Towards Terabit/s Systems: Performance Evaluation of Multi-Rail Systems

Venkatram Vishwanath[1], Takashi Shimizu[2], Kazuaki Obana[2], Makato Takigawa[2], Jason Leigh[1]

[1]Electronic Visualization Laboratory, University of Illinois at Chicago, USA [2]NTT Network Innovation Laboratories, Yokosuka, Japan

venkat@evl.uic.edu, t-shimizu@ieee.org

www.evl.uic.edu/cavern/anr

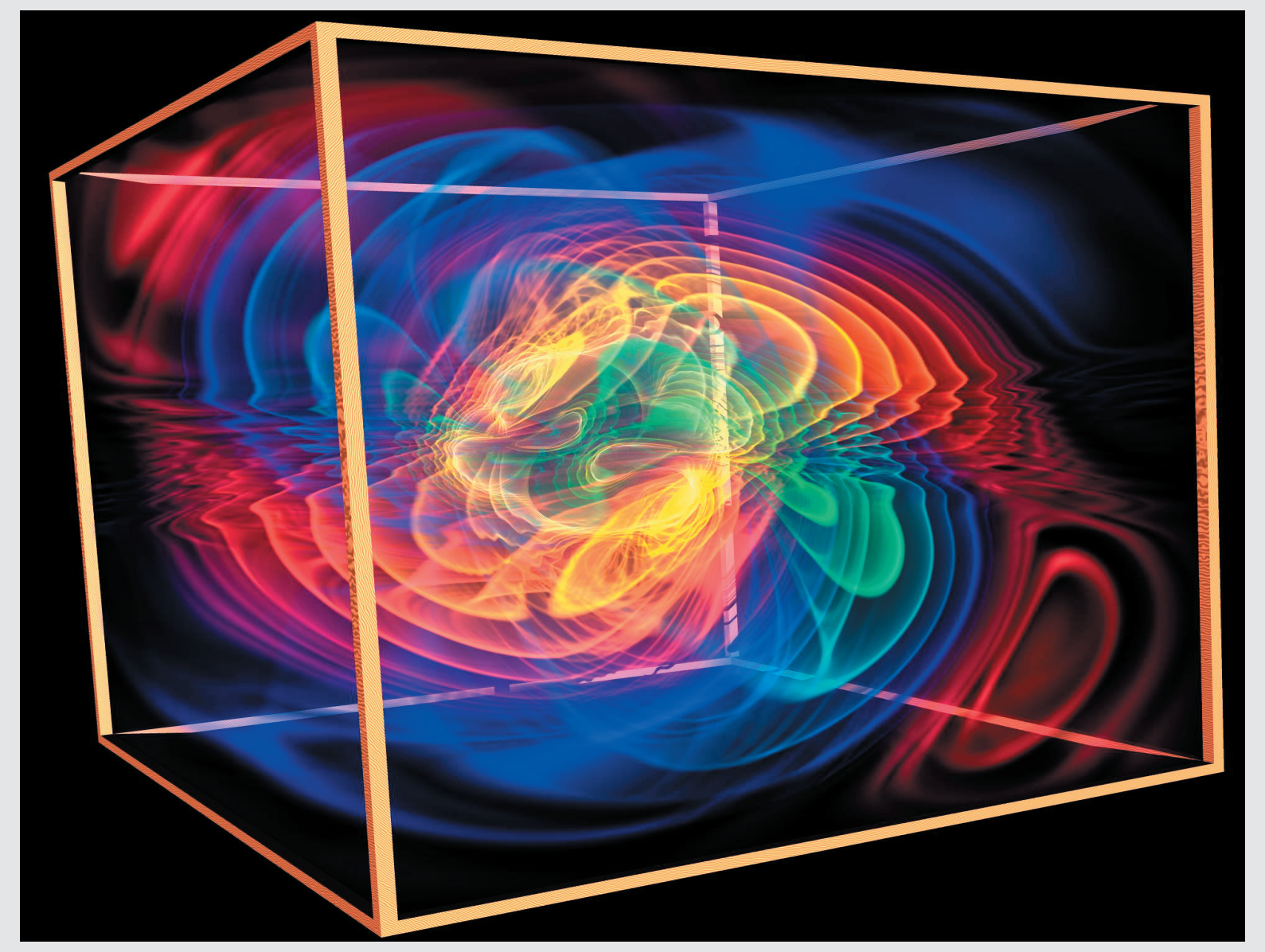
Project Goals: Investigate end-system issues and design novel middleware for applications to scale to Terabit/s

