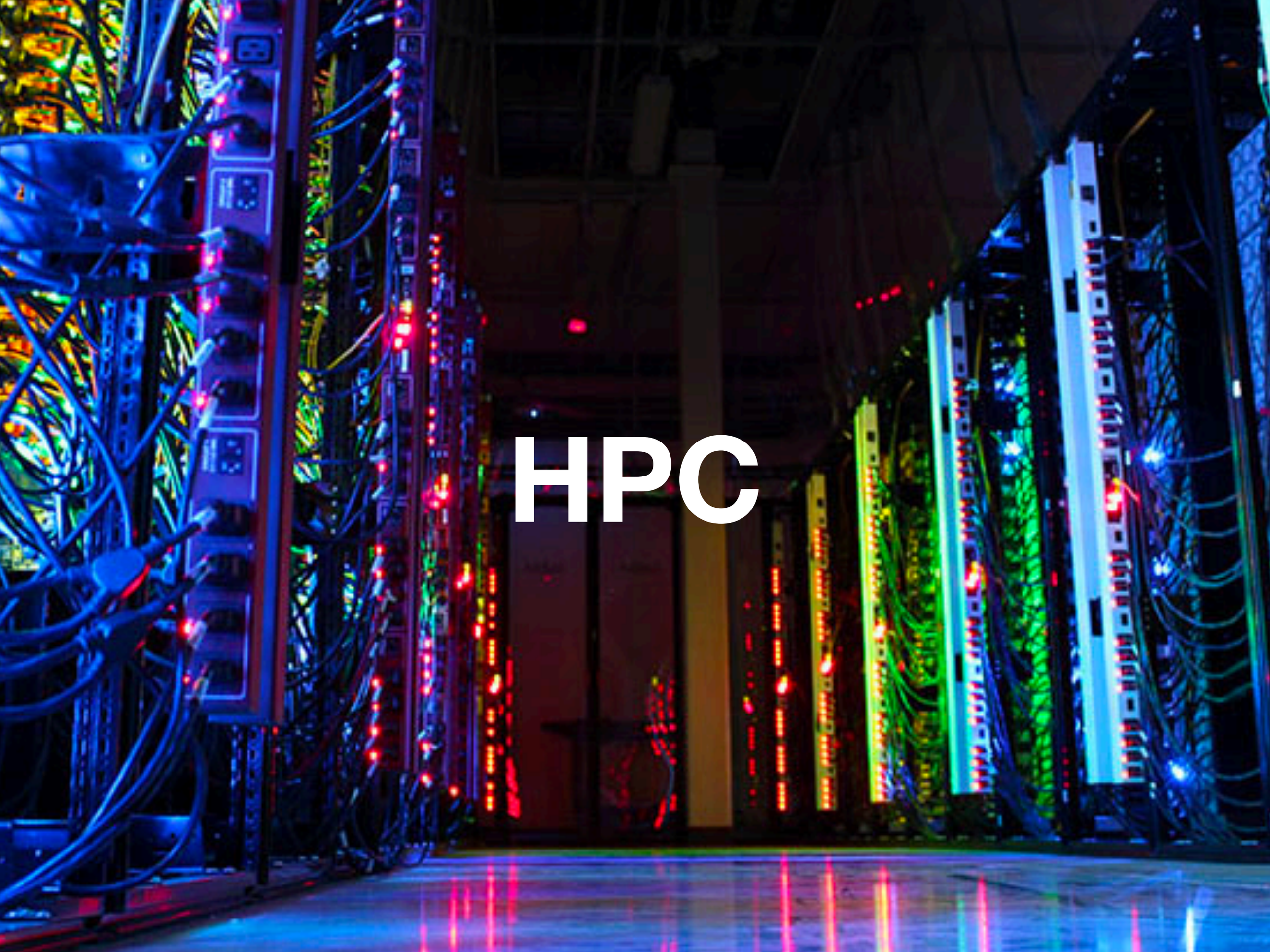
An FPGA-based Acceleration Methodology and Performance Model for Iterative Stencils

Enrico Reggiani, Giuseppe Natale, Carlo Moroni, Marco D. Santambrogio

Reconfigurable Architecture Worskshop

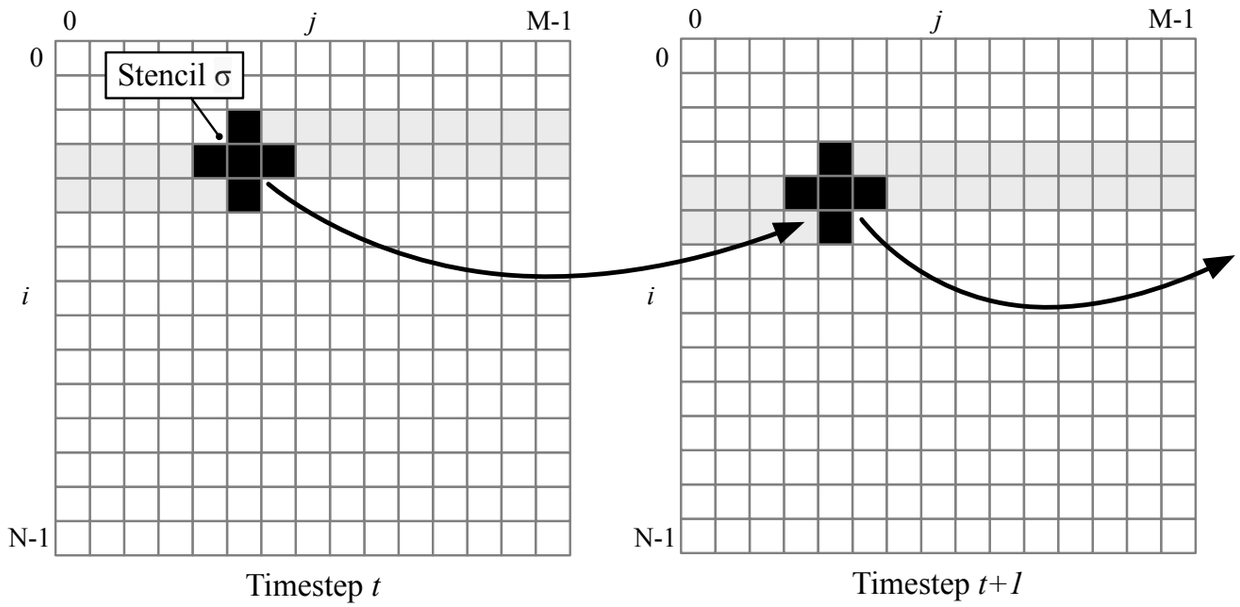
Vancouver, British Columbia, Canada
May 21, 2018

