



Lecture 9: Torus Networks

Abhinav Bhatele, Department of Computer Science



UNIVERSITY OF
MARYLAND

Announcements

- Assignment 2 on OpenMP is online: due on October 7

Summary of last lecture

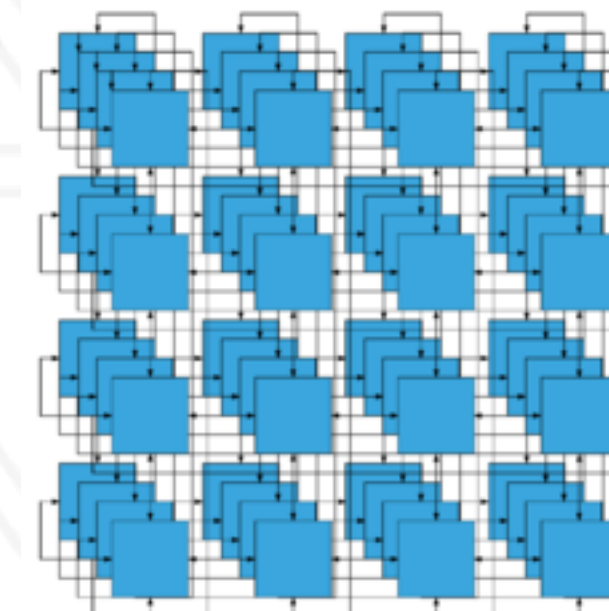
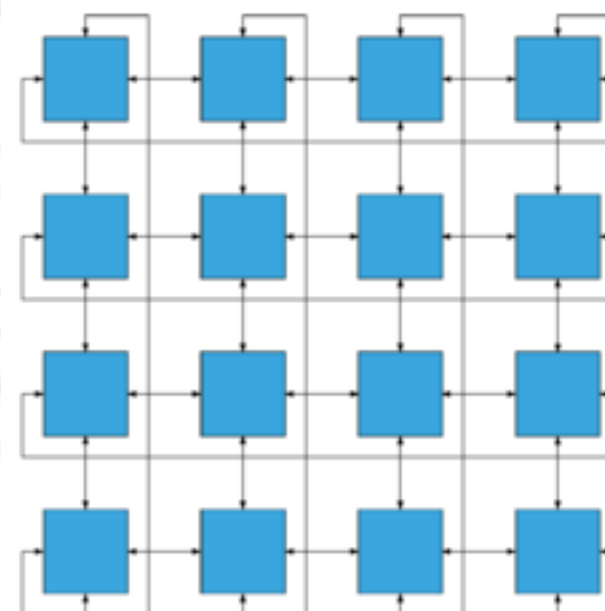
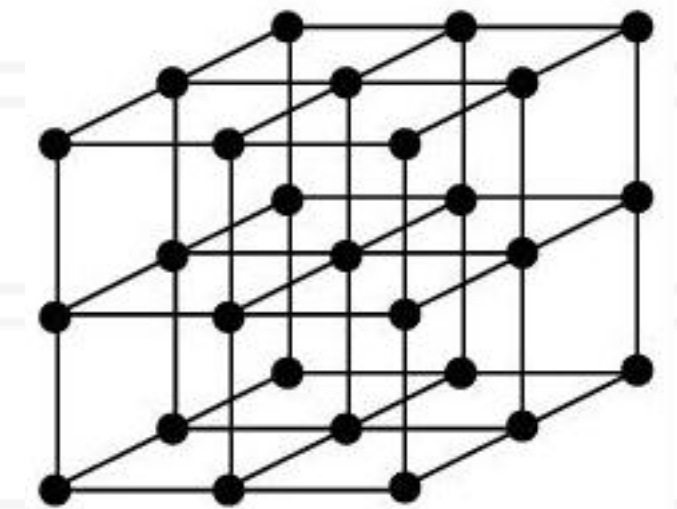
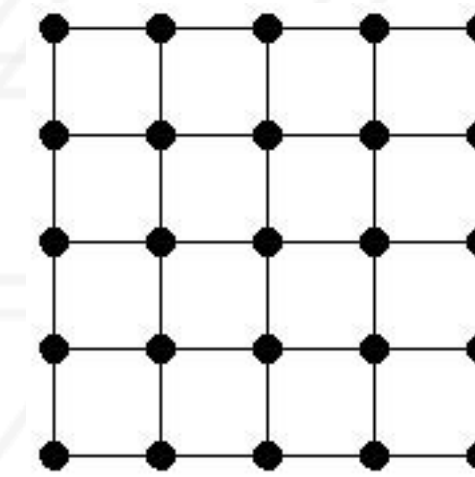
- Shared memory architectures
 - Distributed globally addressable memory
- SGI Origin and Altix series
- Directory-based protocol for cache coherence
- Used hypercube and fat-tree networks

