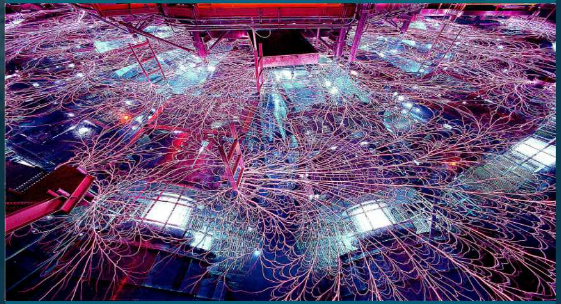
This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

SAND2019-7003C

Evaluating the Marvell ThunderX2 Server Processor for HPC Workloads



PRESENTED BY

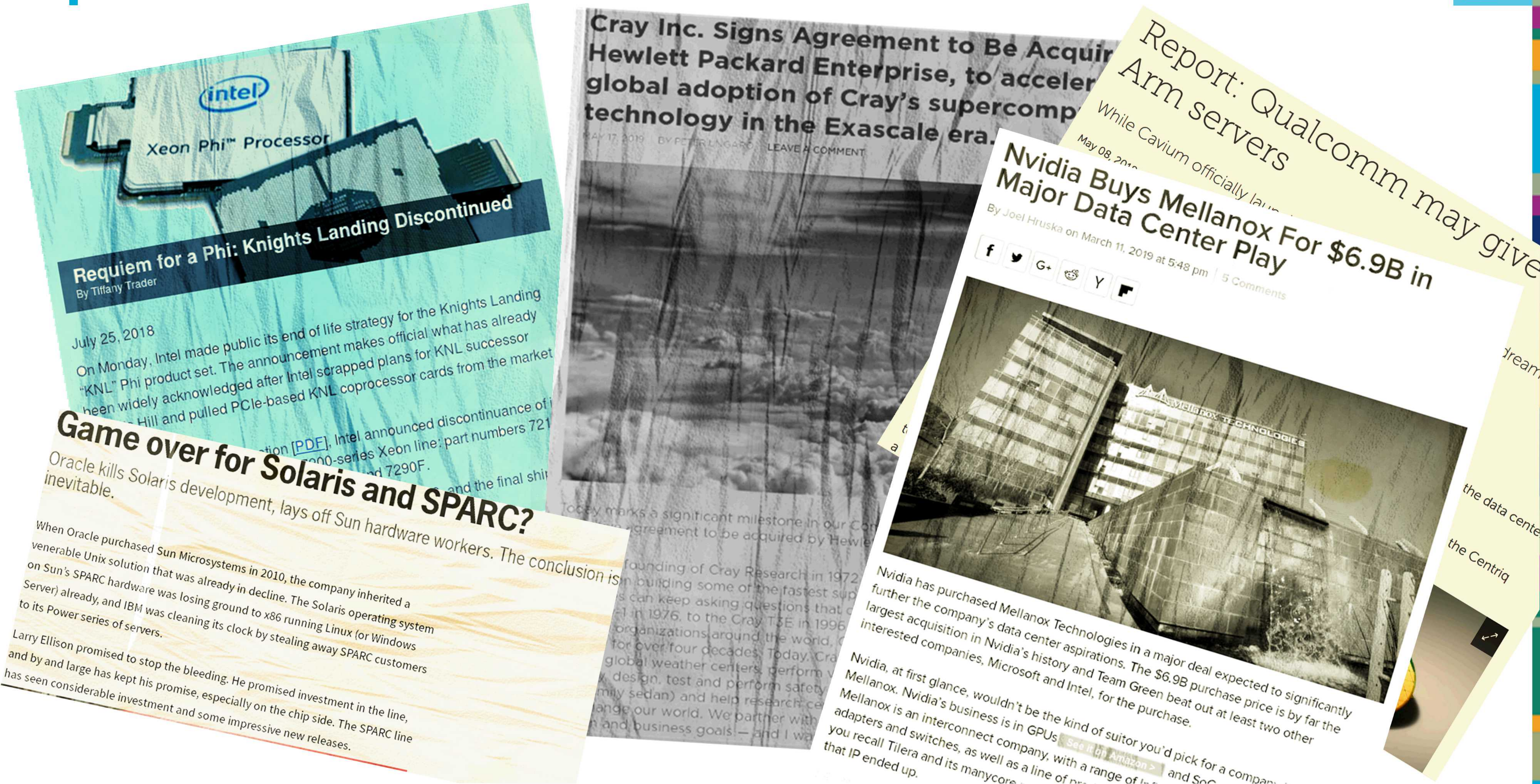
Clayton M. Hughes, Sandia National Laboratories

HPBench 2019
July 15 – 19, 2019

S.D. Hammond, C. Hughes, M.J. Levenhagen, C.T. Vaughan, A.J. Younge, B. Schwaller, M.J. Aguilar, K.T. Pedretti and J.H. Laros

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Heading Toward Grim Times in HPC



Requiem for a Phi: Knights Landing Discontinued

By Tiffany Trader

July 25, 2018

On Monday, Intel made public its end of life strategy for the Knights Landing "KNL" Phi product set. The announcement makes official what has already been widely acknowledged after Intel scrapped plans for KNL successor Hill and pulled PCIe-based KNL coprocessor cards from the market

tion [PDF], Intel announced discontinuance of i 7200-series Xeon line: part numbers 721 nd 7290F.

and the final shi

Game over for Solaris and SPARC?

Oracle kills Solaris development, lays off Sun hardware workers. The conclusion is inevitable.

When Oracle purchased Sun Microsystems in 2010, the company inherited a venerable Unix solution that was already in decline. The Solaris operating system on Sun's SPARC hardware was losing ground to x86 running Linux (or Windows Server) already, and IBM was cleaning its clock by stealing away SPARC customers to its Power series of servers.

Larry Ellison promised to stop the bleeding. He promised investment in the line, and by and large has kept his promise, especially on the chip side. The SPARC line has seen considerable investment and some impressive new releases.

Cray Inc. Signs Agreement to Be Acquir Hewlett Packard Enterprise, to acceler global adoption of Cray's supercomp technology in the Exascale era.

AY 17, 2019 BY PETER UNGARO LEAVE A COMMENT

Today marks a significant milestone in our Co agreement to be acquired by Hewle

founding of Cray Research in 1972 n building some of the fastest sup s can keep asking questions that -1 in 1976, to the Cray T3E in 1996 organizations around the world, for over four decades. Today, Cra global weather centers, perform design, test and perform safety mily sedan) and help research ce ange our world. We partner with n and business goals — and I wa

Report: Qualcomm may give Arm servers

While Cavium officially laun

May 08, 201

dream

the data cente

the Centriq

Nvidia Buys Mellanox For $6.9B in Major Data Center Play

By Joel Hruska on March 11, 2019 at 5:48 pm | 5 Comments

Nvidia has purchased Mellanox Technologies in a major deal expected to significantly further the company's data center aspirations. The $6.9B purchase price is by far the largest acquisition in Nvidia's history and Team Green beat out at least two other interested companies, Microsoft and Intel, for the purchase.

Nvidia, at first glance, wouldn't be the kind of suitor you'd pick for a compan Mellanox. Nvidia's business is in GPUs See it on Amazon > and SoC Mellanox is an interconnect company, with a range of In adapters and switches, as well as a line of pr you recall Tilera and its manycore that IP ended up.

What’s That Mean For The HPC Community?

HPC is a becoming an expensive business

- Cloud and Data Center growth is driving processor features and hardware development
  - Not always in the direction that the science and engineering communities require

Sucking the High-Performance Out of HPC

Fewer integrators in the HPC business

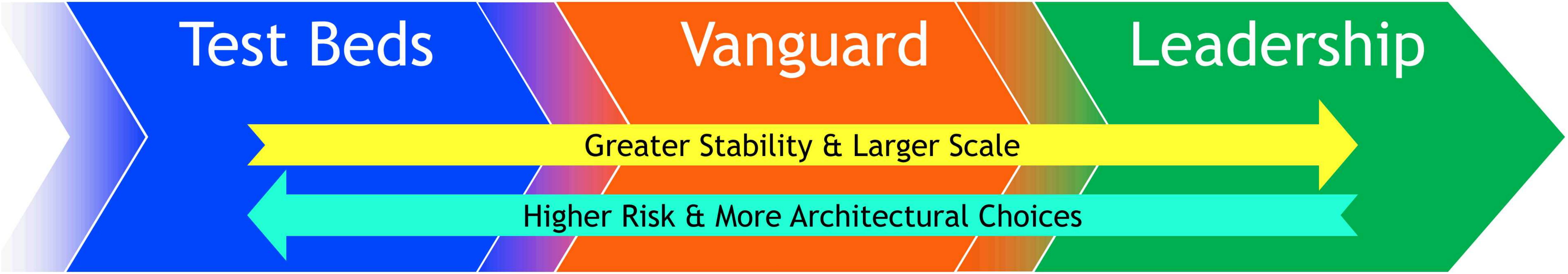
- Developing technology for HPC and solutions in the software space are expensive and need platforms for testing
- Leads to slower pace in innovation and higher cost
  - Greater risk for cutting edge technology required for leadership-class platforms

Less Choice

Higher Risk

Higher Costs

Where Vanguard Fits


Test Beds
Vanguard
Leadership
Greater Stability & Larger Scale
Higher Risk & More Architectural Choices

Test Beds

- Small testbeds (~10-100 nodes)
- Focus on breadth of architectures
- Brave users

Vanguard

- Larger-scale experimental systems
- Focused efforts to mature new technologies
- Broader user-base
- Demonstrate viability for production use
- NNSA Tri-lab resource

ATS/CTS Platforms

- Leadership-class systems (Petascale, Exascale, ...)
- Advanced technologies, sometimes first-of-kind
- Broad user-base
- Production use

