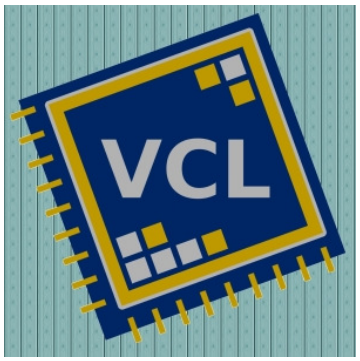
DLABS: a Dual-Lane Buffer-Sharing Router Architecture for Networks on Chip

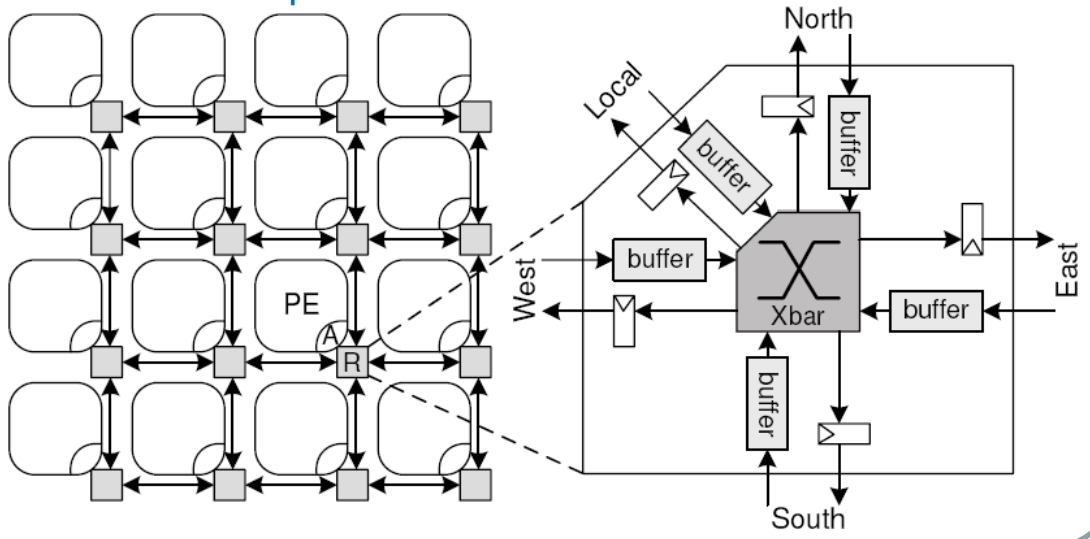


Anh T. Tran, Bevan M. Baas

VLSI Computation Lab
University of California, Davis

Observation & Motivation (1)

a conventional input-buffered wormhole router architecture



✓ More than 60% area and 30% power of the router are spent on its buffers

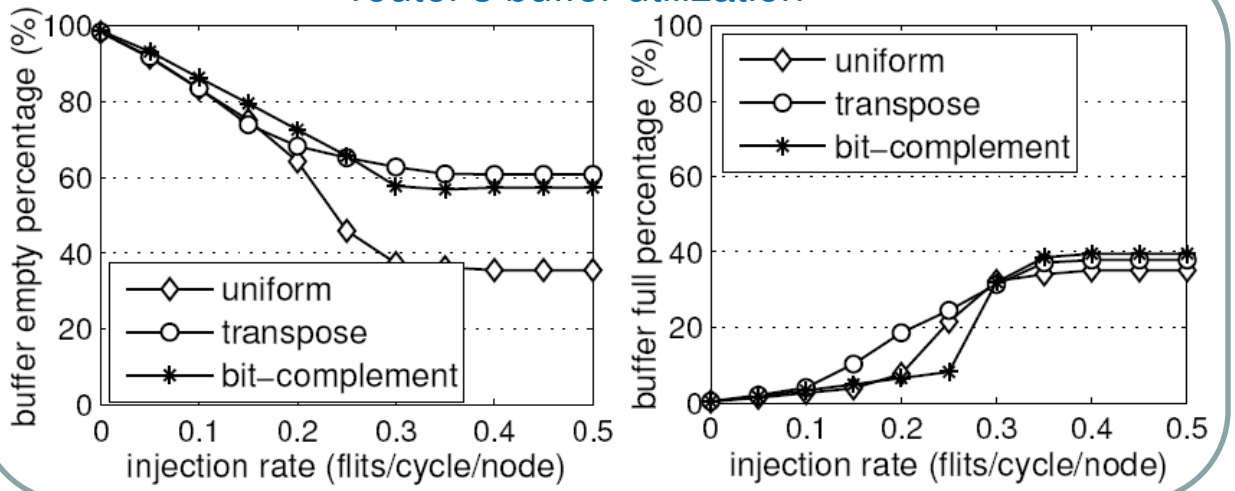
✓ But, a significant amount of these buffers are always empty while running some tested traffic patterns → not efficient

