MVAPICH at Petascale: Experiences in Production on the Stampede System

Dan Stanzione

Executive Director, Texas Advanced Computing Center

August, 2014, Columbus, Ohio

Acknowledgements

- Thanks/kudos to:
  - *Sponsor: National Science Foundation*
    - NSF Grant #OCI-1134872 Stampede Award, "Enabling, Enhancing, and Extending Petascale Computing for Science and Engineering"
    - NSF Grant #OCI-0926574 - "Topology-Aware MPI Collectives and Scheduling"
  - *Many, many Dell, Intel, and Mellanox engineers*
  - *Professor D.K. Panda at OSU (MVAPICH2)*
  - *All my colleagues at TACC who saw way too much of each other during the last year*

Outline

- A long history of collaboration
  - TACC, InfiniBand, and MVAPICH
- Evolving together with Stampede
  - System overview
  - MPI for heterogeneous computing
- And we keep on charging ahead
  - Chameleon

Our history

- We have used MVAPICH on a bunch of systems through a lot of years.
  - Consistently the highest performance MPI
  - Consistently the most reliable
  - Consistently the fastest starting at scale.

  - None of this is by accident.

In each large system, MVAPICH partnership has been key

- Contrary to popular belief, none of this stuff comes up and works the first time from the vendors.
  - Tuning for new interconnect technologies at new scales.
  - Tuning to take advantage of the growth of intra-node communication.
  - Tuning to take advantage of heterogeneous nodes.

Brief Clustering History at TACC

- Like many sites, TACC was deploying small clusters in early 2000 timeframe
- First “large” cluster was Lonestar2 in 2003
  - 300 compute nodes originally
  - Myrinet interconnect
  - debuted at #26 on Top500
- In 2005, we built another small research cluster: ***Wrangler*** (128 compute hosts)
  - 24 hosts had both Myrinet and early IB
  - single 24-port Topspin switch
  - used to evaluate price/performance of commodity Linux Cluster hardware

Early InfiniBand Evaluation

- Try to think back to the 2004/2005 timeframe……
  - only 296 systems on the Top500 list were clusters
  - multiple IB vendors and stacks
  - “multi-core” meant dual-socket
  - we evaluated a variety of stacks across the two interconnects
  - *our first exposure to MVAPICH* (0.9.2 via Topspin and 0.9.5 via Mellanox)

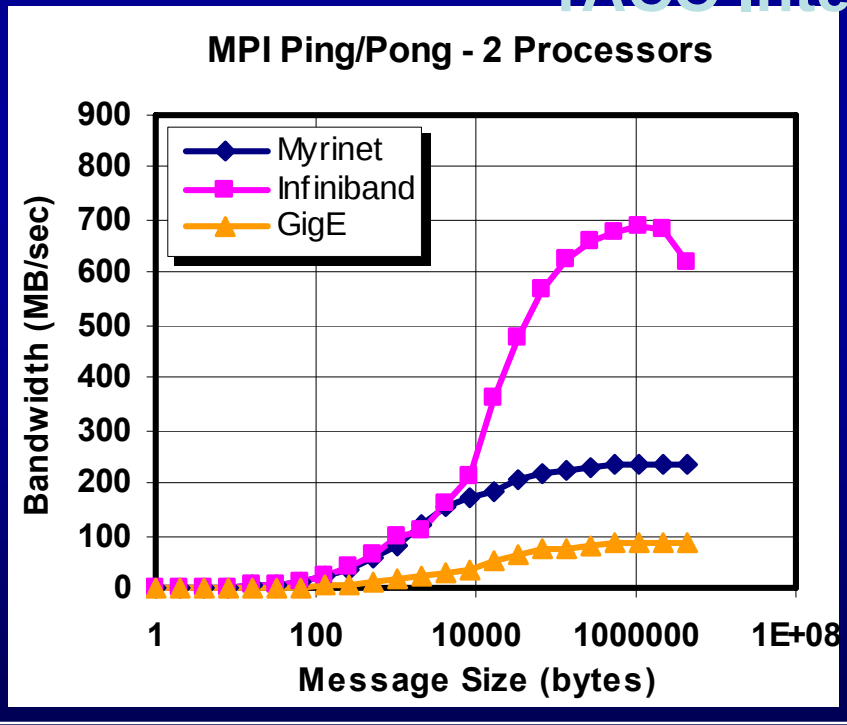
Example MPI Latency Measurements, circa 2005

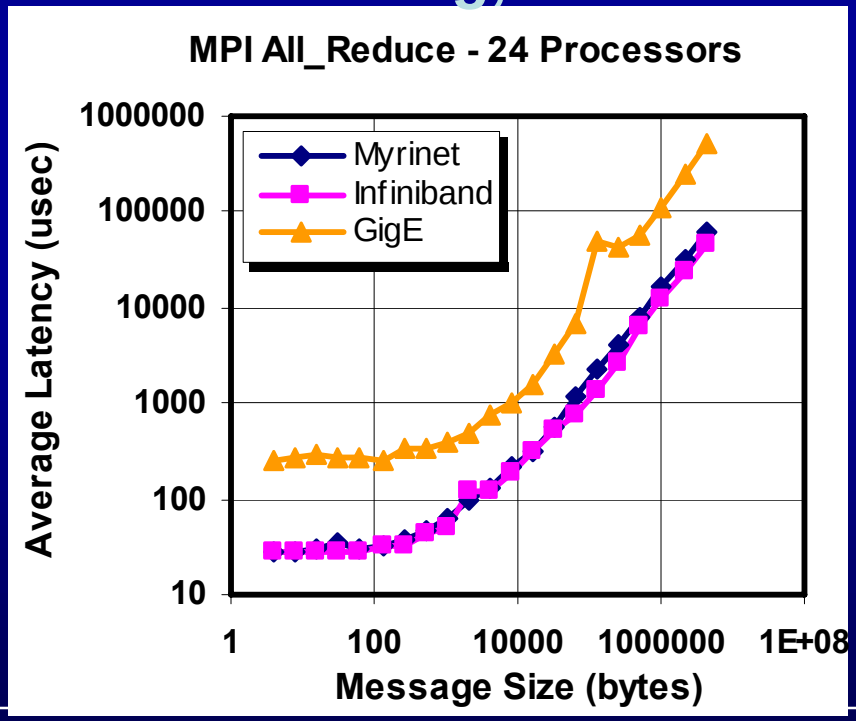
Myrinet	MPICH-GM	7.05 µs
	MPICH-MX	3.25 us
	LAM-GM	7.77 µs
	VMI-GM	8.42 µs
Infiniband	IB-Gold	4.68 µs
	IB-Topspin	5.24 µs
	LAM-IB	14.96 µs
GigE	MPICH	35.06 µs
	LAM-TCP	32.63 µs
	VMI-TCP	34.29 µs

Early InfiniBand Evaluation

- In addition to latency considerations, we were also attracted to BW performance and influence on applications

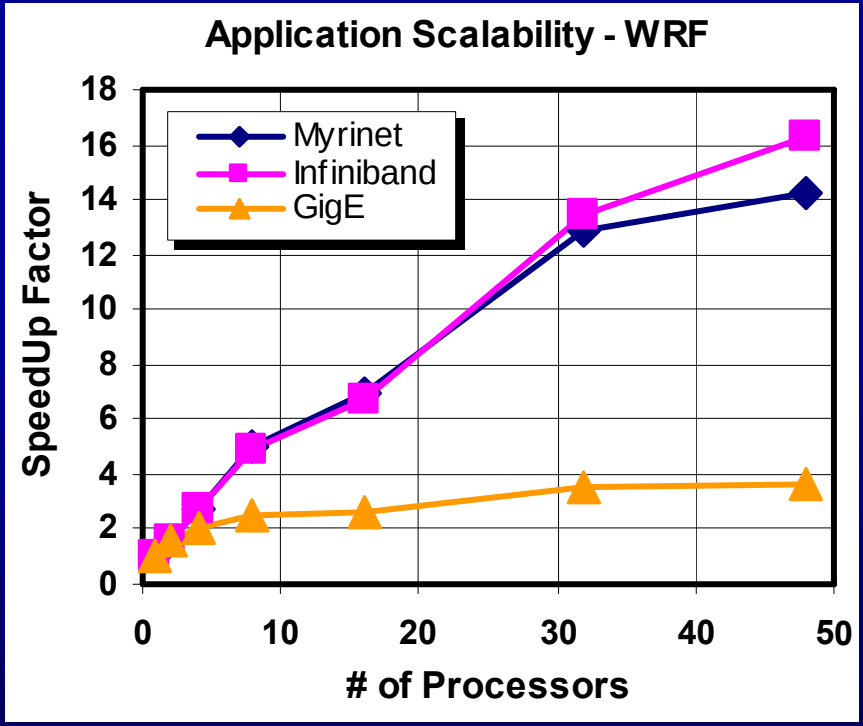
TACC Internal Benchmarking, circa 2005

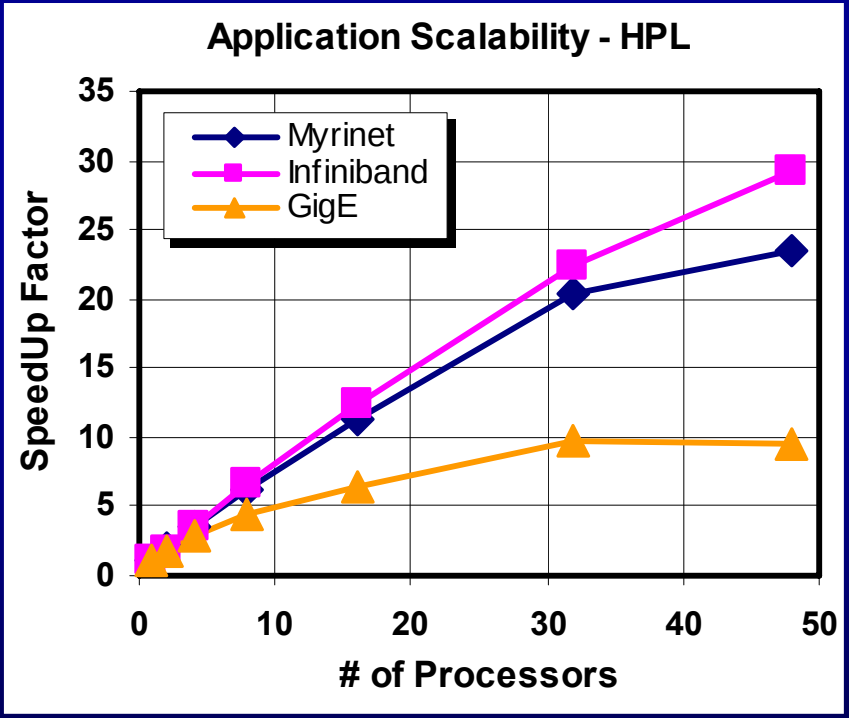




Early InfiniBand Evaluation

TACC Internal Benchmarking, circa 2005





Brief Clustering History at TACC

- Based on these evaluations and others within the community, our next big cluster was IB based

