EE482b: Advanced Computer Organization Lecture #2
Interconnection Networks Architecture and Design
Stanford University Tuesday, 16th of April 2001

Advanced Computer Organization: Interconnection Networks

Lecture #2: Tuesday, 9th of April 2001
Lecturer: Kelly Shaw
Scribe: Amit Gupta and Francois Labonte
Reviewer: Kelly Shaw

1 Review

Kelly started the lecture by reviewing professor Dally's example of a topology compared to a map of the US. Here roads correspond to routes, and junctions with parking spaces correspond to nodes.

2 Introduction to topologies

In choosing a topology various factors need to be taken into account including required number of ports, required bandwidth, duty factor/port, packaging technologies available, etc.

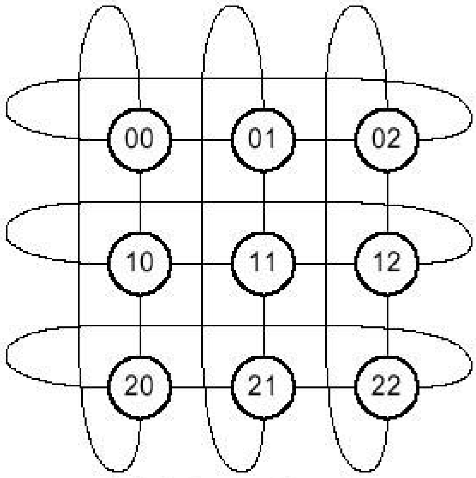
There are two factors in evaluating topologies, without taking into account routing and flow control:

cost — number of chips and their complexity and number and length of connections between chips

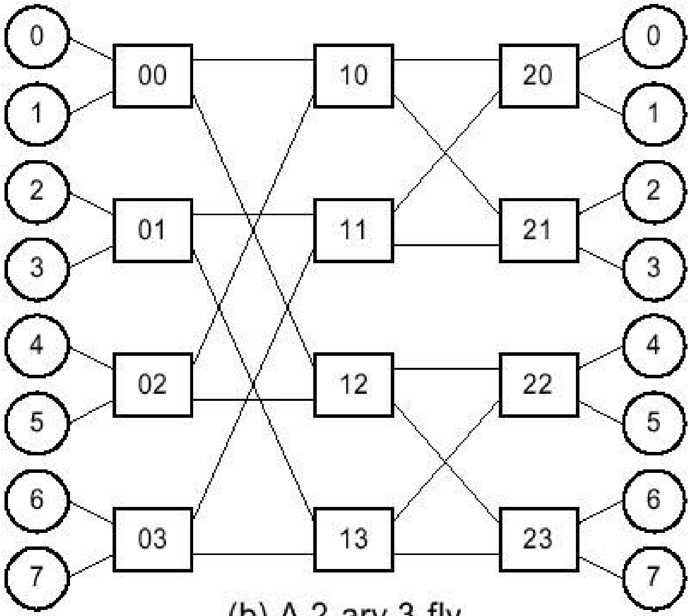
performance — bandwidth, latency

Choosing a topology to match the problem that has to be solved can be bad. Earlier people chose tree topologies, but these are not good because:

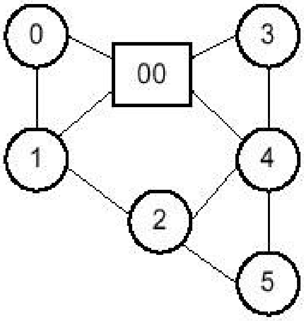
1. load balancing, if problem changes dynamically
2. can lead to long wires and high node degree.
3. if problem size changes it can lead to poor load balance because not matched to machine.
4. not flexible, eg. if algorithm changes



(a) A 3-ary 2-cube



(b) A 2-ary 3-fly



(c) An irregular network

Next we looked at a few different topologies available - these are shown in figure 3.1 of the notes. They correspond to a torus network, a butterfly network and an irregular network.

The difference between a torus and a mesh is that tori have wraparound channels, while meshes do not.

