
# TECNOLOGIE CLOUD NEGLI ESPERIMENTI NON LHC

# ATTIVITA, PROSPETTIVE, IDEE

**Stefano Bagnasco & Davide Salomoni**| Milano, 29 novembre 2013

# ARMONIZZAZIONE E COORDINAMENTO

# CATEGORIE DI RISPOSTE

- “Vogliamo avere un Computing Model basato su tecnologie Cloud, ci stiamo lavorando (ma abbiamo poco manpower)”
- “Vogliamo delle risorse ma non abbiamo il manpower per fare dell’R&D, ce le date sulla Cloud?”
- “Ci interessa potenzialmente, basta non dover investire manpower”
- Non sa/non risponde

# FEEDBACK

- CSN1
  - LHC (Claudio)
  - BELLE-II
  - BES-III
  - COMPASS
- CSN2
  - GERDA
  - ?
- CSN3
  - ALICE (Claudio)
  - CLAS12 (JLab)
  - PANDA
  - AGATA
- CSN4
  - Focus su HPC
  - Varie attività
- CSN5
  - Varie attività

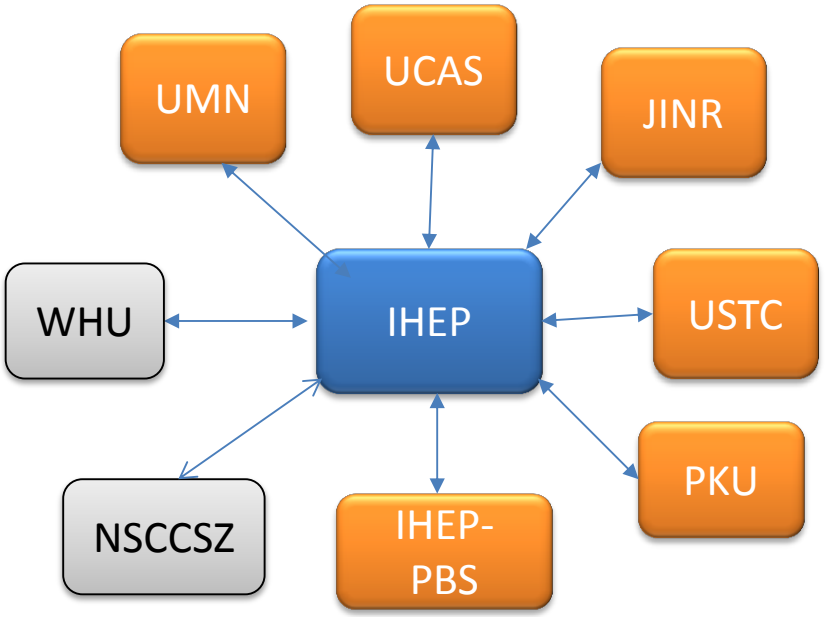
# BESIII – Computing Model

- Reconstruction and Analysis framework: BOSS
  - ROOT based, pretty old: v5.24.00b (Oct. 2009)
  - SLC6 since BOSS 6.6.4
- Batch System @ IHEP based on Torque 2.5.5 + custom tools:
  - 2.5M jobs and 9.2Mh walltime in the last six months
  - currently evaluting Torque 4.2.6 (problems for >20K jobs) and Castor 8.1.0 (ok, testing)
- Storage:
  - current: 3PB disk (lustre 1.x), 4PB tape (castor mod; perf x tape: 60-90MB/s, aggr. 1GB/s)
  - very soon: 5PB disk, 10 pB tape
  - evaluating: Openstack Swift, Loonstore, gLusterfs, Ceph and Luster 2.x
- Distributed computing:
  - current: grid approach
  - BESDIRAC: DIRAC v6r10-pre14 customisation (newly added DFC functions as file-level metadata query, etc)
  - gLite => EMI2
  - SE: StoRM
  - mainly devoted to perform simulations

Marco.Maggiora@to.infn.it

- **IHEP central site**
  - Raw data processing, MC production and analysis
  - Central storage
- **Remote sites**
  - Part of MC production and analysis at peak times
  - 8 sites, 2 (WHU and NSCCSZ) based on virtualization

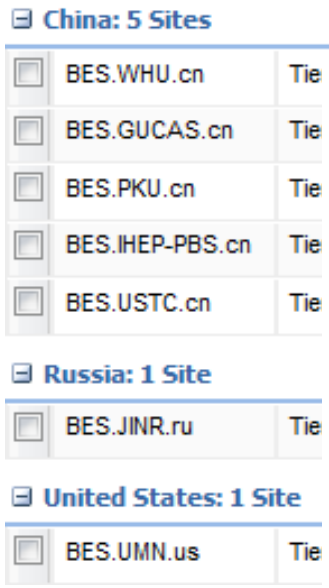
- Current status:
  - MC Production only deployed on remote sites



