NATIONAL RESEARCH CENTRE
«KURCHATOV INSTITUTE»

PETERSBURG NUCLEAR PHYSICS INSTITUTE

NRC "KI" participation in EU DataLake project

A.K. Kiryanov, A.A. Klimentov, A.K. Zarochentsev

Grid’2018, 10-14 September 2018, Dubna, Russia

A Data Lake - why?

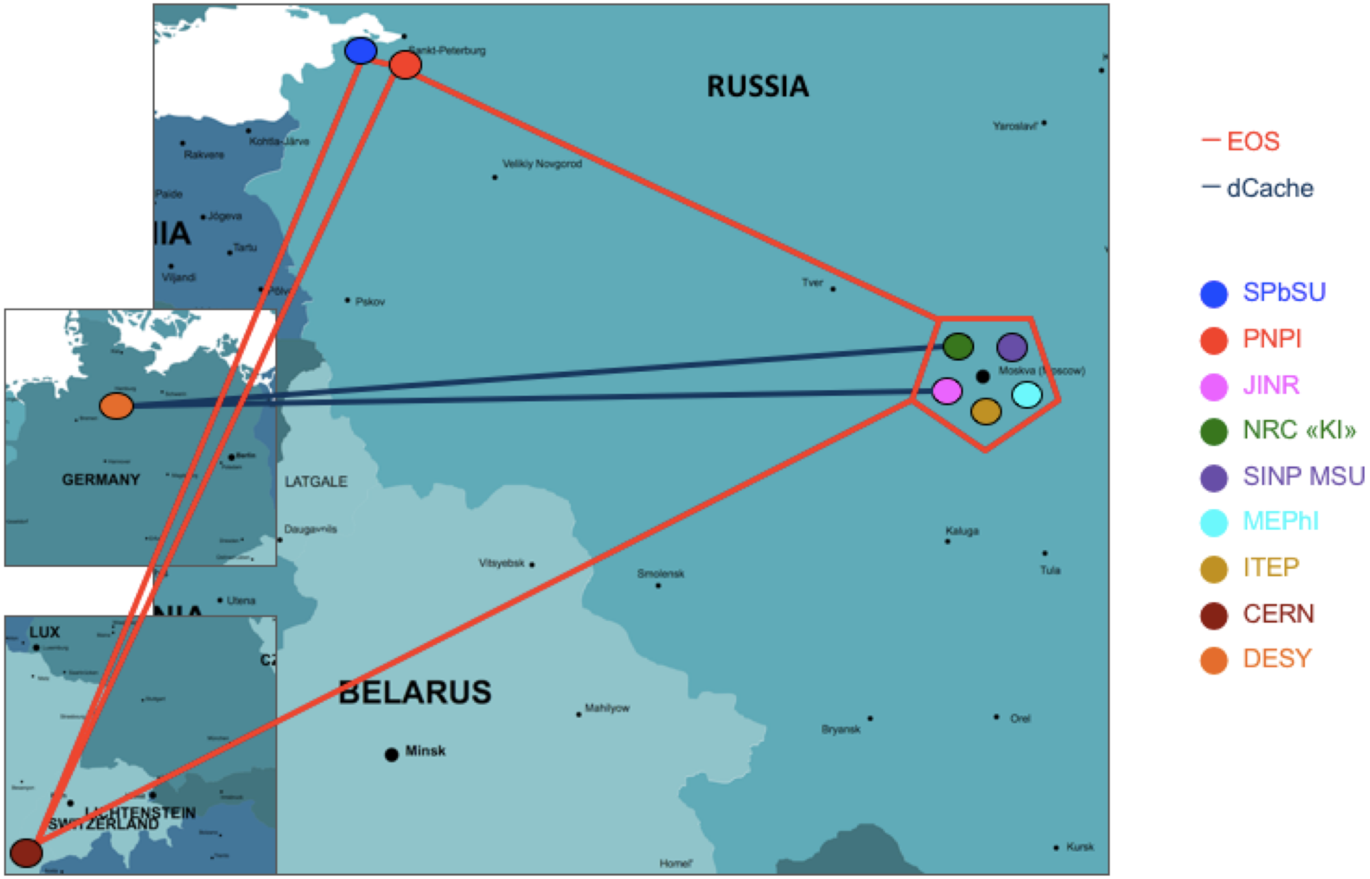
- HL-LHC storage needs are above the expected technology evolution (15%/yr) and funding (flat).

