Clemson University
TigerPrints

Publications | School of Computing

4-2003

AmpNet - A Highly Available Cluster Interconnection Network

Amy Apon
Clemson University, aapon@clemson.edu

Larry Bilbur
Belobox Networks

Follow this and additional works at: https://tigerprints.clemson.edu/computing_pubs

 Part of the Computer Sciences Commons

Recommended Citation
Apon, Amy and Bilbur, Larry, "AmpNet - A Highly Available Cluster Interconnection Network" (2003). *Publications* . 16.
https://tigerprints.clemson.edu/computing_pubs/16

This Conference Proceeding is brought to you for free and open access by the School of Computing at TigerPrints. It has been accepted for inclusion in Publications by an authorized administrator of TigerPrints. For more information, please contact kokeefe@clemson.edu.

AmpNet – A Highly Available Cluster Interconnection Network

Amy Apon, University of Arkansas

aapon@uark.edu

Larry Wilbur, Belobox Networks, Inc.

LarryWilbur@belobox.com

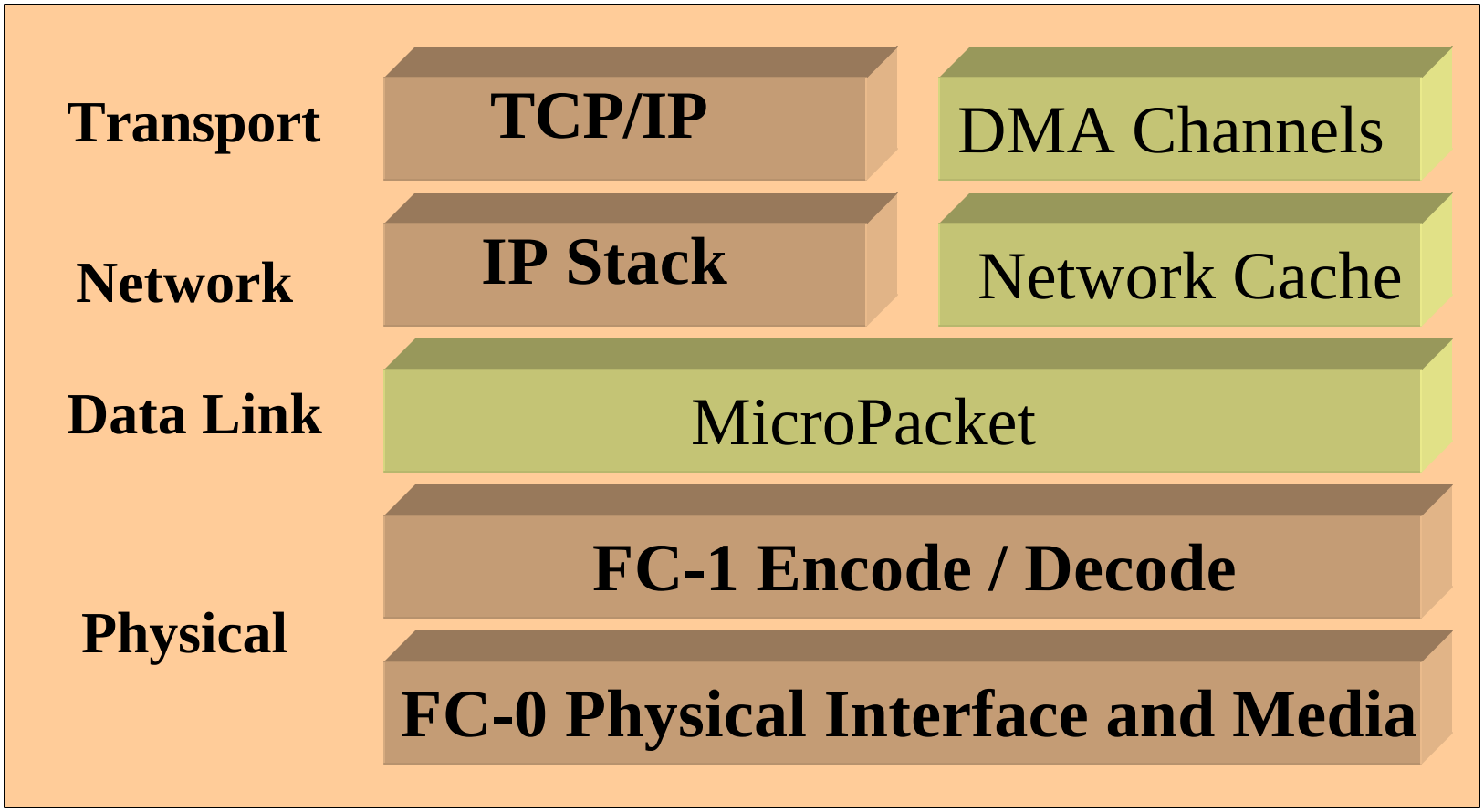
The AmpNet Network is also a Computer

Use Network Cache to keep the same information at every node.

- Nodes can leave and the data is intact.
- New nodes are assimilated with a cache refresh.
- The first network “database” created contains all the information required to operate the network
- The management information is ubiquitous.
- The network is self-managed and can heal after damage
- Applications can use the network to rebuild.

MicroPacket Technology

- **TCP/IP Standards**
- **Fibre Channel (FC-0 and FC-1) Specification**



MicroPacket Types

MicroPacket	Length	Mandatory
Rostering	Fixed	Yes
Data	Fixed	Yes
DMA	Variable	Yes
Interrupt	Fixed	Yes
Diagnostic	Fixed	Yes
D64 Atomic	Fixed	No

