International HPC Summer School on HPC Challenges in Computational Sciences 2016

The Future of HPC through Driving Challenges and Enabling Opportunities

Thomas Sterling

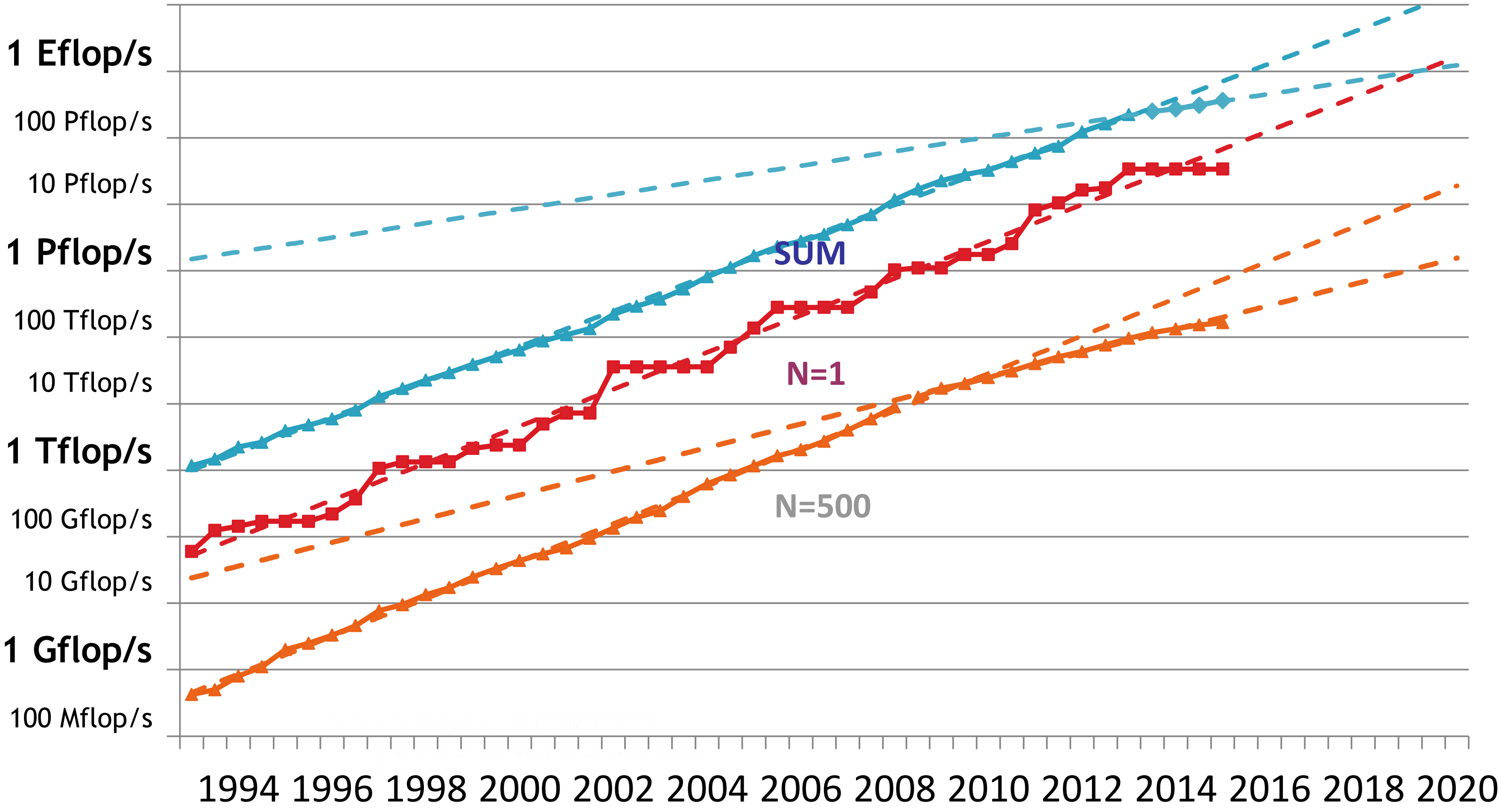
Professor, Informatics and Computing

Center for Research in Extreme Scale Technologies

Indiana University

June 27, 2016

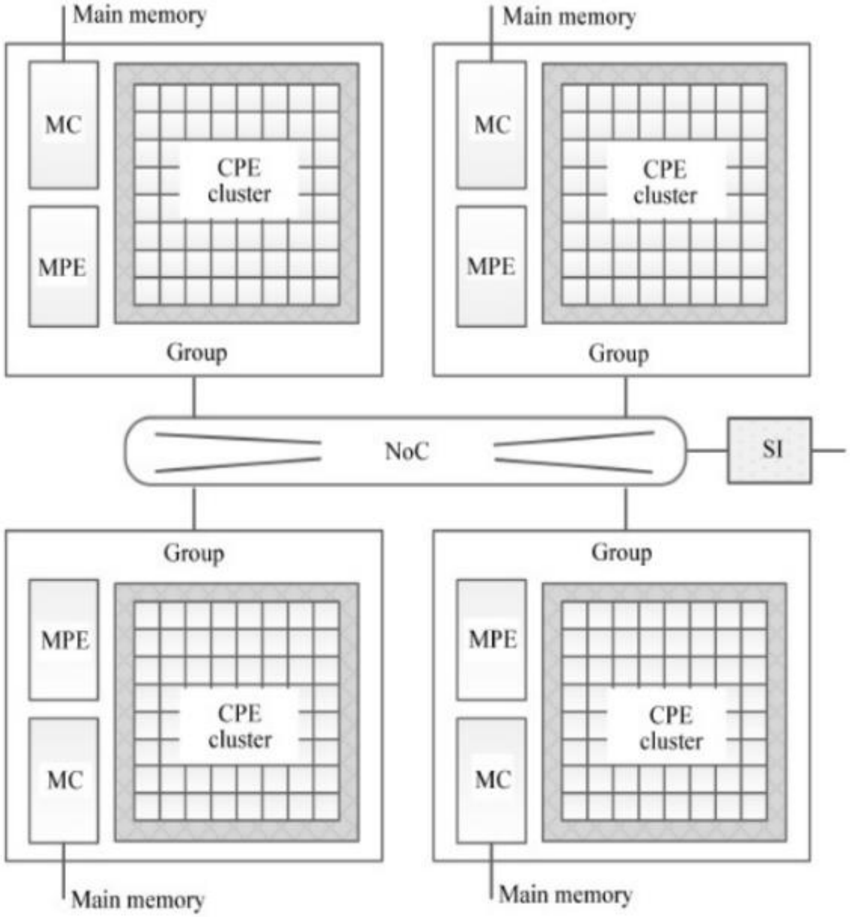
Projected Performance Development



Sunway TaihuLight

- Peak: 125 Petaflops
- Cores: 10 Million, SW26010
- Linpack: 93 Petaflops, 73% eff.
- Clock: 1.45 GHz
- Memory: 1.3 Petabytes
- Power: 15.4 Megawatts
- Located: National Supercomputing Center in Wuxi
- Vendor: NRCPC

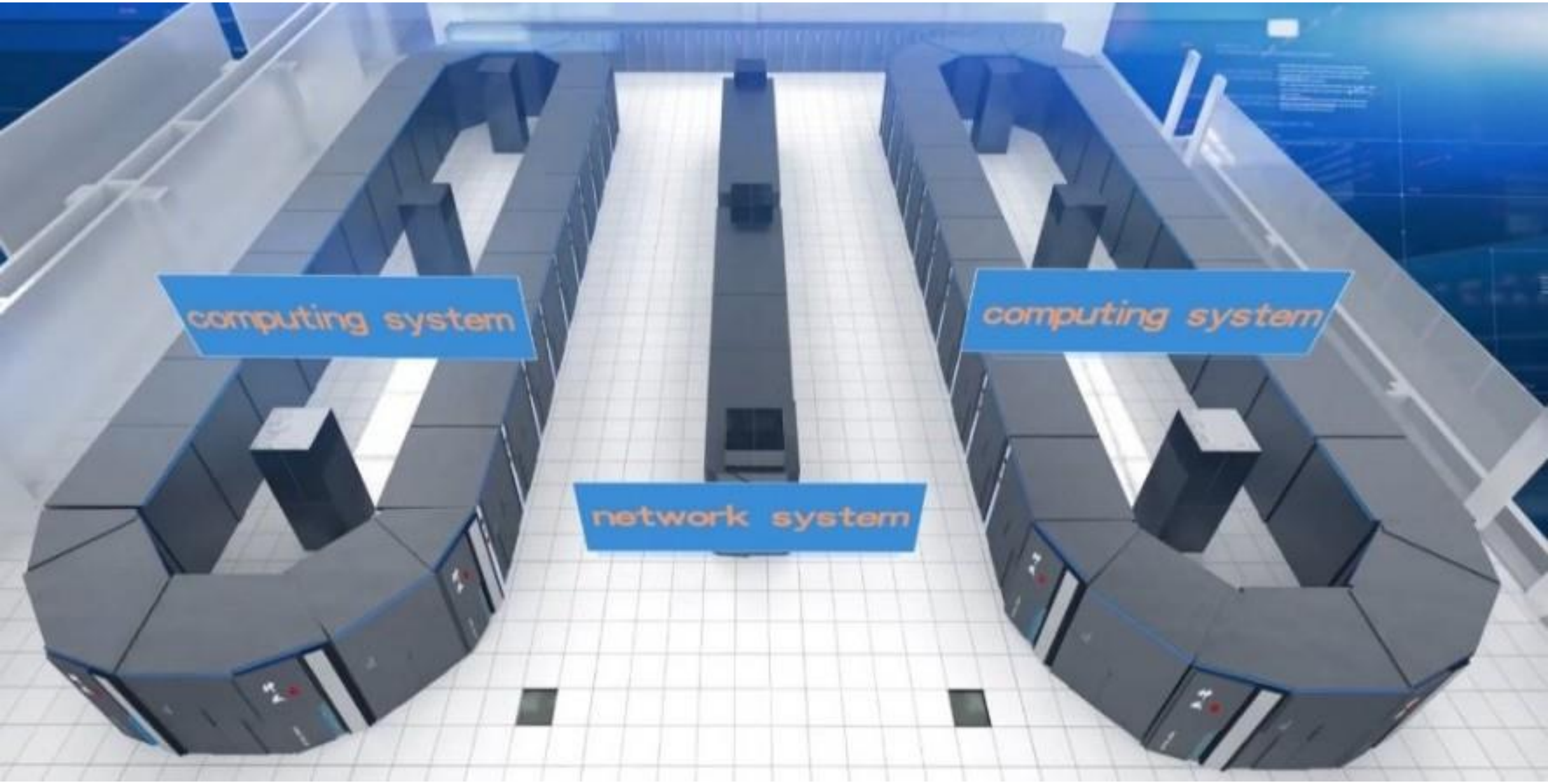
Node Architecture



Source: HPCwire

- 40,960 nodes
  - System Interface – PCIe, 16 GBps
- Node of 4 core groups
  - NoC
  - System Interface (SI) to external devices
  - 32 Gbytes of DDR3 memory
- Each group has
  - a cluster of 64 computing processing elements (CPE)
    - RISC SIMD architecture 8 ops/cycle
    - 64-bit floating point
    - 11.6 Gflops
    - 64KByte scratchpad, 16 Kbyte IC
  - 1 management processing element (MPE)
    - 23.2 Gflops
  - 1 memory controller (MC)
  - Its own memory space
- Designed by the Shanghai High Performance IC Design Center

