Karlsruher Institut für Technologie

Steinbuch Centre for Computing

Performance Comparison of Multi- and "Many"-Core Batch Nodes

Manfred Alef

STEINBUCH CENTRE FOR COMPUTING

Background

- No significant speed-up of single CPU cores since several years
- Servers with multi- and more-core CPUs are providing improved system performance:
  - Until 2005: single-core,
  - 2006 – 2007: dual-core,
  - 2008 – 2009: quad-core,
  - 2010: quad-core with Symmetric Multiprocessing (Hyperthreading) feature,
  - 2011: 12-core, 2 or more CPU sockets (→ up to 48 cores per system)
- Cheap servers with 4 CPU sockets are on the market

Background

- Worker nodes at GridKa (since 2006):

Vendor	CPU *	MHz	L2+L3 Cache (MB) per CPU	Cores	Sockets	Total Cores	
AMD	270	2000	0.5+0	2	2	4	retired
Intel	5148	2333	4	2	2	4	
Intel	5160	3000	4	2	2	4	
Intel	E5345	2333	8+0	4	2	8	
Intel	L5420	2500	12+0	4	2	8	
Intel	$^{E}_{L}$5430	2666	12+0	4	2	8	
Intel	$^{E}_{L}$5520	2266	1+8	4 + HT	2	8	
AMD	6168	1900	6+12	12	2	24	
AMD	6174	2200	6+12	12	4	48	

* In this presentation, the TDP indicator will be omitted, i.e. "5430" is either an "E5430" or a "L5430" chip.

Background

- Worker nodes at GridKa:
  - Hardware details:
    - 2 CPU sockets
      - AMD 6174 box: 4 sockets
    - 2 GB RAM per core
      - Intel 5160: 1.5 GB RAM per core
      - Intel 5520: 3 GB RAM per core
        (12 job slots → 2 GB RAM per job slot)
      - AMD 6168: 3 GB RAM (IO cache)
    - 30 GB local disk scratch space per job slot
      - At least 1 disk drive per 8 job slots

HS06 Scores, Batch Throughput, and More

- What is the performance for realistic applications such as HEP experiments codes? Does it scale with the number of cores?
- To check for possible bottlenecks, e.g. access to local disks or network performance, we have compared
  - HS06 scores,
  - batch throughput,
  - Ganglia monitoring plots,
  - *ps* and *top* output.

General Remarks on CPU Benchmarking

- Scoring of hardware
- Benchmark result should scale with real life applications
- Performance of an application depends on a lot of facts:
  - CPU
    - Clock cycle
    - Architecture
    - Cache size (L2, L3)
  - Memory throughput
  - File access
    - Local disk(s)
    - Remote fileserver(s)
  - Network performance
  - ...
- Application A1 may run faster on machine M1 while A2 is faster on M2

General Remarks on CPU Benchmarking

- HEP benchmarking:
  - HS06 is based on industry standard benchmark suite SPEC[1] CPU2006 ...
    - CPU2006: 12 integer and 17 floating-point applications
  - ... plus benchmarking HowTo provided by HEPiX Benchmarking WG[2]
    - All_cpp subset of CPU2006:
      3 integer and 4 floating-point applications
    - Operating system: the same one which is used at a site
    - Compiler: GNU Compiler Collection (GCC) 4.x
    - Flags (provided by LCG Architects Forum – mandatory!):
      `-O2 -pthread -fPIC -m32`
    - 1 simultaneous benchmark run per core
    - HS06 score of the system is the sum of the geometric means of the 7 individual runs per core

[1] SPEC is a registered trademark of the Standard Performance Evaluation Corporation

[2] Michele Michelotto, Manfred Alef, Alejandro Iribarren, Helge Meinhard, Peter Wegner, Martin Bly, Gabriele Benelli, Franco Brasolin, Hubert Degaudenzi, Alessandro De Salvo, Ian Gable, Andreas Hirstius, Peter Hristov:
A Comparison of HEP code with SPEC benchmarks on multi-core worker nodes. CHEP 2009, Journal of Physics 219 (2010)

