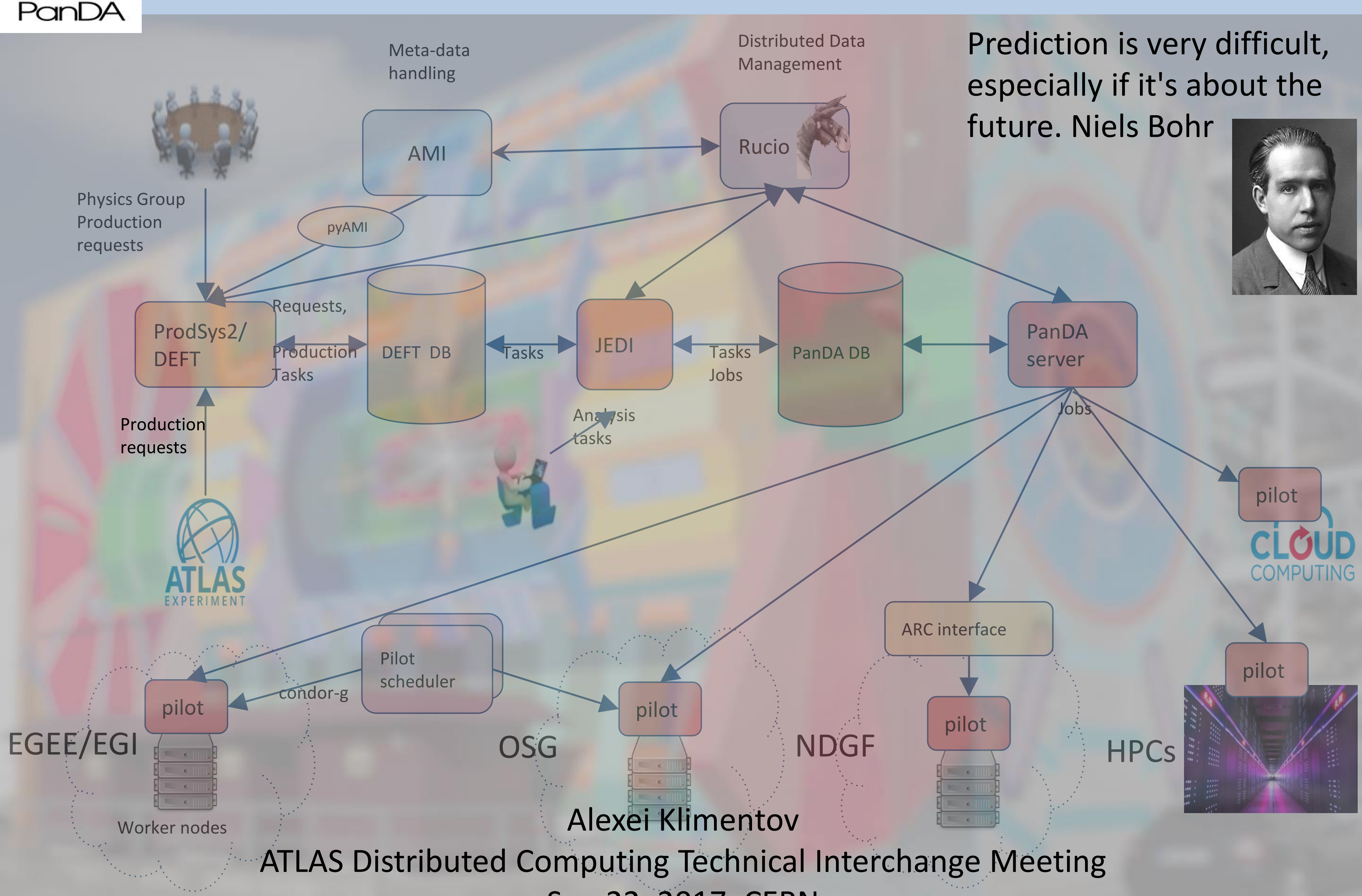
Workflow Management Software

Prediction is very difficult, especially if it's about the future. Niels Bohr



Alexei Klimentov

ATLAS Distributed Computing Technical Interchange Meeting

Sep 22, 2017, CERN

How far back can you remember ?

1998 – Can you remember that far back?