Vanguard: Prototype Systems for Advanced Architectures

Vanguard is a project, not a single platform

Expand the HPC ecosystem by developing emerging, yet-to-be-proven, technologies

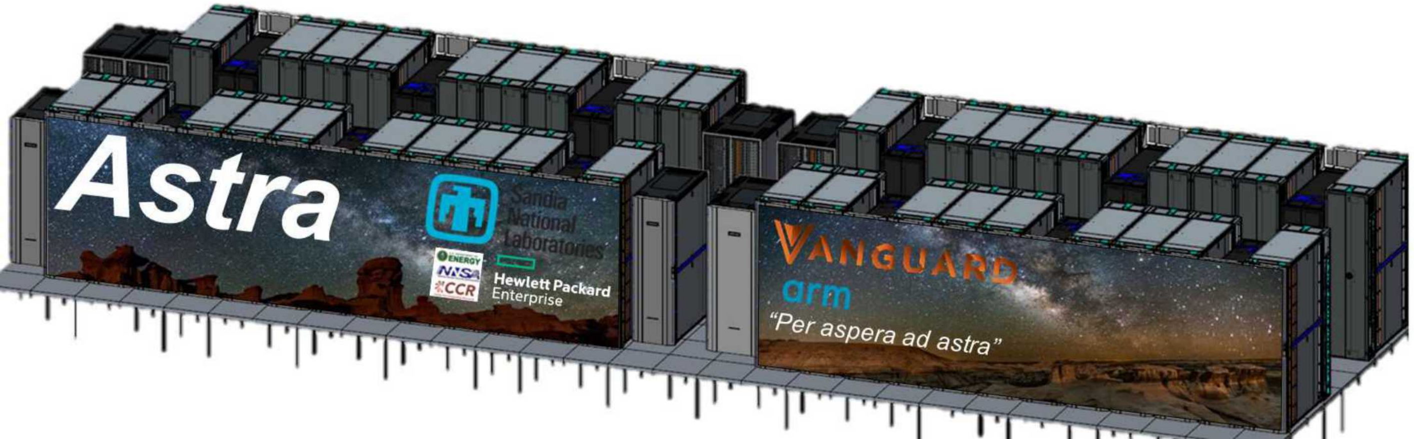
- Increase technology choices, influence HPC community
- Is the technology viable for future production platforms supporting ASC integrated codes?

Address hardware and software technologies together

- If hardware technology is new, gaps in the software stack are certain

Buy down risk before commitment on capability/capacity class investment

Demonstrate viability of ARM for U.S. DOE Supercomputing



per aspera ad astra
Through Difficulties To The Stars

2.3 PFLOPs peak
885 TB/s memory bandwidth peak
332 TB memory
1.2 MW

Contributions

Evaluate the performance of Astra’s Arm-based ThunderX2 processor using microbenchmarks

- Memory bandwidth
- Cache bandwidth
- Floating point performance

Use mini-applications to project the performance of the TX2 on real applications of interest to the scientific community

- Memory bandwidth
- Indirect memory accesses from cache and main memory
- Throughput
- Vectorization and SMT

Marvell ThunderX2 Microarchitecture

Arm 8.1 architecture

Core

- 28 physical PEs @ 2.0GHz
- 4-way SMT

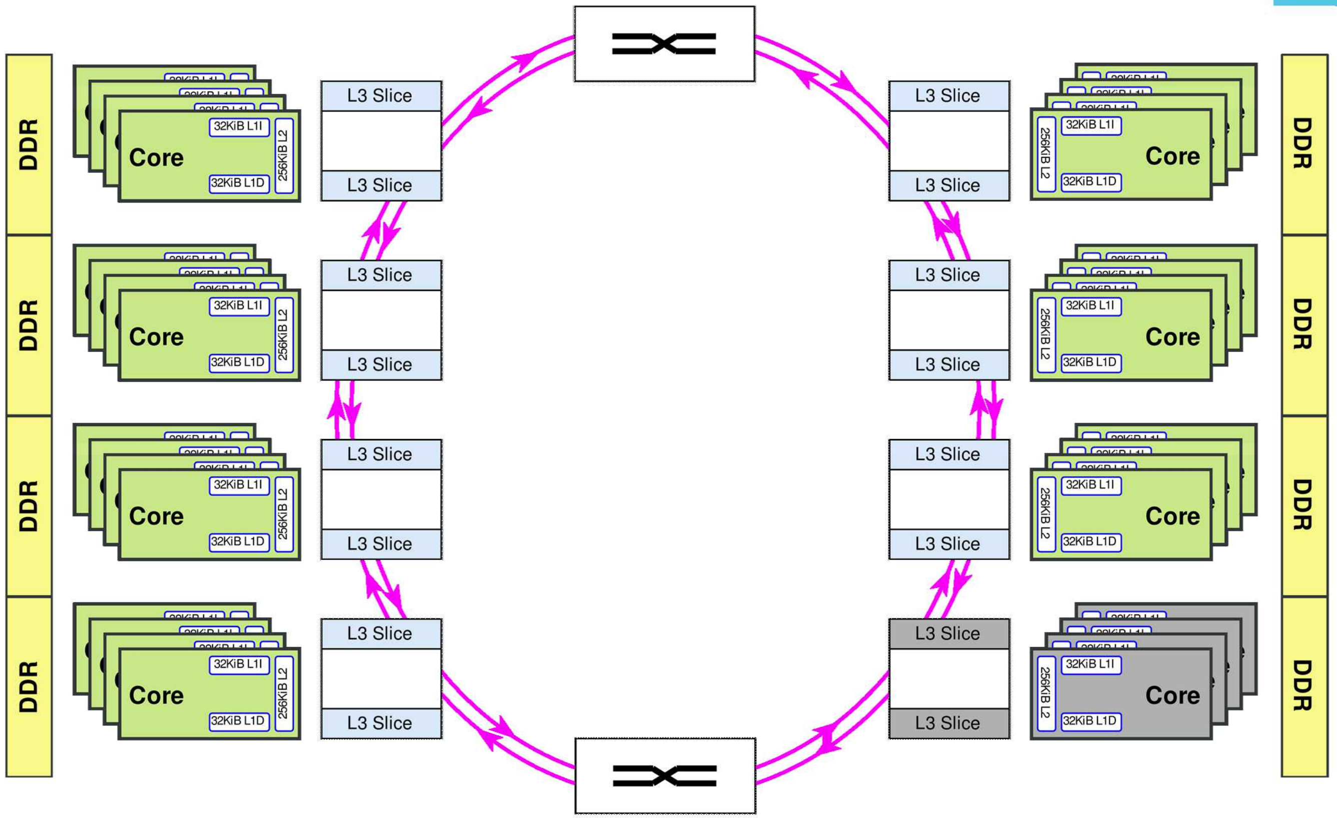
Memory

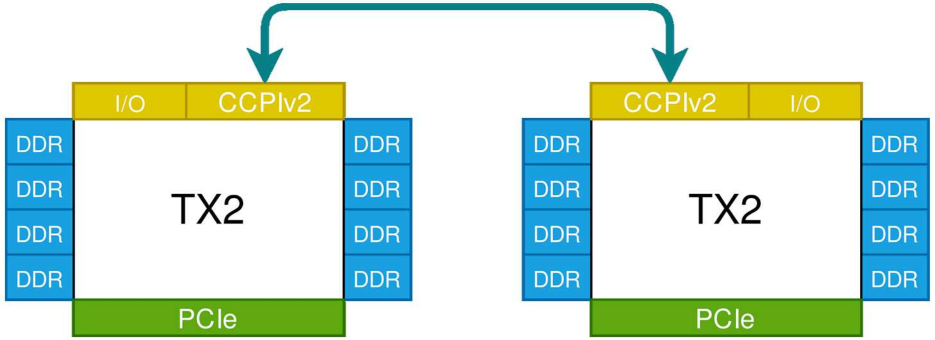
- 32KiB L1 I and D caches
- 256KiB 8-way L2 cache
- 28MiB L3 cache (1MiB per core)

- 64GiB DDR4 (8GiB per channel)

Network

- Ring-based interconnect
- Sockets connected via CCPIv2





Evaluation

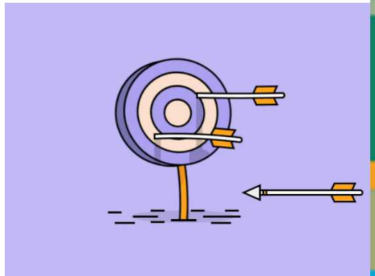
Results averaged over 10 runs with random nodes chosen for each trial

