COMPUTING INSIGHT UK 2022
1 - 2 DECEMBER 2022
Manchester Central, UK
www.stfc.ac.uk/ciuk
UKRI Science and Technology Facilities Council

Performance of Community Codes on Multi-core Processors

An Analysis of Computational Chemistry and Ocean Modelling Applications.

Martyn Guest, Jose Munoz & Thomas Green

Advanced Research Computing @ Cardiff (ARCCA) &

Supercomputing Wales

Introduction and Overview

- Presentation part of our ongoing assessment of the performance of community codes on multi-core processors.
- Focus here on systems featuring **processors from AMD** (EPYC Milan SKUs) and **Intel** (Ice Lake SKUs) with IB and Cornelis Networks interconnects.
  - Baseline cluster: the Skylake (SKL) **Gold 6148/2.4 GHz** and **AMD EPYC Rome 7502 2.5Gz** cluster – "Hawk" – at Cardiff University.
  - **Five** Intel Xeon Ice Lake clusters, the 32-core Platinum **8358** (2.6 GHz) and **8352Y** (2.2 GHz), the 40-core **8380** (2.3 GHz), 38-core **8368Q** (2.6 GHz), 36-core **8360Y** (2.4GHz) plus other Cascade Lake & Cascade Lake-AP systems.
  - **Four** AMD EPYC Milan clusters featuring the 64-core **7713** (2.0 GHz) and **7773X** (2.2 GHz) and the 32-core **7543** (2.8 GHz) and **7573X** (2.8 GHz).
  - Consider performance of both synthetic and **end-user applications**. Latter include molecular simulation (**DL_POLY, AMBER**), materials modelling (**CASTEP, VASP**), & electronic structure (**GAMESS-UK**), plus representative ocean modelling codes including **NEMO** and **FVCOM**.
  - Scalability analysis by **processing elements (cores)** and by **nodes** (ARM Performance Reports). Baselined against **P100** & **V100** NVIDIA GPUs.

Methodology and Approach

1. Provide guidance based on evaluating performance that a **standard user** would experience on the systems
2. Target performance regime – **mid-range clusters**. No real effort invested in optimising the applications having used standard implementations when available
3. All benchmarks run on systems in general production i.e. not dedicated to this exercise – used standard Slurm job schedulers
4. CIUK'22 preparation again **challenging**. Target to evidence performance across a variety on MPI versions using variants of Intel Parallel Studio XE e.g. 2018/4, 2019/5, 2019/12 and 2020/4 and OneAPI proved over ambitious.
   - Number of problems encountered, particularly on **AMD Milan** systems
   - As noted before, working code using 2019/5 on AMD Rome systems failed on Milan, as did attempts to use earlier compilers / MPI libraries. 2019/12 worked on occasion but still led to failures with codes hanging at arbitrary core counts
   - **Intel oneapi resolved many of these** issues. But issues remain with performance compared to earlier variants of Intel Parallel Studio XE. e.g., a major decline in both VASP and CASTEP performance on AMD Rome when moving from "mpi/intel/2018/2" to "mpi/intel/2020/2"

AMD EPYC Milan multi-chip package

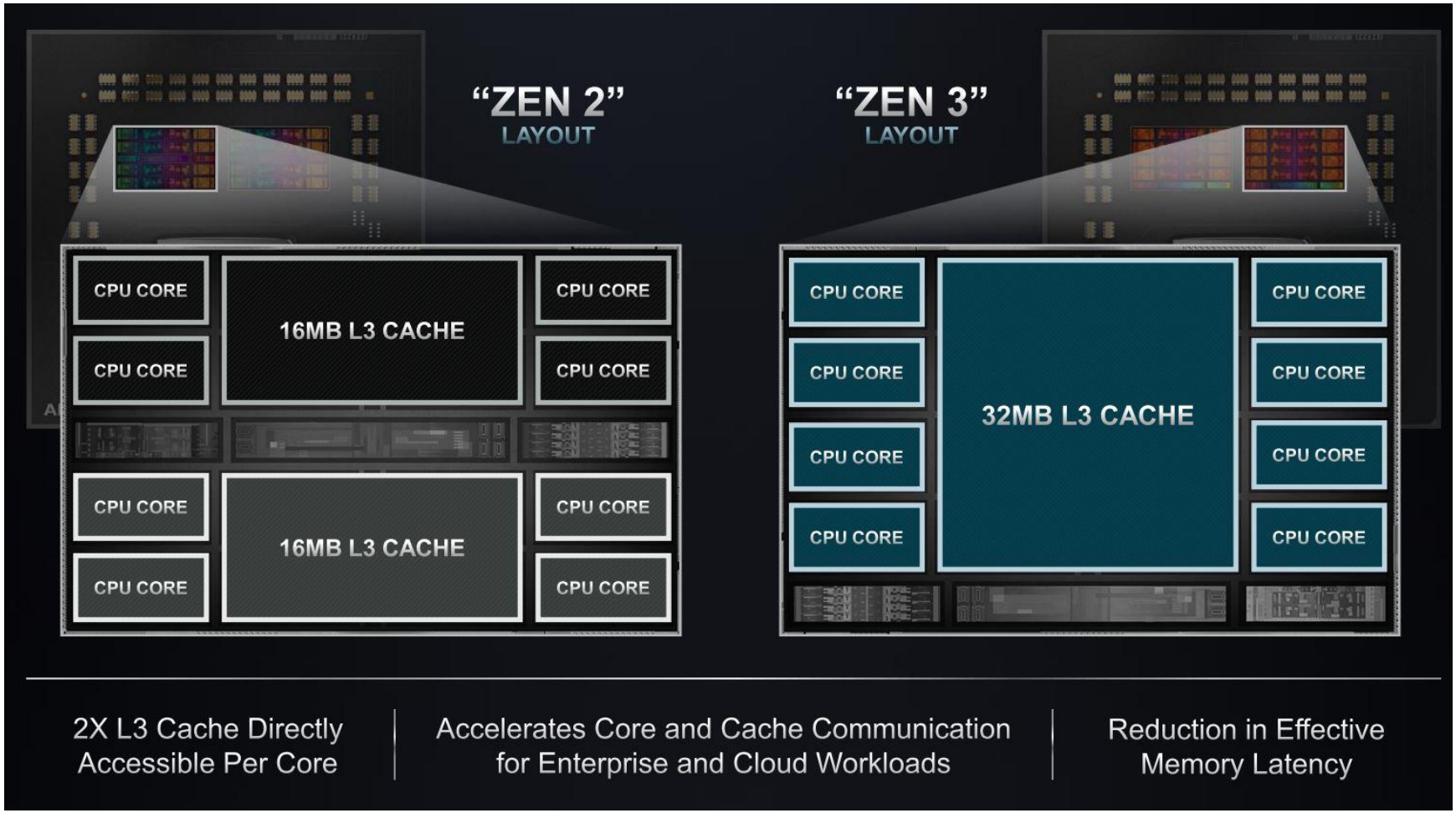


Figure. AMD EPYC Milan multi-chip package.

AMD EPYC Milan-X: Stacked L3 cache

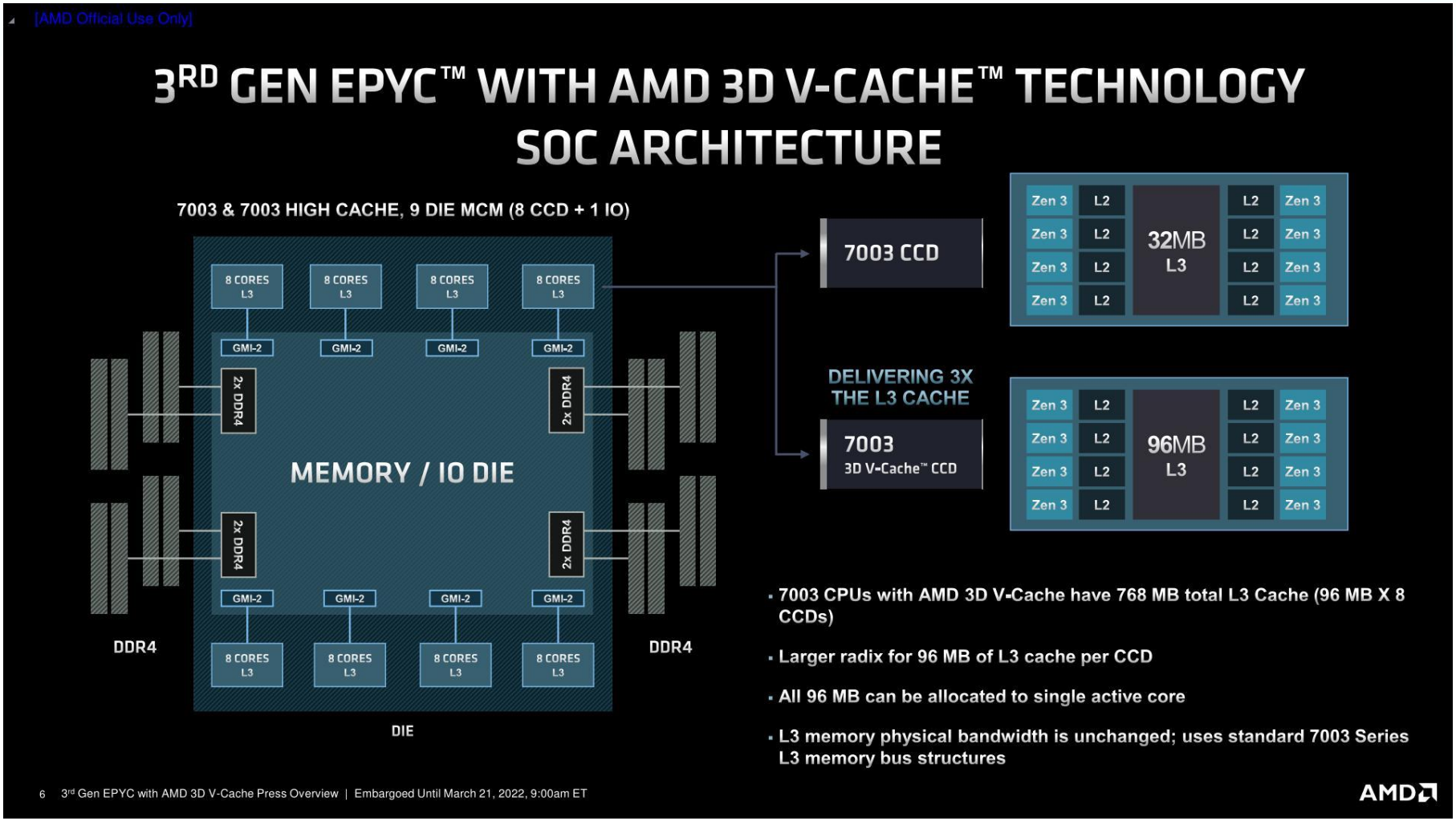
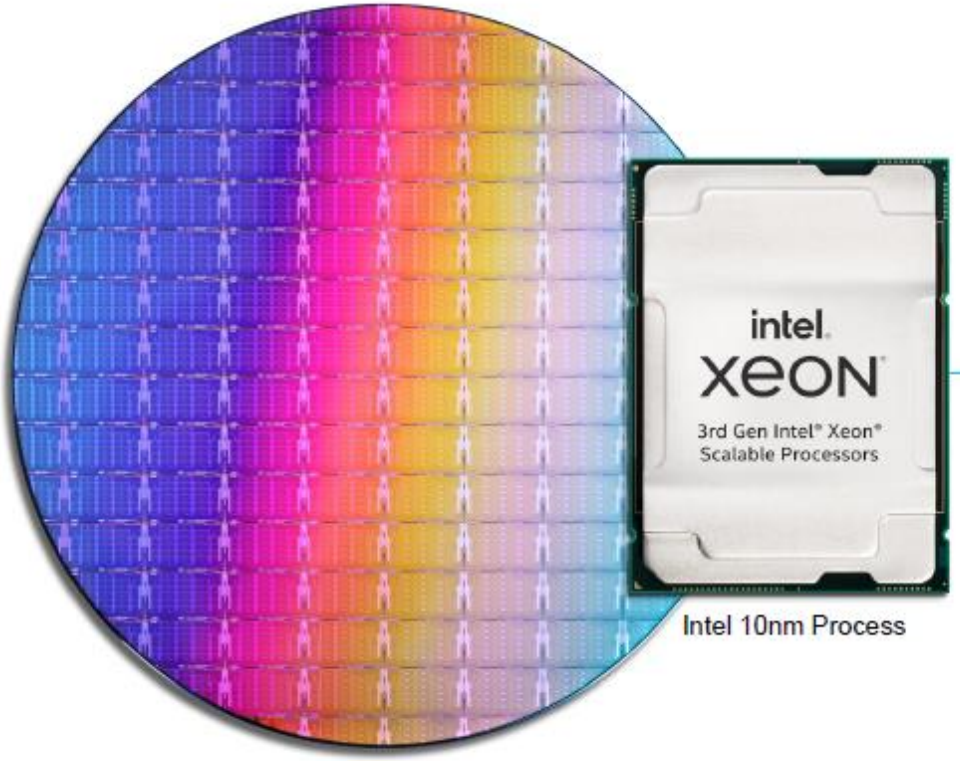


Figure. Milan-X: Upgraded version of Milan using the stacked L3 cache packaging technology. Will use added L3 cache on each chiplet, creating a processor with a total 768 MB of L3 cache e.g., AMD Milan 32c 7573X/2.8 GHz.

Intel 3rd Generation Xeon Scalable Processors

3rd Gen Intel® Xeon® Scalable processors

Performance made flexible



Up to 40 cores
per processor

20% IPC improvement
28 core, ISO Freq, ISO compiler

1.46x average performance increase
Geomean of Integer, Floating Point, Stream Triad, LINPACK
8380 vs. 8280

1.74x AI inference increase
8380 vs. 8280 BERT

2.65x average performance increase
vs. 5-year-old system
8380 vs. E5-2699v4

Intel Xeon Cascade Lake and Ice Lake

	Cascade Lake (per core)	Ice Lake (per core)
Out-of-order Window	224	352
In-flight Loads + Stores	72 + 56	128 + 72
Scheduler Entries	97	160
Register Files – Integer + FP	180 + 168	280 +224
Allocation Queue	64/thread	70/thread; 140/1 thread
L1D Cache (KB)	32	48
L2 Unified TLB (STLB)	1.5K	2K
STLB-IG Page support	**16**	**1024 (shared w/4K)**
STLB-IG Page support	**16**	**1024 (shared w/4K)**
Mid-level Cache (MB)	1	1.25

Performance of Computational Chemistry and Ocean Modelling Codes

Systems, Software and Installation

Baseline Cluster System

Supercomputing Wales "Hawk" Cluster Configuration	
"Phase-1" - Intel Skylake Partition	201 nodes, totalling 8,040 cores, 46.080 TB total memory. • CPU: 2 x Intel(R) Xeon(R) Skylake Gold 6148 CPU @ 2.40GHz with 20 cores each; RAM: 192 GB, 384GB on high memory and GPU nodes; GPU: 26 x nVidia P100 GPUs with 16GB of RAM on 13 nodes. • Mellanox IB/EDR infiniband interconnect.
"Phase-2" AMD Rome Partition	64 nodes, totalling 4,096 cores, 32 TB total memory. • CPU: 2 x AMD EPYC Rome 7502 CPU @ 2.50GHz with 32 cores each; RAM: 512 GB, and GPU nodes; GPU: 30 x nVidia V100 GPUs with 16GB of RAM on 15 nodes
Researcher Funded Partitions	• 4,616 cores – Intel Skylake dedicated researcher expansion • 2,064 cores – Intel Broadwell and Haswell Raven migrated sub-system nodes

The available compute hardware is managed by the **Slurm job scheduler** and organised into 'partitions' of similar type/purpose.

Intel Xeon Cascade Lake Clusters

Cluster / Configuration
Dell Zenith cluster at the Dell Technologies HPC & AI Innovation Lab – Intel Xeon sub-systems with Mellanox HDR interconnect fabric running Slurm. Aging systems that were subject to withdrawal from service, impacting on # nodes available.
• Intel **Xeon Gold 6248 Processor / 2.50 GHz**; # of CPU Cores: **20**; # of Threads: 40; Max Turbo Frequency: 3.90 GHz Base Clock: **2.50 GHz;** Cache 27.5 MB**;** Default TDP / TDP: 150W; Mellanox HDR **200Gb/s**
• Intel **Xeon Platinum 8280 Processor / 2.70 GHz**; # of CPU Cores: **28**; # of Threads: 56; Max Turbo Frequency: 4.00 GHz Base Clock: **2.70 GHz;** Cache 38.5 MB**;** Default TDP / TDP: 205W; Mellanox HDR **200Gb/s**
Cascade Lake-AP cluster at Intel HPC Laboratory with Cornelis OPE fabric running Bright release 8.1, optane filesystem.
• 48 **CLX-AP nodes** (Cascade Lake Advanced Performance**) 9242 processors / 2.3 GHz**; # of CPU Cores: **48**; # of Threads: 96; Max Turbo Frequency: 3.80 GHz Base Clock: **2.30 GHz;** Cache 71.5 MB**;** Default TDP / TDP: 350W

Intel Xeon Ice Lake Clusters

Cluster / Configuration
Dell Zenith cluster at the Dell Technologies HPC & AI Innovation Lab – Intel Xeon sub-systems with Mellanox HDR interconnect fabric running Slurm
• 50 nodes × Intel **Xeon Platinum 8358 Processor / 2.60 GHz**; # of CPU Cores: **32**; # of Threads: 64; Max Turbo Frequency: 3.40 GHz Base Clock: **2.60 GHz;** Cache 48 MB**;** Default TDP / TDP: 250W; **Mellanox HDR 200Gb/s**
• 70 nodes × Intel **Xeon Platinum 8352Y Processor / 2.20 GHz**; # of CPU Cores: **32**; # of Threads: 64; Max Turbo Frequency: 3.40 GHz Base Clock: **2.20 GHz;** Cache 48 MB**;** Default TDP / TDP: 205W; **Mellanox HDR 200Gb/s**
Ice Lake clusters at Intel's OpenHPC Laboratory with Cornelis OPE fabric running Bright release 8.1 and optane filesystem.
• 4 nodes × Intel **Xeon Platinum 8368Q Processor / 2.60 GHz**; # of CPU Cores: **38**; # of Threads: 76; Max Turbo Frequency: 3.70 GHz Base Clock: **2.60 GHz;** Cache 57 MB**;** Default TDP / TDP: 270W; **Cornelis OPE**
• 4 nodes × Intel **Xeon Platinum 8360Y Processor / 2.40 GHz**; # of CPU Cores: **36**; # of Threads: 72; Max Turbo Frequency: 3.50 GHz Base Clock: **2.40 GHz;** Cache 54 MB**;** Default TDP / TDP: 270W; **Cornelis OPE**
Intel's Endeavour cluster with Cornelis OPE fabric running Slurm
• 8 nodes × Intel **Xeon Platinum 8380 Processor / 2.30 GHz**; # of CPU Cores: **40**; # of Threads: 80;
• 10 nodes × Intel **Xeon Platinum 8360Y Processor / 2.40 GHz**; # of CPU Cores: **36**; # of Threads: 72

AMD EPYC Milan Clusters

Cluster / Configuration
Dell Minerva cluster at the Dell Technologies HPC & AI Innovation Lab – AMD EPYC Rome and Milan sub-systems with Mellanox HDR interconnect fabric running Slurm
• **4 nodes × AMD EPYC Milan 7543 / 2.80 GHz**; # of CPU Cores: 32; # of Threads: 64; Max Boost Clock: 3.7 GHz Base Clock: **2.80 GHz; L3** Cache 256 MB**;** Default TDP / TDP: 225W; Mellanox HDR-100 **200Gb/s**
• **6 nodes × AMD EPYC Milan 7573X / 2.80 GHz**; # of CPU Cores: 32; # of Threads: 64; Max Boost Clock: 3.6 GHz Base Clock: **2.80 GHz; L3 Cache 768 MB;** Default TDP / TDP: 280W; Mellanox HDR-100 **200Gb/s**
• **170 nodes × AMD EPYC Milan 7713 / 2.00 GHz**; # of CPU Cores: 64; # of Threads: 128; Max Boost Clock: 3.675 GHz Base Clock: **2.00 GHz; L3** Cache 256 MB**;** Default TDP / TDP: 225W; Mellanox HDR-100 **200Gb/s**
• **4 nodes × AMD EPYC Milan 7763 / 2.45 GHz**; # of CPU Cores: 64; # of Threads: 128; Max Boost Clock: 3.5 GHz Base Clock: **2.45 GHz; L3** Cache 256 MB**;** Default TDP / TDP: 280W; Mellanox HDR-100 **200Gb/s**
SPARTAN cluster at the Atos HPC, AI & QLM Benchmarking Centre – AMD EPYC Rome system with Mellanox ConnectX-6 HDR100 interconnect fabric
• **240 × AMD EPYC Rome 7742 / 2.25 GHz;** # of CPU Cores: 64; # of Threads: 128; Max Boost Clock: 3.35 GHz Base Clock**: 2.25 GHz; L3** Cache 256 MB; Default TDP / TDP: 225W; **Mellanox ConnectX-6 HDR 100 InfiniBand:** Memory: 256GB DDR4 2677MHz RDIMMs per node: **DDN lustre 7990 Storage, NFS**

The Performance Benchmarks

- The ***Test suite*** comprises both **synthetics & end-user applications**. Synthetics limited to **IMB** benchmarks (*http://software.intel.com/en-us/articles/intel-mpi-benchmarks*) and **STREAM**
- Variety of “open source” & commercial end-user application codes:

DL_POLY and AMBER (molecular dynamics)

VASP and CASTEP (ab initio Materials properties)

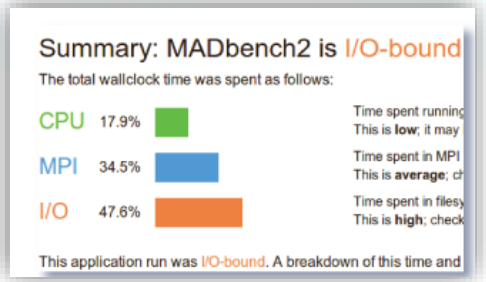
GAMESS-UK (molecular electronic structure)

FVCOM and NEMO (ocean modelling codes)

- These stress various aspects of the architectures under consideration and should provide a level of insight into why particular levels of performance are observed e.g., ***memory bandwidth and latency, node floating point performance and interconnect performance (both latency and B/W) and sustained I/O performance***.

Analysis Software - Allinea|ARM Performance Reports

Provides a mechanism to characterize and understand the performance of HPC application runs through a single-page HTML report.



- Based on Allinea MAP's adaptive sampling technology that keeps data volumes collected and **application overhead low**.
- Modest application slowdown (ca. 5%) even with 1000's of MPI processes.
- **Runs on existing codes: a single command added to execution scripts.**
- If submitted through a batch queuing system, then the submission script is modified to load the Allinea module and add the 'perf-report' command in front of the required mpirun command.

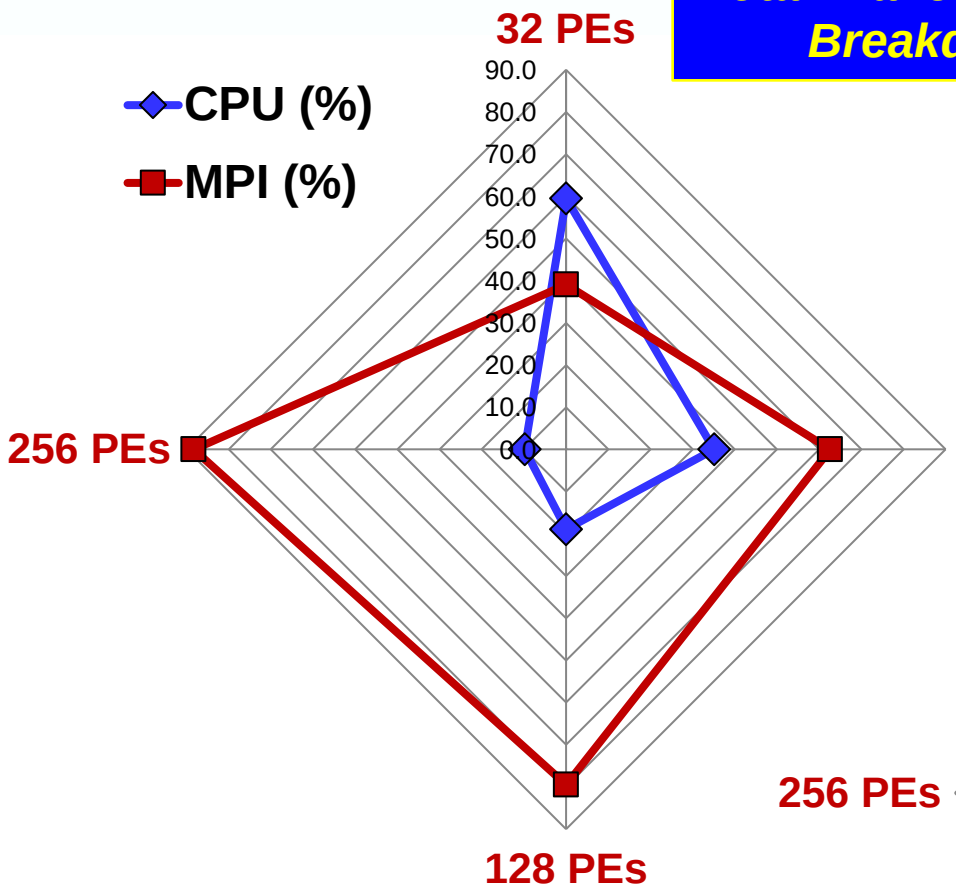
  perf-report mpirun $code

- ***A Report Summary:*** This characterizes how the application's wallclock time was spent, broken down into CPU, MPI and I/O
- All examples from the **Hawk Cluster (SKL Gold 6148 / 2.4GHz)**

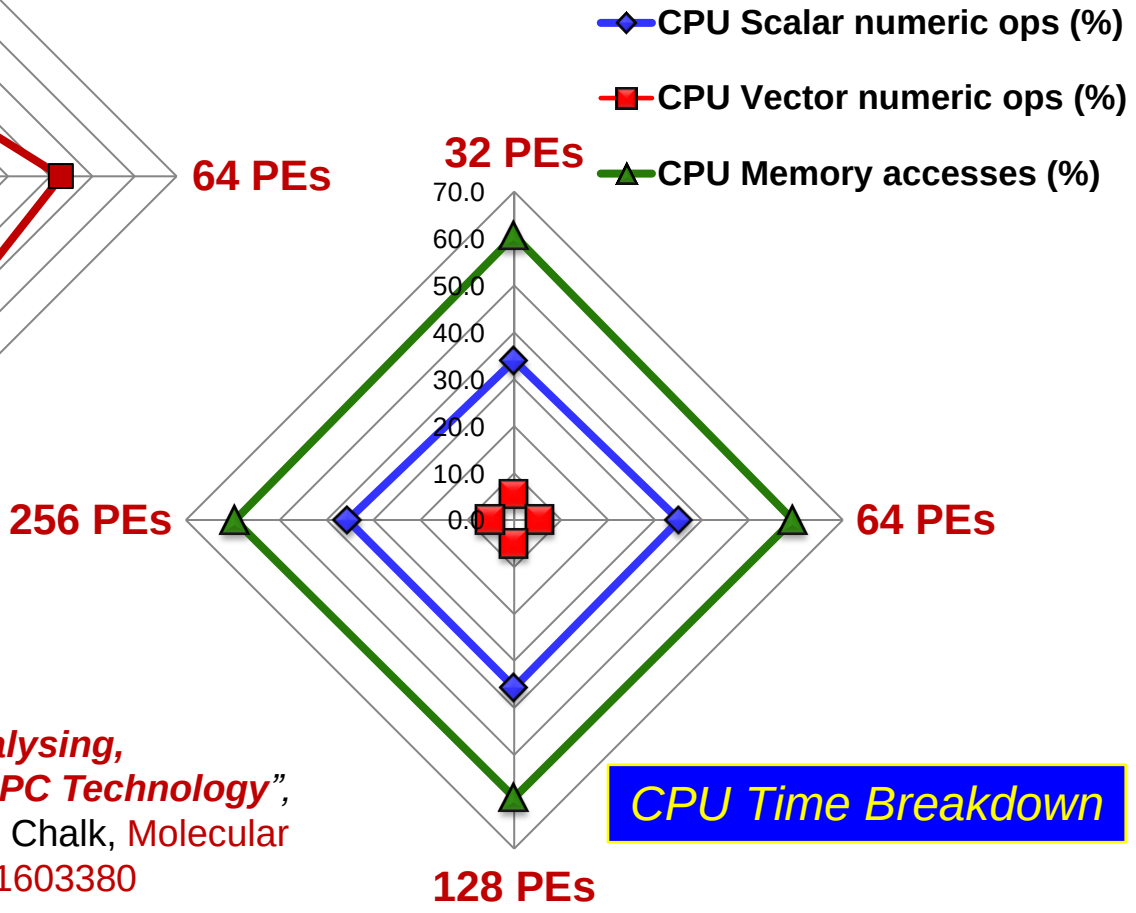
DLPOLY4 – Performance Report

Total Wallclock Time Breakdown

Performance Data (32-256 PEs)



Smooth Particle Mesh Ewald Scheme



CPU Time Breakdown

"DL_POLY - A Performance Overview. Analysing, Understanding and Exploiting available HPC Technology", Martyn F Guest, Alin M Elena and Aidan B G Chalk, Molecular Simulation, (2019) 10.1080/08927022.2019.1603380

EPYC - Compiler and Run-time Options

STREAM (AMD Minerva Cluster):

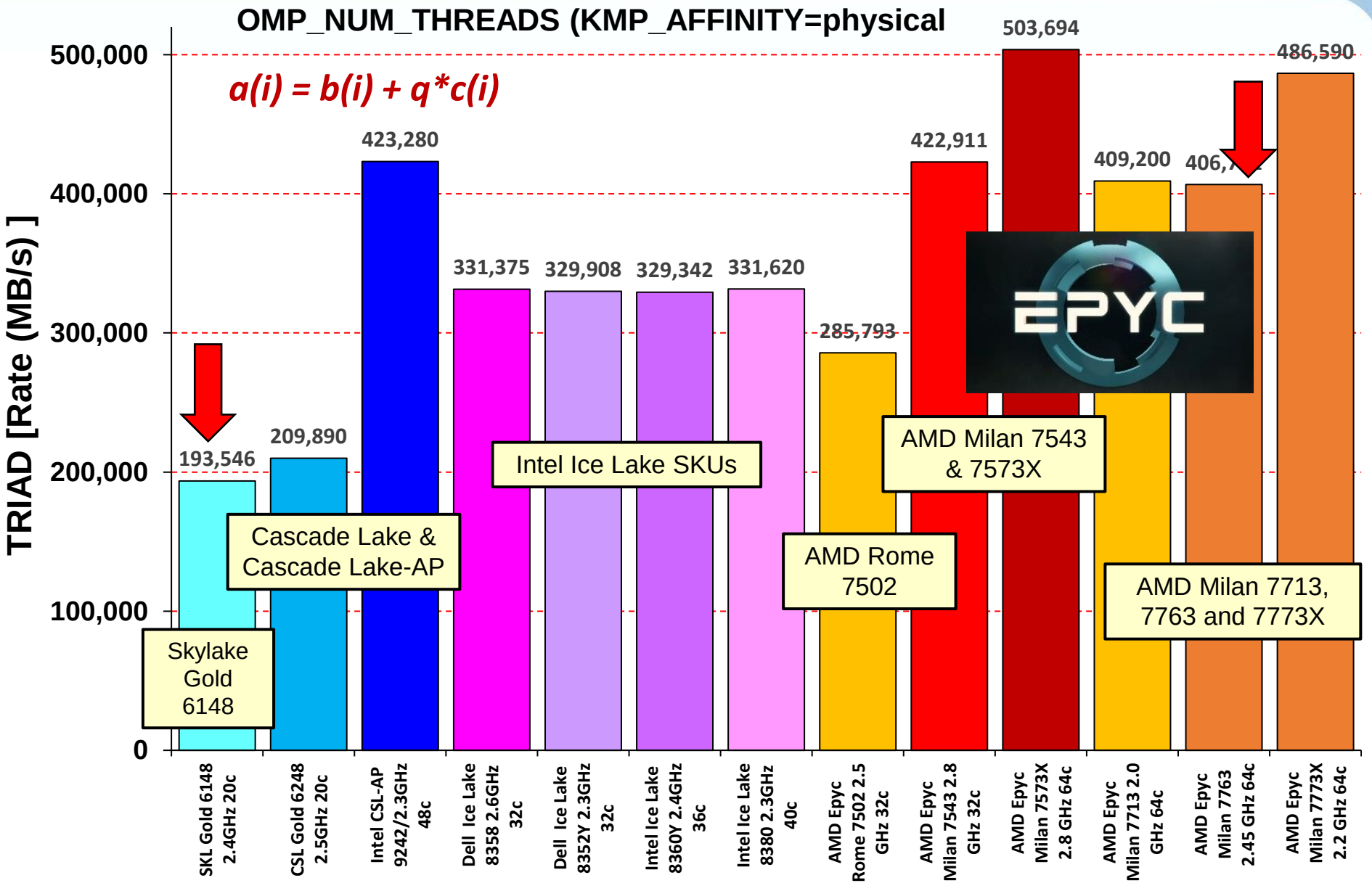
```
icc stream.c -DSTATIC -Ofast -march=core-avx2 -DSTREAM_ARRAY_SIZE=2500000000 -
DNTIMES=10 -mcmodel=large -shared-intel -restrict -qopt-streaming-stores always
-o streamc.Rome
icc stream.c -DSTATIC -Ofast -march=core-avx2 -qopenmp -
DSTREAM_ARRAY_SIZE=2500000000 -DNTIMES=10 -mcmodel=large -shared-intel -restrict
-qopt-streaming-stores always -o streamcp.Rome

# Version of Intel compiler to use and way to source it
source /opt/intel/compilers_and_libraries_2020.2.254/linux/bin/compilervars.sh -
ofi_internal=1 intel64

# Increasing use of oneAPI: e.g., source /opt/intel/oneapi/setvars.sh

# Use of specific version of Intel MKL, further versions do not allow the setting
of AVX2 on non-Intel processors.
source /opt/intel/compilers_and_libraries_2019.6.324/linux/mkl/bin/mklvars.sh
intel64

# When using IntelMPI on AMD Rome/Milan
export I_MPI_FABRICS=shm:ofi
export I_MPI_SHM=clx_avx2
export FI_PROVIDER=mlx

# On AMD Rome/Milan when using Intel MKL
export MKL_DEBUG_CPU_TYPE=5
```

Compilation:

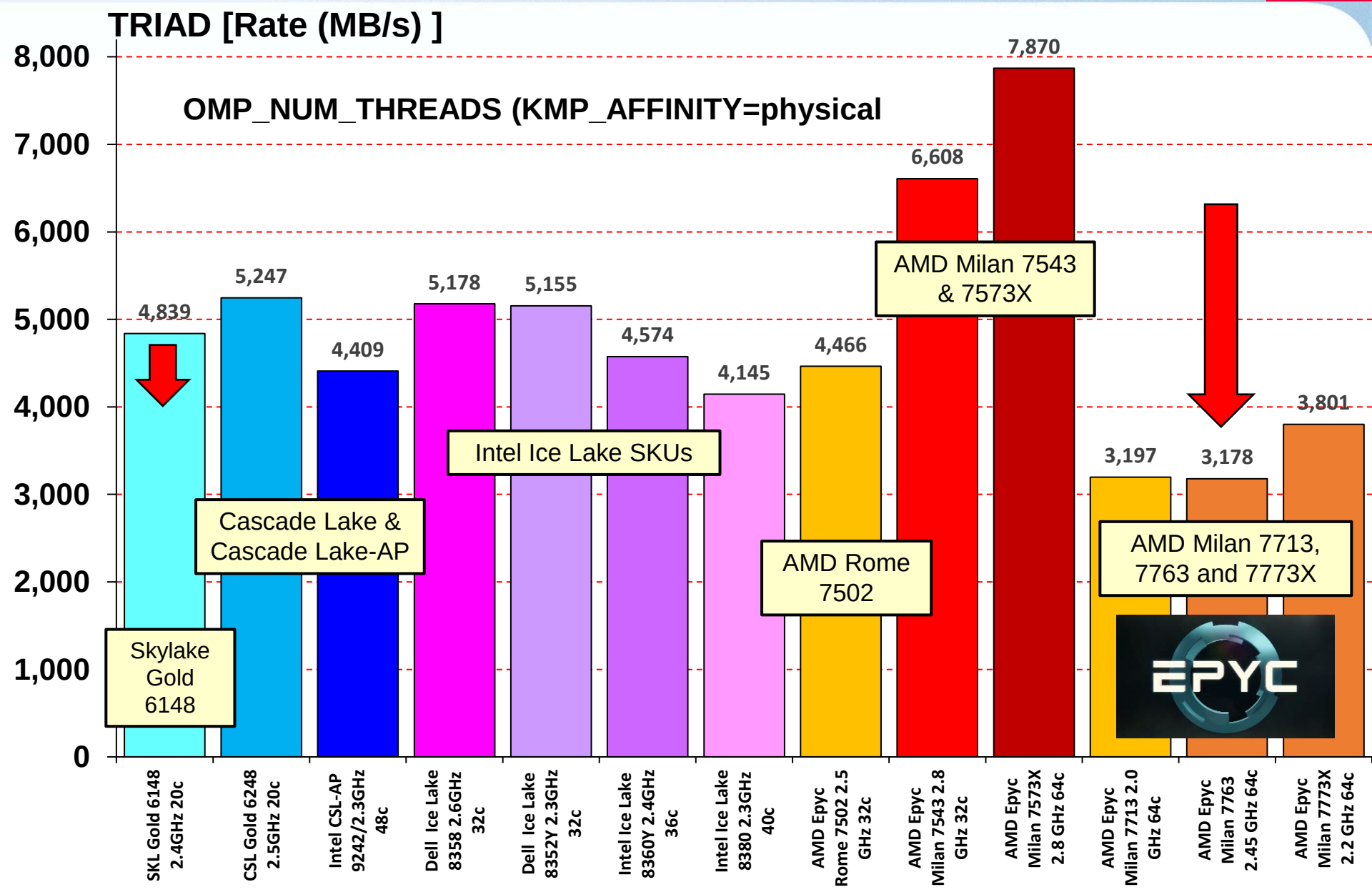
INTEL SKL: –O3 –xCORE-AVX512

AMD EPYC: –O3 –march=core-avx2 -align array64byte -fma -ftz -fomit-frame-pointer

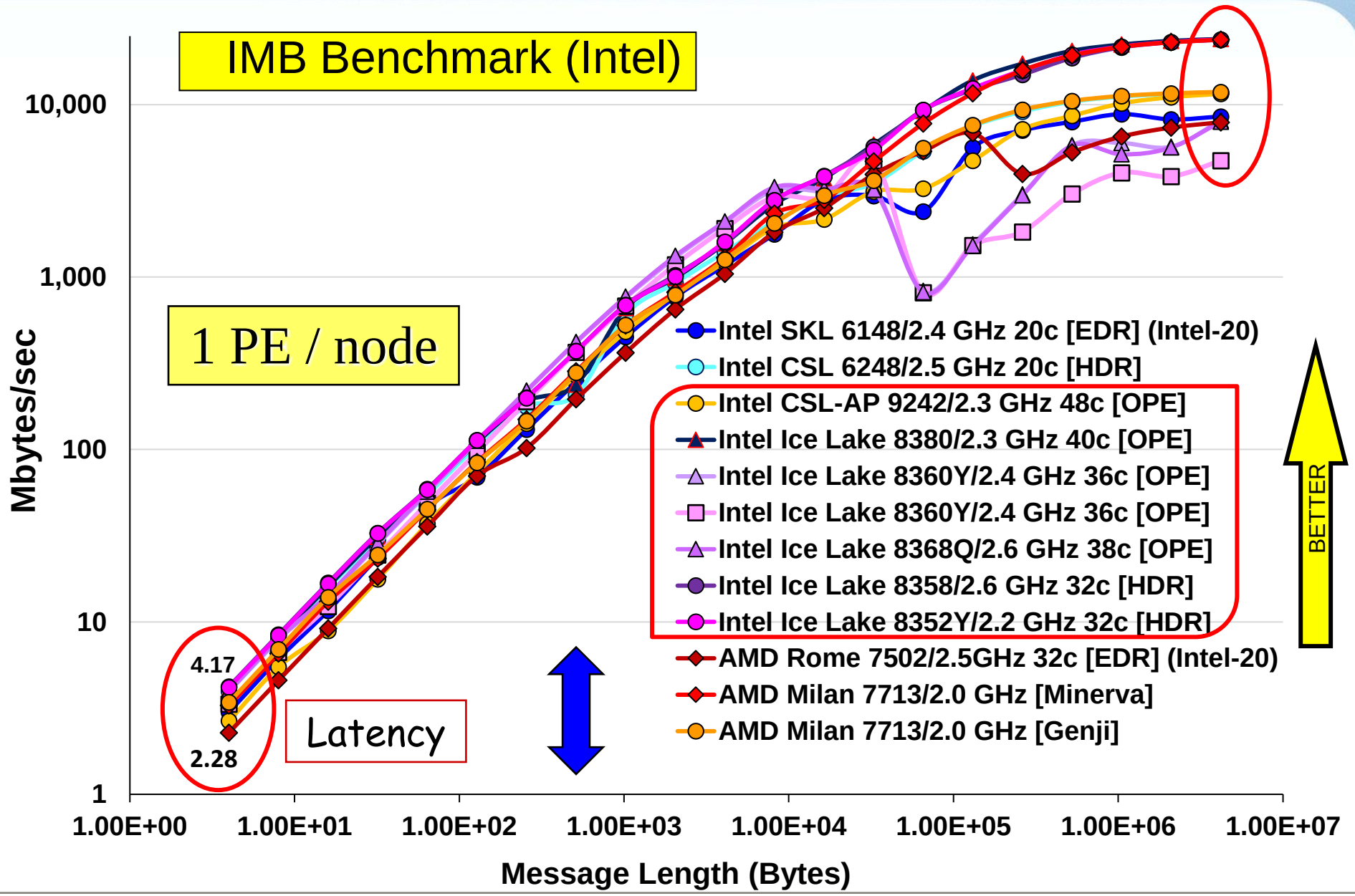
Memory B/W – STREAM performance



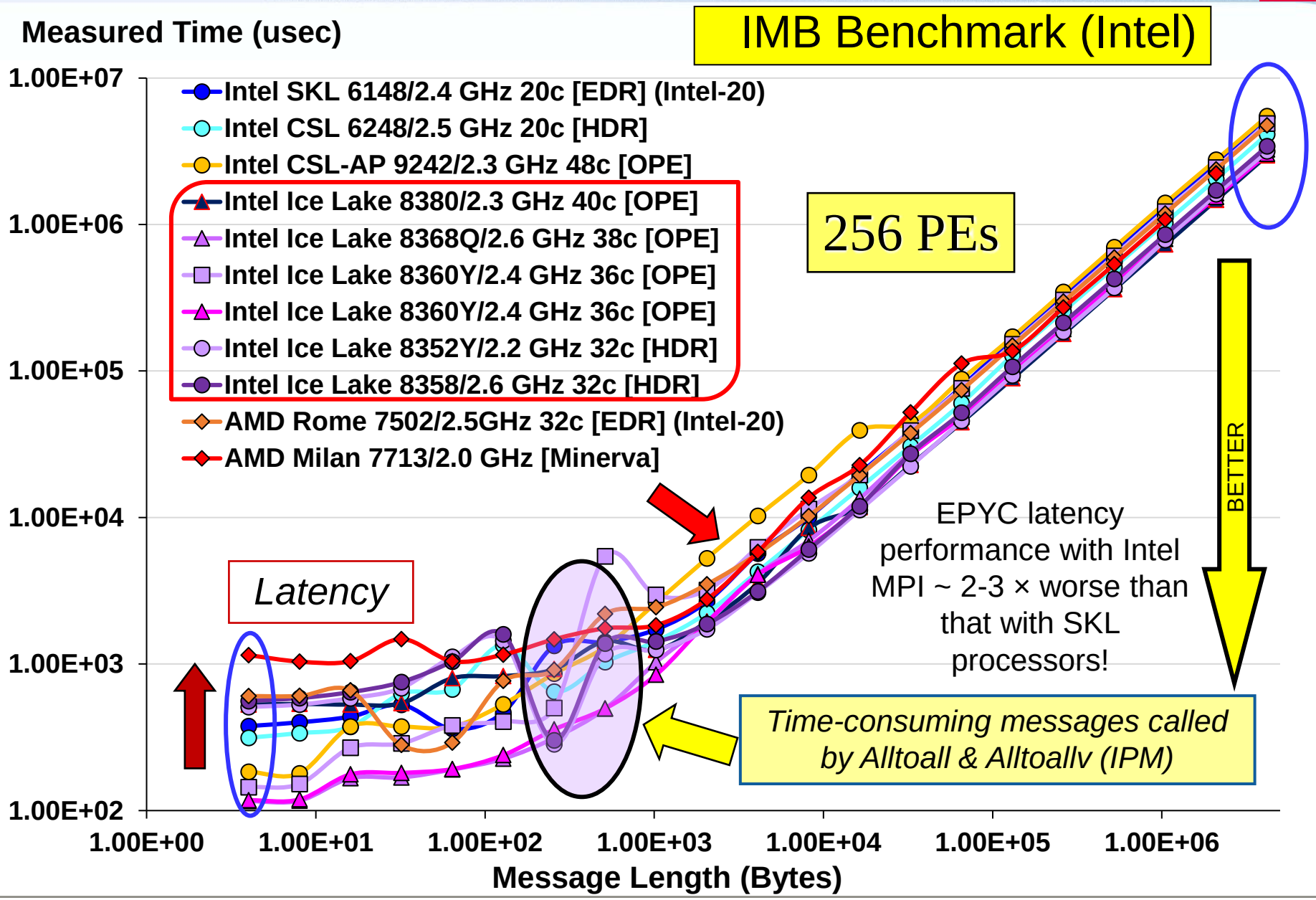
Memory B/W – STREAM / core performance



MPI Performance – PingPong



MPI Collectives – Alltoallv (256 PEs)



Performance Metrics – “Core to Core” & “Node to Node”

- Analysis of performance Metrics across a variety of data sets
  - ❑ “**Core to core**” and “**node to node**” workload comparisons
    - ***Core to core*** comparison i.e. performance for jobs with a fixed number of cores
    - ***Node to Node*** comparison typical of the performance when running a workload (real life production). Expected to reveal the major benefits of increasing core count per socket
  - ❑ Focus on a variety of “**node to node**” and “**core-to-core**” comparisons e.g., :

1	***Hawk - Dell |EMC Skylake Gold 6148 2.4GHz (T) EDR with 40 cores / node***	***AMD EPYC Milan 7713 nodes with128 cores per node. [1-7 nodes]***
2	***Hawk - Dell |EMC Skylake Gold 6148 2.4GHz (T) EDR with 40 cores / node***	***Intel Xeon Platinum Ice Lake 8358 nodes with 64 cores per node. [1-7 nodes]***

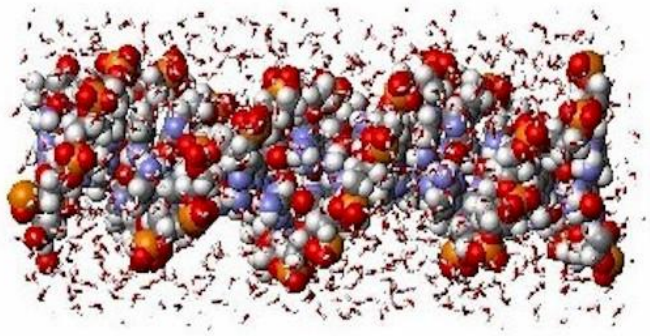
Performance of Computational Chemistry and Ocean Modelling Codes

Molecular Simulation;

1. DL_POLY

Molecular Simulation I. DL_POLY

Molecular Dynamics Codes: AMBER, DL_POLY, CHARMM, NAMD, LAMMPS, GROMACS etc



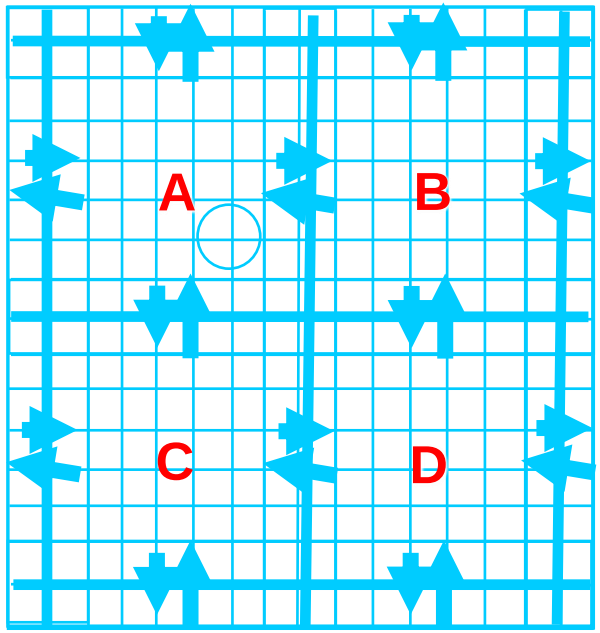
DL_POLY

- Developed as CCP5 parallel MD code by W. Smith, T.R. Forester and I. Todorov
  - UK CCP5 + International user community
  - DLPOLY_classic (replicated data) and DLPOLY_3 & _4 (distributed data – domain decomposition)
- Areas of application:
  - liquids, solutions, spectroscopy, ionic solids, molecular crystals, polymers, glasses, membranes, proteins, metals, solid and liquid interfaces, catalysis, clathrates, liquid crystals, biopolymers, polymer electrolytes.

DL_POLY 4 – Distributed data

Domain Decomposition - Distributed data:

- Distribute atoms, forces across the nodes
  - More memory efficient, can address much larger cases (10^5-10^7)
- Shake and short-ranges forces require only neighbour communication
  - communications scale linearly with number of nodes
- Coulombic energy remains global
  - Adopt **Smooth Particle Mesh Ewald** scheme
    - includes Fourier transform smoothed charge density (reciprocal space grid typically 64x64x64 - 128x128x128)



W. Smith and I. Todorov

Benchmarks

1. NaCl Simulation; 216,000 ions, 200 time steps, Cutoff=12Å

2. Gramicidin in water; rigid bonds + SHAKE: 792,960 ions, 50 time steps

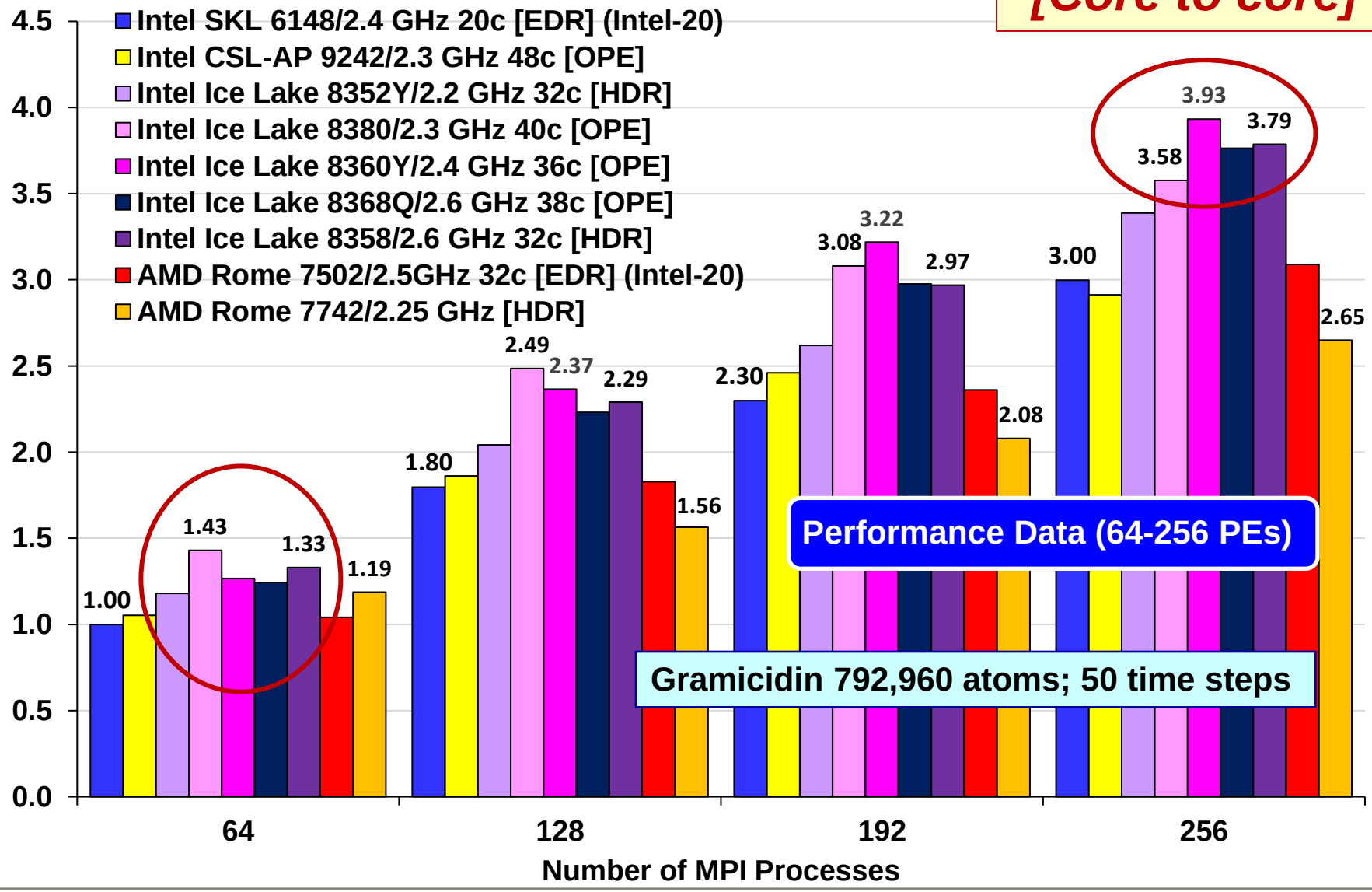
http://www.scd.stfc.ac.uk//research/app/ccg/software/DL_POLY/44516.aspx

DL_POLY 4 – Gramicidin Simulation

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

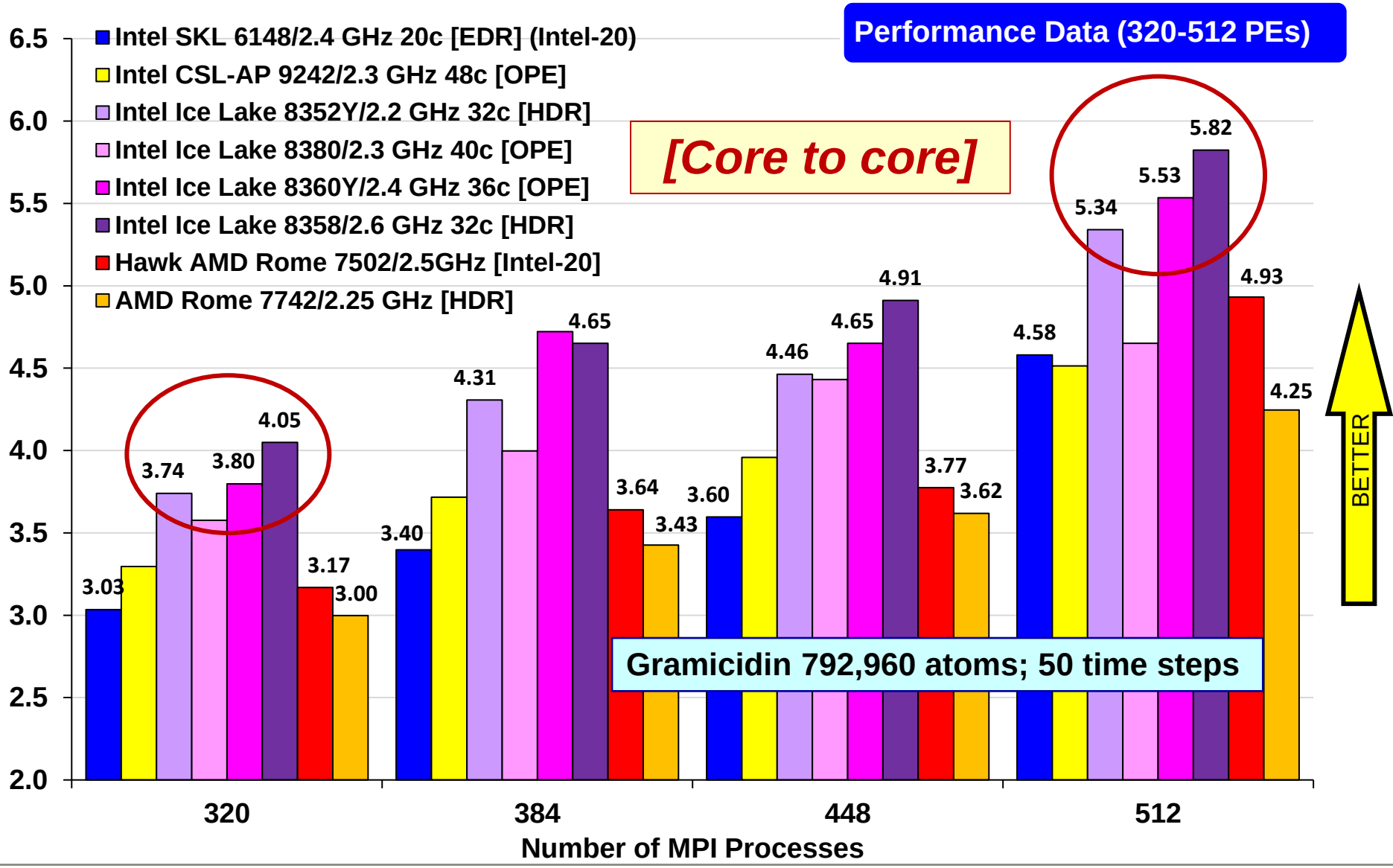
Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*

[Core to core]

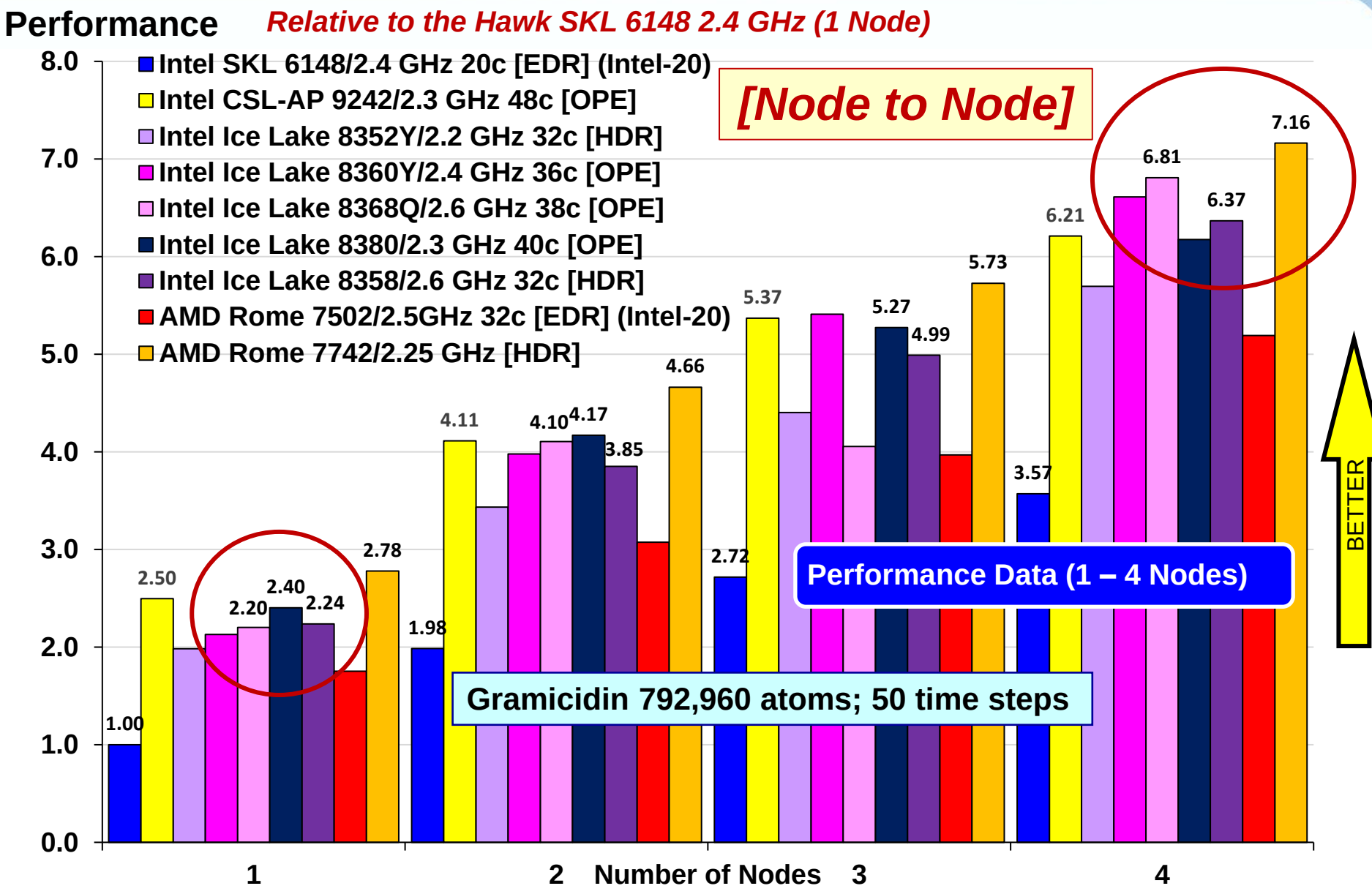


DL_POLY 4 – Gramicidin Simulation

Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*

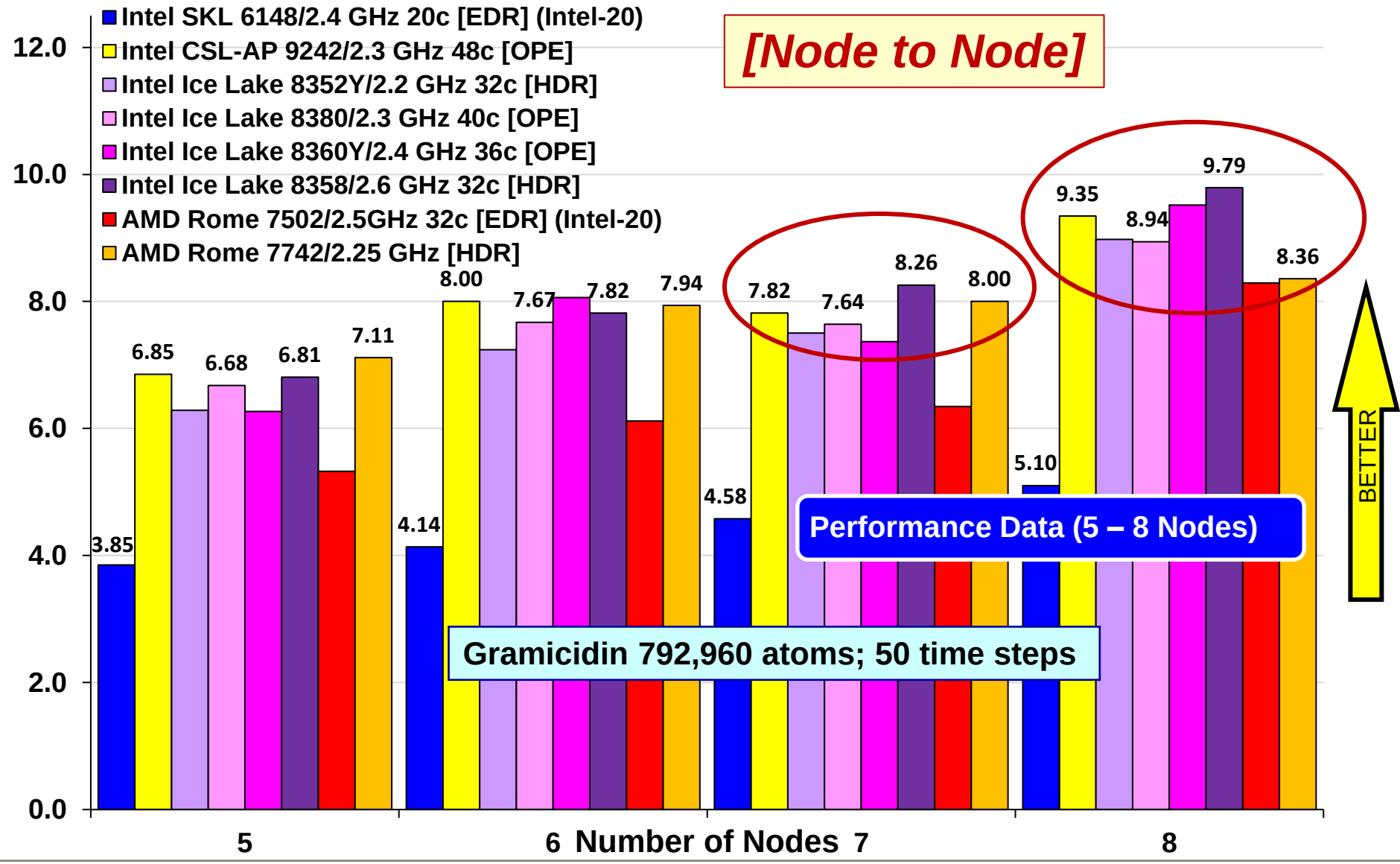


DL_POLY 4 – Gramicidin Simulation

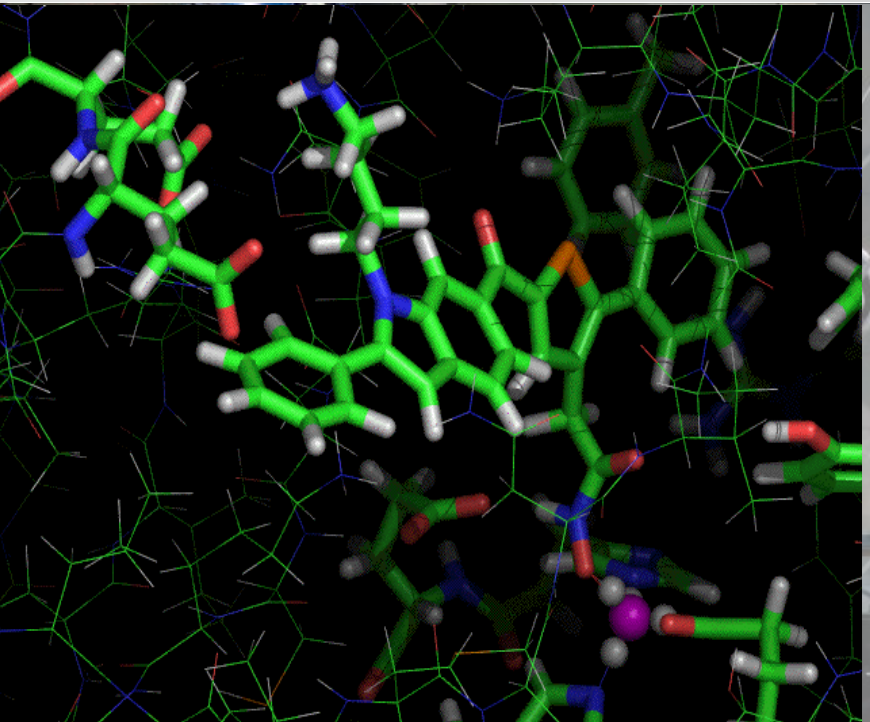


DL_POLY 4 – Gramicidin Simulation

Performance *Relative to the Hawk SKL 6148 2.4 GHz (1 Node)*

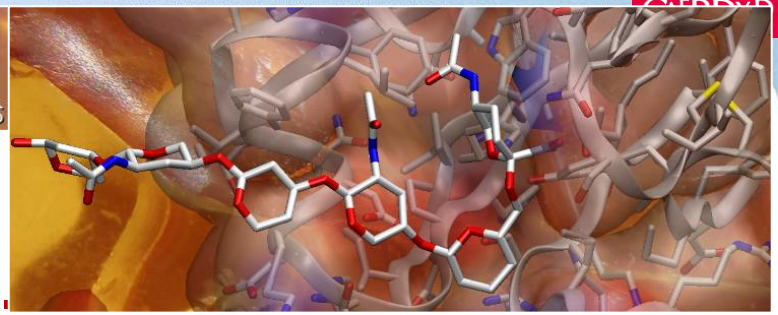


Performance of Computational Chemistry and Ocean Modelling Codes



Molecular Simulation: 2. AMBER

Molecular Simulation II. AMBER



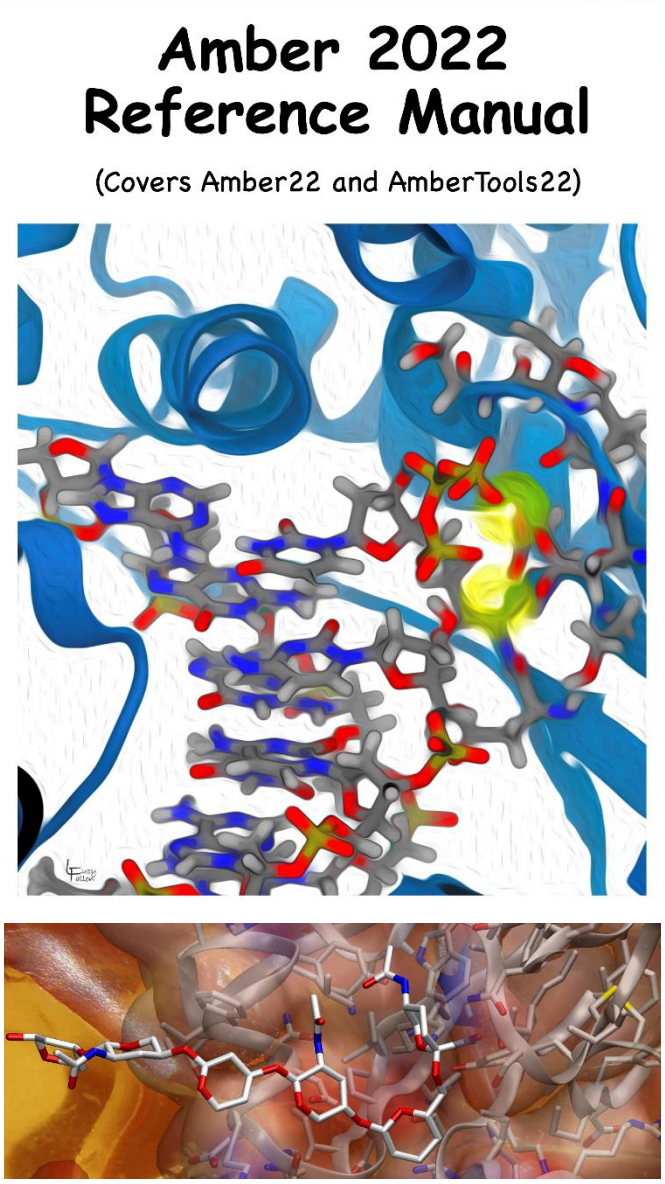
- AMBER18 and AMBER22 used:
  PMEMD & GPU accelerated PMEMD.
- **M01 Benchmark**
  - Major Urinary Protein (MUP) + IBM ligand (21,736 atoms)
- **M06 Benchmark**
  - Cluster of six MUPs (134,013 atoms)
- **M27 Benchmark**
  - **Cluster of 27 MUPs (657,585 atoms)**
- **M45 Benchmark**
  - **Cluster of 45 MUPs (932,751 atoms)**

*All test cases run 30,000 steps * 2fs = 60ps simulation time. Periodic boundary conditions, constant pressure, T=300K. Position data written every 500 steps.*

R. Salomon-Ferrer, D.A. Case, R.C. Walker. An overview of the Amber biomolecular simulation package. WIREs Comput. Mol. Sci. 3, 198-210 (2013).

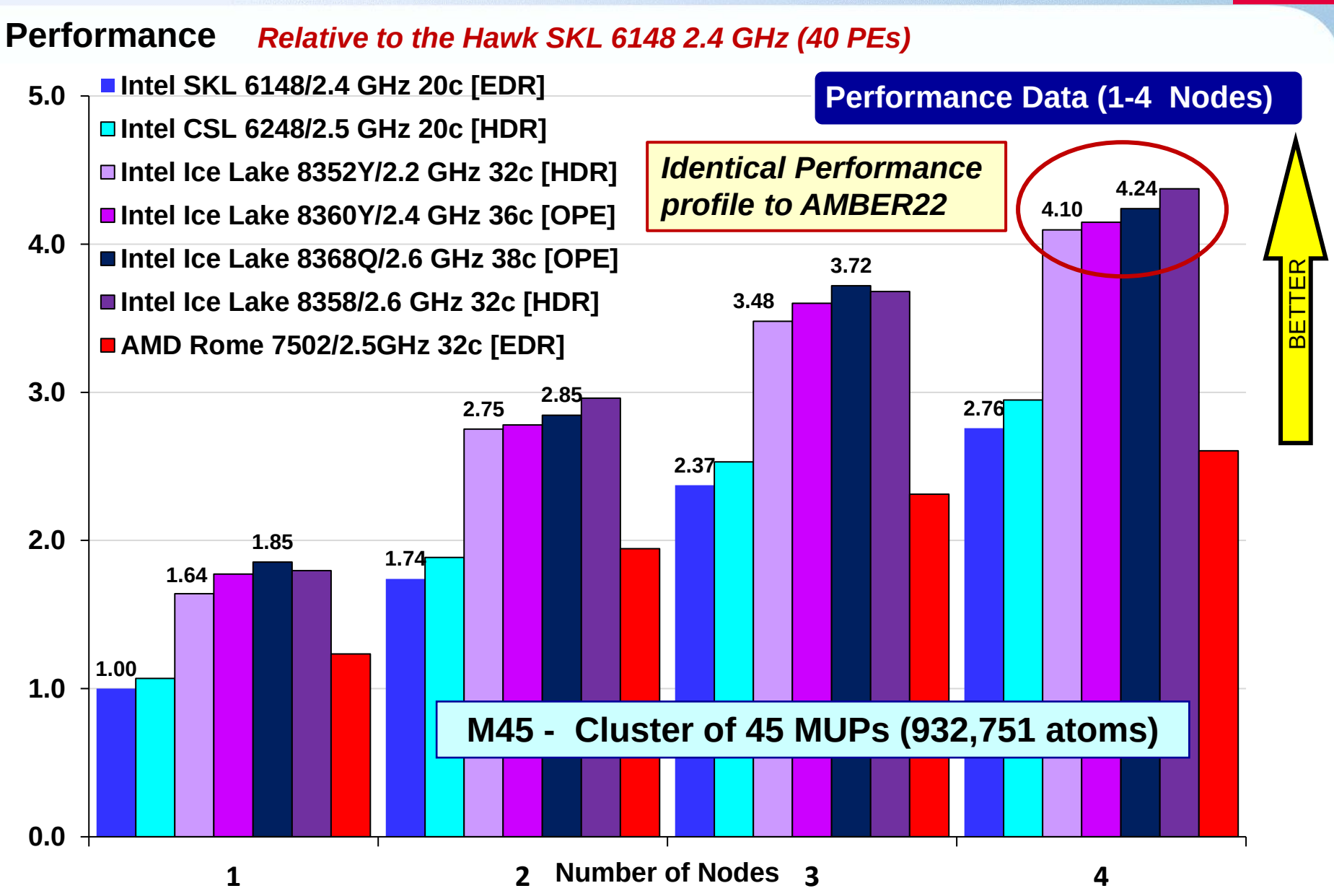
D.A. Case, T.E. Cheatham, III, T. Darden, H. Gohlke, R. Luo, K.M. Merz, Jr., A. Onufriev, C. Simmerling, B. Wang and R. Woods. The Amber biomolecular simulation programs. J. Computat. Chem. 26, 1668-1688 (2005).