HS06 Benchmarking

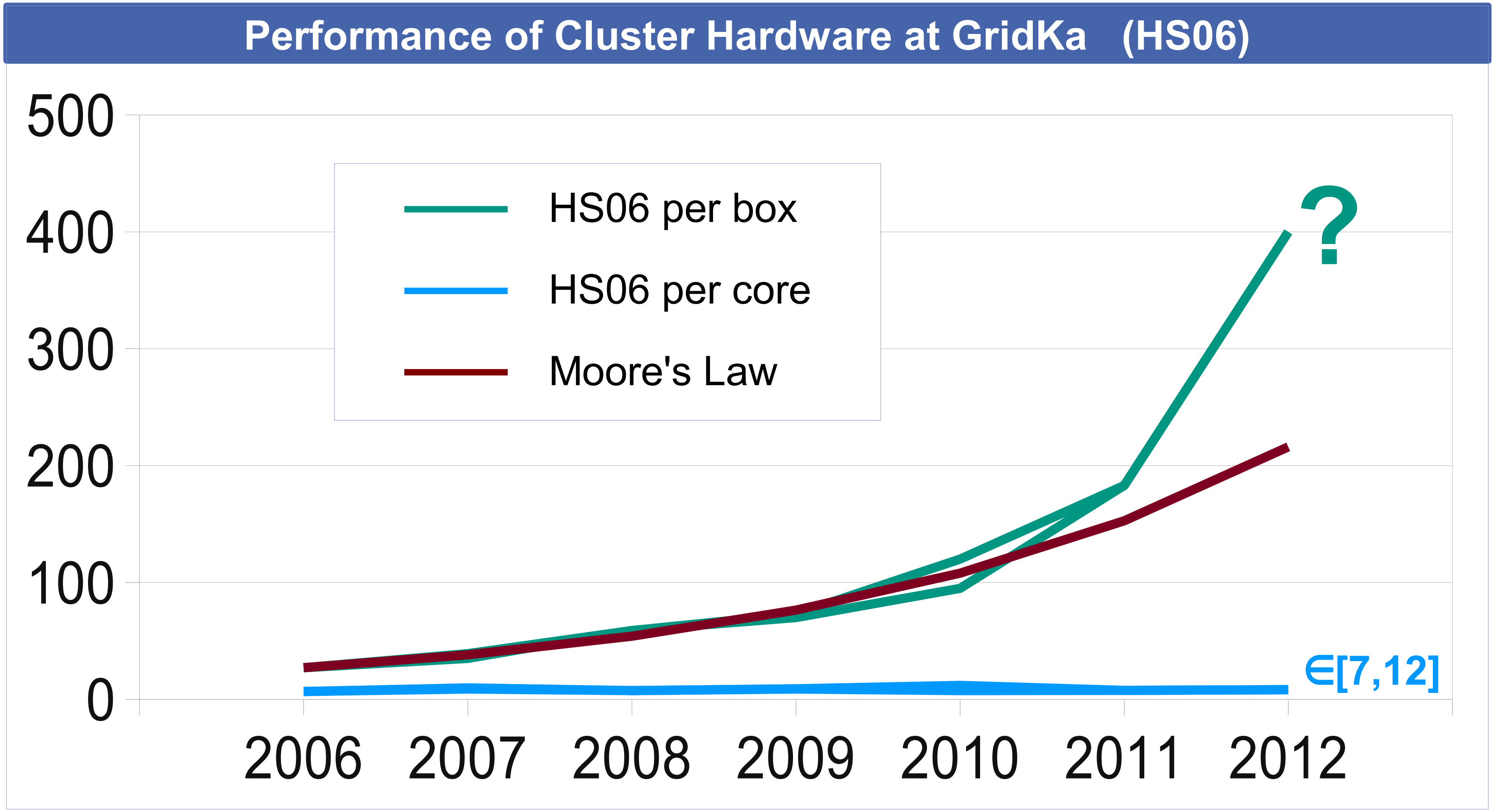
- Benchmark results demonstrate significant speed-up of modern cluster hardware.

