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

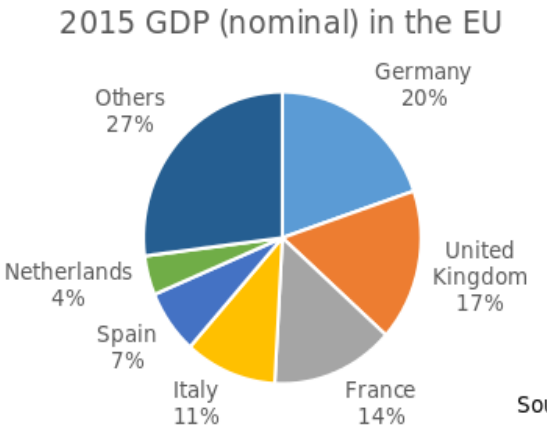
Overview

- ドイツ・コンピューティングの現状、戦略、将来
  - ドイツの Tier-1, Tier-2 センター概要
  - 戦略（ドイツ・ Wuppertal 物理学計算機戦略会議より）
  - 将来
- GoeGrid Tier-2 in Göttingen
  - Göttingen ってどこ？
  - 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

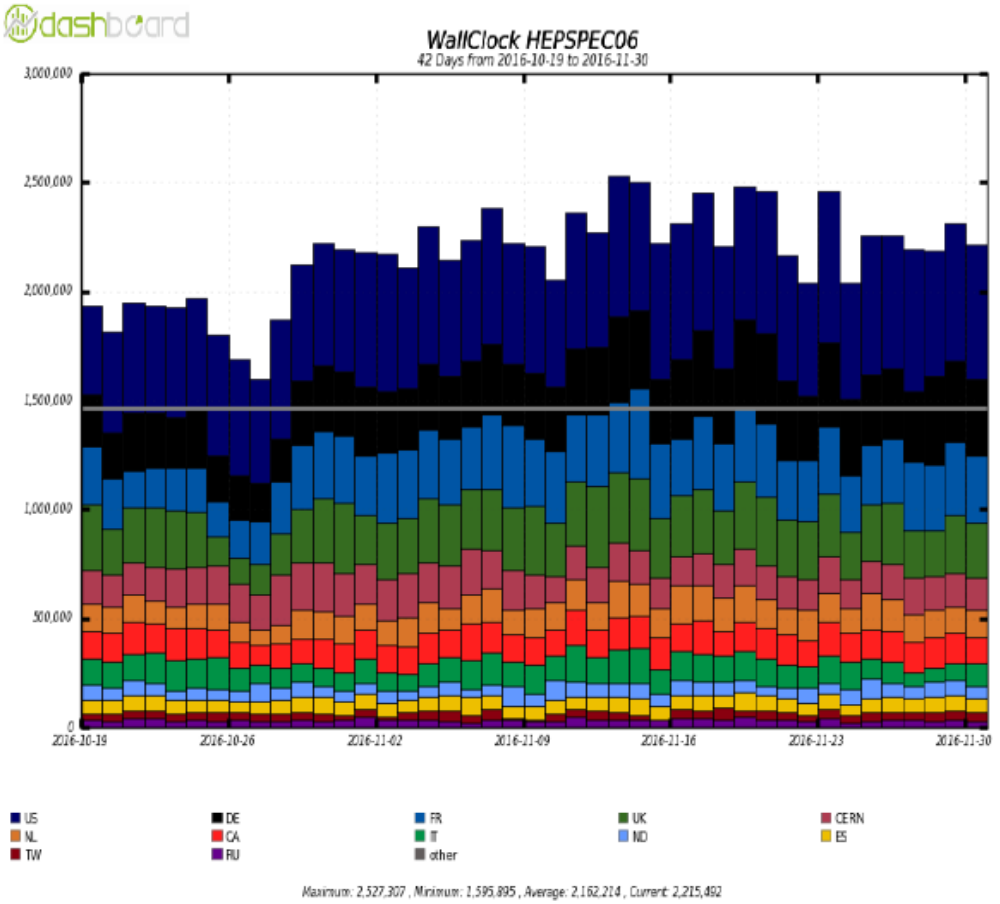
ドイツの Tier-1, Tier-2 センター概要

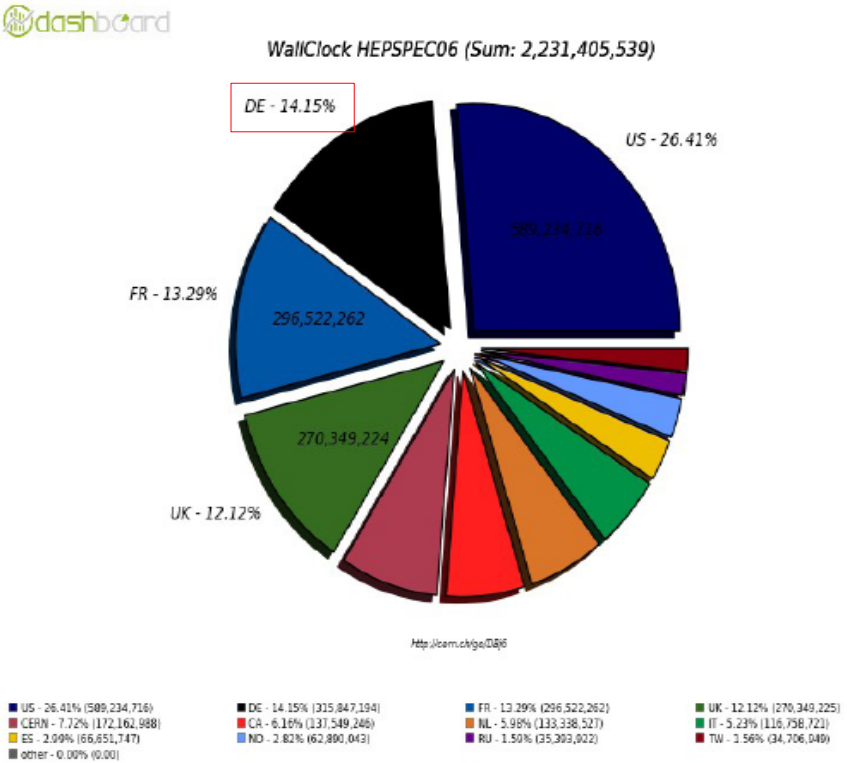
- 基礎データ
  - 欧州連合（ EU ）
    - 単一の国として見た場合世界最大の経済大国
  - ドイツ連邦共和国
    - 技術・科学を基盤とする欧州最大の経済大国
    - EU 内での GDP 比率 約 20%
      - 仮に日本が EU の場合、日本はドイツ＋スペインの規模
    - 連邦制共和国であり、地方分権が進んでいる



ドイツの Tier-1, Tier-2 センター概要

ATLAS Jobs Worldwide Oct 19 – Nov 30, 2016

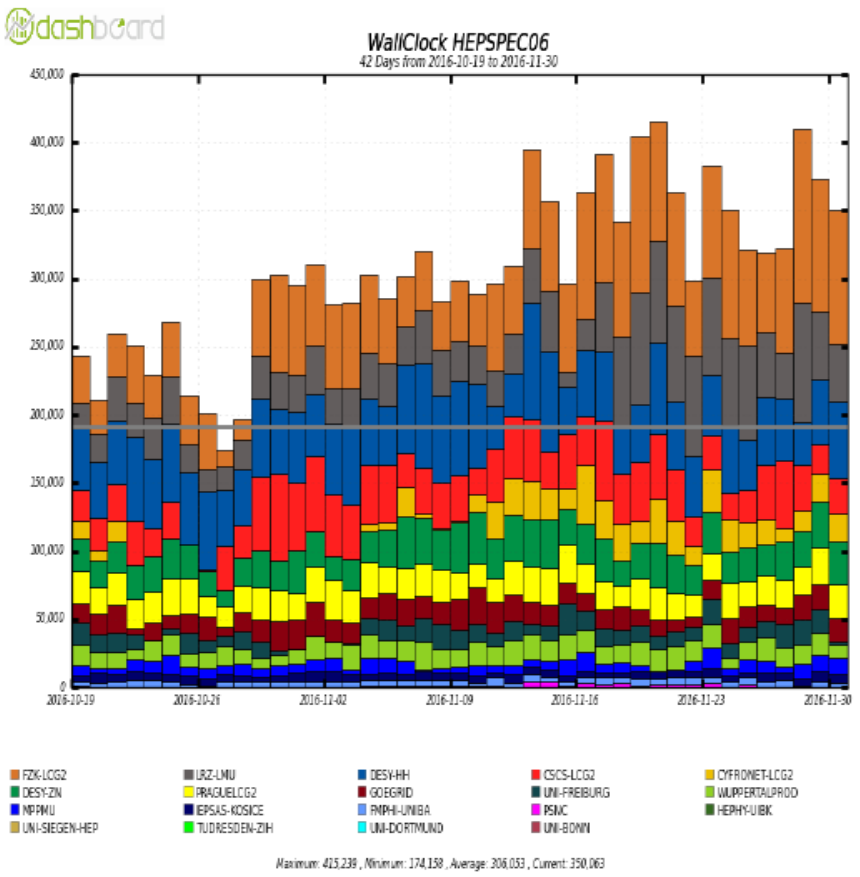


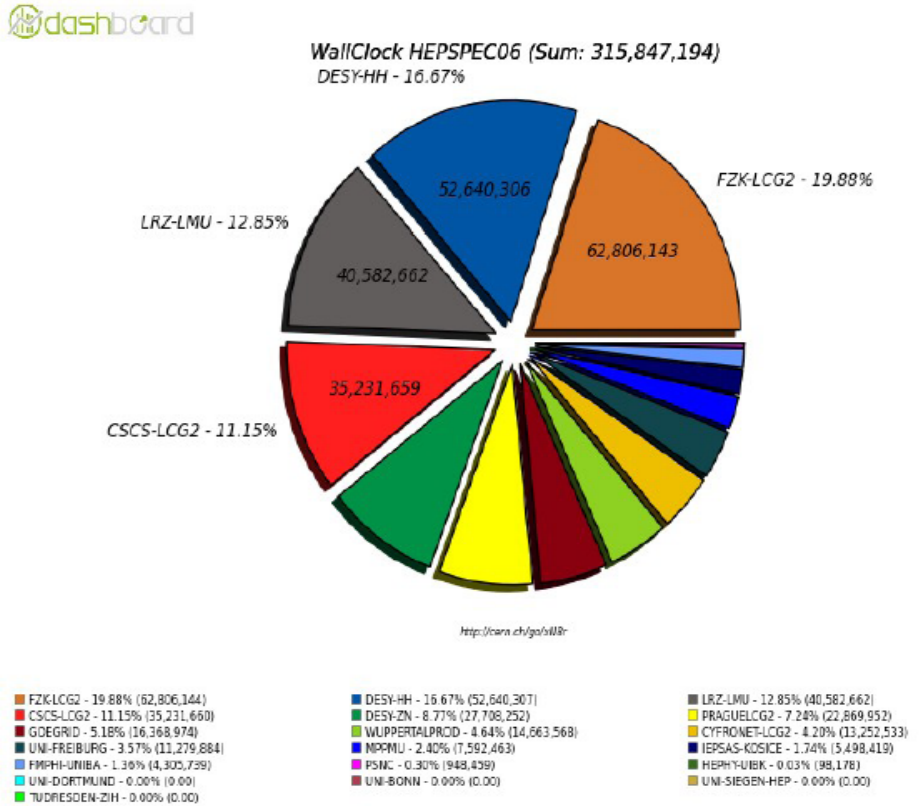


Constantly hi (tot ~150% of 2015 avg)
DE cloud #2 again w/ 14%

ドイツの Tier-1, Tier-2 センター概要

ATLAS jobs DE cloud Oct 19 – Nov 30, 2016

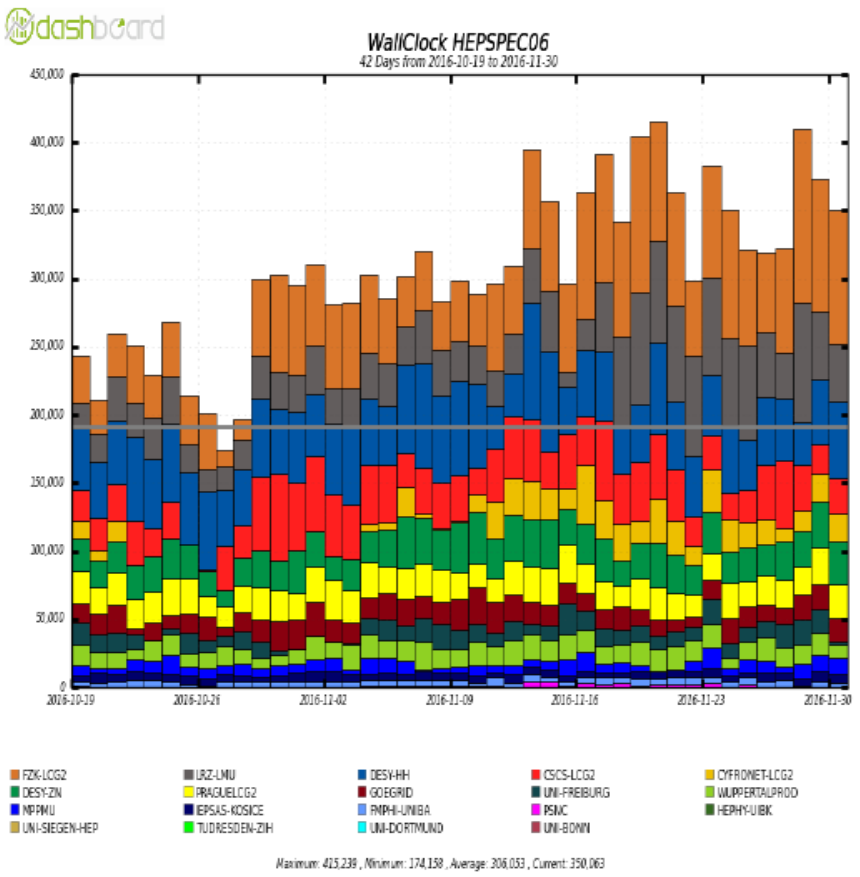


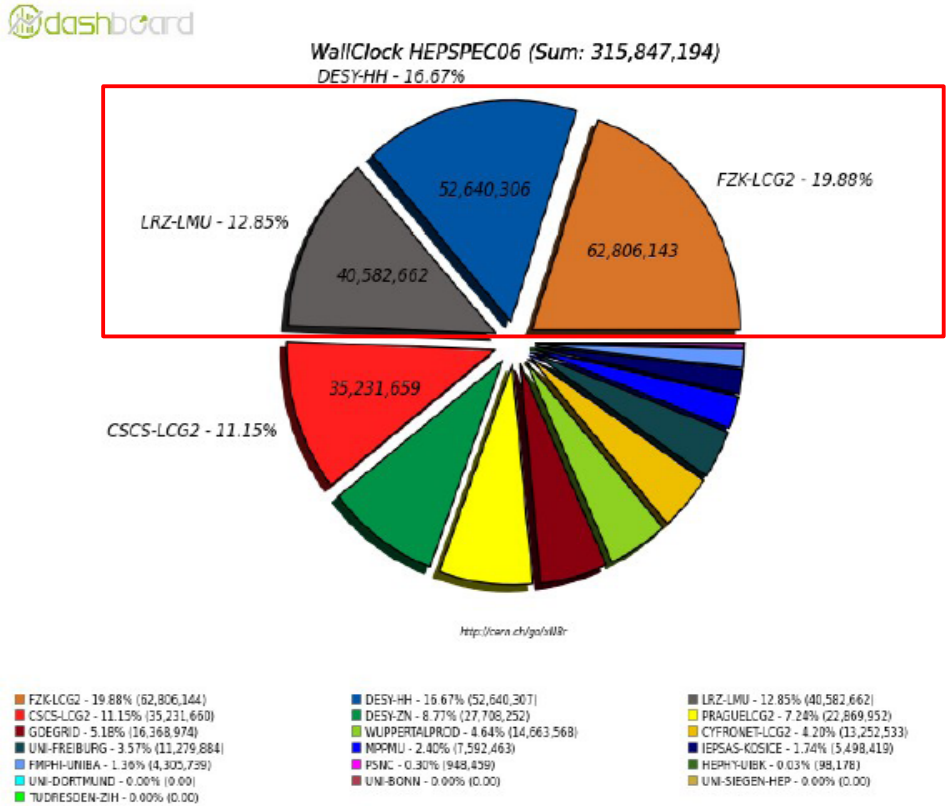


GridKa & Desy-HH leading,
strong contribs by LRZ and CSCS

ドイツの Tier-1, Tier-2 センター概要

ATLAS jobs DE cloud Oct 19 – Nov 30, 2016

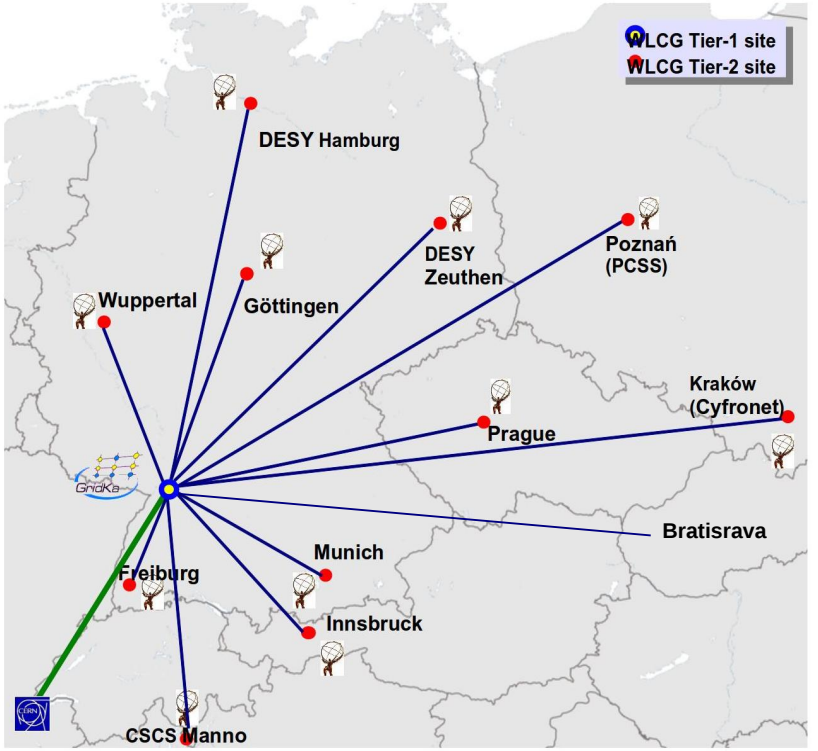




GridKa & Desy-HH leading,
strong contribs by LRZ and CSCS

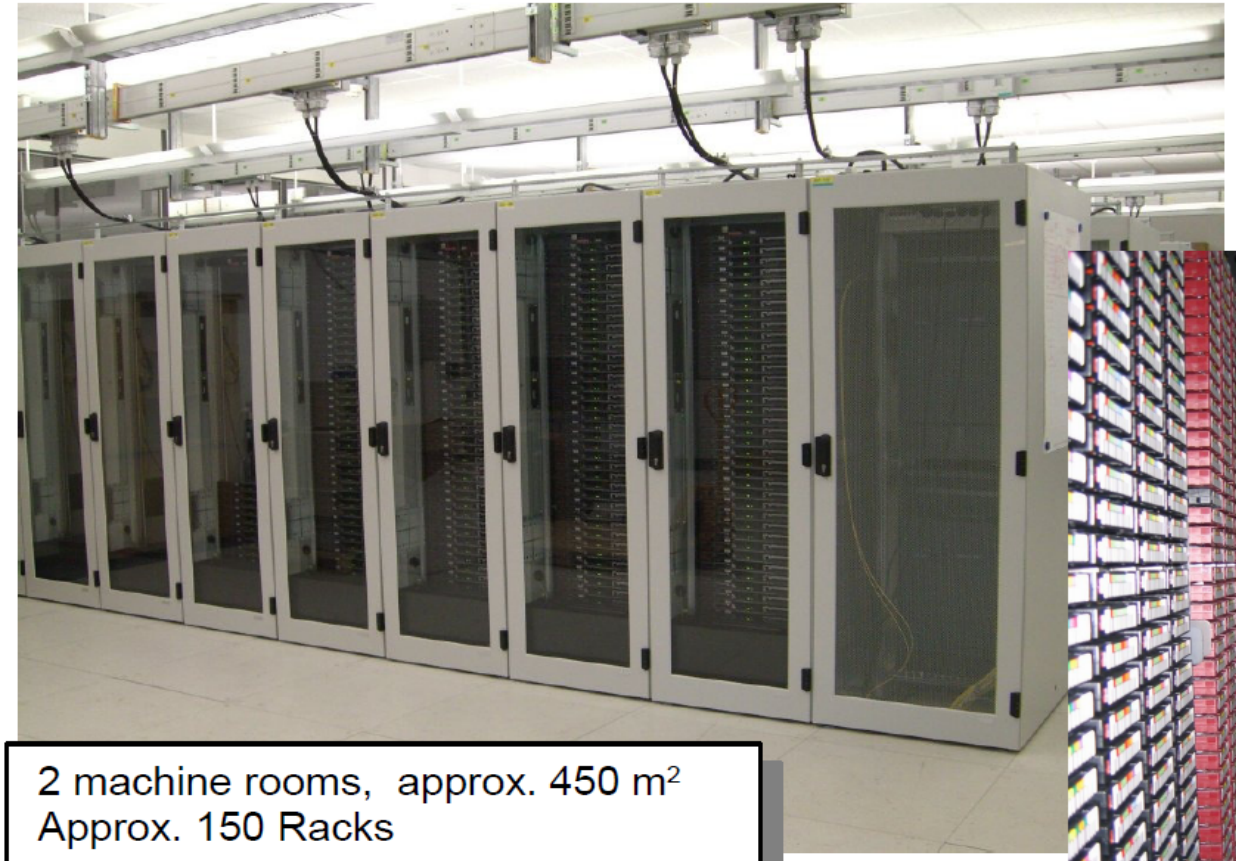
ドイツの Tier-1, Tier-2 センター概要

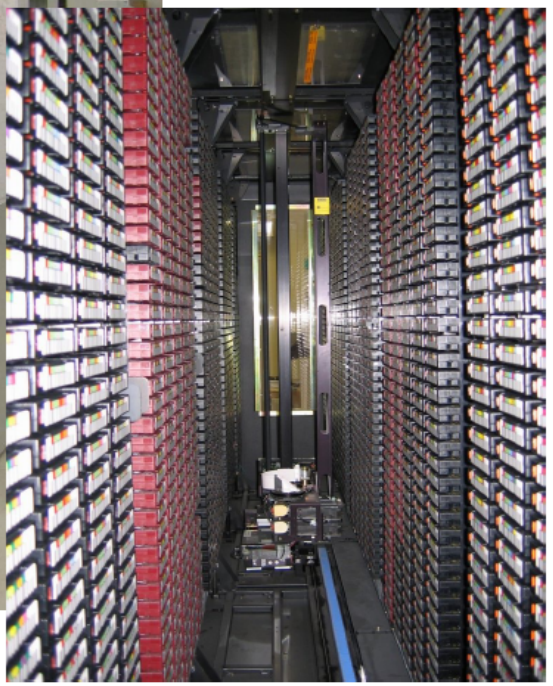
- FZK Tier-1 センター
  - Regional Data and Computing Centre in Germany (RDCCG)
  - 2002 年にスタート
  - Tier-1 support
    - 6.6 FTE
  - ATLAS GridKa
    - 1.5 FTE (9 人体制)
  - 2016.12 の資源
    - CPU cores: 23,773 (310k HEP-SPECs)
    - ストレージ: 10.7PB
- DESY-HH
  - NAF: National Analysis Facility
  - 4 FTE
  - 2016.12 の資源
    - CPU cores: 13,564 (152k HEP-SPECs)
    - ストレージ: 15.2PB



