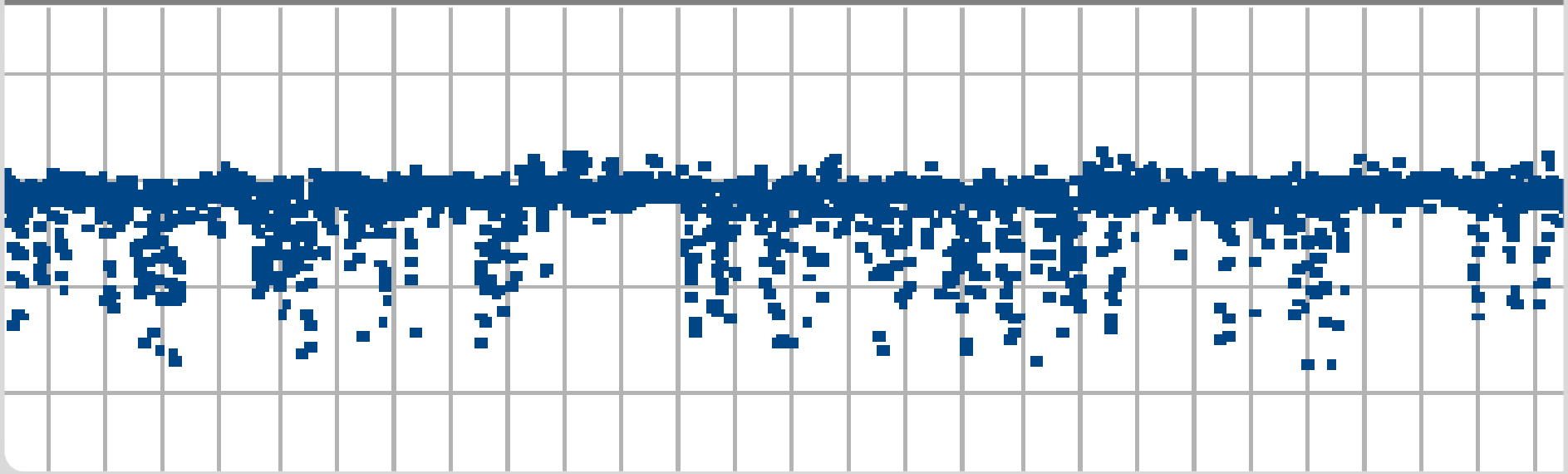
Karlsruhe Institute of Technology

Results of HS06 Scaling Studies at GridKa
WLCG GDB 2015-12-09

Manfred Alef

Steinbuch Centre for Computing (SCC)



Preliminary Remarks

- Performance and benchmarking session at GDB 2015-09-09
  (https://indico.cern.ch/event/319751/)

  - ➔ Philippe Charpentier has demonstrated consistent scaling of LHCb jobs with HS06 power of the provided job slot (via MJF) at GridKa
    (https://indico.cern.ch/event/319751/session/0/contribution/6/attachments/1153280/1656518/150909-MJFandBenchmarking-LHCb.pdf)

