ONEDATA

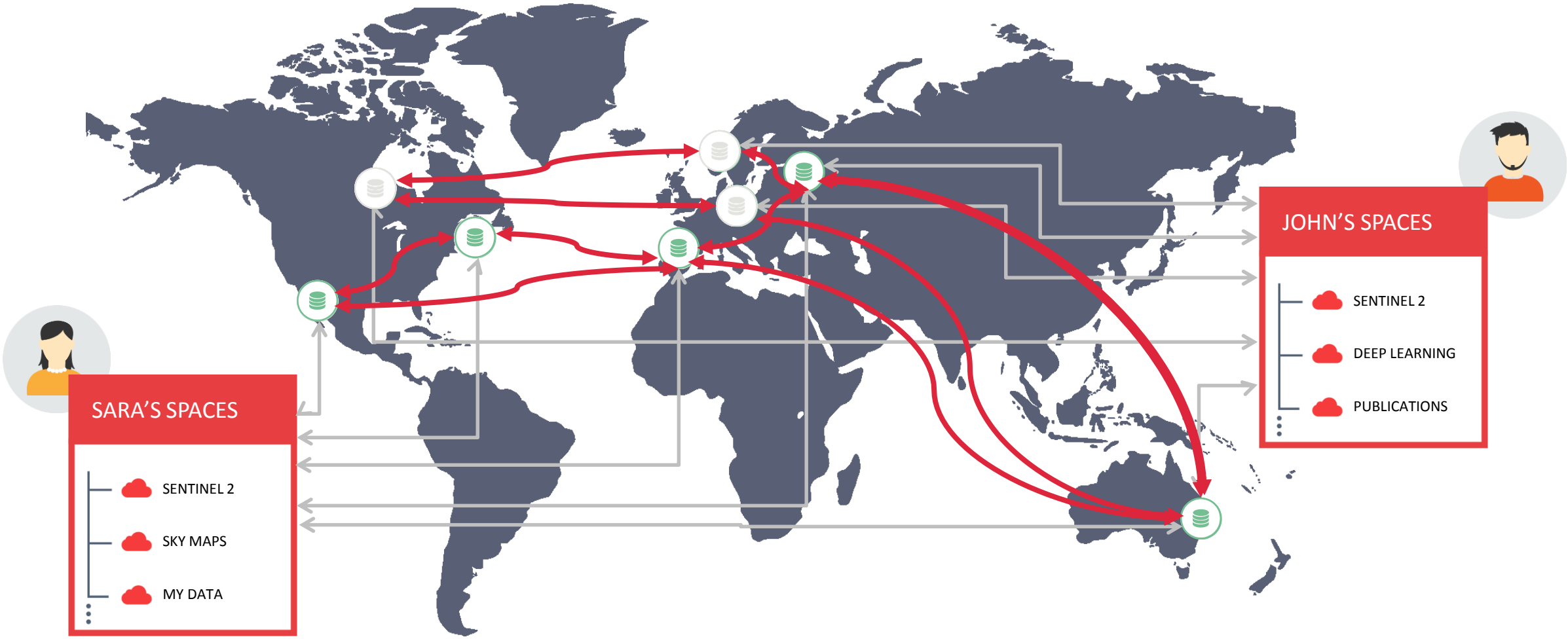
ONEDATA - MANAGING DATA AND METADATA IN HYBRID CLOUDS

Presented by: Lukasz Dutka

WHO WE ARE?

- Group of developers bringing hybrid cloud open source platform to life
- 6+ years devoted development
- Our main goal is:
  - to deliver data management platform for large scale and distributed problems
  - to make the solution decentralized and eventually consistent in order build a mesh of data sources
  - to deliver virtual file system for hybrid cloud
- The work is supported by:

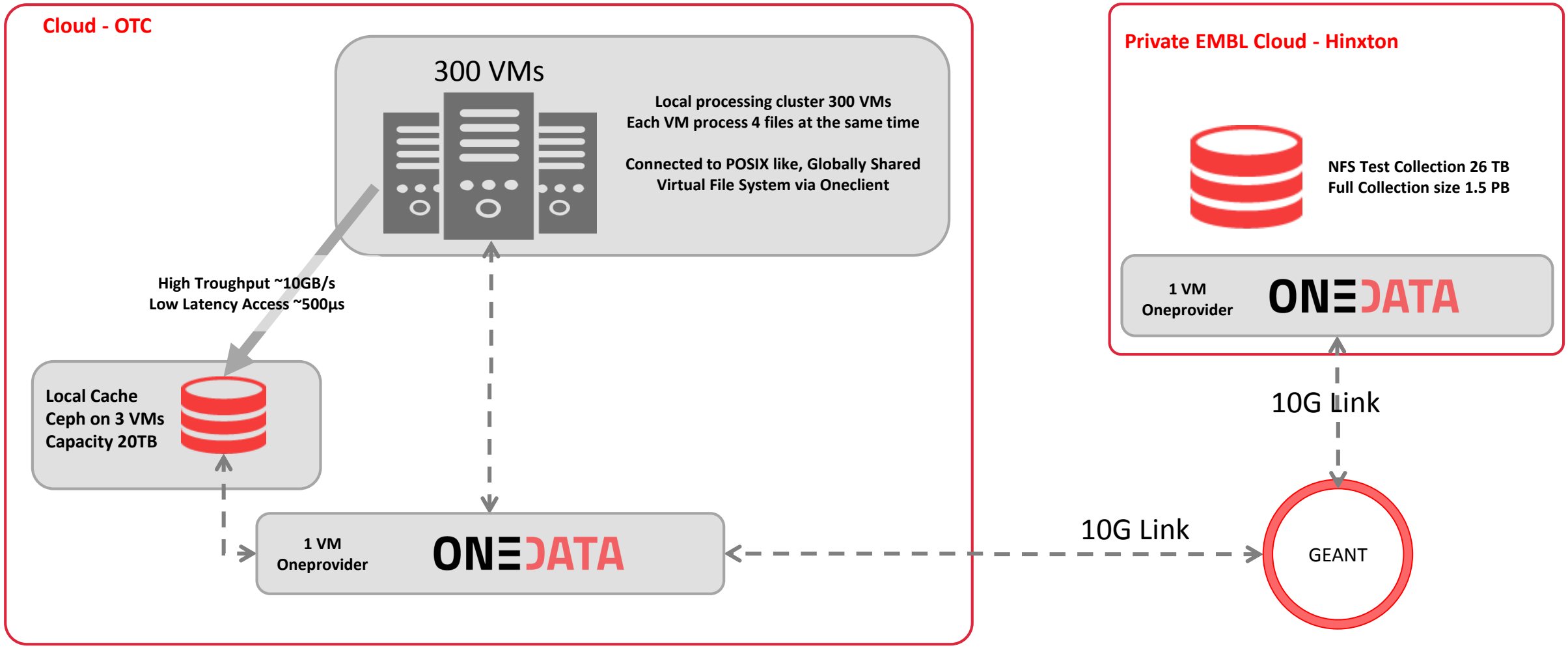
DATA IN HYBRID CLOUD ENVIRONMENTS



MAIN PROBLEMS ADDRESED BY ONEDATA PLATFORM

1. **Lock-in data collection available only locally on local POSIX file system – should be available in multi cloud hybird environments**
2. Multi-protocol transparent access to data “[…] but we want POSIX”
3. High Throughput Data Transfers and Replication
4. **On-the-fly remote data delivery**
5. Replication on demand
6. Heterogeneity of storage technologies
7. **Cache Management**
8. **High-throughput data processing**
9. Authentication and authorization integrated with EDUGAIN

EXAMPLE - EMBL-EBI APPLICATION DEPLOYMENT



REAL DATA TRANSFERS BETWEEN HINXTON – OTC – ON THE FLY DATA DELIVERY FOR 24H OF DATA PROCESSING BY 1200 PARALEL JOBS ON 300 VMs

Experiment Specs:

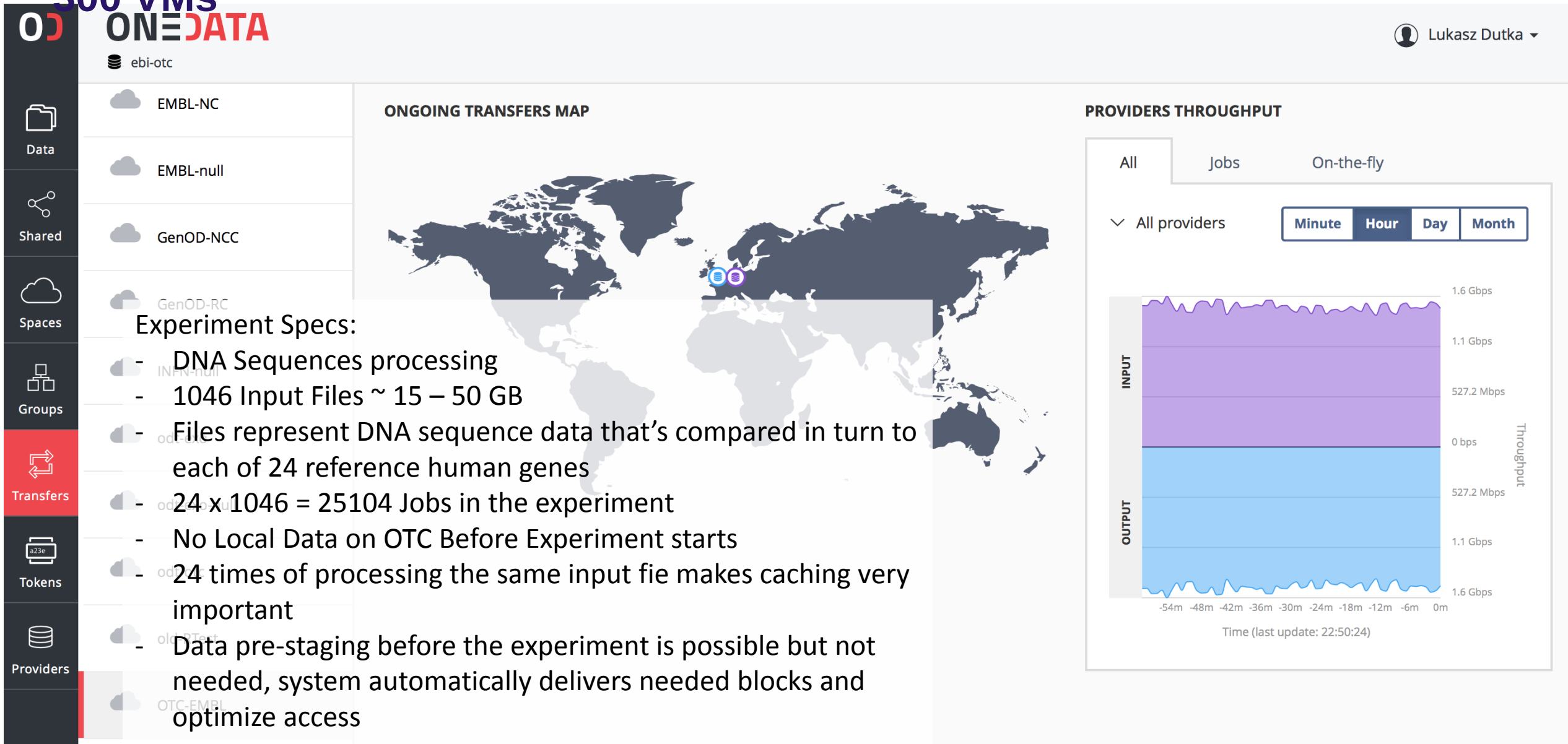
- DNA Sequences processing
- 1046 Input Files ~ 15 – 50 GB
- Files represent DNA sequence data that's compared in turn to each of 24 reference human genes
- 24 x 1046 = 25104 Jobs in the experiment
- No Local Data on OTC Before Experiment starts
- 24 times of processing the same input fie makes caching very important
- Data pre-staging before the experiment is possible but not needed, system automatically delivers needed blocks and optimize access

REAL DATA TRANSFERS BETWEEN HINXTON – OTC – ON THE FLY DATA DELIVERY FOR 24H OF DATA PROCESSING BY 1200 PARALEL JOBS ON 300 VMs

Experiment Specs:

- DNA Sequences processing
- 1046 Input Files ~ 15 – 50 GB
- Files represent DNA sequence data that's compared in turn to each of 24 reference human genes
- 24 x 1046 = 25104 Jobs in the experiment
- No Local Data on OTC Before Experiment starts
- 24 times of processing the same input fie makes caching very important
- Data pre-staging before the experiment is possible but not needed, system automatically delivers needed blocks and optimize access

