NAF Users Meeting Status and Outlook

- Status and prospects of the NAF
- Status of the migration of the batch system to HTCondor
- How to access the NAF from remote, and using graphical tools?
- User feedback and time for discussion

Yves Kemp et al., DESY IT
Hamburg, 28.11.2017

PHYSICS AT THE TERA SCALE
Helmholtz Alliance

A two-slide NAF introduction

... Just in case ...

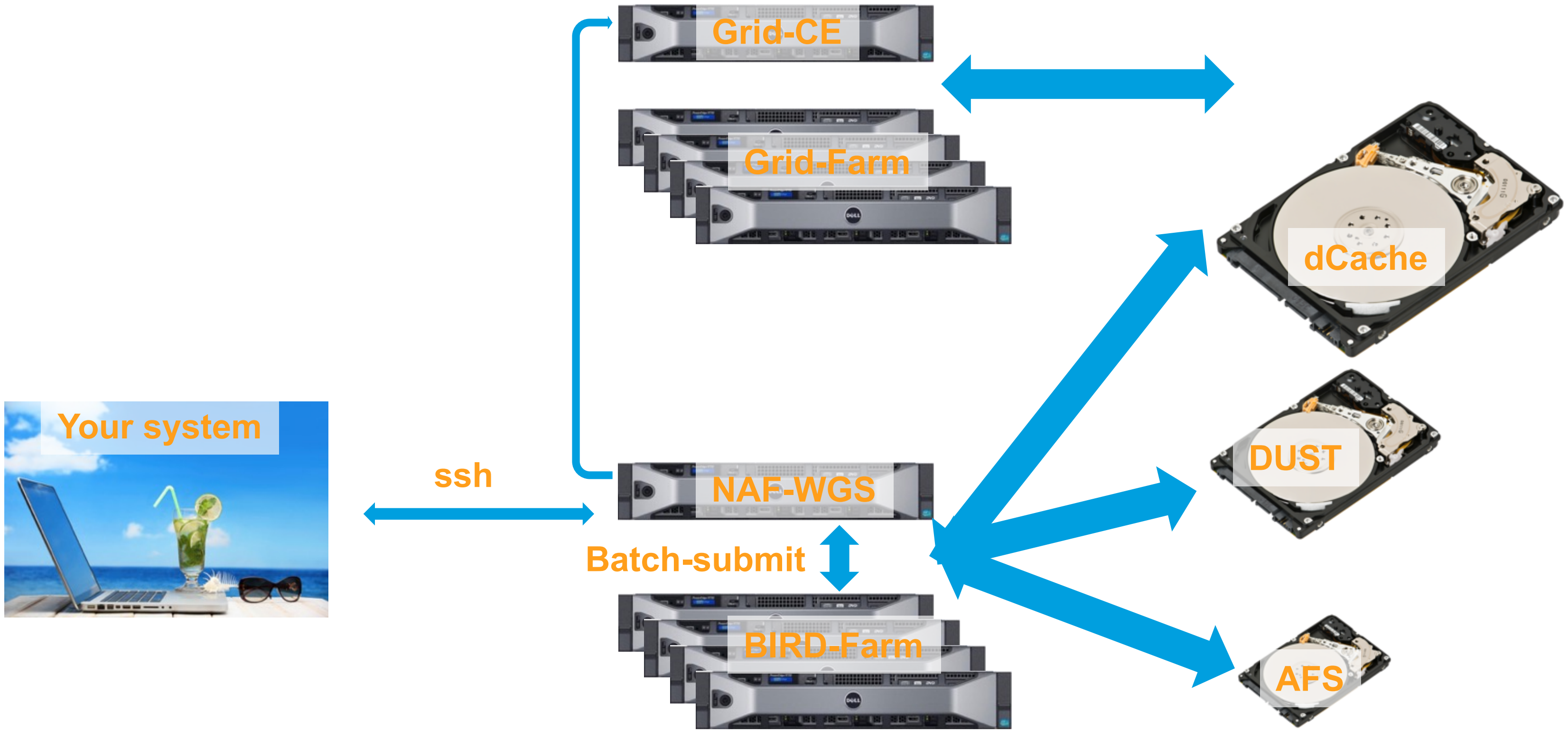
What is the NAF?

