Karlsruhe Institute of Technology

WN Setup and Procurements at GridKa Tier1

Manfred Alef, Max Fischer / pre-GDB 2021-07-13

Batch Farm at GridKa Tier1

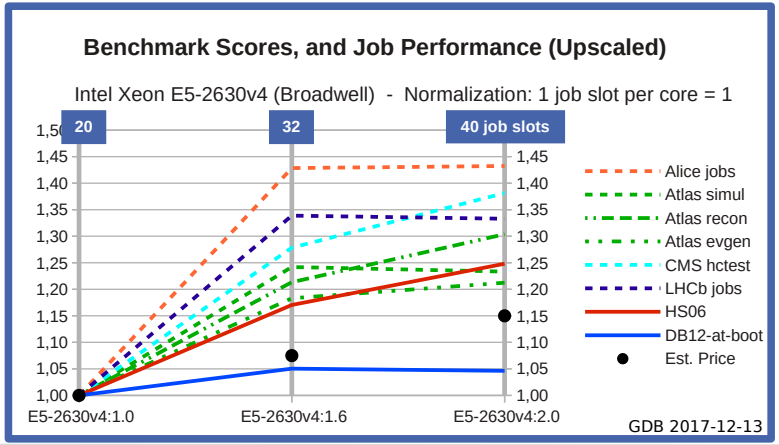
- Approx. 550,000 HS06 (Apr 2021), x86
- Supporting ALICE (26%), ATLAS (30%), CMS (14%), LHCB (20%), Belle 2, Compass, IceCube, WeNMR, ...
- HTCondor (running on bare metal): 1-core, 8-core jobs are supported, 4 HTCondor-CEs
- WN models:
  - Intel E5-2630 v4 (Broadwell, 10-core, 2.2 GHz)
  - AMD EPYC 7702 (Rome, 64-core, 2.0 GHz)
  - AMD EPYC 7742 (Rome, 64-core, 2.25 GHz) coming soon
  - (Some older hardware models still available for benchmarking purposes)

Hardware Details

- Default setup: SMT on, ~1.5 job slots per physical core (multiple of 8 job slots to optimize 8-core job scheduling)
- Memory:
  - At least 4 GB RAM / (phys.) core
    ⇒ ~3 GB RAM / job slot in default setup
  - (Older WN models providing at least 2 GB RAM / job slot)
- Hard disks, SSD(s):
  - At least 30 GB scratch space and swap (> 4 GB) per job slot
  - Latest WN models: SSD(s), older systems: multiple hard disks
- Network connection: 10GbE (older WN models: GbE)

Job Slot Configuration

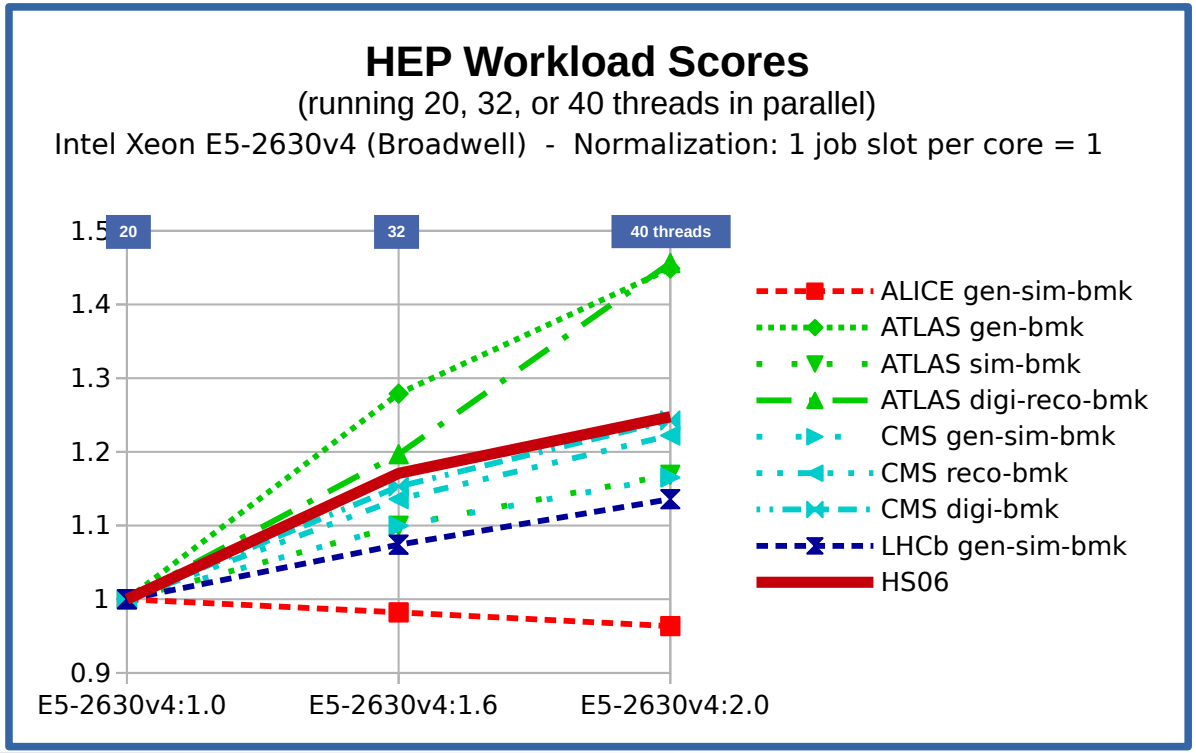
- Default setup: SMT on, ~1.5 job slots per physical core (multiple of 8 job slots to optimize 8-core job scheduling)
- Why that uncommon setup?
  - First WN models with SMT installed in 2009, trying to find out the best configuration:
    - optimizing HS06/€
    - comparing number of jobs per HS06
  - Validation based on job power (events/s) reported by experiments *



