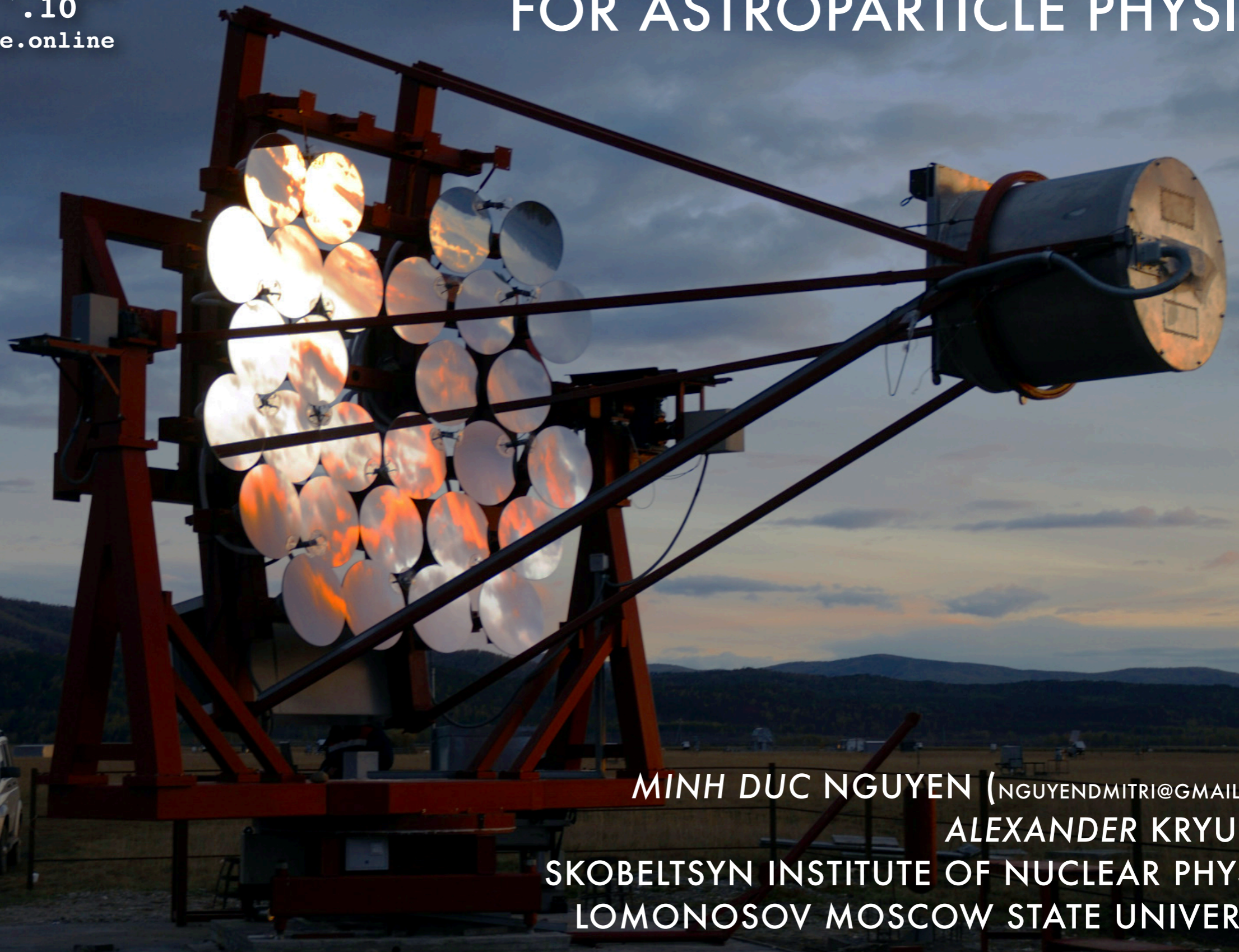


A DISTRIBUTED DATA WAREHOUSE SYSTEM FOR ASTROPARTICLE PHYSICS



MINH DUC NGUYEN (NGUYENDMITRI@GMAIL.COM)

