DiRAC @ Oracle Cloud

Experiences of porting and running an HPC benchmark suite on the Oracle bare metal cloud

Andy Turner, EPCC - DiRAC Technical Manager

a.turner@epcc.ed.ac.uk

DiRAC | epcc | THE UNIVERSITY of EDINBURGH | UNIVERSITY OF CAMBRIDGE | ORACLE

People who **actually** did the work!

- DiRAC RSEs
  - Michael Bareford, EPCC, University of Edinburgh
  - Alexei Borissov, University of Edinburgh
  - Arjen Tamerus, University of Cambridge
- Oracle HPC Team
  - Andy Croft
  - Stuart Leeke
  - Arnaud Froidmont
- Thanks to Oracle (particularly, Paul Morrissey) for arranging access to the resources

DiRAC Benchmark Suite

DiRAC Applications

- Extreme Scaling
  - Grid: Data parallel C++ library for QCD modelling
- Memory Intensive
  - SWIFT: Cosmological modelling
- Data Intensive
  - AREPO: Cosmological modelling
  - RAMSES: Galactic modelling
  - sphNG: Astrophysics modelling
  - TROVE: Molecular rovibrational spectra modelling

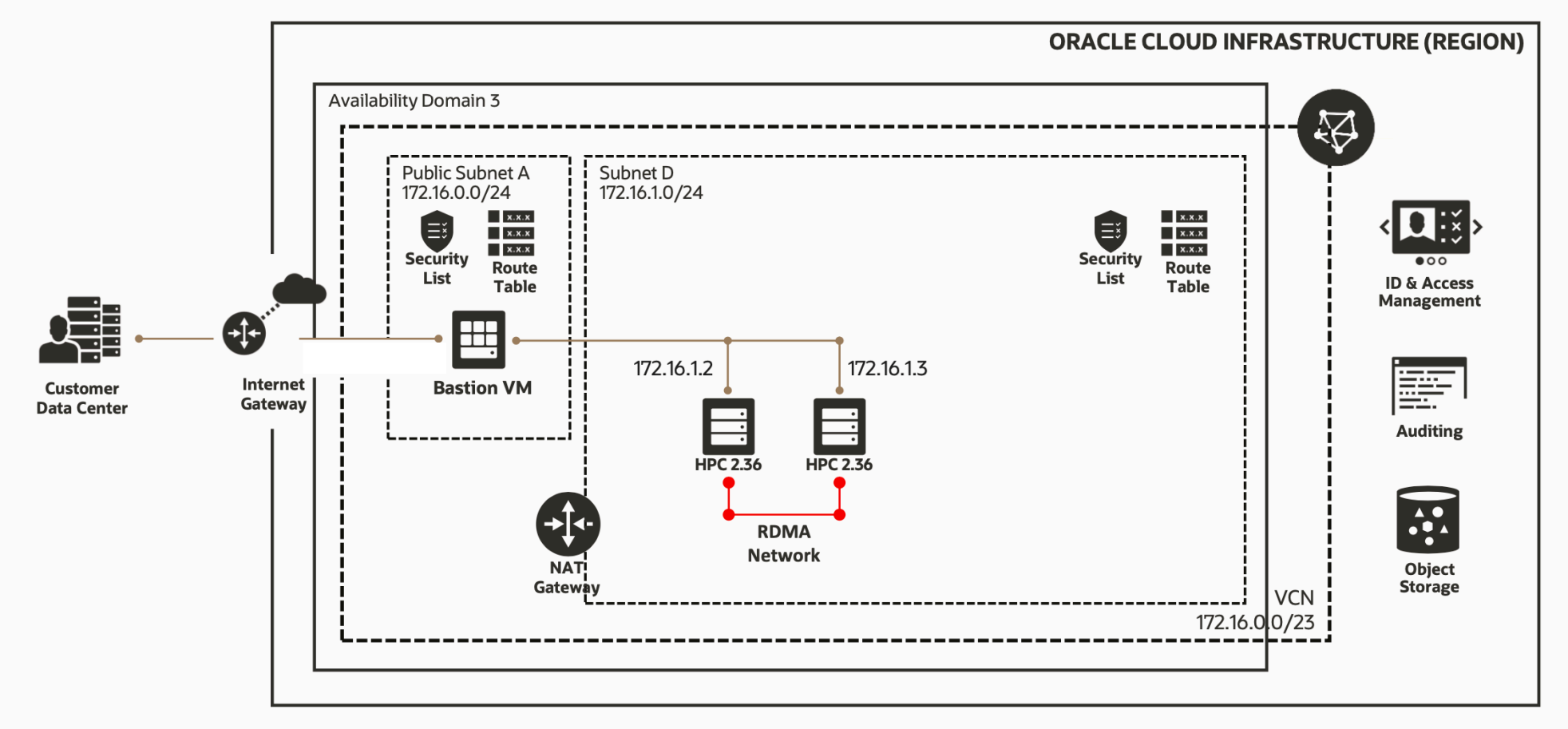
DiRAC

Oracle Cloud

Oracle Bare Metal Cloud

- Each node: BM.HPC2.36 HPC Instance
  - 2x 3.0GHz Xeon 6154 (Skylake), AVX2: 2.6GHz, AVX512: 2.1GHz
  - 384 GB DDR4-2666
  - 6.7 TB NVMe local storage
- Mellanox ConnectX-5, 100 Gbps network interface cards with RDMA over converged Ethernet (RoCE)
- Shared NFS mount
