HRL: Efficient and Flexible Reconfigurable Logic for Near-Data Processing

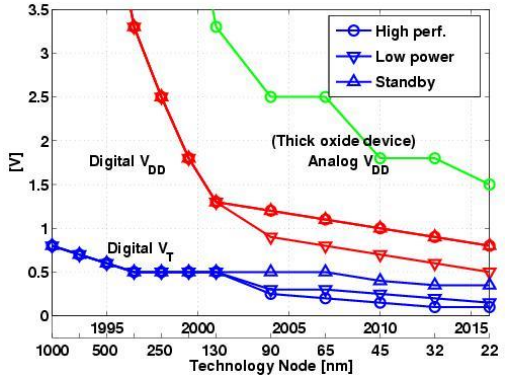
Mingyu Gao and Christos Kozyrakis

Stanford University
http://mast.stanford.edu

Stanford | ENGINEERING
Electrical Engineering

HPCA – March 14, 2016

PIM is Coming Back …



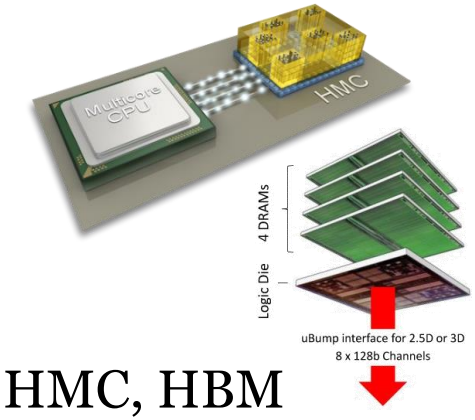
Energy-bound systems

End of Dennard scaling

Near-Data Processing (NDP)

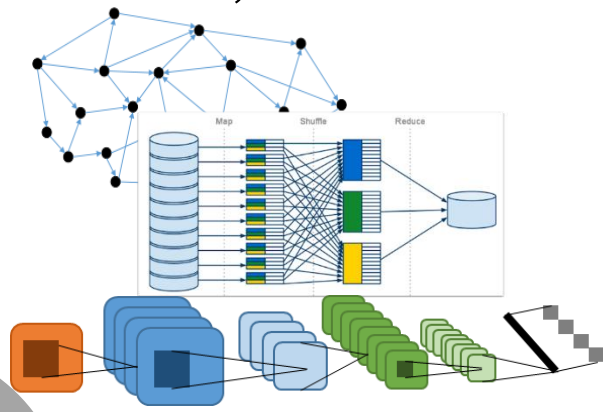
3D stacking

In-memory analytics



HMC, HBM

MapReduce, graph processing, deep neural networks, …



Figs: www.extremetech.com
www.cisl.columbia.edu/grads/tuku/research/
www.oceanaute.blogspot.com/2015/06/how-to-shuffle-sort-mapreduce.html

NDP Logic Requirements

- **Area-efficient**
  - High processing throughput to match the high memory bandwidth
  - 128 GBps per 50 mm² stack → > 32 Gflops → > 0.6 Gflops/mm²
- **Power-efficient**
  - Thermal constraints limit clock frequency
  - 5 W per stack → 100 mW/mm²
- **Flexible**
  - Must amortize manufacturing cost through reuse across apps

NDP Logic Options

	Area Efficiency	Power Efficiency	Flexibility
Programmable cores [IRAM, FlexRAM, NDC, TOP-PIM]	✗	✗	✓
FPGA (fine-grained) [Active Pages]	✗	✓	✓
CGRA (coarse-grained) [NDA]	✓	✗	✓
ASIC [MSA, LiM]	✓	✓	✗

Reconfigurable Logic Challenges

- FPGA
  - Area overhead due to support for bit-level configuration

- CGRA
  - Traditional GGRAs
    - Limited flexibility in interconnects, only for regular computation patterns
  - DySER [HPCA'11] and NDA [HPCA'15]
    - High power due to circuit-switched routing
    - Inefficient for branches and irregular data layouts

Heterogeneity: achieve the best of FPGA and CGRA

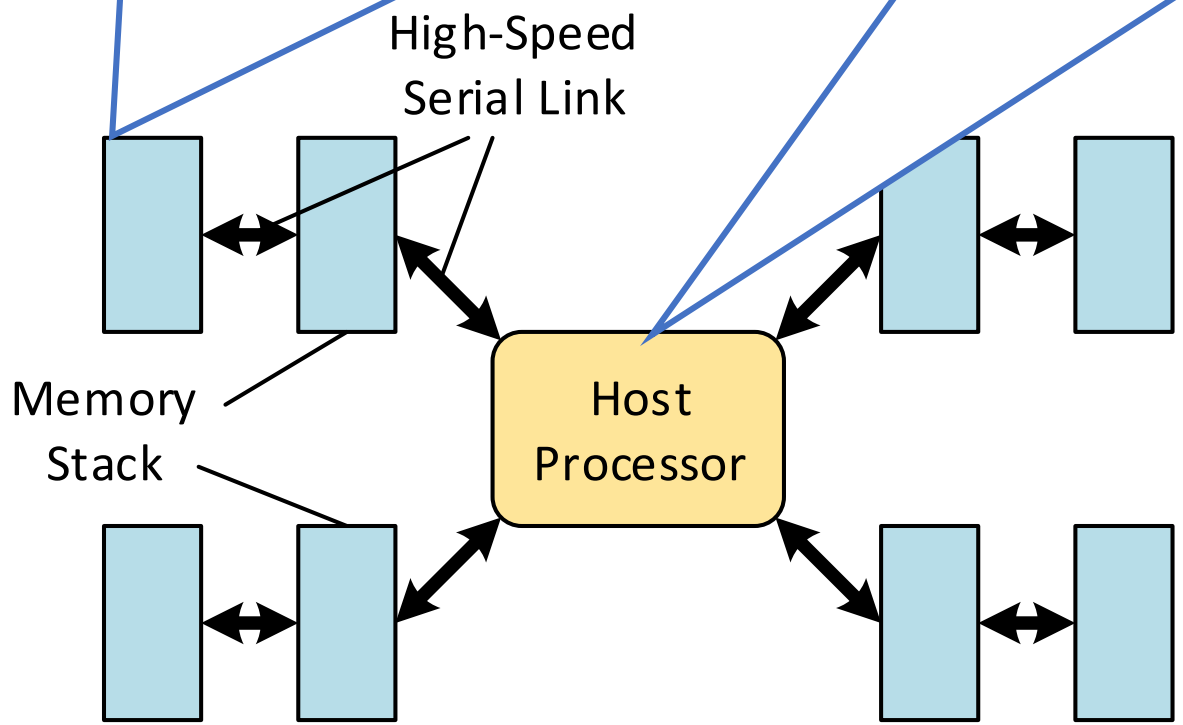
Outline

- Motivation
- NDP System Design
- Heterogeneous Reconfigurable Logic (HRL)
- Evaluation
- Conclusions

