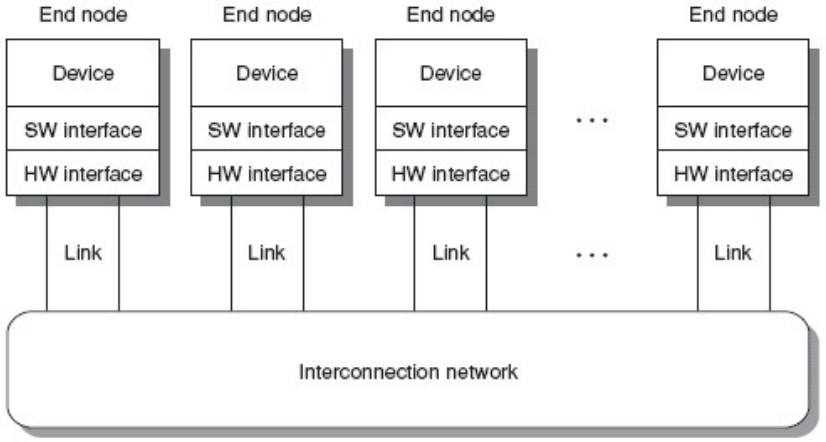
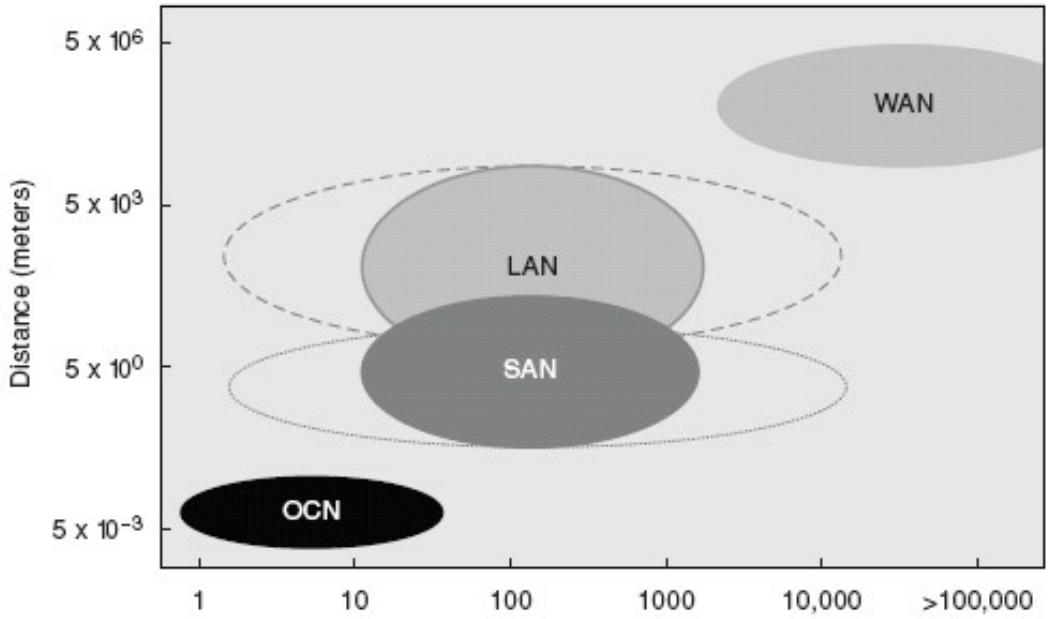
Interconnection Networks



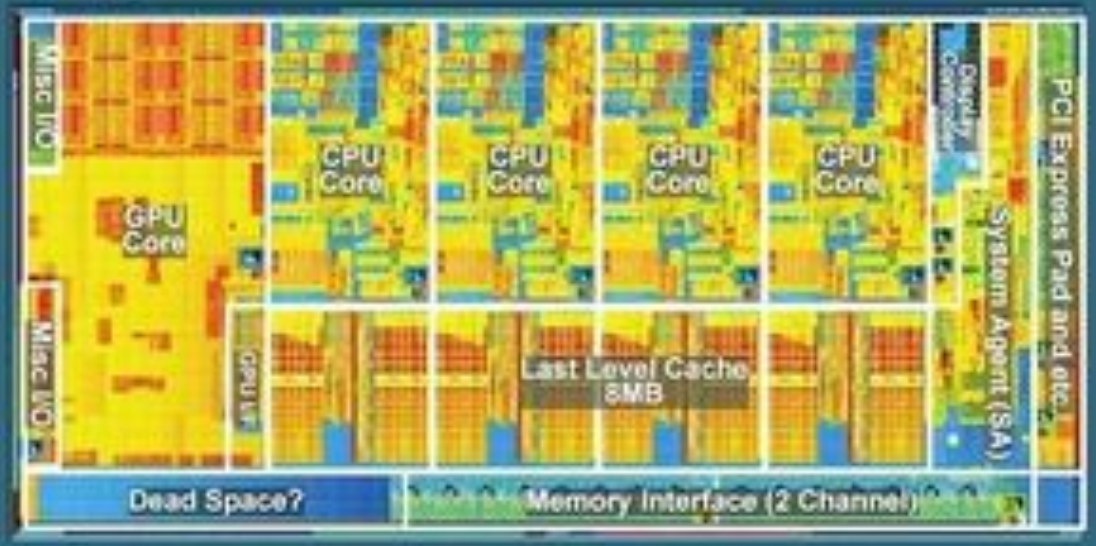


Ivy Bridge and Sandy Bridge Die Layout (Estimated)

Ivy Bridge

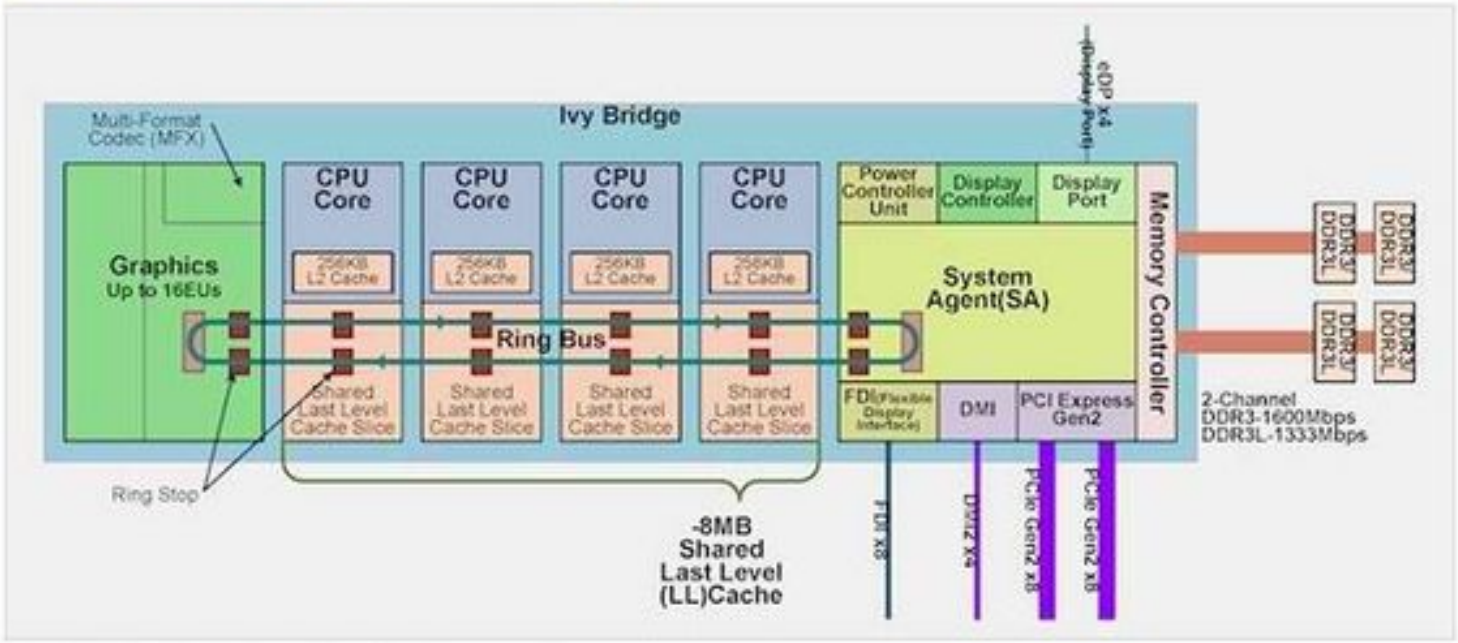
4 CPU cores	22 nm Process
GPU core	1.48B transistors
8MB LL Cache	160 mm2
	TDP

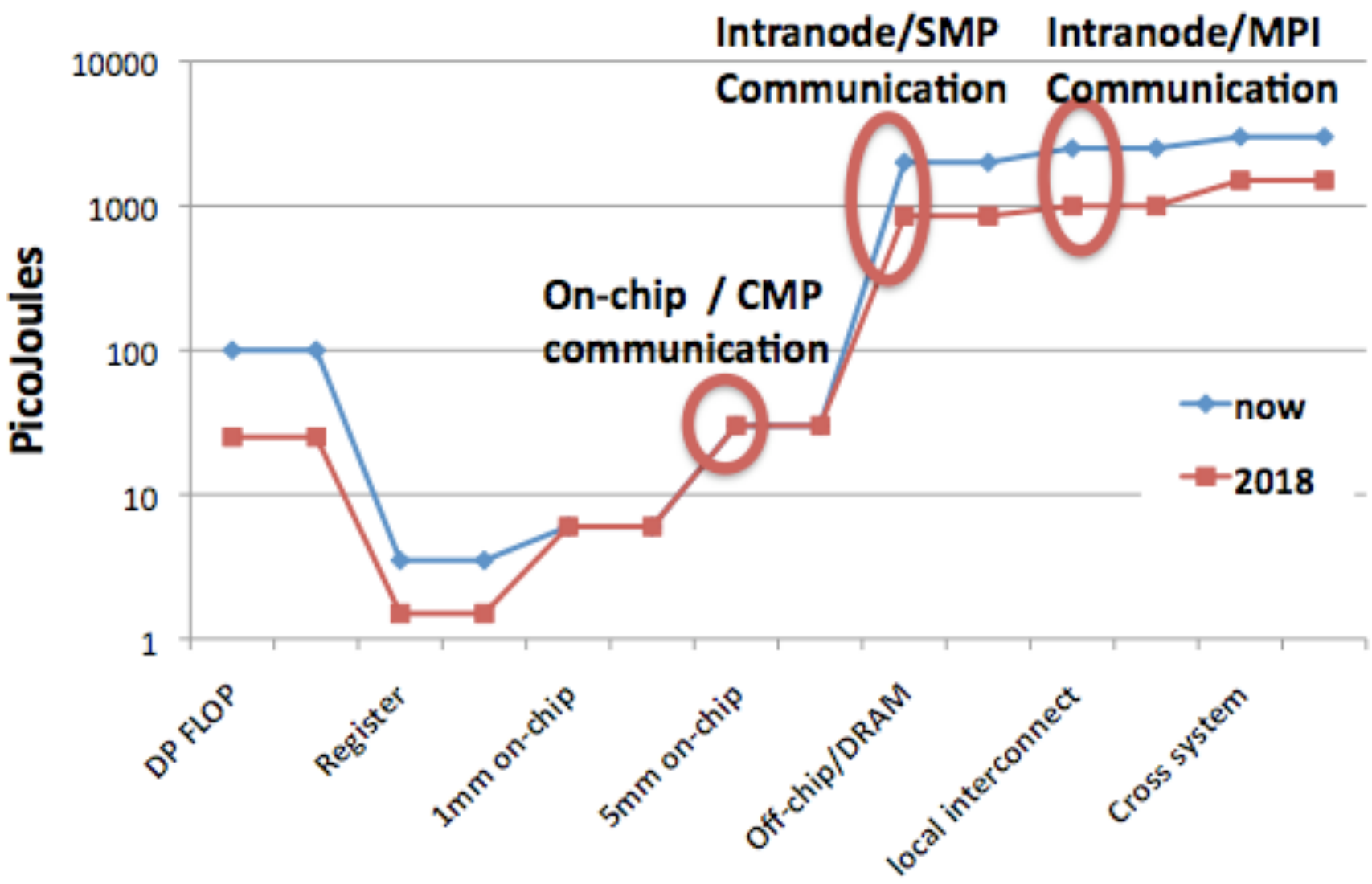
Sandy Bridge



4 CPU cores	32 nm Process
GPU core (12 EUs)	1.16B transistors
8MB LL Cache	216 mm2
PCI Express Gen2 20 Lanes	TDP 95/65/45W(Desktop)

Ivy Bridge 4-core Overview




Intranode/SMP Communication
Intranode/MPI Communication
On-chip / CMP communication
PicoJoules
10000
1000
100
10
1
now
2018
DP FLOP
Register
1mm on-chip
5mm on-chip
Off-chip/DRAM
local interconnect
Cross system

Technology Implications

- Capacity continues to increase
  - DRAM size—2G
  - Disk Size --
  - Transistors on a die
- What happens to latency / delay?
  - Ultimately maxes out at c
    - Where c=3e8 m/s
    - 300ps
    - Speed of light
  - Just better parallelize more links

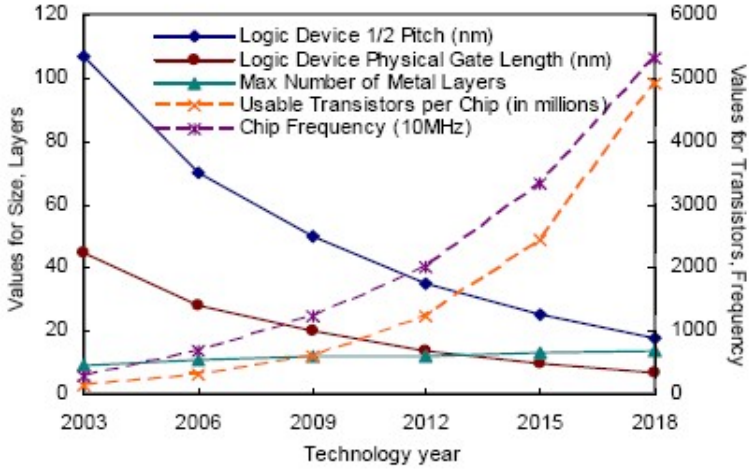


Fig. 1. ITRS-2003 Technology Projections [1].

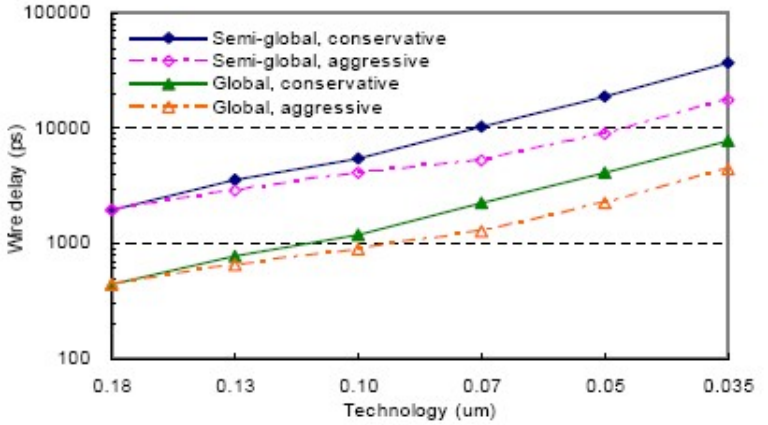


Fig. 2. Unrepeated wire delay (picoseconds) spanning 10 mm distance [4].

Networking Packets

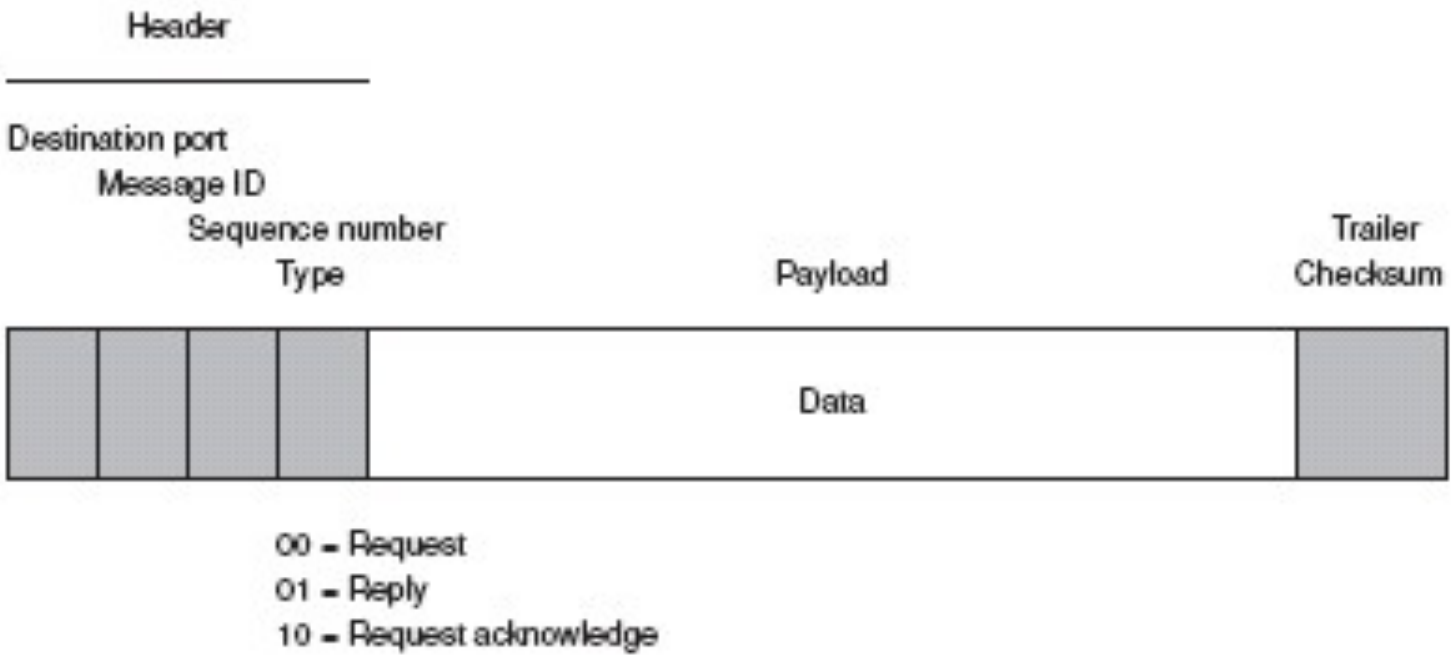


Figure E.4 An example packet format with header, payload, and checksum in the trailer.

Transmission/Latency Overhead

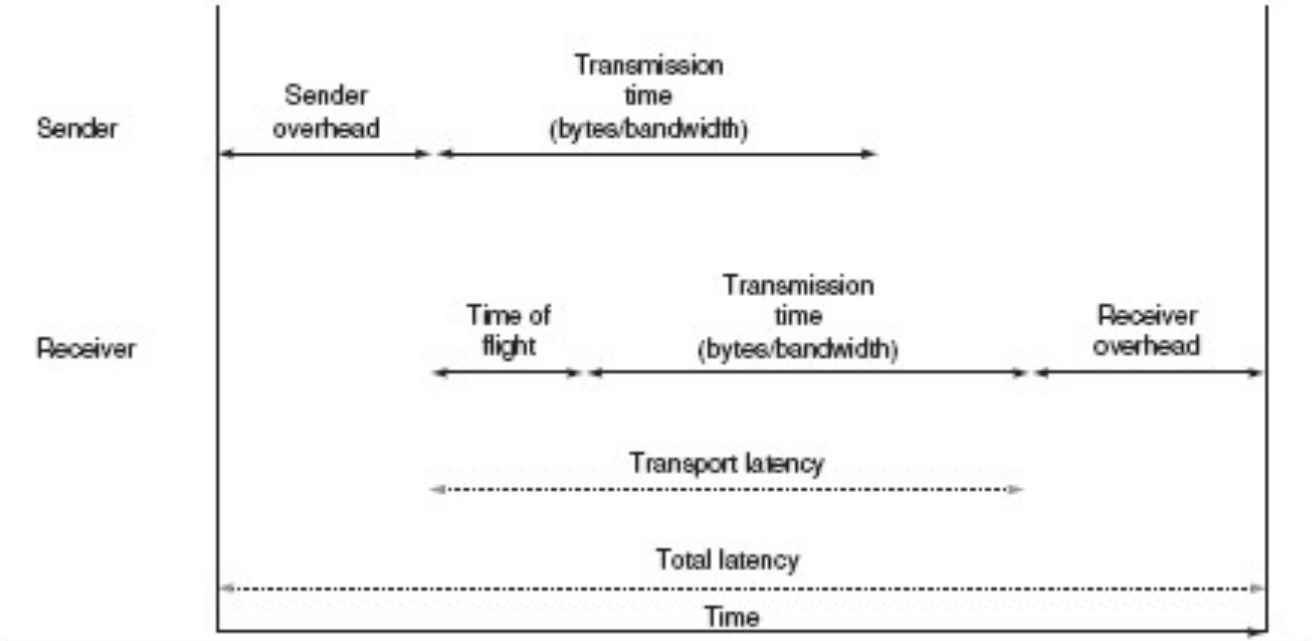


Figure E.5 Components of packet latency. Depending on whether it is an OCN, SAN, LAN, or WAN, the relative amounts of sending and receiving overhead, time of flight, and transmission time are usually quite different from those illustrated here.

