

FAULT TOLERANT SYSTEMS

<http://www.ecs.umass.edu/ece/koren/FaultTolerantSystems>

Part 11 - Networks - 1

Chapter 4 - Fault Tolerant Networks

Part.11 .1

Copyright 2007 Koren & Krishna, Morgan-Kaufman

Types of Interconnection Networks

- ◆ **Shared-memory multiprocessor connecting processors and memories**
 - * processors **read** or **write** in memories
- ◆ **Processors connected in a distributed system**
 - * Each has its own local memory
 - * communicate through messages while executing parts of a common application
- ◆ **Components connected through **links** and **switchboxes****
 - * a switchbox allows a given component to communicate with several other components without separate links
- ◆ **Wide-area networks - large number of processors that operate independently (example - **the Internet**)**
 - * share various types of information through **packets**
 - * more complicated switchboxes called **routers**

Part.11 .2

Copyright 2007 Koren & Krishna, Morgan-Kaufman

Network Topology

- ◆ **Topology** - network organization
- ◆ One or more paths between message sender (source) and receiver (destination)
- ◆ Links and switchboxes - either uni- or bi-directional
- ◆ If single path between a source and a destination - one fault along path will disconnect the pair
- ◆ Fault tolerance achieved by having multiple paths and/or spare units
- ◆ Need to evaluate resilience to faults provided by such redundancy + degradation in network operation as faults accumulate

Graph Theoretical Measures of Network Resilience

- * **Node (link) connectivity** - minimum number of nodes (links) that must be removed to disconnect the graph
- * **Distance between nodes** - smallest number of links
- * **Diameter** - longest distance in the graph
- * **Diameter stability** - rate of increase in diameter when nodes fail

Diameter Stability

- ◆ Distance between nodes increases as nodes, links or switchboxes fail
- ◆ **Diameter Stability** measures this increase
- ◆ **Persistence** - smallest number of nodes that must fail in order for the diameter to increase
- ◆ **Example:** persistence of a cycle graph is 1
 - * one failure increases the diameter
- ◆ **Probabilistic measure** of diameter stability - vector $DS = (P_{d+1}, P_{d+2}, \dots)$
 - P_{d+i} - probability that diameter of the network increases from d to $d+i$ based on some fault probability distribution
 - P_{∞} - probability of diameter becoming infinite (graph being disconnected)

Computer Network Measures

- ◆ **Reliability $R(t)$** - probability that all nodes are operational and can communicate during $[0, t]$
- ◆ **Path Reliability** - same for a specific source-destination pair
 - * both assume no repair
 - * If repair exists - use **availability** instead
- ◆ **Bandwidth** - maximum rate of flow of messages
- ◆ **Connectability $Q(t)$** - expected number of source-destination pairs still connected at time t in the presence of faults

Common Network Topologies

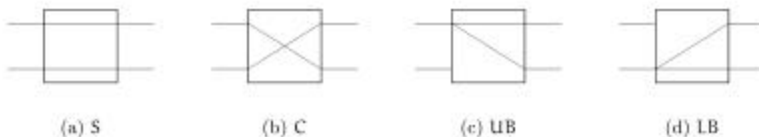
- ◆ **Type 1:** input and output nodes connected through links and switchboxes
 - * multi-stage networks
 - * crossbar
 - * resilience measures - bandwidth and connectability
- ◆ **Type 2:** nodes connected through links - no separate switchboxes, nodes serve as switches
 - * mesh
 - * hypercube
 - * resilience measures - reliability/path reliability (or availability)
- ◆ Resilience obtained through
 - * **multiple paths** connecting message source to destination
 - * **spare nodes** that can replace failed units

Part.11 .7

Copyright 2007 Koren & Krishna, Morgan-Kaufman

Non-Fault-Tolerant Multi-Stage Network

- ◆ **Butterfly Network** - typically built out of **2x2** switches - two inputs and two outputs



- ◆ **Switch has four settings -**
 - * **S - Straight** - top input line connected to top output, same for bottom lines
 - * **C - Cross** - top input line connected to bottom output and bottom input line to top output
 - * **UB - Upper Broadcast** - top input line connected to both output lines
 - * **LB - Lower Broadcast** - bottom input line connected to both output lines

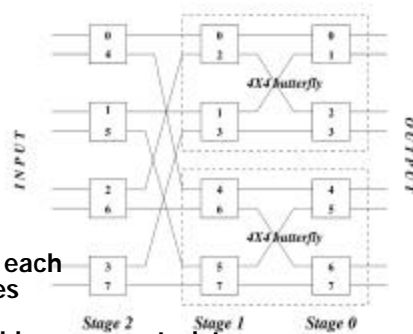
Part.11 .8

Copyright 2007 Koren & Krishna, Morgan-Kaufman

Butterfly Network

◆ k-stage network ($k \geq 3$)

