Off-line data storage and processing system for TAIGA experiment

A.Kryukov (SINP MSU)

kryukov@theory.sinp.msu.ru

Outlook

- Introduction to TAIGA gamma-observatory.
- Data flow in TAIGA (sketch)
- Current off-line computing facilities
  - Hardware
  - Software
- Nearest plan of deployment
- Brief presentation joint RSC-Helmholtz project
- Conclusions

TAIGA gamma-observatory

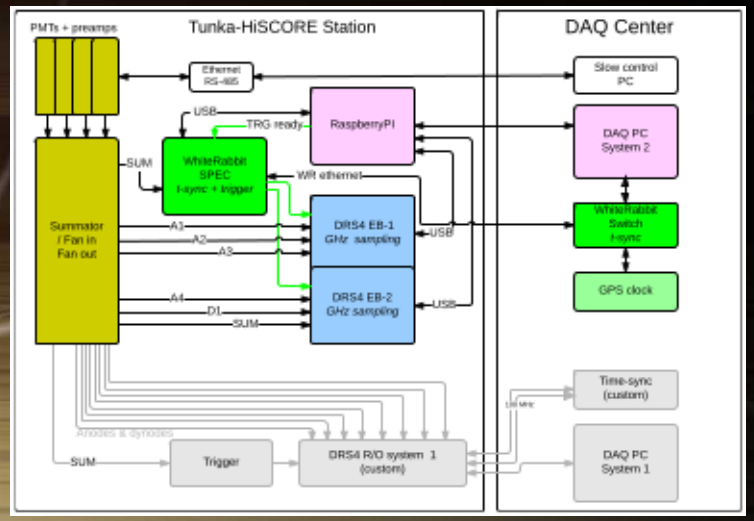
- TAIGA [1] stands for “Tunka Advanced Instrument for cosmic ray physics and Gamma Astronomy” and is a complex, hybrid detector system for ground-based gamma-ray astronomy from a few TeV to several PeV, and for cosmic ray studies from 100 TeV to several 100′s of PeV.
- 500 wide angle optical stations on the 5 km 2 area, energy threshold 30 TeV
- Up to 16 IACT (10 m^2 mirrors).

TUNKA DAQ system (sketch)

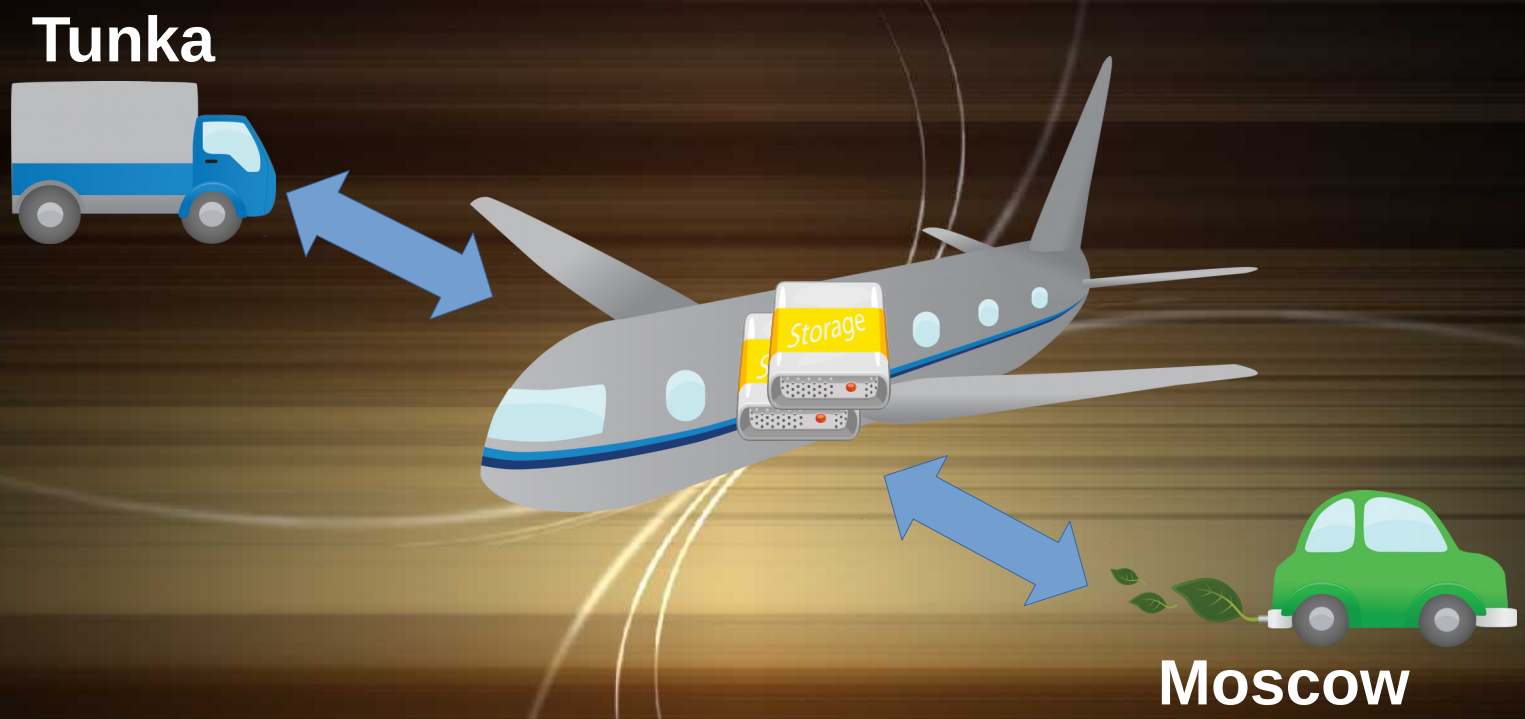
- 28 stations have 8 channels per each.
- One event is 8KB of 20Hz

