

An Asynchronous NoC Router in a 14nm FinFET Library: Comparison to an Industrial Synchronous Counterpart

Weiwei Jiang

Columbia University, USA

Davide Bertozzi

University of Ferrara, Italy

Gabriele Miorandi

University of Ferrara, Italy

Steven M. Nowick

Columbia University, USA

Wayne Burleson

Advanced Micro Devices, USA

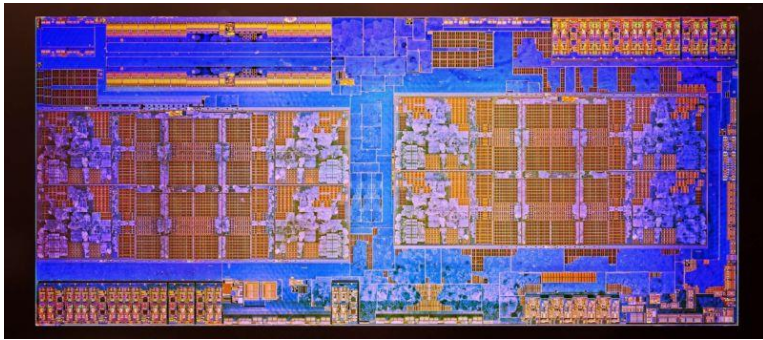
Greg Sadowski

Advanced Micro Devices, USA

ACM/IEEE Design, Automation and Test in Europe (DATE-17)

Motivation for Networks-on-Chip

Future of computing is multi-core



AMD Ryzen 8-core Processor
(March 2017)



- CPU:

8 to 24 cores widely available

- AMD 16-core Opteron 6000 series
- AMD Ryzen 4,6,8,+ cores
- Intel 24-core Xeon-E7
- Intel Xeon Phi – 80+ core

- GPU:

up to 2500-3500 graphics cores

- AMD FirePro series:
 - up to 2560 GCN Stream Processors
- NVIDIA Titan X:
 - 3584 CUDA Cores

Motivation for Networks-on-Chip (Cont.)

➤ NoC separates computation and communication

- Improves scalability
 - global interconnects have high latency and power consumption (e.g. buses and point-to-point wiring)
- Increases performance/energy efficiency
 - share wiring resources between parallel data flows
- Facilitates design reuse
 - optimized IPs can simply plug in ➡ largely decrease design efforts

Potential Advantages of Asynchronous Design

➤ No global clock

- No clock power
 - ➡ less overall power than deeply clock-gated sync designs
- No clock design overhead
 - ➡ no clock generation, distribution, skew analysis, etc.
 - [Gebhardt/Stevens et al., *Comparing energy and latency of asynchronous and synchronous NoCs for embedded SoCs*, NOCS-10]

➤ Greater flexibility/modularity

- Easily integrates multiple timing domains
- Supports reusable components
 - [Bainbridge/Furber, *CHAIN: a delay-insensitive chip area interconnect*, IEEE Micro-02]

➤ Lower system latency

- No per-router clock synchronization ➡ no waiting for clock
 - [Sheibanyrad/Greiner et al., *Multisynchronous and fully asynchronous NoCs for GALS architectures*, IEEE Design & Test of Computers-08]

Recent Commercial Asynchronous NoC Chips

- **Intel's FM5000/6000 Ethernet switches** [IEEE Design & Test 2015]
 - high performance: 640 Gbps max. bandwidth + 400 ns cut-through latency
 - support up to 176 ports
- **IBM's TrueNorth neuromorphic chip** [Science 2014]
 - a 5.4-billion-transistor chip with 4096 neurosynaptic cores
 - models 1M neurons and 256M synapses
 - ultra-low power:
 - only **63 milliwatts** with 400x240 video input at 30 frames/sec.
- **STMicroelectronics' STHORM processor** [DAC-12]
 - A GALS computing accelerator for embedded SoCs
 - connect 4 clusters, each with 16 sync processors
 - improved performance efficiency over several Quadro and Nvidia GPUs

Contributions (1)

➤ First comparison for:

async vs. *commercial sync router* in advanced technology

- Sync baseline is for **high-end** processors and graphics products
 - NoC handles system config and power/performance control
- Sync baseline uses aggressive **clock optimization and fine-grain clock gating**
- Comparison in a 14nm FinFET library
 - not 'textbook' academic technology library
 - state of the art CMOS technology used in commercial products
- Dominating results for asynchronous
 - in **key metrics**: area, latency and idle/active power