- Lonestar3 entered production in 2006:
  - OFED 1.0 was released in June 2006 (and we ran it!)
  - First production Lustre file system (also using IB)
  - MVAPICH was the primary MPI stack
  - workhorse system for local and national researchers, expanded in 2007

  - ☛ Debuted at #12 on Top500

Brief Clustering History at TACC

- These clustering successes ultimately led to our next big deployment in 2008, the first NSF “Track 2” system, ***Ranger:***
  - $30M system acquisition
  - 3,936 Sun four-socket blades
  - 15,744 AMD “Barcelona” processors
  - All IB all the time (SDR) - no ethernet
    - Full non-blocking 7-stage Clos fabric
    - ~4100 endpoint hosts
    - >1350 MT47396 switches
    - *challenges encountered at this scale led to more interactions and collaborations with OSU team*

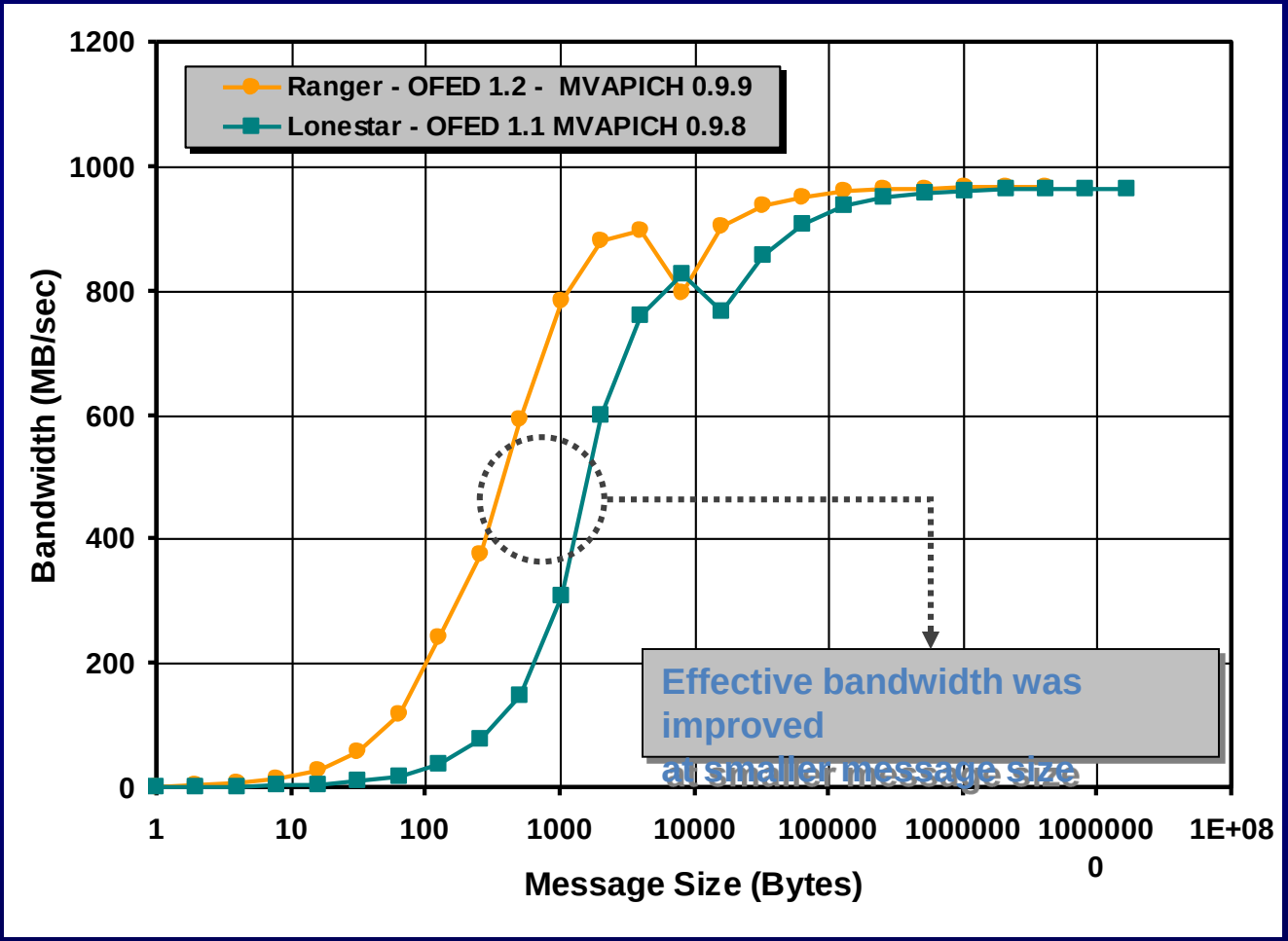
☛ Debuted at #4 on Top500

TACC

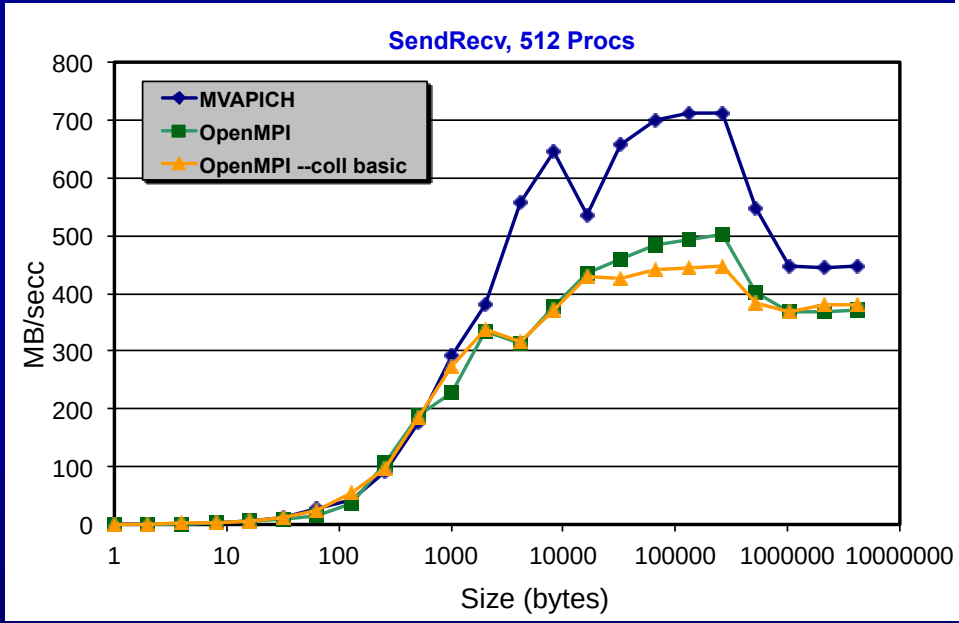
Ranger: MVAPICH Enhancements

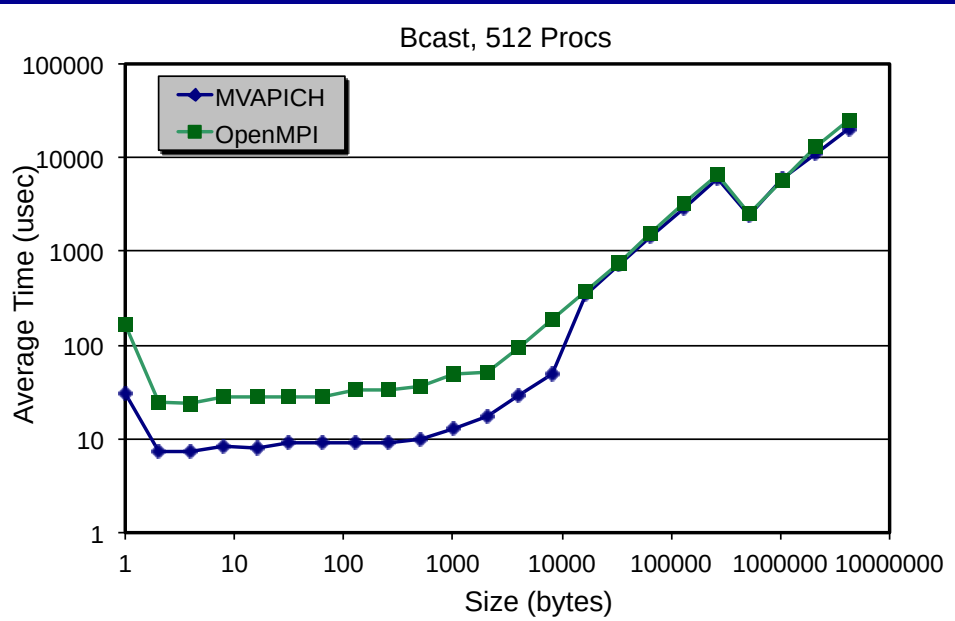
- The challenges encountered at this scale led to more direct interactions with the OSU team
- Direct interaction started after meetings at Cluster 2007.
  - original discussion focused on "`mpirun_rsh`" for which enhancements were released in MVAPICH 1.0
  - subsequent interactions focused on ConnectX collective performance, job startup scalability, SGE integration, shared-memory optimizations, etc.
  - DK and his team relentlessly worked to improve MPI performance and resolve issues at scale; helped to make Ranger a very productive resource with MVAPICH as the default stack for thousands of system users

Ranger: MVAPICH Enhancements



Ranger MPI Comparisons





Stampede

- Funded by NSF as an XSEDE resource through HPC System Acquisition solicitation 11-511 (award #OCI-1134872) in September 2011.
- Stampede was constructed in 2012, and went into production on January 7th, 2013.
- In July, 18 months later, Stampede delivered it's **One Billionth** service unit to a user, and is closing in on it's 4 millionth user job (eclipsing all 5 years of Ranger).

