A Low-Area Interconnect Architecture for Chip Multiprocessors

Zhiyi Yu and Bevan Baas

VLSI Computation Lab
ECE Department, UC Davis

Outline

- **Motivation**
  - Why chip multiprocessors
  - The difference between chip-multiprocessor interconnect and multiple-chip interconnect
- Low-area asymmetric interconnect architecture
- High performance multiple-link architecture

Why Chip Multiprocessors

- Chip Multiprocessor challenges
  - Traditional high performance techniques are less practical (e.g., increased clock frequency)
  - Power dissipation is a key constraint
  - Global wires scale poorly in advanced technologies
- Solution: chip multiprocessors
  - Parallel computation for high performance
  - Reduce the clock freq. and voltage when full rate computation is not needed for high energy efficiency
  - Constrain wires no longer than the size of one core

Interconnect in Chip Multiprocessors vs. Interconnect in Multiple Chips

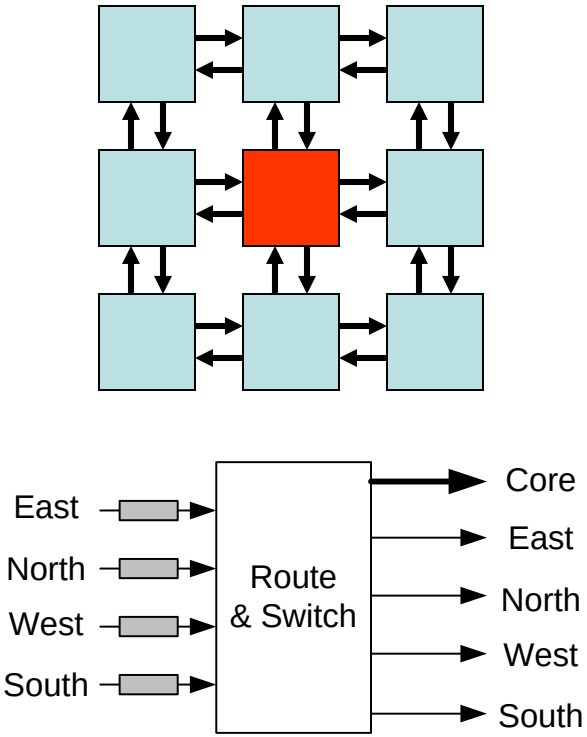
- Chip multiprocessors have relatively limited area resources for interconnect circuitry
  - Try to reduce the area of buffer with little reduction of the interconnect capability
- Chip multiprocessors have relatively abundant wire resources for inter-processor connection
  - Try to use more wire resources to increase the interconnect capability