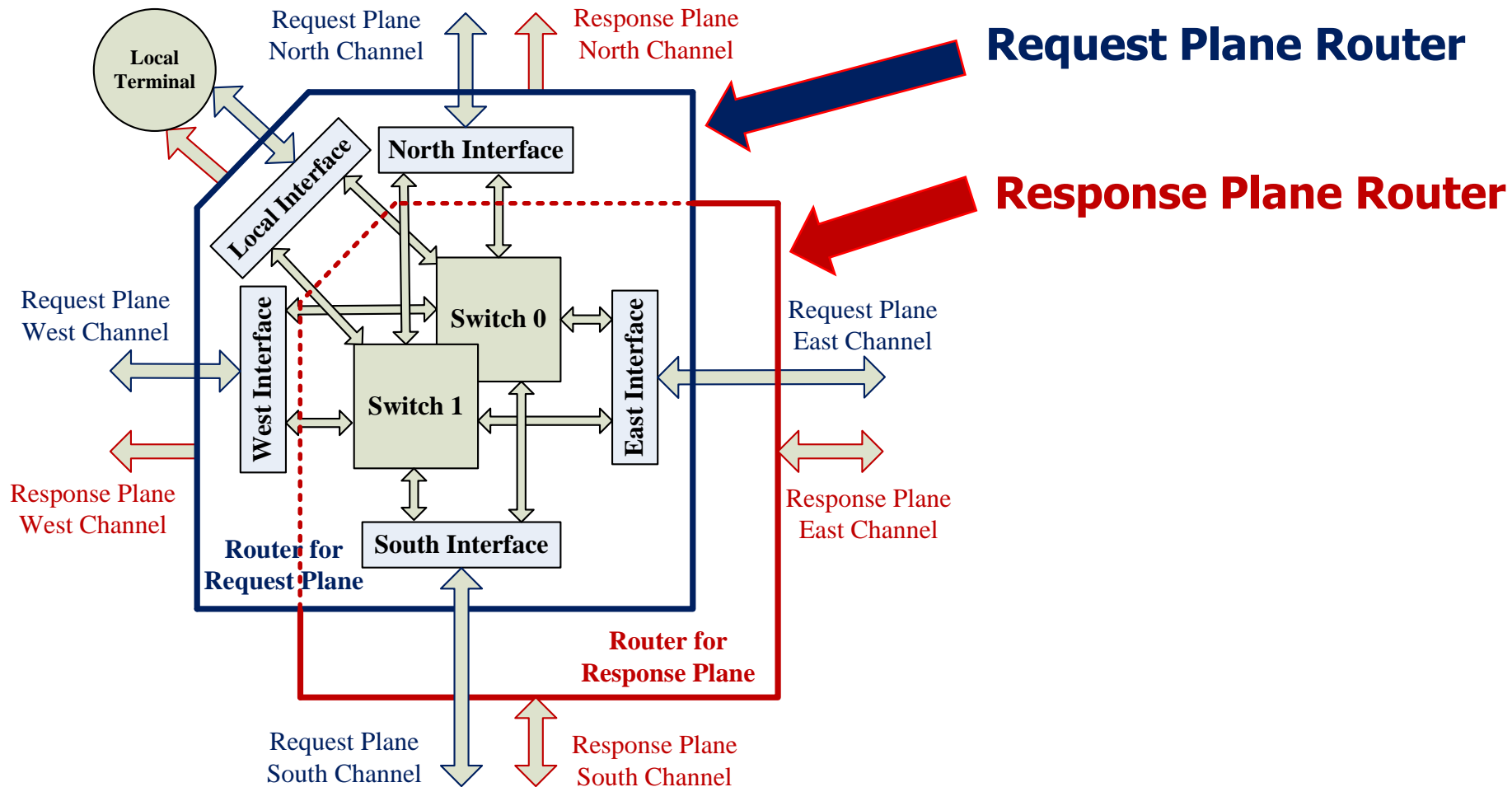
Contributions (2)

- Implementation and validation at **pre- and post-layout**
 - results presented only for pre-layout (**confidentiality reasons**)
- Industrial tools used in async design and validation
 - Functional validation tool (using Synopsys environment)
 - wrapper added for async design for sync environment re-use
 - used for **both pre- and post-layout** implementations
 - Place & Route tool (using AMD's internal tool environment)
 - largely **manual synthesis** + **automated P&R**
 - expect automated logic synthesis can be included with reasonable efforts
(e.g., an existing solution is proposed in [Ghiribaldi/Bertozzi/Nowick DATE-13])

Contributions (3)

- A novel async end-to-end credit-based Virtual Channel control scheme
 - Key idea = **lazy credit-update approach**
 - credit-increments are queued and no immediate update
 - credit updated only with a credit-decrement
 - fewer backward credit synchronization to upstream router
 - Potential increased throughput
 - VC is required for practical industrial usage
 - many existing async NoCs do not include VCs
 - Not the focus of this presentation (see paper for details)