- NAF stands for „National Analysis Facility“
  - *National* means: For people working in institutes in the Terascale Alliance
  - *Analysis* means: Analysing data taken in the Terascale Alliance
  - *Facility* means: Something where you can do real work
- Basically: The NAF is a facility where **YOU** can do your analysis (and stuff around your analysis)

- The current NAF is really „NAF 2.0“ ... Some of you may remember the original NAF, which was finally decommissioned in 2014
- NAF 2.0 is much simpler in that:
  - Only one site, DESY Hamburg
  - No separate „NAF account“
  - Login with normal passwort, instead of X509 certificates

- The NAF comprises
  - Dedicated work-group-servers for login, to do interactive work, testing and development
  - A large batch cluster: currently around 7000 CPU cores for the NAF
    - Shared among ATLAS, CMS, ILC, BELLE, and legacy HERA
  - Additional dCache Grid storage (>5 PB in addition)
  - A dedicated fast file system for scratch purpose, called **DUST**, with ~2.6 PB capacity

You – and the LHC compute & storage at DESY



NAF: Storage Status

Status of the NAF : Storage / DUST

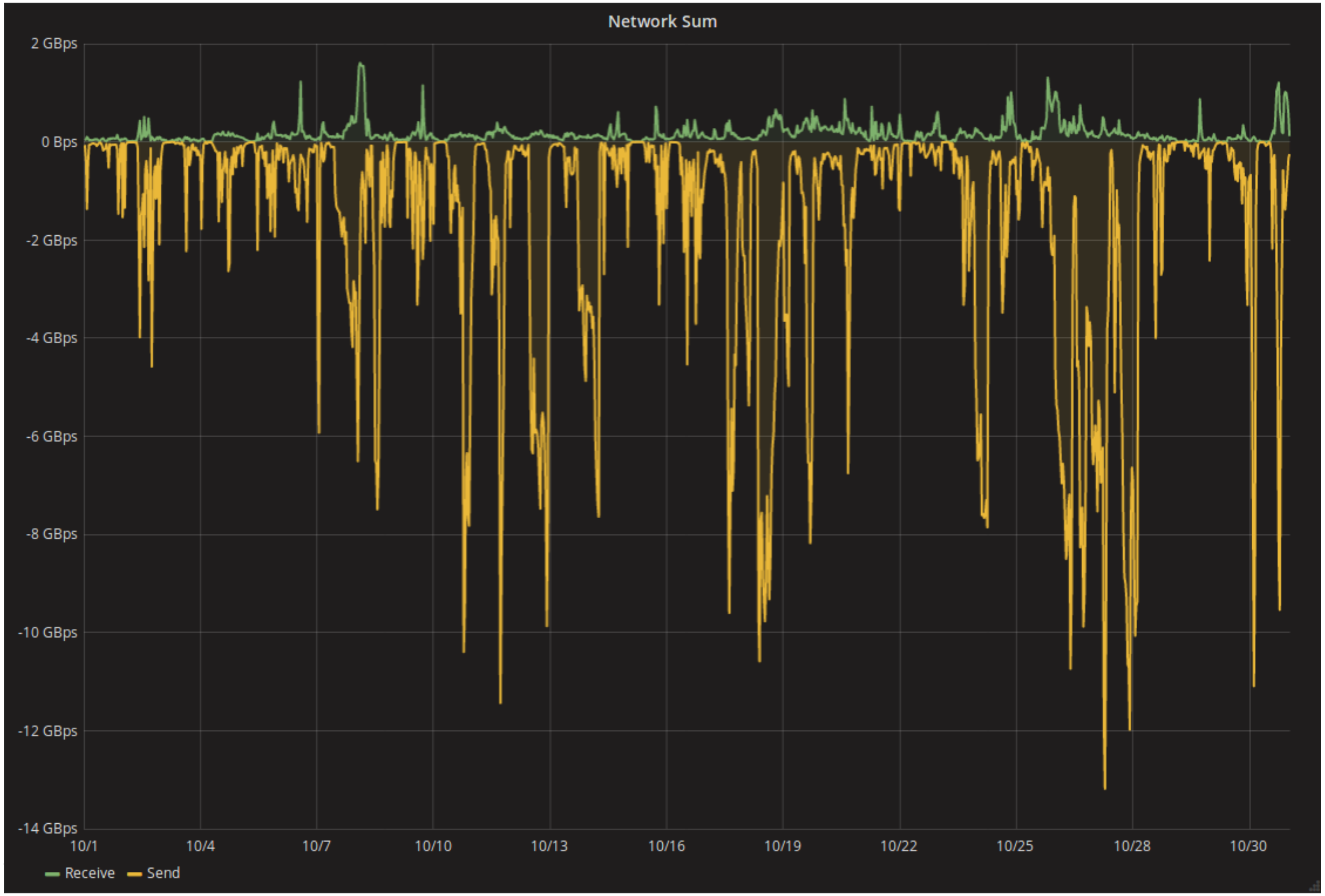
- Since early 2017, all experiments are migrated to DUST
- SONAS (the predecessor) shutdown on 1.2.2017
- Some initial troubles to get DUST stable
  - Close cooperation with the vendor IBM
  - Involvement of developers@IBM
  - Some issues remain, but very rarely seen – and beeing worked on!
  - Please report issues when you encounter them ... Even when we cannot do much!
- Current status:
  - DUST more stable than SONAS
  - DUST has much better performance