Sample Supercomputer Networks

Company	System [network] name	Intro year	Max. number of compute nodes [× # CPUs]	System footprint for max. configuration	Packet [header] max size (bytes)	Injection [reception] node BW in MB/sec	Minimum send/ receive overhead	Maximum copper link length; flow control; error
Intel	ASCI Red Paragon	2001	4510 [× 2]	2500 sq. feet	1984 [4]	400 [400]	few µs	handshaking; CRC + parity
IBM	ASCI White SP Power3 [Colony]	2001	512 [× 16]	10,000 sq. feet	1024 [6]	500 [500]	~ 3 µs	25 m; credit-based; CRC
Intel	Thunder Itanium2 Tiger4 [QsNet^{II}]	2004	1024 [× 4]	120 m^2	2048 [14]	928 [928]	0.240 µs	13 m; credit-based; CRC for link, dest.
Cray	XT3 [SeaStar]	2004	30,508 [× 1]	263.8 m^2	80 [16]	3200 [3200]	few µs	7 m; credit-based; CRC
Cray	X1E	2004	1024 [× 1]	27 m^2	32 [16]	1600 [1600]	0 (direct LD ST accesses)	5 m; credit-based; CRC
IBM	ASC Purple pSeries 575 [Federation]	2005	>1280 [× 8]	6720 sq. feet	2048 [7]	2000 [2000]	~ 1 µs with up to 4 packets processed in ‖	25 m; credit-based; CRC
IBM	Blue Gene/L eServer Sol. [Torus Net.]	2005	65,536 [× 2]	2500 sq. feet (.9 × .9 × 1.9 m^3/1K node rack)	256 [8]	612.5 [1050]	~ 3 µs (2300 cycles)	8.6 m; credit-based; CRC (header/pkt)

Figure E.7 Basic characteristics of interconnection networks in commercial high-performance computer systems.

Shared/Switched Mediums

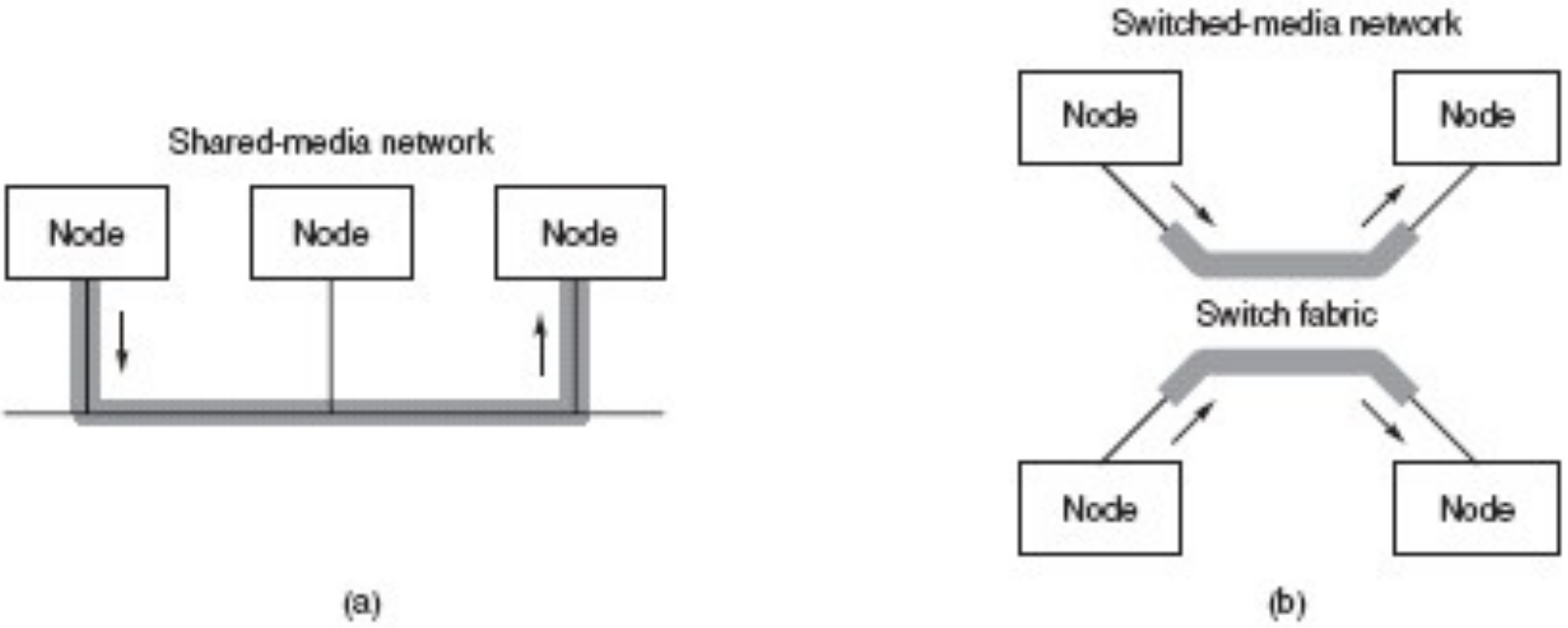


Figure E.8 (a) A shared-media network versus (b) a switched-media network. Ethernet was originally a shared media network, but switched Ethernet is now available. All nodes on the shared-media must dynamically share the raw bandwidth of one link, but switched-media networks can support multiple links, providing higher raw aggregate bandwidth.

Num. of Nodes vs. Latency

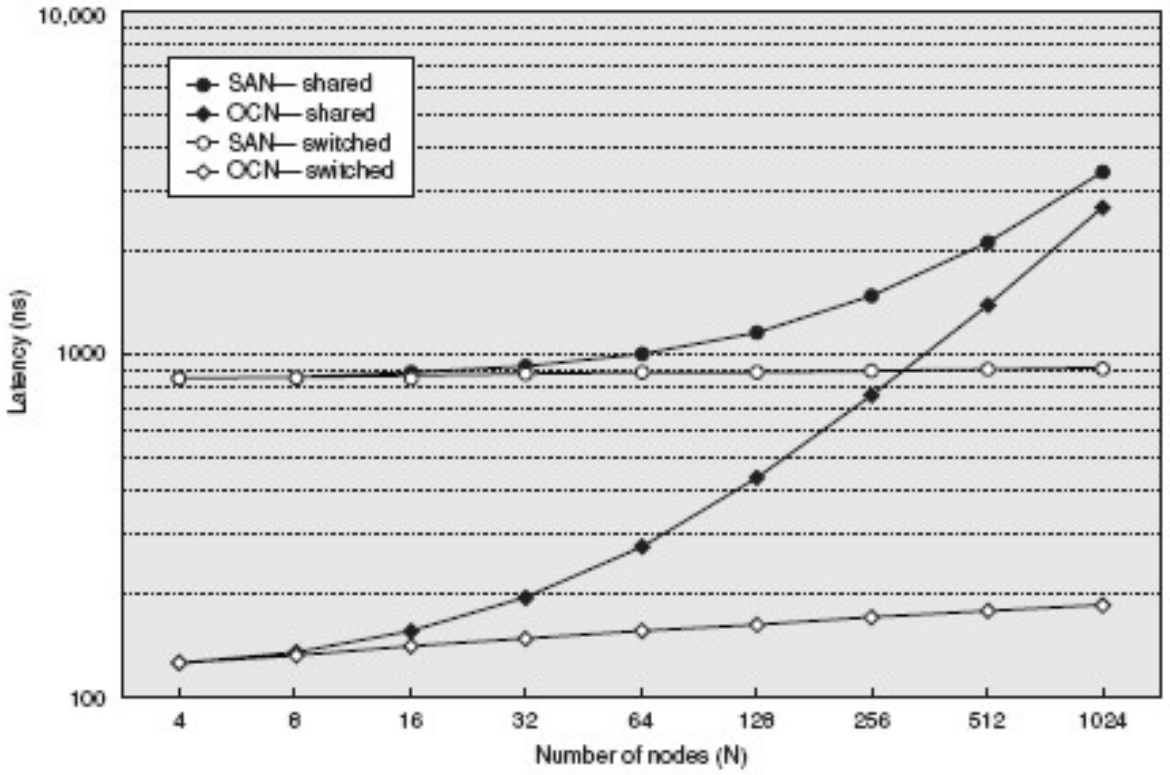
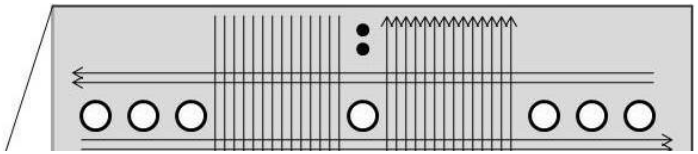


Figure E.9 Latency versus number of interconnected nodes plotted in semi-log form for OCNs and SANs. Routing, arbitration, and switching have more of an impact on latency for networks in these two domains, particularly for networks with a large number of nodes, given the low sending and receiving overheads and low propagation delay.

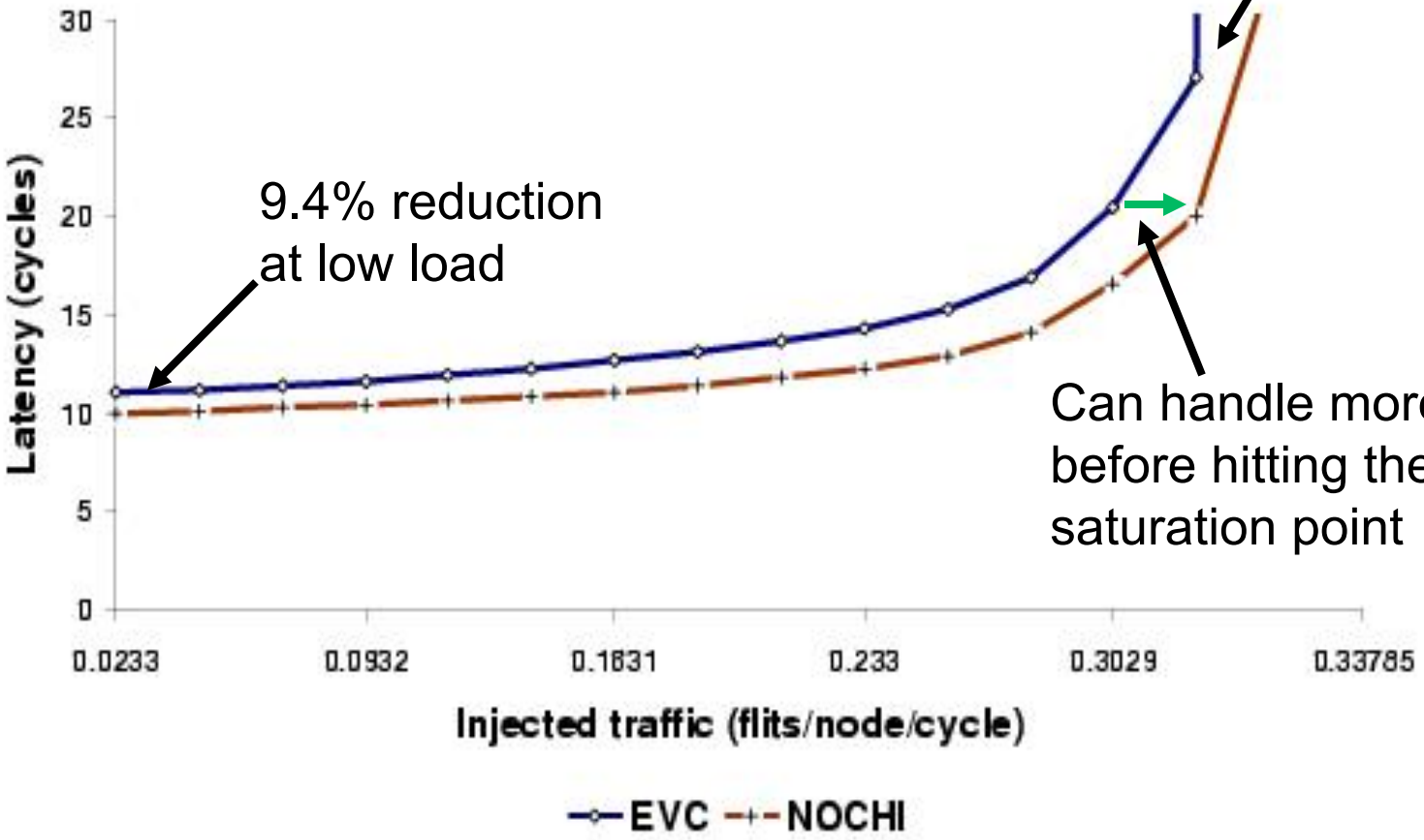
NOCHI: Network-on-chip with hybrid Interconnect



44% reduction
near saturation

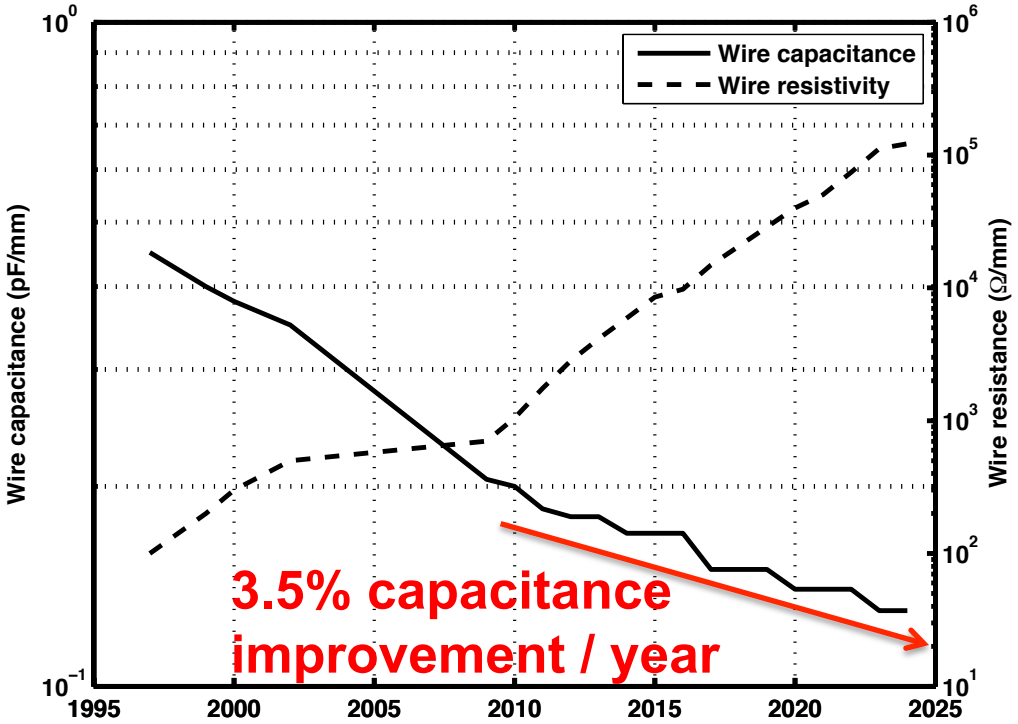
9.4% reduction
at low load

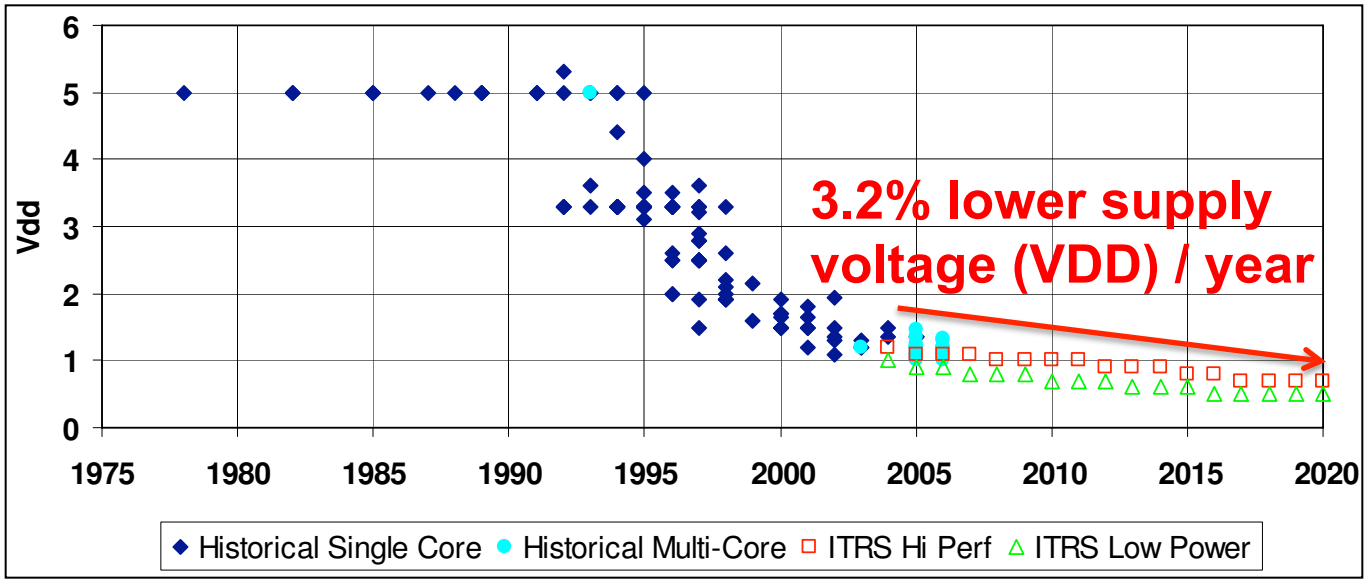
Can handle more load
before hitting the
saturation point



IEEE Micro 2009

(CV^2f) Power not scaling!





