Module 17: "Interconnection Networks"

Lecture 38: "Routing Algorithms"

Interconnection Networks

- Topology
- Routing algorithms
- Deadlock avoidance
- Adaptive routing
- Alpha 21364 µP
- Alpha 21364 router
- Alpha 21364 router

[From Chapter 10 of Culler, Singh, Gupta]

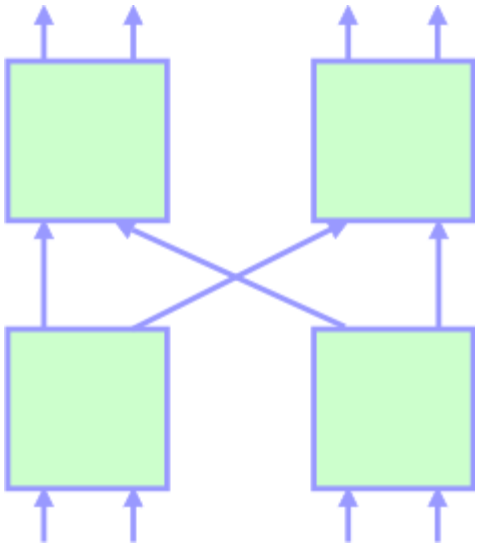
Previous Next

Module 17: "Interconnection Networks"
Lecture 38: "Routing Algorithms"

Topology

- Fully connected
  - A single large switch (could be a bus!)
  - Complete graph
  - Diameter? Average distance? Cost?
- Linear arrays
  - A one-dimensional mesh
  - Unidirectional or bidirectional?
  - Diameter? Average distance? Bisection BW?
- Linear rings
  - One extra connection from end to beginning of linear array
  - Bidirectional is better
  - Diameter? Average distance? Bisection BW?
- Mesh
  - Multi-dimensional generalization of linear array
  - A d-dimensional k-ary mesh has k nodes in each dimension: kd nodes
  - A cube is a 3-dimensional binary mesh
  - A popular deterministic routing algorithm is dimension order routing (DOR) where the message is routed along successive dimensions until it reaches the destination
  - Diameter? Average distance?
- Torus
  - Collection of rings in each dimension
- Hypercube
  - Essentially a d-dimensional binary torus
  - Offers efficient routing via gray codes: current node ^ destination gives you the dimensions to traverse
  - Dimension order routing is referred to as e-cube routing for hypercube
  - Each node connects to d other nodes
  - By connecting 2 hypercubes of d-1 dimension each you$get a d-dimensional hypercube
- Trees

  $
  - Direct and indirect
  - Increased branching factor reduces average distance
- Butterfliew, Benes network and fat trees
  - Tree with many roots
  - Basic butterfly building block



- Unidirectional butterfly introduces conflicts between different routes due to shared edges
- Benes network connects to butterfliew back-to-back to solve this problem: very costly to build
- Fat tree provides a cost-effective solution by folding half of Benes network on the other lalf (bidirectional links)

Routing algorithms

- Types
  - Deterministic vw. adaptive
  - Minimal vs. non-minimal
  - Arithmetic routing algorithm
  - Source-based routing algorithm (large header)
  - Table-based routing (large SRAM storage): $(i_{n+1}, o_n) = R[i_n]$
- Channel dependency and deadlock
  - Common problem in k-ary d-dimensional torus
  - Can happen even with multiple buffers per port
  - Simple solution: reserve buffers for certain packets depending on source and destination (may lead to starvation) [e.g., d > s]
  - Deadlock possibility is higher for wormhole routing

Previous Next

Deadlock avoidance

- Multiple virtual channels per port
  - Break the dependence cycle
  - Allocate virtual channels according to source and destination of packets
- Up*-Down* routing
  - Applies to indirect networks only
  - Logically treat the topology as a spanning tree with processors at leaves
  - Route up to common ancestor and then down
  - No cycles involved
  - Comes for free in trees, fat trees, Butterflies
- Turn model
  - Attacks the fundamental problem: the turns
  - In a 2D mesh there are eight possible turns and that form two different types of cycles (4 turns each)
  - Restrict the use of turns
  - Avoid one turn each from these two cycles: 16 possibilities
  - Out of these 16, 12 are deadlock free
  - West-first: avoids +y to –x and –y to –x turns i.e. cannot turn west
  - North-last: avoids +y to –x and +y to +x turns i.e. cannot turn from north
  - Negative-first: offers a set of choices
  - Dimension order routing is inherently deadlock-free because it disallows all y to x turns (too restrictive)

Adaptive routing

- For fault tolerance and better network utilization
  - Tremendous contention in 2D mesh with DOR for transpose traffic or for accessing locks allocated on a corner node
  - Adaptive routing can make use of other available paths and introduce more concurrency
  - Main idea: decide the next output port dynamically based on switch state (locally or globally gathered)
  - Fully vs. partially adaptive: allow all or some paths
  - Minimal vs. non-minimal adaptive: allow only shortest paths (i.e. hop count goes down monotonically) or allow arbitrary paths
  - Hot-potato routing: example of non-minimal adaptive; misroute one of the contending packets
- Multipath routing
  - A special case of adaptive routing where the decision is static
  - Source-based multipath: source chooses the entire path from the legal set of paths
  - Table-based multipath: routing table provides multiple possible output ports programmed at boot time
- Deadlock avoidance
  - Non-minimal adaptive routing is prone to deadlocks
  - Normal trick is to have a deadlock-free virtual network called the escape route
  - Livelock is a bigger problem: with hot-potato routing a packet may keep on moving around the same path never getting to the destination