- DUST: Increase of capacity to 2.6 PB
  - And also increase meta-data capicity... Users have too many small files
- Reminder: Life-Cycle policy in effect
  - Data ”hidden” on account expiration day, deleted after 6 month if the account is not reactivated.
- Rewrite the “Unused-Data-Monitor”, first beta available at http://www.desy.de/~hannappj/dustUsage/index.html

IBM
Spectrum
Scale

DUST performance

Reading (orange) and writing (green) during October 2017

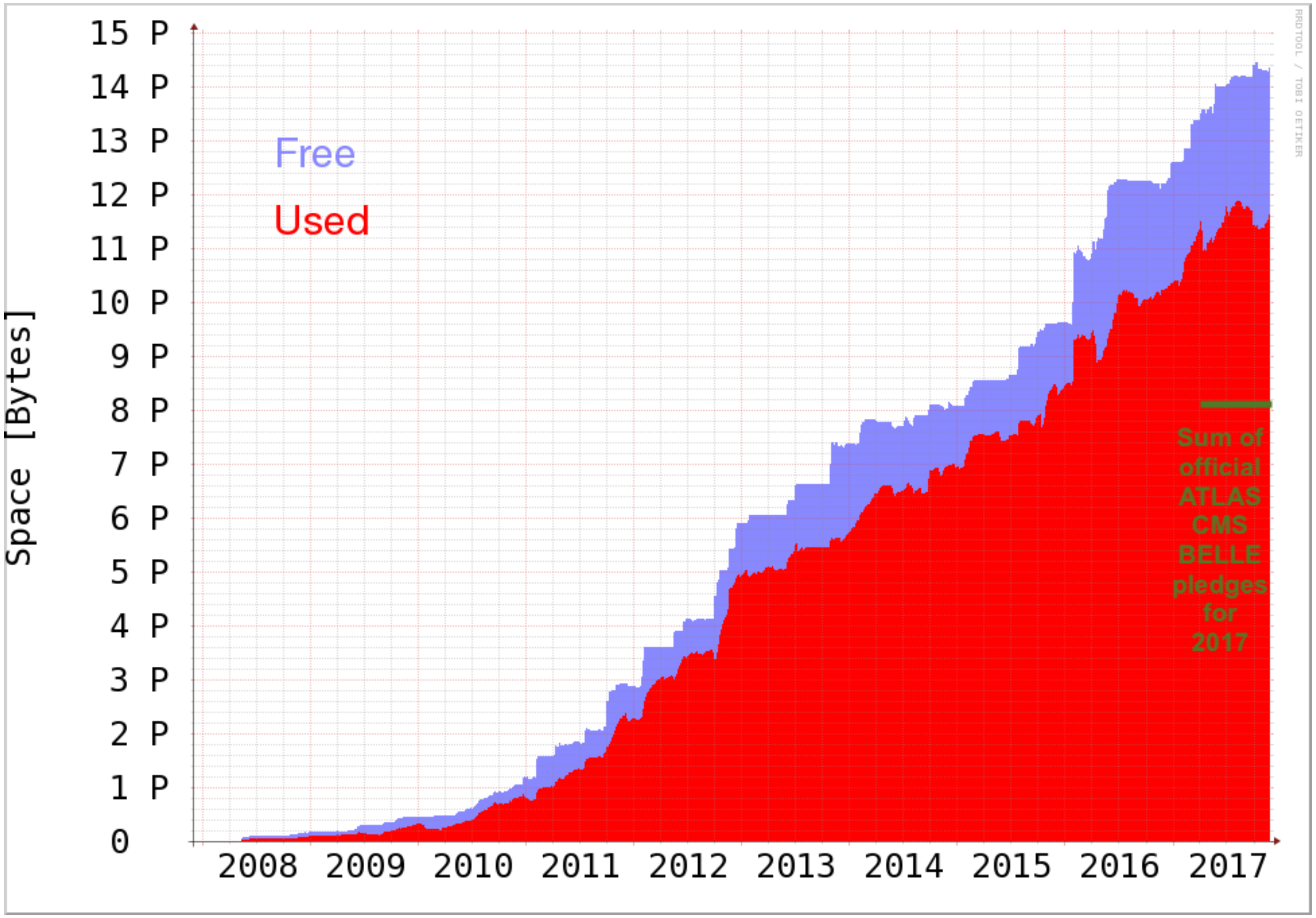


Status of the NAF : Storage / dCache

- dCache: keep going and keep growing
- Large purchases, some used to replace 7y old hardware

- No shortage observed in 2017
- Central data management works well