ドイツの Tier-1, Tier-2 センター概要

- FZK Tier-1 センター

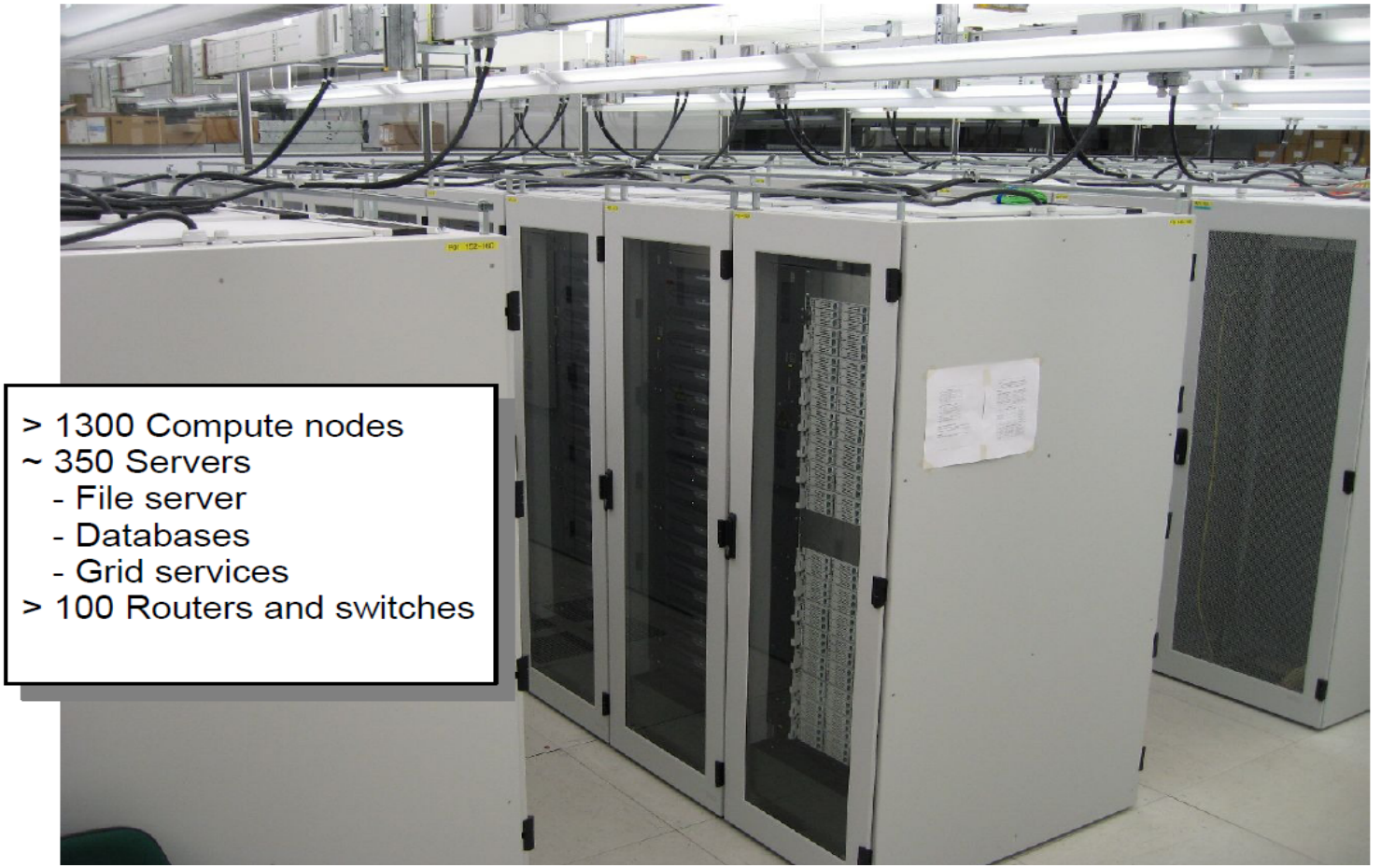




2 machine rooms, approx. 450 m²
Approx. 150 Racks

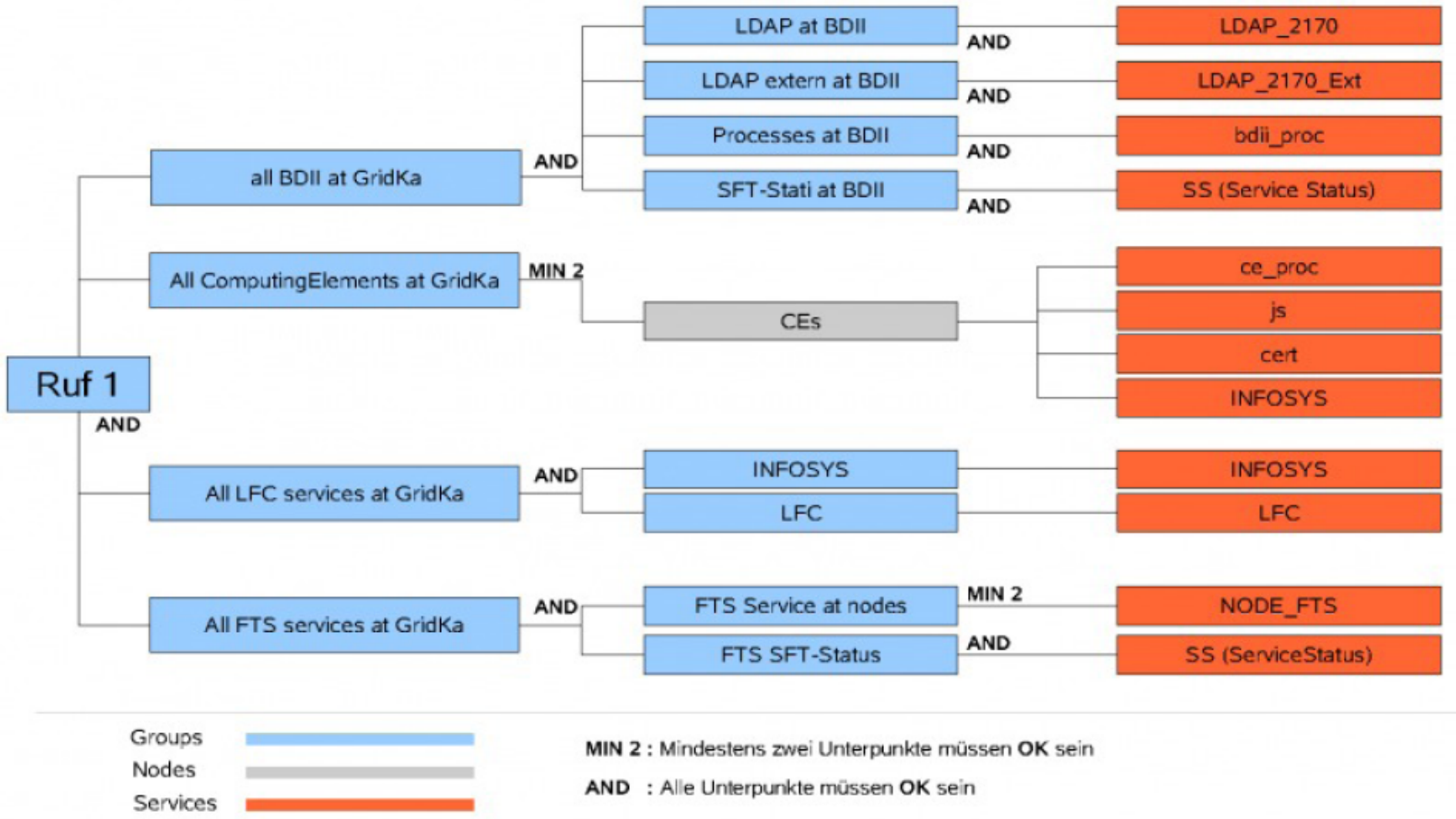
ドイツの Tier-1, Tier-2 センター概要

- FZK Tier-1 センター



ドイツの Tier-1, Tier-2 センター概要

Example: on-call alarm condition for Grid services



ドイツの Tier-1, Tier-2 センター概要

- LHC Run-2 データ取得は大きな成功
  - 各 Tier-2 サイトは 2017 年から約 2 倍増強
    - Tier-1 FZK 計算資源は徐々にシェア減
  - GridKa T1: funding not quite enough to fully match increased 2017 requests. After consulting with ATLAS CoCo agreed on:
    Disk 100%, CPU 40%, Tape 25% of respective increase
  - T2s (2 Desy, 1 MPP, 4 Uni): all match increased 2017 requests (2018 tbd)

	2016			2017		
	CPU	Disk	Tape	CPU	Disk	Tape
Sum Uni T2s	37733	4800		75000	5533	
Desy/MPP T2s	34433	3600		56250	4150	
Sum DE T2s	72167	8400		131250	9683	
T2 ATLAS DE share	12.75%	11.67%		11.67%	11.67%	
GridKa T1	65000	5875	14500	97200	8500	22100
T1 ATLAS DE share	12.50%	12.50%	12.50%	10.55%	12.50%	11.80%

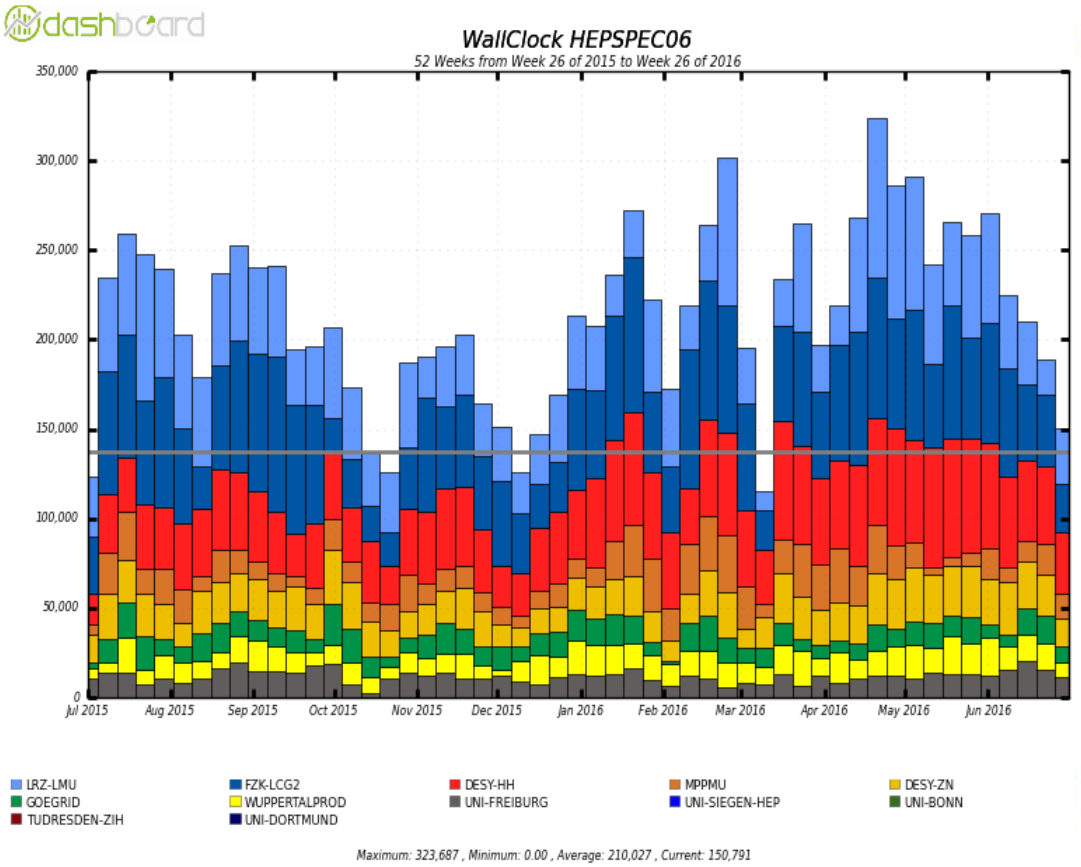
戦略（ドイツ・ Wuppertal 物理学計算機戦略会議より）

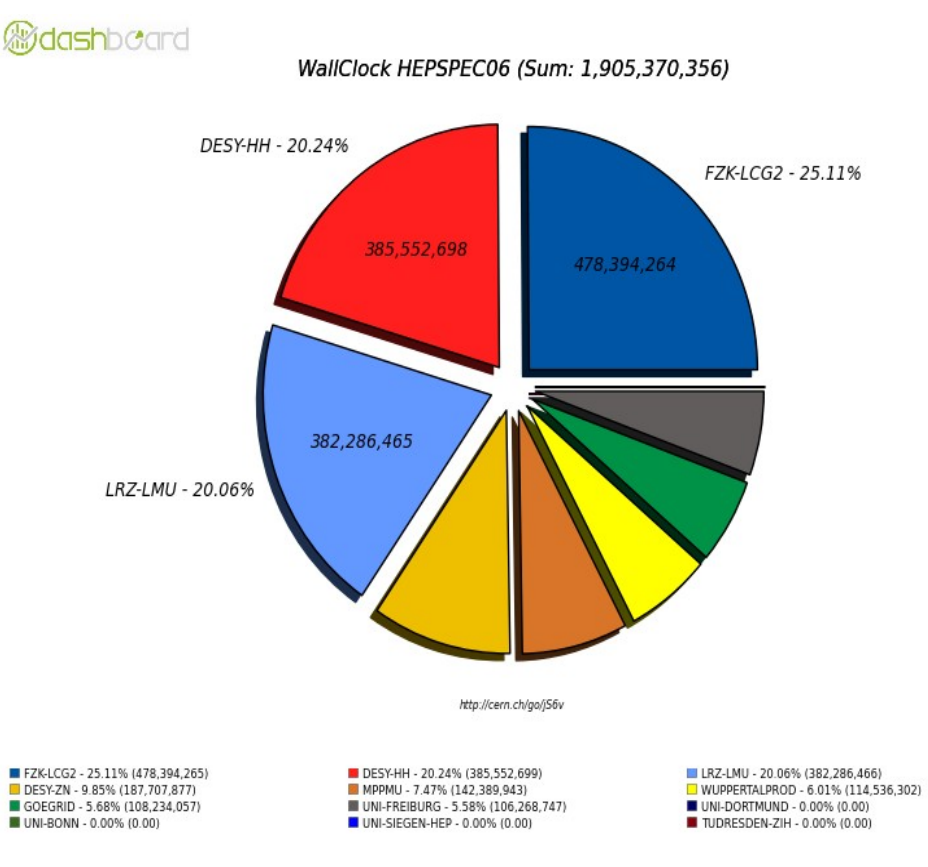
ATLAS-DE T1&T2 July15-June16

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

Sum CPU ~2 x WLCG pledge

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





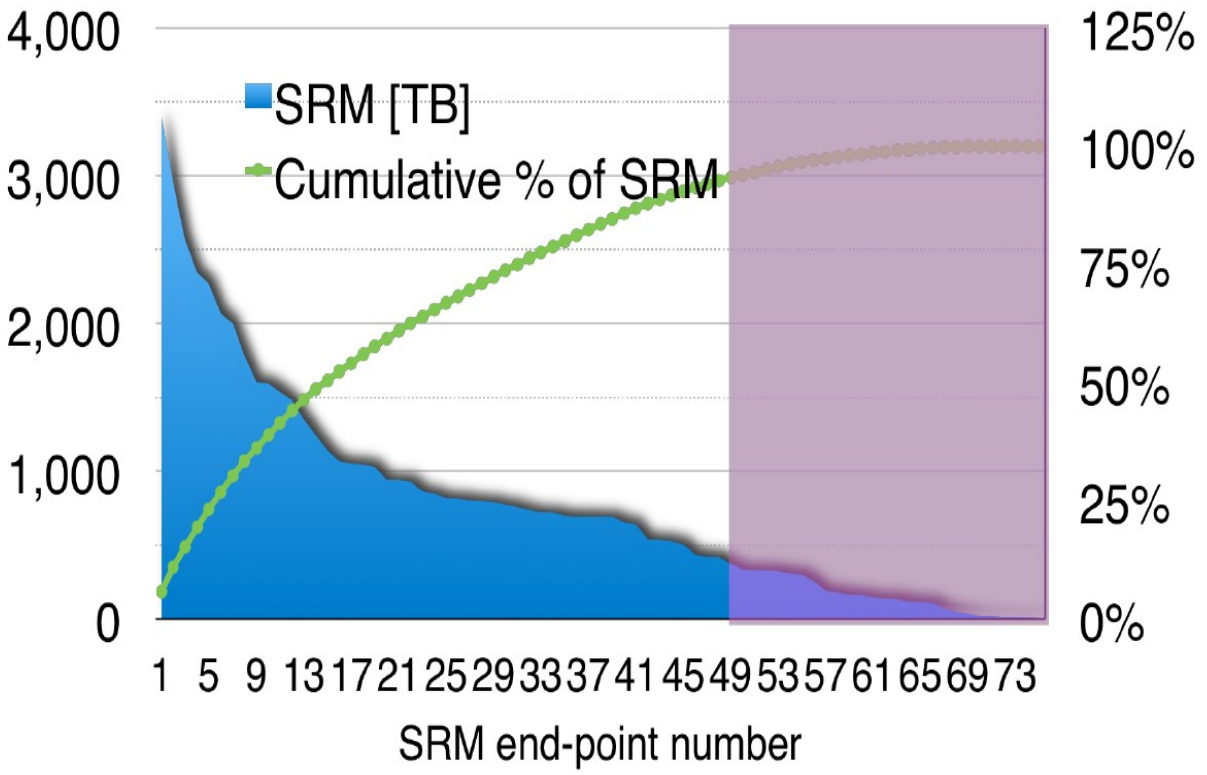
ATLAS pledges & requirements 2016/17

	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

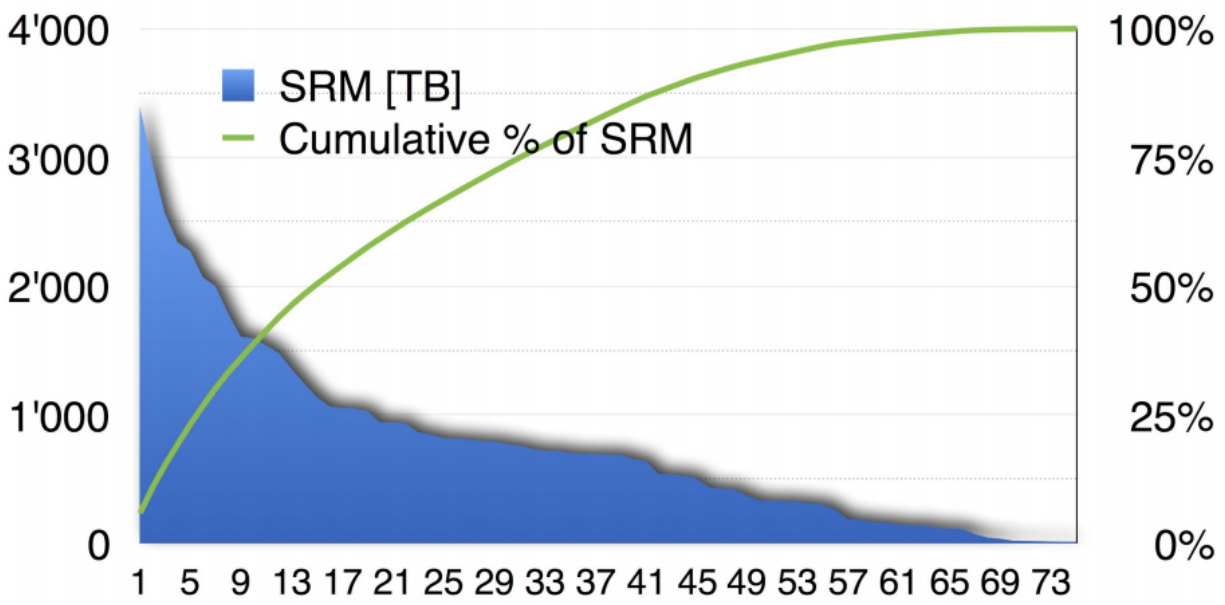
- Reduce large number of sites with small storage:
  - If below 400 TB convert DDM endpoint to cache-only endpoint
  - Focus/invest in CPUs
- Presented to WLCG & C-RRB
- No concern for DE
  - All T1/T2 sites > 1000 TB
- Nothing concrete on larger regional federations (AFAIK)