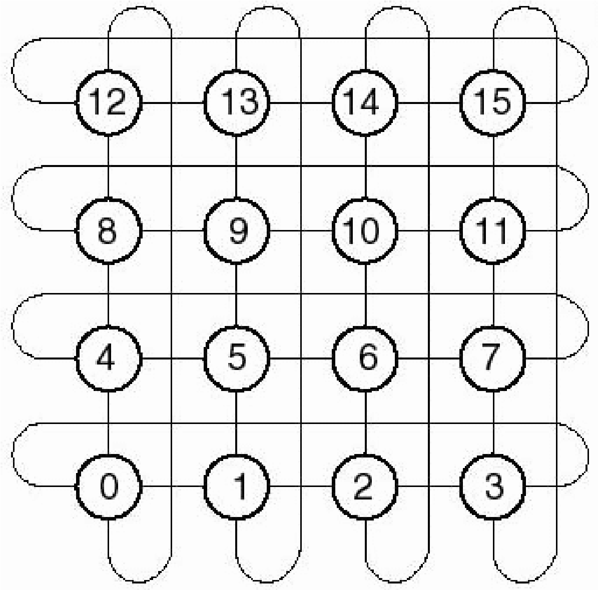
3 Nodes and Channels

A route consists of a series of channels where the destination of one is the source of the next.

A channel has the following characteristics:

width w_{xy} — number of parallel signals

frequency f_{xy} — rate at which bits are transported



latency t_{xy} — time required for bits to travel between 2 nodes

where x is the source and y is the destination.
Latency is proportional to the physical length, and is given by:
$l_c = v \cdot t_c$
Bandwidth, $b_c = w \cdot f_c$
The width, w, is less than the number of pins required by the channel because additional pins are required for timing and control.
The degree of a node is the number of channels a node has:
$\delta_x = \|c_x\| = \|c_{ix}\| + \|c_{ox}\|$

Each edge in the 4-ary 2-cube represents 2 channels, one in each direction. In this torus, there are multiple paths, for different node pairs.

4 Cutting the Network

A cut of the network is achieved by dividing the network into 2 partitions. The cut of the network is represented by all channels you cut, where channels cut are the ones for which the source is on one partition and the destination on the other partition.
size of a set = $\|C(N_1, N_2)\|$ where N_1 nad N_2 are the 2 partitions of the network

Total bandwidth of cut = $B(N_1, N_2)$ =

$$\sum_{c \in C(N_1,N_2)} b_c$$

A bisection is a special type of cut that divides the network into 2 equal parts. Channel bisection is the minimum channel cut over all bisections of the network; ie. minimum number of channels cut in a bisection.
Bisection bandwidth: minimum bandwidth over all bisections of the network
B_B = min B_c over bisections
$B_B = b \cdot B_c$, if network has uniform channel bandwidth.

The 4-ary 2-cube bisection cuts 8 edges or 16 channels. Therefore bisection bandwidth = 16 x channel bandwidth.

5 Paths and Routes

A path is an ordered set of channels:
$\{c_1, c_2 \ldots c_n\}$ where source of c2 is destination of c1.
Length of path is the hop count.
The minimal path from source to destination is the one with the smallest hop count between the 2 nodes. There can be multiple minimal paths between 2 nodes.
R_{xy} = set of all minimal paths between source x and destination y
H_{xy} = hop count of minimal path between x and y
H_{max} = max hop count for network (diameter of network)
H_{avg} = average hop count for network

Distance of path,

$$D(p) = \sum_{c \in P} l_c$$

(sum of lengths of each channel along the path)
where l_c is physical length of channel
Delay,

$$t(p) = \sum_{c \in P} t_c$$

(sum of all the delays for all the channels on the path)
The average and max are defined similar to that for hopcount.

6 Throughput

b — bandwidth

γ_c — maximum channel load

$$\gamma_c = \frac{\text{number of packets going over channel when every node sends a packet to every other node}}{\text{number of nodes in network}}$$

s — speedup

speedup is how fast internal network needs to be deal with external traffic.

γ_{max} - maximum of all γ_c's

$\theta_{ideal} = \frac{b}{\gamma_{max}}$ is the ideal throughput without considering routing/flow control. (assuming perfect flow control)

$s = \frac{\theta_{ideal}}{b_i}$,where bi is the offered input load for each port.

