ドイツ ATLAS Computing, GoeGrid, Göttingen グループ

ATLAS ソフトウェア講習会 2016

河村 元
II.Physikalisches Institut, Universität Göttingen

Overview

- ドイツ・コンピューティングの現状、戦略、将来
  - 欧州中核国ドイツでの Tier-1, Tier-2 センター概要
  - ATLAS German GridKa と役割
  - 戦略（ドイツ・ Wuppertal コンピューティング物理学計算機戦略会議より）
  - 将来
- GoeGrid Tier-2 in Göttingen
  - 概要、大学 Tier-2 計算機センターの役割
  - 現状の計算資源と将来
- Göttingen ATLAS 物理計算グループの研究トピックス
  - ATLAS ソフトウェア資源の ARM アーキテクチャへの移植
  - クラウドコンピューティング
  - Google TensorFlow ライブラリと分散コンピューティング
  - メタモニタリングシステム（ HappyFace, MadFace ）

ドイツ・コンピューティングの現状、戦略、将来

欧州中核国ドイツでの Tier-1, Tier-2 センター概要

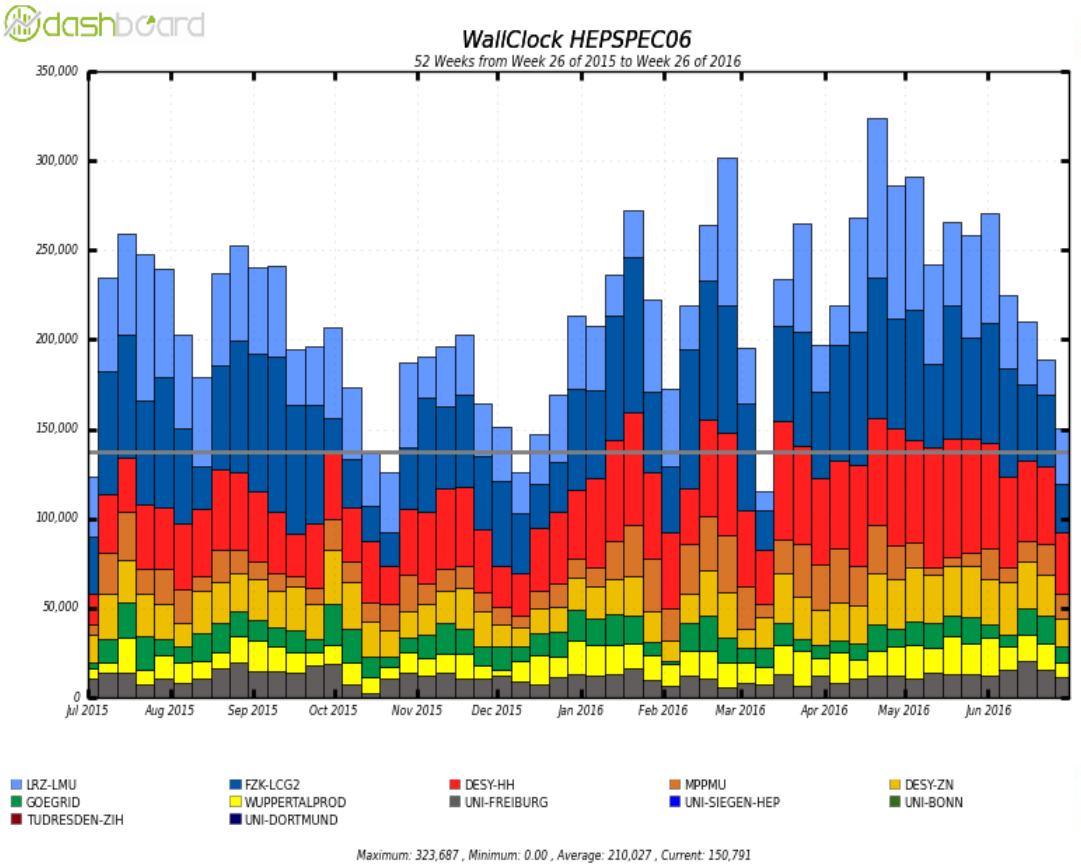
ATLAS German GridKa と役割

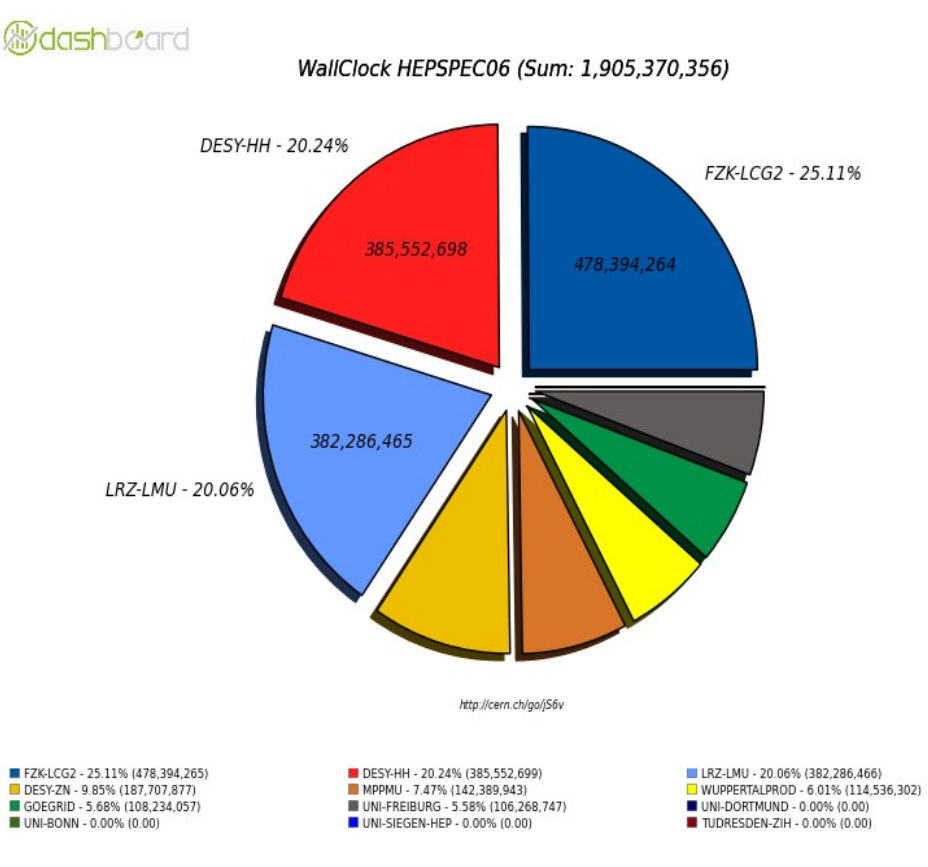
ATLAS-DE T1&T2 July15-June16

- GridKa 25%
- Desy/MPP 38%
- DE Uni T2s 37%

Sum CPU ~2 x WLCG pledge

Worldwide:
DE sites ~11.8%
3rd after US & UK





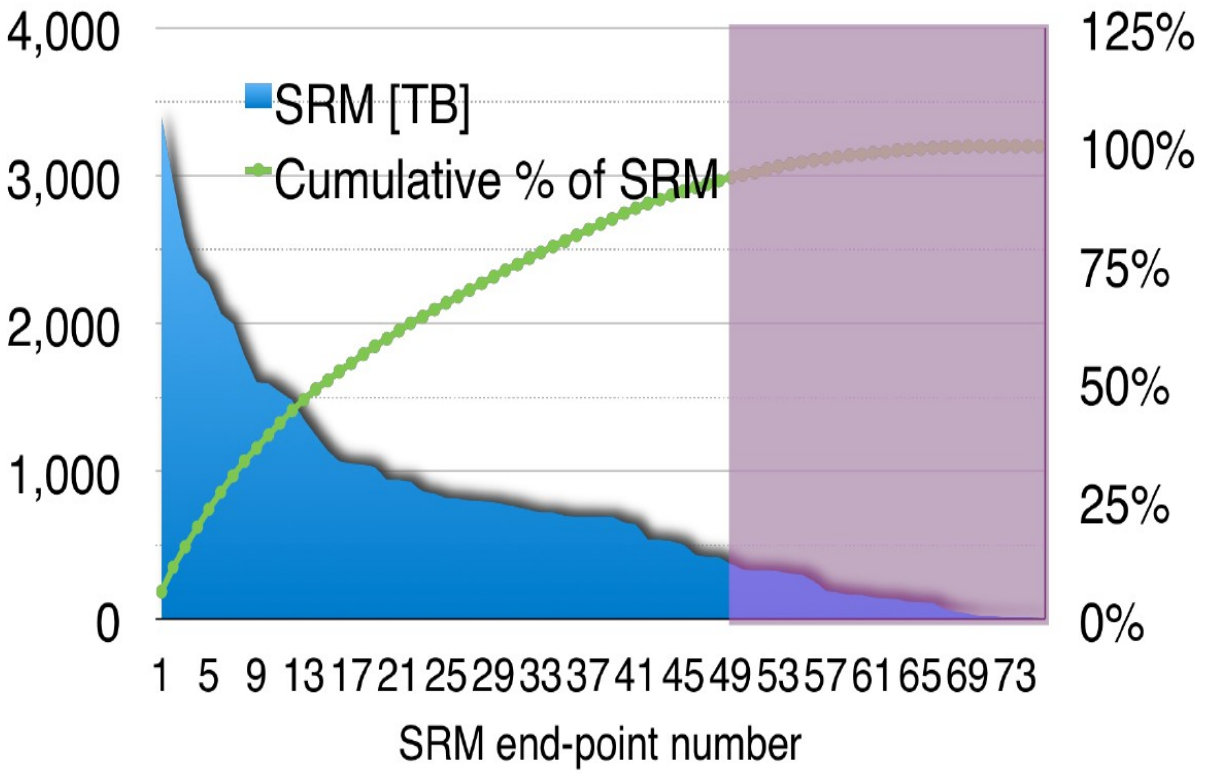
ATLAS pledges & requirements 2016/17

	2016		2017	
	CPU	Disk	CPU	Disk
FR	9433	1200	8261	1050
GOE	9433	1200	8261	1050
LMU	9433	1200	8261	1050
WUP	9433	1200	8261	1050
MPP	9433	1200	14100	1300
DESY	25000	2400	36000	2700
Sum Uni	37732	4800	33044	4200
Desy/MPP	34433	3600	50100	4000
Sum DE	72165	8400	83144	8200
ATLAS reqt	566000	72000	846000	78000
ATLAS DE share	12.75%	11.67%	9.83%	10.51%
GridKa	65000	5875	65000	5875
ATLAS reqt	520000	47000	682000	57000
ATLAS DE share	12.50%	12.50%	9.53%	10.31%

- Notes:
  - Need funding for 2017 @ University T2s to keep up with ATLAS reqts