40 cabinets, 3.1 Pflops each

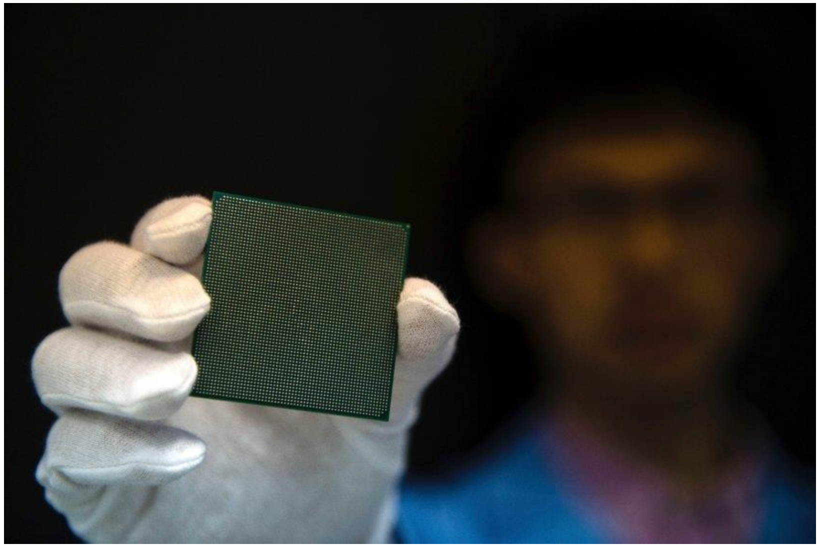


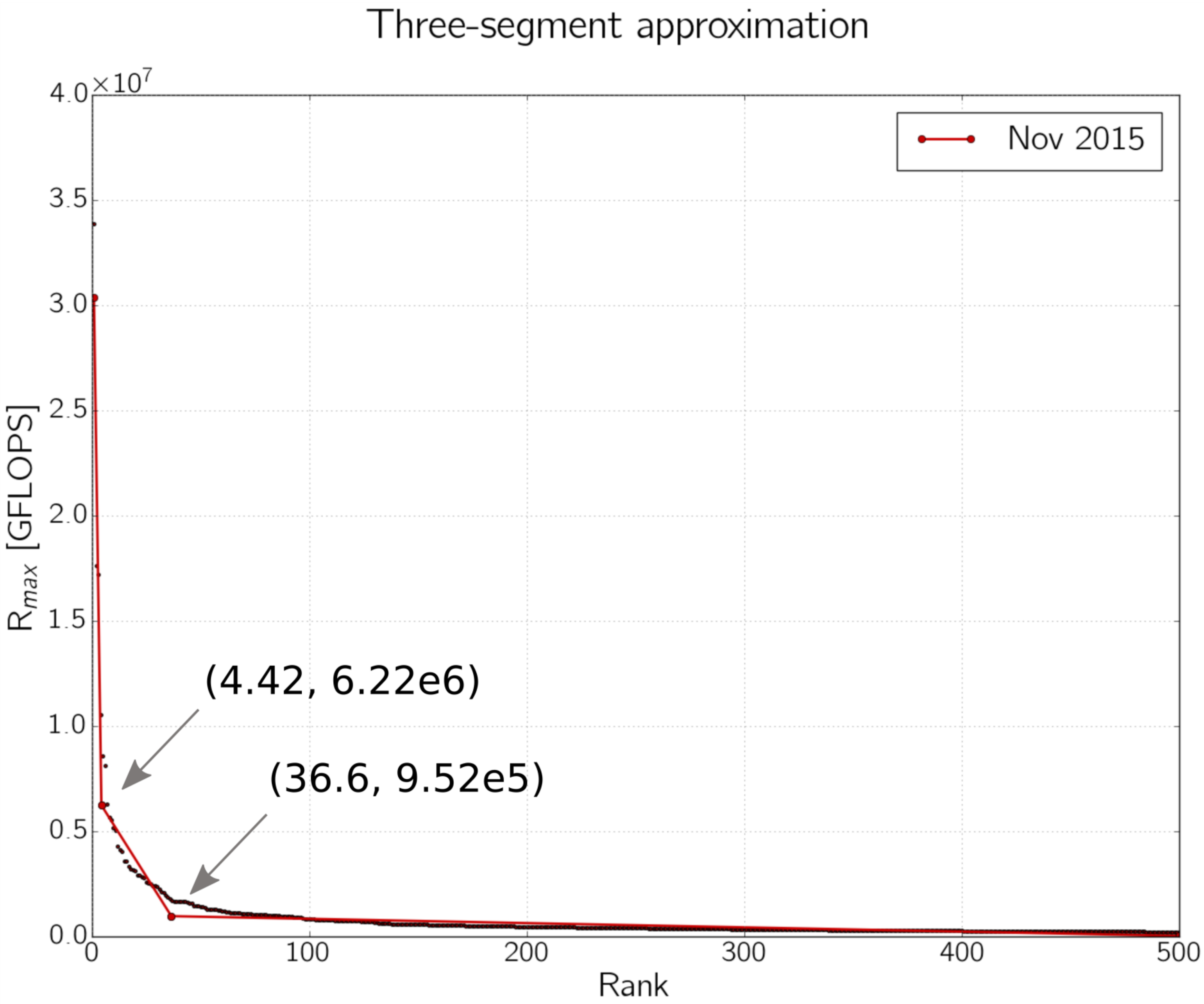
4 super-nodes per cabinet

256 nodes per super-node

Architecture Constraints

- Memory
  - Really lightweight
  - 125 Pflops with only 1.3 Petabytes for a ratio of 100:1 inverse capacity
- Bandwidth
  - 22.4 flops/byte of transfer
- HPCG 0.3% peak
- Cache-less
  - Small instruction cache (12KBytes)
  - Small scratchpad (16KBytes)
- Bi-section Band Width of only 70 Tbytes/sec.
- slow clock rate



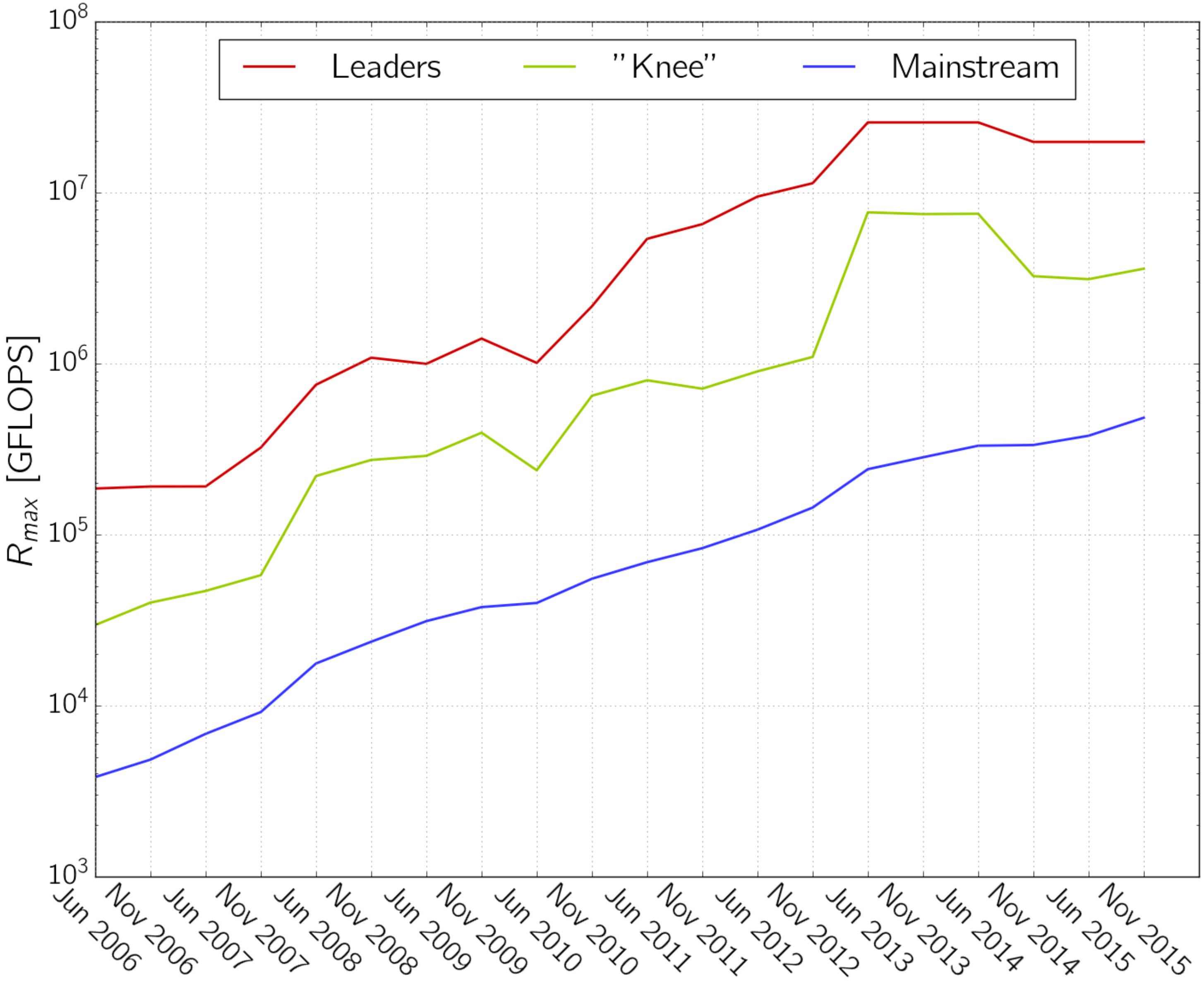


(4.42, 6.22e6)

(36.6, 9.52e5)

Slide courtesy of Maciej Brodowicz, 06/16/2016

Three worlds of supercomputing: average R_{max}



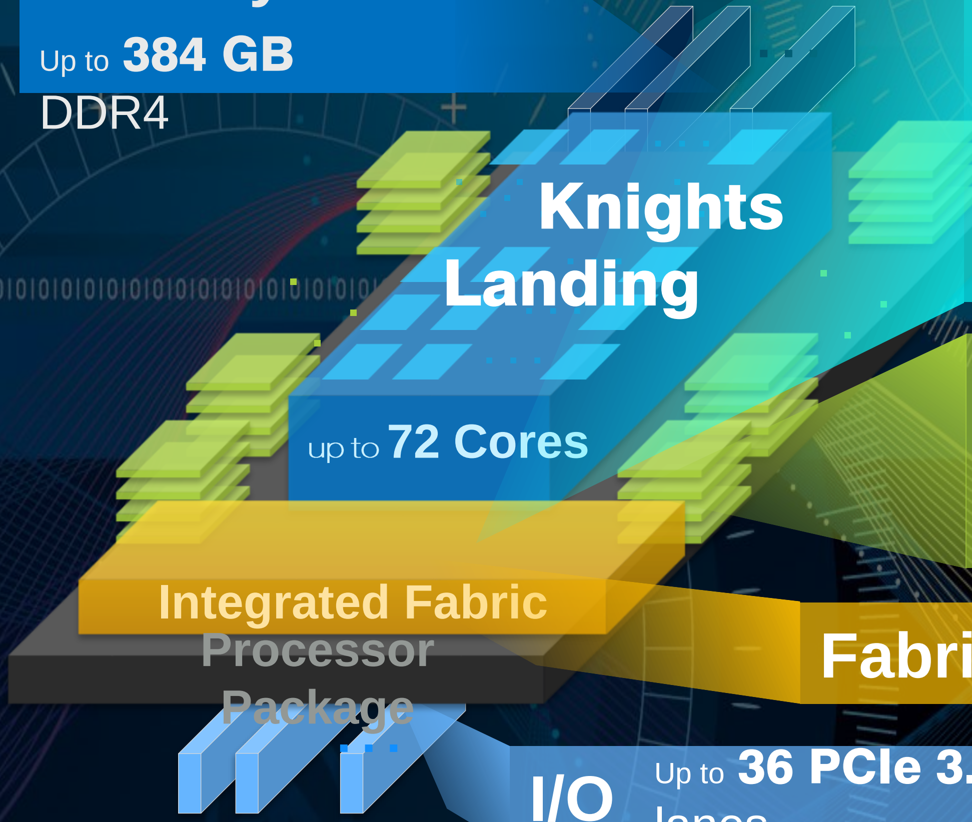
Slide courtesy of Maciej Brodowicz, 06/16/2016