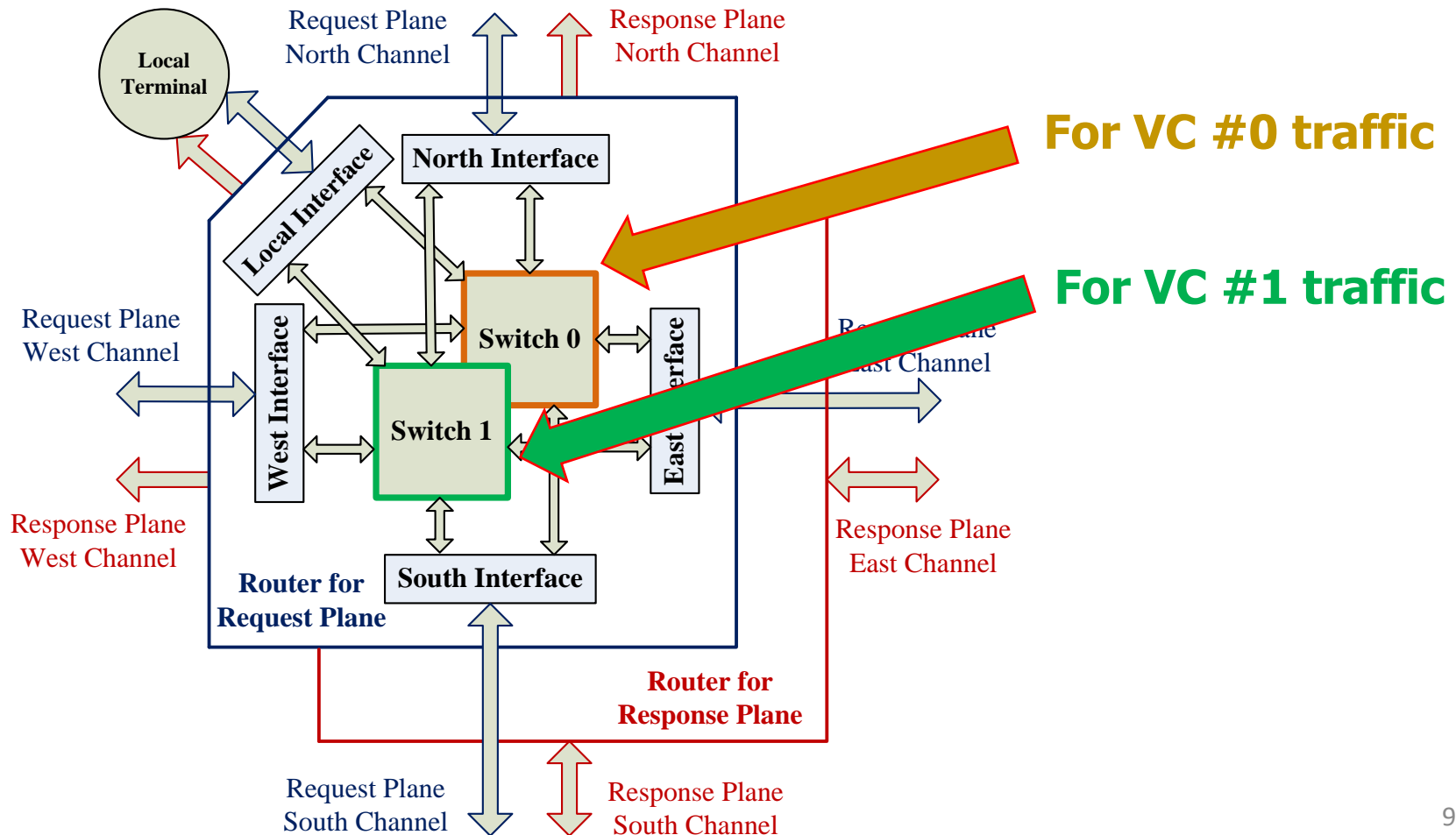
Proposed Asynchronous Node Structure



- Two identical and uncorrelated planes
- Follows AMD sync baseline router architecture

Proposed Asynchronous Node Structure (Cont.)