  - GridKa pledge delayed until end 2016, in particular disk urgently needed

Storage/Site Consolidation

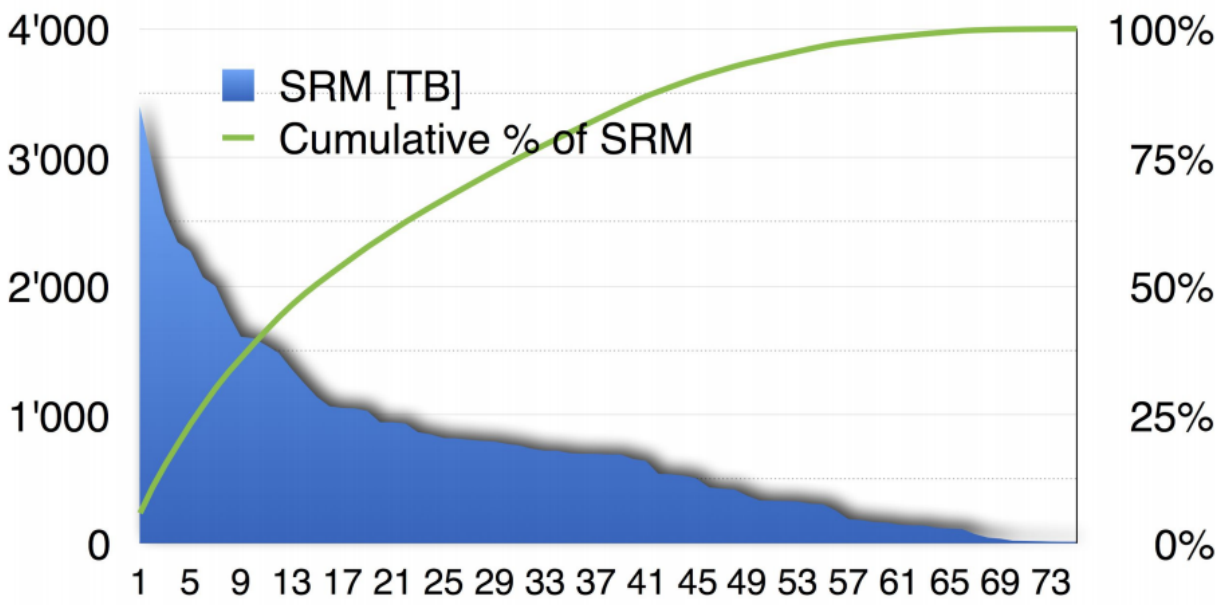
- Reduce large number of sites with small storage:
  - If below 400 TB convert DDM endpoint to cache-only endpoint
  - Focus/invest in CPUs
- Presented to WLCG & C-RRB
- No concern for DE
  - All T1/T2 sites > 1000 TB
- Nothing concrete on larger regional federations (AFAIK)



Disclaimer… Status

- Task/Goal
  - status/plans from the University T2 (WLCG) – no other found ;-)
- small/simple questionnaire sent to Uni contacts to generate overview
  - what type of services you provide or plan to provide (T2, T3, CPU only, with storage - private only or shared with WLCG)
  - what are the top causes (money of course ;-) which could have made the services better scale in quantity and quality (past and current status)
  - what was/is the situation reqruiting personnel for setup and operations
  - scaling plans (quantity and quality) for the period 'until HL LHC' and 'after'
  - concrete plans to react on the (most) current WLCG planning for the next few years (type of services, quantity, ...)
  - what was/are the financial sources (predictable ?) from University and other sources (i.e. BMBF, ...) - absolute numbers are fine, but not required ;-)
  - top 5 solutions, how the German resource providers could strenghten their role and services under the assumption of the magic 'flat funding model'

