Multicore Computing
Lecture15 - Interconnect

남 범 석

bnam@skku.edu

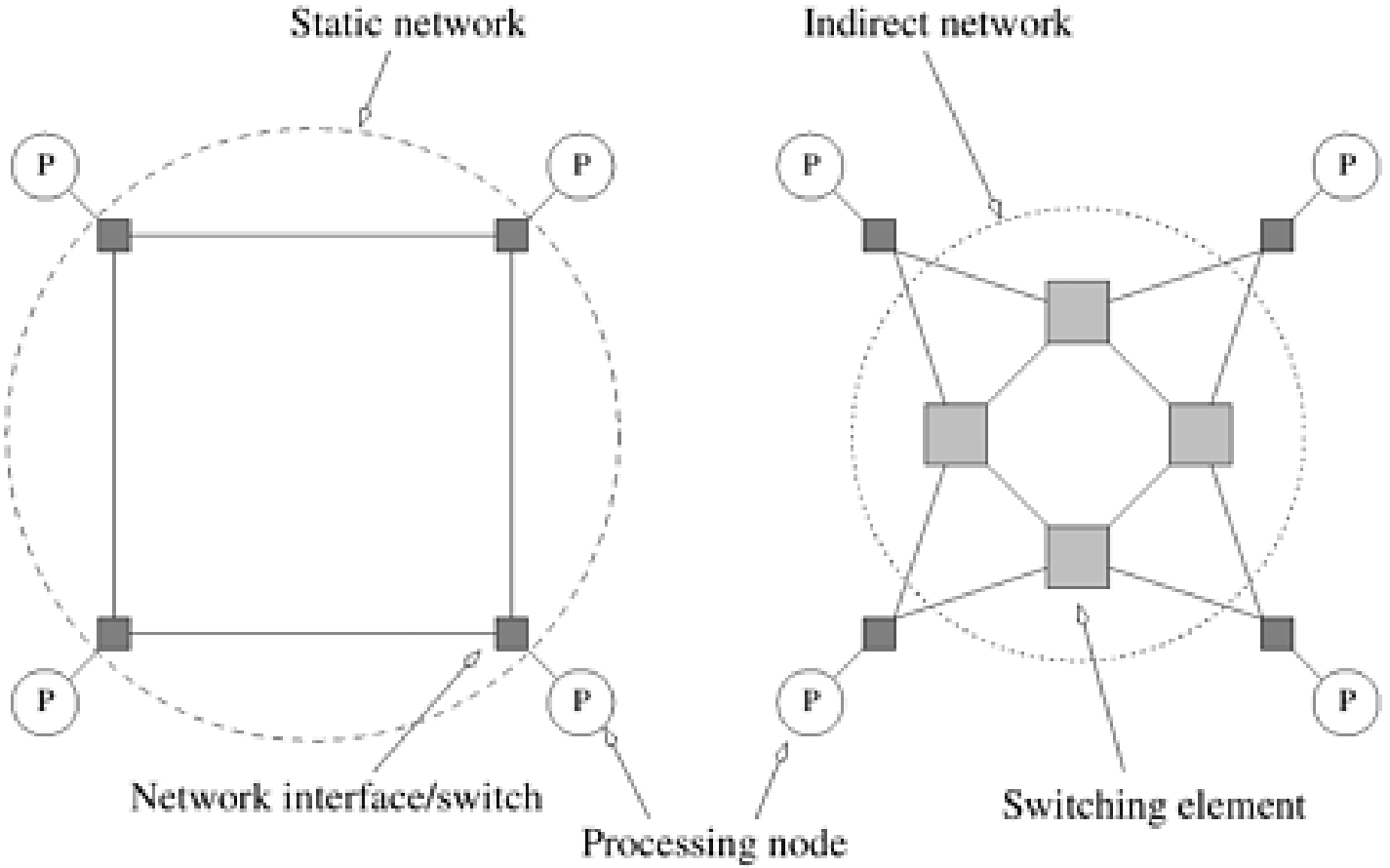
Clusters (Alternative to SMP)

- A group of interconnected whole computers (or SMP's)
- Working together as unified resource
- Illusion of being one machine
- Each computer called a node

- Benefits:
  - Absolute scalability
  - Incremental scalability
  - High availability
  - Superior performance/price

Static vs. Switched Networks

- Static networks: direct connections between compute nodes
- Switched networks: switch is used between compute nodes



