CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

COMPUTING INSIGHT UK 2018
Research Computing Delivers
12 - 13 DECEMBER 2018
MANCHESTER, UK
Science & Technology Facilities Council
UK Research and Innovation

Application Performance on Multi-core processors

Scaling, Throughput and an Historical perspective

M.F. Guest[≠], C.A. Kitchen[≠], M. Foster[†] and D. Cho[§]

[≠] *Cardiff University,* [†]*Atos,* [§] *Mellanox Technologies*

Outline

I. **Performance Benchmarks** and **Cluster Systems**

   a. **Synthetic Code Performance**: *STREAM and IMB*

   b. **Application Code Performance**: *DLPOLY, GROMACS, AMBER,GAMESS_UK, VASP and Quantum Espresso*

   c. **Interconnect Performance**: Intel MPI and Mellanox's HPCX

   d. ***Processor Family and Interconnect*** *– "core to core" and "node to node" benchmarks*

II. **Impact of Environmental Issues *in Cluster Acceptance tests***

   a. **Security patches, turbo mode and Throughput testing**

III. **Performance profile** of **DL_POLY** and **GAMESS-UK** over the **past two decades**

IV. **Acknowledgements and Summary**

Contents

I. Review of parallel application performance featuring synthetics and end-user applications across a variety of clusters

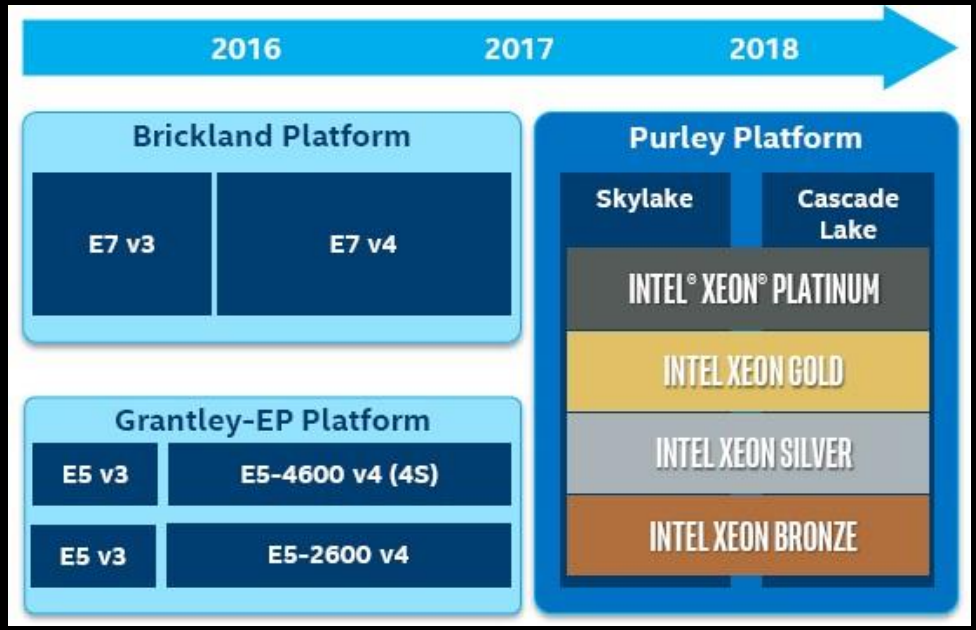
- ¤ End-user Codes – **DL_POLY, GROMACS**, **AMBER,** NAMD, LAMMPS, **GAMESS-UK, Quantum Espresso, VASP**, CP2K, ONETEP & OpenFOAM
- Ongoing Focus on Intel's Xeon Scalable processors ("Skylake"), AMD's Naples EPYC processor plus nVIDIA GPUs, including
- ¤ Clusters with dual-socket nodes - Intel **Xeon Gold 6148** Processor (20c, 27.5M Cache, 2.40 GHz) & **Xeon Gold 6138** Processor (20c, 27.5M Cache, 2.00 GHz) + AMD Naples **EPYC 7551** (2.00 GHz) & **EPYC 7601** (2.20 GHz) CPUs.
- ¤ Updated review of Intel MPI and Mellanox HPCX performance analysis .

II. How these benchmarks have been deployed in the framework of procurement and ***acceptance testing***, dealing with a variety of issues e.g. (a) **security patches, turbo mode etc**. & (ii) **Throughput testing**.

III. An historical perspective of two of these codes – **DL_POLY and GAMESS-UK** – and briefly overview the development and **performance profile** of both over the **past two decades**.

The Xeon Skylake Architecture

- The **architecture of Skylake** is very different from that of the prior "Haswell" and "Broadwell" Xeon chips
- **Three basic variants** that now cover what was formerly the Xeon E5 and Xeon E7 product lines, with Intel **converging the Xeon E5 and E7 chips** into a single socket.



- Product segmentation – **Platinum, Gold, Silver, & Bronze** – with 51 variants of the SP chip
- Also custom versions requested by hyperscale and OEM customers.
- All of these chips differ from each other in a number of ways, including **number of cores, clock speed, L3 cache capacity, number and speed of UltraPath links between sockets, number of sockets supported, main memory capacity, width of the AVX vector units** etc.

Intel Xeon : Westmere - Skylake

	Xeon 5600 (Westmere-EP)	Xeon E5-2600 (Sandy Bridge-EP)	Xeon E5-2600 v4 "Broadwell-EP"	Intel Xeon Scalable Processor "Skylake"
Cores / Threads	**Up to 6 cores / 12 threads**	**Up to 8 cores / 16 threads**	**Up to 22 Cores / 44 threads**	**Up to 28 Cores / 56 threads**
Last-level cache	12 MB	Up to 20 MB	Up to 55 MB	**Up to 38.5 MB (non-inclusive)**
Max memory channels, speed / socket	**3xDDR3 channels, 1333**	**4xDDR3 channels, 1600**	**4 channels of up to 3 RDIMMs, LRDIMMs or 3DS LRDIMMs, 2400 MHz**	**6 channels of up to 2 RDIMMs, LRDIMMs or 3DS LRDIMMs, 2666 MHz**
New instructions	AES-NI	**AVX 1.0 8 DP Flops/Clock**	**AVX 2.0 16 DP Flops/Clock**	**AVX 512 32 DP Flops/Clock**
QPI / UPI Speed (GT/s)	1 QPI channels @ 6.4 GT/s	2 QPI channels @ 8.0 GT/s	2 x QPI channels @ 9.6 GT/s	**Up to 3 x UPI @ 10.4 GT/s**
PCIe Lanes / Controllers / Speed (GT/s)	36 lanes PCIe 2.0 on chipset	40 Lanes / Socket Integrated PCIe 3.0	40 / 10 / PCIe* 3.0 (2.5, 5, 8 GT/s)	**48 / 12 / PCIe* 3.0 (2.5, 5, 8 GT/s)**
Server / Workstation TDP	Server / Workstation: 130W	Up to 130W Server; 150W Workstation	55 - 145W	**70 – 205W**

AMD® Epyc™ 7000 Series - SKU Map and FLOP/cycle

SKU	7601	7551	7501	7451	7401	7351	7301
Freq (base)	2.2	2.0	2.0	2.3	2.0	2.4	2.2
Turboboost All cores active	2.7	2.6	**2.6**	**2.9**	**2.8**	**2.9**	**2.7**
Turboboost On core active	3.2	3.0	3.0	3.2	3.0	2.9	2.7
Cores/socket	32	32	32	24	24	16	16
L3 cache size	64 MB						
Memory Channel	8						
Memory Freq	2667 MT/s						
TDP (W)	180	180	155/170	180	155/170	155/170	155/170

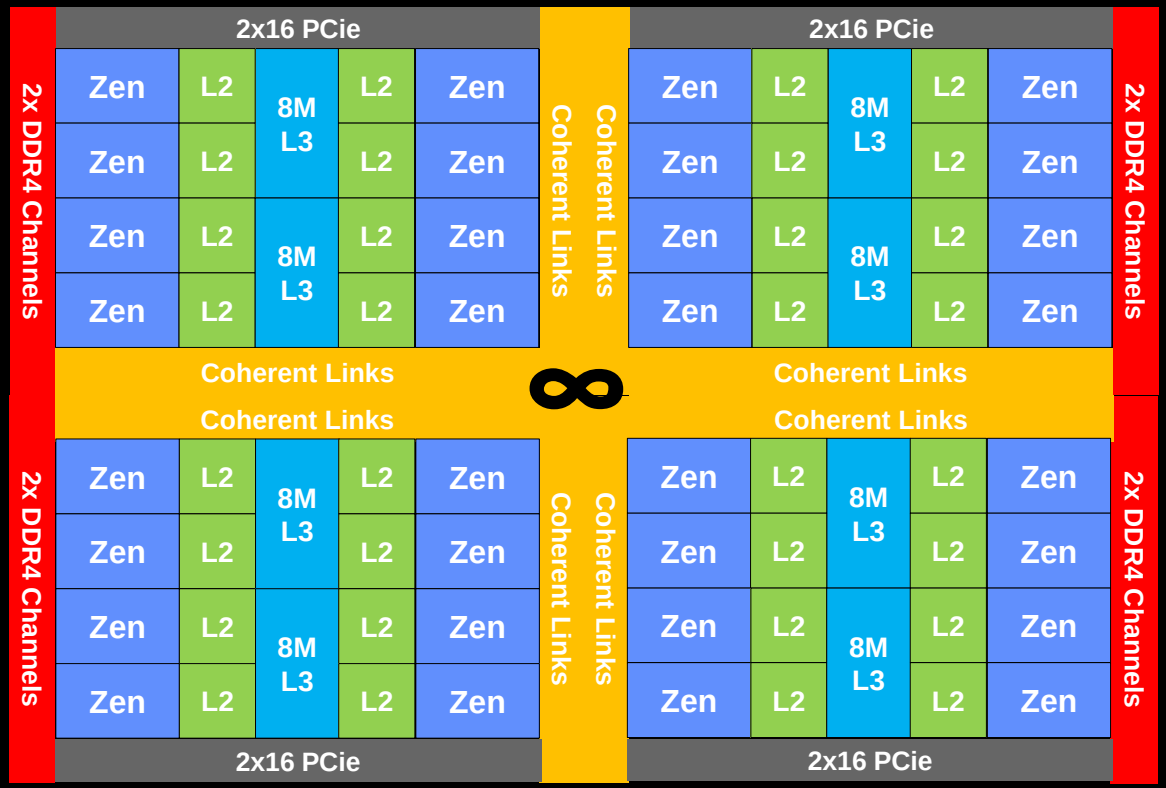
Architecture	Sandy Bridge	Haswell	Skylake	EPYC
ISA*	AVX	AVX2	AVX-512	**AVX2**
op/cycle	2 (1 ADD, 1 MUL)	4 (2 FMA)	4 (2 FMA)	**4 (2 ADD, 2 MUL)**
Vector size (DP = 64-bits)	4	4	8	**2**
FLOP/cycle	8	16	**32**	**8**

The AMD EPYC only supports 2 × 128-bit AVX natively, so there's a large gap with Intel SKL and their 2 × 512-bit FMAs.

Thus the FP peak on AMD is 4 × lower than on Intel SKL.

*** Instruction Set Architecture**

EPYC Architecture - Naples, Zeppelin & CCX



- **Zen cores**
  - Private L1/L2 cache
- **CCX**
  - 4 ZEN cores (or less)
  - 8MB L3 shared cache
- **Zeppelin**
  - 2 CCX (or less)
  - 2 DDR4 channels
  - 2 PCIe 16x
- **Naples**
  - 4 Zeppelin SoC dies fully connected by Infinity Fabric.
  - **4 Numa Nodes !**

- Delivers 32 cores / 64 threads, 16MB L2 cache and 64MB L3 cache per socket.
- Design also means that there are **four NUMA nodes per socket or eight NUMA nodes in a dual socket system** i.e. different memory latencies depending on which die needs data from memory that can be attached to that die or another die on the fabric.
- The key difference with Intel's Skylake SP architecture is that AMD needs to go off die within the same socket where Intel stays on a **single piece of silicon.**

Intel Skylake and AMD EPYC Cluster Systems

Cluster / Configuration
"Hawk" – Supercomputing Wales cluster at Cardiff comprising 201 nodes, totalling 8,040 cores, 46.080 TB total memory • **CPU: 2 x Intel(R) Xeon(R) 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.**
"Helios" – 32 node HPC Advisory Council cluster running SLURM: Mellanox ConnectX-5 • **Supermicro SYS-6029U-TR4 /** Foxconn Groot 1A42USF00-600-G **32-node cluster; Dual Socket Intel(R) Xeon(R) Gold 6138 CPU @ 2.00GHz** • **Mellanox ConnectX-5 EDR 100Gb/s** InfiniBand/VPI adapters with Socket Direct, Mellanox Switch-IB 2 SB7800 36-Port 100Gb/s EDR InfiniBand switches • **Memory: 192GB DDR4 2677MHz RDIMMs per node**
20 node Bull|ATOS AMD EPYC cluster running SLURM; • **AMD EPYC 7551**; # of CPU Cores: 32; # of Threads: 64; Max Boost Clock: 3.2 GHz Base Clock: **2.2 GHz** Default TDP / TDP: 180W; Mellanox EDR **100Gb/s**
32 node Dell|EMC PowerEdge R7425 AMD EPYC cluster running SLURM; • **AMD EPYC 7601**; # of CPU Cores: 32; # of Threads: 64; Max Boost Clock: 3GHz Base Clock: **2.0 GHz** Default TDP / TDP: 180W; Mellanox EDR **100Gb/s**

Baseline Cluster Systems

Cluster	Configuration
	Intel Sandy Bridge Clusters
“Raven”	128 x Bull|ATOS b510 EP-nodes each with 2 Intel Sandy Bridge E5-2670 (2.6 GHz), with Mellanox QDR infiniband.
Supercomputing Wales	384 x Fujitsu CX250 EP-nodes each with 2 Intel Sandy Bridge E5-2670 (2.6 GHz), with Mellanox QDR infiniband.
	Intel Broadwell Clusters
Dell PE R730/R630, Broadwell EP-2697A v4 2.6 GHz 16C	HPC Advisory Council, “Thor” cluster, Dell PowerEdge R730/R630 36-node cluster: **2 x Xeon E5-2697A v4 @ 2.6GHz, 16 Core,** 145W TDP, 40MB Cache,256GB **DDR4 2400MHz**, Interconnect: ConnectX-4 EDR
ATOS Broadwell EP-2680 v4 2.4 GHz 16C	32 node cluster, Node config: **2 x Xeon E5-2680 v4 @ 2.4GHz, 16 Core,** 145W TDP, 40MB Cache,128GB **DDR4 2400MHz**, Interconnect: Mellanox ConnectX-4 EDR; and Intel OPA
	IBM Power 8 S822LC
IBM Power 8 S822LC with Mellanox EDR	20 cores, **3.49 GHz with performance CPU governor**; 256 GB memory ; 1 – IB (EDR) port ; 2 × NVIDIA K80 GPU;
	IBM PE (Parallel Environment) Operating System: RHEL 7.2 LE; Compilers: xlC 13.1.3, xlf 15.1.3, gcc 4.8.5 (Red Hat), gcc 5.2.1 (from IBM Advance Toolchain 9.0)

The Performance Benchmarks

- The ***Test suite*** comprises both **synthetics & end-user applications**. Synthetics include HPCC (*http://icl.cs.utk.edu/hpcc/*) & IMB benchmarks (*http://software.intel.com/en-us/articles/intel-mpi-benchmarks*), IOR and STREAM
- Variety of "open source" & commercial end-user application codes:

GROMACS, LAMMPS, **AMBER,** NAMD, **DL_POLY classic** & **DL_POLY-4** (molecular dynamics)

Quantum Espresso, Siesta**,** CP2K, ONETEP, CASTEP and **VASP** (ab initio Materials properties)

NWChem, GAMESS-US and **GAMESS-UK** (molecular electronic structure)

- 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*.

EPYC - Compiler and Run-time Options

STREAM (Atos Clusters):

```
module load AMD/amd-cputype/1.0
icc -o stream.x stream.c -DSTATIC -
Ofast -xCORE-AVX2 -qopenmp -
DSTREAM_ARRAY_SIZE=800000000 \
-mcmodel=large -shared-intel

export OMP_NUM_THREADS=16
export OMP_PROC_BIND=true
export OMP_PLACES="{0:4:1}:16:4" #1
thread per CCX
export OMP_DISPLAY_ENV=true
```

```
#
# Preload the amd-cputype library to navigate
# the Intel Genuine cpu test
module use /opt/amd/modulefiles
module load AMD/amd-cputype/1.0
export LD_PRELOAD=$AMD_CPUTYPE_LIB
export OMP_PROC_BIND=true
# export KMP_AFFINITY=granularity=fine
export I_MPI_DEBUG=5
export MKL_DEBUG_CPU_TYPE=5
```

STREAM (Dell|EMC EPYC):

```
export OMP_NUM_THREADS=32
export OMP_PROC_BIND=true
export OMP_DISPLAY_ENV=true
export
OMP_PLACES="{0},{16},{8},{24},{2},{1
8},{10},{26},{4},{20},{12},{28},{6},
{22},{14},{30},{1},{17},{9},{25},{3}
,{19},{11},{27},{5},{21},{13},{29},{
7},{23},{15},{31}"
```

Compilation:

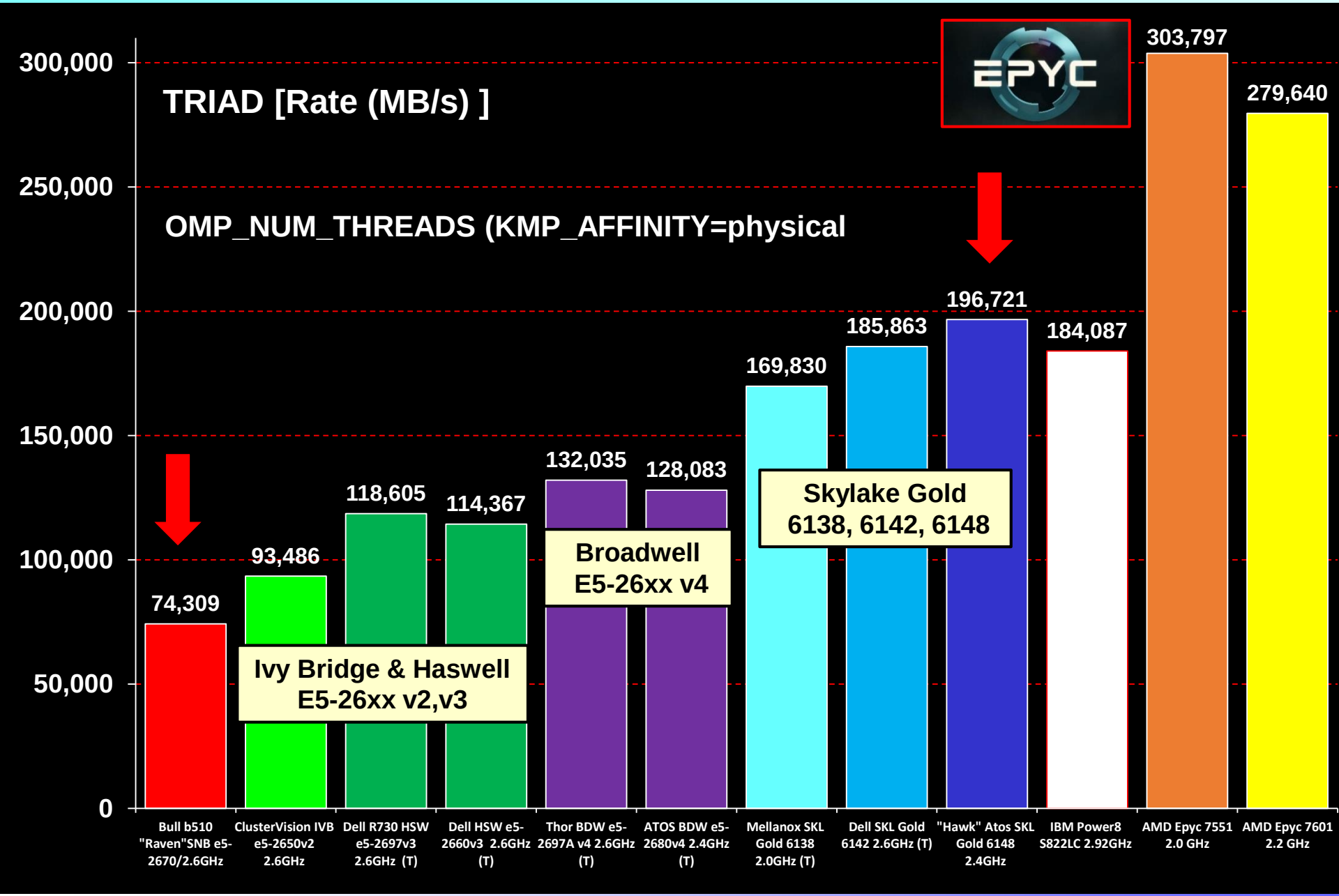
INTEL COMPILERS 2018, IntelMPI 2017 Update 3, FFTW-3.3.5

INTEL SKL: -O3 –xCORE-AVX512

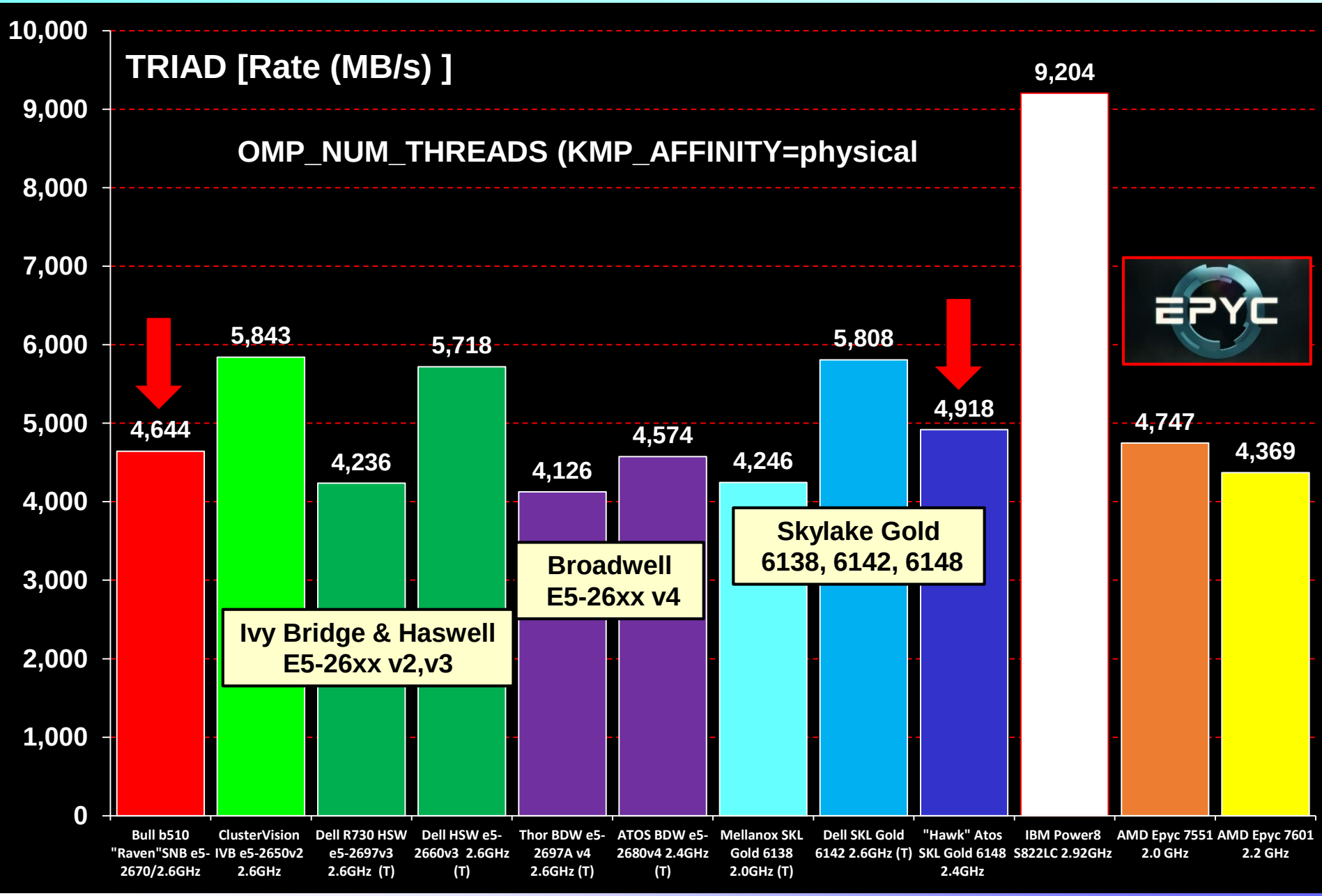
AMD EPYC: –O3 -xAVX2

AMD EPYC: -axCORE-AVX-I

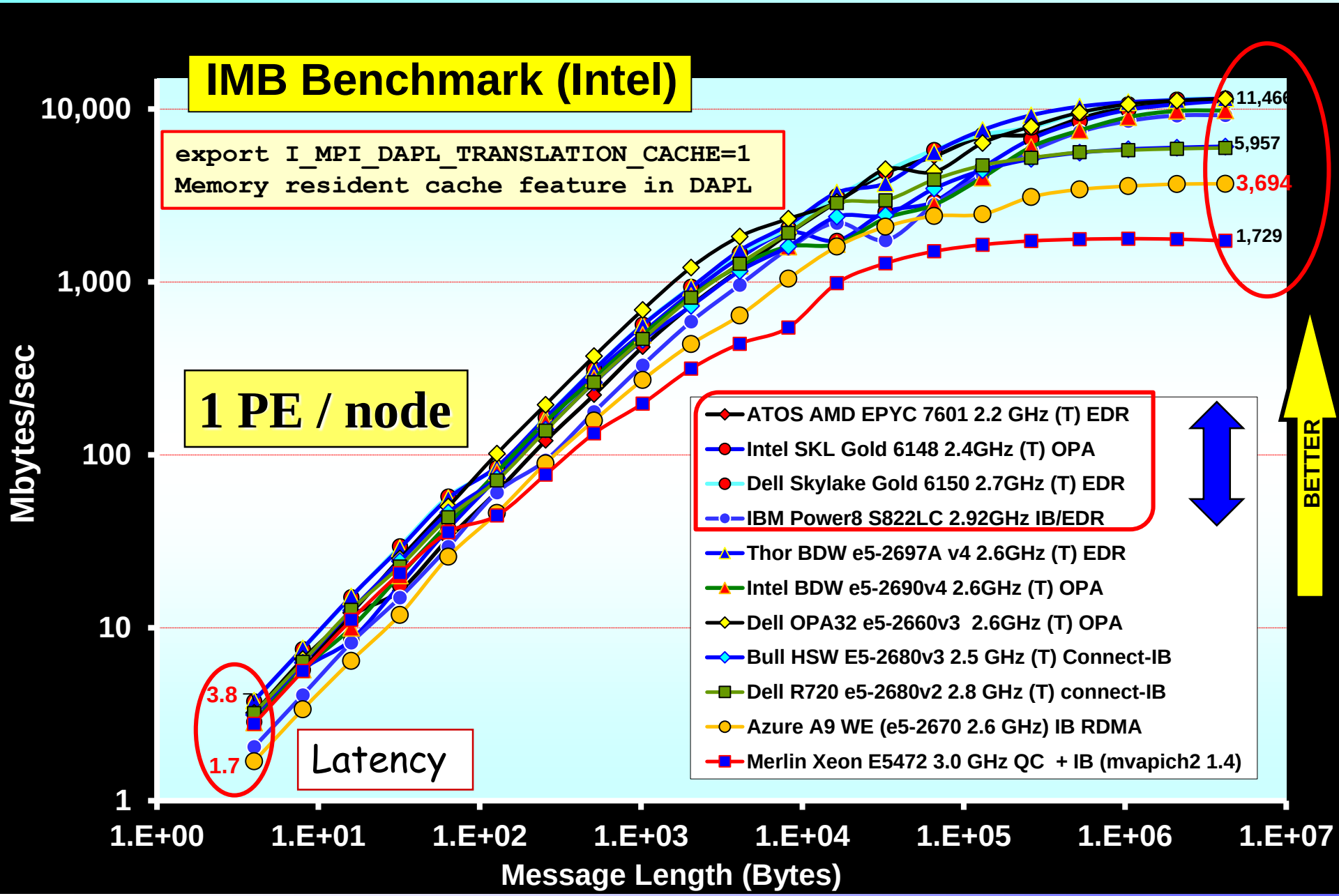
Memory B/W –STREAM performance



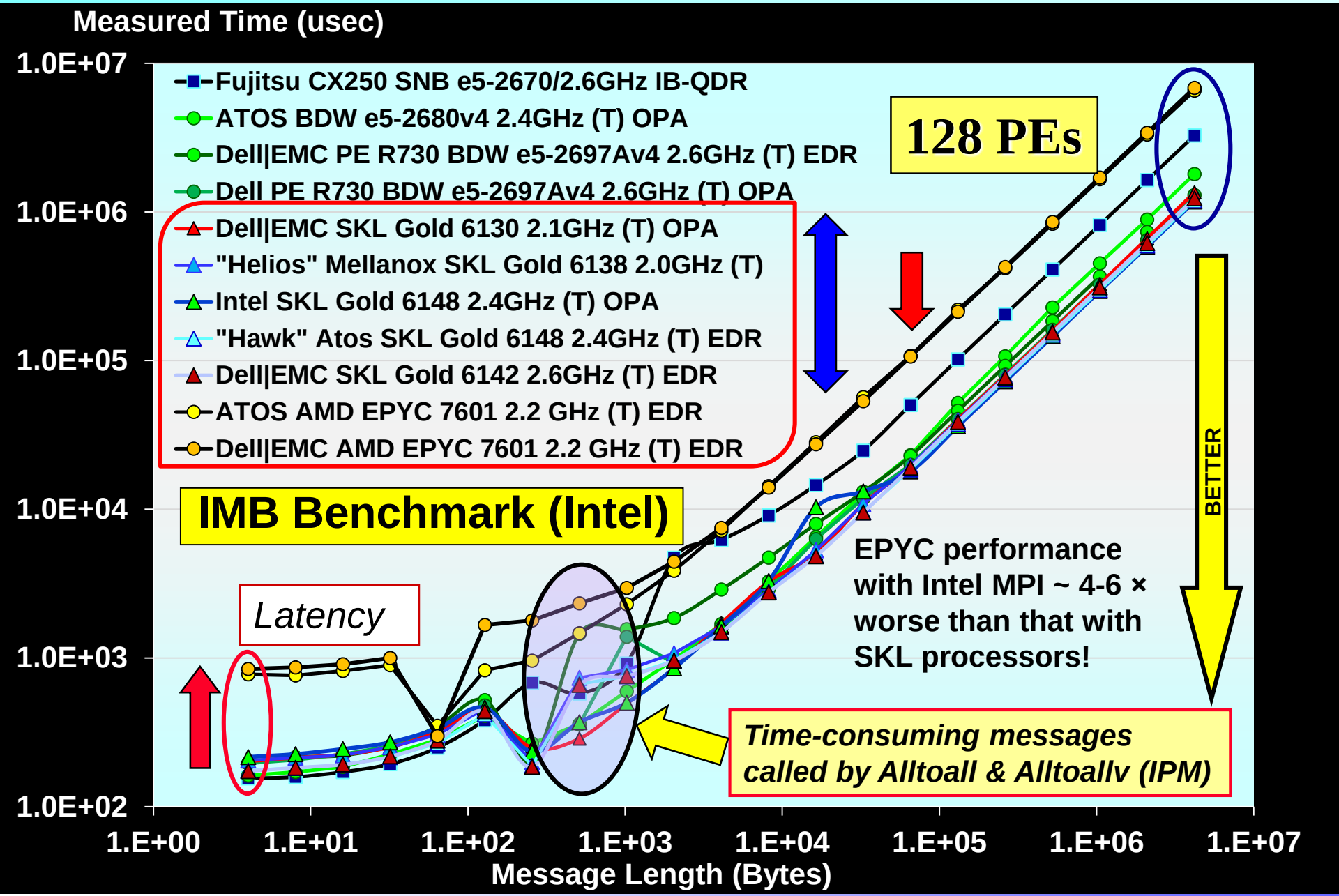
Memory B/W – STREAM / core performance



MPI Performance – PingPong



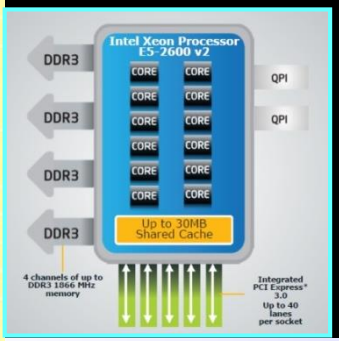
MPI Collectives – Alltoallv (128 PEs)

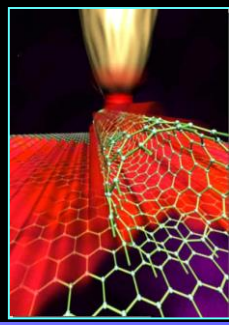


CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Application Performance on Multi-core Processors

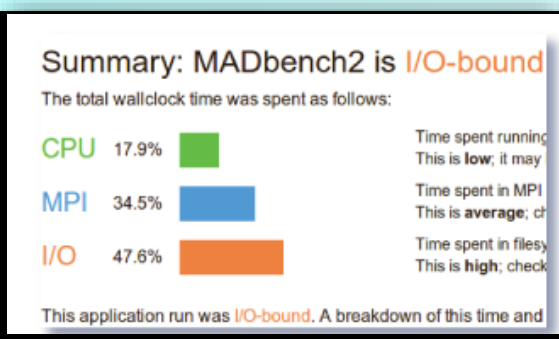
I.1 *THE CODES: DLPOLY, GROMACS, NAMD, LAMMPS, GAMESS, NWChem, GAMESS-UK, ONETEP, VASP, SIESTA, CASTEP, Quantum Espresso, CP2K – on a variety of HPC systems.*





Allinea (ARM) Performance Reports

Allinea 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 mpiexec command.
- **perf-report mpiexec -n 4 $code**
- ***A Report Summary:*** This characterizes how the application's wallclock time was spent, broken down into CPU, MPI and I/O
- All examples updated on Broadwell Mellanox Cluster (E5-2697A v4)

Molecular Simulation I. DL_POLY

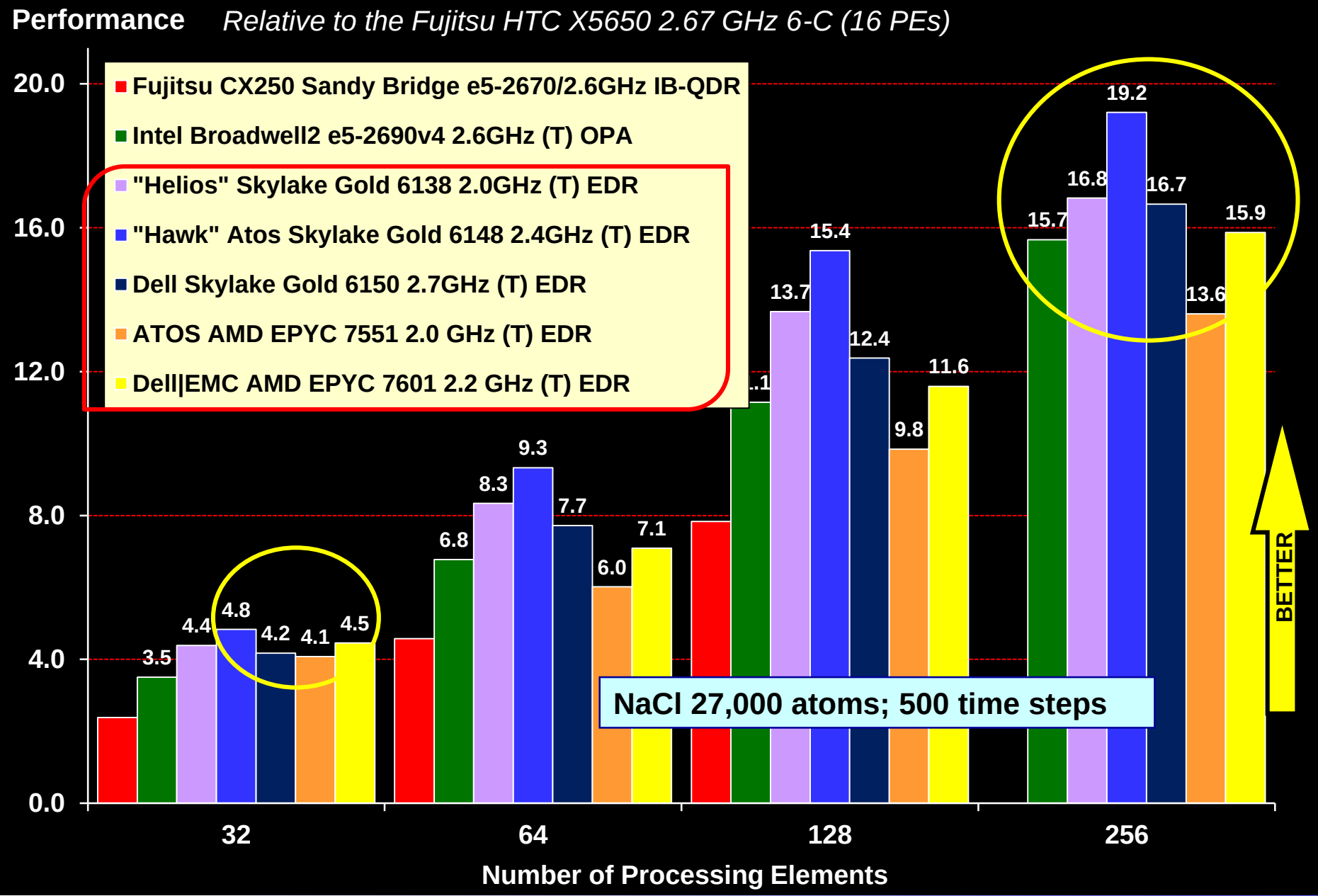
Molecular Dynamics Codes: AMBER, DL_POLY, CHARMM, NAMD, LAMMPS, GROMACS etc

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 Classic – NaCl Simulation

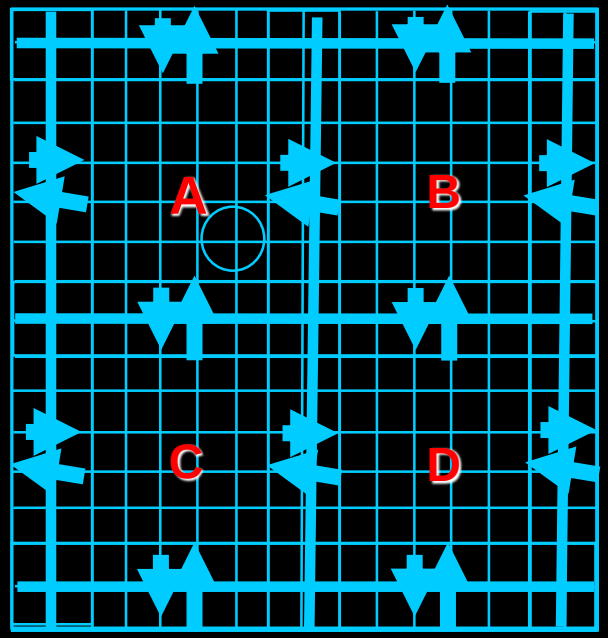
Performance Data (32-256 PEs)



