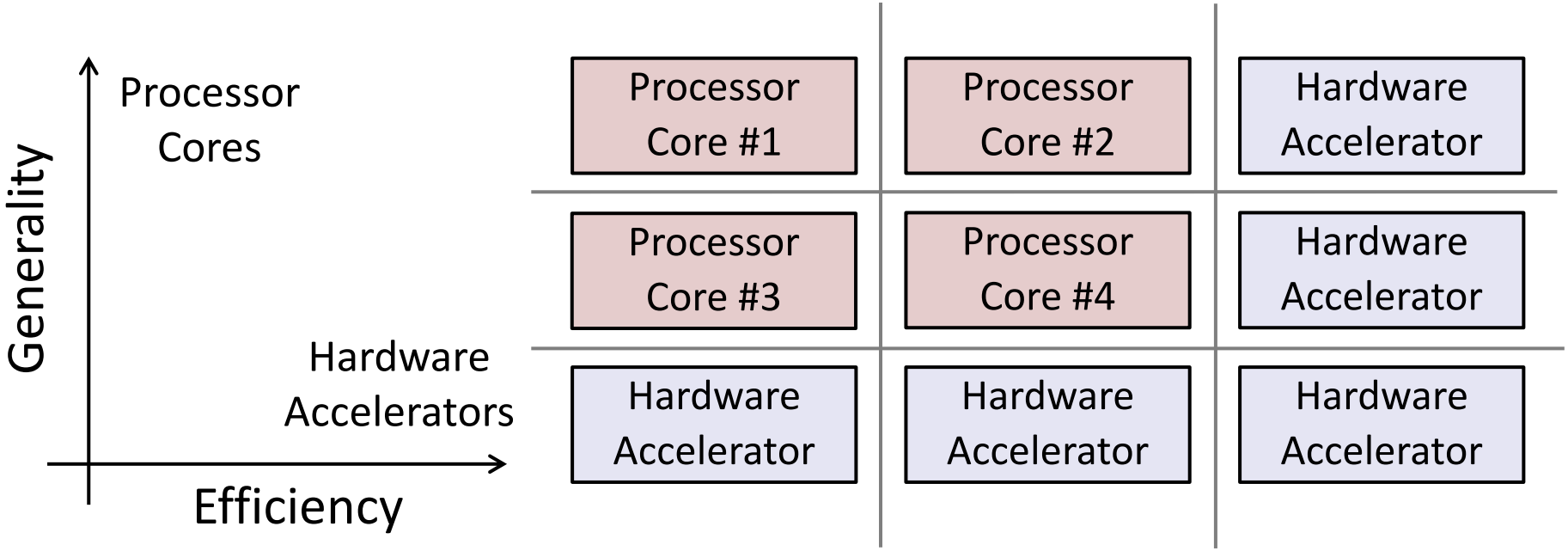
Broadening the Exploration of the Accelerator Design Space in Embedded Scalable Platforms

Luca Piccolboni, Paolo Mantovani,
Giuseppe Di Guglielmo, Luca Carloni
Columbia University, New York, NY, USA

Why Hardware Accelerators?

COLUMBIA UNIVERSITY

- High-performance embedded systems are heterogeneous:
  - they include multiple general-purpose processor cores
  - they include special-function hardware accelerators

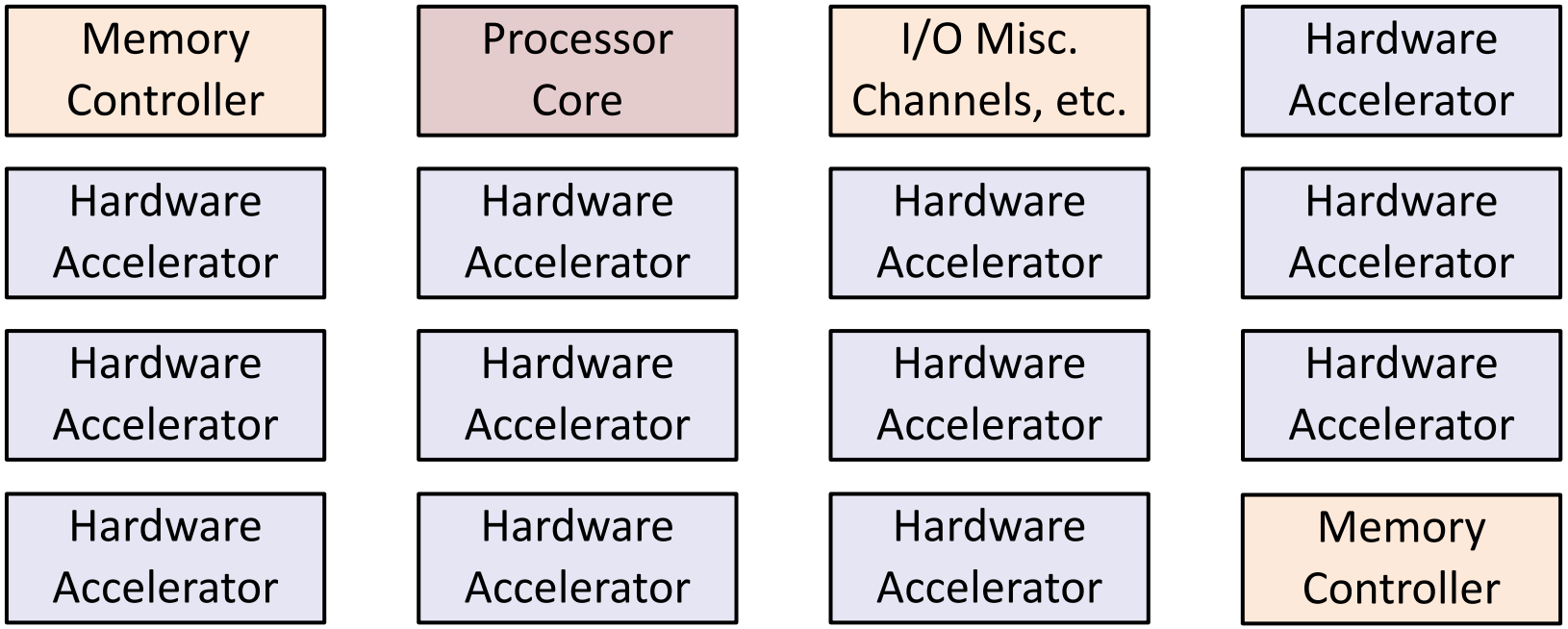


Embedded Scalable Platforms (ESP)

COLUMBIA UNIVERSITY

- To balance the demand for hardware specialization with the need of maintaining helpful degrees of regularity and modularity we proposed: **Embedded Scalable Platforms**

[L. Carloni, "*The Case for Embedded Scalable Platforms*", DAC 2016]



Embedded Scalable Platforms (ESP)

COLUMBIA UNIVERSITY

- To balance the demand for hardware specialization with the need of maintaining helpful degrees of regularity and modularity we proposed: Embedded Scalable Platforms

ESP instance for **WAMI (Wide-Area Motion Imagery)**

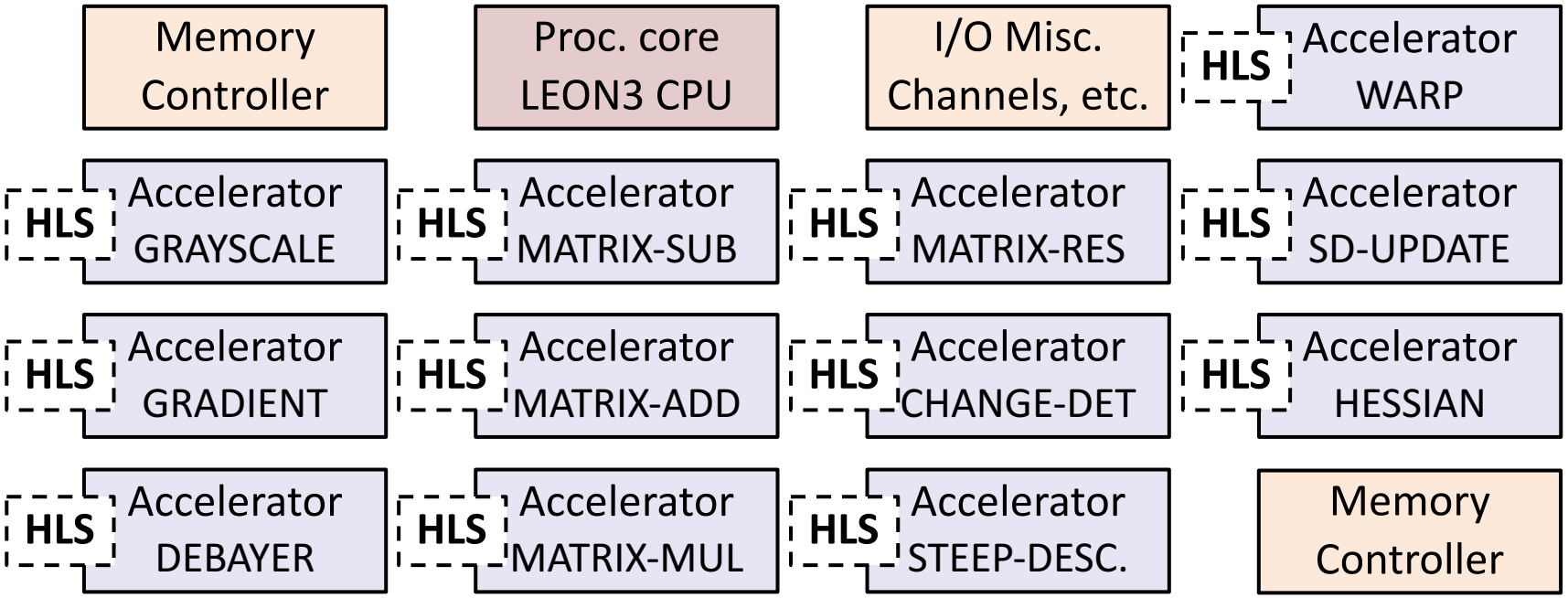
Memory Controller	Proc. core LEON3 CPU	I/O Misc. Channels, etc.	Accelerator WARP
Accelerator GRAYSCALE	Accelerator MATRIX-SUB	Accelerator MATRIX-RES	Accelerator SD-UPDATE
Accelerator GRADIENT	Accelerator MATRIX-ADD	Accelerator CHANGE-DET	Accelerator HESSIAN
Accelerator DEBAYER	Accelerator MATRIX-MUL	Accelerator STEEP-DESC.	Memory Controller

