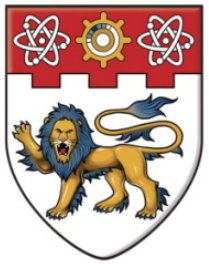


NANYANG TECHNOLOGICAL UNIVERSITY

DeCO: A DSP Block Based FPGA Accelerator Overlay With Low Overhead Interconnect

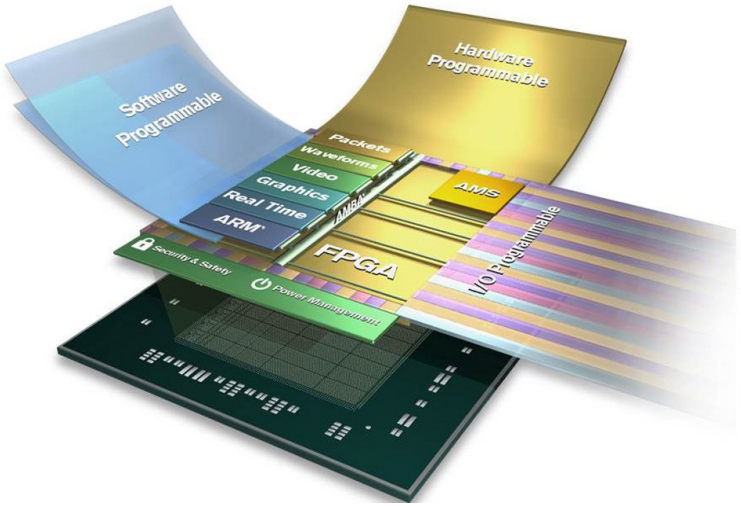
Abhishek Kumar Jain, Xiangwei Li, Pranjul Singhai, Douglas L. Maskell
School of Computer Science and Engineering
Nanyang Technological University (NTU), Singapore

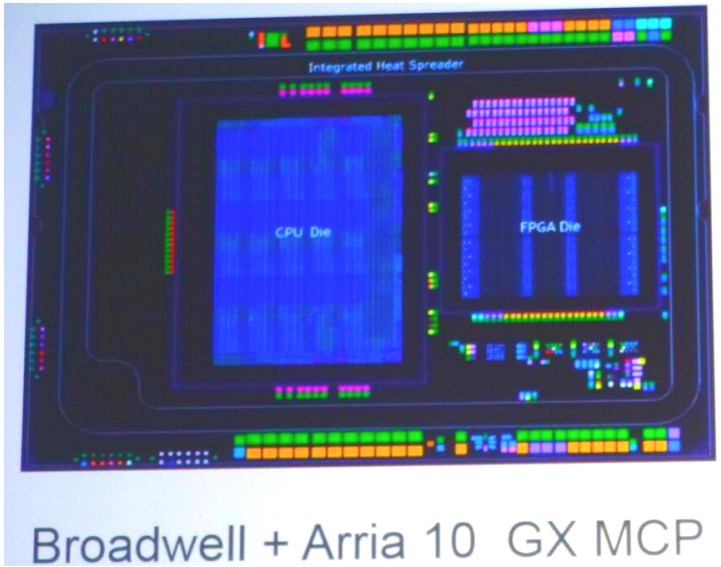
Suhaib A. Fahmy
School of Engineering
University of Warwick, UK

International Symposium on Field-Programmable Custom Computing Machines (FCCM)
2nd May 2016, Washington DC, USA

FPGAs in Heterogeneous Computing Platforms

- Xilinx: FPGAs coupled with ARM (Zynq UltraScale MPSoC)
  - 3500 DSP Blocks in the largest device
  - Peak performance of 5200 Giga-Operations Per Second (GOPS)
- Intel: FPGAs coupled with Xeon
  - 1500 floating point DSP Blocks in the largest device
  - Peak performance of 1300 GFLOPS





Are FPGAs really ready for the mainstream?

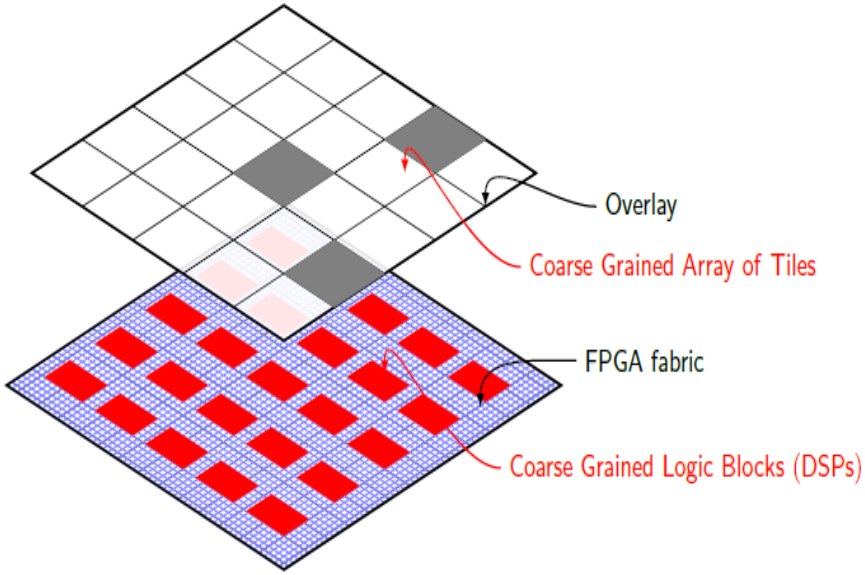
- No, Mainly due to poor design productivity issues
  - Accelerator design at RTL level -> Hardware design expertise

Are FPGAs really ready for the mainstream?

- No, Mainly due to poor design productivity issues
  - Accelerator design at RTL level -> Hardware design expertise
  - Long compilation times of RTL design

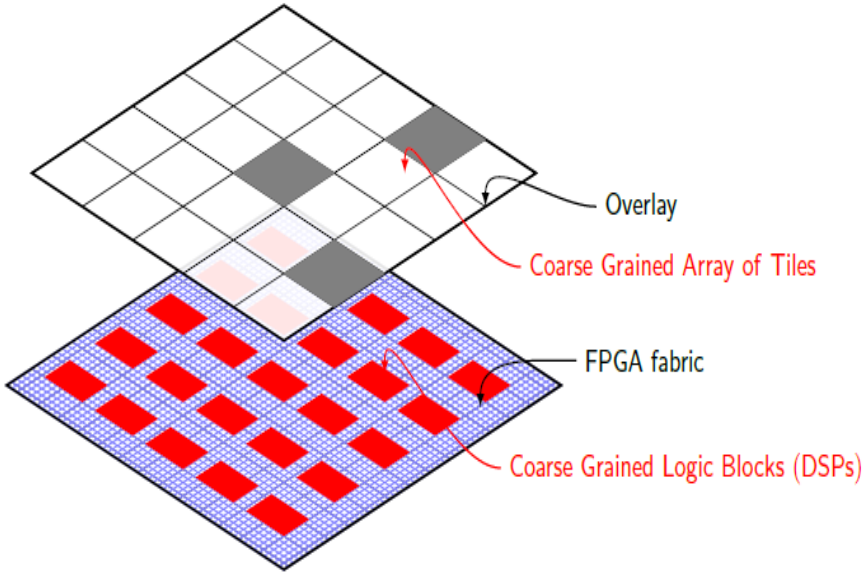
Coarse grained FPGA overlays

- Array of coarse-grained tiles
- Programmable functional unit and interconnect resources



Coarse grained FPGA overlays

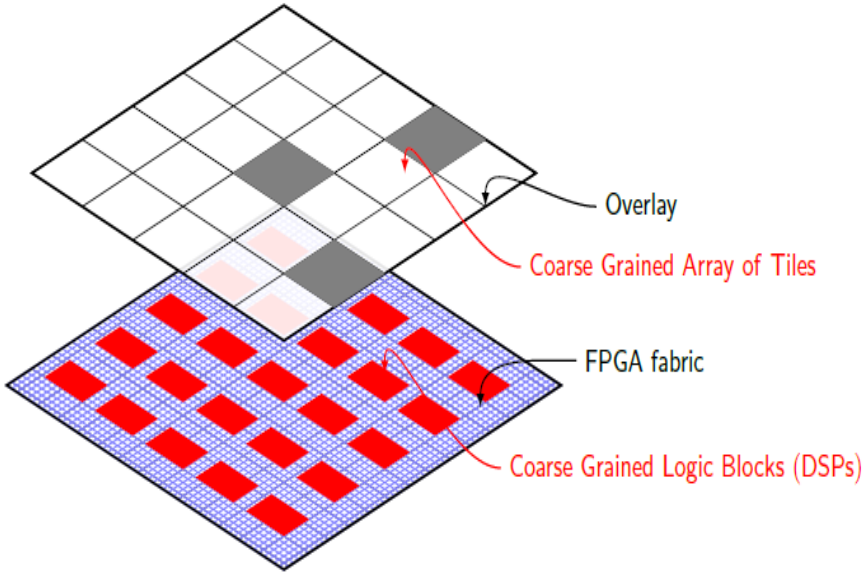
- Array of coarse-grained tiles
- Programmable functional unit and interconnect resources



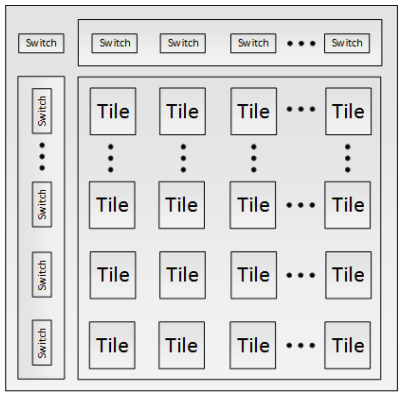
- Benefits:
  - Accelerator design at a higher level of abstraction
  - Fast compilation
  - Fast reconfiguration
  - Improved design productivity

