Commodity Converged Fabrics for Global Address Spaces in Accelerator Clouds

Jeffrey Young, Sudhakar Yalamanchili

School of Electrical and Computer Engineering, Georgia Institute of Technology

Motivation

- Current data warehouse applications process, on average, anywhere from 1 to 50 TB of data
  - Expected to grow at a rate of up to 25% per year [1]
- Accelerators like GPUs can be used to accelerate queries for data warehousing applications
  - Co-processing with GPUs provides 2-27x speedup [2]
- However, GPU processing is limited by the size of on-board memory, 4 – 8 GB
  - Data movement is difficult
    - Requires expensive interconnects or limited performance software layers (TCP/IP)

[1] Independent Oracle Users Group. A New Dimension to Data Warehousing: 2011 IOUG Data Warehousing Survey.
[2] B. He, et. al, "Relational query coprocessing on graphics processors," ACM TODS, 2009

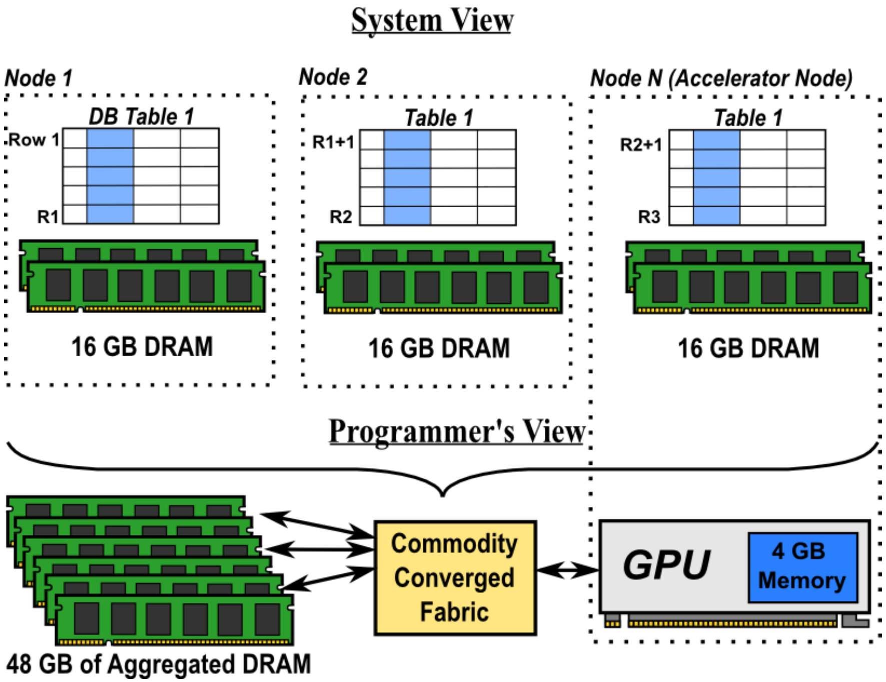
Problem Statement

- **Problem:** How can we cheaply and efficiently move data to GPUs used for data warehousing acceleration?

- **Proposed Solution:** Use commodity, converged fabrics and Global Address Spaces (GAS)
  - Commodity interconnect – widely available interconnect such as Ethernet, InfiniBand, PCIe, HT, QPI
  - Converged fabric – combination of two or more on-chip (HT, QPI, PCIe) and off-chip interconnects (Ethernet, InfiniBand)
  - Global Address Space – physical memory is available for use by remote nodes using either hardware or software support for address translation

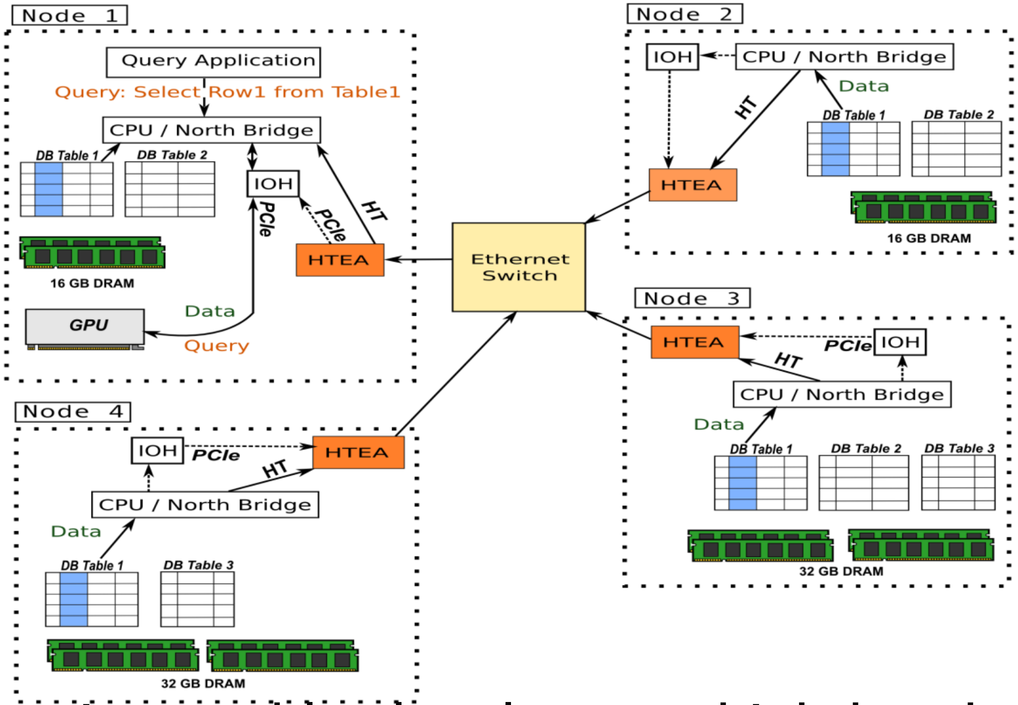
HyperTransport over Ethernet (HToE) and GAS

