A Large Ion Collider Experiment

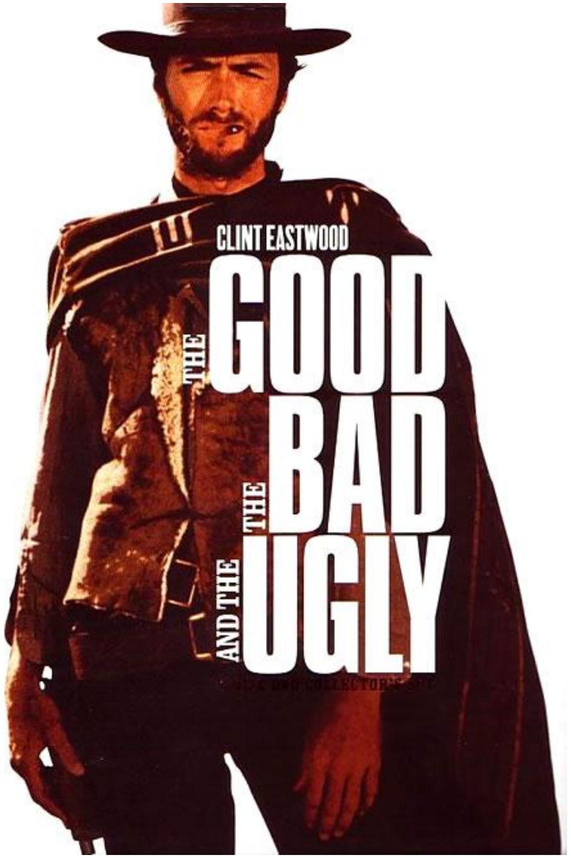
ALICE

AliEn, Clouds and Supercomputers

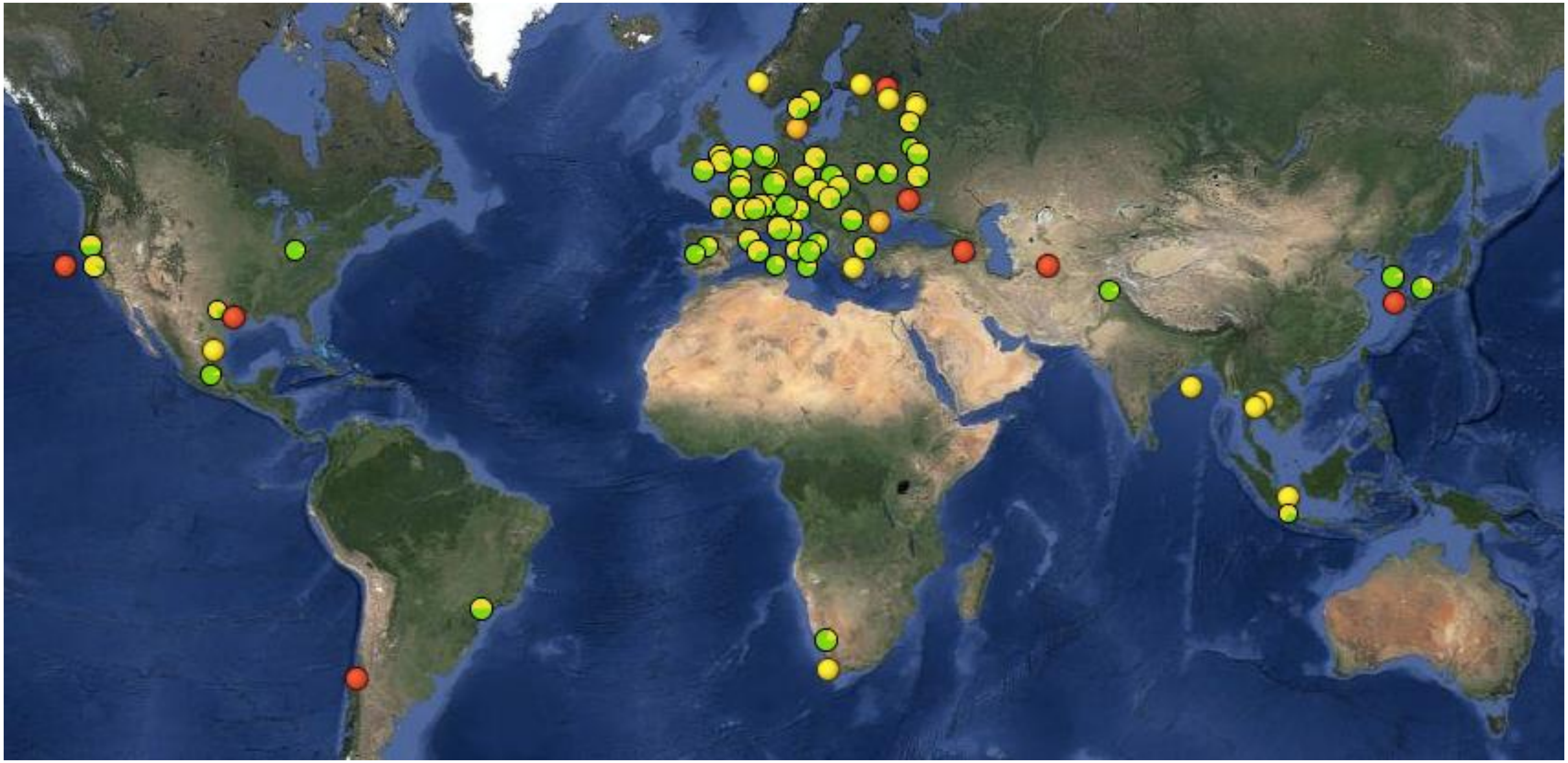
Predrag Buncic

With minor adjustments by Maarten Litmaath for WLCG Collaboration workshop, Nov 11