Giuseppe Natale - giuseppe.natale@polimi.it

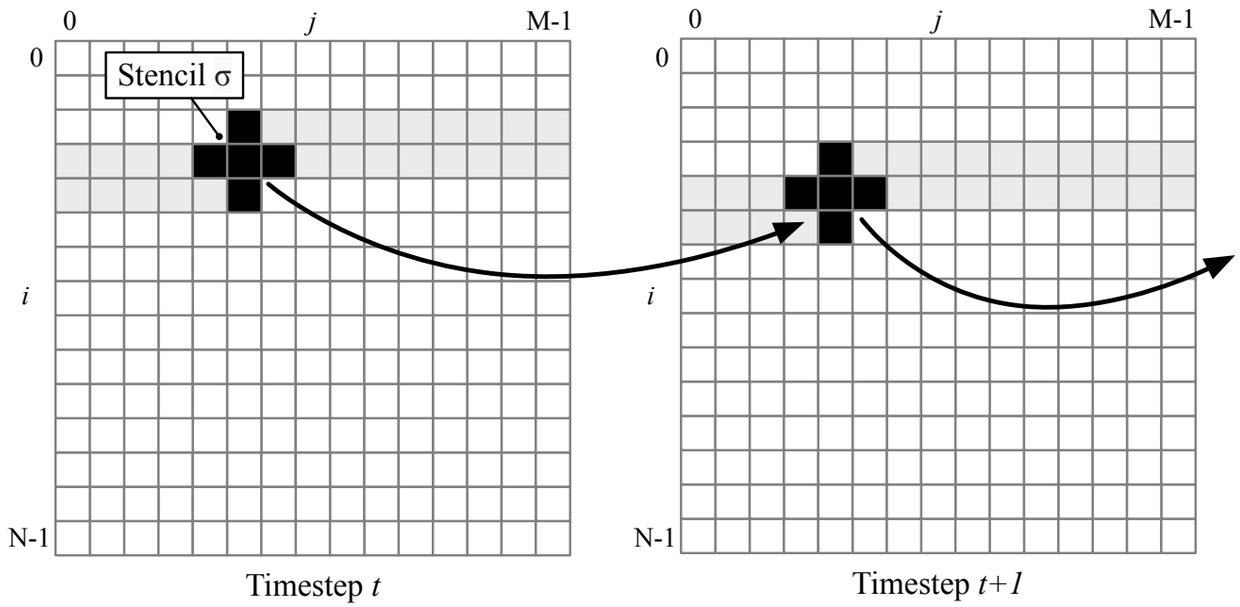


HPC

Iterative Stencil Algorithms

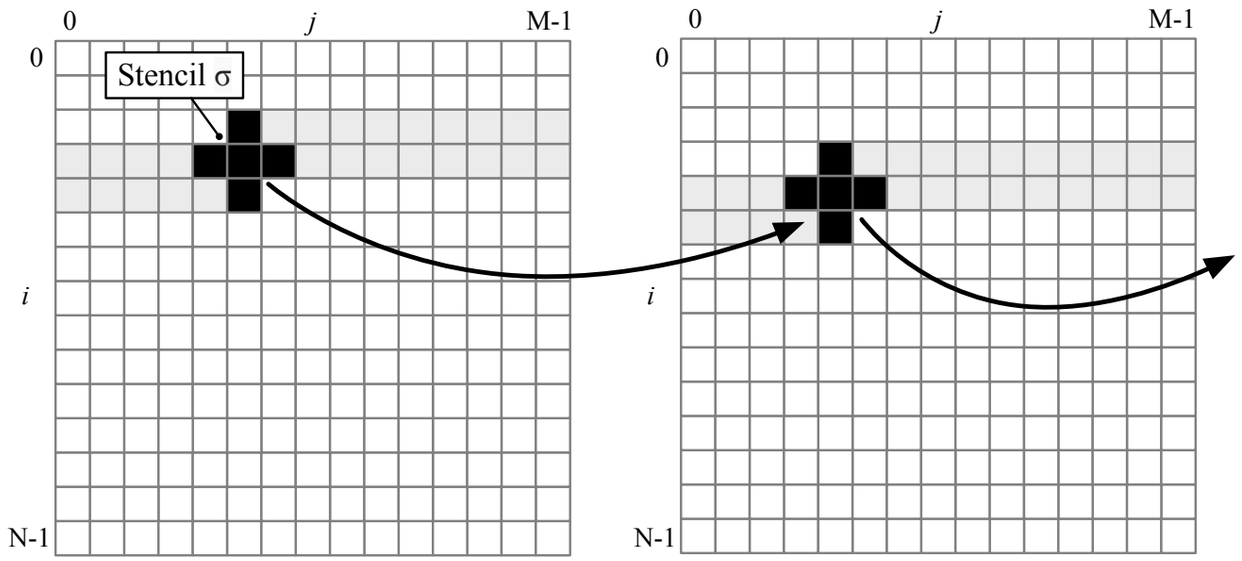


Iterative Stencil Algorithms

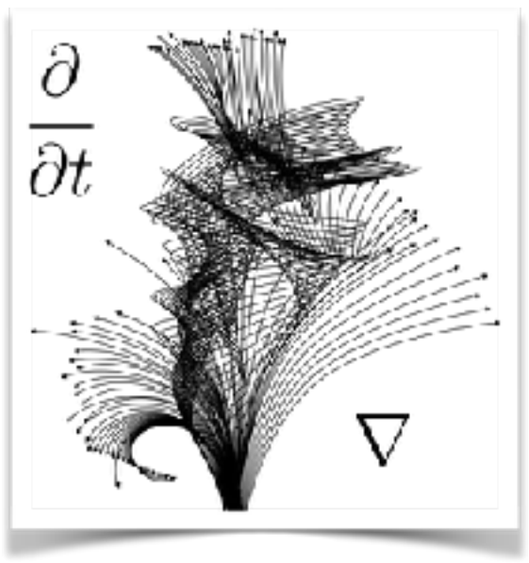


Applications

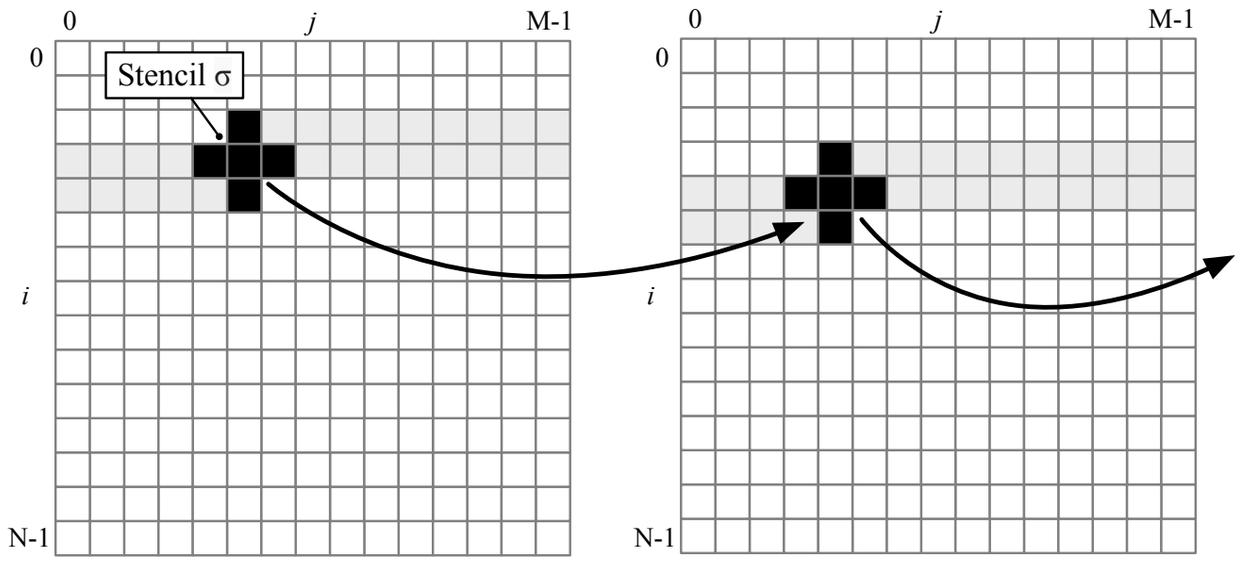
Iterative Stencil Algorithms



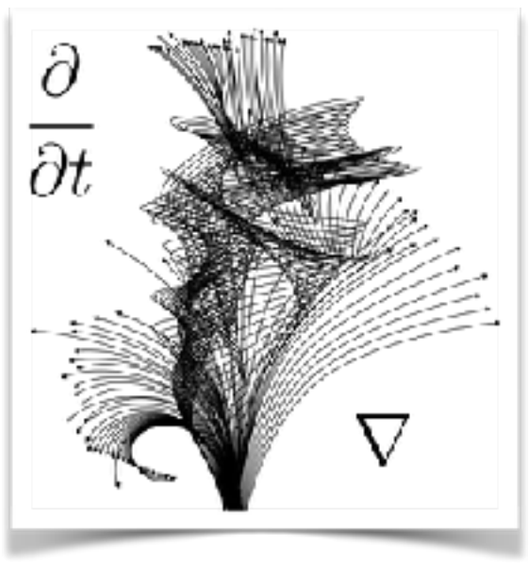
Applications

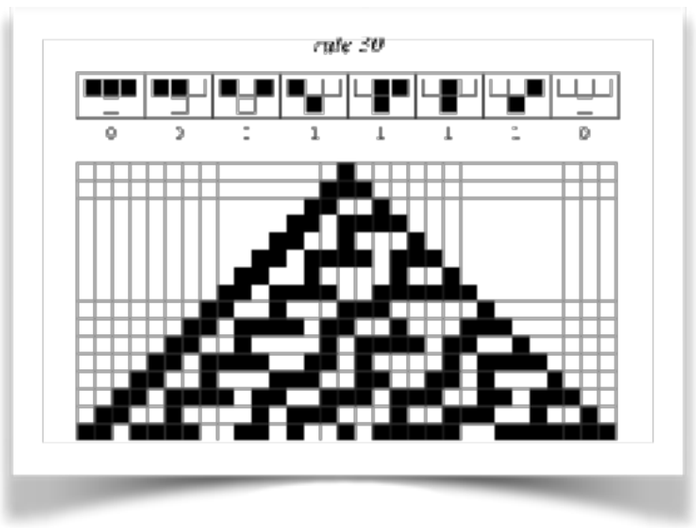


Iterative Stencil Algorithms

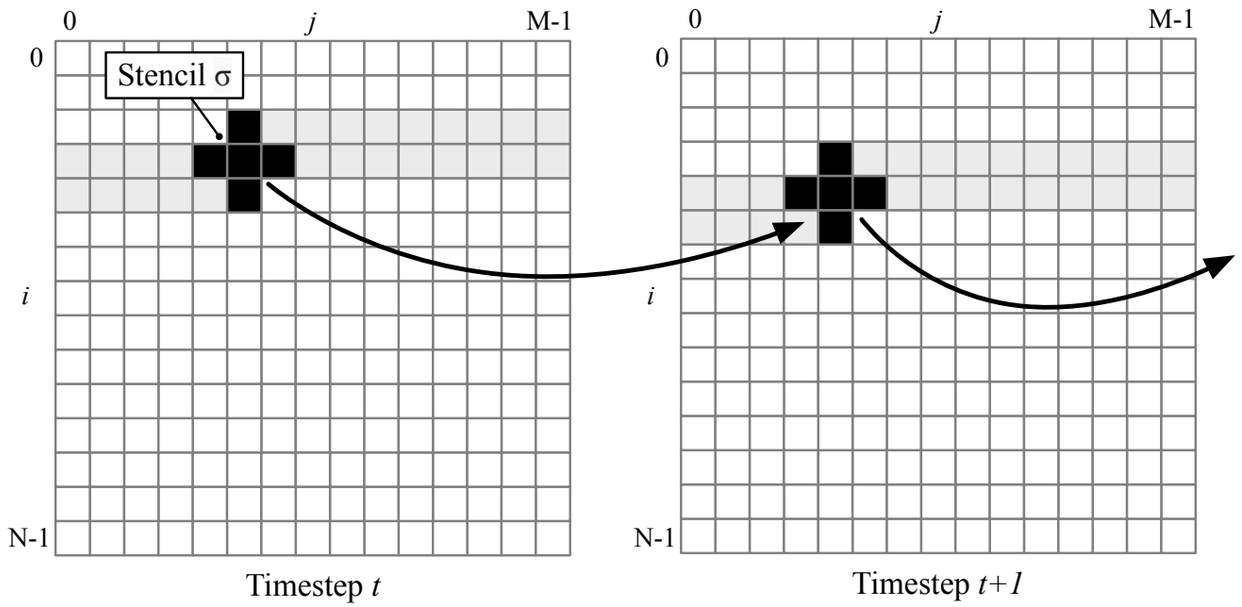


Applications

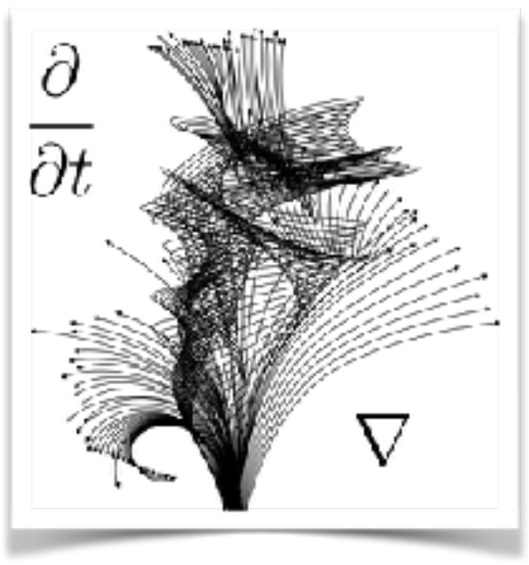


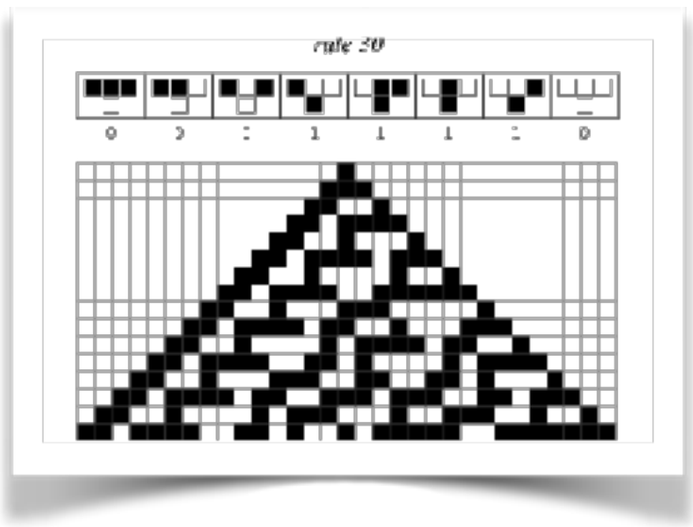


Iterative Stencil Algorithms



Applications



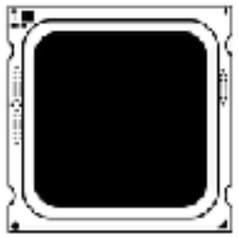