- HT used due to open-source nature; Ethernet is prevalent in data centers
  - Specification details basic requirements
  - This work provides first proposed implementation of specification
- HyperTransport packets are encapsulated in larger Ethernet frames
  - Packet ordering is preserved using credits
- GAS model uses put/get operations that then support applications such as in-core databases
  - Related work: MEMSCALE [3]



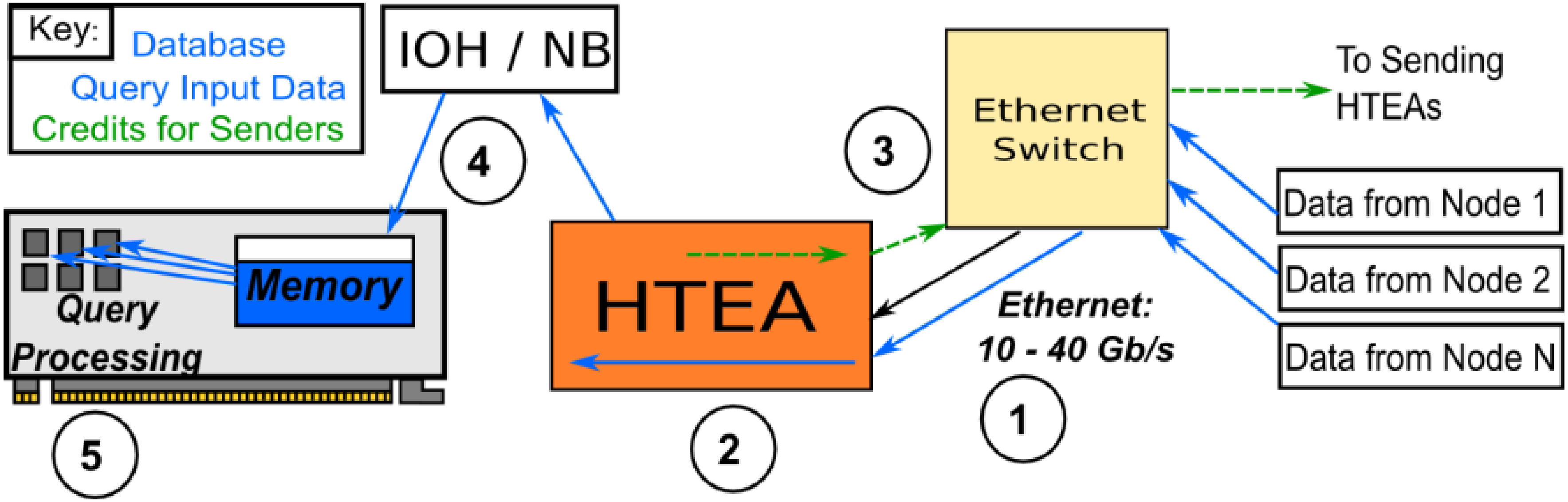
[3] Montaner, H. et al. *MEMSCALE™: A Scalable Environment for Databases.* HPCC, 2011

Commodity Fabrics for Memory Aggregation - HToE System Model



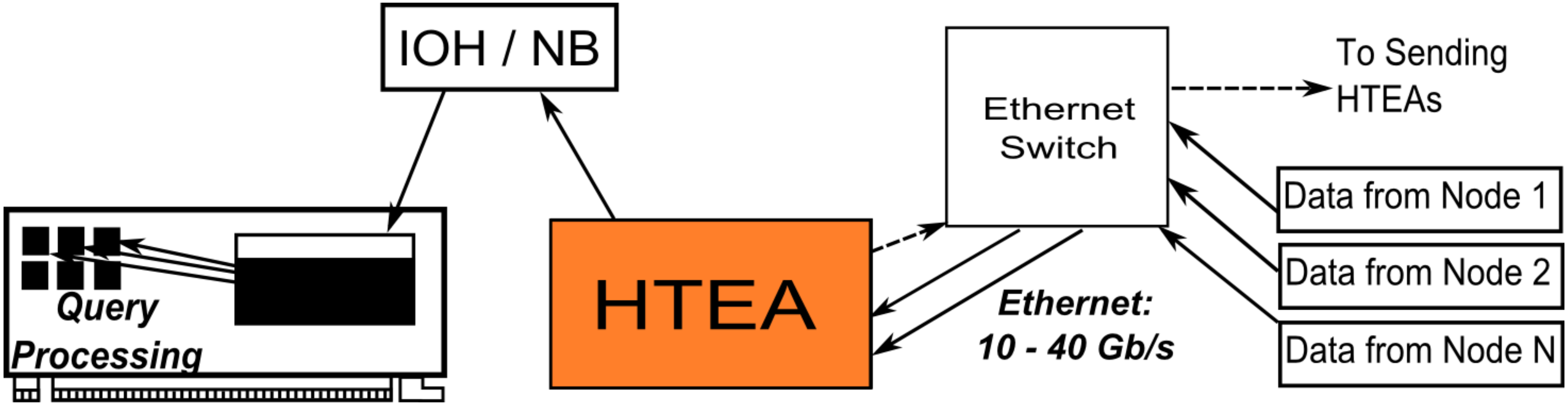
- HToE specification enables low-latency, high-bandwidth Ethernet systems that can be used to create GAS
  - System model stripes in-core database across nodes and allows simultaneous data transfer to an “accelerator” node.
  - Latency: ~1.5 us from adapter to adapter
  - Bandwidth: Up to 24 Gbps for optimized “receiving” adapter
  - ~2.3 ms to transfer 1 GB of data and perform TCP-H query on GPU