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

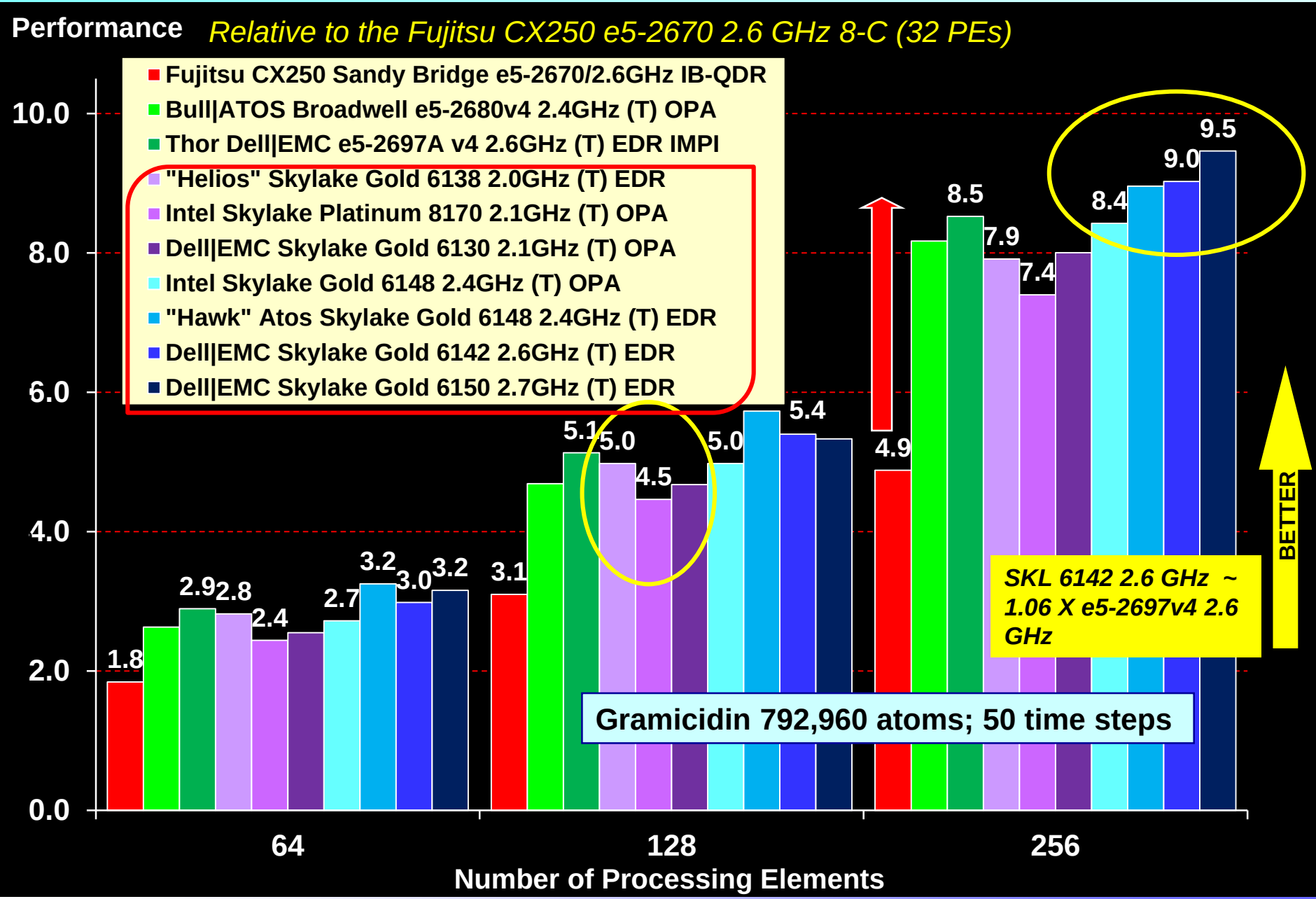
1. NaCl Simulation; 216,000 ions, 200 time steps, Cutoff=12Å

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

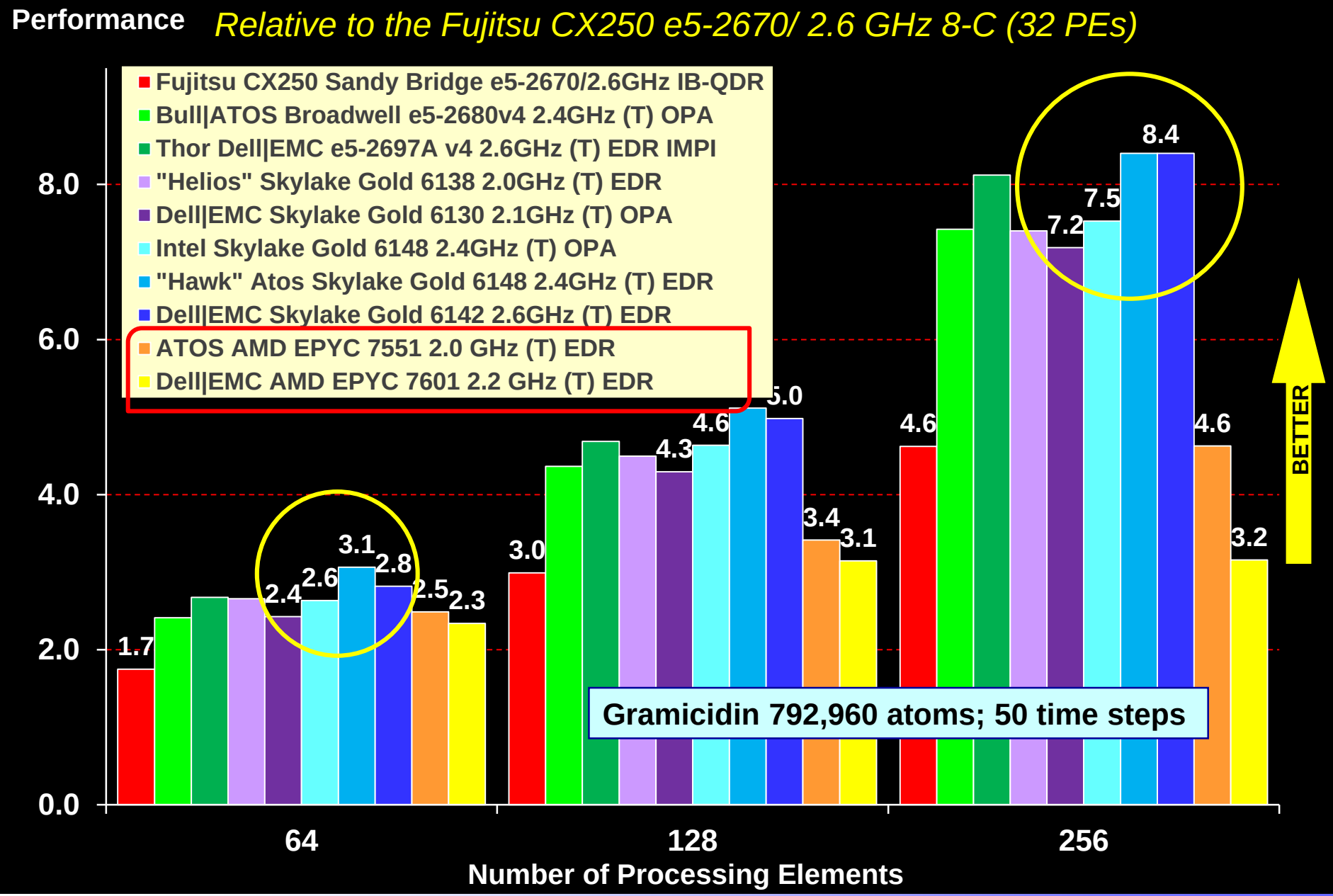
http://www.scd.stfc.ac.uk//research/app/ccg/software/DL_POLY/44516.aspx

DL_POLY 4 – Gramicidin Simulation

Performance Data (64-256 PEs)



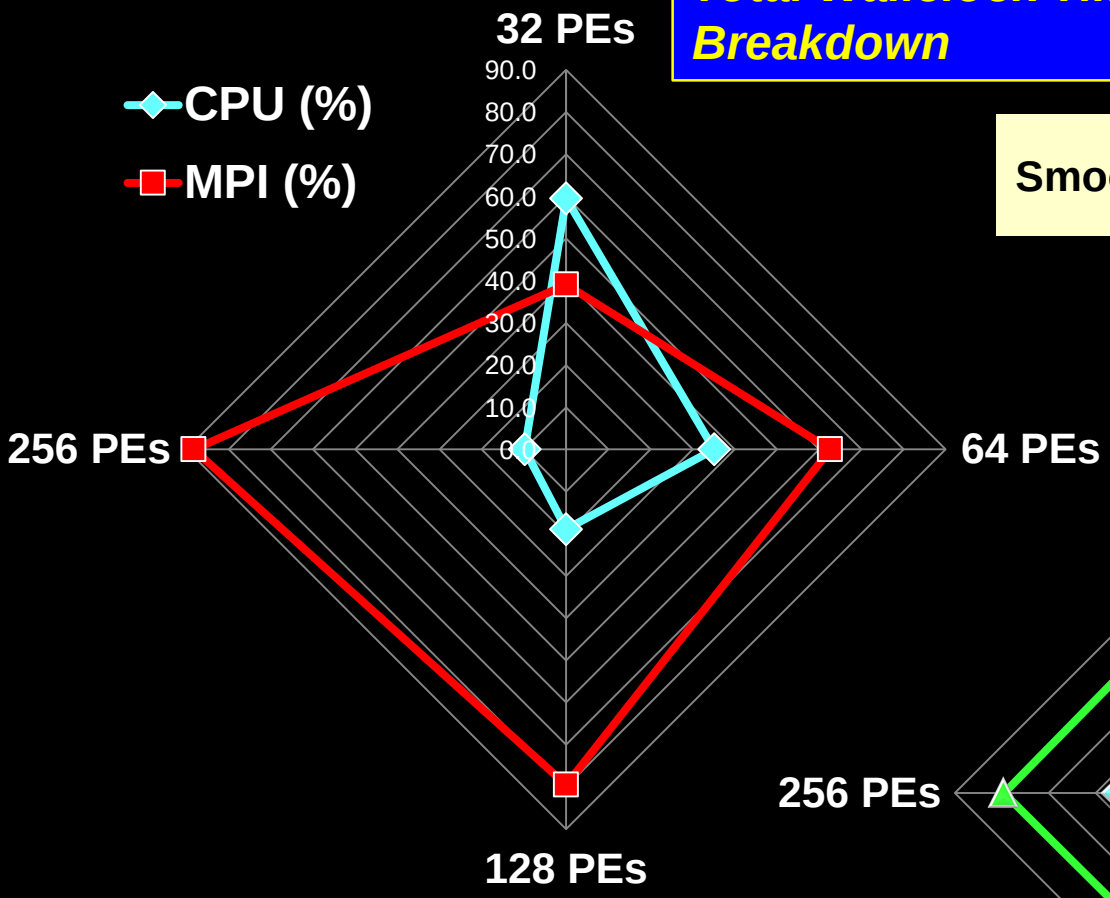
DL_POLY 4 – Gramicidin Simulation – EPYC



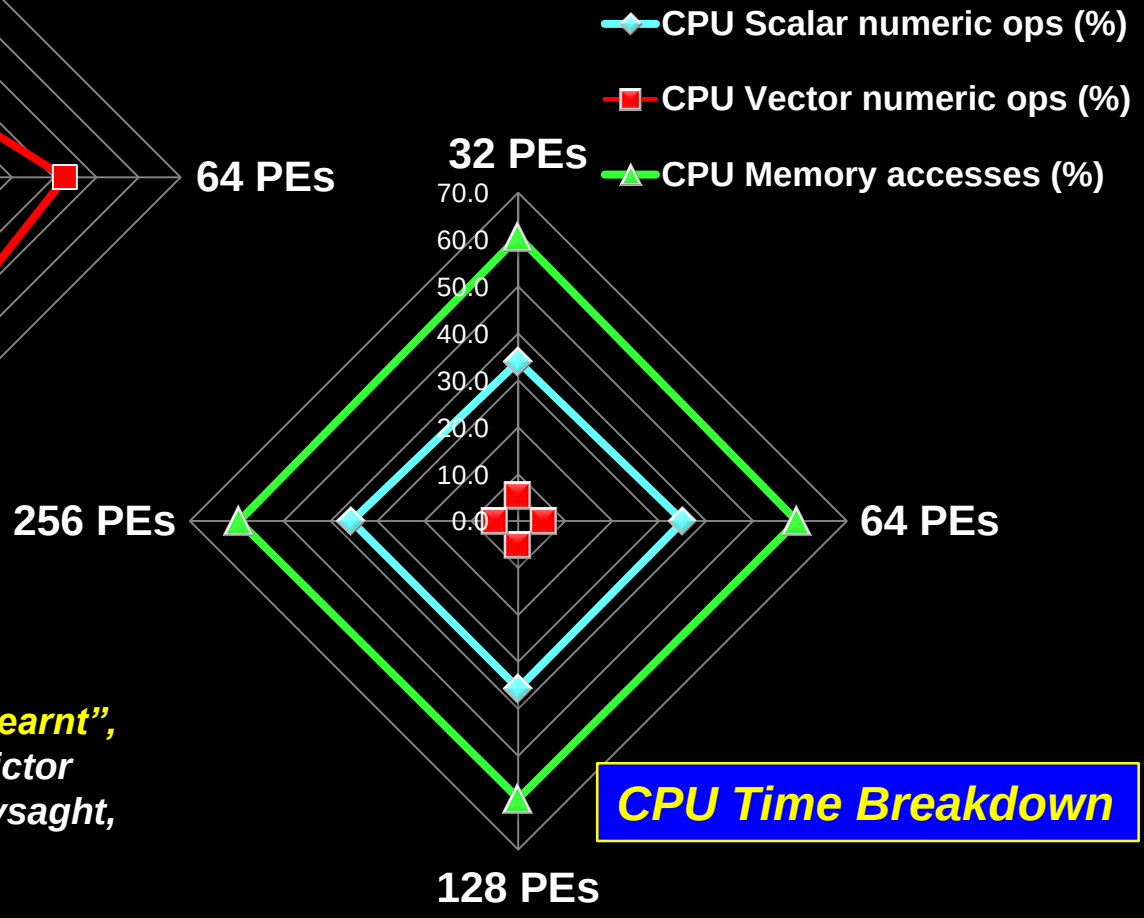
DLPOLY4 – Gramicidin Simulation Performance Report

Total Wallclock Time Breakdown

Performance Data (32-256 PEs)



Smooth Particle Mesh Ewald Scheme



CPU Time Breakdown

"DL_POLY_4 and Xeon Phi: Lessons Learnt", Alin Marin Elena , Christian Lalanne, Victor Gamayunov , Gilles Civario, Michael Lysaght, and IlianTodorov

Molecular Simulation - II GROMACS

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

- Single and Double Precision
- Efficient GPU Implementations

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 (optimised for “Hawk” by Ade Fewings)

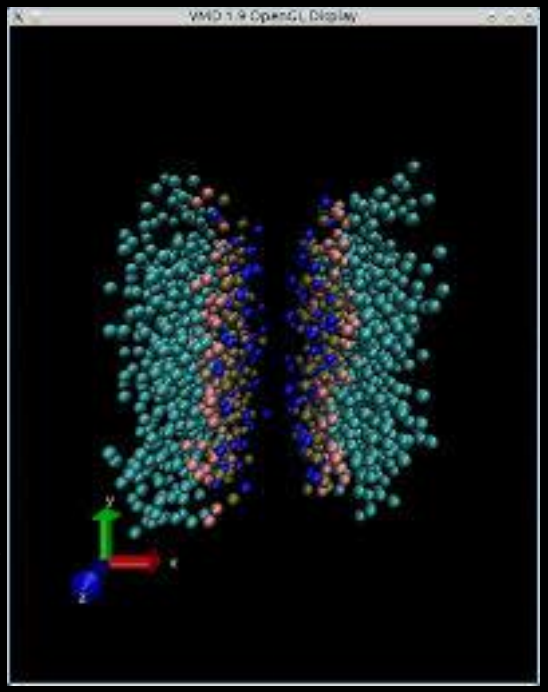
- Berk Hess et al. "GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation". *Journal of Chemical Theory and Computation* 4 (3): 435–447.

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

GROMACS Benchmark Cases

Ion channel system

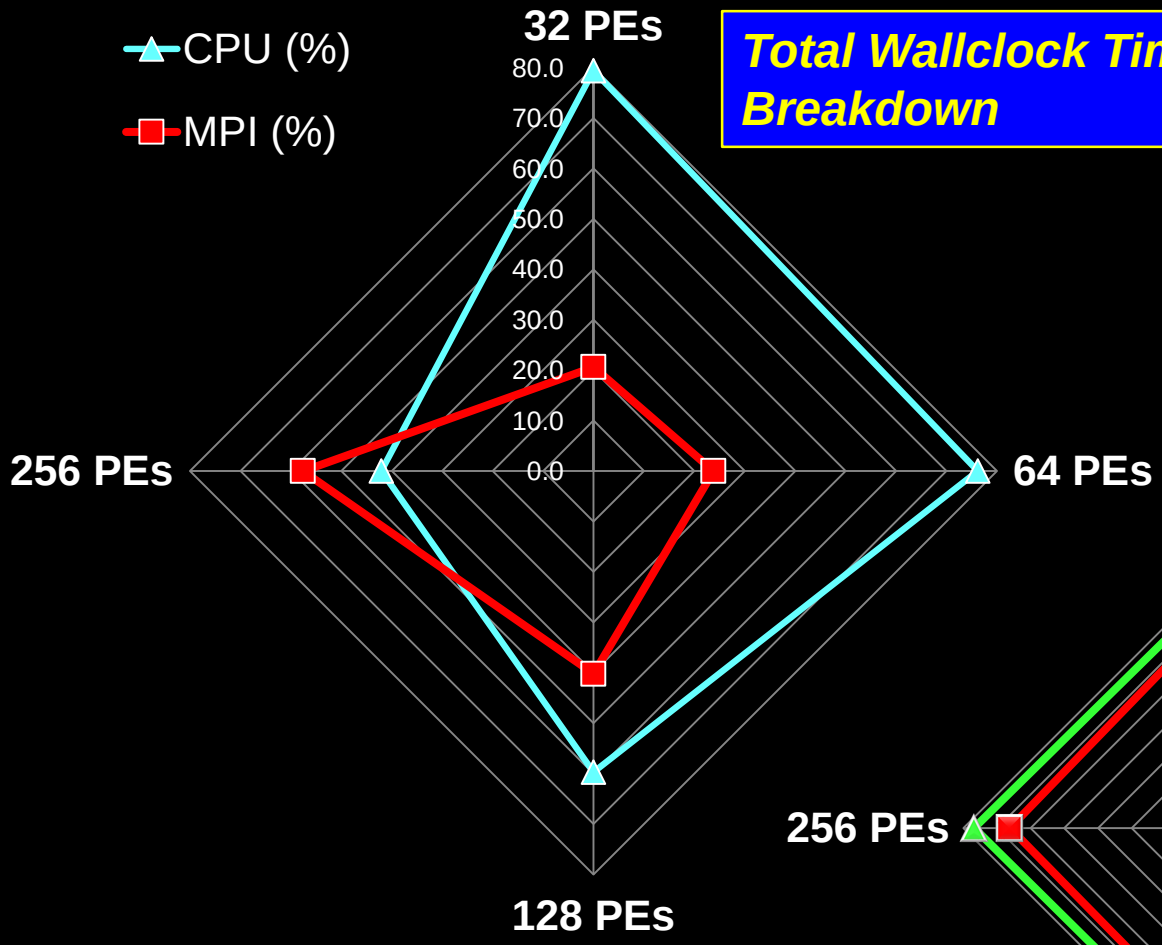
- The 142k particle ion channel system is the membrane protein GluCl - a pentameric chloride channel embedded in a DOPC membrane and solvated in TIP3P water, using the Amber ff99SB-ILDN force field. This system is a challenging parallelization case due to the small size, but is one of the most wanted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



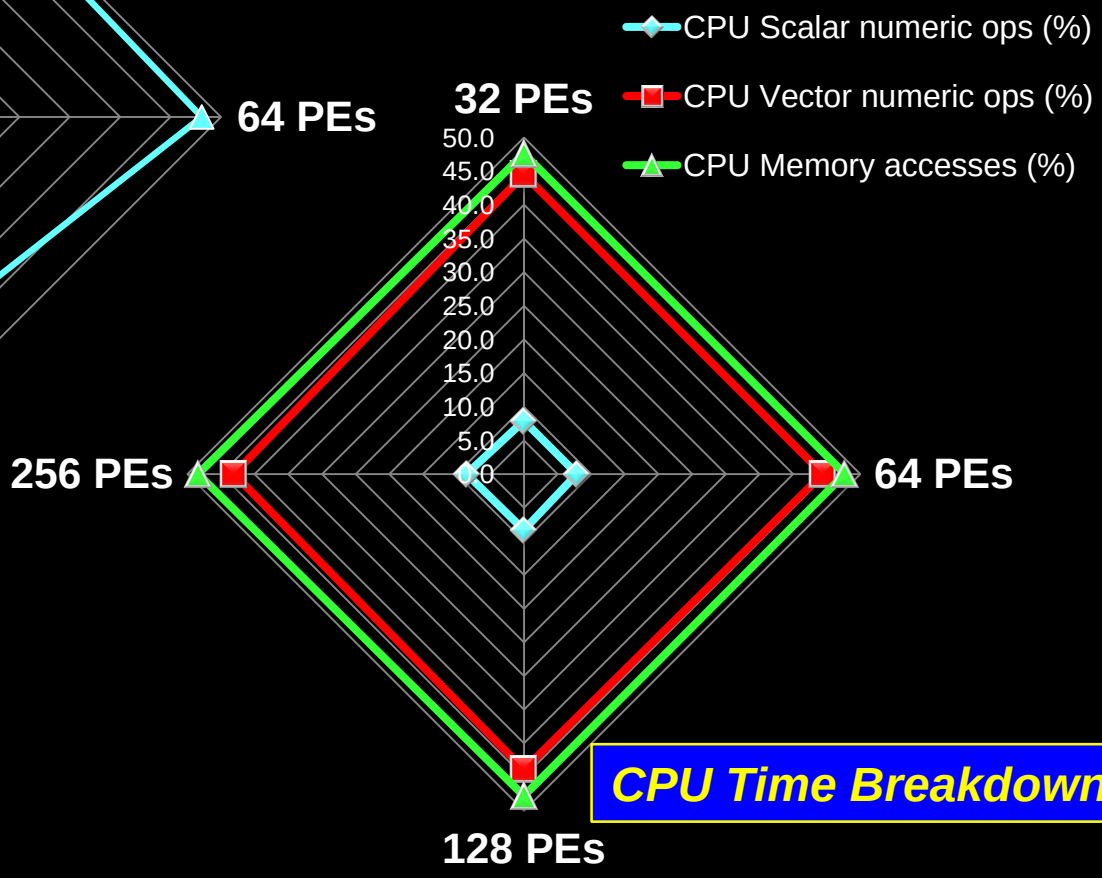
Lignocellulose

- 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 – Ion-channel Performance Report



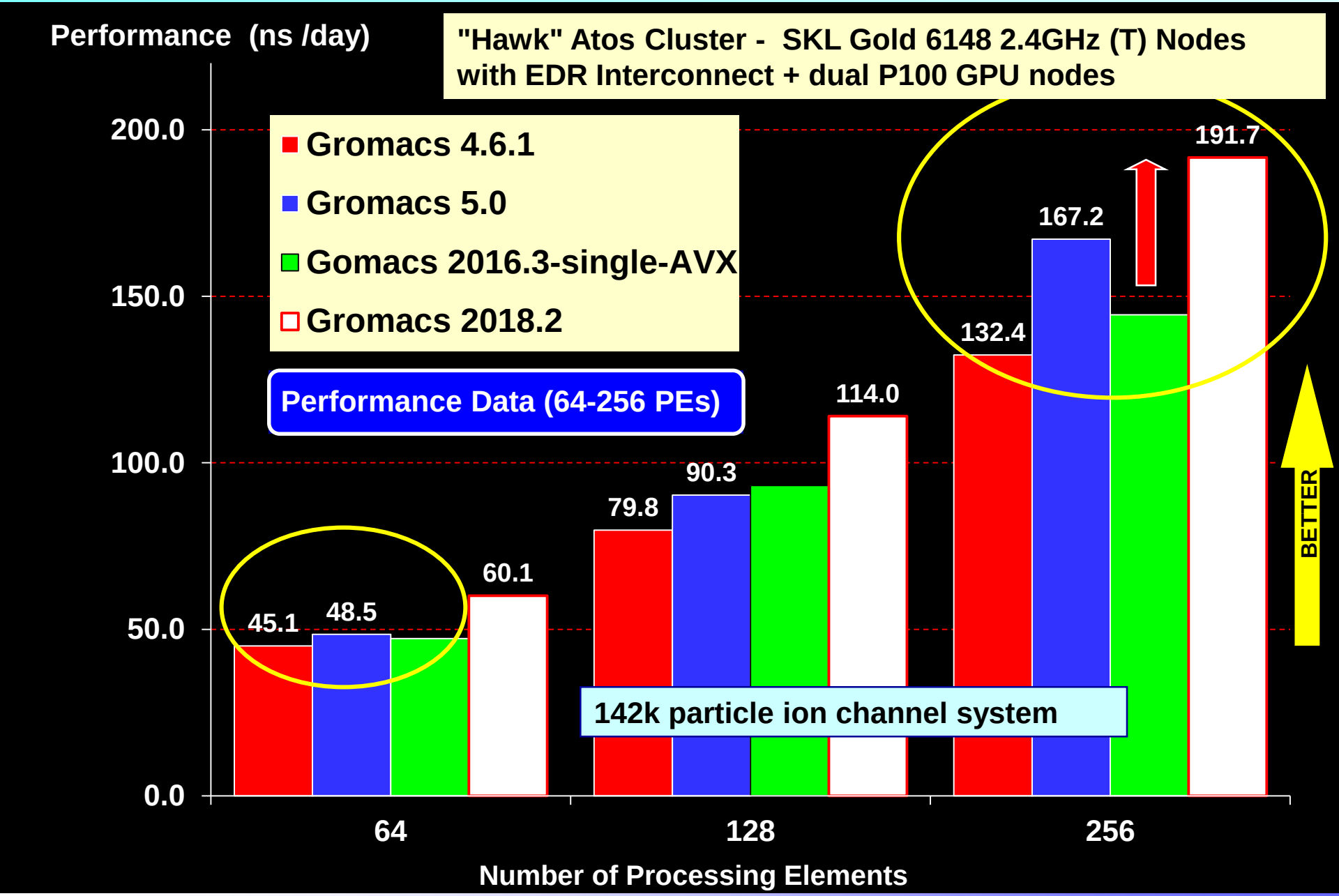
Total Wallclock Time Breakdown



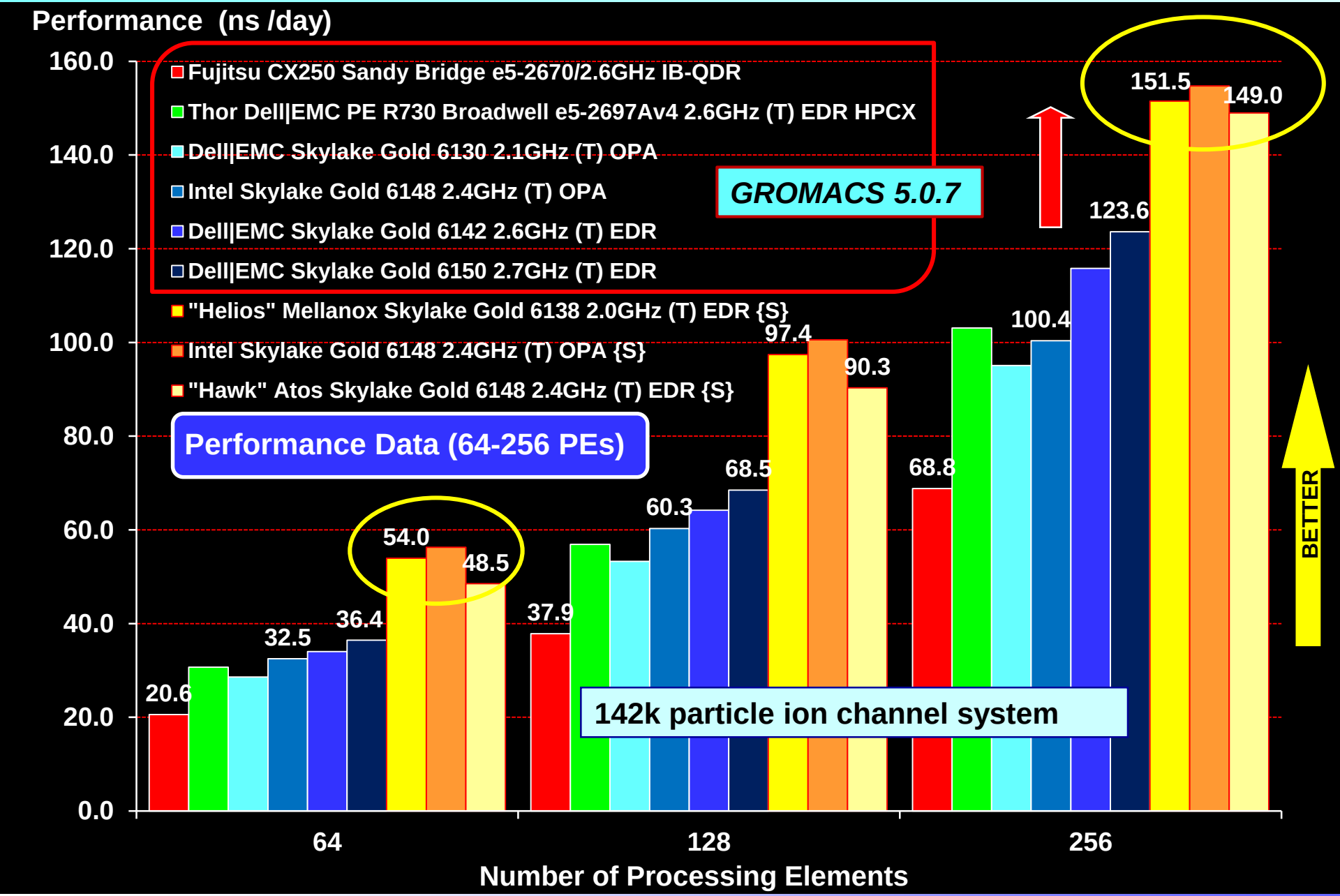
CPU Time Breakdown

Performance Data (32-256 PEs)

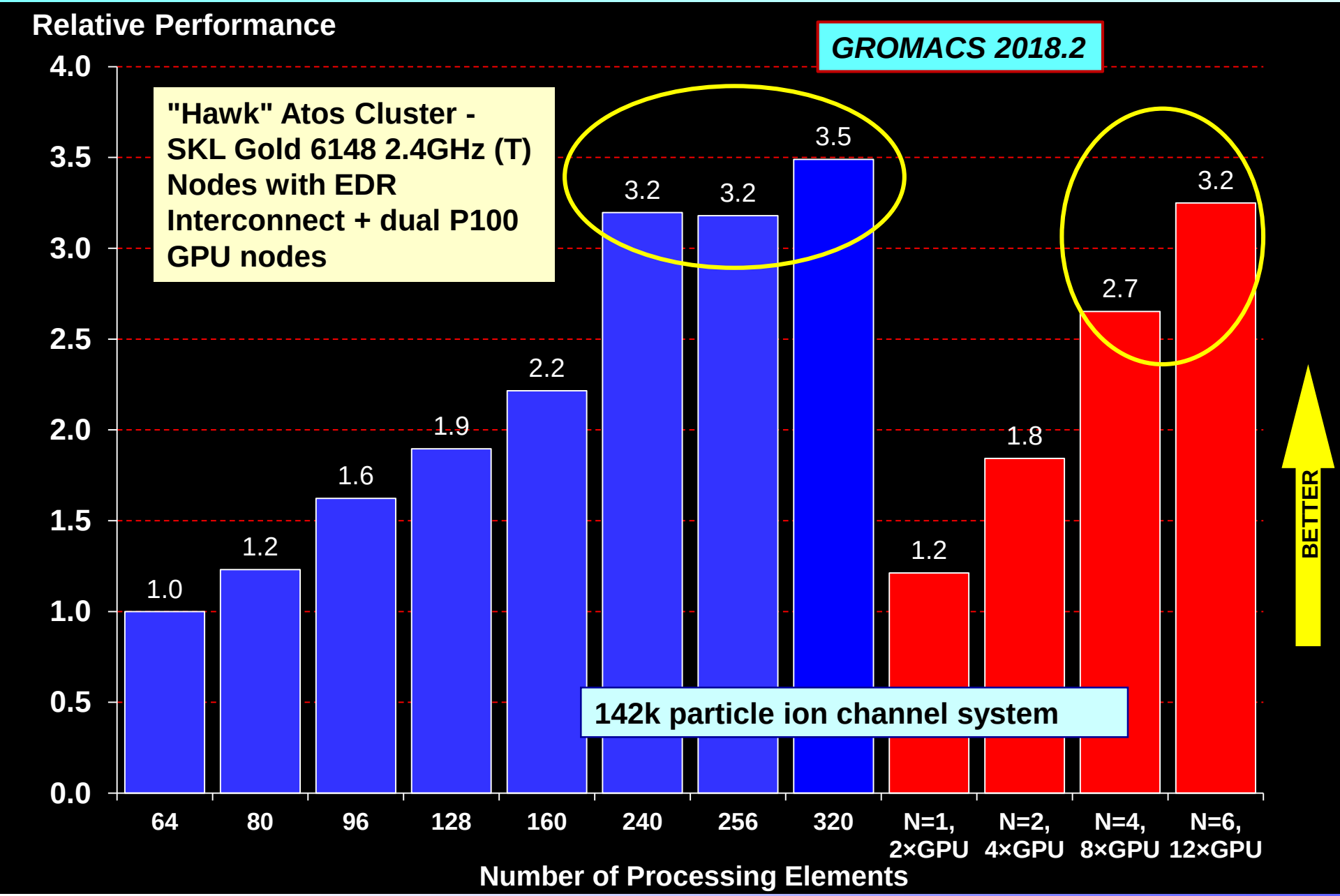
GROMACS – Ion Channel Simulation *Single Precision*



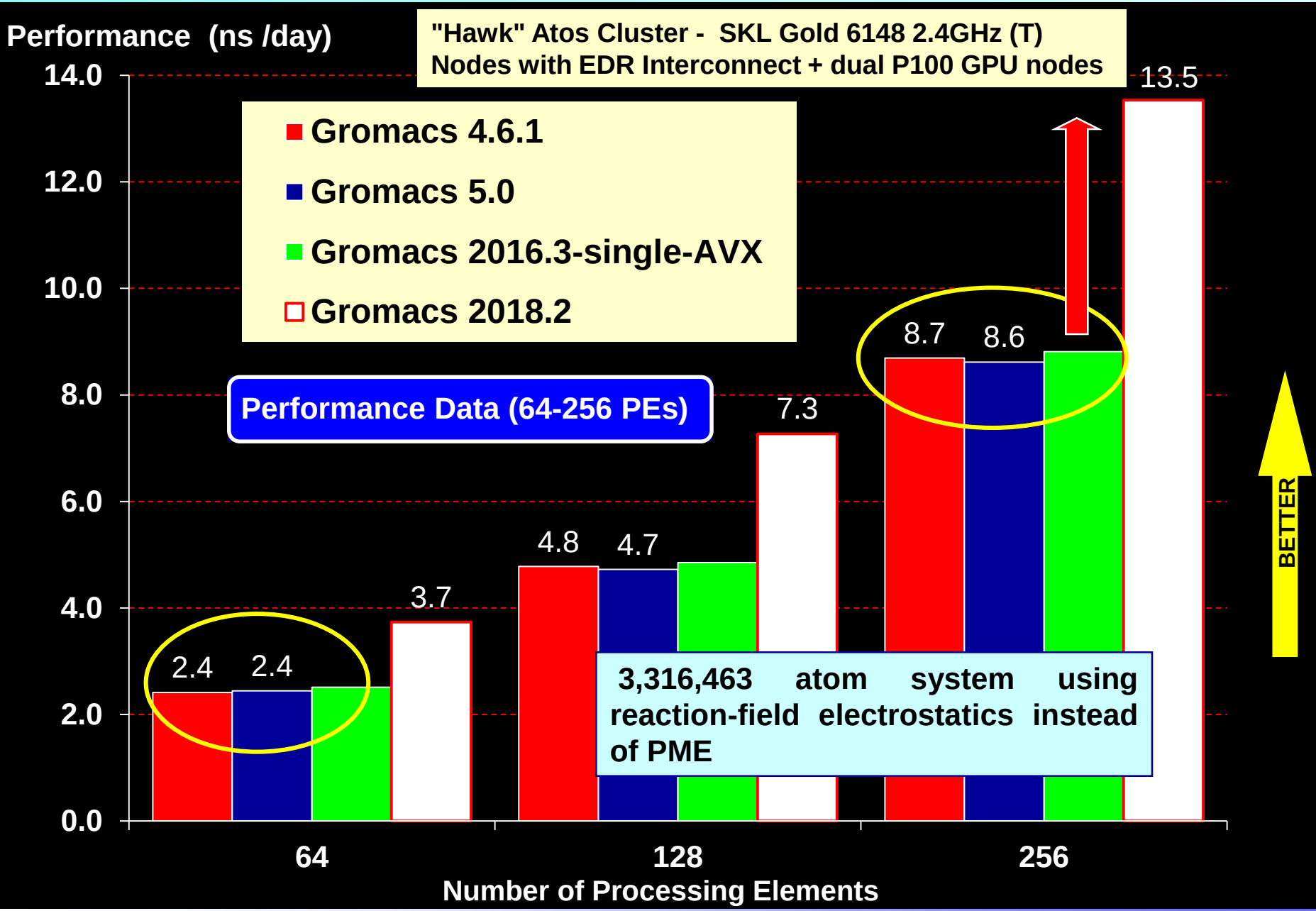
Ion Channel Simulation – Impact of Single Precision