- Oracle Linux (based on RHEL)
- GCC Compilers, OpenMPI

ORACLE®



2 compute node example

ORACLE

Porting


THE #1 PROGRAMMER EXCUSE FOR LEGITIMATELY SLACKING OFF:
"MY CODE'S COMPILING."
HEY! GET BACK TO WORK!
COMPILING!
OH. CARRY ON.

Differences from standard HPC systems

- No software modules
- Root access
- Manage the software installation yourself
- Install newer versions of GCC yourself (6 and 8)
  - GCC 4 available by default
  - sudo yum install devtoolset-8
- Recompile OpenMPI against newer GCC
  - Not strictly needed for C/C++ codes but needed for Fortran MPI modules
- Use Oracle custom scripts to configure and launch MPI jobs

Experience

- Porting was straightforward
  - Needed to install more recent GCC versions
  - Recompile OpenMPI against newer GCC
  - Needed to install performance libraries (BLAS/LAPACK, FFTW) - current versions installed may not be optimal
- Running
  - MPI based on Oracle-provided scripts - could be made more user-friendly with a small amount of work
- An updated HPC image with more built in would be a useful addition
  - Newer compilers, optimised numerical libraries, improved scripts for running MPI jobs

Benchmarks

- sphNG not working yet
  - This has been the most difficult code to port across all systems
  - Not Oracle-specific issues