- User data cleanup works well (from our point of view)

- To show off: Total ATLAS, CMS, ILC & Belle dCache evolution over the past 10 years



Status of the NAF : Storage / AFS

AFS @ DESY

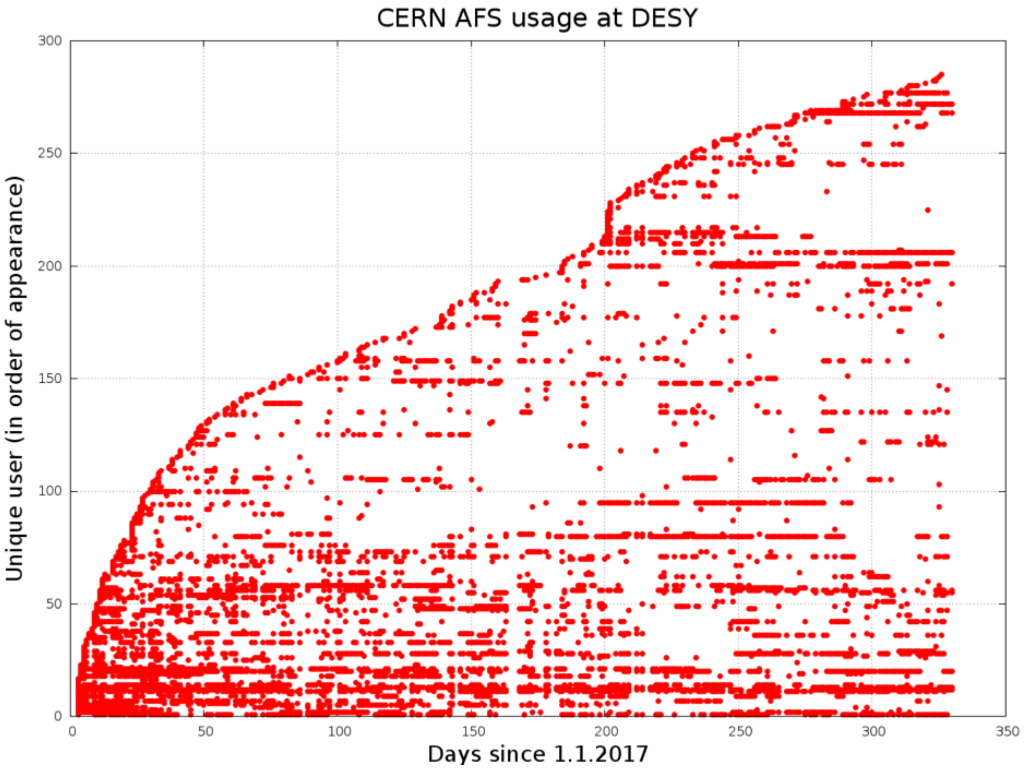
- The OpenAFS project has shown problems to adapt modules to new Kernel versions
- Has led to problems with Ubuntu Linux, not so much with SL/CentOS ... But still
- Thoughts going on to replace OpenAFS by something
  - Different product(s)
  - Different AFS implementation
- Current status:
  - OpenAFS project is „more alive“
  - Difficult to find replacement products
  - DESY evaluating different AFS implementation
  - AFS will most likely stay at DESY for some more time