Knights Landing

Next Generation Intel® Xeon Phi™ Product Family

Platform Memory

Up to **384 GB** DDR4



Compute

- Intel® Xeon® Processor Binary-Compa
- **3+ TF**LOPS[1], **3X ST[2]** (single-thread) pe
- **2D Mesh** Architecture
- **Out-of-Order** Cores

On-Package Memory

- Up to **16 GB** *at launch*
- Over **5x STREAM[3]** *vs. DDR4 at launch*

Fabric (optional)

- **1st** Intel processor to integrate

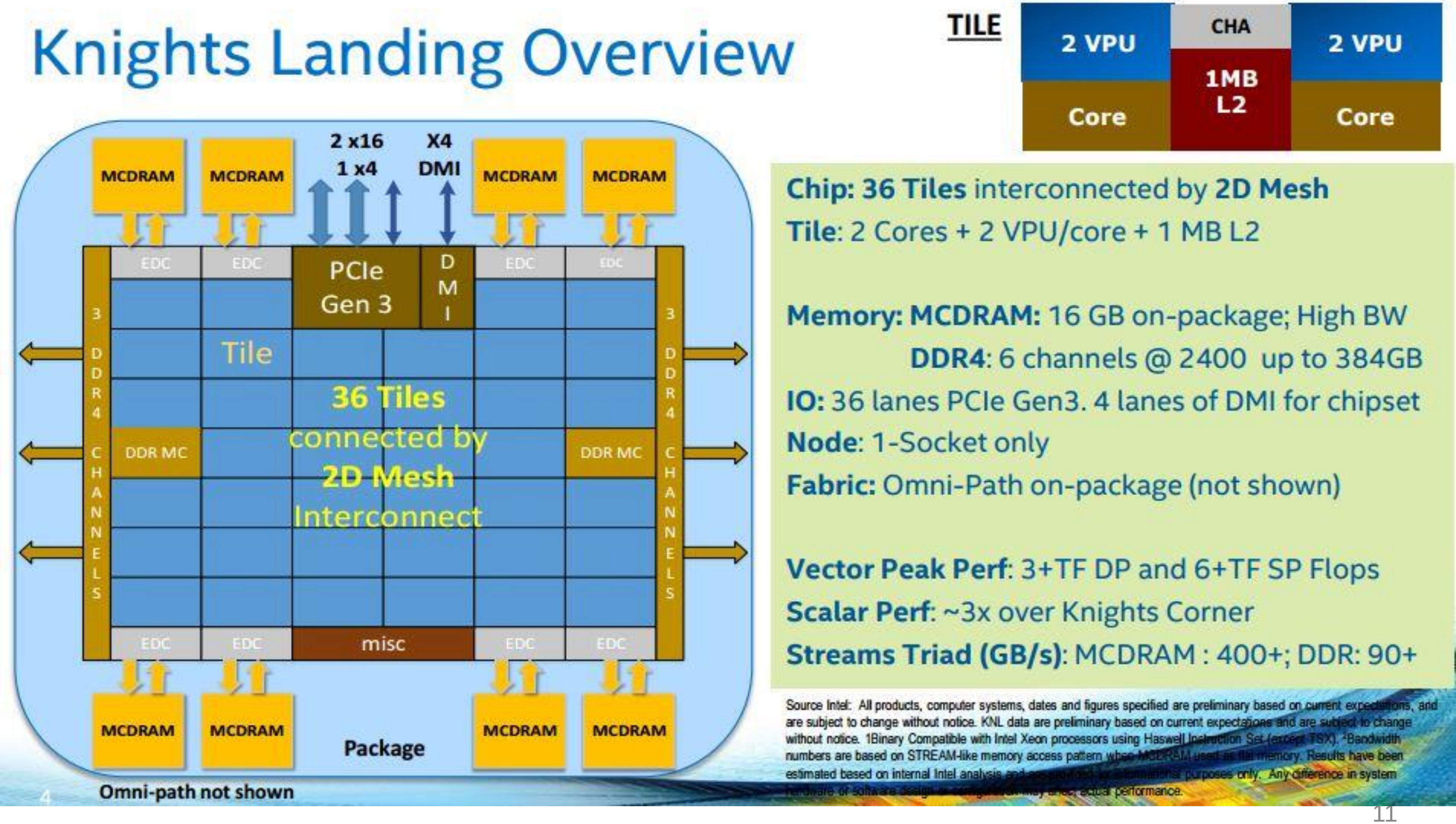
I/O Up to **36 PCIe 3.0** lanes

[1] Over 3 Teraflops of peak theoretical double-precision performance is preliminary and based on current expectations of cores, clock frequency and floating point operations per cycle. FLOPS = cores x clock frequency x floating-point operations per second per cycle. .

[2] Projected peak theoretical single-thread performance relative to 1st Generation Intel® Xeon Phi™ Coprocessor 7120P (formerly codenamed Knights Corner).

[3] Projected result based on internal Intel analysis of STREAM benchmark using a Knights Landing processor with 16GB of ultra high-bandwidth versus DDR4 memory only with all channels populated.

[4] Intel internal estimate



Knights Landing Overview

Chip: 36 Tiles interconnected by **2D Mesh**

Tile: 2 Cores + 2 VPU/core + 1 MB L2

Memory: MCDRAM: 16 GB on-package; High BW

DDR4: 6 channels @ 2400 up to 384GB

IO: 36 lanes PCIe Gen3. 4 lanes of DMI for chipset

Node: 1-Socket only

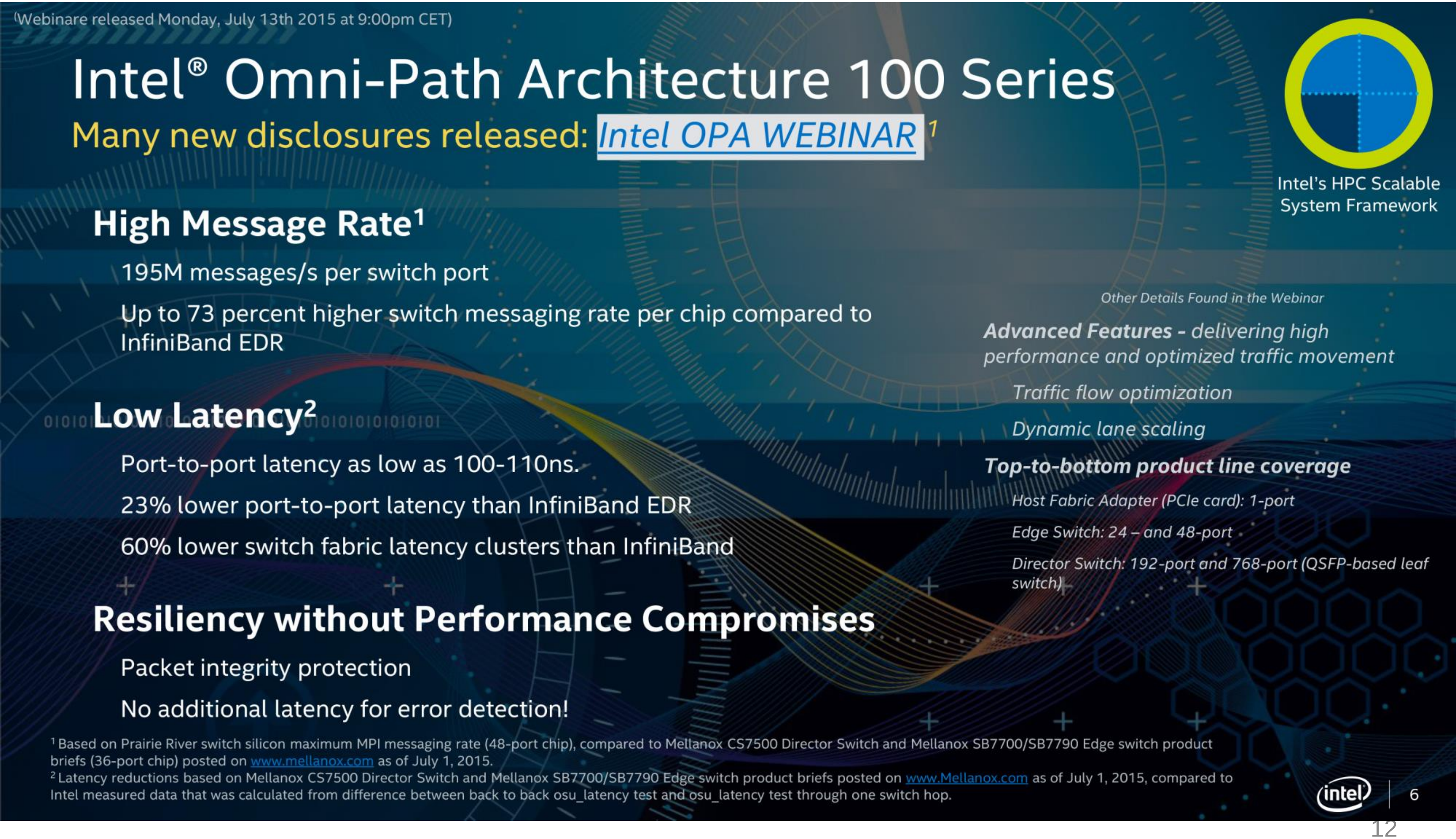
Fabric: Omni-Path on-package (not shown)

Vector Peak Perf: 3+TF DP and 6+TF SP Flops

Scalar Perf: ~3x over Knights Corner

Streams Triad (GB/s): MCDRAM : 400+; DDR: 90+

Source Intel: All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. KNL data are preliminary based on current expectations and are subject to change without notice. 1Binary Compatible with Intel Xeon processors using Haswell Instruction Set (except TSX). ²Bandwidth numbers are based on STREAM-like memory access pattern when MCDRAM used as flat memory. Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.



(Webinare released Monday, July 13th 2015 at 9:00pm CET)

Intel® Omni-Path Architecture 100 Series

Many new disclosures released: Intel OPA WEBINAR [1]

Intel's HPC Scalable System Framework

High Message Rate[1]

- 195M messages/s per switch port
- Up to 73 percent higher switch messaging rate per chip compared to InfiniBand EDR

Low Latency[2]

- Port-to-port latency as low as 100-110ns.
- 23% lower port-to-port latency than InfiniBand EDR
- 60% lower switch fabric latency clusters than InfiniBand

Resiliency without Performance Compromises

- Packet integrity protection
- No additional latency for error detection!

Other Details Found in the Webinar

Advanced Features - *delivering high performance and optimized traffic movement*

- *Traffic flow optimization*
- *Dynamic lane scaling*

Top-to-bottom product line coverage

- *Host Fabric Adapter (PCIe card): 1-port*
- *Edge Switch: 24 – and 48-port*
- *Director Switch: 192-port and 768-port (QSFP-based leaf switch)*

[1] Based on Prairie River switch silicon maximum MPI messaging rate (48-port chip), compared to Mellanox CS7500 Director Switch and Mellanox SB7700/SB7790 Edge switch product briefs (36-port chip) posted on www.mellanox.com as of July 1, 2015.

[2] Latency reductions based on Mellanox CS7500 Director Switch and Mellanox SB7700/SB7790 Edge switch product briefs posted on www.Mellanox.com as of July 1, 2015, compared to Intel measured data that was calculated from difference between back to back osu_latency test and osu_latency test through one switch hop.