Overall System Architecture

Memory stack with NDP capability: runs memory-intensive code

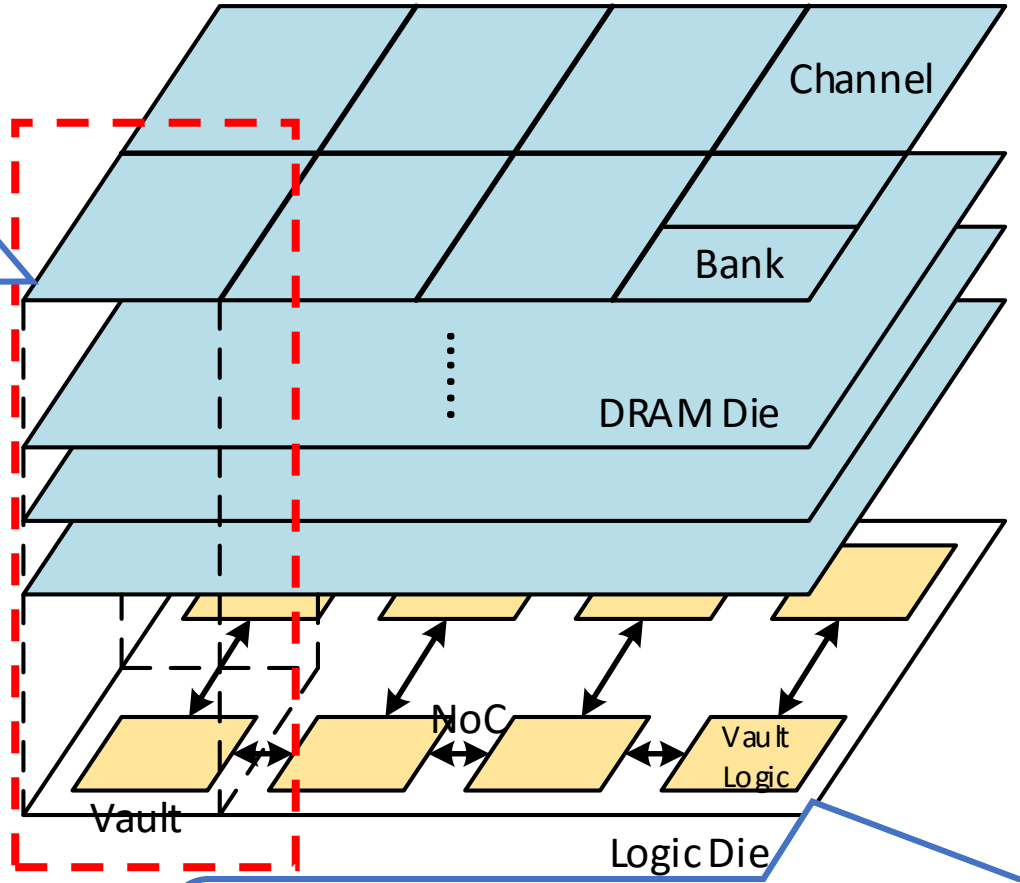
Multi-core chip with cache hierarchy: Runs code with high temporal locality



Multiple stacks linked to host processor through serial links

NDP Stack

Vault:
- Vertical channel
- Dedicated memory controller
- 8 – 16 vaults per stack