Molecular Simulation II. AMBER



- **AMBER22** released (on April 27, 2022).
- The Amber22 package builds on AmberTools22 by adding the pmemd program, which resembles the sander (MD) code in AmberTools, but provides better performance on multiple CPUs, and dramatic speed improvements on GPUs.
- **AMBER18** (released in 2018) also used in this study. In practice we find **also identical performance of the two code releases** when running the M01, M06, M27 and M45 performance test cases.
- Presentation limited to the **M27 and M45** test cases for M01 and M06 are now too small for meaningful analysis

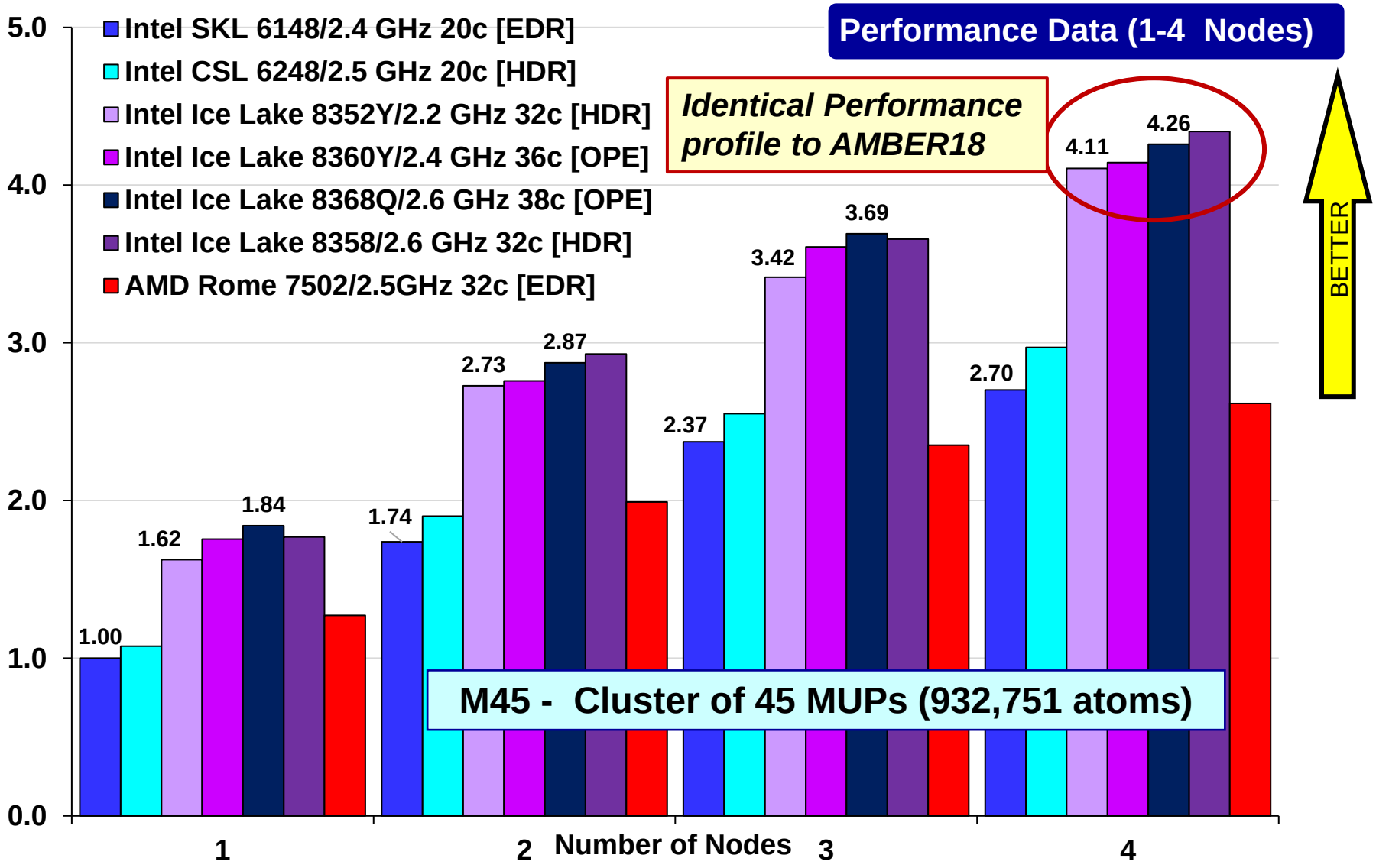
AMBER18 - M45 Performance results



AMBER22 - M45 Performance results

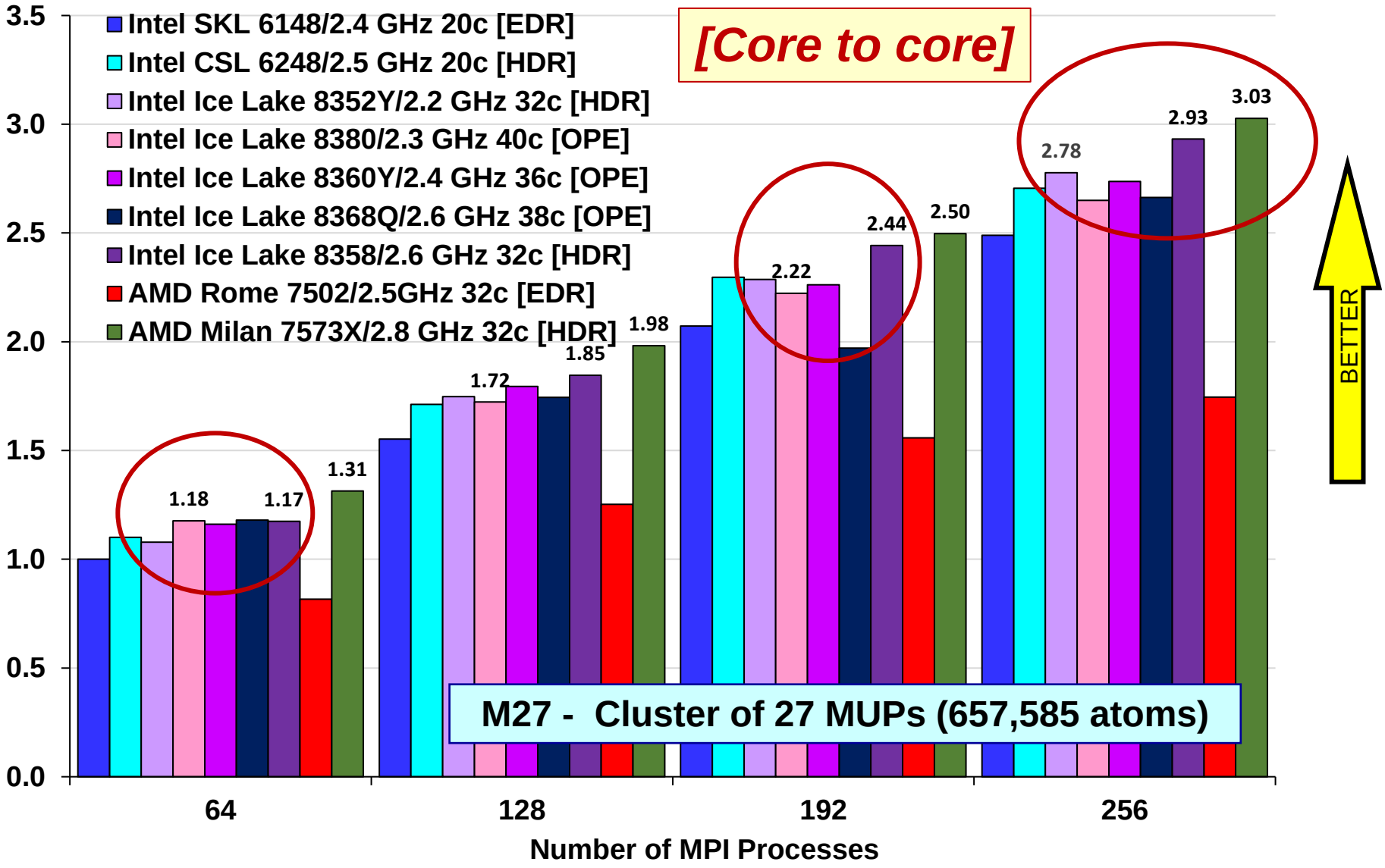
Performance

Relative to the Hawk SKL 6148 2.4 GHz (40 PEs)



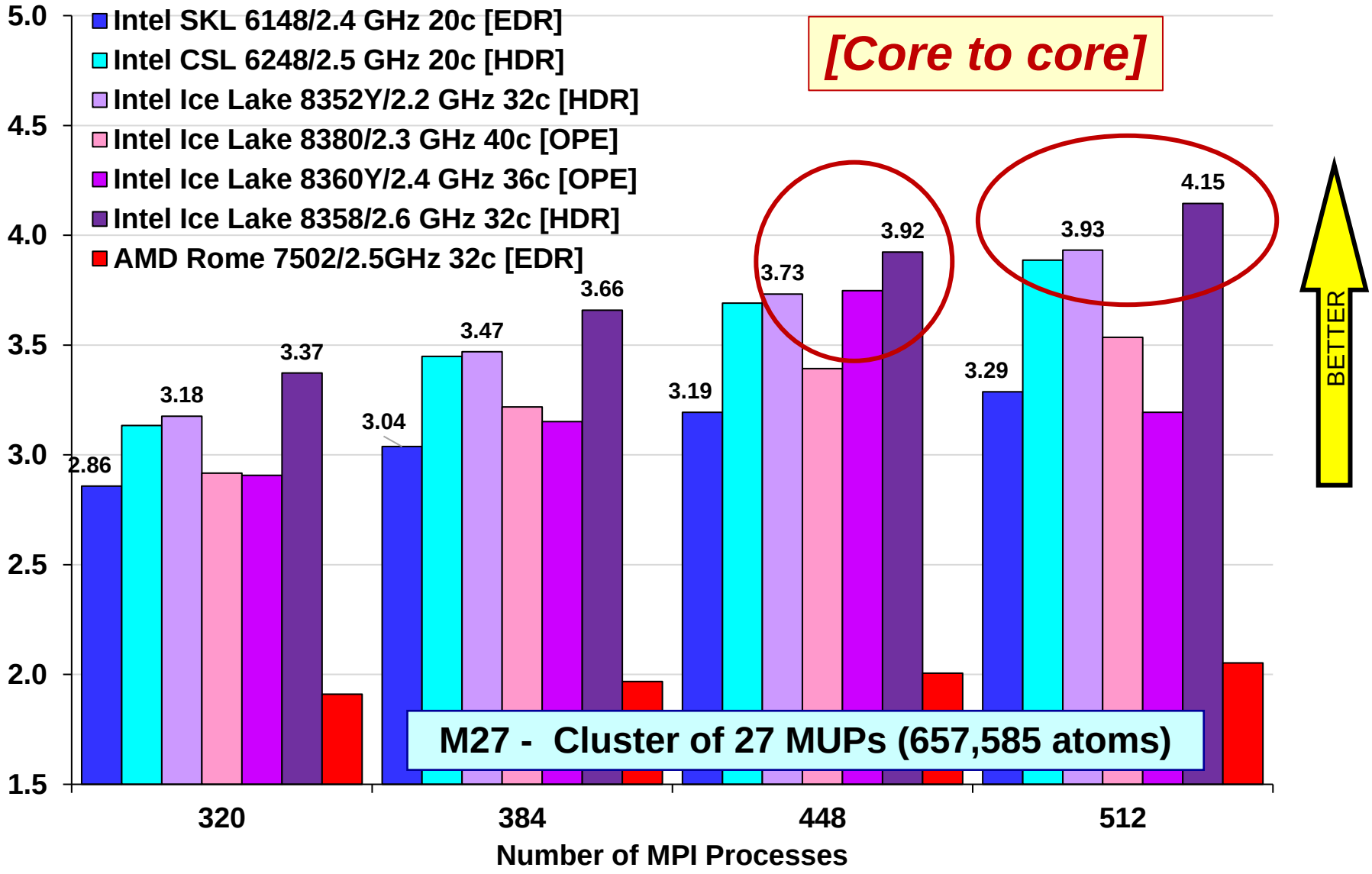
AMBER18 - M27 Performance Analysis

Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*



AMBER18 - M27 Performance Analysis

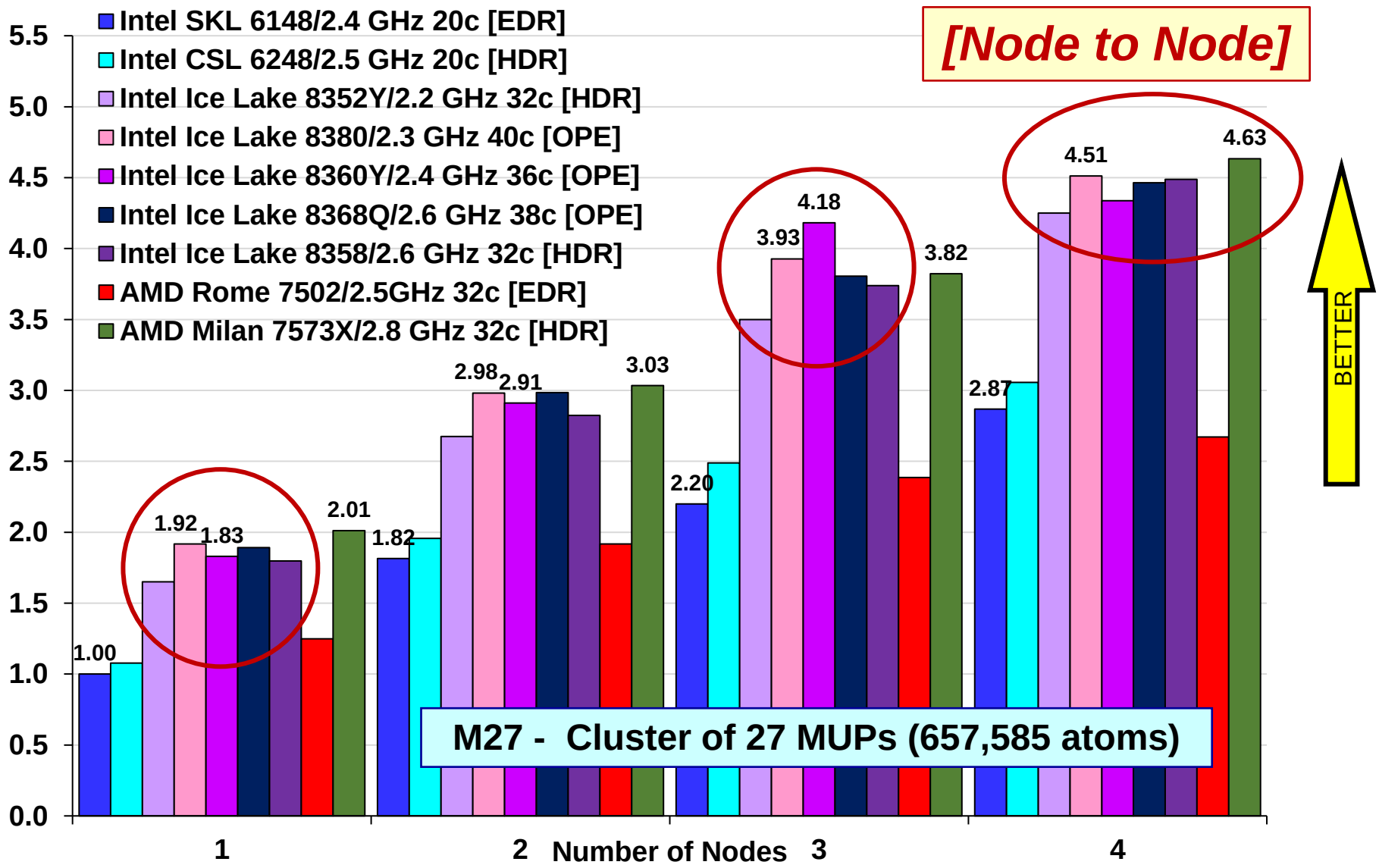
Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*



AMBER18 - M27 Performance Analysis

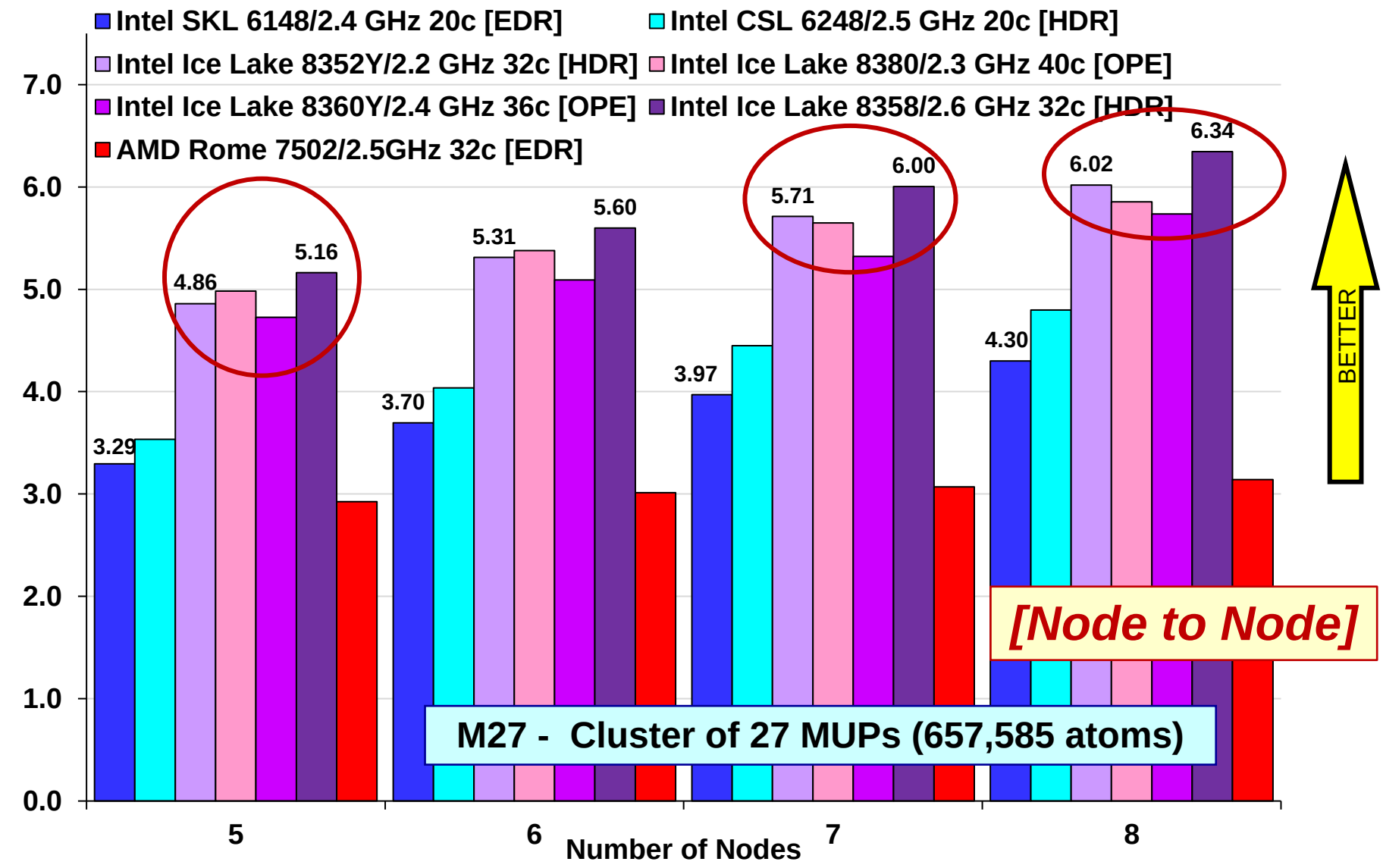
CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*



AMBER18 - M27 Performance Analysis

Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*

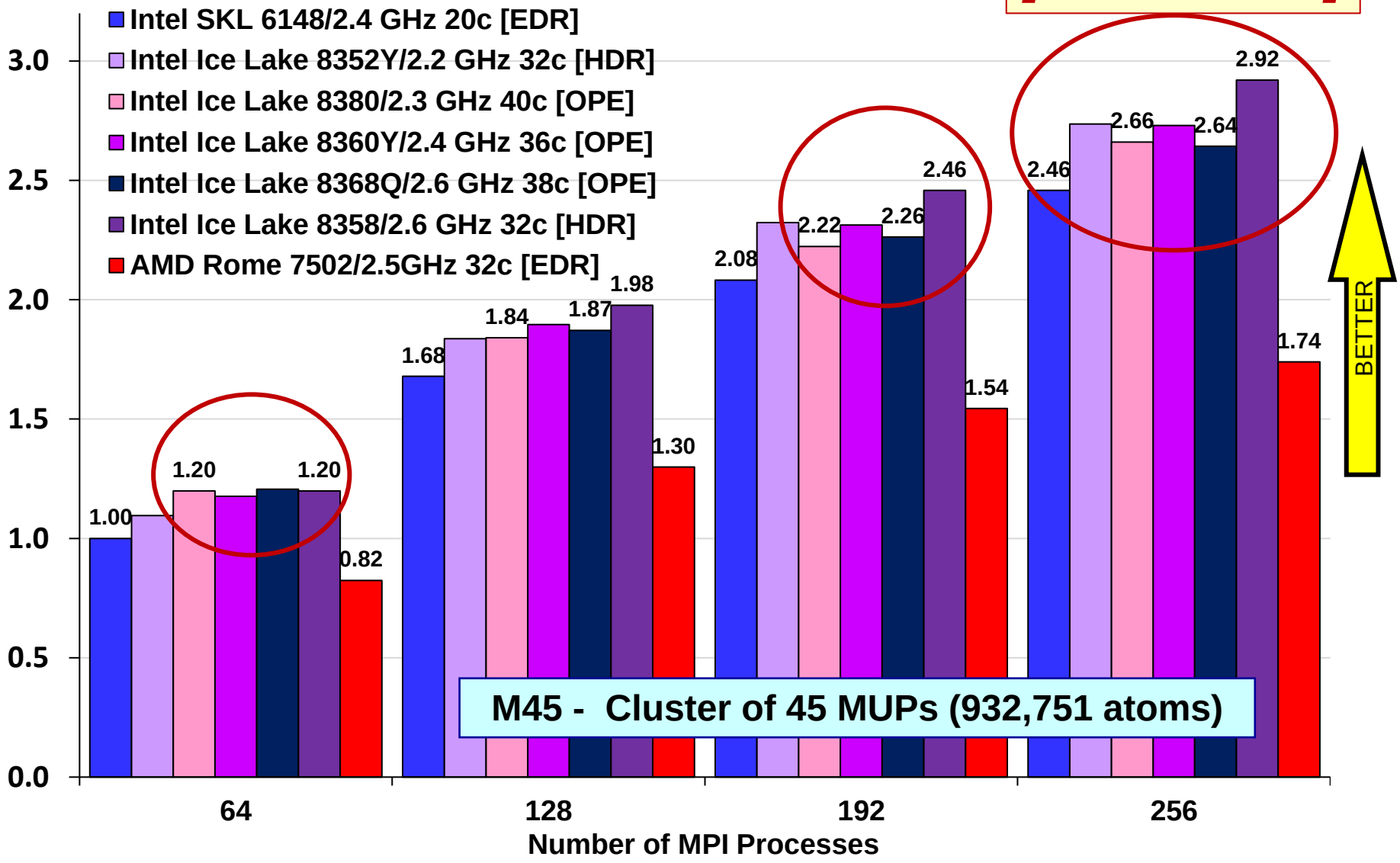


AMBER18 - M45 Performance Analysis

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*

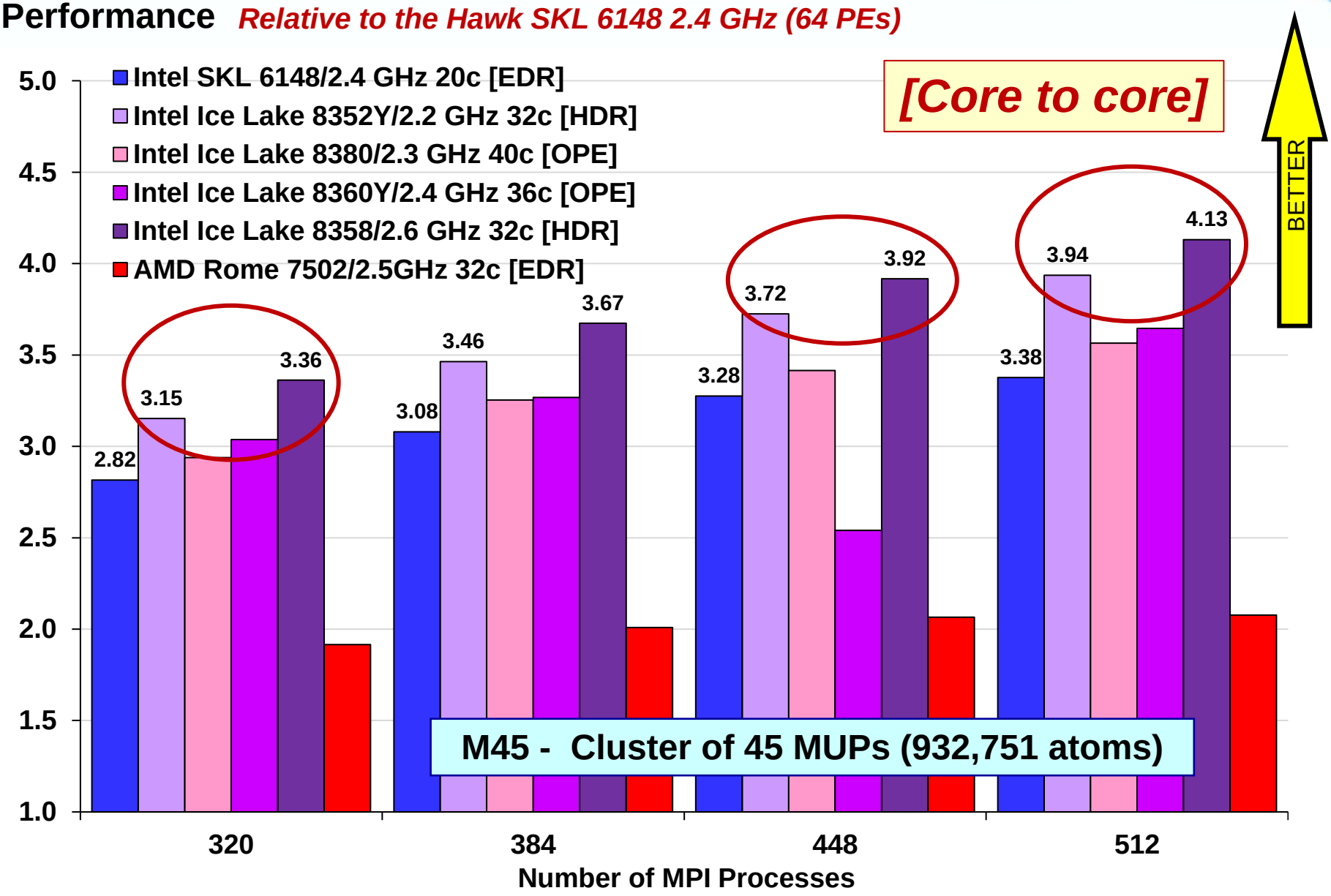
[Core to core]



AMBER18 - M45 Performance Analysis

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

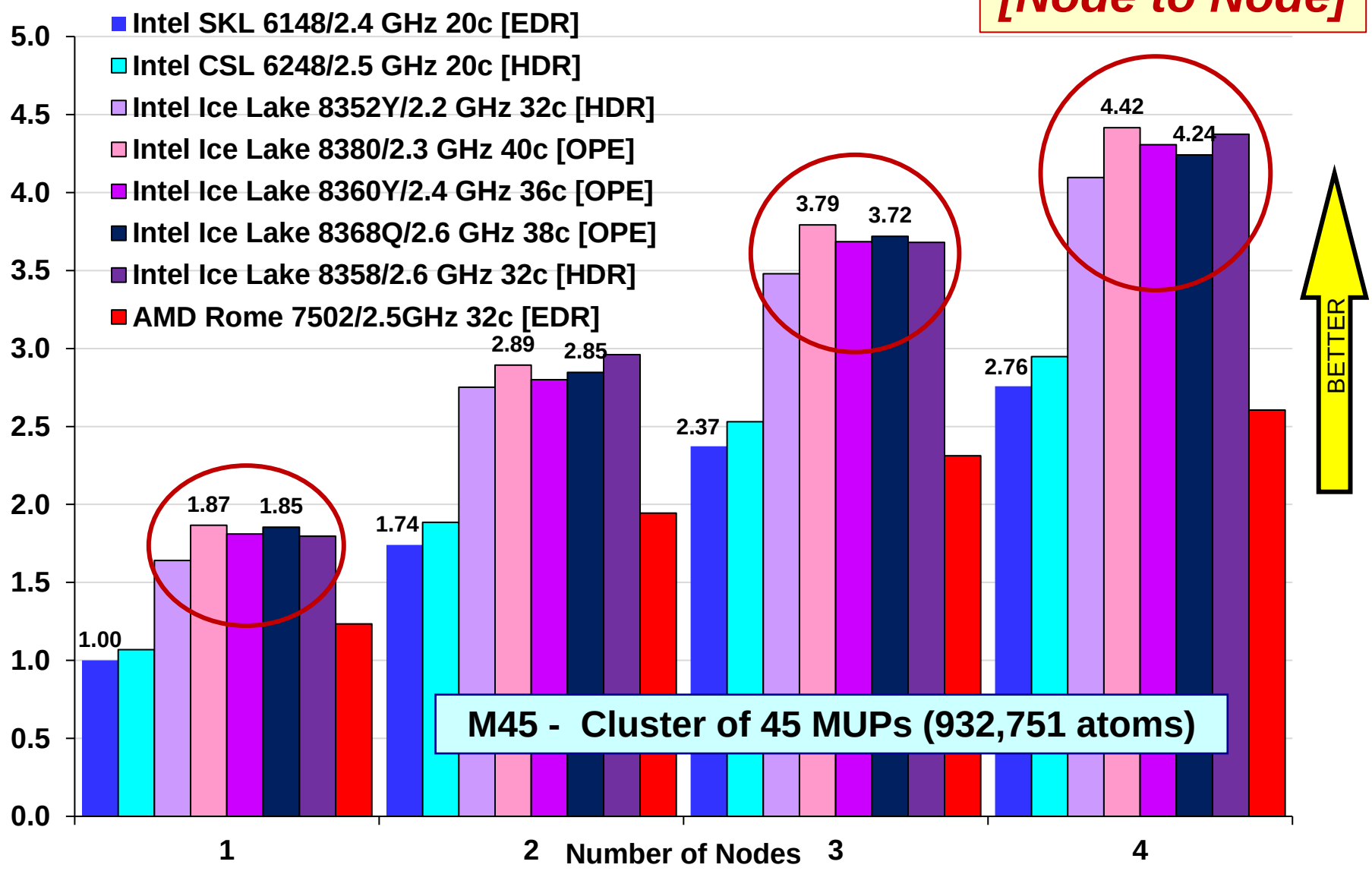
Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*



AMBER18 - M45 Performance Analysis

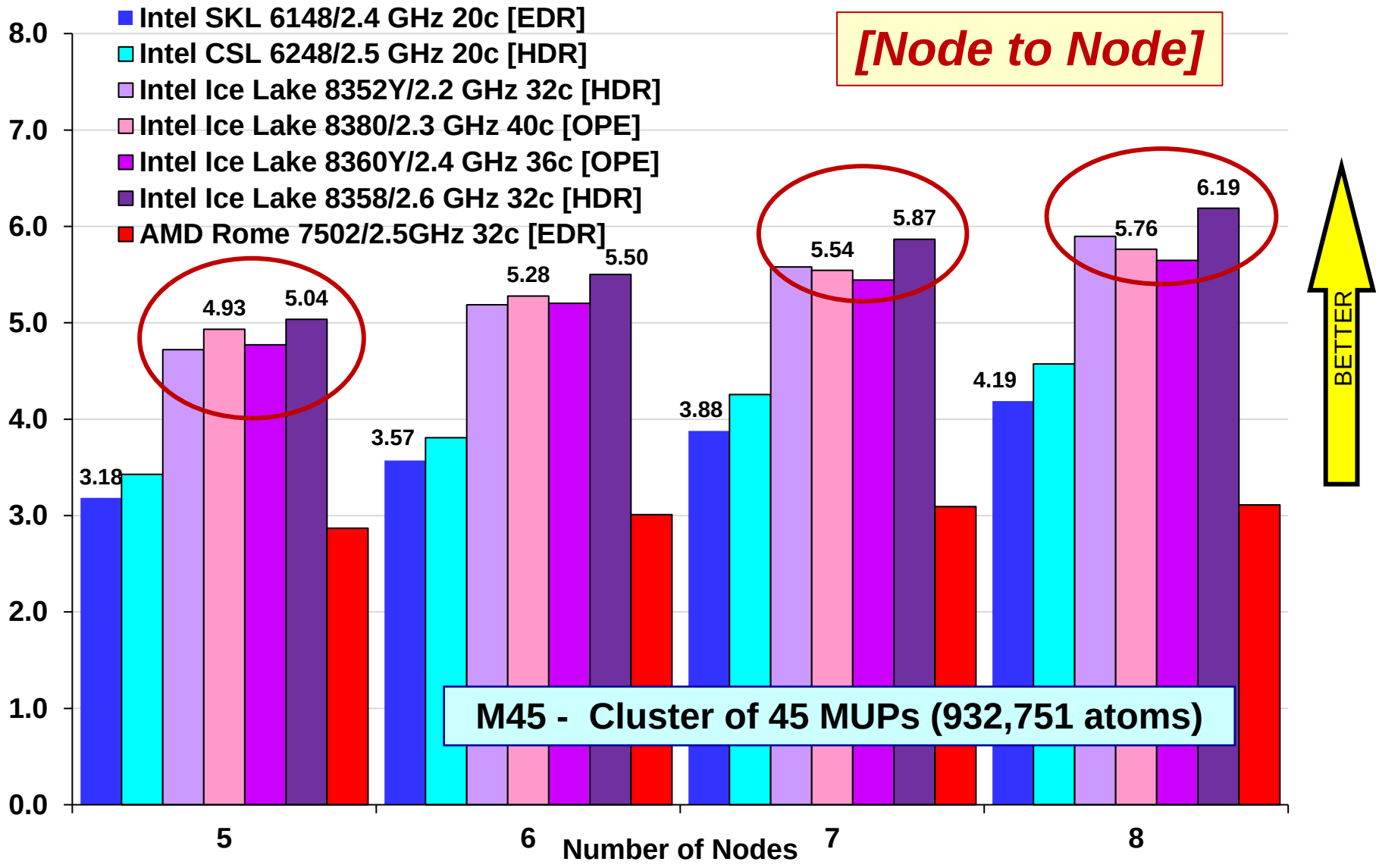
Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*

[Node to Node]

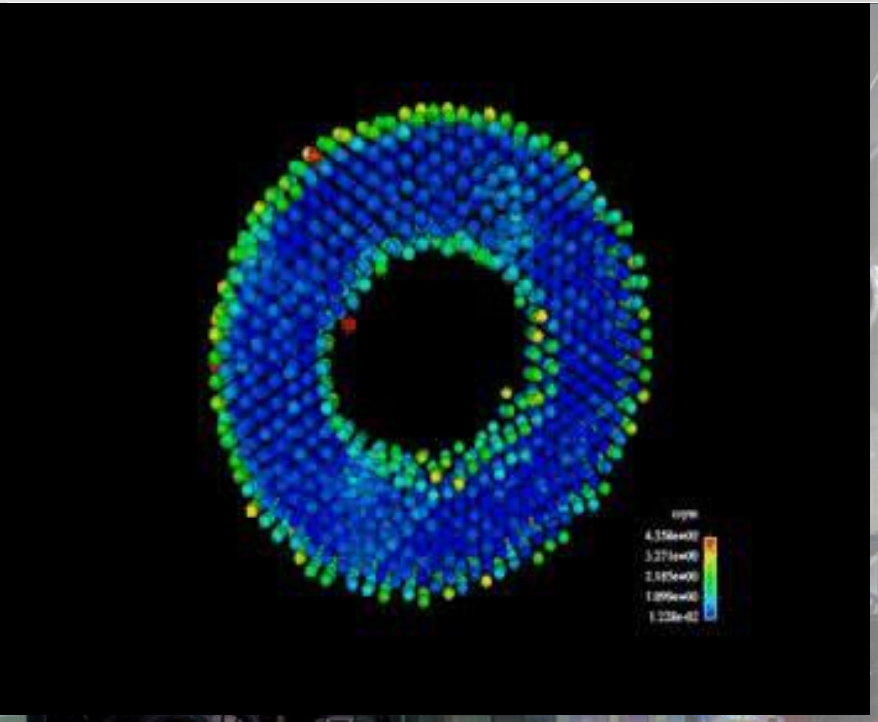


AMBER18 - M45 Performance Analysis

Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*



Performance of Computational Chemistry and Ocean Modelling Codes



Advanced Materials Software:

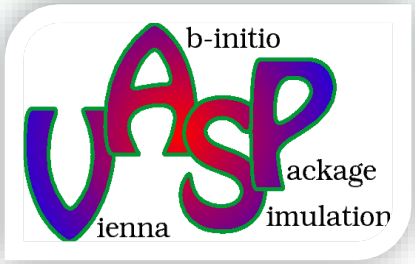
1. VASP

Advanced Materials Software

Computational Materials

- **VASP** – performs ab-initio QM molecular dynamics (MD) simulations using **pseudopotentials** or the projector-augmented wave method and a plane wave basis set.
- **Quantum Espresso** – an integrated suite of Open-Source computer codes for electronic-structure calculations and materials modelling at the nanoscale. It is based on density-functional theory (**DFT**), plane waves, and **pseudopotentials**
- **CASTEP** – a full-featured materials modelling code based on a first-principles QM description of electrons and nuclei. Uses robust methods of a **plane-wave basis set and pseudopotentials**.
- **CP2K** is a program to perform atomistic and molecular simulations of solid state, liquid, molecular, and biological systems. It provides a framework for different methods such as e.g., **DFT** using a mixed Gaussian & plane waves approach (GPW) and classical pair and many-body potentials.
- **ONETEP** (Order-N Electronic Total Energy Package) is a linear-scaling code for quantum-mechanical calculations based on **DFT**.

VASP – Vienna *Ab-initio* Simulation Package



Archer Rank: 1

VASP (**6.3**) performs ab-initio QM molecular dynamics (MD) simulations using pseudopotentials or the projector-augmented wave method and a plane wave basis set.