- Example –
  Compute fabric at GridKa

HS06 Benchmarking

Vendor	CPU	MHz	Cores	Sockets	Runs	In Commission	HS06
AMD	270	2000	2	2	**4**	2006 ... 2010	**27**
Intel	5148	2333	2	2	**4**	2007 ... 2011	**35**
Intel	5160	3000	2	2	**4**	2007 ...	**39**
Intel	5345	2333	4	2	**8**	2008 ...	**59**
Intel	5420	2500	4	2	**8**	2009 ...	**70**
Intel	5430	2666	4	2	**8**	2009 ...	**73**
Intel	5520	2266	4 HT off 4 HT on	2	**8** **16**	2010 ...	**95** **120**
AMD	6168	1900	12	2	**24**	2011 ...	**183**
AMD	6174	2200	12	4	**48**	2011 ...	**400**

HS06 Benchmarking



HS06 Benchmarking

Vendor	CPU	MHz	Cores	Sockets	Runs	In Commission	HS06
AMD	270	2000	2	2	**4**	2006 ... 2010	**27**
Intel	5148	2333	2	2	**4**	2007 ... 2011	**35**
Intel	5160	3000	2	2	**4**	2007	**39**
Intel	5345	2333	4	[illegible]	[illegible]	[illegible]	**59**
Intel	5420	2500	4	2	**8**	2009 ...	**70**
Intel	5430	2666	4	2	**8**	2009 ...	**73**
Intel	5520	2266	4 HT off 4 HT on	2	**8** **16**	2010 ...	**95** **120**
AMD	6168	1900	12	2	**24**	2011 ...	**183**
AMD	6174	2200	12	4	**48**	2011 ...	**400**

Performance issues (insufficient memory bandwith)!

HS06 Scores versus Job Throughput

Karlsruher Institut für Technologie

- How does the number of jobs (per time interval) scale with the HS06 score?
  - Note that the number of jobs running on a particular system is a rough indicator of the performance because some jobs check for the remaining wallclock time and fill up the time slot provided by the batch queue.
  - There are currently no scaling factors configured in the batch system at GridKa.
  - Therefore the jobs-per-HS06 scores may vary similar to the HS06-per-job-slot performance of the host.
- Analysis of PBS accounting records from 16 to 18 April 2011
  - Data processed using Excel sheets

HS06 Scores versus Job Throughput

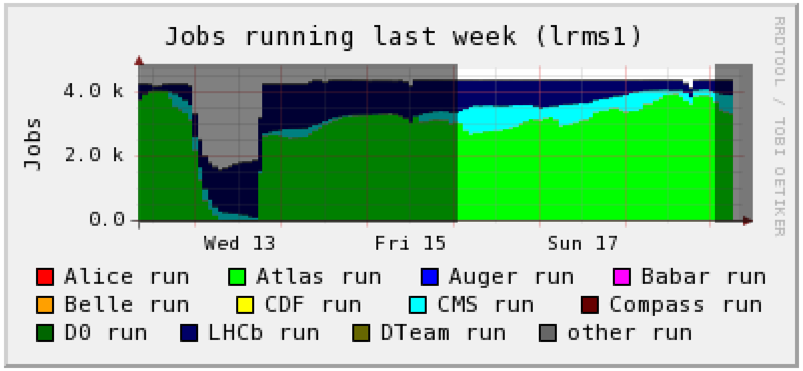
- GridKa WNs are split into 2 PBS sub-clusters
  - Heterogenous hardware in both clusters
  - Restricted VO access in sub-cluster 1

Sub-Cluster	Worker Nodes	Quantity	VOs
1	Intel 5160 Intel 5430 AMD 6168	37 nodes 181 nodes 116 nodes	Atlas, Auger, Belle, CMS, LHCb
2	Intel 5345 Intel 5420 Intel 5430 Intel 5520 HT off Intel 5520 HT on AMD 6174	338 nodes 350 nodes 33 nodes 1 node 218 nodes 1 node	All VOs