Sample Networks

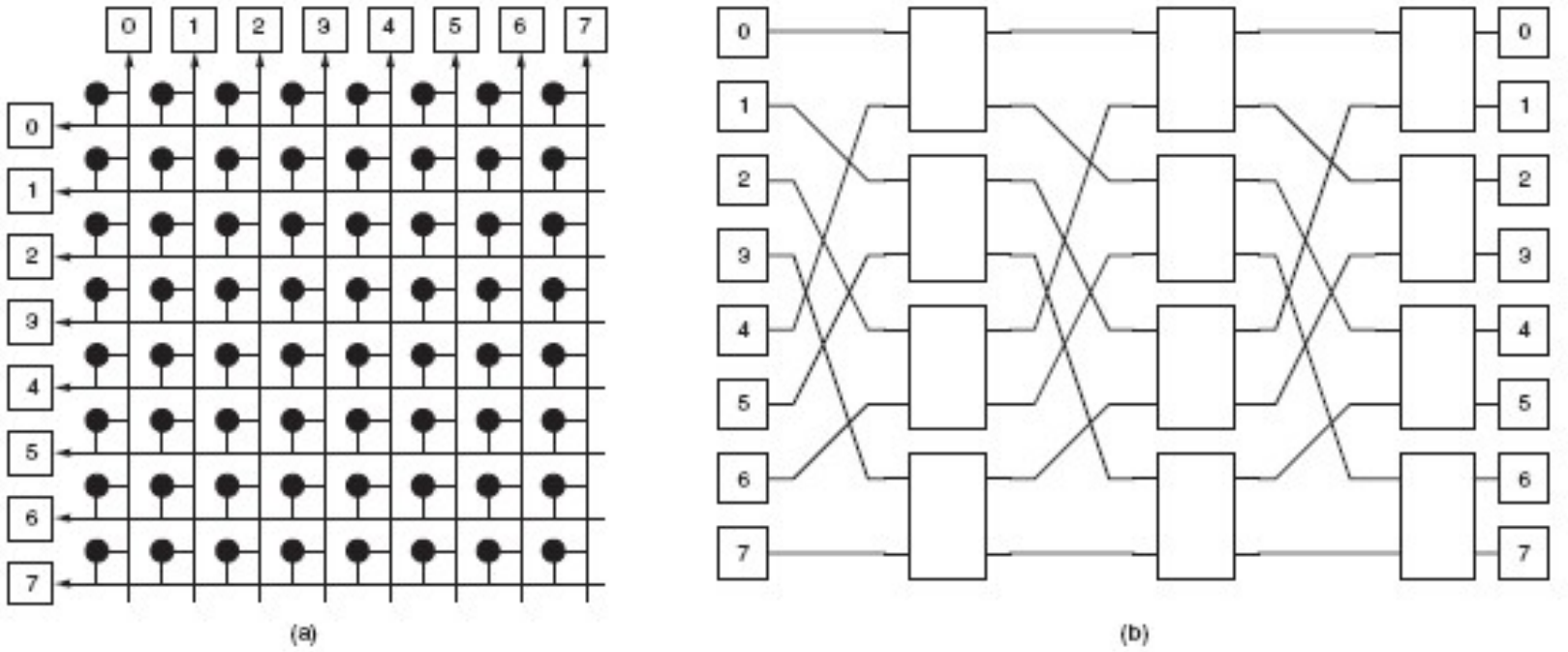


Figure E.11 Popular centralized switched networks: (a) the crossbar network requires N^2 crosspoint switches, shown as black dots; (b) the Omega, a MIN, requires $N/2 \log_2 N$ switches, shown as vertical rectangles. End node devices are shown as numbered squares (total of eight). Links are unidirectional—data enter at the left and exit out the top or right.

Higher Order Networks

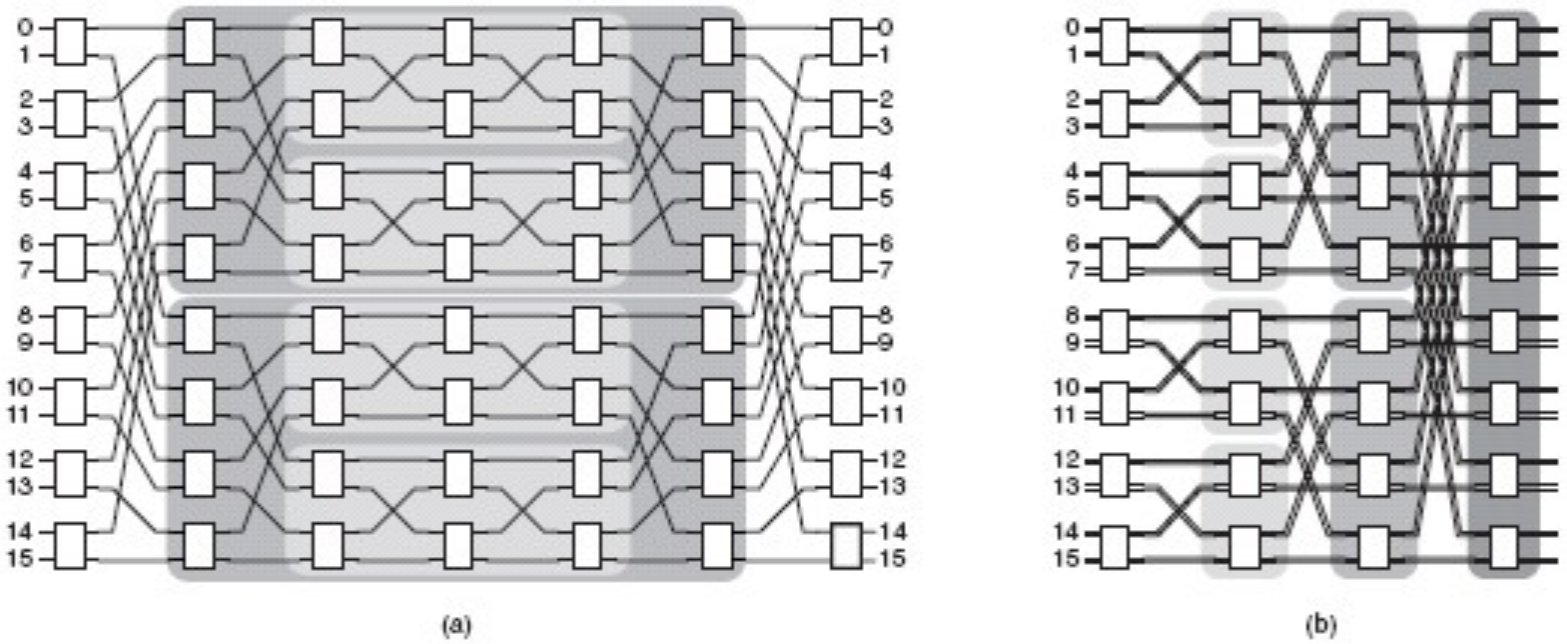


Figure E.12 Two Beneŝ networks. (a) A 16-port Clos topology, where the middle-stage switches shown in the darker shading are implemented with another Clos network whose middle-stage switches shown in the lighter shading are implemented with yet another Clos network, and so on, until a Beneŝ network is produced that uses only 2×2 switches everywhere. (b) A folded Beneŝ network (bidirectional) in which 4×4 switches are used; end nodes attach to the innermost set of the Beneŝ network (unidirectional) switches. This topology is equivalent to a fat tree, where tree vertices are shown in shades.

Others

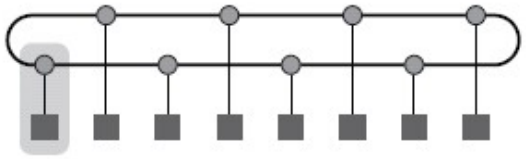
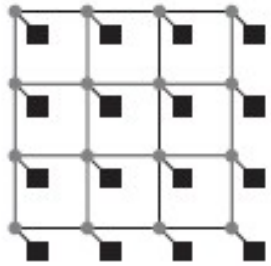
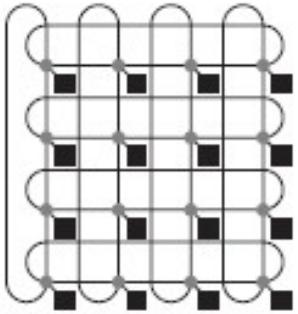


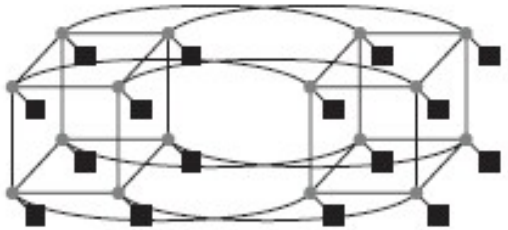
Figure E.13 A ring network topology, folded to reduce the length of the longest link. Shaded circles represent switches, and black squares represent end node devices. The gray rectangle signifies a network node consisting of a switch, a device, and its connecting link.



(a) 2D grid or mesh of 16 nodes



(b) 2D torus of 16 nodes



(c) Hypercube of 16 nodes ($16 = 2^4$ so $n = 4$)

Figure E.14 Direct network topologies that have appeared in commercial systems, mostly supercomputers. The shaded circles represent switches, and the black squares represent end node devices. Switches have many bidirectional network links, but at least one link goes to the end node device. These basic topologies can be supplemented with extra links to improve performance and reliability. For example, connecting the switches on the periphery of the 2-D mesh using the unused ports on each switch forms a 2-D torus. The hypercube topology is an n-dimensional interconnect for 2^n nodes, requiring $n + 1$ ports per switch: one for the n nearest neighbor nodes and one for the end node device.

Bandwidth/Hops for Examples

Evaluation category	Bus	Ring	2D mesh	2D torus	Hypercube	Fat tree	Fully connected
Performance							
$BW_{Bisection}$ in # links	1	2	8	16	32	32	1024
Max (ave.) hop count	1 (1)	32 (16)	14 (7)	8 (4)	6 (3)	11 (9)	1 (1)
Cost							
I/O ports per switch	NA	3	5	5	7	4	64
Number of switches	NA	64	64	64	64	192	64
Number of net. links	1	64	112	128	192	320	2016
Total number of links	1	128	176	192	256	384	2080

Figure E.15 Performance and cost of several network topologies for 64 nodes. The bus is the standard reference at unit network link cost and bisection bandwidth. Values are given in terms of bidirectional links and ports. Hop count includes a switch and its output link, but not the injection link at end nodes. Except for the bus, values are given for the number of network links and total number of links, including injection/reception links between end node devices and the network.

Architecture of Sample Networks

Company	System [network] name	Max. number of nodes [× # CPUs]	Basic network topology	Injection [reception] node BW in MB/sec	# of data bits per link per direction	Raw network link BW per direction in MB/sec	Raw network bisection BW (bidirectional) in GB/sec
Intel	ASCI Red Paragon	4816 [× 2]	2D mesh 64 × 64	400 [400]	16 bits	400	51.2
IBM	ASCI White SP Power3 [Colony]	512 [× 16]	bidirectional MIN with 8-port bidirectional switches (typically a fat tree or Omega)	500 [500]	8 bits (+ 1 bit of control)	500	256
Intel	Thunder Itanium2 Tiger4 [QsNet^{II}]	1024 [× 4]	fat tree with 8-port bidirectional switches	928 [928]	8 bits (+ 2 of control for 4b/5b encoding)	1333	1365
Cray	XT3 [SeaStar]	30,508 [× 1]	3D torus 40 × 32 × 24	3200 [3200]	12 bits	3800	5836.8
Cray	X1E	1024 [× 1]	4-way bristled 2D torus (~ 23 × 11) with express links	1600 [1600]	16 bits	1600	51.2
IBM	ASC Purple pSeries 575 [Federation]	>1280 [× 8]	bidirectional MIN with 8-port bidirectional switches (typically a fat tree or Omega)	2000 [2000]	8 bits (+ 2 bits of control for novel 5b/6b encoding scheme)	2000	2560
IBM	Blue Gene/L eServer Sol. [Torus Net.]	65,536 [× 2]	3D torus 32 × 32 × 64	612.5 [1050]	1 bit (bit serial)	175	358.4

Figure E.16 Topological characteristics of interconnection networks used in commercial high-performance machines.

Network Routing/Switching Types

Company	System [network] name	Max. number of nodes [× # CPUs]	Basic network topology	Switch queuing (buffers)	Network routing algorithm	Switch arbitration technique	Network switching technique
Intel	ASCI Red Paragon	4510 [× 2]	2D mesh 64 × 64	input buffered (1 flit)	distributed dimension-order routing	2-phased RR, distributed across switch	wormhole with no virtual channels
IBM	ASCI White SP Power3 [Colony]	512 [× 16]	bidirectional MIN with 8-port bidirectional switches (typically a fat-tree or Omega)	input and central buffer with output queuing (8-way speedup)	source-based LCA adaptive, shortest-path routing and table-based multicast routing	2-phased RR, centralized and distributed at outputs for bypass paths	buffered wormhole and virtual cut-through for multicasting, no virtual channels
Intel	Thunder Itanium2 Tiger4 [QsNet^{II}]	1024 [× 4]	fat tree with 8-port bidirectional switches	input buffered	source-based LCA adaptive, shortest-path routing	2-phased RR, priority, aging, distributed at output ports	wormhole with 2 virtual channels
Cray	XT3 [SeaStar]	30,508 [× 1]	3D torus 40 × 32 × 24	input with staging output	distributed table-based dimension-order routing	2-phased RR, distributed at output ports	virtual cut-through with 4 virtual channels
Cray	X1E	1024 [× 1]	4-way bristled 2D torus (~ 23 × 11) with express links	input with virtual output queuing	distributed table-based dimension-order routing	2-phased wavefront (pipelined) global arbiter	virtual cut-through with 4 virtual channels
IBM	ASC Purple pSeries 575 [Federation]	>1280 [× 8]	bidirectional MIN with 8-port bidirectional switches (typically a fat tree or Omega)	input and central buffer with output queuing (8-way speedup)	source and distributed table-based LCA adaptive, shortest-path routing and multicast	2-phased RR, centralized and distributed at outputs for bypass paths	buffered wormhole and virtual cut-through for multicasting with 8 virtual channels
IBM	Blue Gene/L eServer Solution [Torus Net.]	65,536 [× 2]	3D torus 32 × 32 × 64	input-output buffered	distributed, adaptive with bubble escape virtual channel	2-phased SLQ, distributed at input and output	virtual cut-through with 4 virtual channels

Figure E.20 Routing, arbitration, and switching characteristics of interconnections networks in commercial machines.

Sample 2D Router/Switch

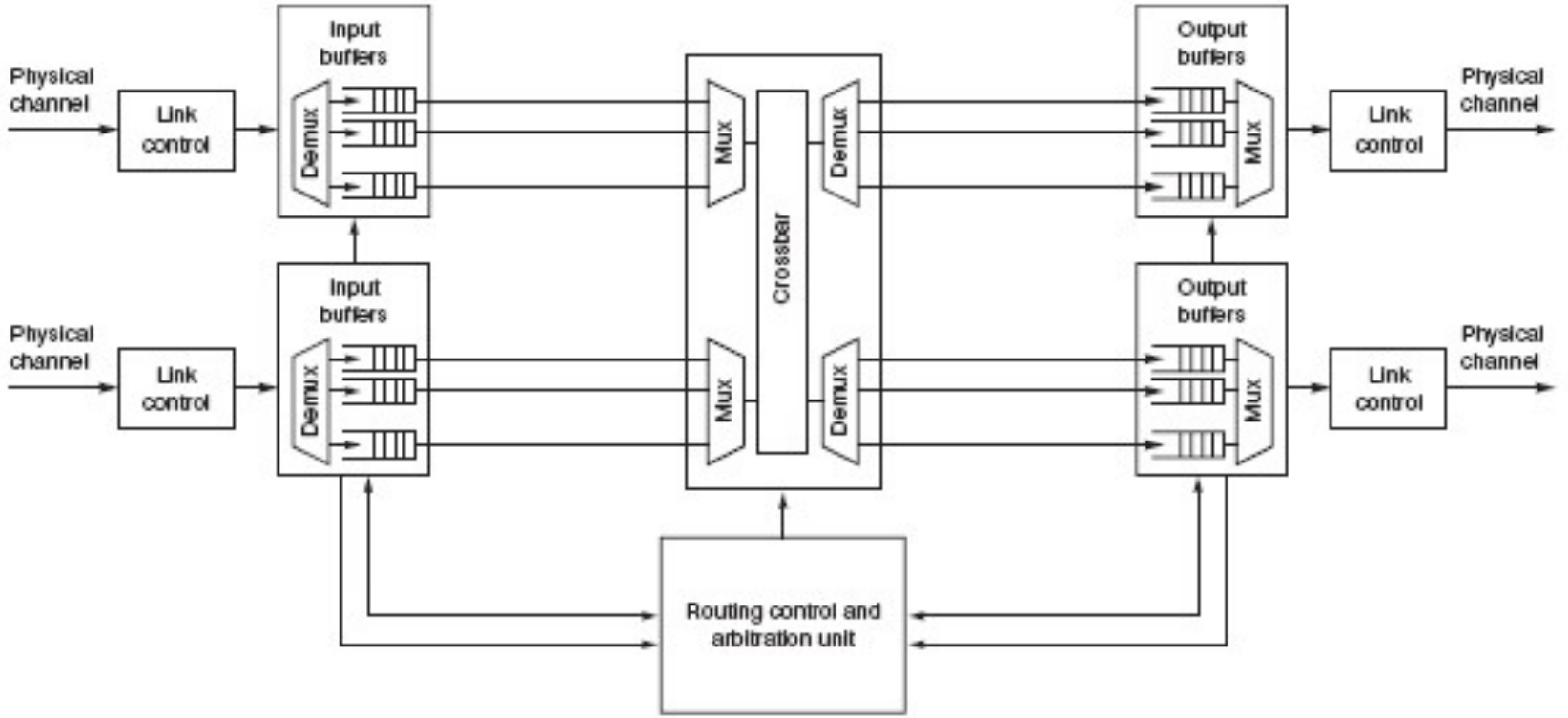


Figure E.21 Basic microarchitectural components of an input-output-buffered switch.

Pipelined Router

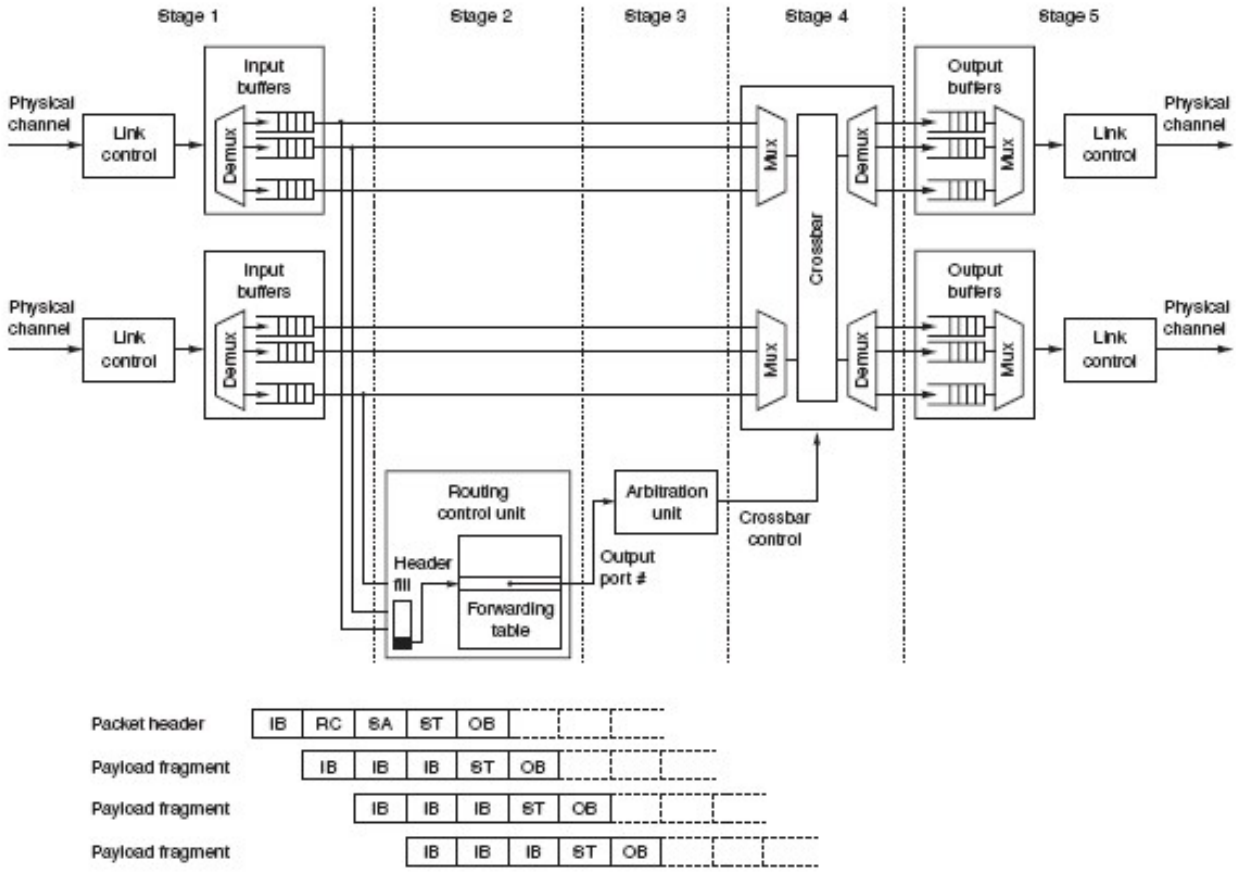


Figure E.23 Pipelined version of the basic input-output-buffered switch. The notation in the figure is as follows: IB is the input link control and buffer stage, RC is the route computation stage, SA is the crossbar switch arbitration stage, ST is the crossbar switch traversal stage, and OB is the output buffer and link control stage. Packet fragments (flits) coming after the header remain in the IB stage until the header is processed and the crossbar switch resources are provided.