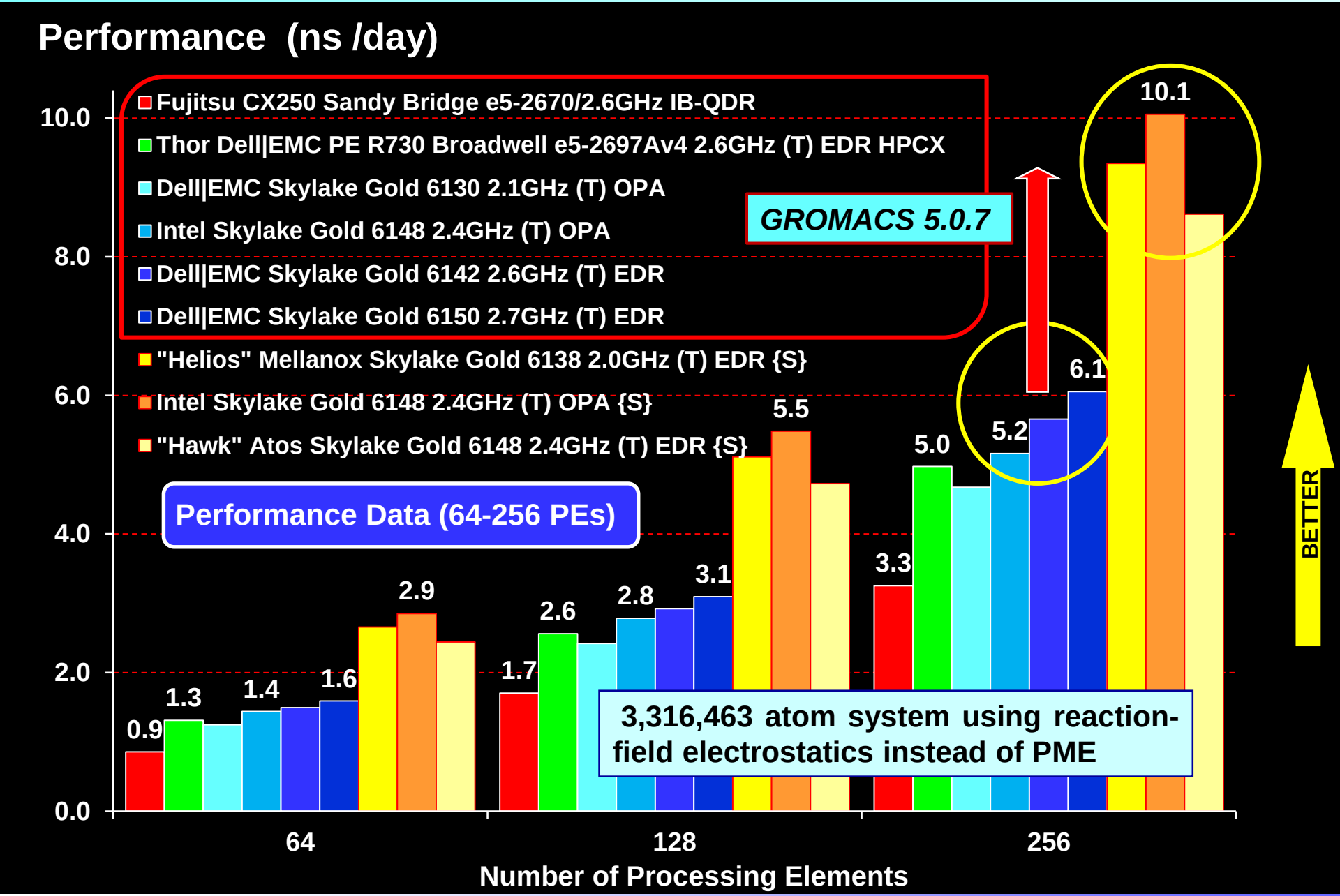
GROMACS – GPU Performance: Ion Channel Simulation



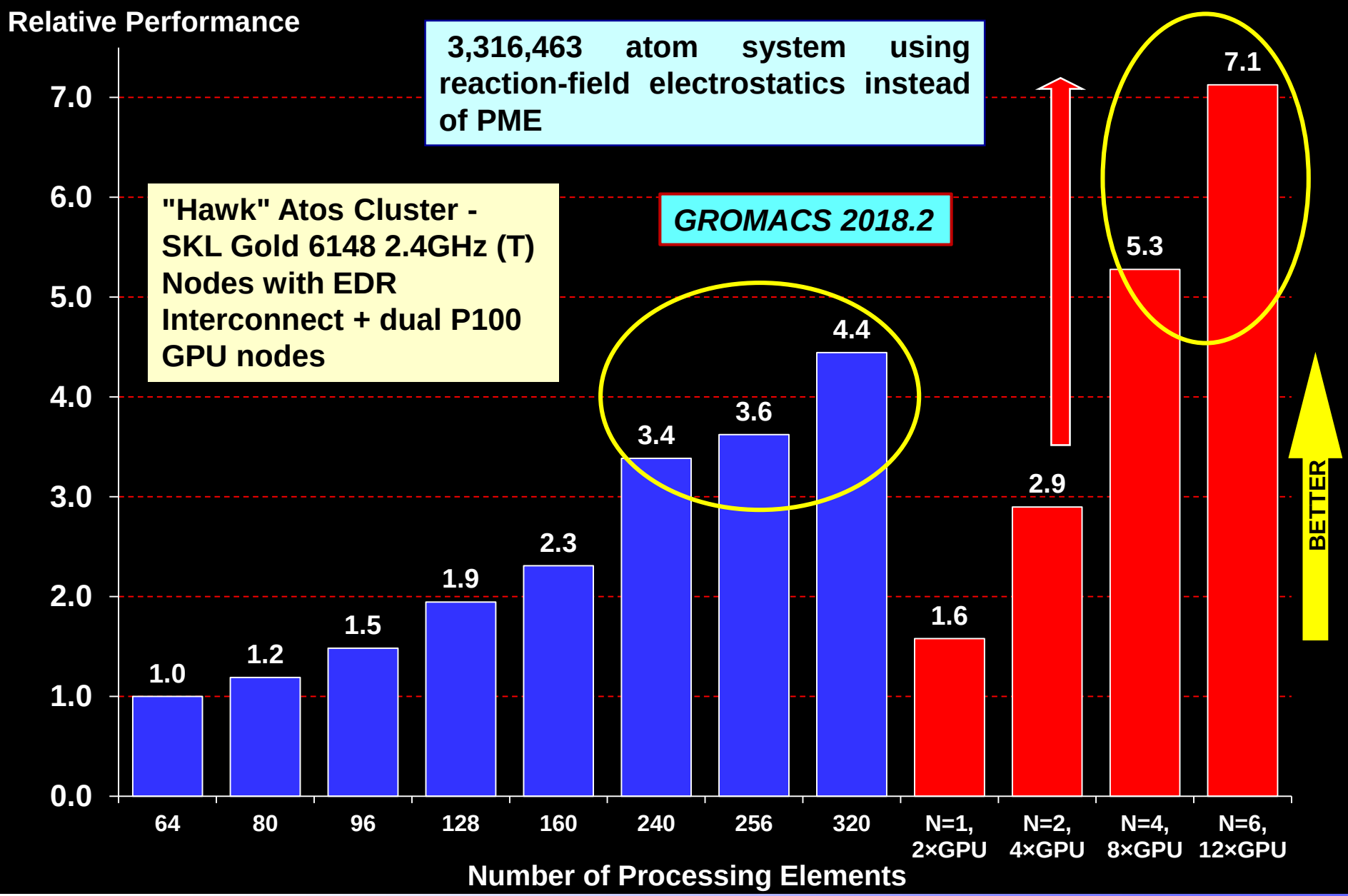
GROMACS – Lignocellulose Simulation *Single Precision*



Lignocellulose Simulation – Impact of Single Precision

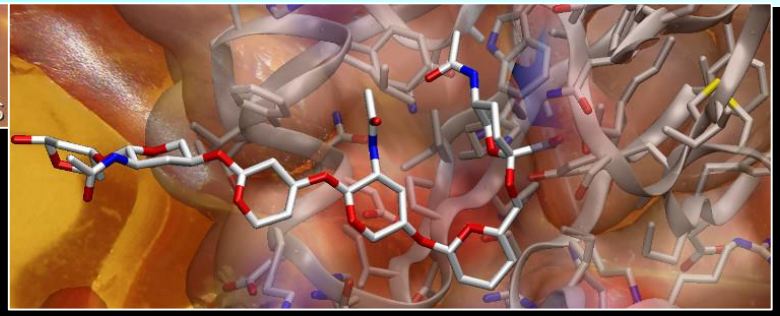


GROMACS – GPU Performance: Lignocellulose Simulation



Molecular Simulation - III The AMBER Benchmarks

The Amber Home Page
Tools for Molecular Simulations



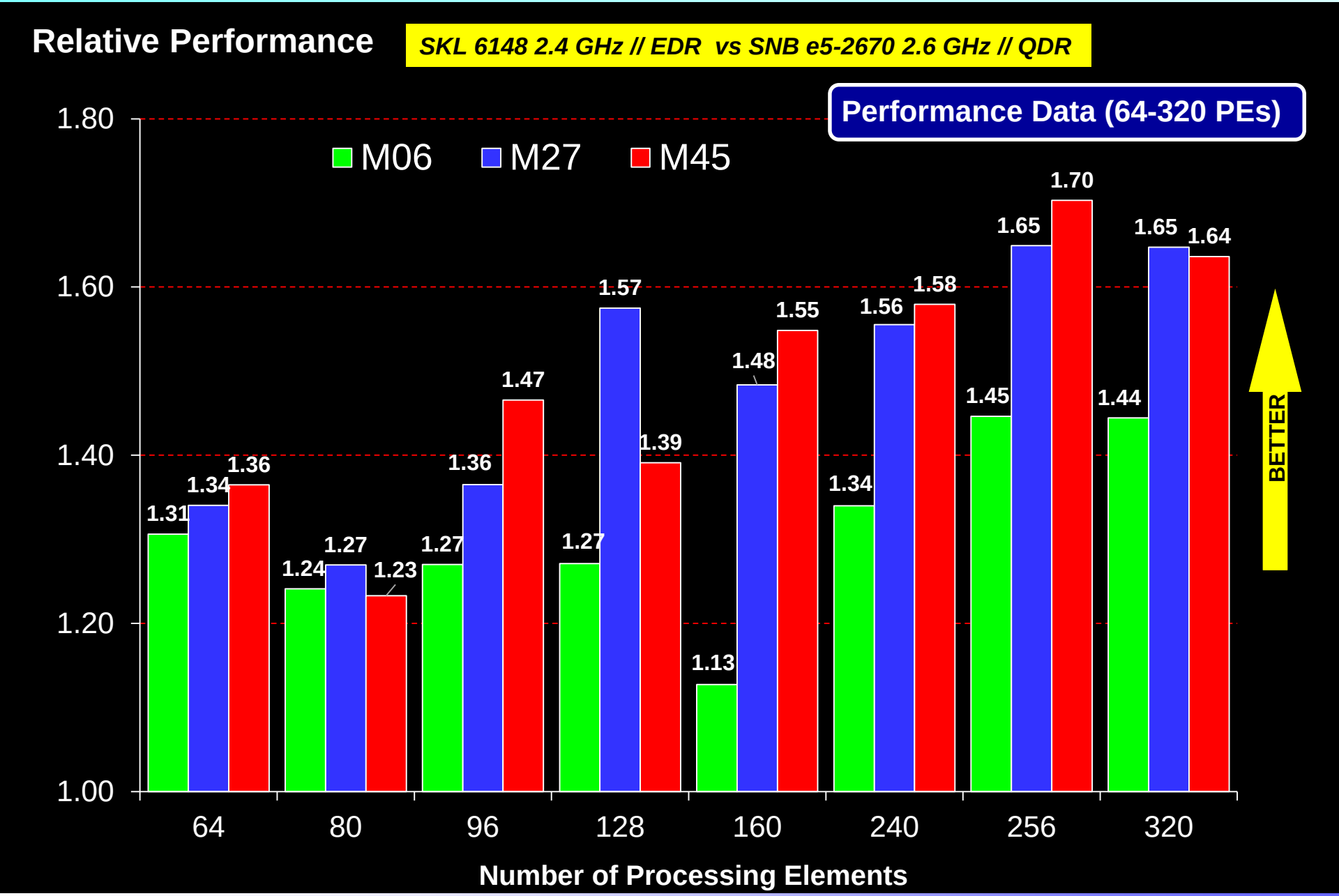
- AMBER 16/1 is used, specifically PMEMD & GPU accelerated PMEMD.
- **M01 Benchmark**
  - Major Urinary Protein (MUP) + IBM ligand (21,736 atoms)
- **M06 Benchmark**
  - Cluster of six MUPs (134,013 atoms)
- M27 Benchmark
  - Cluster of 27 MUPs (657,585 atoms)
- **M45 Benchmark**
  - Cluster of 45 MUPs (932,751 atoms)

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).*

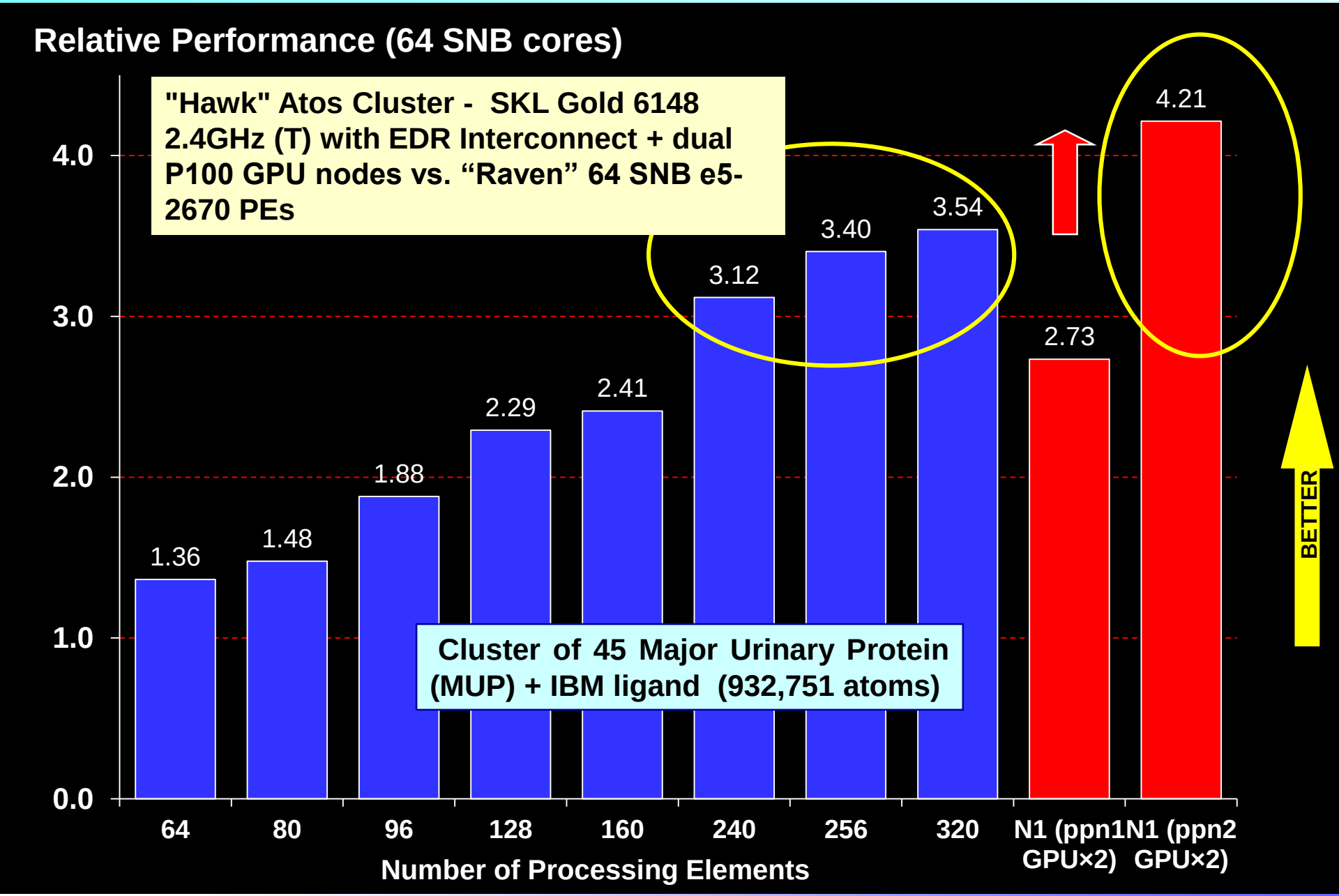
D.A. Case, T.E. Cheatham, III, T. Darden, H. Gohlke, R. Luo, K.M. Merz, Jr., A. Onufriev, C. Simmerling, B. Wang and R. Woods. The Amber biomolecular simulation programs. J. Computat. Chem. ***26,*** *1668-1688 (2005).*

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

AMBER – SKL vs. SNB: M06, M27 and M45

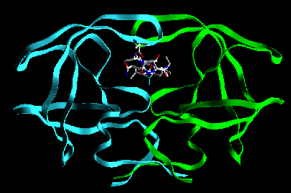


AMBER – GPU Performance: M45 Simulation



GAMESS-UK - Moving to Distributed Data.

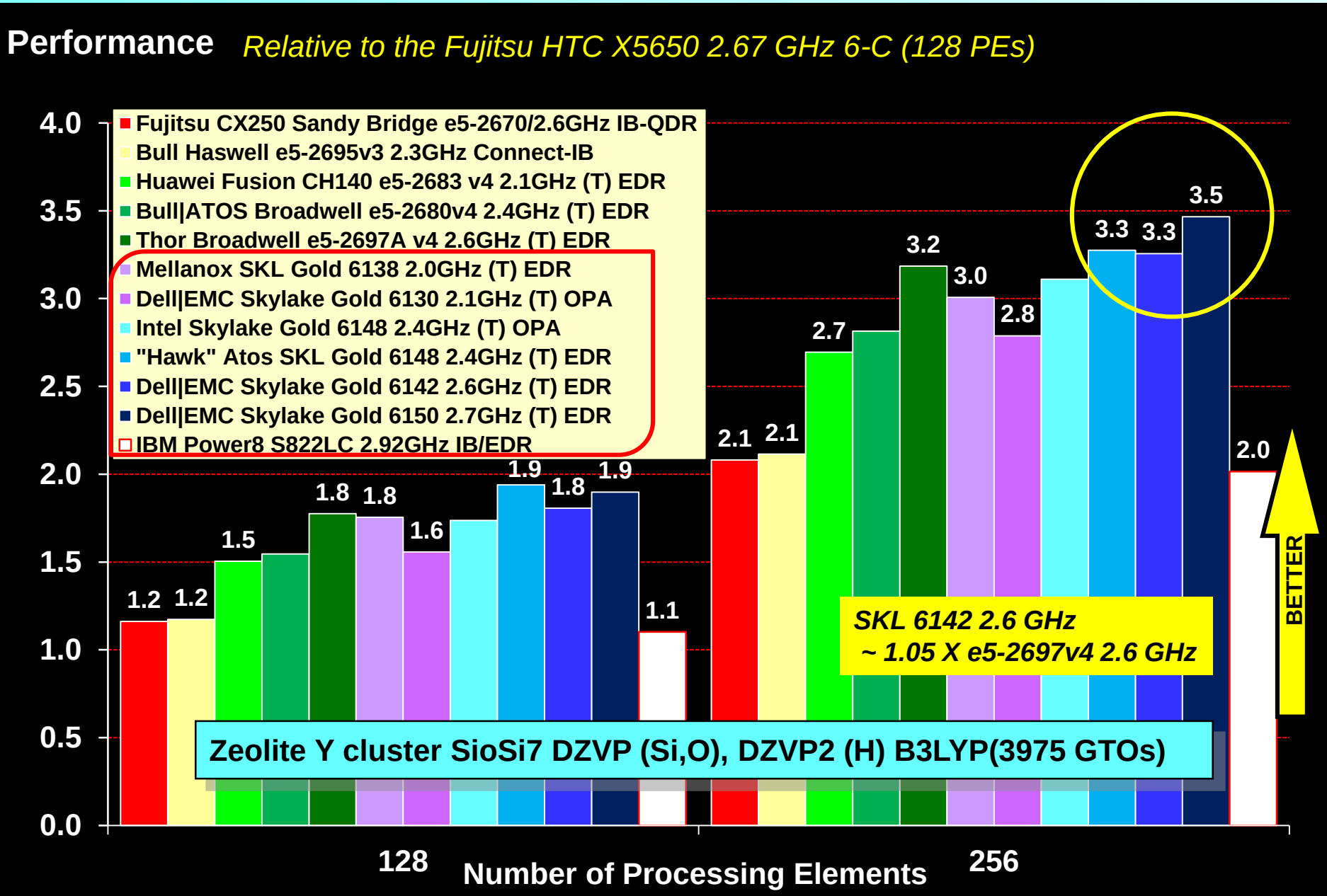
The MPI/ScaLAPACK Implementation of the GAMESS-UK SCF/DFT module



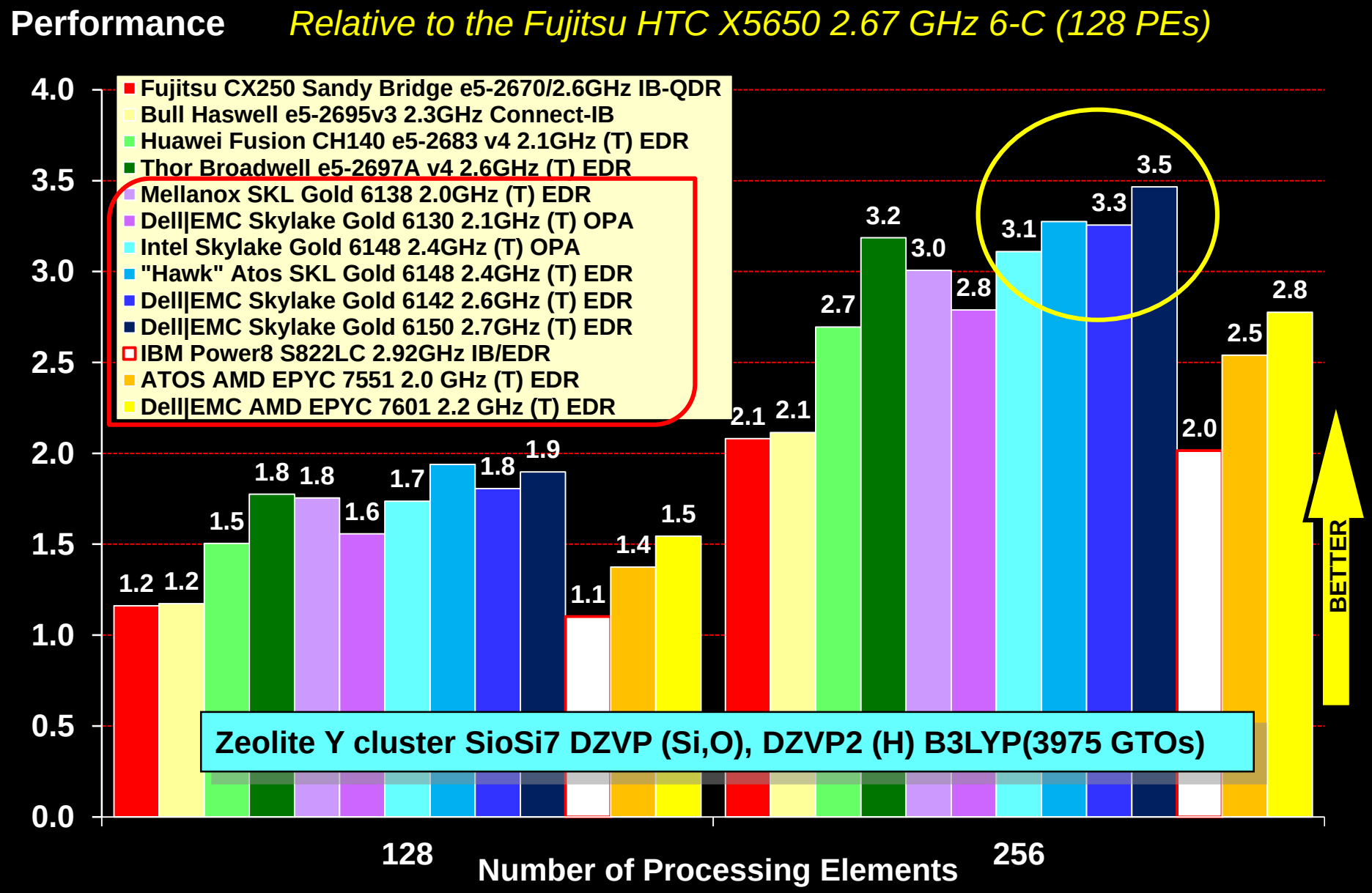
- Pragmatic approach to the replicated data constraints:
  - MPI-based tools (such as ScaLAPACK) used in place of Global Arrays
  - All data structures except those required for the Fock matrix build are fully distributed (F, P)
- Partially distributed model chosen because, in the absence of efficient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

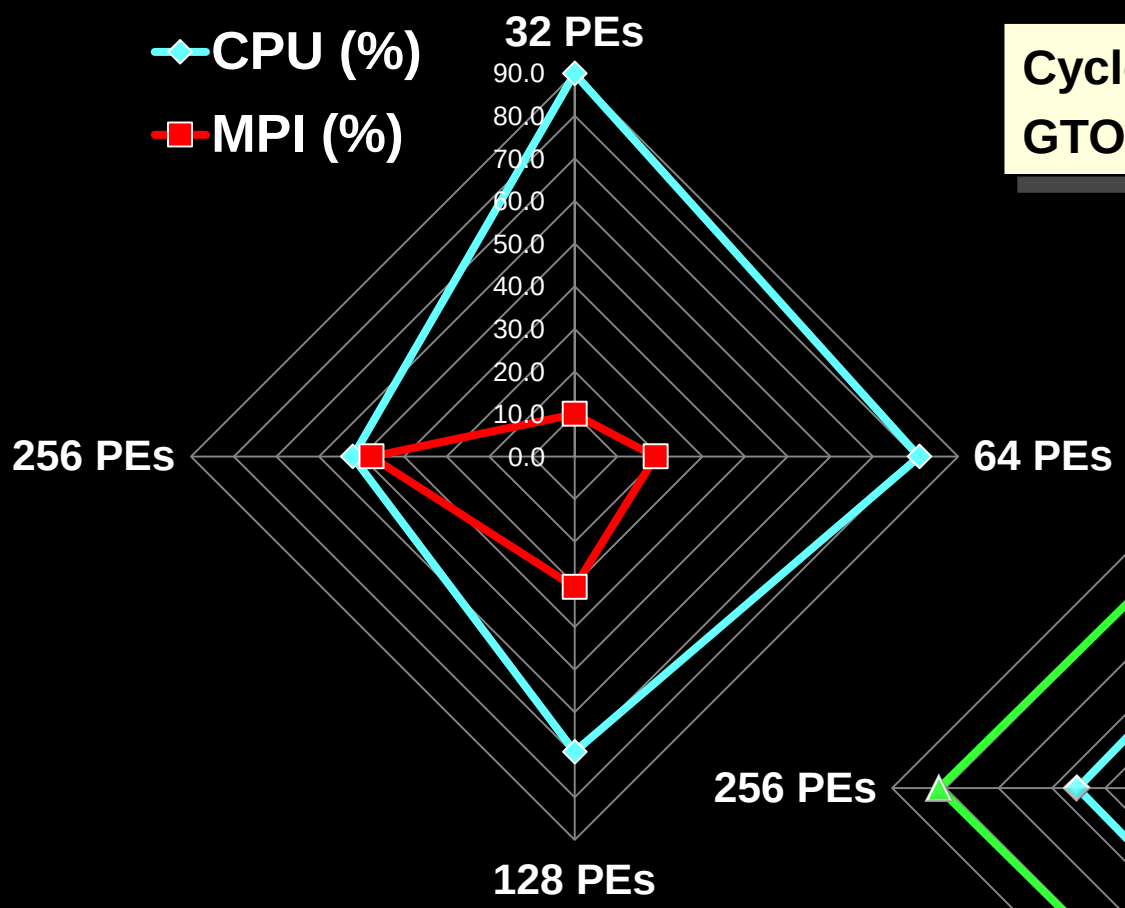
GAMESS-UK Performance - Zeolite Y cluster



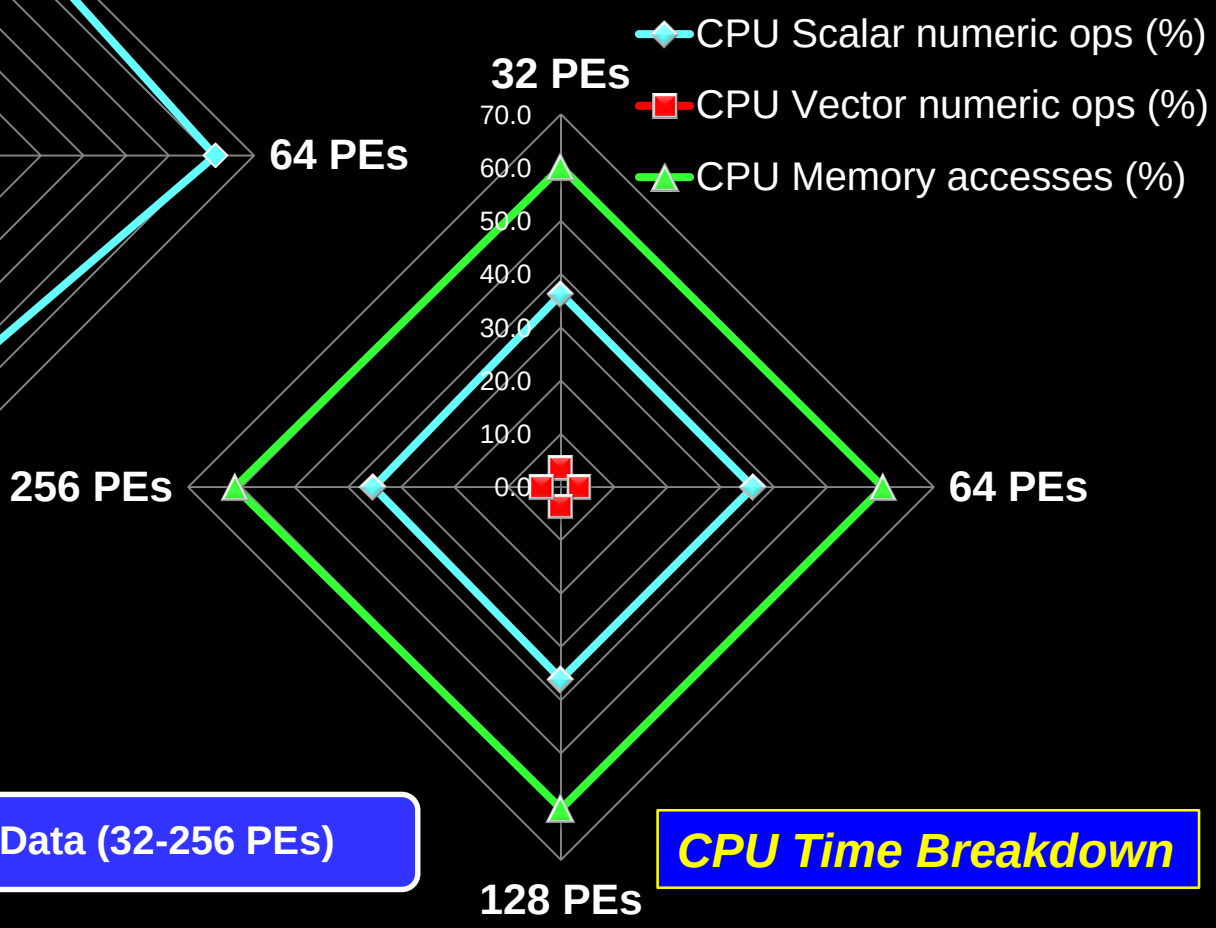
GAMESS-UK MPI/ScaLAPACK code – EPYC Performance



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

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**
- **SIESTA** - an *O*(N) **DFT** code for electronic structure calculations and *ab initio* molecular dynamics simulations for molecules and solids. It uses norm-conserving **pseudopotentials** and linear combination of numerical atomic orbitals (LCAO) basis set.
- **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**.

Quantum Espresso

Ground-state calculations.

Structural Optimization.

Transition states & minimum energy paths.

Ab-initio molecular dynamics.

Response properties (DFPT).

Spectroscopic properties.

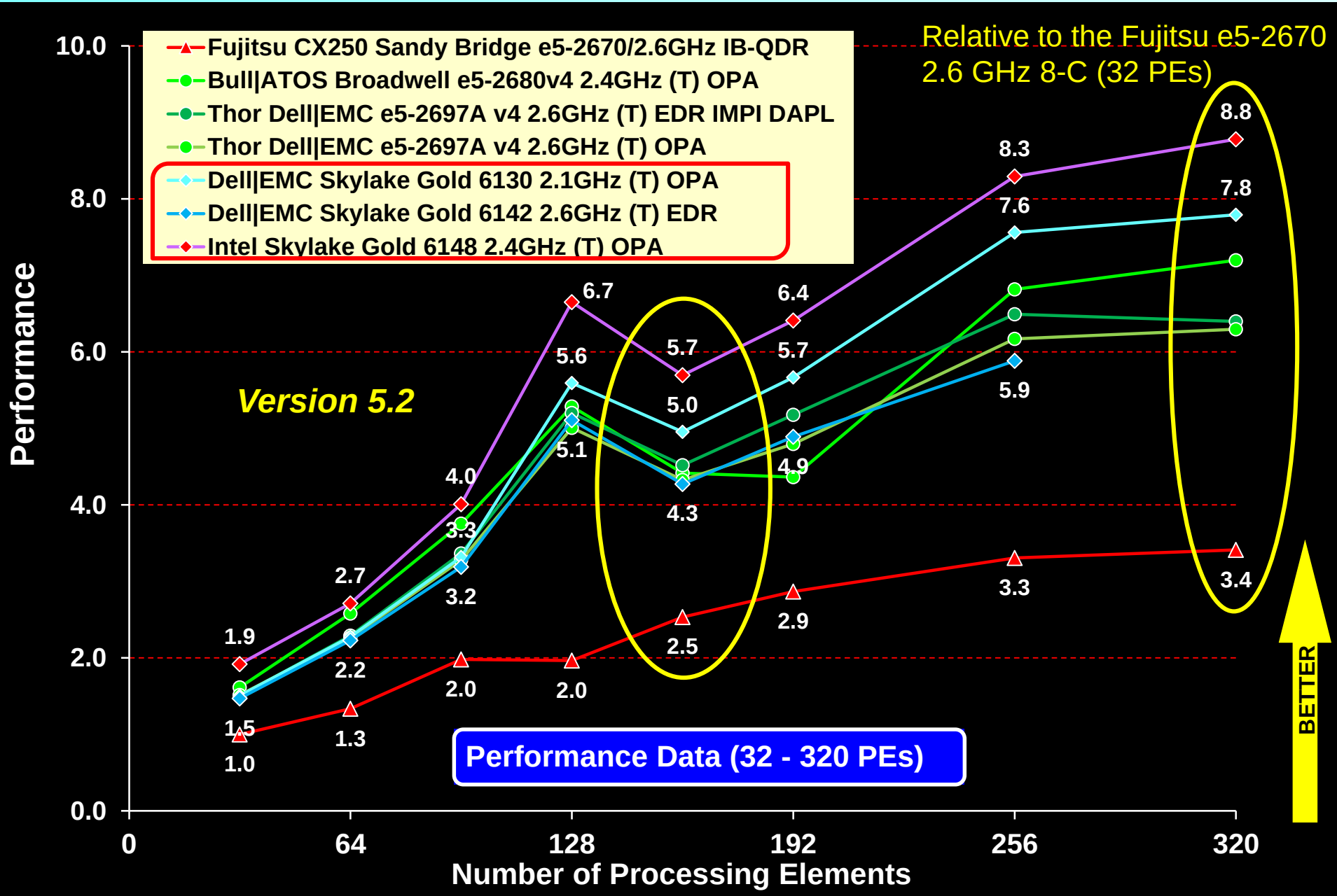
Quantum Transport.

Benchmark	Details
DEISA AU112	Au complex (Au_{112}), 2,158,381 G-vectors, 2 k-points, FFT dimensions: (180, 90, 288)
PRACE GRIR443	Carbon-Iridium complex ($C_{200}Ir_{243}$), 2,233,063 G-vectors, 8 k-points, FFT dimensions: (180, 180, 192)

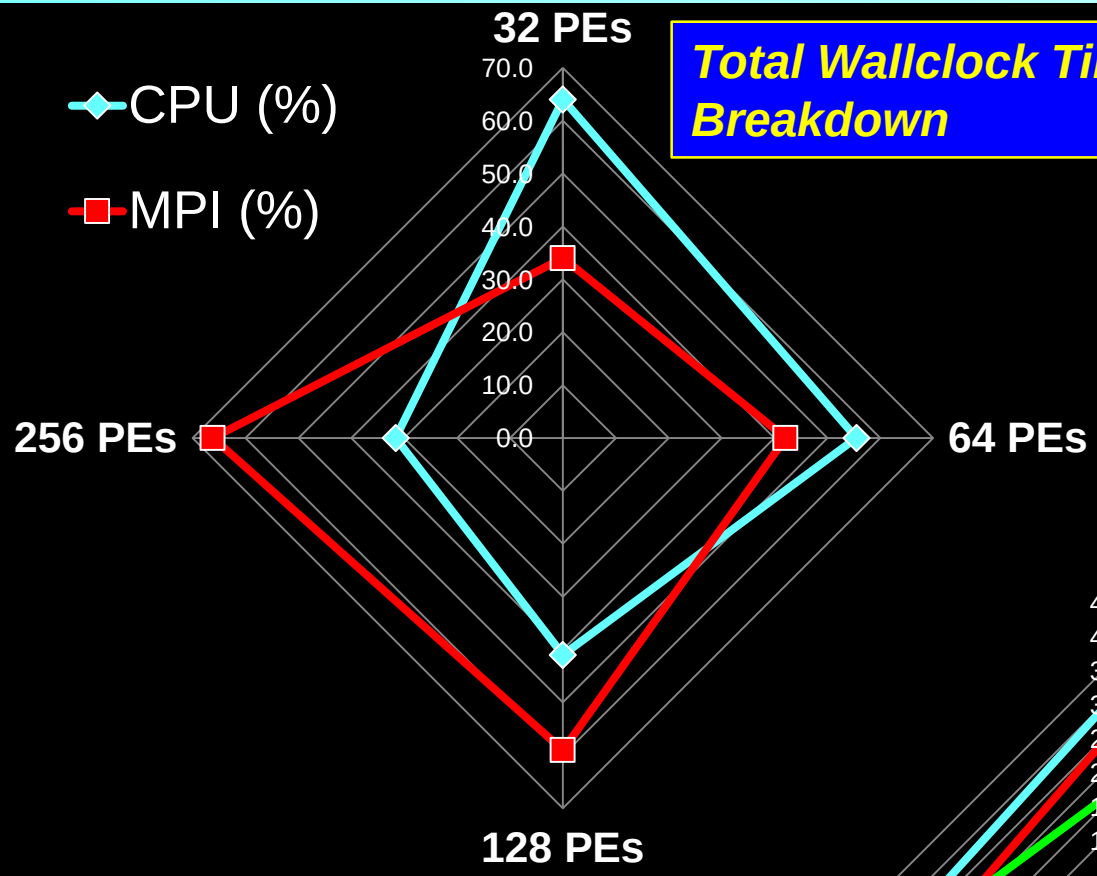
Quantum Espresso is an integrated suite of Open-Source computer codes for electronic-structure calculations and materials modelling at the nanoscale. It is based on density-functional theory, plane waves, and pseudopotentials.

