COMPUTING INSIGHT UK 2023

"Productive Supercomputing"

7 - 8 DECEMBER 2023

Manchester Central, UK

www.ukri.org/CIUK

Performance of Community Codes on Multi-core Processors

An Analysis of Computational Chemistry and Ocean Modelling Applications

Martyn Guest, Jose Munoz Criollo & 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. Regular feature at Daresbury's MEW and successor CIUK conferences.
- Focus on systems featuring **processors from Intel** (Sapphire Rapids & Ice Lake SKUs) and **AMD** (EPYC Genoa & Milan SKUs) with Infiniband (EDR, HDR, NDR) & Cornelis Networks interconnects.
  - Baseline clusters: Skylake (SKL) **Gold 6148/2.4 GHz** and **AMD EPYC Rome 7502 2.5Gz** cluster – "Hawk" – at Cardiff University.
  - **Two** Intel Sapphire Rapids clusters – the 56-core Platinum 8480 (2.0 GHz) and Platinum HBM 9480 (1.9 GHz).
  - **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.

Introduction and Overview

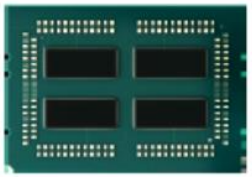
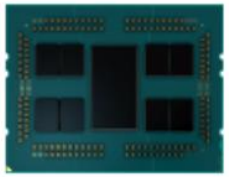
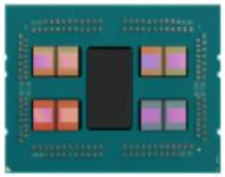
- **Four** AMD EPYC **Milan clusters** featuring the 64-core **7713** (2.0 GHz) and **7773X** (2.2 GHz) and the 32-core **7543** (2.8 GHz) and **7573X** (2.8 GHz).
- **Two** AMD **Genoa clusters** featuring the 32-core **9354** (3.25 GHz) and 48-core **9454** (2.85 GHz) SKUs.

- Consider performance of both synthetic and **end-user applications:**
  - Molecular simulation (**DL_POLY, LAMMPS, AMBER & GROMACS MD codes**);
  - Materials modelling (**VASP, CASTEP**) & electronic structure (**GAMESS-UK**);
  - Ocean modelling codes including **NEMO** and **FVCOM**.
- Scalability analysis by **processing elements (cores)** and by **nodes** (ARM Performance Reports). Baselined against **V100** NVIDIA GPUs.
- **Pricing** – remains of course a key issue but lies outside the scope of this presentation.

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. **Performance comparisons** across a spectrum of MPI versions with Intel Parallel Studio XE e.g. 2018/4, 2019/5, 2019/12 & 2020/4 PLUS OneAPI proved **challenging**.
   - Problems encountered on **AMD Milan** systems. Working code with Intel 2019/5 on AMD Rome systems failed on Milan, with codes hanging at arbitrary core counts. **Intel oneapi resolved many of these** issues.
   - **Performance issues remain** compared to earlier variants of Intel Parallel Studio XE. e.g., a major decline in both VASP and CASTEP performance on AMD EPYC when moving from “mpi/intel/2018/2” to “mpi/intel/2020/2”
5. Consistency through use of **SPACK Package Manager for HPC** demonstrated throughout this analysis.

AMD “GENOA” EPYC SERVER CPUS

	AMD EPYC 7001 'NAPLES'	AMD EPYC 7002 'ROME'	AMD EPYC 7003 'MILAN'	AMD EPYC 9004, 8004 'GENOA', 'SIENA'
				
Core Architecture	'Zen'	'Zen 2'	'Zen 3'	'Zen 4' and 'Zen 4c'
Cores	8 to 32	8 to 64	8 to 64	8 to 128
IPC Improvement Over Prior Generation	N/A	~24% ROM-236	~19% MLN-003	~14% EPYC-038
Max L3 Cache	Up to 64 MB	Up to 256 MB	Up to 256 MB	Up to 384 MB (EPYC 9004) Up to 128 MB (EPYC 8004)
Max L3 Cache with 3D V-Cache™ technology			768 MB	Up to 1152 MB
PCIe® Lanes	Up to 128 Gen 3	Up to 128 Gen 3	Up to 128 Gen 4	Up to 128 Gen 5 8 bonus lanes Gen 3
CPU Process Technology	14nm	7nm	7nm	5nm
I/O Die Process Technology	N/A	14nm	14nm	6nm
Power (Configurable TDP [cTDP])	120-200W	120-280W	155-280W	70-400W
Max Memory Capacity	2 TB DDR3-2400/2666	4 TB DDR4-3200	4 TB DDR4-3200	6 TB DDR5-4800

AMD EPYC Genoa: IPC Improvements

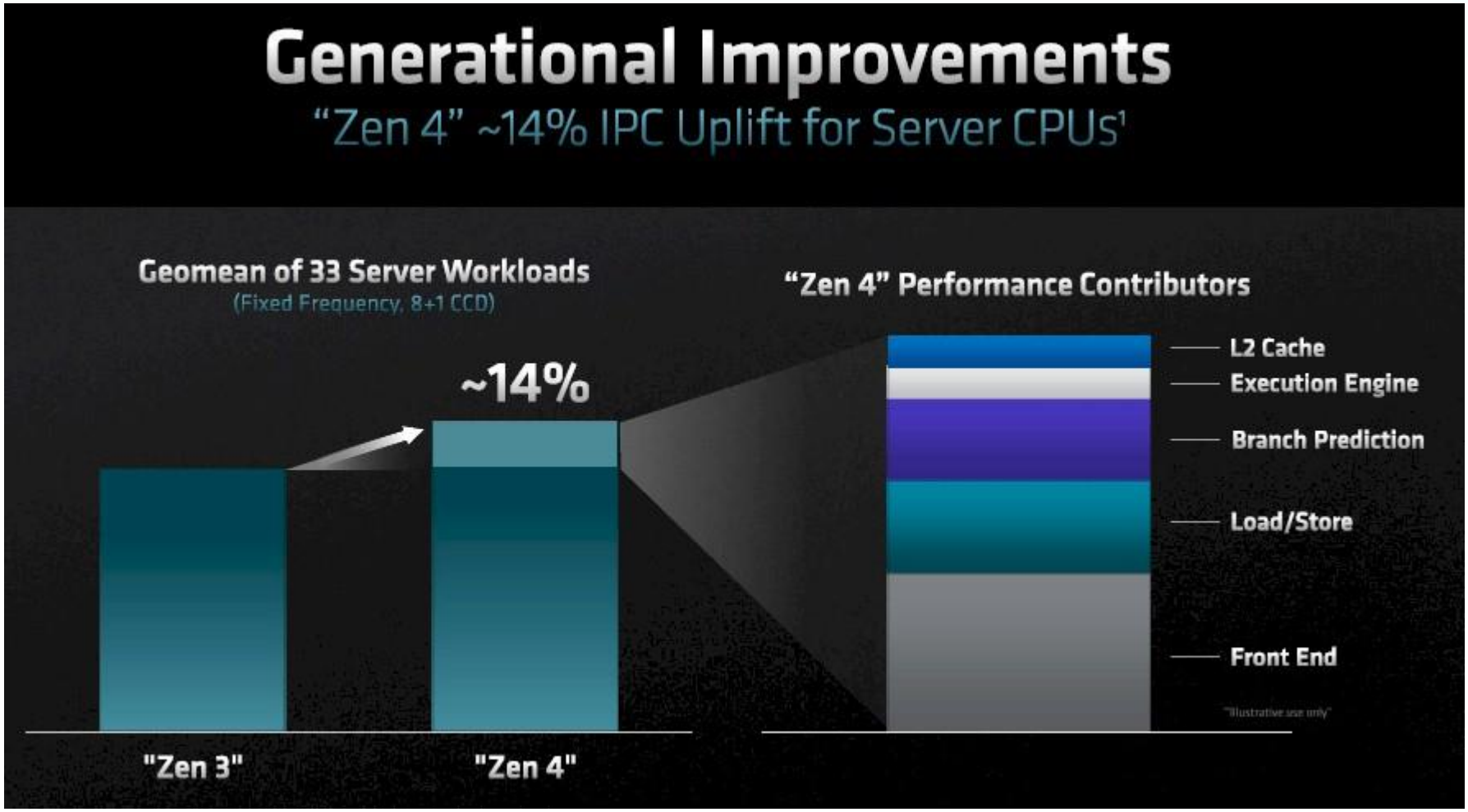
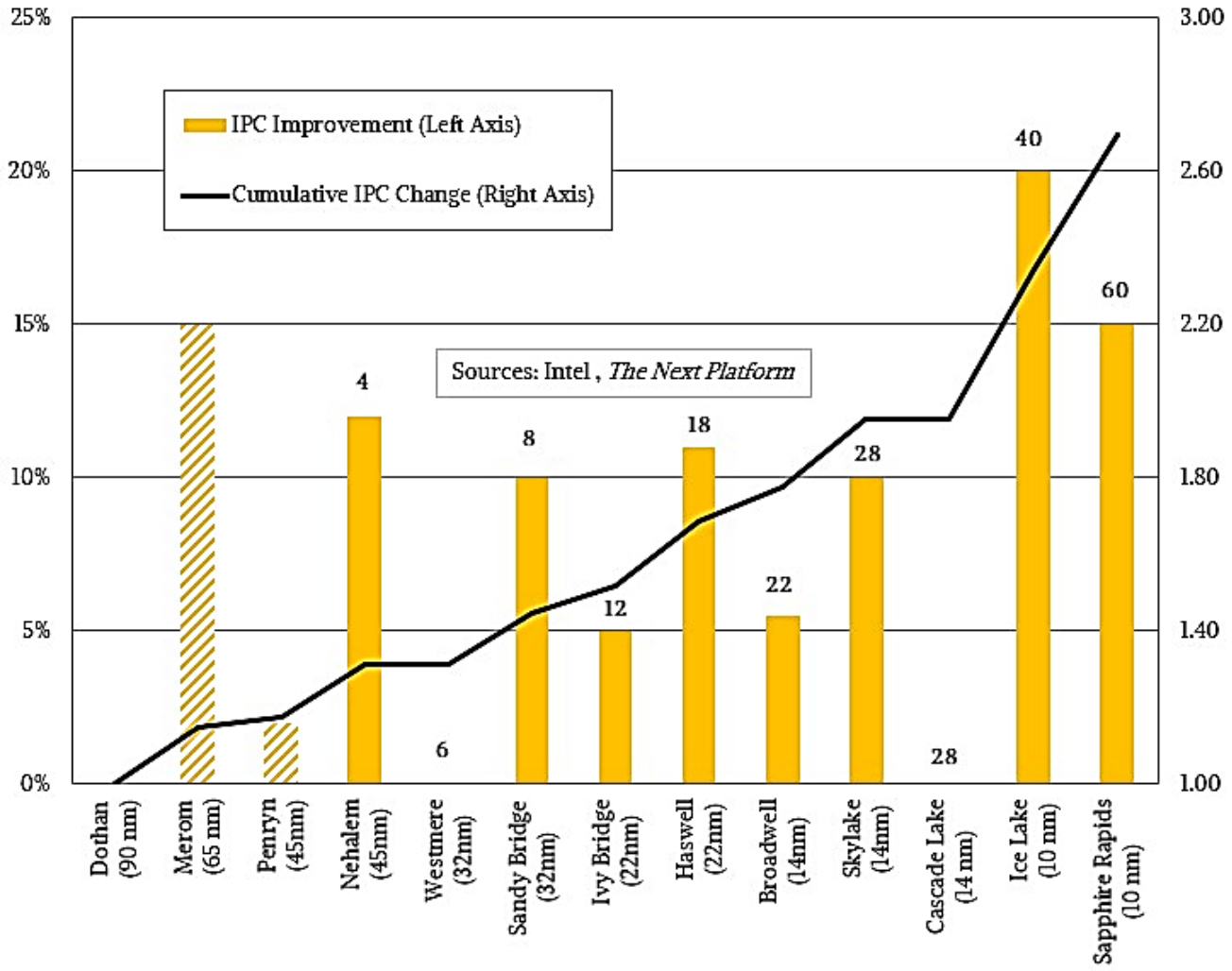


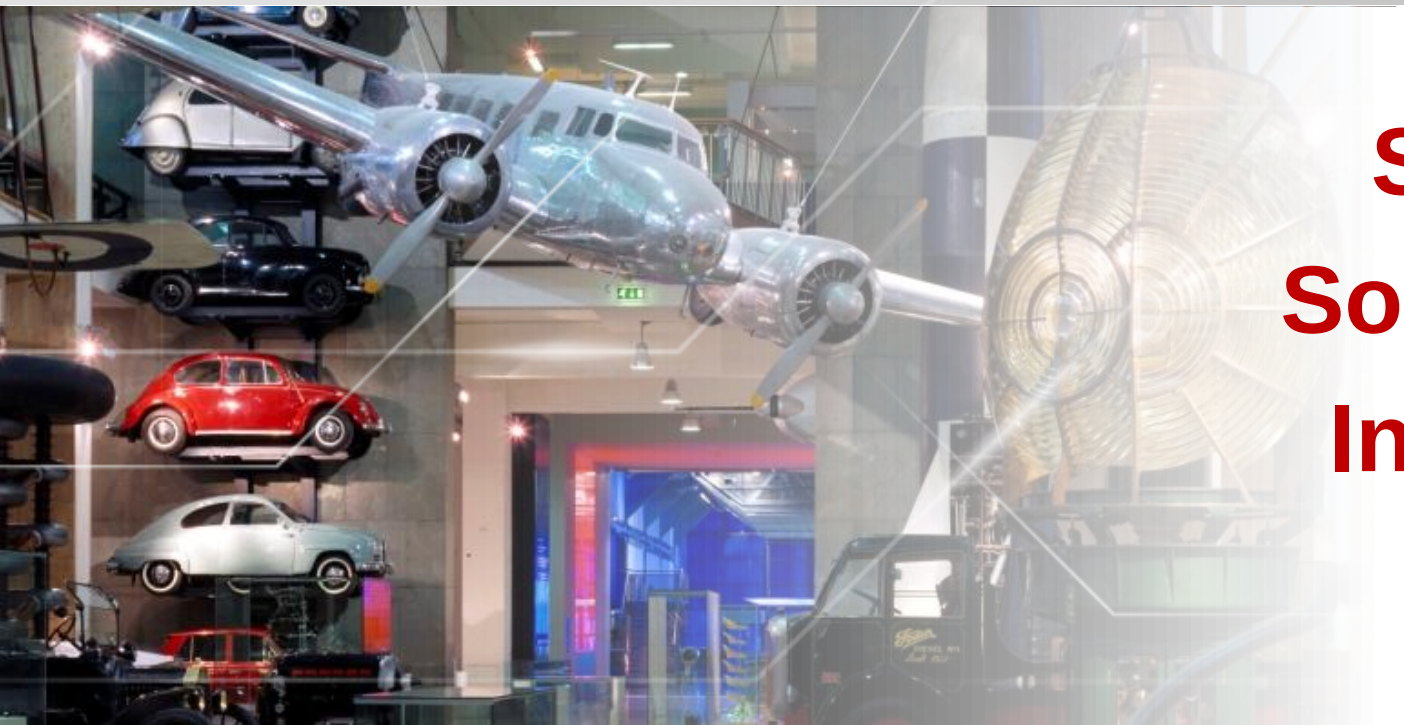
Figure. The move to Genoa is a big leap in performance, starting with the move to the "Zen 4" cores, which are providing a 14 percent increasing in the instructions per clock (IPC) compared to the prior "Zen3" cores used in the Milan Epyc 7003s..

IPC Improvements - Intel Core Generations



Instructions per clock (IPC) improvement per generation versus cumulative IPC over time. Maximum core count per generation shown above the bars for each Xeon chip.

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 Xeon 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 • 5,288 cores – Intel CSL and AMD Milan SKUs • 2,064 cores – Intel Broadwell and Haswell Raven migrated sub-system nodes (no decommissioned)

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

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

Intel Xeon Sapphire Rapids Clusters

Cluster / Configuration
Dell Zenith cluster at the Dell Technologies HPC & AI Innovation Lab – Intel Xeon sub-systems with Mellanox HDR and NDR interconnect fabrics running Slurm
• 50 nodes × Intel **Xeon Platinum 8480 Processor / 2.00 GHz**; # of CPU Cores: **56**; # of Threads: 112; Max Turbo Frequency: 3.80 GHz Base Clock: **2.00 GHz;** Cache **105 MB;** Default TDP / TDP: 350W; **DDR5 4800 MT/s**; **Mellanox NDR 400Gb/s** • The 8480 systems are connected to NDR InfiniBand, configured in a fat tree, with each rack of nodes generally using a single edge switch.
Intel's Endeavour cluster with Mellanox HDR and Cornelis OPE interconnect fabrics running Slurm
• 150 nodes × Intel **Xeon Platinum 8480 Processor / 2.00 GHz**; # of CPU Cores: **56**; # of Threads: 112; Max Turbo Frequency: 3.80 GHz Base Clock: **2.00 GHz;** Cache **105 MB;** Default TDP / TDP: 350W; **DDR5 4800 MT/s; Mellanox HDR 200Gb/s; Cornelis OPE**
• 73 nodes × Intel **Xeon Platinum 9480 Processor / 1.90 GHz**; # of CPU Cores: **56**; # of Threads: 112; Max Turbo Frequency: 3.50 GHz Base Clock: **1.90 GHz;** Cache **112.5 MB;** Default TDP / TDP: 350W; **DDR5 4800 MT/s; [Maximum High Bandwidth Memory (HBM): 64 GB]; Mellanox HDR 200Gb/s; Cornelis OPE**

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**

AMD EPYC Genoa Clusters

Cluster / Configuration
Dell Minerva cluster at the Dell Technologies HPC & AI Innovation Lab – AMD Genoa sub-system with Mellanox NDR interconnect fabric running Slurm
• **22 nodes × AMD EPYC Genoa 9354 / 3.25 GHz**; # of CPU Cores: **32**; # of Threads: 64; Max Turbo Frequency: 3.8 GHz Base Clock: **3.25 GHz;** L3 Cache 256 MB**;** Default TDP / TDP: 280W; **Mellanox NDR 400Gb/s** • The 9354 systems are connected to NDR InfiniBand configured on a single switch.
AMD Genoa cluster at Nottingham University with Mellanox NDR interconnect fabric running Slurm.
• **AMD EPYC Genoa 9454 / 2.75 GHz Processor**; # of CPU Cores: **48**; # of Threads: 96; Max Turbo Frequency: 3.80 GHz Base Clock: **2.75 GHz;** L3 Cache 256 MB**;** Default TDP / TDP: 290W; **Mellanox NDR 400Gb/s.**
• **63 'standard' compute nodes**, 384 GB RAM, 1x NDR200 Dual Port IB HCA: **10 'high mem' compute nodes, 1536 GB RAM**, 1x NDR200 Dual Port IB HCA; 4 'GPU' compute nodes, 2x AMD 9454 48C 2.75GHz CPUs, 768 GB RAM, 8x NVIDIA A100 80GB PCIe Gen4 Passive GPU, 1x NDR200 Dual Port IB HCA. Spectrum Scale (GPFS). SLURM 23.02.4.

NVIDIA HPC-X

NVIDIA HPC-X: Increased use of NVIDIA HPC-X that includes **MPI, SHMEM and PGAS communications libraries**, and various acceleration packages.

- **Key Features**
  - Offloads collectives communications from MPI onto NVIDIA InfiniBand networking hardware
  - Multiple transport support, including Reliable Connection (RC), Dynamic Connected (DC), and Unreliable Datagram (UD)
  - Intra-node shared memory communication
  - Native support for MPI-3
  - Multi-rail support with message striping
  - NVIDIA GPUDirect with CUDA support
  - NCCL-RDMA-SHARP plug-in support
- Experience suggests that this toolkit enables MPI & SHMEM/PGAS programming languages to achieve **higher performance, scalability, and efficiency.**
- Notable performance impact in both **CASTEP and VASP**. ***(Rev 2.16)***

Using the Spack package manager

- Like EasyBuild (1), Spack (2) Spack is a multi-platform package manager that builds and installs multiple versions and configurations of software. **Spack** resolves dependencies and installs them like any other package manager you can find on a linux platform.

- The definition provided by the official documentation is as follows:

  "Spack is a multi-platform package manager that builds and installs multiple versions and configurations of software. It works on Linux, macOS, and many supercomputers. Spack is non-destructive: installing a new version of a package does not break existing installations, so many configurations of the same package can coexist"

- *Spack offers a simple "spec" syntax that allows users to specify versions and configuration options. Package files are written in pure Python, and specs allow package authors to write a single script for many different builds of the same package. With Spack, you can build your software as you wish".*

[1] *https://docs.easybuild.io/installation/*

[2] *https://spack.readthedocs.io/en/latest/index.html#*

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, LAMMPS, AMBER & GROMACS (MD)

VASP and CASTEP (ab initio Materials properties)

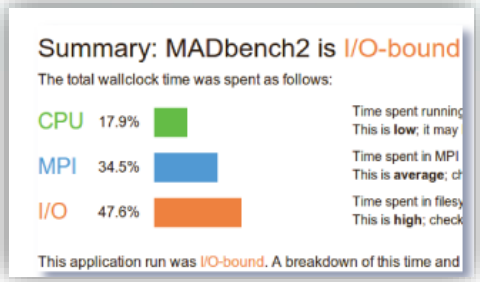
GAMESS-UK (molecular electronic structure)

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|Linaro 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.

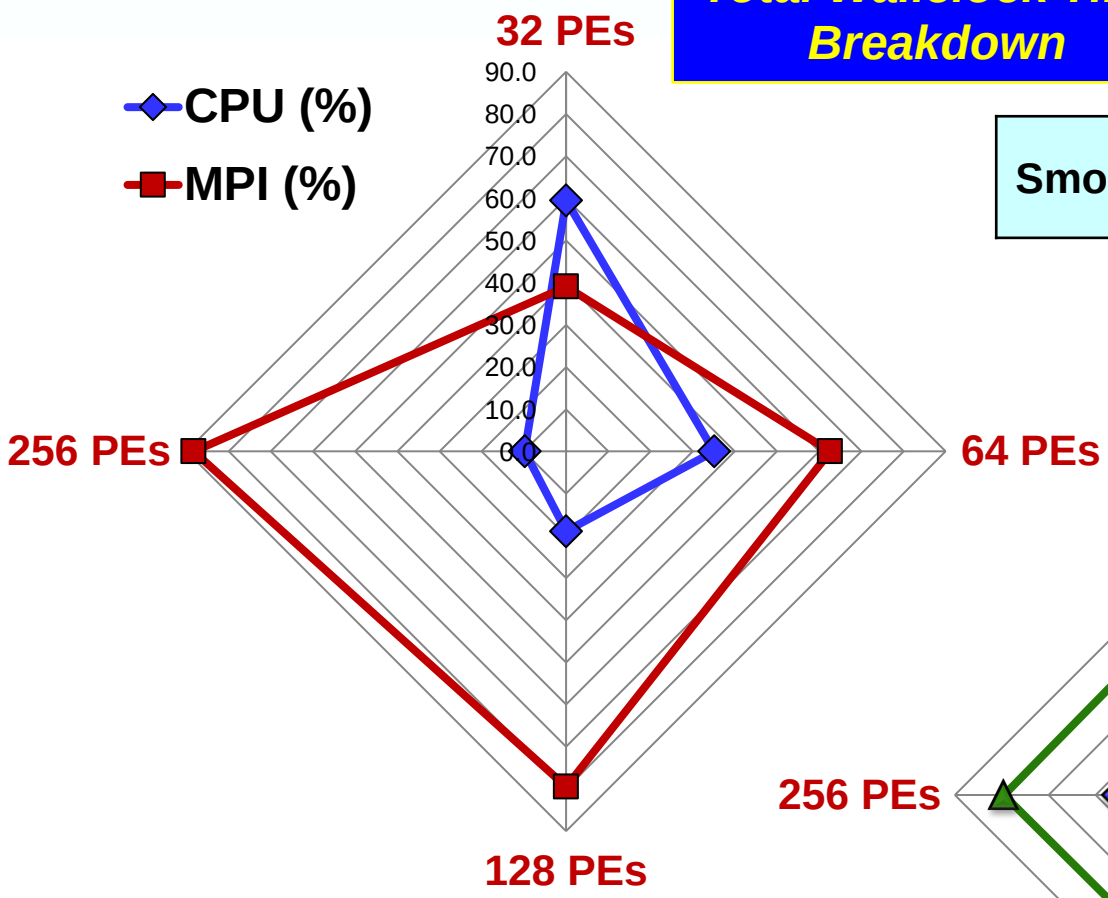
perf-report mpirun $code

- ***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)**

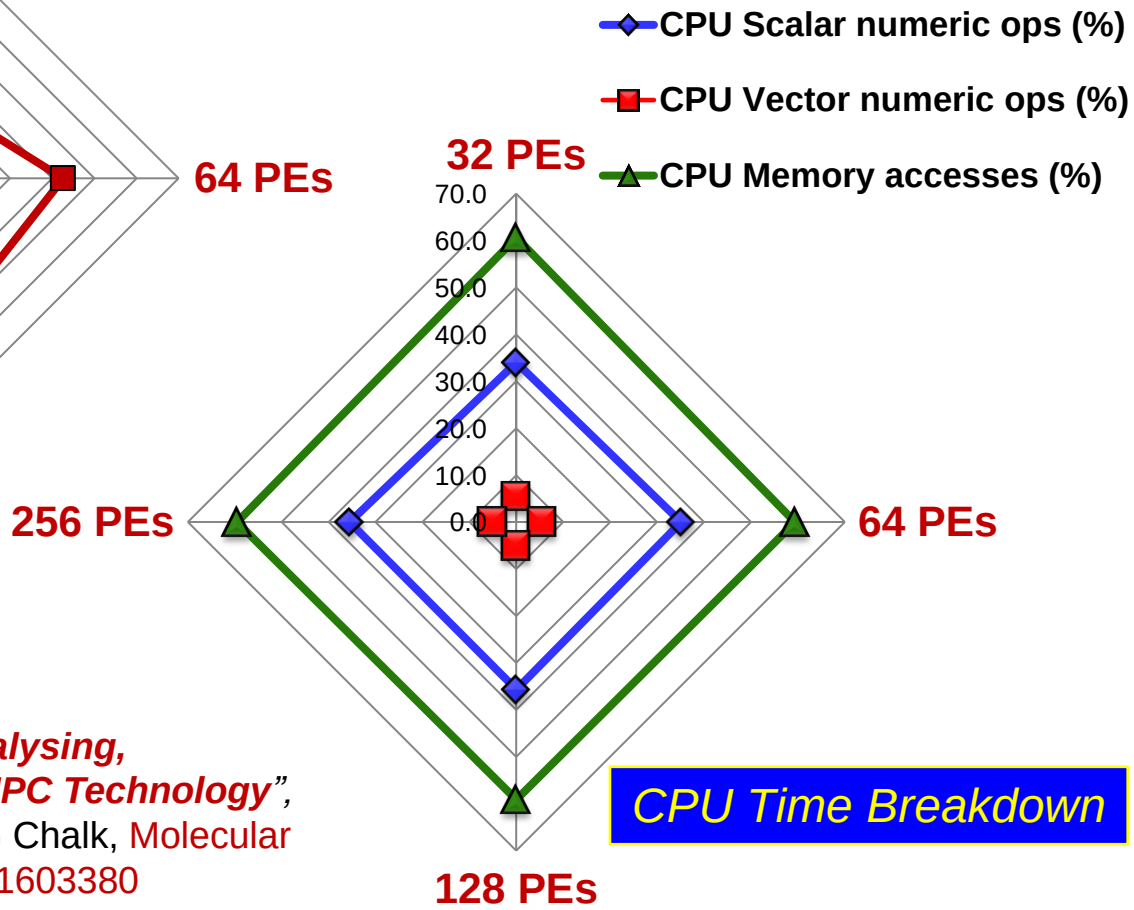
DLPOLY4 – Performance Report

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
```

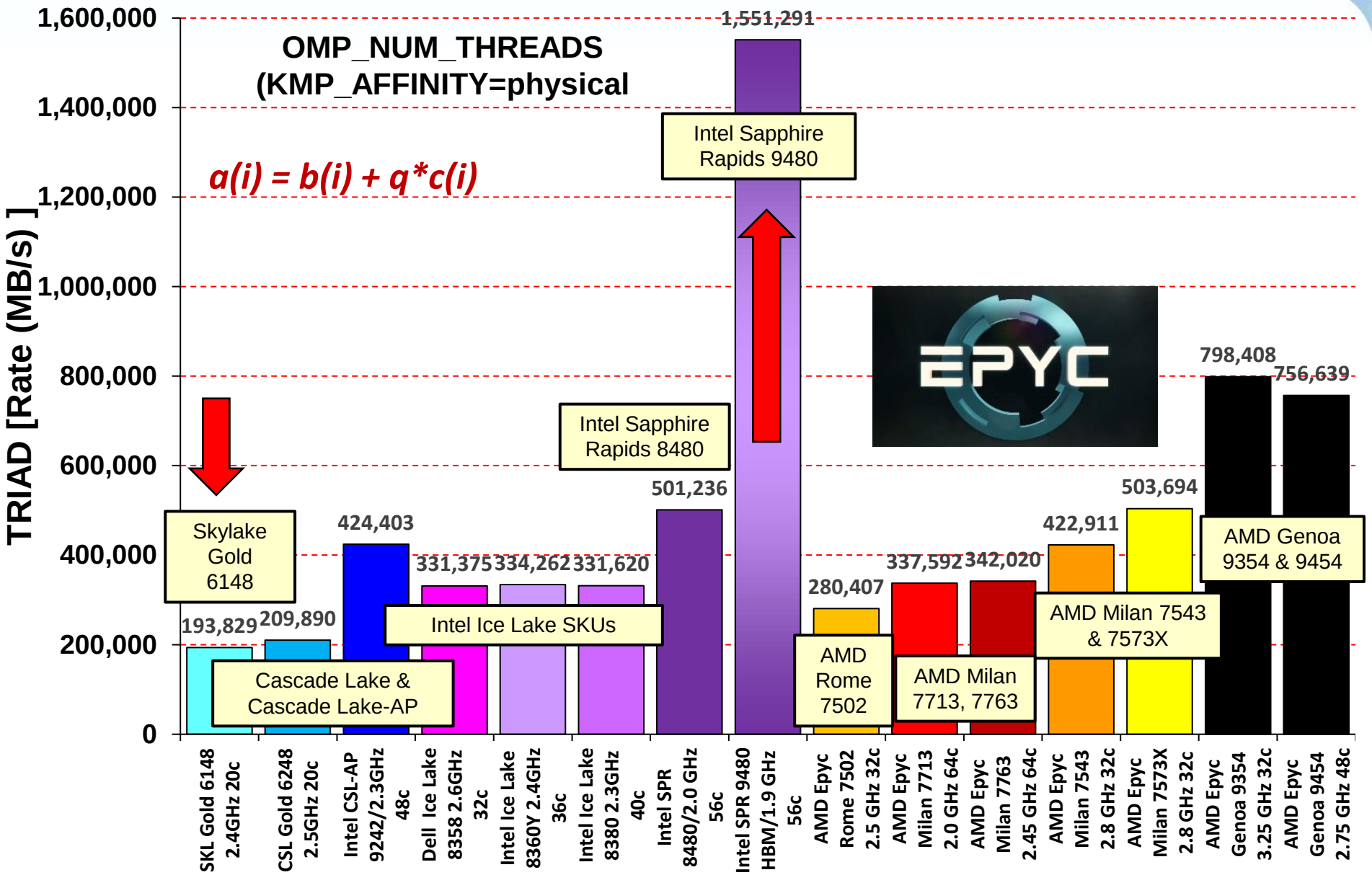
```
# Version of Intel compiler to use and way to source it
source /opt/intel/compilers_and_libraries_2020.2.254/linux/bin/compilervars.sh -
ofi_internal=1 intel64

