ドイツ ATLAS Computing, GoeGrid, Göttingen グループ

ATLAS ソフトウェア講習会 2016
東京大学 2016 年 12 月 28 日

河村 元
II.Physikalisches Institut, Universität Göttingen

Overview

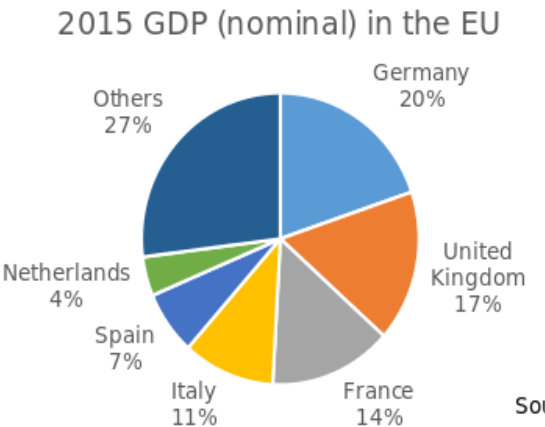
- ドイツ・コンピューティングの現状、戦略、将来
  - ドイツの Tier-1, Tier-2 センター概要
  - ドイツ・コンピューティング戦略
    - コンピューティング・モデル戦略 (Wuppertal 物理学計算機戦略会議より)
    - コンピューティング・リソース戦略 (Wuppertal 物理学計算機戦略会議より)
  - dCache の動向
- GoeGrid Tier-2 in Göttingen
  - Göttingen ってどこ？
  - 概要、大学 Tier-2 計算機センターの役割
  - 現状の計算資源と将来
- Göttingen ATLAS 物理計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

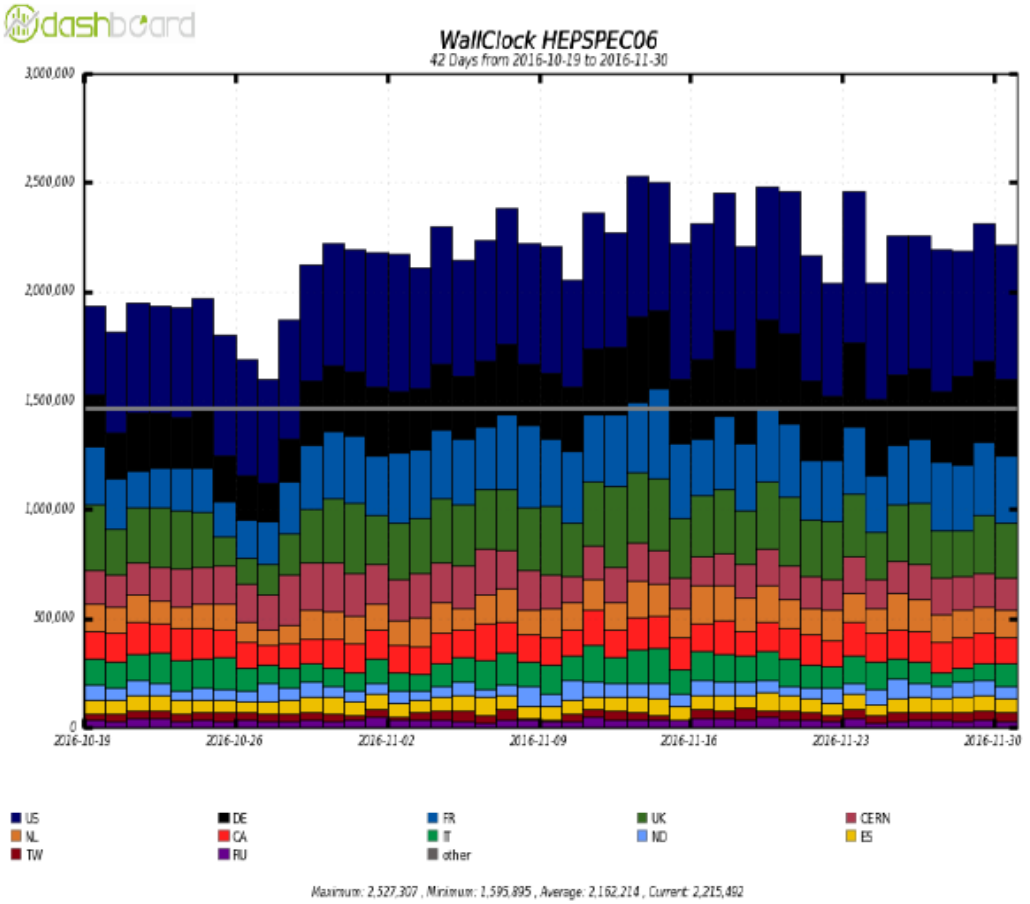
- 基礎データ
  - 欧州連合（ EU ）
    - 単一の国として見た場合、世界最大の経済大国
  - ドイツ連邦共和国
    - CERN の主メンバー国
    - 技術・科学を基盤とする欧州最大の経済大国
    - EU 内での GDP 比率 約 20%
      - 仮に日本が EU の場合、日本はドイツ＋スペイン程の規模
    - 連邦制共和国であり、地方分権が進んでいる

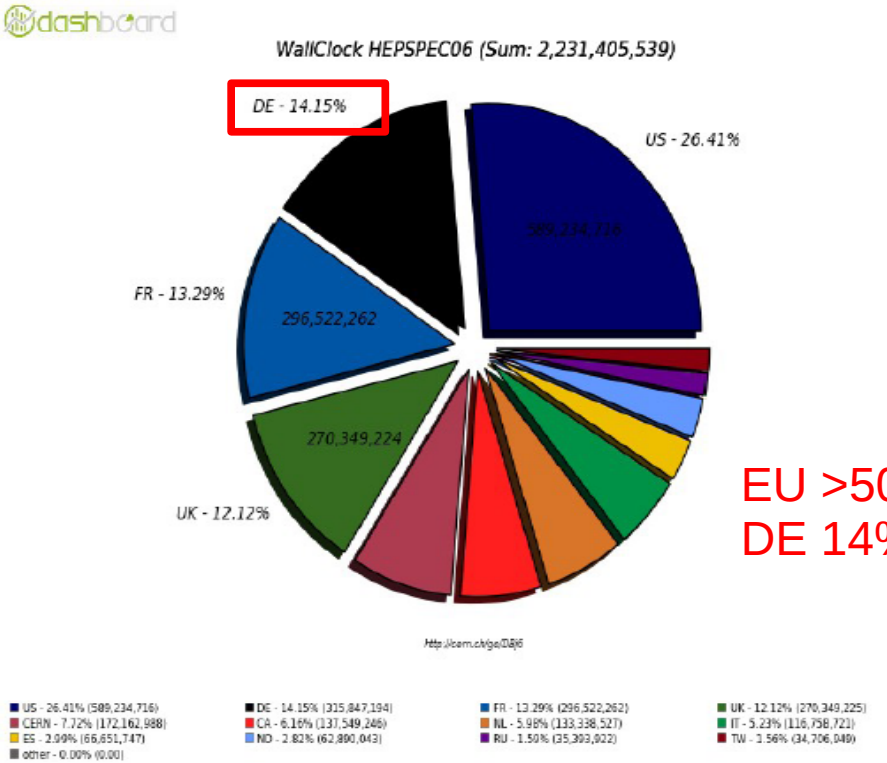
Bundesrepublik Deutschland



ドイツの Tier-1, Tier-2 センター概要

ATLAS Jobs Worldwide Oct 19 – Nov 30, 2016



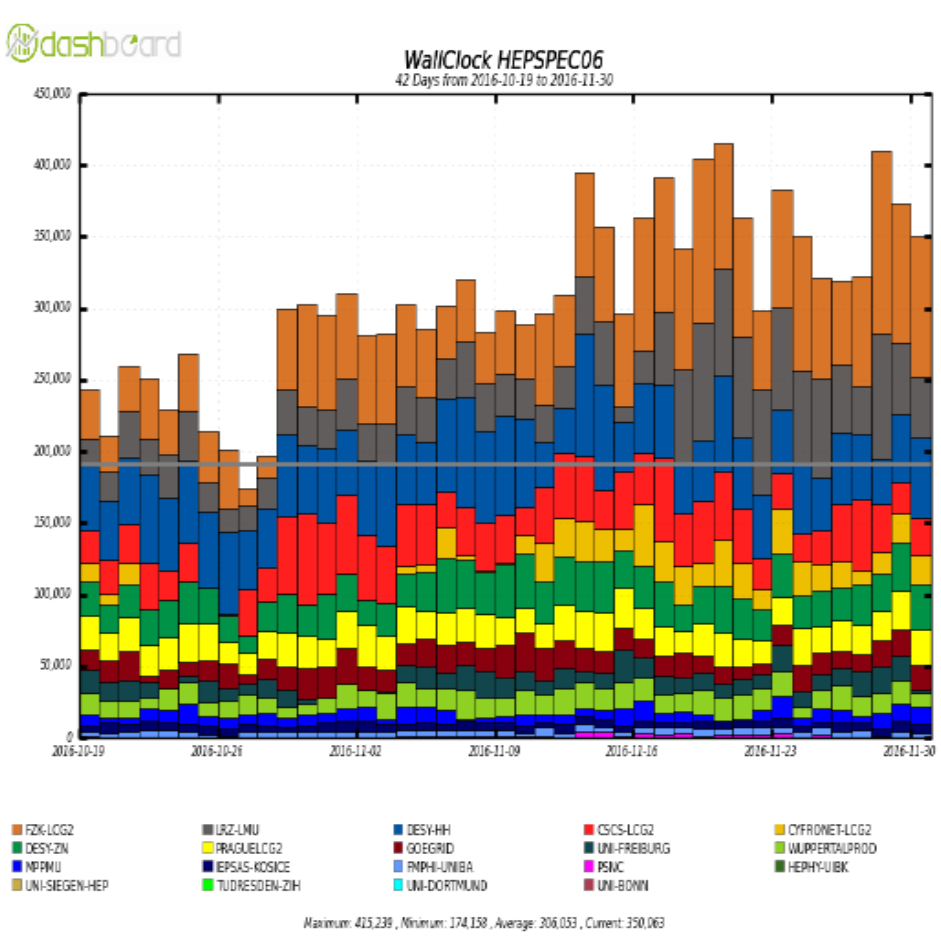


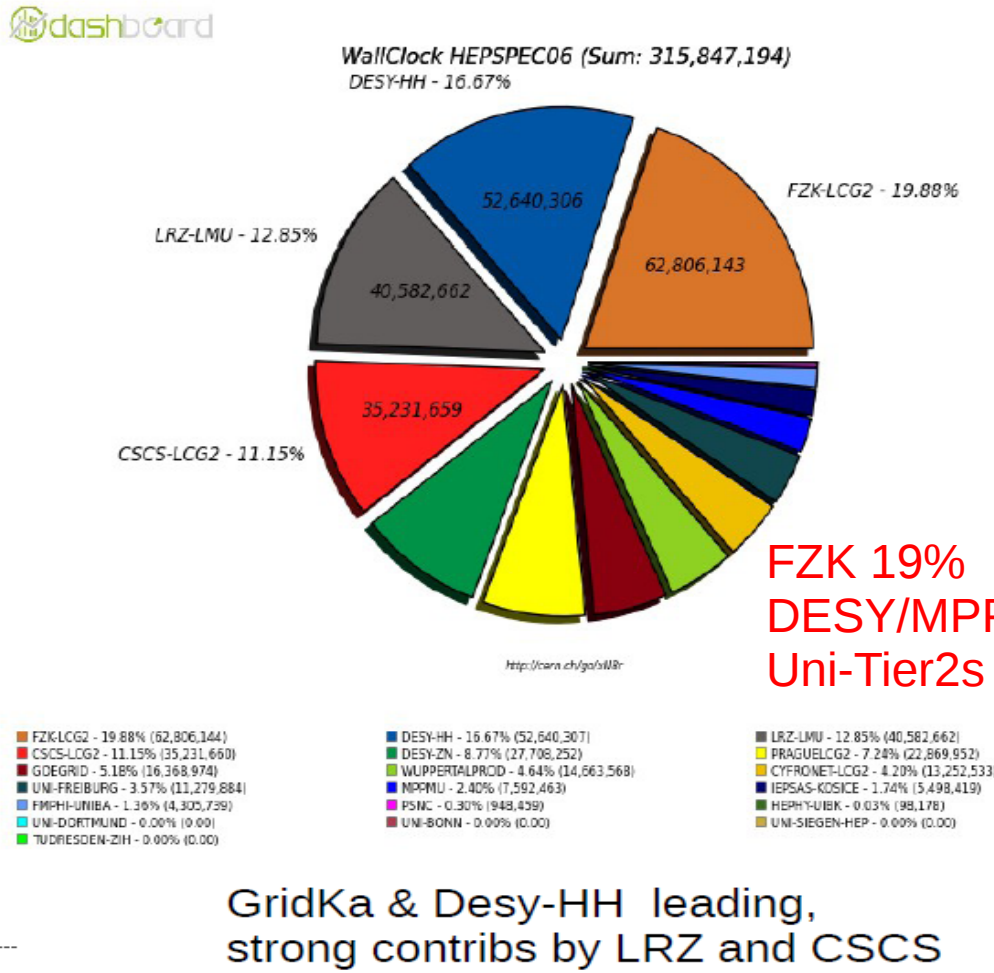
EU >50%
DE 14%

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



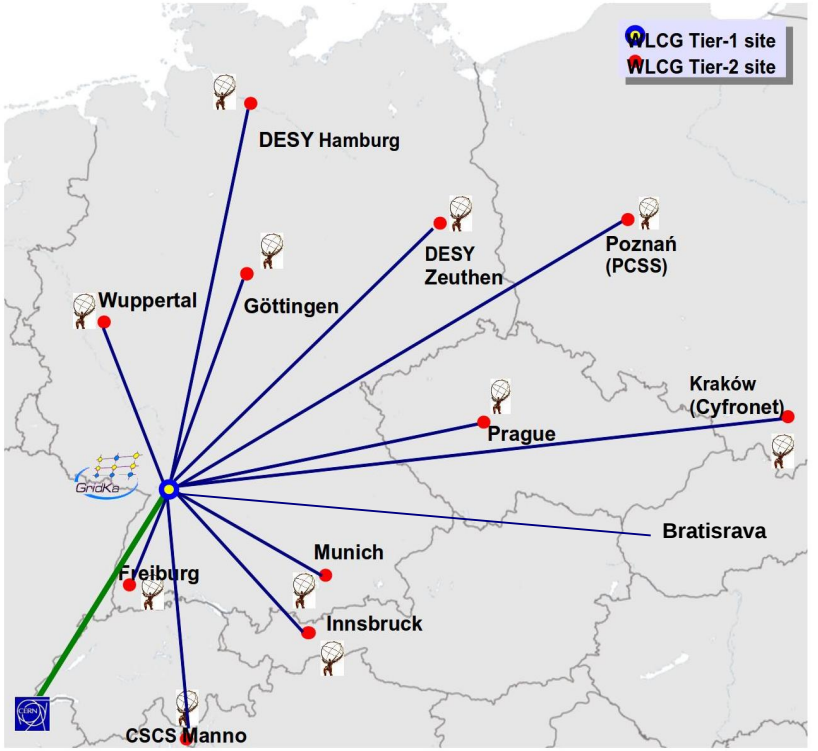


FZK 19%
DESY/MPP 28%
Uni-Tier2s 26%

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

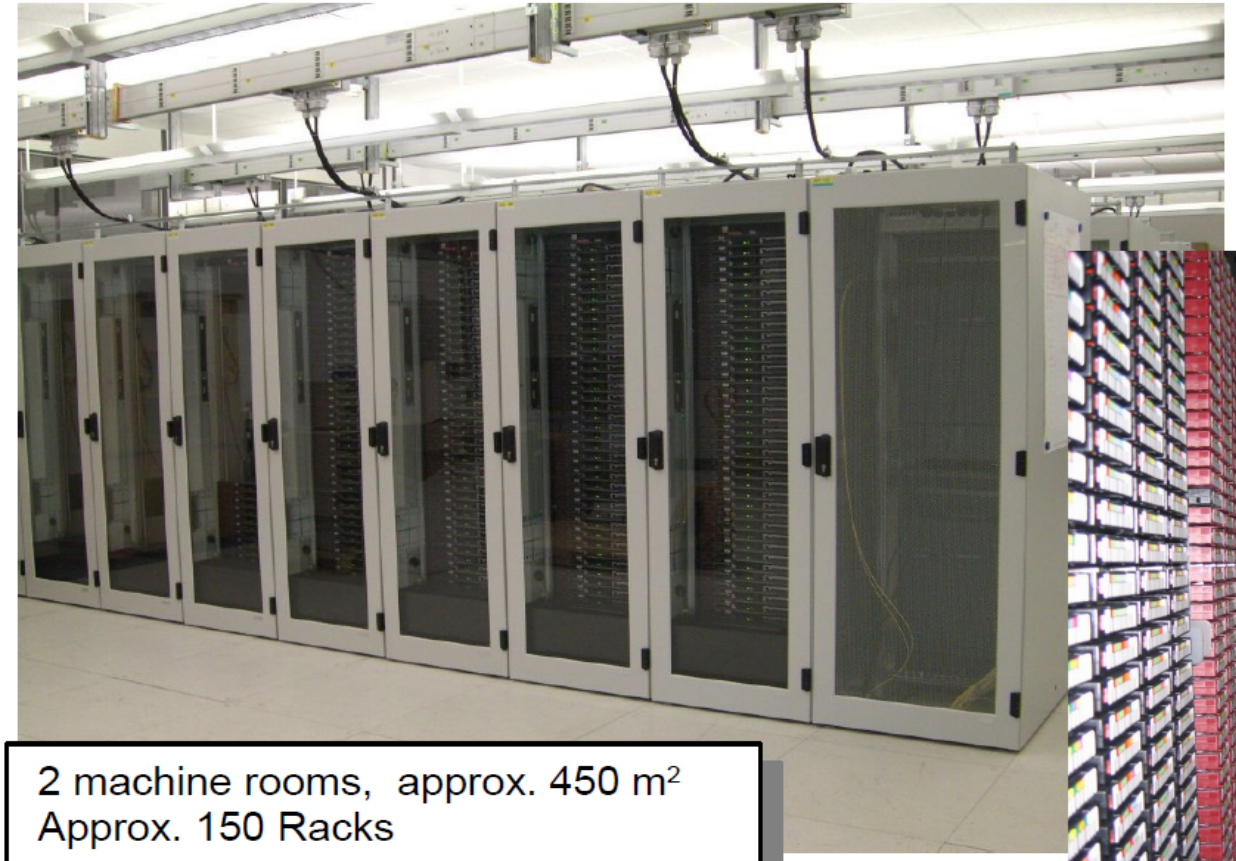
ドイツの Tier-1, Tier-2 センター概要

- FZK Tier-1センター
  - Regional Data and Computing Centre in Germany (RDCCG)
  - 2002年にスタート
  - Tier-1 サポート
    - 6.6 FTE (Full-Time Equivalent)
  - 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 Tier-2 センター
  - 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

