Karlsruhe Institute of Technology

Comparison between MonALISA (Alice) and PBS Accounting at GridKa

Manfred Alef

Steinbuch Centre for Computing

Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft

Universität Karlsruhe (TH)
Research University · founded 1825

Starting Point

- Talk by Yves Schutz ”ALICE Quarterly Report 2008Q1“ about accounting issues (WLCG-MB, 2008-05-27):

Accounting

Accounting Tier 1. Sources: WLCG monthly report and ALICE MonALisa report, April 2008

CPU

Tier 1	WLCG T1 accounting			ALICE MonALIsa			2008 C-RRB Pledges	
	Total Pledged	Delivered to ALICE (wall)	Fraction total	Pledged	Delivered Wall	Fraction	All	ALICE
	KSI2K	KSI2K	%	KSI2K	KSI2K	%	KSI2K	KSI2K
CERN Tier-0+CAF	8’083	1'668	46%	1602	801*	50%	15'851	2'300
CCIN2P3	1'733	506	52%	1060	387	37%	5'740	1'060
CNAF	1'475	49	41%	660	24	4%	3'000	660
FZK-GRIDKA	2'160	674	45%	600	365	61%	5’672	2'500
NDGF	906	149	41%	602	187	31%	2'172	1'102
NL LHC/Tier-1	2'014	41	13%	475	145	31%	4'382	317
RAL	1'505	25	41%	132	51	39%	3'139	132