Sample Cell Processor Interconnect

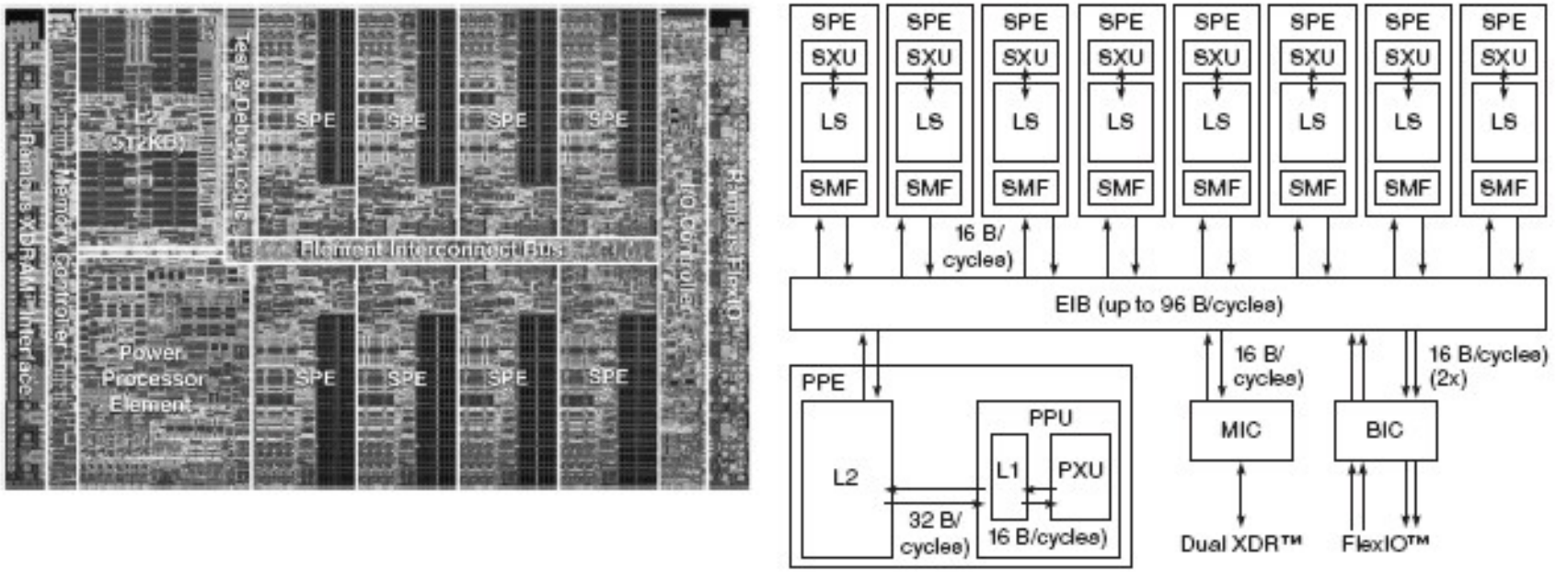


Figure E.25 Cell Broadband Engine (a) die photo and (b) high-level block diagram illustrating the function of the EIB.

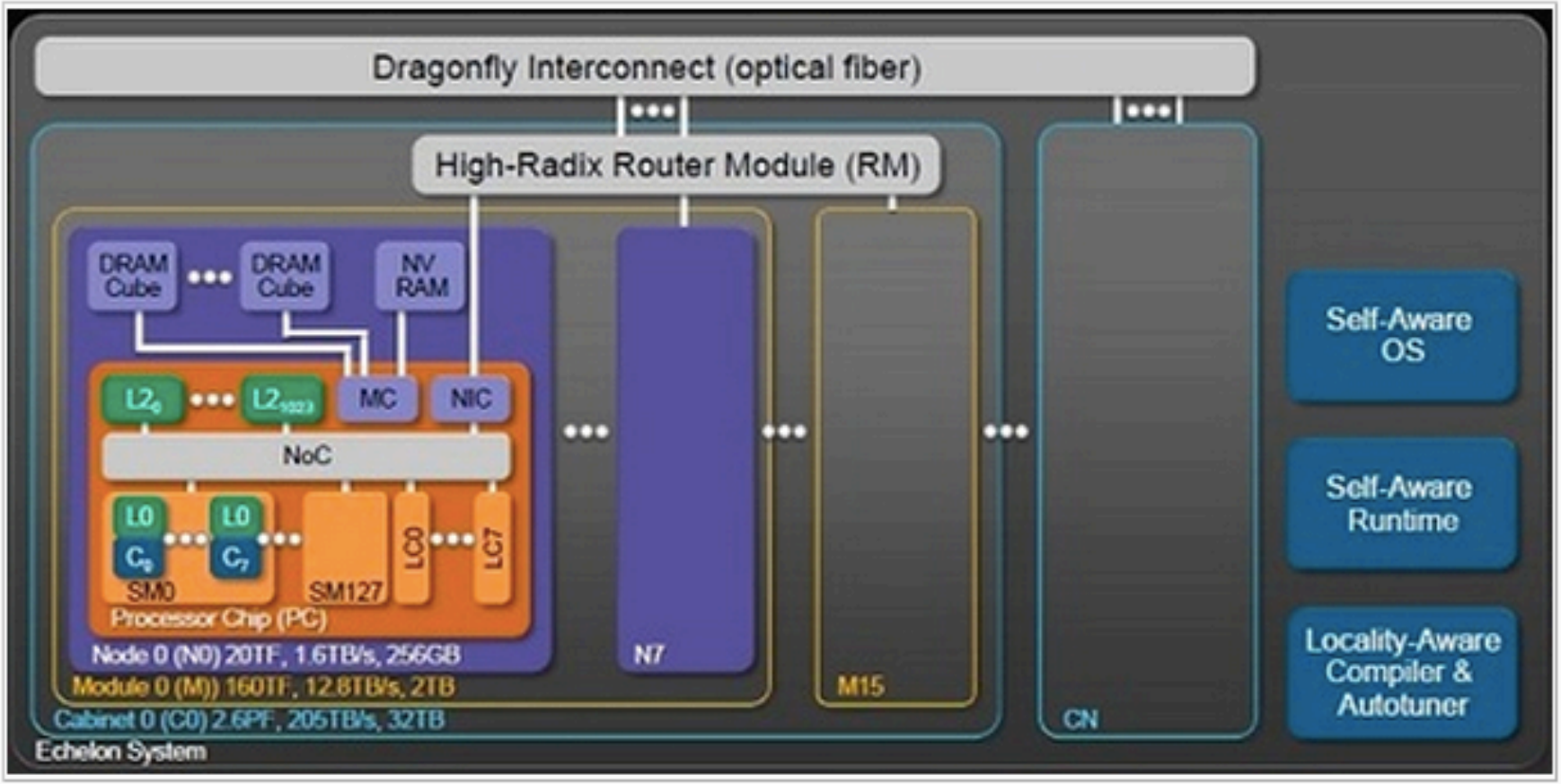
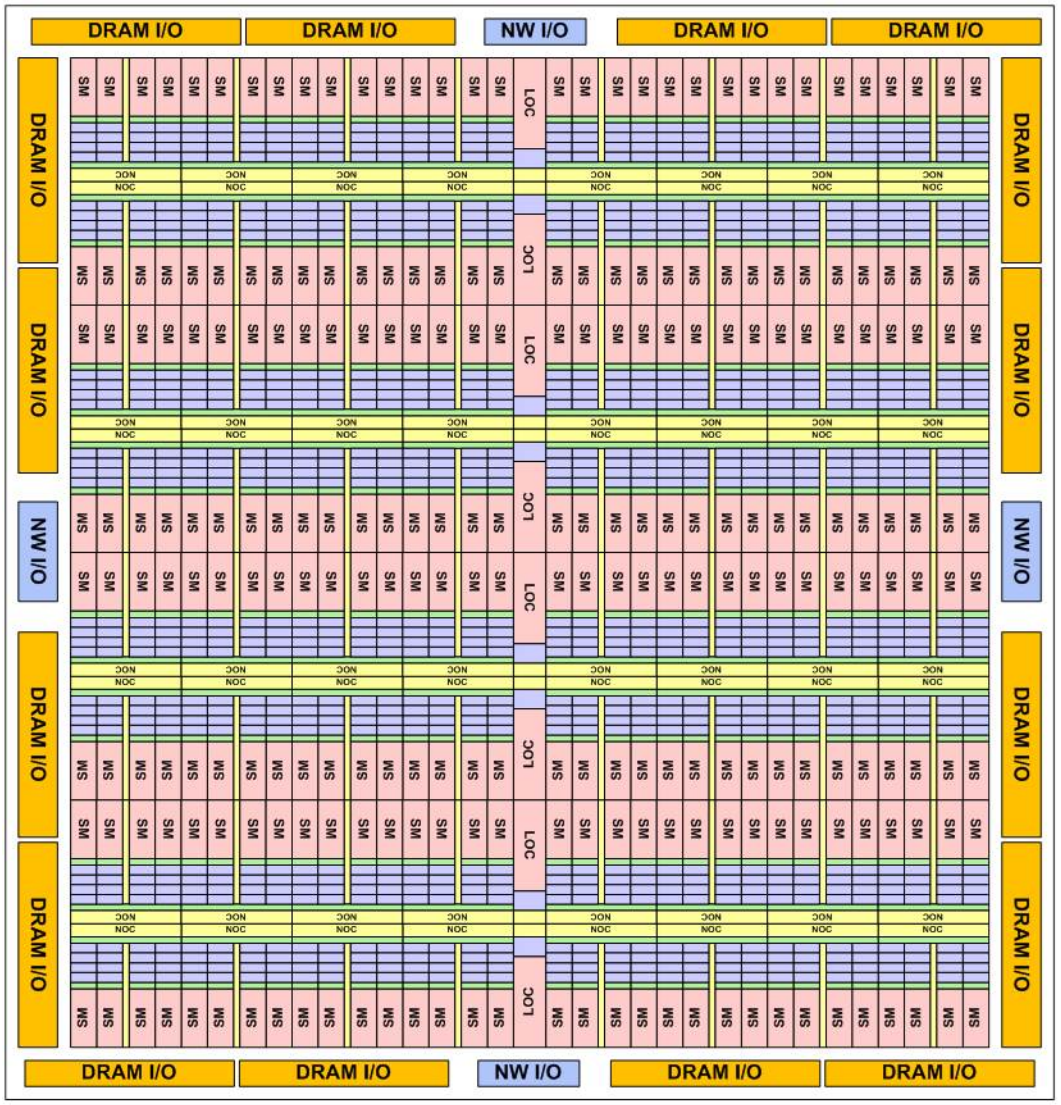
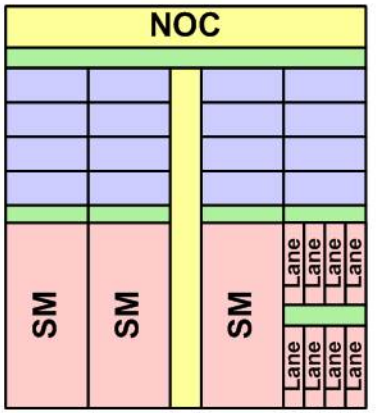


Figure 3: nVIDIA's Echelon Architecture



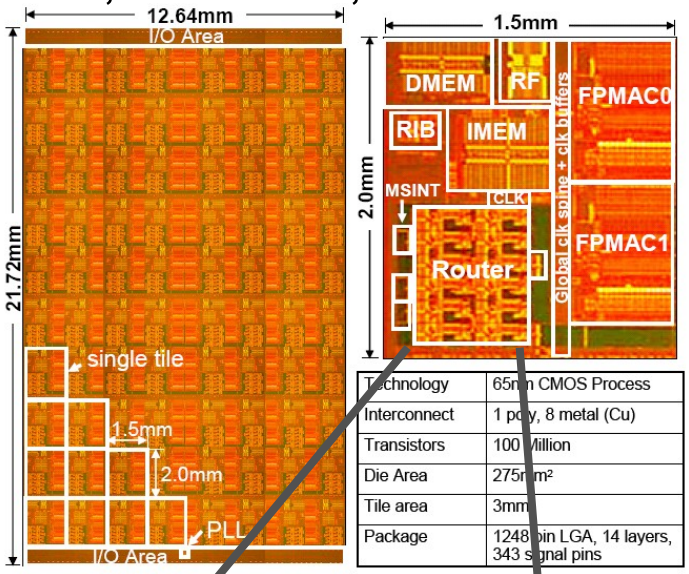
17mm



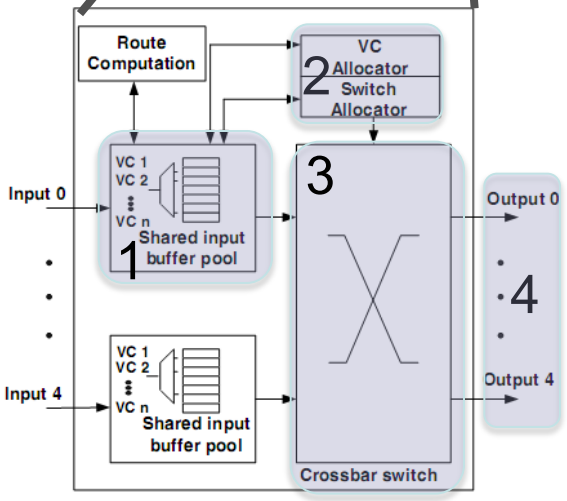
10nm process
290mm^2

(1) Energy-Efficient, On-Chip Links

Intel, 80 Cores, ISSCC 2007



Technology	65nm CMOS Process
Interconnect	1 poly, 8 metal (Cu)
Transistors	100 Million
Die Area	275mm²
Tile area	3mm
Package	1248 pin LGA, 14 layers, 343 signal pins

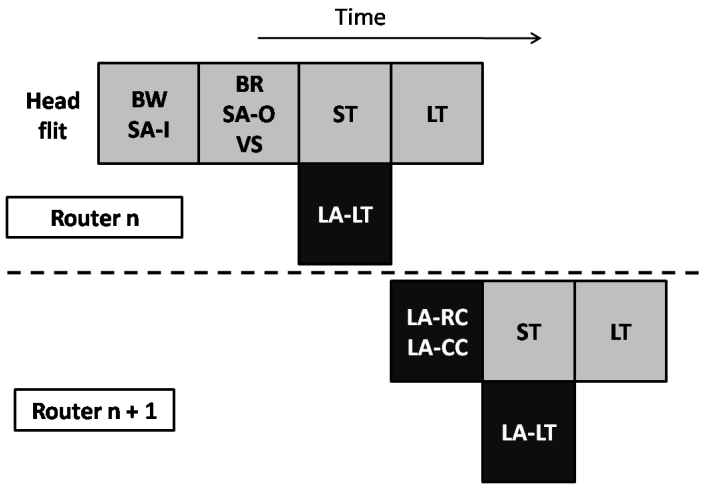


- Router Power:
  - (1) Buffering: 30%
  - (2) Arbitration: 10%
  - (3) XBAR: 30%
  - (4) LINKS: 30%

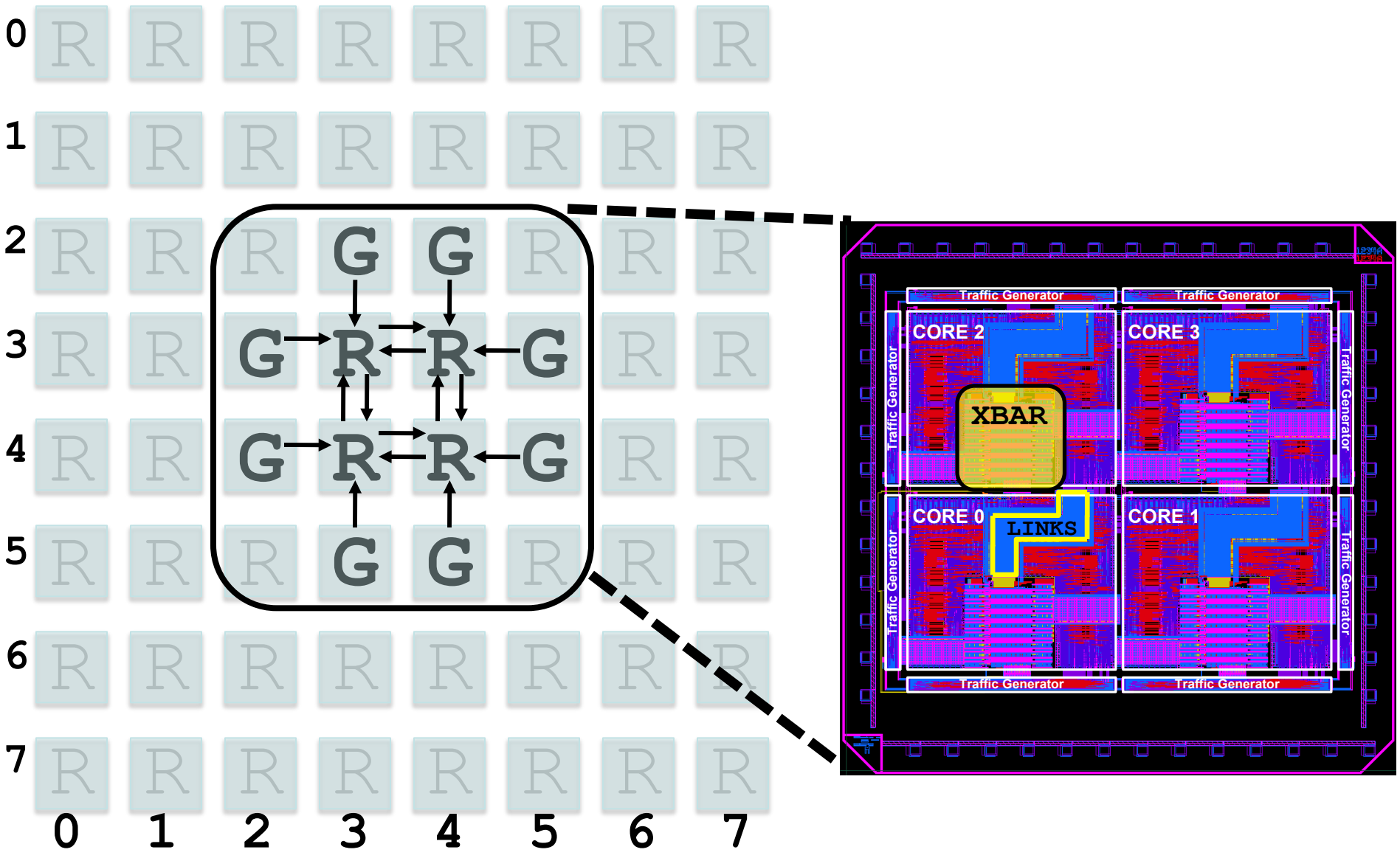
- Our Goal: low-power on-chip links
  - *Analog* low-voltage swing:
    - (3) XBARS
    - (4) LINK TRAVERSAL

Token Flow Control NoC (Li-Shiuan Peh, MIT)

