Fiscal Year 2022

Ver. 2023-01-24a

Course number: CSC.T433
School of Computing,
Graduate major in Computer Science

Advanced Computer Architecture

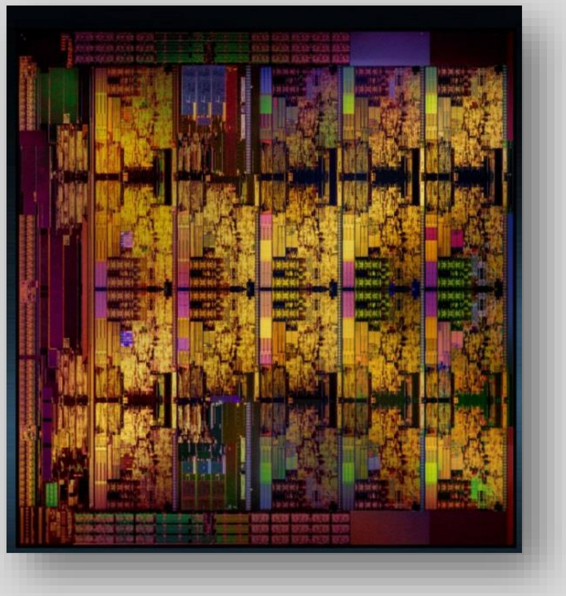
11. Thread Level Parallelism: Interconnection Network

www.arch.cs.titech.ac.jp/lecture/ACA/
Room No.W831, HyFlex
Mon 13:45-15:25, Thr 13:45-15:25

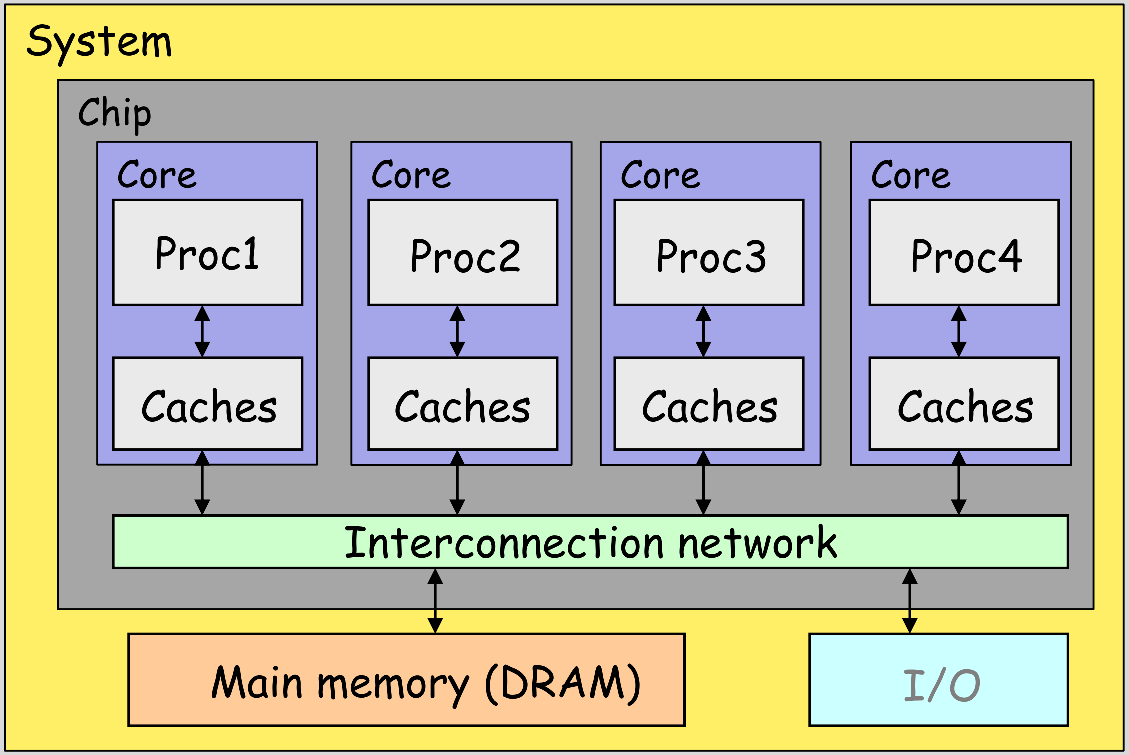
Kenji Kise, Department of Computer Science
kise _at_ c.titech.ac.jp

Shared memory many-core architecture

- The single-chip integrates many cores (conventional processors) and an interconnection network.
- All the processors can access the same address space of the main memory (shared memory) through an interconnection network.
- The shared memory or shared address space (SAS) is used as a means for communication between the processors.

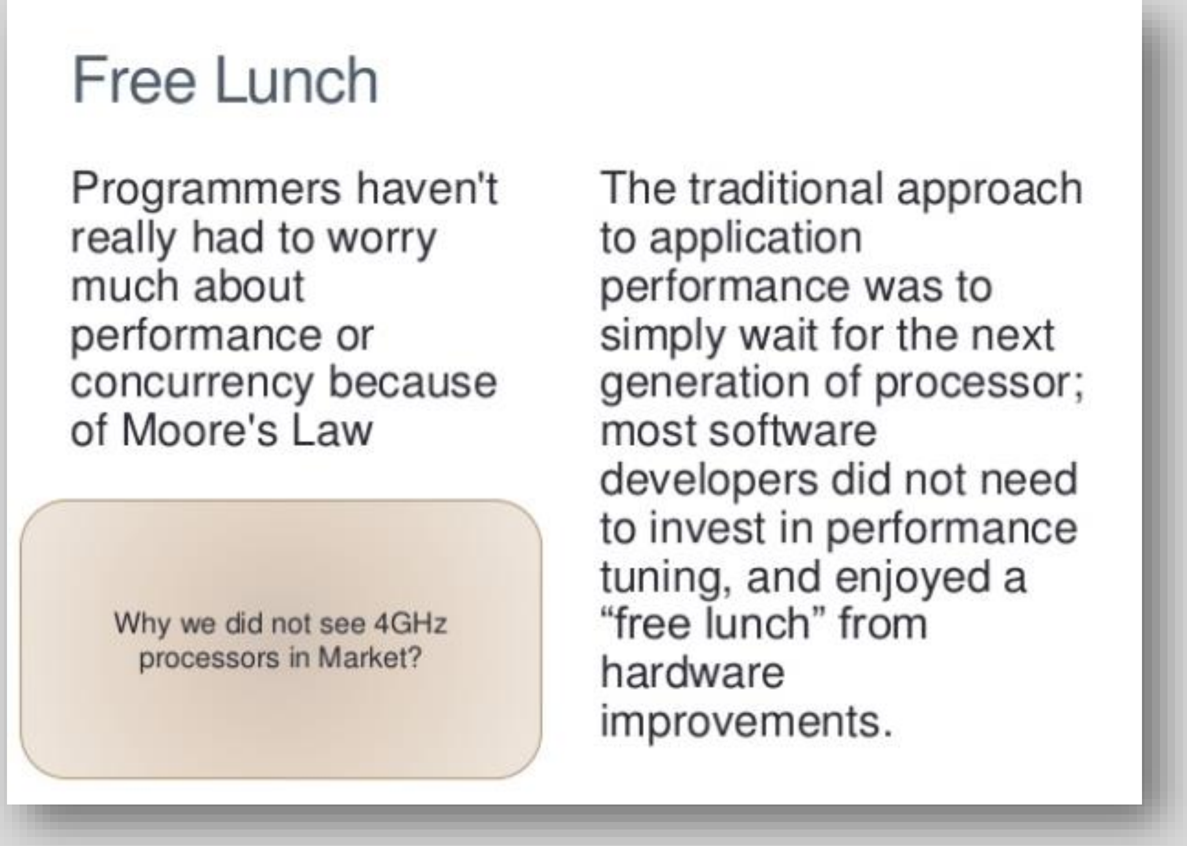


Intel Skylake-X, Core i9-7980XE, 2017



The free lunch is over

- Programmers have to worry much about performance and concurrency
- Parallel programming & multi-processor (multi-core) architecture



Free Lunch

Programmers haven't really had to worry much about performance or concurrency because of Moore's Law

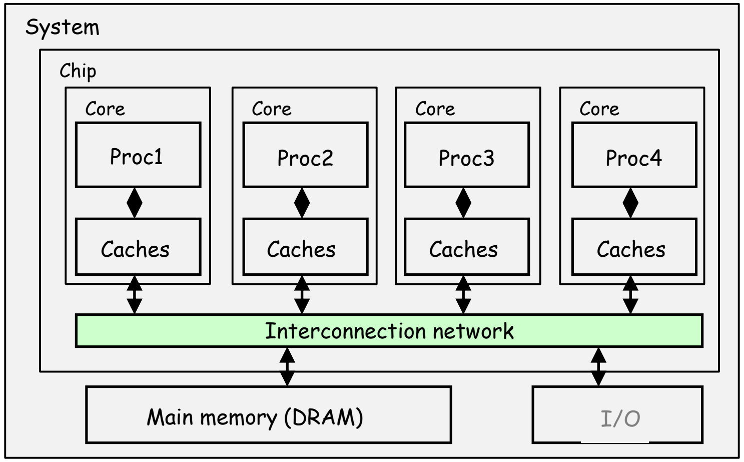
Why we did not see 4GHz processors in Market?

The traditional approach to application performance was to simply wait for the next generation of processor; most software developers did not need to invest in performance tuning, and enjoyed a “free lunch” from hardware improvements.

The Free Lunch Is Over: A Fundamental Turn Toward Concurrency in Software by Herb Sutter, 2005

Key components of many-core processors

- Interconnection network
  - connecting many modules on a chip achieving high throughput and low latency
- Main memory and caches
  - Caches are used to reduce latency and to lower network traffic
  - A parallel program has private data and shared data
  - New issues are cache coherence and memory consistency
- Core
  - High-performance superscalar processor providing a hardware mechanism to support thread synchronization

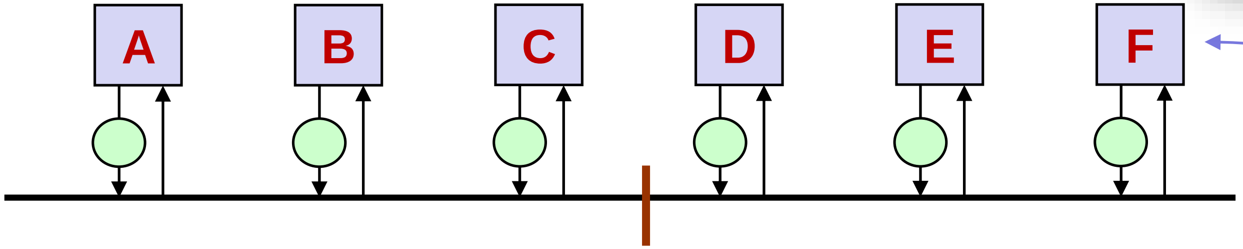


Performance metrics of interconnection network

- Network cost
  - number of links on a switch to connect to the network (plus one link to connect to the processor)
  - width in bits per link, length of link
- Network bandwidth (NB)
  - represents the best case
  - bandwidth of each link x number of links
- Bisection bandwidth (BB)
  - represents the worst case
  - divide the machine in two parts, each with half the nodes and sum the bandwidth of the links that cross the dividing line

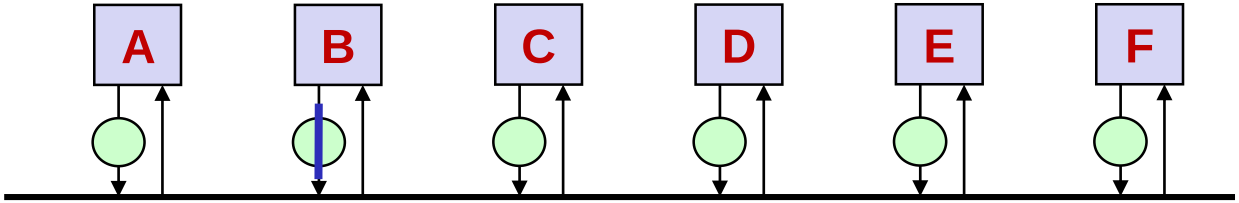
Bus Network

- N cores (□), N switch (○), 1 link (the bus)
- Only 1 simultaneous transfer at a time
  - NB (best case) = link (bus) bandwidth x 1
  - BB (worst case) = link (bus) bandwidth x 1
- All processors can snoop the bus