Benchmark	Details
MFI Zeolite	Zeolite ($Si_{96}O_{192}$), 2 k-points, FFT grid: (65, 65, 43); 181,675 points
Pd-O complex	Palladium-Oxygen complex ($Pd_{75}O_{12}$), 10 k-points, FFT grid: (31, 49, 45), 68,355 points

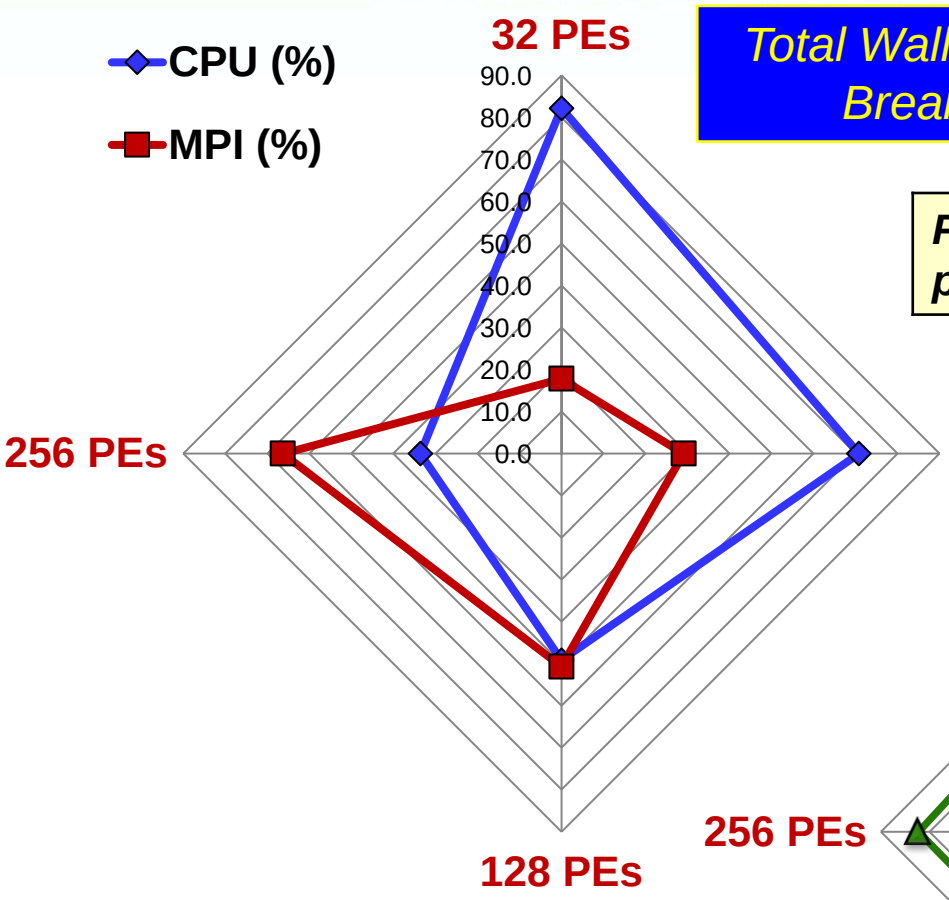
Pd-O Benchmark

- Pd-O complex – $Pd_{75}O_{12}$, 5X4 3-layer supercell running a single point calculation and a planewave cut off of 400eV. Uses the RMM-DIIS algorithm for the SCF and is calculated in real space.
- 10 k-points; maximum number of plane-waves: 34,470
- FFT grid; NGX=31, NGY=49, NGZ=45, giving a total of 68,355 points

Zeolite Benchmark

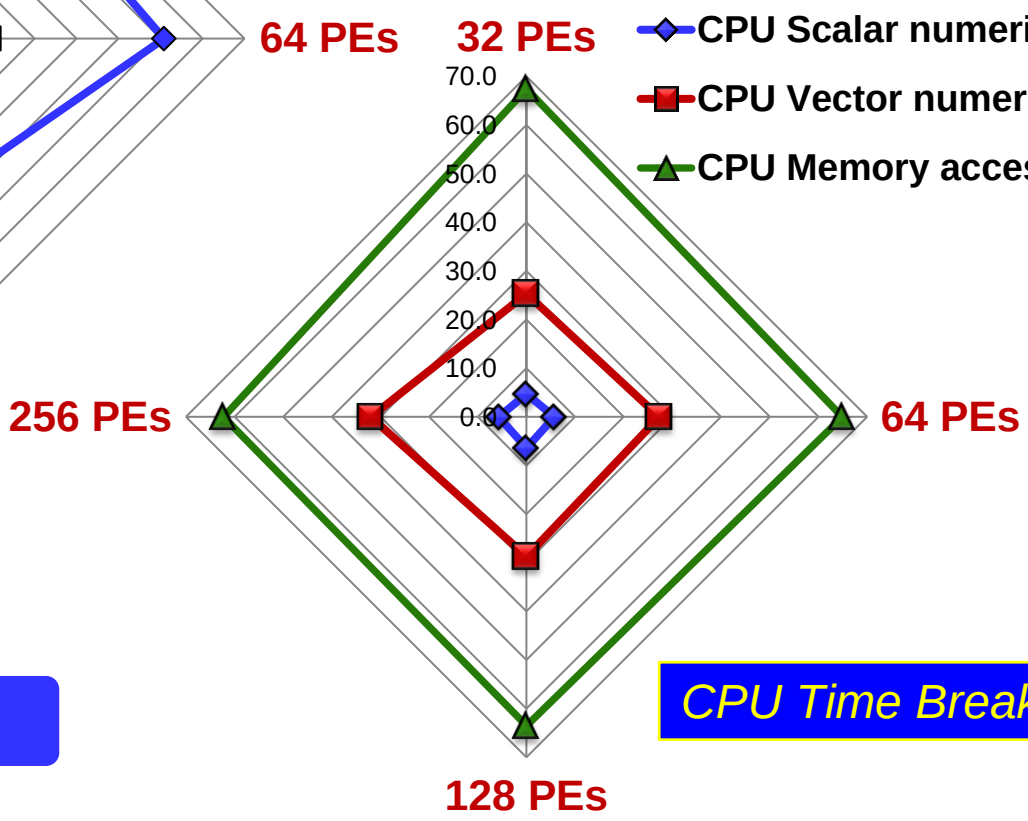
- Zeolite with the MFI structure unit cell running a single point calculation and a planewave cut off of 400eV using the PBE functional
- 2 k-points; maximum number of plane-waves: 96,834
- FFT grid; NGX=65, NGY=65, NGZ=43, giving a total of 181,675 points

VASP – Pd-O Benchmark Performance Report



Total Wallclock Time Breakdown

Palladium-Oxygen complex ($Pd_{75}O_{12}$), 10 k-points, FFT grid: (31, 49, 45), 68,355 points



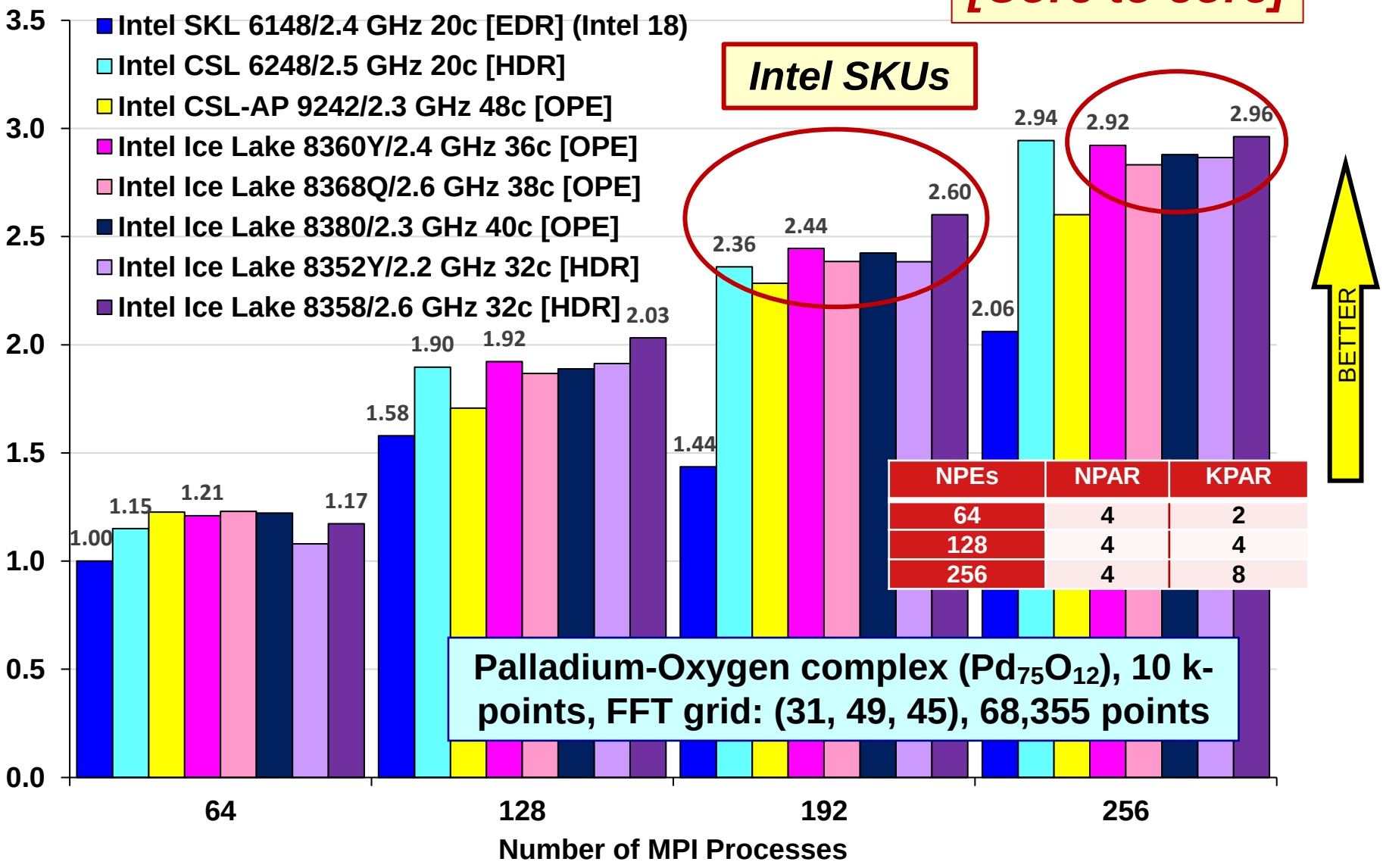
CPU Time Breakdown

Performance Data (32-256 PEs)

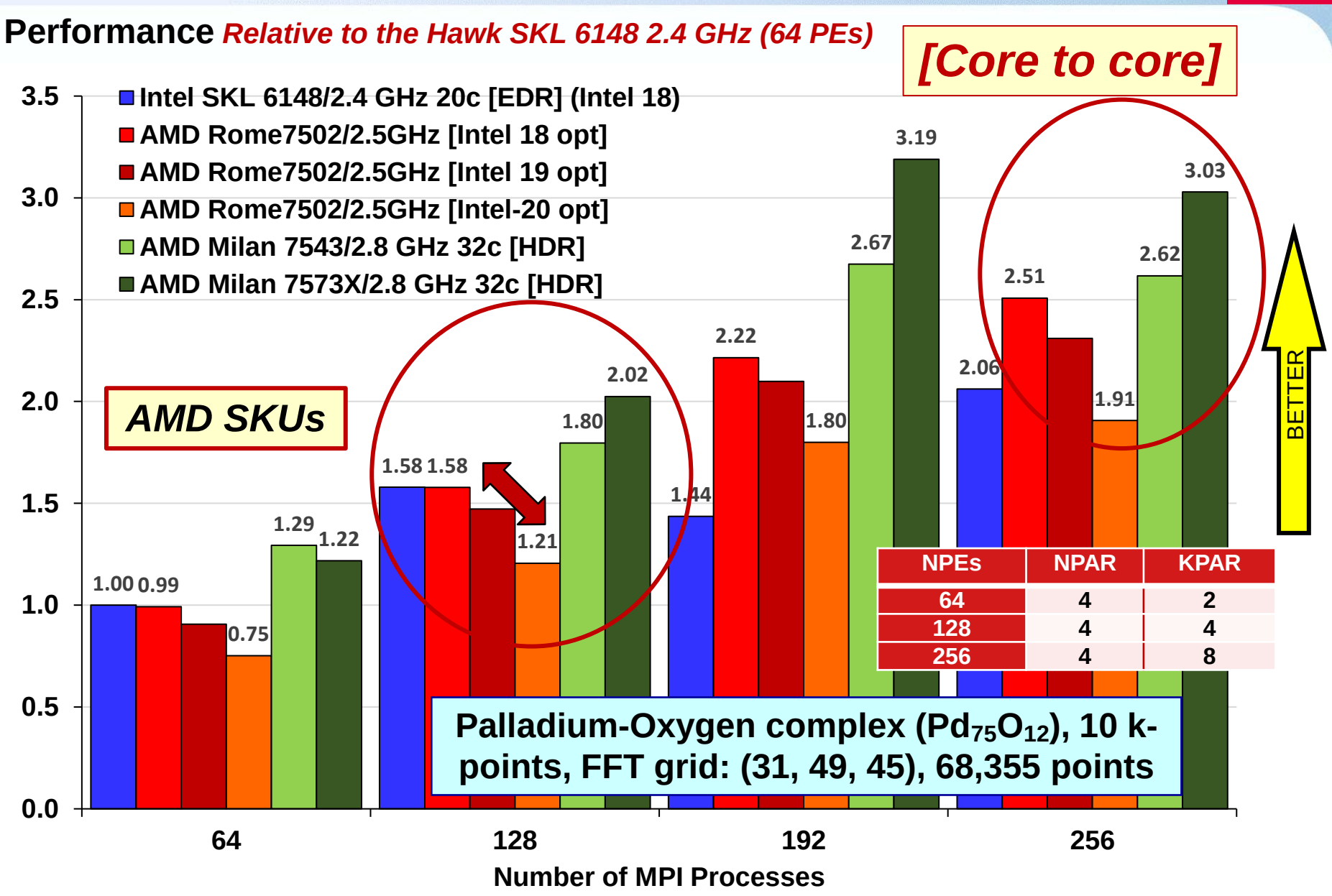
VASP 6.3 – Pd-O Benchmark - Parallelisation on k-points

Performance ***Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)***

[Core to core]



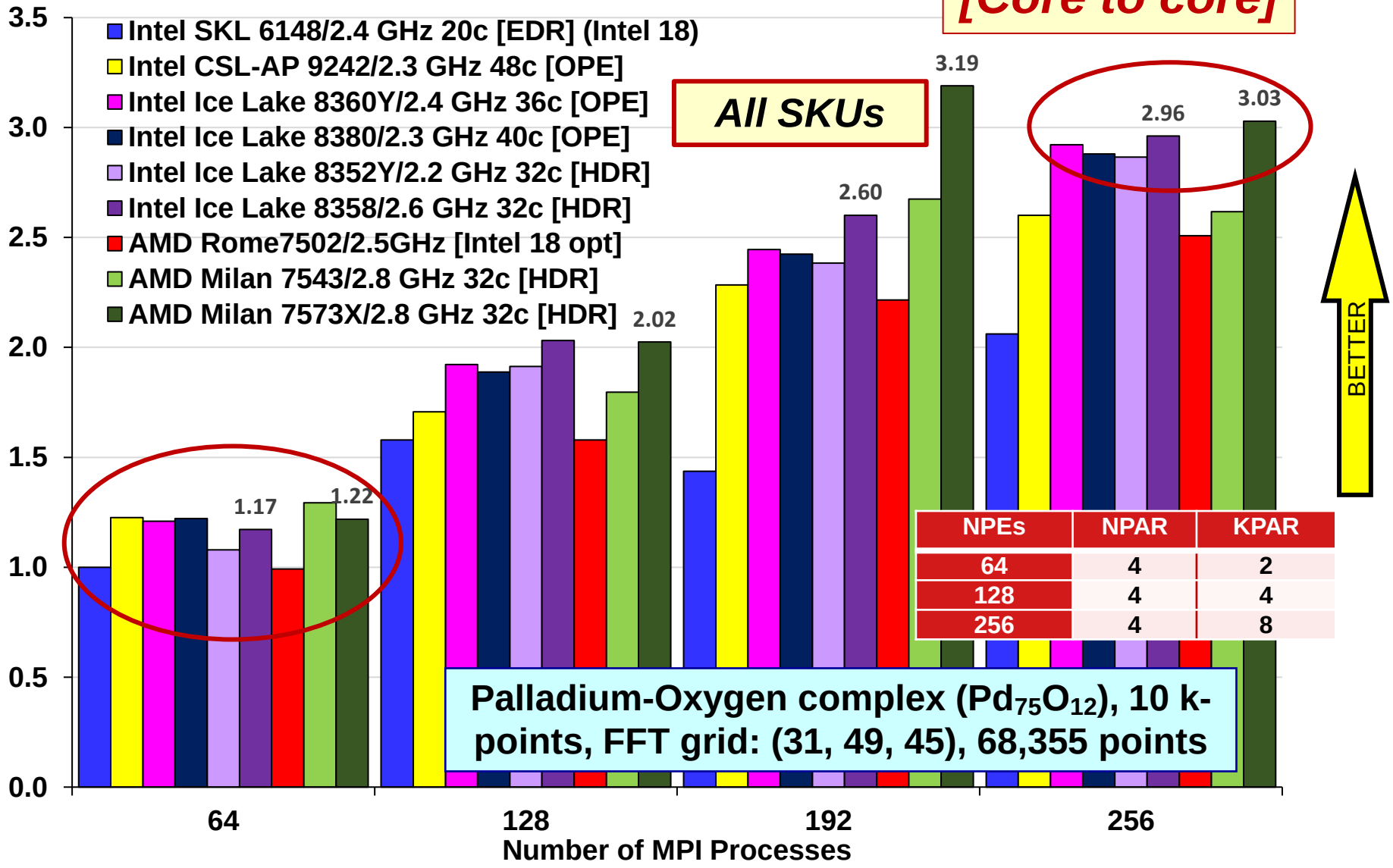
VASP 6.3 – Pd-O Benchmark - Parallelisation on k-points



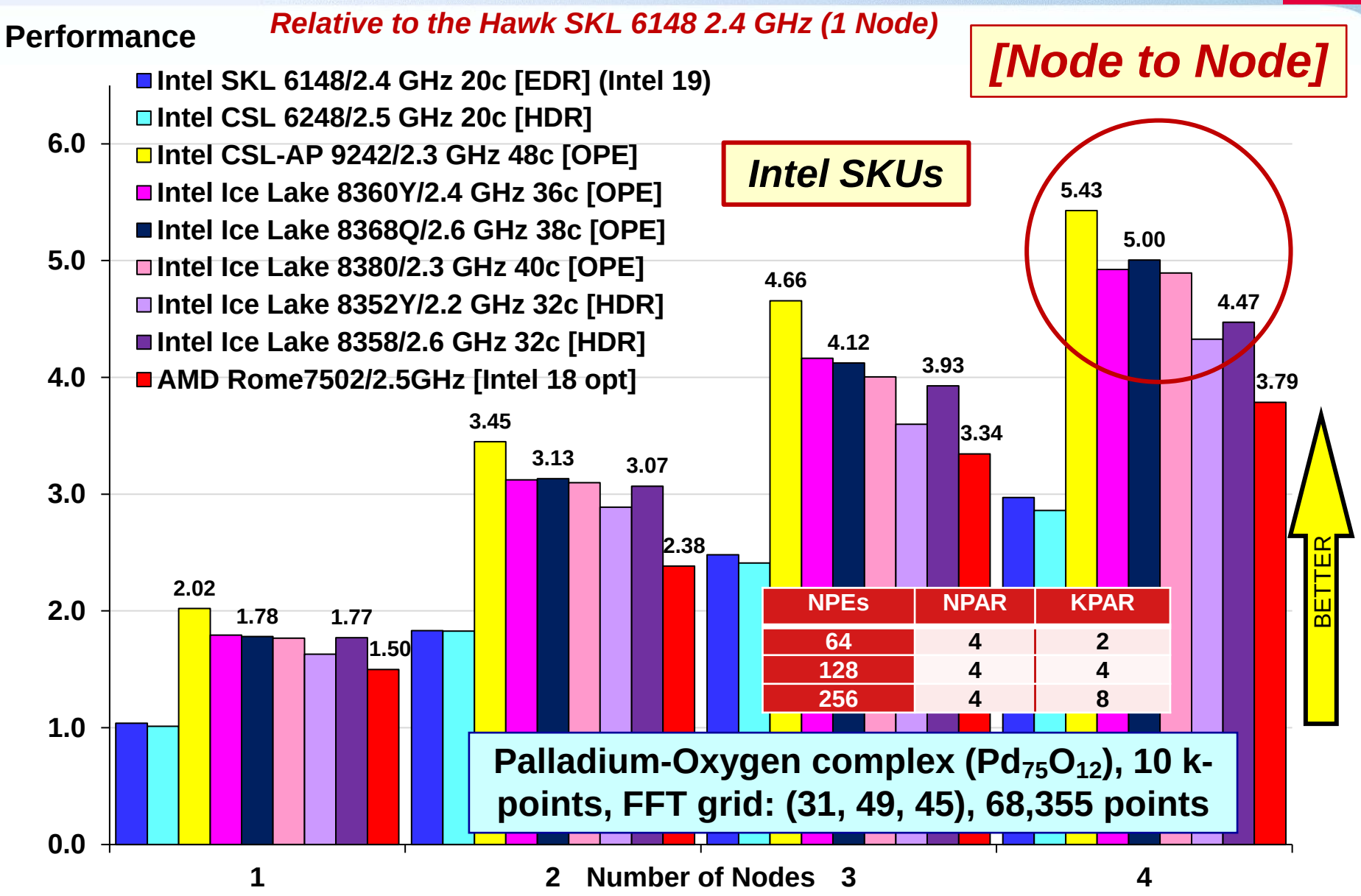
VASP 6.3 – Pd-O Benchmark - Parallelisation on k-points

Performance ***Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)***

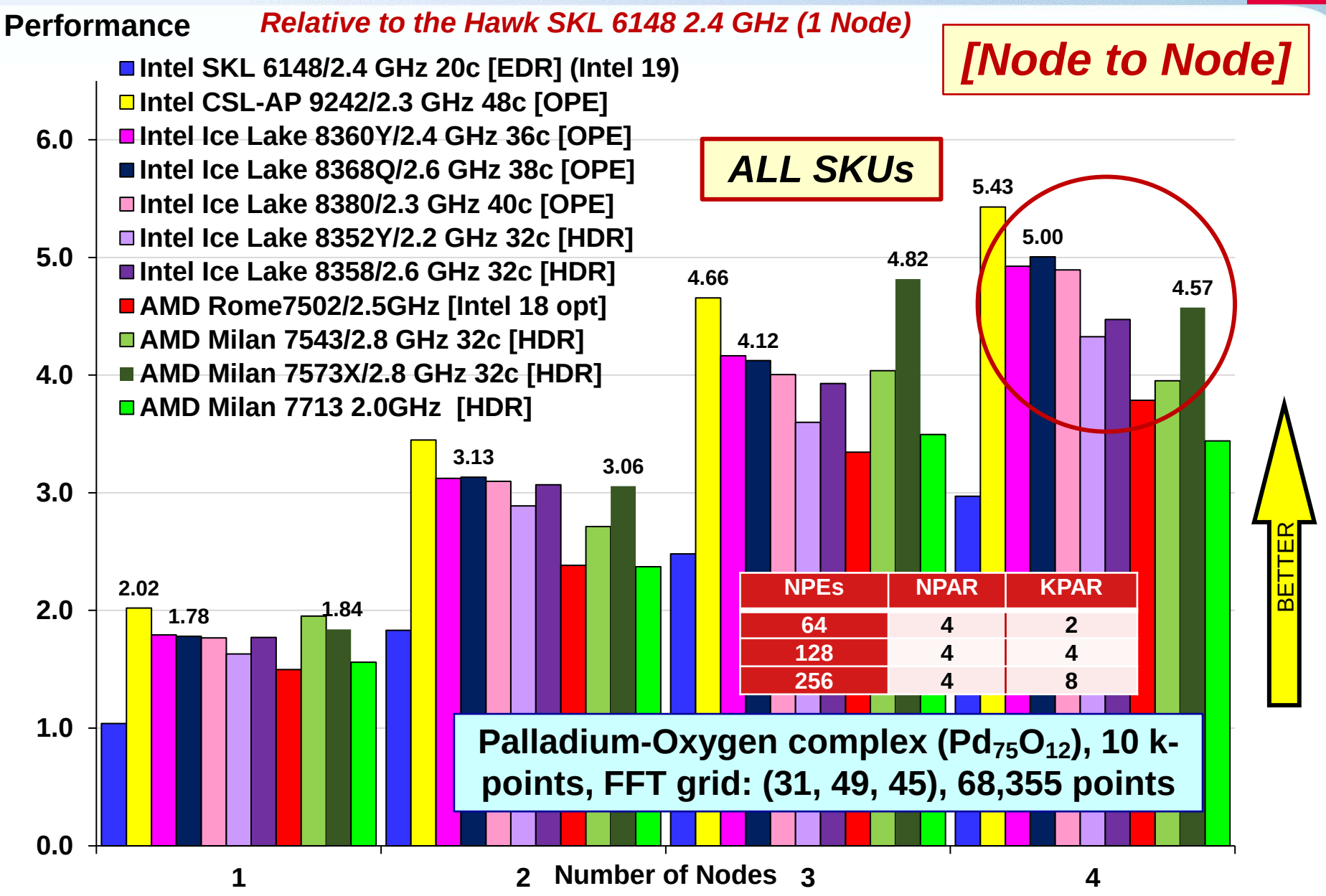
[Core to core]



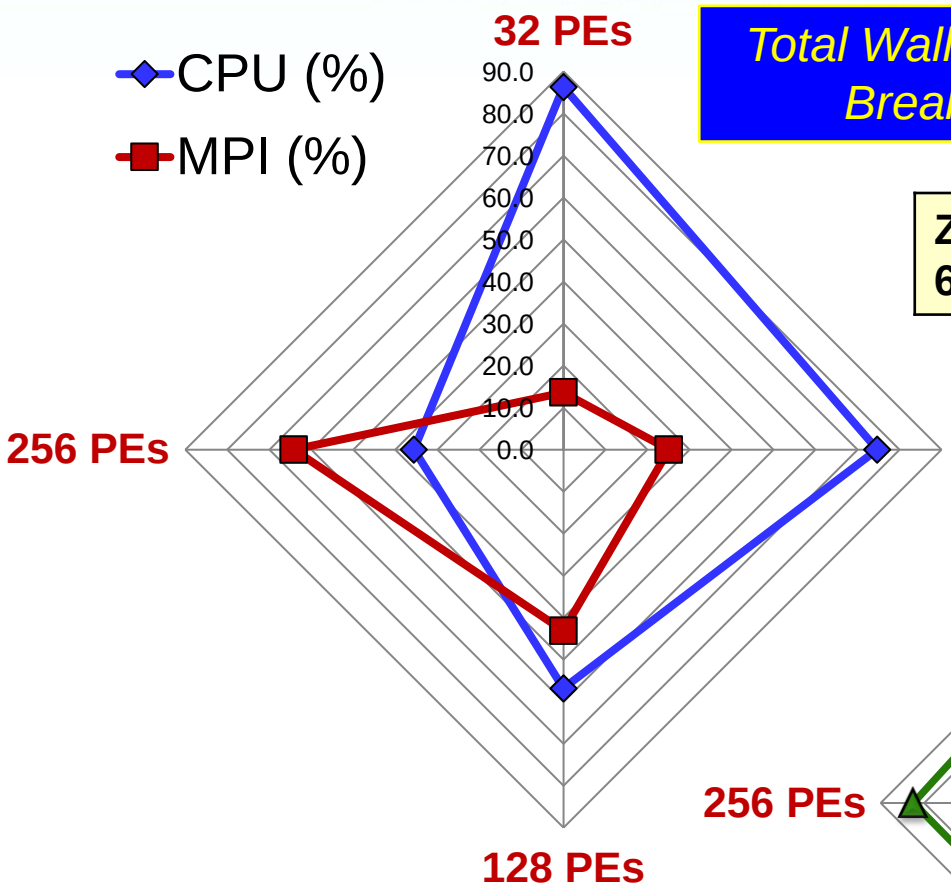
VASP 6.3 – Pd-O Benchmark - Parallelisation on k-points



VASP 6.3 – Pd-O Benchmark - Parallelisation on k-points

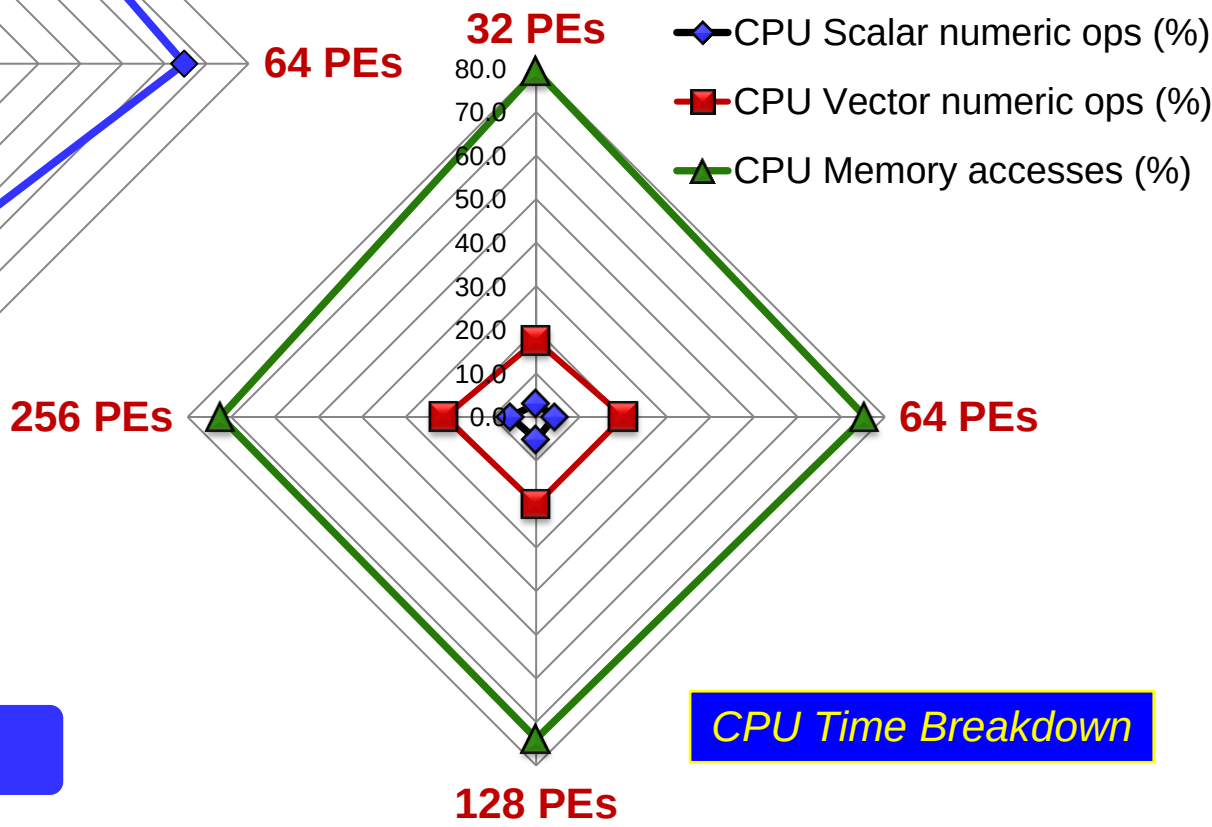


VASP – Zeolite Cluster Performance Report



Total Wallclock Time Breakdown

Zeolite ($Si_{96}O_{192}$), 2 k-points, FFT grid: (65, 65, 43); 181,675 points

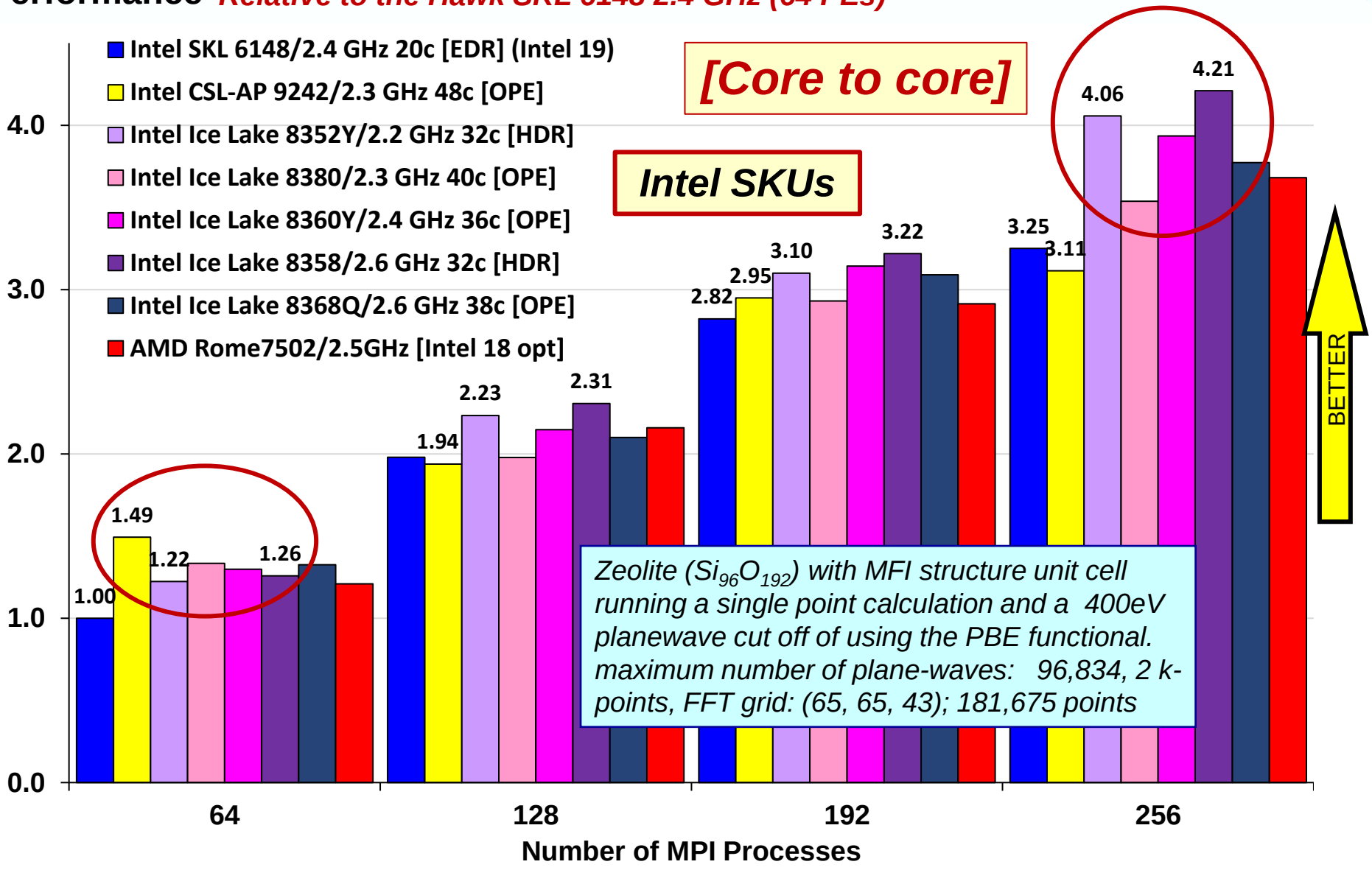


CPU Time Breakdown

Performance Data (32-256 PEs)

VASP 6.3 – Zeolite Benchmark - Parallelisation on k-points

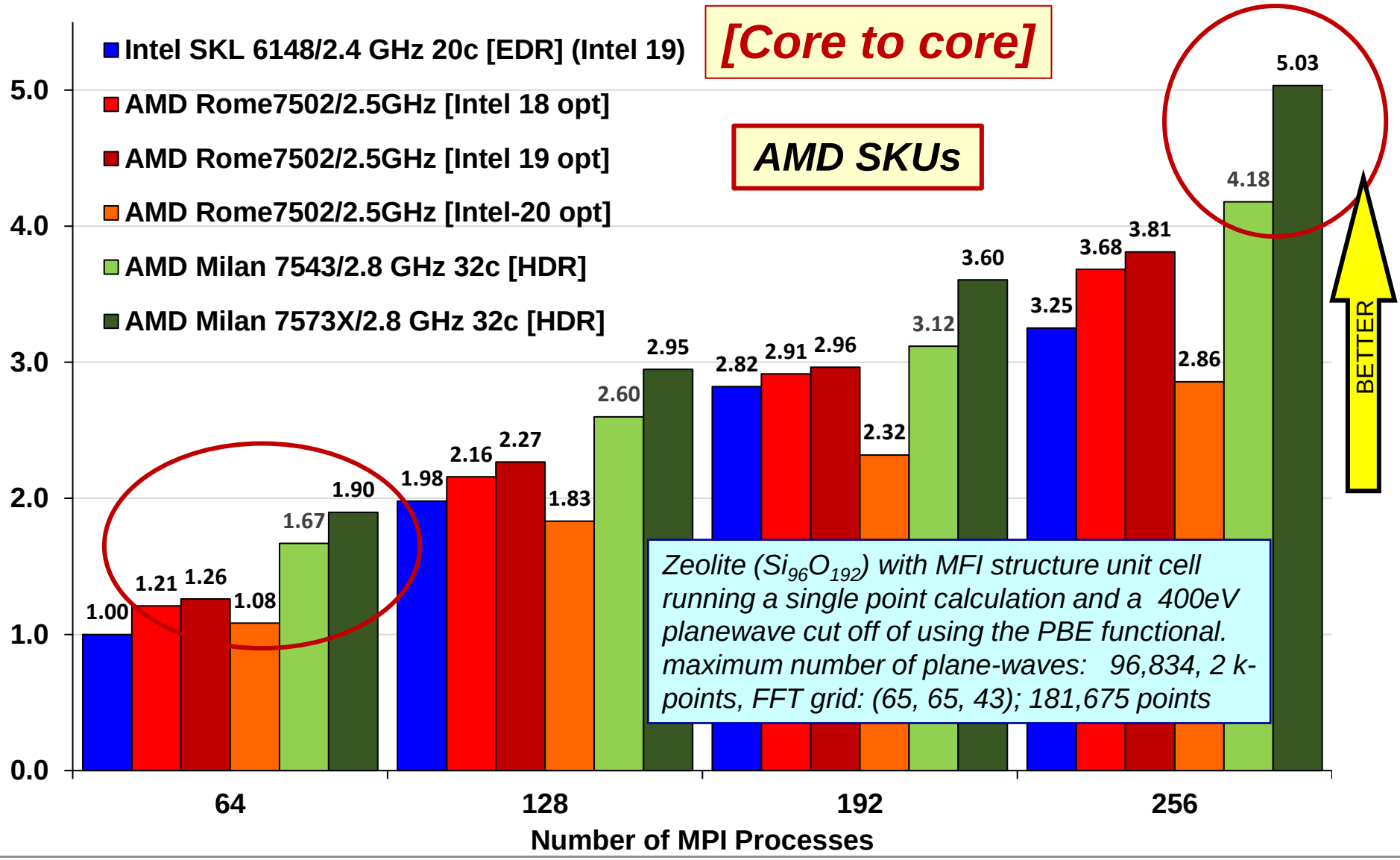
Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*



