

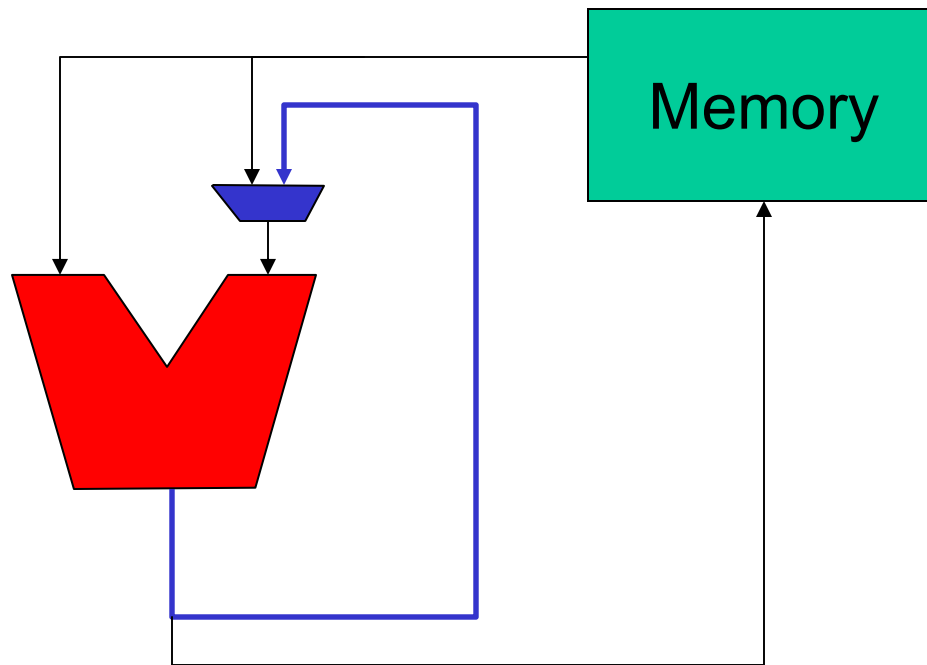
Routed Inter-ALU Networks for ILP Scalability and Performance

