Future of Interconnect Fabric —A Contrarian View

Shekhar Borkar

Intel Corp.

June 13, 2010

Outline

Evolution of interconnect fabric

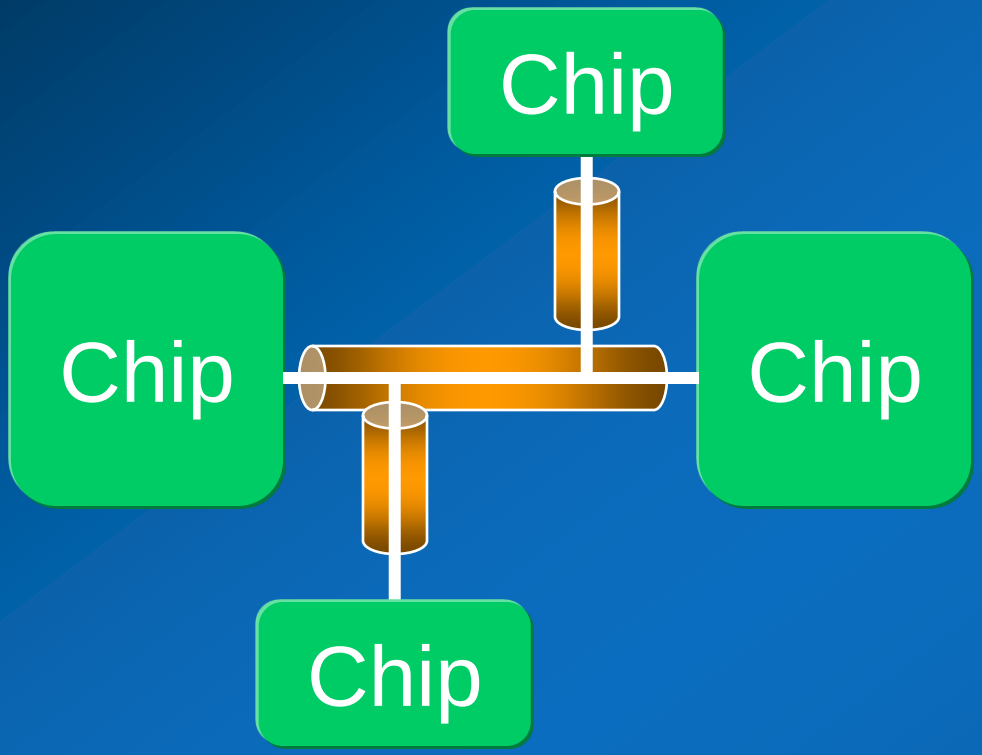
On die network challenges

Some simple contrarian proposals

Evaluation and benefits

Summary

Good-old Bus

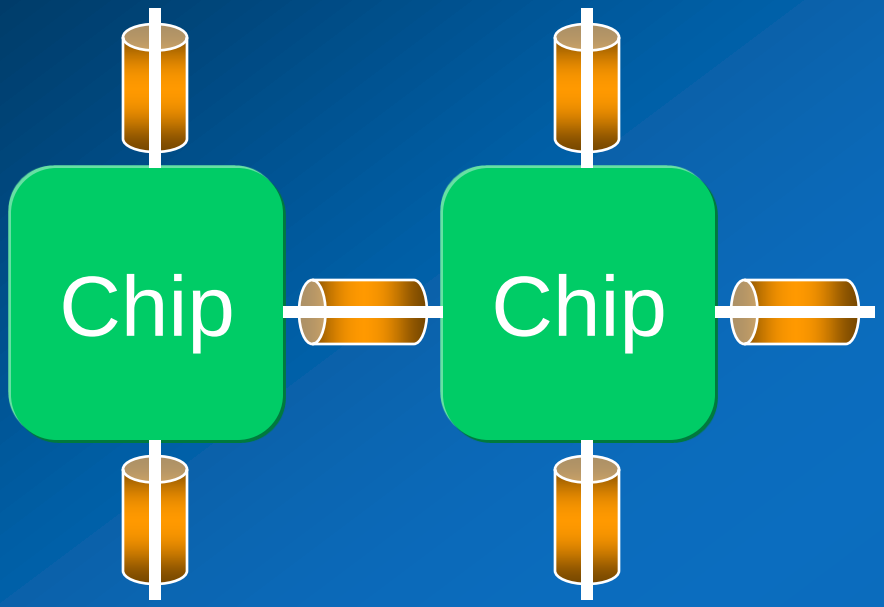


Good at board level, does not extend well

- Transmission lines
- Loss, signal integrity
- Need signal termination
- Limited frequency
- Width limited by pins, board area
- Signal processing power

Frequency	**100-1000 MHz**
Width	64-256 bit
Energy/bit	1 nJ
Data rate	1-10 GB/sec
Power	10+ W

Point-to-point Bus



Fast signaling, long distance

Between chips, boards, racks

High frequency, narrow links

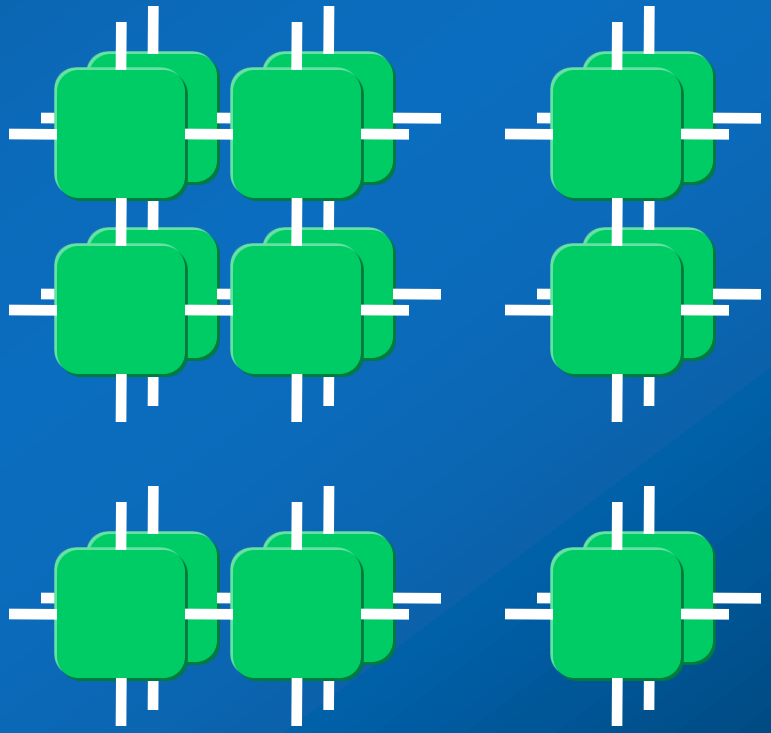
1D Ring, 2D Mesh and Torus to reduce latency

Higher logic complexity and latency in each node

Frequency	**200-3000 MHz**
Width	8-32 bit
Energy/bit	0.5 nJ
Bisection bandwidth	50+ GB/sec
Power	200+ W

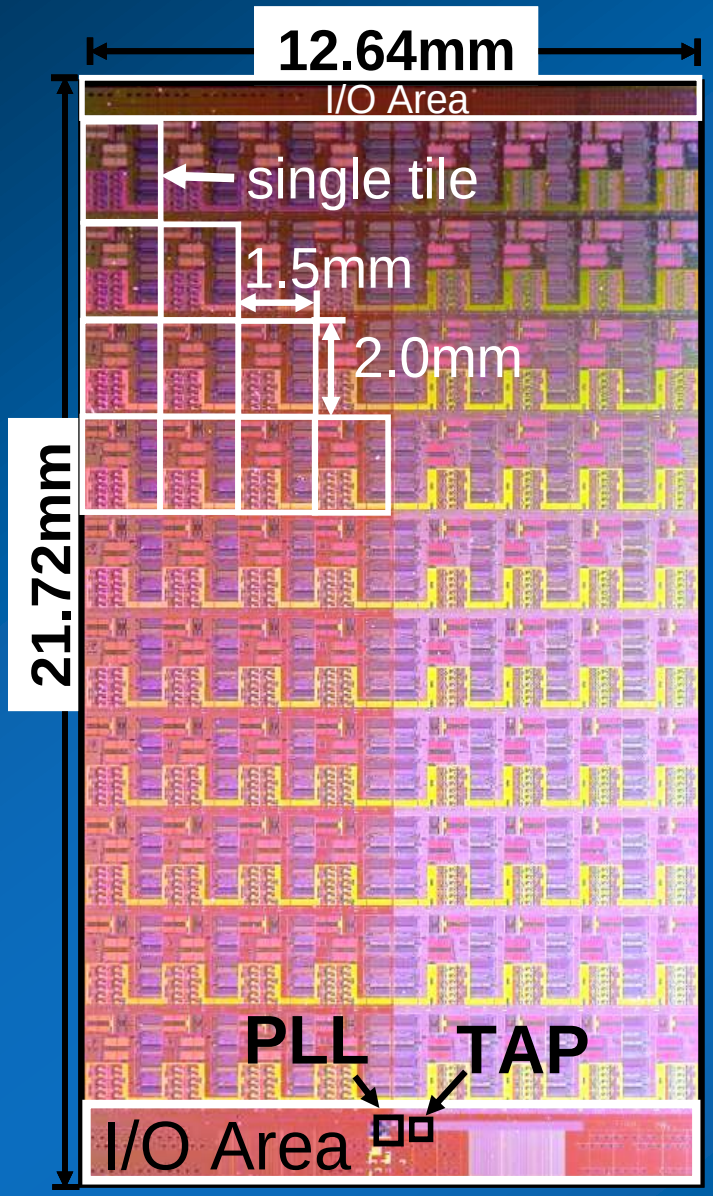
Emergence of packet switched network

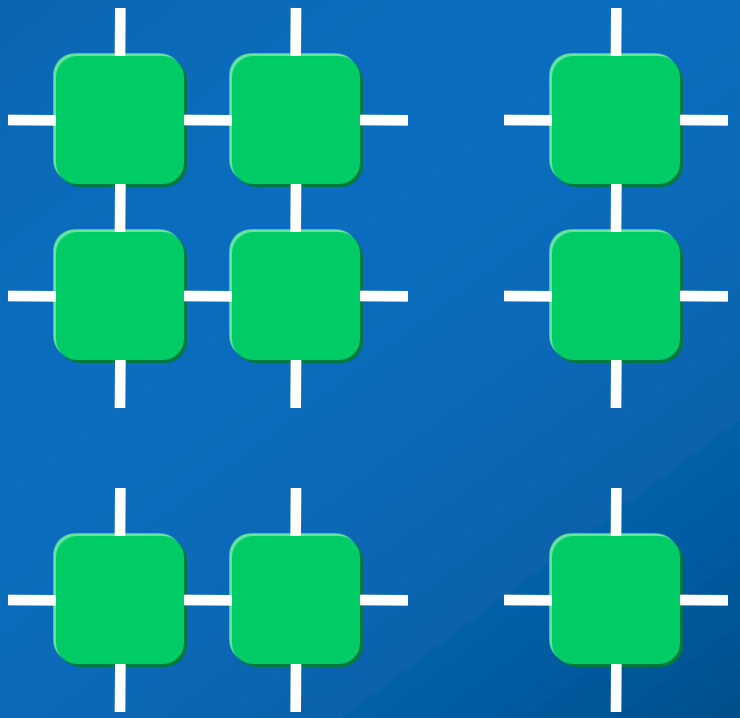
ASCI Red—The first TFLOP



32 X 32 X 2 Mesh Interconnect
Simultaneous bidirectional signaling
50 GB/sec bisection bandwidth

The 80-Core TFLOP Chip (2006)



