HEPIX Fall 2016

Summary Report

Outline

- Key Statistics
- Track highlight
- Next meetings
- Acknowledgements

Key Statistics

- 96 registered participants
  - 12 from Asia Pacific
  - 33 from North America
  - 45 from Europe
  - 1 from Israel
  - 5 sponsors & commercial enterprises
- 74 accepted abstracts
- Four Tony's – a HEPIX record!
- Five incredibly beautiful sunny days – another HEPIX record!
- Several contributions from nearby sites (SLAC, LBL, UCSD, Caltech, TRIUMF, etc)
- Multiple contributions from Asia-Pacific (KEK, IHEP, ASGC, Tokyo, Sidney, Melbourne, CSNS)

Key Statistics (cont'd)

- Tour of new data center generated high level of interest (~60 people in 4 groups?)
- 7 presentations on HPC systems and/or HTC access to HPC systems
- Presentations by tracks
  - Site Reports (25)
  - Security & Networking (11)
  - Storage & Filesystems (12)
  - Grid, Cloud & Virtualization (8)
  - Batch & Computing Services (7)
  - IT Facilities & Business Continuity (4)
  - Basic IT Services (5)
  - End-User IT Services & Operating Systems (2)

Site Reports

- 25 site reports representing major HEP, NP, Photon Science projects, HPC leadership class facilities (LCFs) and other under-represented projects
- GPFS, Hadoop, Dropbox (CERNBox at CERN) used by many sites as storage solutions (useful for building critical mass for community support)
- Lustre widely used (several reports of problems, too)
- AFS slowly dying off with no single one-to-one replacement solution

Site Reports

- Many sites reporting integration of HPC and HTC activities – leveraging effort and experience that can benefit looming computing challenges
- Wider adoption of OpenStack and Docker to facilitate the integration of cloud (private or otherwise) resources for elastic surge (or expansion) of activities
- Monitoring (data center infrastructure, hardware, services, etc) done with multiple solutions – possible area for future collaborations and coordination?

Site Reports

- SDN R&D and initial deployment underway at various sites
- Preparations for migration to IPv6/dual-stack to meet WLCG deadlines
- Many sites upgrading WAN connectivity to meet (mostly LHC-related) needs
- Indirect funding model at SLAC
  - Possible model for HTC sites supporting new HPC activities
  - Need buy-in from stakeholders
- ITER (fusion reactor)
  - Description of computing for control systems
  - Similarities with HEP computing/storage tools -- possible synergies with HEPIX community

End-User IT Services

- SL update by FNAL
  - v7.3 in Beta (targeted at cloud provisioning)
  - support for container/VM provisioning
- DESY hosts 70+ email domains
  - Over 300k emails daily
  - Quarantine to minimize spam, phishing, viruses, etc
  - Replace pure commercial with mostly open-source solution
    - Balanced approach to filtering/quarantine (not too liberal, not too restrictive)

Network & Security

- Network analytics project at U. of Chicago
  - Instrumented to detect anomalies, alarms, etc with existing tools (ELK, perfSONAR, etc)
  - Can be used for data/job placement, improved productivity – applicable beyond ATLAS
- KEK's discussion of WAN (SINET5) upgrade to connect to LHCONE
  - SINET-ESNET-CERN path implemented
  - Driven by Belle II computing requirements

Network & Security

- CERN presentation on intrusion detection
  - Decouples network decision logic from traffic flow
  - SDN-based(OpenFlow, OpenDaylight and BFO)
  - Better control over traffic distribution
- IHEP also investigating SDN
  - Network security implications
  - Relieves bottlenecks in cloud computing
  - Testbed results are promising – plans for expansion

Network & Security

- Report by IPv6 working group
  - Plans for IPv6 only cpu's in 2017
  - HEP (and commercial) support for IPv6 growing
  - WLCG proposed timelines for IPv6 support – migration underway at many levels
  - Need to address security concerns
- Security updates
  - Continuous attacks – now targeting corporations
  - Non-trivial to detect and block malware (ie, Dridex)
  - Analyze and quarantine suspicious emails
  - Issue with ransomware
  - Complications with federated identity management – Sirtfi is a vehicle to address intelligence sharing