Two Intel-based platforms used for comparison

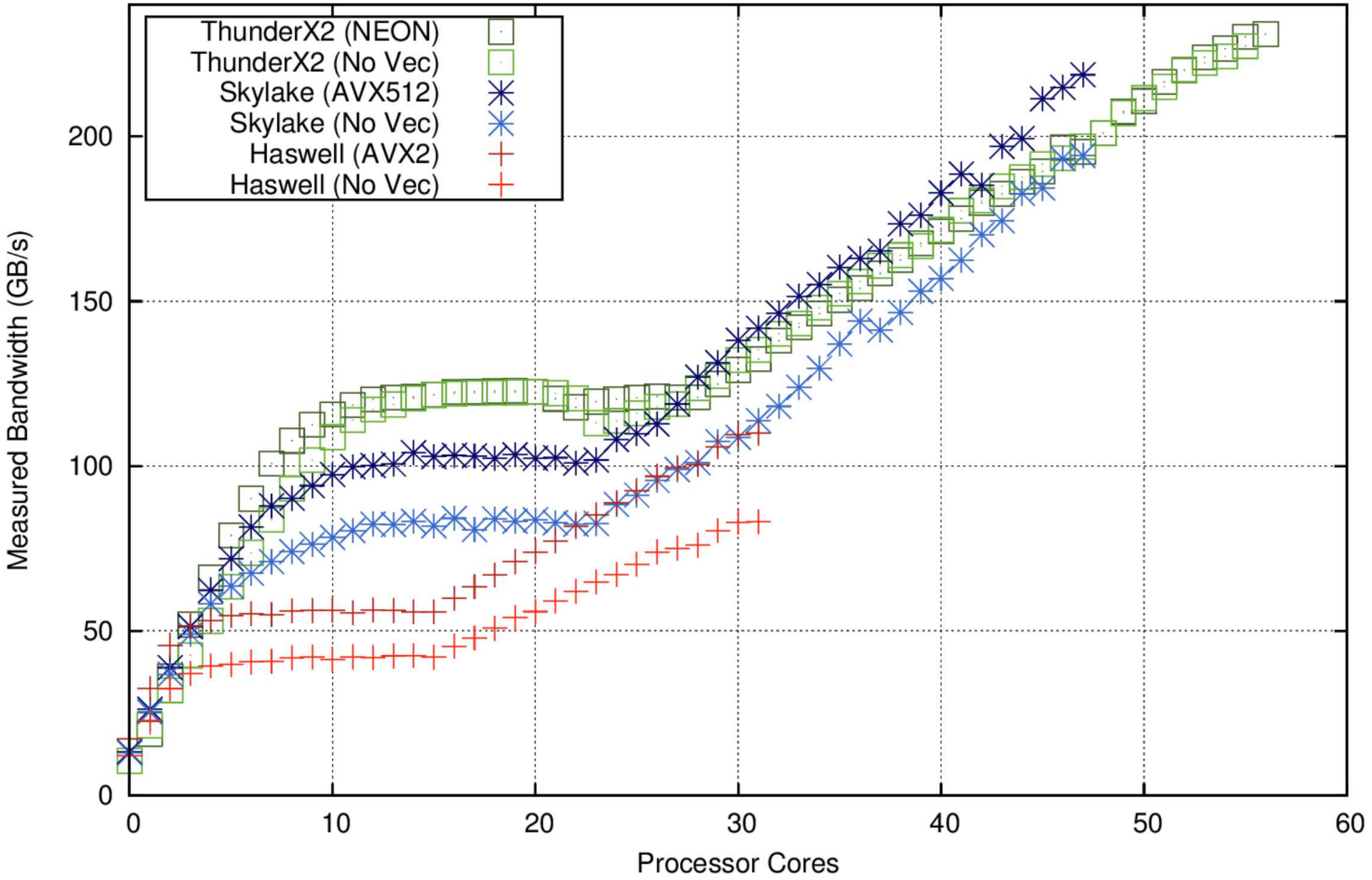
- Mutrino (Haswell)
  - Dual-socket Xeon E5-2697v3
    - 2.3GHz
    - 16 cores with dual SMT
    - 32KiB L1/256KiB L2/40MiB distributed L3
  - 128GB 2133MT/s DDR4
- Blake (Skylake)
  - Dual-socket Xeon Platinum 8160
    - 2.1GHz
    - 24 cores with dual SMT
    - 32KiB L1/1MiB L2/33MiB distributed L3
  - 192GB 2666MT/s DDR4

ICC 18.1.0

- GCC 4.9.3 compatibility
- MKL 18.1



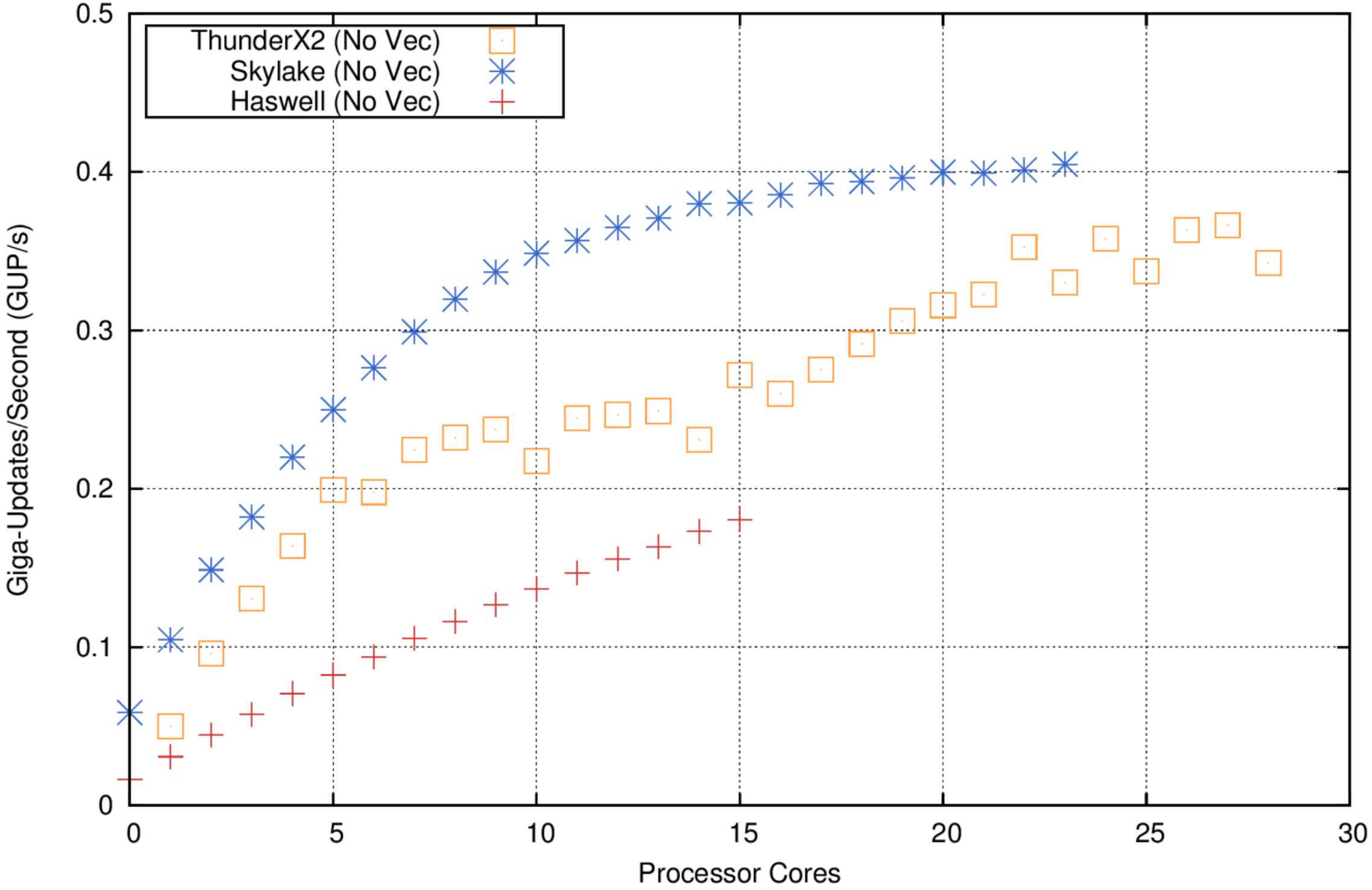
Results – Memory Bandwidth (STREAM)



Extra memory channels on TX2 provide highest memory bandwidth

Vectorization helps the Intel platforms push the efficiency of the memory subsystem on both SKX and HSW but has little effect on the TX2

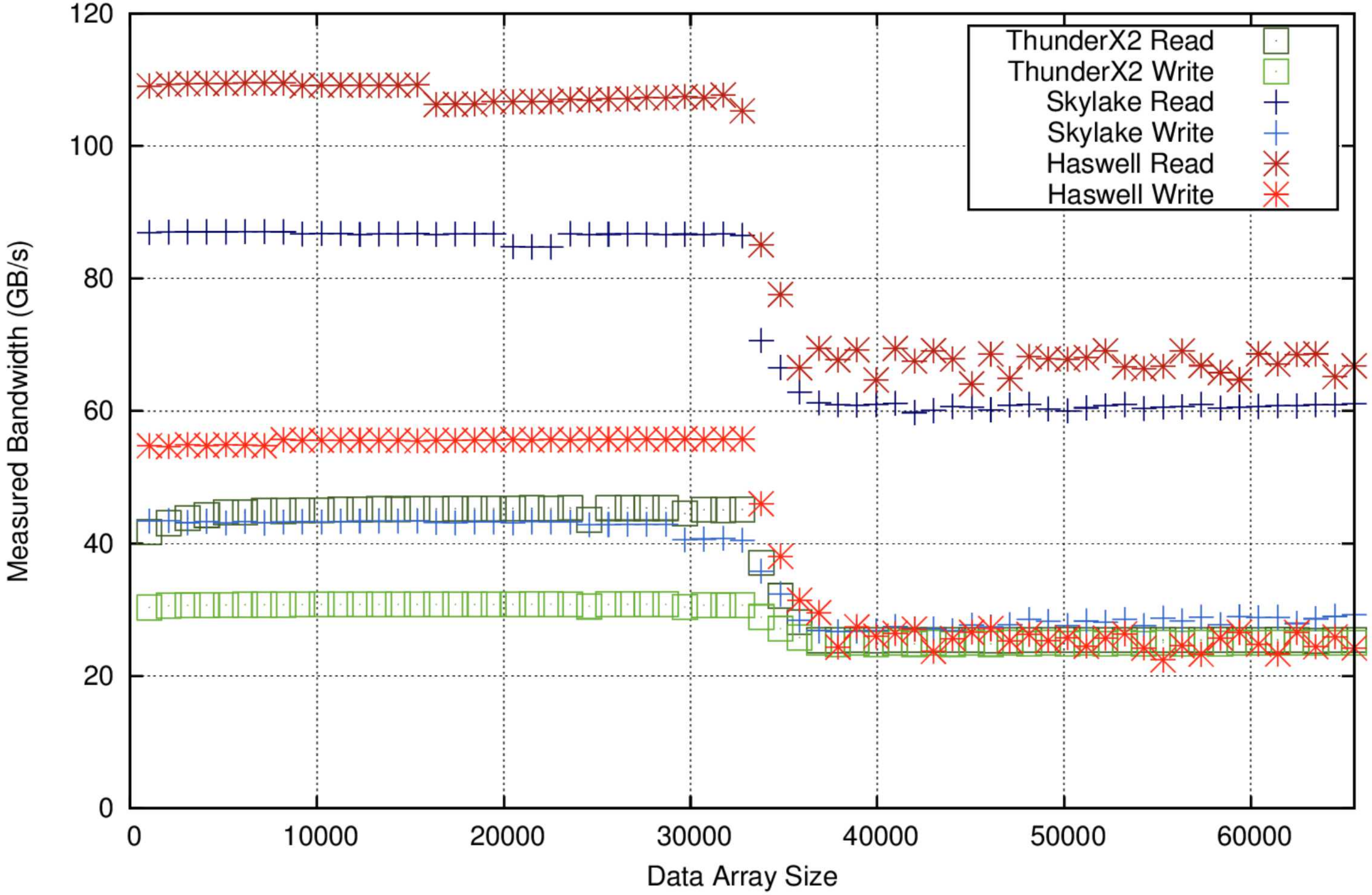
Results – Memory Bandwidth (GUPS)



