Il Calcolo della CSN2/22

CSN2
Fisica delle
Astroparticelle

G. Mazzitelli per il GdL calcolo CSN2

organizzazione calcolo in CSN2

ri-organizzazione 2021—>2022

- **GdL Calcolo** - studio dei **modelli di calcolo e linee guida** per la gli esperimenti della CSN2 (S. Bagnasco, M. Duranti, G. Mazzitelli, A. Menegolli, F. Di Pierro, M. Punturo)

- **Referee calcolo della CSN2** - valutazione delle richieste di **risorse e licenze** al Comitato Nazionale Calcolo - CNC - dell'INFN (M. Duranti e F. Di Pierro)

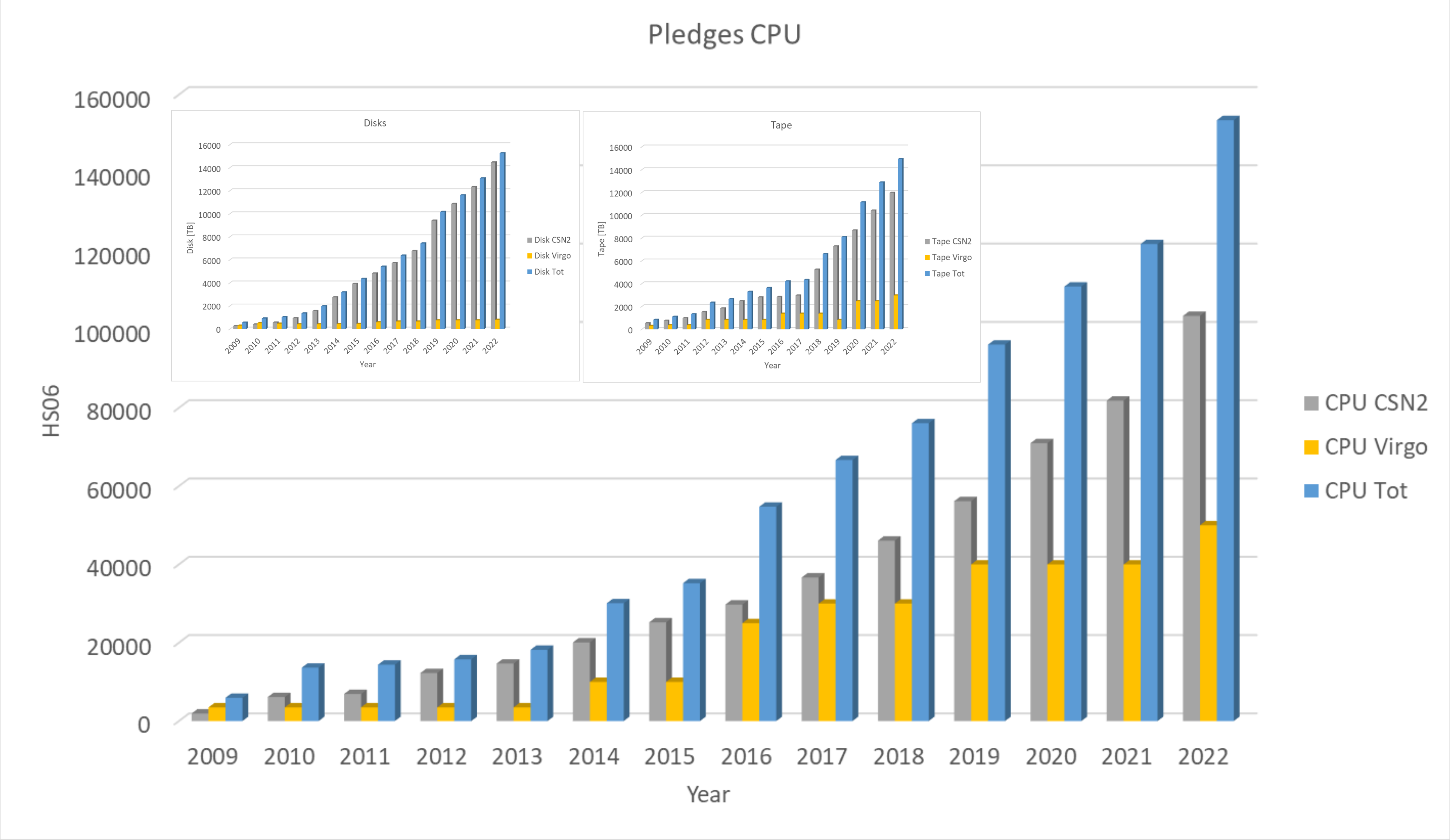
- **Referente nel Comitato di Steering** del Coordinamento Nazionale Calcolo - C3SN - (G. Mazzitelli)
  - WG Computing Model INFN (per la CSN2: G. Mazzitelli, M. Duranti, F. Di Pierro , S. Bagnasco, M. Tenti) coordinamento D. Elia
  - Il WG Computing Model INFN diventa anche **GdL Calcolo** di CSN2 per il 23

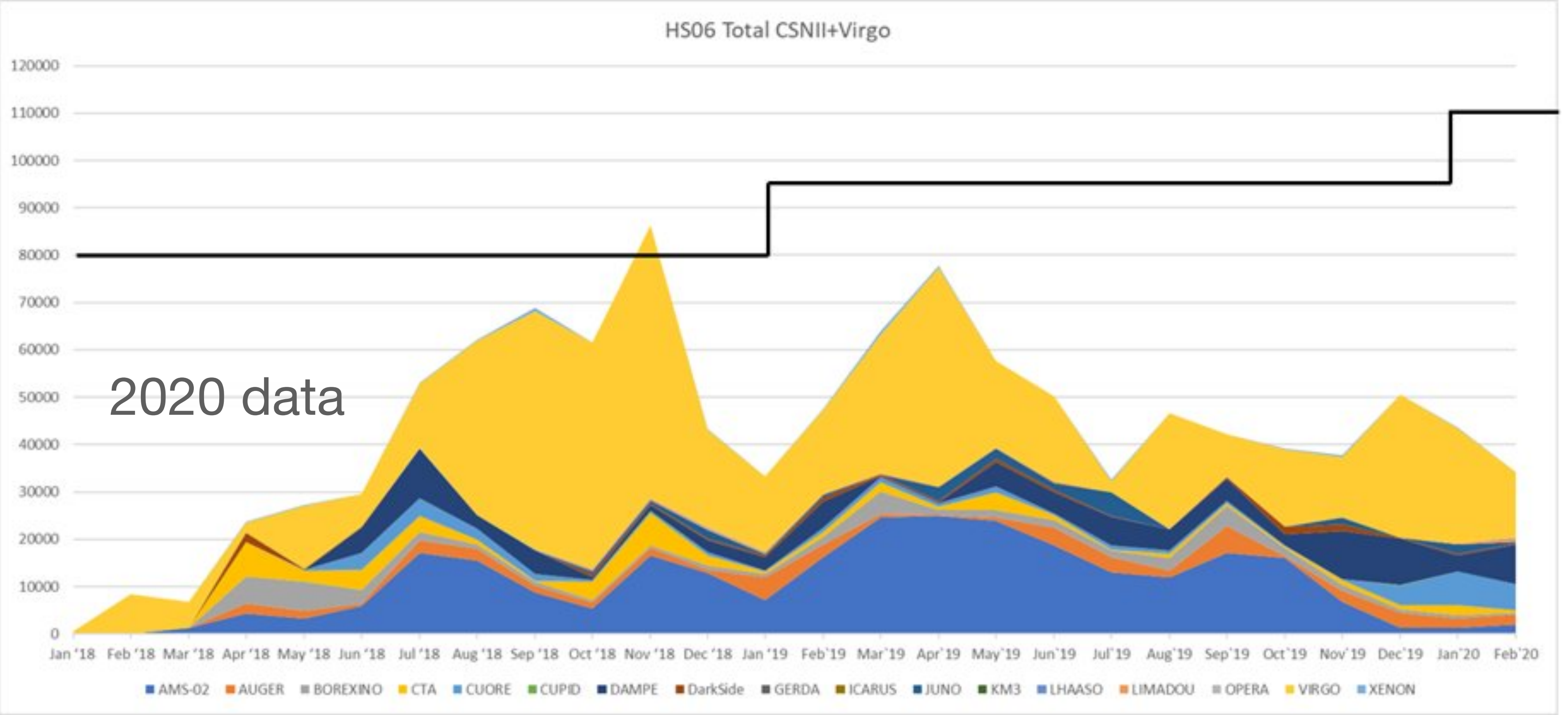
il calcolo in CSN2

la CSN2 ospita 50 sigle composte da O(5 –> 500) FTE

CSN2 Fisica delle Astroparticelle

- il calcolo della CSN2 sta **crescendo** costantemente
- **l'impegno** dell'INFN/CSN2 nei futuri esperimenti sul calcolo e' sempre maggiore (JUNO, DUNE, CTA, HK, VIRGO, …)
- i **costi** sul Tier1 sono paragonabili a quelli della CSN1 (vedi Duranti/Di Pierro)
- l'utilizzo delle pledge allocate su esperimenti di CNS2 dimostra **inefficienza** (~ 75%) e **discontinuità** oltre che una **eterogeneità** notevole delle richieste.





Conclusioni

o meglio riflessioni

CSN2
Fisica delle
Astroparticelle

- il GdL per il calcolo della CSN2 sottolinea la necessità sempre più importante di **definire**, similmente al TIER, **procedure e referenti** per le assegnazione delle risorse HPC e CLOUD

- Ritiene in oltre che questa **trasformazione** in atto dal solo modello di calcolo sulla GRID al'HPC e la CLOUD **vada ben supportata attraverso l'introduzione di figure capaci di ottimizzare e razionalizzare le risorse**

- Suggerisce quindi sia l'inquadramento di un GdL più **tecnico**, capace di mettere in evidenza **sinergie a livello tecnico fra i progetti di CSN2** e possa suggerire **ottimizzazioni e una strategia per gli anni a venire**.

- Suggerisce infine di procedere ad una **ricognizione dei "modelli di calcolo"** (dal DAQ all'analisi) per permettere una analisi più accurata dello stato attuale.

dalla giornata sul calcolo della CSN2 di settembre 2021

Review dei modelli di calcolo

Mandato del comitato 1/2

CSN2 Fisica delle Astroparticelle

- Il grande range dinamico delle necessità di calcolo degli esprimenti di CSN2, la eterogeneità dei requisiti di storage e distribuzione dei dati, la differenza nelle dimensioni degli esperimenti fanno sì **non ci possa essere una unica soluzione valida per tutti gli esperimenti**
- Ma, se gli esperimenti di **medie-grandi dimensioni hanno le risorse per implementare il proprio modello**, gli esperimenti **medio-piccoli possono trarre grande beneficio da una standardizzazione** che possa offrire accesso a nuove soluzioni, come l’uso del cloud INFN

dalla giornata sul calcolo della CSN2 di settembre 2021

Review dei modelli di calcolo

Mandato del comitato 2/2

CSN2
Fisica delle
Astroparticelle

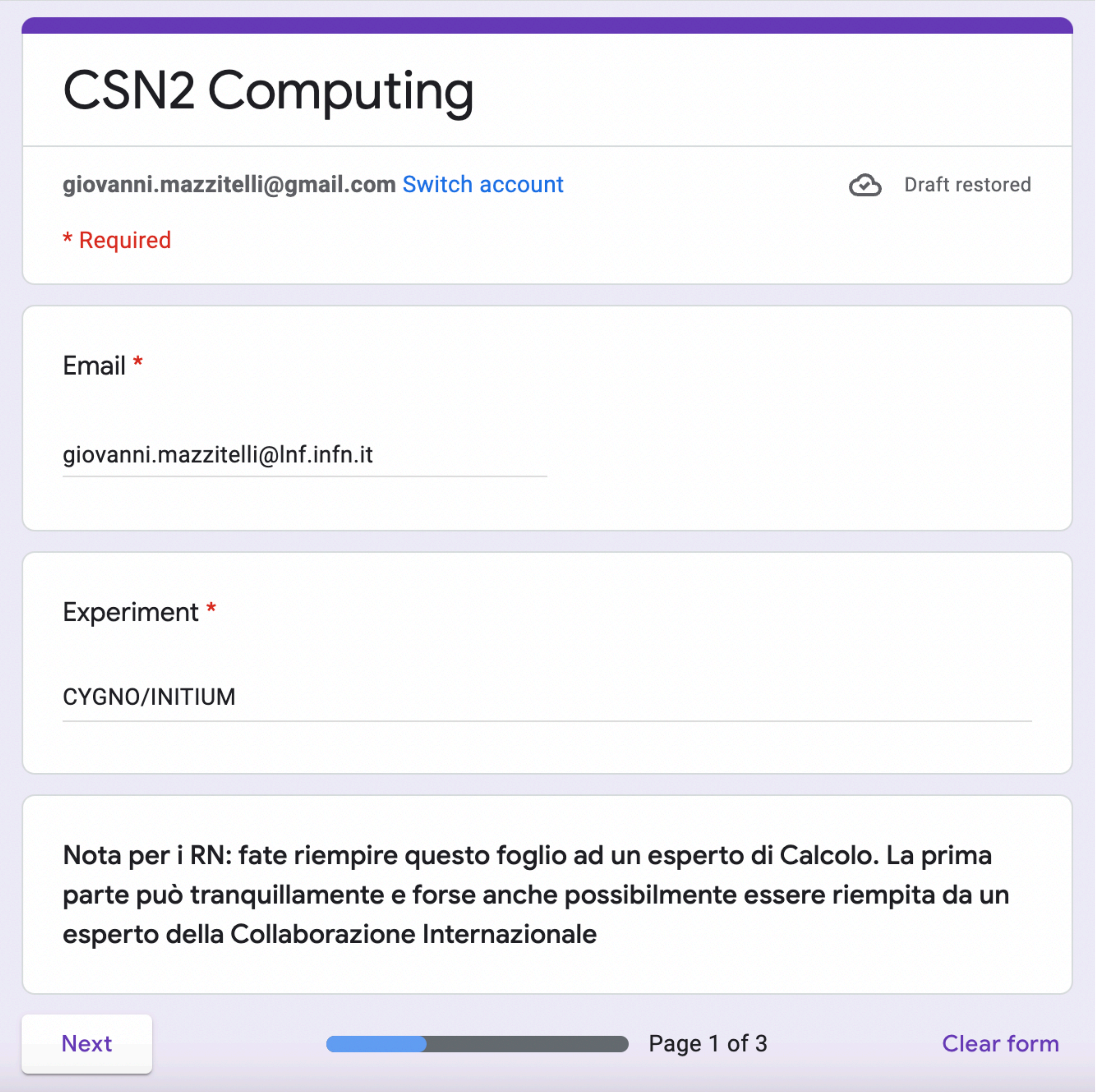
- Per questi motivi si costituisce un comitato con anche esperti esterni, ma «vicini» alla CSN2, che
    - **Raccolga e analizzi lo status quo** delle soluzioni adottate in CSN2 dagli esperimenti
    - Raccolga e analizzi i **desiderata**
    - Formuli dei **suggerimenti d'indirizzo**, non rivolti al **singolo esperimento**, ma che realizzino delle **linee guida** inquadrando, possibilmente, un numero limitato di soluzioni
    - Questo processo dovrebbe raggiungere la prima milestone, producendo un primo draft del **documento di indirizzo**, nell'arco temporale di 6 mesi

dalla giornata sul calcolo della CSN2 di settembre 2021

GdL calcolo CSN2

**S. Bagnasco, F. Di Pierro, M. Duranti, G. Mazzitelli,
A. Menegolli, M. Punturo**

- il 15/4 abbiamo iniziato al raccolta dei dati attraverso un questionario online via google form
- la deadline era il 30/4
- abbiamo raccolto 31 questionari compilati dai maggiori esperimenti
- risultati preliminari…
- nb. attualmente intorno a maggio inviavano anche una form per raccogliere le richieste di risorse (cpu/storage/tape/licenze) ai vari gruppi, che poi venivano referate dalla prendete commissione a luglio inserite nel db.



CSN2 Computing

giovanni.mazzitelli@gmail.com Switch account

Draft restored

* Required

Email *

giovanni.mazzitelli@lnf.infn.it

Experiment *

CYGNO/INITIUM

Nota per i RN: fate riempire questo foglio ad un esperto di Calcolo. La prima parte può tranquillamente e forse anche possibilmente essere riempita da un esperto della Collaborazione Internazionale

Next

Page 1 of 3

Clear form

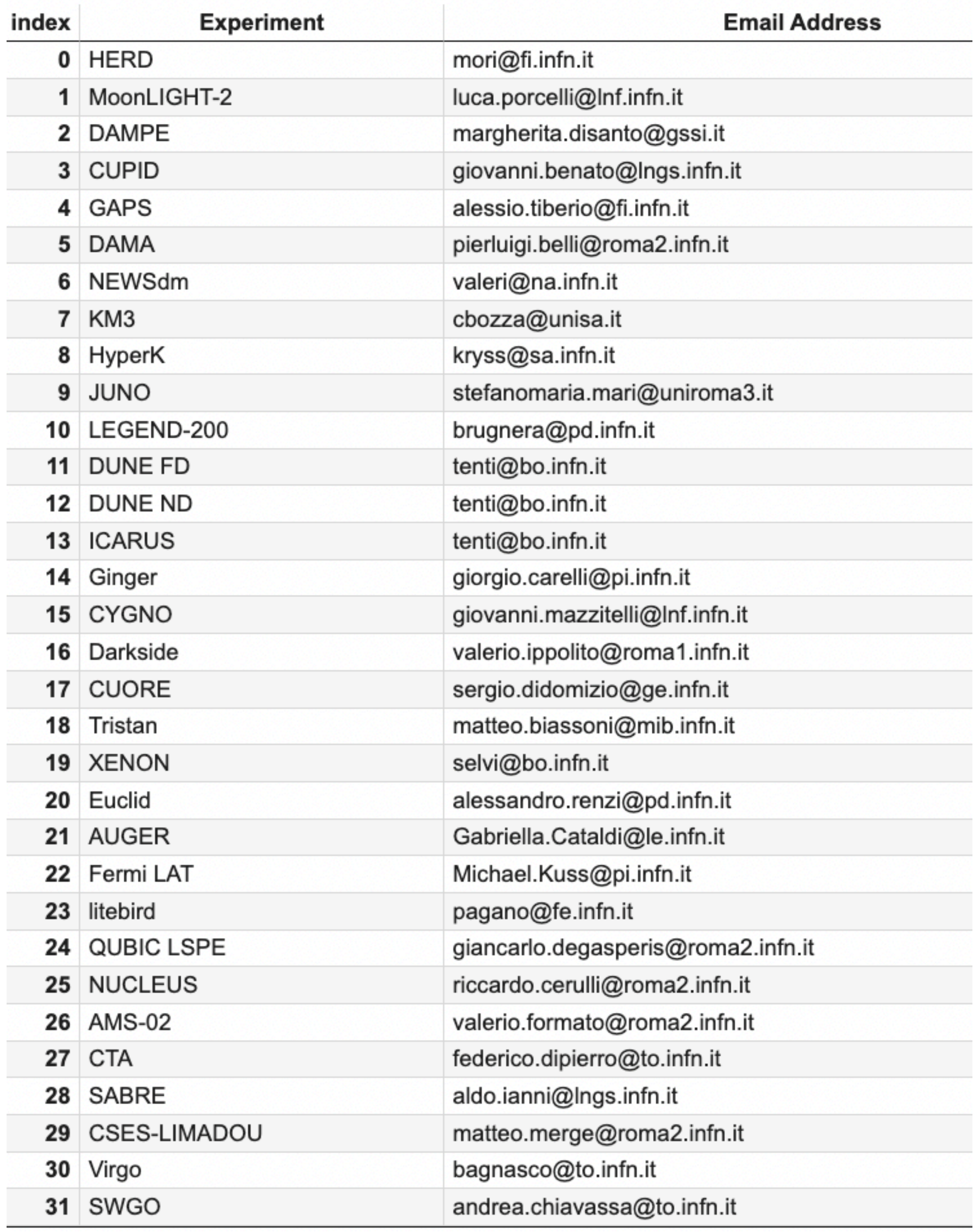
GdL calcolo CSN2

- abbiamo finalmente una lista dei **referti del calcolo** per i vari esperimenti

- abbiamo un database con molte informazioni corrette/confuse e molto “**verbose**”

- per questo motivo l’analisi **quantitativa** non e’ facile perché sarebbe stato necessario fare delle domande più guidate

- il **WP modelli di calcolo dell’INFN** sta partendo dal “survey” della CSN2, imparando da quello che stiamo sottolineando essere stato difficile valutare attraverso quanto chiesto/raccolto.



index	Experiment	Email Address
0	HERD	mori@fi.infn.it
1	MoonLIGHT-2	luca.porcelli@lnf.infn.it
2	DAMPE	margherita.disanto@gssi.it
3	CUPID	giovanni.benato@lngs.infn.it
4	GAPS	alessio.tiberio@fi.infn.it
5	DAMA	pierluigi.belli@roma2.infn.it
6	NEWSdm	valeri@na.infn.it
7	KM3	cbozza@unisa.it
8	HyperK	kryss@sa.infn.it
9	JUNO	stefanomaria.mari@uniroma3.it
10	LEGEND-200	brugnera@pd.infn.it
11	DUNE FD	tenti@bo.infn.it
12	DUNE ND	tenti@bo.infn.it
13	ICARUS	tenti@bo.infn.it
14	Ginger	giorgio.carelli@pi.infn.it
15	CYGNO	giovanni.mazzitelli@lnf.infn.it
16	Darkside	valerio.ippolito@roma1.infn.it
17	CUORE	sergio.didomizio@ge.infn.it
18	Tristan	matteo.biassoni@mib.infn.it
19	XENON	selvi@bo.infn.it
20	Euclid	alessandro.renzi@pd.infn.it
21	AUGER	Gabriella.Cataldi@le.infn.it
22	Fermi LAT	Michael.Kuss@pi.infn.it
23	litebird	pagano@fe.infn.it
24	QUBIC LSPE	giancarlo.degasperis@roma2.infn.it
25	NUCLEUS	riccardo.cerulli@roma2.infn.it
26	AMS-02	valerio.formato@roma2.infn.it
27	CTA	federico.dipierro@to.infn.it
28	SABRE	aldo.ianni@lngs.infn.it
29	CSES-LIMADOU	matteo.merge@roma2.infn.it
30	Virgo	bagnasco@to.infn.it
31	SWGO	andrea.chiavassa@to.infn.it

risultati del survey calcolo CSN2

dati e plot accessibili su https://github.com/gmazzitelli/GSN2/blob/master/rateVsSizeAP.ipynb