Stampede - High Level Overview

- Base Cluster (Dell/Intel/Mellanox):
  - Intel Sandy Bridge processors
  - Dell dual-socket nodes w/32GB RAM (2GB/core)
  - 6,400 nodes
  - 56 Gb/s Mellanox FDR InfiniBand interconnect
  - More than 100,000 cores, 2.2 PF peak performance
- Co-Processors:
  - Intel Xeon Phi "MIC" Many Integrated Core processors
  - Special release of "Knight's Corner" (61 co
  - All MIC cards are on site at TACC
  - 7+ PF peak performance
- *Max Total Concurrency:*
  - *exceeds 500,000 cores*
  - *1.8M threads*

- ***Entered production operations on January 7, 2013***

Additional Integrated Subsystems

- Stampede includes 16 1TB Sandy Bridge shared memory nodes with dual GPUs
- 128 of the compute nodes are also equipped with NVIDIA Kepler K20 GPUs (and MICS for performance bake-offs)
- 16 login, data mover and management servers (batch, subnet manager, provisioning, etc)
- Software included for high throughput computing, remote visualization
- Storage subsystem driven by Dell storage nodes:
  - Aggregate Bandwidth greater than 150GB/s
  - More than 14PB of capacity
  - Similar partitioning of disk space into multiple Lustre filesystems as previous TACC systems ($HOME, $WORK and $SCRATCH)

Innovative Component

- One of the goals of the NSF solicitation was to *"introduce a major new innovative capability component to science and engineering research communities"*

- We proposed the Intel Xeon Phi coprocessor (many integrated core or MIC)
  - one first generation Phi installed per host during initial deployment
  - (Now 480 nodes with two per node)
  - confirmed injection of 1600 future generation MICs in 2015 (5+ PF)

Xeon Phi on Stampede

- In a nutshell, the Phi is a lot of cores (61 per chip), with much longer vector units, more threading, and slower clocks.
- We believe this architecture foreshadows pretty much all future processors – Power makes this inevitable
  - If your code can't vectorize, and you can't scale to a lot of threads, you have a problem
  - GPUs differ (substantially) in the details, but not in the broad strokes.
- Stampede is a chance to start moving codes this direction

Team

- TACC
- Vendors: Intel, Dell, Mellanox
- Academic Partners:
  - Clemson (Campus Bridging)
  - Colorado(Phi support)
  - Cornell (Online Training)
  - Indiana (Campus bridging/Data support)
  - Ohio State (MVAPICH support)
  - UTEP , UT ICES (Technology Insertion)

Timelines

- Solicitation: December 7th, 2010
- Proposal Submitted: March 7th, 2011
- Award: ~September 30th, 2011
- Datacenter completion/Start of system Delivery: August 1st, 2012
- Operations start: January 7th, 2013
- Operations end: January 7th, 2017

Current Status

- Through 19 months of Production Operation, Stampede by all measures is remarkably successful.
  - Over *1 Billion* Service Units delivered.
  - Over 3.65 million jobs
  - 1,771 distinct projects received allocations
  - 5,085 Individuals have actually run a job (~8,000 accounts).
  - 97% cumulative uptime (target: 96).
  - 5,006 User Tickets Resolved
  - 2,000+ users attended training last year.
- Formal requests from the community from XSEDE run 500% available hours

Progress on Xeon Phi

- Stampede was the *first* system to make the Phi available at large scale.
  - Most users saw Phi for the very first time when they got access to Stampede.
- In just 18 months, more than 800 users have run a job in the *production* queues using Phi (hundreds more in development queues).
  - Over 100,000 production Phi jobs.
  - Spanning 300 different projects

Programming Models for MIC

- MIC adopts familiar X86-like instruction set (with 61 cores,244 threads in our case)
- Supports full or partial offloads (offload everything or directive-driven offload)
- Predominant parallel programming model(s) with **MPI**:
  - Fortran: OpenMP, MKL
  - C: OpenMP/Pthreads, MKL, Cilk
  - C++: OpenMP/Pthreads, MKL, Cilk, TBB
- Has familiar Linux environment
  - you can login into it
  - *you can run "top", debuggers, your native binary, etc*

Quick Reminder on Native Execution

```
login1$  srun -p devel --pty /bin/bash -l
c401-102$ cat hello.c
#include<stdio.h>
int main()
{
  printf("Hook 'em Horns!\n");

#ifdef __MIC__
  printf(" --> Ditto from MIC\n");
#endif
}

c401-102$ icc hello.c
c401-102$ ./a.out
Hook 'em Horns!

c401-102$ icc -mmic hello.c
c401-102$ ./a.out
bash: ./a.out: cannot execute binary file

c401-102$ ssh mic0 ./a.out
Hook 'em Horns!
 --> Ditto from MIC
```

Interactive Hello World

- Interactive programming example
  - Request interactive job (srun)
  - Compile on the compute node
  - Using the Intel compiler toolchain
  - Here, we are building a simple hello world…
- First, compile for SNB and run on the host
  - note the __MIC__ macro can be used to isolate MIC only execution, in this case no extra output is generated on the host
- Next, build again and add "-mmic" to ask the compiler to *cross-compile* a binary for native MIC execution
  - note that when we try to run the resulting binary on the host, it throws an error
  - ssh to the MIC (mic0) and run the executable out of $HOME directory
  - this time, we see extra output from within the guarded__MIC__ macro

Quick Reminder on Offload Execution

```
!dec$ offload target(mic:0) in(a, b, c) in(x) out(y)
!$omp parallel
!$omp single
      call system_clock(i1)
!$omp end single
!$omp do
      do j=1, n
        do i=1, n
          y(i,j) = a * (x(i-1,j-1) + x(i-1,j+1) + x(i+1,j-1) + x(i+1,j+1)) + &
                   b * (x(i-0,j-1) + x(i-0,j+1) + x(i-1,j-0) + x(i+1,j+0)) + &
                   c * x(i,j)
        enddo
        do k=1, 10000
          do i=1, n
            y(i,j) = a * (x(i-1,j-1) + x(i-1,j+1) + x(i+1,j-1) + x(i+1,j+1)) + &
                     b * (x(i-0,j-1) + x(i-0,j+1) + x(i-1,j-0) + x(i+1,j+0)) + &
                     c * x(i,j) + y(i,j)
          enddo
        enddo
      enddo
!$omp single
      call system_clock(i2)
!$omp end single
!$omp end parallel
```

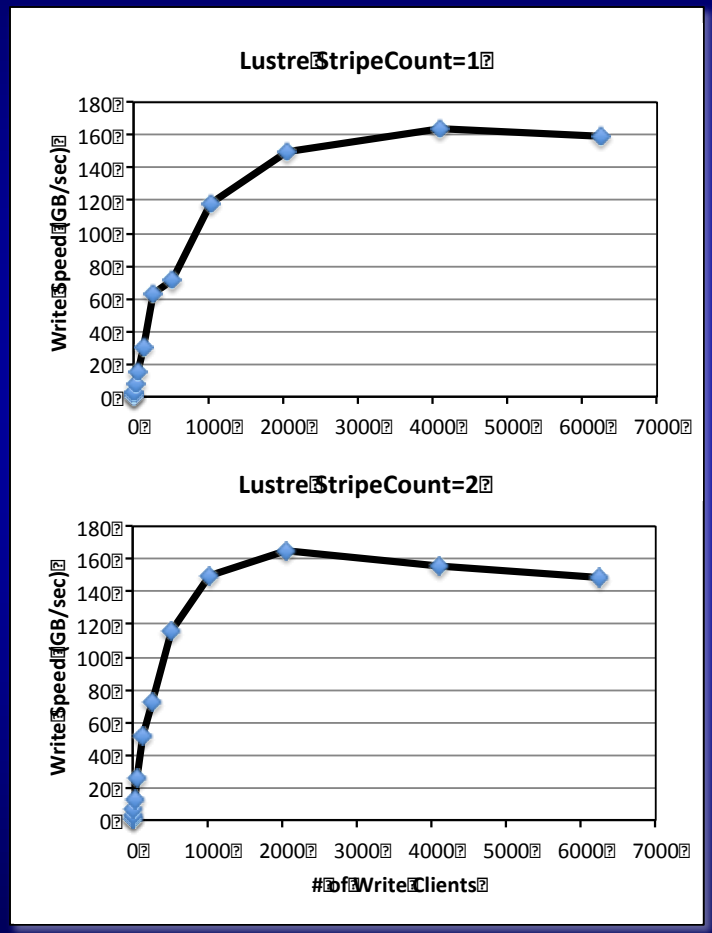
Kernel of stencil code (f90)

File Systems Build-Out: Lustre

- At the Stampede scale, parallel file systems are universally required for all user file systems
  - *Currently running Lustre 2.1.3*
  - *$HOME*
    - 768 TB
    - Permanent user storage; automatically backed up, quota enforced
  - *$WORK*
    - ~2 PB
    - Large allocated storage; not backed up, quota enforced
  - *$SCRATCH*
    - ~11 PB
    - Large temporary storage; not backed up, purged periodically

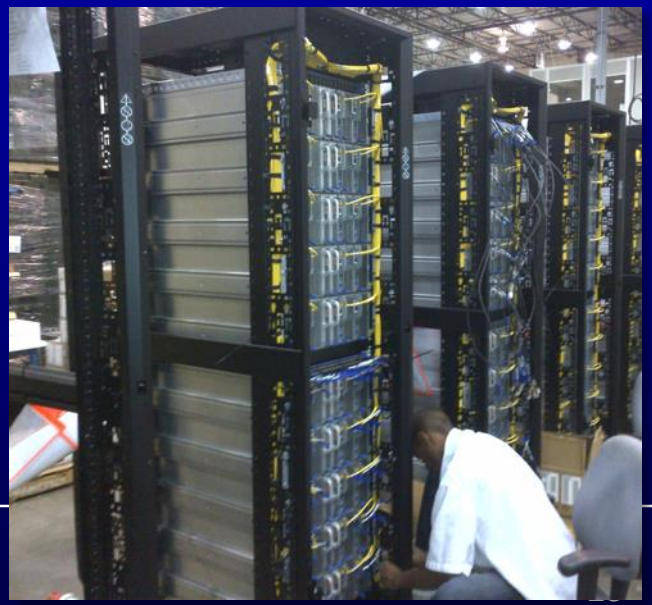
➢ Full System I/O Results ($SCRATCH):

Peak Write = 159 GB/sec
Peak Read = 127.6 GB/sec



InfiniBand Cable Management

- Bringing copious amounts of IB cables into a central location (e.g. 324/648 port switches or larger) requires dedicated cable management:
  - to allow for functional line-card replacement and reasonable air flow
  - to have traceable cable/port mappings
  - to not look horrendous
- Stampede has a large number of cables:
  - 6400 copper cables from computes -> leaf switches
    - Dell responsibility
    - Delivered compute racks fully assembled (including leaf switch IB cabling)
  - > 5100 fiber cables linking leafs to 8 core switches
    - TACC responsibility
    - Designed overhead cable strategy
- TACC had previous experience at large scale with Sun's Magnum switch
  - excellent cable management
  - similar approaches not readily available
  - consequently, we have been designing our own core switch management strategies (Longhorn, Lonestar4, and now Stampede)



Previous Experiences with Sun Data Center Switch (3,456 port switch)

- Visually attractive
- External, side mounted supports with cables coming from under floor
- Homegrown jigs created to organize cables under the floor
- Relatively “easy” to swap line cards