# 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:

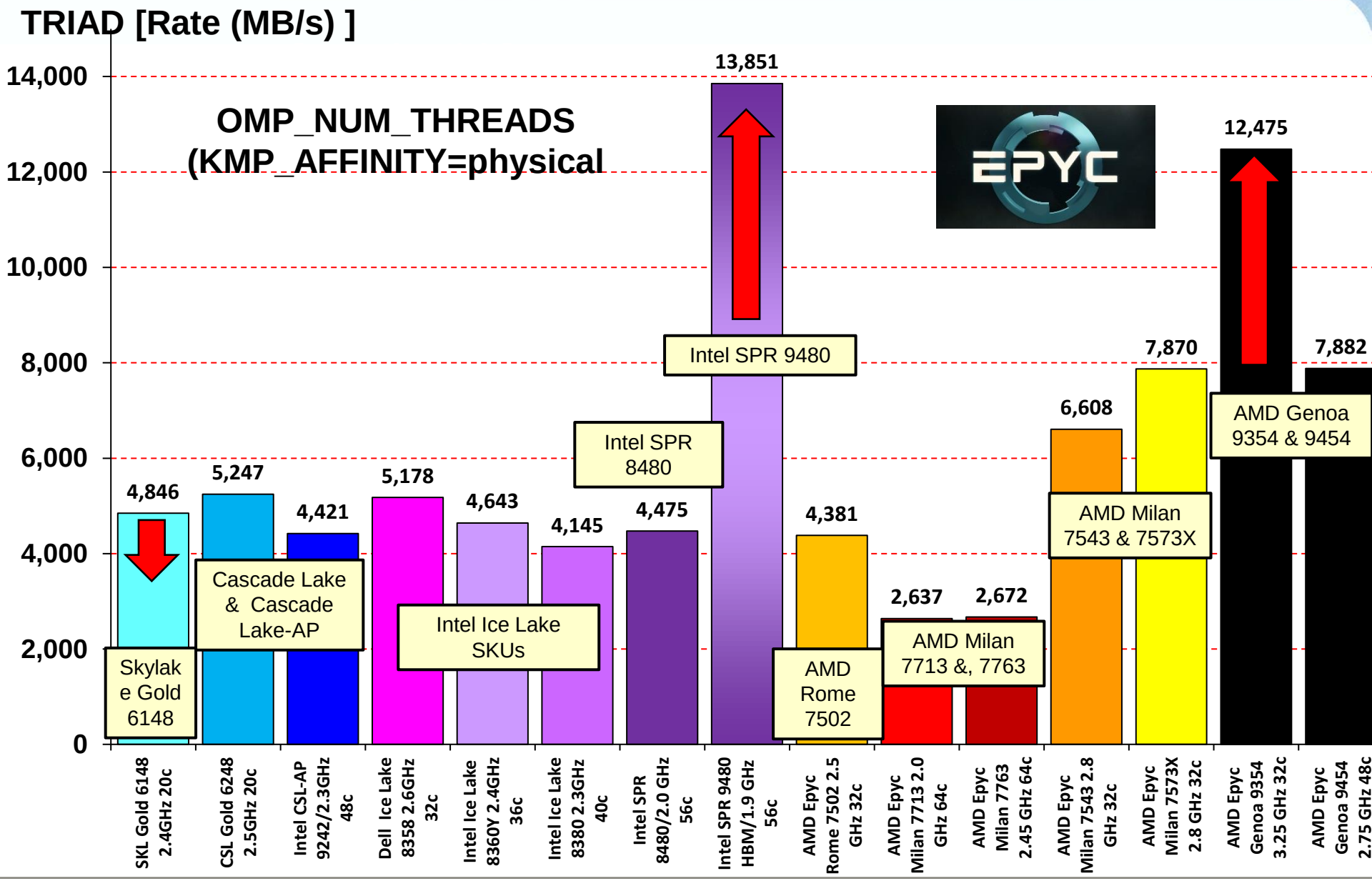
INTEL SKL: –O3 –xCORE-AVX512

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

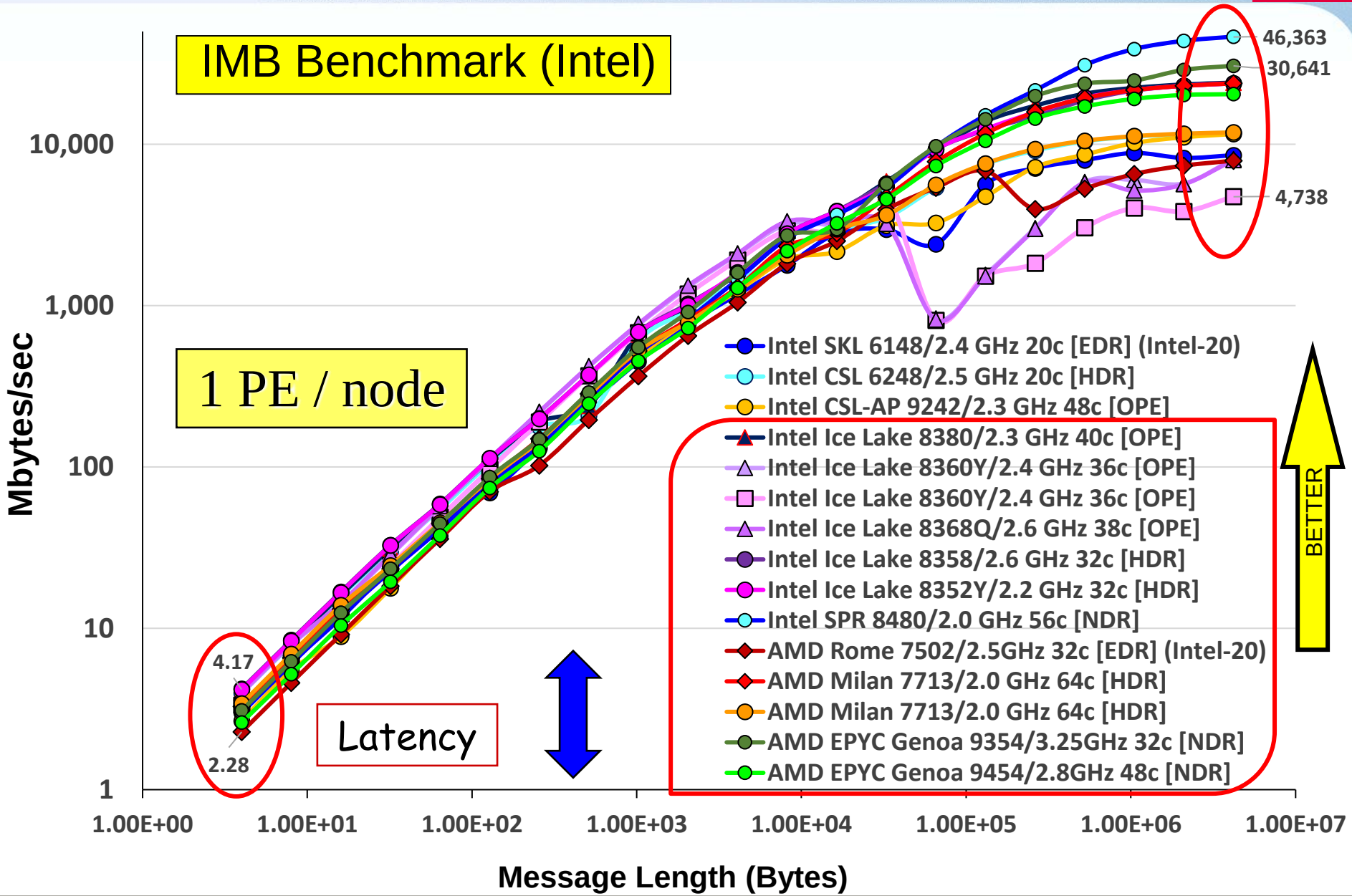
Memory B/W – STREAM performance



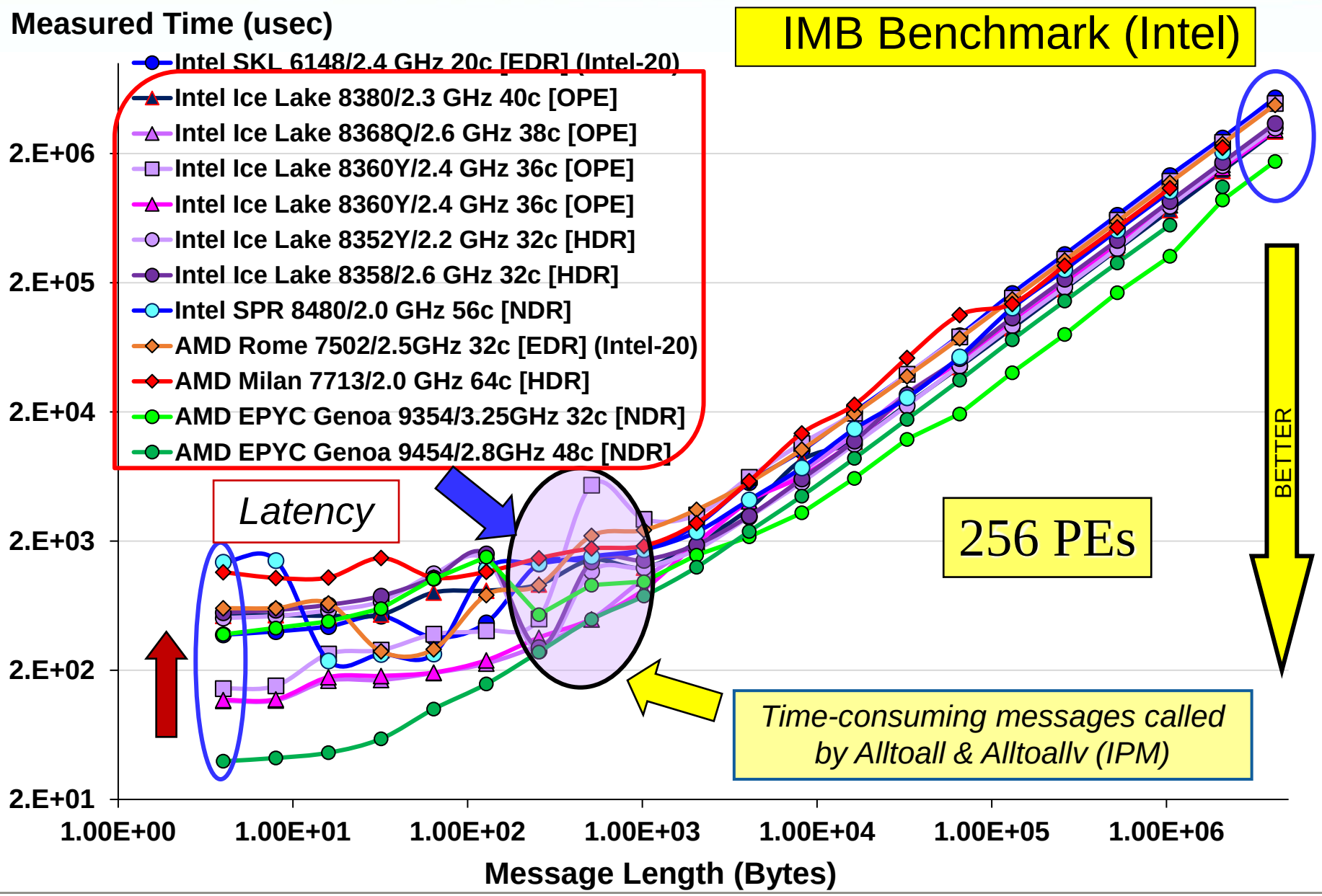
Memory B/W – STREAM / core performance



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 Genoa 9354 nodes with 64 cores per node. [1-8 nodes]***
2	***Hawk - Dell |EMC Skylake Gold 6148 2.4GHz (T) EDR with 40 cores / node***	***Intel Xeon Sapphire Rapids 8480 nodes with 112 cores per node. [1-8 nodes]***

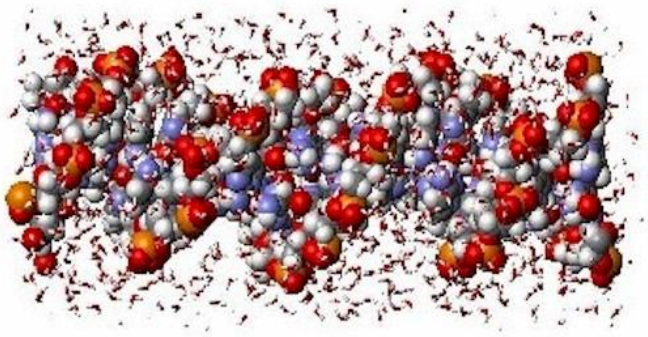
Performance of Computational Chemistry and Ocean Modelling Codes

Molecular Simulation;

1. DL_POLY

Molecular Simulation I. DL_POLY

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



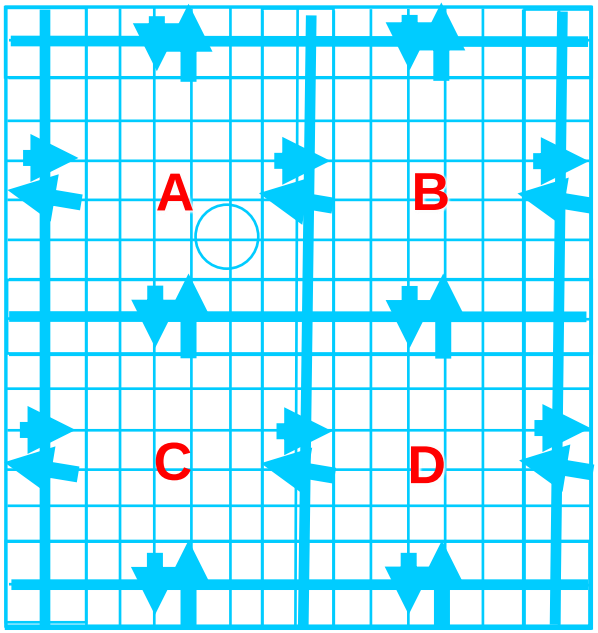
DL_POLY

- 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

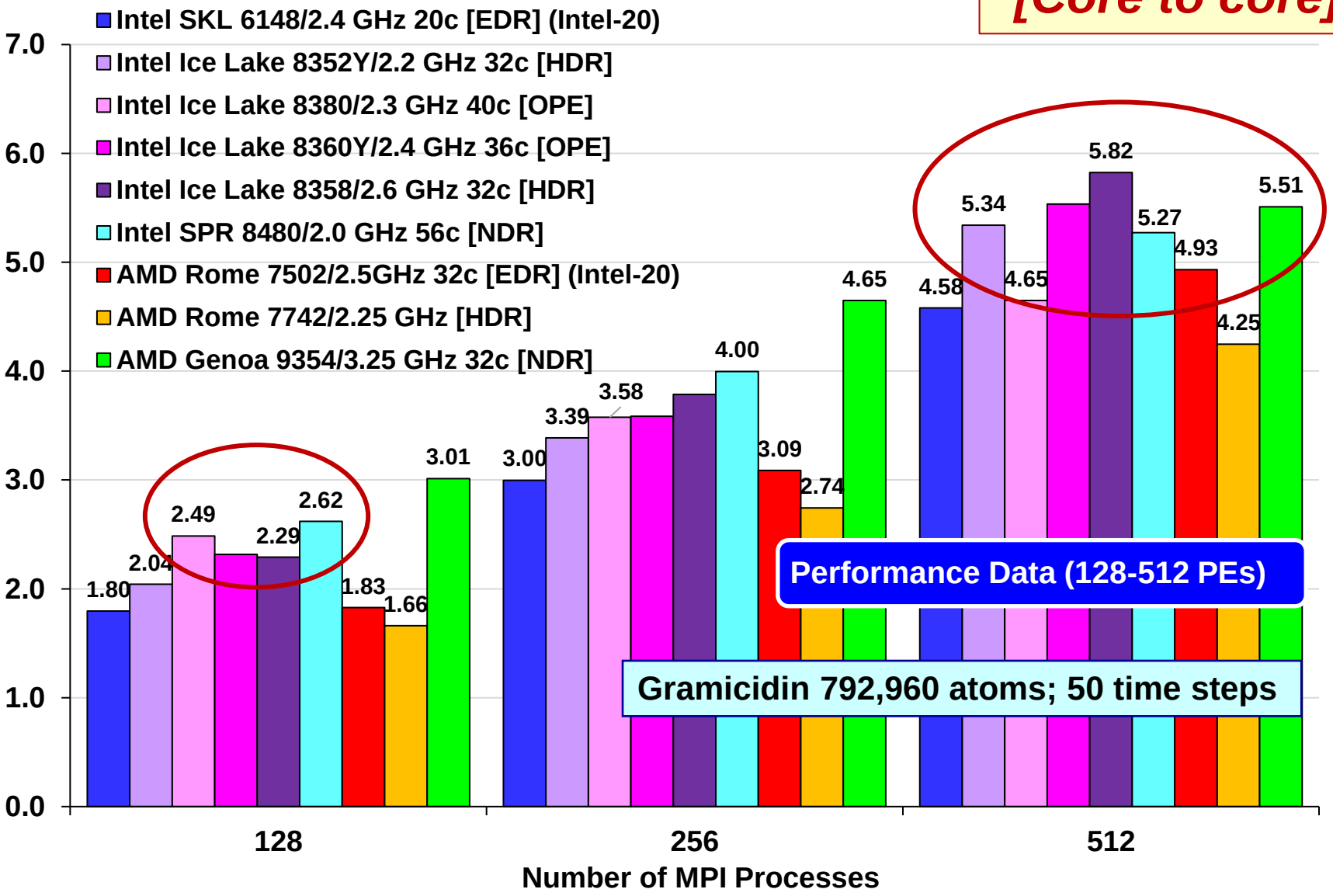
1. **NaCl Simulation; 216,000 ions, 200 time steps, Cutoff=12Å**
2. **Gramicidin in water; rigid bonds + SHAKE: 792,960 ions, 50 time steps**

https://www.scd.stfc.ac.uk/Pages/DL_POLY.aspx

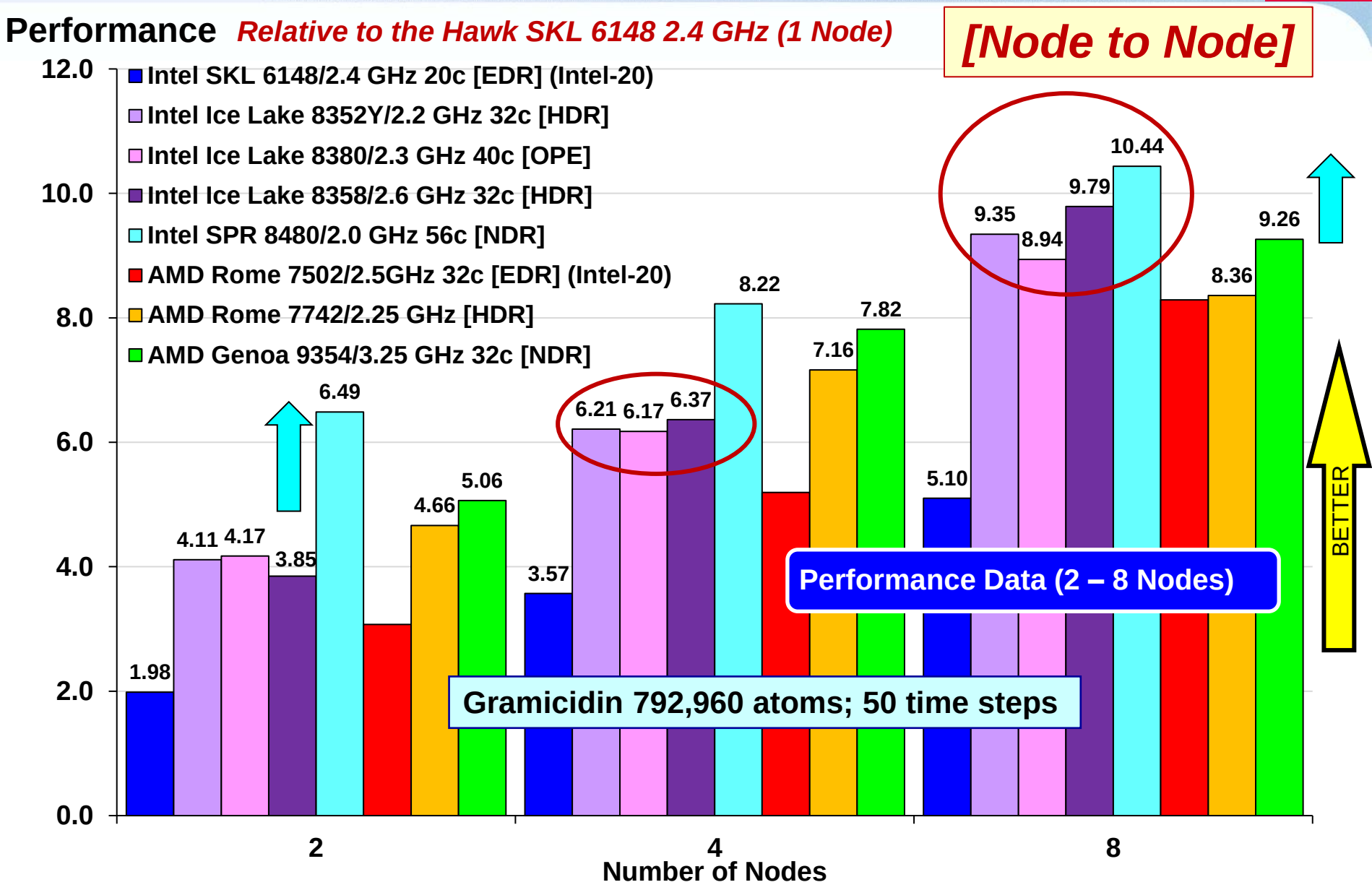
DL_POLY 4 – Gramicidin Simulation

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

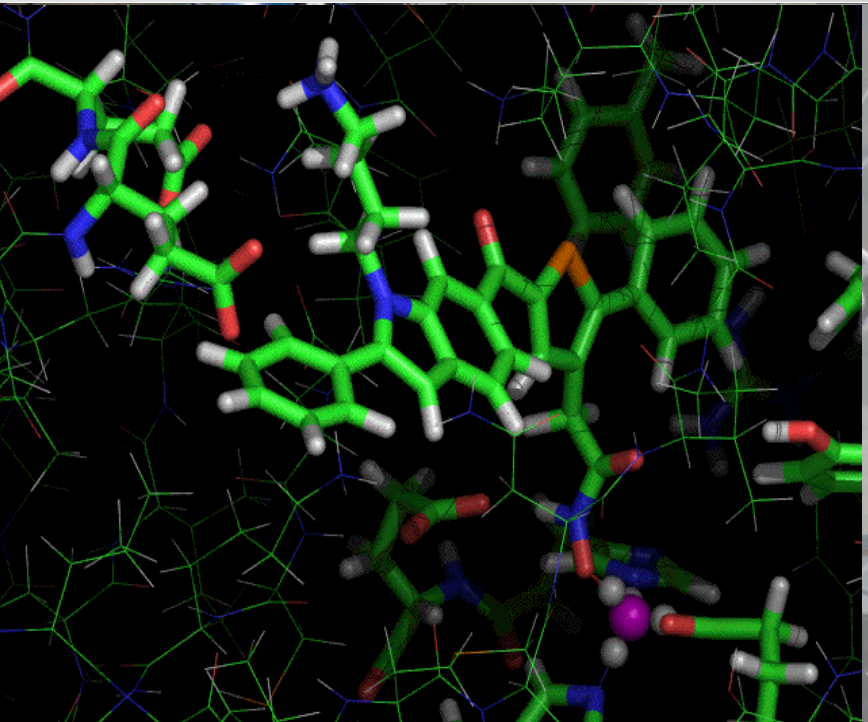
[Core to core]