Switch replication inside each plane
- as many times as the number of VCs

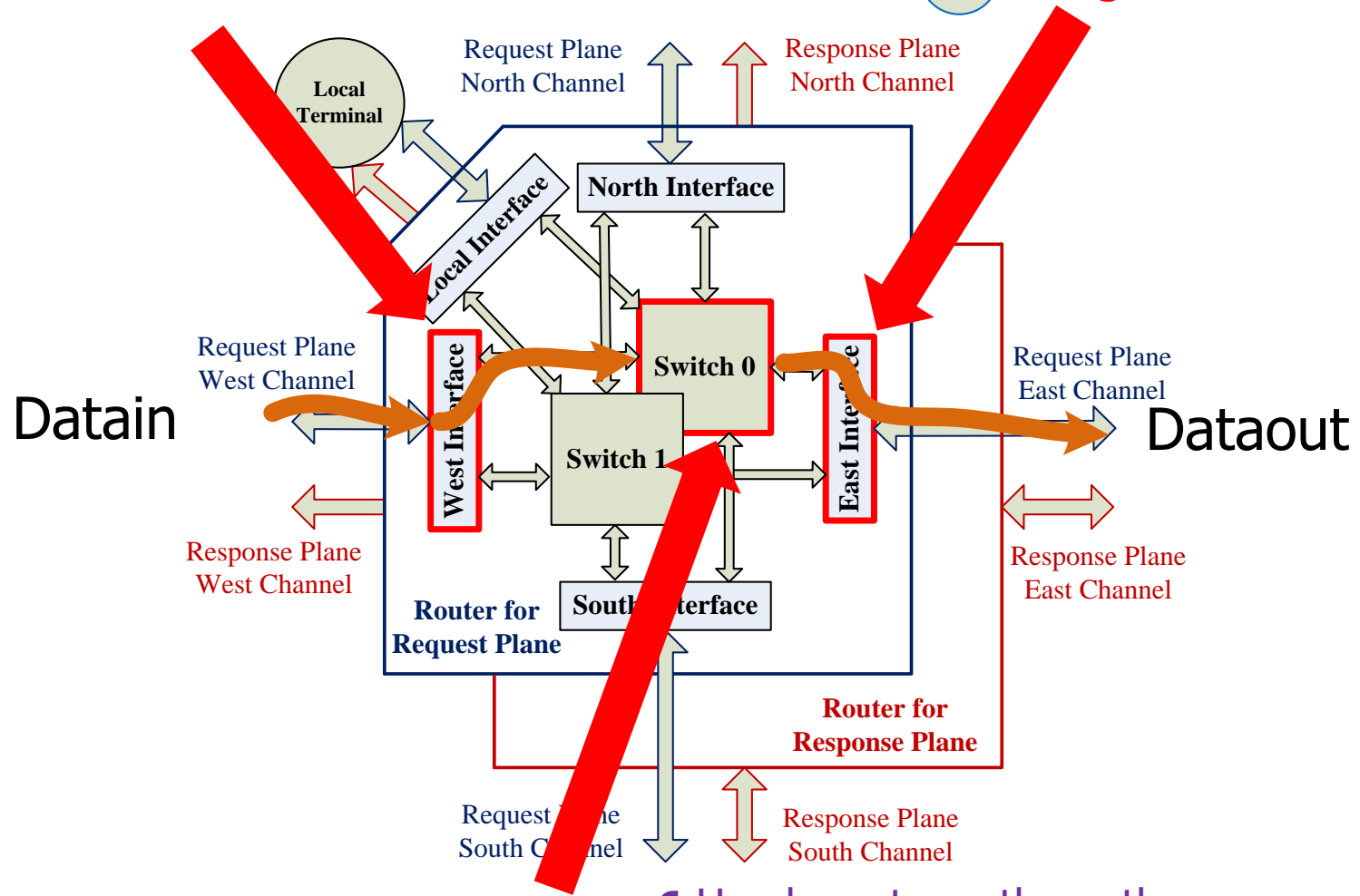


Node Operation

Example: data from west input -> east output

1 De-mux data to a switch

3 Merge data from 2 VCs



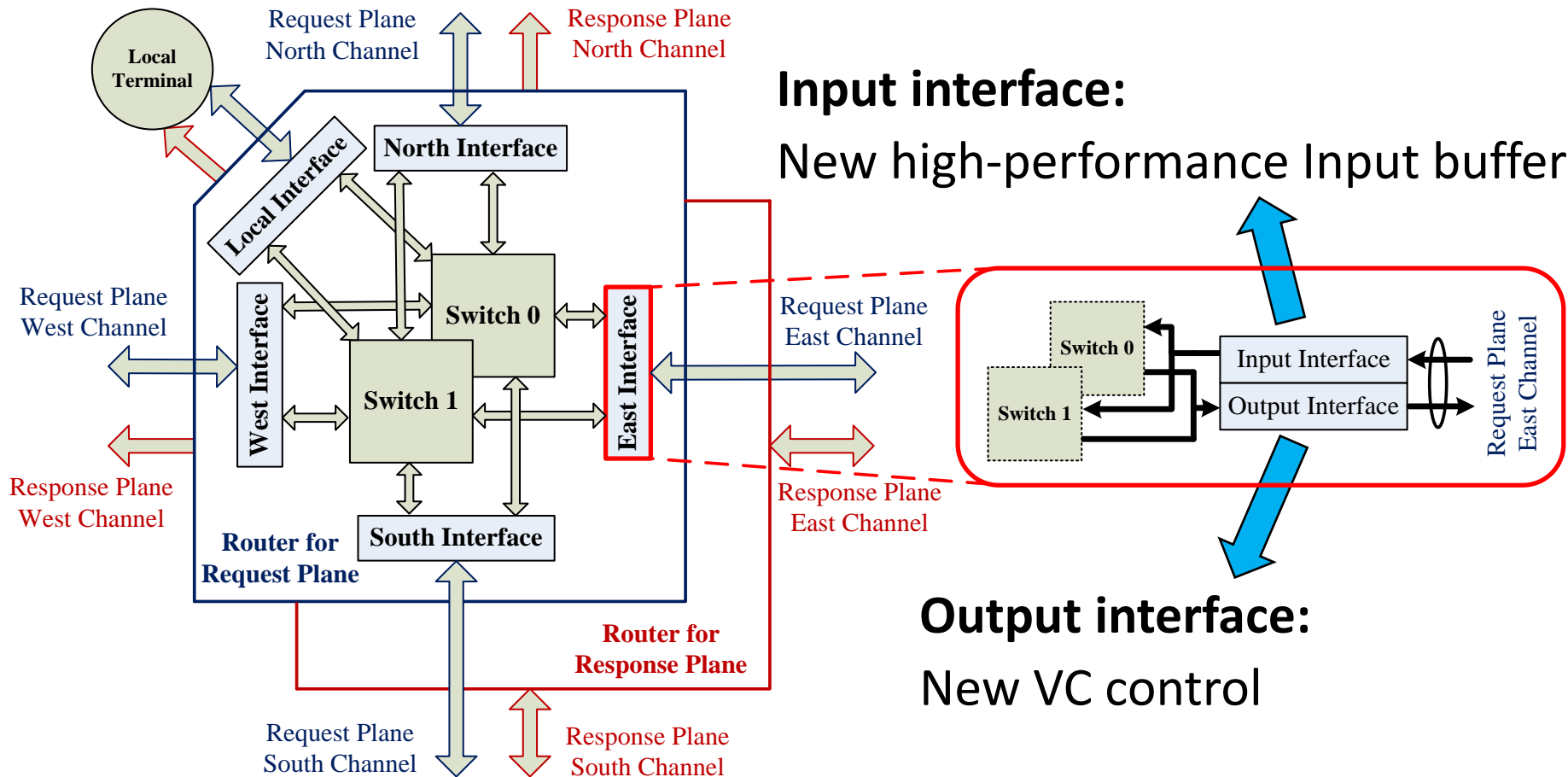
2 Data traverses the switch

{ Header sets up the path
Body/tail flits follow the pre-set up path

New Components in the Async Router

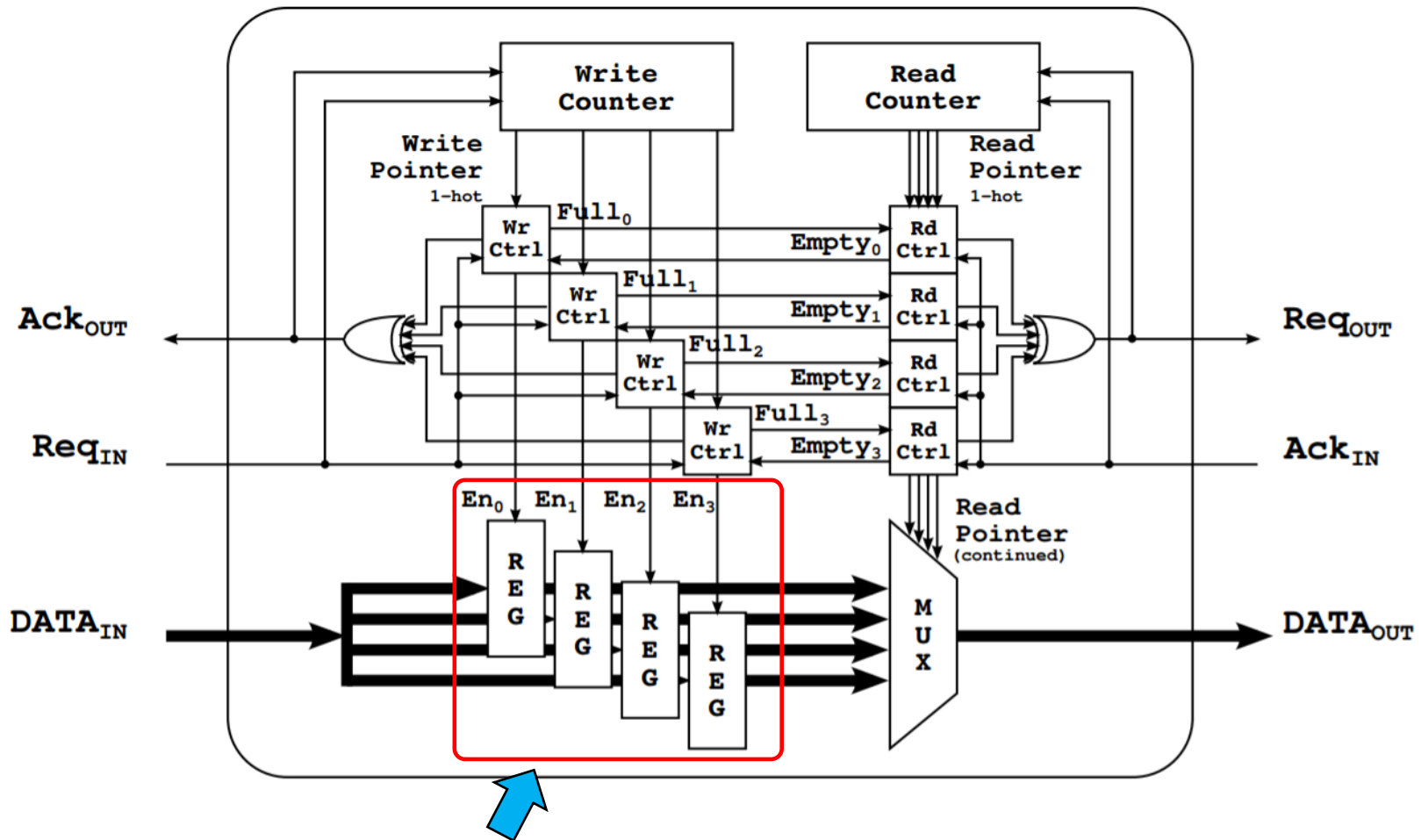
Two new components added on previous DATE-13 async router

