BIOBANKCLOUD

Your PaaS for Biobanking

BiobankCloud – A PaaS for Storage, Analysis and Inter-connection of Biobank Data

Alysson Bessani

Ciências ULisboa

www.biobankcloud.eu is financed by the European Commission 7th Framework Programme.

NGS Data Flood

HiSeq X Ten^ => ~18,000 genomes/year

Volume => ~5.2 PB/year*

Velocity => ~45 MB/sec*

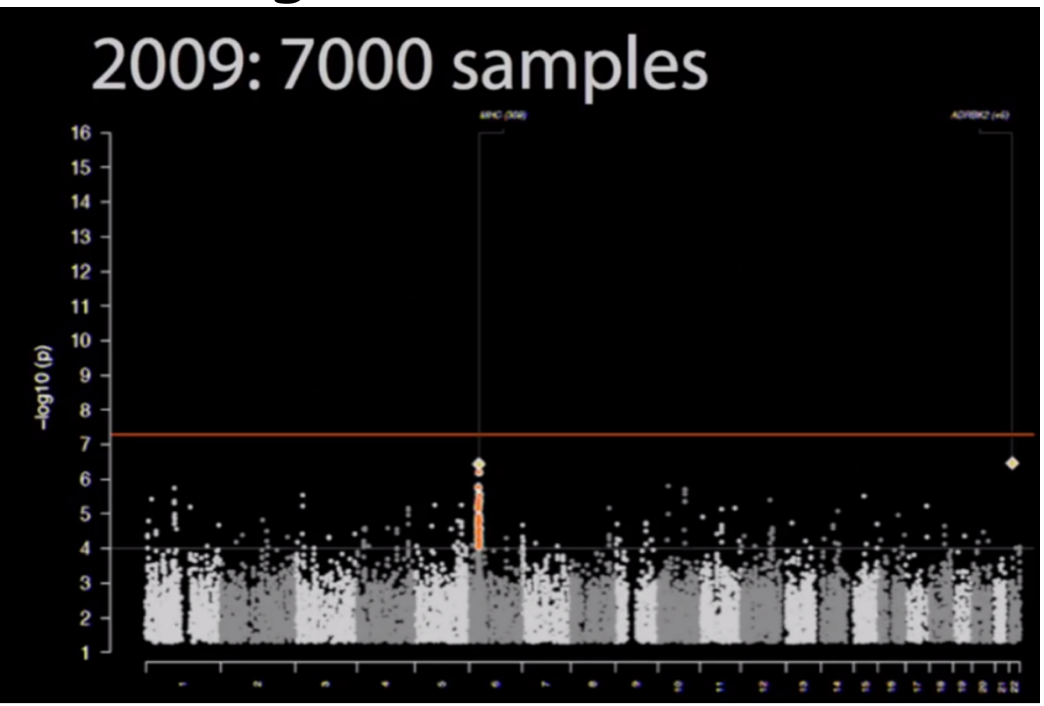
^Cost ~$10 million

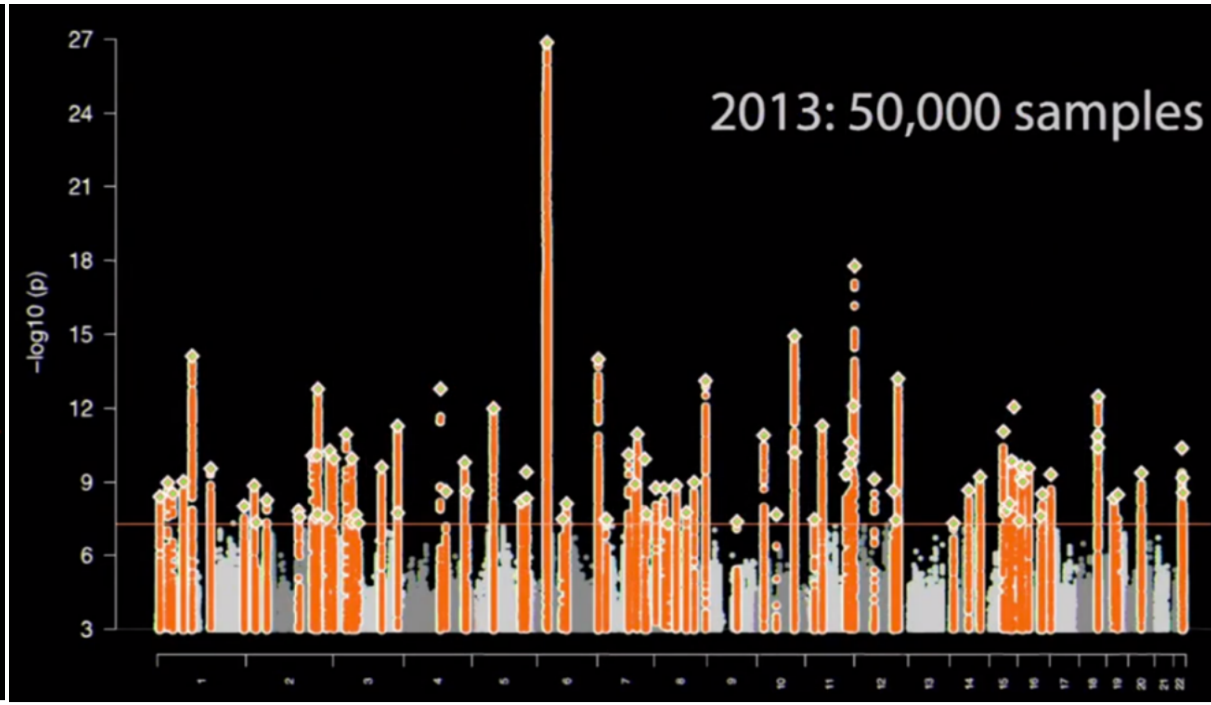
*5.2 PB assumes a replication factor of 3