Network & Security

- Wi-Fi infrastructure @ CERN
  - Gaps in coverage makes roaming on-site difficult
  - Migrate to controller-based solution to improve high-density, seamless roaming
  - Simulation and surveys to map signal strength – tests indicate acceptable performance
  - Pilot deployment underway with global deployment beginning in early 2017
- Research sites & cloud resources
  - Need better bandwidth to cloud providers
  - Issues with NREN carrying cloud traffic
  - Different approaches taken by ESNET and GEANT

Network & Security

- Presentation on EduGAIN (federated identity management worldwide)
  - Policies need to address security incident response
  - Issues with trust, inter-federation transparency, operational support
  - Sirtfi as a vehicle to address some concerns – help from research communities needed
- Security @ NERSC
  - Small # of incidents but high impact – isolating compromised users
  - BRO provides IDS, monitoring & event correlation data
  - Several examples of hacker attacks discussed

Storage & Filesystems

- Sven Oehme's presentation on Spectrum Scale (formerly GPFS) upgrades
  - Possible (partial) replacement for AFS
  - Several changes aimed to improve performance
    - New configuration parameters
    - Better communication (lower latency)
- CephFS presentations by Australia, RAL and BNL
  - Works well, but still buggy – users advised not to rely exclusively on it
  - Data corruption, daemon crashes, etc among problems found
  - Significant development and steep learning curve for system administrators

Storage & Filesystems

- HA dCache presentation by NeIC
  - HA pair deployed recently with two virtual machines – work in progress
- BNL presentation on mass storage
  - Reached cumulative 90 PB in 2016
  - Scheduler (ERADAT) based on code originally from Oak Ridge – requests to open-source it
  - Described massive parallel data staging experience (2 GB/s aggregate, ~150 MB/s per drive)

Storage & Filesystems

- CERN presentation on EOS, DPM and FTS
  - EOS is a high-performance, disk-only storage – description of namespace interface, data structure
  - DPM – storage manager at smaller sites
    - Originated with CASTOR
    - Possibly operate WLCG storage as a cache
  - FTS – data distribution across WLCG infrastructure
    - Deployed at CERN, RAL, BNL and FNAL
    - v3.6 to be released soon
      - better scalability and performance
      - Wider integration with WLCG worflows

Storage and Filesystems

- ZFS on Linux
  - Originally developed for Solaris, but supported later in Linux
  - Combines filesystem, volume manager and raid system (equivalent to xfs, ext4, etc)
  - Performance-wise, it compares well with hardware RAID + xfs/ext4
  - Deployed at some UK sites (local and grid storage)
  - Description of limitations and differences with traditional file systems
- OSiRiS update
  - Software-defined storage (projected supported by NSF involving several US universities) – based on Ceph
  - Provides common storage infrastructure to participants – reduces cost
  - Participants include HEP, Earth Sciences, Life Sciences, etc
  - Latency test at SC16 to measure how far it can be stretched
  - Working to integrate with ATLAS

Storage & Filesystems

- Database services at CERN
  - Support for Oracle, MySQL, PostgreSQL,, etc
  - Currently 415 on-demand DB's supported
  - Upgrades and improvements (SSL encryption, HA features, monitoring, etc) soon
  - Apache Kafka project
    - Data storage in distributed replicated cluster
    - Data back-up for Hadoop

Storage & Filesystems

- Status of AFS at CERN
  - Used since 1990 (35k users and 450 TB of storage)
  - Used for $HOME, working directory, shared space, etc
  - Possibly, combination of CERNBOX, EOS, CVMFS and others – software maturity issues, need to sort out workflow consequences
  - Possible end-of-service ~2019 (before end of LS2)
- AuriStor presentation
  - KAFS within Linux kernel – integrates with AFS and AuriStorFS servers
  - Migration from OpenAFS to AuriStorFS requires no downtime
  - Container support in AuriStorFS
  - AuriStorFS participation in proposed Tennessee Open Cloud Project