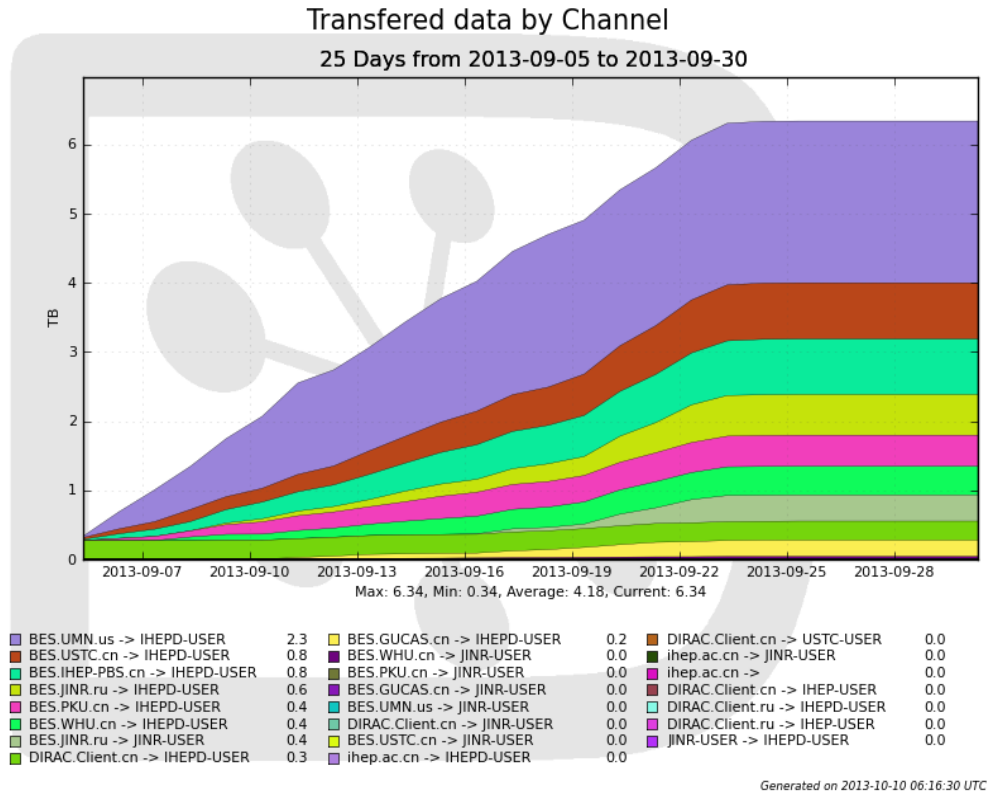

Marco.Maggiora@to.infn.it

# BESIII – Last round GRID simulations

- Simulations of 0.9B J/Ψ
  - 7 sites joined the production




- output: 6TB MC Ra



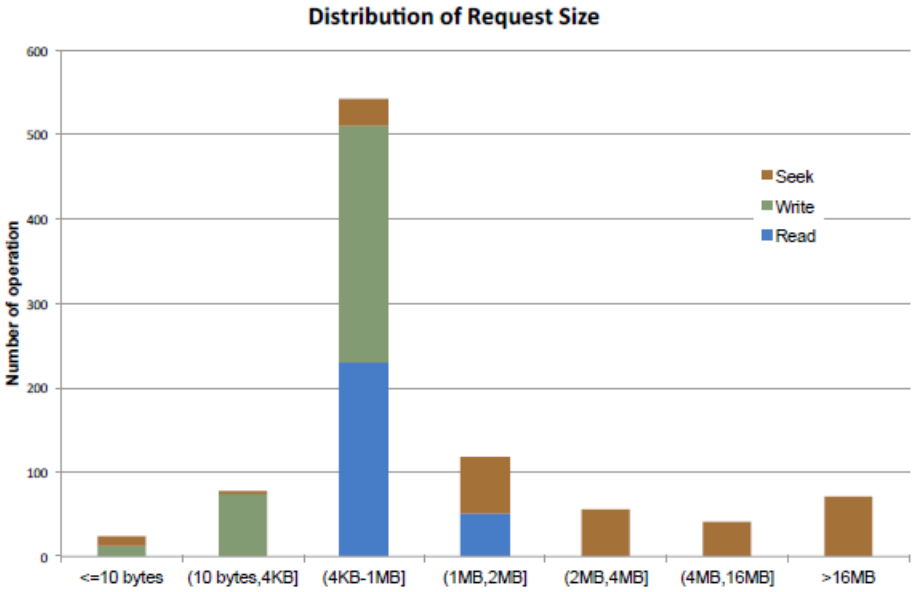

Marco.Maggiora@to.infn.it

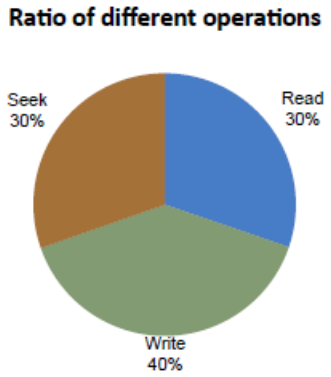
# BESIII – Cloud based SE

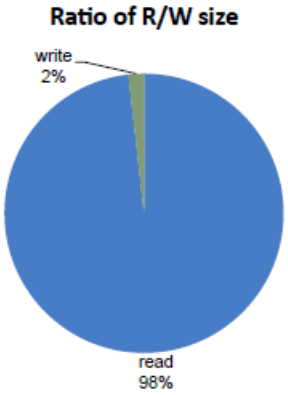
- Use case:
  - write-once, read-many
  - S3 protocol
  - HTTP/HTTPS access
- old (v5.24.00b) ROOT required:
  - SE support built-in since ROOT v.5.34.05 (Feb. 2013)
  - custom extension to ROOT:
    - adding OpenStack Swift support
    - adding S3 support
    - tested vs Amazon S3, Google Storage, Rackspace, Openstack Swift, Huawei UDS
    - no modification to ROOT source needed, backwards to ROOT v5.24
- reconstruction/analysis jobs:
  - custom CLI-based S3 and Swift client
  - evaluating S3fs, FUSE-based developed for Amazon S3
  - testing native Swift vs Amazon S3 (swifts3 plugin); software:
    - server side: OpenStack Swift v1.7.4 (+swift3 module) on SL6
    - client side: s3fs v.1.19, fuse v2.7.4, SL5