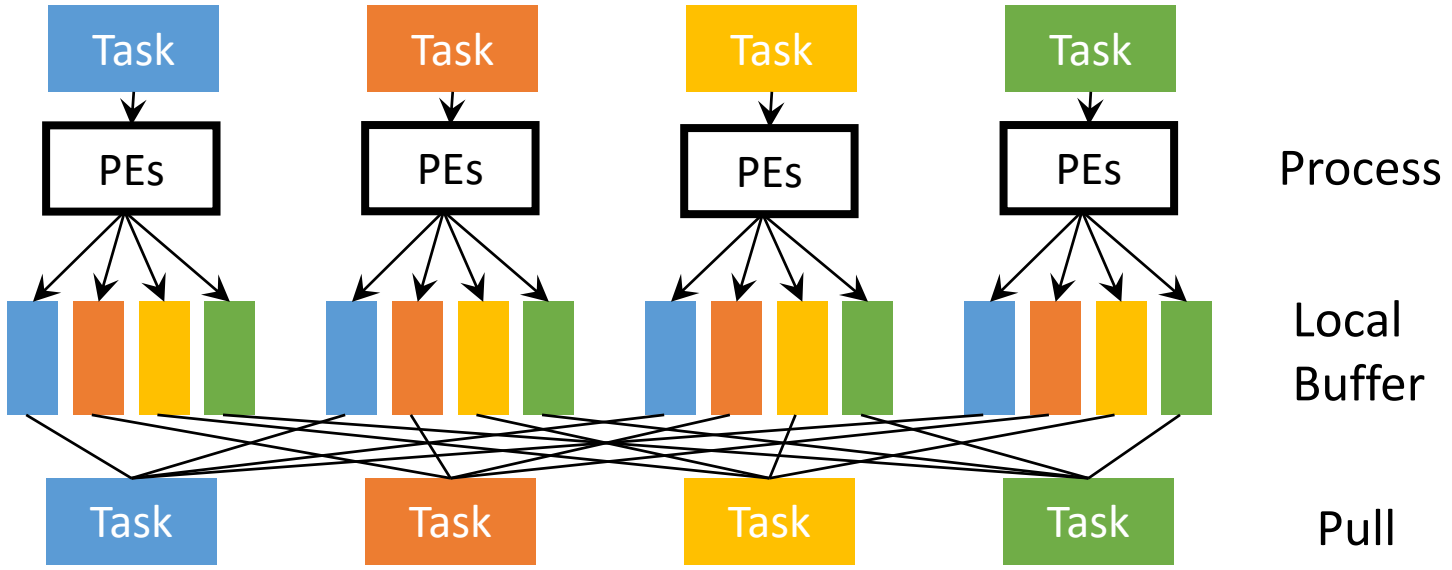


vs. DDR3 channel
- **10x** bandwidth (160 GBps)
- **3-5x** power improvement

Vault logic:
- Multiple PEs + control logic
- NoC to interconnect vaults

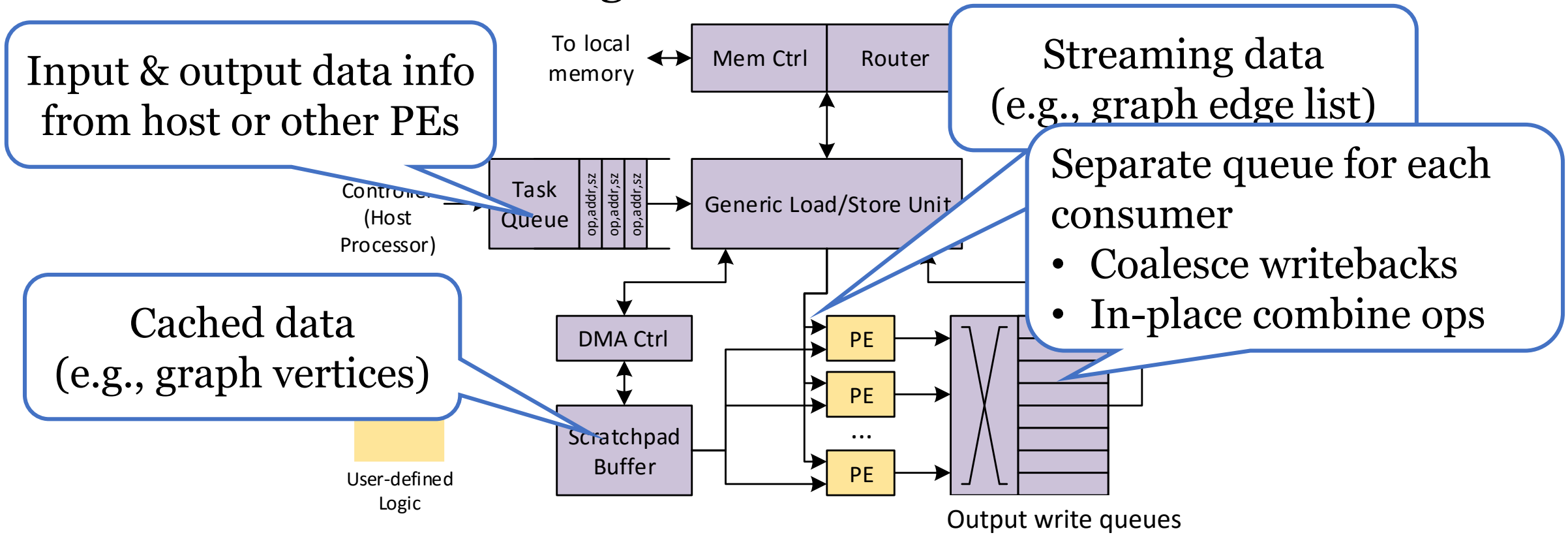
Iterative Execution Flow

- Processing phase: PEs run tasks independently and in parallel
- Communication phase: data exchange and sync b/w PEs [PACT'15]
  - Communication within and across stacks



Vault Logic

- Handles task control and data communication
- Allows the use of reconfigurable or custom PEs



Outline

- Motivation
- NDP System Design
- Heterogeneous Reconfigurable Logic (HRL)
- Evaluation
- Conclusions

HRL Features

- Fine-grained + coarse-grained reconfigurable blocks
  - LUTs for flexible control
  - ALUs for efficient arithmetic
- Static interconnects
  - Wide network for data
  - Separate and narrow network for control
- Special blocks for branches & irregular data layout

Area-efficiency and flexibility

Power-efficiency

Flexibility

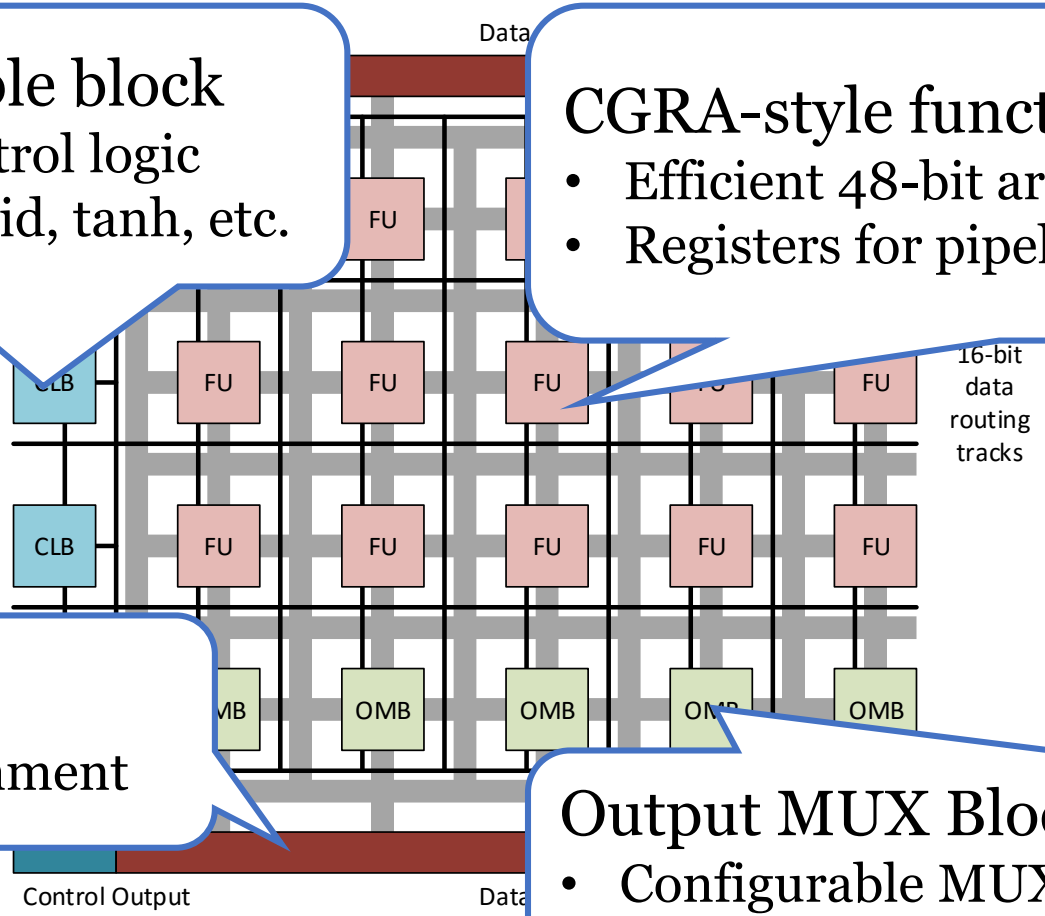
Compute throughput per Watt:

2.2x over FPGA, 1.7x over CGRA

HRL Array: Logic Blocks

FPGA-style configurable block
- LUTs for embedded control logic
- Special functions: sigmoid, tanh, etc.

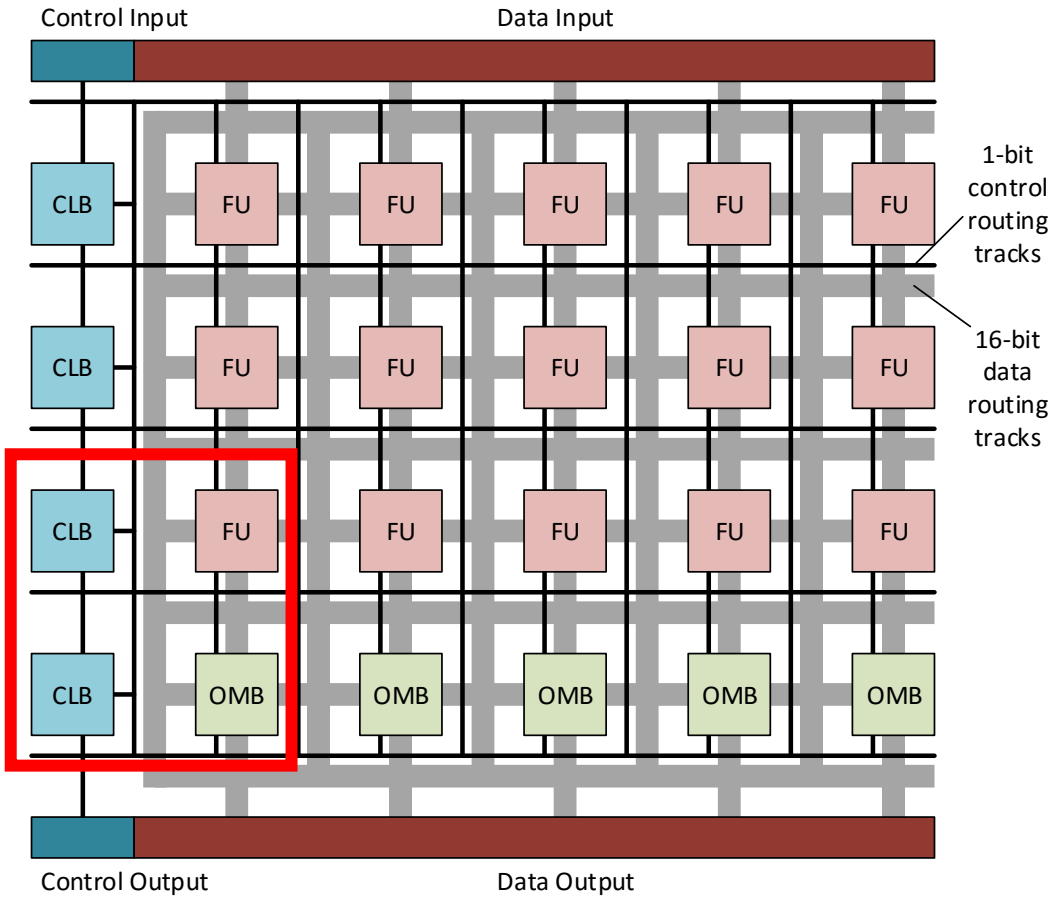
CGRA-style functional unit
- Efficient 48-bit arithmetic/logic ops
- Registers for pipelining and retiming



Flexible IO Interface
- Simple but flexible alignment

Output MUX Block
- Configurable MUXes (tree, cascading, parallel)
- Put close to output, low cost and flexible