Available storage at Tier 2 sites



More efficient to have larger and fewer storage end-points
2 possible categories : 'Cache based' & 'large' Tier 2s

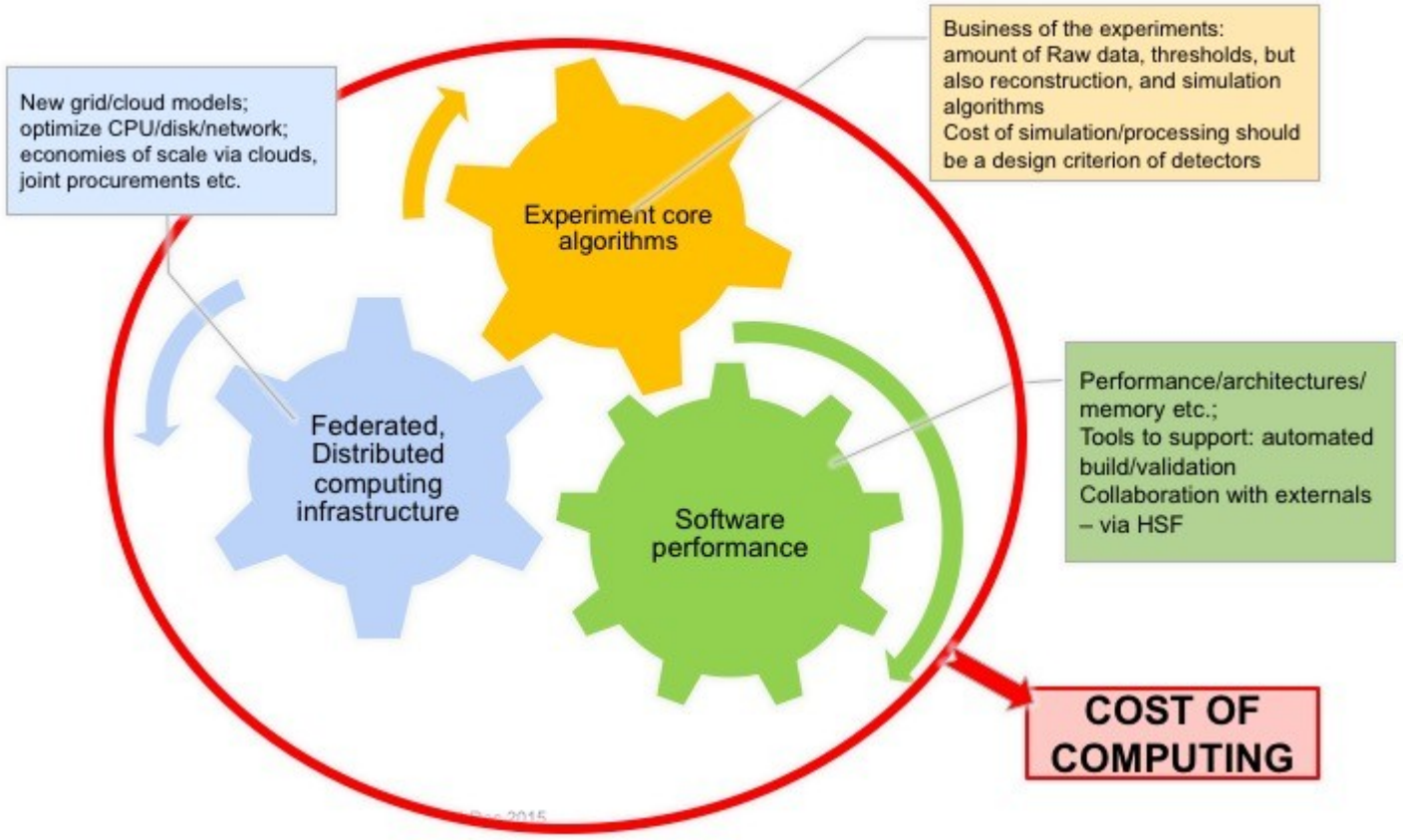
6

Data

A possible medium term plan

- SRM: progress with decommissioning, apart for tapes
- Data access, upload, download:
  - Consolidate around the xrootd protocol (mainstream)
  - Progress with HTTP support, valuable both in the short and medium/long term
- Data Transfer
  - Investigate possible alternatives to gridFTP (e.g. xrootd like Alice, HTTP)
  - Do not forget that data deletion is as challenging as data transfer

HL-LHC cost drivers



CERN WLCG

1) Definition of the upgrade problem

Set up a study group to:

- Firstly:
  - Establish and update estimates of actual computing requirements for HL-LHC, more realistic than previous estimates:
    - what are the baseline numbers for data volumes/rates, CPU needs, etc.?
  - Build a realistic cost model of LHC computing, help to evaluate various models and proposals – this will be a key to guiding direction of solutions
- Secondly:
  - Look at the long term evolution of computing models and large scale infrastructure
    - Need both visionary “revolutionary” model(s) that challenge assumptions, and “evolutionary” alternatives
  - Explore possible models that address (propose strawman models)
    - Today’s shortcomings
    - Try to use best of evolving technologies
    - Address expectations of how the environment may evolve
      - Large scale joint procurements, clouds, interaction with other HEP/Astro-P/other sciences
    - Possible convergence of (the next generation of) main toolsets

2) Software-related activities

- Strengthen the HSF:
  - "Improve software performance" –
    - Need to define what the goals and to define metrics for performance:
      - E.g. time to completion vs throughput vs cost
    - Continue concurrency forum/HSF activities – but try and promote more
    - And other initiatives like reconstruction algorithms etc
  - Techlab
    - expand as a larger scale facility under HSF umbrella
    - Include support tools (profilers, compilers, memory etc)
      - Including support, training, etc
      - openlab can also help here
    - Should be collaborative – CERN + other labs
  - Technology review
    - "PASTA" – reform the activity – make into an ongoing activity, updating report every ~2 years
      - Broad group of interested experts
    - Also under HSF umbrella – strongly related to the above activities
  - What can be done about long term careers and recognition of software development

3) Performance evaluation/"modelling"

- Investigate real-world performance of today's systems:
  - Why is performance so far from simple estimates of what it should be?
  - Different granularities/scales:
    - Application on a machine
    - Site level: bottlenecks, large-scale performance
      - Different scale sites, different workflows
    - Overall distributed system
      - At which level?
      - Are data models and workflows appropriate?
- Once we have a better handle of actual performance – can we derive some useful models/parameterisations etc?
  - Useful enough to guide choices of computing models – don't have to be perfect or complete
  - This feeds into any cost models
- Small team in IT starting to work on this and consolidate existing efforts
  - Define a programme of work to look at current performance and concerns; define initial goals

Grid, Batch, Storage Resources, Status

Grid Hamburg/Zeuthen
- 15000/2350 cores
- Torque+home-build scheduler
- Univa Grid Engine
- HTCondor