DL_POLY 4 – Gramicidin Simulation

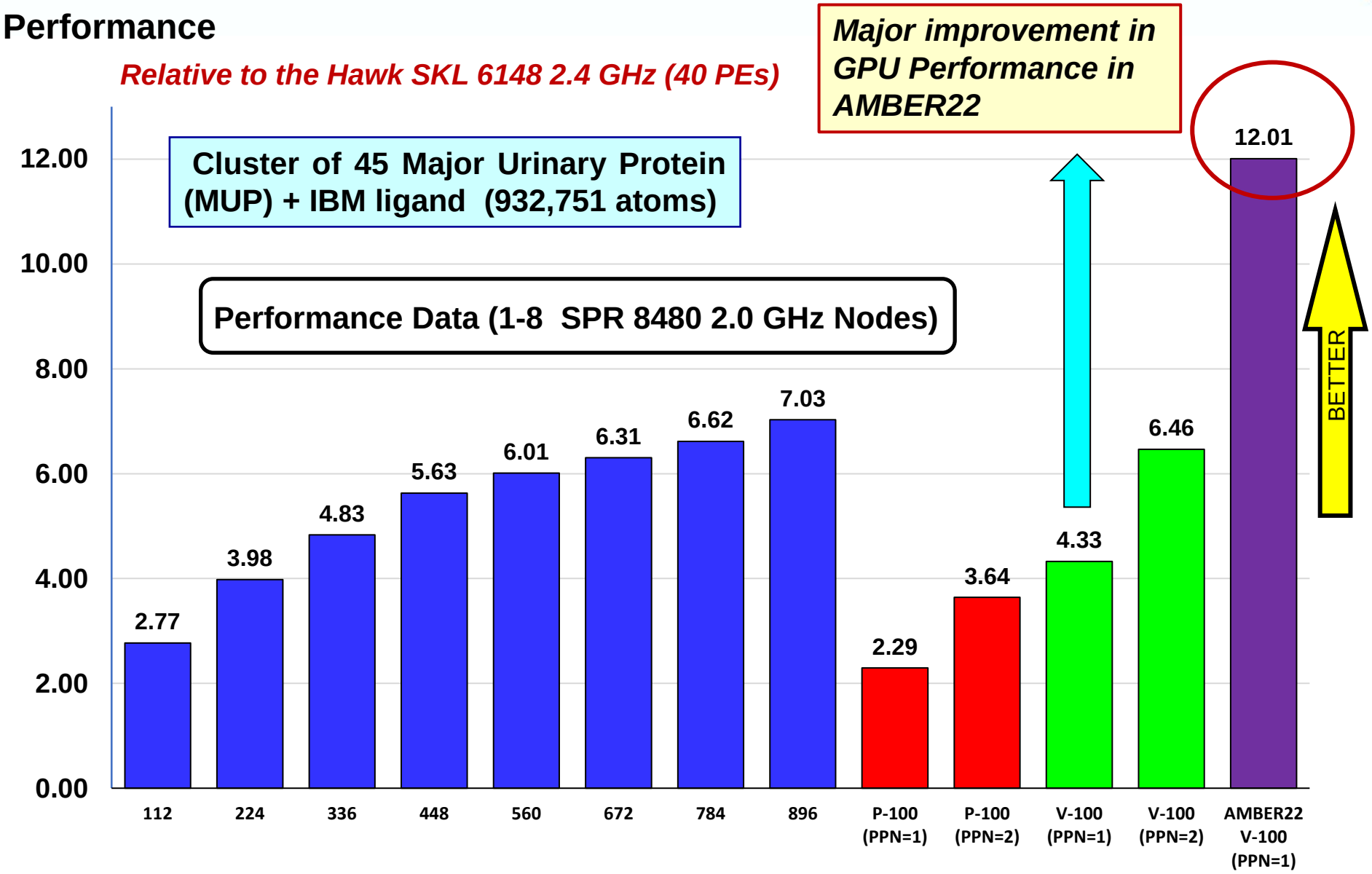


Performance of Computational Chemistry and Ocean Modelling Codes

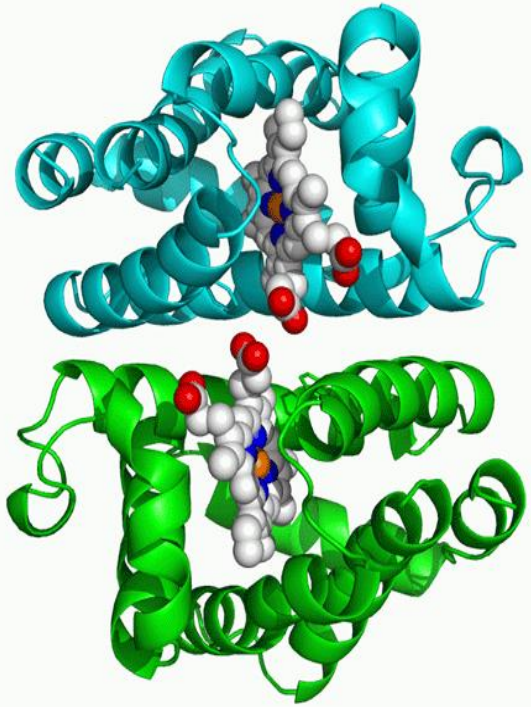


Molecular Simulation: 3. AMBER

AMBER – GPU Performance M45 Simulation



Performance of Computational Chemistry Codes



Molecular Simulation: 4. Gromacs

Molecular Simulation - GROMACS

Archer Rank: 7

GROMACS (GROningen MAchine for Chemical Simulations) is a molecular dynamics package designed for simulations of proteins, lipids and nucleic acids [University of Groningen] .

Versions under Test:

Version 4.6.1 – 5 March 2013

Version 5.0.7 – 14 October 2015

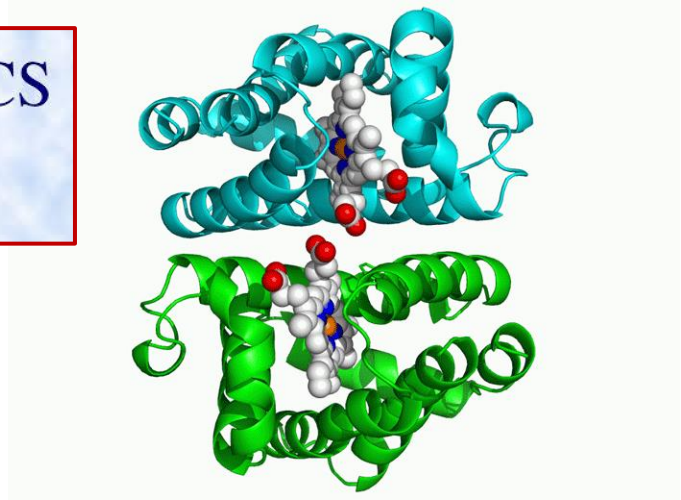
Version 2016.3 – 14 March 2017

Version 2018.2 – 14 June 2018

Version 2019.6 – 28 February 2020

Version 2020.1 – 3 March 2020

Version 2023.1 – 21 April 2023



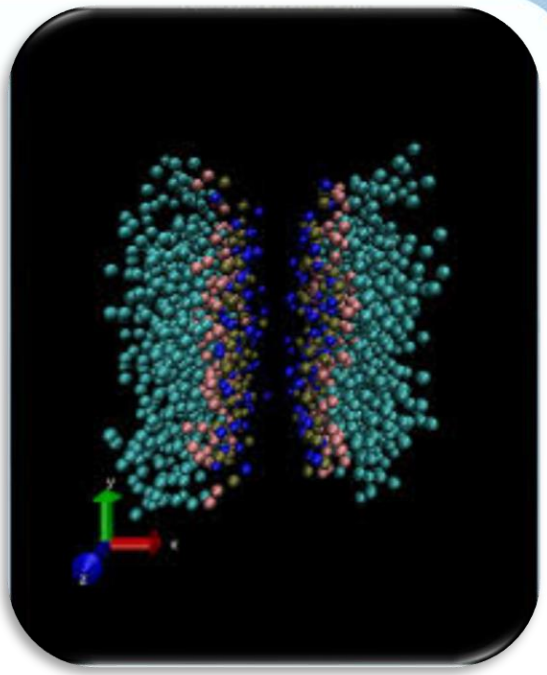
- **Berk Hess et al. "*GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation*". *Journal of Chemical Theory and Computation* 4 (3): 435–447.**

http://manual.gromacs.org/documentation/

GROMACS Benchmark Cases

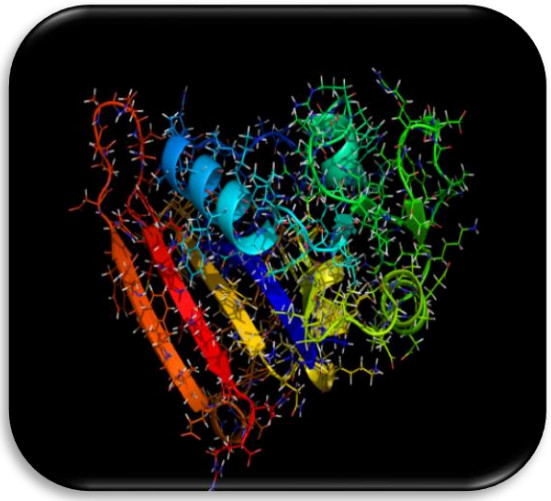
Ion channel system

- The 142k particle ion channel system is the membrane protein GluCl - a pentameric chloride channel embedded in a DOPC membrane and solvated in TIP3P water, using the Amber ff99SB-ILDN force field. This system is a **challenging** parallelization case due to the small size, but was one of the **wanted target sizes** for biomolecular simulations



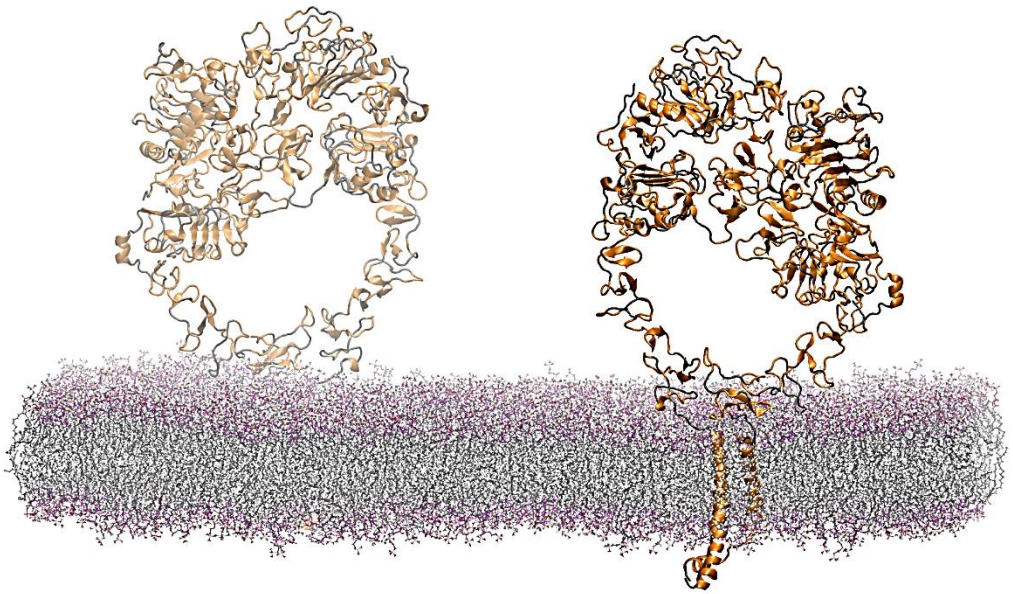
Lignocellulose

- Gromacs Test Case B from the UEA Benchmark Suite. A model of cellulose and lignocellulosic biomass in an aqueous solution. This system of 3.3M atoms is inhomogeneous, and uses **reaction-field electrostatics** instead of PME and therefore should scale well.



GROMACS Benchmark Case II

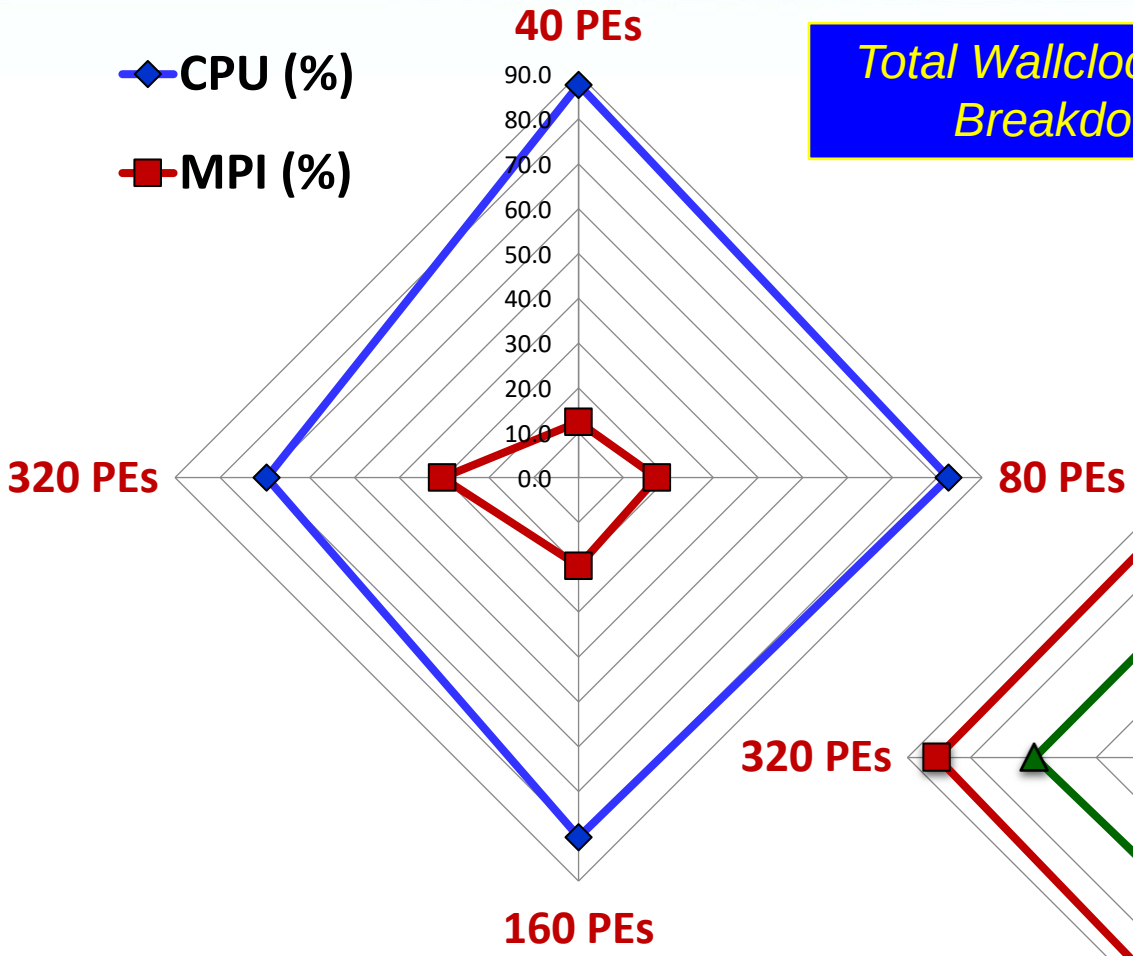
The HECBioSim Benchmarks



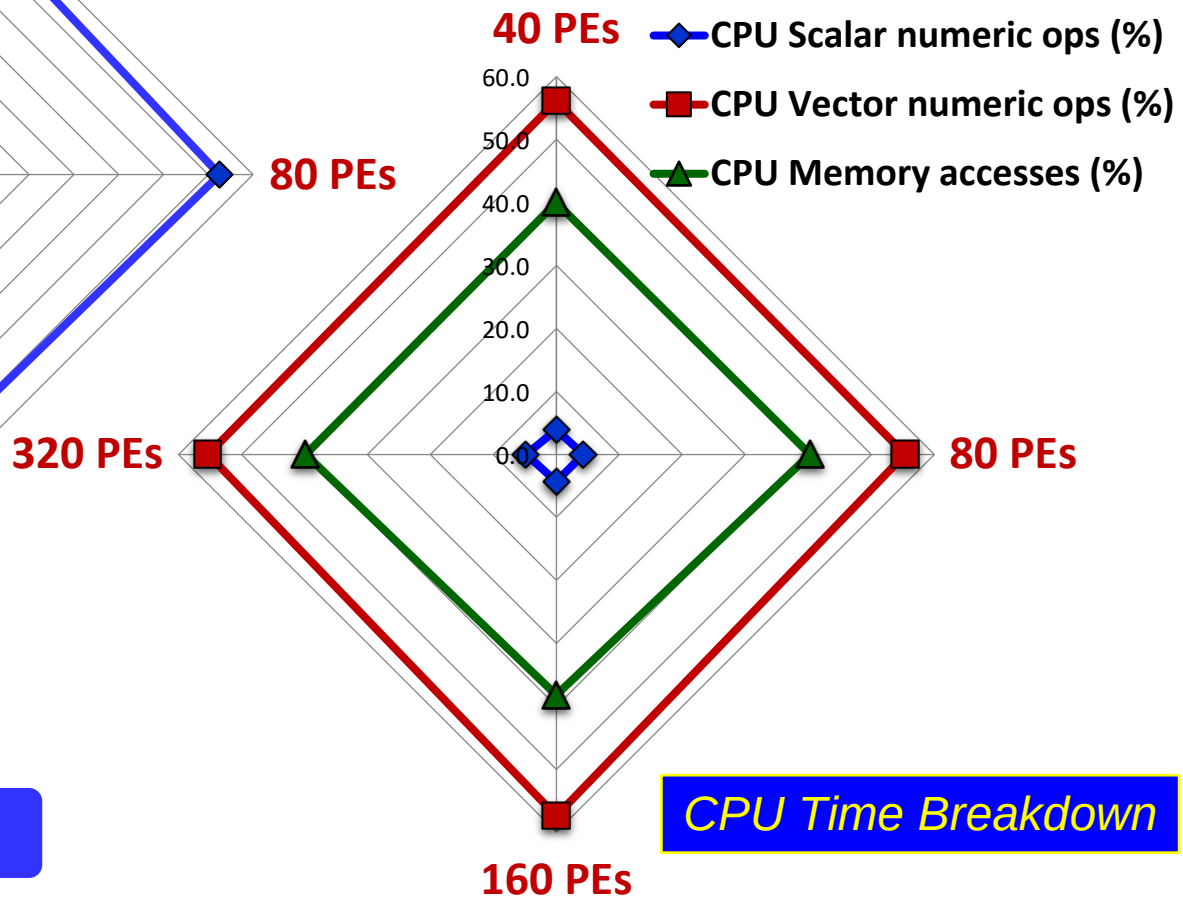
- PME simulation for 1.4M atom system - A Pair of Human Epidermal Growth Factor Receptor (hEGFR) Dimers of 1IVO and 1NQL
- Total number of atoms = **1,403,182**
- Protein atoms = 43,498 Lipid atoms = 235,304 Water atoms = 1,123,392 Ions = 986

https://www.hecbiosim.ac.uk/benchmarks

GROMACS – HECBioSim Performance Report



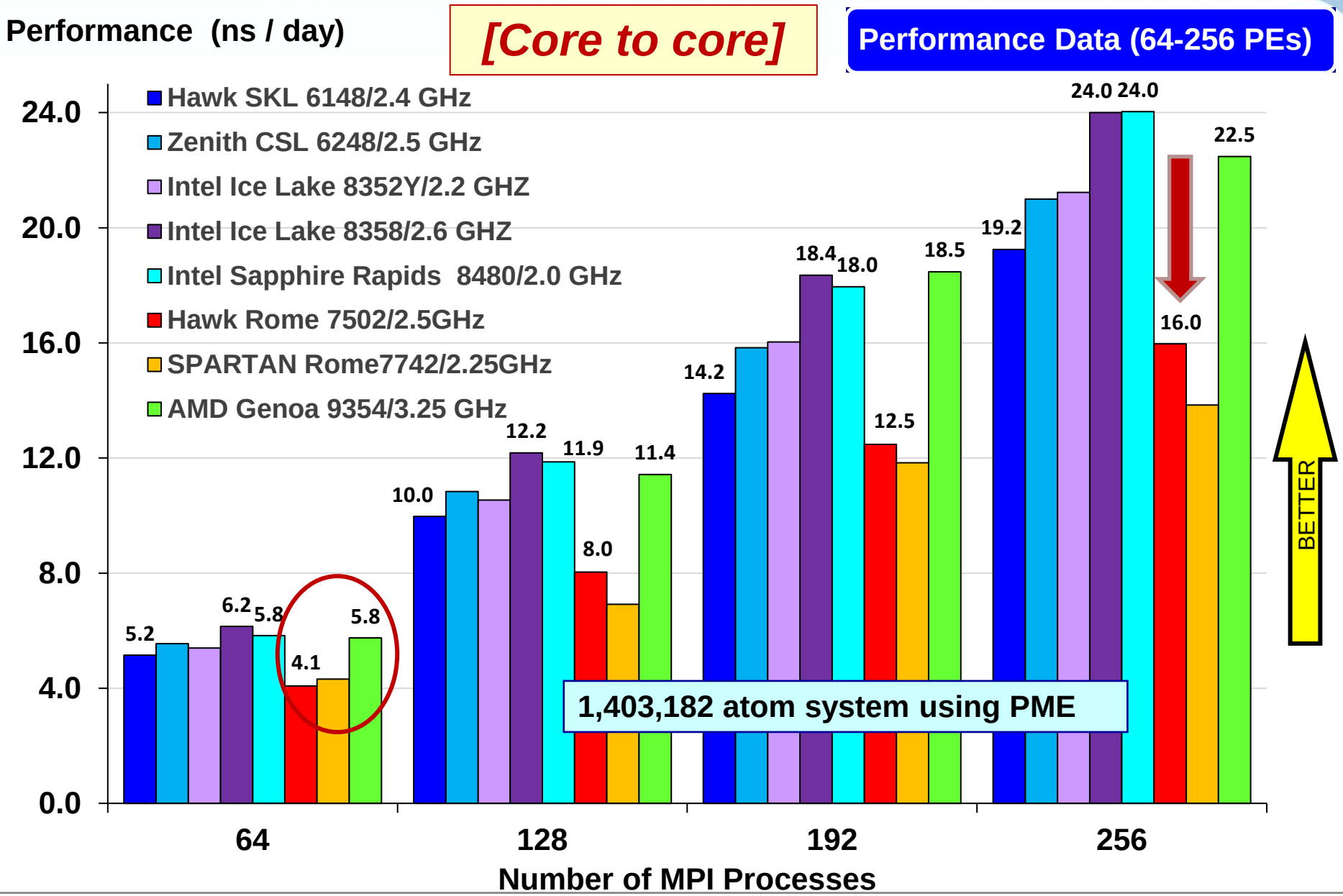
Total Wallclock Time Breakdown



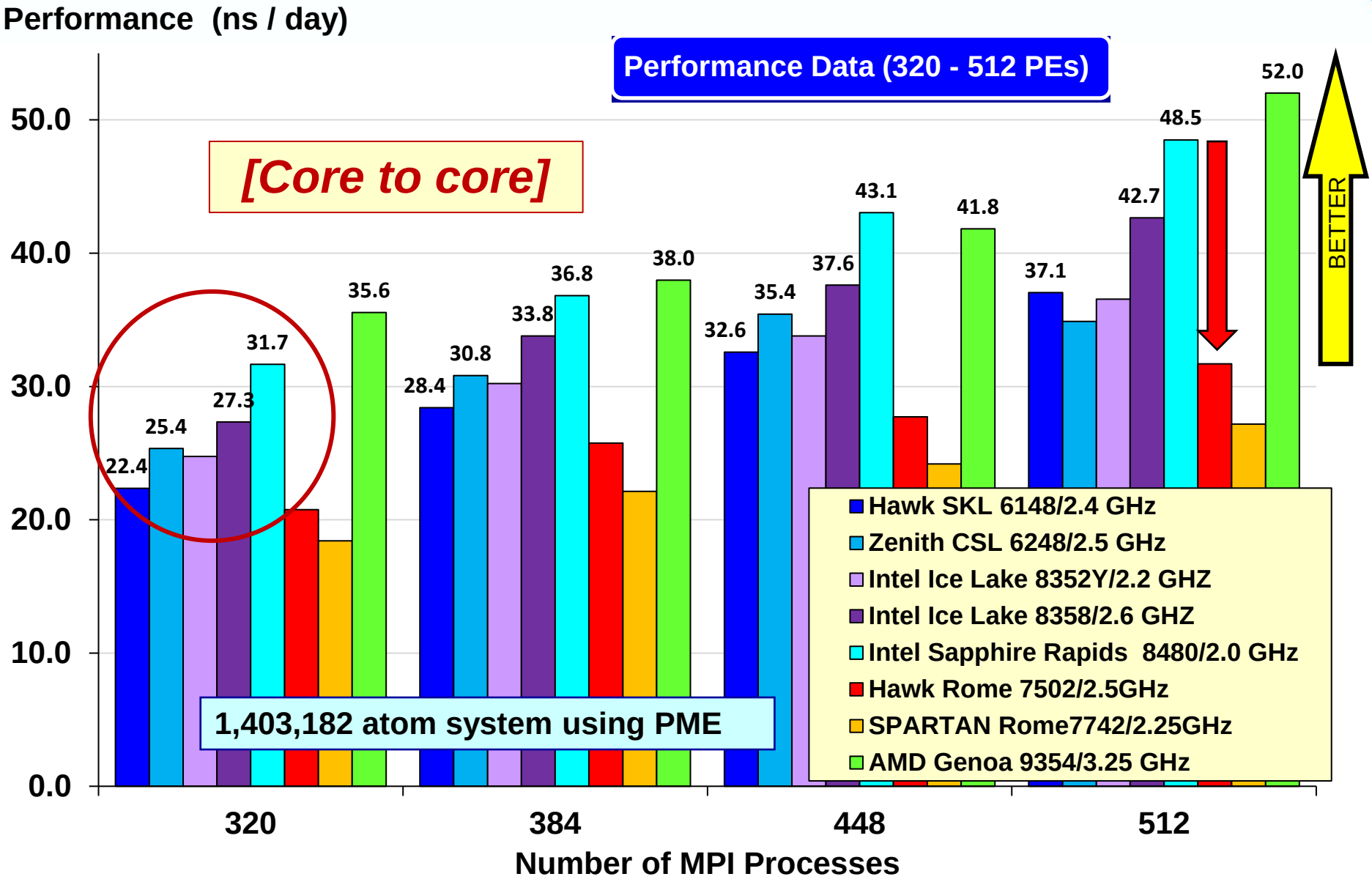
CPU Time Breakdown

Performance Data (40-320 PEs)