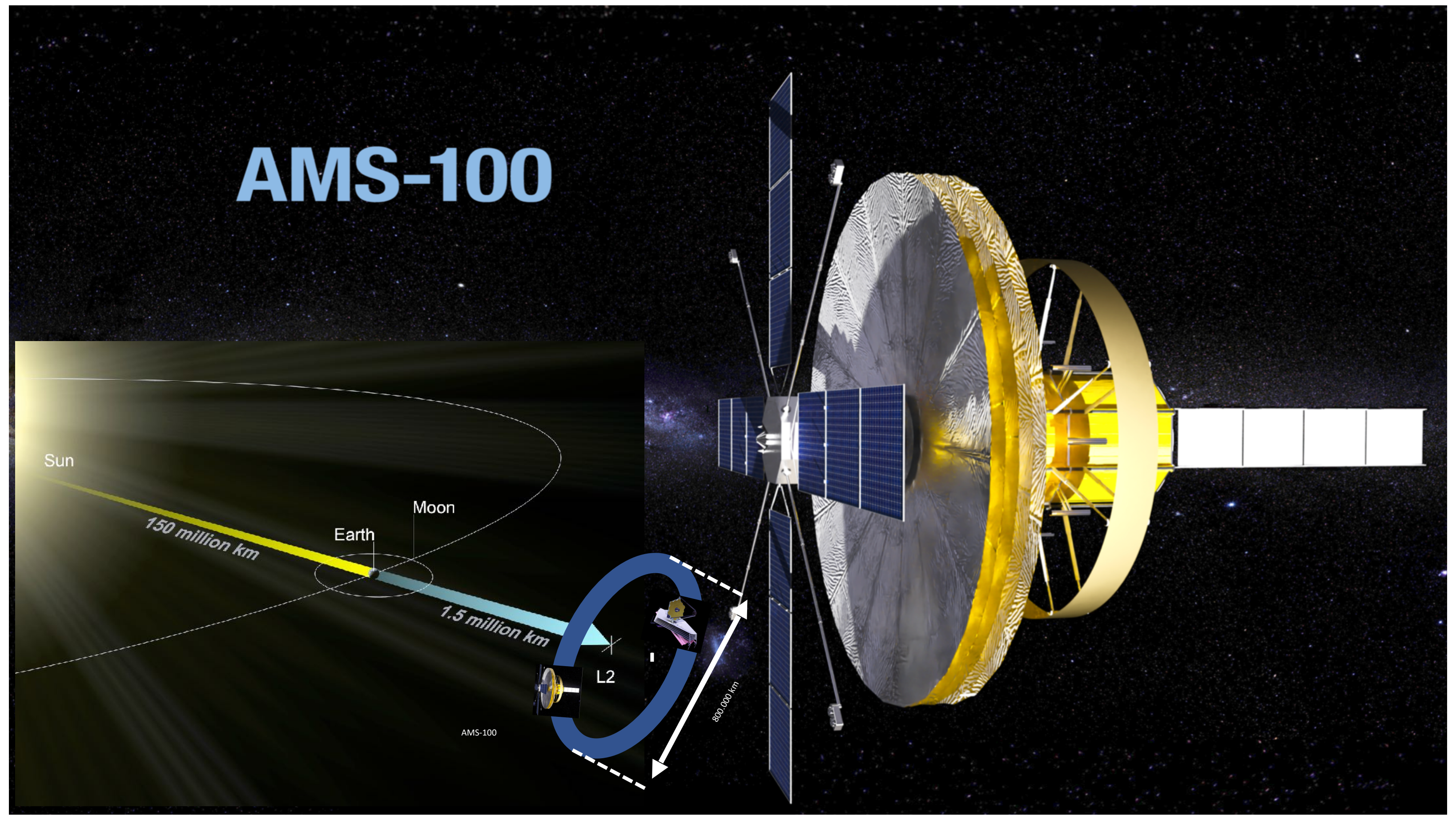
astroparticle data …

commento generale e personale

CSN2
Fisica delle
Astroparticelle

- **unique** and **unrepeatable** data (ex. cosmic events) constraint on uptime/dead-time
- data could be acquired in difficult and **extreme conditions** (ex. space, under water ice, etc) conditioning the possibility of interventions and changes in the setup
- **templates and montecarlo** are needed not only to evaluates systematic but also to identify “candidates” of events. (ex OG, cosmic ray shower, etc) with large request of computing resources
- for many experiment data need to often to be **re-calibrated and reconstructed** many times whit discontinuity and peak in the usage of computing resources

an example of future/futuristic challenge

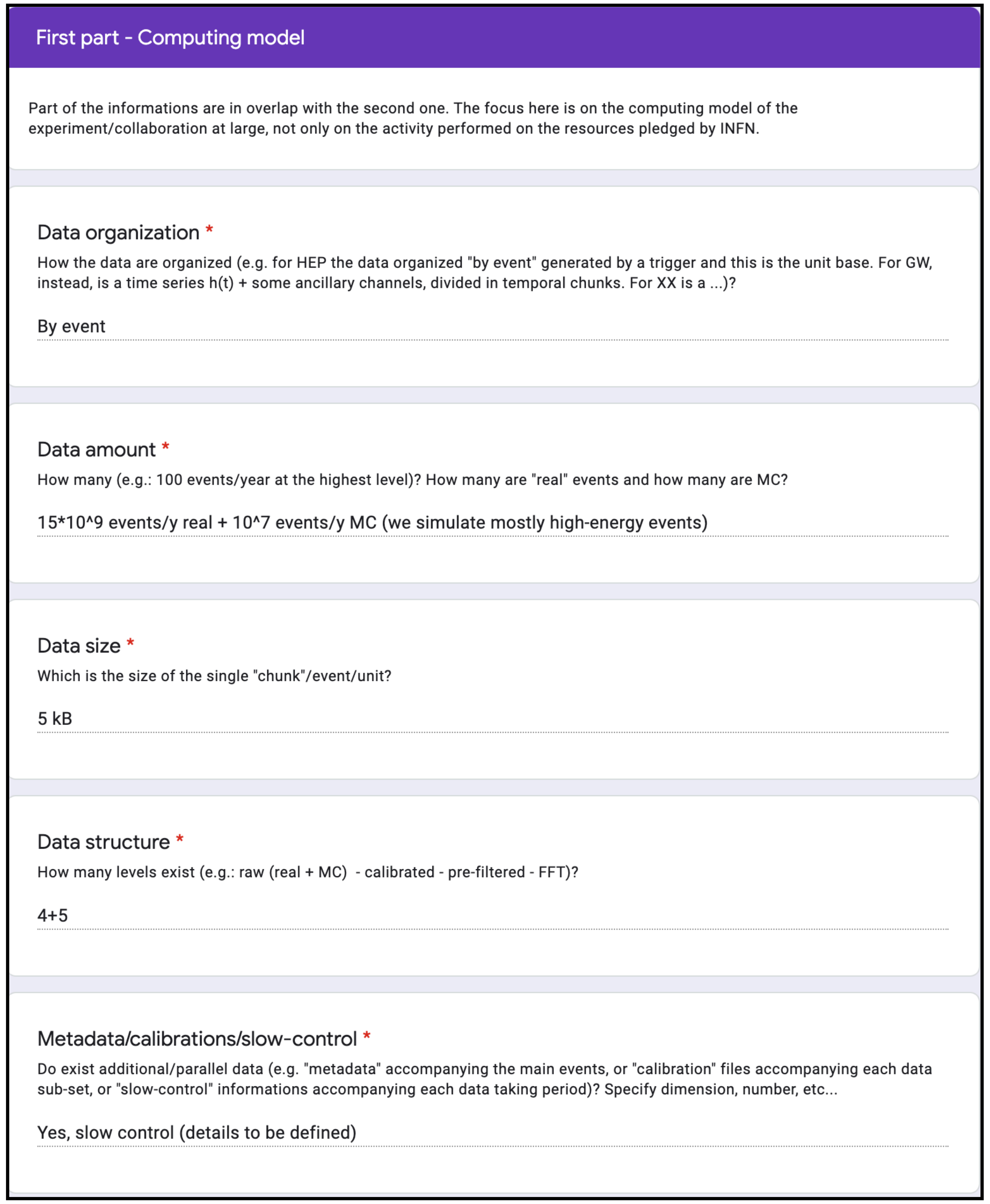


Weight:	40 t
Thin coil Solenoid :	BL2=15 Tm2
Acceptance:	100 m2sr
MDR:	100 TV
Calorimeter:	70 X0, 4λ
Power Consumption:	15 kW
Incoming Particle Rate:	2 MHz
Number Readout Channels:	8 Million
Mission Flight Time:	10 years

survey 1/3

computing model

CSN2 Fisica delle Astroparticelle



First part - Computing model

Part of the informations are in overlap with the second one. The focus here is on the computing model of the experiment/collaboration at large, not only on the activity performed on the resources pledged by INFN.

Data organization *

How the data are organized (e.g. for HEP the data organized "by event" generated by a trigger and this is the unit base. For GW, instead, is a time series h(t) + some ancillary channels, divided in temporal chunks. For XX is a ...)?

By event

Data amount *

How many (e.g.: 100 events/year at the highest level)? How many are "real" events and how many are MC?

15*10^9 events/y real + 10^7 events/y MC (we simulate mostly high-energy events)

Data size *

Which is the size of the single "chunk"/event/unit?

5 kB

Data structure *

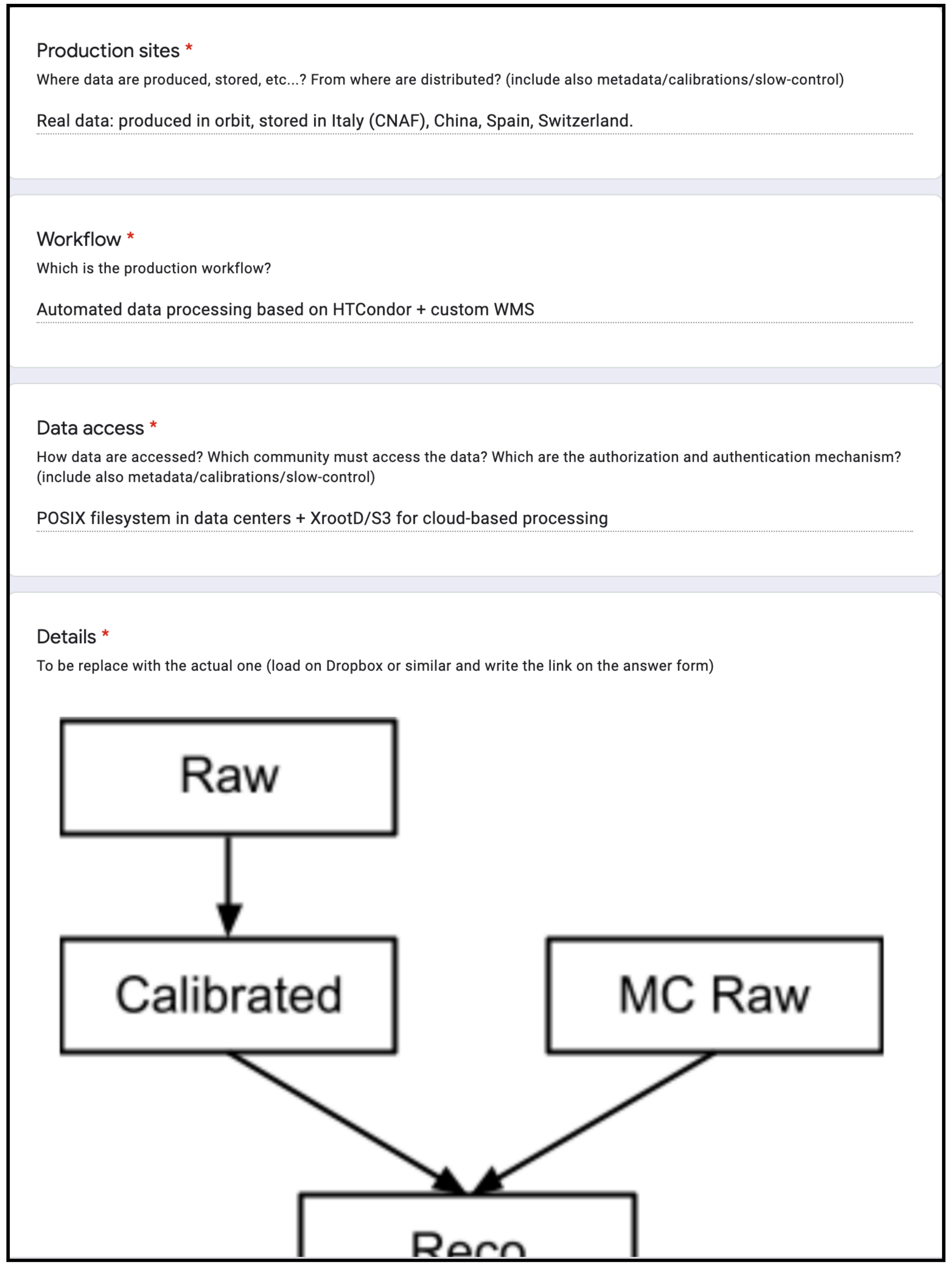
How many levels exist (e.g.: raw (real + MC) - calibrated - pre-filtered - FFT)?

4+5

Metadata/calibrations/slow-control *

Do exist additional/parallel data (e.g. "metadata" accompanying the main events, or "calibration" files accompanying each data sub-set, or "slow-control" informations accompanying each data taking period)? Specify dimension, number, etc...

Yes, slow control (details to be defined)



Production sites *

Where data are produced, stored, etc...? From where are distributed? (include also metadata/calibrations/slow-control)

Real data: produced in orbit, stored in Italy (CNAF), China, Spain, Switzerland.

Workflow *

Which is the production workflow?

Automated data processing based on HTCondor + custom WMS

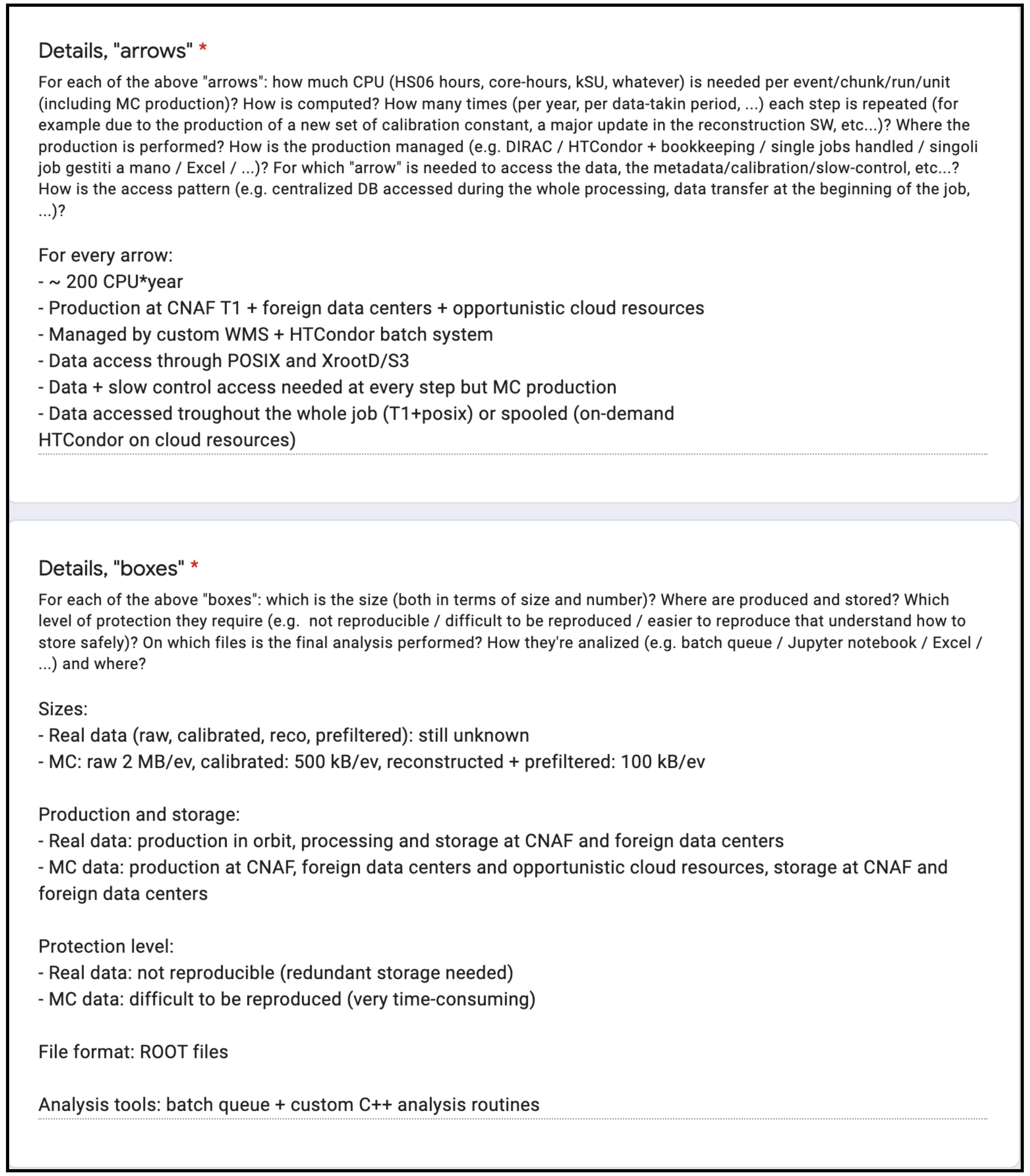
Data access *

How data are accessed? Which community must access the data? Which are the authorization and authentication mechanism? (include also metadata/calibrations/slow-control)

POSIX filesystem in data centers + XrootD/S3 for cloud-based processing

Details *

To be replace with the actual one (load on Dropbox or similar and write the link on the answer form)



Details, "arrows" *

For each of the above "arrows": how much CPU (HS06 hours, core-hours, kSU, whatever) is needed per event/chunk/run/unit (including MC production)? How is computed? How many times (per year, per data-takin period, ...) each step is repeated (for example due to the production of a new set of calibration constant, a major update in the reconstruction SW, etc...)? Where the production is performed? How is the production managed (e.g. DIRAC / HTCondor + bookkeeping / single jobs handled / singoli job gestiti a mano / Excel / ...)? For which "arrow" is needed to access the data, the metadata/calibration/slow-control, etc...? How is the access pattern (e.g. centralized DB accessed during the whole processing, data transfer at the beginning of the job, ...)?

For every arrow:
- ~ 200 CPU*year
- Production at CNAF T1 + foreign data centers + opportunistic cloud resources
- Managed by custom WMS + HTCondor batch system
- Data access through POSIX and XrootD/S3
- Data + slow control access needed at every step but MC production
- Data accessed troughout the whole job (T1+posix) or spooled (on-demand HTCondor on cloud resources)

Details, "boxes" *

For each of the above "boxes": which is the size (both in terms of size and number)? Where are produced and stored? Which level of protection they require (e.g. not reproducible / difficult to be reproduced / easier to reproduce that understand how to store safely)? On which files is the final analysis performed? How they're analized (e.g. batch queue / Jupyter notebook / Excel / ...) and where?

Sizes:
- Real data (raw, calibrated, reco, prefiltered): still unknown
- MC: raw 2 MB/ev, calibrated: 500 kB/ev, reconstructed + prefiltered: 100 kB/ev

Production and storage:
- Real data: production in orbit, processing and storage at CNAF and foreign data centers
- MC data: production at CNAF, foreign data centers and opportunistic cloud resources, storage at CNAF and foreign data centers

Protection level:
- Real data: not reproducible (redundant storage needed)
- MC data: difficult to be reproduced (very time-consuming)

File format: ROOT files

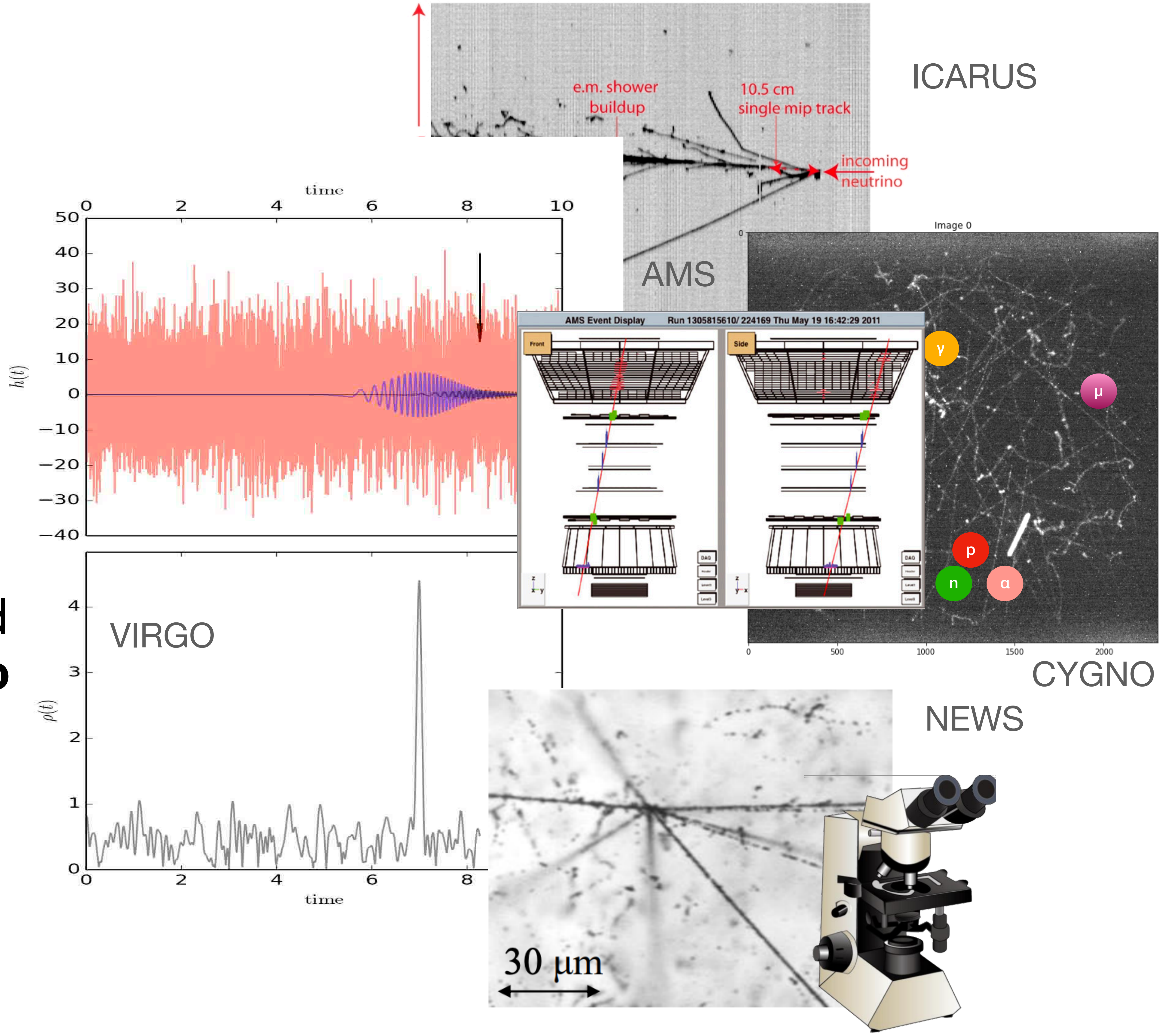
Analysis tools: batch queue + custom C++ analysis routines

data organisation

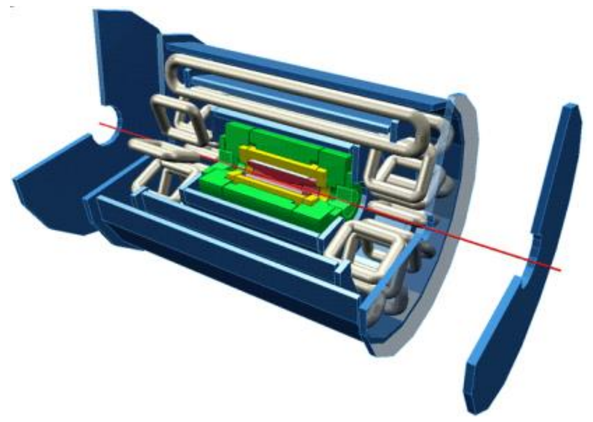
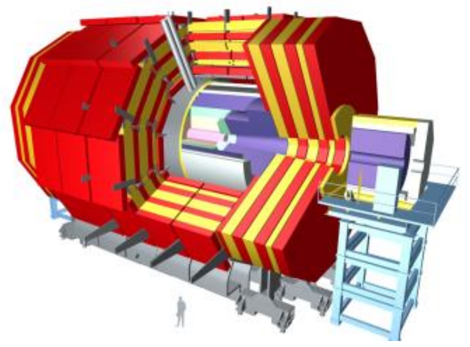
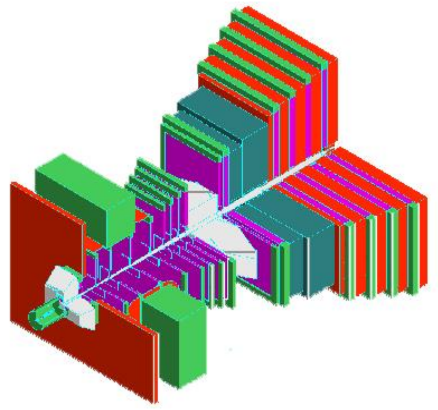
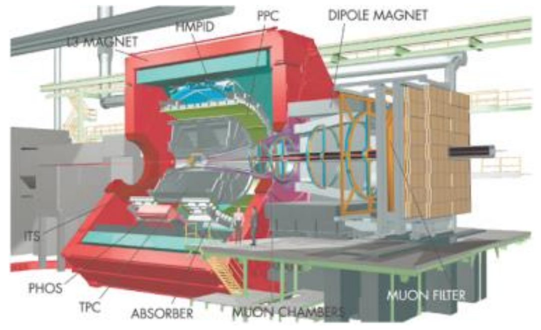
How the data are organized (e.g. for HEP the data organized "by event" generated by a trigger and this is the unit base. For GW, instead, is a time series h(t) + some ancillary channels, divided in temporal chunks. For XX is a ...)?

