Saturating the Transceiver Bandwidth: Switch Fabric Design on FPGAs

FPGA’2012

Zefu Dai, Jianwen Zhu

University of Toronto

Switch is Important

Cisco Catalyst 6500 Series Switches since 1999

Total $: $42 billion
Total Systems: 700,000
Total Ports: 110 million
Total Customers: 25,000

Average price/system: $60K
Average price/port: $381

Commodity 48 1GigE port switch: $52/port

http://www.cisco.com/en/US/prod/collateral/modules/ps2797/ps11878/at_a_glance_c45-652087.pdf

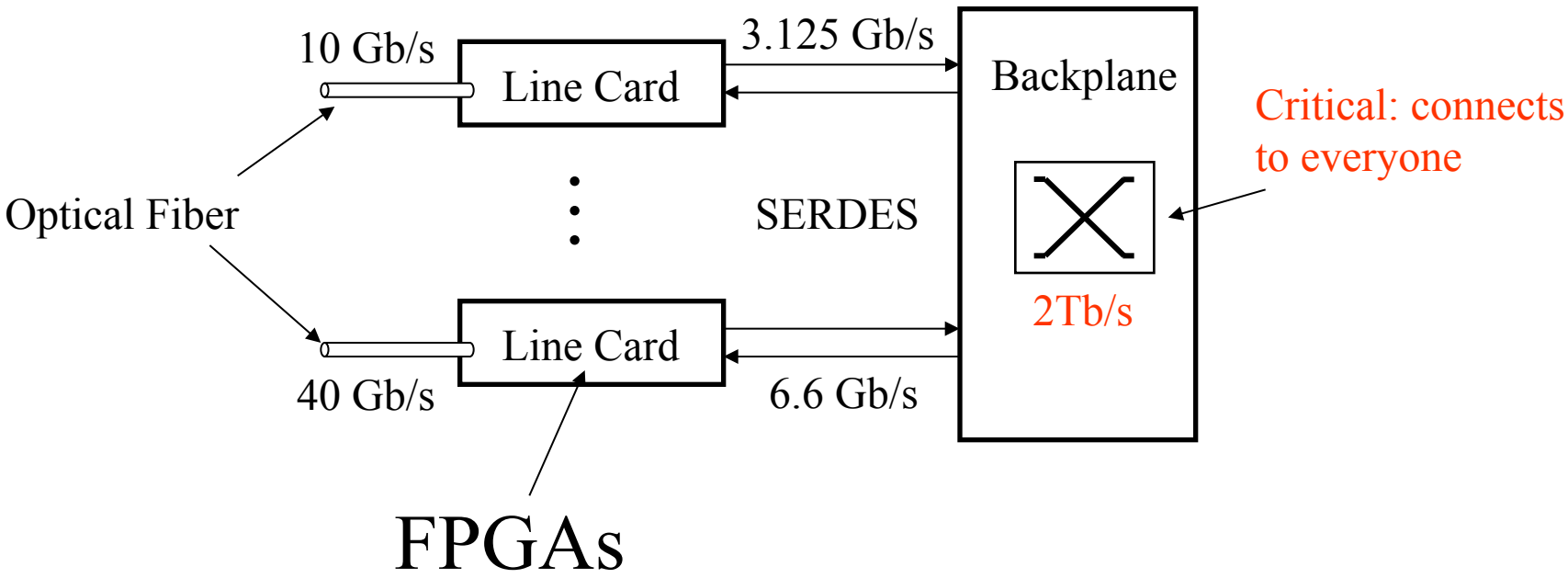
http://www.albawaba.com/cisco-readies-catalyst-6500-tackle-next-decades-networking-challenges-383275

Inside the Box

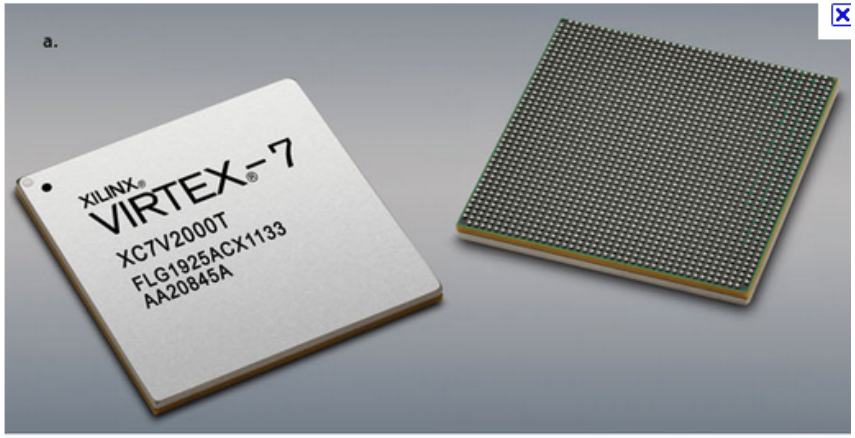
Function blocks:

Virtualization	Admin. Ctrl	Firewall	L2 Bridge	Routing	QoS

Configuration Security Traffic manage

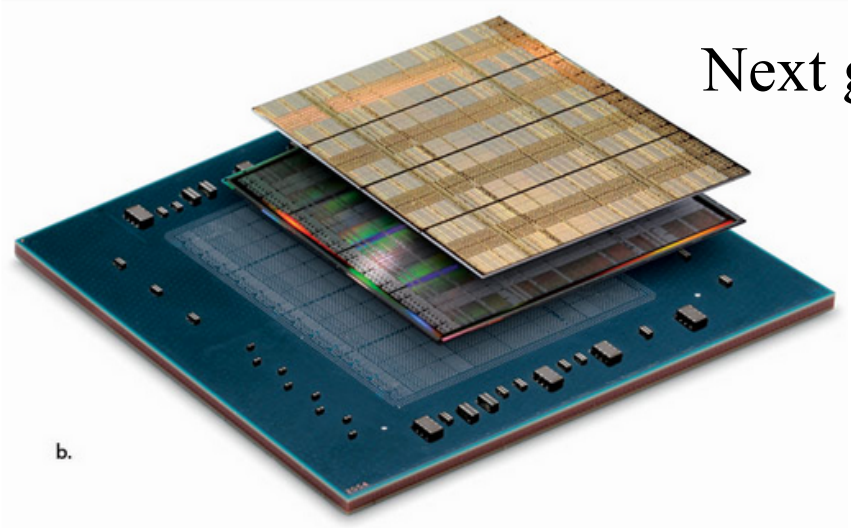


Go Beyond the Line Card

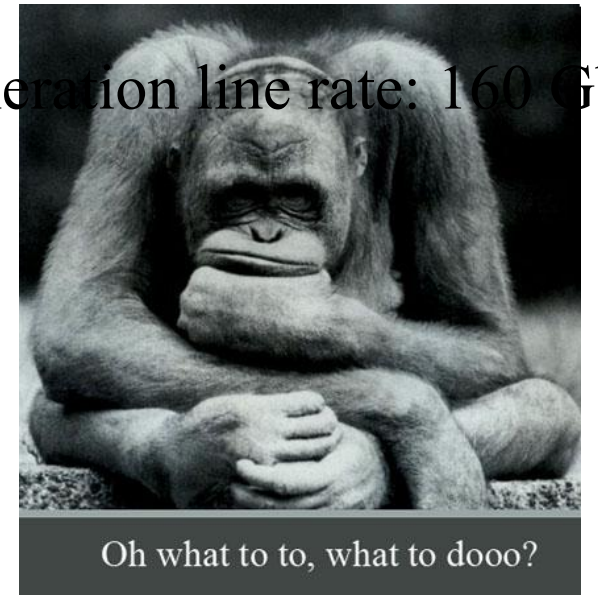


XC7VH870T (single chip):

- 72 GTX: 13.1 Gb/s
- 16 GTZ: 28.05 Gb/s
- Raw total: 1.4 Tb/s



Next generation line rate: 160 Gb/s



Can We Do Switch Fabric?

- ✧ Is single chip switch fabric possible on FPGA that saturate transceiver BW?
- ✧ Is it possible to rival ASIC performance?