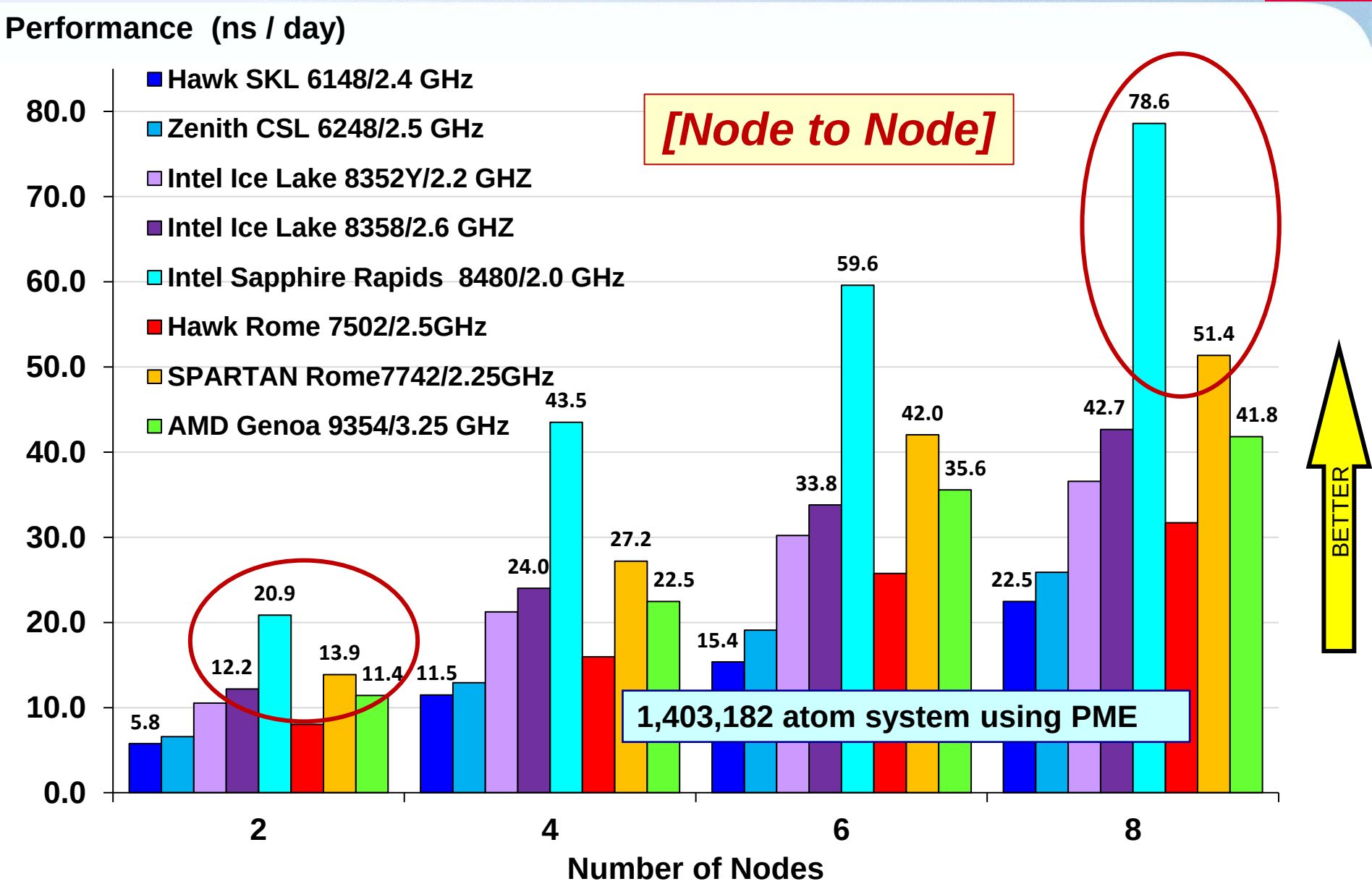
GROMACS – HECBioSim 1.4M Atom System



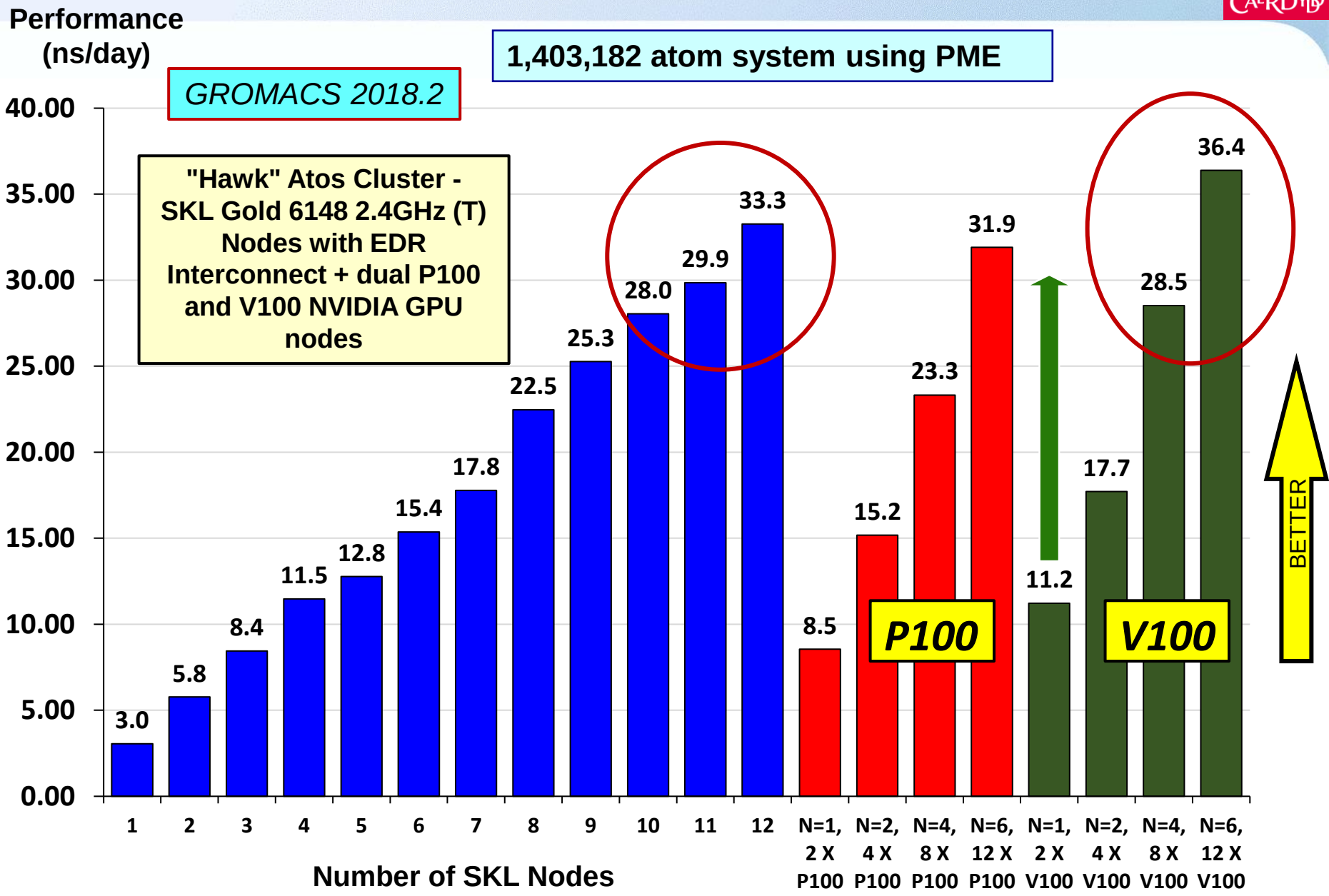
GROMACS – HECBioSim 1.4M Atom System



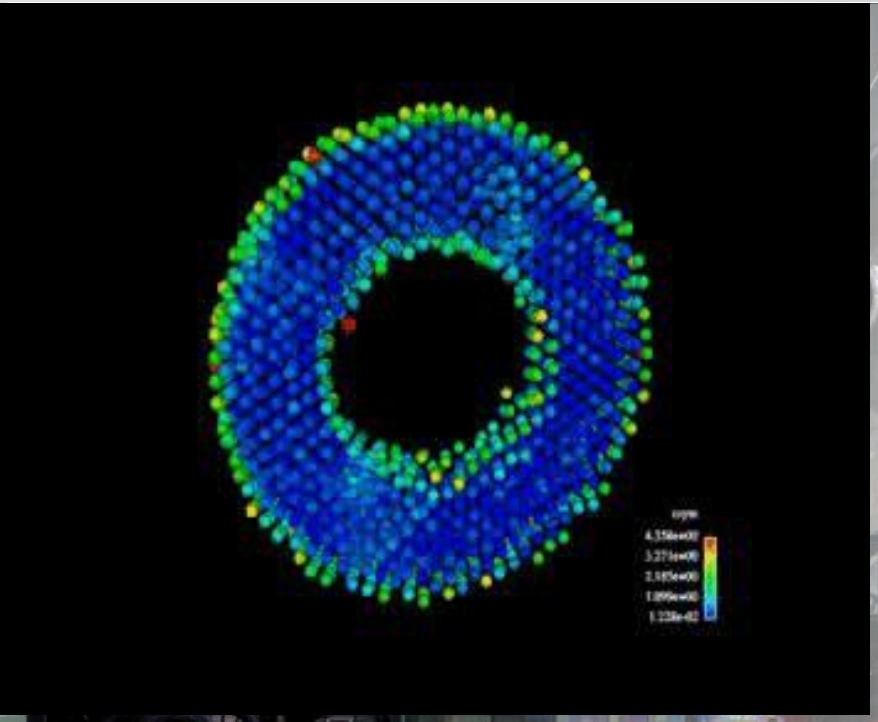
GROMACS – HECBioSim 1.4M Atom System



GROMACS – GPU Performance: HECBioSim Simulation



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

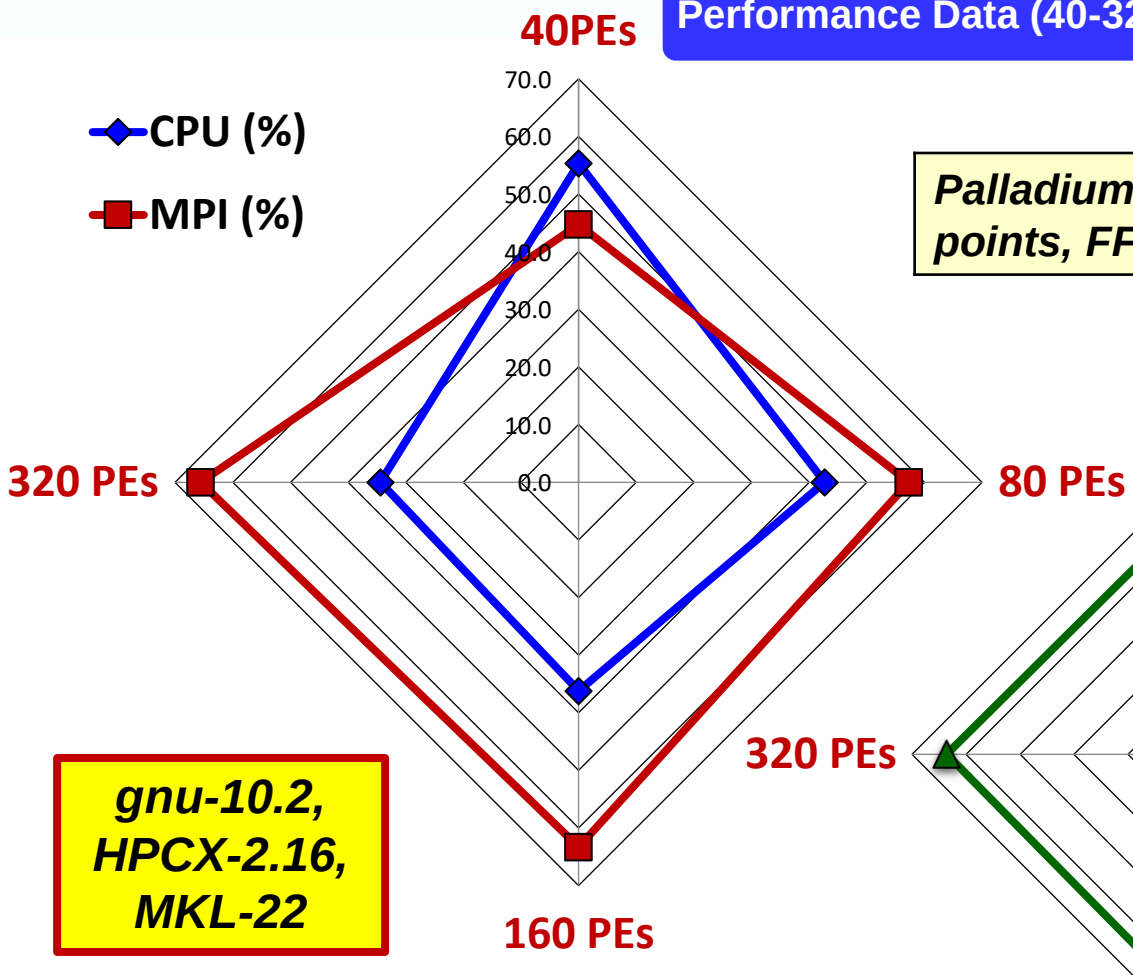
- Pd-O complex – $Pd_{75}O_{12}$, 5X4 3-layer supercell running a single point calculation and a planewave cut off of 400eV. Uses the RMM-DIIS algorithm for the SCF and is calculated in real space.
- 10 k-points; maximum number of plane-waves: 34,470
- FFT grid; NGX=31, NGY=49, NGZ=45, giving a total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

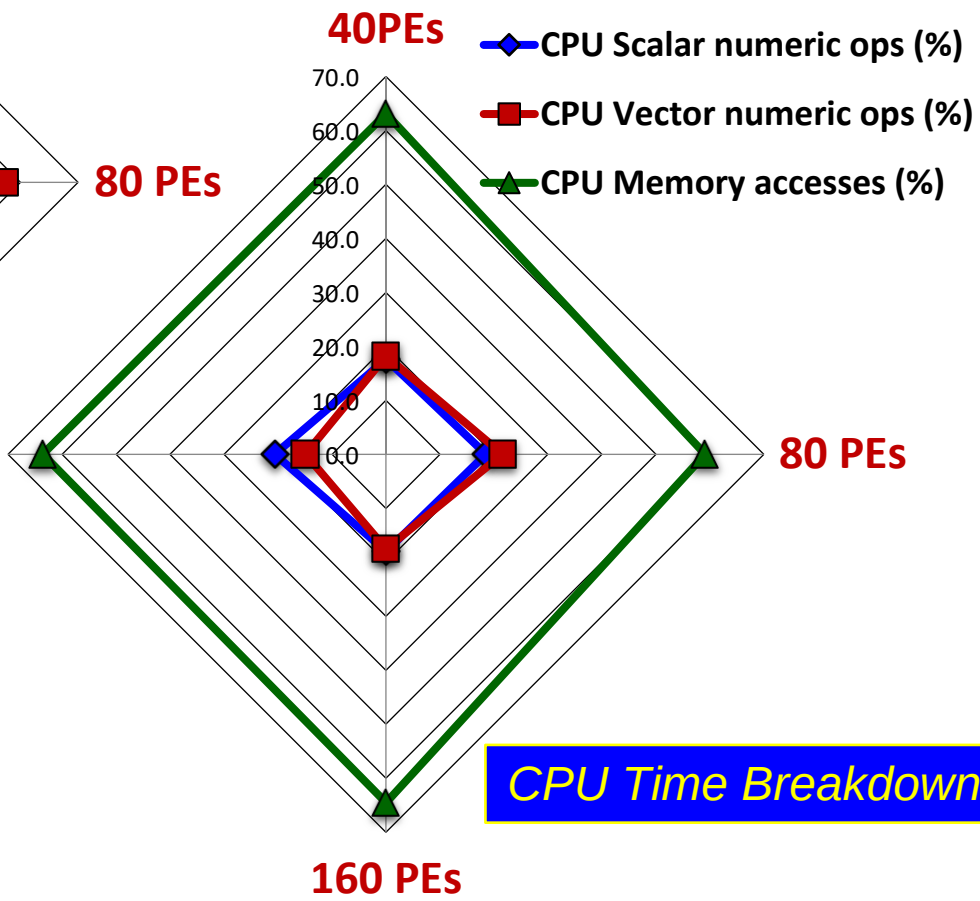
Performance Data (40-320 PEs)



