

Design of a High-Speed Optical Interconnect for Scalable Shared Memory Multiprocessors

Avinash Karanth Kodi and Ahmed Louri

Department of Electrical and Computer Engineering
Optical Networking and Parallel Processing Laboratory
University of Arizona, Tucson, AZ 85721
E-mail: {avinashk,louri}@ece.arizona.edu
<http://www.ece.arizona.edu/~ocppl>

*12th Annual IEEE Symposium on High-Performance Interconnects
25th - 27th August, 2004
Stanford University*

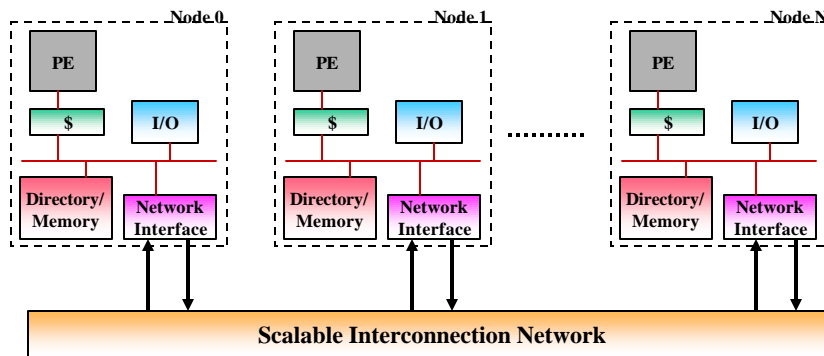


Talk Outline

- Distributed Shared Memory Multiprocessors
 - The Remote Memory Access Latency
- RAPID: Proposed Architecture using Optical Interconnects
 - Wavelength Assignment and Routing Algorithms
 - Scalability Analysis
 - Optical Implementation
- Performance Evaluation of RAPID
 - Simulation Methodology and Results
 - Power Budget Analysis
- Conclusion



Distributed Shared Memory Multiprocessors



- DSMs support **shared address space** by physically distributing the memory
- Communication occurs implicitly as a result of **conventional memory access** instructions



Remote Memory Latency in DSMs

One of the fundamental communication problem in DSMs is:

- **Remote Memory Access Latency (RML)**: the latency in accessing a memory location in a processor other than the one issuing the request
- Remote memory latency takes 1-2 orders of magnitude longer than the local access
- Three critical factors responsible for RML are:
 - **Lack of sufficient remote memory bandwidth**
 - **Long switching/routing Delays**
 - **Cache coherence protocol overhead**



Critical Factors Affecting Remote Latency

- Reducing memory latency and hiding memory latency (prefetching) techniques are commonly used to tolerate large remote latencies
 - => Yet, some of these techniques require much more memory bandwidth and generate more memory traffic
- In addition, as system size increases more processors are incorporated by adding more switching chips
 - => Additional delays occur in additional switching stages increasing RML
- Synchronization operations in parallel programs (locks, barriers) can lead to highly contended accesses
 - => Implementing such synchronization that are essential using broadcast/multicast paradigms are generally expensive



Optical Interconnects

- **Technological Advantages**
 - Absence of electromagnetic wave phenomena (transmission line effects) simplifies or eliminates impedance matching and crosstalk
 - Advantages include reduction of power consumption at high-speeds and voltage isolation
- **Higher Bandwidth**
 - Wavelength Division Multiplexing, along with Space Division Multiplexing, Time Division Multiplexing fully utilizes the huge bandwidth of optics by partitioning into several non-overlapping channels
 - Parallel optical interconnects such as arrays of vertical cavity surface emitting lasers (VCSELs) and photodetectors (PD) provide greater bandwidth-density product



Optical Interconnects

- **Architectural Advantages**

- Optical interconnects can be used to design dense 2-D and 3-D interconnections without significant coupling
- Optical interconnects provide efficient broadcast and multicast functionality at a much lower cost

- **Scalability**

- Optical interconnect based architecture can provide unlimited bandwidth scalability, when more nodes are added, by either increasing the number of channels (adding wavelengths) or by adding additional fibers