What is inside a switch? Crossbar and Multistage Connections

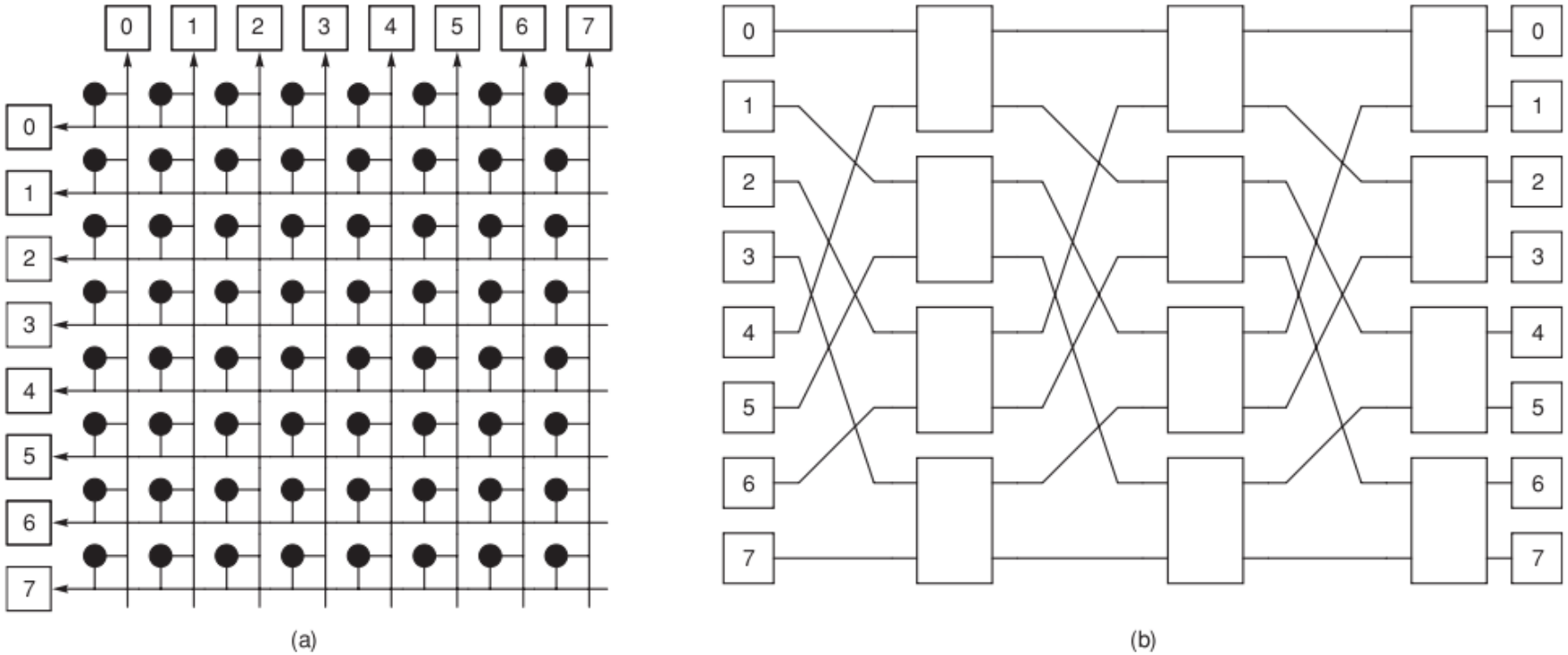
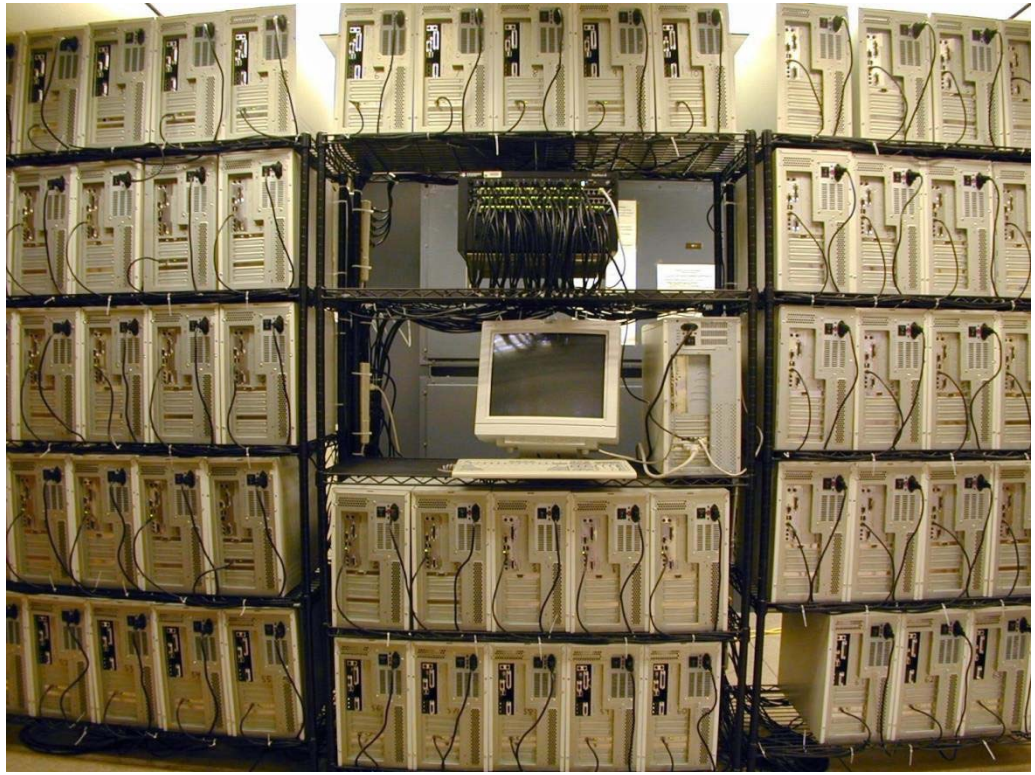


Figure E.11 Popular centralized switched networks: (a) the crossbar network requires N^2 crosspoint switches, shown as black dots; (b) the Omega, a MIN, requires $N/2 \log_2 N$ switches, shown as vertical rectangles. End node devices are shown as numbered squares (total of eight). Links are unidirectional—data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

Hennessy and Patterson

How would you build a cluster computer?

- Allowed components:
  - network cards
  - switches
  - cables
- Propose how to connect 64 nodes with any resources, but only 16-way switches are available



Shopping for a 96-port switch

Cisco Nexus 93128TX Layer 3 Switch

by Cisco

Be the first to review this item

List Price: ~~$26,000.00~~

Price: $23,941.08 + $99.90 shipping

You Save: $2,058.92 (8%)

Note: Not eligible for Amazon Prime.