Karthikeyan Sankaralingam

Vincent Ajay Singh, Stephen W. Keckler, and Doug Burger

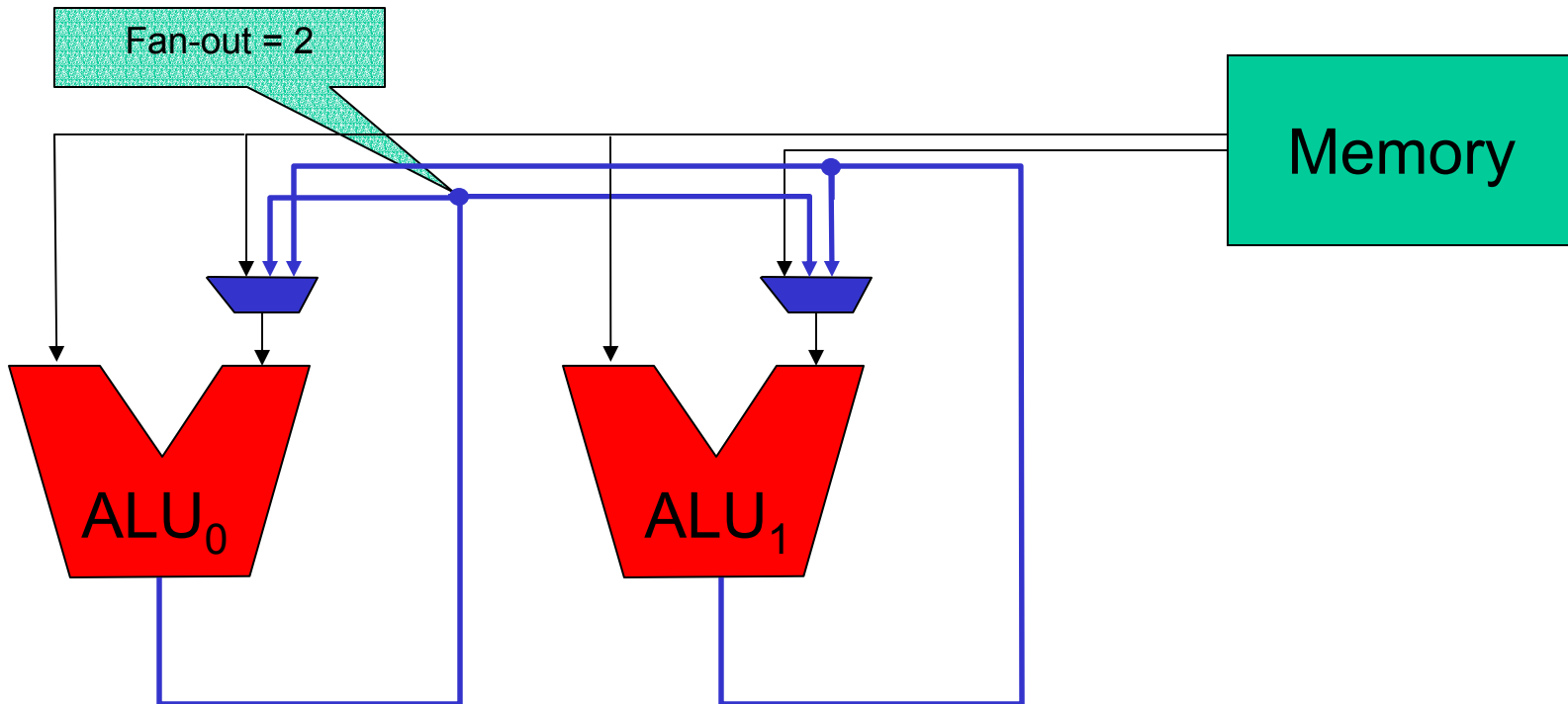
Computer Architecture and Technology Laboratory
Department of Computer Sciences
The University of Texas at Austin

Bypass paths



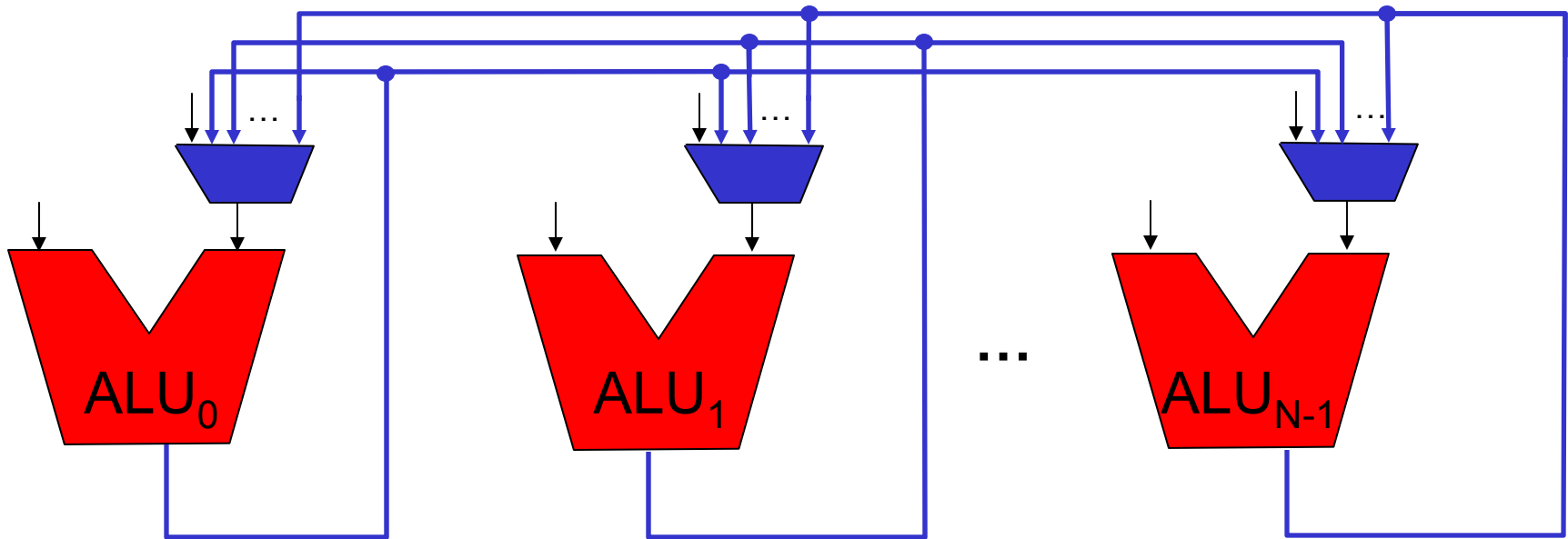
- Single ALU with local bypass
- Fan-out and fan-in = 1