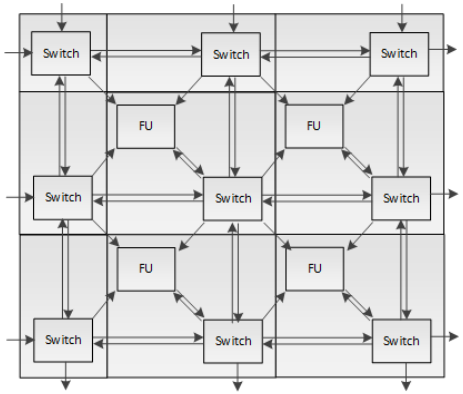
Coarse grained FPGA overlays

- Array of coarse-grained tiles
- Programmable functional unit and interconnect resources



- Benefits:
  - Accelerator design at a higher level of abstraction
  - Fast compilation
  - Fast reconfiguration
  - Improved design productivity





- The major ISSUE is the area and performance overheads

NANYANG TECHNOLOGICAL UNIVERSITY

Coarse grained FPGA overlays

- Two metrics
  - Interconnect area overhead in terms of LUTs/FU
  - Peak performance in terms of GOPS

Overlay	Interconnect Area Overhead	Peak performance
DSP-DySER [HEART2015]	1360 LUTs/FU	6.3 GOPS
DSP-based Island-style [FCCM2015]	437 LUTs/FU	65 GOPS

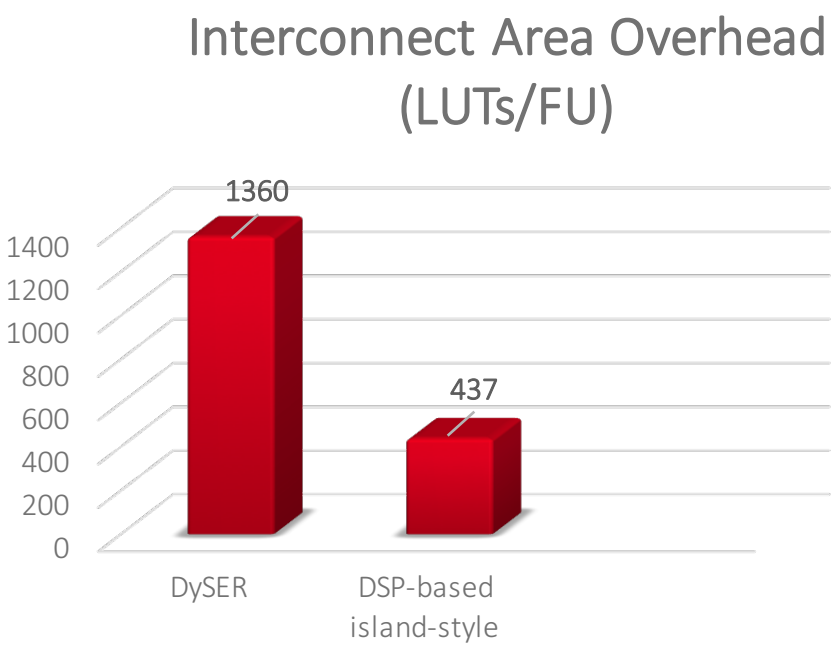
- 3x better (in area overhead)
- 10x better (in peak throughput)

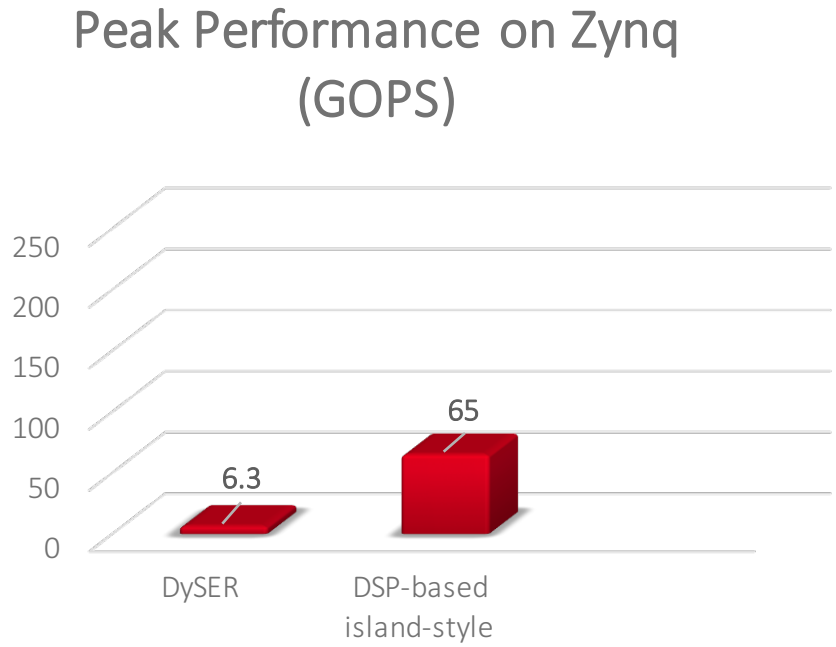
Issues

- Can improve further?
  - On Zynq, an array of 220 DSP blocks can provide 264 GOPS

Issues

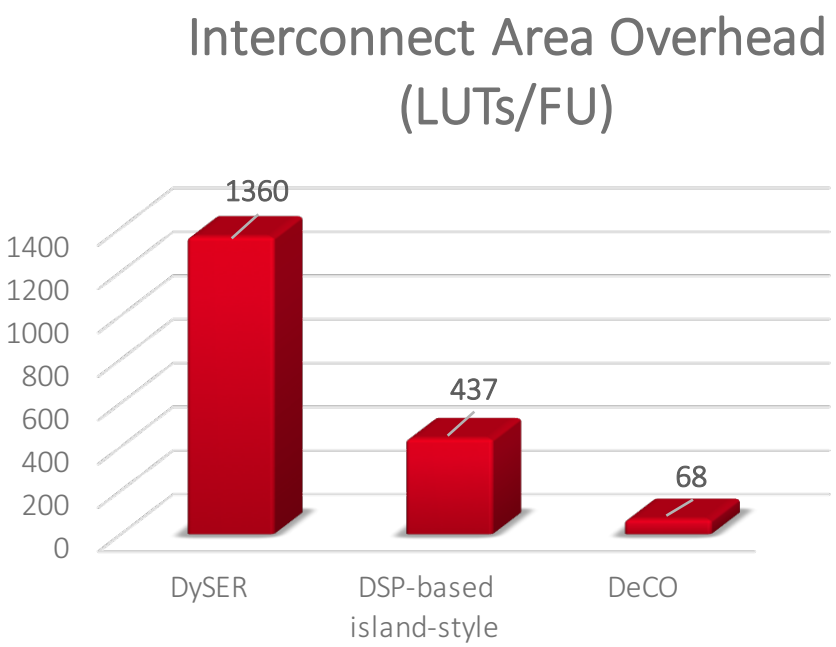
- Can improve further?
  - On Zynq, an array of 220 DSP blocks can provide 264 GOPS
  - Can we reduce interconnect area overhead further to achieve a higher peak performance out of DSP blocks?

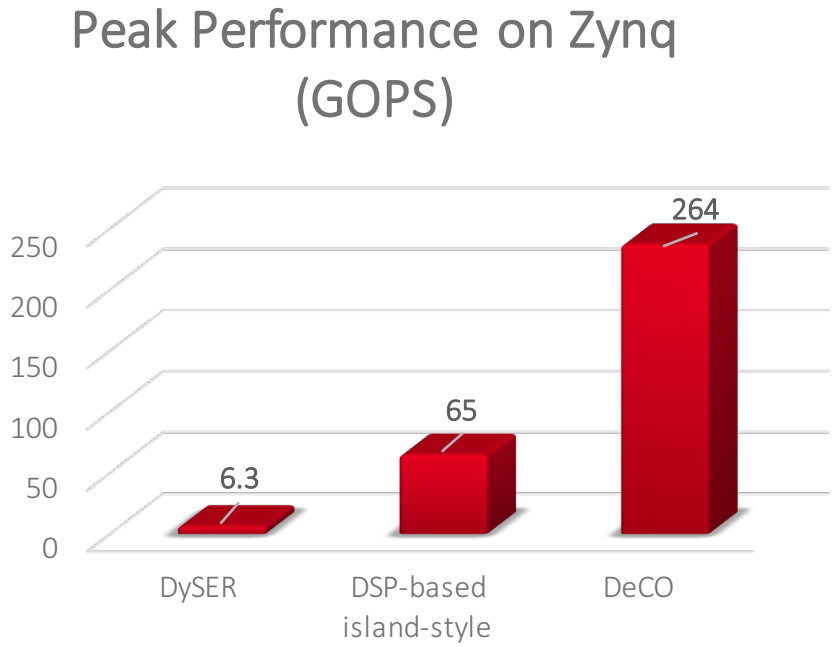




Issues

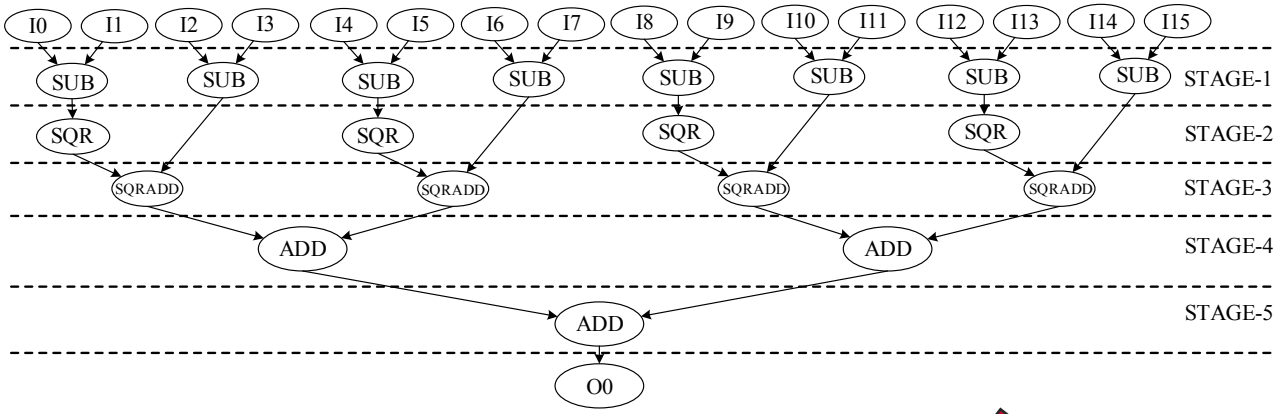
- Can improve further?
  - On Zynq, an array of 220 DSP blocks can provide 264 GOPS
  - Can we reduce interconnect area overhead further to achieve a higher peak performance out of DSP blocks?