The number of buffers which are always empty during 30,000 simulation cycles

Traffic	uniform	transpose	bit-complement
always empty buffers	32	152	144
ratio	10.0%	47.5%	45.0%

Observation & Motivation (2)

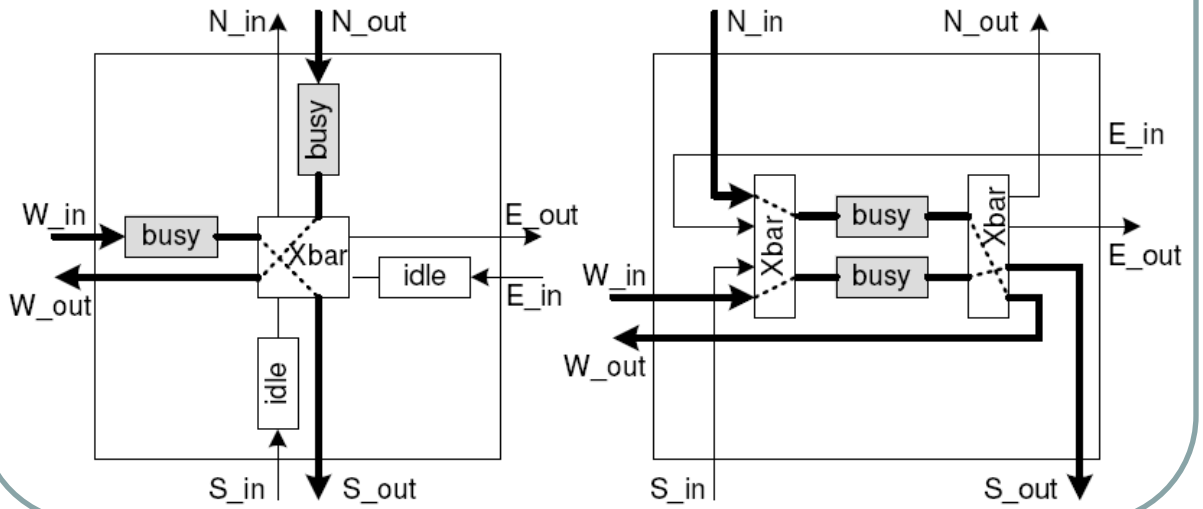
router's buffer utilization



✓ the NoC using typical routers has high buffer empty and low buffer full percentages even at high packet injection rates

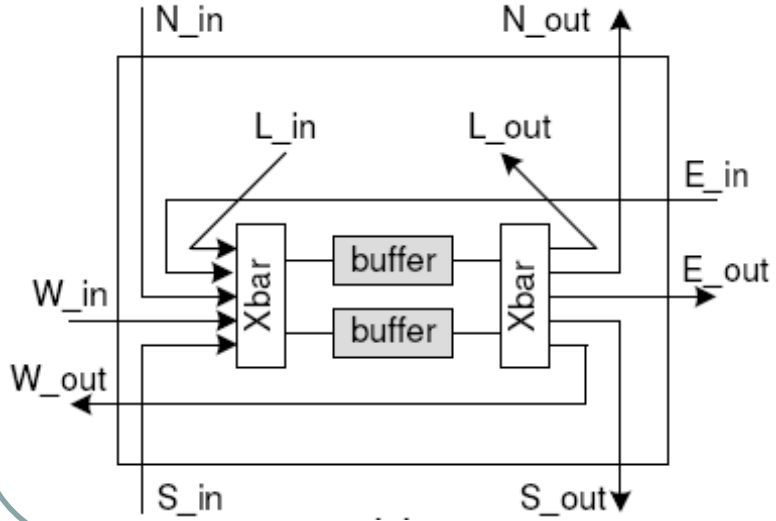
✓ For regular traffic patterns, a router design with less than number of buffers may has an equivalent performance as a typical router