- * 2^k inputs and 2^k outputs
- * k stages of 2^{k-1} switches each
- * Connections follow a recursive pattern from input to output
- * Input stage - top output line of each switchbox connected to input lines of a $2^{k-1} \times 2^{k-1}$ butterfly, and bottom output line of each switchbox connected to input lines of another $2^{k-1} \times 2^{k-1}$ butterfly



◆ 2-stage butterfly

- * Input stage - top output line of each of its two switchboxes connected to one 2×2 switchbox and the bottom output line to another 2×2 switchbox

◆ 1-stage butterfly

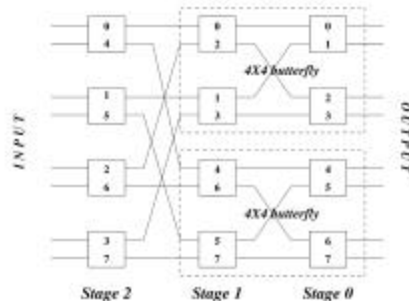
- * Single 2×2 switchbox

Part.11 .9

Copyright 2007 Koren & Krishna, Morgan-Kaufman

Butterfly Network - Details

- ◆ A switchbox in stage i has lines numbered 2^j apart
- ◆ Output line j of every stage goes into input line j of the following stage ($j=0, \dots, 2^{k-1}$)
- ◆ Numbers in any box other than at output stage are both of same (even or odd) parity
- ◆ **Butterfly is not fault-tolerant**: there is only one path from any given input to a specific output
- ◆ If a switchbox in stage i fails - 2^{k-i} inputs will no longer be connected to 2^{i+1} outputs
- ◆ The system can still operate but in a degraded mode



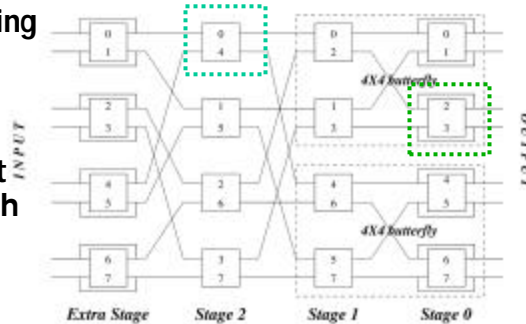
Part.11 .10

Copyright 2007 Koren & Krishna, Morgan-Kaufman

Extra-Stage Networks - Fault Tolerant

- ◆ **Extra stage** - duplicating stage 0 at the input

- ◆ **Bypass multiplexors** around switchboxes at the input and output stages - a failed switch can be bypassed by routing around it



- ◆ **Examples:**

- * **Stage-0 switchbox** carrying lines 2,3 fails - duplicated by the extra stage - failed box is bypassed by the multiplexor
- * **Switchbox in stage-2** carrying lines 0,4 fails - extra stage is set so that input line 0 is switched to output line 1 and input line 4 to output line 5 - bypassing the failed switchbox

- ◆ **Exercise** - Network can remain connected despite the failure of up to one switchbox anywhere in the system

Part.11 .11

Copyright 2007 Koren & Krishna, Morgan-Kaufman

Dependability Measures of a Multi-Stage Network

- ◆ Network connects N processors to N memory units in a shared memory architecture ($N = 2^k$)
- ◆ In the presence of faulty elements - system can operate - possibly in a degraded mode
- ◆ System's resilience as it degrades can be measured
- ◆ **Resilience Measures:**
 - * **Bandwidth**
 - * **Average number of operational paths**
 - * **Metrics of connectivity among processors and memories**
- ◆ **All measures are a function of time t - assuming faults occur and are possibly repaired during $[0, t]$ - t omitted for simplicity**

Part.11 .12

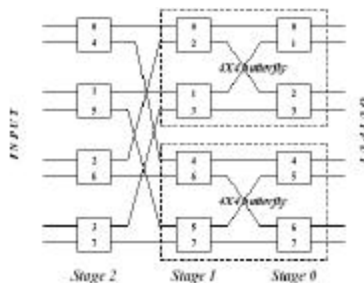
Copyright 2007 Koren & Krishna, Morgan-Kaufman

Butterfly Network - No Failures Bandwidth Calculation Model

- ◆ **Bandwidth (BW)** - expected number of access requests from the processors that reach the memories

- ◆ **Assumptions:**

- * Every processor generates in each cycle, with probability p_r , a request to a memory module, directed to any of the N memory modules with equal probability $1/N$
- * Requests in each cycle are independent from requests in previous cycles - approximation