Embedded Scalable Platforms (ESP)

COLUMBIA UNIVERSITY

- To balance the demand for hardware specialization with the need of maintaining helpful degrees of regularity and modularity we proposed: Embedded Scalable Platforms

System-Level Design with **High-Level Synthesis (HLS)**



Embedded Scalable Platforms (ESP)

COLUMBIA UNIVERSITY

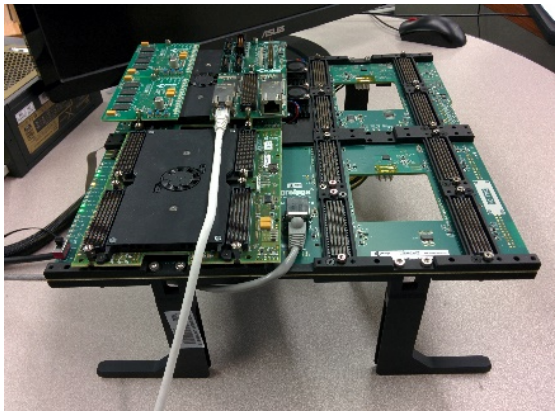
- To balance the demand for hardware specialization with the need of maintaining helpful degrees of regularity and modularity we proposed: Embedded Scalable Platforms

System-Level Design with High-Level Synthesis (HLS)

Memory Controller	Proc. core LEON3 CPU	I/O Misc. Channels, etc.	Accelerator WARP
Accelerator GRAYSCALE	Accelerator MATRIX-SUB	Accelerator MATRIX-RES	Accelerator SD-UPDATE
Accelerator GRADIENT	Accelerator MATRIX-ADD	Accelerator CHANGE-DET	Accelerator HESSIAN
Accelerator DEBAYER	Accelerator MATRIX-MUL	Accelerator STEEP-DESC.	Memory Controller

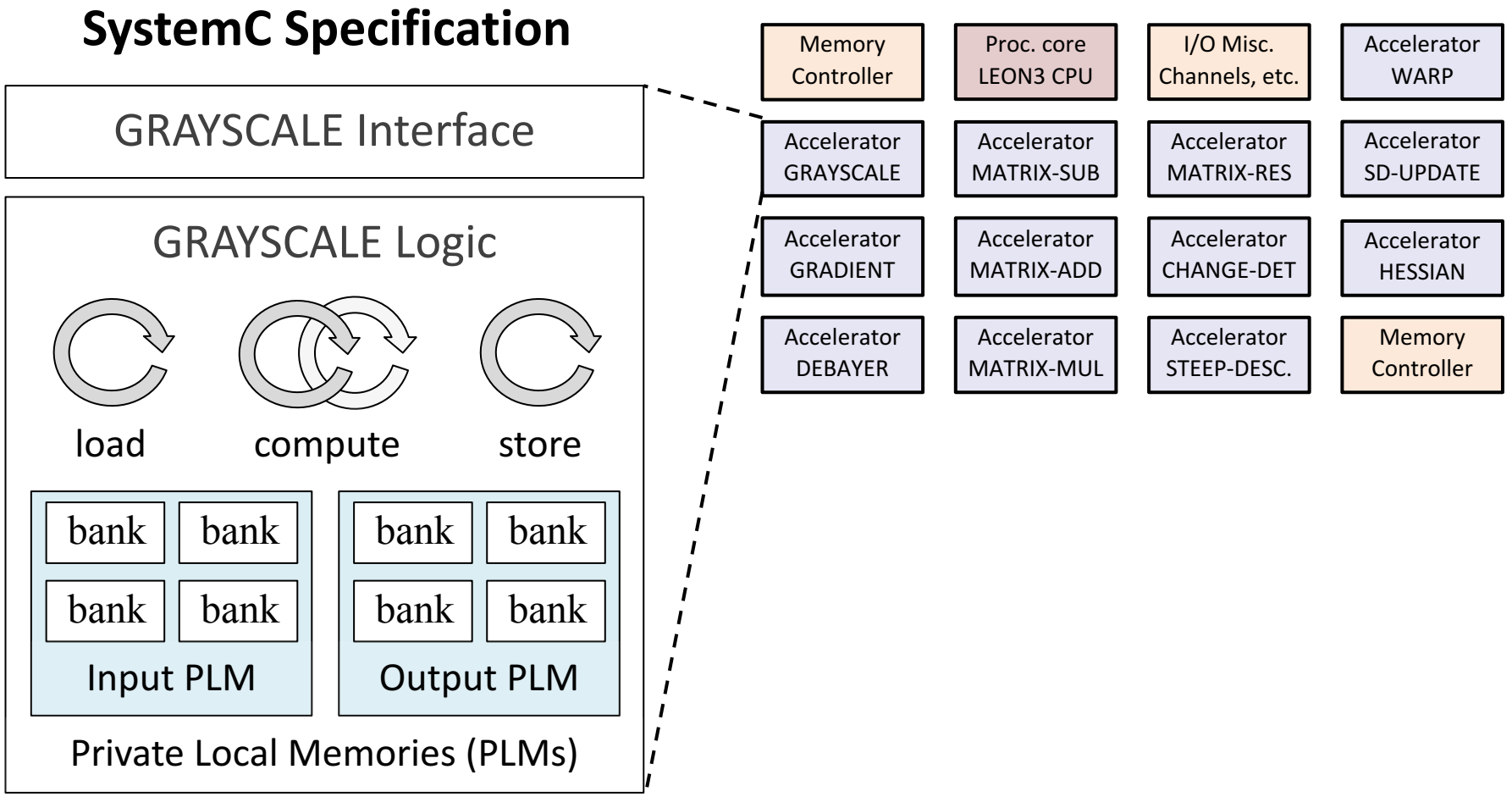
ESP instance for WAMI
(Wide-Area Motion Imagery)