- for uniform traffic :

$$\gamma_c = \frac{1}{\|N\|} \sum_{x \in N} \sum_{y \in N} \sum_{P \in R_{xy}} \begin{cases} \frac{1}{\|R_{xy}\|} & \text{if } c \in P \\ 0 & \text{otherwise} \end{cases}$$

- for arbitrary traffic:

$$\gamma_c = \sum_{x \in N} \sum_{y \in N} \lambda_{xy} \sum_{P \in R_{xy}} \begin{cases} \frac{1}{\|R_{xy}\|} & \text{if } c \in P \\ 0 & \text{otherwise} \end{cases}$$

where $\|R_{xy}\|$ = number of minimal paths,
λ_{xy} = probability x sends to y
Channel width, w is constrained by

1. Node pinout W_n channels can't be made wider than pinout of node allows
$w \leq \frac{W_n}{\delta}$

2. Bisection bandwidth B_s and the maximum number of signals that can cross a minimum cut of the system, W_s
$B_B \leq B_s = f \cdot W_s$
$w \leq \frac{W_s}{B_c}$

$$w \leq min(\frac{W_n}{\delta}, \frac{W_s}{B_C})$$

For any k-ary n-cube $\gamma = \frac{k}{8}$

Example of θ_{ideal}

For a 16-node ring with $b_i = 100$Mb/s, $b = 1$Gb/s

$\gamma = \frac{k}{8} = \frac{16}{8} = 2$
$\theta_{ideal} = \frac{b}{\gamma} = \frac{1Gb/s}{2} = 500Mb/s$

Alternatively,
Number of channels cut in bisection, $B_c = 4$
$\theta_{ideal} = \frac{2B_B}{\|N\|} = \frac{2\cdot(1Gb/sx4)}{16} = 500Mb/s$
where B_B is bisection bandwidth.

7 Latency

Latency without contention includes the head latency T_h and the serialization latency T_s. The serialization latency is just the size of the message L divided by the available bandwidth b. Head latency is made up of router delay T_r, which is the product of the hop count H by a single router delay t_r. The flight delay T_w is the division of the total of the distance of the channels D by the signal speed v.
$T = T_h + T_s$
$T_s = \frac{L}{b}$
$T_h = T_r + T_w$
$T_r = H \cdot t_r$
$T_{r,avg} = H_{avg} \cdot t_r$

When there is no contention (0) ,

$T_{avg}(0) = H_{avg} \cdot t_r + \frac{D_{avg}}{v} + \frac{L}{b}$

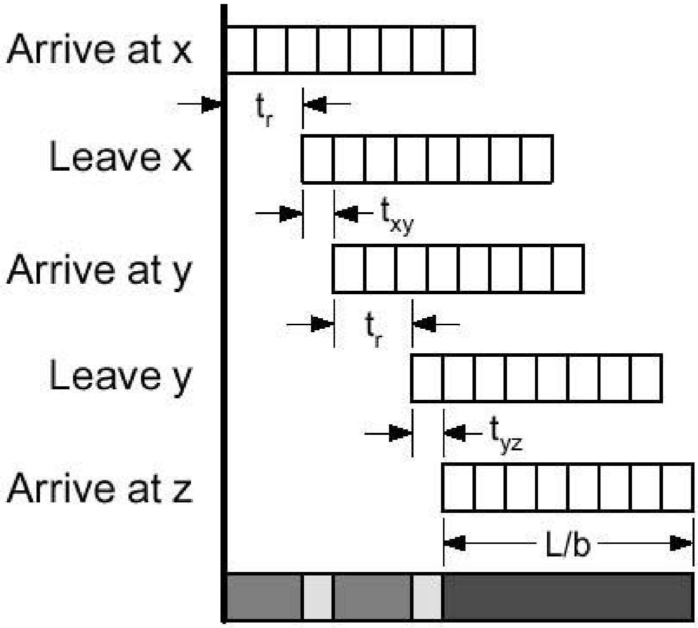


Figure 3.4: Gantt chart showing latency of a packet traversing two channels in the absence of contention

8 Path diversity

R_{xy} is the set of paths from node x to y. A network with $|R_{xy}| > 1$ is more robust because it can possibly avoid a down switch or a heavily loaded channel.