  28*8*20
  or is about **5 MB/s** raw data
- If effective exposition time is 500 hours per year the total number of raw data is

  5*3600*500/10^6 which is about **10TB** raw data per year
- Processed and simulation data require about double and more amount of data



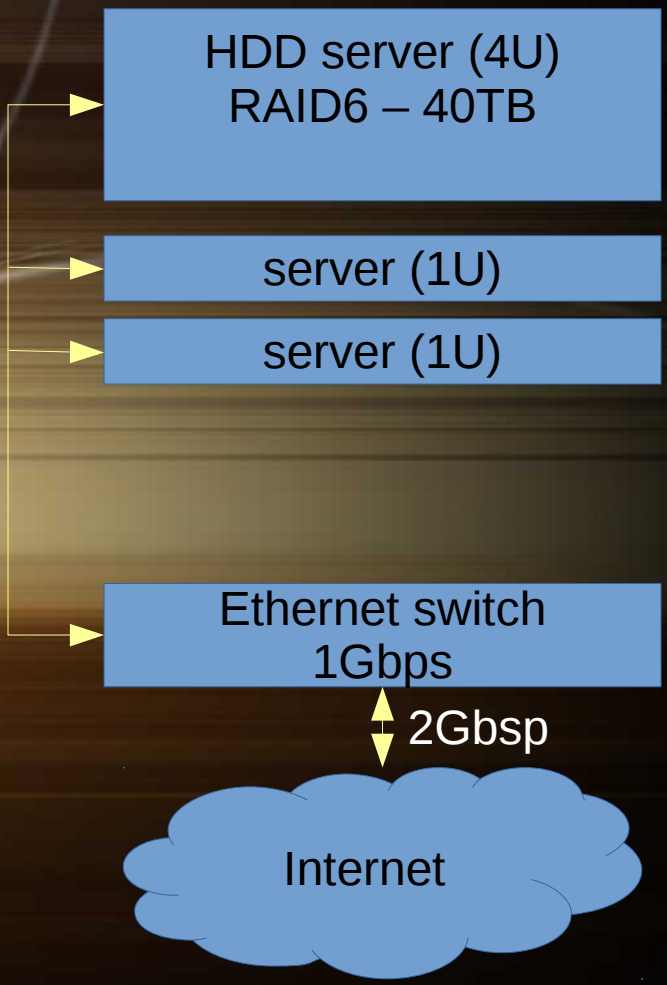
Tunka-Moscow High Speed Data Channel



- Transportation of 6TB HDD during 1 month is equivalent 3HDD*2TB/(3600*24*30) ~ **20Mbps**
- However, increasing the amount of data until 100TB/year closes this way