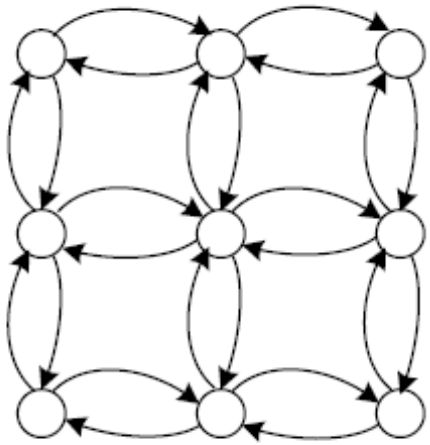
an illustrated example: router's buffer activity corresponding to a regular traffic pattern



Deadlock Potential in a NoC based on Shared-Buffer Routers

a shared-buffer router architecture with 2 buffers;
and the corresponding RDG for a 3x3-array network



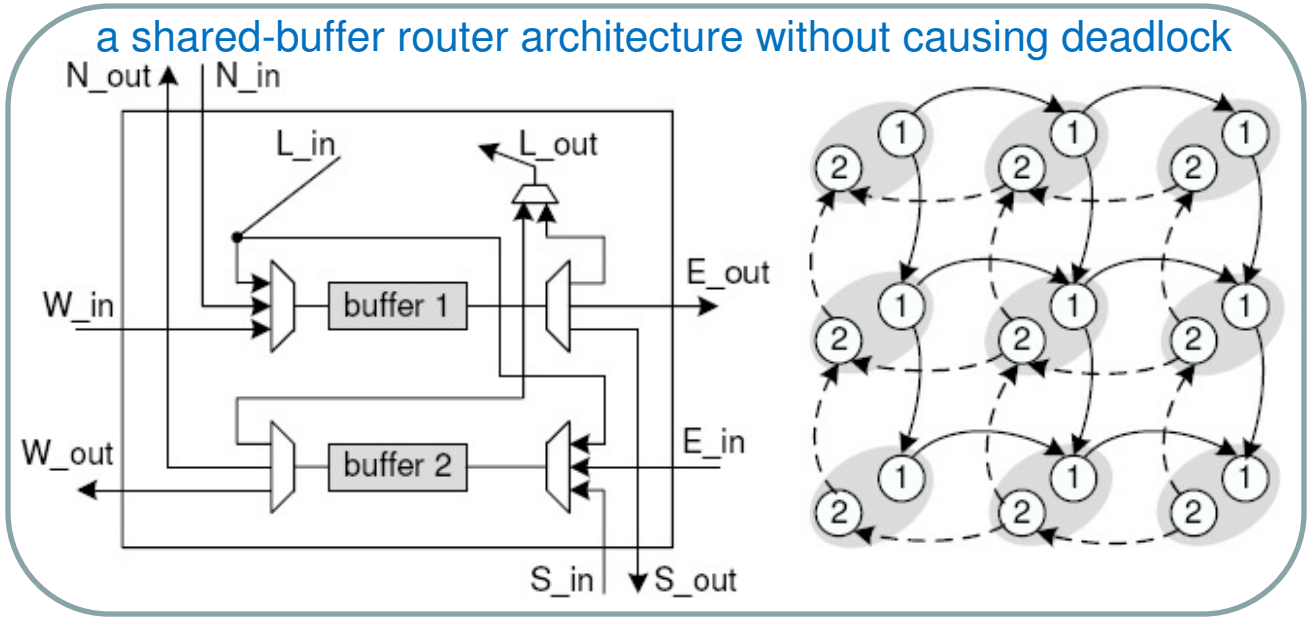


✓ Can build a router with all input ports sharing a group of buffers instead of a buffer per input port