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

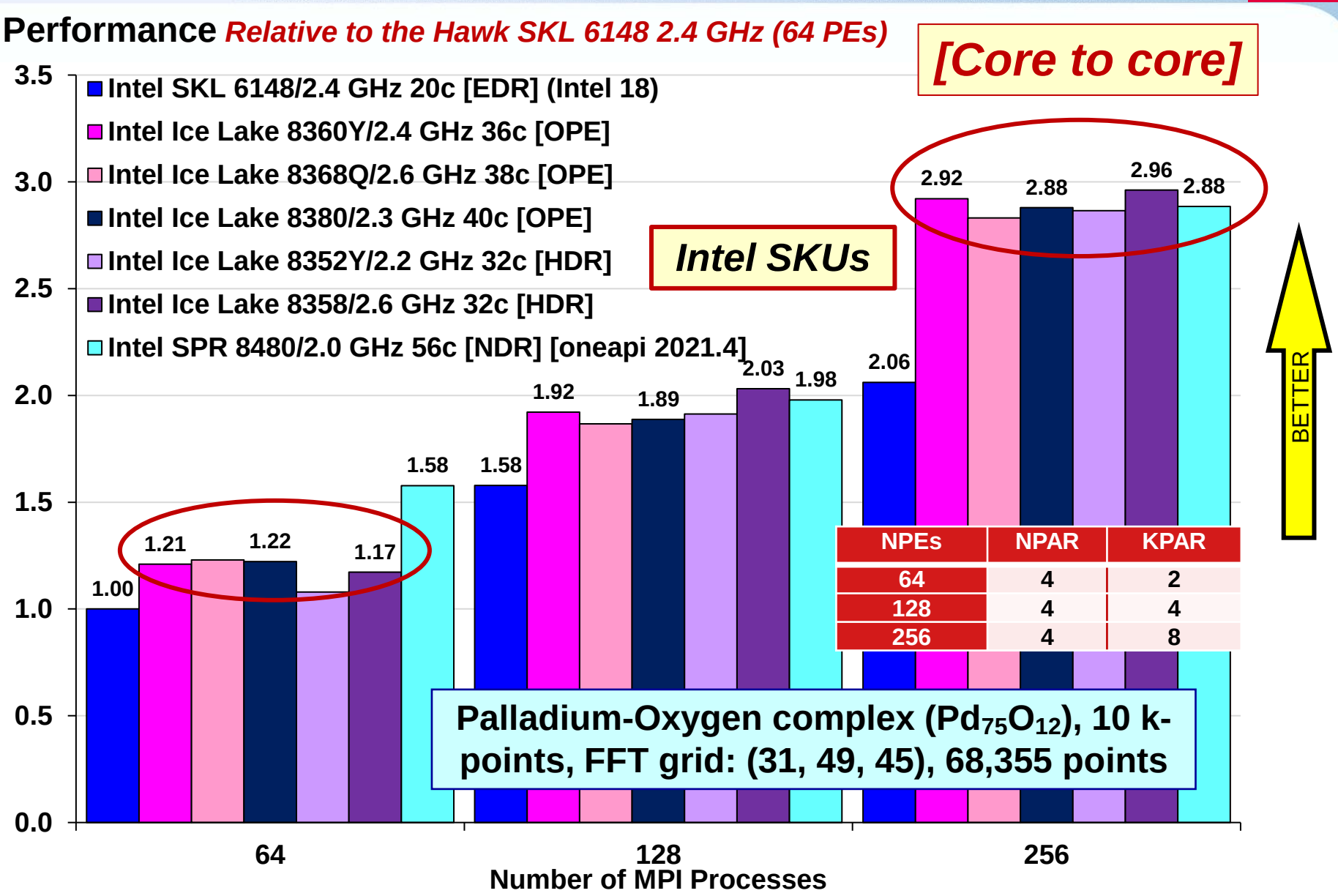
gnu-10.2, HPCX-2.16, MKL-22

Total Wallclock Time Breakdown



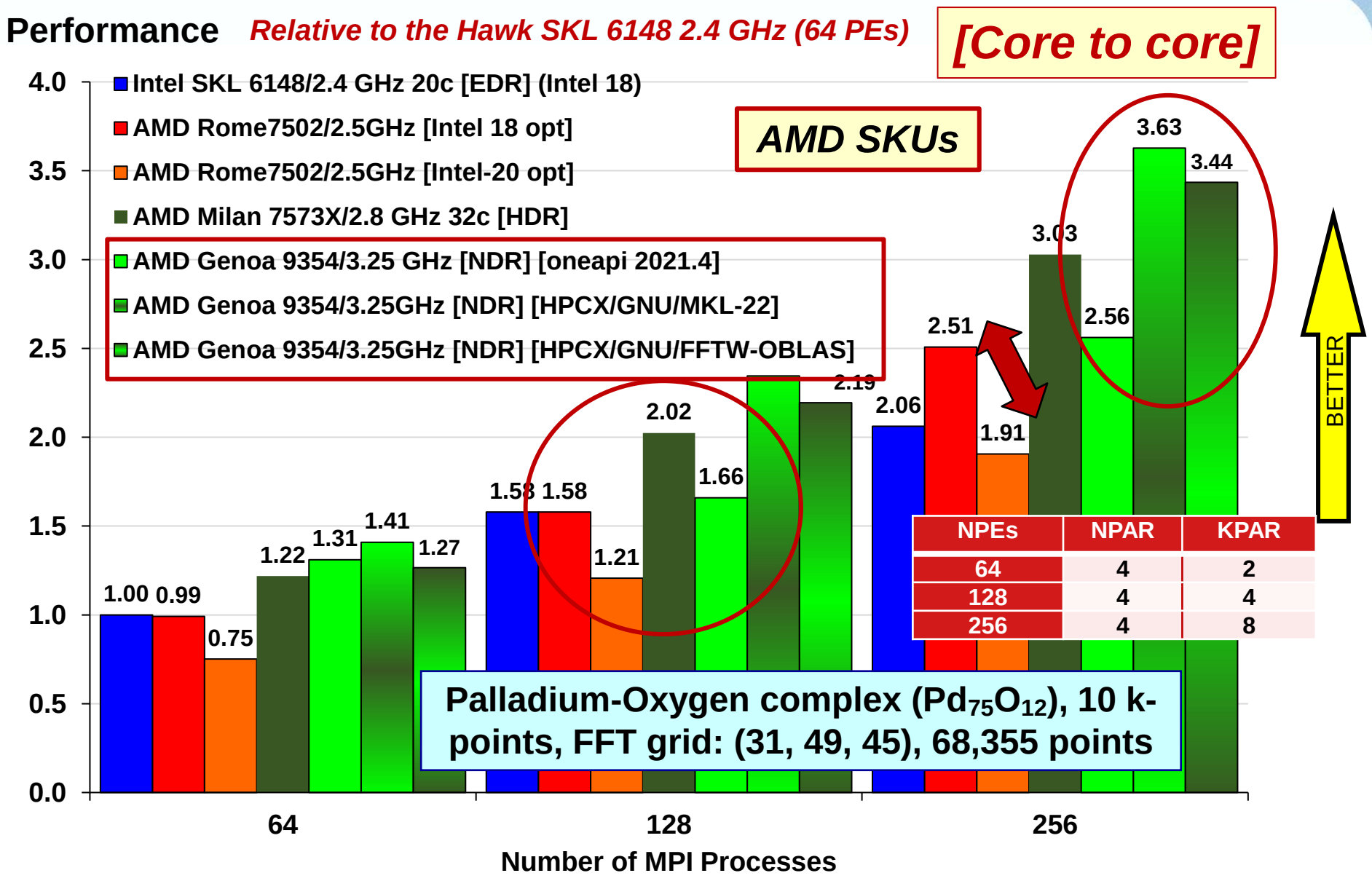
CPU Time Breakdown

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



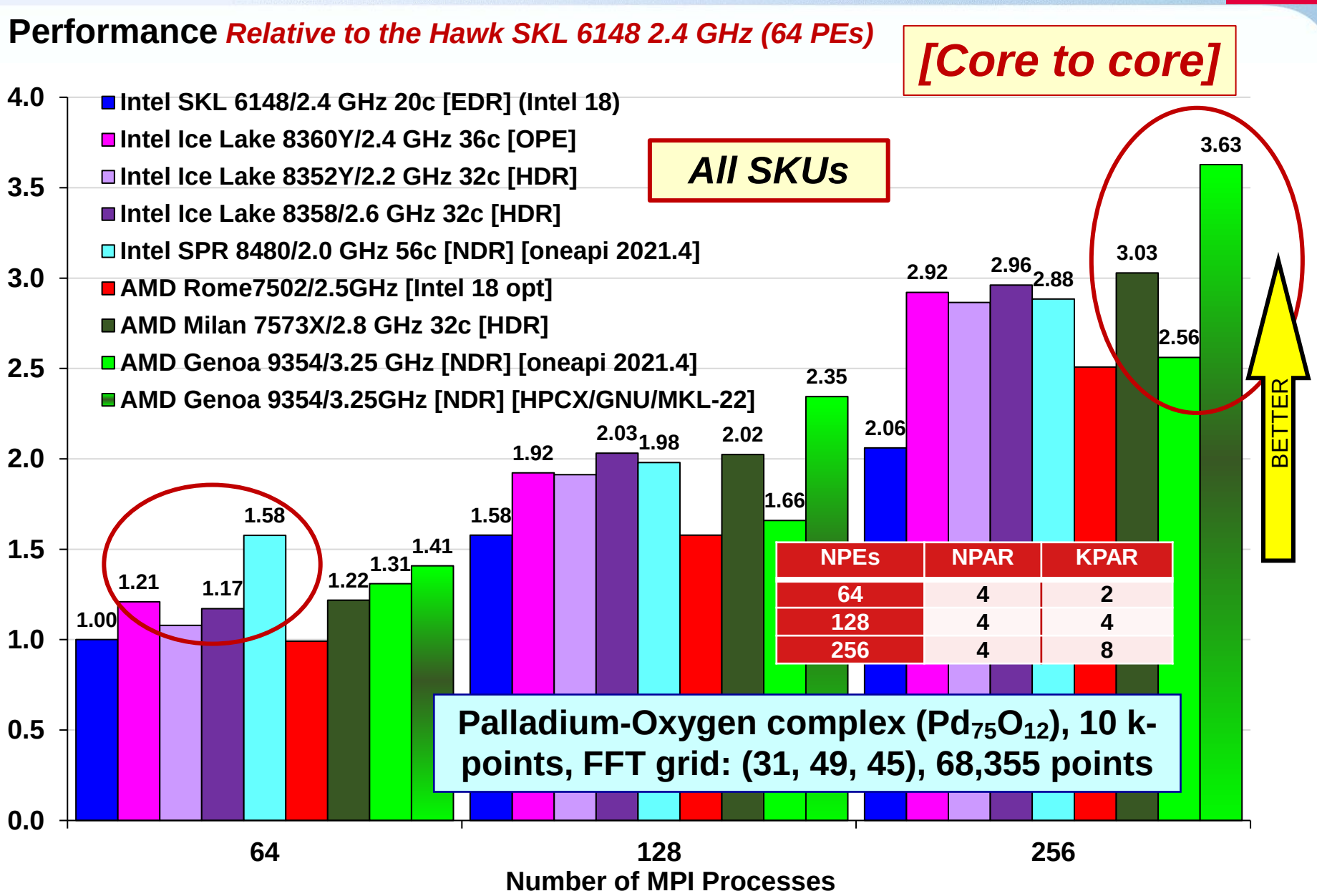
NPEs	NPAR	KPAR
64	4	2
128	4	4
256	4	8

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

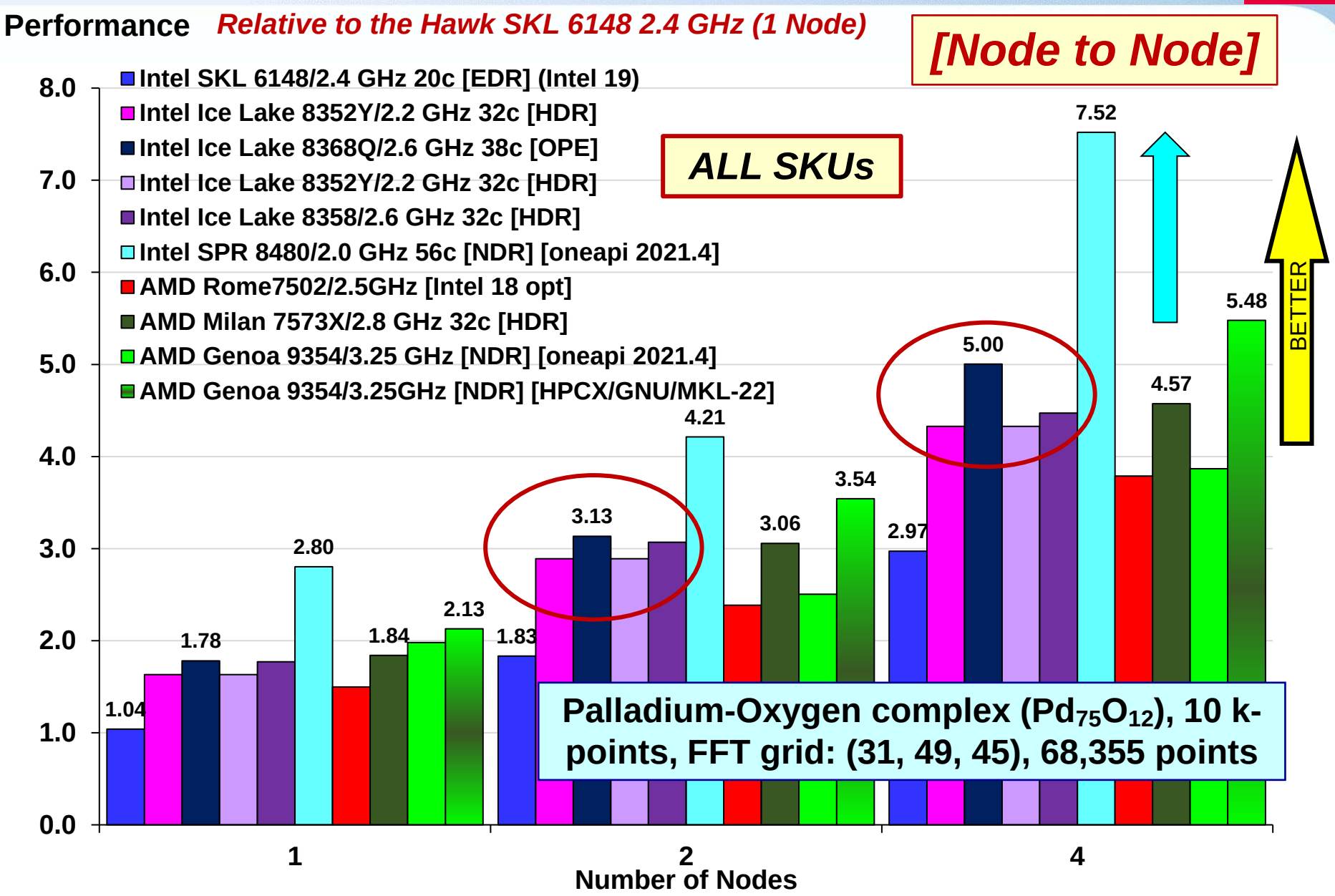


NPEs	NPAR	KPAR
64	4	2
128	4	4
256	4	8

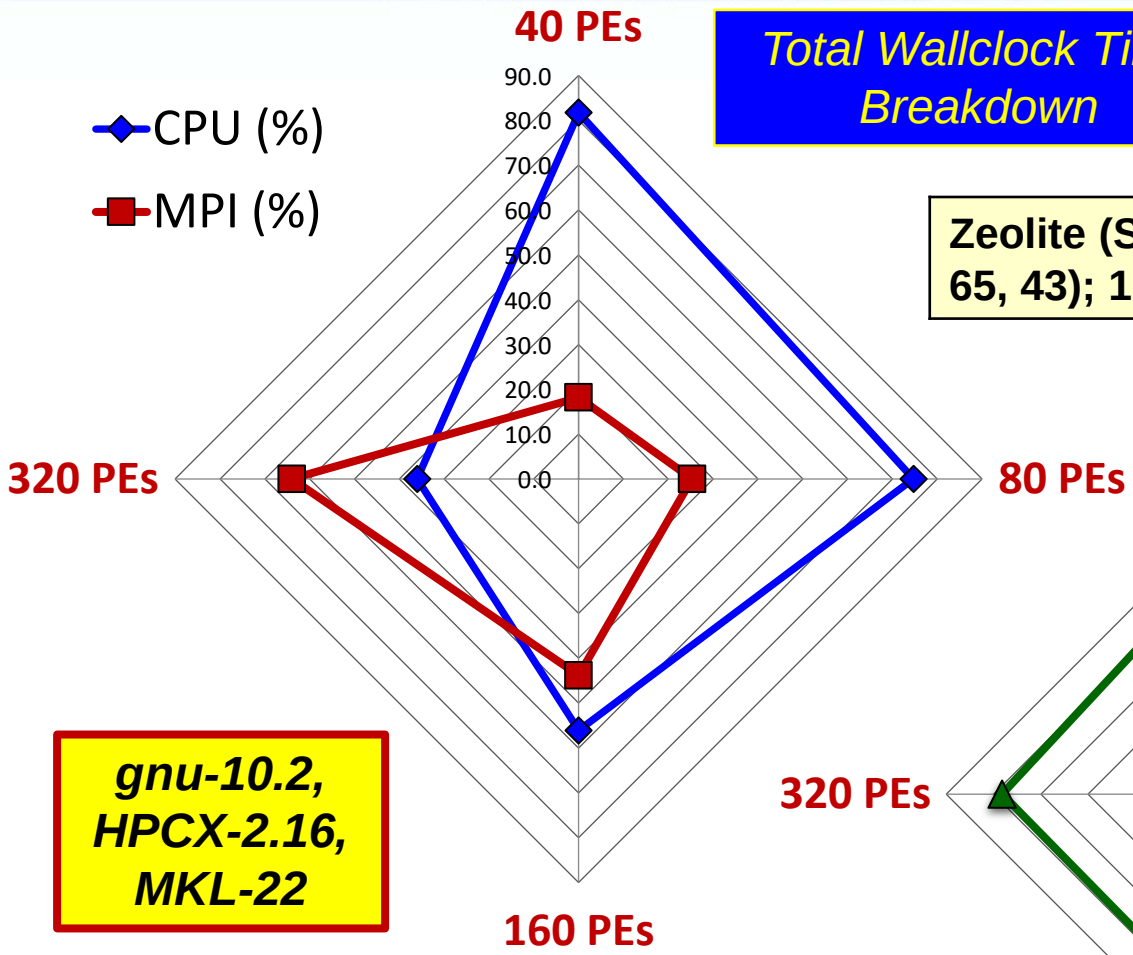
VASP 6.3 – Pd-O Benchmark - Parallelisation on k-points



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

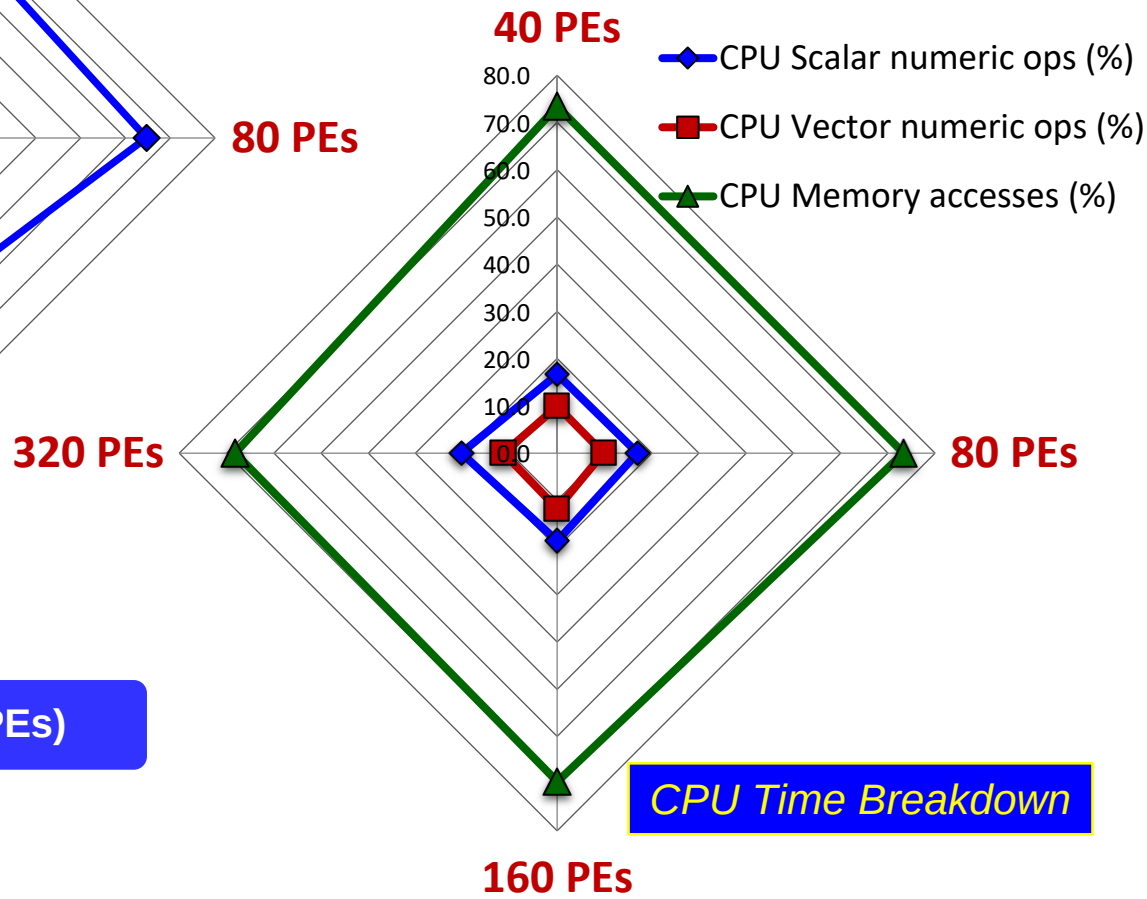


VASP – Zeolite Cluster Performance Report



Total Wallclock Time Breakdown

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



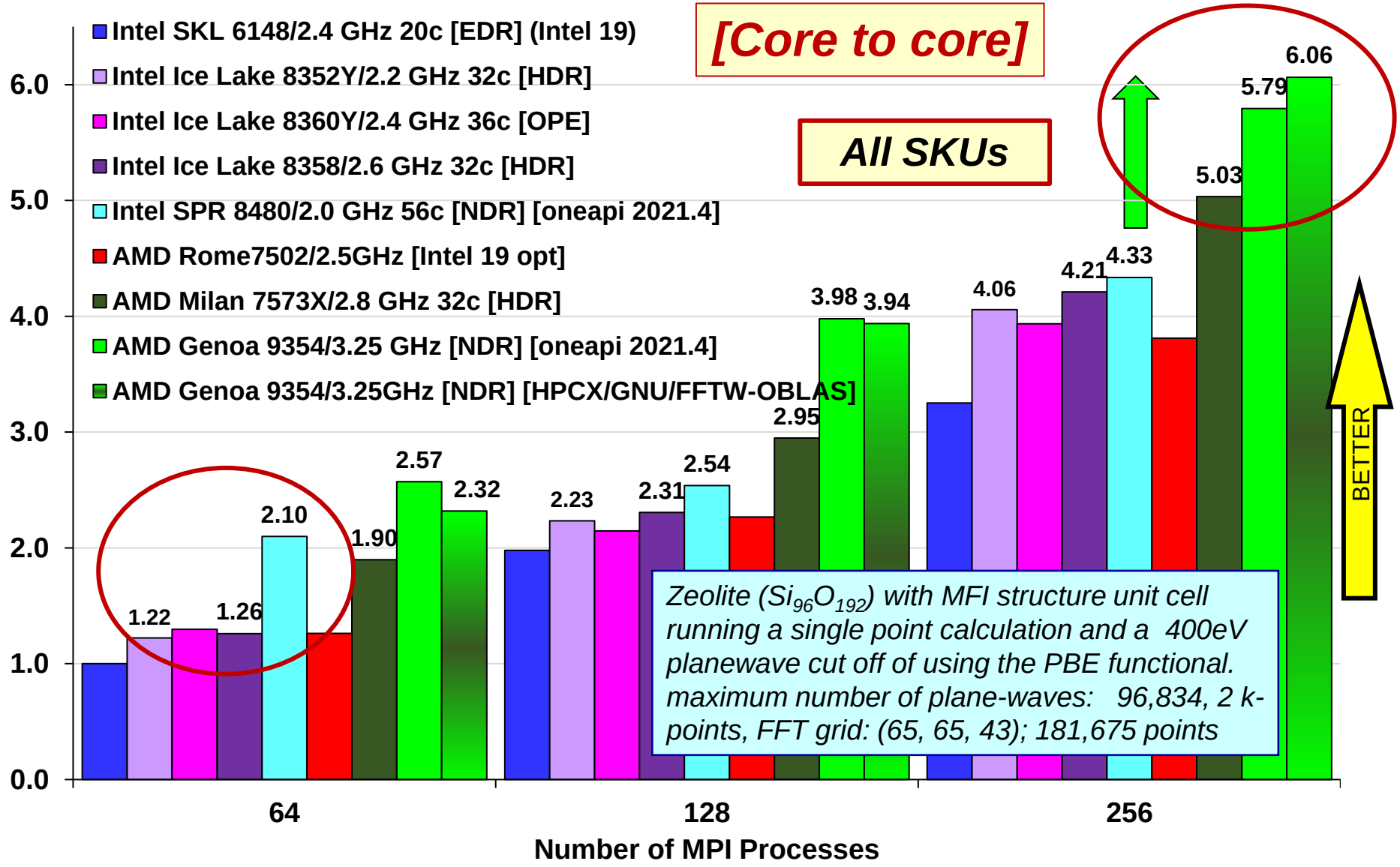
CPU Time Breakdown

gnu-10.2, HPCX-2.16, MKL-22

Performance Data (40-320 PEs)

VASP 6.3 – Zeolite Benchmark - Parallelisation on k-points

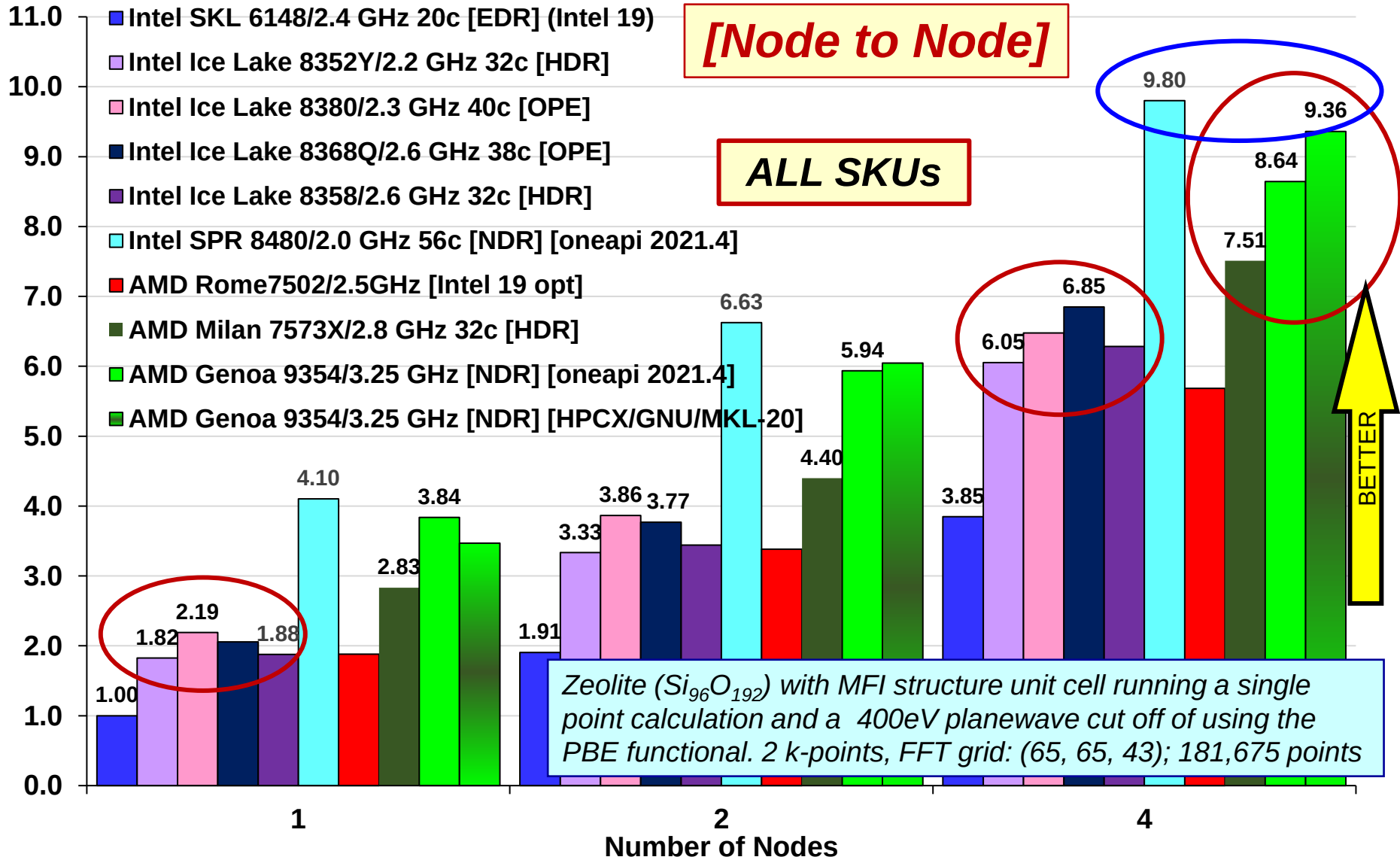
Performance *Relative to the Hawk SKL 6148 2.4 GHz (64 PEs)*



Zeolite ($Si_{96}O_{192}$) with MFI structure unit cell running a single point calculation and a 400eV planewave cut off of using the PBE functional. maximum number of plane-waves: 96,834, 2 k-points, FFT grid: (65, 65, 43); 181,675 points

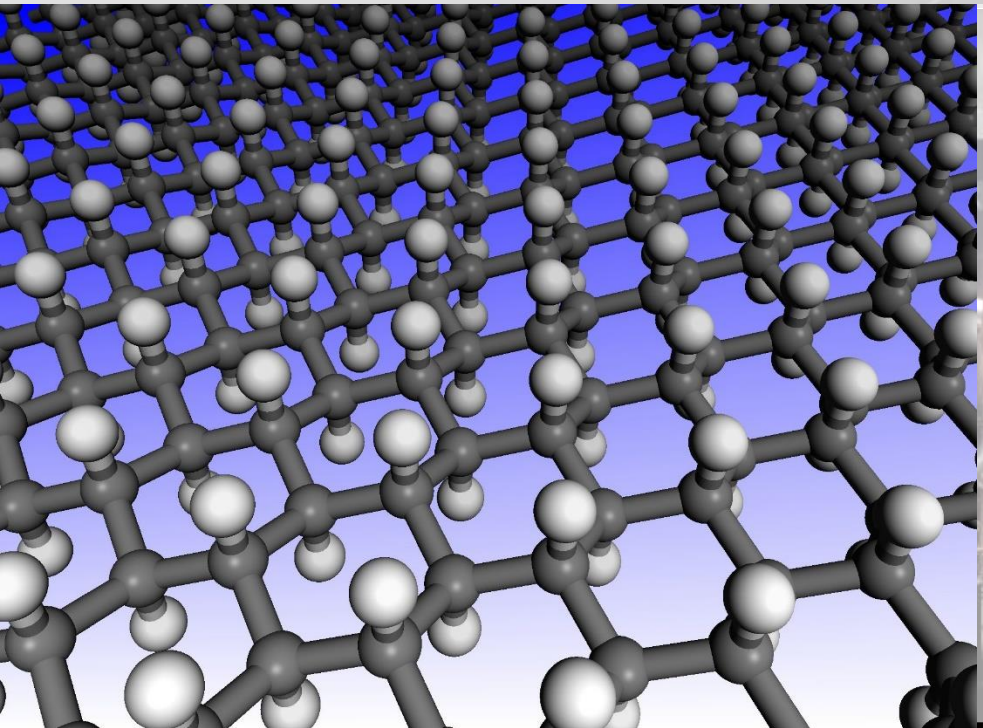
VASP 6.3 – Zeolite Benchmark - Parallelisation on k-points

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



Zeolite ($Si_{96}O_{192}$) with MFI structure unit cell running a single point calculation and a 400eV planewave cut off of using the PBE functional. 2 k-points, FFT grid: (65, 65, 43); 181,675 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**.
- Parallelisation over g-vectors leads to a global data exchange to transpose the FFT grid in 3-dimensions i.e., **MPI_alltoallv.**

- **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**

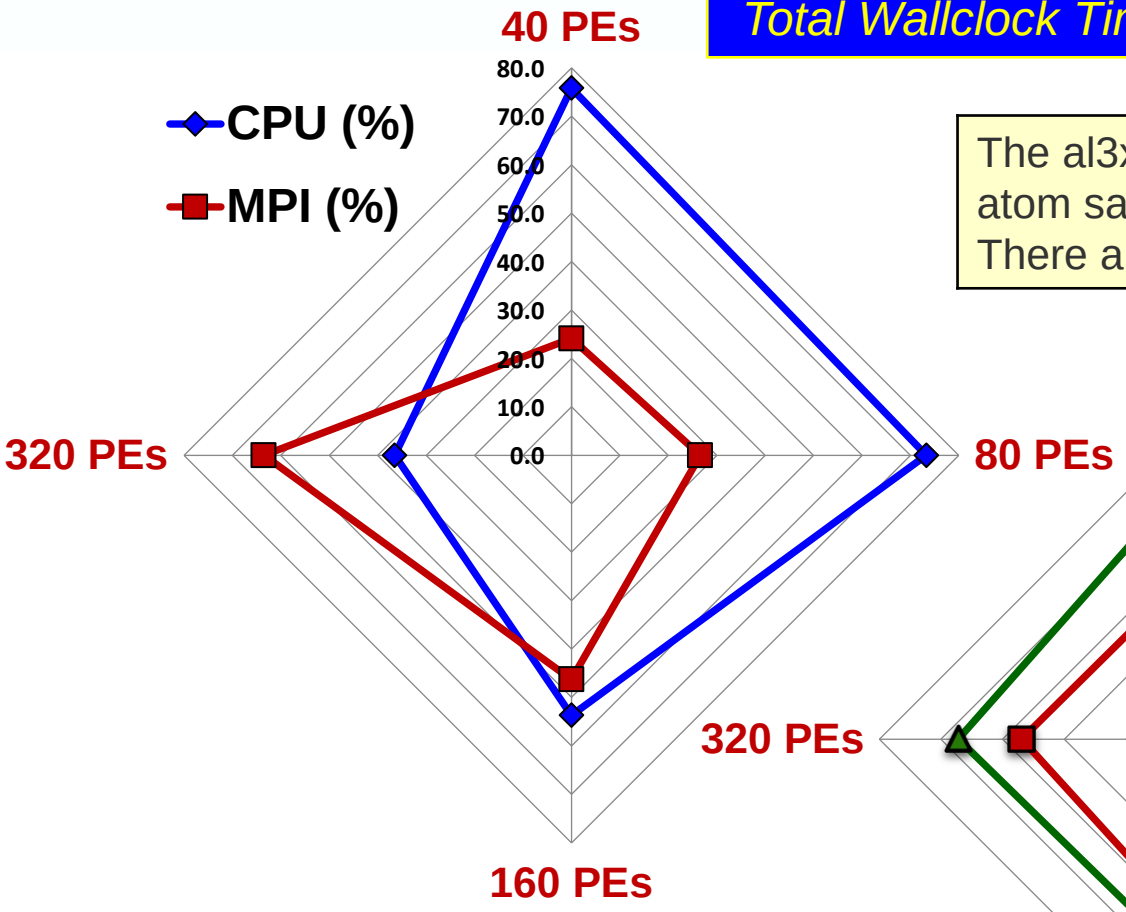
  Bigger calculation (313 electrons and 64 ions) and involves MPI AllToAllV across all processors.

- **IDZ Benchmark**

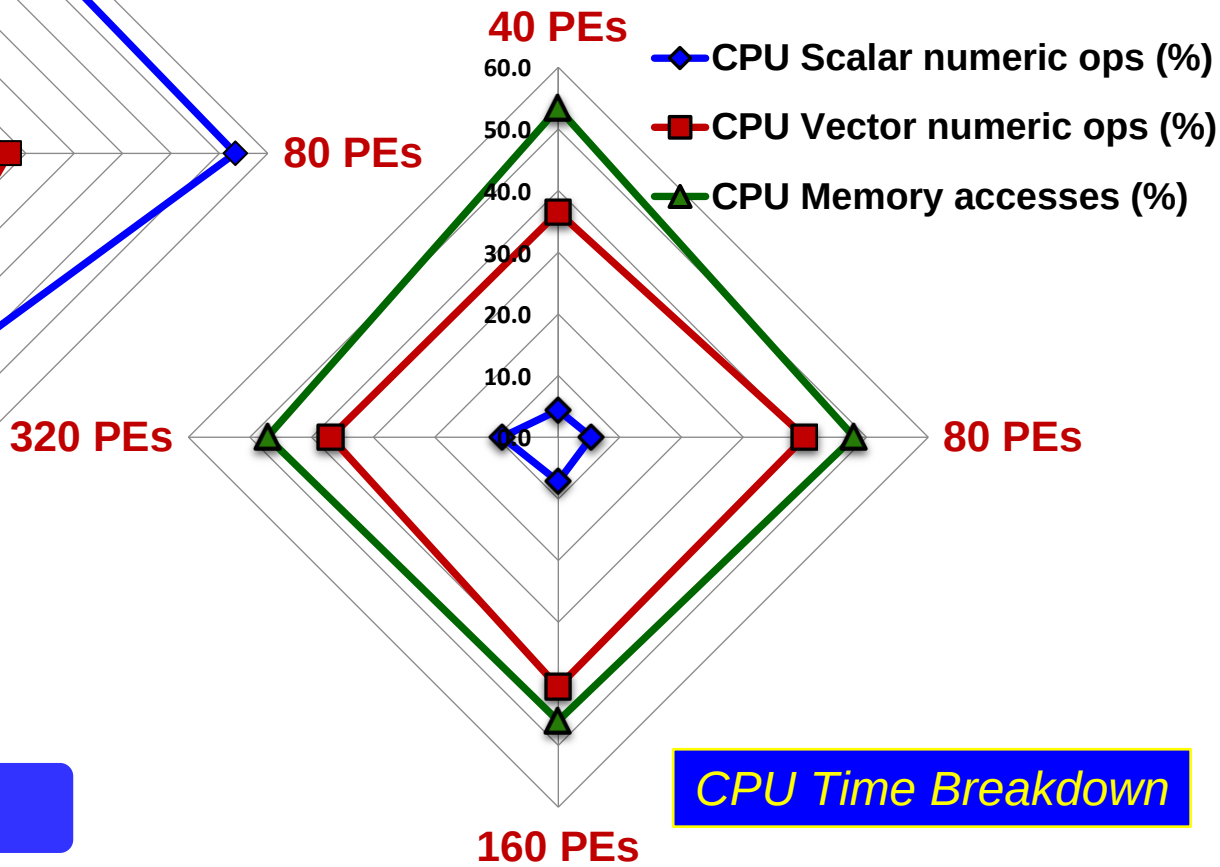
  Longer MD calculation (1104 electrons and 404 ions) requiring several random initializations (16 MD iterations in total).