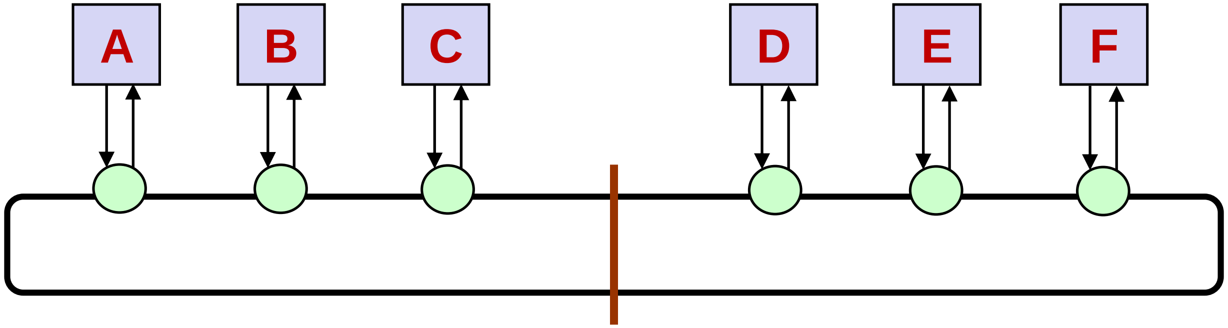
Core or processor node

The case where core B sends a packet to someone

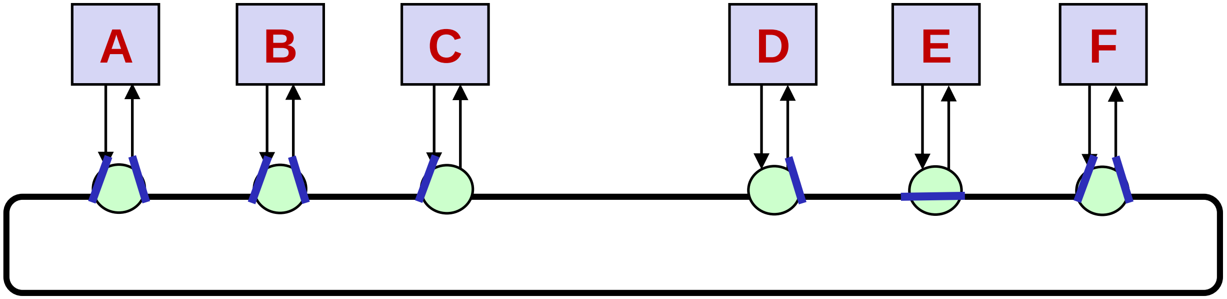


Ring Network

- N cores, N switches, 2 links/switch, N links
- N simultaneous transfers
  - NB (best case) = link bandwidth x N
  - BB (worst case) = link bandwidth x 2
- If a link is as fast as a bus, the ring is only twice as fast as a bus in the worst case, but is N times faster in the best case



The case where
A -> F, B->A, C->B, F->D



Cell Broadband Engine (2005)

- Cell Broadband Engine (2005)
  - 8 core (SPE) + 1 core (PPE)
    - each SPE has 256KB memory
  - PS3, IBM Roadrunner (12k cores)

PlayStation3
from PlaySation.com (Japan)

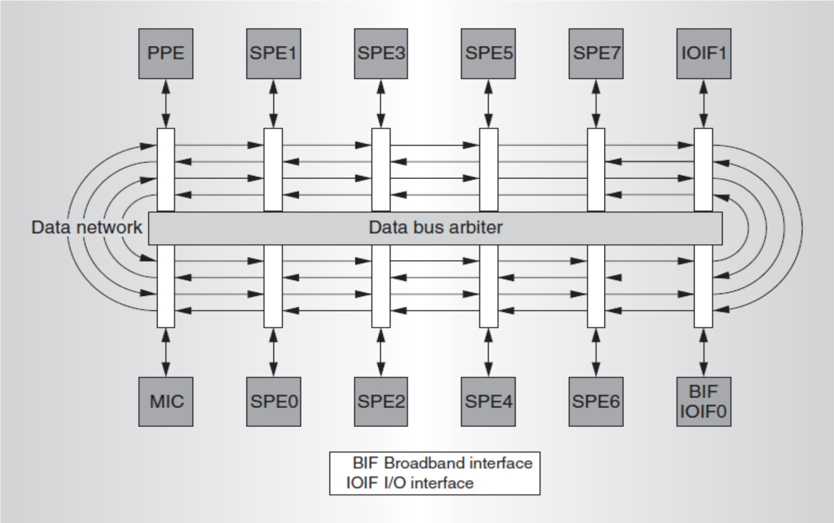


Figure 2. Element interconnect bus (EIB).

IEEE Micro, Cell Multiprocessor Communication Network: Built for Speed

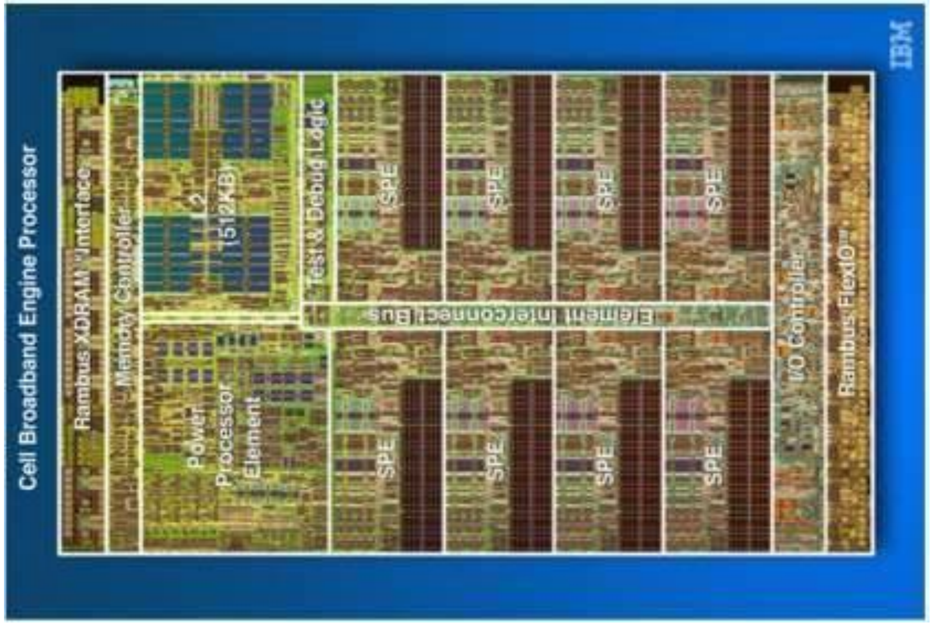
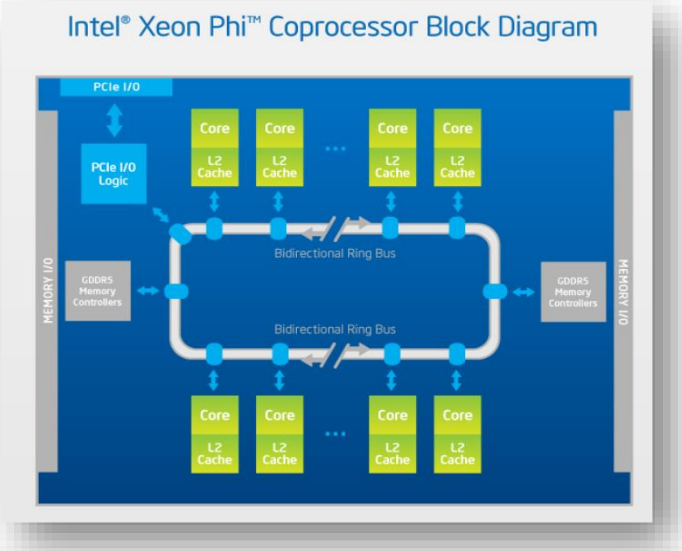


Diagram created by IBM to promote the CBEP, ©2005 from WIKIPEDIA

Intel Xeon Phi (2012)



Intel® Xeon Phi™ Coprocessor Block Diagram

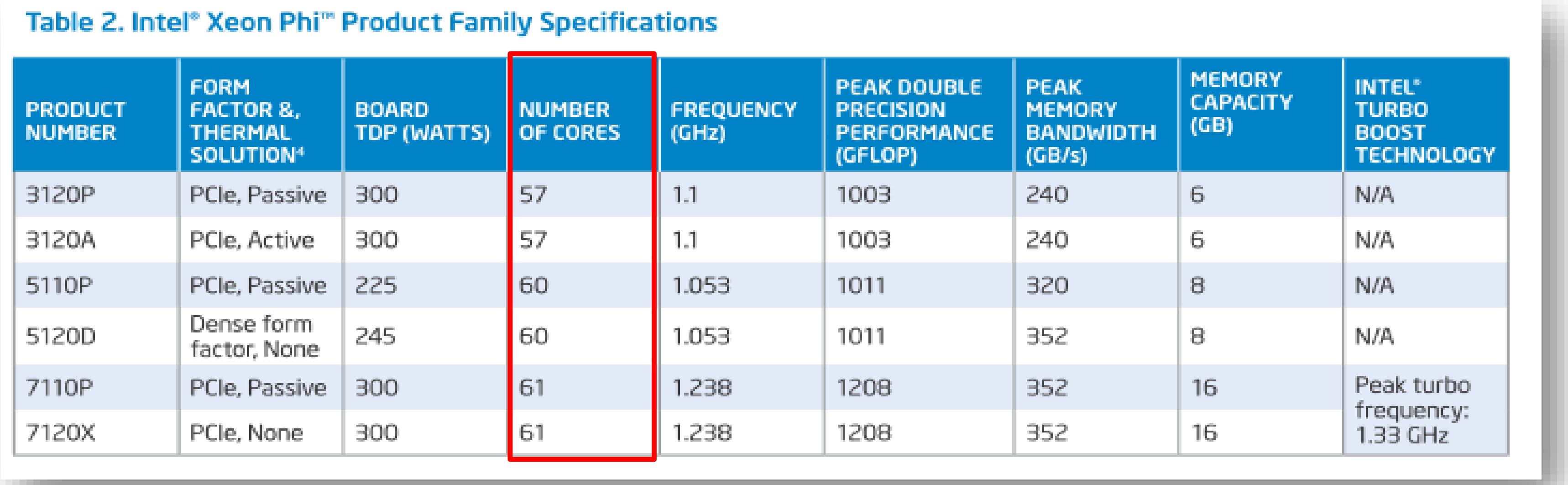
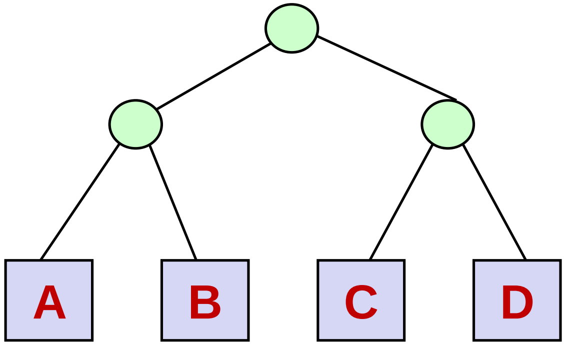


Table 2. Intel® Xeon Phi™ Product Family Specifications

PRODUCT NUMBER	FORM FACTOR &, THERMAL SOLUTION[4]	BOARD TDP (WATTS)	NUMBER OF CORES	FREQUENCY (GHz)	PEAK DOUBLE PRECISION PERFORMANCE (GFLOP)	PEAK MEMORY BANDWIDTH (GB/s)	MEMORY CAPACITY (GB)	INTEL® TURBO BOOST TECHNOLOGY
3120P	PCIe, Passive	300	57	1.1	1003	240	6	N/A
3120A	PCIe, Active	300	57	1.1	1003	240	6	N/A
5110P	PCIe, Passive	225	60	1.053	1011	320	8	N/A
5120D	Dense form factor, None	245	60	1.053	1011	352	8	N/A
7110P	PCIe, Passive	300	61	1.238	1208	352	16	Peak turbo frequency: 1.33 GHz
7120X	PCIe, None	300	61	1.238	1208	352	16	

Fat Tree (1)

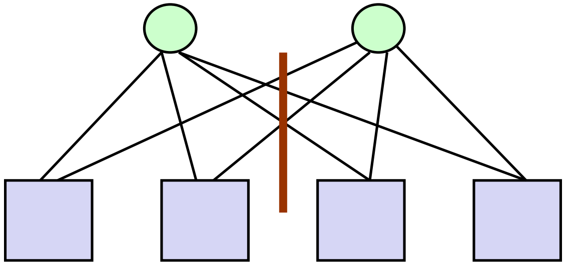
- Trees are good structures. People in CS use them all the time. Suppose we wanted to make a tree network.
- Any time A wants to send to C, it ties up the upper links, so that B can't send to D.
  - The bisection bandwidth on a tree is horrible - 1 link, at all times
- The solution is to 'thicken' the upper links.
  - More links as the tree gets thicker increases the bisection bandwidth