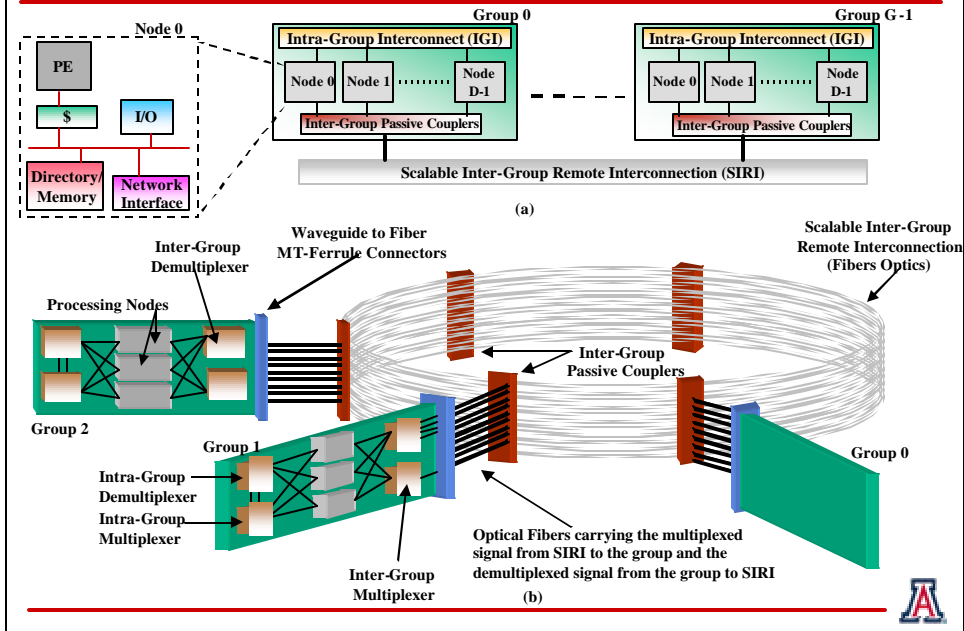
Proposed Network: **RAPID** (Reconfigurable All-Photonic Interconnect for DSMs)

RAPID reduces RML by:

- increasing the **connectivity and maximizing the channel availability** using WDM, TDM and SDM techniques that result in further increased memory bandwidth,
- implementing a **decentralized wavelength allocation scheme and an innovative media access protocol** that not only reduces the diameter of the network but also lowers the queueing/routing delays for packet transmission
- implementing **efficient multicast and broadcast functionality**, which helps to reduce the part of memory latency associated with the implementation of synchronization operations, and
- concentrating only on **passive optical interconnects techniques** such that the optical signal transfer is much faster as there is no switching or conversion



