ARES: Advanced Networking for Distributing Genomic Data

Mauro Femminella, Emilia Nunzi, Gianluca Reali, Dario Valocchi.
University of Perugia

TERENA Networking Conference
Dublin, May 19, 2014
Outline

• Description of ARES
• ARES research and implementation purposes
• Technologies and design choices
• Experimental 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,...

ARES Idea (1/2)

Combined use of CDN and CLOUD/GRID technologies, specifically targeted to genomic data sets, supporting medical needs.
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.9 Networking and service paradigms contributing to the optimized solution
For just 1000 samples!
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

- Not predictable shape, but it never expires!!!!
- Only arrivals process!!!
- Huge implications for CDNs!
content relationships based on gene “affinity”

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
For example, a diagnosis of Colon Cancer could induce further investigation about genetically similar diseases, such as Leukemia. The relevant metadata can be pre-loaded in suitable CDN caches.
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**
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
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 (1/3)

- Medical video interface
- Local DB
- CDN/HTTP VM repos.
- GCM “the brain”
- OpenStackK PoPs
- Open-stack controller
- CDN/ HTTP METADATA server
- NetServ Caches in PoPs + controllers

Components implemented as NetServ bundles filled in red

Load genome and selected diseases

req VMs to DBMS VMs

List of VM servers

req metadata to DBMS metadata

List of metadata servers

NSIS signaling triggered by GCM, from VM and metadata servers to discover caches storing VMs and metadata; Selection of PoPs.

get VMs @ repository selected

req VMs

req VMs

NSIS discovery from selected repository for available caches for VMs

send VMs

VM started

send VMs

get metadata @ repository selected

req metadata

NSIS discovery from selected repository for available caches for metadata

send metadata caches

Processing at VMs

results

optimization problem, optimization function $f$

CACHES POPULATED through advanced NSIS signaling and available for future usage
i.6 Novel cache instantiation algorithms and signaling protocols (2/3)

HTTP Server

N1

NSIS CDN Signaling

N2

HTTP Client

HTTP GET

Setup

200 OK

Probe

N1 Active

N2 Active

Setup N1->Server, N2->N1

200 OK

N2 Active

Setup N1->Server, N2->N1

Probe

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

HTTP Server -> N1
  HTTP REDIRECT TO N2
  HTTP GET
  HTTP DATA

N1 -> N2
  HTTP GET
  HTTP DATA

N2 -> HTTP Client
  HTTP REDIRECT TO N2
  HTTP GET
  HTTP DATA

NSIS Signaling
• Use of a novel NSIS NSLP protocol for discovering bottleneck disjoint paths of NSIS nodes.
  – Off-path NSIS signaling
    • Bubble, Baloon, Hose
• Optimization function $f(g_1, \ldots, g_k)$

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

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

<table>
<thead>
<tr>
<th>Diseases</th>
<th>Time (days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neuroblastoma</td>
<td>2</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>7</td>
</tr>
<tr>
<td>Colon Cancer</td>
<td>7</td>
</tr>
<tr>
<td>Acute Lymphoblastic Leukemia</td>
<td>4</td>
</tr>
<tr>
<td>Leukemias</td>
<td>4</td>
</tr>
<tr>
<td>Lymphomas</td>
<td>4</td>
</tr>
<tr>
<td>Myeloma</td>
<td>7</td>
</tr>
<tr>
<td>Cervical Cancer</td>
<td>7</td>
</tr>
<tr>
<td>Pancreatic Cancer</td>
<td>4</td>
</tr>
</tbody>
</table>
1.9 Networking and service paradigms contributing to the optimized solution

Medical interface for private genome management

SaaS Cloud service

Optimized Solution

NetServ CDN with dynamic caching Genomic Data Sets

IaaS Cloud service Virtualized genome processing software
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.
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 within 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.

User request 1
Service time $T_1$

User request 2
Service time $T_2 < T_1$

\[ \vdots \]

User request $n$
Service time $T_n < T_{n-1}$

\[ \text{CDN service mapping and execution} \rightarrow \text{Processing and metadata creation} \rightarrow \text{Service time} < T_1? \]

\[ \text{CDN service mapping and execution} \rightarrow \text{Processing and metadata creation} \rightarrow \text{Service time} < T_2? \]

\[ \vdots \]

\[ \text{CDN service mapping and execution} \rightarrow \text{Processing and metadata creation} \rightarrow \text{Service time} < T_n? \]

**CONCLUSION:** The CDN can satisfy the same service, using the same data type and volume, also with different and stringent requirements on service time.

**SUCCESS!**
## Experimental Results

<table>
<thead>
<tr>
<th>Pipeline</th>
<th>Configuration</th>
<th>Hypervisor</th>
<th>VM image size</th>
<th>RAM size min</th>
<th># CPU cores min</th>
<th>VM storage size</th>
<th>Auxiliary files size</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNV</td>
<td>BOWTIE aligner (the computing is performed on the whole human genome)</td>
<td>KVM</td>
<td>3.1 GB</td>
<td>8 GB</td>
<td>1</td>
<td>50 GB</td>
<td>3.5 GB</td>
</tr>
<tr>
<td>CNV</td>
<td>BOWTIE aligner (the computing is performed chromosome by chromosome)</td>
<td>KVM</td>
<td>3.1 GB</td>
<td>4 GB</td>
<td>1</td>
<td>50 GB</td>
<td>3.5 GB</td>
</tr>
<tr>
<td>DE</td>
<td>BOWTIE aligner</td>
<td>KVM</td>
<td>3.1 GB</td>
<td>4 GB</td>
<td>1</td>
<td>80 GB</td>
<td>3.5 GB</td>
</tr>
<tr>
<td>DE</td>
<td>STAR aligner</td>
<td>KVM</td>
<td>3.1 GB</td>
<td>32 GB</td>
<td>1</td>
<td>100 GB</td>
<td>26 GB</td>
</tr>
</tbody>
</table>
Experimental Results

![Graph showing processing time vs. number of cores for different memory sizes.](image-url)
Experimental Results
Experimental Results

**Graph 1:**
- **Baseline cloud computing**

**Graph 2:**
- ARES
- ARES, caches in VM only
Thank you for your attention!