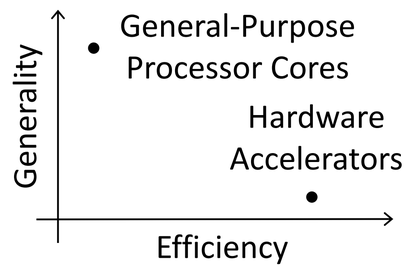
COSMOS: Coordination of High-Level Synthesis and Memory Optimization for Hardware Accelerators

Luca Piccolboni, Paolo Mantovani, Giuseppe Di Guglielmo, Luca Carloni

Columbia University, New York, NY, USA

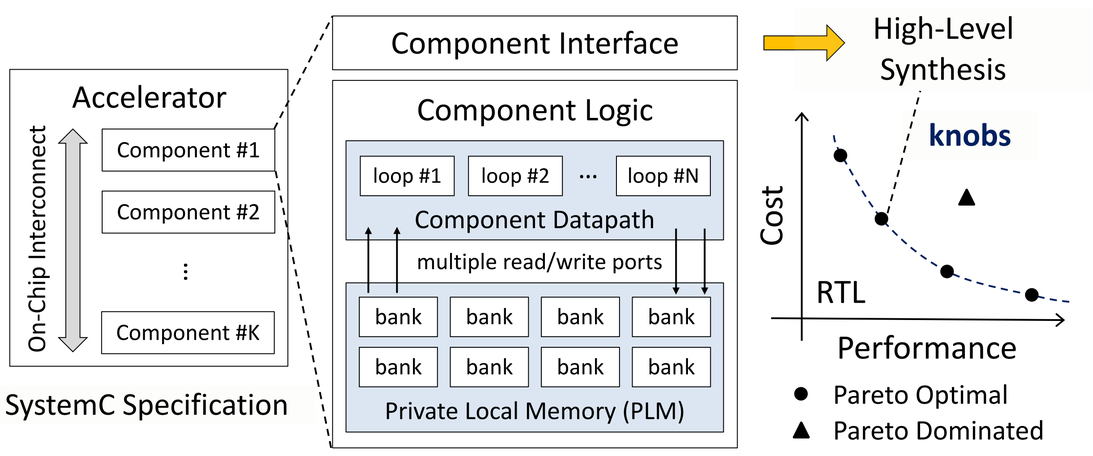
The Need of Accelerator-Rich Computing

- Hardware accelerators are devices that are designed and optimized to execute very specific functionalities
- Hardware accelerators ensure **high performance** and **energy efficiency**



Hardware Accelerators with High-Level Synthesis

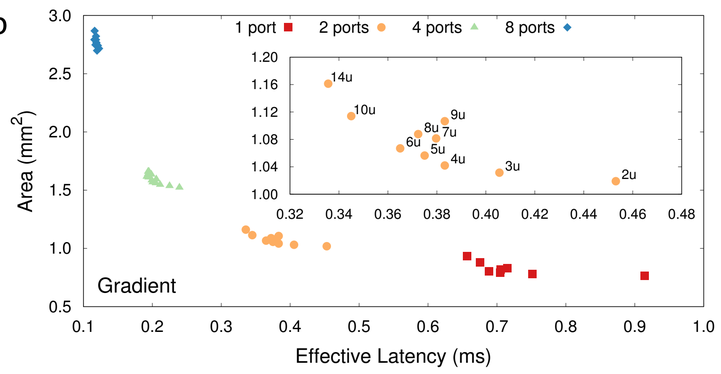
- High-Level Synthesis improves **productivity** and **reusability**