rapid integration and prototyping



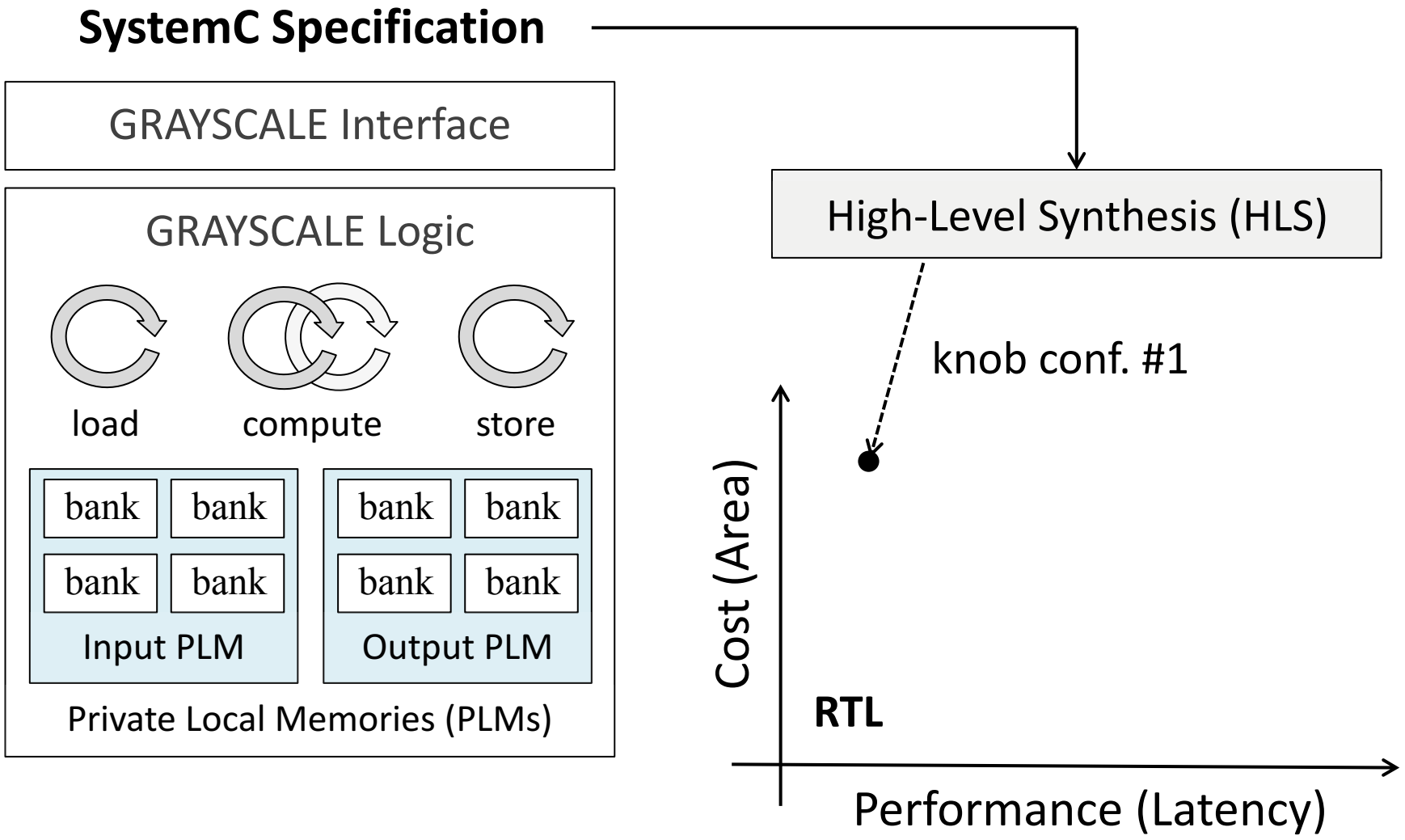
Hardware Accelerators with HLS

COLUMBIA UNIVERSITY



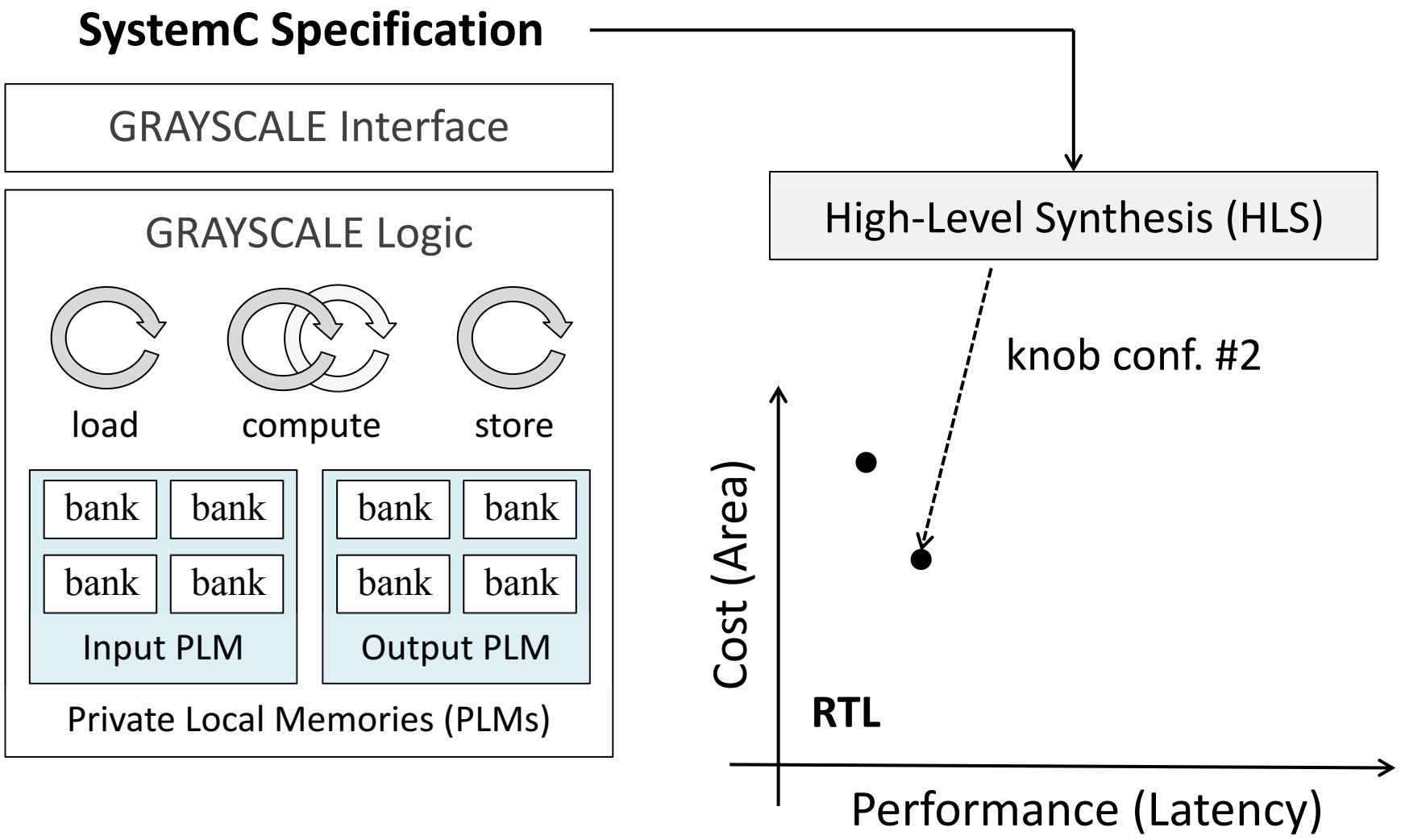
Hardware Accelerators with HLS

COLUMBIA UNIVERSITY



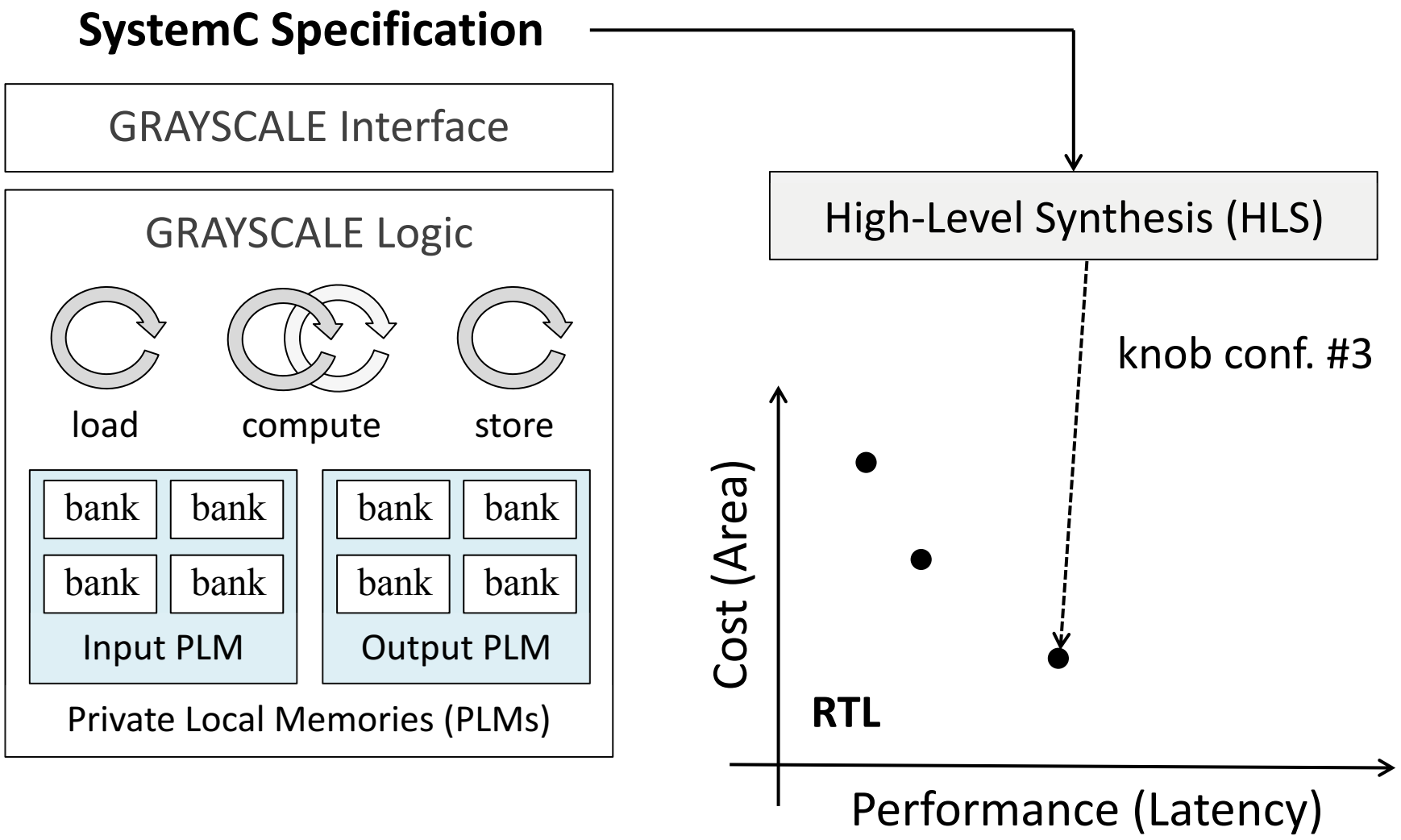
Hardware Accelerators with HLS

COLUMBIA UNIVERSITY