Marco.Maggiora@to.infn.it

# BESIII – I/O patterns of analysis job

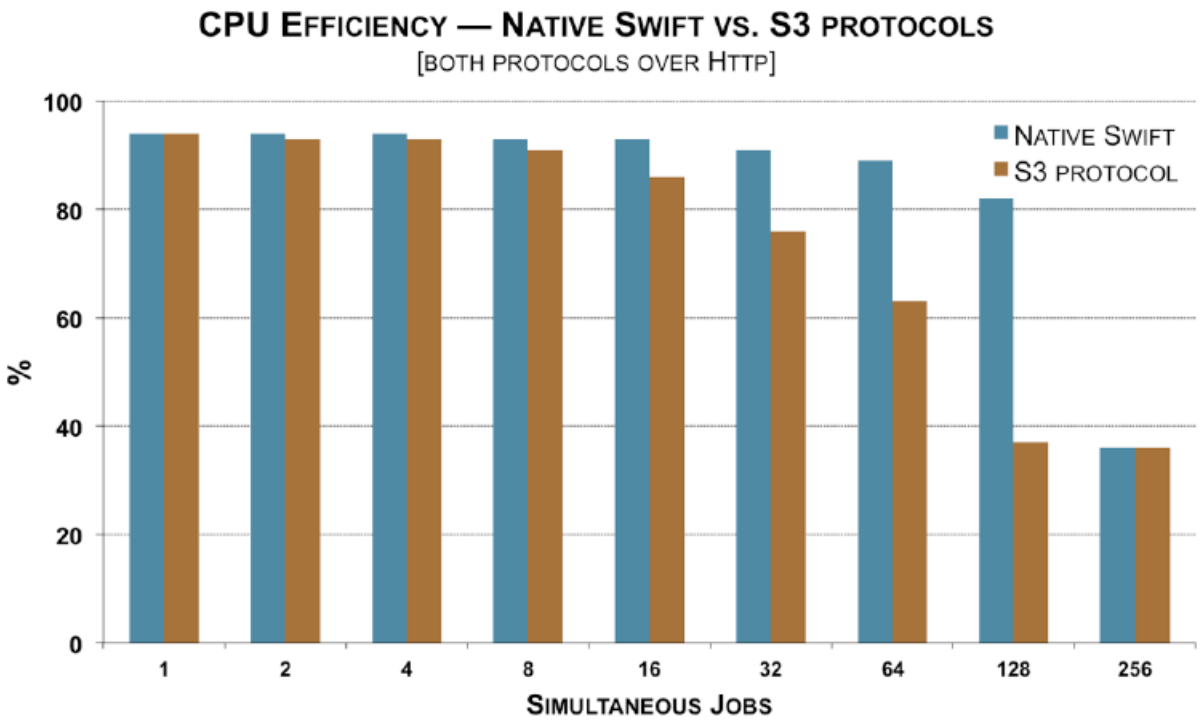



Marco.Maggiora@to.infn.it

# BESIII – Cloud based SE: job efficiency

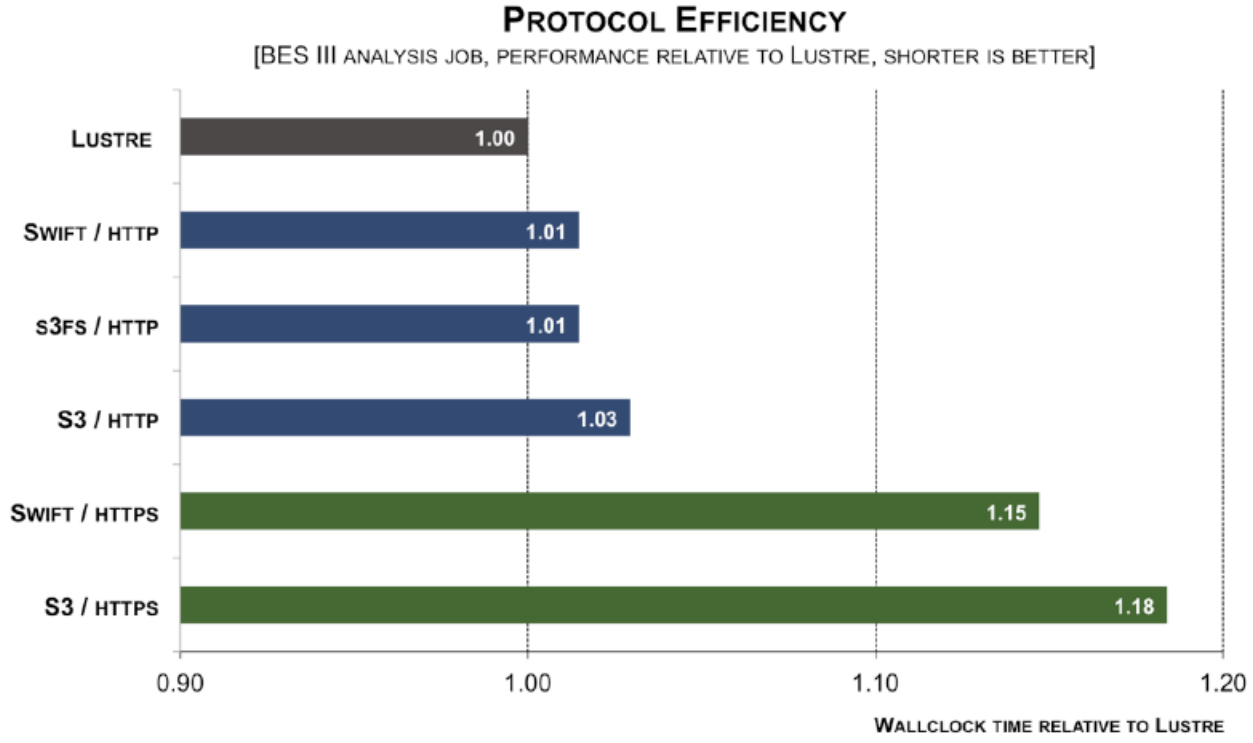



With native Swift protocol, up to 128 jobs can be fed to stay above 80% CPU efficiency. Each job consumes 3.7MB/sec

Lu Wang, Fabio Hernandez, ZiYan Deng @ CHEP2013 (Amsterdam, Oct. 2013)

Marco.Maggiora@to.infn.it

# BESIII – Cloud based SE: job efficiency

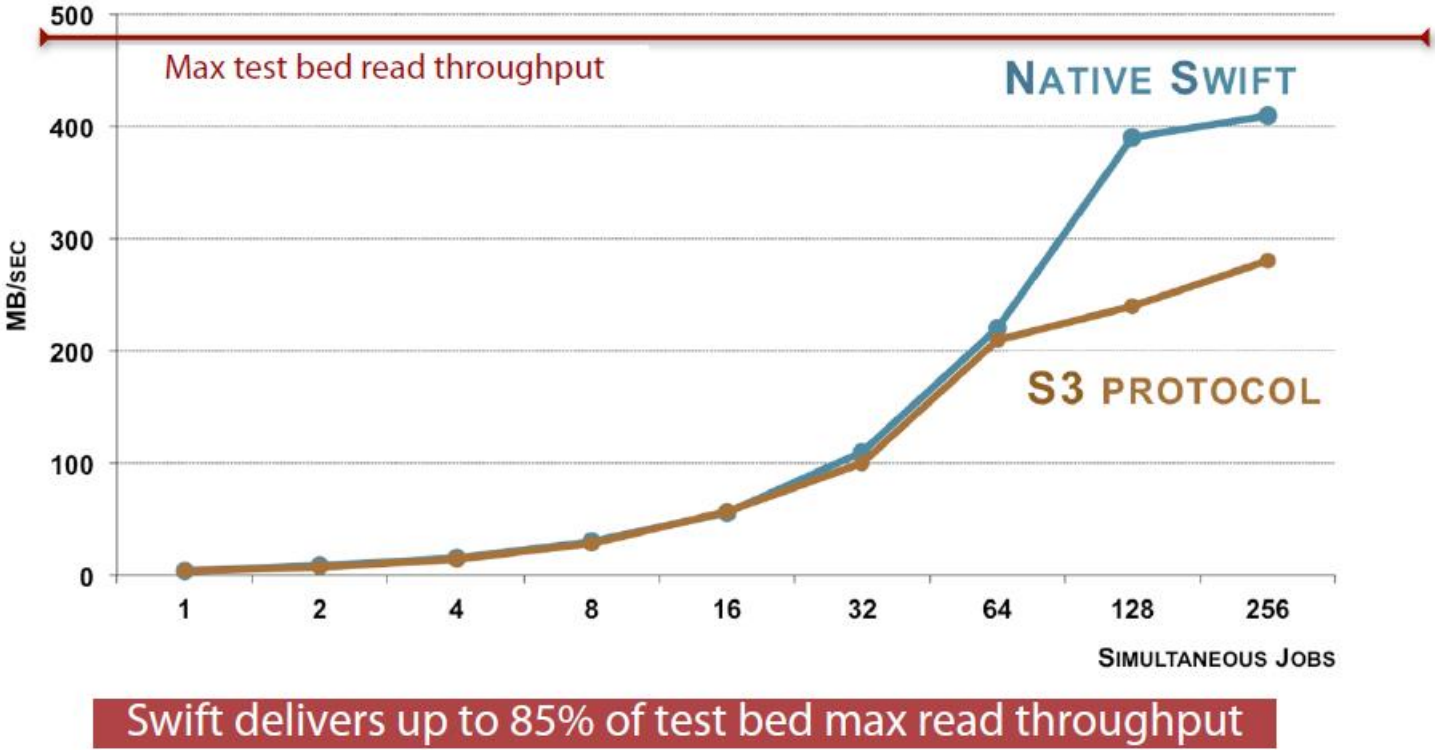



Low overhead of both native Swift and S3 over HTTP
Noticeable penalty when using HTTPS

Lu Wang, Fabio Hernandez, ZiYan Deng @ CHEP2013 (Amsterdam, Oct. 2013)

Marco.Maggiora@to.infn.it

# BESIII – Cloud based SE: job efficiency

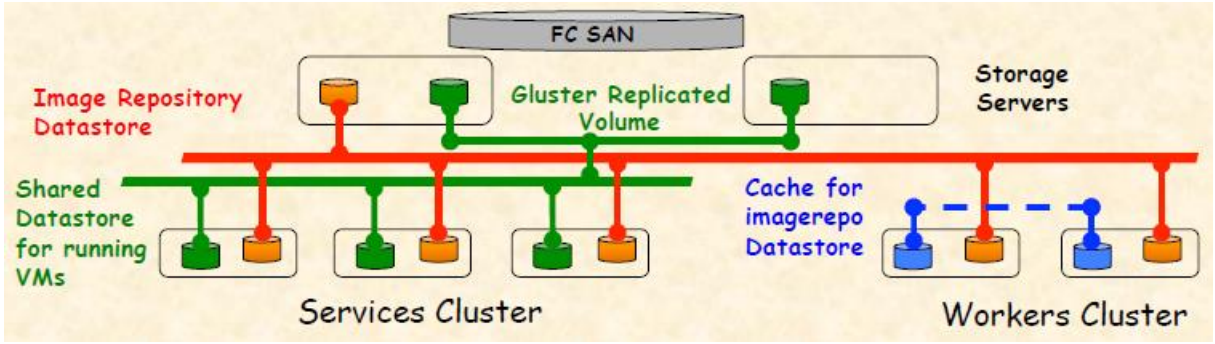



Lu Wang, Fabio Hernandez, ZiYan Deng @ CHEP2013 (Amsterdam, Oct. 2013)