Prof. Jonathan Rose, FPGA’06:

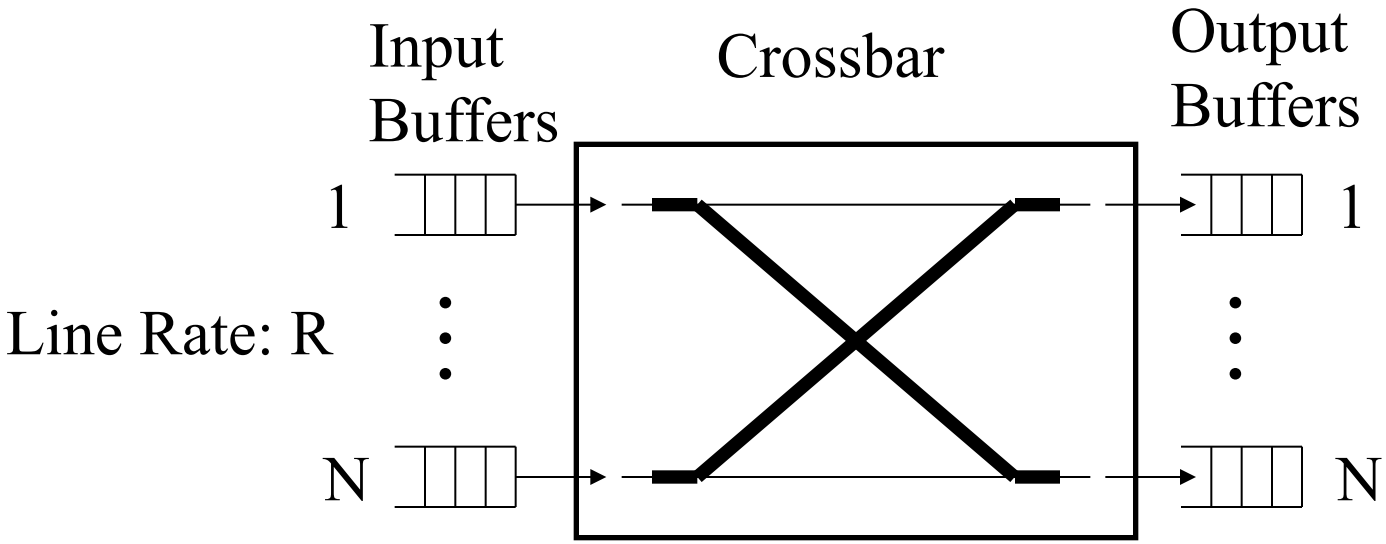
Gap	Area	Speed	Power
FPGA VS. ASIC	40	3	12

What’s Switch Fabric

✧Two major tasks:

- Provide data path connection (crossbar)
- Resolve congestion (buffers)

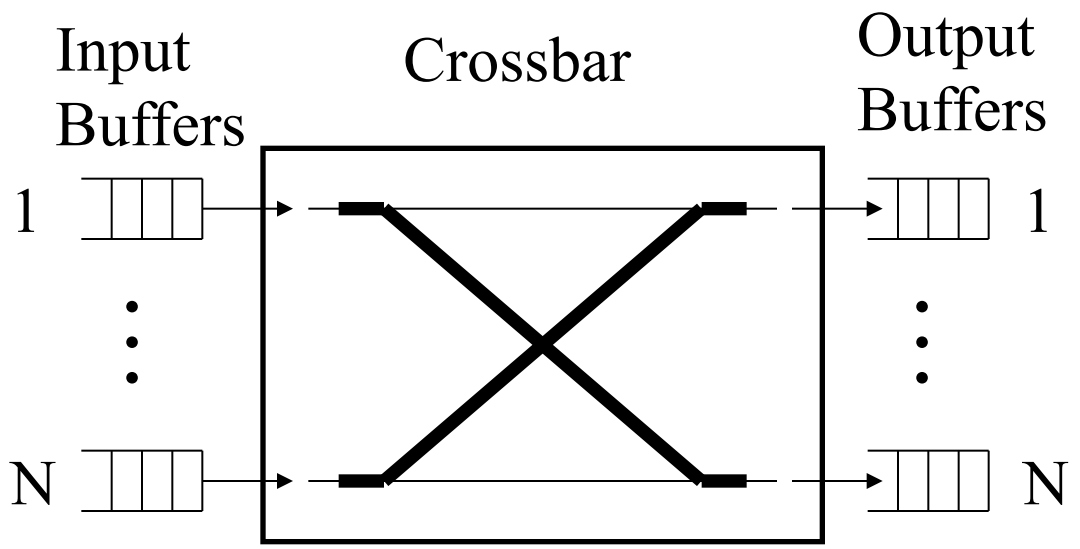
little computational task



Switch Fabric Architectures

✧Buffer location marks the difference:

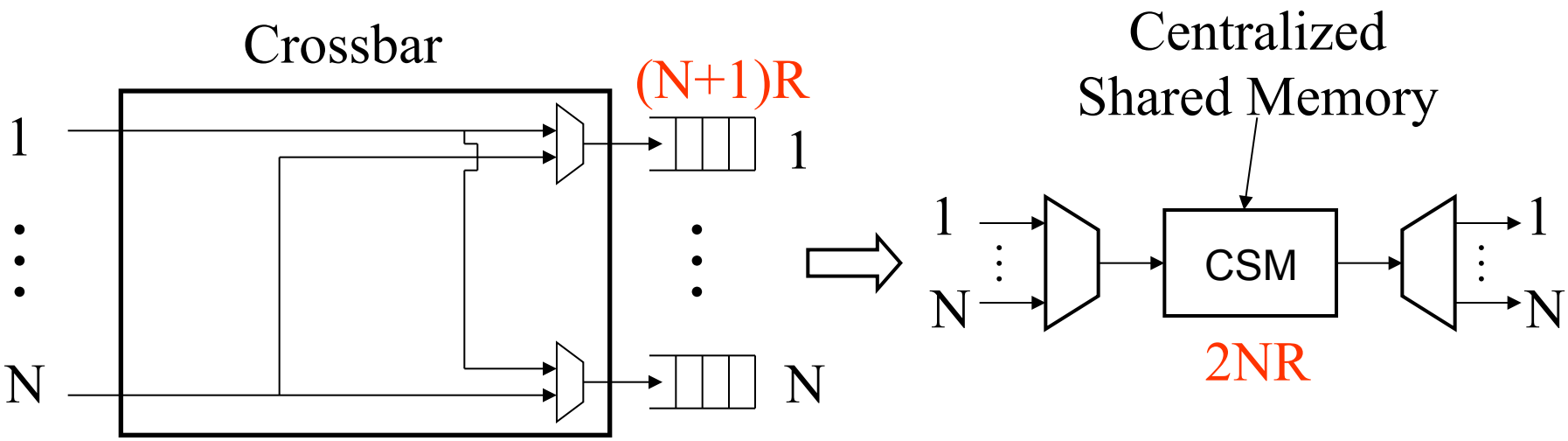
- Output Queued (OQ) switch: buffered at output
- Input Queued (IQ) switch: buffered at input
- Crosspoint Queued (CQ) switch: buffered at XB



Ideal Switch

✧OQ switch

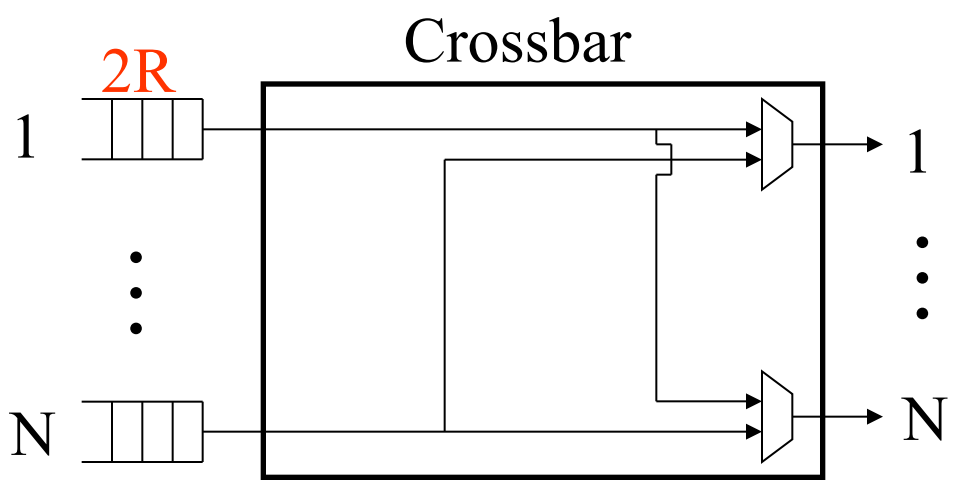
- Best performance
- Simplest crossbar (distributed multiplexers)
- 2NR memory bandwidth requirement