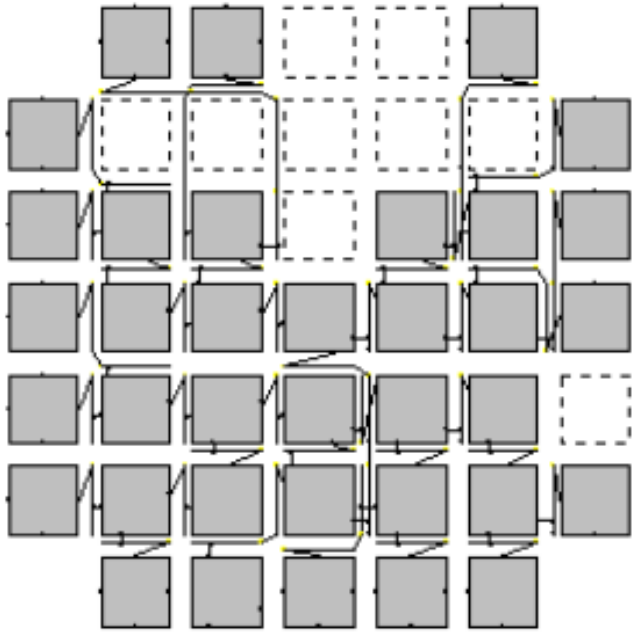




NANYANG TECHNOLOGICAL UNIVERSITY

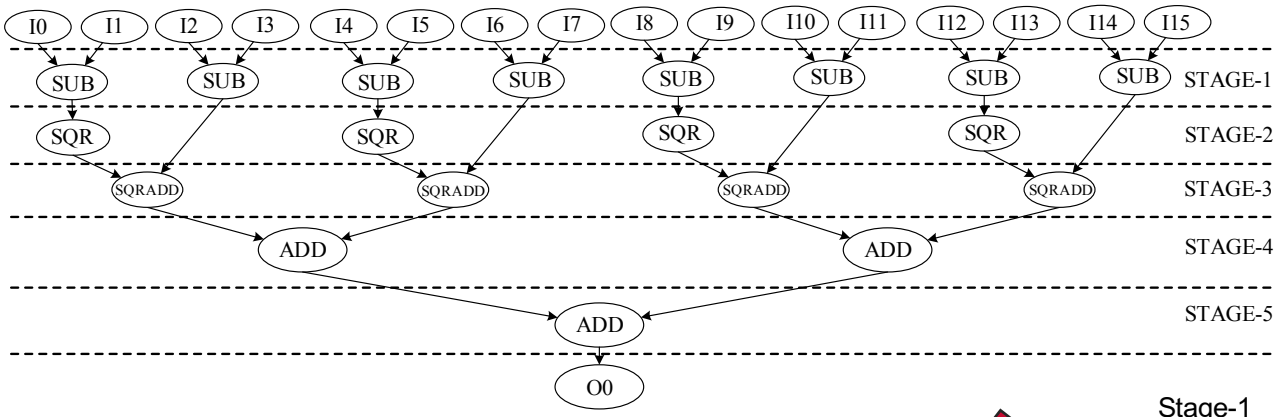
Approach

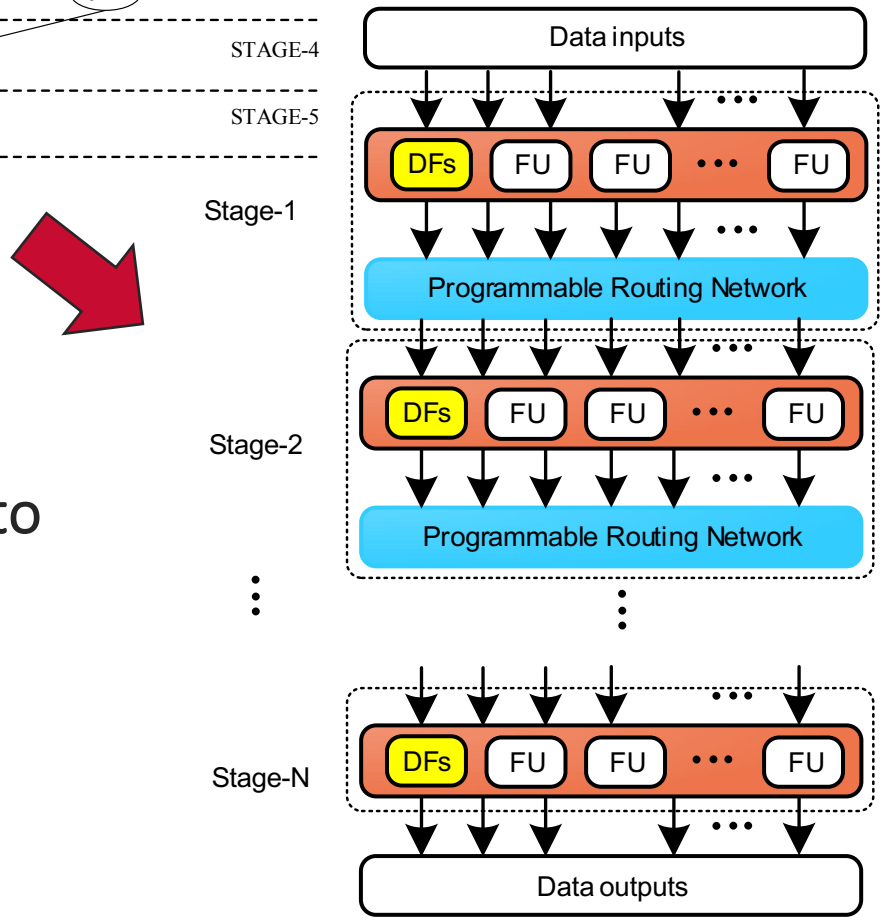




- Island-style interconnect allows communication between any FU to any other FU
- Not required for feed-forward compute kernels

Approach



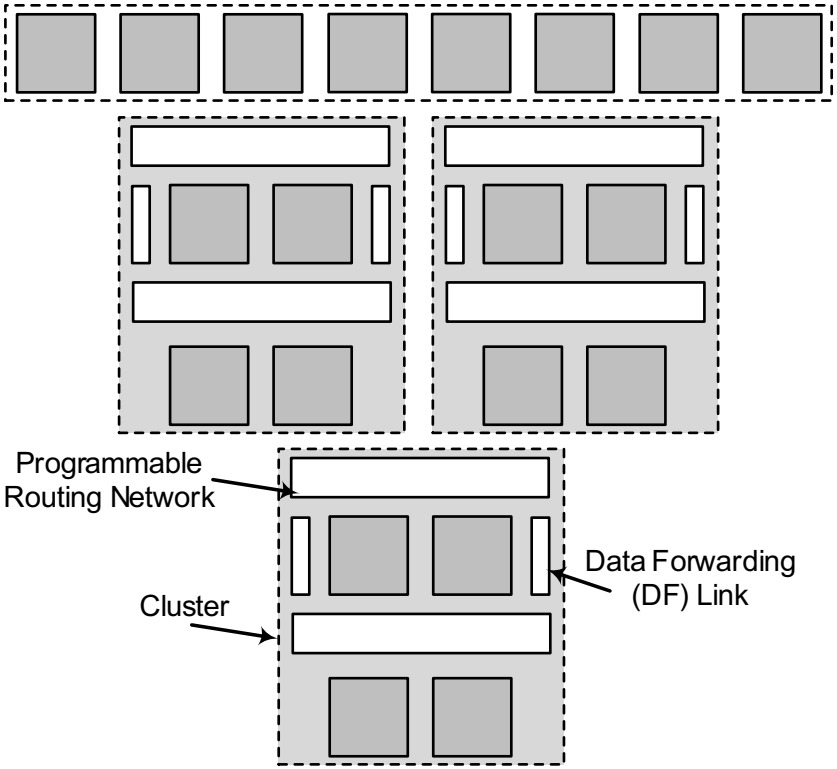


- Island-style interconnect allows communication between any FU to any other FU
- Not required for feed-forward compute kernels