Batch
- 7340/2250+80GPU cores
- SoGE (unstable)
- Univa Grid Engine

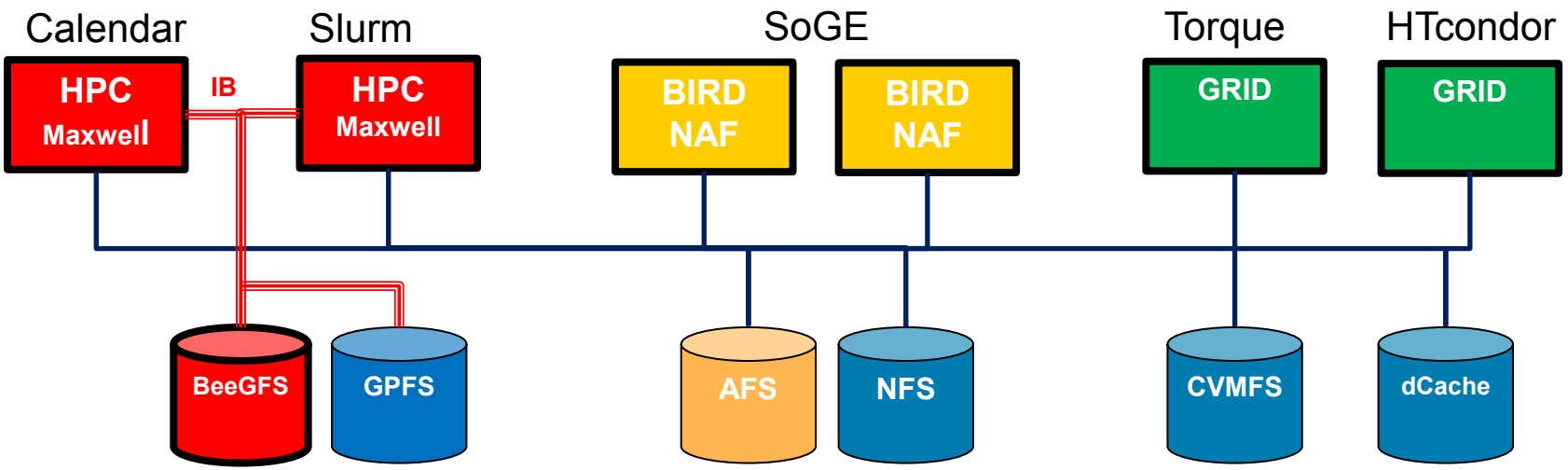
HPC
- Calendar (not scalable)
- Slurm
- Univa Grid Engine

dCache
- 13/3.3 PB

Batch local
- 0.8 PB Sonas
- 1.7 PB Lustre
- 1.4 PB GPFS

Batch Consolidation

Hamburg



50% Grid WN are migrated to Htcondor
HTCondor works fine in production

AFS/Kerberos integration: installation in test

Zeuthen considers migration

HTCondor

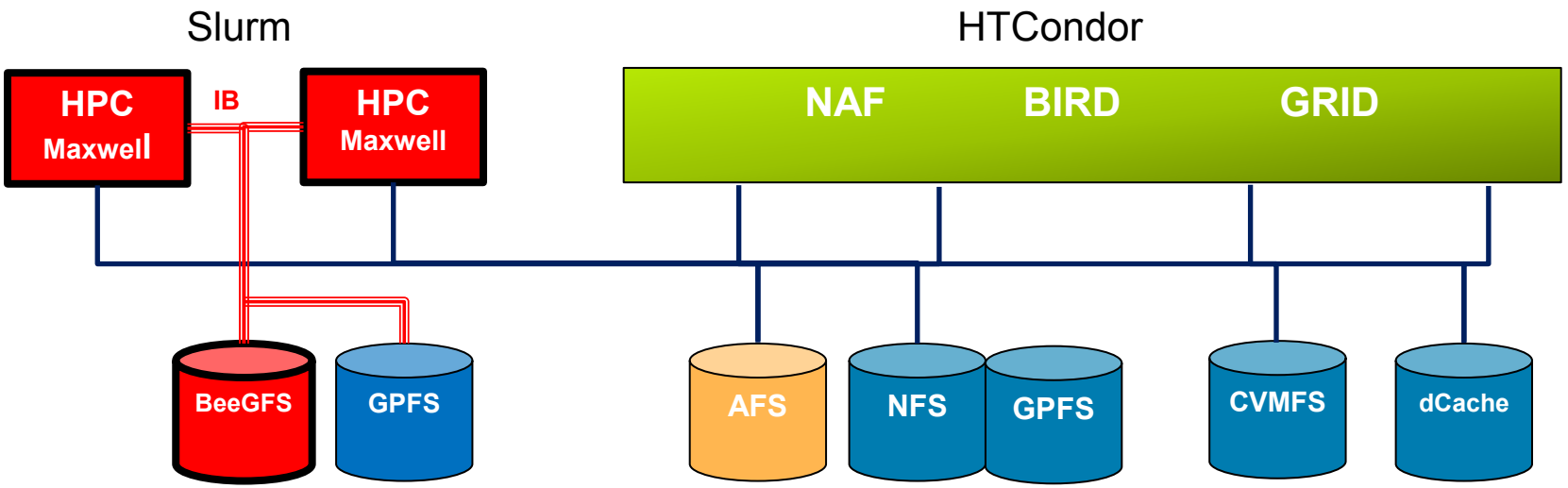
HTCondor is a scheduling and batch system. Main features:

- Extreme scalability (CERN n*100k cores)
- Free and open sources, many adaptes in HEP
- Relatively new code base and modular design
- Optimized for high throughput clusters (as opposed to high performance)

HTCondor @ Grid

- Currently ~8k cores in pool (using hyper-threading)
  This is the half of the resources that is under warranty. The other half still with Torque
- Jobs submitted via 2 ARC CE, one condor master host
- Working: Multicore jobs, quotas, installation via puppet, Grid UI and experiment software via CVMFS, monitoring Icinga and grafana
- Todo: More monitoring, housekeeping ofWNs, ...

Batch Consolidation



Productivity
Scalability
Flexibility

Computing for Astroparticle Physics

Computing support for various Astroparticle Physics projects:
IceCube, Cherenkov Telescope Array (CTA), Veritas, Magic, H.E.S.S., Fermi, ...