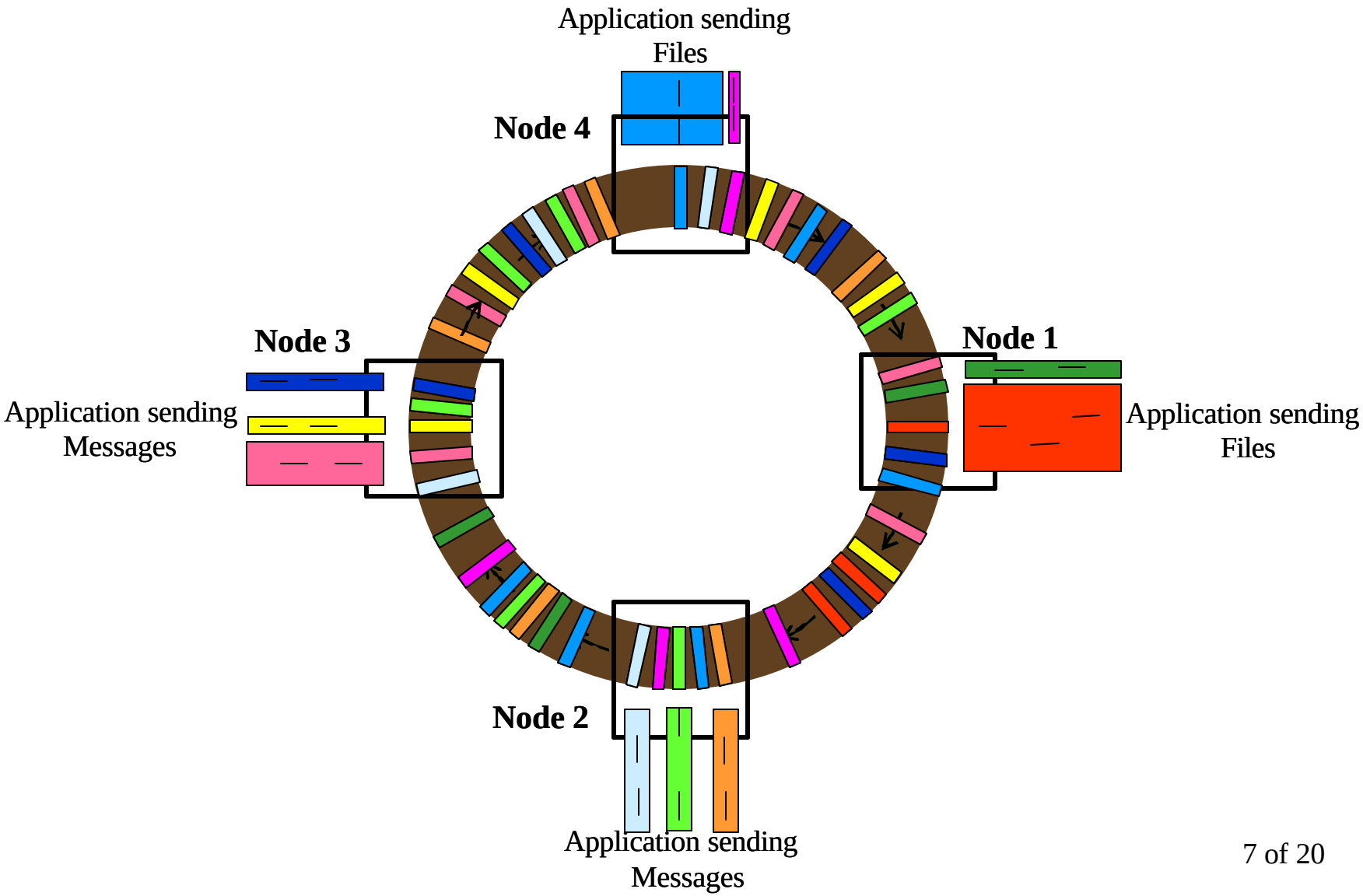
MicroPacket Fixed Format

	Byte 0	Byte 1	Byte 2	Byte 3
Word 0	**Control 0**	**Control 1**	**Control 2**	**Control 3**
Word 1	*Payload 0*	*Payload 1*	*Payload 2*	*Payload 3*
Word 2	*Payload 4*	*Payload 5*	*Payload 6*	*Payload 7*
EOF	**E**	**O**	**F**	**A**

MicroPacket Variable Format

	Byte 0	Byte 1	Byte 2	Byte 3
Word 0	Control 0	Control 1	Control 2	Control 3
Word 1	DMA Ctrl 0	DMA Ctrl 1	DMA Ctrl 2	DMA Ctrl 3
Word 2	DMA Ctrl 4	DMA Ctrl 5	DMA Ctrl 6	DMA Ctrl 7
Word 3	*Payload 0*	*Payload 1*	*Payload 2*	*Payload3*
Word 4	*Payload 4*	*Payload 4*	*Payload 4*	*Payload 4*
. . .	. . .	. . .	. . .	. . .
Word 18	*Payload 60*	*Payload 61*	*Payload 62*	*Payload 63*
EOF	E	O	F	a