  - ➔ Corresponding performance studies at GridKa

Amazing News

Amazing News

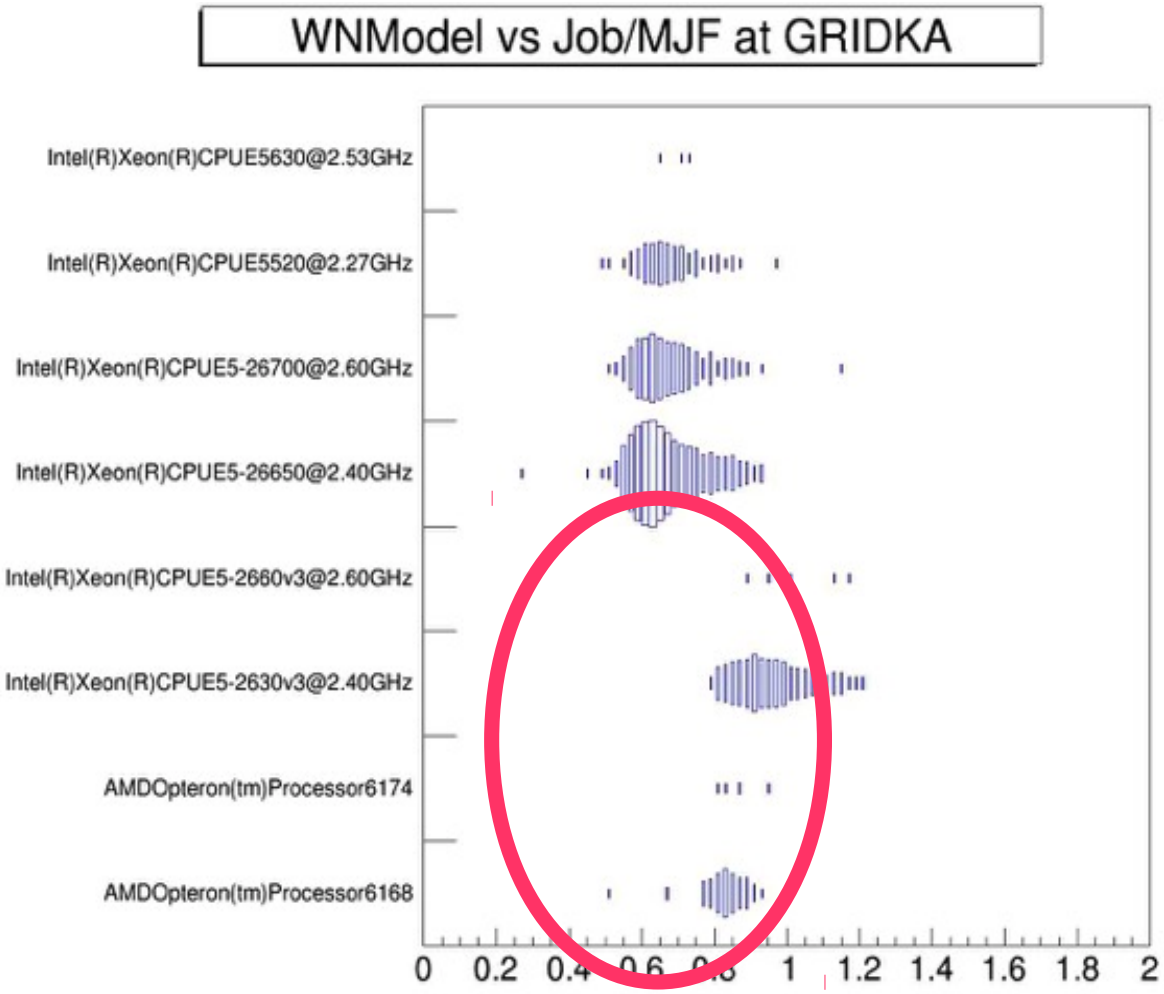
- Bad (?) news from Philippe Charpentier (2015-11-09):
  MJF power seems under-evaluated by 30 to 45%

  - ➔ Affected WN models:
    - AMD Opteron(tm) Processor 6168
    - AMD Opteron(tm) Processor 6174 *
    - Intel(R) Xeon(R) CPU E5-2630v3@2.40GHz
    - Intel(R) Xeon(R) CPU E5-2660v3@2.60GHz *
  - ➔ No differences in performance results on other WN models since September ('WNModel vs Job/MJF': still about 0.63)
  - ➔ No degradation in performance of any WN model at GridKa

* Only few hosts of each type in production at GridKa, therefore excluded

Amazing News

- Bad (?) news from Philippe Charpentier:



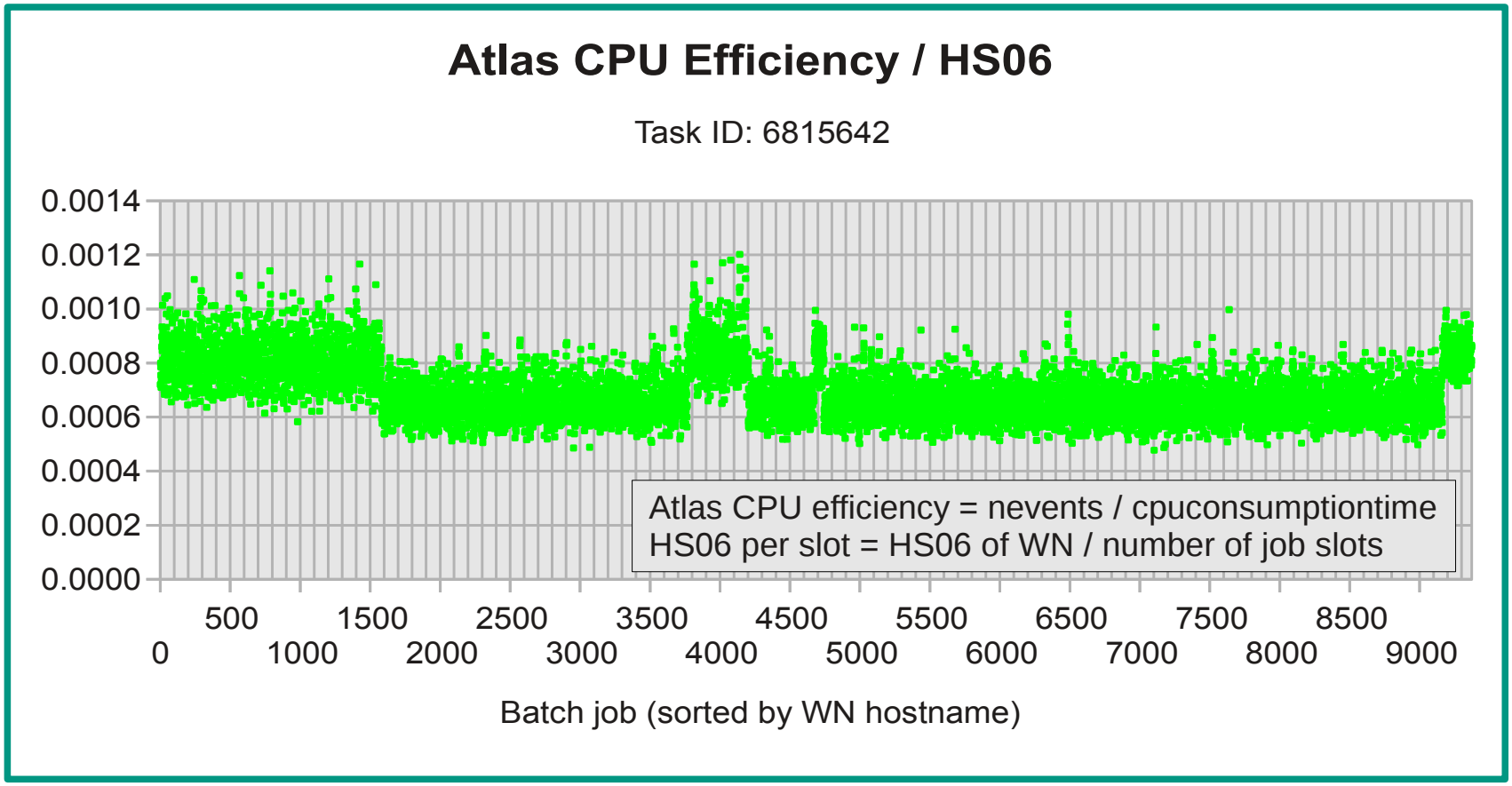
Investigations at GridKa

Investigations at GridKa

- What about jobs from other VOs, e.g. Atlas?
  - ➔ Accounting data (e.g. number of events, CPU time, wallclock time) of each Atlas job stored in Panda
  - ➔ The WN hostname of each job is also available so we can compare with HS06 performance
    - *The 'cpuconsumptionunit' is not a good metric:*
      WNs with the same CPU model name string can differ in several hardware details (memory speed, ...) as well as in individual WN configurations (number of job slots, ...)
  - ➔ Many thanks to Thomas Hartmann (KIT/Atlas) for looking up and downloading the raw job accounting datasets from Panda