HRL Array: Routing

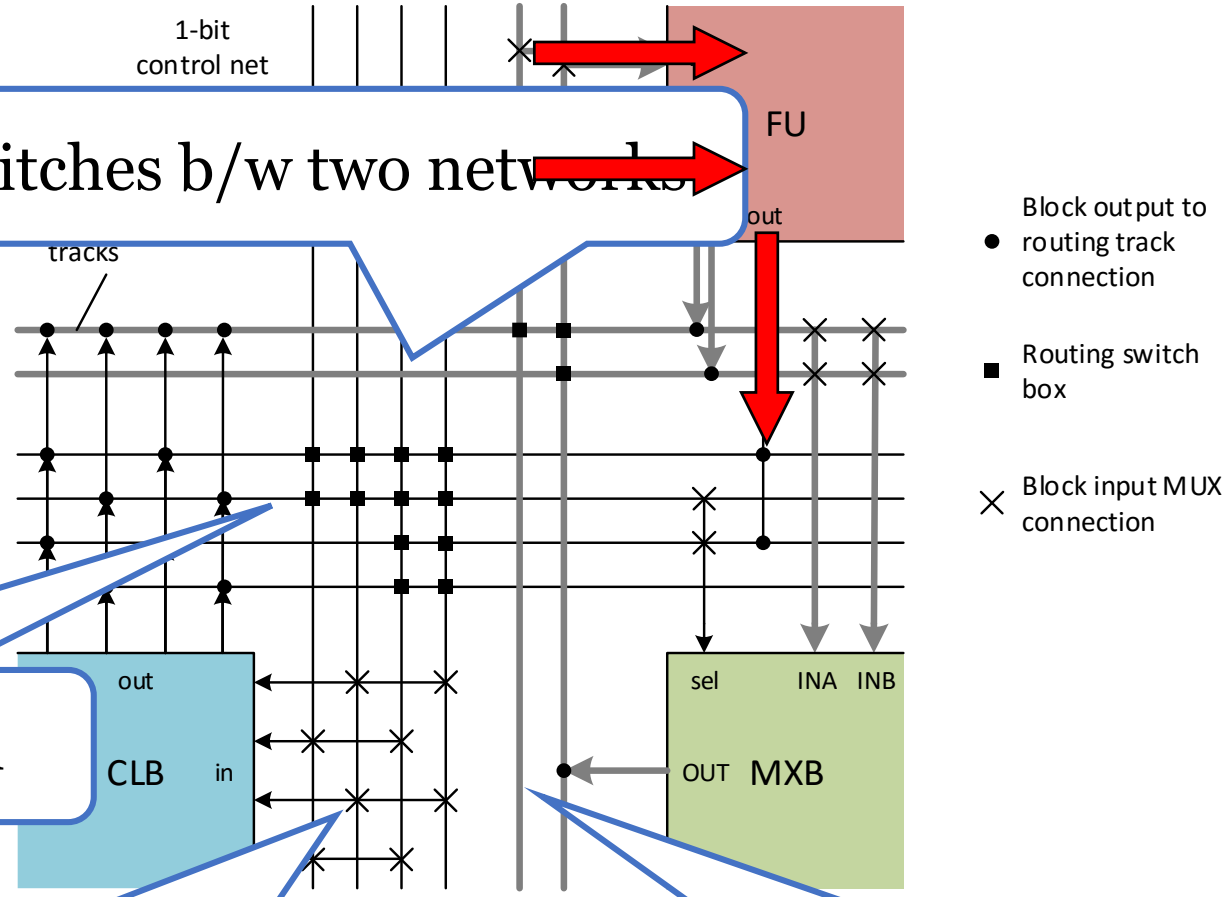


HRL Array: Routing

Fully static for **low pow**

No switches b/w two net

Separate data and control to **reduce area**

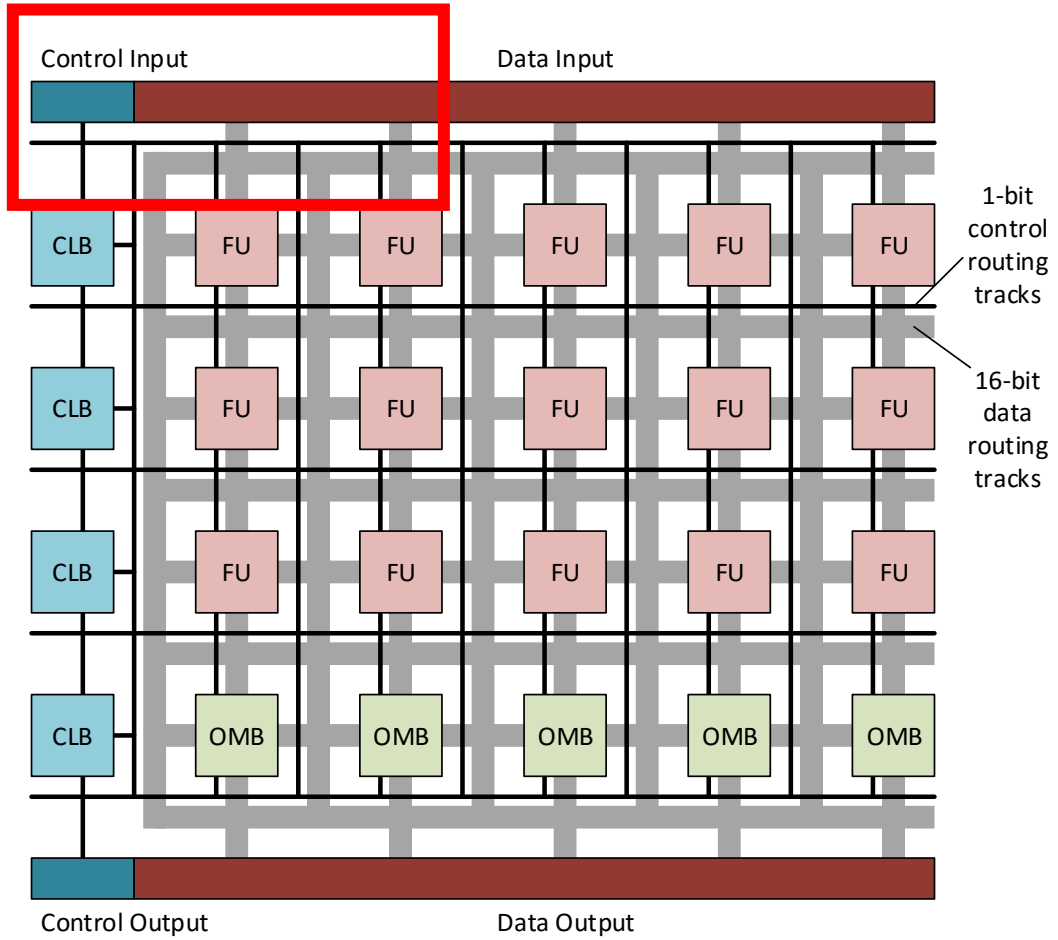


Connection box and switch

1-bit control, few tracks

16-bit data, bus-based

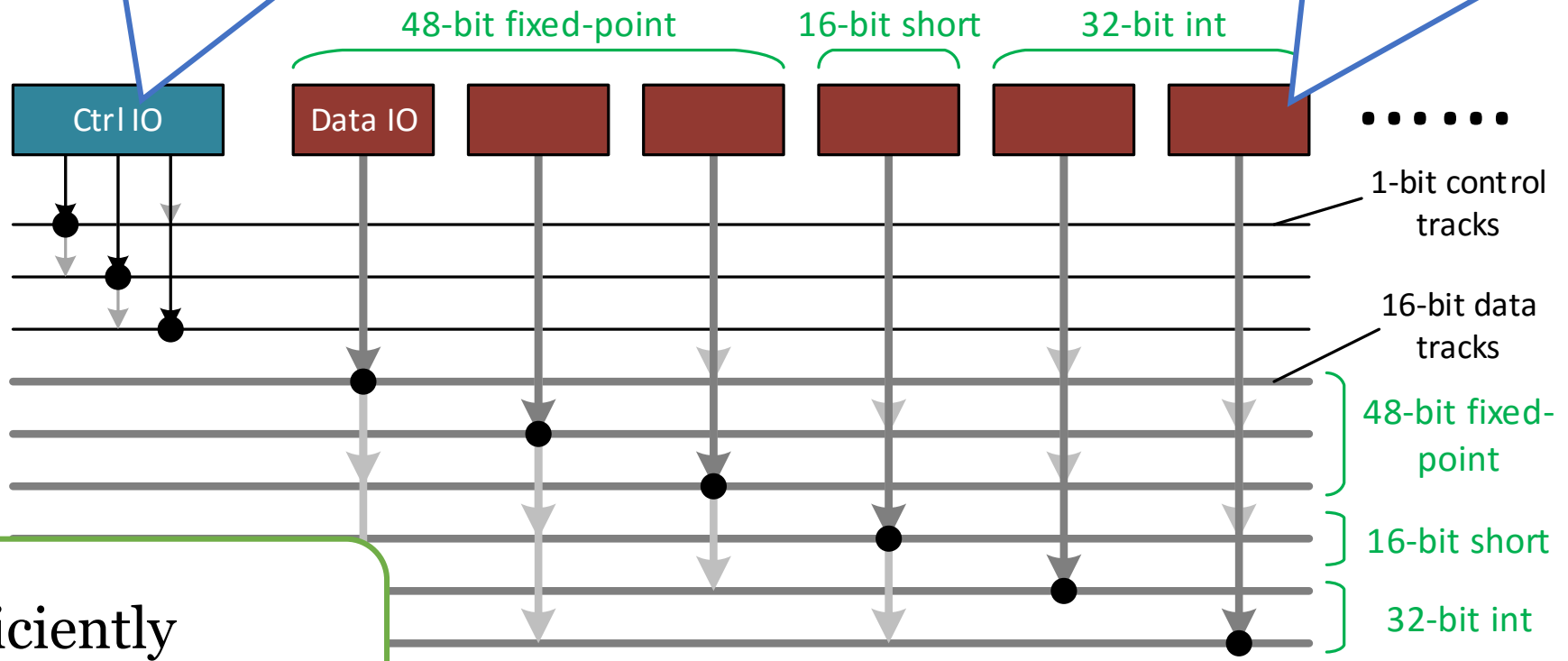
HRL Array: IO


Control Input
Data Input
CLB
FU
OMB
1-bit control routing tracks
16-bit data routing tracks
Control Output
Data Output

HRL Array: IO

Control IO
- Connects to control tracks
- Same as FPGA

Data IO
- Connect to data net
- Simple 16-bit chunk alignment



Low cost and sufficiently **flexible** even for irregular data

Outline

- Motivation
- NDP System Design
- Heterogeneous Reconfigurable Logic (HRL)
- Evaluation
- Conclusions