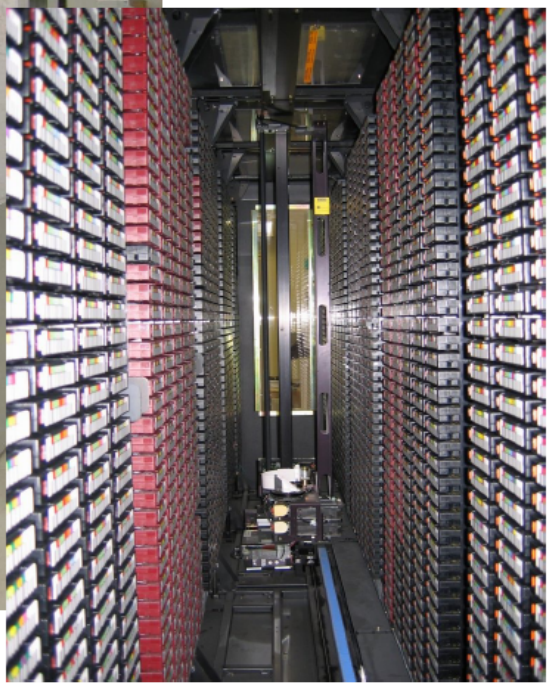


ドイツの Tier-1, Tier-2 センター概要

- FZK Tier-1 センター

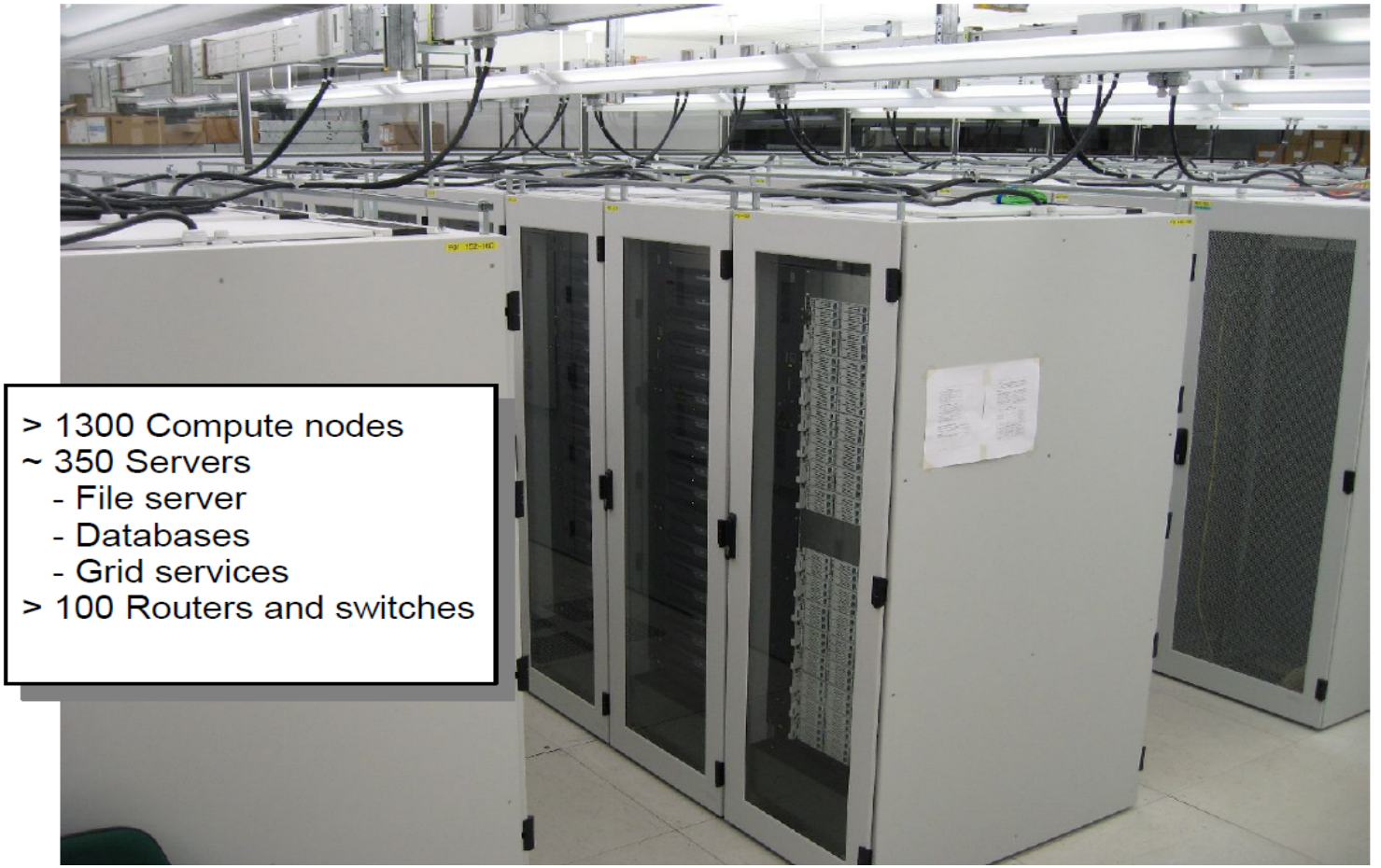


2 machine rooms, approx. 450 m²
Approx. 150 Racks



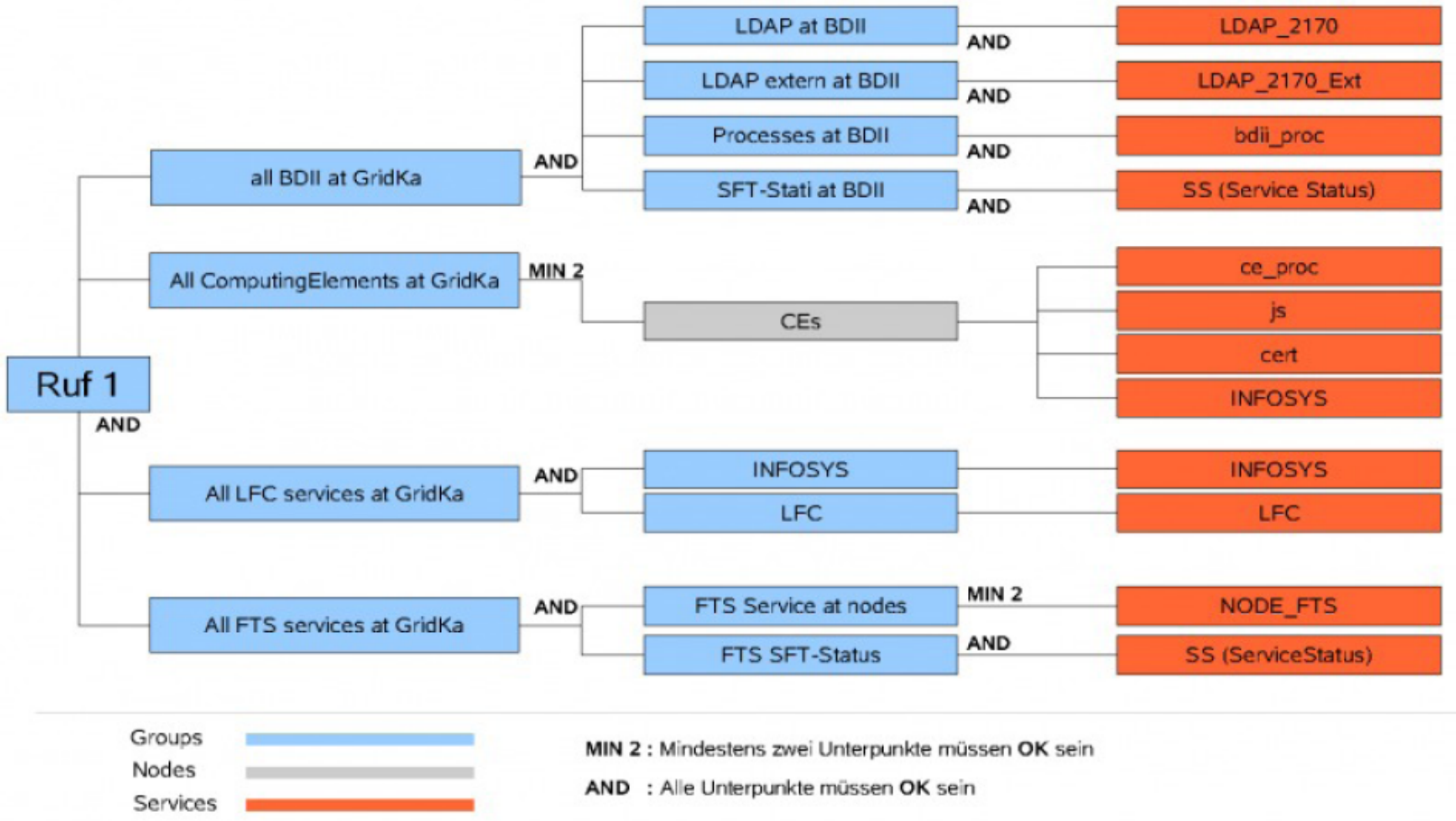
ドイツの Tier-1, Tier-2 センター概要

- 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
  - ドイツの各 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)

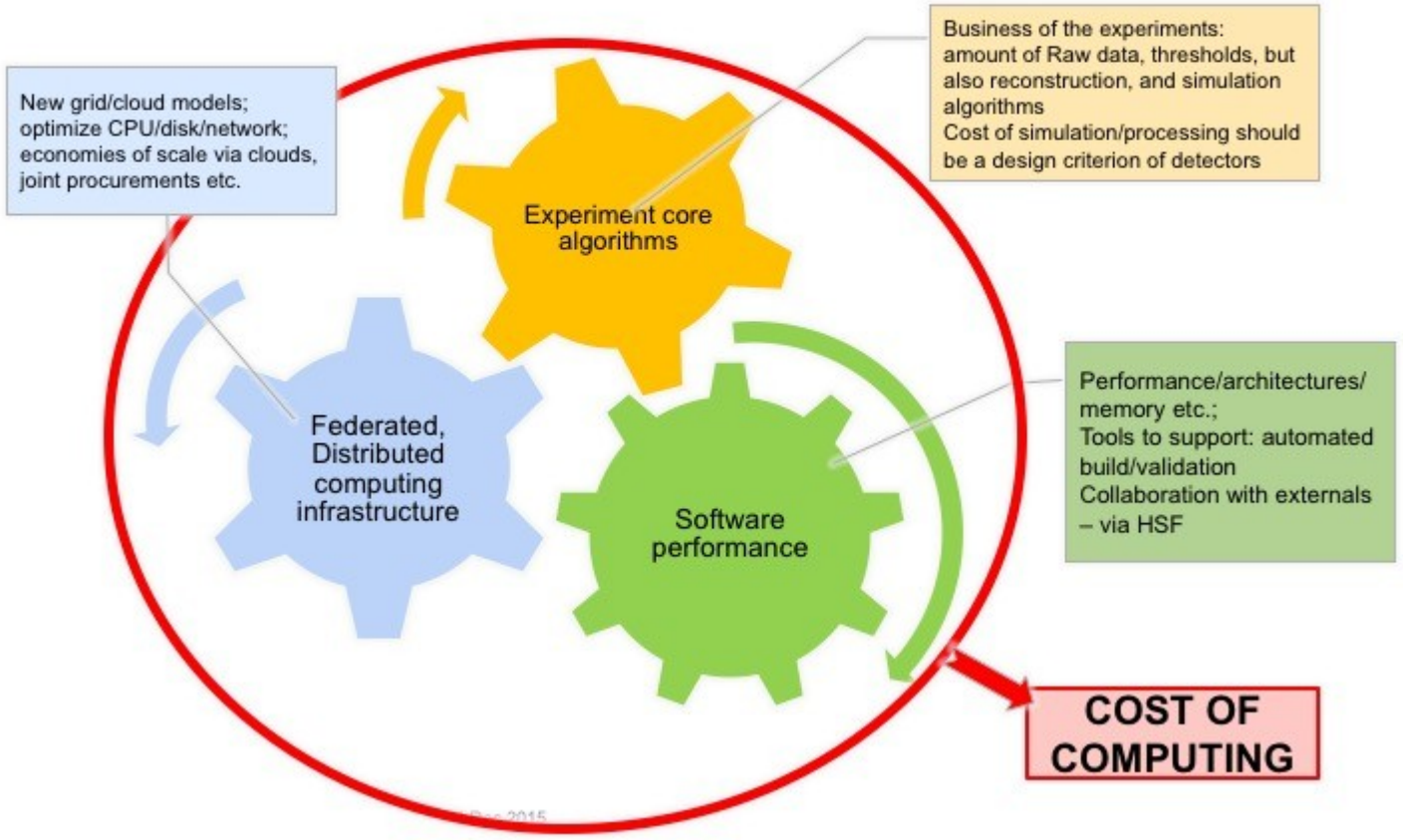
	2016			2017		
	CPU	Disk	Tape	CPU	Disk	Tape
Sum Uni T2s	37733	4800		75000	5533	
Desy/MPP T2s	34433	3600		56250	4150	
Sum DE T2s	72167	8400		131250	9683	
T2 ATLAS DE share	12.75%	11.67%		11.67%	11.67%	
GridKa T1	65000	5875	14500	97200	8500	22100
T1 ATLAS DE share	12.50%	12.50%	12.50%	10.55%	12.50%	11.80%

ドイツ・コンピューティング戦略

コンピューティング・モデル戦略（Wuppertal 物理学計算機戦略会議より）

- 実験とコミュニティのためのコンピューティング・モデル
  - 劇的なソフトウェアの改善が必要
    - パフォーマンス、スレッドの安全性、メモリ消費、 IT プラットフォームへの適応
    - パラレル・コンピューティングと GPU
  - ソフトウェア プロジェクトはドイツ HEP グループの透明性のために良い
- ヘルムホルツ協会・計算センターがリードしている利点
  - 例、 dCache
- ストレージへの努力量は明らかに減少させられるべき
  - 統合型ストレージ、キャッシュストレージ、リモートデータアクセスのみ

HL-LHC cost drivers



CERN WLCG

Observations

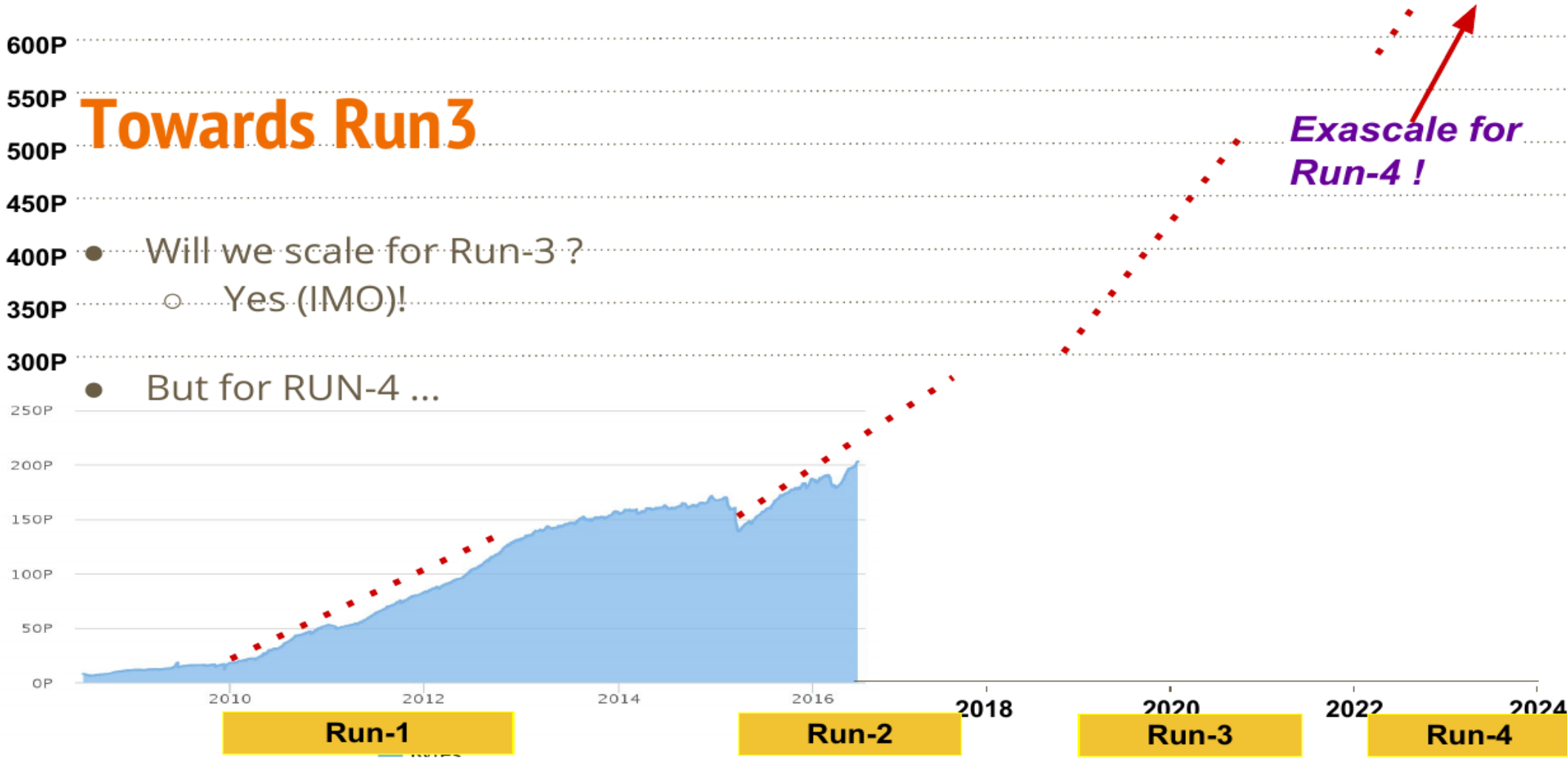
- Probably a lack of clarity over what the situation for Phase 2 upgrades will be:
  - In terms of requirements – what is the real scale of the problem – need better estimates
  - What we can really expect from technology
  - An understanding of the real limitations of the system we have today
- We should also bear in mind that while we potentially need to instigate *revolutionary* changes in computing models, nevertheless we will have to face an *evolutionary* deployment
- Concerns over software and efficiency (in all aspects) will be a significant area of work
- Commonalities may be possible in new tools/services or next generation of existing
- Propose a number of activities to address some of these aspects

コンピューティング・モデル戦略（Wuppertal 物理学計算機戦略会議より）

- LHC Run3/4 で劇的なモデルチェンジが必要か？
  - ALICE + LHCb: LS2 でメジャーアップグレード
    - コンピューティングはかなり大きく変わる
    - オンライン計算ファーム、オンライン再構築、各 Tier への新しい役割
  - ATLAS/CMS: LS3 で大きく変わる
    - ATLAS: Run-3 までは今の資源向上率で OK、 Run-4 は厳しい
    - 実験はアップグレード後なお複雑になる
      - ルミやデータレートが上昇。再構築がなお一層難しくなる
  - フレキシブル vs 特殊 Tier 計算センター
    - NAFs, Grid: dedicated analysis centres, regional data centres, ...

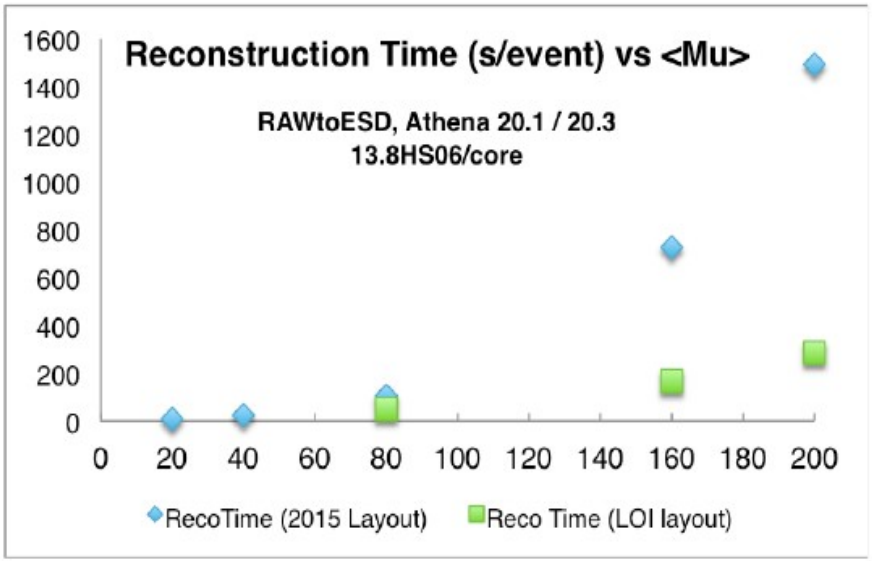
Projections for Run-3 & 4

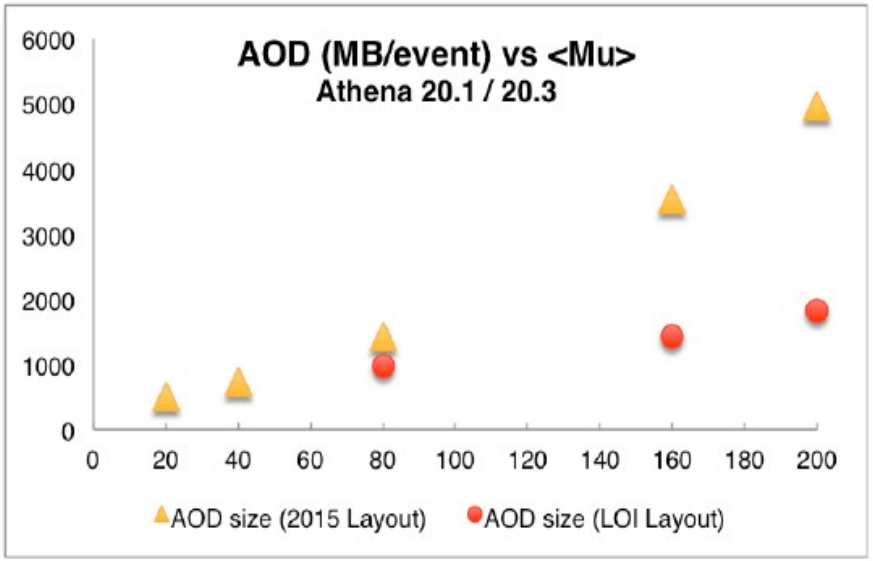
Slide V. Garonne

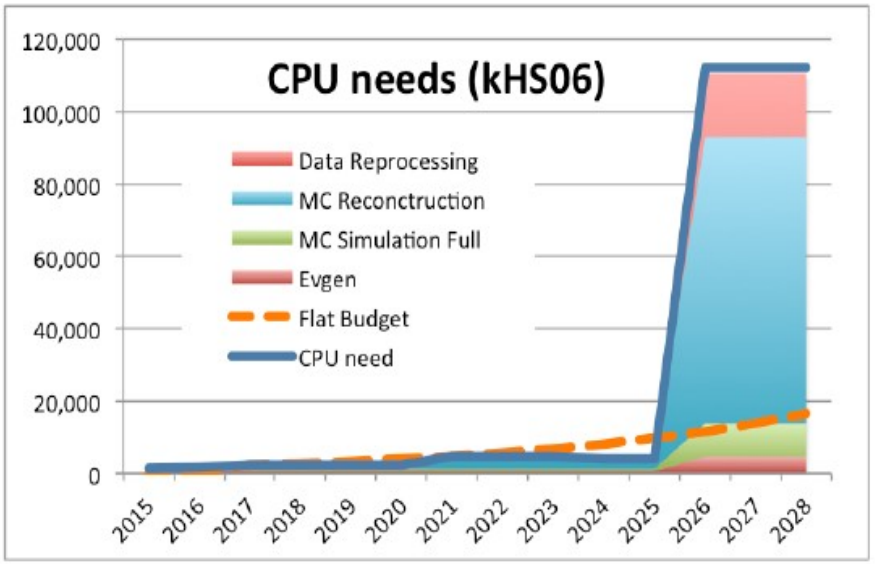


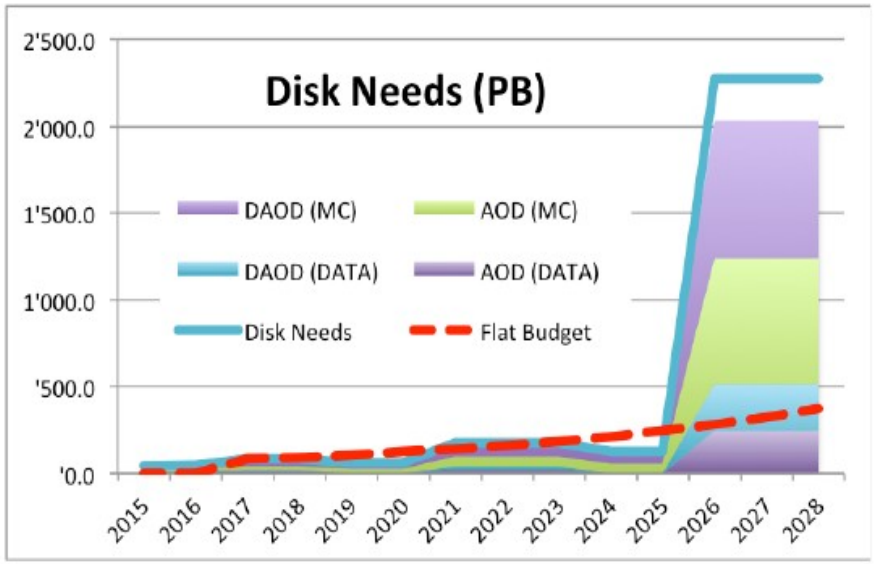
Slide
S. Campana

Initial studies on Computing for HL-LHC





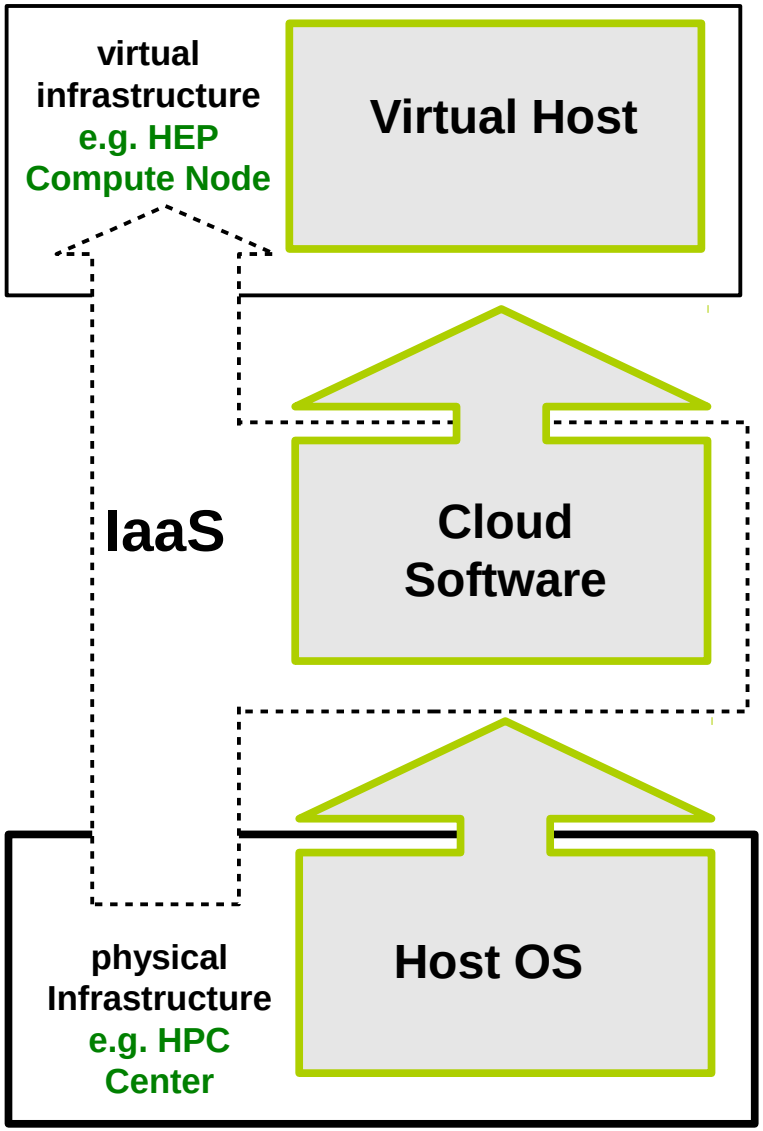




コンピューティング・モデル戦略（Wuppertal 物理学計算機戦略会議より）

- クラウドは誇大広告か、あるいは HEP は効果的に使用できるのか？
  - 市場でスタンダード
  - オープン・ソースも有（OpenStack 等）
  - たくさんの HEP ツールがクラウドに対応
    - CernVM, CVMFS, HTCondor, ...
  - HPC (High Performance Computing) センターはクラウド技術でユーザーへ一般的インターフェースを提供可能
  - Experiments はサイエンス・クラウドへの標準的インターフェースを獲得できる
  - ストレージ資源は今の枠組みで良い
    - 商用クラウドリソース - データ依存ジョブはまだ困難

