LRZ Site Report

Athens, 26.9.- 27.9.2016

Juan Pancorbo Armada
juan.pancorbo@lrz.de

Introduction

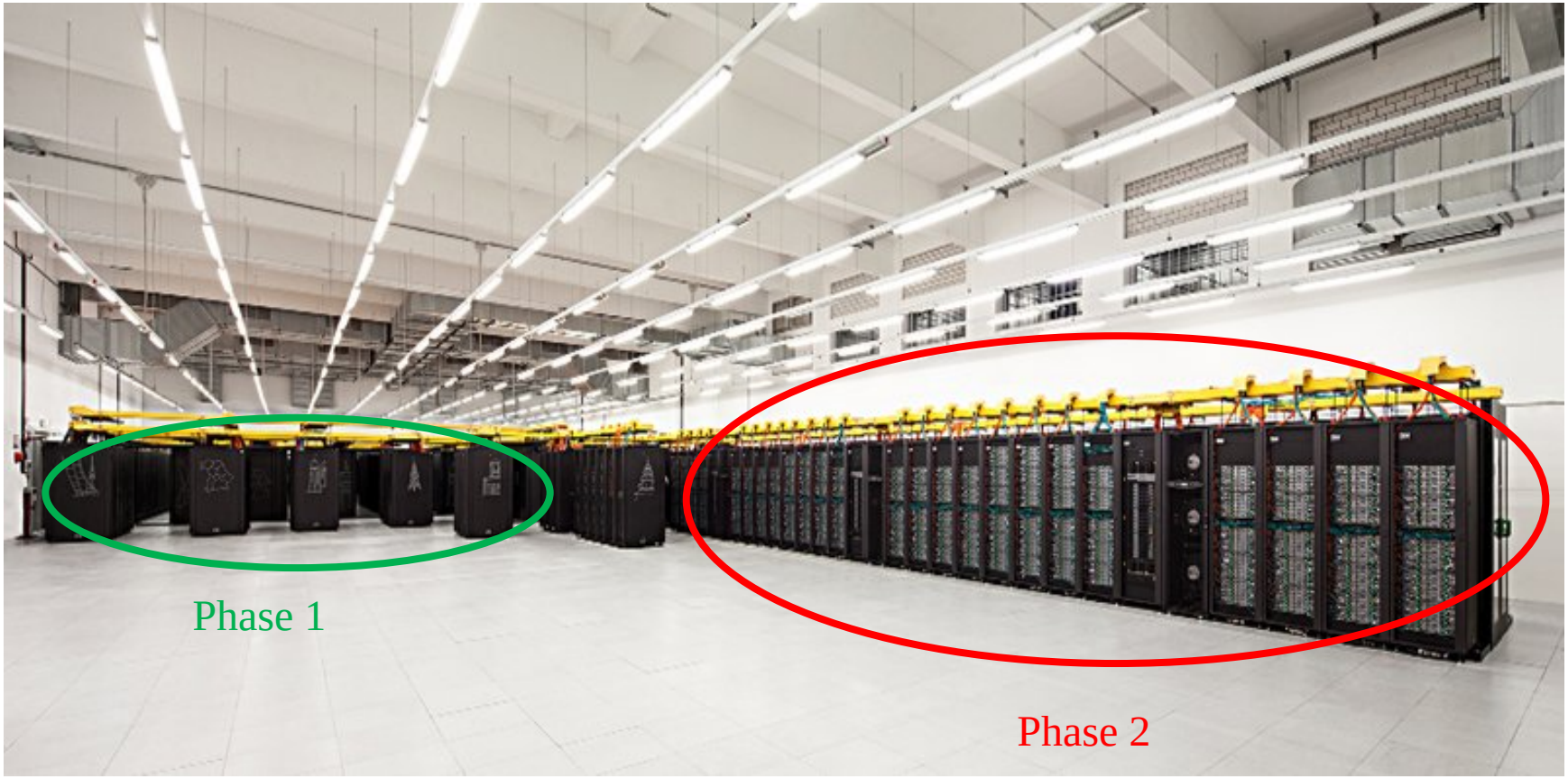
- The LRZ is the computer center for Munich's universities and is part of the Bavarian Academy of Sciences and Humanities.
- As a service provider for scientific high performance computing, LRZ operates compute systems for use by educational institutions in Munich, Bavaria, as well as on the national level.
- LRZ provides own computing resources as well as housing and managing computing resources from other institutions such as Max Planck Institute, Technical University Munich, or Ludwig Maximilians University.
- The LRZ hosts SUPERMUC. A water cooled supercomputer currently ranked on 27th position of top 500 (4th when released in 2012)

SUPERMUC

- In 2015 the phase 2 was installed as a separate instance that has the same performance as the phase 1