* Does not include CAF

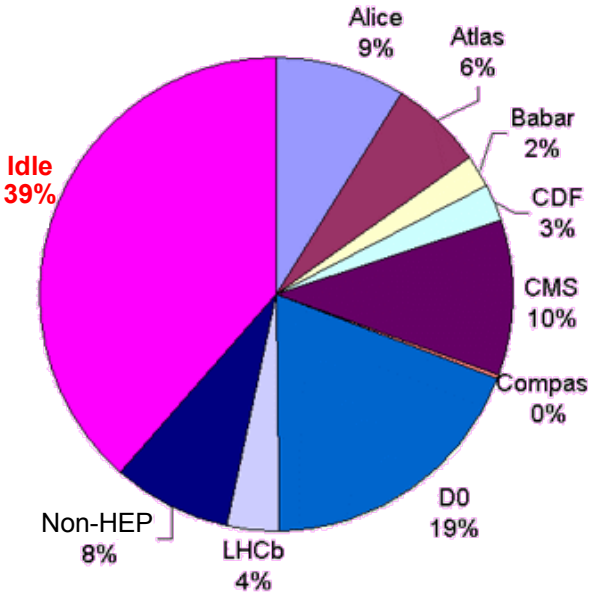
ALICE 1Q2008 report 18

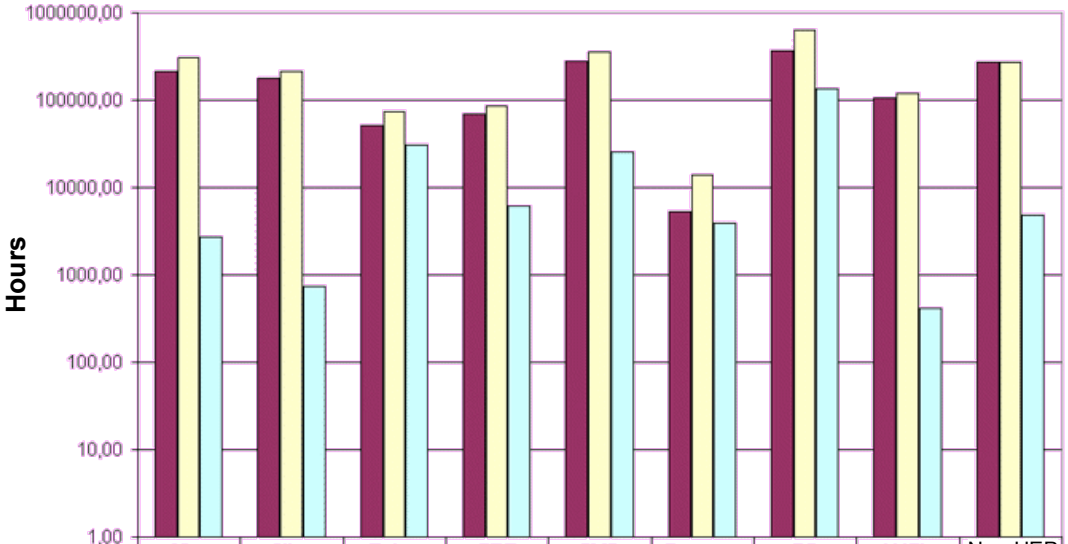
Starting Point

- Abstract (slides 17+18):
  ”In April GridKa has delivered only 365 kSI2k (average walltime) – that's much less than pledged.“
- Substantial differences between ”delivered“ and ”pledged“ resources were reported for other sites, too.
- Investigations at GridKa about the reasons.

Starting Point

- General remarks about ”delivered versus consumed“ CPU time:
  - The GridKa CPU resources have been increased by a factor of about 3 to the milestone 1st April 2008.
  - Usage by the experiments in April was much less than 100%. Always some worker nodes (job slots) have been idle.
  - Therefore Alice has consumed only 365 kSI2k walltime!





	Alice	Atlas	Babar	CDF	CMS	Compass	D0	LHCb	Non-HEP