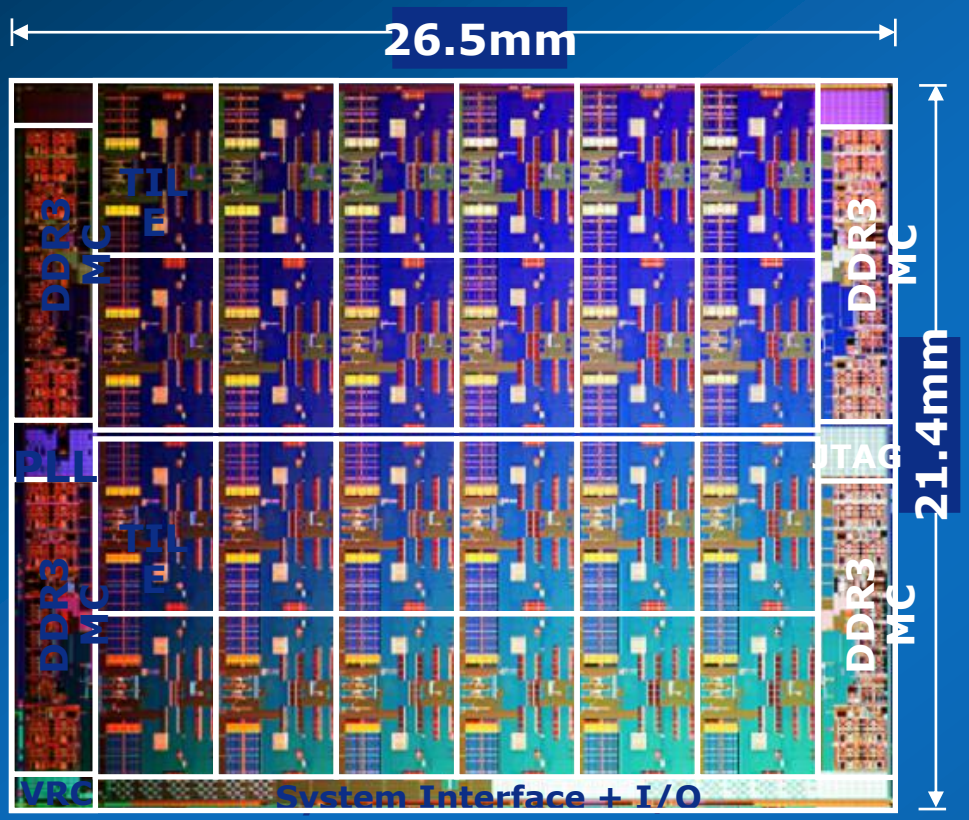


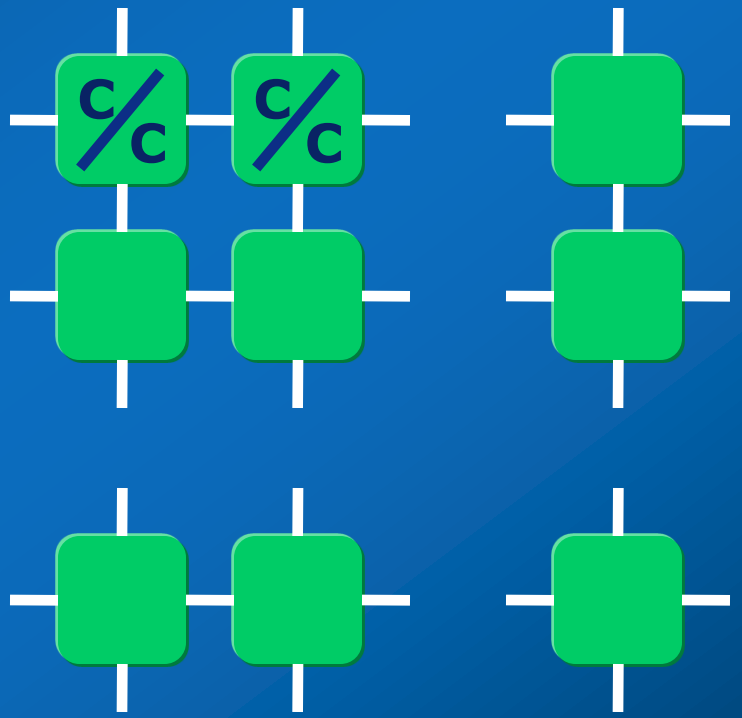
8 X 10 Mesh

32 bit links

320 GB/sec bisection BW @ 5 GHz

48 Core Single Chip Cloud (2009)





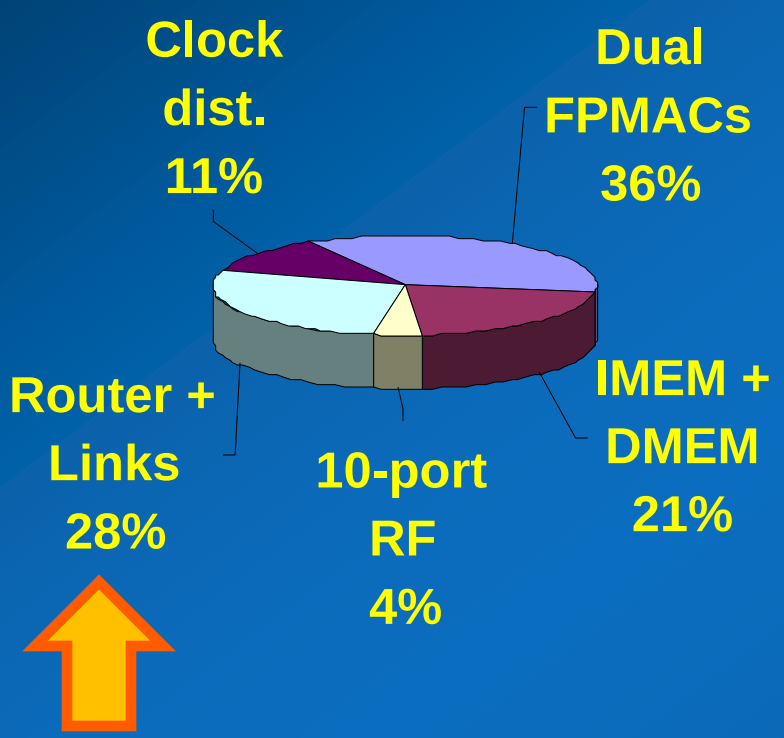
2 Core clusters in 6 X 4 Mesh (why not 6 x 8?)

128 bit links

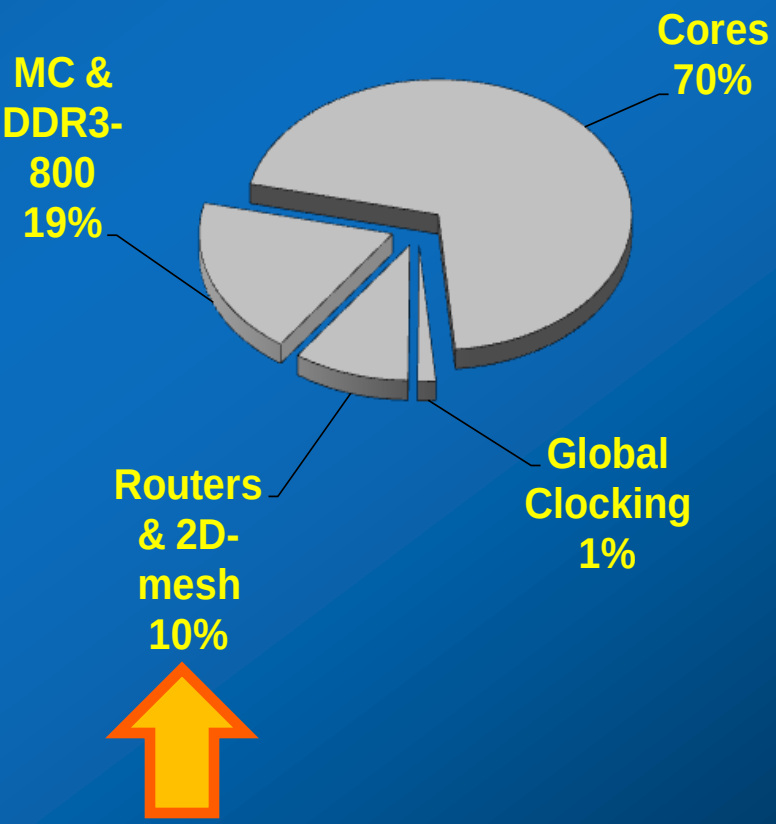
256 GB/sec bisection BW @ 2 GHz

Power Breakdown

80 Core TFLOP Chip

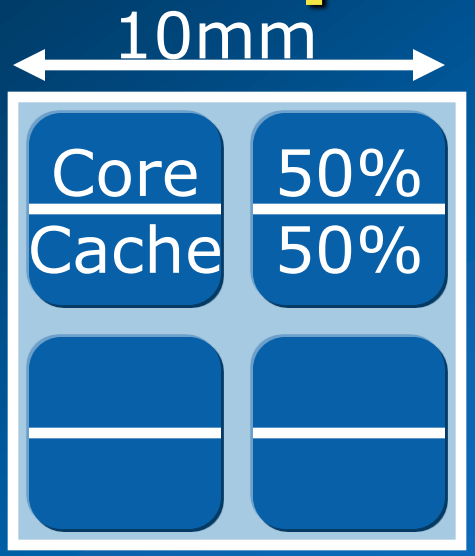


48 Core SCC Chip

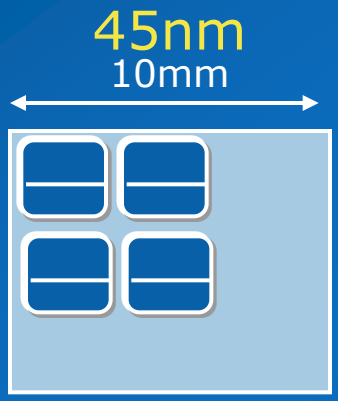


Does pt-to-pt packet switched network on a chip make sense?