See: http://goo.gl/OCgJ36

Genomics needs Big Data

Schizophrenia is not based on genomic variation

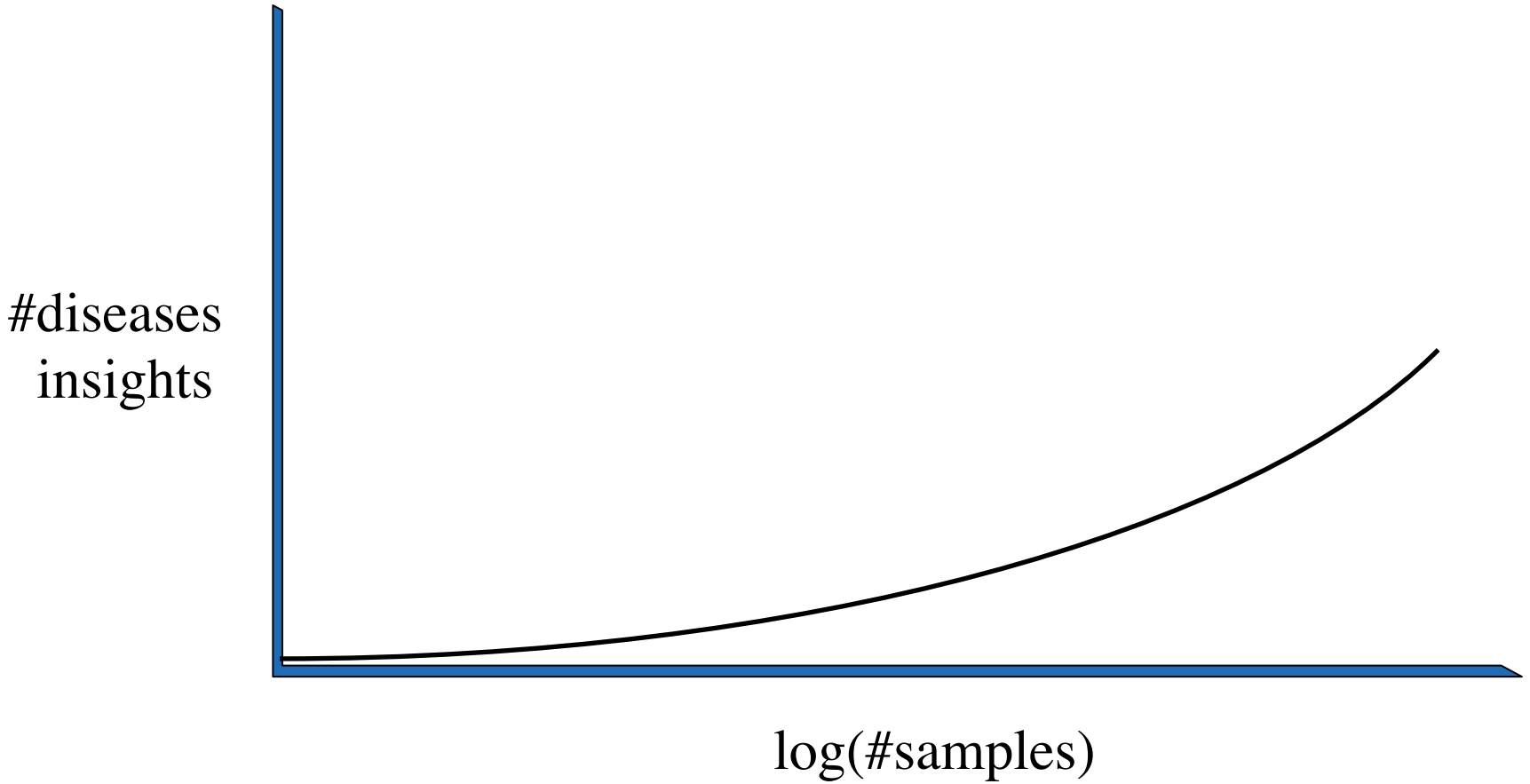




There are 60 mutations associated with schizophrenia

[Images source: Patterson, Fighting the Big C with the Big D, 2014]

Network effects: Biggest Dataset wins!



Centralized dataset: technically feasible, politically hard.
Federated dataset: technically hard, politically feasible.

The BiobankCloud Project

www.biobankcloud.eu

Definition of a Biobank

- The biobank "concept" is defined (by Swedish law) as:

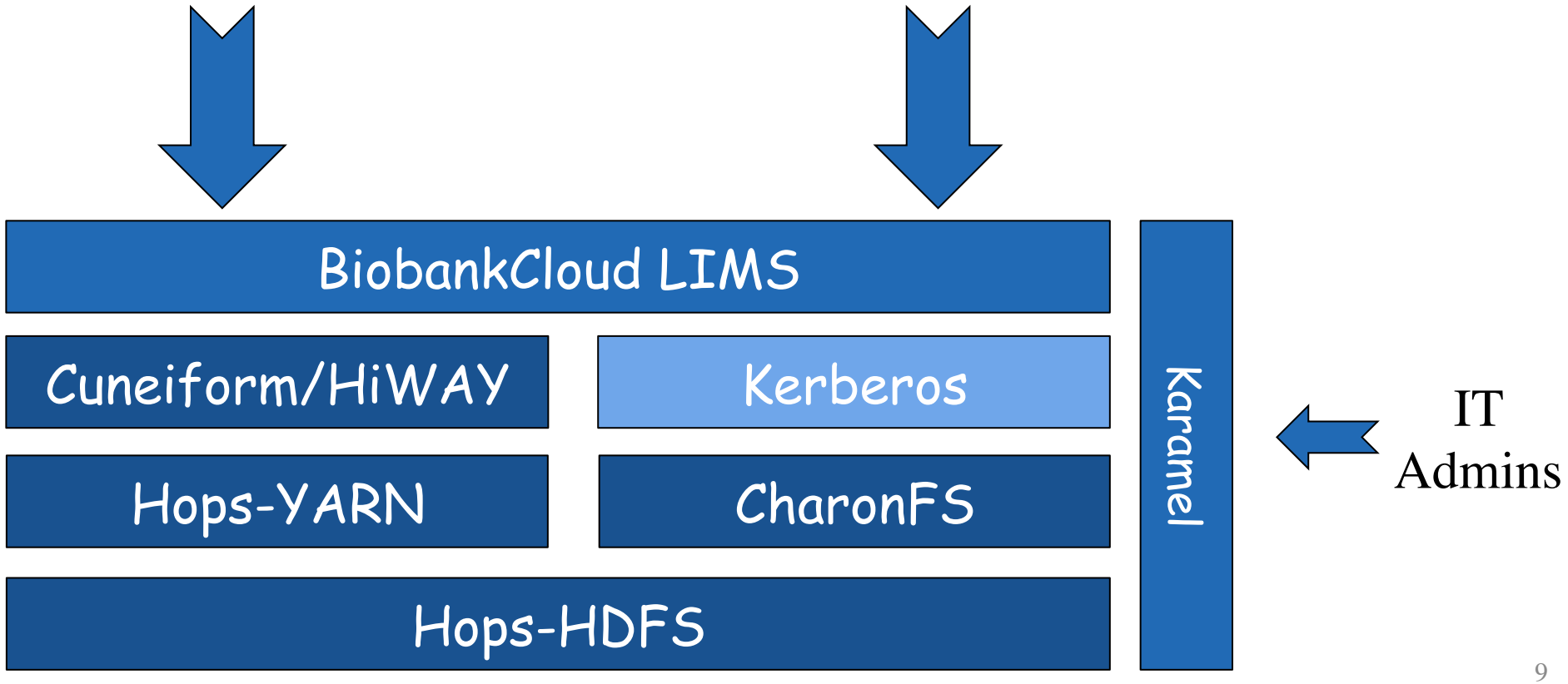
  "biological material from one or several human beings collected and stored indefinitely or for a specified time and whose origin can be traced to the human or humans from whom it originates"

- In Sweden, they have the goal of digitizing biological material in biobanks (*e-biobanking*).