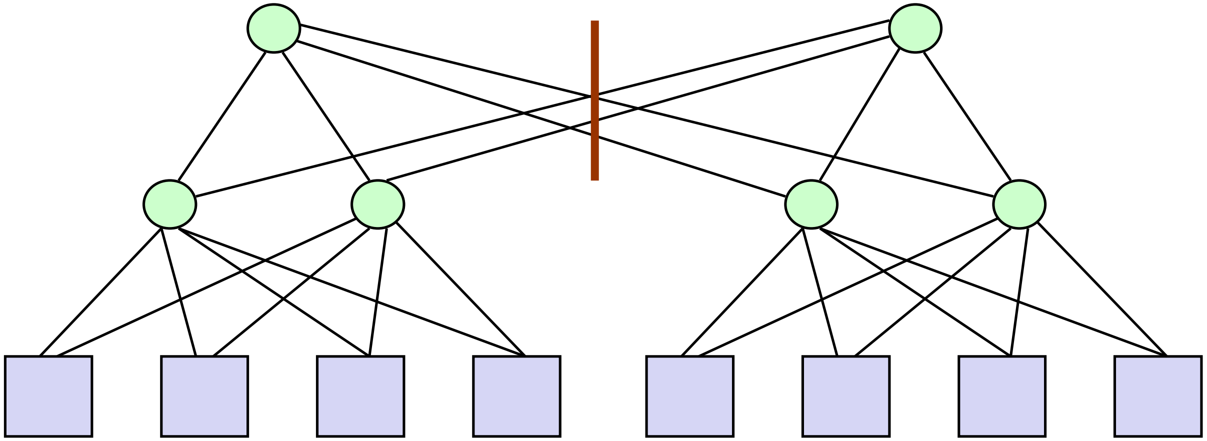
N = 4

Fat Tree

- N cores, log(N-1) x logN switches, 2 up + 4 down = 6 links/switch, N x logN links
- N simultaneous transfers
  - NB = link bandwidth x N log N
  - BB = link bandwidth x 4



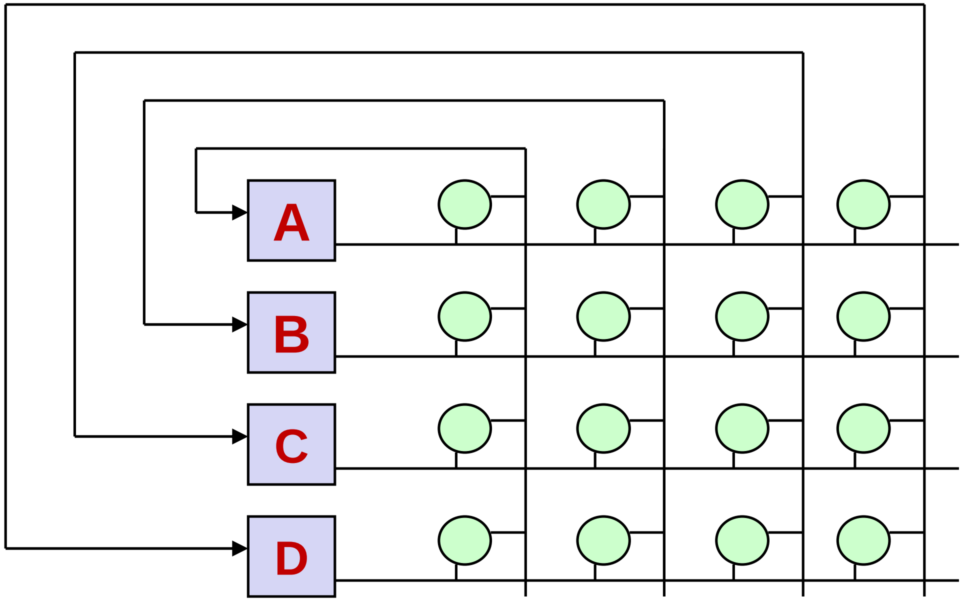
N = 4

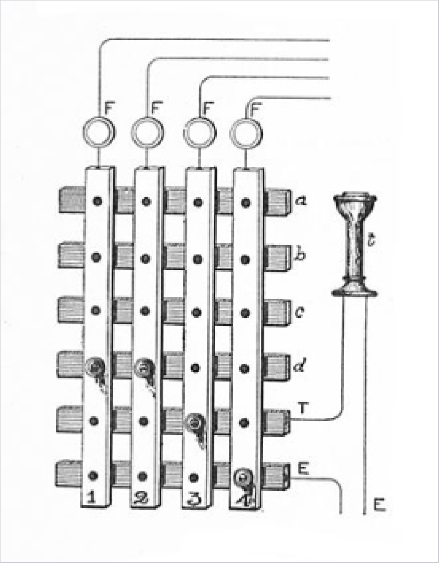


N = 8

Crossbar (Xbar) Network

- N cores, N^2 switches (unidirectional), 2 links/switch, N^2 links
- N simultaneous transfers
  - NB = link bandwidth x N (best case)
  - BB = link bandwidth x N (worst case)



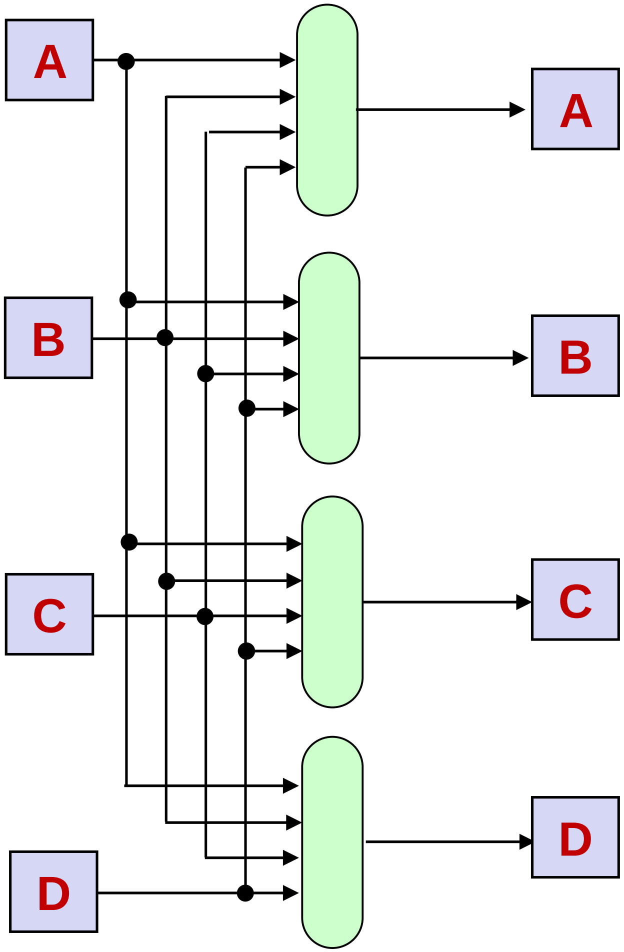


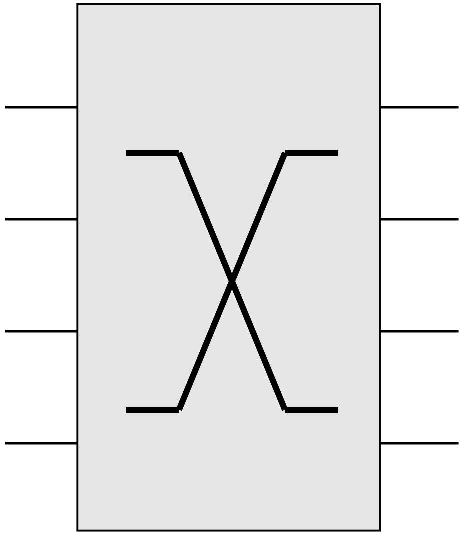
Crossbar telephone exchange of1903 for four subscribers (vertical bars), having four cross-bar talking circuits (horizontal bars), and one bar to connect the operator (T). The lowest cross-bar connects idle stations to ground to enable the signaling indicators (F). The switch is operated manually with metal pins that create a connection between the horizontally and vertically arranged bars.[1]

Wikipedia

Crossbar (Xbar) Network with mux

- N N-input multiplexers

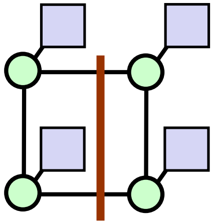




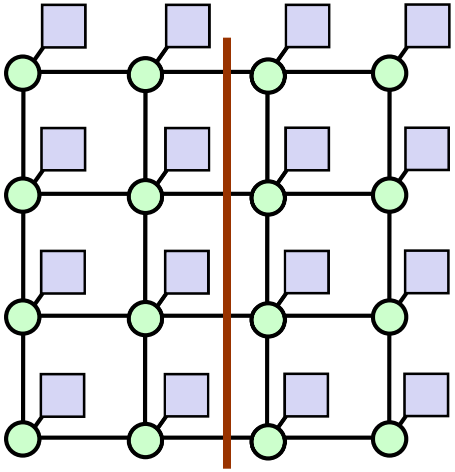
A symbol of Xbar

Mesh Network

- N cores, N switches, 4 links/switch, N x ($N^{1/2}$ – 1) links
- N simultaneous transfers
  - NB = link bandwidth x 2N (best case)
  - BB = link bandwidth x $2N^{1/2}$ (worst case)