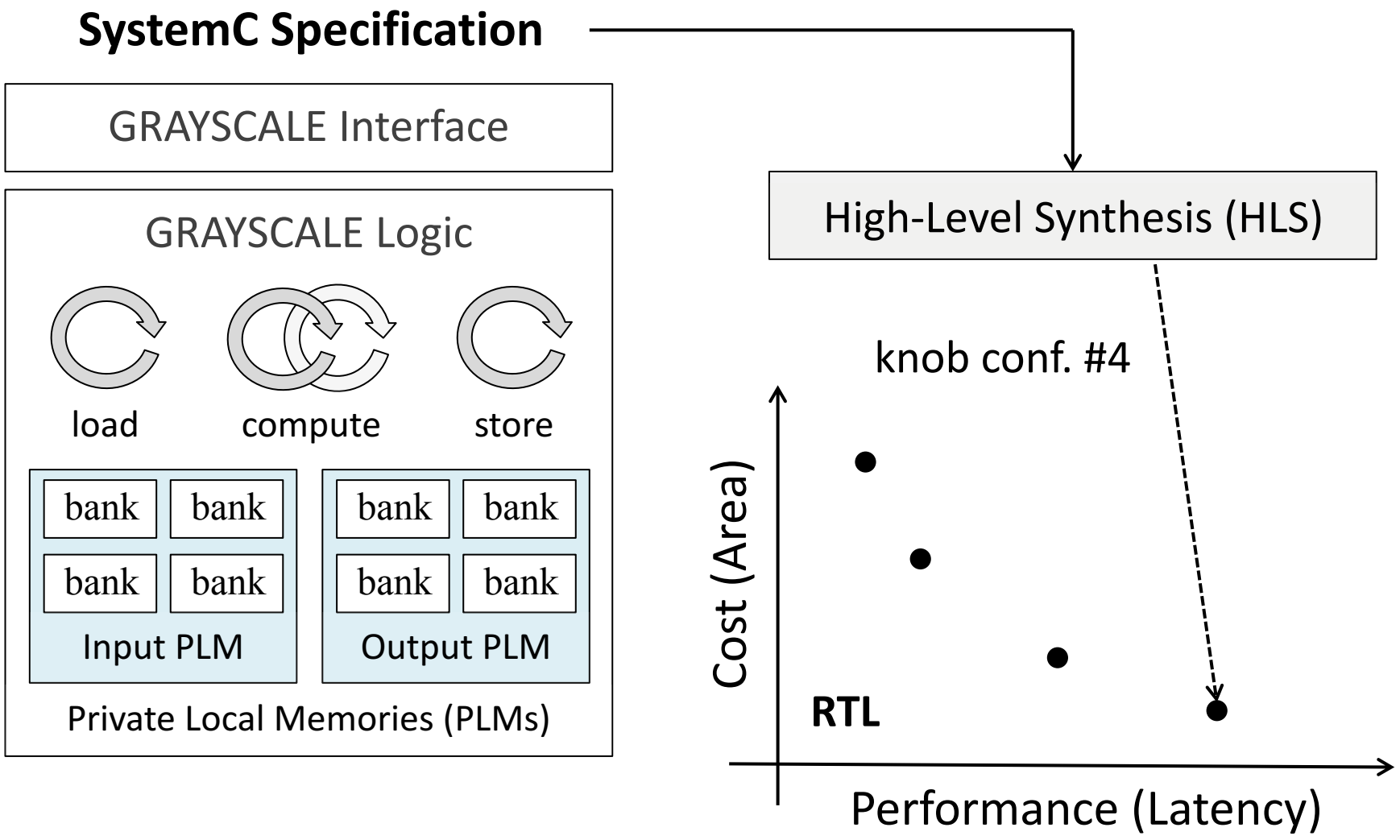
Hardware Accelerators with HLS

COLUMBIA UNIVERSITY



Hardware Accelerators with HLS

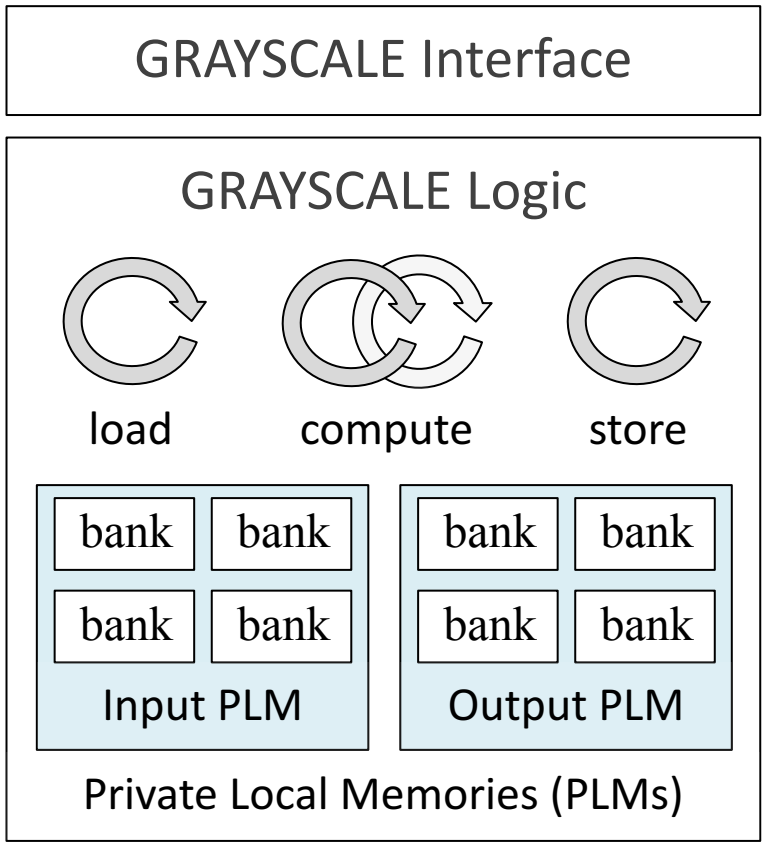
COLUMBIA UNIVERSITY



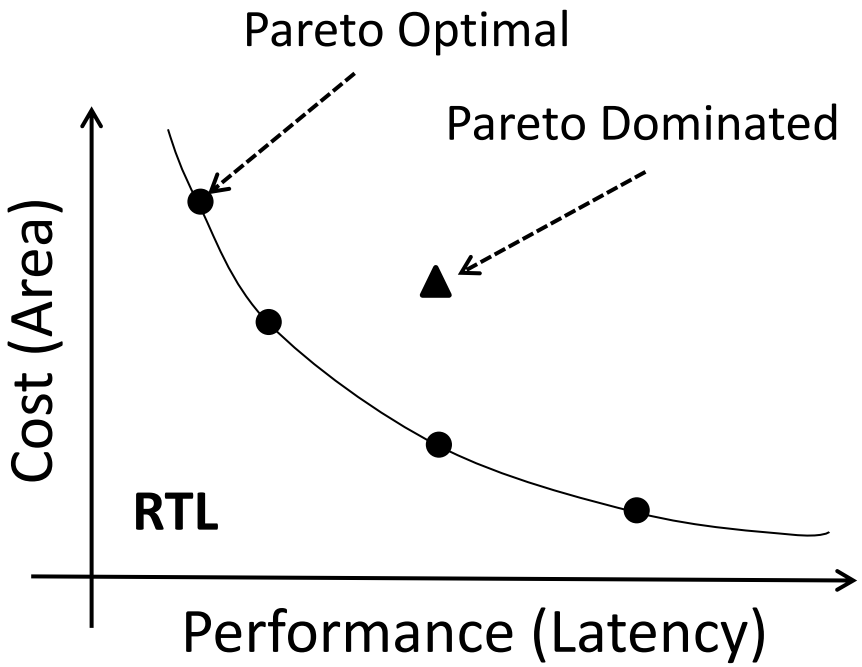
Hardware Accelerators with HLS

COLUMBIA UNIVERSITY

SystemC Specification



High-Level Synthesis (HLS)