Phase	System	Vendor	Total Cores	Rmax (TFlops)	Rpeak (TFlops)	Power (kW)
Phase 1	iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband FDR	IBM/ Lenovo	147,456	2,897.0	3,185.1	3,422.67
Phase 2	Lenovo NeXtScale nx360M5 WCT Xeon E5-2697 v3 2.60GHz Infiniband FDR	IBM/ Lenovo	86,016	2.814	3,580.0	1,722.67

SUPERMUC



Linux cluster

- The tier 2 Linux cluster operated at LRZ is a heterogeneous system with different types of compute nodes, divided into 18 different clusters, each of which is managed by SLURM.
- The various clusters are configured for the different needs and services requested, ranging from single node multiple core NUMAlink shared memory clusters, to a 16-way infiniband-connected cluster for parallel job execution, or an 28-way Gbit Ethernet cluster for serial job execution.
- All the clusters are managed from the same master node using mslurm.

LRZ Resources

LRZ owned clusters			
	Cluster name	Number of Nodes	CPU per Node
1	Inter	14	16,28
2	uv2	1	1920
3	uv3	1	2240
4	mpp1	166	16
5	mpp2	318	28
6	gvs	5	32
7	serial	60	4,8
8	hugemem	12	20
9	superrvs	2	32
10	rsrvd	4	16

LRZ managed & housed clusters			
	Cluster name	Number of Nodes	CPU per Node
11	myri	9	32,48
12	lmu_asc	134	8,16
13	lmu_exc	7	40,64,320
14	tum_geodesy	14	56
15	tum_chem	16	28
16	bsbslurm	34	4
17	lcg	72	24,32,56
18	capp	15	8,48

Slurm @ LRZ

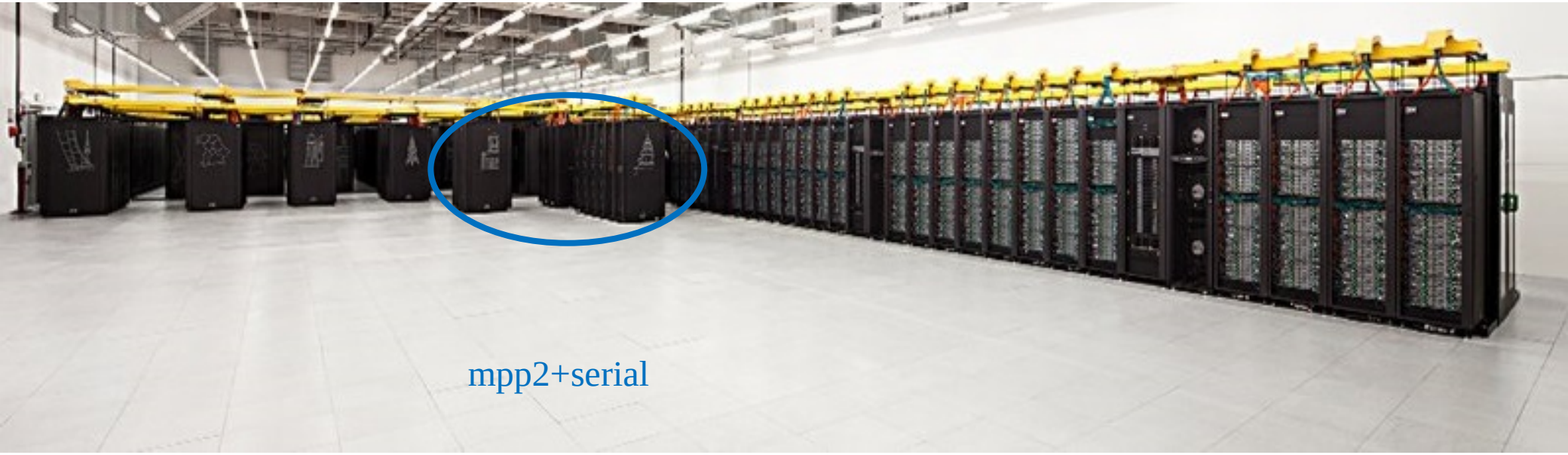
- On May 2016 we upgraded our slurm version from 2.6.9 to 15.08.12
- The best improvement we have noticed is the slurmdbd stability.
  - Only one crash since May and it was because mysql crashed first.
- Also sdiag user information was greatly appreciated
  - Now we can easily detect those “nice” users who send an squeue request each second on a thousands of jobs’ queue
- In the scope of the Montblanc project we are collaborating with Bull/Atos for the development of energy aware batch scheduling algorithms to allow Slurm better energy savings at system software level.

