SAND2013-0014P

Exceptional service in the national interest

Sandia National Laboratories



Finding an On-Ramp to the Exascale Highway

S.D. Hammond

Scalable Computer Architectures, Sandia National Laboratories
sdhammo@sandia.gov

U.S. DEPARTMENT OF ENERGY NNSA

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

This work is from a big team...

Sandia National Laboratories

- **Advanced Test Bed Project Management** - Jim Ang, Jim Laros and Sue Kelly
- **System Administrators** - Jason Repik, Victor Kuhns, Jim Brandt and Ann Gentile
- **PMAT and Mantevo** - Richard Barrett, Courtenay Vaughan, Mike Heroux, Jagan Jayaraj, Li Tang, Christian Trott
- **Trinity Procurement Team** - Doug Doerfler and team
- **SST Team** - Arun Rodrigues, Scott Hemmert, Brian Barrett, Jon Wilson, John Vandyke
- **Local System Administrators** - Bill Goldman and team
- **Vendors** - Intel, AMD, NVIDIA etc
- **Very Patient Nearby Workers** - sorry for the noise!

Talk Outline

Sandia National Laboratories

- Overview of the Advanced Test Bed Project
  - What are we trying to achieve?
- Discussion of two of Sandia's test beds
  - Teller - AMD Fusion APU cluster
  - Compton - Intel MIC / Knights Corner cluster
- Successes and Takeaways
- Feedback?
  - What can we help you with?
  - Concerns - are we missing things?

Warning: Performance Results Notice

- Results are obtained from pre-production hardware and alpha/beta-class software stacks
  - Your mileage will vary, reproducibility is sometimes difficult
- Programming models and tools are changing rapidly; we may not necessarily be able to reproduce with new software stacks
- Focus more towards characteristics of hardware we can exploit
- Opportunity to influence vendors

Legal Information

Sandia National Laboratories

- Intel Knights Ferry is a software development platform for Intel Many Integrated Core (Intel MIC) architecture
- Knights Corner refers to Intel's Xeon Phi Coprocessor (code-named Knights Corner)
- All results included in this work are taken on pre-production Knights Corner cards and pre-production software stacks
  - Your performance improvements will be different
- Host processors refer to Intel's Xeon processor family code-named Sandy Bridge or Ivy Bridge

ADVANCED ARCHITECTURES PROJECT

Finding an On-Ramp to Exascale Computing

Advanced Architectures Project

Sandia National Laboratories

- **Aim:** to be a scout for future computing architectures

- Reduce impact on Sandia and its partner labs from rapid technology change which requires significant rewrites/ modifications in production codes
  - When we make "the change" it is the right move for code longevity, porting efforts, performance *etc*
  - Put another way - to go through all the pain up front so the transition for full codes is made easier

- Procurement of test-bed hardware
- Development of proxy-applications and micro-benchmarks
- Porting, optimization, tools, system-software *etc*

Deployed Architectures

Sandia National Laboratories

- **Teller** (ECN) - AMD Fusion APU ("Trinity") cluster
  - AMD A10-5800K (3.8GHz Quad-core CPU, 800MHz 384-core GPU)
- **Compton** (ECN) - Intel Sandy Bridge and Knights Corner cluster
  - Dual-socket Intel E5-2670 (2.6GHz 8-core CPU)
  - Dual pre-production Knights Corner Co-processor cards
- **Curie** (SRN) - Cray XK6
  - NVIDIA Fermi X2090 (512-core) GPUs (Upgrade due in Q4'12/Q1'13)
- **Shannon** (ECN?) - NVIDIA GPU cluster?
  - Dual NVIDIA Kepler?

Deployed Architectures

Sandia National Laboratories

- **Teller** (ECN) - AMD Fusion APU ("Trinity") cluster
  - AMD A10-5800K (3.8GHz Quad-core CPU, 800MHz 384-core GPU)
- **Compton** (ECN) - Intel Sandy Bridge and Knights Corner cluster
  - Dual-socket Intel E5-2670 (2.6GHz 8-core CPU)
  - Dual pre-production Knights Corner Co-processor cards
- **Curie** (SRN) - Cray XK6
  - NVIDIA Fermi X2090 (512-core) GPUs (Upgrade due in Q4'12/Q1'13)
- **Shannon** (ECN?) - NVIDIA GPU cluster?
  - Dual NVIDIA Kepler?

Application Focus

Sandia National Laboratories

- Focus area for applications is:
  - Mantevo mini-applications
    - **miniFE** - implicit finite element assembly and solve
    - **miniMD** - molecular dynamics (Christian Trott)
    - **miniGhost** - finite difference stencil
  - Office of Science Co-Design Center applications:
    - **ExMatEx** - Materials in Extreme Environments
      - CoMD, LULESH, VPFFT, Voro3D
    - **ExaCT** - Combustion Co-Design Center
      - S3D, SMC, CNS
  - Trinity Procurement partners
    - **NNSA/ASC, LBNL**

SANDIA'S TELLER CLUSTER

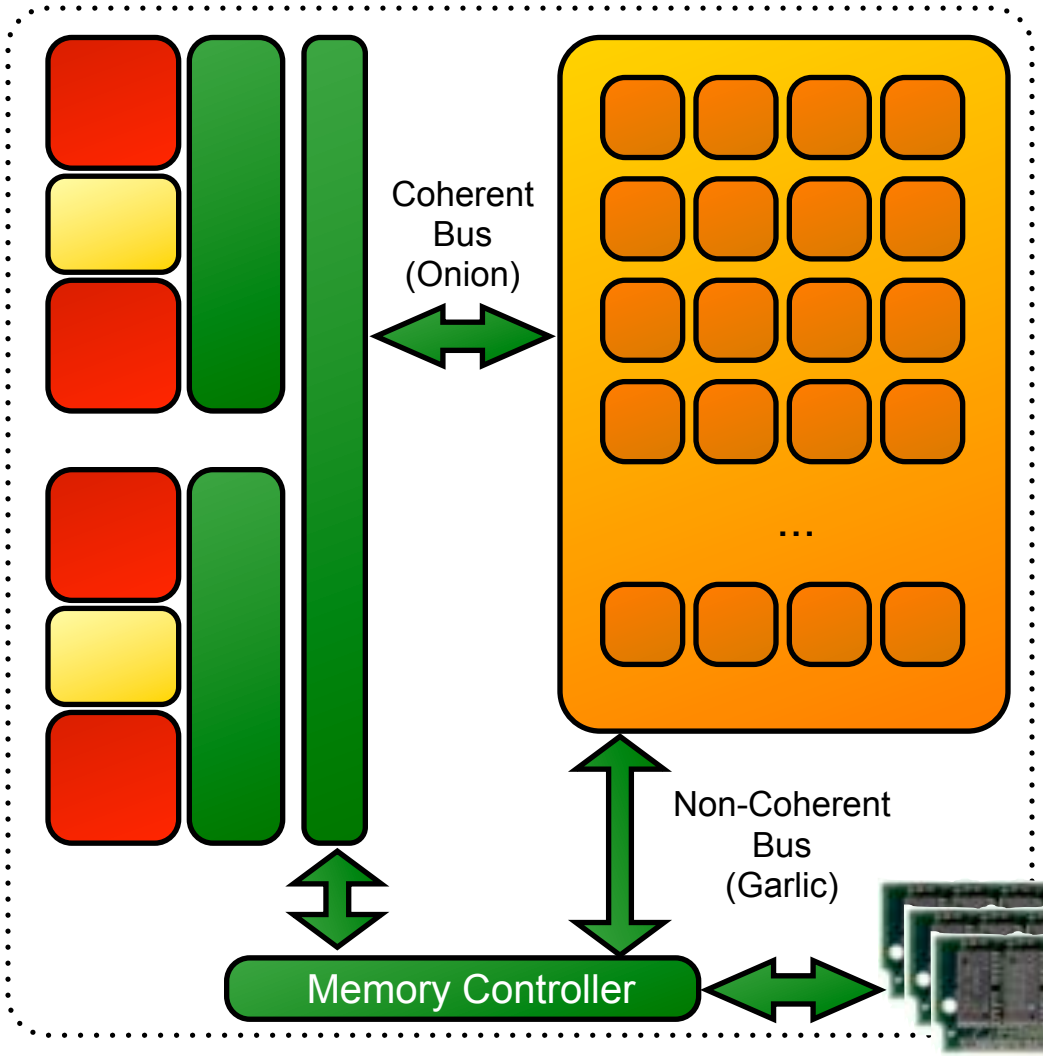
First of a kind AMD Fusion APU Test Bed

AMD Fusion (APU)

