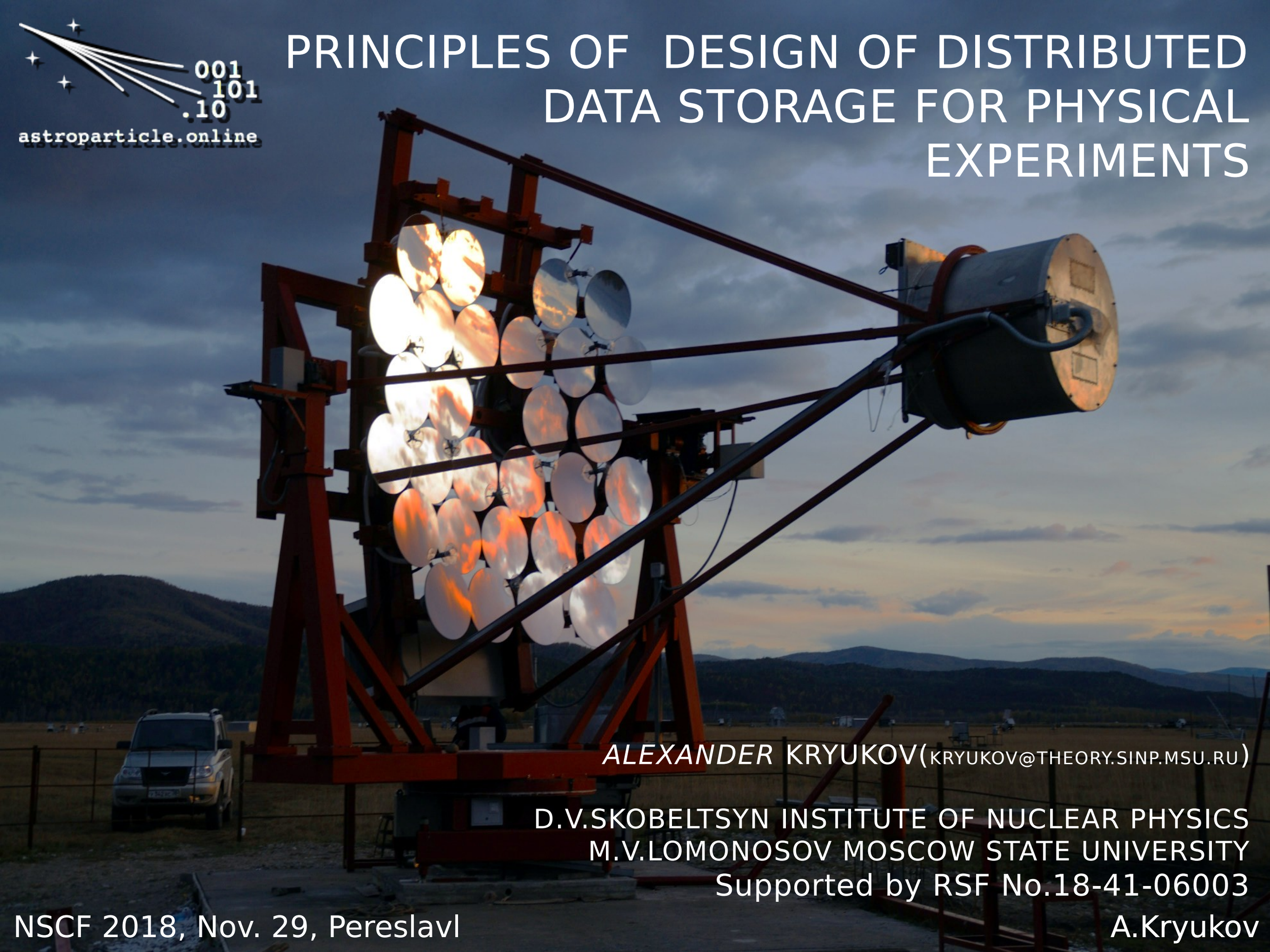


PRINCIPLES OF DESIGN OF DISTRIBUTED DATA STORAGE FOR PHYSICAL EXPERIMENTS

ALEXANDER KRYUKOV(KRYUKOV@THEORY.SINP.MSU.RU)