Performance

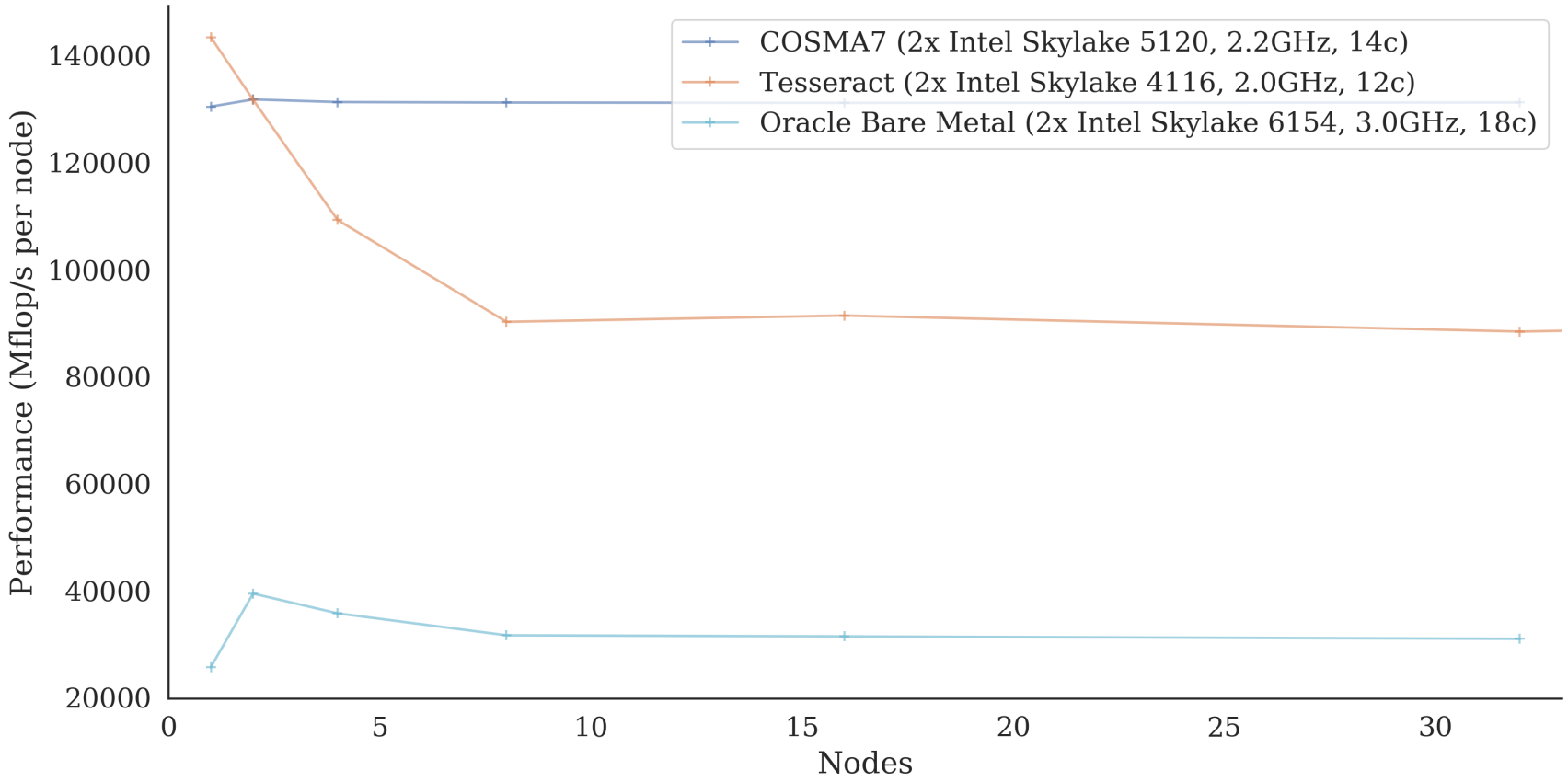
Benchmark Systems

System	Processors	Memory	Interconnect	Notes
Extreme Scaling (Tesseract), Edinburgh	Intel Xeon 4116 (Skylake Silver), 2.2 GHz, 12c	96 GB DDR4-2400	Dual rail Intel OPA	Optimised for interconnect performance
Memory Intensive (COSMA7), Durham	Intel Xeon 5120 (Skylake Gold), 2.2GHz, 14c	512 GB DDR4-2400 (only 4 memory banks populated)	Mellanox EDR	Optimised for memory capacity
Data Intensive (Peta4-Skylake), Cambridge	Intel Xeon 6142 (Skylake Gold), 2.6GHz, 16c	384 GB DDR4-2666	Single rail Intel OPA	General-purpose HPC
Data Intensive (DIaL), Leicester	Intel Xeon 6140 (Skylake Gold), 2.3GHz, 18c	192 GB DDR4-2666	Mellanox EDR	General-purpose HPC
Oracle Cloud	Intel Xeon 6154 (Skylake Gold), 3.0GHz, 18c	384 GB DDR4-2666	RoCA (100 Gbps)	

Grid

- Data parallel C++ library aimed at Quantum chromodynamic (QCD) modelling
- DiRAC_ITT weak scaling benchmark

- Balance of interconnect latency/BW to compute power is key
- Note updated COSMA7 result. Previous poor performance due to pinning issues.