RAPID Architecture

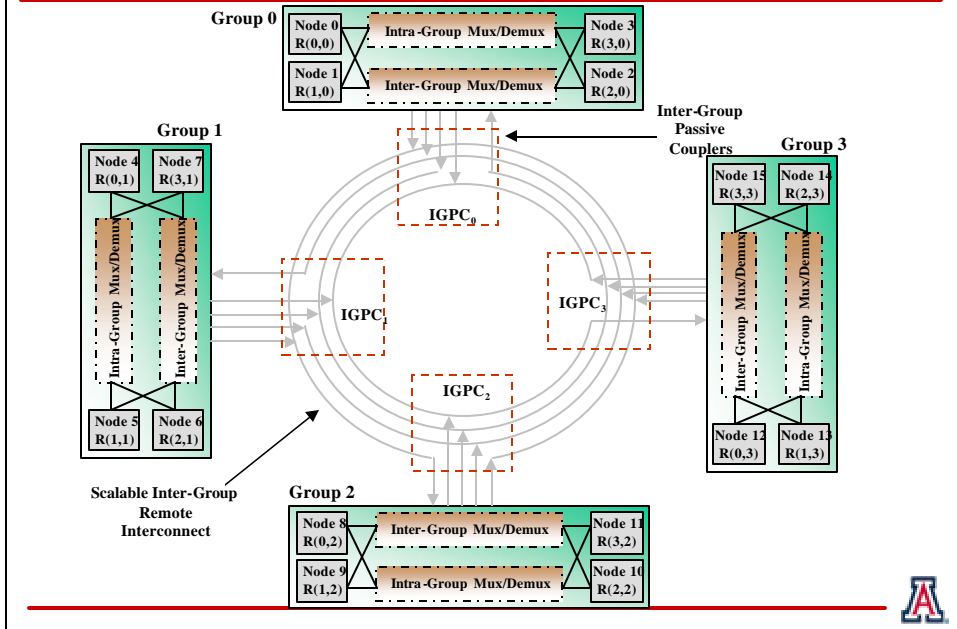


Design of RAPID

- RAPID is defined by a 4-tuple (P,D,G,C) where
 P = Number of Processors/Node, D = Number of Nodes/Group
 G = Number of Groups/Cluster, C = Number of Clusters
- Each node contains up to P processors, each group contains up to D nodes and each cluster contains up to G groups. Initially, we will consider P = 1; C=1, so every node in RAPID is referred to R(d,g) where $0 < d < D-1$ and $0 < g < G-1$
- In what follows is an example of RAPID with D = 4 and G = 4



Functional Design of RAPID ($D = 4, G = 4$)

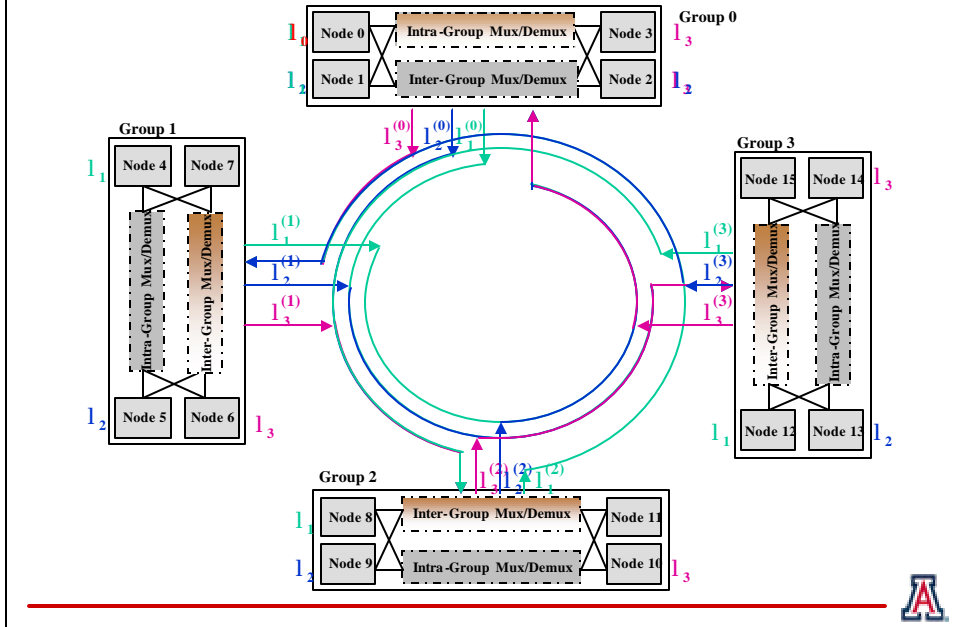


Wavelength Assignment and Routing in RAPID

- RAPID is based on WDM, assigning every node with a wavelength is simple, yet limiting
- RAPID proposes a novel wavelength strategy based on **wavelength re-use and spatial division multiplexing (SDM)**
- Same wavelengths are used for both intra-group and inter-group communication
- Maximum number of wavelength required by RAPID is D (number of nodes/group) wavelengths as explained next



Wavelength Assignment and Routing in RAPID



Wavelength Routing: Inter-Group

- For intra-group communication: Nodes transmit on a given destination wavelength after capturing the token
- For inter-group communication: (2 step protocol)
 - 1) First transmit the request to a “destination group node” - This node need not be the intended destination
 - 2) If not, this intermediate node will forward the request to the actual destination node within the group
- The objective here is to minimize the cost of the network by maintaining a constant degree, yet at the same time to provide maximum connectivity
 - => RAPID requires maximum of only 1-2 hops, that reduces the latency for remote traffic