An alternative talk title…

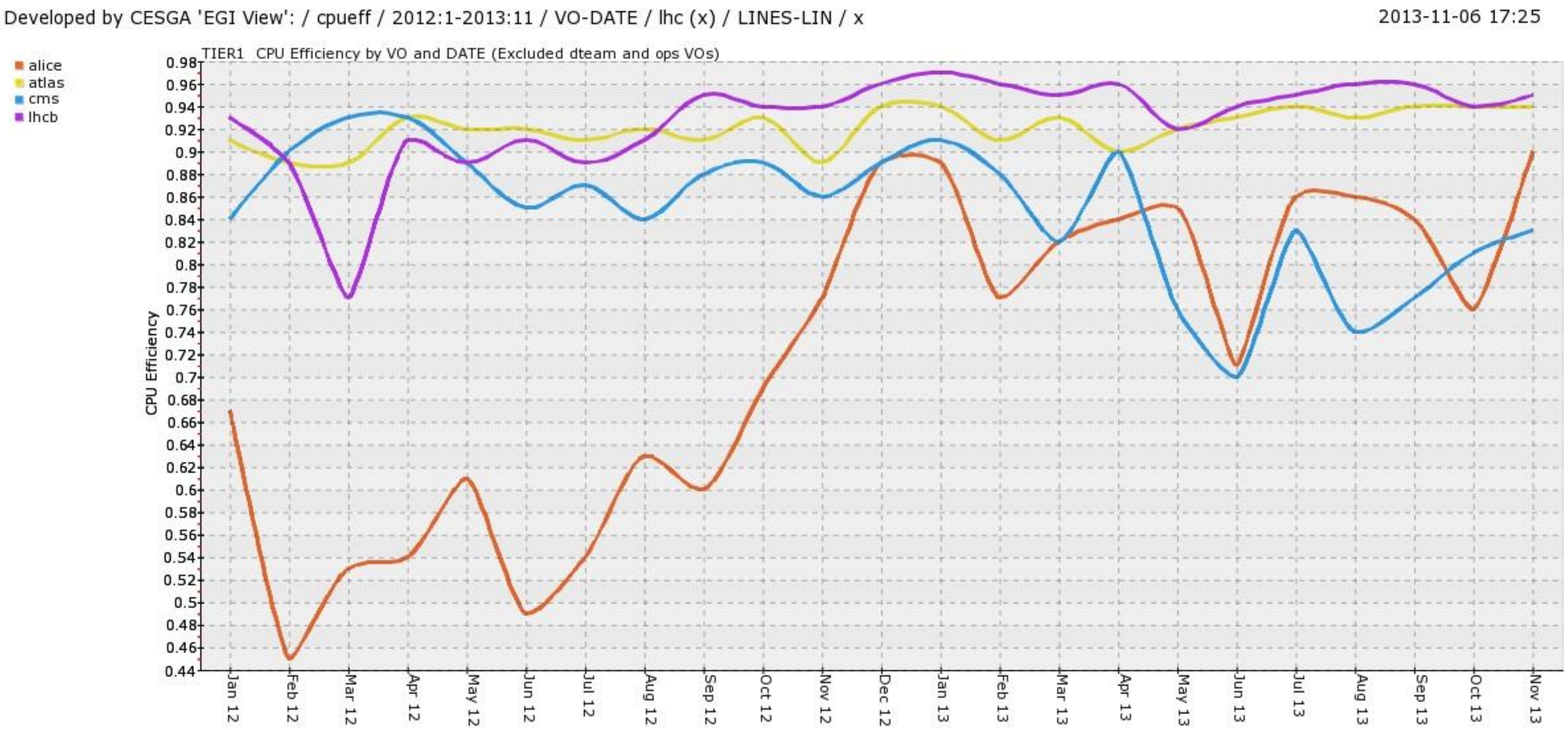


AliEn@Grid



- Highly Distributed infrastructure of ~100 computing centers, 50k cores
- Distributed data, services and operation infrastructure
- Not the easiest thing to master but we learned how to do it

Job Efficiency



- Well done, thank you! Can we still improve?!

Why changes?

- The system currently fulfills all the needs of ALICE users for reconstruction, simulation and analysis
- We know how to operate it and system is tailored to our needs and requirements (that's good)

- However, our needs will grow by x20 during Run3 (driven mostly by the simulation needs)
  - …and even bigger will be needs of other experiments after LS3

- How should the Grid look to match those needs (and still be usable)?