Neither TX2 nor Skylake exhibit linear performance, tapering off quickly

Ring interconnects on TX2 and Haswell limit scalability

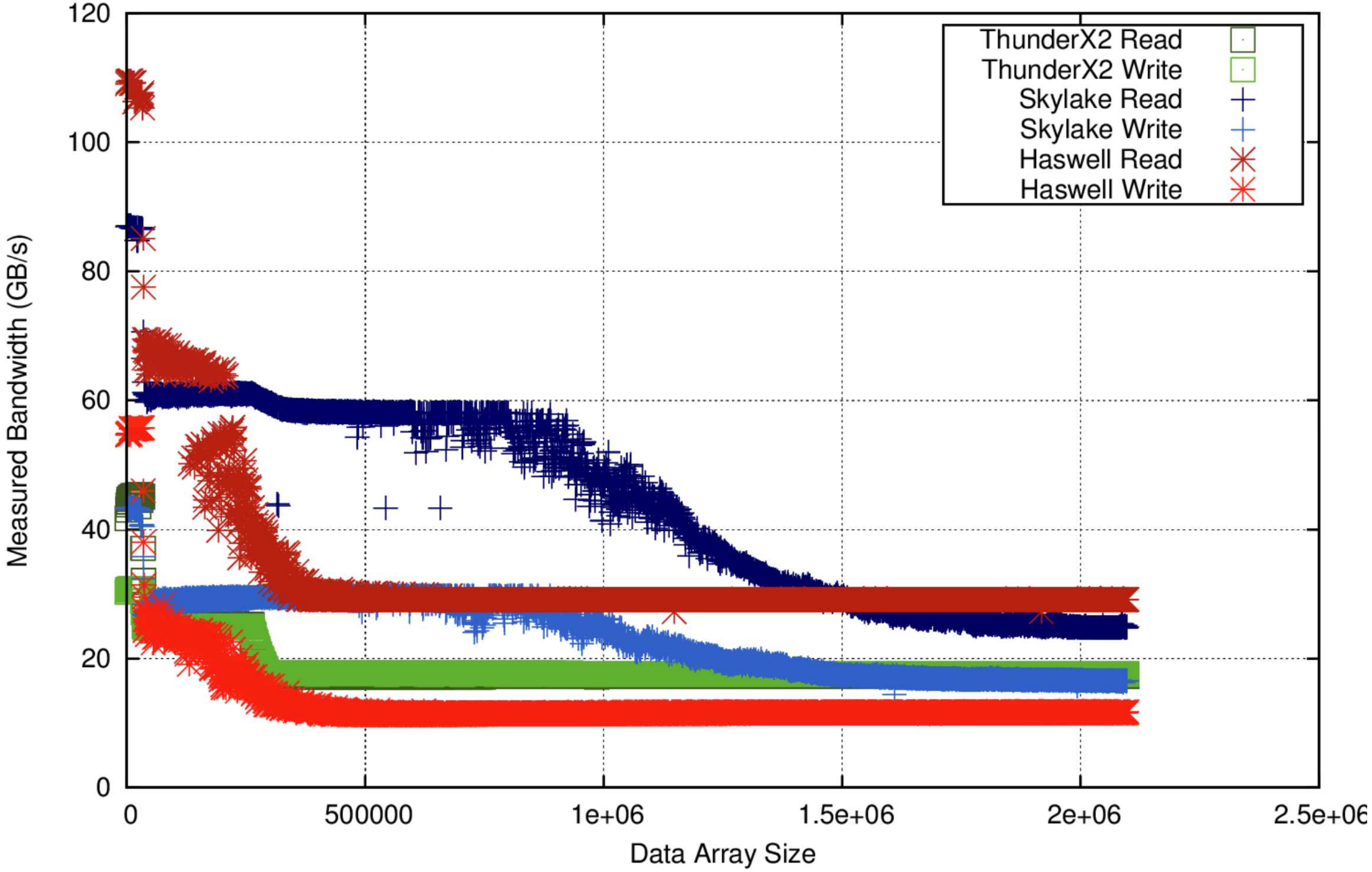
Results – Cache Bandwidth (Small Array)



Haswell and Skylake clock frequencies clearly beneficial for throughput

Smaller R/W BW delta on TX2 due to architectural design decisions

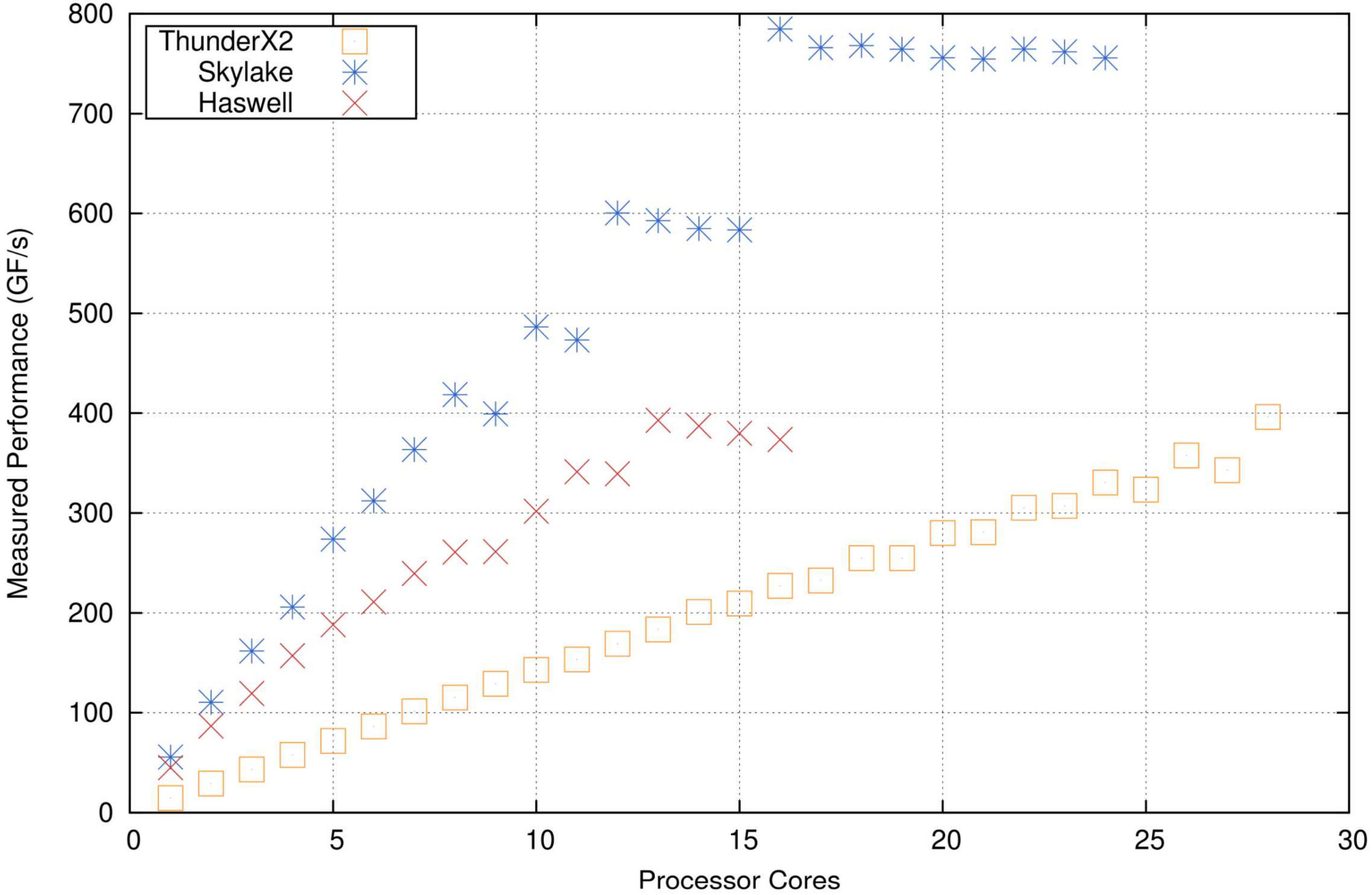
Results – Cache Bandwidth (Large Array)



All three platforms are roughly equal for L3-bound problem sizes

- Haswell likely to be dominant performer for cache read-bound codes
- For applications with large footprints, TX2 provides more even R/W performance

Results – Floating Point Arithmetic (DGEMM)



Skylake outperforms Haswell and TX2 by ~2x

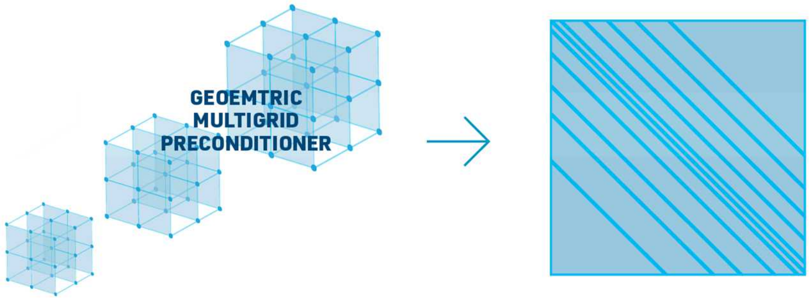
Small SIMD units and limited NEON capabilities clearly hurt TX2