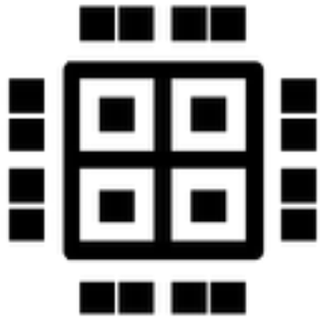
Rationale

Problems with Iterative Stencils

- Low Operational Intensity
- Synchronization between timesteps

Target Architectures



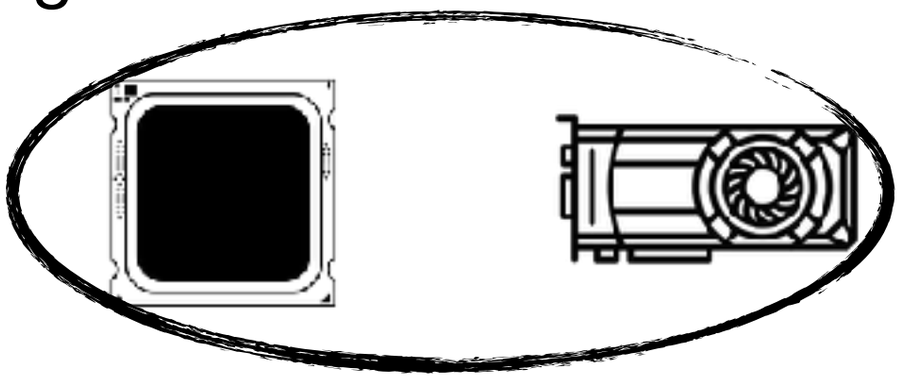


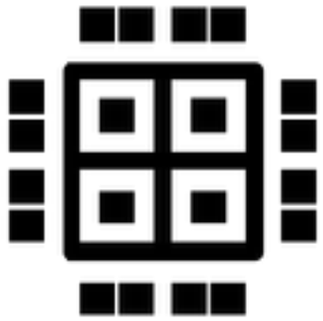
Rationale

Problems with Iterative Stencils

- Low Operational Intensity
- Synchronization between timesteps

Target Architectures



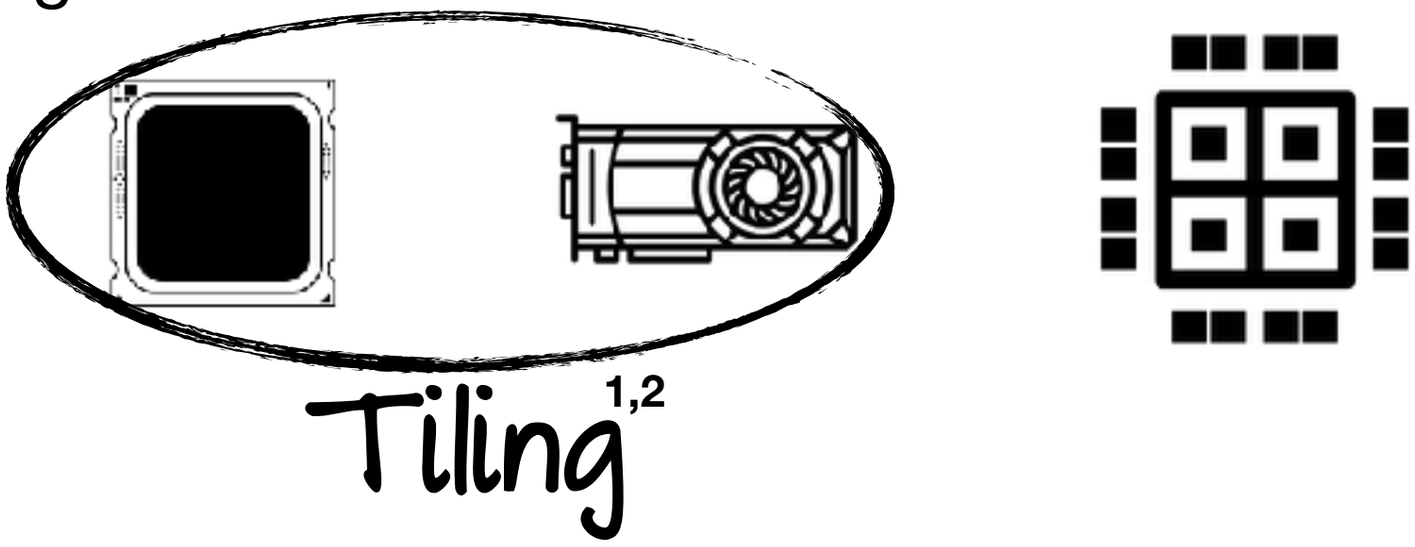


Rationale

Problems with Iterative Stencils

- Low Operational Intensity
- Synchronization between timesteps

Target Architectures

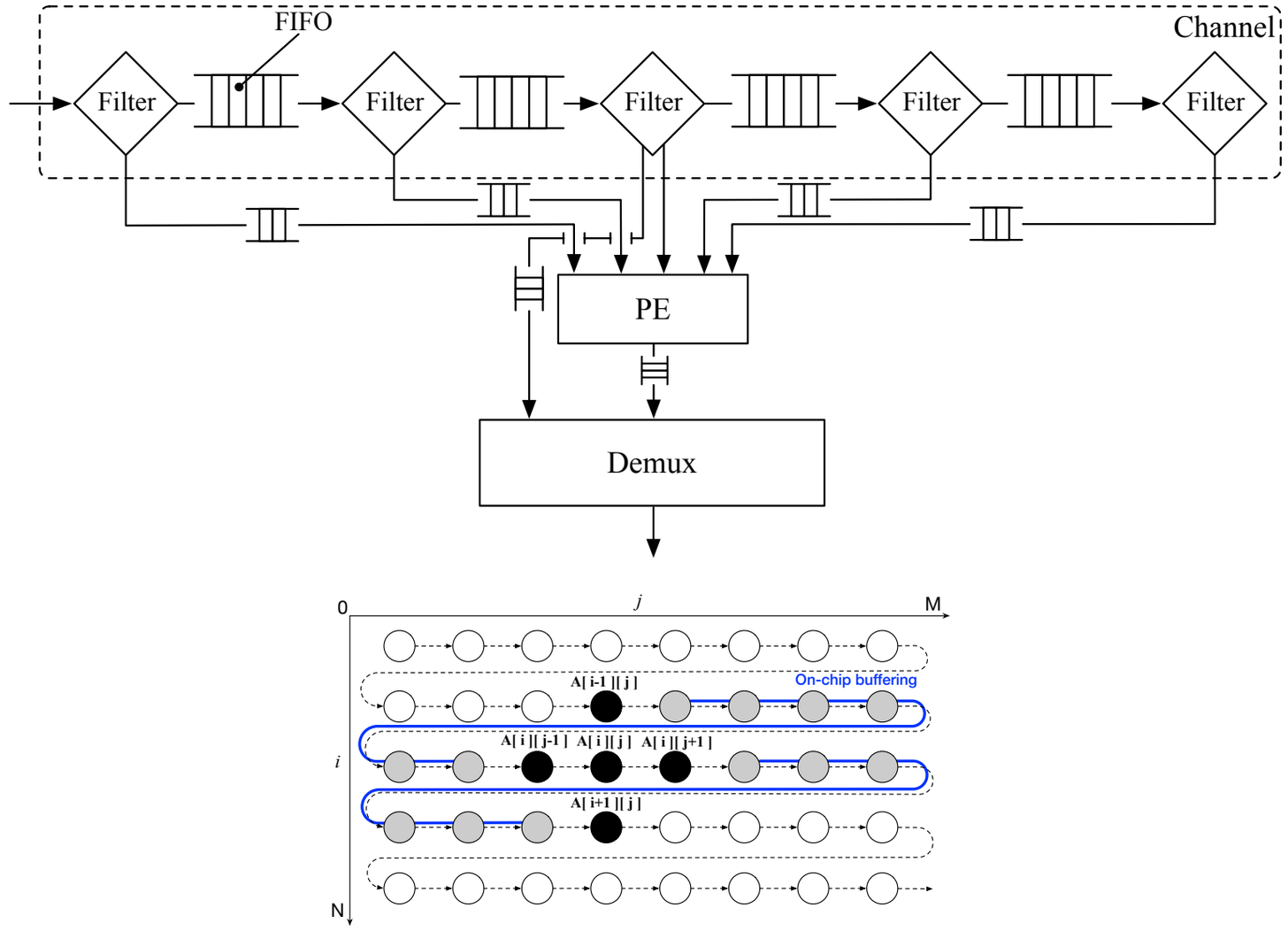


[1] V. Bandishti, I. Pananilath, U. Bodnhugula, **Tiling Stencil Computations to Maximize Parallelism.** SC 2012: 11