Sandia National Laboratories

CPU

2 x Core Pairs
3.8GHz



GPU

HD 7660D
800 MHz
384-cores

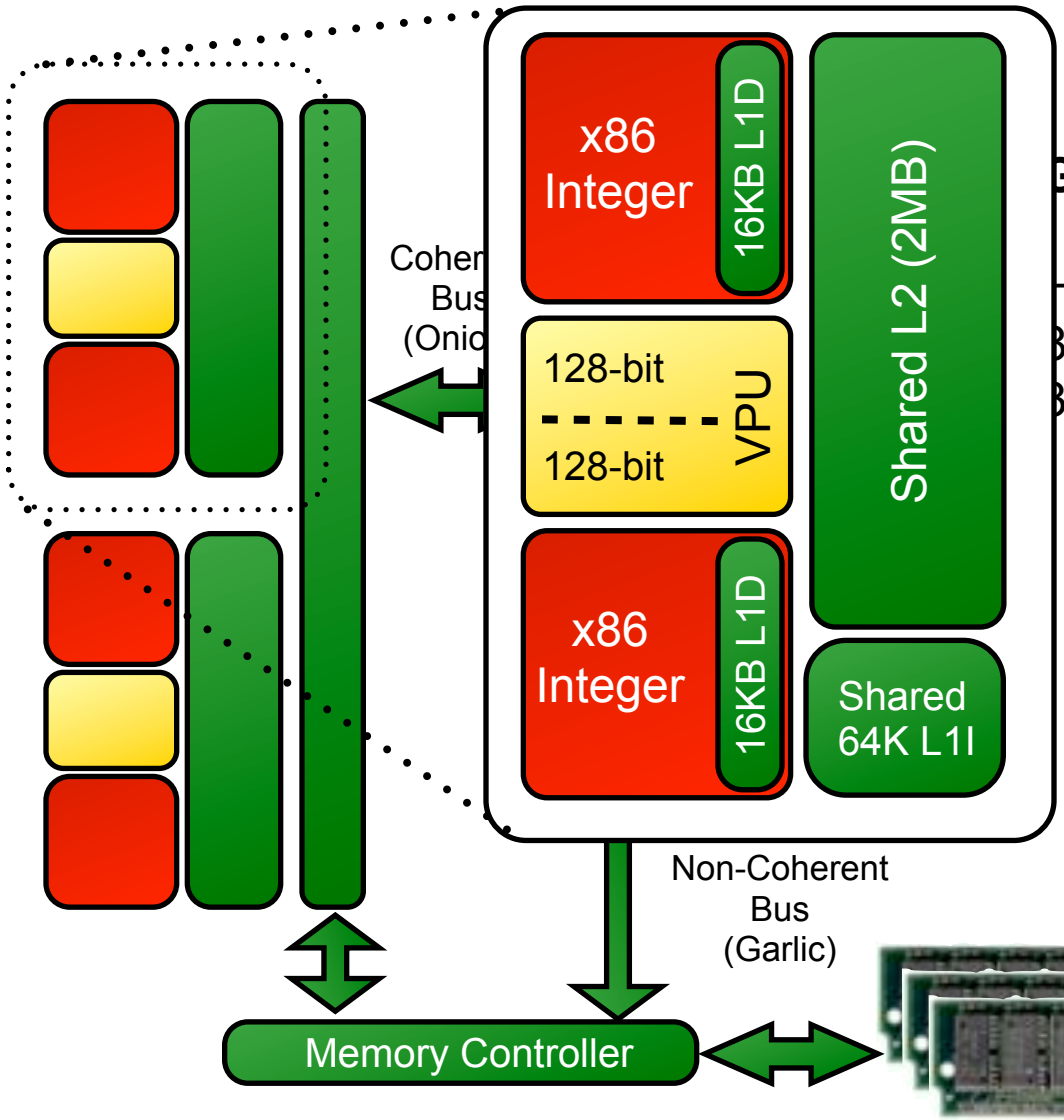
AMD Smarter Choice

AMD Fusion (APU)

Sandia National Laboratories

CPU

2 x Core Pairs
3.8GHz



GPU

HD 7660D
800 MHz
384-cores

AMD Smarter Choice

Teller Specification

Sandia National Laboratories

- 104 nodes of 1 x A10 APU
  - 16GB RAM per node
- InfiniBand QDR Interconnect
- Integrated by Penguin Computers
- Designed for power measurement experiments

Programming APUs

Sandia National Laboratories

- CPU components of APUs run traditional x86 binary
  - AVX compatible (close to Sandy Bridge implementation of AVX 1)
  - FMA-4 implementation (currently only AMD)
- GPU components run OpenCL
  - OpenCL so far has to be written by hand - time consuming process
  - Not currently C++ compatible (non-standard extensions available)
  - CAPS and PGI have announced support for OpenACC
- "Zero-copy" shared memory is not enabled in AMD currently released driver version

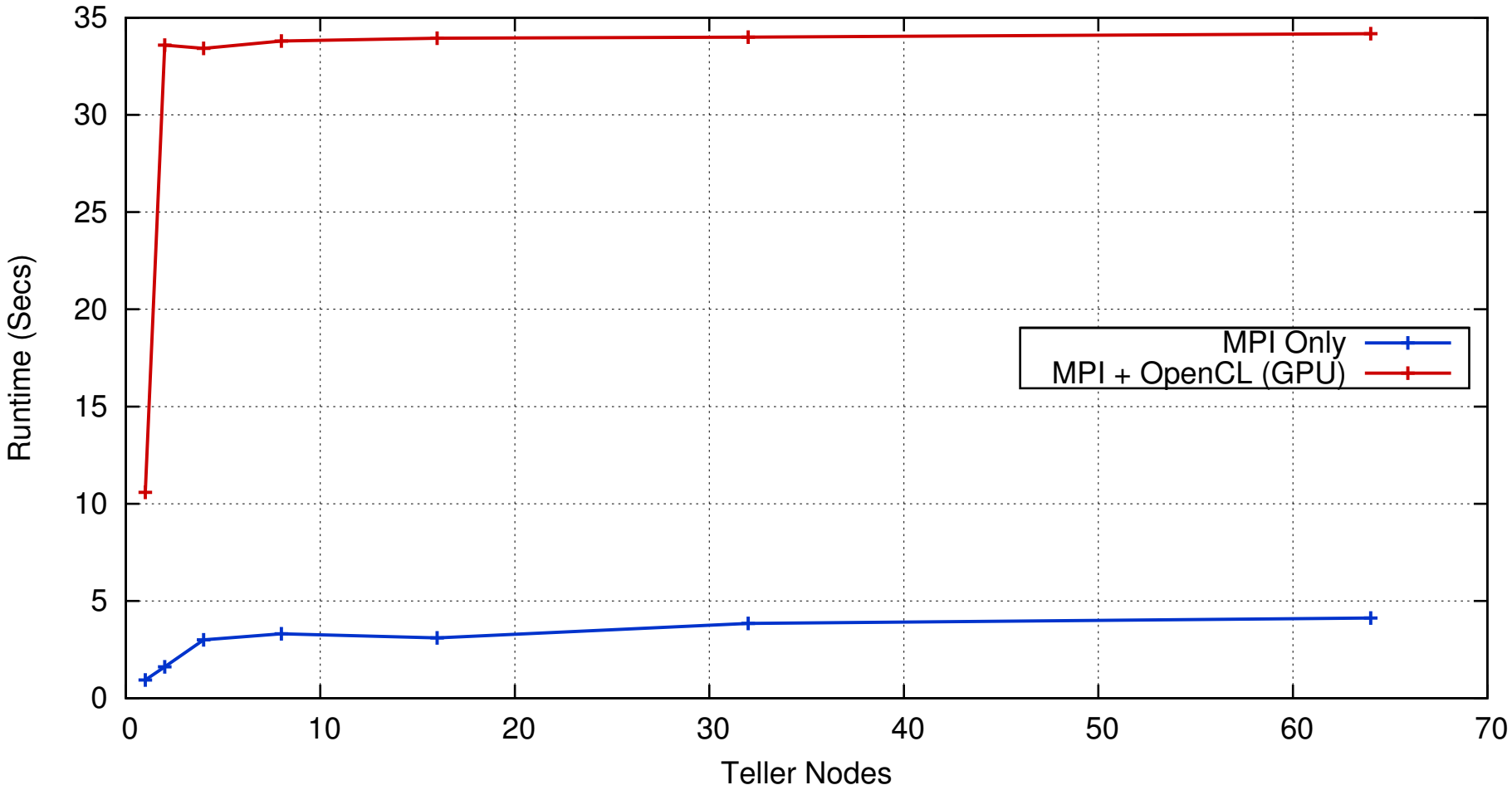
Memory Constraints in APUs

Sandia National Laboratories

- AMD intends to provide 'zero-copy'
  - *Pointer-as-a-pointer* - you can allocate memory on CPU and use on GPU
  - Eventually coherent (by end of kernel)
- Current work is based around driver
  - Limitations are 32-bit, zero-copy bandwidth is not being achieved on our systems - implies problems
  - Current limit is approximately 80% of 256MB of BIOS allocated memory
    - Way too little for our algorithms, made worse by OpenCL 50% allocate
    - High cost associated with streaming blocks to GPU
- Eventually APU memory subsystem is still DDR3
  - STREAM Triad peaks at 2 threads, 13.1 GB/s
  - lmbench estimates 17.58GB/s read, 7.53GB/s write