Bypass paths



- Fan-out, fan-in = 2
- More wires, longer distances

Bypass paths



- Bypassing complexity scales with pipeline width and depth
 - Fan-out and fan-in increases
- Wiring complexity increases with bypassing complexity
 - Area increases due to wiring

Bypass path delays

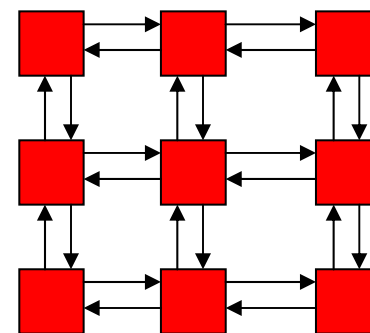
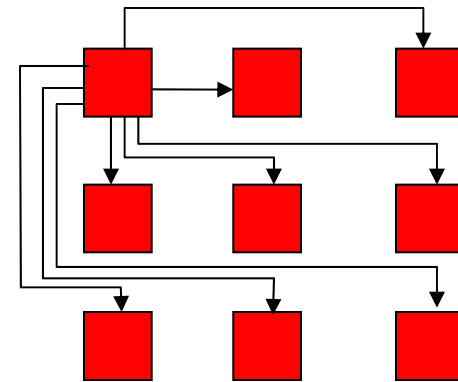
	4-issue	16-issue	64-issue
Shortest path	0.8	0.8	1.3
Longest path	1.5	3.9	8.6

Circuit simulation at 10 F04 clock cycle, 100 nm technology
Delays shown in cycles

- Bypass delays vary with distance and issue width
- In most modern processors, bypass delay is fixed and is a fraction of clock cycle

Solution: Routed Inter-ALU networks

- Current bypass paths are implemented with all-to-all broadcast
- Routed Inter-ALU networks (RIANs)
 - Point-to-point links between neighboring nodes
 - Multiple hops to communicate between distant nodes
- + Fewer wires: lower wiring area overhead
- + Low fan-in and fan-out delay
- Router delay at every hop
- Destinations must be known
- Circuit complexities



Outline

- Taxonomy of Inter-ALU networks
- Circuit analysis
- RIANs in different architectures
 - Dynamically scheduled architectures
 - Statically scheduled VLIW architectures
 - Statically scheduled, dynamically issued Grid Processor architectures
- Conclusion

Taxonomy of Inter-ALU networks

- All-to-all broadcast bypass networks are part of a broad class of Inter-ALU networks
- A taxonomy with 8 types of networks
- Classified along 3 axes

Execution model	Outputs broadcast to all ALUs	Targets specified explicitly, operands sent point-to-point
Network Architecture	Operand sent directly from producer to consumer (single-hop)	Operand may pass through multiple-hops
Router control	Static	Dynamic