- TX2 is still the most efficient, reaching 88.4% of peak

Mini-Applications and Benchmarks

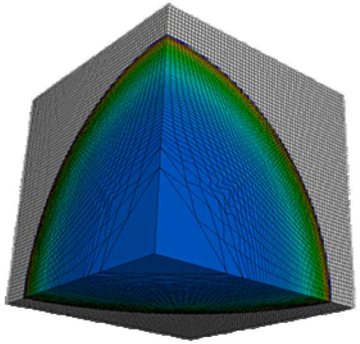
High Performance Conjugate Gradient (HPCG)

- Measures performance of basic HPC operations
- Driven by multigrid preconditioned conjugate gradient algorithm that exercises the key kernels on a nested set of coarse grids



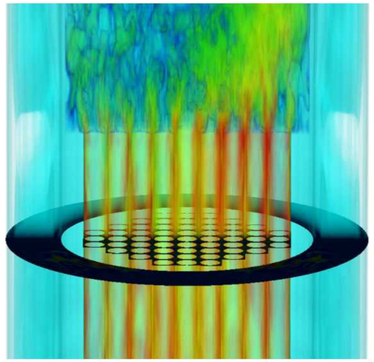
LULESH

- Hydrodynamics over unstructured meshes
- Solves a simple Sedov blast problem with analytic answers



XSBench

- Monte Carlo transport
- Mimics the most computationally expensive steps of a robust nuclear reactor core -- the calculation of macroscopic cross sections



Results – HPCG

Kernel	ThunderX2 (NEON)	ThunderX2 (NoVec)	Skylake (AVX512)	Skylake (NoVec)	Haswell (AVX2)	Haswell (NoVec)
DDOT	20.14	15.96	20.05	30.50	9.87	11.41
WAXBY	23.51	23.52	16.70	16.88	9.53	9.35
SpMV	34.59	34.70	18.56	17.95	10.22	10.20
Multi-Grid	30.97	31.00	18.29	17.94	10.01	9.89
Solve (Total)	30.66	30.51	18.33	18.04	10.03	9.95

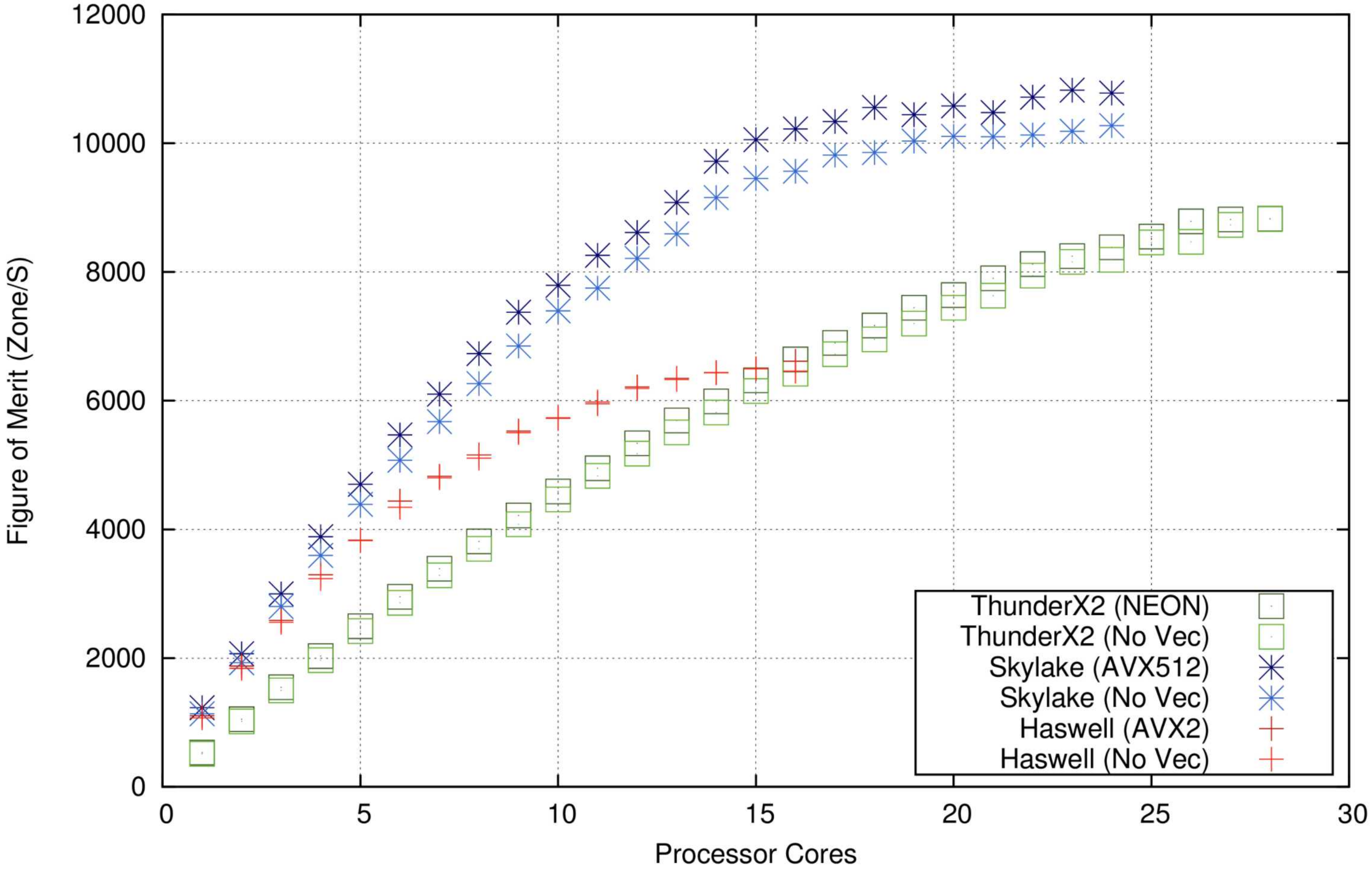
HPCG kernels are considered to be memory bandwidth-bound

- Vectorization makes almost no difference in default implementation
- DDOT is the exception, where vectorization helps on the TX2 but not on the Intel systems

Would expect 1.3-2x performance improvement but observe a 1.4-3.4x improvement

- Driven by larger core count and memory channels

Results – LULESH

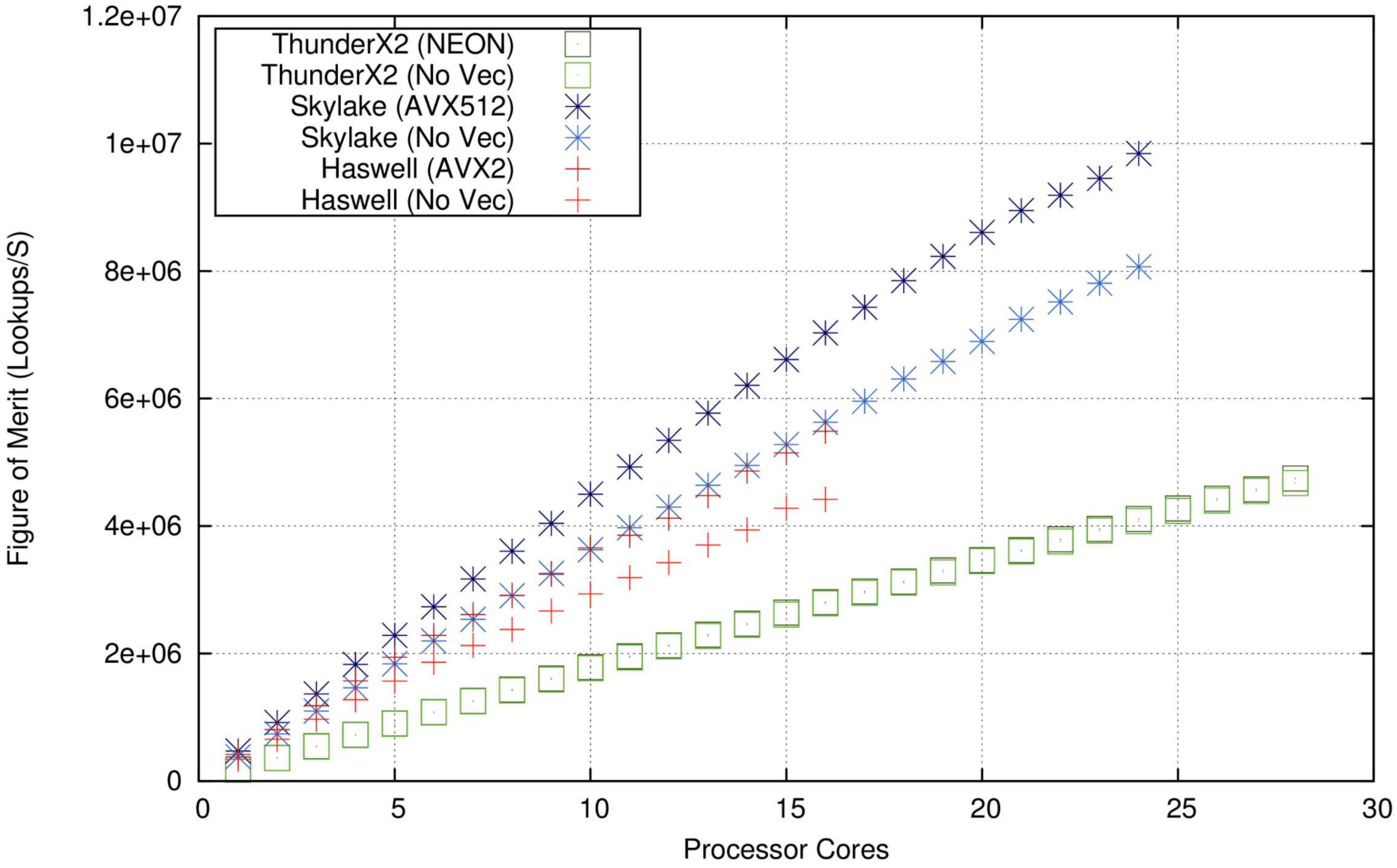


TX2 1.22x more performant than Haswell at 28 PEs; equal at 16 PEs

- Trends are similar, reflecting architectural similarities

Gather/Scatter support in the ISA helps (although only Skylake supports scatter)

