Technical Interchange Meeting :

Exploitation of opportunistic resources. Commercial clouds.

Alexei Klimentov
Brookhaven National Laboratory
October 1, 2018

Disclaimer

- This talk was prepared by me in my personal capacities. The opinions expressed here are my own, and do not necessarily reflect view, policy or position of anybody else.

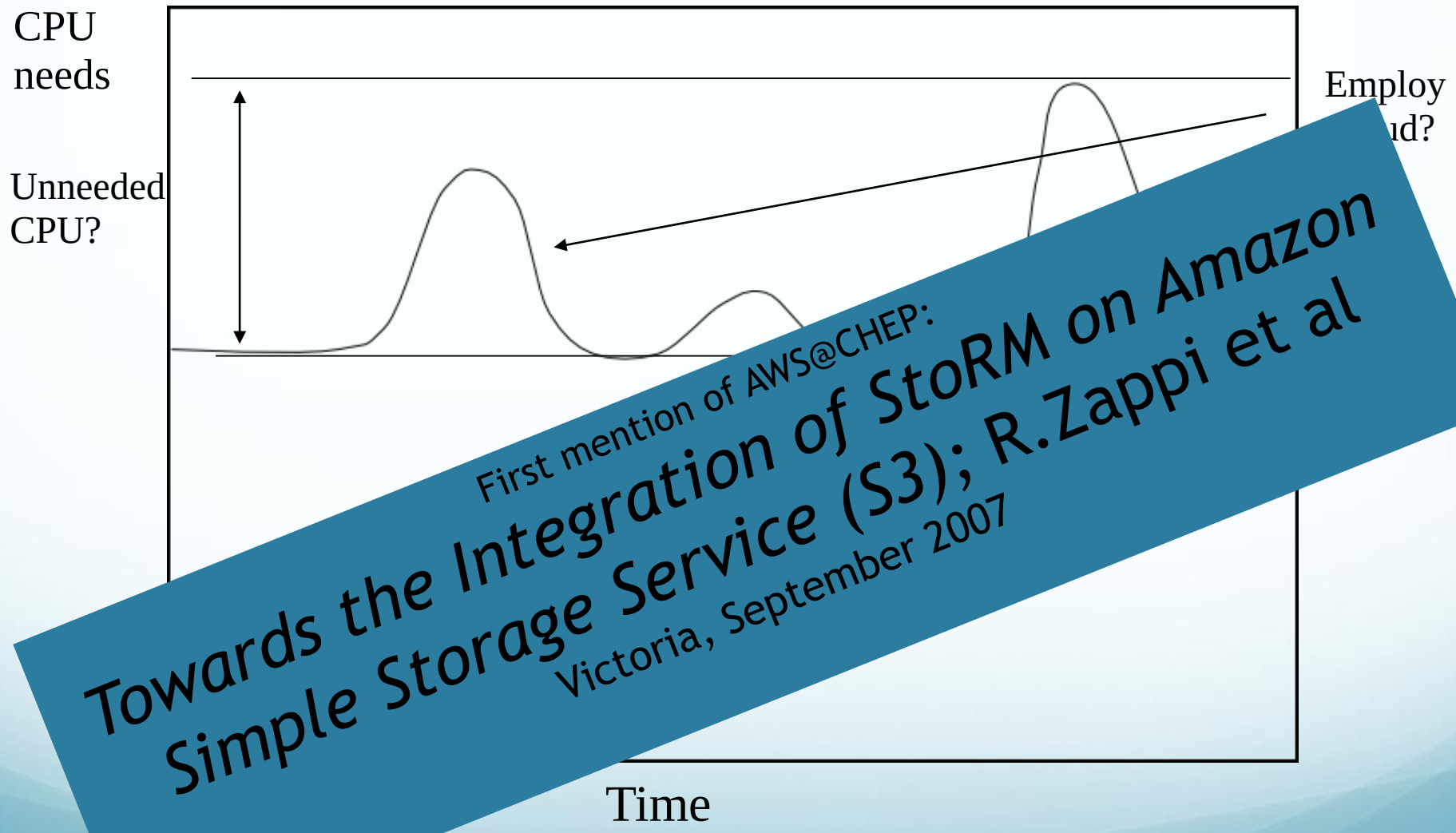
Belle Monte-Carlo production on the Amazon EC2 cloud

Martin Sevior, Tom Fifield (University of Melbourne)

Nobuhiko Katayama (KEK)

17th International Conference on Computing in High Energy and Nuclear Physics
21 - 27 March 2009 Prague, Czech Republic

Particularly useful for Peak Demand



First mention of AWS@CHEP:

Towards the Integration of StoRM on Amazon Simple Storage Service (S3); R.Zappi et al

Victoria, September 2007

BigPanDA Project. Extending the scope. Cloud Computing.