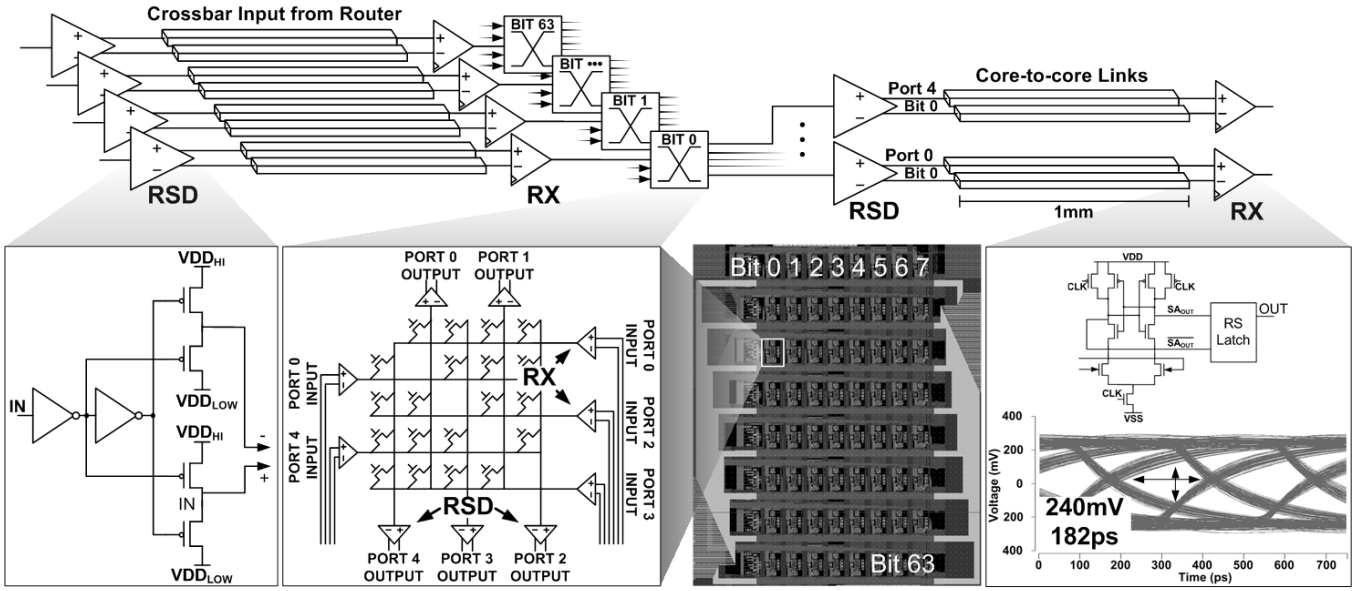
- **Conventional Router:**
  - Each hop requires 4 cycles
- **Proposed TFC Router:**
  - First hop requires 4 cycles
  - Following hops require 2 cycles

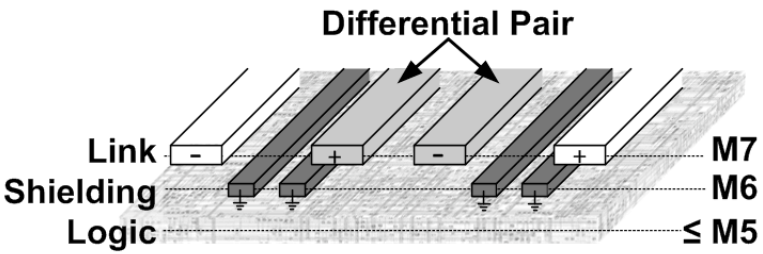


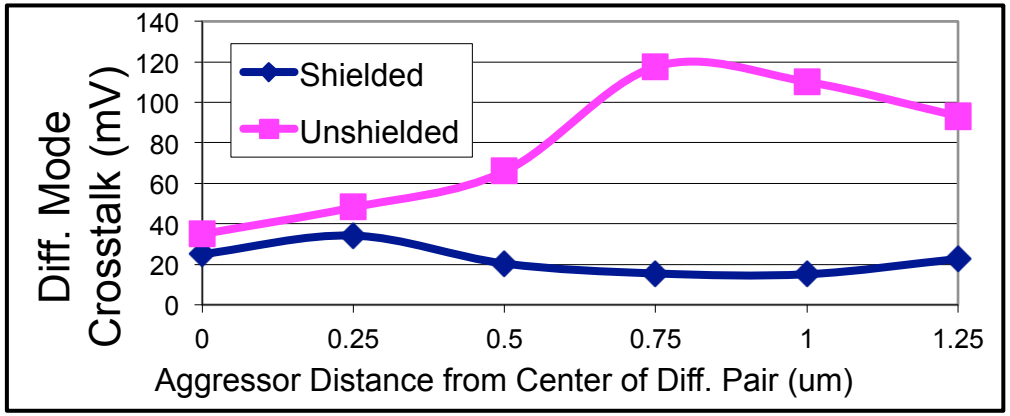
- **Tokens for advance allocation**
  - If little congestion, buffering is skipped
- **NoC power dominated by XBAR and LT**
  - TFC reduces buffer writes


0
1
2
3
4
5
6
7
0 1 2 3 4 5 6 7
R
G
Traffic Generator
CORE 2
CORE 3
CORE 0
CORE 1
XBAR
LINKS

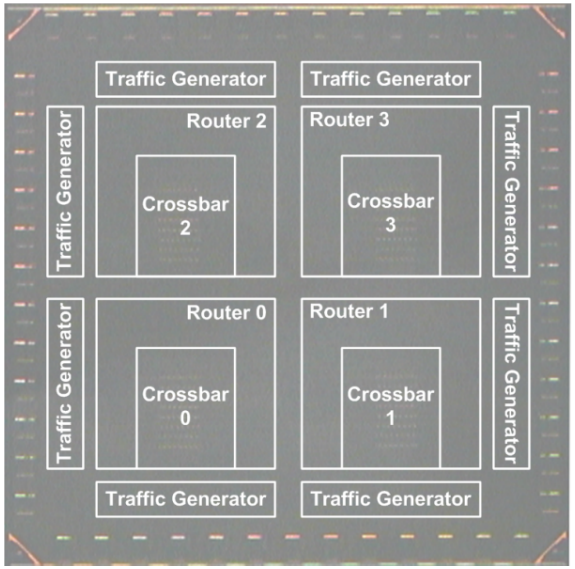
Low-Swing, Bitcell-Based Crossbar

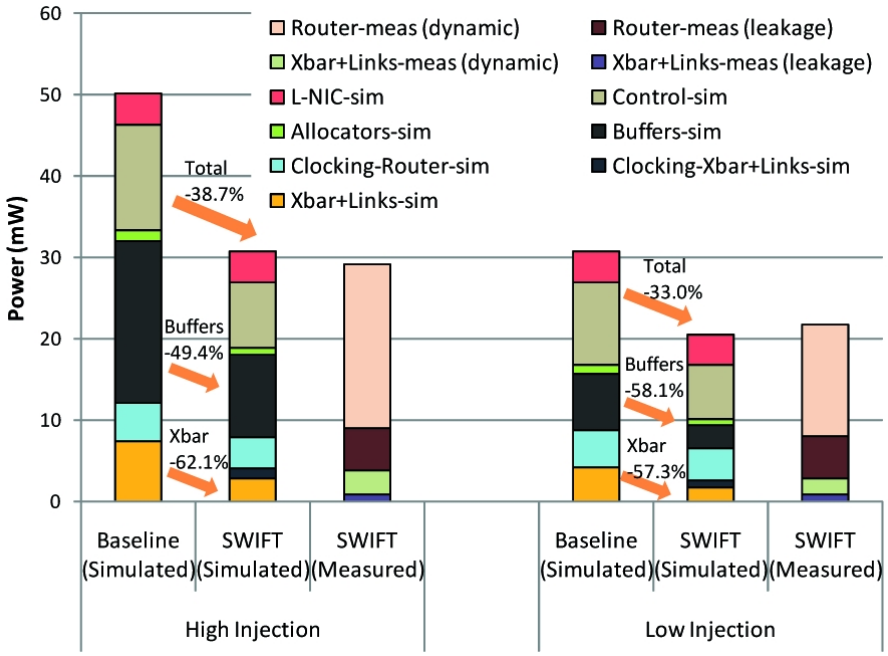






Measurement Summary





Technology	8-Layer, 90nm CMOS
Supply Voltage	1.2V
Chip Size	4mm^2
Transistors – Chip	688k
Frequency	400MHz (low injection) 225MHz (high injection)
Measured Energy/bit (1mm) (signal + clock)	64fJ/bit
Network Latency Reduction*	39%
Power Improvement*	38% Network Total 53% Data Path
Throughput Improvement*	15%

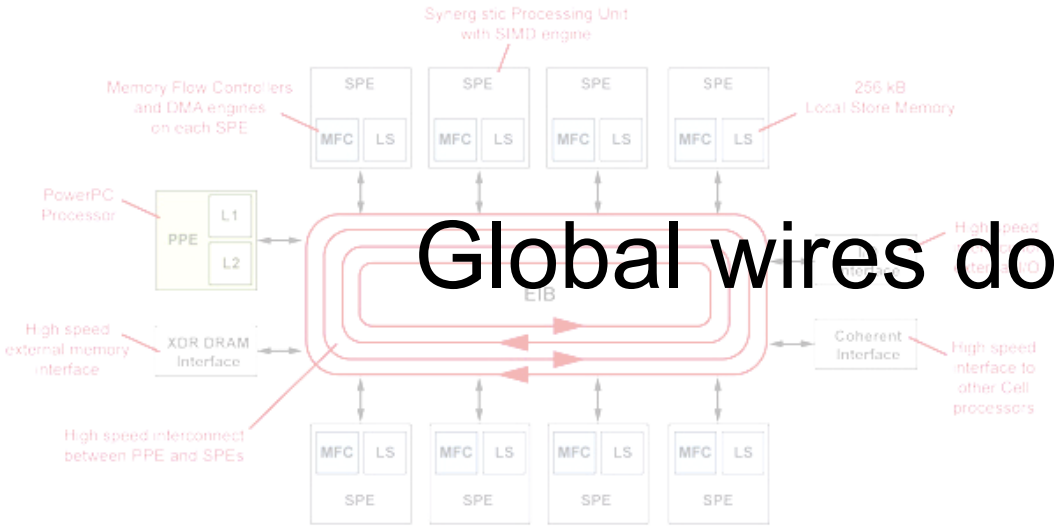
*Relative to a baseline synthesized VC NoC router

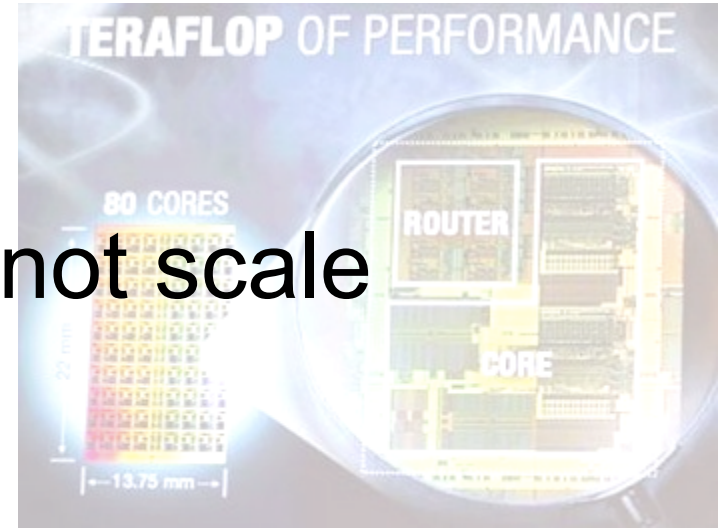
NoC with Near-Ideal Express Virtual Channels Using Global-Line Communication

Tushar Krishna[1], Amit Kumar[1],
Patrick Chiang[2], Mattan Erez[3], Li-Shiuan Peh[1]

[1] Princeton University
[2] Oregon State University
[3]University of Texas, Austin

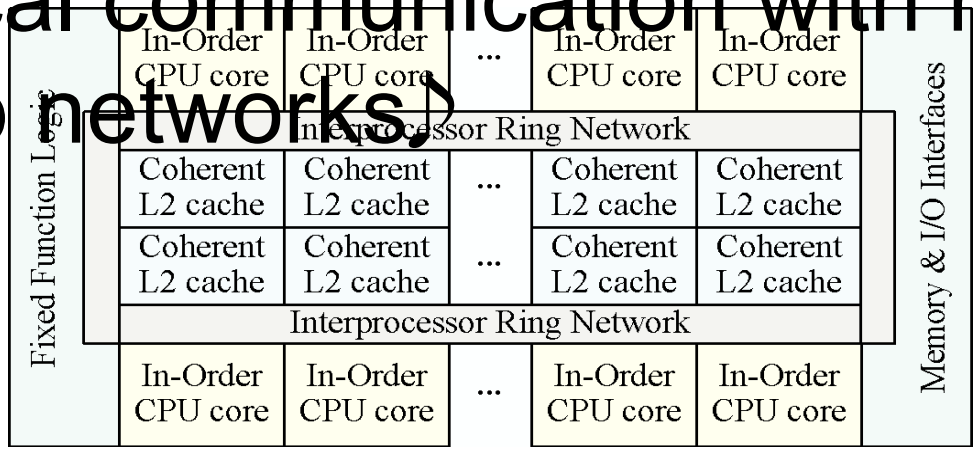
The CMP era…





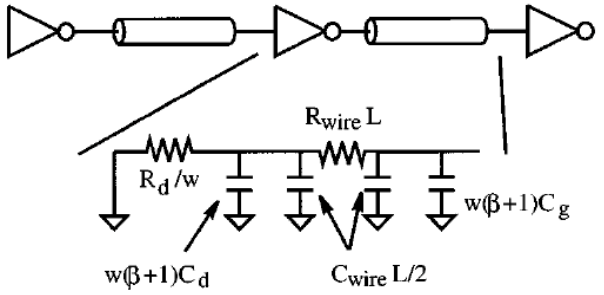
Global wires do not scale

Local communication with multi-hop networks☺



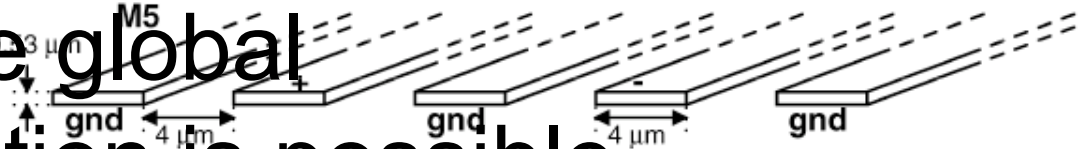
Not all interconnects are equal

- Conventional repeated RC wires
  - R. Ho [2001]
  - Latency several clock cycles across a chip (~3ns / 10mm)
  - High BW for short lengths

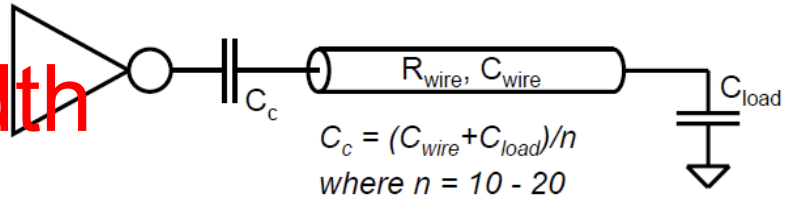


- On-Die Transmission Lines
  - K. Shepard[05,06], Ito[08]
  - Speed of light propagation (~100ps / 10mm)
  - Power and bandwidth density is poor
- Current Sensing/ Capacitive feed-forward
  - R. Ho [07], E. Mensink [07]
  - 5-10x improvement vs. conventional RC wire (~500ps / 10mm)
  - BW density 2–4X lower (vs. short wires)♪

Single-cycle global communication is possible



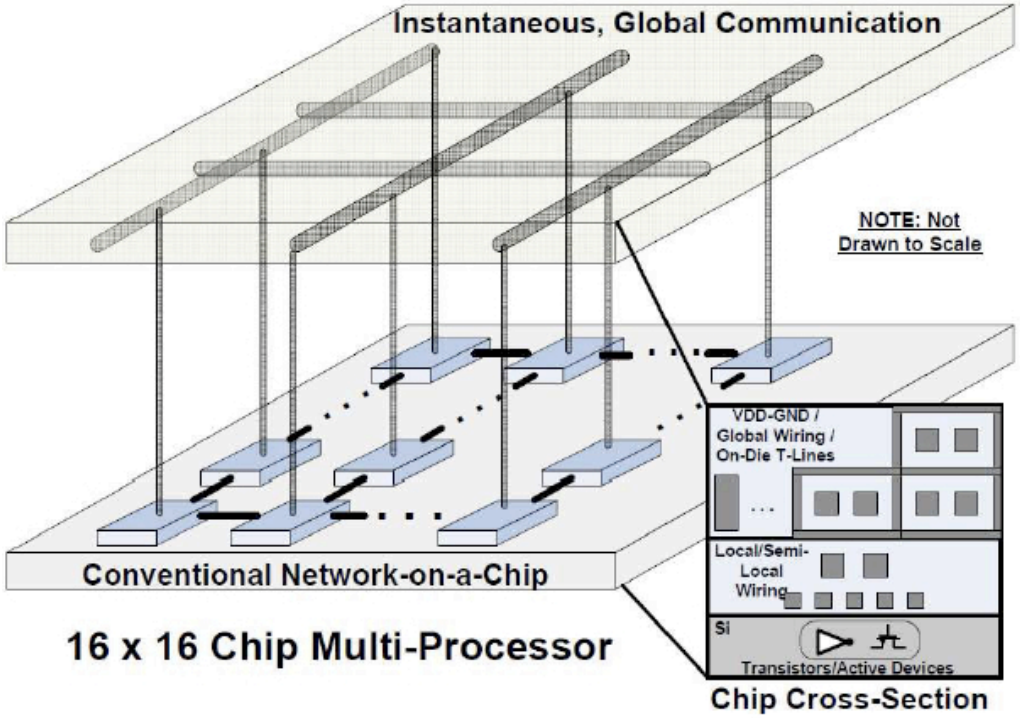
Trade-off latency vs. bandwidth



Not all interconnects are created equal

- Heterogeneous interconnect:
  - Inverters good for short-range and density
  - Transmission lines good for latency and broadcast

- How to build a CMP architecture that exploits different interconnect properties?



Various Interconnect Properties

Table 2: Interconnect Characterization

Process	Energy* (pJ)	Bit rate (Gbps)	Line Width (μm)	Density (Gbps/μm)	Density (Gbps/μm^2)	Latency* (ps)	Transceiver Area (μm^2)	Reference
Conventional RC								
180nm, 1.8V	11	1	1	1	NA	1000	NA	[11]
130nm, 1V	1.84	2.5	1.2	2.08	NA	444	NA	[8]
Low Swing Copper								
180nm, 1.8V	1.05	1	1.8	0.56	NA	NA	NA	[11]
90nm, 1.2V	0.28	2	1.4	1.42	5.28E-003	500	379	[24]
180nm, 1.8V	0.6	3	0.72	4.17	3.38E-003	333	887	[3]
130nm, 1.2V	2	3	1.2	2.5	2.31E-003	440	1300	[31]
90nm, 1.2V	0.36	4	2	2	2.27E-003	250	1760	[18]
Transmission Line								
180nm, 1.8V	2	3	24	0.13	NA	121	NA	[15]
130nm, 1V	0.42	40	13	3.08	NA	NA	NA	[8]
90nm, 1V	1.05	8	14	0.57	4.00E-003	300	2000	[13]
Optical								
65nm	0.55	1280	4	320	NA	NA	NA	[4]
90nm, 1.0V	8.06	16	Off Chip	Off Chip	1.52E-004	NA	105000	[28]
Radio Frequency on Transmission Line								
IBM 90nm	1.5	30	12	2.5	5.33E-003	100	5625	[5]

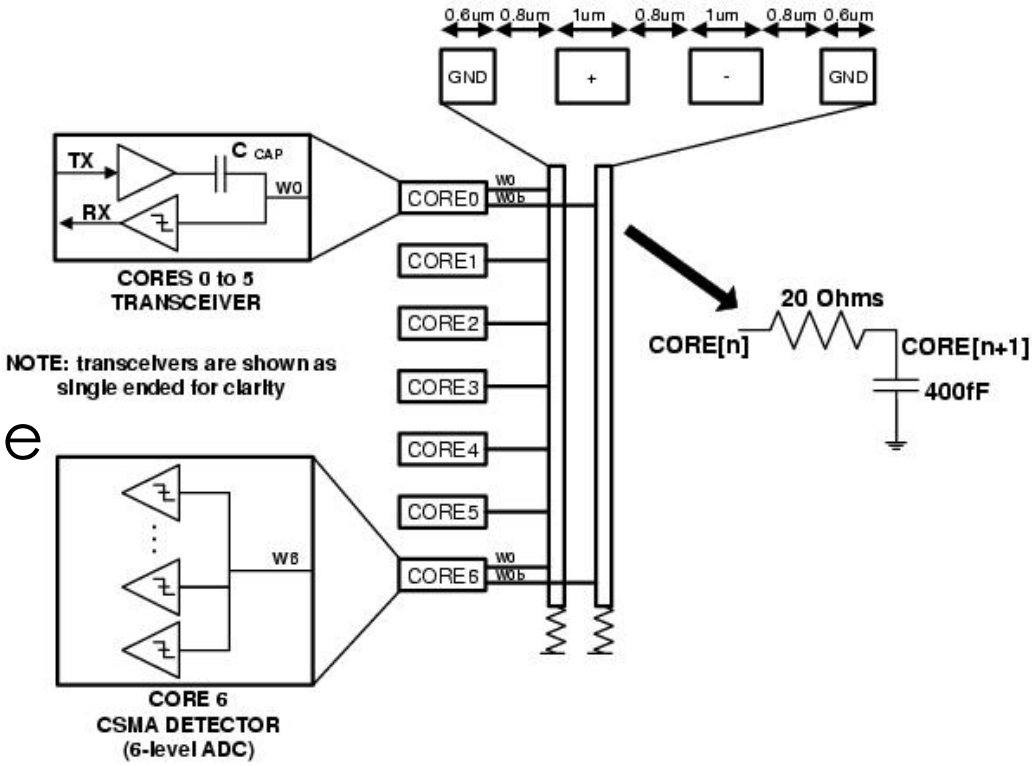
* Normalized to 10mm NA: Not Available

What if single cycle global communication is possible?