HS06 Scores versus Job Throughput

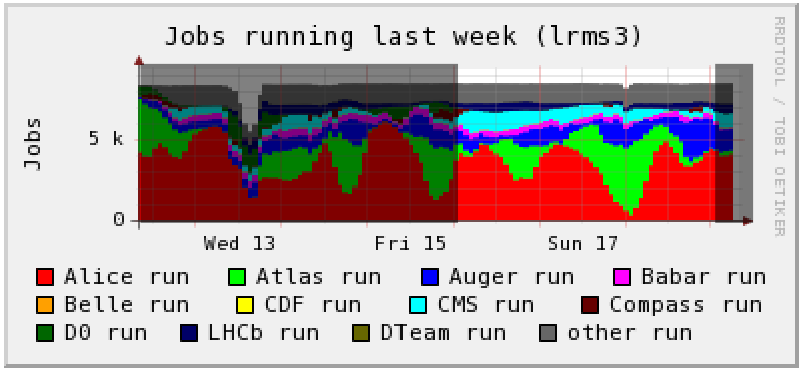
Analysis of Batch Accounting Files

Sub-cluster 1



VOs: Atlas, Auger, Belle, CMS, LHCb

Sub-cluster 2



All VOs

🟥 Alice 🟩 Atlas 🟦 Auger 🟪 BaBar