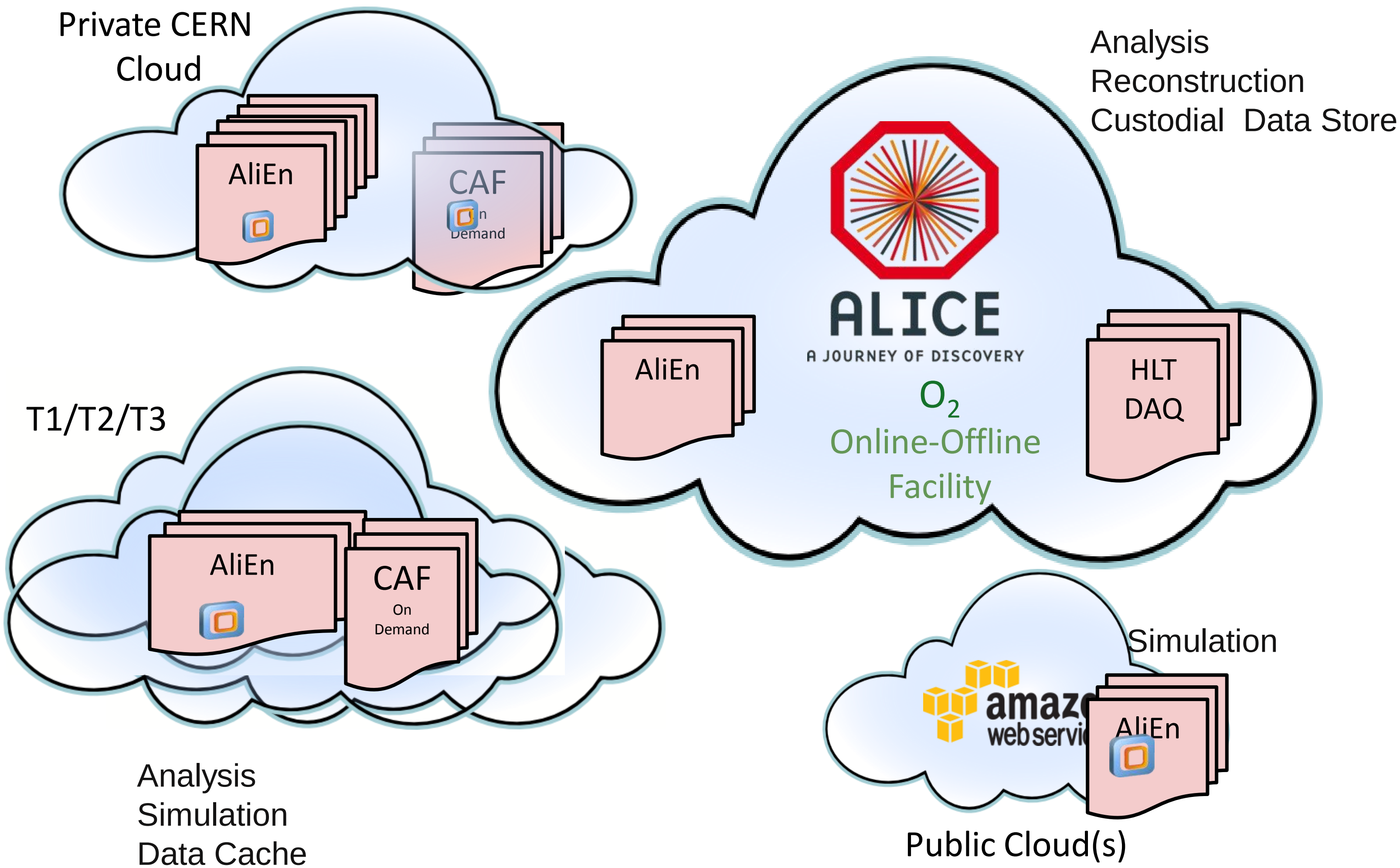
Meet the Cloud…

- At present 9 large sites/zones (up to ~2M CPU cores/site, ~4M total)
- About the size that will be needed for LHC data processing in Run3
- 500x more users, 10x more resources, 0.1x complexity
- Optimized resource usage using virtualizations, IaaS model
- Economy of scale

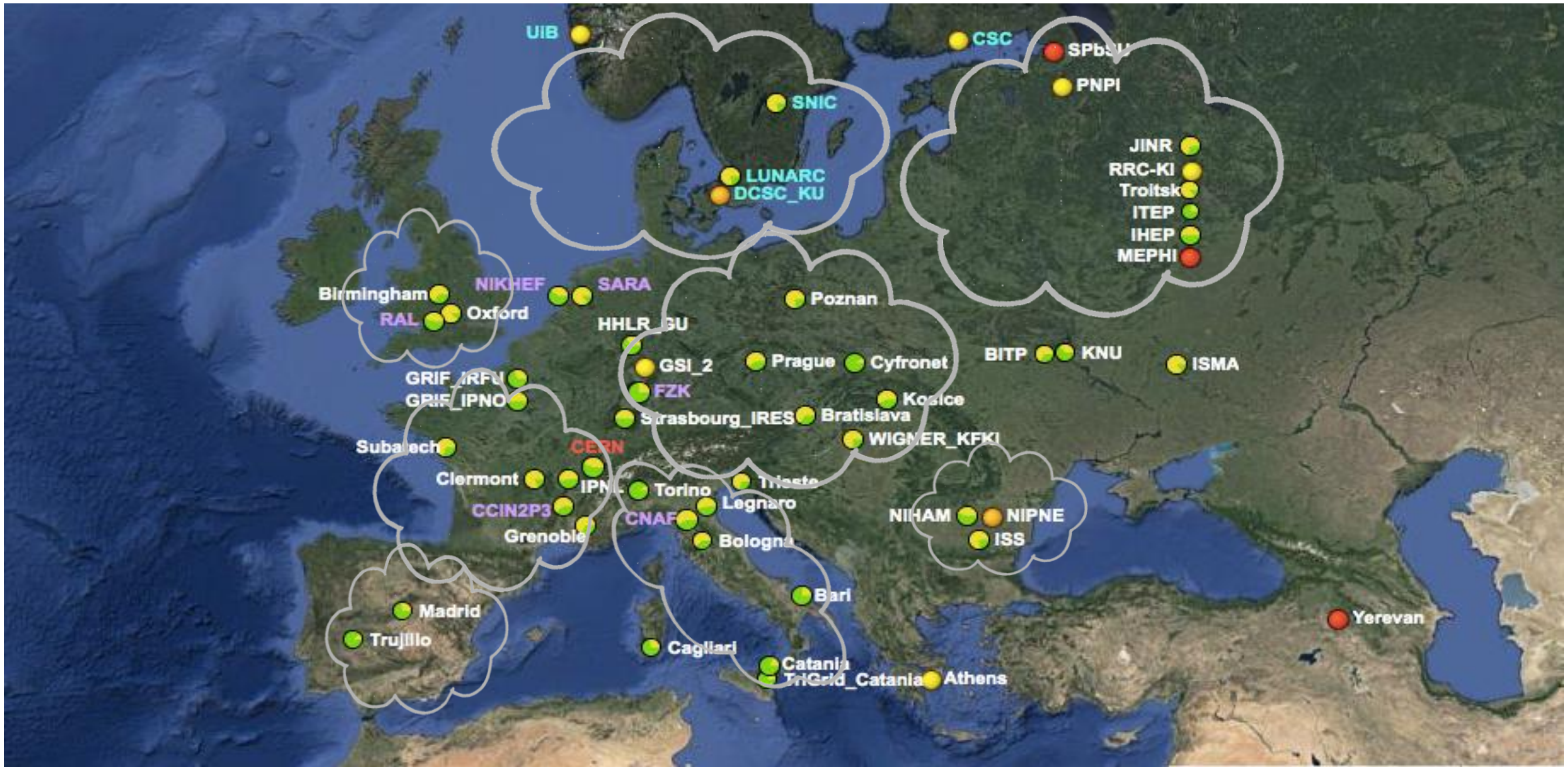
The bad part…

- The Cloud is not End-to-End system
  - It allows us to build one (or reuse one of the existing solutions suitable for given scale)
- We can re-build our Grid on top of Cloud
  - Preserving investment and user interfaces
  - Have to do the development and maintenance of our own middleware
- Data management is already a big problem and is bound to become much bigger
  - Cloud does not come with out of the box solutions for data management that would be appropriate for our problem