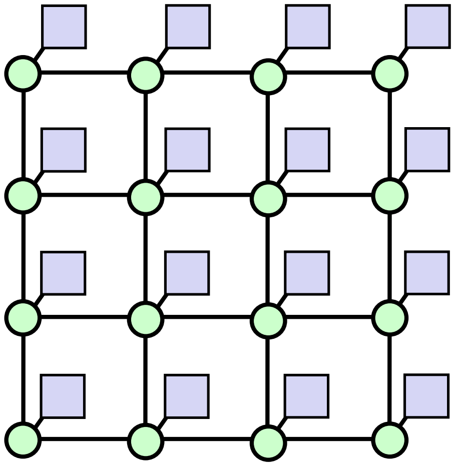


N = 4

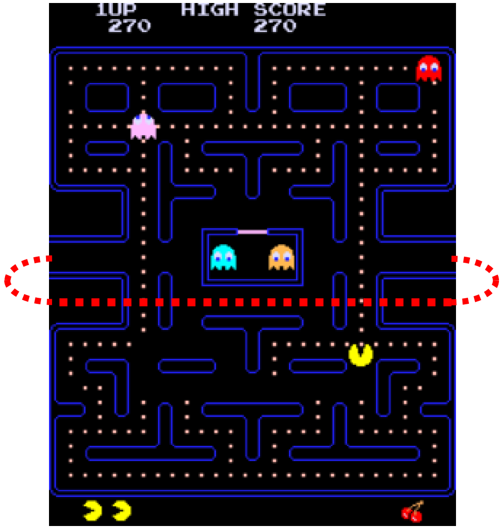


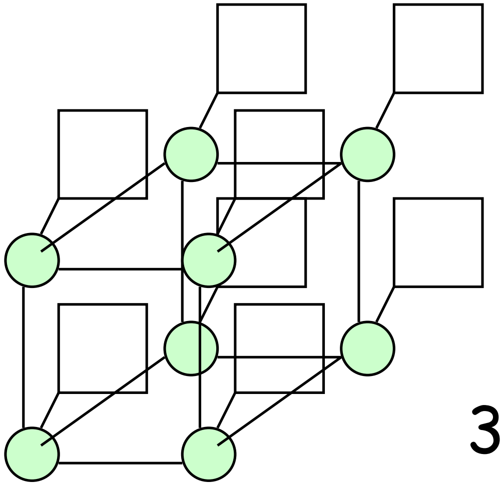
N = 16

2D and 3D Mesh / Torus Network

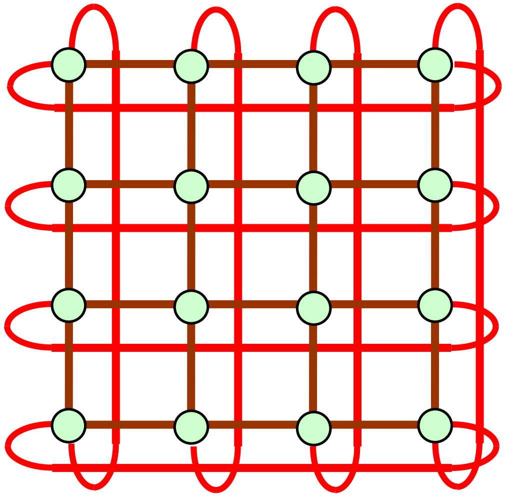


2D Mesh





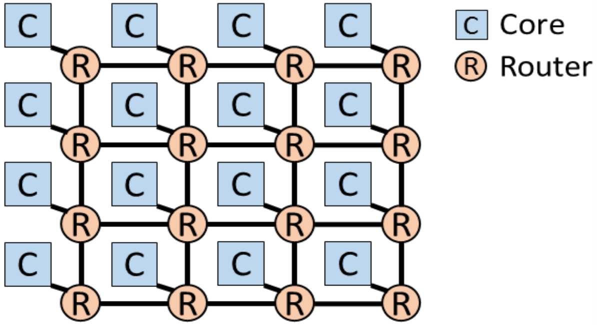
3D Mesh



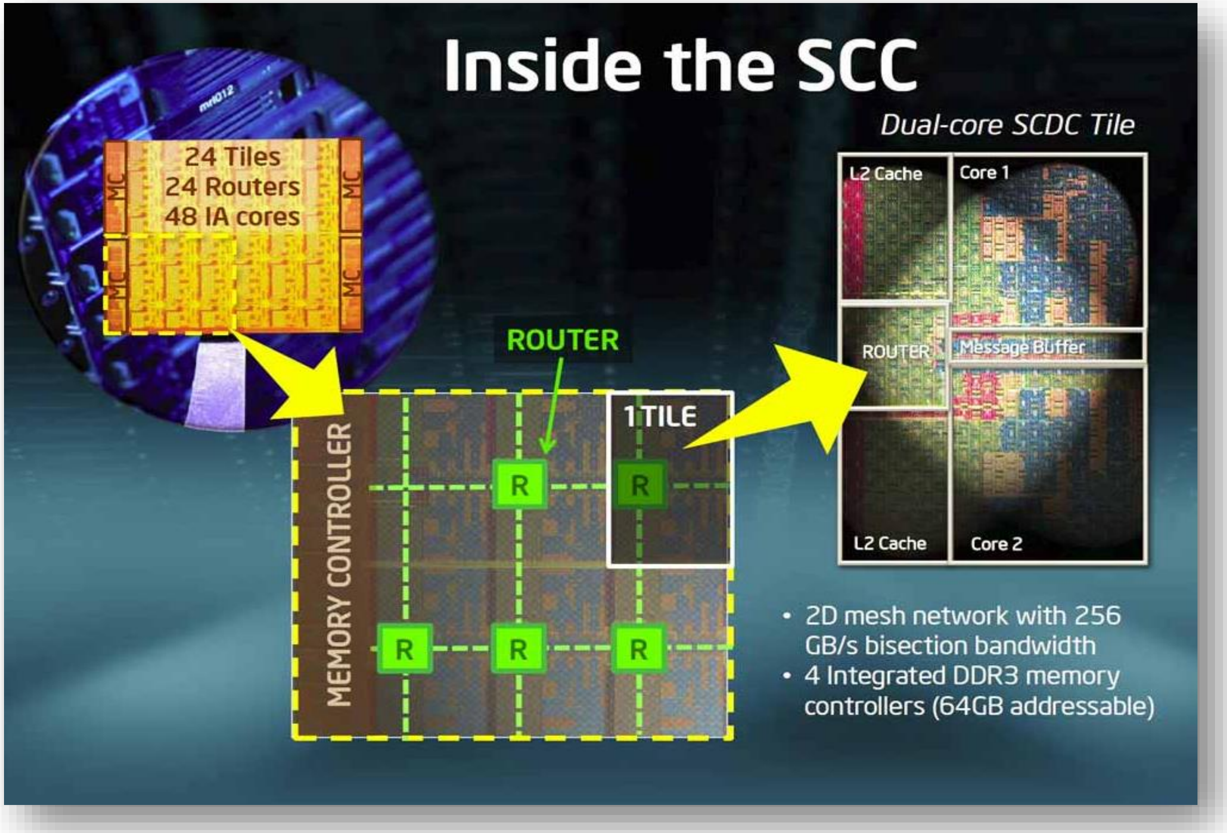
Torus

Intel Single-Chip Cloud Computer (2009)

- To research multi-core processors and parallel processing.



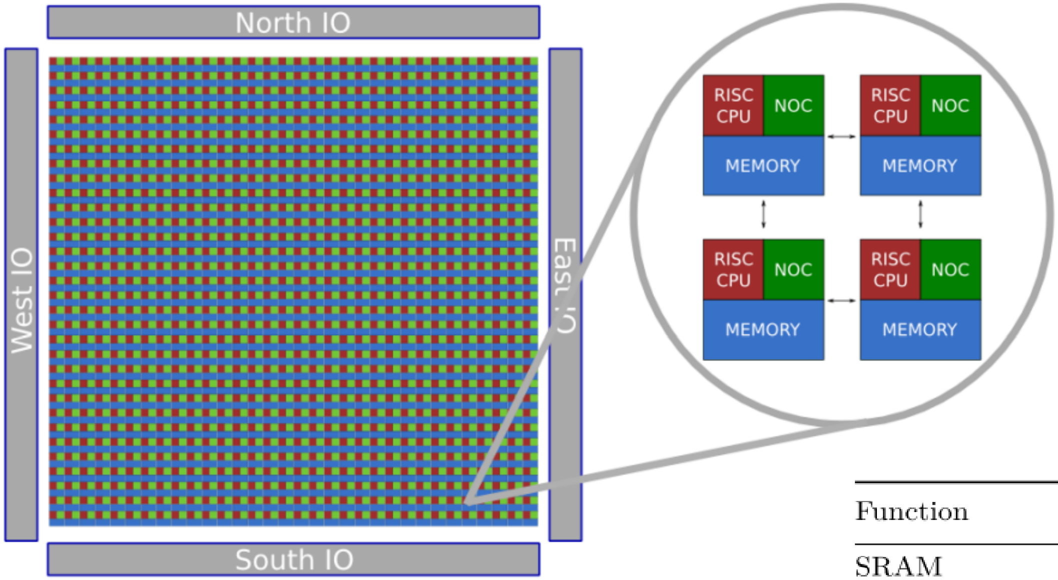
A many-core architecture with 2D Mesh NoC



- 2D mesh network with 256 GB/s bisection bandwidth
- 4 Integrated DDR3 memory controllers (64GB addressable)

Intel Single-Chip Cloud Computer (48 Core)

Epiphany-V: A 1024 core 64-bit RISC SoC (2016)



Summary of Epiphany-V features:

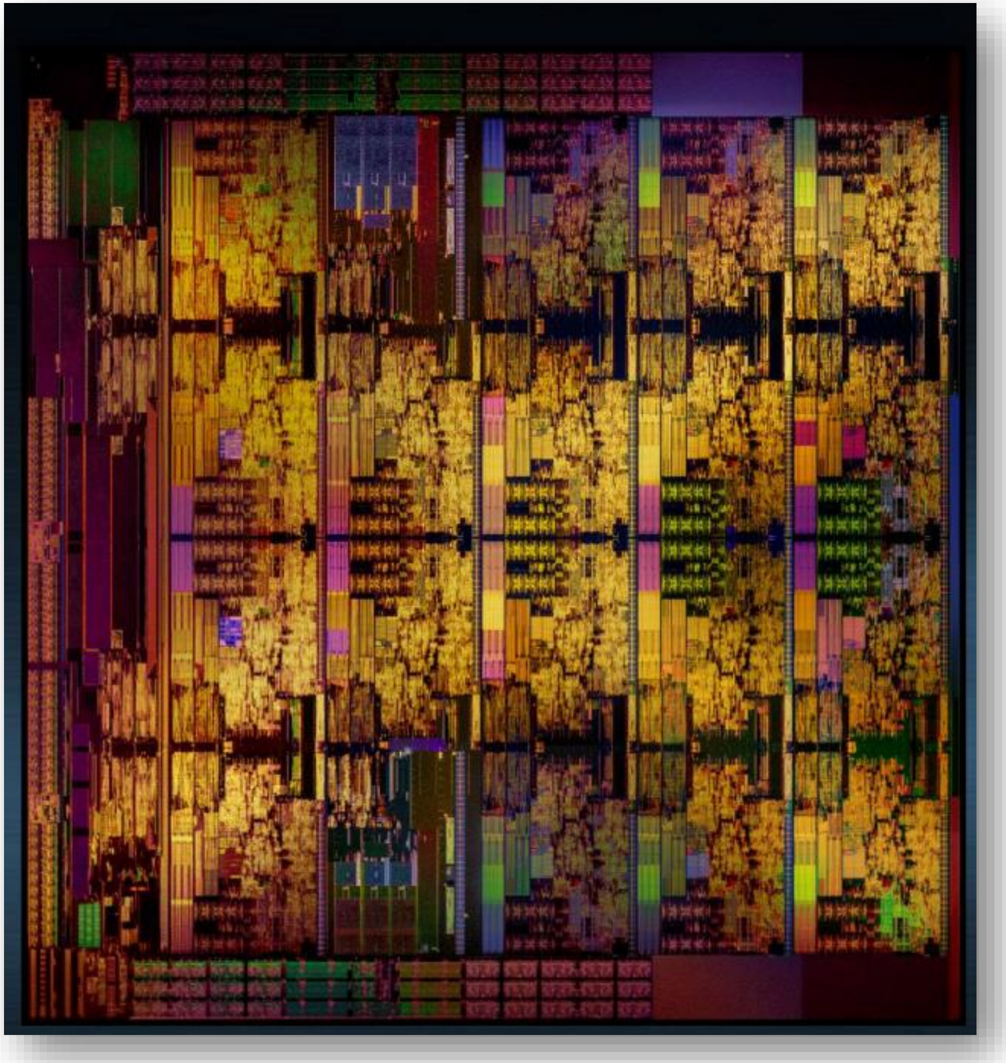
- 1024 64-bit RISC processors
- 64-bit memory architecture
- 64/32-bit IEEE floating point support
- 64MB of distributed on-chip memory
- 1024 programmable I/O signals
- Three 136-bit wide 2D mesh NOCs
- 2052 Independent Power Domains
- Support for up to 1 billion shared memory processors
- Binary compatibility with Epiphany III/IV chips
- Custom ISA extensions for deep learning, communication, and cryptography

Function	Value (mm^2)	Share of Total Die Area
SRAM	62.4	53.3%
Register File	15.1	12.9%
FPU	11.8	10.1%
NOC	12.1	10.3%
IO Logic	6.5	5.6%
“Other” Core Stuff	5.1	4.4%
IO Pads	3.9	3.3%
Always on Logic	0.66	0.6%

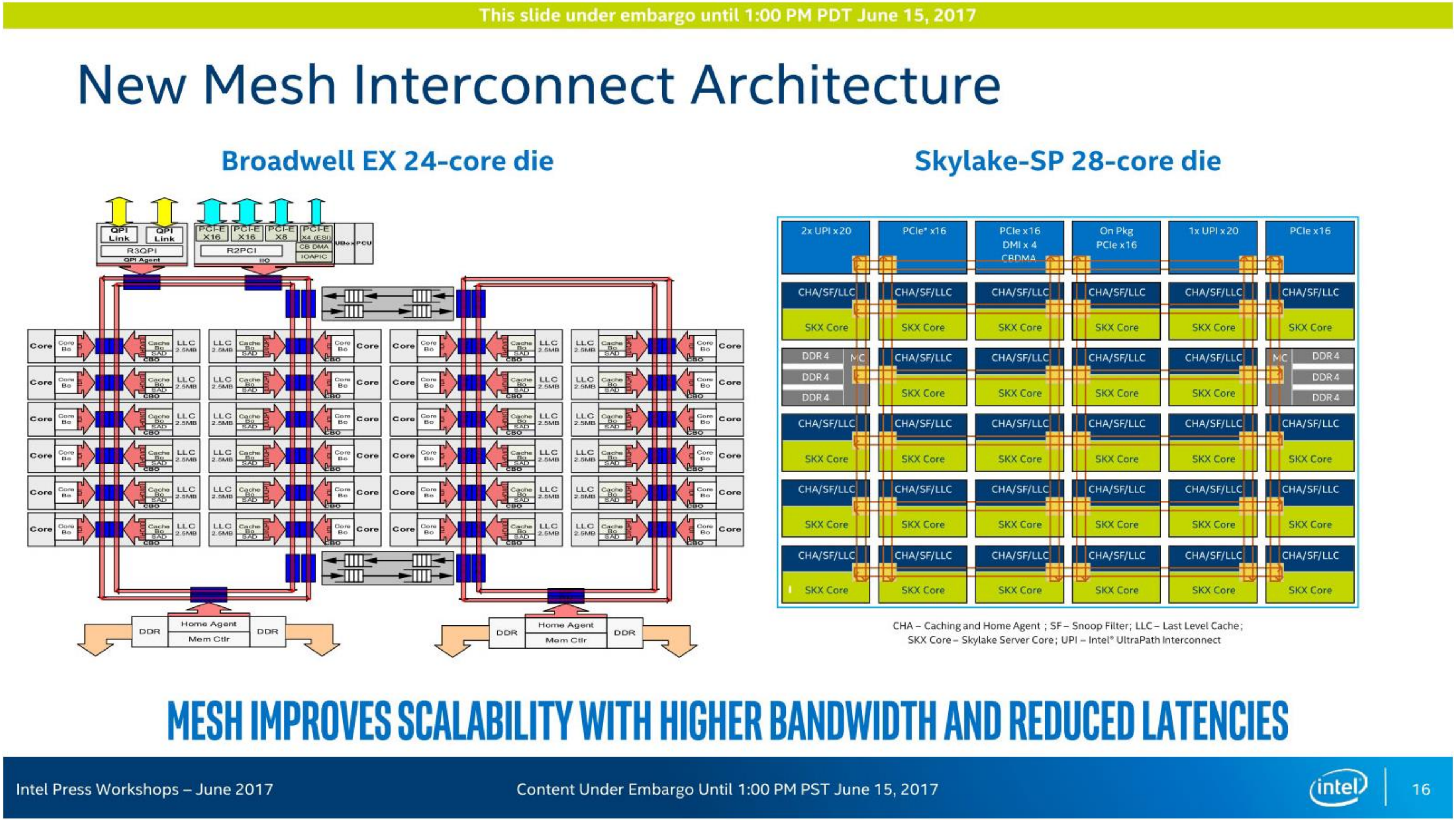
Table 5: Epiphany-V Area Breakdown

Intel Skylake-X, Core i9-7980XE (2017)