2012/13.
Google Compute Engine preview project
The first ATLAS experience with Google

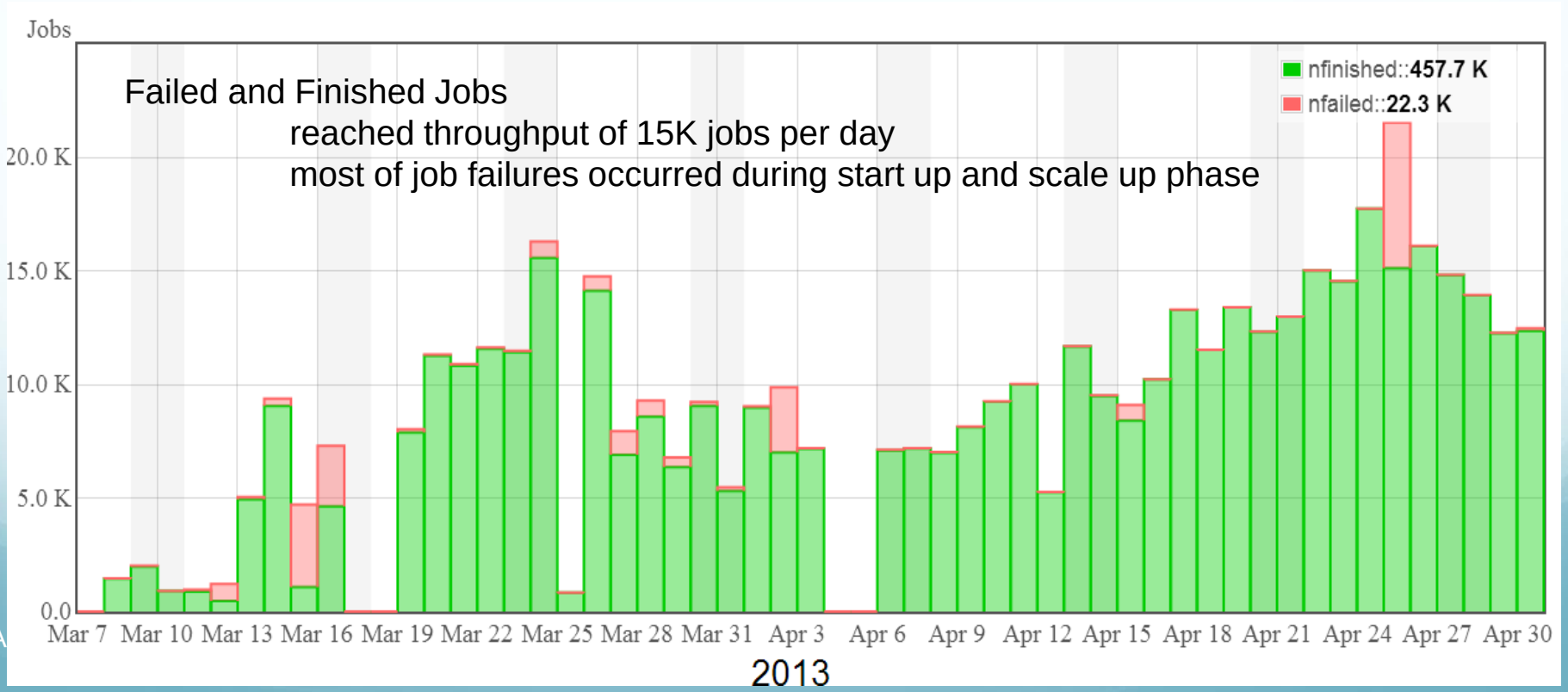
- Google Compute Engine (GCE) preview project
  - Google allocated additional resources for ATLAS for free
    - ~5M cpu hours, 4000 cores for about 2 month, (original preview allocation 1k co
- Resources are organized as HTCondor based PanDA queue
  - Centos 6 based custom built images, with SL5 compatibility libraries to run ATLAS so
  - Condor head node, proxies are at BNL
  - Output exported to BNL SE
- Work on capturing the GCE setup in Puppet
- Transparent inclusion of cloud resources into ATLAS Grid
- The idea was to test long term stability while running a cloud cluster similar in size to Tier 2 site in ATLAS
- Intended for CPU intensive Monte-Carlo simulation workloads
- Planned as a production type of run. Delivered to ATLAS as a resource and not as an R&D platform.
- We also tested high performance PROOF based analysis cluster