Transition from v5.2 to v6.1

Quantum Espresso – Au_{112}

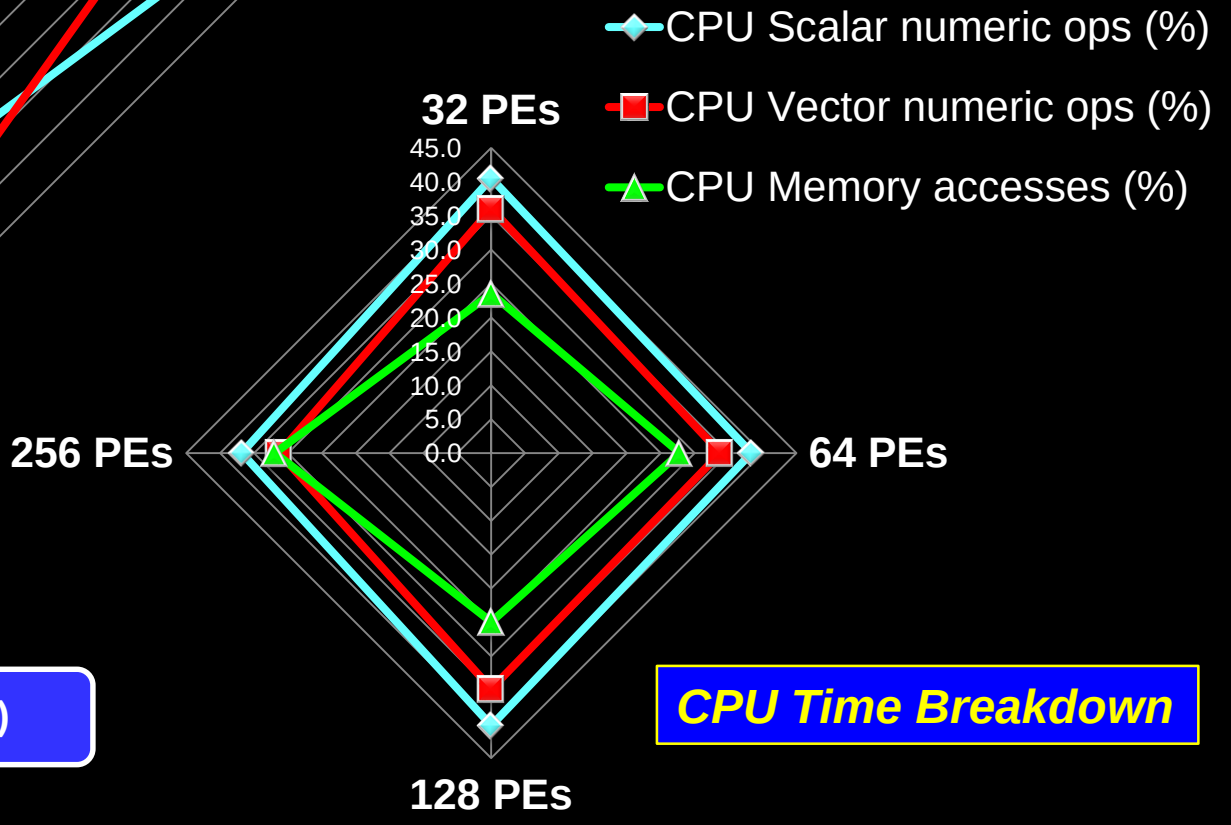


Quantum Espresso – Au_{112} Performance Report



Total Wallclock Time Breakdown

Au complex (Au_{112}), 2,158,381 G-vectors, 2 k-points, FFT dimensions: (180, 90, 288)



Performance Data (32-256 PEs)

CPU Time Breakdown

Parallelism in Quantum Espresso

- Quantum ESPRESSO implements several MPI parallelization levels, with Processors organized in a hierarchy of groups identified by different MPI communicator levels. **Group hierarchy**:
- **images:** Processors divided into different "images", corresponding to a different SCF or linear-response calculation, loosely coupled to others.
- **Pools and bands:** each image can be sub-partitioned into "pools", each taking care of a group of k-points. Each pool is sub-partitioned into "band groups", each taking care of a group of Kohn-Sham orbitals.
- **PW Parallelisation:** orbitals in the PW basis set, as well as charges and density in either reciprocal or real space, distributed across processors. All **linear-algebra operations** on array of PW / real-space grids are automatically and effectively parallelized.
- **tasks:** Allows for good parallelization of the 3D FFT when no. of CPUs exceeds the no. of FFT planes, FFTs on Kohn-Sham states are redistributed to "task".

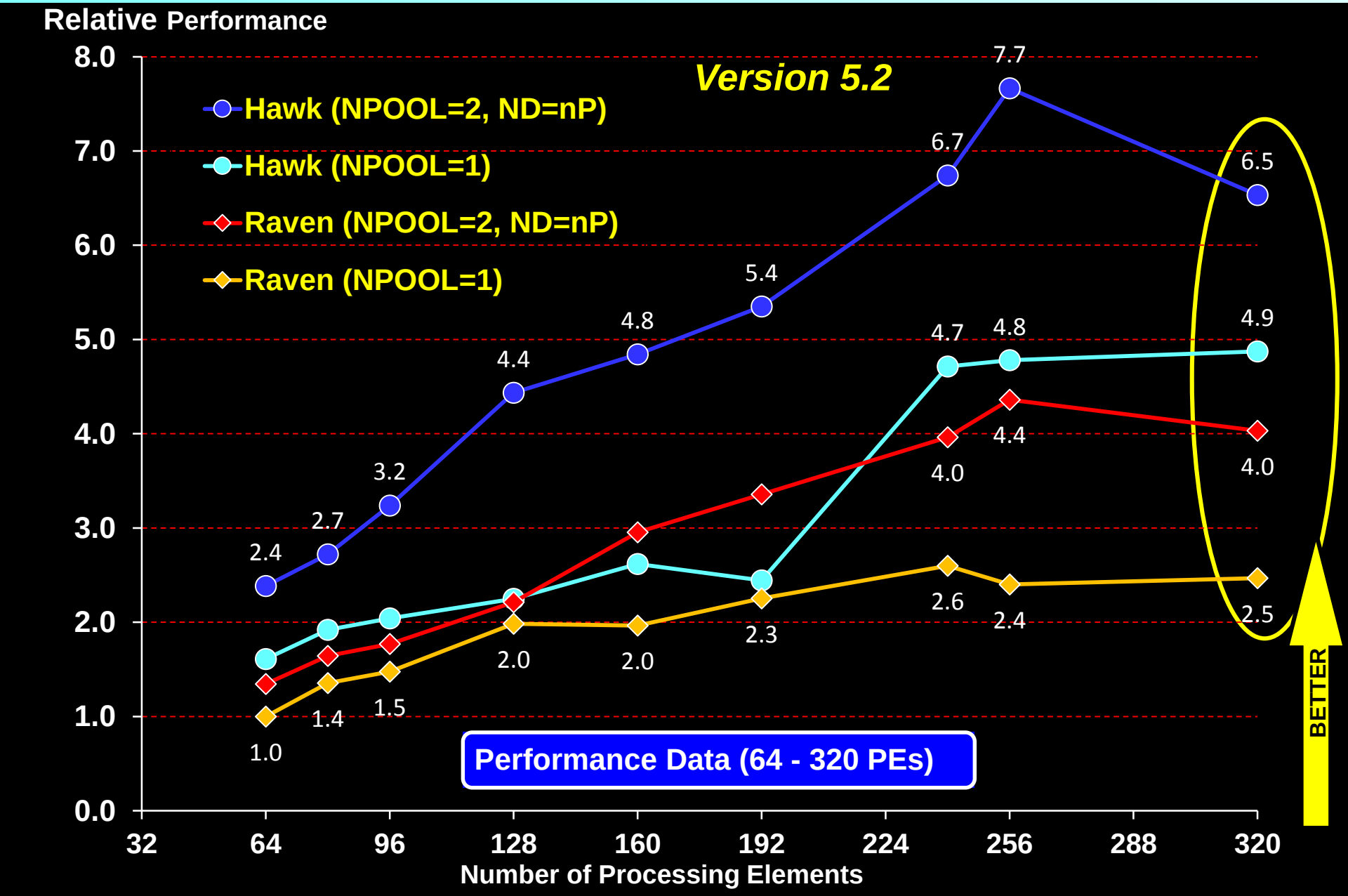
Parallelism in Quantum Espresso

- **linear-algebra group:** A further level, independent on PW or k-point parallelization, is the parallelization of subspace diagonalization / iterative orthonormalization.
- **About communications** Images and pools are loosely coupled and CPUs communicate between different images and pools only once in a while, whereas CPUs within each pool are tightly coupled and communications are significant.
- **Choosing parameters :** To control the no. CPUs in each group, command line switches: -nimage, -npools, -nband, -ntg, -ndiag or –northo. Thus for Au_{112}, use is of the following command line:

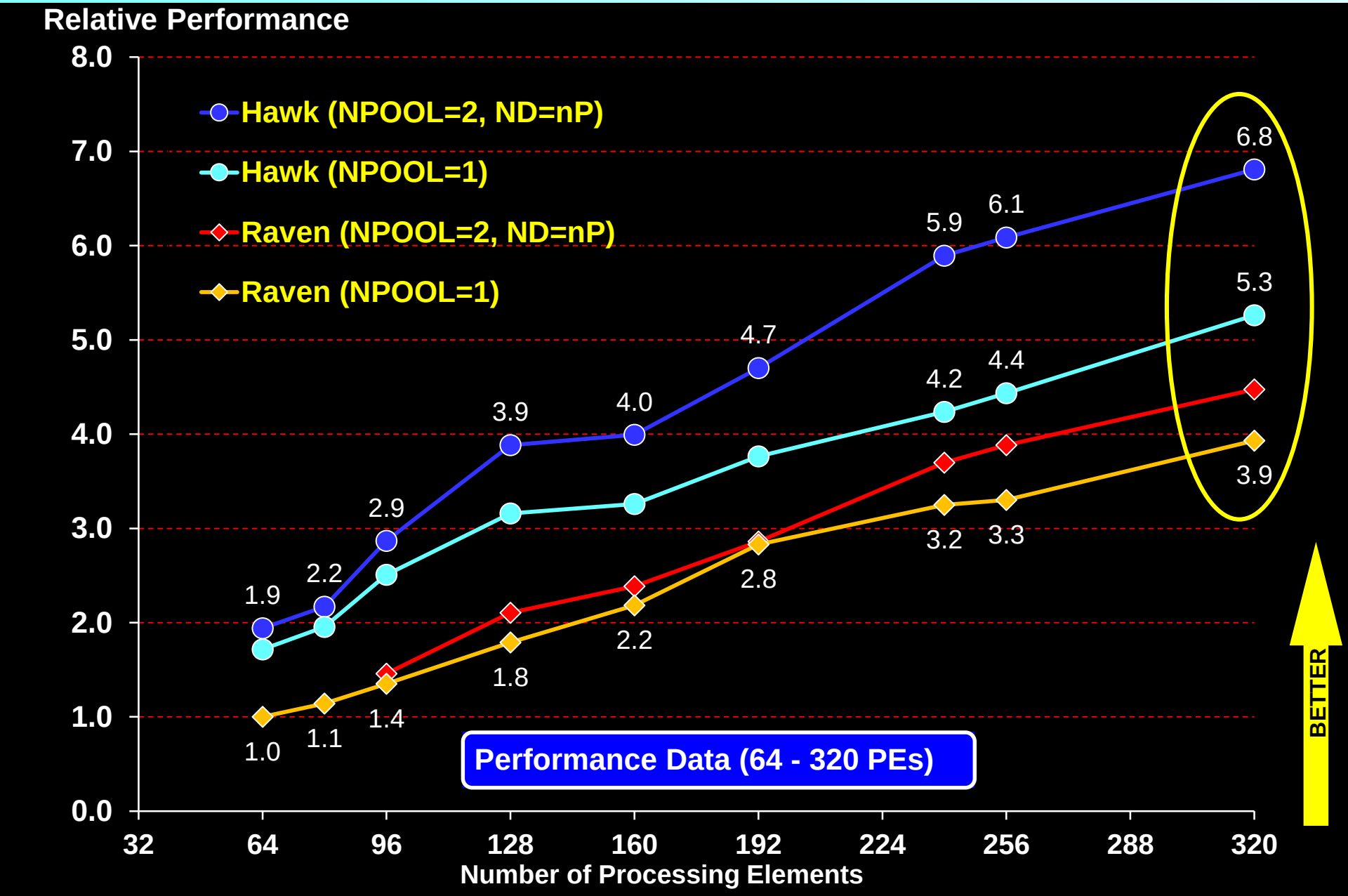
*mpirun $code -inp ausurf.in -npool **$NPOOL** -ntg **$NT** -ndiag **$ND***

- This executes an energy calculation on **$NP** processors, with k-points distributed across **$NPOOL** pools of *$NP/$NPOOL* processors each, 3D FFT is performed using **$NT** task groups, with the diagonalization of the subspace Hamiltonian distributed to a square grid of **$ND** processors.

Impact of npool – Au_{112}

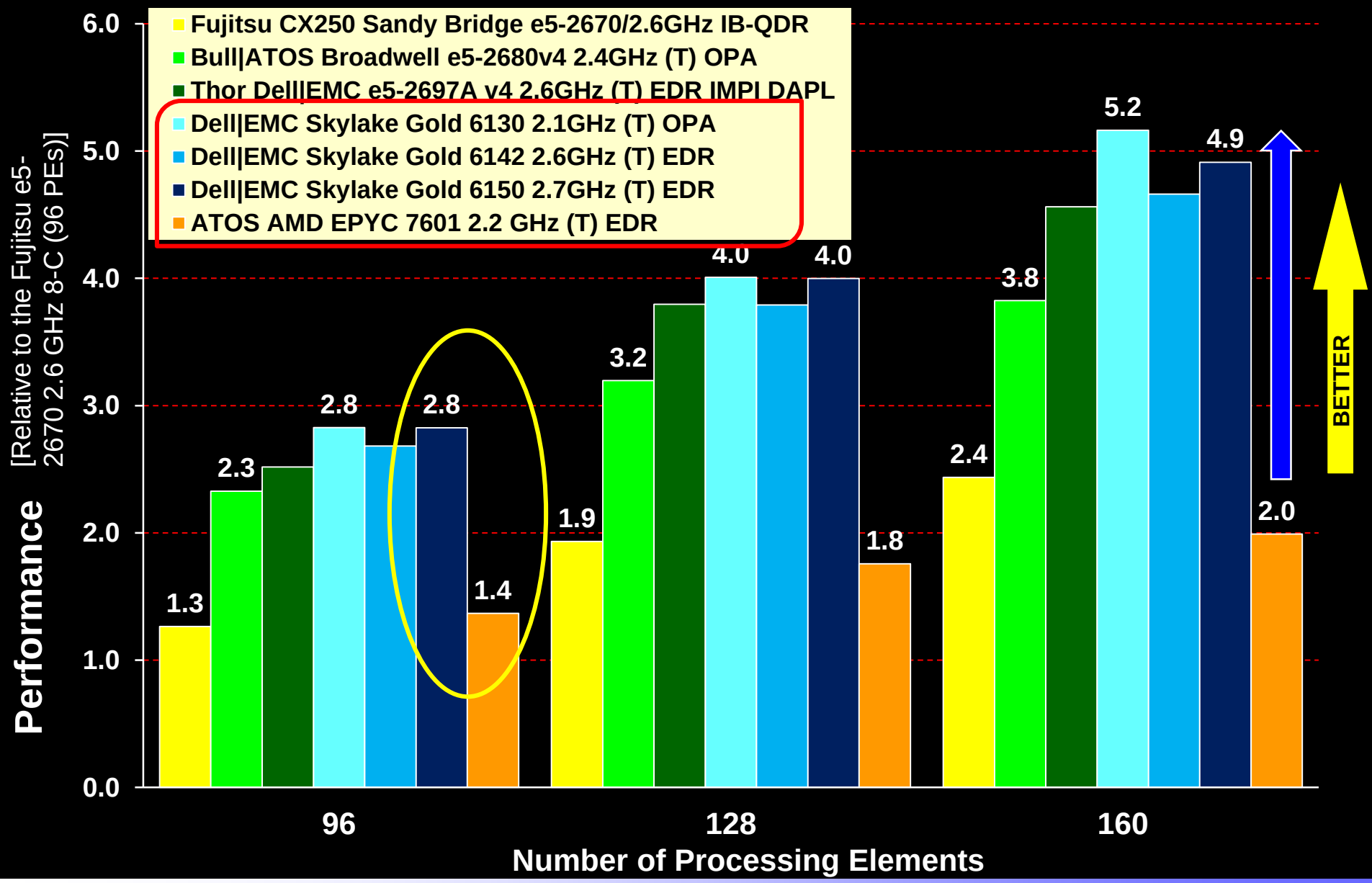


Impact of npool – GRIR443

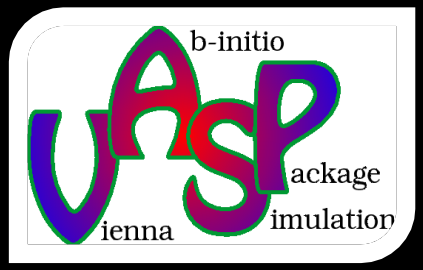


Quantum Espresso – GRIR443

Performance Data (96-160 PEs)



VASP – Vienna *Ab-initio* Simulation Package



VASP (5.4.4) 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

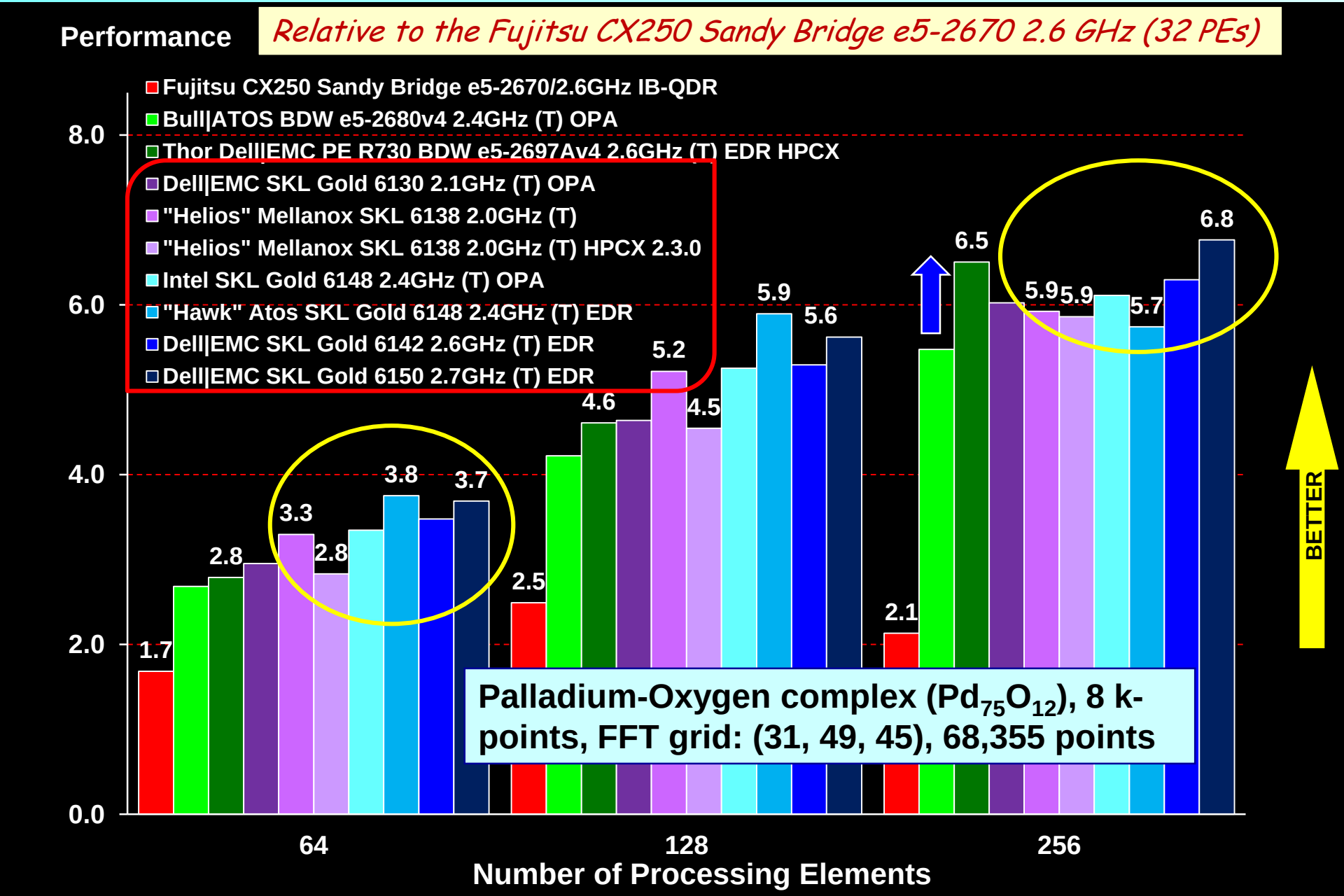
Pd-O Benchmark

- 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,

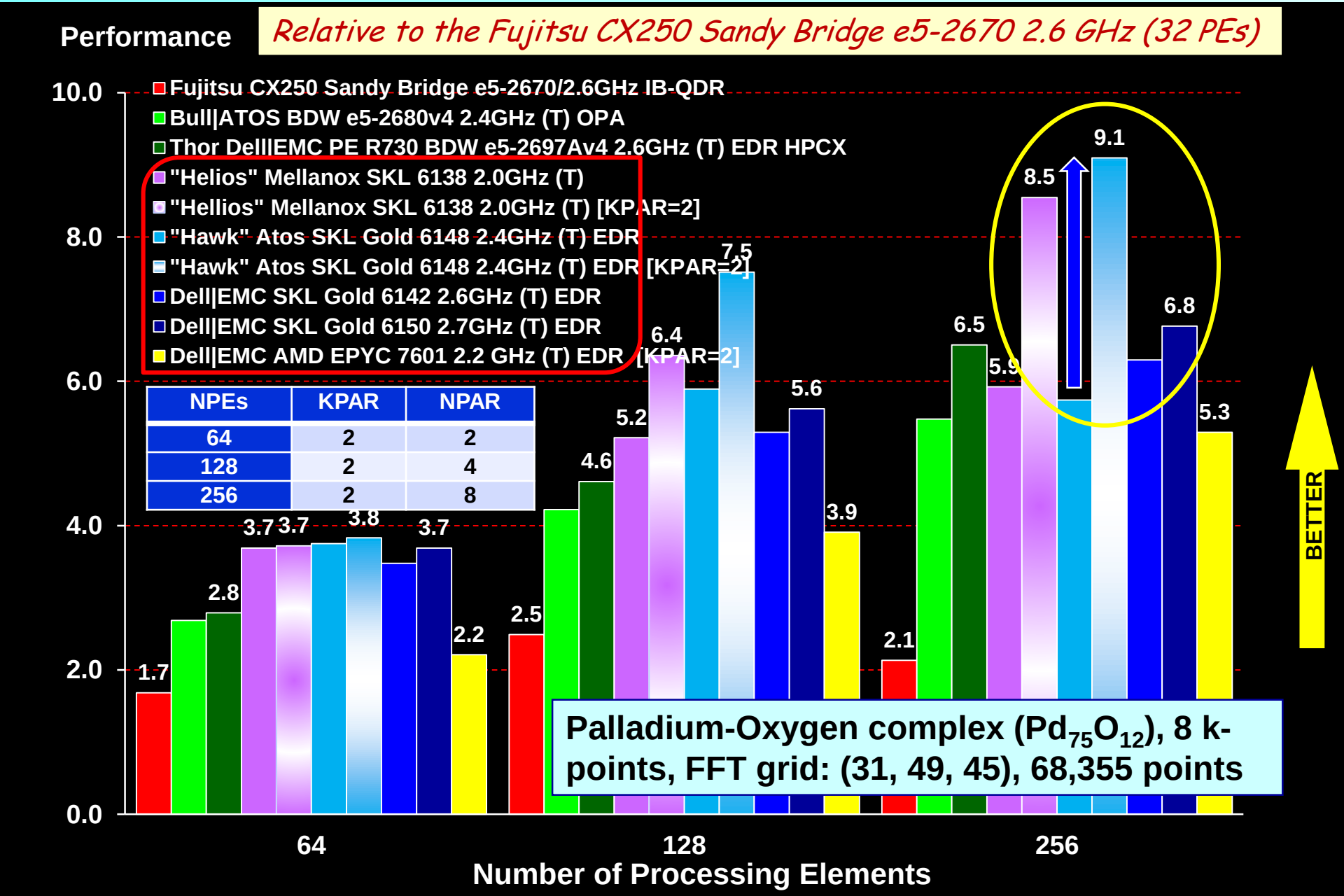
Zeolite Benchmark

- 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,

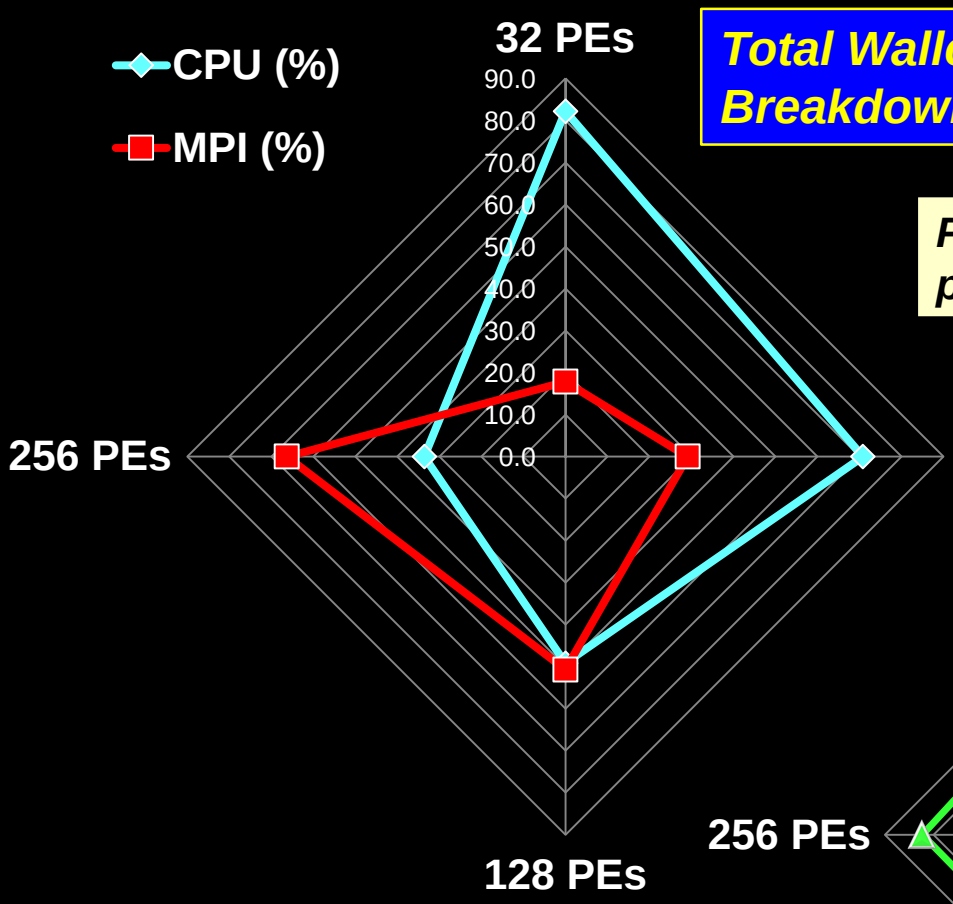
VASP 5.4.4 – Pd-O Benchmark



VASP 5.4.4 – Pd-O Benchmark - Parallelisation on k-points

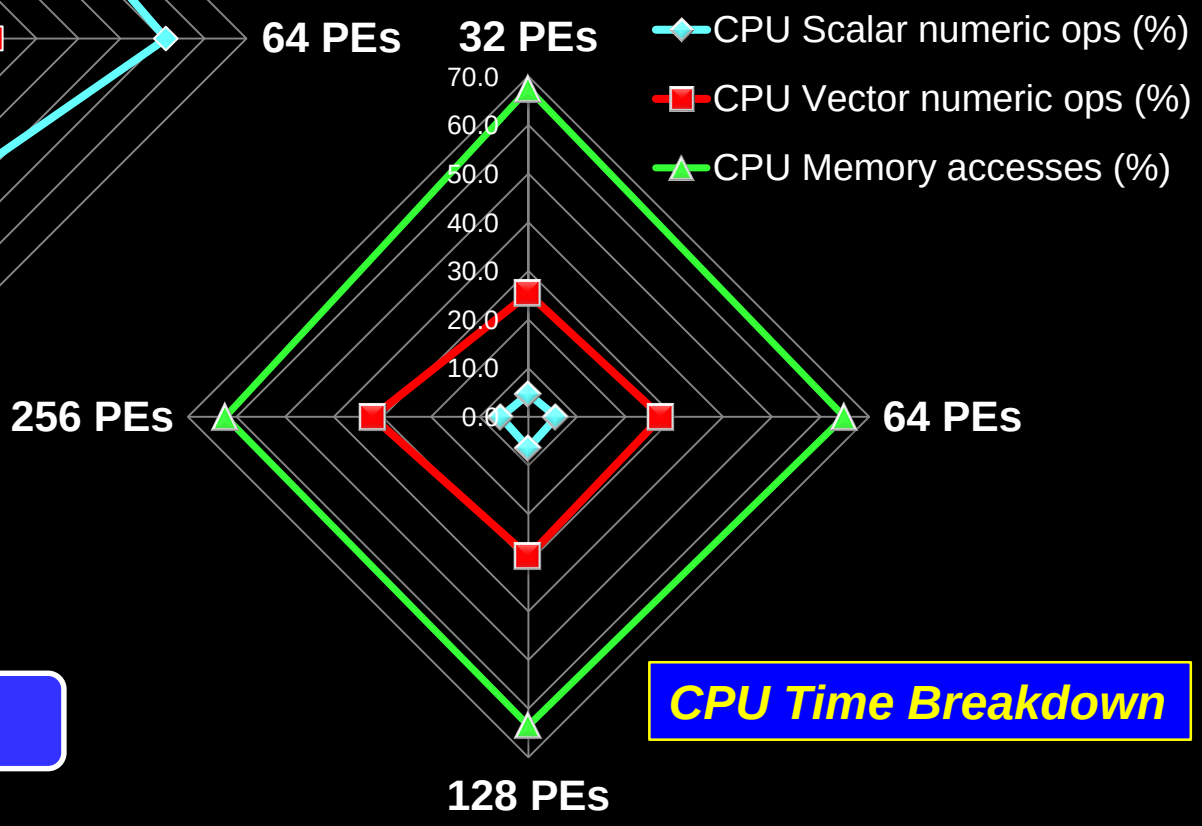


VASP – Pd-O Benchmark Performance Report



Total Wallclock Time Breakdown

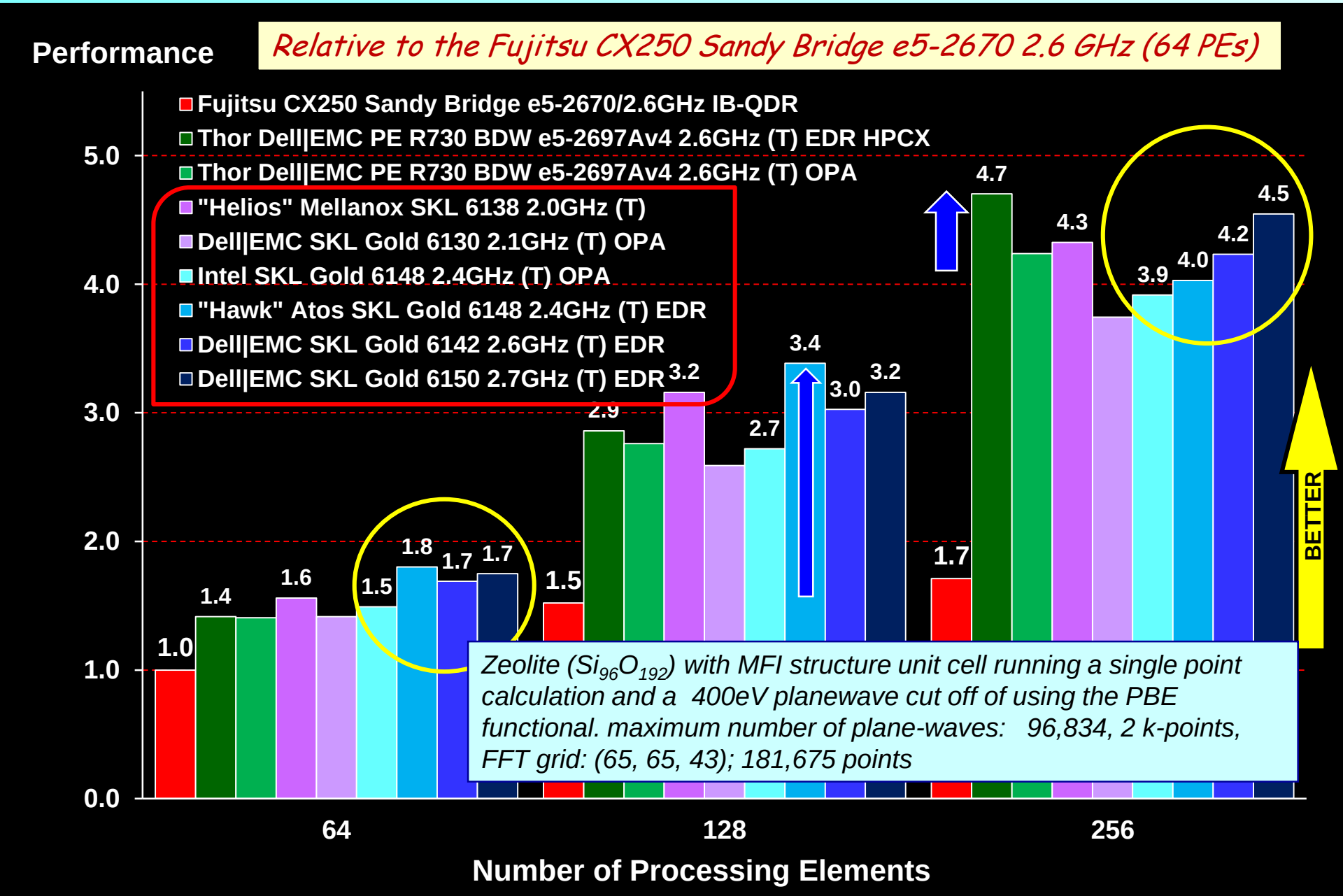
Palladium-Oxygen complex ($Pd_{75}O_{12}$), 8 k-points, FFT grid: (31, 49, 45), 68,355 points



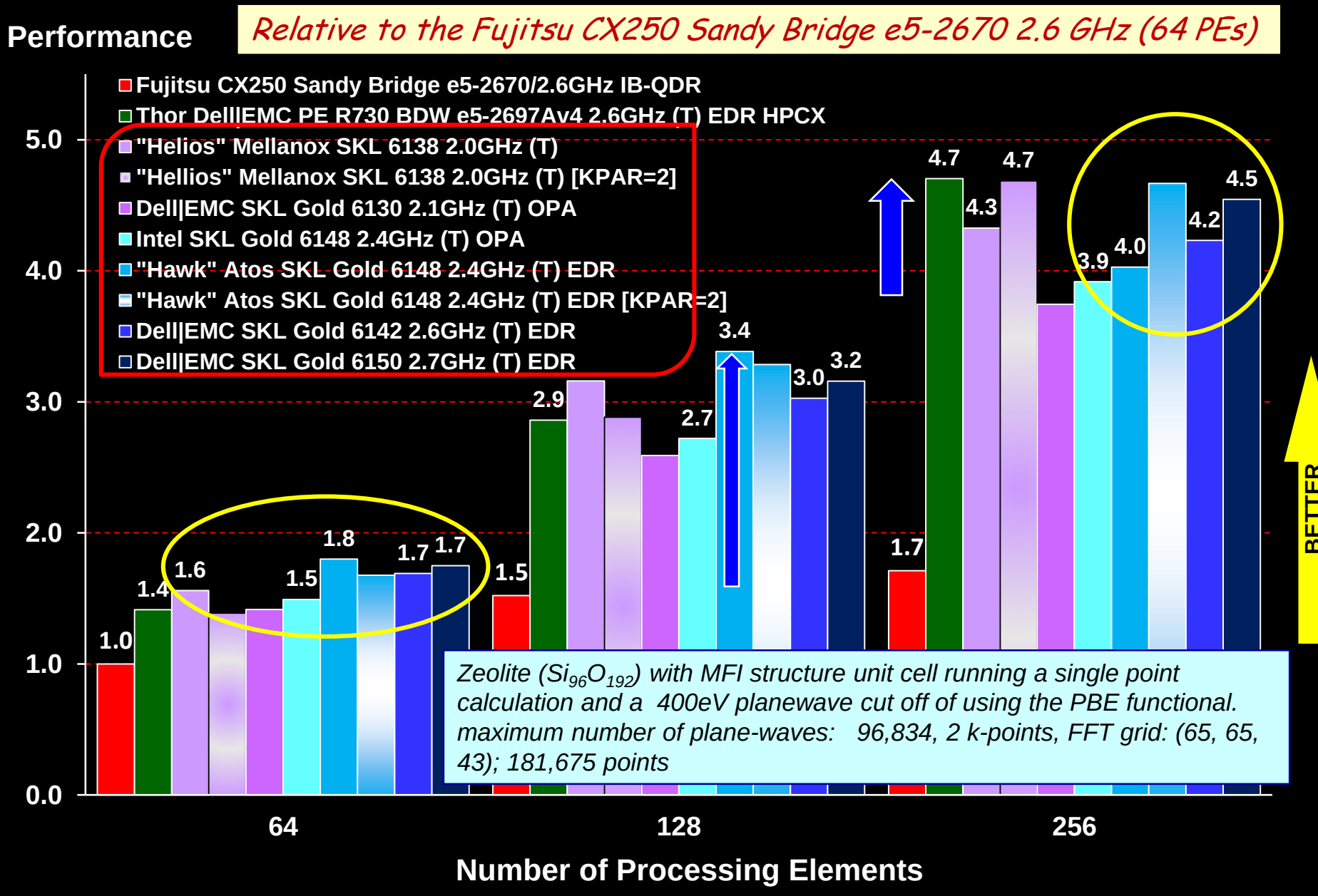
Performance Data (32-256 PEs)

CPU Time Breakdown

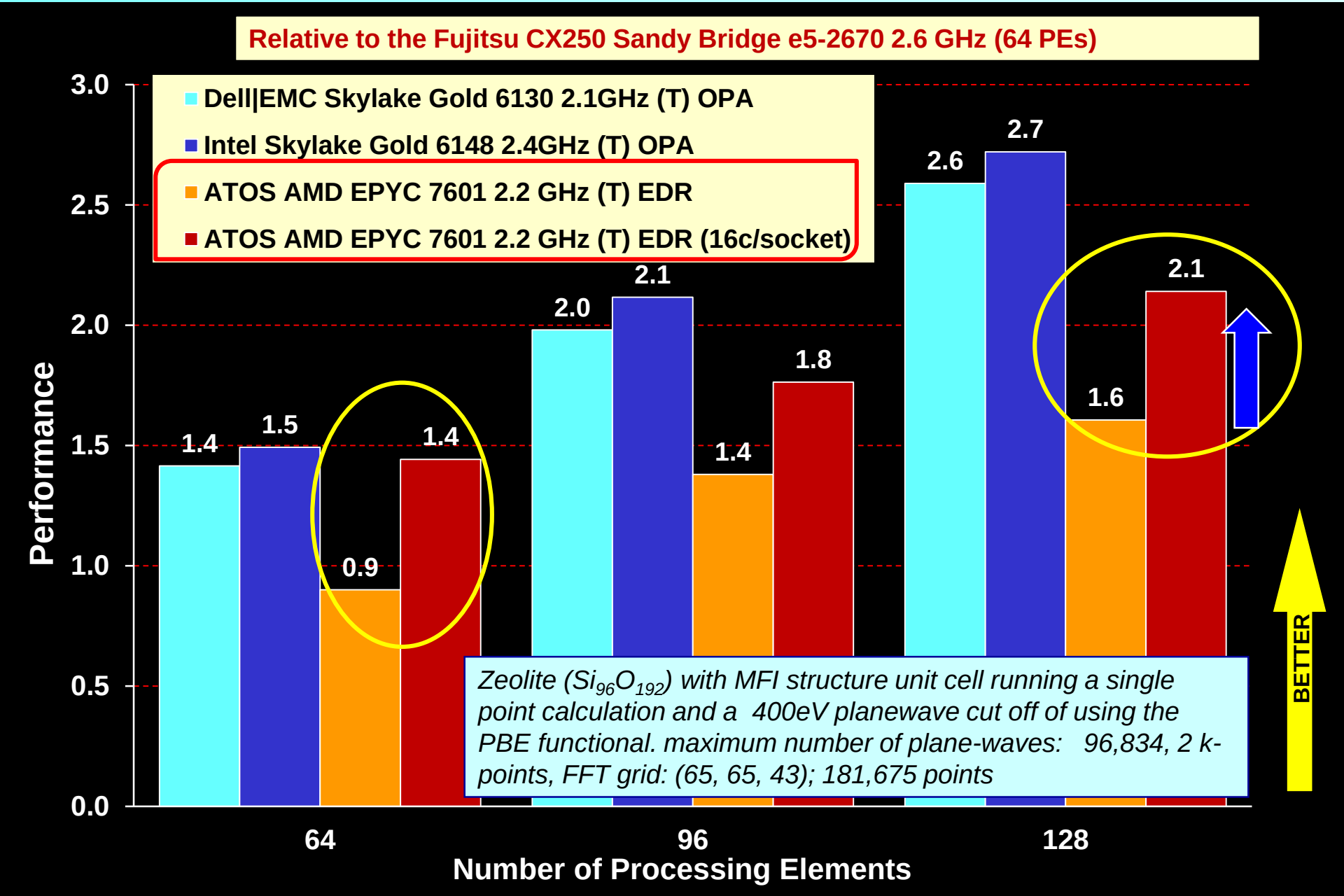
VASP 5.4.4 – Zeolite Benchmark



VASP 5.4.4 – Zeolite Benchmark - Parallelisation on k-points



VASP 5.4.1 – Zeolite Benchmark

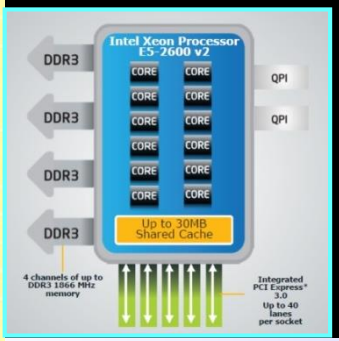


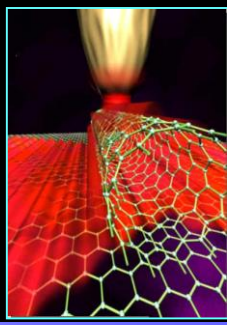
CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Application Performance on Multi-core Processors:

I.2. Selecting Fabrics and Optimising Performance:

Intel MPI and Mellanox HPCX





Selecting Fabrics – MPI Optimisation

- **Intel MPI Library** - can select a communication fabric at runtime without having to recompile the application. By default, it automatically selects the most appropriate fabric based on both S/W and H/W configuration i.e. in most cases you do not have to manually select a fabric.
- Specifying a particular fabric can boost performance. Can specify fabrics for both intra-node and inter-node communications. Following fabrics available:

Fabric	Network hardware and software used
shm	Shared memory (for intra node communication only).
dapl	Direct Access Programming Library (**DAPL**) fabrics, such as InfiniBand (IB) and iWarp (through DAPL).
ofa	OpenFabrics Alliance (**OFA**) fabrics e.g. InfiniBand (through OFED verbs).
tcp	TCP/IP network fabrics, such as Ethernet and InfiniBand (through IPoIB).
tmi	Tag Matching Interface (**TMI**) fabrics, such as Intel True Scale Fabric, Intel Omni Path Architecture and Myrinet (through TMI).
ofi	OpenFabrics Interfaces* (**OFI**) - capable fabrics, such as Intel True Scale Fabric, Intel Omni Path Architecture, IB and Ethernet (through OFI API).

