Gerard's notes from HEPiX April 2016

All presentations available at https://indico.cern.ch/event/466991/timetable/#20160418.detailed

DMS interesting
HPC
Facilities
Computing/HEPCloud

Introduction

- Some new sites showing up (eg: ALBA syncothron)

IRFU Site report
- Now within CEA
- Tested CEPH and not happy without SSD as it doesn't scale well (read 8.5GB/s, write 3.3GB/s with 11 DELL PE R730xd 16x4TB HDD/server.
- Puppet 4
- Moving all SL6 to SL6.7 as having some issues with epel/rhel dependencies (RPMs only on 6.7)
- 31 water cooled racks (20 and 40 kW)

NDGF Site report
- New ARC cache came from funds that needed spending. Requested lots of IOPS and >200TB.
-- Ended up with lots and lots of 2.5" disks but limited due to a 10Gbps uplink.
- dCache pools on Dell PE 730xd, 2U 16x8TB, 3.5" 7k rpm sata RAID6. One 100TiB pool per node. 10GBaseT uplink. Most loads can saturate the 10Gbps (using dCache as server). Servers have warm exhaust (65C).
- Using TSM for tape behind dCache, TSM managing small files on disk. Built an interfaces that can handle 300K requests per dCache pool without overwhelming it, needed to handle small files.

CERN site report (big summary of future talks)
- AFS phase-out, expect final clean-up during LS2 (2019)
-- No single product replacement. CFMFS for SW, EOS+Fuse+CERNbox will cove rmany use cases, discussions bout $HOME ongoing
- Telephony system moved to Asterisk
- Puppet on the desktop, master-less config
- Hadoop service in production(~100 servers), used by experiments for Analysis
- git.cern.ch will shutdown 3Q2016. move from gitolite to GitLab
-- gitlab-ci simpler than jenkins.
- Containers moving to production.
- Puppet: Ruby2 dropped compilation time from 73 to 25sec.
-- Puppet4/foreman1.11
- Centralizing Elasit SEarch service (ESaaS) to consolidate ~20 clusters.
- Rethinking backup system. May consolidate to backups on Castor (dropping TSM et al).

GSI Site report
- The cube finished on time, on budget
-- PUE will be better than 1.1 (12MW cooling capable), will report actual data after 1yr of operation.
-- 2 power plants, 2 independent lines, 2PDUs per rack. No UPS so better PUE.
-- Very careful cabling for proper airflow and pressure.

- Nyx using Lustre with multiple MDS, ZFS in production (1yr experience), more on future talk.
-- 15PB storage, 4Tbps I/O. Limited by OSTs disk speed.
-- They're funded by Intel for Lustre
- Director IB-switches with more than 5k nodes in one fabric with bisec bandwidth

KISTI site report
- Running GLUSTER
- Going towards container based provisioning system

KEK-CRC Site report
- increasing storage up to 180PB by 2024 due to belle II and j-parc
- 10K CPU in 6 racks, with SSD each. Mellanox 4xFDR IB (56gbps).
-- 13 PB Storage on DDN SFA1200 using IBMElastic Storage System (ESS)
-- 70 PB of tape on IBM TSM

IHEP
- Using gLuster (.7PB), Luster (5.7PB) and dCache(.5PB). Testing EOS as a possible future solution.
-- New disk is DELL MD 3860F: 4T * 60 / Disk Array
- Had some issues with Lustre: MDS hung unexpectedly
-- Fix IB network, reduce MDS load, fix SW bugs (on 2.5.3).
- Issues with gluster -- fixed NICs on servers

RAL Site report
- Running pre-emptable jobs on cores which otherwise would be idle (using HTCondor) - (should we do this within our own HPC @FNAL?)
- CVMFS