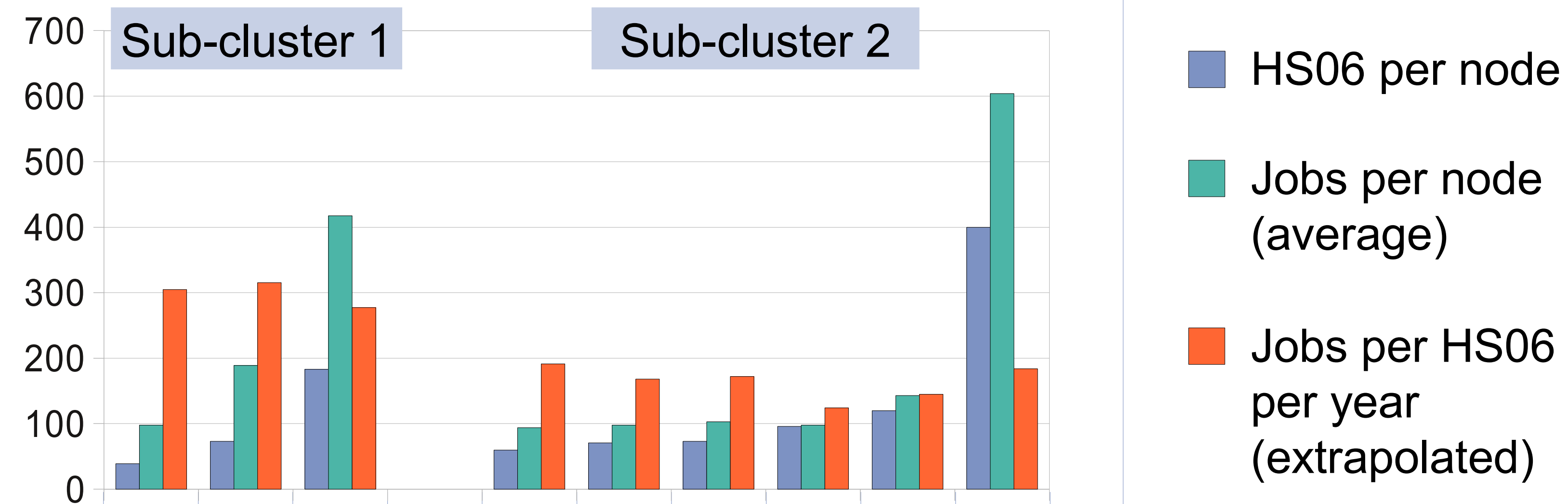
■ CMS 🟫 Compass 🟩 D0 🟦 LHCb

⬛ other

Period investigated: April 16-18, 2011

HS06 Scores versus Job Throughput

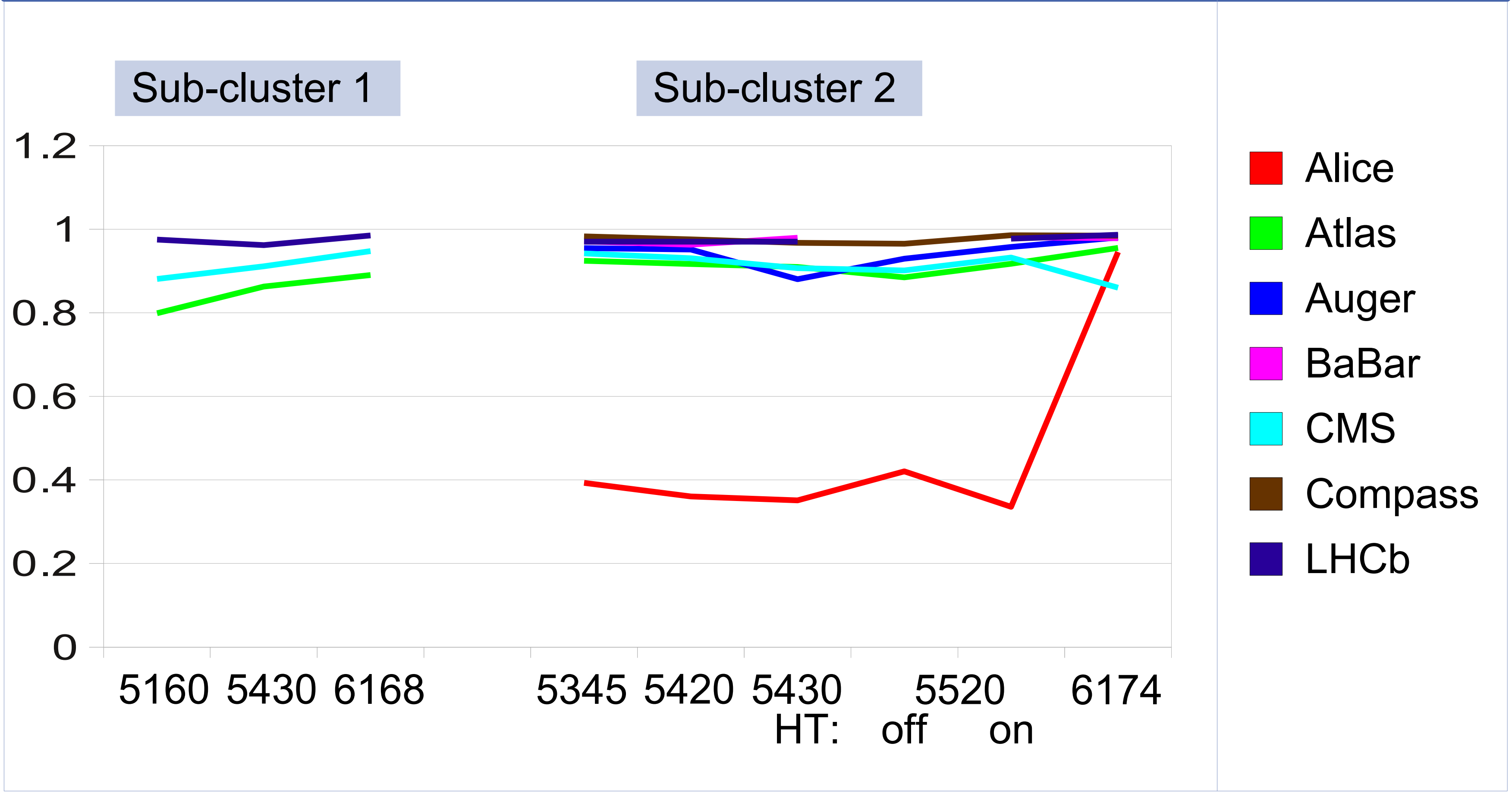
HS06 Score versus Job Count