- For inter-node communication, it uses the first available fabric from the default fabric list. List is defined automatically for each H/W and S/W configuration (see I_MPI_FABRICS_LIST).
- For most configurations, this list is as follows: dapl, ofa, tcp, tmi, ofi

Mellanox HPC-X Toolkit

The **Mellanox HPC-X Toolkit** provides a MPI, SHMEM and UPC software suite for HPC environments. Delivers "*enhancements to significantly increase the scalability & performance of message communications in the network*". Includes:

- Complete MPI, SHMEM, UPC package, including Mellanox MXM and FCA acceleration engines
- *Offload collectives communication from MPI process onto Mellanox interconnect hardware*
- Maximize application performance with underlying hardware architecture. Optimized for Mellanox InfiniBand and VPI interconnects
- Increase application scalability and resource efficiency
- Multiple transport support including RC, DC and UD
- Intra-node shared memory communication

- *Performance comparison conducted on the* ***Mellanox SKL 6138 / 2.00 GHz EDR based "Helios" cluster***

Application Performance & MPI Libraries

Performance comparison exercise undertaken to capture the impact of the latest release of Intel MPI and Mellanox's HPCX.

- In 2017, on the Mellanox HP Proliant- E5-2697A v4 EDR based **Thor** cluster, comparison of Intel MPI and Mellanox HPCX for the following applications (and associated data sets).
  - DLPOLY4 (NaCl and Gramicidin) & GROMACS (Ion Channel and lignocellulose)
  - VASP (PdO Complex & Zeolite System)
  - Quantum ESPRESSO (Au_{112} and GRIR443)
  - OpenFOAM (Cavity 3D-3M)
- Simply compared the time to solution for each application i.e.

  $$T_{HPCX} / T_{Intel\text{-}MPI}$$

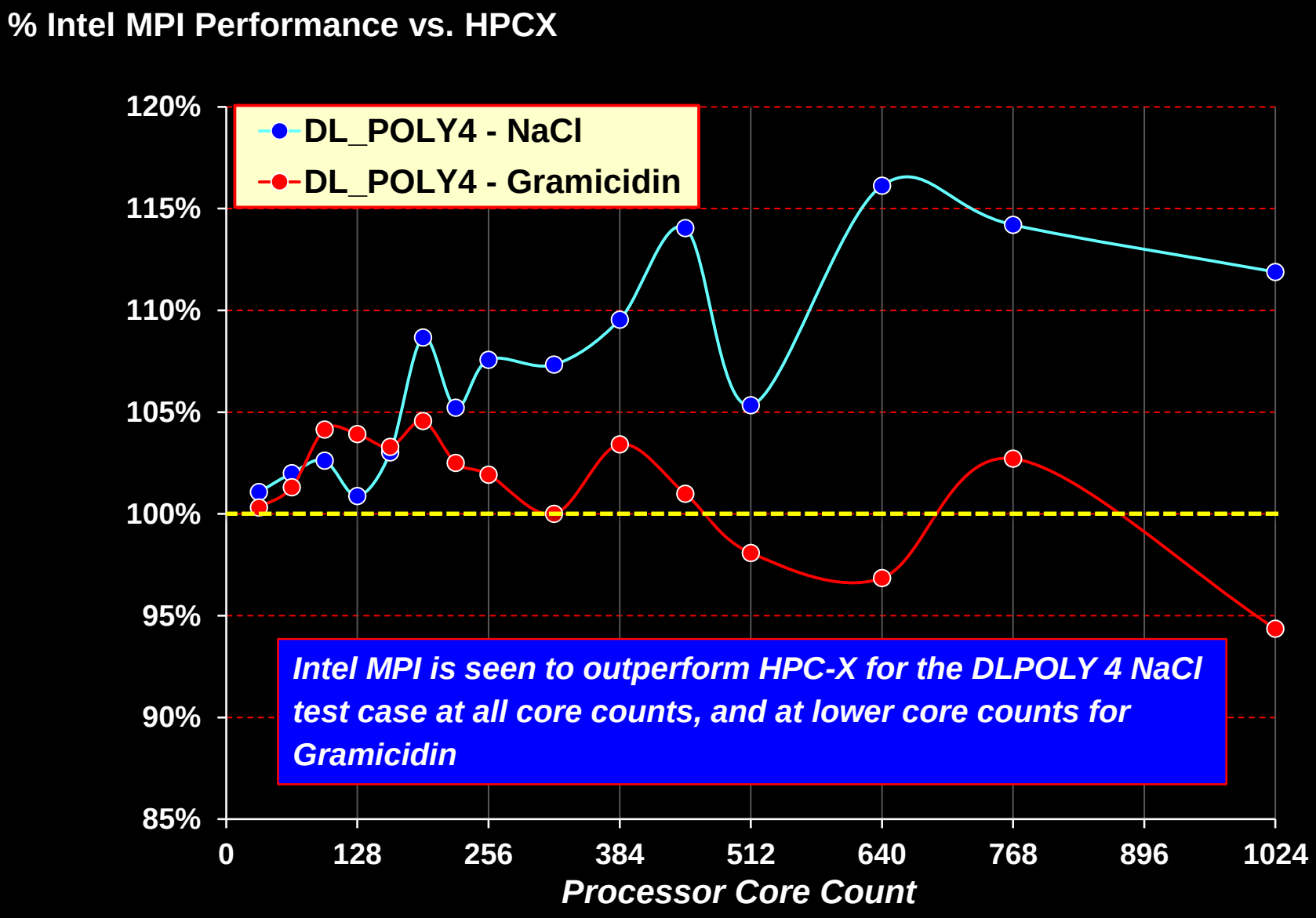
  across multiple core counts

Application Performance & MPI Libraries

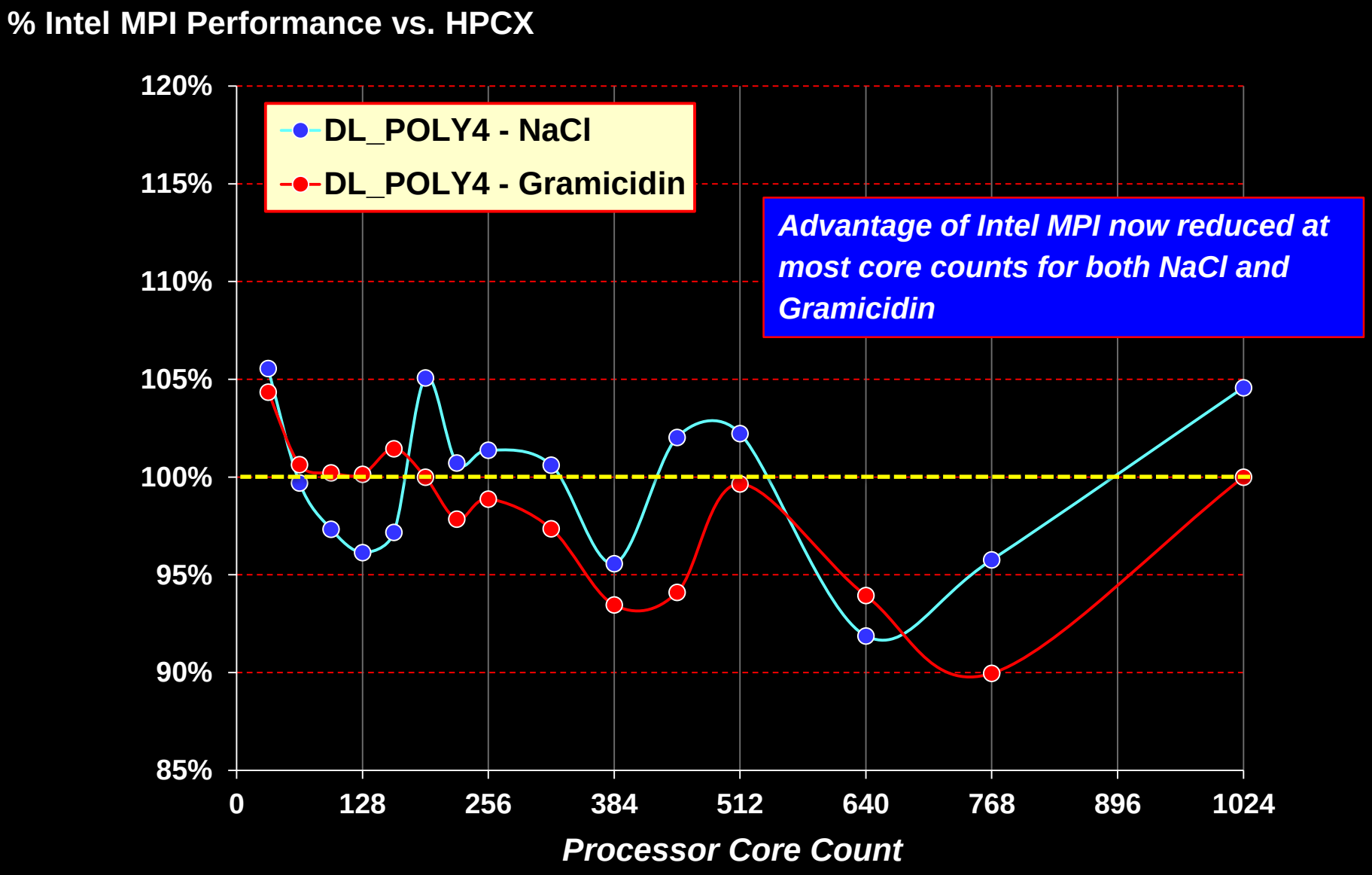
- **Optimum performance** found to be a function of both application and core count.
  - With the **materials-based codes & OpenFOAM**, and at high core count (> 512 cores), **HPCX exhibited a clear performance advantage over Intel MPI**.
  - This was not the case for the **classical MD codes** where **Intel MPI showed a distinct advantage** at all but the highest core counts.
- Repeated the exercise on the Helios partition of the Skylake cluster using latest releases of HPCX v2.2.0 and 2.3.0-pre

http://www.mellanox.com/related-docs/prod_acceleration_software/PB_HPC-X.pdf

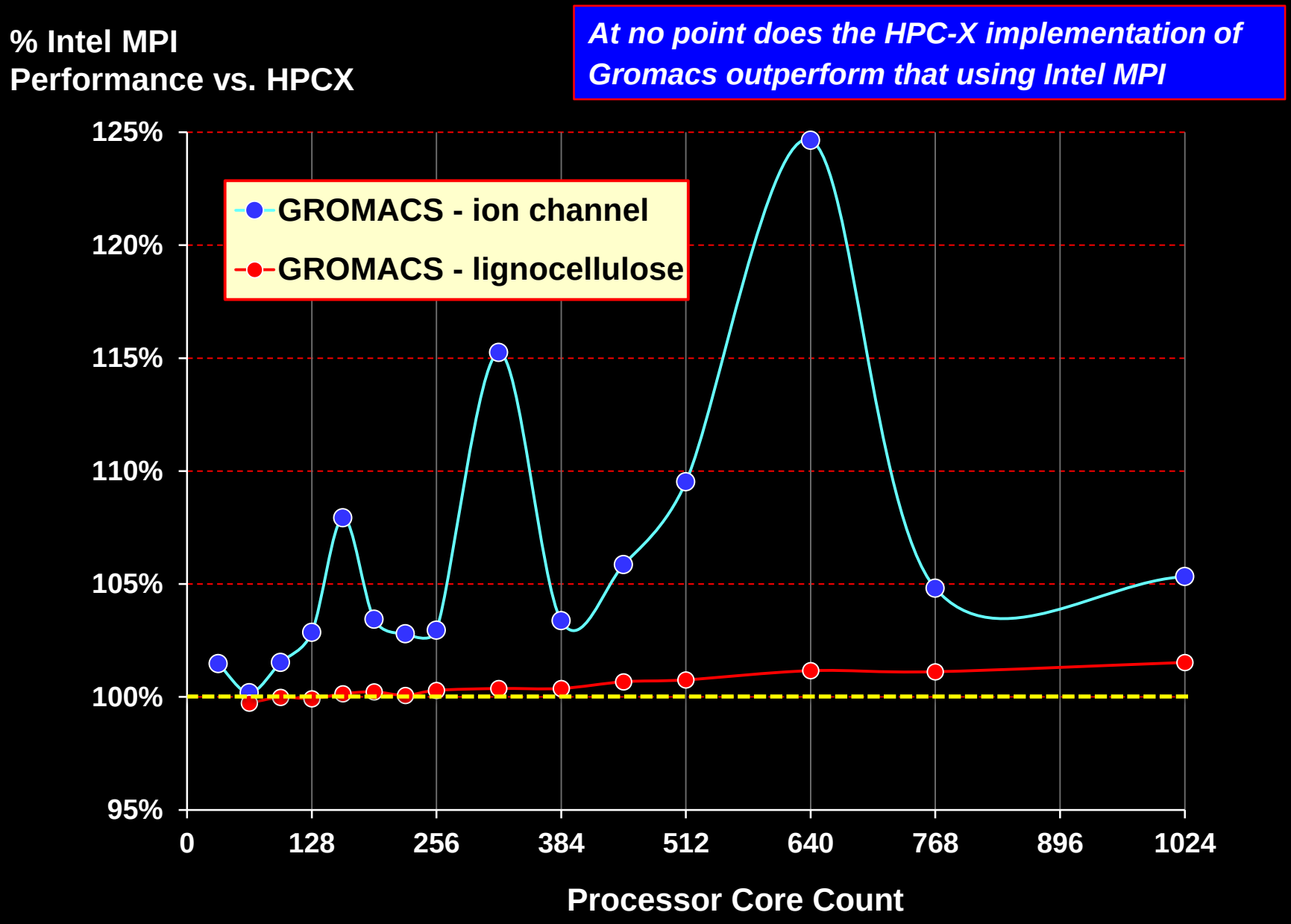
DL_POLY 4 – Intel MPI vs. HPCX – December 2017



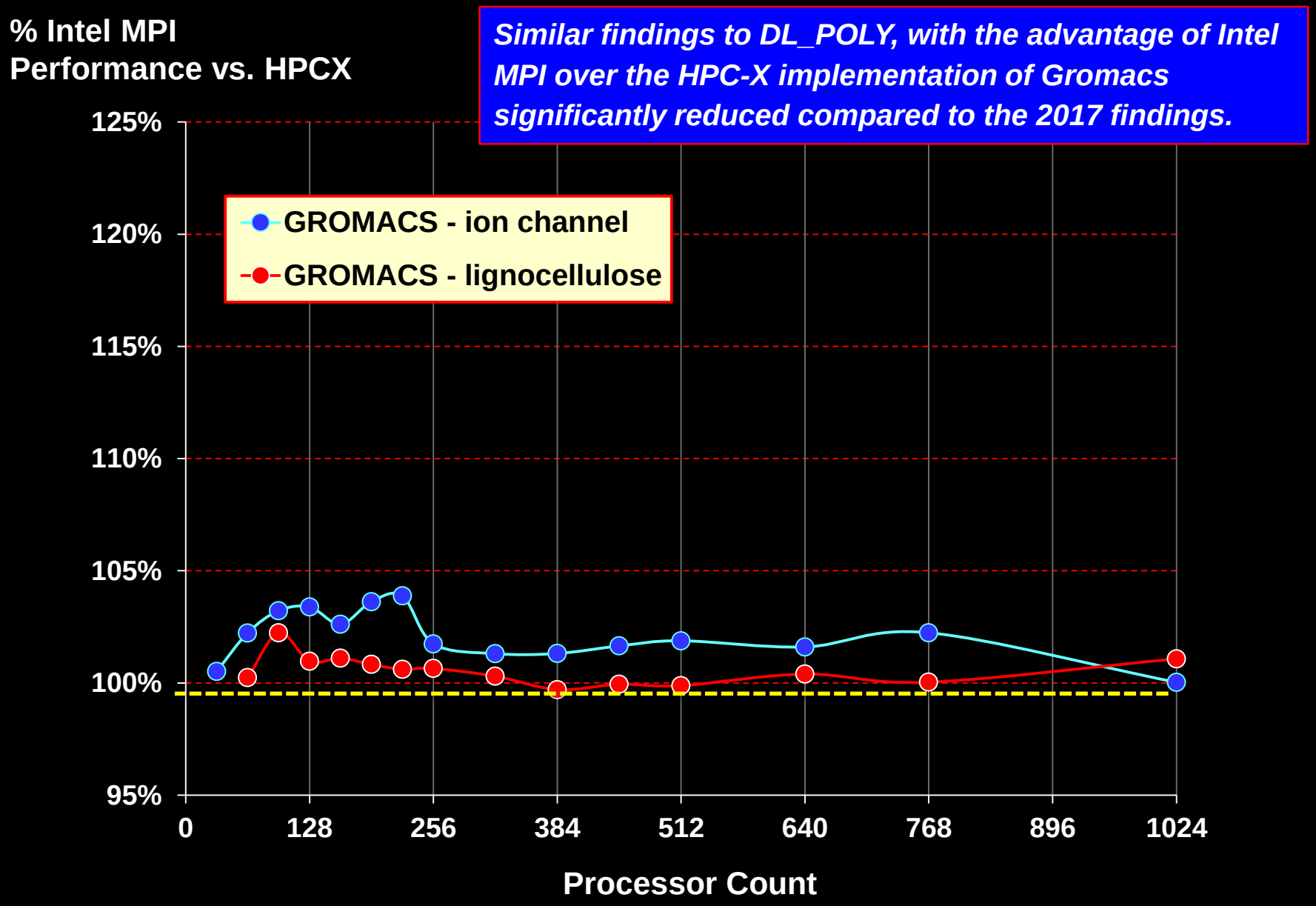
DL_POLY 4 – Intel MPI vs. HPCX – December 2018



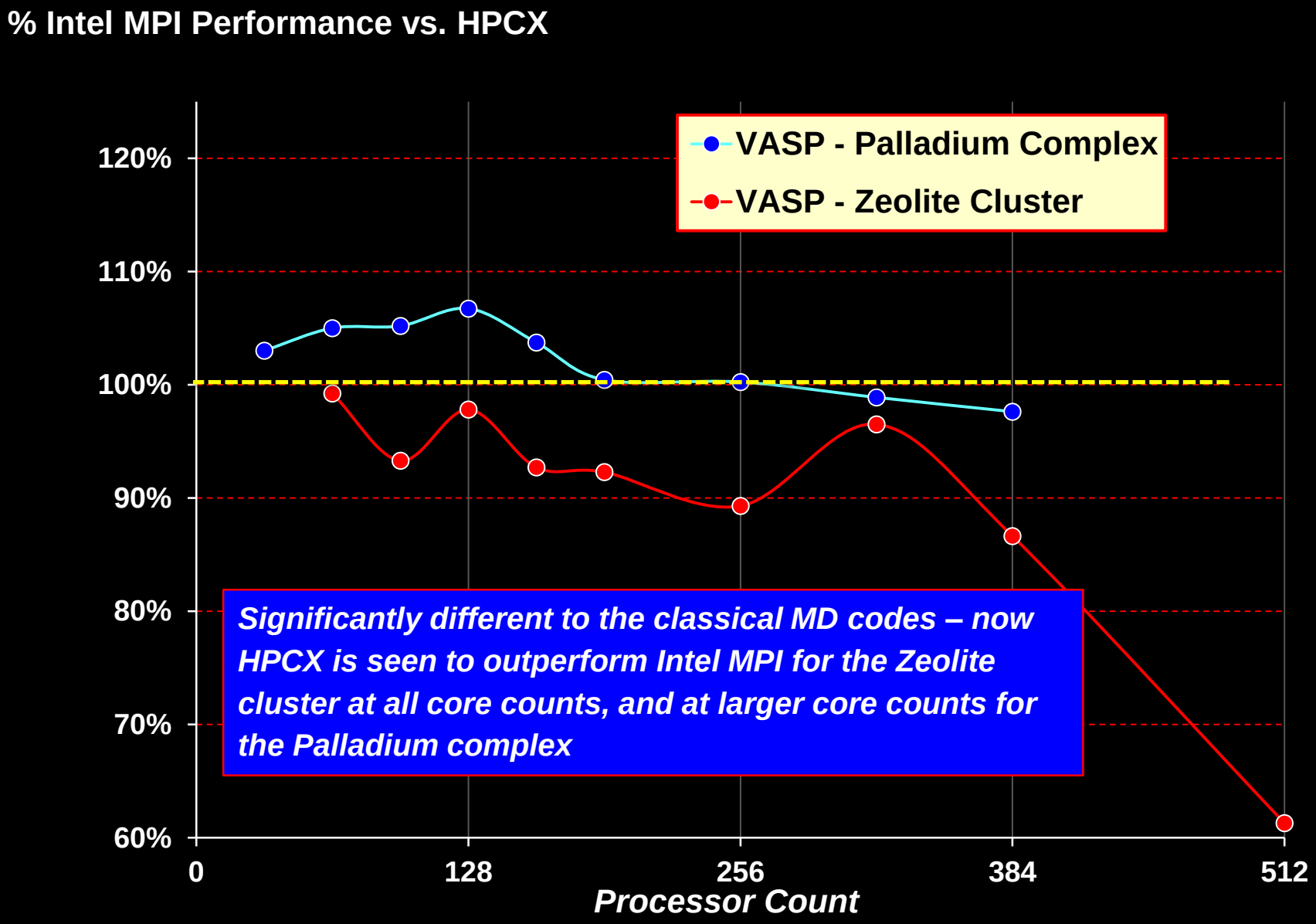
GROMACS – Intel MPI vs. HPCX – December 2017



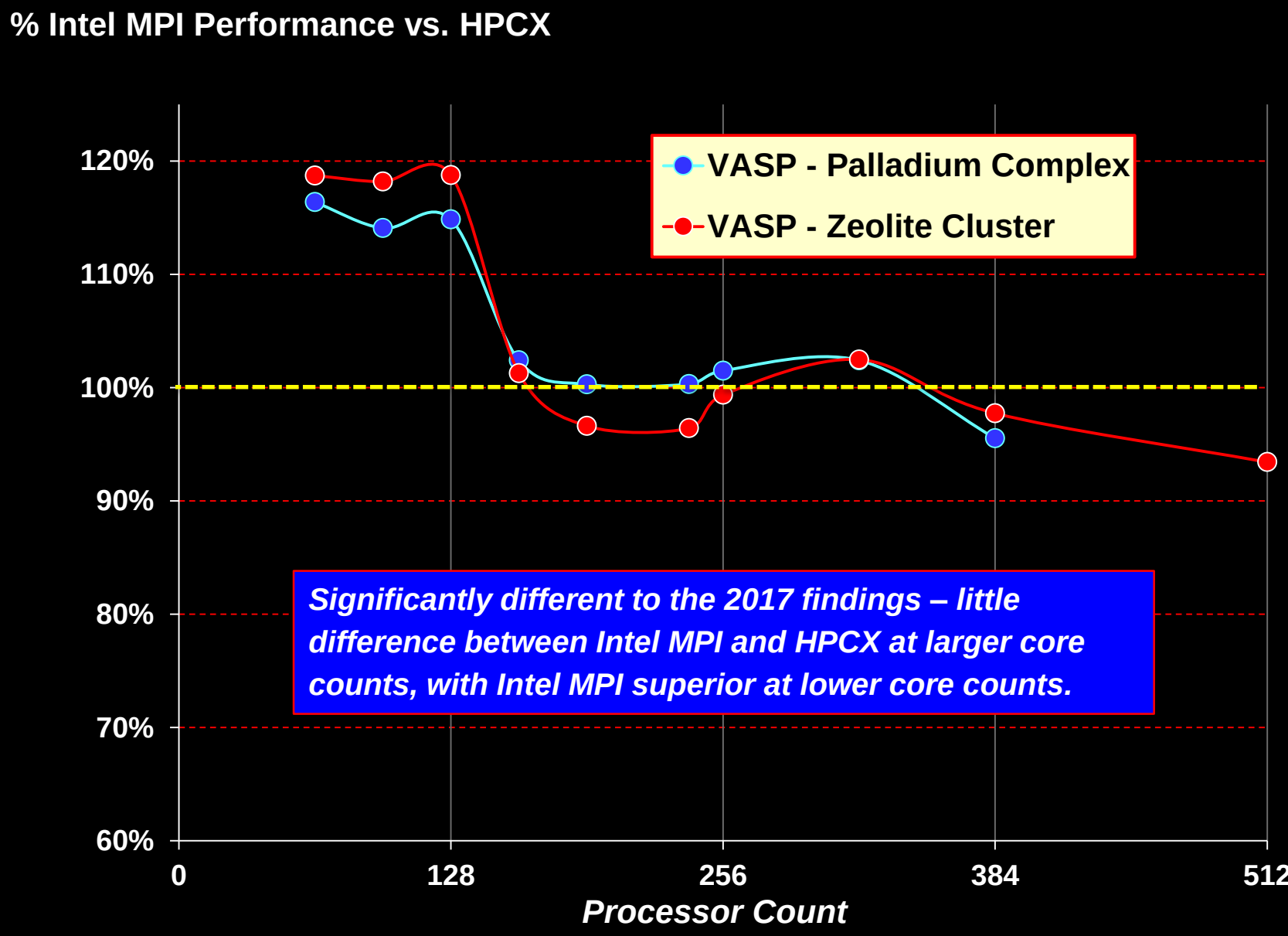
GROMACS – Intel MPI vs. HPCX – December 2018



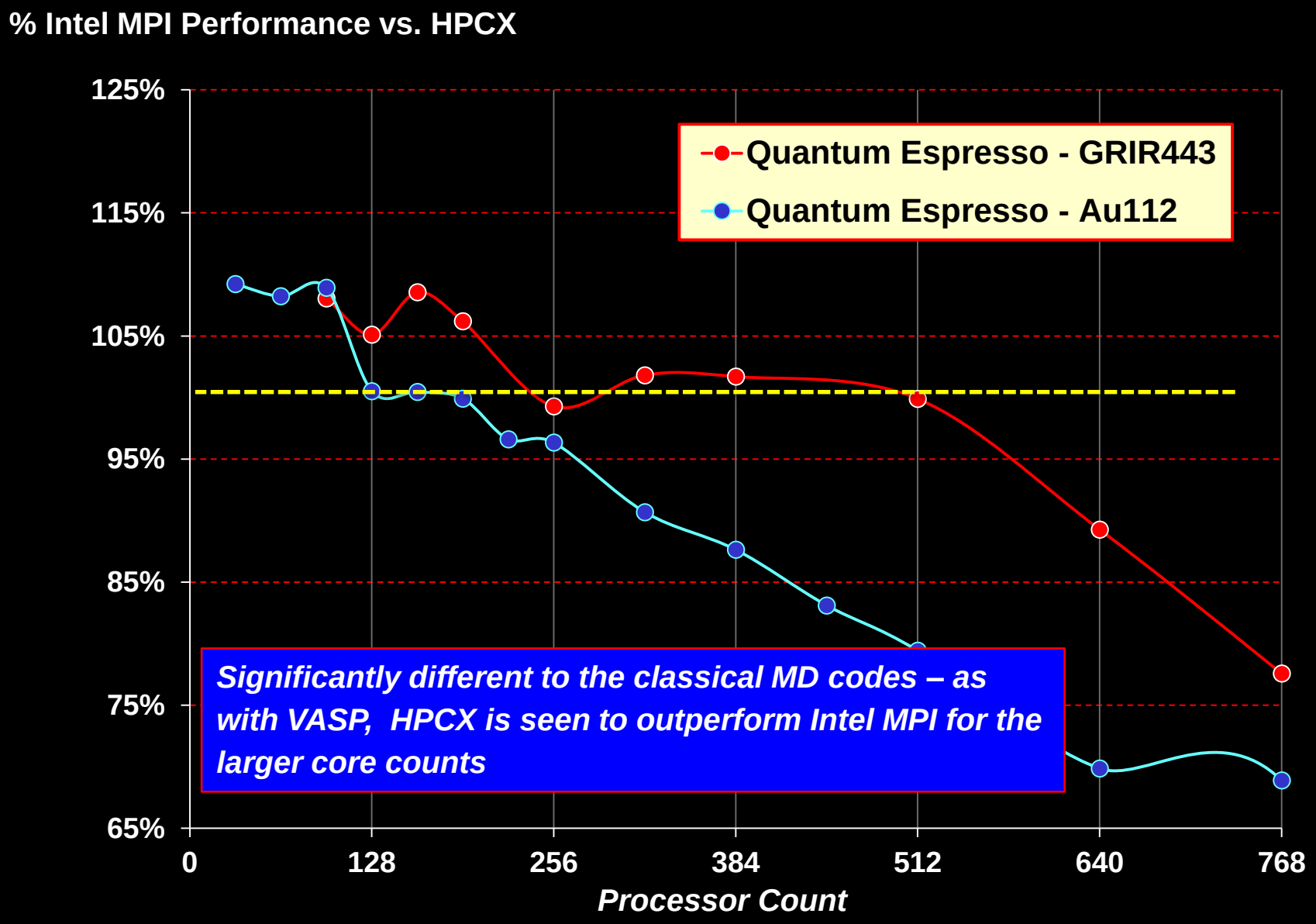
VASP 5.4.1 – Intel MPI vs. HPCX – December 2017



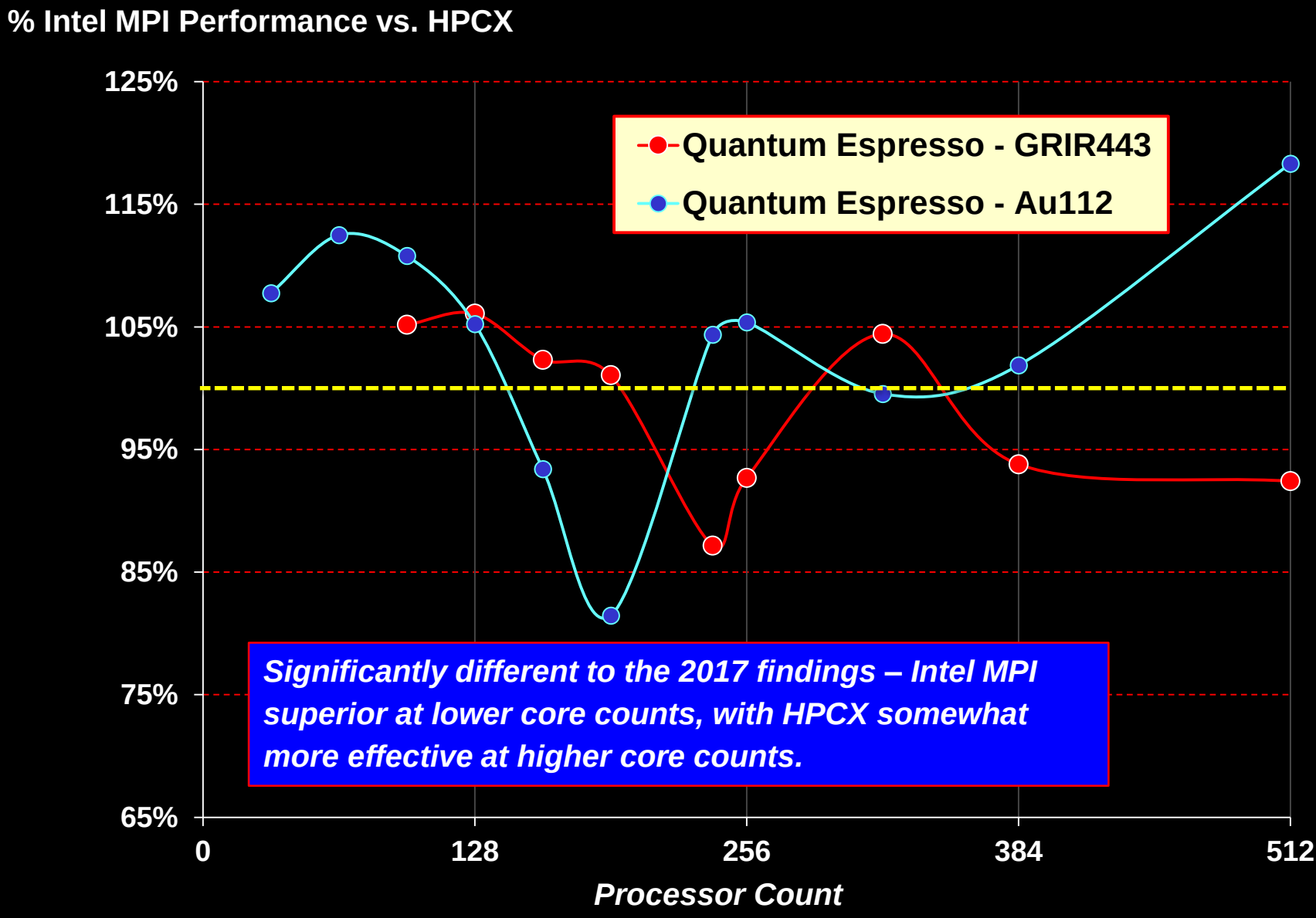
VASP 5.4.4 – Intel MPI vs. HPCX – December 2018



Quantum Espresso v5.2 – Intel MPI vs. HPCX – Dec. 2017



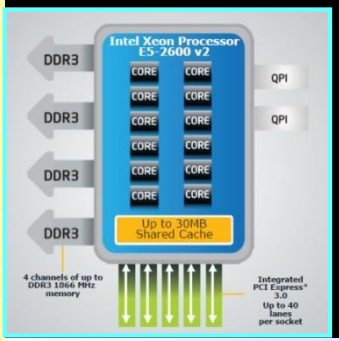
Quantum Espresso v6.1 – Intel MPI vs. HPCX – Dec. 2018

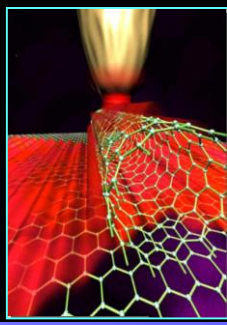


CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

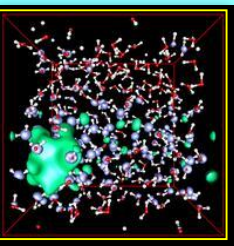
Application Performance on Multi-core Processors

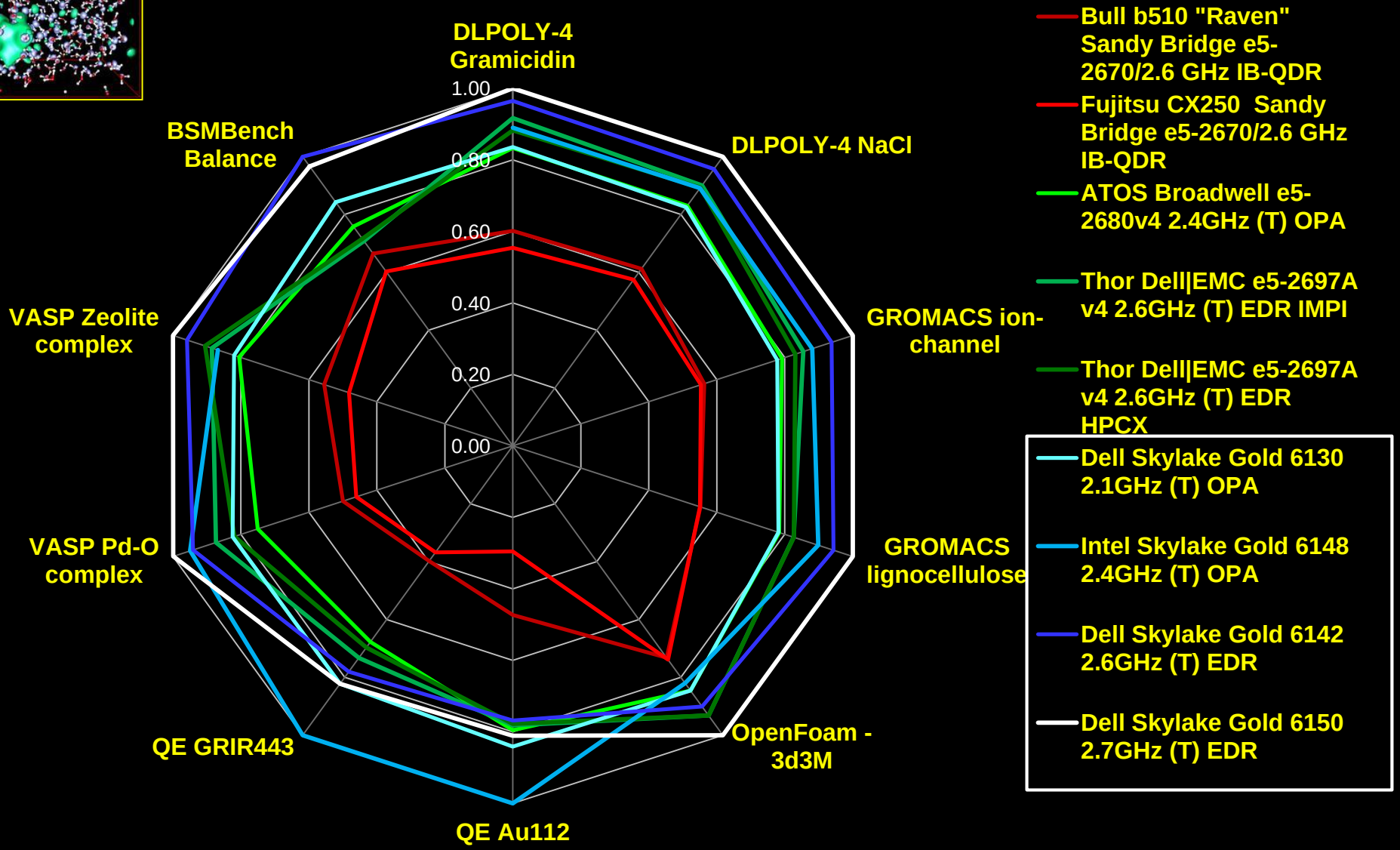
I.3 Relative Performance as a Function of Processor Family and Interconnect – SKL and SNB Clusters.