[2] J. Holewinski, L.-N. Pouchet, P. Sadayappan, **High-performance Code Generation for Stencil Computations on GPU Architectures.** ICS 2012

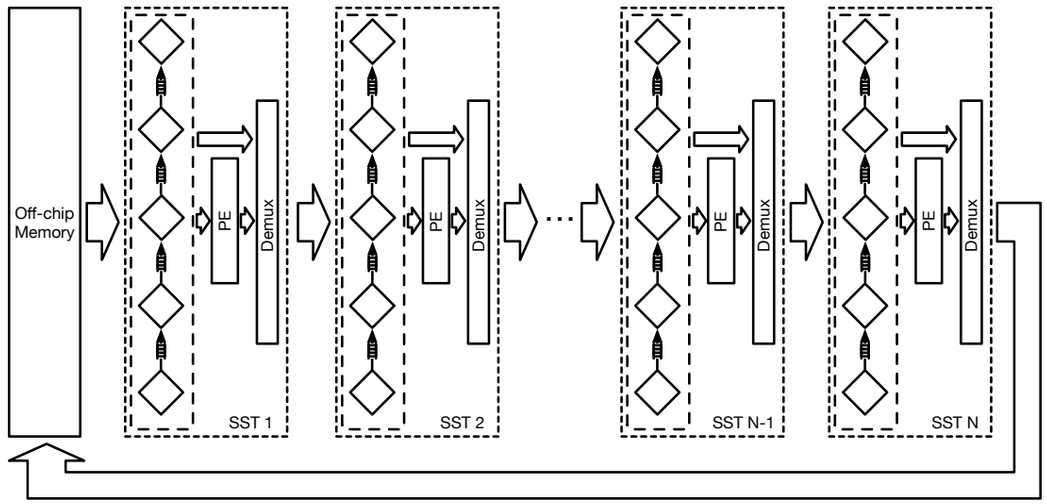
Previous work

- Microarchitecture for single stencil iteration: *Streaming Stencil Timestep*
  - Streaming computation
  - Dataflow blocks
  - Non-uniform memory partitioning



G. Natale, G. Stramondo, P. Bressana, R. Cattaneo, D. Sciuto, M. D. Santambrogio,
A polyhedral model-based framework for dataflow implementation on FPGA devices of iterative stencil loops. ICCAD 2016: 77

Previous work



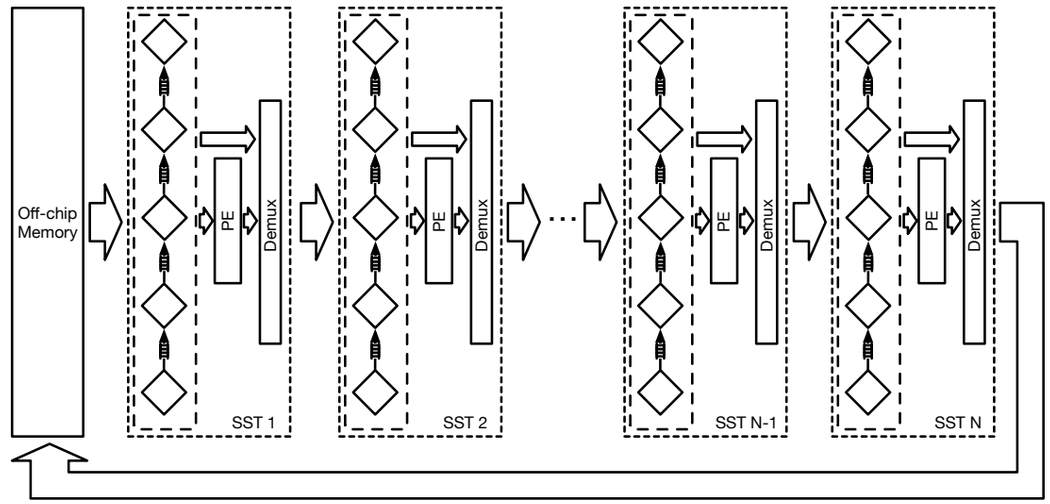
G. Natale, G. Stramondo, P. Bressana, R. Cattaneo, D. Sciuto, M. D. Santambrogio,
A polyhedral model-based framework for dataflow implementation on FPGA devices of iterative stencil loops. ICCAD 2016: 77

Previous work

- Complete accelerator:
  Chain of SSTs

- Constant off-chip BW requirements

- Dataflow Pipelining

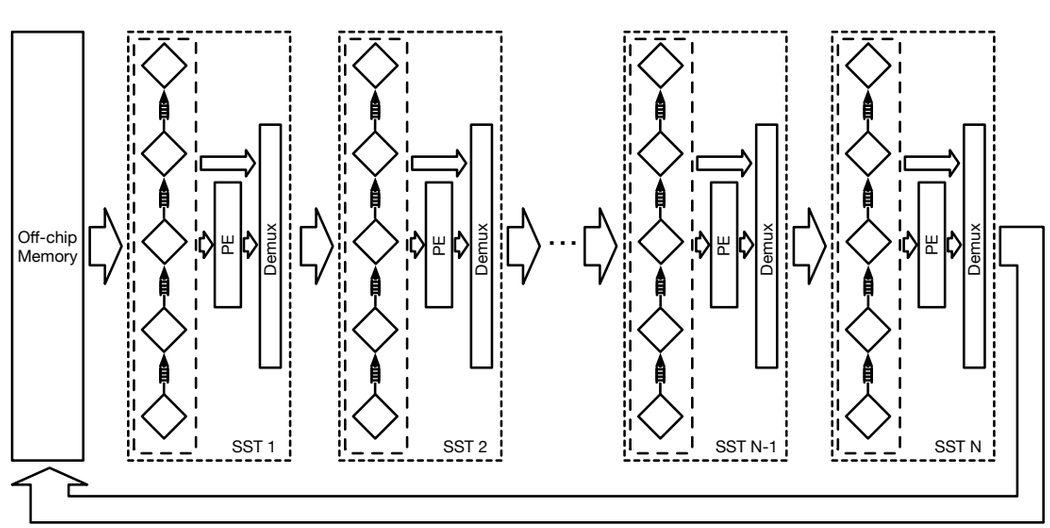


G. Natale, G. Stramondo, P. Bressana, R. Cattaneo, D. Sciuto, M. D. Santambrogio,
A polyhedral model-based framework for dataflow implementation on FPGA devices of iterative stencil loops. ICCAD 2016: 77

Previous work

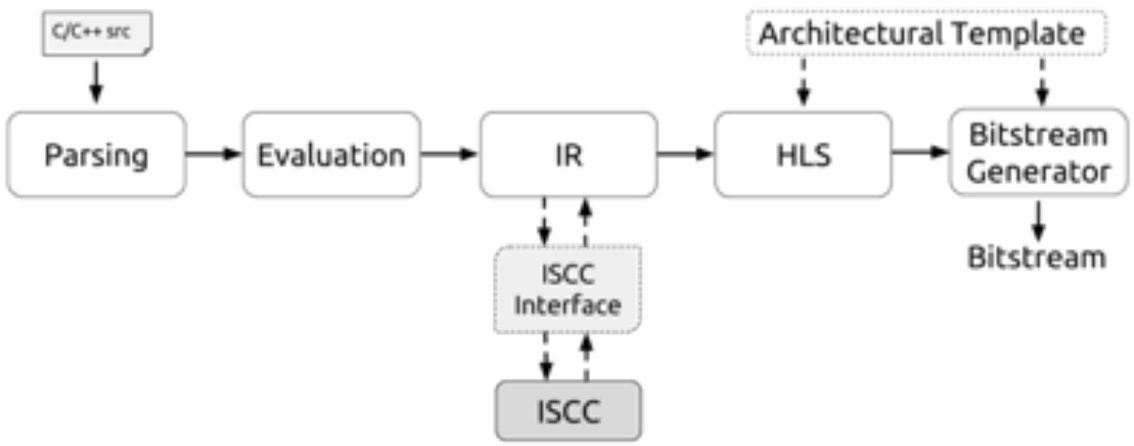
Complete accelerator: *Chain of SSTs*

- Constant off-chip BW requirements
- Dataflow Pipelining



Electronic Design Automation Framework

- Based on the polyhedral model
- Relies on High Level Synthesis
- Supported languages: C/C++



G. Natale, G. Stramondo, P. Bressana, R. Cattaneo, D. Sciuto, M. D. Santambrogio,
A polyhedral model-based framework for dataflow implementation on FPGA devices of iterative stencil loops. ICCAD 2016: 77

