Tier-0 Update

Helge Meinhard, CERN-IT

Grid Deployment Board

04-Nov-2015

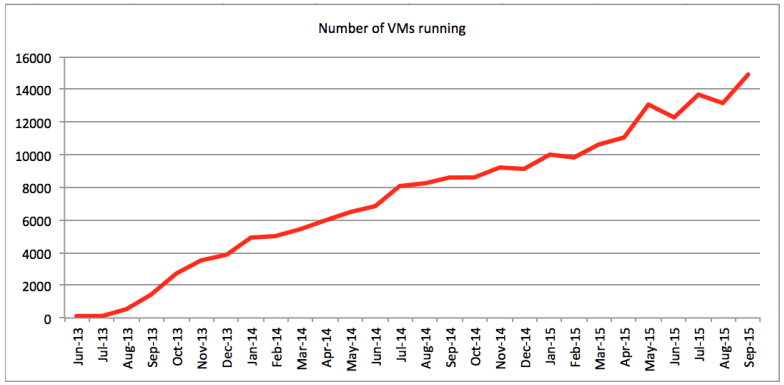
Outline

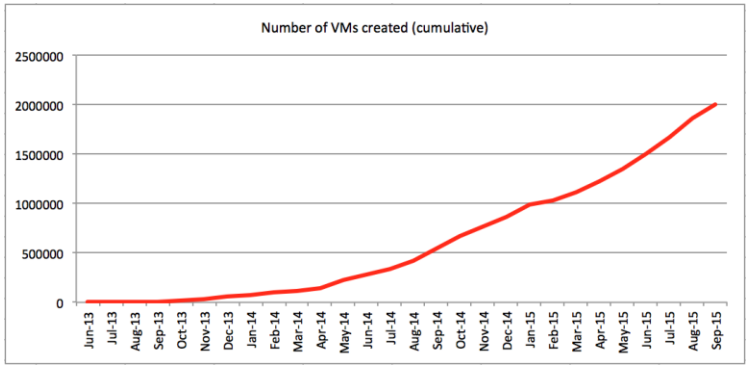
- Cloud
- Databases
- Data and storage
- Network
- Platform services
- Infrastructure

Cloud

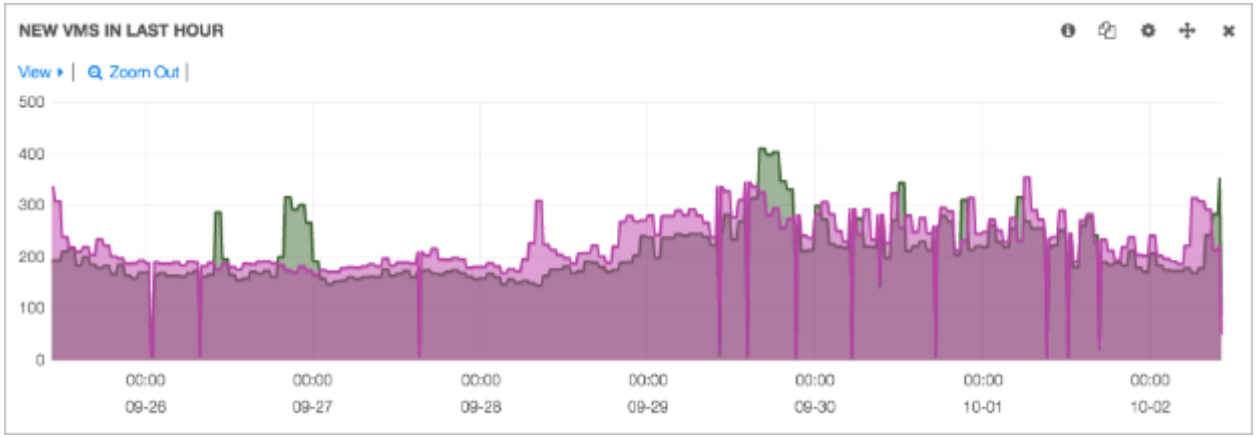
CERN Cloud in Numbers (1)

- 4’800 hypervisors in production (1y ago: 3000)
  - Majority qemu/kvm now on CC7 (~150 Hyper-V hosts) (SLC6)
  - ~2’000 HVs at Wigner in Hungary (batch, compute, services) (batch)
  - 250 HVs on critical power
- 130k Cores (64k)
- 250 TB RAM (128TB)
- ~15’000 VMs (8’000)
- To be increased in 2016!