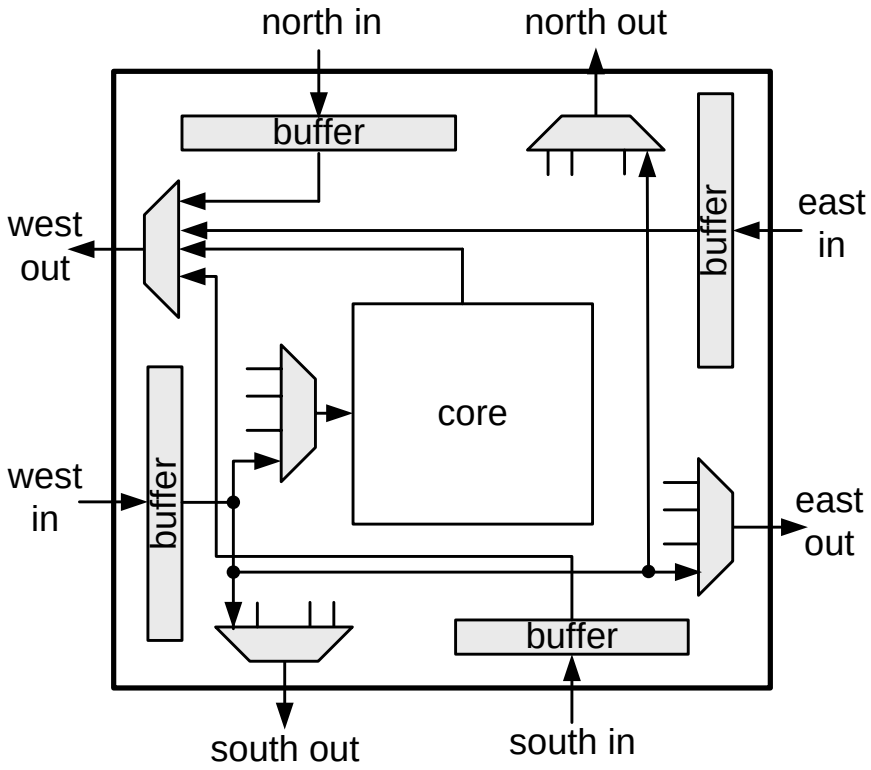
Outline

- Motivation
- **Low-area asymmetric interconnect architecture**
  - Traditional dynamic Network-On-Chip
  - Statically configured Network-On-Chip
  - Proposed asymmetric architecture
- High performance multiple-link architecture

Background: Traditional Network on Chip (NoC) using Dynamic Routing

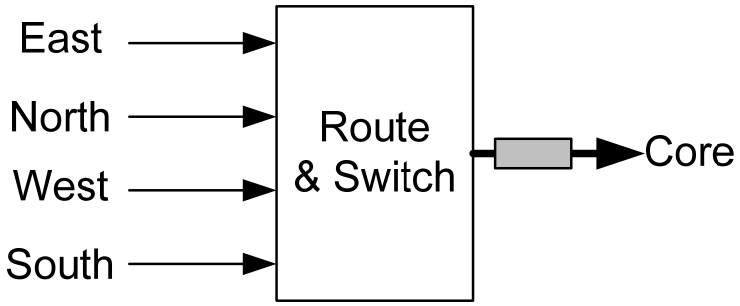
- Advantage: flexible
- Disadvantage: high area and power cost

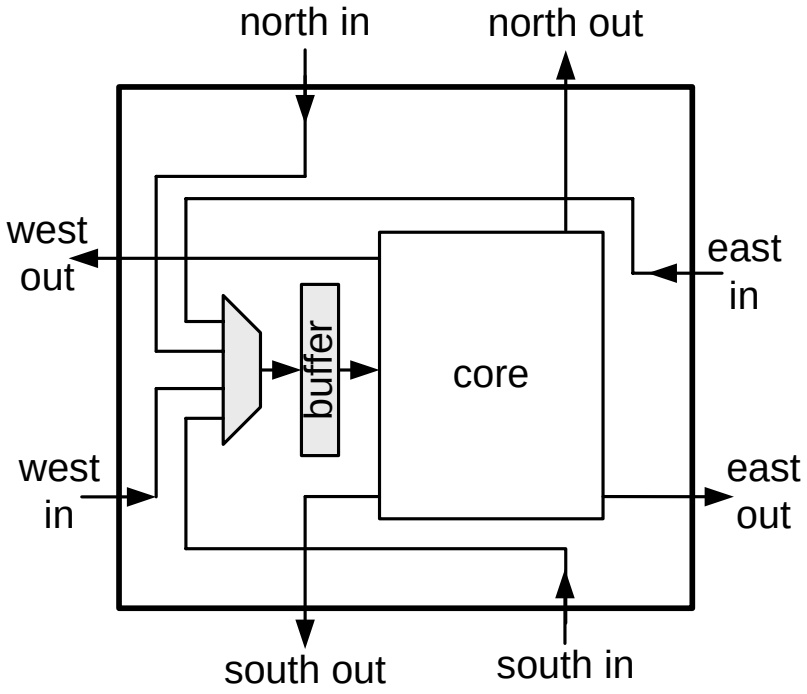




Background: Static Nearest Neighbor Interconnect Architecture

- Low area cost
  - One buffer per processor, not four
- High latency for long distance communication
  - Data passes through each intermediate processor