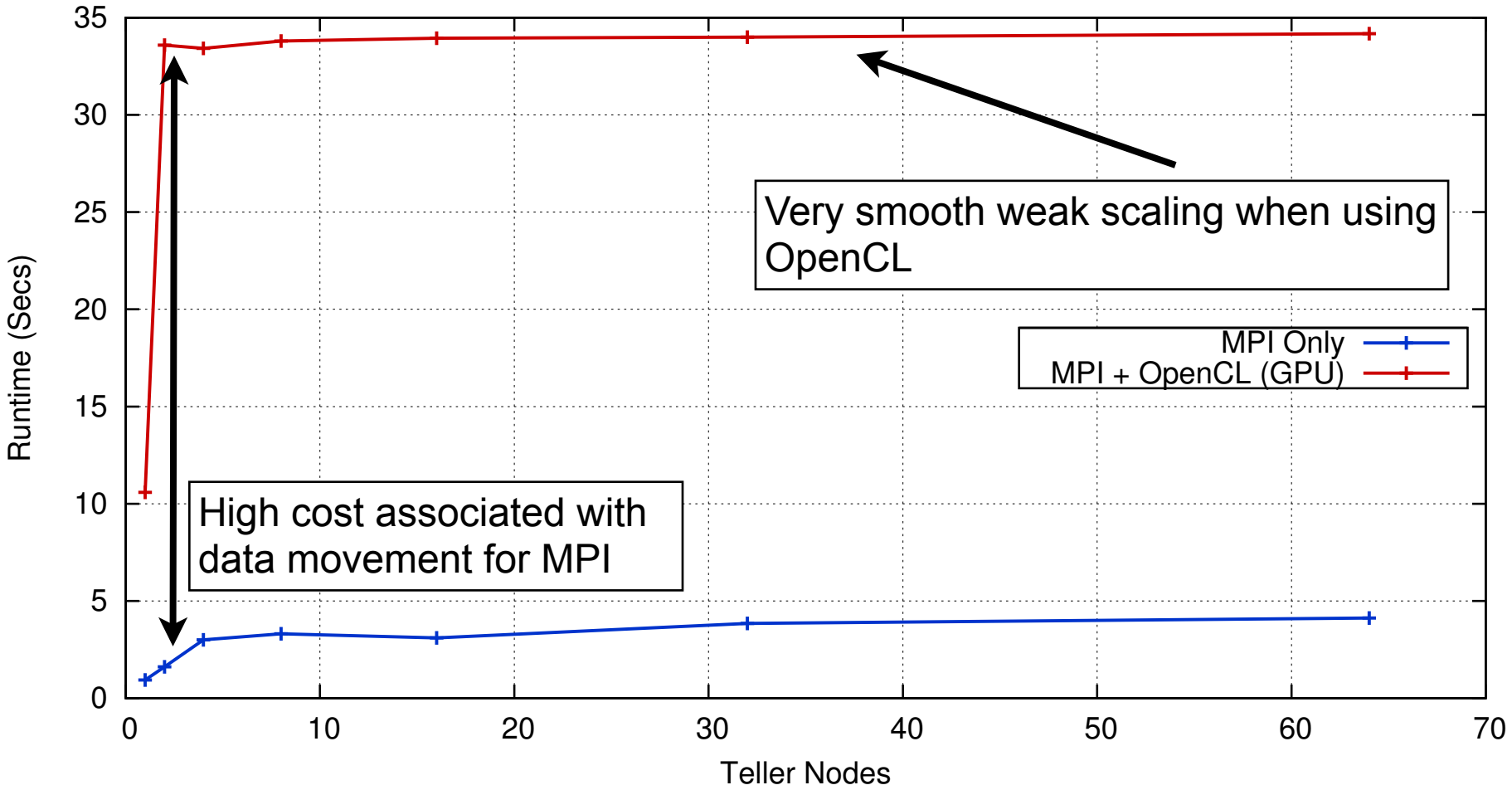
MiniFE Performance

Sandia National Laboratories



MiniFE running a 100 x 100 x 100 problem per node (weak scaled)

MiniFE Performance



MiniFE running a 100 x 100 x 100 problem per node (weak scaled)

CoMD Performance

Sandia National Laboratories

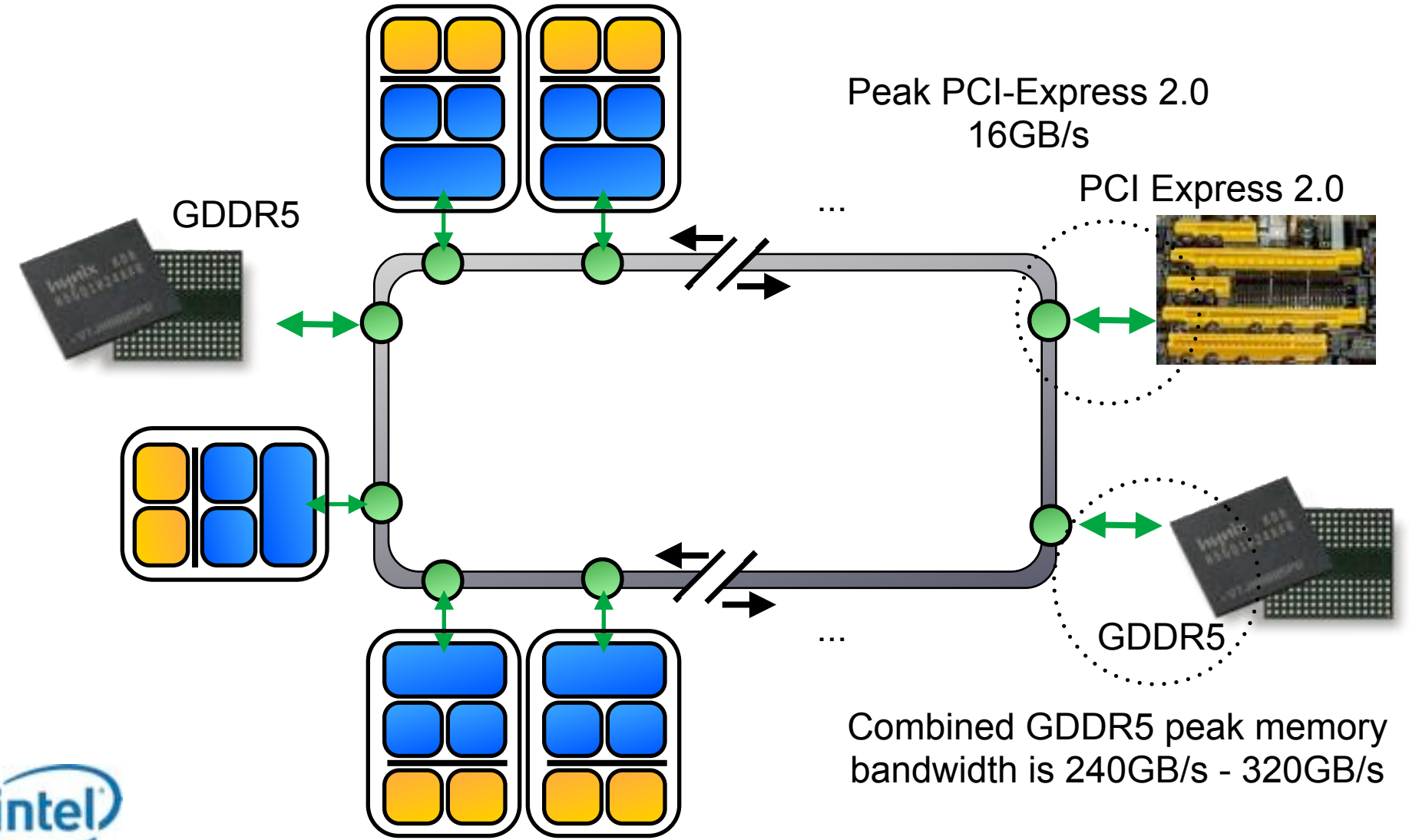
LJ Atom Count	CPU Only	APU-GPU	Discrete GPU
8192	1.14	7.00	1.89
262144	1.17	6.73	1.77

- Code needs significant performance optimization
- APU GPU lags behind discrete offerings from AMD
- Severely limited by memory aperture when GPU really needs more data to account for overhead

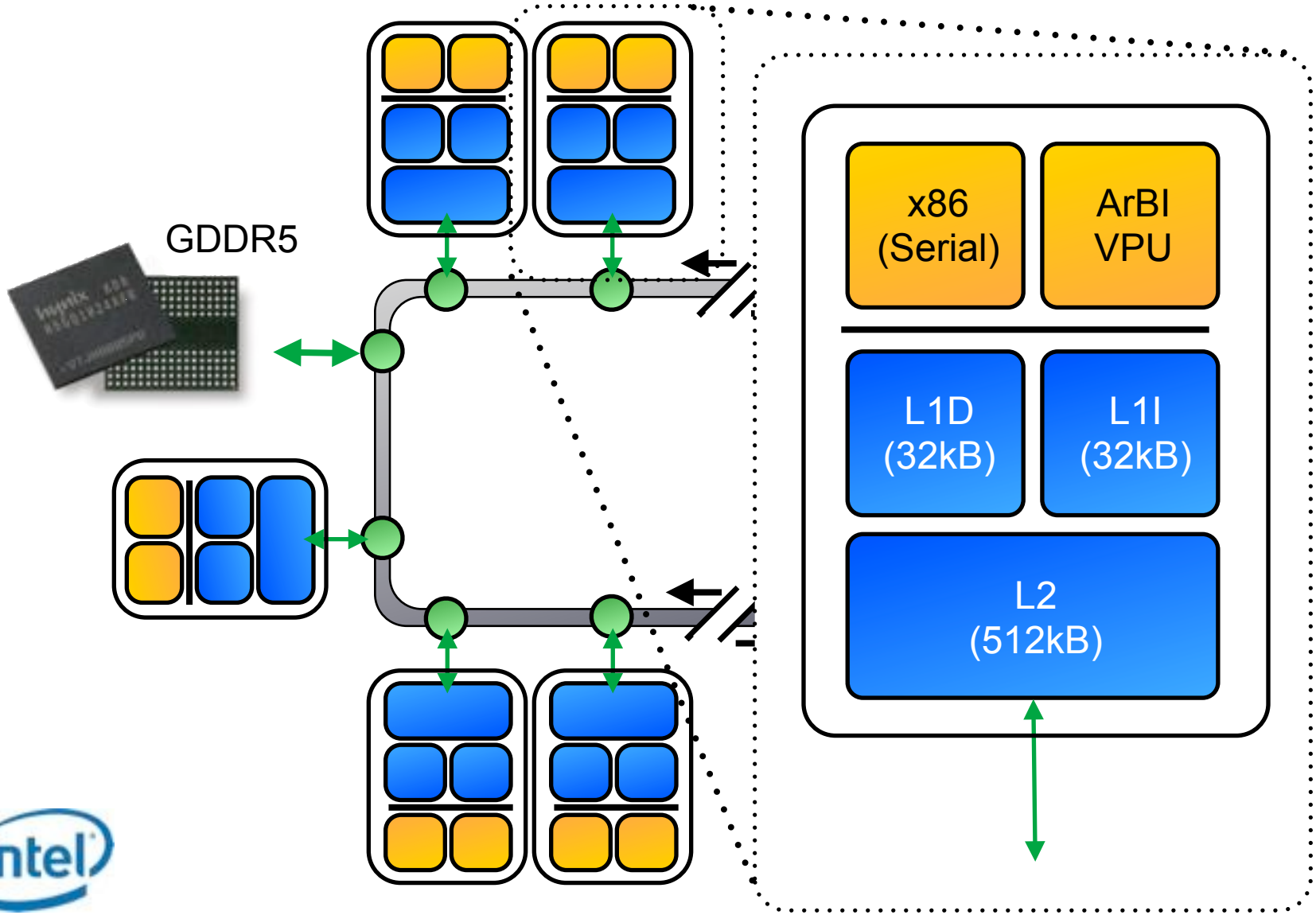
SANDIA'S COMPTON CLUSTER

First of a kind Intel MIC Architecture test bed

Intel MIC Architecture



Intel MIC Architecture



Intel MIC Architecture

Sandia National Laboratories

- Vector Processing Unit ("VPU")
  - 512-bit = 8 doubles, 16 floats
  - Vector Gather and Scatter support (including gather prefetch to L2)
  - Cross-lane operations - reduction operations within VPU register
- x86 cores execute only basic x86
  - Will not permit binary compatibility with Xeon hosts
  - Four-way Hyper-Threading ("Third generation")
    - Single thread can issue at best every second clock cycle
- L2 caches are distributed and directory managed
  - Reduces traffic on the ring bus but *can* increase latency