VASP 6.3 – Zeolite Benchmark - Parallelisation on k-points

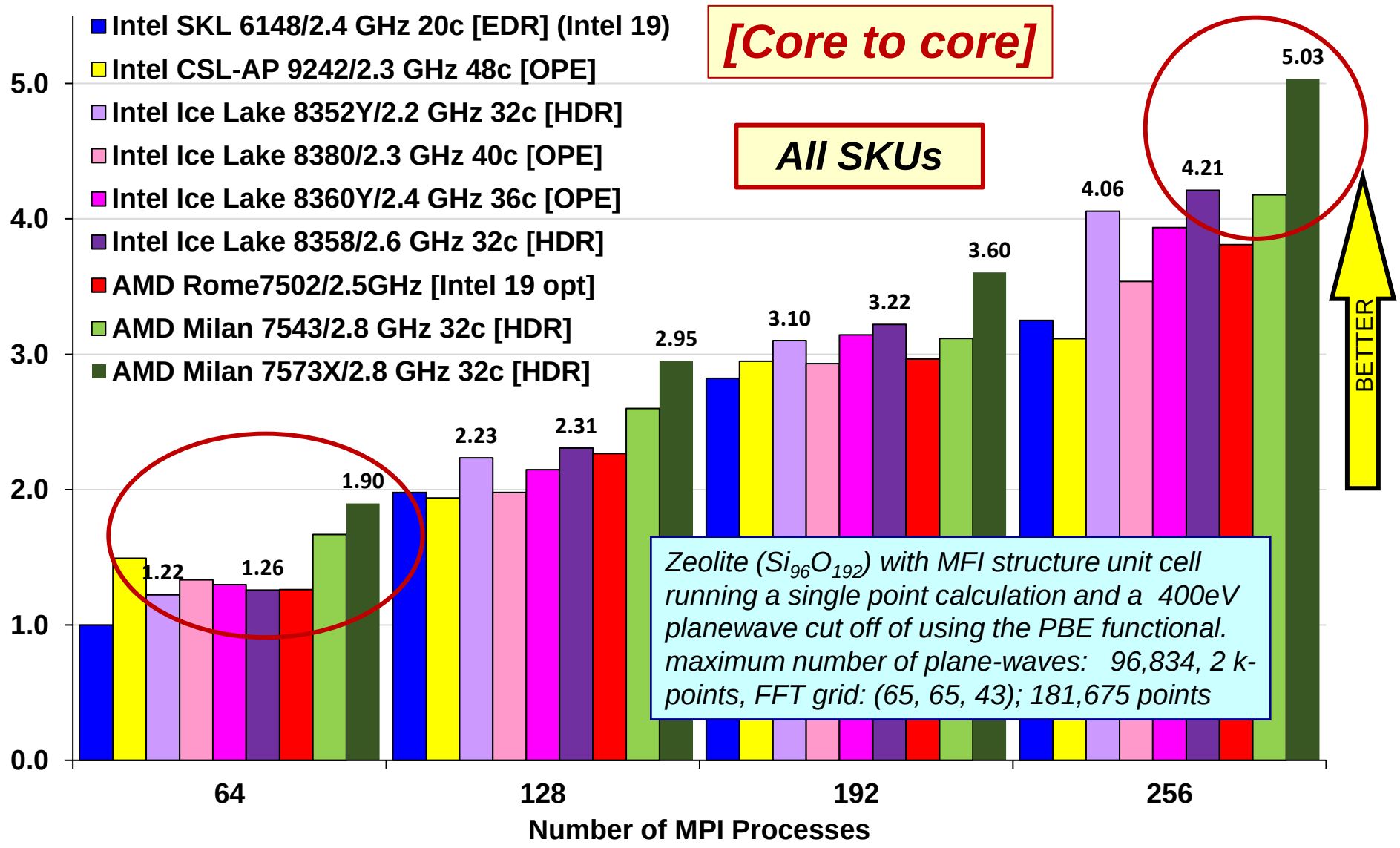
CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*



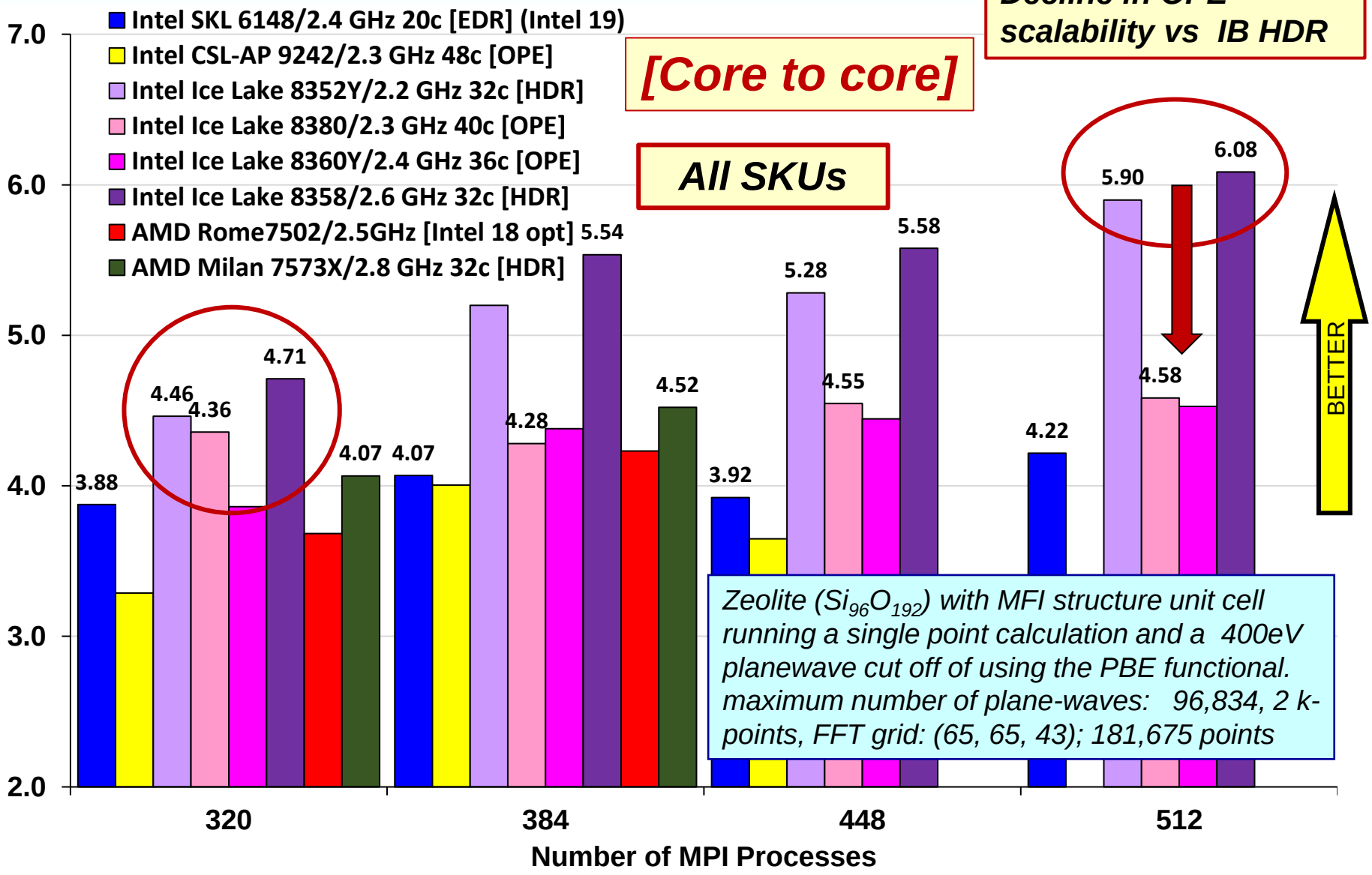
VASP 6.3 – Zeolite Benchmark - Parallelisation on k-points

Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*



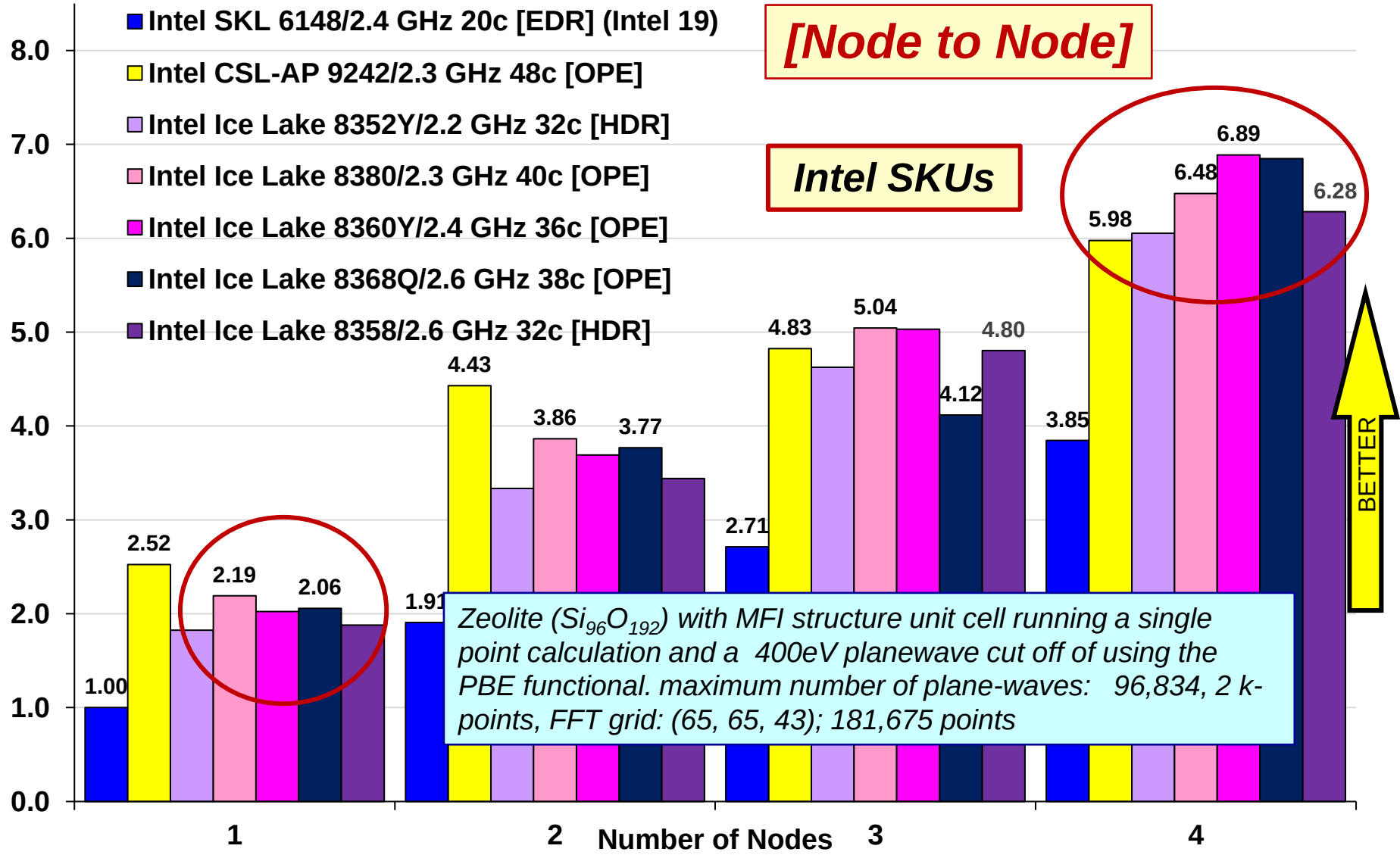
VASP 6.3 – Zeolite Benchmark - Parallelisation on k-points

Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*



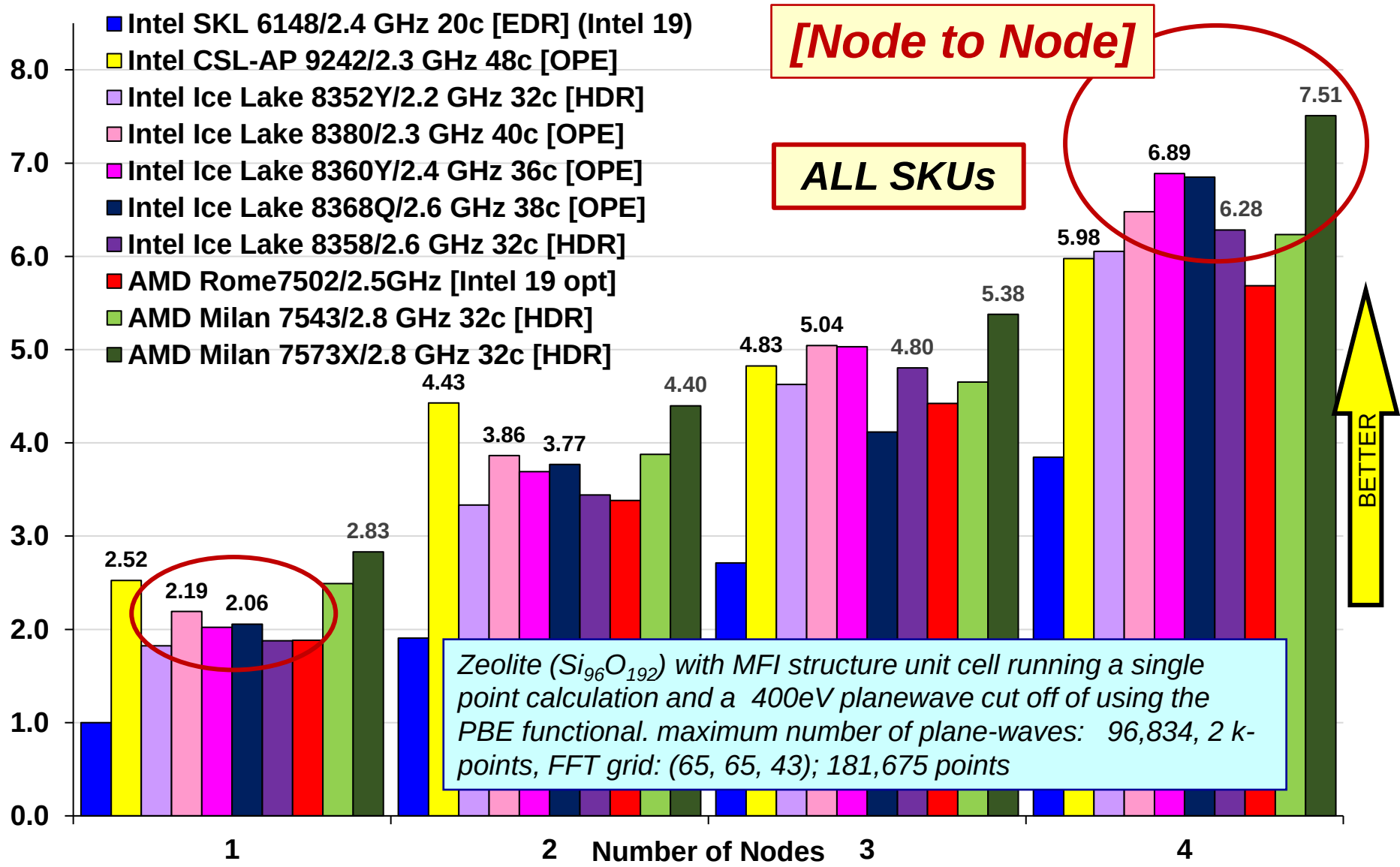
VASP 6.3 – Zeolite Benchmark - Parallelisation on k-points

Performance *Relative to the Hawk SKL 6148 2.4 GHz (1 node)*

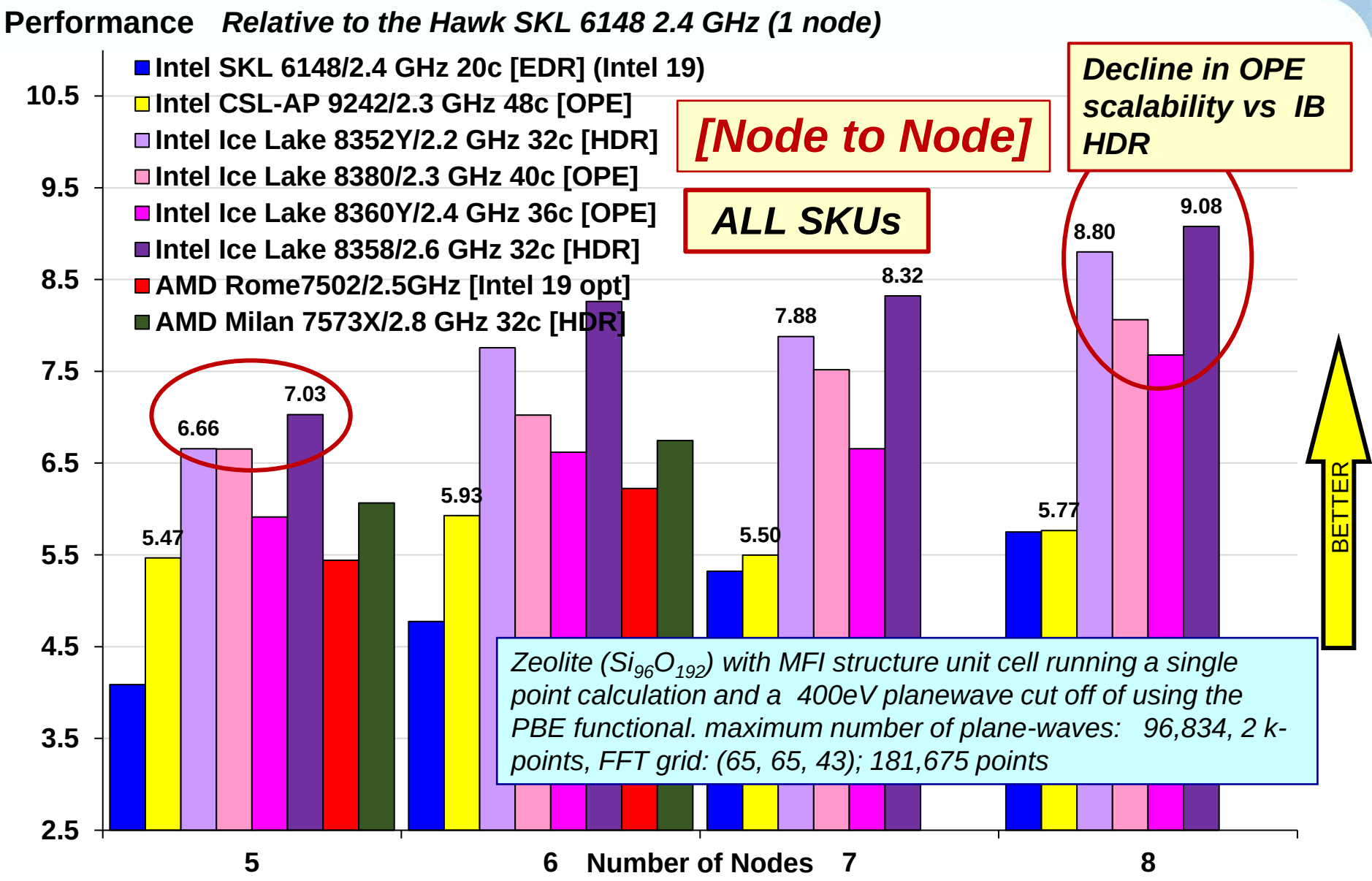


VASP 6.3 – Zeolite Benchmark - Parallelisation on k-points

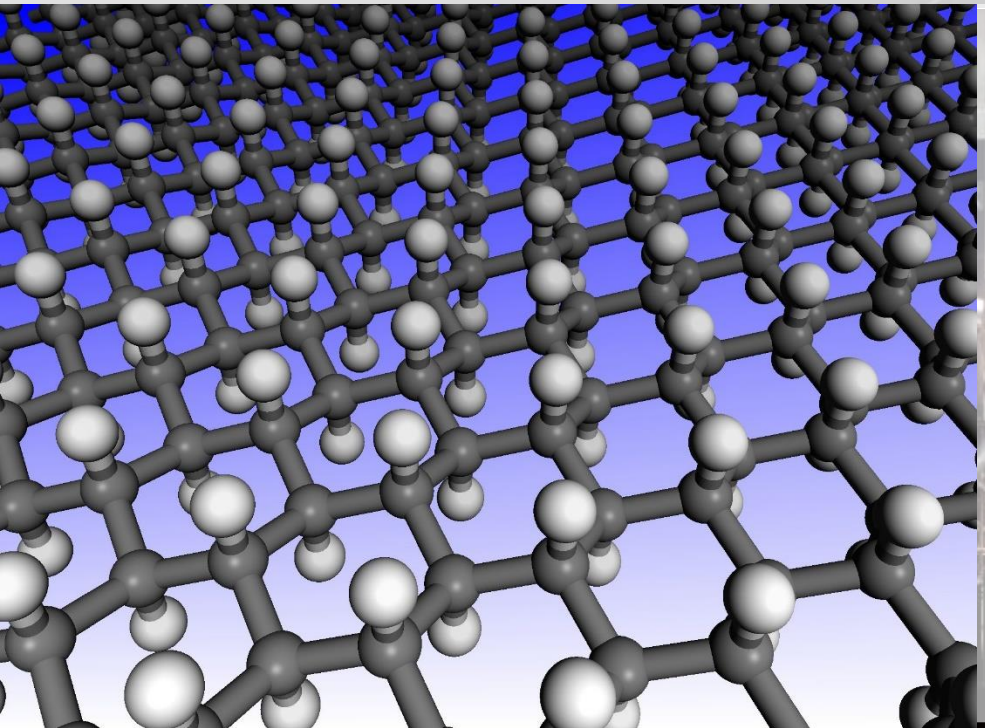
Performance *Relative to the Hawk SKL 6148 2.4 GHz (1 node)*



VASP 6.3 – Zeolite Benchmark - Parallelisation on k-points



Performance of Computational Chemistry and Ocean Modelling Codes



Advanced Materials Software:
2. CASTEP

CASTEP – Materials Modelling

- **CASTEP** is a full-featured materials modelling code based on a first-principles quantum mechanical description of electrons and nuclei. It uses the robust methods of a plane-wave basis set and pseudopotentials.
- Two versions of CASTEP used in this study, **Version 19.1.1** and the current academic release of CASTEP, **Version 21.1.1**.

- **Al3x3 Benchmark**

  The al3x3 simulation cell comprises a 270-atom sapphire surface, with a vacuum gap. There are only 2 k-points, so it is a good test of the performance of CASTEP's other parallelisation strategies.

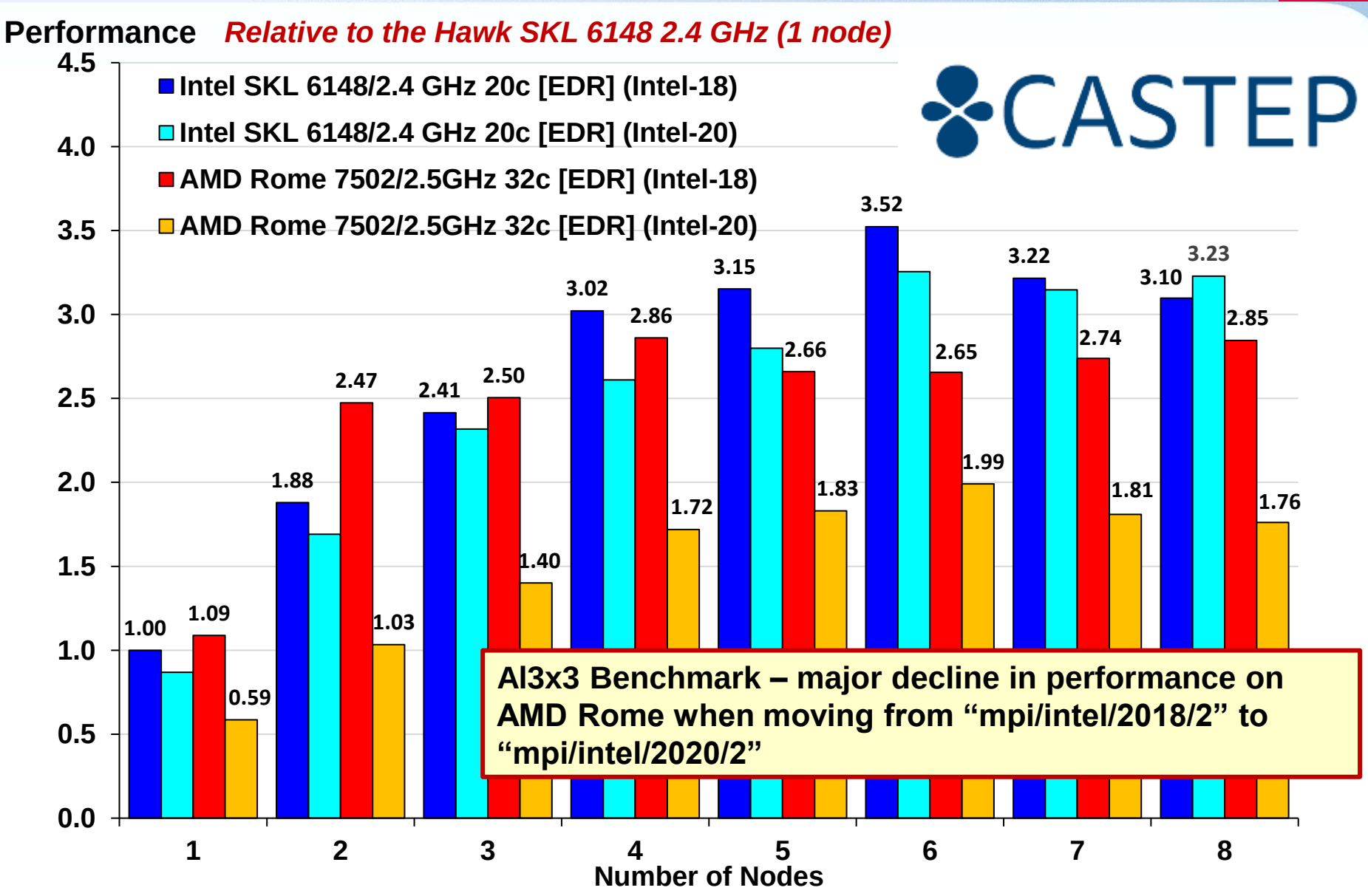
- **MnO_2 Benchmark**

  Bigger calculation (313 electrons and 64 ions) and involves MPI AllToAllV across all processors.

- **IDZ Benchmark**

  Longer MD calculation (1104 electrons and 404 ions) requiring several random initializations (16 MD iterations in total).

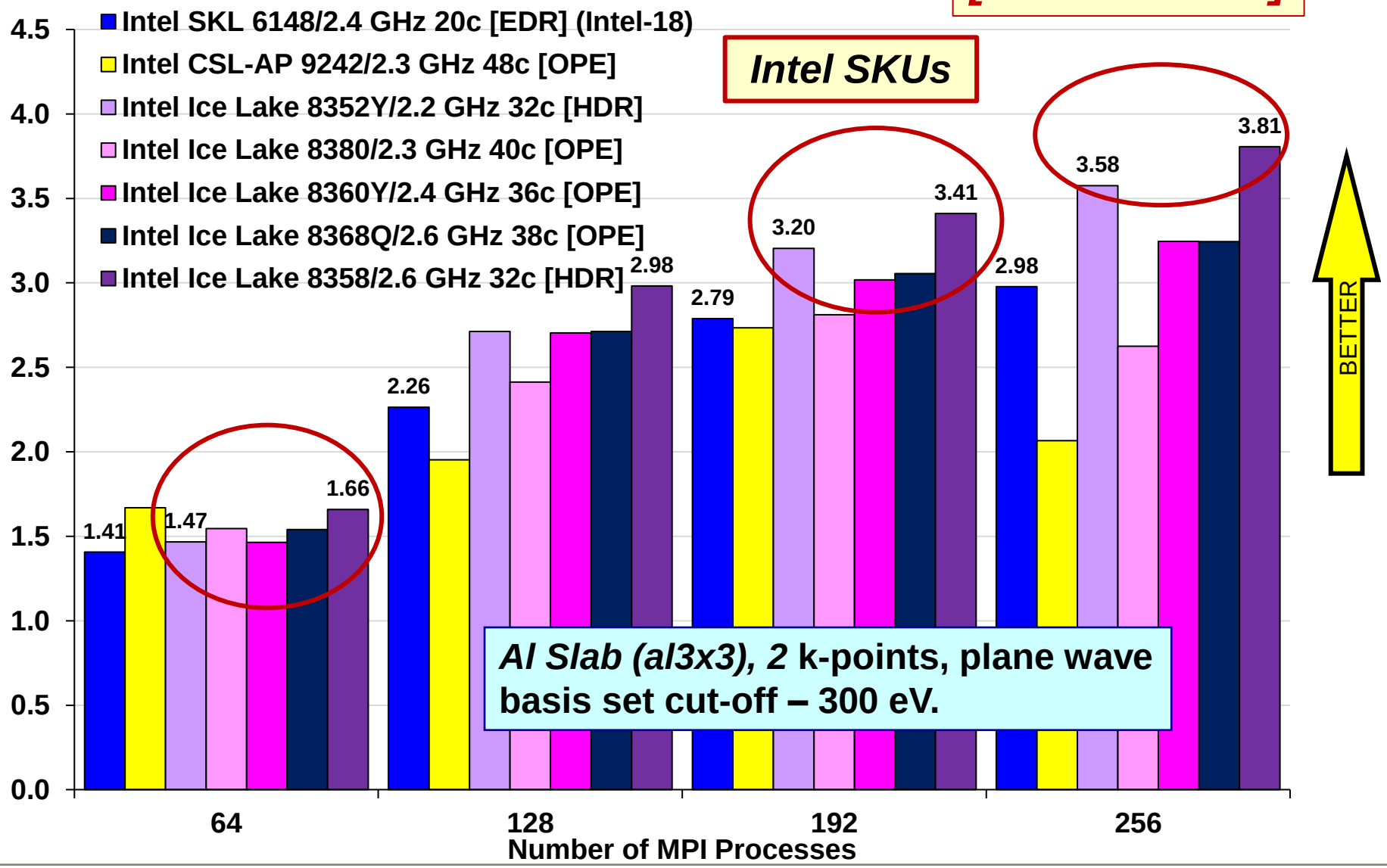
CASTEP – Impact of Intel MPI version on AMD clusters



CASTEP 19 – Al Slab (al3x3) Benchmark

Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*

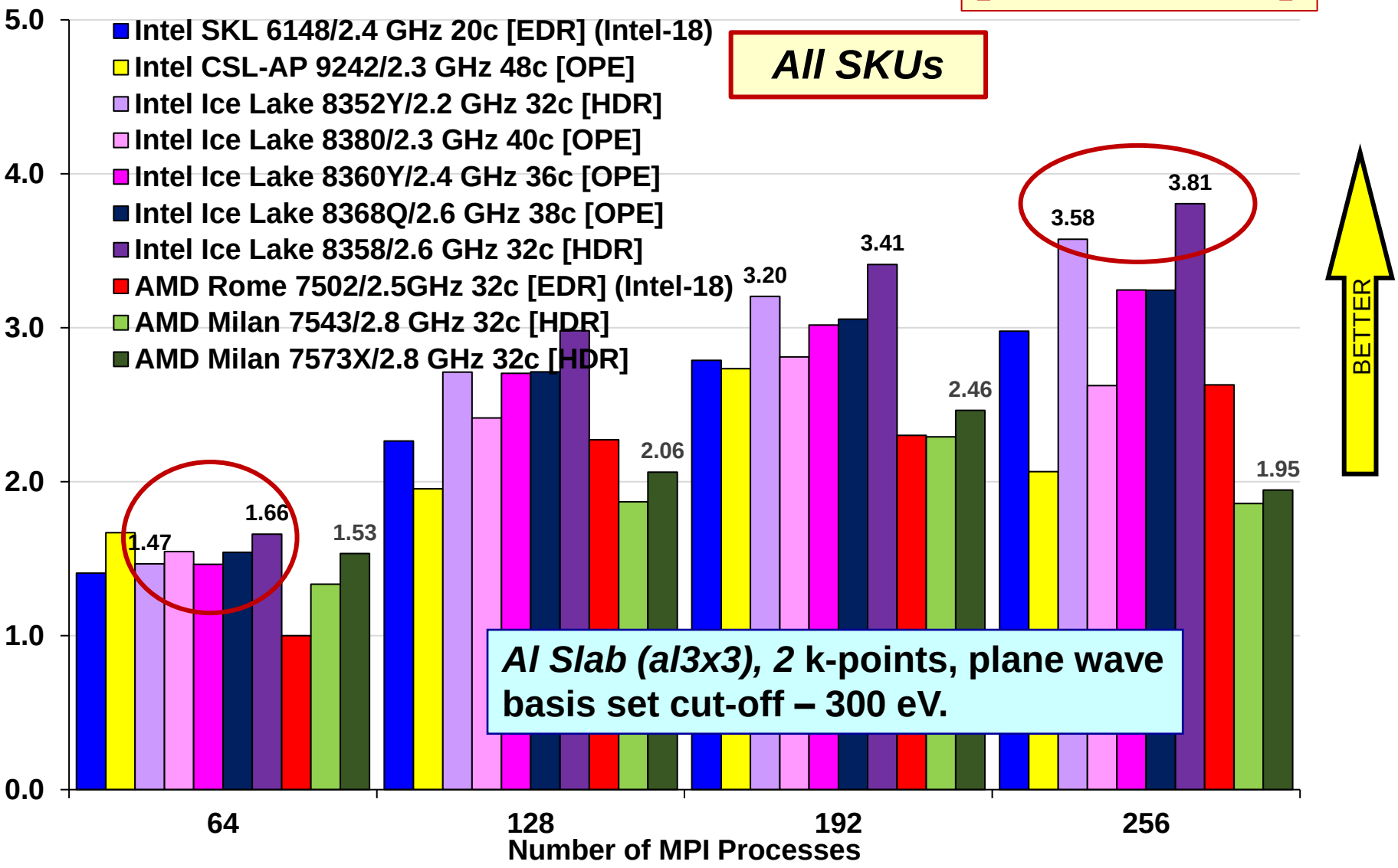
[Core to core]



CASTEP 19 – Al Slab (al3x3) Benchmark

Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*

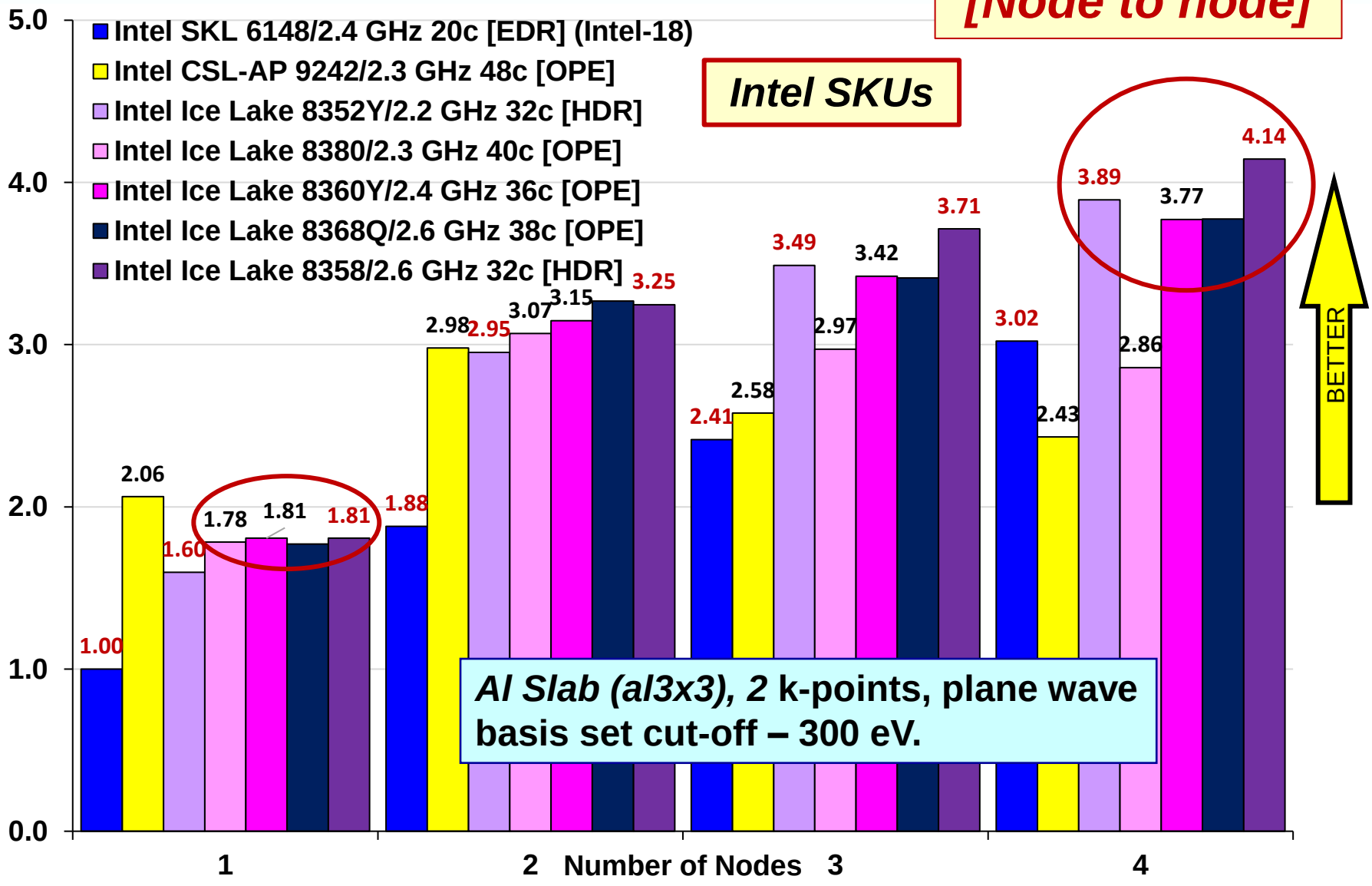
[Core to core]



CASTEP 19 – Al Slab (al3x3) Benchmark

Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*

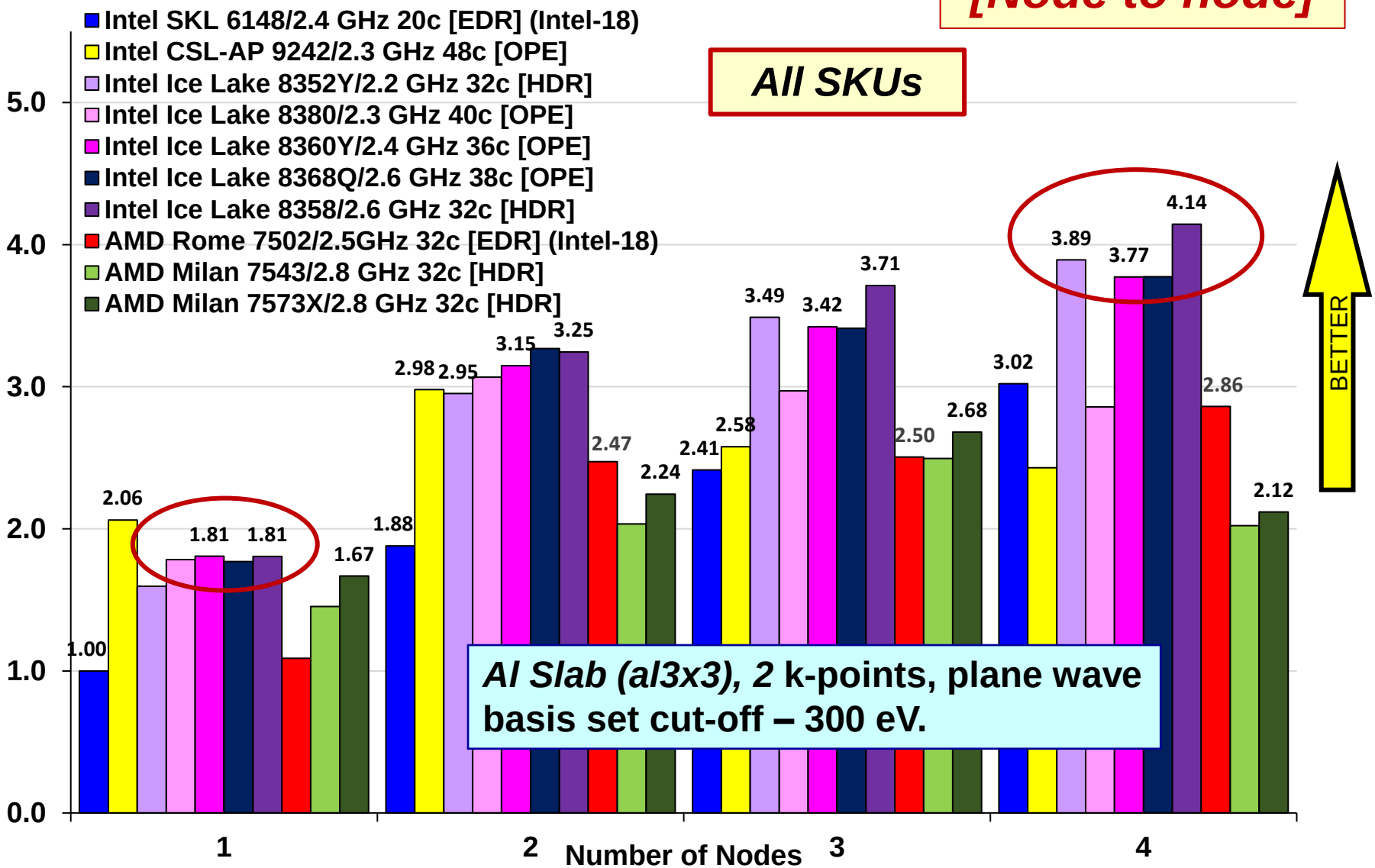
[Node to node]