≈ 500,000 people

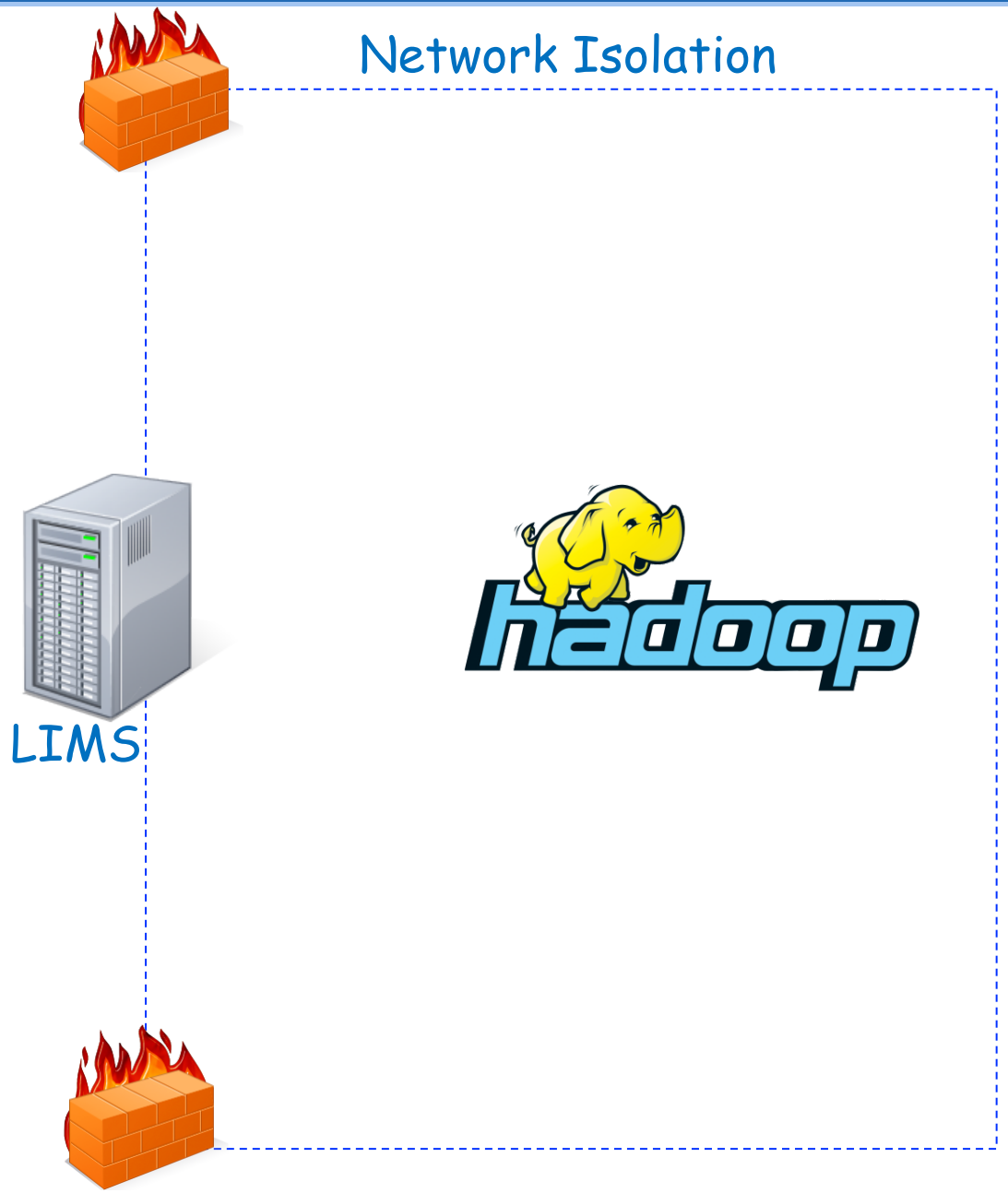
BiobankCloud PaaS

- Biobankers
  - NGS data producers
    - Collections, samples
  - Non-programmers
- Bioinformaticians
  - NGS data analysts
  - Programmers
    - Python, R, Matlab, scripts

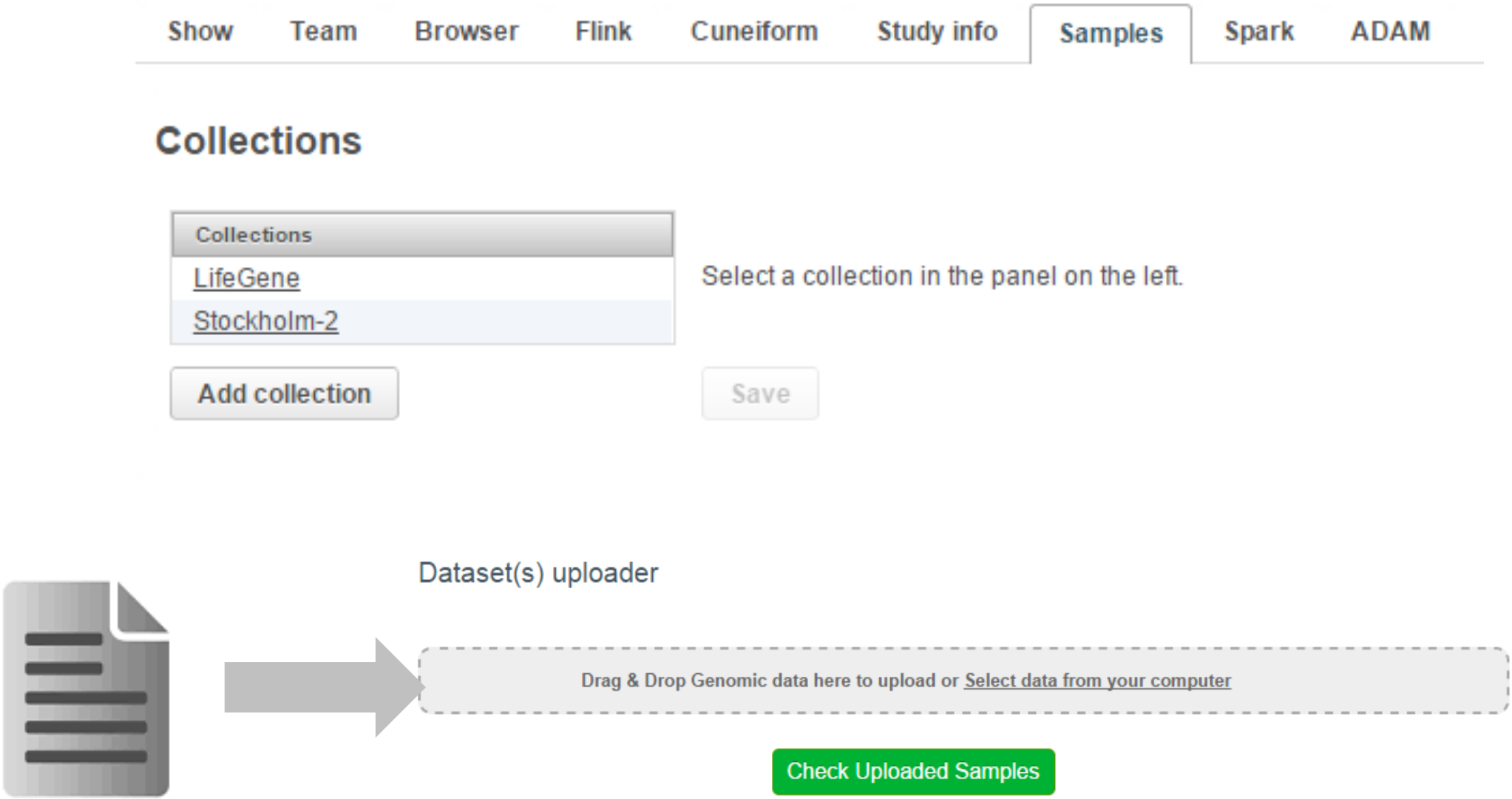


LIMS: Lab Information Management System

- LIMS for NGS Data
  - Multi-tenancy
  - Study-level
  - Role-Based Access Control
  - Two-factor authentication
  - Audit trails
  - Plugin model for apps
  - REST APIs