HPC networks

- Key requirements: extremely low latency, high bandwidth
- Scalable: Adding more nodes shouldn't degrade network properties dramatically
 - Low network diameter, high bisection bandwidth
- Compute nodes connected together in many different logical topologies

n -dimensional Torus Networks

- Specific case of k -ary n -cube networks
 - k = number of nodes in each dimension, n = number of dimensions
- 2-dimensional mesh: k -ary 2-cube
- 3-dimensional mesh: k -ary 3-cube
- Torus networks: add wraparound links to the corresponding mesh network



https://en.wikipedia.org/wiki/Torus_interconnect

Routing protocols

- Minimal hop / shortest-path routing
- Static (dimension-ordered) or dynamic (follow path of least congestion)
- Switching techniques
 - Virtual cut-through, wormhole

Switching techniques: <http://pages.cs.wisc.edu/~tvrdik/7/html/Section7.html#AAAAABasic%20switching%20techniques>

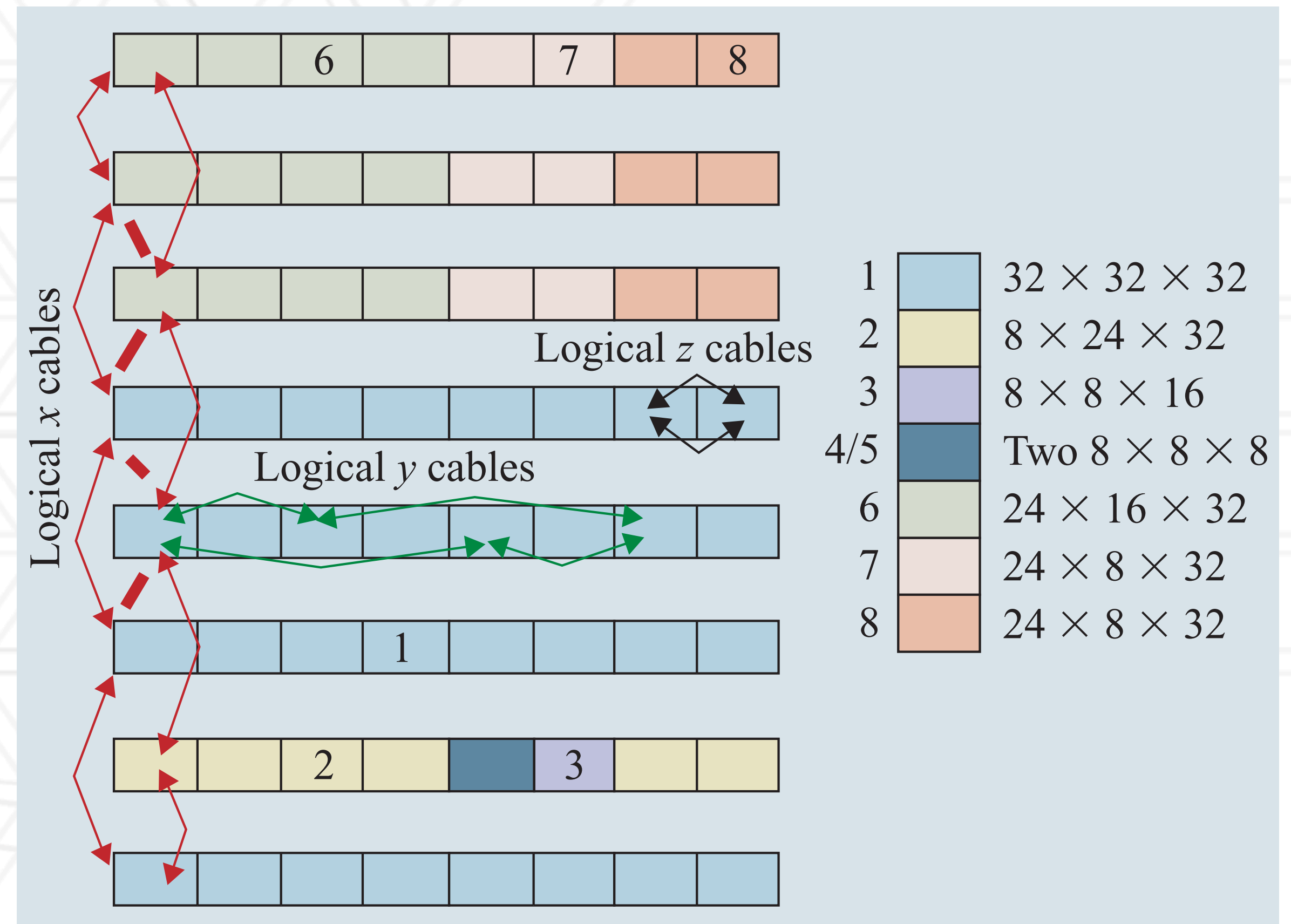
History of torus computers

- Cray T3D was launched in 1993
 - 300 MB/s of bandwidth in each direction
- Cray T3E, XT3/4/5 (SeaStar), XE6/XK7 (Gemini) - 3D tori
- IBM Blue Gene/L/P (3D torus)
- IBM Blue Gene/Q (5D torus with E dimension of size 2)
- Fujitsu Tofu interconnect (6D torus)