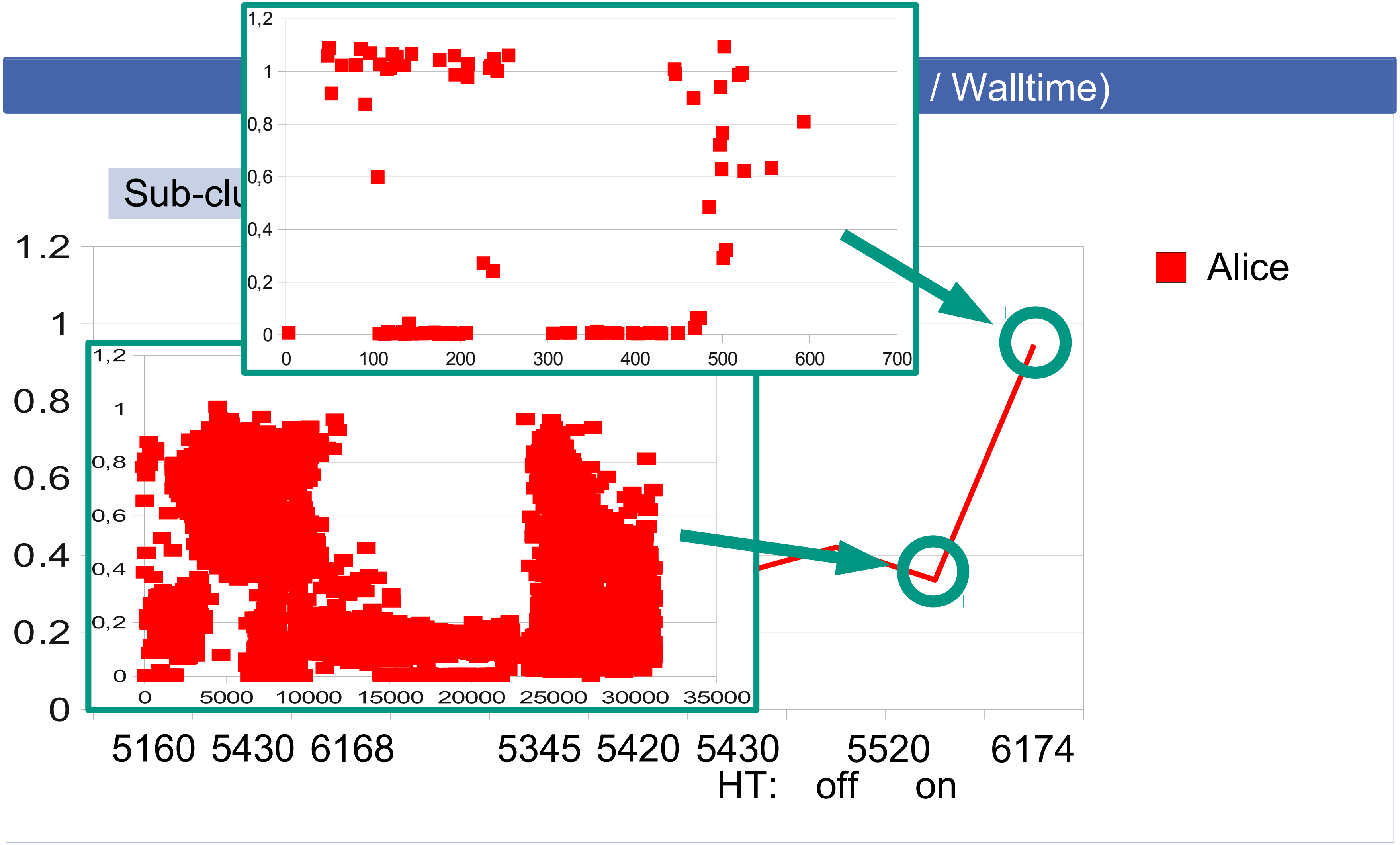
CPU	5160	5430	6168		5345	5420	5430	5520		6174
HT	—	—	—		—	—	—	off	on	—
Cores	4	8	24		8	8	8	8	8	48
Slots	4	8	24		8	8	8	8	12	48

HS06 Scores versus Job Throughput

Job Efficiency (CPU Consumption / Walltime)



