CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

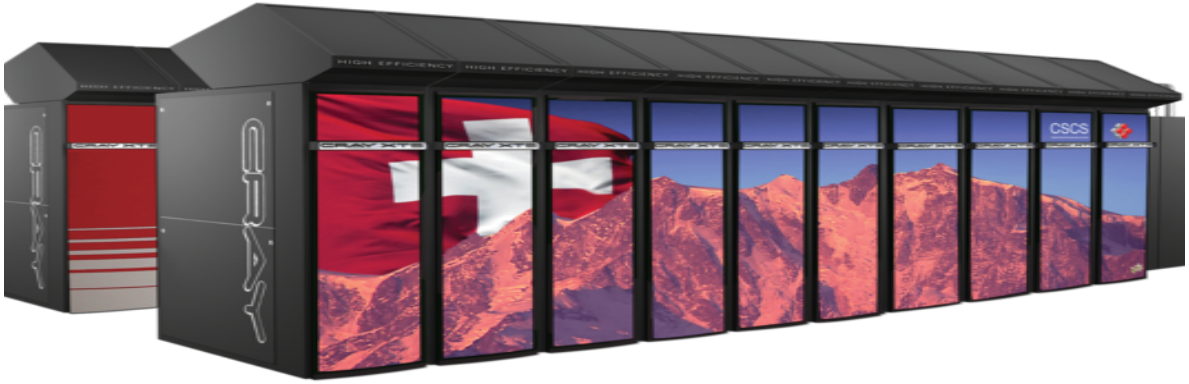
First 12 cabinets of Cray XC30 at CSCS: Scaling and Performance Efficiencies of Applications

Themis Athanassiadou
Tim Robinson, Sadaf Alam, Jeffrey Poznanovic,
Jean-Guillaume Piccinali, Gilles Fourestey, Luca Marcella

CSCS

CUG, May 8th 2013

Overview of Monte Rosa: XE6 at CSCS



- **16 cabinets:** 8 login nodes, 1496 compute nodes and 21 service nodes
- **Compute node:** 2x AMD Opteron 6272 (Interlagos) @ 2.1 GHz
  → 32 cores, 32 GB of memory per node
- **Network:** Gemini communications processor and 3D torus interconnect.

47,872 physical cores, 47 TB memory, 402 TF peak

Overview of Piz Daint : XC30 at CSCS

- **12 cabinets:** 4 login nodes, 5 external login nodes, 2256 compute nodes and 24 service nodes
- **Compute node:** 2x Intel Xeon E5-2670 (SandyBridge) @ 2.60GHz
  → 16 cores, 32 GB of memory per node
- **Network:** Each 4 compute nodes connected to an Aries network chip. Network chips interconnected using a dragonfly topology.

36,096 physical cores, 72 TB memory, 750 TF peak

Programming Environment: Rosa Vs Daint

• The programming environment on Piz Daint is similar to that already found on Monte Rosa:

– Compilers from Cray and Gnu are fully supported by Cray.
Intel supported as 3rd party software
PGI is not available

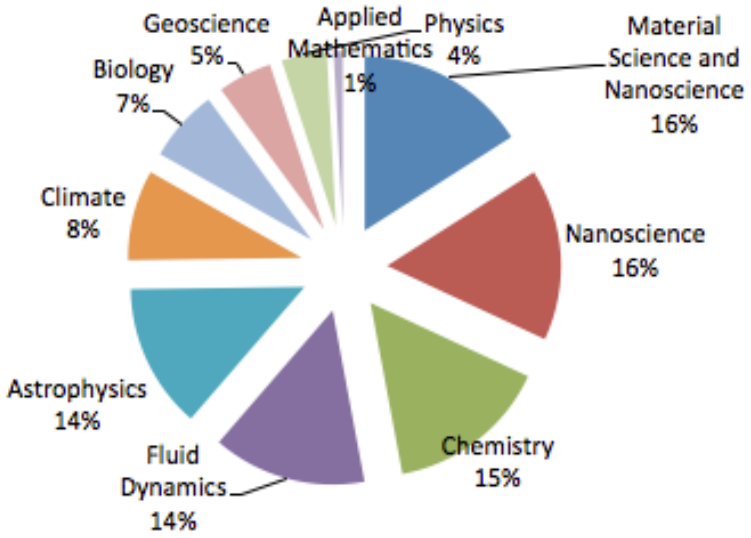
– Usual numerical and scientific libraries are provided