✓ But, it causes deadlock potential in the network (by creating loops in the corresponding resource dependence graph RDG)

[Dally, TC, 1987; Duato, TPDS, 1993]

Breaking Deadlock by a Dual-Lane Router Architecture



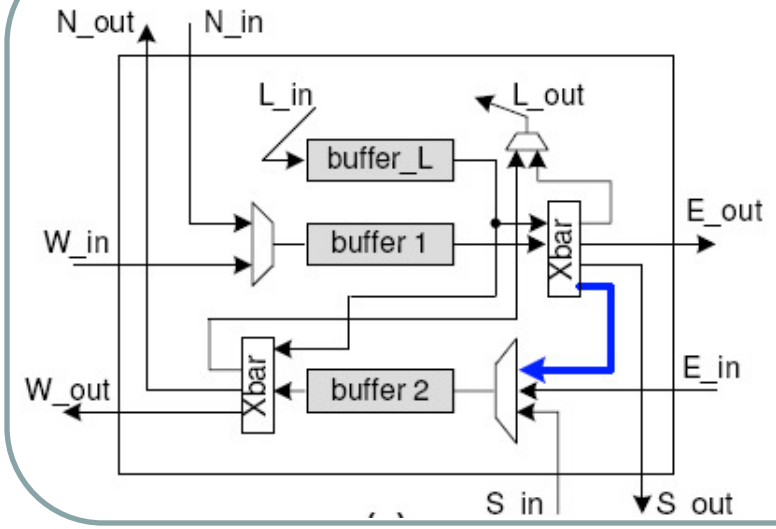
✓Create two separate lanes without loop in the RDG by partitioning buffers into two group:

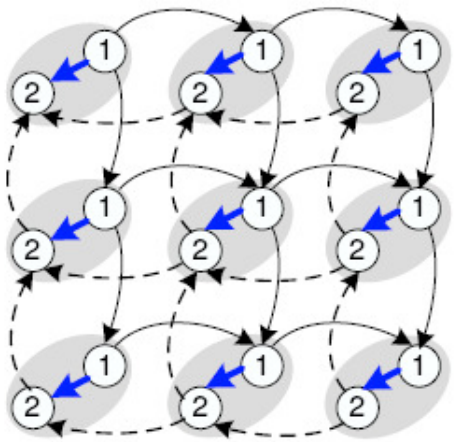
- One group is shared by W_in and N_in ports, and outputs to E_out and S_out ports
- Another group is shared by E_in and S_in ports, and outputs to W_out and N_out ports