HS06 Scores versus Job Throughput



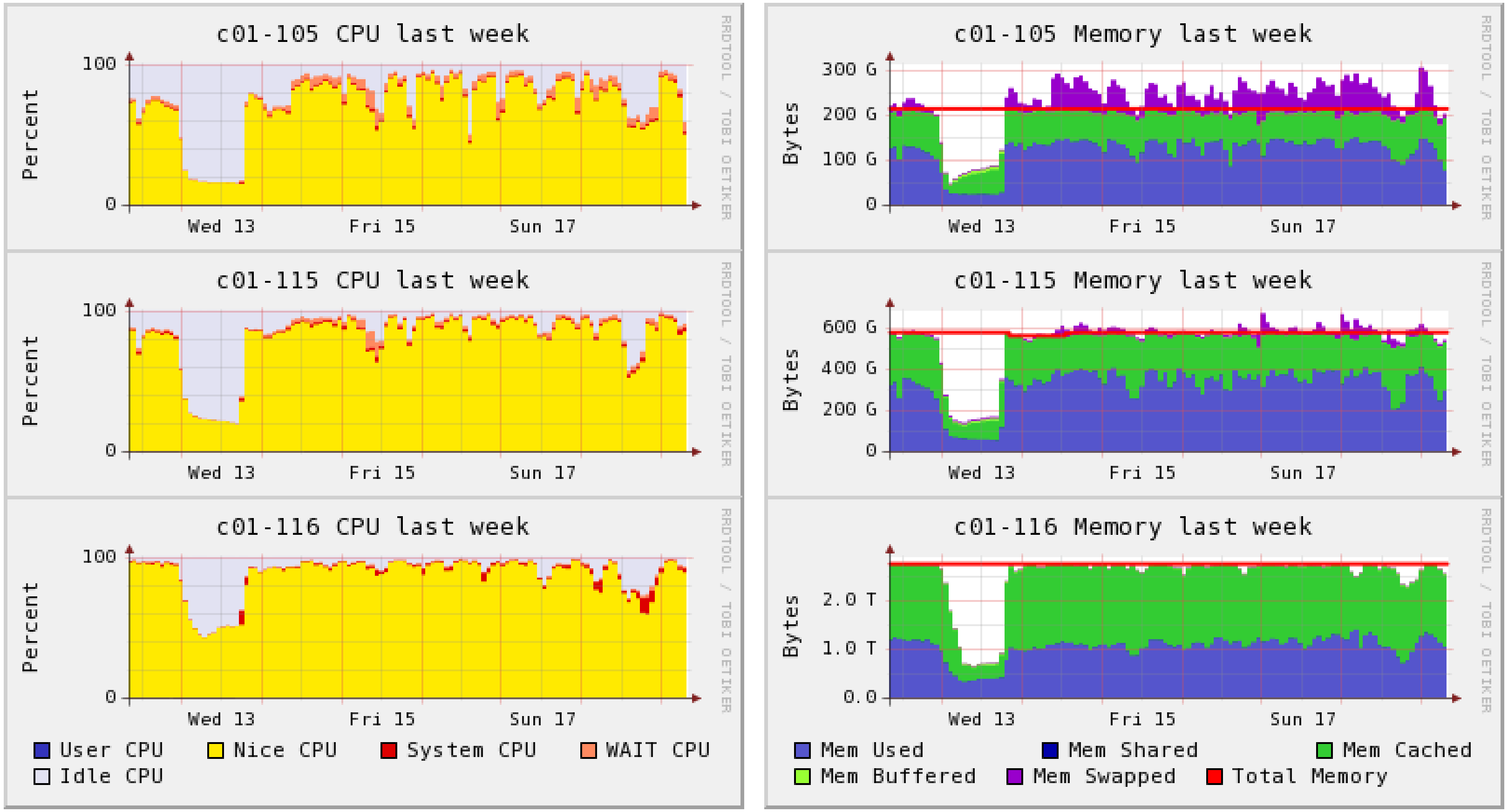
Ganglia and Local Performance Monitoring

Ganglia Performance Plots: Sub-cluster 1

5160 (#4)