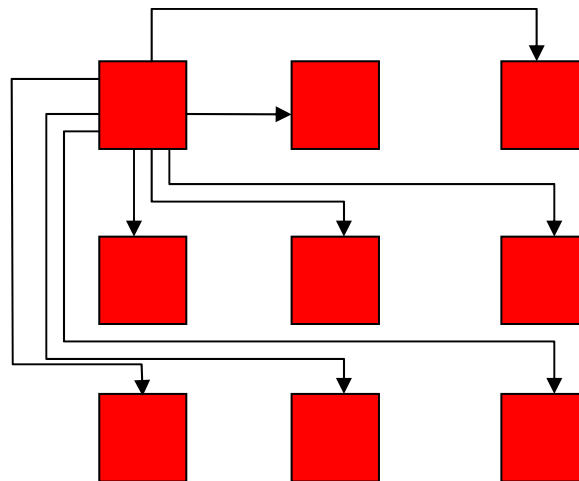
Taxonomy of Inter-ALU networks

Execution model	Network architecture	Router control	Acronym	Examples
Broadcast	Single-hop	Dynamic	BSD	Superscalar
Broadcast	Single-hop	Static	BSS	VLIW
Broadcast	Multi-hop	Dynamic	BMD	Alpha 21264
Broadcast	Multi-hop	Static	BMS	-
Point-to-point	Single-hop	Dynamic	PSD	M-Machine, Multicluster
Point-to-point	Single-hop	Static	PSS	Degenerate PMS
Point-to-point	Multi-hop	Dynamic	PMD (RIAN)	Grid Processor
Point-to-point	Multi-hop	Static	PMS	RAW

Circuit analysis (Network architecture)

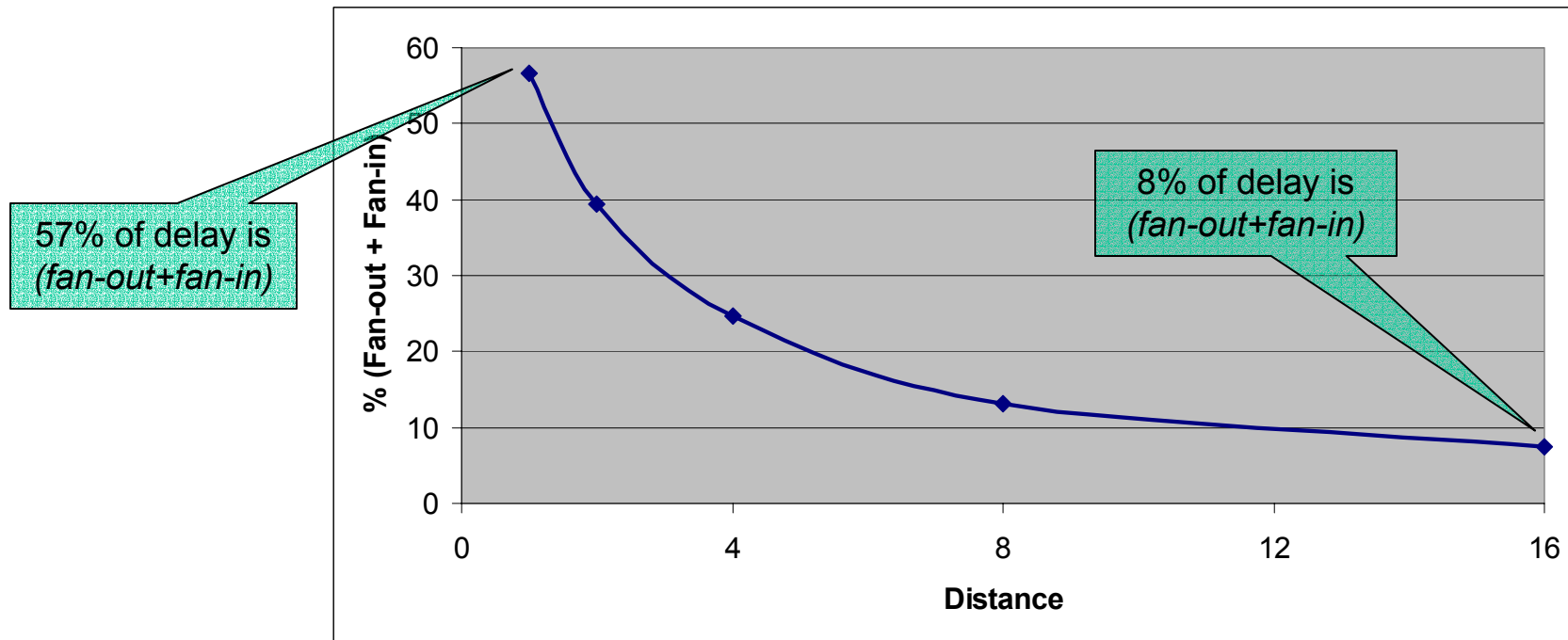
- Analyze single-hop and multi-hop networks
 - Examine circuit delay components:
Wire delay, fan-out delay, fan-in delay
 - Wiring overhead
 - Compare networks
- Modeling
 - SPICE simulation using 100nm technology libraries
 - ALUs arranged in rectangular grid, Manhattan routing
 - Each ALU is $32K\lambda$ on a side
 - Distances measured in *segments*: 1-segment = distance between adjacent ALUs