CERN Cloud in Numbers (2)



Every 10s a VM gets created or deleted in our cloud!

- 2’000 images/snapshots (1’100)
  - Glance on Ceph
- 1’500 volumes (600)
  - Cinder on Ceph (& NetApp)

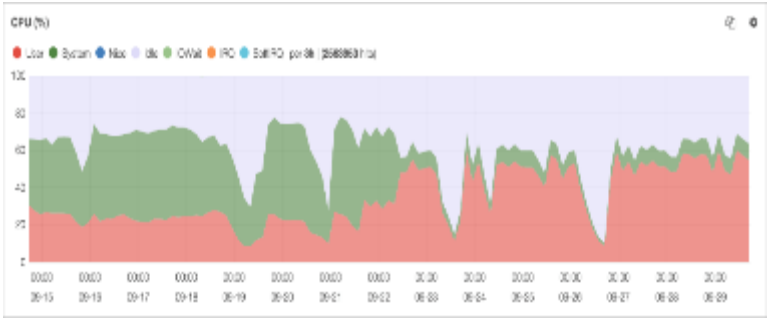
Ongoing service improvements

- Adding hardware
  - Today: ~4800 compute nodes (130K cores)
  - Underway: 25K cores
  - Spring 2016: 60K cores
  - Retirements: being clarified
- Operating system upgrades on the hypervisors
  - 3900 * CC7 hypervisors
  - 700 * SLC6 -> to be upgraded to CC7 in the next 6 months
  - 34 * RHEL6 -> to be upgraded to RHEL7

CPU performance

VM sizes (cores)	Before	After
4x 8	7.8%	3.3% (batch WN)
2x 16	16%	4.6% (batch WN)
1x 24	20%	5.0% (batch WN)
1x 32	20.4%	3-6% (SLC6 … WN)

- Performance
  - Dependence of optimisations from hardware types and other optimisations
  - NUMA, pinning, huge pages, EPT
- Pre-deployment testing not always sufficient
  - Small issues can have major impact
- Performance monitoring
  - Need continuous benchmarks to detect performance changes



- Requires OpenStack Kilo version of Nova
  - Planned for mid-November
- Details presented in previous GDB/HEPiX
  https://indico.cern.ch/event/384358/session/12/contribution/15/attachments/1170139/1689493/Optimisations_of_Compute_Resources_in_the_CERN_Cloud_Service_-_HEPiX14OCT2015.pdf

WIP: Container integration

- Started to look into integration of containers with our OpenStack deployment
  - Initially triggered by the prospect of low performance overheads
  - LXC due to the lack of an upstream Docker driver (not suitable for general purpose)

- We've setup a test cell
  - Performance looks good
  - OpenStack patches for AFS & CVMFS done
  - AFS in containers: kernel access, multiple containers, tokens, …
  - Operational issues still to be understood

- Started to look into OpenStack Magnum
  - Container orchestration via Docker or Kubernetes become first class OpenStack resources

