BatchComputing @ DESY

Christoph Beyer with slides from Thomas Finnern, Stefan Dietrich
HTCondor user meeting
Barcelona, 2016-01-03
IT activities

> Operation of a 'Large Scale Facility' Tier2 and Analysis Center for LHC
> Operation of Tier0 and Tier1 for HERA, IceCube, BELLE 2
> Operation of Tier1 for Cherenkov Telescope Array (CTA)
> Preparations for a future linear collider project (ILC)
many new collaborations and fields of study planned and installed

> CFEL
Max Planck, Univ. Hamburg, DESY

> CSSB
center for structural system biology

> ZOQ
center for optical quantum technology

> EXFEL, Nano, ...

many new buildings for photon science efforts

> Ongoing construction work until at least 2018

> Many more scientists staying at DESY, using DESY infrastructures
Zeuthen

> deploying Univa and started to use UGE 8.2

main motivation is cgroup support & some bug fixes

Hamburg

> Grid

- currently running Torque, with a home-build scheduler

> local batch farm (BIRD/NAF)

- running SoGE, seeing performance and stability issues

> HPC cluster

- calendar based reservation tool - working well, but will not scale for future

- SLURM used for some resources since 2016
Batch Systems – Hamburg site some numbers

> Grid
  - node=280
  - slots=10.054
  - cores=12.196
  - 12.26 HEP-SPEC06 / core

> local batch farm (BIRD/NAF)
  - 577 nodes
  - 7.340 cores

> HPC cluster
  - 3.000 Cores (AMD Bulldozer)
  - 117 TB BeeGFS Clusterfilesystem
Component Topology

Automation

HPC Maxwell
IB
HPC Maxwell

BIRD NAF

BIRD NAF

GRID

GRID

BeeG FS

GPFS

GPU

AFS

NFS

CVM FS
dCache

Complexity

Calendar Reservation

Petra III Data Analysis

Son Of Grid Engine SoGE

Torque +MySched

Cream CE

Batch Infrastructure Resource at DESY

National Analysis Facility

People

Scaling

Support

Robots
We have a Plan!

Now

Test

Calendar Reservation Tool

Son Of Grid Engine SoGE

Torque + MySched

Cream CE

Future

SLURM

HTCondor

ARC CE

HPC Maxwell

BeeGFS

GPFS

BIRD NAF

AFS

NFS

GRID

CVMFS

dCache

- Scaling
- Comply to future HPC

- Scaling and Support
- Common Support Team

- Support
HTCondor tests & preproduction

➢ First setup on a single host - 'Hello World Job' running after 2 hours

➢ Small test cluster 1 master/submit host & 7 workernodes DESY GRID testjobs running fine

➢ Bigger testcluster 800 Cores, 2 master, 1 ARC CE running since oct 2015

➢ BELLE2 jobs (single core monte carlo production) running since nov 2015 turning test- in a preproduction-cluster

➢ Jan 2016 ATLAS, CMS ILC and BELLE2 jobs running concurrently
Setup and experiences

> Very much a relaunch of the work that Andrew did for RAL really !!!
> HT enabled with partitionable/dynamic slots
> Hierarchical accounting groups (%quota)
> $CondorVersion: 8.5.1 (installation/node management through puppet)
> Healthcheck script running as a STARTD_CRON
> Horizontal filling with 'NEGOTIATOR_POST_JOB_RANK'
> Prefere multicor jobs by resorting the grouplist using 'GROUP_SORT_EXPR'
> Monitoring started with simple ganglia plots using 'condor_q' and rrd
> Switched to Grafana (thanks again → Andrew) using python bindings and INFLUXDB (will try to keep data in CSV as a backup)
> HTCondor prooved to handle all the weird misconfigurations we produced on the way without loosing jobs, stability or the central funtionality
Setup and experiences → todos

> Still much work todo in terms of understanding and get to know better
> Multicore job preference still an issue have to think about DEFRAG
> Individual monitoring/dashboards (also for enduser/groups)
> Documentation
> Do more sandboxing with CGROUPS MOUNT_UNDER_SCRATCH etc.
> Job transfer in commercial clouds will most likely become an issue as there is some funding for that purpose expected
> Waiting for AFS/KRB integration to start a testbed for local batch
> Many more individual challenges to be expected then
> Openstack will become more important at DESY most likely
> Integration in the control room monitoring and operational day-to-day business
Conclusions, perspective & questions

> GRID cluster hopefully will be completely migrated to HTCondor during this year

> Once AFS integration is available testing with 'real' users will take some time but will try to migrate local batch in due course

> Looking at the degree of capacity utilisation of the HPC installation closely and consider to run HTCondor as a 'parasitic' second scheduler during idle times

> No educational work had to be done so far (this will change with a local HTCondor batch facility)

> We are happy about the current momentum in the HEP society involving HTCondor and the open minded support by both colleagues and the HTCondor team

> HTCondor gives us the feeling of standing on the shoulder of giants in many ways – thanks for that!

> PS: We are not missing the 'queue-model' as much as we thought

> Questions?