ALEXANDER KRYUKOV

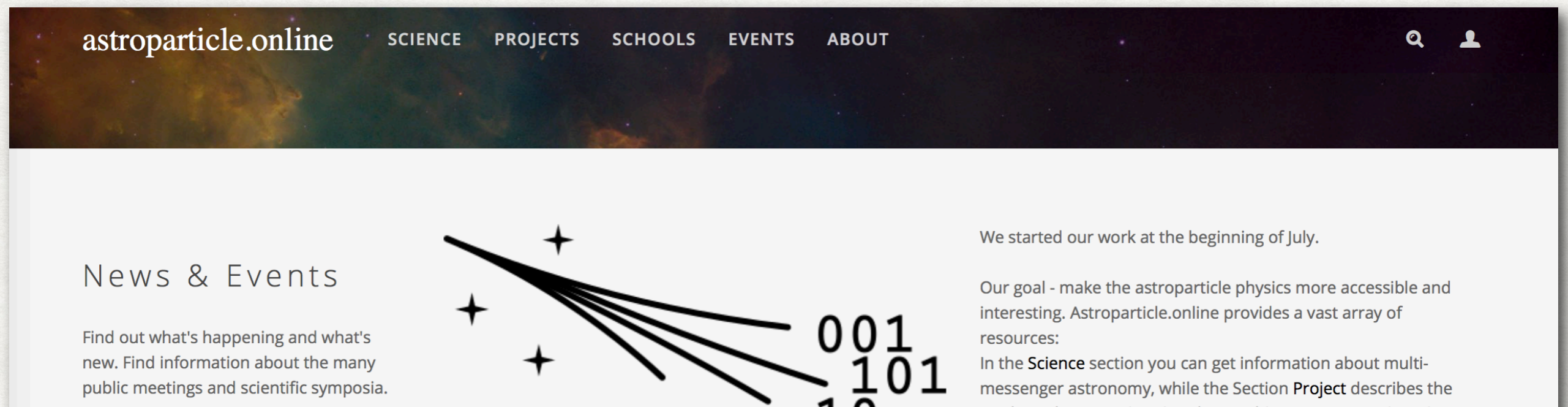
SKOBELTSYN INSTITUTE OF NUCLEAR PHYSICS
LOMONOSOV MOSCOW STATE UNIVERSITY

GRID 2018, September 10 - 14, DUBNA

Supported by RSF #18-41-06003

ASTROPARTICLE.ONLINE

- Karlsruhe-Russian Astroparticle Data Life Cycle Initiative
- Supported by RSF and Helmholtz
- Participants: SINP MSU, ISU, ISDCT SB RAS, KIT



The screenshot shows the homepage of the website astroparticle.online. The header features the site name and navigation links for SCIENCE, PROJECTS, SCHOOLS, EVENTS, and ABOUT. A search icon and a user profile icon are also present. The main content area includes a 'News & Events' section with a sub-header and a paragraph of text. To the right, there is a graphic with binary code (001, 101) and a stylized representation of particle tracks or data lines. Below the graphic, there is a paragraph of text starting with 'We started our work at the beginning of July.' and another paragraph starting with 'Our goal - make the astroparticle physics more accessible and interesting. Astroparticle.online provides a vast array of resources:'. The text continues with 'In the Science section you can get information about multi-messenger astronomy, while the Section Project describes the'.

astroparticle.online SCIENCE PROJECTS SCHOOLS EVENTS ABOUT

News & Events

Find out what's happening and what's new. Find information about the many public meetings and scientific symposia.

We started our work at the beginning of July.

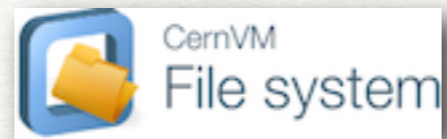
Our goal - make the astroparticle physics more accessible and interesting. Astroparticle.online provides a vast array of resources:

In the **Science** section you can get information about multi-messenger astronomy, while the **Section Project** describes the

REQUIREMENTS FOR THE DATA WAREHOUSE

- Multiple experiments (TAIGA, KASKADE, etc.)
- More than hundreds of terabytes 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