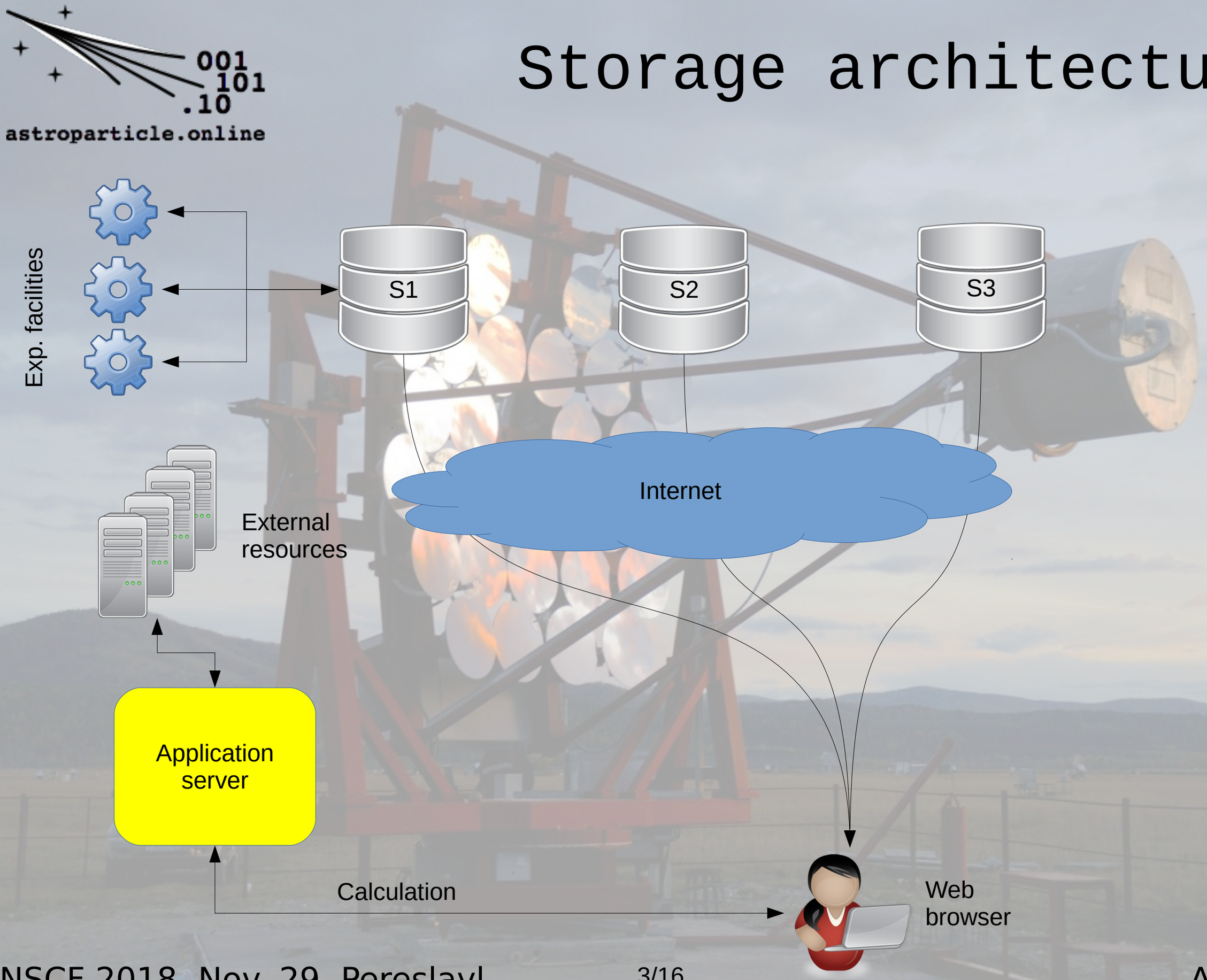
D.V.SKOBELTSYN INSTITUTE OF NUCLEAR PHYSICS
M.V.LOMONOSOV MOSCOW STATE UNIVERSITY