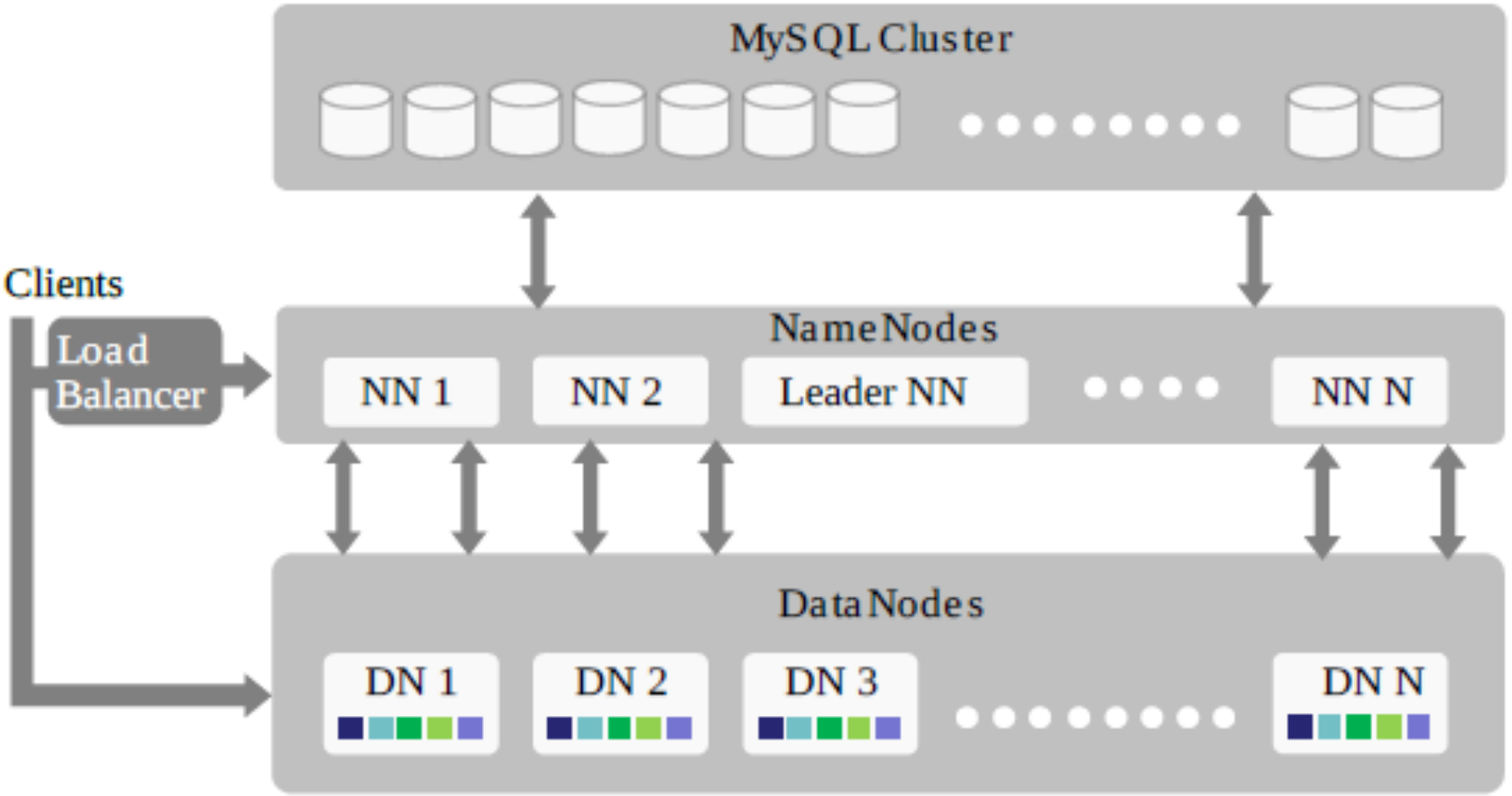


Manage Sample Collections and Studies



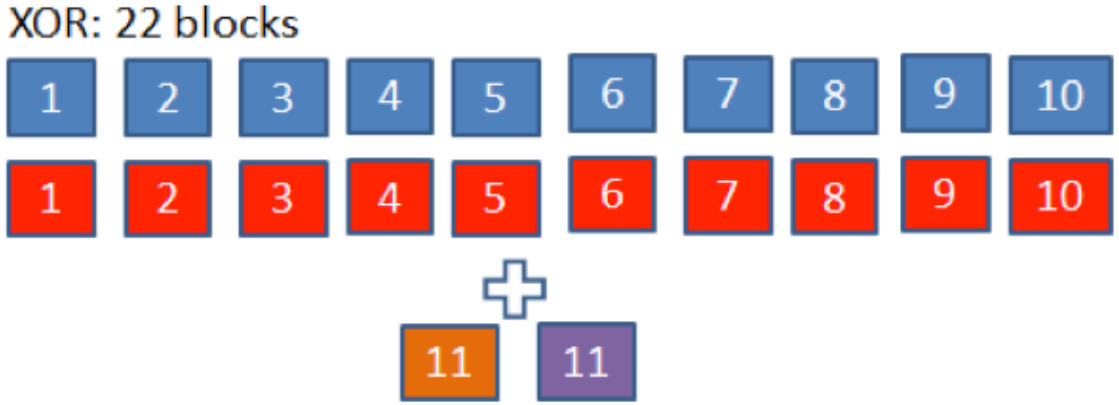
Data Storage: HopsFS

- A modified version of HDFS
- Customizable and Scalable Metadata
- High throughput for read and write operations
- NameNode failover time ≈5 seconds (vs ~1 minute for HDFS)

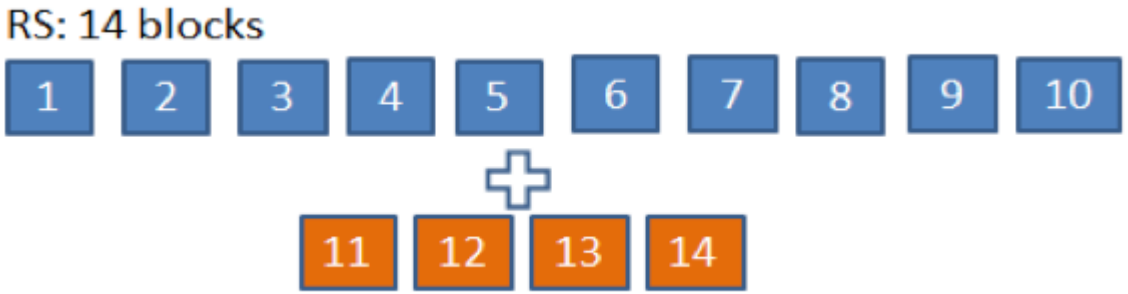


HopsFS Erasure Coding

HDFS 2.x
Triple Replication
(+200%)