Usually ships within 6 to 10 days.

Ships from and sold by Get724.

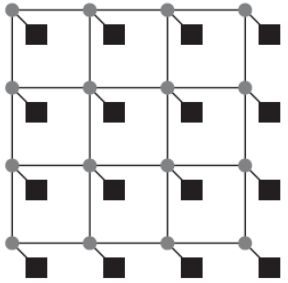
Estimated Delivery Date: Feb. 17 - 25 when you choose Expedited at checkout.

- Device Type: Layer 3 Switch
- Form Factor: Rack-mountable

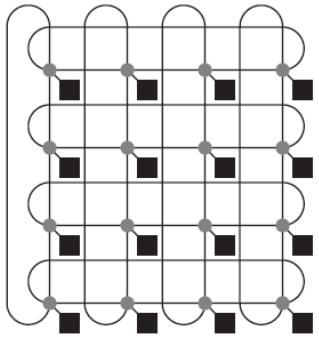
Some Network Topologies

- Trade-off between cost and performance
- Match the topology with the application

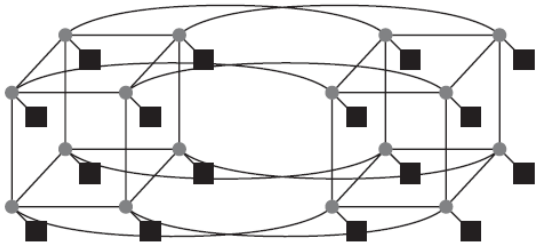
- Ring
- Mesh/Torus
  - Good for applications using nearest-neighbor communication
  - Prevalent for proprietary interconnects
- Hypercube
- Fat tree (uses switches)
  - Popular for commodity clusters
- Dragonfly (switch attached to node)
  - Low diameter network



(a) 2D grid or mesh of 16 nodes

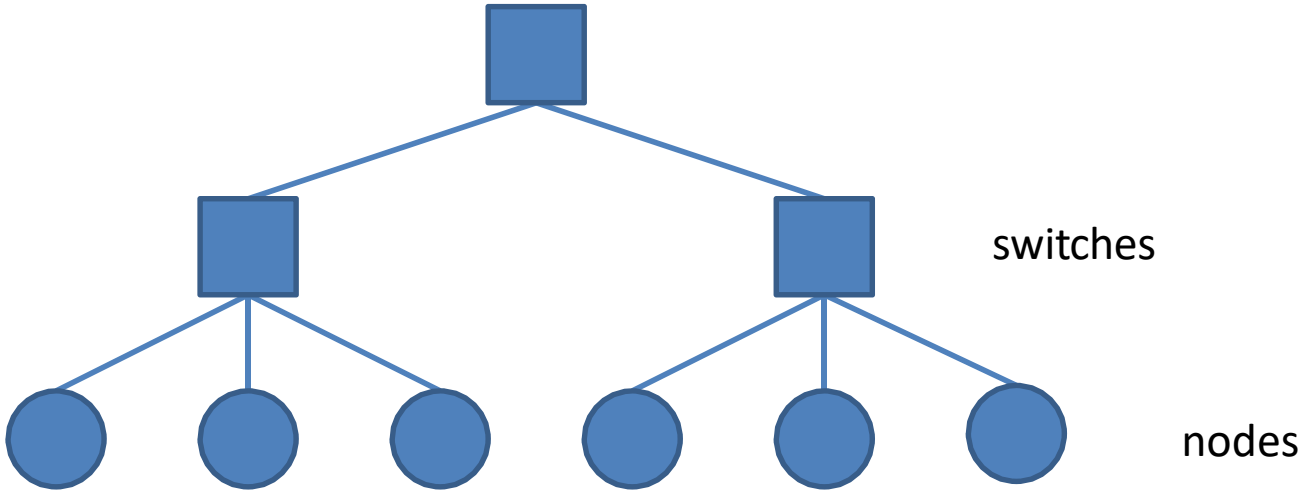


(b) 2D torus of 16 nodes



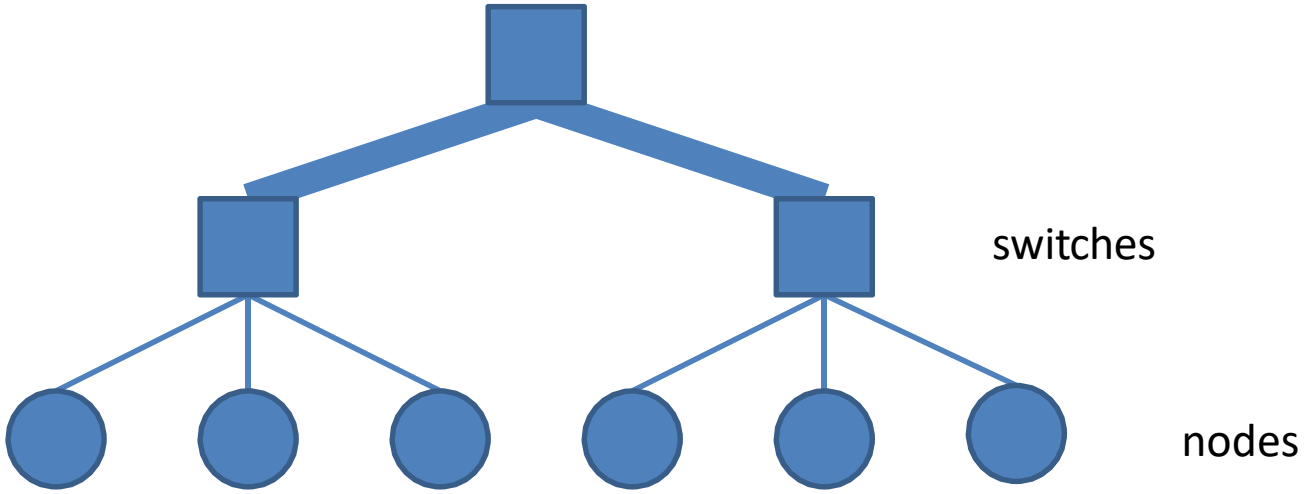
(c) Hypercube of 16 nodes ($16 = 2^4$ so $n = 4$)