- Network-on-chip with hybrid interconnect
  - Data plane
    - Multi-hop network
    - High bandwidth
    - Full-swing
  - Control plane
    - Global lines (G-lines)
    - Ultra-low latency
    - Multi-drop
- Express virtual channels (EVCs) that rely on NOCHI
  - Critical flow control information is shared among routers using G-lines
  - Reduced buffering and power overhead in routers

S-CSMA Circuit Design

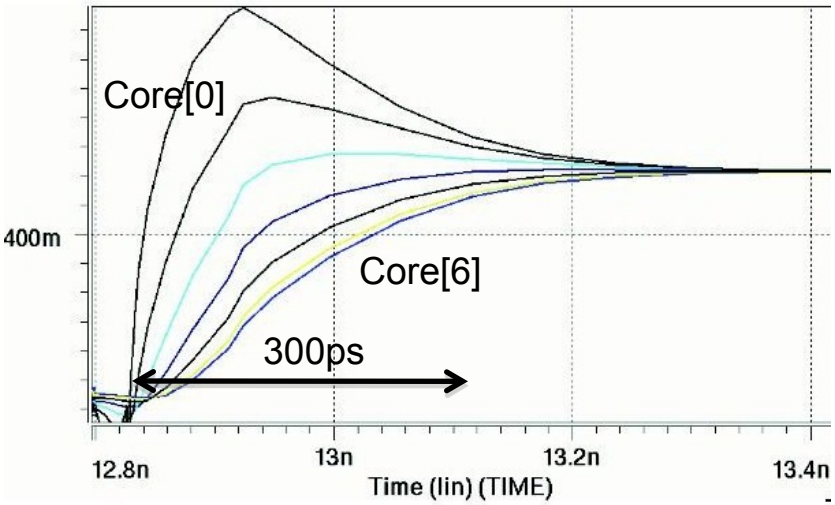
- 7 cores traversing 7mm
- Each TX using C_{CAP}=300fF
- Each Core RX with sense amplifier converting to digital



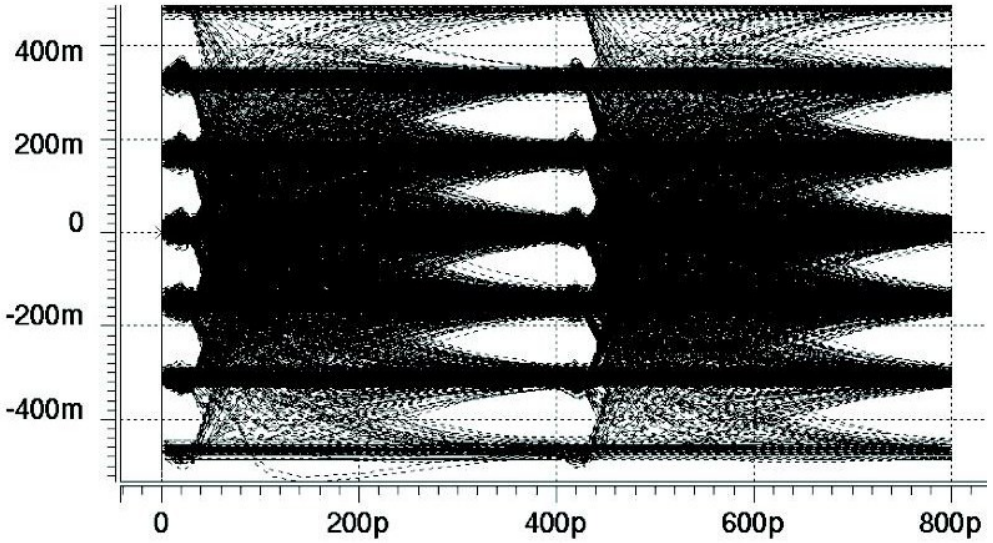
- Last RX uses amplitude detection to determine # of TXs transmitting concurrently
  - Uses Flash ADC (6 sense amplifiers with different ref. voltage)

Simulation Results

- Circuit simulation using 7 metal, 90nm 1.2V, CMOS process
- Pulse response along G-lines (up to 6 cores)

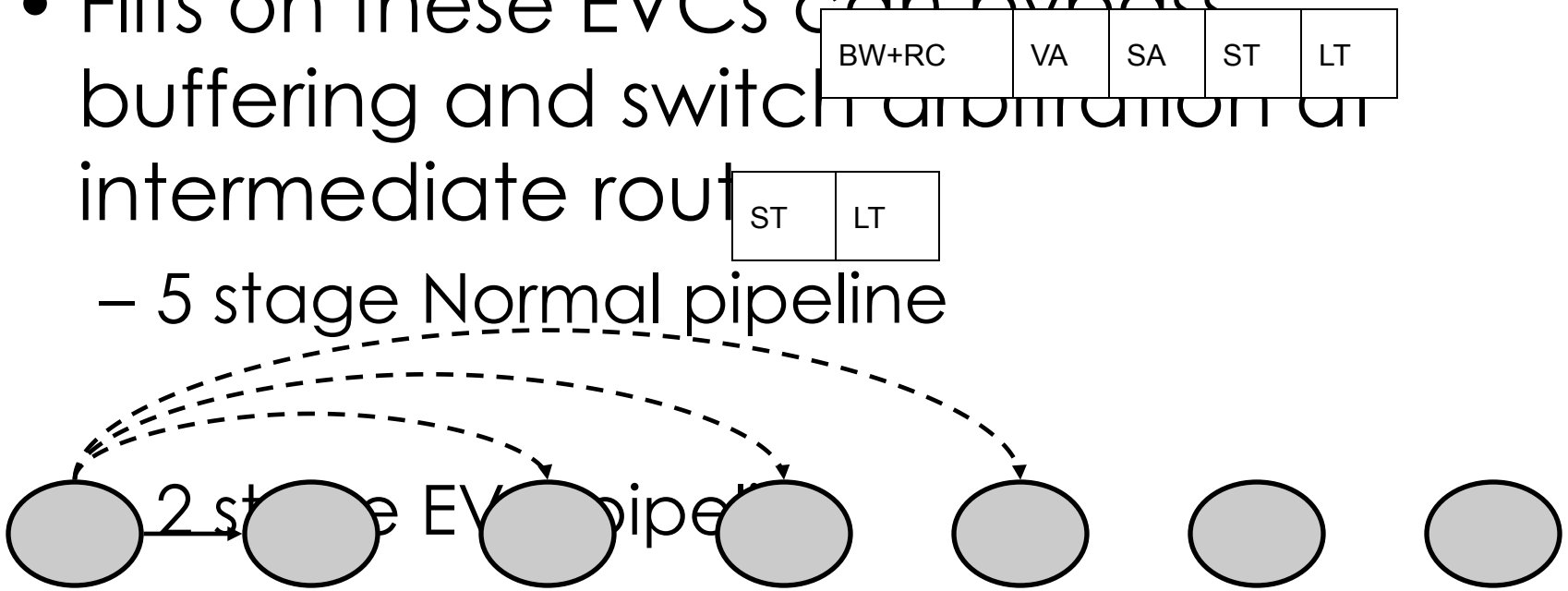


- Eye diagram at Flash ADC input



Express Virtual Channels*

- Virtual *express* lanes in the network
- Flits on these EVCs can bypass buffering and switch arbitration at intermediate routers
  - 5 stage Normal pipeline

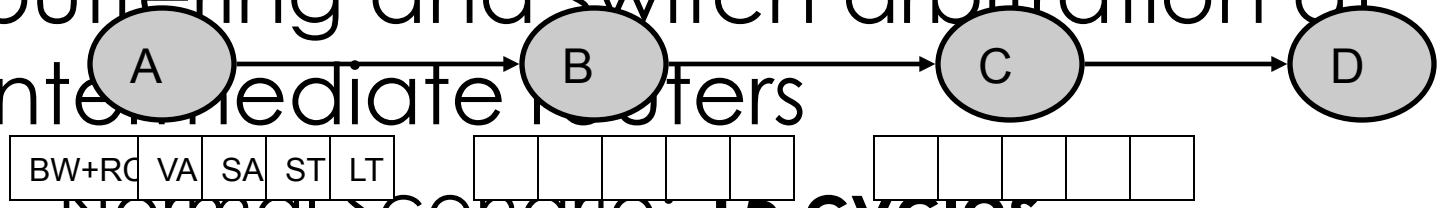


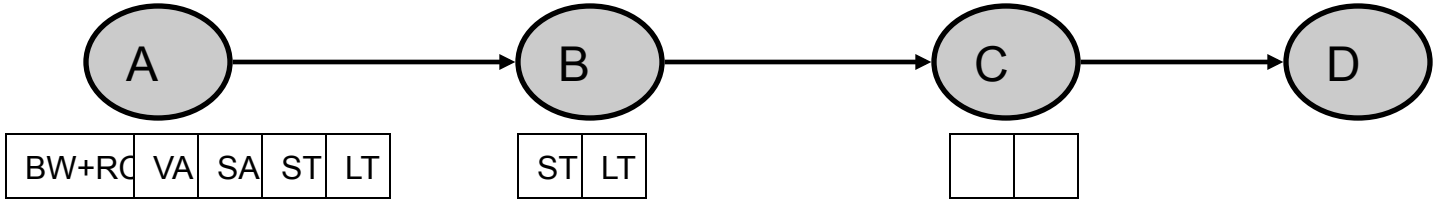
  - 2 stage EVC pipeline

* "Express Virtual Channels: Towards the Ideal Interconnection Fabric", Amit Kumar, Li-Shiuan Peh, Partha Kundu and Niraj K. Jha, *Proc. of the 34th International Symposium on Computer Architecture (ISCA), June 2007.*

Express Virtual Channels (EVCs)*

- Virtual *express* lanes in the network
- Flits on these EVCs can bypass buffering and switch arbitration at intermediate routers
  - Normal scenario: **15 cycles**

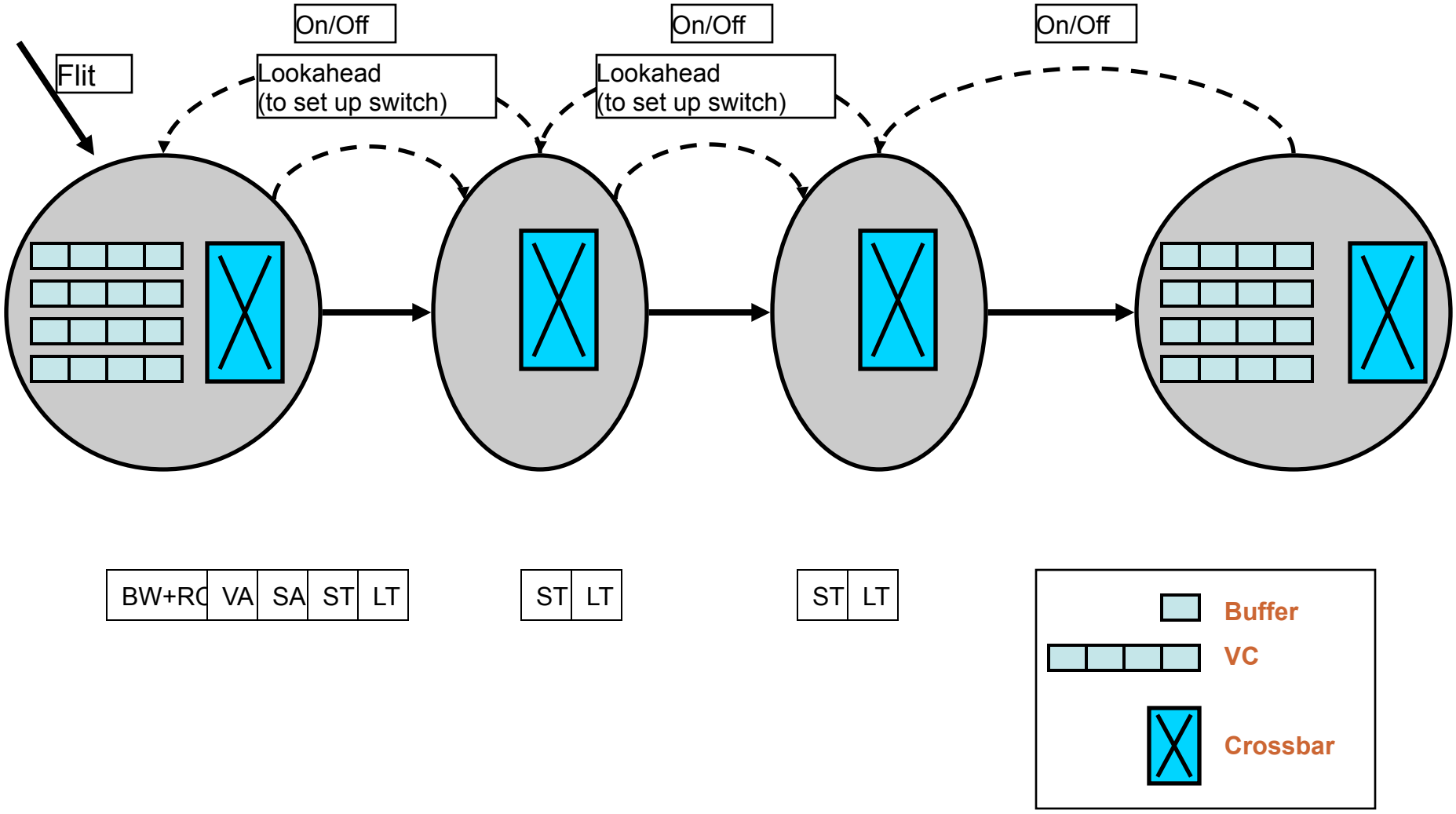




  - Express virtual channel (EVC): **9 cycles**

* "Express Virtual Channels: Towards the Ideal Interconnection Fabric," Amit Kumar, Li-Shiuan Peh, Partha Kundu and Niraj K. Jha, *Proc. of the 34th International Symposium on Computer Architecture (ISCA), June 2007.*

EVCs in Action

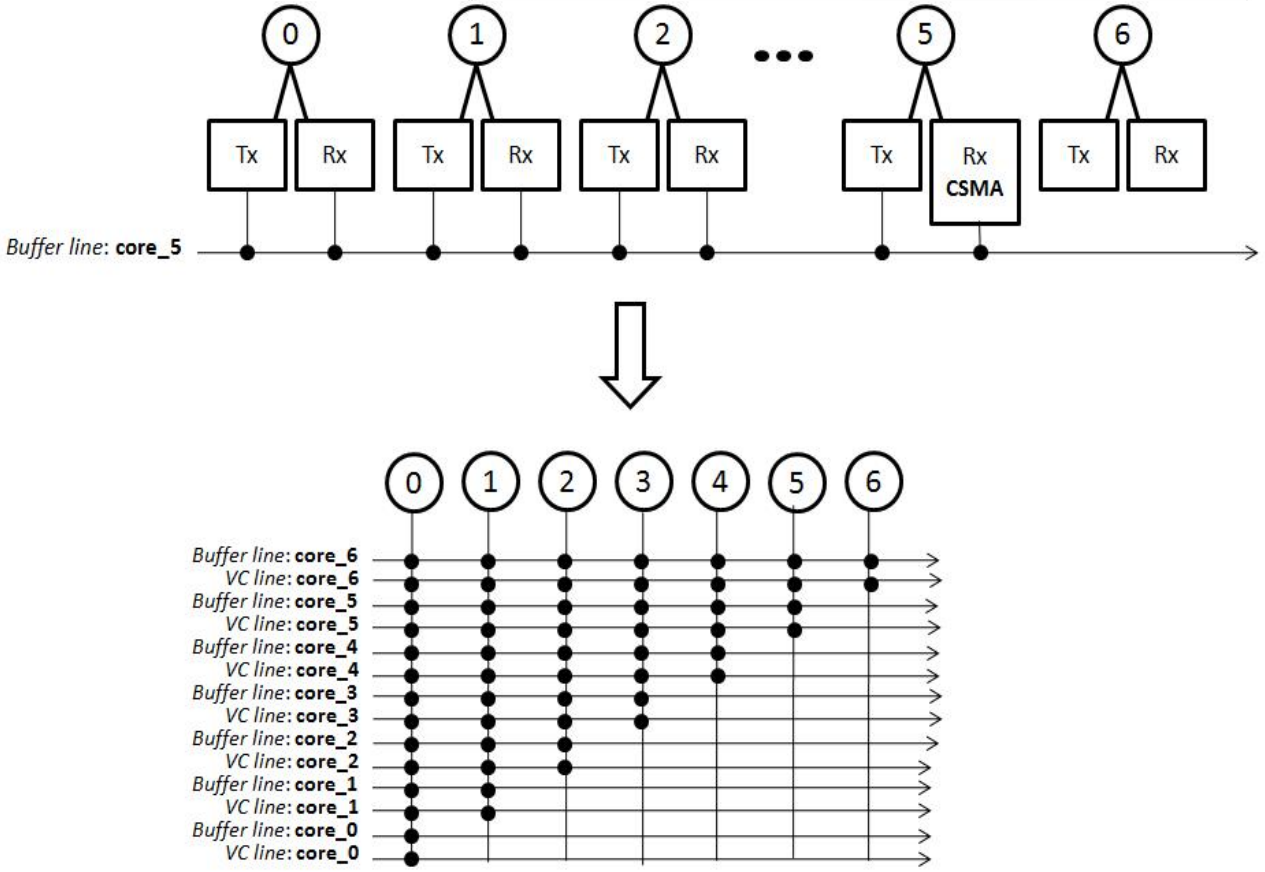


G-line EVCs

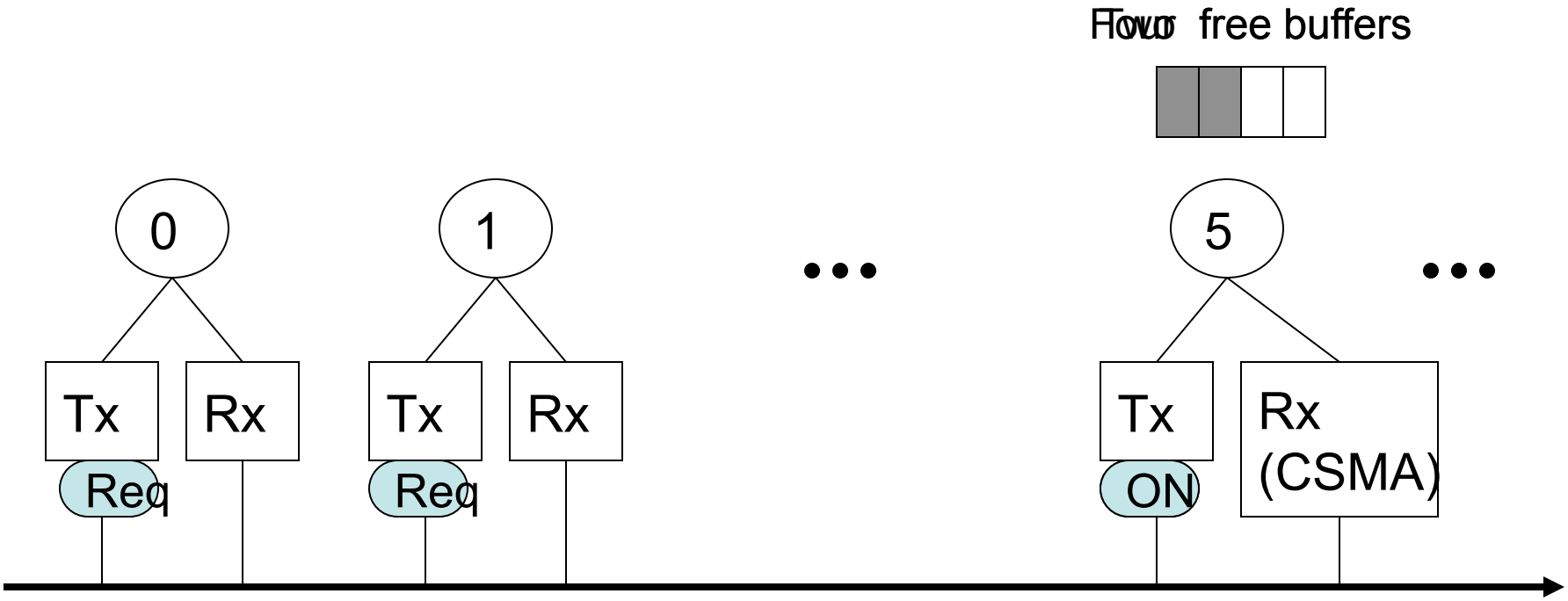
- Use G-lines for global flow-control information
  - G-line control plane and conventional data plane
- Down-stream node broadcasts buffer/VC availability
  - Single cycle to *all* up-stream nodes
  - All nodes have up-to-date information
- Up-stream nodes request resources in 1 cycle
  - Each down-stream node has 1 buffer-request