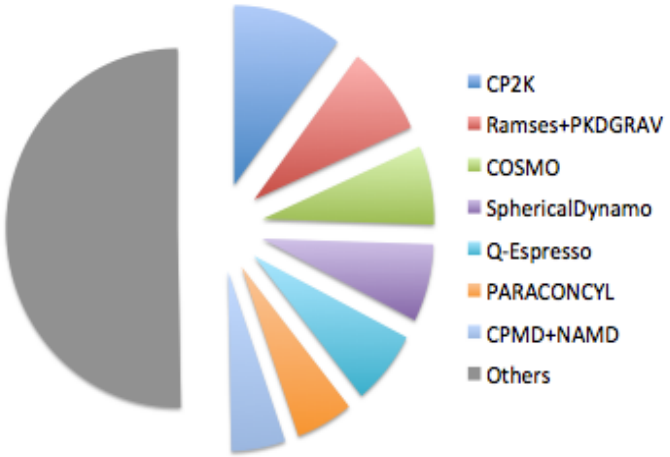
– Modules framework for environment management

– CrayPAT for performance analysis

• Slurm is used as the resource management system

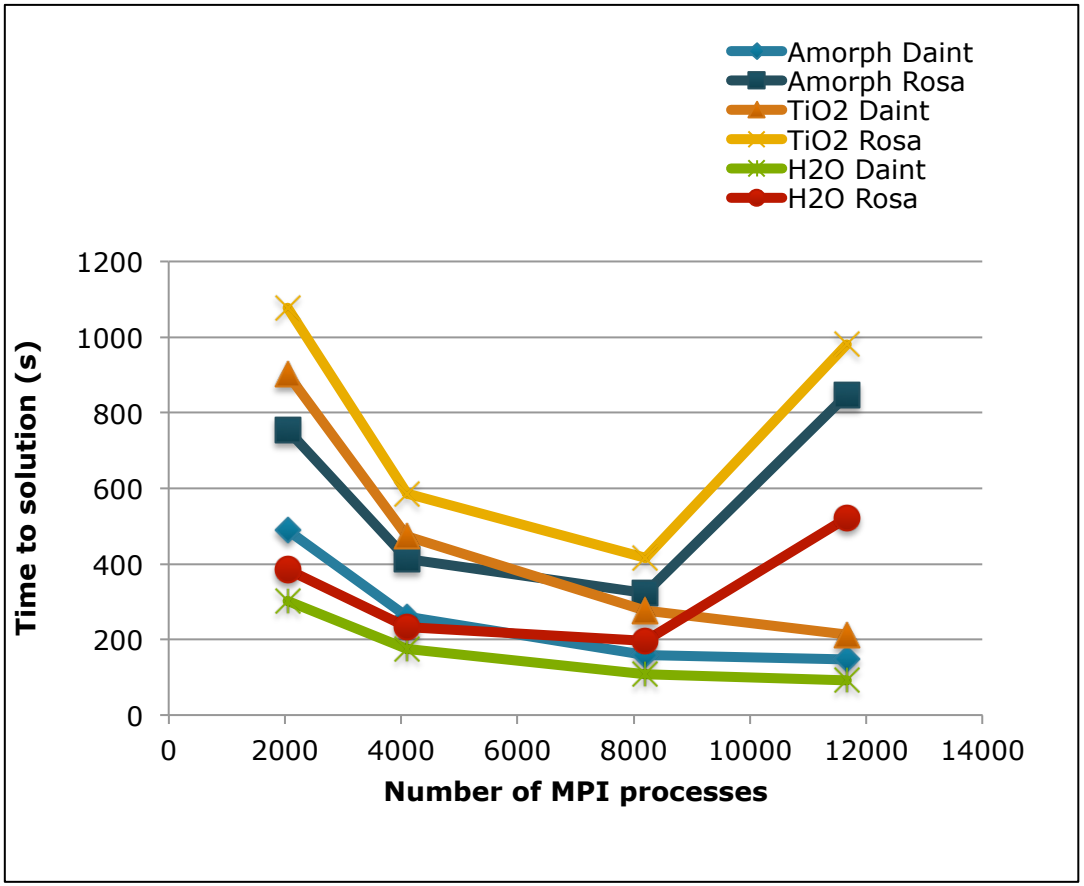
Applications at CSCS





Others: PPM, MRAG, ECHAM, ORBS, OMEN, ELEPHANT, LatField2D, GROMACS, SECRET, LAMMPS, GASOLINE, DLPOLY, IMPACT, CHARMM, 3D-DNS, ARPS, ...