AmpNet Can Insert Multiple Data Streams onto a Segment at Each Node



Network Flow Control

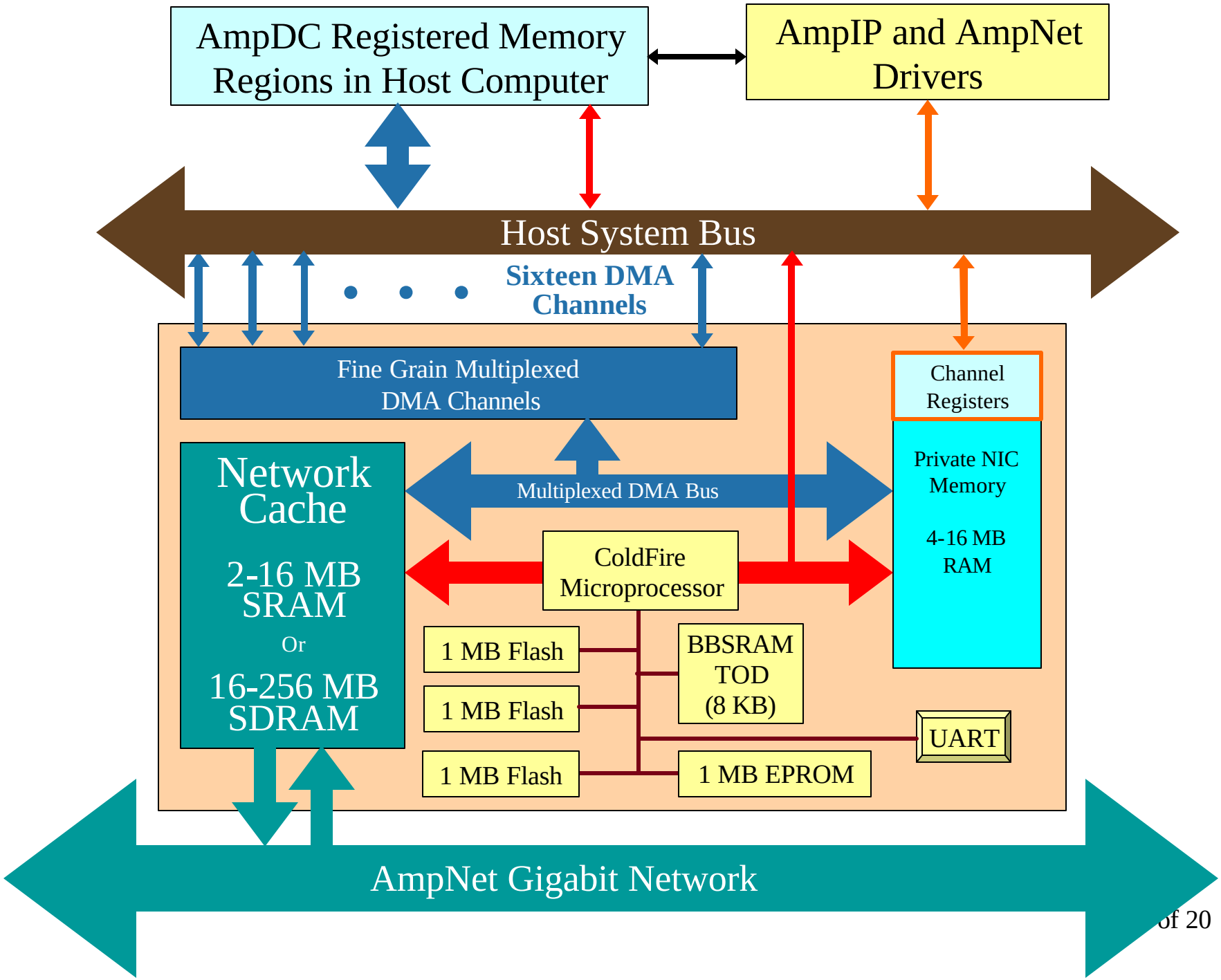
- Uses a variant of a register insertion ring
- Each node monitors its local view if the network and can increase or decrease its contribution to the total flow accordingly
- Even if everyone does a broadcast at the same time (all-to-all broadcast) the network is guaranteed to not drop packets