Single-hop networks



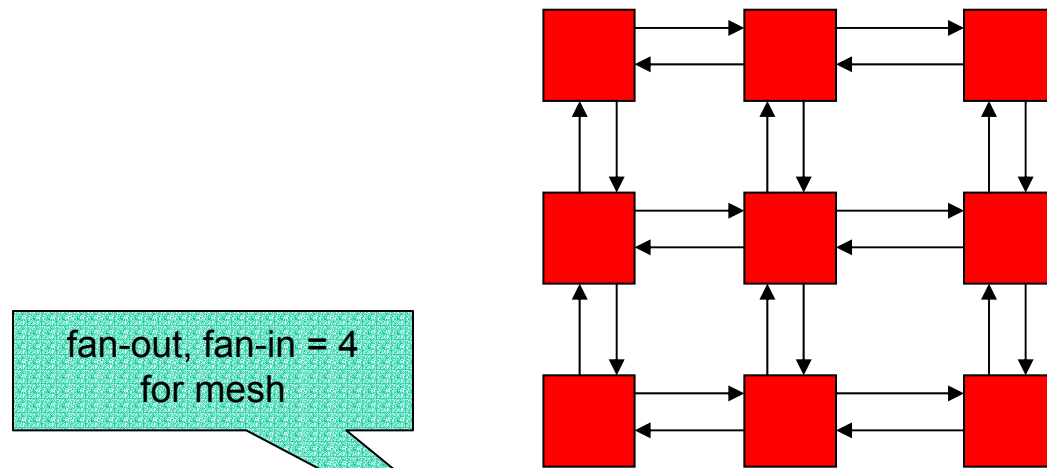
- Delay = Fan-out + Fan-in + Wire
- Fan-out + Fan-in = 240ps on 64-wide network
- Fan-out + Fan-in dominates for short distances
- Wire delay dominates for long distances

Single-hop networks (64 nodes)



- Reducing (*fan-out + fan-in*) helps in short-distance communication

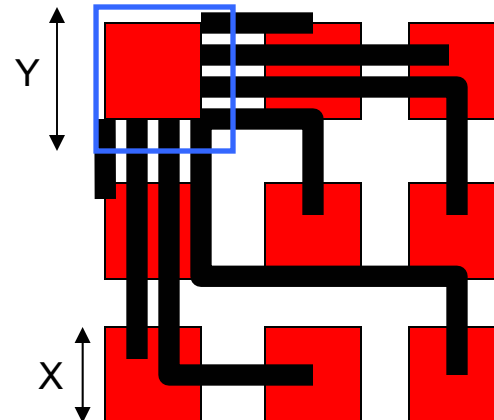
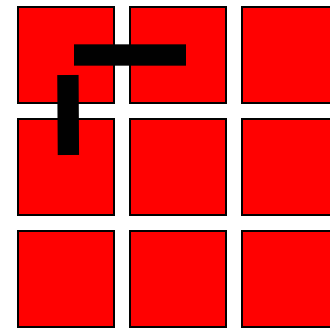
Multi-hop networks