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

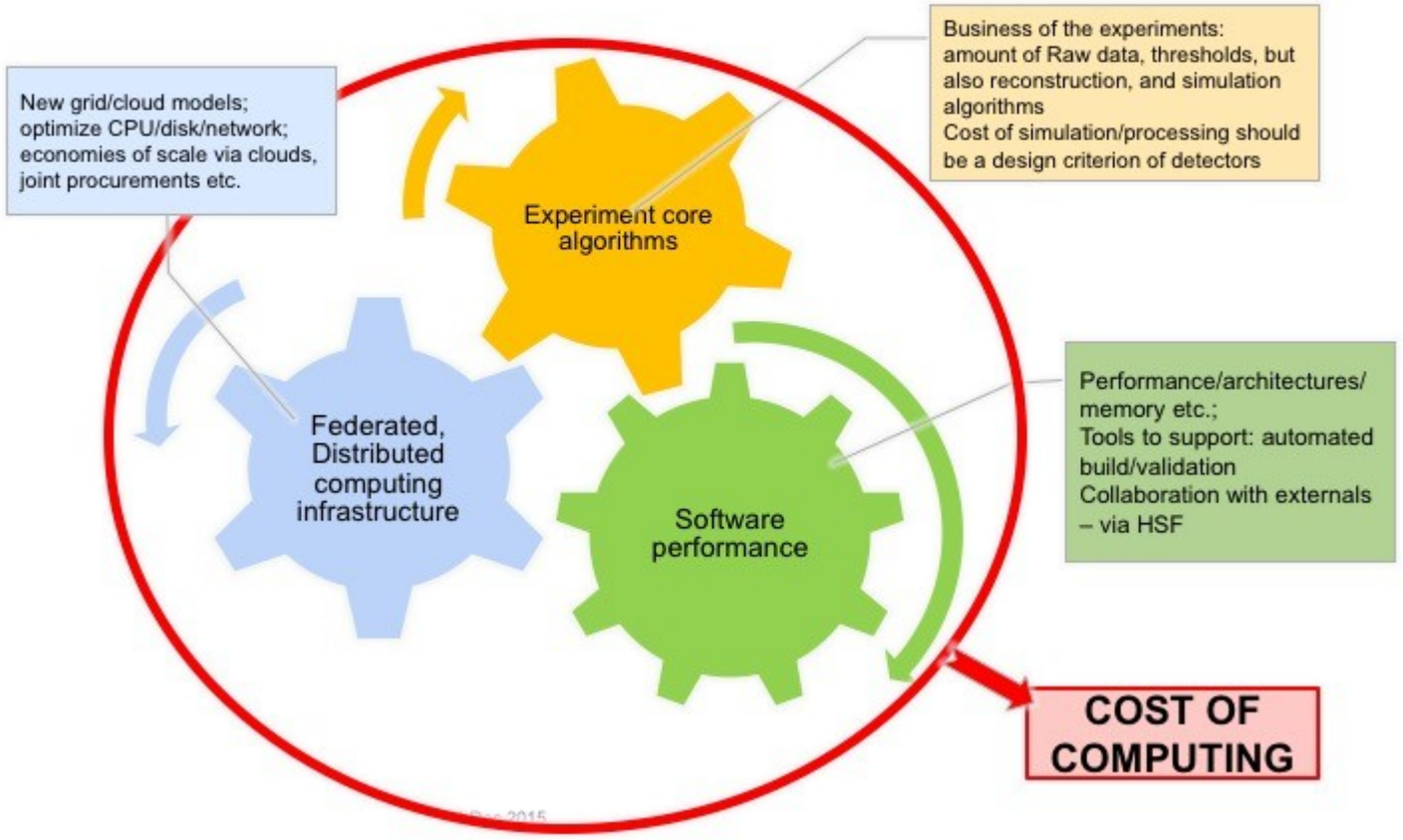
6

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

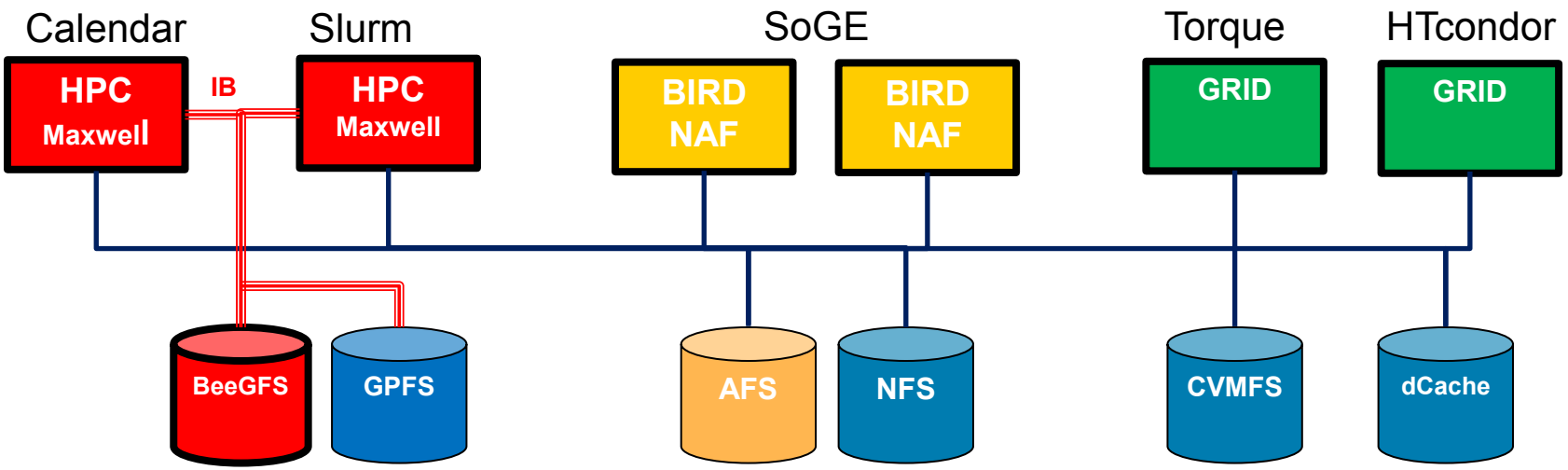
HPC
- Calendar (not scalable)
- Slurm
- Univa Grid Engine

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

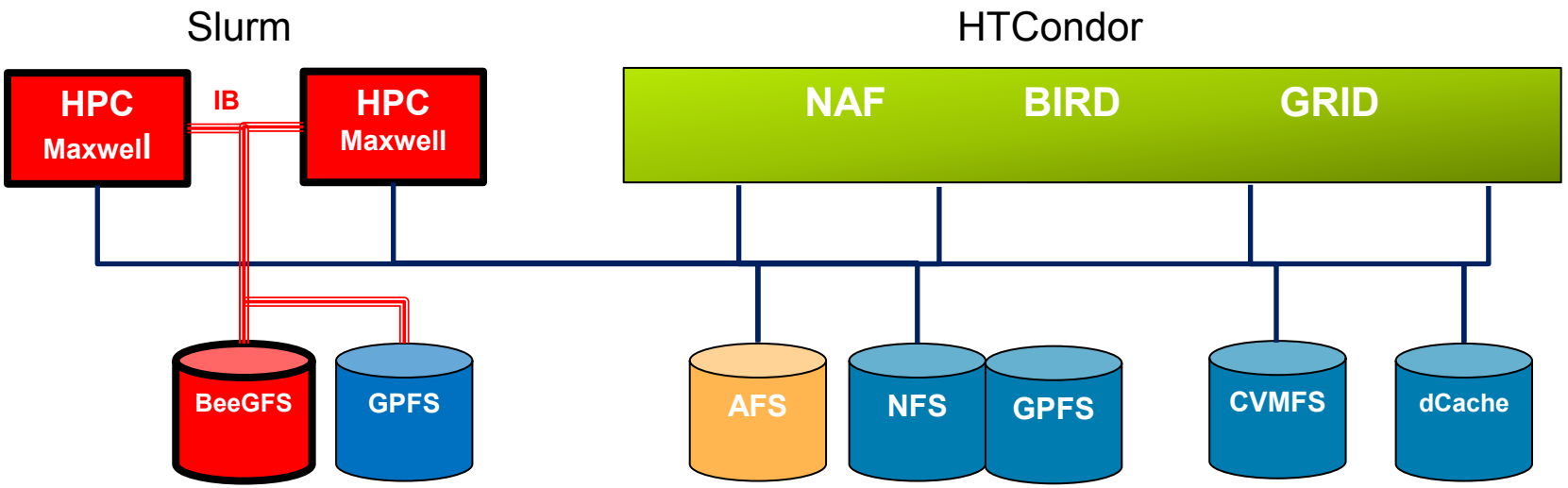
HTCondor is a scheduling and batch system. Main features:

- 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)

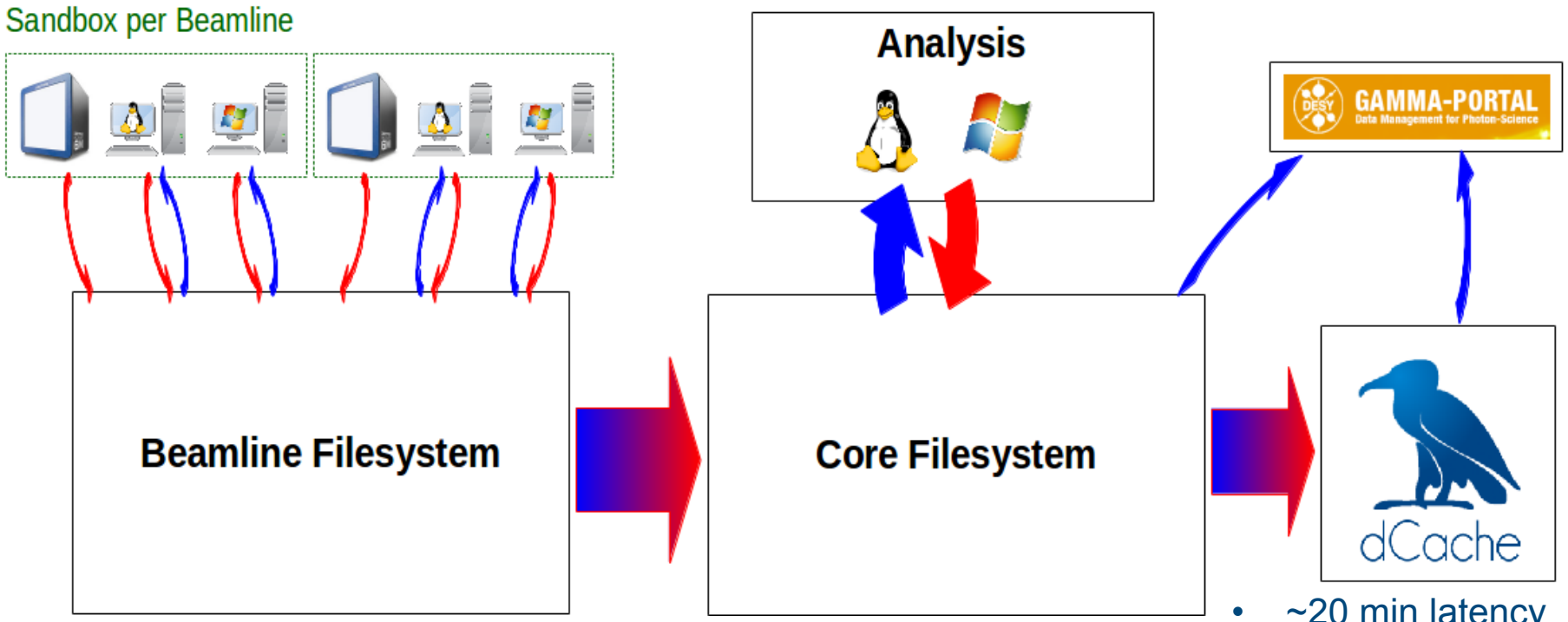
- IceCube simulation production – Grid and local farm computing
  100 GPUs for photon propagation in ice
- Hosting of filtered data (tape-backed), 50% of simulated data (disk-only)
- Continuous data transfer UW Madison --> Zeuthen (300MB/s)
- Acting as disaster recovering center

CTA Science Data Management Center (tbd.)

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

Data Taking Petra3

Logical Dataflow



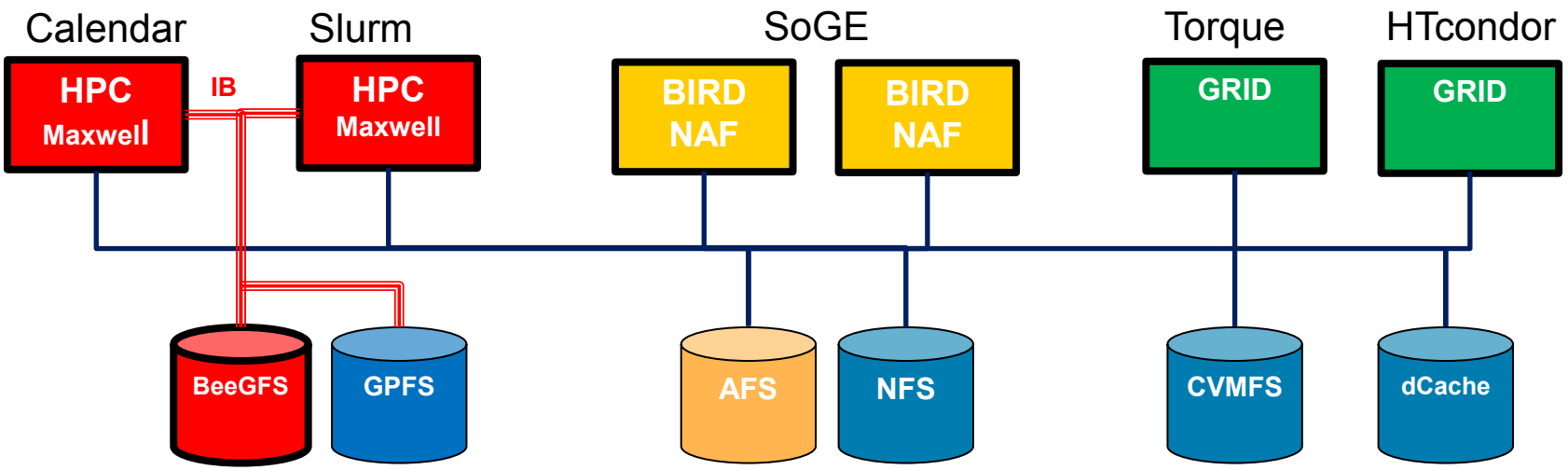
- Low latency
- Low capacity
- Host-based authentication

- 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

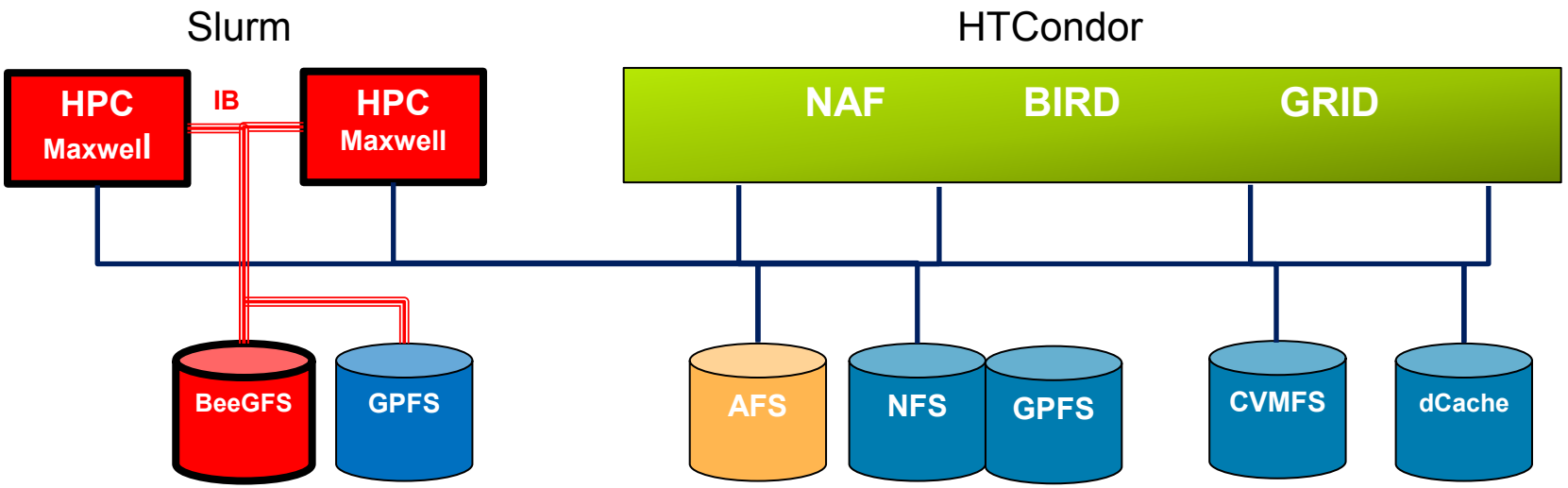
HTCondor is a scheduling and batch system. Main features:

- 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)

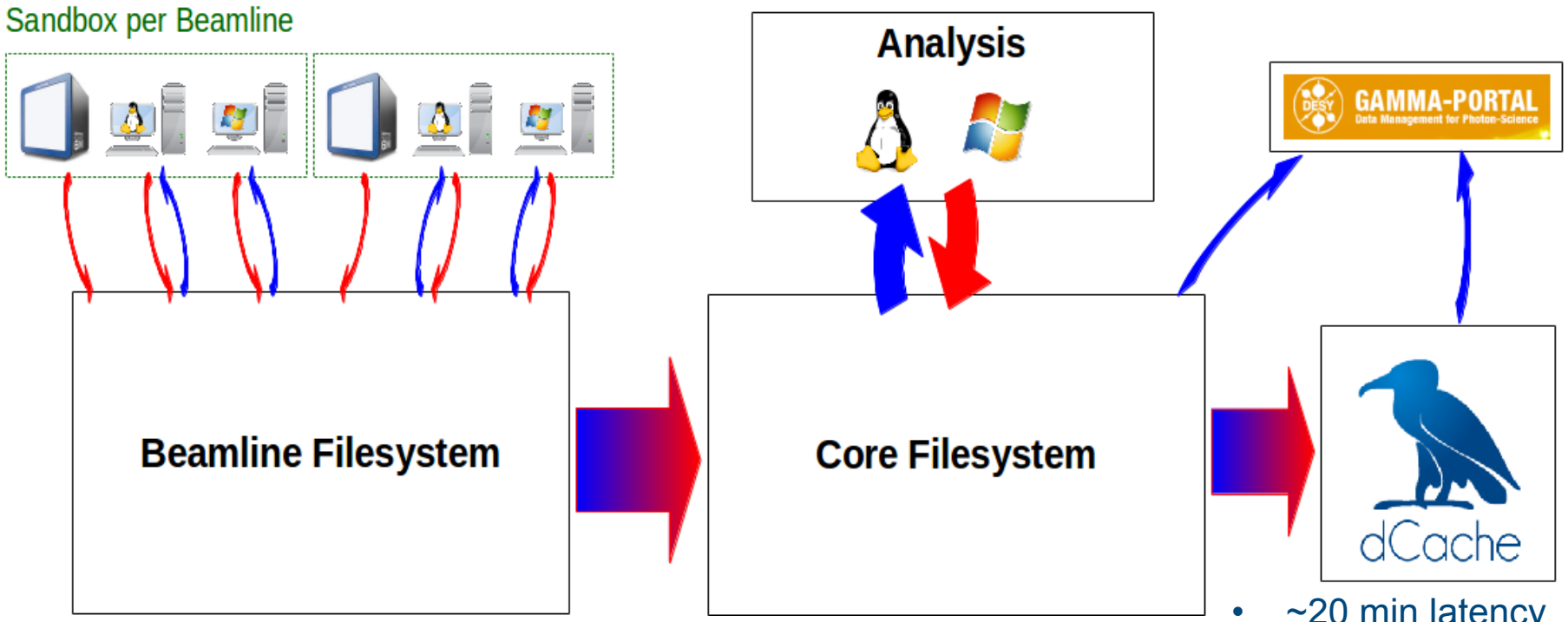
- IceCube simulation production – Grid and local farm computing
  100 GPUs for photon propagation in ice
- Hosting of filtered data (tape-backed), 50% of simulated data (disk-only)
- Continuous data transfer UW Madison --> Zeuthen (300MB/s)
- Acting as disaster recovering center