Stampede IB Cable Management

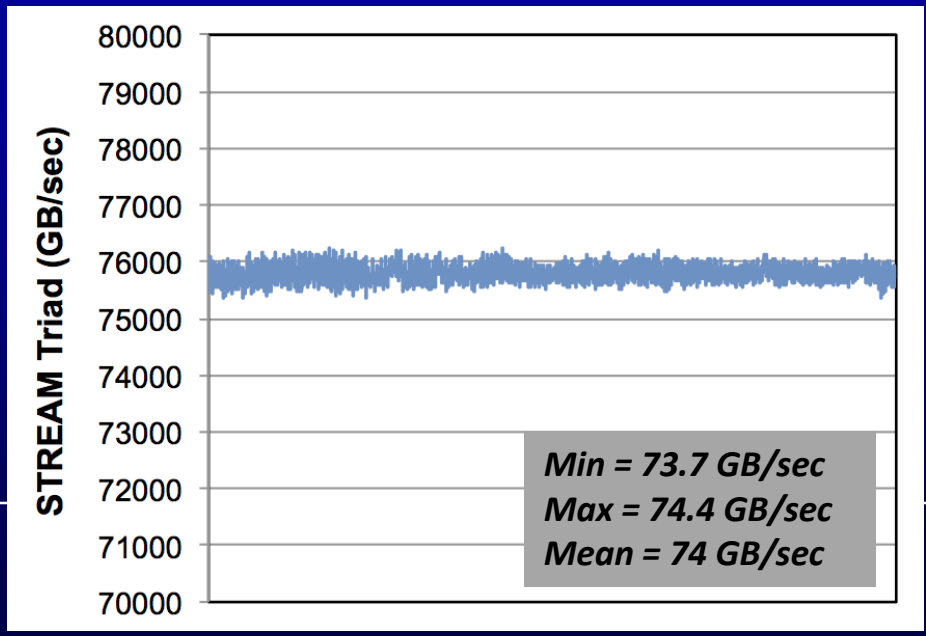
Deployment - Example Gotchas

- Another interesting result observed after repeated node-requalification tests was that our STREAM performance would initially tend to lose performance over time:
  - hosts would all pass during a system testing phase
  - after letting staff and friendly users run lots of jobs, we would observe that some hosts would have slower than expected STREAM numbers
  - a reboot would fix

- Determined to be related to significant file-system caching

➢ *We updated our SLURM epilog mechanism to flush caches between each user job*

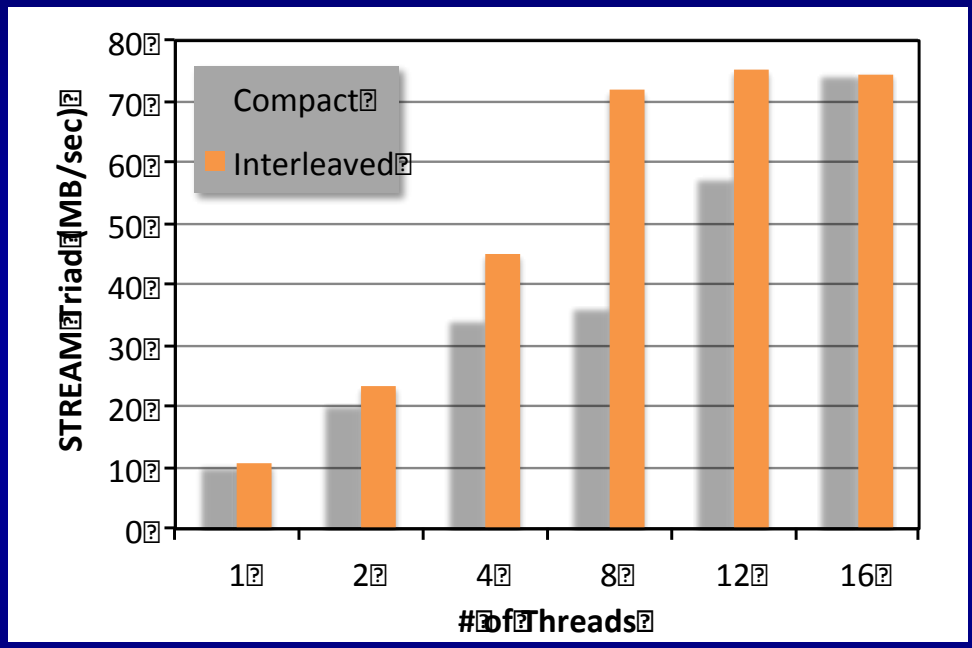
Example Re-certification from 1-22-2013 (6400 hosts)

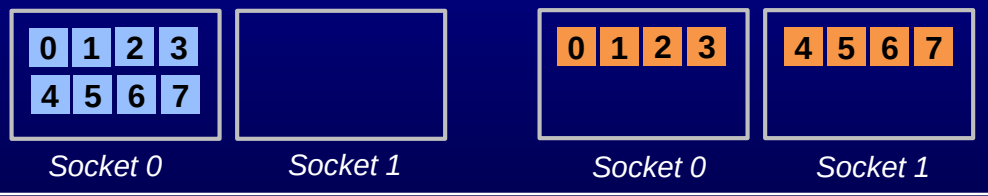


TACC

Speeds and Feeds: Memory Bandwidth (SB)

- Measured STREAM numbers on Stampede compute host shown at the right:
  - measured as a function of thread count
  - compare performance with alternate thread affinities
- Observations:
  - currently sustaining 74.6 GB/sec using all 16 cores
  - single socket saturated STREAM bandwidth is ~37 GB/sec
  - significant fractions of this peak can be achieved using ½ the number of cores

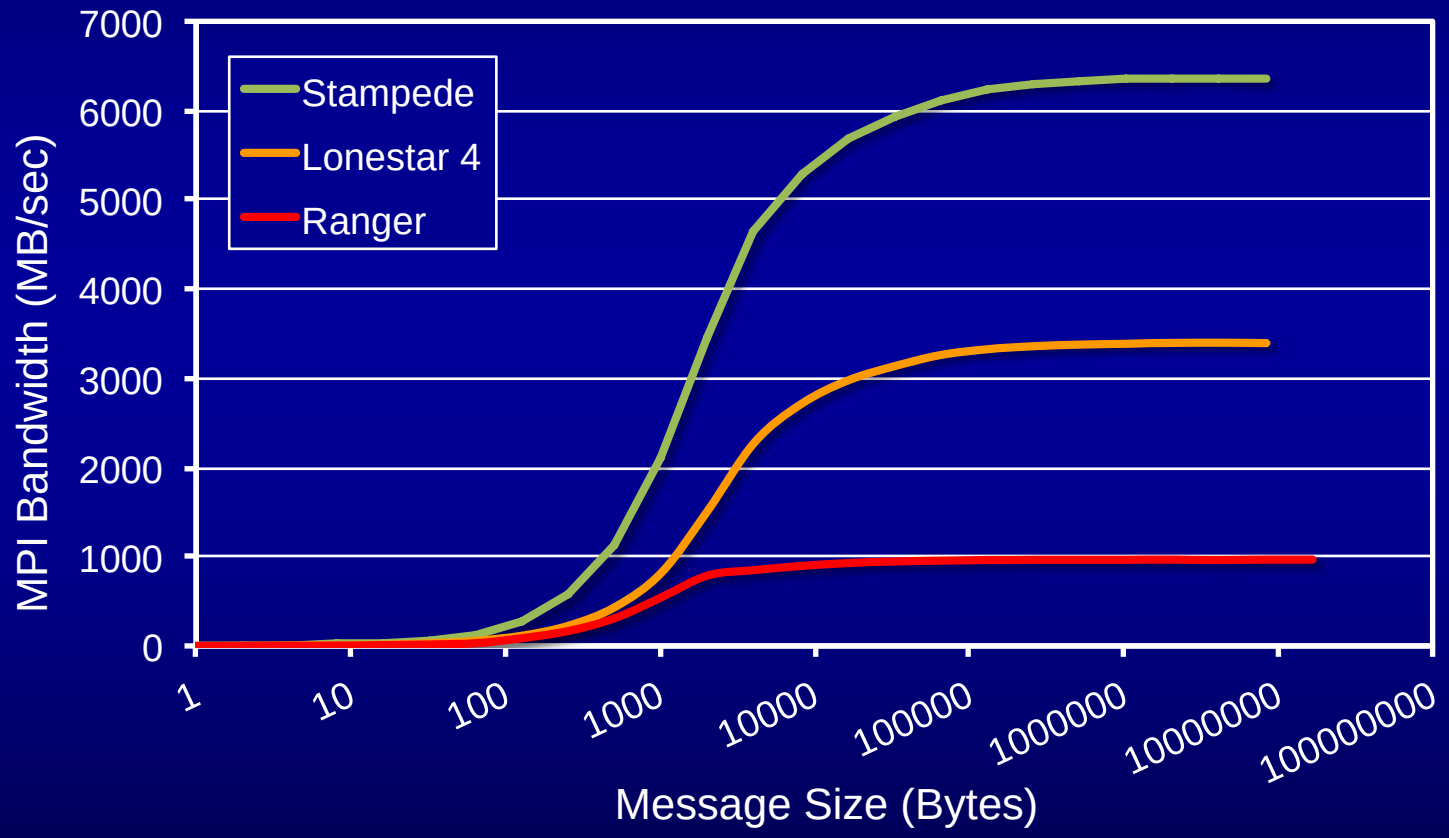




numactl used to control thread placement and memory affinity

Speeds and Feeds: P2P Bandwidth (FDR)

Comparison to previous generation IB fabrics



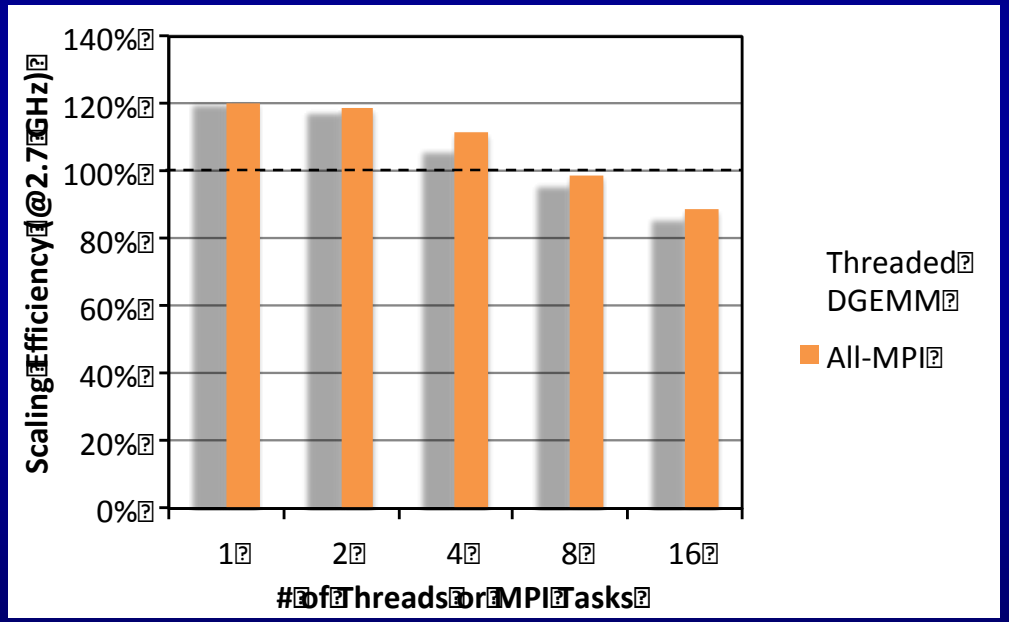
Speeds and Feeds: MPI Latencies

- Minimum value approaching 1 microsecond latency
- Notes:
  - switch hops are not free
  - maximum distance across Stampede fabric is 5 switch hops
- These latency differences continue to motivate our topology-aware efforts (more on that later in the talk)