- PanDA based data processing and workload management
  - Centrally managed queues in the cloud
    - Elastically expand resources transparently to users
  - Institute managed Tier 3 analysis clusters
    - Hosted locally or (more efficiently) at shared facility, T1 or T2
  - Personal analysis queues
    - User managed, low complexity (almost transparent), transient

- Data storage
  - Transient caching to accelerate cloud processing
  - Object storage and archiving in the cloud

Running PanDA on Google Compute Engine

- We ran for about 8 weeks (2 weeks were planned for scaling up)
- Very stable running on the Cloud side. GCE was rock solid.
- Most problems that we had were on the ATLAS side.
- We ran computationally intensive jobs
  - Physics event generators, Fast detector simulation, Full detector simulation
- Completed 458,000 jobs, generated and processed about 214 M events
- Invited talk at Google IO conference



Commercial Clouds. Early days

- ATLAS was invited to participate in the preview of Google Compute Engine in 2012
- The Relativistic Heavy Ion Collider and ATLAS computing facilities at Brookhaven National Laboratory received a grant allocation from Amazon Elastic Compute Cloud (EC2) in 2013. The defining feature of the project was the utilization of Amazon's EC2 Spot market resources. Spot pricing, which trades a willingness to accept abrupt virtual-machine terminations for a significantly reduced cost, has turned commercial public clouds into an economical option, complementary to site-based dedicated resources. The Spot market thus offers an interesting alternative to otherwise expensive commercial clouds.
  - A hybrid cloud set-up includes resources from a large-size ATLAS computing centre at Brookhaven, and the elastic part of Amazon's cloud, spanning their geographically distributed sites. The system can react to the growing demand for computing resources by expanding a number of virtual machines running on Amazon's Spot market — it has proven capable of scaling up to 20,000 simultaneous jobs at peak operation.
    - *S.Panitkin "Look to the clouds and beyond", Nature Physics 11, 373-374 (2015)*
- LHCb-Yandex collaboration (~2014)
- CMS / FNAL HEPCloud project with Google (~2016)
  - Challenge : can we double CMS computing ?
    - Live demo during SuperComputing 2016 conference
    - Expand FNAL facilities for an additional 160,000 core
    - K.Kissell (Google) talk @SMC2017

Looking Forward Looking Back

Seymour Cray :
“supercomputer, it is hard to define, but you know it when you see it”

In 1999...

Vs.