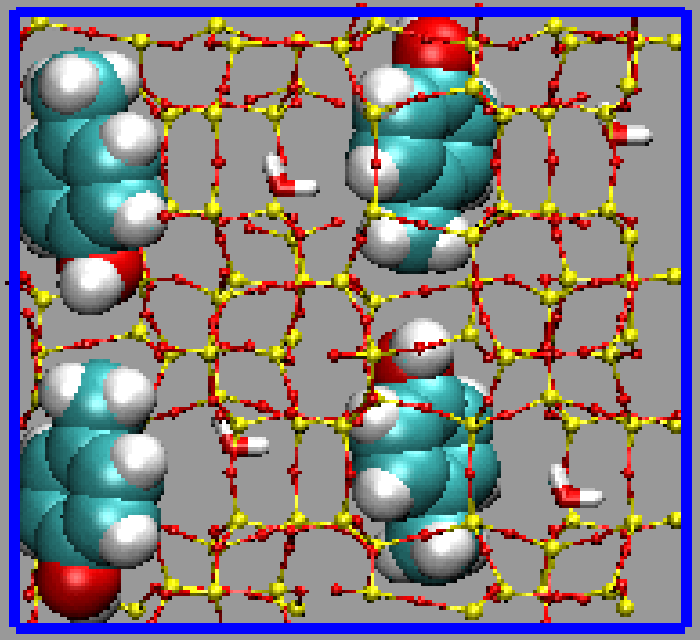
CASTEP 19 – Al Slab (al3x3) Benchmark

Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*

[Node to node]



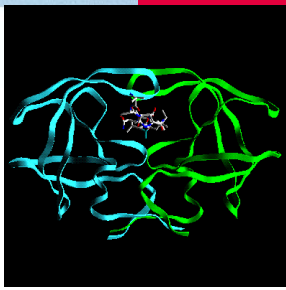
Performance of Computational Chemistry and Ocean Modelling Codes



Electronic Structure

GAMESS -UK

The MPI/ScaLAPACK Implementation of the GAMESS-UK SCF/DFT module

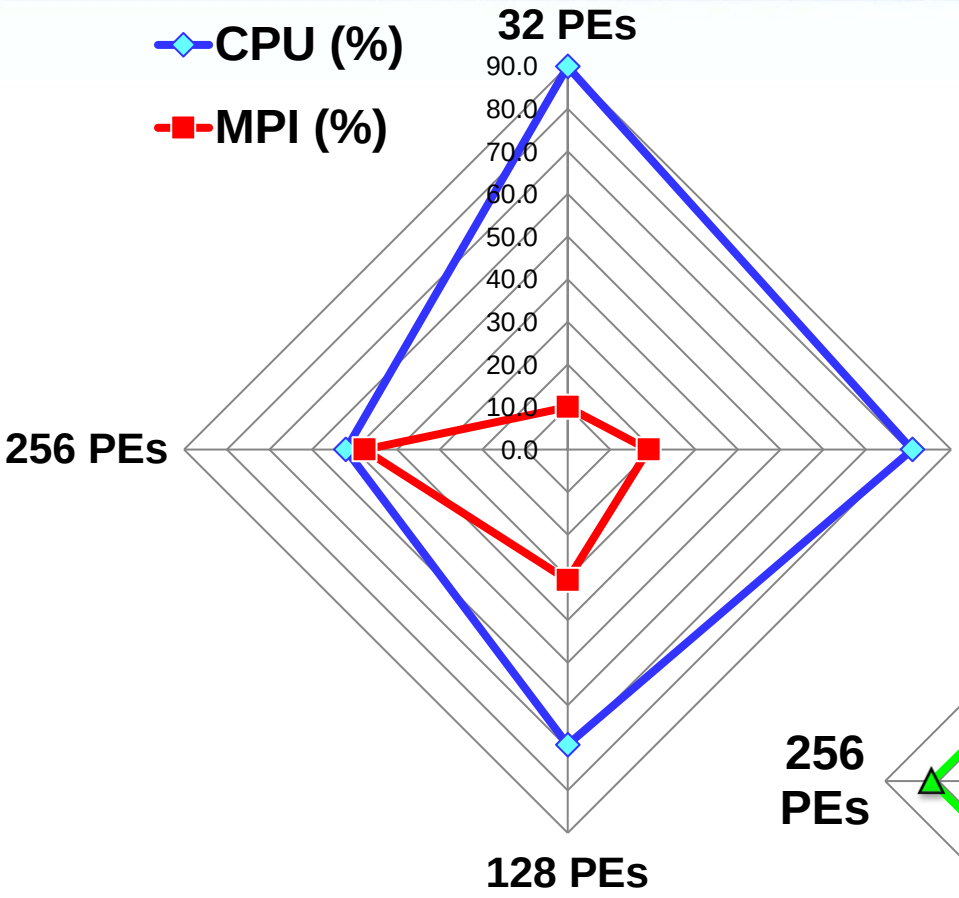


- Pragmatic approach to the replicated data constraints:
- MPI-based tools (such as ScaLAPACK) used in place of Global Arrays
- All data structures except those required for the Fock matrix build are fully distributed (F, P)

- Partially distributed model chosen because, in the absence of efficient one-sided communications it is difficult to efficiently load balance a distributed Fock matrix build.
- Obvious drawback - some large replicated data structures are required.
  - These are kept to a minimum. For a closed shell HF or DFT calculation only **2 replicated matrices** are required, 1 × Fock and 1 × Density (doubled for UHF).

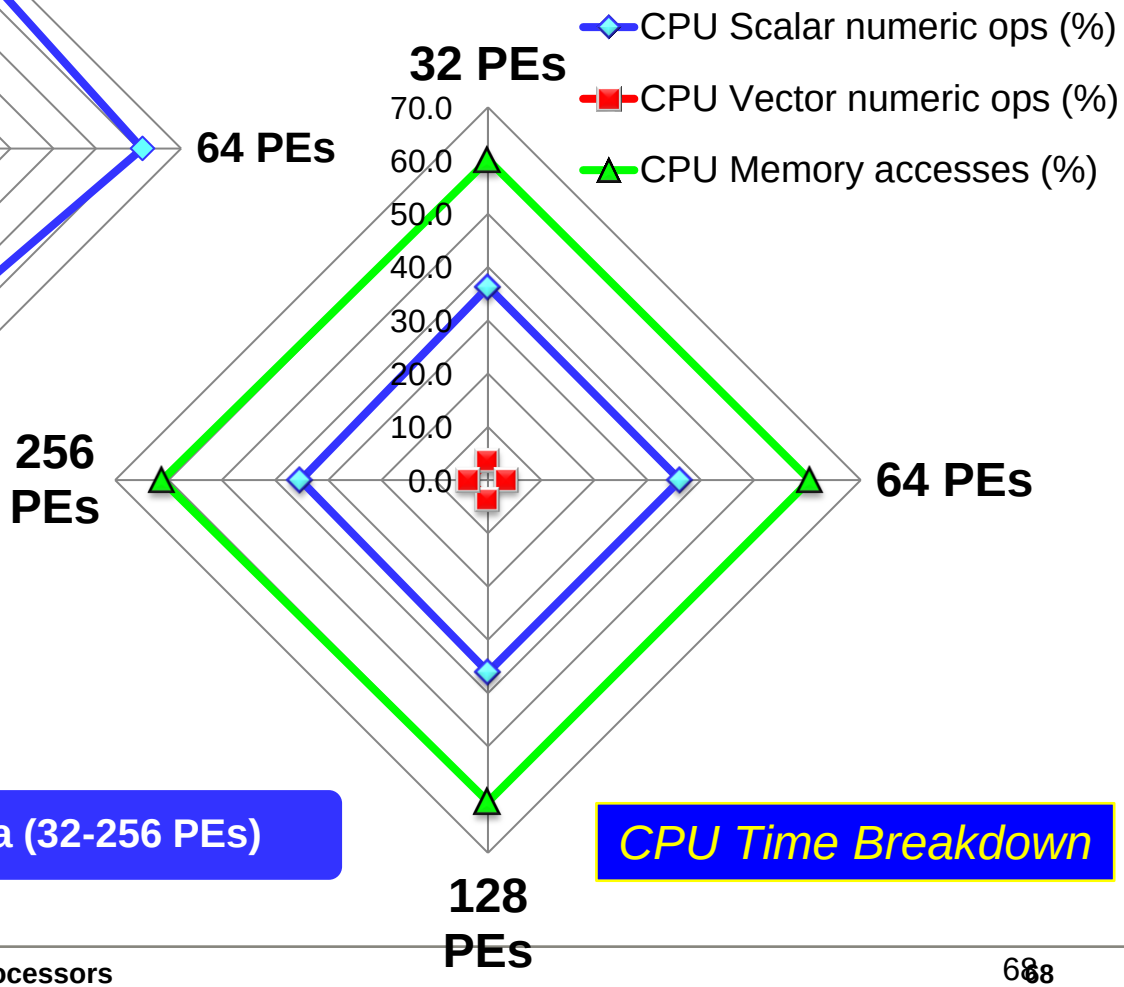
"The GAMESS-UK electronic structure package: algorithms, developments and applications"
M.F. Guest, I. J. Bush, H.J.J. van Dam, P. Sherwood, J.M.H. Thomas, J.H. van Lenthe, R.W.A Havenith, J. Kendrick, Mol. Phys. 103, No. 6-8, 2005, 719-747.

GAMESS-UK.MPI DFT – DFT Performance Report



Cyclosporin 6-31G** basis (1855 GTOs); DFT B3LYP

Total Wallclock Time Breakdown



Performance Data (32-256 PEs)

CPU Time Breakdown

GAMESS-UK Performance - Zeolite Y cluster

[Core to core]

Performance

Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)

- Intel SKL 6148/2.4 GHz 20c [EDR]
- Intel CSL 6248/2.5 GHz 20c [HDR]
- Intel CSL-AP 9242/2.3 GHz 48c [OPE]
- Intel Ice Lake 8352Y/2.2 GHz 32c [HDR]
- Intel Ice Lake 8380/2.3 GHz 40c [OPE]
- Intel Ice Lake 8360Y/2.4 GHz 36c [OPE]
- Intel Ice Lake 8358/2.6 GHz 32c [HDR]
- Intel Ice Lake 8368Q/2.6 GHz 38c [OPE]

Intel SKUs

Performance Data (64-256 PEs)

BETTER

1.00 1.44 1.44 1.46 1.80 2.41 2.66 2.50 2.58 3.43 3.75 3.62 3.29 4.32 4.84 4.51

Zeolite Y cluster SioSi7 DZVP (Si,O), DZVP2 (H) B3LYP(3975 GTOs)

5.5 5.0 4.5 4.0 3.5 3.0 2.5 2.0 1.5 1.0 0.5 0.0

64 128 192 256

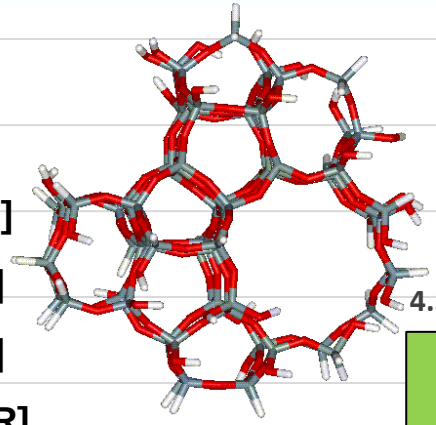
Number of MPI Processes

GAMESS-UK Performance - Zeolite Y cluster

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

[Core to core]

Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)



Performance

- Intel SKL 6148/2.4 GHz 20c [EDR]
- AMD Rome 7502/2.5GHz 32c [EDR]
- AMD Rome 7702/2.0 GHz 64c [HDR]
- AMD Milan 7713/2.0 GHz 64c [HDR]
- AMD Milan 7543/2.8 GHz 32c [HDR]
- AMD Milan 7573X/2.6 GHz 32c [HDR]

AMD SKUs

Number of MPI Processes	64	128	192	256
Intel SKL 6148/2.4 GHz 20c [EDR]	1.00	1.80	2.58	3.29
AMD Rome 7502/2.5GHz 32c [EDR]	1.23	2.19	3.07	3.90
AMD Rome 7702/2.0 GHz 64c [HDR]				
AMD Milan 7713/2.0 GHz 64c [HDR]				
AMD Milan 7543/2.8 GHz 32c [HDR]	1.71	3.07	4.30	5.57
AMD Milan 7573X/2.6 GHz 32c [HDR]	1.78	3.19	4.47	5.66

Performance Data (64-256 PEs)

BETTER

Zeolite Y cluster SioSi7 DZVP (Si,O), DZVP2 (H) B3LYP(3975 GTOs)

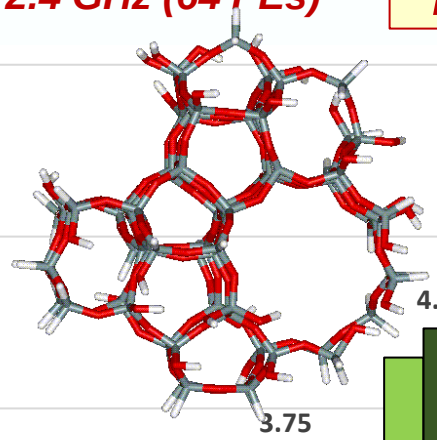
GAMESS-UK Performance - Zeolite Y cluster

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Performance

Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)

[Core to core]



- Intel SKL 6148/2.4 GHz 20c [EDR]
- Intel Ice Lake 8352Y/2.2 GHz 32c [HDR]
- Intel Ice Lake 8380/2.3 GHz 40c [OPE]
- Intel Ice Lake 8360Y/2.4 GHz 36c [OPE]
- Intel Ice Lake 8368Q/2.6 GHz 38c [OPE]
- Intel Ice Lake 8358/2.6 GHz 32c [HDR]
- AMD Rome 7502/2.5GHz 32c [EDR]
- AMD Milan 7713/2.0 GHz 64c [HDR]
- AMD Milan 7543/2.8 GHz 32c [HDR]
- AMD Milan 7573X/2.6 GHz 32c [HDR]

All SKUs

Performance Data (64-256 PEs)

BETTER

Zeolite Y cluster SioSi7 DZVP (Si,O), DZVP2 (H) B3LYP(3975 GTOs)

6.0
5.0
4.0
3.0
2.0
1.0
0.0

1.00 1.44 1.44 1.78 1.80 2.41 2.66 3.19 2.58 3.43 3.75 4.47 3.29 4.32 4.84 5.66

64 128 192 256

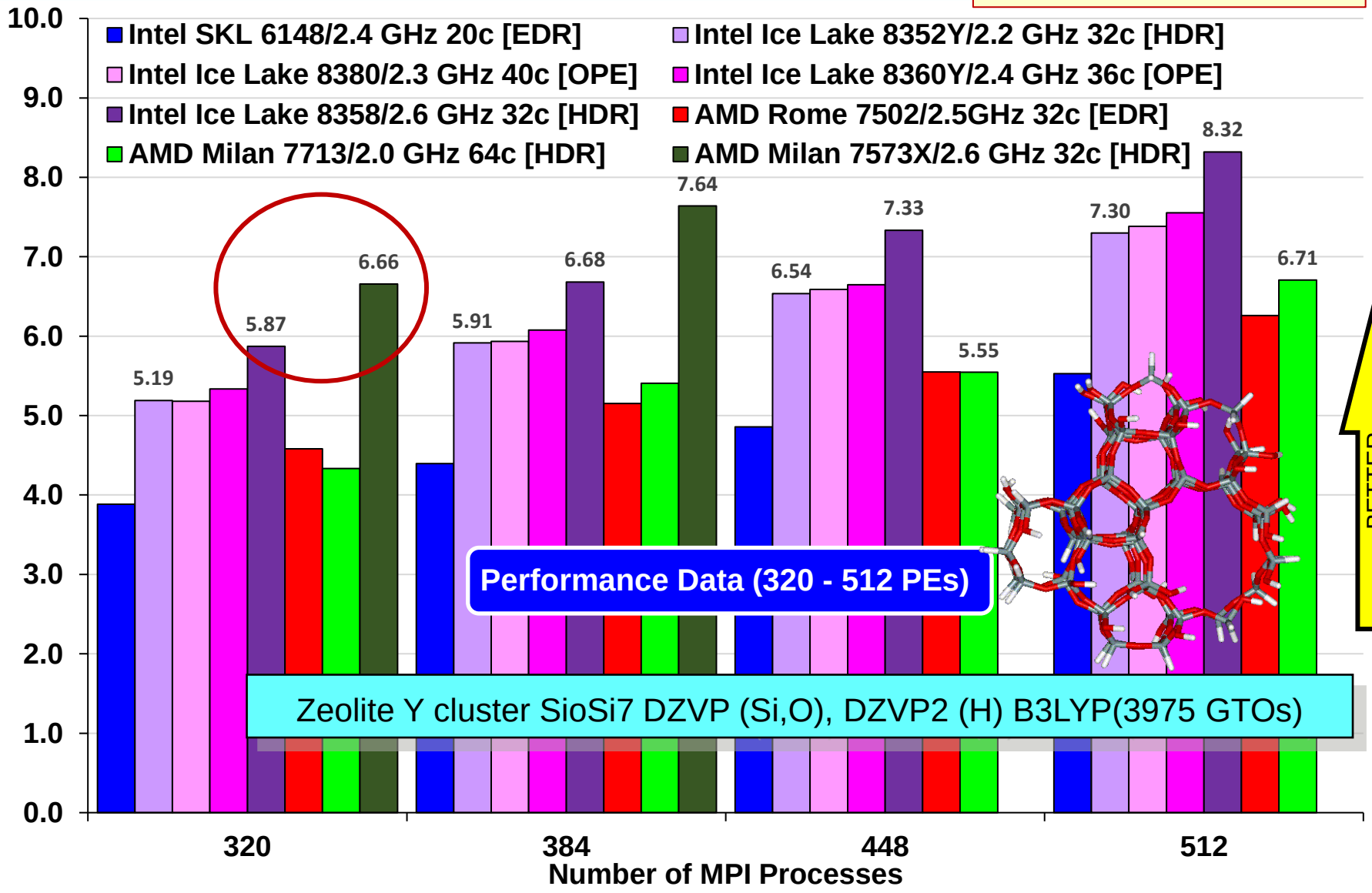
Number of MPI Processes

GAMESS-UK Performance - Zeolite Y cluster

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*

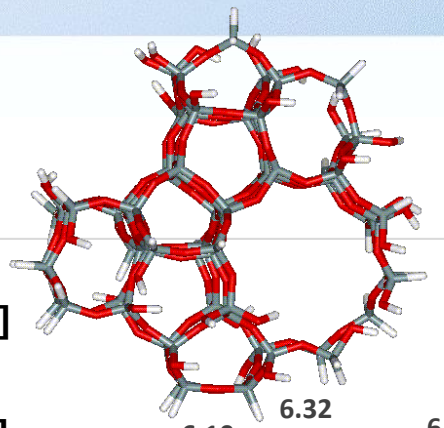
[Core to core]



GAMESS-UK Performance - Zeolite Y cluster

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Relative to the Hawk SKL 6148 2.4 GHz (40 PEs)



Intel SKUs

- Intel SKL 6148/2.4 GHz 20c [EDR]
- Intel CSL 6248/2.5 GHz 20c [HDR]
- Intel CSL-AP 9242/2.3 GHz 48c [OPE]
- Intel Ice Lake 8352Y/2.2 GHz 32c [HDR]
- Intel Ice Lake 8380/2.3 GHz 40c [OPE]
- Intel Ice Lake 8360Y/2.4 GHz 36c [OPE]
- Intel Ice Lake 8358/2.6 GHz 32c [HDR]
- Intel Ice Lake 8368Q/2.6 GHz 38c [OPE]

[Node to Node]

Performance

0.0 2.0 4.0 6.0 8.0

1.00 1.16 2.44 2.07 2.54 2.31 2.31 2.42

1.88 2.24 4.35 3.87 4.72 4.35 4.33 4.64

2.60 3.12 6.10 5.16 6.32 5.90 6.00 6.16

3.51 4.23 7.64 6.78 8.35 7.78 7.75 8.14

1 2 3 4

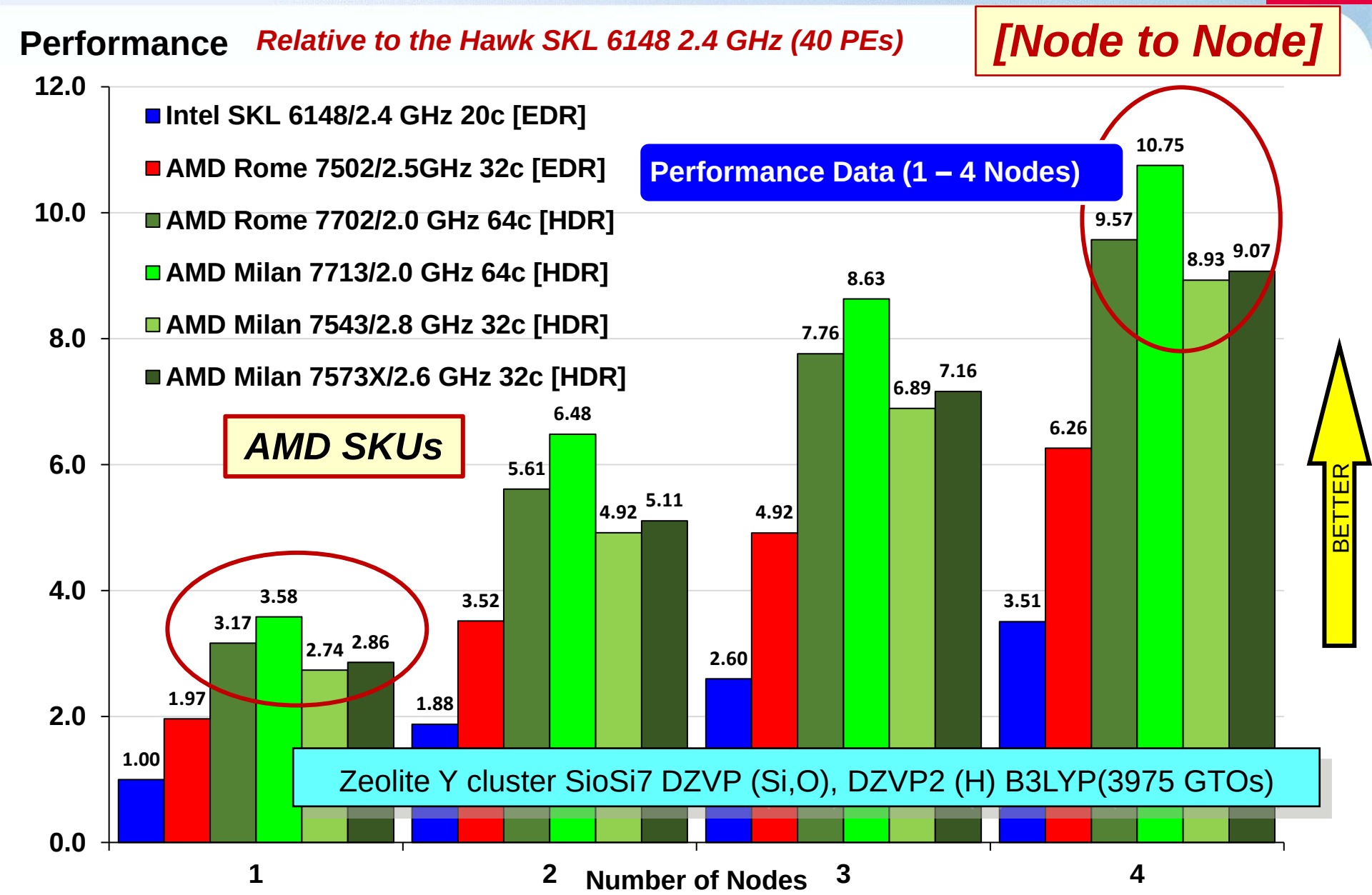
Number of Nodes

BETTER

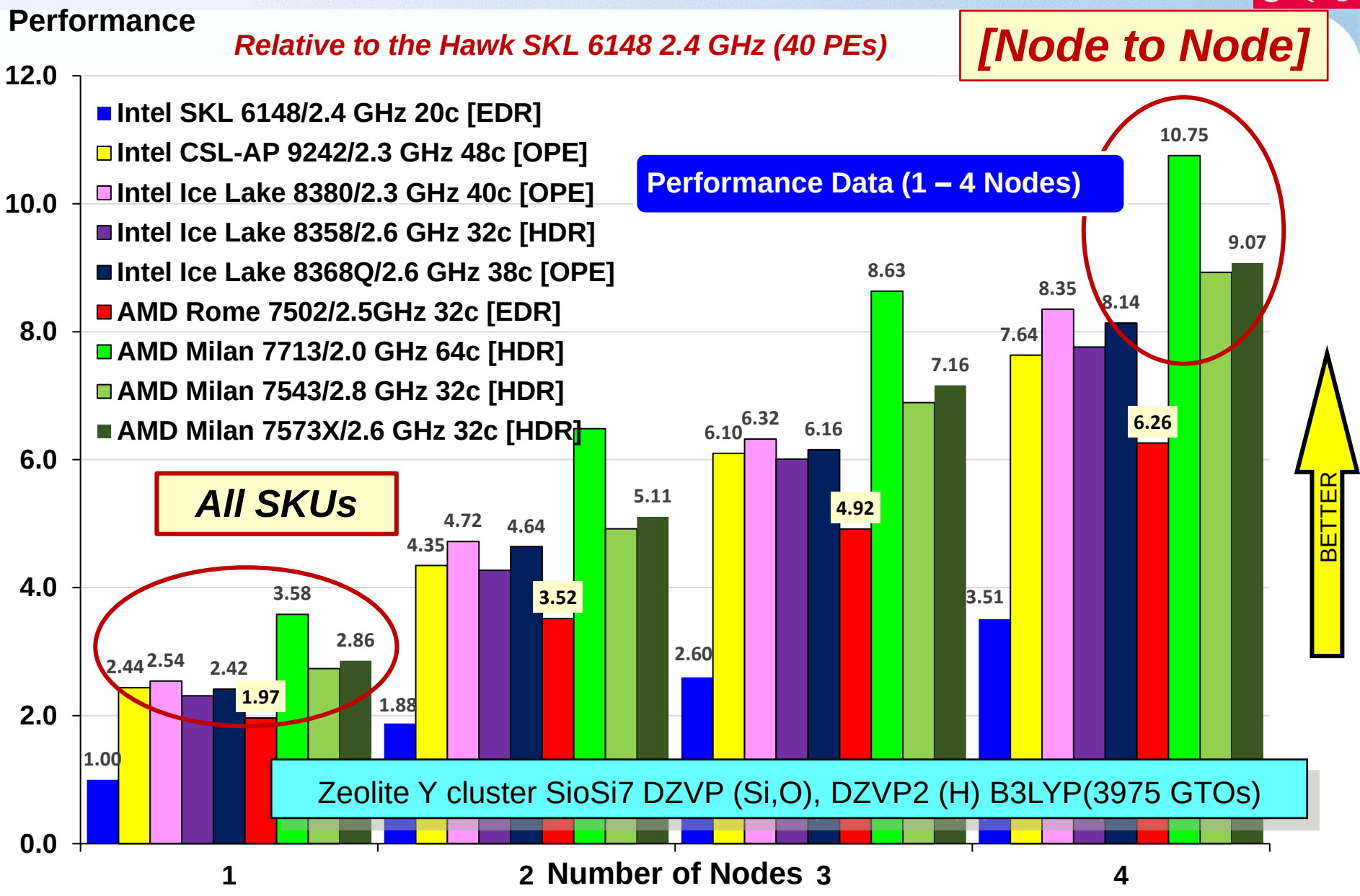
Performance Data (1 – 4 Nodes)

Zeolite Y cluster SioSi7 DZVP (Si,O), DZVP2 (H) B3LYP(3975 GTOs)

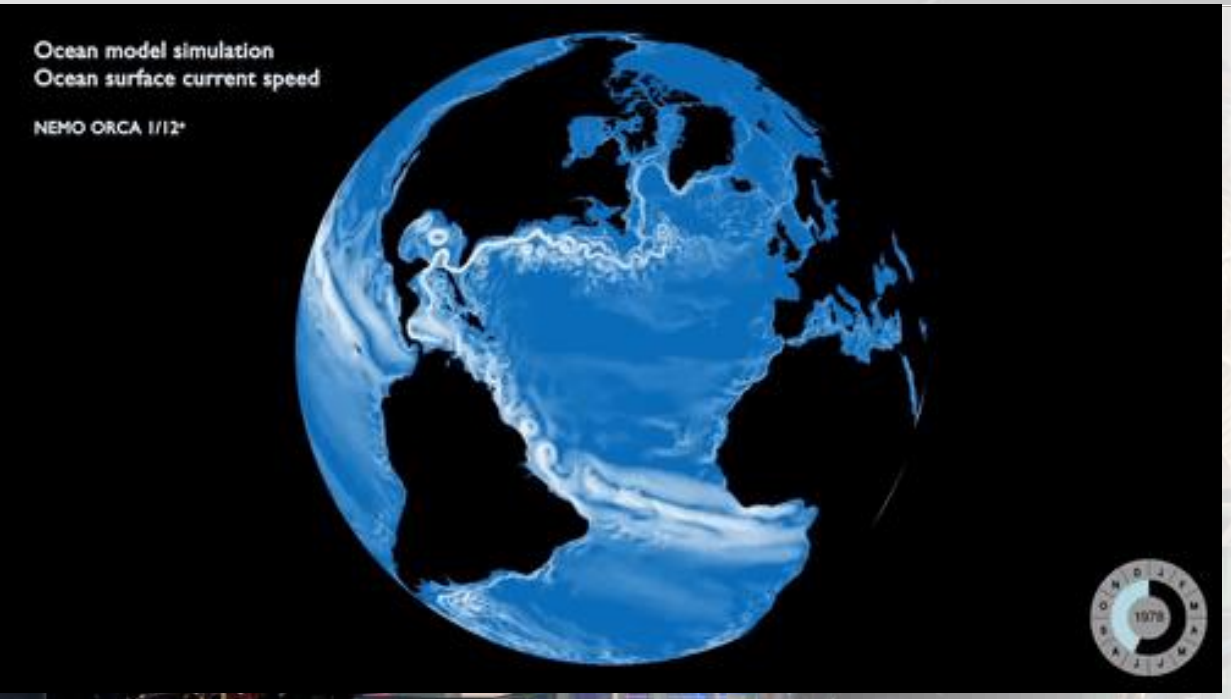
GAMESS-UK Performance - Zeolite Y cluster



GAMESS-UK Performance - Zeolite Y cluster



Performance of Computational Chemistry and Ocean Modelling Codes



Ocean Modelling: NEMO and FVCOM

Performance of Ocean Modelling Codes

- Assistance provided to **The Marine Systems Modelling Group at Plymouth Marine Laboratory.**
- At the heart of much of the group's work are two numerical models of the ocean's circulation:

  The NEMO Community Ocean Model

  A prognostic, primitive equation ocean circulation model for studying problems relating to both the global ocean and marginal seas. Uses a ***structured* model grid.**

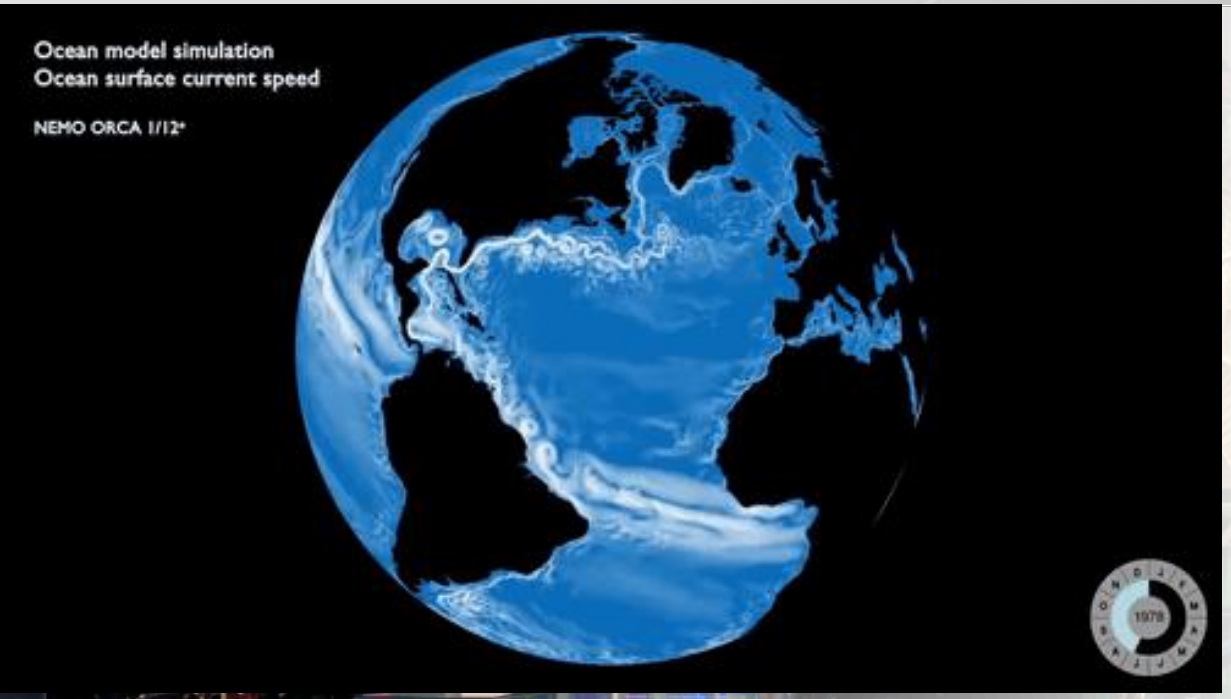
  The Finite Volume Community Ocean Model (FVCOM)

  A prognostic, primitive equation ocean circulation model for (mainly) studying problems relating to estuarine and coastal environments. **Uses an *unstructured* model grid.**

- Both models are often run with a **biogeochemical model called ERSEM -** significantly increases the compute & memory requirements.
- To be run efficiently, both models require a CPU based HPC system

Performance of Ocean Modelling Codes

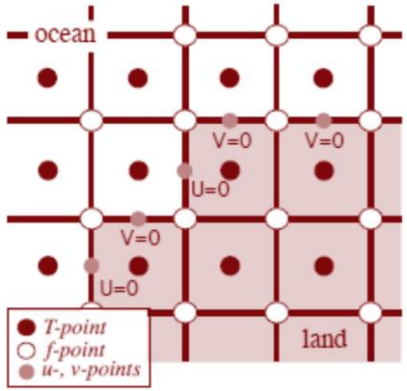
NEMO - Nucleus for European Modelling of the Ocean



1. NEMO

Primary Requirements for PML

NEMO* - Data parallelism through domain decomposition

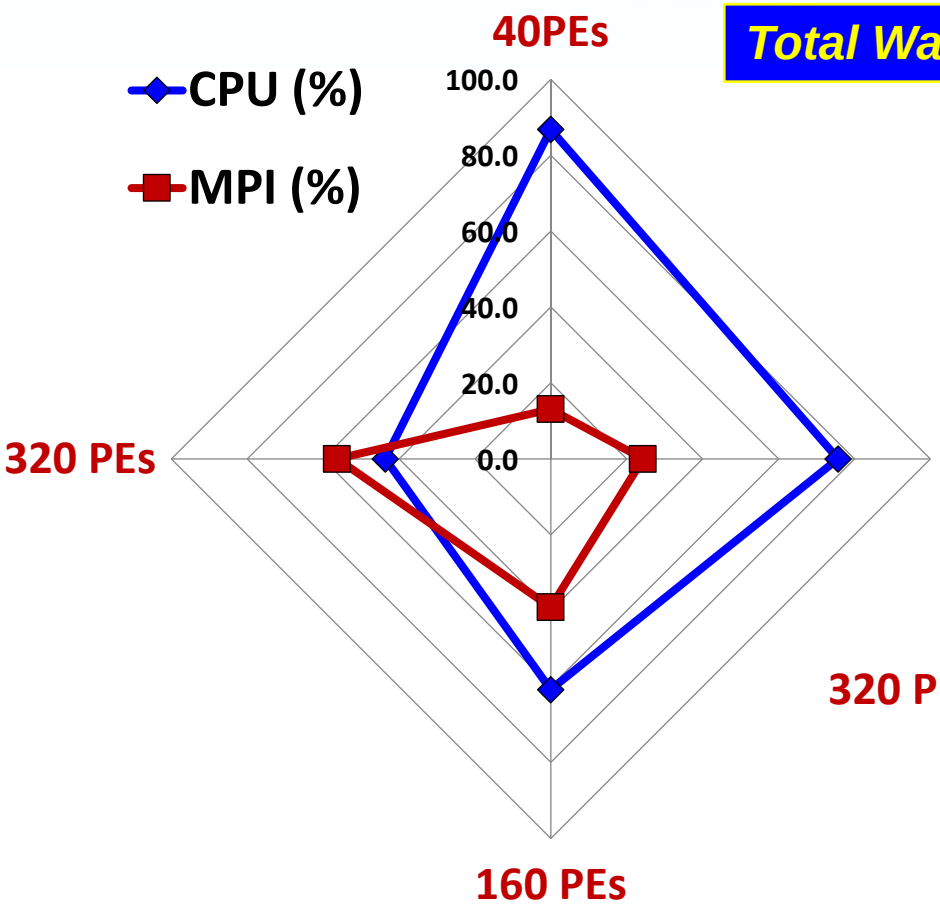


Example: The North Atlantic Ocean

- 773 x 1236 horizontal grid points, multiplied by ‘k’ depth levels.
- Full horizontal domain split into 9 x 20 sub-domains.
- Each subdomain is handled by a separate core during parallel runs.
- MPI for handling communication between subdomains.
- Known memory B/W issues – avoid full node occupancy

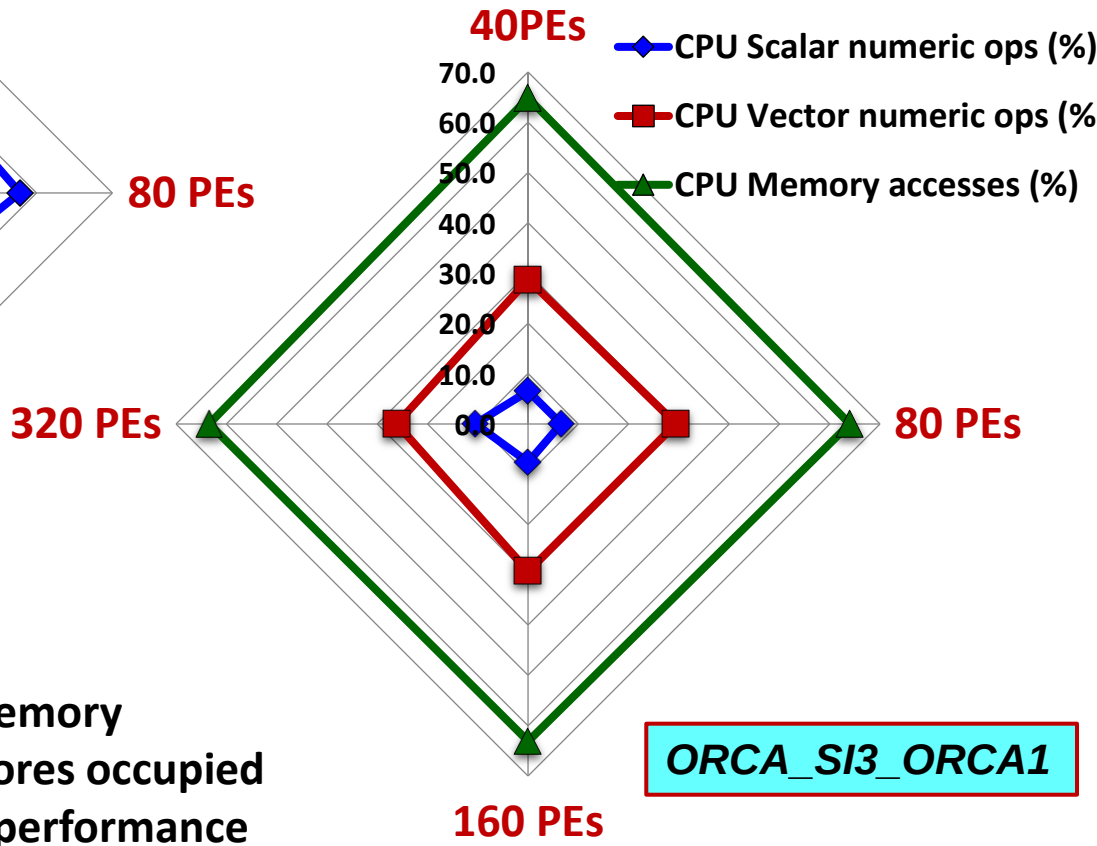
***FVCOM** employs a similar approach to parallelism, albeit based upon an unstructured, triangular horizontal mesh.