#"switch" hops	Avg"Latency" (µsec)
1	1.07
3	1.76
5	2.54

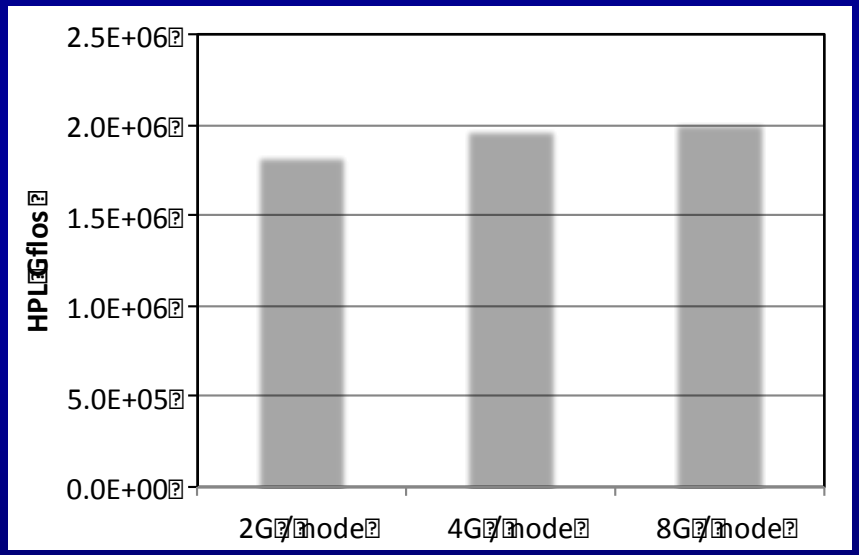
Performance Characteristics: Turbo Effects with DGEMM

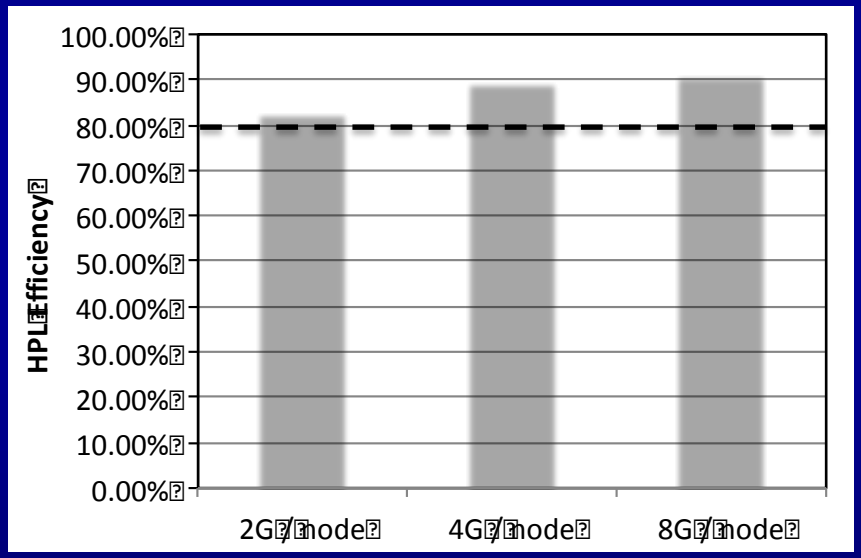
- Recent multi-core designs introduced a clock throttling feature to allow raising the nominal speed when all cores are not in use
- Introduces another knob to think about for performance optimization
- Measured Linpack performance on a Stampede compute node shown here (both threaded and all-MPI versions)
  - Efficiency is based on a 2.7 GHz nominal clock rate
  - Peak turbo frequency for these processors are 3.50 GHz
- Note that frequency scaling allows efficiencies in excess of 100% when not using all cores



Speeds and Feeds: Full System HPL (SB)

- HPL Completed on all 6400 hosts on 12/31/12
- Exceeded 90% efficiency with 8GB/node



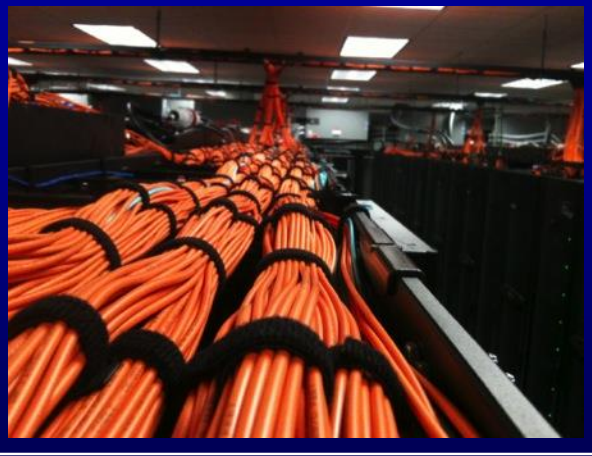


Prior to these full system runs, we also ran a heterogeneous SB+MIC run for submission at SC12 -> Stampede currently ranked 7th

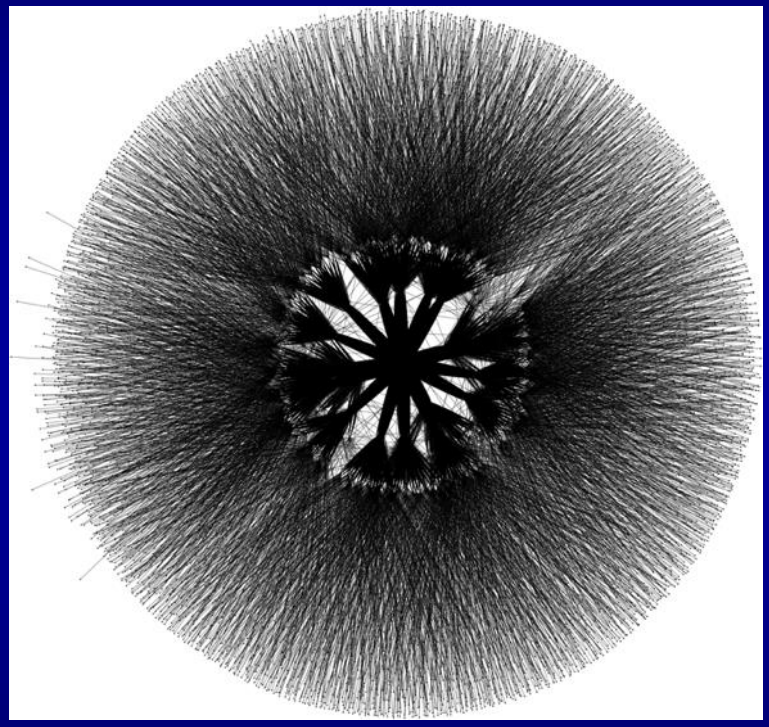
TACC

TOP500 SUPERCOMPUTER SITES

Stampede InfiniBand Topology



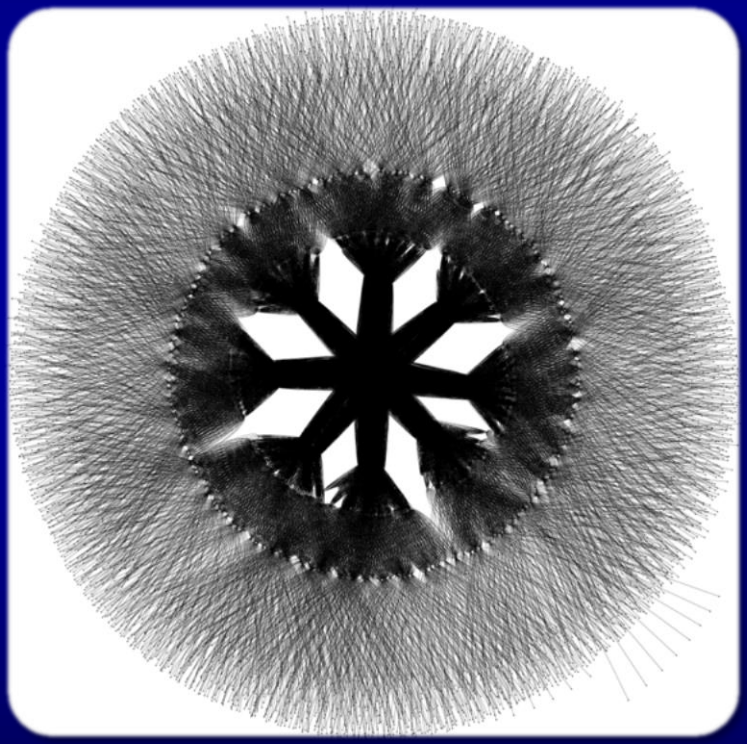
8 Core **S**witches



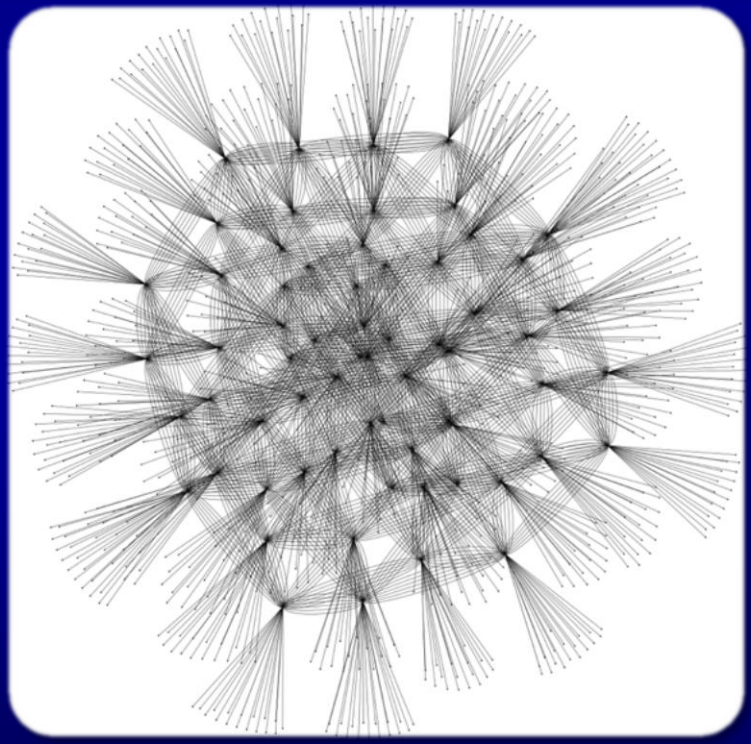
Stampede InfiniBand (fat-tree)
~75 Miles of InfiniBand Cables