Creating a Data Warehousing Accelerator Cloud



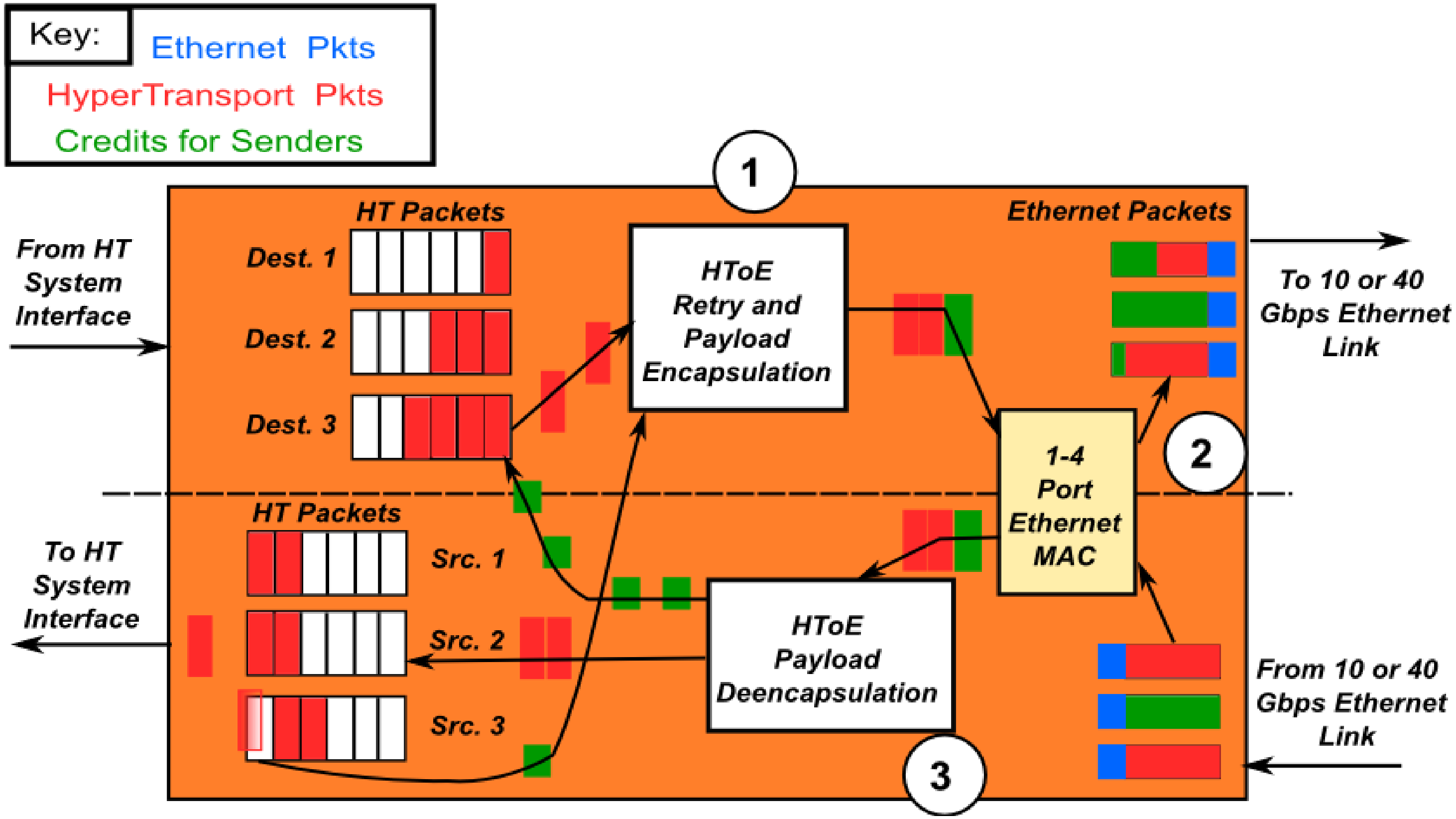
- An accelerator cloud may only contain a few high-end GPUs due to power, cost restraints
- Bottlenecks for multiple-sender, single-receiver system model include, network, receiving adapter, and PCIe transfer BW

Creating a Data Warehousing Accelerator Cloud



- Focus on the design of a HyperTransport Ethernet Adapter (HTEA) that supports HToE specification

Bottlenecks in HToE Adapter Design



- Credits are needed to encapsulate outgoing packets
- We can try to speed up processing of outgoing packets and/or return of credits
- Ethernet MAC can simultaneously process Mod(N) packets

Experimental Setup

- Simulator based infrastructure uses network simulator, NS-3, with synthetic traces based on TCP-H queries
- 1, 3, or 7 nodes contain a striped database table that is 1 GB in size (~16 million HT packets)
  - 1 accelerator (GPU) node receives data from remote nodes via HTEA
- Realistic timing statistics gathered from:
  - Ventoux prototype FPGA from EXTOLL project – runs at 300 MHz with critical path latency of < 1 µs. [4]
  - Ethernet timing gathered from FPGA experiments by Emilio Billi Engineering [5]
  - PCIe transfer time and GPU computation time for queries from Georgia Tech's Red Fox project [6]