Results – XSBench



Vectorization improves performance by approximately 20% on Intel systems

- Gather instructions on Skylake and Haswell prove beneficial to the sparse read patterns while the lack of them on TX2 is a clear detriment

Conclusions

Astra is the first system deployed under the Vanguard program

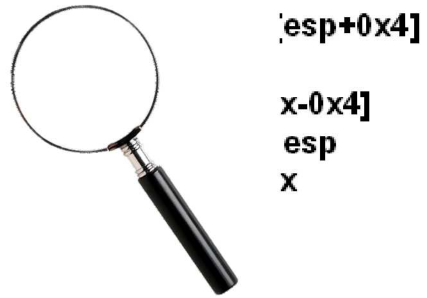
- Vanguard allows the DOE to take risks necessary to ensure a healthy HPC ecosystem for future production mission platforms
  - Increase technology choices
  - Prove ability to run multi-physics production applications at scale on novel hardware

We compared the Astra processors, Marvell ThunderX2, against two Intel offerings

- No single processor was best for the selected kernels and mini-apps
  - Shows that this is a difficult time for the complex workloads in HPC
    - Need a deep understanding of software demands and hardware capabilities
- Demonstrates that the ThunderX2 can deliver exceptional performance for some applications
  - Viable for selection for next-generation supercomputers

Questions?

Motivation



Scientific computing relies on strong server-class processors

- Wide availability of GPUs, many-core processors, and special-purpose accelerators and functional units
  - Majority of calculations still take place on commodity server-class processors
- Many applications still have large regions of serial code
  - Necessitates the need for powerful cores

Two classes of computing platforms for U.S. Department of Energy

- Advanced Technology Systems (ATS)
- Capacity Technology Systems (CTS)

Vanguard: Prototype Systems for Advanced Architectures

Prove viability of advanced technologies for NNSA integrated codes, at scale

Vanguard is a *project*, not a single platform

Expand the HPC ecosystem by developing emerging, yet-to-be-proven, technologies

- Increase technology choices, influence HPC community
- Is the technology viable for future production platforms supporting ASC integrated codes?

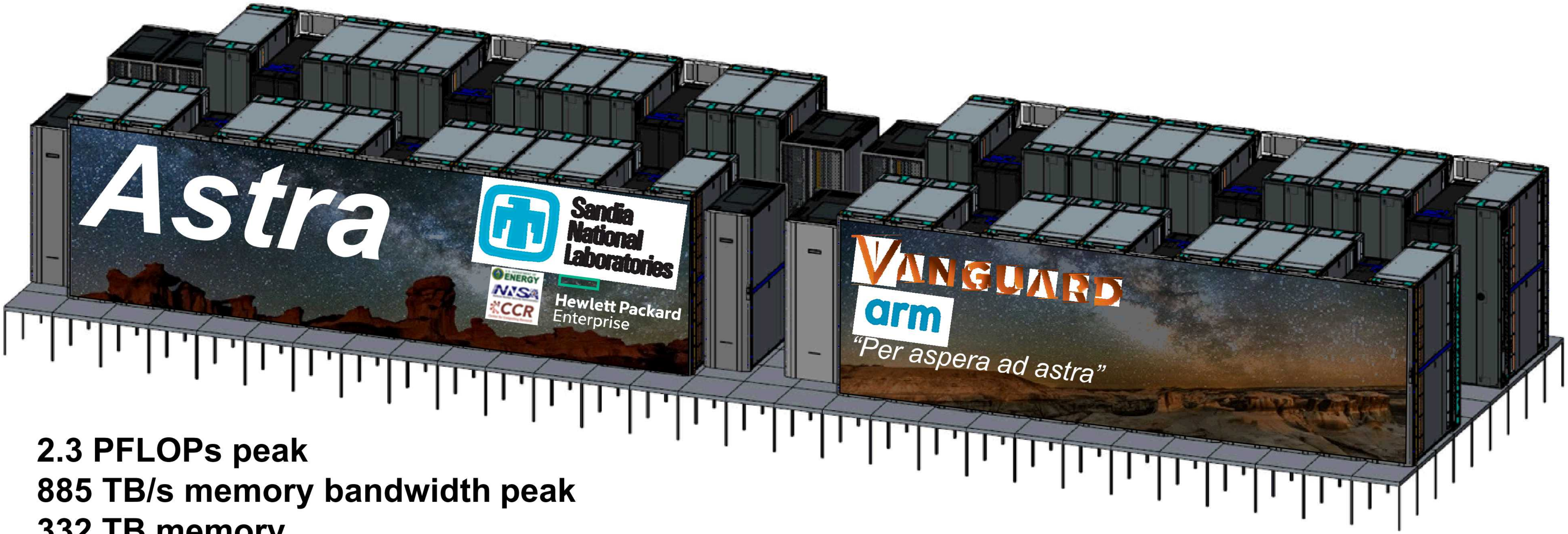
Address hardware and software technologies together

- If hardware technology is new, gaps in the software stack are certain

Buy down risk before commitment on capability/capacity class investment

VANGUARD

per aspera ad astra – Through Difficulties To The Stars



2.3 PFLOPs peak
885 TB/s memory bandwidth peak
332 TB memory
1.2 MW

Demonstrate viability of ARM for U.S. DOE Supercomputing

Astra Architecture

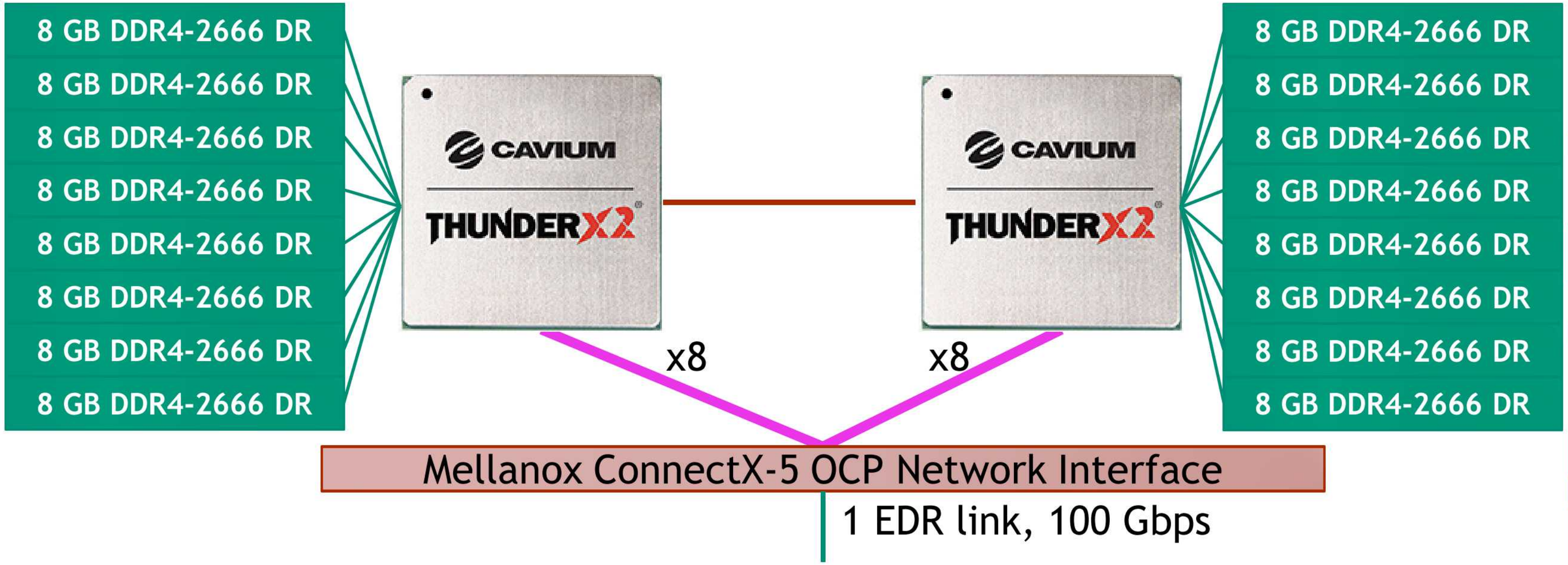
2,592 HPE Apollo 70 compute nodes

- Cavium Thunder-X2 Arm SoC, 28 core, 2.0 GHz
- 5,184 CPUs, 145,152 cores, 2.3 PFLOPs system peak
- 128GB DDR Memory per node (8 memory channels per socket)
- Aggregate capacity: 332 TB, Aggregate Bandwidth: 885 TB/s