Marco.Maggiora@to.infn.it

# BESIII – Grid Tier-2 on Cloud Virtualisation

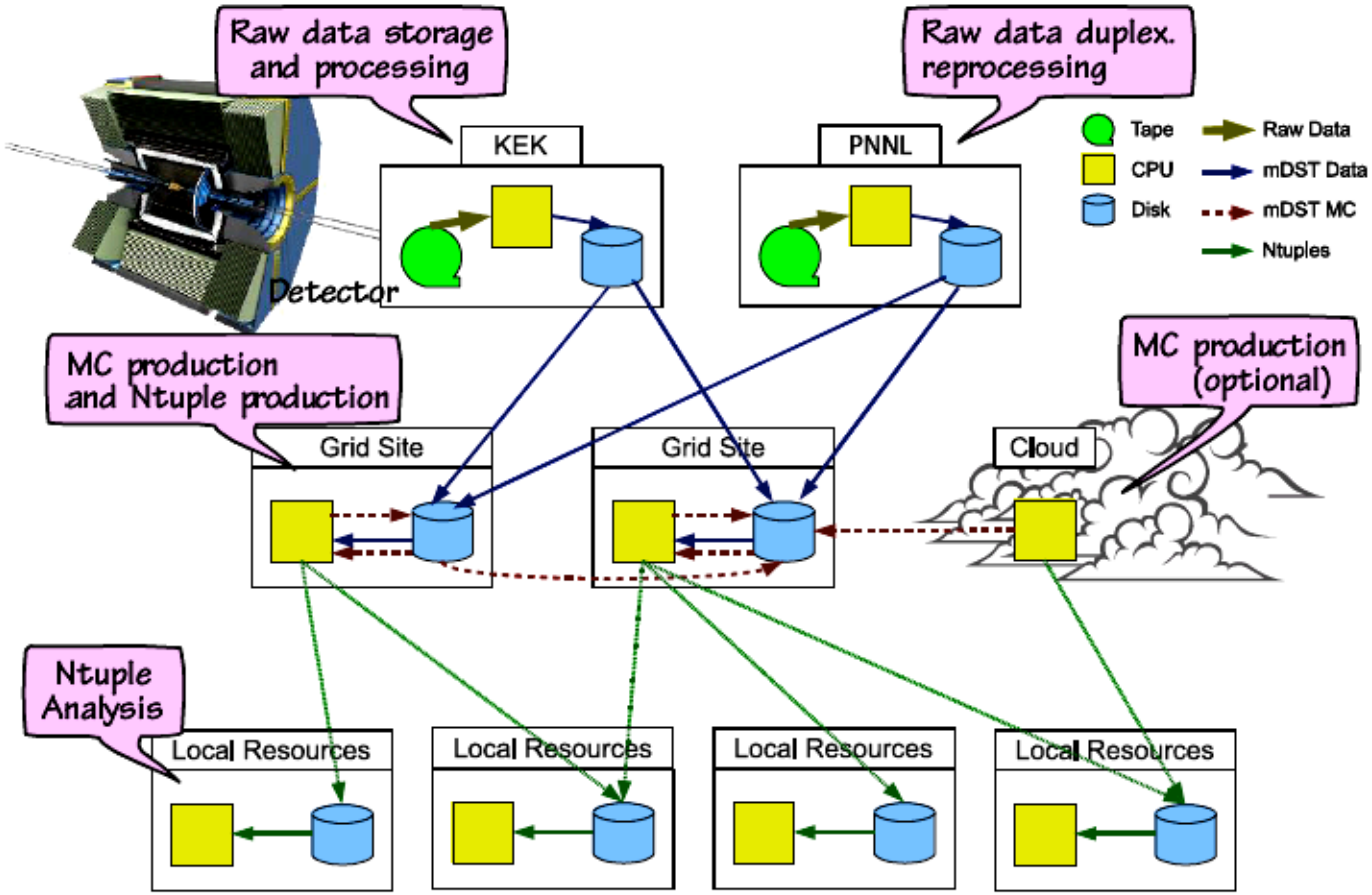
- Use case:
  - virtualised WNs and SE on Cloud infrastructure at remote site
  - exporting BESIII Grid Tier-2 standard remote site
- Turin INFN CdC: test bed
  - critical services VMs: in- & out-bound connectivity, public and private IPs, live migration, no special I/O requirements
  - computing nodes VMs: GRID WNs, private IPs, high storage I/O performance
  - cloud management: OpenNebula 3.6 ( =>3.8 soon), full contestualization via context scripts and puppet (CloudInit)
  - backend storage: gLusterfs




  - VM network: OpenWRT
  - monitoring: Zabbix

Marco.Maggiora@to.infn.it

# Belle II: Computing Model




Fabrizio.Bianchi@to.infn.it

# Belle II: Computing Model

- In sintesi:
  - Seconda copia dei raw data a PNNL
  - Ricostruzione dei raw data a KEK
  - Produzione MC su Grid e Cloud
  - Prevista la necessita' di reprocessing dei raw data (a PNNL) e rigenerazione di MC
- Accesso ai microDST via Grid e/o Cloud
  - Usando il tool DIRAC per produrre n-tuple
- Analisi n-tuple su risorse locali