ALICE@Run3


Private CERN Cloud
AliEn
CAF On Demand
Analysis
Reconstruction
Custodial Data Store
ALICE
A JOURNEY OF DISCOVERY
AliEn
O_2
Online-Offline Facility
HLT
DAQ
T1/T2/T3
AliEn
CAF
On Demand
Analysis
Simulation
Data Cache
Simulation
AliEn
Public Cloud(s)

From Grid to Cloud(s)

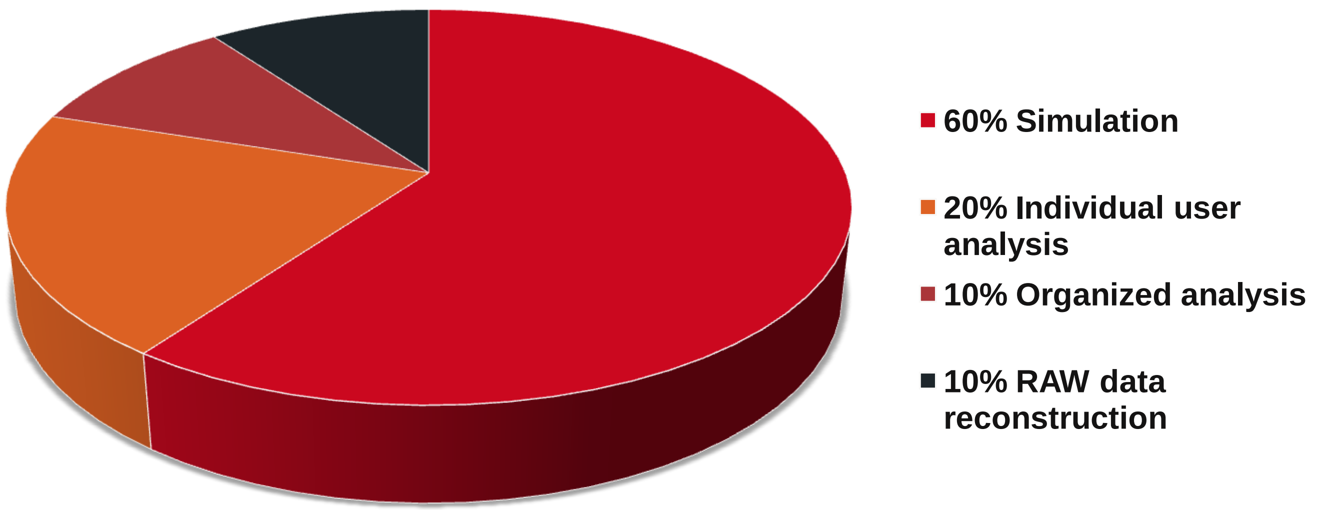


- In order to reduce complexity national or regional T1/T2 centers could transform themselves into Cloud regions
  - Providing IaaS and reliable data services with very good network between the sites, dedicated links to T0

Reducing the complexity

- This would allow us to reduce complexity
  - Deal with handful of clouds/regions instead of individual sites
- Each cloud/region would provide reliable data management and sufficient processing capability
  - What gets created in a given cloud, stays in it and gets analyzed there
- This could dramatically simplify scheduling and high level data management

Simulation



- Even with efficient use of Clouds and our own O2 facility we might be still short of resources for simulation
  - Need to be able to use all possible sources of CPU cycles

Simulation strategy

- Migrate from G3 to G4
  - G3 is not supported
  - G4 is x2 slower for ALICE use case
- Need to work with G4 experts on performance
- Expect to profit from future G4 developments
  - Multithreaded G4, G4 on GPU…
- Must work on fast (parameterized) simulation
  - Basic support exists in the current framework
- Make more use of embedding, event mixing…

