

*Advanced Networking for the EU genomic **research***

ARES

Gianluca Reali – coordinator
University of Perugia



2nd TERENA Network Architects Workshop
Prague, November 13-14, 2013

Outline

- Description of ARES
- ARES research and implementation purposes
- Technologies and design choices
- Expected Results

ARES Partners

– University of Perugia (UoP)

- To design and deploy the ARES CDN network;
- To deploy software instances to manage both the network and the processing tools;
- Execution of experiments (network side);

– Polo d'Innovazione di Genomica, Genetica e Biologia SCARL (GGB)

- Definition of experimental scenarios and relevant metrological procedures.
- Execution of experiments as a CDN customers;
- Evaluation of the grade of received network service,

Why ARES?

Future P4 medicine framework: **proactive, personalized, predictive, and participatory** [1].



Berge Minassian, Hospital for Sick Children in Toronto, “I am certain that in the next few years patients walking into children’s hospitals will have their whole genomes sequenced,”[2].
FUTURE NEED OF SEQUENCING, STORING, MAKING AVAILABLE, CONTINUOUSLY ANALYZING THE GENOME OF EACH INDIVIDUAL through real-time knowledge of the latest findings!!!



tremendous volume of data:

NEED OF SUITABLE STORAGE, NETWORKS, PROTOCOL ARCHITECTURES, APPLICATIONS,...

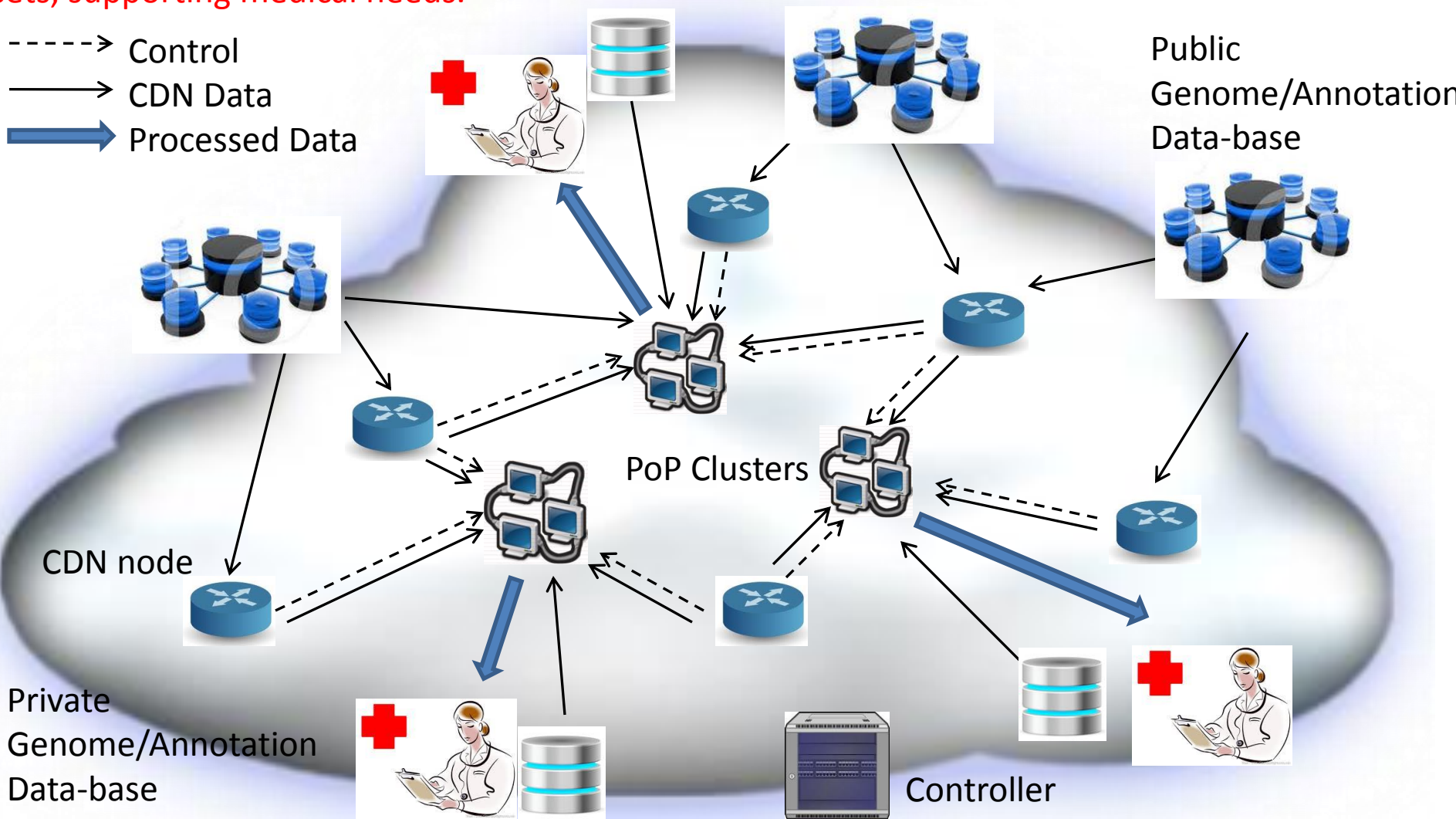
[1] Hood, L., Balling, R., and Auffray, C. (2012). Revolutionizing medicine in the 21st century through systems approaches. *Biotechnol. J.*, 7:1-10.

[2] <http://blogs.nature.com/spoonful/2013/01/gene-sequencing-yields-breakthrough-for-children-with-rare-parkinsons-like-disorder.html>

ARES Idea (1/2)

Combined use of CDN and CLOUD/GRID technologies , specifically targeted to genomic data sets, supporting medical needs.

-----> Control
 —————> CDN Data
 —————> Processed Data



Reasoning behind technology and design choices

- Original aspects of genomic data sets
 - i.1 Content growth
 - i.2 Content popularity
 - i.3 Logical content relationships
- Advanced CDN features
 - i.4 Content distribution logic
 - i.5 Suitably integration with cloud storage and processing services
 - i.6 Novel cache instantiation procedure
 - i.7 Parallel download algorithm
 - i.8 Multiple classes of network services supporting different medical needs.

i.1 Content growth (1/2)

1000 Genomes

A Deep Catalog of Human Genetic Variation

[Home](#)
[About](#)
[Data](#)
[Analysis](#)
[Participants](#)
[Contact](#)
[Browser](#)
[Wiki](#)
[FTP search](#)

[Home](#)

How many individuals will be sequenced?

[Data Access](#)
[statistics](#)

Answer:
The project aims to sequence 2500 individuals in total both low coverage whole genome sequencing and exome sequencing. So far more than 1000 samples have been sequenced.

Related questions:
[What Sequencing Platforms were used for the 1000 genomes project](#)
[Which samples are you sequencing?](#)
[How do I find out about new 1000 genomes releases](#)

How much disk space is used by the 1000 genomes project?