Compton Specification

Sandia National Laboratories

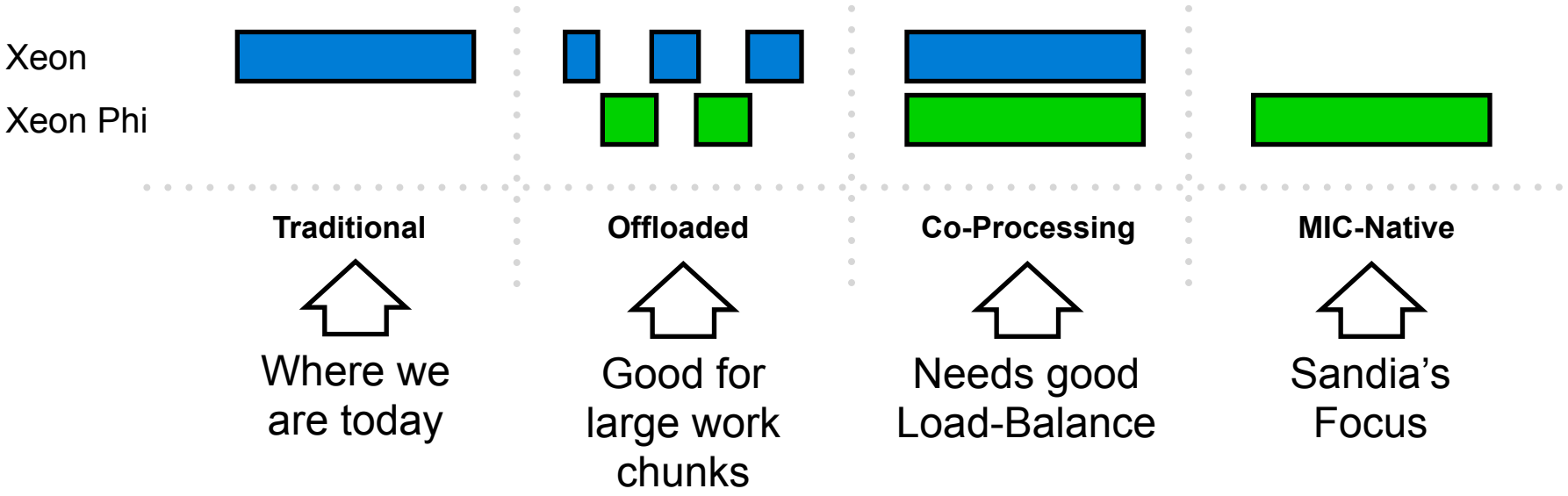
- Cards used in Compton are pre-production B0/B1 cards
  - Intel will offer the 5110P (Q4'12) and 3100 (Q2'13)
- Specifications:
  - 57-cores, 4-way Hyper-threaded = 228 threads @ 1.1GHz
  - 6GB GDDR-5 memory running peak of approx. 240GB/s
  - 1 TFLOP/s of double precision peak
    - Intel have demonstrated a KNC running LINPACK at 1 TFLOP/s
  - 225 - 300W TDP (includes memory, PCIe I/O, co-processor *etc*)
- Dual cards per node, one host Sandy Bridge E5 per Knights Corner (highest PCIe bandwidth)

Programming Knights Corner

Sandia National Laboratories

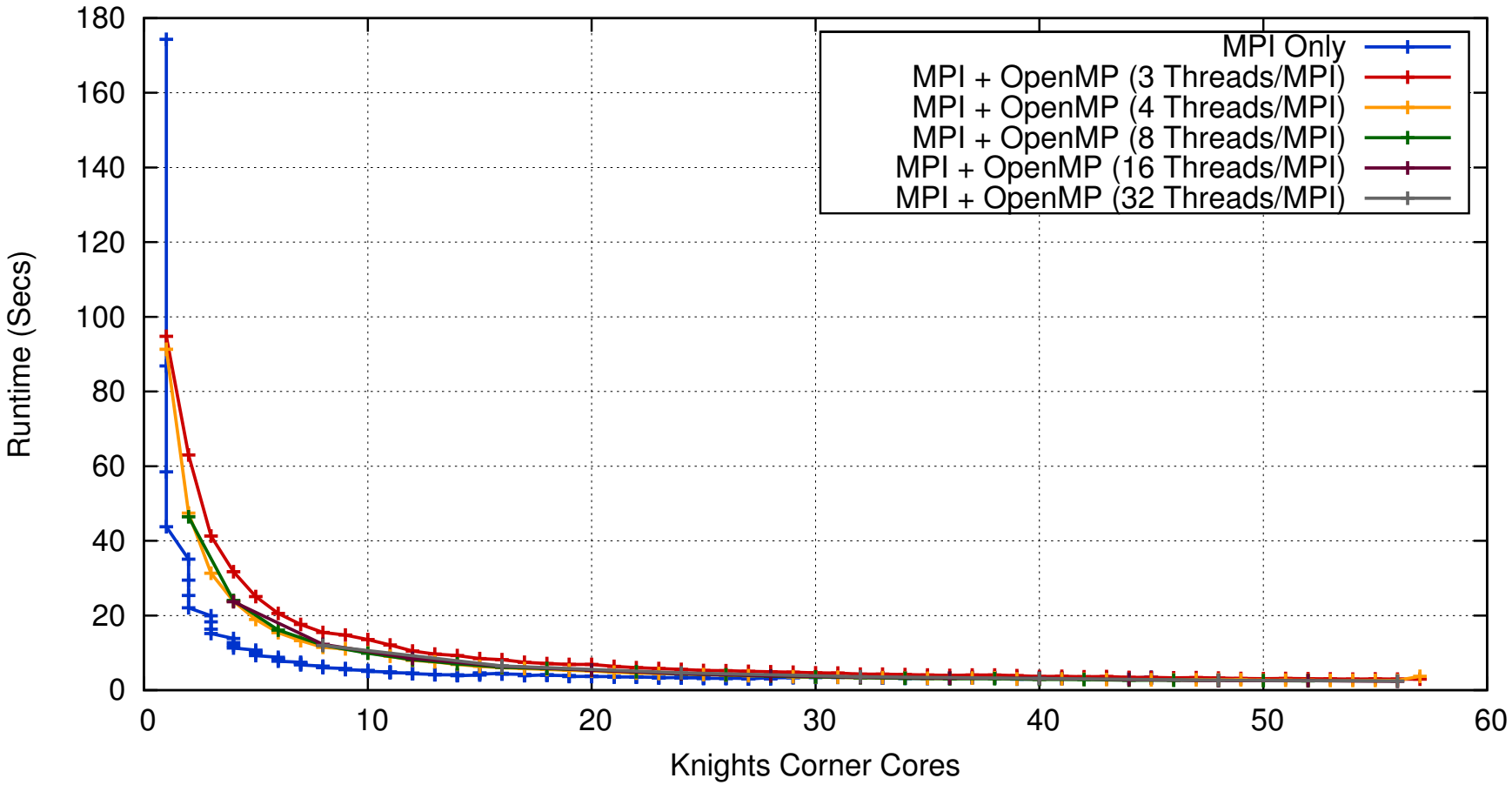
- Knights Corner has been designed to be compatible with traditional Xeon host processors (in so far as programming)
  - This does not mean the same code is always performant
  - This does not imply binary compatibility
  - General rule is code which performs well on Xeon Phi works well on Xeon, but not necessarily the inverse
- Programming models:
  - Threads, threads and more threads!
    - Intel Cilk Plus, Intel Thread Building Blocks, pthreads, OpenMP
  - MPI
    - On-card and between cards

Where to run what?