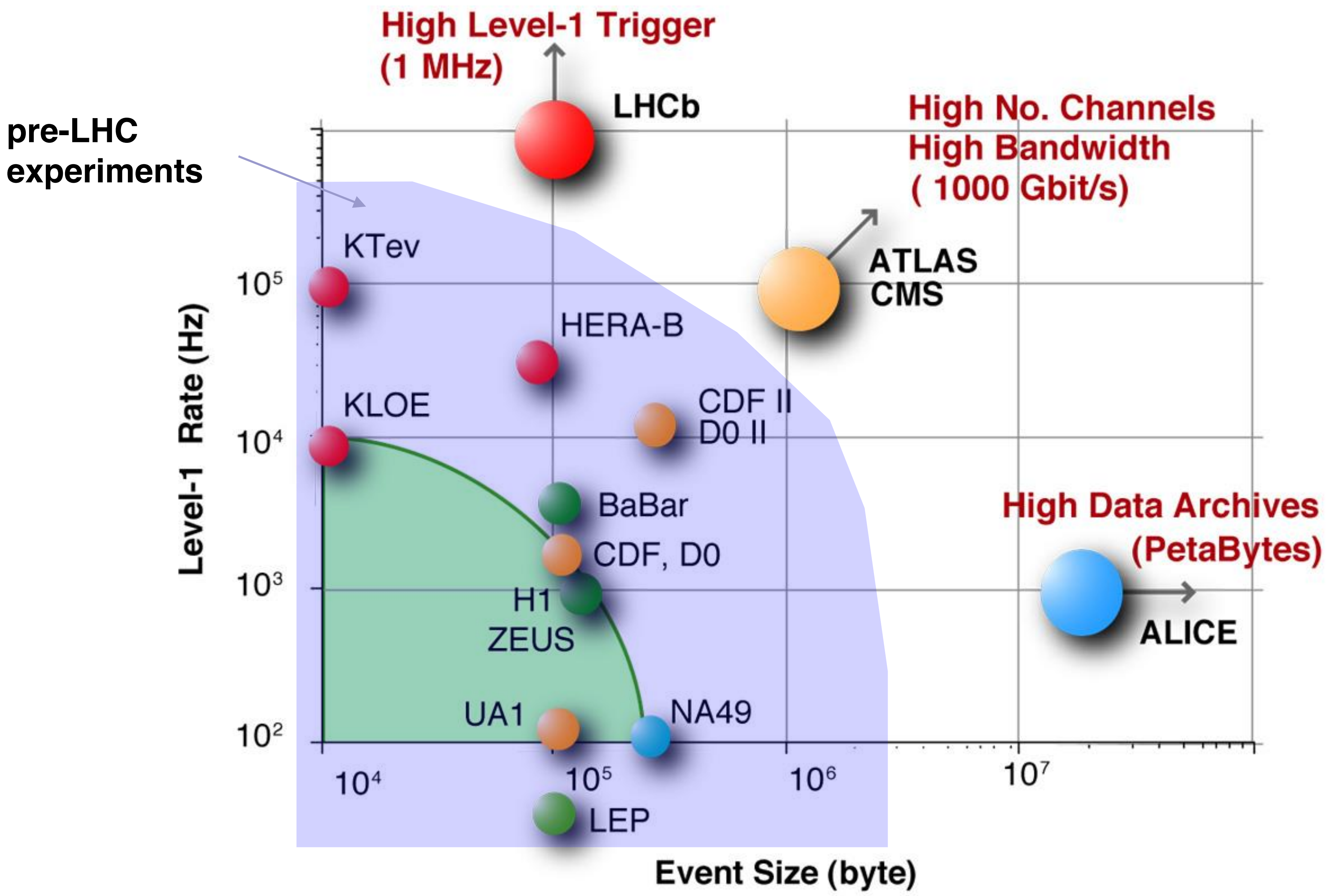
CSN2 Fisica delle Astroparticelle

- **triggered**: a logic (hw or sw) requirements enable data collection: **a trigger acquire one event**, signal candidate are selected (eg: cosmic ray detectors, space experiments, neutrino experiments)
- **post triggered**: a logic (hw or sw) requirements enable to acquire many information: **a trigger acquire many events**, a post processing is needed to extract the signal/noise (eg: DM experiments)
- **continuous streaming**: continues time series post processed to extract data, an **elaborated post processing is needed to extract events** (eg: double beta, GW experiments)
- **post processed**: the **detector it self "store" data** and and a complex procedure extract events (eg: nuclear emulsion)



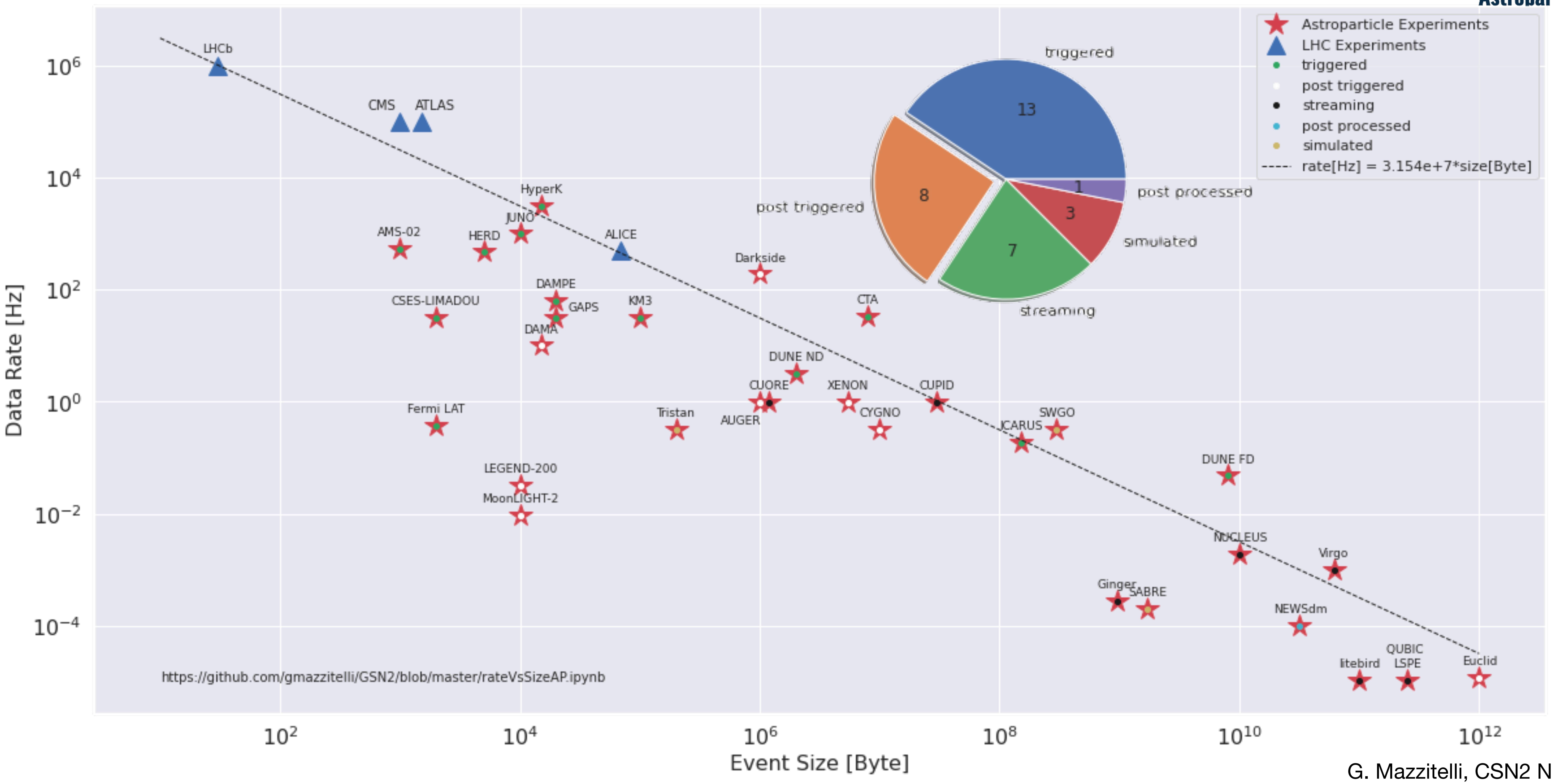
LHC experiments

	No.Levels	Lvl 0,1,2	Event	Evt Build.	HLT Out (design)
	Trigger	Rate (Hz)	Size (Byte)	Bandw.(GB/s)	MB/s (Event/s)
	3	LV-1 10^5 LV-2 $6x10^3$	1.5 MB	5.25	300 (200)
	2	LV-1 10^5	1.0 MB Pb-Pb 1500MB/s	100	300 (200)
	2	LV-0 10^6	30 kB	40	60 (2 kHz)
	4	Pb-Pb 500	70 MB	2	1250 (100)



H. Sakulin, 2014 https://indico.cern.ch/event/282492/contributions/1631039/attachments/520206/717680/2014_03_25_DAQ_HS.pdf

data amount vs data size



https://github.com/gmazzitelli/GSN2/blob/master/rateVsSizeAP.ipynb

G. Mazzitelli, CSN2 Novembre 2022

data structure and metadata/calibrations/slow-control

CSN2
Fisica delle
Astroparticelle

How many levels exist (e.g.: raw (real + MC) - calibrated - pre-filtered - FFT)?
Do exist additional/parallel data (e.g. "metadata" accompanying the main events, or "calibration" files accompanying each data sub-set, or "slow-control" informations accompanying each data taking period)? Specify dimension, number, etc…?

- il minimo di livelli riscontrati e' 3 (raw—>reco—>analisi) ma e' probabile che si siano dimenticati almeno la **calibrazione**, e inoltre molti di noi non hanno riportato le **pipeline della simulazione**

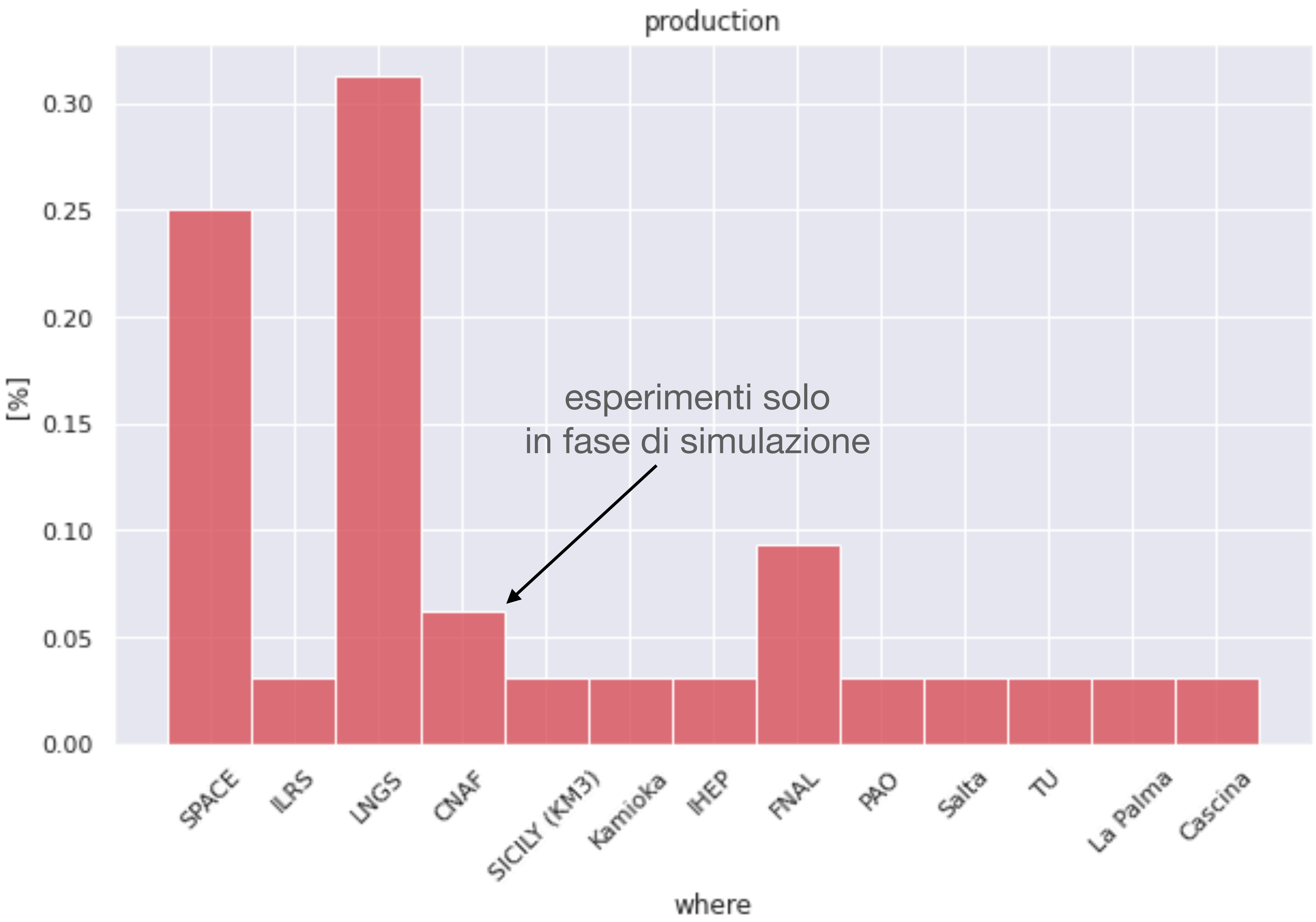
- tipicamente si arriva **intorno 5/6 livelli** di trattazione dei dati

- alcuni esempi delle riposte per sottolineare anche una difficolta' di **comunicazione** nella quale non abbiamo una base di termini comuni.
  - **VIRGO**: Full bandwidth raw (not saved); Bulk raw (downsampled signal + control channels); Reduced Data Sets (very downsampled raw); aggregated h(t), calibrated and uncalibrated, in several flavours (Frame File format); derived data (usually FFT of aggregated h(t))
  - **AMS-02**: 3 + 4
  - **CTA**: RAW, DL0, DL1, DL2, DL3, MC
  - **Darkside**: raw, calibrated, analysis ntuple (to be finalised by computing TDR submission)
  - **CUORE**: Real data: raw (continuous) -> denoised (continuous) -> triggered -> processed MC: Geant4 output files, negligible size

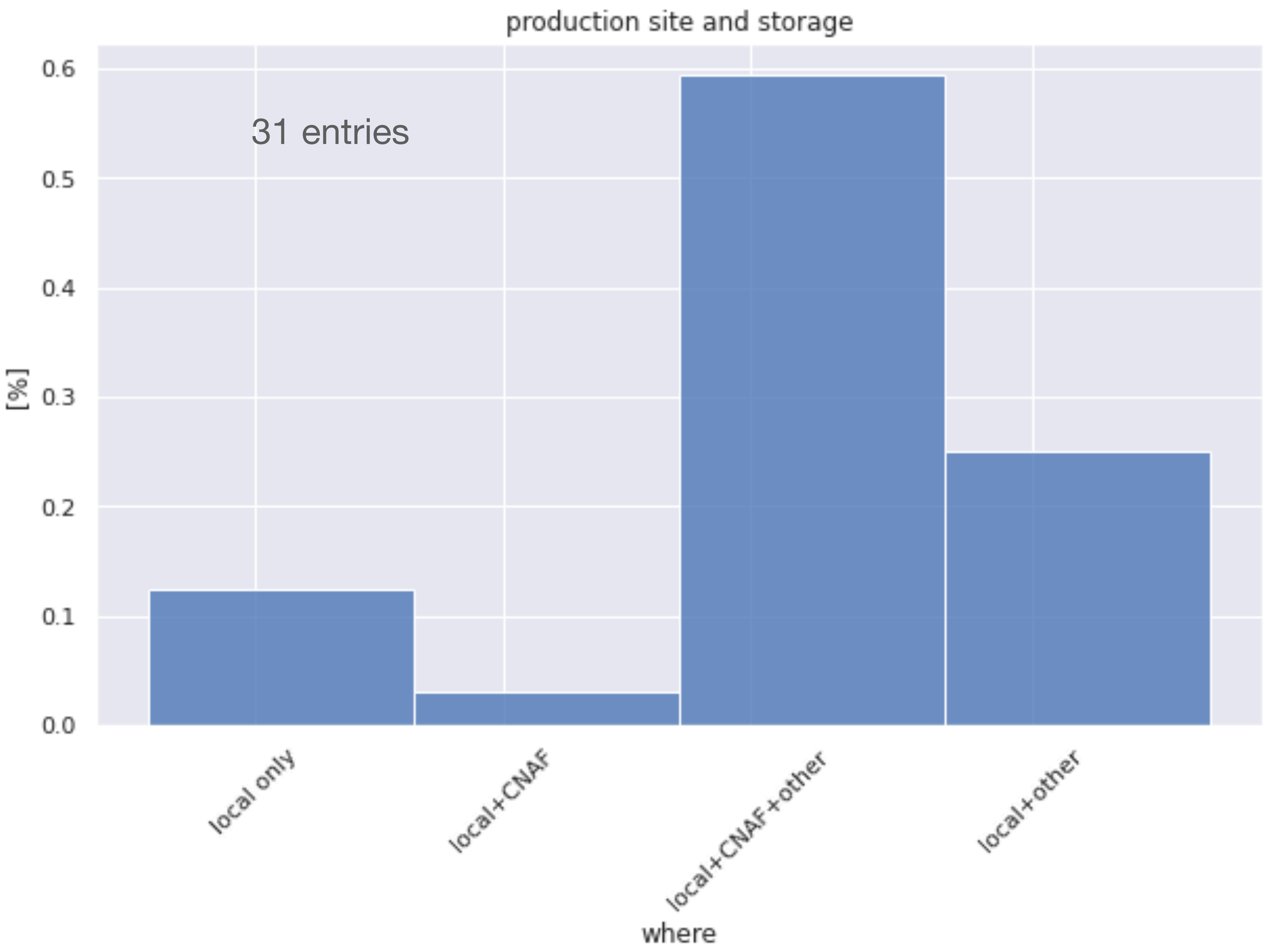
- la maggior parte degli esperimenti ricorre a **metadati** e dati di **calibrazione** attraverso DB relazionali dove sono salvati anche i le infrazioni ausiliarie degli **slow control**

production sites

Where data are produced, stored, etc…? From where are distributed? (include also metadata/calibrations/slow-control)



esperimenti solo in fase di simulazione



31 entries

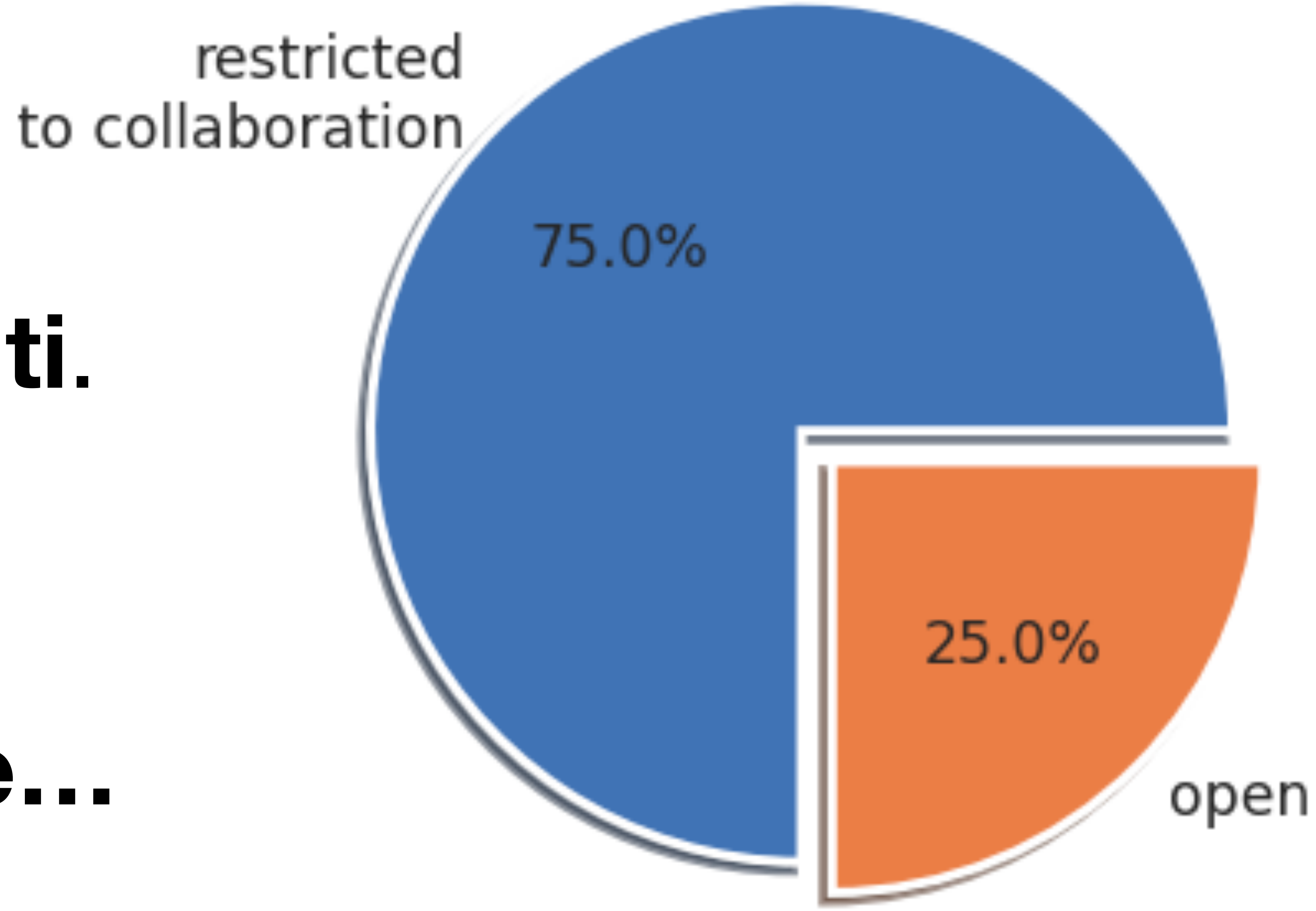
- il > 65% degli esperimenti hanno dati in copia al CNAF o infrastrutture italiane; un ~ 12% ha solo copie locali; ~ 25 % non sembra usare il CNAF o infrastrutture italiane

data access

How data are accessed? Which community must access the data? Which are the authorization and authentication mechanism? (include also metadata/calibrations/slow-control)

CSN2 Fisica delle Astroparticelle

- solo il 25 % dei dati e' accessibile al di fuori della collaborazione (web site) in modo **open**
- di questi solo un paio danno accesso anche ai raw, altrimenti si parla sempre di **dati ricostruiti**.
- la maggior parte e' accessibile solo via **SSH**!!!
- alcuni (<20%) usano **XrootD/DIRAC/LCG-like…**



workflow (qualche esempio)

dati e plot accessibili su https://github.com/gmazzitelli/GSN2/blob/master/rateVsSizeAP.ipynb