[Data Access](#)
[statistics](#)

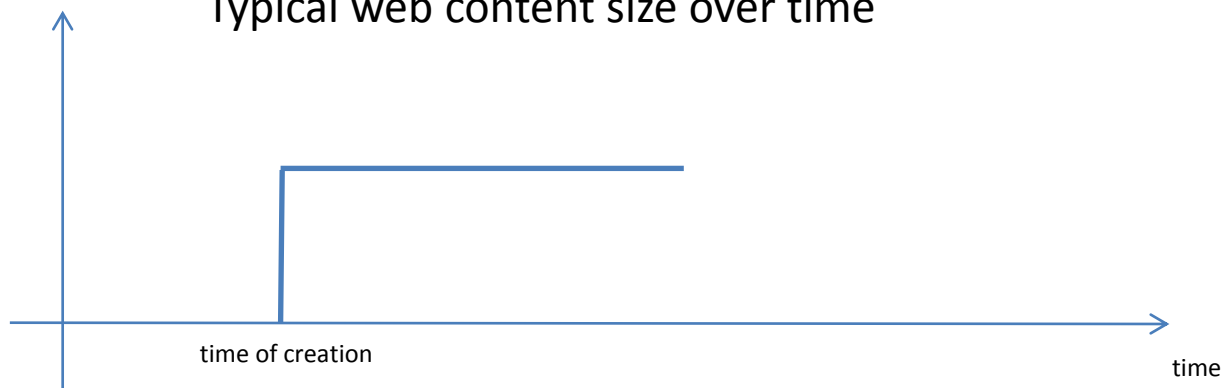
Answer:
As of March 2013 our ftp site is 464Tbytes and continuing to grow. To find the most accurate size of our site at anyone time it is best to look at the file sizes in our [current tree file](#). This file contains the relative path, size, md5 and last updated time of every file on our ftp site.

Related questions:
[What to](#)
[Can I g](#)
[How do](#)

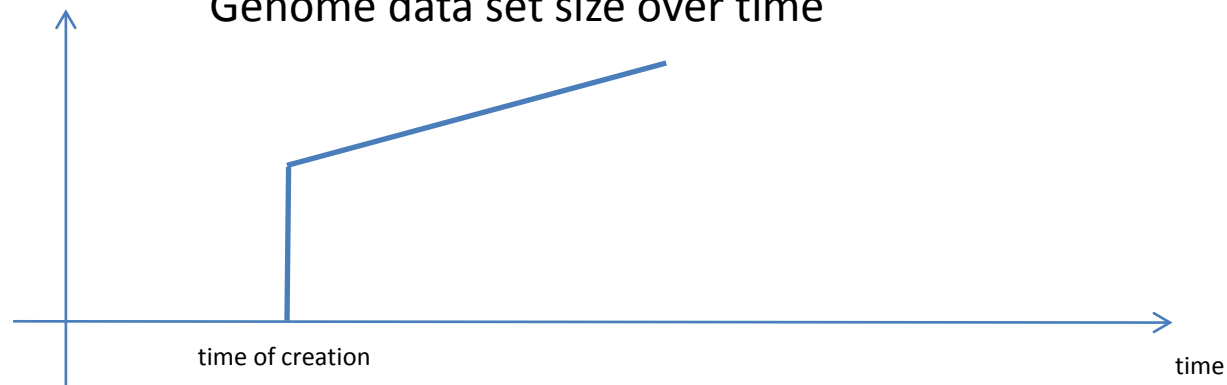
For just 1000 samples!

i.1 Content growth (2/2)

Typical web content size over time



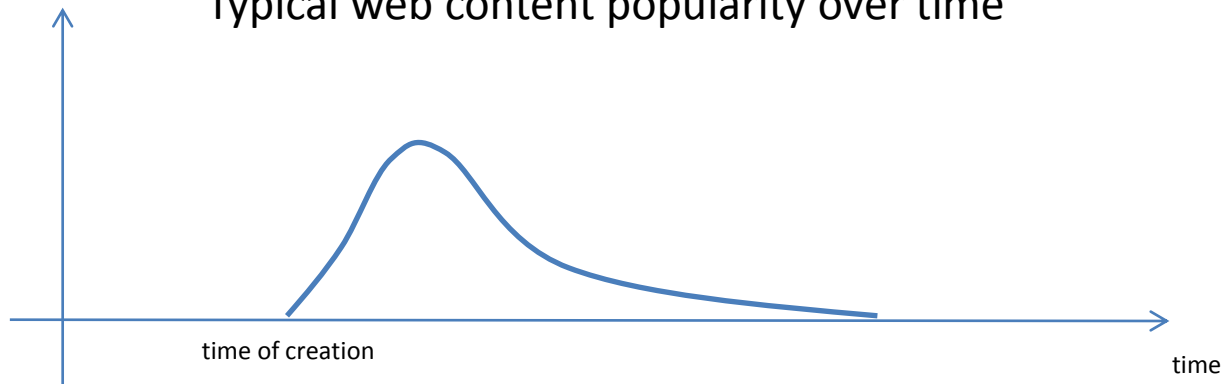
Genome data set size over time



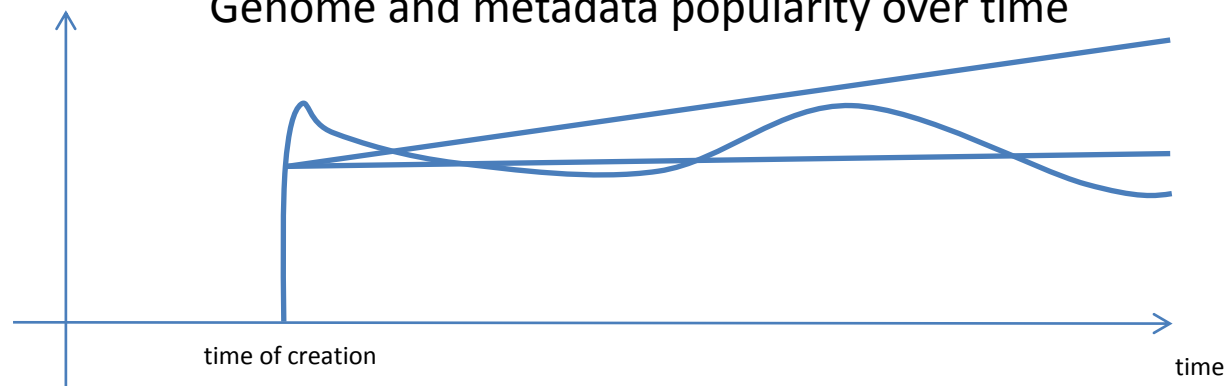
Any genome is a huge source of information to be still unveiled !!!
 Research will produce a significant increase of the genomic data set
 for each patient!

i.2 Content Popularity

Typical web content popularity over time



Genome and metadata popularity over time



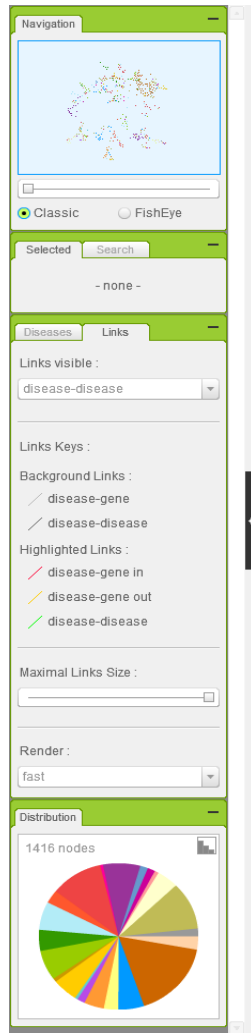
Not predictable
shape, but it
never expires!!!!

Only arrivals
process!!!