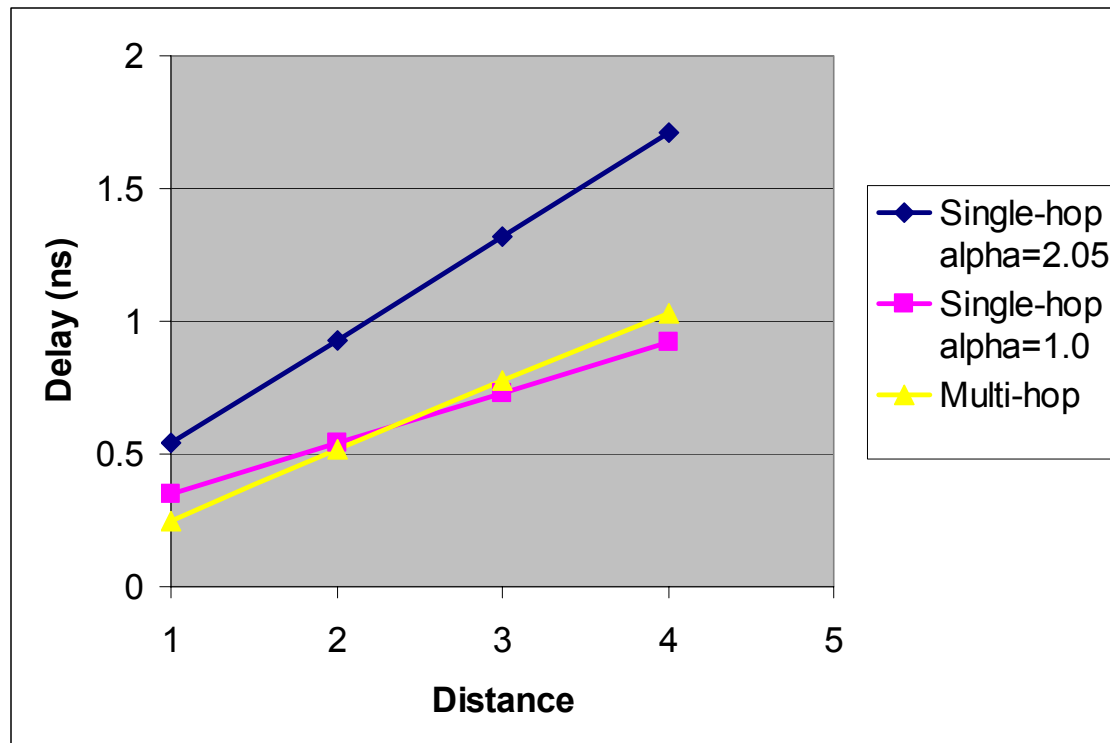
- Delay = (Fan-out + Fan-in + ~~Router~~) * n_{hops} + Wire
- Fan-out + Fan-in = 100ps for mesh network
- Far fewer wires, no wiring overhead
- Router delays accumulate for long distances
 - Can be effectively hidden through a lookahead arbitration scheme
 - Routed Inter-ALU Network (RIAN)

Wiring area overhead

- $\text{area}_{\text{wiring tracks}} < \text{area}_{\text{ALUs}}$
 - No wiring overhead, $\alpha = 1$
- When $\text{area}_{\text{wiring tracks}} > \text{area}_{\text{ALUs}}$
 - α indicates increase in wiring length
 - $\alpha = Y/X$
- Conservative area models show:
 - $\alpha = 1$, for all networks ≤ 32 nodes
 - $\alpha = 2.05$, for single-hop 64-wide networks



Single- vs. Multi-hop networks (64 nodes)