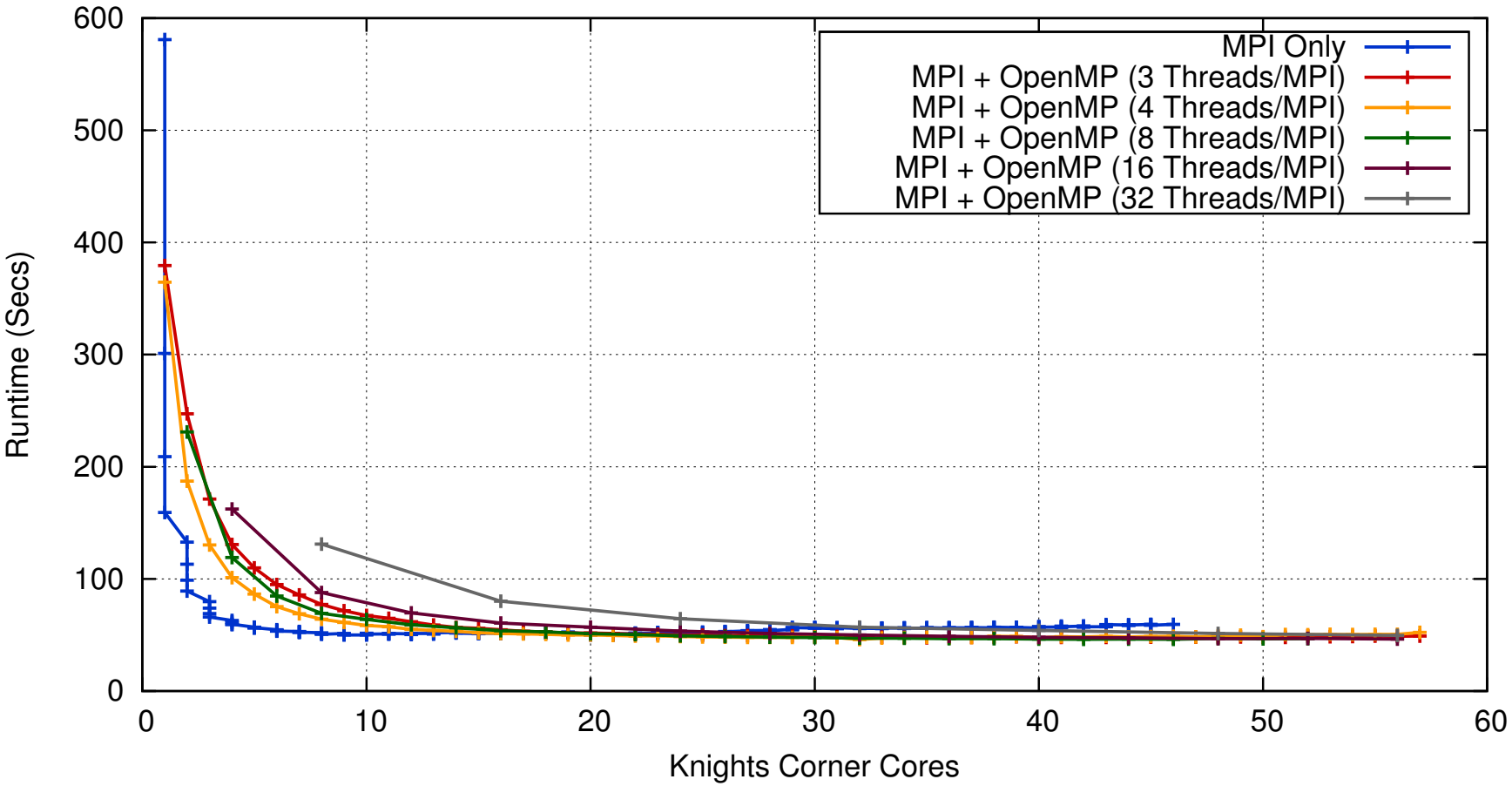


- Knights Corner functions as a node (think node within a node)
  - SSH directly to card, scp files etc (full TCP stack)
  - Run `main(int argc, char* argv[])` directly on card
  - No need for host interactions (get the Xeon out of the way)