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

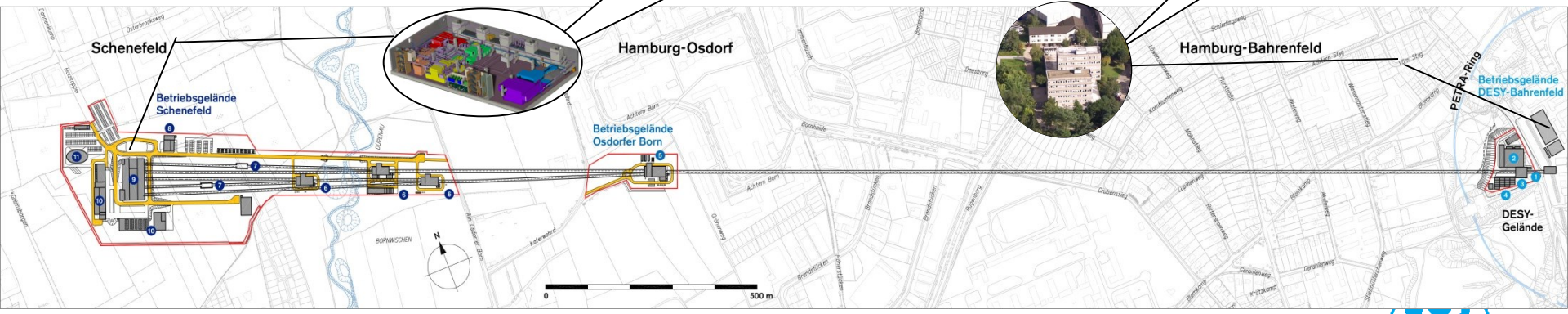
- ~20 min latency
- Very high capacity, tape
- Full user authentication

Preparation for European XFEL

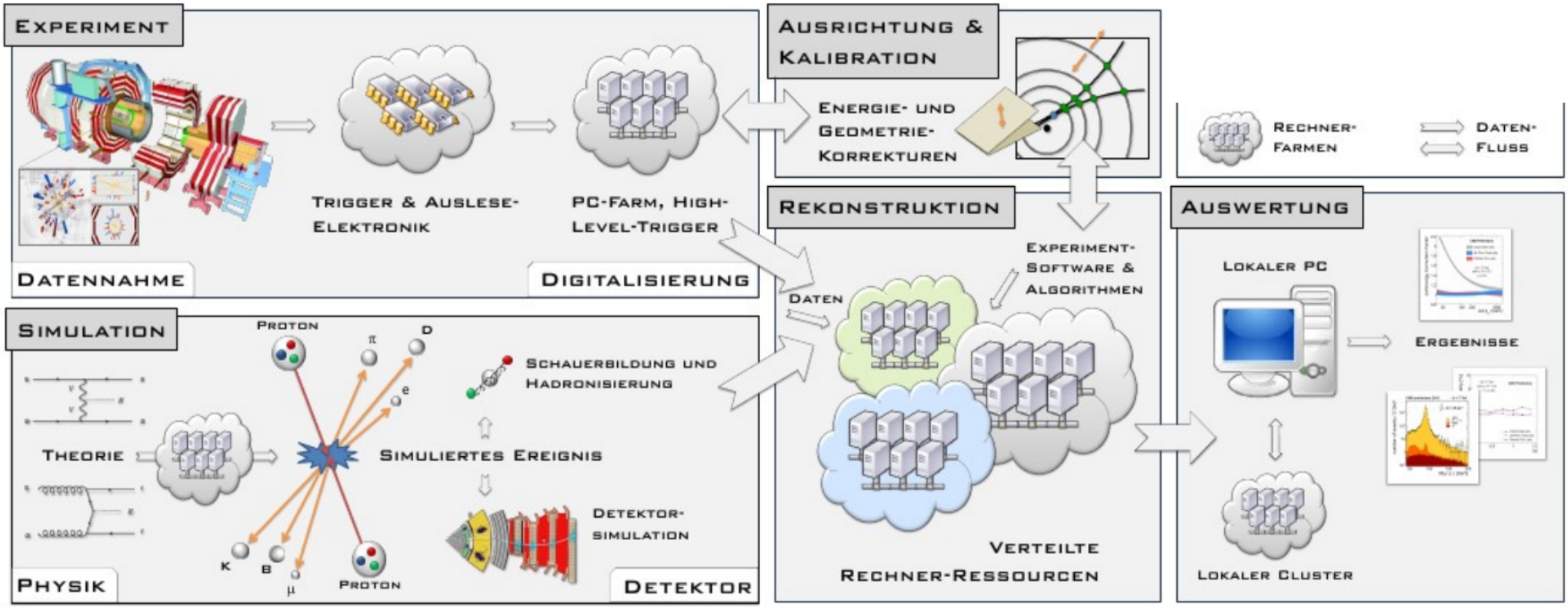
- Use the system developed for PETRA III
- as a blueprint for XFEL
- Setup of distributed GPFS system in
- Schenefeld and at DESY
- Management of the dataflow
- High-performance and reliable coupling
- between the locations

GPFS system for online storage and computing

GPFS system for offline storage and computing

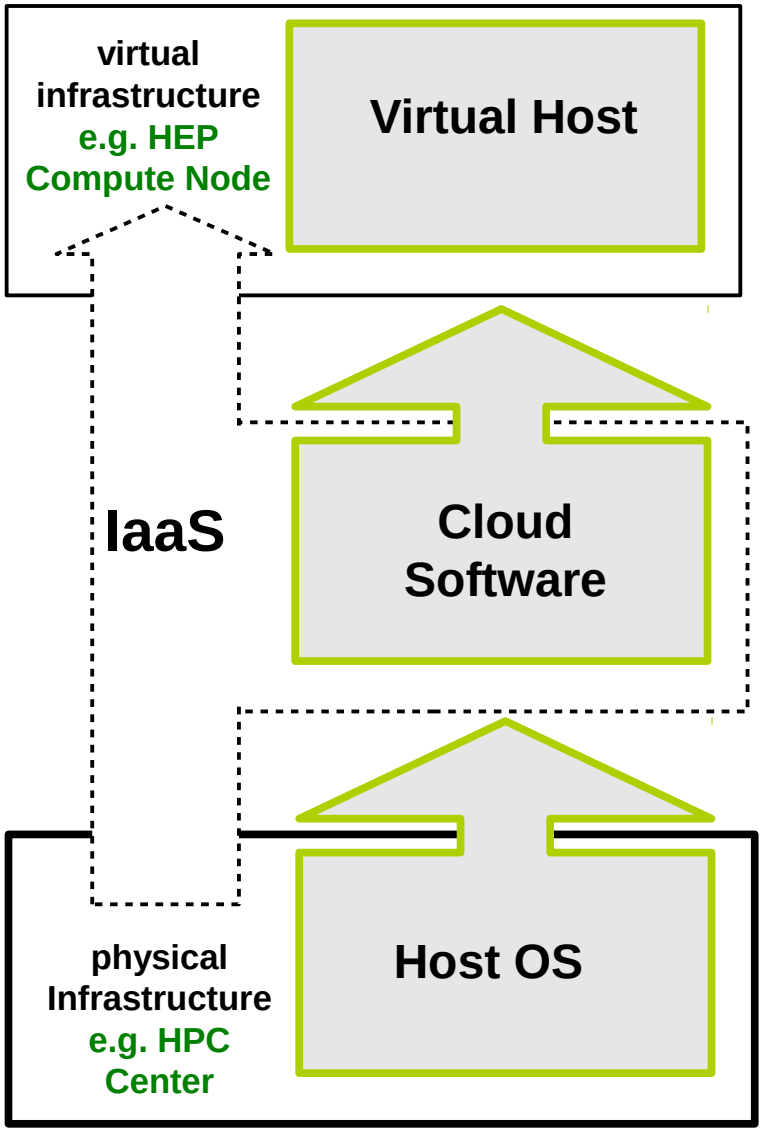


Diverse Computing in HEP



- 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)

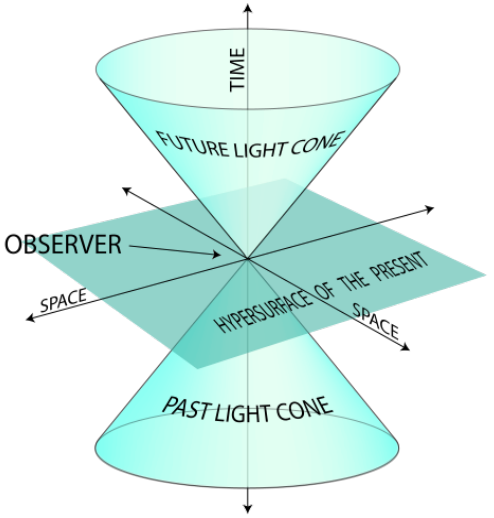
- Used for gird job submission and data management by LHCb and Belle II [1]

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

要約

GoeGrid Tier-2 in Göttingen

Deutsches Elektronen SYnchrotron



Minkovsky space-time

Hisenberg

Göttingen ってどこ？

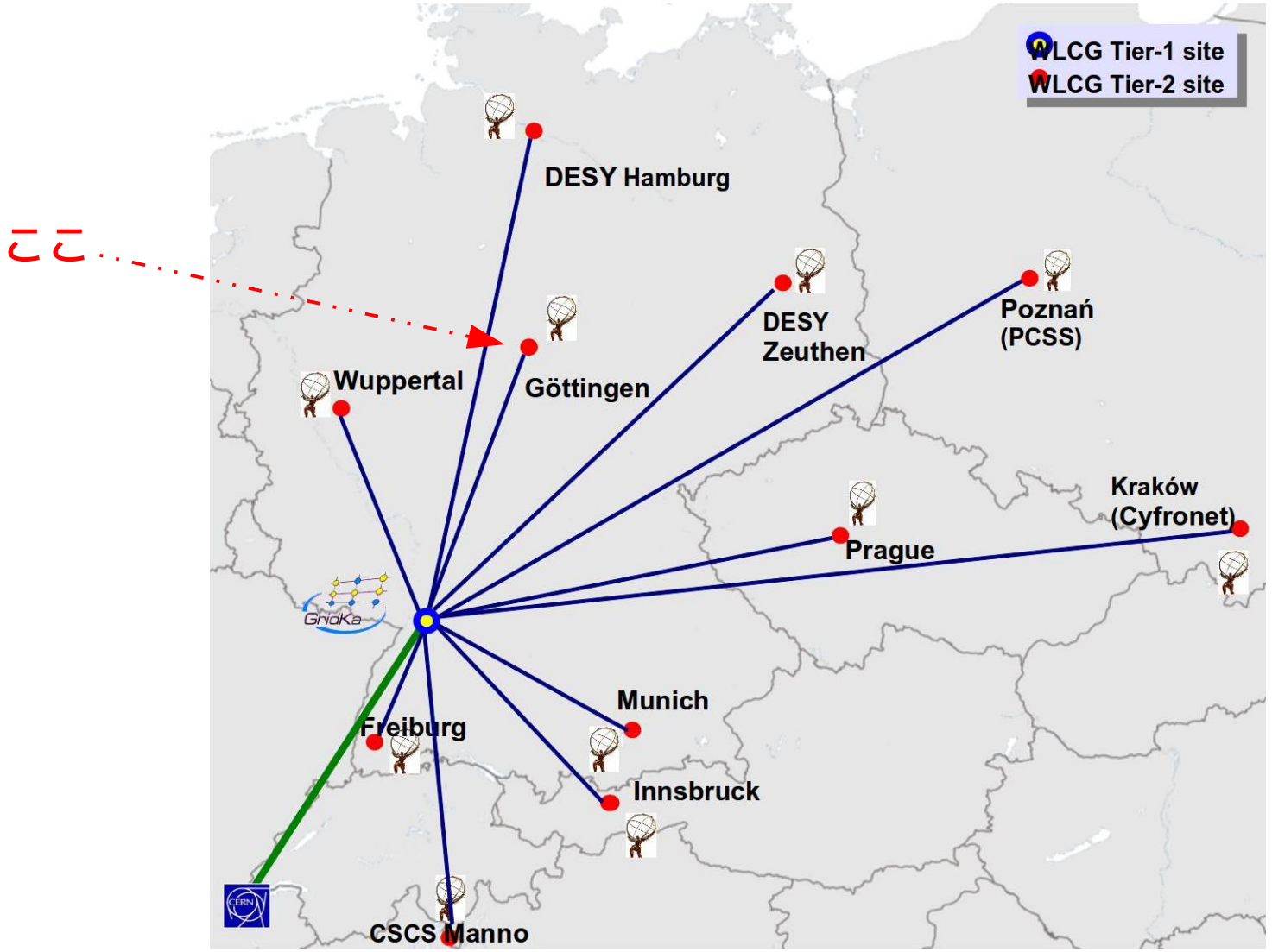
- ドイツ連邦共和国ニーダーザクセン州
  - 人口約３万人の大学都市、ドイツエリート大学の一つ
  - 高地ドイツ語（標準ドイツ語）の中心
  - 近くの都市はカッセル、ハノーバー
  - ネアンデルタール人が住んでた洞窟等が近くにある
- Wikipedia: George-Augstust-Universität Göttingen より
  - 伝統的に物理・数学・哲学が強い
  - ドイツでは最大数のノーベル賞受賞者を輩出
  - 中道左派、緑の党の中心
  - 物理の教科書で有名な所ではハイゼンベルク、ボルン、プランク、ミンコフスキー、リーマン、ガウス等
  - 日本人だと理研所長の仁科芳雄氏がいたところ
  - ケンブリッジ大学との紳士協定により、他の街とは違い第二次世界大戦での破壊を免れる
- 戦前は核物理学の世界的中心地 → ナチスにより地位を失墜

概要、大学 Tier-2 計算機センターの役割

- GoeGrid はドイツ大学用 ATLAS 広域計算機の一つ
  - 教育・研究用インフラ。
  - グリッド・クラスタのノウハウの普及
  - ドイツ大学計算機トータルで ATLAS ドイツ全体の約２０％ほどの計算機資源を供給
  - 書類上は DESY-HH との連合サイト。システム設定など共通化可能
    - 現在は独立運営
- LHC ATLAS Tier-2 サイトの役割
  - MC プロダクション
  - No Tape ドライブ
  - より公的なインフラの側面
    - 使用可能時間をより長くとる
    - High Through Computing (HTC)
    - High Performance Computing (HPC) ではない
      - 使用可能時間 ＞ 性能
    - いくつかの大学では実験的に HPC システムを統合中

GoeGrid Tier-2

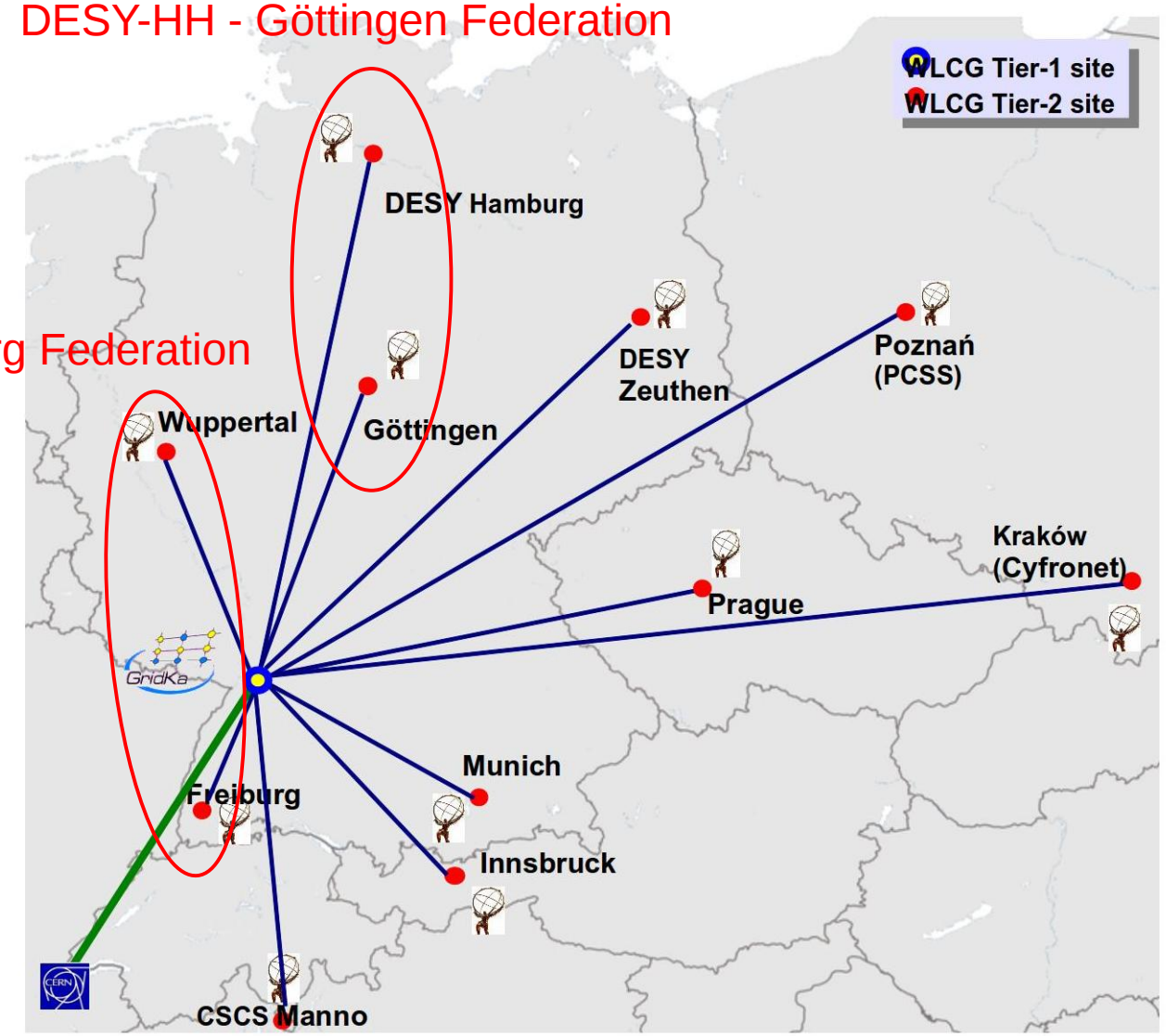
概要、大学 Tier-2 計算機センターの役割



概要、大学 Tier-2 計算機センターの役割

DESY-HH - Göttingen Federation

Wuppertal - Freiburg Federation



計算機施設

- Göttingen 大学 Max-Plank-Institut のすぐ隣
  - フロアスペース、クーリング設備、電力、簡単なハードウェアメンテナンスは GWDG

GWDG
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

マンパワー

- 運用は 3 人（ LHC 実験のための実質要員は 2 人）
  - 合計 1 FTE 程度？

? FTE

ハードウェアメンテ
GWDG: Tim Ehlers

0.3FTE

Local Batch System
Teory group: Dr. Juergen Holm

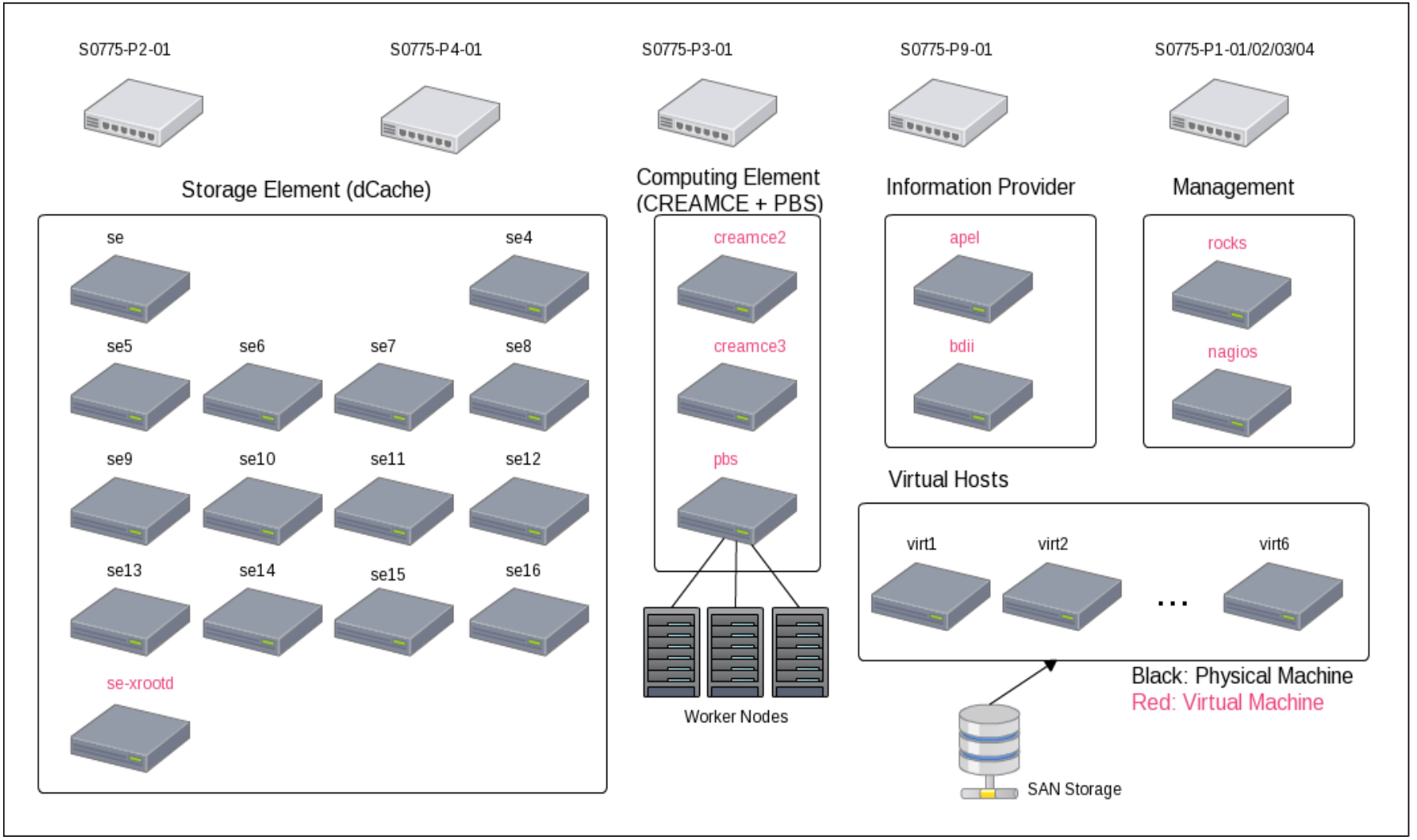
0.6FTE

Grid, Storage, ATLAS exp
ATLAS: Dr. Gen Kawamura

現状の計算資源と将来

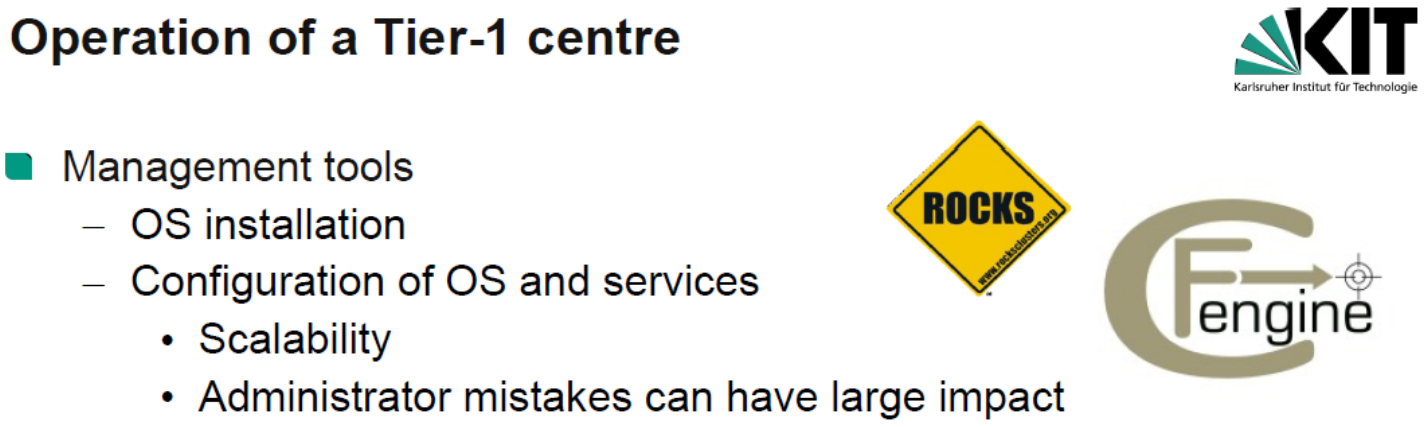
- GoeGrid is a WLCG ATLAS tier-2 resource centre Inaugurated in 2008 Usage of resources in the elds of astrophysics, biomedical sciences, grid development, high energy physics, theoretical physics, and the humanities Resource sharing based on the amount of the contribution Hosted at GWDG Part of the 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)