AUGER

The Pierre Auger Observatory

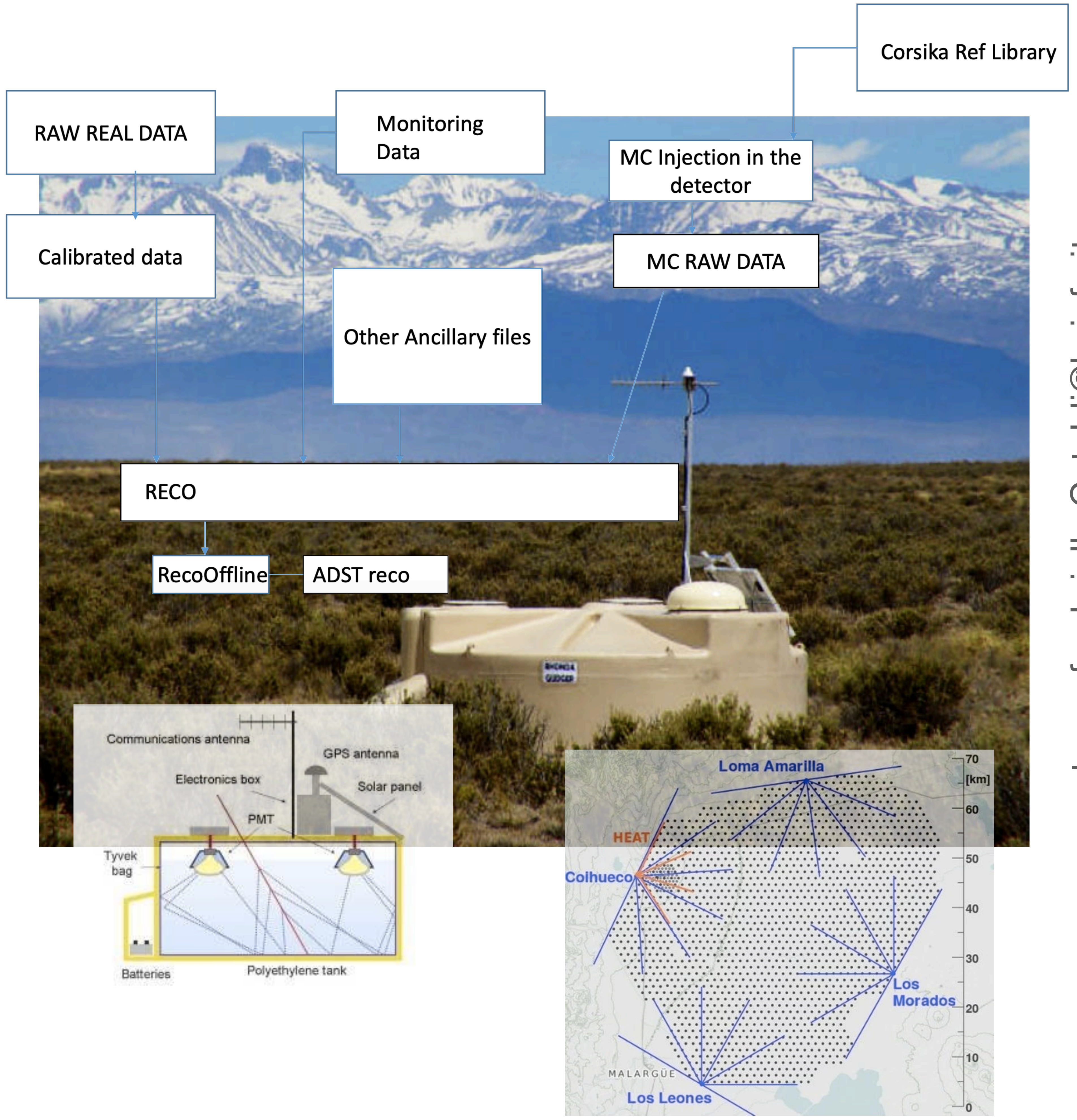
- The Pierre Auger Observatory, Pampa Amarilla in western Argentina, devoted to studies the highest-energy cosmic rays, in operation since **2008** with an exposure ~ 40,000 km^2 sr yr.

- **raw** triggered data have a modest size of **1MB/event**, with a rate ~ 1Hz; All data stored since 2008 are about 50 TB: raw, monitoring, and offline, i.e., high-level data (detector position, atmospheric info, etc) typically used in all the physics analyses

- raw data are stored locally in Malargüe and mirrored every 3 hours to the IN2P3 Computing Center in Lyon.

- On the contrary the Monte Carlo simulations is the large ammonite of data **300MB/event** the experimental method of studying ultra-high energy cosmic rays is an indirect one.

- **The Monte Carlo simulation is used to define the characteristics of extensive air showers** (EAS) and to obtained information to infer the properties of the original particle, its energy, type, direction etc.



Corsika Ref Library

RAW REAL DATA

Monitoring Data

MC Injection in the detector

Calibrated data

MC RAW DATA

Other Ancillary files

RECO

RecoOffline

ADST reco

courtesy of gabriella.Cataldi@le.infn.it

CTA

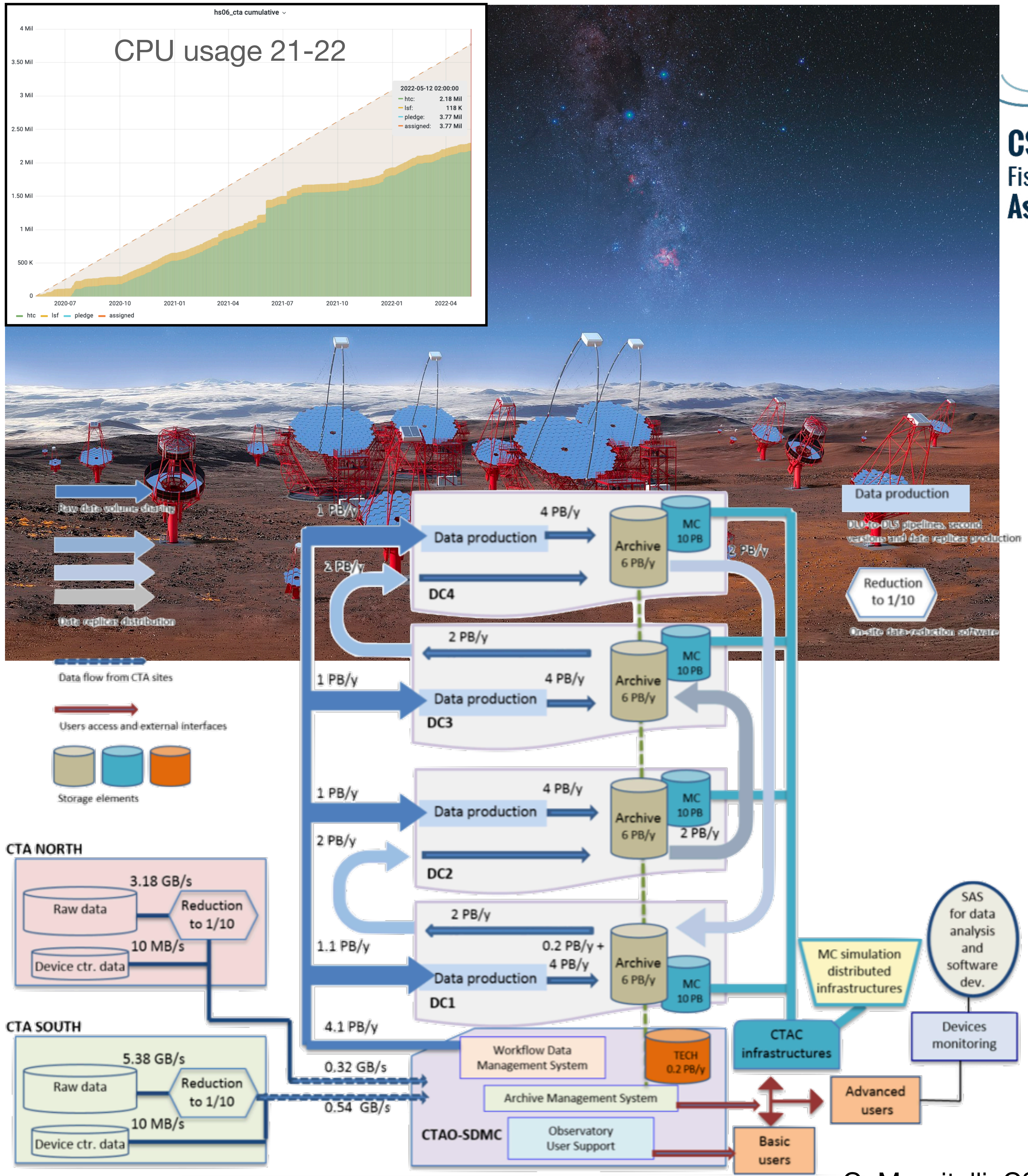
Cherenkov Telescope Array

- The raw event data rate is estimated to **6.3 Gbytes/s** for the CTAO-South and 17.1

- Gbytes/s for the CTAO-North prior to any data volume reduction.

- An additional 20% is added to that accounting for monitoring and service data.

- Data will be taken during 1314 hours observation time per year (corresponding to an annual duty cycle of ~15%) and a maximum of 12 hours of data acquisition per day



courtesy of federico.dipierro@to.infn.it

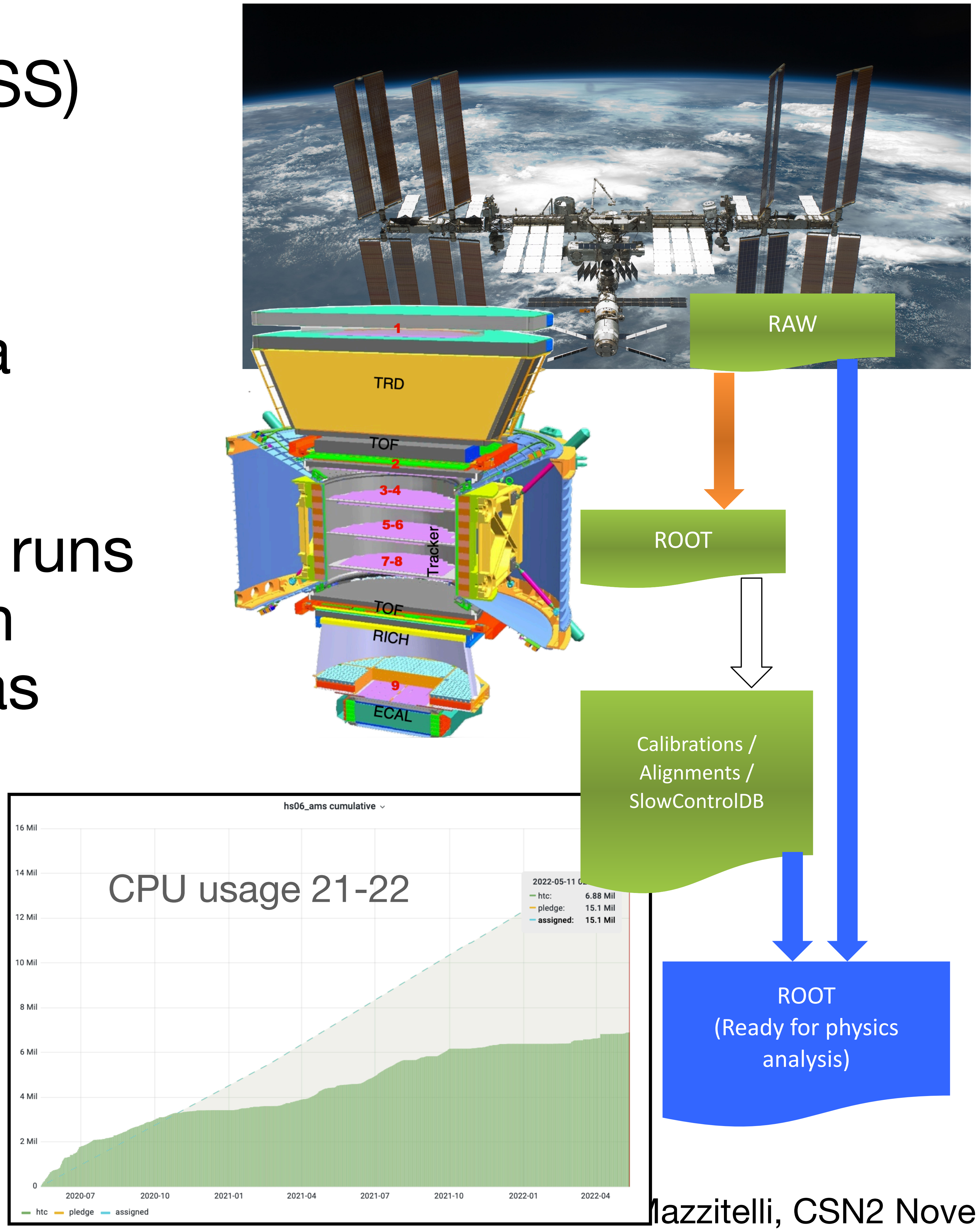
G. Mazzitelli, CSN2 Novembre 2022

AMS-2

Alpha Magnetic Spectrometer

CSN2 Fisica delle Astroparticelle

- experiment mounted on the International Space Station (ISS) to study cosmic radiation, antimatter, dark matter, strangelets…
- **211 billion triggers acquired and 35 TB/year** of raw data between 2011 to 2022 (1.7E10 ev/y)
- **first production**: raw data arrive every 2 hours divided in runs acquired every hours every days, then data calibrated with flight info are available for the second production as well as quick performance evaluation
- **second production**: full reconstruction that take care of calibrations, alignments, ancillary data, etc is produced incrementally **every 6 months**



CPU usage 21-22

courtesy of matteo.duranti@pg.infn.it

DUNE

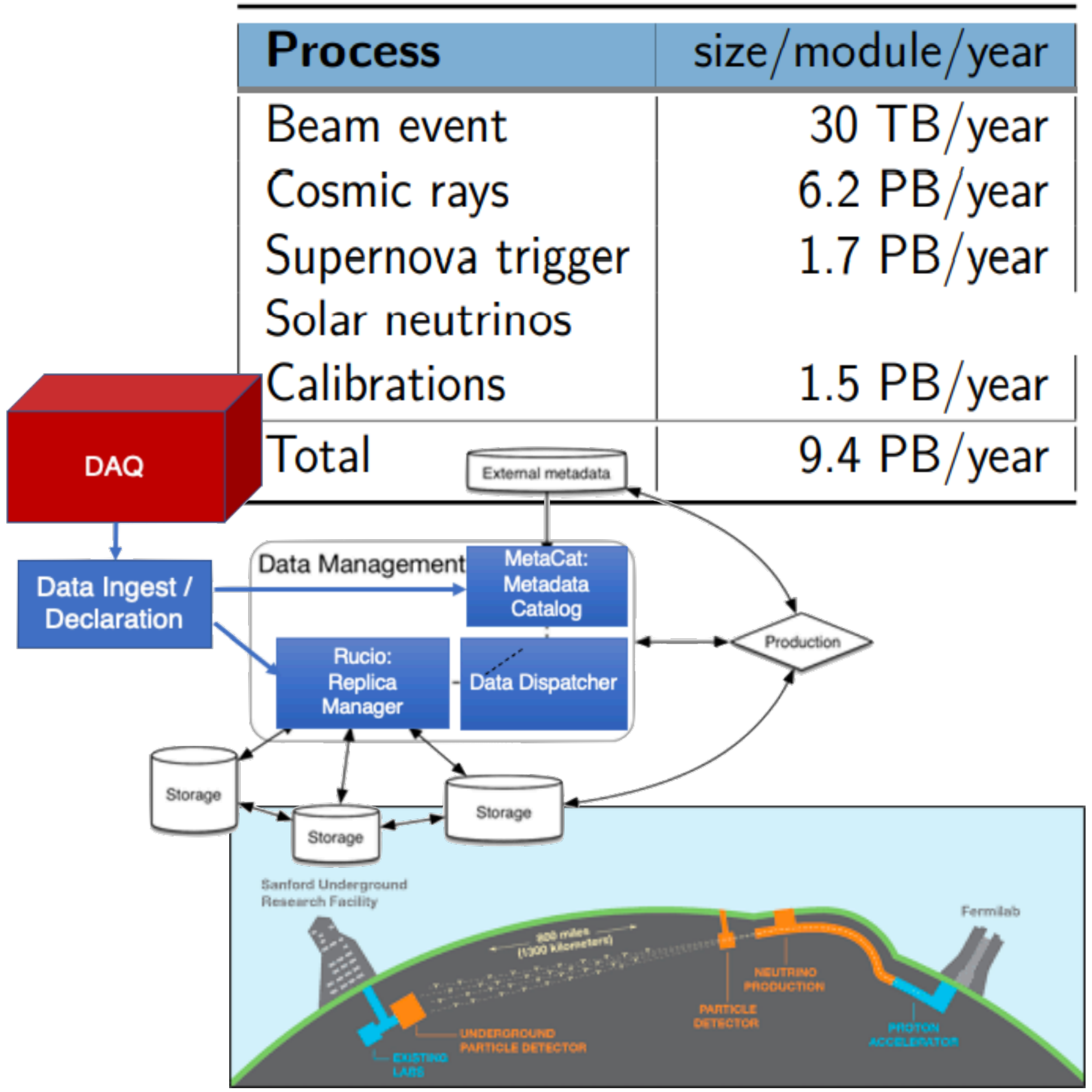
Deep Underground Neutrino Experiment

CSN2 Fisica delle Astroparticelle

neutrino science and proton decay at Fermi National Accelerator Laboratory in Batavia, Illinois FNAL and Sanford Underground Research Laboratory in Lead, South Dakota — 1,300 kilometers form FNAL beam

- **Far Detector (FD):**
  - localized and high-energy: for beam, cosmic and nucleon decay events **6.5 GB/event**
  - Extended and low-energy: for **supernovae burst 150GB/event**
- **Near Detector (ND):**
  - beam flux and monitor, systematics, high precision neutrino physics, etc a **few MB/event**
- **raw data storage FNAL/CERN**
- **computing resources** (CPU and storage) largely contributed by **collaborating institutions**

- Production Operations Management System (POMS) developed and maintained by Fermilab interfaces both with DUNE-dedicated at FNAL and with opportunistic OSG and WLCG resources, and possibly HPC resources within the HEPCloud infrastructure



Process	size/module/year
Beam event	30 TB/year
Cosmic rays	6.2 PB/year
Supernova trigger	1.7 PB/year
Solar neutrinos	
Calibrations	1.5 PB/year
Total	9.4 PB/year

courtesy of matteo.tenti@bo.infn.it

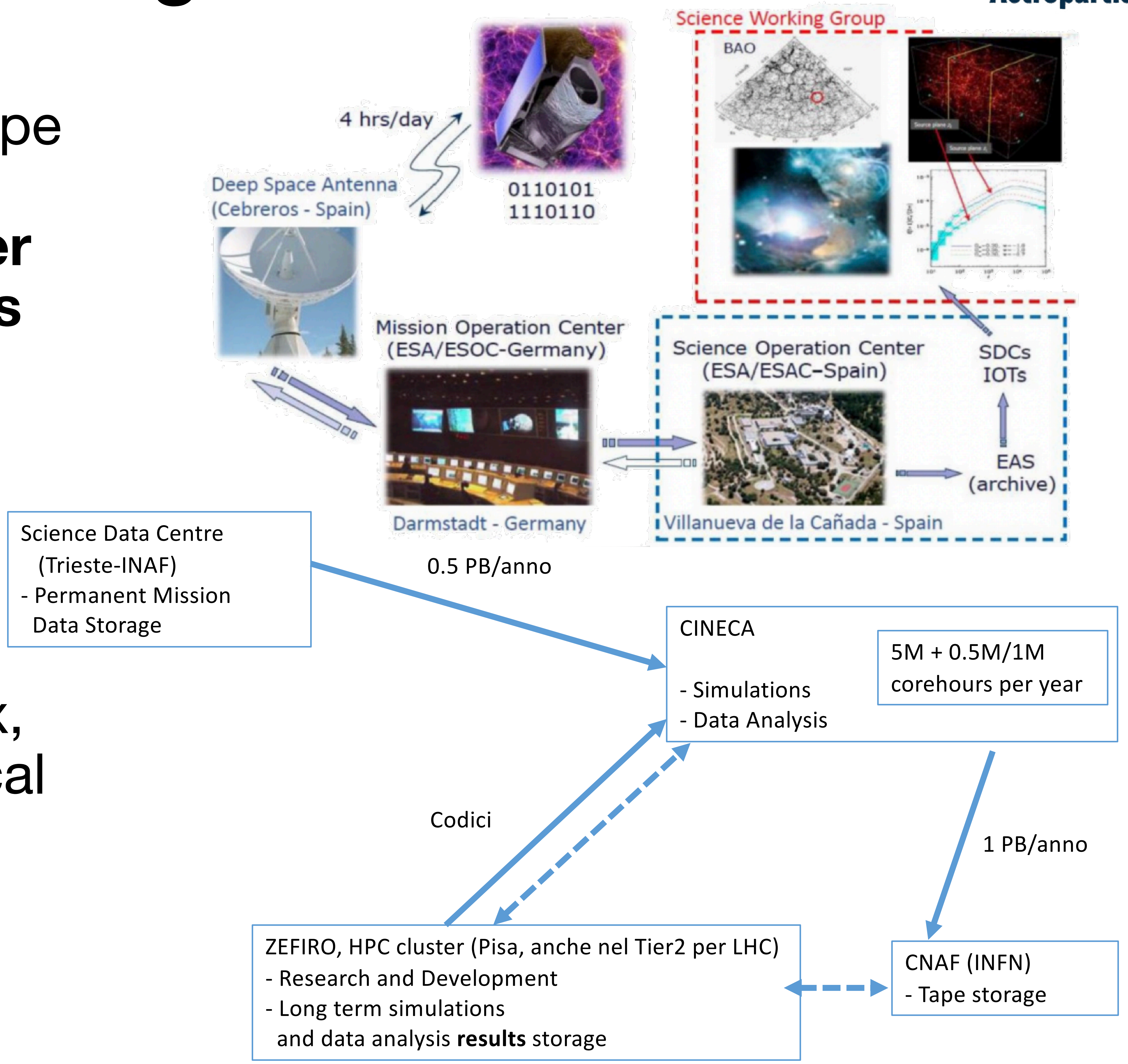
G. Mazzitelli, CSN2 Novembre 2022

CSN2
Fisica delle
Astroparticelle

EUCLID

Euclid and the origin of the accelerating universe

- 1.2 m diameter Silicon Carbide (SiC) mirror telescope made by Airbus (Defence and Space) feeding 2 instruments, high quality **panoramic visible imager** (VIS) and **near infrared, photometer and a slitless spectrograph** (NISP)
- Euclid Consortium: data taking, data reduction, scientific output of NISP e VIS instruments
- 2D/3D images generating large scale structure catalogs of observed objects (position, speed, flux, etc.) to be compere with EuclidN-body cosmological simulations: **catalog + simulation 100PB/mission**
- O(10M) cpu-hours-per-year.
- **HPC** @ ReCaS for prototyping



Science Data Centre
(Trieste-INAF)
- Permanent Mission
Data Storage

0.5 PB/anno

CINECA
- Simulations
- Data Analysis

5M + 0.5M/1M
corehours per year

Codici

1 PB/anno

ZEFIRO, HPC cluster (Pisa, anche nel Tier2 per LHC)
- Research and Development
- Long term simulations
and data analysis **results** storage