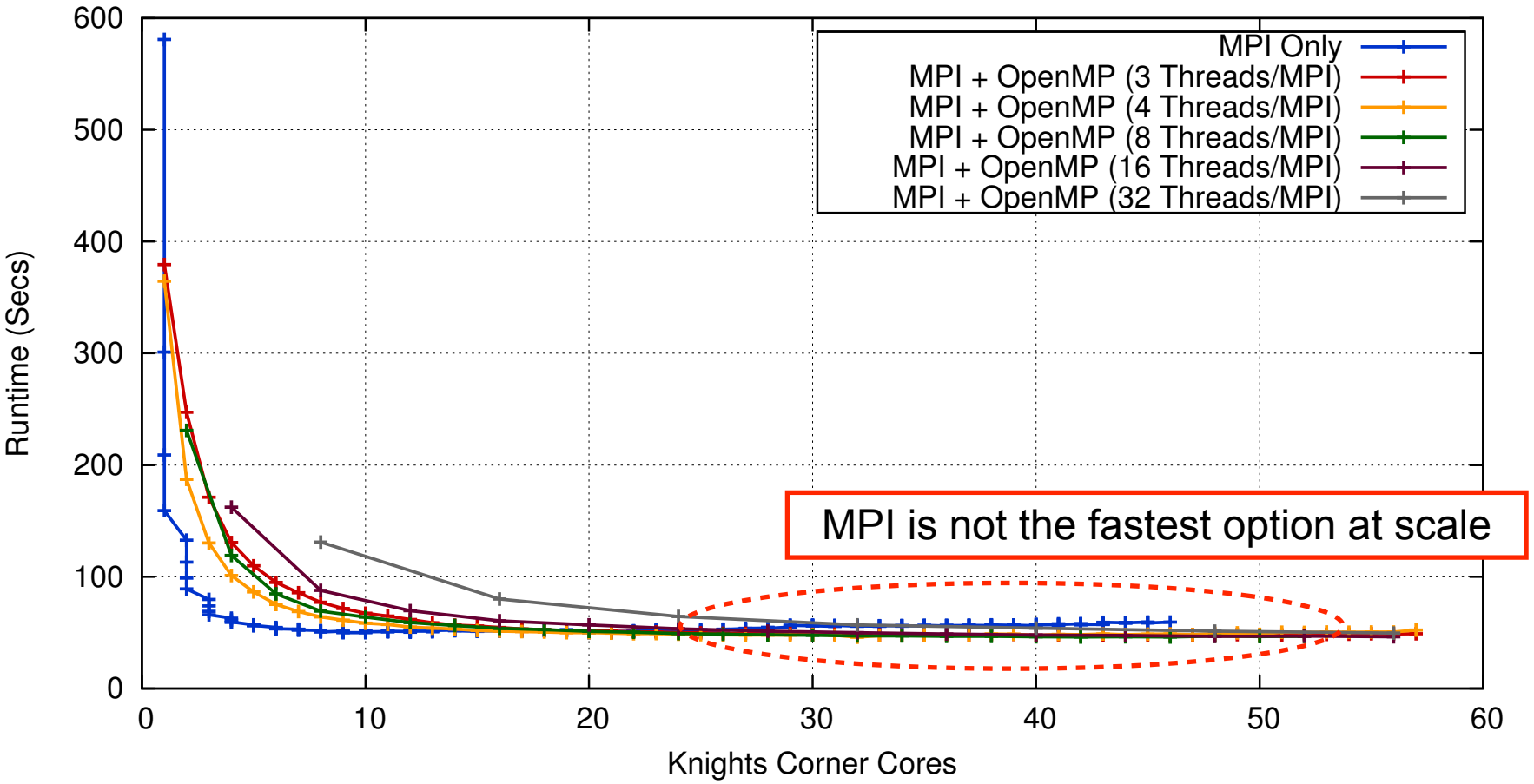
On-Card Scaling for MiniFE-CG



On-Card Scaling for MiniFE



On-Card Scaling for MiniFE

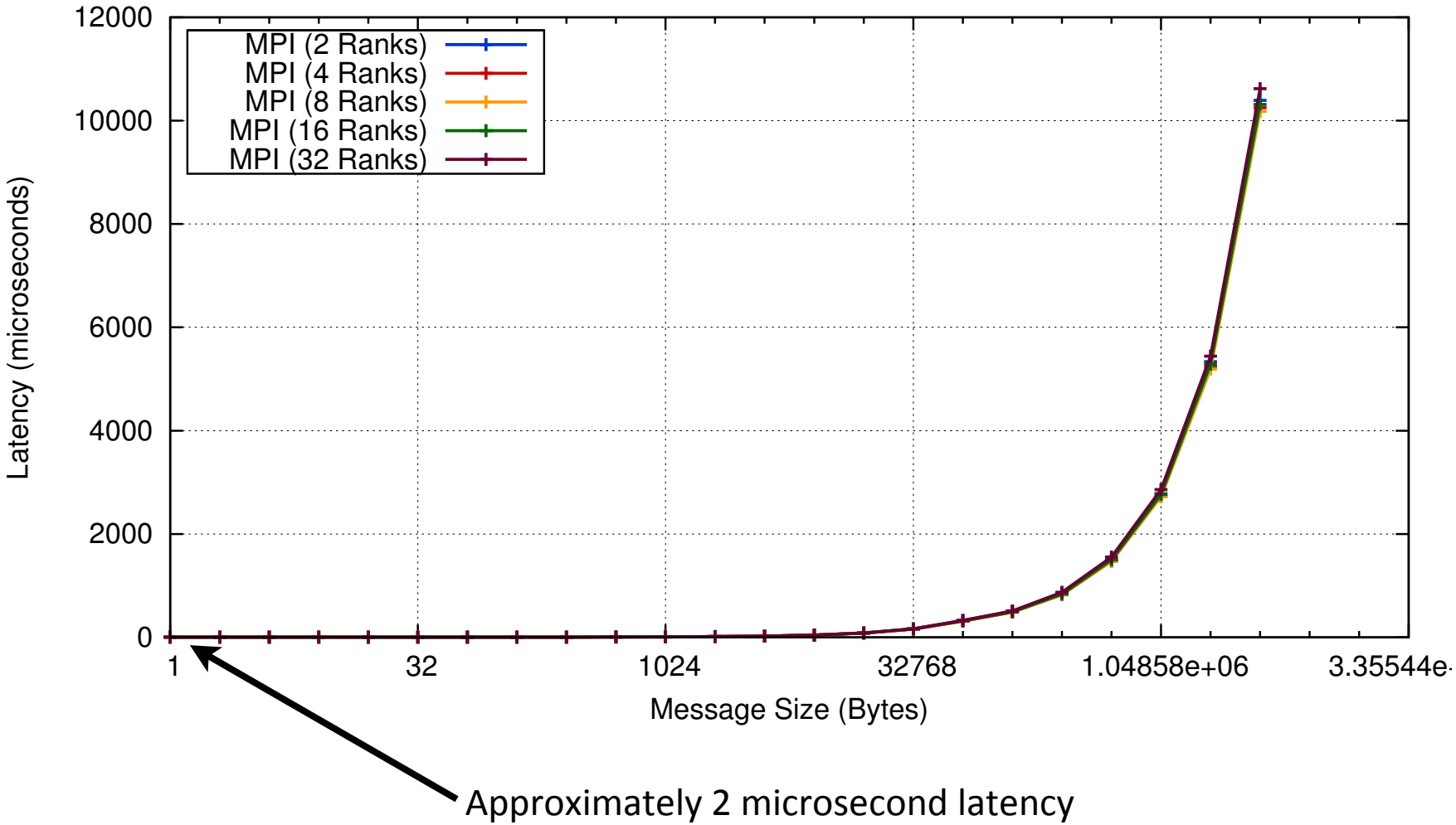


Knights Corner and Sandy Bridge

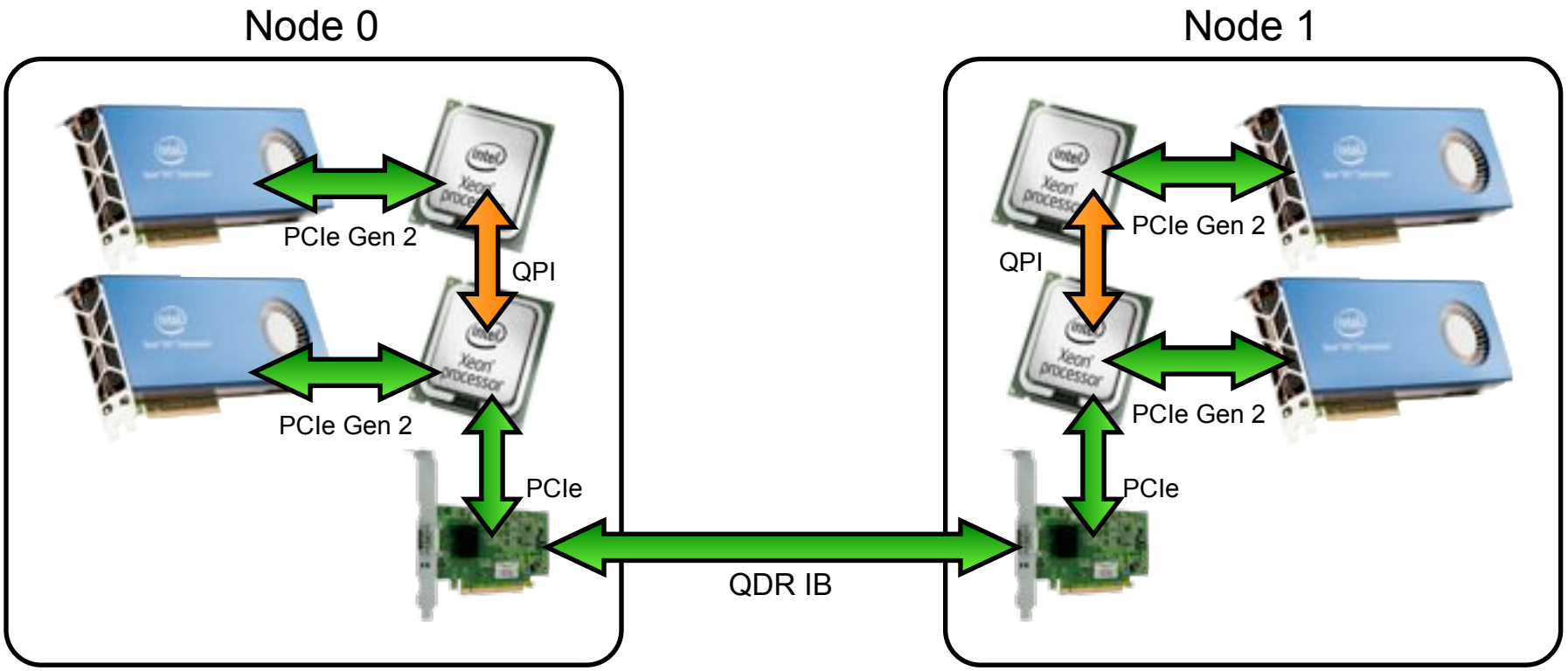
Sandia National Laboratories

- Best CG runtime on a dual-socket Sandy Bridge
  - Classic miniFE, MPI everywhere, 16-ranks
  - 2.38808 seconds for a 200 x 200 x 200 problem
- Best CG runtime on a single Knights Corner
  - Knights Corner ported miniFE, 16 OpenMP Threads per MPI, 14 MPI
  - 2.3503 seconds for a 200 x 200 x 200 problem
- Runtimes roughly equal
  - 225 - 300W TDP including memory for Knights Corner
  - ~300W TDP for dual-socket Sandy Bridge including memory