In spite of all this we might fall short of resources

ALICE

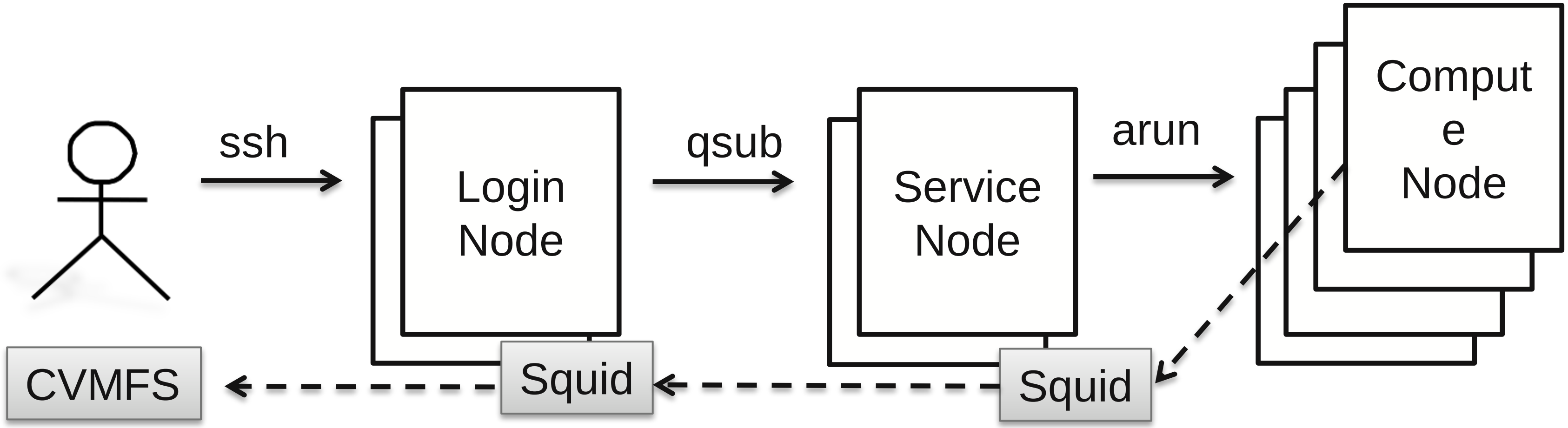
Top 500

The following table gives the Top 10 positions of the 41st TOP500 on June 16, 2013.

Rank	Rmax Rpeak (Pflops)	Name	Computer design Processor type, interconnect	Vendor	Site Country, year	Operating system
1	33.863 54.902	*Tianhe-2*	**NUDT** Xeon E5–2692 + Xeon Phi, Custom	NUDT	National Supercomputing Center in Guangzhou China, 2013	Linux (Kylin)
2	17.590 27.113	*Titan*	**Cray XK7** Opteron 6274 + Tesla K20X, Custom	Cray	Oak Ridge National Laboratory United States, 2012	Linux (CLE, SLES based)
3	17.173 20.133	*Sequoia*	**Blue Gene/Q** PowerPC A2, Custom	IBM	Lawrence Livermore National Laboratory United States, 2013	Linux (RHEL and CNK)
4	10.510 11.280	*K computer*	**RIKEN** SPARC64 VIIIfx, Tofu	Fujitsu	RIKEN Japan, 2011	Linux
5	8.586 10.066	*Mira*	**Blue Gene/Q** PowerPC A2, Custom	IBM	Argonne National Laboratory United States, 2013	Linux (RHEL and CNK)
6	5.168 8.520	*Stampede*	**PowerEdge C8220** Xeon E5–2680 + Xeon Phi, Infiniband	Dell	Texas Advanced Computing Center United States, 2013	Linux
7	5.008 5.872	*JUQUEEN*	**Blue Gene/Q** PowerPC A2, Custom	IBM	Forschungszentrum Jülich Germany, 2013	Linux (RHEL and CNK)
8	4.293 5.033	*Vulcan*	**Blue Gene/Q** PowerPC A2, Custom	IBM	Lawrence Livermore National Laboratory United States, 2013	Linux (RHEL and CNK)
9	2.897 3.185	*SuperMUC*	**iDataPlex DX360M4** Xeon E5–2680, Infiniband	IBM	Leibniz-Rechenzentrum Germany, 2012	Linux
10	2.566 4.701	*Tianhe-1A*	**NUDT YH Cluster** Xeon 5670 + Tesla 2050, Arch[7]	NUDT	National Supercomputing Center of Tianjin China, 2010	Linux