JLAB Site report
- talking about the USQCD ~$1M procurement
-- Intel XEon Phi or NVIDIA Pascal GPU or Intel Broadwell CPU. It all depend on benchmarking.
- Upgrading to 100Gbps either Omnipath or IB.
- 2016 procuement will require 12MB/s/TB
- Intending 3 Lustre pools
-- Faste with 6SSD in new HW
-- Production all RAID6 ro RAID z2
-- Work - smaller files, 3 slower systems.
- Details about the 4 2014 fileservers (some did not have lustre)
-- CERN is seeing similar issues with Intel 2680 processors too
-- BIOS, disk controller and zfs_arc_max upgrade (Supermicro just released new BIOS)
- Need to move IBM robots (again), IBM asks them to remove the tapes.
- Investigating CentOS7 with lustre 2.x client

Evolution of CERN Printing Services
- CERN golden rule: everyone at CERN can print (staff, visitors, conference participants, anyone!)
- 60-65K pages printed per working day (14k print jobs)
- New procurement led to ~30% savings.
- Neat 'add printer' web based system with a location map to help locate printers
- Support confidential printing (requires pin typed in before the paper goes out and print job gets removed after print).

DCD - Desktop Chromodynamics, or: Linux on DESY Desktops

- DESY provides support for linux desktops (but not laptops)
-- ~800 linux vs ~4k windows.
-- Zoo of linux destributions: Ubuntu & SL
- Now supporting Ubuntu in both managed (puppet) and unmanaged (user is root) fashion.
-- For managed systems the user groups still have the possibility to make changes themselves and apply them via the config management system (puppet), then they're still within the managed realm.

State of OpenAFS
- Project is still alive, no IPv6 support soon, clients falling behind
- Lack of manpower (users moving to auristor), lots of users have or are planning to drop soon

Running virtualized Hadoop, does it make sense?
- Why? Easy scaling up&down. Clusters optimized as per user needs (separate clusters per users).
-- Automation is key (makes many instances manageable)
- Fast CPU and memory are ok, also some SSD available
- Storage options
-- local drives
-- shared storage
-- external HDFS
- To sum up, performance is ok with properly tunned VMs, some benchmark is still work in progress.

The OSiRIS Project: Meeting the Multi-Institutional Data Collaboration Challenge
- Shows HW (all Dell) purchased for the project.
- 3 Michigan universities
- The idea is to benefit of economy-of-scale. Building a globus-online+CEPH+monitoring (checkmk OMD&)
-- Running it all on VMs on top of RHEV plus puppet to setup the services.
- Different reserach groups within each univerity, all using the shared resource
- using www.openproject.org to manage the project, happy with it.
- So far staging the users so that they don't use the infrastructure at the same time (not enough effort to support users)

Why so Sirius? Ceph backed storage at the RAL Tier-1

- Sirius (low-latency storage to STFC private cloud) & Echo (to replace disk-only castor @RAL)
- Echo is a bunch of ServiceX-to-CEPH wrappers, where services so far are XrootD, globus-gridftp-server
- Echo is yet another storage solution github.com/stfc/
- Demostrator by July 1st and authorization to go (or not) in production by October.

DESY site report

- F5 HW end of lifetime, investigatin A10 Hard and software.
- Belle II moving services to DESY IT (from KEK)
- Expanding HPC due to experiment needs (Petra-III)

BNL RACF Site report

- Unifying HPC and HTC operations (BNL interdisciplinary HPC Institutional Cluster)
- Deprecating Bluearc Titan 3200, replacing with GPFS which performs much better.
- ATLAS providing dCache NFSv4.1 for Tier3 use with 2.14

- ~1PB of CEPH, including cephfs
-- CEPH used by the ATALS Event Service
- Test ELK framework, primarily to monitor dCache billing logs.
- Using puppet 3.8, planning upgrade to 4.
-- Automated Jenkins testing in place to make sure changes work
- AWS/EC2 test with 45000 slots using spot pricing succesful.
- Evaluating Slurm (15.08) for HPC because Condor is not great for HPC

US ATLAST SWT2 report (Oklahoma)

- 10368 preemptable cores comming in, will be used 2h a day for weather forecast so there is high potential for opportunistic slots.