構成


S0775-P2-01
S0775-P4-01
S0775-P3-01
S0775-P9-01
S0775-P1-01/02/03/04
Storage Element (dCache)
Computing Element
(CREAMCE + PBS)
Information Provider
Management
se
se4
se5
se6
se7
se8
se9
se10
se11
se12
se13
se14
se15
se16
se-xrootd
creamce2
creamce3
pbs
Worker Nodes
apel
bdii
rocks
nagios
Virtual Hosts
virt1
virt2
virt6
Black: Physical Machine
Red: Virtual Machine
SAN Storage

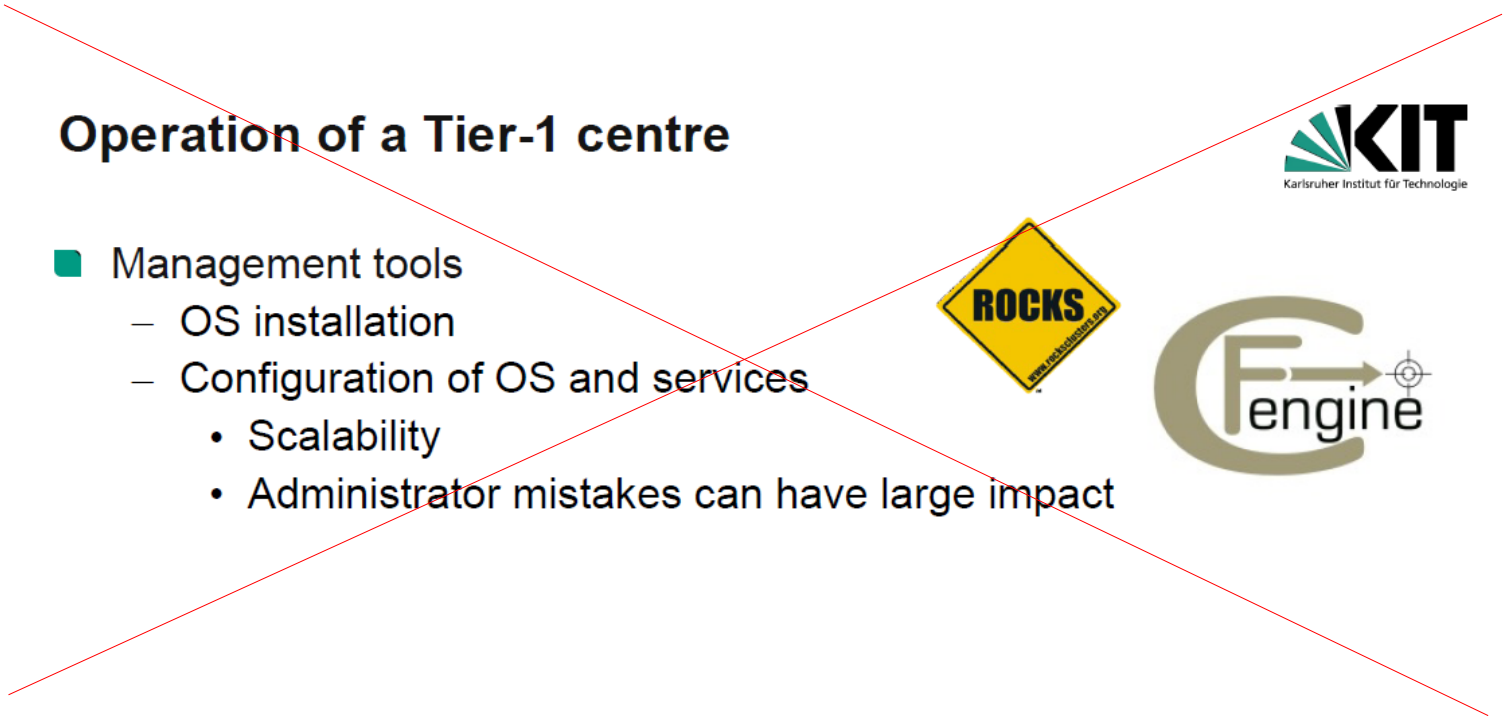
GoeGrid 1.0

- 過去の構成（Cfengine + ROCKS)
  - FZK, DESY-HH に倣った構成



GoeGrid 2.0

- Foreman + Puppet
  - HTCondor, ARC-CE + CREAMCE



～ 2018 ハードウェア予定

可視化

EGI クラウド

EGI クラウド

クラウド

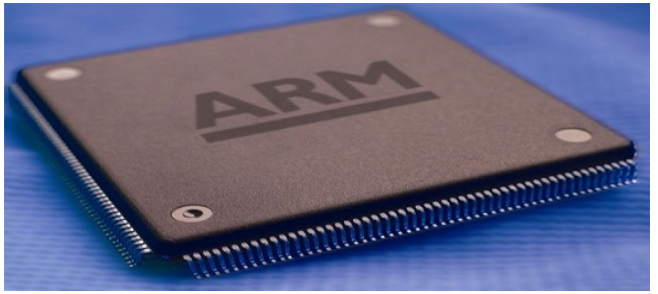
クラウド

GoeGrid Tier-2 まとめ

- ドイツの中堅 Tier-2 サイト
  - 5 ～ 8% 程度の ATLAS 計算資源を供給

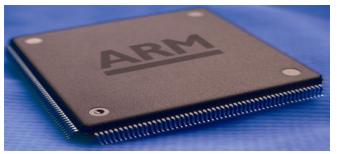
Göttingen ATLAS 物理計算グループの研究トピックス

Cloud Computing



方向

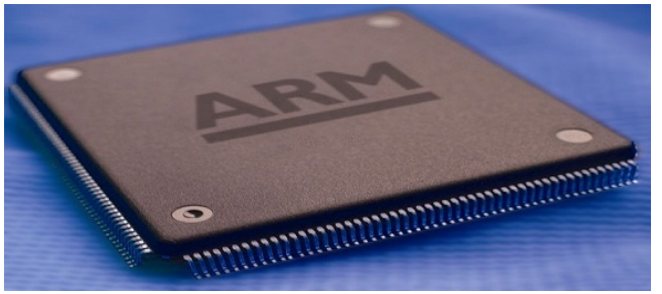
- 将来必要になりそうなもの
- （高エネ解析の視点から） ATLAS の役に立ちそうなもの
- 実用性・有用性・汎用性があるもの
- コンピューター市場の動向に一致しているもの
  - ミーハーは市場無視より良い
- 面白そうなもの
- 新規性があるもの
- *そのうち自分たちが楽が出来そうなもの*



ATLAS ソフトウェア資源の ARM アーキテクチャへの移植

- モチベーション
  - ATLAS のソフトウェア資源は Intel x86_64 で構築されている
  - ARM アーキテクチャへの移植
- 利点
  - ARM プロセッサは既に巨大なマーケット
    - 例　スマートフォン、タブレット、 IoT デバイス
  - 低消費電力
  - ARM クラスタを利用できる
  - 将来性あり
    - ソフトバンクが多大な投資をする？
- 欠点
  - ATLAS のソフトウェア資源は巨大すぎる
  - 移植のみでも大仕事

MSc. Joshua Wyatt Smith



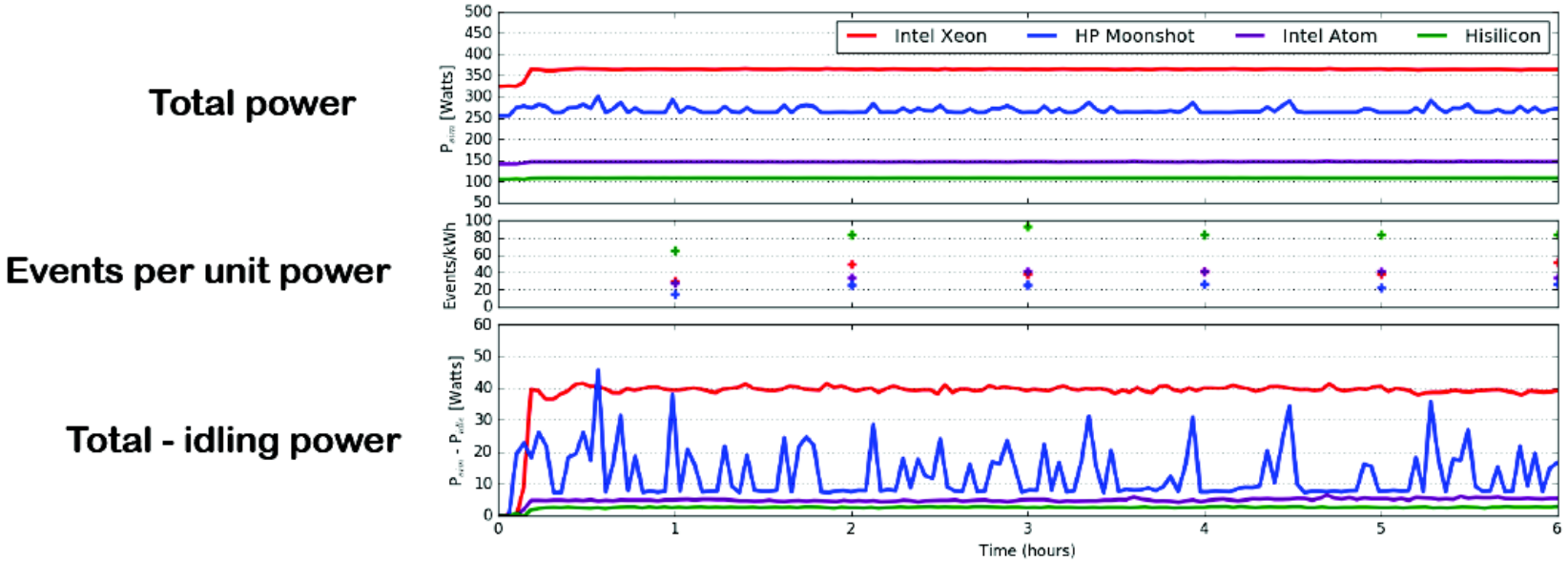
ATLAS ソフトウェア資源の ARM アーキテクチャへの移植

- What did we port?
  - Athena stack overkill
  - Picked AthSimulation
    - A fraction of the packages of Athena (~345 compared to ~2400)
    - Much quicker compile time
    - Potential for errors in port decreases
    - Geant4 gives a good CPU load
    - Good for simulation and validation
    - Implement this in Jenkins
    - Build this using CMake - we were pioneers