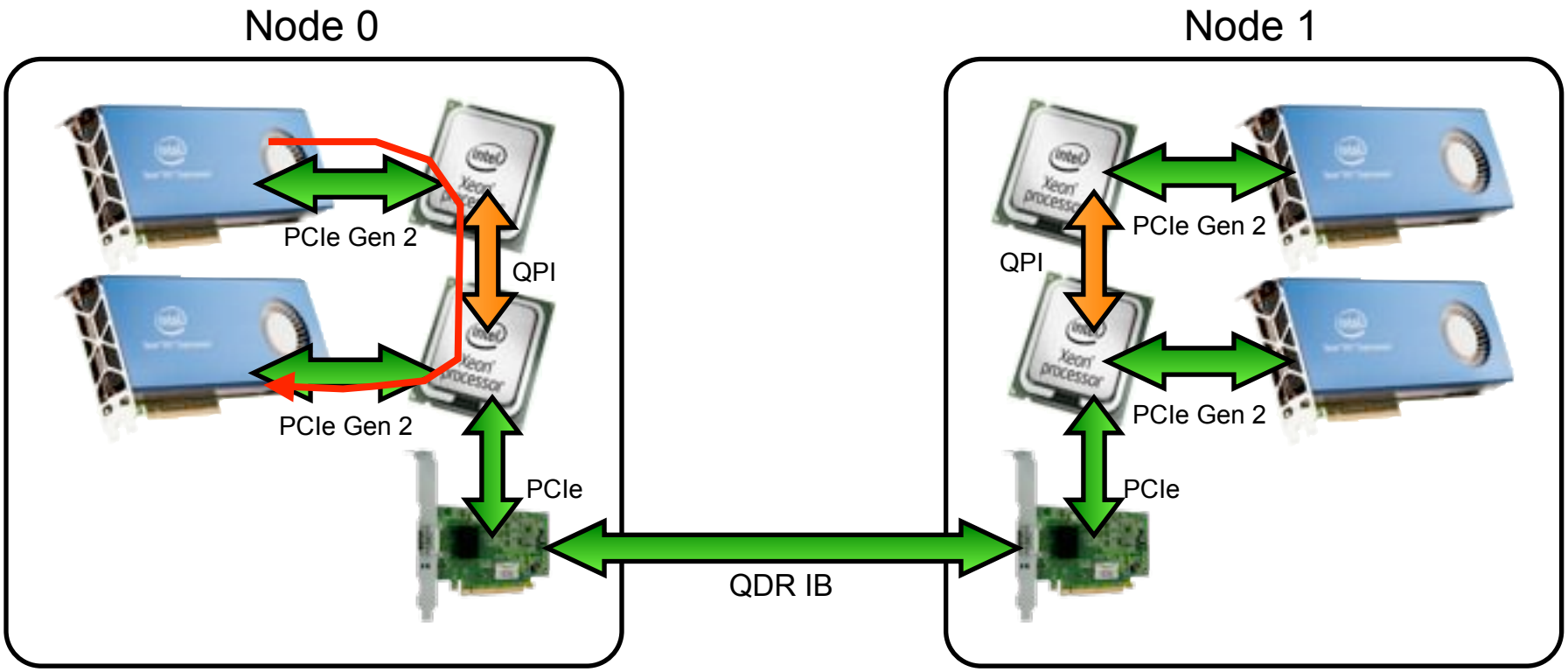
On-Card MPI Performance



Connecting Multiple Cards

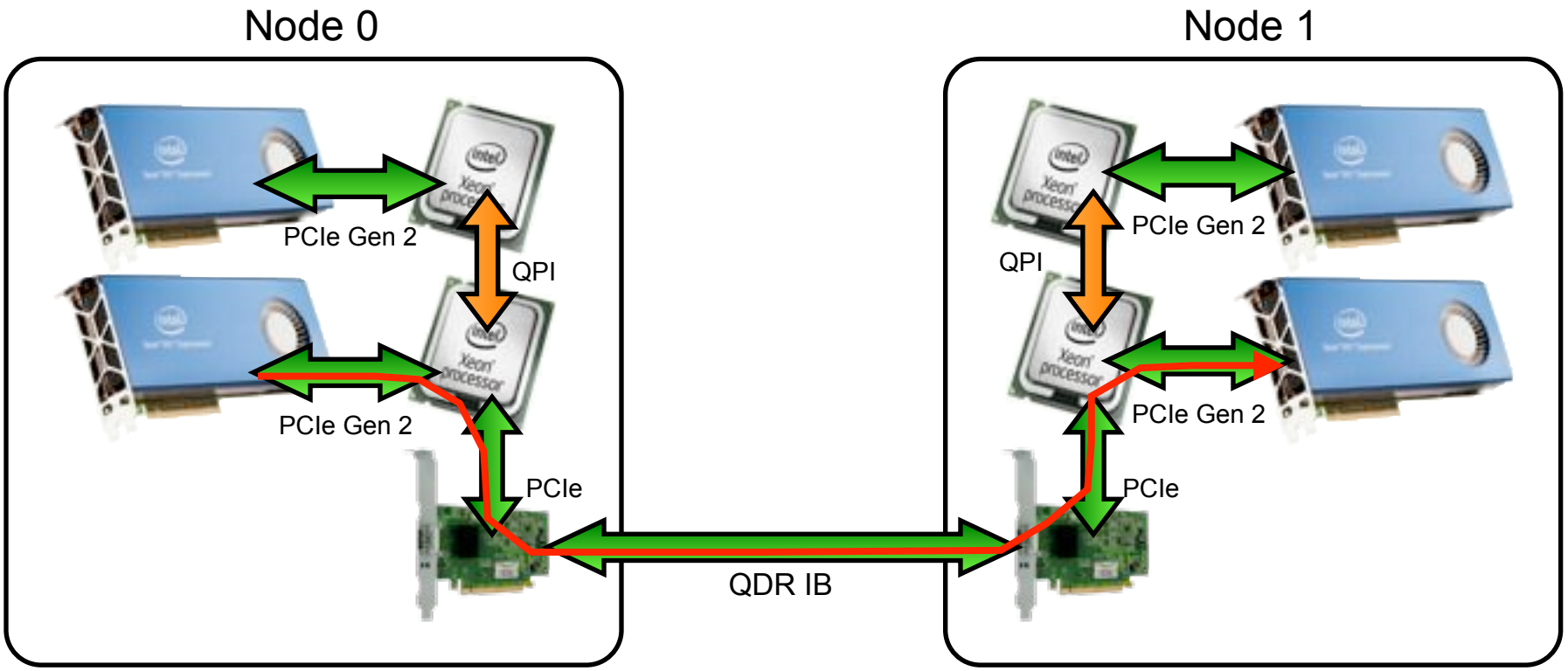


Connecting Multiple Cards



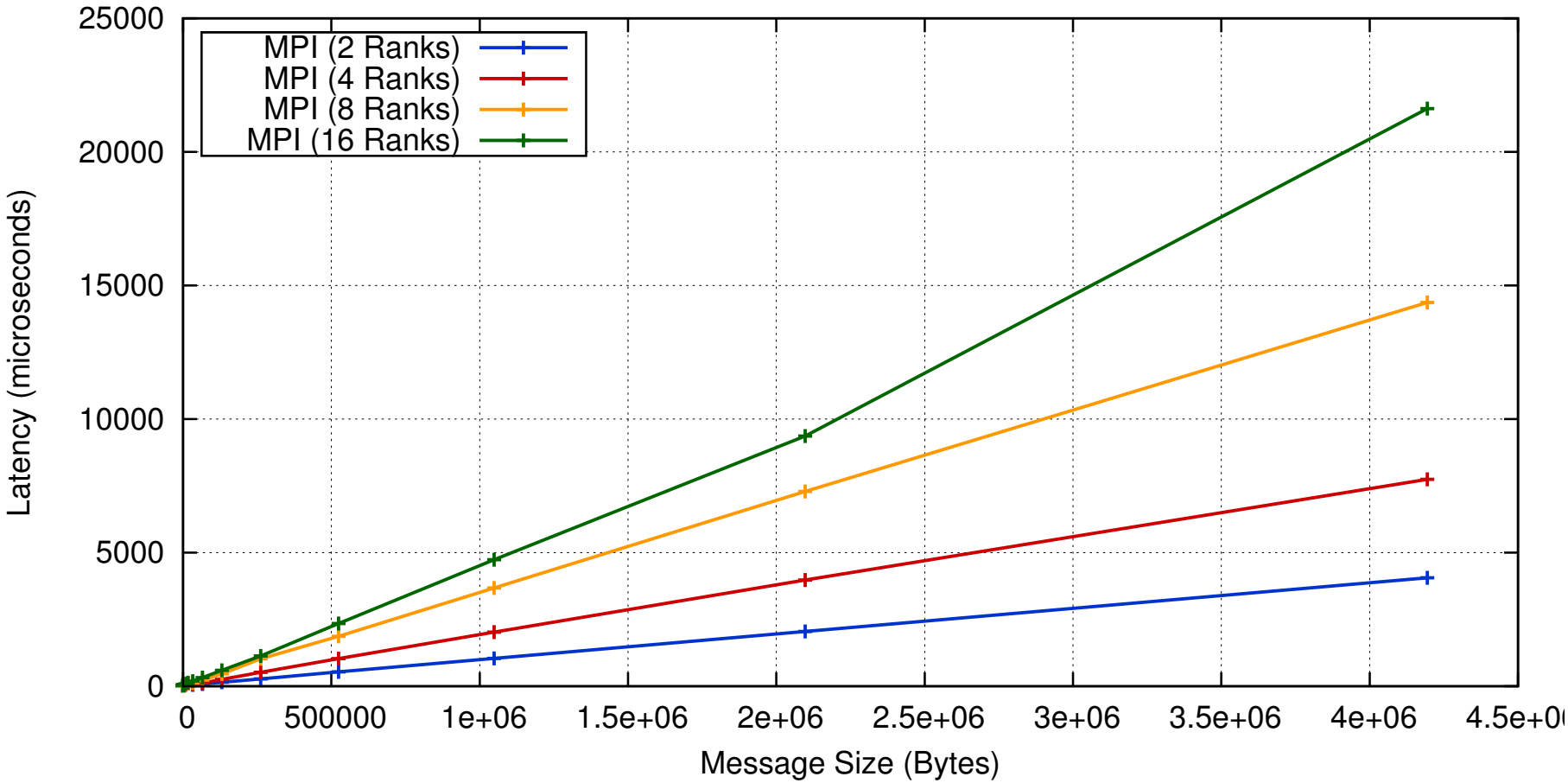
Similar to NVIDIA's GPU Direct concept

Connecting Multiple Cards

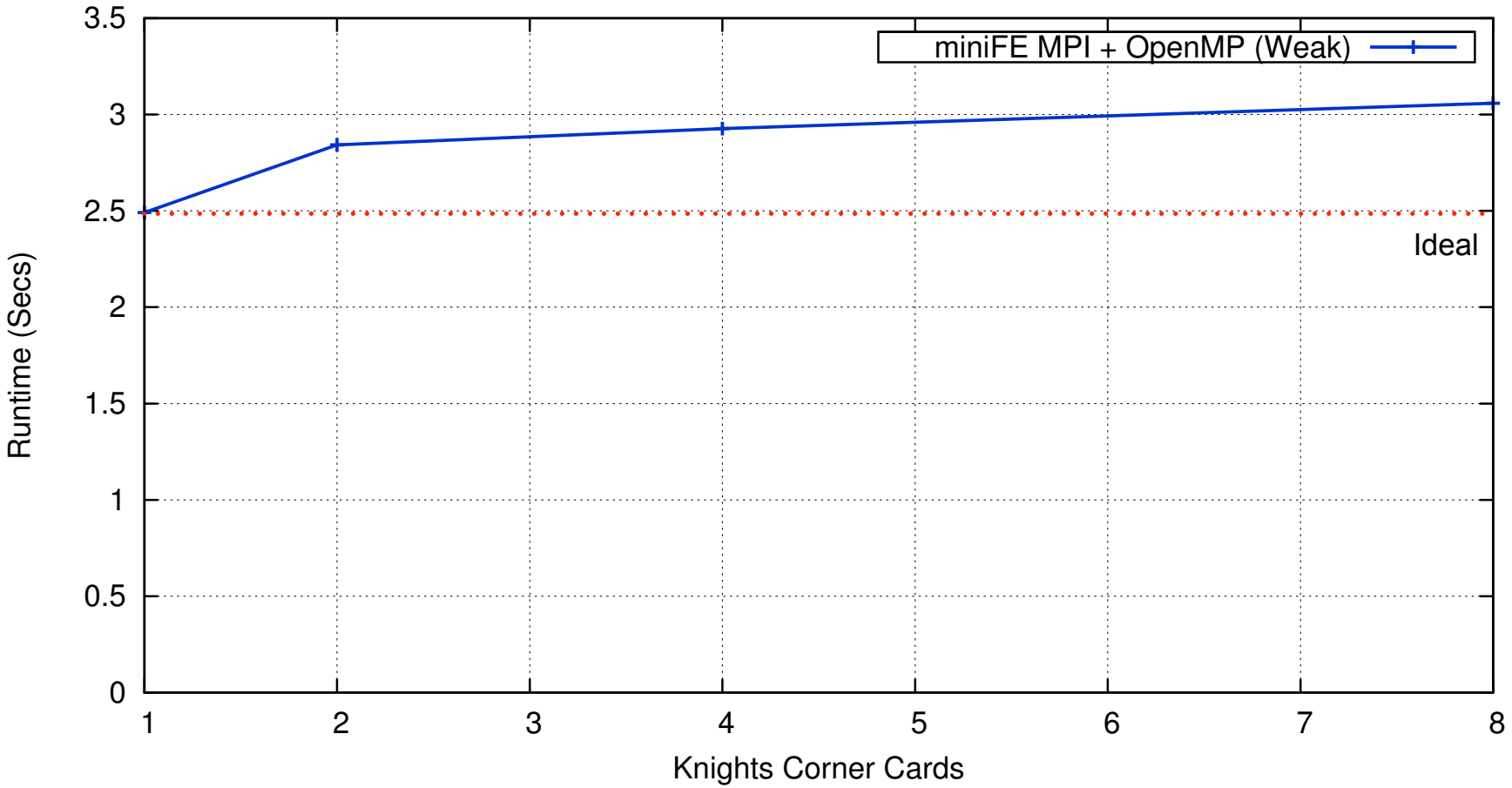


Similar to NVIDIA's GPU Direct concept

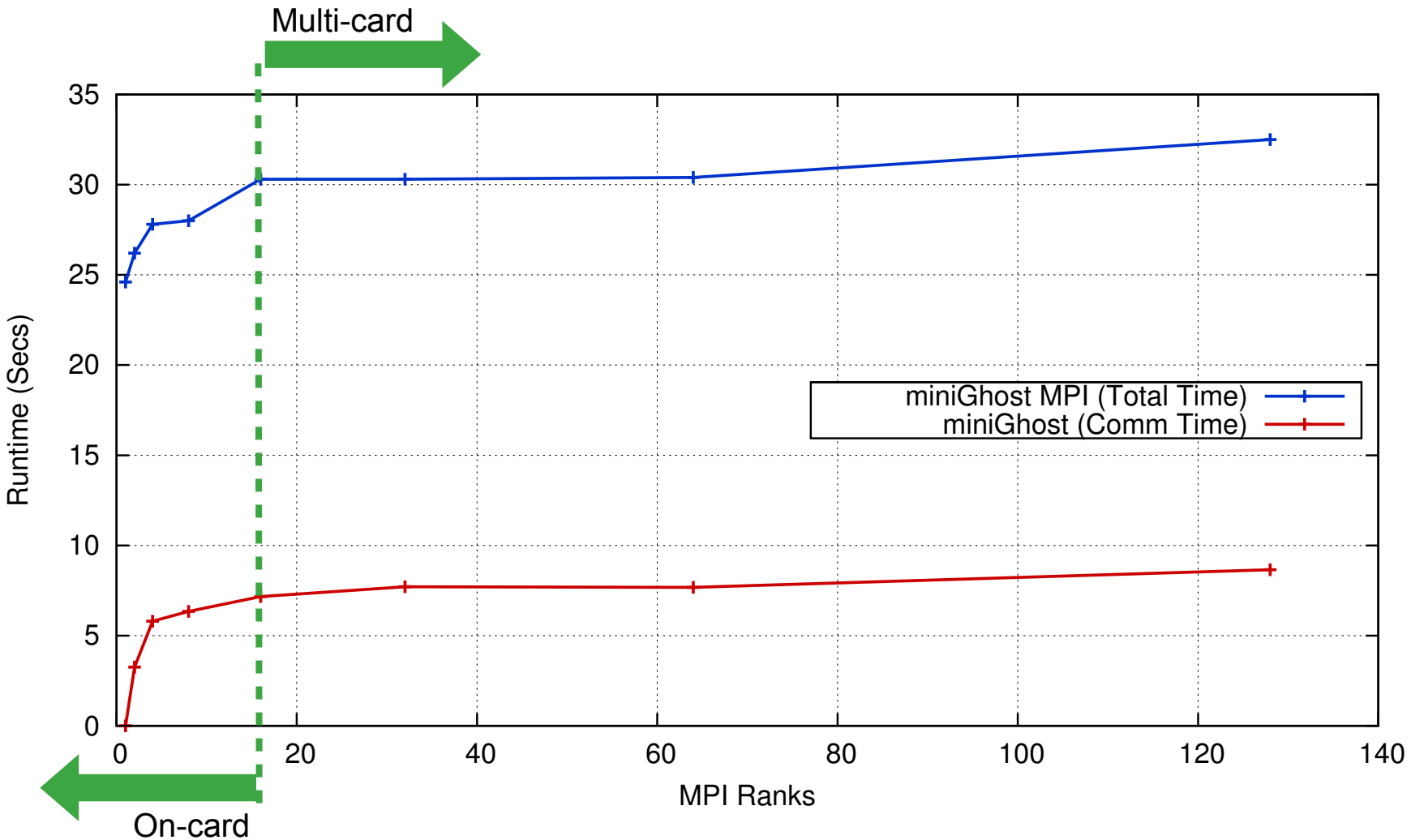
Off-card MPI Performance

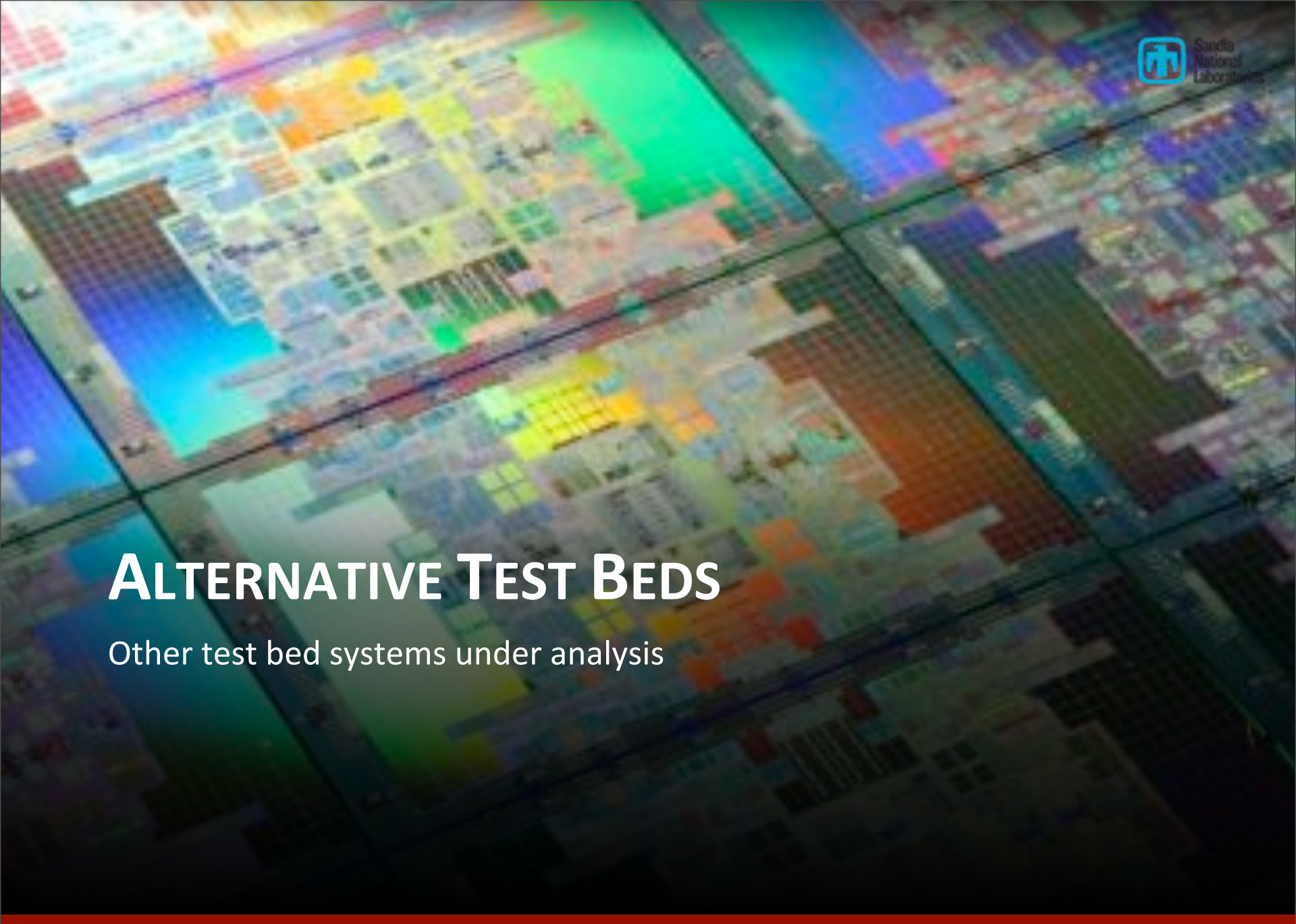


Multi-card Scaling (MiniFE)



On-Card Scaling for MiniGhost





Alternative Test Beds

Other test bed systems under analysis

What other systems are there?

Sandia National Laboratories

- Calxeda
  - ARM-based quad-core processor
  - Proprietary on-board interconnect
  - Extremely low power (< 5 Watts for processor, memory and NIC)
  - OpenMP + MPI?
- NVIDIA GPUs
  - Shannon cluster and upgraded Cray XK6 due 2013
  - CUDA, OpenACC?
- IBM POWER
  - Heavy throughput cores with high memory bandwidth
  - OpenMP + MPI (+ intrinsics?)

THE ON-RAMP TO EXASCALE?

What are we learning about the future?

Snapshot of Technologies

Sandia National Laboratories

	MPI	OpenMP	OpenCL	OpenACC	Intrinsics	CUDA	TBB	Cilk+	ArBB/ CEAN	pthreads	qthreads	MKL/ Math Lib.	Adv. Lang.	DSLs
NVIDIA (GPU)		?												?
AMD (CPU)														?
AMD (APU)					?									?
Intel (CPU)														?
Intel (MIC)														?
IBM (BG)							?							?
ARM														?

miniFE														
miniMD														
miniGhost														
LULESH														
S3D														

Long Road to Exascale

Sandia National Laboratories

- We all would like a "happy Exascale future"
  - 1000x increasing in application performance without changing our algorithms (or even our code)
  - Few people (if any) believe this is possible in the foreseeable future

- What can we do to lessen the load?
  - **Evaluate platforms early** - find out what the characteristics are that we like, what are going to be problems
  - **Working with vendors to explore alternatives** - vendors can tune many different ways, we can be effective if we know what we want
  - **Improve system software, compilers and programming models from the beginning** - actively engage in the development for our needs

Our successes?

Sandia National Laboratories

- Sandia was the first laboratory to engage in Intel's MIC Dungeon program
  - Explored performance issues on Knights Ferry and Corner
  - Specific issues relating to poor compilation experience