Topology Considerations

- At scale, process mapping with respect to topology can have significant impact on applications



Full fat-ree (Stampede, TACC)

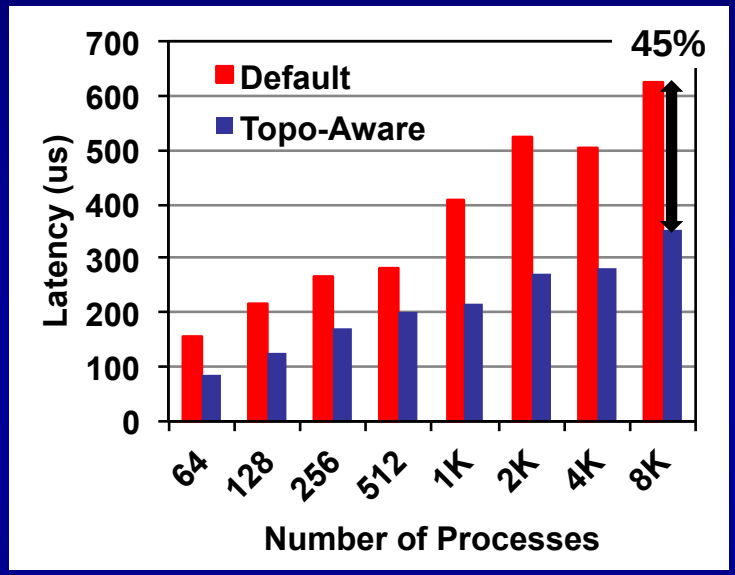


4x4x4 3D Torus (Gordon, SDSC)

Topology Considerations

- Topology query service (now in production on Stampede) - NSF STCI with OSU, SDSC
  - caches the entire linear forwarding table (LFT) for each IB switch - via OpenSM plugin or *ibnetdiscover* tools
  - exposed via network (socket) interface such that an MPI stack (or user application) can query the service remotely
  - can return # of hops between each host or full directed route between any two hosts

Nearest neighbor application benchmark from Stampede [courtesy H. Subramoni, SC 12]




```
query c401-101:c405-101
     c401-101 0x0002c90300776490 0x0002c903006f9010 0x0002c9030077c090 c405-101
```

- We will also be leveraging this service to perform topology-aware scheduling so that smaller user jobs will have their nodes placed closer together topologically
  - have created simple tool to create SLURM topology config file using above query service
  - works, but slows interactivity when users specifiy maximum # of switch hops desired during job submission

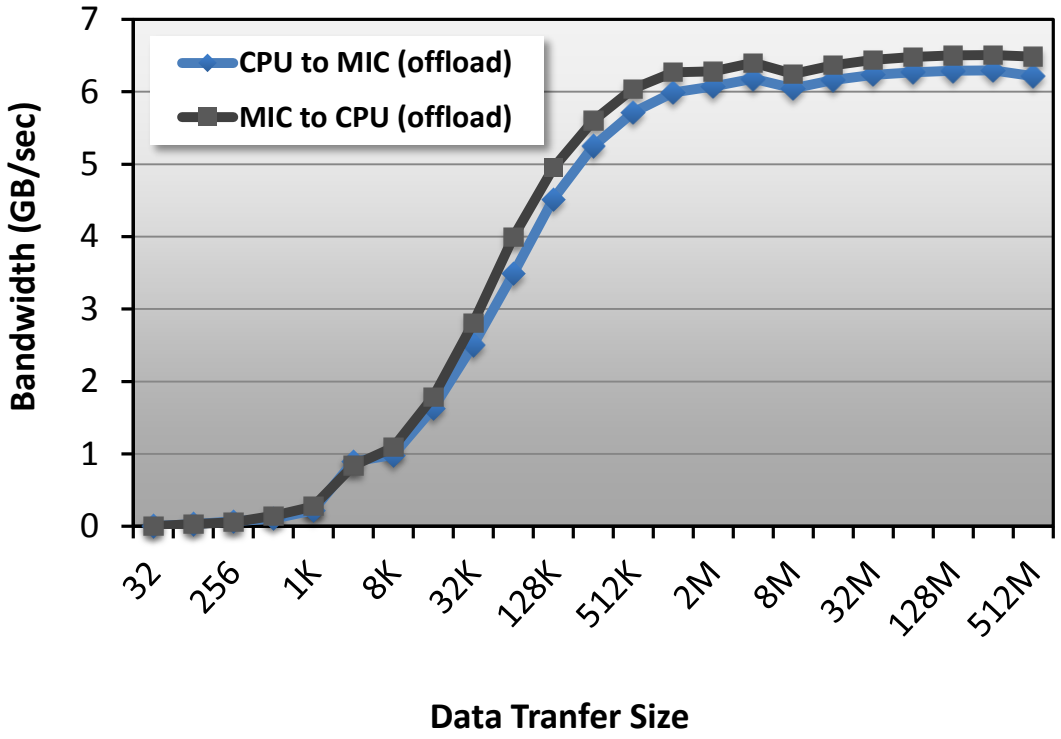
Big Systems are never really done

- There is continuous work and improvement through the system lifecycle to make the environment better.
  - This is one reason you need a great team, not just a vendor.
- For instance, over the first 6 months, we identified and worked with several partners to greatly enhance some of the Xeon Phi data transfers…

Stampede Data Movement

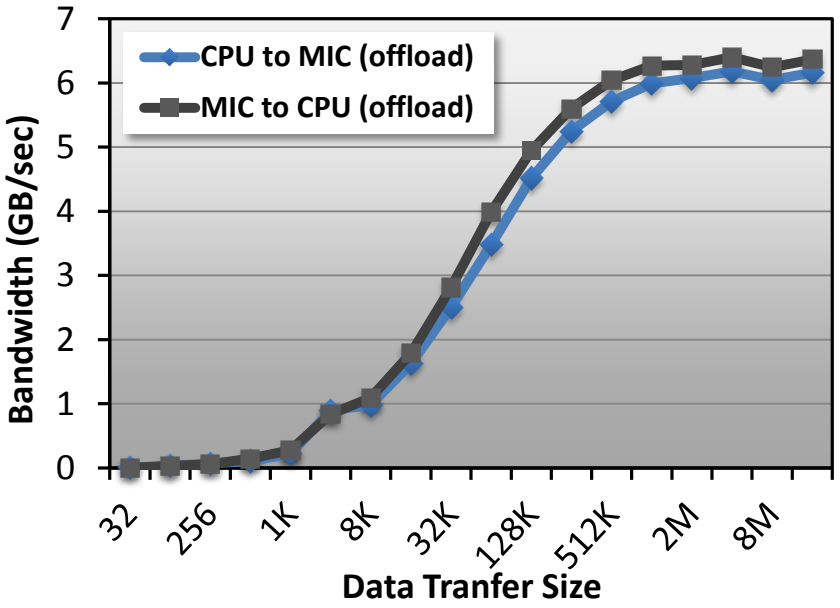
New MPI Features

- Efficient data movement is also critical in a heterogeneous compute environment (SB+MIC)
- Let's look at current throughput between host CPU and MIC using standard "offload" semantics
  - bandwidth measurements are likely what you would expect
  - symmetric data exchange rates
  - capped by PCI XFER max

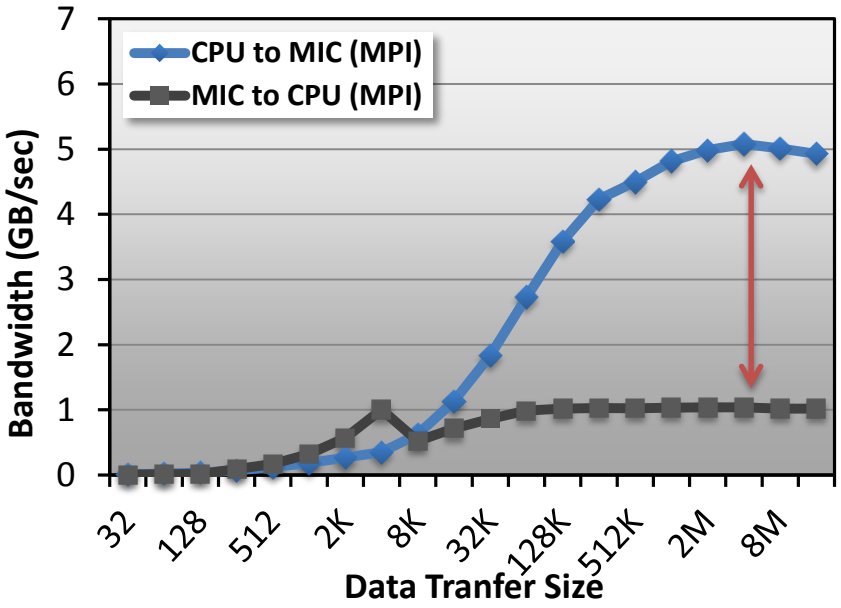


Phi Data Movement

Offload Test (Baseline)



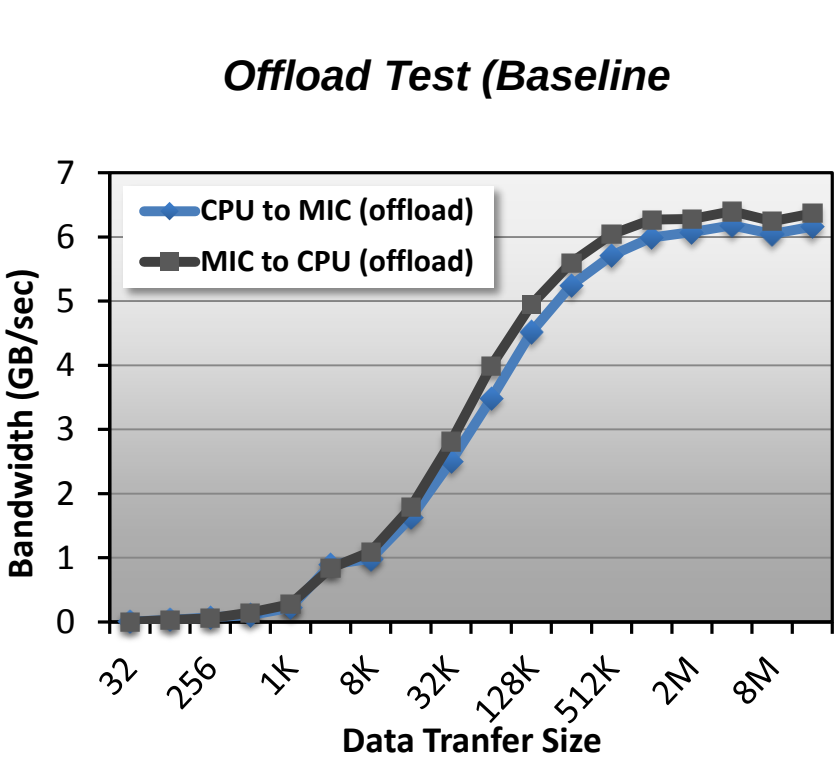
OSU Bandwidth Test
Intel MPI 4.1.0.030 (Feb 2013)
DAPL: ofa-v2-mlx4_0-1u