PDSF (NERSC) Site report - James Botts (Computational Systems Group)

- Using SLURM, happy with it.
-- Thinking to move their HTC from Univa's Grid Engine to SLURM too.
- Biggest 'trouble maker' is Lustre (MDS failover not working, failing OSS)
- Moving tape on 2020 due to dust issues on the new site.
- Using RDMA vs IPoIB to access GPFS improved x3 bandwidth
- ~1PB of storage deployed for ALICE xROOTd storage - first EOS at NERSC
-- 1/2 rack 3 netapp e5560. 3 FSTs, 1 MGM/MQ server
-- Now regret not having gone with JBOD
-- 'complains' that the EOS documentation is somehow scattered around
- Greater use of container technologies (shifter), includes CVMFS
- cfengine -> ansible for config. management. The CRAY uses ansible

CERN Cloud Status Update

- 5800 hypervisors in production, 155k cores, 18K VMs.
-- Every 10s a VM gets created or deleted on CERN's cloud
-- Biggest performance difference was due to NUMA misalignment on the VMs
-- New hypervisors all have SSDs. Had to introduce new VM flavors with smaller disks.
-- Using dm-cache in front of cinder on CEPH to have fast remote disks.
- Reported some KVM/CEPH issues
- Future plans to Investigate Ironic (baremetal provisioning for openstack) and replace hiper-v by qemu/kvm

Deploying service with Mesos at RAL

- Apache Mesos is a UC Berkeley developed batch scheduler
- Have a loadbalancer and floating service IPs
- Run services on docker containers and then use Mesos to start/stop those images as needed (eg if machine crashes, starts on another server).
-- Autoscaling is much faster with containers than VMs

SCD Cloud at STFC
- Migrating off OpenNebula to OpenStack
-- There is an interesting Gap Analysis in case we want to do the same at FNAL.
- The architecture they're running is the same CERN was running a while back and CERN recommends changing it.

Recent work of the HEPiX IPv6 Working Group
- CERN still has some IPv4
-- ATLAS pushing to get IPv6 support
- Targeting 2017 to have IPv6 only WN.
- New sites (ASIA) have far more restrictions with IPv4 and will require IPv6 sooner

perfSONAR Status in WLCG/OSG
- Some issues get magnified with increased RTTs
-- A single packet lost out of 22K can heavily impact WAN bandwidth (examples in presentation, going down to 60Mbps)
-- packet reordering and jitter cause a similar issues
- Perfsonar makes measurements for both to ensure there are no issues on the network.

VPN over SDN - A virtual private network based on software-defined network architecture for high energy physics scientific data exchange (by IHEP)

- IPv4 bandwidth is limited while IPv6 is not.
- IPv4overIPv6 reaches ~95% efficiency

Computer Security update

- Angler remains the most successful Exploit Kit attack.
- 88% of the rootkit attackers comes from malvertising (for example affecting 20min.ch, Guardian and Lenovo websites)
- Ransomeware (encryption of your files) as a service exists
- OS X affected too

A Network Security self-service platform (NSSP) of IHEP

- This is a 'self scanner' service, marketed as 'security as a service'.
- developed in house by students

Identitiy Federation for HEP (AARC)
- EU funded project within EU HORIZON2020: https://aarc-project.eu/
- Want to leverage on eduGAIN+VOMS for authorization and then move towards x509 free auth.

Future HEPIX meetings
- October 17th 2016 @LBNL
- April 24th 2017 @Wigner Budapest
- October 16th 2017 @KEK

Lustre/ZFS Development @ GSI
- Interesting talk to HPC & DMS (anyone running ZFS): https://indico.cern.ch/event/466991/contributions/1143595/attachments/1260659/1863031/LustreZFS.pdf
- GSI is getting funding from Intel as an Intel Parallel Computing Center: software.intel.com/de-de/ipcc
- Using ZFS RAIDZ2 with no RAID controllers on the Lustre OSTs
-- They've scripted global spares/resilvering.