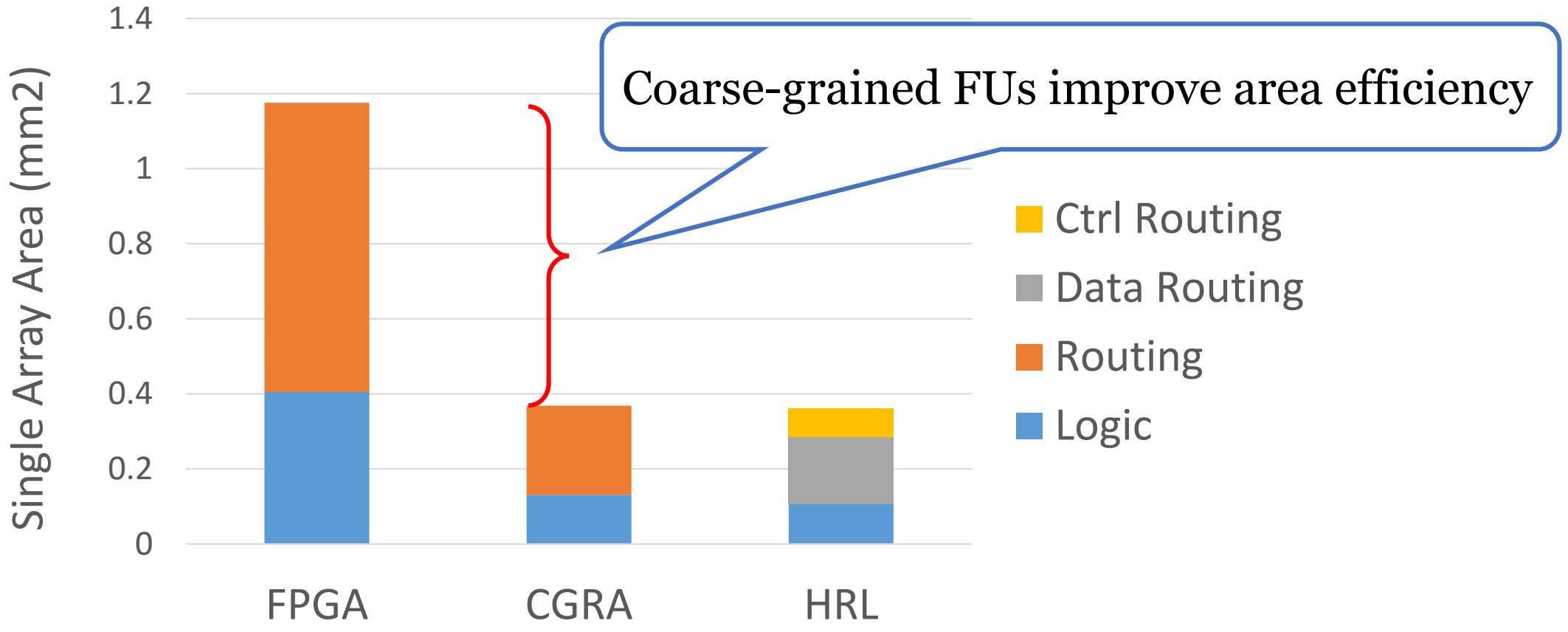
Methodology

- Workloads
  - 3 analytics frameworks: MapReduce, graph, DNN
  - 9 representative applications, 11 kernel circuits (KCs)

- Technology
  - 45 nm area and power model
  - CGRA: DySER as in NDA [HPCA'11, HPCA'15]
  - FPGA: Xilinx Virtex-6

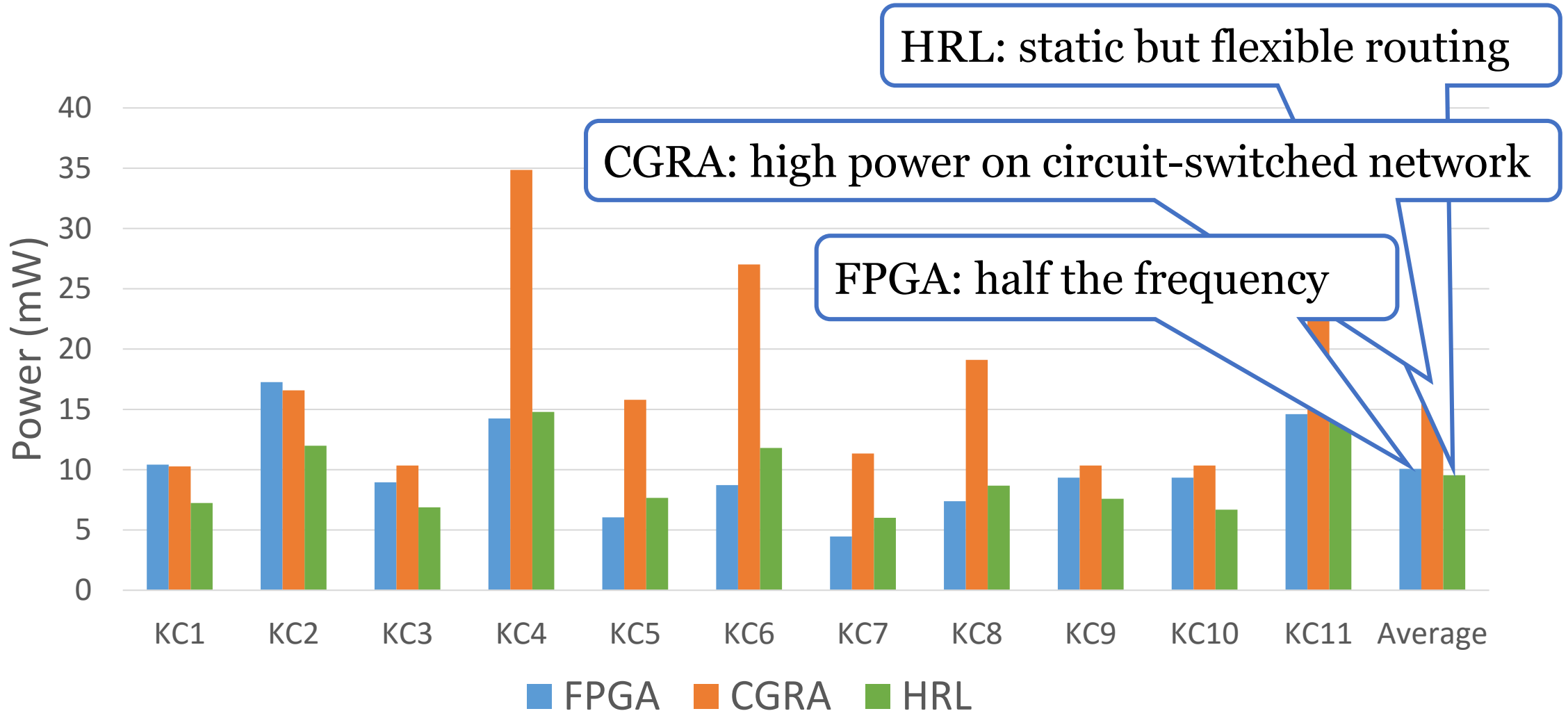
- Tools
  - Synthesize, place & route by Yosys + VTR
  - System simulation by zsim