- We have started working with ORNL and TACC, hosts of #2 resp. #6

Lots of CPU, no internet, minimal OS

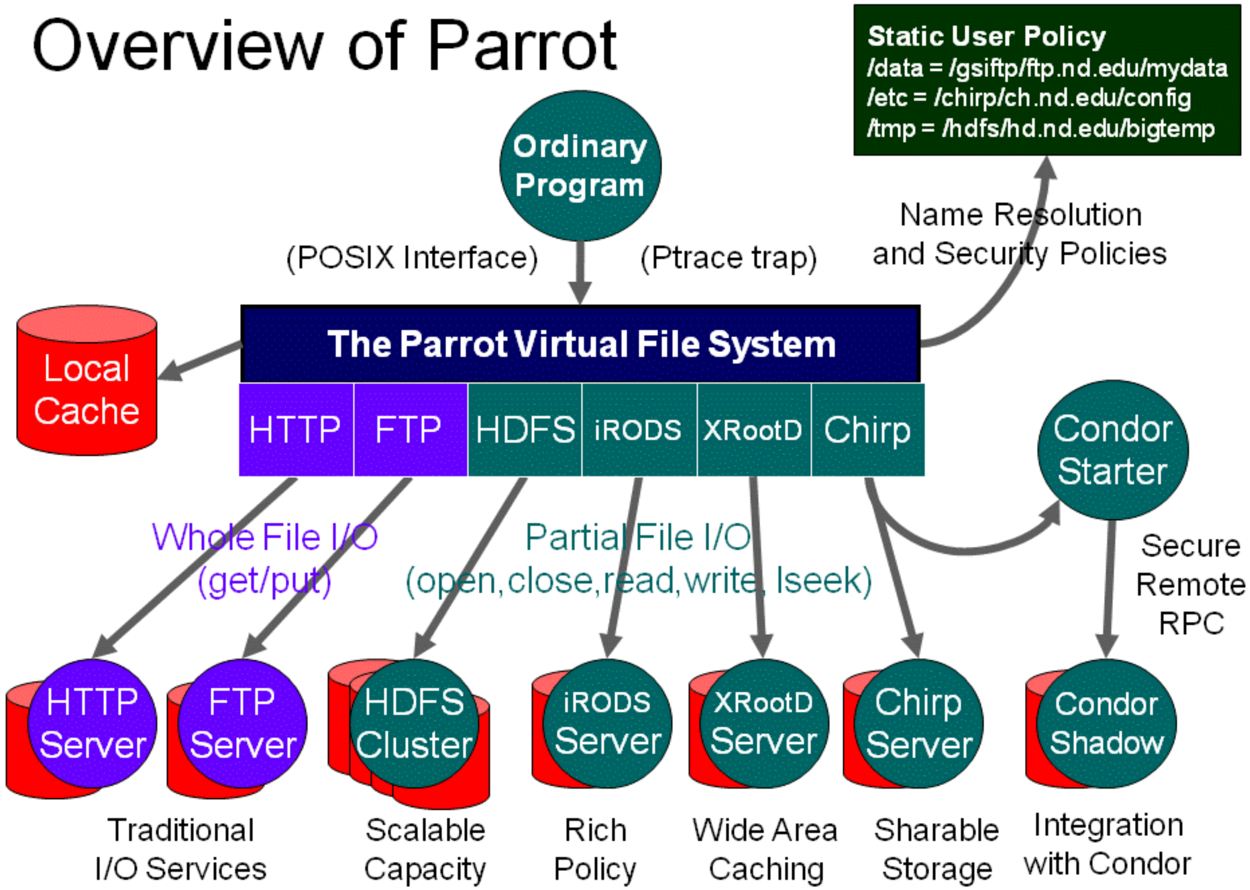



```
$ [titan-batch6][12:12:09][/tmp/work/atj/cvmfs_install/bin]$ aprun ./parrot_run -t/tmp/scratch /cvmfs/alice.cern.ch/bin/alienv setenv AliRoot -c aliroot -b
*******************************************
*       W E L C O M E  to  R O O T       *
*   Version   5.34/08      31 May 2013   *
*  You are welcome to visit our Web site
*          http://root.cern.ch           *
*                                        *
*******************************************

ROOT 5.34/08 (v5-34-08@v5-34-08, Jun 11 2013, 10:26:13 on linuxx8664gcc)

CINT/ROOT C/C++ Interpreter version 5.18.00, July 2, 2010
Type ? for help. Commands must be C++ statements.
Enclose multiple statements between { }.
2+2
(const int)4
```

Thanks to CMS parrot (from cctools) extended to allow access to CVMFS (some restrictions apply)