[redacted]

From Xavier Espinal's slides at Joint WLCG and HSF workshop, Napoli, 26-29 March 2018

DataLake – early prototype

In the fall of 2015 the “Big Data Technologies for Mega-Science Class Projects" laboratory at NRC "KI" has started work on a storage federation prototype for geographically distributed data centers located in Moscow, Dubna, St. Petersburg, and Gatchina (all are members of Russian Data Intensive Grid and WLCG).



- A working prototype was developed with multiple storage technologies (EOS, dCache)
- A set of metrics was established along with a toolchain and a testing methodology

WLCG DataLake R&D project goals

Explore distributed storage **evolution** to improve overall costs (storage and ops):

- Common namespace and interoperability
- Co-existence of different QoS (storage media)
- Geo-awareness
- File transitioning based on namespace rules
- File layout flexibility
- Distributed redundancy

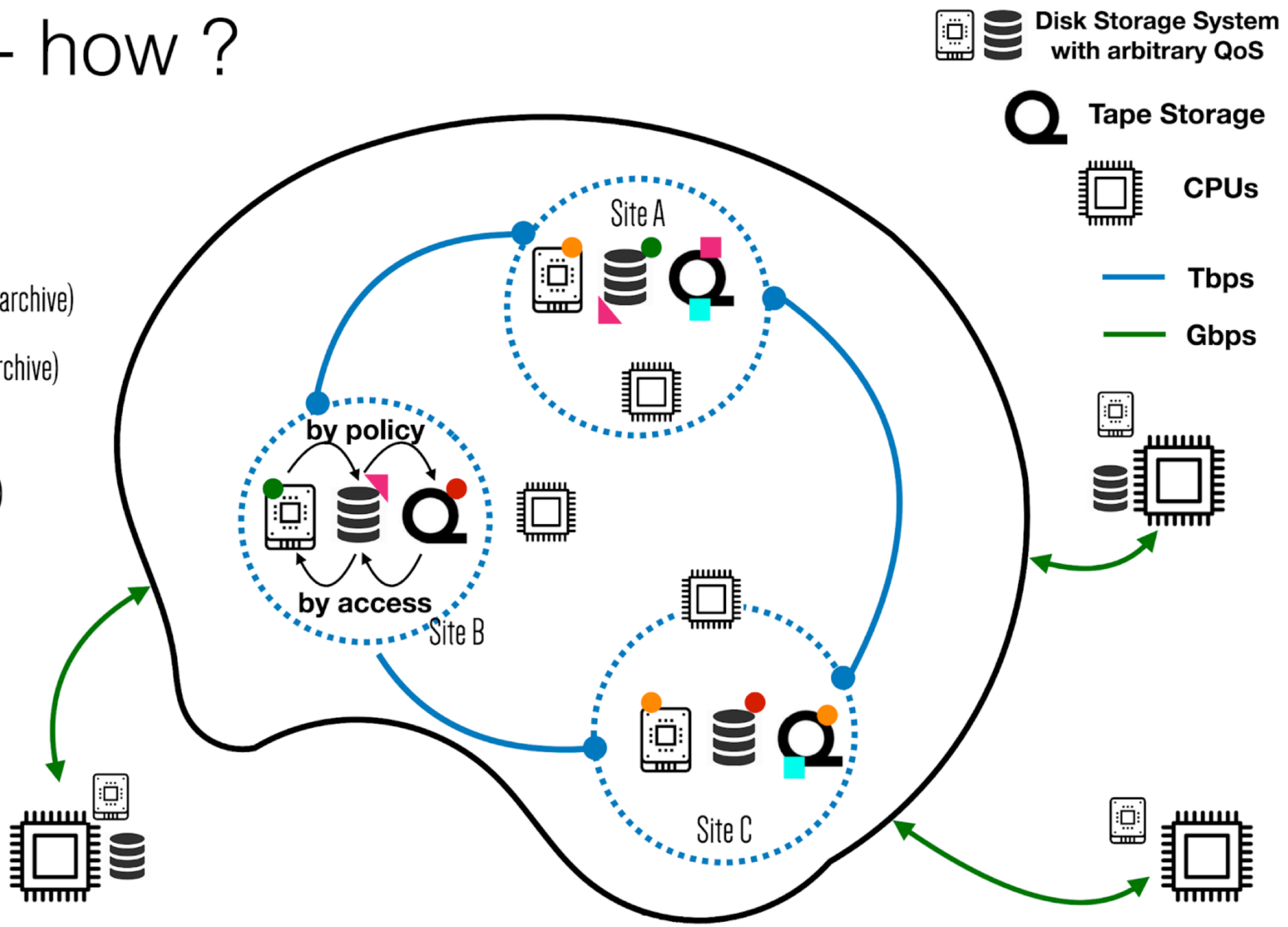
R+D project aims to **demonstrate** that a dynamically distributed storage system with a common namespace:

- Has the potential to lower the cost of stored data
- Has the potential to ease local administration and world-wide operations
- Has the acceptable efficiencies in performance, reliability and resilience
- Is compatible with HL-LHC computing models

Data Lake - how ?