Chemistry/material science: CP2K



CP2K: predictive quantum simulation tool based on Density Functional Theory and used in material science and chemistry.

Benchmark setup:

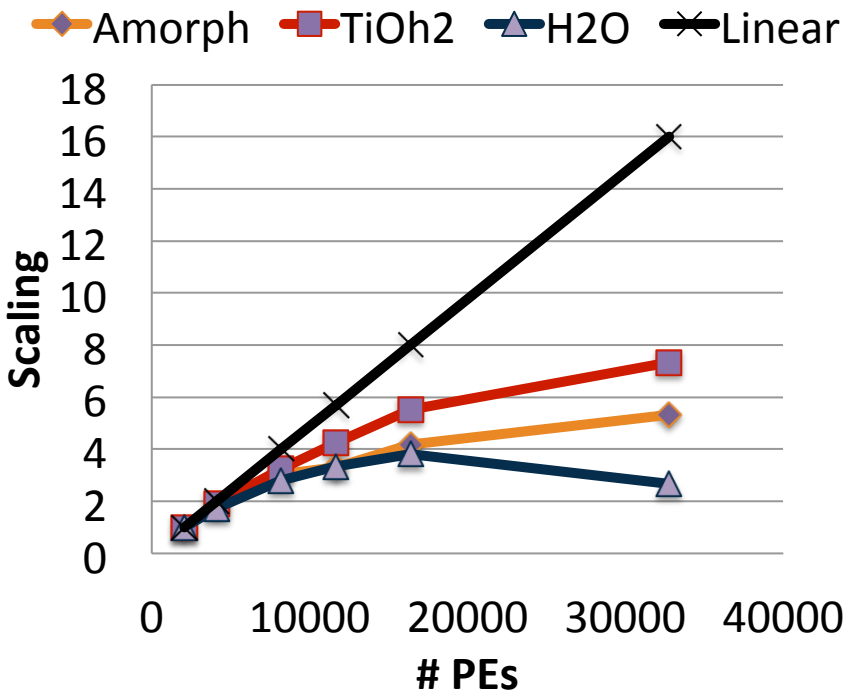
- On Daint: 1 MPI process per core
- On Rosa: 1 MPI process + 2 threads per module.

- Speed-up is between 1.2 (low node count) and 3 (high node count)

- Weird behavior on rosa when the number of PEs is not a power of 2 does not exist on Daint