CPU-Time	212418,43	178424,13	51856,87	69863,02	276036,27	5276,53	364883,94	105061,45	271757,34
Wall-Time	303233,74	215488,88	73425,17	86872,34	351691,60	13780,22	640737,06	120926,80	274523,06
Wait-Time	2724,37	733,27	30308,23	6135,35	25422,04	3900,40	137438,98	417,57	4820,95

jobs: 39040

Investigations at GridKa

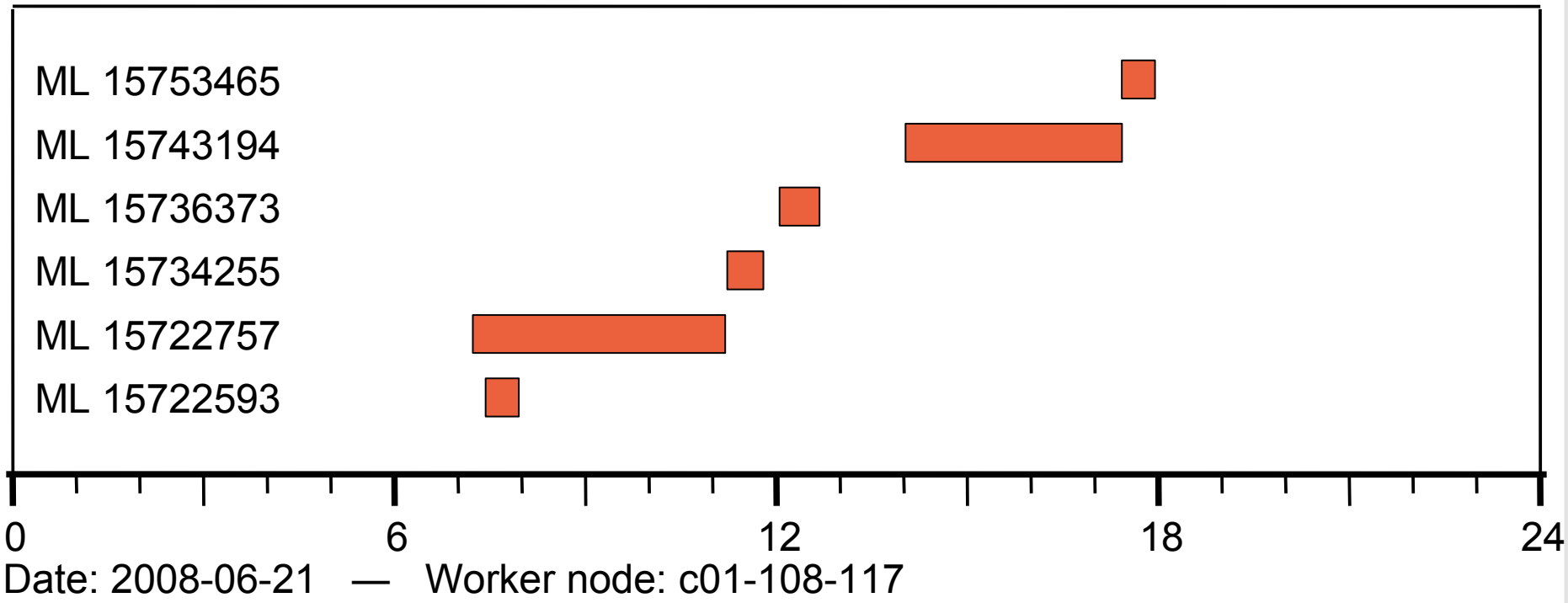
- MonALISA accounting by Alice:
  - There is no local MonALISA logging (at site level) enabled per default.
    - MonALISA data of a particular job are only available at run time (http://pcalimonitor.cern.ch).
  - In order to check the logfiles and compare the MonALISA data with our PBS logs, local MonALISA logging (on a per-job basis) has been enabled on the VObox alice-fzk.gridka.de.
    - *Many thanks to Kilian Schwarz (GSI) and Costin Grigoras (CERN) for switching it on!*
- GridKa accounting:
  - Based on the log files of the local batch system (PBS Professional).