Abstract

We present a novel multi-rail approach that is necessary for future E-Science applications to effectively exploit Terabit/s networks. This multi-rail approach consists of creating parallel "rails" through every aspect of an end-system: from processing on the multiple and many cores, generation of multiple application data flows, and streaming over multiple-lanes, multi-wavelength NICs connected via a parallel interconnect. We present the evaluation of end-systems parameters that impact the efficiency of multi-rail systems such as interrupt, memory, thread, and core affinities. These evaluations were tested on the ability of individual cluster nodes to achieve TCP and UDP throughput at 10Gbps and 20Gbps rates. We analyze the additive effects of the parameters - a key property for achieving scalable performance towards Terabits/s. Thread and Interrupt affinity together was found to have an additive effect and plays a critical role in achieving maximum throughput.

Terabit/s Applications

Terabit/s is a characteristic of many applications including a number of Petascale applications. These applications generate myriad data flows. In the very near future, these applications will run on multi-cores systems with multi-lane system interconnects and multiple network interfaces.



Gravitational field of a grazing collision of black holes

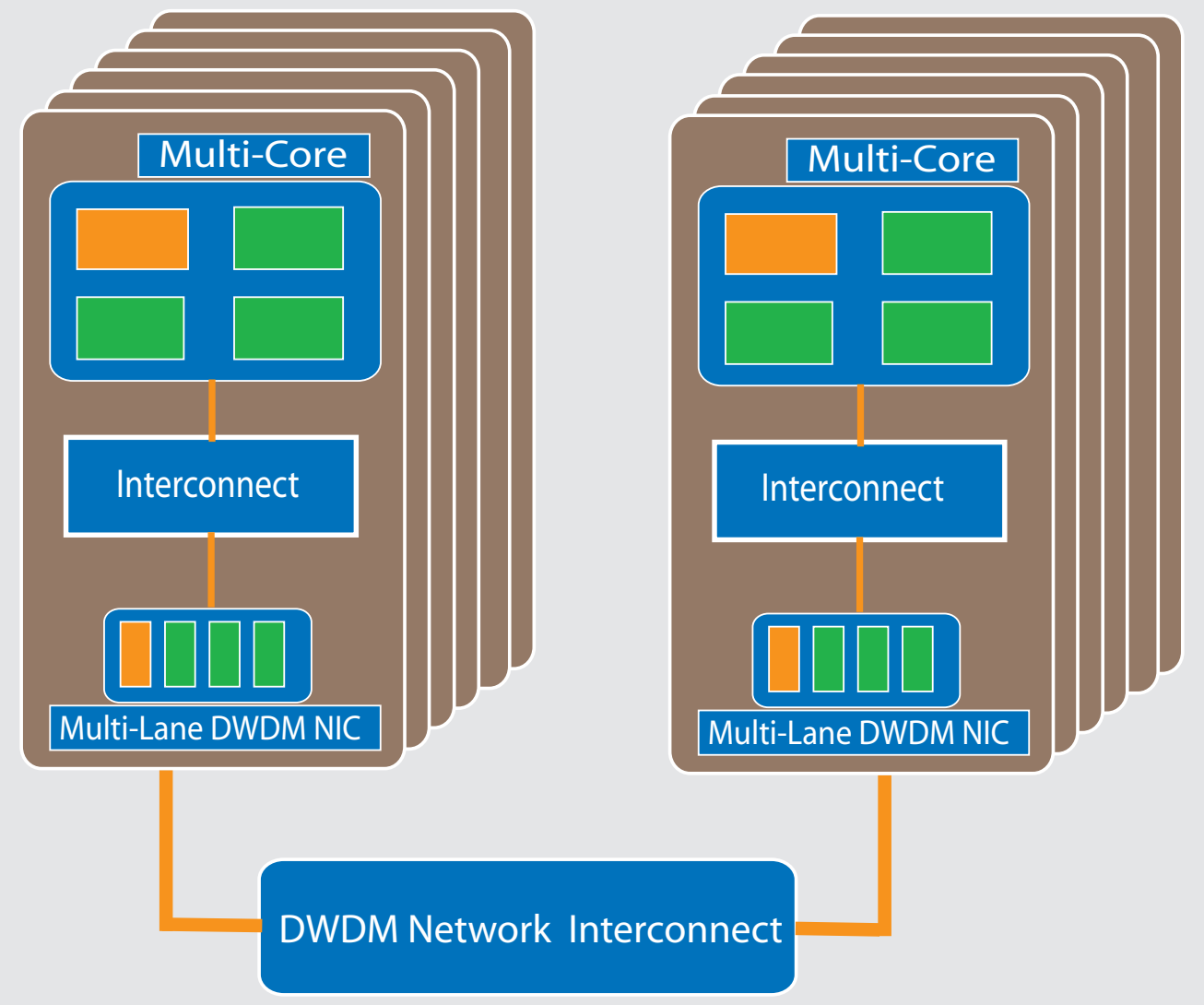
Numerical simulation: Ed Seidel, Numerical Relativity Group, Max Planck Institute for Gravitational Physics (Albert Einstein Institute) and Center for Computation and Technology at Louisiana State University (CCT/LSU) using the CACTUS framework; Visualization: Werner Benger, CCT, Zuse Institue Berlin and AEI using the Light++ raytracer