Standard HLS Knobs

COLUMBIA UNIVERSITY

Standard knobs provided by the current HLS tools

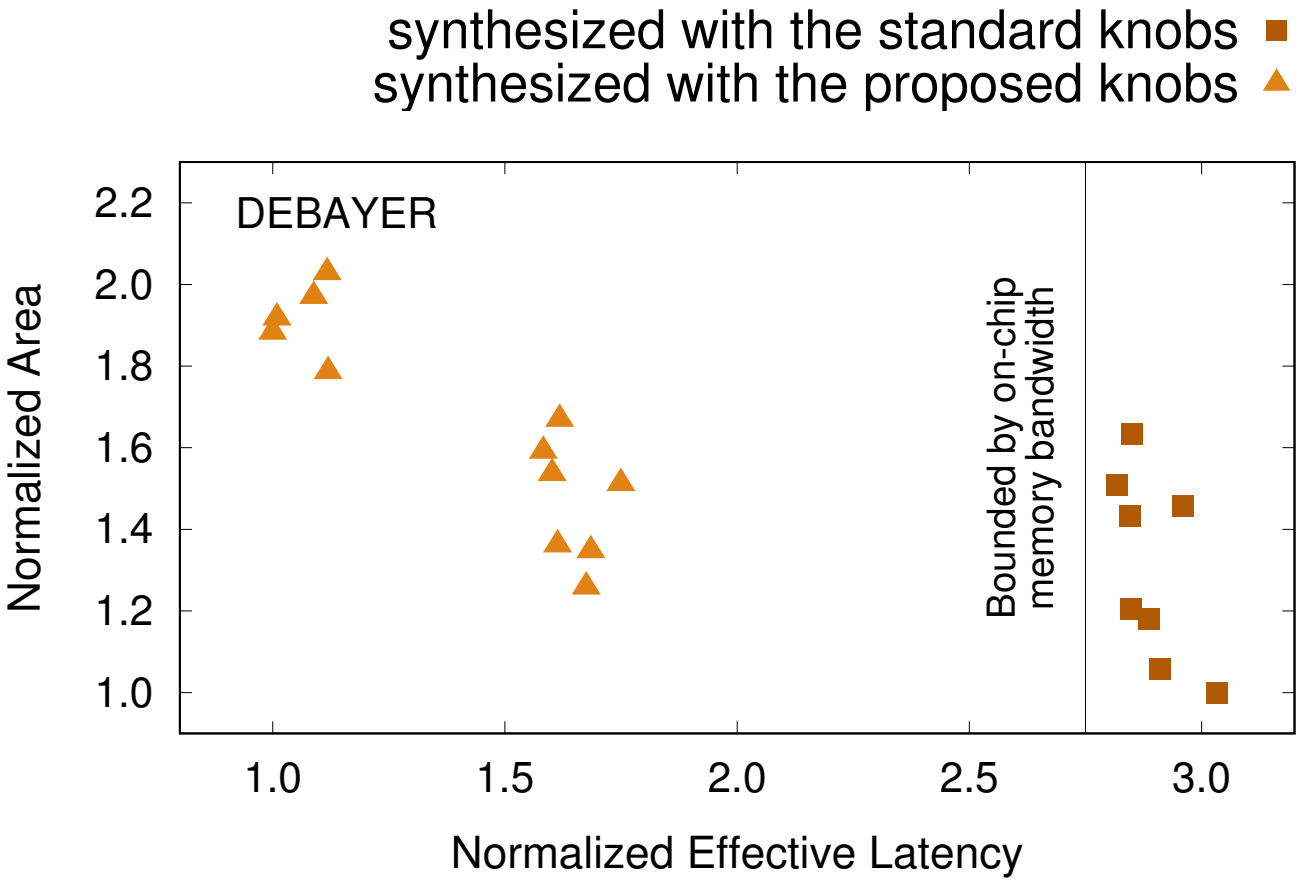
Knob	Settings and Effects
Loop manipulations	Unrolls, pipelines or breaks the body of loops
Array mappings	Maps arrays to registers or on-chip memories
Clock period	Sets the target clock period for synthesis

- These knobs enable already a rich design-space exploration
- However, they are not sufficient for exploring accelerators

  ↳ We need other knobs to broaden the exploration

Motivational Example #1

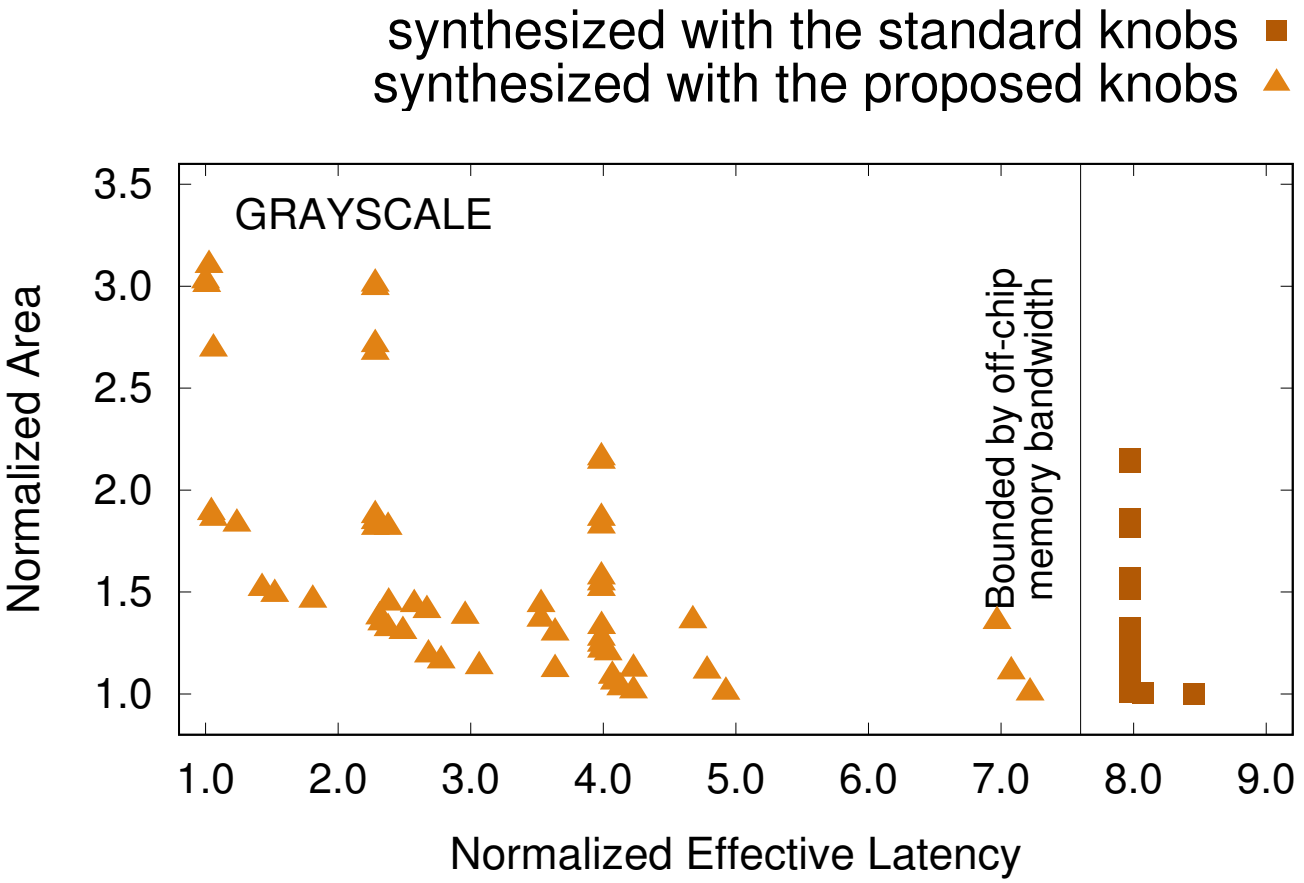
COLUMBIA UNIVERSITY



- **Limiting factor**: limited bandwidth to the on-chip memory
- **We need** knobs to tailor the PLM to the accelerator needs

Motivational Example #2

COLUMBIA UNIVERSITY



- **Limiting factor**: limited bandwidth to the off-chip memory
- **We need** knobs to operate on the communication interfaces