CNAF (INFN)
- Tape storage

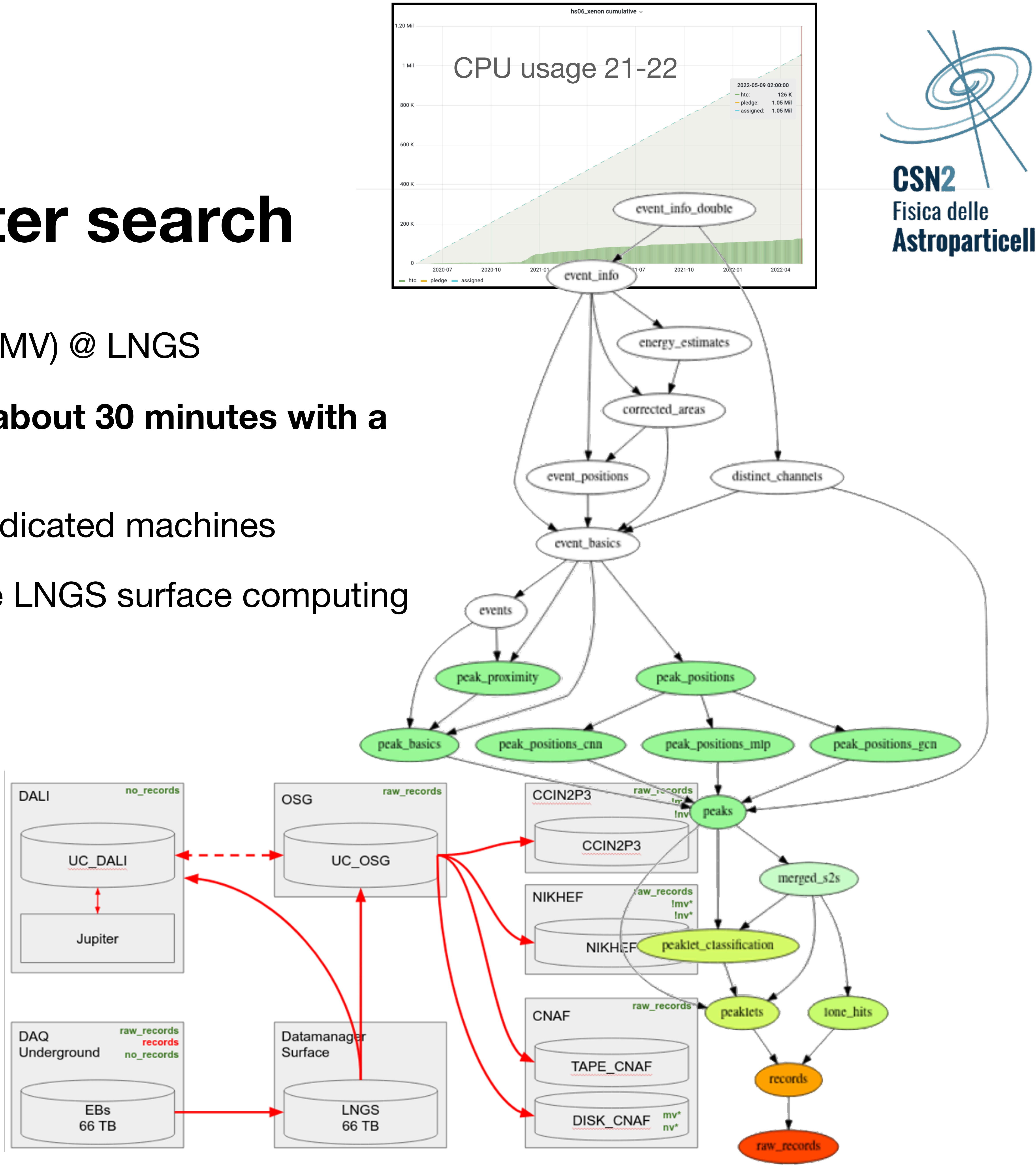
courtesy of alessandro.renzi@pd.infn.it

XENON

Liquid Xenon TPC for Dark Matter search

CSN2 Fisica delle Astroparticelle

- **datasets**: Liquid Xenon TPC, Neutron Veto (NV), Muon Veto (MV) @ LNGS
- raw data are divided in chunks of a few GB costing a **run of about 30 minutes with a typical size of 10GB** (100GB during calibration)
- All data from DAQ are processed and monitored online by dedicated machines
- raw and processed data are buffered 66TB disk hosted in the LNGS surface computing center
- Then sent through **GRID on several external sites**:
  - **CNAF** (Italy): Tape storage for a full backup of raw data, plus a disk storage to keep all NV and MV data only
  - **CC-IN2P3** (France) and NIKHEF (Netherland): copy of TPC raw data
  - **UC_OSG** (Uchicago) and SD_OSG (U San Diego): second copy of all raw data
  - **DALI** (UChicago): analysis hub for all users


CPU usage 21-22

courtesy of scotto@lpnhe.in2p3.fr

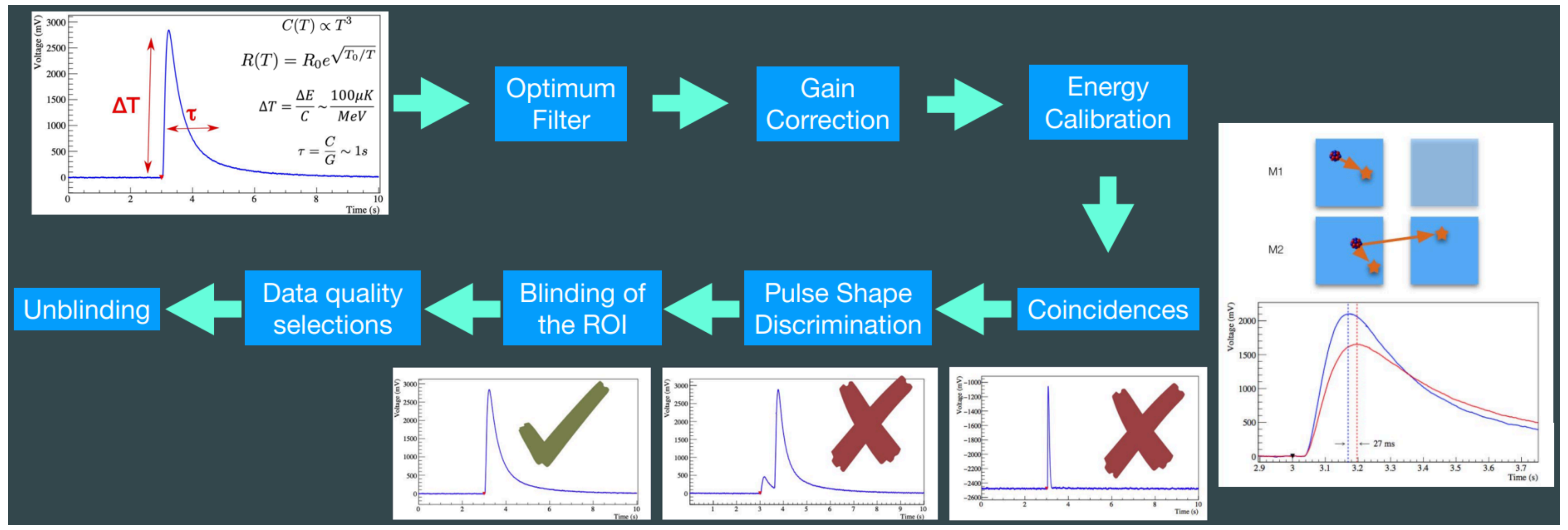
CUORE/CUPID

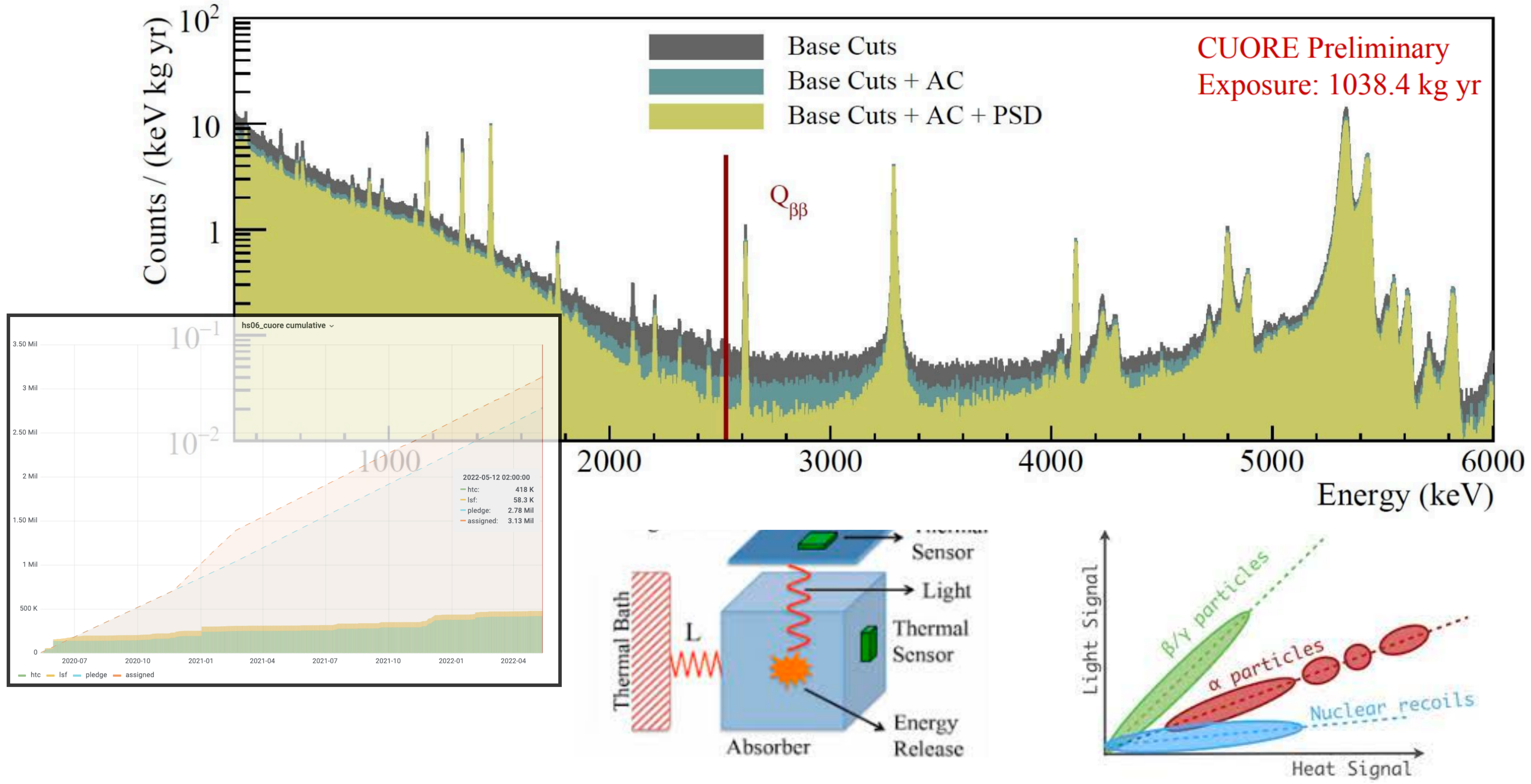
search for neutrinoless double beta decay

- **CUORE** is a ton scale bolometric experiment for the search of neutrinoless **double beta decay**; 1Hz data stream of O(10^3) channels 200 MB/day

- **CUPID** (CUORE Upgrade with Particle Identification) **scintillating bolometers** able to discriminate better discriminate background; 1Hz data stream, 1TB/day

- events **reconstruction** flow consists in two steps: 1) event-based **quantities** are evaluated (pulse amplitude estimation, detector gain correction, energy calibration,) 2) **energy** spectra are produced. Event candidates - pulse-shape parameters - are **coincidence** among multiple bolometers





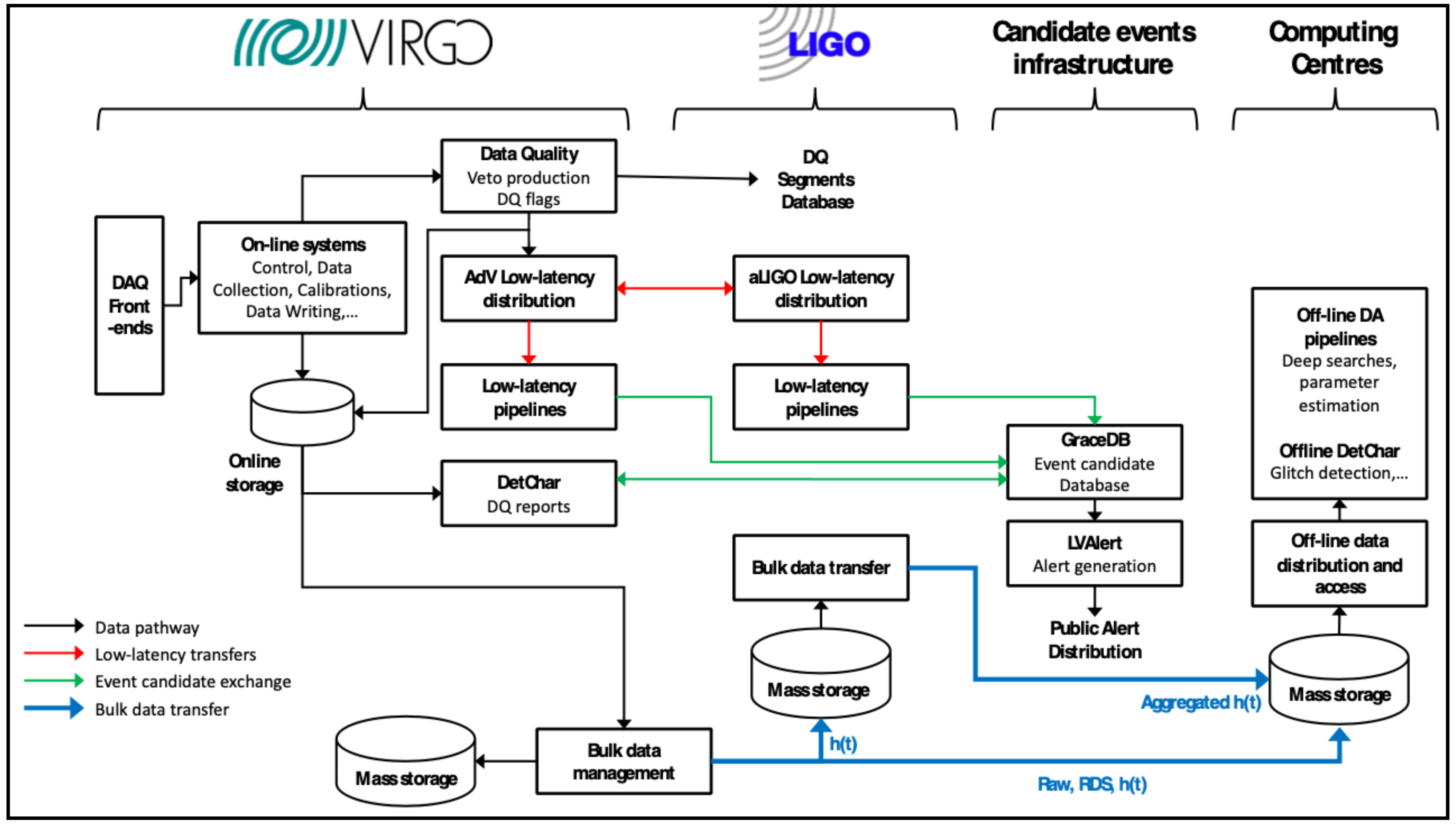
1500 +1500 thermal + light signal

courtesy of sergio.didomizio@ge.infn.it
giovanni.benato@lngs.infn.it

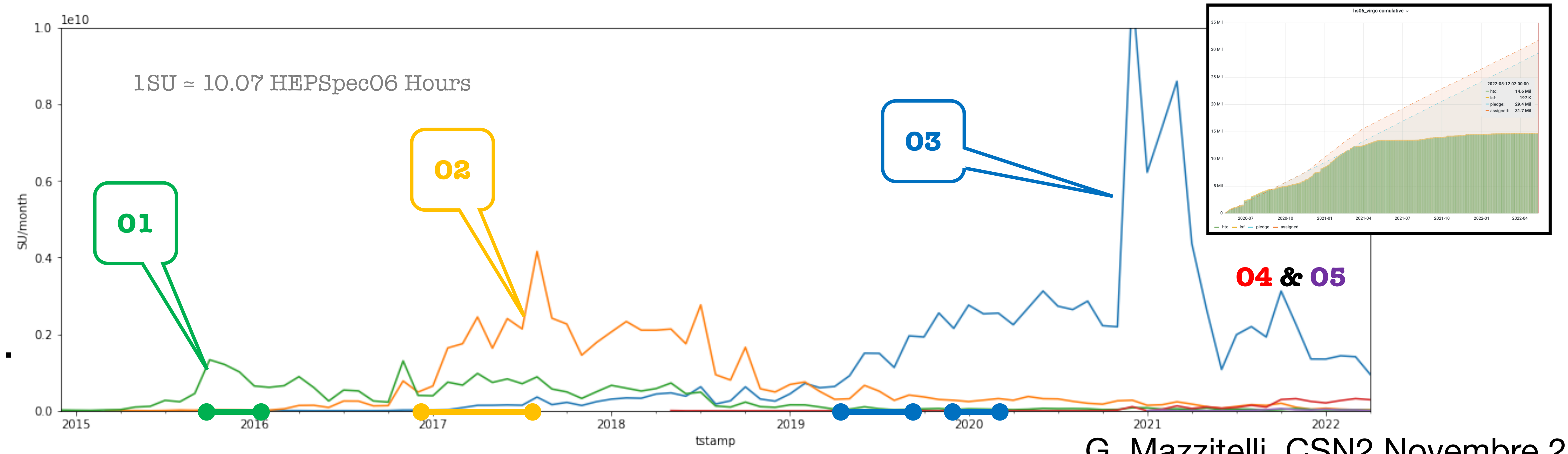
VIRGO

large data streaming

- Raw data stored in circular buffer at EGO and shipped to custodial centres (CNAF and CC-IN2P3): **1-2 PB/year**
- **LIGO and Virgo exchange low-latency h(t)** data; LL searches are run on dedicated EGO resources and the **multi-messenger alert** system is managed by LIGO
- Aggregated h(t) data for offline analysis copied to **CNAF and CC-IN2P3,** distributed to all centres through CVMFS/StashCache: 1 kS frame files, 3-4 TB/year/detector
- **Offline data processing** (searches, PE, noise glitch detection,…) either via local jobs or grid
    - CNAF, CC-IN2P3, NIKHEF/SurfSARA, PIC, UCLouvain, Wigner,…
    - Grid currently about 10%
    - Overall projection for O4: o(11 MHS06 hours) incl. LL, 1.5×WRT 03


VIRGO
LIGO
Candidate events infrastructure
Computing Centres
DAQ Front -ends
On-line systems Control, Data Collection, Calibrations, Data Writing,...
Data Quality Veto production DQ flags
DQ Segments Database
AdV Low-latency distribution
aLIGO Low-latency distribution
Low-latency pipelines
Low-latency pipelines
Online storage
GraceDB Event candidate Database
DetChar DQ reports
LVAlert Alert generation
Public Alert Distribution
Off-line DA pipelines Deep searches, parameter estimation
Offline DetChar Glitch detection,...
Off-line data distribution and access
Bulk data transfer
Mass storage
Mass storage
Mass storage
Bulk data management
h(t)
Aggregated h(t)
Raw, RDS, h(t)
Data pathway
Low-latency transfers
Event candidate exchange
Bulk data transfer

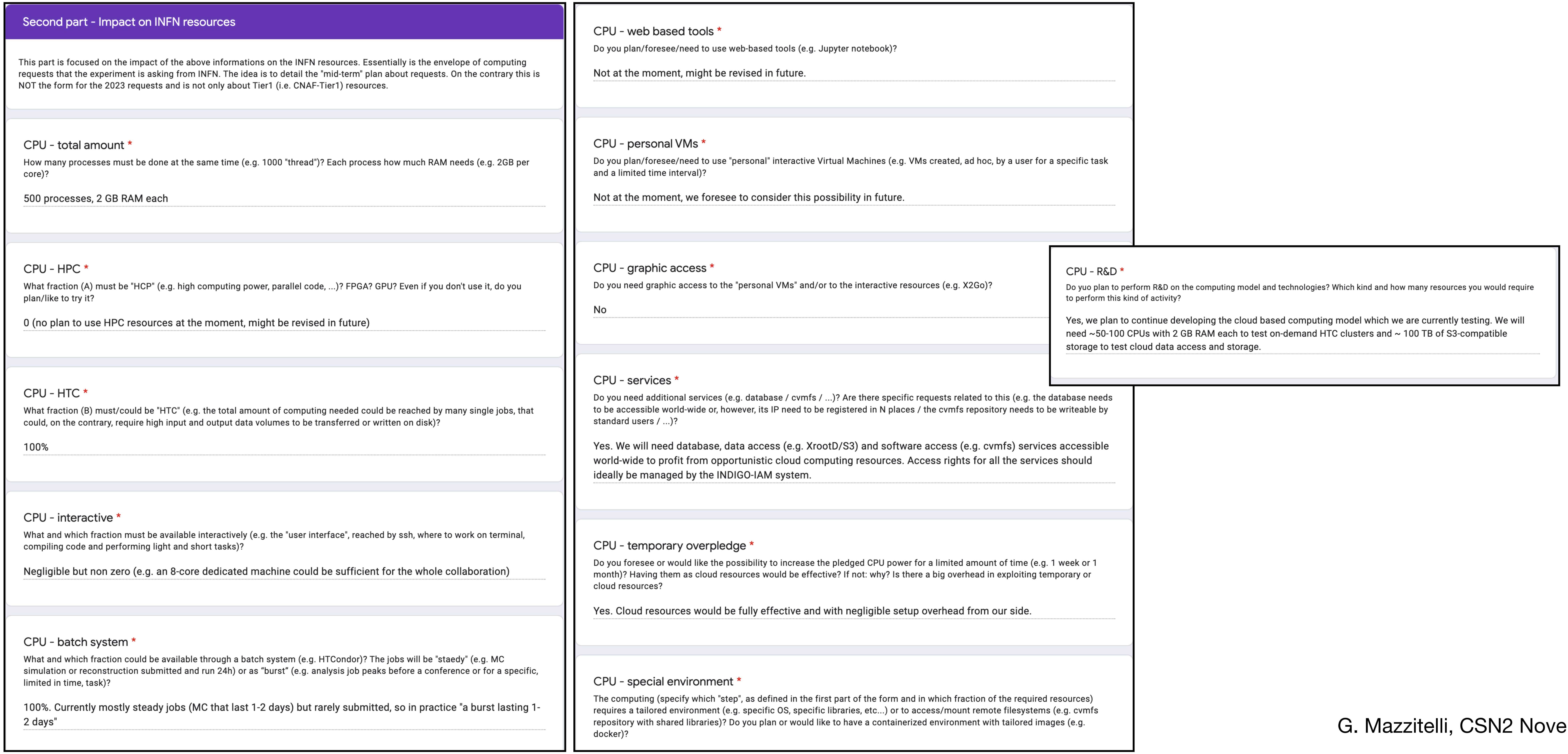
courtesy of bagnasco@to.infn.it



Survey 2/3

impact on INFN resources (identify the "mid-term" plan about requests)

CSN2 Fisica delle Astroparticelle



Second part - Impact on INFN resources

This part is focused on the impact of the above informations on the INFN resources. Essentially is the envelope of computing requests that the experiment is asking from INFN. The idea is to detail the "mid-term" plan about requests. On the contrary this is NOT the form for the 2023 requests and is not only about Tier1 (i.e. CNAF-Tier1) resources.

CPU - total amount *
How many processes must be done at the same time (e.g. 1000 "thread")? Each process how much RAM needs (e.g. 2GB per core)?

500 processes, 2 GB RAM each

CPU - HPC *
What fraction (A) must be "HCP" (e.g. high computing power, parallel code, ...)? FPGA? GPU? Even if you don't use it, do you plan/like to try it?