Trees and Fat Trees



- Tree (above figure)

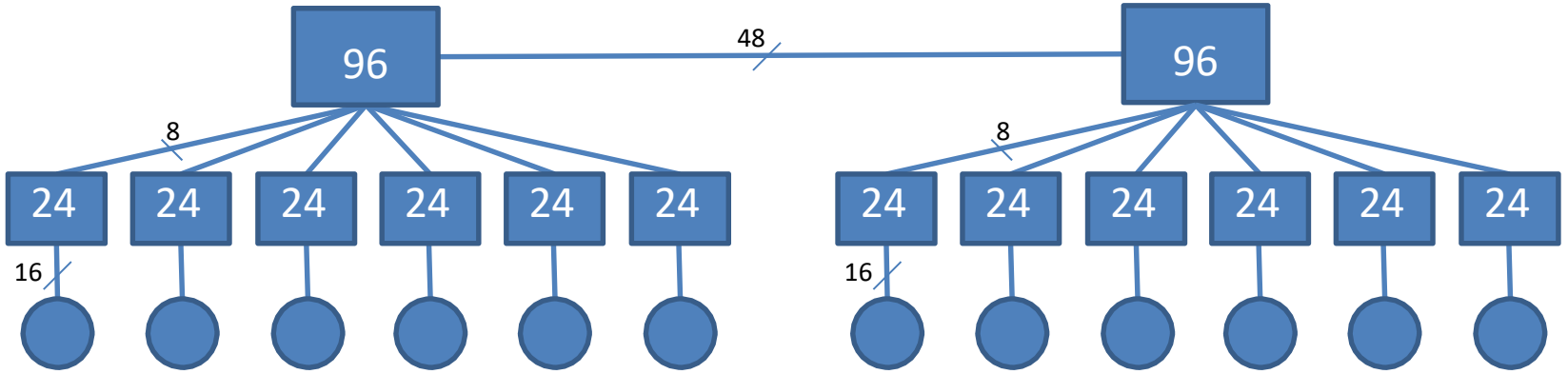
Trees and Fat Trees



- Fat Tree: additional links to increase the bandwidth near the root of the tree

Fat Tree Challenge

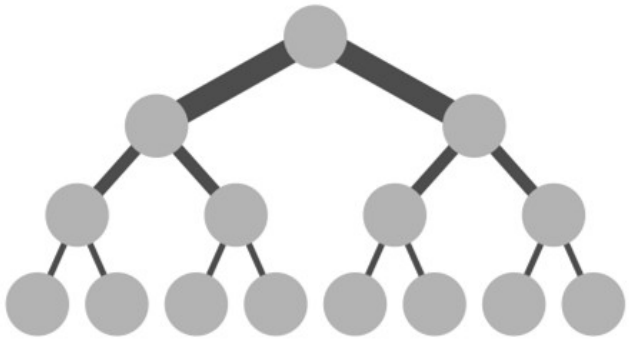
- Build a 192 node fat tree cluster with two 96-way switches and any number of 24-way switches.