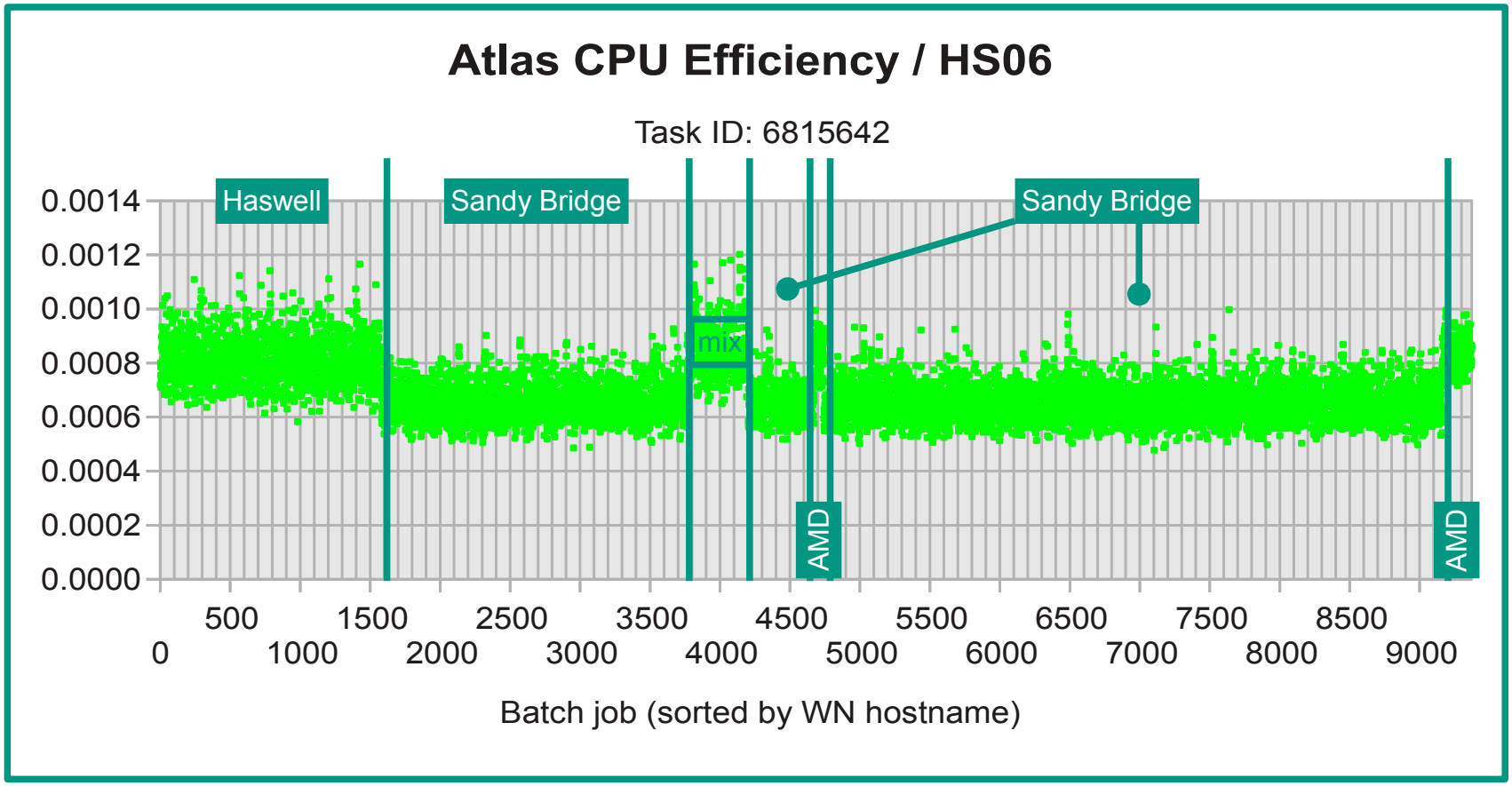
Investigations at GridKa

- What about jobs from other VOs, e.g. Atlas?



Investigations at GridKa

- What about jobs from other VOs, e.g. Atlas?



Investigations at GridKa

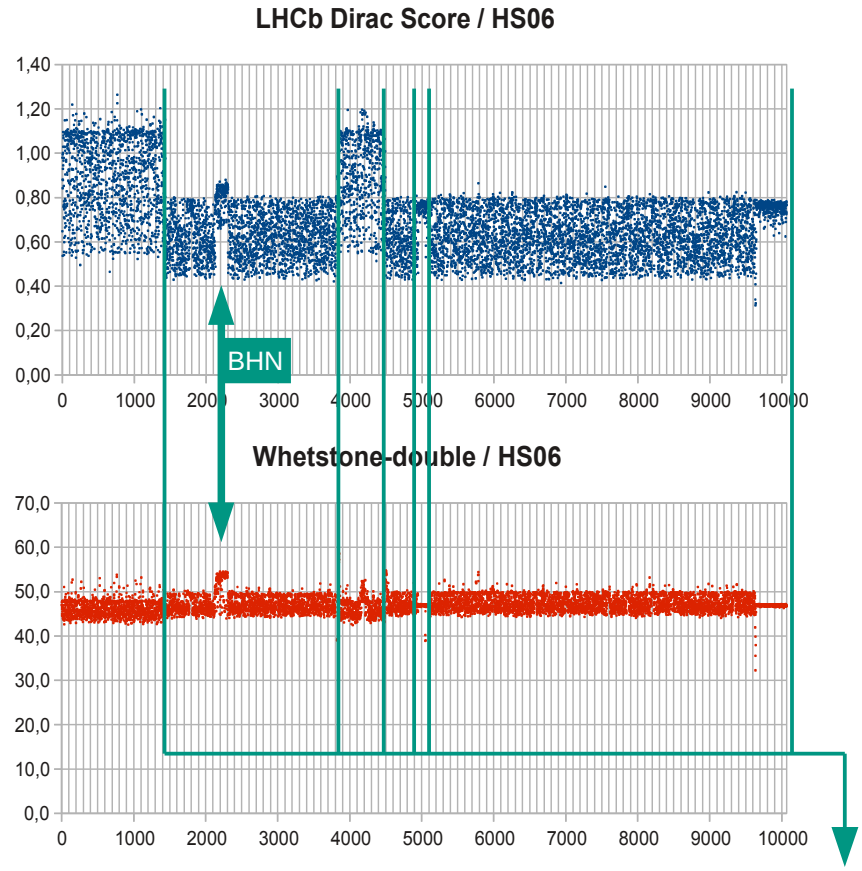
- What about the scaling with other applications, e.g. fast benchmarks?

  - From time to time:
    running series of batch jobs starting fast benchmarks
    - LHCb Dirac Python script (runtime: < 1 min)
    - Whetstone-double (runtime: 2...3 min)
      from UnixBench (https://github.com/kdlucas/byte-unixbench)
    - HS06 (querying MJF implementation at GridKa)

  - Number of jobs per study:
    - September 2015: ~ 4000 jobs (1 job per 30 seconds → 2 days)
      (WN model E5520 excluded because of inconsistent MJF setting)
    - November 2015: ~ 10000 jobs (1 job per minute → 1 week)

Investigations at GridKa

- Scaling of HS06 with fast benchmarks:



	Min	Avg.	Max
• Haswell:	0.47	0.89	1.26
• Sandy Bridge:	0.42	0.63	0.87
• AMD:	0.31	0.75	0.79

	Min	Avg.	Max
• Haswell:	42.5	46.2	53.8
• Sandy Bridge:	44.0	47.4	54.5
• AMD:	32.3	46.8	47.2