- 18 core
- 2D mesh topology



Intel Xeon Scalable Processor



This slide under embargo until 1:00 PM PDT June 15, 2017

New Mesh Interconnect Architecture

Broadwell EX 24-core die

Skylake-SP 28-core die

2x UPI x20	PCIe* x16	PCIe x16 DMI x 4 CBDMA	On Pkg PCIe x16	1x UPI x20	PCIe x16
CHA/SF/LLC SKX Core	CHA/SF/LLC SKX Core	CHA/SF/LLC SKX Core	CHA/SF/LLC SKX Core	CHA/SF/LLC SKX Core	CHA/SF/LLC SKX Core
DDR4 DDR4 DDR4 MC	CHA/SF/LLC SKX Core	CHA/SF/LLC SKX Core	CHA/SF/LLC SKX Core	CHA/SF/LLC SKX Core	MC DDR4 DDR4 DDR4
CHA/SF/LLC SKX Core	CHA/SF/LLC SKX Core	CHA/SF/LLC SKX Core	CHA/SF/LLC SKX Core	CHA/SF/LLC SKX Core	CHA/SF/LLC SKX Core
CHA/SF/LLC SKX Core	CHA/SF/LLC SKX Core	CHA/SF/LLC SKX Core	CHA/SF/LLC SKX Core	CHA/SF/LLC SKX Core	CHA/SF/LLC SKX Core
CHA/SF/LLC SKX Core	CHA/SF/LLC SKX Core	CHA/SF/LLC SKX Core	CHA/SF/LLC SKX Core	CHA/SF/LLC SKX Core	CHA/SF/LLC SKX Core

CHA – Caching and Home Agent ; SF – Snoop Filter; LLC – Last Level Cache;
SKX Core – Skylake Server Core; UPI – Intel® UltraPath Interconnect

MESH IMPROVES SCALABILITY WITH HIGHER BANDWIDTH AND REDUCED LATENCIES

Intel Press Workshops – June 2017 Content Under Embargo Until 1:00 PM PST June 15, 2017 intel 16

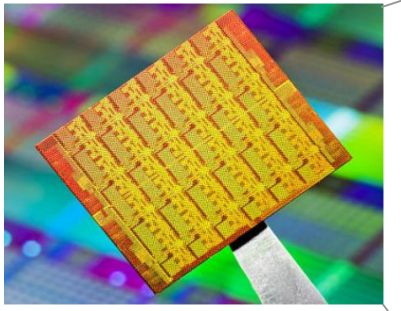
this slide is to be used as a whiteboard

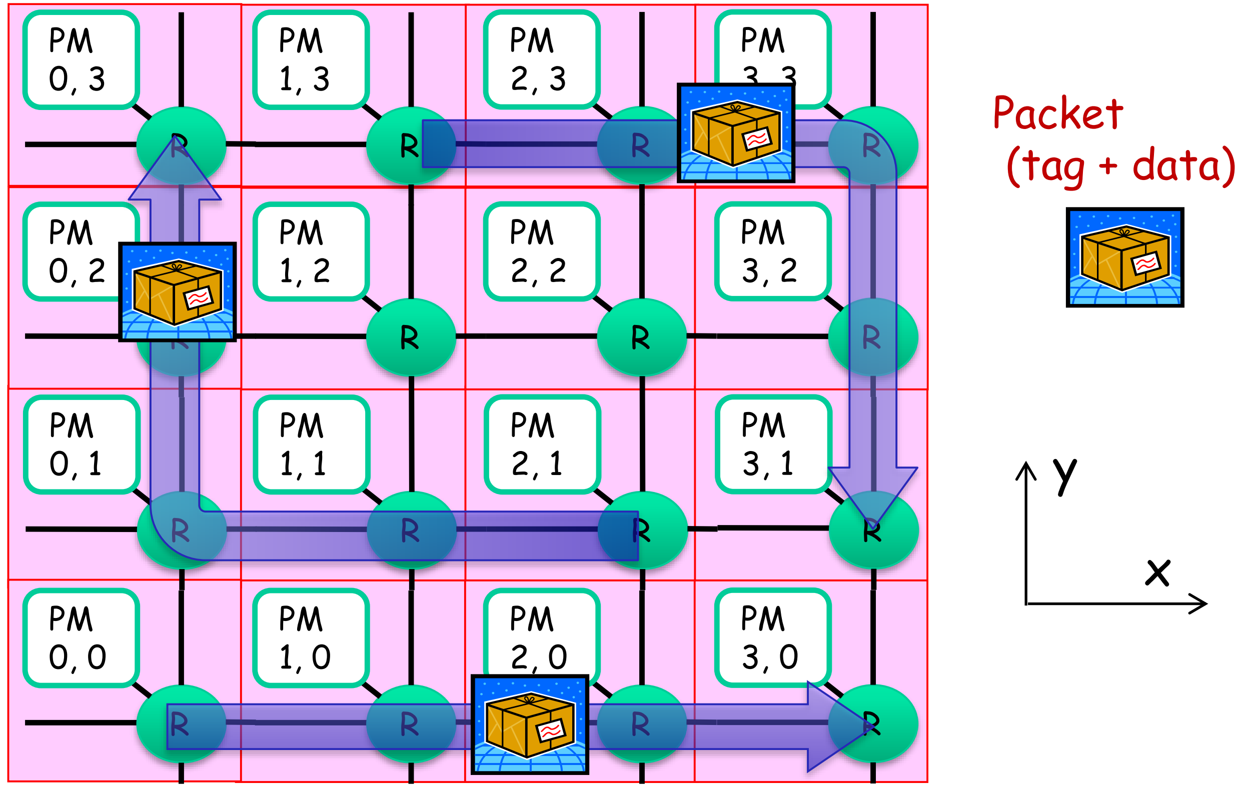
Bus vs. Networks on Chip (NoC) of mesh topology

intersection

Typical NoC architecture of mesh topology

- NoC requirements: low latency, high throughput, low cost
- Packet based data transmission via NoC routers and XY-dimension order routing



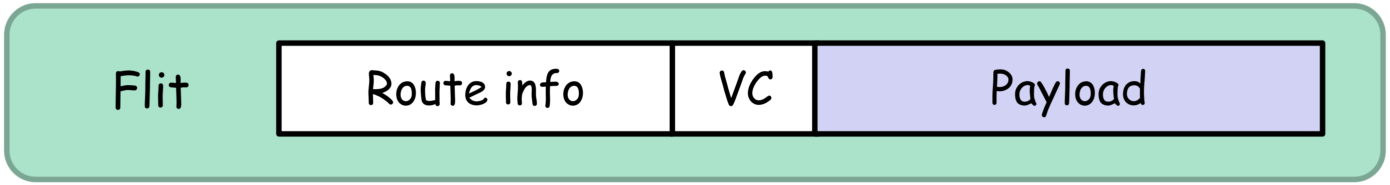


PM: Processing Module or Core,
R: Router

Packet organization (Flit encoding)

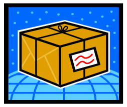
- A flit (flow control unit or flow control digit) is a link-level atomic piece that forms a network packet.
  - A packet has one head flit and some body flits.
- For simplicity, assume that a packet has only one flit.
  - Later we see a packet which has some flits.
- Each flit has typical three fields:
  - Payload (data)
  - Route information
  - Virtual channel identifier (VC)

Packet (tag + data)



Packet organization (Flit encoding)

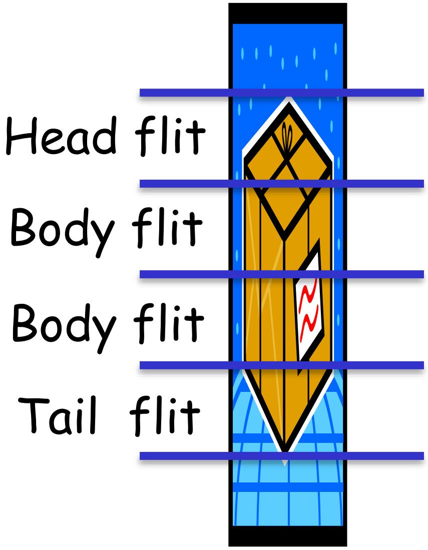
- A flit (flow control unit or flow control digit) is a link-level atomic piece that forms a network packet.
  - A packet has one head flit and some body flits.
- Each flit has typical three fields:
  - payload(data) or route information(tag)
  - flit type : head, body, tail, etc.
  - virtual channel identifier



Packet (tag + data)

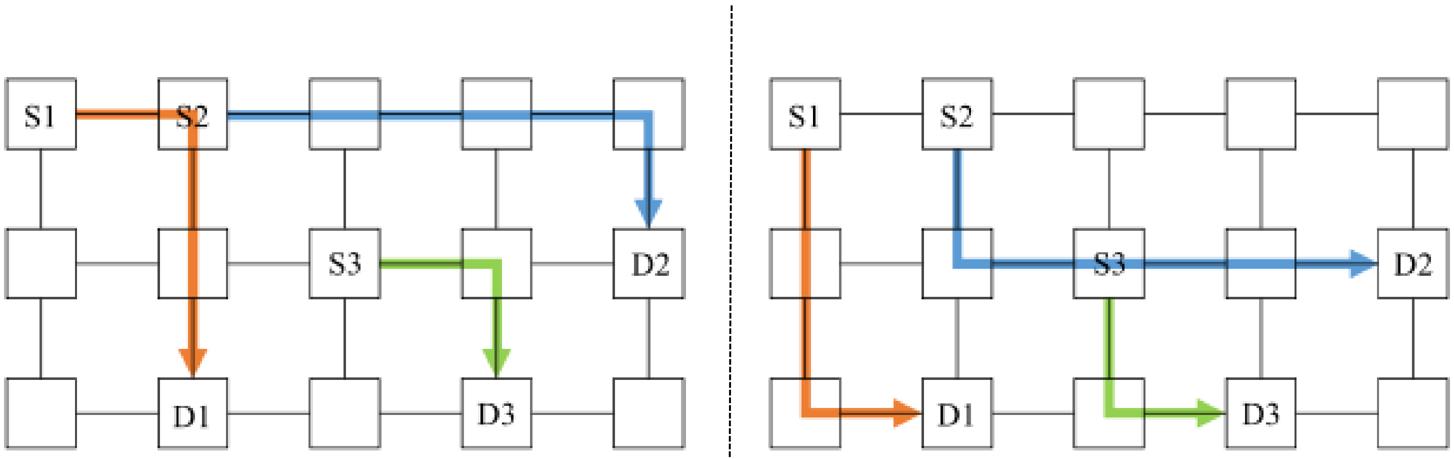
Head flit	VC	Type		Route info
Body flit	VC	Type	Payload	