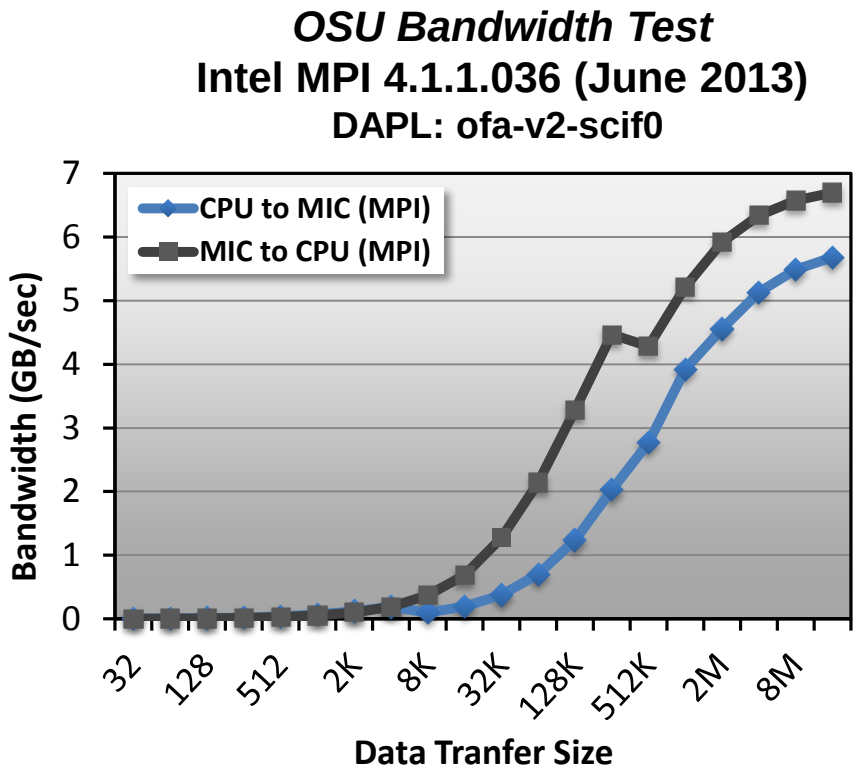


asymmetry undesired for tightly coupled scientific applications…

TACC

Phi Data Movement (improvement)

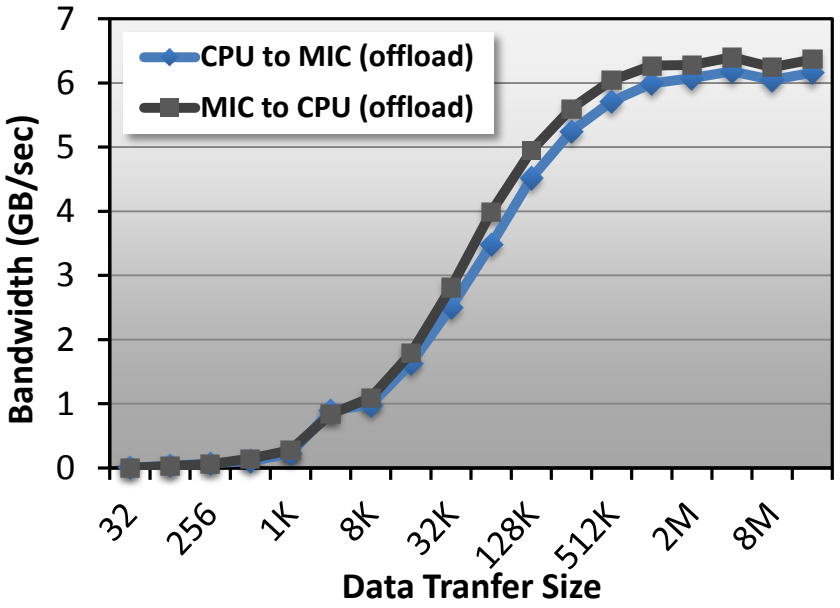




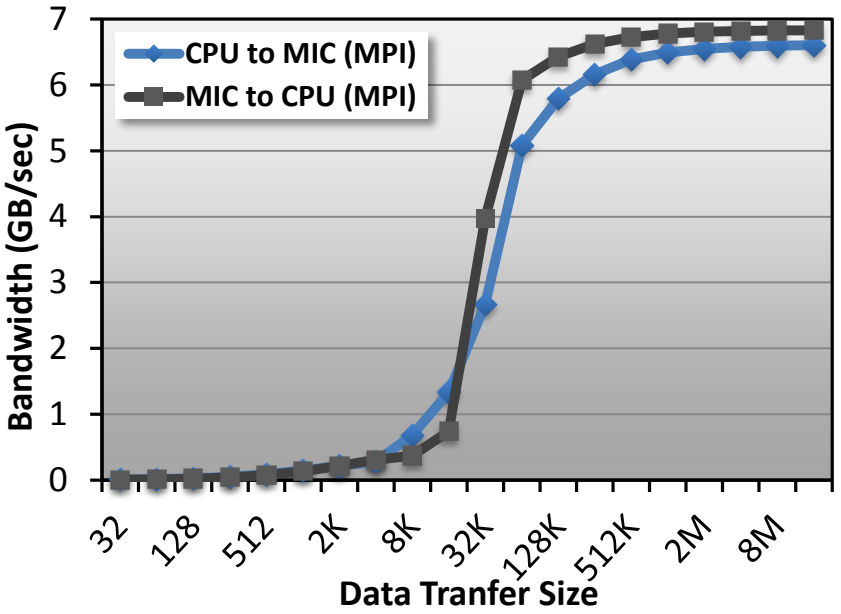
TACC

Phi Data Movement (improvement)

Offload Test (Baseline)



OSU Bandwidth Test
MVAPICH2 Dev Version (July 2013



New developments to proxy messages through HOST

TACC

Work on Improvements still continues

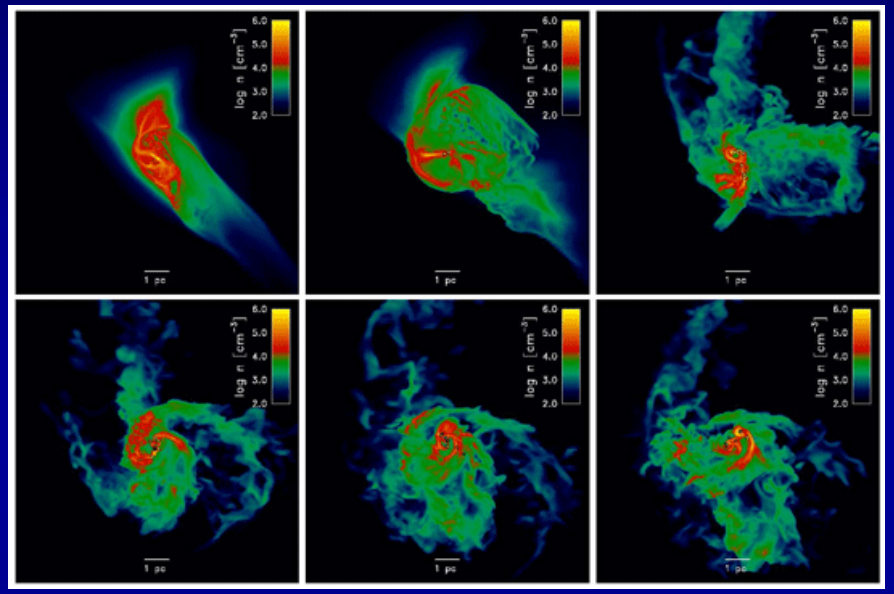
- In conjunction with MVAPICH team, workoing on CMA (Cross Memory Attach) to optimize inter-node communication.
  - Jerome to present tomorrow.
- Working to keep up with MPSS (MIC software stack) changes in latest versions of MVAPICH.
- MVAPICH-MIC developed with help from the Stampede project is now available on other Phi systems!

Stampede Recent Science Highlights

Heavy Metal in the Early Cosmos

- Researchers at UT-Austin (Bromm et al) used Stampede to perform ab initio simulations refining how the first galaxies are formed, and how metals in stellar nurseries influenced characteristics of first stars in the galaxy.

- These simulations are making predictions that can be validated in 2018 by the James Webb Space Telescope (and in fact determine how JWST is used).



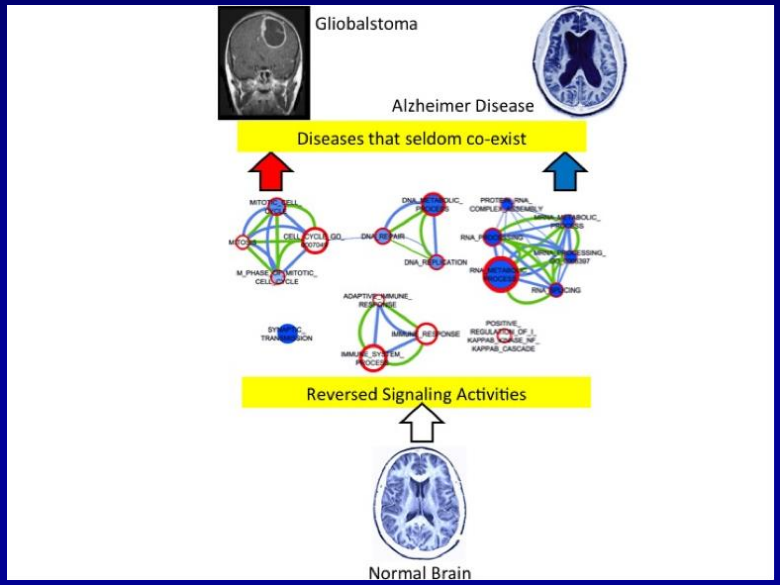
"It is a really exciting time for the field of cosmology. We are now ready to collect, simulate and analyze the next level of precision data...there's more to high performance computing science than we have yet accomplished."

Astronomer and Nobel Laureate Saul Perlmutter,Supercomputing '13 keynote address

Stampede Recent Science Highlights

A link between Alzheimer's and Cancer

- A team led by Houston Methodist Hospital (with researchers in Harvard, Taiwan, and Italy) used Stampede to find a link between Alzheimer's and GBM, one of the most aggressive forms of brain cancer.
- This systems biology approach has uncovered linked signaling pathways, and identified 15 gene ontology terms relating the diseases.