Contributions: Xknobs

COLUMBIA UNIVERSITY

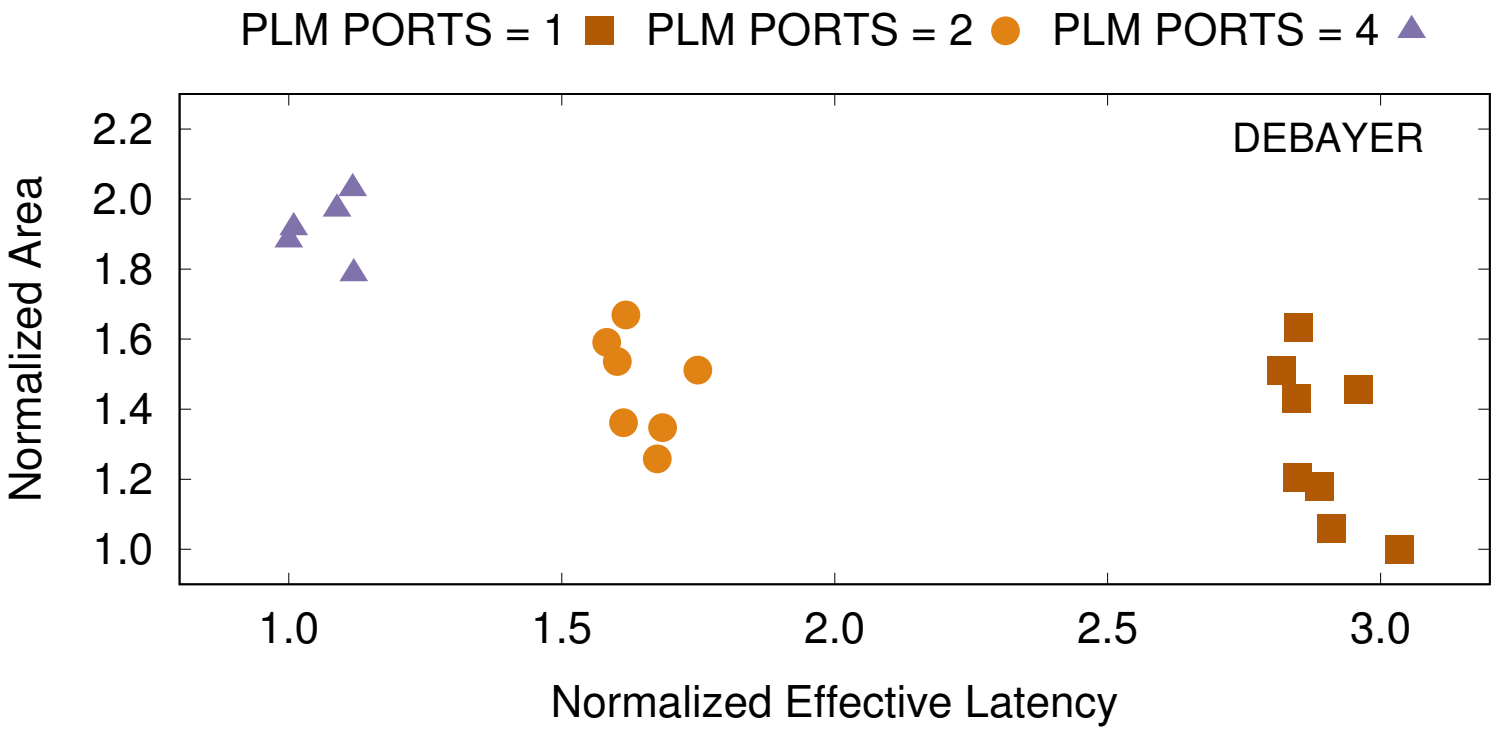
eXtended Knobs for High-Level Synthesis

XKnob	Settings and Effects
PLM PORTS	Sets the on-chip memory bandwidth
DMA WIDTH	Sets the off-chip memory bandwidth
DMA CHUNK	Sets the size of the input and output PLM

Xknob #1: PLM PORTS

COLUMBIA UNIVERSITY

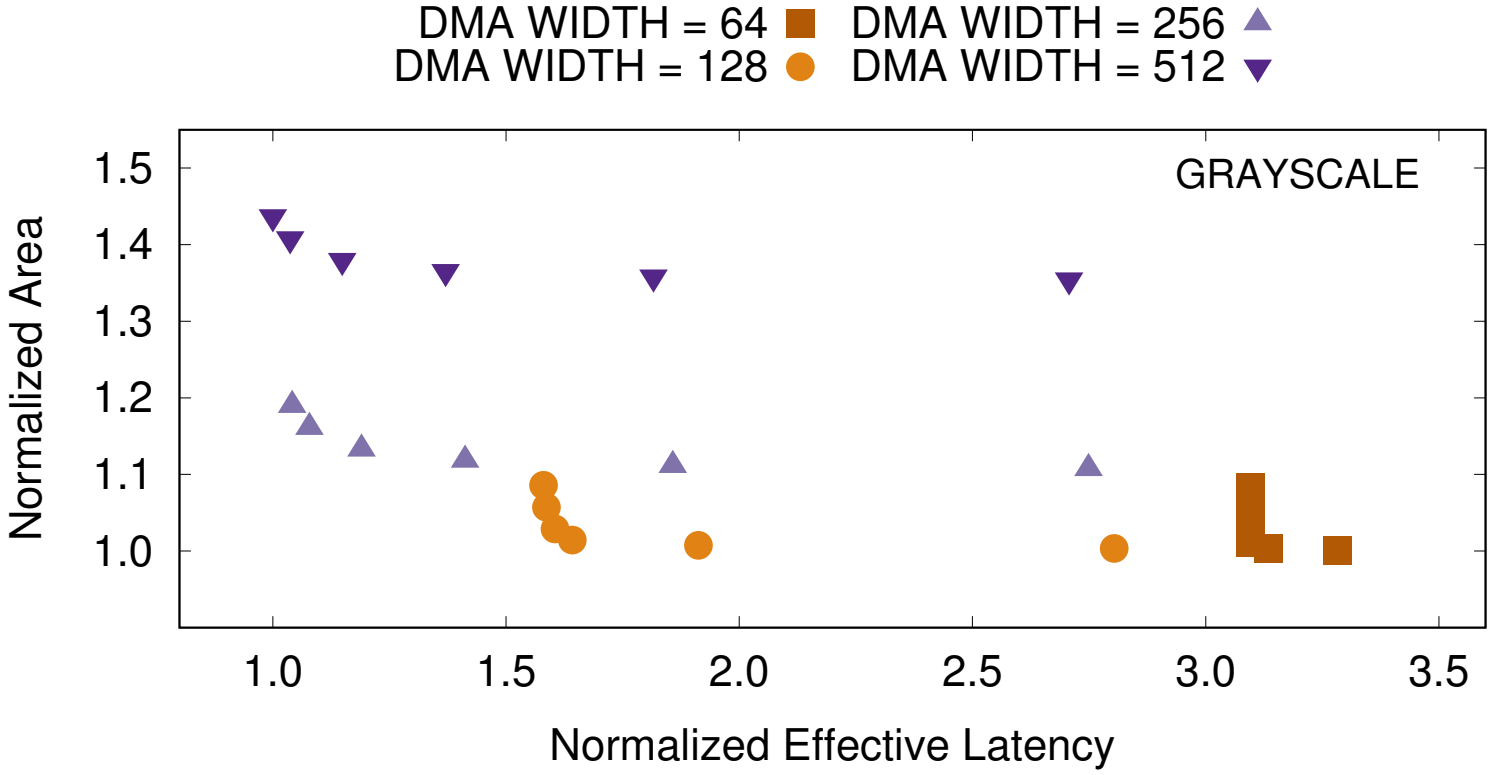
- Sets the number of read/write ports of input/output PLMs
- Higher values of PLM PORTS → more read/write accesses
- Higher values of PLM PORTS → higher area (more banks)



Xknob #2: DMA WIDTH

COLUMBIA UNIVERSITY

- Set the size in bits of the DMA communication channels
- Higher values of DMA WIDTH → higher mem. throughput
- Higher values of DMA WIDTH → higher area (more banks) (higher number of write/read ports of input/output PLMs)



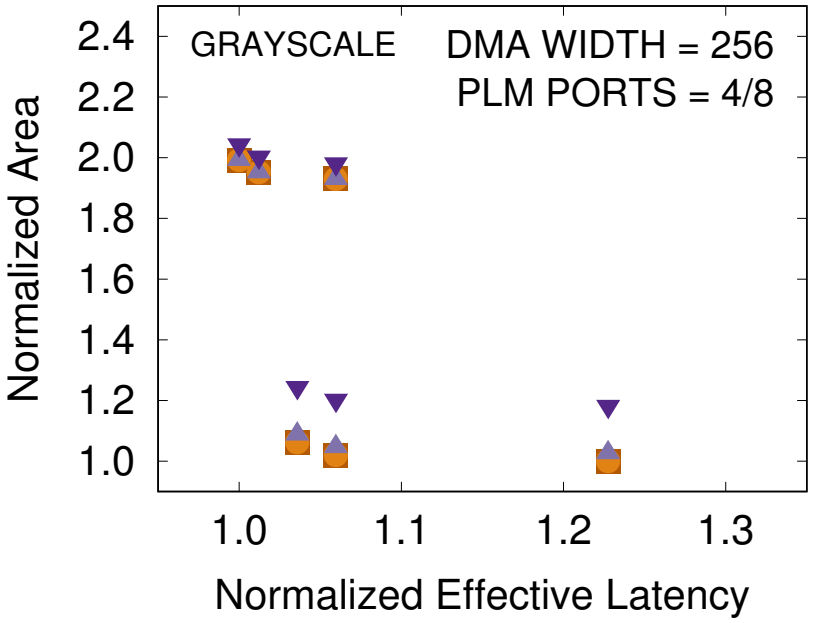
Xknob #3: DMA CHUNK

COLUMBIA UNIVERSITY

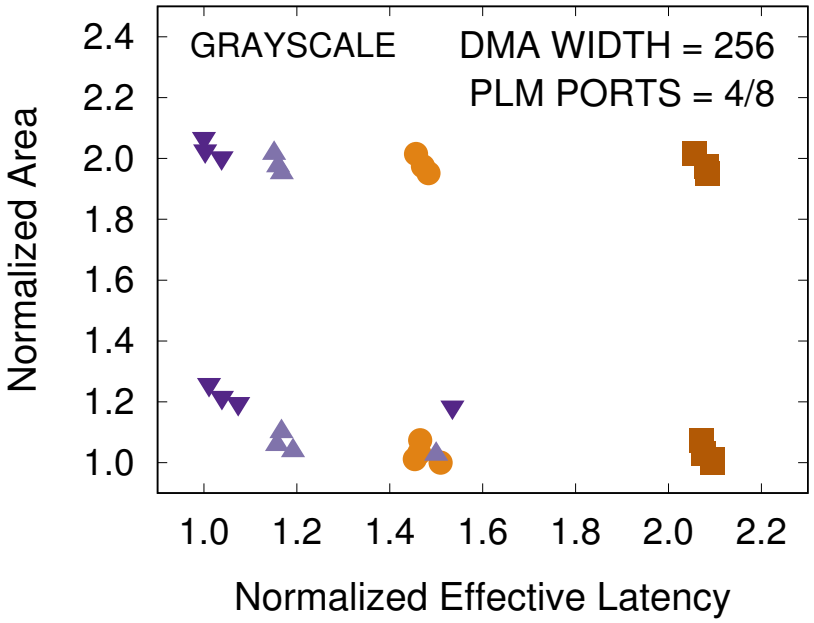
- Set the size of the PLM in multiple of the stored data type
- Higher values of DMA CHUNK → optimized communication
- Higher values of DMA CHUNK → higher area (for the PLM)