CASTEP 21 – al3x3 Benchmark Performance Report

Total Wallclock Time Breakdown



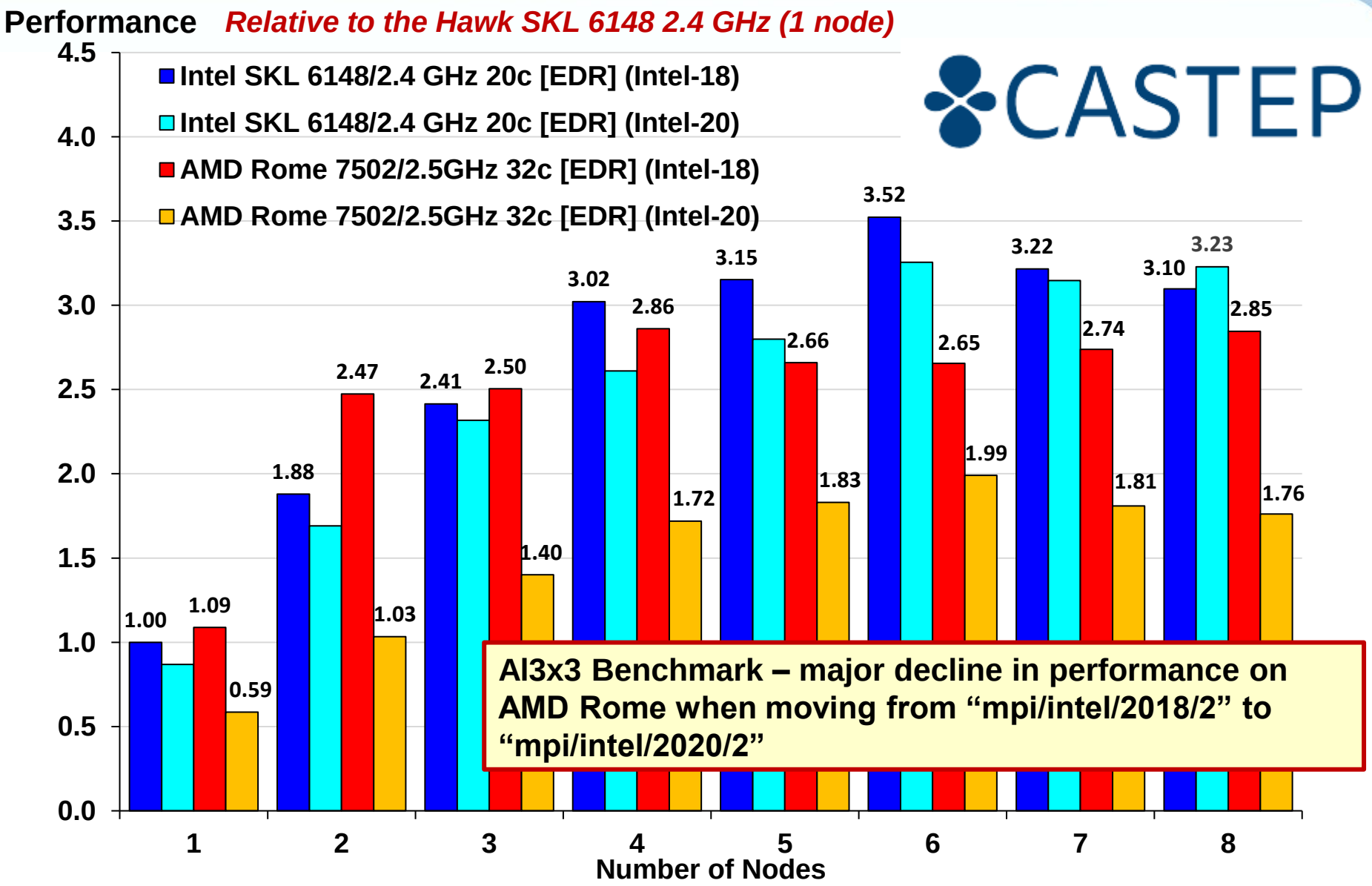
The al3x3 simulation cell comprises a 270-atom sapphire surface, with a vacuum gap. There are only 2 k-points.



CPU Time Breakdown

Performance Data (40-320 PEs)

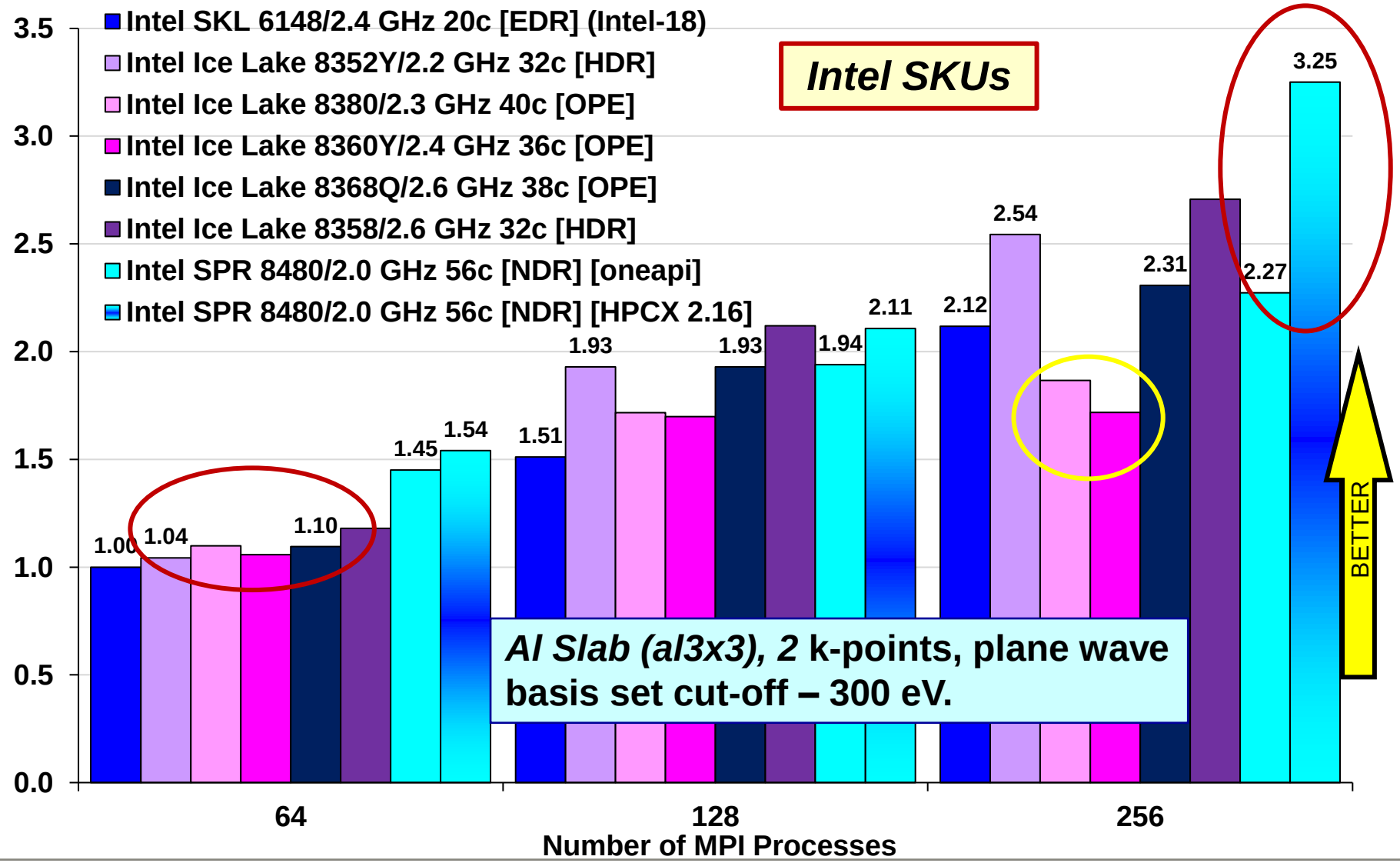
CASTEP – Impact of Intel MPI version on AMD clusters



CASTEP 19 – Al Slab (al3x3) Benchmark

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

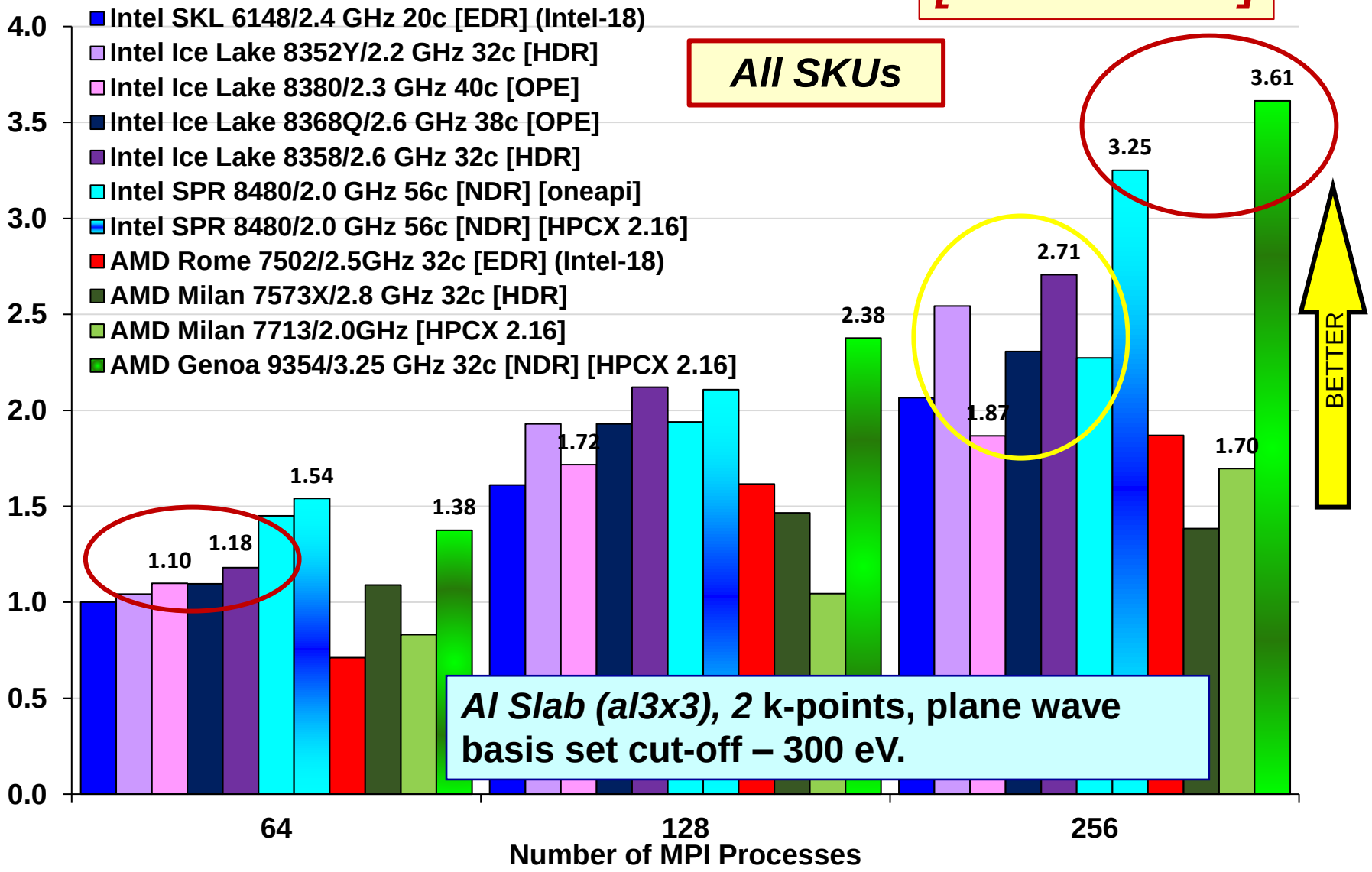
[Core to core]



CASTEP 19 – Al Slab (al3x3) Benchmark

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

[Core to core]

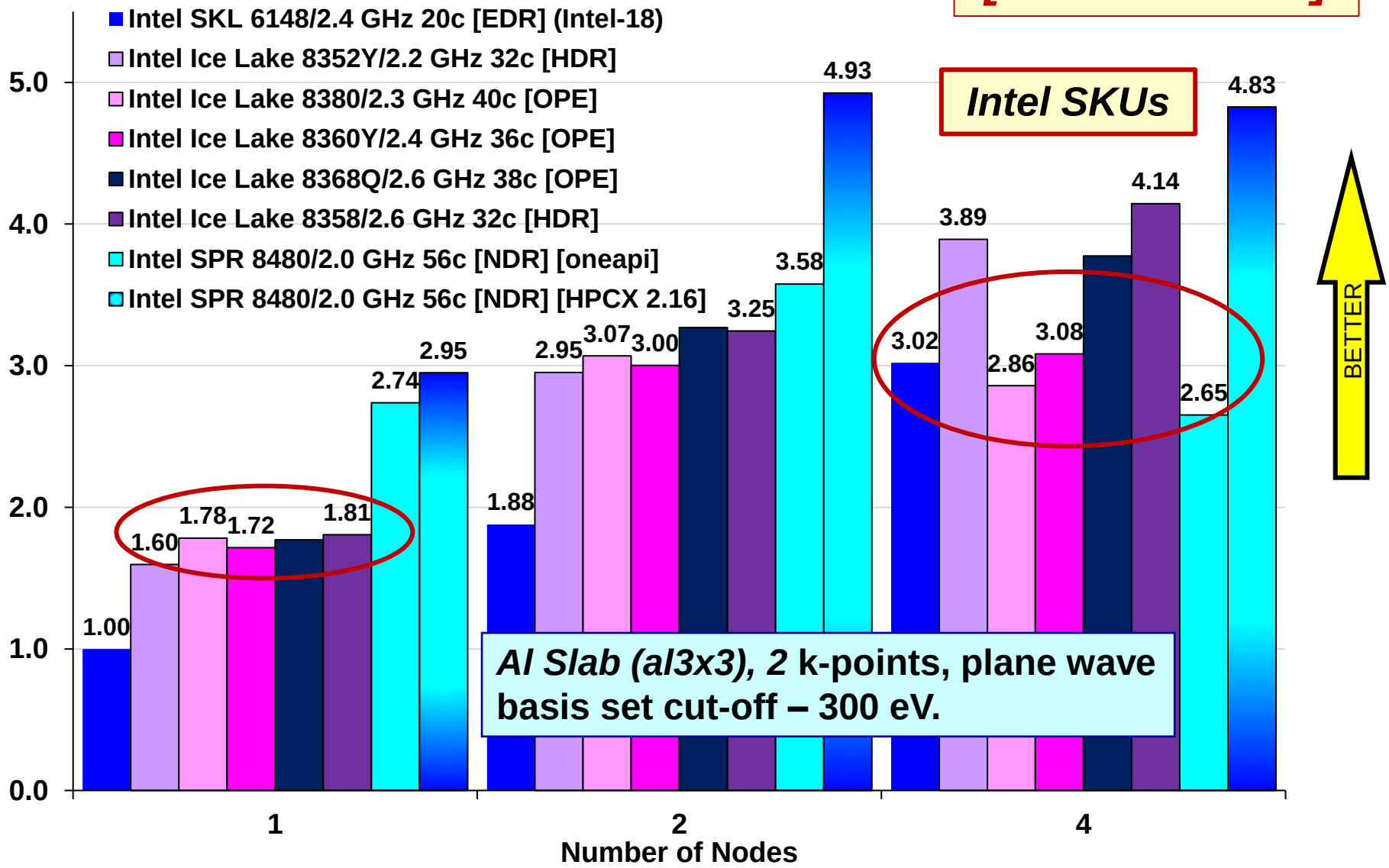


Al Slab (al3x3), 2 **k-points, plane wave basis set cut-off – 300 eV.**

CASTEP 19 – Al Slab (al3x3) Benchmark

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

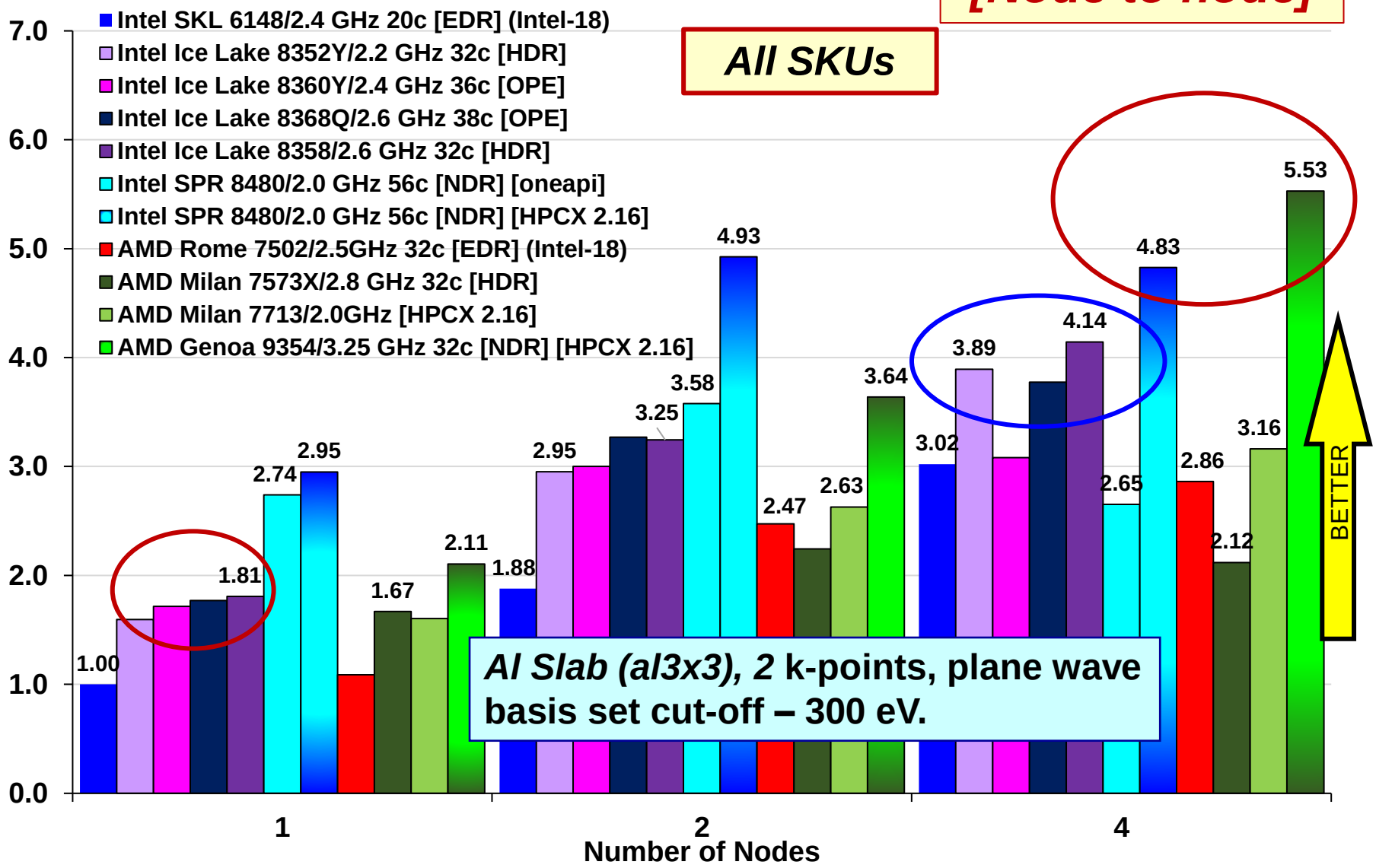
[Node to node]



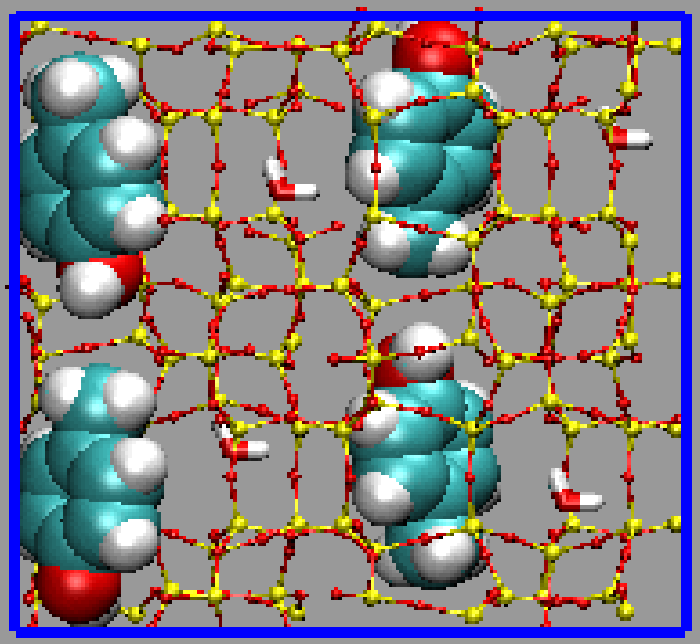
CASTEP 19 – Al Slab (al3x3) Benchmark

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

[Node to node]



Performance of Computational Chemistry and Ocean Modelling Codes

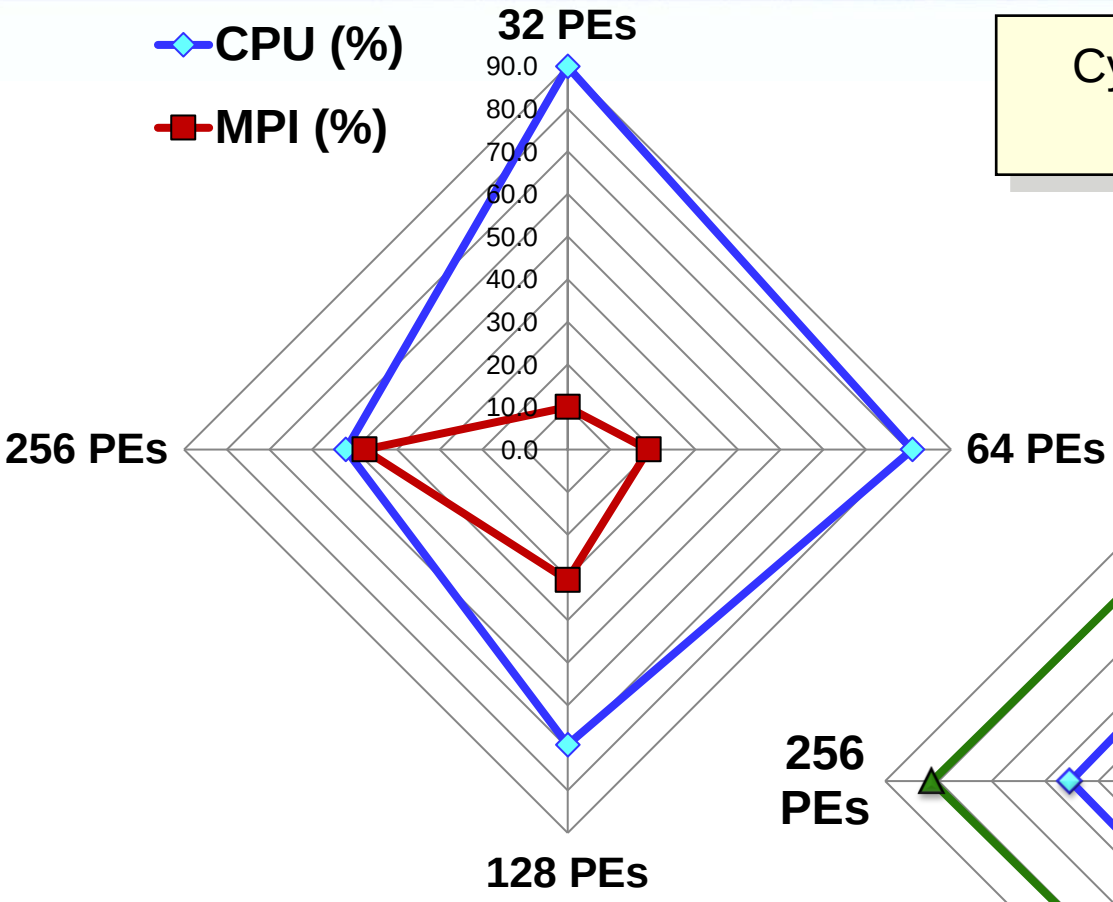


Electronic Structure

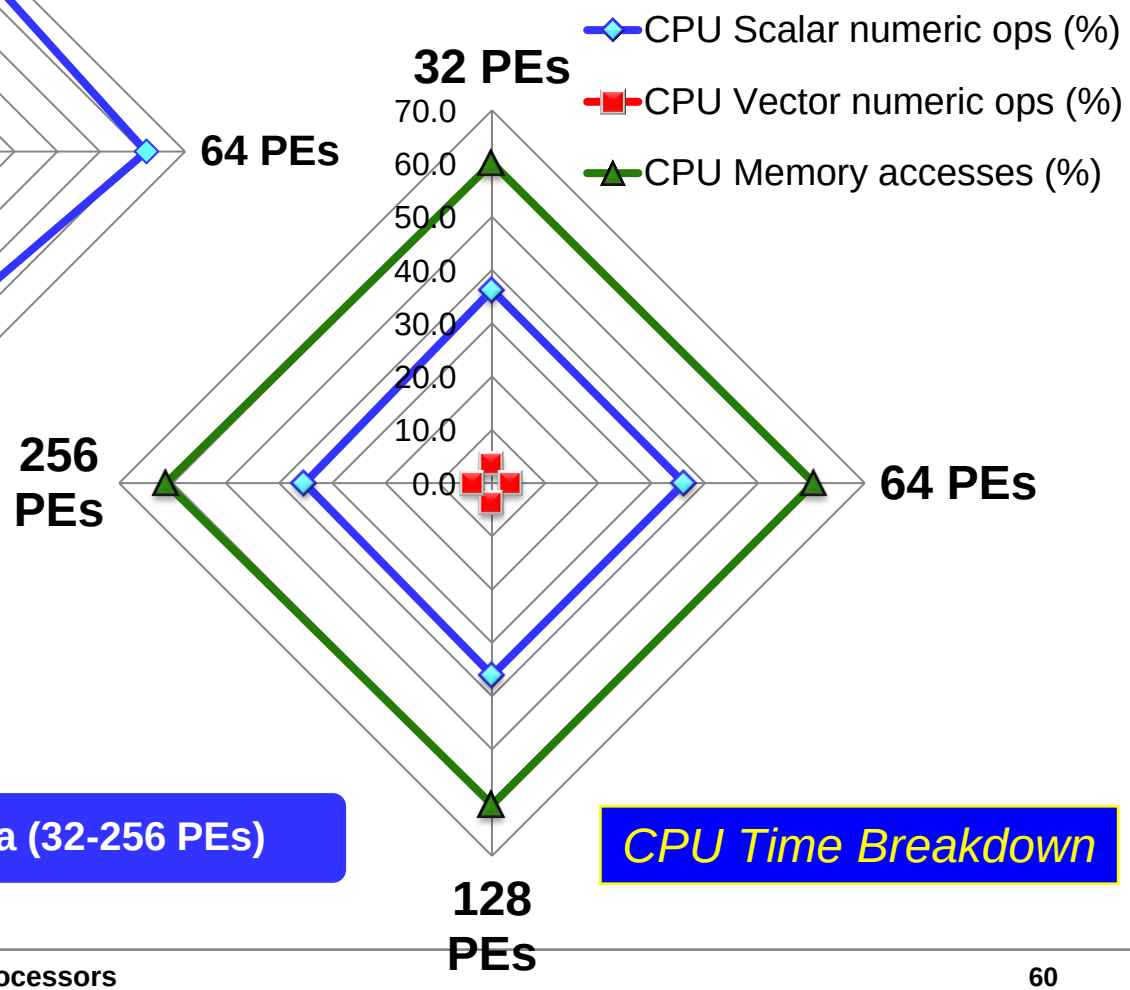
GAMESS -UK

GAMESS-UK.MPI DFT – DFT Performance Report

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



Total Wallclock Time Breakdown



Performance Data (32-256 PEs)