[Ghiribaldi/Bertozzi/Nowick DATE-13]



Identical switches; new components in 'router interfaces'

Input Buffer Circular FIFO: Storage Element



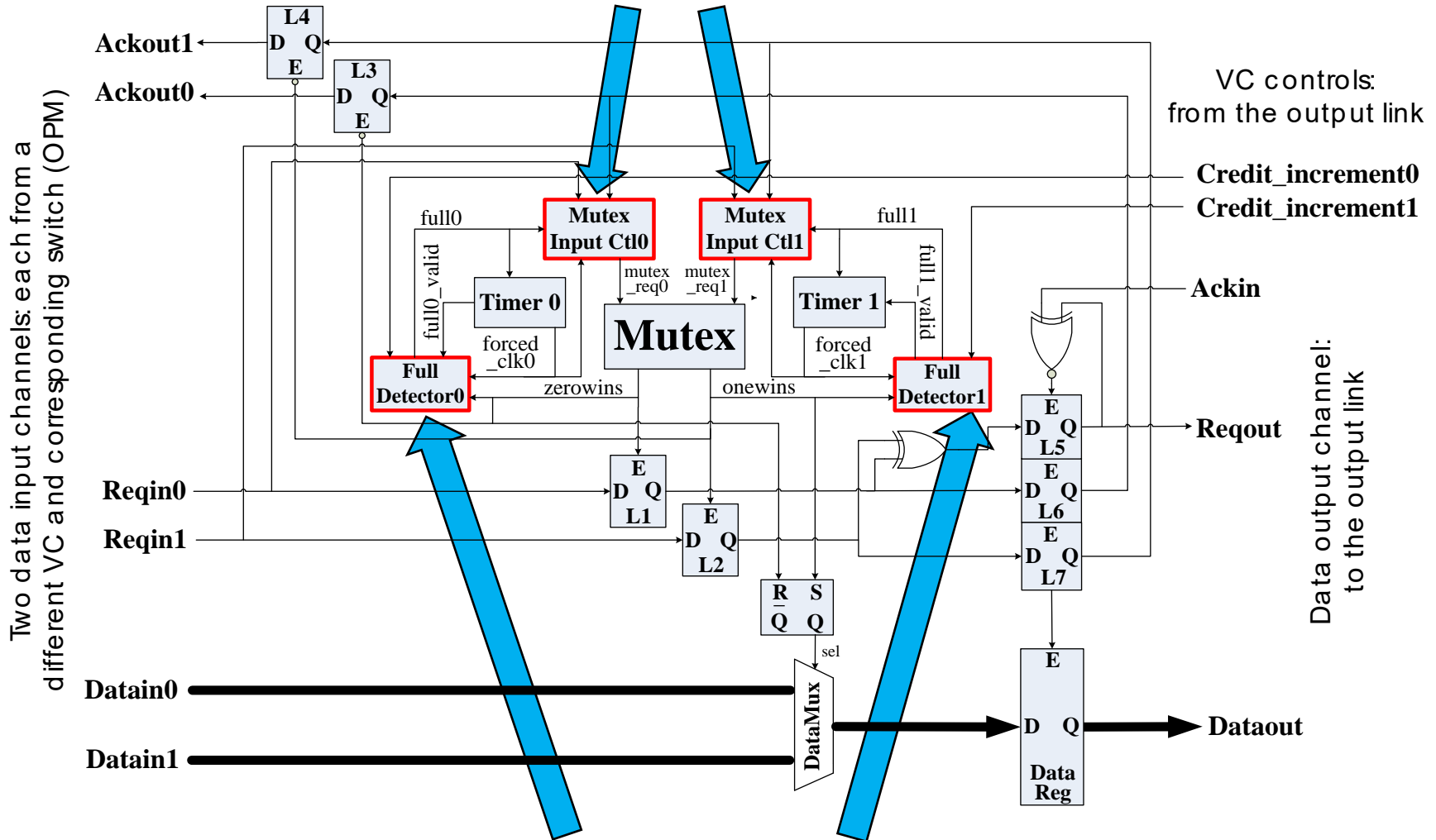
Each async storage element = single level-sensitive D-latch register

- Each latch register has **full storage capacity**
- **Half area/power cost** as a typical Flip-Flop storage in sync