A Sample Multi-core System

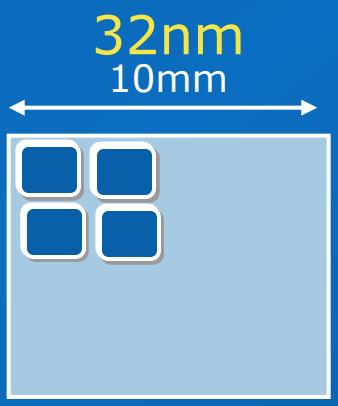


65nm, 4 Cores
1V, 3GHz
10mm die, 5mm each core
Core Logic: 6MT, Cache: 44MT
Total transistors: 200M



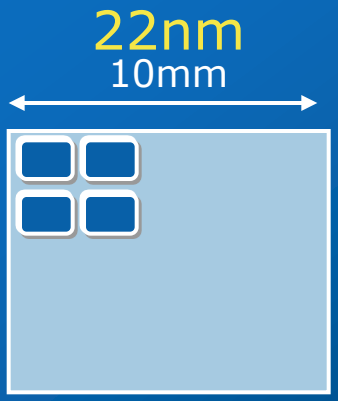
8 Cores, 1V, 3GHz
3.5mm each core

Total: 400MT



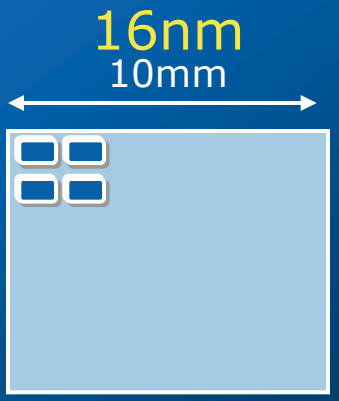
16 Cores, 1V, 3GHz
2.5mm each core

Total: 800MT



32 Cores, 1V, 3GHz
1.8mm each core

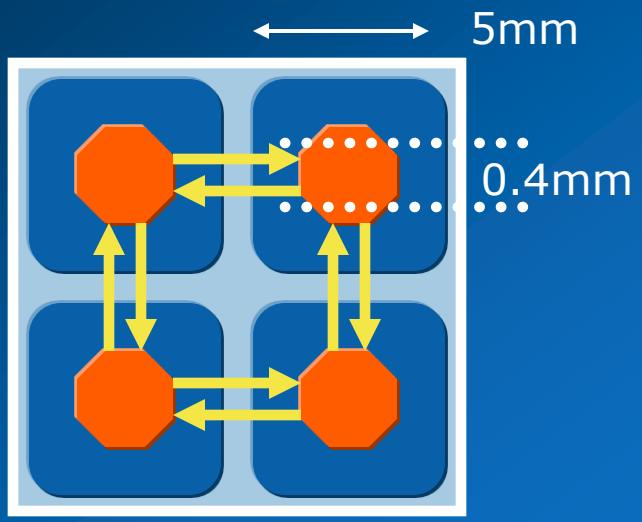
Total: 1.6BT



64 Cores, 1V, 3GHz
1.3mm each core

Total: 3.2BT

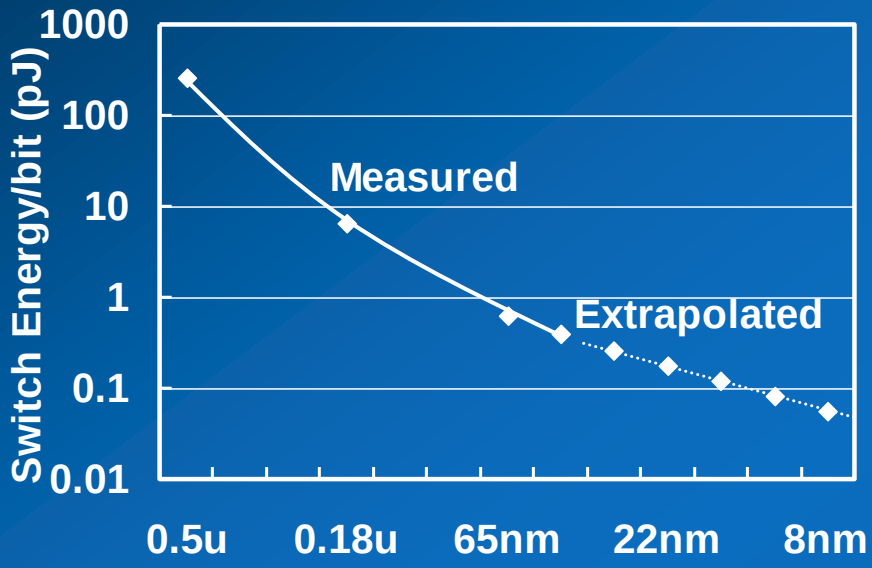
A Sample MC Network

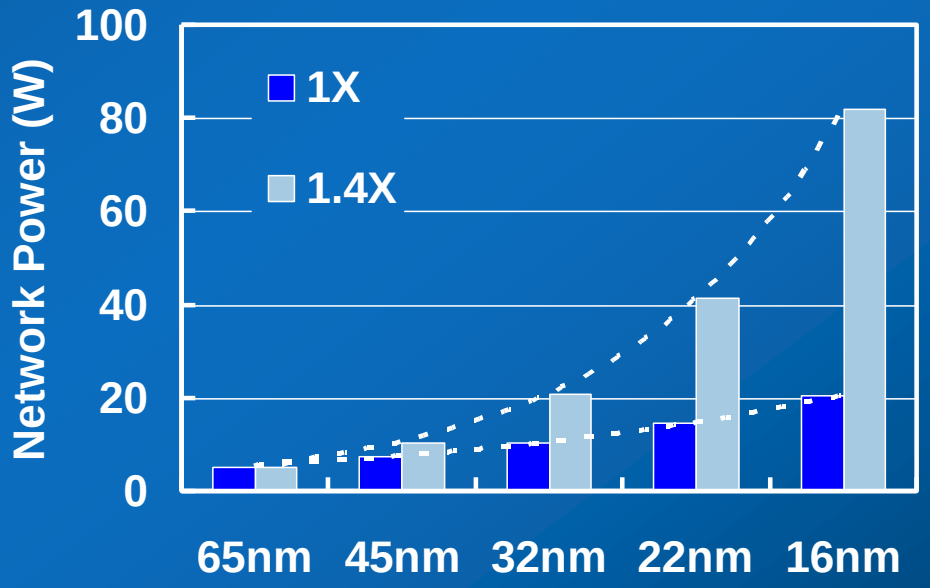


Packet Switched Mesh
16B=128 bit each direction
0.4mm @ 1.5u pitch
192GB/s Bisection BW

Tech	Core (mm)	Port size (mm)	Bisection BW GB/sec@3GHz
65nm	5	0.4	192
45nm	3.5	0.4	272
32nm	2.5	0.4	384
22nm	1.8	0.4	543
16nm	1.3	0.4	768