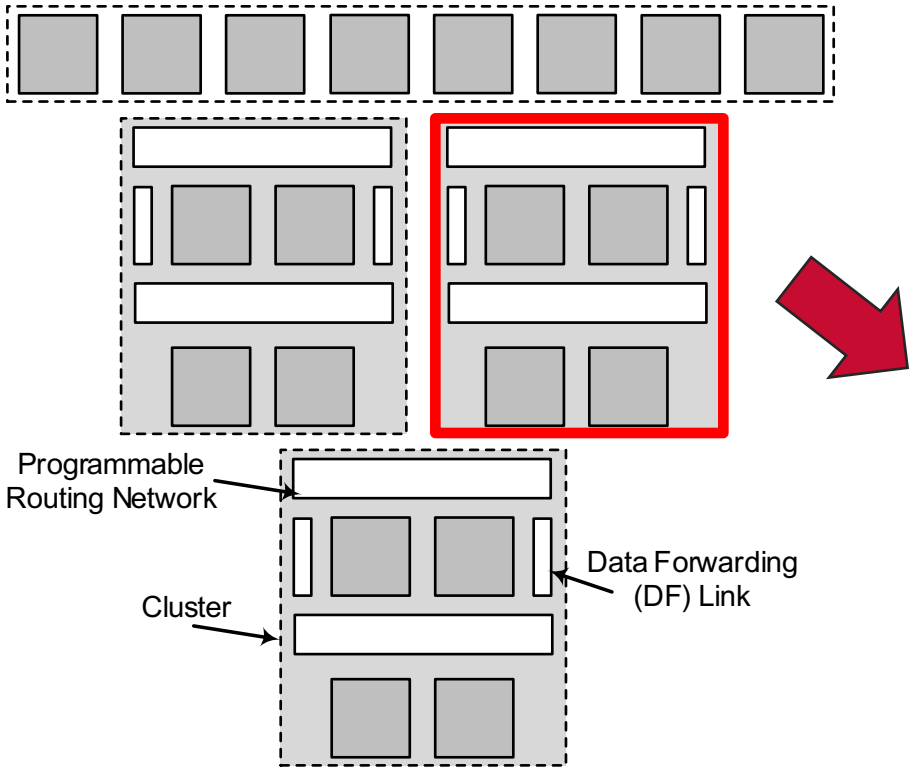
Kernel Set Characteristics

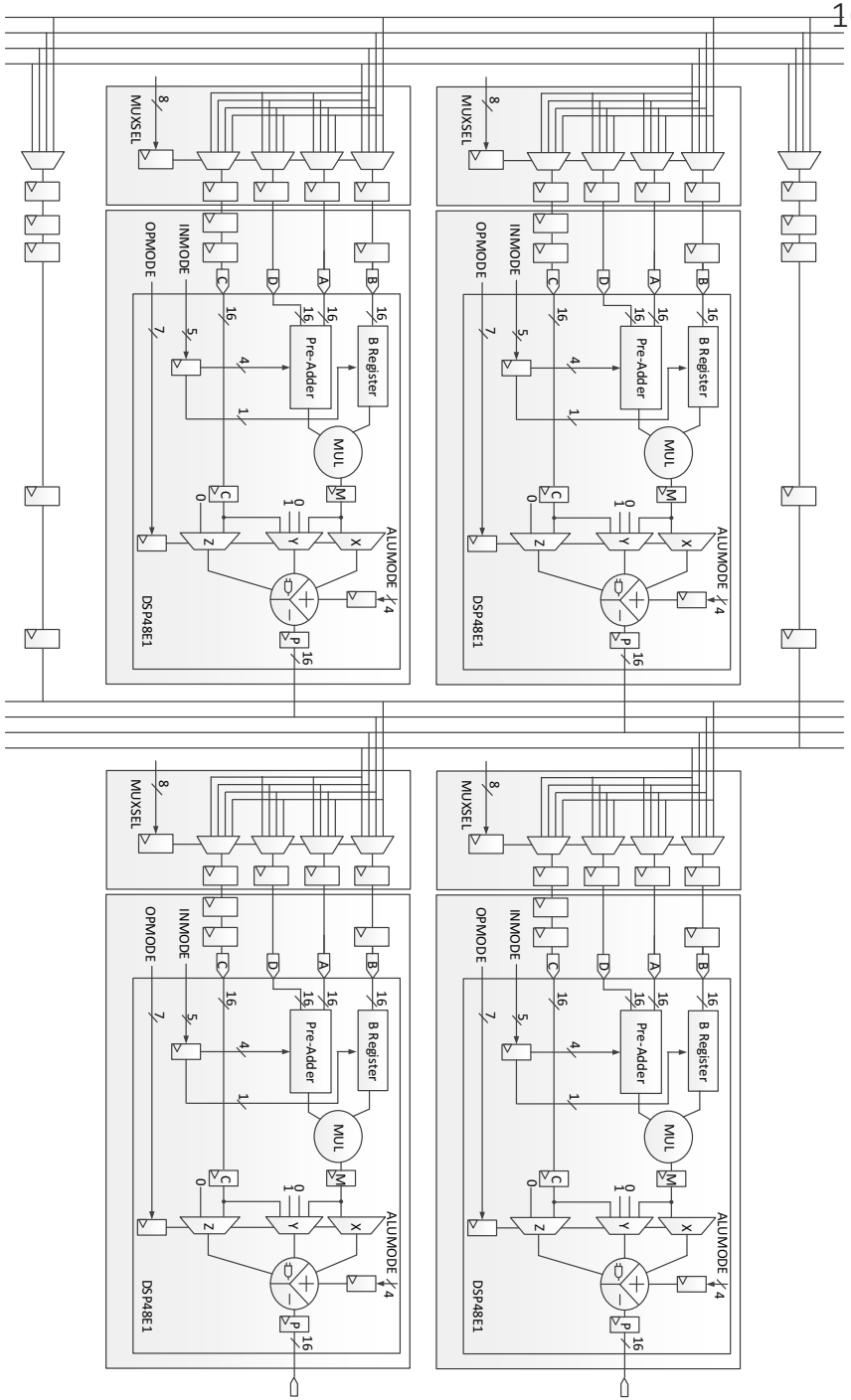
Kernels	I/O nodes	Before Transformation	
		OP nodes	DFG depth
fft	6/4	10	3
kmeans	16/1	23	9
mm	16/1	15	8
spmv	16/2	14	4
mri	11/2	11	6
stencil	15/2	14	5

Designed Overlay



Designed Overlay



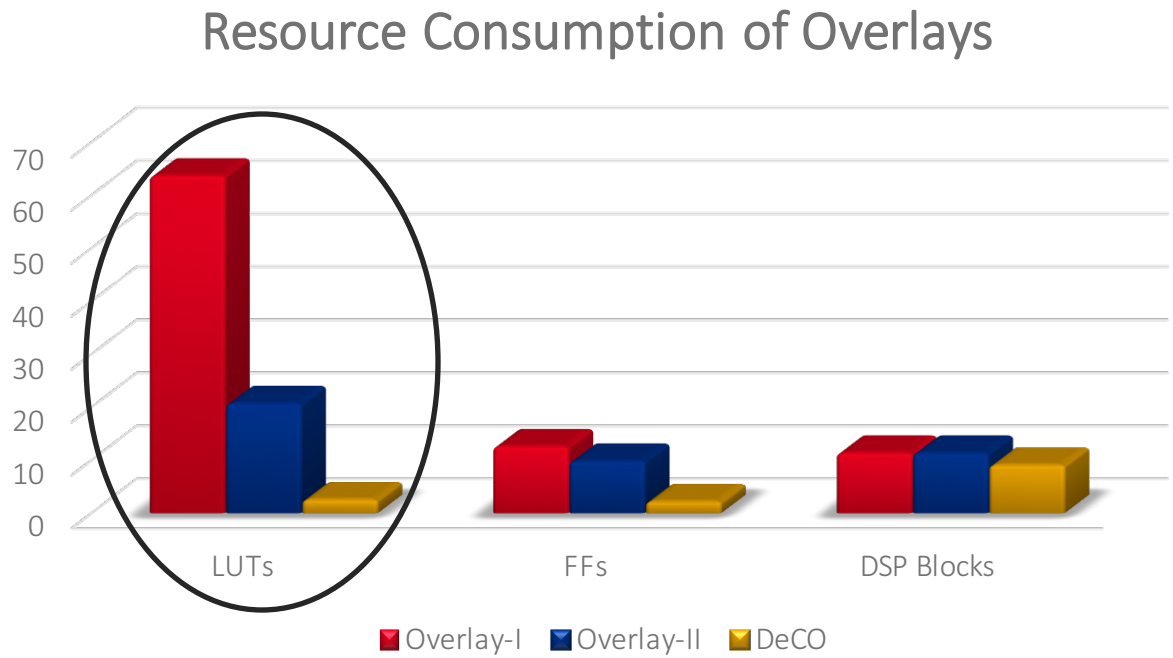


Comparison of Overlays for the kernel set

- Prototyped DeCO and two other overlays for the kernel set
  - 5x5 DSP-Based DySER overlay (Overlay-I)
  - 5x5 DSP block based island-style overlay (Overlay-II)

Comparison of Overlays for the kernel set

- Prototyped DeCO and two other overlays for the kernel set
  - 5x5 DSP-Based DySER overlay (Overlay-I)
  - 5x5 DSP block based island-style overlay (Overlay-II)
- Significant savings in LUT requirements
  - 96% compared to Overlay-I
  - 87% compared to Overlay-II



Mapping Kernels onto DeCO

- FU utilization of up to 95%

Kernels	Required No. of Cones	% FU Utilization	Achievable GOPS
fft	1	40%	3.95
kmeans	1	95%	9.08
mm	1	75%	5.92
spmv	1	70%	5.53
mri	1	75%	4.34
stencil	1	80%	5.53

Mapping Kernels onto DeCO

- FU utilization of up to 95%
- Can replicate small kernels and map

Kernels	Required No. of Cones	% FU Utilization	Achievable GOPS
fft	1	40%	3.95
kmeans	1	95%	9.08
mm	1	75%	5.92
spmv	1	70%	5.53
mri	1	75%	4.34
stencil	1	80%	5.53
gradient	0.5	90%	4.34
chebyshev	0.5	40%	5.53