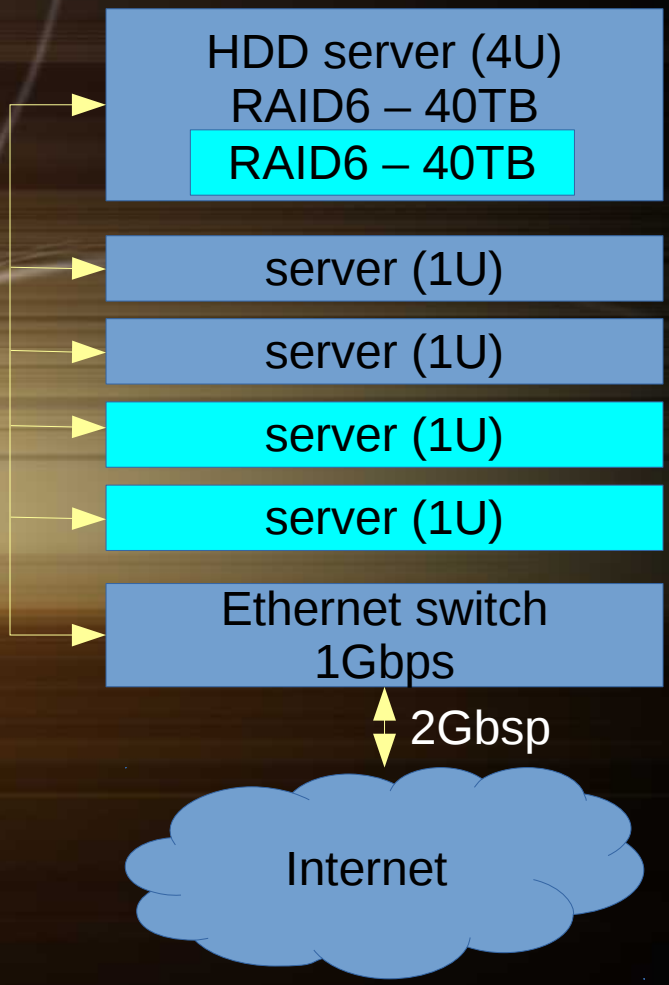
Current status of off-line data storage and processing system

1) HDD server (4U)
   - RAID6, 40 TB
   - **New occupy about 80%**
2) 2 computing server
   - 2 Intel CPU, 2.3 GHz, 4 cores
   - RAM – 16GB
   - Ethernet – 1 Gbps
3) Ethernet switch HP
   - 64 ports, 1 Gbps
   - 4 optical ports un to 10 Gbps



Off-line data storage and processing system in 2017-2018

1) HDD server (4U)
   - +RAID6, 40 TB
2) +2 computing server (mini cluster)
   - 2 Intel CPU, 2.3 GHz, 4 cores
   - RAM – 16GB
   - Ethernet – 1 Gbps

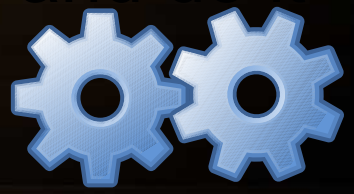


Software

- Now:
  - CentOS-6
  - FTP (over ssh) server to distribute data
  - NFS for cluster
  - GCC family compilers (C/C++, F90/F95)
  - PAW, CORSIKA
- 2018:
  - NextCloud – cloud storage
  - Torque batch system
  - Git version and collaboration system
    - Also may be GitLab
  - ROOT

Software improvements

- FORTRAN-77
- There were made improvements of the using software.
  - As a results the speed of data processing was increase more then 10 times
- There was redesigned the Fortran program.
  - Now all work with data, calibrations and other files eliminate to the bash scripts.
  - This is to simplify logic of Fortran program and do it more clear and modifiable.



Karlsruhe-Russian Astroparticle Data Life Cycle Initiative

- The application was submitted to the joint competition of the RNF and the Helmholtz Association. The result of the competition will be announced in September 2017.
- Main participants:
  - Russia: SINP MSU, ISU, ISDCT SB RAS
  - Germany: KIT
  - Team leaders: A. Kryukov (SINP MSU) and A. Haungs (KIT)
- Duration: 2018-2020
- Financial request
  - RSF – 18*10^6 Rub.
  - Helmholtz – 390*10^3 Euro