    - Has led to specific changes in the Intel compiler for templated C++
  - Discussions regarding MPI and how we can use multiple cards
  - To be repeated in January 2013
- AMD, Sandia and Oak Ridge (Keeneland) are the first laboratories to explore zero-copy for APUs within HPC
  - Investigating driver performance
  - Memory limits (currently absolute maximum is 32-bit memory limit)
  - Programming models
    - Directly involved with AMD on HSA and OpenACC tools

Take-aways

Sandia National Laboratories

- Getting today's applications to run on tomorrows hardware is going to be challenging
  - New programming models
  - Shift in bottlenecks
  - Need for extreme parallelism (which may not be present in algorithms)
  - Slow serial performance

- Test-beds combined with mini-apps and simulators give us a chance to see what characteristics may be issues in the future
  - Performance
  - Programmability
  - Portability

Take-aways

Sandia National Laboratories

- Sandia is leading the DoE/NNSA in analyzing future systems
  - High degree of vendor interaction
  - Gaining feedback on applications and hardware
  - Collaborating with others to make systems useful

- Experiences:
  - Finding parallelism is the key
    - Many cases parallelism is portable across architectures
    - Vectorization is a big issue regardless of platform
  - Software stacks are not mature
    - Debuggers, profilers, libraries will need a lot of effort
    - Programming models are a big question

Activities Under Way

Sandia National Laboratories

- Benchmarking of energy use
  - Li Tang and Sharon Hu (Notre Dame)
  - Teller test-bed and AMD FastForward
- New programming languages and models
  - qthreads and SPR for XGC
  - MiniFE Chapel
- Advanced program transformations
  - Jagan Jayaraj and Paul Woodward (Univ. of Minnesota)
- Modeling using Sandia's Structural Simulation Toolkit
- Investigating mini-app performance with ARM
- Lakeside Code Tools - vectorization, bit flips and memory

Want to join in?

Sandia National Laboratories

- Be part of Mantevo
  - Having mini-apps helps us engage vendors
  - Opportunities for collaboration
  - Mini-apps are *the* enabling tool to help us understand the hardware
- Systems are open on the ECN
  - Get an account through WebCARS today
  - Encourage your collaborators to be a part of the test bed project
- Suggestions for future systems?
- What tools do you want to see in the future?

Sandia National Laboratories

Exceptional service in the national interest

sdhammo@sandia.gov