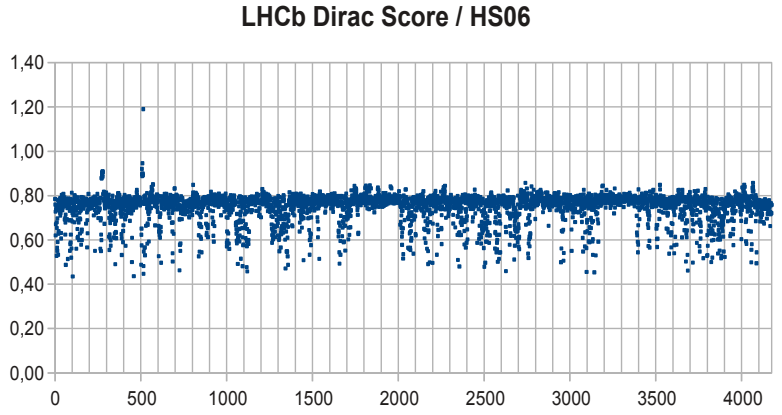
Investigations at GridKa

- And what about 2 months ago when Philippe Charpentier had prepared his GDB talk?

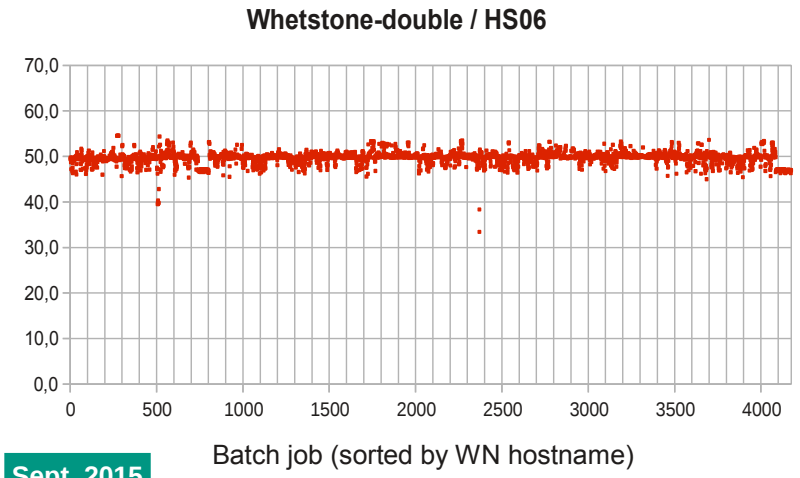
Investigations at GridKa

- Scaling of HS06 with fast benchmarks:

	Min	Avg.	Max
• Haswell (n.a.):			
• Sandy Bridge:	0.45	0.75	0.86
• AMD:	0.67	0.75	0.79



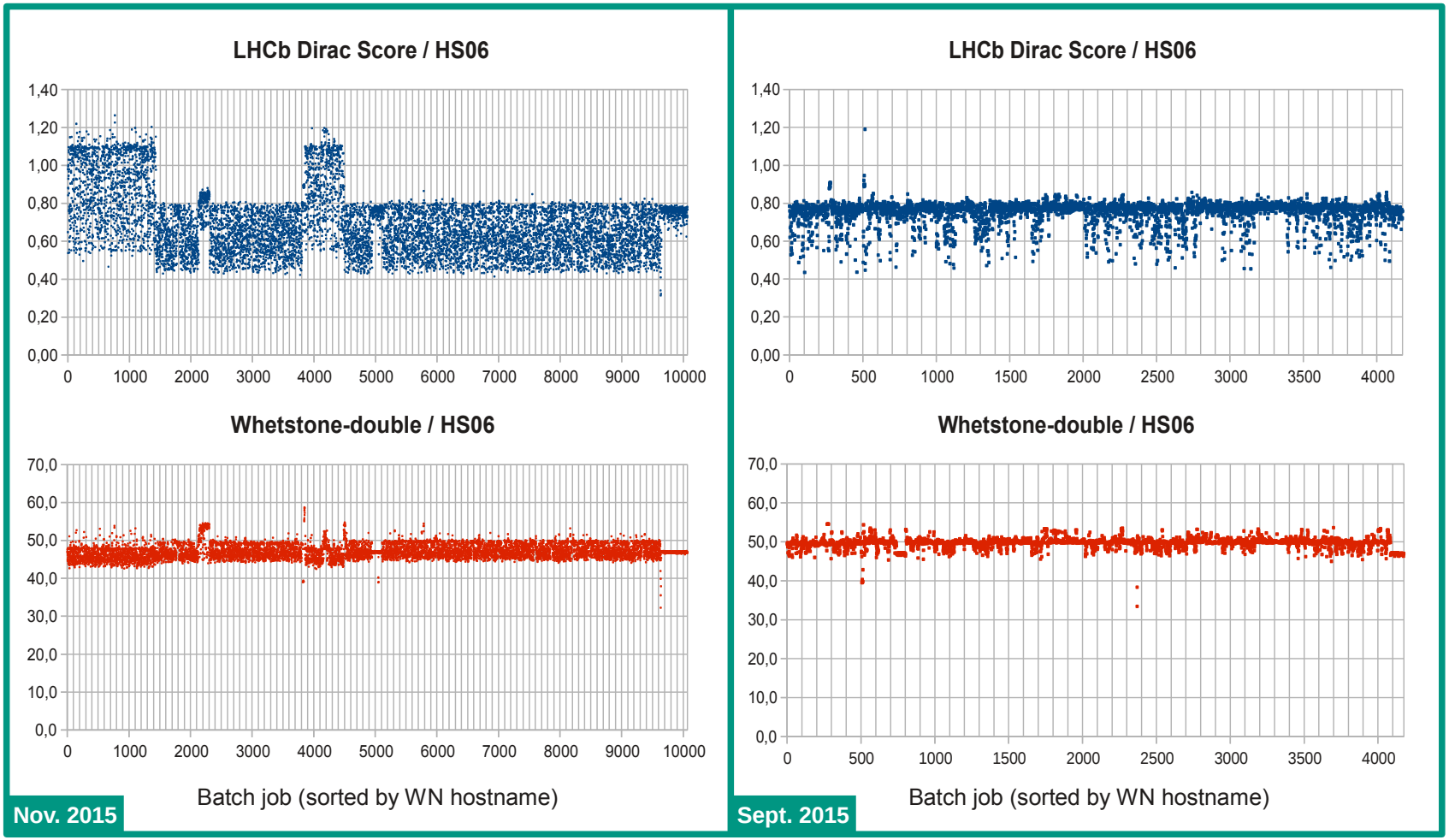
	Min	Avg.	Max
• Haswell (n.a.):			
• Sandy Bridge:	33.4	49.7	53,7
• AMD:	46.5	46.9	47.2