✓ But, the network does not cover all destination-source patterns

DLABS_1+1

DLABS_1+1 router architecture – version 1



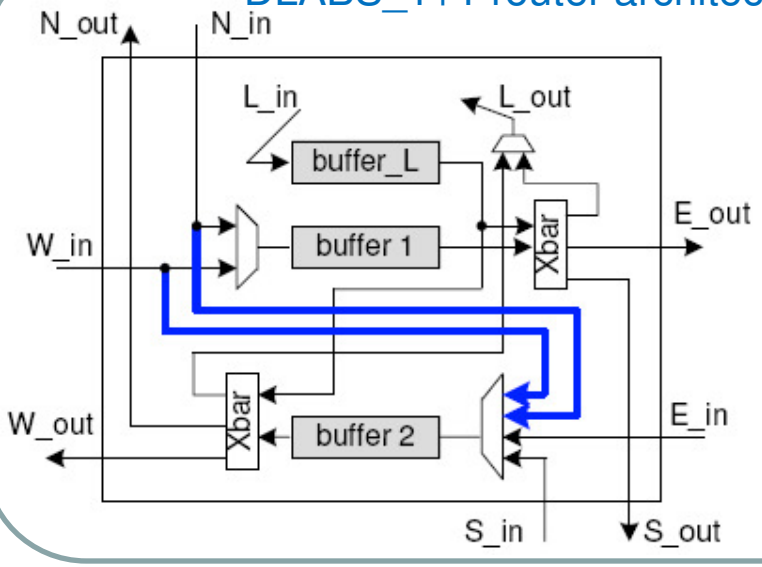


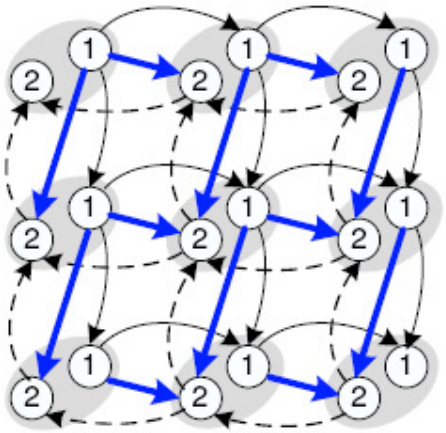
✓ Allow a packet to be sent from lane 1 to lane 2, but not in the reversed way

✓ Has poor performance due to potential of buffering a packet two times in a router

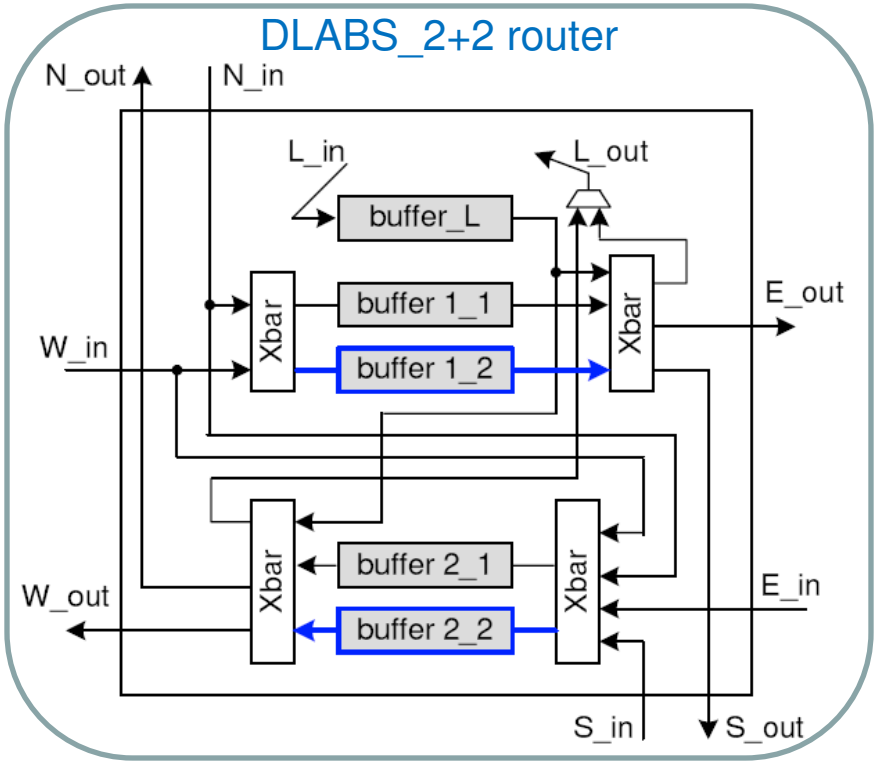
✓ A faster design: allows a packet from buffer 1 of a router to be sent directly to buffer 2 of the next downstream router