Random traffic is usually the best case. Permutation traffic is when each node sends to exactly one other node. Path diversity can be determined by evaluating γ_C for different traffic patterns. Butterfly has no path diversity while torus has some. Non-minimal routing exploits path diversity to improve throughput.

Example from Chapter 3, page 37: In a 2-ary 4-fly network a shuffle permutation where node {b3, b2, b1, b0} sends to {b2, b1, b0, b3}, $\gamma_{max} = 4$, $\Theta_{ideal} = 0.25$

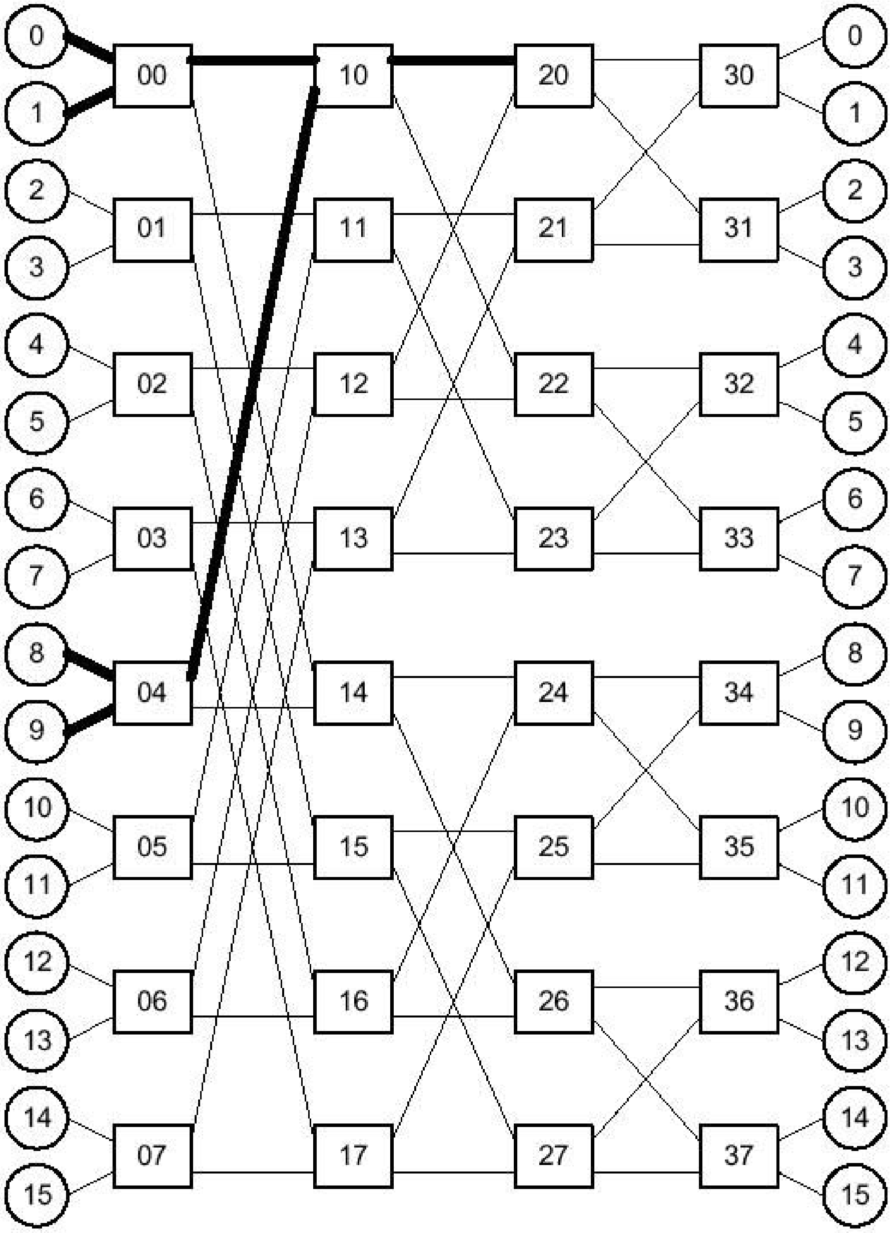


Figure 3.5: Routing a shuffle permutation on a 2-ary 4-fly results in all of the traffic concentrating on one quarter of the channels at the center of the network, degrading performance by a factor of four.