Computing & Batch

- Report by benchmarking working group
  - Regular Vidyo meetings held
  - A few candidates for fast benchmarks
  - Development of a CERN benchmark suite for cloud and batch
    - Support by LHC experiments
    - Scaling and correlation studies ongoing
    - Investigation of reliable HS06 estimators with fast benchmarks
  - Expect proposed fast benchmark by early 2017 and development of long-running benchmark in mid-2017
  - Additional contributors (ie, neutrino code) welcome

Computing & Batch

- Tony Wildish's presentation on genomics
  - Data sizes comparable to LHC experiments
  - Growth not linear like HEP
  - Not as processing-latency sensitive as HEP, but considerable variety in data types
  - Data production and distribution not as well organized or controlled as HEP
  - Cost of data production falling steeply
  - Larger communities
  - No reliable predictors of computing needs for data analysis
  - Very heterogeneous set of software tools
  - Gov't shaping more organized effort on data and computing management—requires cultural shift in community

Computing & Batch

- JLAB presentation on USQCD computing project
  - KNL cluster production allocation began recently
  - Adding additional hardware to KNL cluster soon
  - Rated at 329 Tflops without additional hardware—will rerun soon
- Integration of ARM64 and Power8 at CERN
  - Benchmarked with HS06 – comparable to Xeon processors but many more cores
  - Significant higher power consumption – not as efficient as Xeon
- UCSD presentation on dynamic provisioning of cloud resources using HTCondor
  - Elastic expansion for cloud resources
  - Focused on AWS access
  - Proof of concept to evaluate feasibility of moving workflow to clouds

Computing & Batch

- HTCondor presentation on latest updates
  - Slurm support
  - Improved support for OpenStack and AWS
  - Support for Singularity (container system like Docker)
- Data intensive applications at PDSF and Genepool (NERSC)
  - Workflows pose different challenges than MPI-based applications at Cori
  - Supported by Slurm (batch) and GPFS (storage)

Facilities & Infrastructure

- CERN OpenCompute project
  - Simplify and standardize hardware procurement to lower costs and minimize customization
  - DC power at rack-level
  - OpenCompute responses not competitive in regular procurements (yet)
- New Data Centers at CERN
  - Renovate existing building for 2nd site with core network equipment only (48 racks and 120 kW capacity) to insure high-availability
  - On UPS and ready by next summer
  - New DC to accommodate more T0 capacity and HLTF's for LHCb and Alice
  - Consider GSI's Green Cube model – want it ready by 2020

Facilities & Infrastructure

- GSI green cube status report
  - Operational since Spring 2016 – no major infrastructure issues since
  - 13k cores and 20 PB over 50 racks by end of year
  - Data center monitoring being re-designed
    - Better reporting (status, alarms, etc)
    - Increased automation

Basic IT Services

- BNL presentation on HTCondor monitoring
  - Bricolage of various pieces adapted from open-source tools (Ganglia, Nagios, etc)
  - Migration to Graphite/Grafana – contribution to HEPIX working group
- Load-balancing at RAL
  - Hardware failure on HA systems with fixed DNS leads to degraded service
    - DNS round-robin not fool-proof (ie, first host picked up is down)
    - Build a load-balancer built with HAProxy and KeepAlived (floating IP)
    - Instrument process to check if back-up server is healthy
    - Developing scalable solution to address multiple failure scenarios
  - Works well with FTS3, OpenStack now
    - in the future add GridFTP, S3 API
    - Other services (CASTOR, etc) could benefit

Basic IT Services

- Puppet at the Australian ATLAS sites
  - Migrated from Cfengine to Puppet a few years ago
  - Gradually puppetizing most (not all) servers – legacy reasons
  - Migration to CentOS7, KVM prompted adoption of “best practices” – upgrade Puppet and puppetize everything
- ELK deployment at KEK
  - Issue of access control features – Kerberos-based controls installed but upgrades to ELK broke them
  - Develop plugins for better portability
  - Overhead of these enhancements is 80-120 ms/query (13% overhead on indexing throughput)
- RH Satellite 6 for lifecycle management (FNAL)
  - Originally only for workstations – request for wider support but limited manpower
  - Heavy upfront effort required and several bugs and deficiencies in early versions
  - Learning curve for traditional sysadmins