DLABS_1+1 router architecture – version 2



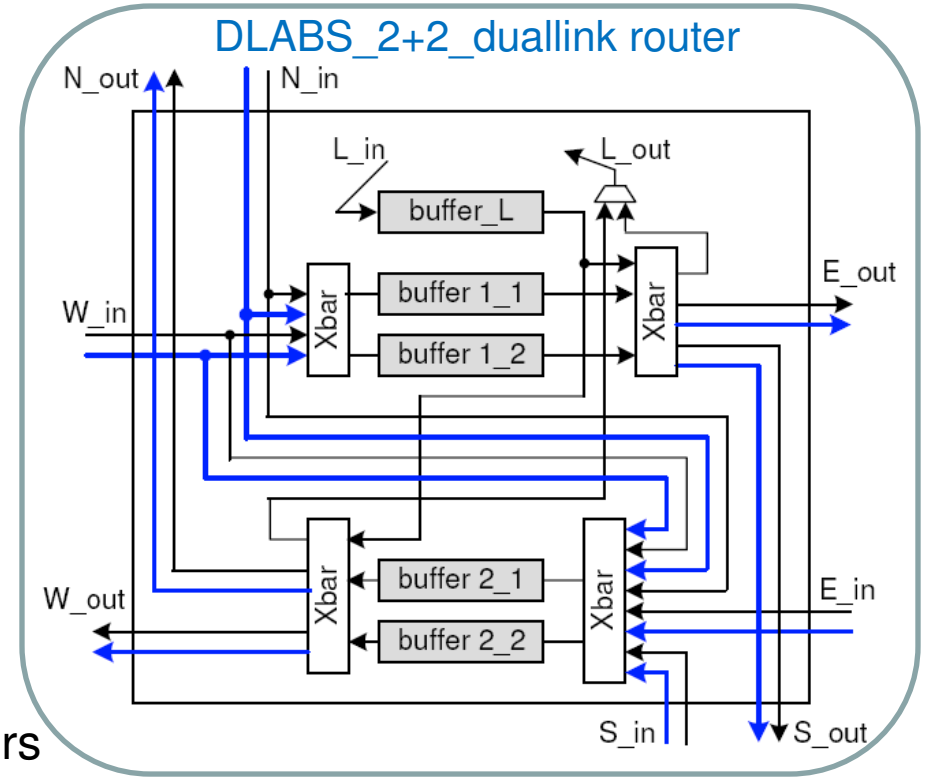


Enhanced DLABS Router Architectures



✓ DLABS_2+2: two buffers per lane to have the same number of buffers as in a typical router

✓ Output physical links may become performance bottlenecks



✓ DLABS_2+2_duallink: two physical links per input/output ports

✓ Assumed on-chip interconnect wires are cheap:

- multi-layer metal wires
- routed only between nearest neighbors

Experimental Setup

Five router architectures in evaluation

Architecture	Typical wormhole	DLABS_1+1	DLABS_2+2	DLABS_2+2 _duallink
Total buffers	5	3	5	5
Buffer depth (flits)	8	8	8	8
I/O links per port	2	2	2	4

- ✓ Implement of all four router architectures in Verilog RTL
- ✓ Cycle-accurate simulation with Cadence NC Verilog
- ✓ Evaluate and compare their performance over three synthetic traffic patterns: uniform random, transpose, and bit-complement
- ✓ The simulated network consists of 8x8 nodes; each node = PE + router
- ✓ Run each simulation for 30,000 cycles
- ✓ Each packet is 10 FLITs in length
- ✓ Activity of each router is recorded cycle-by-cycle for evaluation

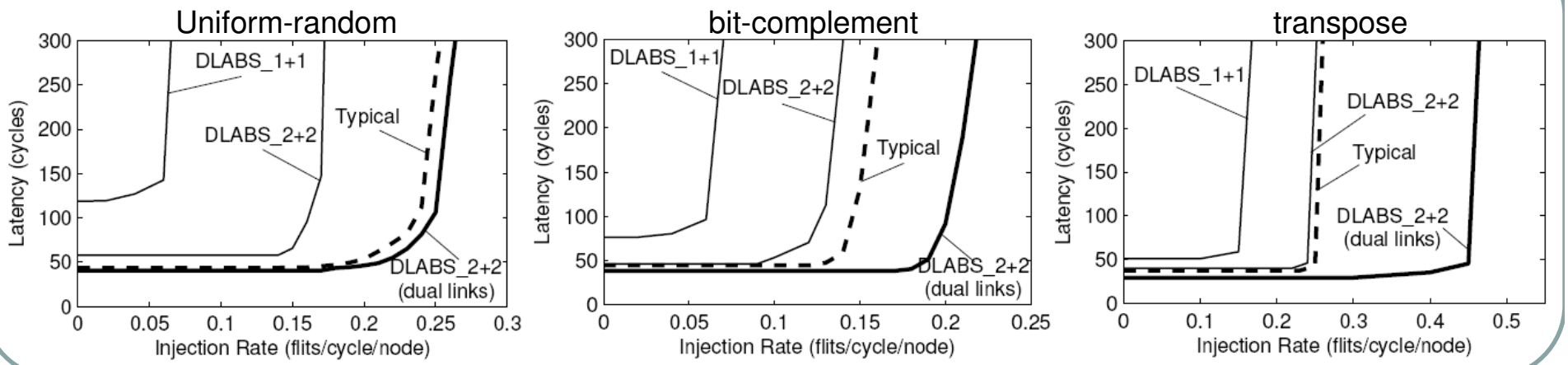
Results (1)