- Clinton was President and the Lewinsky scandal broke in January
- Amazon had been selling books online for only three years
- The Google name was not registered until September
- Intel introduced the Pentium II on 250 nm technology, with 7.5M transistors, and up to 450 MHz clock *(2008 – Core 2 – 45 nm, 410M transistors, 3.2GHz)*
- Top of the TOP 500 list was the ASCI Red at Sandia with 9,152 processors delivering 1.3TFLOPS

2017 :

  - *WLCG : 220,000 x86 compute cores*
  - *Titan : 300,000 x86 compute cores and 18,000 GPU cores*

- The European Research Network backbone was upgraded from 34 to 155 Mbps in December
- 64 kbps was an *excellent* home network connection
- WiFi prototypes were just appearing (IEEE 802.11-1997)
- and GPRS data services on GSM phones were yet to be launched

The landscape has changed

Have we updated the maps?

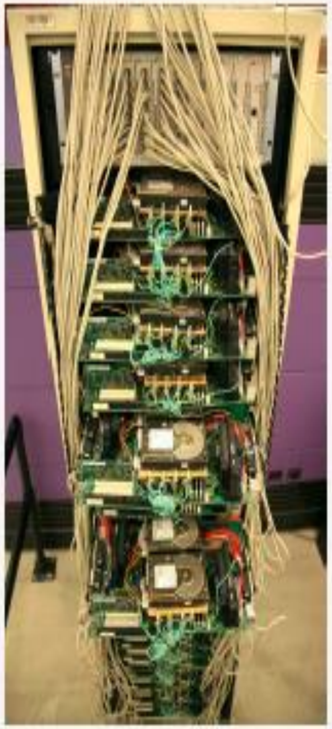
Three issues:

- **Energy**
- **Virtualisation**
- **Clouds**
- **Networks and mobility**

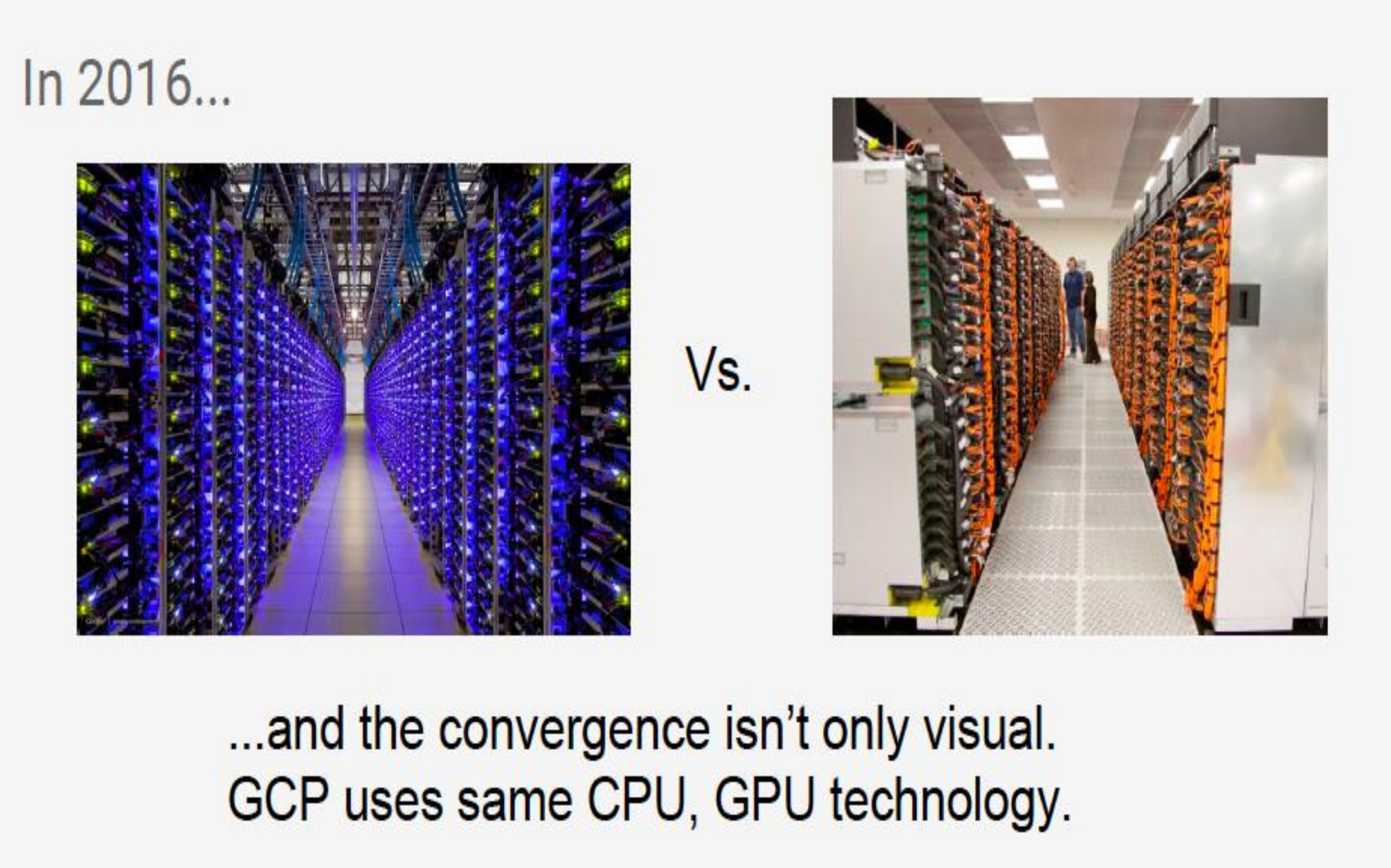
Les Robertson. CHEP plenary talk. Prague 2009