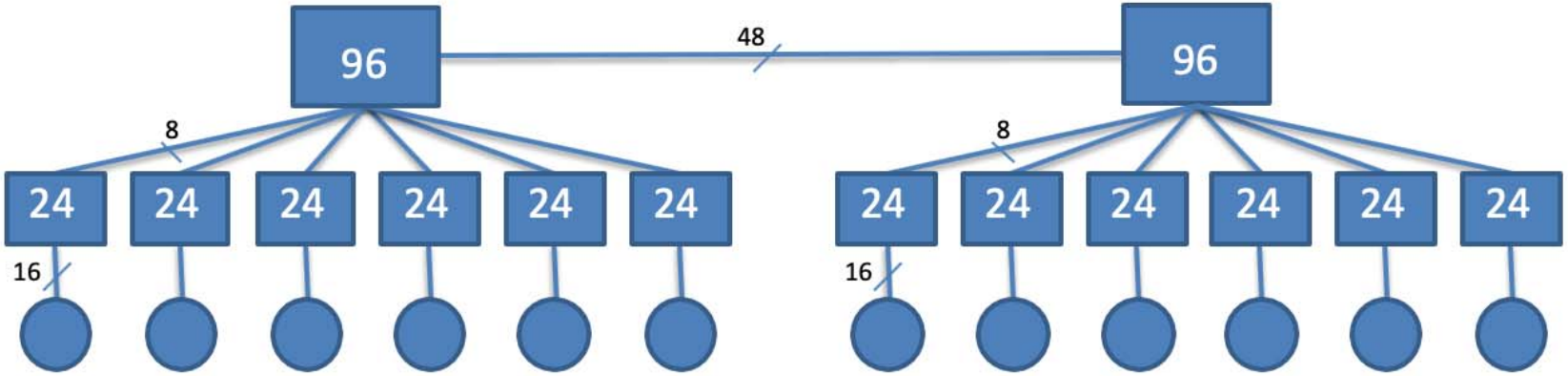
Possible solution

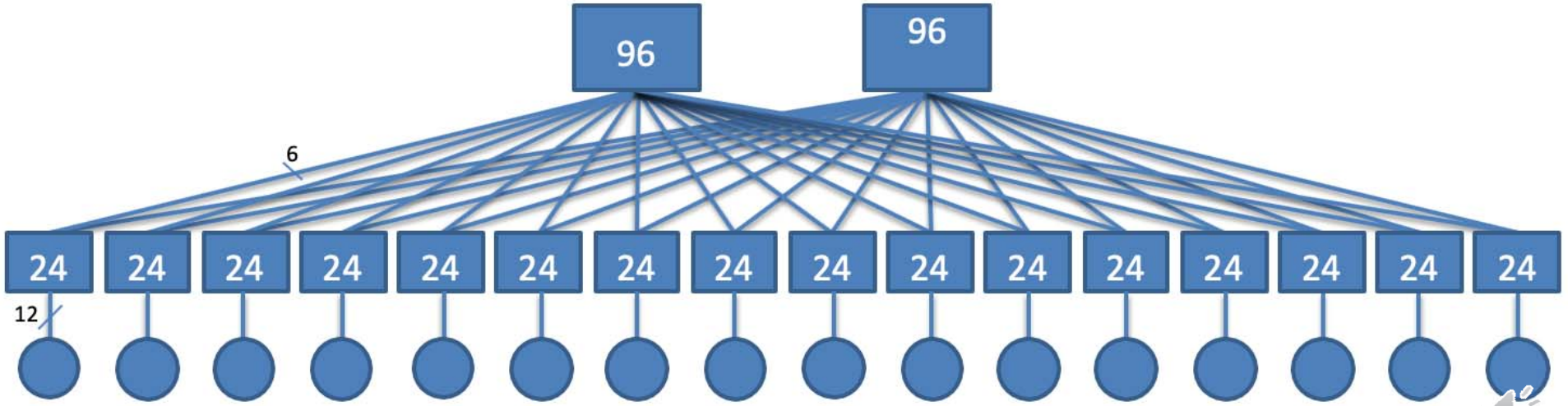
Network Properties

- Diameter = max number of hops between two nodes
- Bisection width = minimum number of links cut to divide the network into two halves
- Bisection bandwidth = bandwidth of above
- Full bisection bandwidth allows one half of nodes to communicate simultaneously with other half of nodes
  - Suppose a half of the nodes inject data into the network at a rate B Mbps. When the bisection bandwidth is B then the network has **full bisection bandwidth**.



What is the bisection bandwidth in these two examples?





Network Properties

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.

One port per node; nodes attached to switches.
Hennessy and Patterson, 2007.

InfiniBand

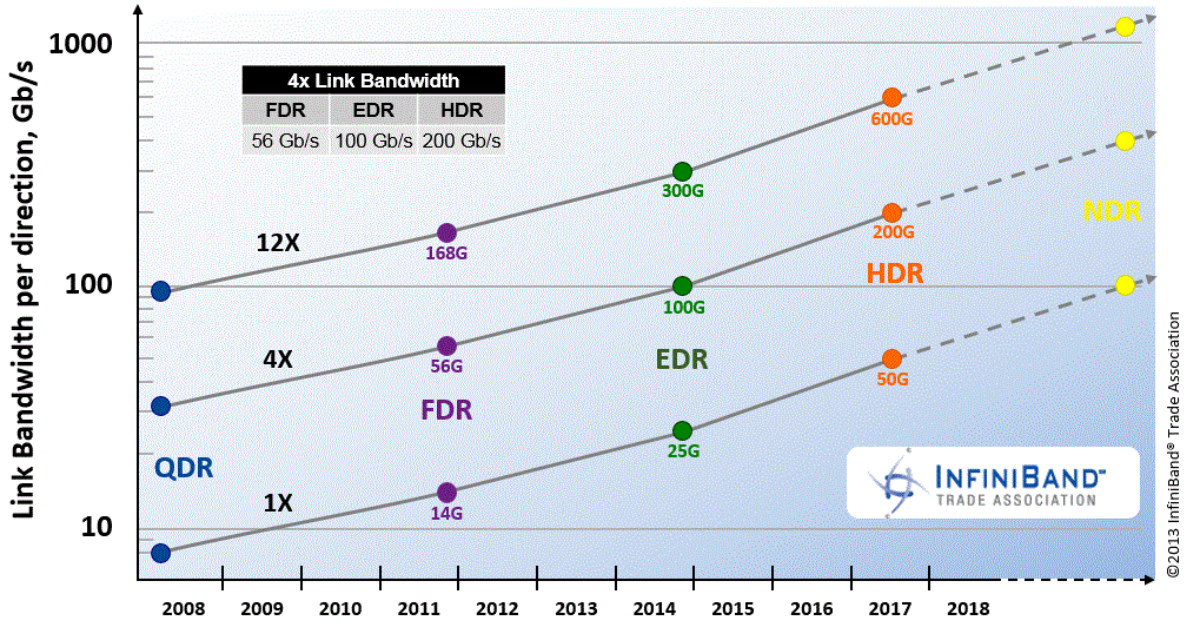
- Was originally designed as a “system area network”: connecting CPUs and I/O devices.
  - A larger role: replaceing all I/O standards for data centers: PCI, Fibre Channel, and Ethernet: everything connects through InfiniBand.
  - A less role: Low latency, high bandwidth, low overhead interconnect for commercial datacenters between servers and storage.
- Can form local area or even large area networks.
- Has become the de-facto interconnect for high performance clusters (100+ systems in top 500 supercomputer list).