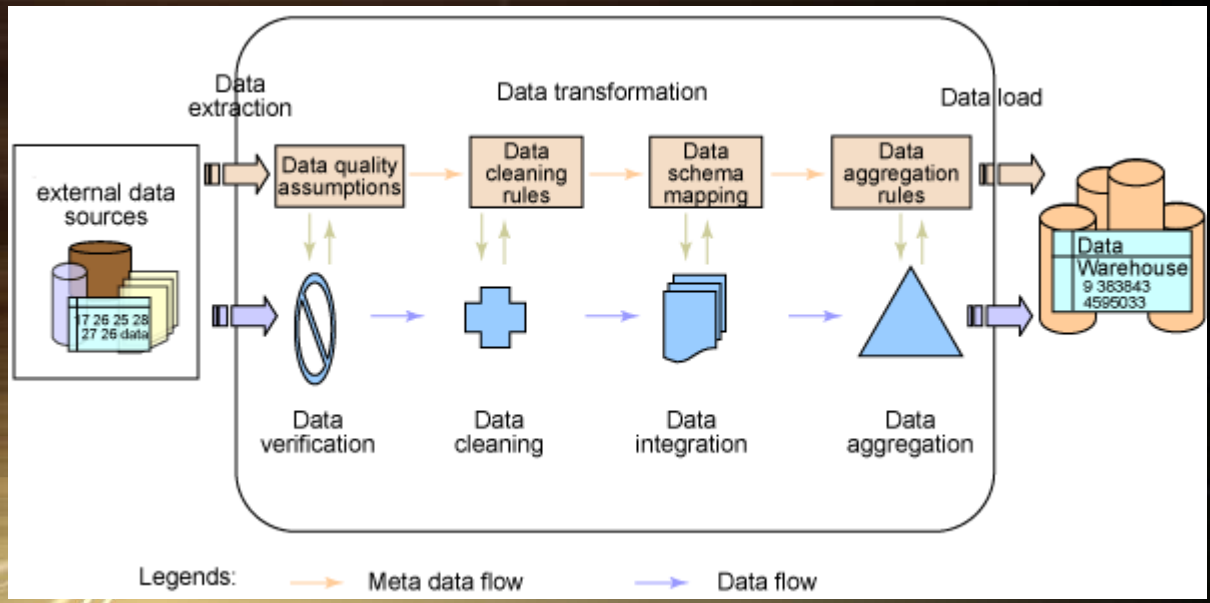
Data life cycle [2,3]

- Data created/collected
- Data shared/processed
- Data analyzed
- Data published
- Data archived
- Data re-used

Data Center in Astroparticle Physics

Data availability | Analysis | Simulations & Methods development | open access | Education in Data Science | Data archive

Extract-Transform-Load systems[4]



- Data extraction
  - data verification
- Data transformation
  - data cleaning,
  - data integration
- Data load
  - data aggregation

Main targets of the project

- The project will strive to develop an open science system to be able to collect, store, and analyze astrophysical data having the TAIGA and KASCADE experiments as the examples.
- The novelty of the proposed approach can be seen in developing integrated solutions including:
  - development and adaptation of distributed data storage algorithms and techniques with a common meta-catalog to provide a common information space of the distributed repository;
  - development and adaptation of data transmission algorithms as well as simultaneous data transmission from several data repositories thus significantly reducing load time;