File placement by QoS

- ● Hot custodial file (2 fast copies+archive)
- ● Warm custodial file (disk copy+archive)
- ■ Cold custodial file (archive)
- ● Hot ephemeral file (2 fast copies)
- ◩ Warm ephemeral file ("Rain")



From Xavier Espinal's slides at Joint WLCG and HSF workshop, Napoli, 26-29 March 2018

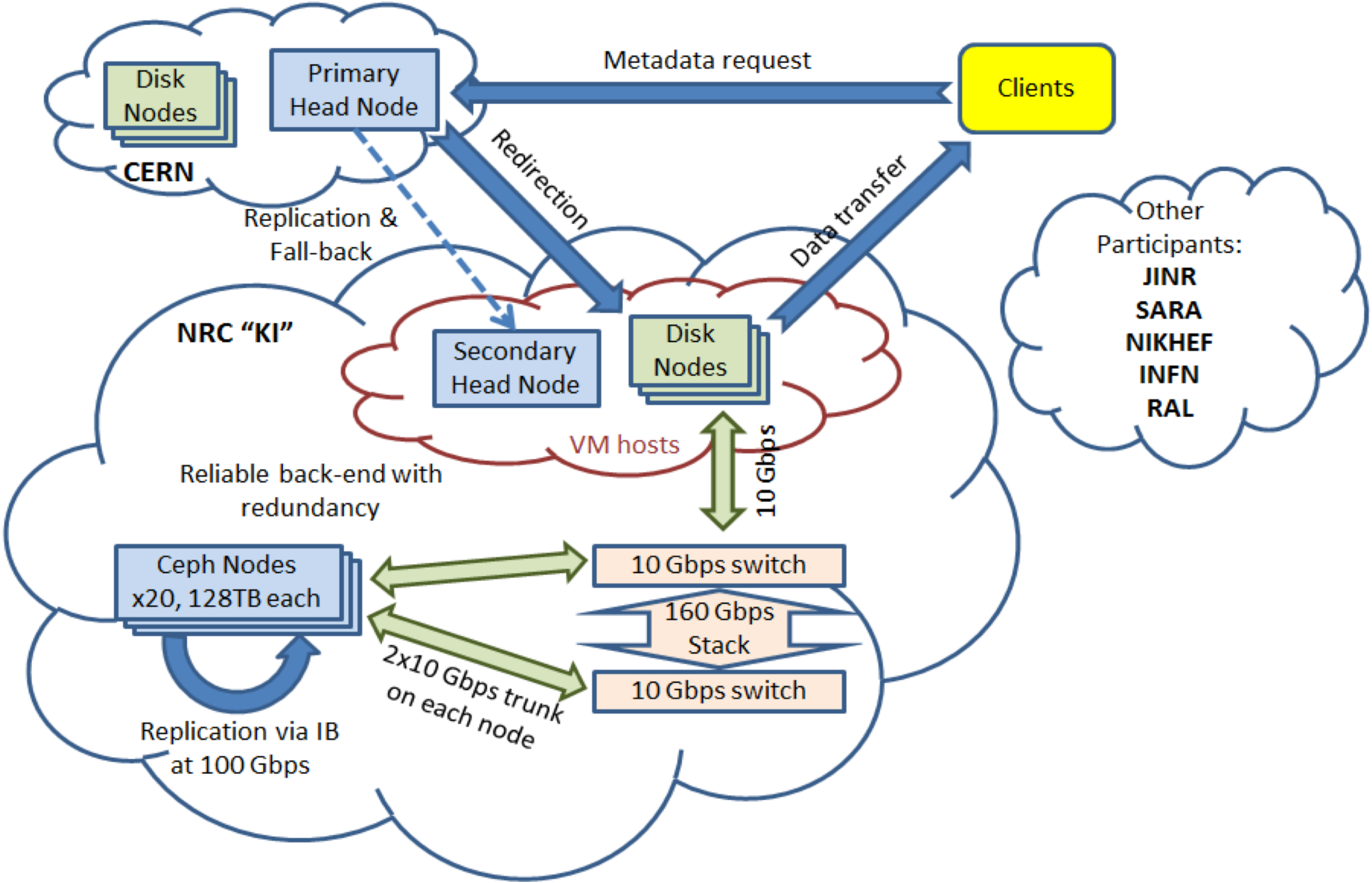
WLCG DataLake prototype (eulake)