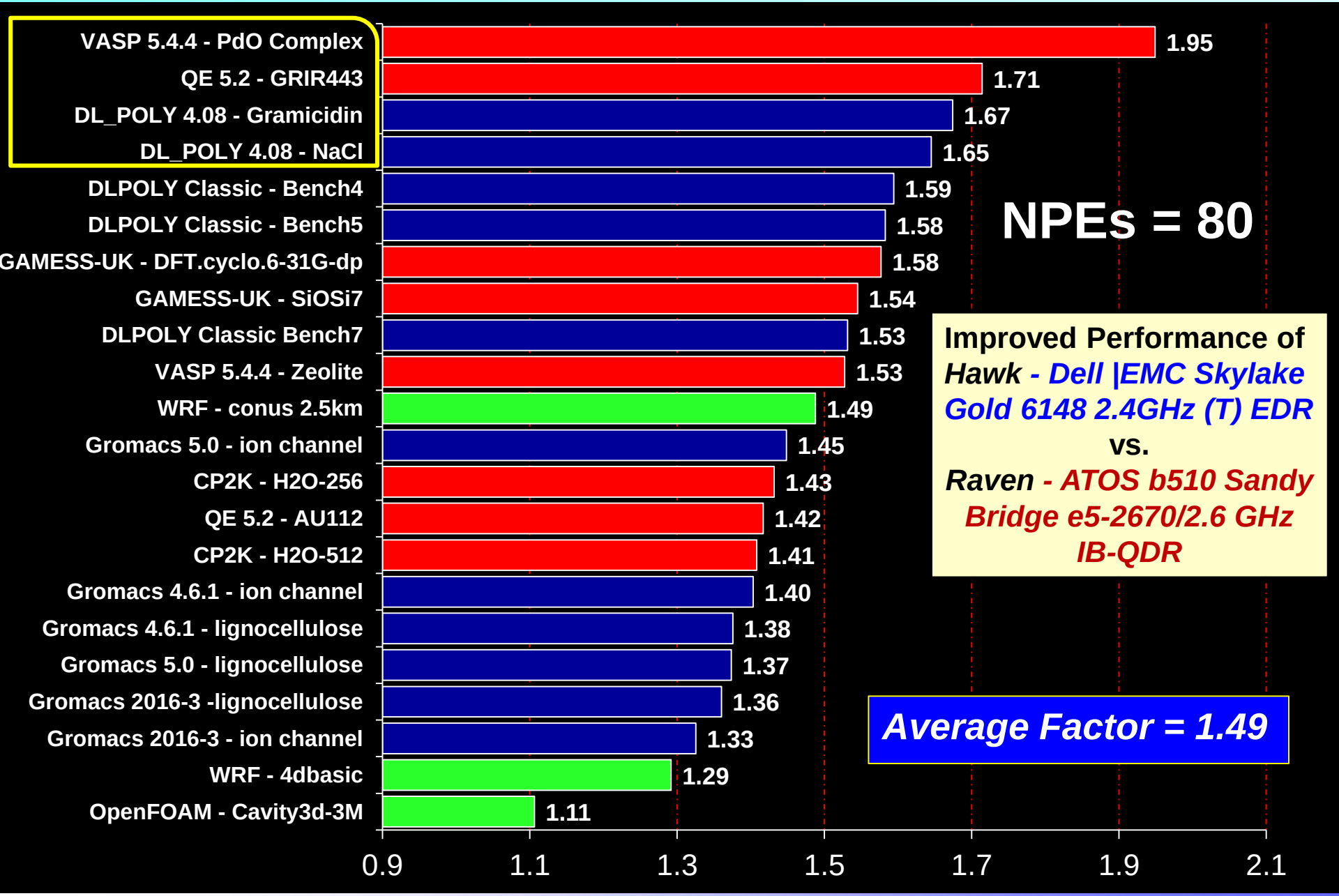


Target Codes and Data Sets – 128 PEs

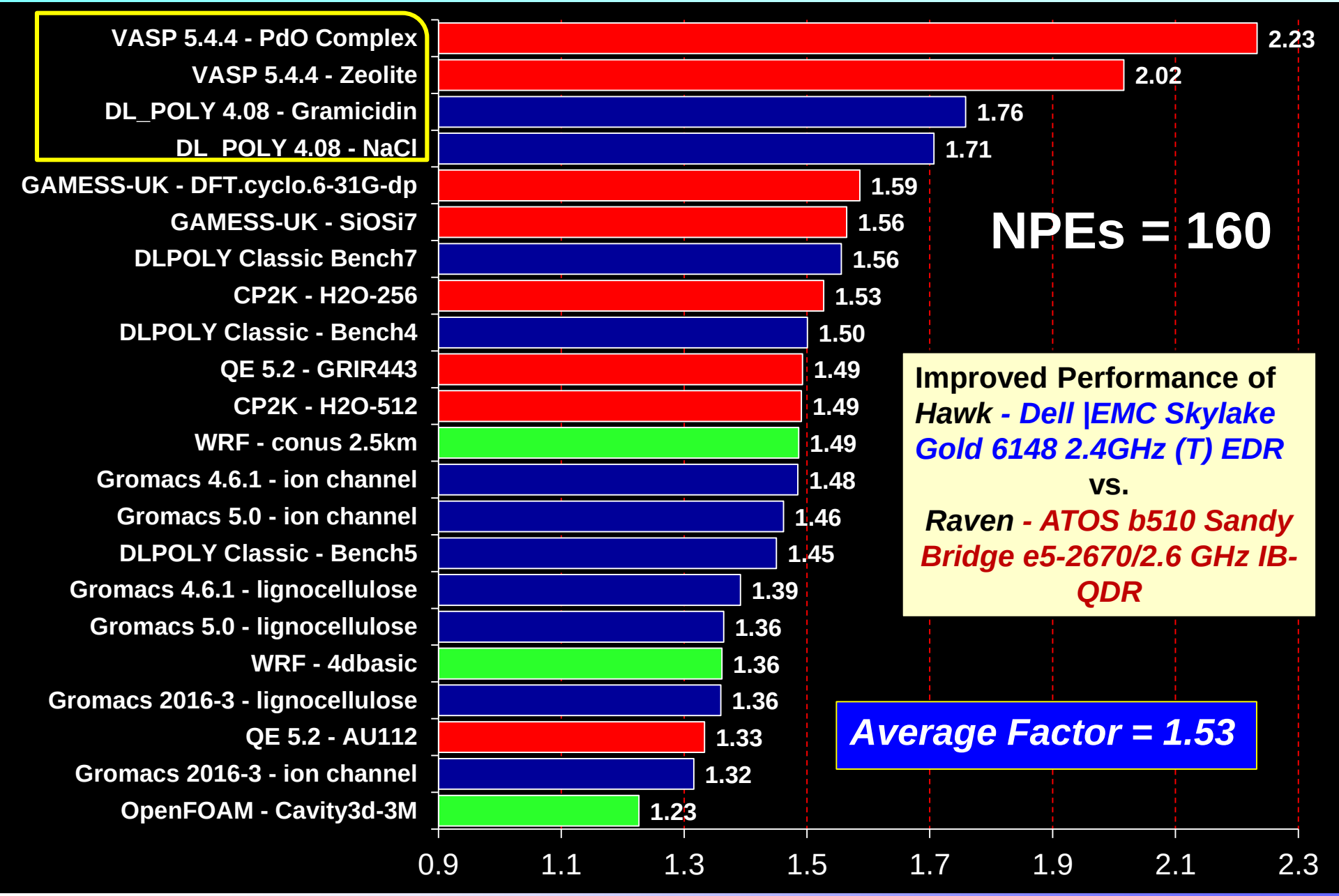




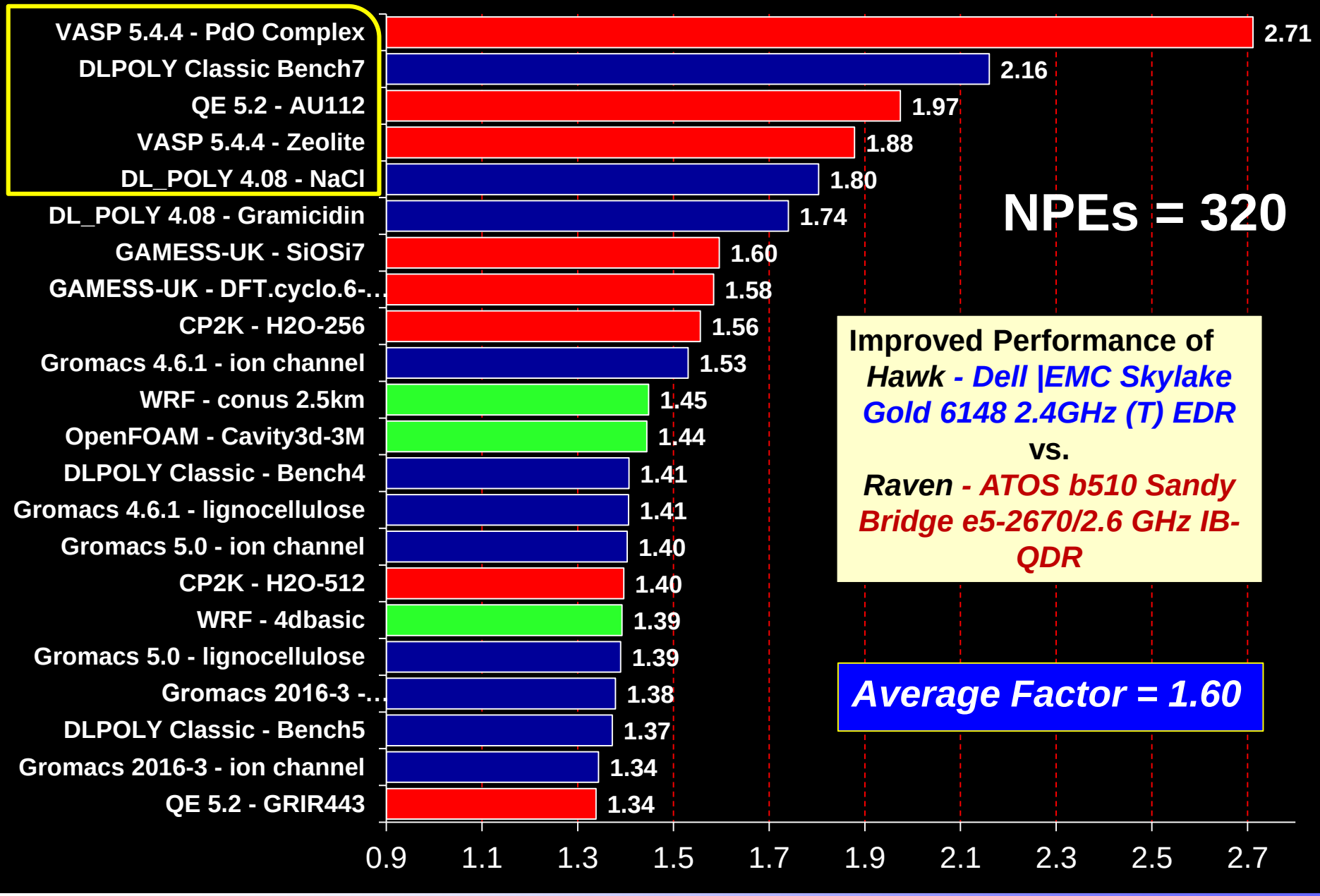
SKL "Gold" 6148 2.4 GHz EDR vs. SNB e5-2670 2.6 GHz QDR



SKL "Gold" 6148 2.4 GHz EDR vs. SNB e5-2670 2.6 GHz QDR



SKL "Gold" 6148 2.4 GHz EDR vs. SNB e5-2670 2.6 GHz QDR



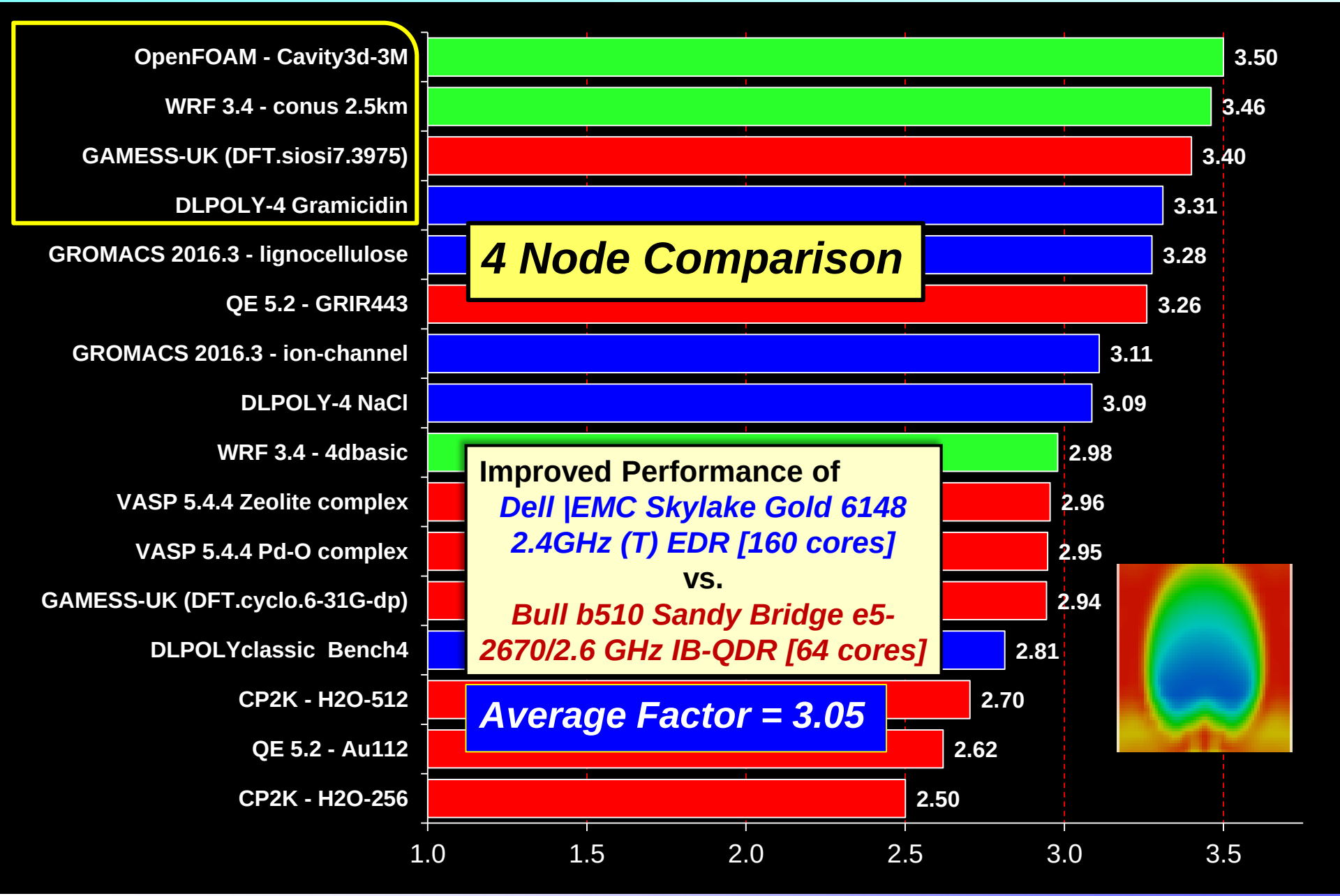
Performance Benchmarks – Node to Node

- Analysis of performance Metrics across a variety of data sets
  - ¤ "Core to core" and "node to node" workload comparisons
    - Previous charts based on *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 4 and 6 "node to node" comparison of the following:

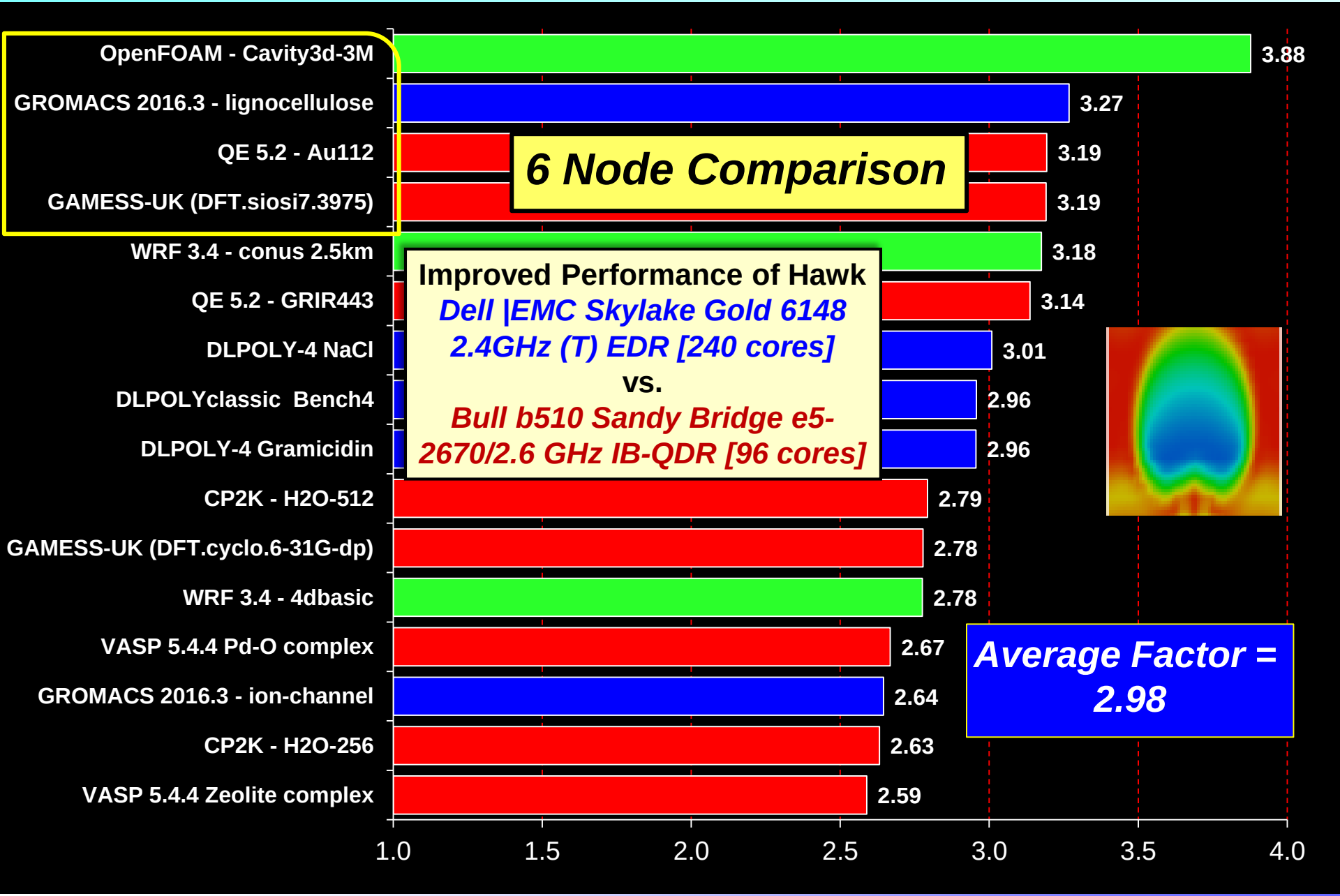
1	***Raven - Bull b510 Sandy Bridge e5-2670/2.6 GHz IB-QDR [64 cores]***	***Hawk - Dell |EMC Skylake Gold 6148 2.4GHz (T) EDR [160 cores]***
2	***Raven - Bull b510 Sandy Bridge e5-2670/2.6 GHz IB-QDR [96 cores]***	***Hawk - Dell |EMC Skylake Gold 6148 2.4GHz (T) EDR [240 cores]***

  - ¤ Benchmarks based on set of 10 applications & 19 data sets.

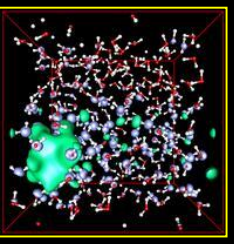
SKL "Gold" 6148 2.4 GHz EDR vs. SB e5-2670 2.6 GHz QDR



SKL "Gold" 6148 2.4 GHz EDR vs. SNB e5-2670 2.6 GHz QDR



EPYC - Target Codes and Data Sets – 128 PEs



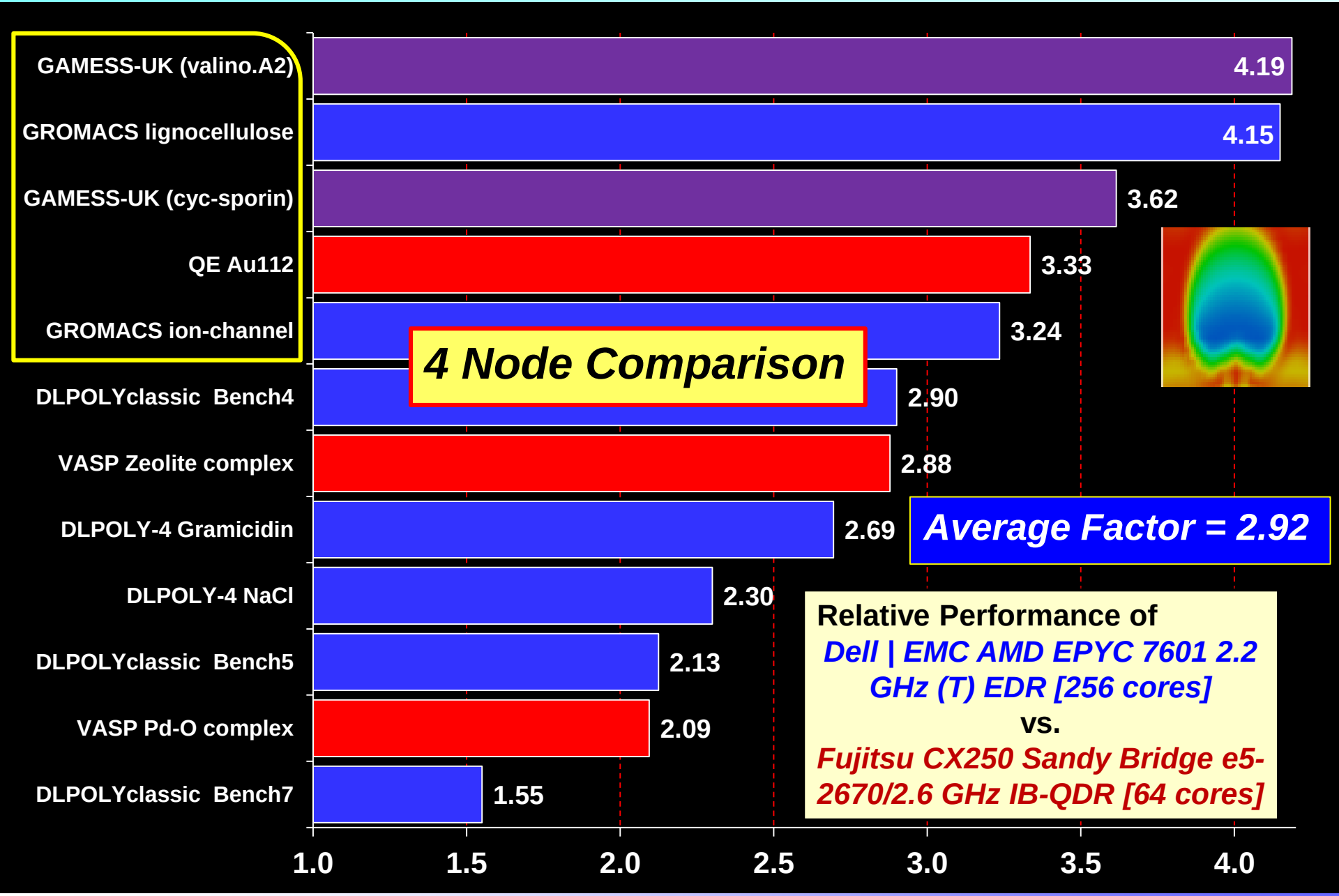
Performance Benchmarks – Node to Node

- Analysis of performance Metrics across a variety of data sets
  - ¤ "Core to core" and "node to node" workload comparisons
    - Previous EPYC charts based on *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 "node to node" comparison of the following:

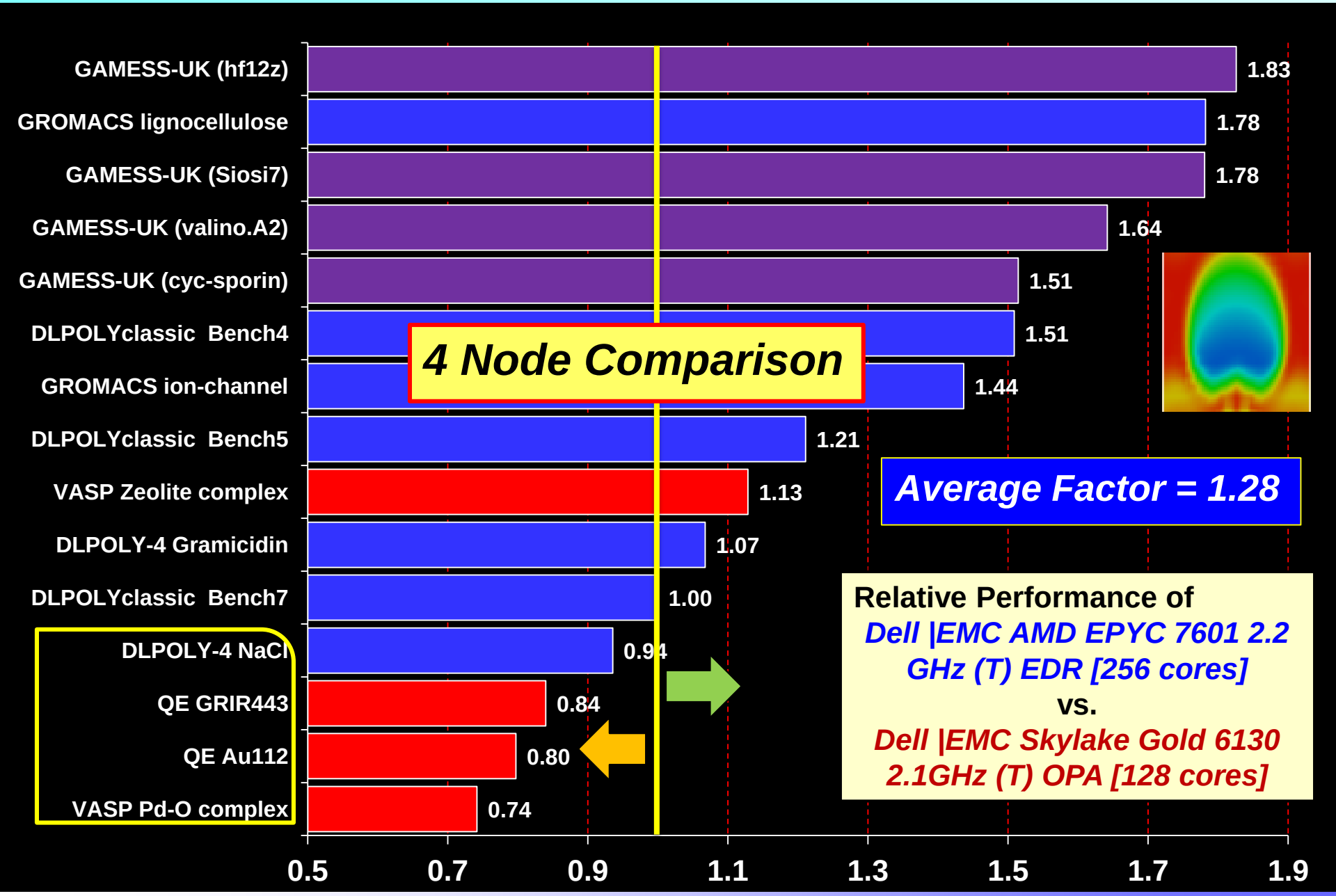
1	***Fujitsu CX250 Sandy Bridge e5-2670/2.6 GHz IB-QDR [64 cores]***	***Dell |EMC AMD EPYC 7601 2.2 GHz (T) EDR [256 cores]***
2	***Dell |EMC Skylake Gold 6130 2.1GHz (T) OPA [128 cores]***	***Dell |EMC AMD EPYC 7601 2.2 GHz (T) EDR [256 cores]***

  - ¤ Benchmarks based on set of 6 applications & 15 data sets.

Dell|EMC EPYC 7601 2.2 GHz (T) EDR vs. SB e5-2670 2.6 GHz QDR



SKL "Gold" 6130 2.1 GHz OPA vs. AMD EPYC 7601 2.2 GHz (T) EDR



Summary

- Ongoing Focus on performance benchmarks and clusters featuring **Intel's SKL** processors, with the addition of the **"Gold" 6138**, 2.0 GHz [20c] and **6148**, 2.4 GHz [20c] alongside the **6142**, 2.6 GHz [16c] ; and **6150**, 2.7 GHz [18c]).
- Performance comparison with current **SNB** systems and those based on **dual Intel BDW processor EP nodes** (16-core, 14-core) with Mellanox **EDR** and Intel's Omnipath **OPA interconnects**.
- Measurements of parallel application performance based on synthetic and end user applications – **DLPOLY, Gromacs, Amber, GAMESS-UK, Quantum ESPRESSO and VASP**.
  - Use of Allinea Performance reports to guide analysis, and updated comparison of Mellanox's HPC-X and Intel MPI on EDR-based systems
- Results augmented through consideration of two AMD Naples EPYC clusters, featuring the 7601 (2.20 GHz) and 7551 (2.00 GHz) processors.

Summary II

- Relative Code Performance: *Processor Family and Interconnect – "core to core" and "node to node" benchmarks.*
- A ***Core-to-Core*** comparison focusing on the Skylake "Gold" 6148 cluster (EDR) across 19 data sets (7 applications) suggests average speedups between **1.49** (80 cores) through **1.60** (320 cores) when comparing the to the Sandy Bridge-based "Raven" e5-2670 2.6GHz cluster with QDR environment.
  - Some applications however show **much higher factors** e.g. GROMACS and VASP depending on the level of optimisation undertaken on Hawk.
- A ***Node-to-Node comparison*** typical of the performance when running a workload shows increased factors.
  - A 4-node benchmark (*160 cores*) based on examples from 9 applications and 16 data sets show average improvement factors of **3.05** compared to the corresponding 4 node runs (*64 cores*) on the Raven cluster.
  - This factor is reduced somewhat, to **2.98**, when using 6 node benchmarks, comparing 240 SKL cores to 96 SNB cores.

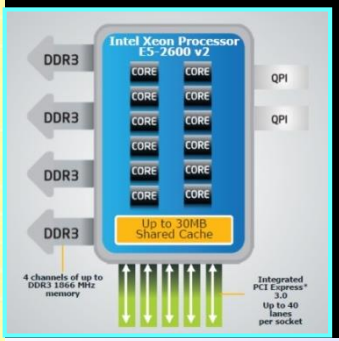
Summary III

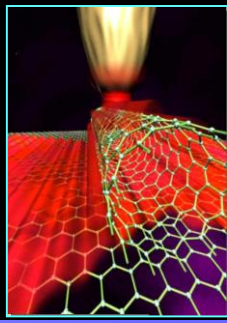
- An updated comparison of **Intel MPI and Mellanox's HPCX** conducted on the "Helios" cluster suggests that the clear delineation between MD (DLPOLY, GROMACS) and Materials-based codes (VASP, Quantum Espresso) is no longer evident.
- Ongoing studies on the EPYC 2701 shows a complex performance dependency on EPYC architecture.
  - Codes with high usage of vector instructions (Gromacs, VASP and Quantum Espresso) perform at best in somewhat modest fashion.
  - The AMD EPYC only supports 2 × 128-bit AVX natively, so there's a large gap with Intel and their 2 × 512-bit FMAs.
  - The floating point peak on AMD is 4 × lower than Intel and given that e.g., GROMACS has a native AVX-512 kernel for Skylake, performance inevitably suffers.