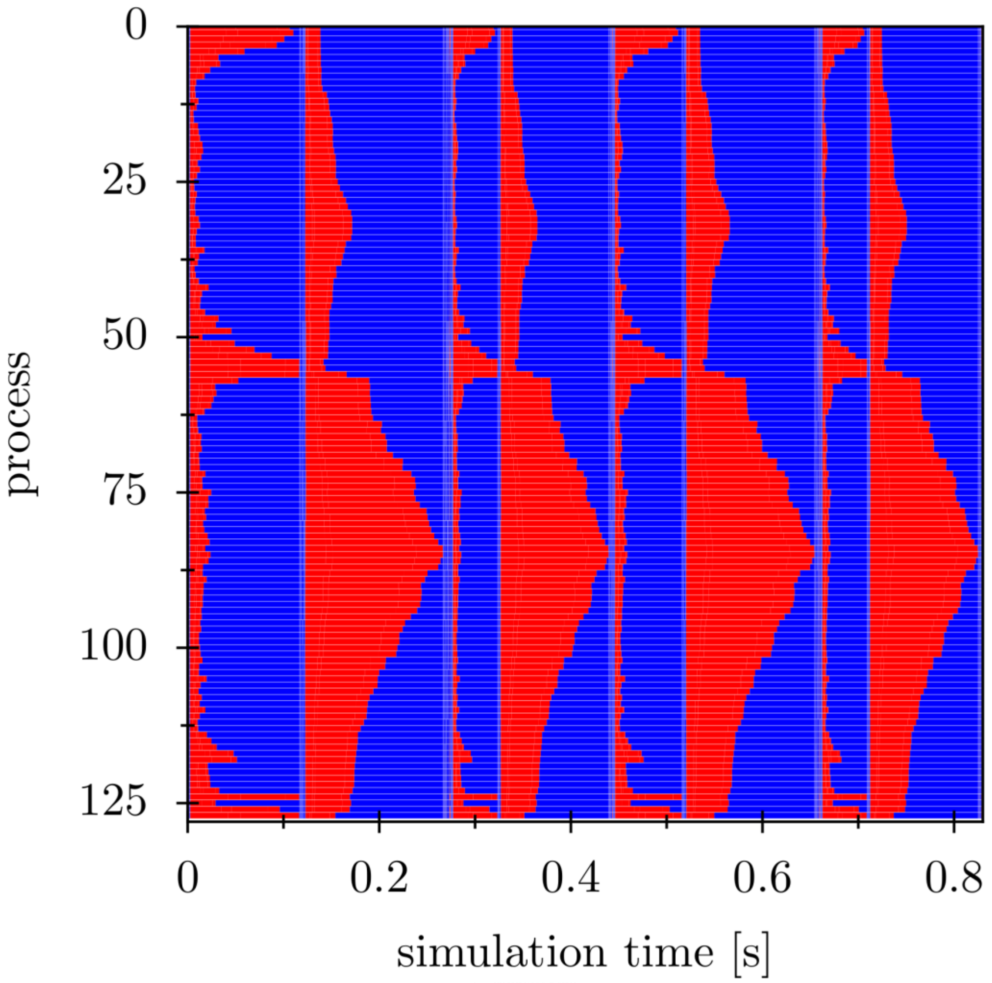
ANL Aurora

Argonne NATIONAL LABORATORY | Leadership Computing Facility

- Cray Shasta architecture
- Over 50,000 nodes
- Peak performance: 180 PFLOPS
- 3rd generation Xeon Phi cores (Knight's Hill)
- Over 7PB of DRAM and persistent system memory
- Intel interconnect based on 2nd generation Omni-Path architecture with silicon photonics
- 150+PB data storage using Lustre with >1TB/s throughput
- 13MW peak power consumption
- Software environment includes MPI+OpenMP 4.x, Intel compilers and optimization tools, and Cray compilers and libraries
- Cost: $200 million
- Located at Argonne Leadership Computing Facility (ALCF)
- Delivery in 2018 with anticipated start of production phase Q2 2019

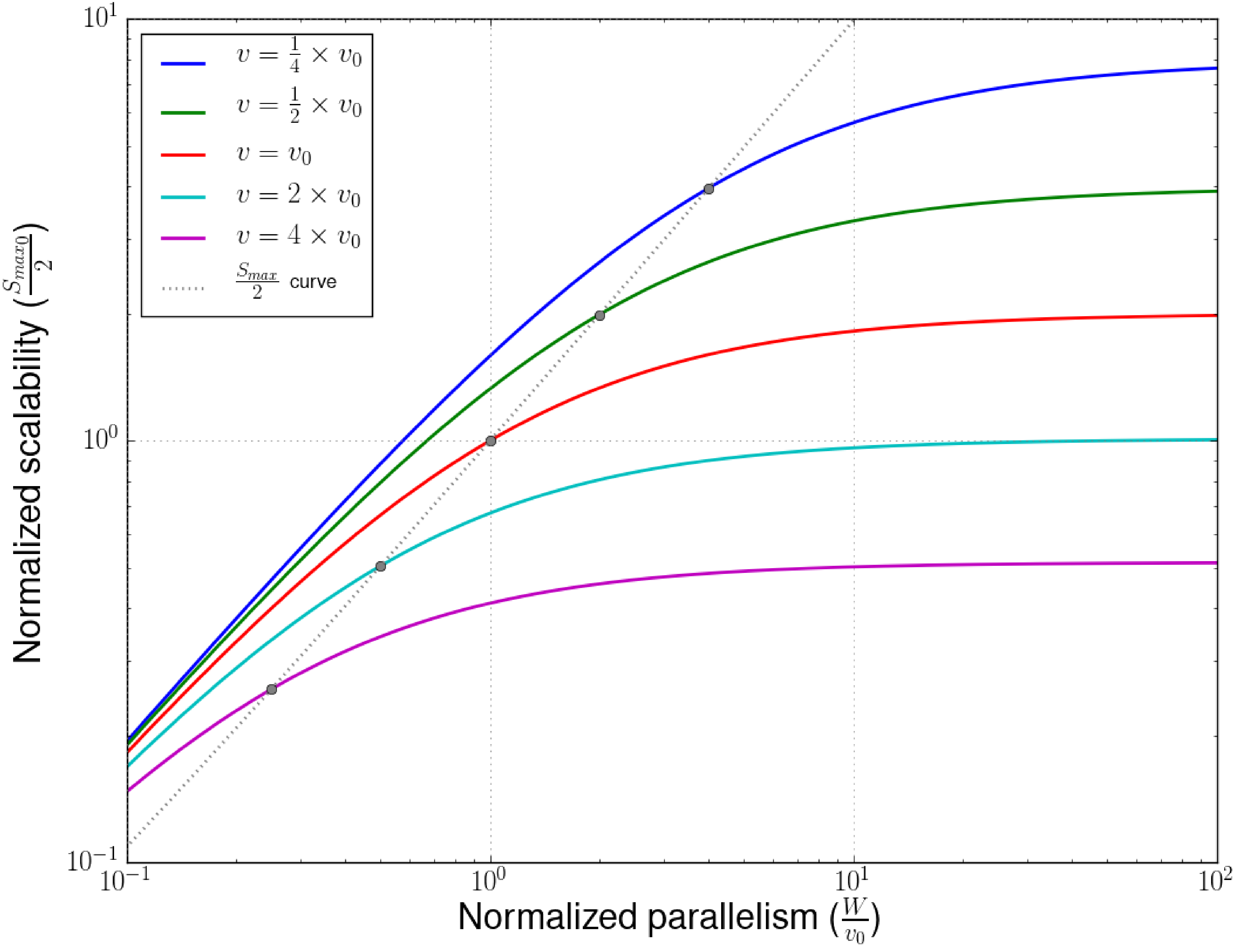
Slide courtesy of Maciej Brodowicz, IU

The Negative Impact of Global Barriers in Astrophysics Codes

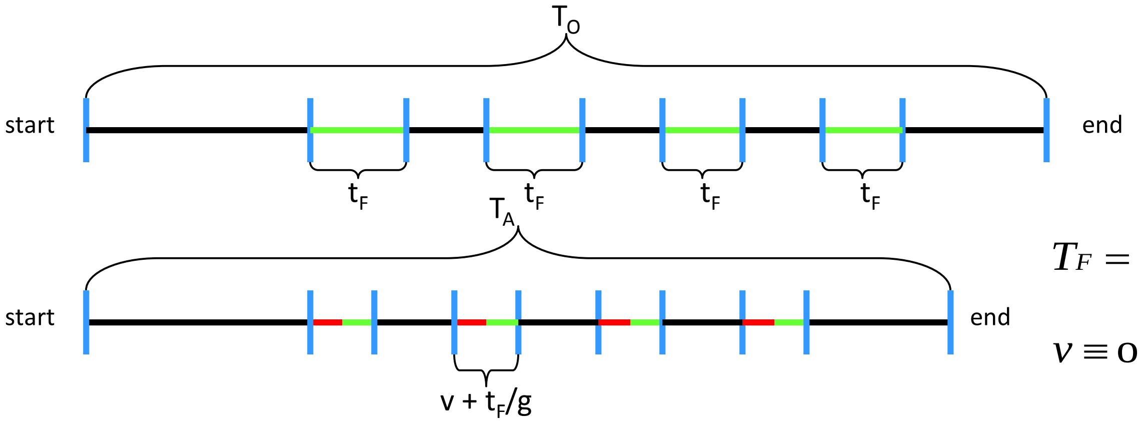


Computational phase diagram from the MPI based GADGET code (used for N-body and SPH simulations) using 1M particles over four time steps on 128 procs.

Red indicates computation
Blue indicates waiting for communication


$v = \frac{1}{4} \times v_0$
$v = \frac{1}{2} \times v_0$
$v = v_0$
$v = 2 \times v_0$
$v = 4 \times v_0$
$\frac{S_{max}}{2}$ curve
Normalized scalability $(\frac{S_{max_0}}{2})$
Normalized parallelism $(\frac{W}{v_0})$
10^1
10^0
10^{-1}
10^{-1}
10^0
10^1
10^2

Amdahl’s Law with Overhead



$$T_F = \sum_i^n t_{Fi}$$

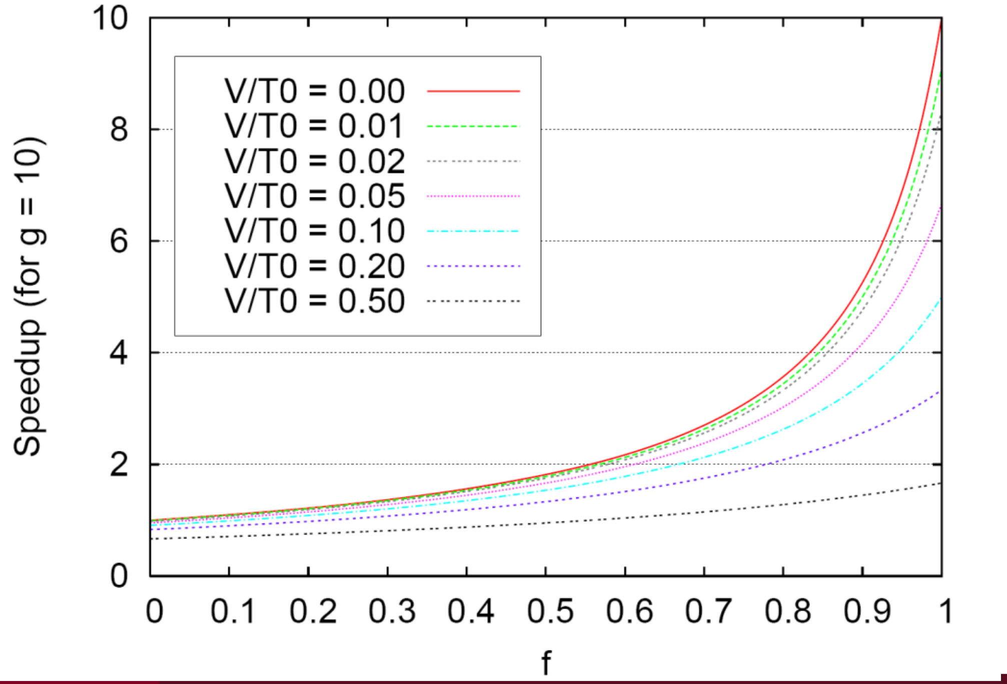
$v \equiv$ overhead of accelerated work segment