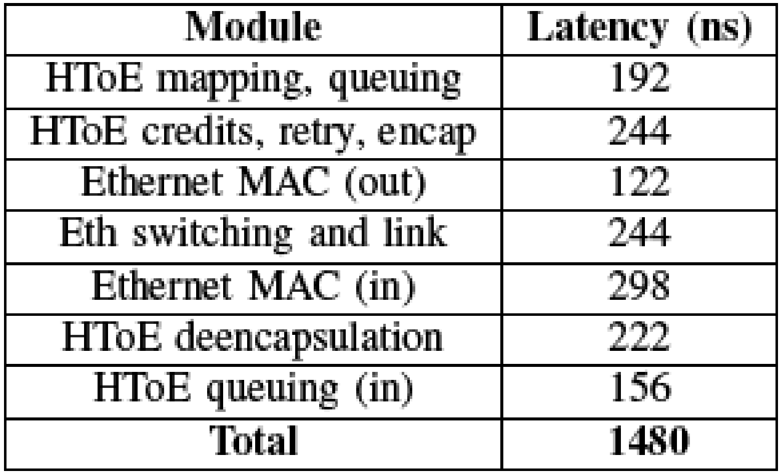
[4] Fröning, H. *EXTOLL Introduction*. HPC User Forum, HLRS and SARA Meeting, October 2010

[5] Xilinx FPGA simulation with 10 Gbps Ethernet MAC - obtained from tests run by Emilio Billi of EBE. 2012. http://www.emiliobilli.com/.

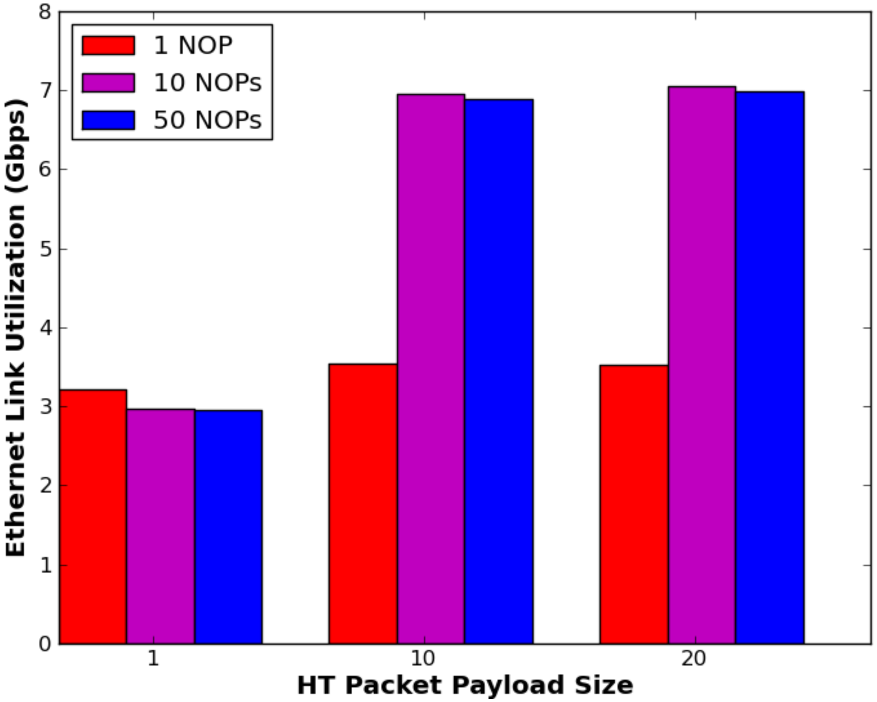
[6] Wu, H. et al., *Optimizing Data Warehousing Applications for GPUs Using Kernel Fusion/Fission*, in IPDPS-PLC, 2012.