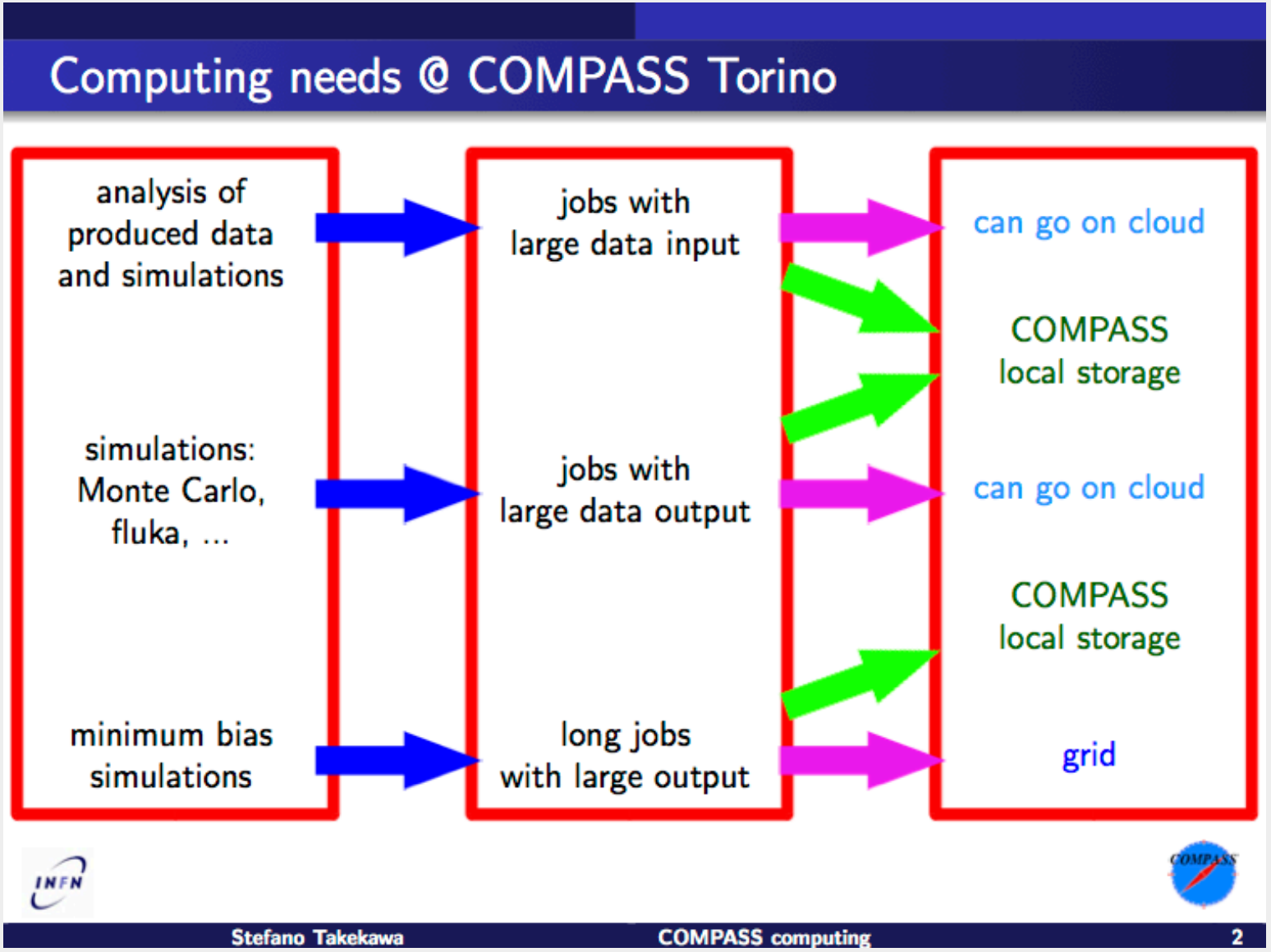
Fabrizio.Bianchi@to.infn.it

# Belle II: Uso del Cloud Computing

- Test in corso a Melbourne ed in Canada (Uvic) per eseguire il framework di Belle II su Cloud
  - Disegno delle macchine virtuali
- Il nostro piano e' di aspettare che Melbourne e Canada sviluppino un oggetto funzionante e poi portarlo sulla Cloud INFN.



Fabrizio.Bianchi@to.infn.it

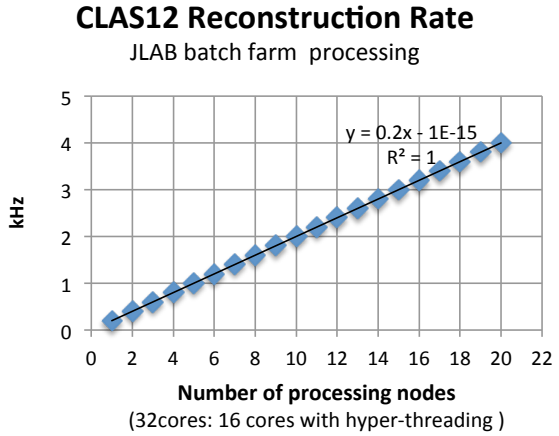


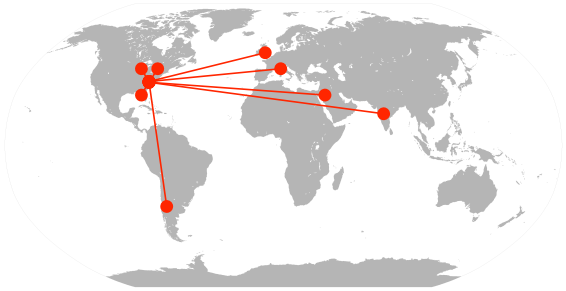

Stefano.Takekawa@to.infn.it

# CLARA

## CLAS12 Reconstruction and Analysis Framework

- Cloud computing framework.
  - Implements SOA architecture
- Data processing major software components as services (SaaS)
  - Multilingual support
    - Services can be written in C++, Java and Python
- Data (storage and persistency) as a services (IaaS)
- Supports both traditional and cloud computing models
  - Single process as well as distributed application design modes
  - Centralized batch processing
  - Distributed cloud processing
- Utilization of multicore processor systems
  - Built in Multi-threading of a user service
    - Requires thread safety of a service code
- Ability to expand computing power with minimal capital expenditure
  - Dynamic elasticity.
  - Utilization of IT resources of collaborating Universities.
  - Take advantage of available commercial computing resources.
- Summary
  - Thin clients to organize and conduct data cloud processing
  - On-demand data processing.
  - Location independent resource pooling.
  - Software agility



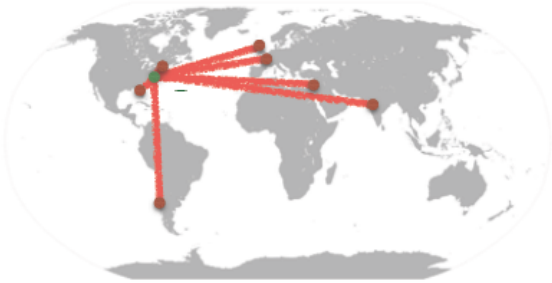
Cloud processing.
CLAS data on the ODU data-center is being actively analyzed by users from Scotland, Germany, Chile, India, Israel, MIT, FSU, ODU and others.

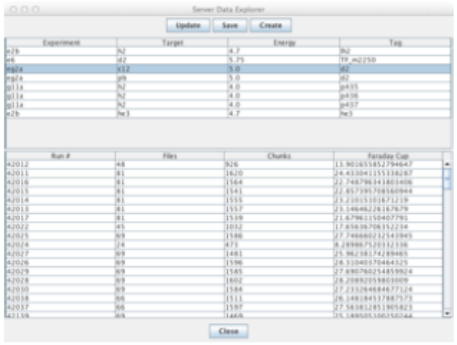
JLab/CLAS12

Marco.Battaglieri@ge.infn.it

# COAT Data Management System

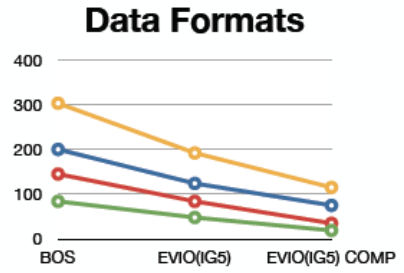
- CLARA based data management, distribution and analysis system for CLAS:
  - World wide access to the data
  - Analysis Services for specific data sets
  - On-Demand data processing and skimming
  - Ability to download predefined 4-vectors run full analysis on cloud servers.





- Tagged File System:
  - File description services for searching data over many clusters.
  - Tagging interface for categorizing files by experiment type (beam, target)
  - Run condition database for tags describing experiment run conditions.
  - GUI for file search and download

- Efficient Data formats:
  - Custom developed data formats for CLAS data.
  - Buffered data stream for efficient transfer between services.
  - Lossless compression for storage efficiency




JLab/CLAS12

Marco.Battaglieri@ge.infn.it

# GAMMA (CSN III) : typical data analysis

| | standard experiments @ various facilities in Italy/Europe/Japan/USA |
|---|---|
| A) | * Single experiments in dedicated campaigns<br>•Max 10 days of data taking<br>➔ event size ~0.1 kB<br>➔ max 1-5 TB/experiment RAW data + 0.5-2 TB "replayed" data |
| B) | * Analysis involves 3-4 people for 1-4 years<br>* Dedicated analysis SW based on self-developed packages (GASPWARE/ RADWARE…) or, lately, ROOT-based<br>* Usually every analysis runs on a dedicated workstations<br>*Typical HW: 4/8 cores + 32 GB RAM + disk space |

AGATA experiments are quite different in character and requirements
we cover 2 roles: user and, from time to time, also host