In 1999...

Vs.



Not Difficult to Spot the Supercomputer



In 2016...

Vs.

...and the convergence isn't only visual.
GCP uses same CPU, GPU technology.

Options for Future Computing

The ultimate question

- How will data be processed and analyzed in 7-10 years and beyond ?

◆ Buy facilities

- ✓ Pro : Own it! No impediment to running at full capacity when needed
- ✓ Con : Must invest for peak utilization, even if not used

◆ Use services from other providers :

- ✓ Pro : Others make capital investments
- ✓ Con : Will usage be available/affordable when needed ?

We worked hard during last years to provide examples of infrastructure not owned by ATLAS and to integrate HPC with HTC

◆ Hybrid model

- ❑ Own baseline resources that will be used at full capacity
- ❑ Use service providers for peak cycles when needed

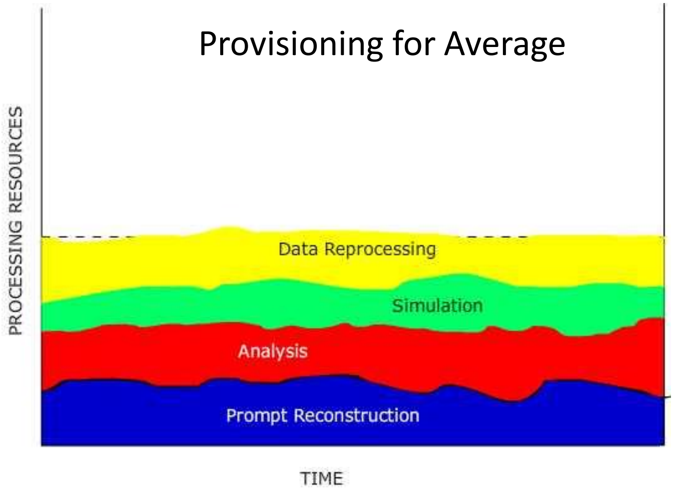
Impact on Workflow Management

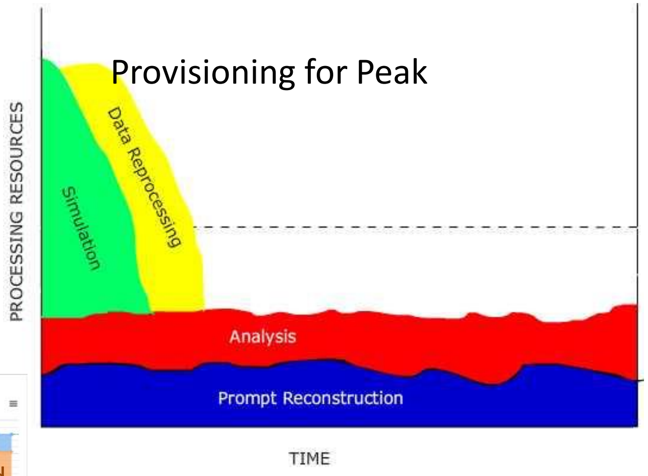
One of the biggest improvements in joining to a much larger pool of resources is breaking the idea we need to lay out our resources for average load