Sept. 2015

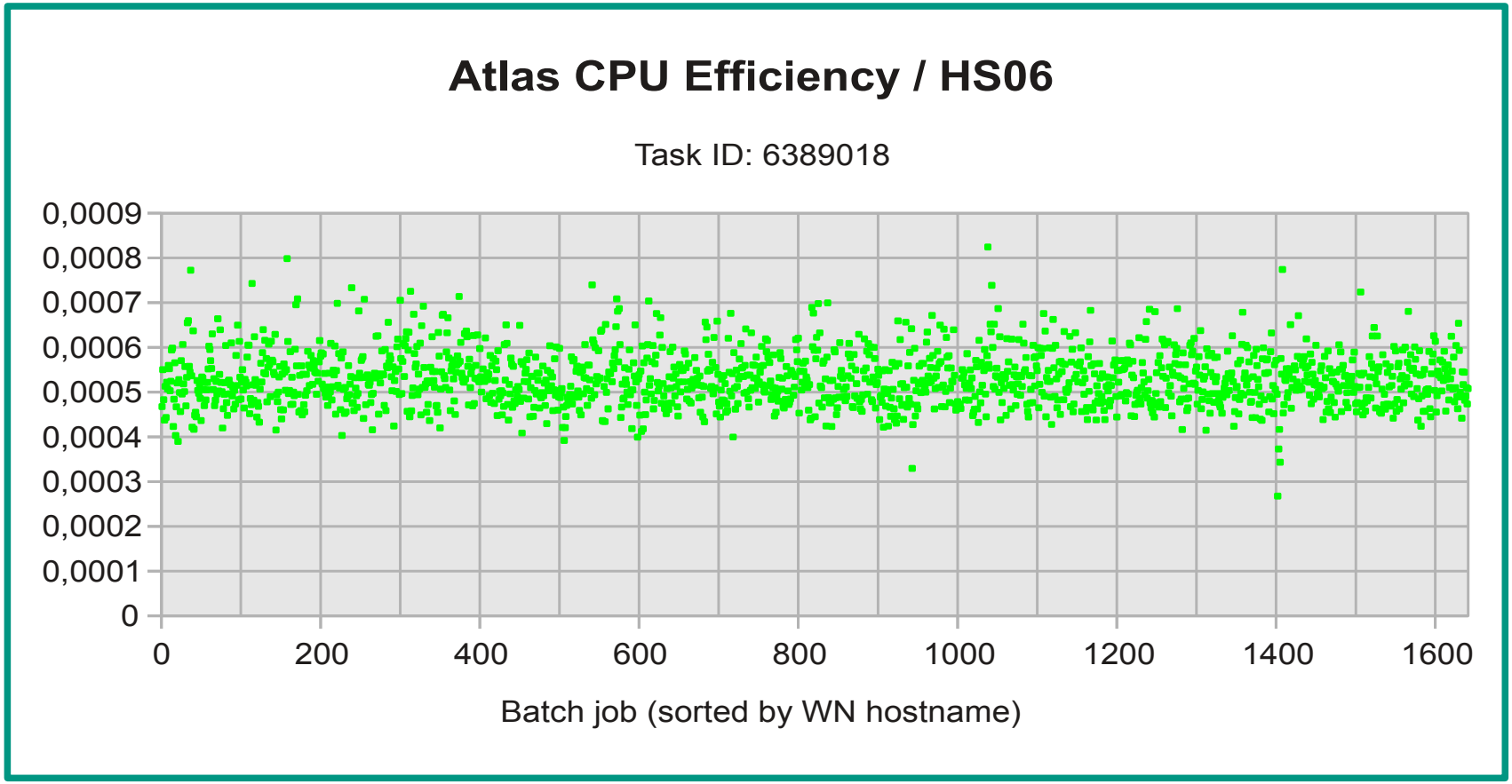
Investigations at GridKa

- Scaling of HS06 with fast benchmarks:



Investigations at GridKa

- Finally – Atlas in September:



(Only small differences found between WN models in several Panda datasets)

Analysis

Analysis

- HS06 scaling with several applications changed from September to November 2015:
  - LHCb: better on WNs with Haswell or AMD processors than with Sandy Bridge chips
  - Atlas: ditto
  - LHCb Dirac fast benchmark: ditto
  - Whetstone: obviously no difference, still good scaling with HS06

Analysis

- Affected WN classes:
  - Haswell (Intel Xeon E5-2630 v3) are the newest generation of WNs at GridKa
    - In production since October 2015
  - AMD (Opteron 6168) are the oldest WNs at GridKa

- Common hardware feature of both batches of WNs:
  king-size RAM:
  - Haswell: 4 GB per job slot
  - AMD: 3 GB per job slot
    - AMD: already in production since 2011, therefore not the only cause of the different performance scaling

Analysis

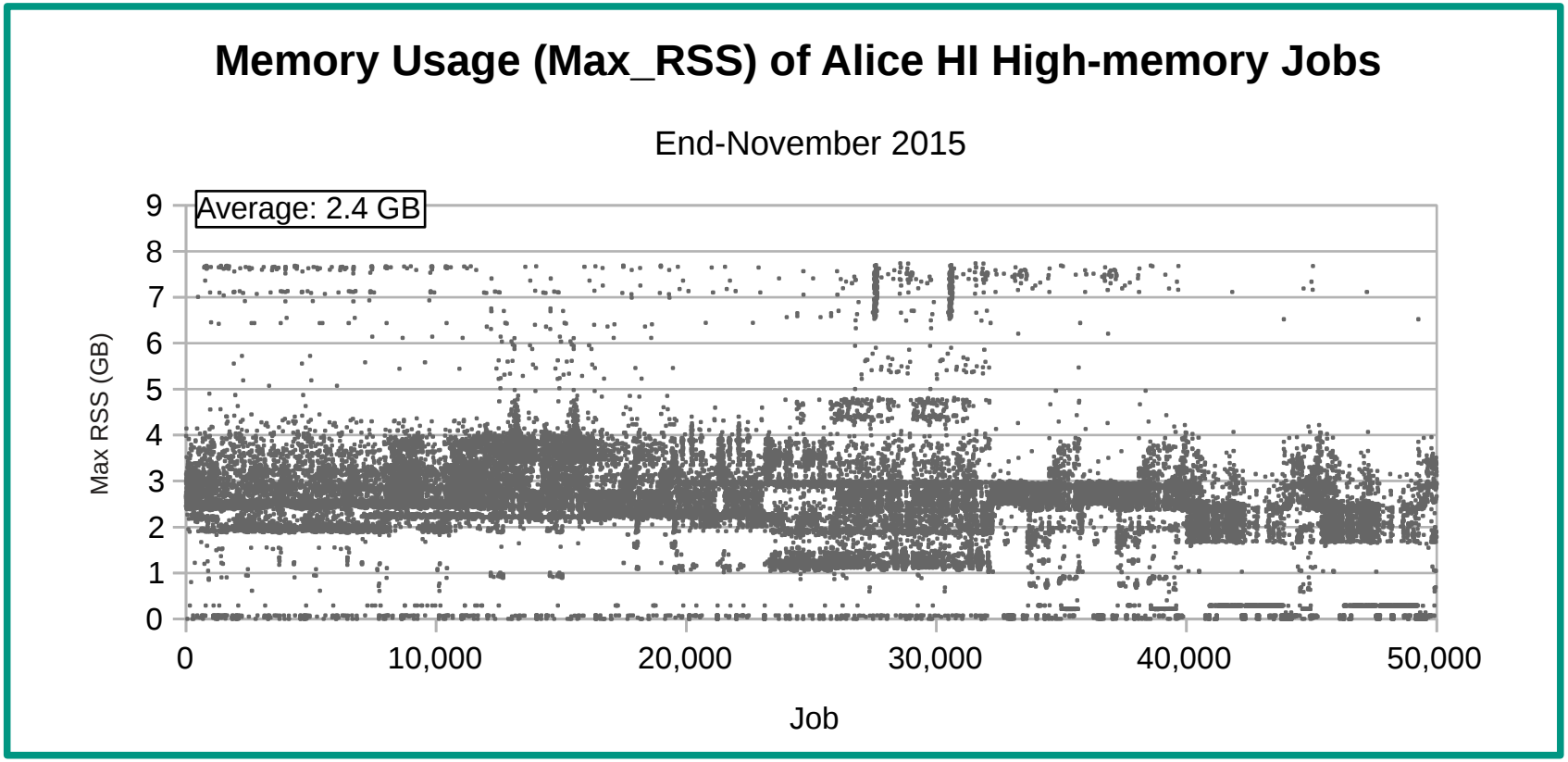
- HS06 benchmark results (HT enabled, 1.5 copies per physical core):
  - Sandy Bridge (Intel Xeon E5-26xx):
    ~ 5.3 HS06/jobslot/GHz
  - Haswell (Intel Xeon E5-26xx v3):
    ~ 5.7 HS06/jobslot/GHz (about 7.5 % better)
    (major enhancement: AVX2, requires special compiler flags)