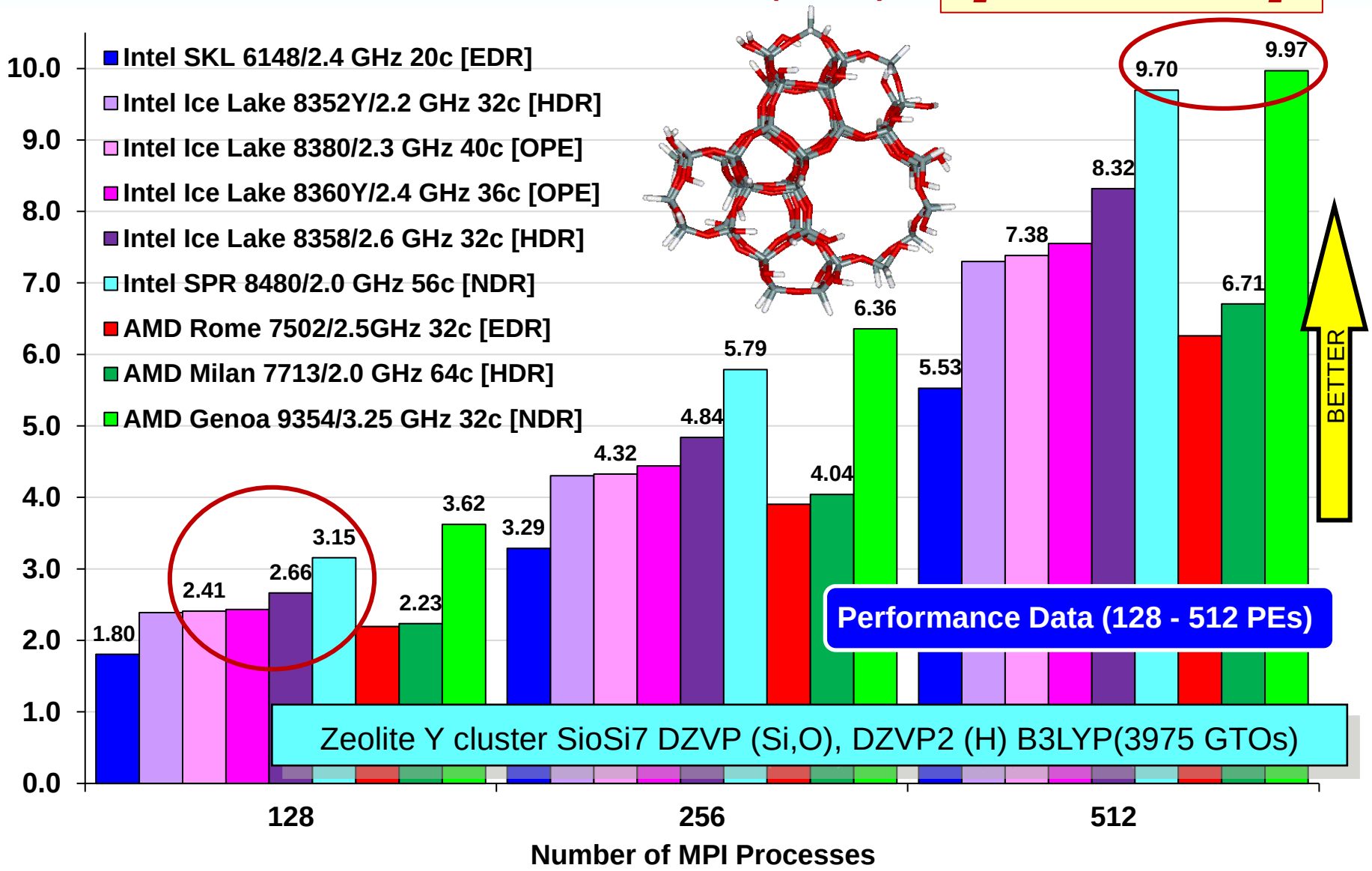
CPU Time Breakdown

GAMESS-UK Performance - Zeolite Y cluster

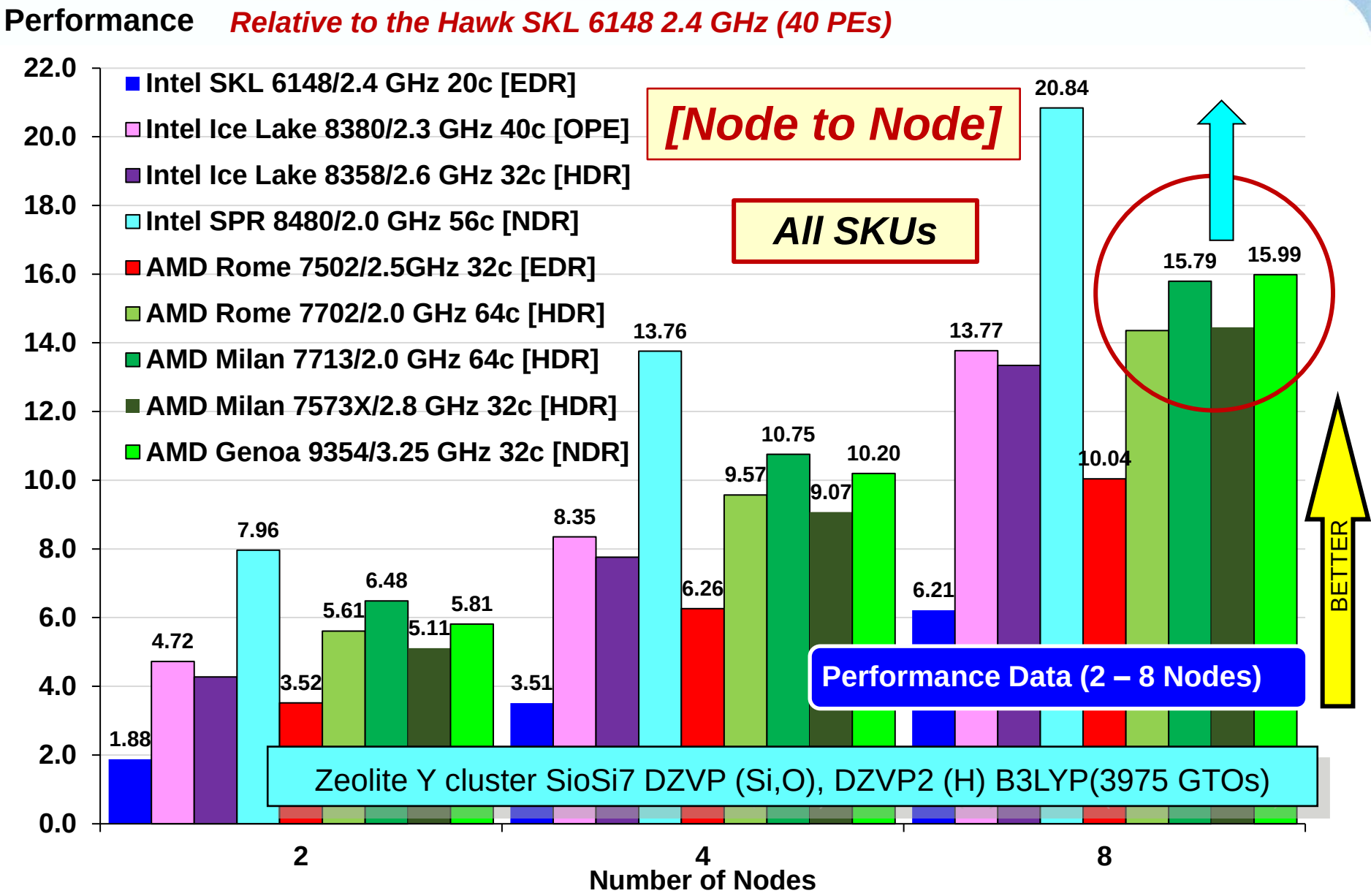
CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

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

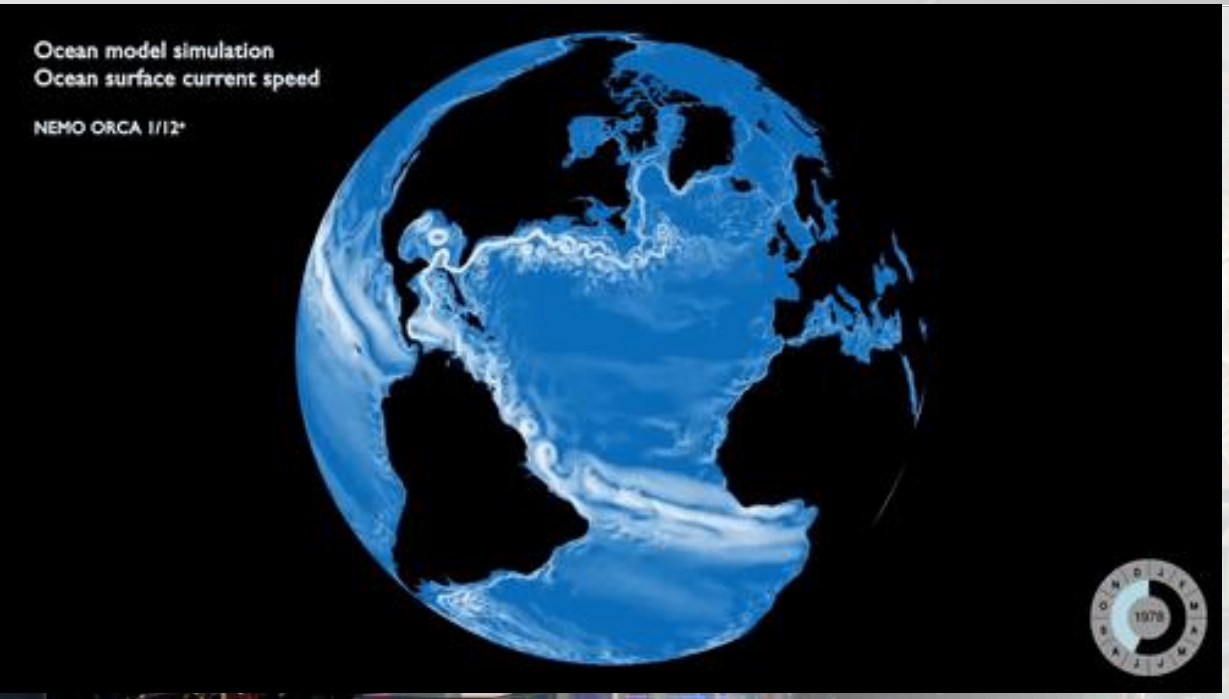
[Core to core]



GAMESS-UK Performance - Zeolite Y cluster



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

The NEMO-ERSEM Benchmark

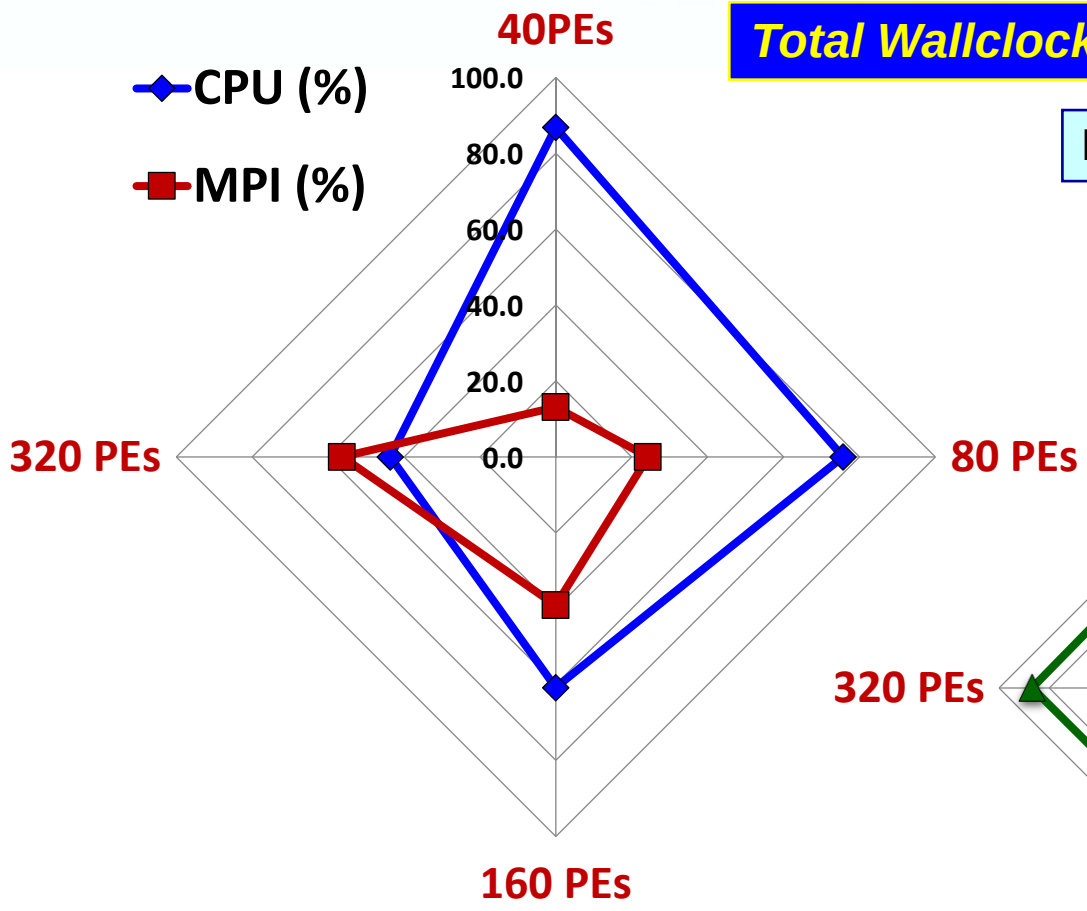
- NEMO, "Nucleus for European Modelling of the Ocean" is a modelling framework for research activities and forecasting services in ocean and climate sciences, developed by a European consortium. (https://www.nemo-ocean.eu)
- NEMO is a **memory-bandwidth limited code** where performance can be improved 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 (Atlantic Margin Model) 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:

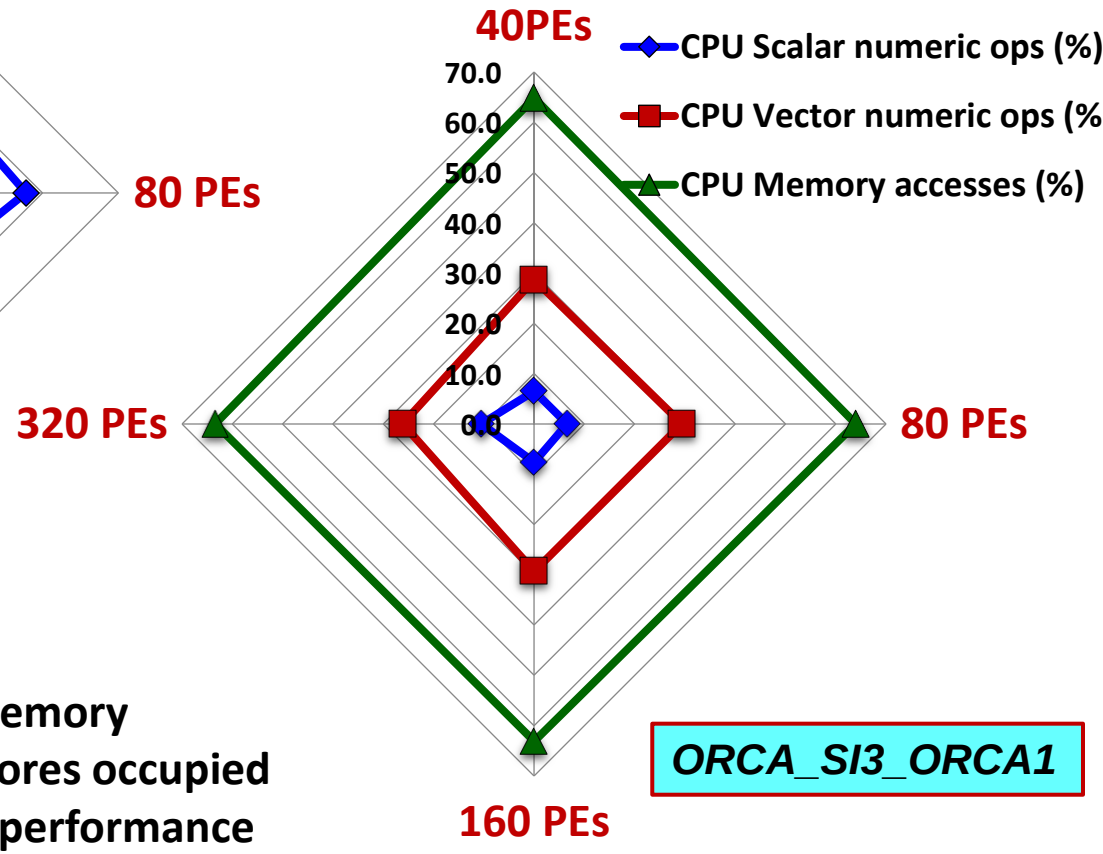
```
mpirun -n $XIOSCORES $code_xios : -n $OCEANCORES $code_nemo
```

NEMO – ORCA_SI3 Model Performance Report

Total Wallclock Time Breakdown

horizontal resolutions of 1-degree



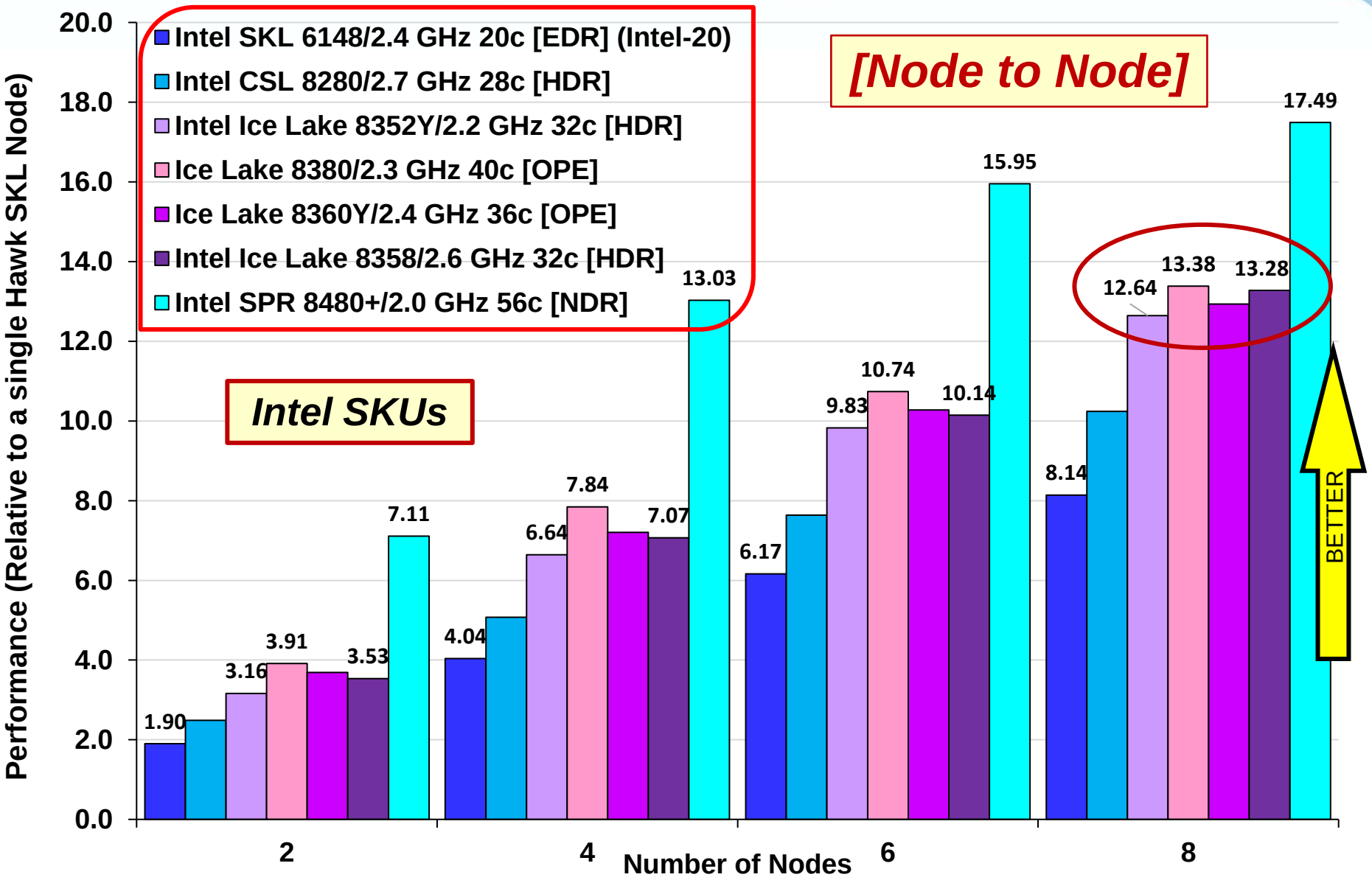


ORCA_SI3_ORCA1

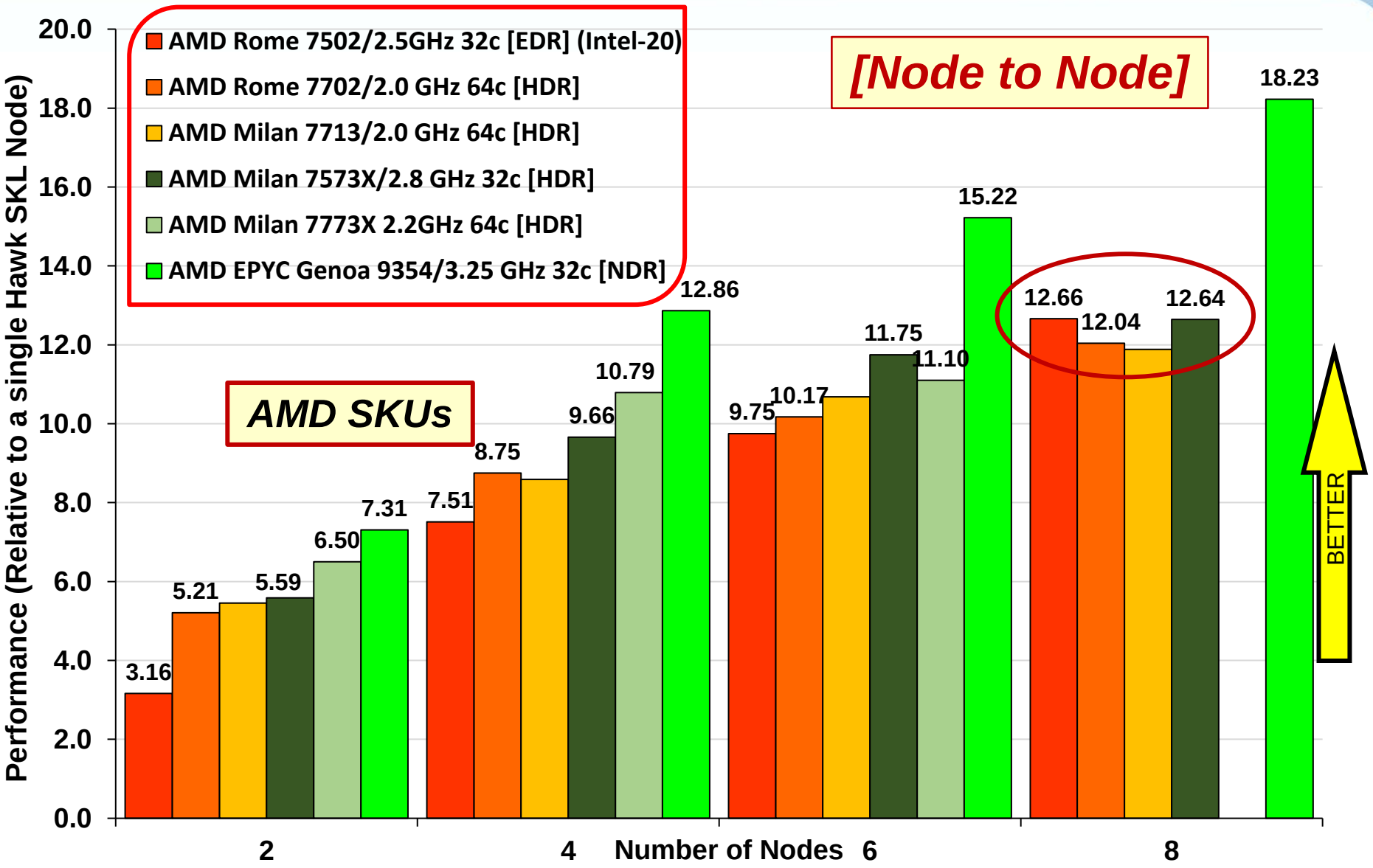
CPU Time Breakdown

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.

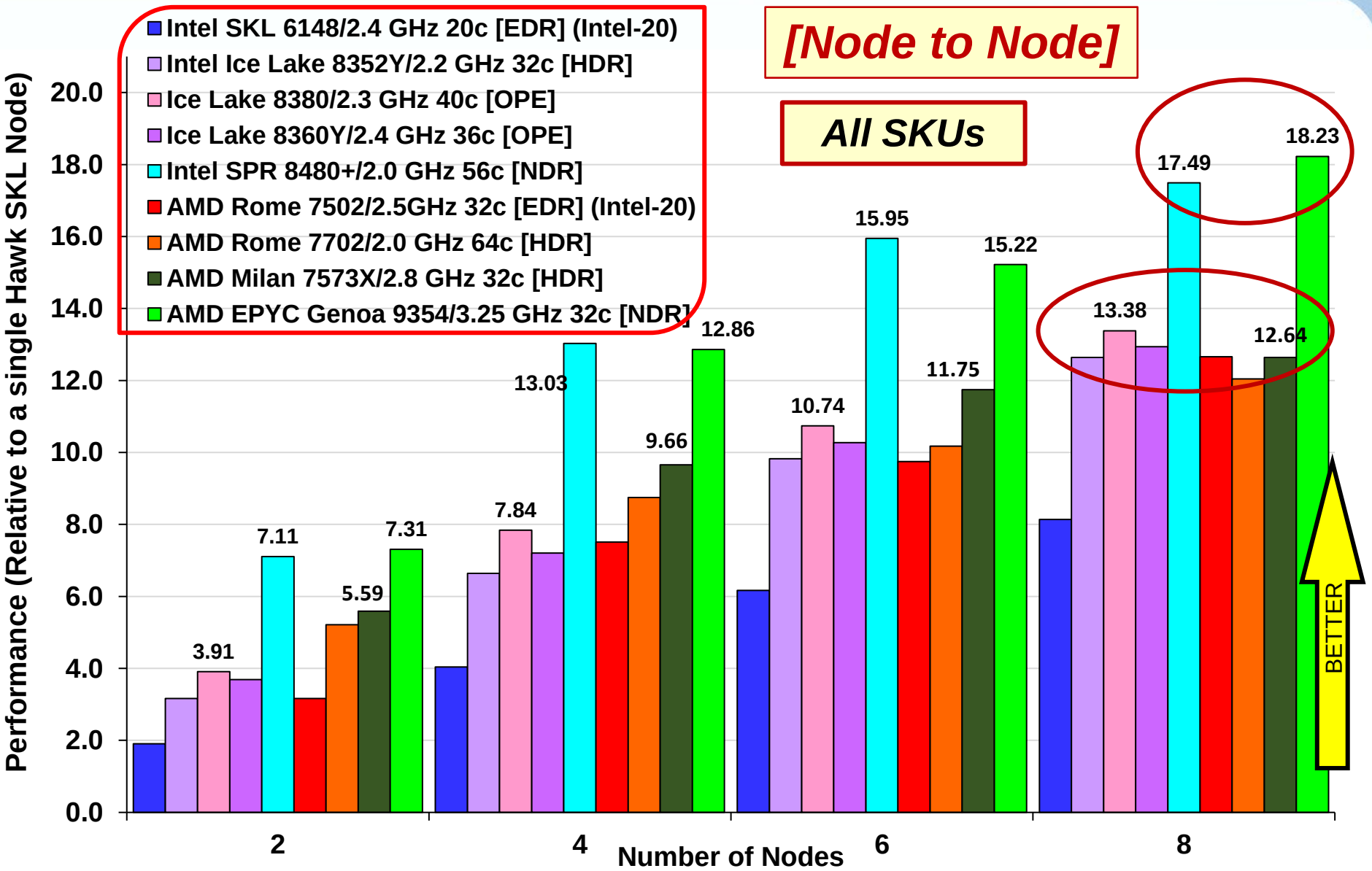
NEMO-FABM-ERSEM (AMM7) – Node Performance