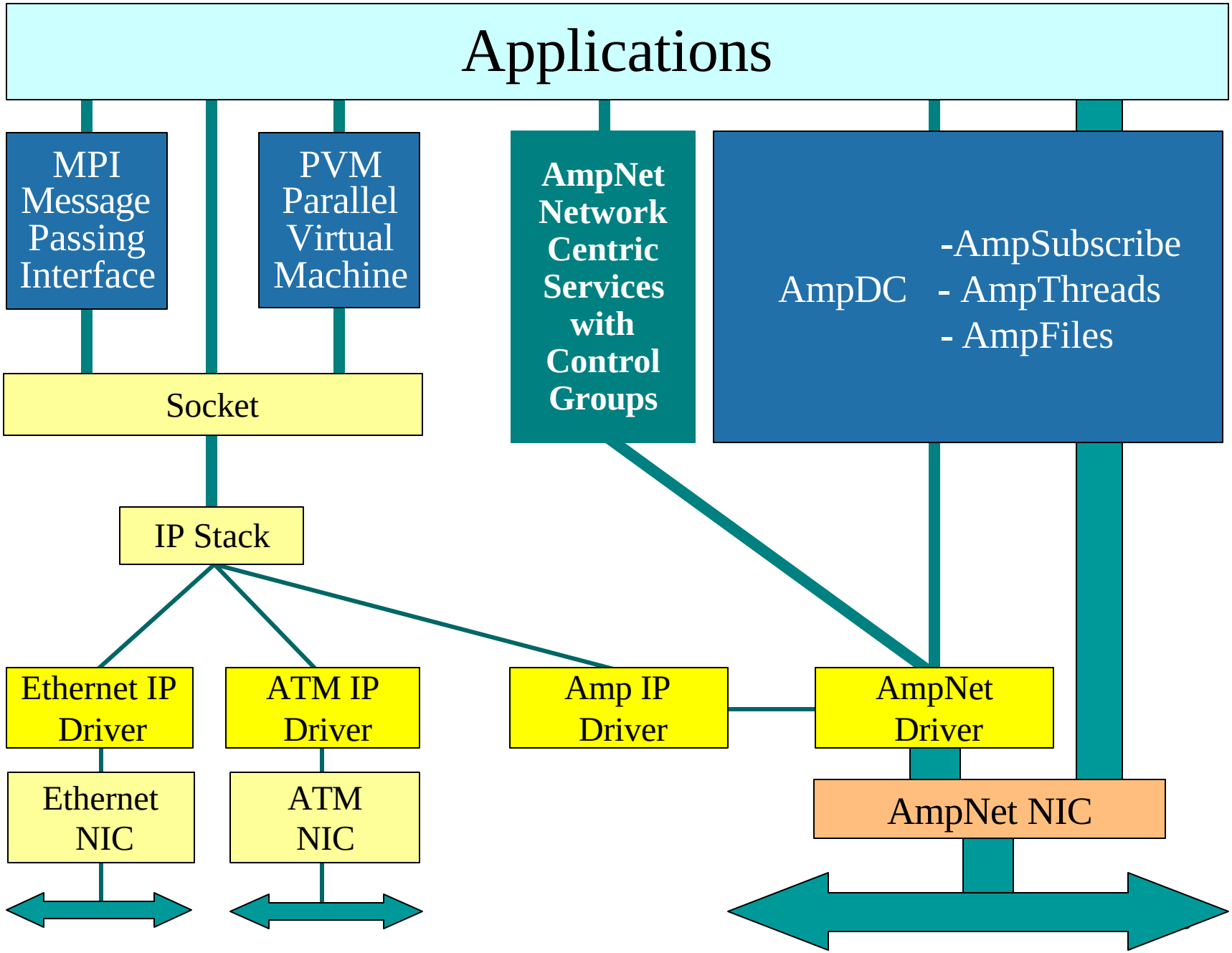
Cache consistency

- Two counters, at the start and end of every message – “Lamport counters”
- To read
  - **Start**: Read first counter, read last counter
  - If they agree, read data, else wait and go to **Start**
  - Read first counter, if changed go to **Start**
- To write
  - Just write

Cache coherence

- Write conflicts are handled at the user level using AmpNet locking primitives implemented in software (network semaphores)
- To maintain coherence between memory mapped from host memory to AmpNet NIC memory, updates in host memory are written through to AmpNet NIC memory – no caching is allowed in local host cache


AmpDC Registered Memory Regions in Host Computer
AmpIP and AmpNet Drivers
Host System Bus
Sixteen DMA Channels
Fine Grain Multiplexed DMA Channels
Channel Registers
Network Cache
2-16 MB SRAM
Or
16-256 MB SDRAM
Multiplexed DMA Bus
Private NIC Memory
4-16 MB RAM
ColdFire Microprocessor
1 MB Flash
1 MB Flash
1 MB Flash
BBSRAM TOD (8 KB)
UART
1 MB EPROM
AmpNet Gigabit Network