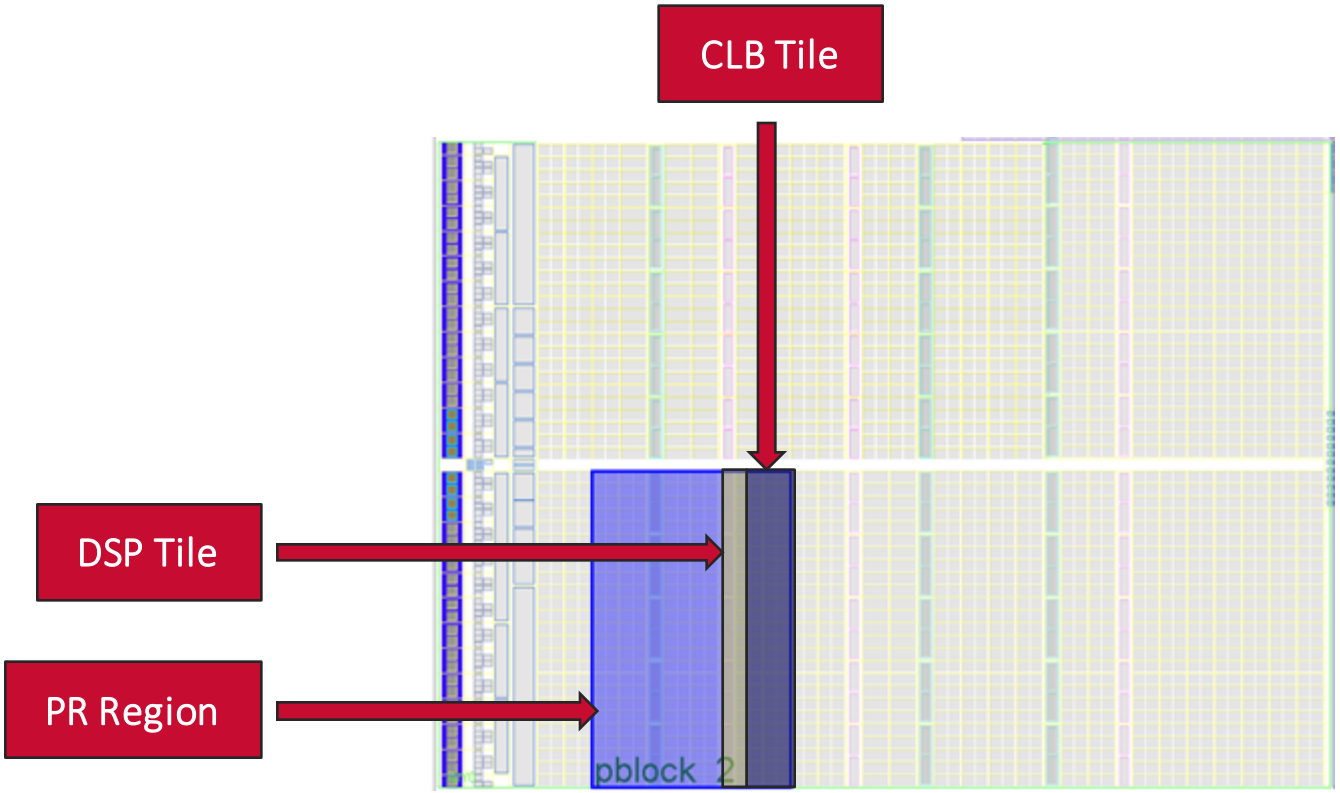
Mapping Kernels onto DeCO

- FU utilization of up to 95%
- Can replicate small kernels and map
- Multiple cones can be used to map large kernels

Kernels	Required No. of Cones	% FU Utilization	Achievable GOPS
fft	1	40%	3.95
kmeans	1	95%	9.08
mm	1	75%	5.92
spmv	1	70%	5.53
mri	1	75%	4.34
stencil	1	80%	5.53
gradient	0.5	90%	4.34
chebyshev	0.5	40%	5.53
bicg	3	50%	11.85
trmm	4.5	60%	21.33
syrk	4.5	80%	28.44

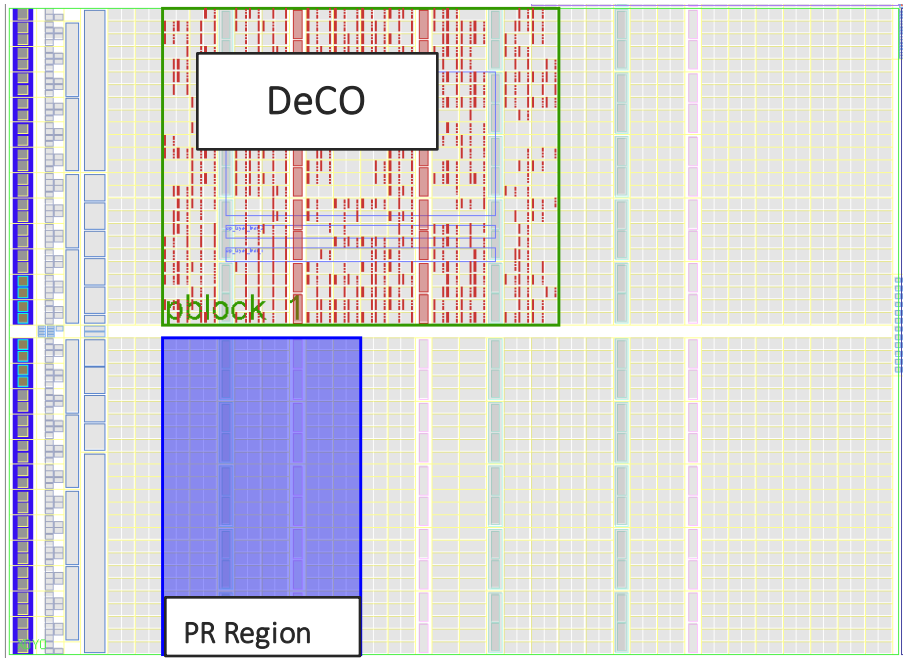
Comparison to HLS

- Compare DeCO with a Vivado HLS implementations (implemented in PR region)
  - For the kernel set HLS required 1 DSP and 3 CLB tiles



Comparison to HLS

- Compare DeCO with a Vivado HLS implementations (implemented in PR region)
  - For the kernel set HLS required 1 DSP and 3 CLB tiles
  - DeCO requires 2 DSP and 6 CLB tiles.

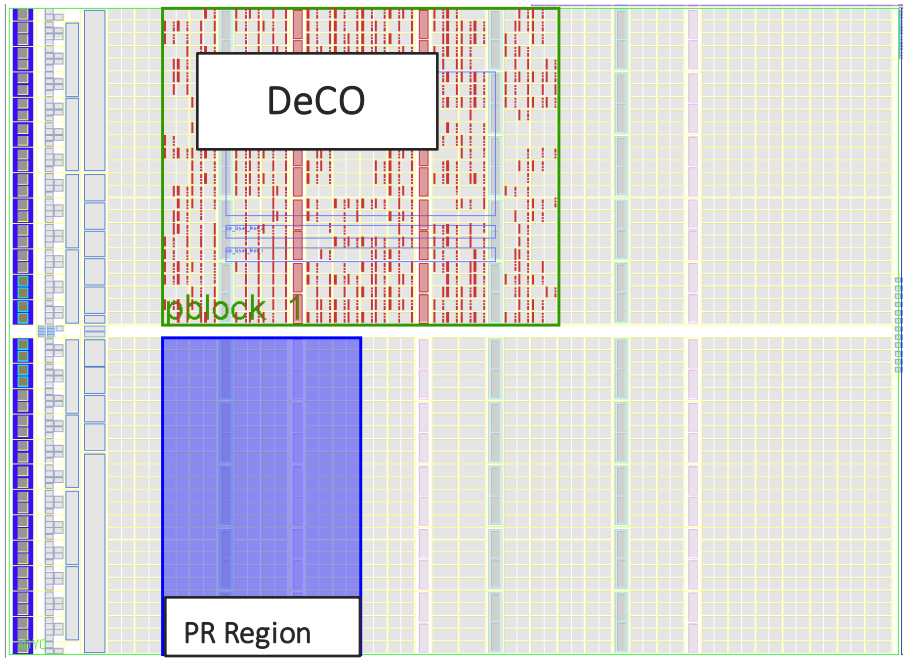


Comparison to HLS

- Compare DeCO with a Vivado HLS implementations (implemented in PR region)
  - For the kernel set HLS required 1 DSP and 3 CLB tiles
  - DeCO requires 2 DSP and 6 CLB tiles.

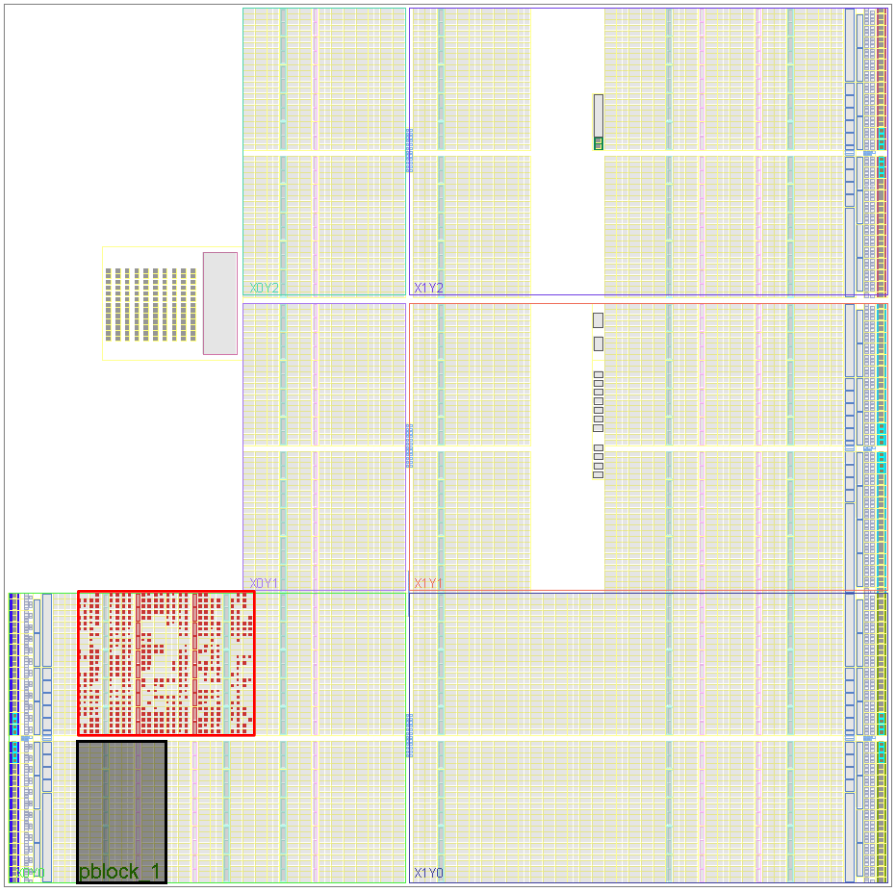
	Vivado-HLS	DeCO
Configuration Data Size	49000 Bytes	53.5 Bytes
Configuration time	382 us	2 us