$V \equiv$ total overhead for accelerated work $= \sum_i^n v_i$

$$T_A = (1 - f) \times T_O + \frac{f}{g} \times T_O + n \times v$$

$$S = \frac{T_O}{T_A} = \frac{T_O}{(1 - f) \times T_O + \frac{f}{g} \times T_O + n \times v}$$

$$\boxed{S = \frac{1}{(1 - f) + \frac{f}{g} + \frac{n \times v}{T_O}}}$$



Head room, margins, potential innovations
All architectures are von Neumann derivatives
Control is sequential instruction issue, IP

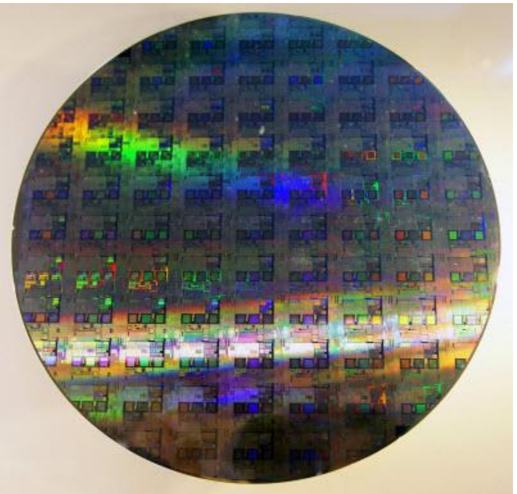
- Costs and burdens
  - Variants: out of order, vector, SIMD, MPPs and clusters
  - Flow control bottlenecks
  - Control state limited to program counters, fork-joins
  - Loss of operational precedence
  - Not effective in asynchronous operation
- Alternatives
  - DAGs
  - Dataflow
  - Systolic arrays
  - unums

Head room, margins, potential innovations
Floating point ALU optimized resource

- Costs and burdens:
  - Cache hierarchy
  - Branch prediction
  - Speculative execution
  - Out of order flow control reservation stations, …
  - Prefetching, many simultaneous in-flight requests
- Alternatives:
  - Emphasis on memory access throughput
  - Response time to incidence of external messages
  - Scratch pad memory
  - Multi-threading
  - Dataflow ISA
  - Asynchronous flow control

Head room, margins, potential innovations

- Separation of CPU and main memory
  - Major bottleneck
  - Worse with multi/many core processor sockets
  - A driver for need for cache
  - Processor in Memory (PIM)
  - On-chip scratch pad memory
- Silicon based semiconductor technology
  - Moore’s Law will flat-line by end of decade, ~ 5 nm feature size
  - Superconducting single flux quantum logic at 100 – 200 GHz, 100X energy advantage
  - Leakage current a challenge
  - Graphene of interest
- CSP/MPI (well, not unquestioned)
  - MPI + X, where X = OpenMP maybe
  - Fork-joins impose Amdahl bottlenecks
  - X could also be DAGs
  - Asynchronous Multi-Task execution models

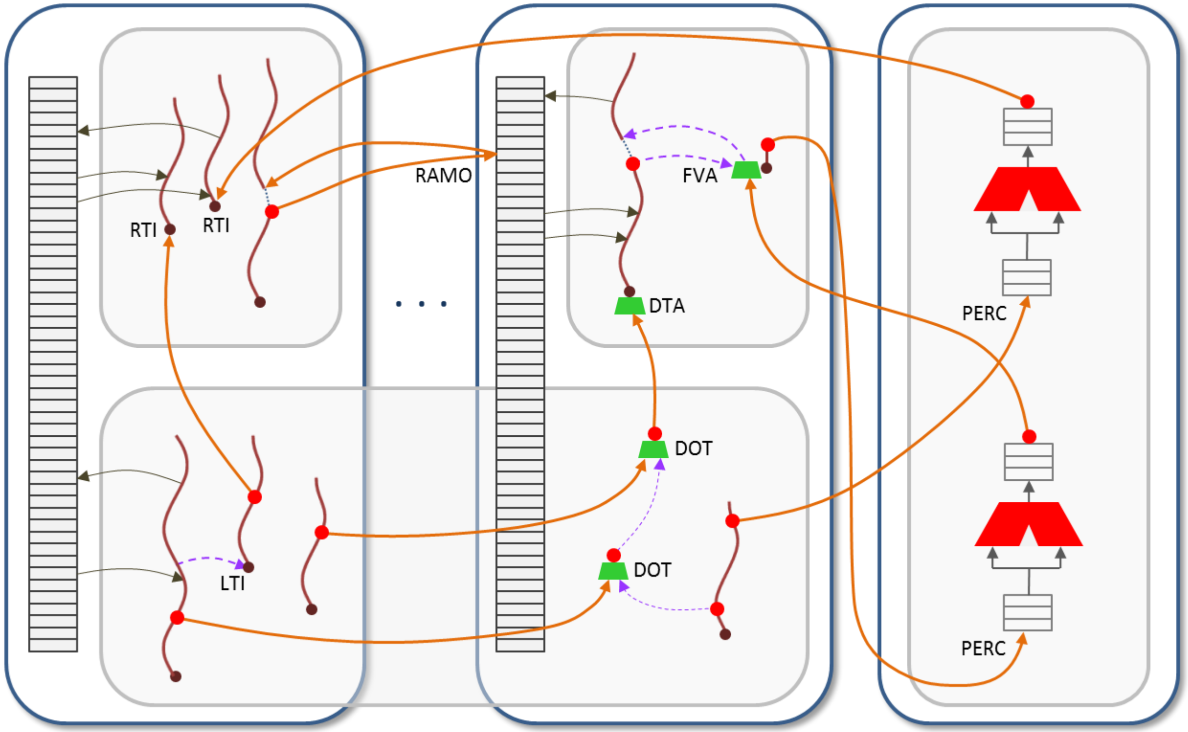


Old Technology Characteristics	New Technology Characteristics
Peak clock frequency as primary limiter for performance improvement	**Power** is primary design constraint for future HPC system design
Cost - FLOPs are biggest cost for system: optimize for compute	**Cost - Data movement dominates:** optimize to minimize data movement
Concurrency - Modest growth of parallelism by adding nodes	**Concurrency: Exponential growth** of parallelism within chips
Memory scaling maintain byte per flop capacity and bandwidth	**Memory Scaling: Compute growing 2x faster** than capacity or bandwidth
Locality: MPI+X model (uniform costs within node & between nodes)	**Locality**: must reason about data locality and possibly topology
Uniformity: Assume uniform system performance	**Heterogeneity**: Architectural and performance non-uniformity increase
Reliability: It's the hardware's problem	**Reliability: Cannot depend on hardware protection alone**

Game Changer – Runtime System

- Runtime system
  - is: ephemeral, dedicated to and exists only with an application
  - is not: the OS, persistent and dedicated to the hardware system
- Moves us from *static* to *dynamic* operational regime
  - Exploits situational awareness for causality-driven adaptation
  - Guided-missile with continuous course correction rather than a fired projectile with fixed-trajectory
- Based on foundational assumption
  - More computational work will yield reduced time and lower power
  - Untapped system resources to be harvested
  - Opportunities for enhanced efficiencies discovered only in flight
  - New methods of control to deliver superior scalability

Distinguishing Features of ParalleX/HPX



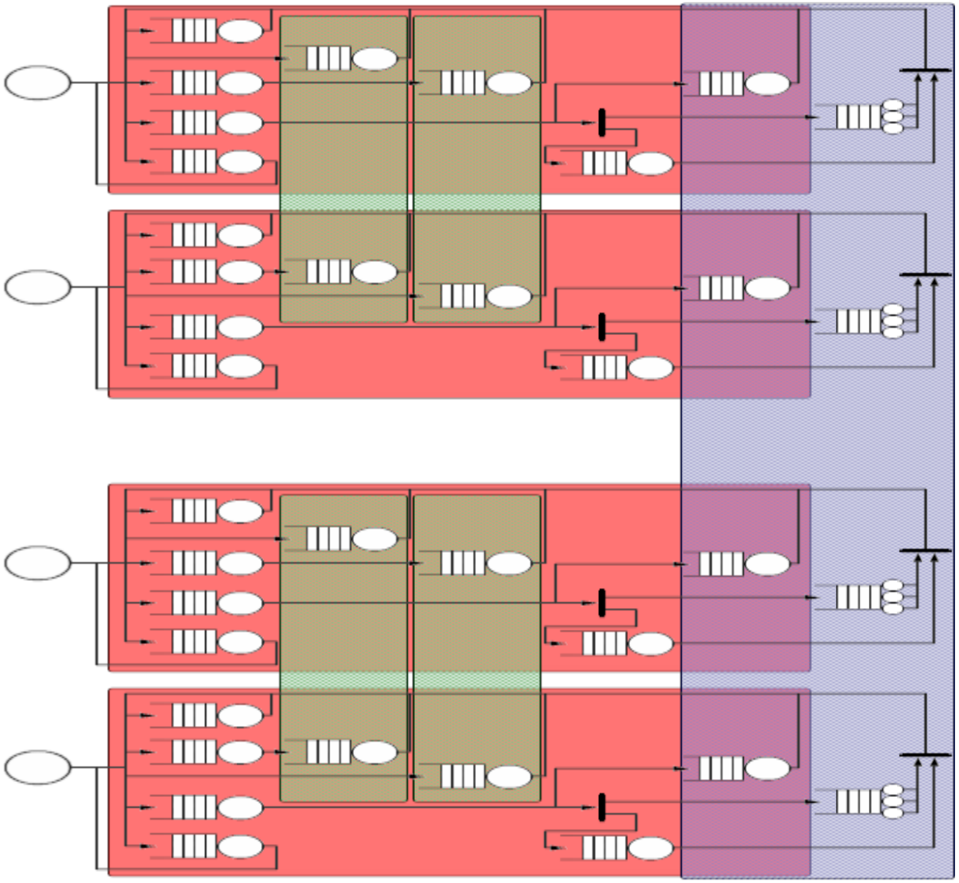
- ▭ Locality
- ▭ Process
- ▤ Local Memory
- 🟩 LCO
- 🟥 Accelerator

- ∼ Thread
- ---- Suspended Thread
- → Local Memory Access
- ● AGAS Address Lookup
- - - → Local Action
- → Parcel

LTI: local thread instantiation
RTI: remote thread instantiation
RAMO: remote atomic memory operation
DTA: depleted thread activation
DOT: dataflow object trigger
FVA: future value access
PERC: percolation