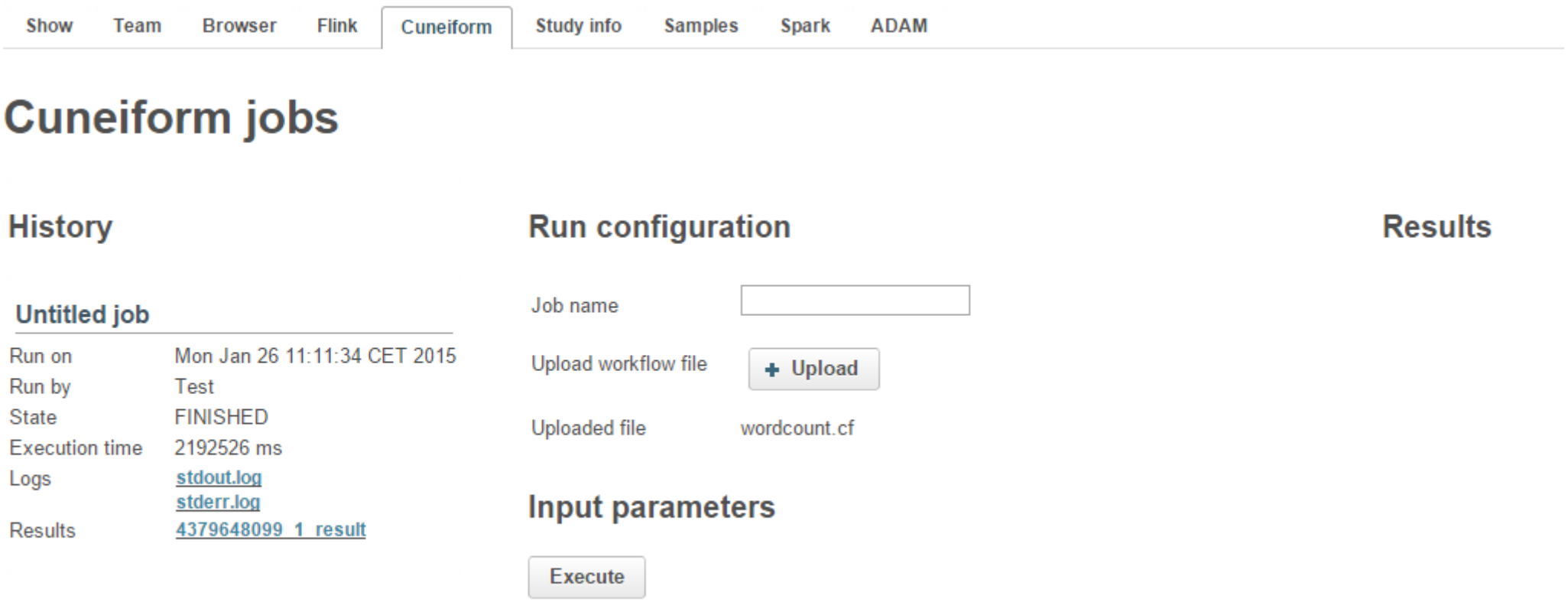


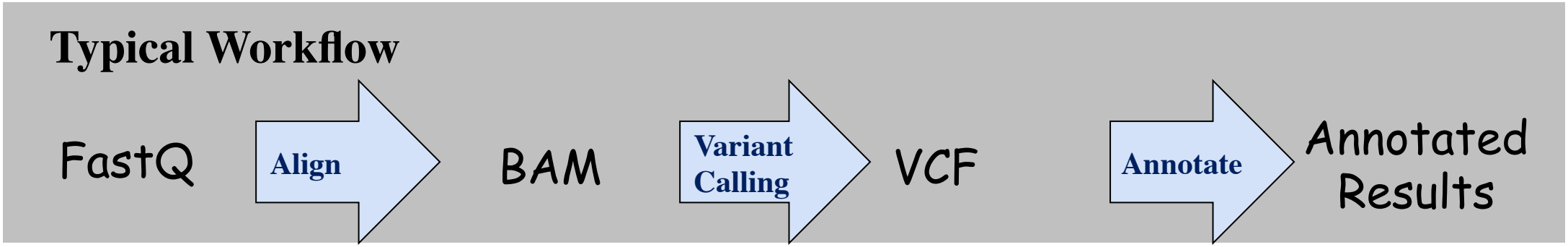
2x Replication + XOR
(+120%)



Reed-Solomon
(+40%)

Run Cuneiform Workflows on YARN





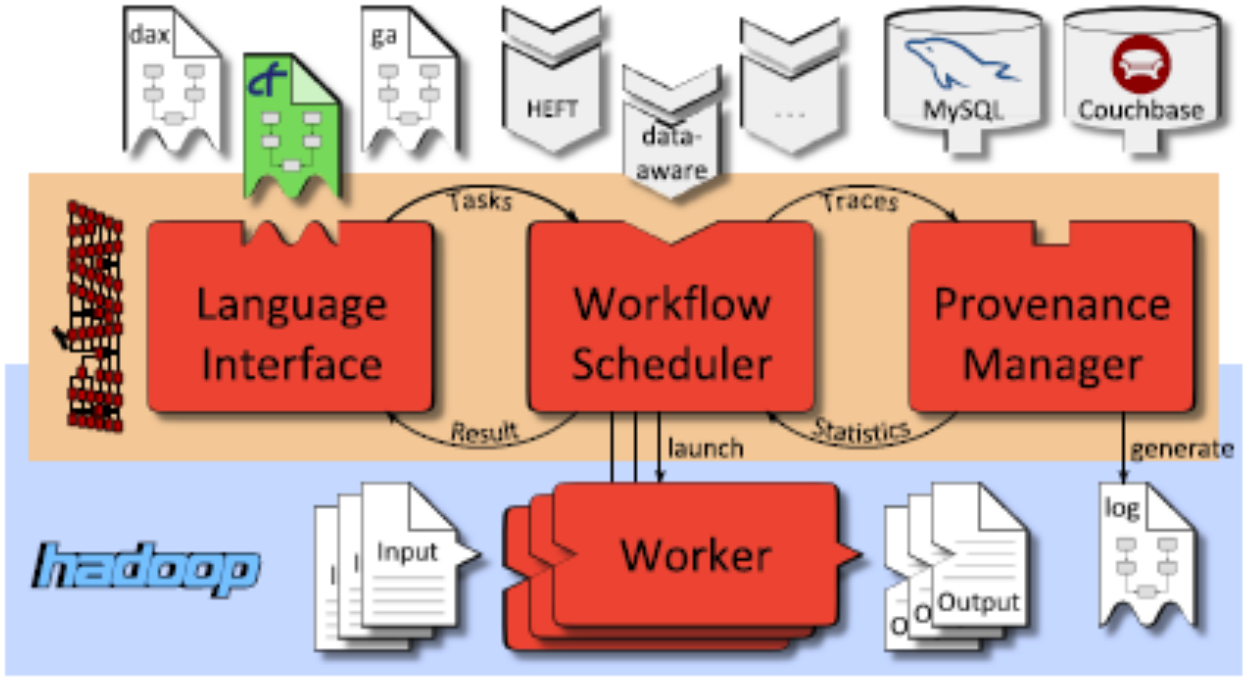
Cuneiform

- Light-weight statically typed functional dataflow language
- Compiles into dynamic pipelines of black-box tools
- Aims
  - Make foreign code integration as easy as possible
  - Allow complex, iterative workflows
  - Deduce options for parallelism automatically

```
deftask per-chromosome(
    vcf( File )
  : fa( File )
    [fastq1( File ) fastq2( File )] ) {

  bt2idx = bowtie2-build( fa: fa );
  fai = samtools-faidx( fa: fa );

  sam = bowtie2-align(
    idx:    bt2idx
    fastq1: fastq1
    fastq2: fastq2 );

  bam = samtools-view( sam: sam );

  sortedbam = samtools-sort( bam: bam );

  mpileup = samtools-mpileup(
    sortedbam: sortedbam
    fa:        fa
    fai:       fai );

  vcf = varscan( mpileup: mpileup );
}
```

Hi-Way

- Hi-Way Workflow Application Master for YARN
- Executes workflows on Hadoop YARN
  - Scalability, maintenance, fault tolerance, …
- Full provenance tracing, executable provenance
- Runs **Cuneiform**, Galaxy, Pegasus (DAX)
- Various (adaptive) schedulers
- Dynamic workflow interface