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

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

Private (Research) Clouds

OpenStack
Cloud Operating System

- Complete open source **IaaS framework**
- Backed by lots of big companies (IBM, Red Hat, Rackspace, HP, ...)
- **Standardized API** (Amazon AWS, fits with HEP workflow tools)
- Acceptable to HPC Centers (already in use)
- Active Involvement of the HEP Community (CERN Personnel in OpenStack Foundation Board)

proven software

long term support can be expected

コンピューティング・リソース戦略（Wuppertal 物理学計算機戦略会議より）

- HPC (High Performance Computing) リソースを WLCG へ統合するのにいくら必要か？
  - 巨大な資源がドイツ HPC に存在
  - LMU の経験では申請書だけで”無料資源”
  - クラウド型か Grid 型のリソースか？
- HPC での GPU の利用
- Reconstruction at ALICE-HLT, full workflow by ICECube, Geant V, ALICE Kalman filter, ...

EU & China HPC

- LRZ SuperMUC
  - Phase 1: 150k cores, Sandybridge
  - Phase 2: 86k cores, Haswell
  - 19Mcore hours used from 20M allocation
    - effectively open-ended allocation if preempt-only
- Max Planck Institute computer centre: Hydra
  - 83k core Sandybridge + 28k core extension
- UK Supercomputer - Edinburgh
- China – various
  - start small and move up

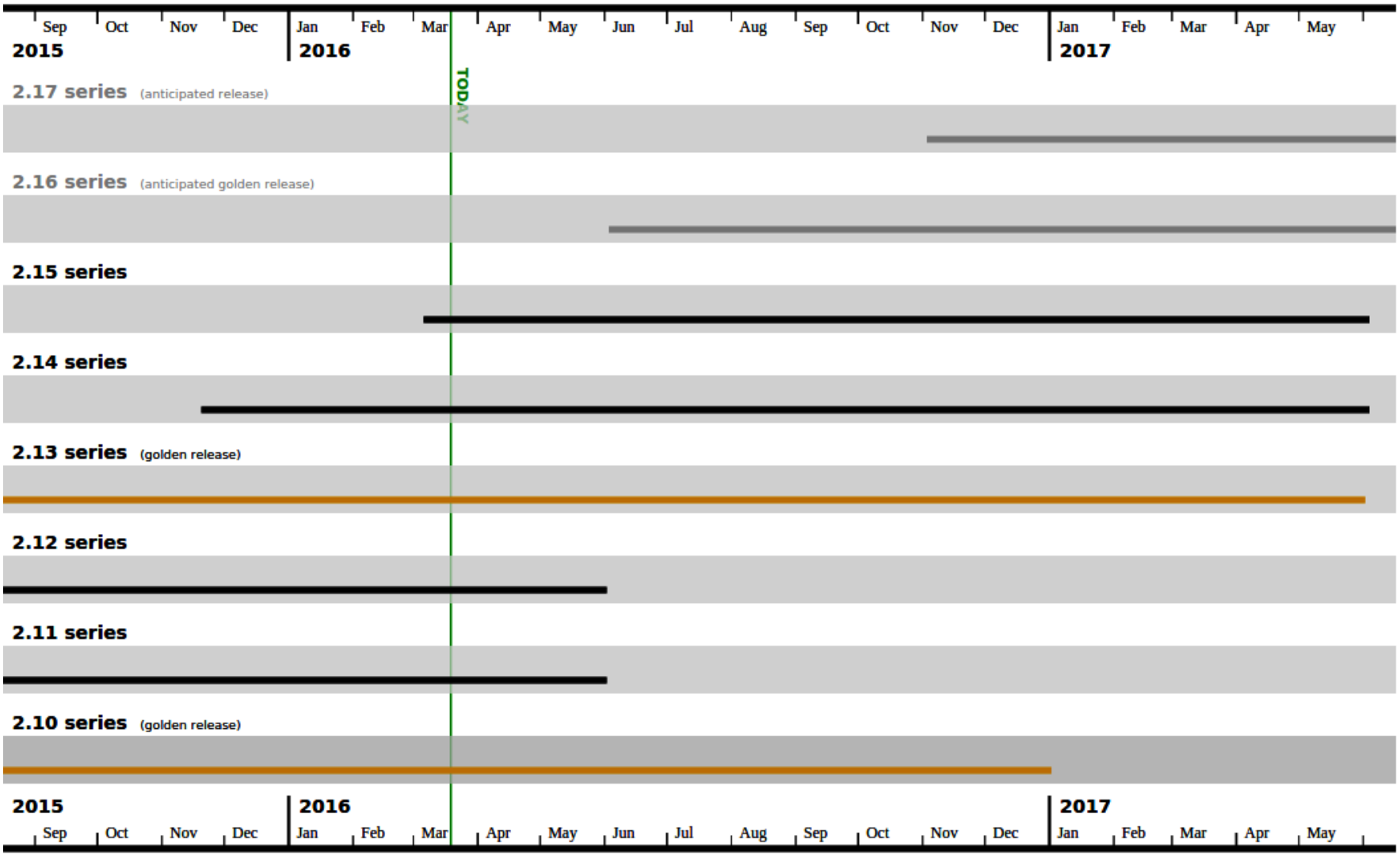
dCache の動向

dCache = コンピューティングクラスタ用の PB クラス・ストレージの管理用ミドルウェア。ディスクサーバーを結合し巨大で一元化されたストレージ資源を提供可能。WLCG Tier-1 ではほぼ標準 Grid ストレージ用ミドルウェア。多彩なプロトコルに対応。 Grid 対応。クラウドストレージプロトコル等にも対応中。

開発元：ドイツ DESY
メイン開発言語： Java
DB バックエンド： PostgreSQL
質問等は support@dcache.org

dCache server releases

... along with the series support durations.



dCache 運用コストの削減

- High Availability (HA) のサポート
  - いかなるサブシステムの停止でも人間を介在させずに復帰させる
    - 運用に必要なマンパワーを縮小できる
  - HA 対応は dCache v3.1 以降
  - Light-Out Management
    - REST ウェブサービスへシステムマネージメントとデータ管理のインターフェースを可能に
    - 既存の管理システムやモニタリングシステムと容易に統合可能に

INDIGO - DataCloud

RIA-653549

The First INDIGO-DataCloud Software Release

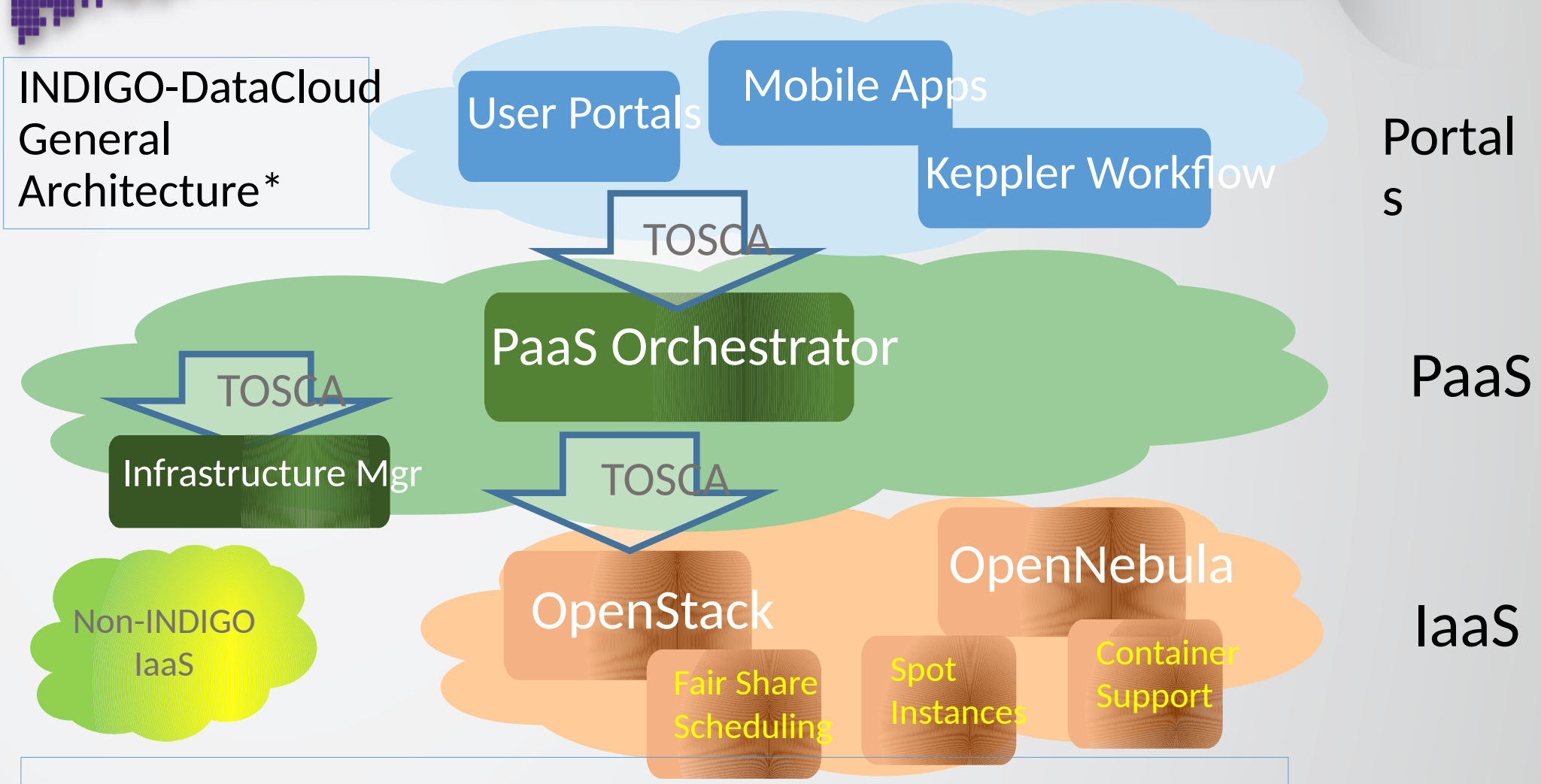
Presenter : Paul Millar for
Patrick Fuhrmann for
Davide Salomoni
INDIGO-DataCloud Project Coordinator
HGF “Physics at the Terascale”, November 2016, DESY
davide.salomoni@cnaf.infn.it

INDIGO-DataCloud is co-founded by the
Horizon 2020Framework Programme

The long road to the release, from the architecture…

INDIGO - DataCloud

INDIGO-DataCloud General Architecture*



*: see details in http://arxiv.org/abs/1603.09536 or in https://www.indigo-datacloud.eu/documents-deliverables

What this means for dCache

- INDIGO-DataCloud providing some €0.5M of project money to dCache team:

  Funding 3 FTEs working on dCache.

- Work follows the DoW:

  HEP (through WLCG) has supplied use-cases

- Improvements within dCache include:
  - Adding OpenID Connect support to dCache
  - Improving Quality of Service management options,

    Media, number of copies, …
  - Improving Quality of Service management,

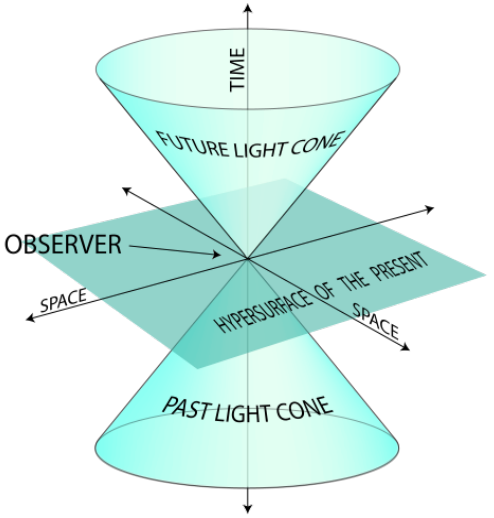
    Focus is on QoS management via CDMI.

要約

- ドイツ・コンピューティングはWLCGへ多大な貢献をしている
  - ドイツTier-1はFZK、 Tier-2最大サイトはDESY-HH
  - Run-2 計算資源アップグレード進行中
- 将来的（HL-LHC）に劇的な何かが必要
  - ATLASは特にRun-4
  - ソフトウェア？？
- ATLAS資源へのHPC統合は（ドイツでは）お得
- dCacheはさらに開発中
  - メンテナンスフリー
  - クラウド対応
  - DESY INDIGO プロジェクト始動

GoeGrid Tier-2 in Göttingen

Deutsches Elektronen SYnchrotron



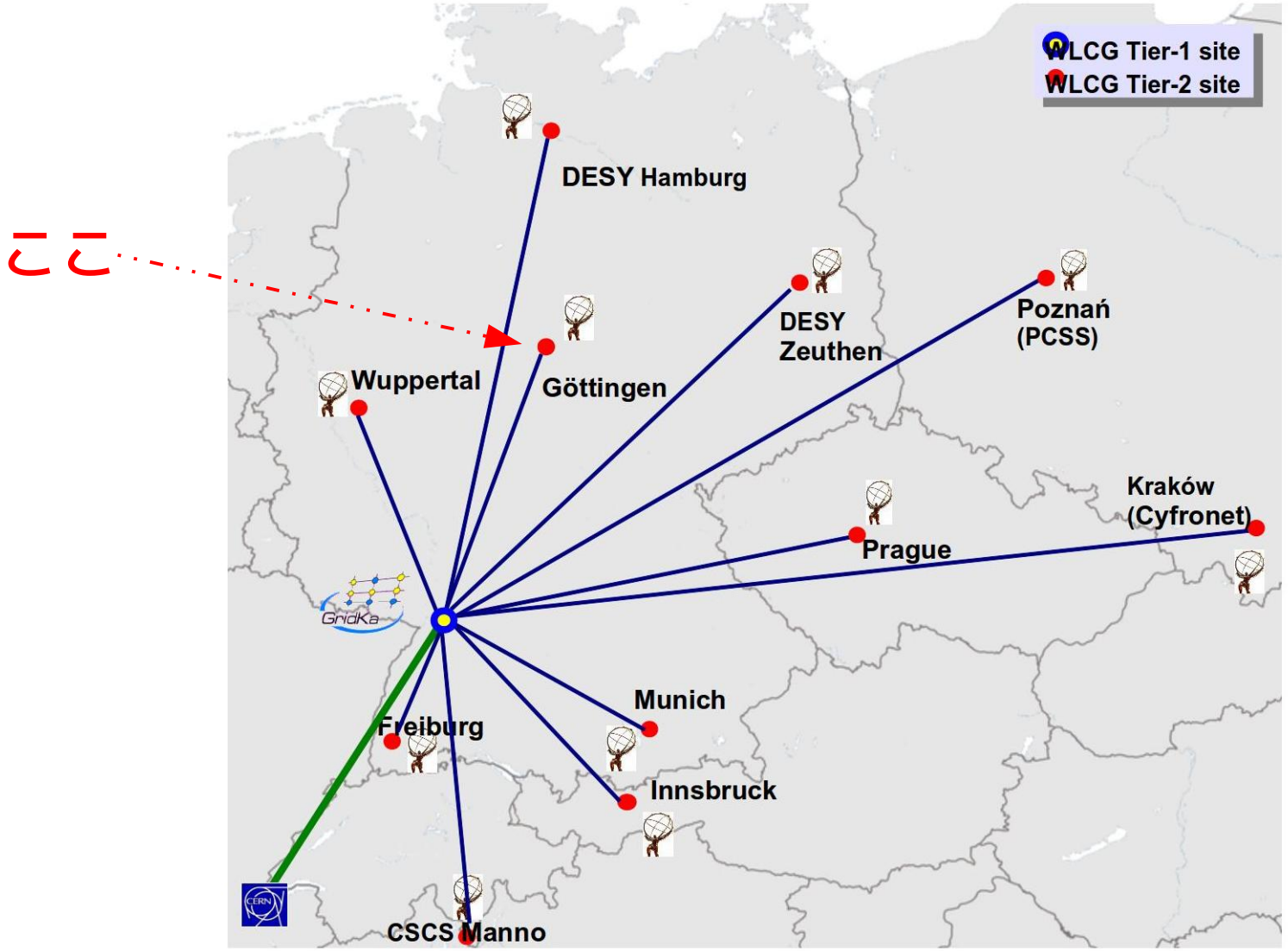
Minkovsky space-time

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10万人の大学都市
  - 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、ハノーファー
  - 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
  - ドイツエリート大学の一つ
  - 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、ドイツ緑の党の中心
  - 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（日本の現代物理学の父 → 仁科記念賞）
  - ケンブリッジ大学との紳士協定により、ドイツの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・原子核物理学の世界的中心地 → ナチスにより地位を失墜
  - 大戦中の世界の３つの核爆弾プロジェクト（牽引者は全員ゲッティンゲンにいた）
    - マンハッタン計画（連合国：ロバート・オッペンハイマー）
    - 二号計画（日：仁科芳雄）
    - ウラン・クラブ （独：ヴェルナー・ハイゼンベルク）

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

- LHC ATLAS Tier-2 サイトの役割
  - MC プロダクション
  - Tape ドライブなし
  - 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 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

GoeGrid Tier-2

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ナンス
GWDG: Tim Ehlers

0.3FTE

Batch System, Provisioning
Teory group: Dr. Juergen Holm

0.6FTE

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

現状の計算資源と将来

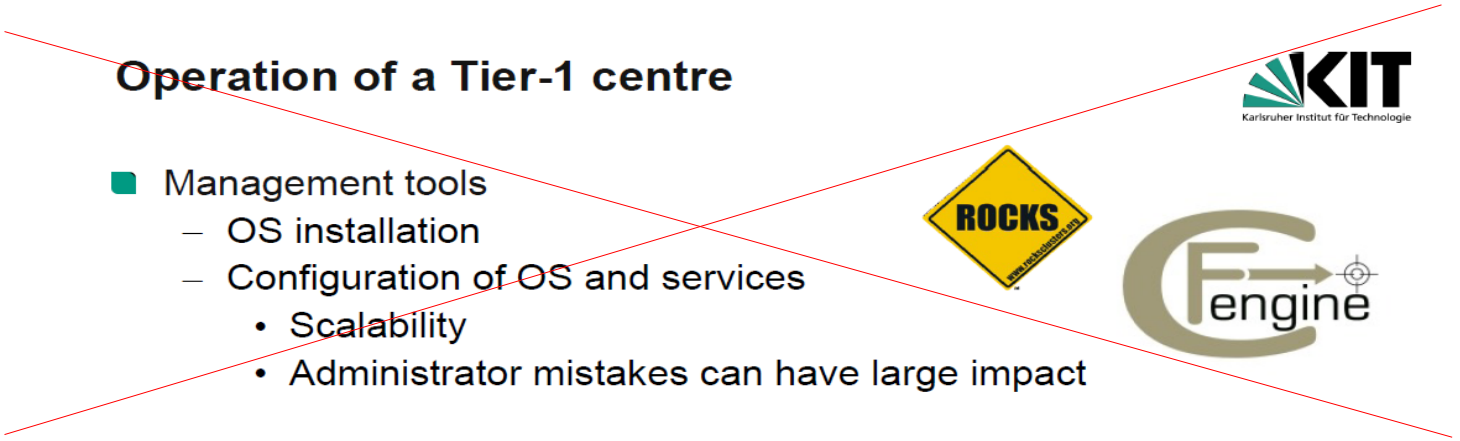
- 2008 年に WLCG の ATLAS 専用 Tier-2 として開始
  - 当初は天文物理、バイオ医療科学、グリッド開発、理論物理等も参加
- 現在
  - 約 20% の計算資源はゲッティンゲン理論グループ、 ATLAS は約 80%
  - ストレージ資源は 100%ATLAS
  - ドイツ中堅レベルの WLCG ATLAS 大学計算サイト
- ハードウェア構成
  - ラック x 13
  - サービスホストサーバー x 24
  - ストレージサーバー x 15
    - 1.2 PB ディスク
  - 3024 Logical CPUs