Supported by RSF No.18-41-06003

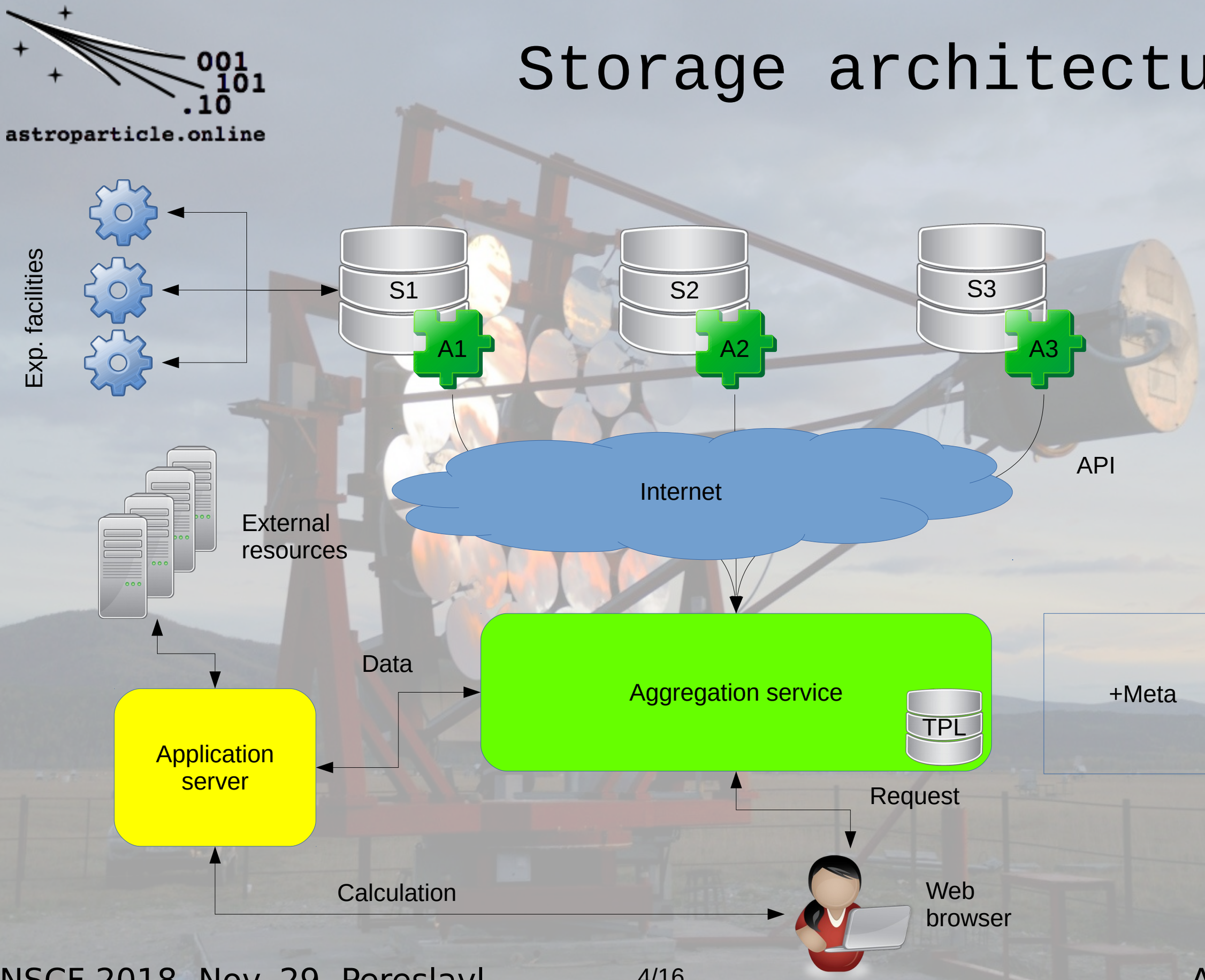
REQUIREMENTS FOR THE DATA STORAGE

- Multiple experiments (TAIGA, KASCADE, etc.)
- Hundreds of terabytes and more of raw data at each site
- Remote access to data as local file systems
- On-demand data transfer by requests only
- Automatic real-time updates
- No change to existing site infrastructure, only add-ons