Performance Model, Full Example System

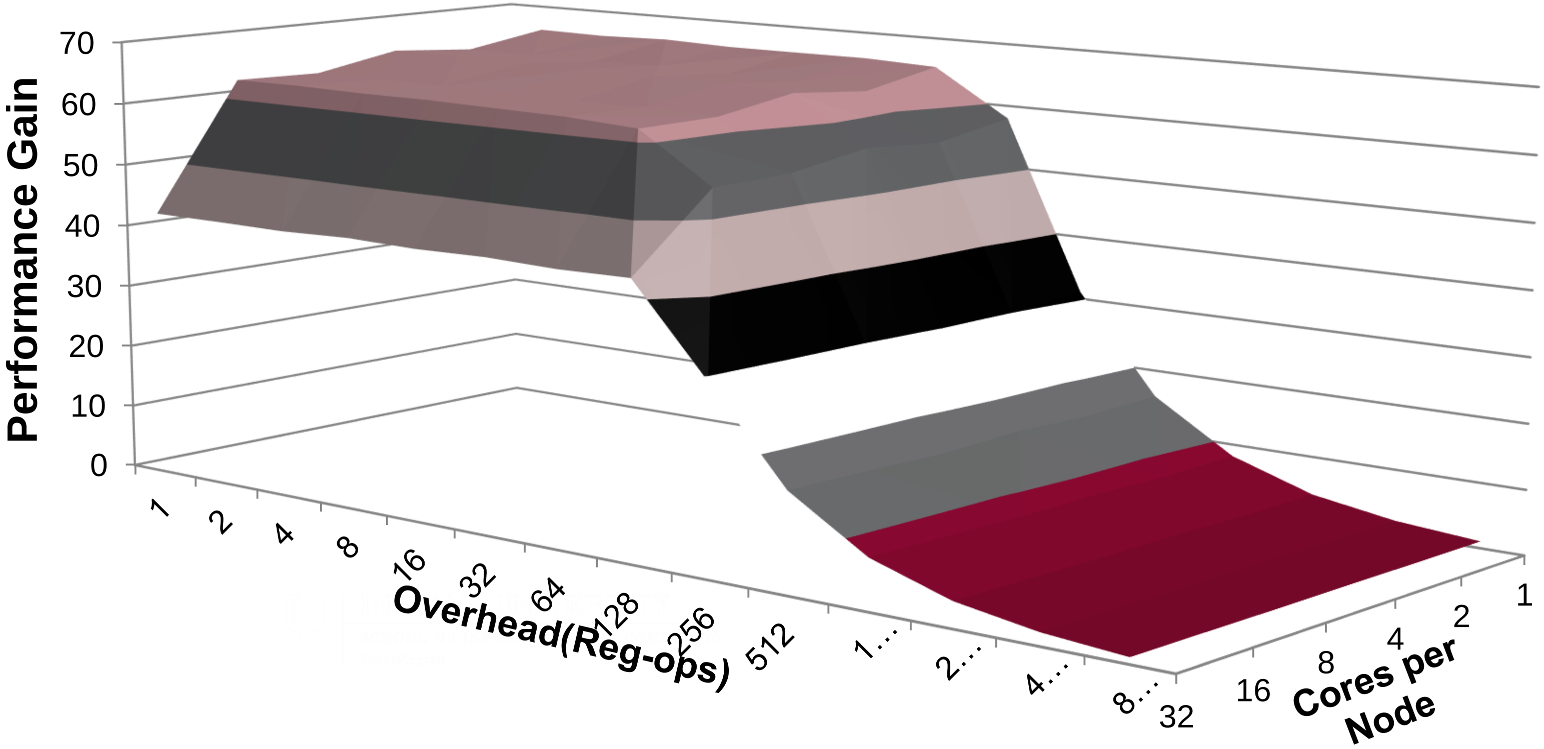
Modeling the full example system



- Example system:
  - 2 nodes,
  - 2 cores per node,
  - 2 memory banks per node
- Accounts for:
  - Functional unit workload
  - Memory workload/latency
  - Network overhead/latency
  - Context switch overhead
  - Lightweight task management (red regions can have one active task at a time)
  - Memory contention (green regions allow only a single memory access at a time)
  - Network contention (blue region represents bandwidth cap)
  - NUMA affinity of cores
- Assumes:
  - Balanced workload
  - Homogenous system
  - Flat network

Gain with Respect to Cores per Node and Overhead; Latency of 8192 reg-ops, 64 Tasks per Core

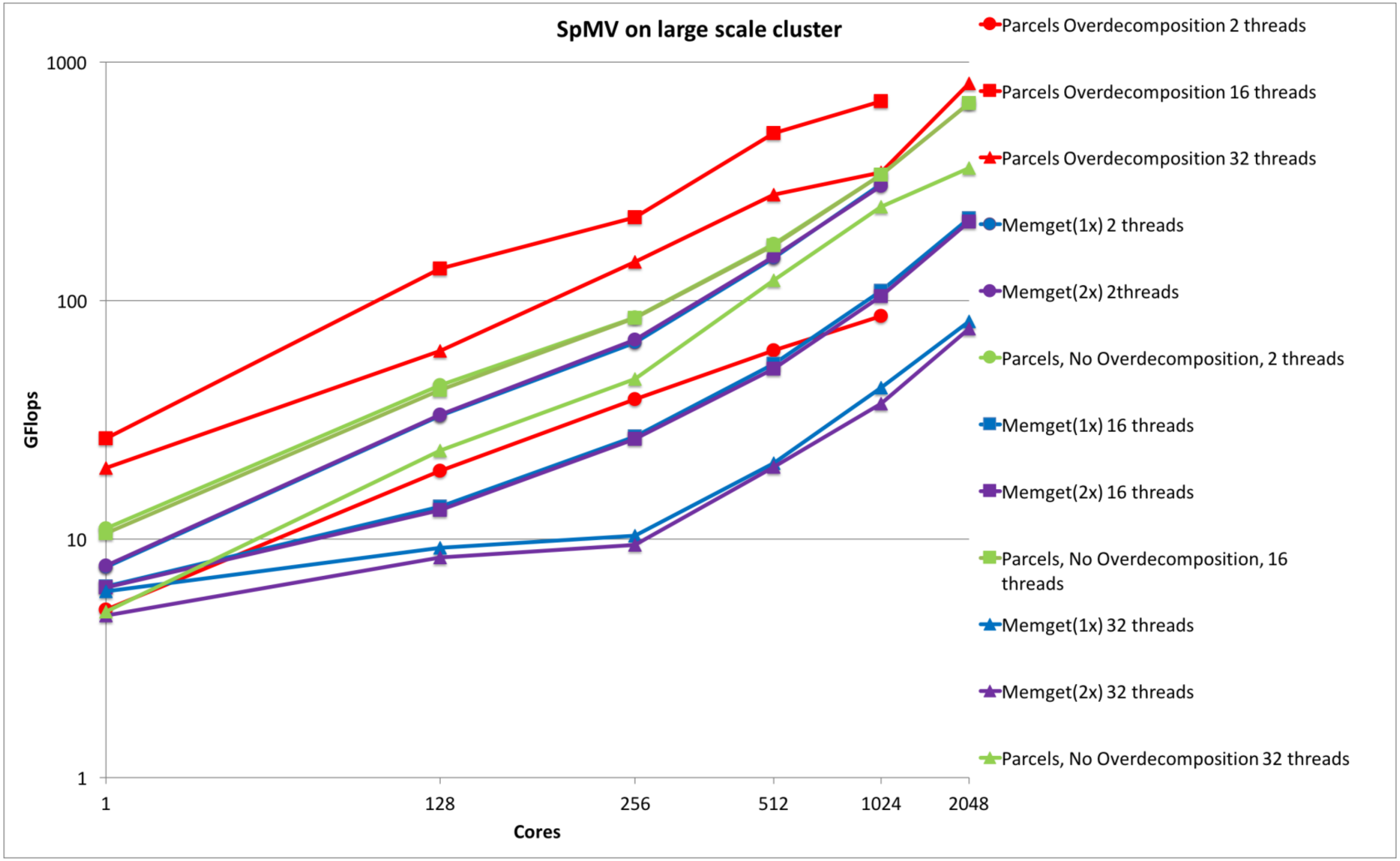
Performance Gain of Non-Blocking Programs over Blocking Programs with Varying Core Counts (Memory Contention) and Overheads



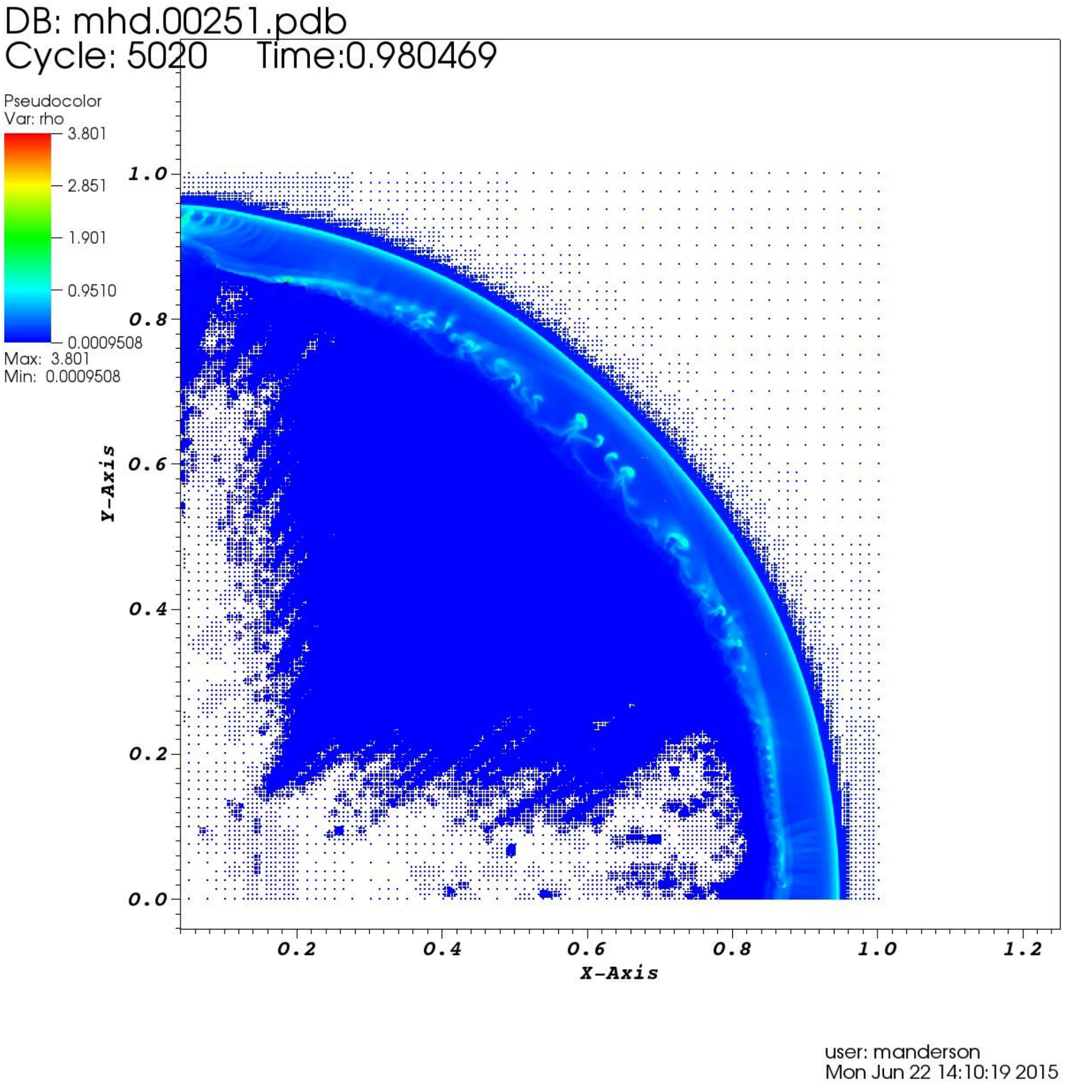
Motivation for HPX

- Exploit runtime information through introspection to discover parallelism for **scalability** and dynamically manage resources to demand for **efficiency**
- Expose limitations of conventional computer **architecture** and devise mechanisms for lower overhead and latency
- Based on a crosscutting **execution model** to determine respective roles, responsibilities, and interoperability
- Serve as a **research** platform to explore utility, generality, opportunity, and challenges/limitations
- Target and enabler for parallel **programming models**
- Operation in the presence of uncertainty of **asynchrony**
- First conceived in support of HTMT project and Cascade