Giovanna.Benzoni@mi.infn.it

| | AGATA |
|---|---|
| A) | * Long campaigns in different labs ( LNL / GSI / Ganil )<br>* ~ 10 days of data taking/ experiment: event rate 100 MB/s<br>➔ raw event size ~10 kB<br>➔ up to 30 TB /experiment RAW data + 5-10 TB "replayed" data |
| | • Crucial analysis step is the Pulse Shape Analysis (PSA)<br>➔ Very CPU intensive<br>~ 20 workstations with ~200 (LNL) o 500 (GSI) cores<br>➔ After PSA data is γ-ray tracked and saved in analysis format<br>factor ~ 20 reduction of data size<br>• In principle this is done ON-LINE but the quality is not sufficient; therefore, we save the raw data and repeat the process off-line<br>• Still evaluating if the off-line replay can be done on the GRID |
| B) | Same as of standard experiments |

Giovanna.Benzoni@mi.infn.it

# Main Issues

Storage ➔ local farms (~200 TB) for standard exp. (Mi/Pd/LNL)
➔ ~150 TB/y raw data of our interest from the AGATA campaigns
➔ copied locally from the ~300 TB archived so far at CNAF and Lyon
➔ 100 TB/y for GALILEO campaigns @ LNL from 2015
➔ ~300 TB for the next AGATA campaign at LNL
➔ archived data should be available for ~10 years

Backup of data at "replay" level ➔ managed locally with different solutions

Fast and Frequent access to data (at "replay" level)

Use of standardize workstations dressed with typical SW (➔ virtual machines)

Possible need to re-analyze data after years (➔ virtual machines)

AGATA working group working to port the AGATA PSA to the GRID

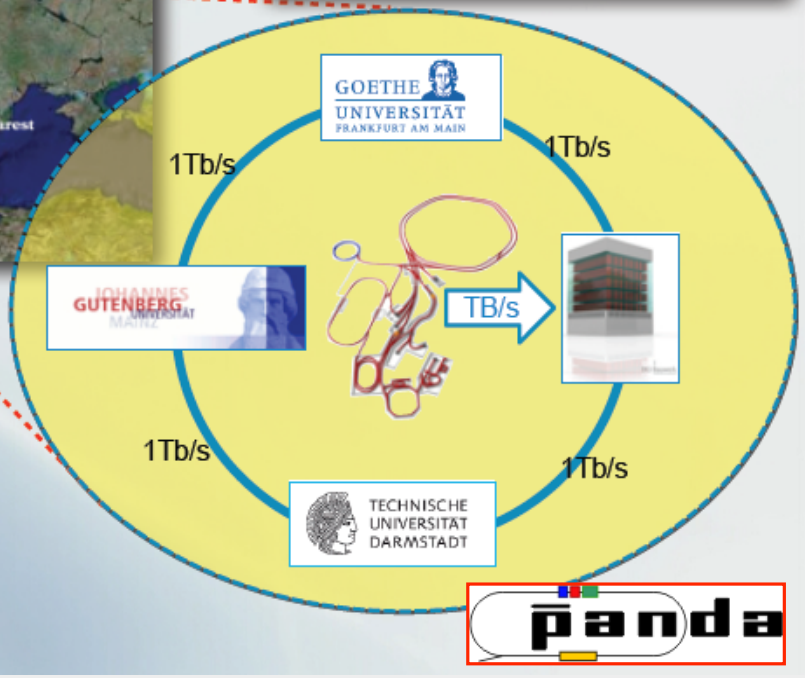
Giovanna.Benzoni@mi.infn.it

# The current PANDA Computing Model

## Distributed T0/T1 centre embedded in Grid/Cloud


APPA
CBM
LQCD
NuStar
Panda (66k cores,12PB disks,12PB/y tape)

in 2018
300000 cores + grid
40 PB disk
40PB/y archive




computing center for high-energy, hadron, nuclear, and atomic physics

serve >20 collaborations

Stefano.Spataro@to.infn.it

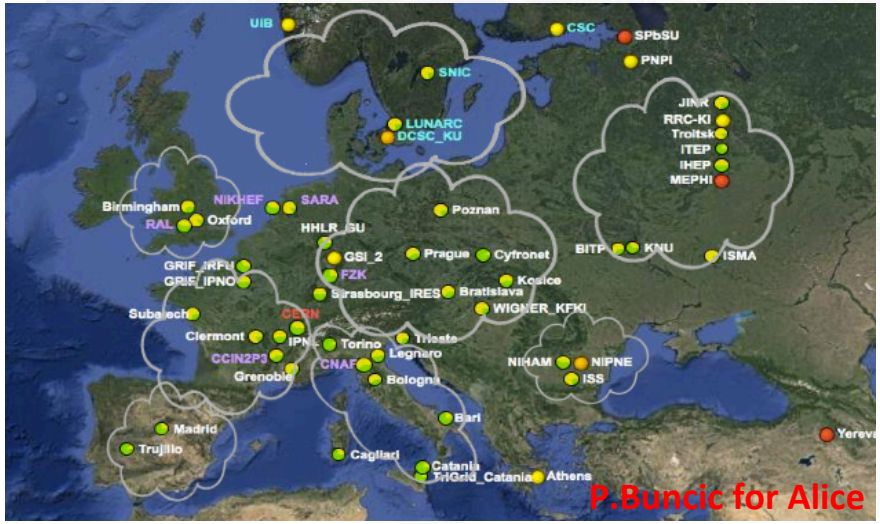
# The "Final" PANDA Computing Model ?

- Original PANDA Computing Model written in 2008
- Now the Computing Model is under revision for 2018 data taking

→ Computing TDR (end of 2014?)

Not enough PANDA computing manpower
to develop our own system or to try many different concepts

Few big supercomputing farms? (less infrastructure costs?)
Distributed computing: A cloudish Grid?




INFN Cloud?

Reduce complexity
Easier (?) to maintain
Reduced experiment support

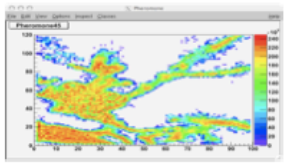
**In Torino Panda jobs are already running on our private cloud**

Stefano.Spataro@to.infn.it

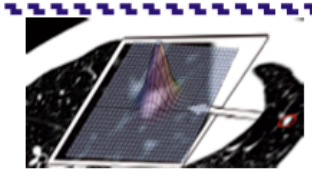
# the M5L algorithms

Lung Segmentation → Candidate Nodule identification → Candidate Nodule Feature Extraction → Candidate Nodule Classification

Region Growing



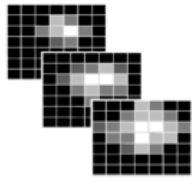
Virtual Ants



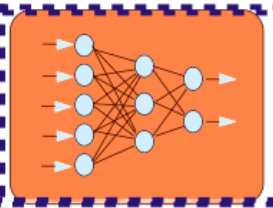
3D Multiscale Gaussian Filter + Pleura Surface Normals

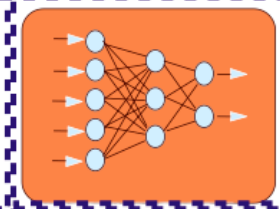
≈ 10-20 features

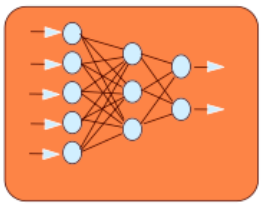
Volume,
Sfericity,
Ellipticity,
Compactness,
Shannon's Entropy,
...