Huge implications
for CDNs!

i.3 Logical content relationships (1/2)

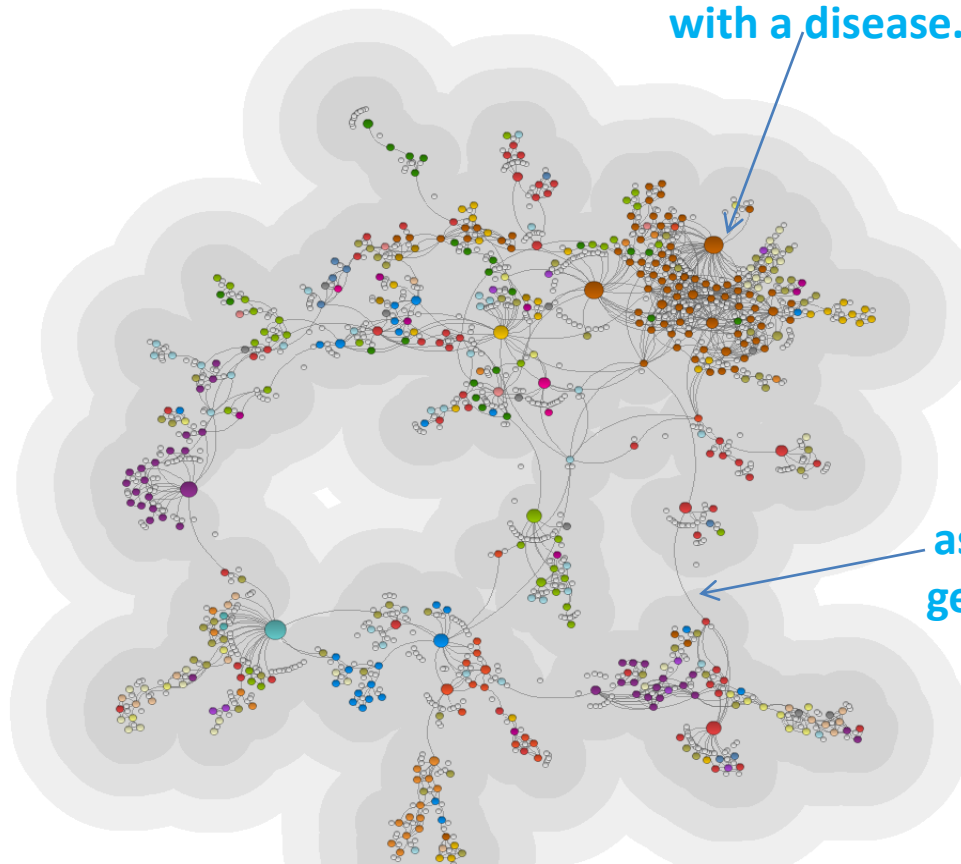
content relationships based on gene “affinity”



Diseaseome

Each circle is associated with a disease.

Each arch is associated with a gene relationship.

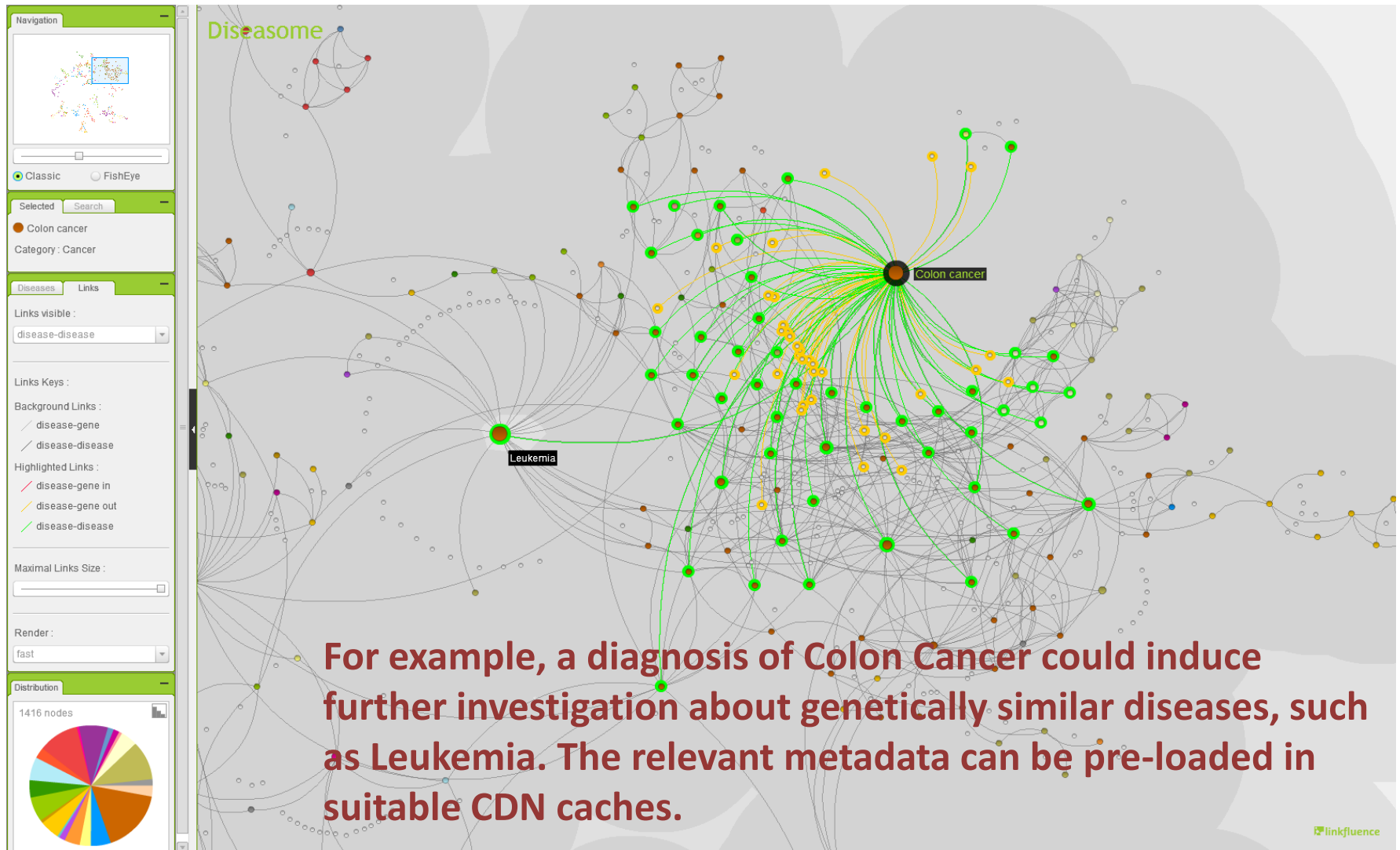


Diseases may show degree of generic similarity.

Information useful for driving diagnostic investigations, thus for managing data in CDNs

i.3 Logical content relationships (2/2)

e.g. genomic links with colon cancer.



i.4 Content Distribution Logic (1/3)

- Based on **NSIS** advanced discovery algorithms and signaling
- Based on **differentiated** medical needs, that is the time required for downloading data according to the seriousness of a disease (better illustrated in what follows)
- Leveraging on **cloud** services
- Original management of virtualization services through **NetServ**

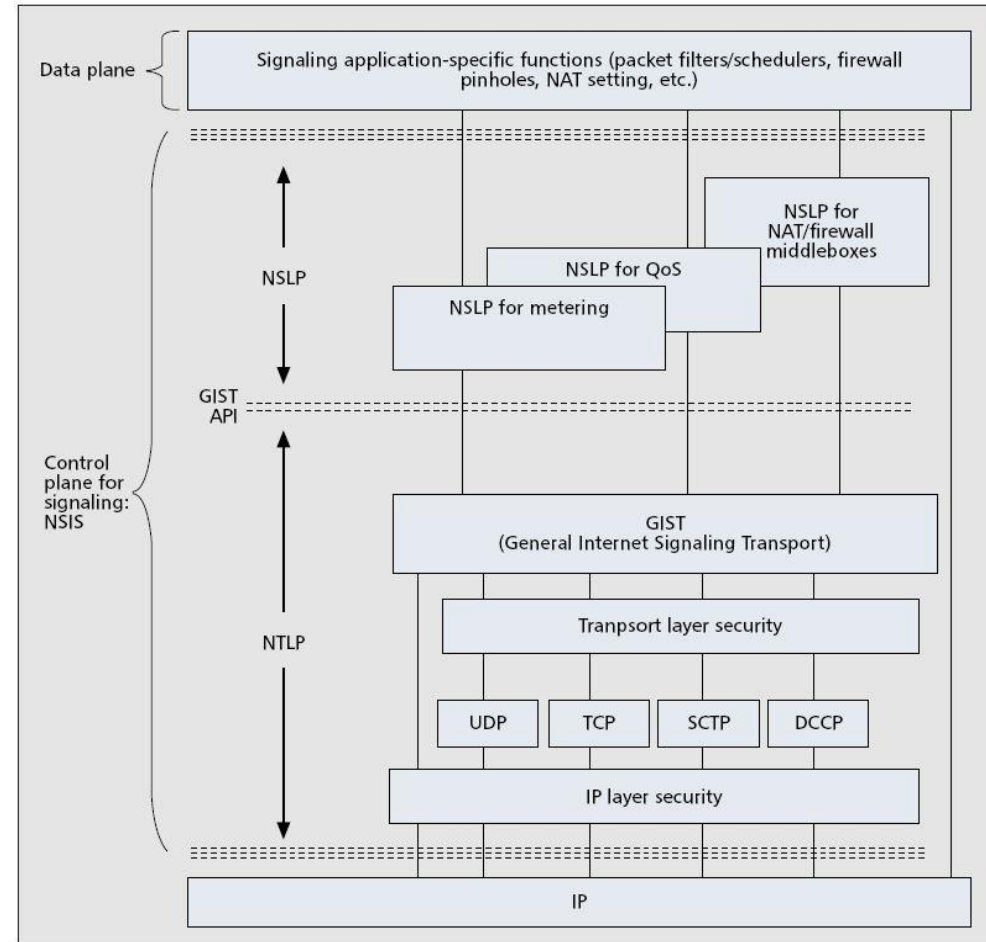
i.4 Content Distribution Logic (2/3)