Integration with Public Clouds

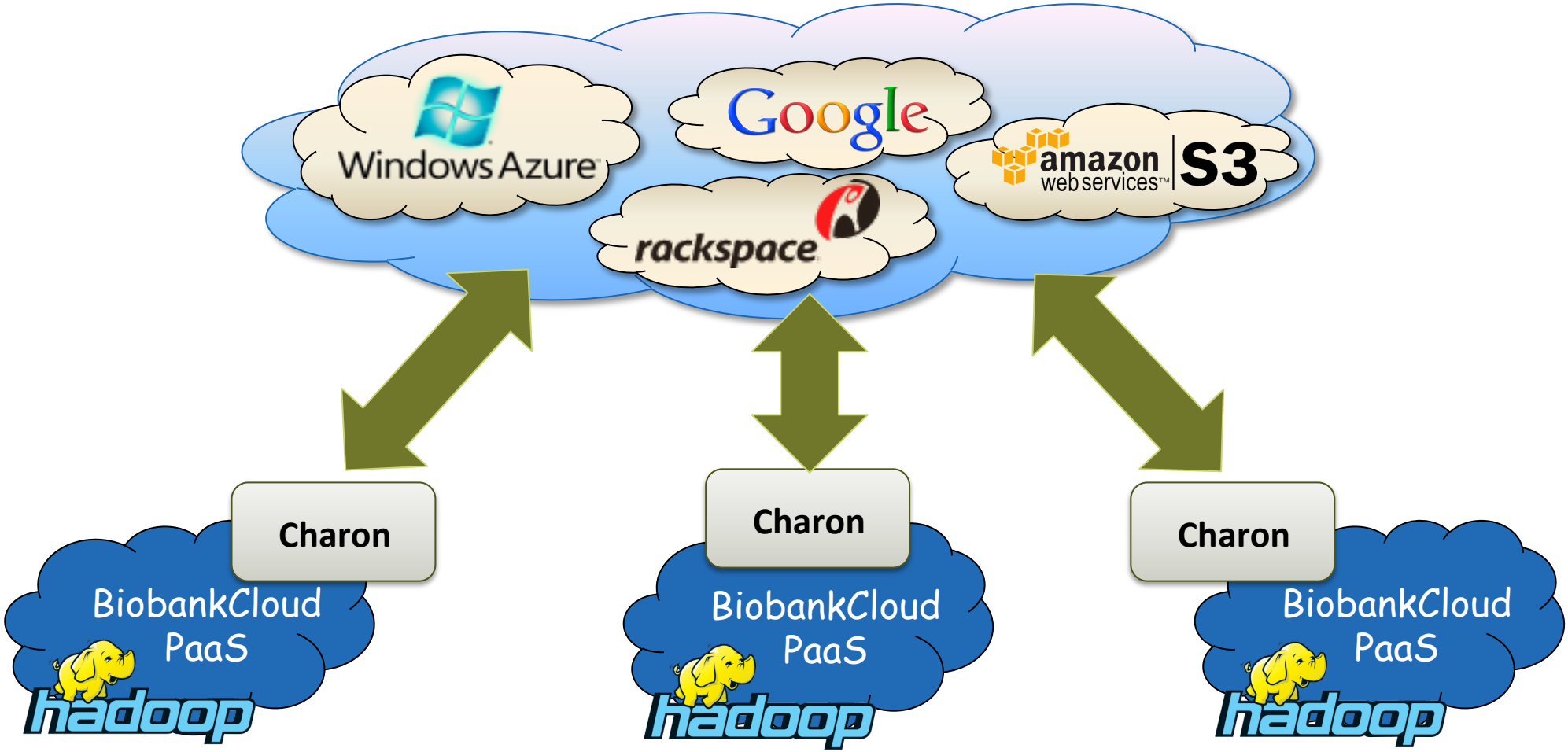
Two Problems…

- How to use public clouds to increase the storage capacity of BiobankCloud platform deployments?

- How to interconnect different BiobankCloud platform deployments in a federation?

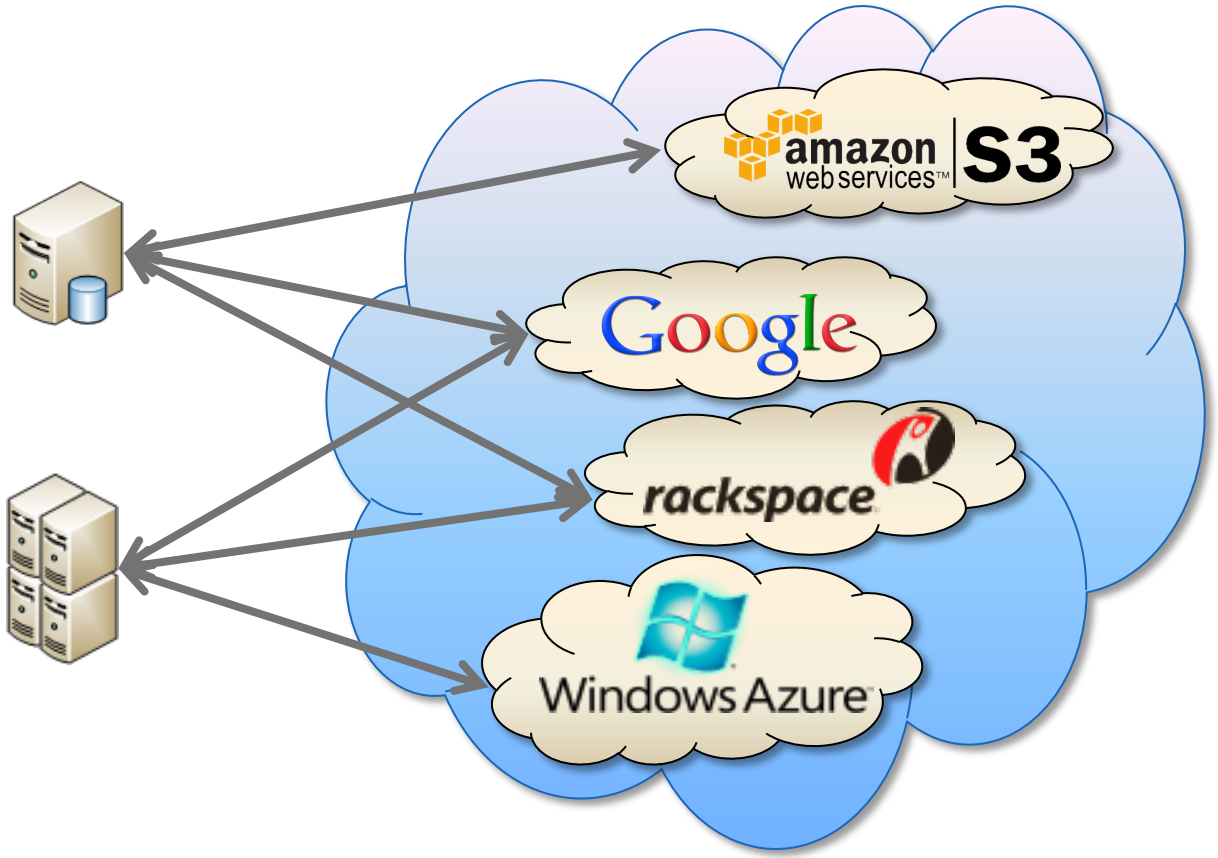
*(Important: how to do that without endangering **security** and with minimal **management effort**?)*

Charon FS


Windows Azure
Google
rackspace
amazon web services
S3
Charon
BiobankCloud PaaS
hadoop
Charon
BiobankCloud PaaS
hadoop
Charon
BiobankCloud PaaS
hadoop

