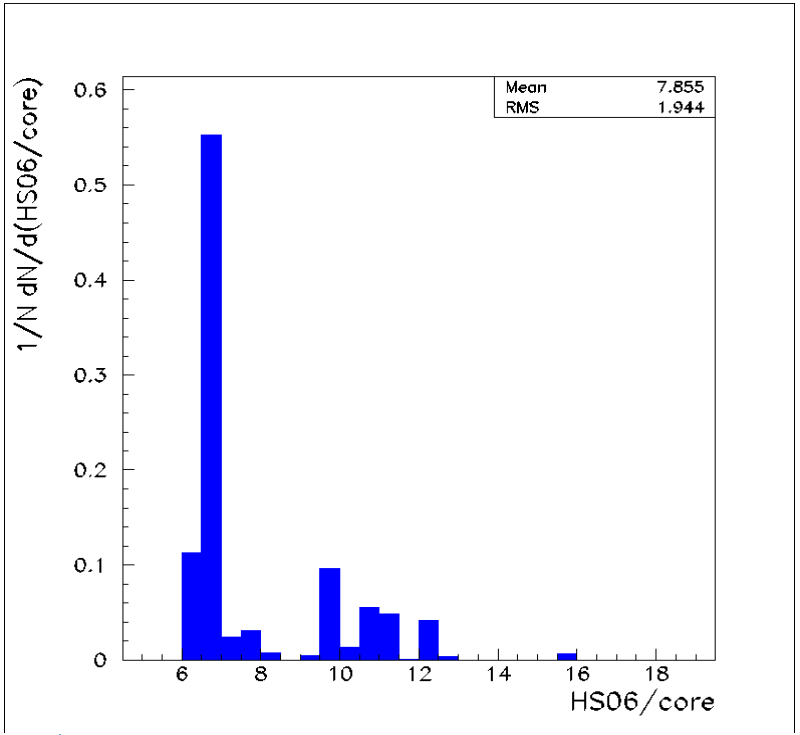

CERN

VM benchmarking: update on CERN approach

This is a follow-up from the slides I showed in September 2014

Update: CERN IaaS for batch

- Most hardware resources under IaaS
  - Heterogenious hardware
  - Complexity partly hidden by virtualization
  - Hypervisor and its performance hidden
  - Still large spread of per core performance
- CERN LSF batch farm:
  - About 3700 nodes in total, ~3100 VMs
  - About 2900 in public resources
    - Got rid of old physical worker nodes
    - 92% on virtual machines now
  - Traditional GRID worker nodes
  - Traditional APEL based accounting (HS06)
- In addition dedicated IaaS projects for experiments



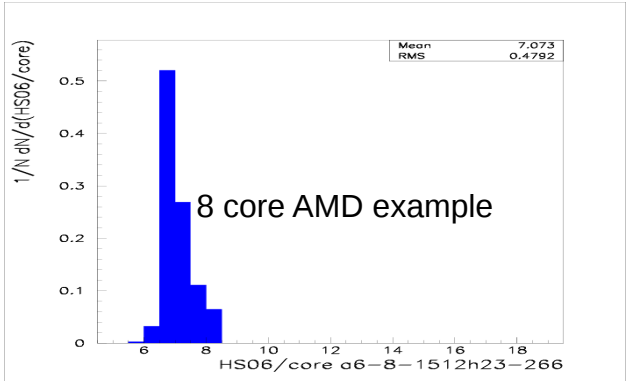
Reminder: Classification of worker nodes

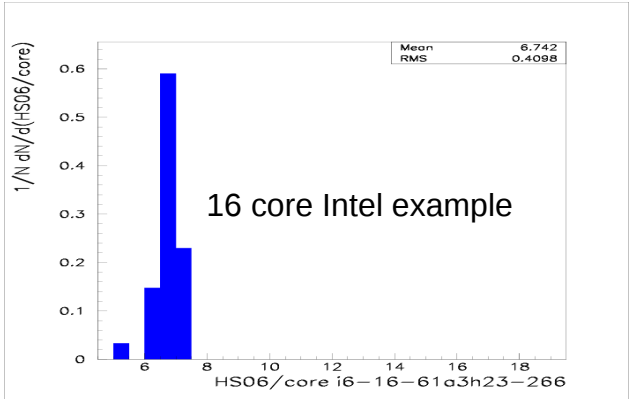
- Idea for CERN batch farm:
  - Classify VM cores using info from
    - Operating system
    - /proc/cpuinfo
    - dmidecode
  - Benchmark a large number of VMs per class
  - Be pessimistic about the performance
    - jobs may run faster than expected
    - capacity counts will be pessimistic

Classification of worker nodes

Example: i6_8_61a5h16_1066

- Intel Nehalem based machine
- SLC6
- 8 cores
- CPU-ID 61a5h
- 1600 MHz
- memory speed 1066 MHz





Remark : Details of the machine:

Memspeed = 1066
Cpuspeed = 1.6
Cpucache = 8192
Cpuvendor = GenuineIntel
CPUfamily = 6
CPUmodel = 26
CPUstepping = 5

=> CPUID = 61a5h

Updates since September at CERN

- Retirement of old nodes from the farm
- Improvements on the batch tenants
  - Ensure CPU pass-through is enabled on dedicated tenants
  - I/O improvements
  - Memory management for the hypervisor
  - Re-basing of VMs
- Rollout requires restart or re-creation of VMs
- Memory speed still not available inside VMs

Status at CERN

- Fully deployed on CERNs batch farm
  - Physical and virtual machines
  - Mapping hardware type → HS06 rating used for
    - Local accounting
    - WLCG accounting via APEL
    - LSF scheduling via CPU factors

Status at CERN

- Local cloud accounting
  - Includes non-managed and VO owned VMs
  - Based on Ceilometer:
    - It does not provide enough information to determine the hardware type
    - Still some issues on stability in large installations
  - Use estimate of HS06 based on the performance of the hypervisors

Status at CERN

- APEL based cloud accounting
  - Based on Ceilometer
  - Currently cannot report performance
  - Discussions ongoing about extensions of the SSM schema to support
    - VMPerformance reporting
    - Ways to fix the long running VM issue

Summary

- Established new schema to classify virtual batch worker nodes
  - Used for scheduling and accounting in LSF
  - usable by anybody having access to a VM
  - Reflects changes in the underlying hardware
- Rolling out a bunch of changes which will
  - Improve I/O and CPU efficiency
  - Improve classification a bit
- Local cloud accounting
  - Not all information available via Ceilometer
  - Using the hypervisor performance for now
  - Aim is to change that at the longer term

www.cern.ch