CP2K Scalability on Piz Daint

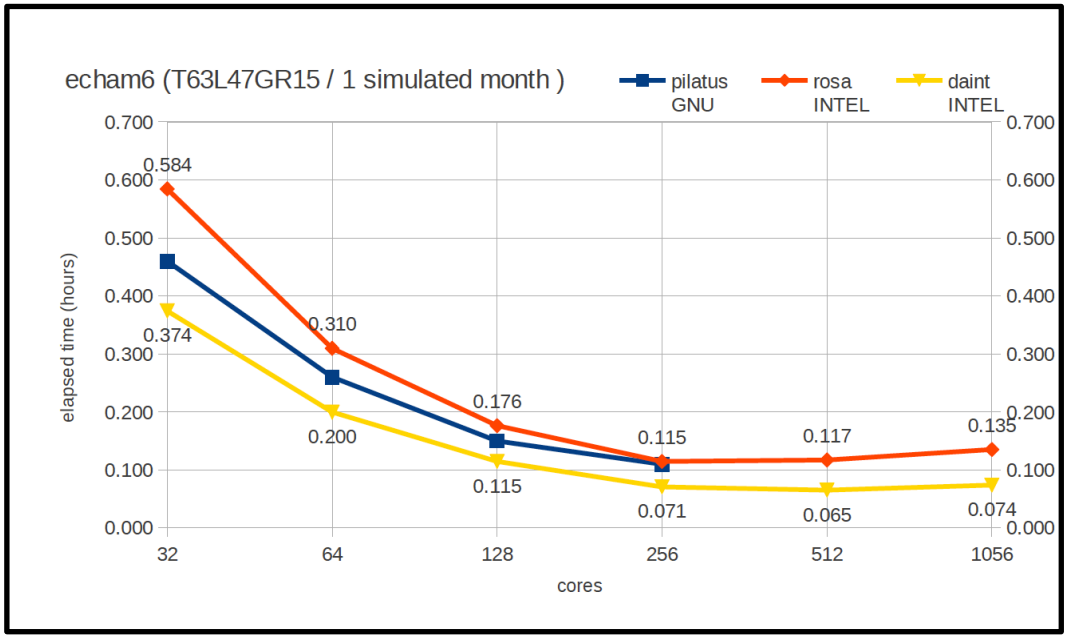
# Pes	Amorph	TiOh2	H2O
32768	91	123	113
16384	116.6	163	79
11664	146	213	90
8192	158.8	277	108
4096	259.5	472	174
2048	485.5	902	302



Performance on Daint (Time To Solution in s.)
and Scalability with respect to 2048 PEs

Scalability plateaus for a large number of MPI processes possibly due to the problem size (not enough data for large number of PEs)

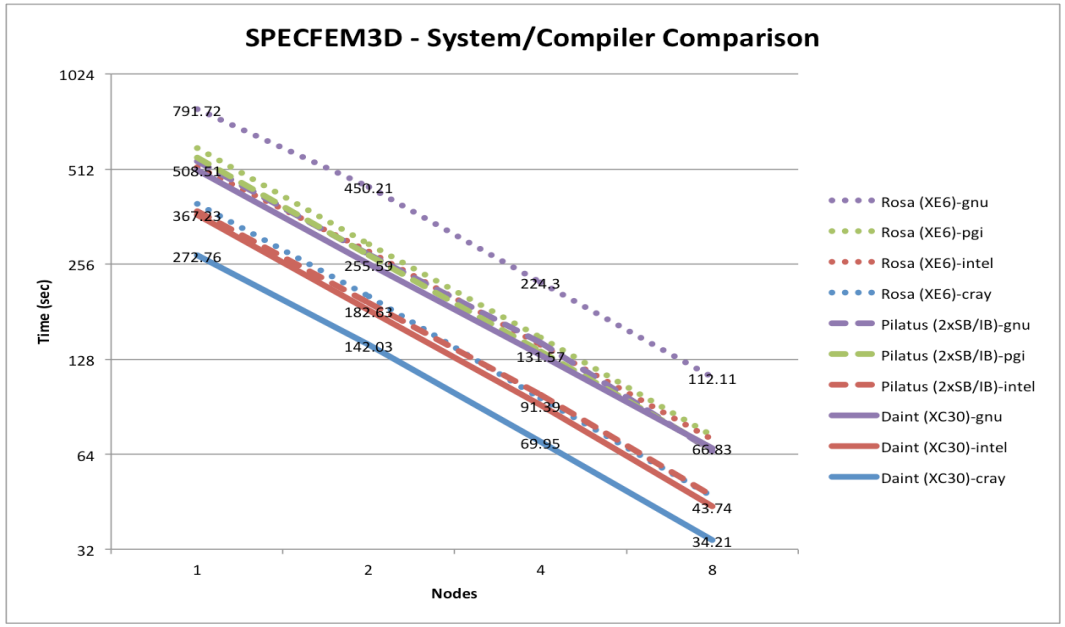
Climate: echam6



- Global climate model (Max Plank institute)
- MPI bottleneck from mpi_waitall and mpi_bcast
- No gain from hyperthreading on Daint beyond 8 nodes.

The best overall performance is obtained with the INTEL compiler on DAINT which is outperforming Monte Rosa by up to **45% faster** runtime which can be explained by the faster interconnect (aries). Echam6 scales up to 256 cores.