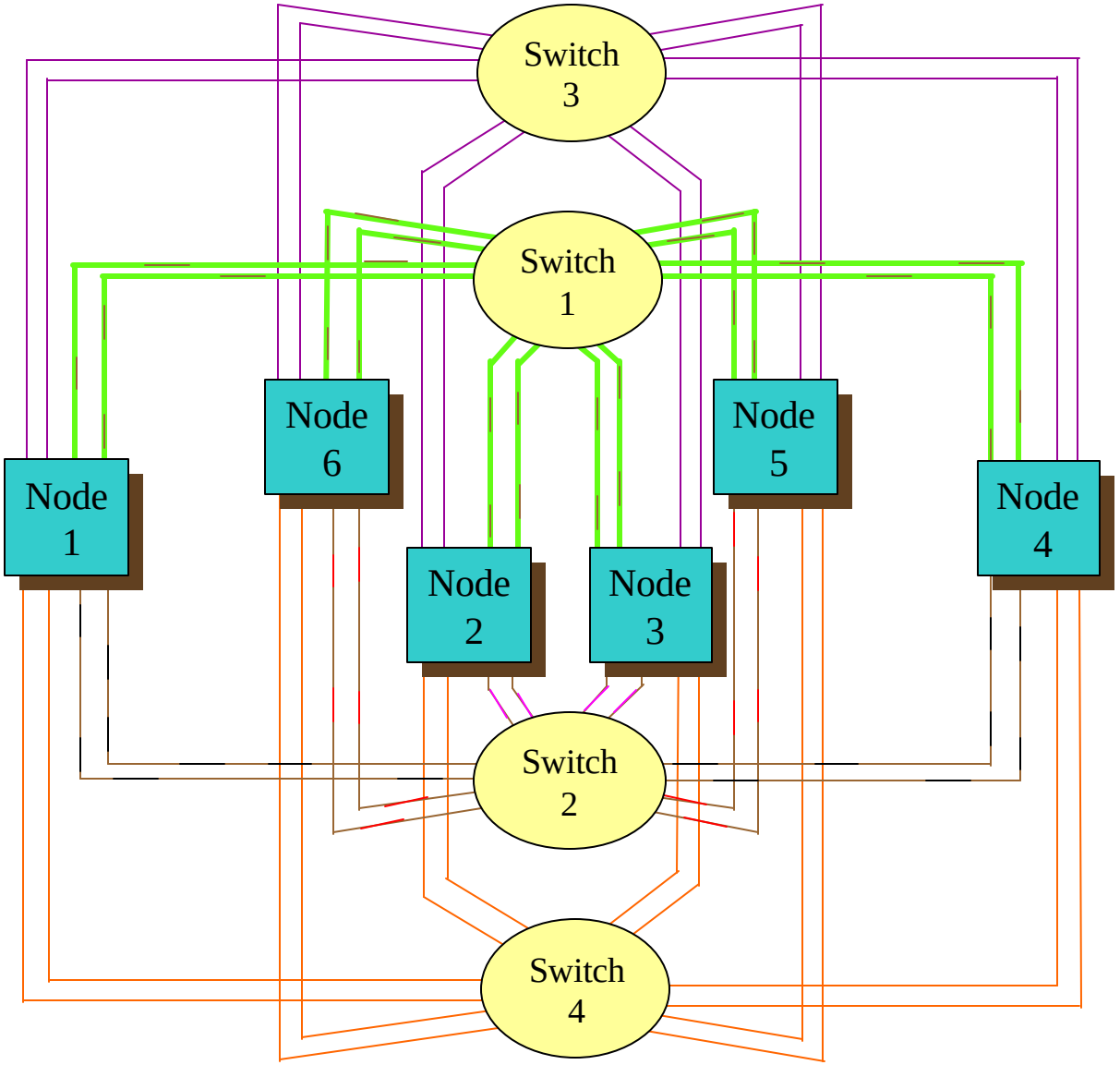

Applications
MPI
Message
Passing
Interface
PVM
Parallel
Virtual
Machine
AmpNet
Network
Centric
Services
with
Control
Groups
AmpDC
-AmpSubscribe
- AmpThreads
- AmpFiles
Socket
IP Stack
Ethernet IP
Driver
ATM IP
Driver
Amp IP
Driver
AmpNet
Driver
Ethernet
NIC
ATM
NIC
AmpNet NIC