Storage architecture



Storage architecture



astroparticle.online

POSSIBLE SOLUTIONS

CernVM File system

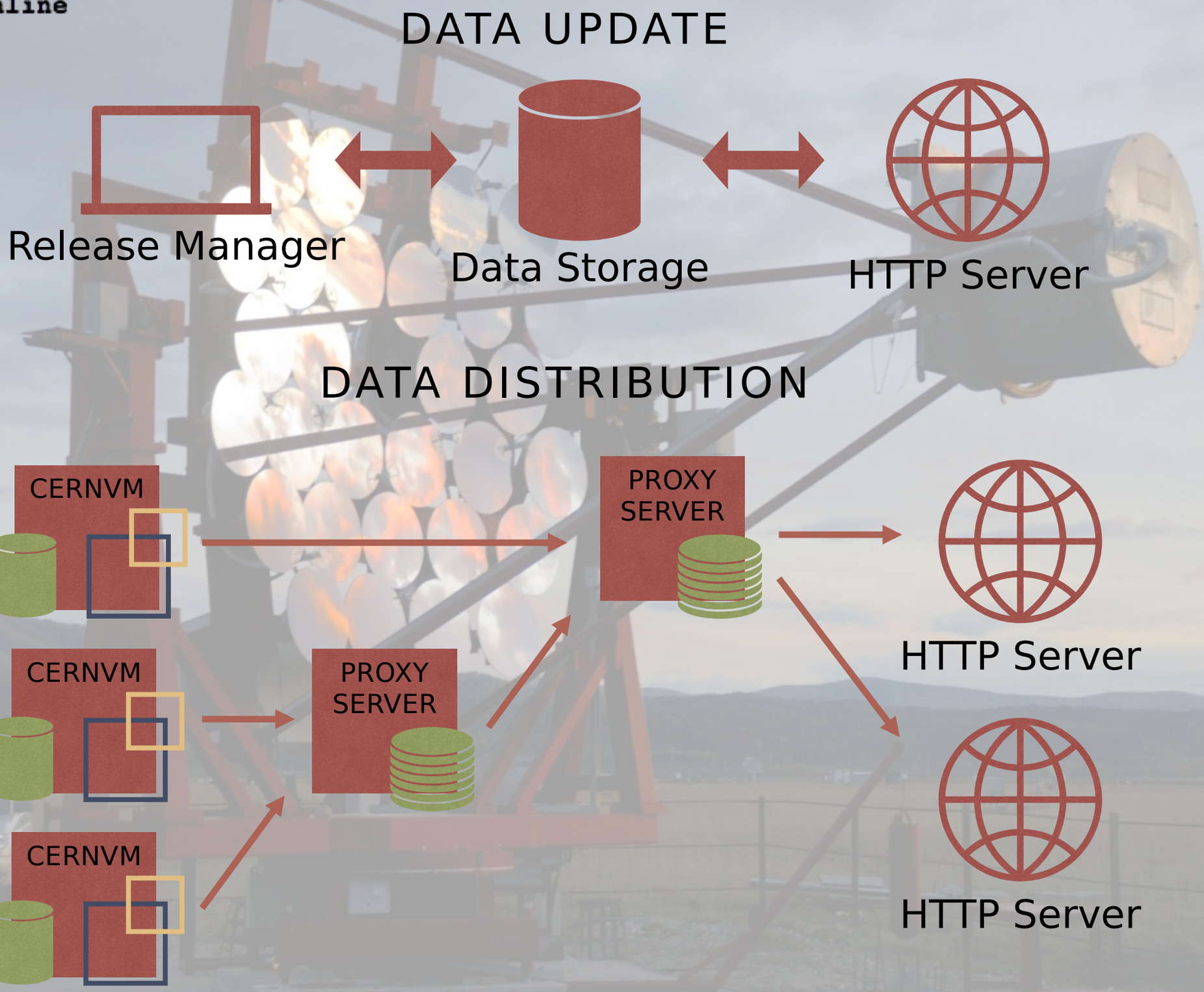
APACHE hadoop

OpenAFS

CERNVM-FS

- Data are left untouched in their own file system
- CernVM-FS indexes the data and changes, stores only the metadata (indices, checksums, locations, etc.) and data tree
- CernVM-FS uses HTTP as the data transfer protocol, so there’s no firewall problem
- Data transfer starts only on actual reads
- Multilevel cache-proxy servers

CERNVM-FS


DATA UPDATE
Release Manager
Data Storage
HTTP Server
DATA DISTRIBUTION
CERNVM
CERNVM
CERNVM
PROXY SERVER
PROXY SERVER
HTTP Server
HTTP Server

astroparticle.online

CERNVM-FS

TAIGA

CVMFS SERVER

KASCADE

CVMFS SERVER

AGGREGATOR

PROXY

AO DATA SERVICE

- METADATA
- CACHE
- SEARCH
- ACCESS POLICY

CVMFS CLIENT

CVMFS CLIENT

```
|— TAIGA
|    |— TAIGA_FILE 1
|    |— TAIGA_FILE 2
|— KASCADE
|    |— KASCADE_FILE 1
|    |— KASCADE_FILE 2
```