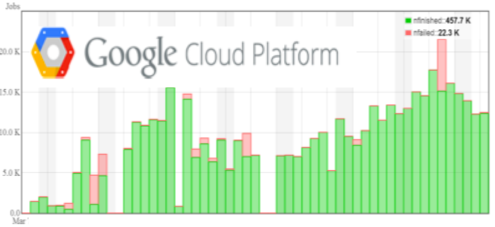
Workflows could be completed as they are defined and not over months

In these processing models the workflow system needs to be able to scale to 5-10 times the average load (~10-20M jobs/day)

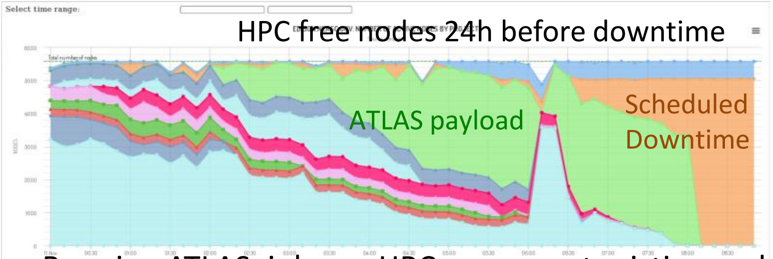
- We want to be able to burst to high values
- The least expense time to be delivered resources might be all at the same







Running ATLAS jobs on GCC



Running ATLAS jobs on HPC on opportunistic mode

Impact on Workflow Management. Cont’d

WFMS is a complex system whose behavior is intrinsically difficult to model due to the dependencies, relationships, and interactions between their parts or between WFMS and Rucio, AMI, core SW, etc, etc

- Need to focus on future automation
  - new operational scenario
    - No heroic efforts
    - Less meetings
- New approach in monitoring/analytics/ Users I/F. We should study our system.
  - Metrics
  - Anomalies
  - Inconsistencies
  - Tailes, etc, etc
  - Database schemas and performance

  No need to wait, R&Ds can be started now
- New approach how requests should be handled
- Better planning and prediction tools
- Review all distributed SW components and rid of redundancy
- Review nomenclature
  - Production steps, formats, workflow

Impact on Workflow Management. Cont’d

If one is using commercially provided computing faults turn into real money

- Need to focus on potentially wasteful things
  - Infinite loops
  - Giant log output that trigger data export charges
  - CPU efficiency loss

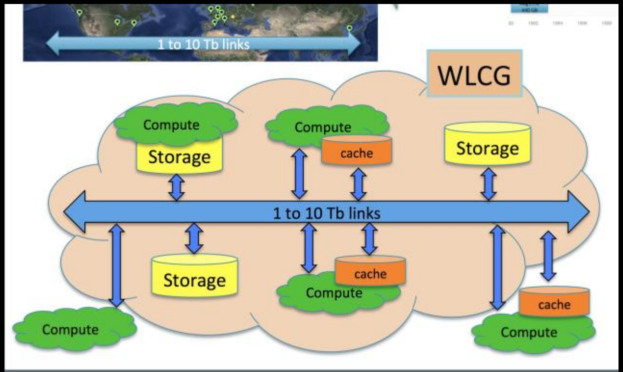
The same is true for HPC allocations

All things we probably should have been worrying about with our dedicated systems, but somehow when you are directly paying for the resources you are a bit more careful