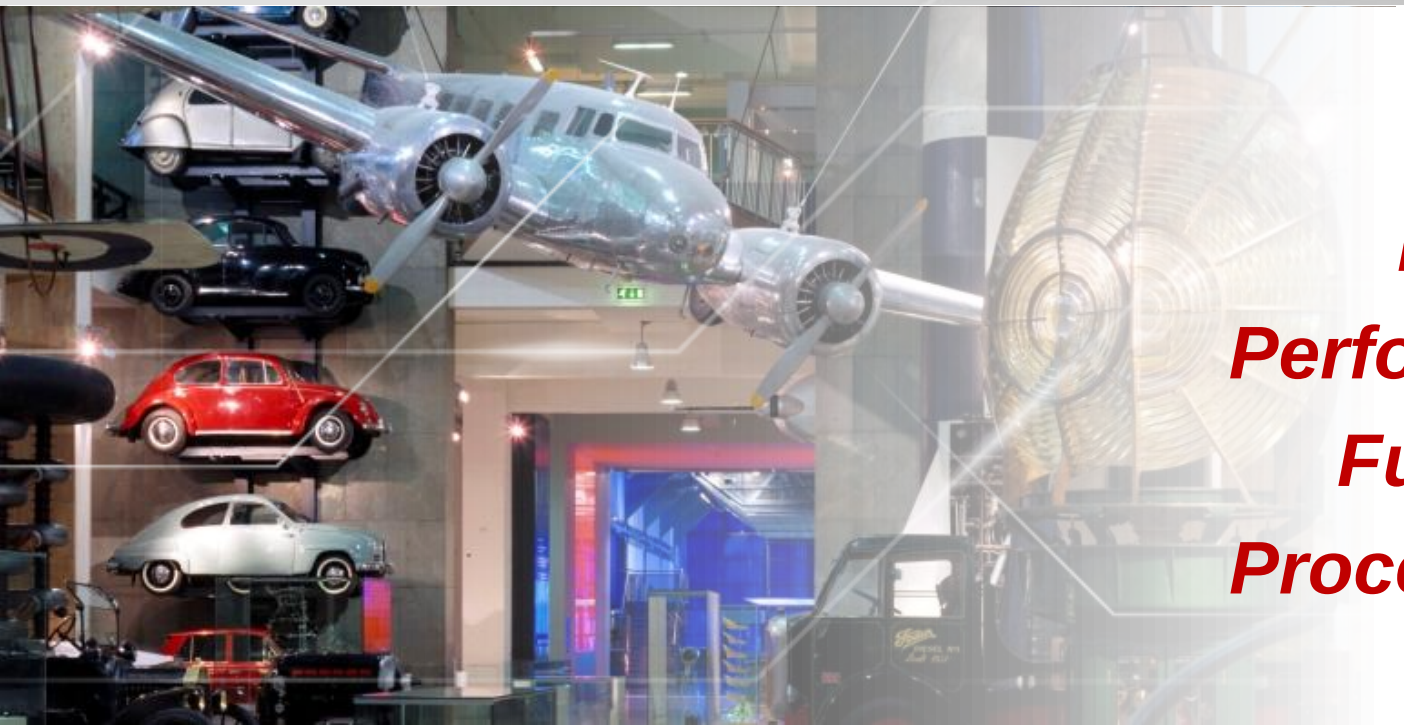
NEMO-FABM-ERSEM (AMM7) – Node Performance



NEMO-FABM-ERSEM (AMM7) – Node Performance

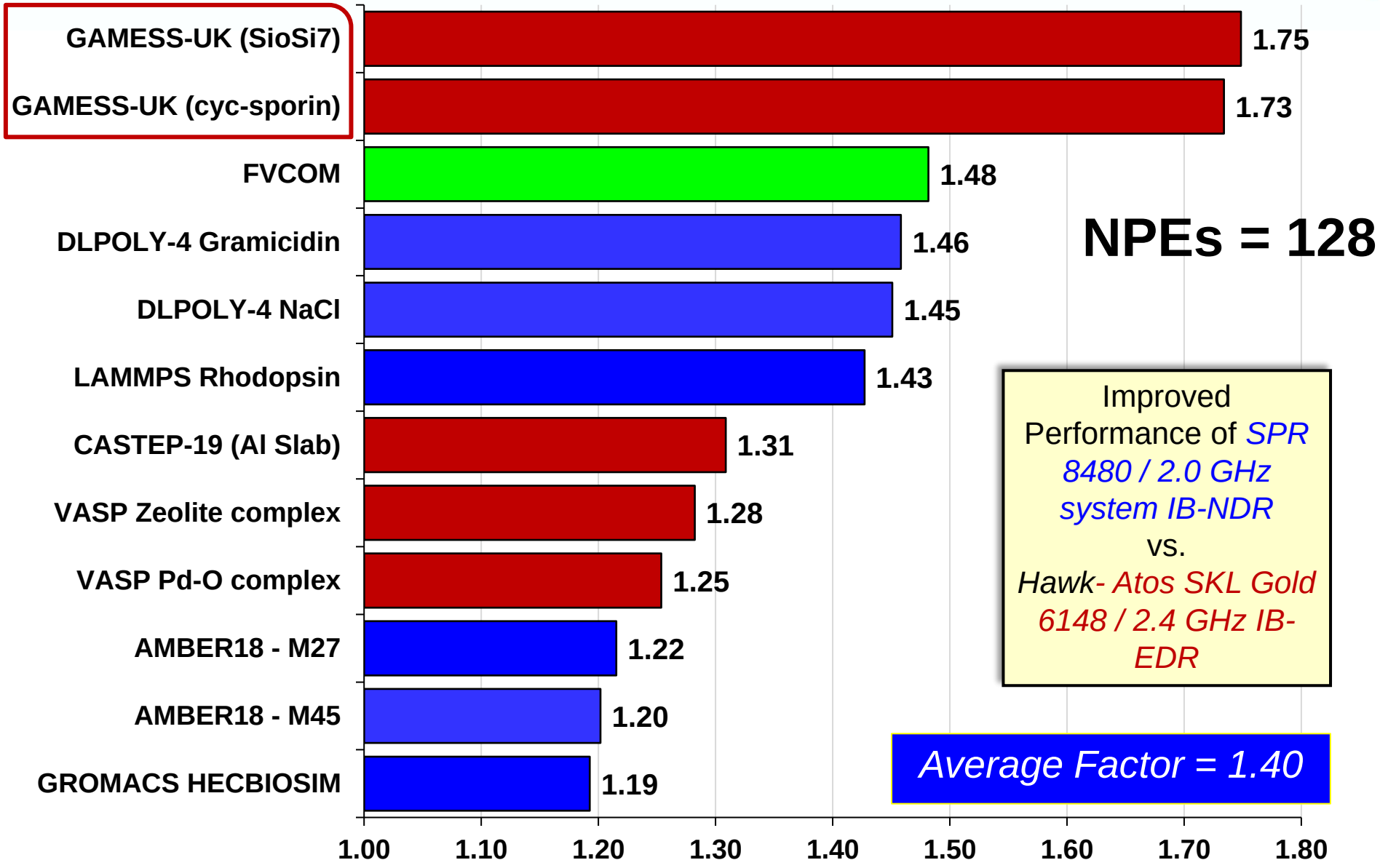


Performance of Computational Chemistry and Ocean Modelling Codes

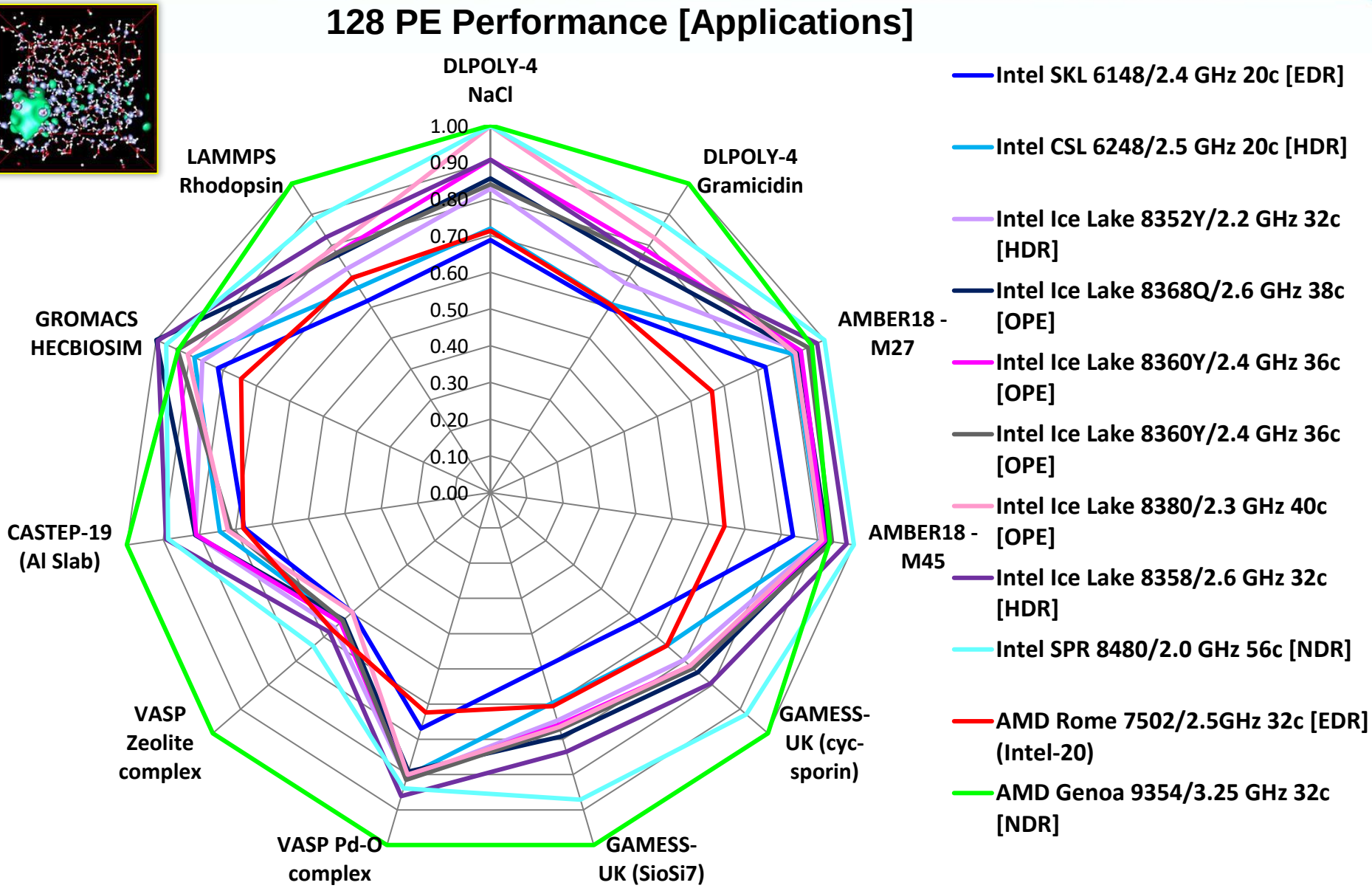


Relative Performance as a Function of Processor Family

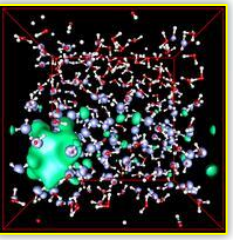
Sapphire Rapids 8480 2.0 GHz NDR vs. SKL 6148 2.4 GHz EDR

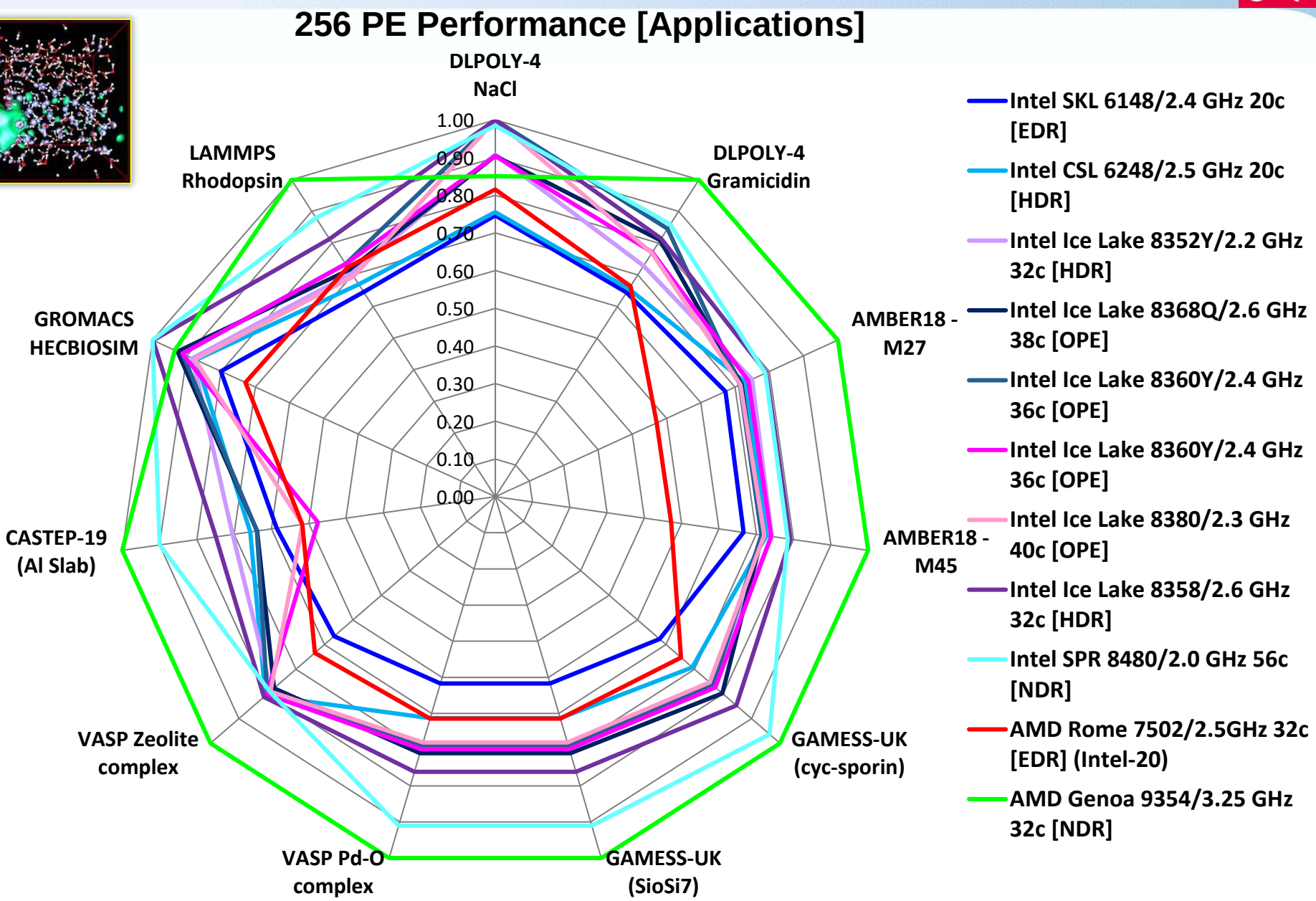


Target Codes and Data Sets – 128 PEs

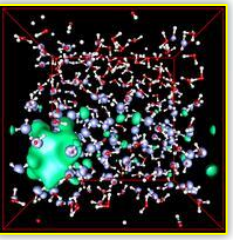


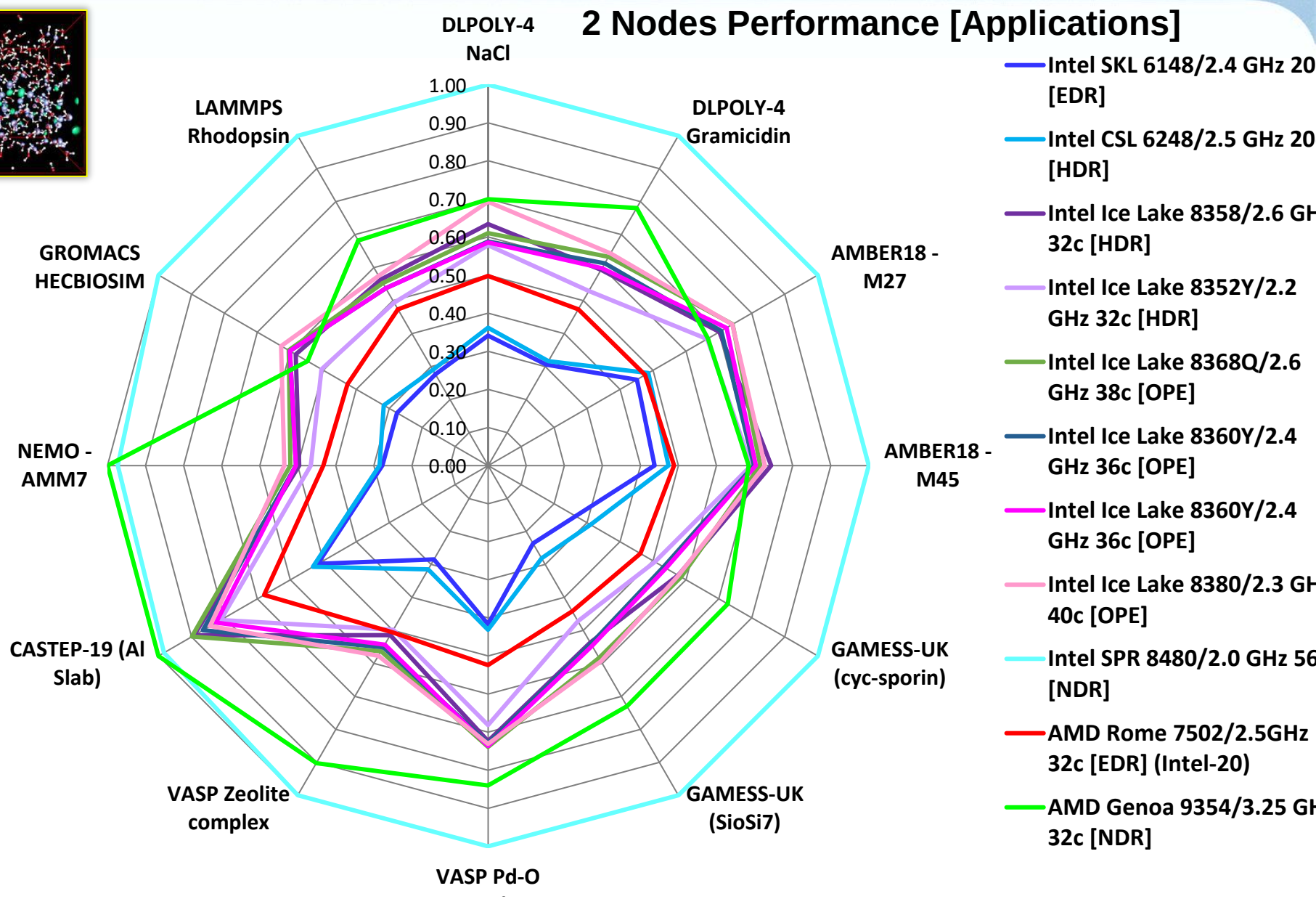
Target Codes and Data Sets – 256 PEs



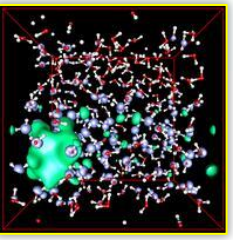


Target Codes and Data Sets – 2 Nodes

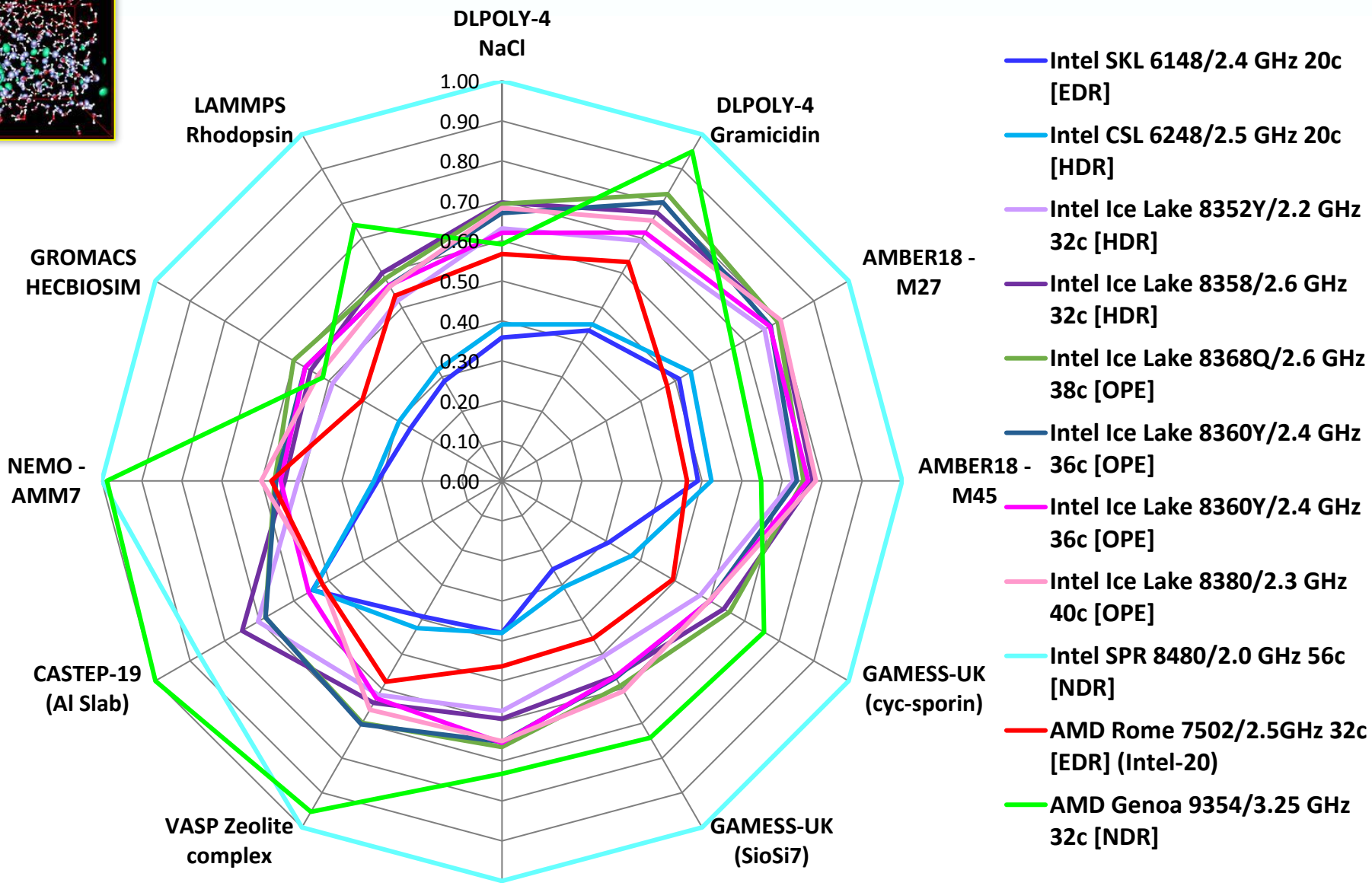




Target Codes and Data Sets – 4 Nodes



4 Nodes Performance [Applications]



Conclusions – Core-to-Core Comparisons

- ***Core-to-Core* comparisons** suggests that the **AMD Genoa 9354 32c 3.25 GHz** outperforms the **Intel SPR 8480 2.0 GHz SKU** in **most cases**, **The exceptions being the** Gromacs 1.4M atom HECBIOSIM & DLPOLY4 NaCl simulations**.**
- The **Intel SPR 8480 2.0 GHz SKU** outperforms all other Intel SKUs (cf. CASTEP), with relative performance sensitive to use of AVX instructions. Low utilisation of AVX-512 leads to weaker performance of the SKL, CSL & Ice Lake CPUs and **better performance of the AMD Milan-based clusters** e.g. DLPOLY, GAMESS-UK, LAMMPS.
- Superior performance of **AMD Genoa 9354** compared to their Milan predecessors.
- Major performance improvement of CASTEP when using the **HPC-X MPI** library on **both Intel and AMD clusters**.
- With **significant AVX-512 utilisation**, Intel **Ice Lake systems** outperform the **AMD Milan systems** e.g.. Gromacs. **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 Ice Lake SKUs used in this study.
- Baselined in part across the **V100** NVIDIA GPU performance.

Conclusions – Node-to-Node Comparisons

- Given superior core performance, a ***Node-to-Node comparison*** typical of the performance when running a workload shows the **SPR 8480** delivering **far superior performance** compared to (i) the SKL Gold 6148 (112 cores vs. 40 cores). Average improvements factors of **3.2** (2-node) and **2.8** (4-nodes) across all applications.
- A ***Node-to-Node comparison*** *shows the* ***SPR 8480*** delivering on average ***superior performance*** *compared to* ***the*** **AMD Genoa 9354 32c** (112 cores vs. 64 cores – **1.75**) of **1.30** (2-nodes) and **1.25** (4-nodes). The **NEMO-AMM7** and **CASTEP-19 (Al-slab)** position the Genoa 9354 ahead.
- Performance of the **AMD Milan 7713, 7763 and 7773X** **(128 core nodes)** is disappointing.
- In contrast to the core-to-core comparisons, the higher core count **Ice Lake systems** – **38c 8368Q & 40c 8380** – now perform on a par with the **32c 8358**.
- Relative to the Ice Lake systems, 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.

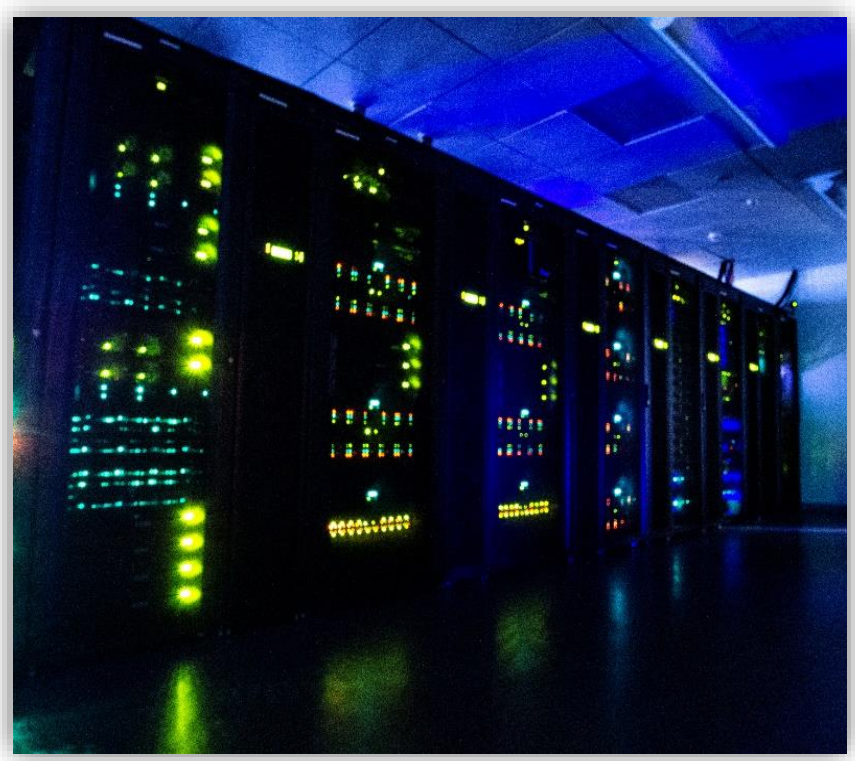
Acknowledgements

- **Joseph Stanfield, Joshua Weage,** Martin Hilgeman, Benoit Lodej, Mark Mendez & Dave Coughlin for access to, and assistance with, the variety of AMD EPYC & Intel Xeon 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*** for 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 on systems featuring **processors from Intel** (Sapphire Rapids & Ice Lake SKUs) and **AMD** (Genoa & Milan SKUs) with IB & Cornelis Networks.
  - Baseline clusters: Skylake (SKL) **Gold 6148/2.4 GHz** and **AMD EPYC Rome 7502 2.5Gz** cluster – "Hawk" – at Cardiff University.
  - **Two** Intel Sapphire Rapids clusters – the 56-core Platinum 8480 and Platinum HBM 9480 plus **five** Intel Xeon Ice Lake clusters, and their Cascade Lake & Cascade Lake-AP counterparts.
  - **Four** AMD EPYC Milan clusters featuring the 64-core **7713** & **7773X** and the 32-core **7543** & **7573X**. **Two** AMD Genoa systems, the 9354 & 9454.
- **Performance** of both synthetic and **end-user applications,** including molecular simulation (**DL_POLY, AMBER, LAMMPS & GROMACS MD codes**), materials modelling (**CASTEP, VASP**), & electronic structure (**GAMESS-UK**), plus the **NEMO** and **FVCOM** ocean modelling codes.
- **Scalability analysis** by **processing elements (cores)** and by **nodes** (ARM Performance Reports). Baselined against **V100** NVIDIA GPUs.
- **Pricing** – remains of course a key issue but lies outside the scope of this presentation.

Any Questions?



Martyn Guest *GuestMF@Cardiff.ac.uk*

Jose Munoz Criollo *MunozCriolloJJ@cardiff.ac.uk*

Thomas Green *Greent10@cardiff.ac.uk*