Lowest Memory Requirement Switch

✧IQ switch

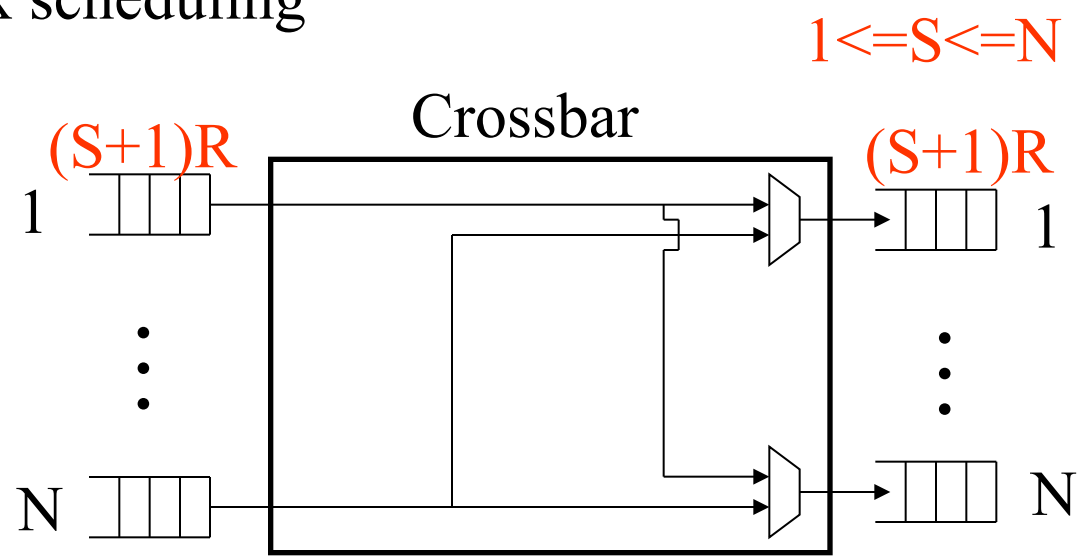
- 2R memory bandwidth
- Low performance, 58% throughput cap (HOL)
- Maximum bipartite matching problem



Most Popular Switch

✧ Combined Input and Output Queued (CIOQ) :

- Internal speedy up *S: read (write) S cells from (to) memory*
- Emulate OQ switch with *S*=2
- (S+1)R memory bandwidth
- Complex scheduling

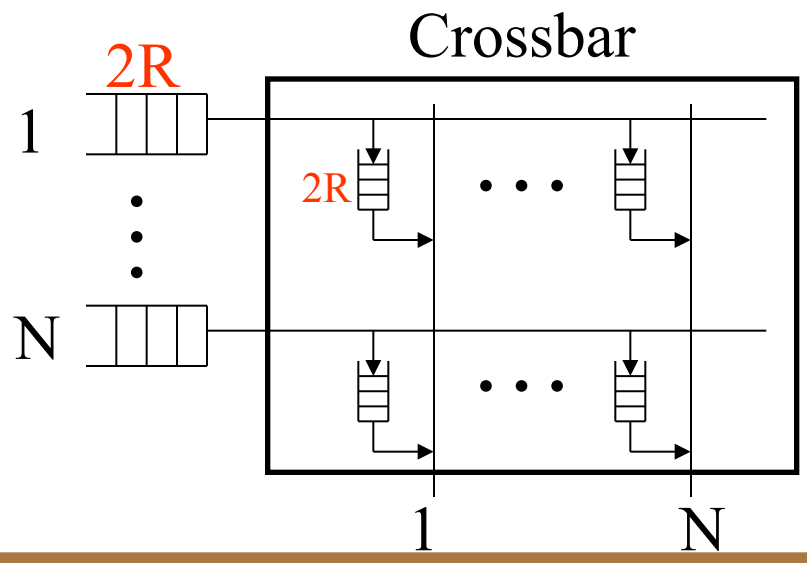


State-of-the-art Switch

- Combined Input and Crosspoint Queued (CICQ)
  - High performance, close to OQ switch
  - Simple scheduling
  - 2R memory bandwidth,
  - N^2 buffers

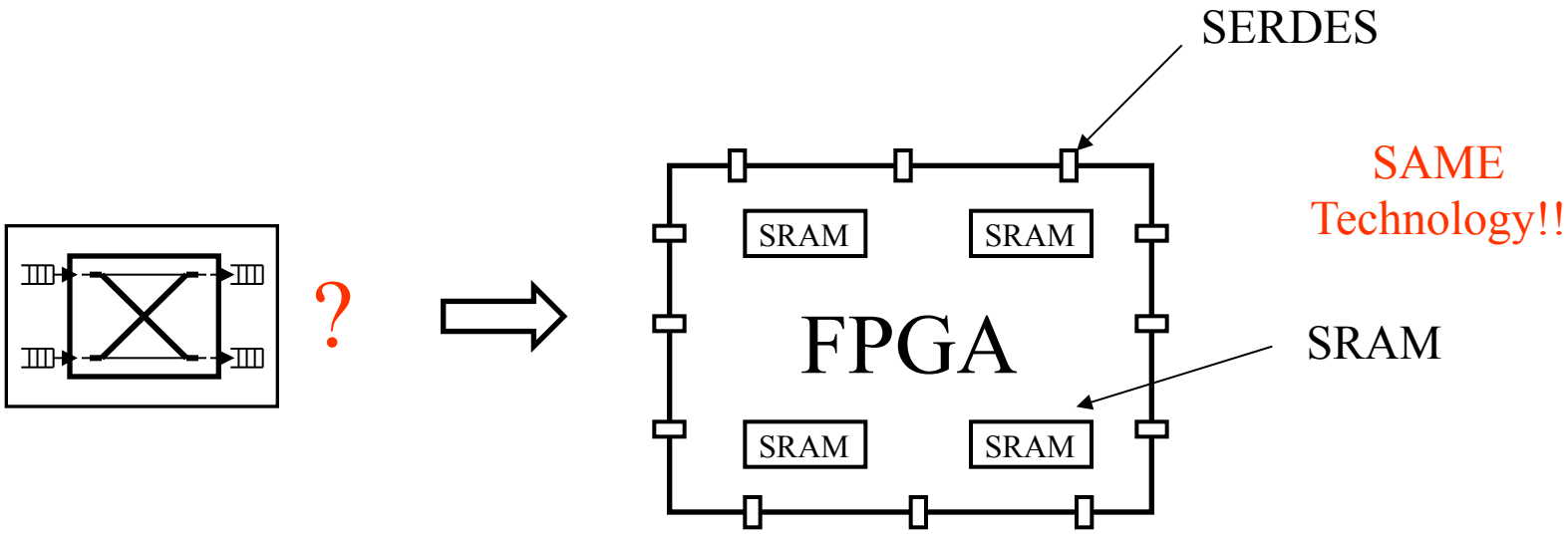
IBM Prizma switch (2004)

2Tb/s



Switch Fabric on FPGA?

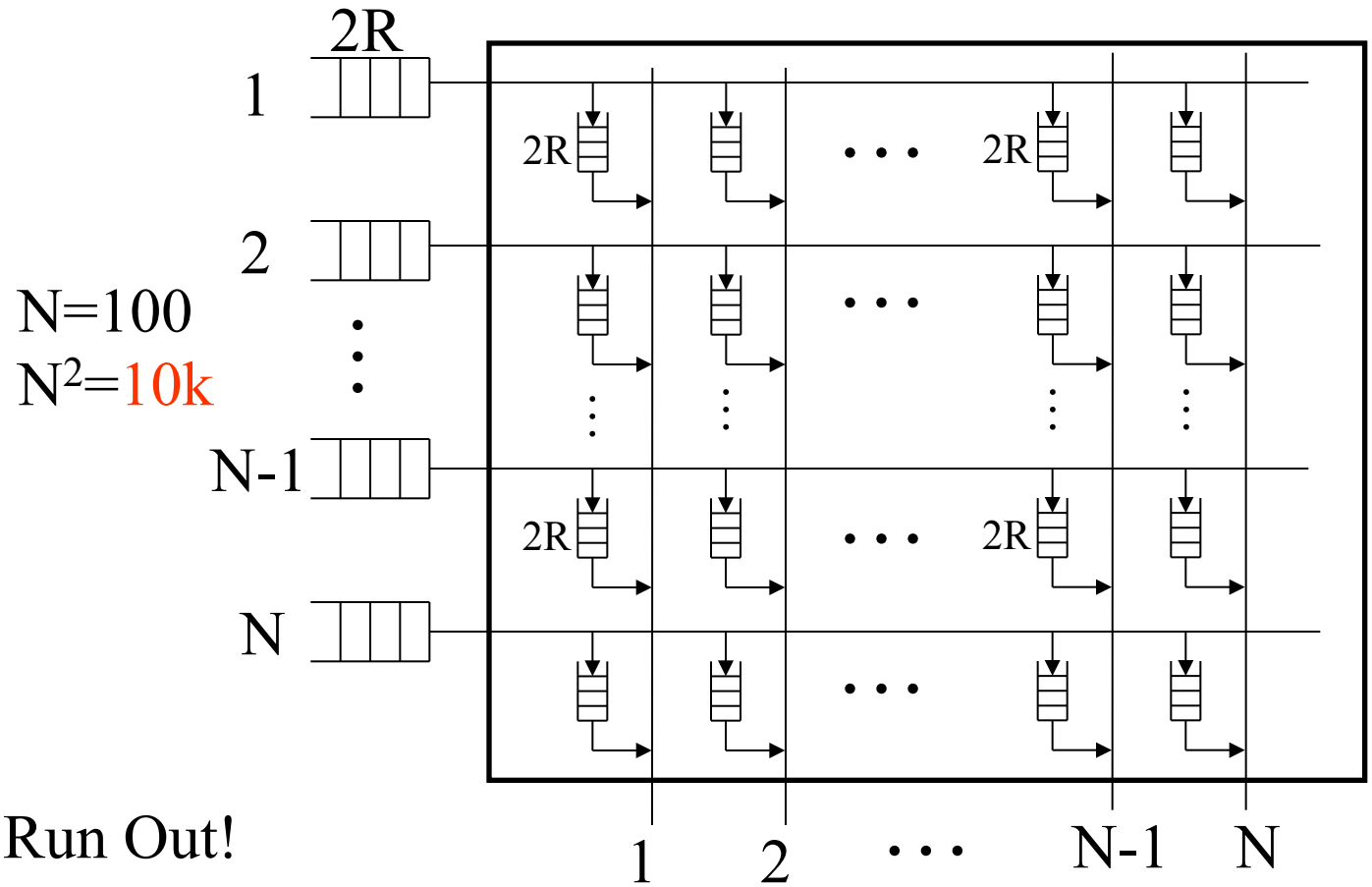
- ✧FPGA resources:
  - LUTs, wires, DSPs, …
- ✧Idea: Memory is the Switch!