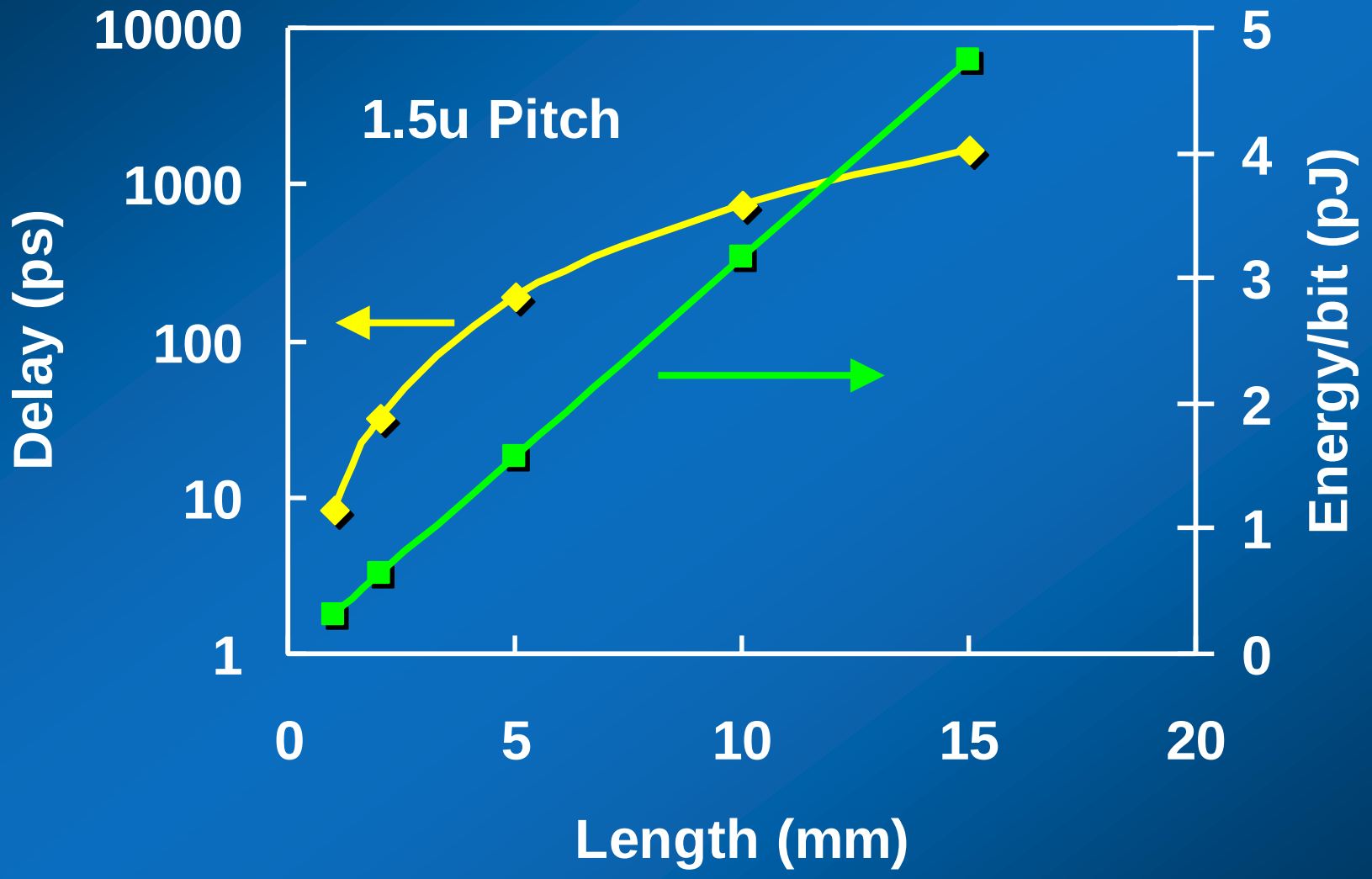
Mesh Power @ 3GHz, 1V



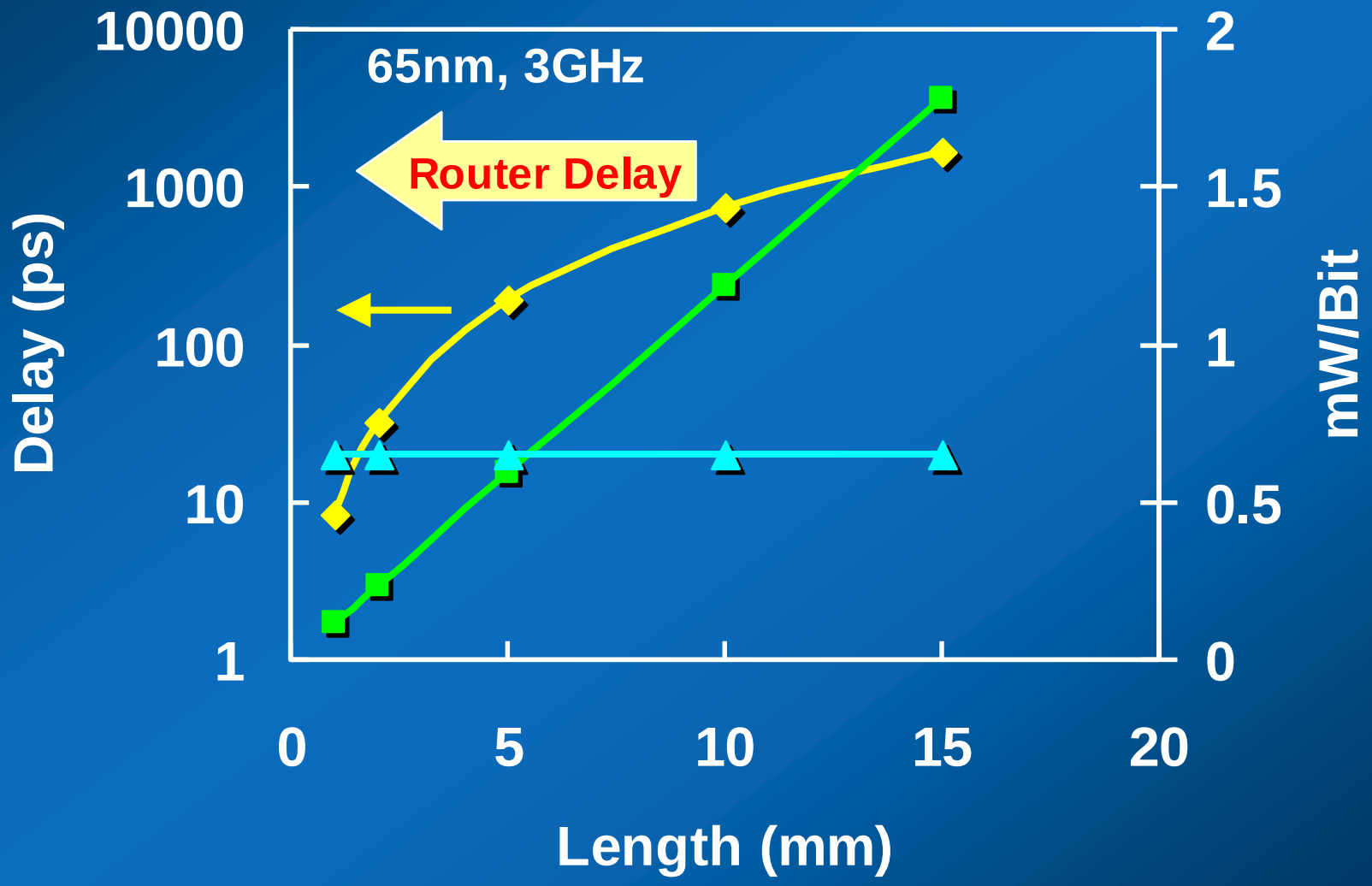


1. Network power will be too high
2. Worse if link width scales up each generation
3. Cache coherency mechanism is complex

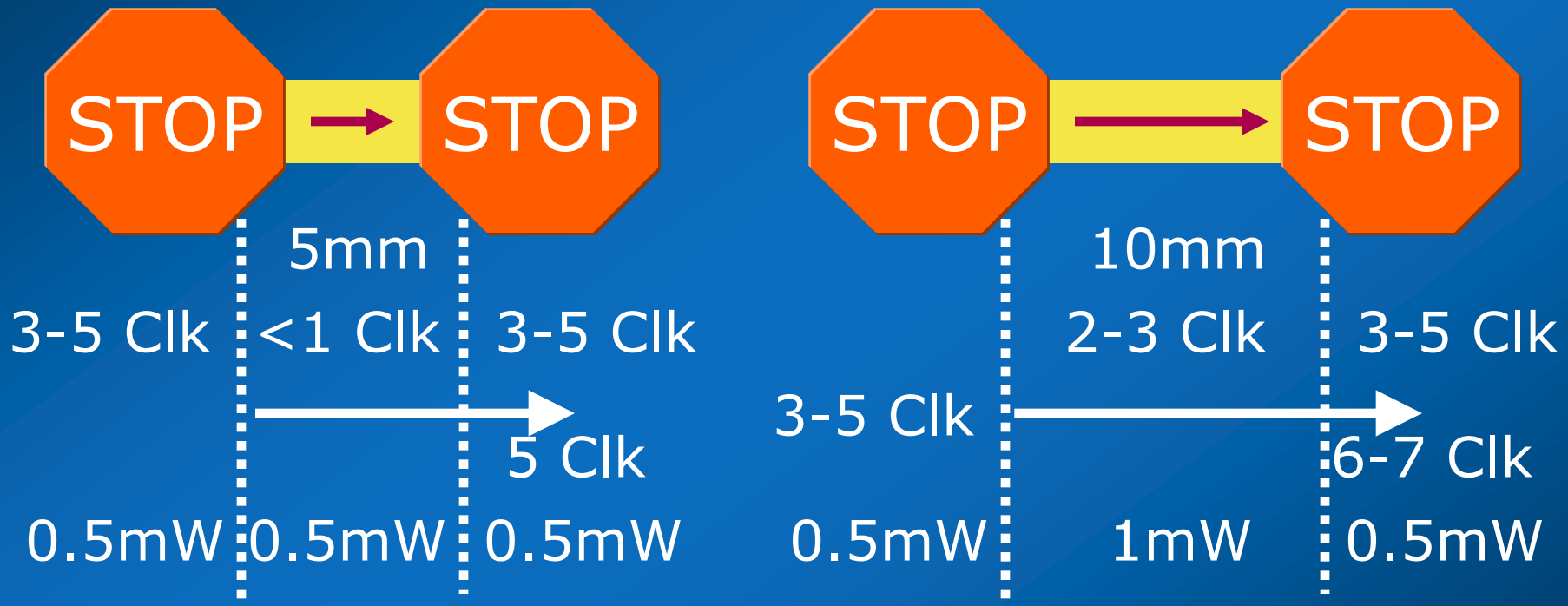
Interconnect Delay & Energy



Interconnect Delay & Power

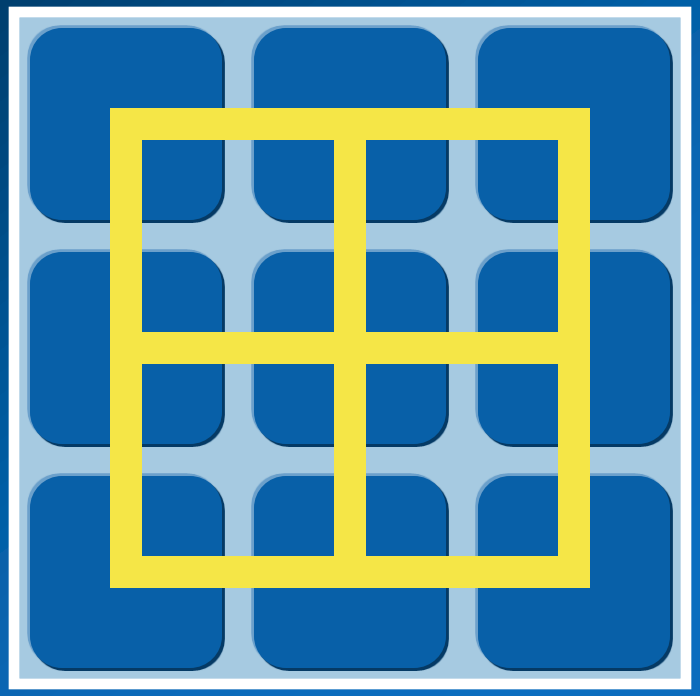


Packet Switched Interconnect



1. **Router acts like STOP signs—adds latency**
2. **Each hop consumes power (unnecessary)**