現状の計算資源と将来

- ミドルウェア・管理システム
  - dCache
  - CREAMCE x 2 + PBS (メイン)
  - CREAMCE x 1 + HTCondor (テストシステム)
  - Scheduler は Göttingen Tuned (by Gen Kawamura)
    - PBS のスケーラビリティ確保のため。3000 コアで標準 PBS は無理
  - ROCKS プロビジョニング・管理エンジン
  - APEL, BDII, etc...
  - SAN ストレージ 10TB
  - VM ホストサーバー x 4
    - dCache 以外の Grid サービスは VM
  - Frontier server x 2
  - モニタリング： Nagios, Ganglia、 HappyFace、 MadFace

GoeGrid 1.0 → GoeGrid 2.0

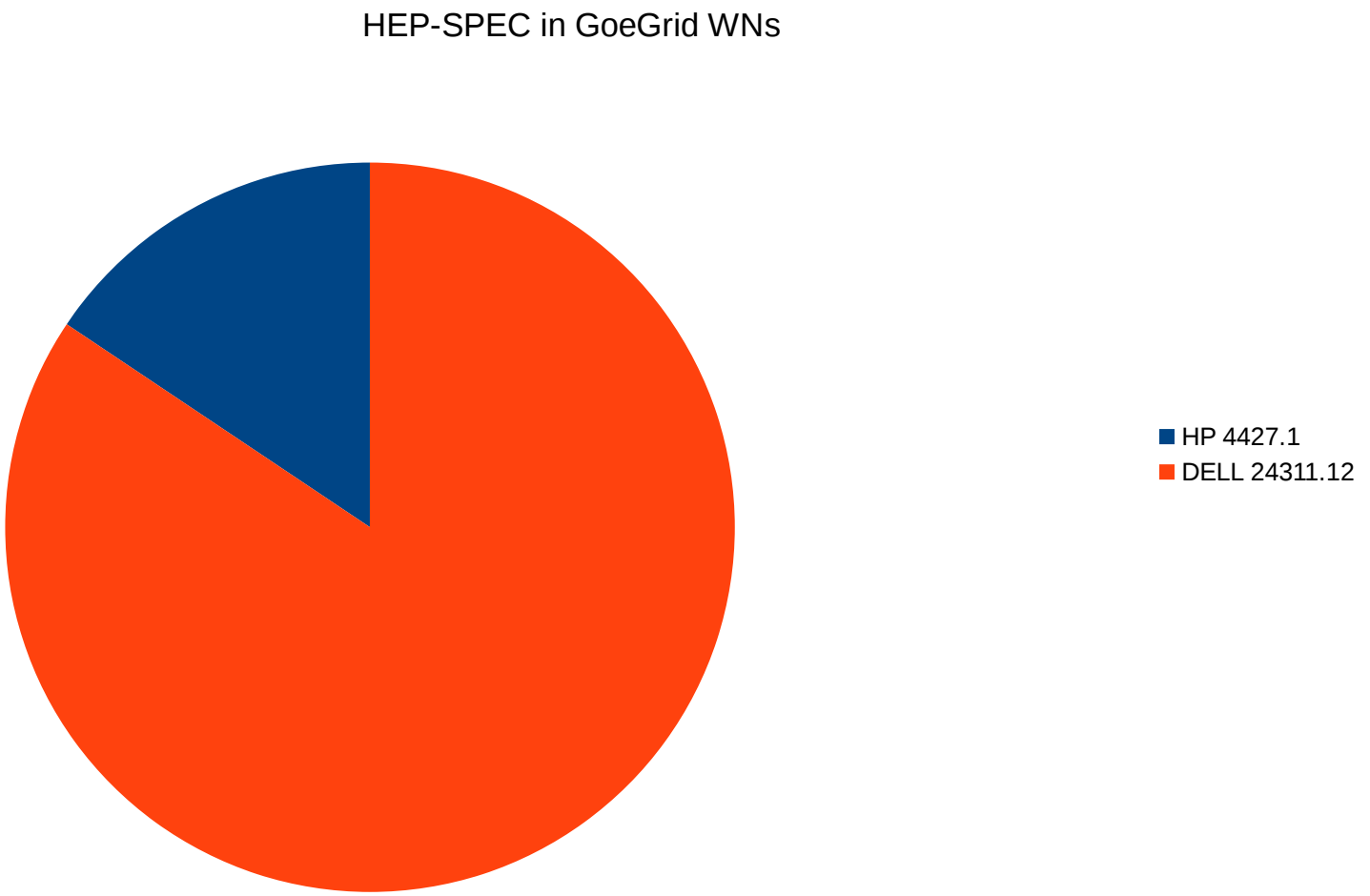
- 1.0: 過去の構成 (Cfengine + ROCKS)
  - FZK, DESY-HH に倣った構成
  - PBS + CREAMCE + dCache



- 2.0: Foreman + Puppet
  - HTCondor + ARC-CE + CREAMCE + dCache

～ 2018 ハードウェア予定

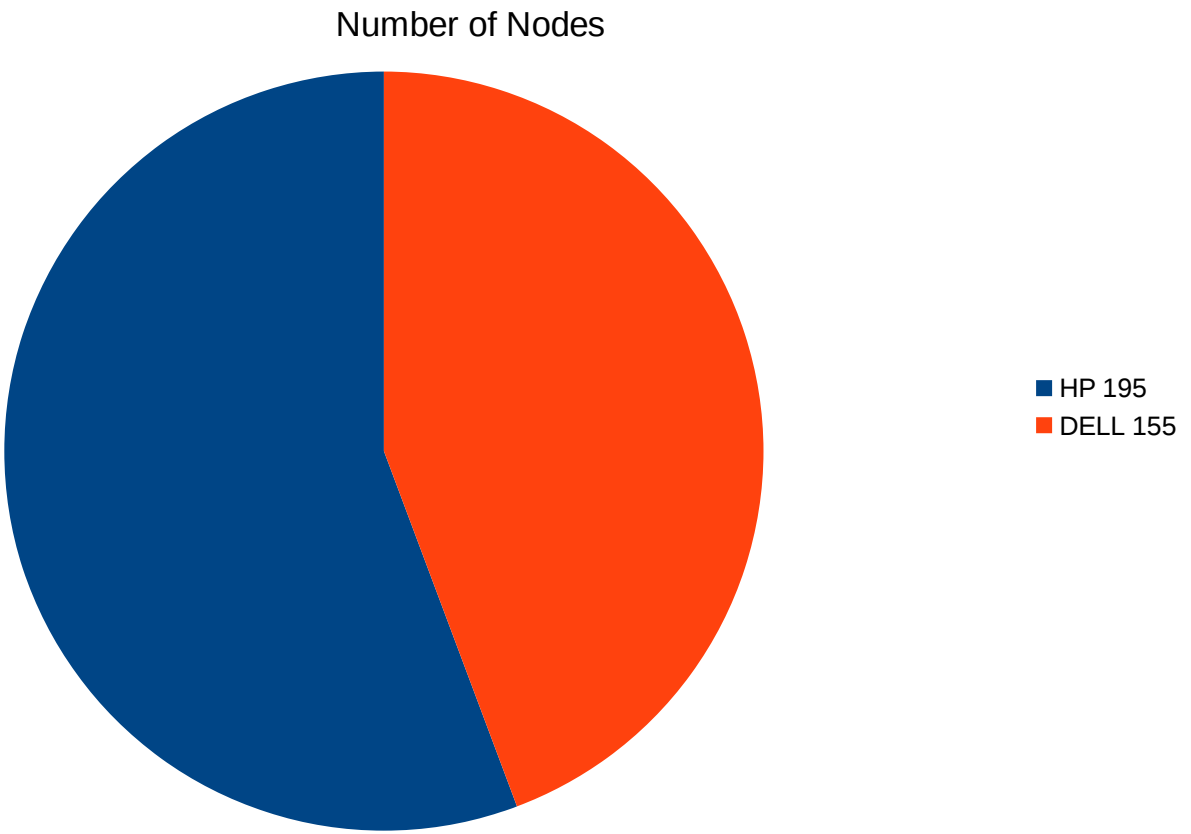
計算ノードの HEP-SPEC 比率



～ 2018 ハードウェア予定

計算ノードの比率

古い HP の計算ノードは低性能＋占有スペース大



2016.12 購入済 - DELL

約 10,000 HEP-SPECs

DELL enclosure + blade servers

M1000e x 1 + M630 x 16 blades

DELL enclosure + blade servers

M1000e x 1 + M630 x 16 blades

DELL storage servers + SAS controllers

PowerEdge R730 x 2

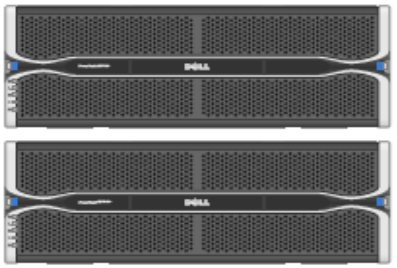
DELL server

PowerEdge R730 x 1

236,896 EUR (including TAX)

2016.12 購入済 - EUROstar

800TB



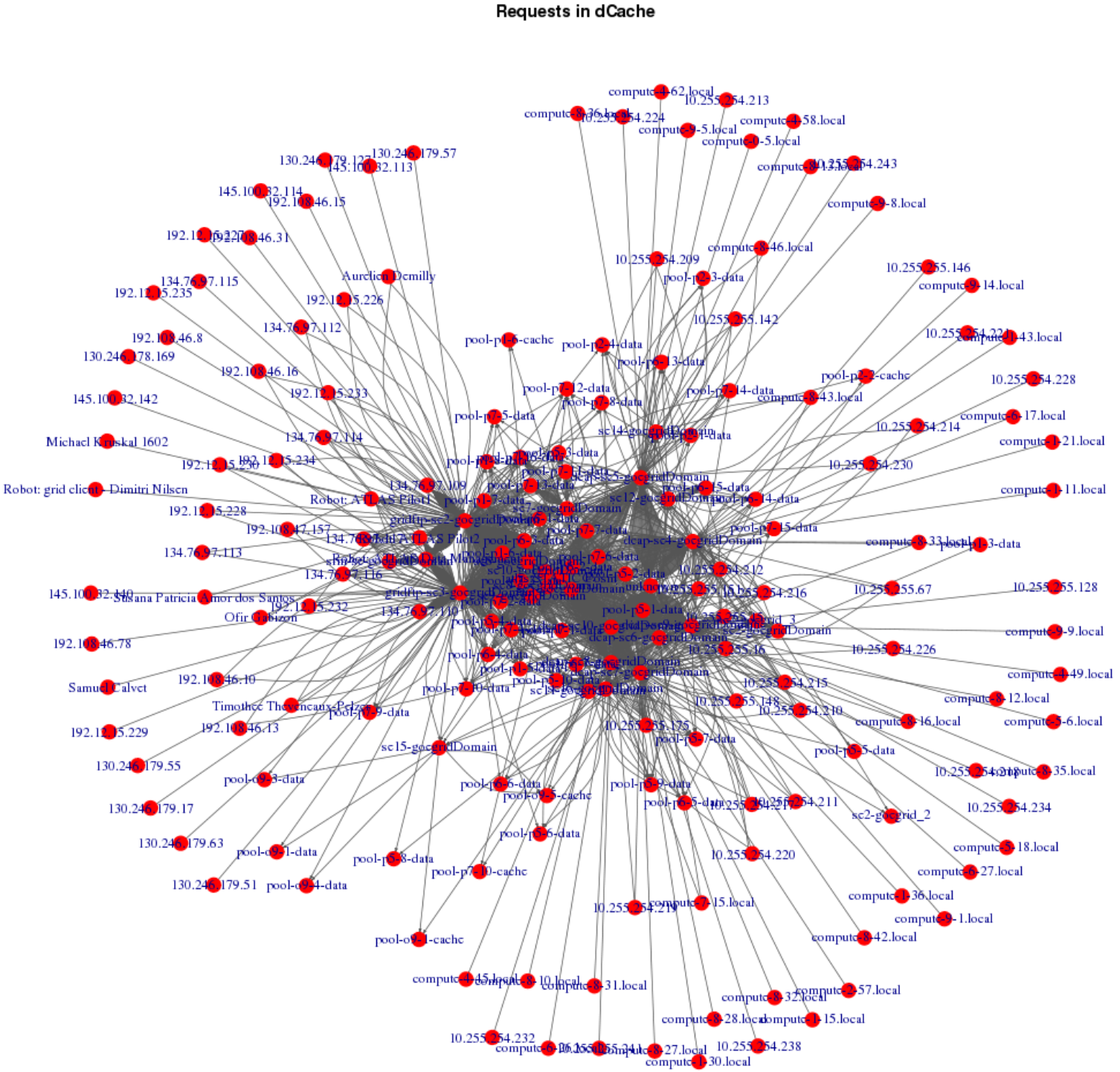
EUROstar Disk arrays

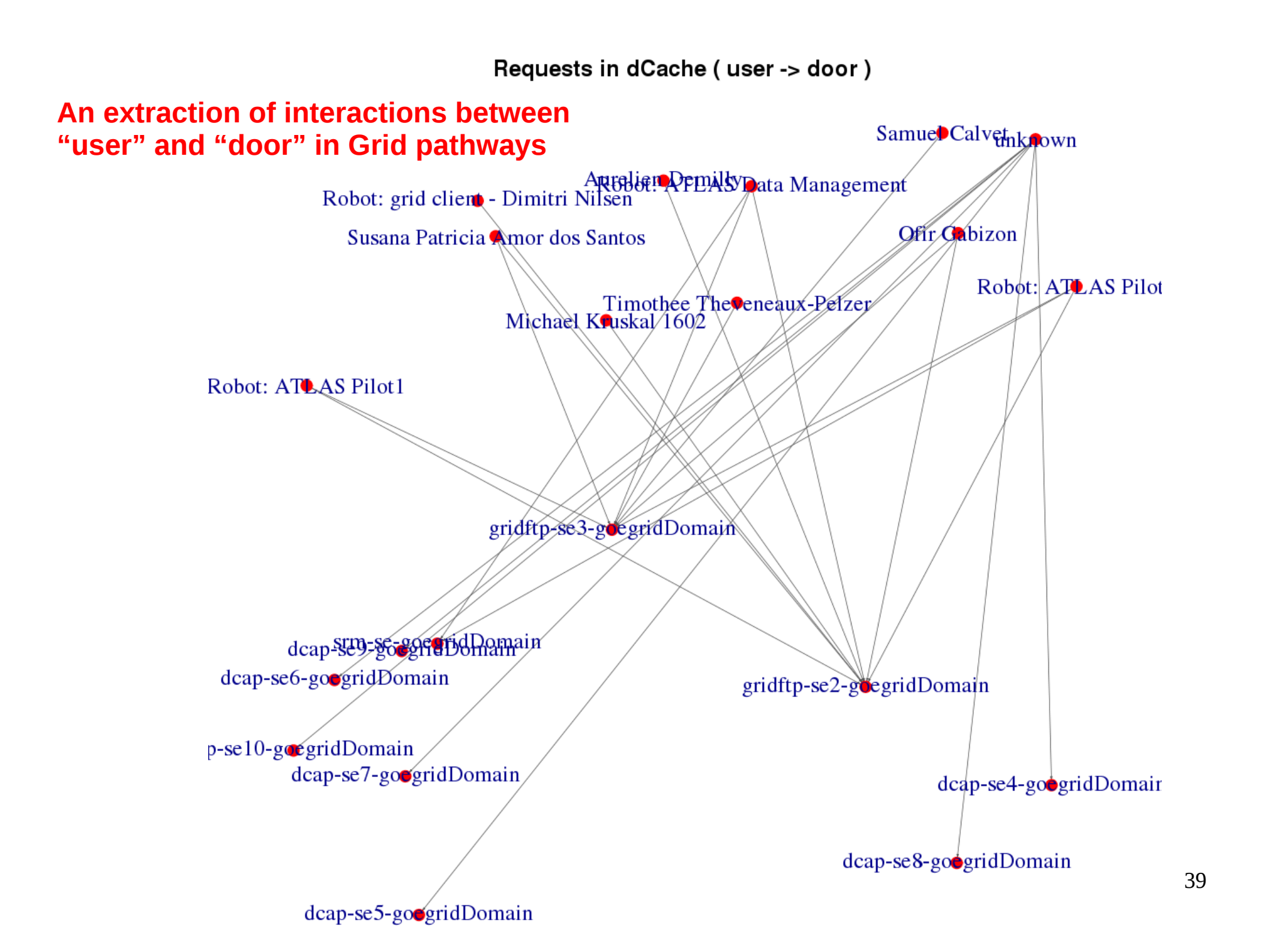
M1000e x 1 + M630 x 16 blades
+ SAS cable

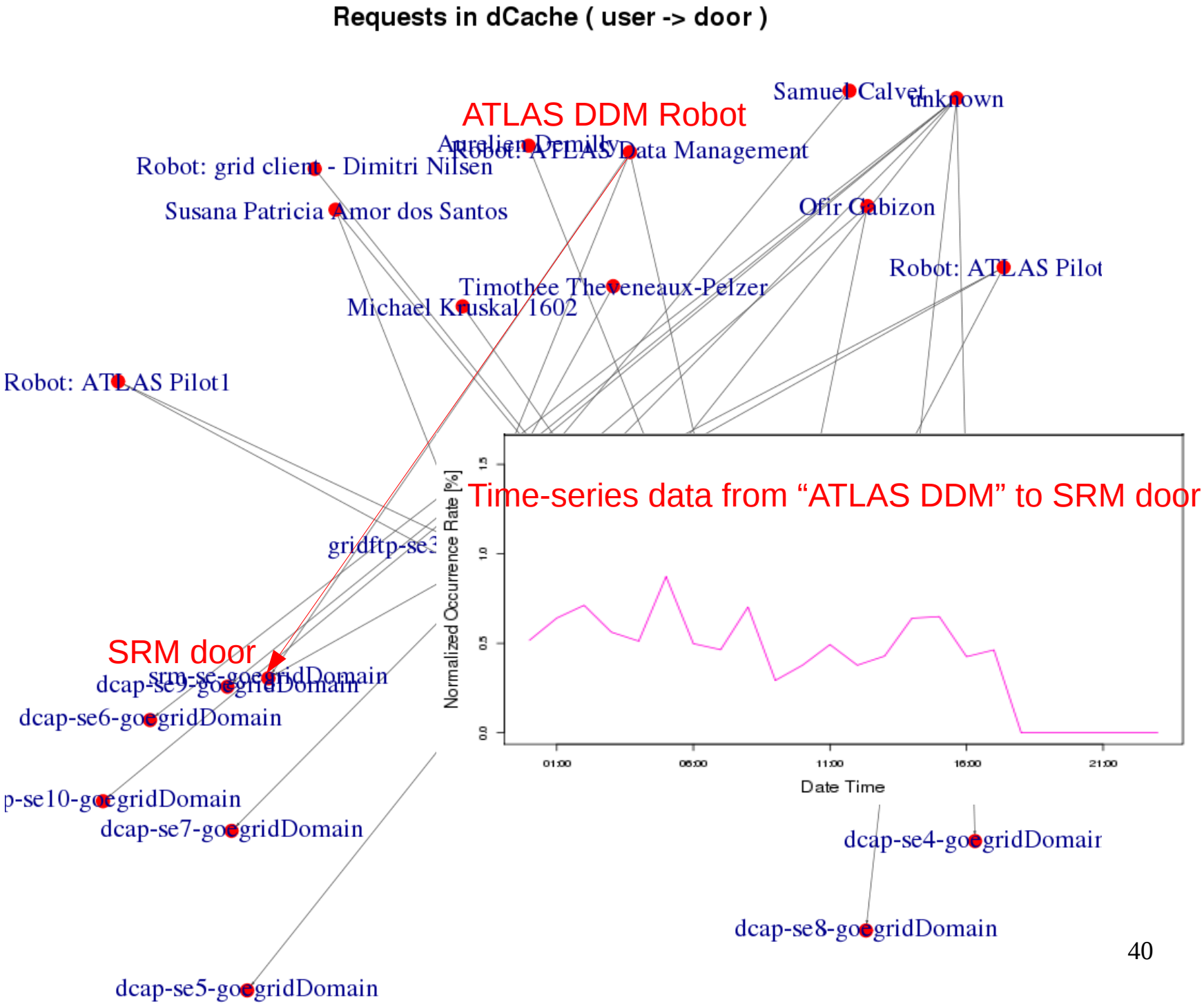
66,033 EUR (including TAX)