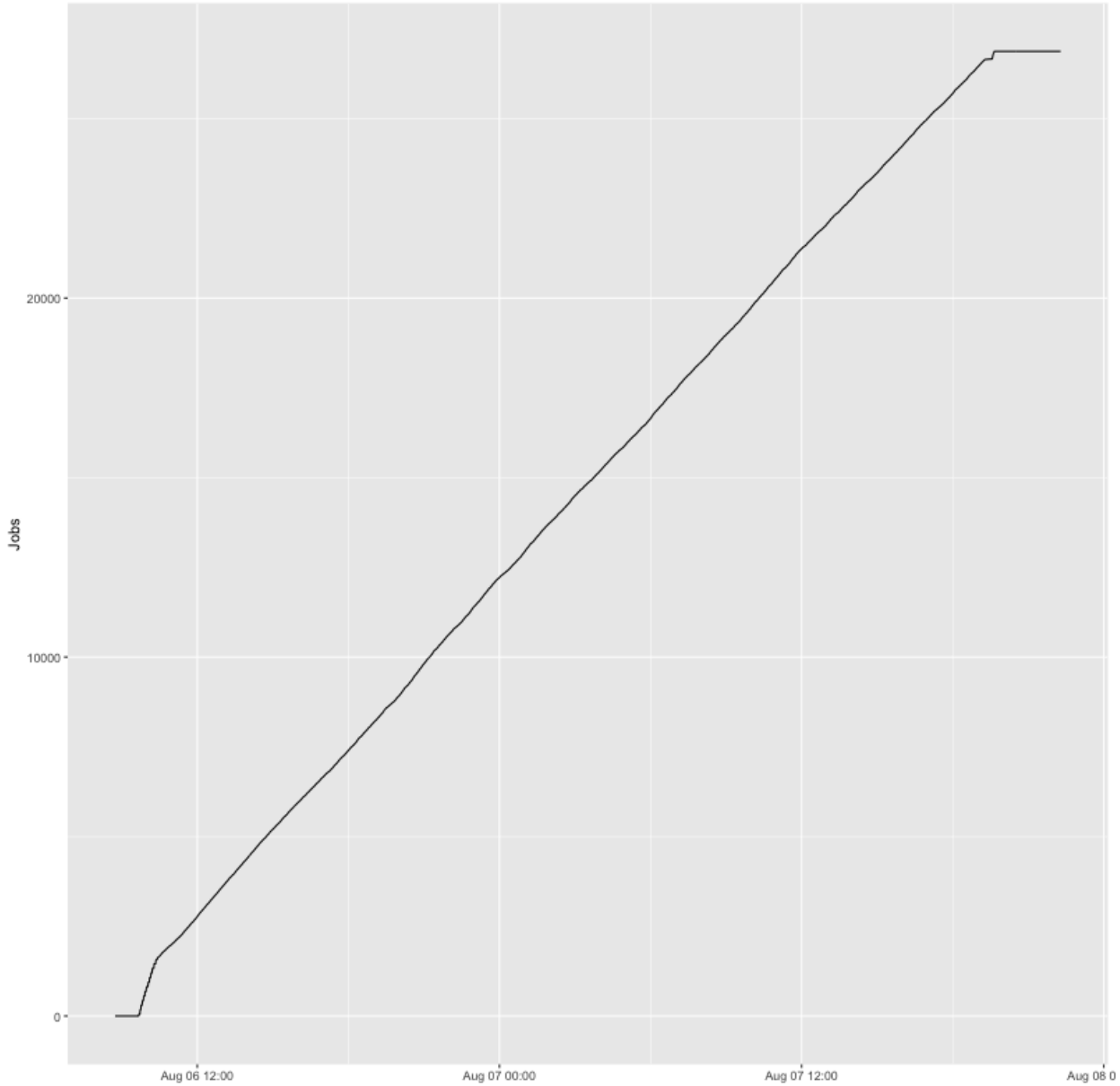
REAL DATA TRANSFERS BETWEEN HINXTON – OTC – ON THE FLY DATA DELIVERY FOR 24H OF DATA PROCESSING BY 1200 PARALEL JOBS ON 300 VMs