Bus—The Other Extreme…



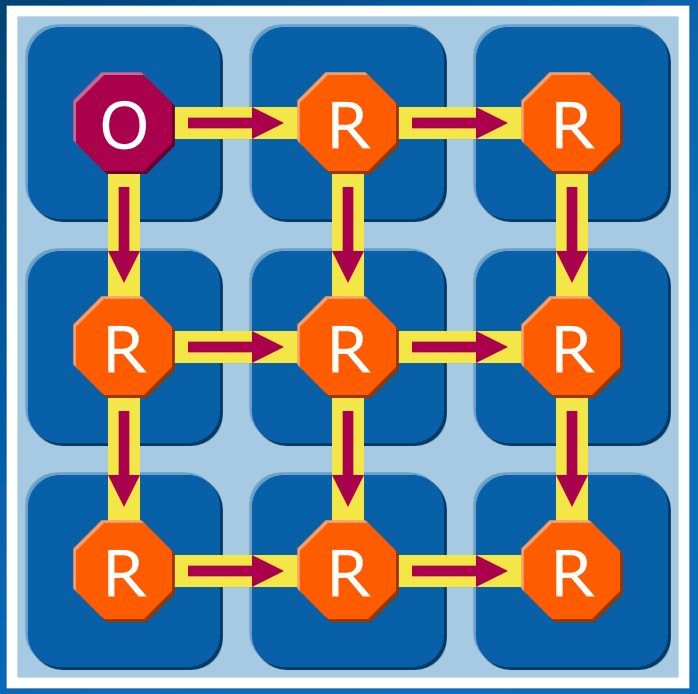
Issues:
Slow, < 300MHz
Shared, limited scalability?

Solutions:
Repeaters to increase freq
Wide busses for bandwidth
Multiple busses for scalability

Benefits:
Power?
Simpler cache coherency

Move away from frequency, embrace parallelism

Repeated Bus (Circuit Switched)



Arbitration:
Each cycle for the next cycle
Decision visible to all nodes

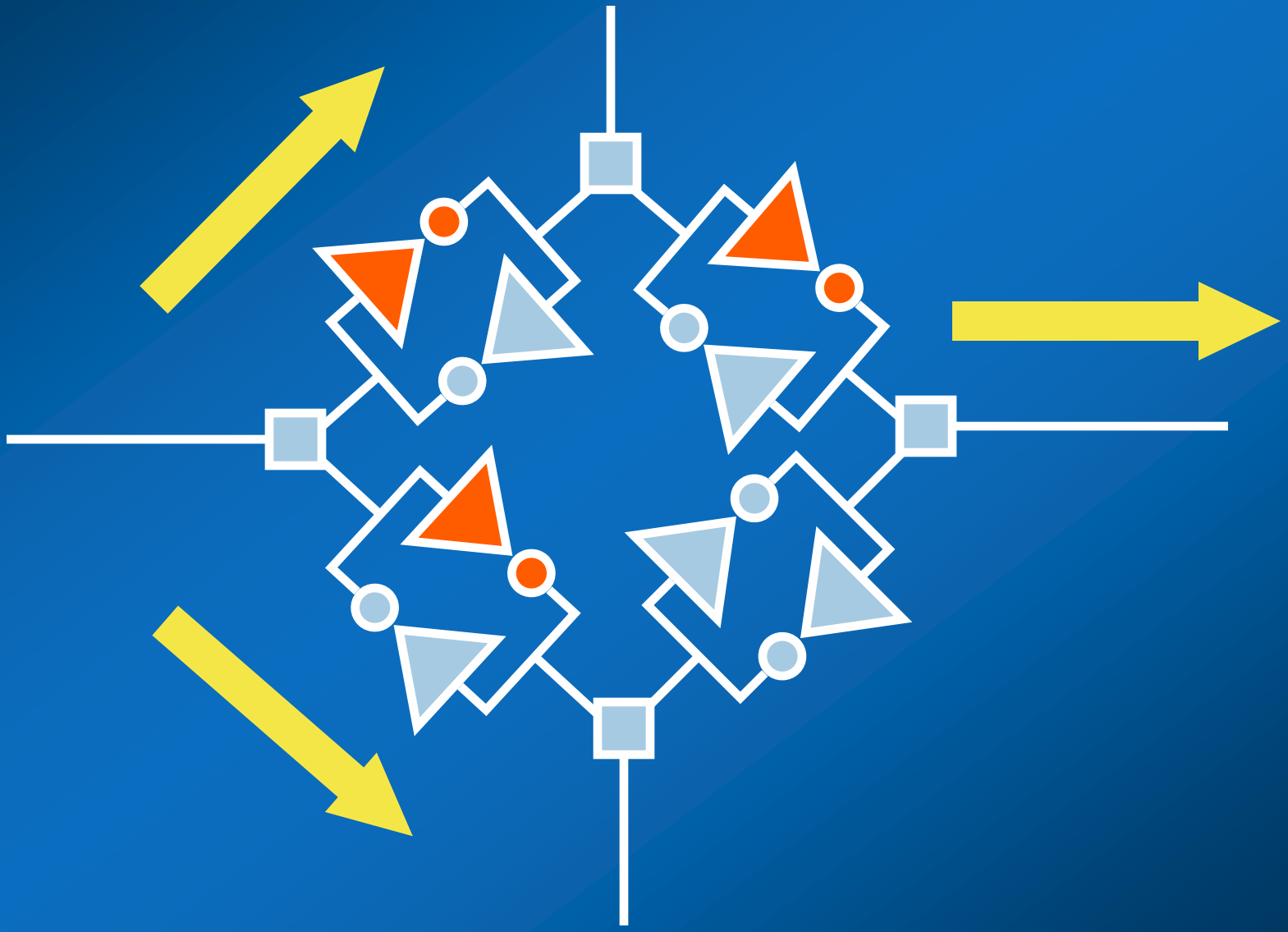
Repeaters:
Align repeater direction
No driving contention

	Core (mm)	Bus Seg Delay (ps)	Max Bus Freq (GHz)
65nm	5	195	2.2
45nm	3.5	99	2
32nm	2.5	51	1.8
22nm	1.8	26	1.5
16nm	1.3	13	1.2

Assume:
10mm die,
1.5u bus pitch
50ps repeater delay

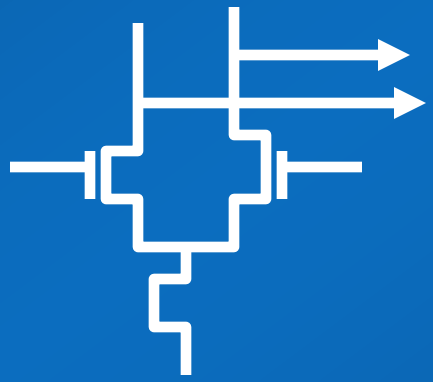
Anders et al, A 2.9Tb/s 8W 64-Core Circuit-switched Network-on-Chip in 45nm CMOS, ESSCIRC 2008
Anders et al, A 4.1Tb/s Bisection-Bandwidth 560Gb/s/W Streaming Circuit-Switched 8×8 Mesh Network-on-Chip in 45nm CMOS, ISSCC 2008

Example of a Bus Repeater

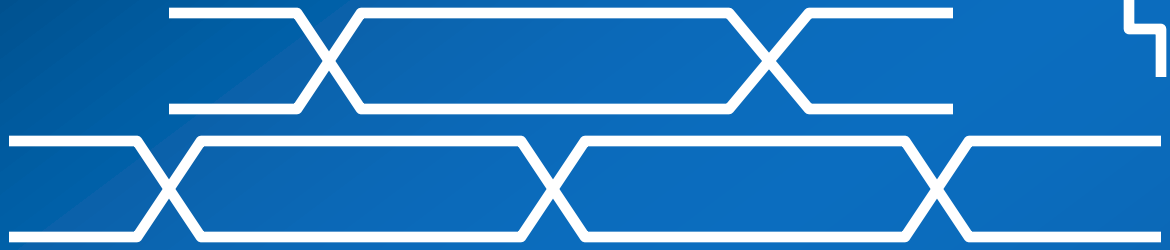


Other Bus Enhancements

Differential, low voltage swing



Twisted to reduce cross-talk



Optimal repeater placement

- Not necessarily at the core
- Higher bus frequency

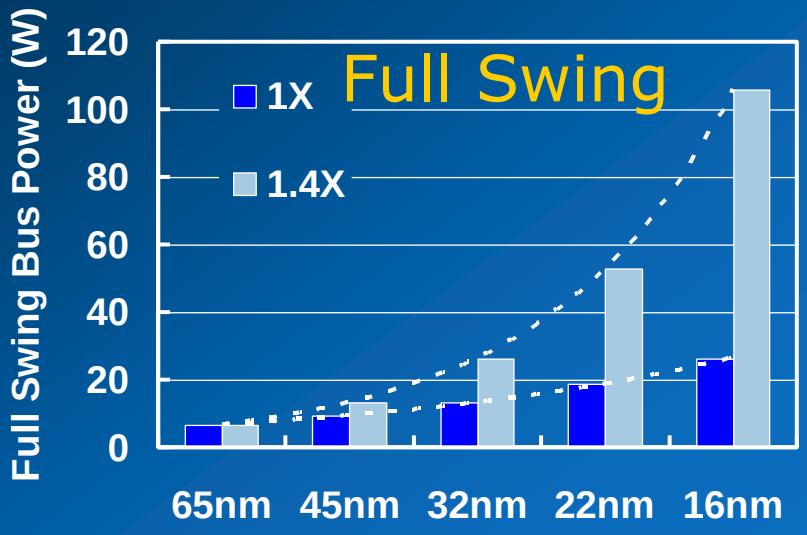
Wide bus, 1024 bit or more, transfer lots of data in one cycle