Percentage of number of buffers which are always idle during the whole simulation time

Architect.	Typical wormhole	DLABS_1+1	DLABS_2+2	DLABS_2+2 _duallink
random	10.0%	1.0%	0.9%	0.9%
transpose	47.5%	16.2%	16.9%	16.9%
bit-comp.	45.0%	8.3%	9.8%	9.8%

✓ All DLABS routers utilize well their buffers

Average network latency

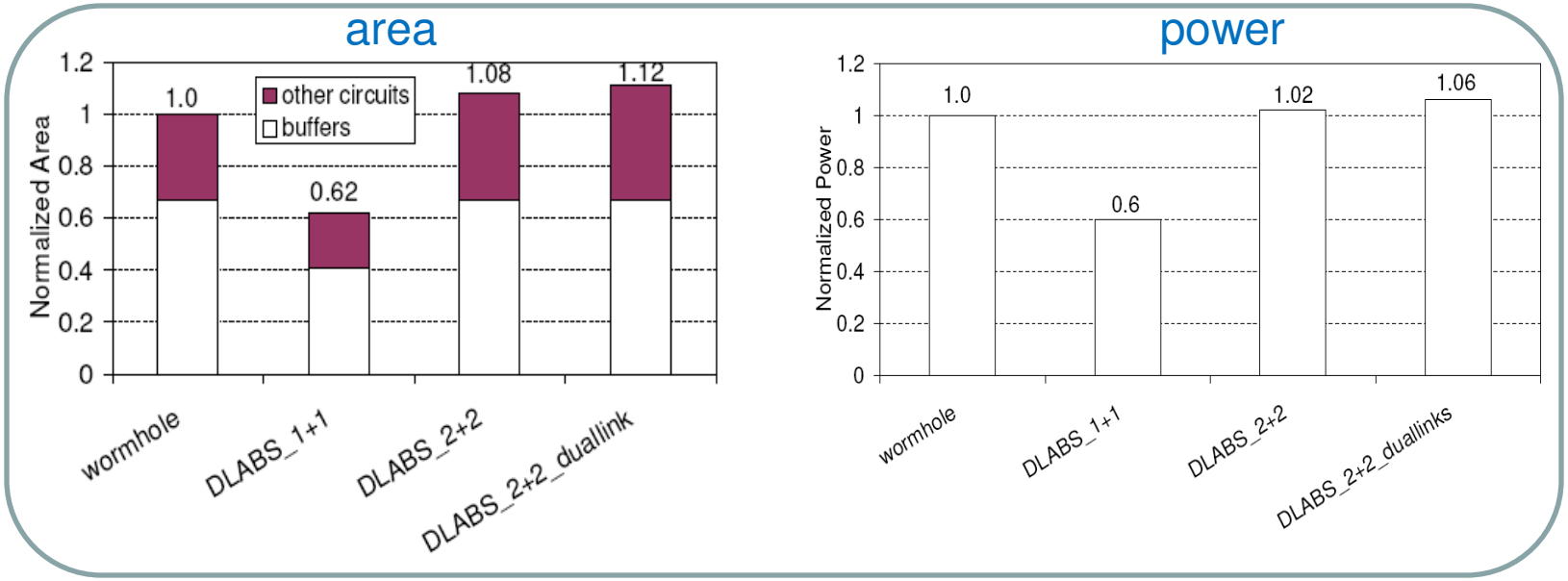


✓ DLABS_1+1: poorest performance due to congestion on shared buffers

✓ DLABS_2+2: better, but still poor due to congestion on interconnect links

✓ DLABS_2+2_duallink: best performance; especially in regular traffic patterns (not random)