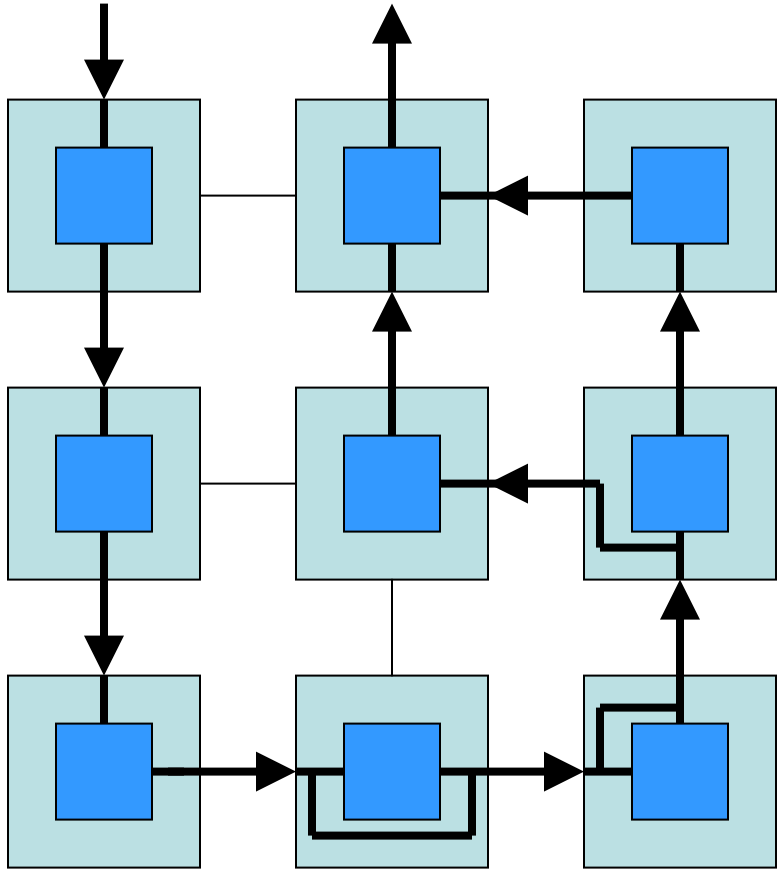


Asymmetric Data Traffic in Inter-processor Communication

- Data traffic of routers in a 9-processor JPEG encoder
- 80% of traffic goes to processor core, 20% passes by core

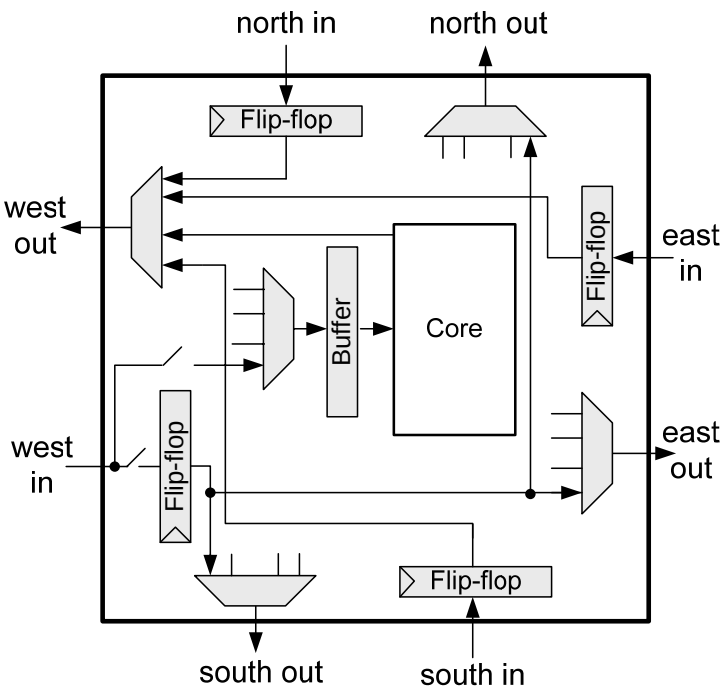
	Network data words of **input** ports of router			
	East	North	West	South
Relative	9%	26%	22%	43%