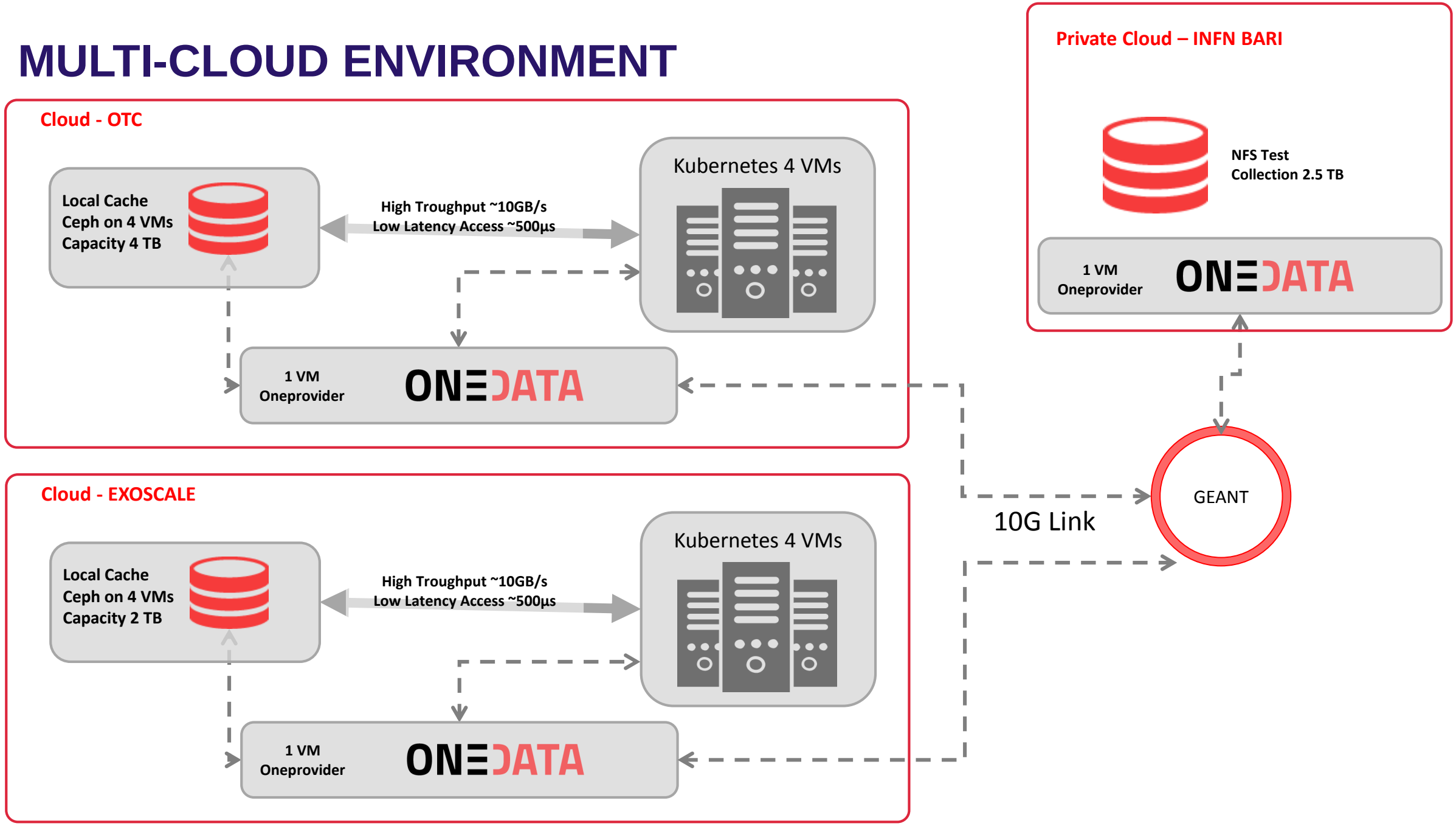
Experiment Specs:

- DNA Sequences processing
- 1046 Input Files ~ 15 – 50 GB
- Files represent DNA sequence data that's compared in turn to each of 24 reference human genes
- 24 x 1046 = 25104 Jobs in the experiment
- No Local Data on OTC Before Experiment starts
- 24 times of processing the same input fie makes caching very important
- Data pre-staging before the experiment is possible but not needed, system automatically delivers needed blocks and optimize access