InfiniBand Bandwidth

- Link speed (signal rate):
  - Single data rate (SDR): 2.5Gbps (1X), 10Gbps (4X), and 30Gbps (12X).
  - Double data rate (DDR): 5Gbps (1X), 20 Gbps (4X), 60Gbps(12X)
  - Quad data rate (QDR): 10Gbps (1X), 40Gbps(4X), 120Gbps(12X)
  - Fourteen data rate (FDR): 14Gbps(1X), 56Gbps(4X), 168Gbps(12X)
  - Enhanced data rate (EDR): 25Gbps(1X), 100Gbps(4X), 300Gbps(12X)

- 8b/10b enconding in SDR, DDR, and QDR
- 64b/66b enconding in FDR and EDR

InfiniBand link speed



Infiniband Roadmap from InfiniBand trade association http://www.infinibandta.org/content/pages.php?pg=technology_overview

Layer architecture: somewhat similar to TCP/IP

- Physical layer
- Link layer
  - Error detection (CRC checksum)
  - flow control (credit based)
  - switching, virtual lanes (VL),
  - forwarding table computed by subnet manager
    - Not adaptive
- Network layer: across subnets.
  - No use for the cluster environment
- Transport layer
  - Reliable/unreliable, connection/datagram
- Verbs: interface between adaptors and OS/Users

Verbs

- OS/Users access the adaptor through verbs
- Communication mechanism: Queue Pair (QP)
  - Users can queue up a set of instructions that the hardware executes.
  - A pair of queues in each QP: one for send, one for receive.
  - Users can post send requests to the send queue and receive requests to the receive queue.
  - Three types of send operations: SEND, RDMA-(WRITE, READ, ATOMIC), MEMORY-BINDING
  - One receive operation (matching SEND)