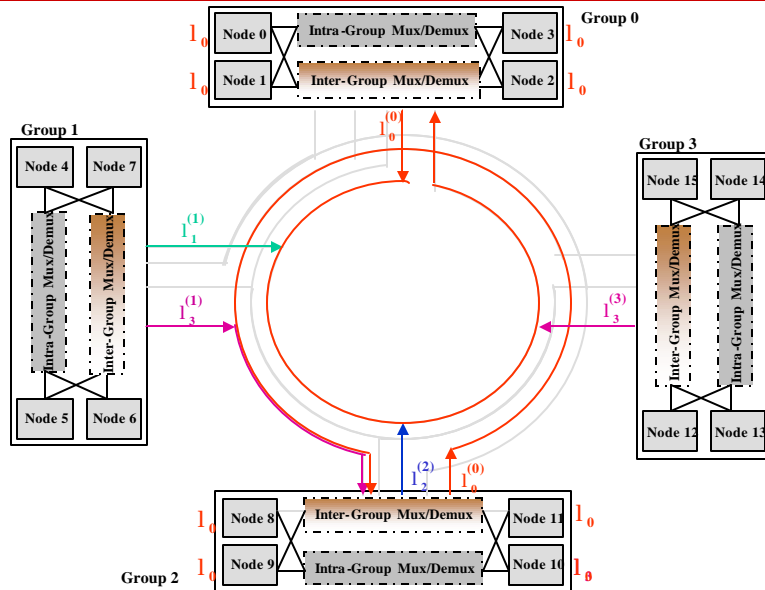


Media Access Protocol

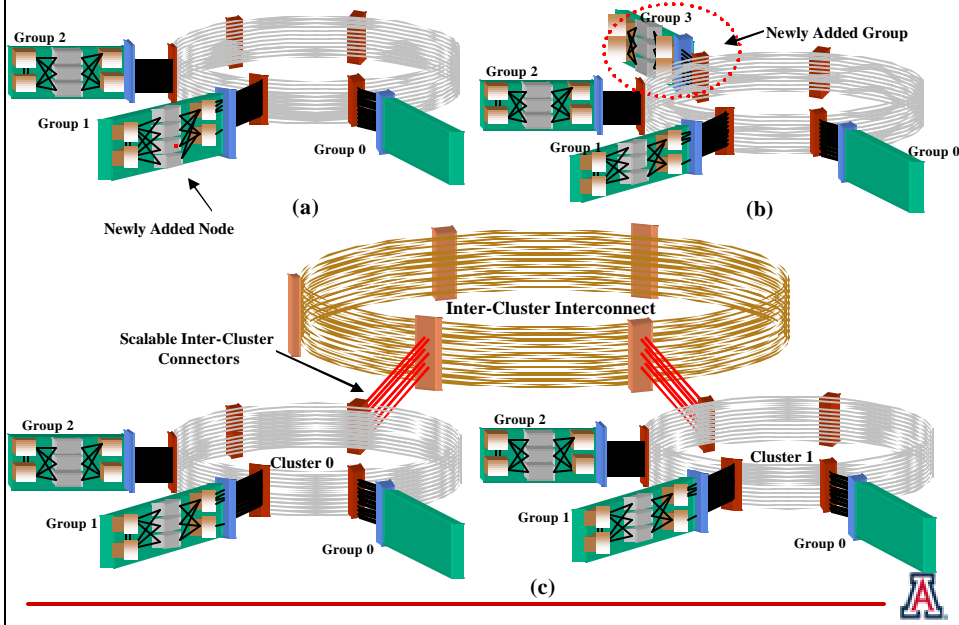
- Token-based media access protocol: Nodes communicate only when they have captured the token for the destination node
- Generate 2 sets of tokens; for intra- and inter-group communication
=> Key idea: These tokens are shared only among the locally connected nodes within the group. Even inter-group communication tokens are shared among only the locally connected nodes within the group
- In RAPID, under worst case scenario, a node waits only for (D - transmissions before transmitting its request where D is number nodes
=> This greatly reduces the waiting time for packet transmission thereby reduces the RML