Status of the NAF : Storage / AFS

AFS @ CERN ... Used at DESY

- Same boundary conditions at CERN
- Different conclusion: AFS shutdown 2019, migration ongoing
- One „external disconnection test“ took place
- DESY is actively scanning usage of CERN AFS, and informs users by email
- Usage is not really decreasing

- Alternatives: (in a nutshell)
  - Software: Use CVMFS
  - Own software: Use code management tools (git, ...)
  - Larger files / data: Use the Grid
  - CernBox: Currently no idea from CERN on how to access from remote compute clusters, no real option
  - http://linux.desy.de/linux__desy_for_users/access_to__afs_cernch/



Batch, compute and OS

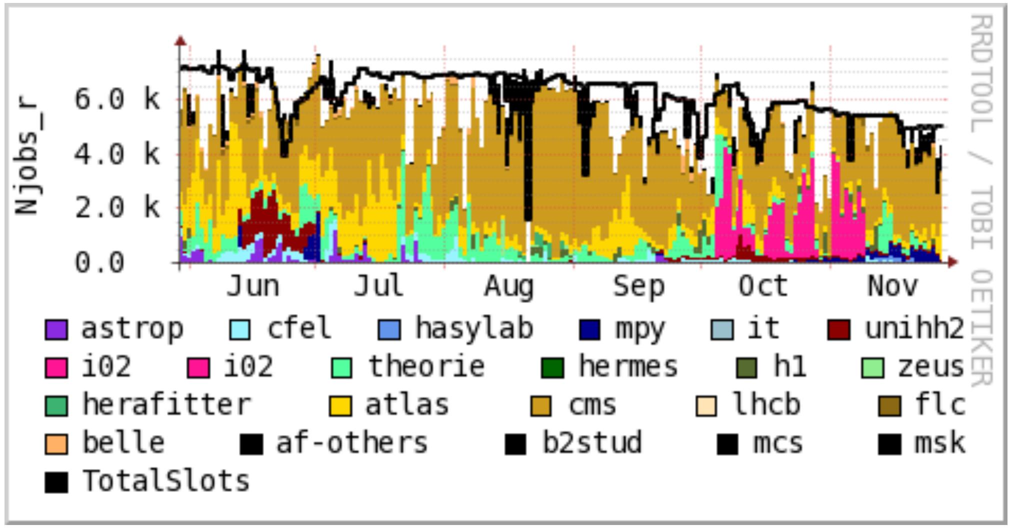
Migration to HTCondor

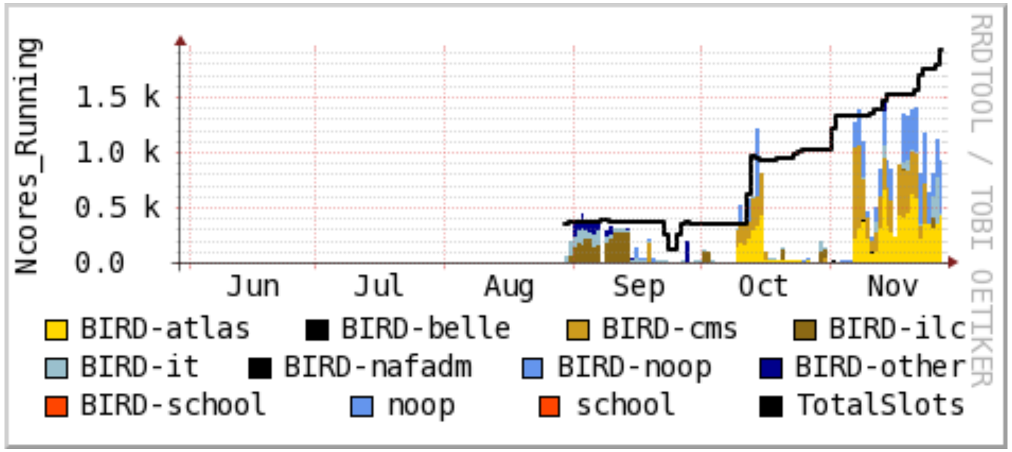
BIRD running SGE

- Going from ~7000 CPU cores to ~5000
- Main production workhorse
- Unstable from time to time, needs a lot of babysitting
- Want to decommission 1.4.2018

BIRD HTCondor pilot system

- HTCondor developers fixing AFS and Kerberos ~September
- October: Substantial amount of resources migrated
- Currently 1750 CPU cores, and increasing