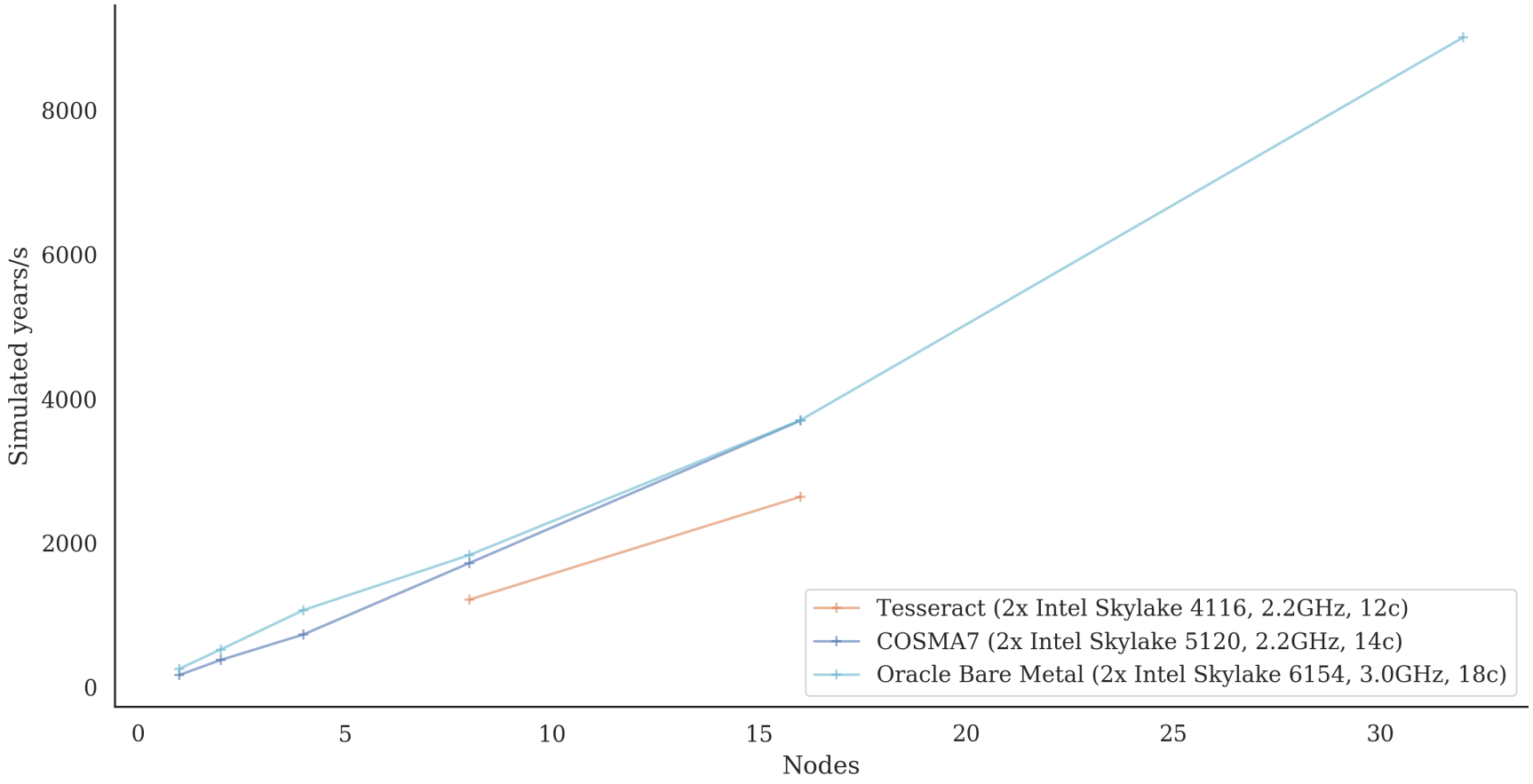


https://github.com/paboyle/Grid/wiki/Dirac-ITT-Benchmarks

AREPO

- Cosmological magnetohydrodynamical moving-mesh simulation code
- Employs both N-body dynamics and grid-based Fourier methods
- DiRAC strong scaling benchmark