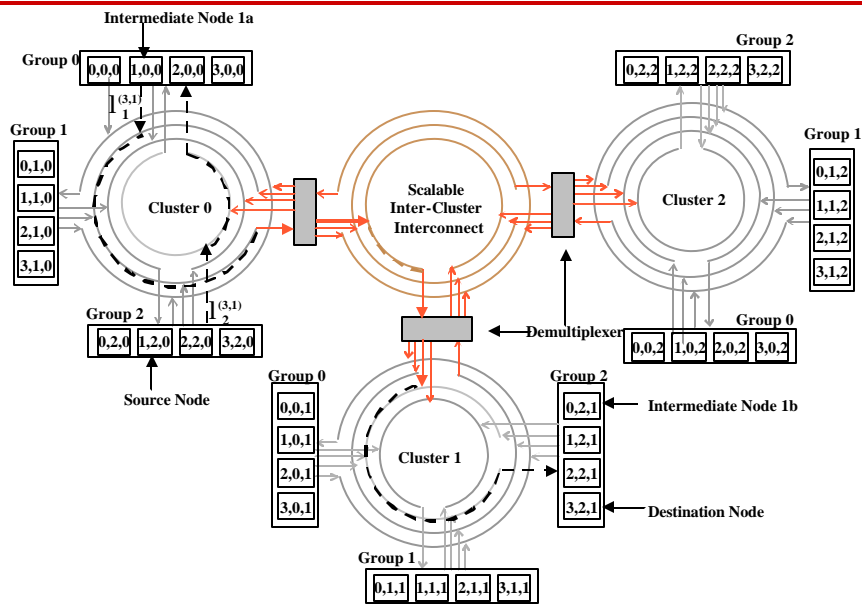
Multicast and Broadcast Communication



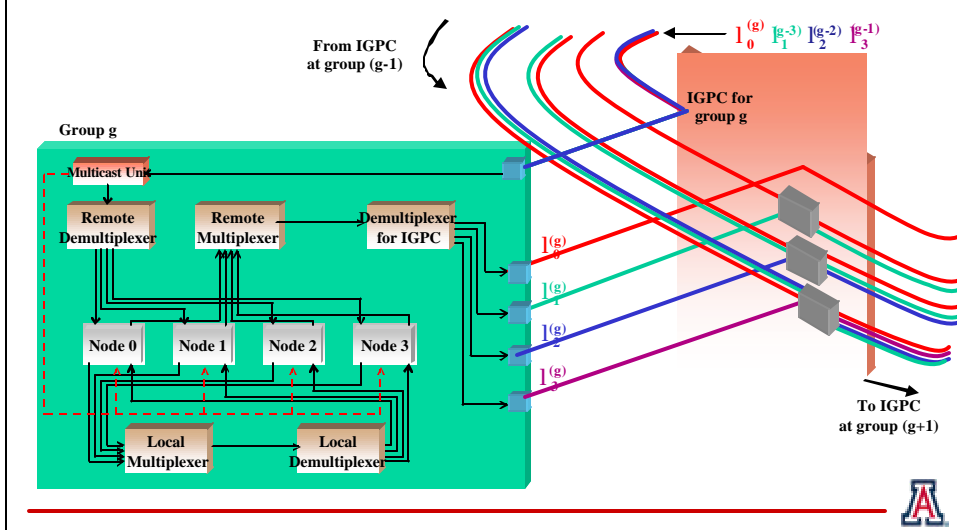
RAPID: Scalability Analysis



R(d,g,c): Inter-Cluster Communication



RAPID: Example Group Implementation



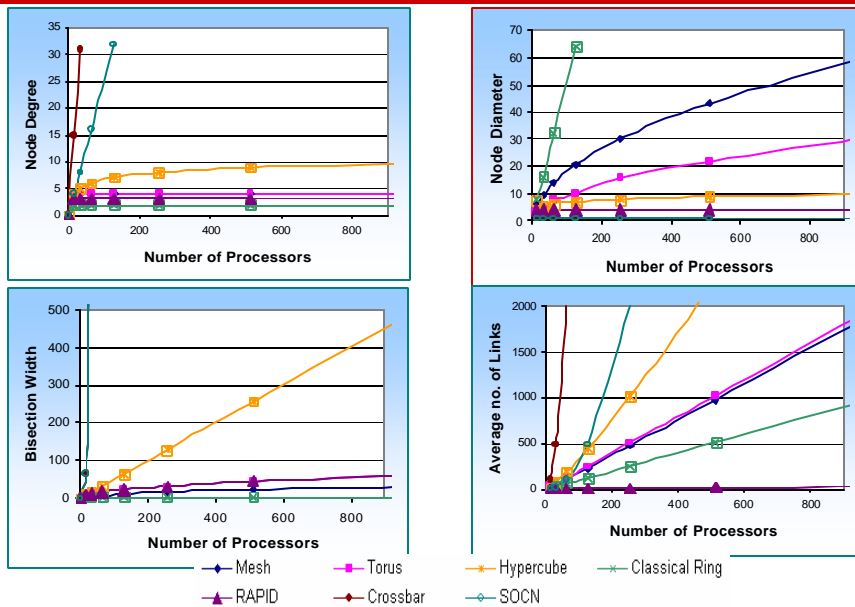
Performance Evaluation

The performance of RAPID was analyzed based on the following

- **Network Analysis:** Node degree, diameter, average bisection width, and number of links required - compared against Mesh, Torus Hypercube and the Classical Ring
- **CSIM Simulation:** Simulation based on timing information by modeling contention at the network interface and interconnect (theoretical analysis)