0 (no plan to use HPC resources at the moment, might be revised in future)

CPU - HTC *
What fraction (B) must/could be "HTC" (e.g. the total amount of computing needed could be reached by many single jobs, that could, on the contrary, require high input and output data volumes to be transferred or written on disk)?

100%

CPU - interactive *
What and which fraction must be available interactively (e.g. the "user interface", reached by ssh, where to work on terminal, compiling code and performing light and short tasks)?

Negligible but non zero (e.g. an 8-core dedicated machine could be sufficient for the whole collaboration)

CPU - batch system *
What and which fraction could be available through a batch system (e.g. HTCondor)? The jobs will be "staedy" (e.g. MC simulation or reconstruction submitted and run 24h) or as "burst" (e.g. analysis job peaks before a conference or for a specific, limited in time, task)?

100%. Currently mostly steady jobs (MC that last 1-2 days) but rarely submitted, so in practice "a burst lasting 1-2 days"

CPU - web based tools *
Do you plan/foresee/need to use web-based tools (e.g. Jupyter notebook)?

Not at the moment, might be revised in future.

CPU - personal VMs *
Do you plan/foresee/need to use "personal" interactive Virtual Machines (e.g. VMs created, ad hoc, by a user for a specific task and a limited time interval)?

Not at the moment, we foresee to consider this possibility in future.

CPU - graphic access *
Do you need graphic access to the "personal VMs" and/or to the interactive resources (e.g. X2Go)?

No

CPU - services *
Do you need additional services (e.g. database / cvmfs / ...)? Are there specific requests related to this (e.g. the database needs to be accessible world-wide or, however, its IP need to be registered in N places / the cvmfs repository needs to be writeable by standard users / ...)?

Yes. We will need database, data access (e.g. XrootD/S3) and software access (e.g. cvmfs) services accessible world-wide to profit from opportunistic cloud computing resources. Access rights for all the services should ideally be managed by the INDIGO-IAM system.

CPU - temporary overpledge *
Do you foresee or would like the possibility to increase the pledged CPU power for a limited amount of time (e.g. 1 week or 1 month)? Having them as cloud resources would be effective? If not: why? Is there a big overhead in exploiting temporary or cloud resources?

Yes. Cloud resources would be fully effective and with negligible setup overhead from our side.

CPU - special environment *
The computing (specify which "step", as defined in the first part of the form and in which fraction of the required resources) requires a tailored environment (e.g. specific OS, specific libraries, etc...) or to access/mount remote filesystems (e.g. cvmfs repository with shared libraries)? Do you plan or would like to have a containerized environment with tailored images (e.g. docker)?

CPU - R&D *
Do yuo plan to perform R&D on the computing model and technologies? Which kind and how many resources you would require to perform this kind of activity?

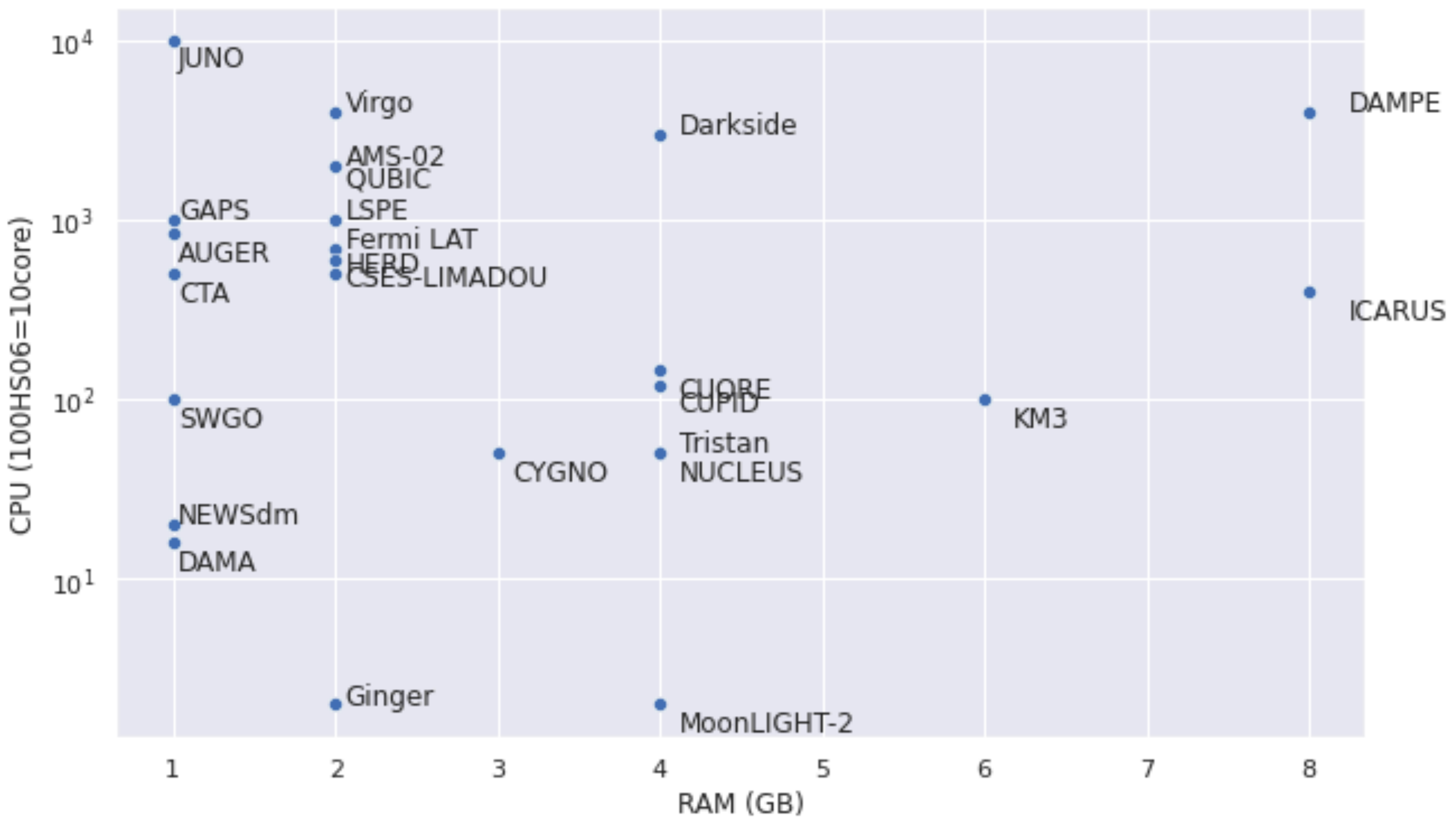
Yes, we plan to continue developing the cloud based computing model which we are currently testing. We will need ~50-100 CPUs with 2 GB RAM each to test on-demand HTC clusters and ~ 100 TB of S3-compatible storage to test cloud data access and storage.

CPU - total amount

CSN2
Fisica delle
Astroparticelle

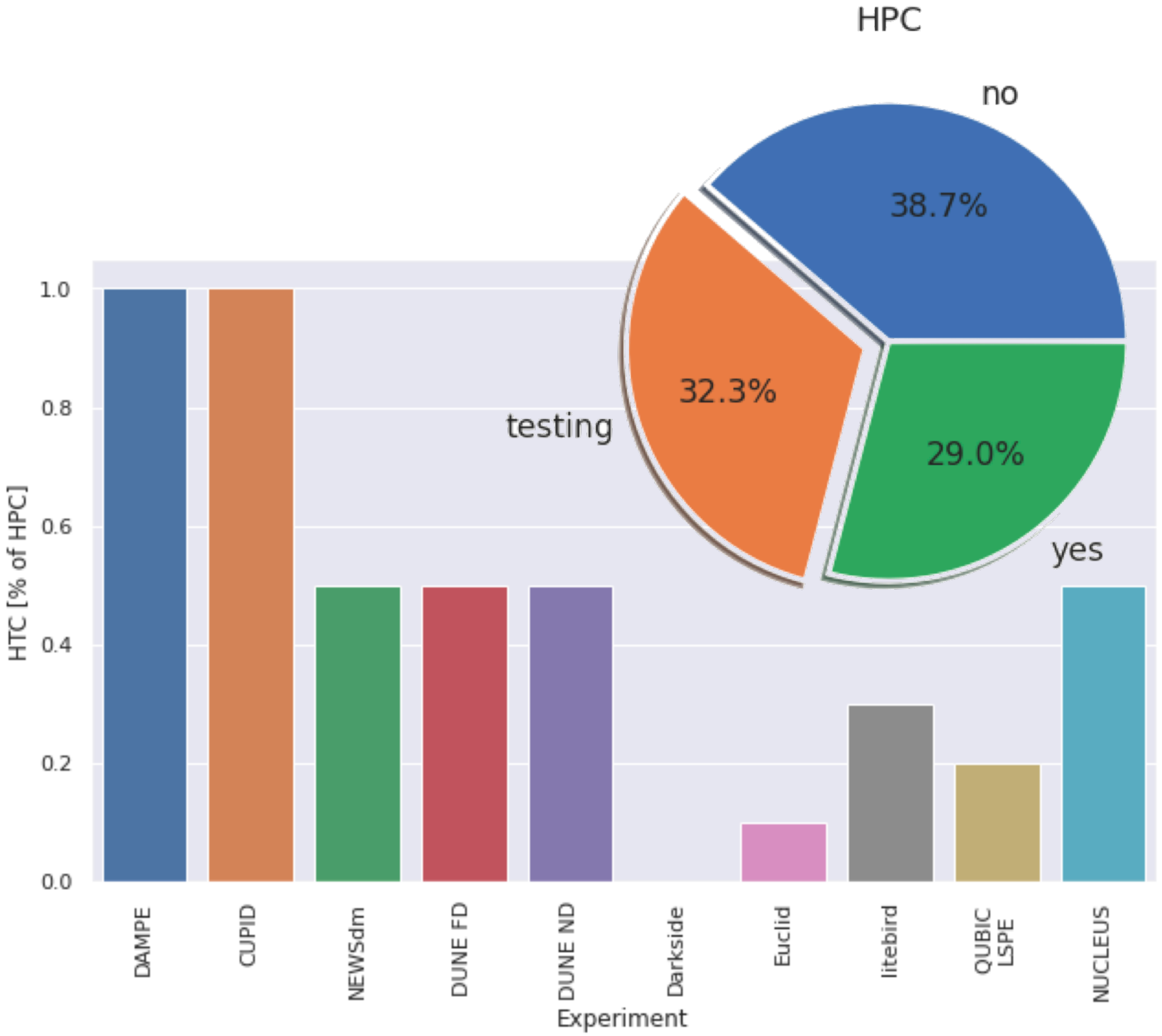
CPU - total amount How many processes must be done at the same time (e.g. 1000 "thread")? Each process how much RAM needs (e.g. 2GB per core)?

- [-1] per alcuni esperimenti non sono stati riportati questi dati (medio piccoli o in fase di approvazione)
- [1] sono prevalentemente esperimenti che non hanno dichiarato esplicitamente la RAM, ma solo i cores/jobs
- [>1] c'e' sicuramente un po' di confusione fra cio' che serve per la **reco** e/o per la **sim** per un totale di circa **300 kHS06**
- probabilmente la conclusione (anche incrociando le informazioni) e' che buona parte della **ricostruzione non viene fatta in Italia**.



CPU - HPC/HTC

What fraction of CPU must be “HPC"? FPGA? GPU? do you plan/like to try it?
What fraction HPC must/could be "HTC"?



- e’ probabile che a breve la maggior parte degli esprimenti utilizzerà HPC, con aumento dei **costi anche di corrente elettrica**…

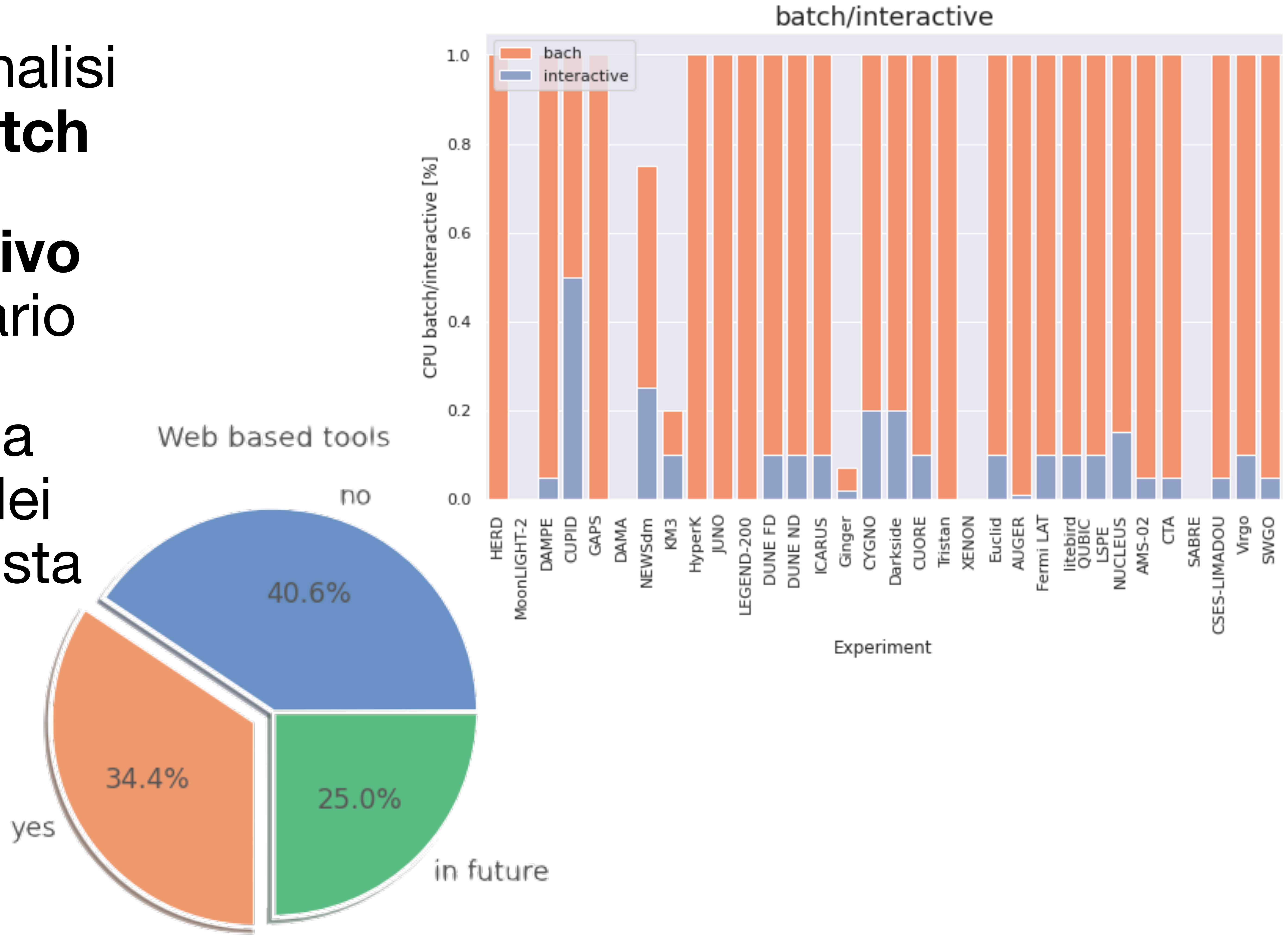
- quindi e’ sempre più importante introdurre un **coretto sfruttamento delle risorse**, lo sfruttamento attuale rimane < 30%: assegnazione dinamica basta potenza integrata e non di picco.

- questo deve essere ottimizzato a livello di infrastruttura (Spiga, di la tua)

CPU - batch/interactive

CSN2 Fisica delle Astroparticelle

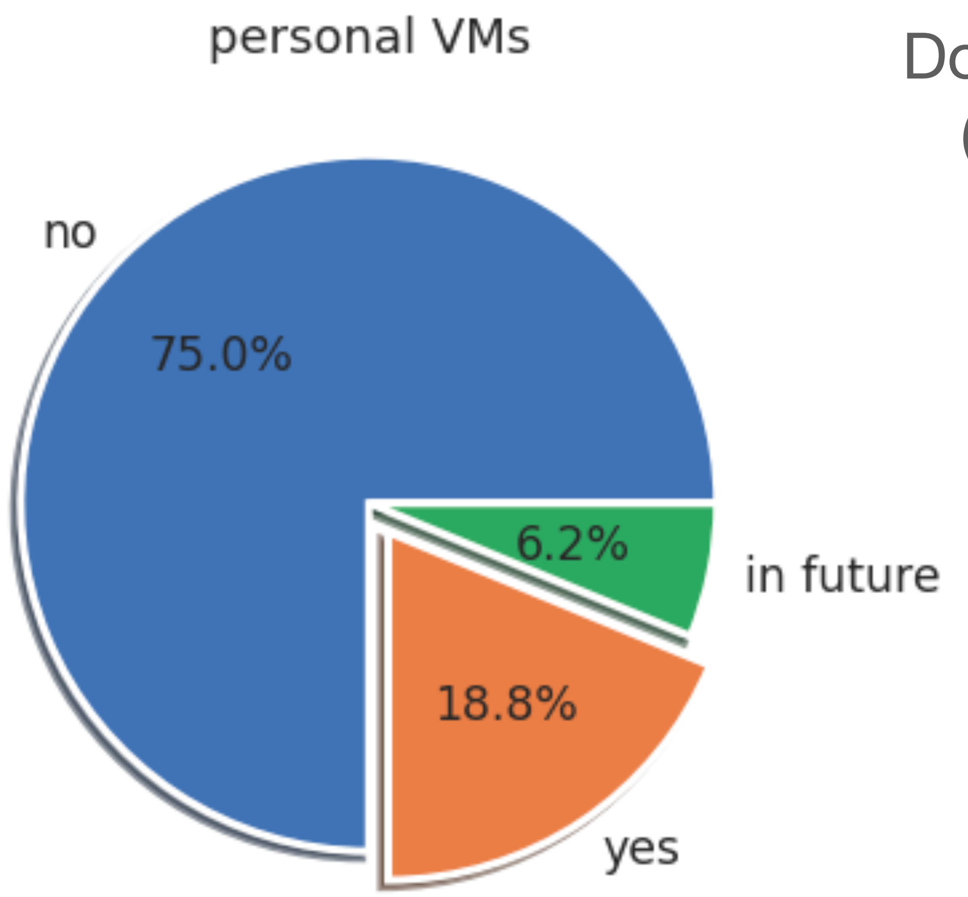
- ovviamente la gran parte dell'analisi e ricostruzione viene fatta in **batch**
- comunque un **10-20% in iterativo** e' +/- sempre richiesto/necessario
- l'utilizzo di **strumenti web** per la visualizzazione (eg jupytehub) dei risultai o per l'analisi interattiva sta diventando preponderante.



G. Mazzitelli, CSN2 Novembre 2022

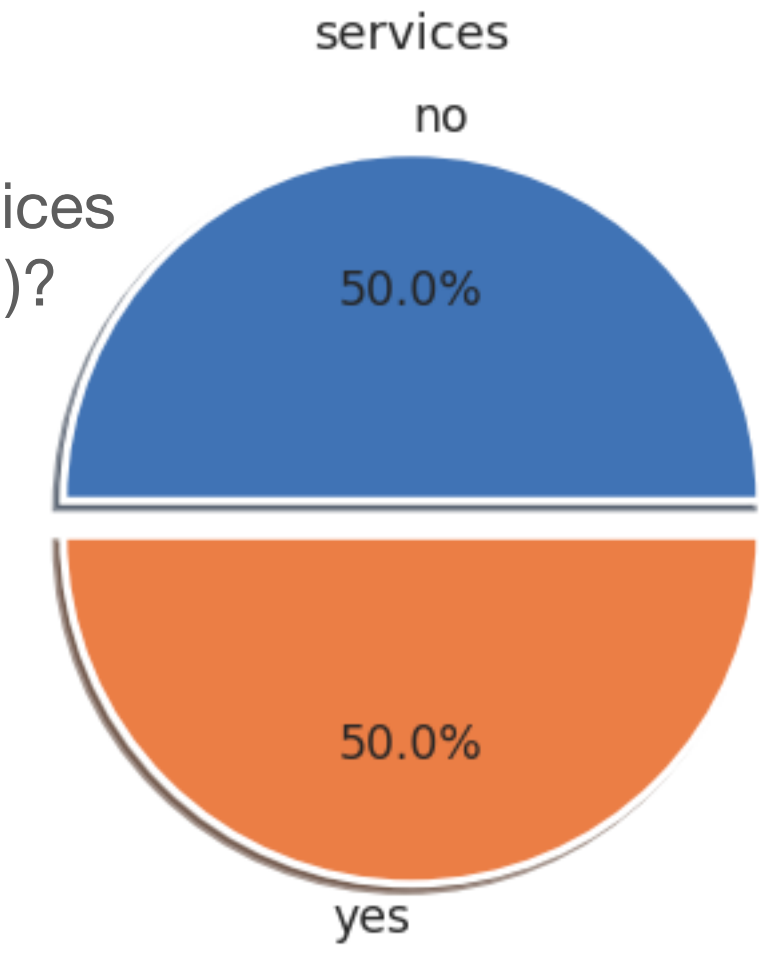
CPU

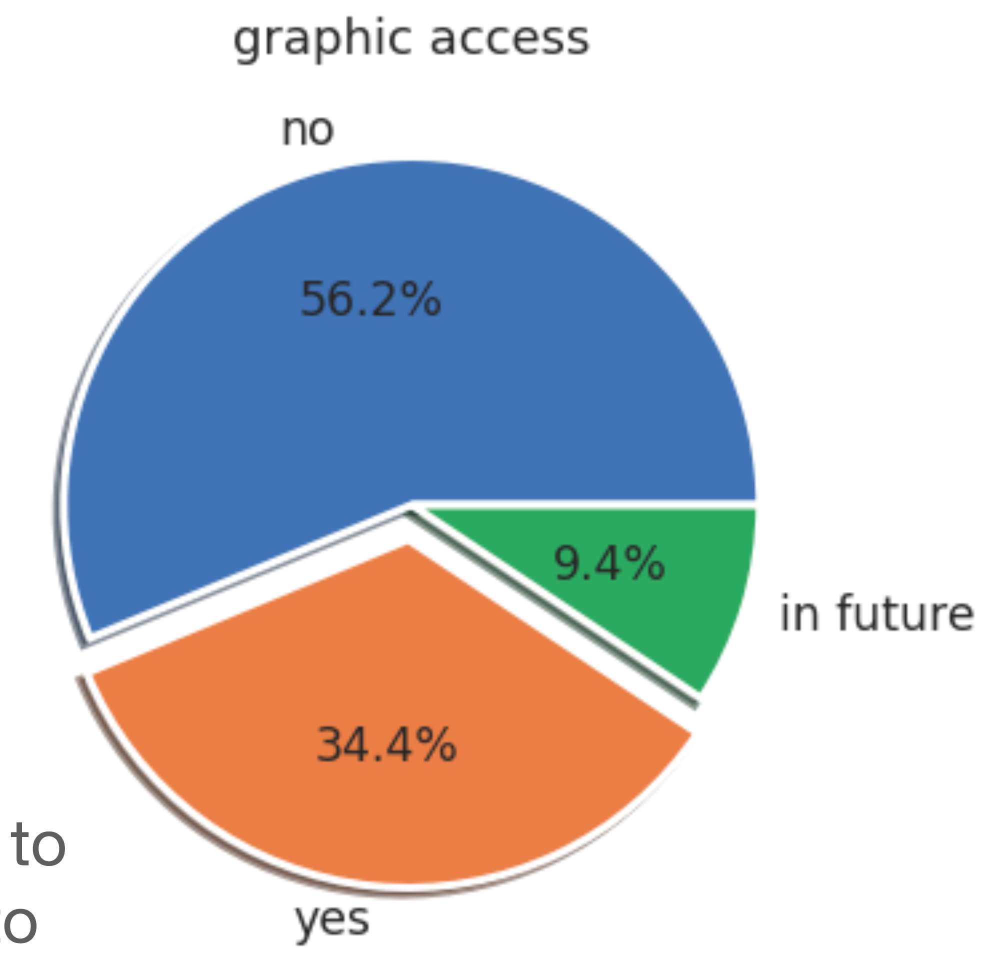
CSN2
Fisica delle
Astroparticelle



Do you plan/foresee/need to use "personal" interactive Virtual Machines

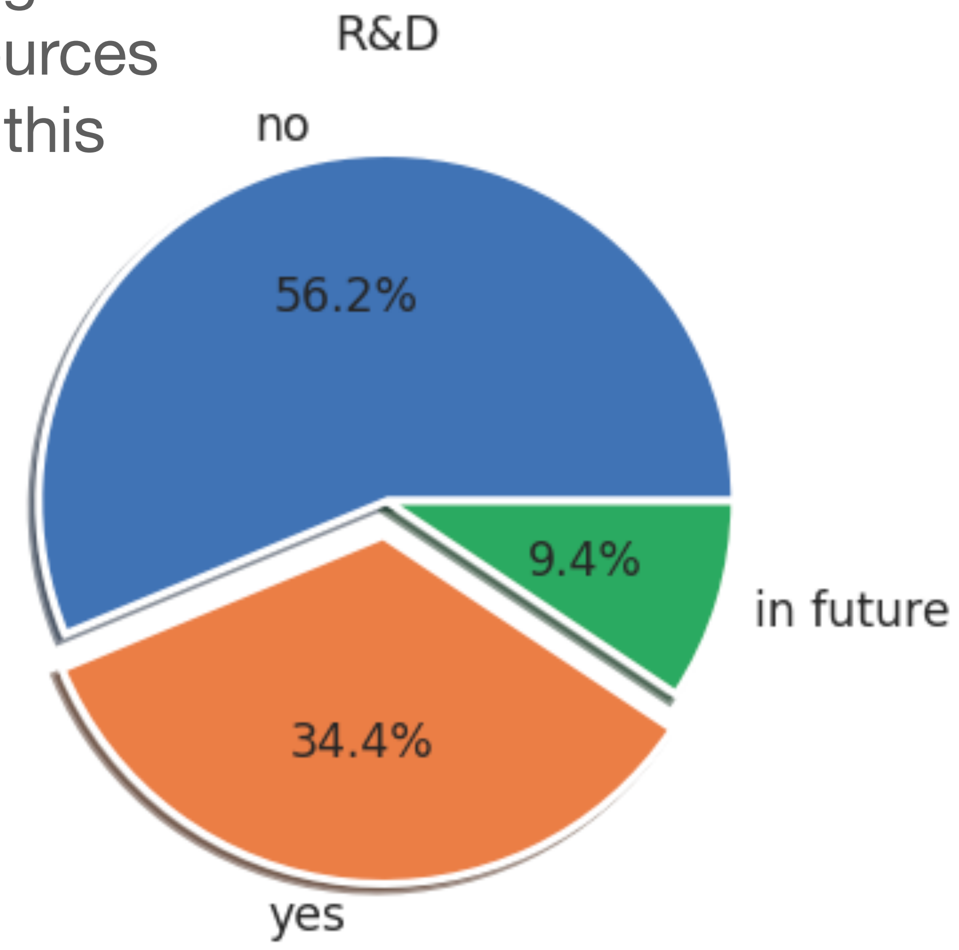
Do you need additional services (e.g. database / cvmfs / ...)?