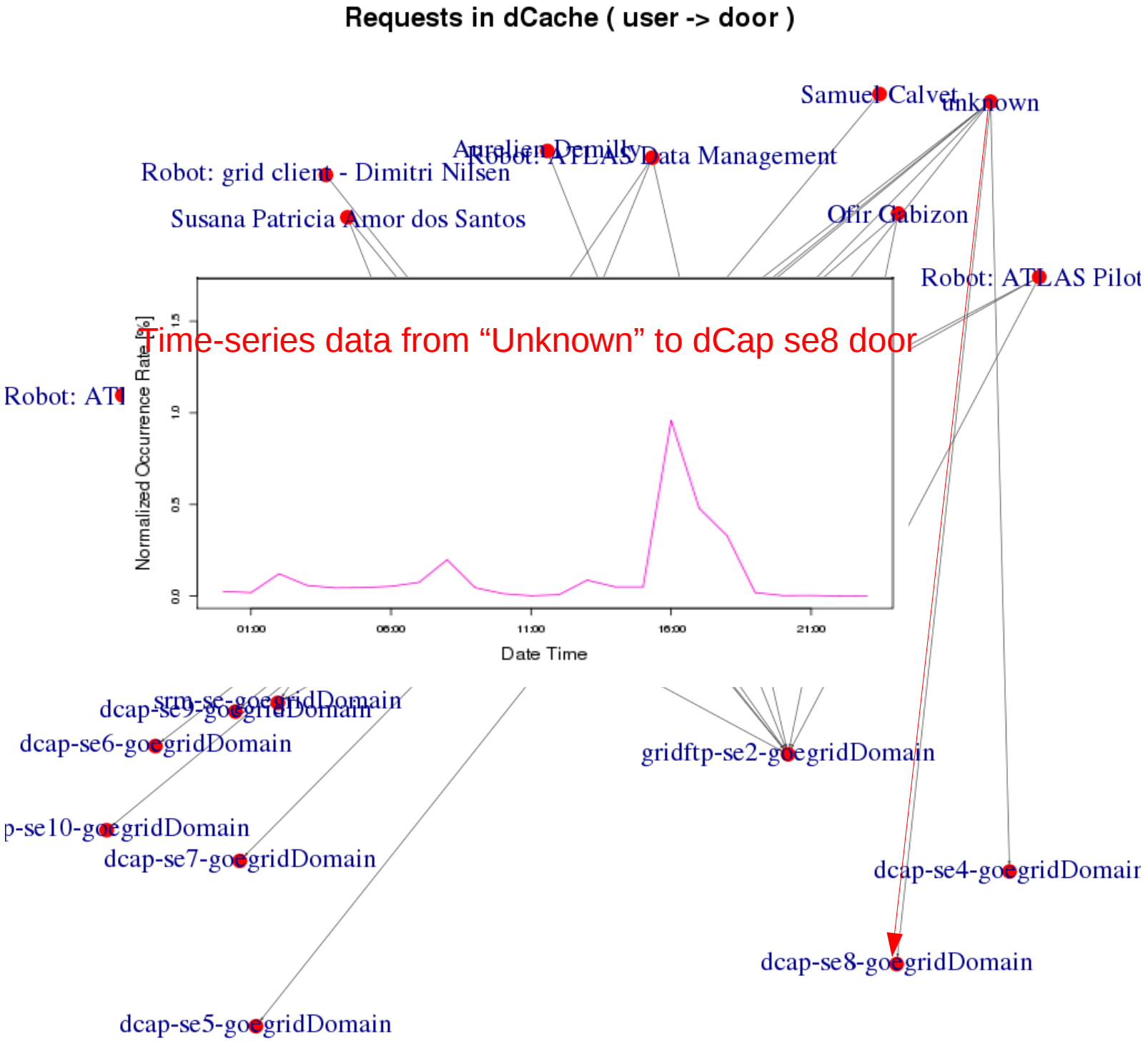
GoeGrid の可視化

- dCache ストレージログの可視化
  - 実装は R + graph 構造可視化ライブラリ
  - 1% のログデータを使用
    - 統計的検証はブートストラップ法を使用
    - 1% のログで信頼区間 95% 以内


Requests in dCache


Requests in dCache (user -> door)
An extraction of interactions between
“user” and “door” in Grid pathways
Samuel Calvet
unknown
Aurelien Demilly
Robot: ATLAS Data Management
Robot: grid client - Dimitri Nilsen
Susana Patricia Amor dos Santos
Ofir Gabizon
Robot: ATLAS Pilot
Timothee Theveneaux-Pelzer
Michael Kruskal 1602
Robot: ATLAS Pilot1
gridftp-se3-goegridDomain
srm-se-goegridDomain
dcap-se9-goegridDomain
dcap-se6-goegridDomain
gridftp-se2-goegridDomain
p-se10-goegridDomain
dcap-se7-goegridDomain
dcap-se4-goegridDomain
dcap-se8-goegridDomain
dcap-se5-goegridDomain


Requests in dCache (user -> door)
ATLAS DDM Robot
Samuel Calvet
unknown
Aurelien Demilly
Robot: ATLAS Data Management
Robot: grid client - Dimitri Nilsen
Susana Patricia Amor dos Santos
Ofir Gabizon
Robot: ATLAS Pilot
Timothee Theveneaux-Pelzer
Michael Kruskal 1602
Robot: ATLAS Pilot1
Time-series data from “ATLAS DDM” to SRM door
Normalized Occurrence Rate [%]
0.0
0.5
1.0
1.5
01:00
06:00
11:00
16:00
21:00
Date Time
SRM door
gridftp-se3
srm-se-goegridDomain
dcap-se9-goegridDomain
dcap-se6-goegridDomain
p-se10-goegridDomain
dcap-se7-goegridDomain
dcap-se4-goegridDomain
dcap-se8-goegridDomain
dcap-se5-goegridDomain


Requests in dCache (user -> door)
Samuel Calvet
unknown
Aurelien Demilly
Robot: ATLAS Data Management
Robot: grid client - Dimitri Nilsen
Susana Patricia Amor dos Santos
Ofir Gabizon
Robot: ATLAS Pilot
Time-series data from “Unknown” to dCap se8 door
Normalized Occurrence Rate [%]
0.0
0.5
1.0
1.5
01:00
06:00
11:00
16:00
21:00
Date Time
Robot: ATL
srm-se-goegridDomain
dcap-se9-goegridDomain
dcap-se6-goegridDomain
gridftp-se2-goegridDomain
p-se10-goegridDomain
dcap-se7-goegridDomain
dcap-se4-goegridDomain
dcap-se8-goegridDomain
dcap-se5-goegridDomain

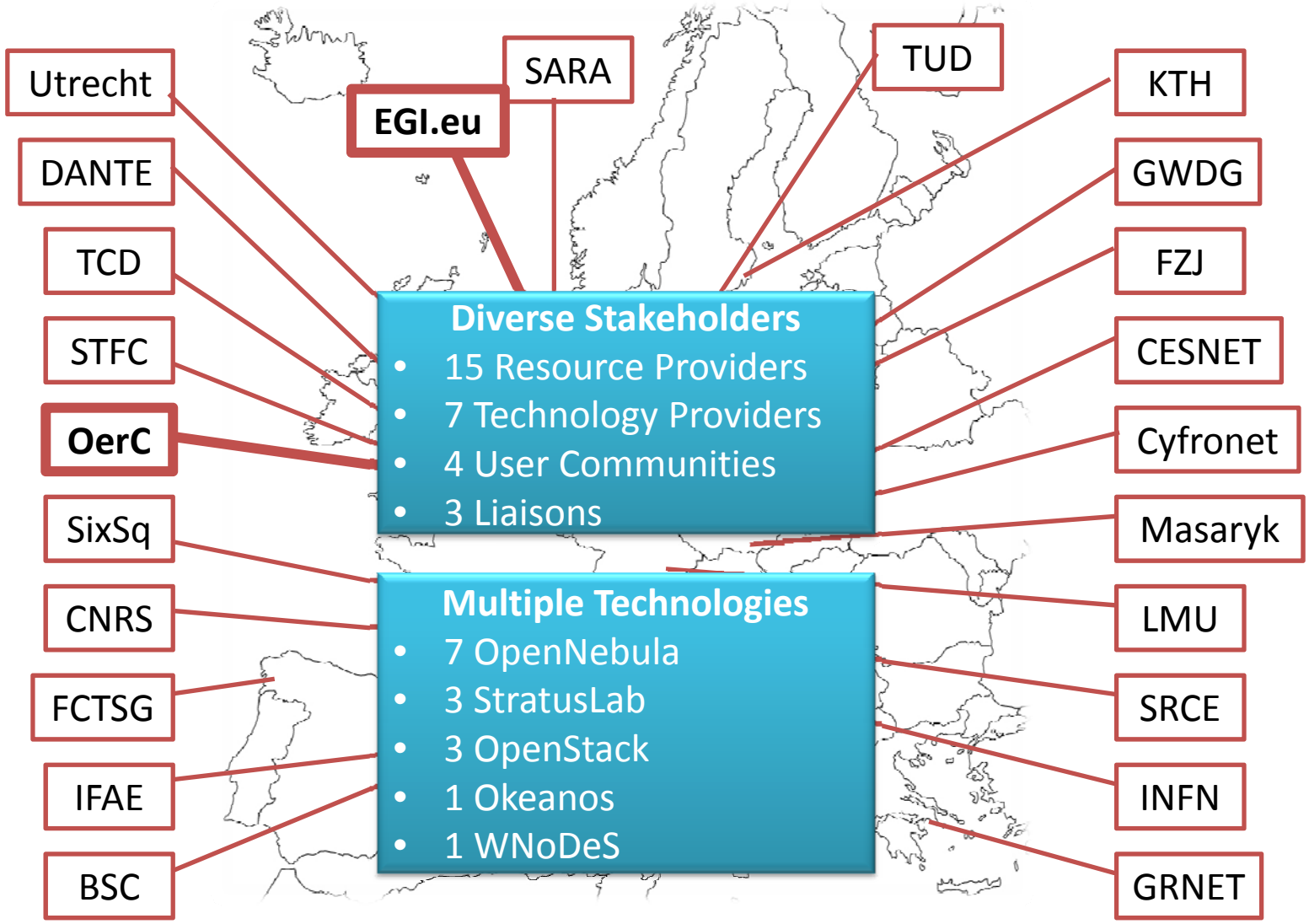
EGI 2020 and the Globus Community

Steven Newhouse

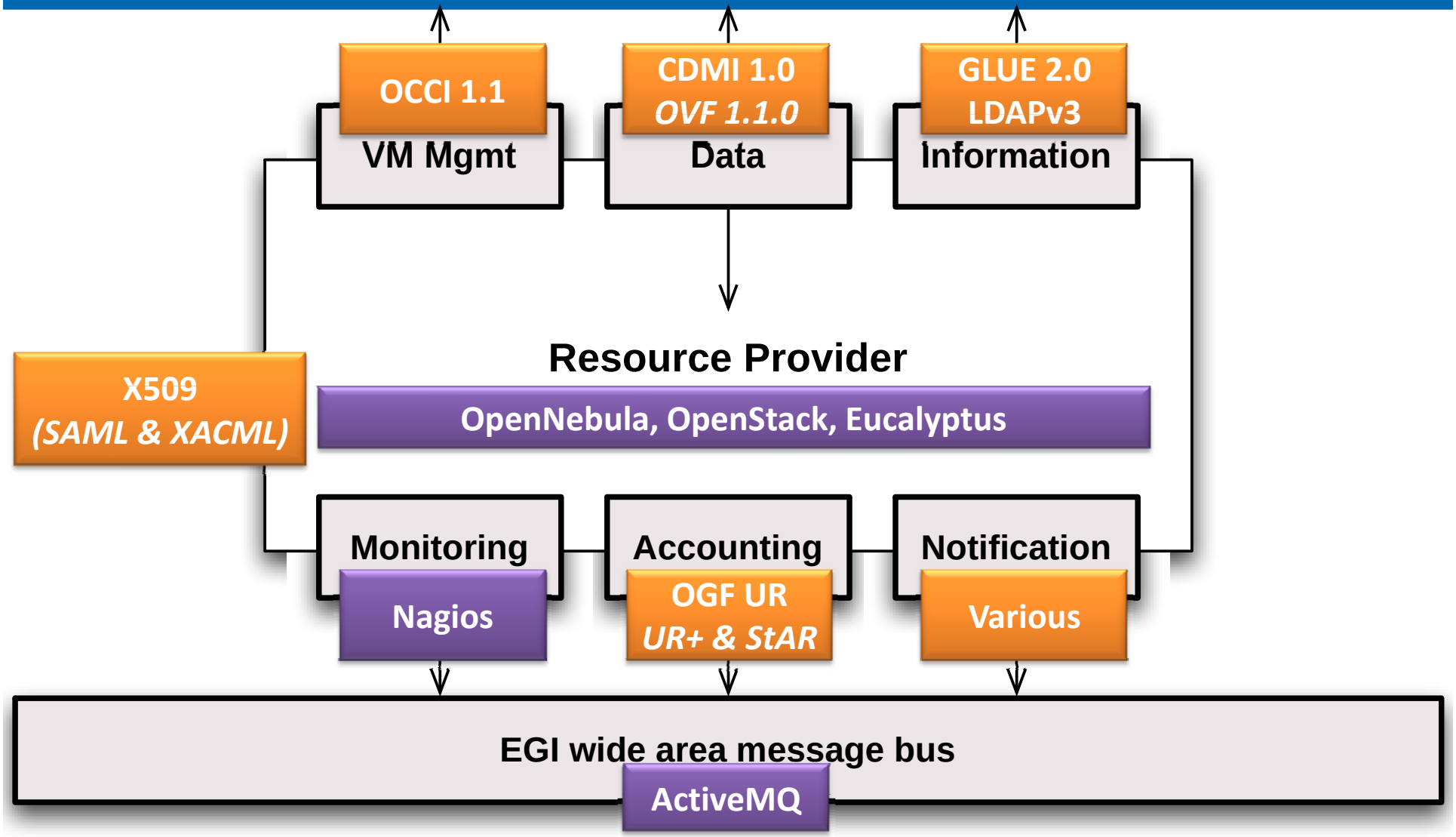
Director, EGI.eu

e-infrastructure

Federated Clouds



Standards & Technology



Test-bed status

Host	Service	Status	Last Check	Duration	Attempt	Status Information
149.156.10.30	StratusLab OpenNebula proxy	OK				
cagnode42.cs.tcd.ie	StratusLab OpenNebula proxy	OK				
carach5.ics.muni.cz	OCCI 0.8	OK				
	OCCI 1.1	OK				

List of records contained in the cloud accounting database.

Page last updated: 2012-03-23 14:01:23.220392

					User				

LDAP - GLUE2EndpointID=http://cdmi.cloud.cesga.de:4001/_CDMI,GLUE2ServiceID=cloud.service.CESGA-cloud_service,GLUE2GroupID=resource,GLUE2

LDAP Browser

- DIT
  - Root DSE (2)
    - o=glue (2)
      - GLUE2GroupID=resource
      - GLUE2GroupID=cloud (5)
        - GLUE2DomainID=CESGA-cloud (1)
          - GLUE2GroupID=resource (1)
            - GLUE2ServiceID=cloud.serv...e.CESGA-cloud_service (6)
              - GLUE2ResourceID=CESGA-cloud_ScientifcLinux
              - GLUE2ManagerID=cloud.service.CESGA-cloud_manager
              - GLUE2EndpointID=http://cd...cloud.cesga.de:4001/_CDMI
              - GLUE2EndpointID=http://oc...cloud.cesga.de:3200/_OCCI
              - GLUE2EndpointID=http://oc...cloud.cesga.de:3400/_OCCI
              - GLUE2EndpointID=https://o...d.cesga.de:8443/_SunStone
        - GLUE2DomainID=Cesnet-cloud
        - GLUE2DomainID=Cyfronet-cloud
        - GLUE2DomainID=GWDG-cloud
        - GLUE2DomainID=Kth-PDC-cloud
- Searches
- Bookmarks

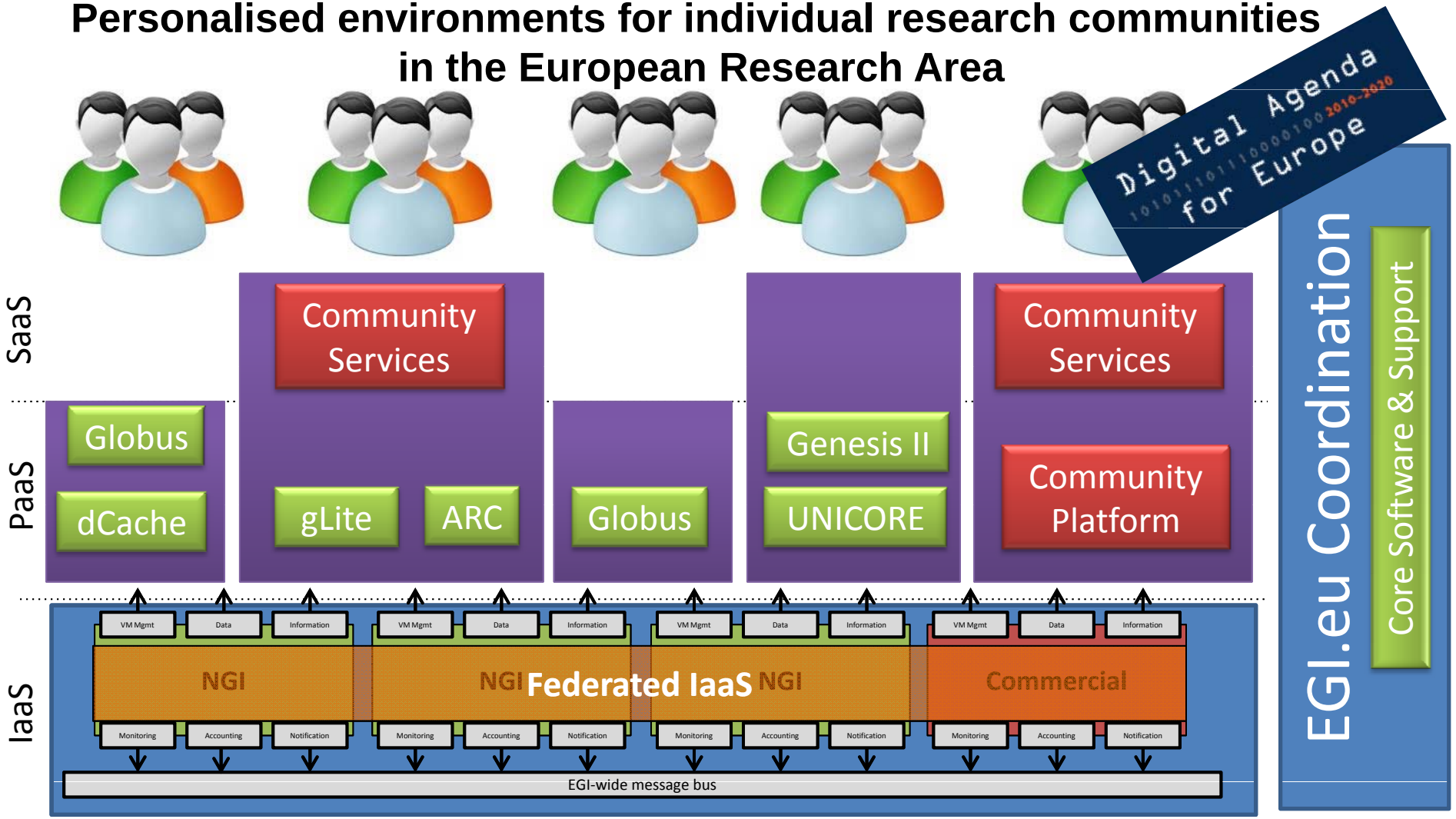
GLUE2EndpointID=http://cdmi.cloud.cesga.de:4001/_CDMI,GLUE2

DN: GLUE2EndpointID=http://cdmi.cloud.cesga.de:4001/_CDMI

Attribute Description	Value
objectClass	*GLUE2ComputingEndpoint (au*
objectClass	*GLUE2Endpoint (structural)*
objectClass	*GLUE2Entity (abstract)*
GLUE2EndpointHealthState	ok
GLUE2EndpointID	http://cdmi.cloud.cesga.de:40
GLUE2EndpointInterfaceName	CDMI
GLUE2EndpointQualityLevel	production
GLUE2EndpointServiceForeignK	cloud.service.CESGA-cloud_se
GLUE2EndpointServingState	production
GLUE2EndpointURL	http://cdmi.cloud.cesga.de:40
GLUE2ComputingEndpointCompu	cloud.service.CESGA-cloud_ser
GLUE2EndpointCapability	cloud.managementSystem,clou
GLUE2EndpointImplementationNa	OpenNebula
GLUE2EndpointImplementationVe	3.2
GLUE2EndpointImplementor	OpenNebula
GLUE2EndpointInterfaceVersion	NA
GLUE2EndpointTechnology	REST

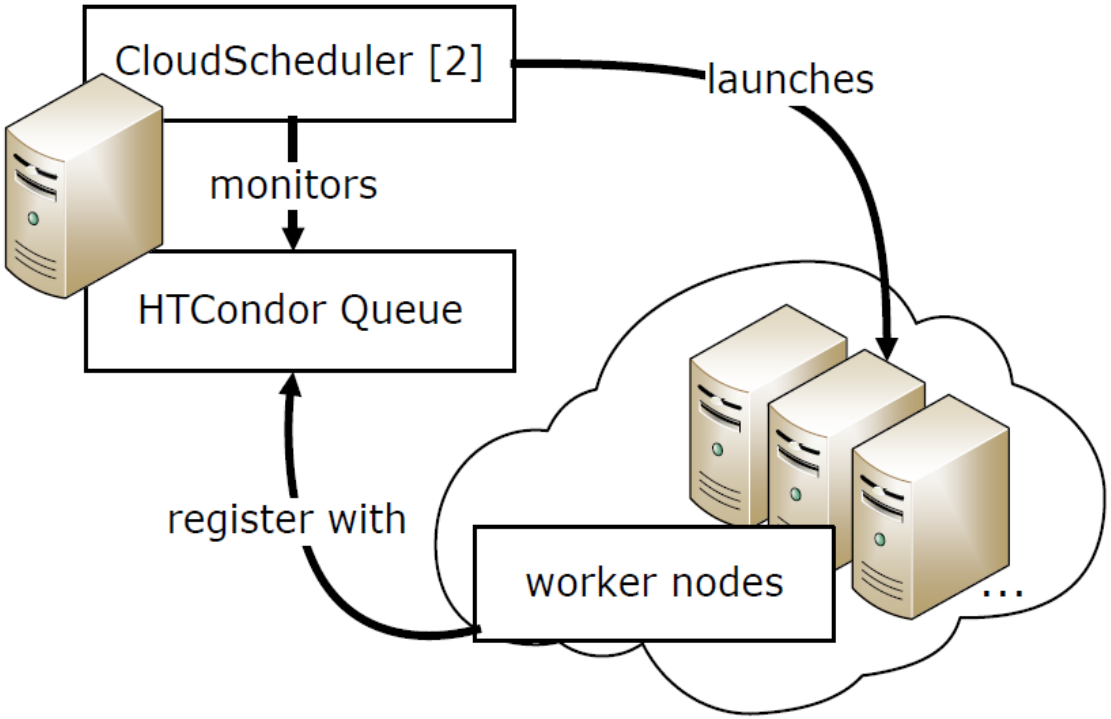
Personalised Environments

Personalised environments for individual research communities in the European Research Area



クラウドテスト

- GWDG Cloud
  - 結合テストベッド
  - HTCondor+CREAMCE
  - FAX でデータ転送



GoeGrid Tier-2 Computing Center

GoeGrid racks at GWDG

The GoeGrid Cluster is integrated in the WLCG as a Tier-2 center.