Analysis

- What has changed at GridKa between September and November '15?
  - New WNs in production (Haswell), 4 GB RAM per job slot
  - UGE update since end-September, now using cgroups to limit memory consumption (RSS soft limit)
  - GridKa is now attracting a lot of high-memory jobs:
    - Alice: > 2200 high-memory (5 GB) job slots permanently busy
    - Atlas: opportunistic high-memory jobs (up to 6 GB)
    - Inhomogenous job scheduling because of high-memory jobs
  - True high-memory job scheduling, managed by cgroups, no arbitrary conversion to multicore jobs
    - "Memory Tetris"
      - 16*5 GB + 8*2 GB on WN with 24 slots and 96 GB
    - No degradation in the number of usable job slots, almost never idle slots

Analysis

- Memory usage pattern?
  - ➔ Alice HI high memory jobs:



Average memory consumption: < 1 GB

Analysis

- Memory usage pattern?
  - ➔ Whetstone:
    homeopathic memory footprint < L3 cache size of modern CPUs

Analysis

- Other causes?
  - ➔ Philippe Charpentier has now presented new results demonstrating outstanding performance of WNs with Intel E5-26xx v3 (Haswell) processors at several sites
    (https://indico.cern.ch/event/319754/session/0/contribution/8/attachments/1202029/1749779/151209-MJFUpdate-LHCb.pdf)
    - Strange differences in 'WNModel vs Job/MJF' performance results of WNs with different generations of Intel E5-26xx chips ...
      - Sandy Bridge, Ivy Bridge: ~ 1
      - Haswell: ~ 1.4
    - ... and AMD Opteron:
      - Magny-Cours (61xx): ~ 1.3
      - Interlagos (6276): ~ 1
      - Abu Dhabi (6376): ~ 1.3

Analysis

- Other causes?
  - ➔ Compiler flags:
    - Are all WLCG experiments still using the default gcc flags when compiling their software?
      - `-O2 -pthread -fPIC -m32`
    - The LHCb Dirac fast benchmark is a Python script; Python RPM in SL6 is build with
      - `-pthread -g -O2 -O3`
    - Whetstone (coming with UnixBench package) has been compiled using default flags
      - `-O2 -ffast-math`

Conclusions

Conclusions

- Scaling issue of HS06 versus several applications detected
  - → LHCb results indicate performance boost (bonus), no degradation
- On the other hand, a fast benchmark with very small memory footprint is (in average) still scaling well with HS06
- Possible causes:
  - → Available memory per job slot
  - → Cgroups
  - → Inhomogeneous job scheduling because of high-memory jobs
  - → Compiler flags?
- Not fully understood yet
  - → Big differences between similar hardware models found by Philippe Charpentier
  - → Issue with HS06 itself, or with the operating environments?

Questions, Comments?

Appendix

- WN hardware models at GridKa:
  - ➔ Production:
    - Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40GHz
      - 2*8 cores+HT, 96 GB (DDR4-2133), 24 job slots
    - Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60GHz
      - 2*8 cores+HT, 48 GB (DDR3-1600), 24 job slots
    - Intel(R) Xeon(R) CPU E5-2665 0 @ 2.40GHz
      - 2*8 cores+HT, 48 GB (DDR3-1333), 24 job slots
    - AMD Opteron(tm) Processor 6168 (@ 1.9 GHz)
      - 2*12 cores, 72 GB (DDR3-1333), 24 job slots

Appendix

- WN hardware models at GridKa:
  - ➔ Benchmarking:
    - AMD Opteron(tm) Processor 6174 (@ 2.2 GHz)
      - 4*12 cores, 96 GB (DDR3-1333), 48 job slots
    - AMD Opteron(tm) Processor 6376 (@ 2.3 GHz)
      - 4*16 cores, 128 GB (DDR3-1600), 32 or 64 job slots
    - Intel(R) Xeon(R) CPU E5-2660 v3 @ 2.60GHz
      - 2*10 cores, 128 GB (DDR4-2133), 20 job slots
      - 2*10 cores+HT, 128 GB (DDR4-2133), 32 job slots
    - Intel(R) Xeon(R) CPU E5630 @ 2.53GHz
      - 2*4 cores+HT, 24 GB (DDR3-1333), 12 job slots
    - Intel(R) Xeon(R) CPU E5520 @ 2.27GHz
      - 2*4 cores, 24 GB (DDR3-1333), 8 job slots
      - 2*4 cores+HT, 24 GB (DDR3-1333), 12 job slots