Array Area

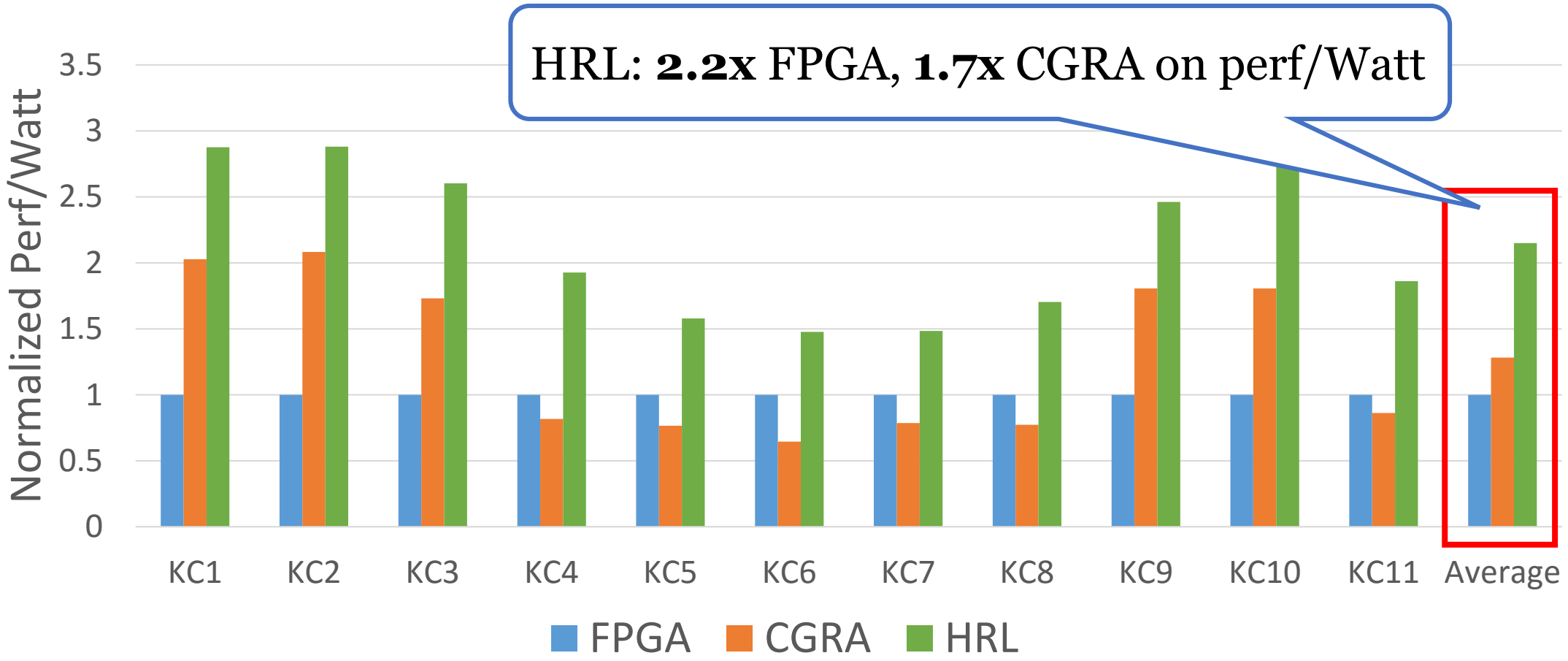


- Same logic capacity for each type array

Array Power

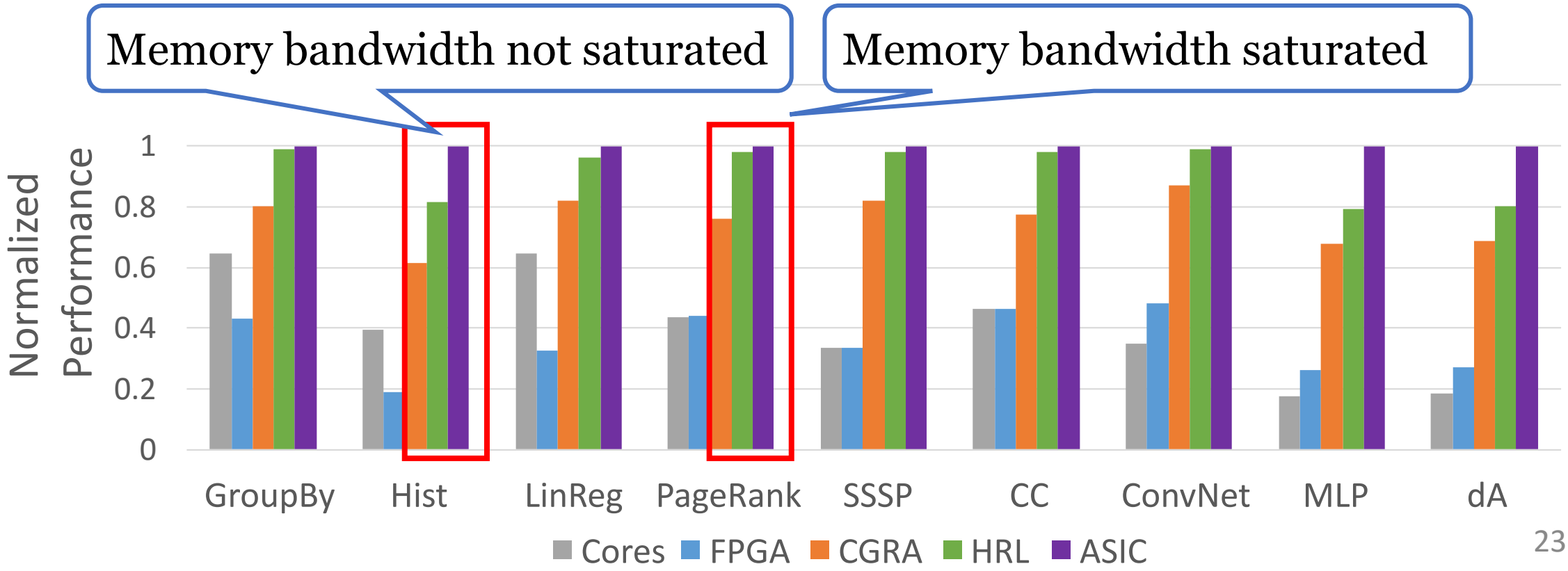

HRL: static but flexible routing
CGRA: high power on circuit-switched network
FPGA: half the frequency
Power (mW)
40
35
30
25
20
15
10
5
0
KC1
KC2
KC3
KC4
KC5
KC6
KC7
KC8
KC9
KC10
KC11
Average
FPGA
CGRA
HRL

Vault Power Efficiency



Overall Performance

- ASIC represents the upper bound of efficiency
- Cores, FPGA, CGRA only match 30% to 80% of ASIC
- HRL has 92% of ASIC performance on average



Conclusions

- NDP logic requirements: area + power efficiency, flexibility
- Heterogeneous reconfigurable logic (HRL)
  - Fine-grained + coarse-grained logic blocks
  - Static and separate data and control networks
  - Special blocks for branching and layout management
  - Vault logic handles communication and control
- HRL for in-memory analytics
  - 2.2x performance/Watt over FPGA and 1.7x over CGRA
  - Within 92% of ASIC performance

Thanks!

Questions?