Metadata management

- Two type MD
  - File level MD (F)
    - Data/time
    - Source (detector)
    - ....
  - Event level MD (E)
    - Energy of shower
    - Particle type
    - Distance
    - ...
- User requests have also two types
  - File level
  - Event level

astroparticle.online

Exp. facilities

MD

S1

S2

S3

A1

A2

A3

Internet

API

F+E

External resources

Data

F(F+E)

F(E2)

TPL

F+E

+Meta (TimeScale DB)

Application server

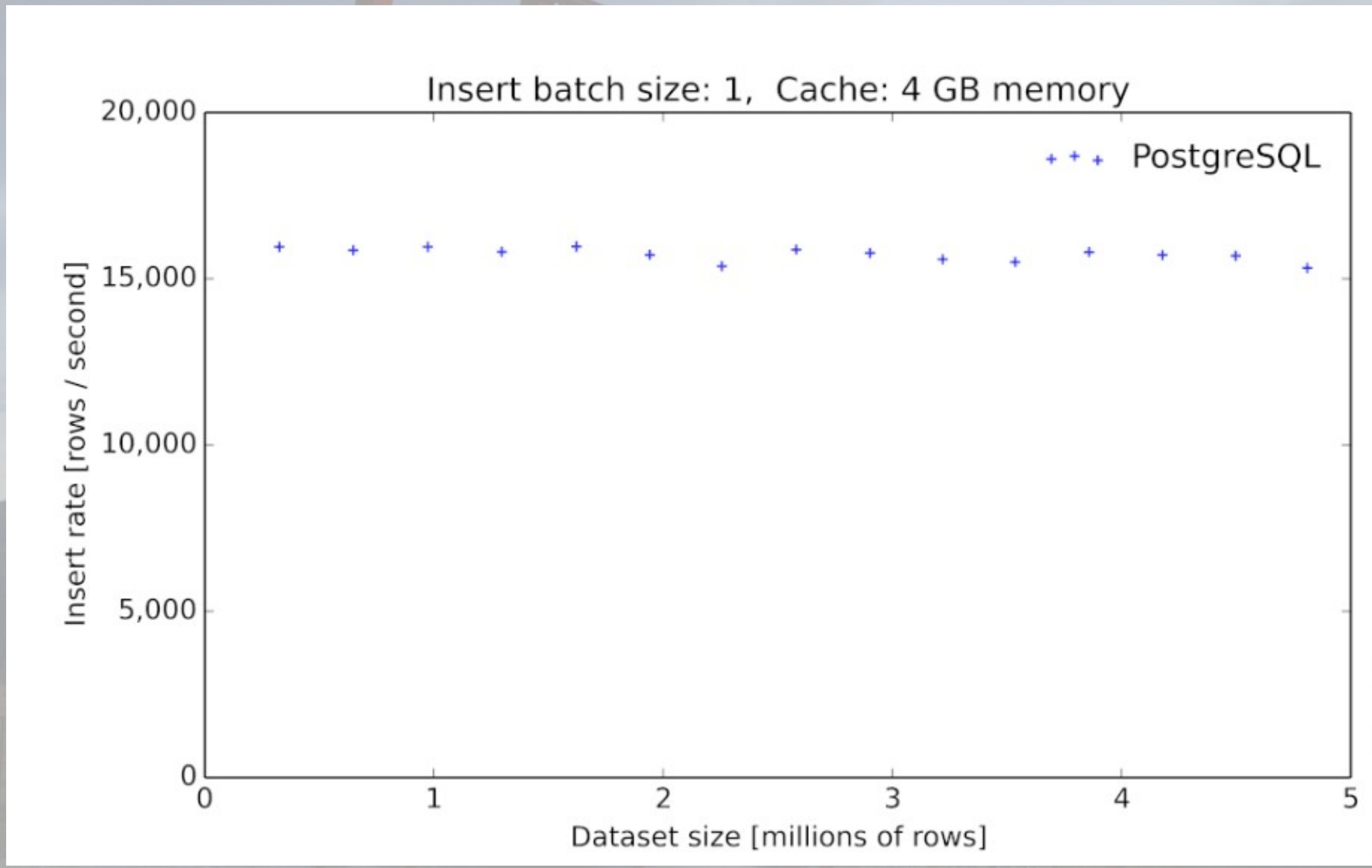
Request F+E

Calculation

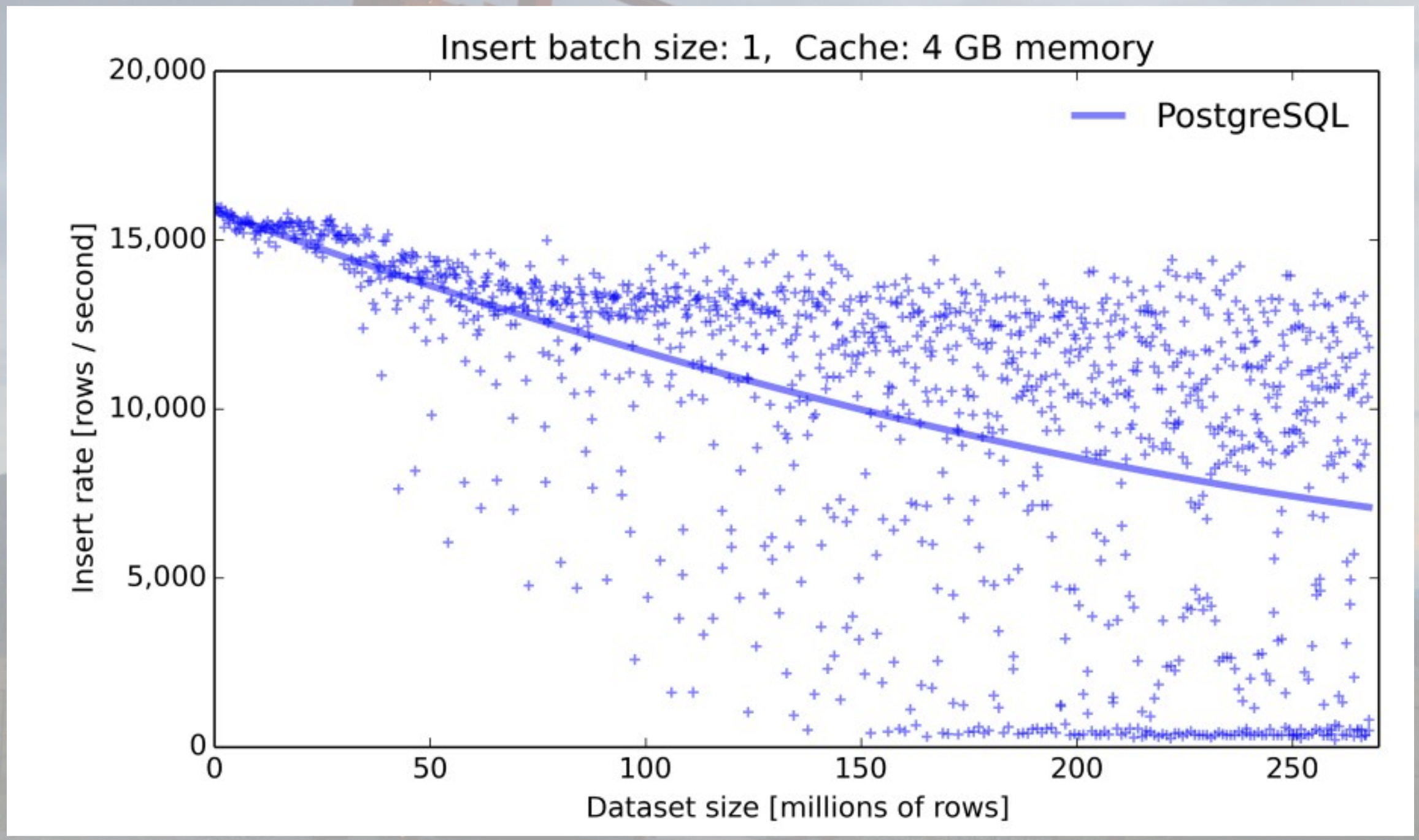
Web browser

astroparticle.online

Time series DB

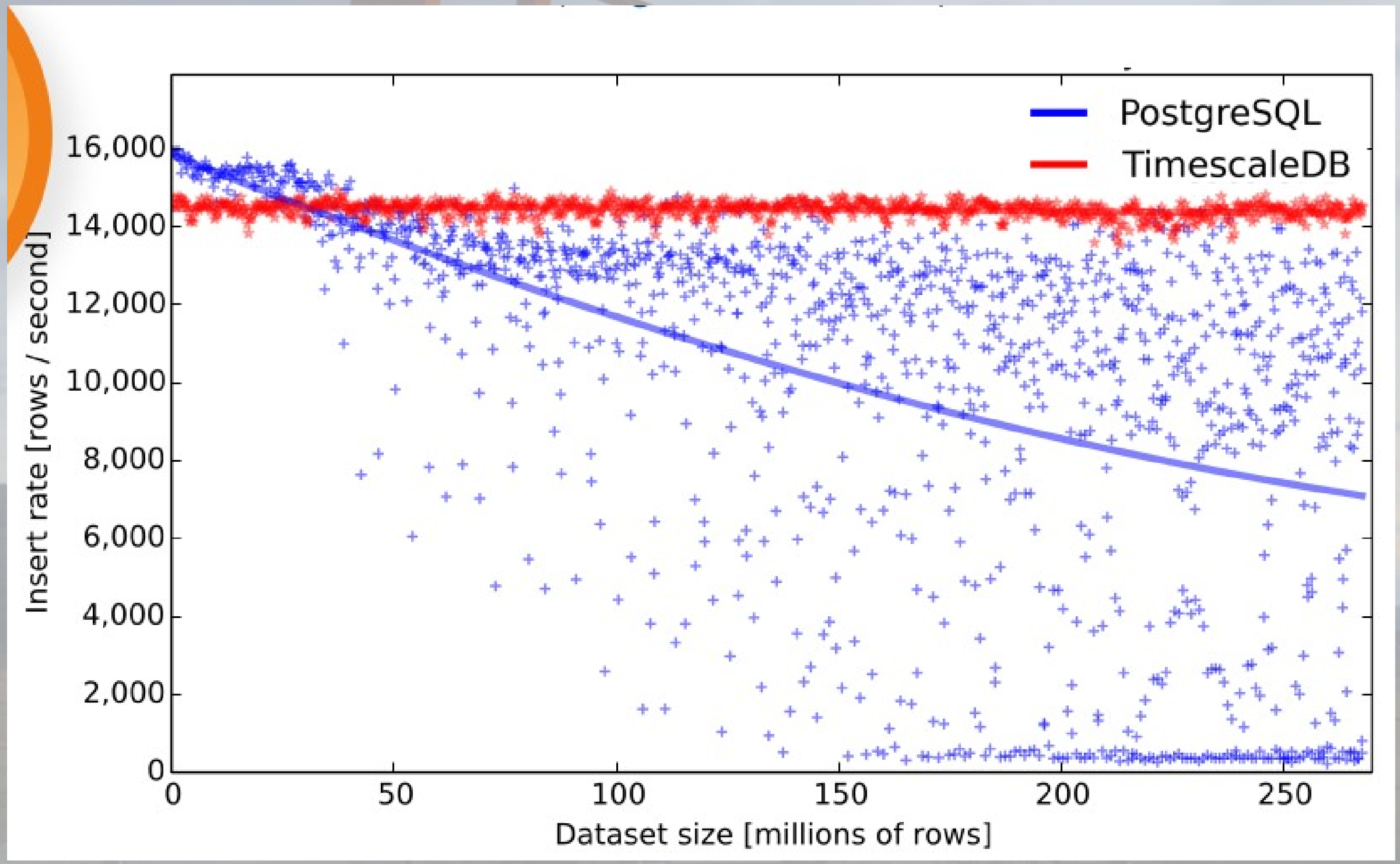


Time series DB



astroparticle.online

Time series DB



CURRENT STATUS

- ✓ Used CernVM-FS to export the existing data storage of each site as is without changing the file system
- ✓ Merged different data trees to a single one at the aggregation server level
- Metadata search and API (in progress)
- Access policy (in progress)

FUTURE WORK

- Sub-tree export (build a CVM-FS middleware module or an independent bridging module?)
- Data access policy and API (RESTful API or GraphQL?)
- Metadata indexing and parameterized search (Time series DB)
- Benchmark

THANK YOU!

QUESTIONS?