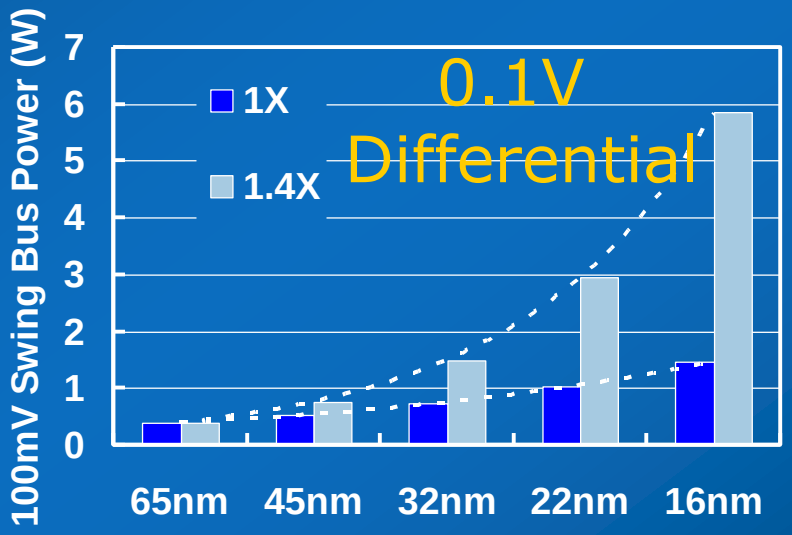
Multiple busses for concurrency

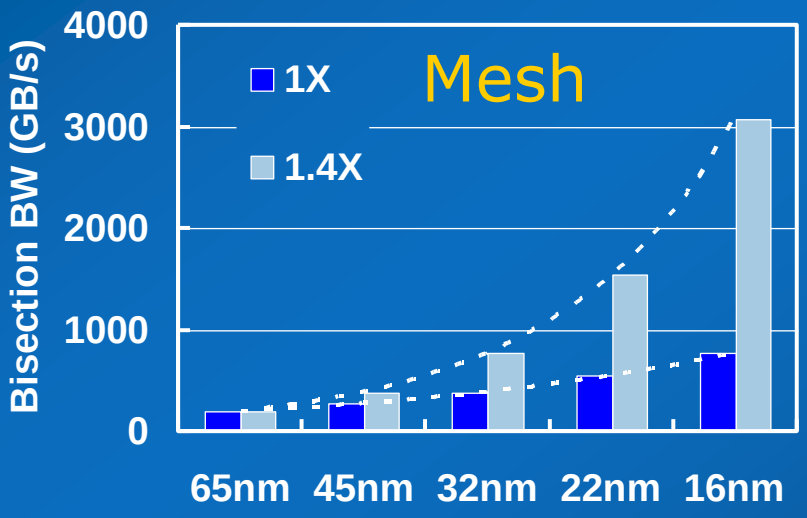
Employ interconnect engineering techniques