- Very little used. The colors are mainly IT jobs, running under false flag. Real user jobs so far ~1.5 days in total.





Migration to HTCondor

- We IT are confident that we prepared everything for a smooth and stable operation
  - Experience tells us, that some problems only show up when scaling up
- First user tests found some bugs, that we fixed
- No major user problems reported so far
- However, we really would like a slowramp-up, and therefore, need some real users doing real work and putting real pressure on the system

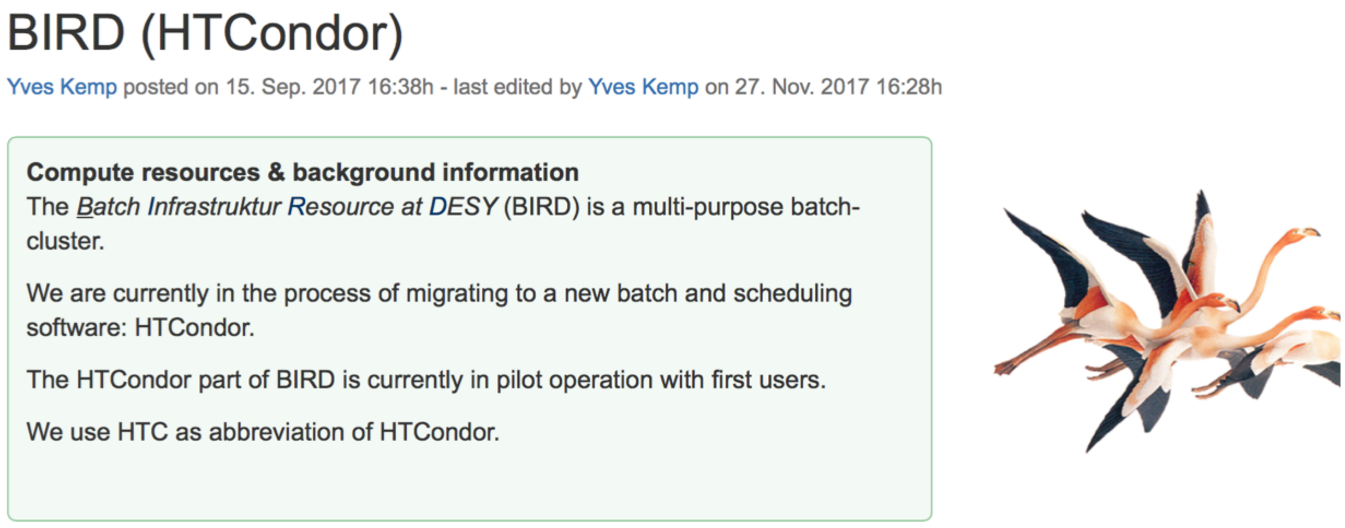
- Hardware and forecoming events:
  - 60 new WNs will arrive soon, put to HTCondor
  - Will replace old hardware
  - On 13.12. partial power-off (repairs), affects ~40 nodes in SGE and ~20 nodes in HTCondor. Convert the 40 nodes to HTCondor nodes
  - Total: Will move another ~1500 cores to HTCondor

- **Proposal**: Perform these migrations, and then wait until real user load in HTCondor starts.
  - Roughly 50/50 repartition
  - Take the time it takes
- Further migration as HTCondor gets used

- ... Or should we force the 80% HTCondor on 1.1.2018?

