



# HyperTransport Over Ethernet - A Scalable, Commodity Standard for Resource Sharing in the Data Center

Jeffrey Young, Sudhakar Yalamanchili

*School of ECE*

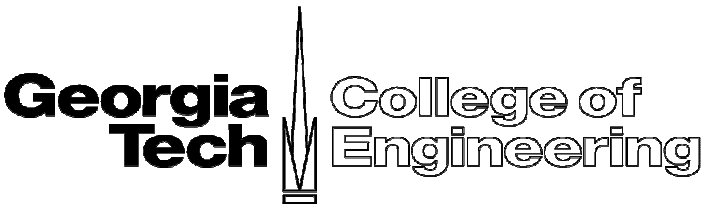
*Georgia Institute of Technology*

Brian Holden, Mario Cavalli

*HyperTransport Consortium*

Paul Miranda

*AMD*


# Talk Outline

- Why do we need another HT-related standard?
- Why combine HyperTransport with Ethernet?
  - Performance
  - Cost and Market Share
  - Scalability
- What does a HyperTransport over Ethernet (HToE) specification require?
  - What solutions does the specification propose?
- How can HToE be used?

# Motivation

- HyperTransport has gained traction as a low-latency, on-package interconnect
- High Node Count (HNC) specification and HTX boards have created opportunities for system-wide usage and better resource sharing
- However…
  - Limitations on coherent HT links
  - Requires custom HyperTransport cables
- How do we further promote the usage of HT in data centers?

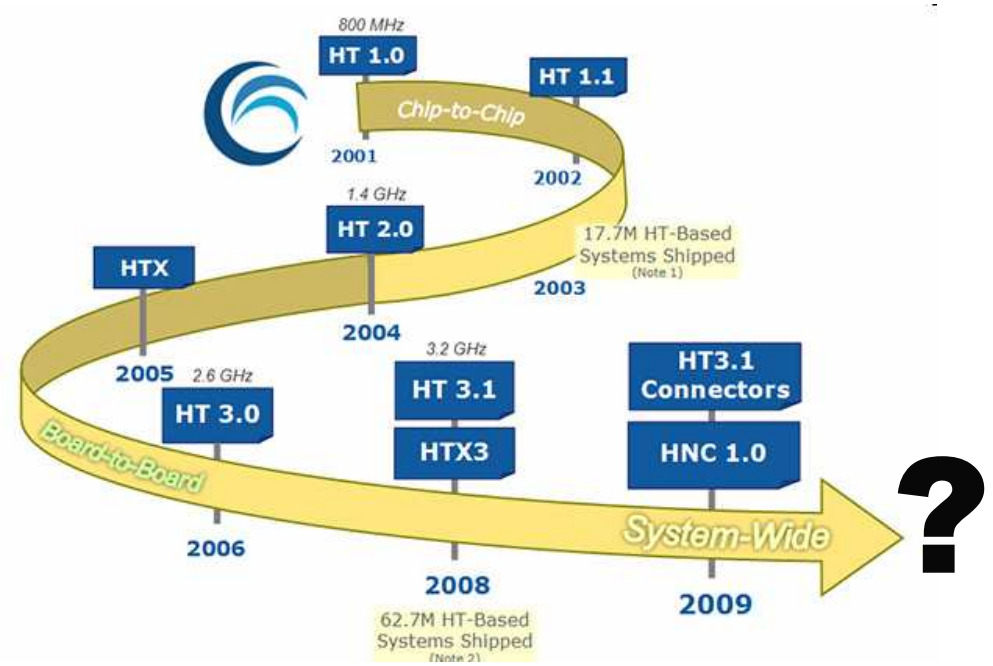



*Photo Source: http://www.hypertransport.org*

# Towards a Solution – HToE Focus

- Problem: Coherent HyperTransport links are limited
  - Solution: Focus on non-coherent resource sharing, including memory, disks, and accelerators
- Problem: Custom HyperTransport HW required to connect blades
  - Solution: Co-opt existing *commodity* data center interconnects like Ethernet and InfiniBand, while still promoting HNC

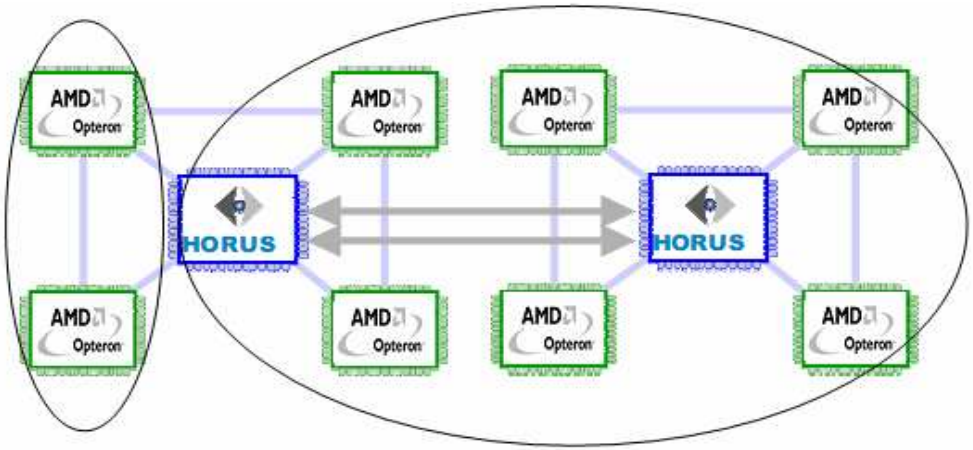



*Photo Source: http://www.hypertransport.org/docs/tech/horus_external_white_paper_final.pdf*

# Towards a Solution – HyperShare and HToE


- HyperShare is a design platform that aims to promote non-coherent resource sharing in data centers and HPC clusters

- Focuses on three commodity interconnects:
  - 3D torus via PCI Express-enabled HNC cards
  - InfiniBand via HToIB
  - Ethernet via HToE

- This talk addresses HyperTransport over Ethernet using 10, 40, and/or 100 Gigabit Ethernet

*Photo Source: http://www.hypertransport.org*

# Motivation for HToE

- IEEE Standard 802.3ba was ratified in June, 2010
  - Provides for 40 and 100 Gbps Ethernet (GE) MAC rates
  - Backwards compatible with previous Ethernet standards
  - Defines physical standards for copper and optical substrates
- Performance
  - Latency improvements
- Cost and Market Share
  - Ethernet continues to be widely deployed, despite cost issues
- Scalability
  - Recent research focused on large Ethernet deployments

ethernet alliance

*Photo Source: http://www.ethernetalliance.org*

# Ethernet Performance

- Recent studies have demonstrated low-latency 10GE
  - 8 - 10 microseconds with iWARP[1]
  - Comparable to IB at 2 - 6 microseconds
- In many cases, TCP/IP increases latency
  - TCP overhead can be up to 1 - 2 microseconds for HPC applications[2]
- To improve performance, HToE tries to reduce OS and network stack latency:
  - Focuses on Layer 2 (L2) solution
  - Uses non-coherent, global address space to reduce need for OS intervention in memory sharing


1) *Swamy N. Kandadai and Xinghong He. Performance of hpc applications over infiniband, 10 gb and 1 gb ethernet. 2010.*
2) *Pavan Balaji, Wu-chun Feng, and Dhabaleswar K Panda. Bridging the ethernet-ethernot performance gap IEEE Micro, 26:24–40, May 2006.*

*Photo Source: http://www.intel.com/Products/Server/Adapters/Server-Cluster/Server-Cluster-overview.htm*

# Ethernet Market Share and Cost

- Slow adoption for 10GE, but 1GE and 10GE still have 45.6% share of Top500 supercomputer list
  - InfiniBand has made inroads in HPC due to high costs of 10GE
- Large invested infrastructure for Ethernet
  - Supports many standard TCP/IP applications
- Specifications for converged fabrics
  - Data Center Bridging (2011?)
  - Fibre Channel over Ethernet (2010)
  - RDMA over Converged Ethernet (2010)
    - InfiniBand over Ethernet!