Head and body flit formats



Routing

- XY dimension order routing (DOR), YX DOR

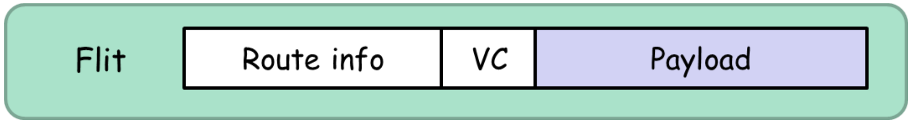


(a) XY routing

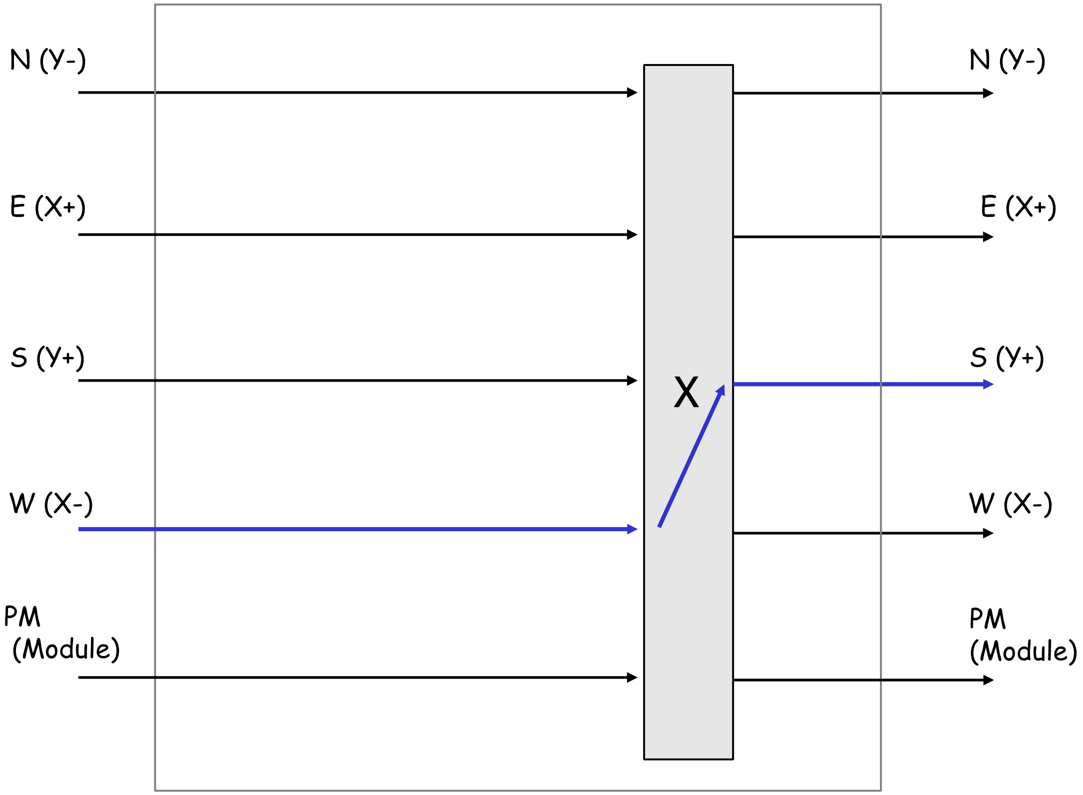
(b) YX routing

Simple NoC router architecture

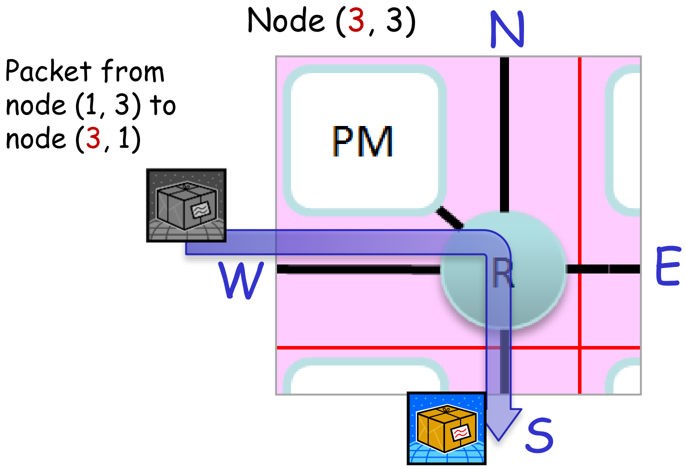
- Routing computation for XY-dimension order



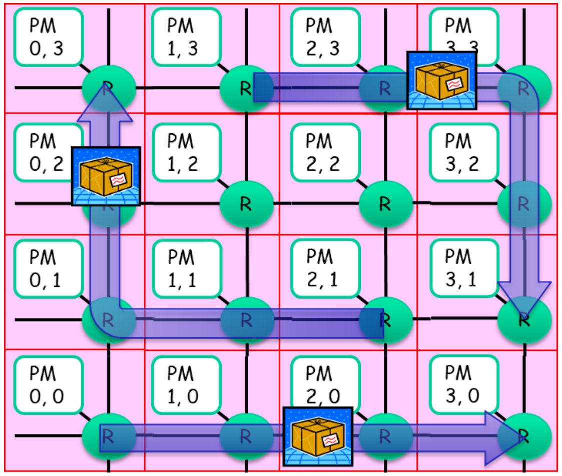
Flit | Route info | VC | Payload



NoC router

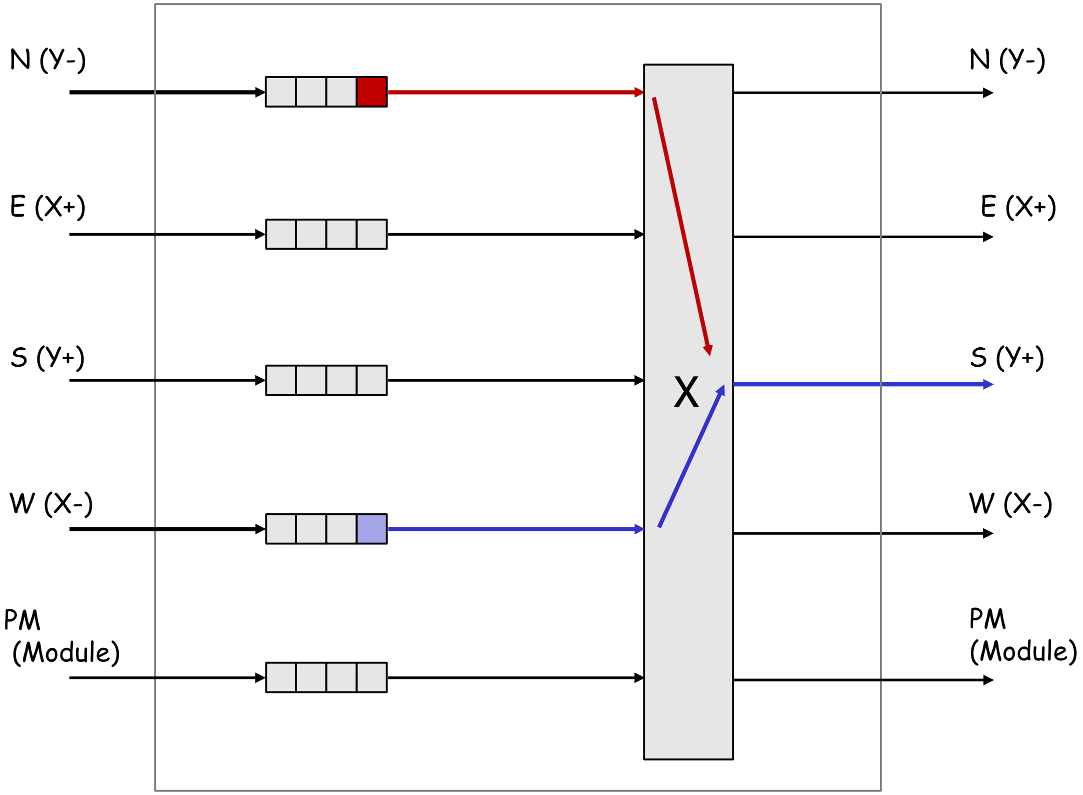


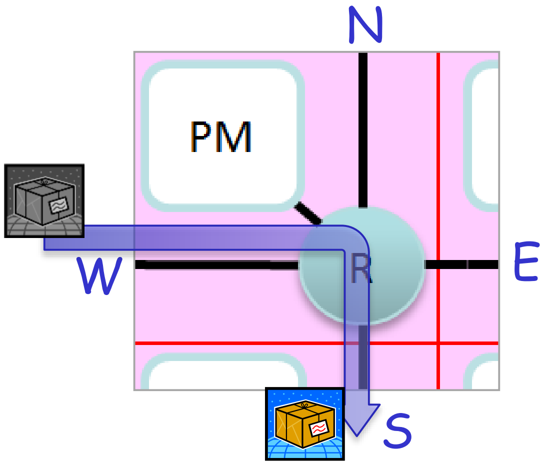
Packet from node (1, 3) to node (3, 1)



Simple NoC router architecture

- Buffering and arbitration
    - time stamp based, round robin, etc.