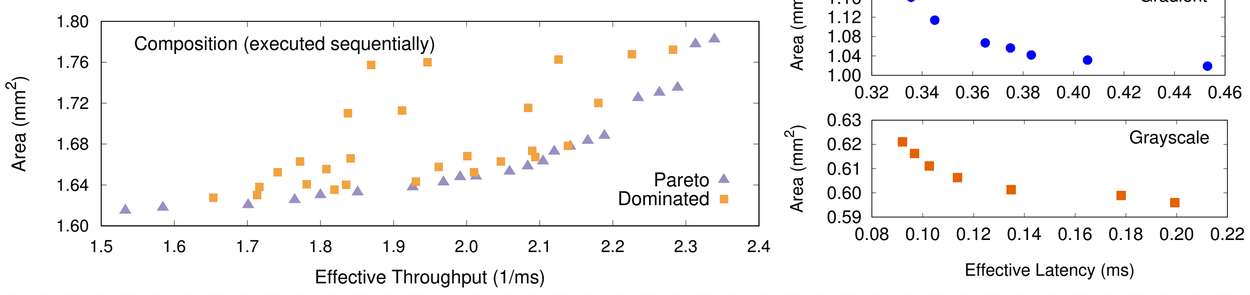
Limitations in Exploring the Design Space of Accelerators

- **Design-Space Exploration**: from a *single* SystemC specification obtain *many* RTL designs with different characteristics in terms of cost and performance

1. Several HLS tools do not take into account **Private Local Memories**
   - spans no mem: area 1.2x, latency 1.4x
   - spans w/ mem: area 3.7x, latency 7.9x
2. **Heuristics** used by the HLS tools make it difficult to set the knobs
   - increasing the number of unrolls can lead to Pareto-dominated designs (7u)

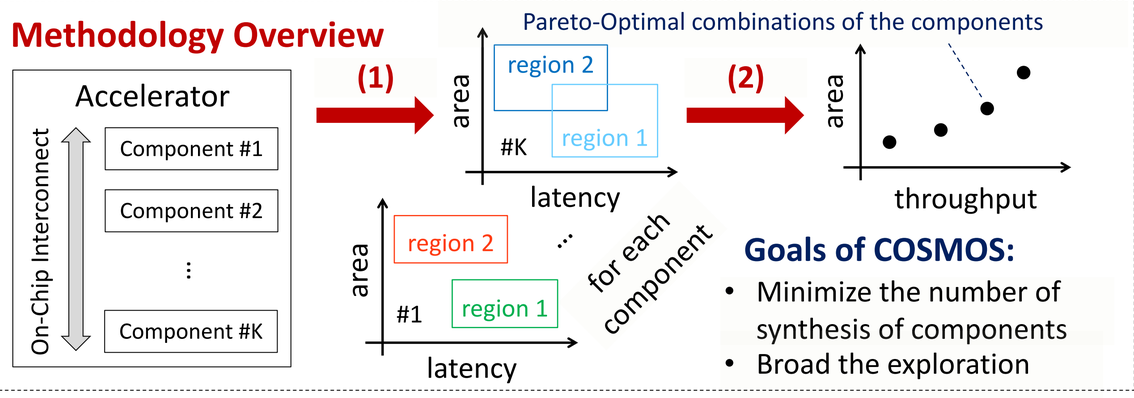


3. HLS tools cannot simultaneously synthesize and optimize **multiple components**
   - it is important to determine the critical components

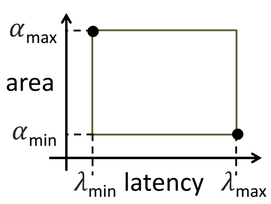


COSMOS: A Design-Space Exploration Methodology for Hardware Accelerators

Methodology Overview



(1) Component Characterization



- a region includes designs with the *same number of ports* in the local memory
- $\lambda\_constraint_{ports}(u)$ is **sat** if the number of states of the loop is $\leq h_{ports}(u)$
- Function h estimates the *number of states* for 1 iteration:

$$h_{ports}(u) = \left\lceil \frac{u * \gamma_{read}}{ports} \right\rceil + \left\lceil \frac{\gamma_{write}}{ports} \right\rceil + \eta$$