NSIS signaling

- suite of protocols envisioned to support various signaling application
- IETF RFC 4080

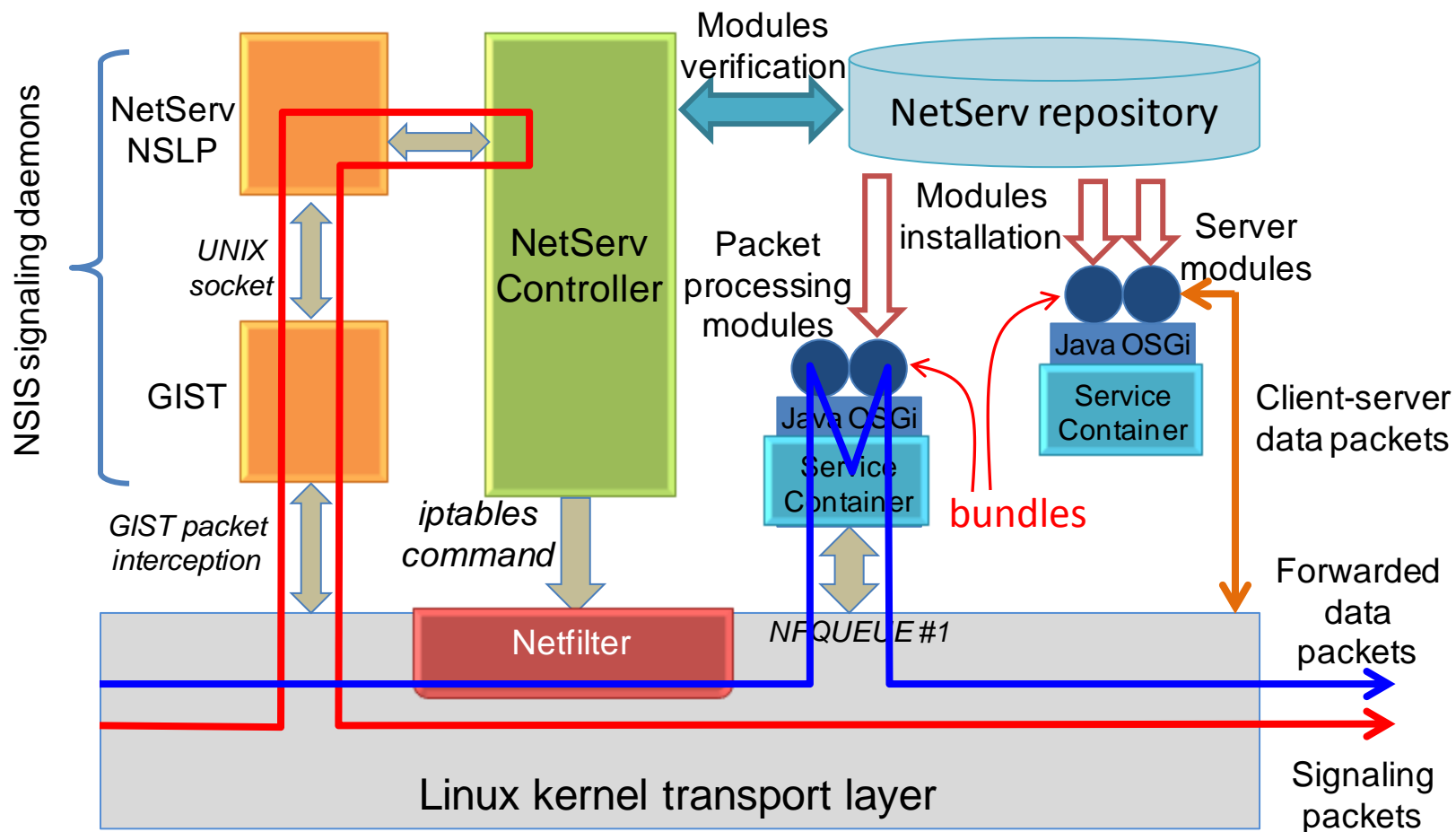
Two layers:

- NTLP: NSIS Transport Layer Protocol
 - GIST (Generic Internet Signaling Transport)
- NSLP: NSIS Signaling Layer Protocol
 - NetServ-specific NSLP
 - On-path based signaling
 - Three messages
 - » SETUP + ACK
 - » PROBE REQUEST/RESPONSE
 - » REMOVE + ACK



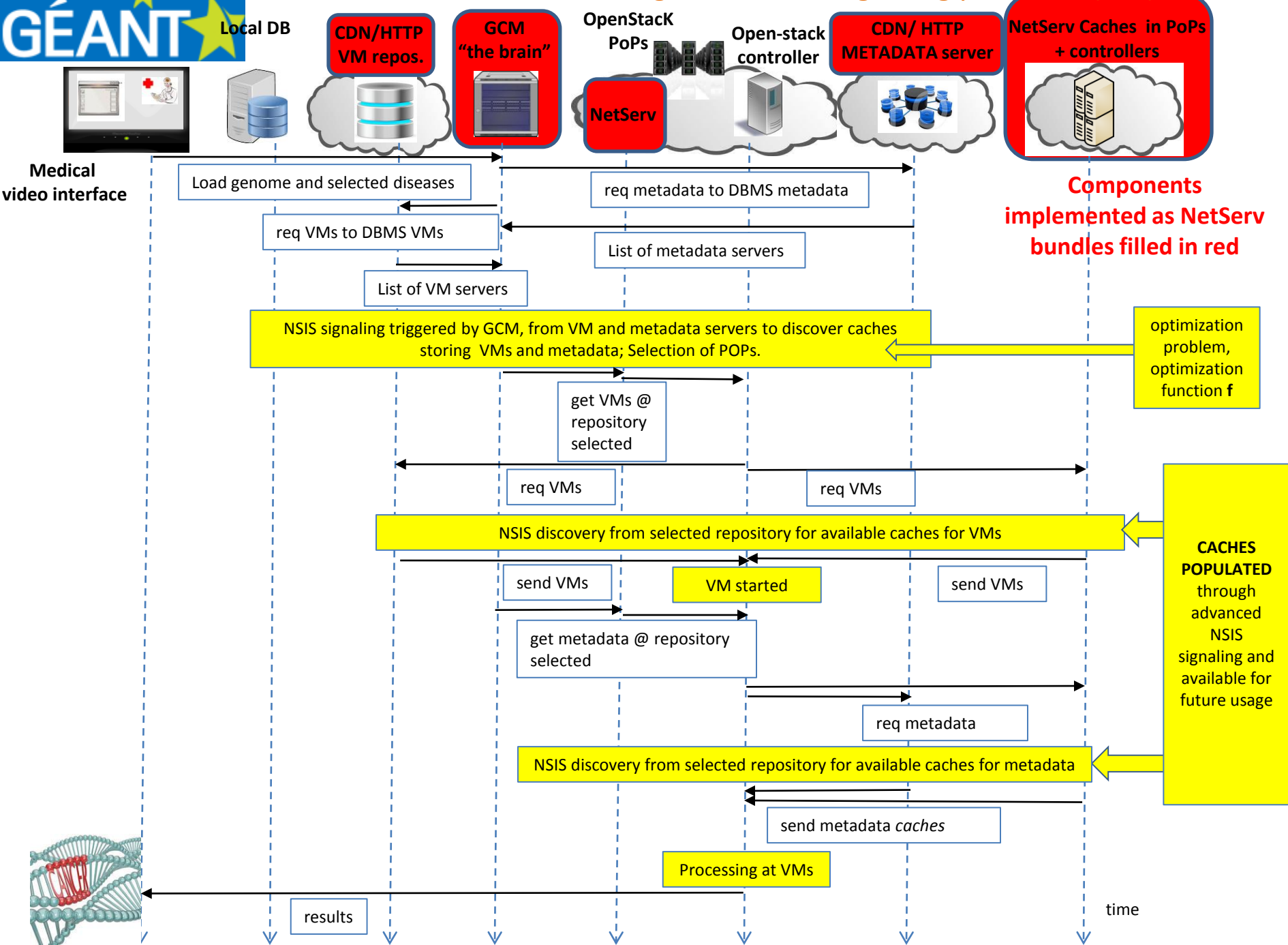
i.4 Content Distribution Logic (3/3)