Contributions Overview

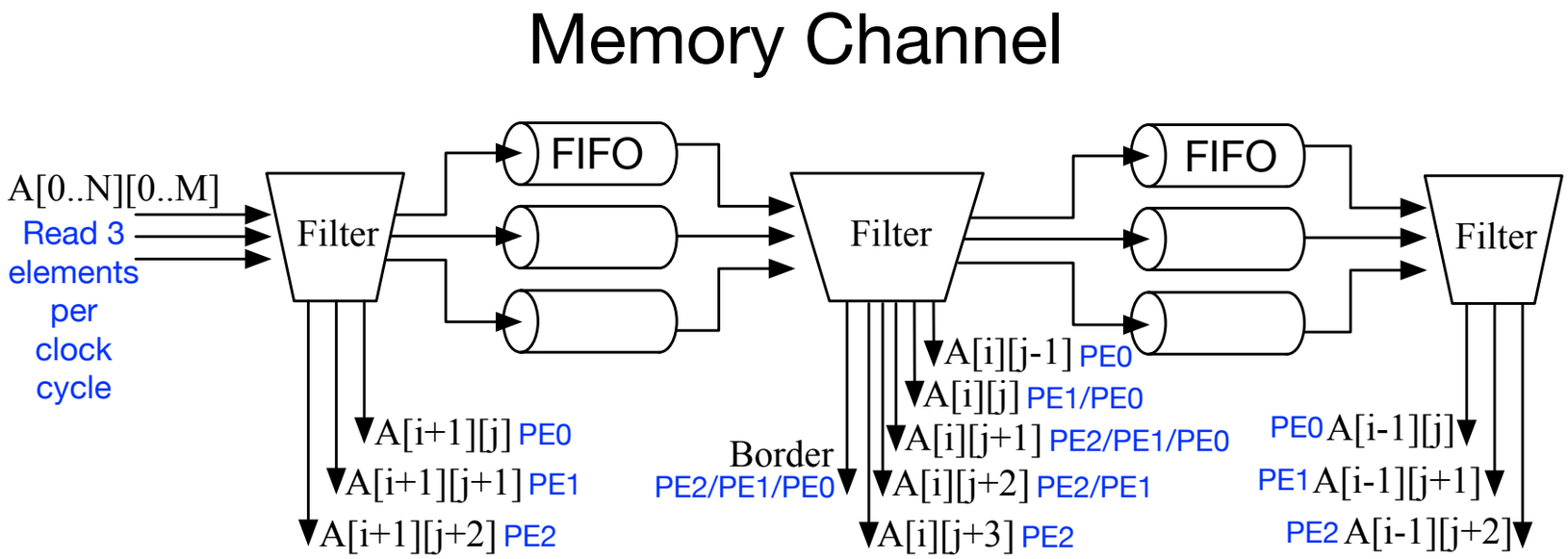
Issues with previous work

- Relies on HLS
- No intra-iterations parallelism

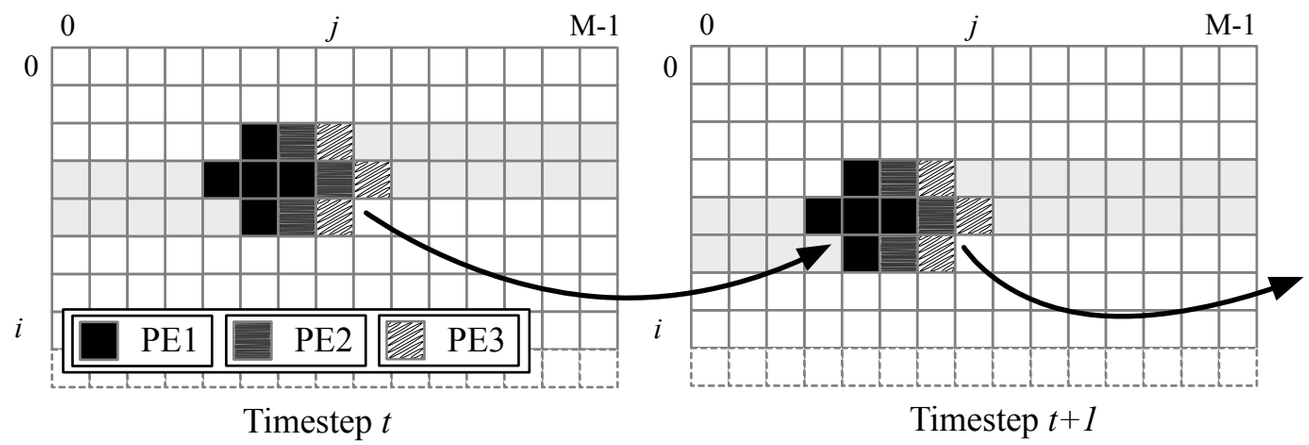
Proposed Improvements

- HDL-based design
- Intra-iterations parallelization strategy
- Performance model

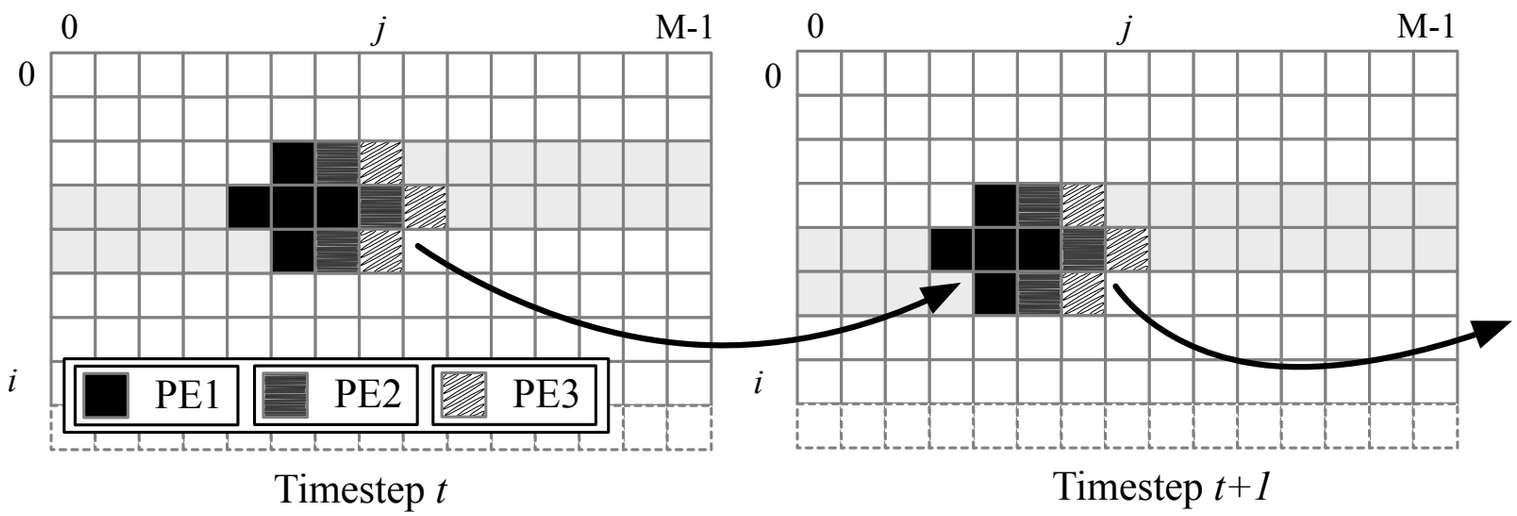
Intra-iterations parallelization



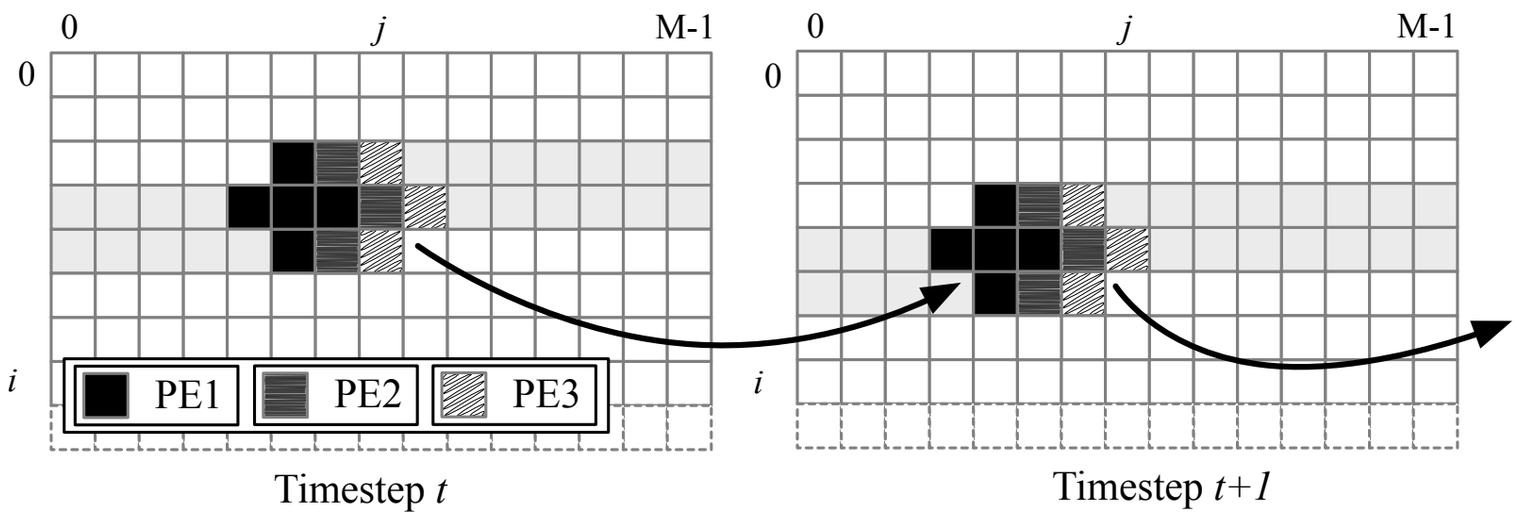
Parallelization Pattern: consecutive updates