Configuration

- All of our clusters have multifactor plugin activated (except for two of our housed clusters).
  - Fairshare and job age are activated on all of them.
  - On our owned clusters qos is also activated so that the support staff get additional priority when solving tickets
  - Jobsize is also activated on 4 of them.
- Limits, associations, and qos enforced via slurmdbd
- Access restricted to housed clusters via allowgroups parameter on cluster partitions.
- Enforce partition limits activated to avoid jobs waiting in queue for ever.
- Triggers configured to restart the slurmdbd, slurmctlds in case they crash, and to report down and drain states on the nodes.

Energy plugin

- In 2015 after the installation of SUPERMUC phase 2 a new extension for linux cluster of 384 nodes with the same hardware as SUPERMUC phase 2 was acquired.
- 60 of them replaced the old nodes on serial cluster and the rest were included in the new mpp2 (coolmuc2) cluster.
- mpp2 is dedicated to parallel job processing

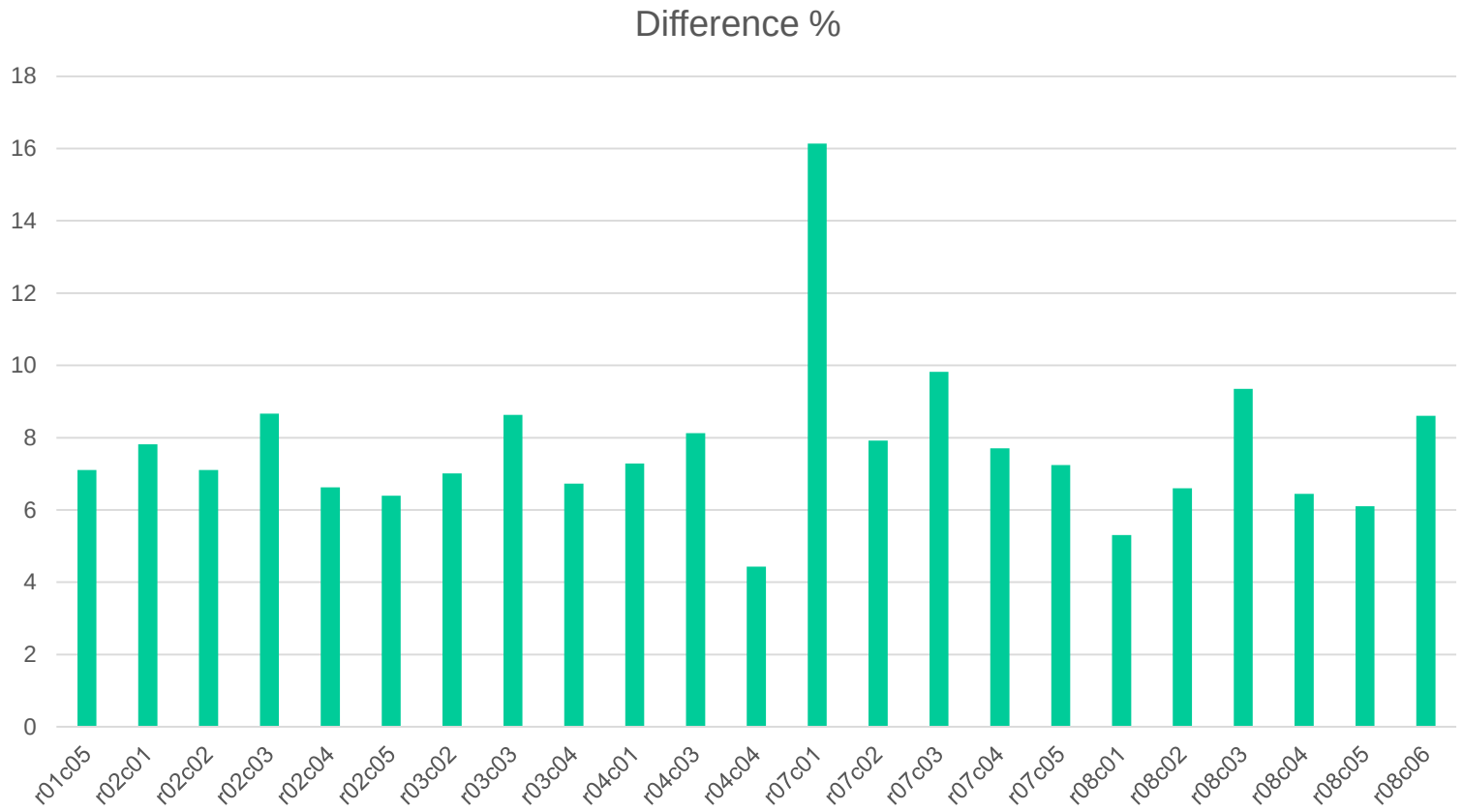


mpp2+serial

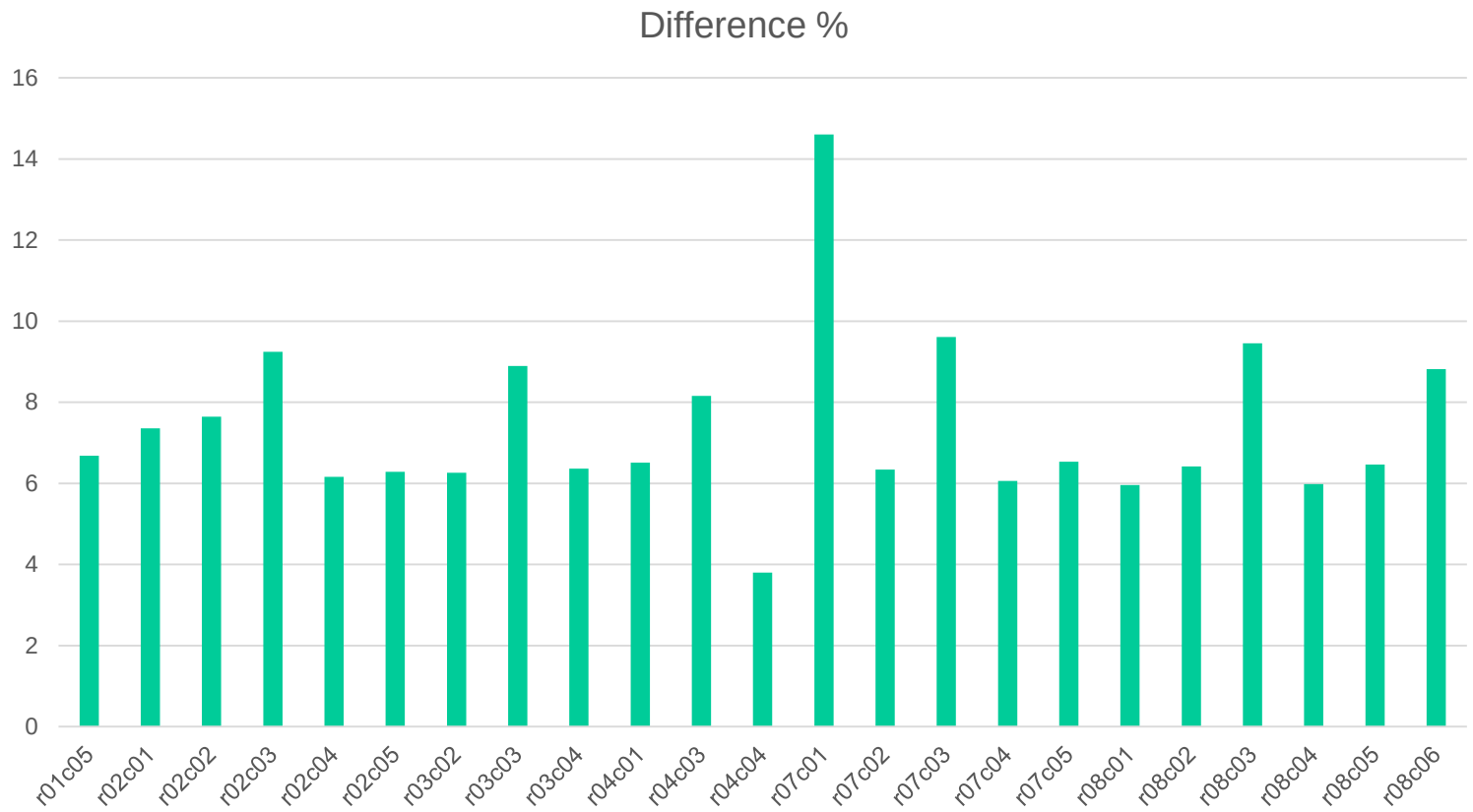
Energy plugin

- The nodes of mpp2 are Lenovo NeXtScale nx360M5 WCT, and have installed IBM Active Energy Manager.
- As part of our research on energy aware scheduling Axel Auweter wrote the slurm energy plugin ibmaem, which reads the energy and power counters of the Active Energy Manager on the nodes.
- To check the results of the energy measurements we ran the Firestarter benchmark (from TU Dresden)to have cpu load.
- In order to have an external energy measurement we took the energy measurements from PDU chassis on the racks. Each rack have up to 6 chassis and each chassi had 12 compute node.
- So we had to run 12 nodes jobs in order to gather the energy in a similar level from slurm and from the PDU’s

Energy measurement test 100% load



Energy measurement test 50% load



References & Credits

- Bibliography:
  - *D. Hackenberg, R. Oldenburg, D. Molka, and R. Schone. Introducing FIRESTARTER: A processor stress test utility. In Proceedings of the International Green Computing Conference(IGCC), pages 1–9, June 2013.*

- Axel Auweter for his implementation of the ibmaem energy plugin
- LRZ staff for their help and support running the tests

Thank you for you attention