Data Management Evolution

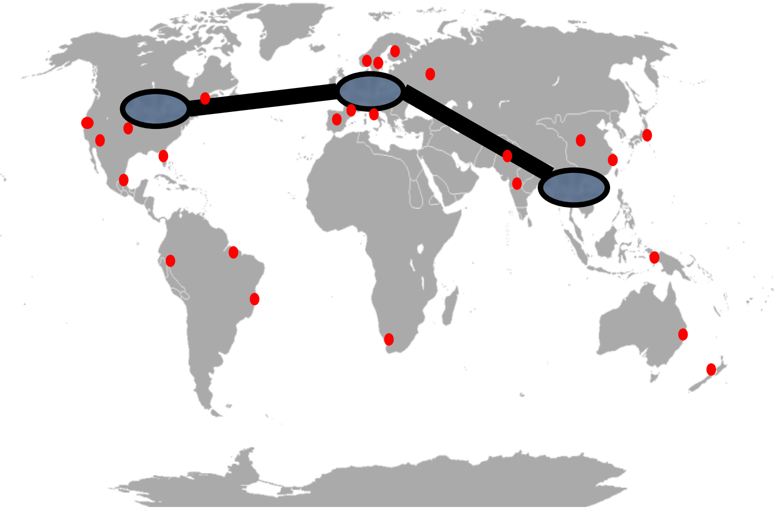
- Storage and Compute loosely coupled but connected through fast network
  - Heterogeneous computing facilities in and outside the cloud
  - Different centers with different capabilities, for different use cases
- We want to keep control of data
  - Need to be able to deploy data to a diverse set of resources
    - Clouds, dedicated sites, HPC centers, etc
  - Will need to a combination of real time delivery and advanced data caching

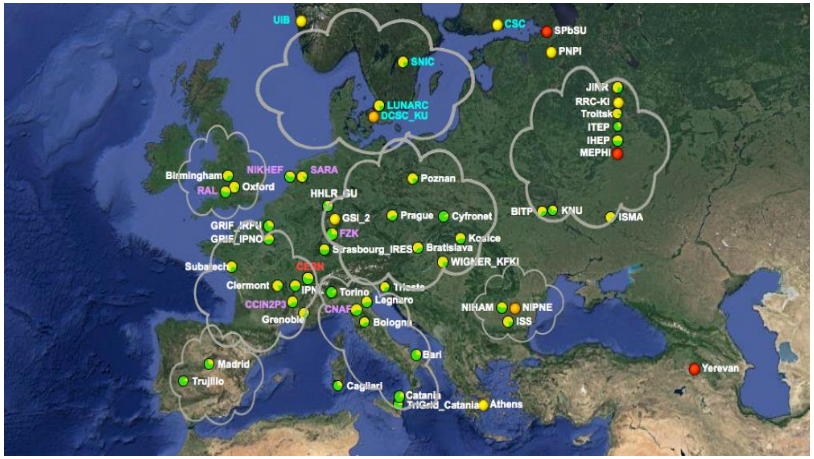
In order to replicate samples of hundreds TB in hours we will need the systems optimized end-to-end and a very high capacity network in between.