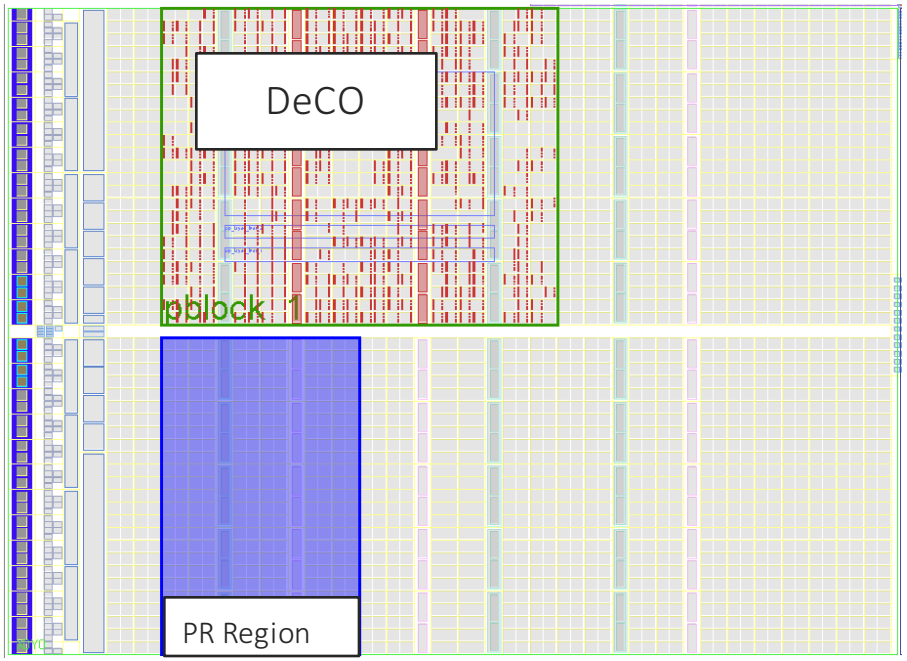
190x faster

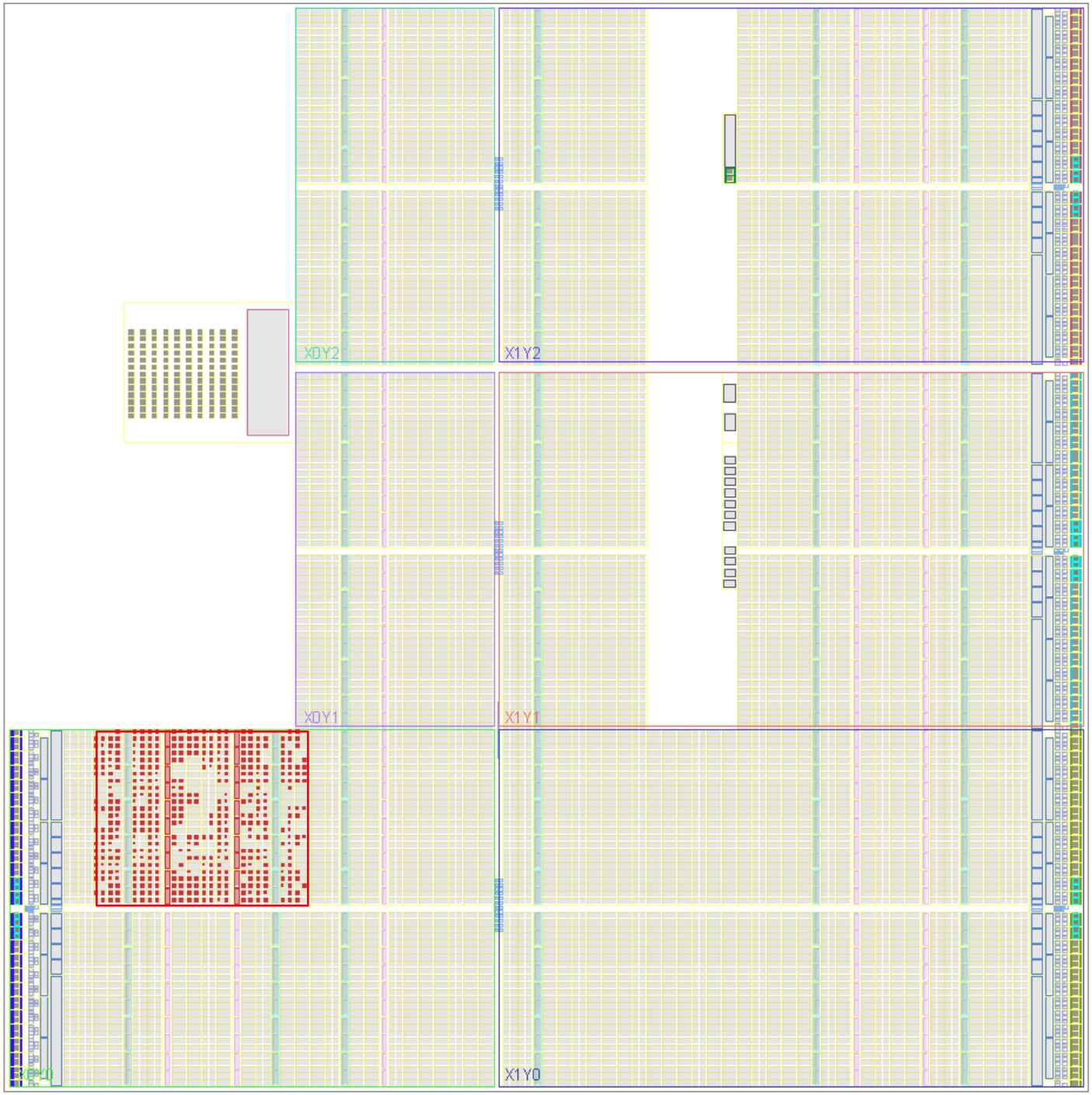


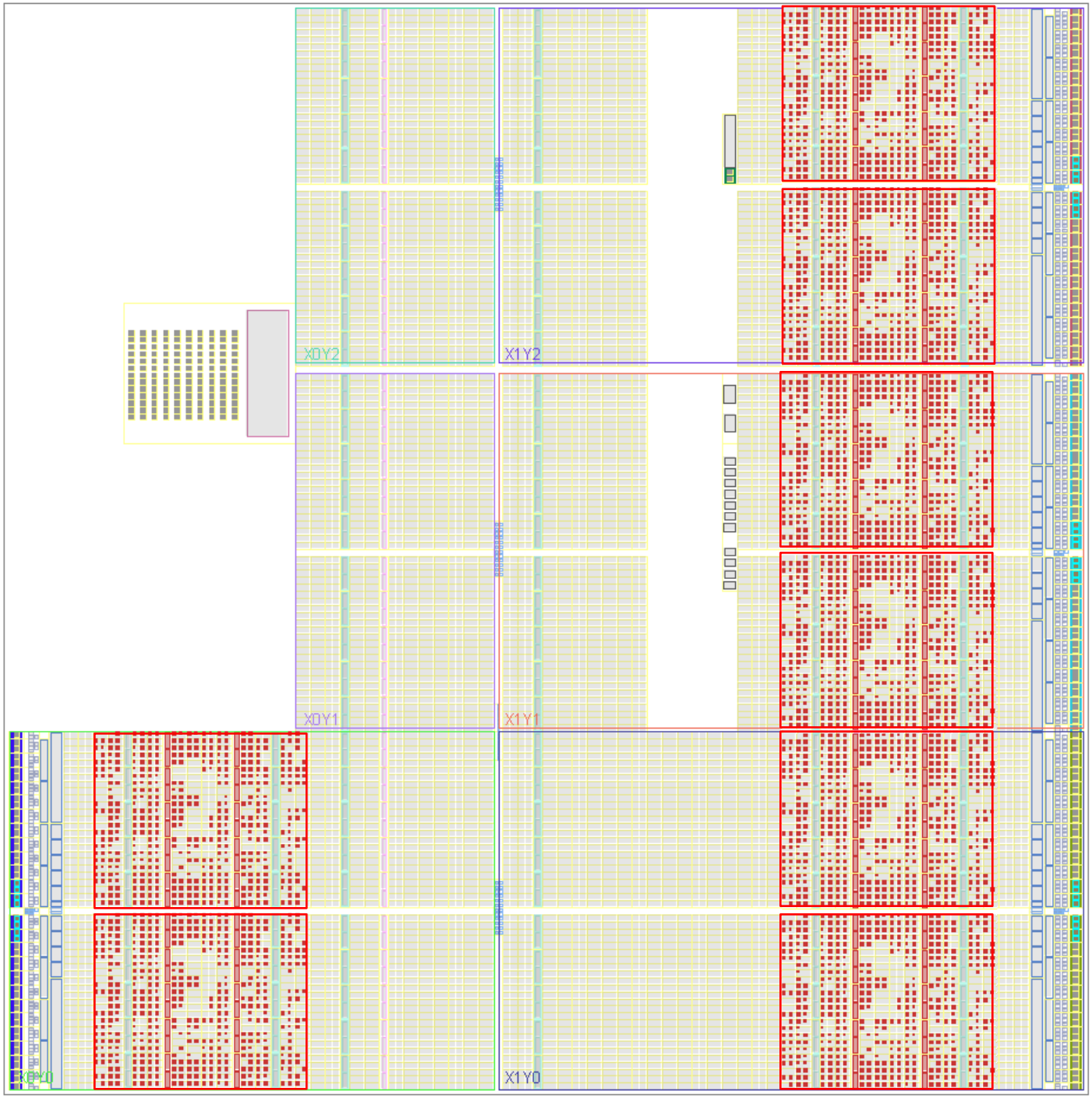
NANYANG TECHNOLOGICAL UNIVERSITY

Mapping to Zynq





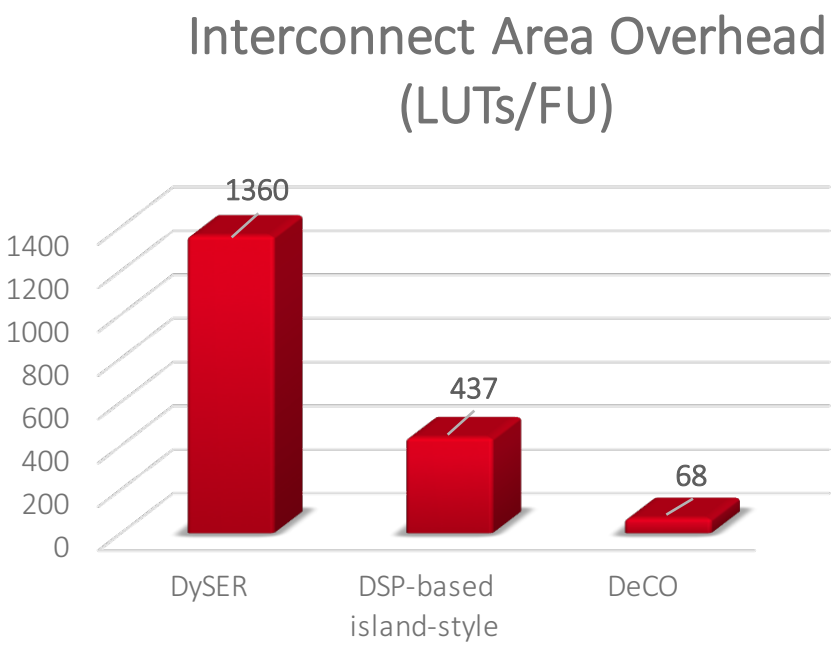

X0Y2
X1Y2
X0Y1
X1Y1
X0Y0
X1Y0

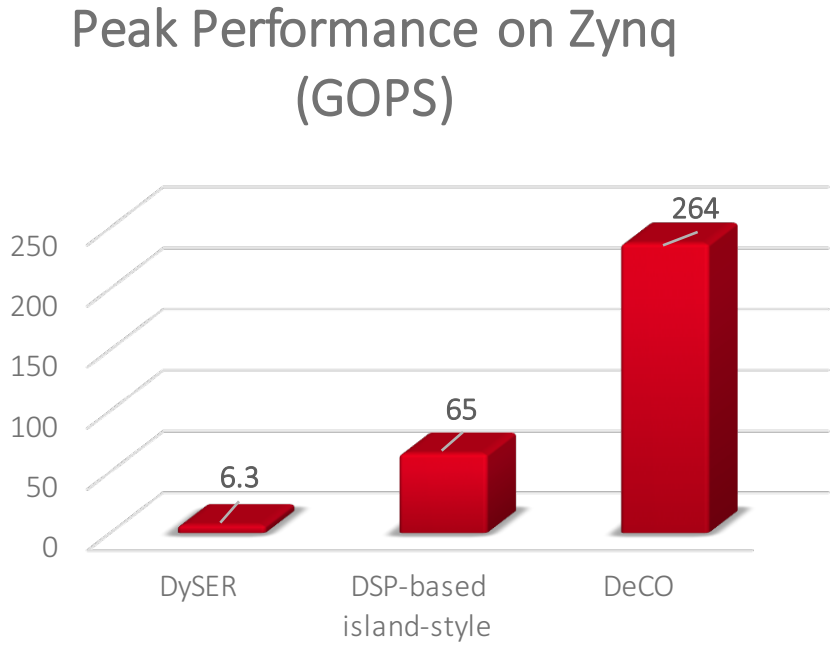


NANYANG TECHNOLOGICAL UNIVERSITY

Conclusion

- Presented DeCO, an overlay with
  - Lower Interconnect area overhead
  - Higher peak performance

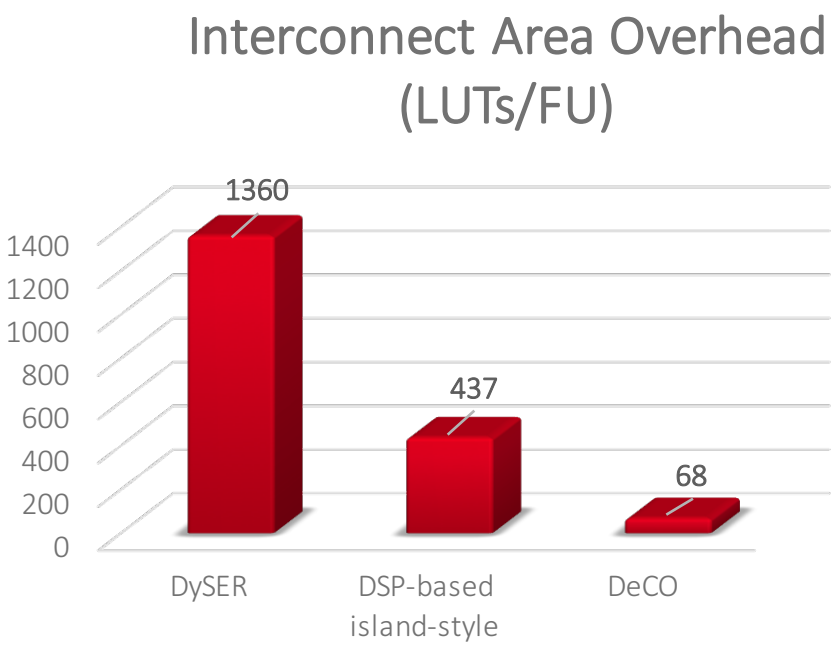


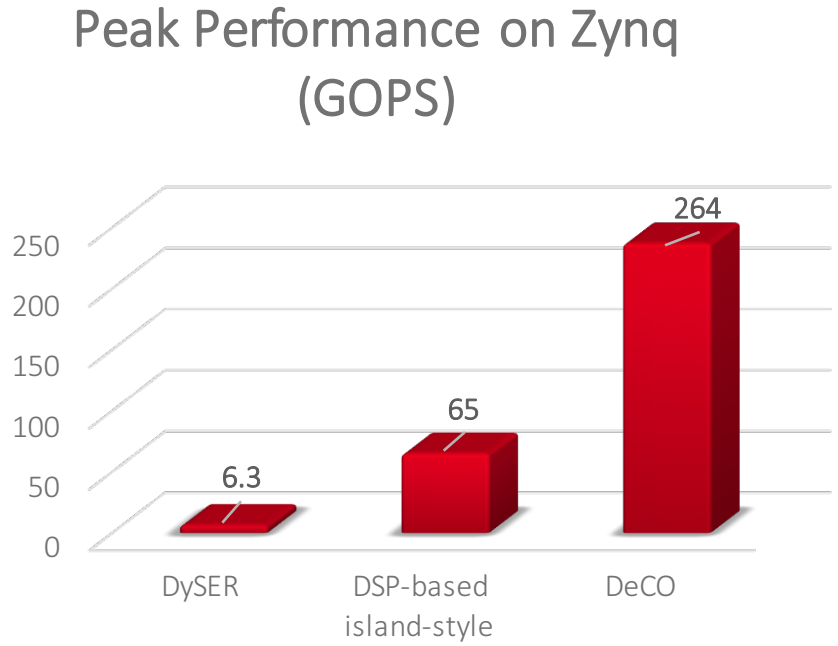


NANYANG TECHNOLOGICAL UNIVERSITY

Conclusion

- Presented DeCO, an overlay with
  - Lower Interconnect area overhead
  - Higher peak performance
- Compared to HLS generated PR-based implementation