Setup

- 7x7 packet-switched mesh network
- Two G-lines per node & direction for buffer/VC request



The signaling mechanism♪



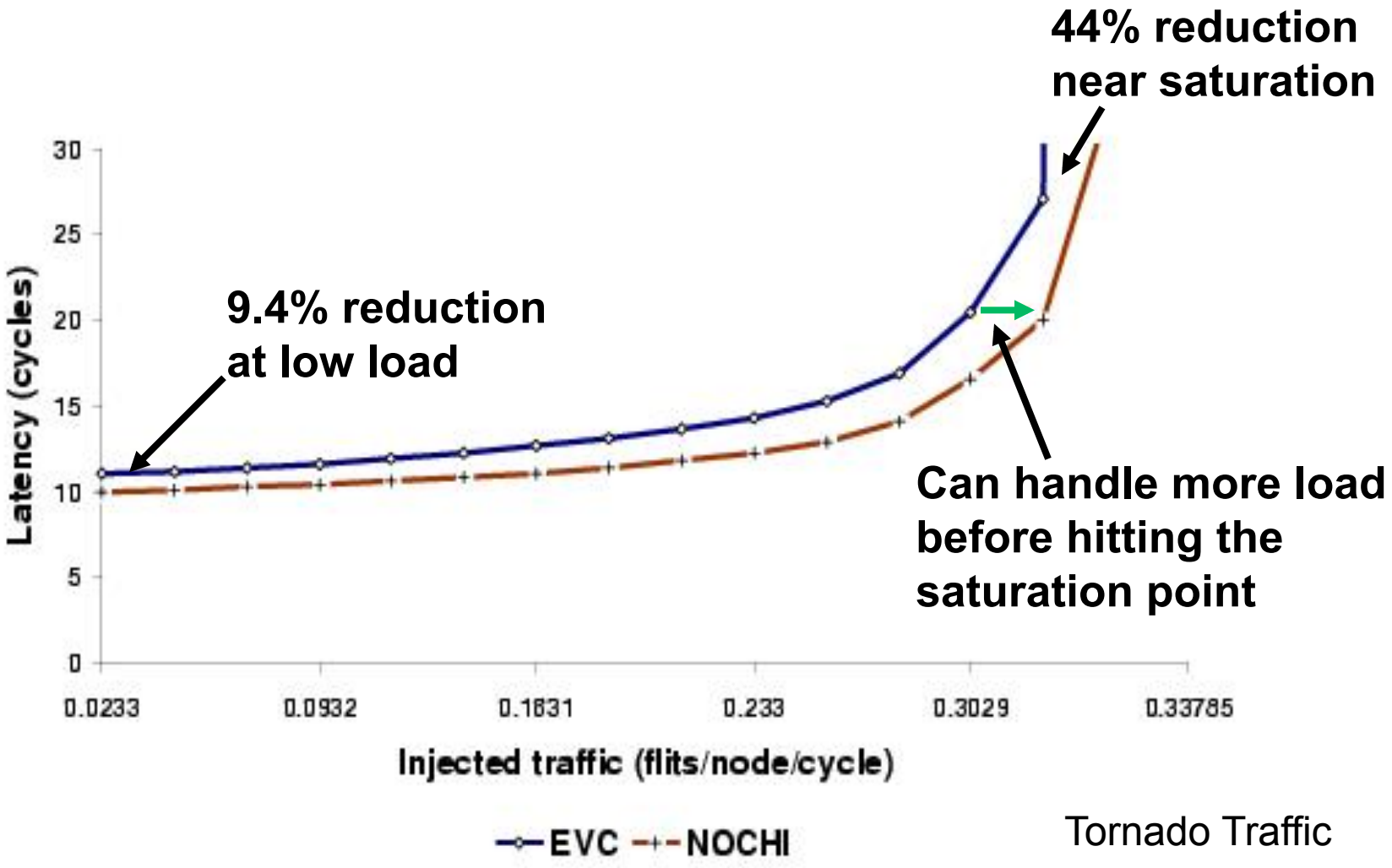
Number of granted requests are sent to the upstream on the data plane

Benefits of G-line EVCs

- Instantaneous global information
  - Aggressive buffer management
  - Original EVC reserved buffers for signal traversal time
- Broadcast medium
  - Enables flexible, dynamic EVCs of any length
  - Original EVC limited by signaling cost
    - Partition VCs into k-hop bins
    - Limits EVCs to short lengths (< 3 hops)

Network evaluation (1)

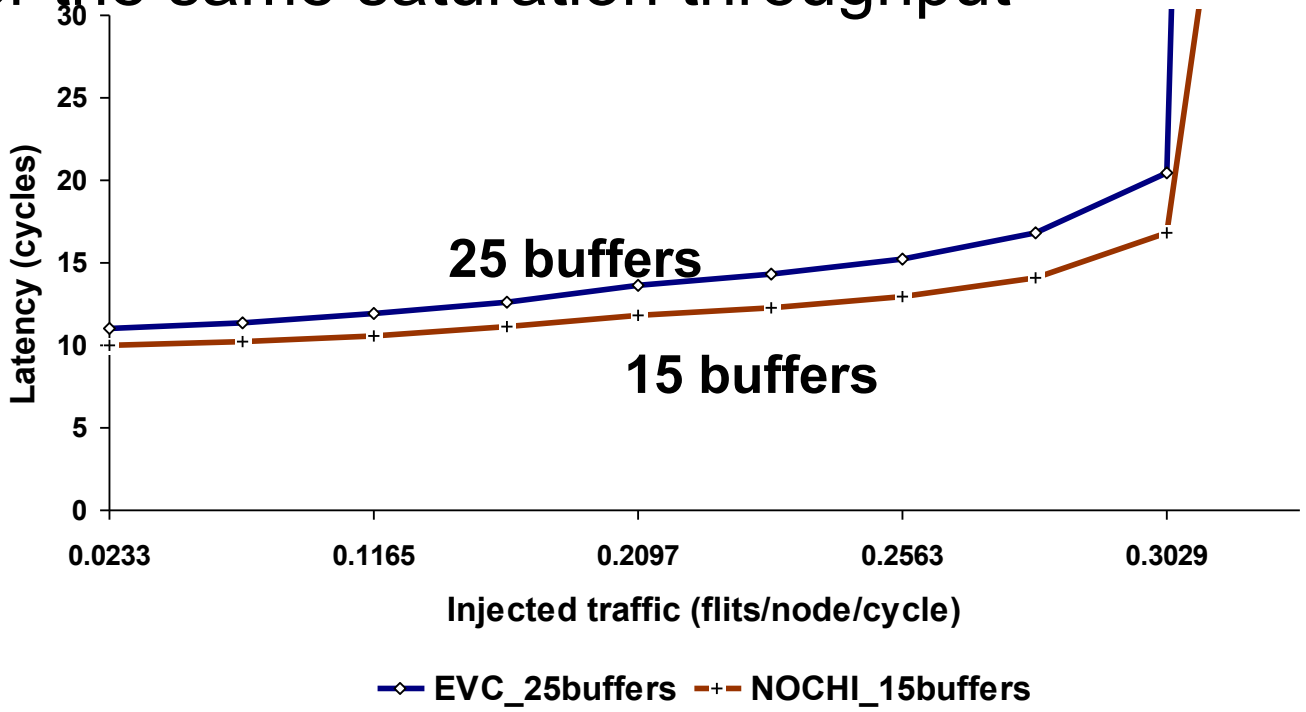
- For the same number of buffers per port



Tornado Traffic

Network evaluation (2)

- For the same saturation throughput



Network power for original EVC	Network power for G-line EVC	Net power reduction
47.99W	43.76W	4.23 W (**8.8%**)

* Power numbers based on extrapolation of Intel Polaris 80-core network

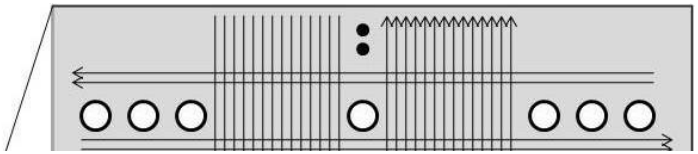
Conclusions

- Effective long-range communication is possible
  - 1 cycle cross-chip with reasonable power
  - Lower overall bandwidth
- NOCHI utilizes low-latency control plane and high-bandwidth data plane
  - Single cycle, multi-drop, broadcast for control
  - Full-swing multi-hop network for data
- The advantages of EVCs (latency and

Thanks

Backup

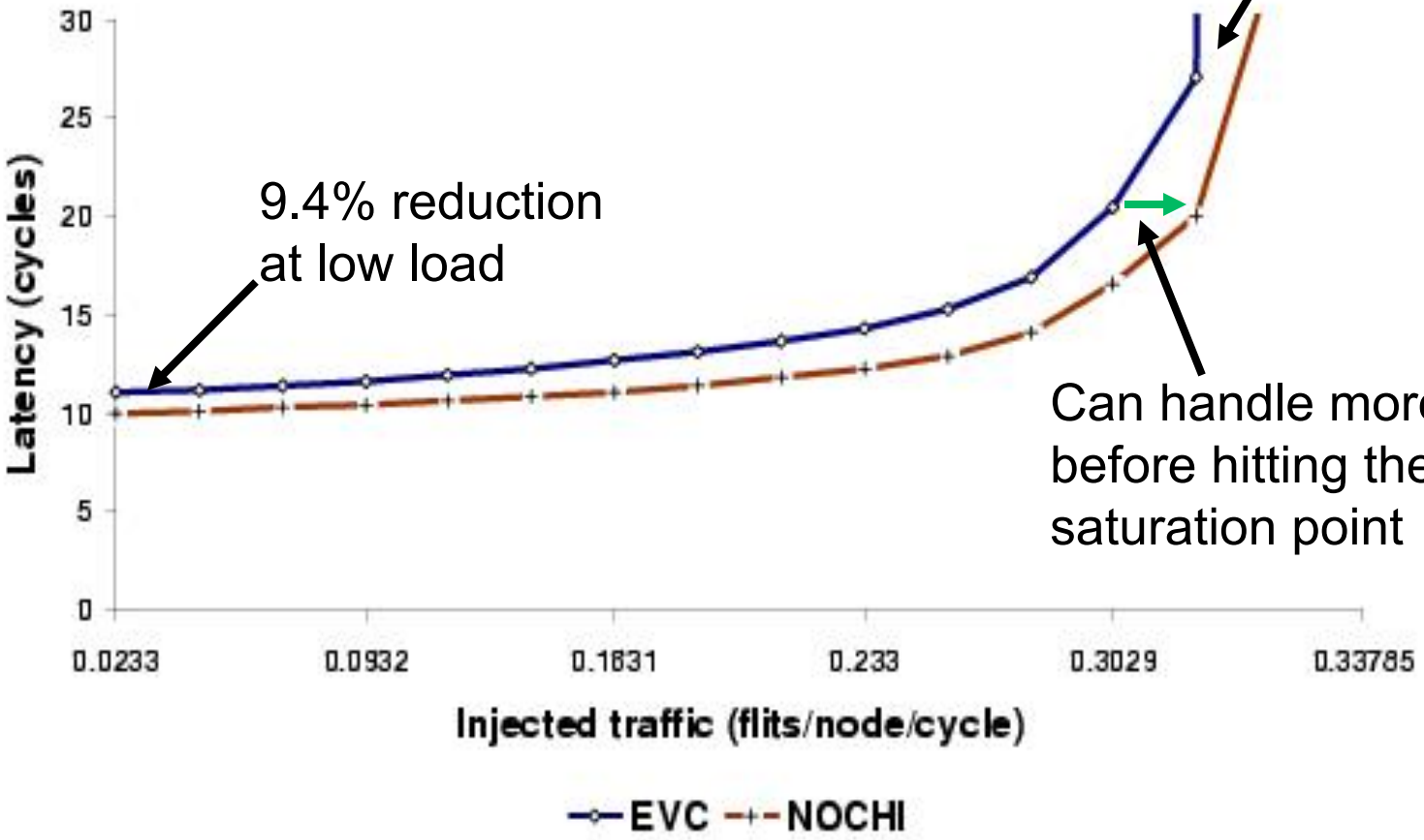
NOCHI: Network-on-chip with hybrid Interconnect



44% reduction
near saturation

9.4% reduction
at low load

Can handle more load
before hitting the
saturation point



IEEE Micro 2009

Overview

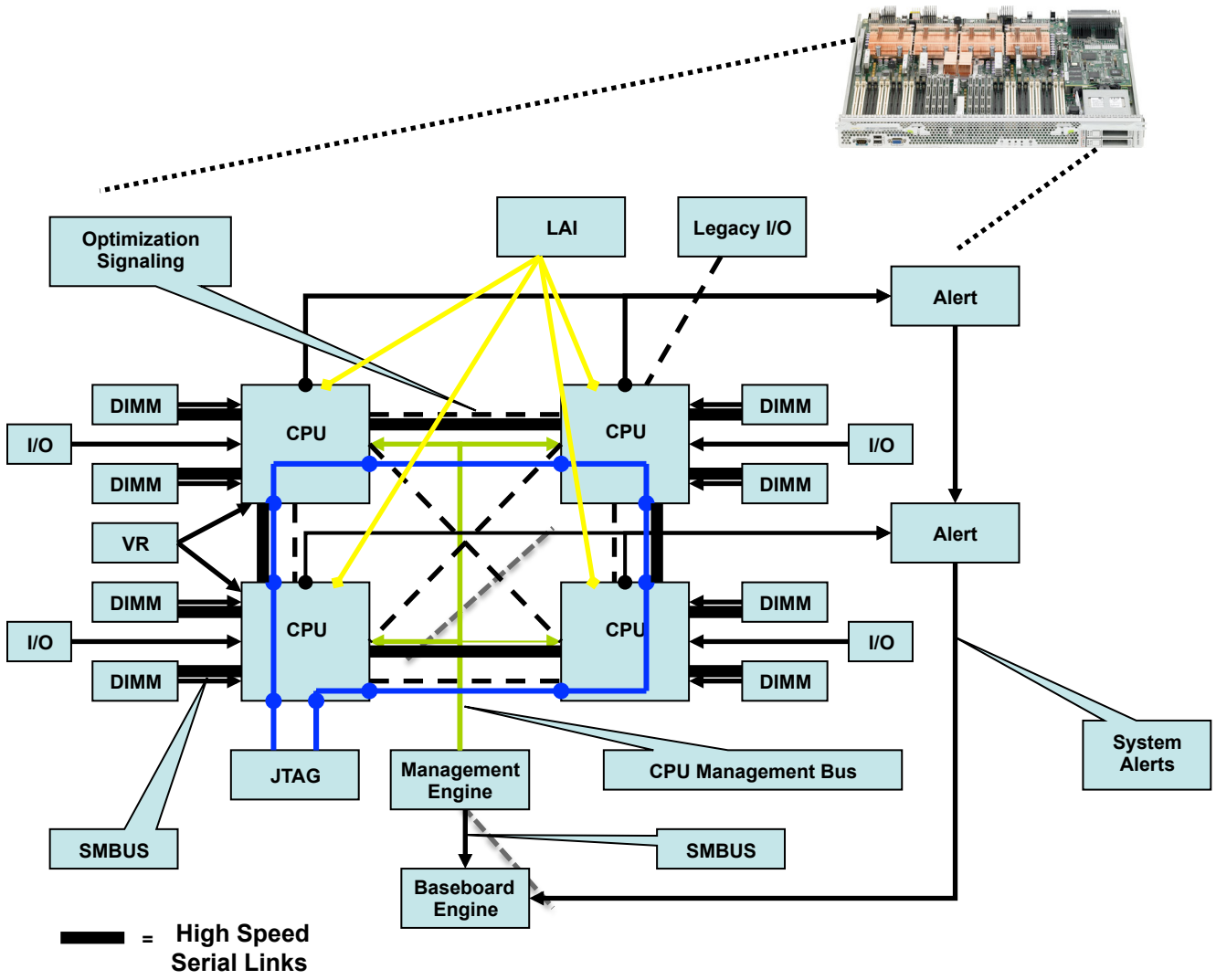
1. Energy-Efficient On-Chip I/O
   - Network-on-a-Chip with Reduced-Swing Interconnect
   - Fundamental limits to low-voltage swing
   - Heterogeneous Interconnect Architectures