Migration to HTCondor

- Entry point for documentation and so on:
- http://bird.desy.de/htc/
- Entry point for support is (still) a mailing list, will be changed when more users come



BIRD (HTCondor)

Yves Kemp posted on 15. Sep. 2017 16:38h - last edited by Yves Kemp on 27. Nov. 2017 16:28h

Compute resources & background information
The *Batch Infrastruktur Resource at DESY* (BIRD) is a multi-purpose batch-cluster.

We are currently in the process of migrating to a new batch and scheduling software: HTCondor.

The HTCondor part of BIRD is currently in pilot operation with first users.

We use HTC as abbreviation of HTCondor.

HTCondor Future plans

- Once (!) we go into production, and SGE BIRD is shut down!
- Container technologies: Current idea:
  - Docker only for sys-admins (OS flavours: SL6, EL7, Ubuntu 16.04). HTCondor will not act as Docker-Container-Deployment tool!
  - No user-Docker in „OS-Docker“ possible
  - Singularity (or rkt) for user containers
- Merge BIRD HTCondor farm with Grid HTCondor farm
- Benefits for:
  - IT: Easier management
  - Users: Larger cluster with more “entropy“
- We plan for a transparent migration

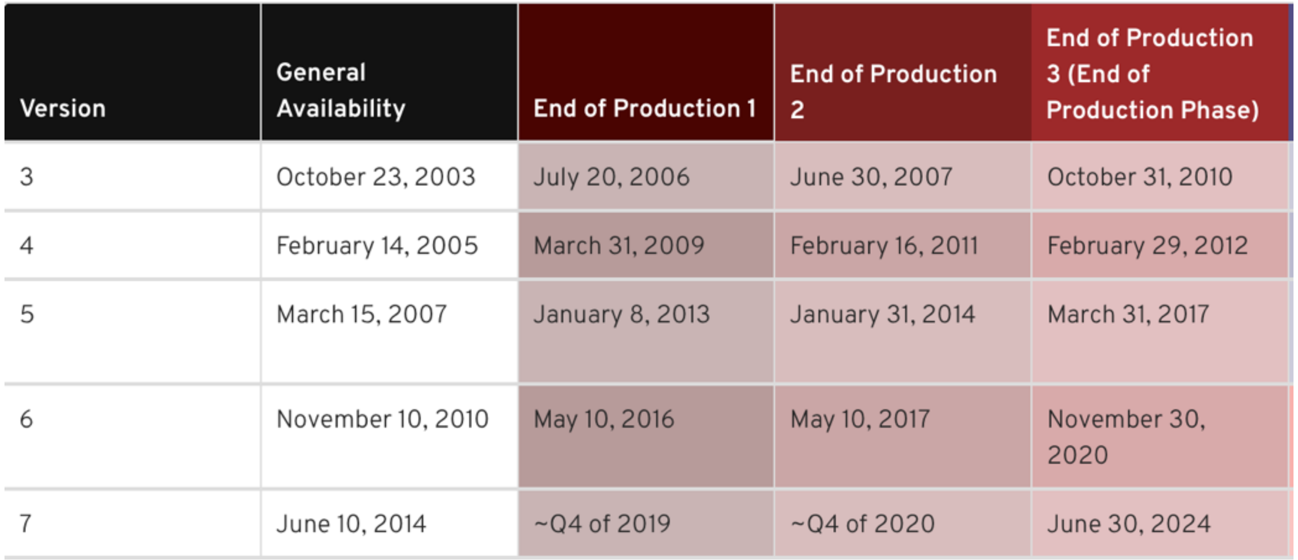
Around batch: the WGS

- We need to provide new WGS, as we want a WGS to only serve one batch system
- ATLAS and CMS provided with two physical machines
- ILC and Belle with one
- Will change with increasing workload
- See docu for entry points

Operating Systems on the NAF

- RHEL 6 (and hence SL 6) have entered „Production 3“ on 10.5.2017
- RHEL 7 (and hence CentOS 7) is there and well established
- Test CentOS 7: First test: Check whether an SL6 compiled binary runs on EL7
  - SGE BIRD: qsub –l os=el7 ...
- Keep bugging your experiment software coordinators

- “Production 3“ means“
  - During the Production 3 Phase, Critical impact Security Advisories (RHSAs) and selected Urgent Priority Bug Fix Advisories (RHBAs) may be released as they become available. Other errata advisories may be delivered as appropriate.
  - New functionality and new hardware enablement are not planned for availability in the Production 3 Phase. Minor releases with updated installation images may be made available in this Phase.
  - MIGRATE NOW!