  - 2 times area penalty but 200 times faster reconfiguration





NANYANG TECHNOLOGICAL UNIVERSITY

Future Work

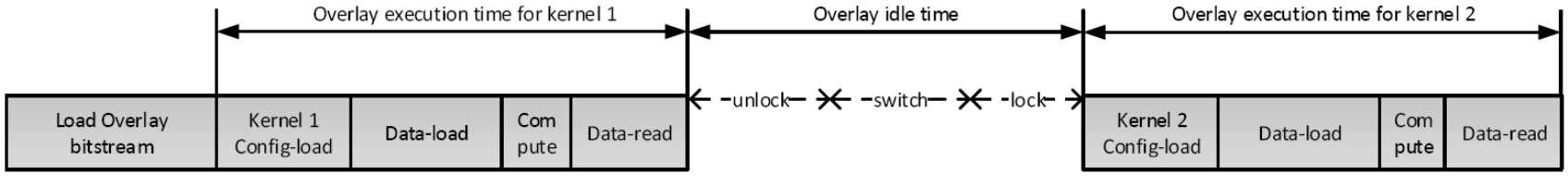
- Releasing DeCO, as a programmable accelerator within Zynq

Future Work

- Releasing DeCO, as a programmable accelerator within Zynq
- OpenCL compiler for DeCO

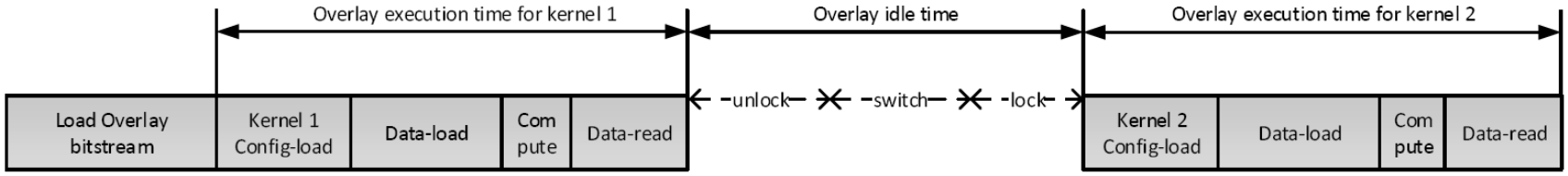
Future Work

- Releasing DeCO, as a programmable accelerator within Zynq
- OpenCL compiler for DeCO
- Fast context switching under control of OS/Hypervisor



Future Work

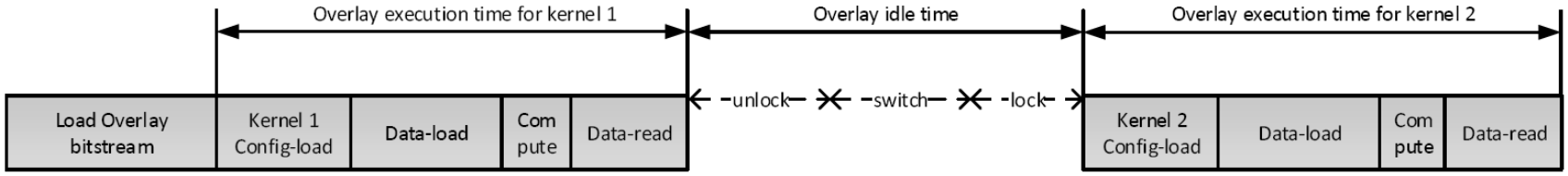
- Releasing DeCO, as a programmable accelerator within Zynq
- OpenCL compiler for DeCO
- Fast context switching under control of OS/Hypervisor



Demo of compiler for island-style overlays tonight!