The NetServ Architecture (developed in collaboration with Columbia University)

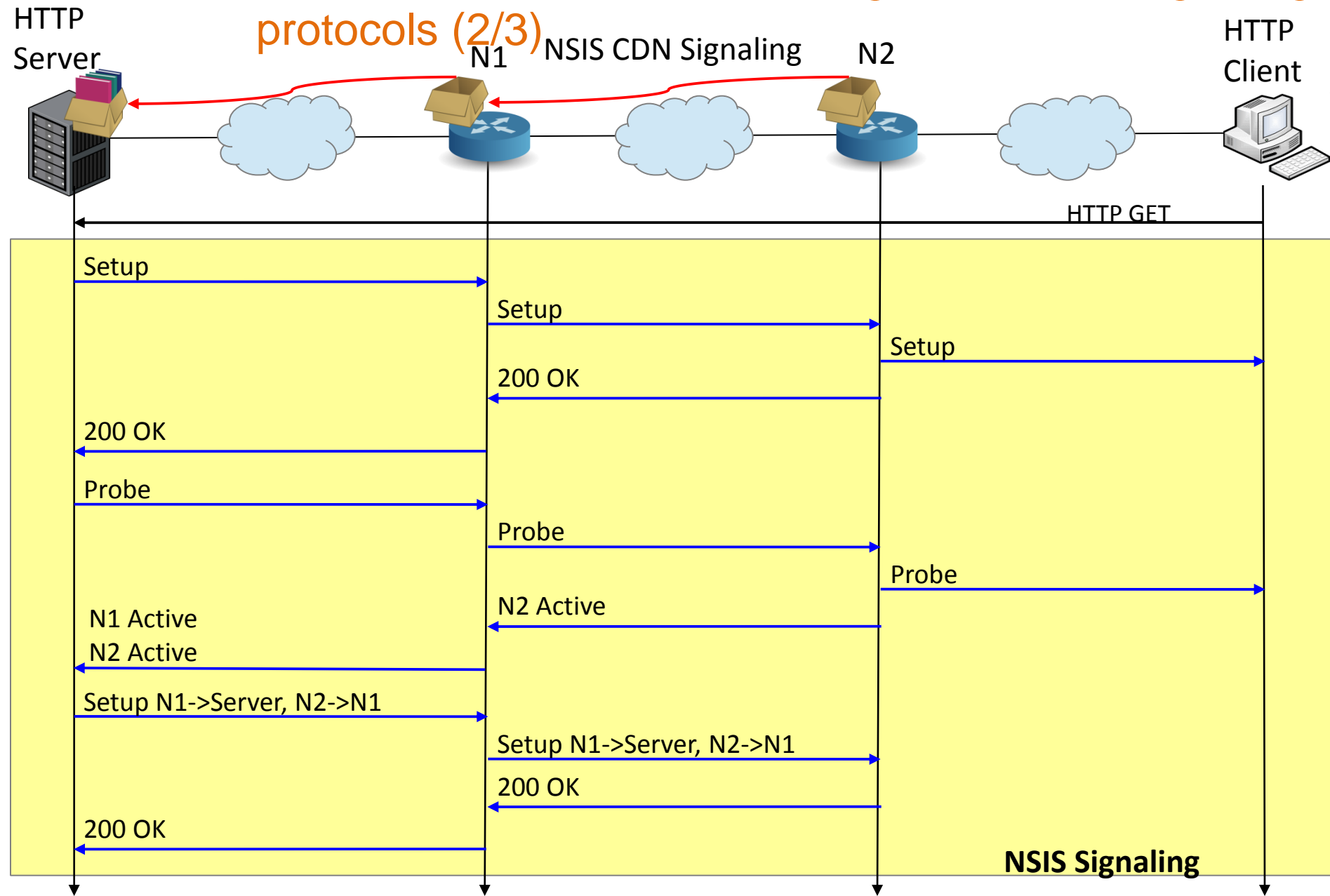


i.5 Suitable integration with cloud storage and processing services

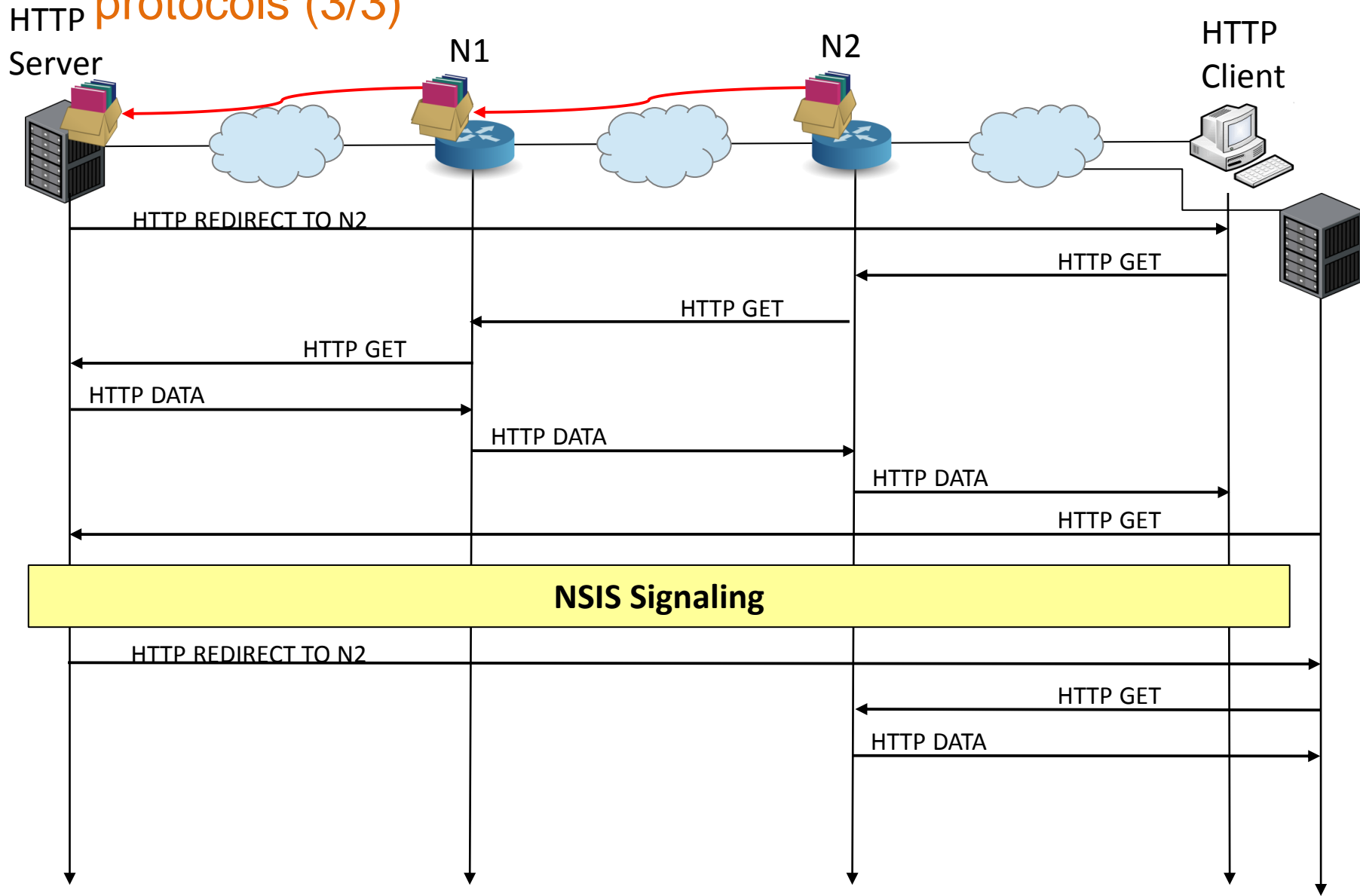
- The NSIS driven caching allows accessing data, suitably located, through a cloud-like interface.
- Extensive virtualization through the IaaS OpenStack service allows aggregating computing resources and storage.



i.6 Novel cache instantiation algorithms and signaling protocols (2/3)

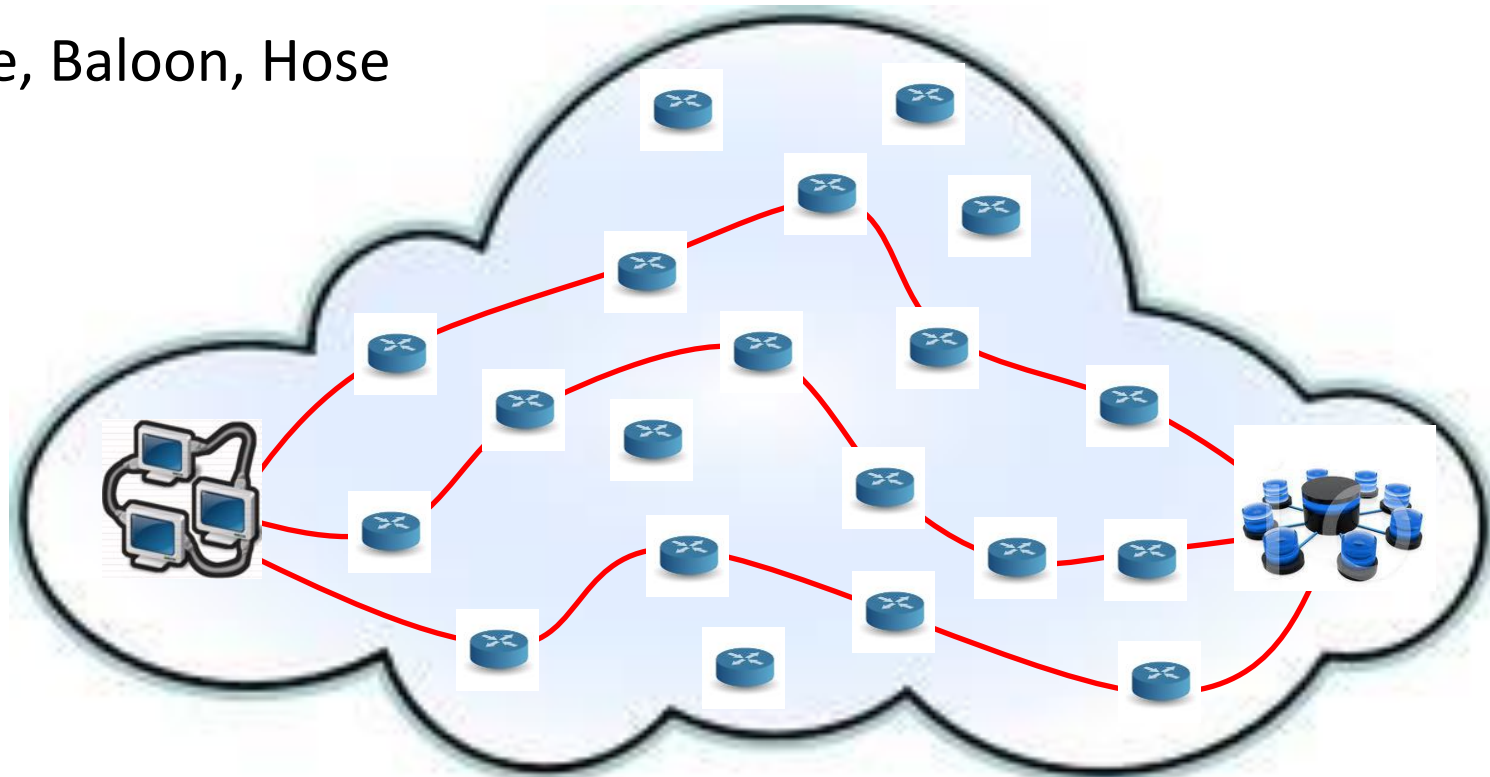


i.6 Novel cache instantiation algorithms and signaling protocols (3/3)



i.7 Parallel downloading (1/2)

- Use of a novel NSIS NSLP protocol for discovering bottleneck disjoint paths of NSIS nodes.
 - Off-path NSIS signaling
 - Bubble, Baloon, Hose



i.7 Parallel downloading (2/2)

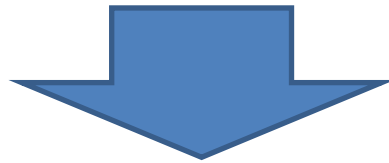
- Optimization function $f(g_1, \dots, g_k)$

being g_i a function of the i^{th} medical service request

- g_i (genome size, metadata size and location, VM size, network topology and link bandwidths, required clinical service time, quality of the sequencing machine, processing reliability, download parallelization capabilities...)

i.8 Multiple classes of network services supporting different medical needs (1/2).

- e.g. peripheral neuroblastic tumours (Neuroblastoma, Ganglioneuroblastoma, Ganglioneuroma) must be diagnosed immediately, breast cancer may be handled in some days, diabetes diagnosis can be done in two weeks



- **Different CDN services** must be provided, such as:
 - **Minimum delay CDN services** for handling urgent situations.
 - **Short delay CDN services** for handling less urgent situations.
 - **Balanced network load CDN services** for handling all other situations.

i.8 Multiple classes of network services supporting different medical needs (2/2).

The table below shows **some examples** of tolerable times for medical personnel requiring support from the project. These tolerable times include the CDN service time, in addition to other times which depends on other medical requirements, such as the type of the sequencing, the portion of the genome to be analyzed, the processing software used and the reliability of results. Through the expertise of the researchers involved in ARES, we will translate these times in CDN service classes.