- Prototype is based on EOS
  - Other storage technologies and their possible interoperability are also considered
- Primary namespace server (MGR) is at CERN
  - Secondary namespace server will be deployed at NRC “KI”
- Storage endpoints run a simple EOS filesystem (FST) daemon
  - Storage endpoints deployed at SARA, NIKHEF, RAL, JINR, NRC “KI”, PIC, CNAF and Aarnet
- perfSONAR endpoints are deployed at participating sites
- Monitoring is hooked up to Grafana
- Performance tests ready to be run in continuous mode

NRC “KI” participation in eulake

- Why?
  - Extensive expertise in deployment and testing of distributed storages
  - A similar prototype was successfully deployed on Russian sites (Russian Federated Storage Project) – still alive!
  - An appealing universal storage technology may be useful not only for HL-LHC and HEP, but also for other experiments and fields of science (NICA, PIK, XFEL)
  - We just cannot let this happen without us involved
- NRC “KI” equipment for eulake is located at PIK Data Centre in Gatchina
  - 10 Gbps connection, IPv6
  - 100 TB of Ceph storage as a backend
  - EOS endpoints on VMs

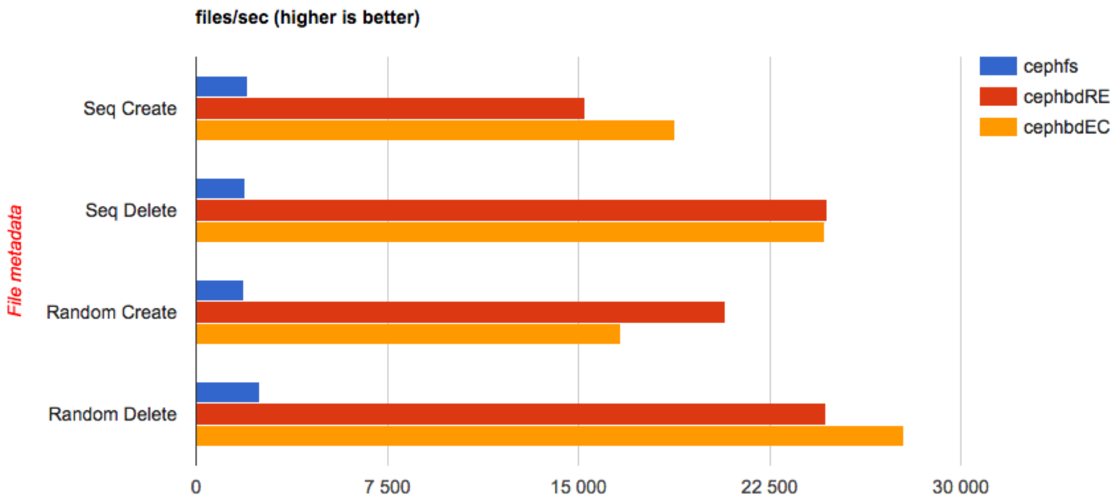
NRC “KI” equipment for eulake



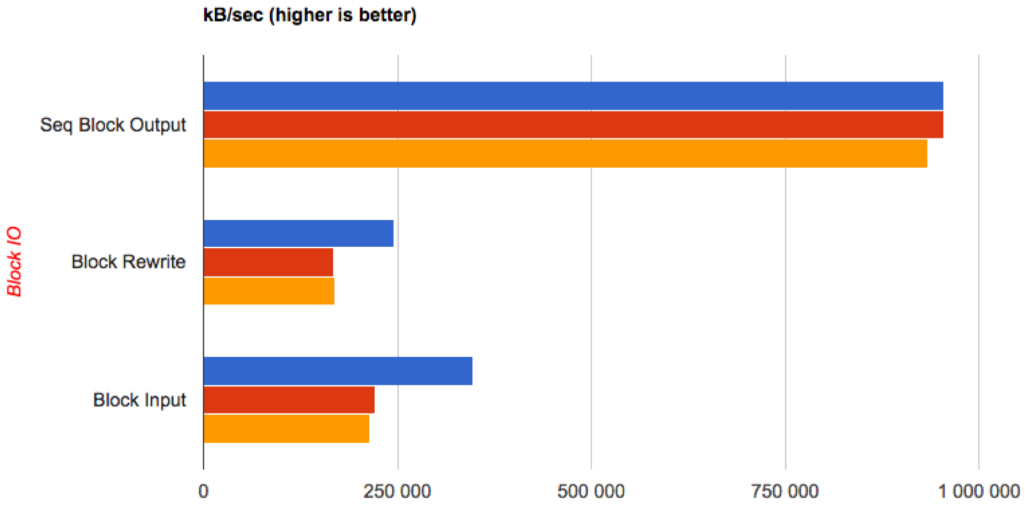
Highlights

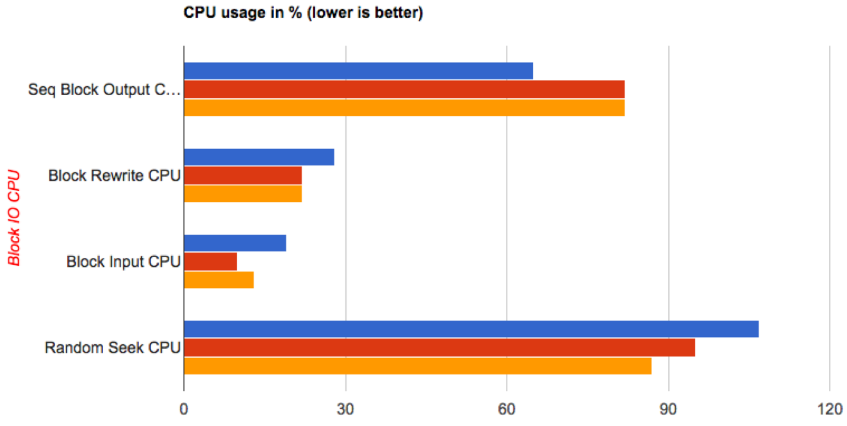
- Why Ceph?
  - Deploying EOS on physical storage is perfectly suitable for CERN, *but*
  - PIK Data Centre is not a dedicated facility for HEP computing
  - Ceph adds necessary flexibility in storage management as we also use it for other purposes
  - Out of 2.5 PB of Ceph storage we provide 100 TB for eulake prototype
- Storage configuration
  - We have started with Luminous but quickly moved to Mimic
    - CephFS performance improved significantly in the new release
  - We have four different “types” of Ceph storage exposed to EOS:
    - CephFS with replicated data pool
    - CephFS with Erasure Coded data pool