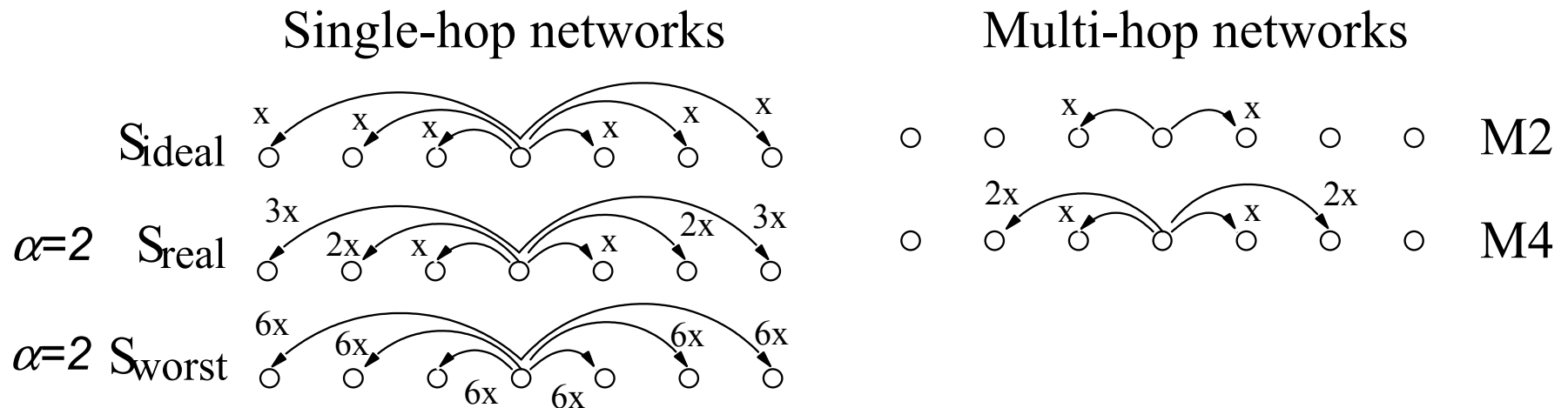
α	Crossover dist.
1.0	3
1.2	4
1.4	10
1.5	Never

- With a conservative wiring overhead model, single-hop networks never outperform multi-hop networks.
- Detailed modeling is required to accurately determine α

Evaluation on different architectures

- Compare single-hop (BSD,BSS) and multi-hop (PMD) networks
 - Three different architectures
 - Dynamically scheduled superscalar processors
 - Statically scheduled VLIW processors
 - Statically scheduled, dynamically issued Grid Processors
 - Different network configurations
- Benchmarks:
 - gzip, mcf, parser, ammp, art, equake, dct, adpcm, mpeg2encode, radar
- Machine configuration:
 - Alpha 21264 like functional unit latencies, 3-cycle, 64-KB L1-cache, 12-cycle, 2MB L2-cache, 2-level branch predictor
 - 10 F04 clock cycle, 100nm technology
 - Wiring overhead = 2 for single-hop networks
- Performance metrics
 - Routing latency, contention, # of hops, and IPC