9 Butterfly

The Butterfly network is an indirect network referred to as a k-ary n-fly where n is the number of stages, k is the radix of the nodes ($\delta/2$). The number of I/O nodes N, is k^n, each stage has k^{n-1} kxk crossbars

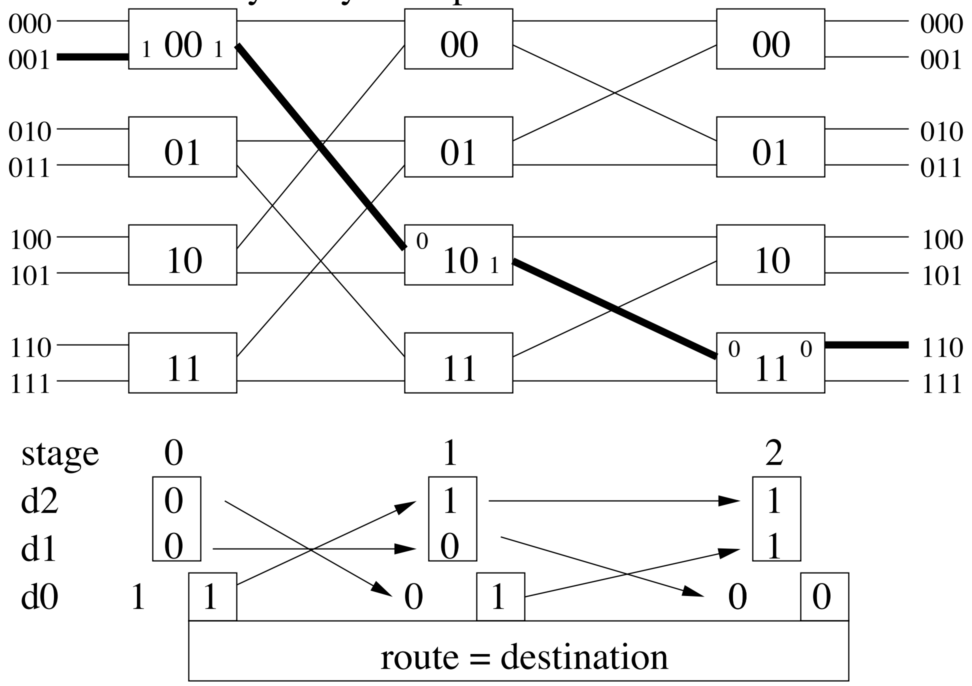
Advantages:

- Minimum diameter for N node network, $H = log_k(N+1)$
- Switches of degree $\delta = 2k$
- Simple to route

Disadvantages:

- No path diversity $|R_{xy}| = 1$
- Cannot be physically implemented without long wires that cross half the machine.