NEMO – ORCA_SI3 Model Performance Report



Total Wallclock Time Breakdown

horizontal resolutions of 1-degree



NEMO performance is dominated by memory bandwidth – running with 50% of the cores occupied on each Hawk node typically improves performance by ca. 1.6 for a fixed number of MPI processes.

ORCA_SI3_ORCA1

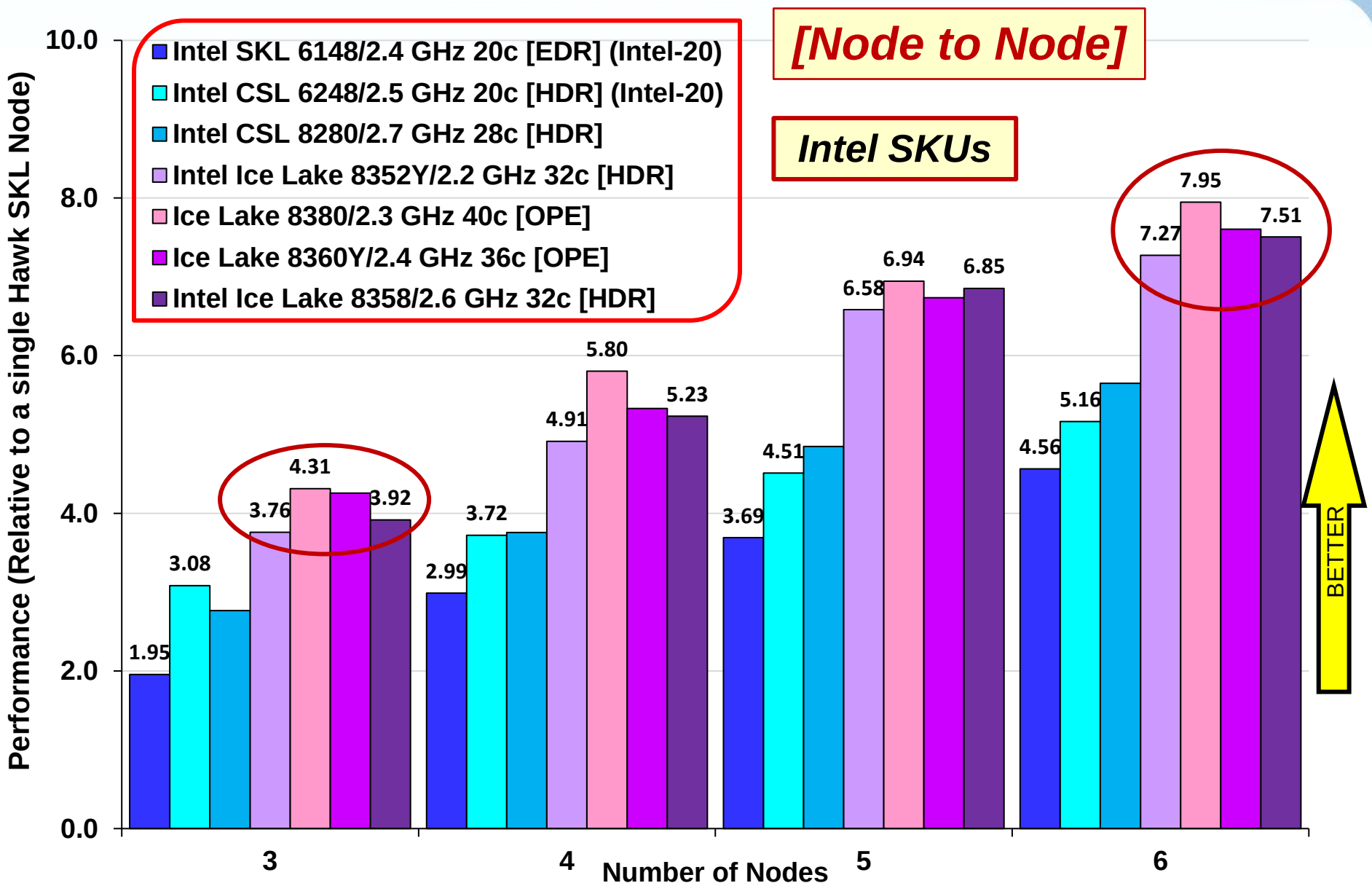
CPU Time Breakdown

The NEMO-ERSEM Benchmark

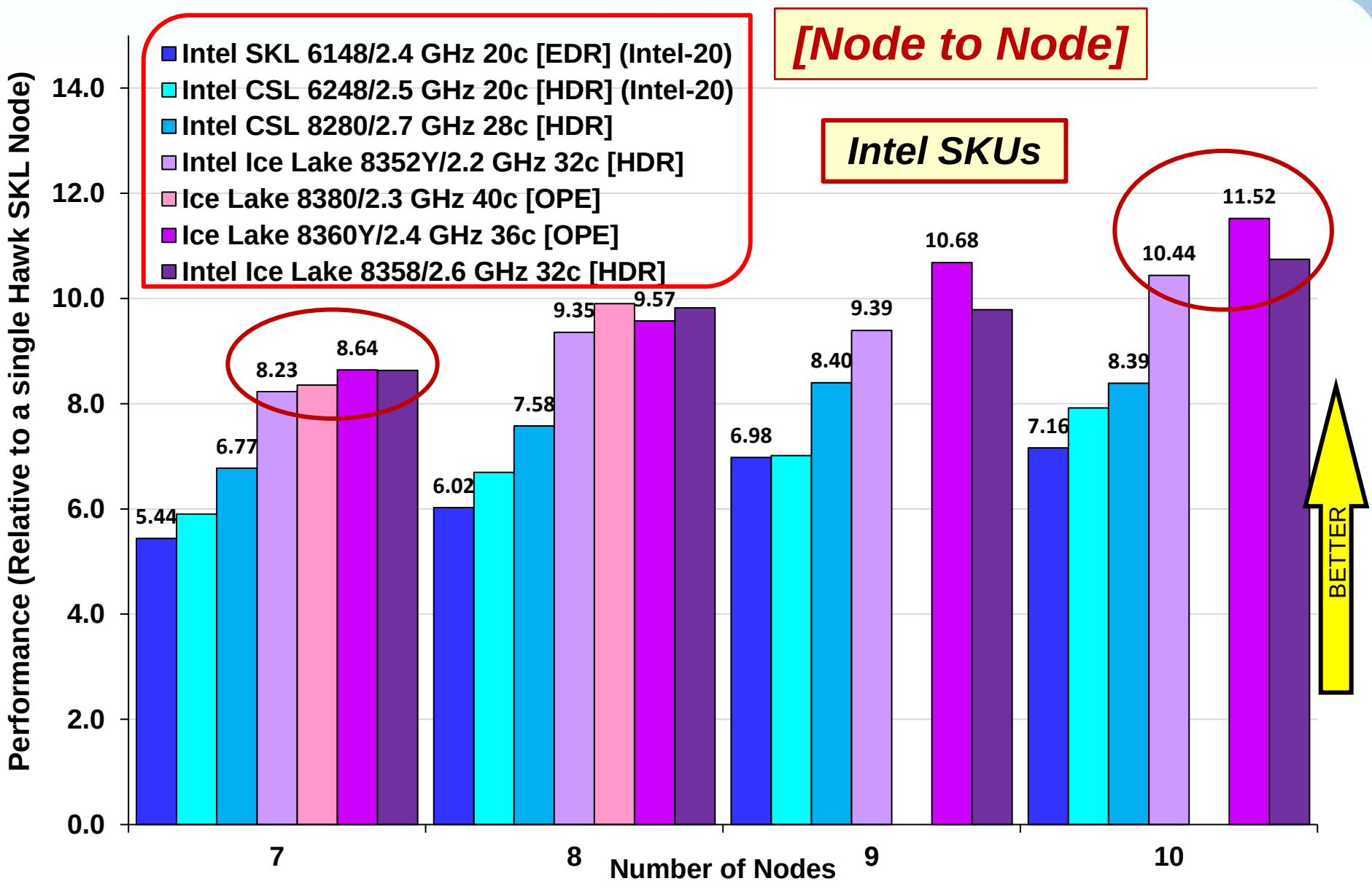
- NEMO, "Nucleus for European Modelling of the Ocean" is a modelling framework for research activities and forecasting services in ocean and climate sciences, developed by a European consortium. (https://www.nemo-ocean.eu)
- NEMO is a **memory-bandwidth limited code** where performance can be improved by part-populating nodes.
- ERSEM, “European Regional Seas Ecosystem Model” is a bio-geochemical and ecosystem mode, developed at PML (https://github.com/pmlmodelling/ersem)
- **Benchmark Case:** NEMO-FABM-ERSEM on the AMM7 domain covering the NW European shelf at ca. 7 km resolution. Four elements to the code (a) **XIOS**: an I/O library, (b) **ERSEM:** Biogeochemical model code, (c) **FABM:** Interface between ERSEM and NEMO and (d) **NEMO**.
- Compilation requires **parallel netcdf and hdf5 libraries**. Several cores are allocated to the I/O server XIOS, with remainder allocated to NEMO:

mpirun -n *$XIOSCORES* $code_xios : -n *$OCEANCORES* $code_nemo

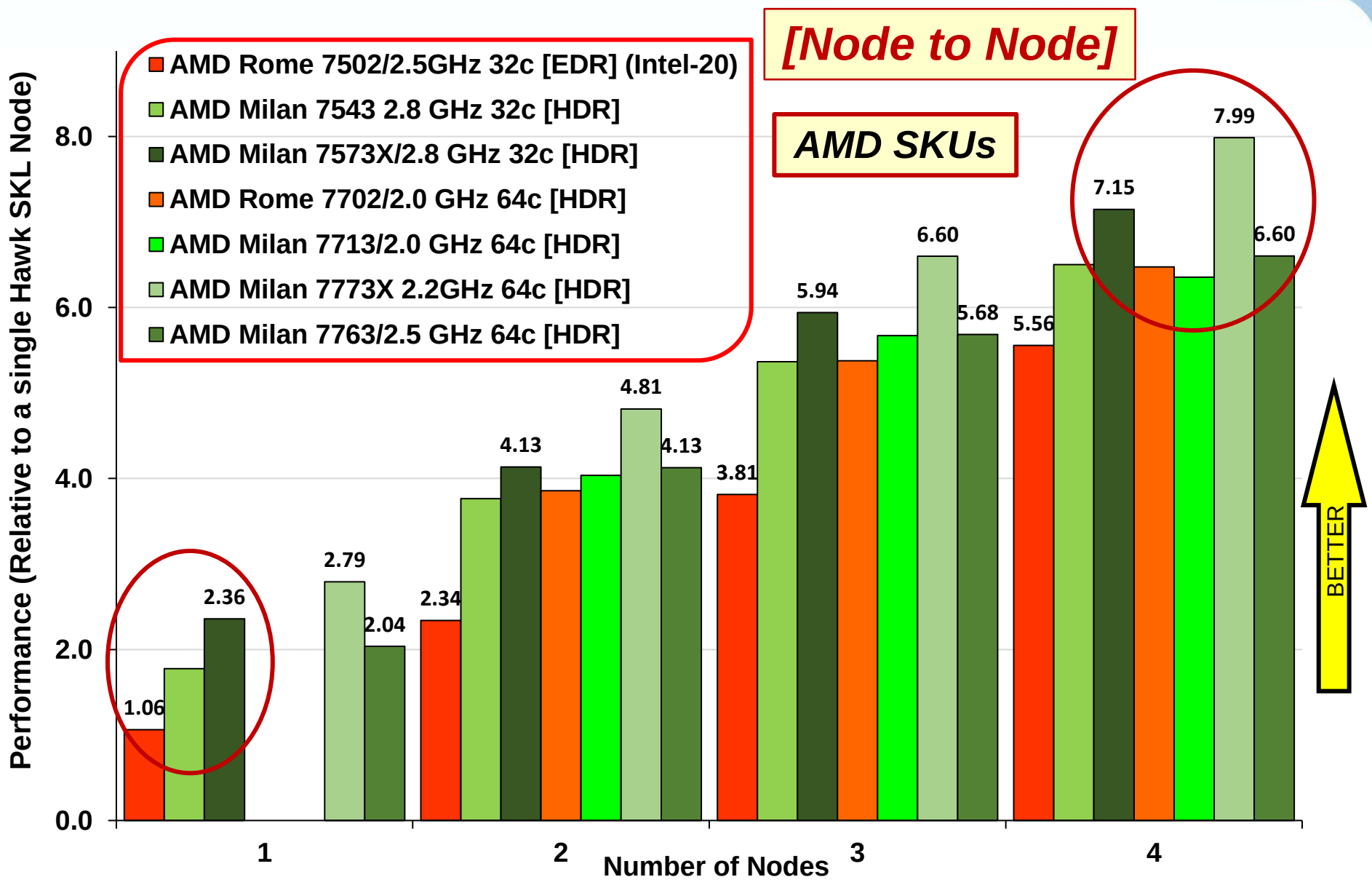
NEMO-FABM-ERSEM (AMM7) – Node Performance



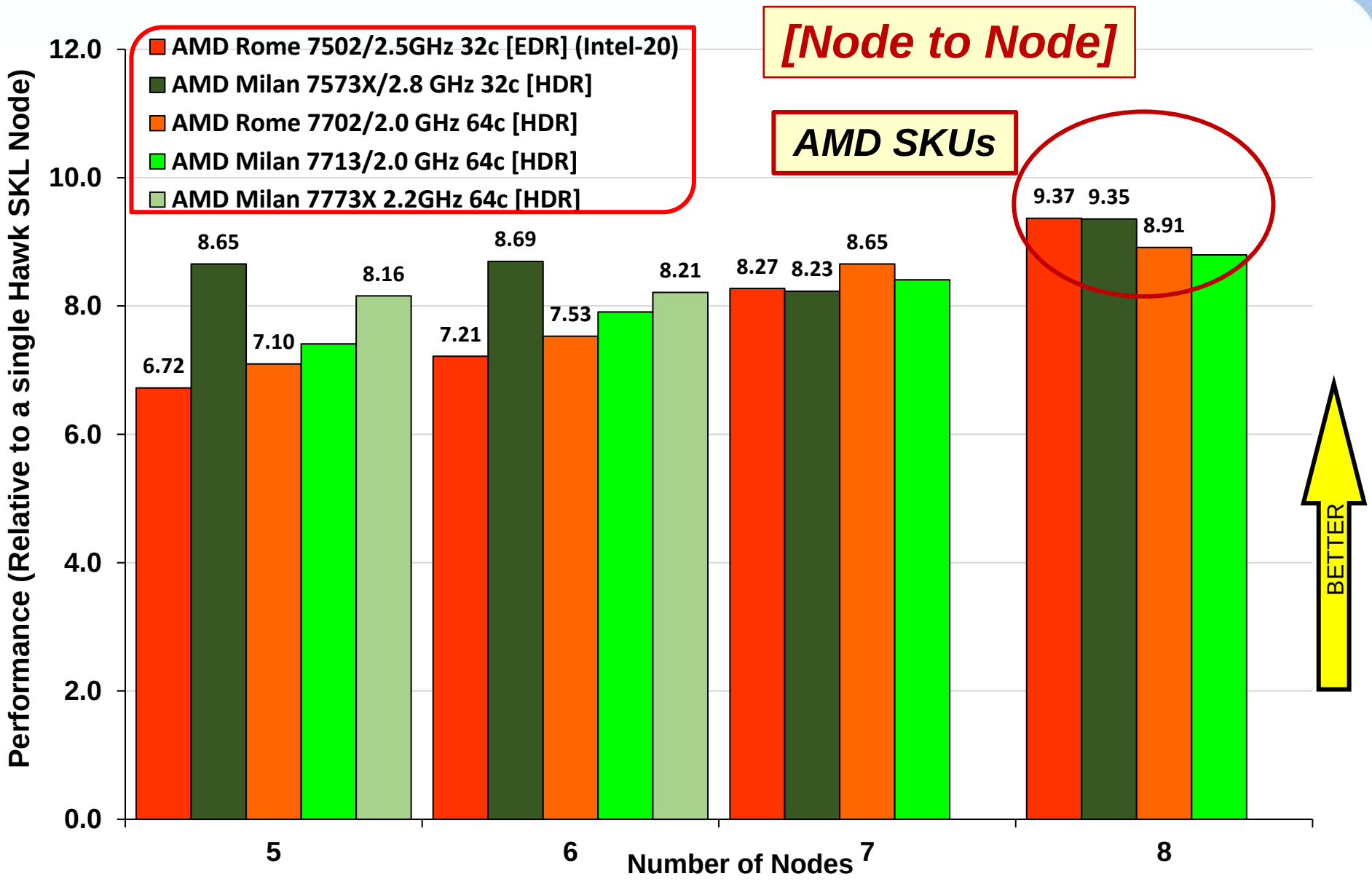
NEMO-FABM-ERSEM (AMM7) – Node Performance



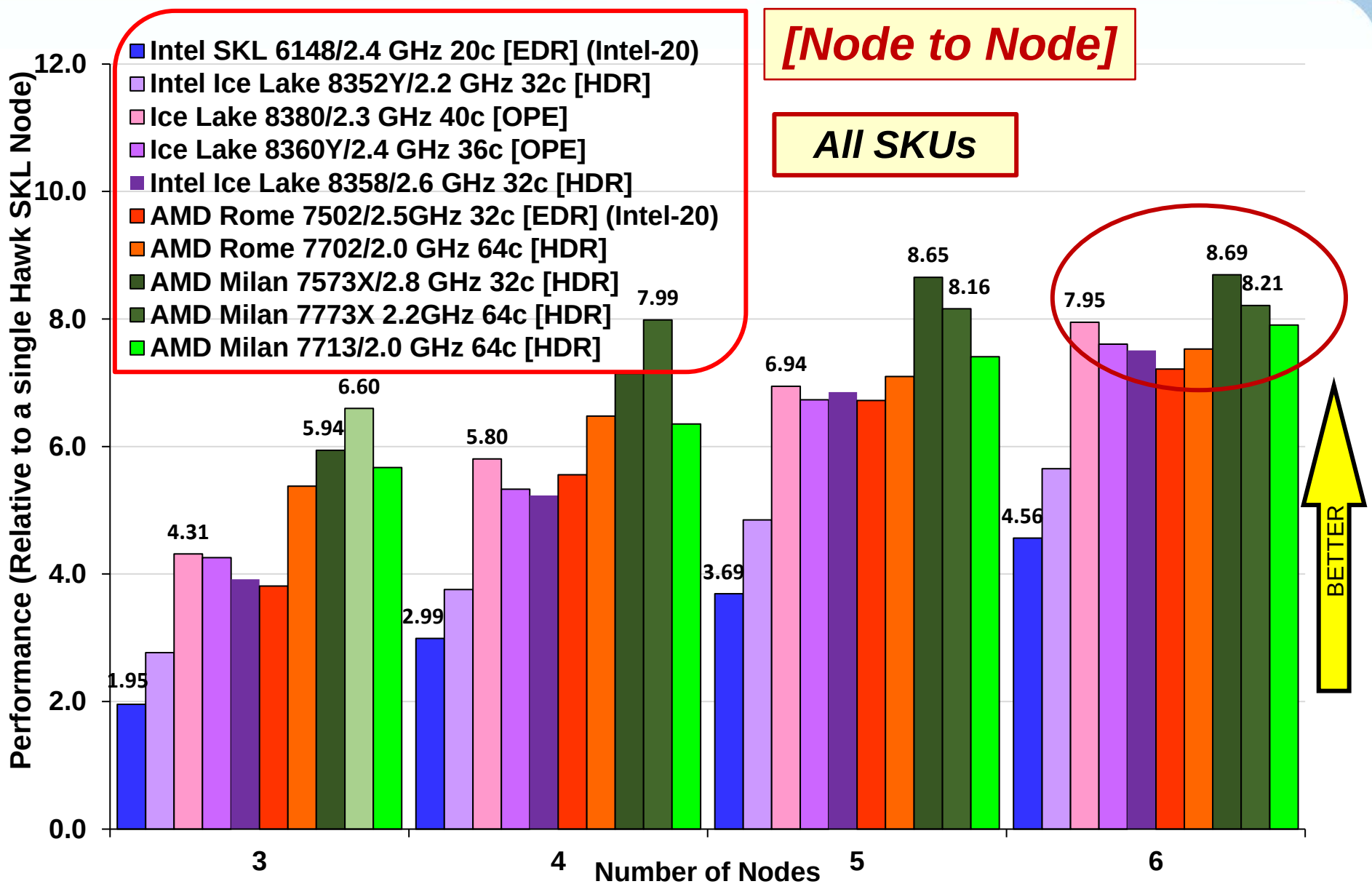
NEMO-FABM-ERSEM (AMM7) – Node Performance



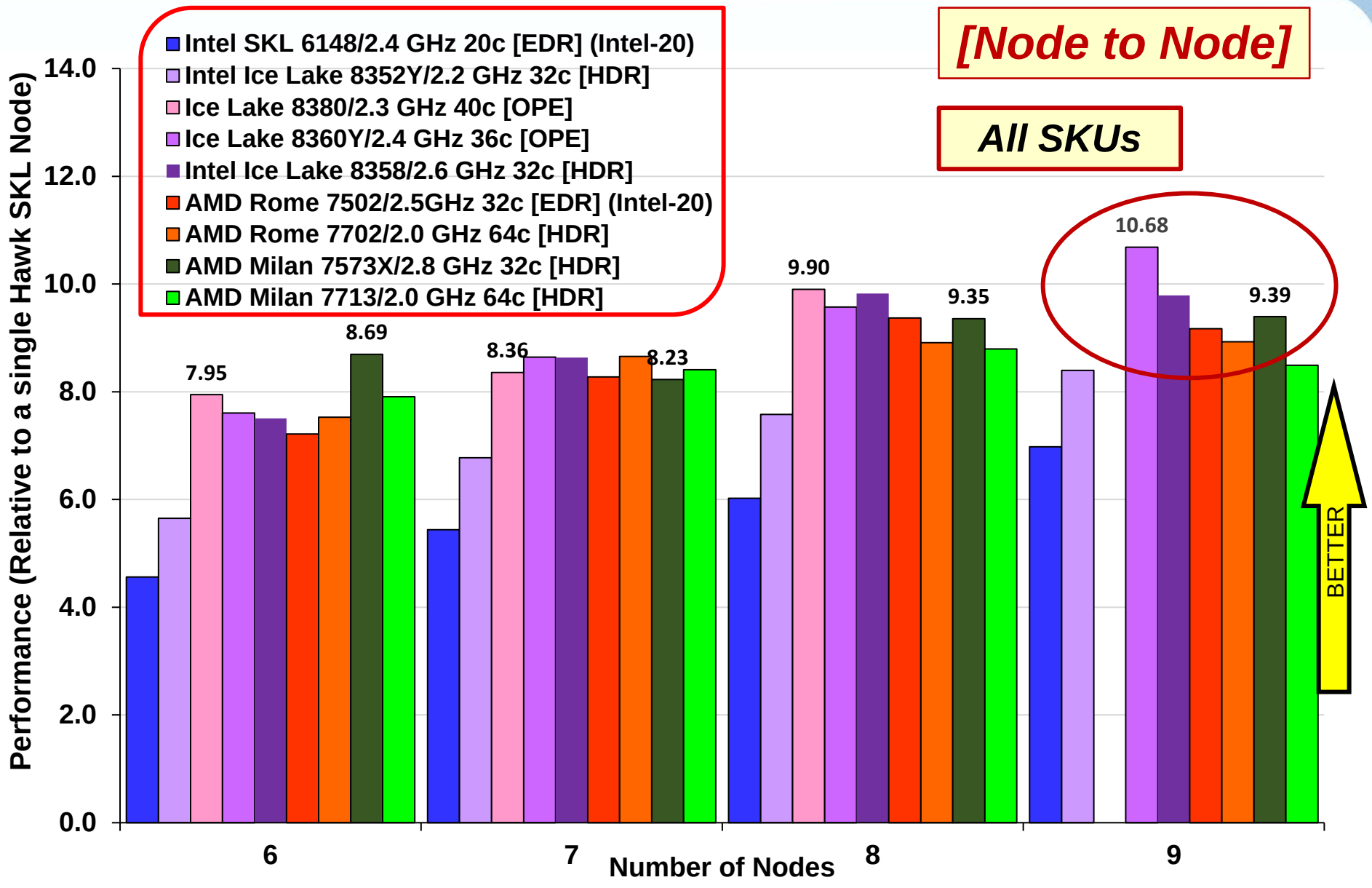
NEMO-FABM-ERSEM (AMM7) – Node Performance



NEMO-FABM-ERSEM (AMM7) – Node Performance

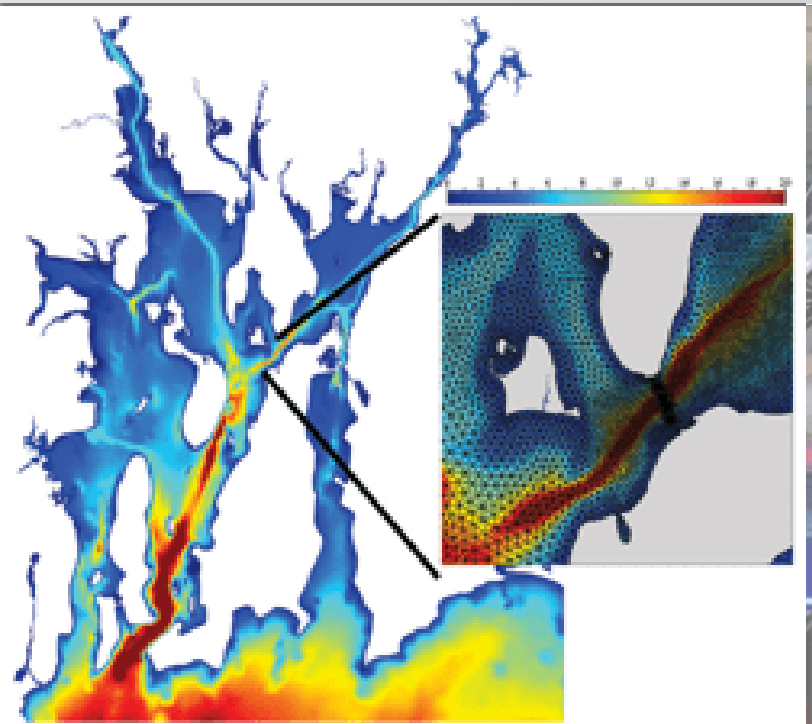


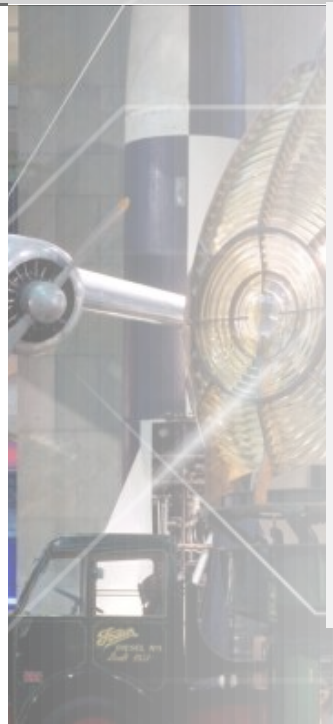
NEMO-FABM-ERSEM (AMM7) – Node Performance



Performance of Ocean Modelling Codes

The Finite Volume Community Ocean Model (FVCOM)





2. FVCOM

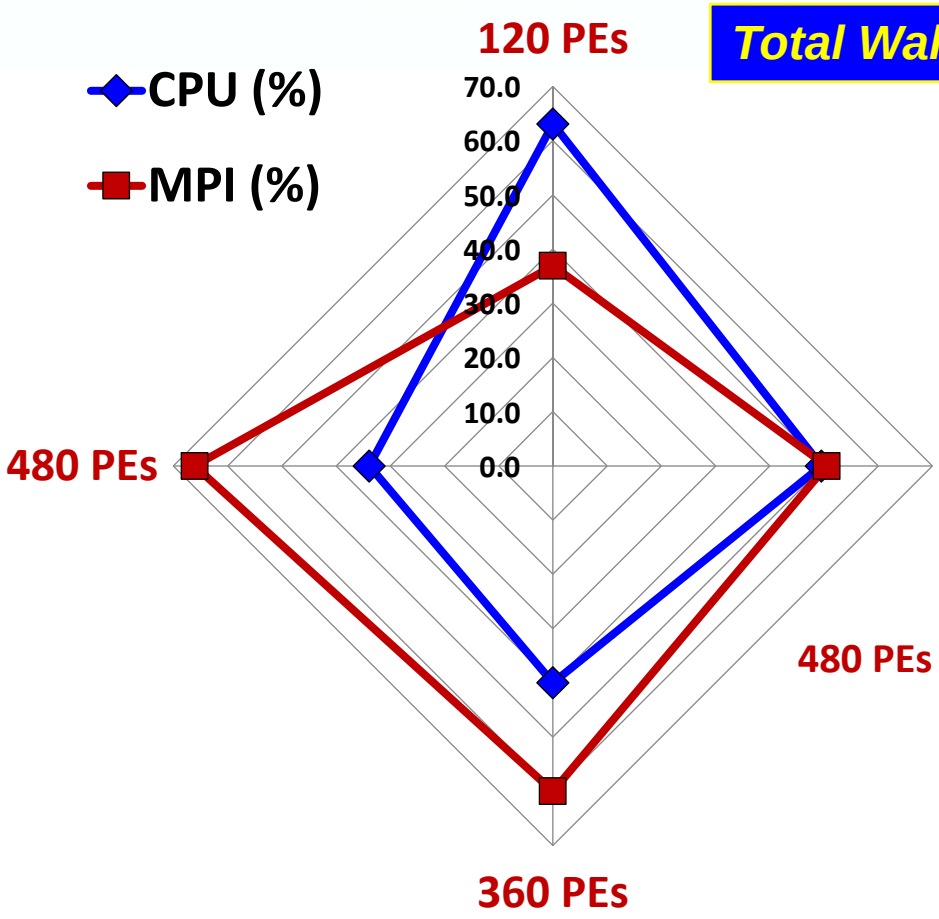
An ocean circulation model for (mainly) studying problems relating to estuarine and coastal environments.

Uses an *unstructured* model grid.

The FVCOM-ERSEM Benchmark

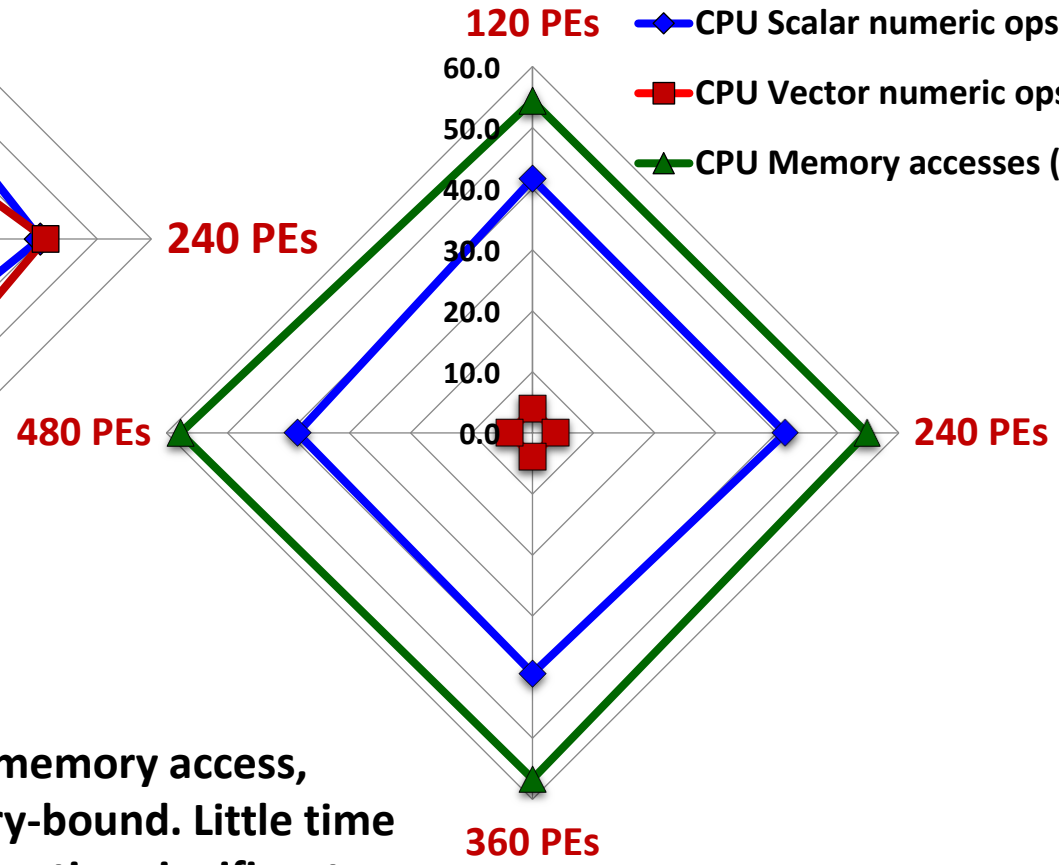
- **FVCOM**, "Finite Volume Community Ocean model" is a prognostic, unstructured-grid, finite-volume, free-surface, 3-D primitive equation coastal ocean circulation model developed by UMASSD-WHOI in the US and based on a triangular mesh. (http://fvcom.smast.umassd.edu/fvcom/)
- **ERSEM**, "European Regional Seas Ecosystem Model" is a bio-geochemical and ecosystem mode, developed at PML (https://github.com/pmlmodelling/ersem)
- Compilation requires **parallel netcdf and hdf5 libraries**.
- *Performance Report* highlights major features of the code. Performance **dominated by memory access**, with the per-core performance memory-bound.
- **Little time spent in vectorized instructions**, suggesting significant opportunities for improving code performance.