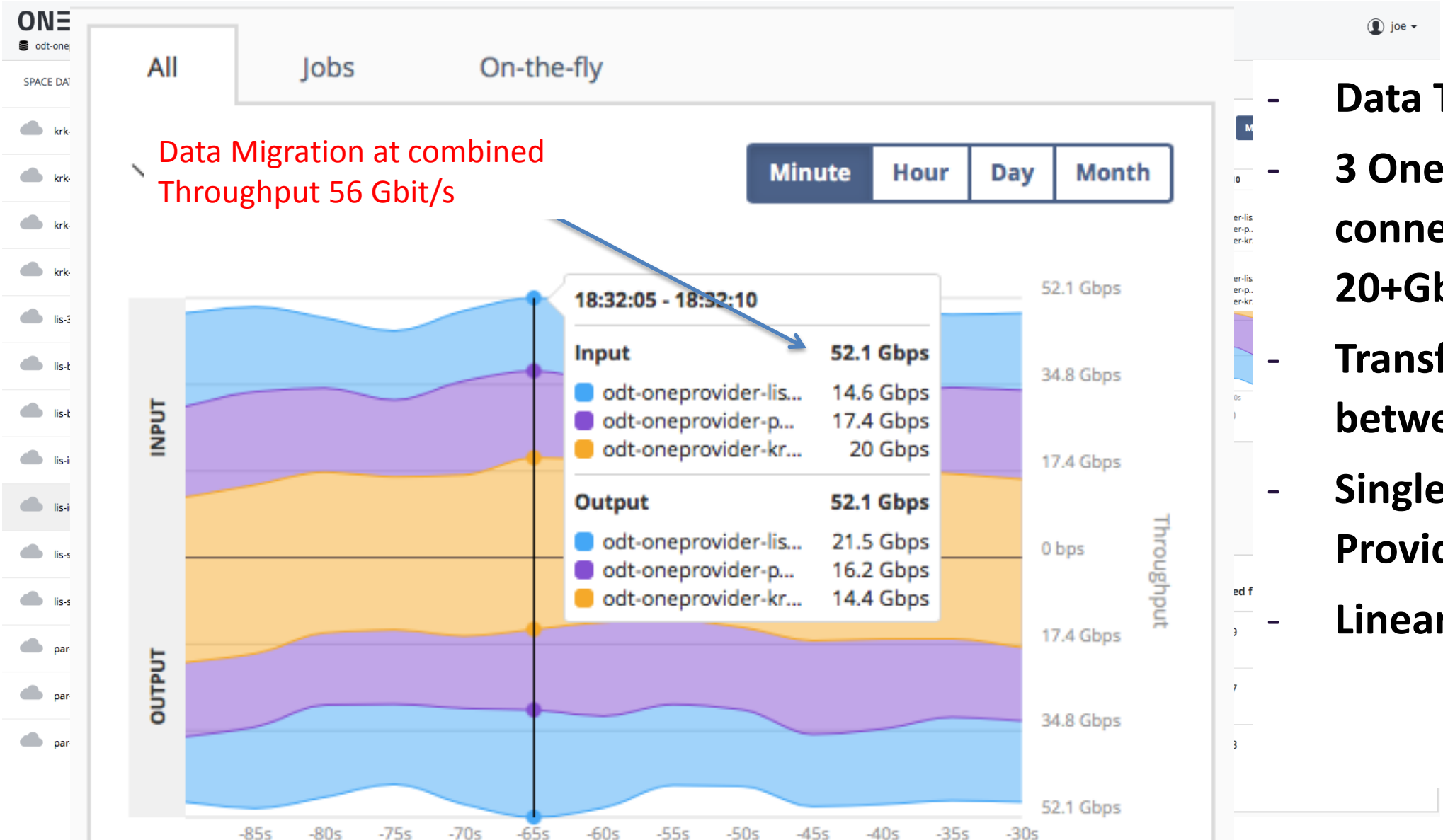
JOBS PROCESSING RATE WHEN DATA DELIVERED ON THE FLY