SpMV for parcels and memget



Wavelet Adaptive Multiresoultion



Courtesy of Matt Anderson, IU

Time Required to Check if Memory Address is Local or Remote in HPX5

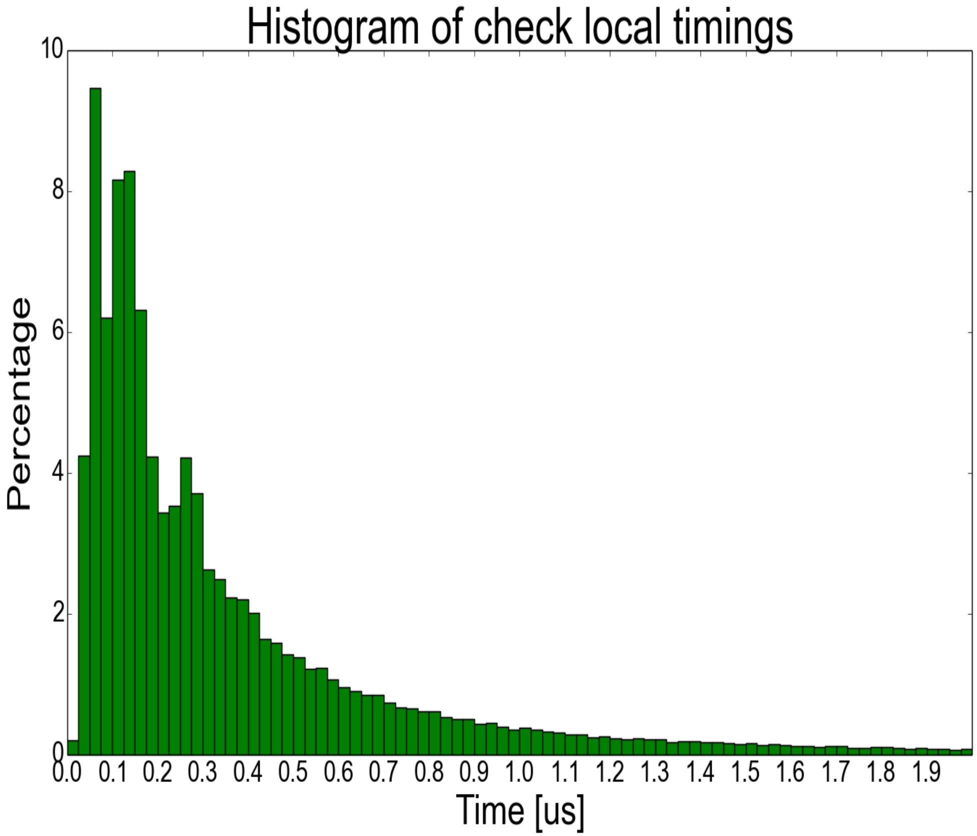


Chart courtesy of Daniel Kogler, IU

Time Required to Perform a Context Switch Between Lightweight Threads in HPX5

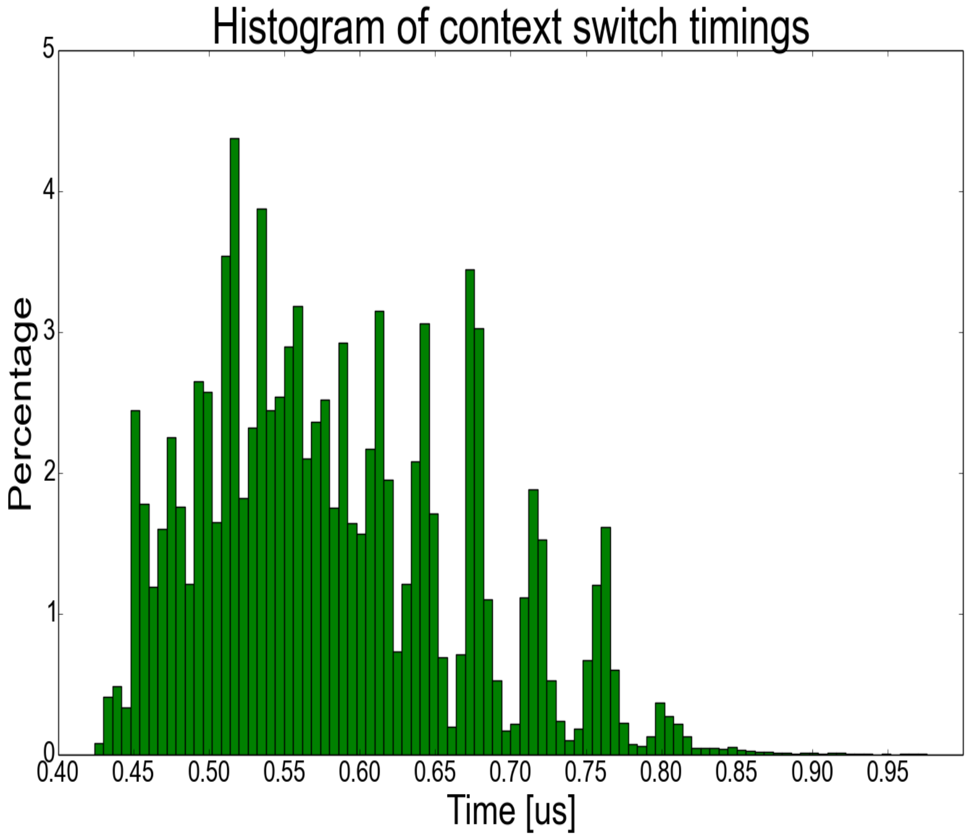


Chart courtesy of Daniel Kogler, IU

Time Required to Create a New Lightweight Thread in HPX5

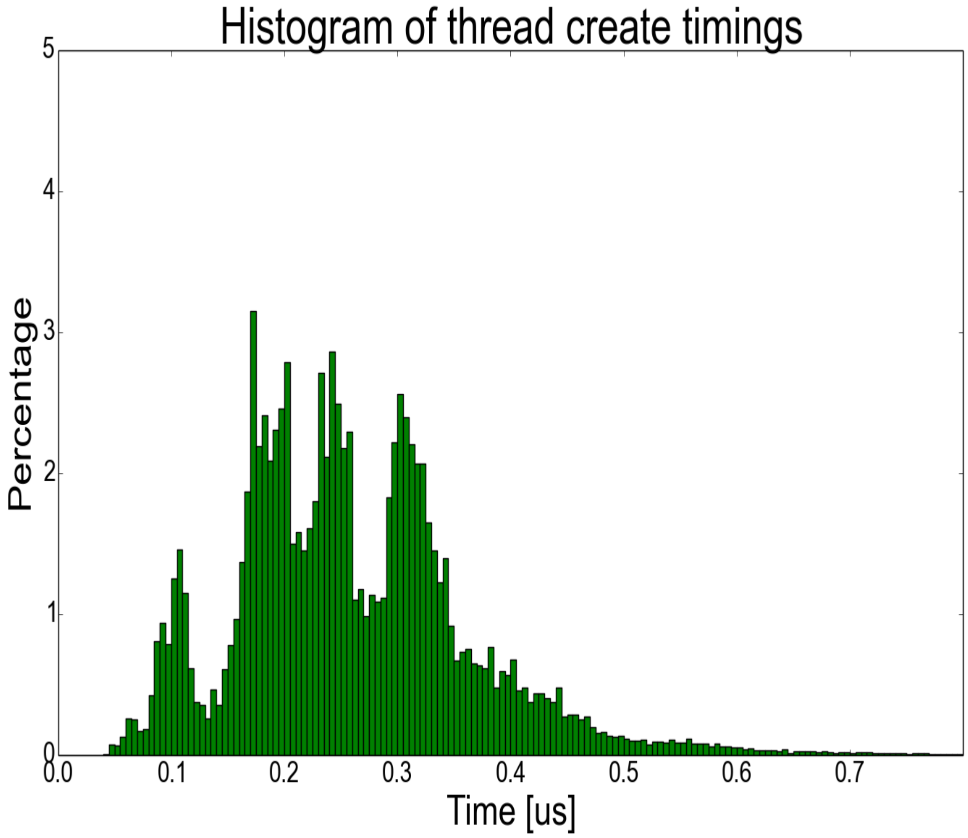
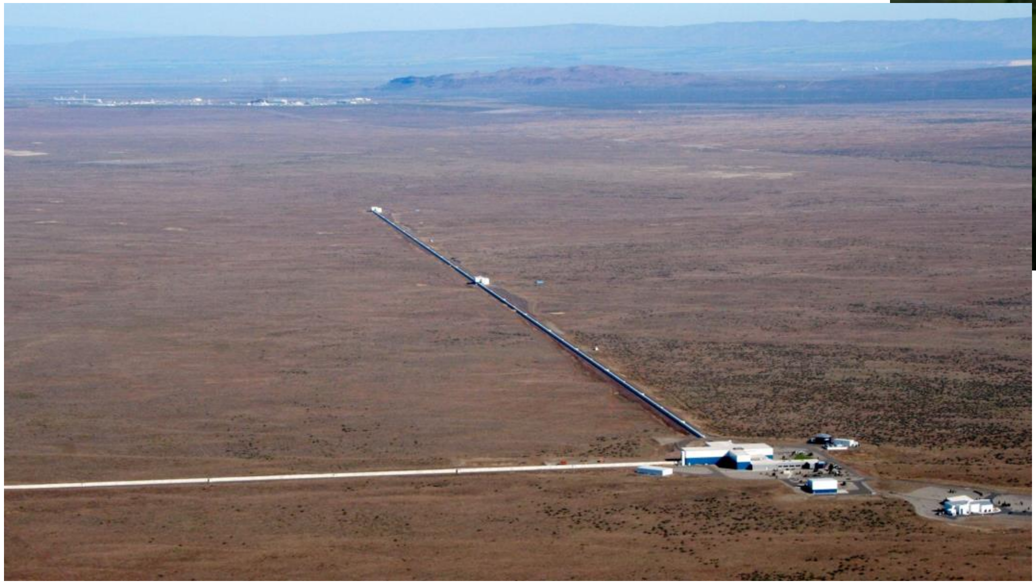