- Worked on new algorithms to RAID on Lustre.
-- The graphs are testing the code performance on RAM, no disks involved.
-- No production experience on this yet, still working on the code, but intend to contribute the code.
-- Production results presented on the next HEPIX
-- presented on LUG last week (Portland)
- LBUG automated analysis
-- Markov model
-- Helped locate and fix bugs, now LBUG free since 2013
- Working to integrate TSM with Lustre HSM capabilities
- More details on the ldiskfs on the MDTs due to ZFS performance issues. Did you try ZFS on SSD?
-- Chief ZFS developer is working into this
-- They tried on SSDs and the MDT performance was much lower on ZFS (test was mid 2014)
- Running 2.5.3 Lustre, working with Intel

Storage Chamaleons by Xavier Espinal (CERN)
- Castor evolved to a tape oriented system focusing an a per stream speed
- moved Castor disk caches form RAID1 to RAID60 (100MB/s to >350MB/s per stream)
- Evaluating a common disk layer (multiple experiments)
- Tape policies and per experiment/user/group resources
- EOS namespace still fits on memory but worried about it.
-- 40 minutes to boot bigger instances
- 6 Instances out of CERN (Fermi, Russia-T1, EsNET, Aarnet, JRC, Univ. Vienna, INFN Trieste, IHEP)
- Productizing it with Openlab/COMTRADE
- Working on an embedded Web browser ROOT viewer.
- CEPH 30PB, 40K OSDs
- Multi-site production with 3PB@wigner + 1PB@meyrin
-- Lots of CERN code development for CEPH
- evaluating RADOSfs for Castor backend.
- Tape is still very much in the roadmap for CERN as it's the most cost efficient solution (and keeps improving with time)
- With the goal to make data access easy to the experiments Storage@CERN is thinking to MOUNT everything. Nice slide showing it, EOS being the center piece:
-- /cernbox for users' data, and papers, and
-- /eos/$VO for experiments' data
-- /cvmfs/$VO for experiment's SW
- One of the targers is to allow new data management and analysis models, LHC@myPC with sync&Share
- To support this, CERN has been focusing on FUSE for EOS since March 2016
-- Note that this matches with CERN fixing the bugs FNAL filled ~1yr ago.
- Data corruption case on EOS
-- There was a broken network equipment, that corrupted TCP packets and caused retransmission and some TCP checksum collisions
-- ADLER32 was correct although the file was corrupted
- There is a 0.22% silent corruption rate detected with production ATLAS data
-- More details on the data corruption at cern.ch/go/wr8j
- Could not reproduce yet, but there is a complete match between the network issue and the data corruption windows.

Mass Storage at CC-IN2P3

- Storage on: dCache, xRootd, iRods and Direct Access
- They're implementing a SW to work with HPSS for queue ordering for tape reads, user fair share, etc.

ASAP3: Status update and activities for XFEL
- SPEED project (DESY & IBM collaboration), to setup a new system to improve storage system capacities.
-- Solution based on IBM Spectum Scale and ESS
- Using a Maxwell GPU + IB FDR fabric with 2:1 blocking (192 ports) and a wire speed fabric (36 ports)
-- 3 top and 8 leaf IB switches
- Testing ZeroMQ to move data from detector (POSIX I/O, detector specific) to a GPFS system.
- Most Operational Issues appeared on GPCF and Infiniband
-- Issues with Mellanox OFED on ppc64
-- Firmware bugs on the IB HCAs which caused data corruption
- XFEL dataflow for online/offline shown
-- Planning to use dCache's pNFSv4.1 to access offline data

IBM Spectrum Scale support for technical workflows

- GPFS with multiple protocol support: POSIX NFS/SMB, HDFS, SWIFT, S3
- Automated rule based policies for multiple layer storage (flash/ssd/hdd/tape)
- This is a mature product that started development ~1993
- Future roadmap includes
-- Cloud storage integration (eg: move data from cloud to local with rules)
-- integration with dCache (not sure how serious this is)
- Transparent cloud triggering is freely available to try out, link in the talk.
- There will be a talk in Chicago on June.