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Application Performance on Multi-core Processors

II. Acceptance Test Challenges and the Impact of Environment.





Background - Supercomputing Wales, New HPC Systems

- Multi-million £ **procurement exercise for new hubs** agreed by all partners
- **Tender issued in May 2017** following 6-9 month review of research community requirements and development of technical reference design
- **Budgetary challenges** due to currency devaluation and increase in component costs since budgets agreed in 2016
- **Contracts awarded to Atos, March 2018**. Hubs now installed and operational, based on **Intel Skylake Gold 6148**, supported by Nvidia GPU accelerators:
  - **Lot 1** – "**Hawk**" system - Cardiff hub. 7,000 HPC + 1,040 HTC cores
  - **Lot 2** – "**Sunbird**" system - Swansea hub. 5,000 HPC cores
  - **Lot 3** – "**Sparrow**" – Cardiff High Performance Data Analytics development system
- Suppliers to provide development opportunities and other activities through a programme of **Community Benefits**

Performance Acceptances Tests

1. Consideration of the **Performance Acceptance tests** undertaken as part of the Supercomputing Wales procurement. Carried out by Atos on the **"Hawk" HPC Skylake 6148 Cluster** at Cardiff University.
2. Performance targets built on benchmarks specified in the ITT – but developments impacted on the subsequent testing e.g., **SPECTRE / Meltdown**.
3. Assess Performance through analyses of results generated through three distinct run time environment variables, characterised by :
   - Turbo Mode – ON or OFF. Impact considerably more complicated with Skylake compared to previous Intel processor families.
   - Security patches – DISABLED or ENABLED on the Skylake 6148 compute nodes
   - Distribution of processing cores – PACKED or UNPACKED on each node e.g. 256 cores on either 7 or 8 × 40-core nodes.
4. Total of 8 Combinations – Impact on Performance ?
   - ITT defined that all – *"Application benchmarks should be in "PACKED" mode; HPCC in non-turbo mode*"

Process Adopted

1. Performance benchmark results generated by Atos (Martyn Foster) on the Hawk HPC Skylake 6148 Cluster at Cardiff University
2. MF adopted a systematic approach to assessing performance through the analyses of results generated **across four distinct environments** (a subset of the 8 possible environments)
   - “base (switch contained)” – Turbo mode off, security patches disabled on the Skylake 6148 compute nodes
   - “turbo + packed” - Turbo mode activated, with packed nodes – Slurm default, with 40 cores per Skylake 6148 node
   - “turbo + spread” - Turbo mode activated, de-populated nodes (32 cores / node)
   - “base + spectre” – base configuration above with security patches enabled
3. Identify those applications where the committed performance from the SCW ITT submission (“Target”) is not achieved.10% shortfall allowed

GLOBAL_SETTINGS

```
        export SPECTRE="clush -b -w $SLURM_NODELIST sudo /apps/slurm/disablekpti"
        export SPEC="disable"
###### OR ######
        export SPECTRE="clush -b -w $SLURM_NODELIST sudo /apps/slurm/enablekpti"
        export SPEC="enable"

        export TURBO="clush -b -w $SLURM_NODELIST sudo /apps/slurm/turbo_on" ;
        export TSTR=TURBO
###### OR ######
        export TURBO="clush -b -w $SLURM_NODELIST sudo /apps/slurm/turbo_off" ;
        export TSTR=OFF

        export SRUN_PACKING="-m Pack" ; export PSTR=Packed
###### OR ######
        export SRUN_PACKING="-m NoPack"; export PSTR=Spread

export LAUNCHER="srun ${SRUN_PACKING} --cpu_bind=verbose,cores --export
LD_LIBRARY_PATH"
```

SCW Application Performance Benchmarks

- The ***Benchmark suite*** comprises both **synthetics & end-user applications**. Synthetics include HPCC (*http://icl.cs.utk.edu/hpcc*) & IMB benchmarks (*http://software.intel.com/en-us/articles/intel-mpi-benchmarks*), IOR and STREAM
- Variety of "open source" & commercial end-user application codes:

GROMACS and DL_POLY-4 (molecular dynamics)

Quantum Espresso and **VASP** (ab initio Materials properties)

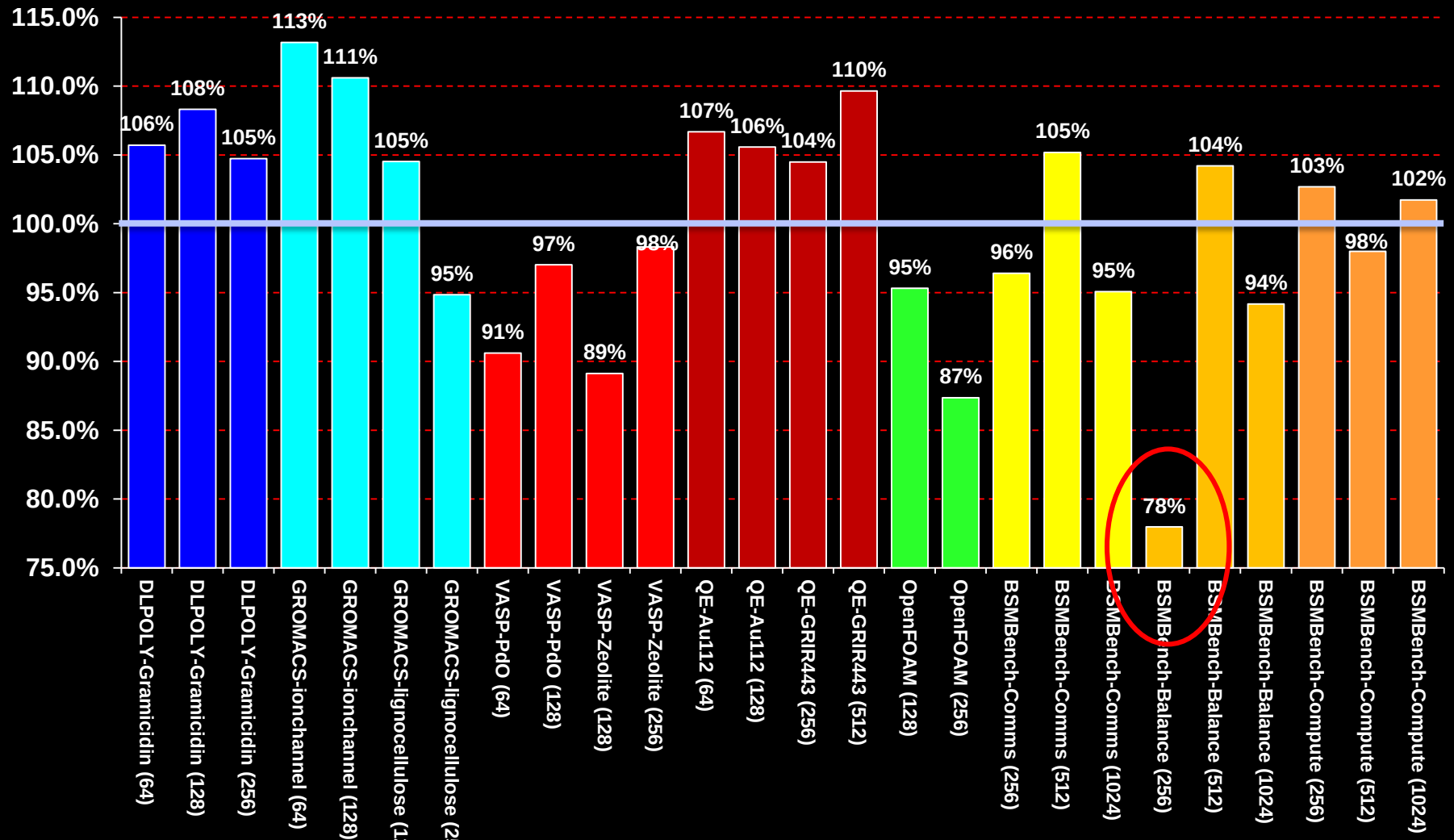
BSMBench (particle physics – Lattice Gauge Theory Benchmarks)

OpenFOAM (computational engineering)

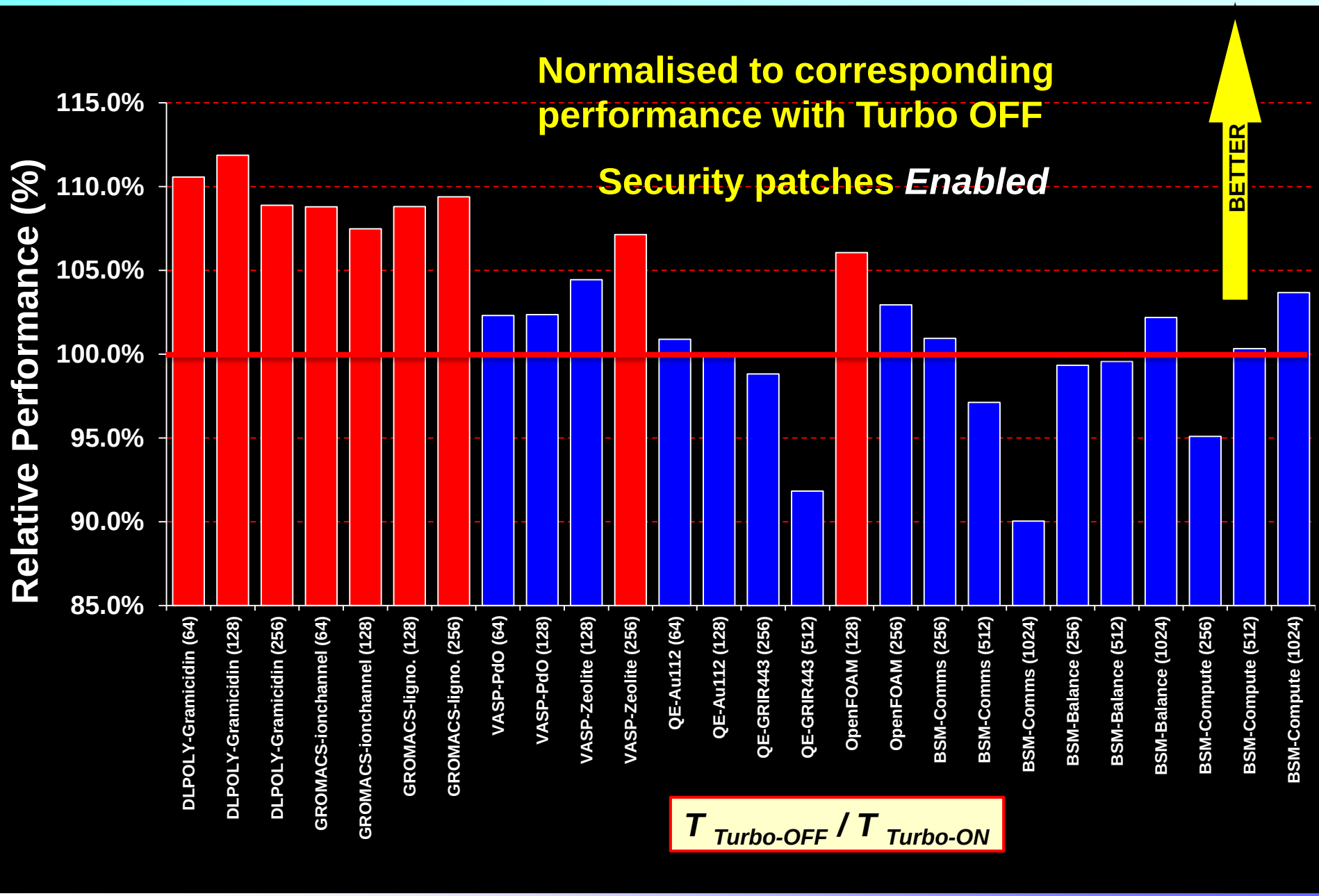
- These stress various aspects of the architectures under consideration and should provide a level of insight into why particular levels of performance are observed.

“Sunbird” Acceptance Tests – User Applications

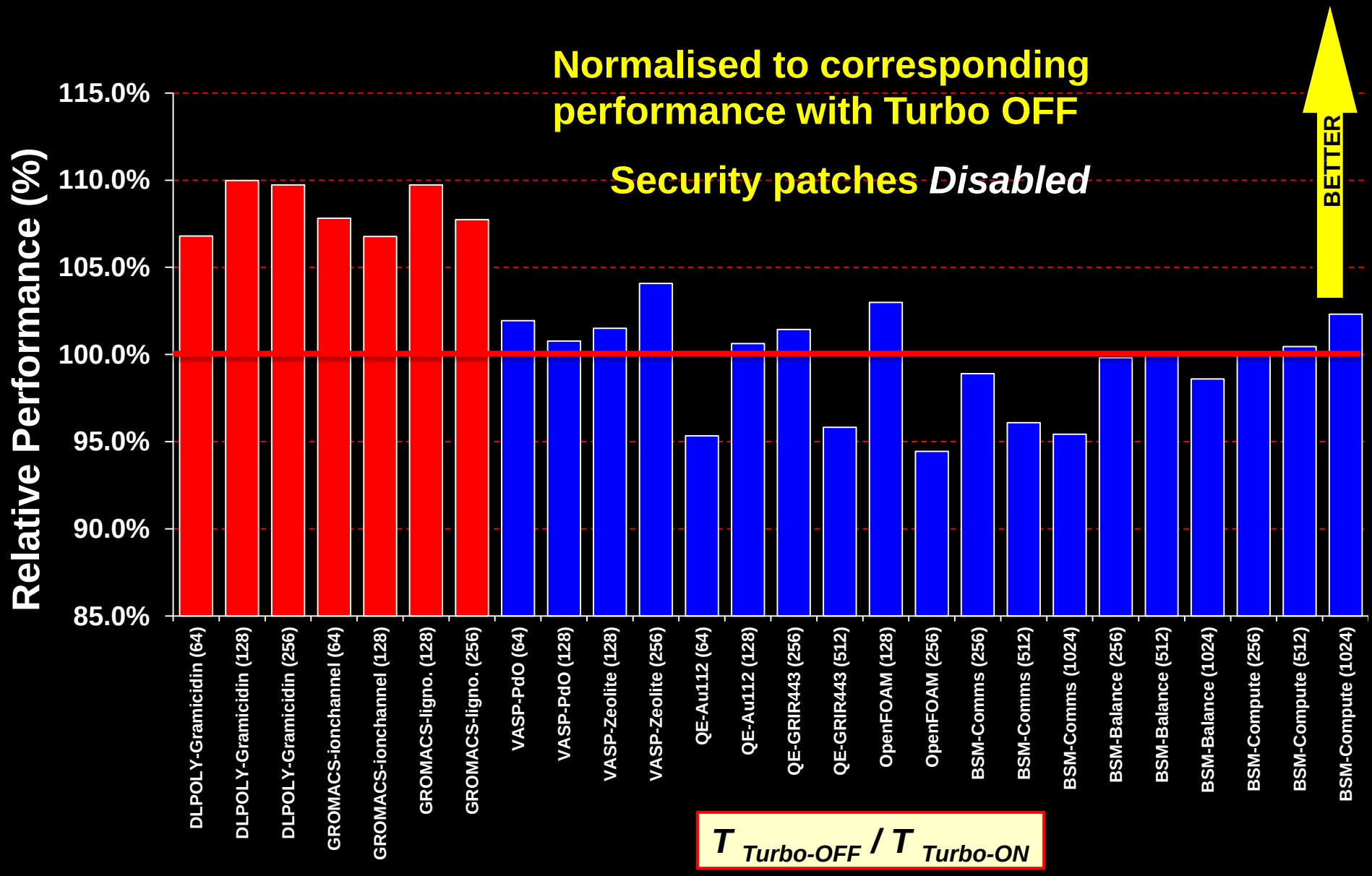
Basket of Synthetic (HPCC, IOR, STREAM, IMB) and end-user application codes – DL_POLY, GROMACS, VASP, ESPRESSO, OpenFOAM & BSMBENCH)



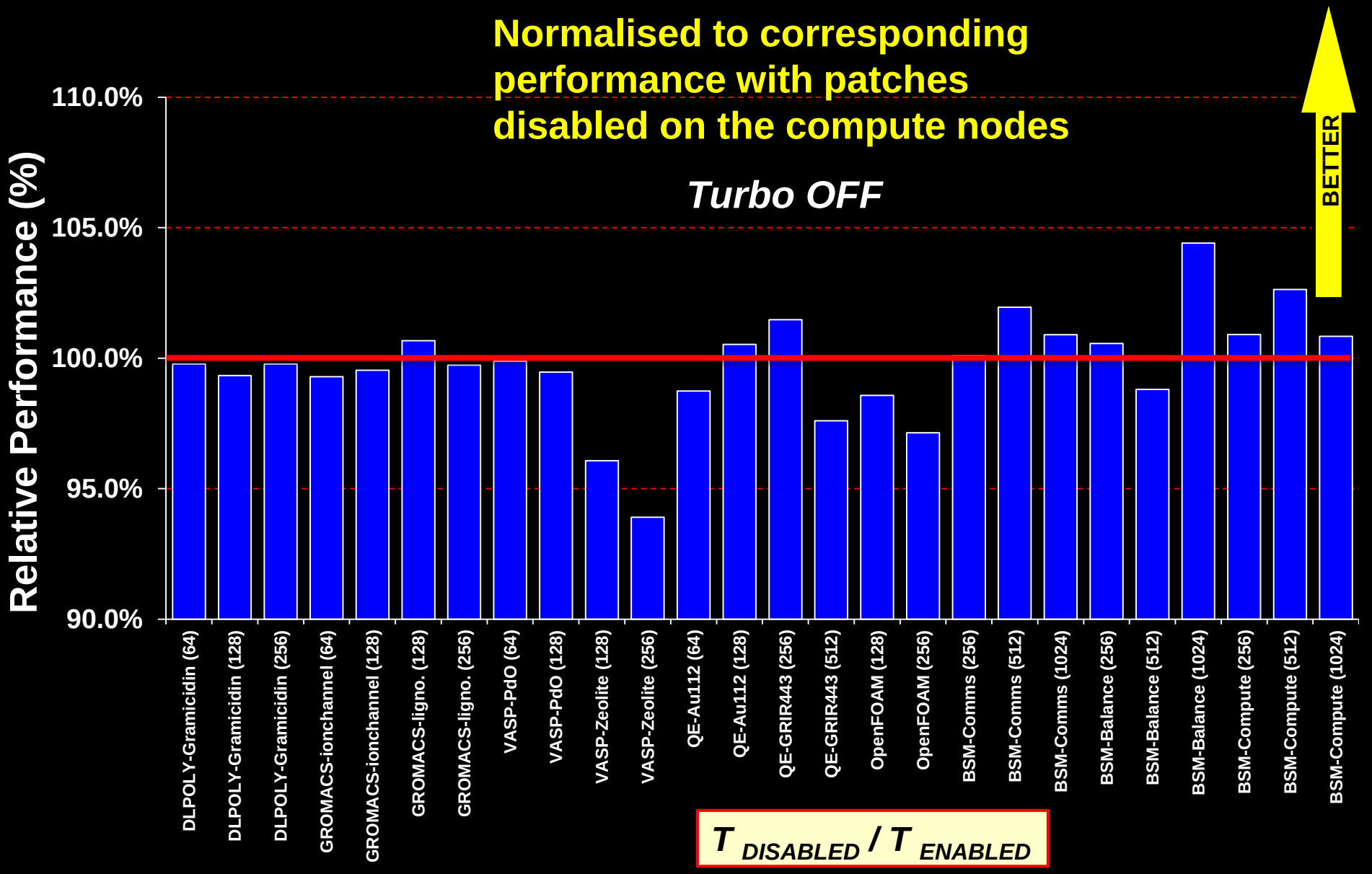
Impact of Turbo Mode on Performance (Security Patches Enabled)



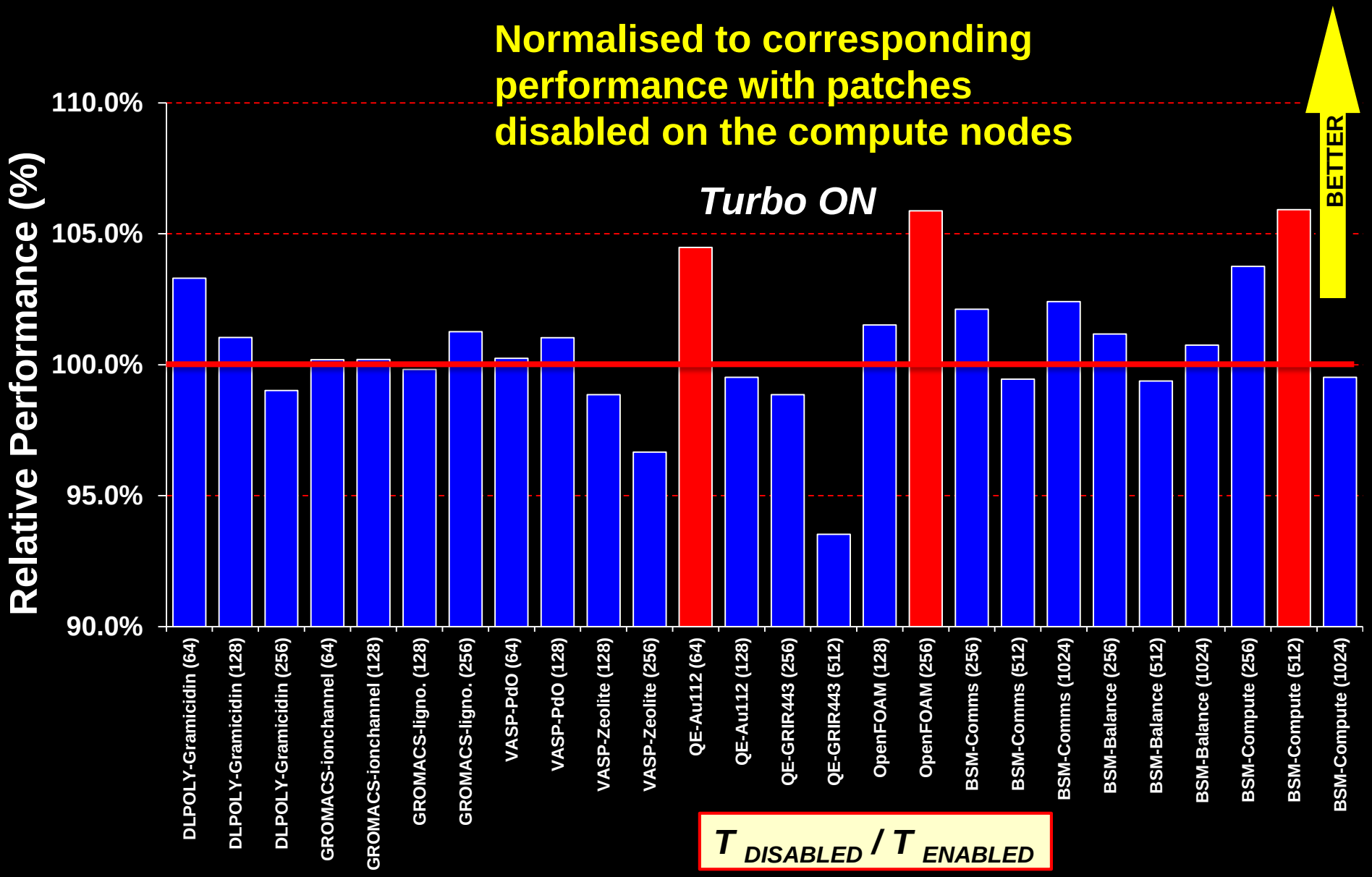
Impact of Turbo Mode on Performance (Security Patches Disabled)



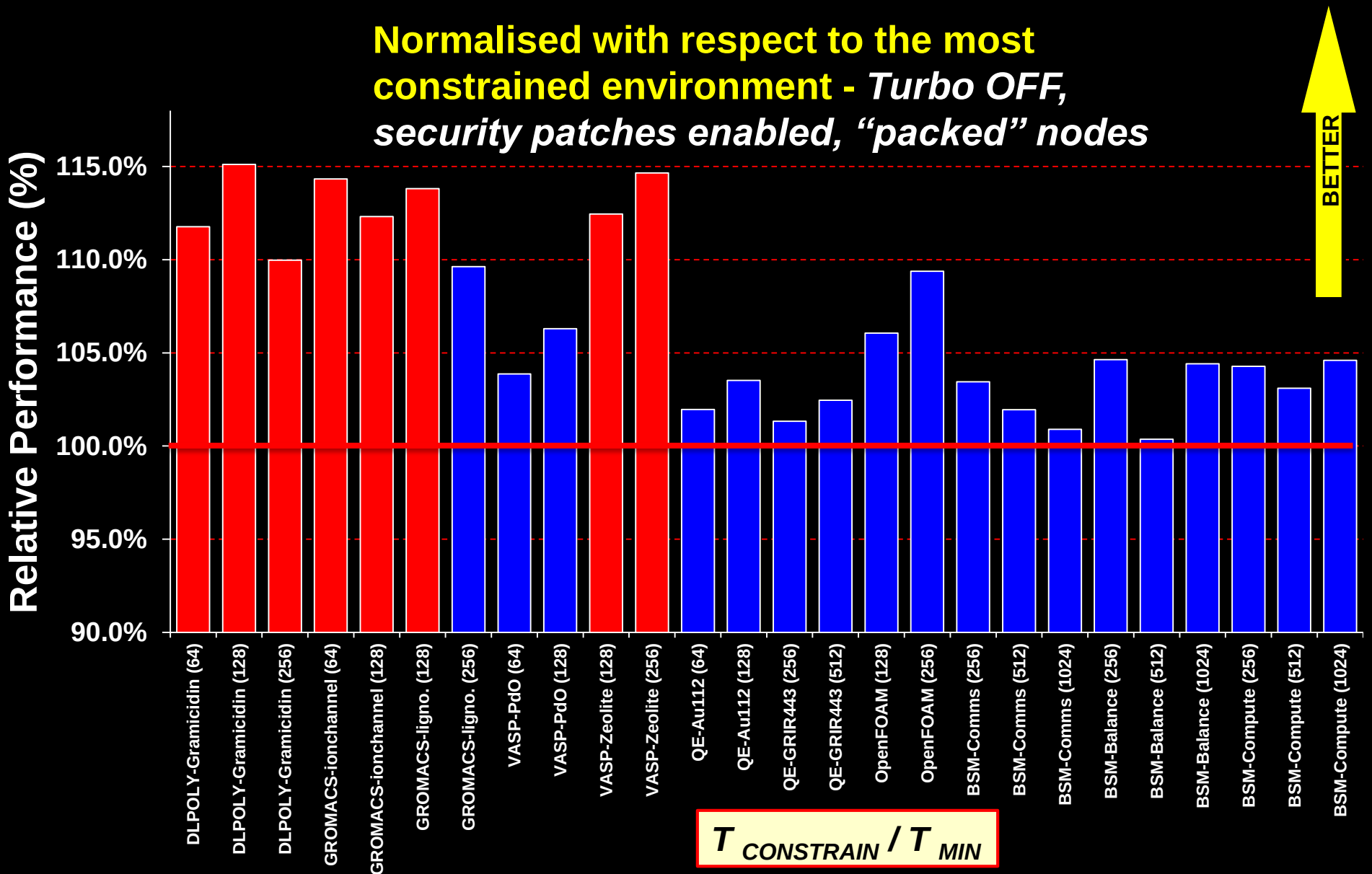
Impact of Security Patches on Performance (Turbo Mode OFF)



Impact of Security Patches on Performance (Turbo Mode ON)



Overall Impact of Environment on Performance



Workload validation and Throughput tests

- **Aim:** Throughput designed to **illustrate the Stability of the system** over an observed period of a week, while **hardening** the system
- Benchmarks based on **multiple, concurrent instantiations** of a number of data sets associated with five of the end user application codes and two of the synthetic benchmarks.
- Each data set is run a number of times on a variety of processor (core) counts - typically 40, 80, 160, 320, 640 and 1024. This combination of jobs has been **designed to run for approximately 6 hours (elapsed time)** on a 2720-core, 68 node cluster partition.
- Note that the metrics for success of these tests are twofold:
  1. All jobs comprising a given run **complete successfully** and
  2. There is a **consistency of run time** across each of the tests. The measured time is simply the time at which the first of the jobs is launched through the time that the last jobs finishes.

Workload validation and Throughput tests

- Based around multiple instantiations of a number of data sets associated with the five codes, DLPOLY4, Gromacs (v5.2), Quantum Espresso, OpenFOAM and VASP, and the two synthetic benchmarks, IMB and IOR.
  - DLPOLY4 - NaCl & Gramicidin
  - Gromacs - ion_channel & lignocellulose
  - QE 6.1 - AUSURF112 & GRIR443
  - OpenFOAM - cavity3d-3M
  - VASP 5.4.4 – PdO complex and Zeolite

SLURM Scripts

DLPOLY4.test2+test8.SCW.40.q
DLPOLY4.test2+test8.SCW.80.q
DLPOLY4.test2+test8.SCW.160.q
DLPOLY4.test2+test8.SCW.320.q
DLPOLY4.test2+test8.SCW.640.q
GROMACS.All.SCW.80.q
GROMACS.All.SCW.160.q
GROMACS.All.SCW.320.q
GROMACS.All.SCW.640.q
GROMACS.All.SCW.1024.q
IMB3.SCW.160.q
IMB3.SCW.320.q
IOR.SCW.4.q
IOR.SCW.8.q
OpenFOAM_cavity3d-3M.SCW.80.q
OpenFOAM_cavity3d-3M.SCW.160.q
OpenFOAM_cavity3d-3M.SCW.320.q
OpenFOAM_cavity3d-3M.SCW.640.q
QE.AUSURF112.SCW.160.q
QE.AUSURF112.SCW.320.q
QE.GRIR443.SCW.320.q
QE.GRIR443.SCW.640.q
VASP.example3.SCW.80.q
VASP.example3.SCW.160.q
VASP.example3.SCW.320.q
VASP.example4.SCW.160.q
VASP.example4.SCW.320.q

Throughput Tests – Hawk System – Two partition Approach

The throughput tests were undertaken on two separate partitions of the Hawk cluster – *compute64* and *compute64b* – to enable other testing and early pilot user service. Each partition comprised 68 nodes.

Partition 1 – Compute 64 (68 Nodes)

- The first set of trial runs was executed between **12-14 May**. A number of the runs failed to complete, subsequently attributed to an apparent VASP related error peculiar to the lustre file system:

```
forrtl: severe (121): Cannot access current working directory for unit 18, file "Unknown"
Image              PC                Routine            Line        Source
vasp_std           00000000014F3E09  Unknown               Unknown  Unknown
vasp_std           000000000150E10F  Unknown               Unknown  Unknown
vasp_std           000000000134C950  Unknown               Unknown  Unknown
vasp_std           000000000040AF5E  Unknown               Unknown  Unknown
libc-2.17.so       00002B450F32EC05  __libc_start_main     Unknown  Unknown
vasp_std           000000000040AE69  Unknown               Unknown  Unknown
forrtl: error (76): Abort trap signal
```

- This transient error affected perhaps one in twenty identical jobs, and although reported into the appropriate Level 3 service regimes, has still not been formally addressed. A workaround module was developed by Cardiff's Tom Green when it became clear that the formal channels were struggling.

```
module load lustre_getcwd_fix
```

Throughput Tests – Hawk System II

Partition 1:

- A second set of trial runs were carried out over the bank holiday weekend and successfully passed the associated tests over the period **30 May – 3 June**.

Run #	Start Time		Finish Time		Total Elapsed Time (hours:Mins)
6	30May	21-21	31May	03-25	6:02
7	31May	23-33	01Jun	05-38	6:05
8	02Jun	00-04	02Jun	06-06	6:02
9	02Jun	13:24	02Jun	19:27	6:04
10	02Jun	22-57	03Jun	05-00	6:03
11	03Jun	05-45	03Jun	11-47	6:02
12	03Jun	15-57	03Jun	22-12	6:15

- **Partition 2:**
- **Runs 11 -22:** Initial runs using *compute64b* conducted between **7 – 10 June** revealed a number of issues pointing to readiness of the nodes. Timings from the first completed run suggested some variability in run times for a given application/core count, with the total run time significantly longer than those on compute64.

Throughput Tests – Hawk System III

- Following a lustre upgrade, a further set of runs were undertaken between **21 June and 25 June**. Runs 8 - 12 actually ran OK, so formally *compute 64b*, along with compute64, can be judged to have passed the Acceptance Test throughput requirement of five consecutive error-free runs, although the variations in the individual run times are perhaps larger than hoped.

Run #	Start Time		Finish Time		Total Elapsed Time (hours:Mins)
8	23Jun	15-44	23Jun	20-57	5:13
9	23Jun	21-35	24Jun	02-55	5:20
10	24Jun	11-34	24Jun	17-06	5:32
11	24Jun	18-56	25Jun	00-16	5:20
12	25Jun	00-31	25Jun	06-03	5:32

- Testing on Hawk commenced on **12 May 2018** and was finally completed on the **25 June 2018**.

Throughput Tests – Sunbird System – Two partition approach

Partition 1: Runs 1 – 4:

- **Run 3** did not complete with JOBID #11050 hanging, while JOBID #11372 of **Run 4** suffered the same fate. Both jobs failed with the all too familiar **VASP/lustre error diagnostics**. The scripts used were identical to those used on Hawk in June, and did not include the workaround introduced at the time.

- **Runs 5 -10:** Completed successfully, with two of the VASP/Lustre partitions trapped though the added module

```
module load lustre_getcwd_fix
```

Partition 2: Runs 11 - 22: Three jobs in one of the runs hung when hitting problems on scs0105. That node had been taken out for when setting up the user-facing file systems and needed the playbooks running. Several of the runs showed the impact of the lustre issue with VASP.

- However, there were **significant variations in the overall run times**.
  - At least three of the nodes appeared to be either defective or possess some **different bios settings** (scs0064,scs0092 and scs0096). These were subsequently removed from service.
  - **Turbo in inconsistent state** across the compute nodes. Usually reset by the Slurm prologue scripts, but they appear to have been commented out.

Throughput Tests – Sunbird System

- Runs 23 – 30: Runs certainly acceptable from the metric of job completion, for all completed successfully. Note there was no reoccurrence of the lustre-related issue during this set of runs.

Run #	Start Time		Finish Time		Total Elapsed Time (hours:Mins)
23	17Aug	17:52	17Aug	23:07	5:15
24	17Aug	23:54	18Aug	05:11	5:17
25	18Aug	05:19	18Aug	10:36	5:17
26	18Aug	14:00	18Aug	19:16	5:16
27	18Aug	19:51	19Aug	01:08	5:17
28	19Aug	02:45	19Aug	07-57	5:12
29	19Aug	13:21	19Aug	18:32	5:11
30	19Aug	19:06	20Aug	00:26	5:20 (SLURM CG Issue

- Testing on Sunbird commenced on **10 August 2018**, and was finally completed on the evening of **19 August 2018**

Throughput Tests – Nottingham OCF Cluster

- Tests modified to run on two partition of the OCF cluster at Nottingham, “*martyn*" and "*colin*", each comprising **50 nodes** with EDR interconnect. All component nodes comprised dual **Gold 6138** 2.0GHz 20c SKL processors
- Initial runs of the workload failed to complete successfully, with each of the **8 x 320-core IMB jobs** hanging, consuming all of their allocated time. Traced to an issue with the ***gatherv*** collective that failed to complete across all specified msglens.
- Navigated around the issue by removing those environment variables deemed likely to trigger the problem, specifically:
  - export I_MPI_JOB_FAST_STARTUP=enable
  - export I_MPI_SCALABLE_OPTIMIZATION=enable
  - export I_MPI_DAPL_UD=enable
  - export I_MPI_TIMER_KIND=rdtsc
- With these removed, runs proceeded to complete successfully.
- One of the allocated nodes (*compute099*) rendered unusable as a result of tests - removed from service. Thus the subsequent runs used 49 nodes, rather that the intended 50.

Throughput Tests – Acceptance Achieved (OCF System)

Table. Overall run times for the throughput runs on the "*martyn*" partition.

Run #	Start Time		Finish Time		Total Elapsed Time (hours:Mins)
2	31Jul	18-04	01Aug	00-49	6:45
3	01Aug	01-22	01Aug	08-07	6:45
4	01Aug	08-32	01Aug	15-18	6:46
5	01Aug	17-08	01Aug	23-52	6:44
6	02Aug	03-20	02Aug	10-06	6:46

Table. Overall run times for the throughput runs on the "colin" partition.

Run #	Start Time		Finish Time		Total Elapsed Time (hours:Mins)
1	02Aug	12-24	02Aug	19-05	6:41
2	03Aug	02-41	03Aug	09-18	6:37
3	03Aug	10-34	03Aug	17-18	6:44
4	03Aug	17-35	04Aug	00-26	6:51
5	04Aug	01-00	04Aug	07-45	6:45

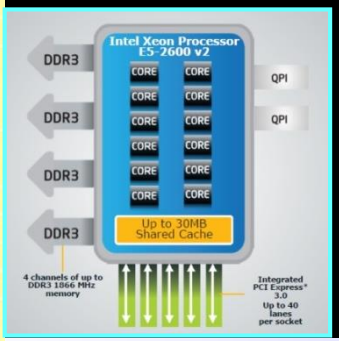
Results of "throughput benchmarks" carried out on the new OCF Skylake cluster at Nottingham University between 31 July and 4 August 2018.

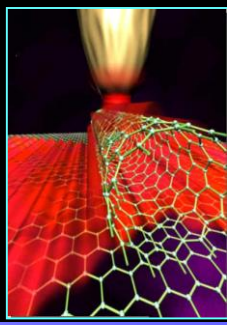
CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Application Performance on Multi-core Processors

III. The Performance Evolution of two Community Codes, DL_POLY and GAMESS-UK

.





Outline and Contents

1. Introduction – DL_POLY and GAMESS-UK
   - ¤ Background and Flagship community codes for the UK's CCP5 & CCP1 – Collaboration!
2. HPC Technology – Impact of Processor & Interconnect developments
   - ¤ The last 10 years of Intel dominance – Nehalem to Skylake
3. DL_POLY and GAMESS-UK Performance
   - ¤ Benchmarks & Test Cases
   - ¤ Overview of two decades of Code Performance: From the Cray T3E/900 to Intel Skylake clusters

"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, Accepted for publication (2019).

The Story of Two Community Codes

DL_POLY and GAMESS-UK - A Performance Overview

HPC Technology – Processor and Networks

Computer Systems

- Benchmark timings - a wide variety of systems, starting with the **Cray T3E/1200 in 1999**. Access initially undertaken as part of Daresbury's Distributed Computing support programme (**DiSCO**), with the benchmarks presented at the annual **Machine Evaluation Workshops** (1989-2014) and STFC's successor **Computing Insight (CIUK)** conferences (2015 onwards).
  - Access typically **short-lived** as systems provided by suppliers to enhance their profile at the MEW Workshops - limited opportunity for in depth benchmarking.
- Systems include a **wide range of CPU offerings**. Representatives from over a dozen generations of Intel processors, from the early days of single processor nodes housing **Pentium 3 and Pentium 4** CPUs, through dual processor nodes featuring dual-core **Woodcrest**, quad-core **Clovertown** & **Harpertown** processors, along with the **Itanium and Itanium2** CPUs, through to the extensive range of multi-core offerings **Westmere - Skylake**.

Computer Systems

- A variety of processors from **AMD** (Athlon, Opteron, MagnyCours, Interlagos etc.) along with the **"power" processors from the IBM pSeries** have also featured (typically dual processor configurations).
- In the same way a wide variety of processors feature, so too is the appearance of a range of **network interconnects**. **Fast Ethernet and GBit Ethernet** were rapidly superseded by the increasing capabilities of the family of **Infiniband** interconnects from **Voltaire and Mellanox** (SDR, DDR, QDR, FDR, EDR and soon HDR), along with the now defunct offerings from **Myrinet, Quadrics and QLogic**. The **Truescale** interconnect from Intel, along with its successor, **Omnipath**, also feature.
- Dating from the appearance of Intel's SNB processors, many of the timings generated with the **Turbo mode feature** enabled by the system administrators. Such systems are tagged with "(T)" notation.
- As for **software**, most of the commodity clusters featuring Intel CPUs used successive generation of **Intel compilers** along with **Intel MPI**, although a range of MPI libraries have been used – OpenMPI, MPICH, MVAPICH and MVAPICH2. Proprietary systems (Cray and IBM) used system specific compilers and associated MPI libraries.