MULTI-CLOUD ENVIRONMENT

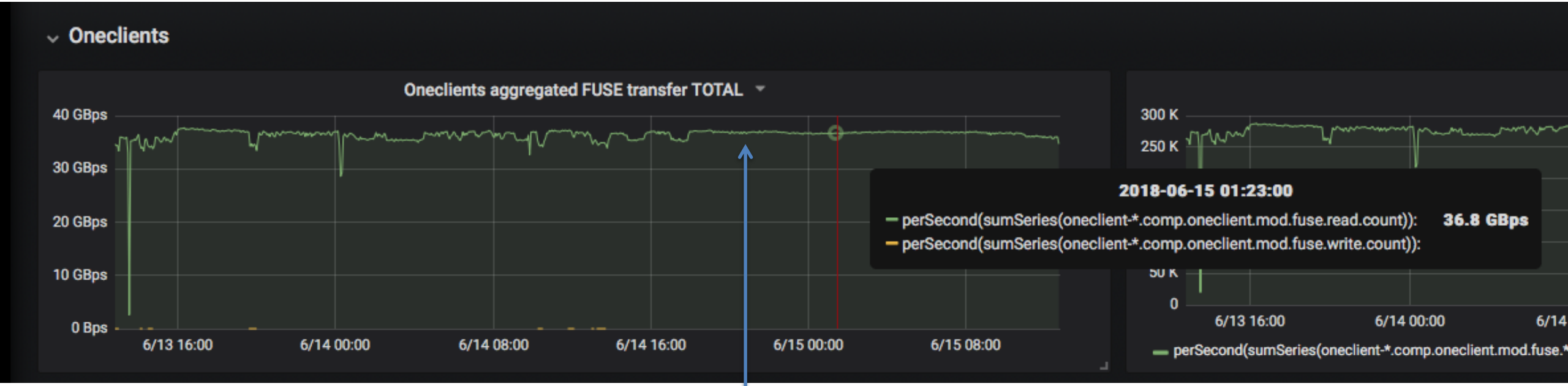


HIGH TROUGHPUT DATA MIGRATION IN PEER TO PEER MESH



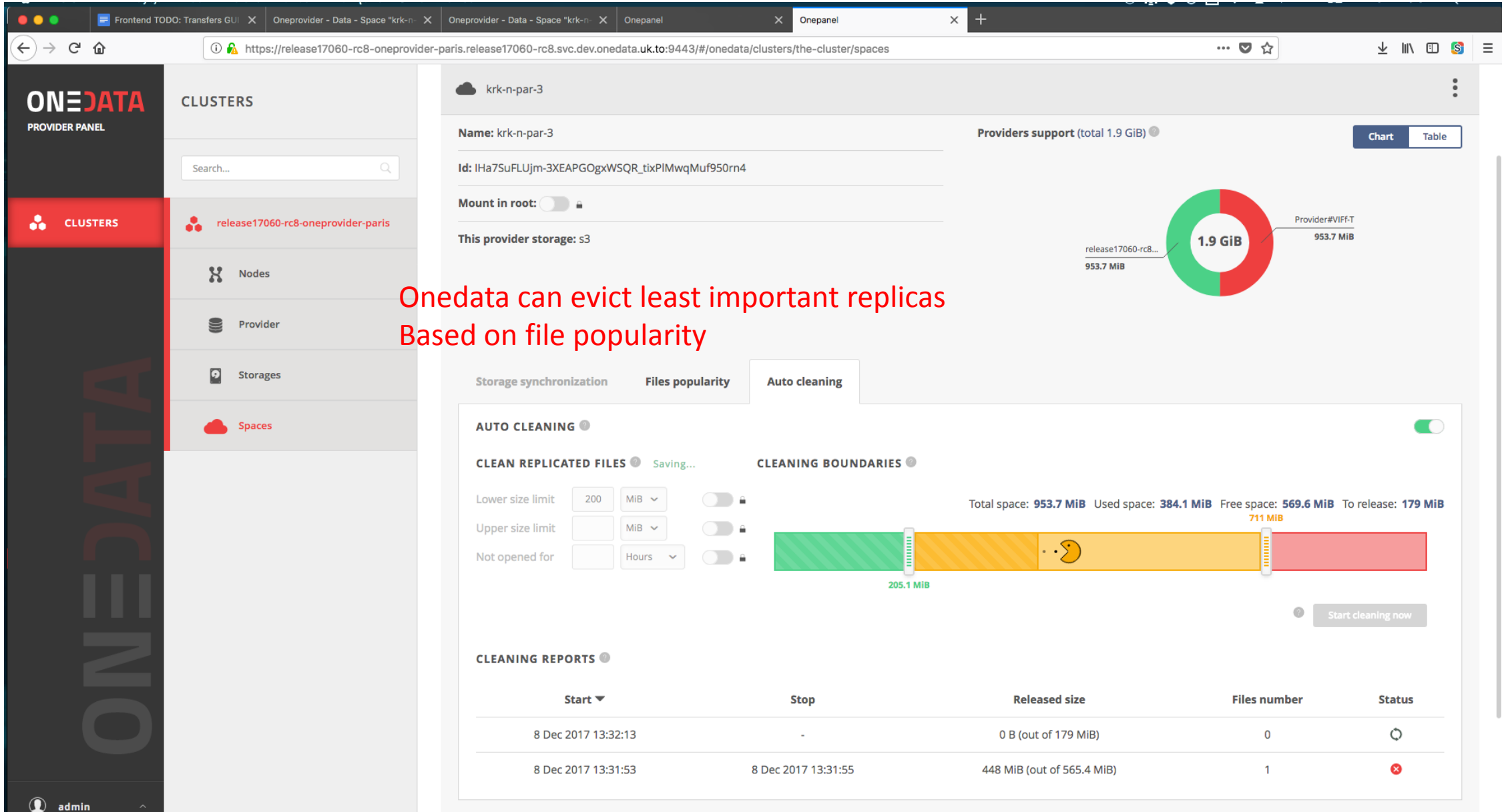
- Data Transfer Mesh
- 3 Oneproviders connected by 20+Gbit/s links
- Transfer data between all them
- Single VM Node per Provider
- Linear scalability