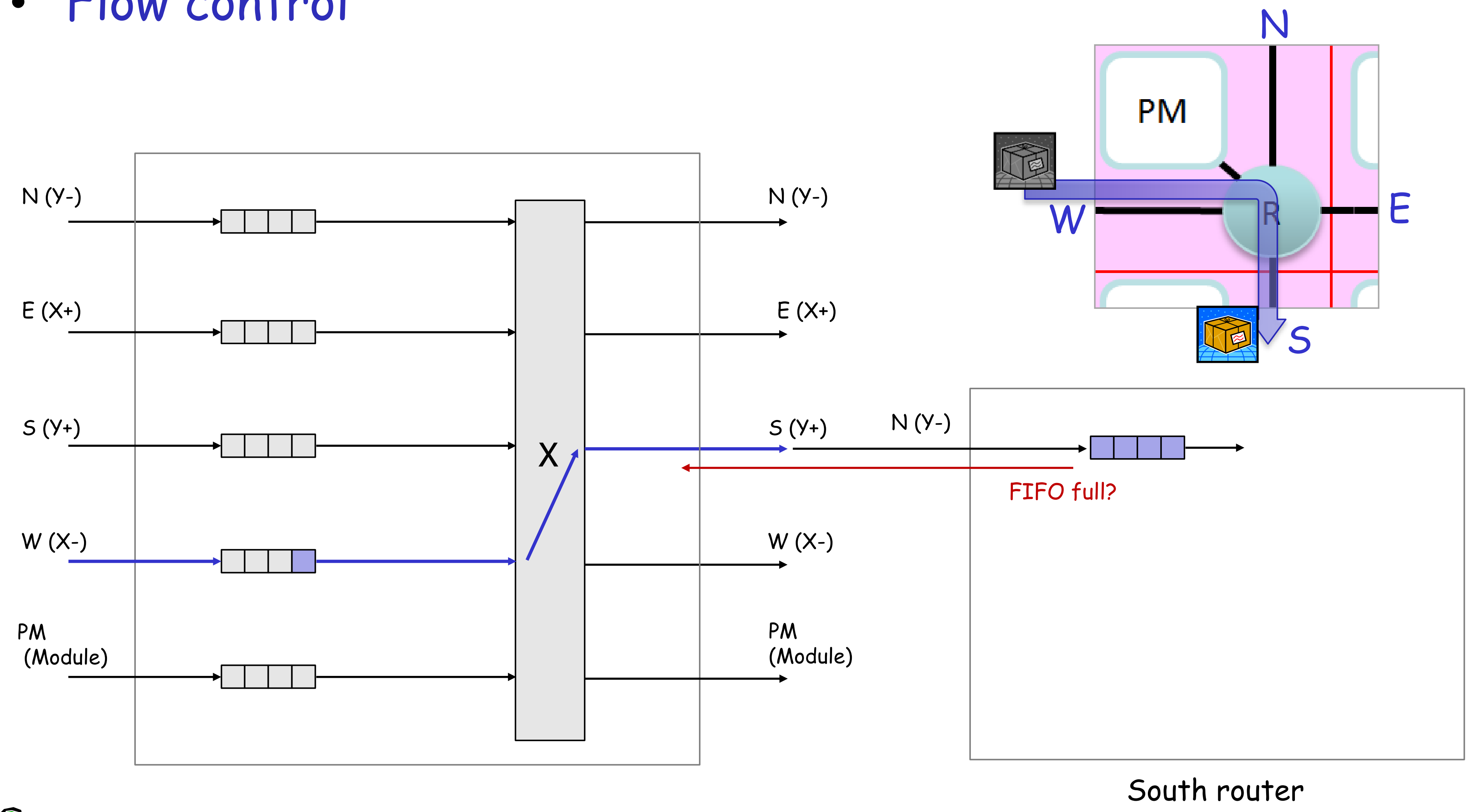




this slide is to be used as a whiteboard

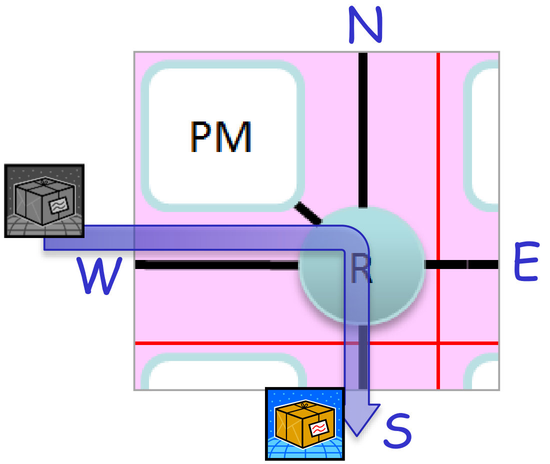
Simple NoC router architecture

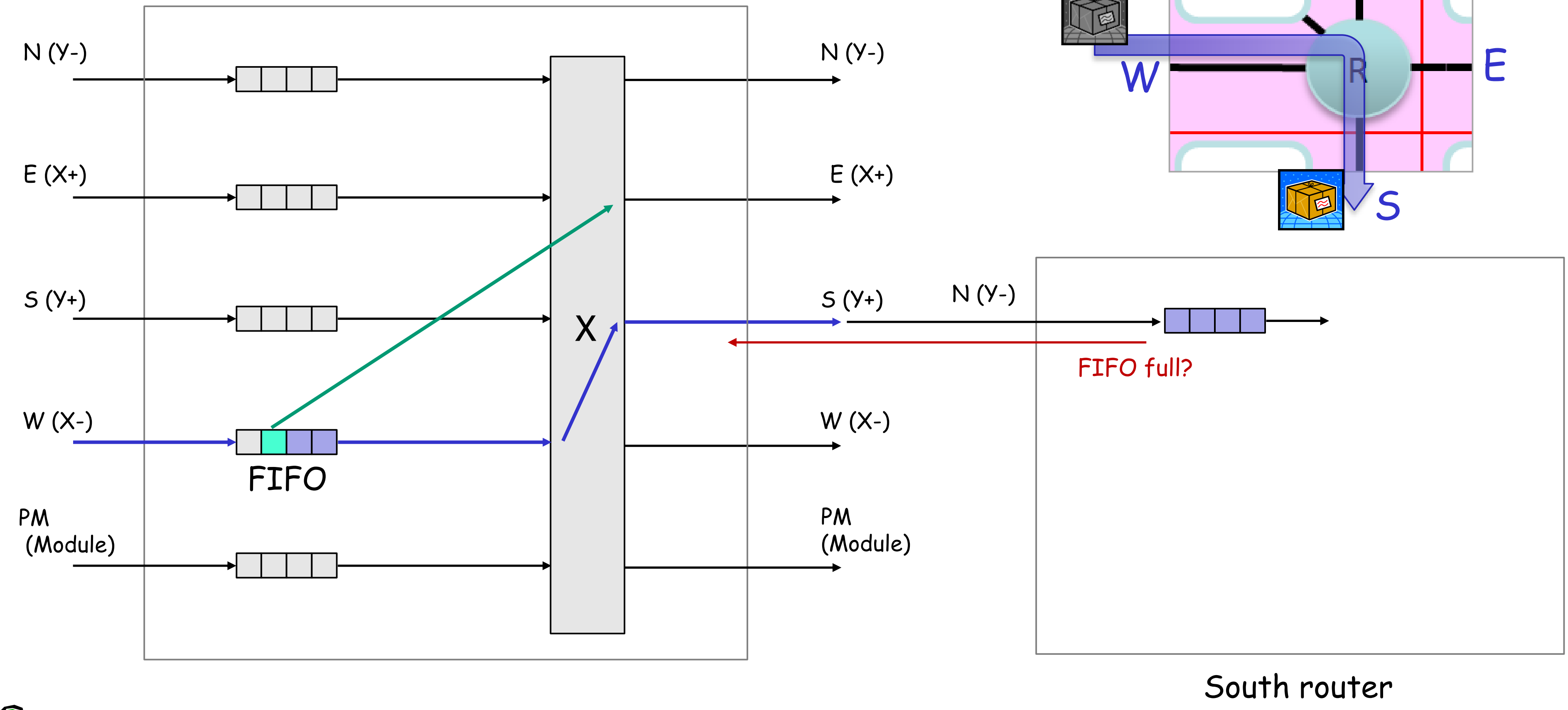
- Flow control



Simple NoC router architecture

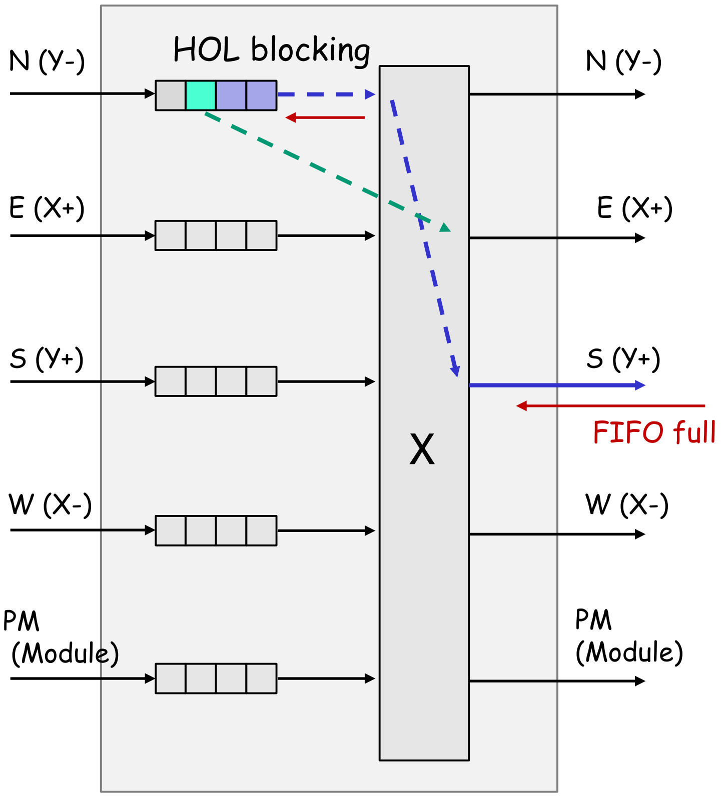
- Problem: Head-of-line (HOL) blocking



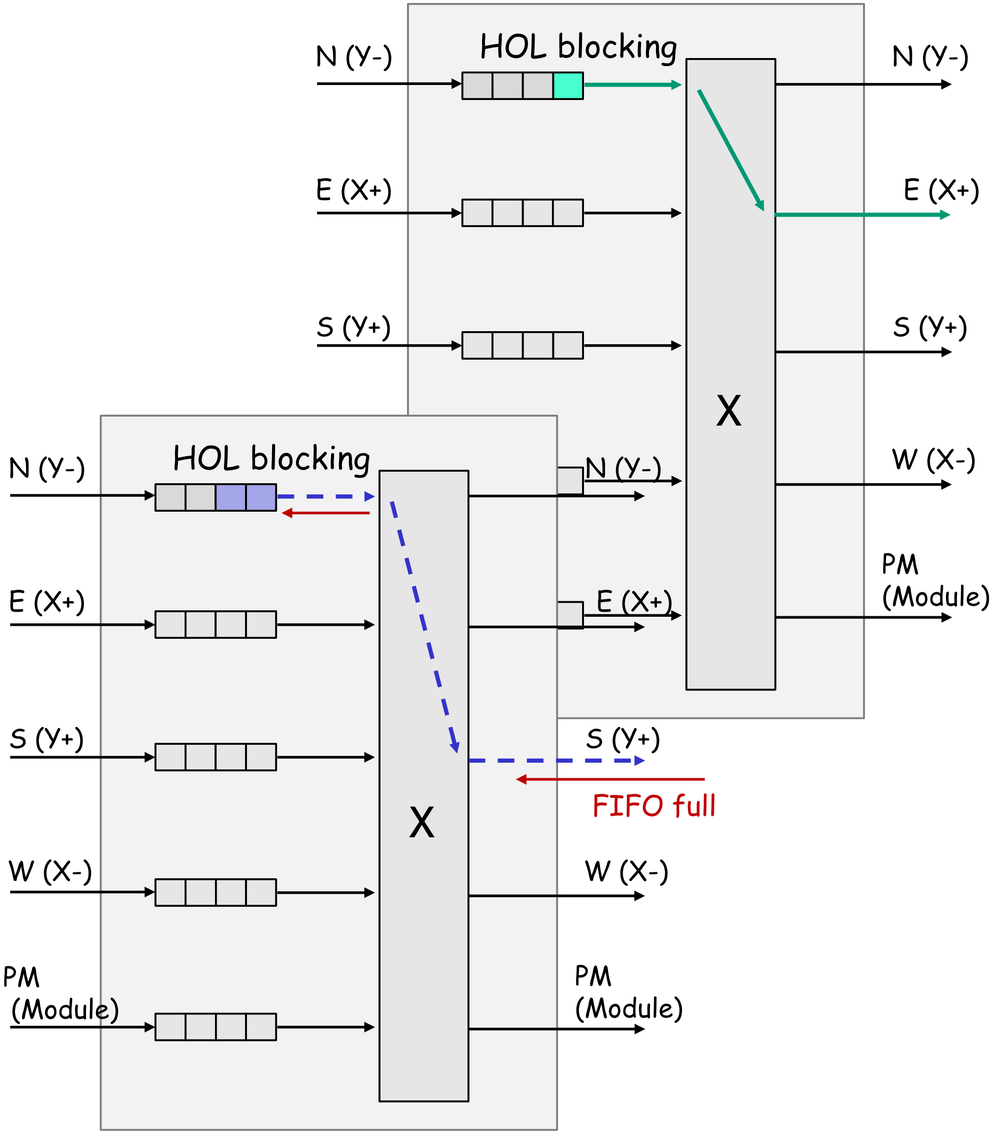


this slide is to be used as a whiteboard

Two (physical) networks to mitigate HOL ?



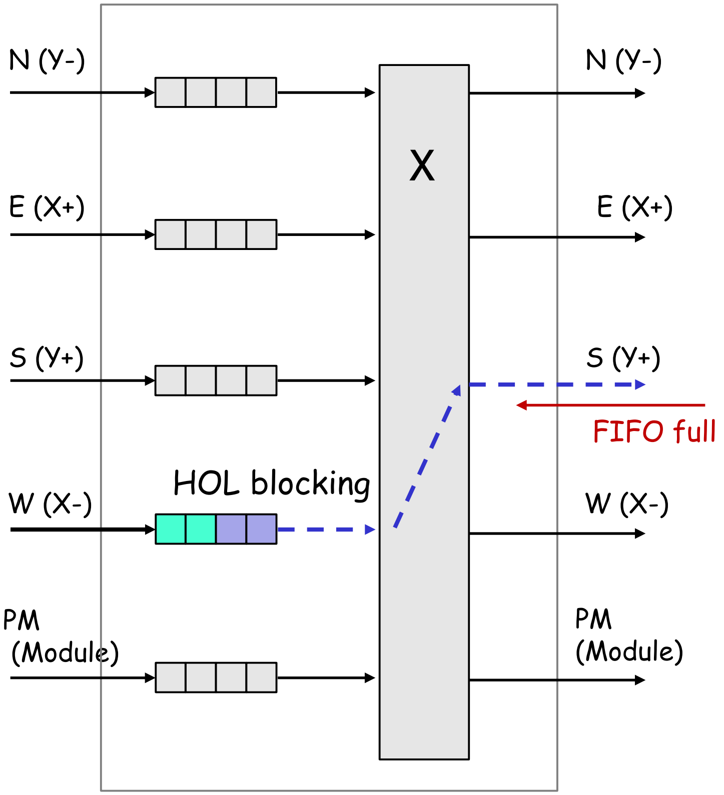
Simple NoC router



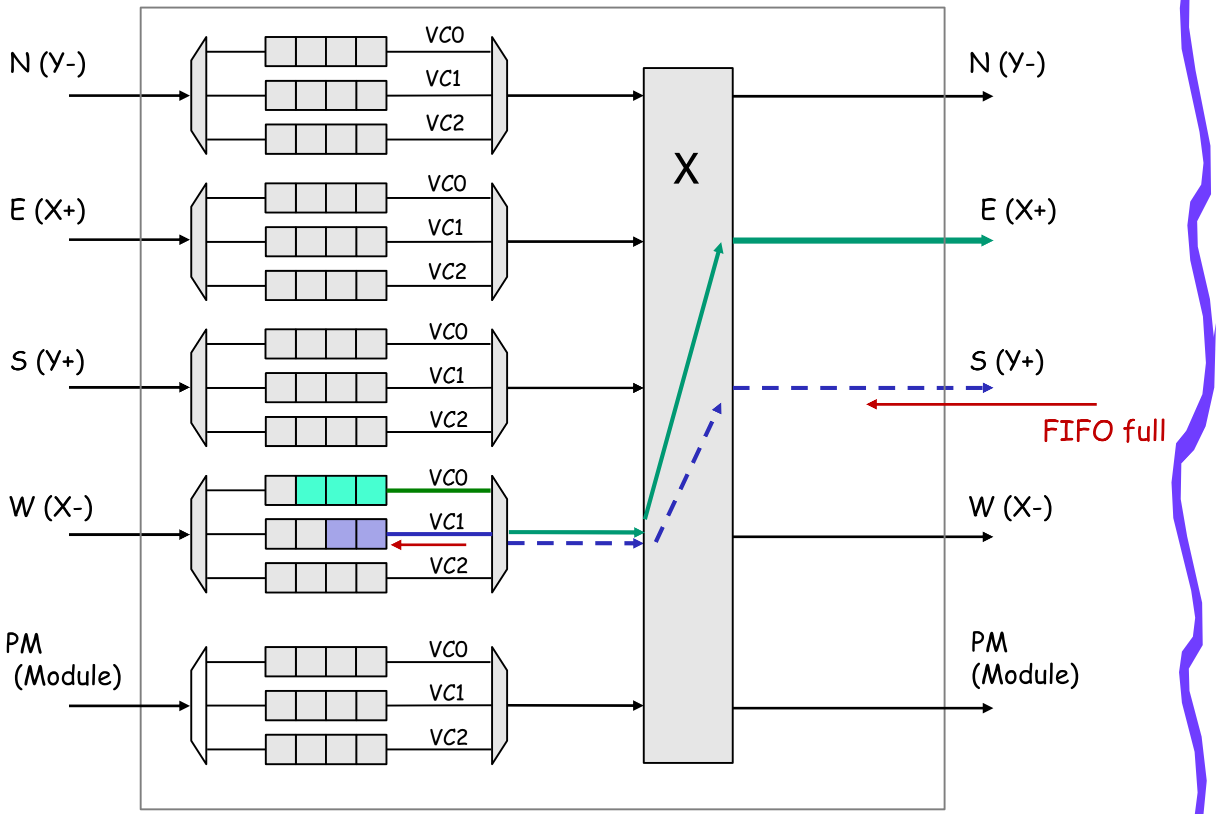
Datapath of Virtual Channel (VC) NoC router