**Top 500 Interconnect Family Share, Nov. 2010**

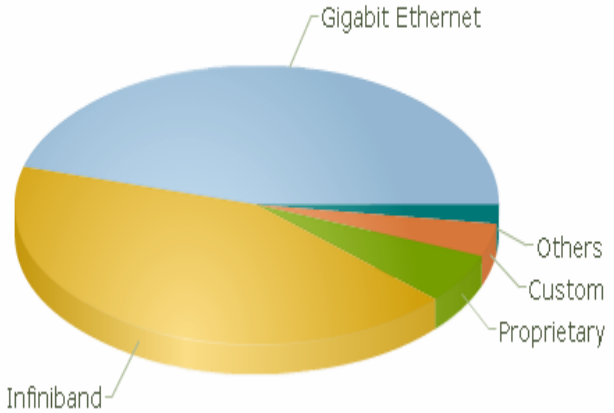



*Photo Source: http://www.top500.org/charts/list/36/connfam*

# Ethernet Scalability

- Research into multipath routing for L2 switches improves scalability[3]
  - Reduces issues with Spanning Tree Protocol
- Switch makers focused on scalable Fat Tree networks
  - Fulcrum, Fortinet (Woven), etc.
- Fibre Channel over Ethernet implementers working on scalable, converged fabric
  - Reduces need to switch between IB, Ethernet, and FC for different parts of data center

*3) M. Schlansker, J. Tourrilhes, Y. Turner, and J.R. Santos. Killer fabrics for scalable datacenters. In Communications (ICC), 2010 IEEE International Conference on, pages 1 –6, May 2010.*

# HToE Motivation – Bringing it Together

- Performance
  - Ethernet has similar latency to InfiniBand; in many cases it is good enough
- Cost, Market Share
  - 10 GE has suffered due to lack of adoption at the hands of IB, *but…*
  - 40, 100 GE specification gaining early implementers
  - Cost likely will come down due to widespread adoption of 1 GE and relevant IT staff
  - FCoE and RDMA over Converged Ethernet point to big bets in favor of Ethernet
- Scalability
  - Research and Data Center Bridging standards leverage existing knowledge of 1 and 10 GE-based networks

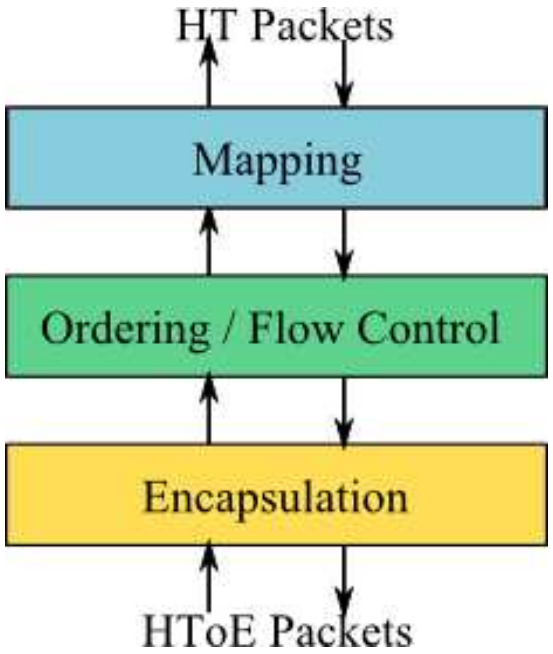
# HToE Motivation – Bringing it Together

It makes sense to encapsulate HyperTransport over a switched Ethernet fabric!


*Photo Sources: http://www.hypertransport.org, http://en.wikipedia.org/wiki/File:Cat_5.jpg*

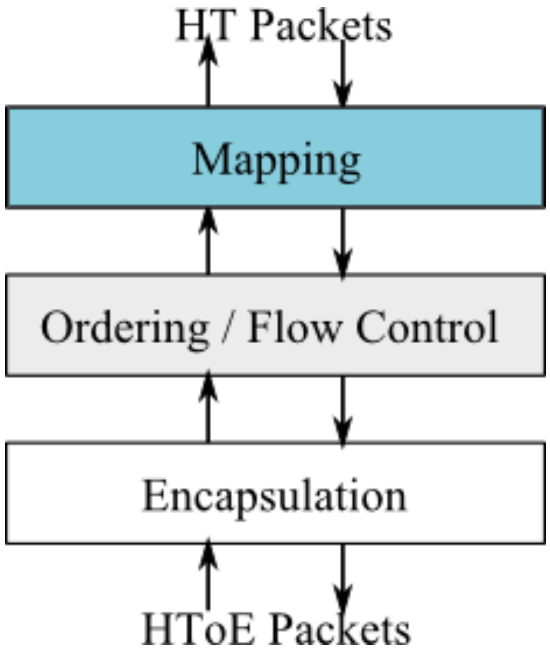
# HToE Specification – Core Functionality

- On-package and off-package addressing
  - "Mapping" layer handles global address space mapping
- Ordering/Flow Control
  - Translate credits from on-package to Ethernet
  - Maintain HyperTransport Ordering
- Encapsulation
  - Handles packing HT packets, retry, and error handling