- Grid (CREAM x 2, dCache)
- 3000 cores
- 1.2PB storage
- SAN connected
- 10Gbit network

FAX

GWDG Compute Cloud

The GWDG Compute Cloud is a private Infrastructure as a Service Cloud based on OpenStack which can be utilized by researchers of the University of Göttingen and the local Institutes of the Max-Planck Society.
Current Size:

- 52 nodes
- ca. 3300 cores,
- ca. 15 TB memory
- ca. 300 TB NFS storage,
- 12 times 10 Gbit network (redundant)

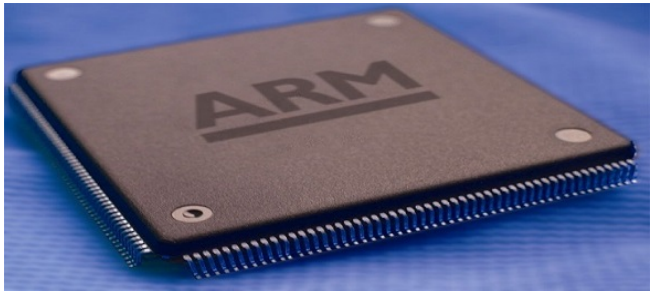
GWDG
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

GoeGrid Tier-2 まとめ

- ドイツの中堅 Tier-2 サイト
  - 5 ～ 8% 程度の ATLAS DE 資源を供給
  - マンパワーは約 1FTE (通常 DE UNI-T2 は 2 FTE)
  - Run-2 データ解析のため資源を約 2 倍に増強中
    - 9,433 HS06 CPU + 1.2PB →19,300 HS06 CPU + 2.0PB
  - ミドルウェア・管理システムをアップグレード中
  - EGI クラウド統合タスクフォースなどのテストベッド

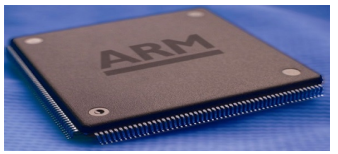
Göttingen ATLAS 物理計算グループの研究トピックス

Cloud Computing



方向

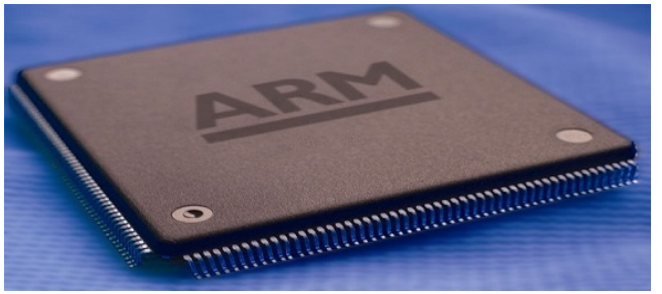
- 将来必要になりそうなもの
- （高エネ解析の視点から） ATLAS の役に立ちそうなもの
- 実用性・有用性・汎用性があるもの
- コンピューター市場の動向に一致しているもの
  - ミーハーは市場無視より良い
- 面白そうなもの
- 新規性があるもの
- *そのうち自分たちが楽が出来そうなもの*



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



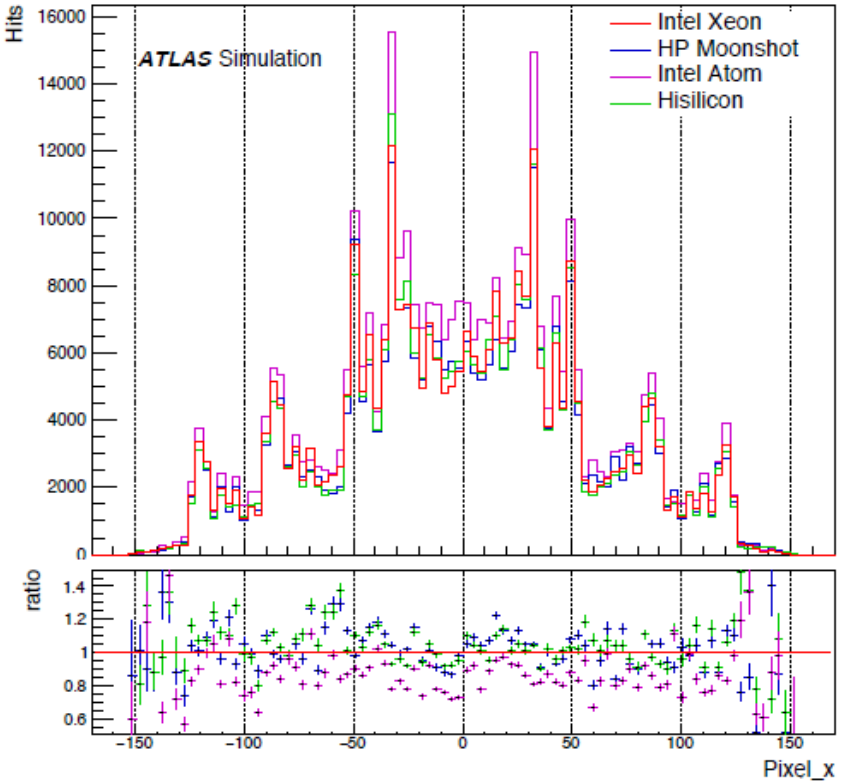
ATLAS ソフトウェア資源の ARM アーキテクチャへの移植

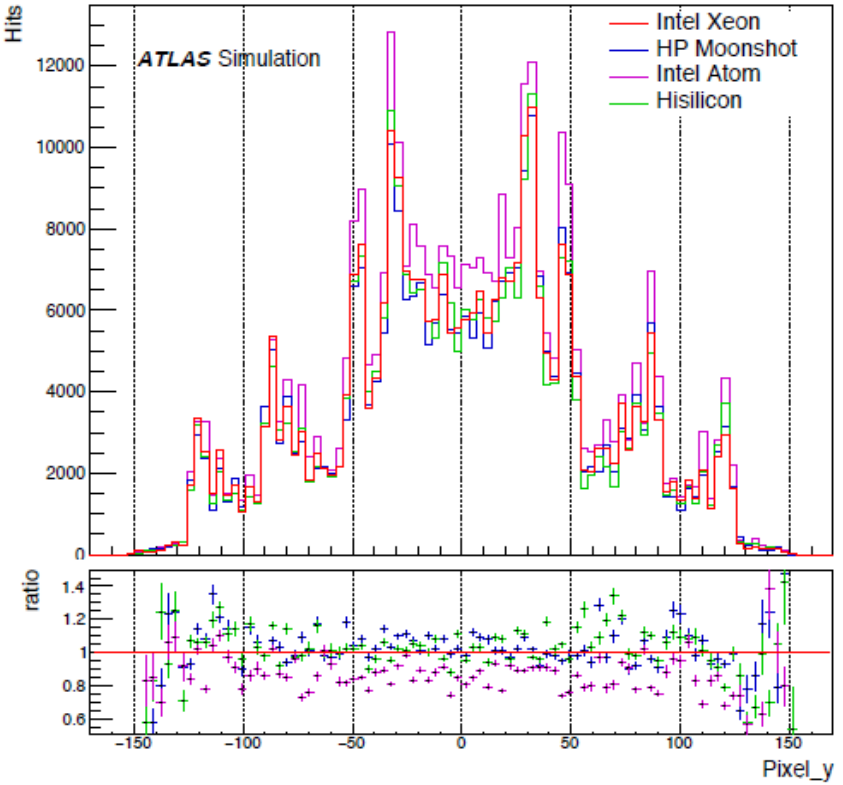
- What did we port?
  - Athena stack overkill
  - Picked AthSimulation
    - A fraction of the packages of Athena (~345 compared to ~2400)
    - Much quicker compile time
    - Potential for errors in port decreases
    - Geant4 gives a good CPU load
    - Good for simulation and validation
    - Implement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

Remove CMT!

ATLAS ソフトウェア資源の ARM アーキテクチャへの移植

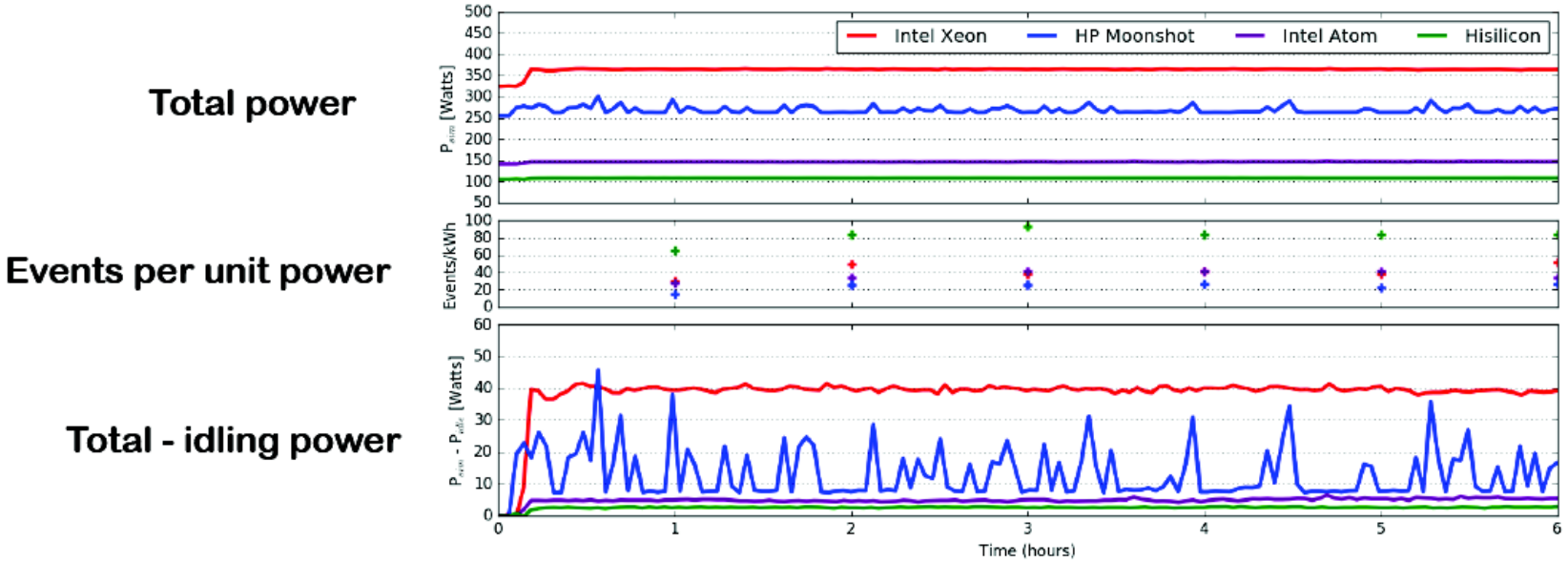
- Event 生成





ATLAS ソフトウェア資源の ARM アーキテクチャへの移植

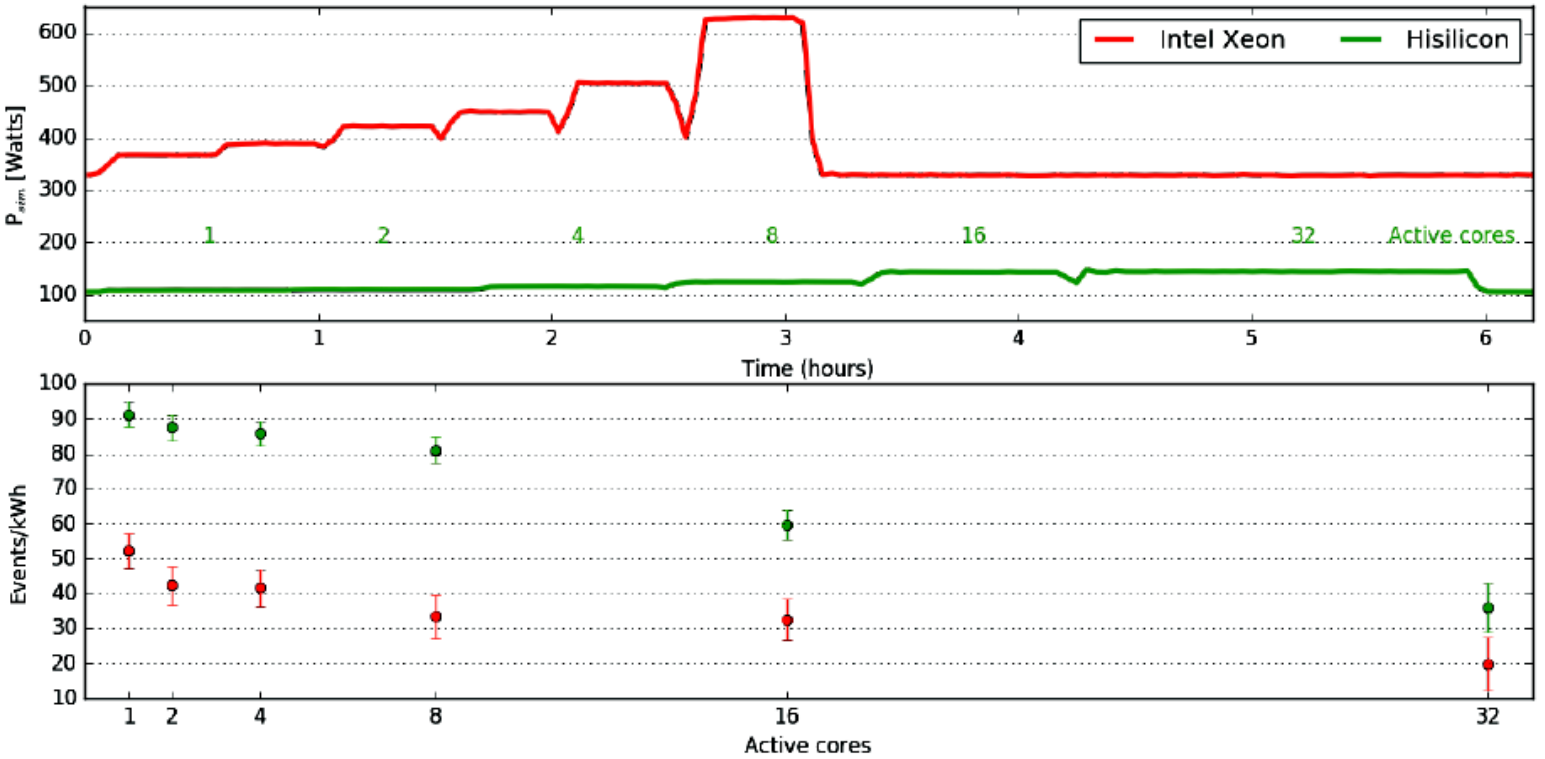
- ttbar イベント生成と消費電力



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



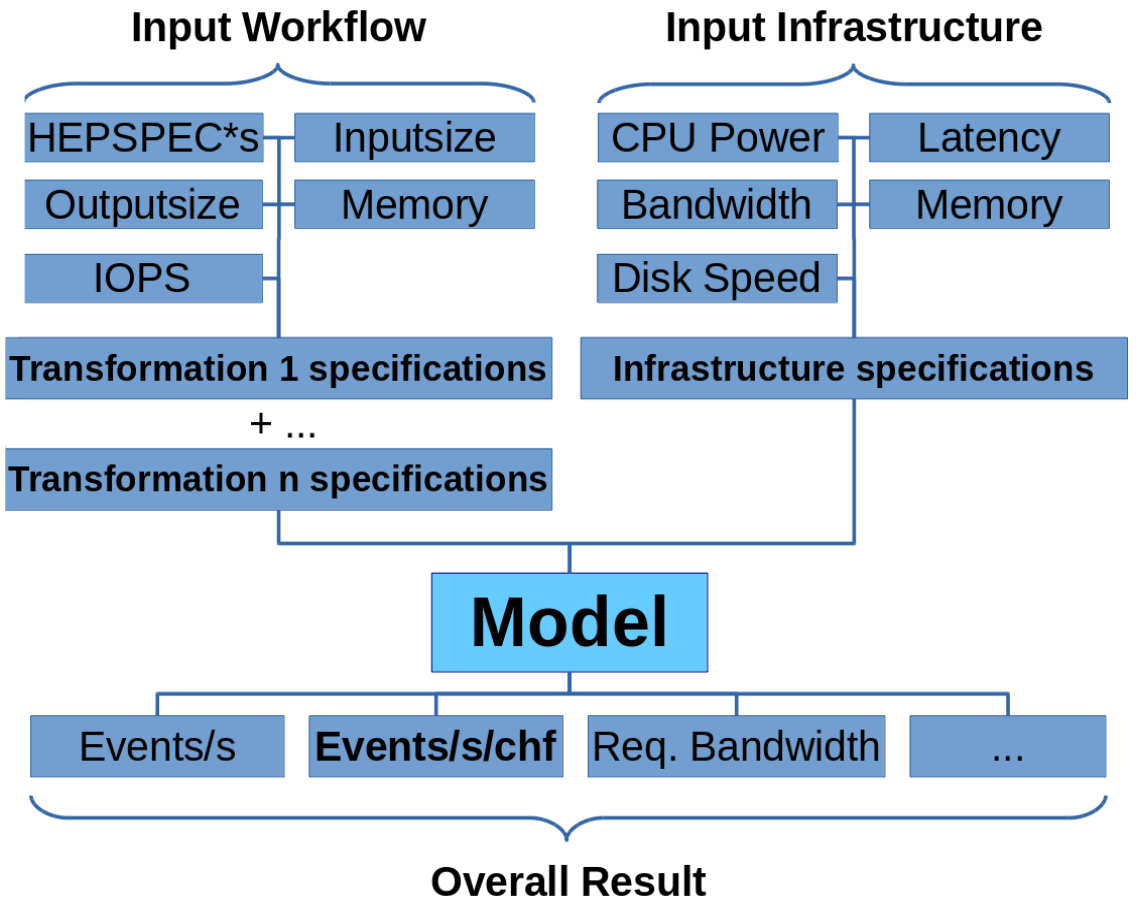
クラウドコンピューティング

- 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

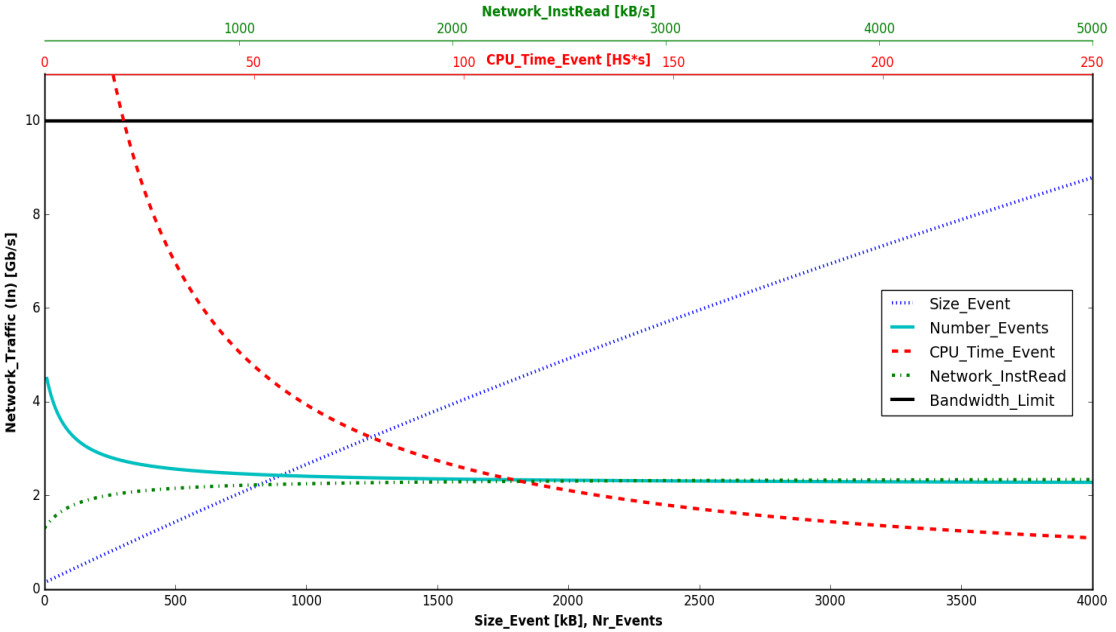
クラウドコンピューティング

- 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

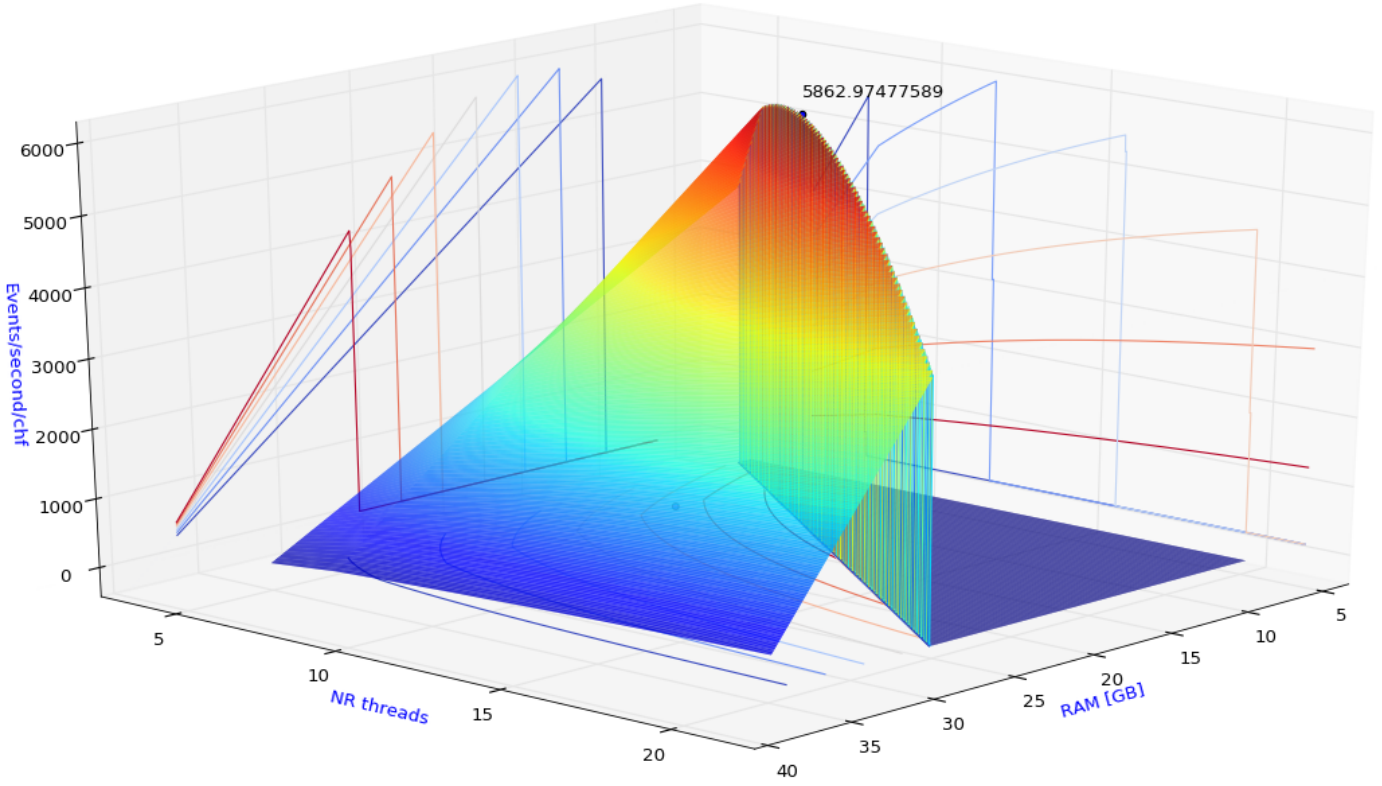
- 最適値を探索
  - モデル探索
  - Overcommittment
    - I/O 待機時間を減らす



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

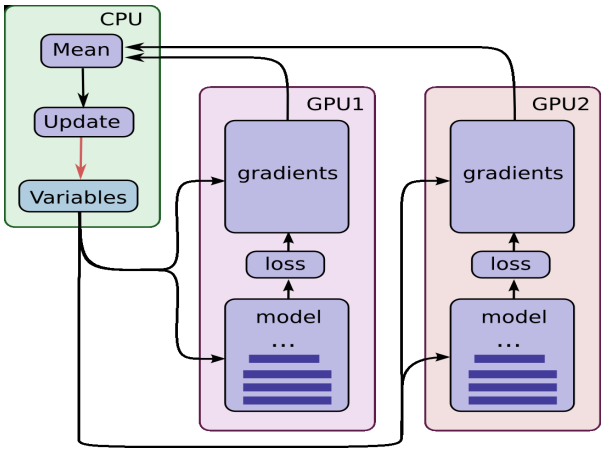
クラウドコンピューティング

- 最高値 5863 Events/s/chf
  - 14 GB RAM/machine
  - Overcommittment: 11 threads/machine



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

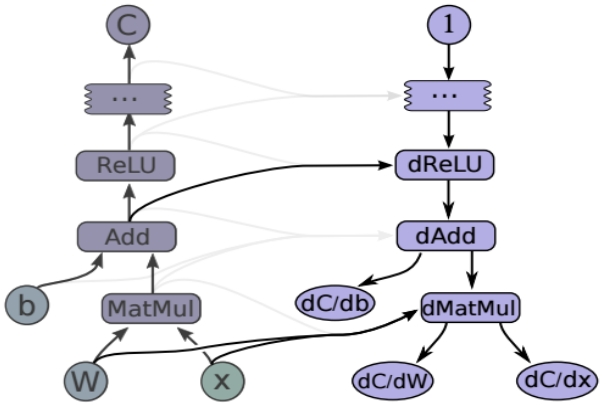
Table 1 | Performance for Higgs benchmark.