- Performance seems to broadly follow floating point performance

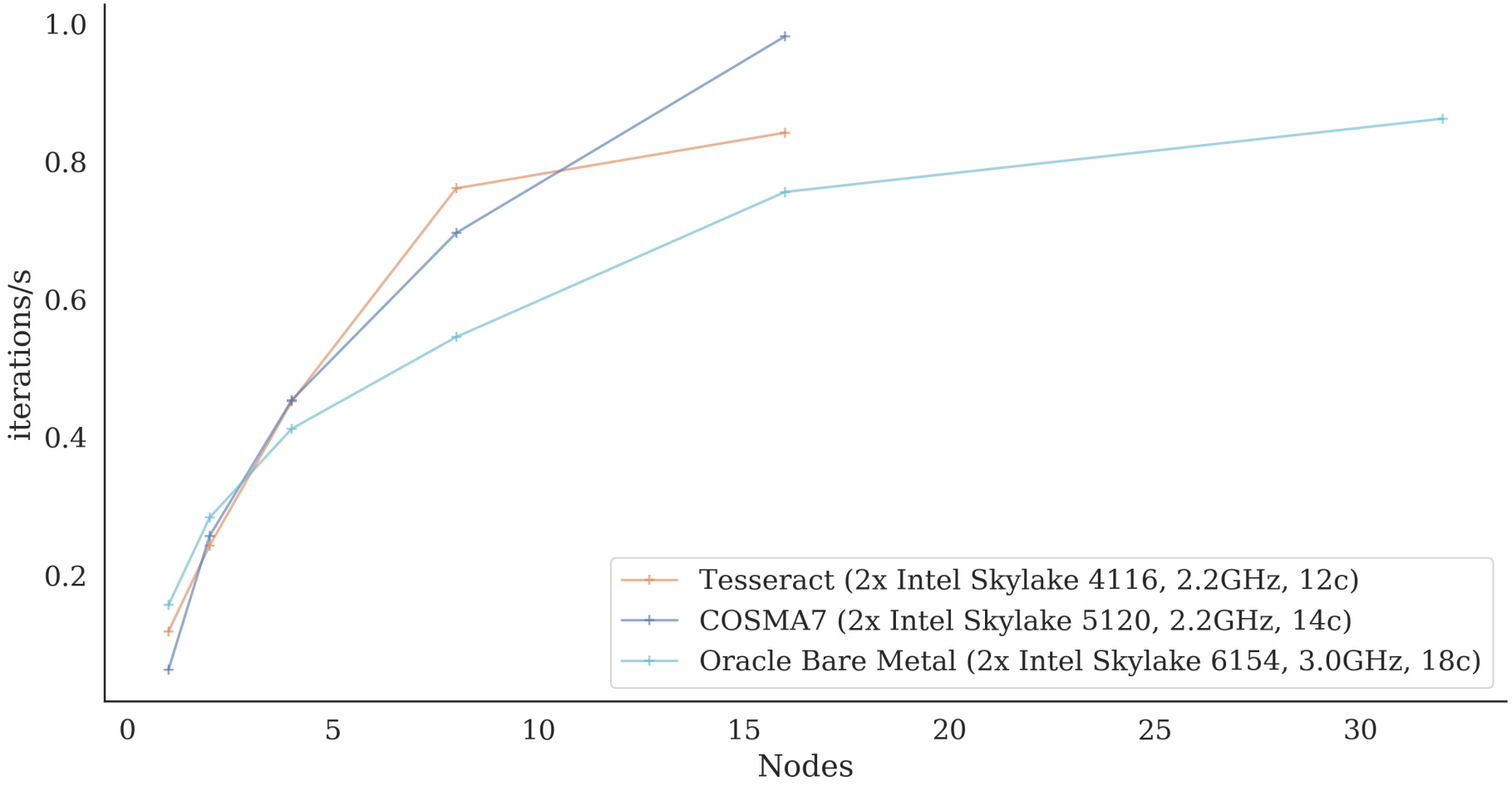


https://arepo-code.org/

RAMSES

- Astrophysics (galactic structure/dynamics)
- Adaptive mesh refinement
- DiRAC strong scaling benchmark