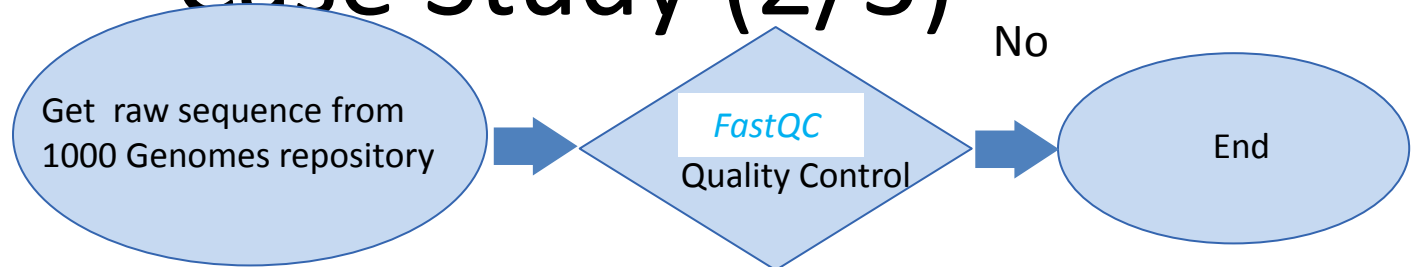
<u>Diseases</u>	<u>Time (days)</u>
Neuroblastoma	2
Breast Cancer	7
Colon Cancer	7
Acute Lymphoblastic Leukemia	4
Leukemias	4
Lymphomas	4
Myeloma	7
Cervical Cancer	7
Pancreatic Cancer	4

Case study(1/3)

Sample case study:

1. A doctor needs to investigate the occurrence of a gene mutation.
2. Assume that a **Copy Number Variation (CNV)** analysis is needed for this purpose.
3. The **appropriate CDN service** provide the data needed
4. The CNV analysis can start, as shown in what follows.
5. Outcome for measuring the **client-side** success of the procedure: achievement of results within the pre-established timeframe, compliant with the CDN service deployed.

Case Study (2/3)



Sample case study on genome mutation: find Copy Number Variation (CNV)

FastQC OSS is used for quality control.

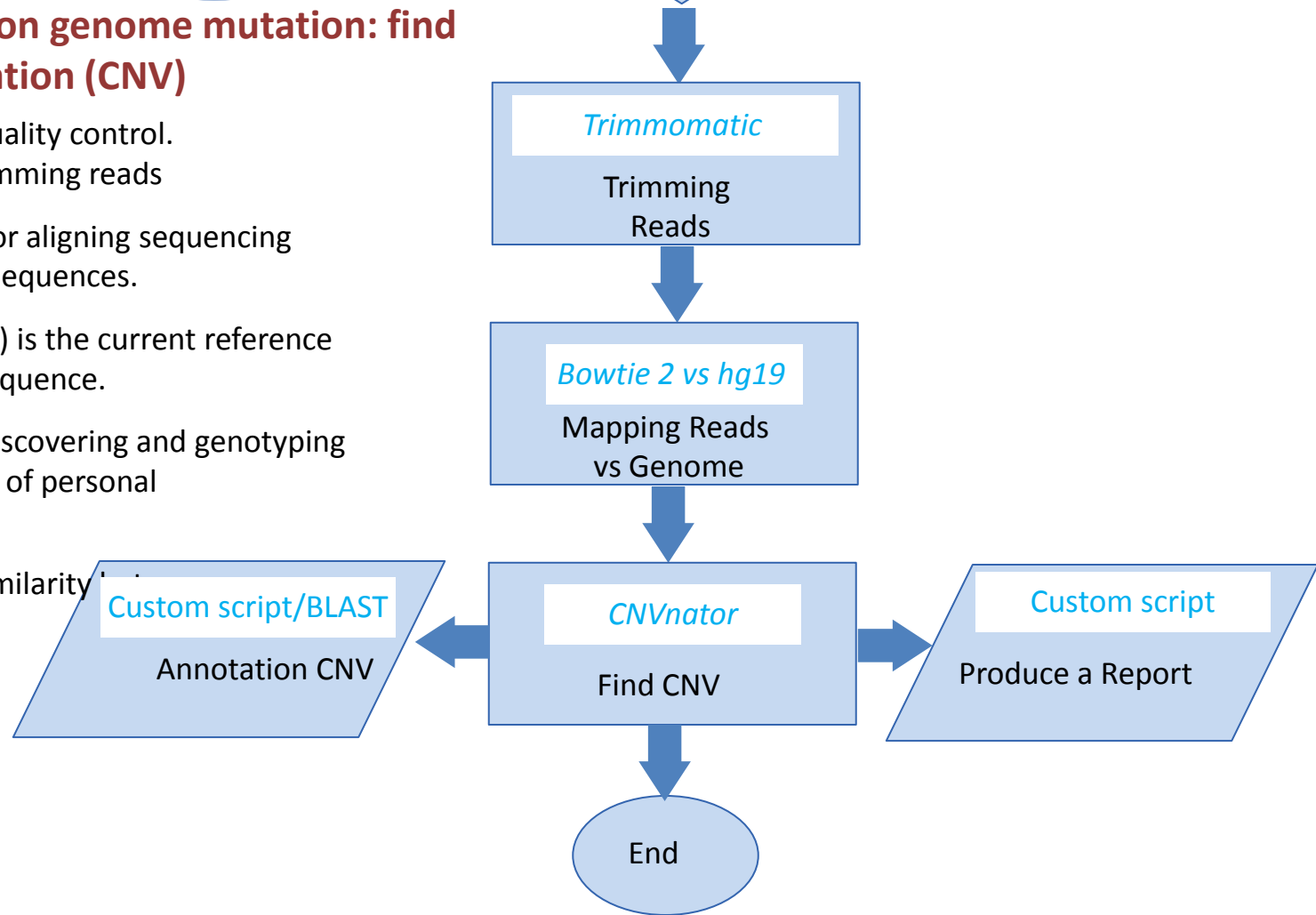
Trimmomatic OSS for trimming reads

Bowtie 2 is an OSS tool for aligning sequencing reads to long reference sequences.

hg19 (human genome 19) is the current reference to the human genome sequence.