Update from Database Services @CERN
- Basically an Oracle talk.
- In memory DB x5-10 performance improvement on certain queries. Transparent deployment for Oracle.

Batch Computing at DESY

- HPC cluster 4592 instances (slashed in 2)
- 117 TB BeeGFS scratch clusterfilesystem and GPFS for physics data and $HOME
- Good slide that shows automation/complexity on HTC_grid vs HPC https://indico.cern.ch/event/466991/contributions/1143623/attachments/1260821/1863379/htcondor_hepix_2016.pdf
- The plan for Batch Computing is:
-- SLURM for HPC. Condor lacks MPI features.
-- HTCondor for Grid (replacing SoGE, Torque+mySched+CreamCE). Slurm is not happy with thousands of single core jobs.

xBatch: Extending the CERN Batch Service Into the Public Cloud

- goal is to Deploy Condor worker nodes in external clouds. Transparent to the user, same entry point
- Tested IBM SoftLayer

- Testing several provisioning mechanism (including HTCondor instantiating VMs), config management done with puppet (distributed puppet masters with centralized GIT repos).
- Accounting is specially important because can't do benchmarking of the HW beforehand and high variance is expected.
- Terraform looks promising for orchestration

Computing and Storage for Life Science at MDC (Alf Wachsmann CIO)

- They have nasty data patterns (random access with small blocks on 3TB datasets)
-- NFS does not scale, use GPFS
- Planning big ID Management to 'unify' departments within hospitals plus laboratory IDs (eg: rats, bugs, etc). Planning to use SAP HANA for 'on memory DB analytics', or IBM Watson or similar

CPU Benchmarking at GridKa (Update April 2016)

- Found wide differences between VO benchmarks and HS06 (over 45% differences when comparing different CPU generations)
-- Differences could be explained by GCC version used and optimization levels, being the older version used by HS06 and not being able to use most recent CPU improvements.
-- Lacks CMS information on benchmarking, maybe someone from CMS should contact Mandred Alef (KIT) to let him know what CMS uses (or could use) to benchmark.

Computing on Low Power Architectures (Michele Michelotto)
- Peter Elmer (CMS, Princeton University) tried some low power cExynos4412 prime with very good W/event processing numbers
- Annual shipment of ARM grew linearly since 2005 while x86 growth is almost flat.
- Atom C2750 is a lot more power efficient than any other solution out there (HS06/Watt), also better than AMD A1100 (newly made ARM based server). Note Knights Landing is missing.
- https://indico.cern.ch/event/466991/contributions/1143635/attachments/1260937/1863715/hepix-desy-zeuthen-Michelotto.pdf

First Experiences with Container Orchestration in the CERN Cloud
- Orchestration with Docker Swarm (clustering)/Compose (orchestration)
-- uses Docker API
-- Has scheduler filters (eg. run on SSD nodes) and labels (eg production/dev)
-- With compose one can define the whole application in a single file.
- Kubernetes (by Google)
-- more complete, ha auto-scaling
- OpenStack Magnum
-- Deploys container engine cluster (swarm, kubernetes, mesos) using heat
- CERN used Docker Swarm for
-- CI in GitLab, transparently to GitLab
-- FTS, to scale service by service
- Now testing integration of CVMFS so that many containers can share the same CVMFS
- 1h presentation available at cern, see link in talk.