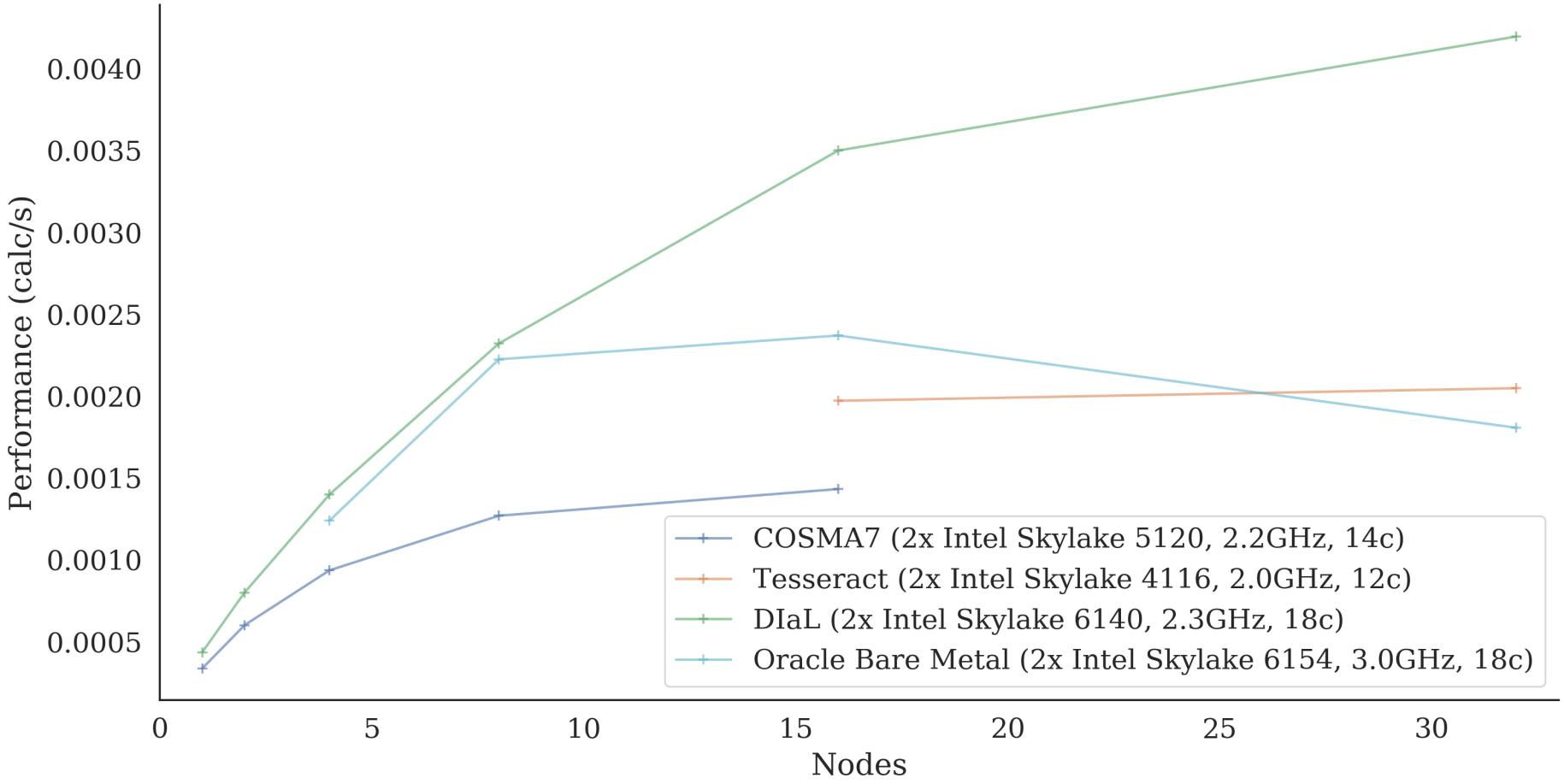
- Performance and scaling dependency seems more complex than other benchmarks