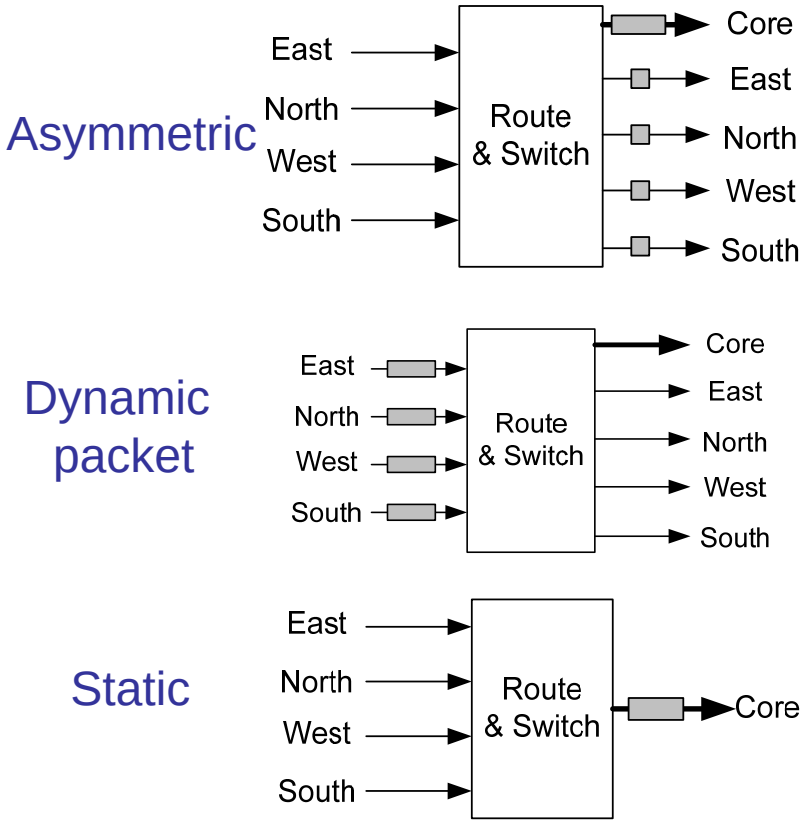
	Network data words of **output** ports of router				
	Core	East	North	West	South
Relative	**80%**	8%	4%	8%	0%



Asymmetrically-Buffered Interconnect Architecture

- Large buffer only for the processing core
- Support flexible direct switch/route interconnect