Voxel Based Analysis







RGVP CAM VBNA

CARS 2012

Pisa, June, 28th, 2012 Piergiorgio Cerello (cerello@to.infn.it)

INFN Istituto Nazionale di Fisica Nucleare

Piergiorgio.Cerello@to.infn.it

# the WEB-based M5L

diXit

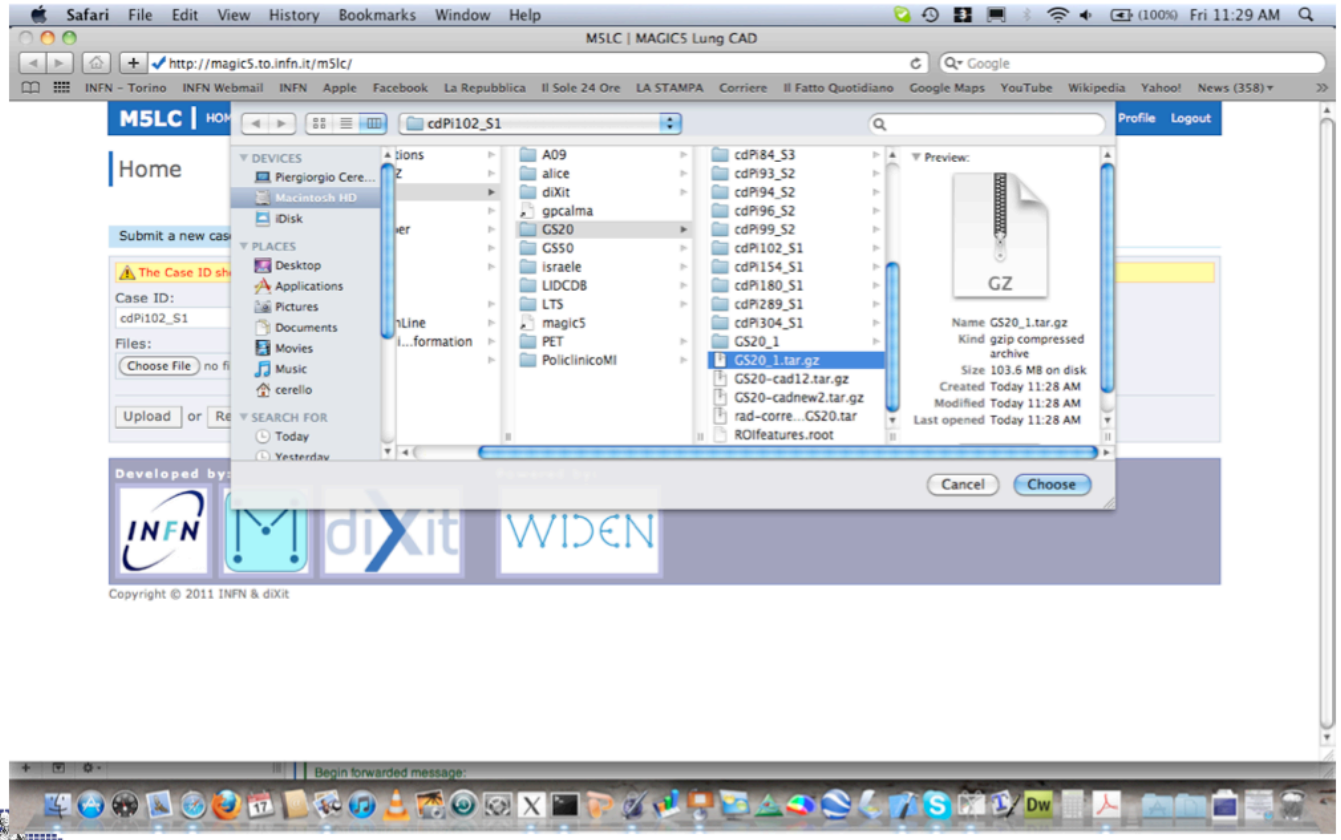
- **the concept: M5L CADe as a service**
  - WEB-interface for registration/CT upload
    - VBNA
    - RGVP
    - CAM
  - (Link to) Combined result sent to the user/reviewer as e-mail


Pisa, June, 28th, 2012 Piergiorgio Cerello (cerello@to.infn.it)

INFN Istituto Nazionale di Fisica Nucleare

Piergiorgio.Cerello@to.infn.it

# the WEB-based M5L

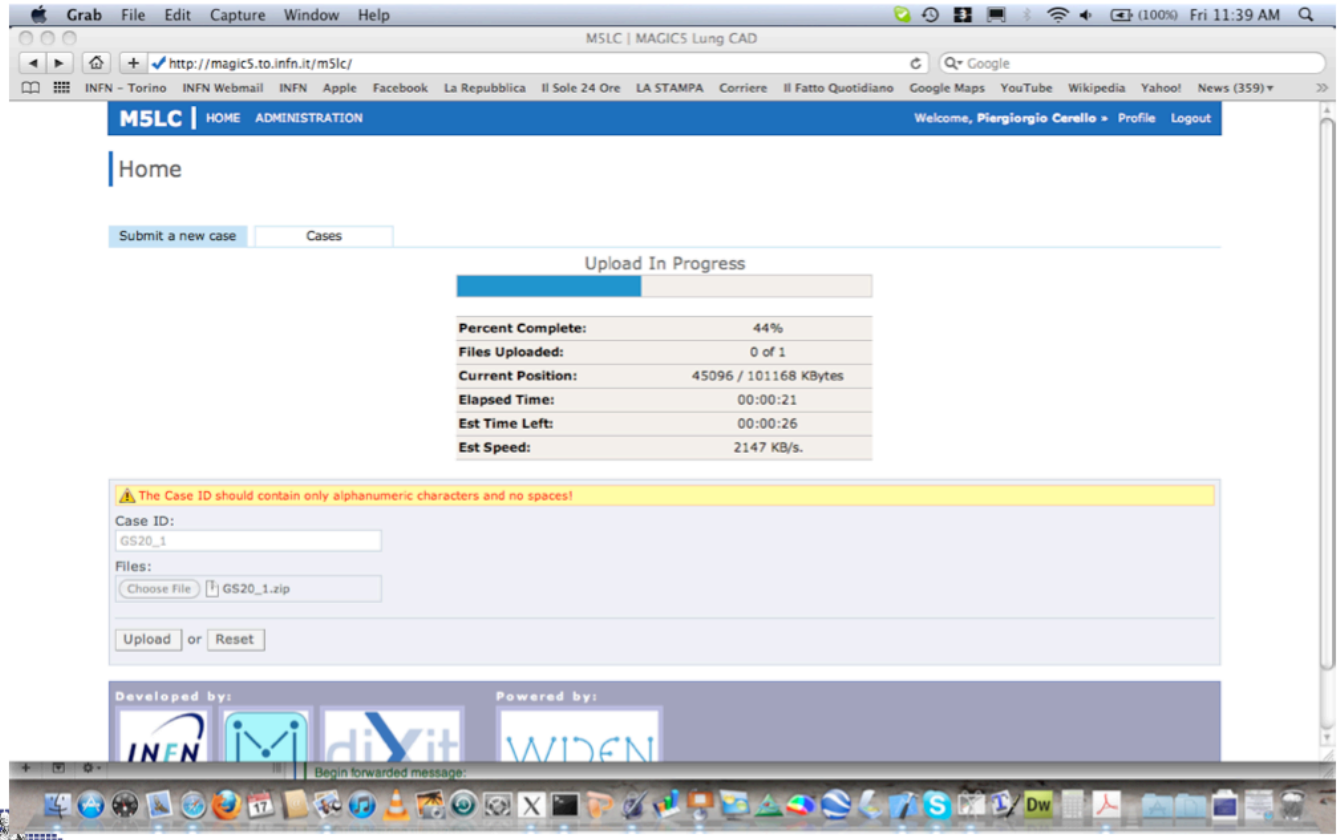



Pisa, June, 28th, 2012

Piergiorgio Cerello (cerello@to.infn.it)