Send Queue/Receive Queue

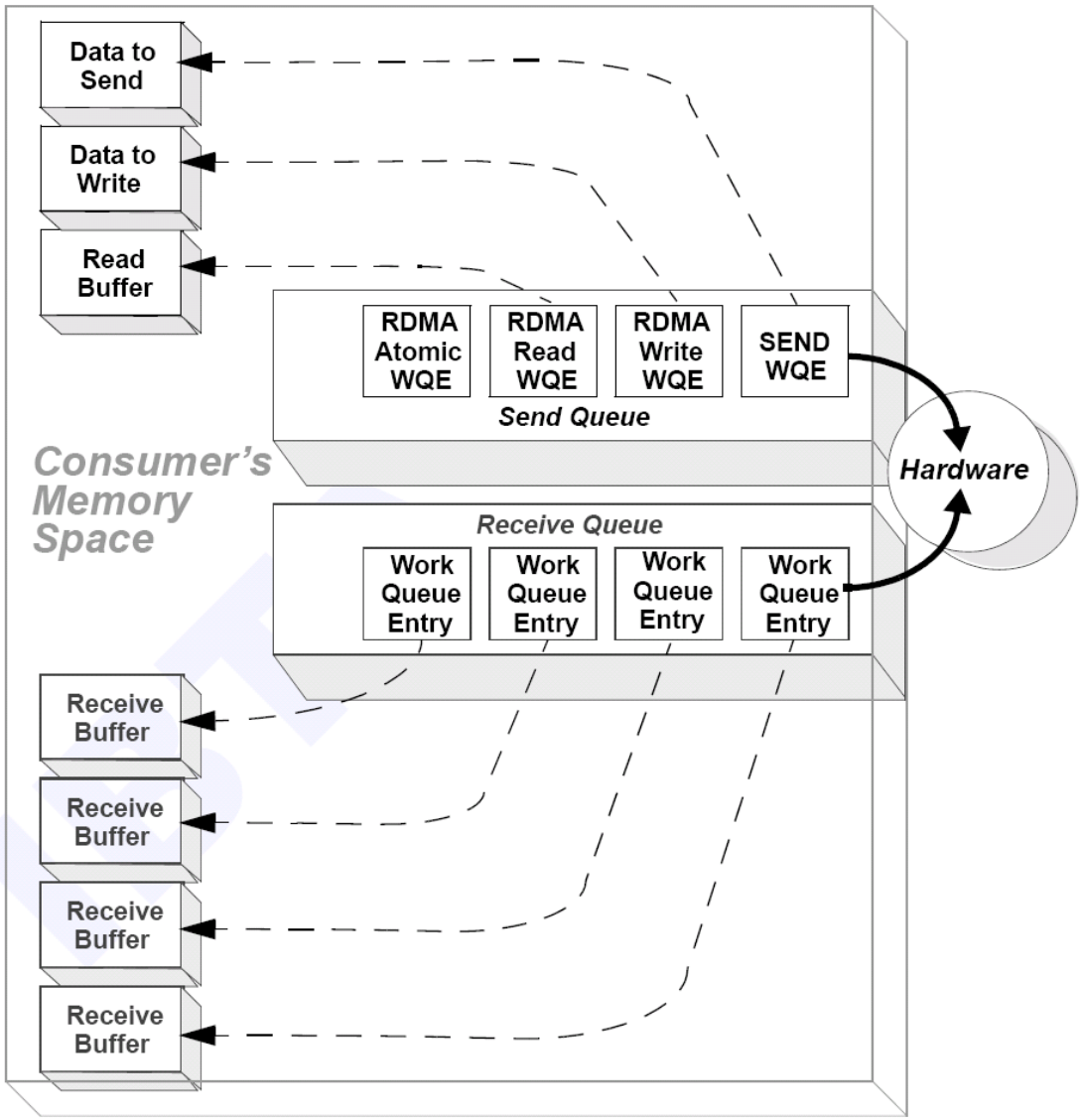


Figure 12 Work Queue Operations

Completion Queue

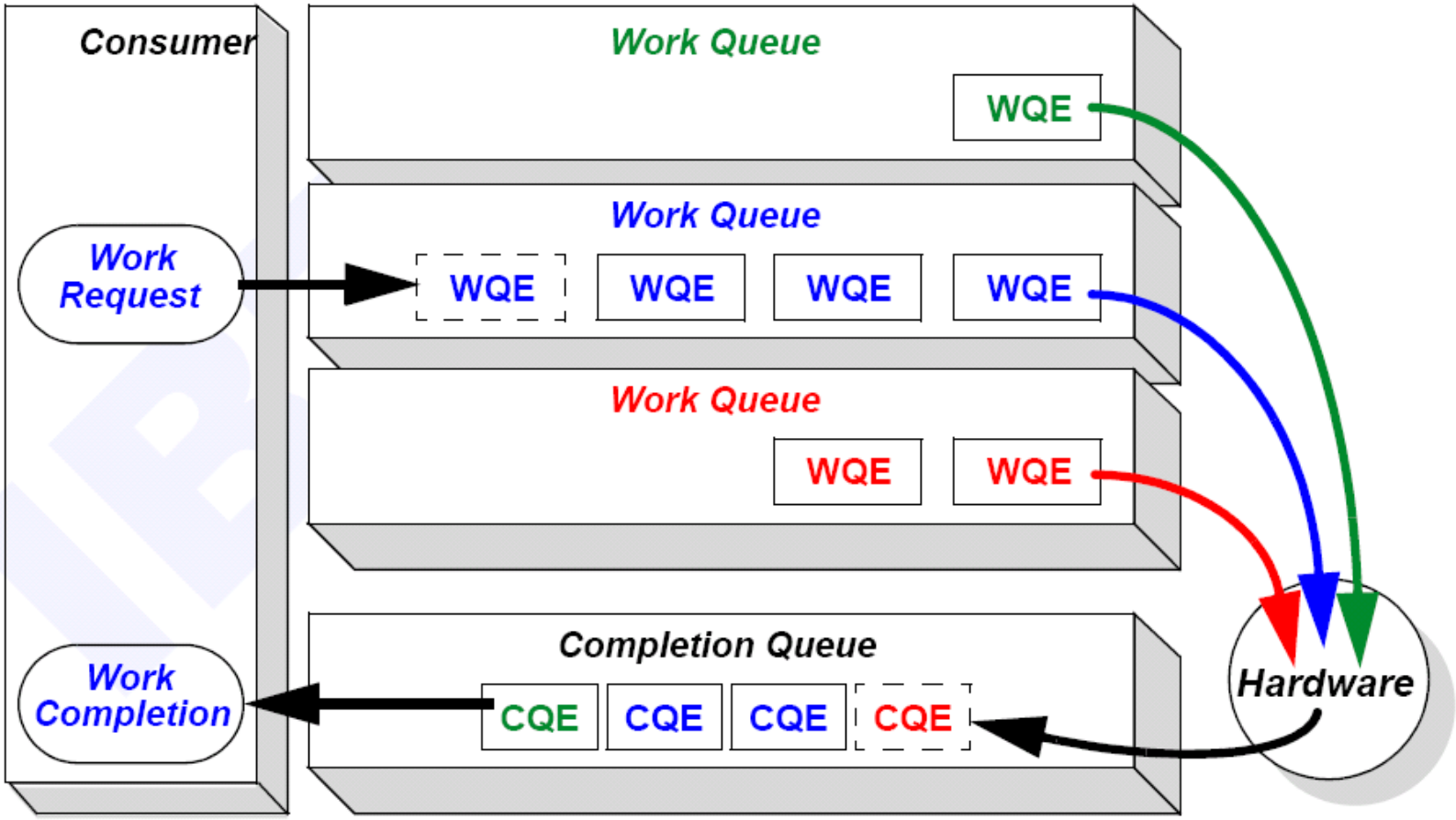
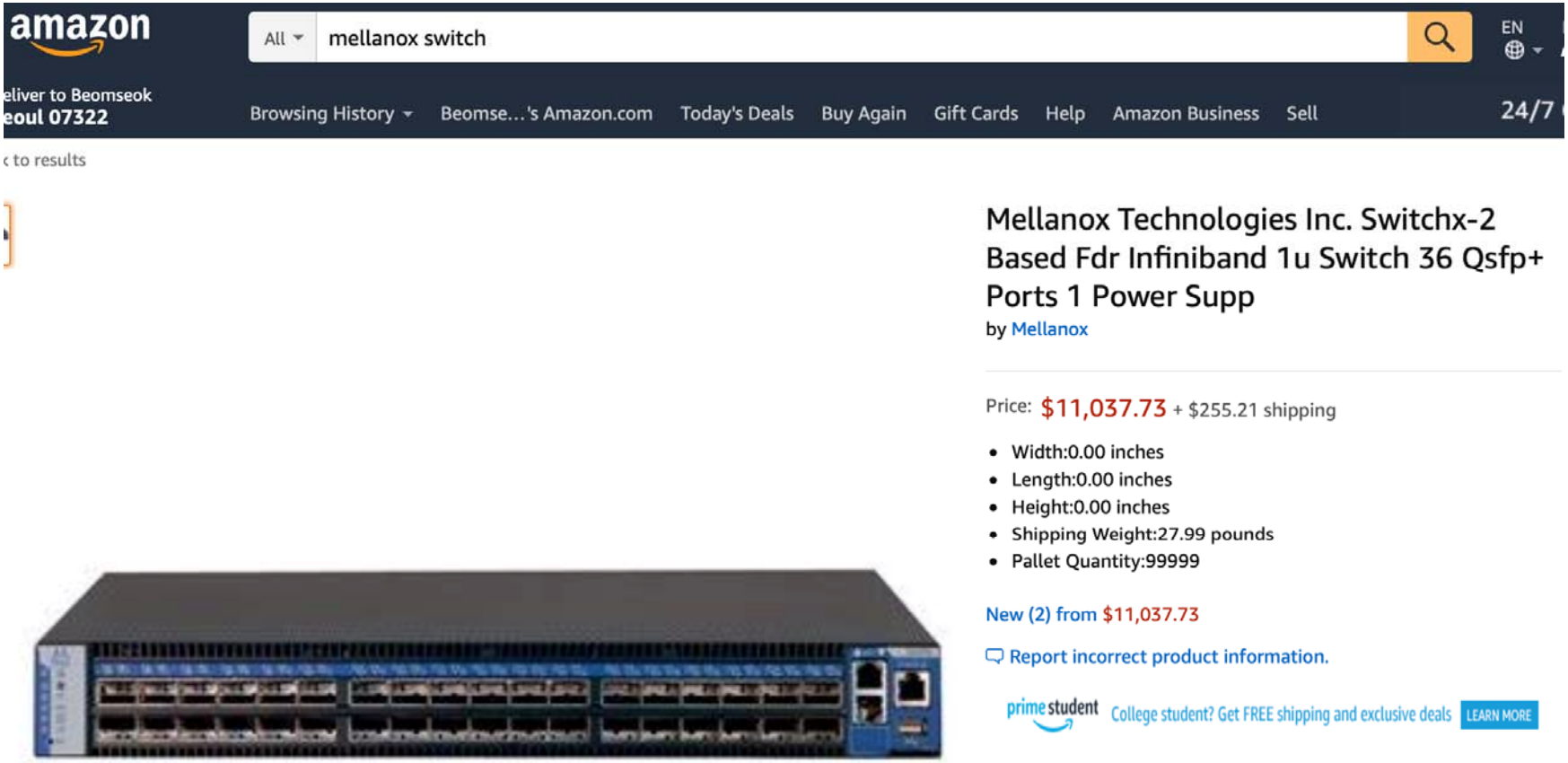