DepSky: Dependable Cloud-of-Clouds Object Storage

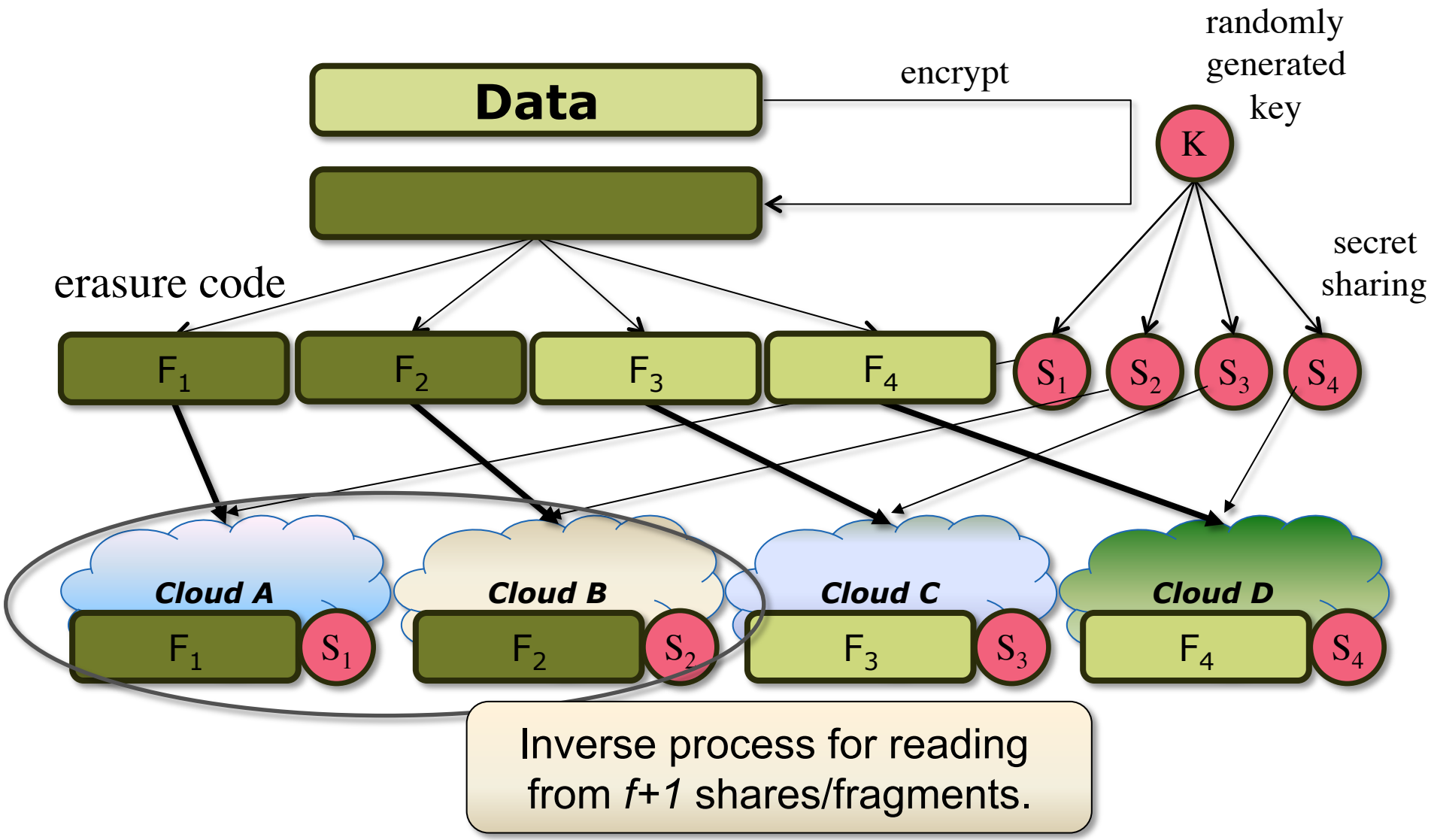
ACM Transactions on Storage 9(4). Nov. 2013 (earlier version in EuroSys'11).



Why multiple clouds?

Survive cloud-wide outages
Avoid vendor lock-in
Better read performance
Tolerance to data corruption
- Bugs
- Malicious insiders
- Attacks and intrusions

DepSky Storage Efficiency and Confidentiality



Why a File System?

1. Most bioinformatics tools also work directly over files, avoiding thus import/export operations

2. The popularity of Dropbox-like services evidences that working with files is simple and intuitive

3. A file system allows one to transparently store files in several locations (e.g., local storage, remote sites, clouds, cloud-of-clouds)

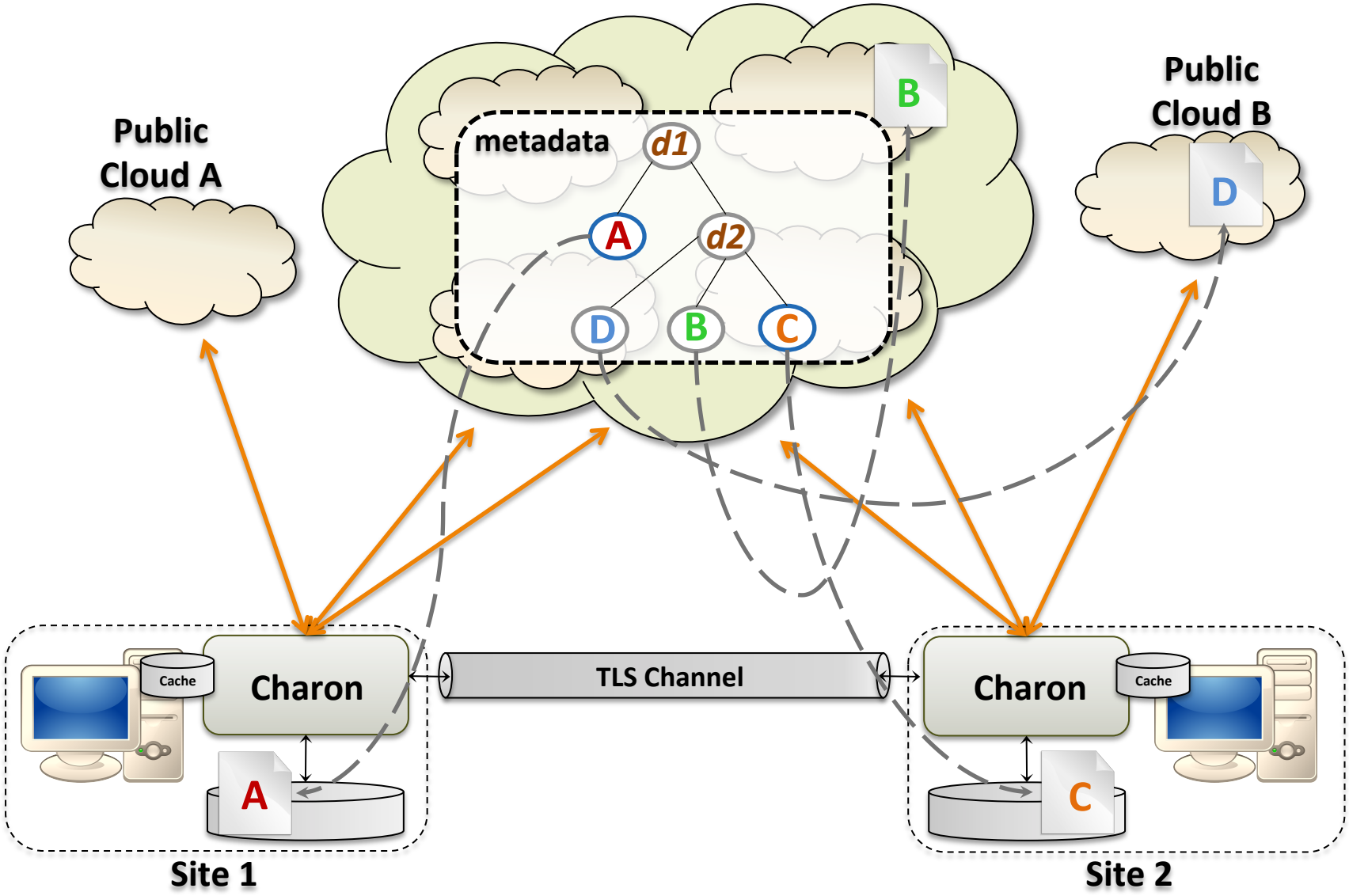
Charon Design Principles

- Separation of file data and metadata
  - Metadata: file name, size, permissions, stored in *name service* objects
  - Data: file contents in blocks of up to 16MB
- Files can be stored in different locations
  - Using semantic cues as in WheelFS (MIT). Examples:

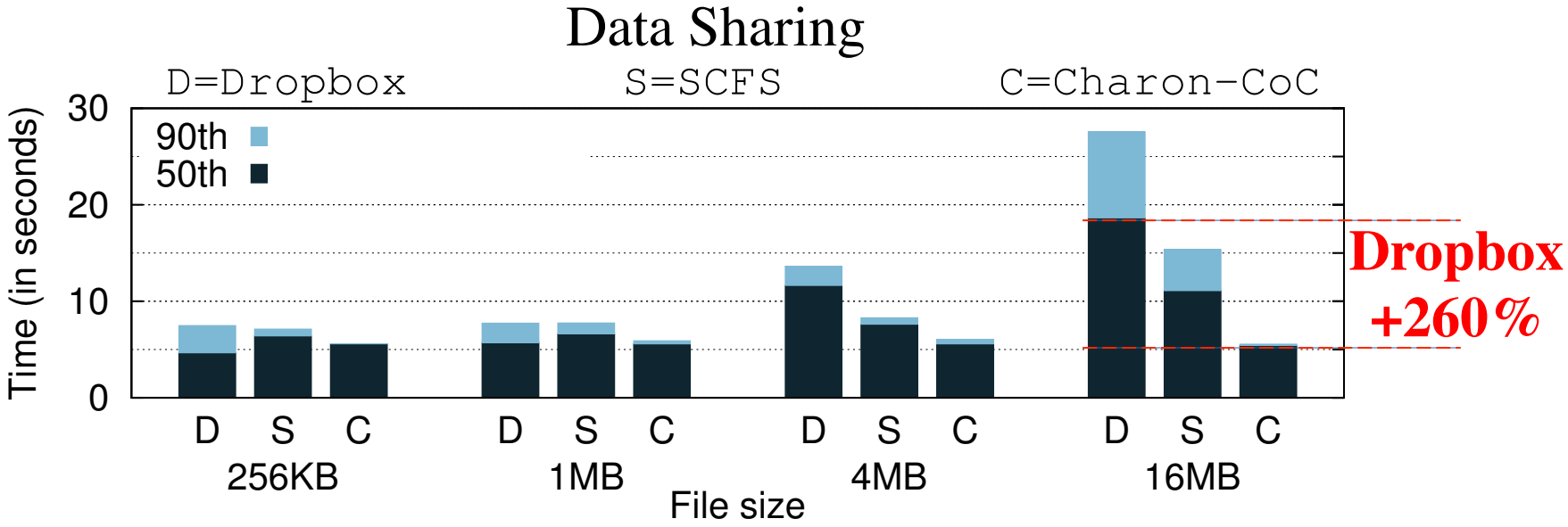
    `/shared/studies/.Site=CoC/rawreads`

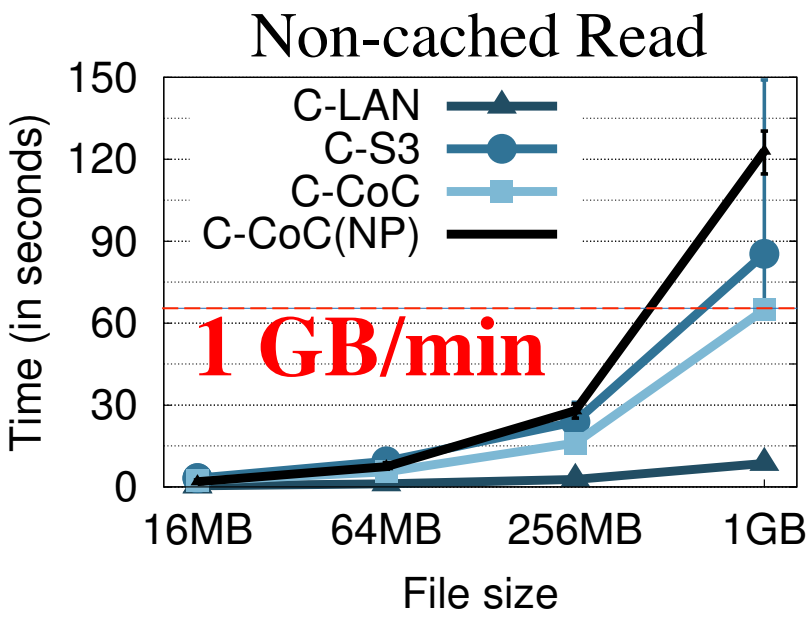
    `/shared/studies/.Site=local/vcfs`
- No write/write file conflicts
  - Files and directories are **locked** for write
- Serverless design
  - There are no explicit servers in a Charon infrastructure, only cloud (storage) services
  - All storage and coordination protocols are **data-centric**, based on the services cloud providers have to offer

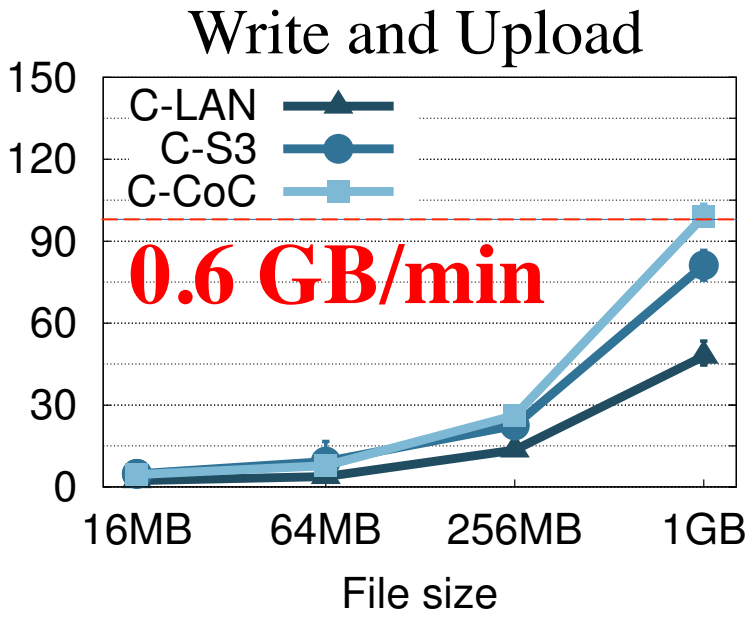
Charon Architecture



Charon Performance







Summary

- BiobankCloud PaaS
  - One-click Cluster deployment
    - Chef/Karamel
  - Big Data Storage
    - HopsFS – modified HDFS with scalable metadata and erasure code
  - Workflow Execution Engine
    - Cuneiform – workflow language
    - Hi-way – execution engine on top of YARN
  - Multi-cloud Storage and Big Data Sharing
    - Charon – Serverless Cloud-of-Clouds File System
    - Secure integration of public cloud storage to the system
    - Minimal infrastructure requirements

- The software will be made available in the next months

The Team

www.biobankcloud.eu

- **KTH**
  Jim Dowling, Salman Niazi, Mahmoud Ismail, Kamal Hakimzadeh, Ali Gholami, Erwin Laure

- **Karolinska Institute**
  Jan-Eric Litton, Roxana Martinez, Jane Reichel, Mats Hansson

- **University of Lisbon (FC)**
  Alysson Bessani, Vinicius Cogo, Tiago Oliveira, Ricardo Mendes

- **Humboldt University**
  Ulf Leser, Jörgen Brandt, Marc Bux, Sebastian Wandelt, Johannes Starlinger

- **Charité University Hospital**
  Michael Hummel, Lora Dimitrova, Karen Zimmermann

Financed by the European Commission 7th Framework Programme.