History: <https://www.extremetech.com/extreme/125271-the-history-of-supercomputers>

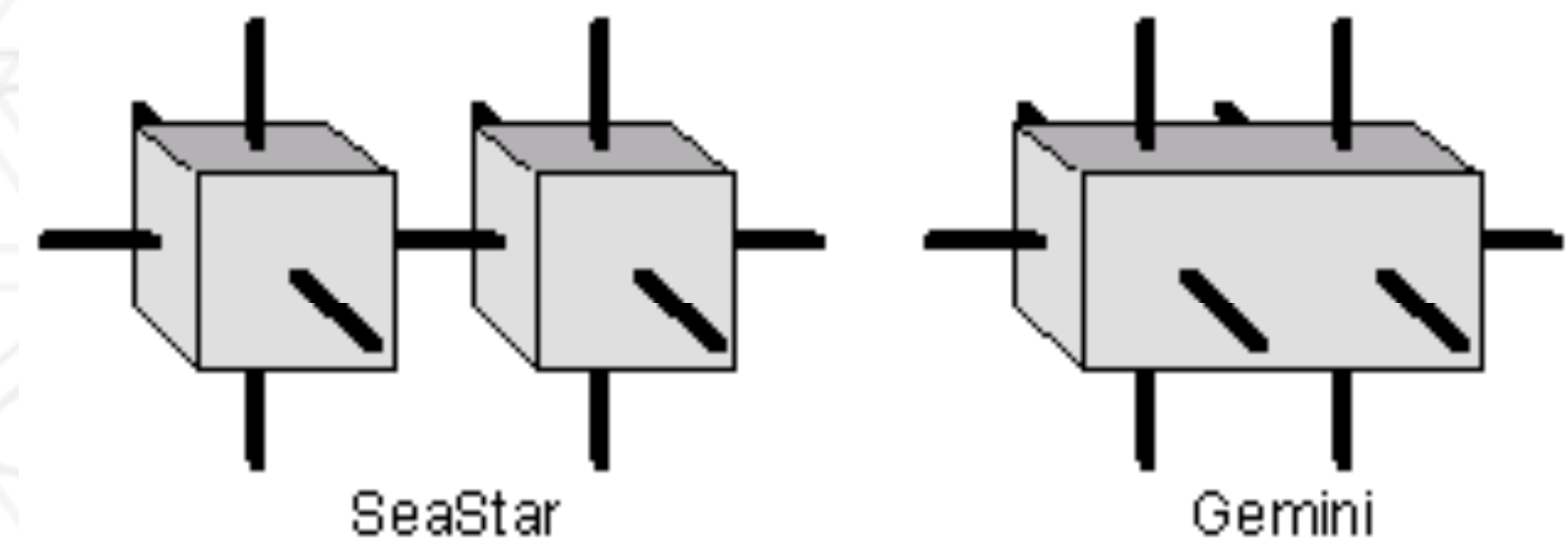
Blue Gene/L: Five networks

- 3-dimensional torus: $64 \times 32 \times 32 = 65,536$ nodes
 - Build block: 1 mid plane of $8 \times 8 \times 8$ nodes
- Collective network
 - Integer reductions, broadcast
- Barrier network
- Gigabit Ethernet
 - Parallel I/O
- Control system network (Ethernet)



Cray Gemini network

- Each Gemini router switch has 2 nodes attached to it
- 2 pairs of links in the X and Z dimensions, one in the Y dimension



Questions

Blue Gene/L torus interconnection network

- What are CRC codes?
- How do mesh network topologies deadlock? How does the bubble escape set of rules help?
- What does it mean to connect each rack with its next-to-nearest neighbor
- Why packets can be forwarded before being entirely received?

Questions

The Gemini System Interconnect

- Why did the Cray designers choose to have phits be composed of 24 bits? Does it have to do with the number of lanes in a link (3) sending a byte each?
- Does the Cray system use a CRC code to check integrity of the header phit?
- When do the costs of more complex headers for data transmission outweigh the bandwidth losses?
- Since Gemini supports global address space programming, does it mean that we can use some shared memory programming model on it?
- How is “transfer data directly between nodes without OS intervention” achieved?

Questions?



UNIVERSITY OF
MARYLAND

Abhinav Bhatele

5218 Brendan Iribe Center (IRB) / College Park, MD 20742

phone: 301.405.4507 / e-mail: bhatele@cs.umd.edu