Intel Xeon : Westmere - Skylake

	Xeon 5600 (Westmere-EP)	Xeon E5-2600 (Sandy Bridge-EP)	Xeon E5-2600 v4 "Broadwell-EP"	Intel Xeon Scalable Processor "Skylake"
Cores / Threads	**Up to 6 cores / 12 threads**	**Up to 8 cores / 16 threads**	**Up to 22 Cores / 44 threads**	**Up to 28 Cores / 56 threads**
Last-level cache	12 MB	Up to 20 MB	Up to 55 MB	**Up to 38.5 MB (non-inclusive)**
Max memory channels, speed / socket	**3xDDR3 channels, 1333**	**4xDDR3 channels, 1600**	**4 channels of up to 3 RDIMMs, LRDIMMs or 3DS LRDIMMs, 2400 MHz**	**6 channels of up to 2 RDIMMs, LRDIMMs or 3DS LRDIMMs, 2666 MHz**
New instructions	AES-NI	**AVX 1.0 8 DP Flops/Clock**	**AVX 2.0 16 DP Flops/Clock**	**AVX 512 32 DP Flops/Clock**
QPI / UPI Speed (GT/s)	1 QPI channels @ 6.4 GT/s	2 QPI channels @ 8.0 GT/s	2 x QPI channels @ 9.6 GT/s	**Up to 3 x UPI @ 10.4 GT/s**
PCIe Lanes / Controllers / Speed (GT/s)	36 lanes PCIe 2.0 on chipset	40 Lanes / Socket Integrated PCIe 3.0	40 / 10 / PCIe* 3.0 (2.5, 5, 8 GT/s)	**48 / 12 / PCIe* 3.0 (2.5, 5, 8 GT/s)**
Server / Workstation TDP	Server / Workstation: 130W	Up to 130W Server; 150W Workstation	55 - 145W	**70 – 205W**

The Story of Two Community Codes

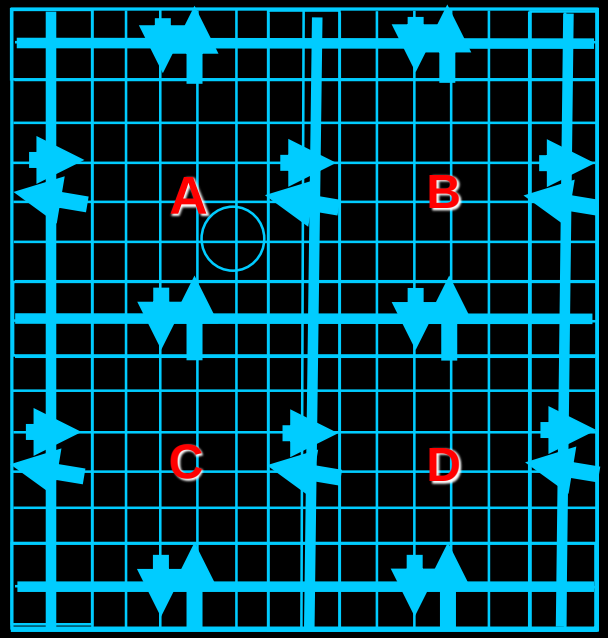
DL_POLY and GAMESS-UK - A Performance Overview

Overview of two decades of DL_POLY Performance

DL_POLY 3/4 – 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

The DLPOLY Benchmarks

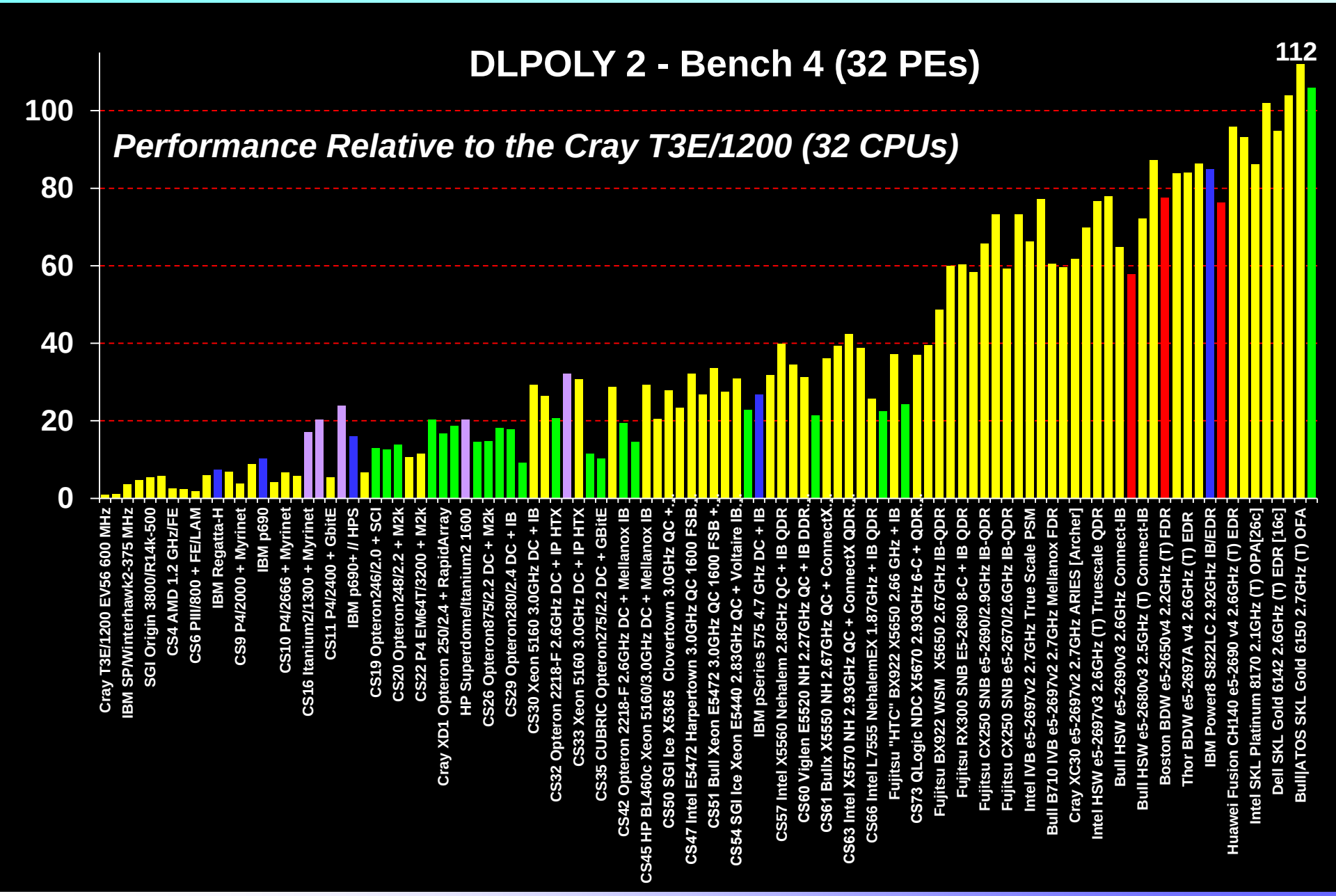
DL_POLY Classic

- Bench4
  - NaCl Melt Simulation with Ewald sum electrostatics & a MTS algorithm. 27,000 atoms; 500 time steps.
- Bench5
  - Potassium disilicate glass (with 3-body forces). 8,640 atoms: 3,000 time steps
- Bench7
  - *Simulation of gramicidin A molecule in 4012 water molecules using neutral group electrostatics. 12,390 atoms: 5,000 time steps*

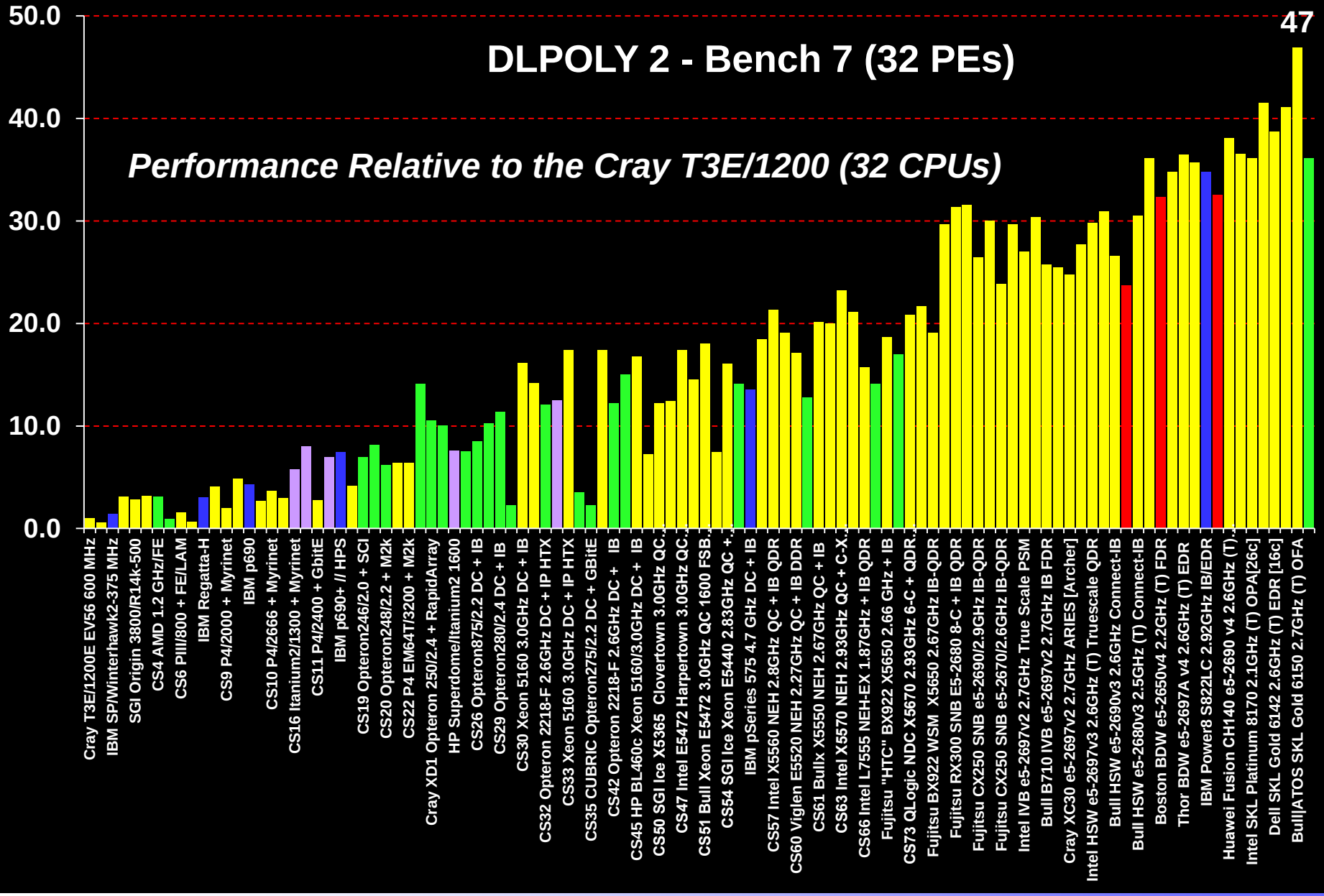
DL_POLY 4

- Test2 Benchmark
  - NaCl Simulation; 216,000 ions, 200 time steps, Cutoff=12Å
- Test8 Benchmark
  - *Gramicidin in water; rigid bonds + SHAKE: 792,960 ions, 50 time steps*

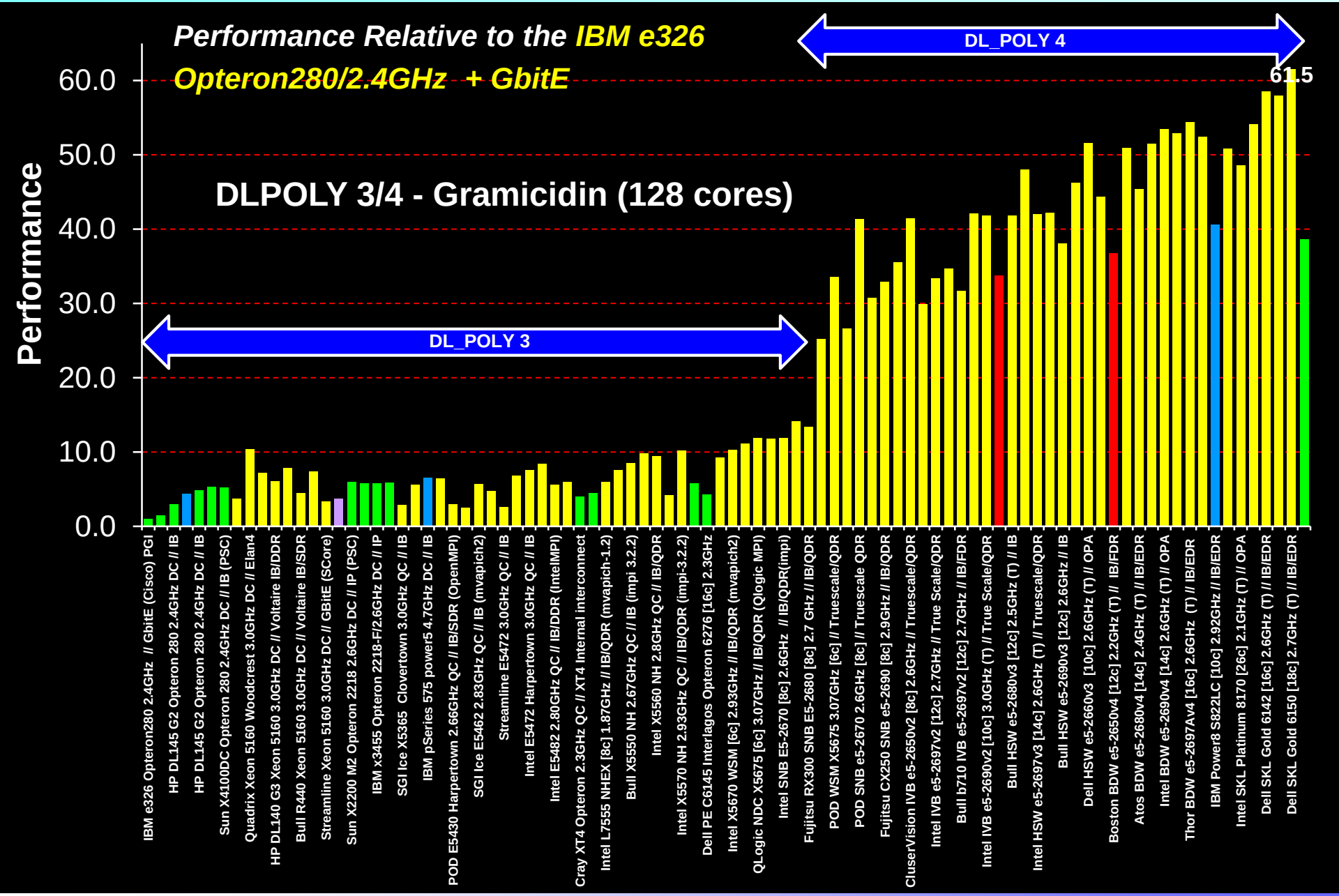
DL_POLY Classic: Bench 4



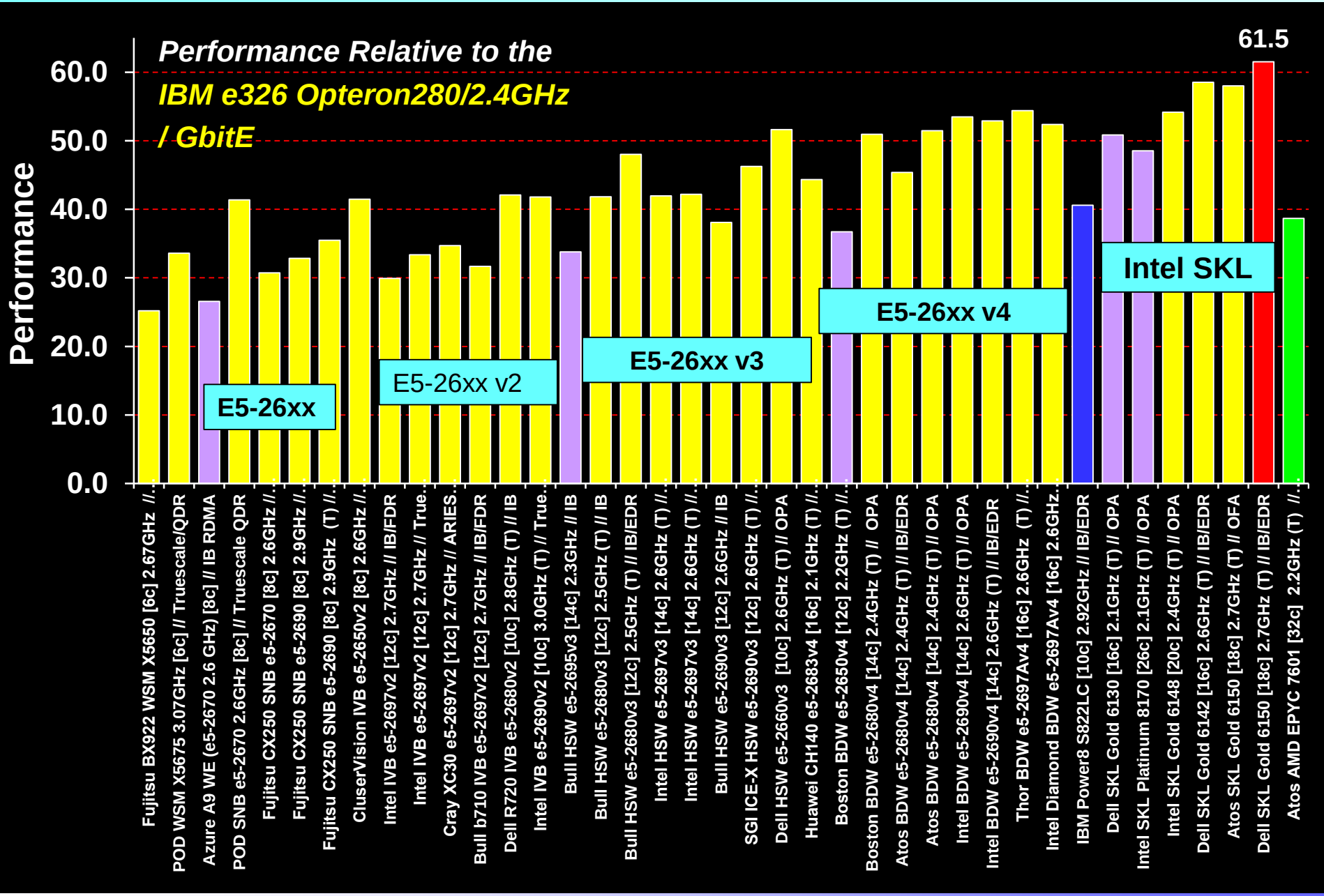
DL_POLY V2: Bench 7



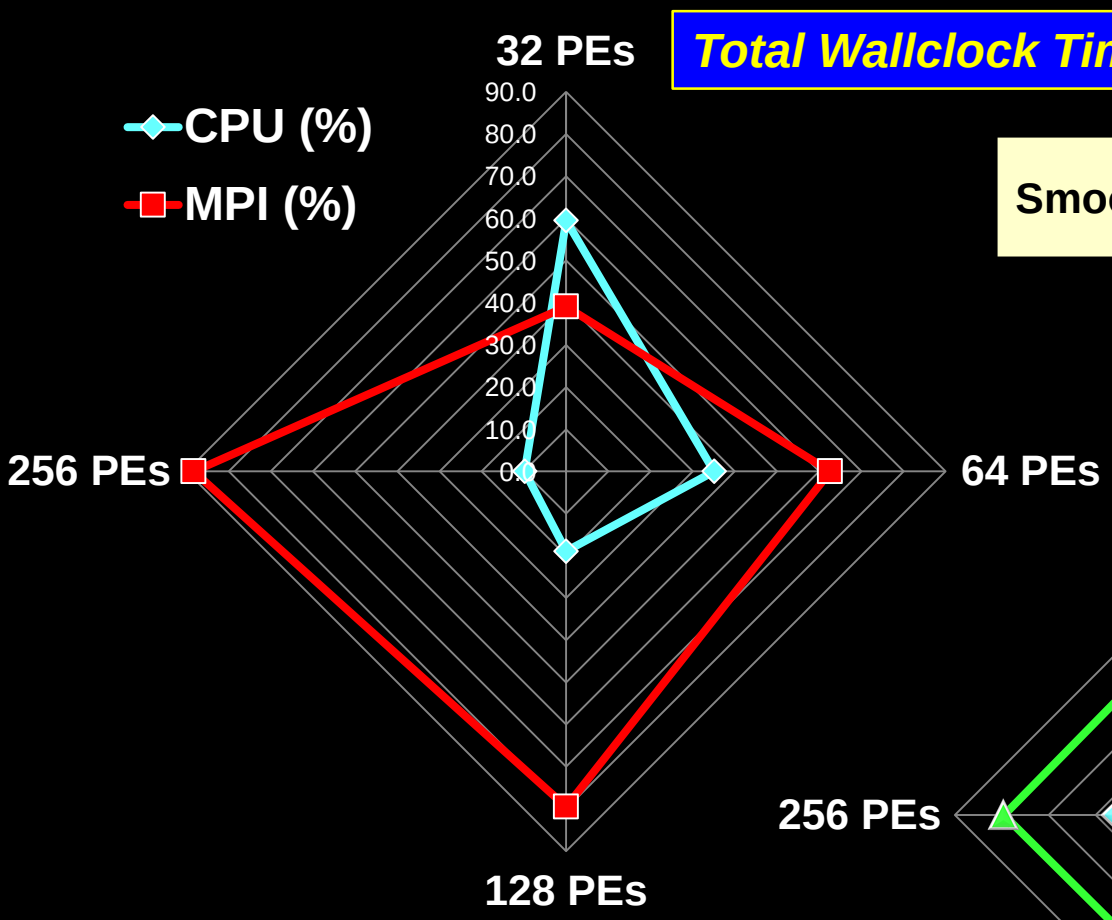
DL_POLY 3/4 – Gramicidin (128 Cores)



DL_POLY 4 – Gramicidin (128 cores)

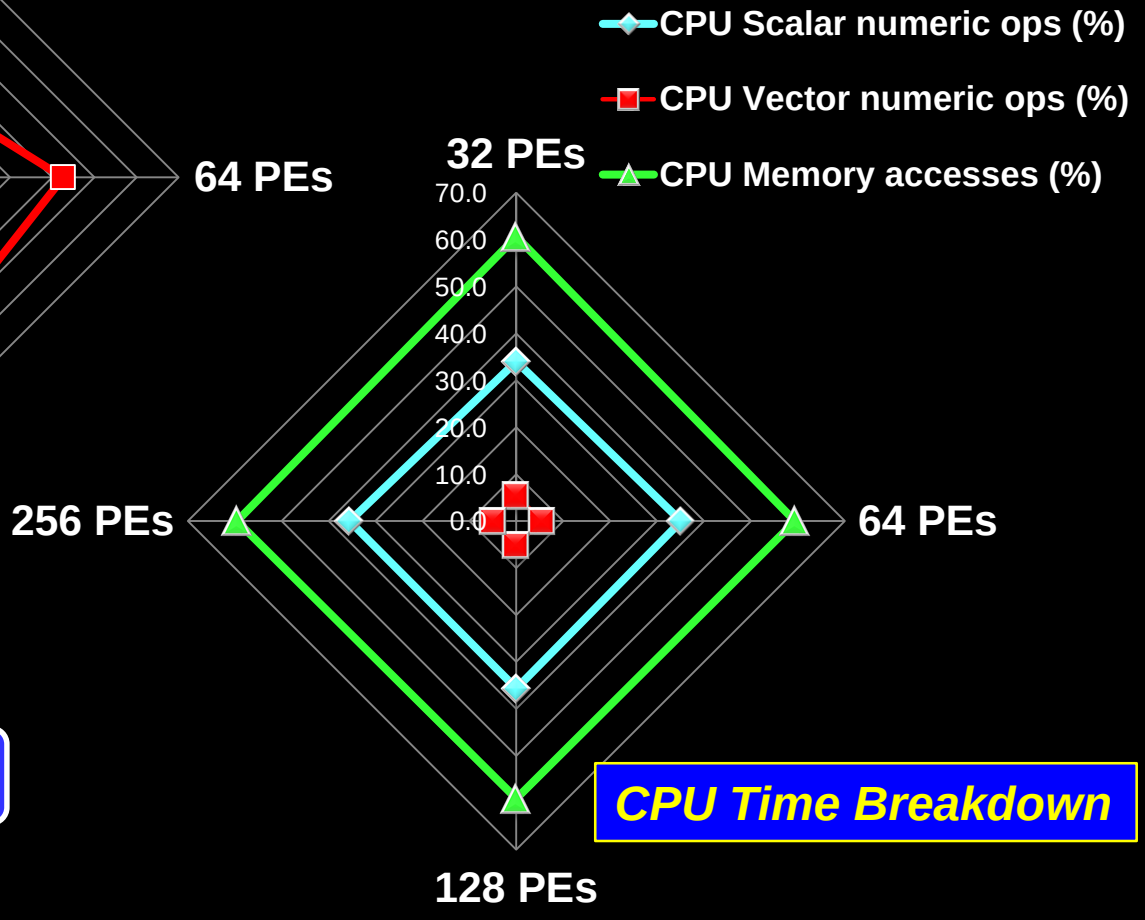


DL_POLY4 – Gramicidin Perf Report



Total Wallclock Time Breakdown

Smooth Particle Mesh Ewald Scheme



Performance Data (32-256 PEs)

CPU Time Breakdown

The Story of Two Community Codes

DL_POLY and GAMESS-UK - A Performance Overview

Overview of two decades of GAMESS-UK Performance

Large-Scale Parallel *Ab-Initio* Calculations

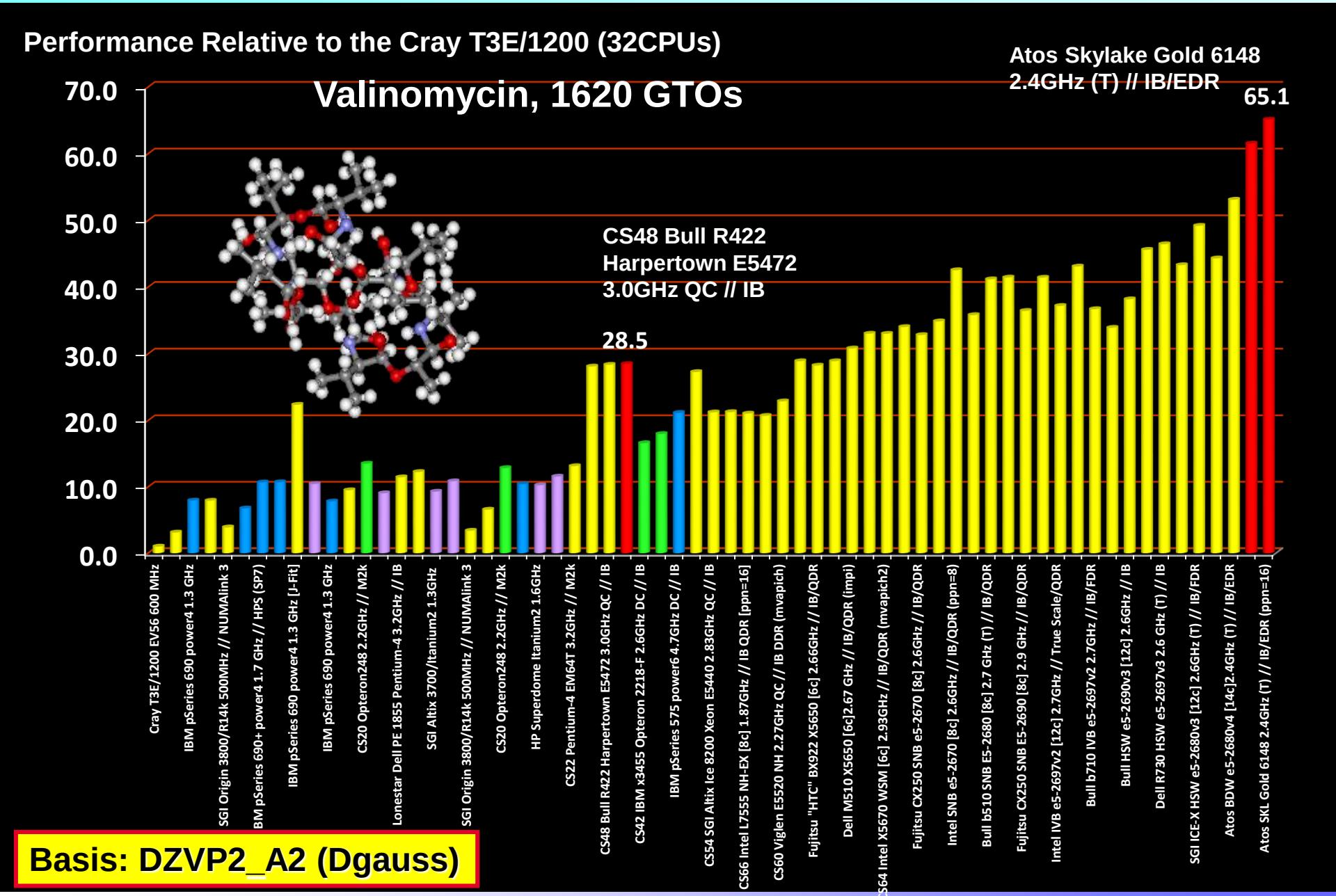
- GAMESS-UK now has **two parallelisation** schemes:
  - The traditional version based on the Global Array tools
    - **retains a lot of replicated data**
    - **limited to about 4000 atomic basis functions**
  - Subsequent developments by **Ian Bush** (High Performance Applications Group, Daresbury, now at Oxford University via NAG Ltd.) have extended the system sizes available for treatment by both GAMESS-UK (molecular systems) and CRYSTAL (periodic systems)
    - **Partial introduction of "Distributed Data" architecture…**
    - **MPI/ScaLAPACK based**

The GAMESS-UK Benchmarks

Five representative examples of increasing complexity.

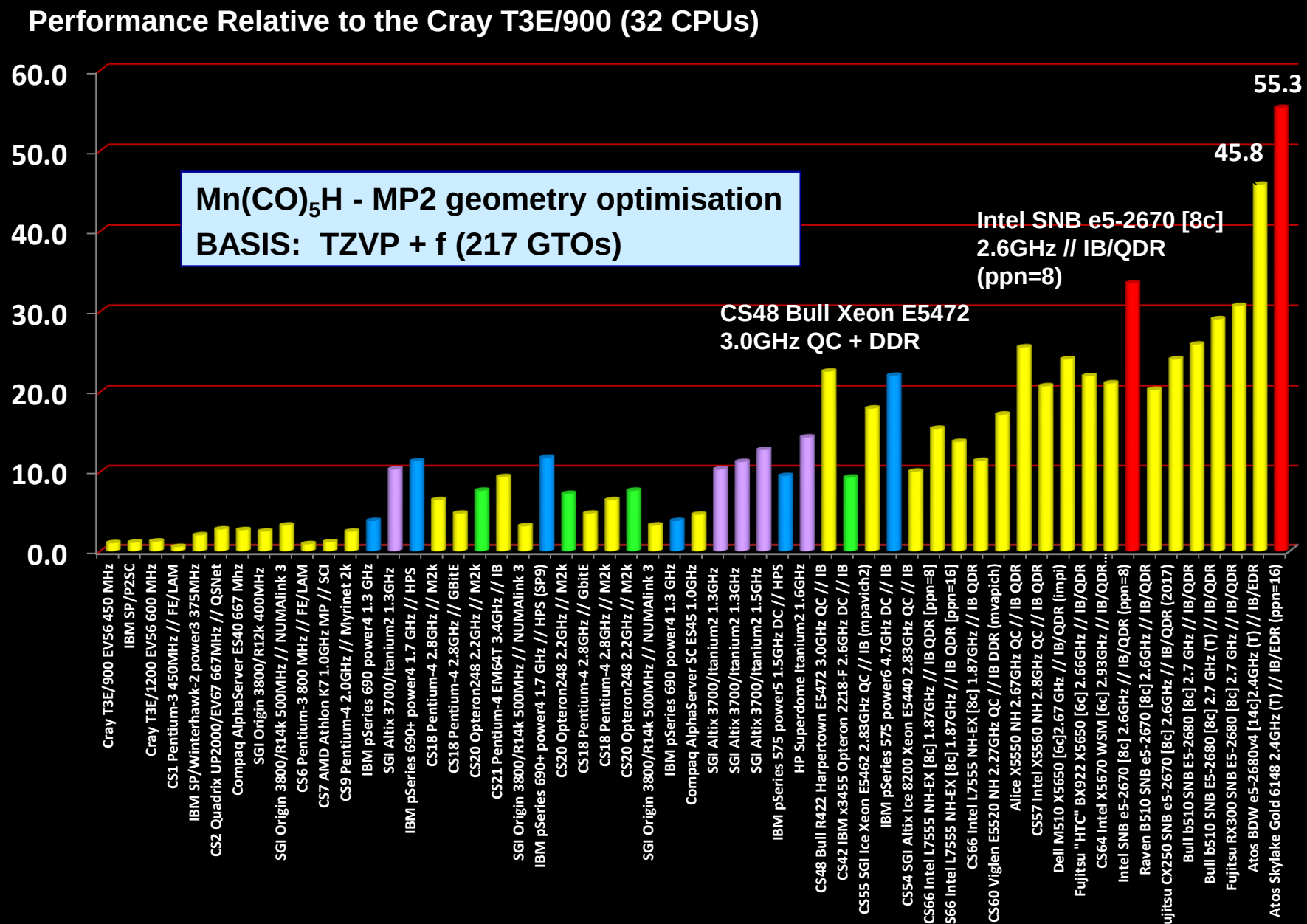
- **Cyclosporin** 6-31g basis (1000 GTOs) **DFT B3LYP** (direct SCF)
- **Cyclosporin** 6-31g-dp basis (1855 GTOs) **DFT B3LYP** (direct SCF)
- **Valinomycin** (dodecadepsipeptide) in water; DZVP2 DFT basis, **HCTH** functional (1620 GTOs) (direct SCF)
- **$Mn(CO)_5H$** TZVP/DZP **MP2** - geometry optimization
- **$((C_6H_4(CF_3))_2$** 6-31g basis DFT B3LYP opt geom + analytic 2nd Derivatives

GAMESS-UK. DFT B3LYP Performance

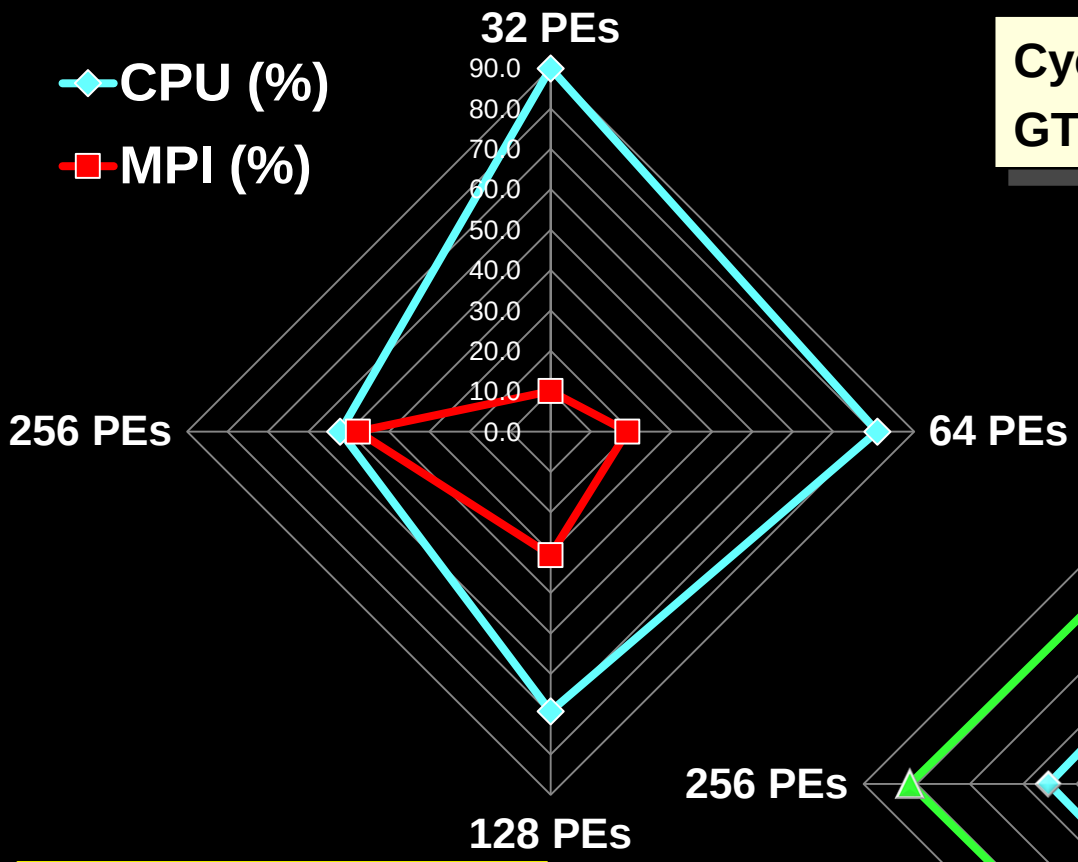


Basis: DZVP2_A2 (Dgauss)

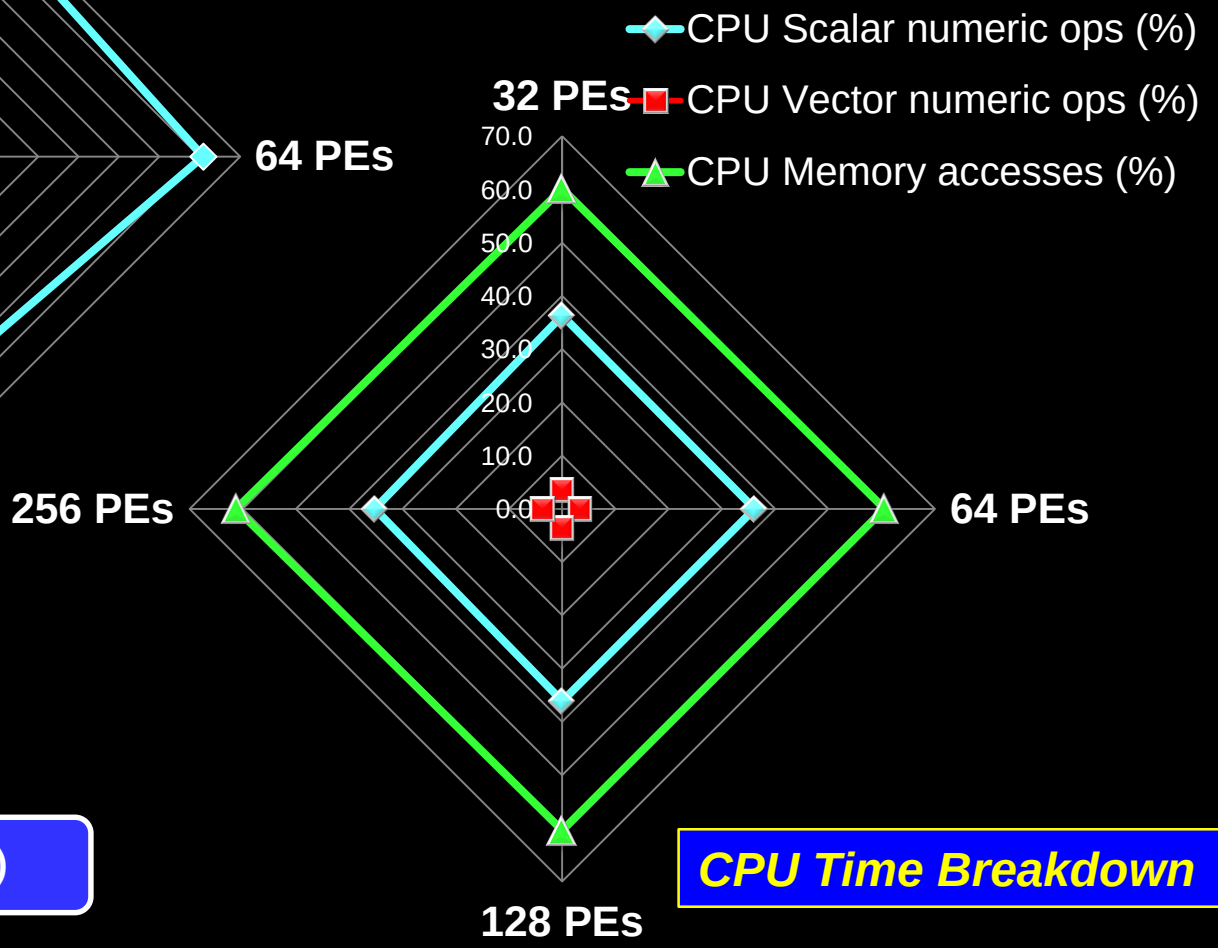
Performance of MP2 Gradient Module



GAMESS-UK – DFT Performance Report



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

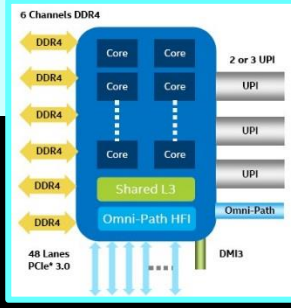


Total Wallclock Time Breakdown

Performance Data (32-256 PEs)

CPU Time Breakdown

Summary



1. Introduction – DL_POLY and GAMESS-UK
   - Background and Flagship codes for the UK's CCP5 & CCP1
   - Critical role of collaborative developments
2. HPC Technology - Processor & Interconnect Technologies
   - The last 10 years of Intel dominance – Nehalem to Skylake
3. DL_POLY and GAMESS-UK Performance
   - Benchmarks & Test Cases
   - Overview of two decades of Code Performance: From T3E/1200E to Intel Skylake clusters
4. Understanding Performance – Useful Tools
5. Acknowledgements and Summary

Acknowledgements

- *Ludovic Sauge,* ***Enguerrand Petit****,* ***Martyn Foster*** *and Nick Allsopp and John Humphries* (Bull/ATOS) for informative discussions and access to the Skylake & EPYC clusters at the Bull HPC Competency Centre.
- ***David Cho****, Gilad Shainer, Colin Bridger & Steve Davey* for access to and considerable assistance with the "*Helios*" cluster at the HPC Advisory Council.
- Joshua Weage, **Martin Hilgeman**, Dave Coughlin, Gilles Civario and Christopher Huggins for access to, and assistance with, the variety of Skylake and EPYC SKUs at the Dell Benchmarking Centre.
- Alin Marin Elena and Ilian Todorov (STFC) for discussions around the DL_POLY software
- The DisCO programme at Daresbury Laboratory.

Final Thoughts & Summary

I. **Performance Benchmarks** and **Cluster Systems**
   a. **Synthetic Code Performance**: *STREAM and IMB*
   b. **Application Code Performance**: *DLPOLY, GROMACS, AMBER,GAMESS_UK, VASP and Quantum Espresso*
   c. **Interconnect Performance**: Intel MPI and Mellanox's HPCX
   d. ***Processor Family and Interconnect*** *– "core to core" and "node to node" benchmarks*

II. **Impact of Environmental Issues *in Cluster acceptance tests.***
   a. **Security patches, turbo mode and Throughput testing**

III. **Performance profile** of **DL_POLY** and **GAMESS-UK** over the **past two decades**