Data Management Evolution

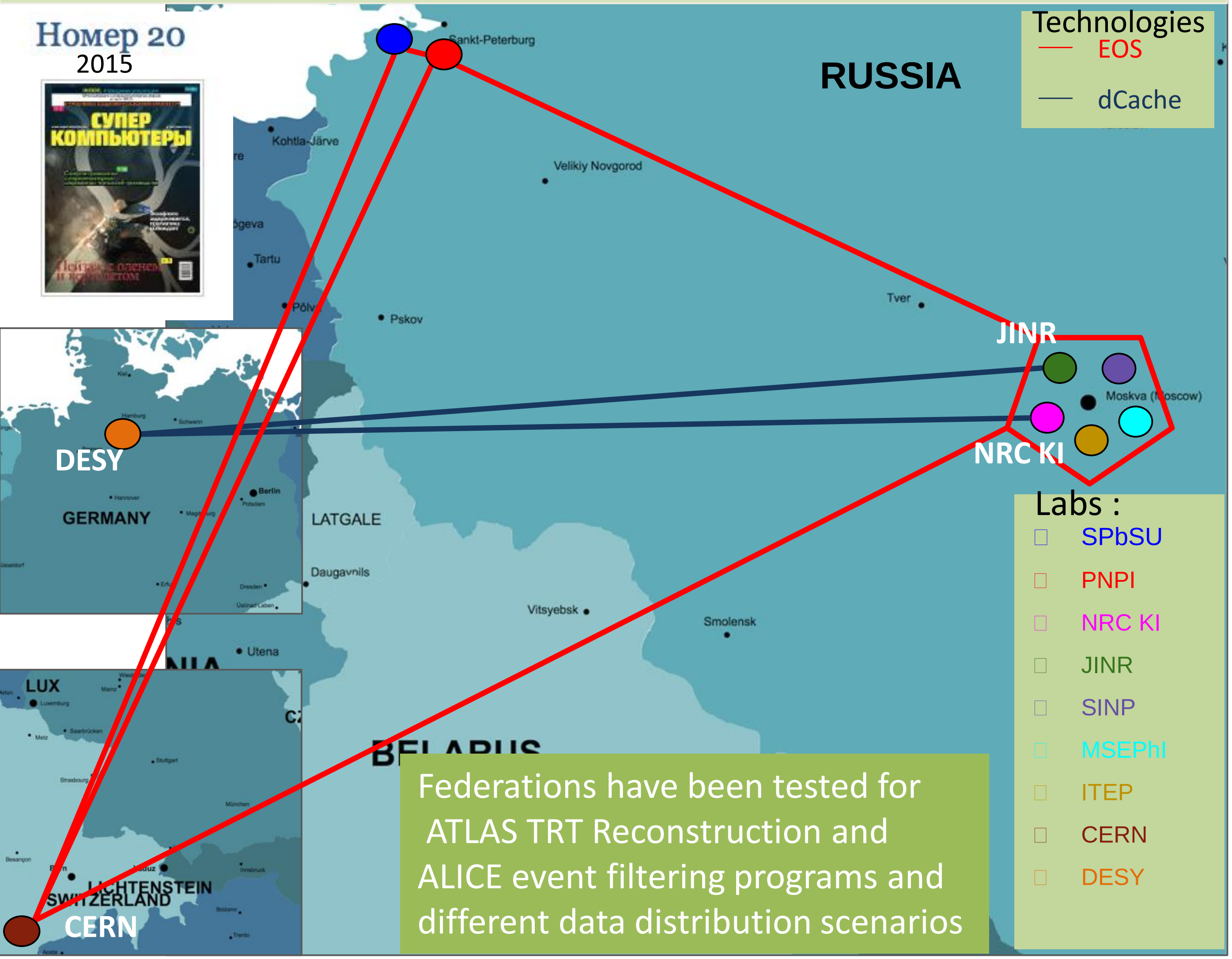
- Big centers for data reduction impacts workflow and data management
- Data selection workflow sits on top of “big data” tools
  - Focusing effort on reproducibility and shared selection criteria
- Data Management involves moving small samples to end sites
- Activity is triggered automatically
  - Needs throttling mechanisms
- The bulk of the data is placed at big sites
  - Reduced samples are moved and replicated
- Still a push to enable the processing on a variety of resources
  - Ability to burst to high capacity becomes even more important when access can trigger processing





Russian Fund for Basic Research Award. Federated Storage

CERN, DESY, NRC KI, JINR, PNPI, SPbSU, MSEPhI, ….; ATLAS, ALICE (LHC), NICA (JINR)
EOS technology : NRC-KI, JINR, T2 (ATLAS, PNPI, Gatchina), T2 (ALICE, SPbSU, Petergof), CERN
dCache technology : NRC-KI, JINR, DESY
P.Fuhrmann, I.Kadochnikov, A.Kiryanov, A.Klimentov, D.Krasnopevtsev, A.Kryukov, M.Lamanna, A.Peters, A.Petrosyan, E.Ryabinkin S.Smirnov, A.Zarochentsev, D.Duelmann



Technologies
— EOS
— dCache

Labs :
□ SPbSU
□ PNPI
□ NRC KI
□ JINR
□ SINP
□ MSEPhI
□ ITEP
□ CERN
□ DESY

Federations have been tested for ATLAS TRT Reconstruction and ALICE event filtering programs and different data distribution scenarios

R&D Project Motivation

Computing models for the Run3 and HL-LHC era anticipate a growth of storage needs.

The reliable operation of large scale data facilities need a clear economy of scale.

A distributed heterogeneous system of independent storage systems is difficult to be used efficiently by user communities and couples the application level software stacks with the provisioning technology at sites.

- Federating the data centers provides a logical homogeneous and consistent reliable resource for the end users

Small institutions have no enough people to support fully-fledged software stack.

- In our project we try to analyze how to set up distributed storage in one region and how it can be used from Grid sites, from HPC, academic and commercial clouds, etc.

Thanks

- This talk drew on presentations, discussions, comments, input from many
- Thanks to all, including those I’ve missed
  - *F.Barreiro, P.Buncic, S.Campana, K.De, I.Fisk, M.Grigorieva, A.Kiryanov, S.Panitkin, A.Patwa, L.Robertson, V.Tsulaia, T.Wenaus, A.Zarochentsev …*