Investigations at GridKa

- Comparison based on all jobs which have finished on Saturday, 21st June 2008.
- All jobs which were executed on the sample worker node c01-108-117 have been checked in more detail.

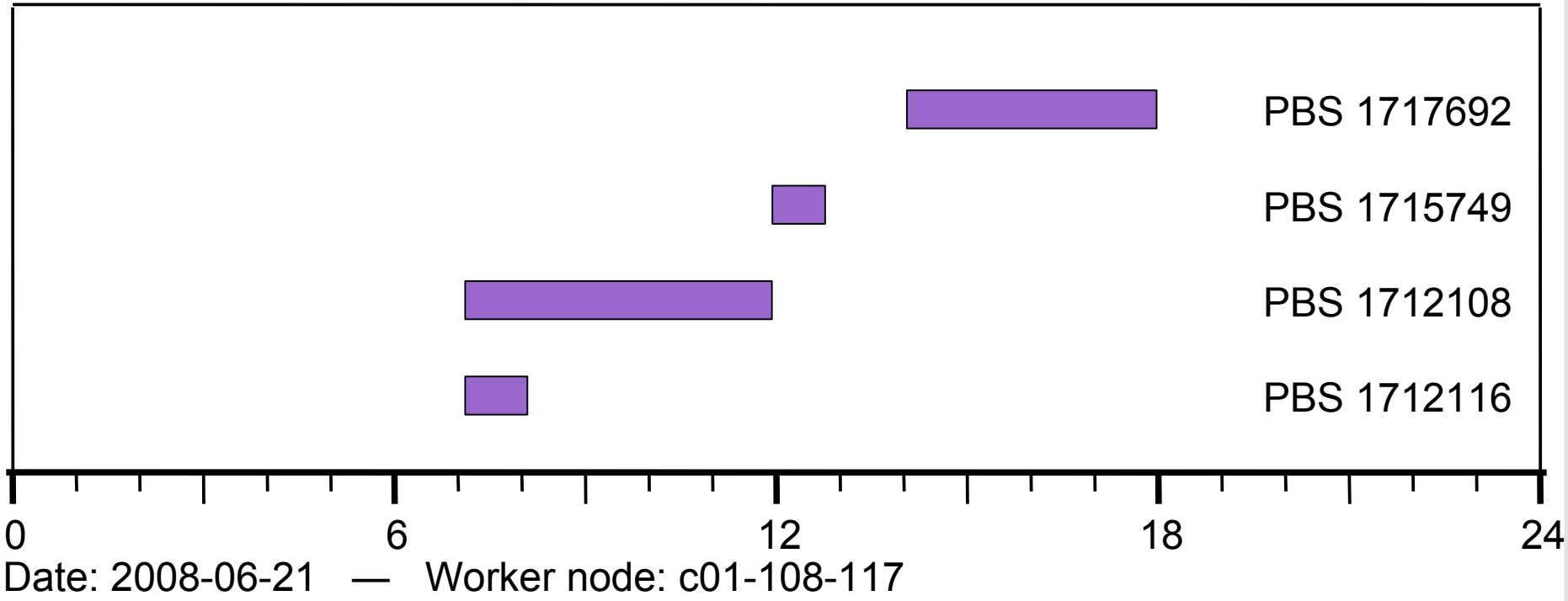
Investigations at GridKa

- Checking MonALISA and PBS files – **pilot jobs**:
  - Matches found in the log files:
    - MonALISA: 6 jobs (of 7278)



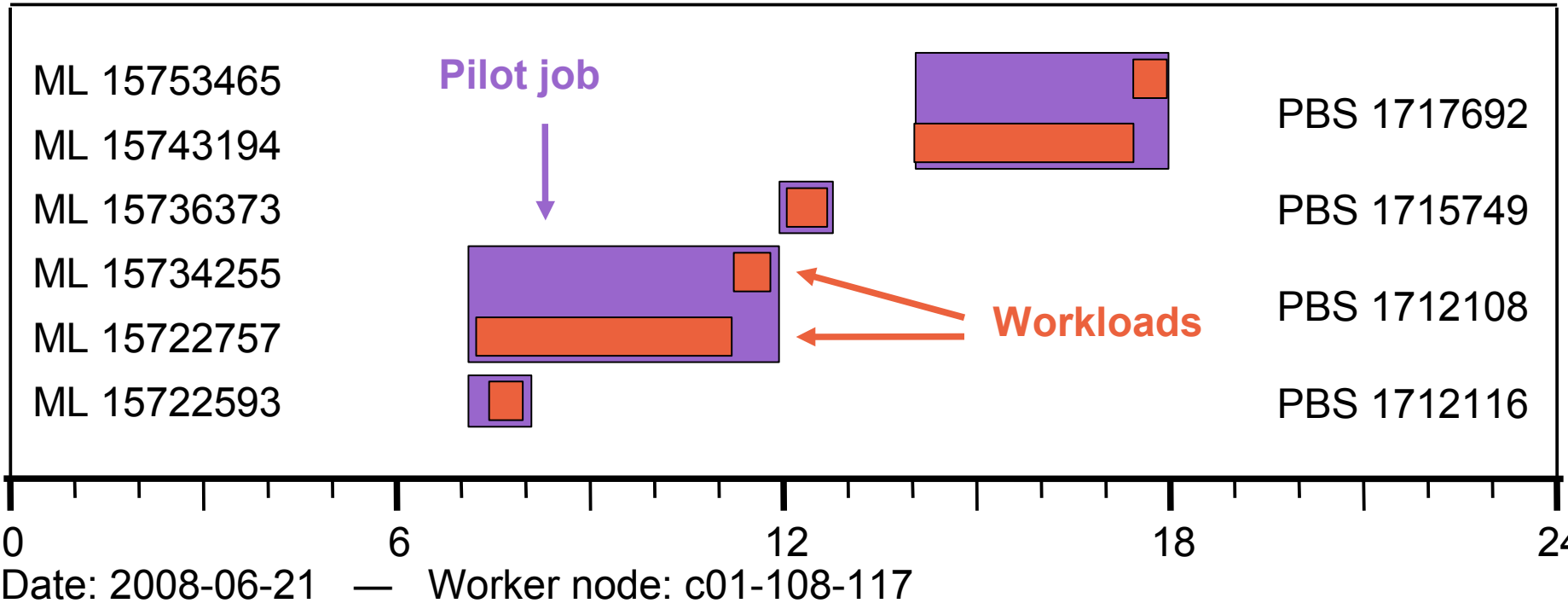
Investigations at GridKa

- Checking MonALISA and PBS files – **pilot jobs**:
  - Matches found in the log files:
    - PBS: 4 jobs (of 4339)