POSSIBLE SOLUTIONS

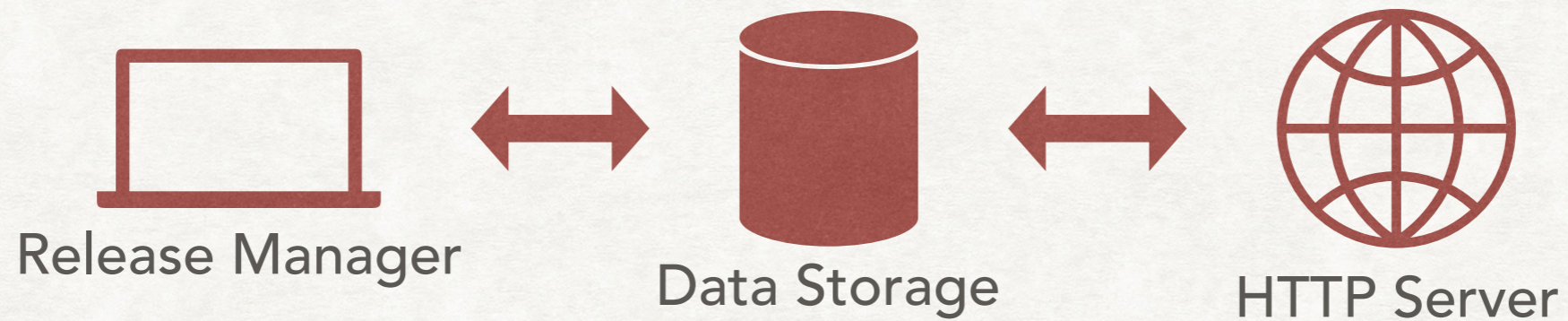


CERNVM-FS

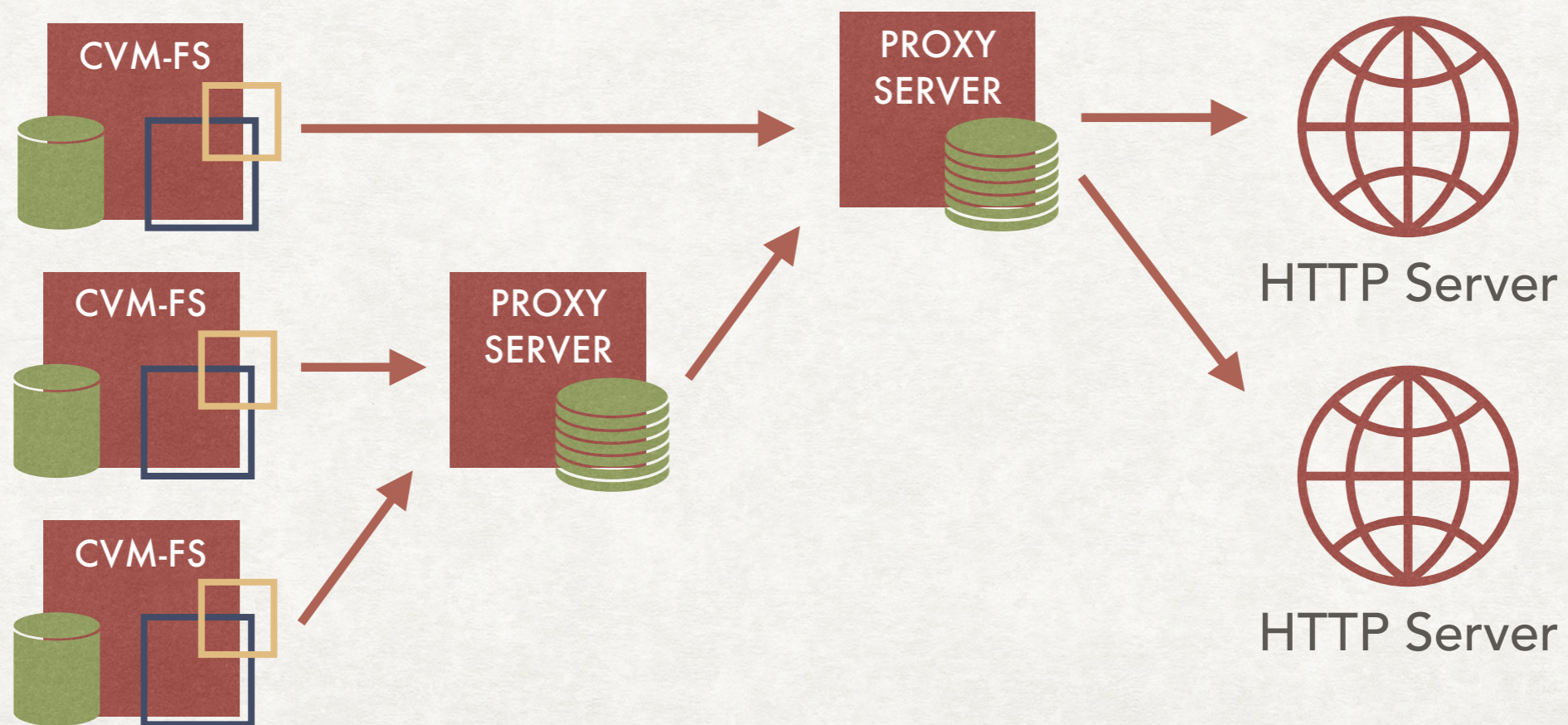
- Data are left untouched in their own file system
- CernVM-FS indexes the data and changes, stores only the metadata (indices, checksums, locations, etc.) and data tree
- CernVM-FS uses HTTP as the data transfer protocol, so there's no firewall problem
- Data transfer starts only on actual reads
- Multilevel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