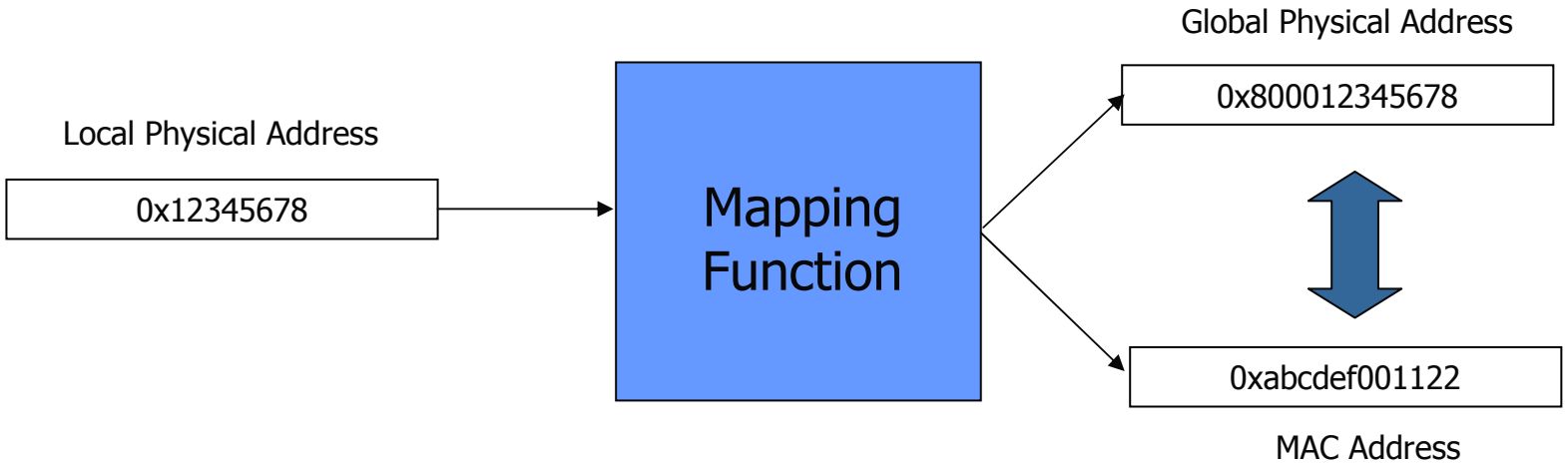
# HToE Specification – Mapping Layer

- On-package and off-package addressing
  - "Mapping" layer handles global address space mapping
- Ordering/Flow Control
  - Translate credits from on-package to Ethernet
  - Maintain HyperTransport Ordering
- Encapsulation
  - Handles packing HT packets, retry, and error handling



# HToE Requirement: Mapping

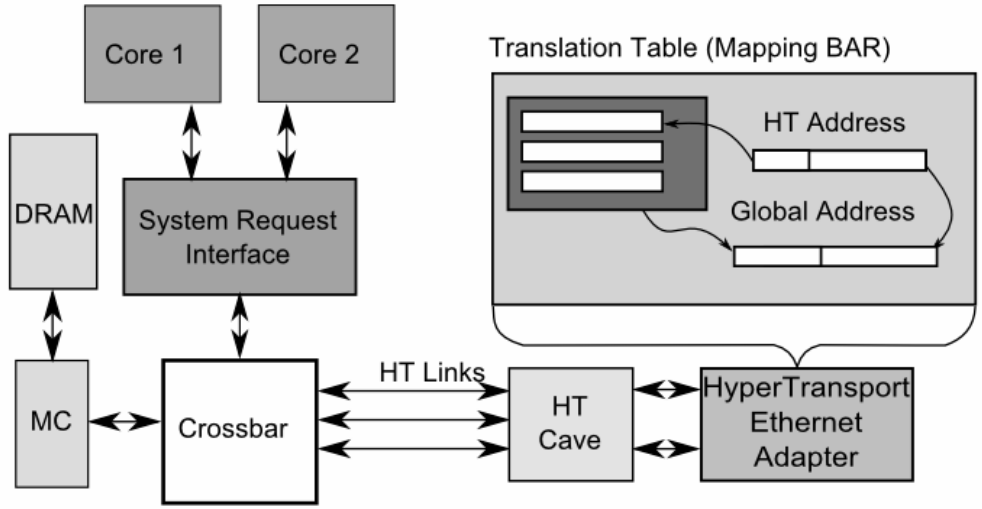
- Would like to map I/O devices and DRAM into a global physical address space
- Needs to translate between local HT address and global address
  - Associate global address with Ethernet MAC address
- Mapping must not change how local HT links interpret addresses



# HToE Specification – Mapping

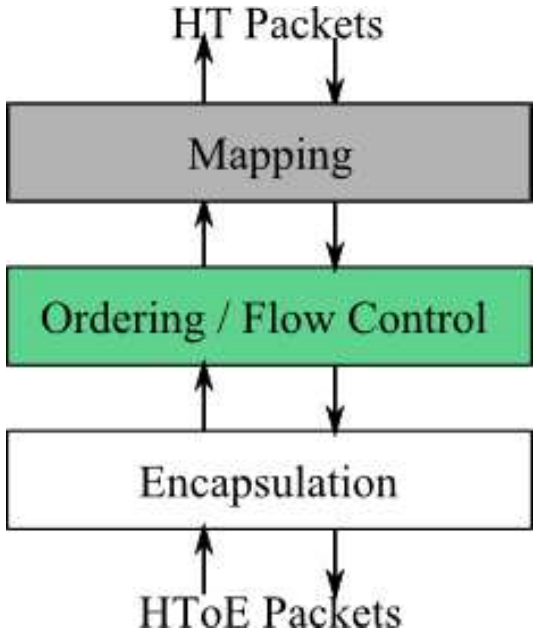
- Memory addresses on each node form subset of 64 bit, global physical address space
  - Mapping function allows for remote puts/gets
  - Specifics of mapping table left up to implementers
  - Global mappings can be communicated with Data Center Bridging Exchange Protocol (DCBX)

- Tag remapping optimization
  - Allows for increased scalability for incoming requests by remapping Source Tag bits



# HToE Specification – Ordering and Flow Control

- On-package and off-package addressing
  - "Mapping" layer handles global address space mapping
- Ordering/Flow Control
  - Translate credits from on-package to Ethernet
  - Maintain HyperTransport Ordering
- Encapsulation
  - Handles packing HT packets, retry, and error handling

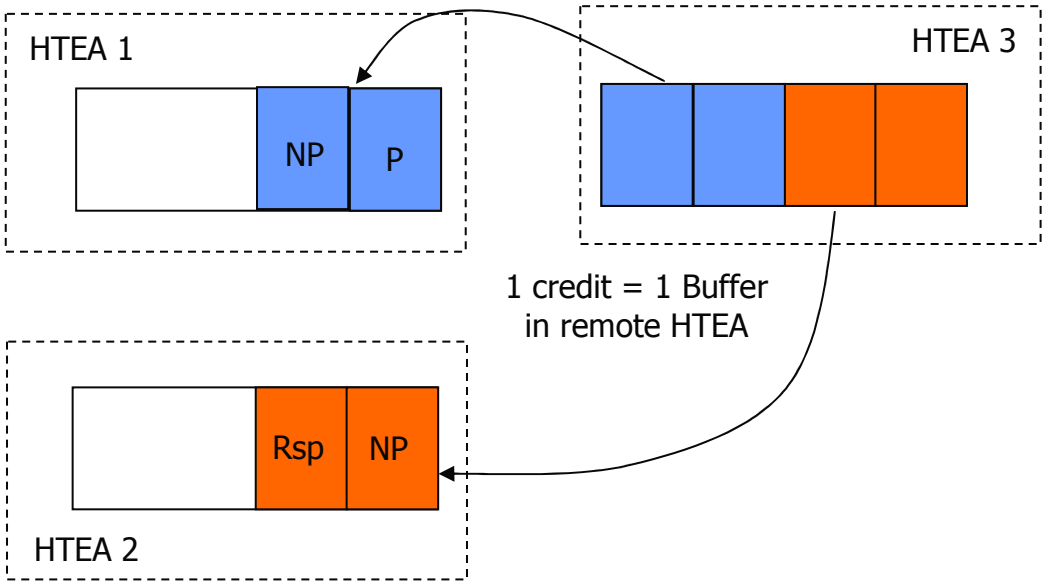


# HToE Requirement: Ordering and Flow Control

- HyperTransport is a credit-based, point-to-point protocol
  - We must allocate credits to multiple senders with switched Ethernet
- HToE must implement flow control when HT credits run out
  - We could use DCB per-flow flow control, but is this sufficient?
- HT Ordering Requirements must preserve ordering on all HT virtual channels
  - Must maintain deadlock freedom in face of potential Ethernet packet loss