➡ **key source** for performance/area/power benefits

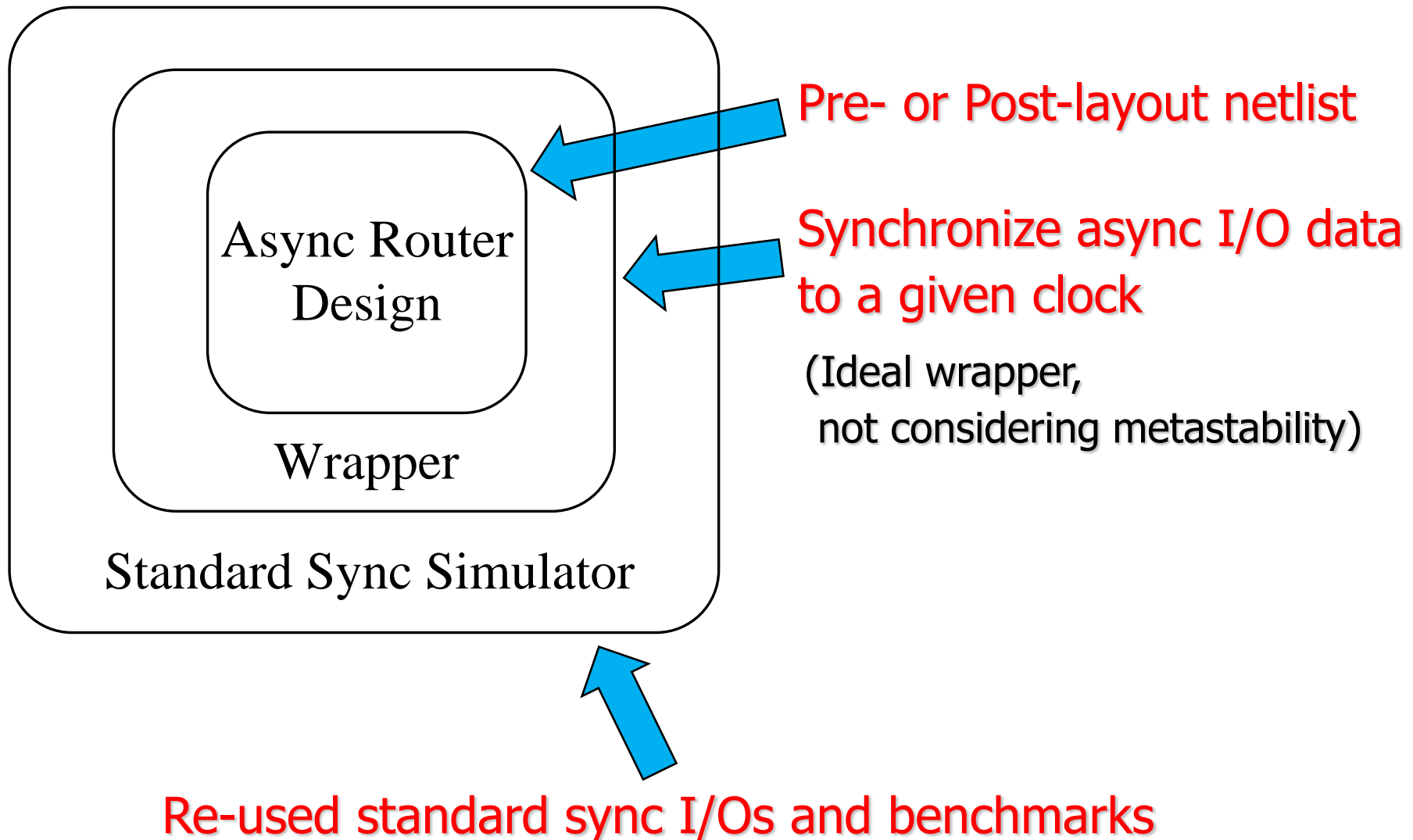
Output Interface Design: Proposed VC Control

Blocks or allows output traffic for a particular VC



Updates downstream credits only every time a flit is sent out
(See details in the paper)