Saturate the transceiver BW

Start Point

✧ CICQ switch: rely on **large** amount of **small** buffers



N=100
N^2=10k

R=10 Gb/s
2R = 20 Gb/s

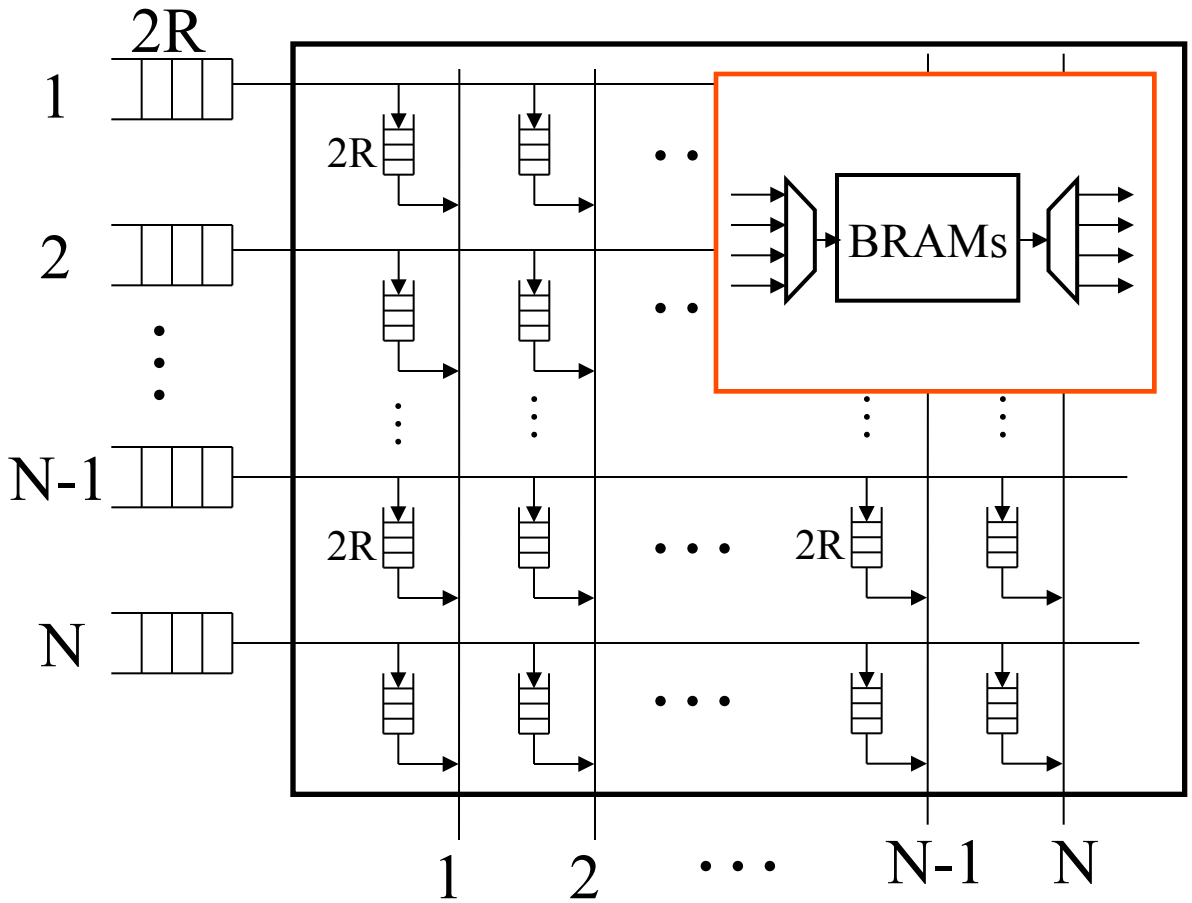
Xilinx 18kbit dual-port BRAM (500MHz): 36Gb/s

Run Out!

Waste of BW!

Memory Can be Shared

✧ Use *x* BRAMs to emulate *y* crosspoint buffers (x<y)



x=5, y=9

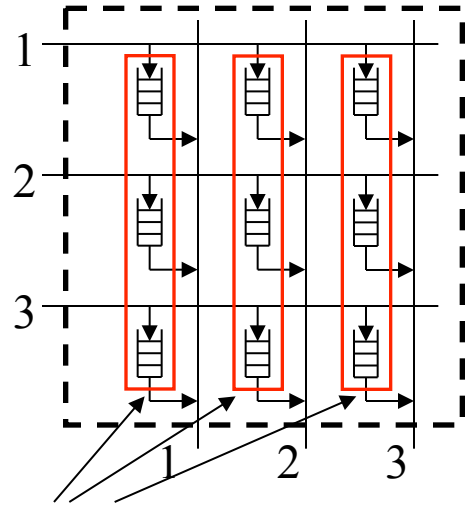
5*36 = 9*20

Total: $5*(N/3)^2$

Xilinx 18kbit dual-port BRAM (500MHz): 36Gb/s

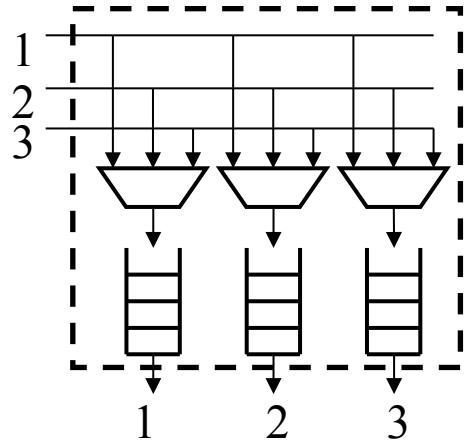
Memory is the Switch

Logical View



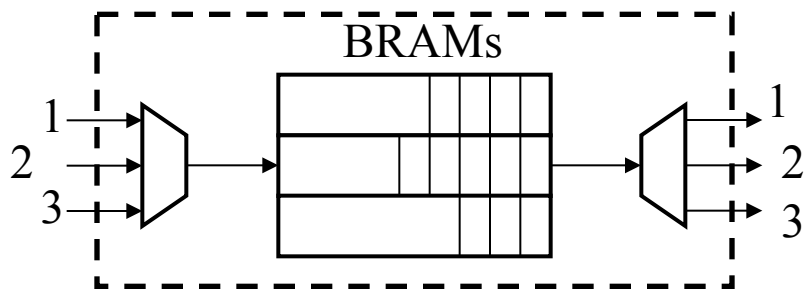
In the same memory

Logical View



Small OQ Switch!