    - Block device from a replicated pool
    - Block device from an Erasure Coded pool
  - Functional and performance tests are ongoing
- Auxiliary infrastructure
  - Repository with stable EOS releases (CERN repo changes too fast, sometimes breaking the functionality)
  - Web server with a visualization framework and a test results storage

Ceph performance measurements



- Metadata performance of CephFS is much slower than of a dedicated RBD (this is expected)
- Block I/O performance is on par, but CPU usage is lower with CephFS





Ultimate plans

- Evaluate the fusion of local (Ceph) and global (EOS) storage technologies
  - Figure out the strong and weak points
  - Come out with a high-performance, flexible yet easily manageable storage solution for major scientific centers participating in multiple collaborations
  - Further plans on testing converged solutions (compute + storage). More details in David’s CHEP 2018 talk: https://indico.cern.ch/event/587955/contributions/2937728/
- Evaluate DataLake as a storage platform for Russian scientific centers and major experiments
  - NICA, XFEL, PIK
  - Possibility to have dedicated storage resources with configurable redundancy in a global system
  - Geolocation awareness and dynamic storage optimization
  - Data relocation & replication with a proper use of fast networks
  - Federated system with inter-operable storage endpoints based on different solutions (EOS + dCache?)

ATLAS+Google DataOcean

- An R&D project for evaluating and adopting modern IT technologies
  - Allow ATLAS to explore the use of different computing models to prepare for HL-LHC
  - Allow ATLAS user analysis to benefit from the Google infrastructure
  - Give Google real science use-cases to improve their cloud platform

- More details in Mario’s CHEP 2018 talk:

  https://indico.cern.ch/event/587955/contributions/2947395/

Thank you!