Figure 11 Consumer Queuing Model

- Queue Pair:
  - The status of the result of an operation (send/receive) is stored in the complete queue.
  - Send/receive queues can bind to different complete queues.
- Related system level verbs:
  - Open QP, create complete queue, Open HCA, open protection domain, register memory, allocate memory window, etc
- User level verbs:
  - post send/receive request, poll for completion.

- To communicate:
  - Make system calls to setup everything (open QP, bind QP to port, bind complete queues, connect local QP to remote QP, register memory, etc).
  - Post send/receive requests.
  - Check completion.

Shopping for a high-speed Infiniband switch



Latency Numbers Every Programmer Should Know

Latency Numbers Every Programmer Should Know

latency.txt

```
Latency Comparison Numbers (~2012)
----------------------------------
L1 cache reference                           0.5 ns
Branch mispredict                            5   ns
L2 cache reference                           7   ns                      14x L1 cache
Mutex lock/unlock                           25   ns
Main memory reference                      100   ns                      20x L2 cache, 200x L1 cache
Compress 1K bytes with Zippy             3,000   ns        3 us
Send 1K bytes over 1 Gbps network       10,000   ns       10 us
Read 4K randomly from SSD*             150,000   ns      150 us          ~1GB/sec SSD
Read 1 MB sequentially from memory     250,000   ns      250 us
Round trip within same datacenter      500,000   ns      500 us
Read 1 MB sequentially from SSD*     1,000,000   ns    1,000 us    1 ms  ~1GB/sec SSD, 4X memory
Disk seek                           10,000,000   ns   10,000 us   10 ms  20x datacenter roundtrip
Read 1 MB sequentially from disk    20,000,000   ns   20,000 us   20 ms  80x memory, 20X SSD
Send packet CA->Netherlands->CA    150,000,000   ns  150,000 us  150 ms
```

- QPI Latency: 40 nsec
- InfiniBand Latency: 1.3 usec
- 10G Ethernet Latency: 5 usec

- QPI Bandwidth: 12.8 Gbps
- InfiniBand Bandwidth: 40~100 Gbps
- 10G Ethernet Bandwidth: 10Gbps

InfiniBand in Titan Cluster

- Each host has two IPs.
  - One for 1G Ethernet
    - titan1, titan2, titan3, titan3
  - The other for IPoIB (IP over InfiniBand)
    - titan1-ib, titan2-ib, titan3-ib, titan4-ib
    - You should use IPoIB for your project #3.