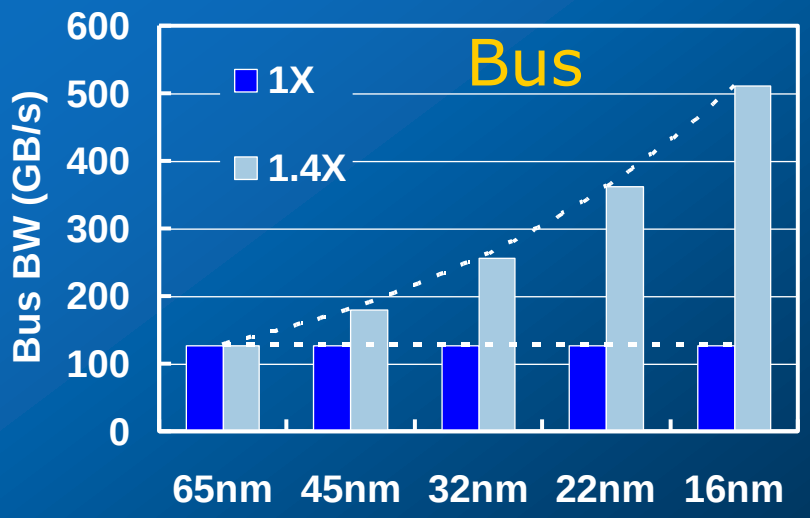
Bus Power and Bandwidth

Includes bus and repeater power

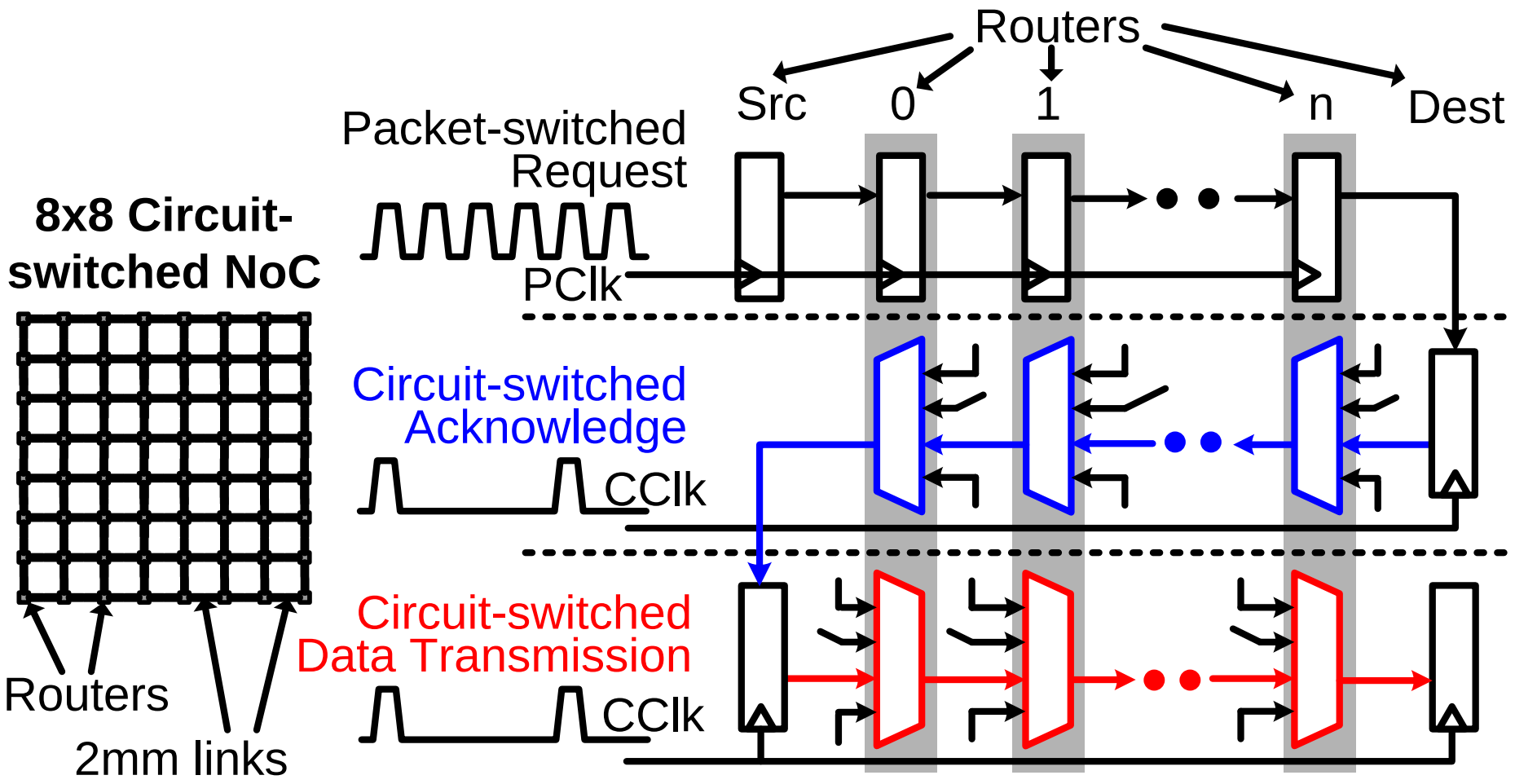








A Circuit Switched Network

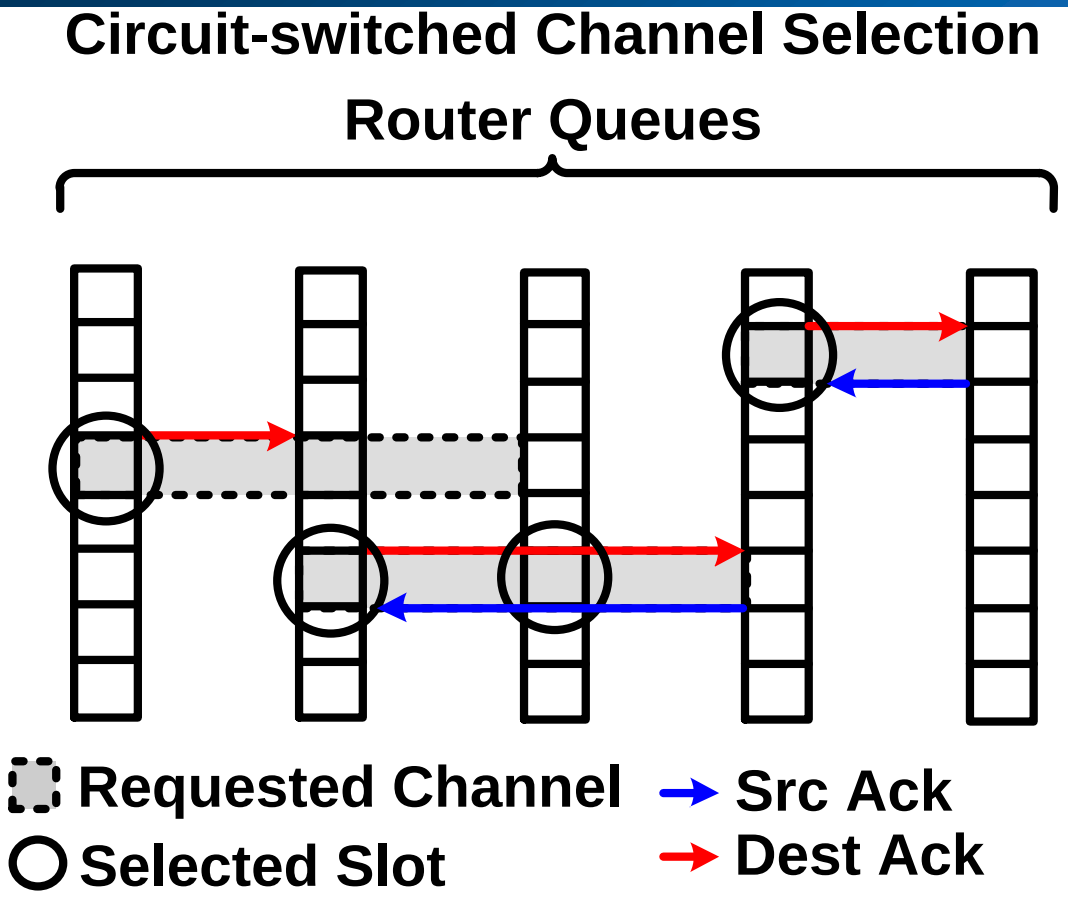


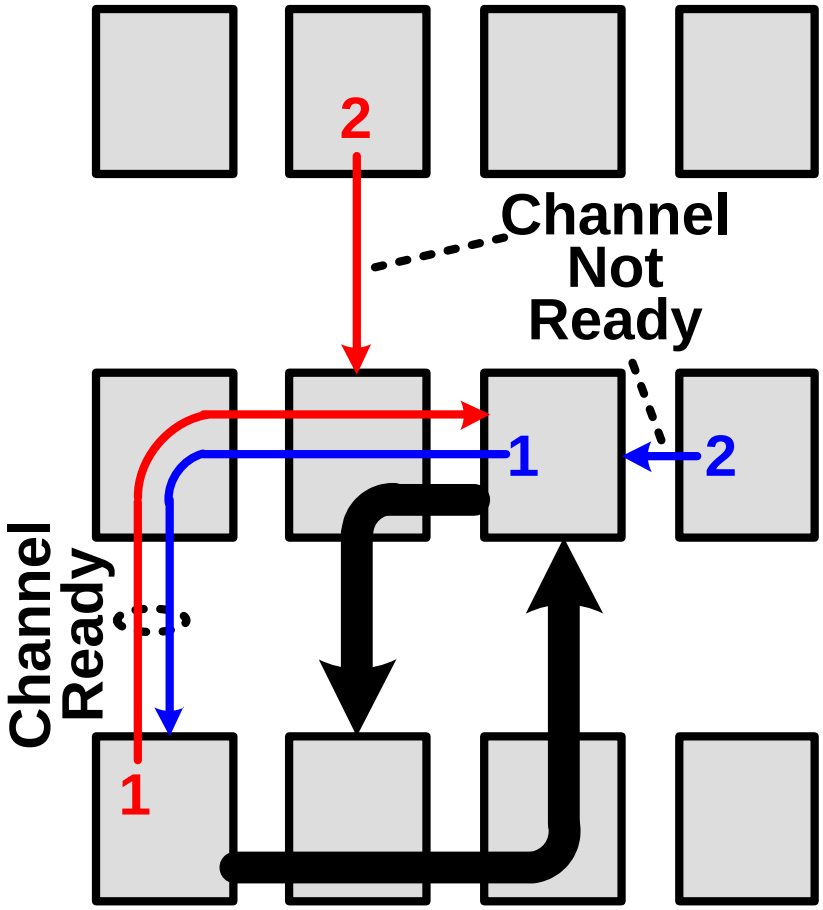
- **Circuit-switched NoC eliminates intra-route data storage**
- **Packet-switching used only for channel requests**

⇒ **High bandwidth and energy efficiency (1.6 to 0.6 pJ/bit)**

Anders et al, A 4.1Tb/s Bisection-Bandwidth 560Gb/s/W Streaming Circuit-Switched 8×8 Mesh Network-on-Chip in 45nm CMOS, ISSCC 2008

Channel Selection & Transmission



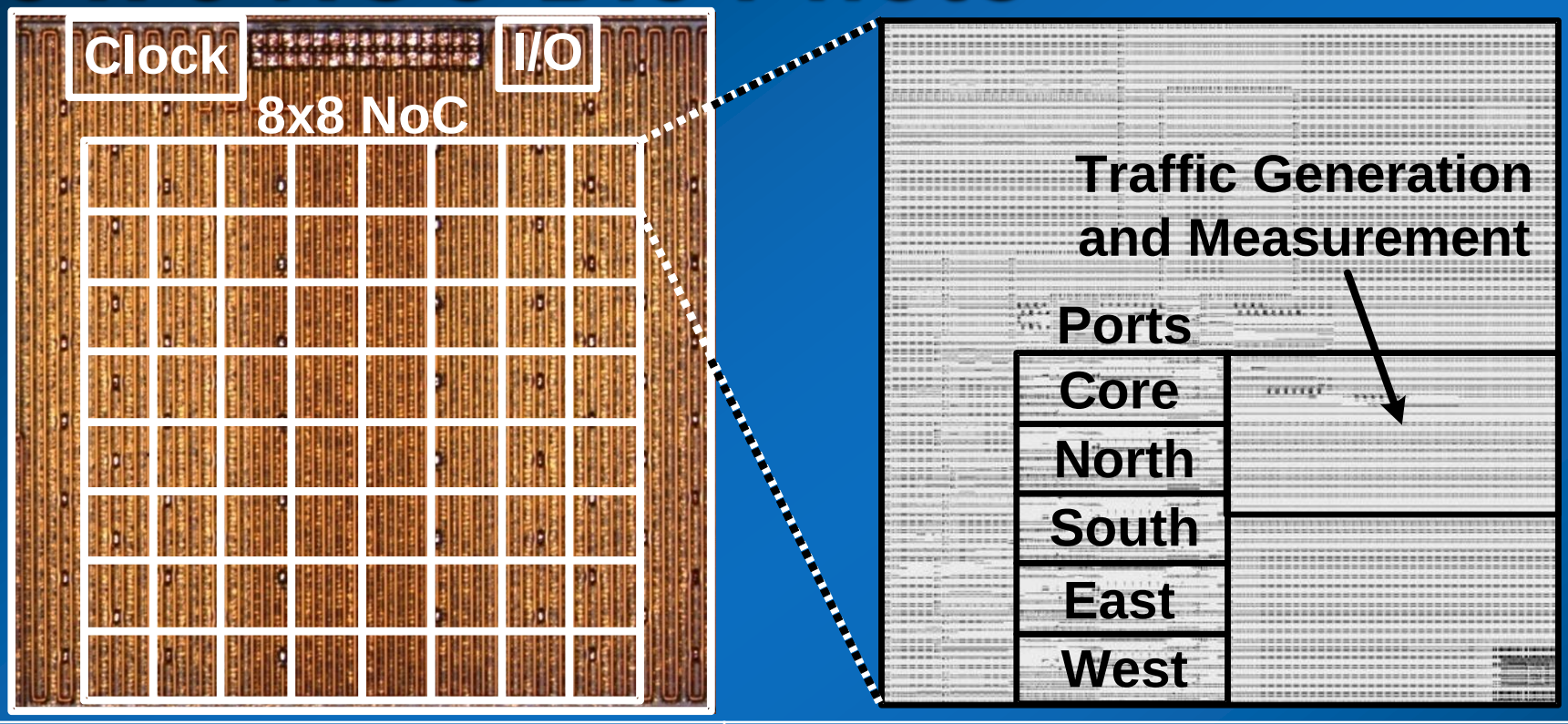


Queue slots store multiple channel requests

Two acknowledges required to indicate ready channel

More possible channels ⇒ 87% throughput increase

8 x 8 NOC Die Photo

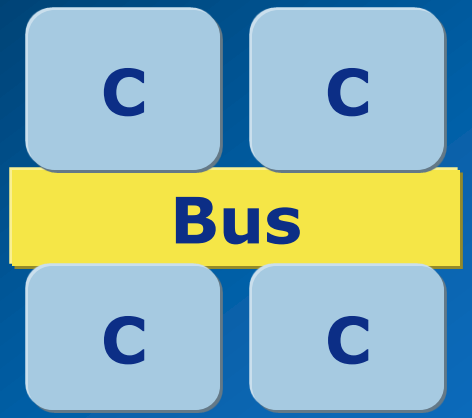


Process	45nm Hi-K/MG CMOS, 9 Metal Cu
Nominal Supply	1.1V
Arbitration and Router Logic	Supports 512b data
Data Bus Width	1b (2mm link distance)
Number of Transistors	2.85M
Die Area	6.25mm^2

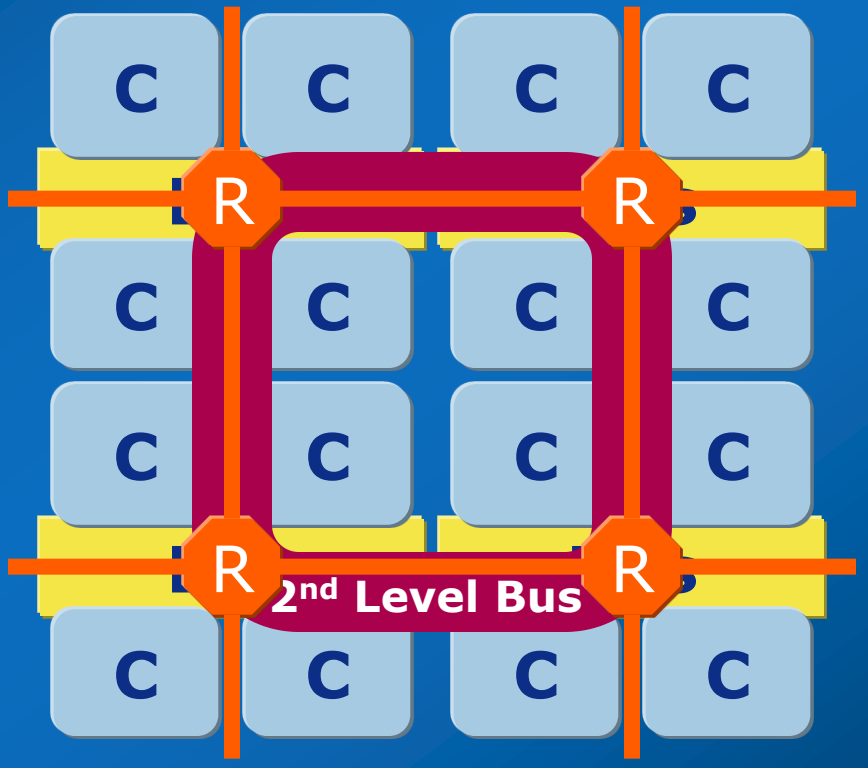
Factors Affecting Latency

Packet Switched	Circuit Switched
Arbitration in each node, multiple arbitration cycles	Single arbitration for entire transaction
Multiple hops from source to destination	Pipelined data flow
3-5 Clock latency in each node	One time latency to establish a circuit
Fast clock (3 GHz)	Slow clock (1 GHz)
One source and destination	Broadcast possible