2. Energy-Efficient Off-Chip I/O
   - Sub-1mW/Gbps Off-Chip I/O

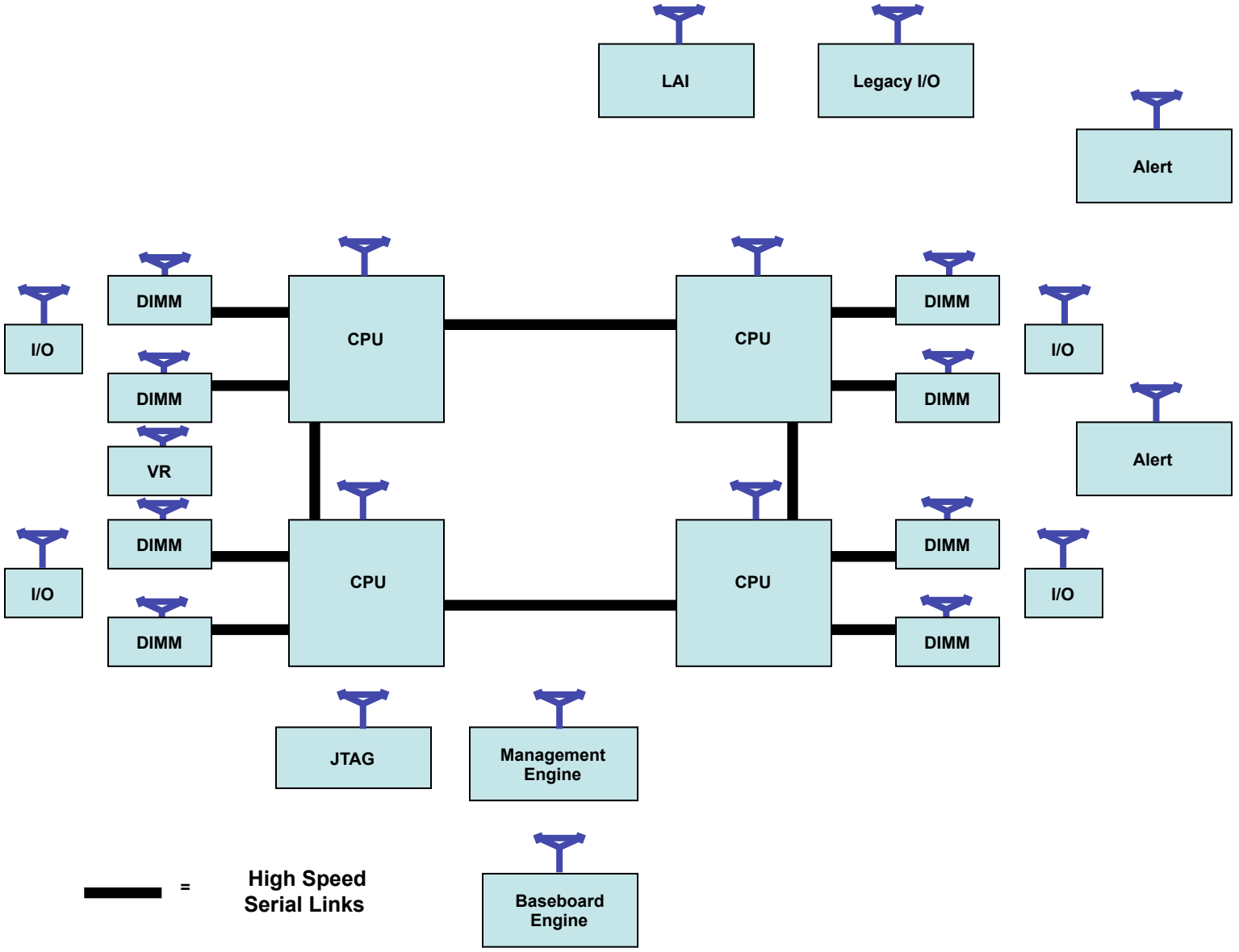
3. **Short-range, UWB Transceivers**

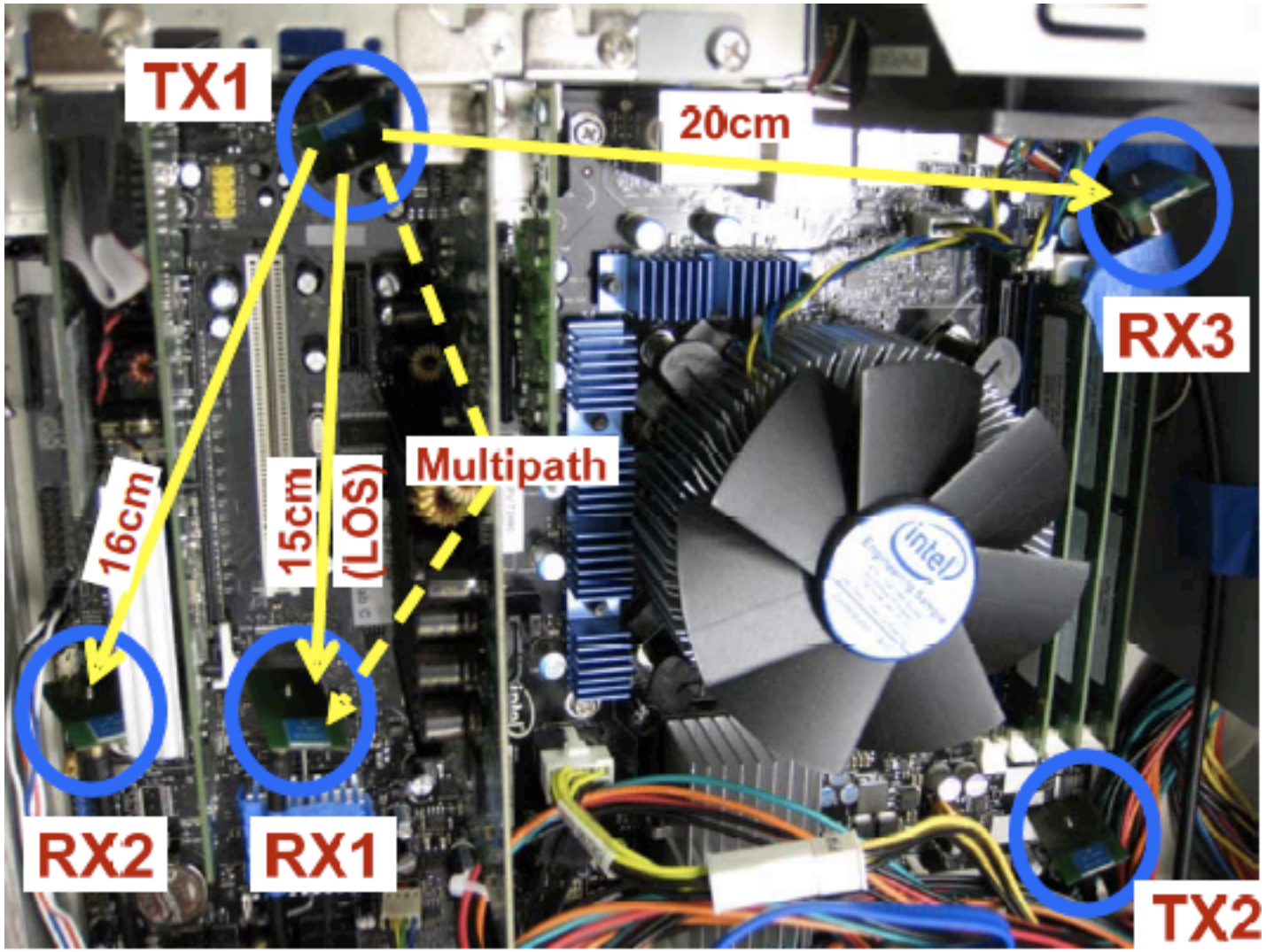
(Time): Near-threshold process variation tolerance

Conventional, 4-Socket Intel Blade Server

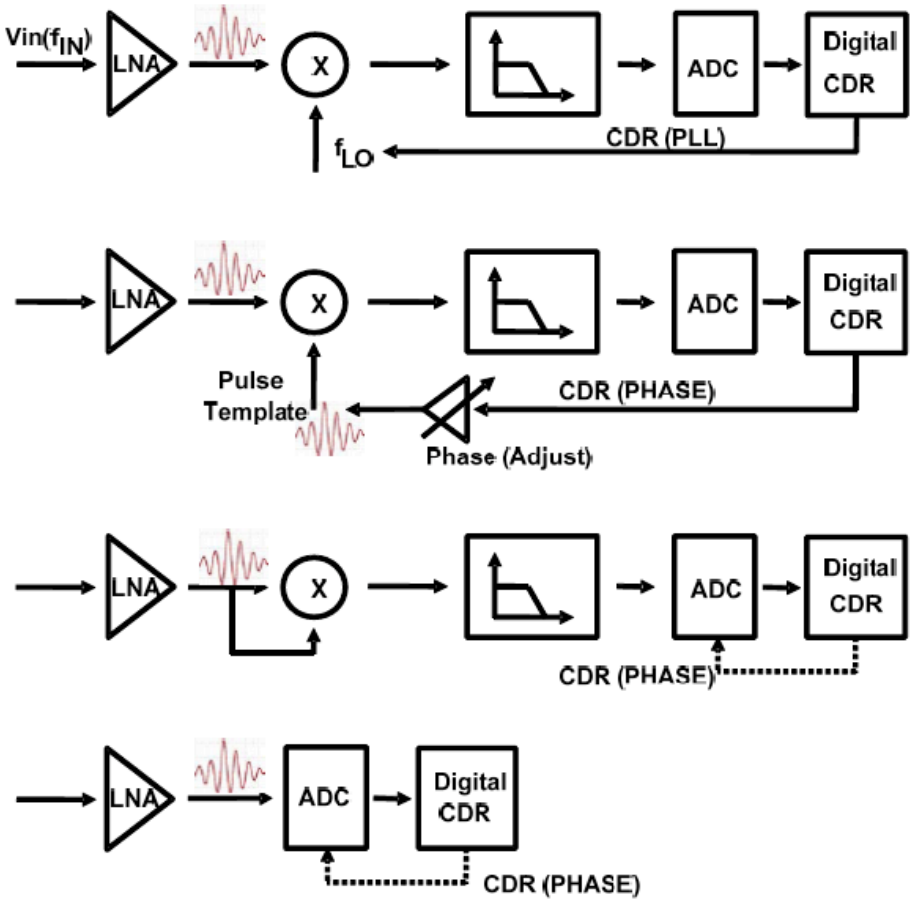


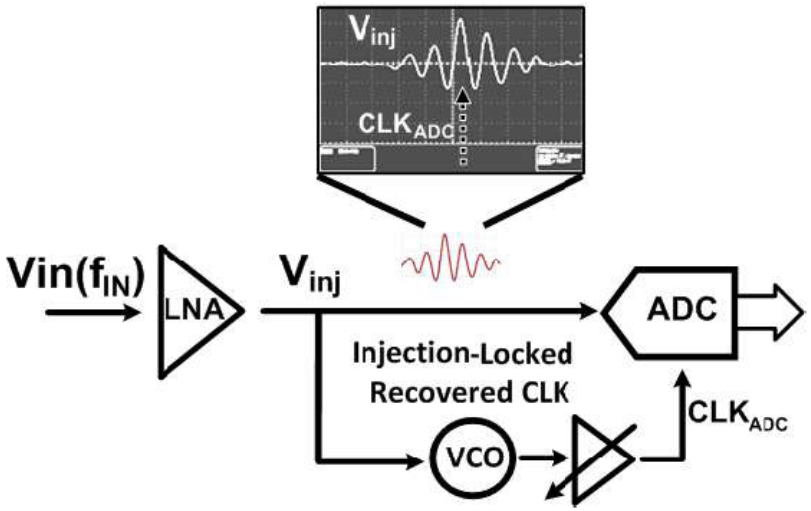
Proposed, Wireless Interconnect within Chassis

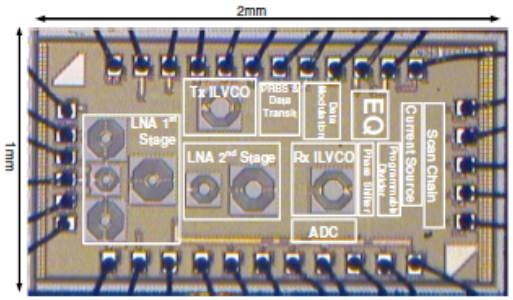



TX1
20cm
RX3
Multipath
16cm
15cm
(LOS)
RX2
RX1
TX2

A Fully-Integrated IR-UWB Transceiver Using Pulse Injection-Locking for Receiver Phase Synchronization

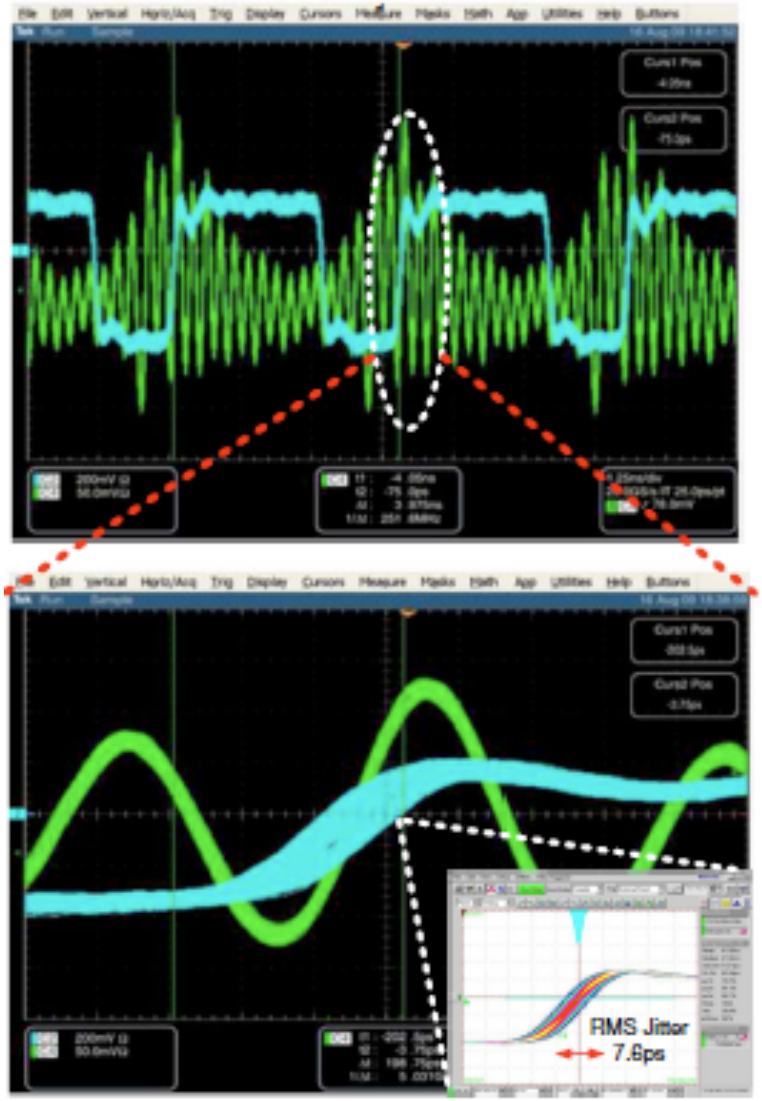


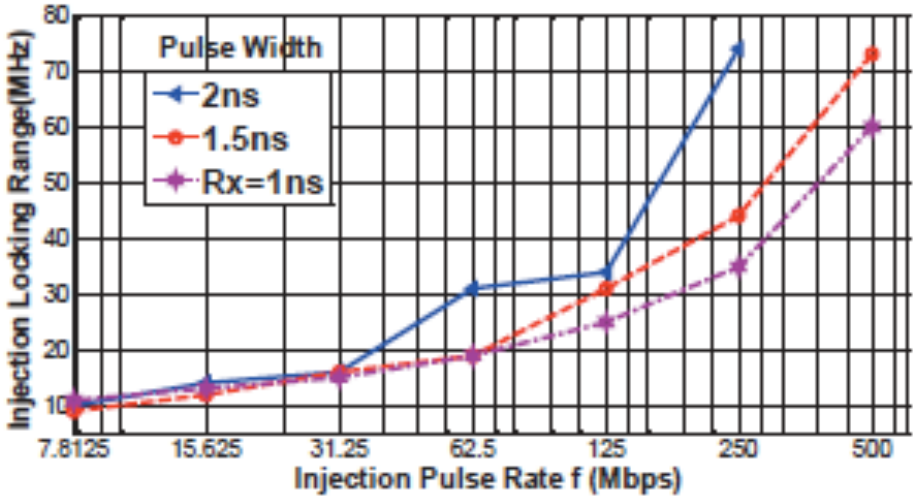




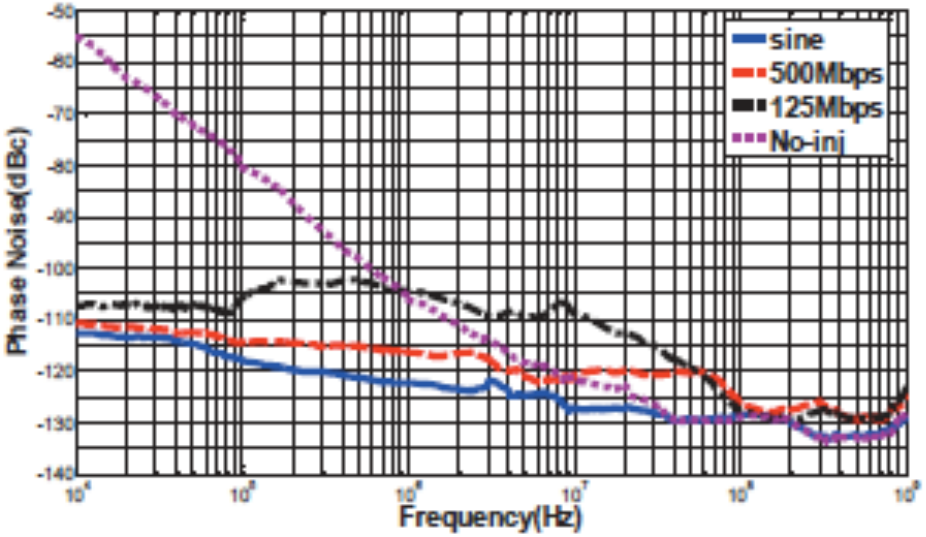
RFIC2010, JSSC2011

Measurement Results of Pulse Injection-Locking

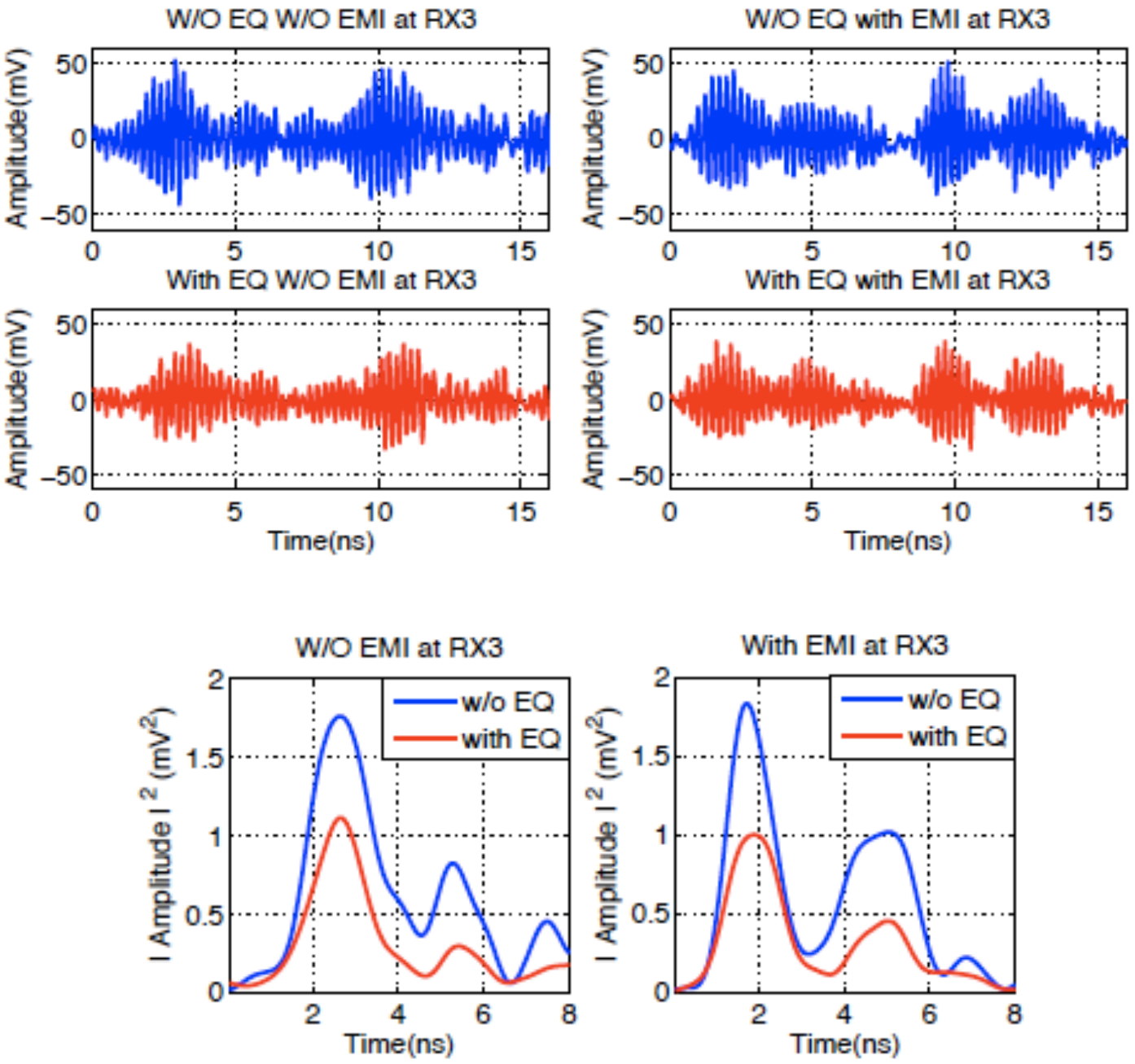




(a) Pulse injection lock range vs. pulse width and pulse rate



(b) Pulse injection locked VCO phase noise vs. pulse rate


W/O EQ W/O EMI at RX3
W/O EQ with EMI at RX3
With EQ W/O EMI at RX3
With EQ with EMI at RX3
Amplitude(mV)
Time(ns)
W/O EMI at RX3
With EMI at RX3
w/o EQ
with EQ
I Amplitude I 2 (mV2)