2−ary 3−fly example route from 001 to 110

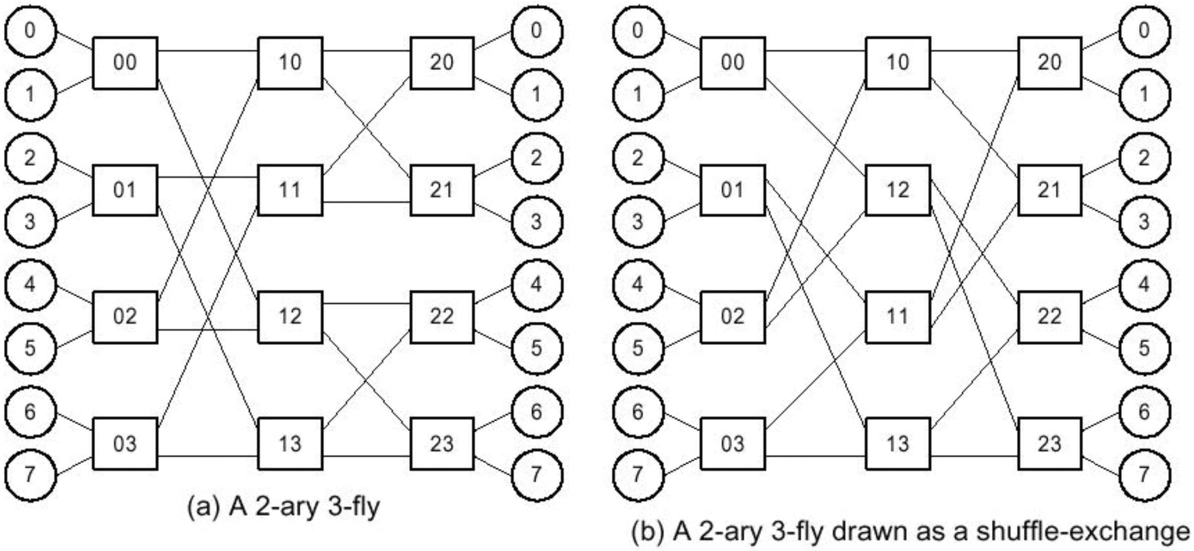


In the figure, each switch is labeled a n-digit radix-k number $\{d_{n-1}, ..., d_1\}$. An extra digit d_0 can identify a terminal (input or output). The wiring between stages i-1 (output) and i (input) permutes the digits d_{n-i} and d_0. At stage i, for a destination of $\{b_{n-1}, ..., b_0\}$ the output port chosen is b_{n-i}.

9.1 Isomorphic Butterflies

A lot of networks are butterflies with renumbered nodes. Two networks with sets of channels and nodes $K_1(C_1, N_1)$ and $K_2(C_2, N_2)$ are isomorphic,

$K_1 \simeq K_2$, iff $\exists f, f(N_1) = N_2$ and $g(C_1) = C_2$



(a) A 2-ary 3-fly

(b) A 2-ary 3-fly drawn as a shuffle-exchange

Figure 4.1: A 2-ary 3-fly drawn two ways (a) as a conventional butterfly, and (b) as a shuffle exchange network. The only difference is the position of switch nodes 11 and 12.

9.2 Butterfly Performance

Topology is characterized by throughput, latency and path diversity. The bisection of a Butterfly, $B_{C,fly}$ is half the number of nodes, the number of channels that span cut between upper and lower halves of network. Under uniform traffic, the channel load is one (half of the traffic crosses the bisection). Under reverse traffic, where all the traffic crosses the bisection, it goes up to 2. The ideal bandwidth is the ratio of the channel bandwidth by the maximum channel load.

$$B_{C,fly} = \frac{N}{2}$$
$$\gamma_{uniform} = 1$$
$$\gamma_{reverse} = 2$$
$$\Theta_{ideal} = \frac{b}{\gamma_{max}}$$

The channel width is the lesser of the pinout and bisection limitation.

$$w_{fly} \leq min(\frac{W_n}{\delta}, \frac{W_s}{B_C})$$
$$B_C = \frac{N}{2}$$
$$\delta = 2k$$
$$w_{fly} \leq min(\frac{W_n}{2k}, \frac{2W_s}{N})$$

Thus with uniform loading ($\gamma = 1$)

$\Theta_{ideal} = wf$
$\Theta_{ideal} \leq min(\frac{W_n f}{2k}, \frac{2W_s f}{N})$
$\Theta_{ideal} \leq min(\frac{B_n}{2k}, \frac{2B_s}{N})$

To maximize throughput and minimize diameter, k is chosen as the largest value for which the network is still bisection-limited.

$k = \frac{NB_n}{4B_s}$

The number of hop counts is always the number of stages plus 1:

$H_{fly} = n + 1$

Example from Chapter 4, page 47 is rederived here:
A k-ary n-fly, using the formula for optimal k, derive metrics

$N = 2^{12}$
$L = 512bits$
$W_s = 2^{14}$
$W_n = 2^8$
$f = 1Gb/s$
$k = \frac{2^{12} \times (1Gb/s \times 2^8)}{4 \times (1Gb/s \times 2^{14})} = 16$

The optimal configuration is thus a 16-ary 3-fly

$\delta = 32$
$w = \frac{W_n}{\delta} = 8$
$H = 3 + 1 = 4$
$\Theta_{ideal} = \frac{b}{\gamma} = \frac{wf}{\gamma} = \frac{8 \times 1Gb/s}{1} = 8Gb/s$
$T_s = \frac{L}{b} = \frac{512}{8Gb/s} = 64ns$