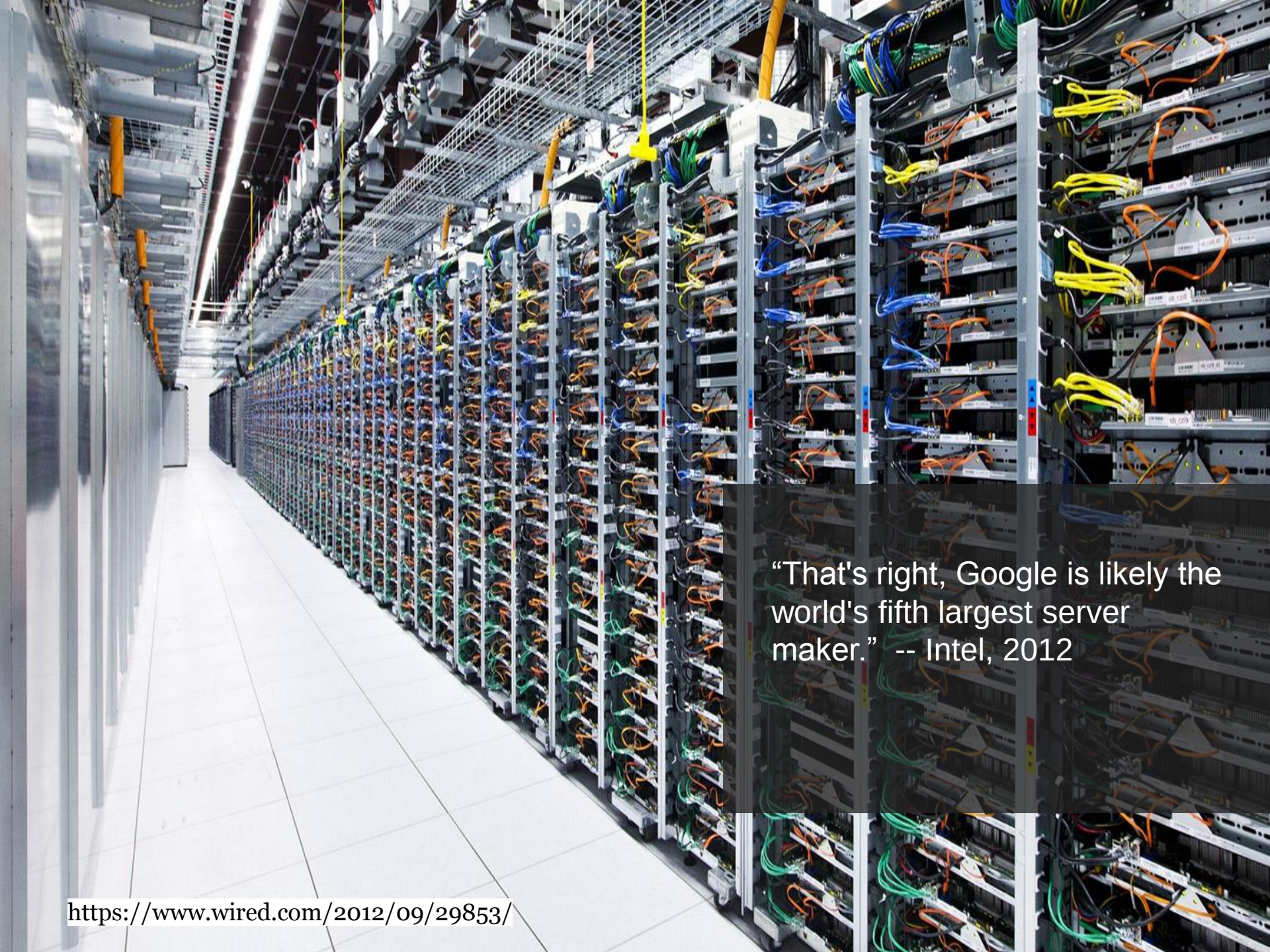
Not Difficult to Spot the Supercomputer

In 2016...

Vs.

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

K.Kissel, Google, SMC2017 talk


“That's right, Google is likely the world's fifth largest server maker.” -- Intel, 2012
https://www.wired.com/2012/09/29853/

Looking Forward, Looking Back

Body of Existing SMP, MPI Codes Must Be Supported

But New Models Will Be Required for the Future

Cloud HPC, Exascale Share Many Problems

Reason to Hope for Common Solutions

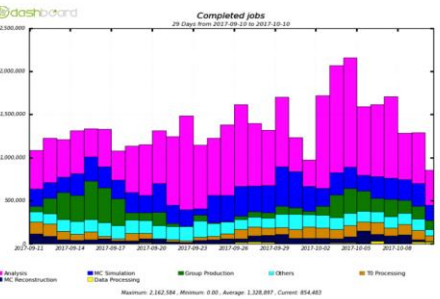
2018. ATLAS / Google “Data Ocean” R&D project (PoC phase)

First contact at Smoky Mountains Conference, brainstorming discussions in Nov – Dec 2017, F2F meetings at SC17, at BNL, Google NYC and at CERN to define project goals and deliverables.

Status and results were presented to ATLAS community, at CHEP and at Google NEXT conferences. Three main ideas :

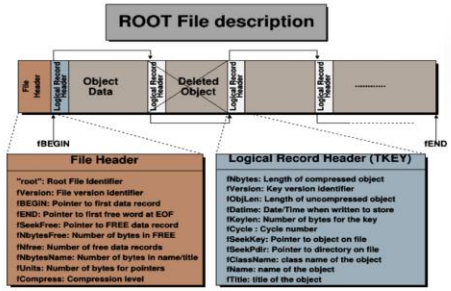
- User analysis
  - Place copies of analysis output on GCP for reliable user access
  - Serves as cache with limited lifetime
- Data placement, replication, and popularity
  - Store the final derivation of MC and reprocessing data campaigns
  - Use Google Network to make data available globally (e.g., ingest in Europe but job reads from US)
  - Incorporate cloud access patterns into popularity measurements
- Data marshaling and streaming
  - Evaluate necessary compute for generation of sub-file products (branches/events from ROOT files)
  - Job performance and network behavior for very small sample streaming