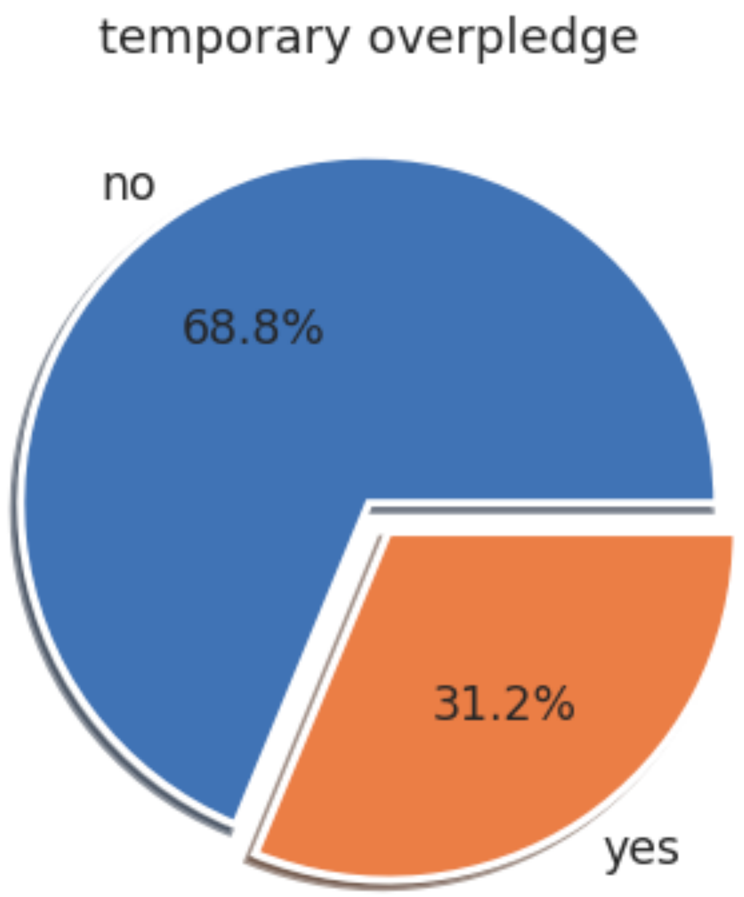




Do you need graphic access to the "personal VMs" and/or to the interactive resources

Do yuo plan to perform R&D on the computing model and technologies? Which kind and how many resources you would require to perform this kind of activity?





Do you foresee or would like the possibility to increase the pledged CPU power for a limited amount of time?

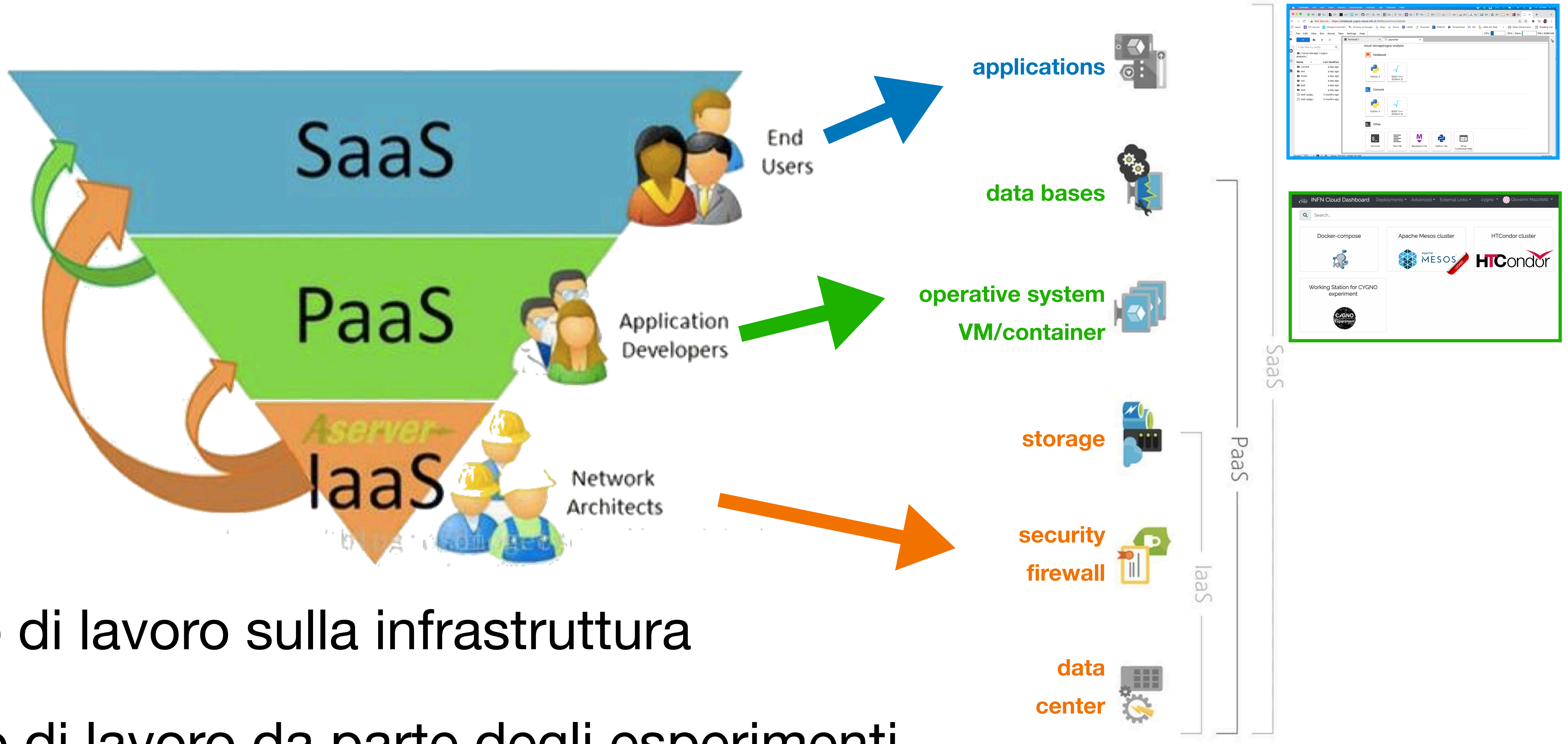
G. Mazzitelli, CSN2 Novembre 2022

qualche nota sulla Cloud

CSN2 Fisica delle Astroparticelle

- nell'INFN **l'infrastruttura** diventerà unica, HPC/Grid/Cloud divengono quindi solo pladge gestite, servite e referate in modo unico e centralizzato.

- quindi se ancora avremo bisogno di distinguerle, il modello **cloud** puo' offrire alcuni vantaggi:



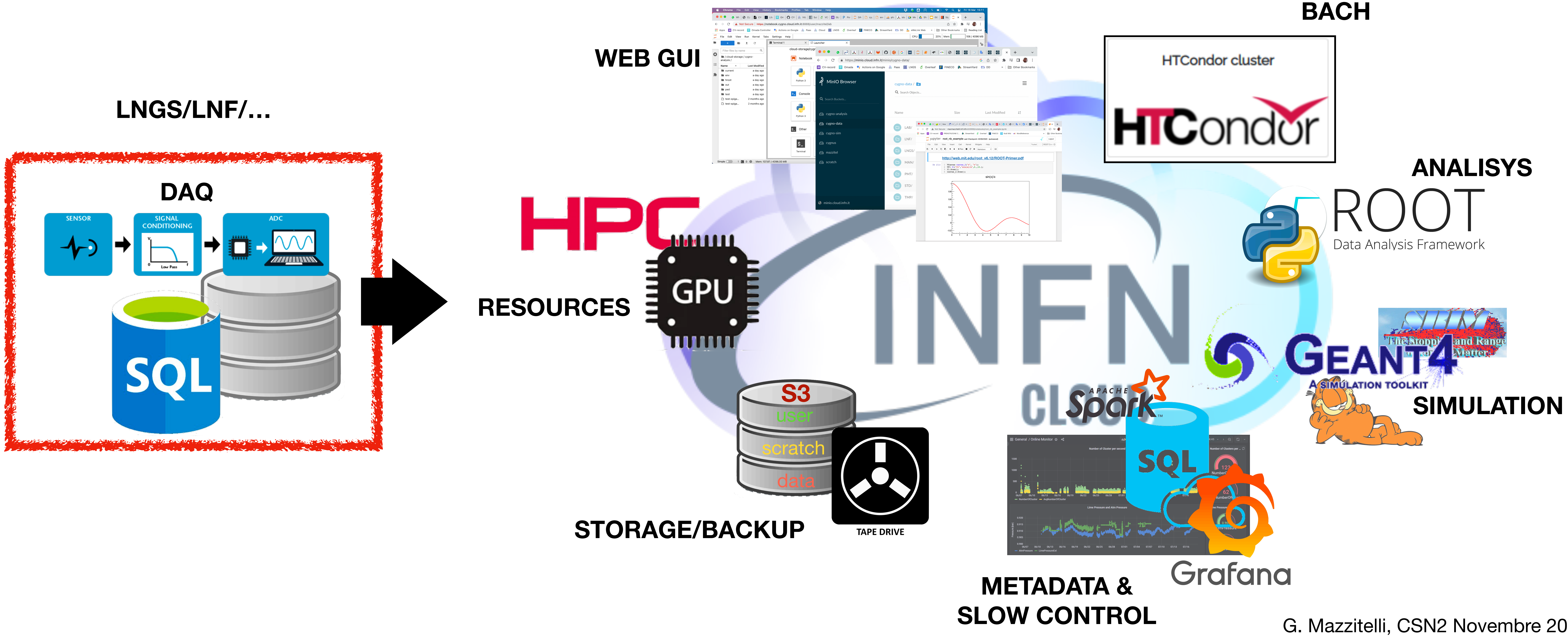
  - standardizzazione delle **risorse** —> risparmio di lavoro sulla infrastruttura
  - standardizzazione del **software** —> risparmio di lavoro da parte degli esperimenti
  - gestione delle **autenticazioni** (in particolare degli user nazionali ed internazionali) integrata
  - maggiore **sicurezza**
  - gestione "on demand" —> **risparmio di risorse e corrente** (almeno in parte)

Implementare soluzioni su misura: comporre dei servizi

D. Ciangottini on behalf of WP5 (CCR 21)

- **Composizione di due servizi esistenti:**
  - Vorrei uno jupyterHub che interagisca (per authZ o per accesso ai dati) con un servizio che instanzio allo stesso tempo su docker-compose
- **Composizione di un servizio esistente con uno nuovo:**
  - Vorrei che la mia applicazione che gira su docker-compose sia combinata alla creazione di un DB su volume persistente su una macchina dedicata
- **Customizzazione di una ricetta esistente:**
  - Docker-compose con un particolare volume driver per caricare i miei dati da uno storage esterno
- **Integrazione con cloud storage INFN Cloud per gestione persistenza:**
  - ML-INFN: soluzione su docker compose, ma che necessita di configurazioni ad-hoc per accedere a GPU e un monitoring dedicato
  - CYGNO: soluzione distribuita su cluster K8s che offre un endpoint JHUB ma con integrazione di un batch system integrato, tape, GPU…

CYGNO/QUAX … model



LNGS/LNF/...

DAQ

RESOURCES

WEB GUI

BACH

ANALISYS

SIMULATION

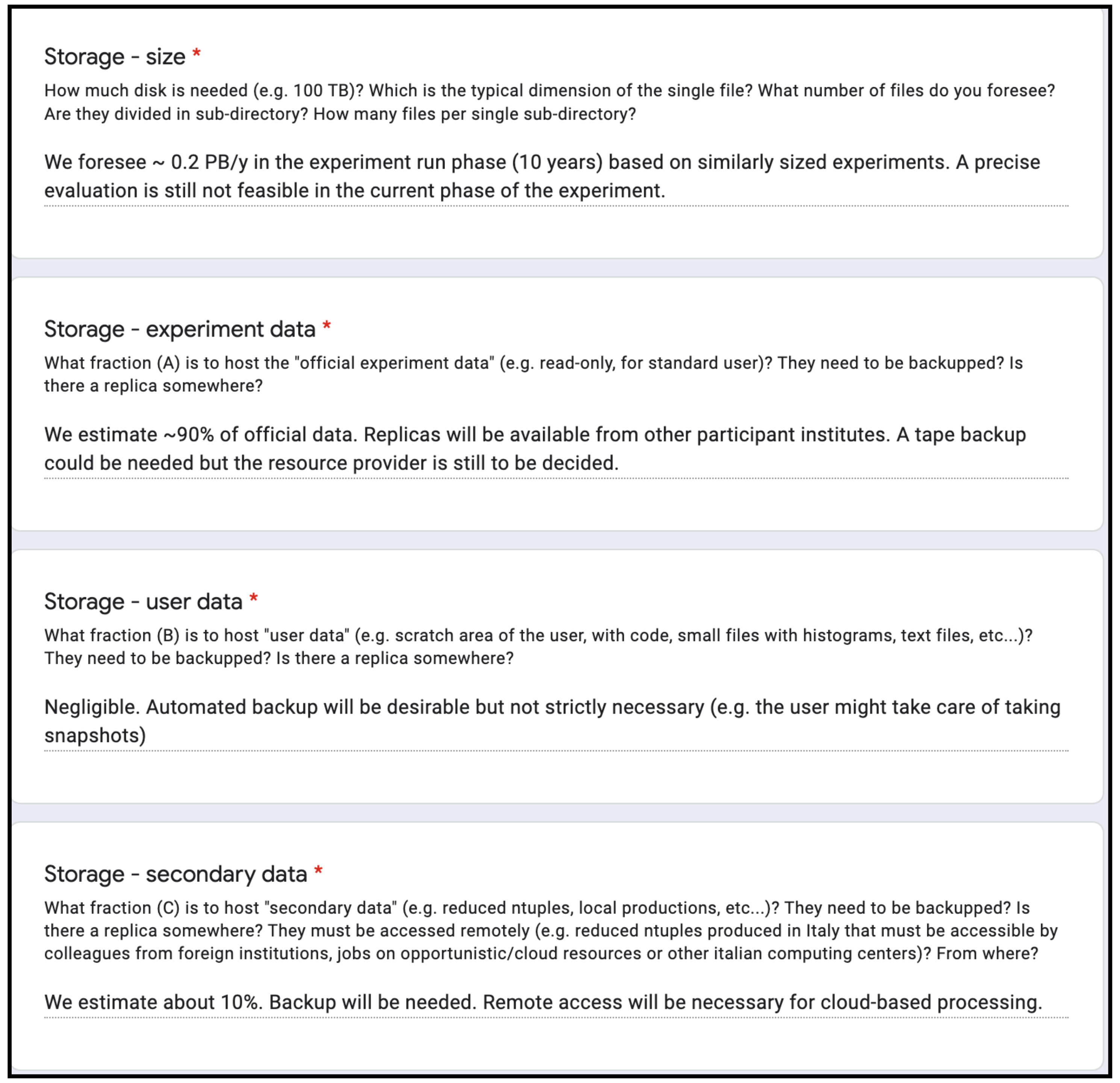
STORAGE/BACKUP

METADATA &
SLOW CONTROL

Survey 3/3

impact on INFN resources Storage/Tape

CSN2 Fisica delle Astroparticelle



Storage - size *

How much disk is needed (e.g. 100 TB)? Which is the typical dimension of the single file? What number of files do you foresee? Are they divided in sub-directory? How many files per single sub-directory?

We foresee ~ 0.2 PB/y in the experiment run phase (10 years) based on similarly sized experiments. A precise evaluation is still not feasible in the current phase of the experiment.

Storage - experiment data *

What fraction (A) is to host the "official experiment data" (e.g. read-only, for standard user)? They need to be backupped? Is there a replica somewhere?

We estimate ~90% of official data. Replicas will be available from other participant institutes. A tape backup could be needed but the resource provider is still to be decided.

Storage - user data *

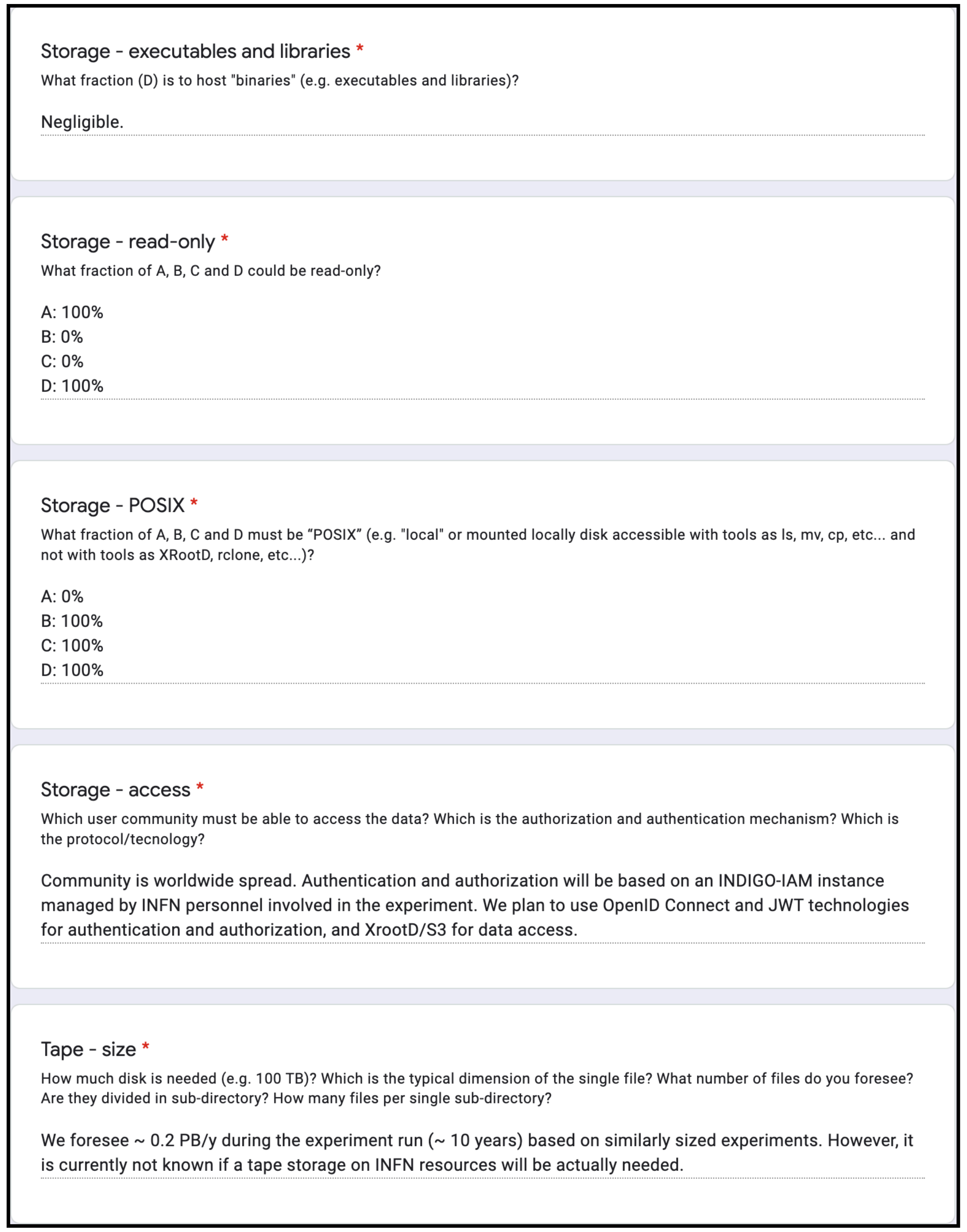
What fraction (B) is to host "user data" (e.g. scratch area of the user, with code, small files with histograms, text files, etc...)? They need to be backupped? Is there a replica somewhere?

Negligible. Automated backup will be desirable but not strictly necessary (e.g. the user might take care of taking snapshots)

Storage - secondary data *

What fraction (C) is to host "secondary data" (e.g. reduced ntuples, local productions, etc...)? They need to be backupped? Is there a replica somewhere? They must be accessed remotely (e.g. reduced ntuples produced in Italy that must be accessible by colleagues from foreign institutions, jobs on opportunistic/cloud resources or other italian computing centers)? From where?

We estimate about 10%. Backup will be needed. Remote access will be necessary for cloud-based processing.



Storage - executables and libraries *

What fraction (D) is to host "binaries" (e.g. executables and libraries)?

Negligible.

Storage - read-only *

What fraction of A, B, C and D could be read-only?