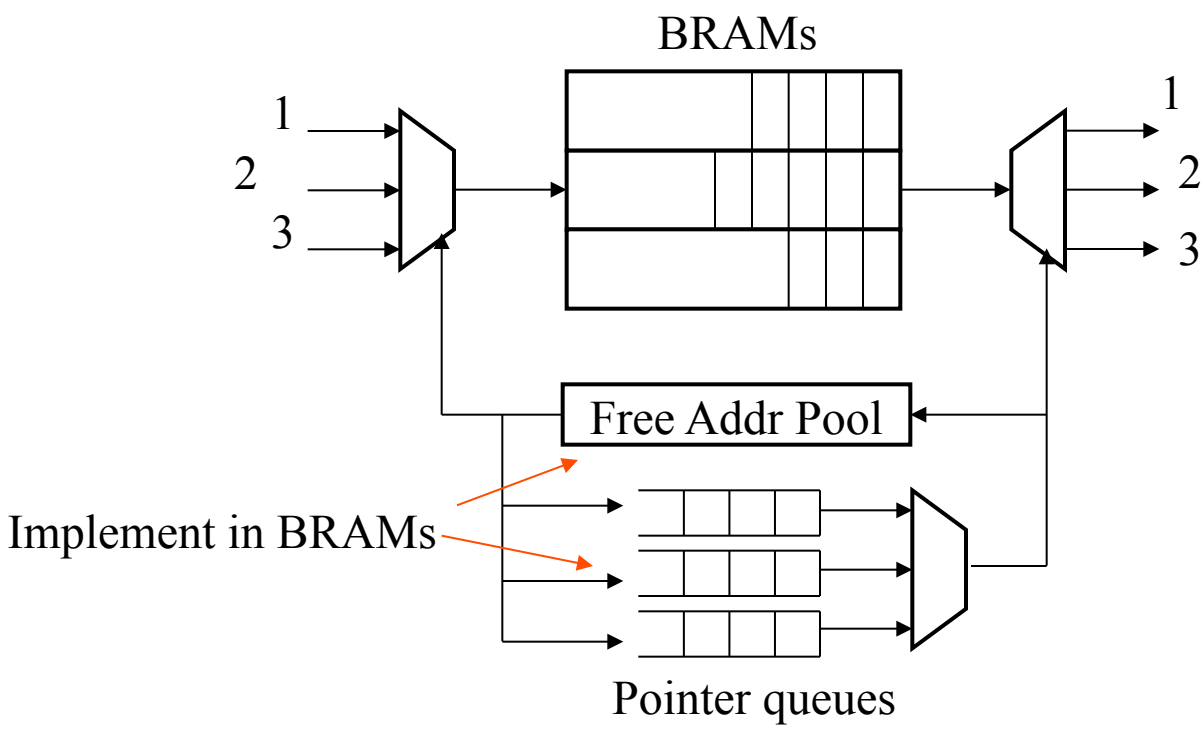
Physical View



address decoder and sense amp act as crossbar

Memory Can be Borrowed

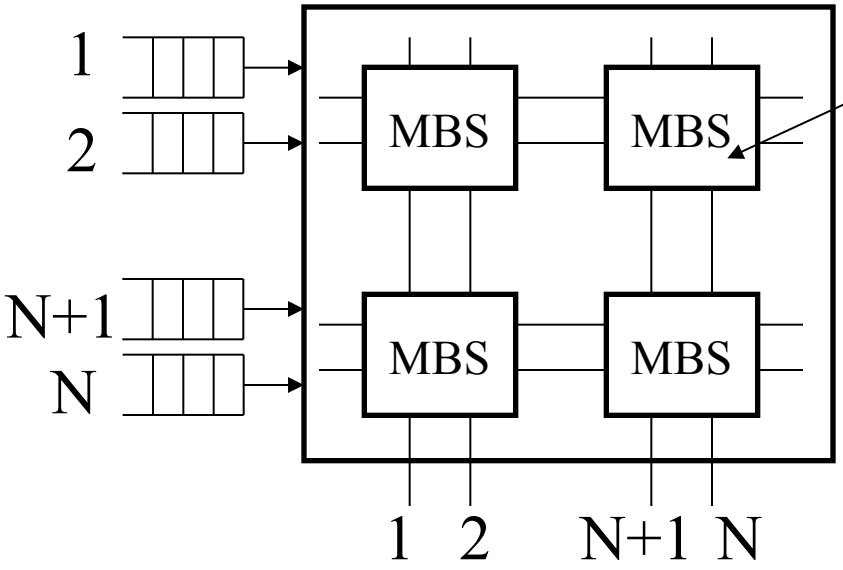
- ✧Give busy buffer more memory space
- ✧Enable efficient multicast support



Group Crosspoint Queued (GCQ) Switch

✧ Similar to Dally's HC switch, but:

- Higher memory requirement for simpler scheduling and better performance
- Use SRAMs as small switch (not just buffer)
- Multicast support



Memory Based Sub-switch

-Efficient utilization of BRAMs

-simple scheduler (time muxing)

Resource Savings

- Each *SxS* sub-switch requires *P* BRAMS:
  - P*B= 2SR; (R: line rate; B: BW of BRAM)
  - P/S = 2R/B; (constant)
- Total BRAMs required for entire crossbar:
  - $P*(N/S)^2 = (2N^2R/B)/S = C/S$ *(C is constant)*
  - Savings: N^2-C/S; (larger S, more saving)
- Max S limited by minimum packet size
  - Aggregate data width of BRAMs <= minimum packet size
  - TCP packet (40 bytes): max S = 8

Hardware Implementation

	Virtex6-240T	Spartan6-150T
N	16	9
S	4	3
Data Width	256 bit	256 bit
Registers	36945 (12%)	27028(14%)
LUTs	49537 (32%)	37285 (40%)
BRAMS	224 (27%)	95 (36%)
Savings	288(56%)	67 (41%)

Saturate the transceiver BW

Small S due to automatic P&R (BRAM frequency as low as 160MHz)

Clock domain crossing (CDC)

Source synchronous CDC technique

Performance Evaluation

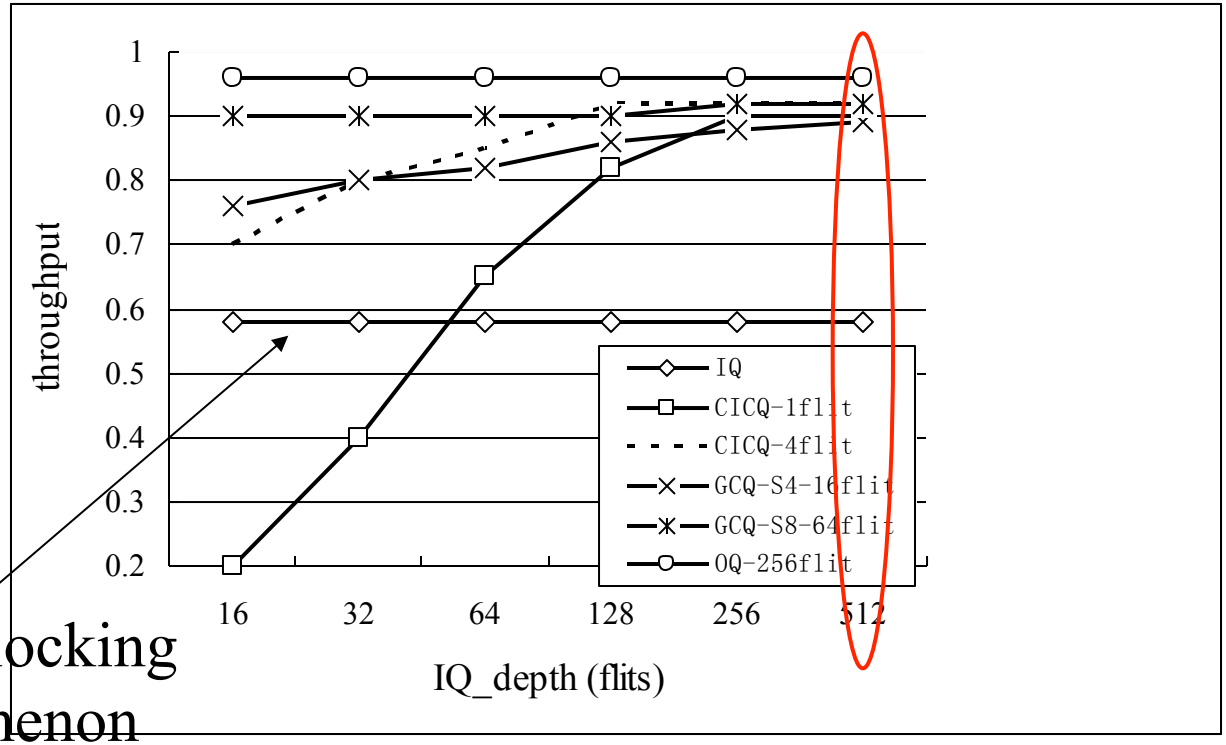
✧ Experimental setup

- Booksim: network simulator by Dally’s group
- Tested switches: IQ, OQ, CICQ and different configuration of GCQ with different S

Flit Delay	2
Credit Delay	2
N	16
Flit Size	32 bytes
Packet size	1 flit, 16 flits
Traffic	Uniform
Network topology	Fat tree
Routing	Nearest common ancestor