2 Node Results – Average Link Bandwidth and Number of Credits

1 HT NOP = up to 3 HT credits

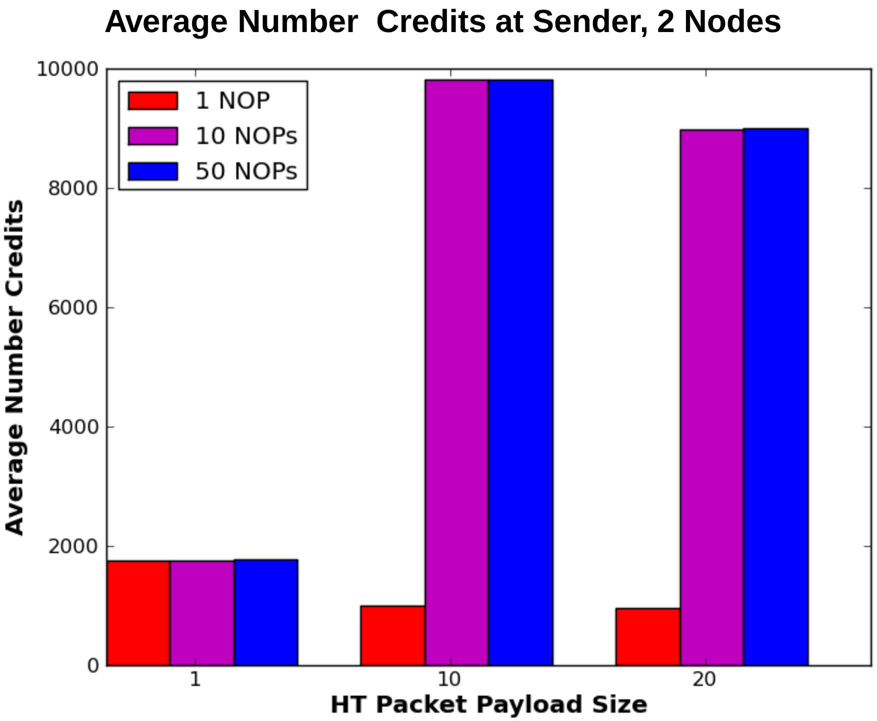


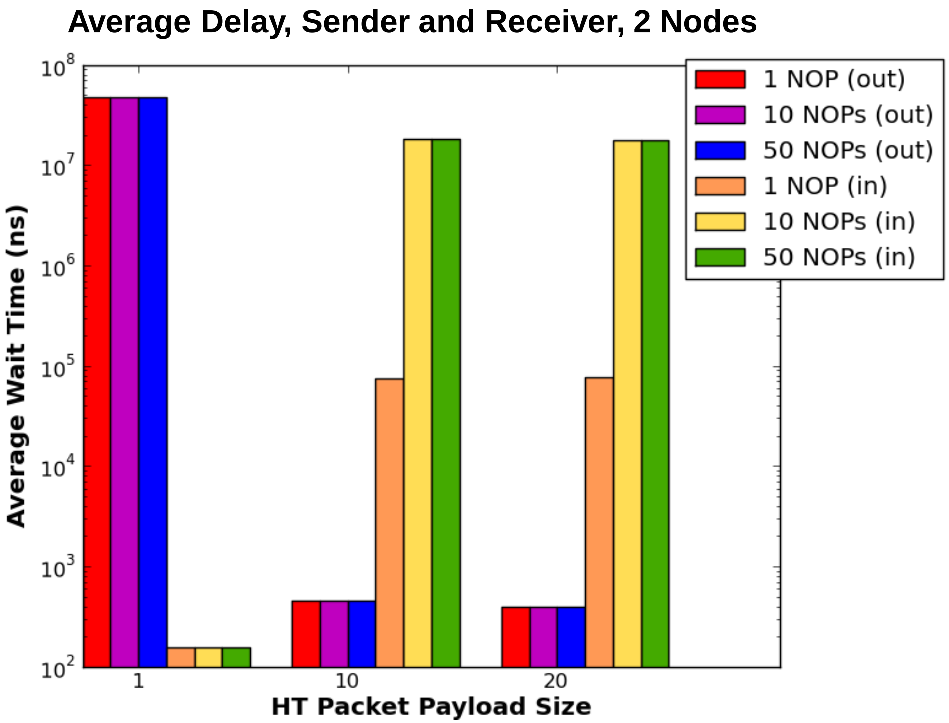
Module	Latency (ns)
HToE mapping, queuing	192
HToE credits, retry, encap	244
Ethernet MAC (out)	122
Eth switching and link	244
Ethernet MAC (in)	298
HToE deencapsulation	222
HToE queuing (in)	156
Total	**1480**



- Deencapsulation latency higher than encapsulation latency
  - 1.5 µs adapter to adapter latency
- Using small HT packet and credit payload sizes impairs the link utilization

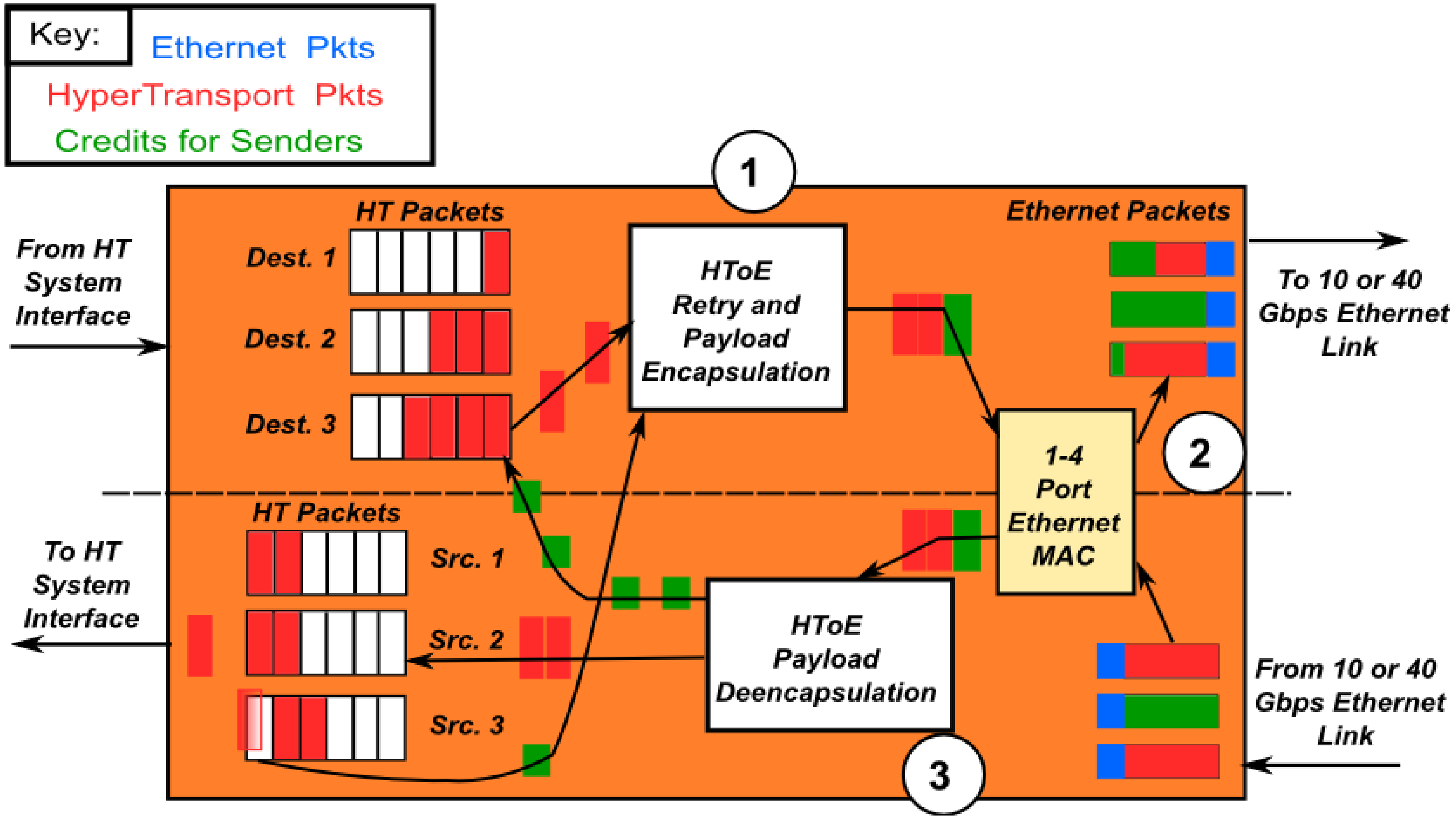
2 Node - Credit Availability and Encapsulation Delays