Earth science: SPECFEM3D



-SPECFEM3D simulates three-dimensional seismic wave propagation

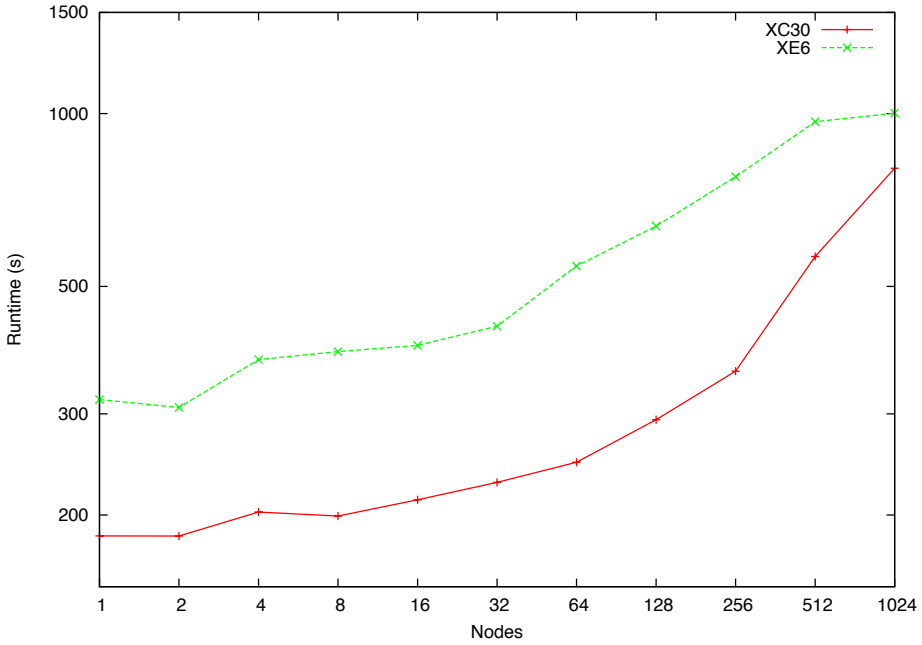
-When comparing with the same compilers, SandyBridge is significantly faster than Interlagos for this particular application

-The crayftn and ifort compilers give the fastest results, while gfortran gives slower results

- When running larger jobs up to 256 nodes, application developers recently reported a significant speedup when running on the XC30 when compared to the XE6

SPECFEM3D performance comparison between SandyBridge, Interlagos and various compilers on a moderate-sized 300K element European dataset

Biology: Gromacs



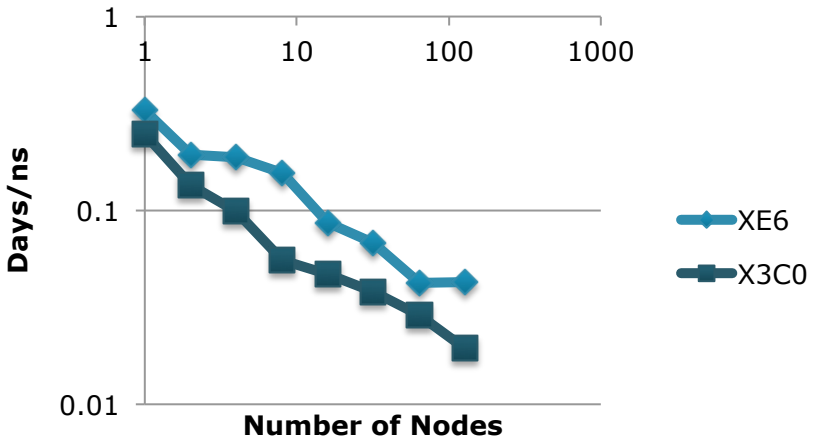
The benchmark compares the performance on the Cray XC30 versus the Cray XE6 of the hybrid OpenMP/MPI Gromacs 4.6, running with 16 MPI tasks per node and 2 OpenMP threads per MPI task.