Maximum Throughput Test

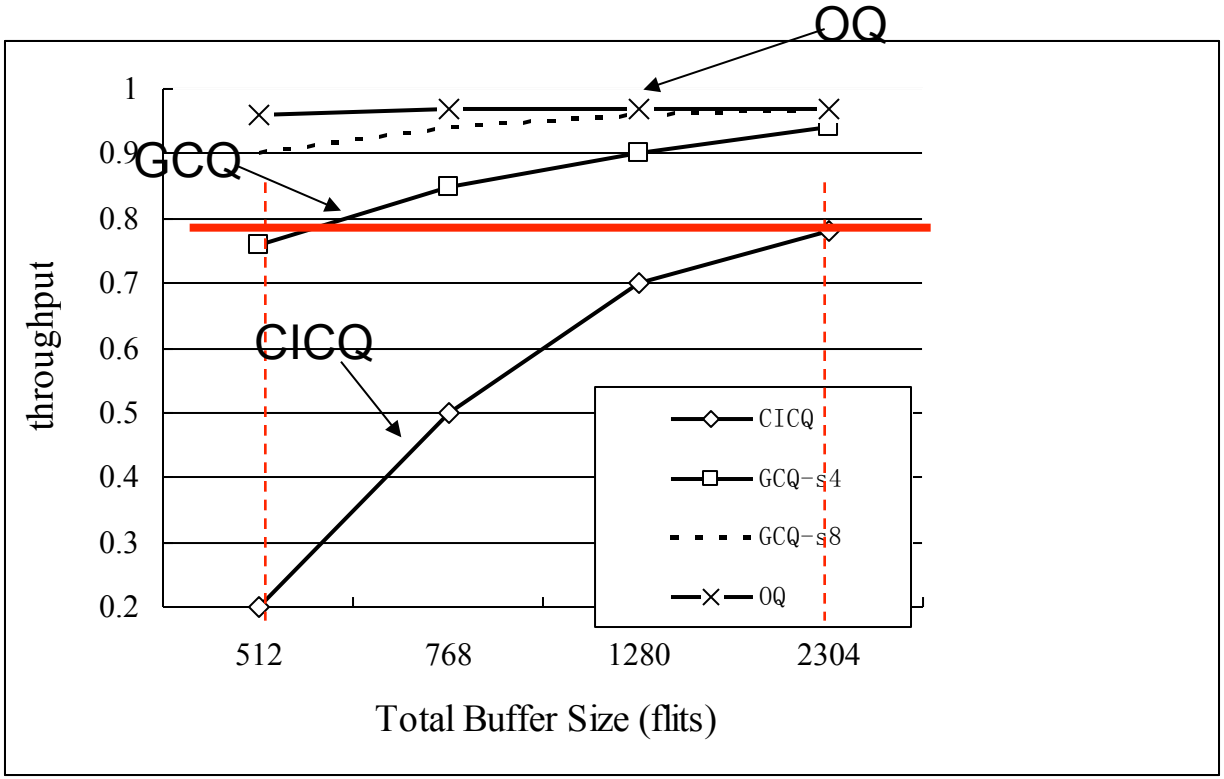
Keep minimum buffer in the crossbar, sweep different Input Buffer depth