European Tier1 for IceCube (part of IceCube Maintenance and Operations MoU)

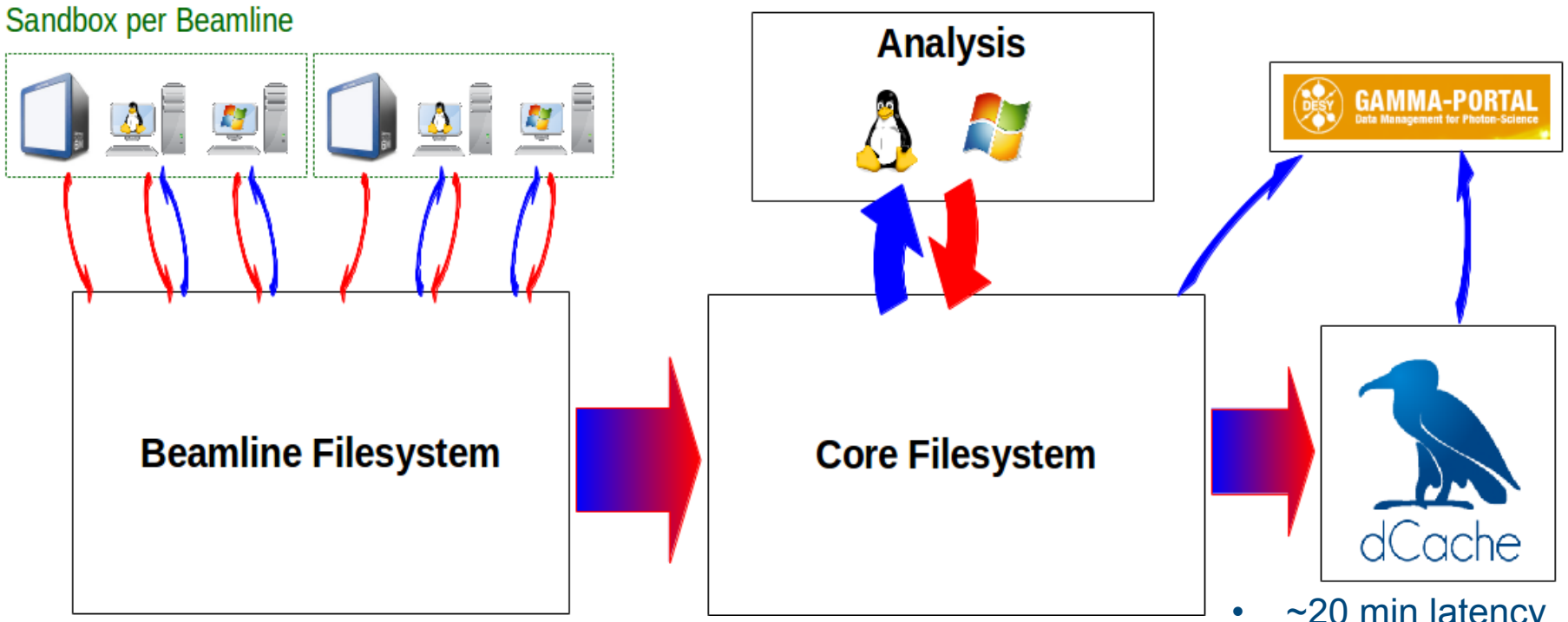
- IceCube simulation production – Grid and local farm computing
  100 GPUs for photon propagation in ice
- Hosting of filtered data (tape-backed), 50% of simulated data (disk-only)
- Continuous data transfer UW Madison --> Zeuthen (300MB/s)
- Acting as disaster recovering center

CTA Science Data Management Center (tbd.)

- Science coordination including software maintenance and data processing for the Observatory

Data Taking Petra3

Logical Dataflow



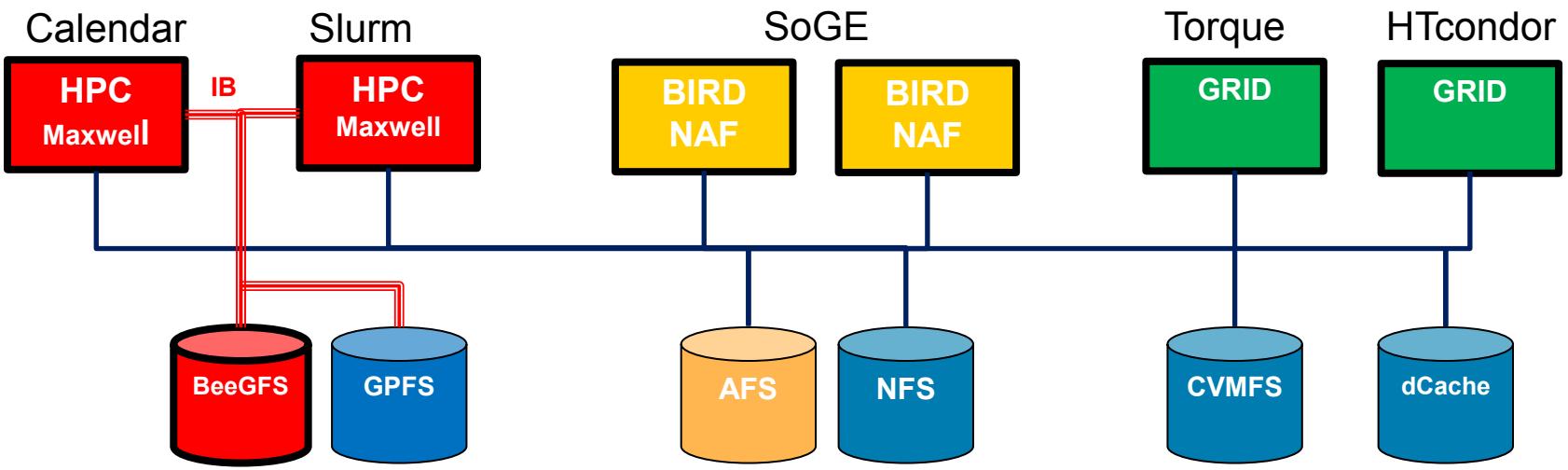
- Low latency
- Low capacity
- Host-based authentication

- 4 min latency
- High capacity
- Full user authentication

- ~20 min latency
- Very high capacity, tape
- Full user authentication

Batch Consolidation

Hamburg



50% Grid WN are migrated to Htcondor
HTCondor works fine in production

AFS/Kerberos integration: installation in test

Zeuthen considers migration

HTCondor

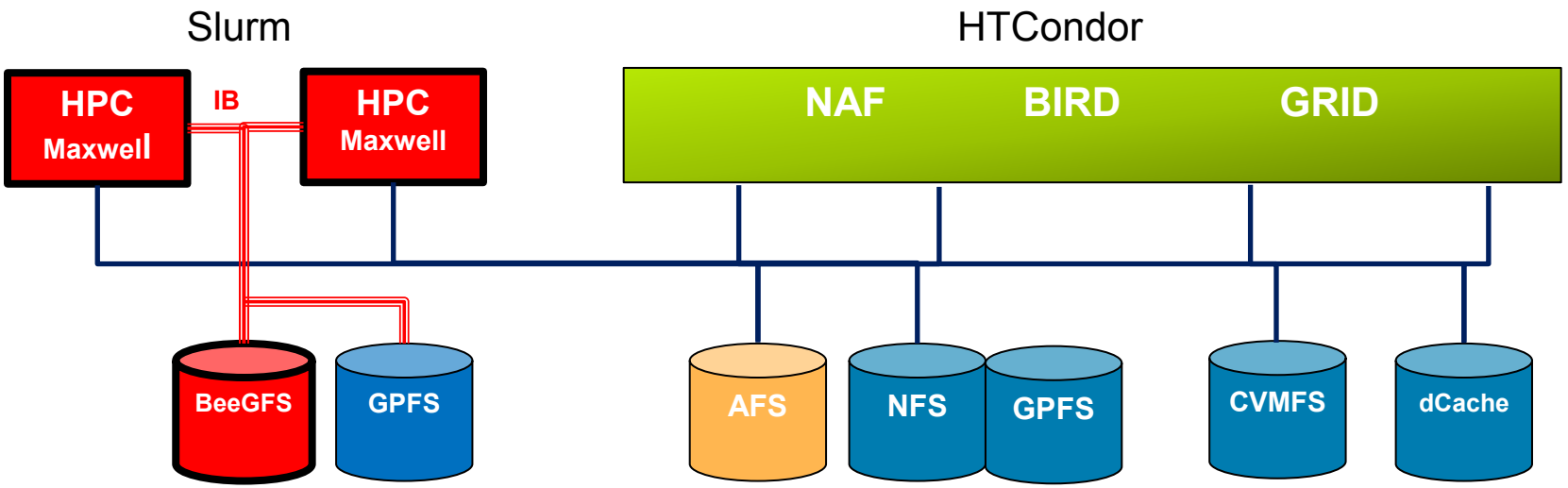
HTCondor is a scheduling and batch system. Main features:

- Extreme scalability (CERN n*100k cores)
- Free and open sources, many adaptes in HEP
- Relatively new code base and modular design
- Optimized for high throughput clusters (as opposed to high performance)

HTCondor @ Grid

- Currently ~8k cores in pool (using hyper-threading)
  This is the half of the resources that is under warranty. The other half still with Torque
- Jobs submitted via 2 ARC CE, one condor master host
- Working: Multicore jobs, quotas, installation via puppet, Grid UI and experiment software via CVMFS, monitoring Icinga and grafana
- Todo: More monitoring, housekeeping ofWNs, ...

Batch Consolidation



Productivity
Scalability
Flexibility

Computing for Astroparticle Physics

Computing support for various Astroparticle Physics projects:
IceCube, Cherenkov Telescope Array (CTA), Veritas, Magic, H.E.S.S., Fermi, ...

European Tier1 for IceCube (part of IceCube Maintenance and Operations MoU)

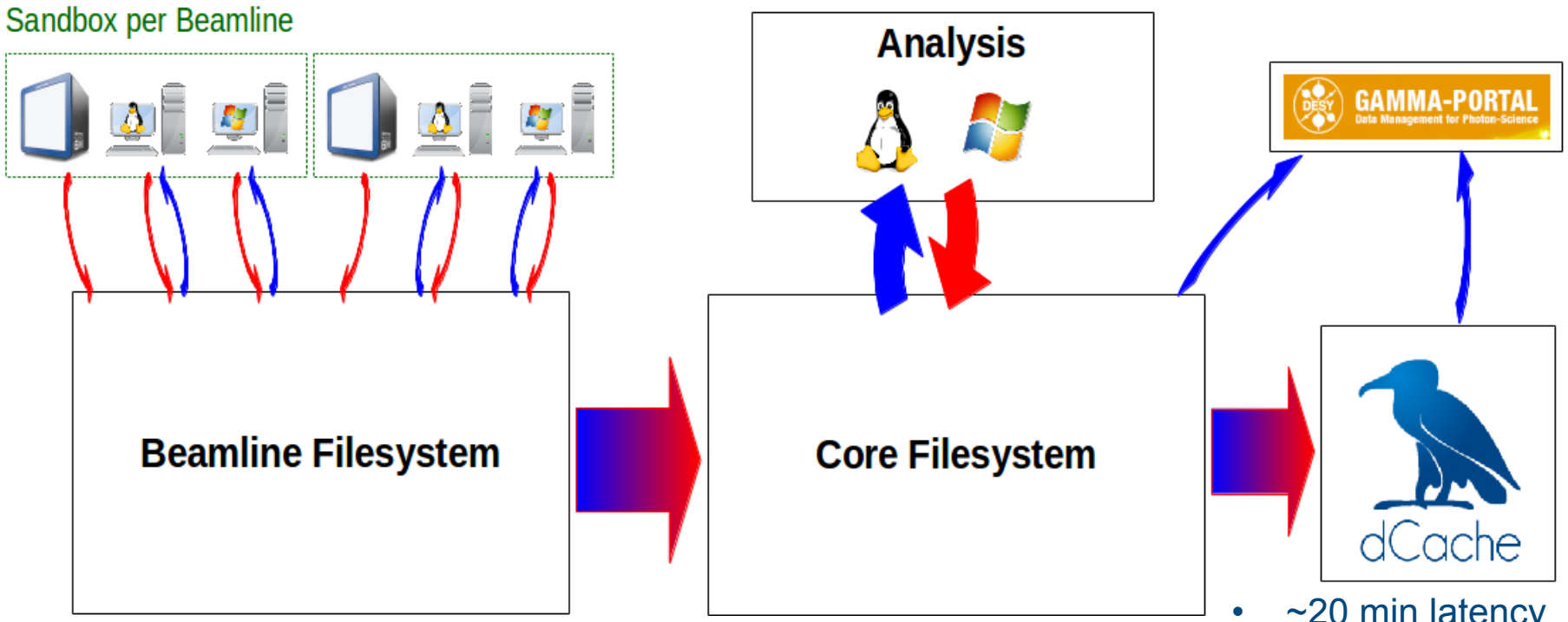
- IceCube simulation production – Grid and local farm computing
  100 GPUs for photon propagation in ice
- Hosting of filtered data (tape-backed), 50% of simulated data (disk-only)
- Continuous data transfer UW Madison --> Zeuthen (300MB/s)
- Acting as disaster recovering center

CTA Science Data Management Center (tbd.)

- Science coordination including software maintenance and data processing for the Observatory

Data Taking Petra3

Logical Dataflow



- Low latency
- Low capacity
- Host-based authentication

- 4 min latency
- High capacity
- Full user authentication

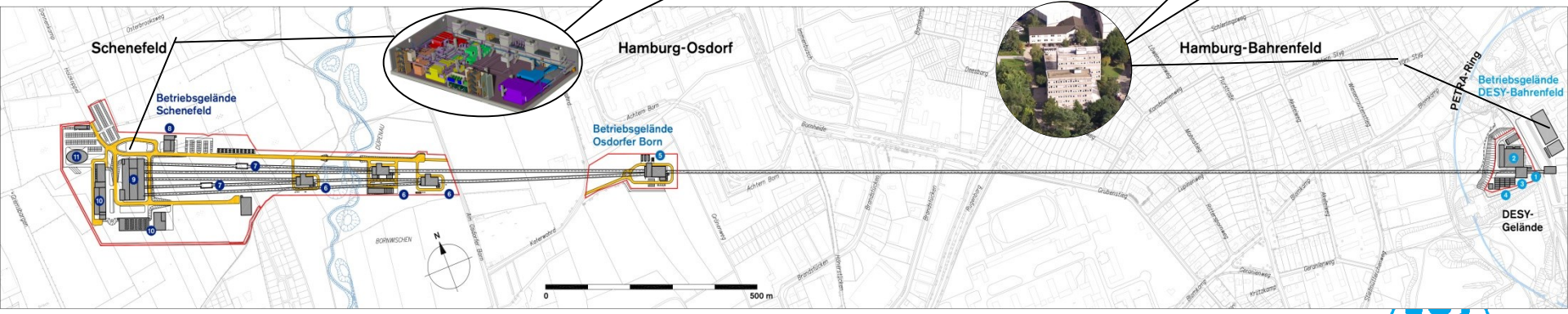
- ~20 min latency
- Very high capacity, tape
- Full user authentication