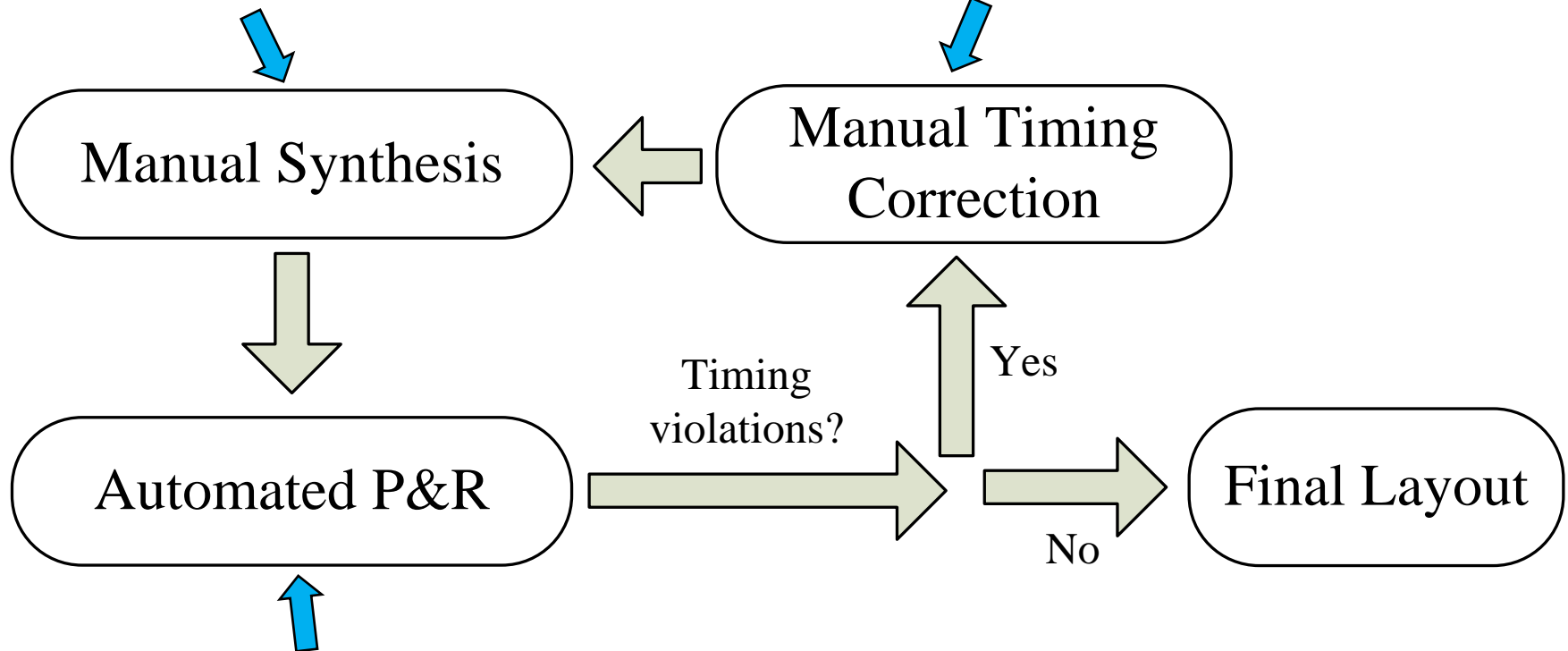
Design Validation Tool



Design Flow and Place & Route Tool

Manually derive gate netlist

Manually add inverter chains

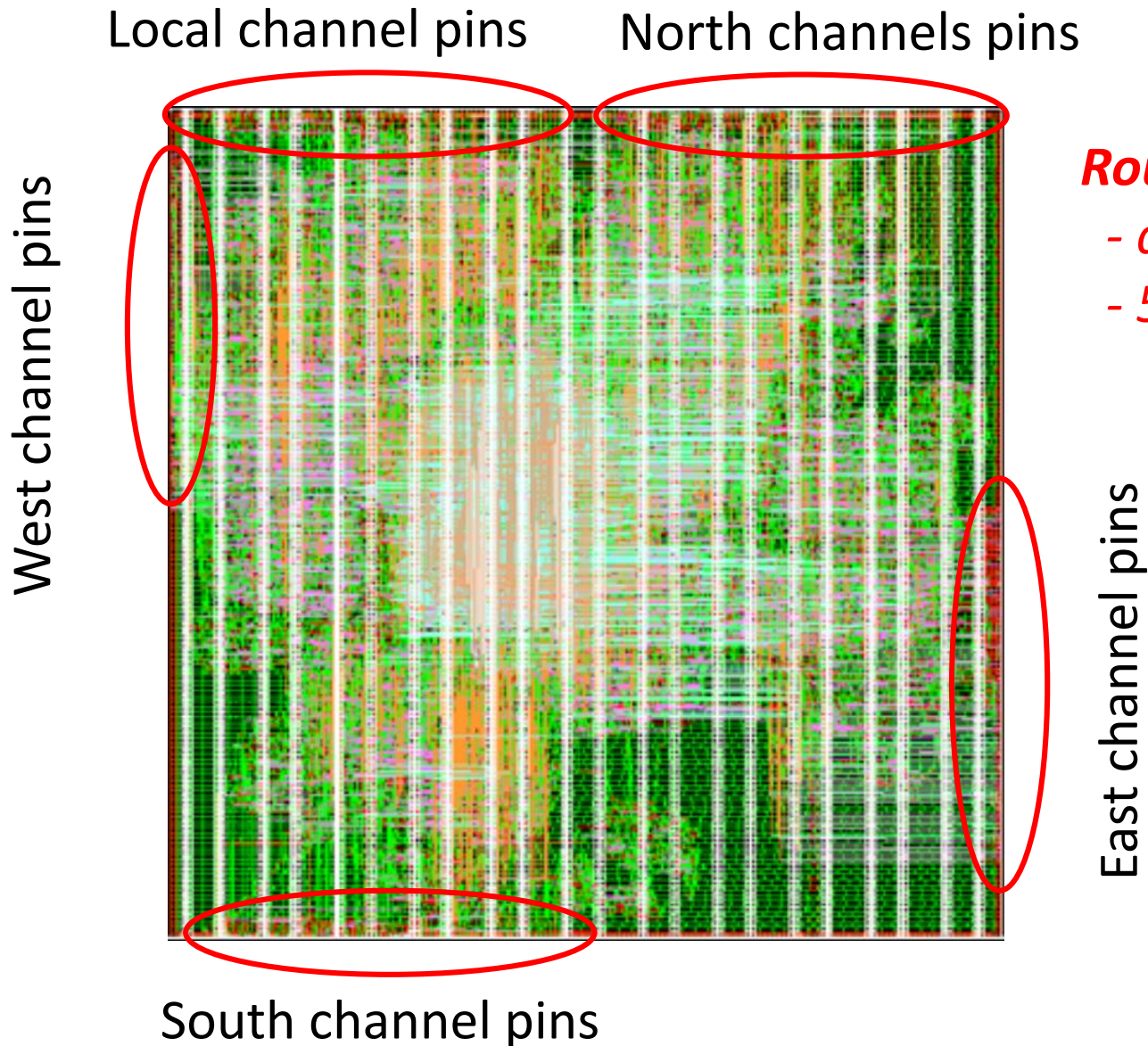


Standard sync P&R with 'don't touch' everything

Expect further synthesis automation can be included with reasonable effort

- An async logic synthesis solution was proposed in [Ghiribaldi/Bertozzi/Nowick DATE-13]

Actual Layout for Asynchronous Router



Router config.:

- double-plane router
- 5 port + 2 VCs

Experimental Results: Overview

➤ AMD commercial sync router vs. proposed async router

- Identical router configuration for both routers
 - 5-port + 2 VCs
 - buffer depth = 7 for each VC
- Pre-layout results only (for confidentiality reasons)
 - post-layout comparisons **expected to be similar** for small designs
- One testing benchmark: activating all switch ports
 - evenly distributed traffic from all inputs to all outputs
 - sufficient for initial router-level results
- Testing corner: 14nm FinFET library (0.65V, TT)

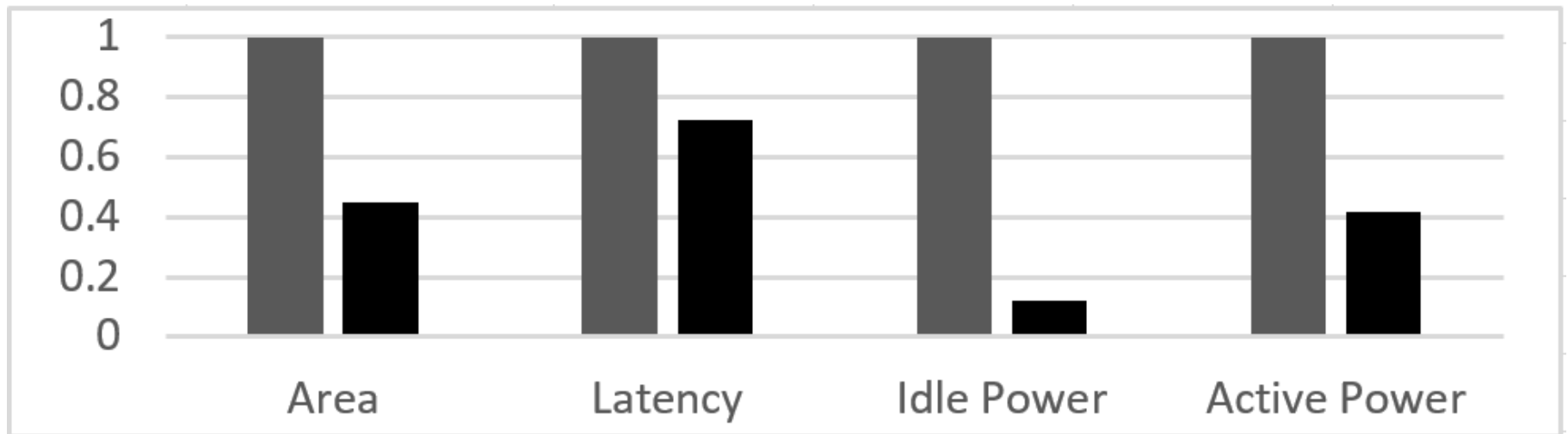
➤ Additional projected results for more complex routers