Using Docker container virtualization in DESY HPC environment

- For each job they want to create a cluster of Docker containers to isolate the user from the system
- This way users have root privileges from their jobs. This could be good for the Wilson Cluster to allow users to manage their own SW.

- Since RH7.2 --enable-user-namespace=1 prevents user to be root on the host system
-- Using default bridge network (IPoIB)
-- Sharing parallel filesystem from host machine in the container. Works with nfs, gpfs, beegfs (after patching)

The HELIX NEBULA Science Cloud project (Heldge Meinhart @CERN)

- 10 year horizon from LHC Run3&4 means >x50 growth in computing and storage needs
- Technology evolution with a flat budget is not enough
- Same trend for other particle phyiics needs and also outside of physics too
- Improvement on the code (better algorithms and HW optimized compilation), but also need more resources
- CERN-IT evaluation of Microsoft Azure cloud IaaS
-- Wayne Salter went visit the data centers en EU
- Current procurement mechanisms and cloud missmatch
- HNSciCloud Joint Pre-Commercial Procurement
-- Not HEP specific
-- Total procurement commitment >5M euro (1.6M from institutes like CERN, IFAE, etc)
- Project timeline is January'16 - June'18
-- preparing tender now, will launch in 7 months.
- Targetting caching at provider's site. They want to do data heavy calculations (not only MC!).
-- Network via GEANT, authentication via eduGAIN

Virtual Cluster Computing in IHEPCloud
- 60% resource utilization because they're not shared across experiments and provisioned for peak computing

Using Ganeti for running highly available virtualized services (NDGF)
- a SW stack for managing VMs (like vmware, openstack, libvirt, etc)
-- From google for it's internal use
- outside google on Debian, ndgf-t1, lufthansa, etc.
- It's good at running HA on modest HW.
- Cluster size 3-50
-- multiple clusters integrate well
- Native support for CEPH RBD
- NDGF using on 2 HW nodes with 10Gbps direct connections and running the dCache headnodes

Computing and Software for data-intensive physics (Yves Kemp @DESY)

- Proposing to creating an Academic Journal, peer reviewed.
- There is general support to do it for physics instead of science.
- Claims CHEP and Advanced Computing and Analysis Techniques in physics research (ACAT) are not good enough (either not frequent enough or focused).

Authorization extension for the secure use of ElasticSearch and Kibana
- Technical talk showing how to add user authentication and access control to ELK with LDAP and Kerberos5, looks good!

Automating operational procedures with Rundeck (Daniel Fernandez from CERN)
- CERN has shown this in the past, seems to be production ready.

- List of plugins at http://rundeck.org/plugins/. There is a puppet module to deploy it github.com/voxpupuli/puppet-rundeck
- It keeps credentials secure.
- You can write your jobs on whichever scripting language you want
- CERN runs rundeck from within Tomcat to be able to do the Single Sign On.
- The product is open source and one can get commercial support if need to
- This could be very useful for SCD groups to interoperate on each other systems while keeping a well defined role distinction.
-- EG: create/remove a VM from the system (create VM, add it to monitoring, etc.). Sysadmins can execute, 'cloud managers' can execute and modify
-- Other examples: rename phyisical host, enable/disable compute host, HW intervention on a Compute Node. Good detailed example that perhaps the DMS-ECF interactions could benefit of.

CHEF @GSI revisited
- Cultural SVN vs GIT issues in the beginning
- https://github.com/GSI-HPC/sys-chef-cookbook
- This talk does not make CHEF appealing at all.

Grid Computing System in the KEK Central Computer System
- Rental system. Every 4-5 years they launch a procurement that includes HW and operations personnel.
- Most of their usage is Belle II computing (grid jobs on a supercomputer)
- They have WLCG-FTS,

Server gerontology in LHCb

- Running large amount of online resources (+2K servers)
- Have plenty of space (in the experiment pit)
- Babysitting old HW to have it up:
-- Still running Dell 1950 in production. Replacing burned capacitors on RAID cards ($500 to replace on all systems vs $400 a piece to replace card). They've counted the cost per hour across time (0.28 for it's warranty live period, now 0.11)
-- Supermicro

Monitoring at LHCb: Migrating to Icinga2,Puppet, Hiera and Foreman Stack for Monitoring.
- boot from net most of the systems.