Cyber discovery enabled by Scalable Adaptive Graphics Environment (SAGE)

Image courtesy: Electronic Visualization Laboratory, University of Illinois at Chicago

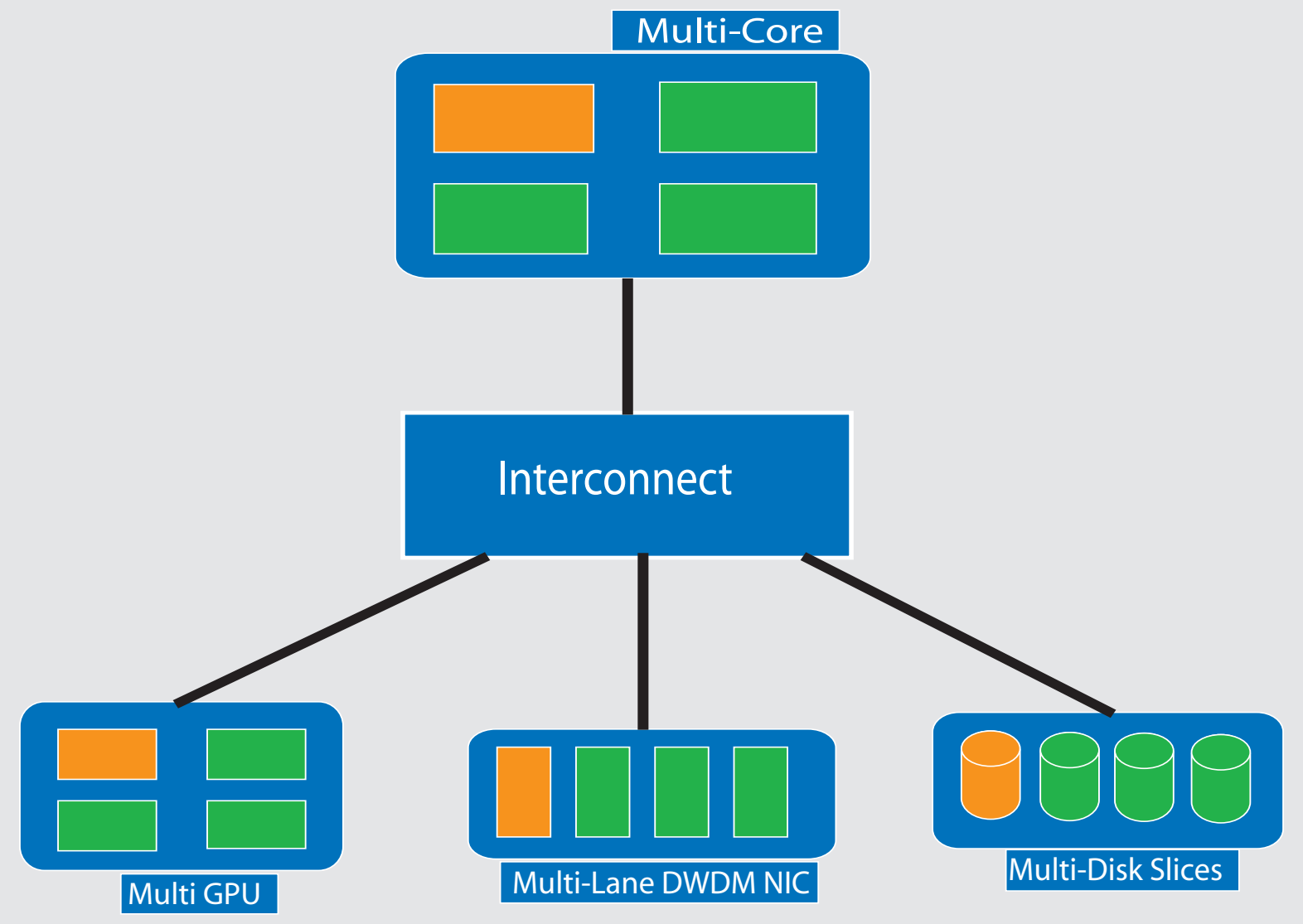
Architectural Trends

Processor	Multi-Core Teraflops Chips
System Interconnects	Multi-Lane System Interconnects including HyperTransport and CSI
Network Cards	DWDM-based Multi-Lane NICs at 40Gbps and 100Gbps
Network Interconnect	DWDM-based Network Interconnects
Applications	Multi-Threaded with Multiple Data Flows



Multi-Rail Systems

We investigate system characteristics that would play a critical role in such future Terabit/s systems.The goal was to study the additive properties of the system parameters and their combinations as we scale systems towards Terabit/s by increasing the number of network cards and processors i.e. multiple-rails.



In a multi-rail system, each rail consists of a processor core connected to a lane on a NIC via a dedicated interconnect.

The multi-rail approach can be expanded to exploit parallelism in other subsystems including graphics and disk.

We envision future Teraflops chip architectures to include a multi-rail subsystem, as part of its system architecture, for scaling performance towards Terabits/s

The multi-rail concept can be easily realized with currently available processor architecture.

Let **T** be the achievable performance of a single rail, In an **N** rail system, the expected performance would be:

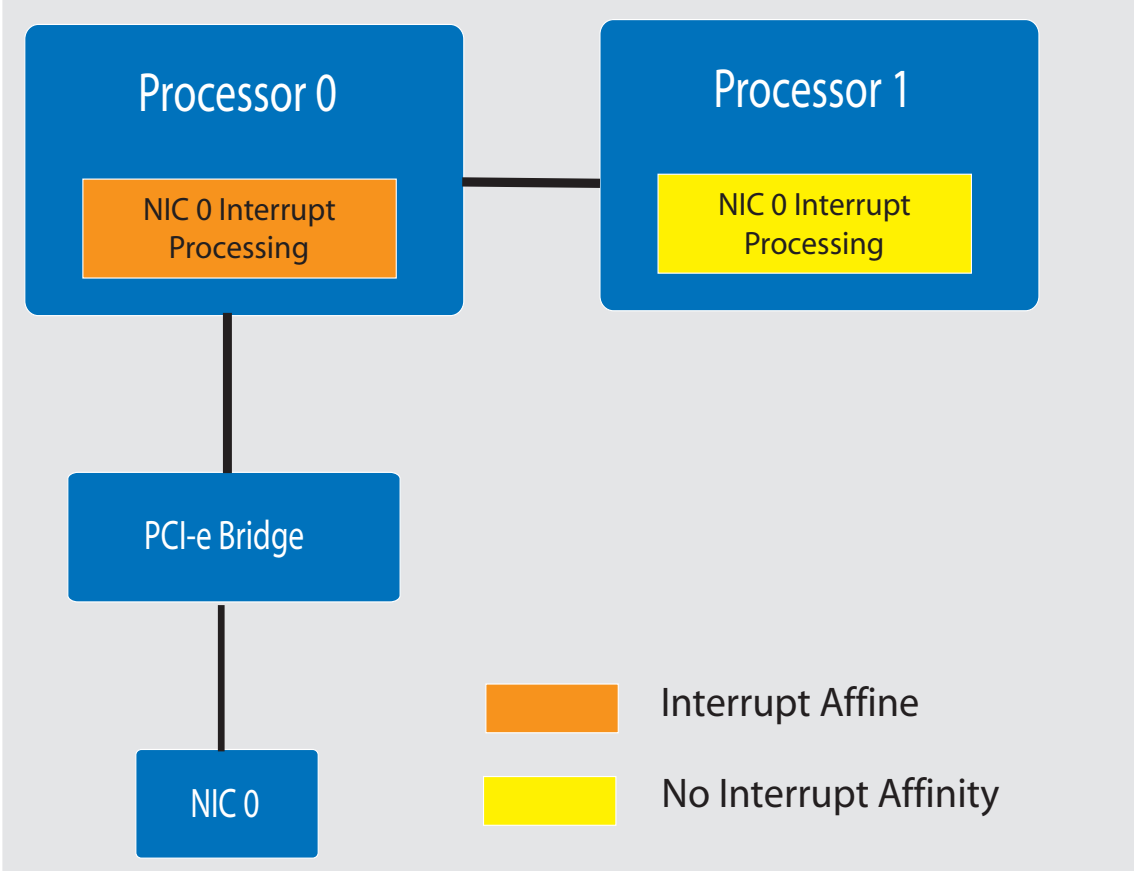
$$N \times T \times \partial$$

where ∂ is the parallel efficiency. In an ideal parallel system, $\partial \to 1$, and this system exhibits additive performance.

Approach

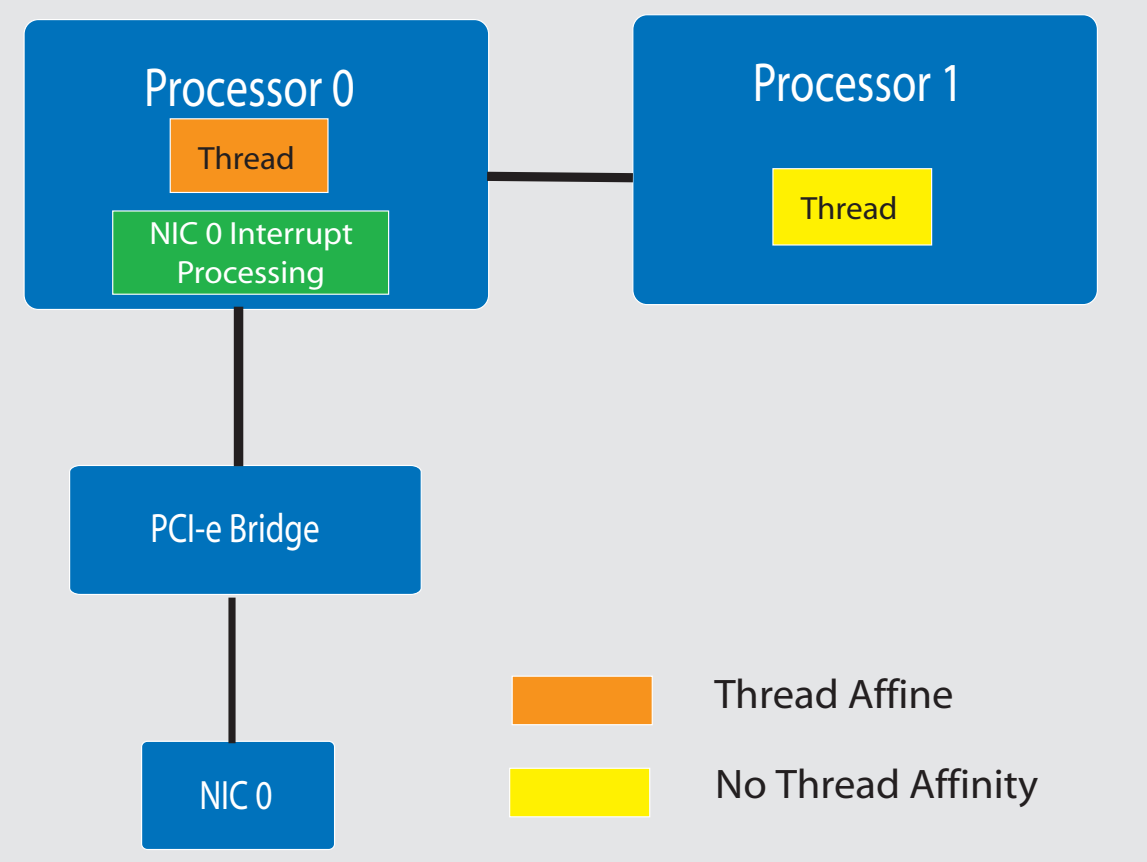
We investigate the system characteristics that will play a critical role in such future Terabit/s systems. Specifically, we investigate system effects such as **Thread Affinity, Interrupt Affinity, Memory Affinity, and Core Affinity on the throughput of network-intensive applications**. Identifying the effects of these system parameters, towards achieving additive performance, will aid in scaling performance towards Terabit/s.