Databases

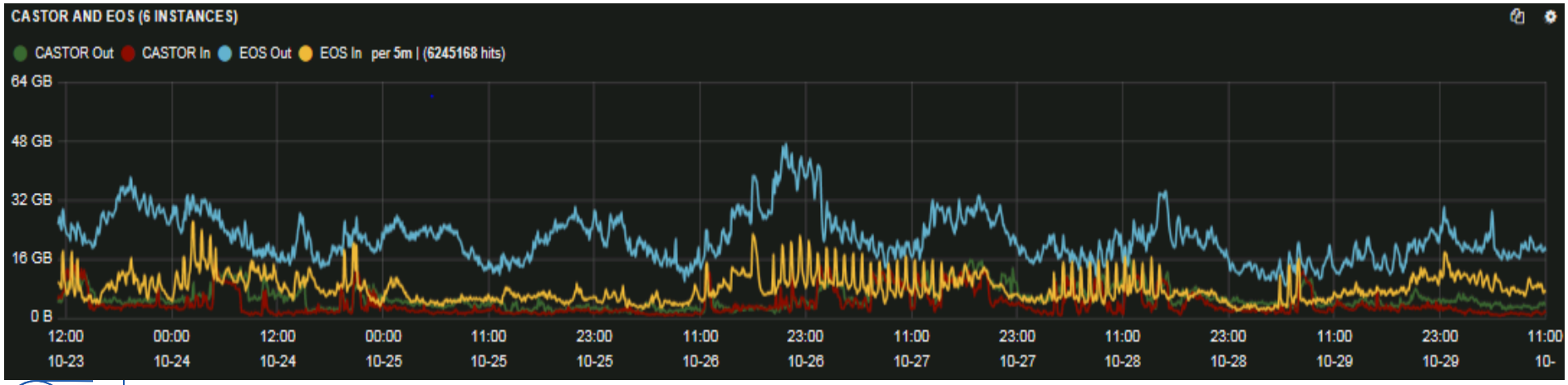
Hadoop, Scale-Out Databases

- Working together with ATLAS on optimizing the design and performance for Event Index application based on Hadoop infrastructure
- Started the Hadoop Users' Forum, with fortnightly meetings aimed at sharing experience on using Hadoop components at CERN, between users and with IT service providers

Data and Storage Services

Data services for Tier0 (1)

- **CASTOR**
  - Located in Meyrin, notably the 120 PB tape archive
- **EOS**
  - 70 PB disk space available; evenly distributed across Meyrin and Wigner
  - About 80% of the disk resources for Tier0 (remaining 20%: CASTOR)



Data services for Tier0 (2)

- **Different T0 workflows**
  - ATLAS and CMS → EOS
    - EOS is the distribution centre for T1 export, batch processing and tape (EOS→CASTOR)
  - ALICE and LHCb kept the Run 1 model
    - Experiment → CASTOR (which is the distribution centre for RAW data for T1 export and batch processing). EOS used only for analysis
  - Tier0 rates:
    - EOS reading often exceeding ~30 GB/s (Tier0 + analysis)
    - CASTOR in beam max ~15 GB/s (input dominated by raw from LHC, output mainly tape writing)
  - Smooth data taking
    - Important reconfiguration to prepare for the ALICE PbPb run (provide better isolation between recording and raw reconstruction)

Networking

Networking

- Will reduce number of servers exposed to the LHCOPN and LHCONE
  - Requires major reconfiguration of the datacentre network
  - Foreseen for end of February 2016 (during a scheduled technical stop)
  - Tier1s to update their configuration (see http://cern.ch/go/6Whh)
- Third 100 Gbps link between Geneva and Budapest expected to go into production in December 2015 or January 2016
- PoP established by NORDUnet at CERN Geneva
  - Peered with CERN with two 10 Gbps links
  - More network redundancy and capacity to the Nordic countries

Platform and Infrastructure Services

Batch (1)

- LSF being upgraded from 7 to 9
  - Master: First attempt on 13 October failed; this morning: all successfully completed within less than 1.5 hours
  - Worker/submission nodes to follow now (usual QA procedure)
- HTCondor in production since 02-Nov-2015
  - 2 HTCondor CEs
  - 2 ARC CEs obsolete; one to be retired next week, one before end 2015

Batch (2)

- Transition from LSF to HTCondor
  - HTCondor: currently 4% of total batch
  - Will move capacity from LSF to HTCondor to cover basic Grid submission load (20…25%)
    - Weekly steps of 500…1000 cores
  - Kerberos ticket handling: prototype before end 2015, production to be defined
    - Documentation, training, consultancy to help migrating locally submitted jobs
  - Capacity to be moved further once users can submit jobs locally
  - Deadline: No LSF-based services at end of Run 2

Other Platform Services

- Middleware: Some recent issues with FTS and ARGUS
  - Good collaboration with developers
  - Some ARGUS outages due to deployment choices or bad input
- Lxplus: frequent crashes understood and fixed
  - Tracked down to incompatibility between cgroups and a range of kernels
- Volunteer computing: Good progress by CMS, continued high level of ATLAS usage

Infrastructure Services

- GitLab-based service set up at CERN
  - Very fast take-up: more than 1’500 projects already
  - Complements GitHub for projects requiring restricted access or tight integration with CERN’s environment
- Plan to phase out git(olite) and SVN
  - New repositories require moderator approval
  - Plan to be aggressive on git, but more subtle on SVN (rundown targeted some time during LS2)
- Continuous integration very popular

Acknowledgements

- Tony Cass
- Eva Dafonte Perez
- Massimo Lamanna
- Jan van Eldik
- Arne Wiebalck

Questions?