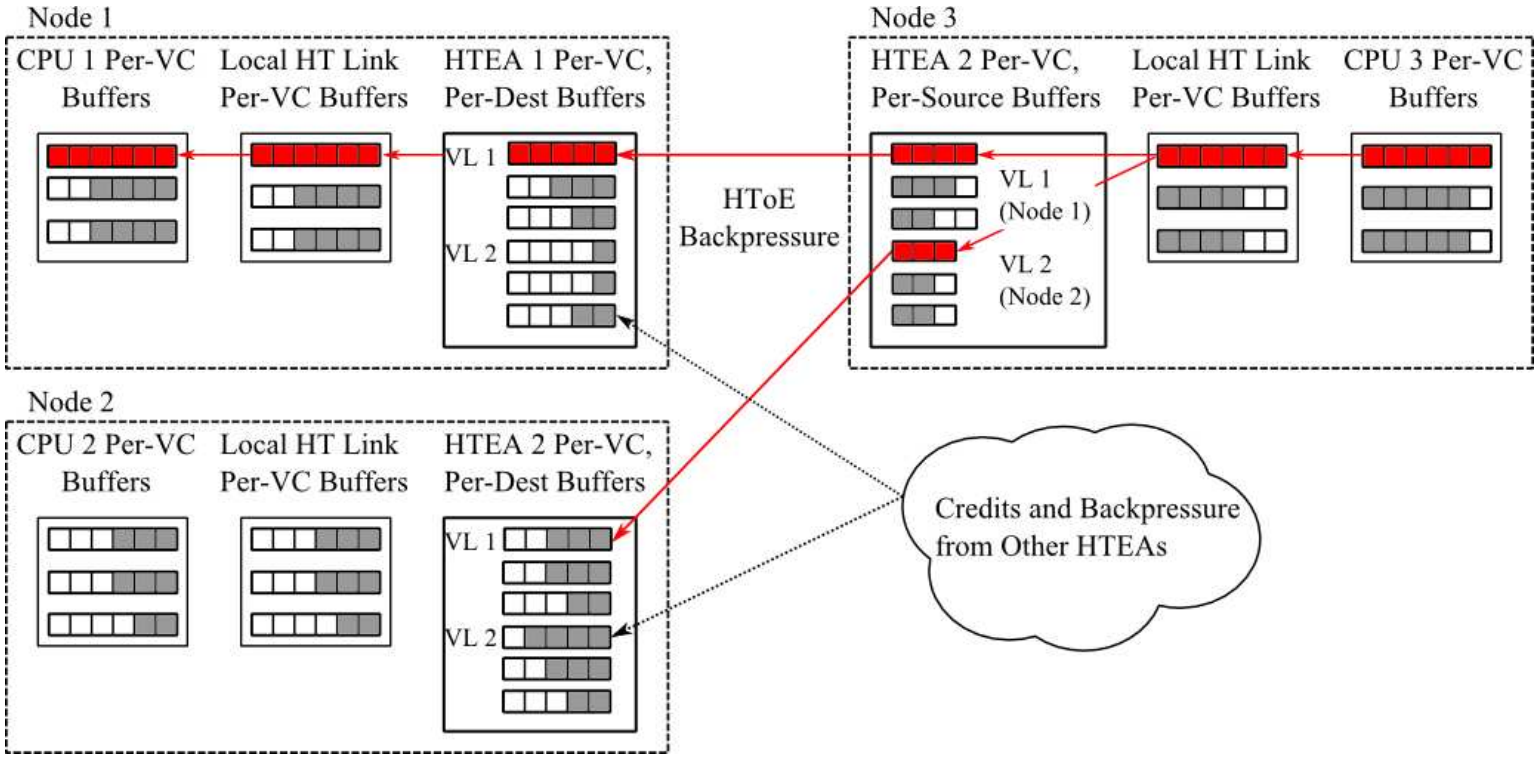
# HToE Specification – Ordering and Flow Control

- HT packets must have a credit from remote HTEA before being encapsulated into Ethernet packet
  - Preserves ordering between HT packets on source-destination pair
  - Each credit equals one physical buffer in the remote HTEA
- End-to-end flow control links HT credits to HToE traffic
  - Lack of physical buffers results in backpressure



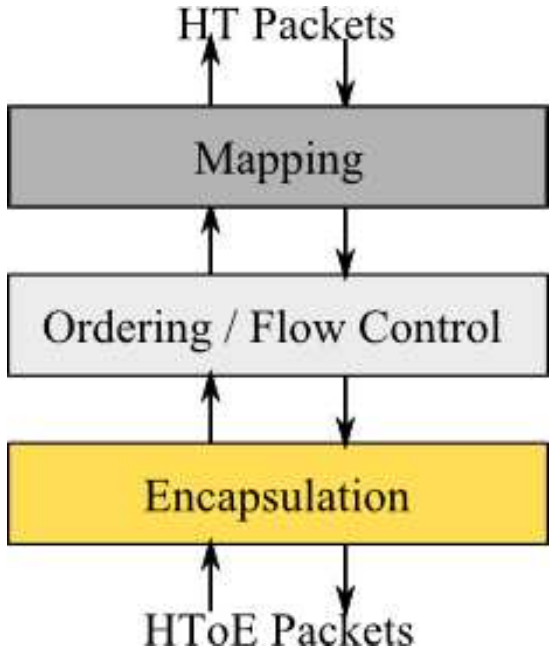
# HToE Specification – Backpressure

- Lack of credits (remote HTEA buffers) prevents HToE packets from being sent across the network



# HToE Specification – Encapsulation

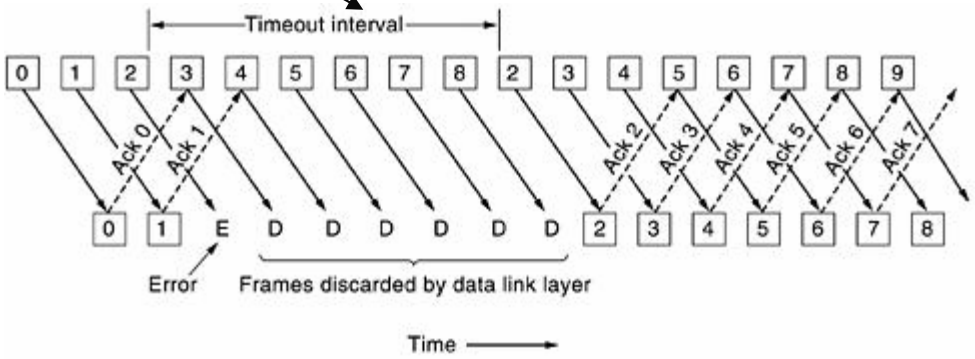
- On-package and off-package addressing
  - "Mapping" layer handles global address space mapping
- Ordering/Flow Control
  - Translate credits from on-package to Ethernet
  - Maintain HyperTransport Ordering
- Encapsulation
  - Handles packing HT packets, retry, and error handling