Interrupt Affinity (IA)



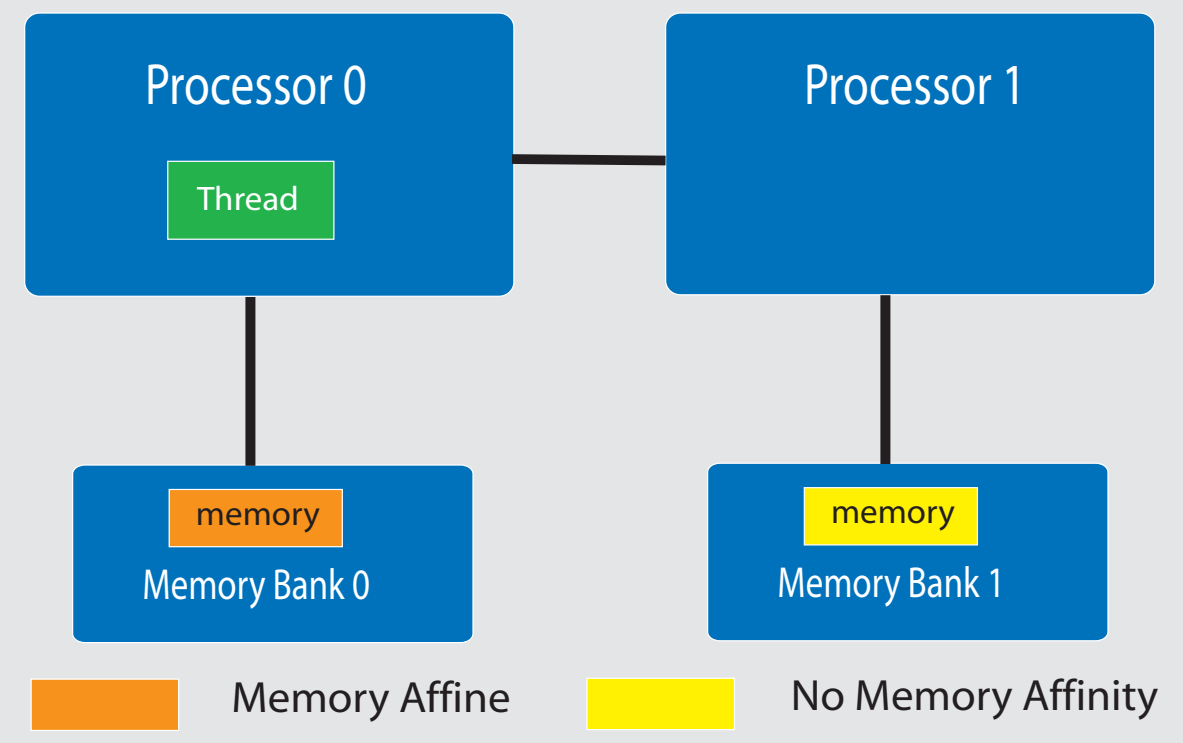
We consider a system to be Interrupt Affine if the Interrupt processing is done by the processor to which the interrupt is physically bound.

Thread Affinity (TA)



We consider a system to be Thread Affine if the network application thread is bound to the processor where the Interrupt processing of the network traffic occurs.

Memory Affinity (MA)



We refer to a system as Memory affine system as one where the memory used by an application thread is allocated on the memory bank with the lowest access latency. **SMP** systems are **Memory Affine.**