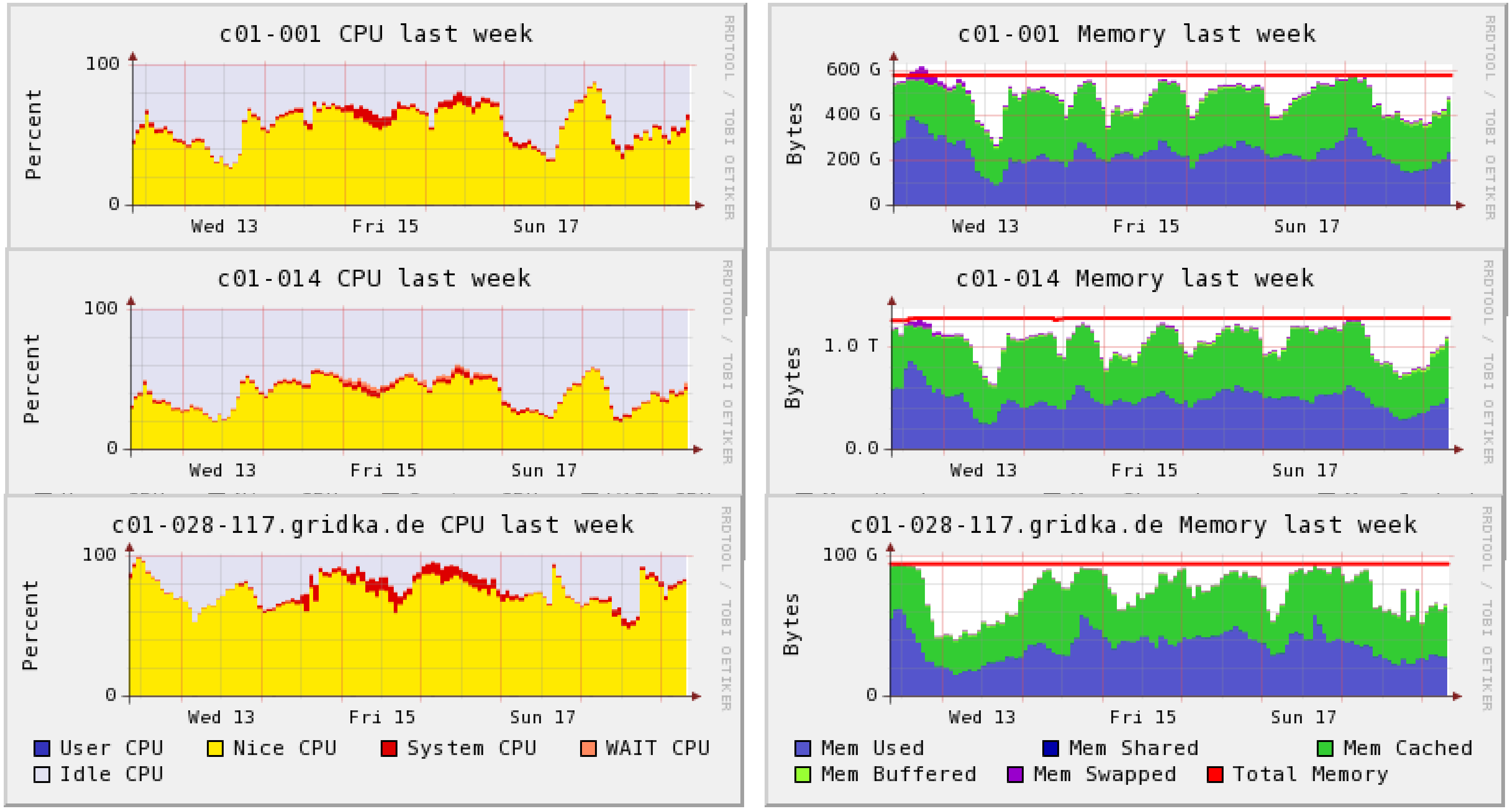
5430 (#8)

6168 (#24)



Ganglia and Local Performance Monitoring

Ganglia Performance Plots: Sub-cluster 2



5345 (#8)

5520 (#12)

6174 (#48)

Ganglia and Local Performance Monitoring

Local Performance Monitoring: 'top' and 'ps' Output

Most time-consuming processes running on the 48-core node (AMD 6174)

```
[alef@c01-028-117 ~]$ uptime ; ps -uroot | sort -k3 -r | head
 14:04:13 up 34 days, 22:09,  2 users,  load average: 43.54, 43.44, 43.33
  PID TTY          TIME CMD
 6894 ?        03:30:31 kjournald
10171 ?        01:36:30 pbs_mom
14208 ?        00:19:00 pdflush
10993 ?        00:14:22 pdflush
 8132 ?        00:07:54 rpciod/47
 5428 ?        00:07:16 nfsiod
 8560 ?        00:05:31 snmpd
 8131 ?        00:05:24 rpciod/46
 8130 ?        00:04:39 rpciod/45
[alef@c01-028-117 ~]$
```

Conclusions

- New batch workers are coming with more and more CPU cores.
- The performance level per core has been frozen at around 10 HS06.
- Boxes with up to 4x12=48 cores are on the market.

- Performance investigations have not found any real show-stoppers:
  - HS06 scores scale well with the number of CPU cores per system.
  - Number of jobs started on particular nodes scale with HS06 performance.
  - Performance monitoring tools, like Ganglia plots or local system commands, don't show serious bottlenecks.