# HToE Requirement: Failure Recovery and Retry

- HT 3.1 recovery
  - "Poison" bit in response packets
  - Retry algorithm that uses Go-Back-N
  - Warm, cold reset of links

| Bit-Time | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 |
|---|---|---|---|---|---|---|---|---|
| 0 | Isoc | Rsv | Cmd[5:0]: 110000 | | | | | |
| 1 | PassPW | Bridge | Rsv | UnitID[4:0] | | | | |
| 2 | Count[1:0] | | Error0 | SrcTag[4:0] | | | | |
| 3 | Rsv/RqUID | | Error1 | Rsv/RspVCSet | | | Count[3:2] | |




- HToE recovery
  - Must report errors to remote requesting node
  - Must communicate resets to remote node
  - Resets must only affect one source and one destination (virtual link)

*Photo Sources: http://www.hypertransport.org, http://en.wikibooks.org/wiki/File:Go_Back_N.jpg*

# HToE Specification – Retry and Flow Control

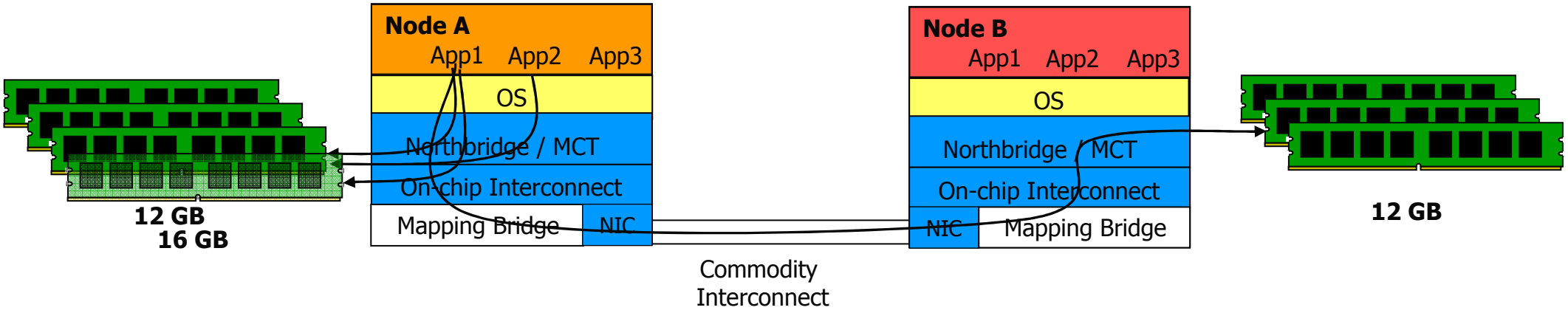
- Repurpose HT 3.1 retry algorithm to apply to **Ethernet** packets
  - Removes need for Ethernet retry mechanism
- Cold and warm reset apply only to one virtual link
  - Spec defines packet header for transmitting resets
  - Removes chance that reset could impact other connections
- Keep track of sent HT packets to allow for error notification
  - Remote failure triggers response with "poison" bit set

# HToE Specification – Potential Use Cases

- Resource sharing is game-changing application for data centers
  - Traditional data centers are overprovisioned to meet peak demand
- PGAS Support for Virtual DIMMs
  - "Virtualize" Dimms to allow for average provisioning of DRAM
- Pooled Accelerators for Reducing Cluster TCO and Power
  - Remote accelerators accessed via HToE allow for more reduced power, TCO in the data center

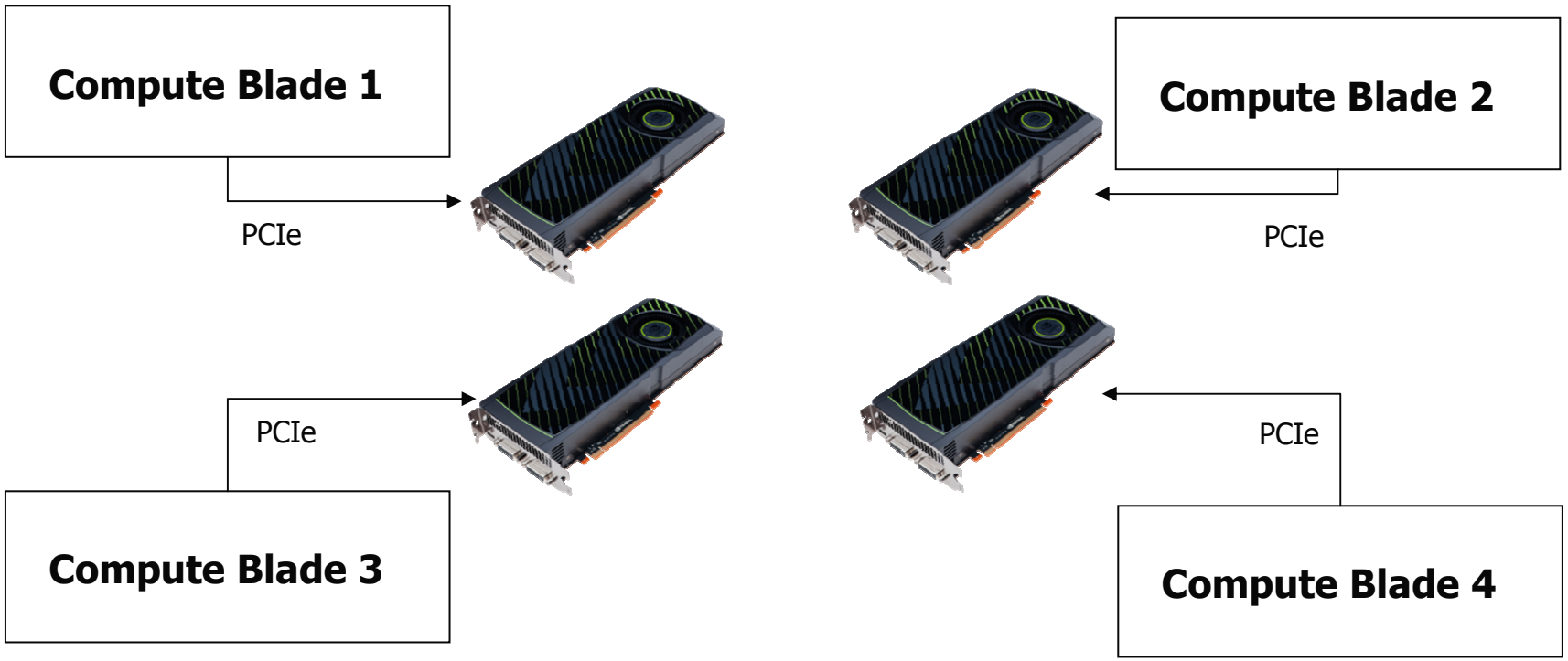
*Photo source: http://eetd.lbl.gov*

# HToE Applications – Virtual DIMMs for DRAM Power Efficiency




- Create a "virtual DIMM" abstraction that allows for transparent, low-latency DRAM sharing over commodity interconnects
  - Remote access is handled at hardware layer with OS control path interaction for setup
- Cost and power savings for 10,000 core cluster with 50% reduction:
  - Reduce memory from 64 GB to 32 GB (1.5 V DDR3)
  - Savings: $750,000 and 8,500 Watts