- **RSIM Simulation:** Complete system level simulation based on Splash-2 benchmarks and compared against the 2-D Mesh (preliminary results)
- **Power Budget Analysis:** Loss estimation and BER



Network Analysis: Degree, Diameter, Bisection Width and Number of links

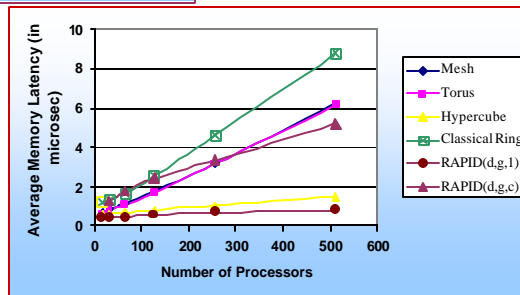
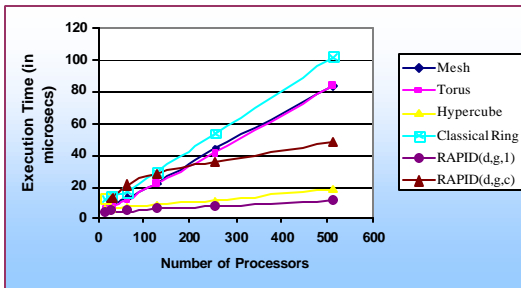


Network Performance: CSIM Simulation

- CSIM is a **process-oriented discrete-event simulator** that provides precise timing to evaluate RAPID
- RAPID is evaluated with some scalable networks, such as 2-D Mesh, 2-D Torus, Hypercube and Classical Ring based on execution time and remote latency
- A probabilistic model was used for simulation and was run on synthetic workloads
- Contention was modeled at all resources for both optical and electrical networks