Preparation for European XFEL

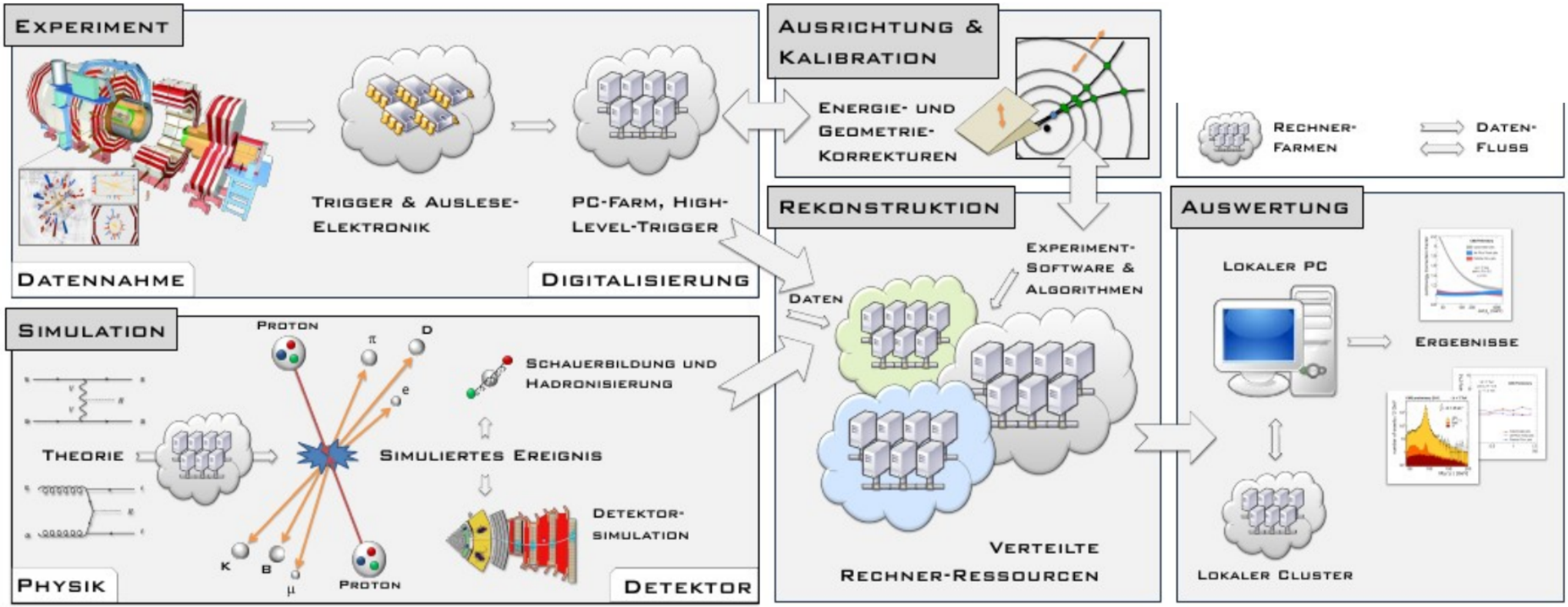
- Use the system developed for PETRA III
- as a blueprint for XFEL
- Setup of distributed GPFS system in
- Schenefeld and at DESY
- Management of the dataflow
- High-performance and reliable coupling
- between the locations

GPFS system for online storage and computing

GPFS system for offline storage and computing

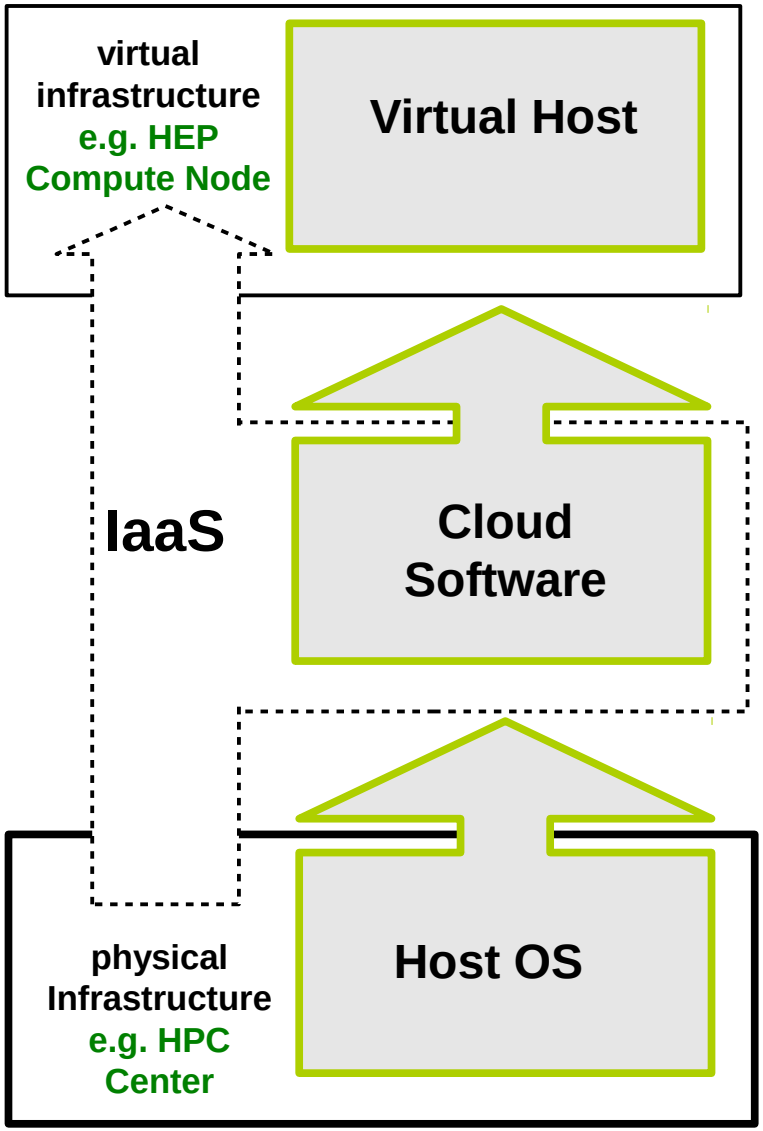


Diverse Computing in HEP



- HEP software applications are very diverse with different requirements in terms of I/O and compute
- Some applications need to be located at specific sites
  - HLT farm must be close to the detector
  - Event reconstruction needs fast access to the measurements stored on disk
- Simulation and analysis can be located more freely
- Fast WAN connections blur the strict hierarchical layering of data. Any Data, Anytime, Anywhere (AAA) via xrootd becomes a reality.