Intra-iterations parallelization

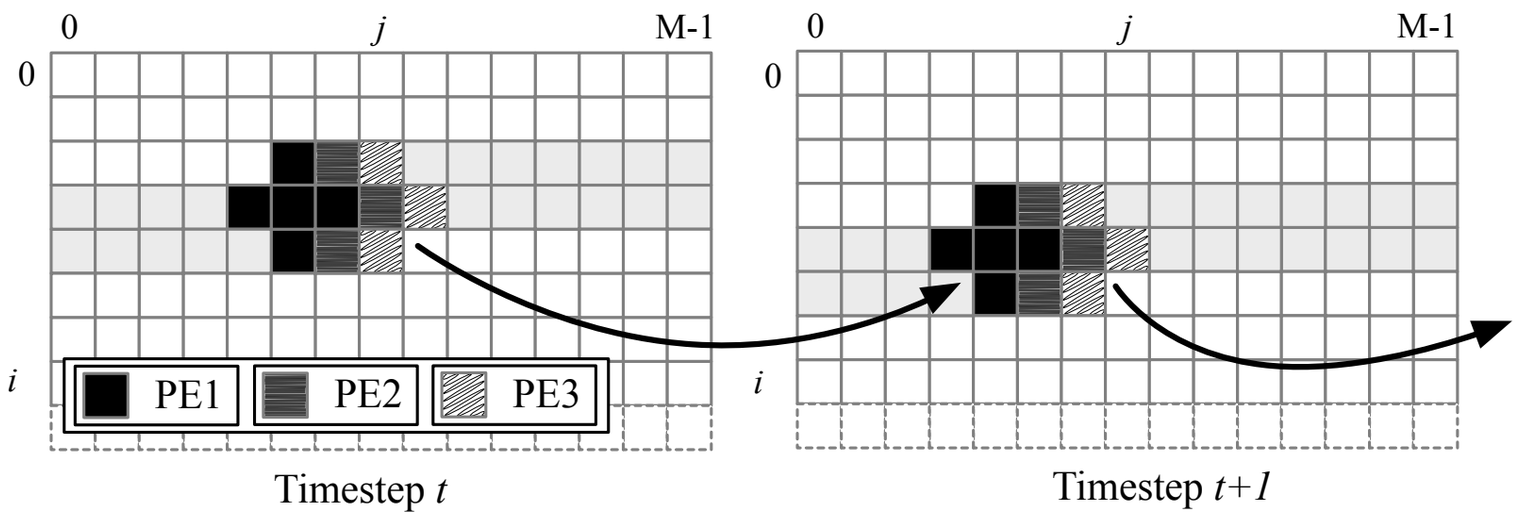


Intra-iterations parallelization



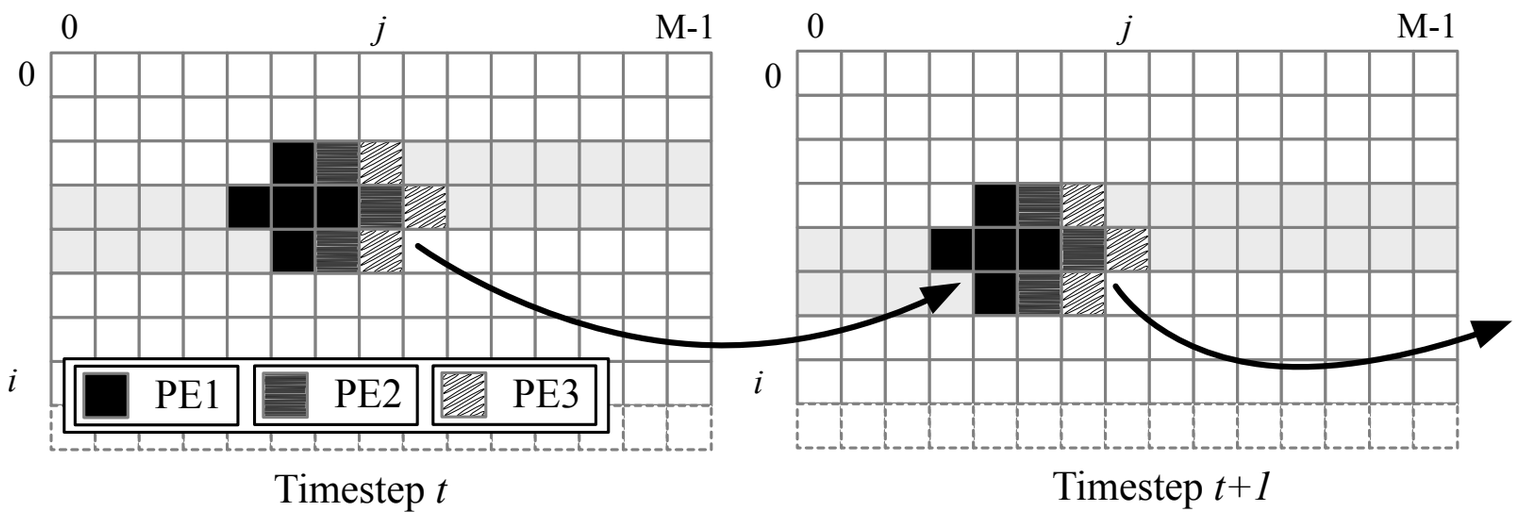
- Minimum impact on-chip memory requirements

Intra-iterations parallelization



- Minimum impact on-chip memory requirements
- Data reuse among PEs

Intra-iterations parallelization



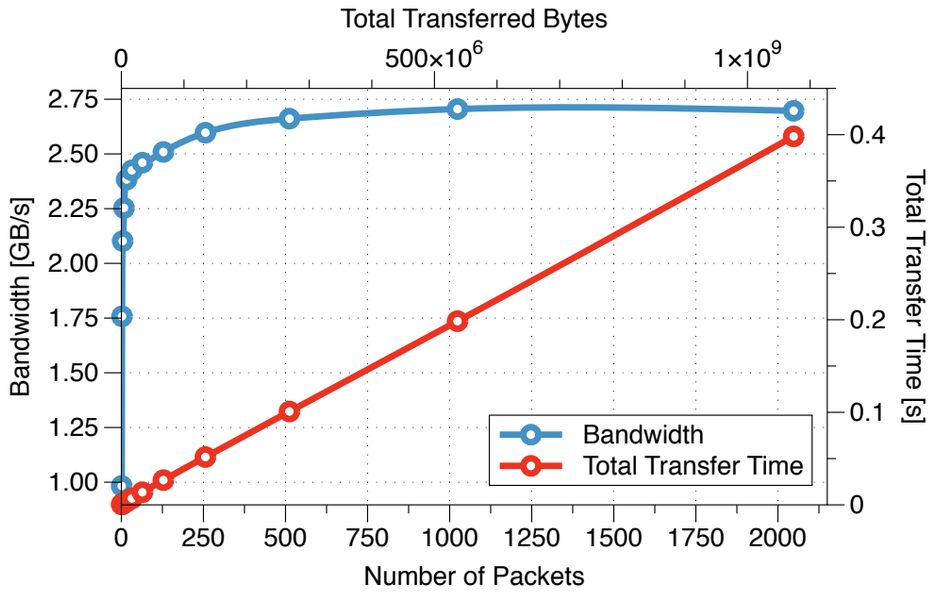
- Minimum impact on-chip memory requirements
- Data reuse among PEs
- Maximize off-chip BW usage

Performance Model

$$\tau = \max(\tau_{\text{transf}}, \tau_{\text{comp}})$$

$$\Delta = p_{\text{len}} \cdot n_p$$

$$n_p = \left\lceil \frac{\Delta}{p_{\text{len}}} \right\rceil$$

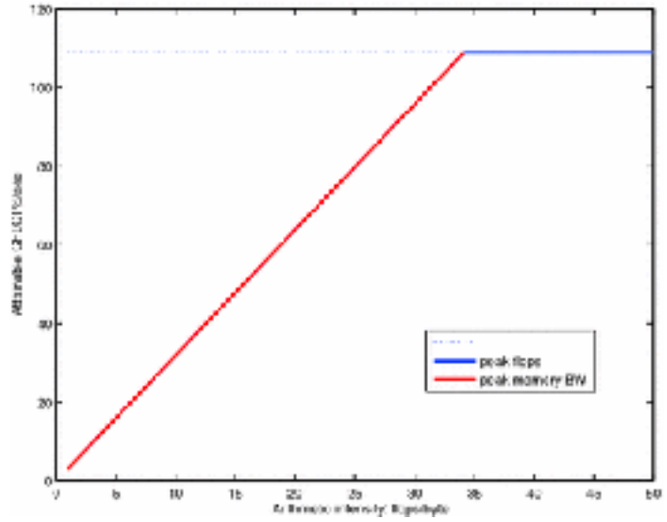


$$\tau_{\text{comp}} = \frac{(L_{\text{mem}} + L_{\text{PE}}) \cdot q + (\text{II}_{\text{filter}} + \text{II}_{\text{PE}}) \cdot \frac{N}{\lambda}}{\Phi}$$

$$L_{\text{mem}} = L_{\text{filter}} + \frac{\text{II}_{\text{filter}} \cdot \beta_{\text{size}}}{\lambda}$$

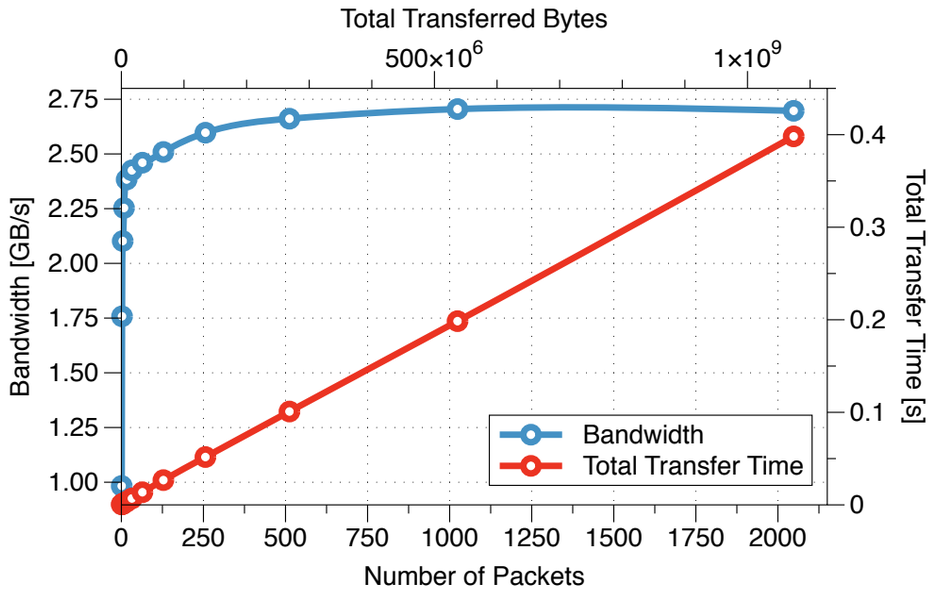
Performance Model

$$\tau = \max(\tau_{\text{transf}}, \tau_{\text{comp}})$$



$$\Delta = p_{\text{len}} \cdot n_p$$

$$n_p = \left\lceil \frac{\Delta}{p_{\text{len}}} \right\rceil$$



$$\tau_{\text{comp}} = \frac{(L_{\text{mem}} + L_{\text{PE}}) \cdot q + (\text{II}_{\text{filter}} + \text{II}_{\text{PE}}) \cdot \frac{N}{\lambda}}{\Phi}$$

$$L_{\text{mem}} = L_{\text{filter}} + \frac{\text{II}_{\text{filter}} \cdot \beta_{\text{size}}}{\lambda}$$