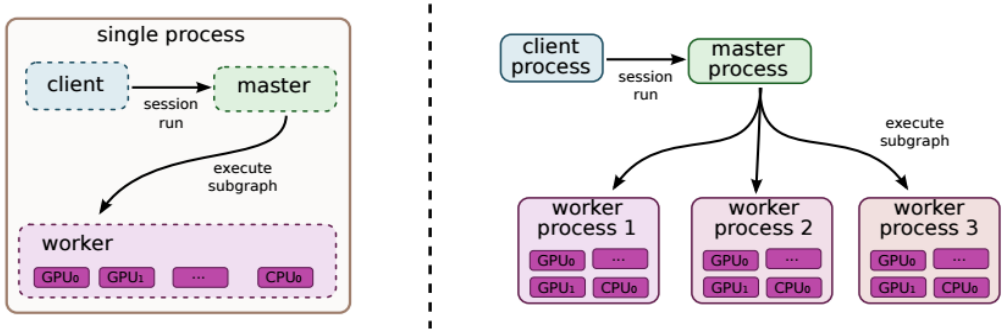
Technique	Low-level	High-level	Complete
AUC			
BDT	0.73 (0.01)	0.78 (0.01)	0.81 (0.01)
NN	0.733 (0.007)	0.777 (0.001)	0.816 (0.004)
DN	0.880 (0.001)	0.800 (<0.001)	0.885 (0.002)
Discovery significance			
NN	2.5σ	3.1σ	3.7σ
DN	4.9σ	3.6σ	5.0σ

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

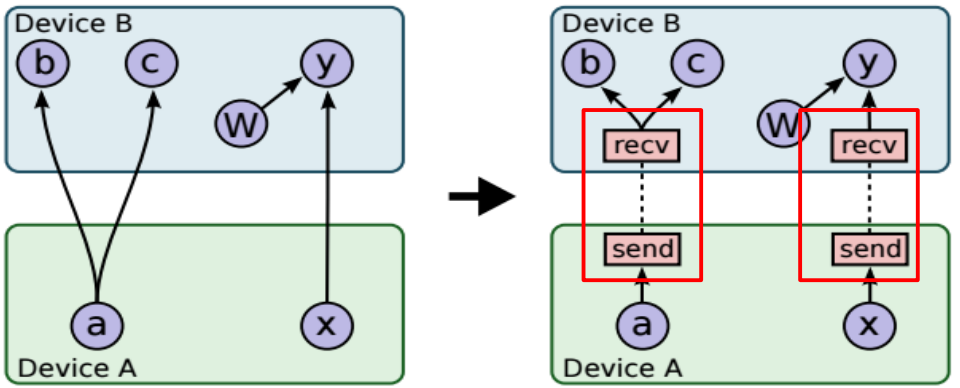
Google TensorFlow ライブラリと分散コンピューティング

- 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





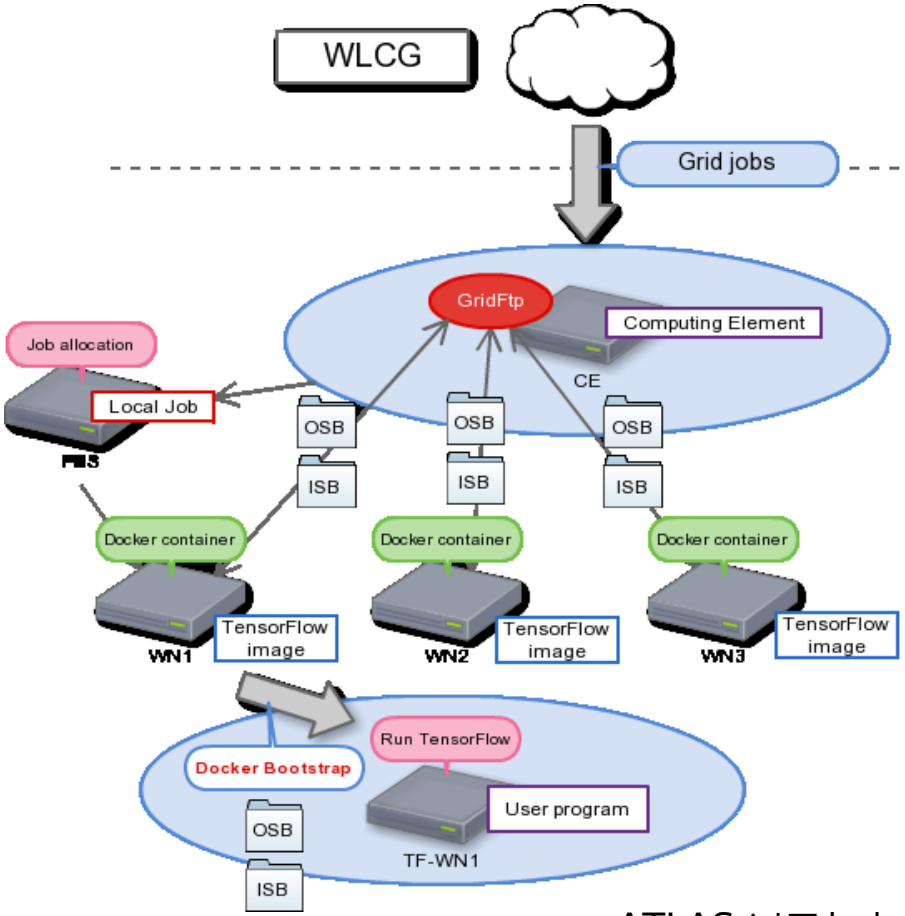
計算処理の分割実行と結合



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

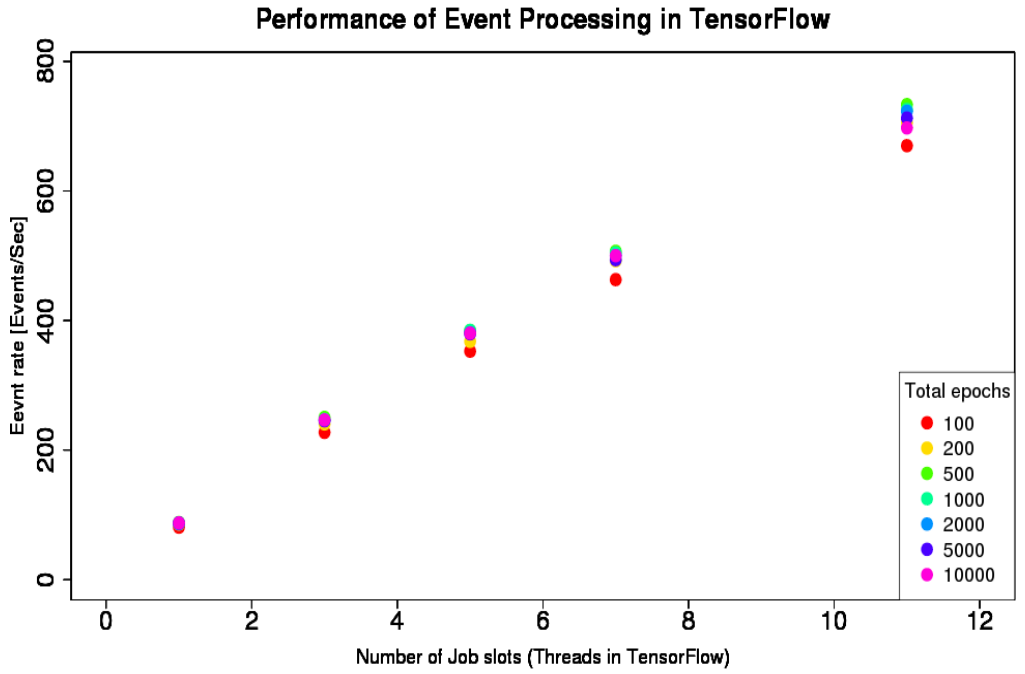
- Grid Docker 環境用テストキュー
  - Grid CE から自作ブートストラップで Docker 用 TensorFlow 環境を計算ノードへロード

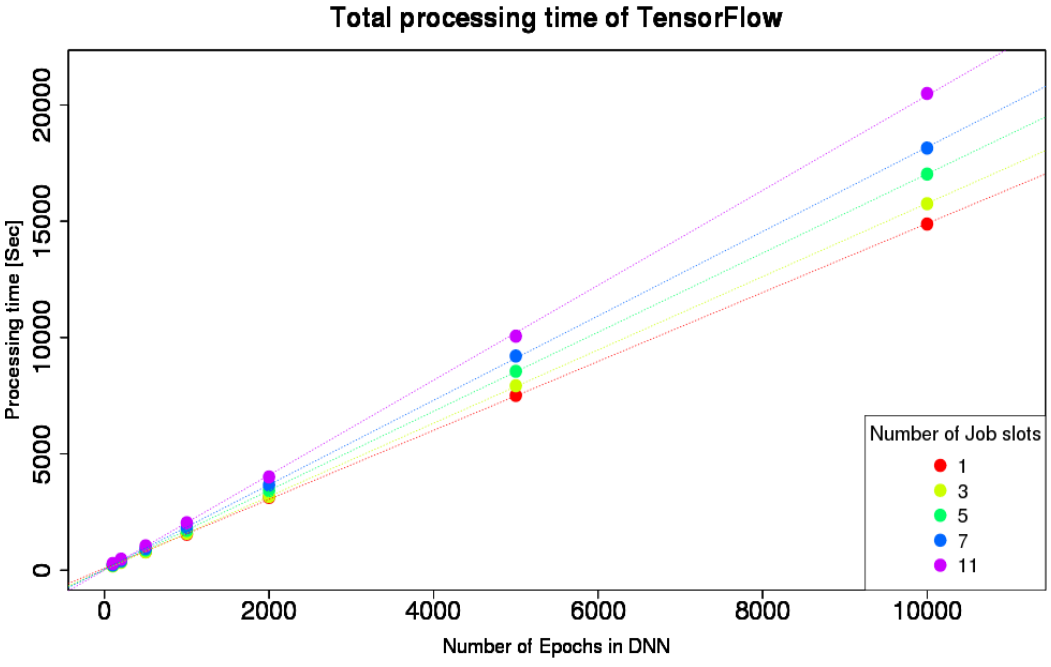


- プロダクション・レベルでの考慮点
  - アカウンティング（使用量の記録）
  - 安全性 (！)
  - 堅牢性

Google TensorFlow ライブラリと分散コンピューティング

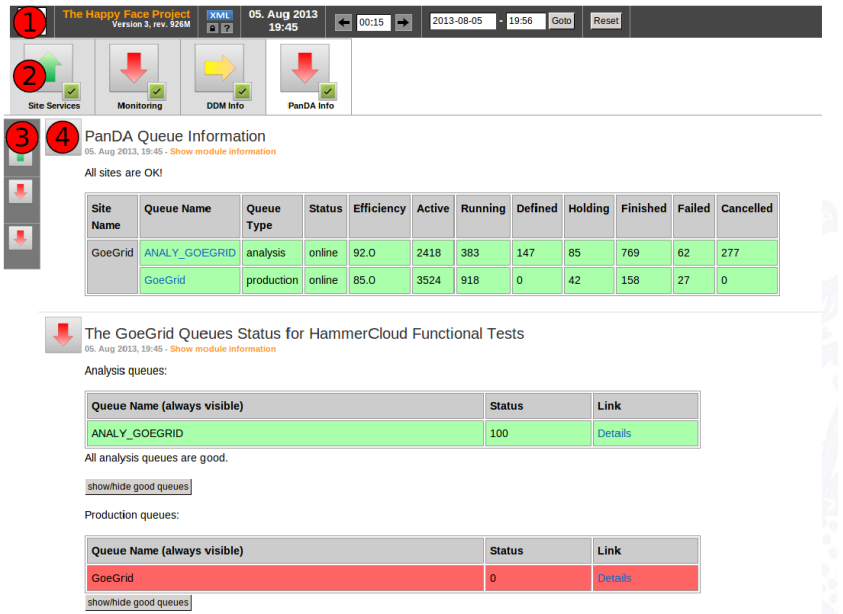
- Convolutional DNN でのイメージ識別学習時間
  - TensorFlow v0.8 CPU モードでテスト (1 thread / 1 docker VM)
  - Event processing rate はリニアに上昇
  - Event processing 以外での（通信）遅延が大きい
    - > 10 jobs で TensorFlow (CPU mode) 自体がまだ不安定。





メタモニタリングシステム（ HappyFace, MadFace ）

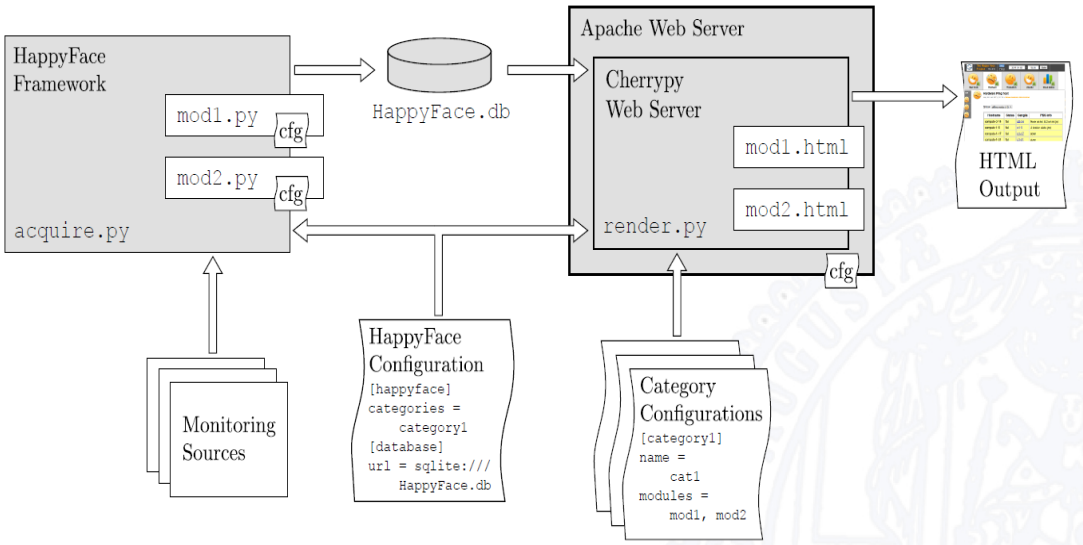
- メタモニタリングシステムとは？
  - モニタリングのモニタリングのためのシステム
- HappyFace meta-monitoring system in CMS
  - ドイツ Tier1 モニタリングのためのメタモニタリングシステムとして開発 (KIT)
    - Python フルで実装
    - SQL DB backend
    - 1 web server
    - モジュラー構造



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

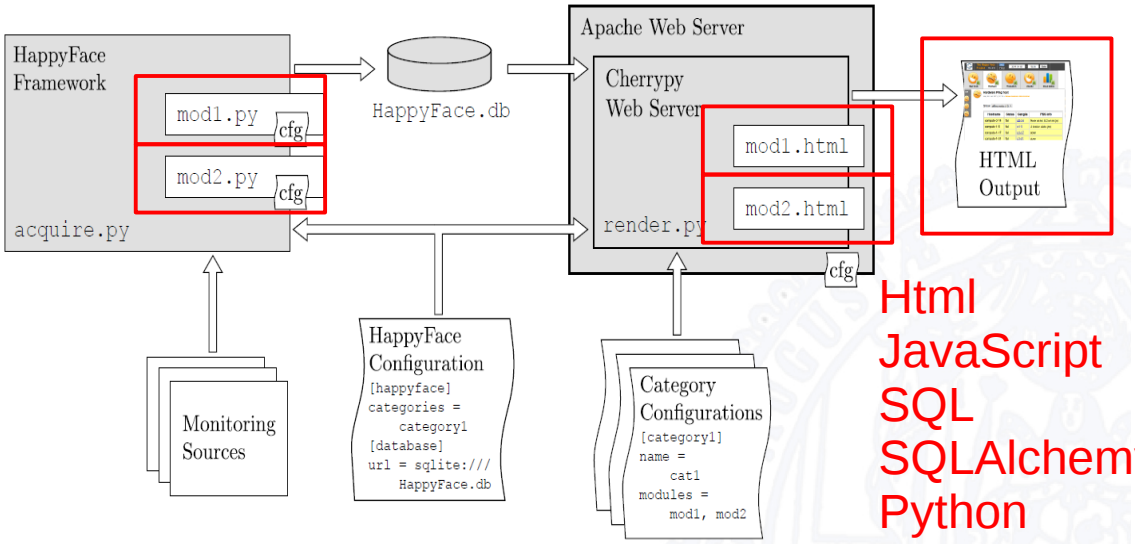
- HappyFace meta-monitoring system in CMS
  - ドイツ Tier1 モニタリングのためのメタモニタリングシステムとして開発 (KIT)
    - Python フルで実装
    - SQL DB backend
    - 1 web server
    - モジュラー構造



メタモニタリングシステム（ HappyFace, MadFace ）

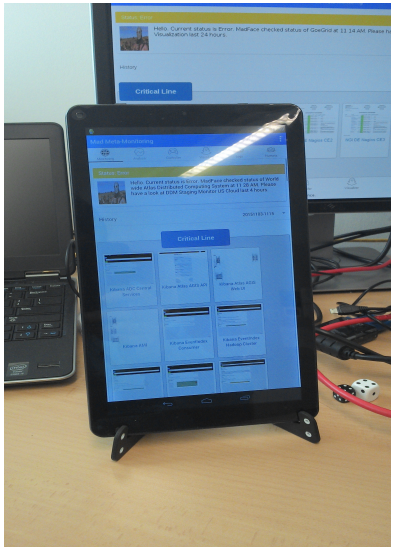
- メタモニタリングシステムとは？
  - モニタリングのモニタリングのためのシステム
- HappyFace meta-monitoring system in CMS
  - ドイツ Tier1 モニタリングのためのメタモニタリングシステムとして開発 (KIT)
    - Python フルで実装
    - SQL DB backend
    - 1 web server
    - モジュラー構造

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 ）

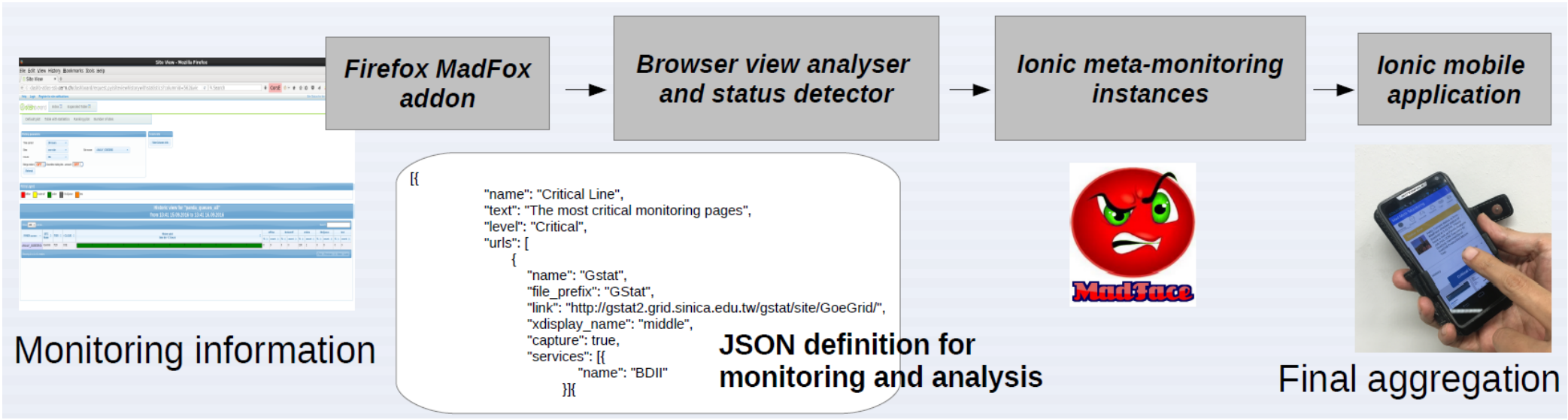
- MadFace メタモニタリングシステム
- 最新のモバイルフレームワークで開発
  - Web + mobile フレームワーク
    - Ionic, AngularJS framework
    - Server-side JavaScript technology
    - Firefox + ***Madfox*** (JetPack Manager)
  - オンラインベイズ分析機
    - R, bayesian change process
    - Bayesian network
- プロトタイプのコーディング時間
  - 約 300 時間 (By Gen Kawamura)