* https://indico.cern.ch/event/578993/contributions/2801058/attachments/1574522/2485835/Benchmarking-update-2017-12-13.pdf

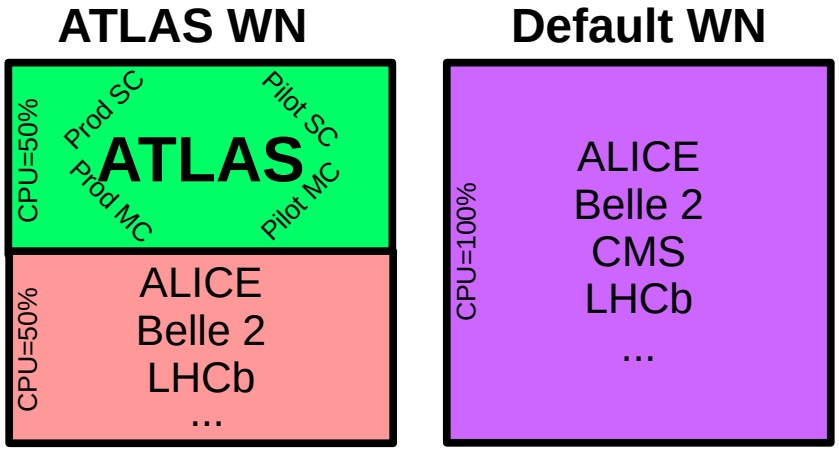
Benchmarking ("2nd Dimension")

- Results of the individual HEP workloads (running 20, 32, or 40 threads in parallel)



ATLAS Unified SC/MC Queue Support

- Modified setup optimizing ATLAS unified SC/MC queue scheduling
  - Several issues caused by frequent MC-SC-MC transitions
  - Solution (since autumn 2020):
    dedicated ATLAS farm partition, still maintaining job mix by implementing 2 HTCondor partitionable slots on subset of WNs, serving either only ATLAS, or only the other VOs



Procurements of WN Hardware

- Inviting tenders for the requested amount of HS06 (not for #boxes)
  - x86 CPU, memory, SSD (TBW: 50TB/core), network interface, BMC, ...
- Vendors run benchmark (optimizing HS06/€)
- Ordering the tenders: based on estimated lifecycle cost considering additional boundary conditions, e.g. power consumption, rack space, network ports, ...
- Acceptance:
  - Running HS06 in a loop (burn-in + HS06 re-measure)
  - General burn-in

Opportunistic Resource Support

- Opportunistically provide resources from external sources
  - Dedicated HTCondor cluster separate from Tier1 cluster
  - CE+Scheduler managed by Tier1, best-effort support for resources
  - Integration via COBalD/TARDIS "pilots", running VO pilots as payload
- Wide range of resource shapes and kinds
  - 1 – 96 Slots, 2 GB – 4.5 GB RAM/slot, sometimes GPUs
  - Changing availability, volume and partitioning of resources
- Not yet used to improve scheduling, benchmarking, ...

Questions, Comments