Piergiorgio.Cerello@to.infn.it

# the WEB-based M5L

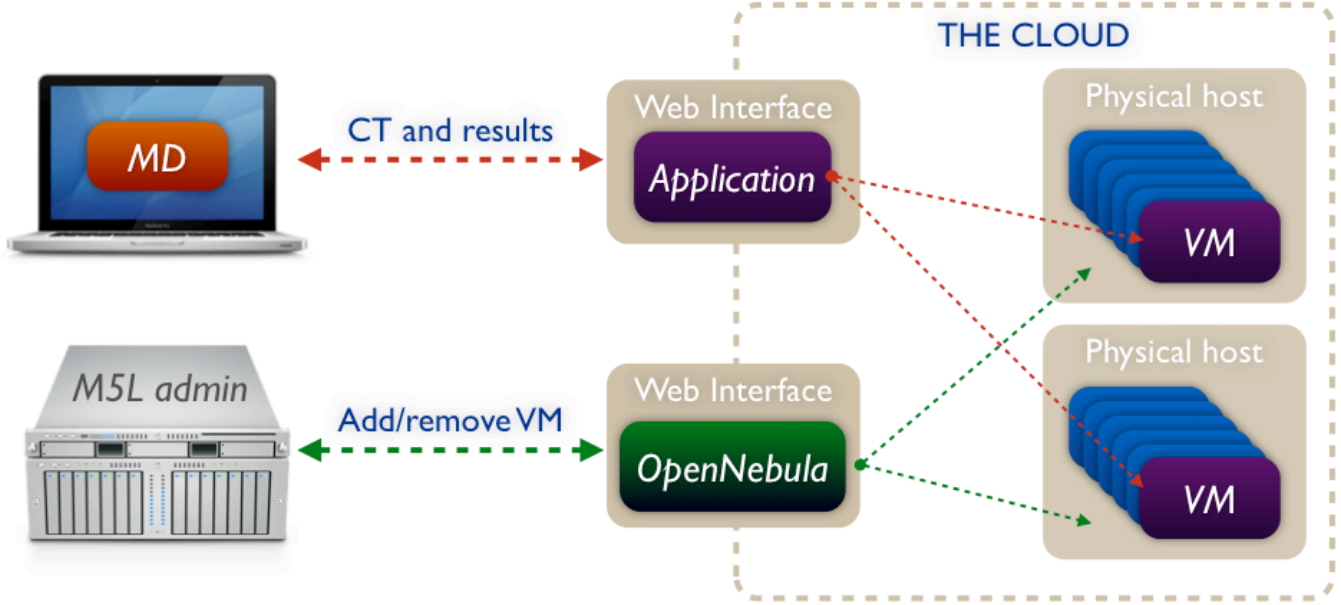



Pisa, June, 28th, 2012

Piergiorgio Cerello (cerello@to.infn.it)

Piergiorgio.Cerello@to.infn.it

# the 'cloud' engine




- **Software as a Service** → the **Medical Doctor** only sees the Web interface: s/he uses a software which is not installed on her/his client, but **subscribes** to a remote service
- **Infrastructure as a Service** → **M5L-CADe admins** do not own physical machines: they **subscribe** in turn to a service which adds or removes VMs on demand

Pisa, June, 28th, 2012 Piergiorgio Cerello (cerello@to.infn.it)

Piergiorgio.Cerello@to.infn.it

# Development


- **Step0: M5L validation**
  - ✦ successful validation on public datasets (ANODE09, LIDC)
- **Step1: prototype service**
  - ✦ feedback required to improve the stability and functionality
  - ✦ the WEB/Cloud-based system is intrinsically scalable
  - ✦ INFN - Torino is providing a cloud computing facility


Pisa, June, 28th, 2012 Piergiorgio Cerello (cerello@to.infn.it)

INFN Istituto Nazionale di Fisica Nucleare

Piergiorgio.Cerello@to.infn.it

# Development

- **Step2: full service?**
  - a tool to speed-up/improve the diagnosis
    - ✦ integrated with inter/national screening programs
    - ✦ for everyday clinical practice
  - INFN Istituto Nazionale di Fisica Nucleare as infrastructure provider
  - diXit as service provider


Pisa, June, 28th, 2012 Piergiorgio Cerello (cerello@to.infn.it)

INFN Istituto Nazionale di Fisica Nucleare

Piergiorgio.Cerello@to.infn.it

# Development

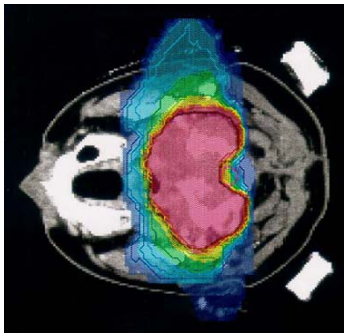
- **Step2: full service?**
  - a tool to speed-up/improve the diagnosis
    - integrated with inter/national screening programs
    - for everyday clinical practice
  - INFN as infrastructure provider
  - dIXit as service provider

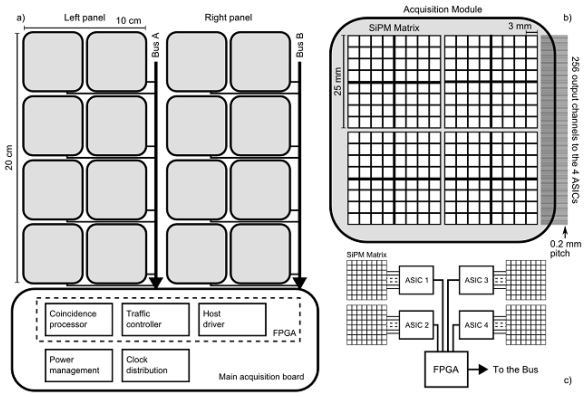

Pisa, June, 28th, 2012 Piergiorgio Cerello (cerello@to.infn.it)

Piergiorgio.Cerello@to.infn.it

# INSIDE: use case

- The INSIDE project will build a PET scanner and a fiber tracker to be used for real-time in-beam monitoring in **hadrontherapy**, so as to verify the agreement between the treatment plan and the actual dose distribution.

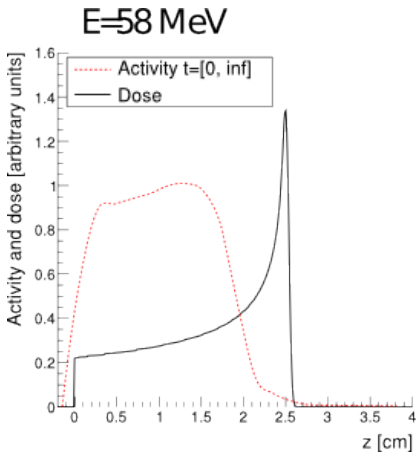




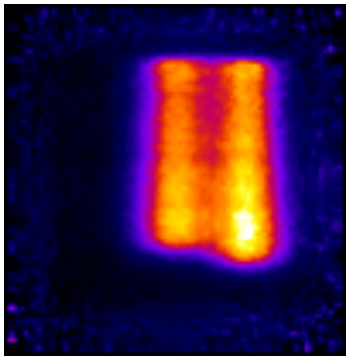

Francesco.Pennazio@gmail.com

# INSIDE: use case

- Detector and Analysis optimization require the Monte Carlo simulation of full treatment plans. (more than $10^{10}$ particles / treatment)
- Data analysis and reconstruction




Dose (black) and activity (red) distribution



Reconstructed activity for two spills at different energy

- Parallelization: each beam spot can be simulated separately, so a cloud-based environment would be very welcome
- Time scale (2.5 years)

Francesco.Pennazio@gmail.com

# DISCUSSIONE!