User Analysis



Data Analysis, Replication and Placement

Data Streaming



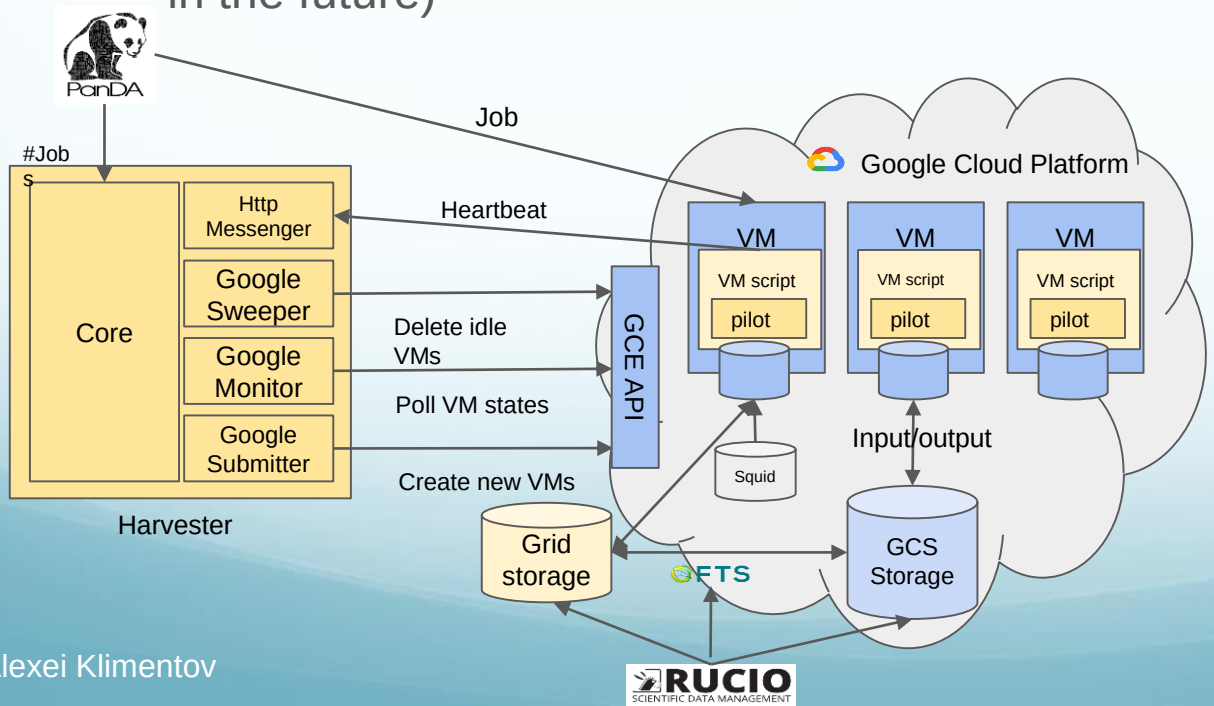
Google Cloud

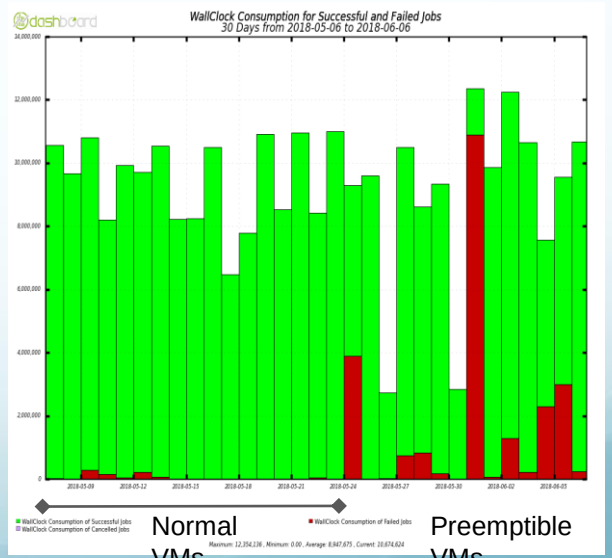
PoC Project Status. Data Management

- Developed an interface to cloud storage in Rucio compatible with regular non-cloud storage systems. GCS would show as "just another" endpoint, without the user having to know/worry.
- Developed a native implementation of the interface in Rucio to interact with GCS, using cryptographically secure authentication, authorisation, and transport. (The protocol-open implementation based on S3 that was developed can also be used, however it won't get the full throughput)
- Transparent inclusion of GCS for all orchestrated data activities (i.e., putting rules and lifetime). This was demonstrated by putting several Terabytes of data into US and EU GCS endpoints. The limit on these transfers was the available bandwidth of the NRENs.

PoC Project Status. GCE integration with PanDA

- Harvester is integrated with GCE by adding the resource facing plugins using the python API (GCE Submitter, GCE Monitor, GCE Sweeper and HTTP Messenger) and a python module to run on the worker node at startup.
- It's a 100% PanDA implementation and there are no intermediate layers, it is not rely on Condor. everything is done through contextualisation and our own modules. Contextualisation possibilities depend strongly on the OS and cloud.
- ATLAS MC simulation jobs were ran successfully in Apr-Jun. The I/O was going to CERN_DATADISK. Normal and preemptive VM modes.
  - Preemptive VM cannot be used for analysis (Rucio will support Google upload in the future)





Efficiency of preemptible VMs can be optimized through usage of Event Service.