ONEDATA HIGH TROUGHPUT DATA PROCESSING ON HNSC

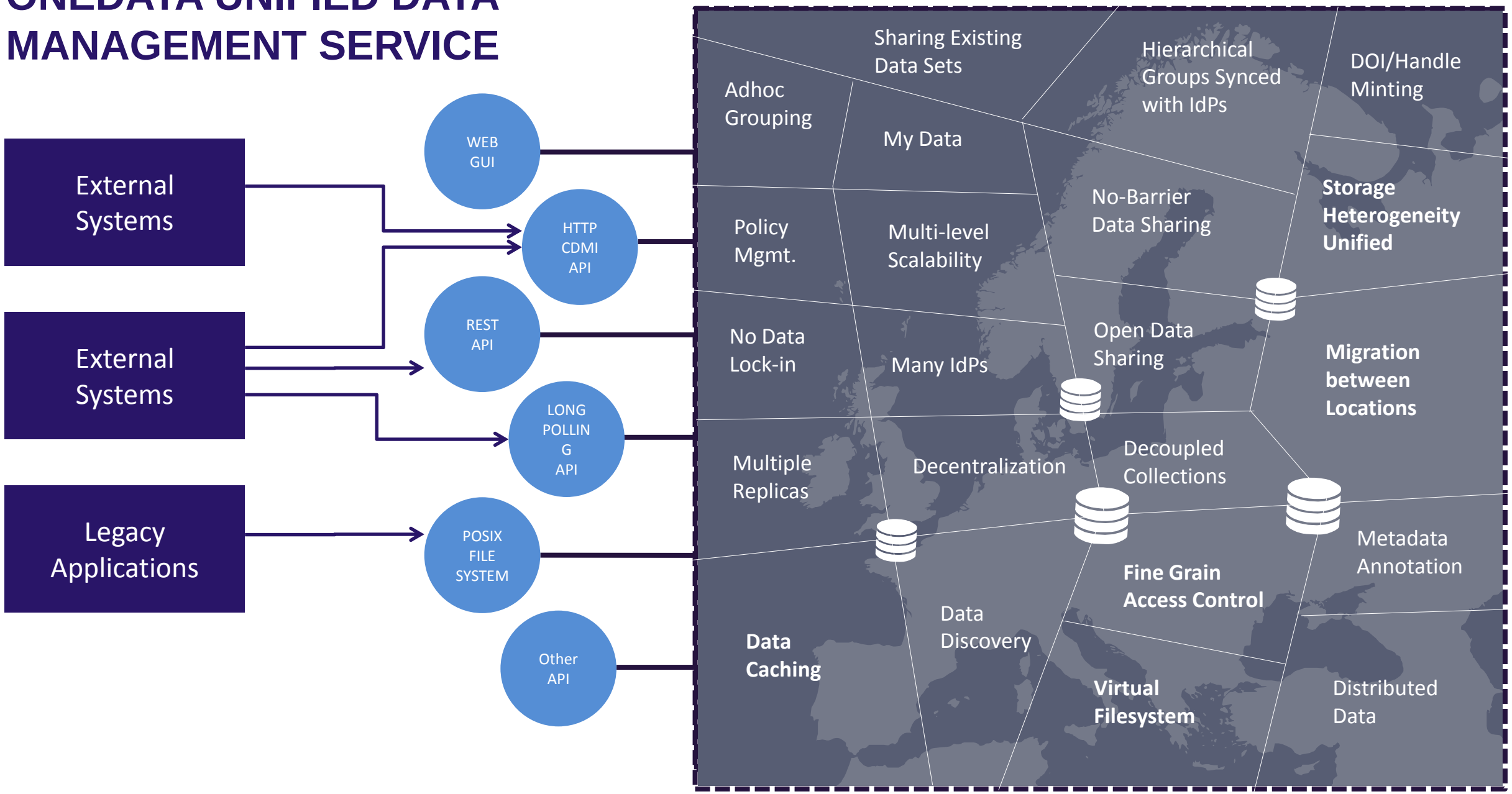


Onedata Transparent POSIX File System
Processing transparently cached data - 37GBytes/sec

ONEDATA FILE POPULARITY AND SMART CACHING



ONEDATA UNIFIED DATA MANAGEMENT SERVICE



QUESTIONS?

Please visit:

www.onedata.org