From Physical to Virtual Infrastructure



The Infrastructure-as-a-Service (Iaas) model

- Infrastructure (e.g. machines, networks) is virtualized
- Decouples complexities of hardware maintenance and specific software setup

- The life cycle of this virtual infrastructure is managed by a Cloud system:
  - Virtual machine images are managed
  - The user can upload and start custom virtual machines
  - Storage blocks can be attached to these VMs

The Cloud vs. The Cloud

An important distinction needs to be drawn between Cloud Technologies, Hardware and Software (closed source, open source) and the Cloud Hoster (Companies)

Cloud Technologies

Cloud Hoster

Recent Development: Docker (docker.io)

Uses Linux kernel to ensure encapsulation of applications (kernel cgroups)

- Lightweight alternative to full virtualization
- Resource saving, because no additional memory usage by virtualized OS
- User applications (and dependencies) are bundled in containers
  - For example: Belle II automatically generates Docker container with the full experiment software stack for new releases
- Also all user land software and libraries of a OS flavour (for example Scientific Linux) can be bundled into one container
  - Successfully executed Scientific Linux-based application (CMS software framework) on Ubuntu system without any adaptations
- **Some limitations apply:**
  - No cross-architecture support (i686 vs. x86 32-bit vs. x86 64 bit): no problem as HEP computing uses x86-64 almost exclusively
  - Security guarantees not as strong as full virtualization: docker container creator need to be trusted
  - Advanced features like virtualized network not provided as core docker feature

The core concept of Hard- and Software abstraction can also be provided via Docker. The opportunities for HEP Computing are equivalent to fully virtualized systems.

Standardized access to Computing resources

Apart from the operational aspects (who runs the cloud service etc.):

Cloud Technologies can act for us a HEP community as an (industry-) standardized entry point to various resource providers

Cloud-Hoster provides:

- Configuration of machine type (Number of cores, memory etc.)
- Allocation of storage resources
- Scalability of allocated resources
- Network configuration and encapsulation
- Billing and accounting
- API for automated resource allocation and configuration

Customization required for HEP usage:

- HEP community members can use this base-layer of offerings and APIs to implement the software layer best suited for the task at hand: Monte Carlo production, user analysis etc.
- Usage of a customized VM image (or docker container or) reuse of existing virtual machine images (for example CERN VM)

Cloud-ready technologies in HEP (today !)

This is an (incomplete) list of software used in the HEP domain, that is already targeting the Cloud computing area or works excellent in such an environment.

CernVM (https://cernvm.cern.ch/)

- Virtual Machine Image based on Scientific Linux maintained by CERN
- Very lightweight and can be directly deployed on various cloud sites

CernVM-FS (https://cernvm.cern.ch/portal/filesystem)

- On-demand file system using HTTP protocol to download files from central repository
- Many big experiments use CernVM-FS today to deploy new software versions to compute centers of the WLCG
- CernVM-FS works excellent also on cloud sites (via HTTP Proxy)

HTCondor (https://research.cs.wisc.edu/htcondor/)

- Free and open-source batch system
- Excellent with integrating worker dynamic worker nodes (even behind NATed networks)

DIRAC / VMDIRAC (https://github.com/DIRACGrid/VMDIRAC/wiki)

- Used for gird job submission and data management by LHCb and Belle II [1]

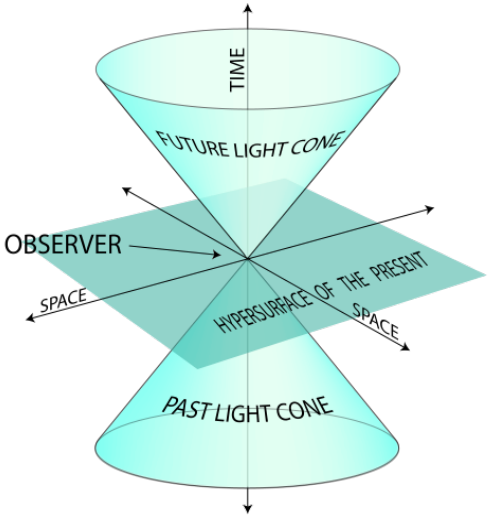
[1] http://iopscience.iop.org/article/10.1088/1742-6596/664/2/022021

戦略（ドイツ・ Wuppertal コンピューティング物理学計算機戦略会議より）

将来

GoeGrid Tier-2 in Göttingen

Deutsches Elektronen SYnchrotron



Minkovsky space-time

Hisenberg

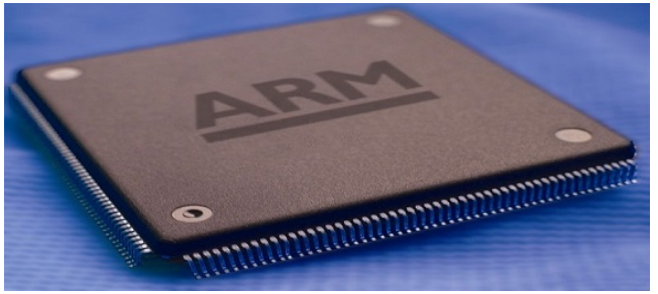
概要、大学 Tier-2 計算機センターの役割

現状の計算資源と将来

- GoeGrid is a WLCG ATLAS tier-2 resource centre Inaugurated in 2008 Usage of resources in the elds of astrophysics, biomedical sciences, grid development, high energy physics, theoretical physics, and the humanities Resource sharing based on the amount of the contribution Hosted at GWDG Part of the D-Grid initiative and the European Grid Infrastructure (EGI) 24 servers (9 virtual servers, 15 hardware storage servers) 305 compute nodes, 2508 logical CPUs 1:1Petabyte disk storage (dCache for HEP)

Göttingen ATLAS 物理計算グループの研究トピックス

Cloud Computing



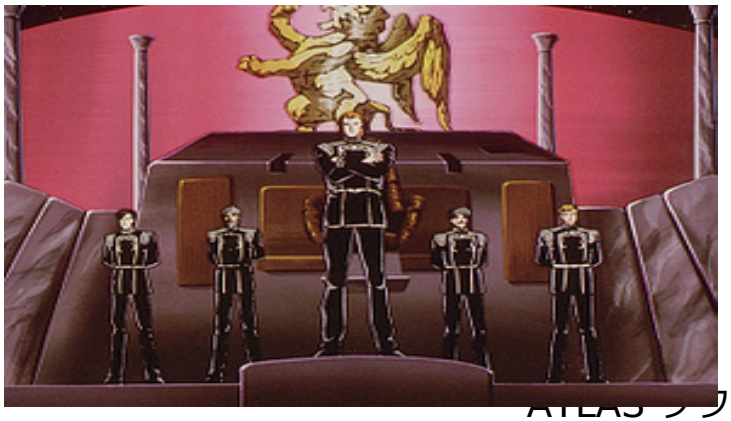
ATLAS ソフトウェア資源の ARM アーキテクチャへの移植

クラウドコンピューティング

Google TensorFlow ライブラリと分散コンピューティング

メタモニタリングシステム
（ HappyFace, MadFace ）

独逸物理計算機英雄伝説
Heldensagen von Deutschen Physikalischen Rechenmaschinen



Fragen?
質問？