Version	General Availability	End of Production 1	End of Production 2	End of Production 3 (End of Production Phase)
3	October 23, 2003	July 20, 2006	June 30, 2007	October 31, 2010
4	February 14, 2005	March 31, 2009	February 16, 2011	February 29, 2012
5	March 15, 2007	January 8, 2013	January 31, 2014	March 31, 2017
6	November 10, 2010	May 10, 2016	May 10, 2017	November 30, 2020
7	June 10, 2014	~Q4 of 2019	~Q4 of 2020	June 30, 2024

Extract from https://access.redhat.com/support/policy/updates/errata

Services on the NAF

NAF Remote Desktop: FastX

- FastX demo
- ... Or look at
- https://confluence.desy.de/display/IS/Using+FastX+as+NAF+Remote+Desktop
- (linked from http://bird.desy.de/htc/)

Jupyter / ROOT notebooks

- BELLE2 is already using this technology
- Plan to start a prototype in 2018
  - First with server-only resources
  - If accepted, plan to use HTCondor as a backend (will need some investigation)

HPC and GPU computing

- DESY has set up the Maxwell HPC cluster
- Used for
  - Real parallel applications (not only „trivial parallel“ applications like HEP analysis)
  - Fast data analysis from Photon Science
- Makes uses of an additional, fast interconnect
  - Which makes it more expensive
- If special projects from the NAF *really* need HPC ressources, contact us
- GPU computing currently done in Maxwell
- One project from around the NAF started to use (and contribute) GPU systems
- Follow this effort
- If there is a strong need for GPU computing on the NAF, this can be added in future

DESY and NAF and Cloud

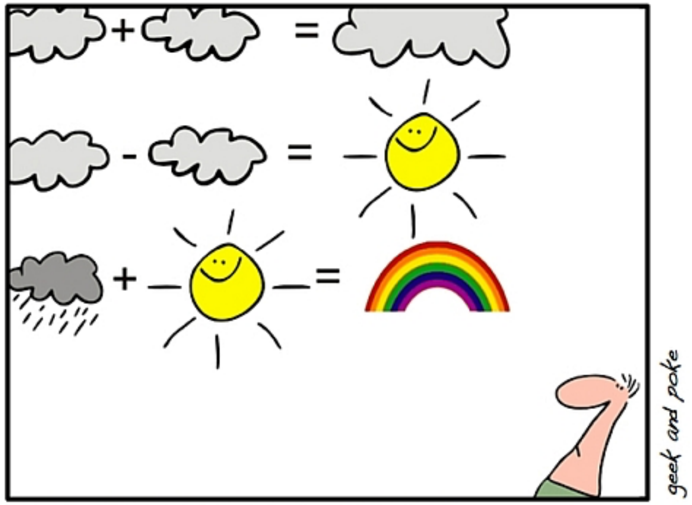
- **Compute Cloud**
- Based on OpenStack, currently in internal pilot stage
- Used in/for/with different EU projects, but also internal projects
- Integration into batch system not our current priority
- More to come in2018
- ... And we participate in HNSciCloud activities

- **Storage Cloud**
- “DropBox-like“, currently in late pilot stage
- Integration into batch system not our current priority
- More to come in 2018

HELIX NEBULA THESCIENCECLOUD

openstack.

DESY DESY CLOUD dCache.org



SIMPLY EXPLAINED - PART 17:
CLOUD COMPUTING

Schools on the NAF

- Just a reminder:
- The NAF provides a powerful infrastructure for schools
- Contact us if you are planning a workshop or school with computing needs

... And now: Your feedback!