η: latency of operations that do not access the memory

$\gamma_{read}, \gamma_{write}$: max number of r/w accesses to the same array

Algorithm: component characterization

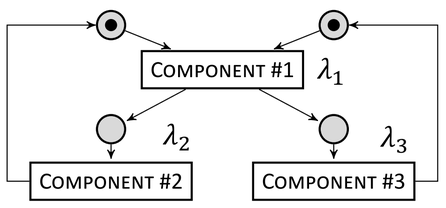
Input: *clock, max_ports, max_unrolls*

Output: regions (λ_{max}, α_{min}, λ_{min}, α_{max})

```
for ports = 1 up to max_ports do
  // 1. Find the lower-right point of the region
  (λ_max , α_min) = hls_tool(ports, ports, clock);
  // 2. Identify the upper-left point of the region
  for unrolls = max_unrolls down to ports + 1 do
    (λ_min , α_max) = hls_tool(unrolls, ports, clock);
    if λ_constraint_ports(unrolls) is sat then break;
  // 3. Generate the private local memory
  α_plm = mem(ports); α_min += α_plm; α_max += α_plm;
```

(2) Design-Space Exploration

- We use **timed marked graphs**, a subclass of Petri nets, to model the accelerators



N transitions, M places

- minimum cycle time → $max\ (D_k / N_k)$ with $k \in \mathbf{K}$, **K** is the set of cycles of the graph, D_k is the sum of the latencies in cycle k, and N_k is the number of tokens (●) in cycle k
- effective throughput ϑ → reciprocal of min. cycle time

- **Synthesis Planning:** we use a ϑ-constrained cost-minimization LP formulation:

f_i → calculates the estimated area of the *i*-th component
λ^- → vector $\mathbb{R}^N$ with the latencies of the N components
$\lambda^-_{min}, \lambda^-_{max}$ → min/max latencies from characterization **(1)**
M_0 → vector $\mathbb{N}^M$ with the number of tokens in the M places
σ → vector $\mathbb{R}^M$ with the transition-firing initiation-time values
A[i, j] → +/- 1 if trans. j is an output/input of place i, 0 otherwise

$$min \sum_{i=1}^{n} f_i(\lambda_i)$$
$$\text{s.t. } A\sigma + M_0/\vartheta \geq \lambda^-$$
$$\lambda^-_{min} \leq \lambda^- \leq \lambda^-_{max}$$

[Liu et al., ACM/IEEE DATE '12]

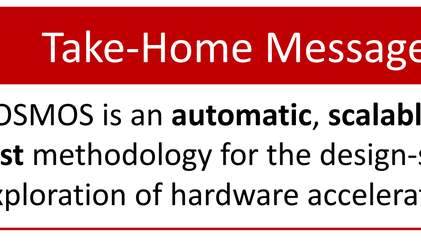
- **Synthesis Mapping:** we need to map the LP solutions to knob settings:

$$S = \frac{1}{(1-P) + \frac{P}{F}}$$

Amdahl's Law

$$P = \frac{\mu_{target} - \mu_{min}}{\mu_{max} - \mu_{min}}$$

$$S = \frac{\lambda_{target}}{\lambda_{max}} \quad F = \frac{\lambda_{min}}{\lambda_{max}}$$



Experimental Results

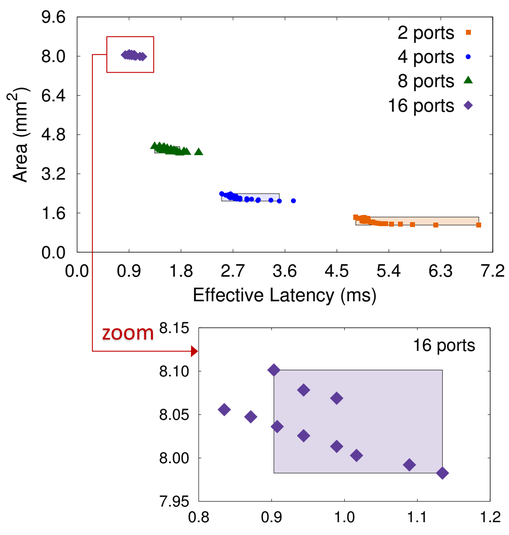
(1) Component Characterization

- Characterization of the WAMI components:

	COSMOS			No Memory	
Component	**reg.**	**λ_{span}**	**α_{span}**	**λ_{span}**	**α_{span}**
DEBAYER	3	2.89x	1.99x	1.04x	1.36x
GRAYSCALE	4	6.91x	3.41x	2.75x	1.14x
GRADIENT	4	7.89x	3.65x	1.39x	1.22x
HESSIAN	4	7.70x	7.30x	1.44x	1.30x
SD-UPDATE	4	9.87x	2.01x	2.78x	1.79x
MATRIX-SUB	4	2.75x	3.98x	1.88x	1.05x
MATRIX-ADD	3	1.53x	1.01x	1.26x	1.01x
MATRIX-MUL	3	2.88x	3.05x	1.92x	1.14x
MATRIX-RESH	1	1.02x	1.04x	1.02x	1.04x
STEEP.-DESCENT	1	1.95x	1.46x	1.95x	1.46x
CHANGE-DET.	1	2.21x	1.04x	2.21x	1.04x
WARP	1	1.09x	1.03x	1.09x	1.03x
Average	-	4.06x	2.58x	1.73x	1.22x

- "reg." is the number of regions, "λ_{span}" is the latency span, "α_{span}" is the area span

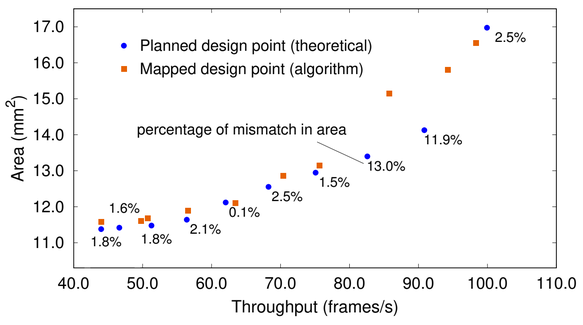
- Characterization of HESSIAN:



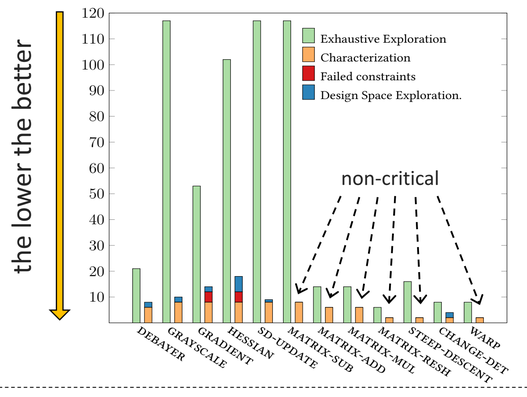
- Some Pareto-dominated designs cannot be avoided by using only the proposed λ-constraint
- Some Pareto-optimal designs can be outside the regions, but they can still be used in step **(2)**

(2) Design-Space Exploration

- Pareto for the WAMI accelerator:

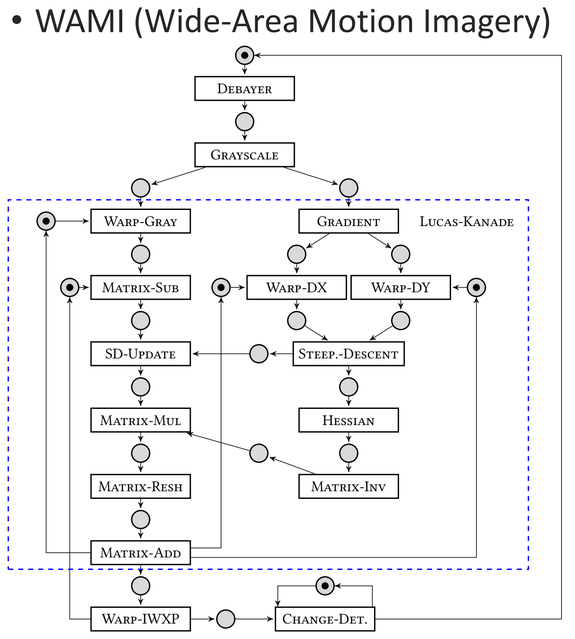


- Number of calls to the HLS tool:



Case Study

- WAMI (Wide-Area Motion Imagery)



Take-Home Message

COSMOS is an **automatic**, **scalable** and **fast** methodology for the design-space exploration of hardware accelerators