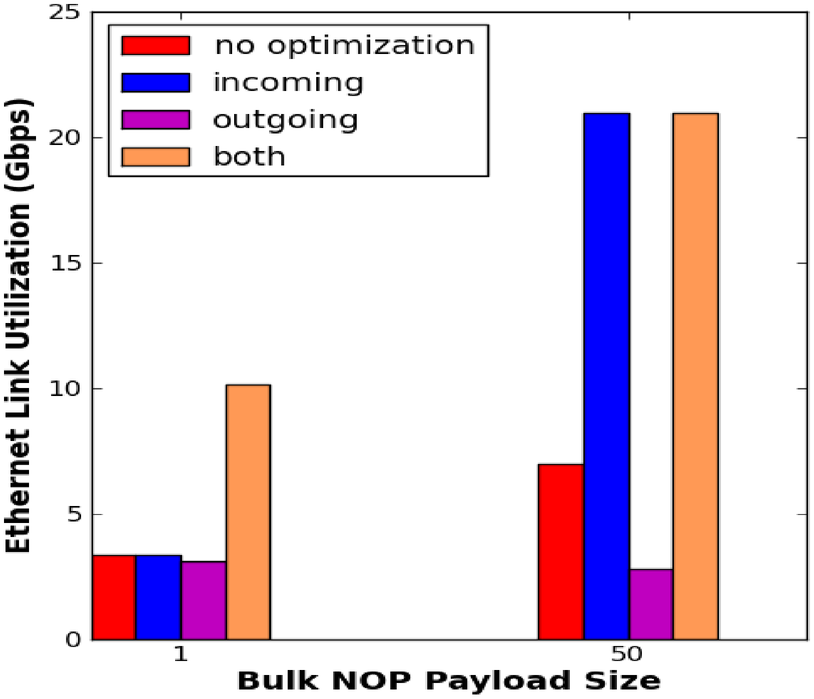
- Large packet and NOP payloads result in more credits being available to senders
  - Less head-of-line blocking and better link utilization
- However, larger payloads have higher delay at receiver because of complexity of deencapsulation

HTEA Optimizations (2)