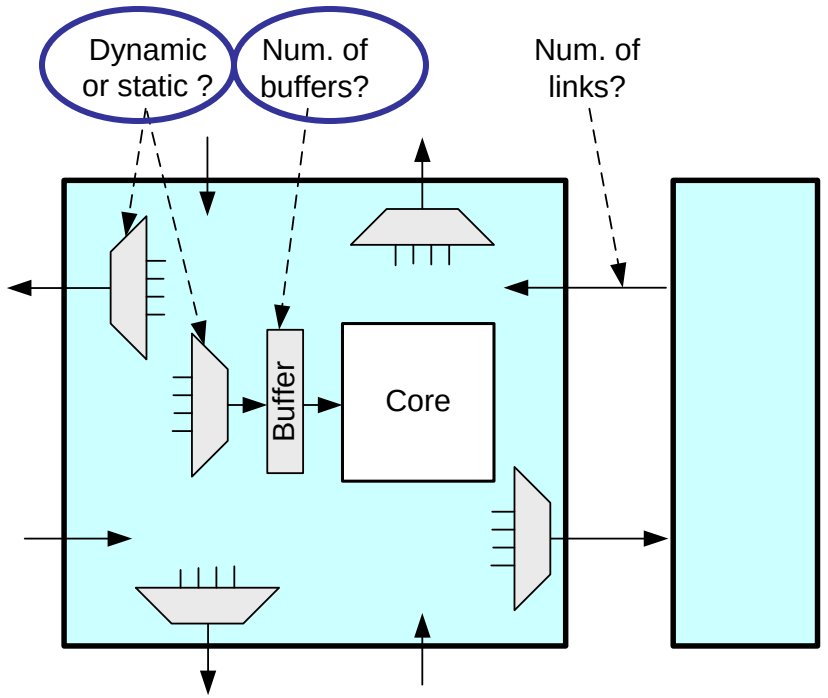


Outline

- Motivation
- Low-area asymmetric interconnect architecture
- **High performance multiple-link architecture**

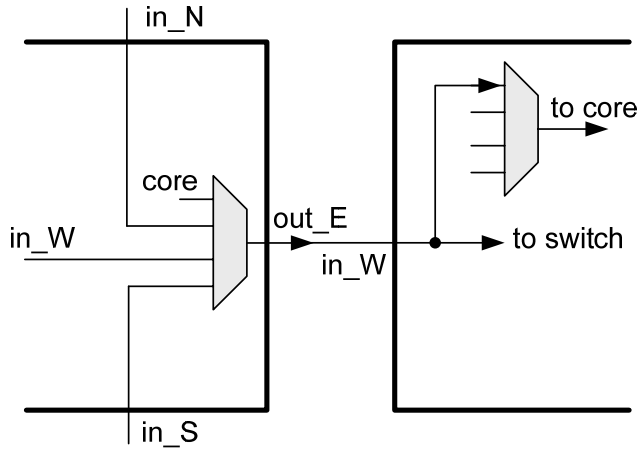
Design Option Exploration

- Choose static routing architecture
  - Much smaller area and power required
- Assign two buffers (ports) for each processing core
  - Natural fit with 2-input, 1-output instructions (e.g., *C=A+B*)

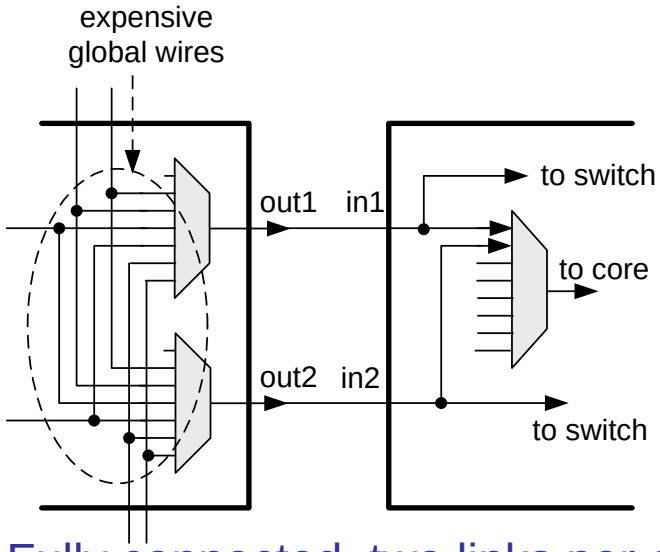


Single Link vs. Multiple Links

- Why consider multi-link architectures?
  - Single link might be inefficient in some cases
  - There are many link/wire resources on chip
- Fully connected multi-link architectures have huge numbers of long wires
  - Alternative simplified architectures are needed