PoC Project Status. End User Analysis

- Setup Rucio RSE on Google storage
- Analysis queue with harvester talking directly to Google API. It isn't trivial to describe it in AGIS
- Replicate DAOD from ATLAS RSE to Google RSE
  - A real use-case from A.Dubreuil. Skimmed DAOD RPVILL, 1M events, 450 GB
  - A real use-case from F.Fischer (lepton isolation), O(1TB)
- Run analysis jobs via PanDA
- Retrieve results from Google RSE with Rucio

Preliminary Company *X* 2018 Prices.

- Have to pay for:
  - CPU: n1-standard-1, 1 CPU, 3.75GB RAM, $0.0475/hour, ($0.0100/hour preemptible (not useful for analysis))
  - Storage of data: Multi-Regional Storage $0.026 per GB/month down to Coldline Storage $0.007 per GB/month
  - Network: Ingress is free, Egress $0.12 per GB/month
  - Access operations: 1M class A ops $5-10 per month
- Storing 1 PB of data: $26000 per month
- Storing/processing 50 TB with 5 TB out in a month (a day):
  - Total: $4000 ($143)
  - includes: Storage: $1000 ($43), 100 VMs with 73k hours $2400, 5 TB network out $600 ($20)

J.Elmsheuser, July 2018

Company *Y* 2018 Prices.

Preliminary

CE:

1 core vCPU : 1c / hour

1 GB RAM: 0.3c / hour

1 GB HDD: 3c / month

SE:

1 GB hot storage: 1.8c / month

1 GB cold storage: 1c / month

1 GB input traffic: 0c

1 GB output traffic: 2c

Network:

1 GB outgoing traffic: 2c

- Storing 1PB of data/month
  - Hot storage $18000
  - Cold storage $10000
- Storing/processing 50 TB with 5 TB out in month (day)
  - Total : $2280 ($76)
  - Includes
    - Storage $900 ($30)
    - 100 VM with 73k hours $1280 ($42.5)
    - 5 TB network out : $100 ($3)

- You can always find a cheaper (in this particular case ~40%) resources provider, but even the cheapest one should be trustable.

EC2 2011 Bill.

Part of actual EC2 Bill for Panda related activities Apr. 2011

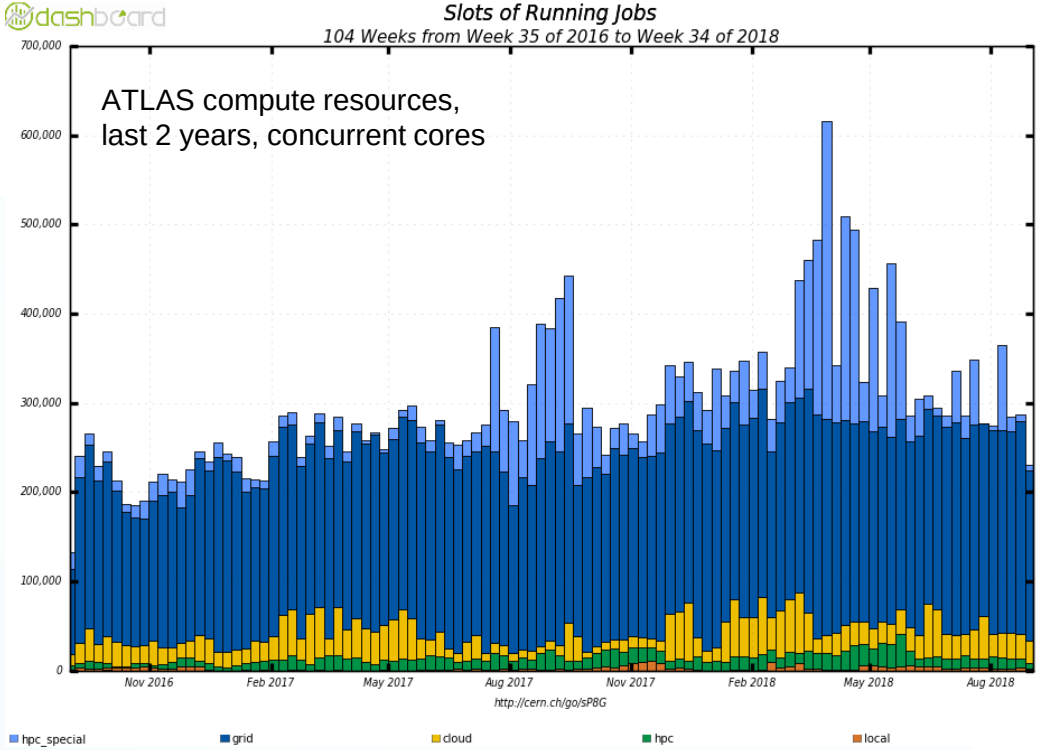
		Totals
Amazon Elastic Compute Cloud		
US East (Northern Virginia) Region		
Elastic IP Addresses		
$0.01 per non-attached Elastic IP address per complete hour	744 Hrs	7.44
»		7.44
Amazon Simple Storage Service		
US Standard Region		
$0.140 per GB - first 1 TB / month of storage used	387.825 GB-Mo	54.30
$0.01 per 1,000 PUT, COPY, POST, or LIST requests	4,759 Requests	0.05
$0.01 per 10,000 GET and all other requests	62,627 Requests	0.06
»		54.41
Amazon Simple Notification Service		
»		0.00
Amazon Virtual Private Cloud		
»		0.00
AWS Data Transfer (excluding Amazon CloudFront)		
$0.100 per GB - data transfer in per month	129.790 GB	12.98
$0.000 per GB - first 1 GB of data transferred out per month	1.000 GB	0.00
$0.150 per GB - up to 10 TB / month data transfer out	2,533.798 GB	380.07
		393.05
Bill Summary		
Usage charges and monthly recurring fees during this billing cycle (More Info)		$454.90