- To mitigate head-of-line (HOL) blocking, virtual channels are used

Flit	Route info	VC	Payload



Simple NoC router



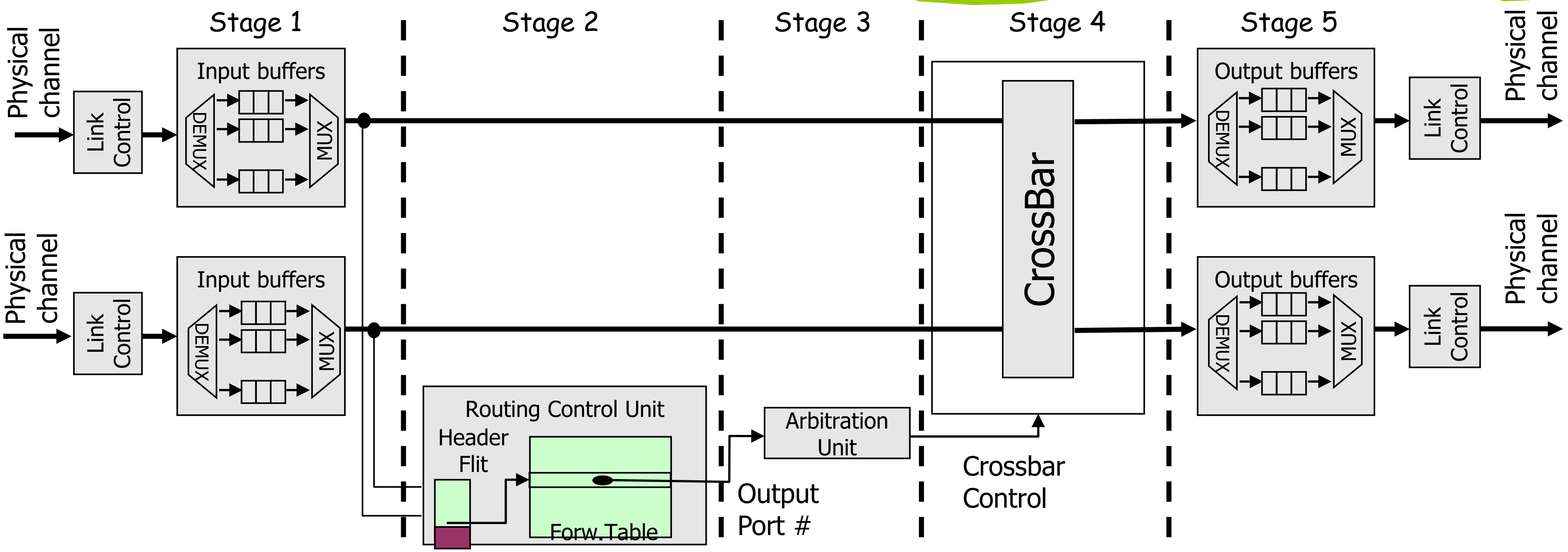
VC NoC router

Bus vs. Networks on Chip (NoC) of mesh topology

To mitigate
head-of-line (HOL) blocking

Virtual Channel

Pipelining the NoC router microarchitecture



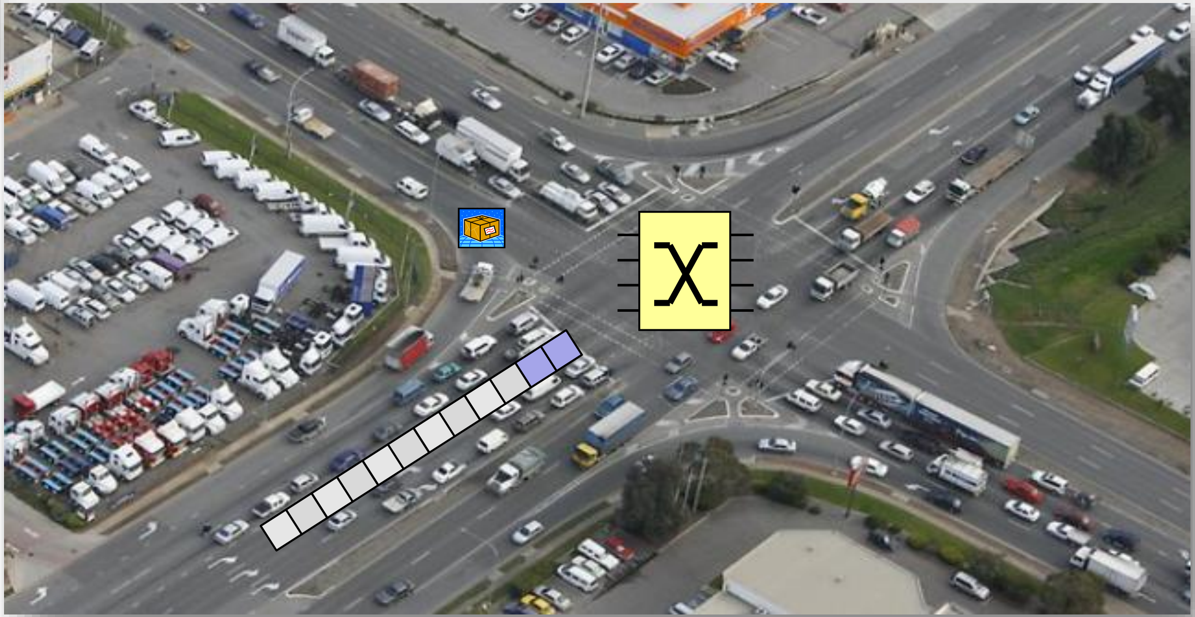
IB (Input Buffering)	RC (Route Computation)	SA (Switch Arb) - VCA (VC Arb) -	ST (Switch Trasv)	OB (Output Buffering)

Head flit	IB	RC	SA	ST	OB			
Body flit		IB	IB	IB	ST	OB		
Body flit			IB	IB	IB	ST	OB	
Body flit				IB	IB	IB	ST	OB

"A Delay Model and Speculative Architecture for Pipelined Routers," L. S. Peh and W. J. Dally, *Proc. of the 7th Int'l Symposium on High Performance Computer Architecture*, January, 2001.

Bus vs. Networks on Chip (NoC) of mesh topology

Packet
(tag + data)



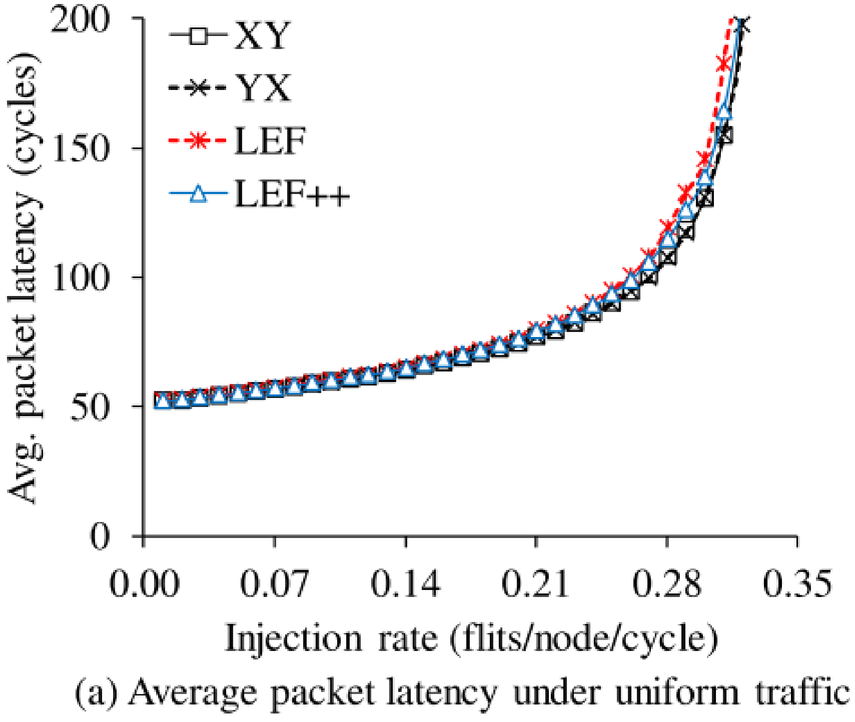
Distributed system

FIFO

intersection

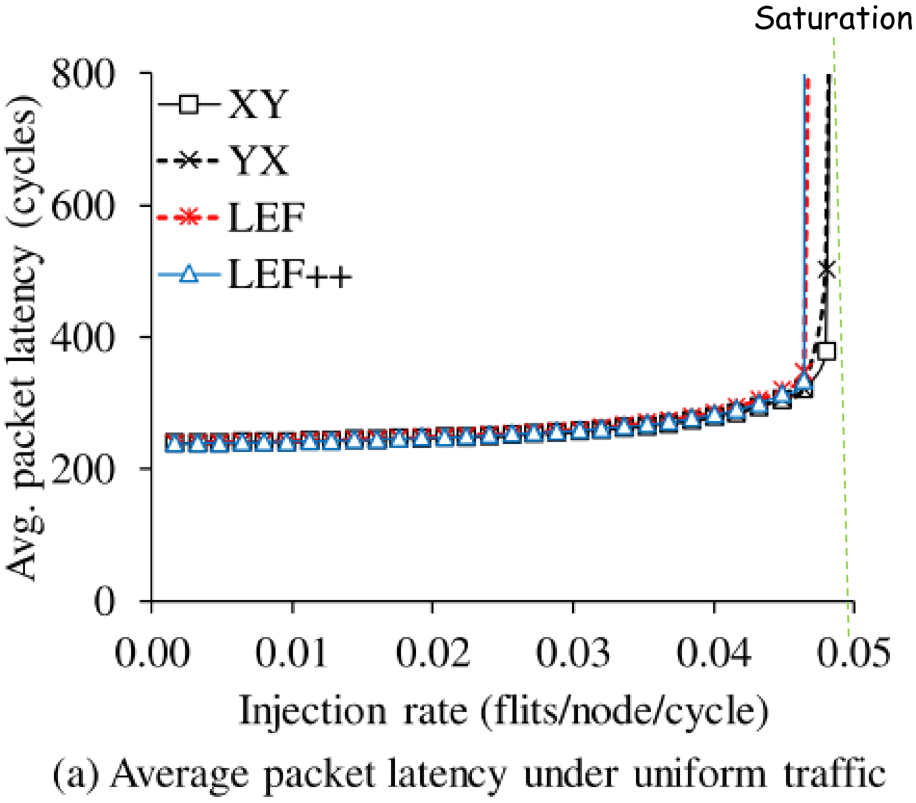
Average packet latency of mesh NoCs

- 5 stage router pipeline
- Uniform traffic (destination nodes are selected randomly)



(a) Average packet latency under uniform traffic

8x8 NoC



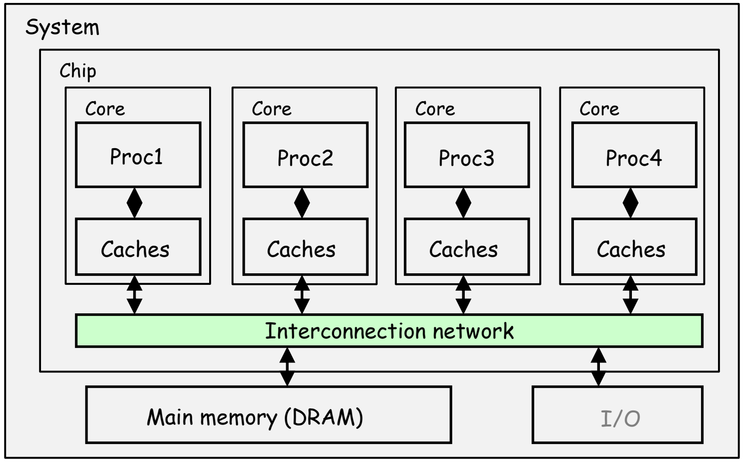
(a) Average packet latency under uniform traffic

64x64 NoC (4096 nodes)

Thiem Van Chu, Myeonggu Kang, Shi FA and Kenji Kise: Enhanced Long Edge First Routing Algorithm and Evaluation in Large-Scale Networks-on-Chip, IEEE 11th International Symposium on Embedded Multicore/Many-core Systems-on-Chip, (September 2017).

Key components of many-core processors

- Interconnection network
  - connecting many modules on a chip achieving high throughput and low latency
- Main memory and caches
  - Caches are used to reduce latency and to lower network traffic
  - A parallel program has private data and shared data
  - New issues are cache coherence and memory consistency
- Core
  - High-performance superscalar processor providing a hardware mechanism to support thread synchronization



this slide is to be used as a whiteboard