Investigations at GridKa

- Checking MonALISA and PBS files – **pilot jobs**:
  - Matches found in the log files:
    - MonALISA: 6 jobs (of 7278)
    - PBS: 4 jobs (of 4339)



Investigations at GridKa

- Checking MonALISA and PBS files **– pilot jobs:**
  - **Alice jobs are pilots. In contrast to PBS, only the real workload is considered in the MonALISA accounting.**

Investigations at GridKa

- Checking MonALISA files (example 1) – **walltime**:

MonALISA job ID	Timestamp (ms)		
15722757	1214025271790	host	c01-108-117.gridka.de
15722757	1214025292187	cpu_ksi2k	0.0
15722757	1214025292187	cpu_time	0.0
15722757	1214025292187	run_ksi2k	116.028
15722757	1214025292187	run_time	66.0
15722757	1214025352989	cpu_ksi2k	0.0
...			
15722757	1214038739669	cpu_ksi2k	19181.538000000008
15722757	1214038739669	cpu_time	10911.0
15722757	1214038739669	run_ksi2k	23755.854000000018
15722757	1214038739669	run_time	13513.0
15722757	1214038739670	host	c01-108-117.gridka.de
15722757	1214039526563	host	c01-108-117.gridka.de

cpu_time:	CPU time consumed so far
cpu_ksi2k:	cpu_time * ksi2k of the node
run_time:	wall time consumed so far
run_ksi2k:	run_time * ksi2k of the node

The last MonALISA walltime entry is 13513.0 s. However, the time interval from the first to the last timestamp is 14254.8 s!

Investigations at GridKa

- Checking MonALISA files (example 1) – **walltime:**
  - **MonALISA accounting conceals some overhead within the Alice workload tasks!**

Investigations at GridKa

- Checking MonALISA files (example 2) – **kSI2k factors**:

MonALISA job ID	Timestamp (ms)		
15722757	1214025271790	host	c01-108-117.gridka.de
15722757	1214025292187	cpu_ksi2k	0.0
15722757	1214025292187	cpu_time	0.0
15722757	1214025292187	run_ksi2k	116.028
15722757	1214025292187	run_time	66.0
15722757	1214025352989	cpu_ksi2k	0.0
...			
15722757	1214038739669	cpu_ksi2k	19181.538000000008
15722757	1214038739669	cpu_time	10911.0
15722757	1214038739669	run_ksi2k	23755.854000000018
15722757	1214038739669	run_time	13513.0
15722757	1214038739670	host	c01-108-117.gridka.de
15722757	1214039526563	host	c01-108-117.gridka.de

cpu_time:	CPU time consumed so far
cpu_ksi2k:	cpu_time * ksi2k of the node
run_time:	wall time consumed so far
run_ksi2k:	run_time * ksi2k of the node

The ksi2k factor used by MonALISA is:
cpu_ksi2k : cpu_time

Investigations at GridKa

- Checking MonALISA files (example 2) – **kSI2k factors:**
  - **The kSI2k factors used in MonALISA differ from the official numbers used at GridKa*!**

CPU type	kSI2k used by ... GridKa	MonALISA
AMD Opteron 246 (2.0 GHz SC)	1.016	1.289
AMD Opteron 270 (2.0 GHz DC)	1.002	1.452
Intel Xeon 5148 (2.33 GHz DC)	1.676	1.454
Intel Xeon 5160 (3.00 GHz DC)	2.071	1.627 ... 2.535
Intel Xeon E5345 (2.33 GHz QC)	1.641	1.336 ... 1.626
Intel Xeon E5430 (2.66 GHz QC)	2.044	1.479 ... 2.629
Average (arithmetic means over all Alice jobs)	1.726	1.578

* SPECint_base2000 ”L“ numbers (i386) from http://hepix.caspur.it/processors, multiplied with 1.5, divided by the number of cores (= number of job slots)

Investigations at GridKa

- Checking MonALISA files (example 3) – **idle workloads**:

MonALISA job ID	Timestamp (ms)		
15722593	1214026045451	host	c01-108-117.gridka.de
15722593	1214026065732	cpu_time	0.0
15722593	1214026065732	run_time	200.0
15722593	1214026126345	cpu_time	0.0
...			
15722593	1214026490048	run_time	625.0
15722593	1214026550668	cpu_time	0.0
15722593	1214026550668	run_time	685.0
15722593	1214026604990	host	c01-108-117.gridka.de
15722593	1214026611273	cpu_time	0.0
15722593	1214026611273	run_time	746.0
15722593	1214026627447	host	c01-108-117.gridka.de
15722593	1214026641589	host	c01-108-117.gridka.de
15722593	1214027925188	host	c01-108-117.gridka.de

`cpu_time`: CPU time consumed so far
`run_time`: wall time consumed so far

No run_ksi2k, no cpu_ksi2k entries! -- Average total run_time of idle workload tasks is about 20 mins, walltime about 30 mins (arithmetic means).

Investigations at GridKa

- Checking MonALISA files (example 3) – **idle workloads:**
  - **Some workload tasks don't consume CPU time at all. Left unconsidered in MonALISA?**
  - *(If you assume that these jobs are failing due to system issues, please open GGUS ticket.)*

Investigations at GridKa

- Checking MonALISA files (example 4) – **no kSI2k conversion**:

MonALISA job ID	Timestamp (ms)		
15705551	1213991981100	host	c01-005-131.gridka.de
15705551	1213992165939	host	c01-005-131.gridka.de
15705551	1213992169551	host	c01-005-131.gridka.de
15705551	1213992243945	cpu_time	65.0
15705551	1213992243945	run_time	263.0
...			
15705551	1214006397379	cpu_time	13529.0
15705551	1214006397379	run_time	14417.0
15705551	1214006397381	host	c01-005-131.gridka.de
15705551	1214006817119	host	c01-005-131.gridka.de

`cpu_time`: CPU time consumed so far
`run_time`: wall time consumed so far

No run_ksi2k, no cpu_ksi2k entries! -- Nevertheless the job was run and had consumed CPU time.

Investigations at GridKa

- Checking MonALISA files (example 4) – **no kSI2k conversion:**
  - **CPU consumption of these jobs probably left unconsidered in MonALISA (~ 15%)!**

Investigations at GridKa

- Results (based on run_time and walltime of all jobs of this day):
  - Sum of highest run_ksi2k entries from MonALISA log files (column Q in Excel sheet ”MonALISA 20080621“): 25679 h * kSI2k
  - Sum of walltime (computed from time stamps in MonALISA files, multiplied with the right kSI2k factor – column R in that Excel sheet): 38944 h * kSI2k
  - Sum of PBS accounting records (column K in Excel sheet ”PBS 20080621“): 50217 h * kSI2k

Summary

- Several reasons for differences between accounting data computed by Alice (MonALISA) and GridKa:
  - The GridKa accounting measures the whole time interval while the worker node (job slot) is occupied by the (Alice) job.
  - Alice jobs are pilots. In contrast to PBS, only the real workload is considered in the MonALISA accounting.
  - The walltime of the workload tasks computed in MonALISA is less than the real walltime (difference between first and last timestamp in logfile).
  - Wrong kSI2k factors are used.
  - Many idle workload tasks (0 seconds CPU usage) probably not accounted by MonALISA.
  - Some Alice jobs don't report their kSI2k usage, although they have consumed CPU time.

Summary

- Several reasons for differences between accounting data computed by Alice (MonALISA) and GridKa:
  - The GridKa accounting measures the whole time interval while the worker node (job slot) is occupied by the (Alice) job.
  - Alice jobs are pilots. In contrast to PBS, only the real workload is considered in the MonALISA accounting. **-20%**
  - The walltime of the workload tasks computed in MonALISA is less than the real walltime (difference between first and last timestamp in logfile). **-6%**
  - Wrong kSI2k factors are used. **-7%**
  - Many idle workload tasks (0 seconds CPU usage) probably not accounted by MonALISA. **-4%**
  - Some Alice jobs don't report their kSI2k usage, although they have consumed CPU time. **-15%**

Questions?