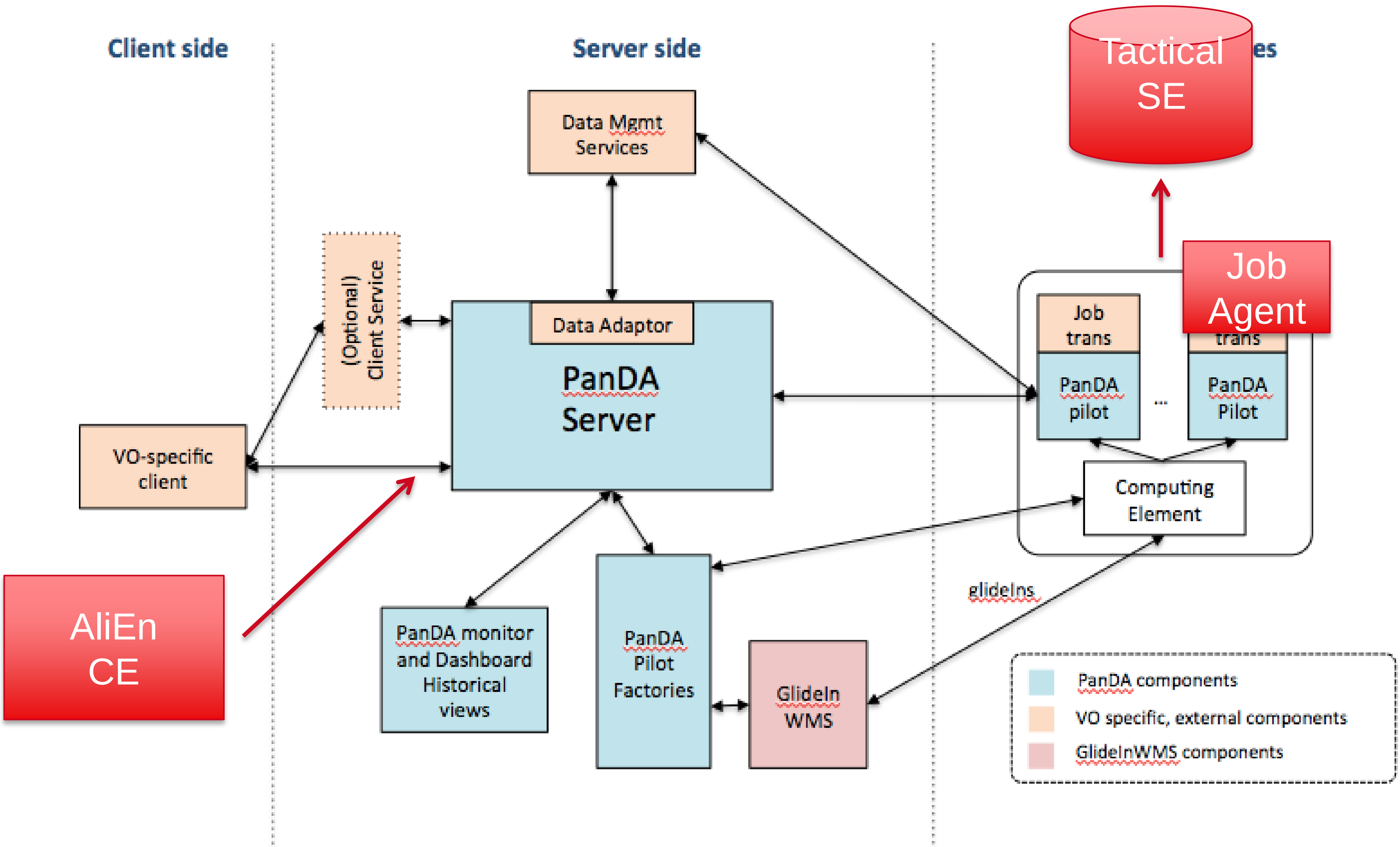
Adam Simpson, ORNL

Overview of Parrot



Parrot is a transparent user-level virtual filesystem that allows any ordinary program to be attached to many different remote storage systems, including HDFS, iRODS, Chirp, and FTP **+ CVMFS**

Interfacing with PanDA



Conclusions

- Run3+ will impose much higher computing requirements
  - 100x more events to handle
- Simply scaling the current Grid won't work
  - We need to reduce the complexity
- Regional clouds may be an answer
  - We are **not** abandoning the grid
  - "Our grid on the cloud" model
- In order to complement these resource we might have to tap into flagship HPC installations
  - Not exactly grid friendly environment, needs some work(arounds)
- Looking for synergies and collaboration with other experiments and IT
  - Transition to clouds, CVMFS, location aware service, edge/proxy services, data management, release validation, monitoring, pilot jobs, distributed analysis, use of opportunistic resources, volunteer computing…