Results (2)



✓ Routers are synthesized targeting 65-nm ST Microelectronics standard cells using Synopsis DC Compiler

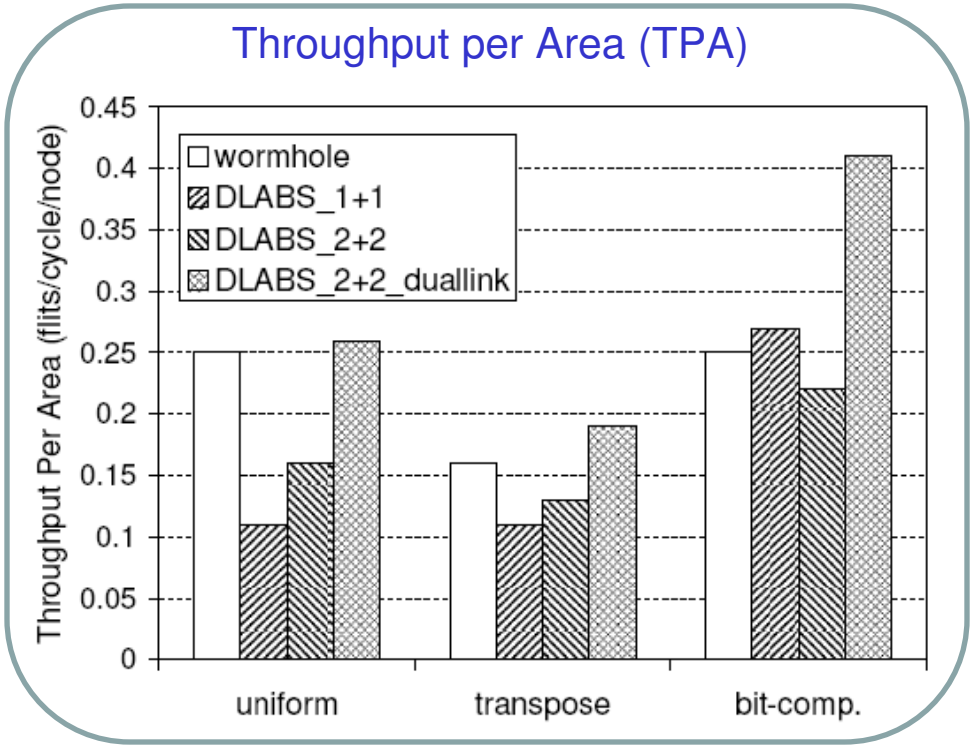
✓ In typical router, 66% area is spent on its buffers

✓ Sharing buffers makes DLABS_1+1 router's area only 62% of the typical router

✓ Having the same buffer area as the typical router, but DLABS_2+2 and DLABS_2+2_duallink are 8% and 12% bigger, respectively due to additional control logic circuits

✓ A similar result is observed in the synthesis power

Results (3)



✓ TPA shows the silicon area using efficiency of a router design
✓ Throughput is measured when the network has an average latency of 200 cycles (near saturated)
✓ DLABS_2+2_duallink has greatest TPA over all traffic patterns
✓ Especially, for the bit-complement pattern, the DLABS_1+1 and DLABS_2+2-duallink are 8% and 64% greater than the typical router

Conclusion

✓ Achieve higher area efficiency by sharing buffers of a router for multiple ports
✓ Resolve deadlock problem by a dual-lane architecture, named DLABS
✓ DLABS_1+1 has 62% area compared to a typical wormhole router, but has low performance
✓ DLBAS_2+2_duallink has 12% area greater than a typical one, but achieves much higher performance and throughput per area, especially in regular traffic patterns

Future Work

✓ Evaluate DLABS routers over other traffic patterns
✓ Exploit other shared-buffer router architectures
✓ Compare with virtual-channel routers and other architectures
✓ Consider to use bidirectional interconnect links

Acknowledgements

✓ NSF Grant 430090, 903549; CAREER award 546907
✓ SRC GRC Grant 1598, 1971; CSR Grant 1659
✓ a VEF Fellowship
✓ ST Microelectronics
✓ Intel
✓ Intellasys