Panda server IP address

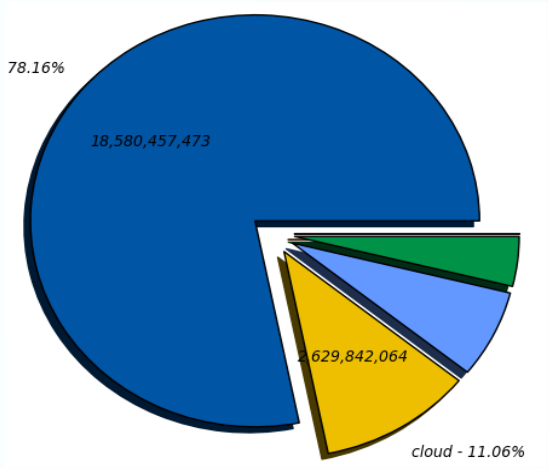
Panda monitor data

Panda monitor data transfers

S.Panitkin



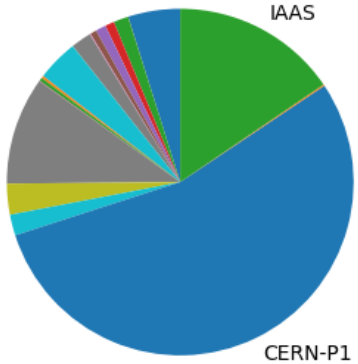
CPU HS06 shares, last year

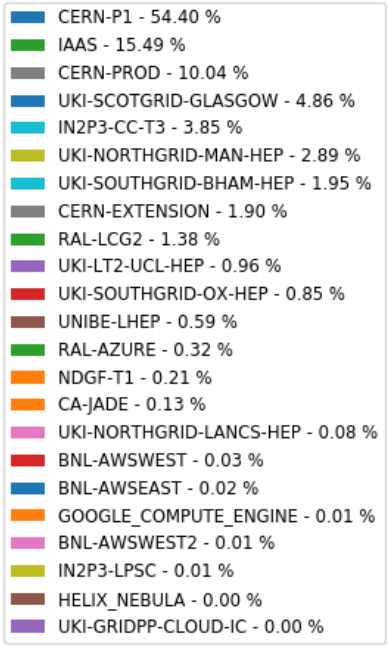


Grid: 78%
Cloud, HLT: 11%
HPC special: 7%
HPC regular: 4%

Breakdown of the cloud facilities (2016-2018)

- CERN: 64.4%
  - P1 : 54.4%
  - Prod : 10.0%
- IAAS : 15.49%





A.Klimenov

Looking Forward Looking Back

- This year conference talks
  - ISGS
  - CHEP
  - HSF-WLCG workshop /CWP
  - Grid

The European Open Science Cloud Project

- The EOSC: "**a model for the use of a cloud in the private and public sectors**" (European Parliament resolution on the European Cloud Initiative, Feb 2017)
- **Why the EOSC?**
  - To facilitate scientific developments and make the EU a center for global research
  - EOSC: "S" as in "Science", but with its user base to be extended to industry and governments
  - To foster the growth of the European Digital Economy → competiveness, global market positioning (esp. for SMEs)