HOL blocking phenomenon

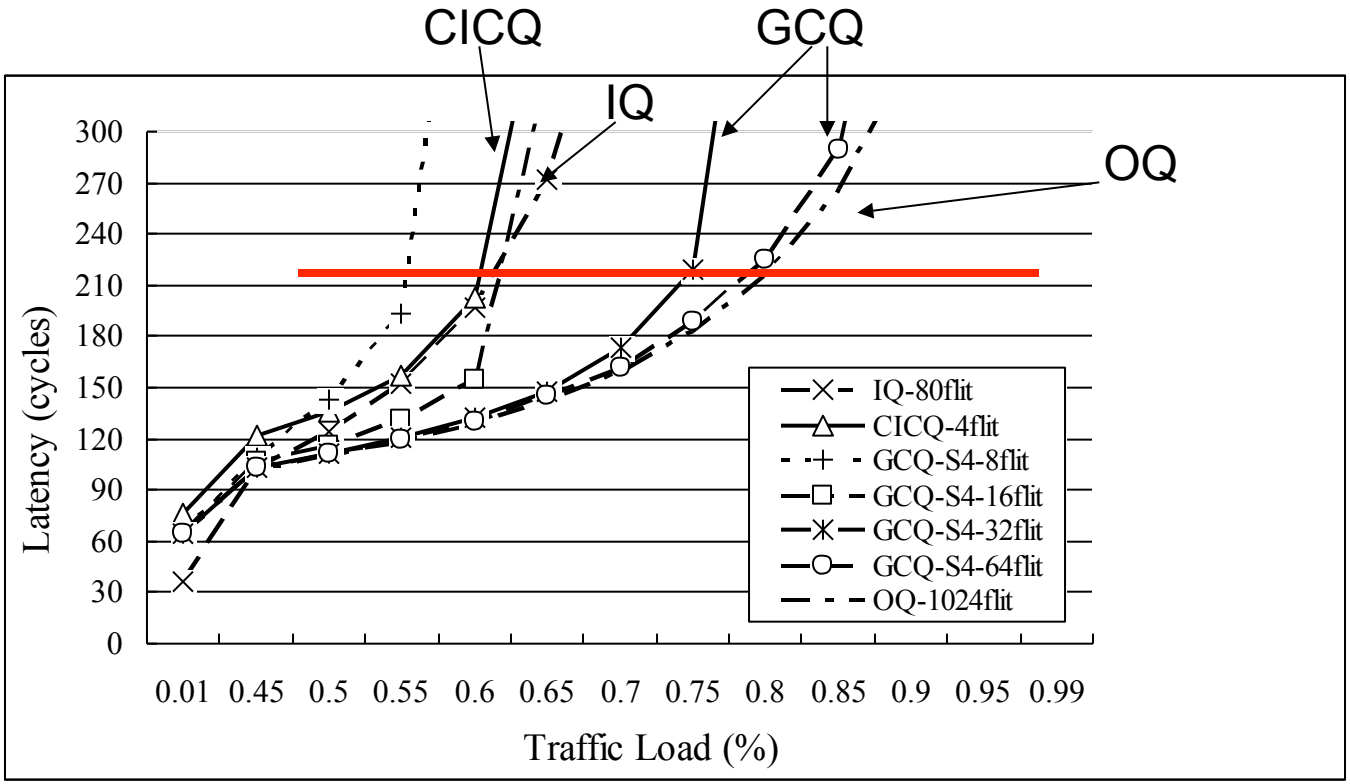
Buffer Memory Efficiency

Keep minimum buffer in the Input Buffer, sweep different crossbar buffer size



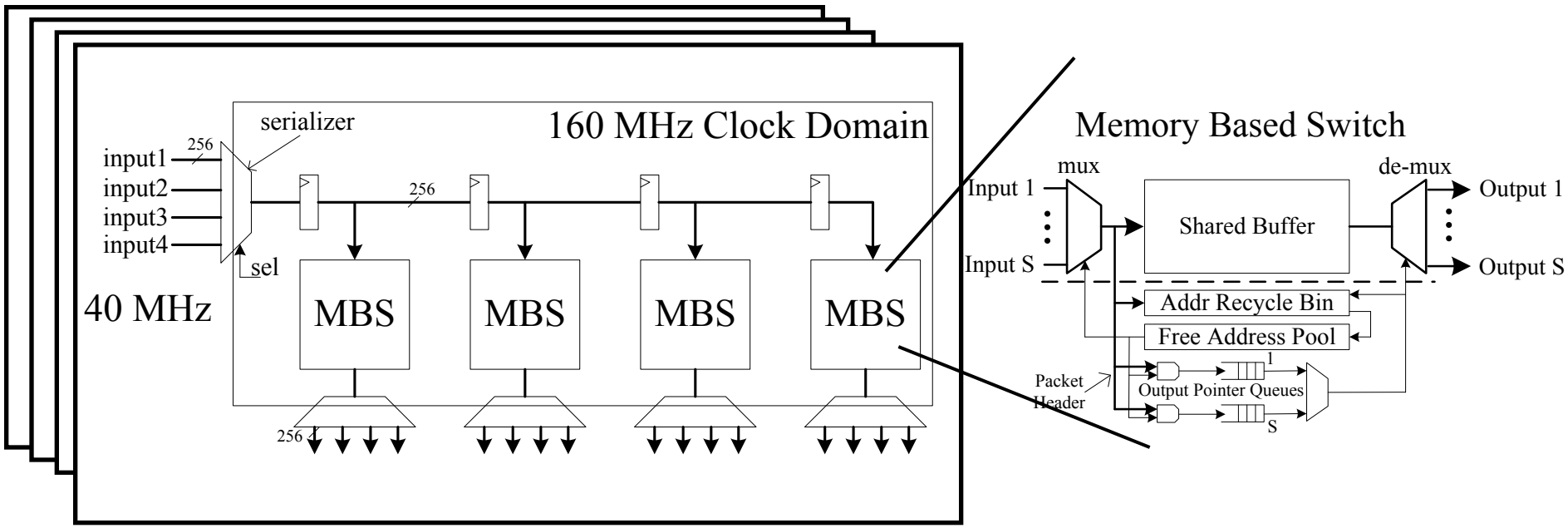
Packet Latency Test

Limit the total buffer size to 1k flits



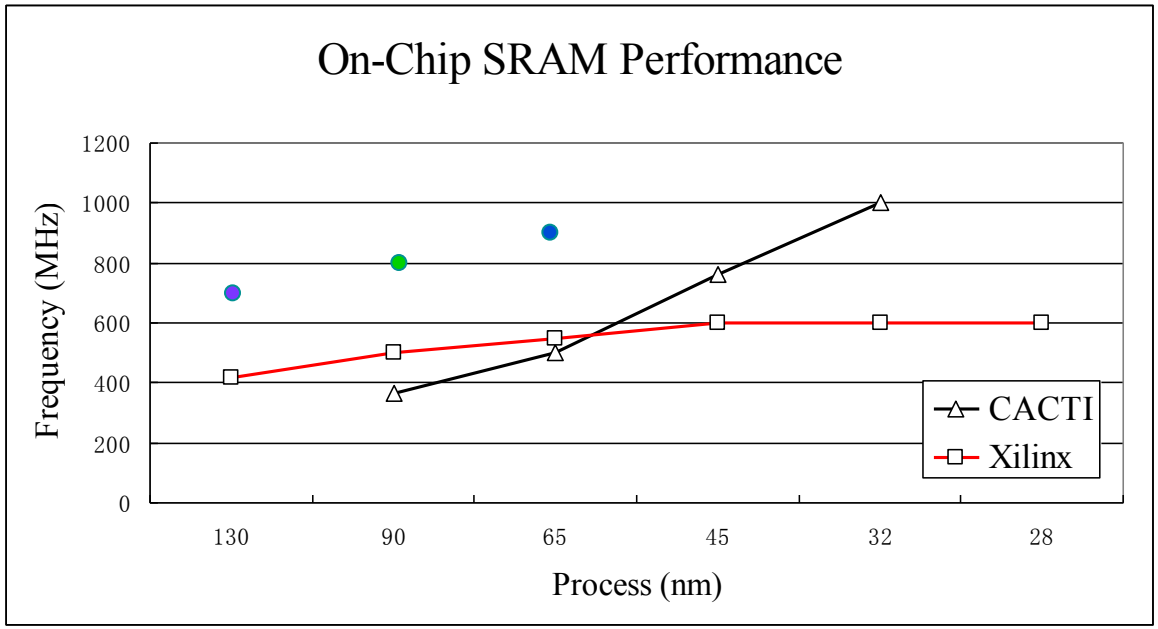
Discussion: Why ISE cannot meet timing?

- Hierarchical compilation, place and router
  - The design is regular and symmetric



Discuss: Why BRAM is slow?

- ✧ CACTI prediction based on ITRS-LSTP data
- ✧ Xilinx BRAM (18Kbit) performance
  - 45 nm → 28 nm, with little improvement on BRAM performance ?



- Fulcrum FM4000 130nm, 2MB: 720MHz
- Sun SPARC 90nm 4MB L2: 800MHz
- Intel Xeon 65nm 16MB L3: 850MHz

Discussion: Extrapolation on Virtex7

- ✧Virtex-7 XC7VH870T :
  - Max BRAM frequency: 600 MHz
  - Total of 2820 18kbit BRAMS
  - 1.4 Tb/s transceiver
- ✧CICQ requirement: > 10k BRAMs
- ✧Proposed GCQ switch:
  - Assume 400MHz BRAM frequency: 2R/B = 0.69
  - $(1-P/S^2) = 91\%$
  - 1014 18kbit BRAMS (36%)

Questions?

Conclusions

- ✧FPGAs can rival ASICs on switch fabric design
  - Remain resource for other functions
- ✧Big room for improvement in FPGA’s on-chip SRAM performance
- ✧FPGA CAD tools can do a better job.

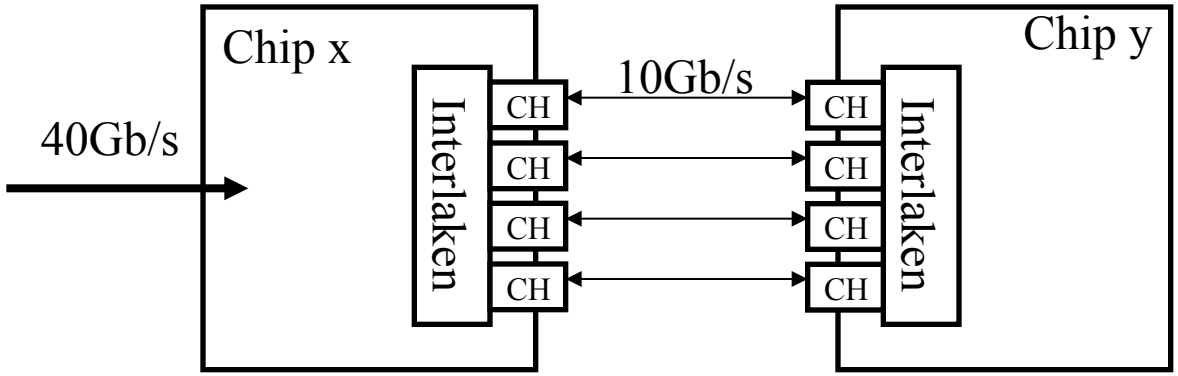
Roadmap for Data Link Speed

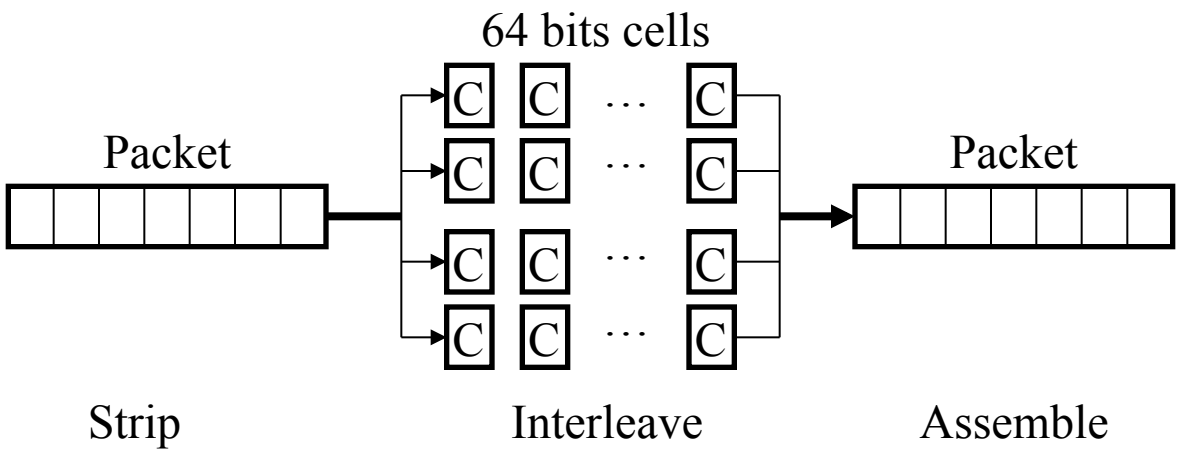
Nathan Binkert *et al.* @ ISCA 2011

	Process	nm	45	32	22
Link Characteristics	Link Speedy	Gb/s	80	160	320
	Max link length	m	10		
	in flight data	Bytes	1107	2214	4428
Optical Link parameters	Data wavelengths		8	16	32
	Optical data rate	Gb/s	**10**		
Electric link parameters	SERDES speed	Gb/s	10	20	32
	SERDES channels		8	8	10