Experimental Setup

- Xilinx VC707 board, Virtex 7 FPGA, PCI-e 2.0 X8
- 250 MHz Target Frequency (200 MHz prev. work)
- 2 GB/s BW (800 MB/s prev. work)

Benchmark	Source
Jacobi-2D	Polybench [1]
Jacobi-3D	Parboil [2]
Heat-3D	Pluto [3]
3d7pt	Pluto [3]
American Put Option	Pluto [3]
Game of Life	Pluto [3]

[1] S. Grauer-Gray, L. Xu et al., **Auto-tuning a high-level language targeted to GPU codes**. InPar 2012

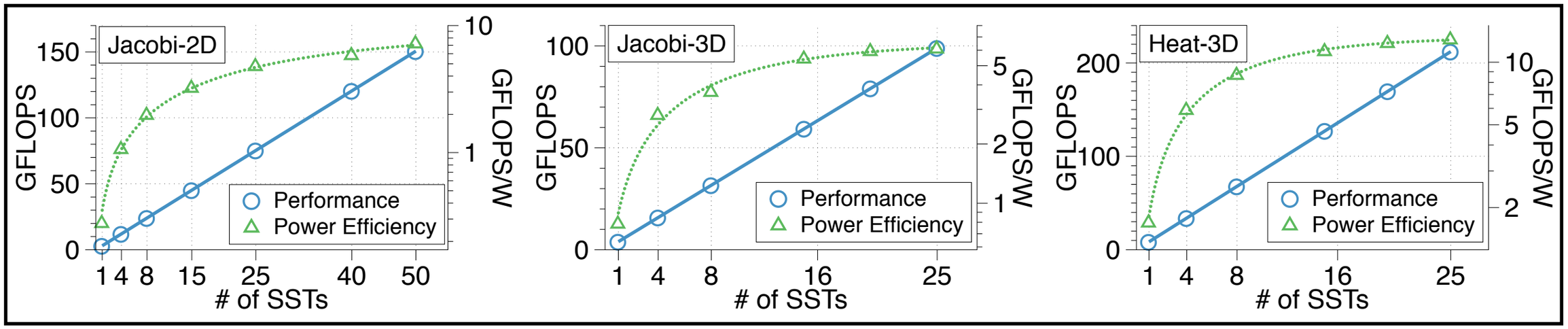
[2] J. A. Stratton et al., **Parboil: A revised benchmark suite for scientific and commercial throughput computing.** Center for Reliable and High-Performance Computing, vol 127, 2012

[3] V. Bandishti, I. Pananilath, U. Bodnhugula, **Tiling Stencil Computations to Maximize Parallelism.** SC 2012: 11

Results (1)

Benchmark	[1] Timesteps Queued	[1] GFLOPS	[1] GFLOPS/W	This Work Timesteps Queued	This Work GFLOPS	This Work GFLOPS/W	Parallel.
Jacobi-2D	72	31.20	4.65	50	150.63	7.28	4
Jacobi-3D	48	5.69	0.82	25	99.00	6.12	4
Heat-3D	32	5.42	0.69	15	212.16	12.89	4
3d7pt	8	3.396	1.04	24	108.63	6.30	4
American Put Option	–	–	–	200	209.89	9.46	1
		GOPS	**GOPS/W**		**GOPS**	**GOPS/W**	
Game of Life	48	29.60	4.65	112	472.38	22.98	4

Performance and Energy Efficiency Trend



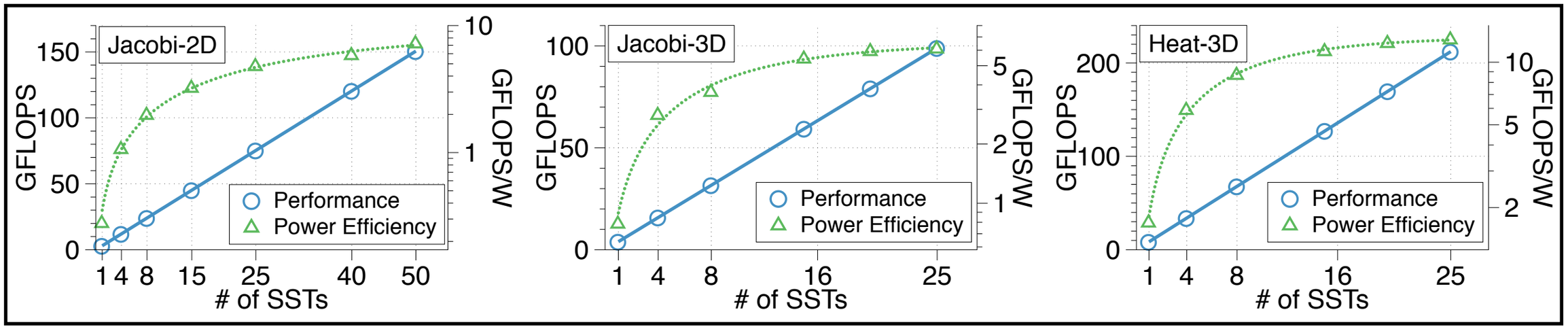
[1] G. Natale, G. Stramondo, P. Bressana, R. Cattaneo, D. Sciuto, M. D. Santambrogio,
A polyhedral model-based framework for dataflow implementation on FPGA devices of iterative stencil loops. ICCAD 2016: 77

Results (1)

Benchmark	[1] Timesteps Queued	[1] GFLOPS	[1] GFLOPS/W	This Work Timesteps Queued	This Work GFLOPS	This Work GFLOPS/W	This Work Parallel.
Jacobi-2D	72	31.20	4.65	50	150.63	7.28	4
Jacobi-3D	48	5.69	0.82	25	99.00	6.12	4
Heat-3D	32	5.42	0.69	15	212.16	12.89	4
3d7pt	8	3.396	1.04	24	108.63	6.30	4
American Put Option	–	–	–	200	209.89	9.46	1
		GOPS	**GOPS/W**		**GOPS**	**GOPS/W**	
Game of Life	48	29.60	4.65	112	472.38	22.98	4

~22x Performance

Performance and Energy Efficiency Trend



[1] G. Natale, G. Stramondo, P. Bressana, R. Cattaneo, D. Sciuto, M. D. Santambrogio,
A polyhedral model-based framework for dataflow implementation on FPGA devices of iterative stencil loops. ICCAD 2016: 77

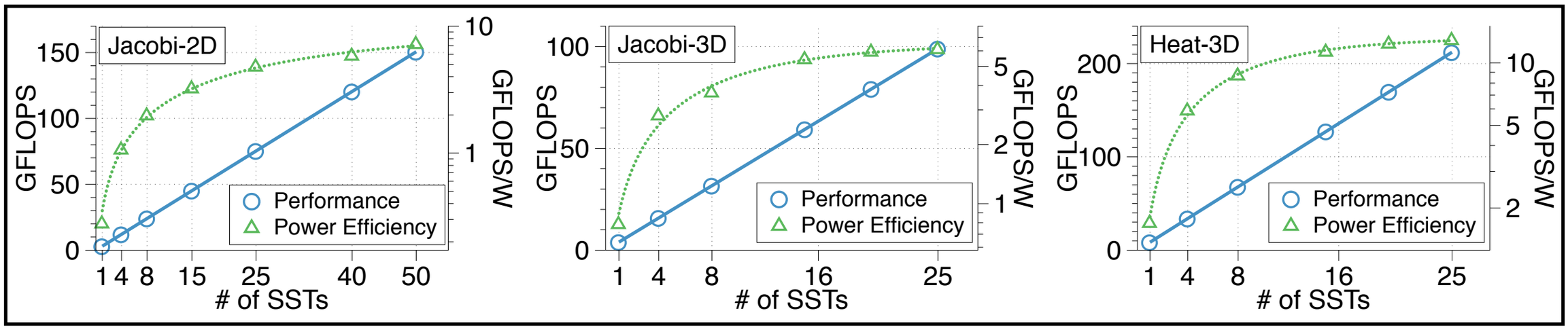
Results (1)

Benchmark	[1] Timesteps Queued	[1] GFLOPS	[1] GFLOPS/W	This Work Timesteps Queued	This Work GFLOPS	This Work GFLOPS/W	This Work Parallel.
Jacobi-2D	72	31.20	4.65	50	150.63	7.28	4
Jacobi-3D	48	5.69	0.82	25	99.00	6.12	4
Heat-3D	32	5.42	0.69	15	212.16	12.89	4
3d7pt	8	3.396	1.04	24	108.63	6.30	4
American Put Option	–	–	–	200	209.89	9.46	1
		GOPS	**GOPS/W**		**GOPS**	**GOPS/W**	
Game of Life	48	29.60	4.65	112	472.38	22.98	4