The computer can be fault tolerant and self-healing with no data loss

This means it never goes down
and never loses your data

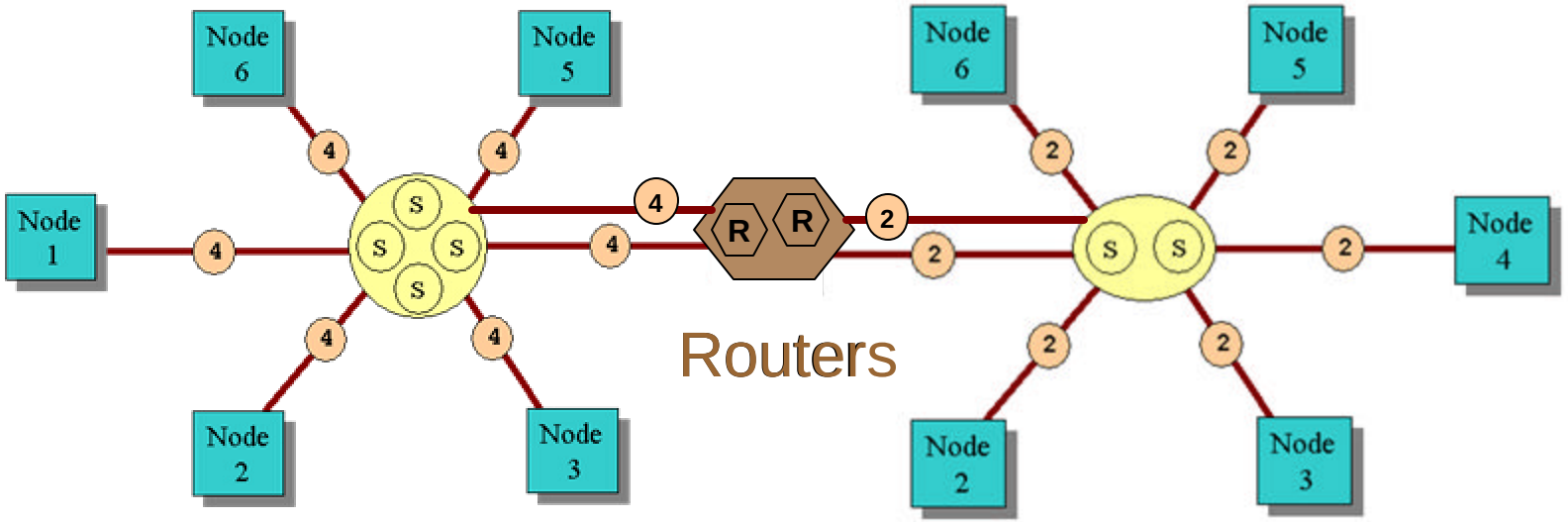
- **Requires Multiple Paths**
- **Requires a Rostering Algorithm**
- **Requires Network Cache**

Quad-Redundant Switched Network - **Physical**



Quad-Redundant Segment

Dual-Redundant Segment



Rostering

- Algorithm starts automatically whenever a failure is detected
- A modified flooding algorithm that explores the network for available paths and allows the creation of the largest possible logical ring
- Packets are forwarded according to rostering rules
- Rostering completes in two ring-tour times
  - 1 to 2 milliseconds, depending on the number of nodes and the length of the fiber

Nodes Easily Enter the Network

- Every node is a real-time Micro Computer
- Managed by AmpNet Distributed Kernel (AmpDK)
- Instantly Self-Boots - Doesn’t need a Host
- Conforms to assimilation rules before coming online
  - Enforces version compatibilities across the network
  - Enforces the same rules for all computers (VxWorks, Linux, Windows 2000, etc.)
- Supports embedded multi-threaded application processes

Application Failover is Possible

- Network failures detected by hardware
  - Hardware Self Heals via a Rostering Algorithm
- Distributed Micro-Kernel provides Self Reconstitution
- Built-in diagnostics certify new configuration
- Smart Data Recovery is supported by Cache Refresh
- Cached Database reflects new configuration

Node Re-Entry and Application Failover

- Millisecond application failure detection
- Application definable fail-over period
- Control passes to the best qualified computer
- Applies Application Rules of Recovery

No down time and no loss of data!

AmpNet® real-time Computing Fabric℠

- Application Survivability
- Network Centric Services
- Fault and Damage Tolerance
- Self Healing
- No Data Loss
- No Down Time
- No Loss of Service

Unlimited computing capacity and seamless growth for non-stop real-time applications

NETWORKS that EVOLVE℠

Built on FPGA soft logic, our AmpNet gigabit network conforms to evolving standards and meets tomorrow's requirements

BELOBOX

Follow the evolution at: belobox.com

Belobox Networks, Inc.
18 Technology Drive, Suite 130, Irvine, CA 92618-2310
phone (949) 727-4115 or 1-800-BELOBOX, fax (949) 727-2149