Grows in channel count NOT channel speed

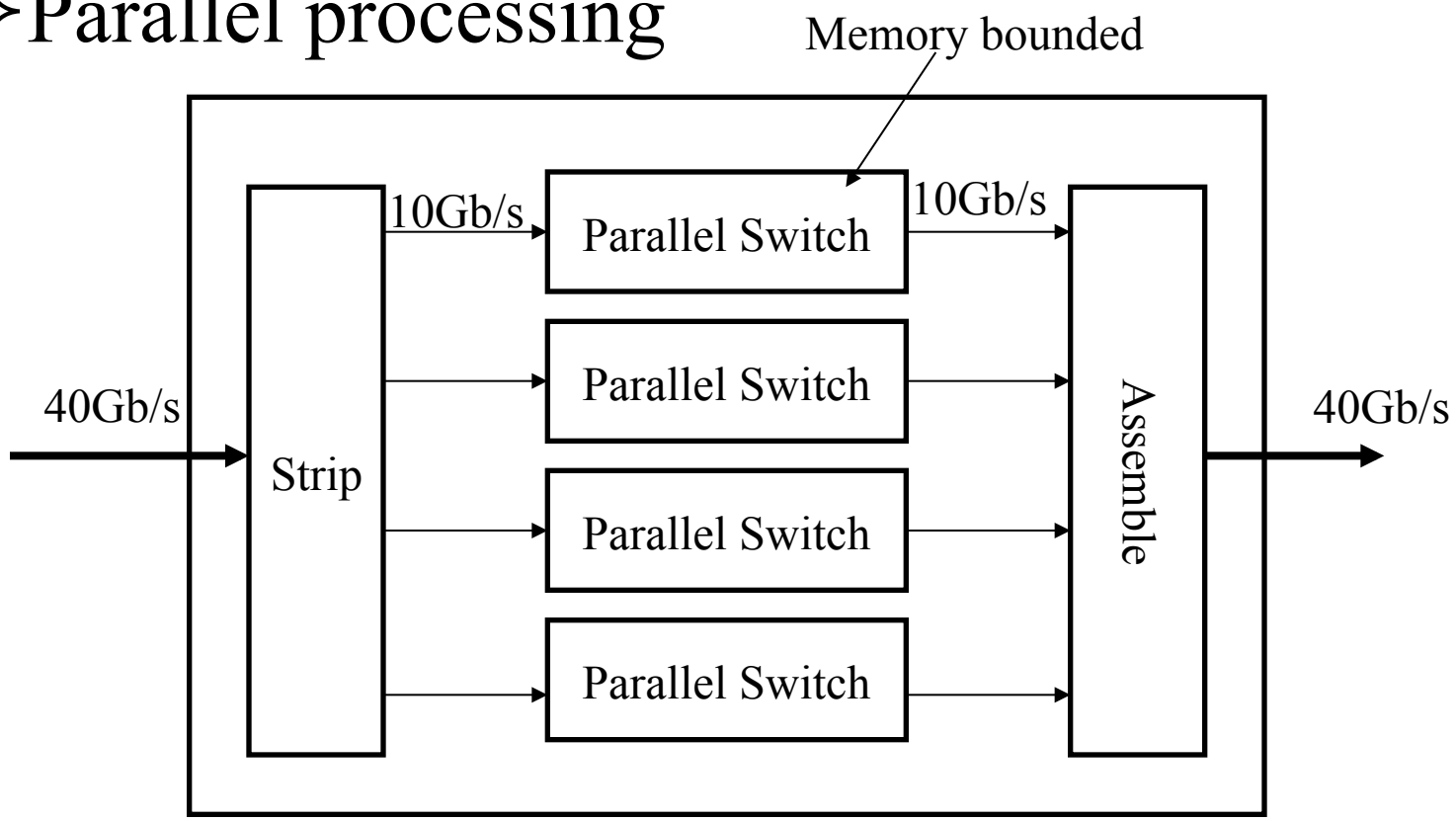
Interlaken


Chip x
40Gb/s
Interlaken
CH
CH
CH
CH
10Gb/s
CH
CH
CH
CH
Interlaken
Chip y


64 bits cells
Packet
C
C
C
C
C
C
C
C
C
C
C
C
Packet
Strip
Interleave
Assemble

High Link Speed Processing

✧Parallel processing



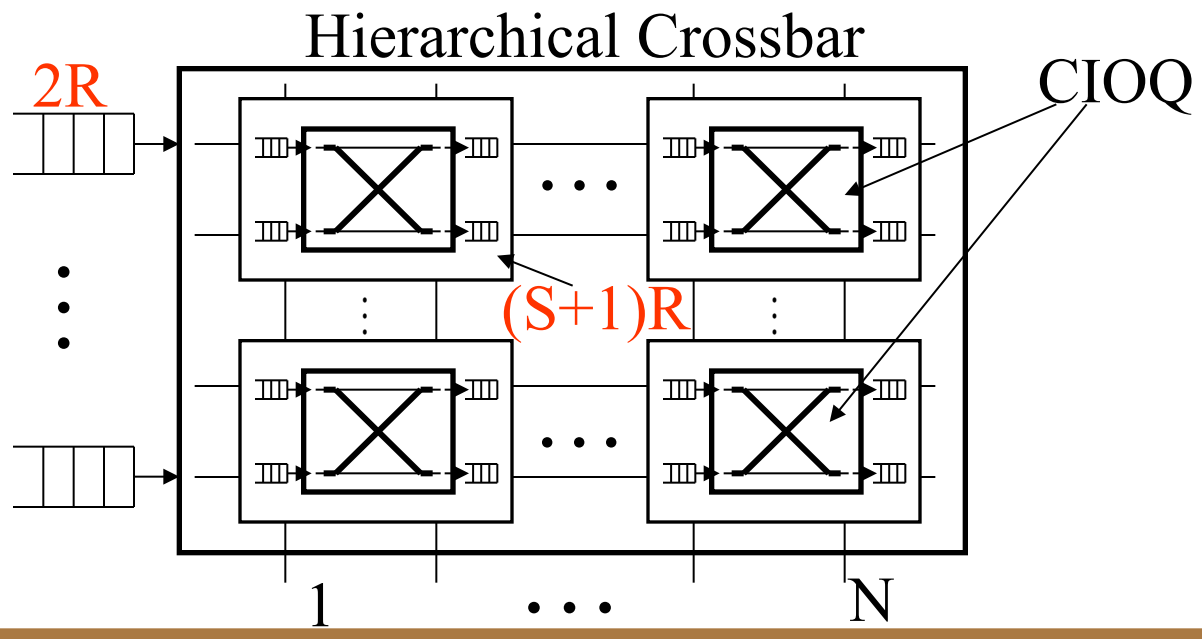
Prefer High radix switch with **thin** ports over Low radix switch with **fat** ports

Latest Improvement Upon CICQ

✧Hierarchical Crossbar --- Dally:

- Require (40%) less buffers
- Higher requirement on memory BW and scheduling of sub-switches

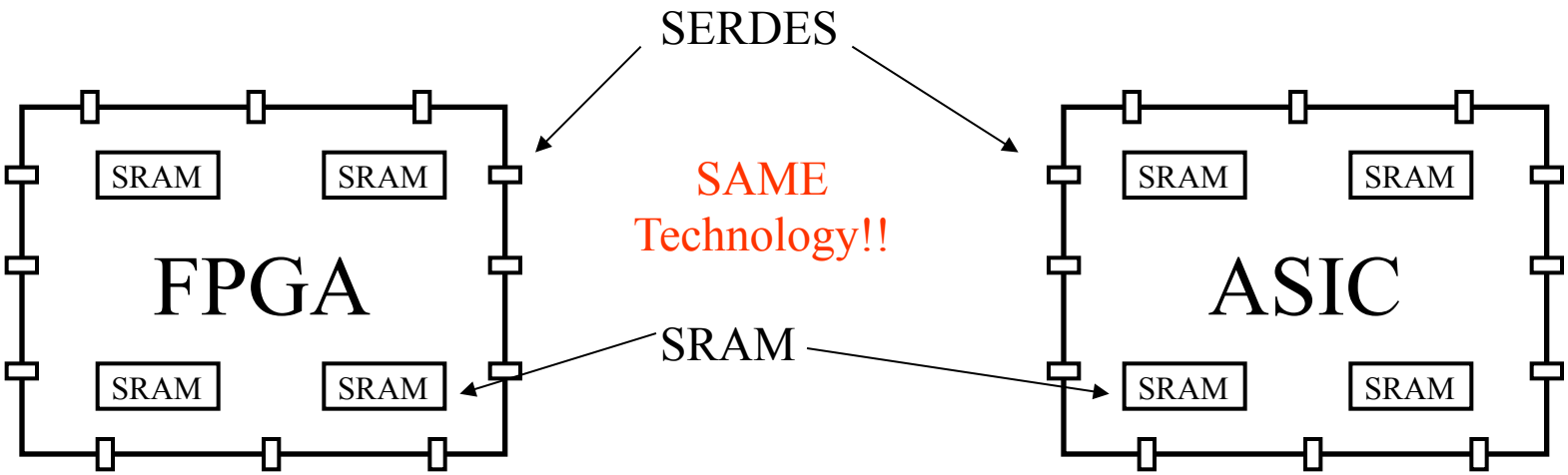
Cray BlackWidow (2006): 2Tb/s



Can We Do Single-Chip Switch Fabric?

✧FPGA VS. ASIC:

- Comparable transceiver BW ☺
- Abundant on-chip SRAMs ☺ (same as ASICs)



Saturate the transceiver BW

Can’t do better