Performance Evaluation of Multi-Rail Systems

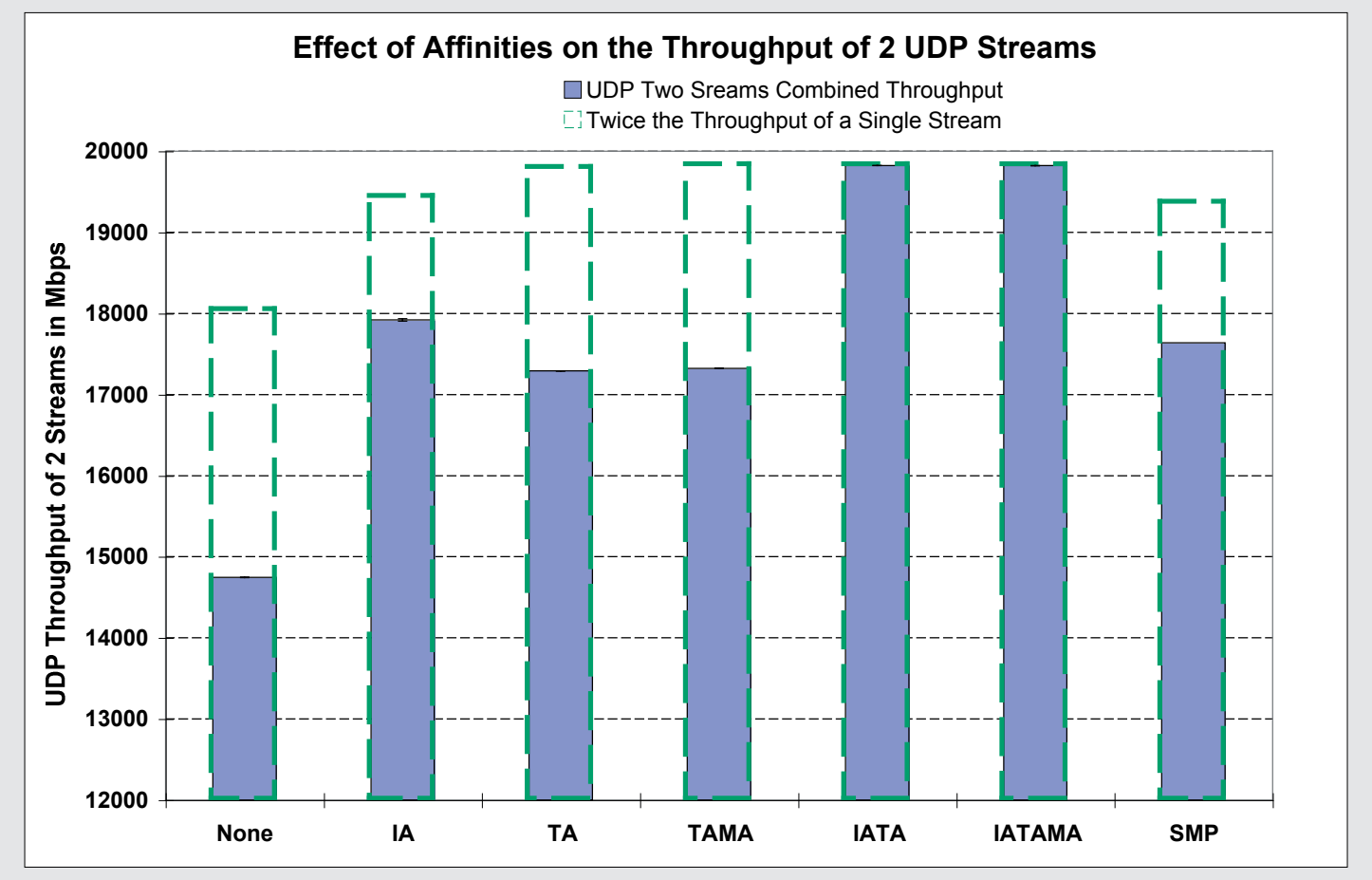
Exhaustive evaluation of the effects of Thread Affinity, Interrupt Affinity, Memory Affinity and Core Affinity and their additive properties.