Truth in between…



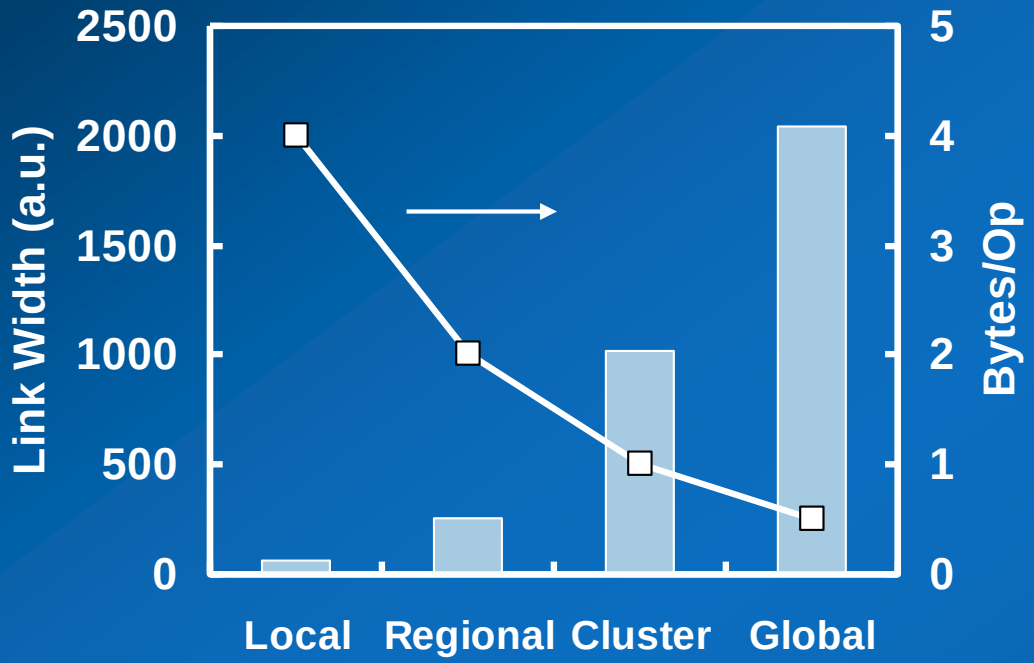
Bus to connect over short distances

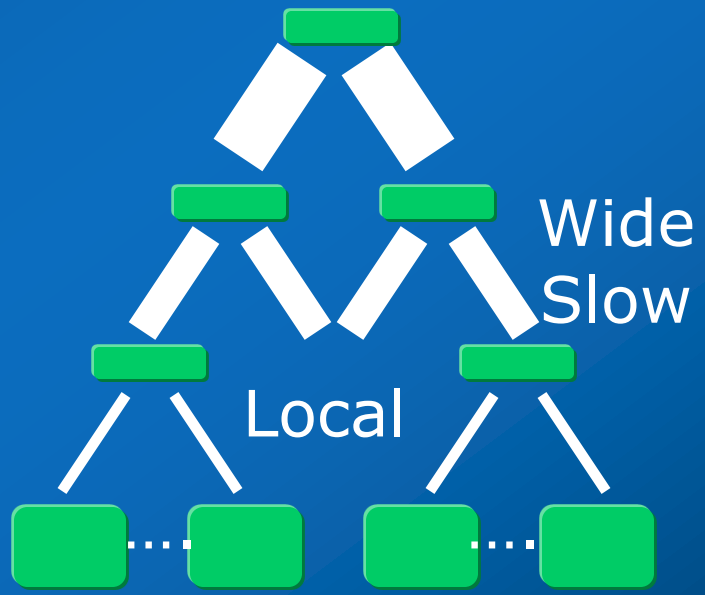


Hierarchy of Buses

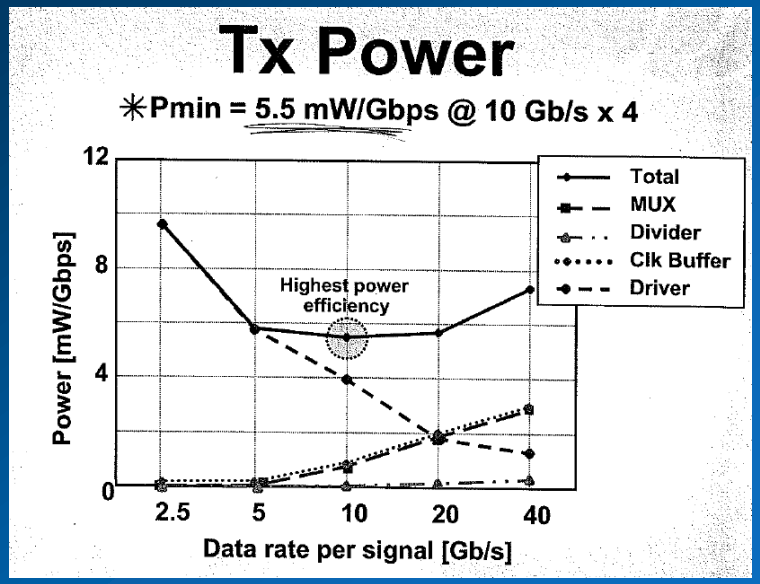
Or hierarchical circuit and packet switched networks

Hierarchical & Heterogeneous

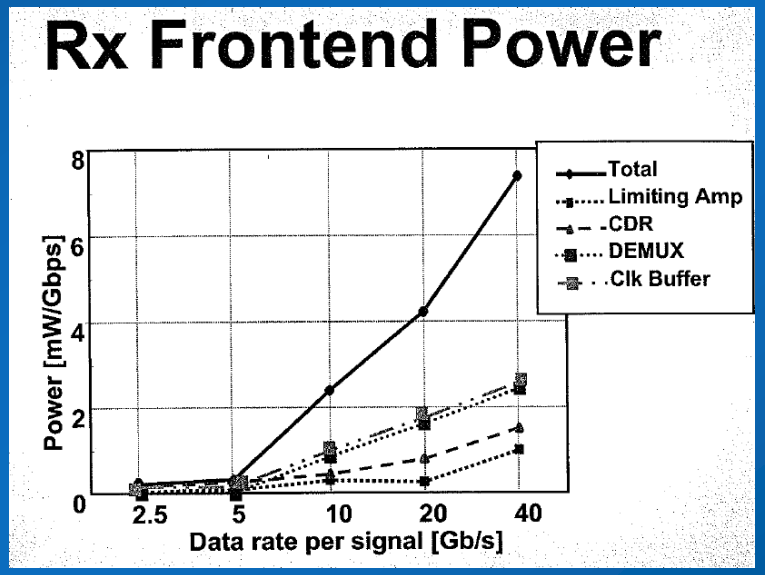


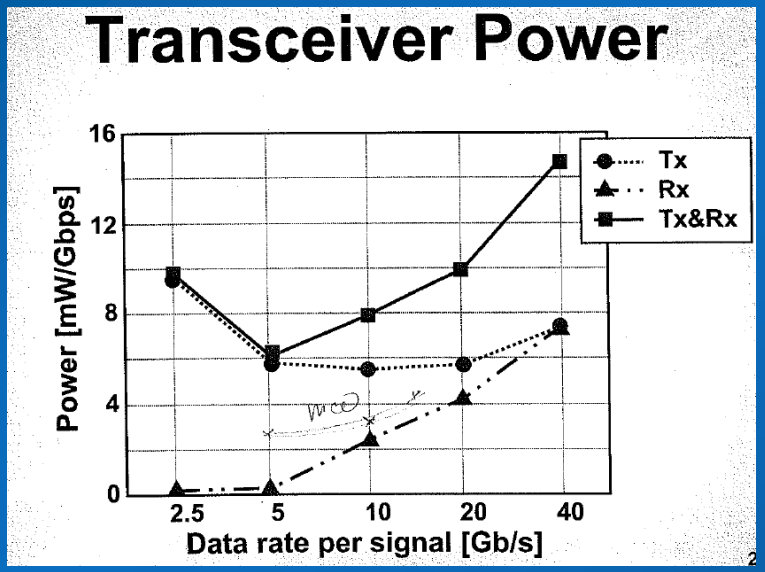


But wait, what about Optical?



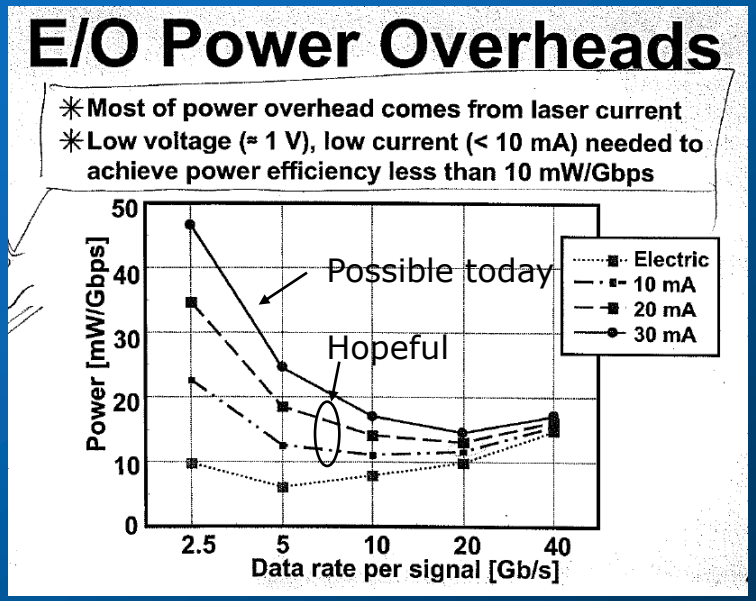
65nm





Optical:
Pre-driver
Driver
VCSEL

TIA
LA
DeMUX
CDR
Clk Buffer



Source: Hirotaka Tamure (Fujitsu), ISSCC 08 Workshop on HS Transceivers

Summary

Point to point busses are not necessary for on-die networks

Rings and meshes were devised for point to point busses over long distances—overkill for on chip network?

Router power could be prohibitive

Wide busses, circuit switched networks, show promise

Hierarchical, heterogeneous, tapered, circuit and packet switched schemes suitable for on-die networks

Go slower, wider, simpler, and efficient