https://www.ics.uzh.ch/~teyssier/ramses/RAMSES.html

TROVE

- Molecular line structure modelling
- Large matrix diagonalization problem

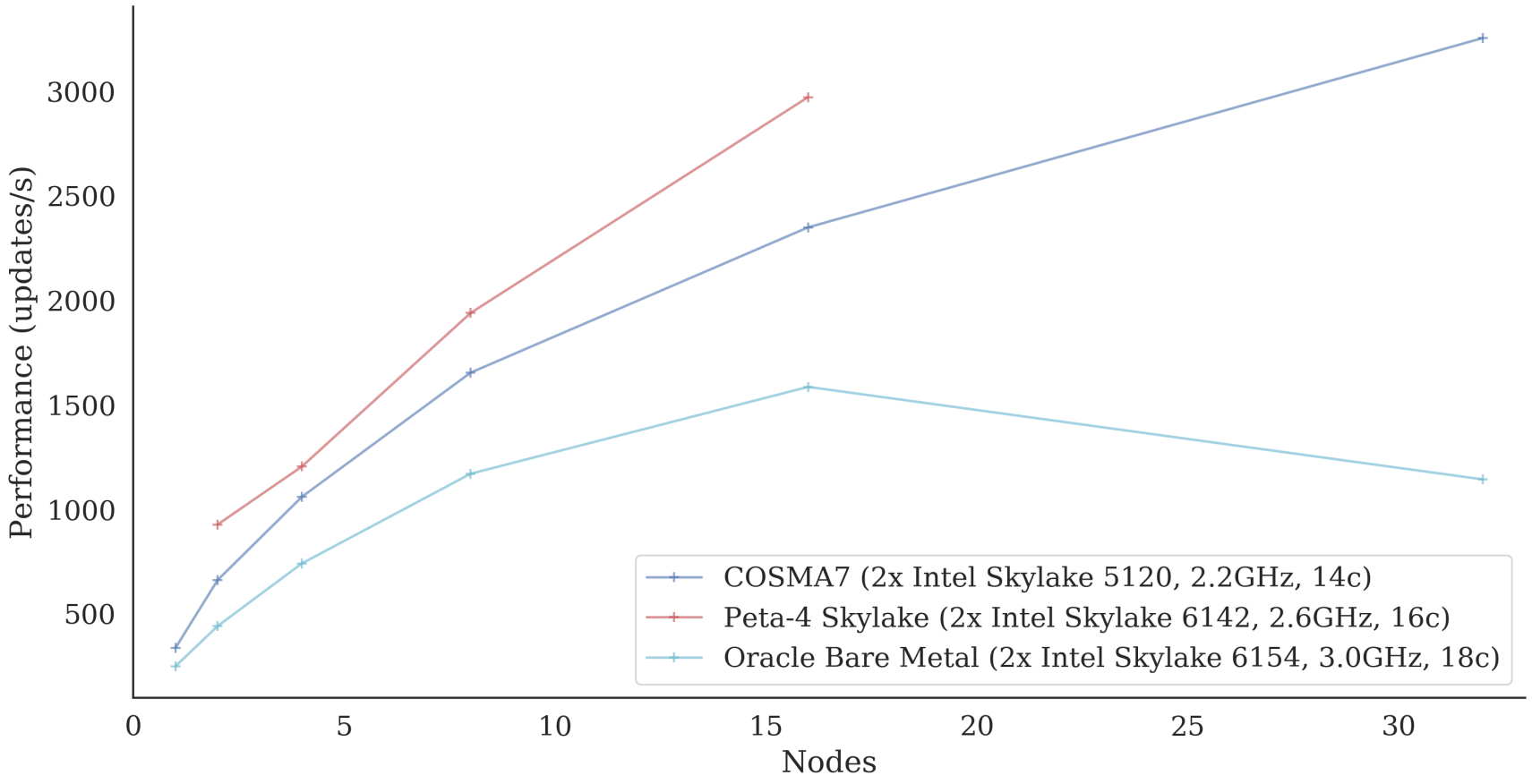


https://www.ucl.ac.uk/~ucapsy0/

SWIFT

- Smoothed Particle Hydrodynamics (SPH) and gravity code for astrophysics and cosmology
- DiRAC strong scaling benchmark

- Performance differences mostly attributed to compiler and library differences – Intel compilers and libraries not currently available on Oracle HPC test platform



http://swift.dur.ac.uk/

SWIFT I/O Benchmarking

- Parallel I/O is a large part of SWIFT use:
  - Snapshot files: 370 GB in benchmark
  - Restart files: 994 GB in benchmark