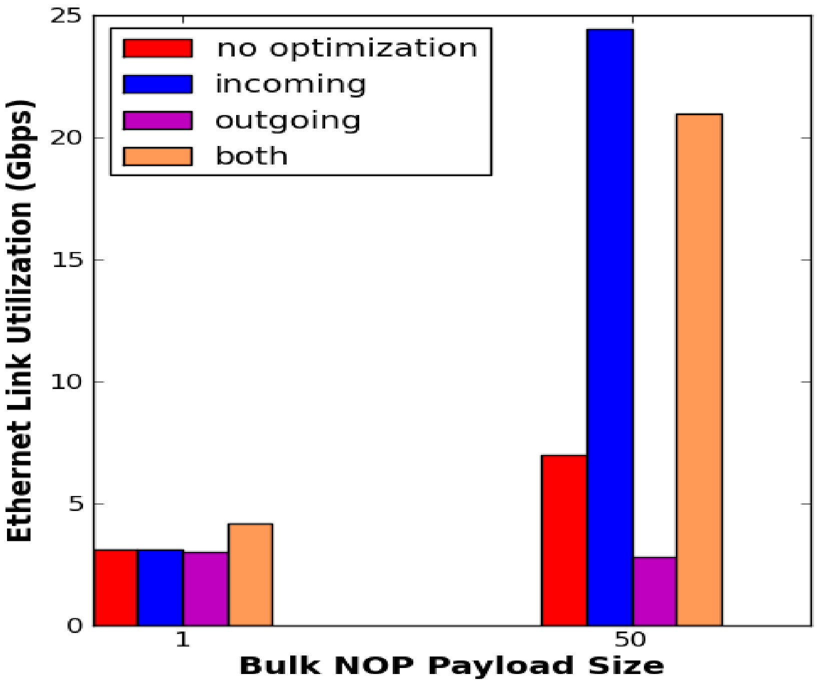


4 and 8 Node Results – HTEA Optimizations

Link Utilization, 4 Nodes

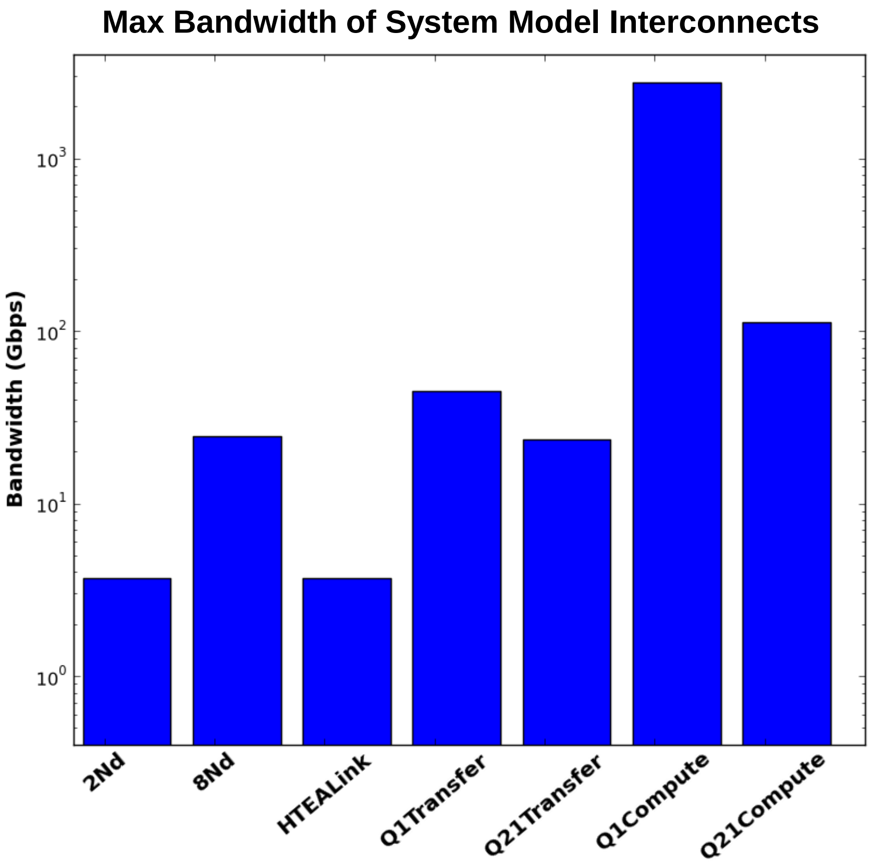


Link Utilization, 8 Nodes



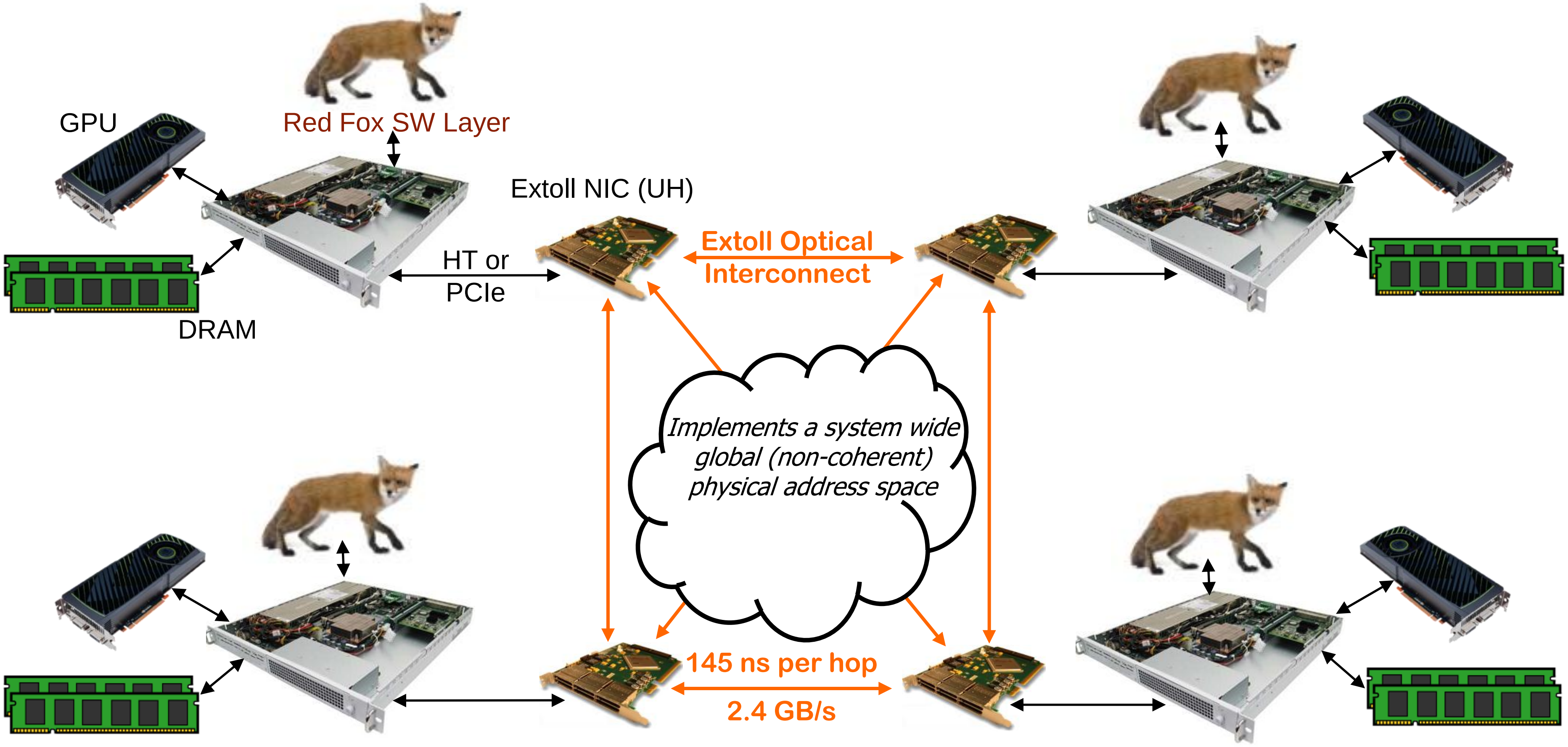
- More parallelization for deencapsulation has a much greater impact than for encapsulation
  - In some cases, (50 NOP, outgoing) sending rate far exceeds return of credits, decreasing overall utilization
- Use of bulk credit payloads enables faster HT packet transfer with wider pipelines