- 7-port router with 2 VCs ➡ **for 3D stacking**
- 5-port router with 8 VCs ➡ **more realistic VC configuration**

Basic comparison: 5-port router with 2 VCs

- Asynchronous router dominates in area, latency and power

Comparison for 5-port router with 2 VCs



55% lower

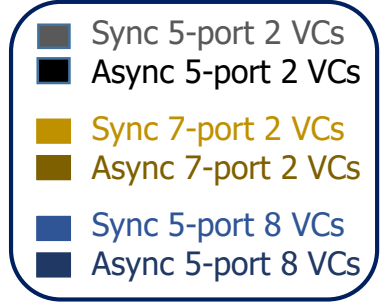
28% lower

88% lower

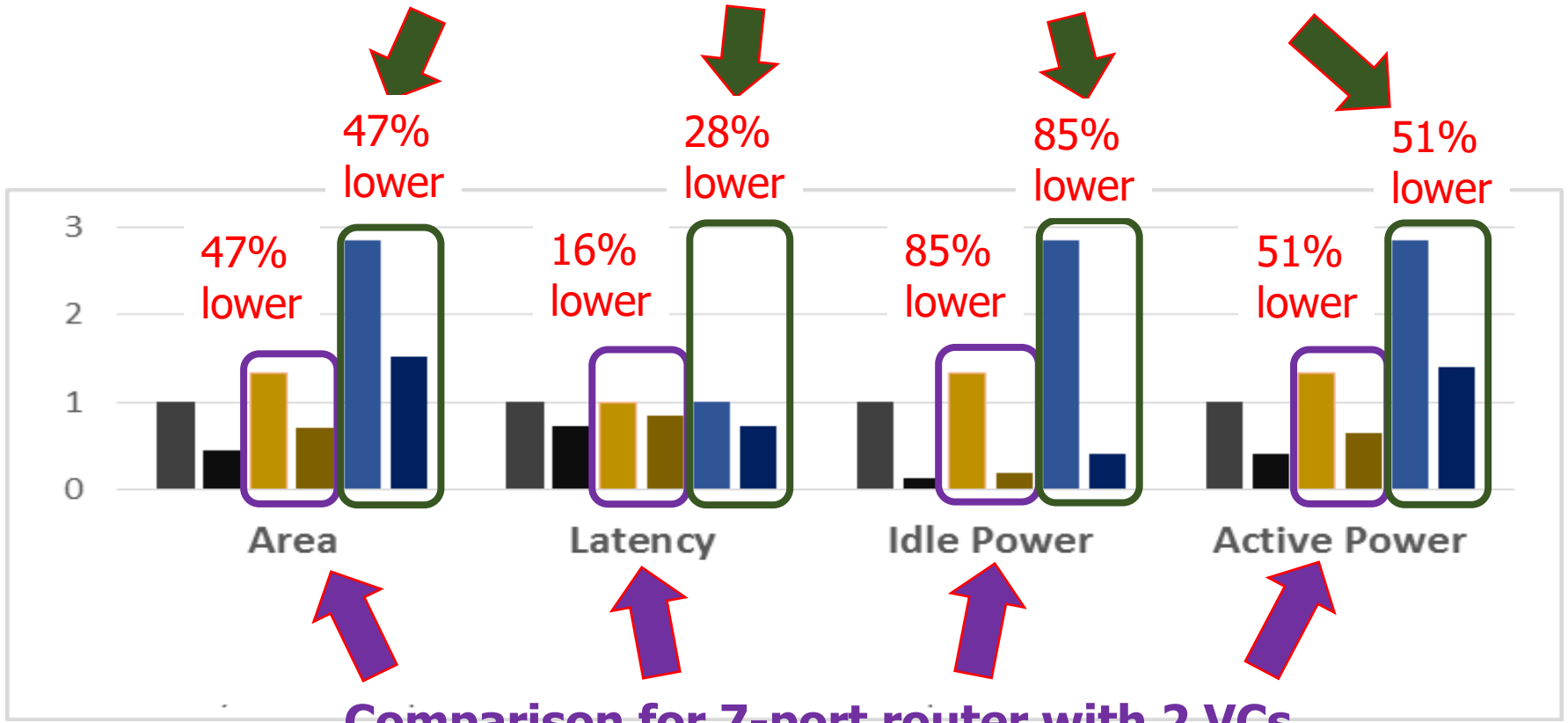
58% lower

Projected Results for More Complex Routers

- Absolute area and power costs are noticeably increased - due to higher radix or more VCs
- Relative asynchronous benefits are largely maintained




Comparison for 5-port router with 8 VCs



Comparison for 7-port router with 2 VCs

Conclusions

- First “async vs. commercial sync router” in advanced library
 - Sync router optimized for high-end products with fine-grain clock-gating
 - Comparison in 14nm FinFET library
- Industrial tools for async design and validation
 - Design validation tool: sync testing environments are largely re-used
 - Manual synthesis + automated P&R
 - synthesis automation can be further included with some effort
 - Shows opportunity for industrial asynchronous designs
 - Some remaining tool challenges for full automation
- A novel async end-to-end  credit-based VC control approach
 - Lazy credit-update approach potential higher throughput
- Results: async router shows significant benefits
 - In key metrics: area, latency and power