DMA CHUNK = 256 ■ DMA CHUNK = 512 ● DMA CHUNK = 1024 ▲ DMA CHUNK = 2048 ▼

without contention



with contention



Experimental Results

COLUMBIA UNIVERSITY

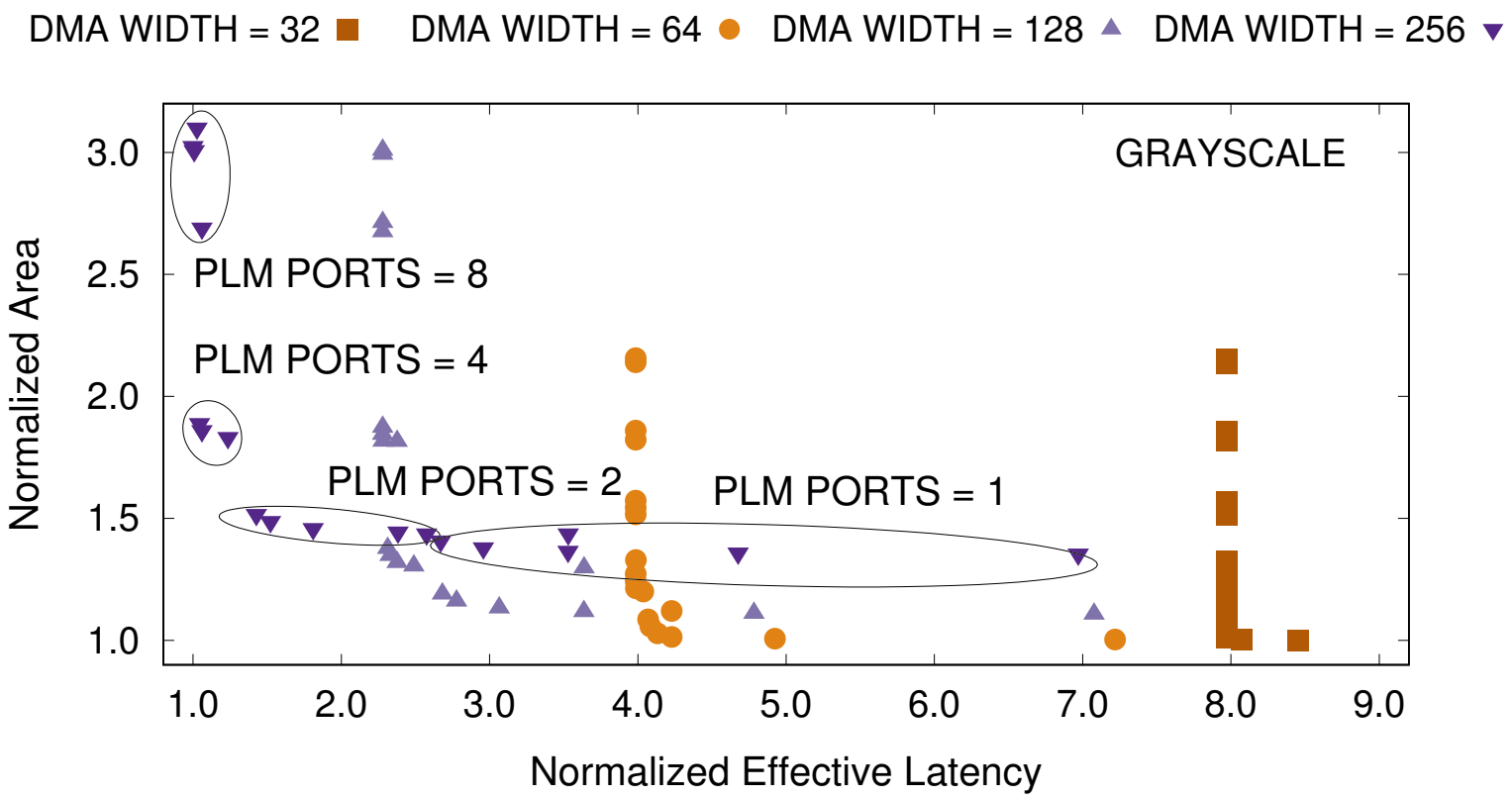
- We evaluate the combined effects of the **XKnobs** by using:
  - **GRAYSCALE** → accelerator limited by communication
  - **DEBAYER** → accelerator limited by computation

- The other WAMI accelerators behave similarly to either the GRAYSCALE accelerator or the DEBAYER accelerator