Type1: one link per edge

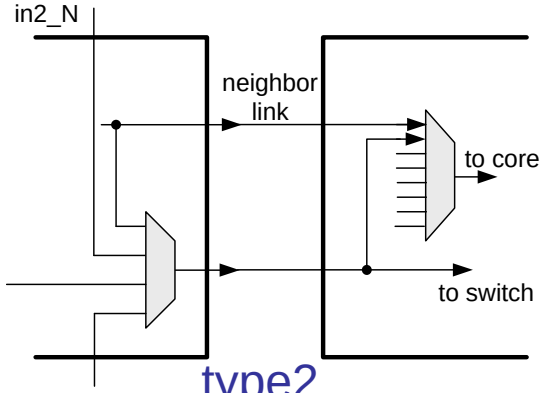


Fully connected, two links per edge

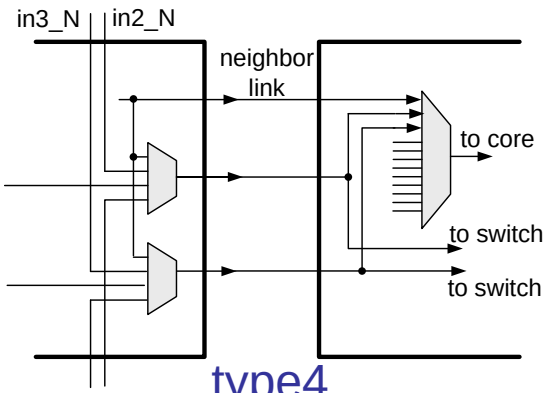
Architectures with Multiple Links

Two links per edge | Three links per edge | Four links per edge

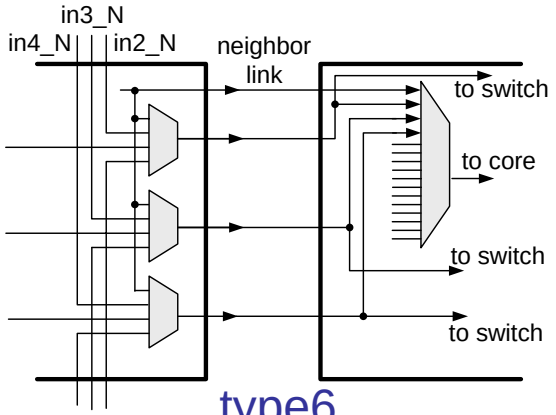
One link dedicated to neighbor link



type2

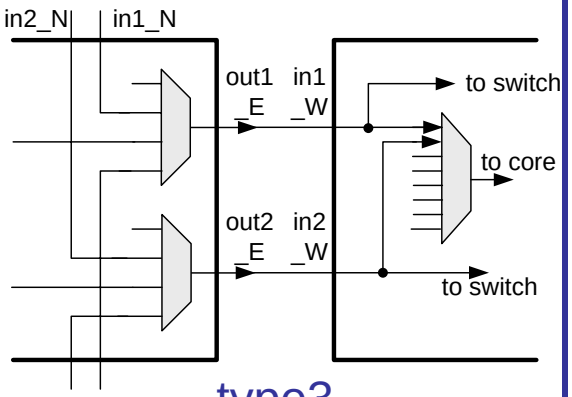


type4

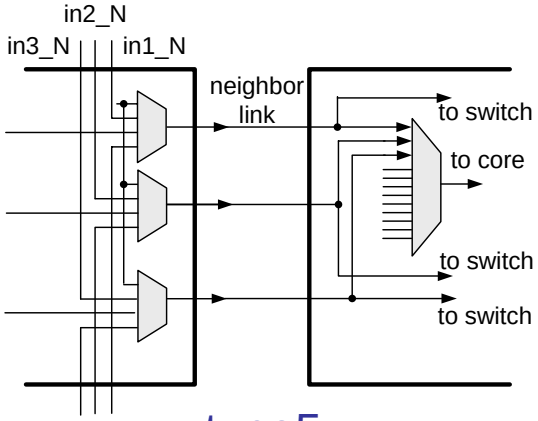


type6

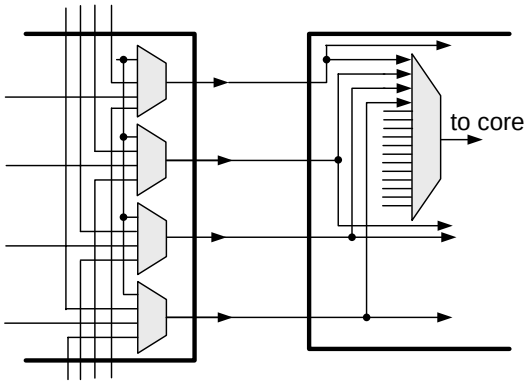
All links the same



type3



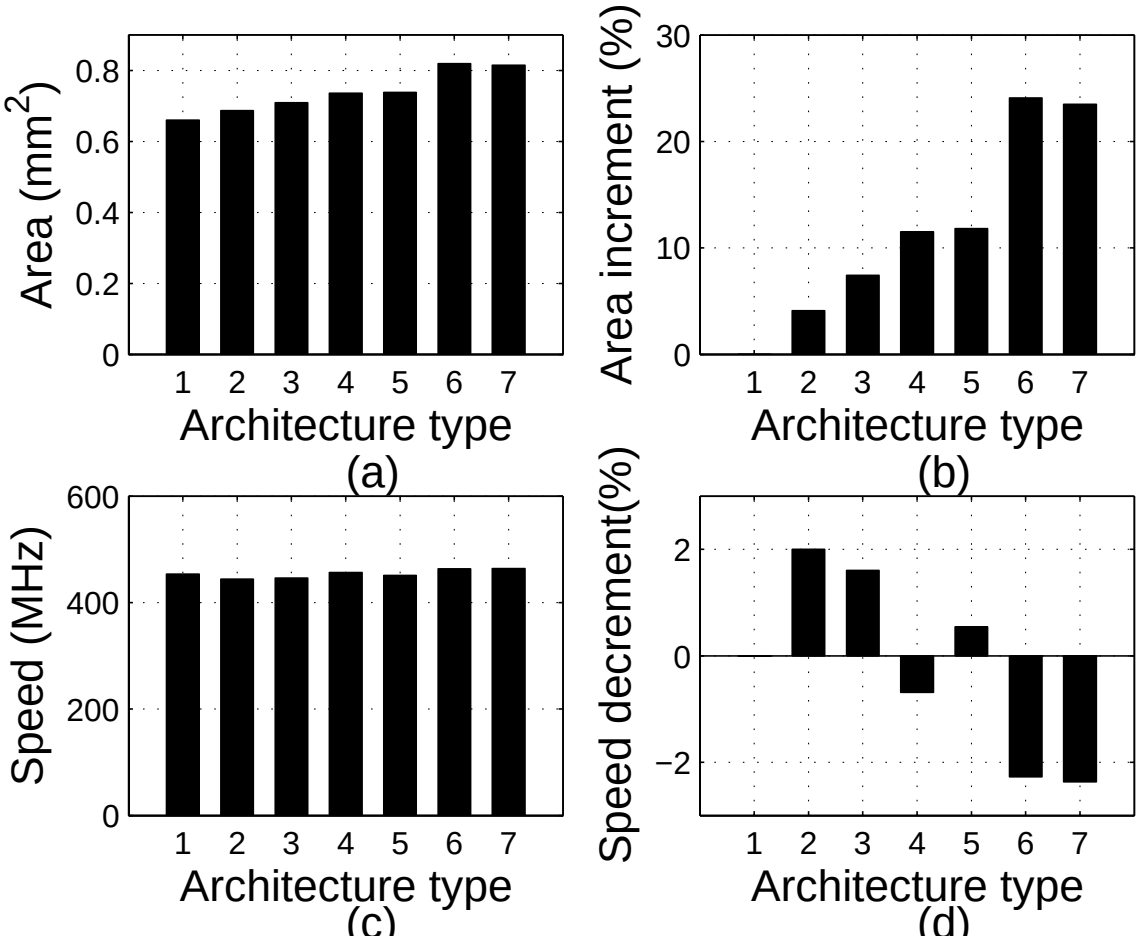
type5



type7

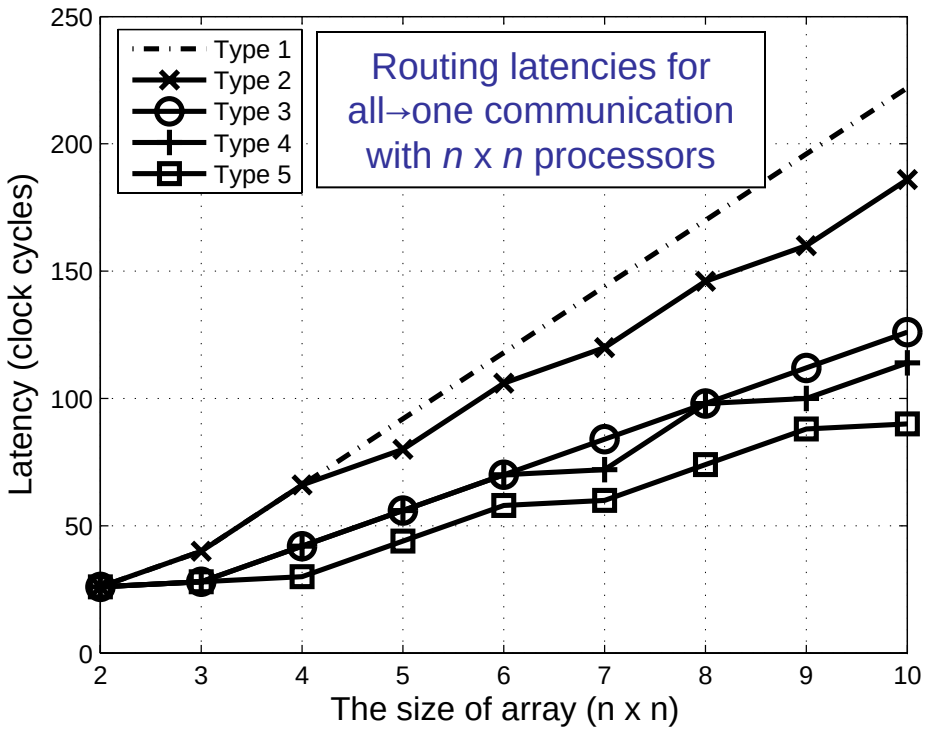
Area and Speed of Seven Proposed Network Architectures