      Avoid CMT

ATLAS ソフトウェア資源の ARM アーキテクチャへの移植

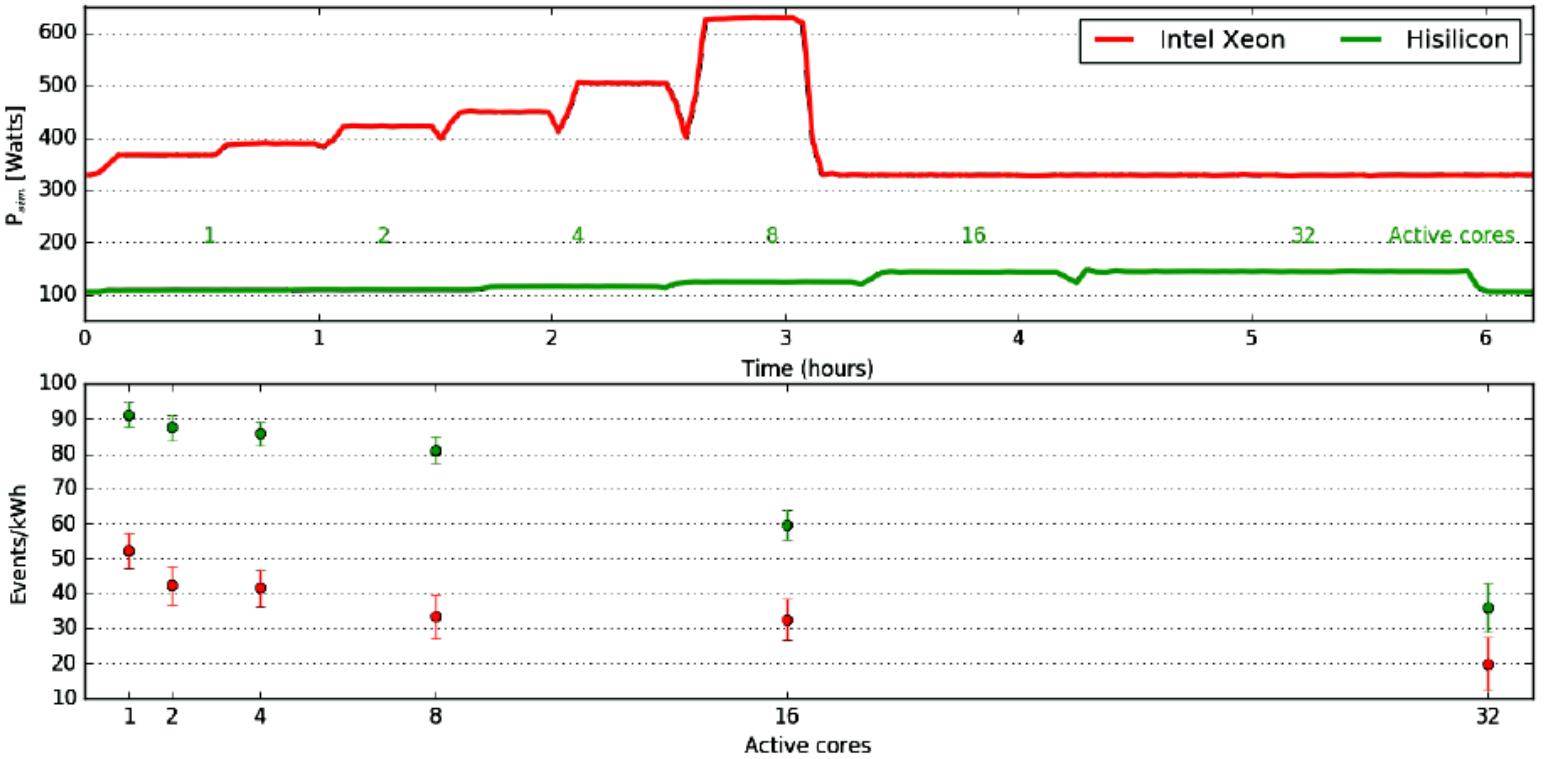
- ttbar イベント生成と消費電力



Name	Time (Hours)
HP Moonshot	15.10
Hisilicon	10.46
Intel Atom	18.03
Intel	6.33

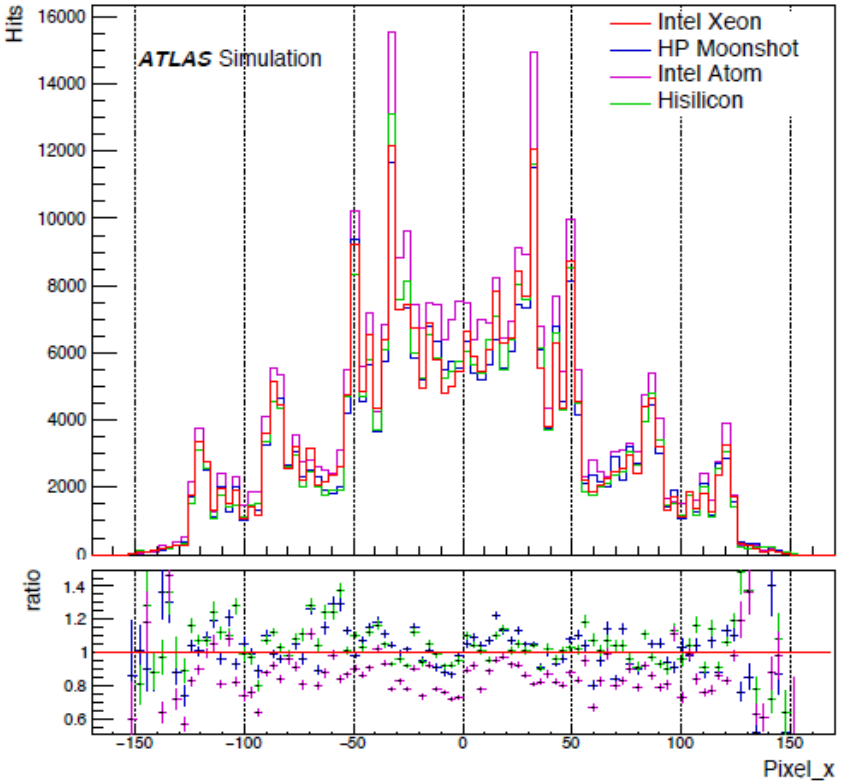
ATLAS ソフトウェア資源の ARM アーキテクチャへの移植

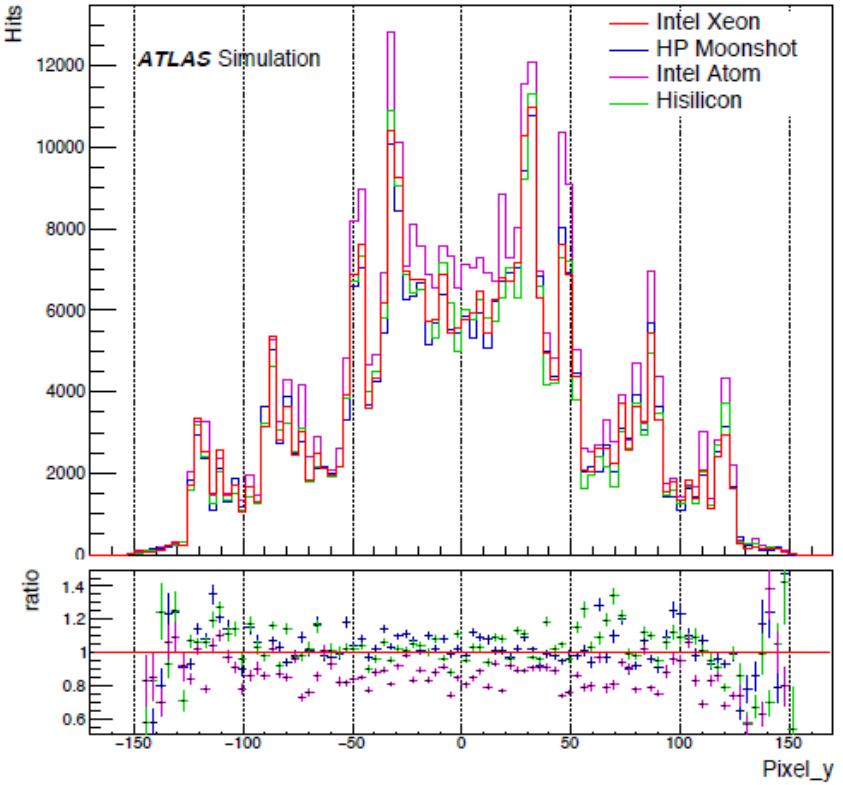
- Event 生成と実行時間・消費電力



ATLAS ソフトウェア資源の ARM アーキテクチャへの移植

- Event シミュレーション





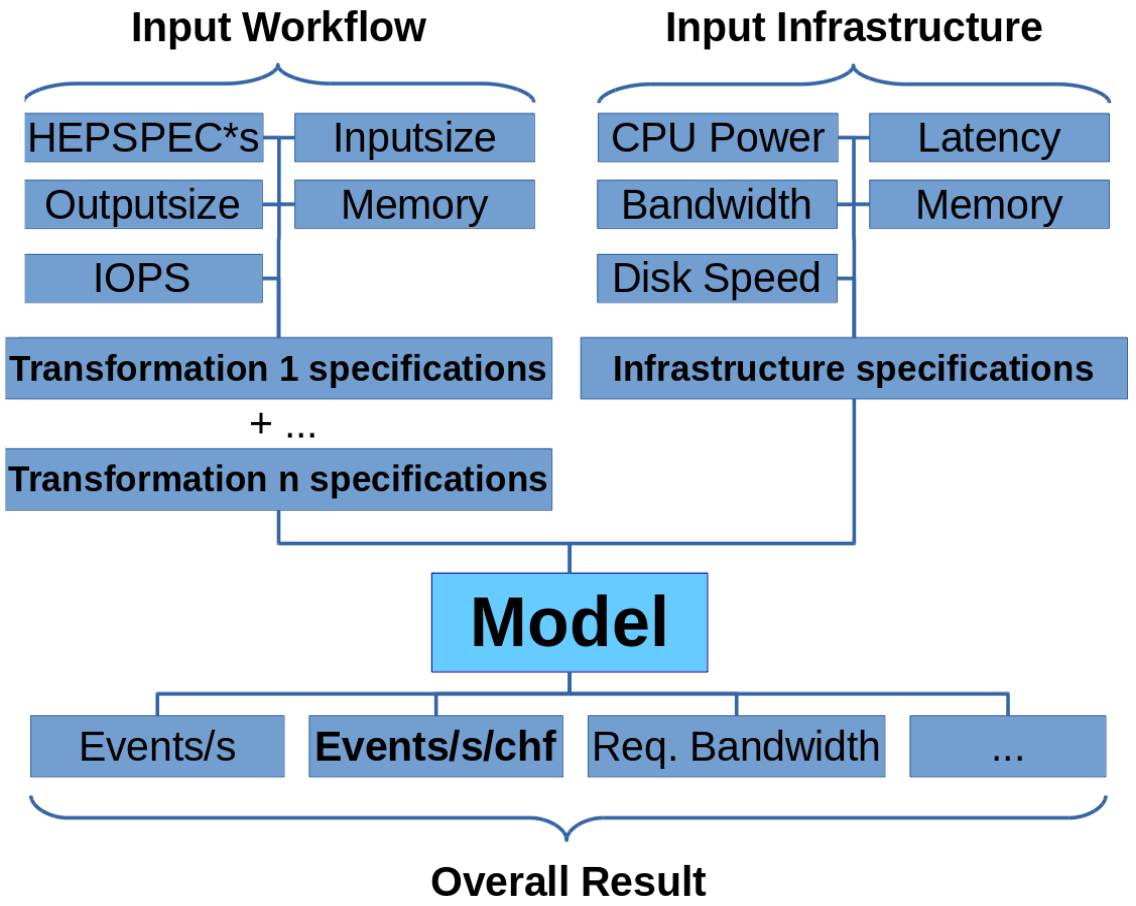
クラウドコンピューティング

- モチベーション
  - 商用クライドに最適な ATLAS の計算モデルは？
  - 商用クラウドのコスト評価モデル
    - 各サイトを持つということの妥当性自体を評価
- 利点
  - 商用クラウドマーケットは急成長中
  - 将来的に計算単価をさらに劇的に下げる可能性あり
- 欠点
  - データ依存型ジョブは当面考えない
    - ストレージ資源とストレージ IO の問題
  - LHC に蓄積した人的資源、ノウハウを将来的に失う可能性あり

MSc. Gerhard Fredinand Rzehorz

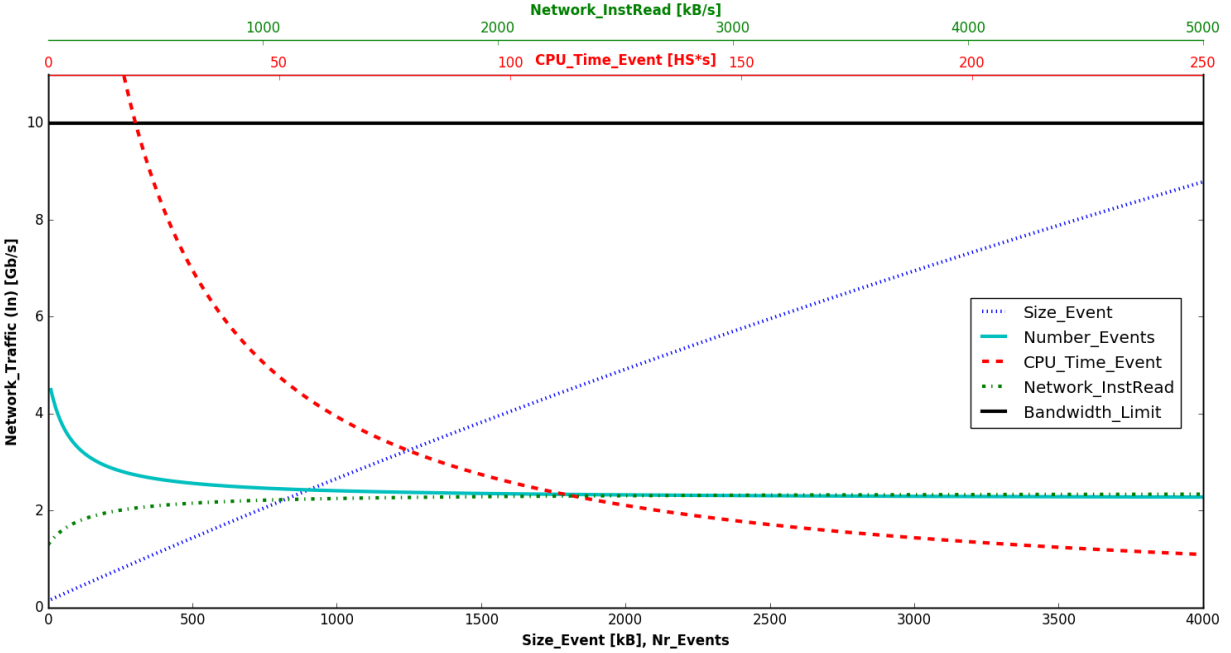
クラウドコンピューティング

- ジョブパラメタからイベントレートやコスト (CHF) などを算出



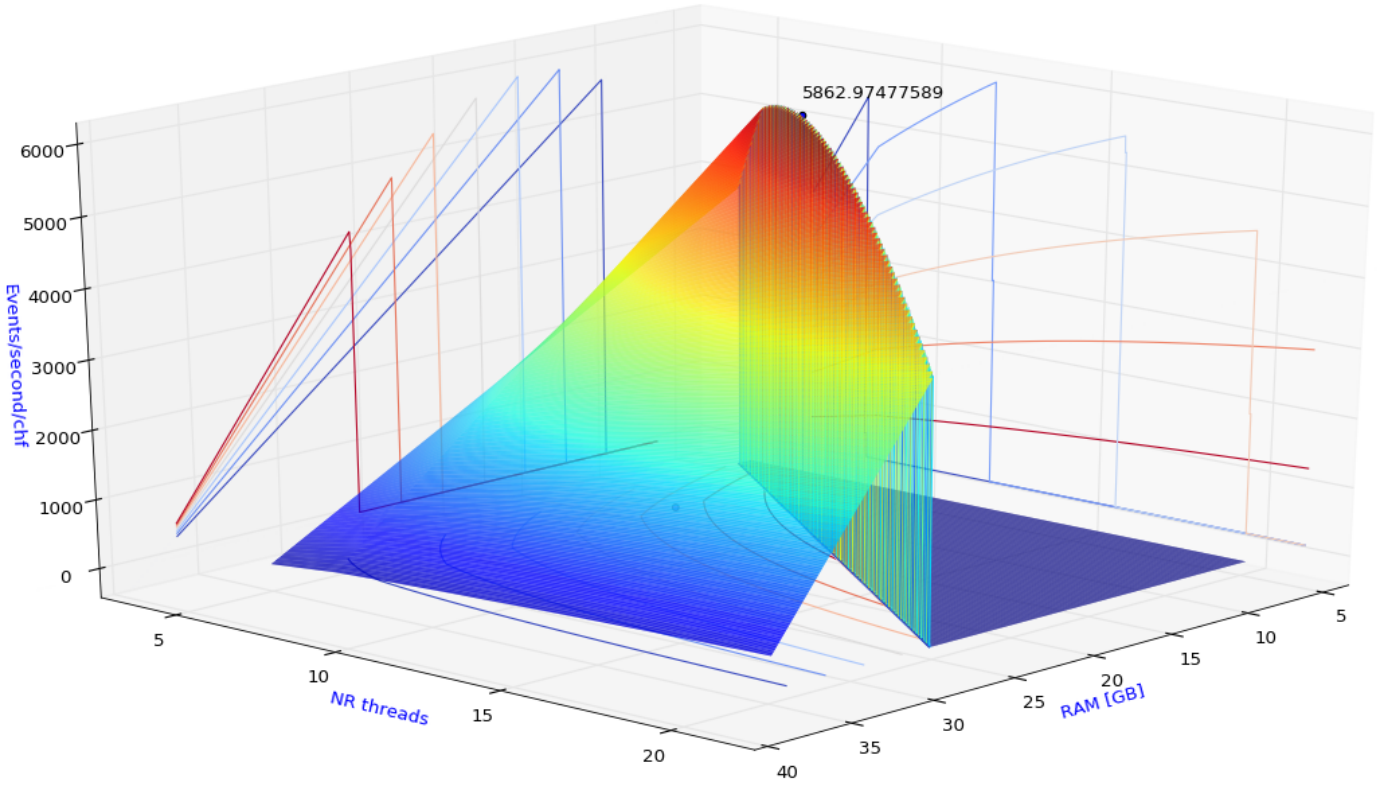
クラウドコンピューティング

- 最適値を探索
  - Overcommitting
  - I/O 待機時間を減らす



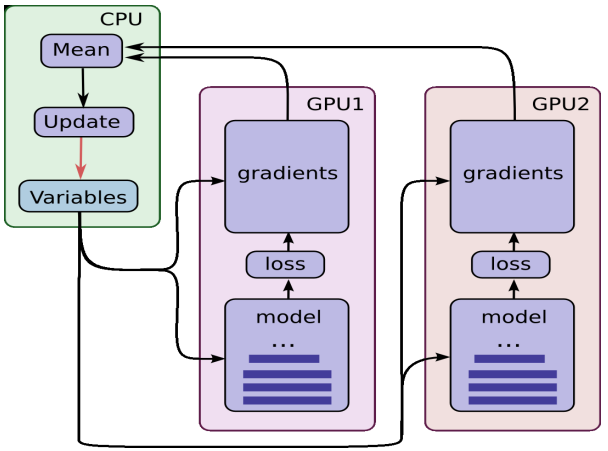
ATLAS Real Data Reconstruction					
Number of processes	RAM [GB]	Data location	Overall node throughput [s/event]	Overcommit improvement [%]	Duration improvement to standard [%]
8	32	BNL	4,19 ± 0,05		-32
2x8	32	BNL	2,55 ± 0,01	39	19
8	16	BNL	4,31 ± 0,08		-36
2x8	16	BNL	3,51 ± 0,08	19	-11
8	32	local	3,07 ± 0,04		3
2x8	32	local	2,24 ± 0,01	27	29
8	16	local	3,17 ± 0,09		0
2x8	16	local	3,33 ± 0,01	-5	-5

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

- TensorFlow とは？
  - 2015 年末にリリースされた Google の最新分散処理用ディープニューラルネット（ DNN ）用ライブラリ
    - DNN に特化ではなく、一般的分散計算処理ライブラリ
- モチベーション
  - WLCG グリッド・クラスタで動かすには？ (GPU なし)
- Google Cluster = 巨大 GPU cluster ?



Google TensorFlow ライブラリと分散コンピューティング

- DNN っていいの？
- Low-level vs High-level 特徴量
  - Low-level は特徴量抽出前の物理パラメタ

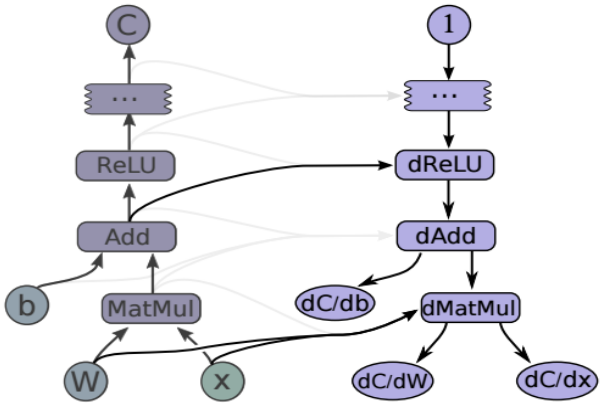
Table 1 | Performance for Higgs benchmark.

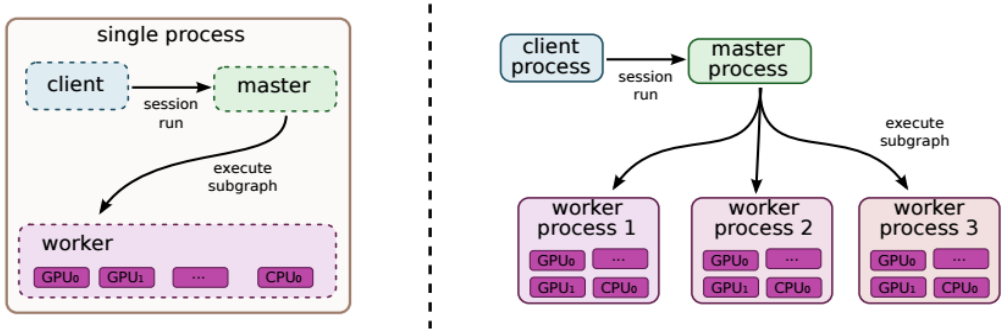
Technique	Low-level	High-level	Complete
AUC			
BDT	0.73 (0.01)	0.78 (0.01)	0.81 (0.01)
NN	0.733 (0.007)	0.777 (0.001)	0.816 (0.004)
DN	0.880 (0.001)	0.800 (<0.001)	0.885 (0.002)
Discovery significance			
NN	2.5σ	3.1σ	3.7σ
DN	4.9σ	3.6σ	5.0σ

*) Baldi, Pierre, Peter Sadowski, and Daniel Whiteson. "Searching for exotic particles in high-energy physics with deep learning." Nature communications 5 (2014).

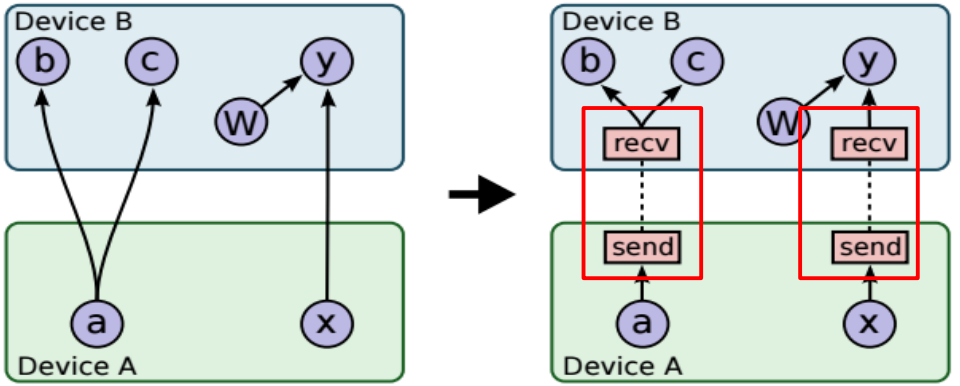
Google TensorFlow ライブラリと分散コンピューティング

- TensorFlow のアーキテクチャと設計
  - Graph ベースの計算クラス定義・記法
  - 変数は Tensor
  - Gradient は分割して一括実行（ Flow ）
  - Graph を複数計算デバイスへ分割可能
- 当初 GPU モードのみサポート、 CPU モードは v0.8 （今年夏）以降





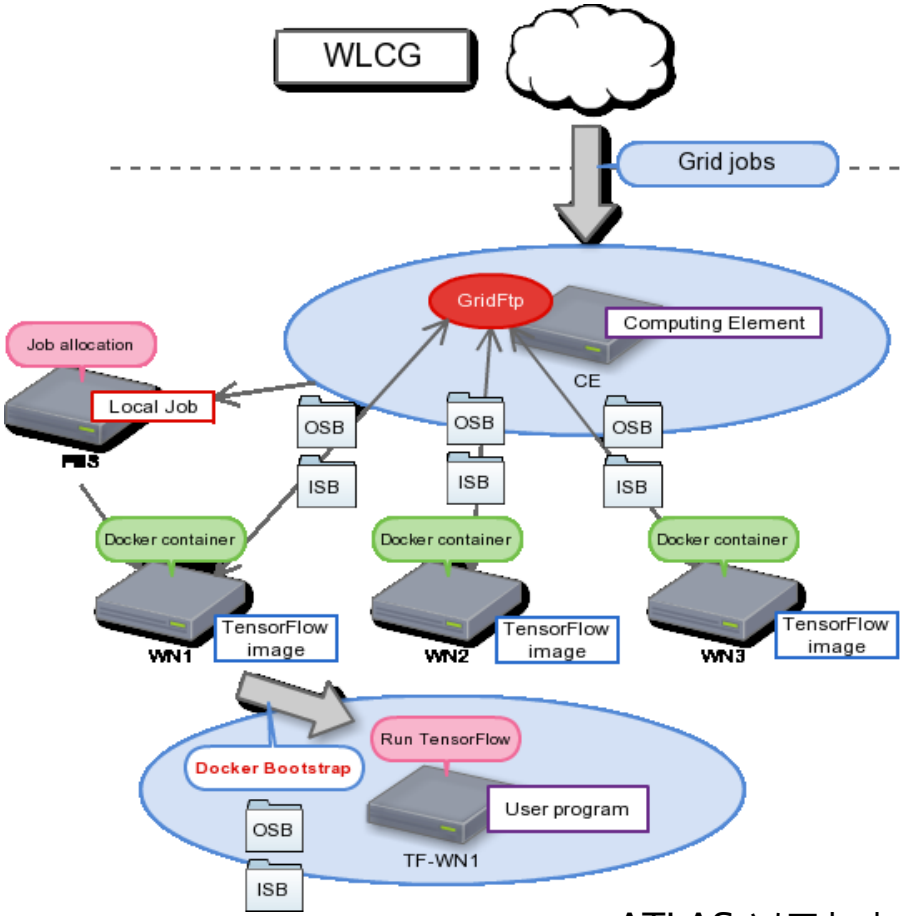
計算処理の分割実行と結合



計算デバイスへ分割

Google TensorFlow ライブラリと分散コンピューティング

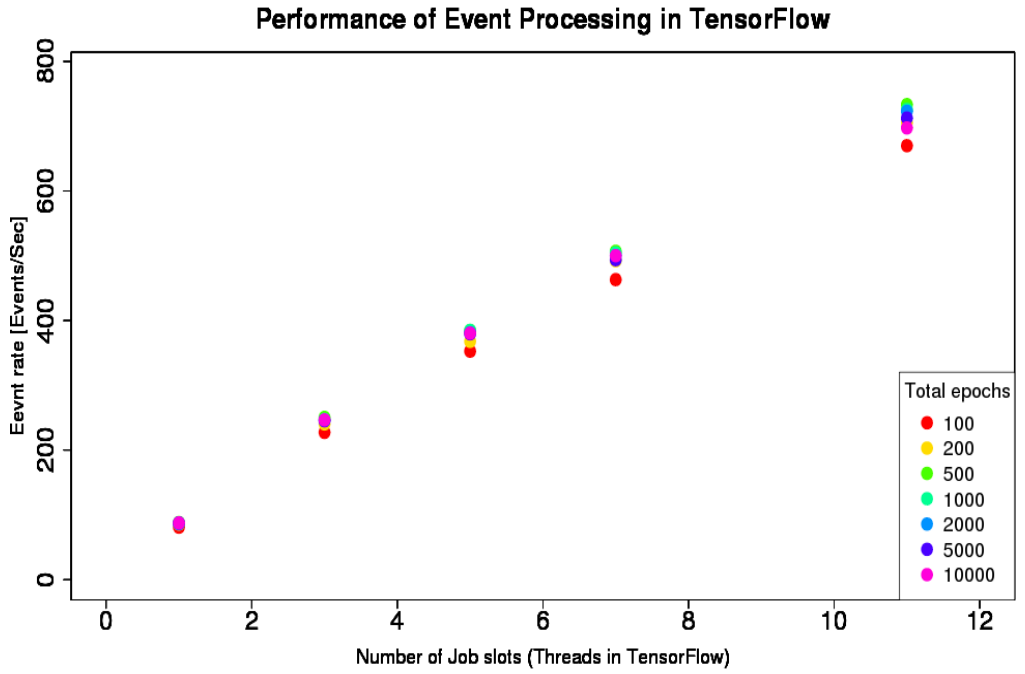
- Grid Docker 環境用テストキュー
  - Grid CE から自作ブートストラップで Docker 用 TensorFlow 環境を計算ノードへロード