Experiment #1

COLUMBIA UNIVERSITY

- We consider two **XKnobs**: PLM PORTS and DMA WIDTH

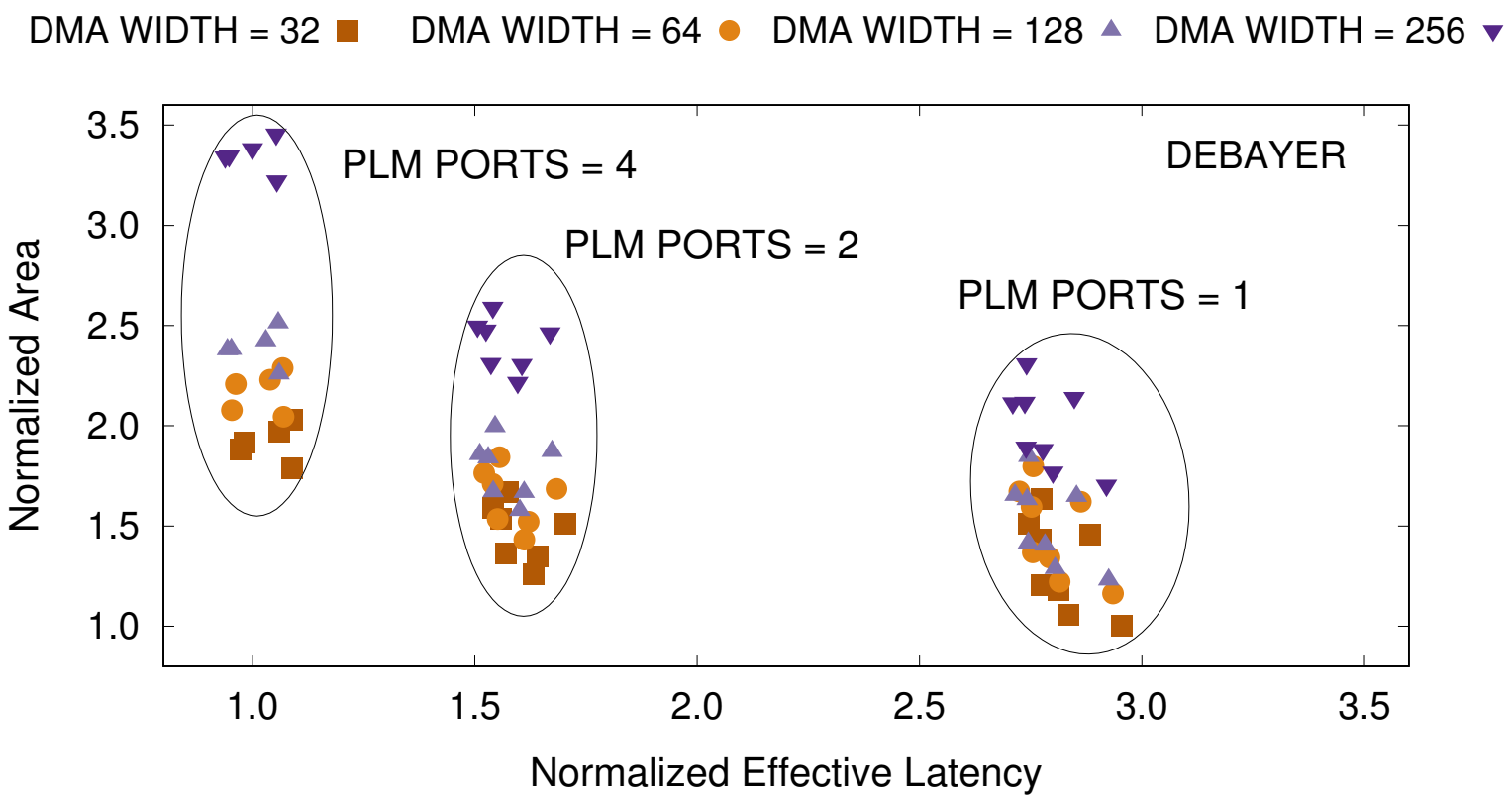


- **GRAYSCALE** → accelerator limited by communication

Experiment #1

COLUMBIA UNIVERSITY

- We consider two **XKnobs**: PLM PORTS and DMA WIDTH

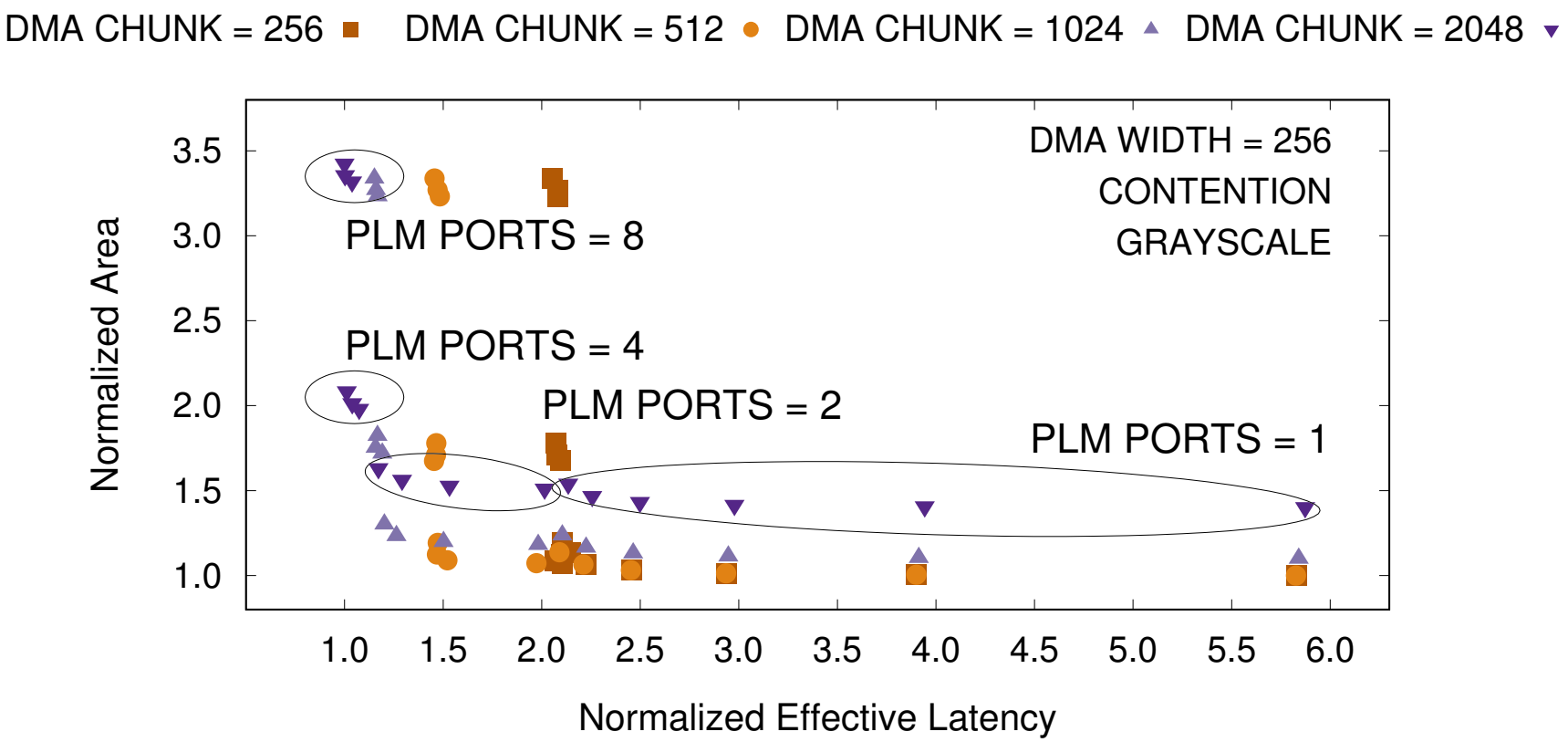


- **DEBAYER** → accelerator limited by computation

Experiment #2

COLUMBIA UNIVERSITY

- We consider two **XKnobs**: PLM PORTS and DMA CHUNK

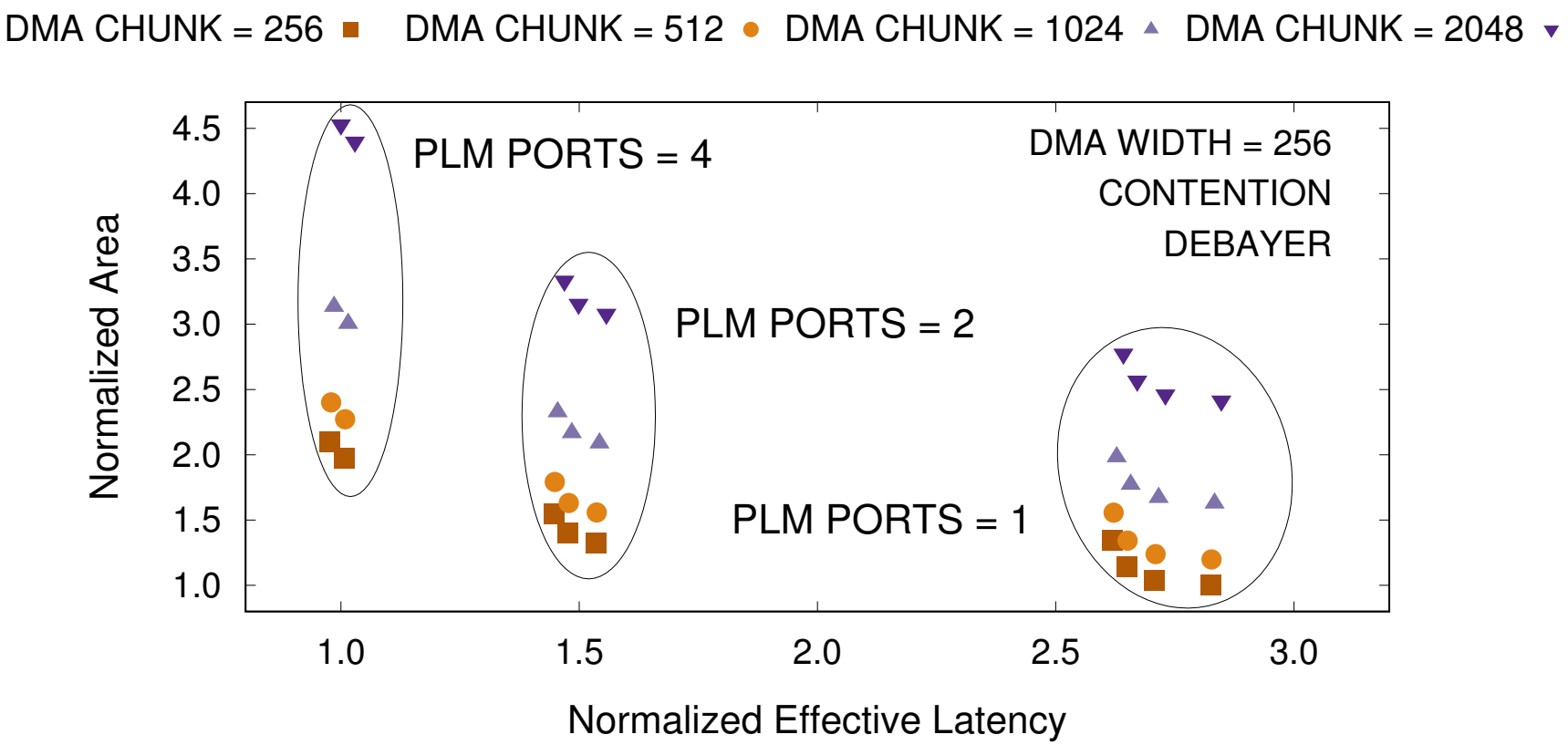


- **GRAYSCALE** → accelerator limited by communication

Experiment #2

COLUMBIA UNIVERSITY

- We consider two **XKnobs**: PLM PORTS and DMA CHUNK



- **DEBAYER** → accelerator limited by computation

Concluding Remarks

COLUMBIA UNIVERSITY

- We presented the **XKnobs**
  - a set of knobs that aims at extending the standard knobs used in current HLS tools
- For WAMI, the **Xknobs** broaden the design space by up to **8.5x** for performance and **3.5x** for cost
- The **XKnobs** can be integrated in any HLS tools and design-space exploration methodologies to enrich the set of Pareto-optimal implementations of hardware accelerators

IEEE High Performance Extreme Computing Conference (HPEC), 2017

Broadening the Exploration of the Accelerator Design Space in Embedded Scalable Platforms

Thank you for the attention!

Speaker: **Luca Piccolboni**
Columbia University, NY