Mellanox IB EDR, ConnectX-5

HPE Apollo 4520 All–flash storage, Lustre parallel file-system

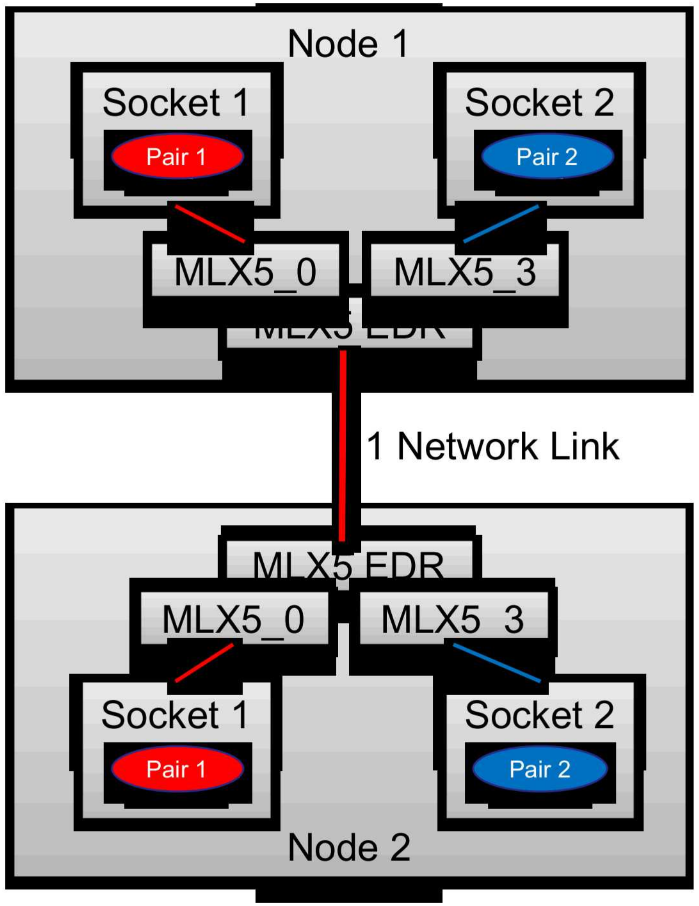
- Capacity: 403 TB (usable)
- Bandwidth 244 GB/s



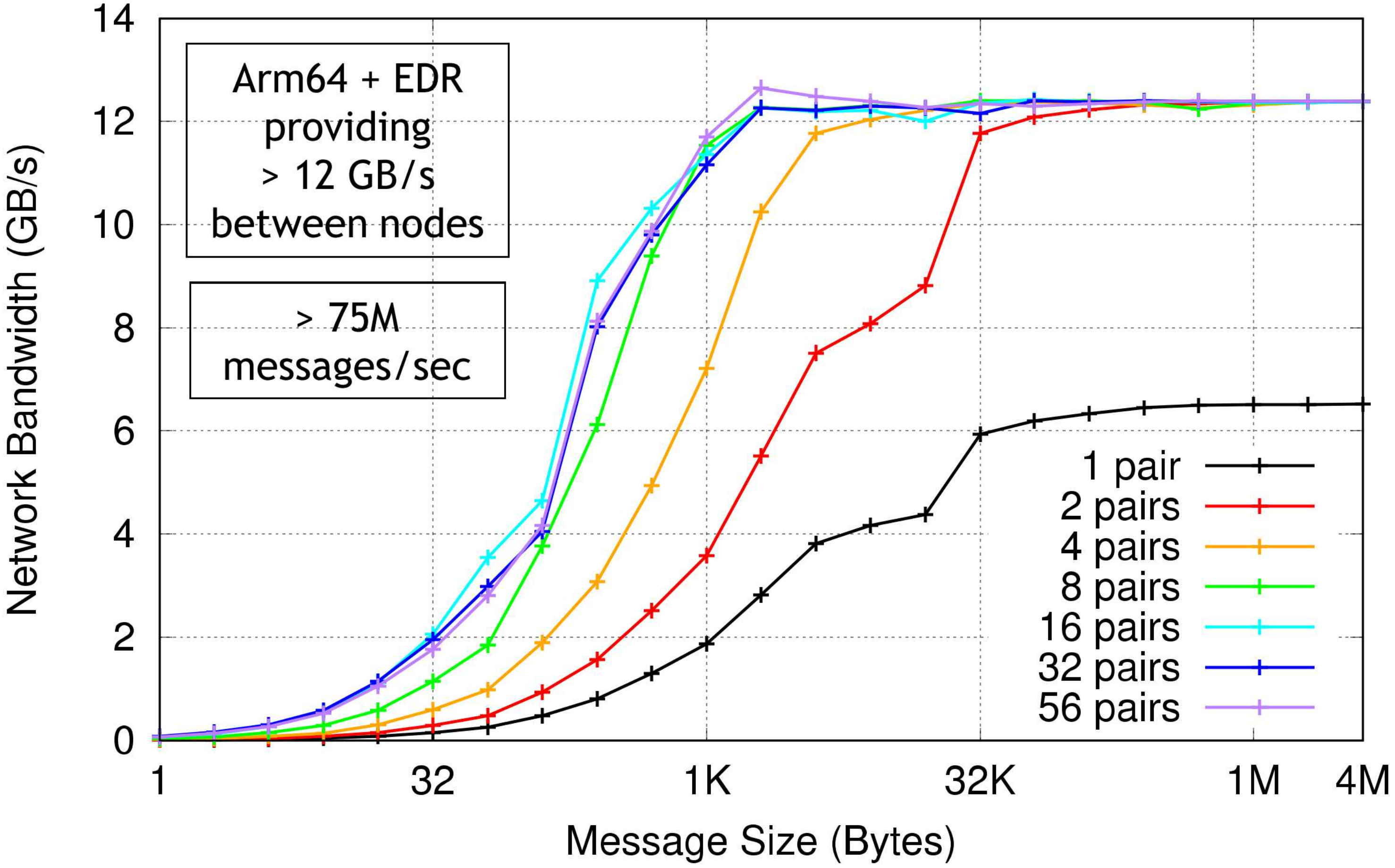
Network Bandwidth

ThunderX2 + Mellanox MLX5 EDR with Socket Direct

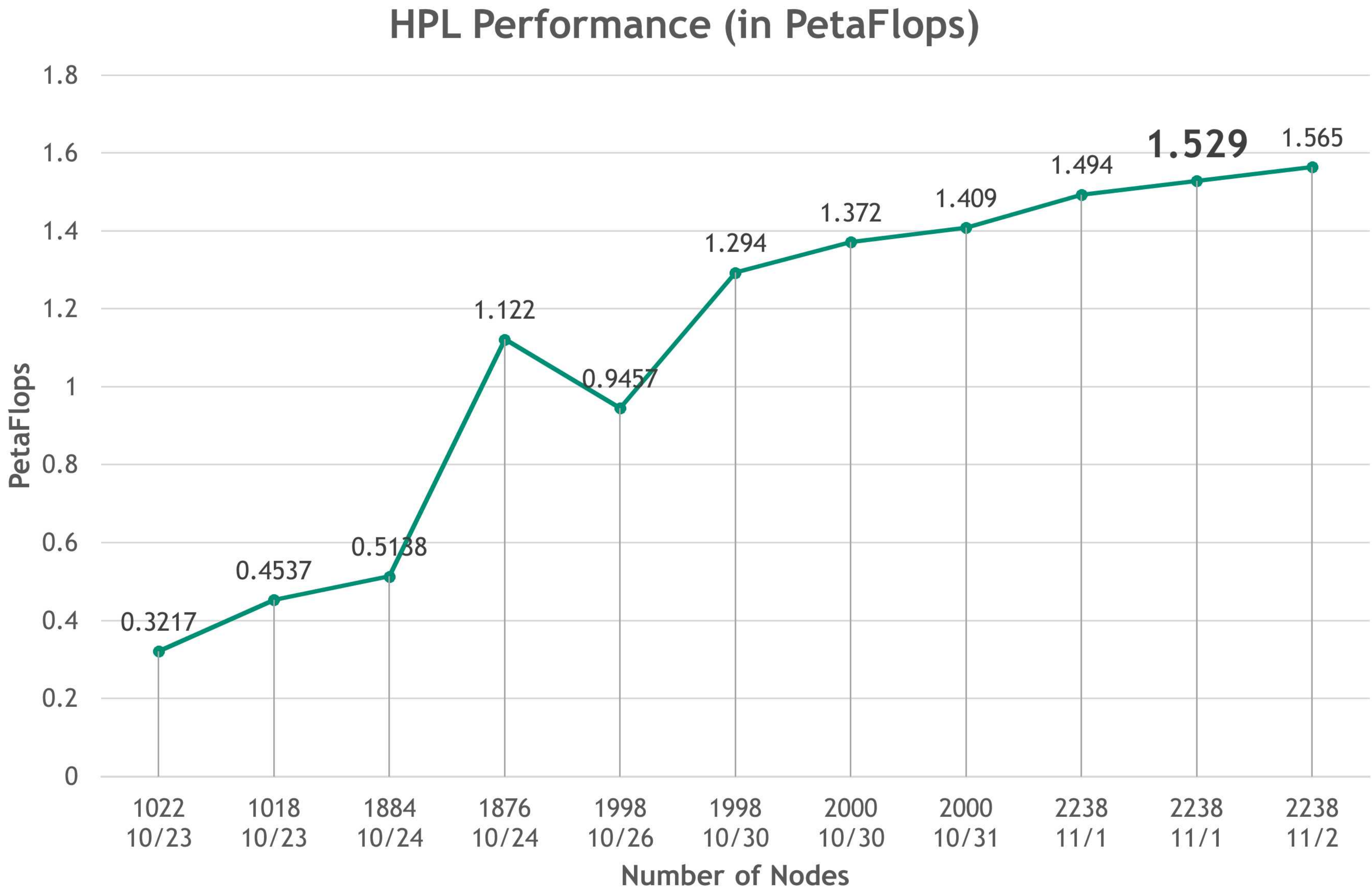
Socket Direct – Each socket has dedicated path to the NIC