Monitoring at scale: a needle in the haystack (@IN2P3)
- Aggregates all system logs
- Analyze all system logs with ELK & websocket+reimann-dash
- Looks nice&powerful

A slice of the NERSC data collect system
- Issues with old facilities having propertary (industrial) control systems. Turned out to be very limited (limited sensors, limited capabilities, and even SW vulnerabilities!)
- New facility has many many sensors, collectors and networking
-- 1600+ temperature sensors
-- 800+ power meters
-- 600+ power strips (raritan, 50/60)
-- others like sismometers
- Store all data forever (HPSS)

- Many requirements (no plugins like java or flash etc, mobile support)
-- They build their own monitoring system that reads and stores all those sensors' data.
- Rack instrumentation is ~$200 per rack.

More NERSC Data Collect (Cary Whitney)
- Pushing vendors to improve data collection
- NERSC, Oak Ridge, Sandia, Los Alamos and NCSA working together for data collection organization
- DDN has a collectd module for GPFS.
-- CSCS Lugano is helping write a gmond collection for GPFS
- Cray/Sonexion data gather. Progress report in 2 weeks at CUG (Sonexion is Seagate's Lustre product)
- Thinking about what it means to have 90% full of a 1PB filesystem...

Consolidating Scientific Computing services at BNL (Tony Wong)
- Reorganized from a de-centralized computing services (originated by the funding model) towards the bnl.gov/compsci
-- RACF taking many more duties, will be served by the 'Institutional Cluster'
- The BNL IC
-- 20 kW/rack
-- Will try to have HTCondor and Slurm co-exist on the IC
-- Possible back-fill by RHIC, USATLAS, OSG, etc. Will do study this summer. Tony going to CondorWeek to investigate more on this.
- Got CD-0 for a DOE-funded Core Facility Revitalization program
-- 10MW, 32000 ft2, 5K ft2 for offices
-- PUE must be 1.2-1.4

DESY Hamburg infrastructure
- Now near max capacity for power and cooling, but need to increase 1.5MW for new projects in 1 year.
- Building a cold water ring to cool their 3 datacenters.

GSI GreenITCube
- 16 million euro for a 27*30*22 meter building, with power and cooling lines included (11.5 just building). Racks not included
- 6 floors with 128 racks each (~16kW per rack right now). Cooling via passive rear-door water good up to 35kW.
- built from December 2014 to January 2016. First cluster running on March 2016
-- Paperwork & planning took longer than construction.
- Tested the room with cheap fan heaters, up to 2.5MW and the only issue was that the fans got too hot (cheap).
-- PUE under 1.1 during all test
- Nice home grown facility monitoring (whole building&systems, per rack).
-- Using Ganglia to collect data

ForHLR - New Energy Efficient HPC System at KIT with Warm Water Cooling - KIT HPC

- FDR HCAs (56Gbps), EDR switches (100Gbps)
-- Infiniband Mellanox MetroX 6240
-- 35km IB WAN interconnects
- DDN ES7K Storage, running luster 2.5 (IB interconnected)

- HPC system bought with the water cooled racks, racked already (1 ton each!)

Workshop wrap-up
- https://indico.cern.ch/event/466991/contributions/1143581/attachments/1262244/1866326/2016-04-22-Wrapup.pdf
- Good to see trend of topics on tracks
-- Lustre
-- Containers
- Highlights ?requirement? for all T1 WLCG sites to be at the IPv6 WG

AOB
- Should we deploy http://xdmod.sourceforge.net/ to monitor our HPC clusters?