Previous Next

Alpha 21364 µP

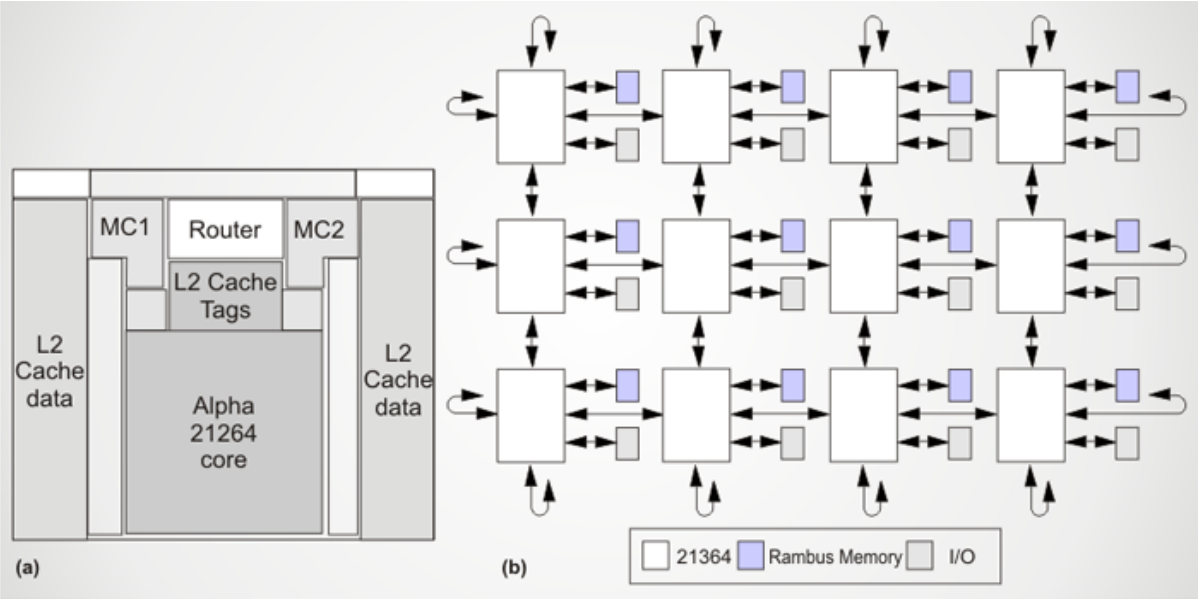


Figure 1. Alpha 21364 floor plan (a) and a 12-processor configuration of Alpha 21364s (b).

Reproduced from IEEE Micro

Alpha 21364 router

- Integrated on-chip router
  - Supports adaptive routing over a 2D torus
  - Clocked at 1.2 GHz and has a 13-cycle routing delay from input to output port (10.8 ns)
  - Eight input and seven output ports
  - Input ports: NEWS, two memory controllers, cache, I/O
  - Output ports: NEWS, L1, L2, I/O
  - Each port has 3.2 GB/s link bandwidth (links are clocked at 0.8 GHz)
  - Flit size is 39 bits: 32 bits of data/control, 7 bits of ECC
  - Seven coherence message classes: Request (3 flits), Forwarded request (3 flits), Data response (18 or 19 flits), Dataless response (2 or 3 flits), I/O write (19 flits), I/O read (3 flits), Special (1 or 3 flits; mostly used for flow control)
- Implements virtual cut-through adaptive routing
  - Blocking router buffers all flits and it is its responsible to restart the routing later; router can buffer 316 messages
  - Simple partial adaptive routing: two choices for each input flit; either continue in same dimension (i.e. east to west or north to south) or take a turn within the minimal rectangle containing the source and destination
  - Preference is given for continuing in the same dimension
  - Deadlock avoidance in coherence protocol: seven distinct virtual networks are provided for seven message classes
  - Deadlock avoidance in routing: three virtual channels within each virtual network; two for deadlock-free torus routing and one for adaptive routing
- The router pipeline
  - Input and output ports are divided into three types each: local (cache and memory

controller), inter-processor (NEWS), I/O

- Depending on input and output ports of a packet it goes through one of the nine logically different routing pipelines (implemented in a single seven-stage pipe)
- Six additional cycles are needed to account for synchronization between router's internal clock and external link clock (runs at 0.8 GHz), pad receiver and driver delay, transport delay from pin (inbound) to router and from router to pin (outbound)
- Uses table-based routing to select output port and virtual channel (one 128-entry routing table and one virtual channel table giving assignments for three output virtual channels)

Previous Next