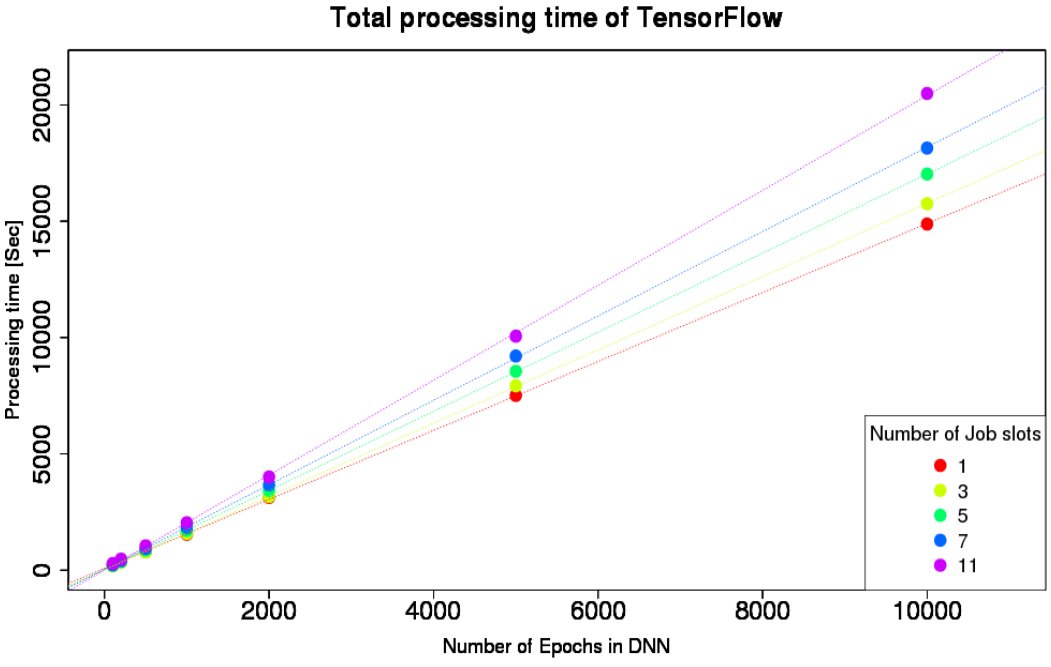


- プロダクション・レベルでの考慮点
  - アカウンティング（使用量の記録）
  - 安全性 (！)
  - 堅牢性

Google TensorFlow ライブラリと分散コンピューティング

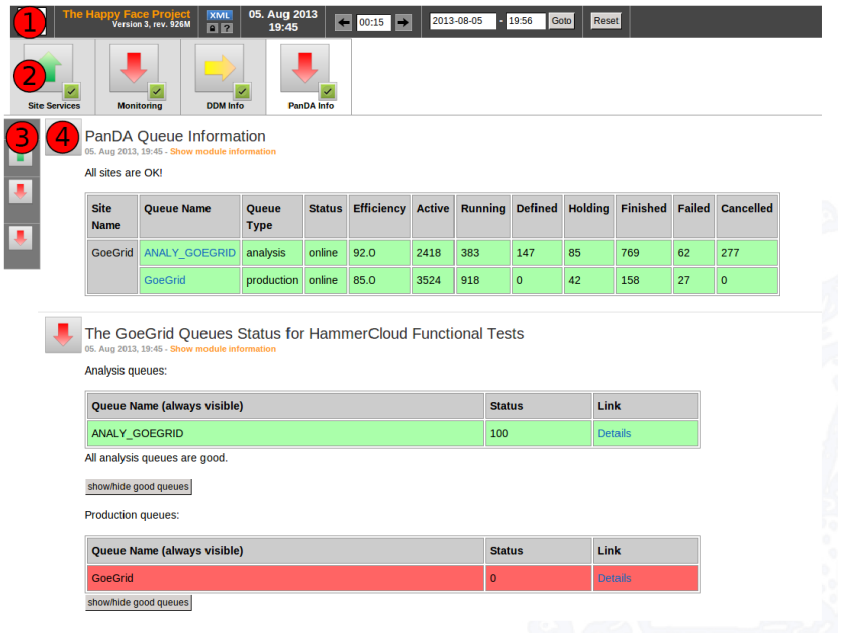
- Convolutional DNN でのイメージ識別学習時間
  - TensorFlow v0.8 CPU モードでテスト (1 thread / 1 docker VM)
  - Event processing rate はリニアに上昇
  - Event processing 以外での（通信）遅延が大きい
    - > 10 jobs で TensorFlow (CPU mode) 自体がまだ不安定。





メタモニタリングシステム（ HappyFace, MadFace ）

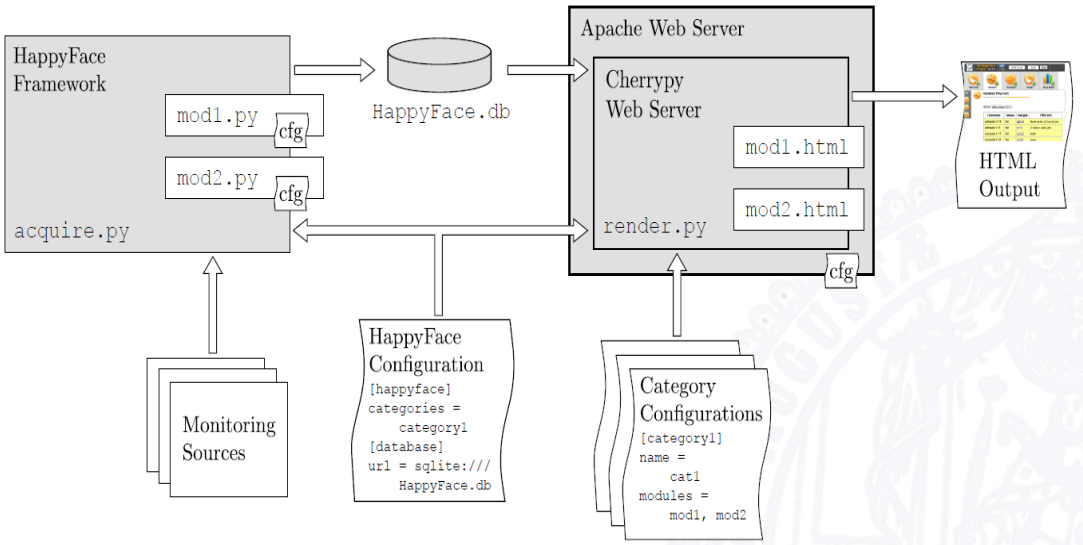
- メタモニタリングシステムとは？
  - モニタリングのモニタリングのためのシステム
- HappyFace meta-monitoring system in CMS
  - ドイツ Tier1 モニタリングのためのメタモニタリングシステムとして開発 (KIT)
    - Python フルで実装
    - SQL DB backend
    - 1 web server
    - モジュラー構造



メタモニタリングシステム（ HappyFace, MadFace ）

- メタモニタリングシステムとは？
  - モニタリングのモニタリングのためのシステム

- HappyFace meta-monitoring system in CMS
  - ドイツ Tier1 モニタリングのためのメタモニタリングシステムとして開発 (KIT)
    - Python フルで実装
    - SQL DB backend
    - 1 web server
    - モジュラー構造

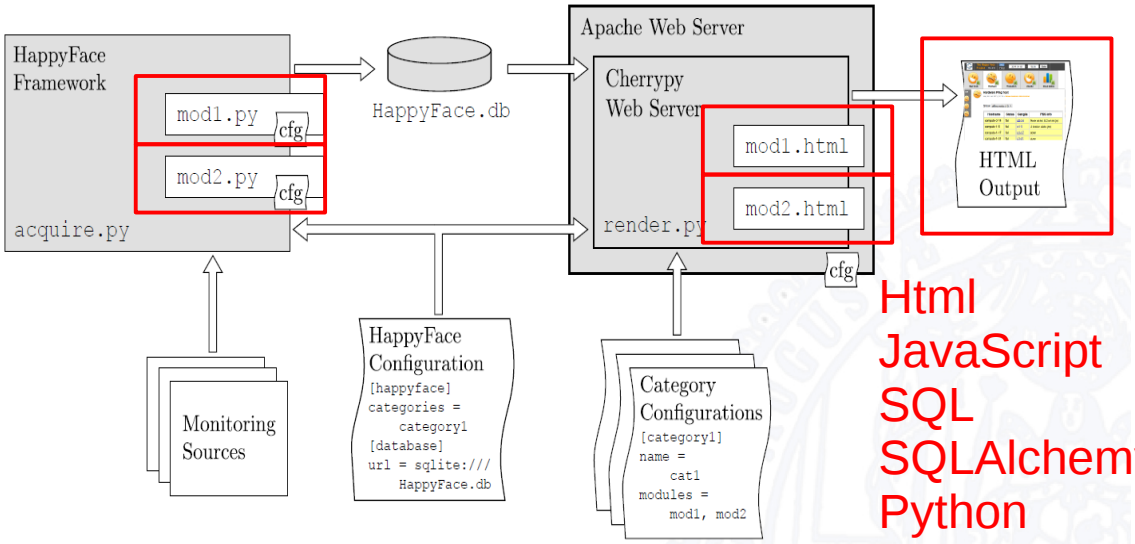


メタモニタリングシステム（ HappyFace, MadFace ）

- メタモニタリングシステムとは？
  - モニタリングのモニタリングのためのシステム

- HappyFace meta-monitoring system in CMS
  - ドイツ Tier1 モニタリングのためのメタモニタリングシステムとして開発 (KIT)
    - Python フルで実装
    - SQL DB backend
    - 1 web server
    - モジュラー構造

本質的に常に開発・メンテの労力が必要な箇所



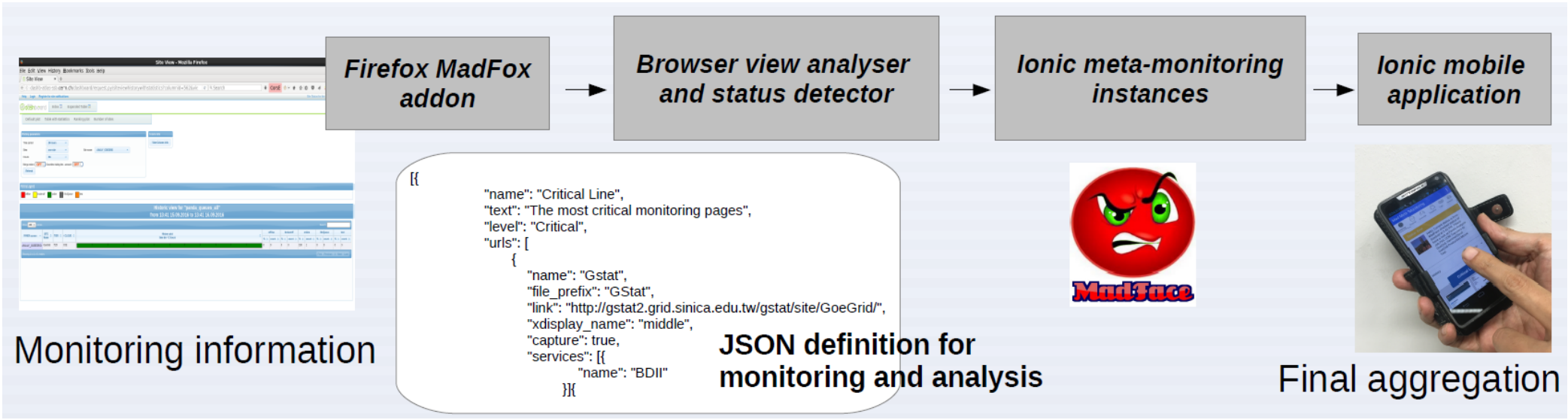
メタモニタリングシステム（ HappyFace, MadFace ）

- MadFace メタモニタリングシステム
- 最新のモバイルフレームワークで開発
  - Web + mobile フレームワーク
    - Ionic, AngularJS framework
    - Server-side JavaScript technology
    - Firefox + ***Madfox*** (JetPack Manager)
  - オンラインベイズ分析機
    - R, bayesian change process
    - Bayesian network
- コーディング時間
  - 約 300 時間

グリッドクラスタのモバイル管理

メタモニタリングシステム（ HappyFace, MadFace ）

- Web サーバーと Mobile アプリが同じフレームワークなので開発時間が大幅短縮
  - 依存言語は JavaScript と R
  - データ定義は JSON

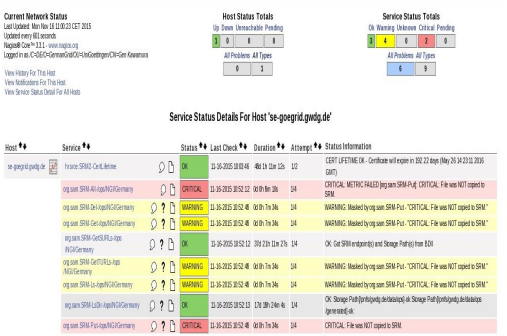


メタモニタリングシステム（ HappyFace, MadFace ）

- 開発やシステム管理等に労力のかかる部分（情報取得と状態識別）を完全自動化

ブラウザ画面 t1

change



ブラウザ画面 t2

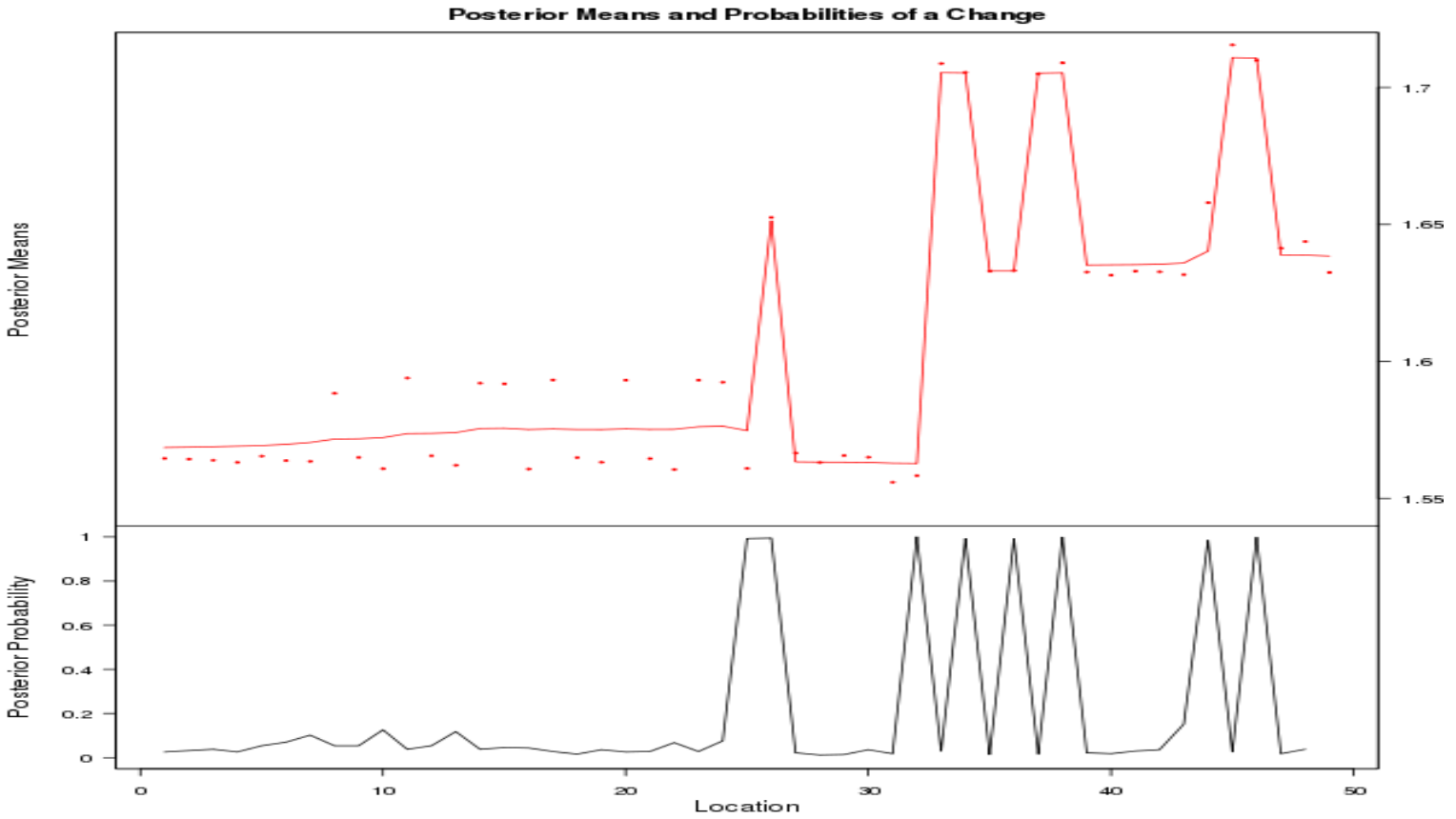
- ブラウザ画面を情報量へ変換し、システムの状態を検出
  - 画面情報を参照画面情報との情報量の差へ変換（ KL 偽距離）
  - KL 偽距離の変化状態をバイナリコーディングで近似化
  - ベイジアン事後変化確率を計算
    - > 0.8 ならベイジアン事後確率変化点
  - ブラウザサイトの重要度のグルーピングにより情報を補強
    - G1 = Sign(w1 B1+w2 B2+ w3 B3 …)
    - 例えば、重要なウェブページのうち２つが変化していたら”重要な変化”

メタモニタリングシステム（ HappyFace, MadFace ）

- 開発やシステム管理等に労力のかかる部分（情報取得と状態識別）を完全自動化

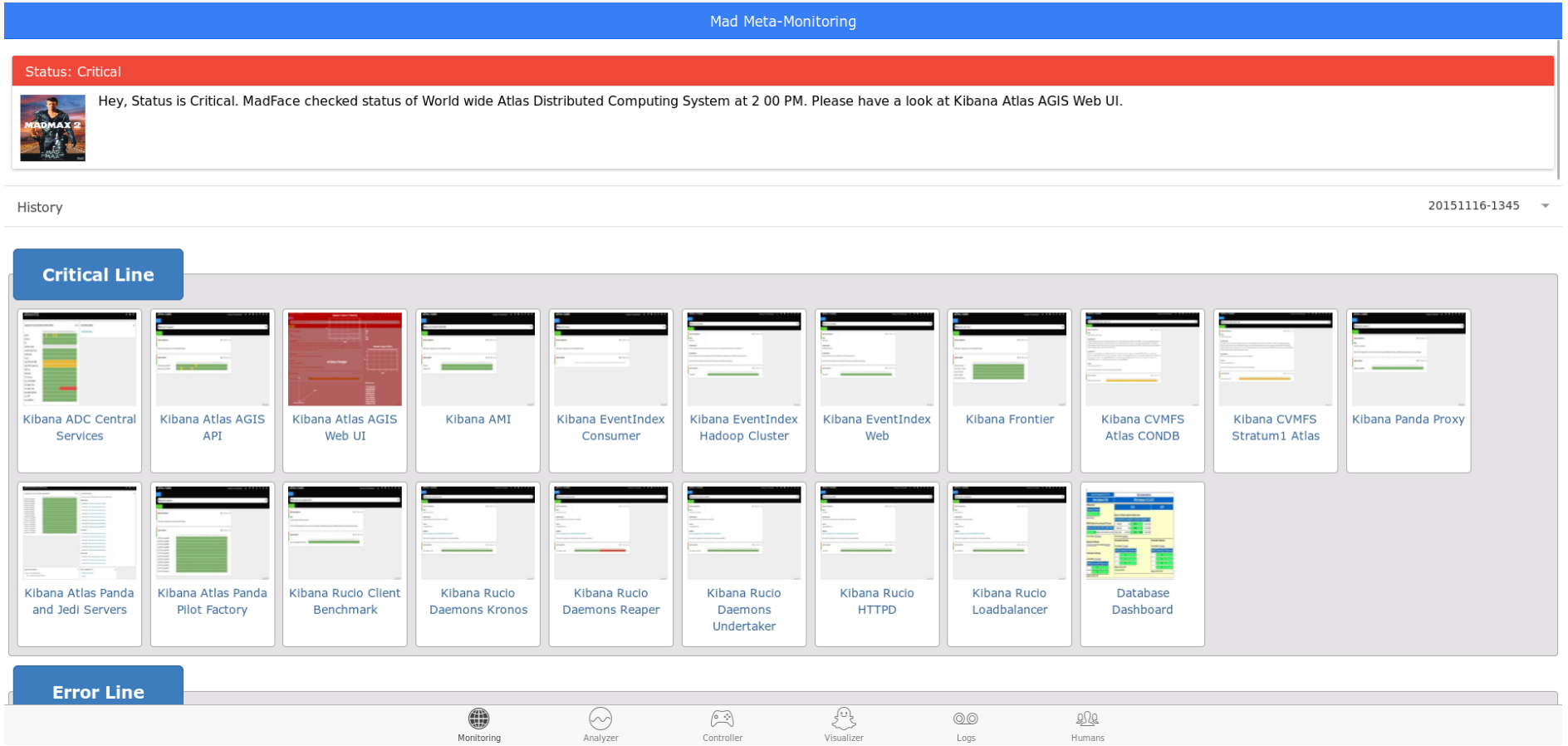
KL 偽距離

ベイズ事後確率



What does it look like now?