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

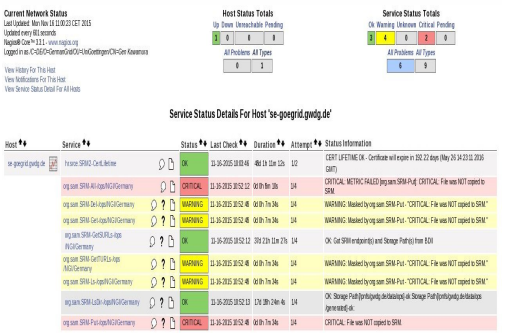


メタモニタリングシステム（ HappyFace, MadFace ）

- 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

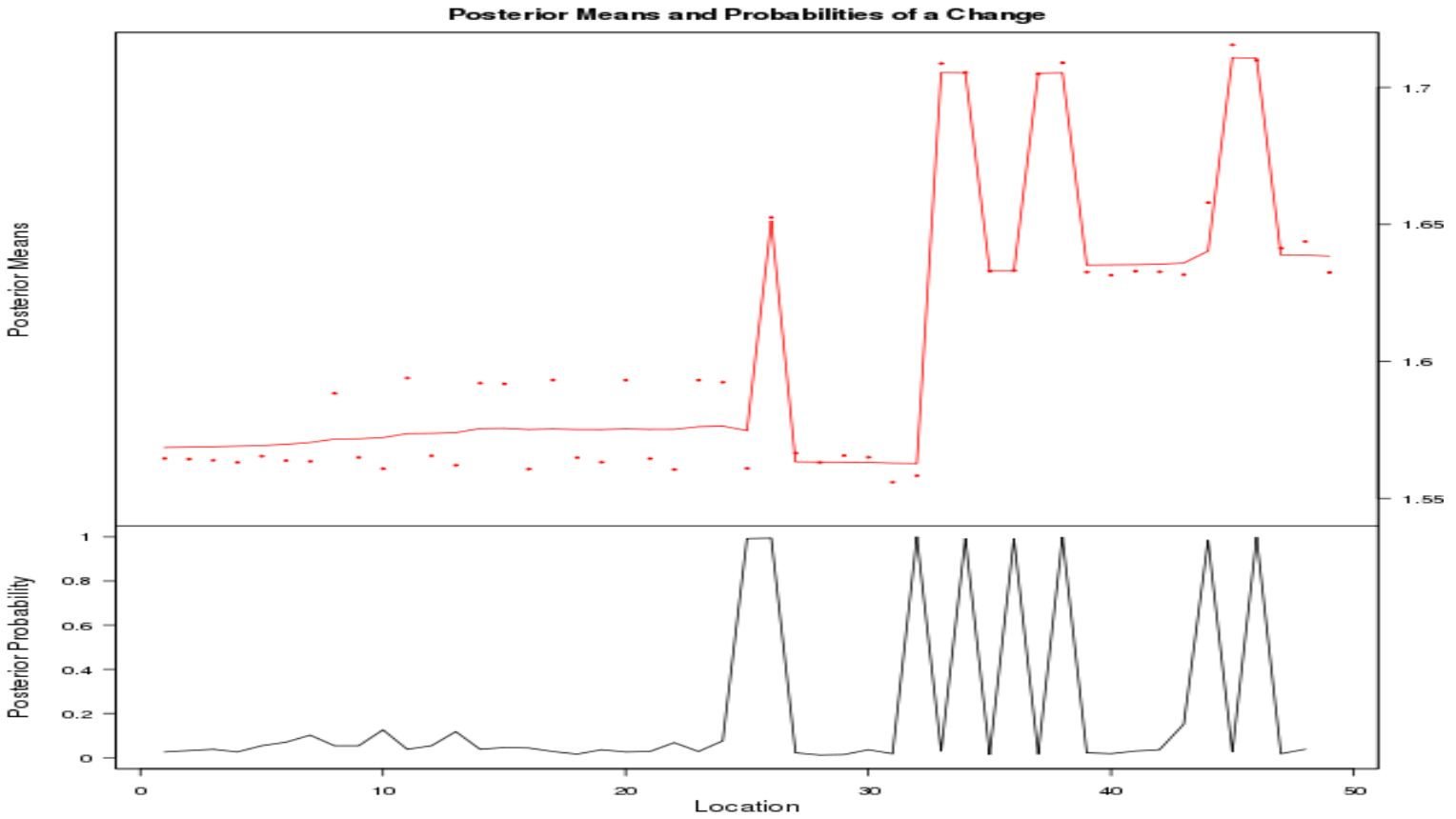
- ブラウザ画面を情報量へ変換し、システムの状態を検出
  - 画面情報を参照画面情報との情報量の差へ変換（ KL 偽距離）
  - KL 偽距離の変化状態をバイナリコーディングで近似化
  - ベイジアン事後変化確率を計算
    - > 0.8 ならベイジアン事後確率変化点
  - ブラウザサイトの重要度のグルーピングにより情報を補強
    - G1 = Sign(w1 B1+w2 B2+ w3 B3 …)
    - 例えば、重要なウェブページのうち２つが変化していたら”重要な変化”

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

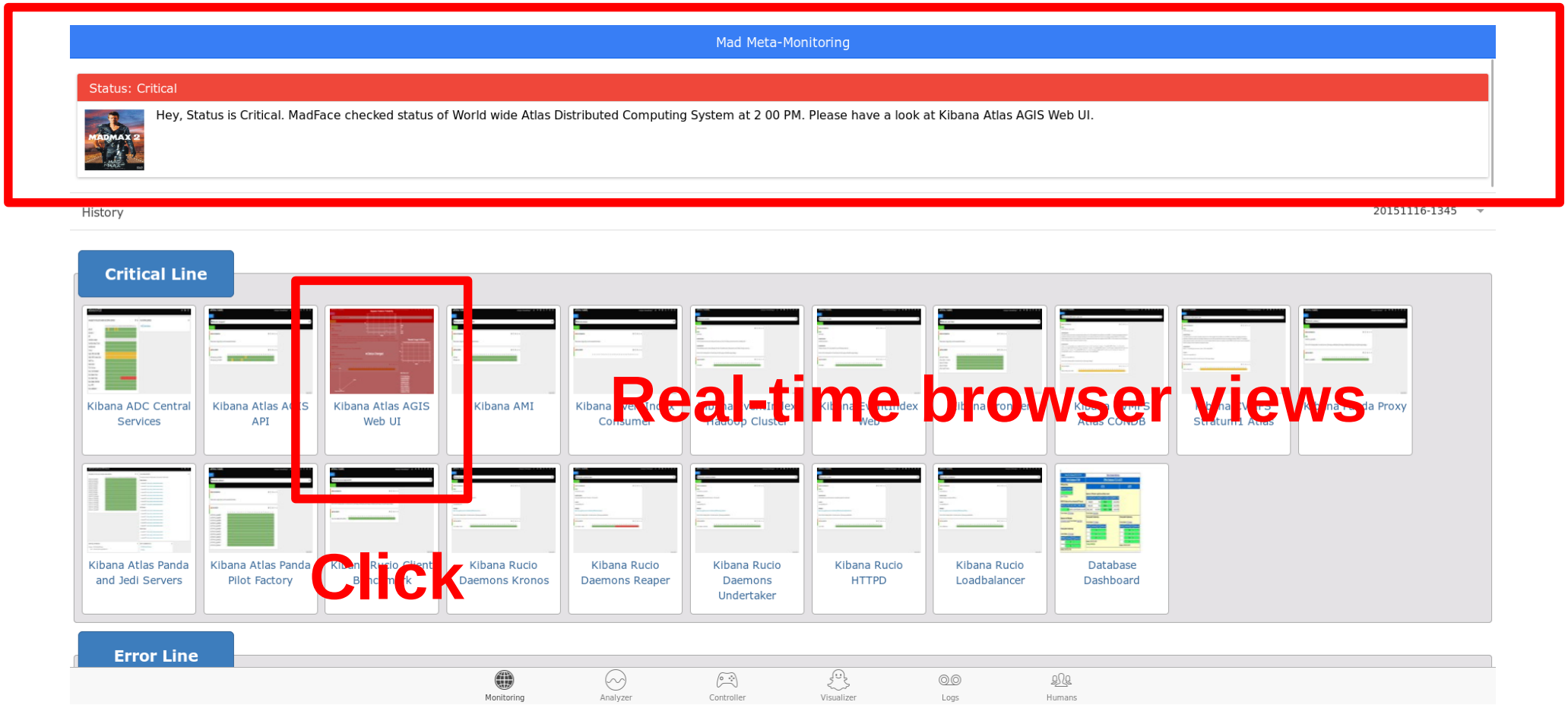
ベイズ事後確率



What does it look like now?

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

Human-readable status summary

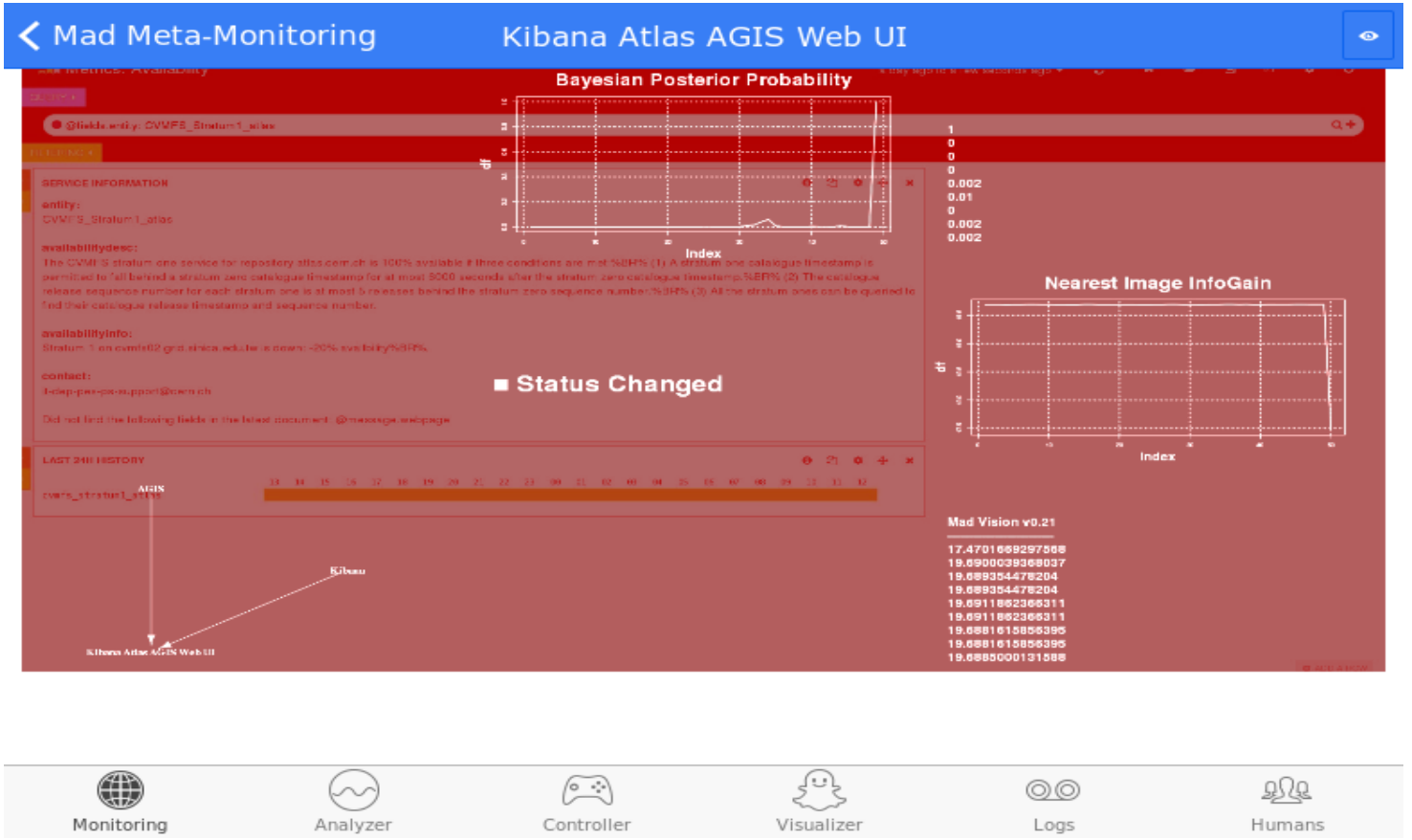


Real-time browser views

Click

What does it look like now?

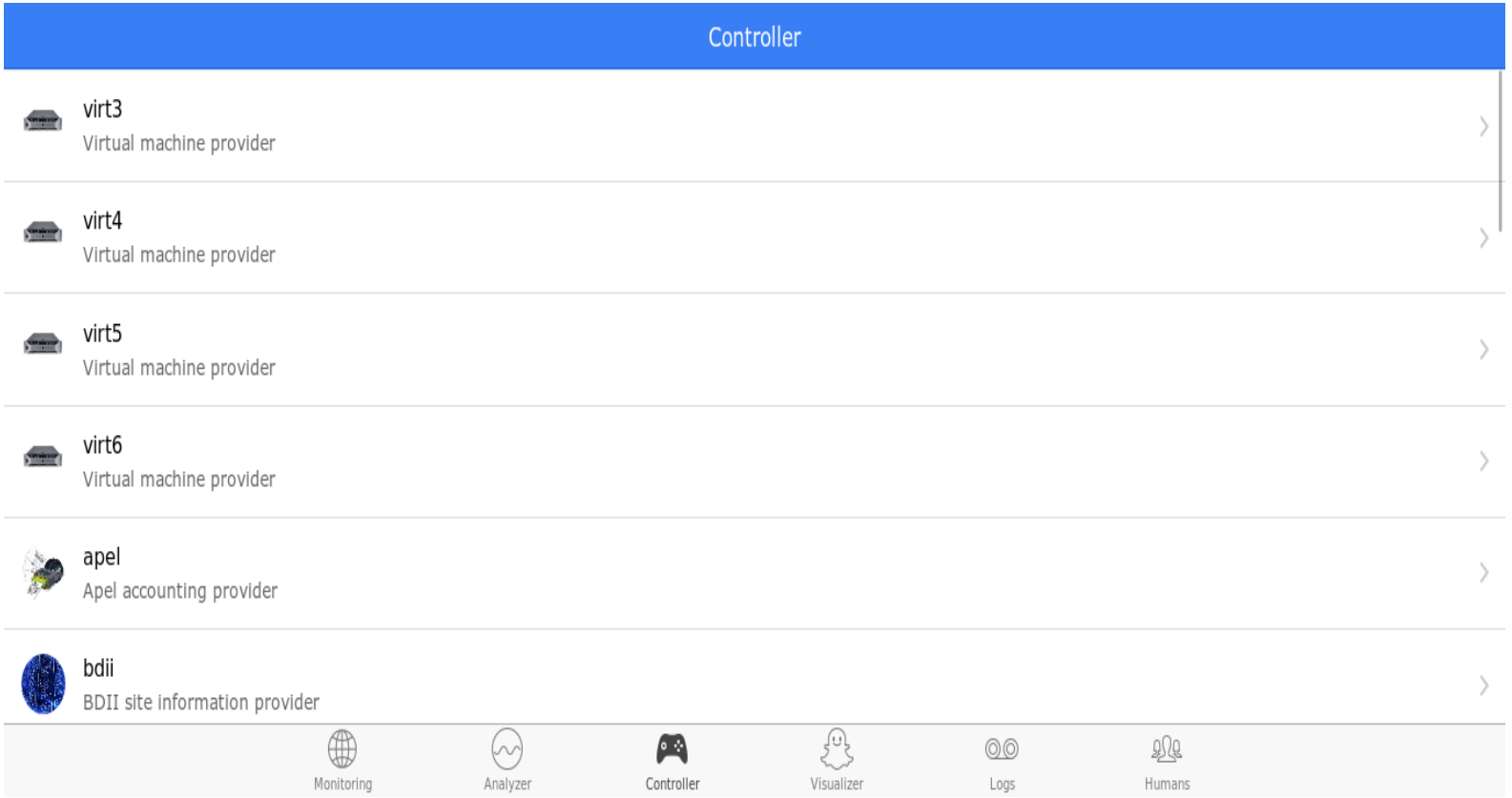
Mad Vision



What does it look like now?

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

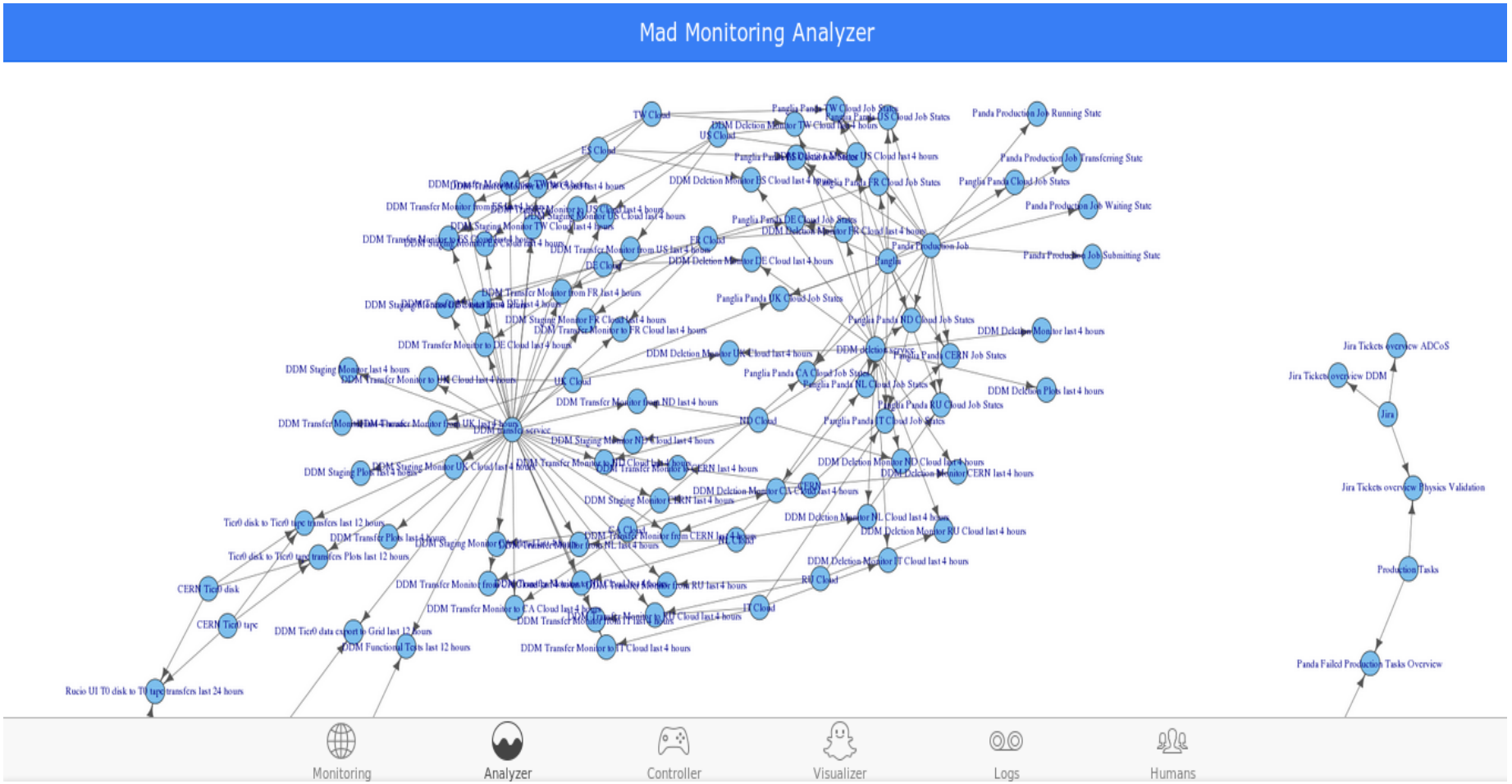
1-click controller



What does it look like now?

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

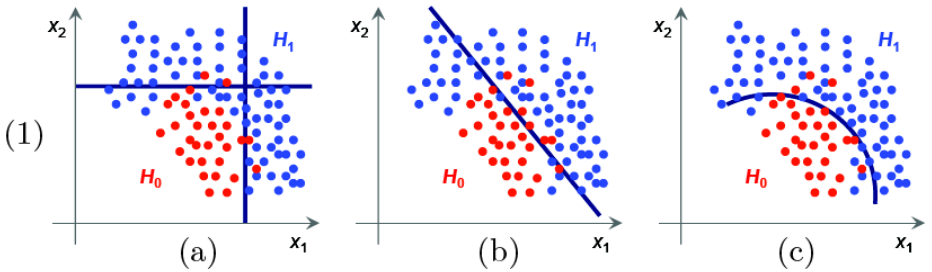
System analyzer skeleton

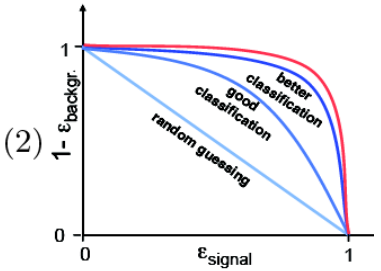


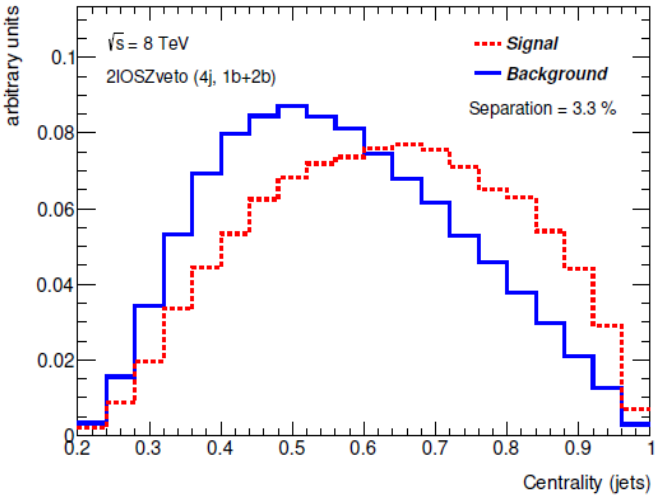
一般化されたブラウザ画面識別手法

- 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
黄金獅子旗の下に
原作・田中芳樹
総監督・石黒昇

Fragen?
質問？

おまけ

Bier (Krombacher Radler)

Blutwurst