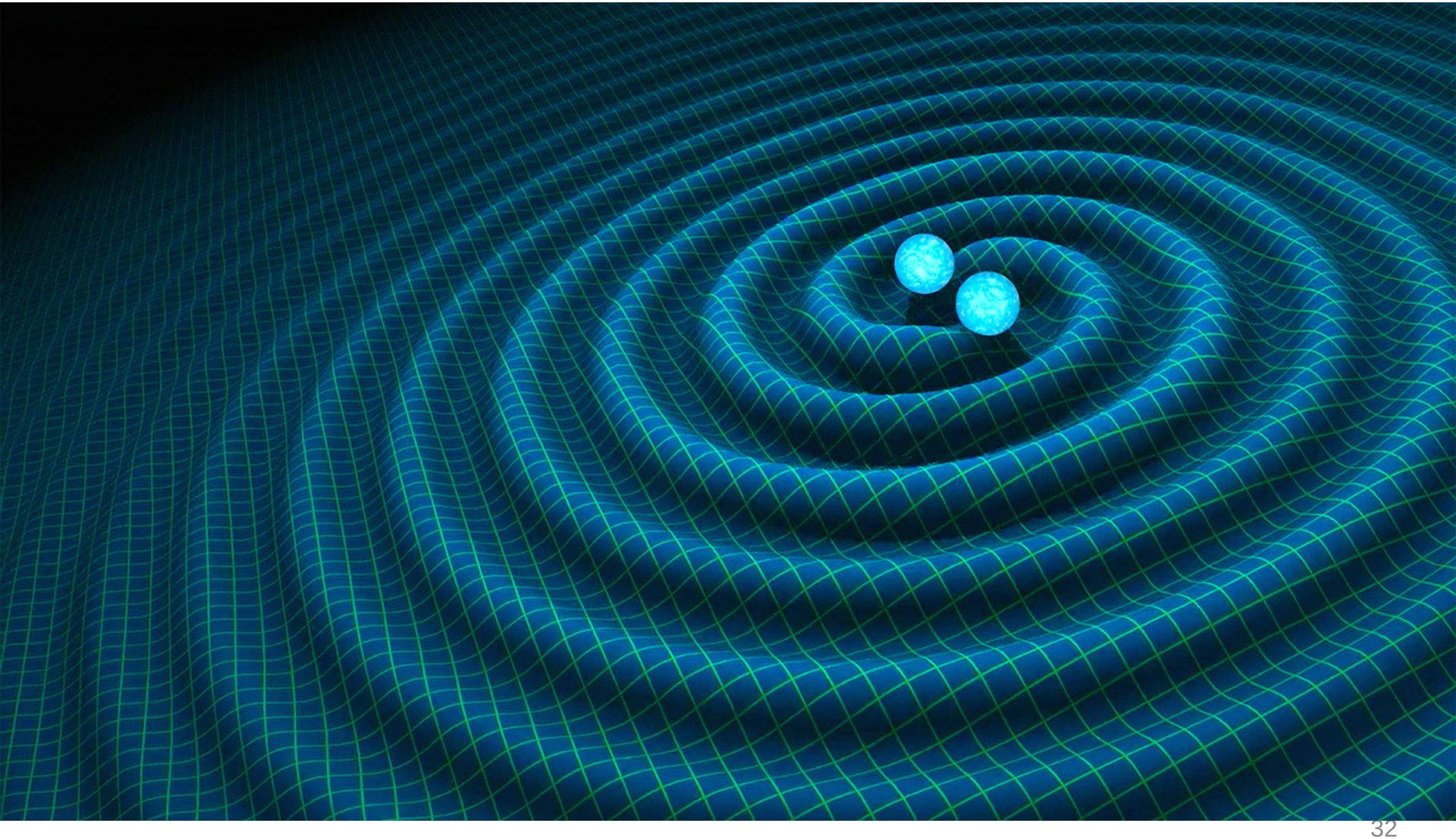
Chart courtesy of Daniel Kogler, IU

Laser Interferometric Gravitational-wave Observatory (LIGO)

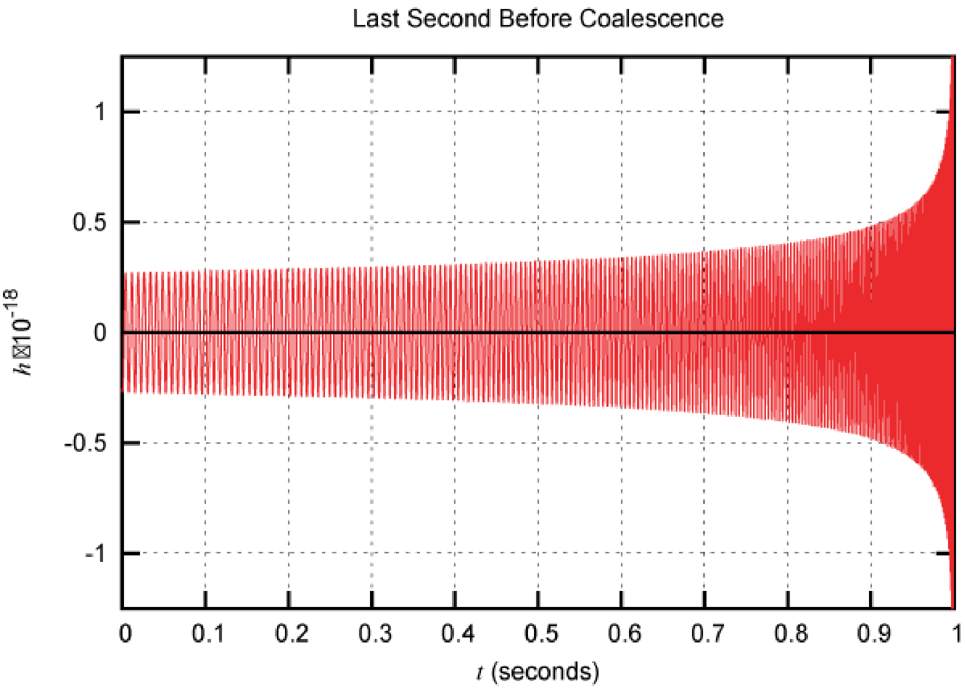
Hanford, WA



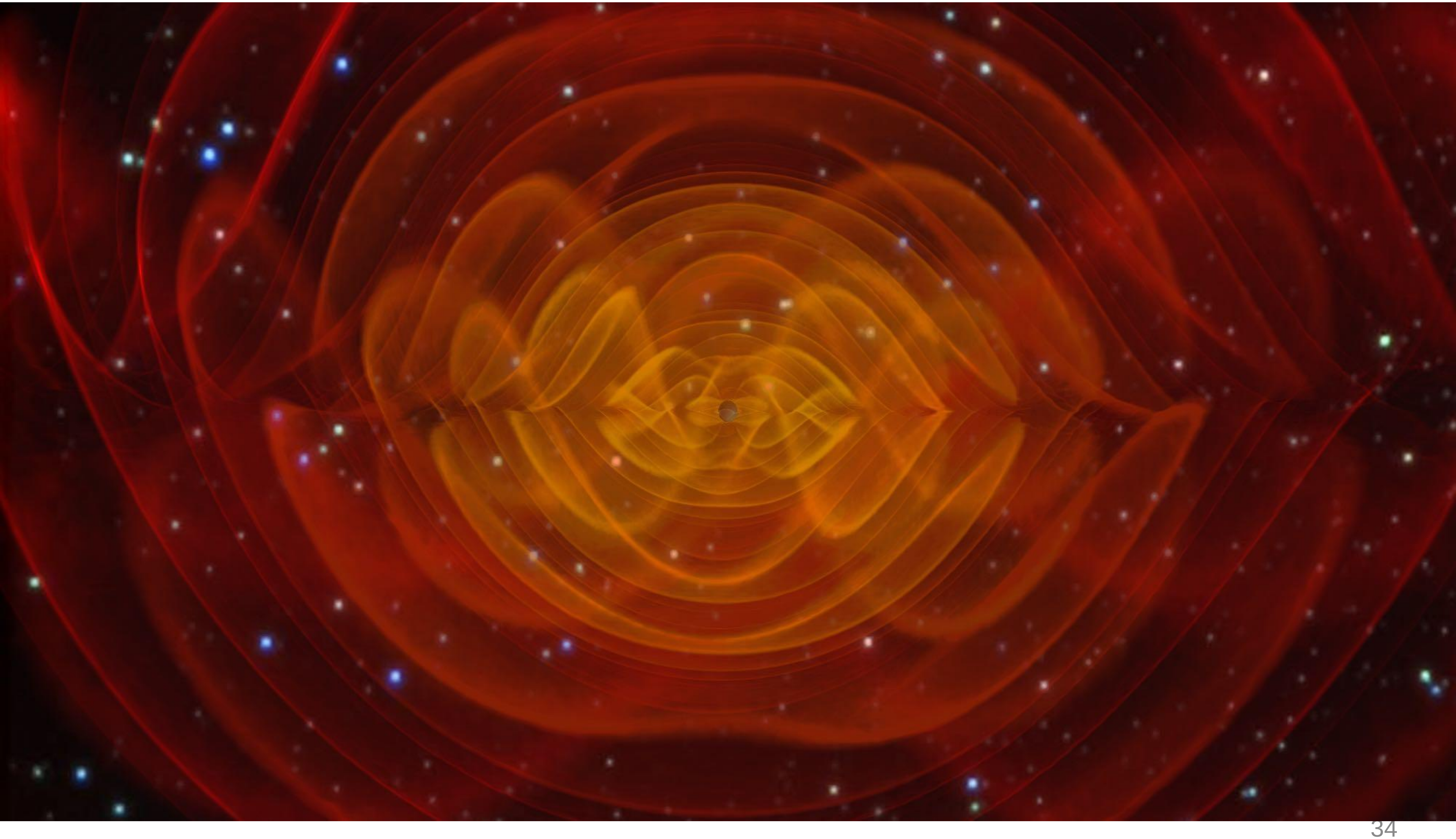
Livingston, LA



LIGO Chirp Filter for Signal Target

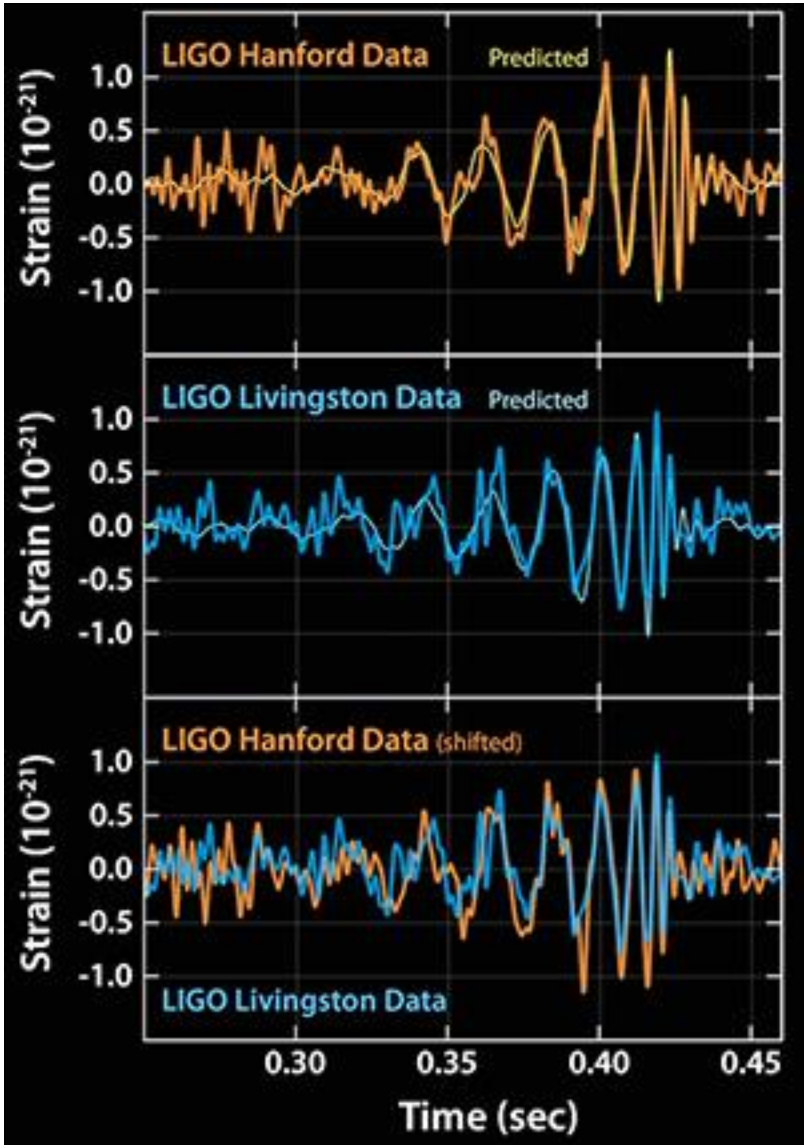


Kaplowee!!!



Discovery

- 14 September, 2015
- Combined objects of 29 and 36 solar masses
- Produced a black hole of 62 solar masses.
- Missing 3 solar masses converted to gravitational waves
- Travelled 1.3 billion years to Earth
- 50X all the power of all the stars in the universe



INDIANA UNIVERSITY
CardCow.com