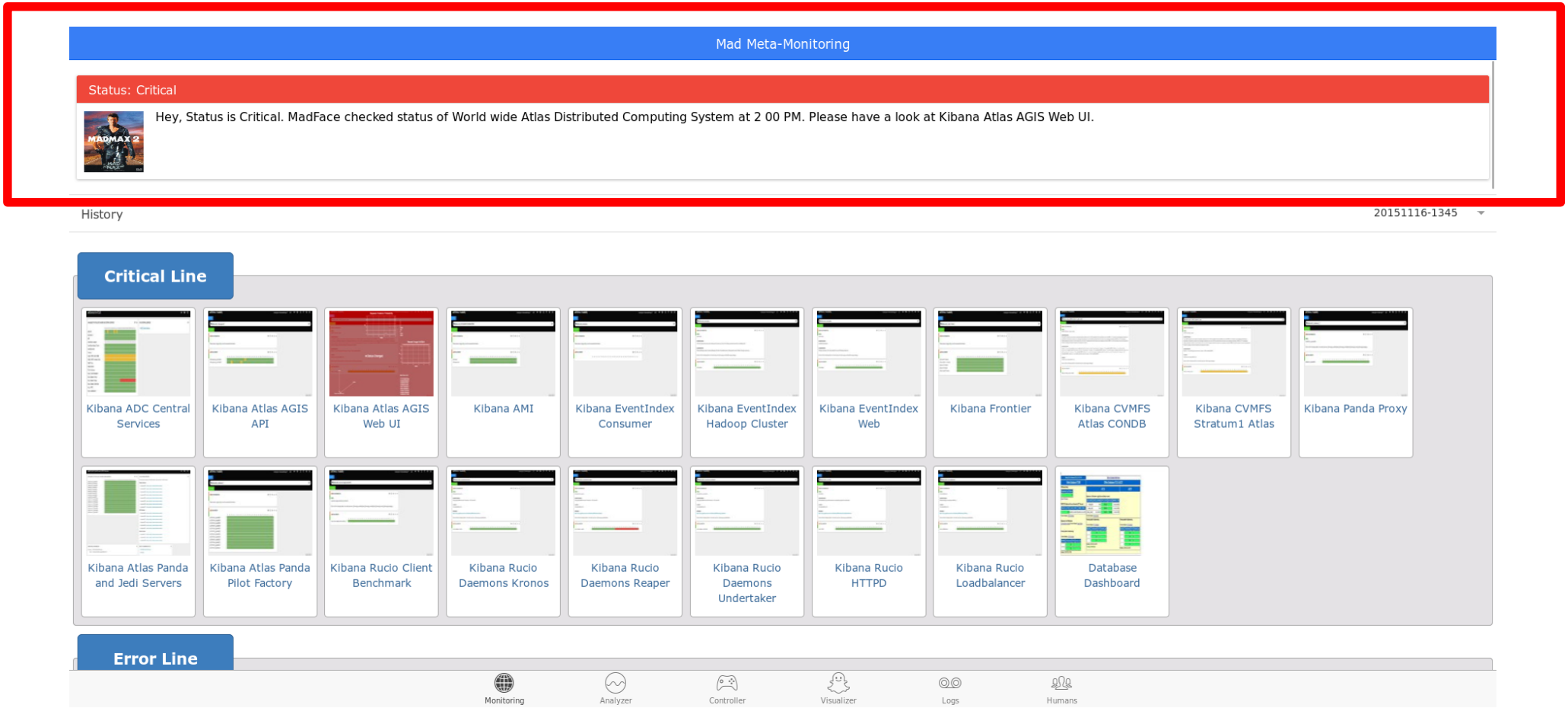
ATLAS Distributed Computing Operation Shifter Mode



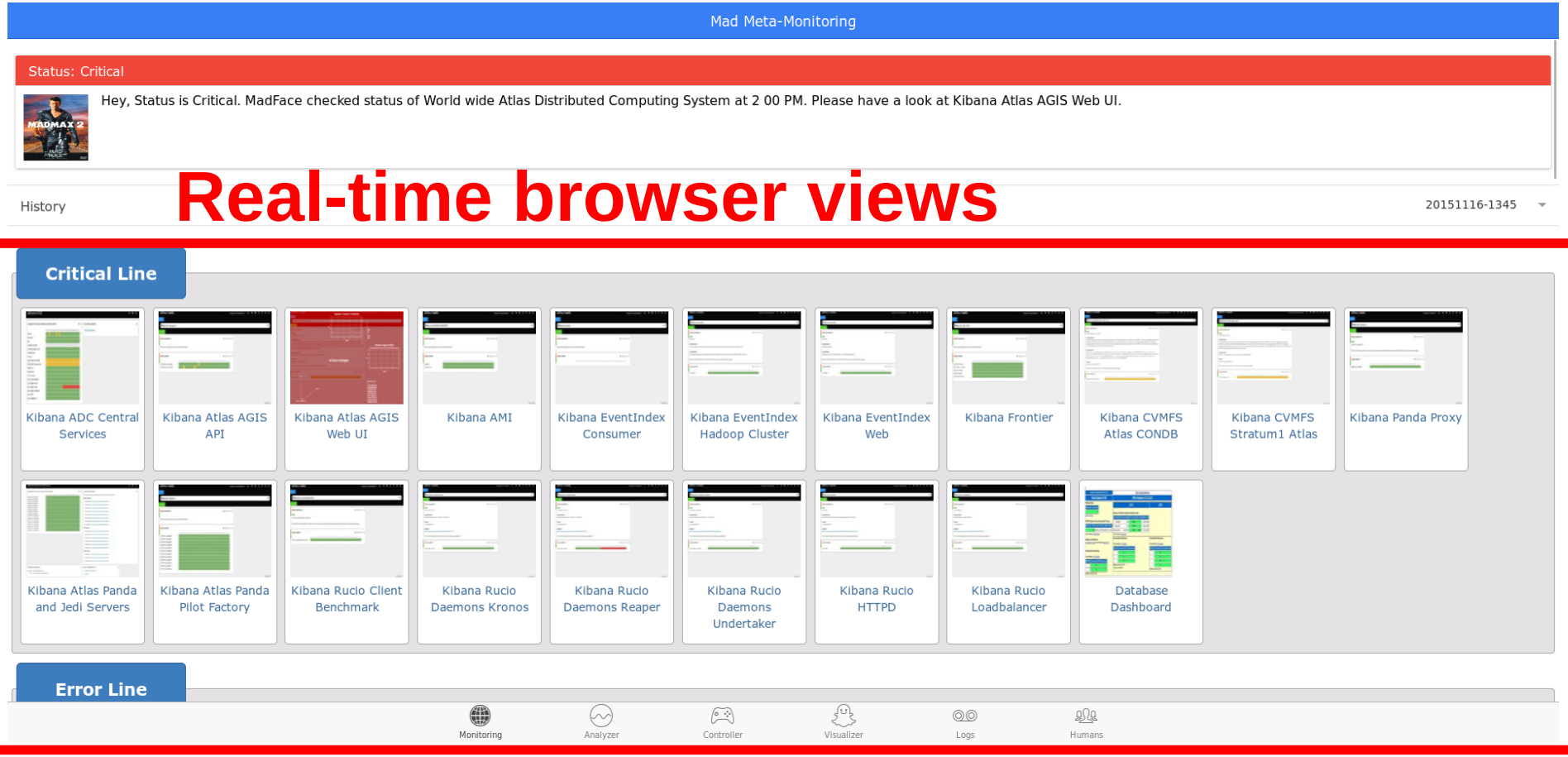
What does it look like now?

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

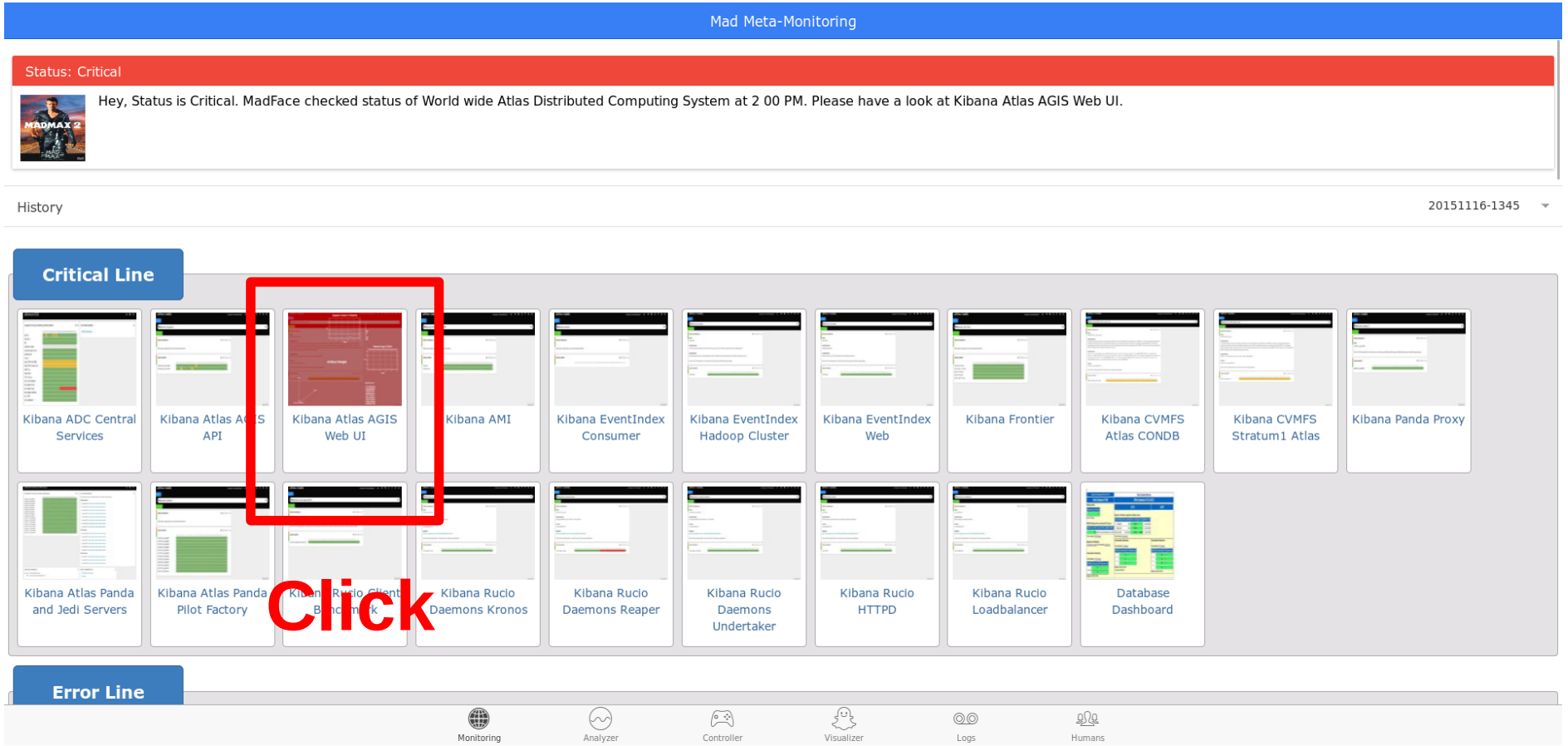
Human-readable status summary



What does it look like now?

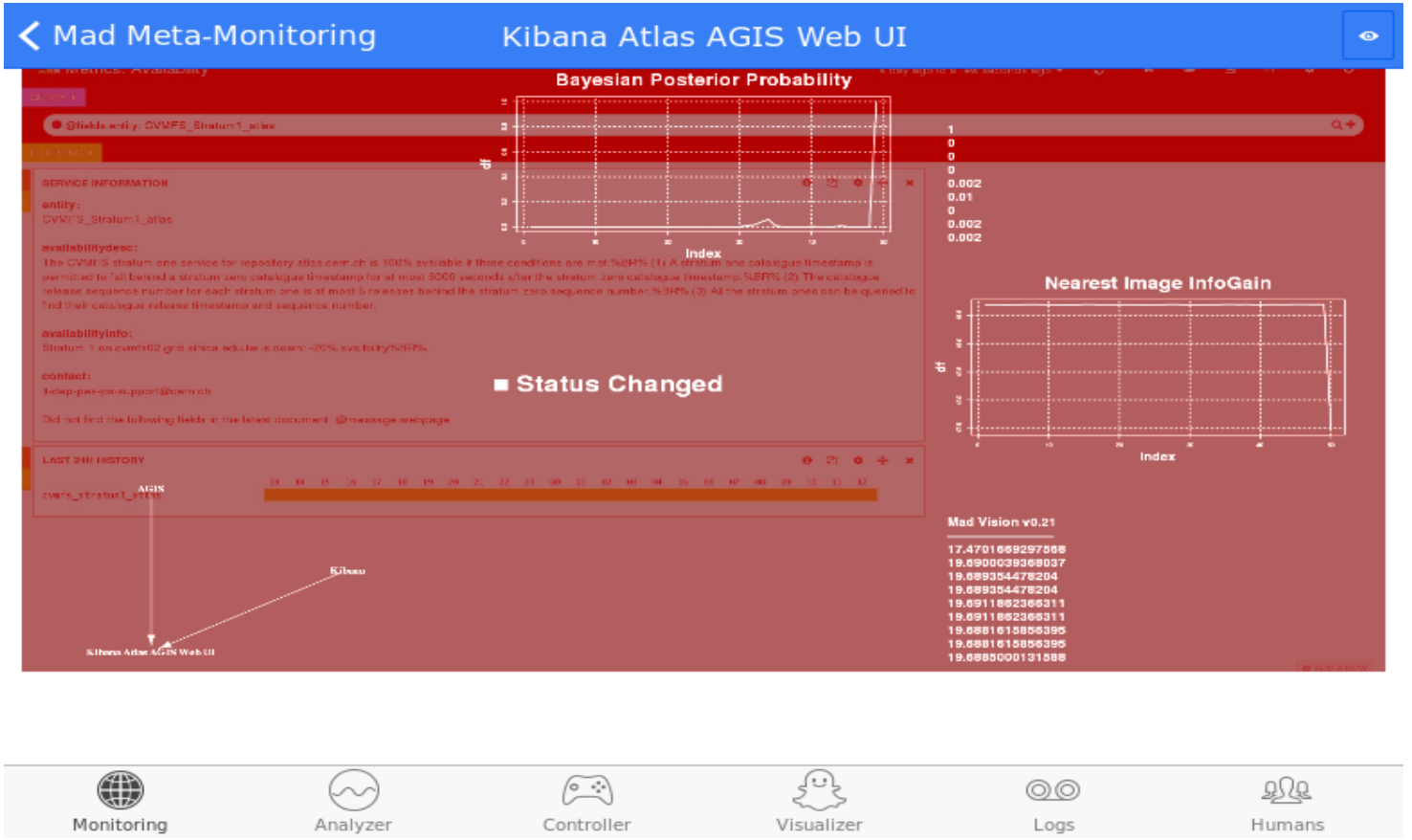


What does it look like now?



What does it look like now?

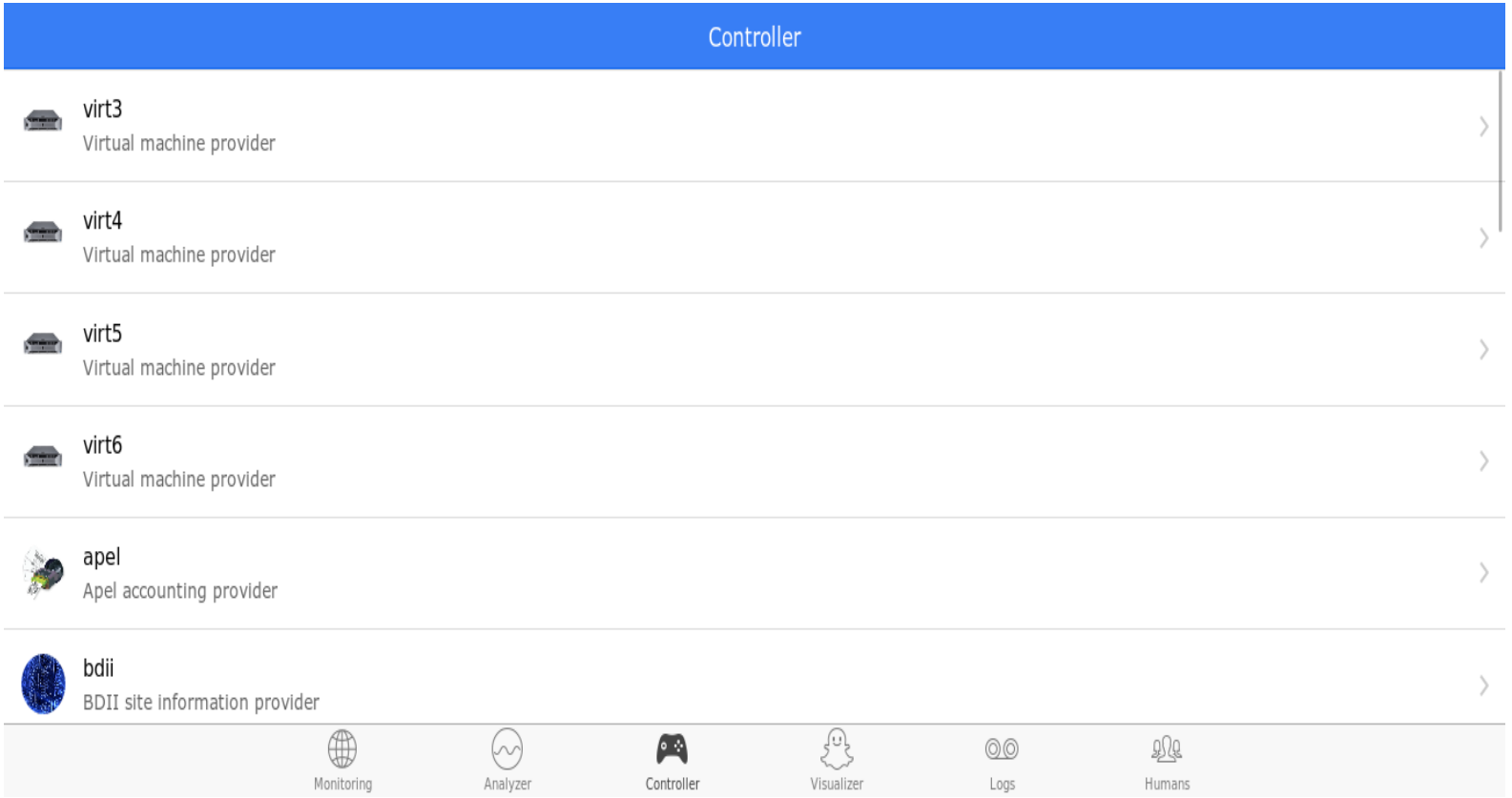
Mad Vision



What does it look like now?

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

1-click controller



What does it look like now?

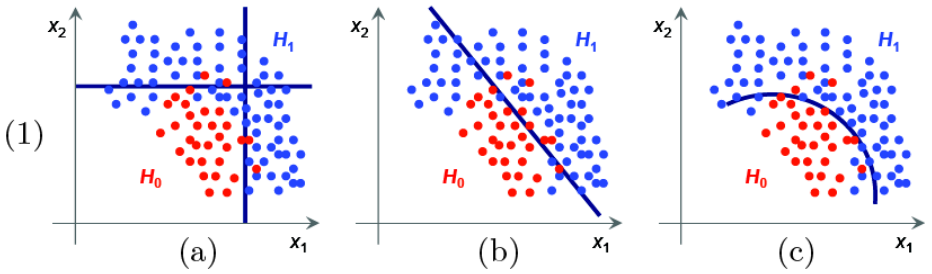
GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

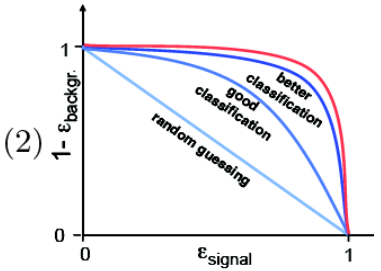
System analyzer

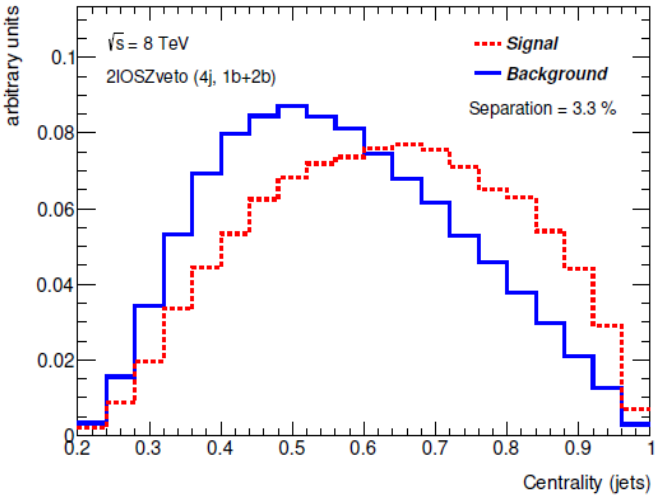
一般化されたブラウザ画面識別手法

- ブラウザ画面の特徴量を多次元化してクラスタ化
  - 手法： MeanShift クラスタリング
  - 最近傍参照画面との情報量距離を計算
    - 単純に閾値以下 (< h) であれば既知状態
      - Background → Normal (既知)
      - Signal → Error
  - 物理イベント S/B 識別とほぼ同様

Mr. Matyas Halasz







要約

- Göttingen コンピューティンググループはアクティブに活動中
  - ATLAS ソフトウェアは ARM アーキテクチャで駆動可能
    - ATLAS も推進中。＋ 1 マンパワー
  - クラウドは CPU のみかつコストベースだと Grid クラスタに比肩
    - ストレージ IO は当面課題
  - TensorFlow + DNN の性能は侮れない
    - ただし学習時に計算能力を馬鹿食いする
    - Grid クラスタでの実行はなお工夫を要する
  - モニタリングと管理の自動化は省力化のキーポイント
    - 一般化されたのですべてのサイト（大から小まで）で駆動できる
    - 実装と状態の自動検出は可能
    - 実際かなりの助けになる（実感では 0.5FTE 以上）

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


BERGISCHE
UNIVERSITÄT
WUPPERTAL


ビデオ・アニメーション・シリーズ
銀河英雄伝説
Vol.66
黄金獅子旗（ゴールデンルーウェ）の下に
原作・田中芳樹
総監督・石黒昇

Fragen?
質問？