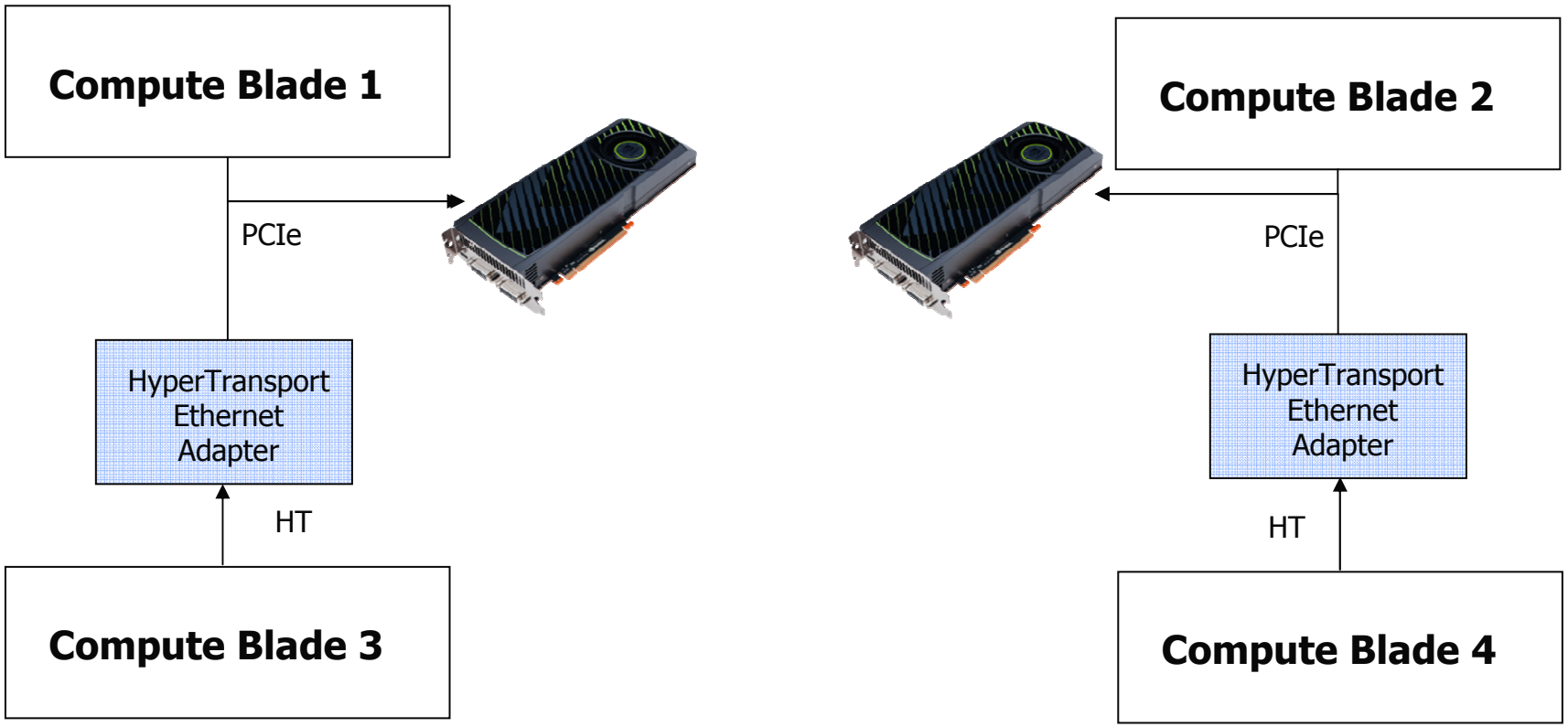
# HToE Applications – Pooled Accelerators to Reduce TCO and Power Usage




- Remote access to GPUs currently limited to OS-level approaches
  - HToE can enable low-latency access to remote hardware via PCI reads/writes

*Photo source: http://www.nvidia.com/object/product-geforce-gtx-570-us.html*

# HToE Applications – Pooled Accelerators to Reduce TCO and Power Usage




- Sharing one GPU between two blades could lead to 75% reduction in capital costs, power usage, capital cost
  - $328,125 cost savings for GeForce GTX 570 with 10,000 core data center
  - 28,125 Watt power reduction (assumes GPU uses 30 Watts at idle)

# Conclusions

- HyperTransport over Ethernet specification is aimed at data centers and HPC clusters
  - Uses commodity interconnect to boost appeal of HT as system-wide interconnect
  - Performance makes it feasible for HPC
- HToE specification proposes several engineering solutions to HT encapsulation
  - Tag remapping, credit-based backpressure, retry and recovery
- HToE shows potential for resource sharing
  - DRAM, accelerators, possibly other models

# More Information

- HyperTransport over Ethernet Specification and HyperShare information:

  http://www.hypertransport.org/

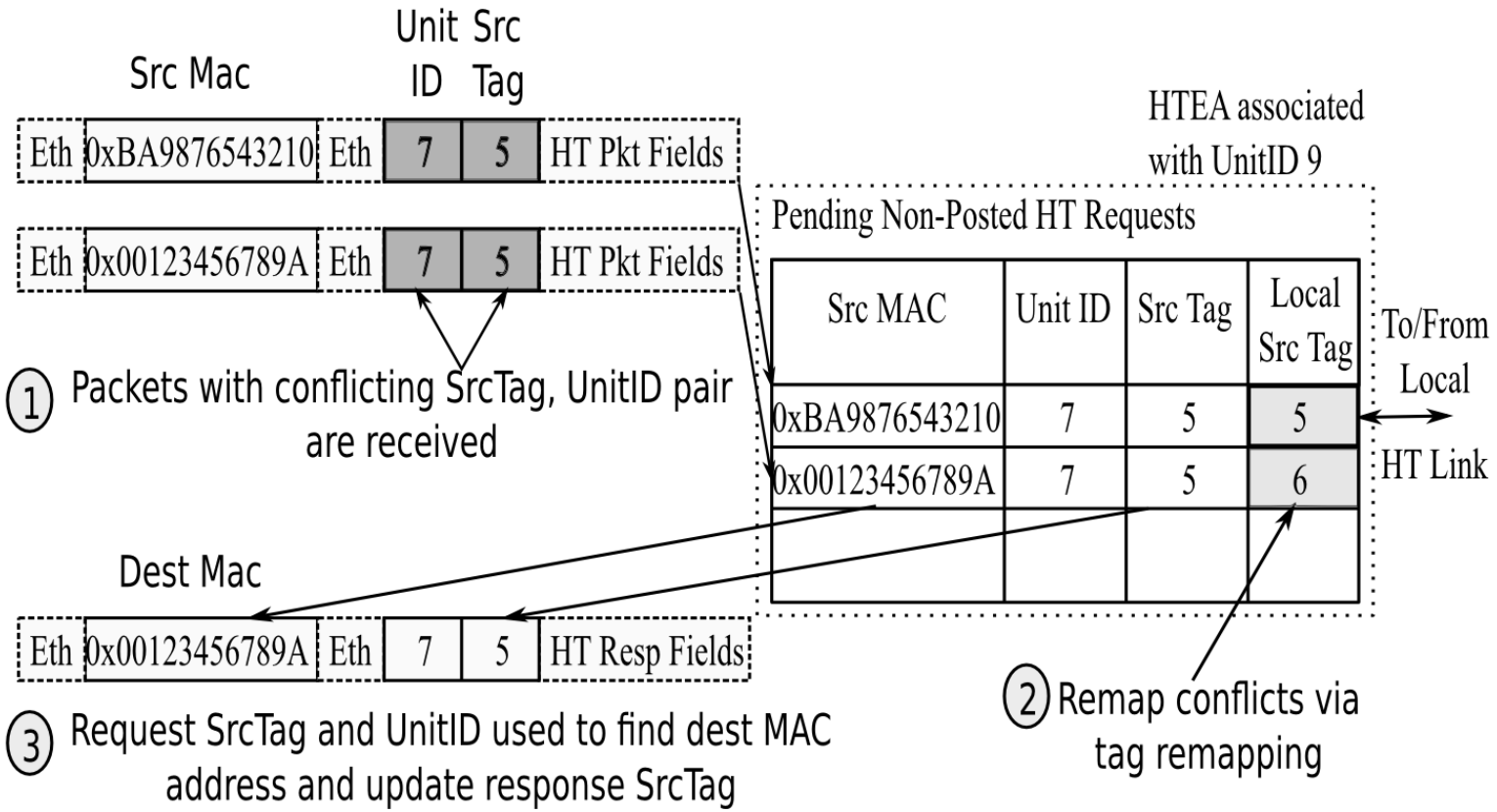
- HToE-related research at Georgia Tech:

  http://www.ece.gatech.edu/research/labs/casl/netmem.html

***Thank You!***

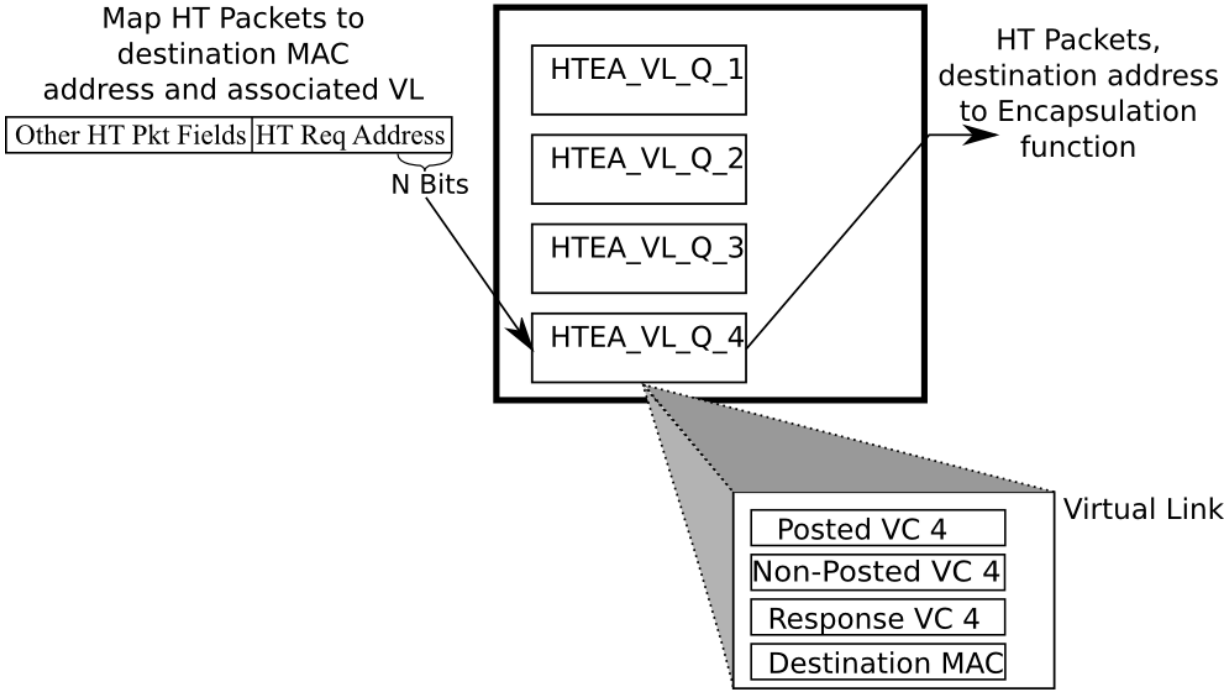
# Backup Slides

# HToE Specification – Tag Remapping

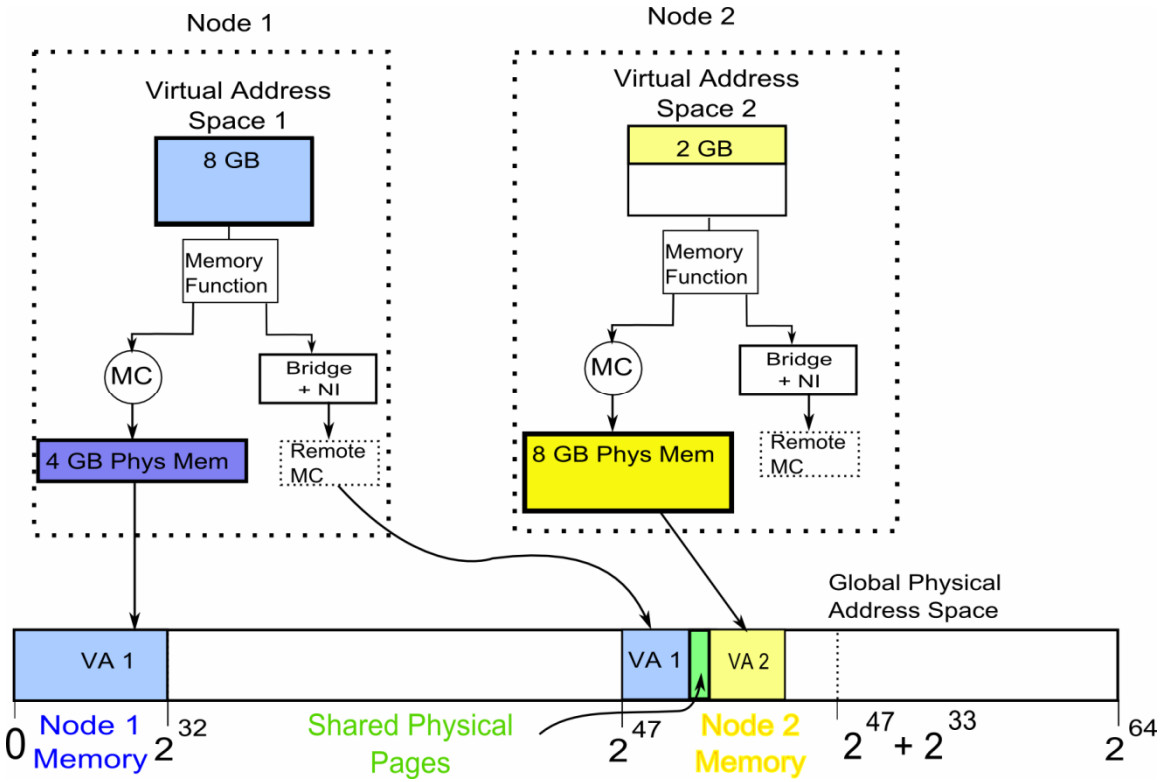


# HToE Specification - Mapping

- Virtual Link – couples 3 HT virtual channels with MAC address to create source/destination pair
  - Allows for mapping HT credits and physical buffers to specific connection