CERNVM-FS

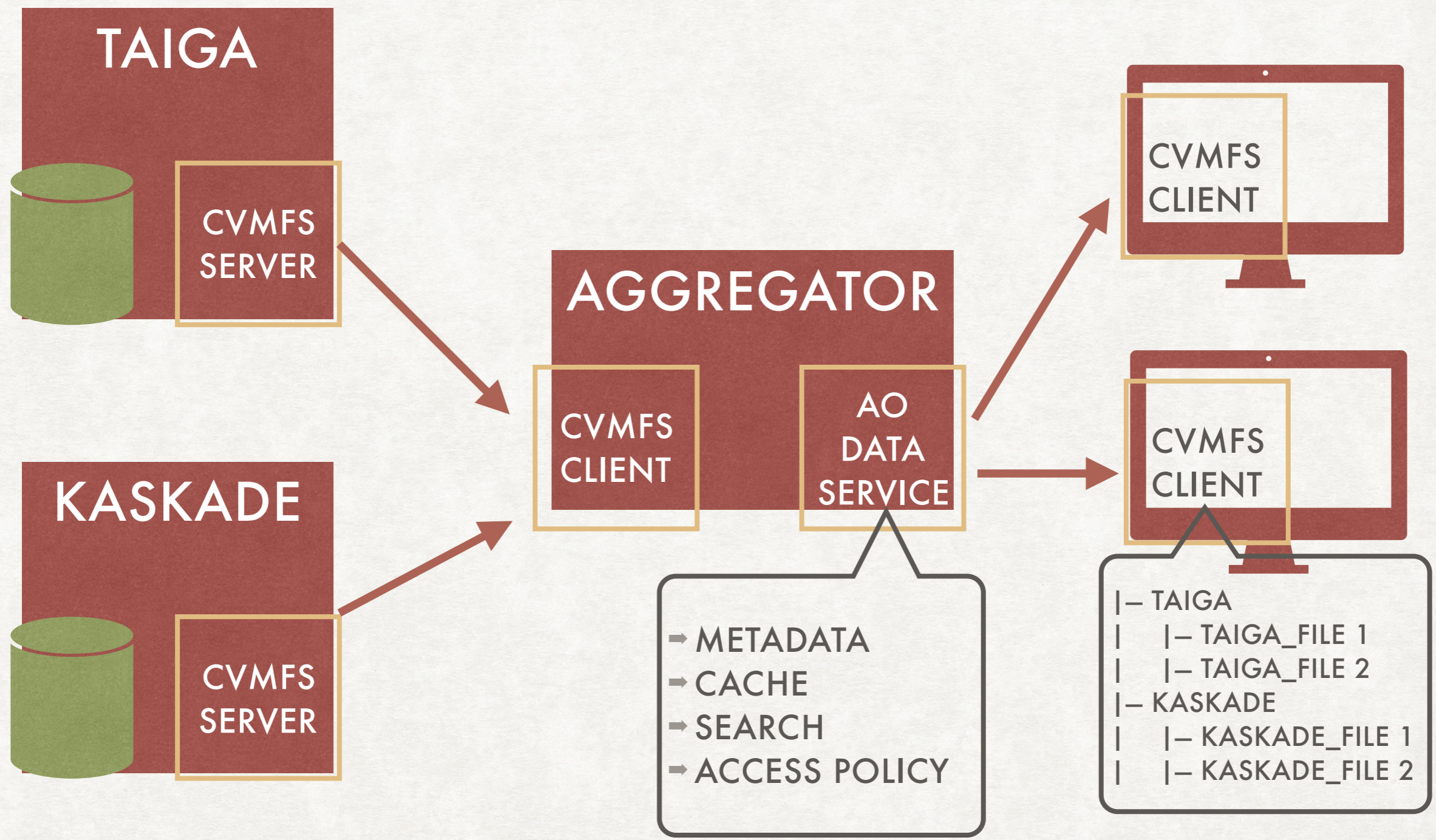
DATA UPDATE



DATA DISTRIBUTION



ASTROPARTICLE.ONLINE + CERNVM-FS



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. Currently, the whole data tree is accessible for everyone)

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 parameterised search (RDBMS (PostgreSQL) or NoSQL (column-based or row-based)?)
- HDFS-prototype and AFS-prototype
- Benchmark

“

THANK YOU!

— *Minh Duc Nguyen* <nguyendmitri@gmail.com>

”