Part.11 .13

Copyright 2007 Koren & Krishna, Morgan-Kaufman

Recursive Bandwidth Calculation (No failures)

- ◆ $p_r^{(i)}$ - probability that a link at stage i carries a request - calculated recursively from processors ($i=k-1$) to memories ($i=0$)
- ◆ **Stage $k-1$** - two processors feeding each link - union of two independent events with probability $p_r/2$ each

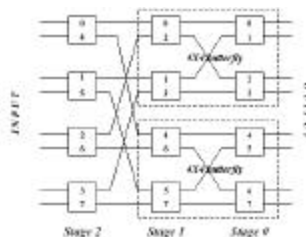
$$p_r^{(k-1)} = p_r/2 + p_r/2 - (p_r/2)^2 = p_r - p_r^2/4$$

- ◆ Recursively for a link in **stage $i-1$** ($i=k-1, \dots, 1$)

$$p_r^{(i-1)} = p_r^{(i)} - (p_r^{(i)})^2/4$$

- ◆ and

$$BW = N p_r^{(0)}$$



Part.11 .14

Copyright 2007 Koren & Krishna, Morgan-Kaufman

Butterfly Network - Bandwidth Calculation - Including Failures

q_l - probability of link failure ; $p_l = 1 - q_l$

- ◆ Failure probability of a switchbox included in q_l
- ◆ Probability that a request at the input line to stage $i-1$ will propagate to an output in stage i is

$$p_l p_r^{(i)}/2$$

- ◆ Resulting recursive equation -

$$p_r^{(i-1)} = p_l p_r^{(i)} - (p_l p_r^{(i)})^2/4$$

- ◆ Setting $p_r^{(k)} = p_r$, calculate $p_r^{(0)}$ recursively, and substitute in

$$BW = N p_r^{(0)}$$

Butterfly - Connectability Analysis

- ◆ **Connectability - Q** - expected number of connected processor-memory pairs
- ◆ Assuming fault-free processors and memories
- ◆ Exactly one path between a processor and a memory
- ◆ Each path has $k+1$ links and k switchboxes
- ◆ Probability of a link/switchbox failure is q_l / q_s

$$p_l = 1 - q_l \quad p_s = 1 - q_s$$

- ◆ Probability of a fault-free path is $p_l^{k+1} p_s^k$
- ◆ There are $2^{2k} = N^2$ processor-memory pairs

$$Q = 2^{2k} p_l^{k+1} p_s^k = N^2 p_l^{k+1} p_s^k$$

Accessibility

- ◆ **Connectability** does not indicate number A_c of **accessible processors** - still connected to at least one memory (nor the number of accessible memories)
- ◆ We calculate recursively the probability that a given processor is accessible
- ◆ $\Phi(i)$ - probability that at least one fault-free path exists from a switchbox in stage i to the output end of the network
- ◆ $\Phi(0) = 1 - q_i^2$ - probability that at least one link out of a switchbox at the output stage is functional
- ◆ The recursive equation is $\phi(i) = 1 - (1 - p_\ell \phi(i-1))^2$
- ◆ The probability that a given processor can connect to the output end is $p_i \Phi(k)$
and the expected number of **accessible processors** (equal to expected number of **accessible memories** due to symmetry) is $A_c = 2^k p_\ell \phi(k)$

Part.11 .17

Copyright 2007 Koren & Krishna, Morgan-Kaufman

Extra-Stage Network - Connectability

- ◆ Each processor-memory pair connected by two disjoint paths
 - * Probability of at least one fault-free path
 $= \text{Prob}\{\text{1st path is fault-free}\} + \text{Prob}\{\text{2nd path is fault-free}\} - \text{Prob}\{\text{both paths are fault-free}\}$
- ◆ This probability can assume one of the following two expressions (see paths from 0 to 0 or from 0 to 1)

$$A = (1 - q_\ell^2) p_\ell^k (1 - q_\ell^2) + p_\ell^{k+2} - p_\ell^{2k+2} (1 - q_\ell^2)^2$$

$$B = 2(1 - q_\ell^2) p_\ell^{k+1} - p_\ell^{2k+2} (1 - q_\ell^2)^2$$

- ◆ $(1 - q_\ell^2)$ - probability that for a switchbox with a bypass multiplexer at least one link is operational
- ◆ Since there are $2^{2k} = N^2$ pairs,

$$Q = (A + B) 2^{2k} / 2 = (A + B) N^2 / 2$$

Part.11 .18

Copyright 2007 Koren & Krishna, Morgan-Kaufman