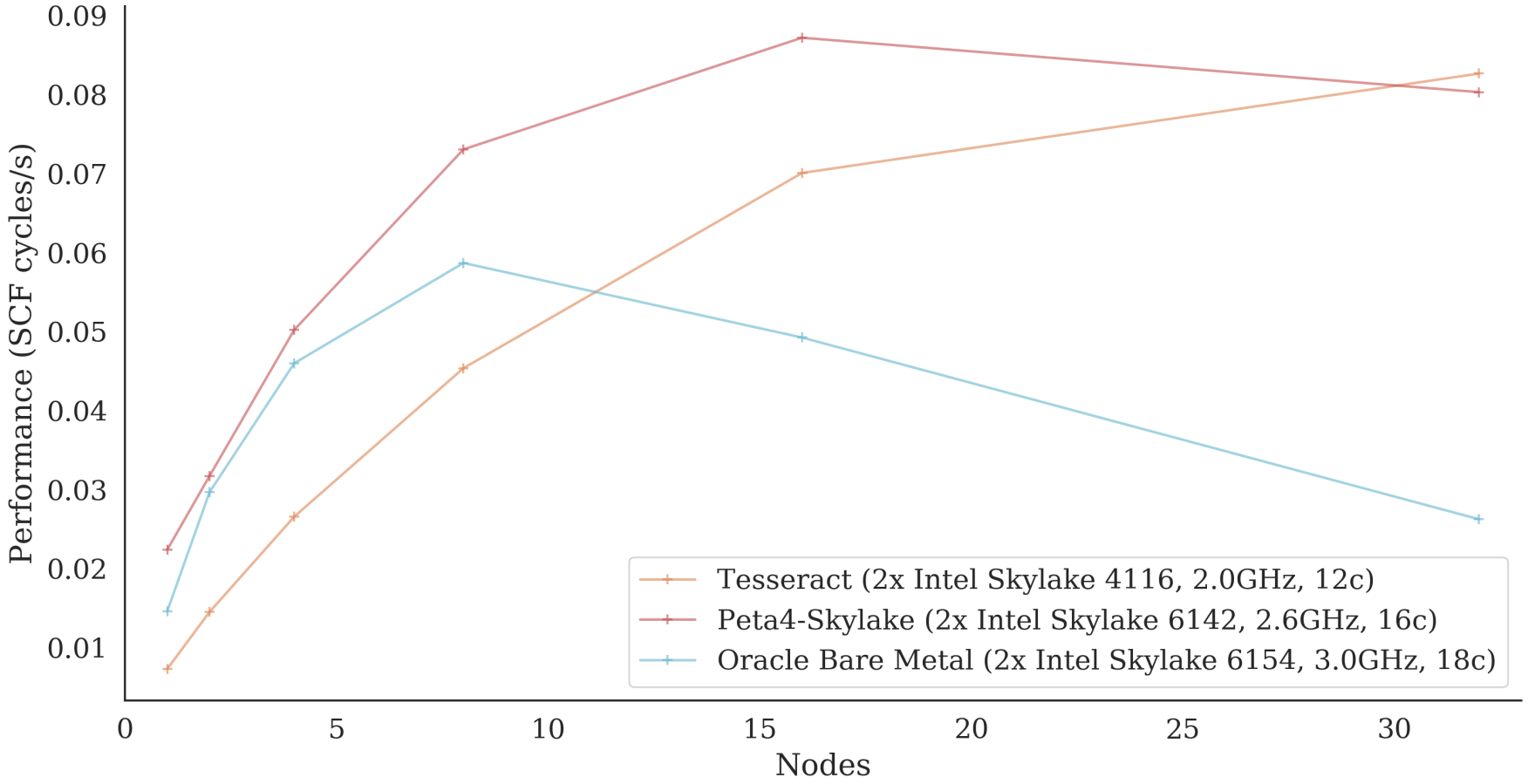
System	Snapshot write time (BW)	Restart write time (BW)
COSMA7 (16 nodes)	53s (7.0 GB/s)	34s (29.2 GB/s)
Oracle Cloud (16 nodes)	59s (6.2 GB/s)	14s (70 GB/s)

- COSMA7: Lustre over EDR IB, NVMe-based
- Oracle Cloud: BeeGFS over RoCE, NVMe-based

CASTEP

- Plane wave DFT
- Materials modelling
- Al Slab strong scaling benchmark

- Compute/memory-bound at low node counts (LAPACK ZGEMM) – Oracle performance due to not well optimised BLAS/LAPACK in this regime
- Becomes MPI latency-bound at higher node counts (MPI_Alltoallv)



http://www.castep.org/CASTEP/Al3x3

Next steps

Next steps

- Publish benchmarking report
- Publish DiRAC benchmark repositories, including:
  - Benchmarks themselves
  - Information on how we compiled them
  - Job submission scripts
  - Full output from benchmark runs
  - Analysis scripts – to show how we got our results
- Publish results from benchmarking on AMD Rome
- Approach other cloud providers to benchmark on their HPC offerings
- Collaborate more closely with other benchmarking exercises: e.g. ARCHER2 and ExCALIBUR

Summary

Performance and scaling

- Single node performance
  - No discernable performance overhead compared to running natively
- Multi-node performance and scaling
  - Similar to single rail IB (EDR/OPA) performance for most benchmarks
  - Some benchmarks show drop off in performance as nodes increase – needs further investigation
  - Plan to compare performance on new HPC interconnect technology (HDR IB and Cray Slingshot) once these systems are available
- Parallel IO performance on BeeGFS over RoCE using NVMe was very good

Things we learned

- Gap in expectation/understanding between DiRAC and Oracle was large at the start of the project
  - DiRAC were expecting something closer to a standard HPC environment
  - Oracle had less experience with what standard HPC environment looked like and the technical skill level of a typical DiRAC HPC user
- Had an extremely positive working relationship
  - We now both understand each others' experience/expectations much better
  - Both sides have learned a lot from each other – which was one of the major points of this exercise!
- Porting was straightforward once we understood the different environment
- Better documentation on compiling and running MPI on bare metal instances would have been useful