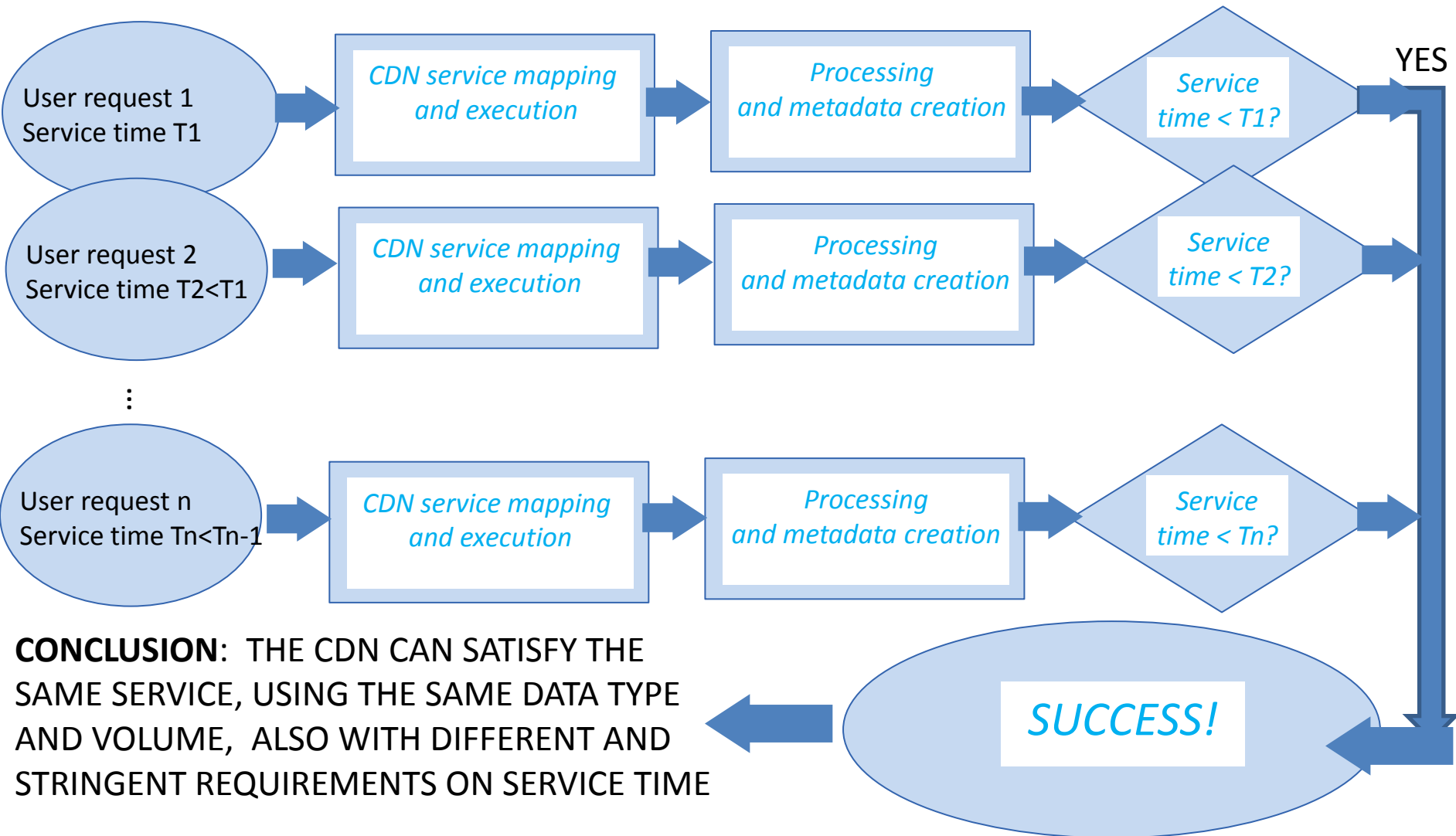
CNVnator is an OSS for discovering and genotyping from read-depth analysis of personal genome sequencing.

BLAST finds regions of similarity biological sequences

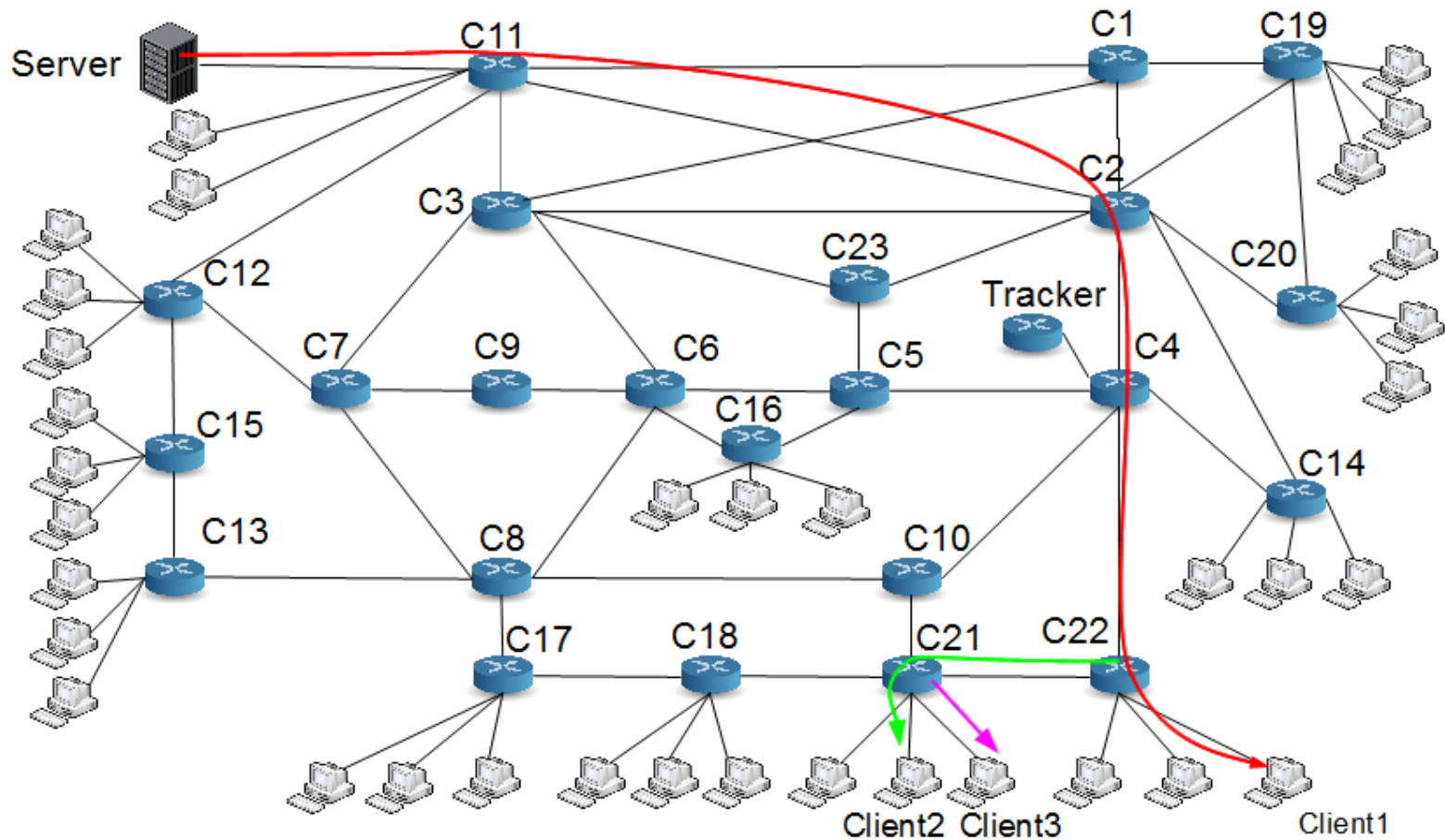


Case Study (3/3)

METROLOGICAL VALIDATION TEST: EXECUTION OF THE SAME DATA PROCESSING REQUIRING DIFFERENT TIME SPECIFICATIONS SO AS TO STRESS THE NETWORK CAPABILITIES.



Demo @ UoP Networking Lab



Expected Results

Similar metrological approaches, based on the **GUM** (Guide to the expression of uncertainty in measurement) specifications, will be implemented through multiple experiments, used to collect also **network-side** metrics.

- **Access transparency:** the set of CDN services are accessible regardless the user locations, to be verified experimentally. Success= successful verification for all locations.
- **Location transparency:** the NSIS signaling provides transparency to any change of the repository locations. Success=transparency verified for all PoPs.
- **Availability:** according to the *CAP theorem*, a distributed information system *cannot* guarantee consistency, availability, and partition-tolerance at the same time. The achievable availability for all CDN classes will be investigated in relation to the tolerable service time and the metrics illustrated below.
- **Failure transparency or Partition tolerance:** CDN service are robust to PoP and router failures. We will show how the system can manage and overcome node failures. In particular, the client programs will operate correctly after a server or repository failure. Repeated failures will be emulated so as to investigate and maximize the actual robustness. This metric is strictly related to access transparency.
- **Consistency:** the cache instantiation and update procedures will guarantee metadata consistency. This metric is strictly related to location transparency. Repeated experiments, also in the presence of node failures, will be executed. Any experiment will be considered successful if all caches are synchronized with the relevant metadata.
- **Scalability:** CDN services will allow increasing the tolerable network load and also scale gracefully to huge ones. Scalability will be analyzed and optimized in relation to the suitable trade-off induced by the CAP theorem.



Thank you for your attention!