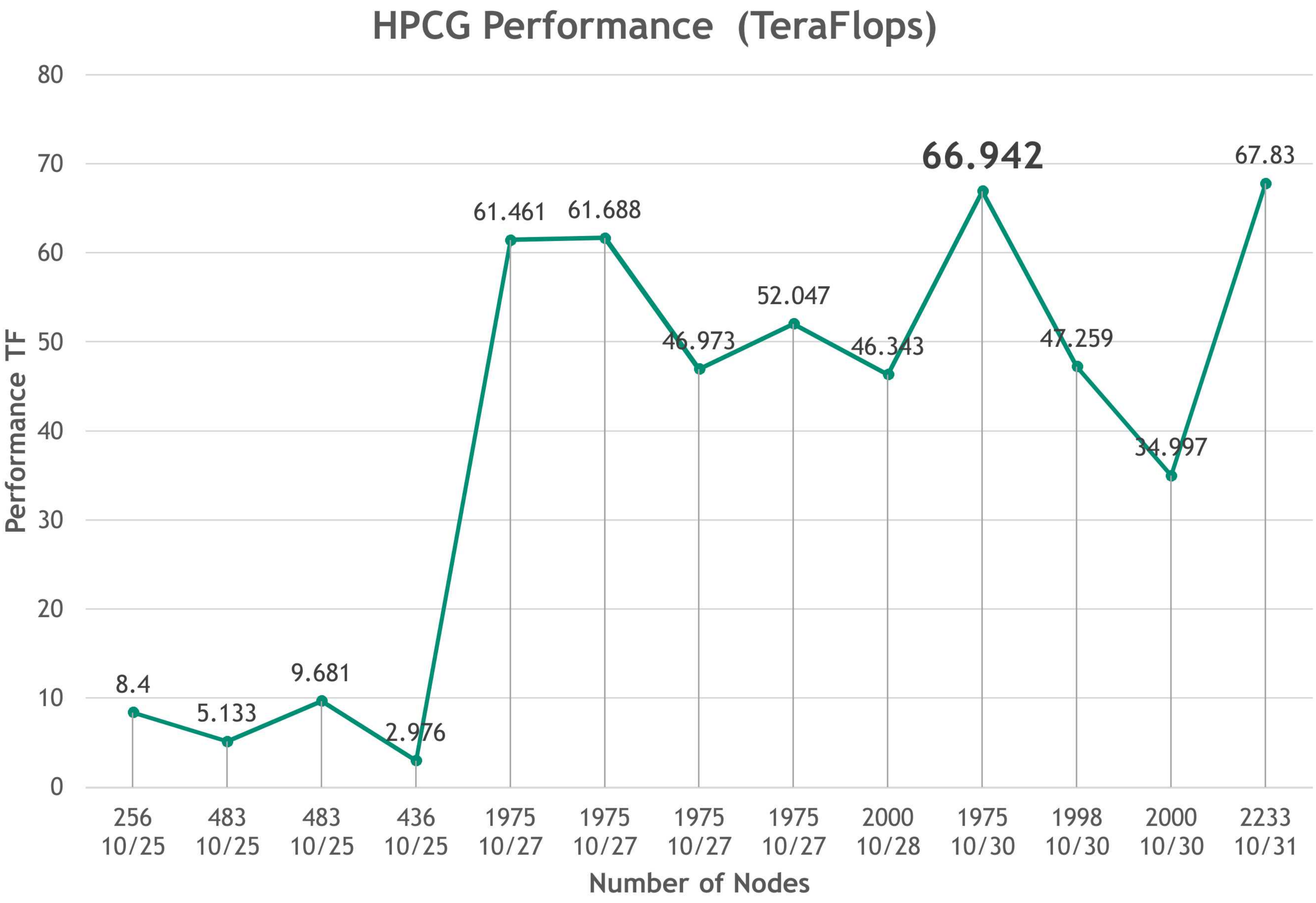
OSU MPI Multi-Network Bandwidth



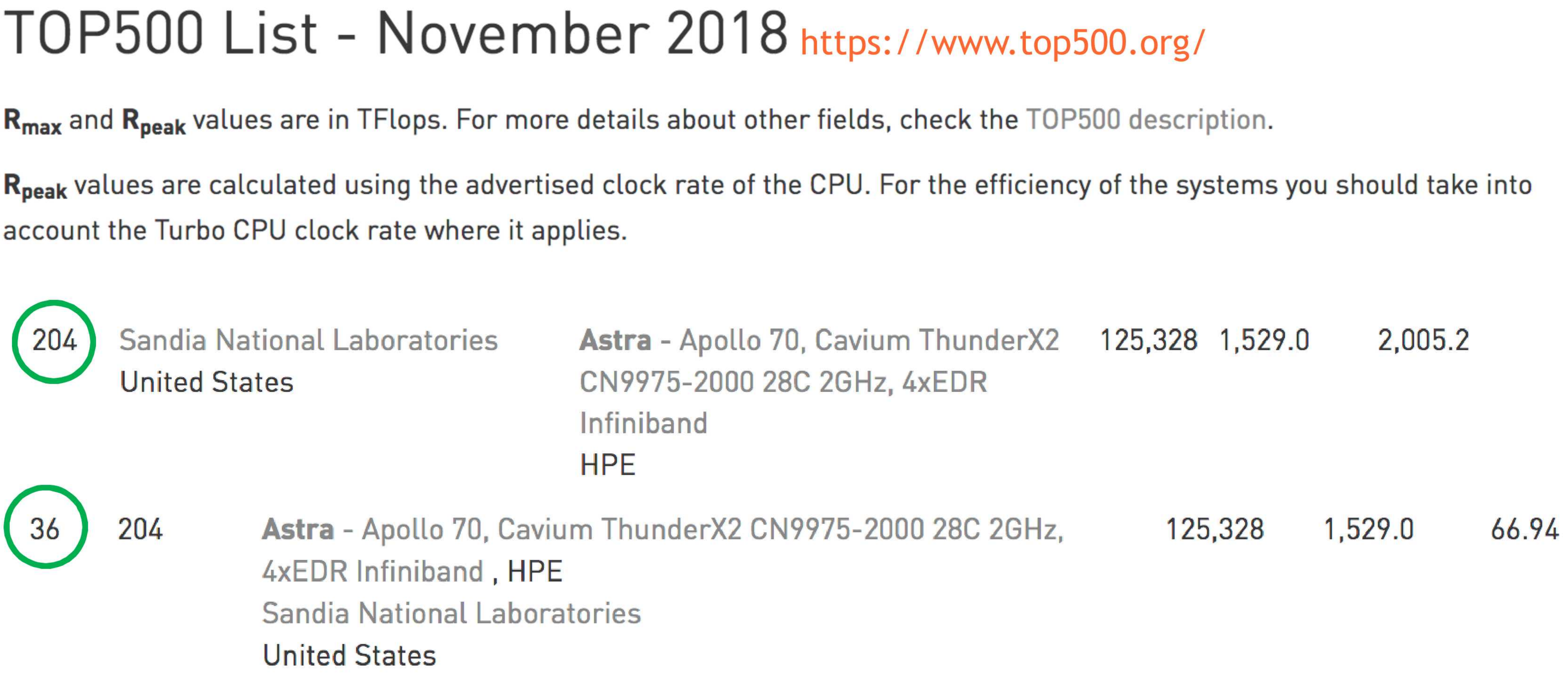
Initial Large Scale Testing and Benchmarks (HPL)



Initial Large Scale Testing and Benchmarks (HPCG)



Top500



TOP500 List - November 2018 https://www.top500.org/

R$_{\textbf{max}}$ and **R**$_{\textbf{peak}}$ values are in TFlops. For more details about other fields, check the TOP500 description.

R$_{\textbf{peak}}$ values are calculated using the advertised clock rate of the CPU. For the efficiency of the systems you should take into account the Turbo CPU clock rate where it applies.

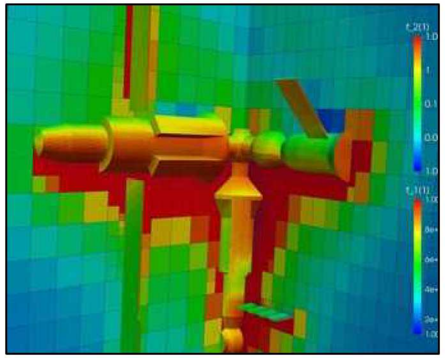
204	Sandia National Laboratories United States	**Astra** - Apollo 70, Cavium ThunderX2 CN9975-2000 28C 2GHz, 4xEDR Infiniband HPE	125,328	1,529.0	2,005.2

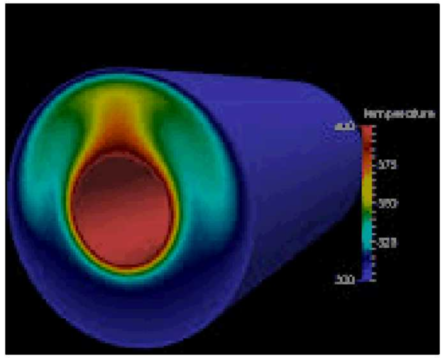
36	204	**Astra** - Apollo 70, Cavium ThunderX2 CN9975-2000 28C 2GHz, 4xEDR Infiniband , HPE Sandia National Laboratories United States	125,328	1,529.0	66.94

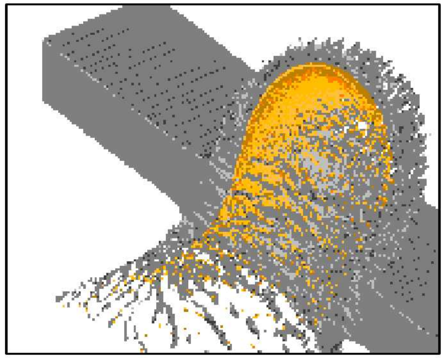
204 on HPL and 36 on HPCG, November 2018

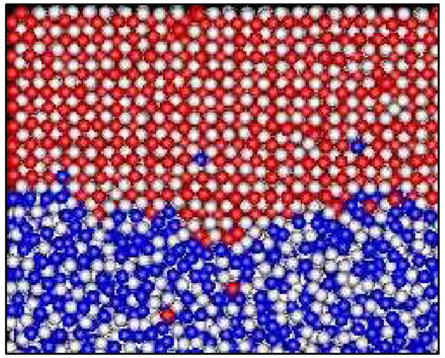
Early Results from Astra

Baseline: Trinity ASC Platform (Current Production), dual-socket Haswell









Monte Carlo	CFD Models	Hydrodynamics	Molecular Dynamics	Linear Solvers
1.60x	1.45x	1.30x	1.42x	1.87x

RM SUPERCOMPUTER

ASTRA

"Per aspera ad astra"

ENERGY NASA

CLICK TO EDIT MASTER TITLE STYLE

Slide Left Blank