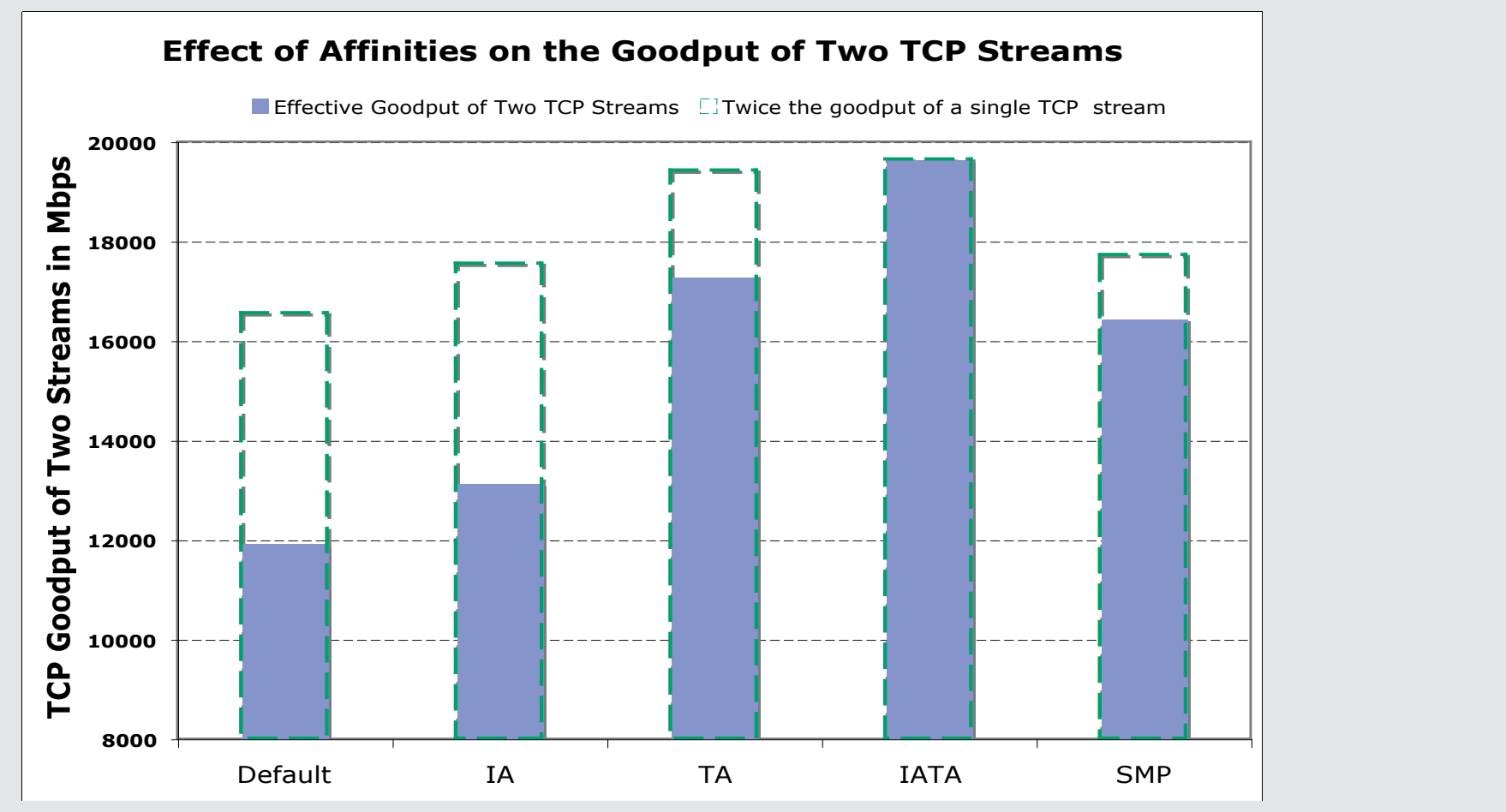
Thread and Interrupt Affinity, together, play a key role in achieving Additive performance

Additive performance achievable on current NUMA based multi-rail systems

The experimental testbed consisted of:

- Two Dual-Core, Dual-Processor AMD 2.6 GHz Opteron TYAN 2895 systems with 4GB RAM and two PCI-express 16x slots. The two machines were connected back-to-back with two Myrinet 10G NICs.
- Two Dual-Core, Dual-Processor 3.0Ghz Intel Xeon IBM x3500 systems with 4GB RAM and two PCI-express 16x slots. The two machines were connected back-to-back with two 10G Myrinet NICs.
- The Linux kernel version used was 2.6.18 with MSI enabled. "nuttcp" 5.5.4 was used as the Network Intensive workload. The MTU used for the experiments was 9000bytes.





Other Significant Results

- Memory Affinity plays a critical role on the achievable goodput for large data transfers.
- Thread Affinitty plays a key role in reducing CPU utilization and improving throughput.
- In case of TCP, distributing the interrupt processing and thread processing on the cores improves the achivable throughput. In case of UDP, core affinity helps in reducing CPU utlization.
- Ethernet Channel bonding does not scale at 10Gbps rates and higher.

Ongoing Work

Multi-Rail Framework for HPC Applications

API to enable applications run on multi-rail system

Leverages existing open-source libraries including pthreads and libnuma

Given an application profile, a multi-rail daemon can optimize the performance on a given node

Can be integrated with job schedulers to improve the performance of applications on multi-rail systems

Celeritas - Multi-Rail Aware Transport Protocol

A multi-rail aware, end-system aware, application-level, compositional transport protocol framework

Support reliable and unreliable data transfer and streaming over Local and Wide-area Networks

Uses Task parallelism to distribute the transport protocol computation, including rate control, over multiple cores

Uses Data parallelism to stripe the data using multiple connections over multiple network interfaces