Key points of EOSC

- **A data infrastructure common and a federation of existing resources**, which will:
  - Re-use existing building blocks & state-of-the-art services and solutions whenever possible
  - Develop and offer services based on user needs
- In summary: leverage past investment in research data infrastructures to **add value in terms of scale, interdisciplinarity and faster innovation, with a clear business model for sustainability.**

Conclusion : Europe has been investing a large amount of money to create EOSC

HNSciCloud
a Hybrid Cloud for Science

To **provide a common cloud platform for the European research community**

Via a collective effort of 10 procurer Research Organisations forming the **Buyers Group**

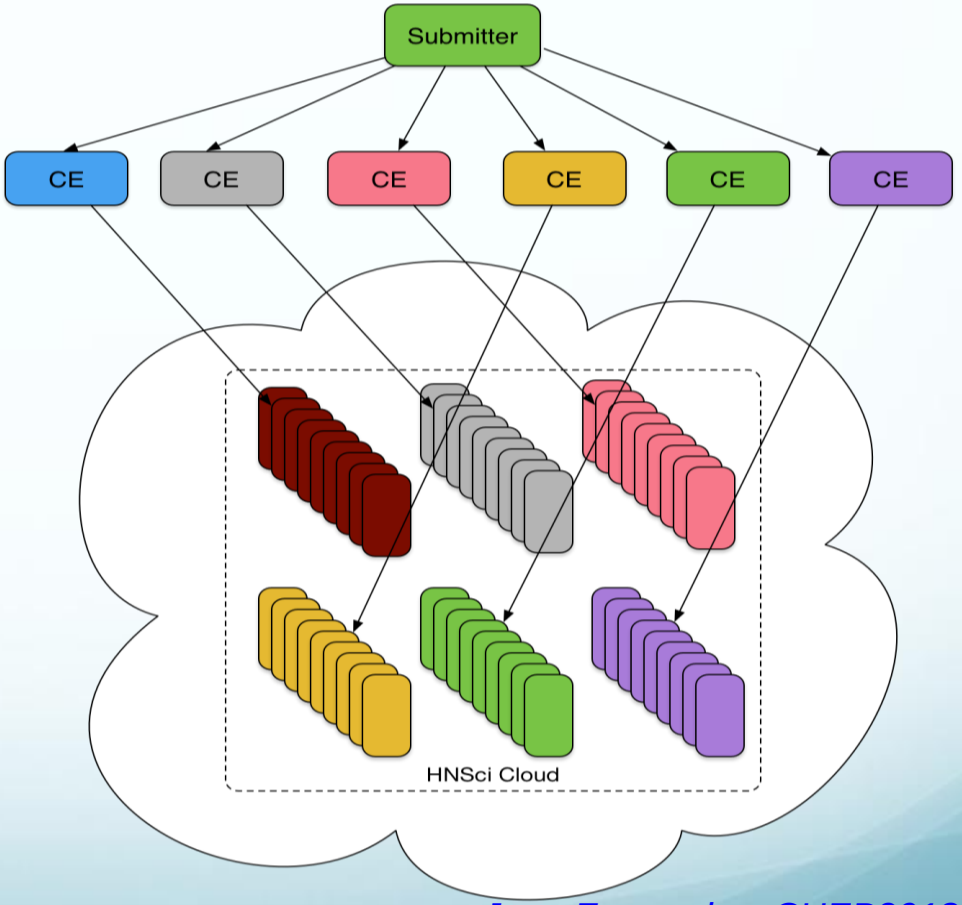
Bringing together:

- Research Organisations
- Data Providers
- Publicly funded e-infrastructures
- Commercial cloud providers

With: Procurement and Governance suitable for the dynamic cloud market

Pilot phase : till Dec 2018

Contractual Relationship: Important to understand how to integrate commercial cloud services into scientific activities



Joao Fernandes. CHEP2018

Possible R&D topic

- Integration of Cloud with ATLAS computing facilities: in addition to leveraging cloud directly through WLCG, a number of sites have expressed interest in independently using cloud for their needs, but integrating into the WLCG infrastructure. Additional R&D is needed to determine how to support third party billing and provisioning in workload management system.

Acknowledgements

- Special thanks for contributions/materials from : Alexander Alexeev, Fernando Barreiro, Tony Cass, Kaushik De, Johannes Elmsheuser, Joao Fernandes, Rob Gardner, Kevin Kissel, Siarhei Padolski, Sergey Panitkin, Davide Salomoni, Martin Sevior, Torre Wenaus