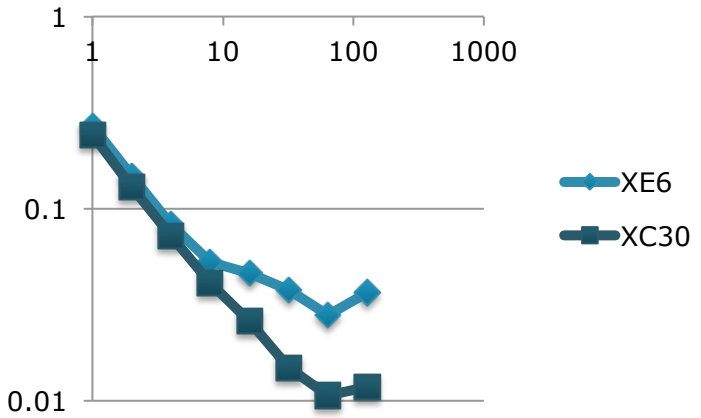
The XC30 shows a better performance than the XE6 by a factor ranging from ~1.7 on one and two nodes, which quickly grows increasing the number of nodes and reaches 2.18 at 256 nodes.

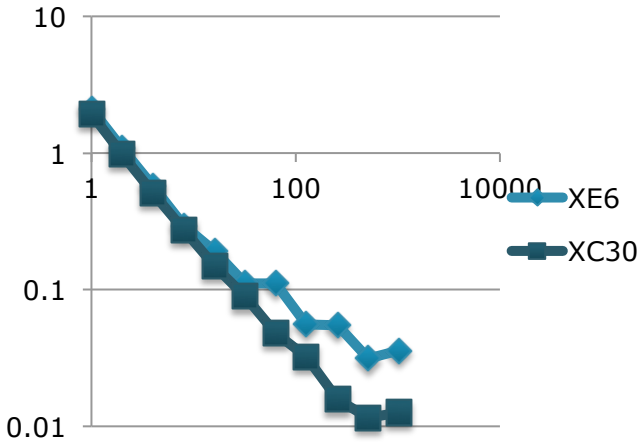
GROMACS is a versatile package to perform molecular dynamics, i.e. simulating the Newtonian equations of motion for systems with hundreds to millions of particles.

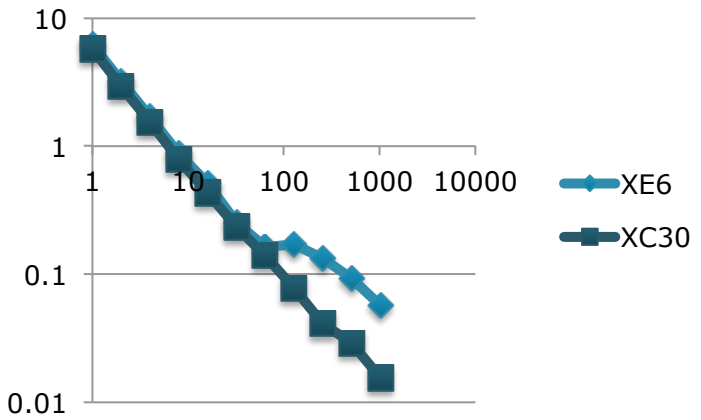
The weak scaling experiment on the XC30 and the XE6 at CSCS has been carried out using water molecules, starting with a box of 32000 water molecules at standard conditions on a single node and increasing the size of the system proportionally with the number of nodes, up to 32768000 water molecules for 1024 nodes.

Biology: NAMD – 2OK, 60K, 0.5M, 1.4M Atoms









Conclusions

Piz Daint has been opened to users since April 1st 2013

- **350 million allocation units/6 months available**
- **Applications have been ported without much difficulty**
- **Increase in performance and scalability over a wide range of applications**
- **As more users get on the system, we will have more results**

Q & A

Network: Aries vs Gemini

Aries features:

- 3X increase in sustained injection BW (10 GB/sec)
  – improves efficiency of communication intensive applications
- Increase in global BW (from 3X at the low end to 20X at high end
  – benefits applications with complex communication patterns.
- Hardware support for collectives
  – helpful on large jobs utilizing collectives

AMD Interlagos VS Intel SandyBridge

- 32 physical cores vs. 16 cores with 32 hyper threads
- 4 NUMA vs. 2 NUMA memories
- 256-bit AVX + FMA4 vs. 256-AVX