"This work of Dr. Wong's is quite exciting in that it shows connections between two of the most intractable diseases in society. And while our focus is on cancer, the great hope is that as we make these connections we can leverage that to find new targets and opportunities that can provide meaningful intervention for either disease."

- Dan Gallahan,NIH, deputy director,
National Cancer Institute

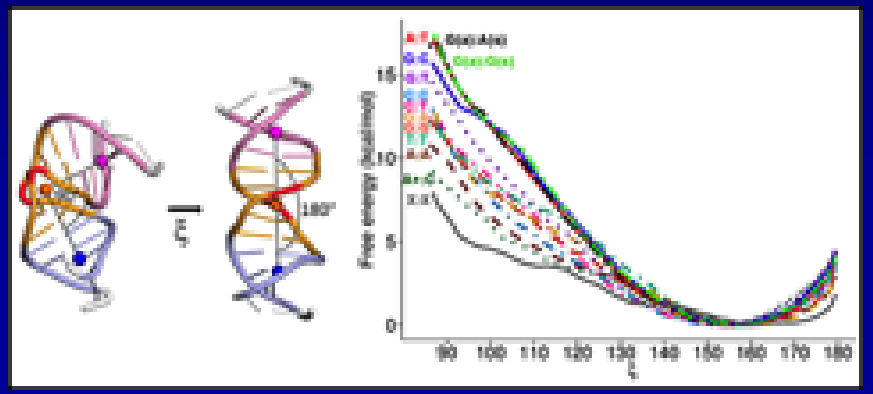
The gene sequencing data size would easily be 1000-fold larger than the microarray data in the reported study, which means the need to use TACC's Stampede supercomputing cluster for number crunching is even more eminent."

- Stephen Wong,
Houston Methodist Research Institute

Stampede Recent Science Highlights

How DNA Repair Helps Prevent Cancer

- Researchers from Michigan State University used Stampede to understand DNA repair and bending mechanisms.
- Numerical simulations provide a detailed view down to the atomistic level of how MutS and MSH2-MSH6 scan DNA and identify which DNA needs to be repaired (tens of millions of CPU core hours per year).
- https://www.tacc.utexas.edu/news/feature-stories/2013/how-dna-repair-helps-prevent-cancer



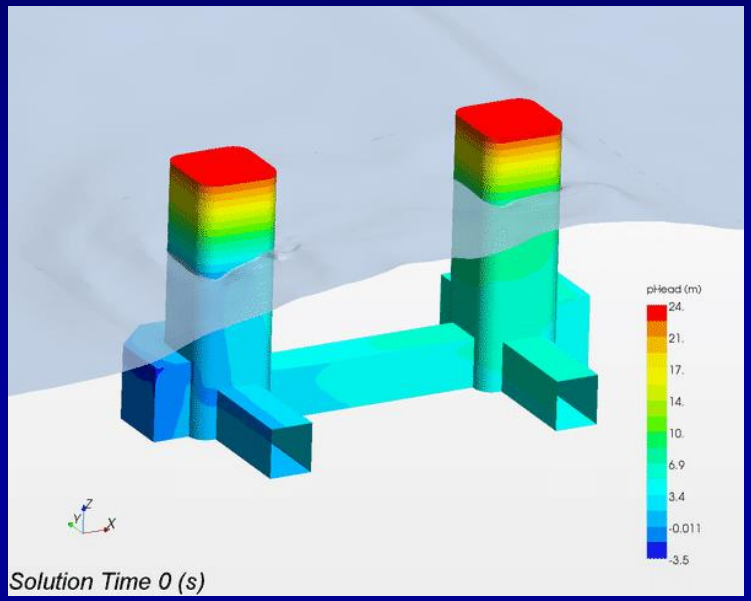
"We need high-level atomic resolution simulations to get insights into the answers we are searching for and we cannot run them on ordinary desktops. These are expensive calculations for which we need hundreds of CPUs to work simultaneously and TACC resources made that possible."

- Michael Feig, Michigan State University

Stampede Recent Science Highlights

Design of Offshore Oil Platforms

- An industrial partner, Technip, used Stampede to run full-scale simulations using a numerical wave basin to design offshore floating oil platforms.
- Technip's business is to design, construct and install offshore platforms for major oil companies such as BP, Shell, Chevron, and ExxonMobil.
- Modeling has replaced wave tank tests that take up to a year to perform.



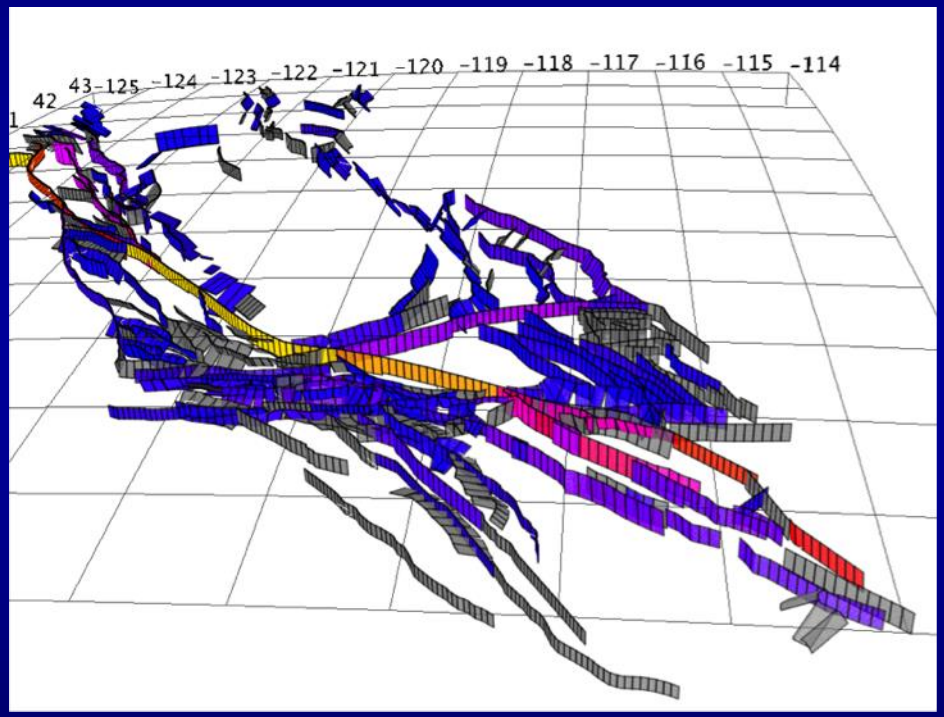
"Technip has one of the largest computer clusters among engineering companies, but the simulations would take weeks to complete", said Jang Kim, Chief Technical Advisor at Technip. "Stampede has allowed us to run simulations in one day or even overnight. We are then able to get these safer designs out into use faster."

Stampede Early Science Highlights

Predicting Earthquakes in California

- Researchers from the Southern California Earthquake Center used Stampede to predict the frequency of damaging earthquakes in California for the latest Uniform California Earthquake Rupture Forecast (UCERT3).

- An Earthquake Rupture Forecasts gives the probability of all possible, damaging earthquakes throughout a region and over a specified time span.



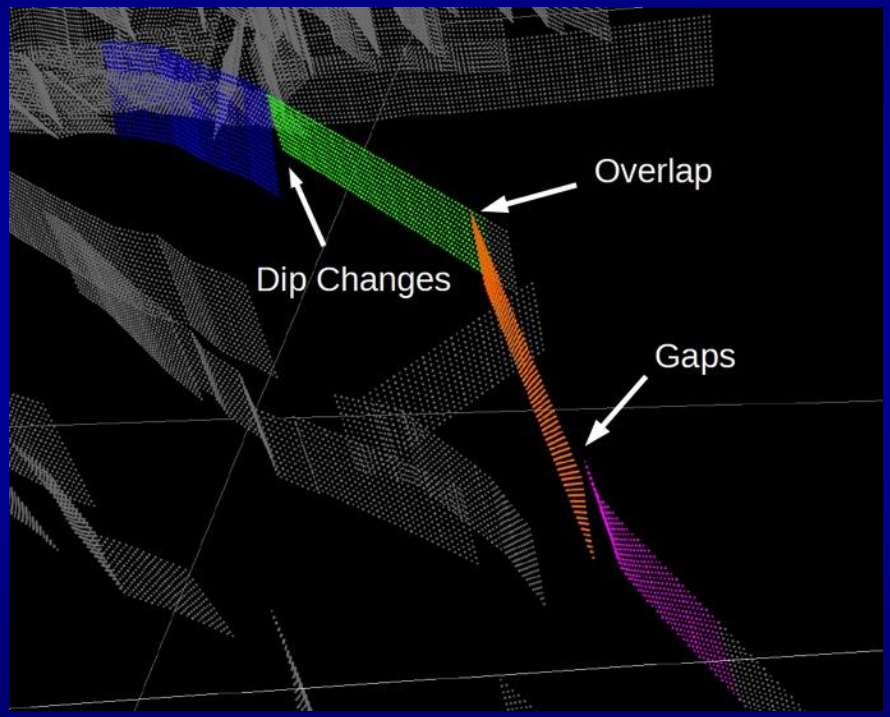
Stampede Early Science Highlights

Predicting Earthquakes in California

- The results of the simulations will be incorporated into USGS's National Seismic Hazard Maps which are used to set building codes and insurance rates.

- *"We do a lot of HPC calculations, but it's rare that any of them have this level of potential impact."*

  Thomas Jordan, Director Southern California Earthquake Center



Next System Up: Chameleon

- Available in summer 2015
- Partnership of U of Chicago, TACC, Northwestern, UTSA, and of course the MVAPICH team!
- A testbed for cloud and computer science research
- Low level access for those studying system design tradeoffs, including such issues as:
  - Advanced network in clouds
  - Use of SR-IOV, and virtualization + Advanced networks.

And then

- The “Stampede 1.5” upgrade plan will likely be modified to include *self-hosted*, standalone, second generation Xeon Phi nodes.
  - Another technology for MVAPICH to conquer.

What does the future hold

- Some form of container/virtualization inevitable.
  - Looking forward to adopting SR-IOV soon.
- Interconnect is evolving ; big changes in store
  - New generations of Infiniband
  - Intel OmniScale, other options coming
- Will the HCA go away at the server level?

Will the HCA go away at the server level?

- The HCA could move into the processor die
  - Like the PCI controller, memory controller, basic graphics controller, and other things before it, the HCA could integrate with the processor
  - PCI slot/pin availability could suffer in this approach.
- The HCA could move into the switch
  - Direct connect PCI cables could connect to a switch at the rack level
  - Infiniband/ethernet from switch to other racks.
- Either way, lots of optimization work for MVAPICH of the future!

Thank You!

And

Thanks to DK and team for the uplift MVAPICH provides for everyone!!!!

Dan Stanzione

dan@tacc.utexas.edu