Future Work

- Releasing DeCO, as a programmable accelerator within Zynq
- OpenCL compiler for DeCO
- Fast context switching under control of OS/Hypervisor



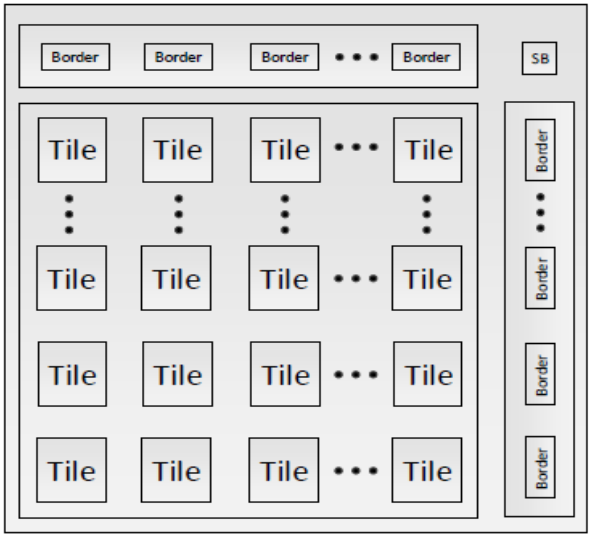
Demo of compiler for island-style overlays tonight!

NANYANG TECHNOLOGICAL UNIVERSITY

Back-up Slides

DSP based overlays (Island-style network)

- Fully Pipelined DSP Block based Island-style Overlay[1]
  - An 8 × 8 array of FUs (64 DSPs) on Xilinx Zynq
  - 28K LUTs required to implement interconnect
  - Fmax = 338MHz, peak throughput of 65 GOPS
  - Interconnect area overhead: 437 LUTs/FU

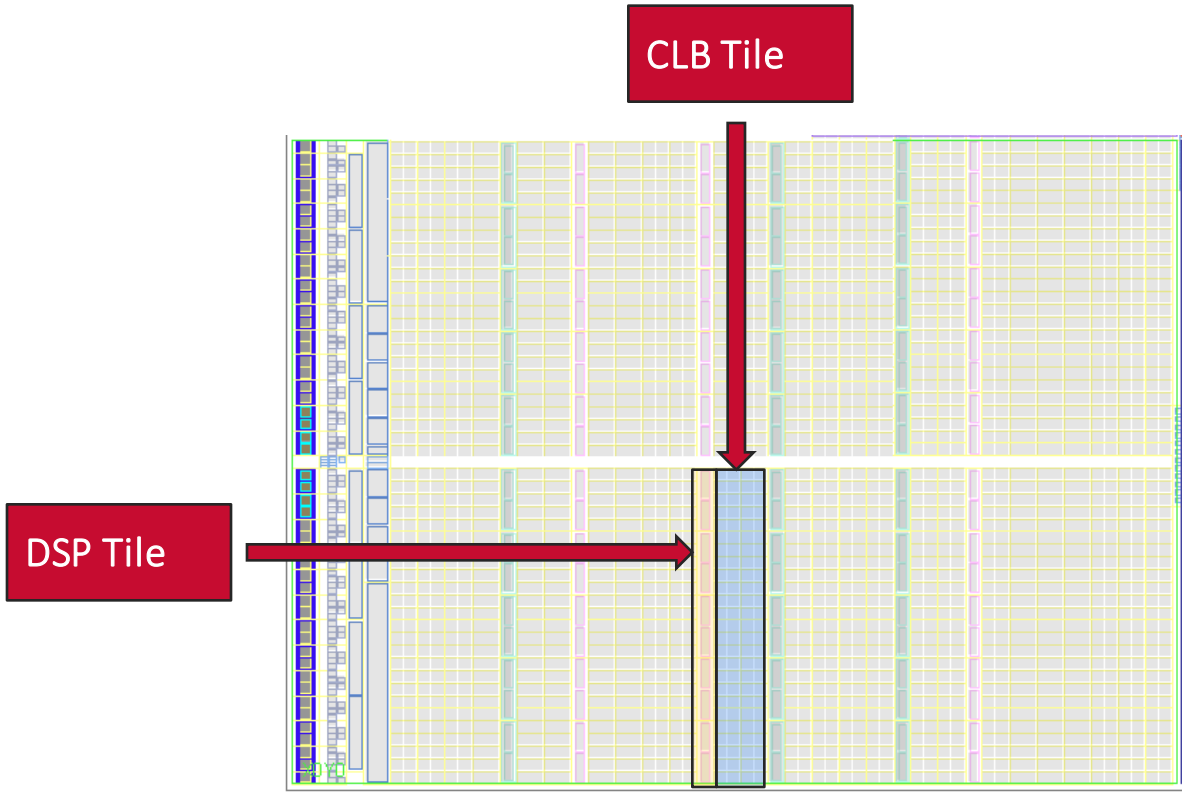


- Another Overlay: DySER[2]
  - An 6 × 6 array of FUs (36 DSPs) on Xilinx Zynq
  - 48K LUTs required to implement interconnect
  - Fmax = 175MHz, peak throughput of 6.3 GOPS
  - Interconnect area overhead: 1360 LUTs/FU
- Proposed design was 10x better (in peak throughput) and 3x better (in area overhead) then DySER

[1] A. K. Jain, S. A. Fahmy, and D. L. Maskell, "Efficient Overlay architecture based on DSP blocks," in FCCM, 2015.
[2] A. K. Jain, X. Li, S. A. Fahmy, and D. L. Maskell, "Adapting the DySER architecture with DSP blocks as an Overlay for the Xilinx Zynq," in HEART, 2015.

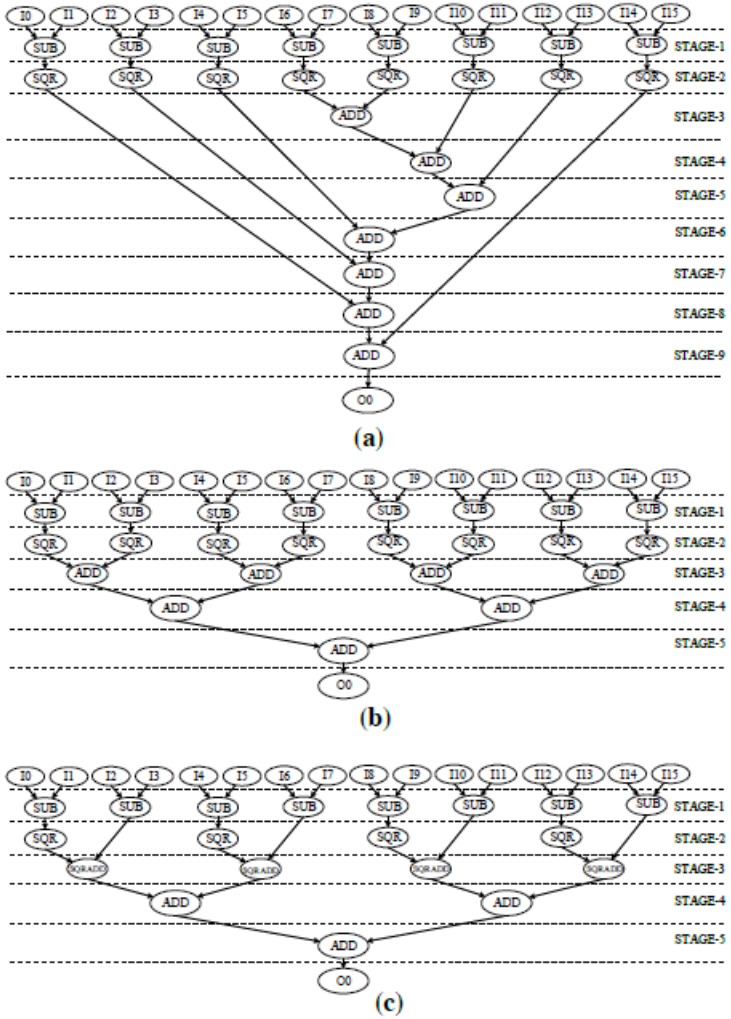
Comparison to HLS

- Compare DeCO with a direct PR-based FPGA implementation using Vivado HLS.



Programmability Overhead Modeling

- As a function of DSPs and DFs requirements in each tile
- Programmability Overhead (PO)
  - 6.25 LUTs/bit per FU for (a)
  - 4.75 LUTs/bit per FU for (b)
  - 4.30 LUTs/bit per FU for (c)
- Choose (c) as candidate DFG for overlay design process



Kernel Set Characteristics

- Perform transformation on each kernel to get the best candidate DFG

Kernels	I/O nodes	Before Transformation		After Transformation	
		OP nodes	DFG depth	OP nodes	DFG depth
fft	6/4	10	3	8	3
kmeans	16/1	23	9	19	5
mm	16/1	15	8	15	4
spmv	16/2	14	4	14	3
mri	11/2	11	6	9	5
stencil	15/2	14	5	8	3