Grid, Cloud & Virtualization

- ANL presentation Chameleon
  - Facility for CS experimentation (650 nodes, 14.5k cores, 5 PB over 2 sites connected with 100G network)
  - 1000+ users/200+ projects on testbed built with commodity components
  - IB, SSD's NVMe's, GPU's, FPGA's, ARM, Atom, etc available in testbeds
  - Bare metal or VM images – user configurable, too
- CNAF's extension into external resources
  - Testbed (CMS jobs) with Aruba (Italian cloud provider)
    - Works but much lower efficiency (49% vs 80%)
  - Use Bari resources
    - Appears as local resources – good efficiency
    - Increased CNAF throughput by ~8%

Grid, Cloud & Virtualization

- KIT also working on extending local resources
  - HPC resources and commercial cloud providers
  - Resource integration with HTCondor/ROCED (latter is cloud scheduler with VM provisioning)
  - HPC & virtualization worked well – continued development
- Helix Nebula Science (HNSCi) Cloud report
  - 30-month project to address looming computing & data challenges – develop hybrid cloud model for European Scientific community
  - Many institutions and science fields participating
  - More prominent usage of public cloud resources—several technical challenges to be addressed
- IHEP Cloud Project
  - Support for various projects -- 2k cores, 900 TB storage and 200 users
  - OpenStack-based provisioning – self-service virtual platforms and computing environment
  - VPManager provides dynamic scheduling architecture –worked well and there are future expansion plans

Grid, Cloud & Virtualization

- HEP workloads on NERSC HPC resources
  - Shifter (now open-source) enables Docker-like containers to provide static environment suitable for HEP workflow
  - Burst buffer – dynamic allocation of high-performance filesystems for data-intensive workflows
  - Slurm and SDN used to improve bandwidth to local resources
- Container orchestration at RAL
  - Aim to manage existing services and provide more services with less effort
  - Apache Mesos for service discovery & management and Docker-based images in private registry – scalable so far
  - Future plans
    - Integration with configuration management system
    - Ceph-based persistent storage for containers
    - OpenStack hypervisors in containers – cloud/batch share resources

Grid, Cloud & Virtualization

- CSNS presentation
  - Accelerator-based neutron source (branch of IHEP) in China – begins operations in 2018
  - Openstack-based provisioning in data center
  - Building up software infrastructure to support operations (image storage, authentication, network, OpenStack management, monitoring,etc)
  - Collaboration with and leveraging of HEPIX experience and expertise helpful

BOF on HPC

- Focus on sharing procurement and porting software among sites—Leadership Class Facilities (LFC's) and smaller sites
  - Informal discussions – create mailing list
  - Future BOF's likely
- Computer accounts at NERSC and JLAB for system administrators at other sites are possible if this helps to understand operational issues
- Procurement practices vary among sites – information confidentiality limits sharing
- CVMFS mentioned as a potential vehicle to share software among sites
- Operational issues with KNL systems discussed – long reboot times, setting memory modes, effects on SLURM scheduling, etc
- Vendor(s) can help connect other labs (ie, Sandia) to the effort

Next meetings

- Next Spring meeting
  - Apr. 24-28, 2017 @ Wigner (Budapest, Hungary)
  - HEPIX website updated
- Next Fall meeting
  - HEPIX returns to Asia
  - Oct. 16-20, 2017 @ KEK (Tsukuba, Japan)
  - Mark your calendars

Acknowledgements

- Thanks to LBL/NERSC for making facilities available to this HEPIX meeting
- Thanks to James Botts and the entire local organizing committee – great job!
- Thanks to our sponsors
  - Seagate
  - Intel
  - Penguin Computing
- Thanks to all of you for coming!