~22x Performance

~8x Power Efficiency

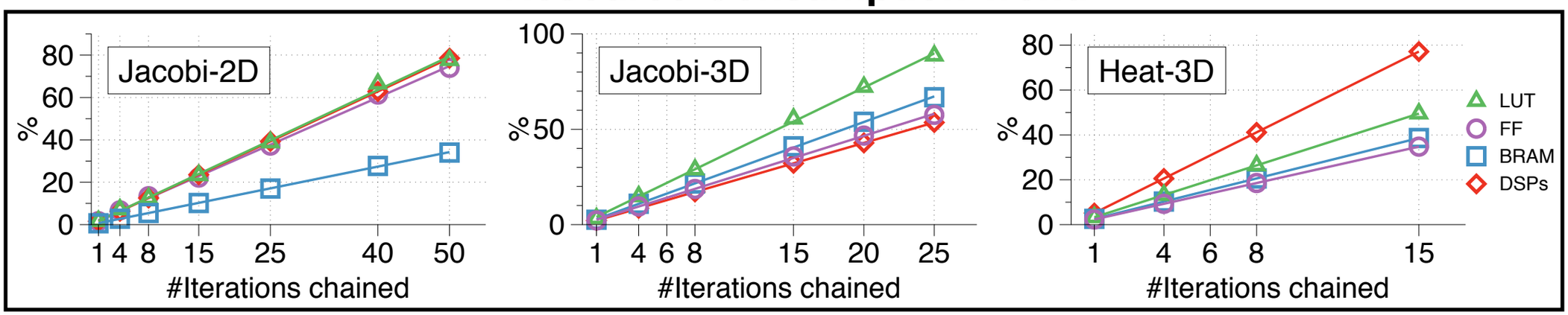
Performance and Energy Efficiency Trend



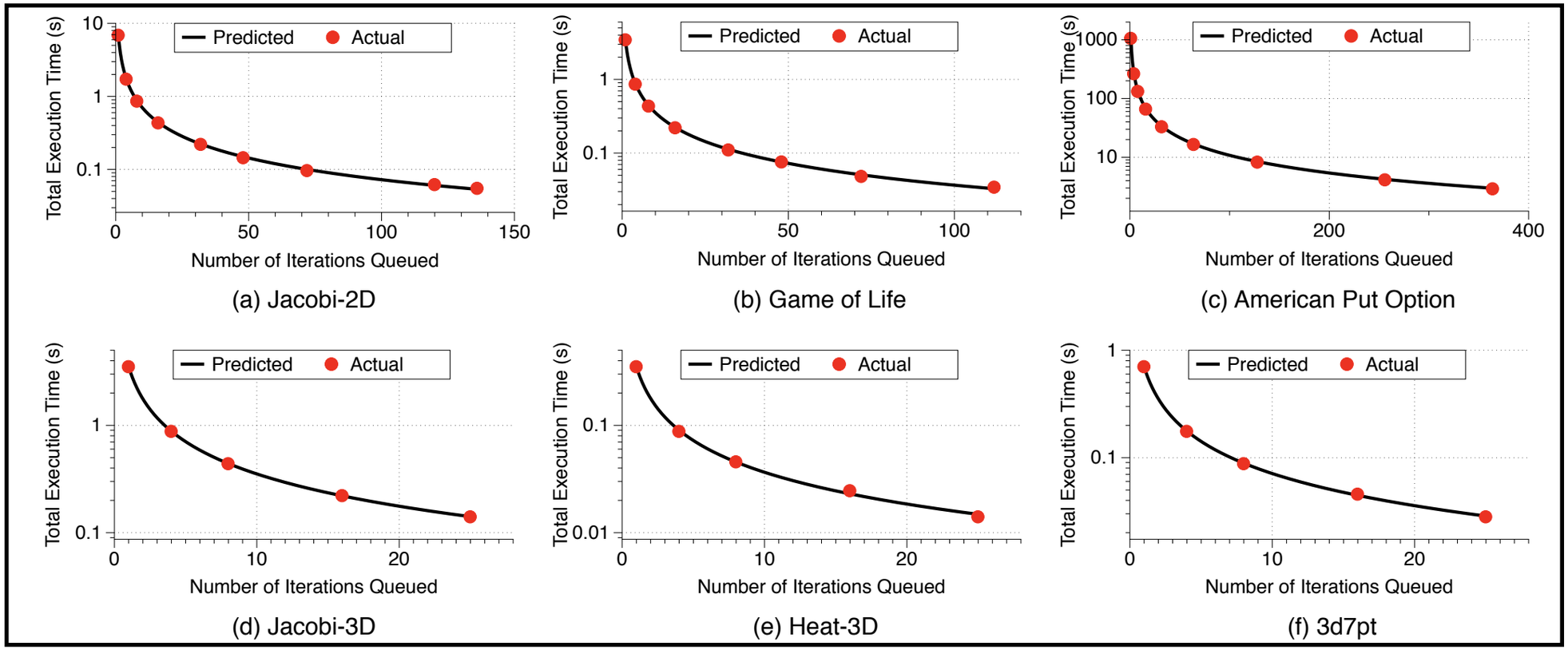
[1] G. Natale, G. Stramondo, P. Bressana, R. Cattaneo, D. Sciuto, M. D. Santambrogio,
A polyhedral model-based framework for dataflow implementation on FPGA devices of iterative stencil loops. ICCAD 2016: 77

Results (2)

Resource Consumption Trend



Model Accuracy



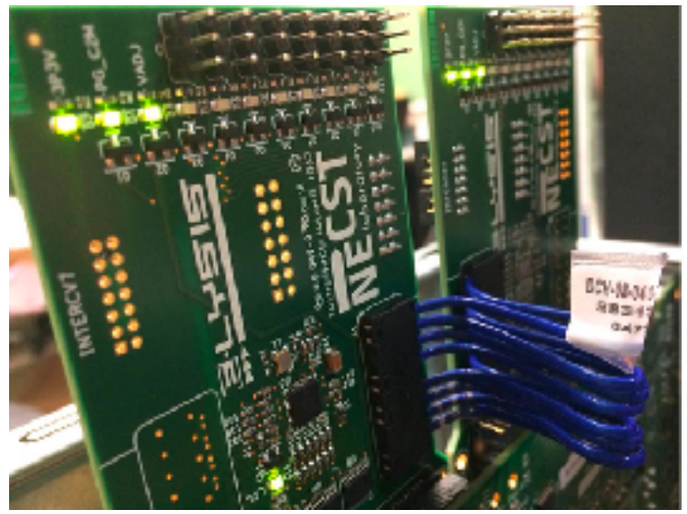
Conclusions

Acceleration Methodology for Iterative Stencils on FPGAs

- Exploit Intra and inter-iterations parallelism
- Efficient on-chip storage
- Performance Model

Future Work

- Scale on a multi-FPGA system using custom interconnection boards designed in collaboration with Elysis



Slides will be available @

www.slideshare.net/necstlab

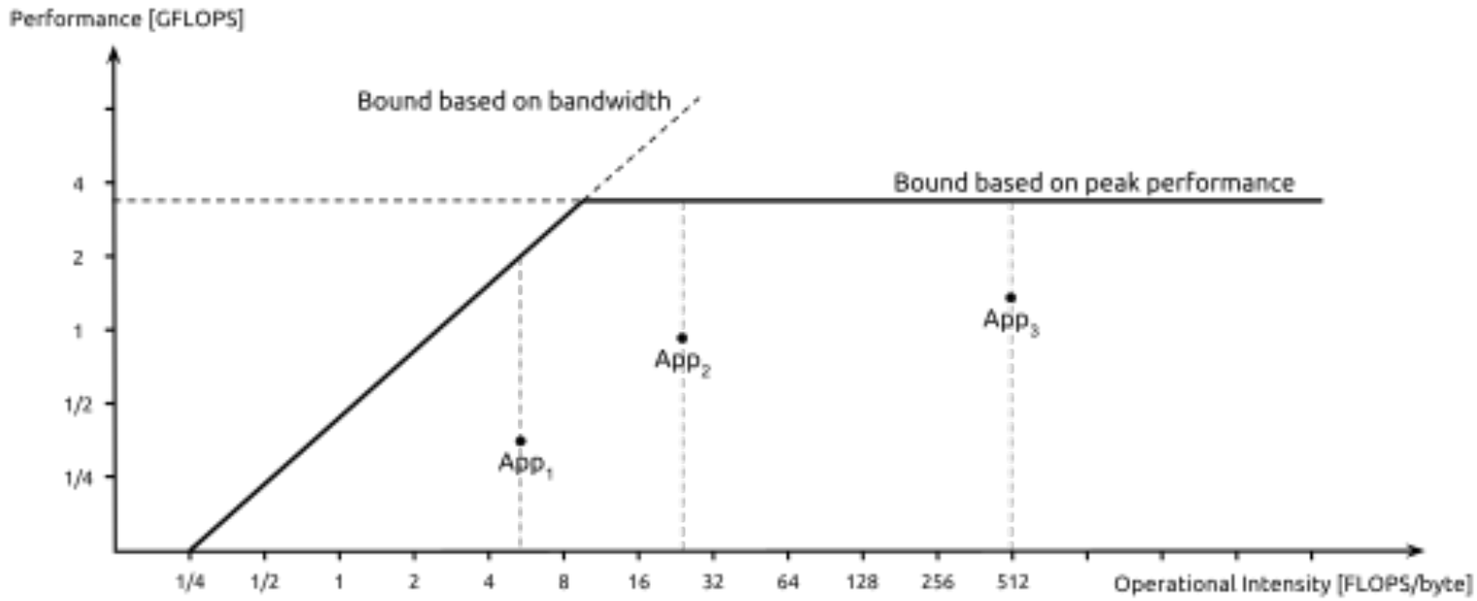
facebook.com/groups/ReconfigurableArchitecturesWorkshop

An FPGA-based Acceleration Methodology and Performance Model for Iterative Stencils

Enrico Reggiani, Giuseppe Natale, Carlo Moroni, Marco D. Santambrogio

Giuseppe Natale - giuseppe.natale@polimi.it

Roofline Model



$$P = \min \begin{cases} \pi \\ \beta \times I \end{cases}$$

$$I = \frac{W}{Q}$$

P : attainable performance
π : peak performance
β : peak bandwidth
I : operational Intensity
W : work
Q : memory Traffic

Williams, Samuel, Andrew Waterman, and David Patterson. **Roofline: an insightful visual performance model for multicore architectures.** *Communications of the ACM* 52.4 (2009): 65-76.