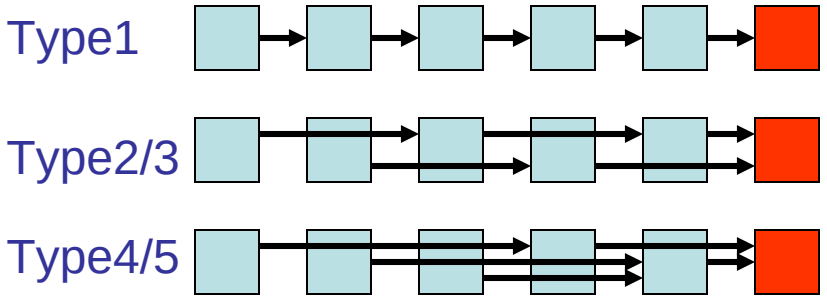
- All seven architectures implemented in hardware
- Four-link architectures (types 6 and 7) require almost 25% more area
- Clock rates for all are within approx. 2%



Comparing Routing Latency using Communication Models

- All architectures have the same routing latency for the one→one, one→all, and all→all communication
- Different types of architectures behave differently for the all-one communication





Example all→one linear communication for six processors

Summary

- Asymmetric buffer allocation topology uses 1 or 2 large buffers instead of 4 large buffers to support long distance communication
  - Provides **2** to **4** times area savings with little performance reduction
- Using **2** or **3** links per edge achieves good area/performance tradeoffs for chip multiprocessors containing simple single-issue processors
  - Provides **2** to **3** times higher all→one communication capacity with 5%-10% increased area

Acknowledgements

- Funding
  - Intel Corporation
  - UC Micro
  - NSF Grant No. 0430090
  - CAREER award 0546907
  - SRC GRC Grant 1598
  - IntellaSys Corporation
  - S Machines
- Special thanks
  - Members of the VCL, R. Krishnamurthy, M. Anders, and S. Mathew
  - ST Microelectronics and Artisan