FVCOM – Performance Report



Total Wallclock Time Breakdown

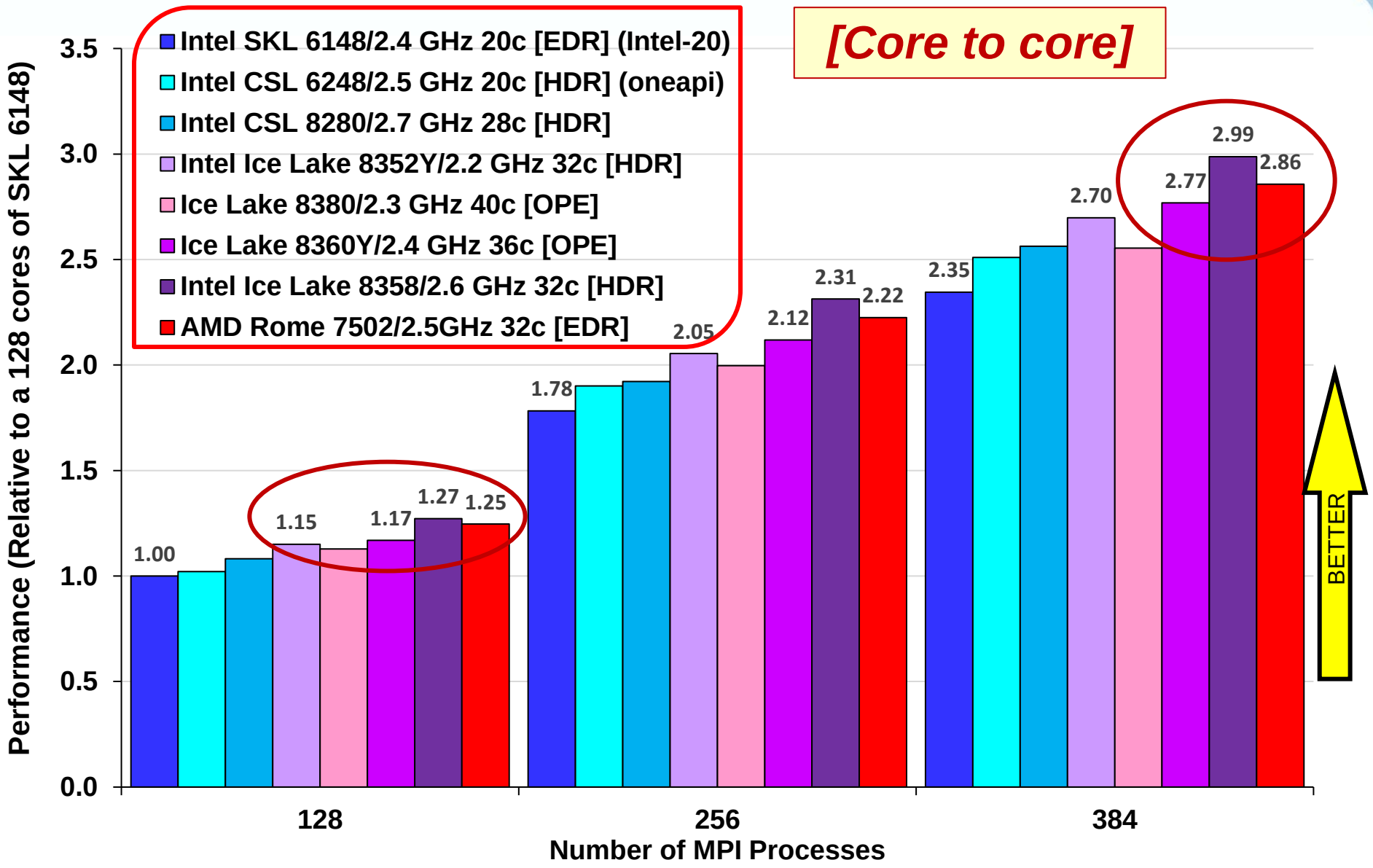
Performance Data (120-480 PEs)



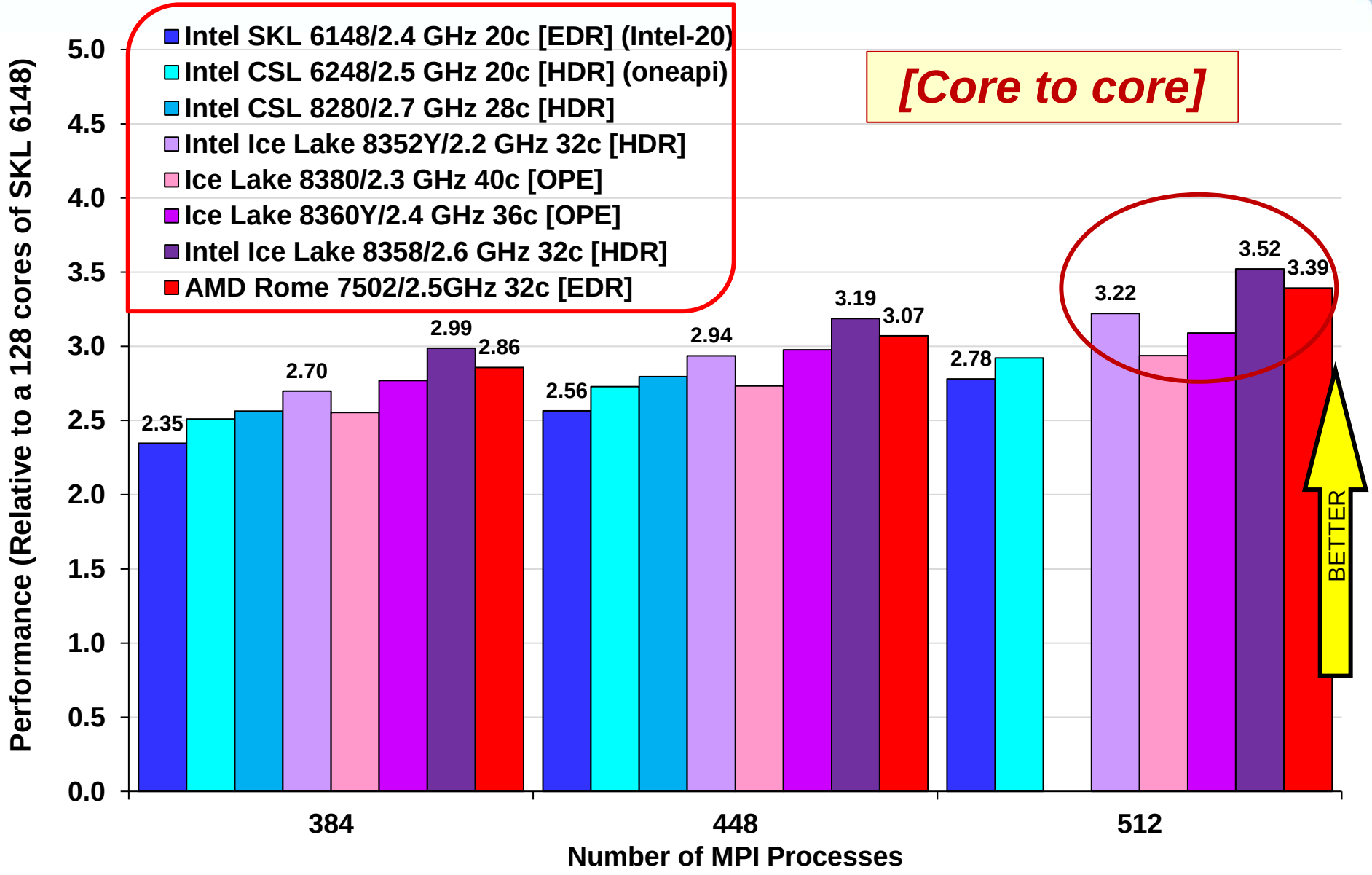
CPU Time Breakdown

FVCOM performance is dominated by memory access, with the per-core performance memory-bound. Little time is spent in vectorized instructions, suggesting significant opportunities for improving code performance.

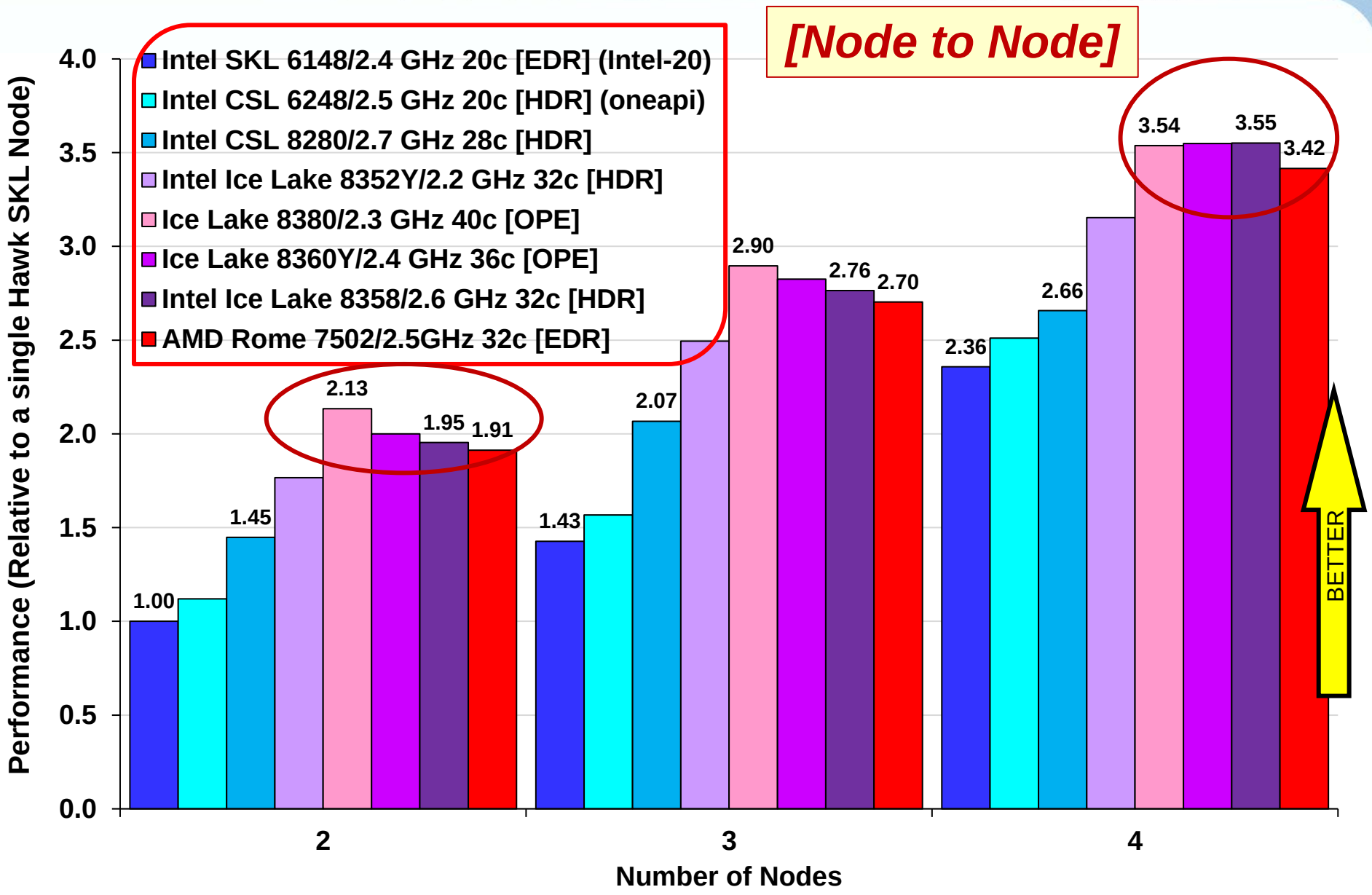
FVCOM – Performance Analysis [Core to Core]



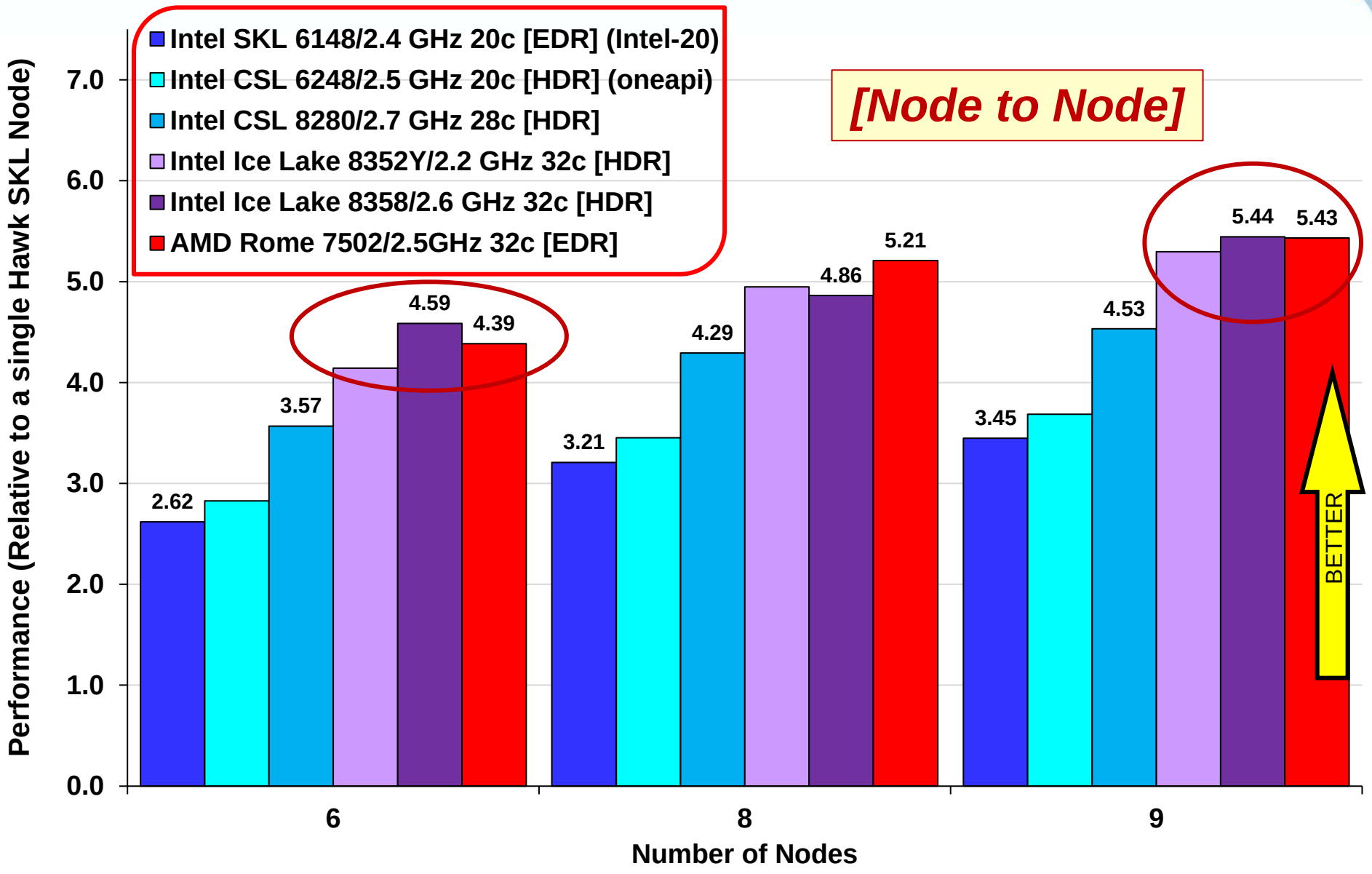
FVCOM – Performance Analysis [Core to Core]



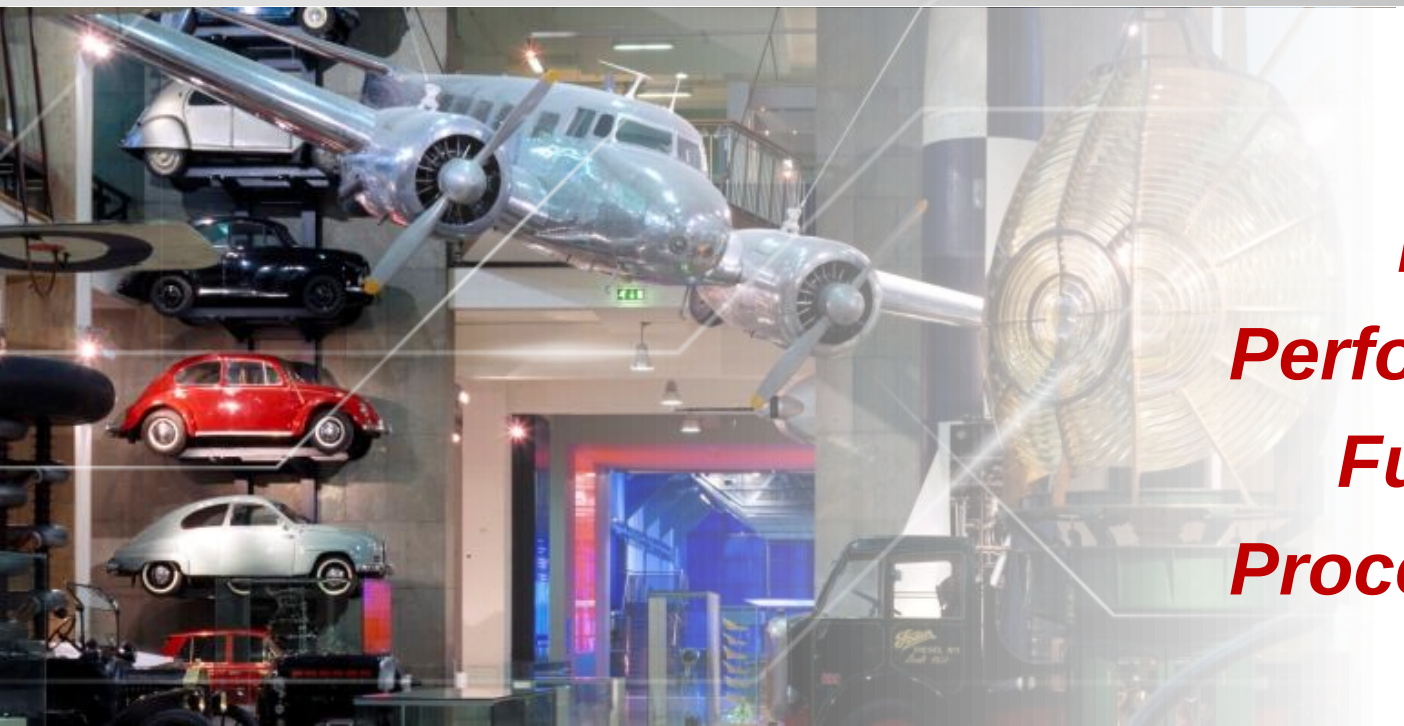
FVCOM – Performance Analysis [Node to Node]



FVCOM – Performance Analysis [Node to Node]

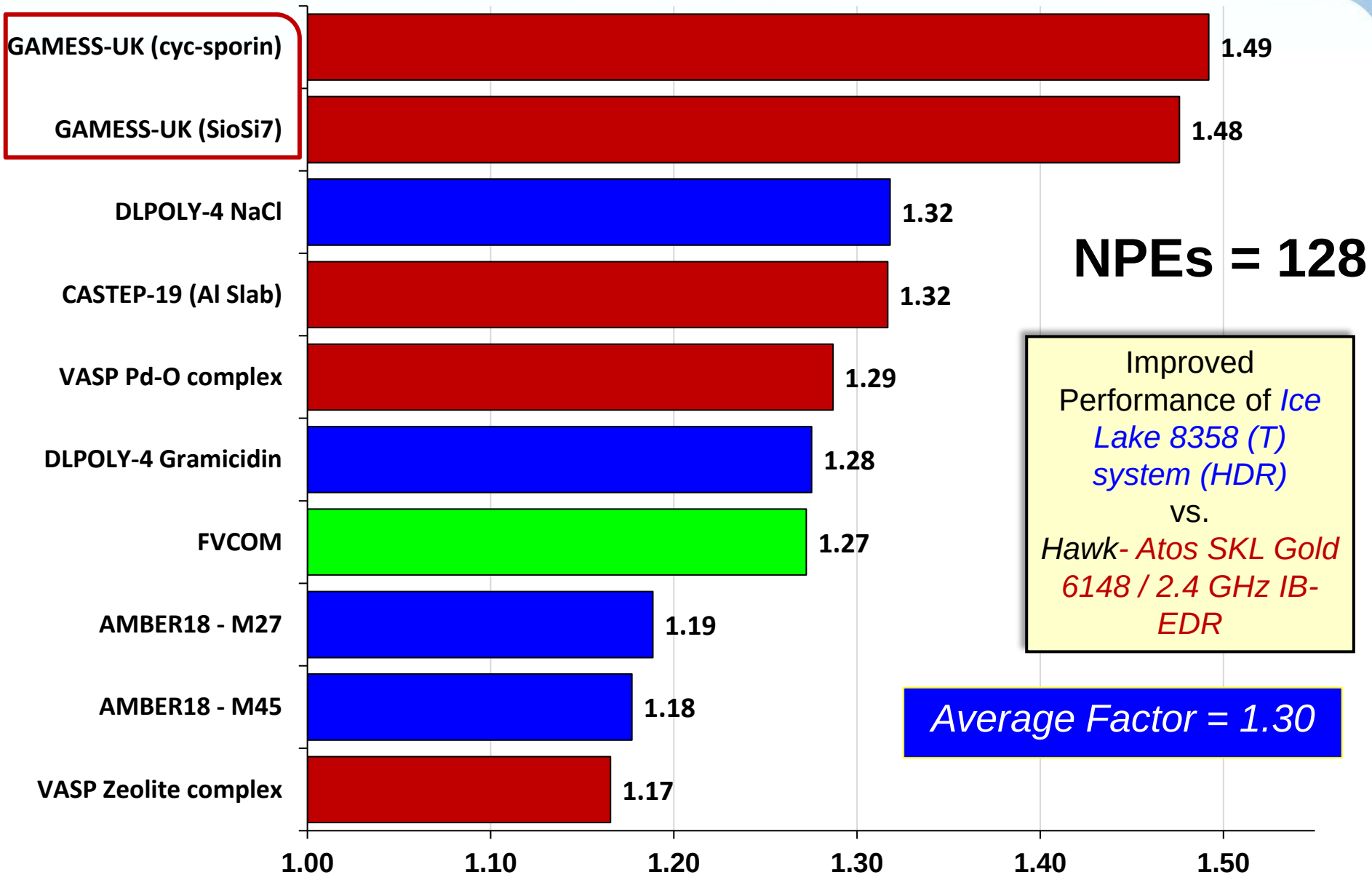


Performance of Computational Chemistry and Ocean Modelling Codes

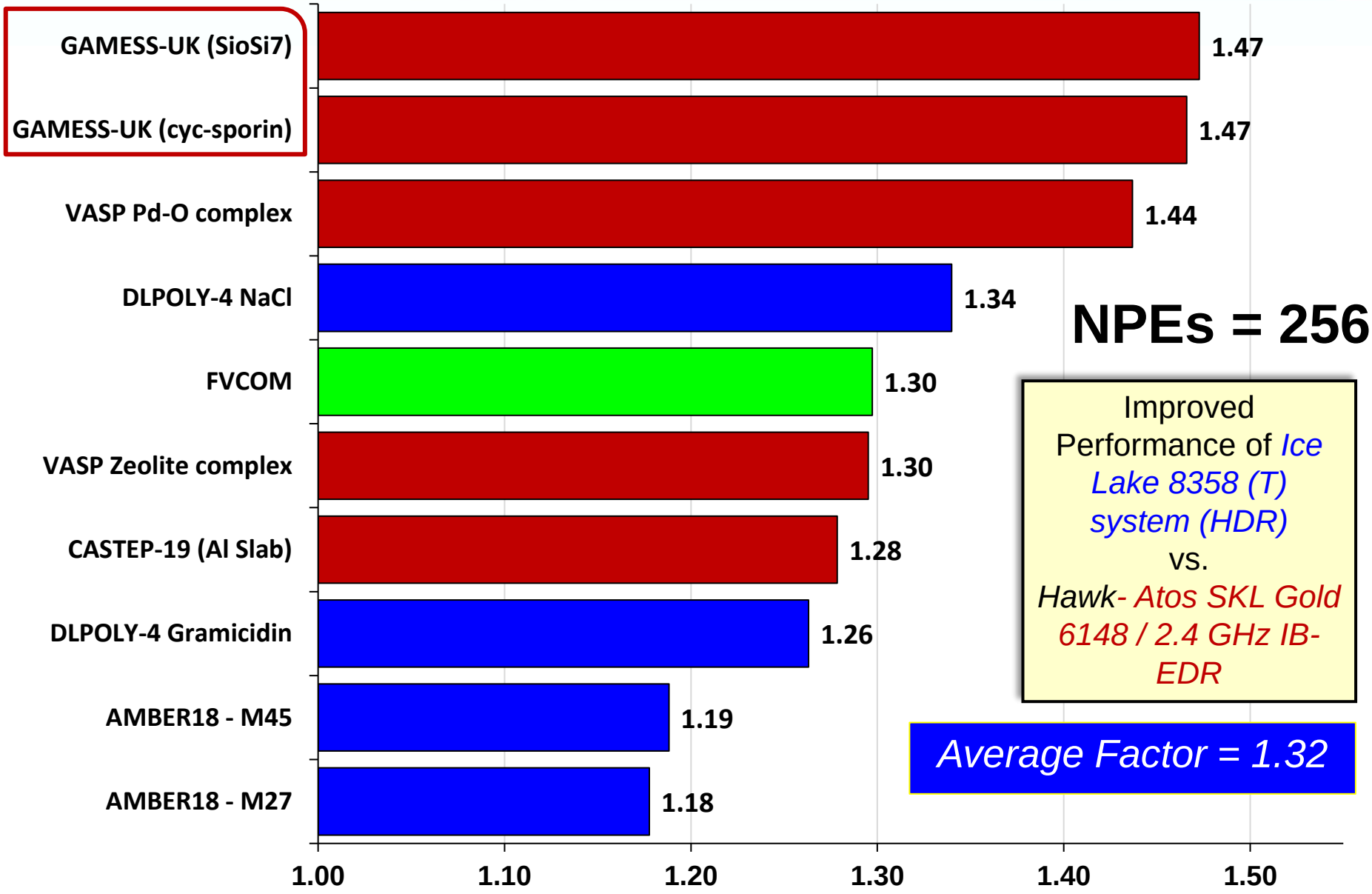


Relative Performance as a Function of Processor Family

Ice Lake 8358 2.6 GHz HDR vs. SKL 6148 2.4 GHz EDR



Ice Lake 8358 2.6 GHz HDR vs. SKL 6148 2.4 GHz EDR

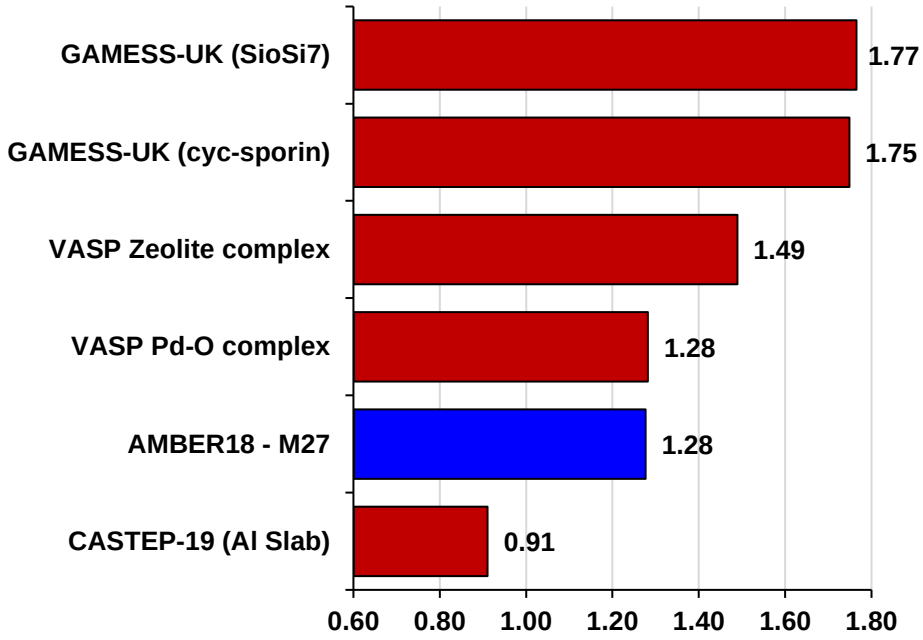


Performance of the AMD Milan 7573X 2.5 GHz HDR

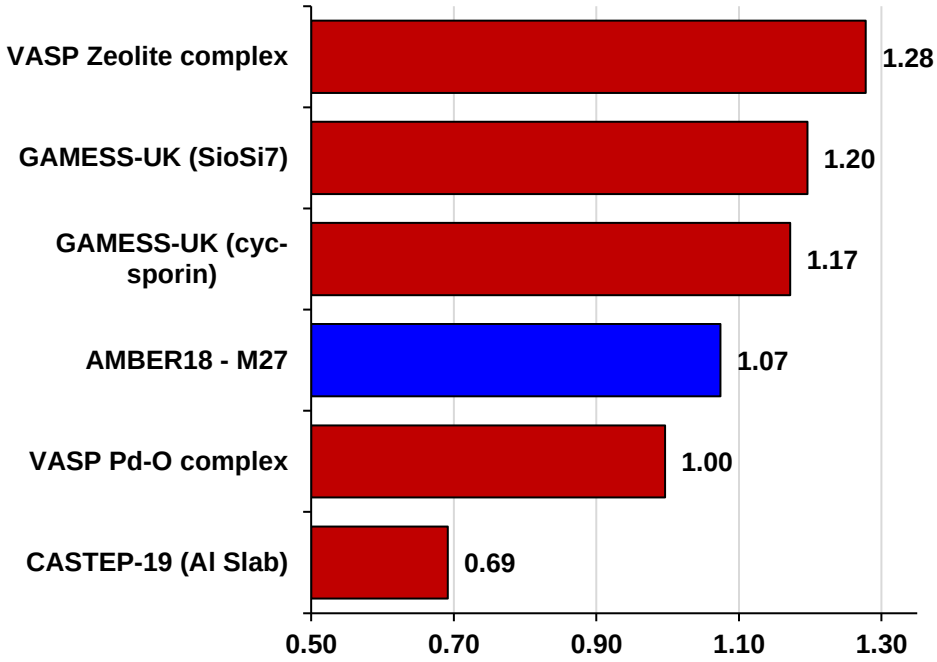
Performance of the AMD Milan 7573X / 2.8 GHz *(HDR)* vs. *Intel SKL Gold 6148 / 2.4 GHz (EDR)*

NPEs = 128

Performance of the AMD Milan 7573X / 2.8 GHz *(HDR)* vs. *Intel Ice Lake 8358 / 2.6 GHz system (HDR)*



Average Factor = 1.41



Average Factor = 1.07

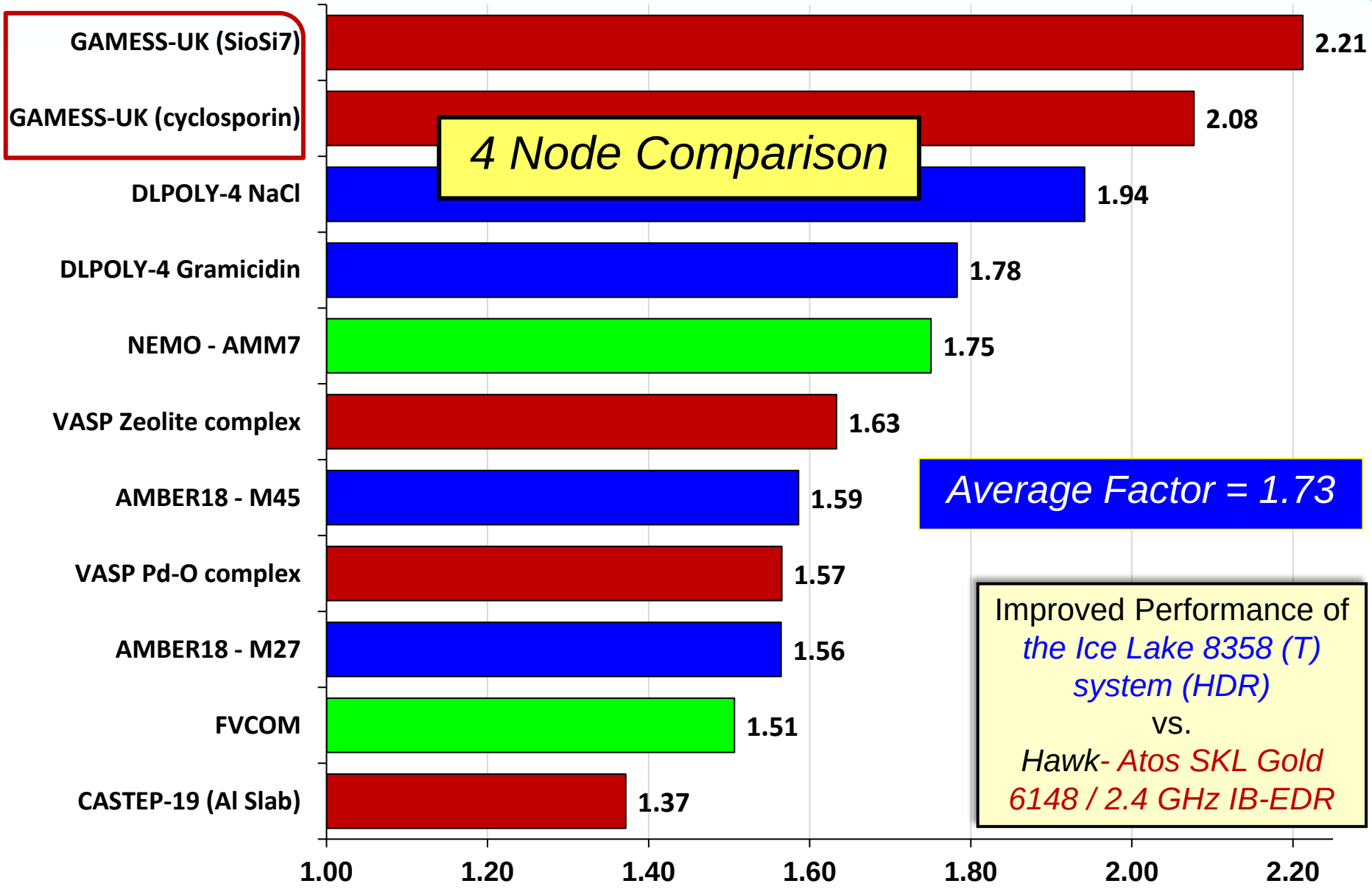
Summary – Core-to-Core Comparisons

- A ***Core-to-Core* comparison** suggests on average that the Intel Ice Lake 8358 2.6 GHz SKU outperforms all other Intel SKUs, although relative performance is sensitive to effective use of the AVX instructions.
- Low utilisation of AVX-512 leads to weaker performance of the SKL, CSL and Ice Lake CPUs and **better performance of the Milan-based clusters** e.g. DLPOLY, GAMESS-UK
- With significant AVX-512 utilisation, Ice Lake Lake systems outperform the AMD Milan systems in core-to-core comparisons e.g. Gromacs, notwithstanding the use of AVX2-256.
- **Exception** is the **AMD Milan 7573X / 2.8 GHz that outperforms the Intel Ice Lake** SKUs in a number of applications.
- With the possible exception of the Intel Ice Lake 8358, there is little to choose between the variety of Intel-based SKUs used in this study, the 36c 8360Y/2.4 GHz, the 38c 8368Q/2.6 GHz & 40c 8380/2.3 GHz.
- Baselined in part across **P100** and **V100** NVIDIA GPU performance.

Summary – Node-to-Node Comparisons

- Given superior core performance, a ***Node-to-Node comparison*** typical of the performance when running a workload shows the Ice Lake 8358 delivering **superior performance** compared to (i) the SKL Gold 6148 (64 cores vs. 40 cores) by a factor of between 1.4 – 2.2 across all applications.
- The **AMD Milan 7713, 7763 and 7773X (128 core nodes)** are the dominant systems given the “high” core counts. e,g,. GROMACS and GAMESS-UK.
- In contrast to the core-to-core comparisons, the higher core count Ice Lake systems – the **38c 8368Q and 40c 8380** – are now performing on a par with the 32c 8358.
- The 32c AMD Milan 7573X is ranked first in four of the 4-node application benchmarks.
- **Pricing** – remains of course a key issue, but lies outside the scope of this presentation.

Ice Lake 8358 2.6 GHz HDR vs. SKL 6148 2.4 GHz EDR



Acknowledgements

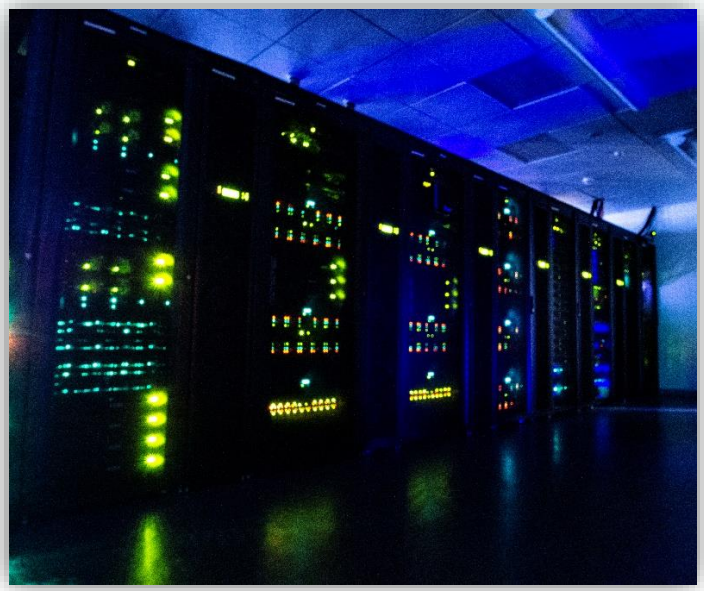
- **Joseph Stanfield and Joshua Weage,**, Dave Coughlin, Derek Rattansey for access to, and assistance with, the variety of AMD EPYC and Ice Lake SKUs at the Dell Benchmarking Centre.
- **Toby Smith, Ian Lloyd and Adam Roe** for access to and assistance with the CXL-AP and Ice Lake clusters at the Swindon Benchmarking Lab
- **Erwin James and John Swinburne** for implementing the NETCDF and XIOS-5 libraries on the Endeavour cluster for testing both the NEMO and FVCOM applications
- ***Okba Hamitou, Luis Cebamanos and Chrisophe Bertherlot*** and access to the SPARTAN and Ice Lake & Milan systems (Genji) at the Atos HPC, AI & QLM Benchmarking Centre
- **Jim Clark, Dale Partridge, Gary Holder and Jerry Blackford** at Plymouth Marine Laboratory for discussions on NEMO & FVCOM performance.

Summary

Focus here on systems featuring **processors from AMD** (EPYC Milan SKUs) and **Intel** (Ice Lake SKUs) with IB and Cornelis Networks interconnects.

- Baseline cluster: the Skylake (SKL) **Gold 6148/2.4 GHz** and **AMD EPYC Rome 7502 2.5Gz** cluster – "Hawk" – at Cardiff University.
- **Five** Intel Xeon Ice Lake clusters, the 32-core Platinum **8358** (2.6 GHz) and **8352Y** (2.2 GHz), the 40-core **8380** (2.3 GHz), 38-core **8368Q** (2.6 GHz), 36-core **8360Y** (2.4GHz) plus other Cascade Lake & Cascade Lake-AP systems.
- **Four** AMD EPYC Milan clusters featuring the 64-core **7713** (2.0 GHz) and **7773X** (2.2 GHz) and the 32-core **7543** (2.8 GHz) and **7573X** (2.8 GHz).
- Consider performance of both synthetic and **end-user applications**. Latter include molecular simulation (**DL_POLY, AMBER**), materials modelling (**CASTEP, VASP**), & electronic structure (**GAMESS-UK**), plus representative ocean modelling codes including **NEMO** and **FVCOM**.
- Scalability analysis by **processing elements (cores)** and by **nodes** (ARM Performance Reports). Baselined against **P100** & **V100** NVIDIA GPUs.

Any Questions?



Martyn Guest *029-208-79319*

Jose Munoz *029-208-70626*

Thomas Green *029-208-79269*