A: 100%
B: 0%
C: 0%
D: 100%

Storage - POSIX *

What fraction of A, B, C and D must be "POSIX" (e.g. "local" or mounted locally disk accessible with tools as ls, mv, cp, etc... and not with tools as XRootD, rclone, etc...)?

A: 0%
B: 100%
C: 100%
D: 100%

Storage - access *

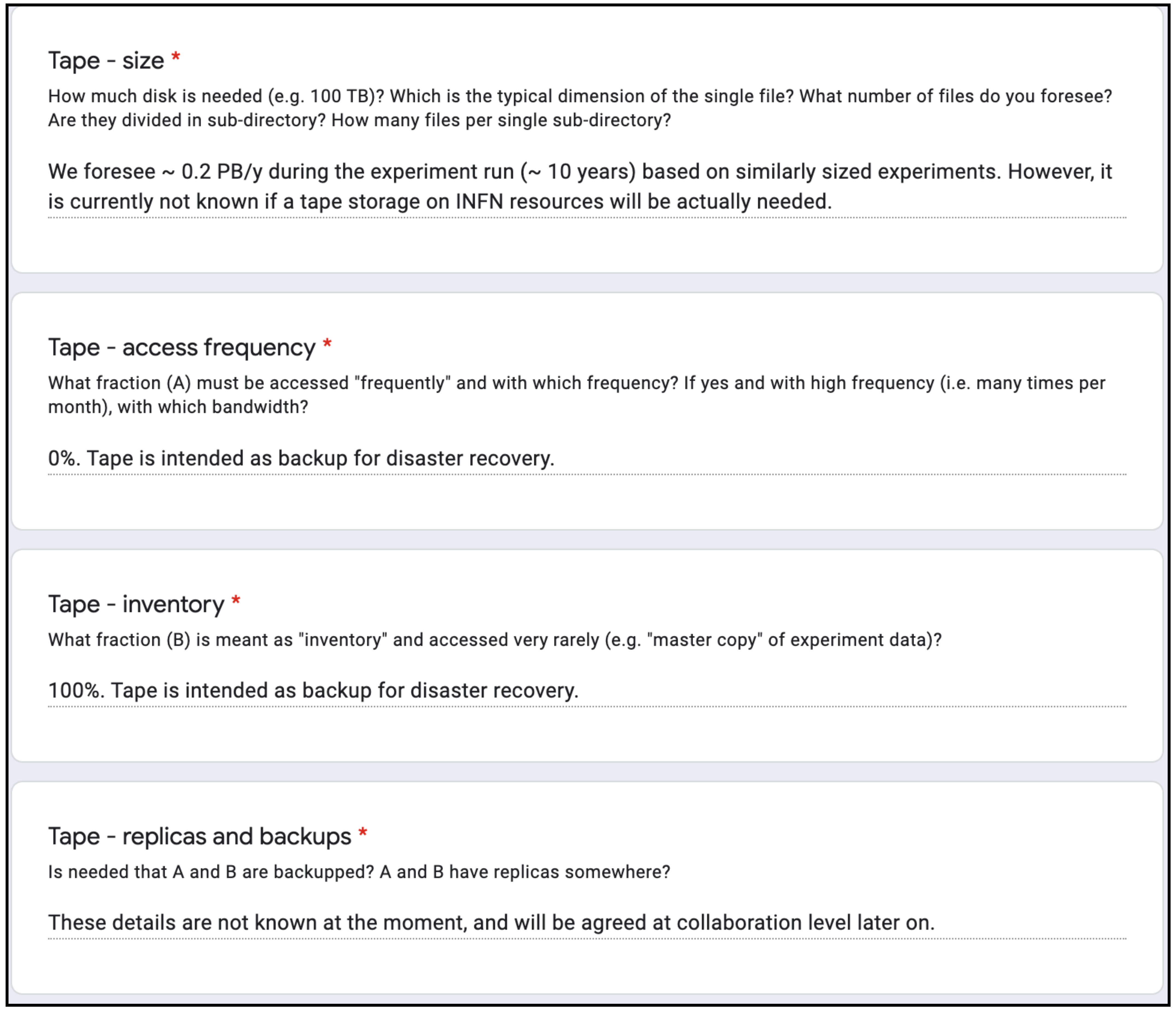
Which user community must be able to access the data? Which is the authorization and authentication mechanism? Which is the protocol/tecnology?

Community is worldwide spread. Authentication and authorization will be based on an INDIGO-IAM instance managed by INFN personnel involved in the experiment. We plan to use OpenID Connect and JWT technologies for authentication and authorization, and XrootD/S3 for data access.

Tape - size *

How much disk is needed (e.g. 100 TB)? Which is the typical dimension of the single file? What number of files do you foresee? Are they divided in sub-directory? How many files per single sub-directory?

We foresee ~ 0.2 PB/y during the experiment run (~ 10 years) based on similarly sized experiments. However, it is currently not known if a tape storage on INFN resources will be actually needed.



Tape - size *

How much disk is needed (e.g. 100 TB)? Which is the typical dimension of the single file? What number of files do you foresee? Are they divided in sub-directory? How many files per single sub-directory?

We foresee ~ 0.2 PB/y during the experiment run (~ 10 years) based on similarly sized experiments. However, it is currently not known if a tape storage on INFN resources will be actually needed.

Tape - access frequency *

What fraction (A) must be accessed "frequently" and with which frequency? If yes and with high frequency (i.e. many times per month), with which bandwidth?

0%. Tape is intended as backup for disaster recovery.

Tape - inventory *

What fraction (B) is meant as "inventory" and accessed very rarely (e.g. "master copy" of experiment data)?

100%. Tape is intended as backup for disaster recovery.

Tape - replicas and backups *

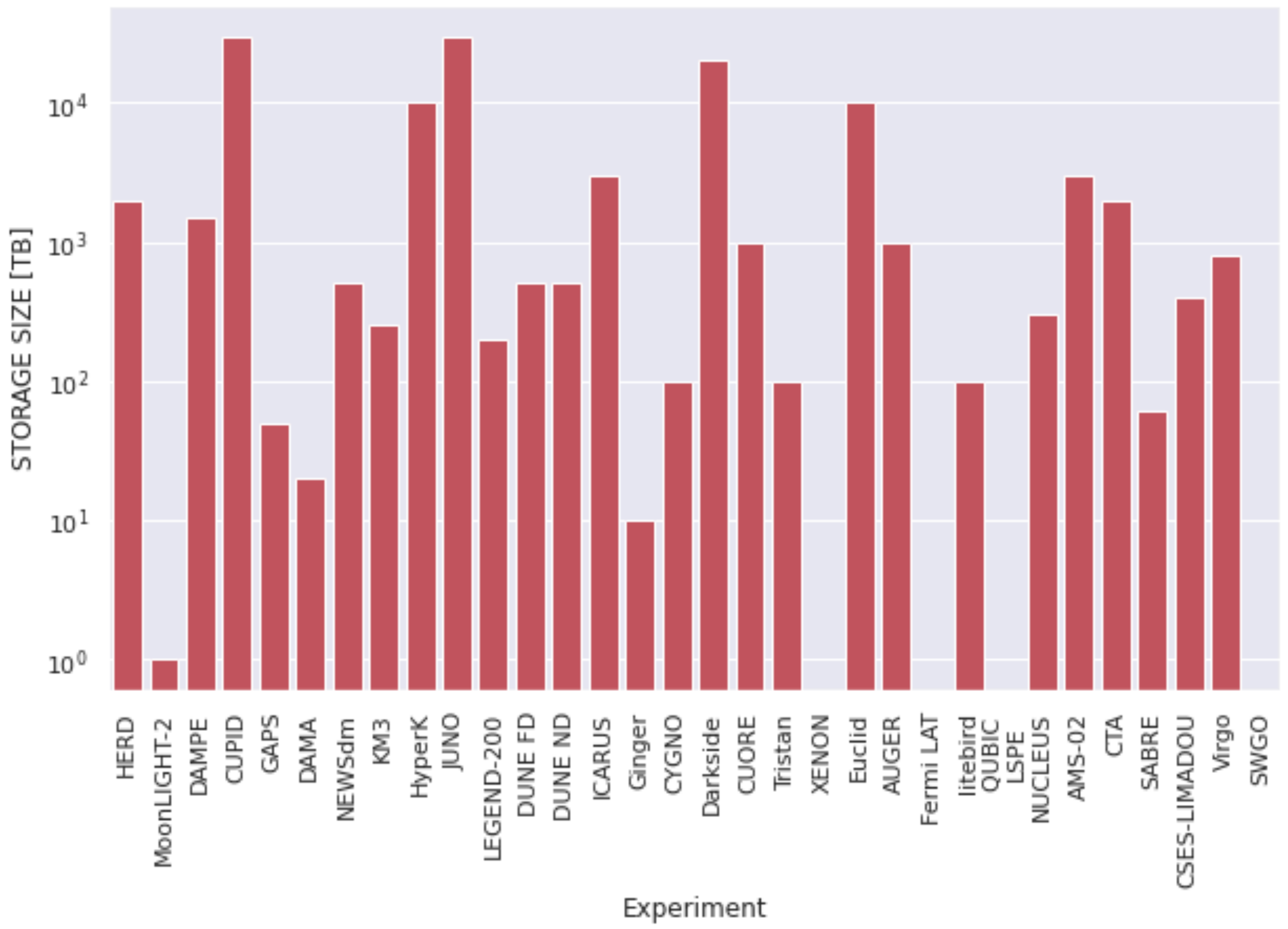
Is needed that A and B are backupped? A and B have replicas somewhere?

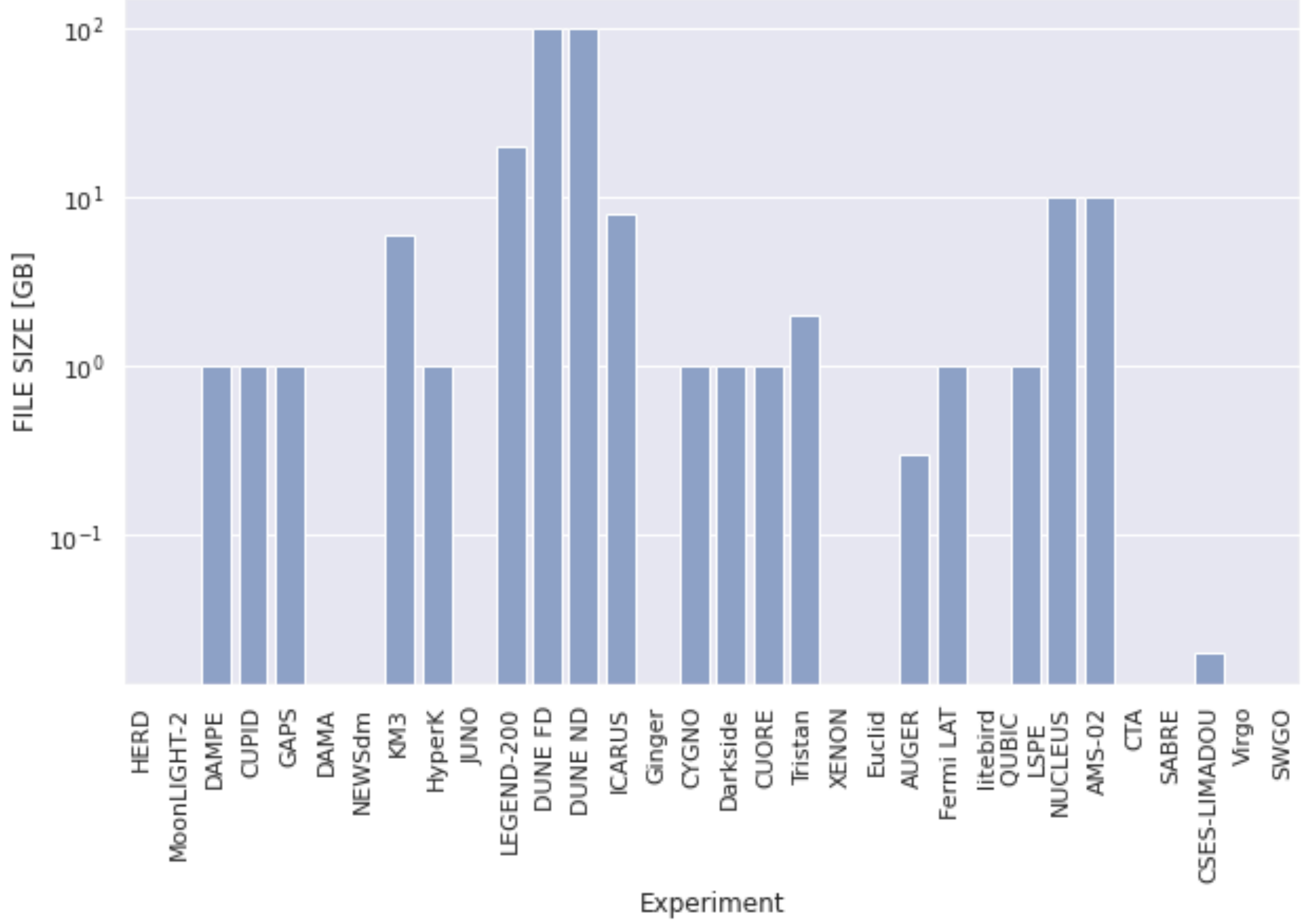
These details are not known at the moment, and will be agreed at collaboration level later on.

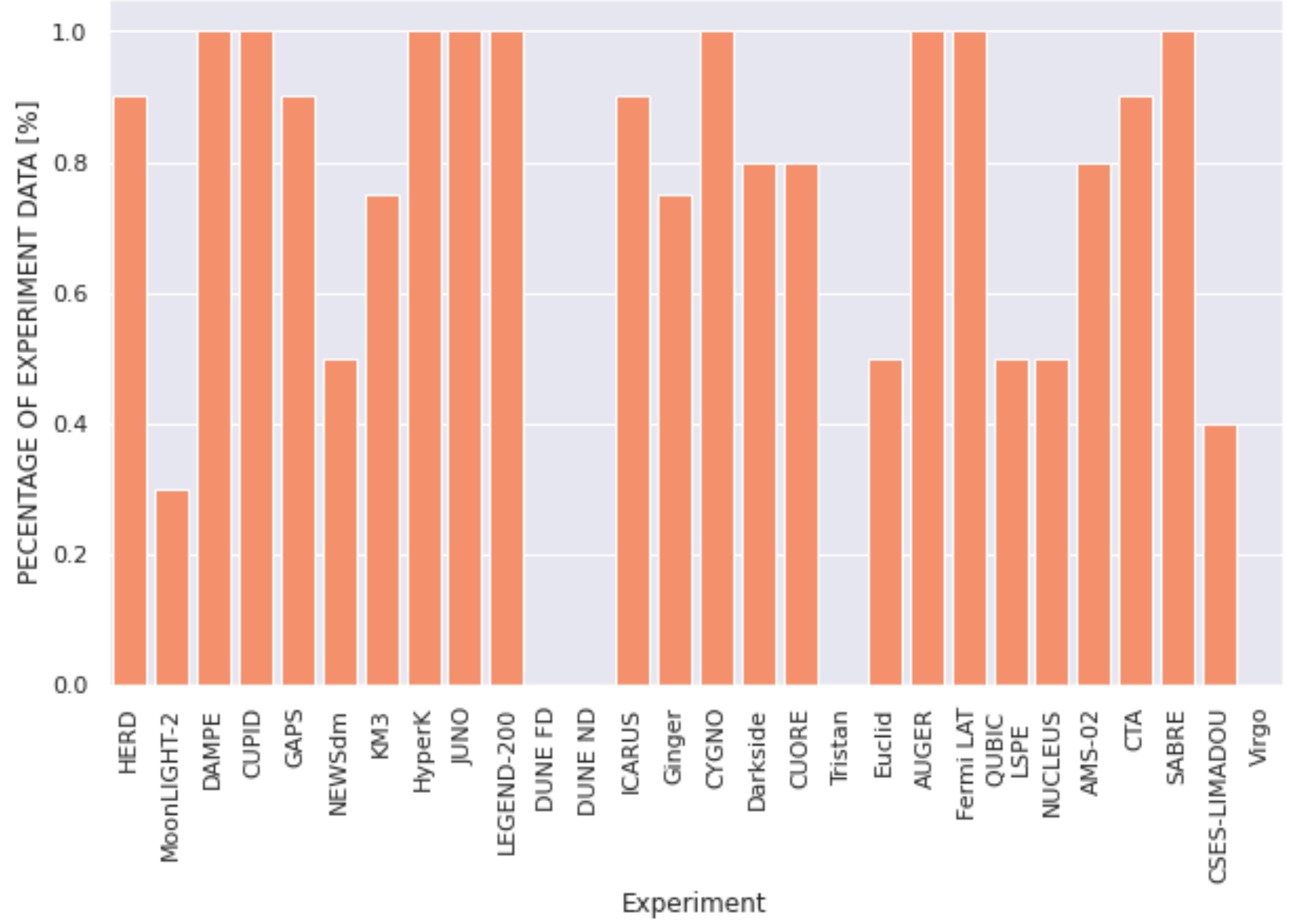
Storage/TAPE

in sintesi

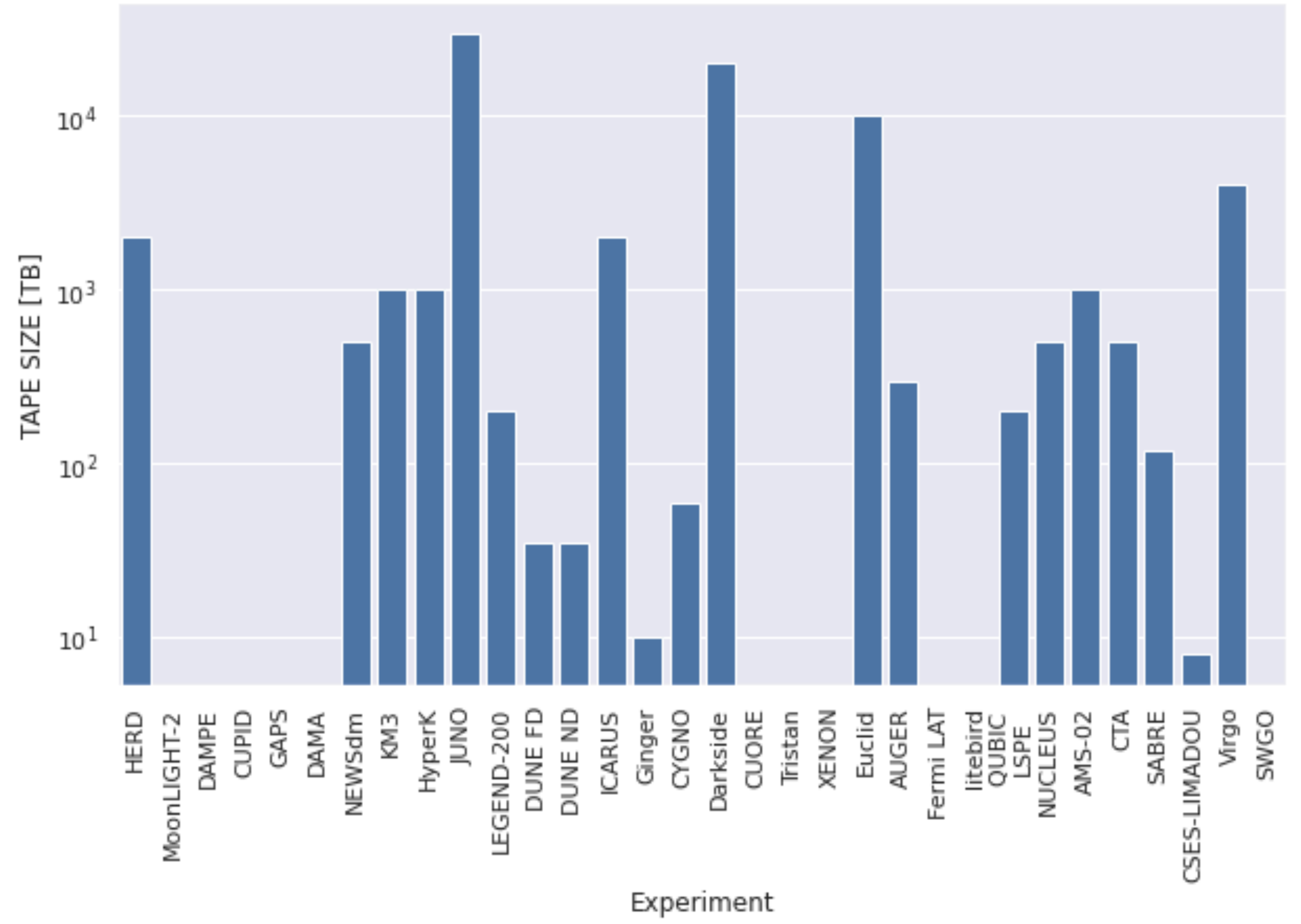
CSN2
Fisica delle
Astroparticelle







- c'e' una necessita' > **100 PB totali**, determinata soprattutto dai futuri grandi esperimenti
- buona parte degli esprimenti hanno **files con size O(GB)**, ottimali per file system ad oggetti
- buona parte degli esperimenti ospitano in Italia solo **parte dei dati dell'esperimento**
- quasi tutti usano **POSIX con accesso limitato** principalmente alla collaborazione
- le richieste sul TAPE sono minori **O(70) PB** ma non trascurabili.



conclusioni

CSN2
Fisica delle
Astroparticelle

- l’INFN sta effettuando una grande rivoluzione sul calcolo, che riguarda sia **l’infrastruttura**, che le **persone** che le **pratiche** da seguire. Questa rivoluzione e’ detta dalla nascita dell’ICSC, finanziata dal PNRR e guidata dalla CNC

- le nostre conclusioni sul survey possono solo essere **preliminari**, sia per **l’eterogeneità** dei dati raccolti, utilissima, ma solo parzialmente sfruttabile, sia per l’imminente **survey INFN** che sarà alla base delle **scelte infrastrutturali dell’INFN**

- appare chiaro che i **modelli di calcolo** degli esperimenti afferenti alla CSN2 sono spesso **complessi e specifici**, e non sia facile trovare dei **fattori comuni** anche perché non tutti usano le risorse INFN al medesimo scopo (daq/reco/sim/meta/ecc), a volte sono **tecnicamente superati**.

- oltre il 50% degli esperimenti che hanno partecipato al survey sicuramente oggi **potrebbe essere disegnato** in altro modo e sfruttare **più efficientemente** le risorse assegnate/disponibili, ma non c’e’ grande interesse in questo (vedi domanda sulle pladge dinamiche, non e’ un problema degli esperimenti oggi?)

conclusioni

CSN2 Fisica delle Astroparticelle

- ovviamente da questo rimangono esclusi i **grandi esperimenti** (Virgo, JUNO, DUNE, CTA, DarkSide, Euclid, ecc) ai quali possiamo solo raccomandare di essere **partecipi della rivoluzione in atto nell'INFN**, che può essere un grande vantaggio nel futuro visti gli sforzi dell'INFN e che se non condivisa/accettata potrebbe creare dei problemi in futuro.

- sicuramente si vede un trend chiaro vero **l'HPC e l'R&D sul calcolo** anche se gli esperimenti ancora bassano molto le loro attività su **standard noti e consolidati** (es POSIX, SSH, ecc), inoltre le **richieste** per i prossimi anni non sono trascurabili e avranno dei **costi** notevoli sia per la **mole** che per la **tipologia** con lo sfruttamento crescente dell'HPC/HTC

- nella maggior parte dei casi i dati sono accessibili **solo alla collaborazione**, con modalità poco pratiche, solo il 25% sembra essere adeguato a scambi inter-collaborazioni se non a servire almeno parte dei dati come "**open-data**"

- la richiesta indicativa per i prossimi anni della CSN2 equivale +/- a **300 kHS06**, **100 PB** di **disco**, **70 PB** di **tape**, +/- un mezzo CMS…

non sottovalutare con quale strumento tecnicamente fare il sondaggio