VLIW architectures

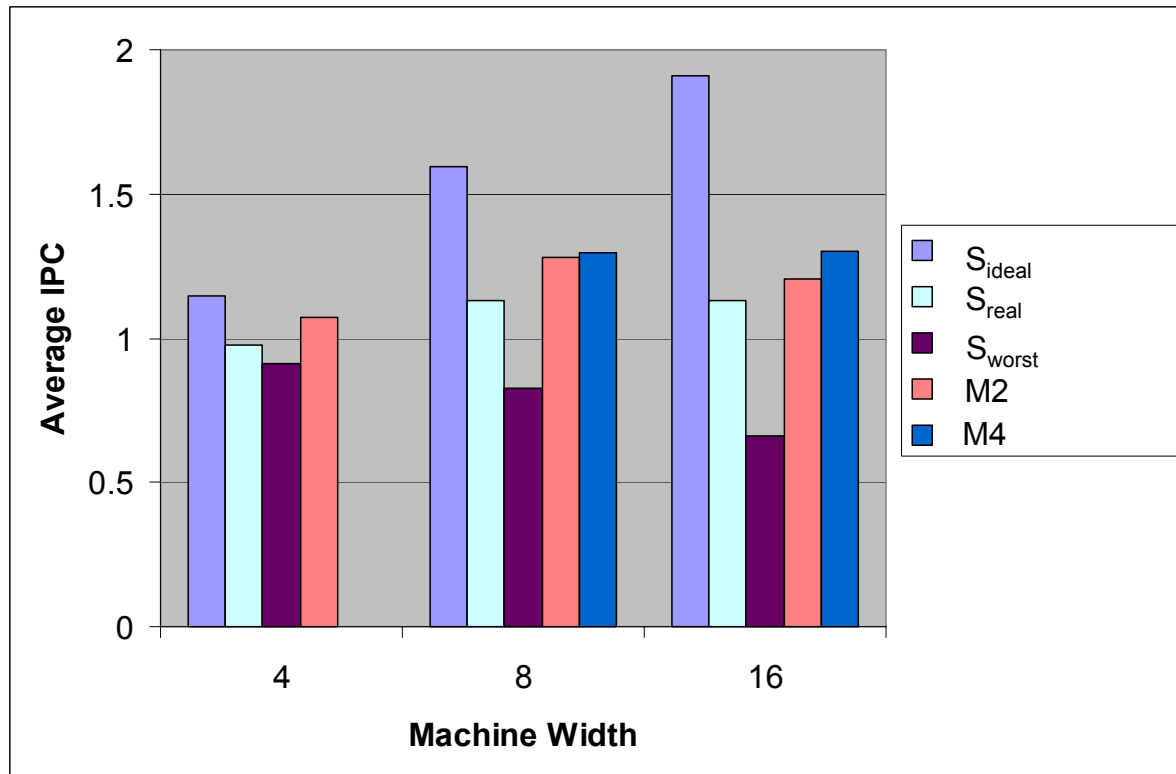


- 4-, 8-, and 16-wide machines simulated
- Wire, fan-out, and fan-in delays used from circuit analysis

VLIW architectures: Results

- Routing latency
 - 1 to 2.4 cycles on Multi-hop networks
 - Multi-hop networks almost always better S_{real} ; always better than S_{worst}
- Contention
 - Between 12% and 26% of latency is due to contention in M2 and M4 networks
- Number of hops:
 - Between 1.2 and 2.1
 - Scheduler is effective at placing consumers close to producers

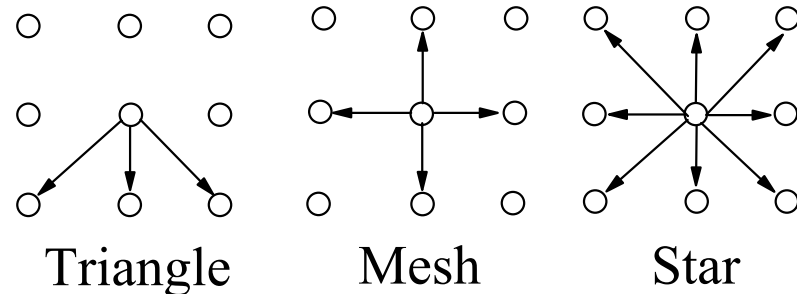
VLIW architectures: Results



- M4 is within 60% of S_{ideal}
- M2 and M4 always better than S_{real} , S_{worst}

Grid Processor Architectures

- Statically scheduled, dynamically issued
- Array of ALUs with a RIAN
- Fast clock rate and high ILP
- 8x8 array
- Delays used from circuit analysis
- Compare with s_{ideal} , S_{real} , S_{worst}



Grid Processor Architectures: Results

Network	Latency (cycles)	Contention (%)	# of hops	IPC
S_{ideal}	0.25	0	1	7.8
S_{real}	4.2	0	1	3.6
S_{worst}	15.58	0	1	1.6
Star	3.26	13	2.5	4.2
Mesh	5.11	27	3.2	3.7
Triangle	7.8	43	2.8	3.7

Conclusions

- Traditional operand transmission networks
 - Large wiring overhead
 - Delay increases as architectures and technology scales
- RIAs outperform traditional broadcast
 - Scale to 10s of ALUs, using a low bandwidth interconnect
 - Always faster than broadcast network circuits
 - Better performance: IPC and routing latency
- Projections for future RIAs
 - Dynamically scheduled architectures require rescheduling techniques
 - Statically scheduled architectures require destinations encoded in ISA

Questions