CSIM Results: Execution Time, Remote Latency

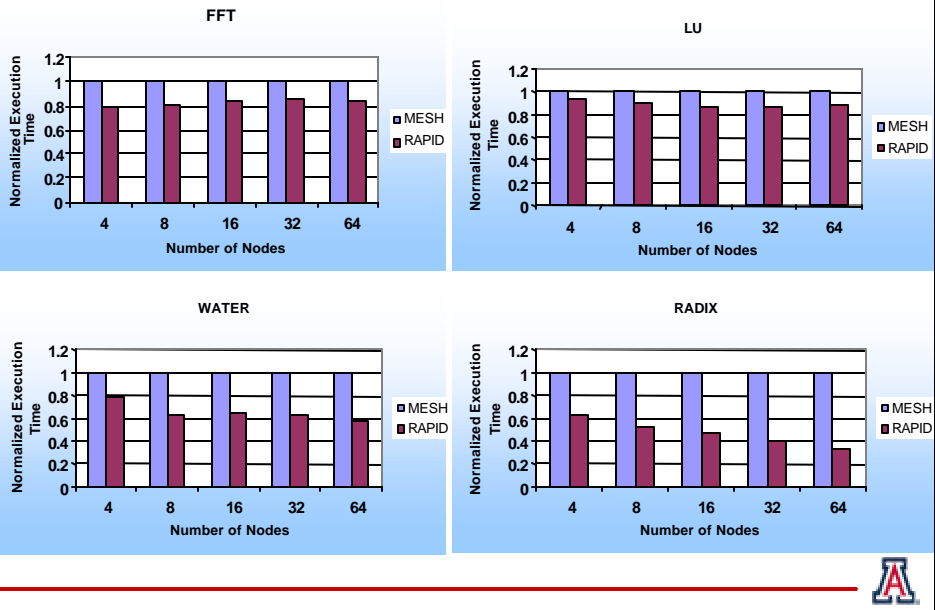


Network Performance: RSIM Simulation

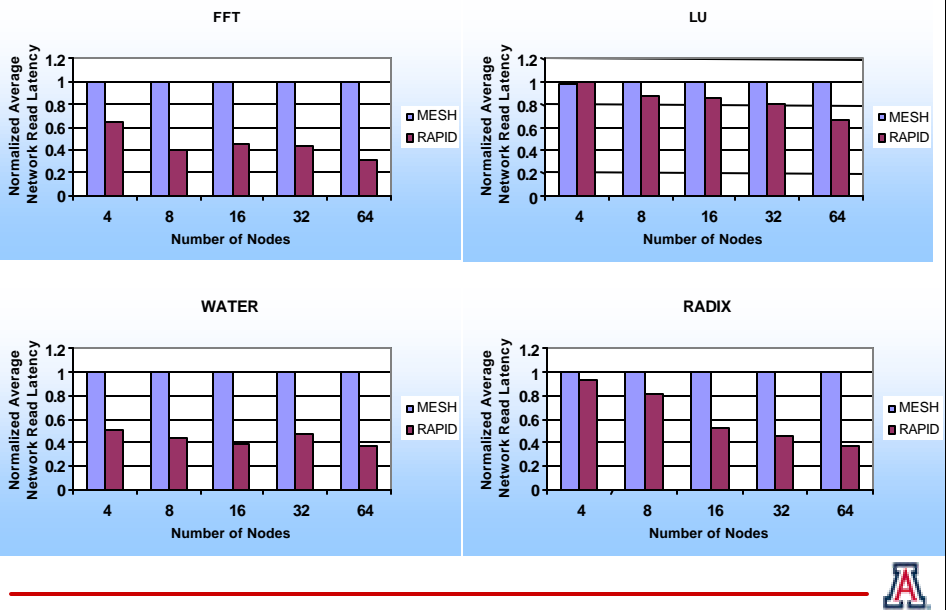
- RSIM is an **execution-driven simulator** that models an out-of-order super-scalar processor pipeline, 2-level cache hierarchy, split transaction bus on each processor
=> **Complete System Level Simulation**
- Benchmarks: Splash-2 suite applications that cover a spectrum of memory sharing and access patterns (FFT, LU, WATER, RADIX)
- Modified version of RSIM was used to evaluate R(d,g,1) network with 2-D mesh interconnect based on execution time and average network latency



RSIM Simulation: Execution Time



RSIM Simulation: Average Memory Latency



Power Budget Analysis

Intra-Group Losses	
VCSEL-waveguide coupling	0.2 dB
Fiber/Waveguide coupling	0.5 dB
Arrayed Waveguide Grating Loss	2.1×2
Directional Couplers Loss	$3 \text{ dB} \times \log_2(D)$
Receiver Coupling	0.2 dB
total loss (local)	$3 + 3 \times \log_2(D)$
Inter-Group Losses	
Additional Waveguide/Fiber	0.5 dB
Additional AWG Loss	2.1 dB
Directional Coupler Loss at IGPC	$0.225 \times G$
Circulator	0.5 dB
Fiber Bragg Grating	0.5 dB
Waveguide-to-Fiber	0.5 dB
total loss (Remote)	$10.1 + 3 \times \log_2(D) + 0.225 \times G$

- For intra-group interconnect, total loss is $-3 \text{ dB} - 3 \text{ dB} \times \log_2(D)$ and for the inter-group interconnect, the total loss is $10.1 - 3 \times \log_2(D) - 0.225 \times G$

=>With 32 wavelengths, we can have $32 \times 32 = 1024$ nodes using RAPID



Conclusion and Future Work

- RAPID reduced remote memory access latency in DSMs by maximizing the memory bandwidth, using innovative media access protocol and decentralized wavelength allocation
- RAPID $R(d,g,1)$ can reduce the latency for smaller system configurations by using more wavelengths and maintaining low diameter. Additionally, RAPID $R(d,g,c)$ can scale to very large configurations, yet provide low latency by using minimal wavelengths.
- In the future, we intend to build a prototype of a DSM interconnect using commercially available optical components