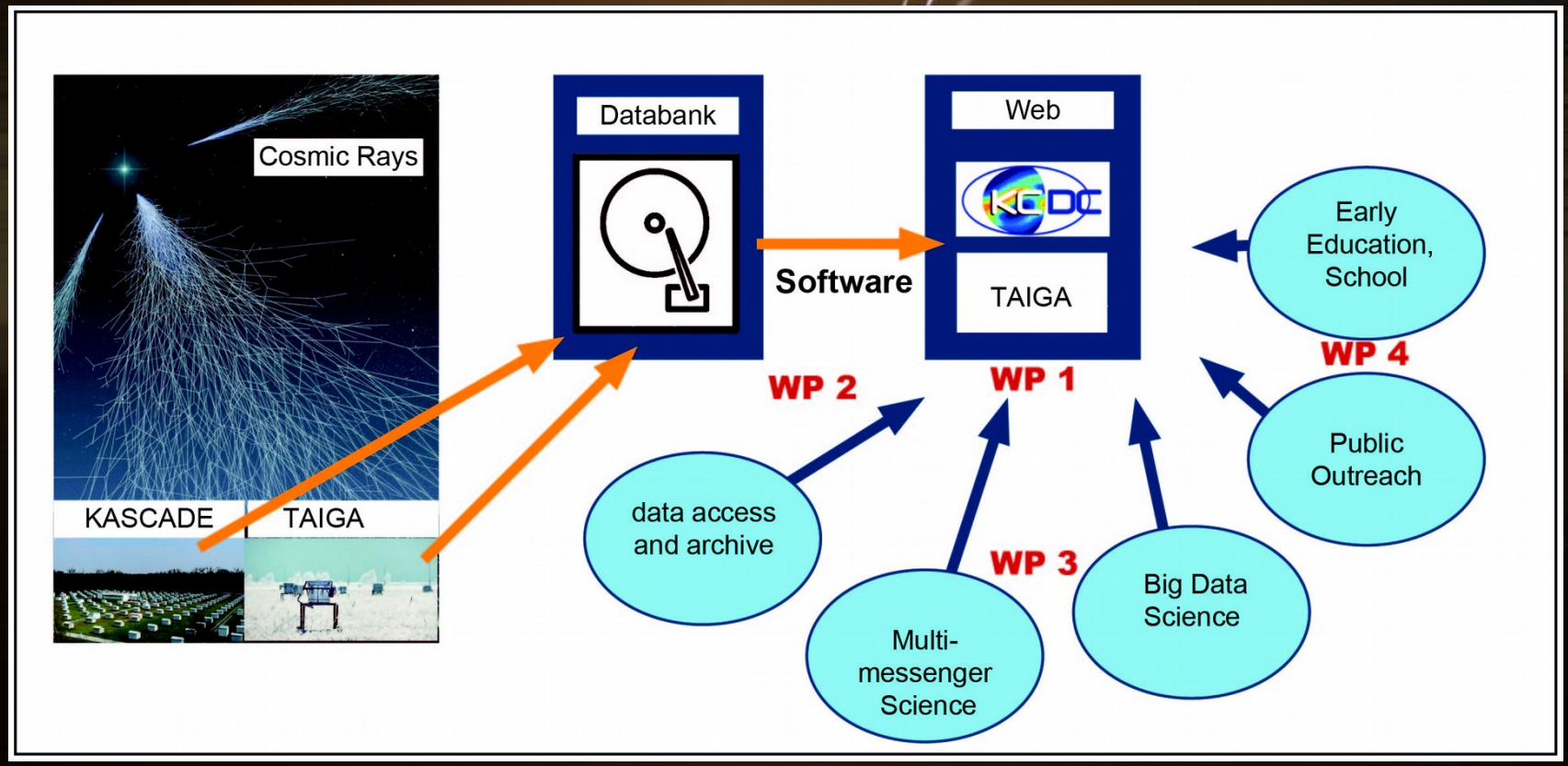
Main targets of the project (cont.)

- development of machine-learning techniques for identifying mass groups of particles and their properties in a fully remote access mode;
- deployment of the KCDC-based prototype system of Big Data analysis and exporting the experimental data from KASKADE and TAIGA for testing technology of data life cycle management.

- We will also create an educational subsystem on the HubZero platform dedicated to astroparticle physics.

KASCADE Cosmic Ray Data Centre

- KCDC [5] is the installation and establishment of a public data centre for high-energy astroparticle physics based on the data of the KASCADE experiment.
- The KCDC system will be use as a prototype of future system



The novelties in the project

- This is an innovative approach that will be used in astroparticle physics research for the first time.
- Plans are underway to expand the number of experiments by exporting data from other scientific collaborations
  - it will rapidly advance the research of fundamental properties of matter and the universe.
- It's noteworthy that the suggested approach can be used not only in the specified field of science but also adapted to other scientific disciplines.

Expected main results

- A distributed system for the large astrophysical data collecting and processing on the basis of the existing KCDC system will be created.
  - The main idea to achieve this goal is the concept of so called "data life cycle lab".
- Software for the big data intelligent analysis in particle astrophysics.
- A methodology for the verification of the scientific results reliability based on the comprehensive data analysis of many types and from many sources will be developed.
- Open data access for the scientific community.

Conclusions

- Our off-line computing facilities ready to collect, store and make physics analysis in 2017 year.

- We can increase our facilities twice and more depends on available money

- The Karlsruher-Russian project will open new horizon of the computing in astroparticle.

References

1. Budnev, N.N. and etc. TAIGA the Tunka Advanced Instrument for cosmic ray physics and Gamma Astronomy — present status and perspectives, Journal of Instrumentation, Vol. 9, Sep. 2014
2. Boliang He and etc. AstroCloud, a Cyber-Infrastructure for Astronomy Research: Data Archiving and Quality Control, ArXiv:1411.5071
3. Jos van Wezel and etc. Data Life Cycle Labs, A New Concept to Support Data-Intensive Science, arXiv:1212.5596
4. P. Vassiliadis. Survey of extract-transform-load technology, Int. J. of Datawarehouse and Mining, 5(3), pp.1-27, 2009
5. KCDC, https://kcdc.ikp.kit.edu/

THANK YOU!