System Model Bottleneck Analysis



- GPU compute bandwidth far outstrips data transfer bandwidth
  - Asynchronous transfer model e.g., cudaDMA) may be useful to keep GPU busy
- HTEA optimizations improve transfer bandwidth but may need further improvements

Future Work - Oncilla


GPU
Red Fox SW Layer
Extoll NIC (UH)
Extoll Optical Interconnect
HT or PCIe
DRAM
Implements a system wide global (non-coherent) physical address space
145 ns per hop
2.4 GB/s

- Low-latency, commodity hardware (EXTOLL) allows for efficient memory and GPU aggregation and Red Fox SW layer supports DB queries on remote nodes
- Collaboration with University of Heidelberg (UH), Polytechnic University of Valencia, AIC Inc., LogicBlox Inc.

Conclusions

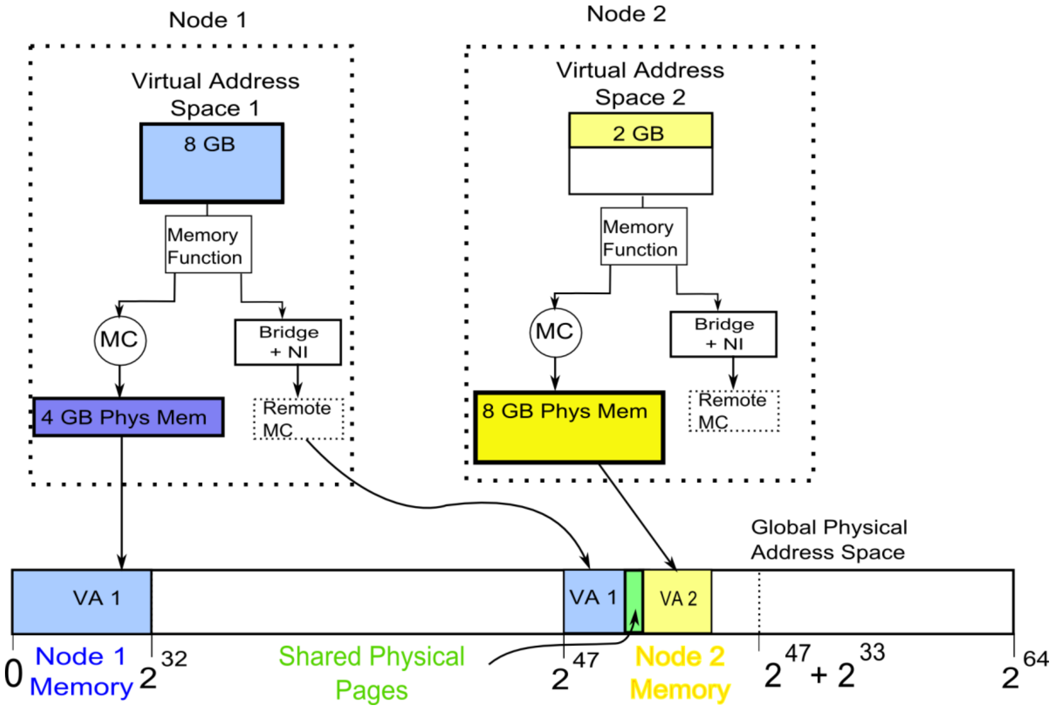
- Commodity, converged fabrics can be used for “accelerator clouds”
  - Demonstrated an implementation of HToE specification
- However, optimization of adapters may be necessary depending on hardware and data distribution
- Computational BW of today’s GPUs far outpace bandwidth of data transfer
  - New interconnect technologies and software models will be needed to address this “GPU memory wall”

Questions?

- Special thanks to HT Consortium members: Emilio Billi, Mario Cavalli, Brian Holden, Paul Miranda
- More related work at http://gpuocelot.gatech.edu/projects/compiler-projects/
- Contact: jyoung9@gatech.edu, sudha@ece.gatech.edu

Backup Slides

Global Address Space Overview



- 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