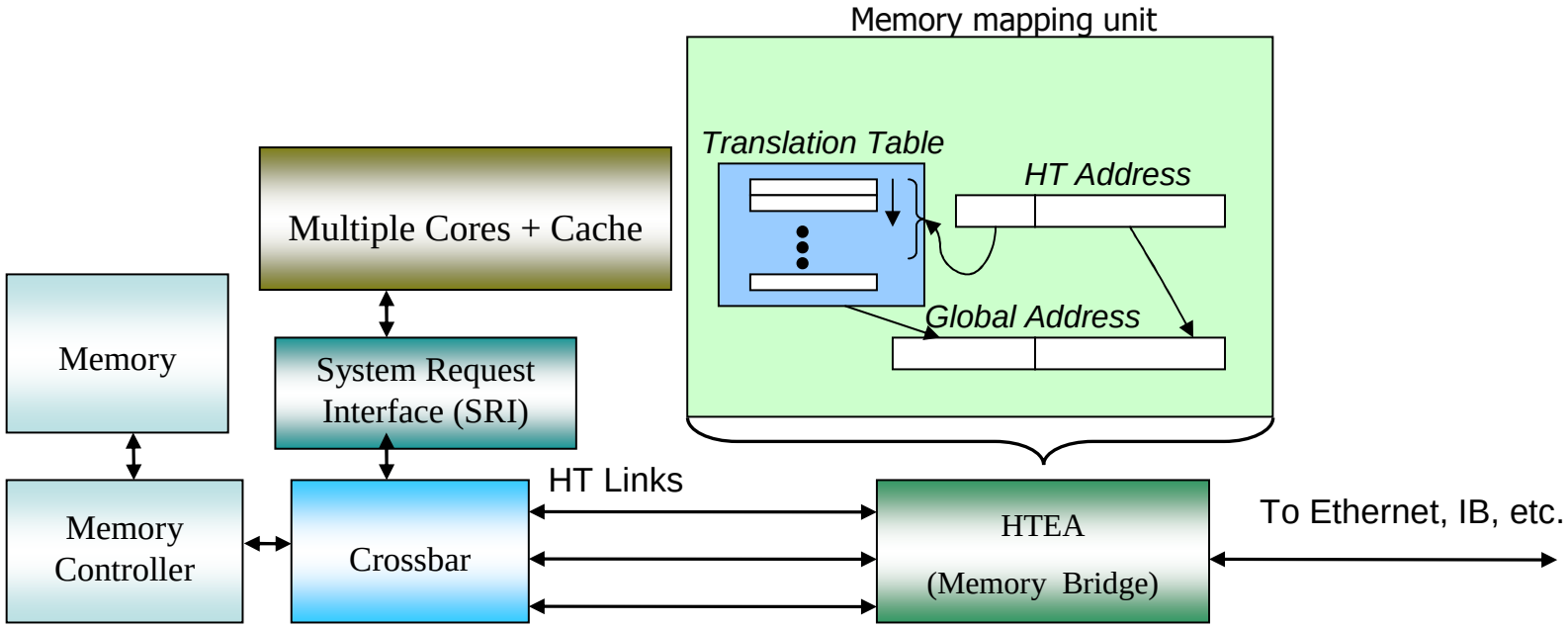


# DPGAS System View




- Portion of the virtual address space mapped to remote physical memory
- Protection issues handled by virtual memory system
  - Bridge mapping handled and coordinated by OS
- "Dynamic" updates allow for flexibility in sharing

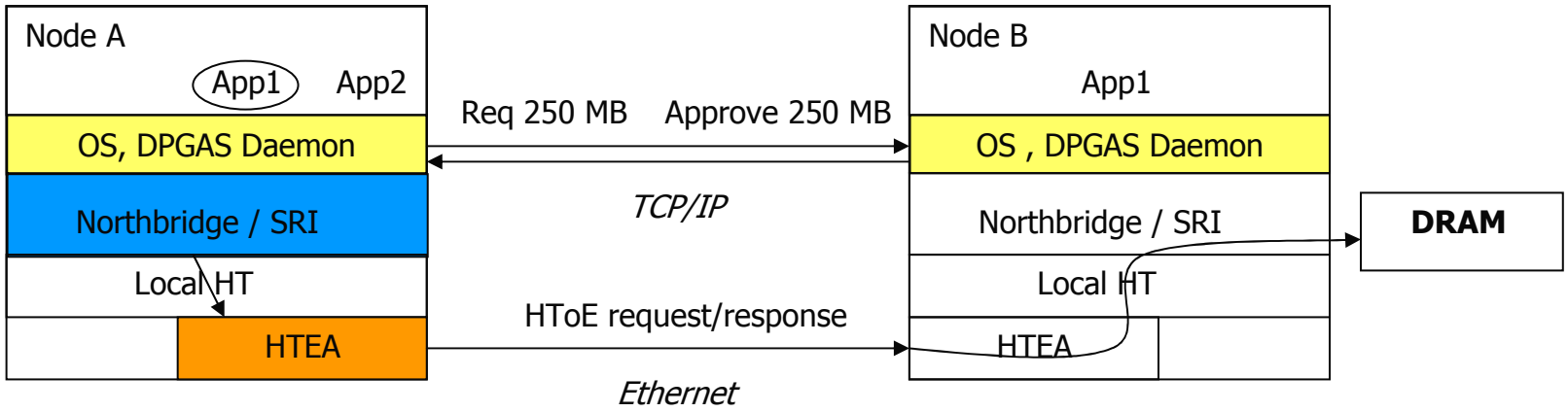
# Architectural Support – Reference Path[1]




- Address translated into a node address and remote local memory address
- Low latency memory bridge: encapsulation takes 24 – 72 ns in current FPGA implementation
  - Referred to as HyperTransport Ethernet Adapter (HTEA)
- Bridge ➔ 1300-1500 FPGA slices (Virtex 4 FX140)

1) *J.Young, et al., A HyperTransport-enabled global memory model for improved memory efficiency, WHTRA '09*

# Memory Allocation with DPGAS – Control Path




- Node A hits memory threshold (pg faults or % of physical memory)
- Node A requests to "spill" to Node B via OS daemon
- Node B approves and agrees to "receive" remote accesses from Node A
  - OS or hypervisor updates available memory (possibly with libnuma hints)
  - System Request Interface is updated to direct requests to HTEA
  - HToE mapping table is updated on Node A
  - If memory is to be unshared, Node B OS updates its available physical memory
- Node A can make remote accesses to Node B's memory via the HTEA