Return on Experience on Cloud Computing Issues

… a stairway to clouds …

Experts Workshop

Nov. 21st, 2013

InGeoCloudS
Inspired GEOdata CLOUD Services

Agenda

- InGeoCloudS Software Stack
- InGeoCloudS Elasticity and Scalability
  - Elastic File Server
  - Elastic Database Server
  - Elastic Web Server
  - Elastic Map Server
  - Elastic Linked Data Store
- InGeoCloudS Monitoring and Accounting

InGeoCloudS
Inspired GEOdata CLOUD Services

What is Cloud Computing

- Cloud computing comes from the convergence of:
  - ***service oriented architectures***
    - ... loose coupling of services with operating systems and technologies ...
  - ***parallel computing***
    - large scale data analysis, up to thousands of machines
  - ***virtualization***
    - independence from physical hardware

> Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. (NIST)

http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

InGeoCloudS Challenges and Cloud Computing

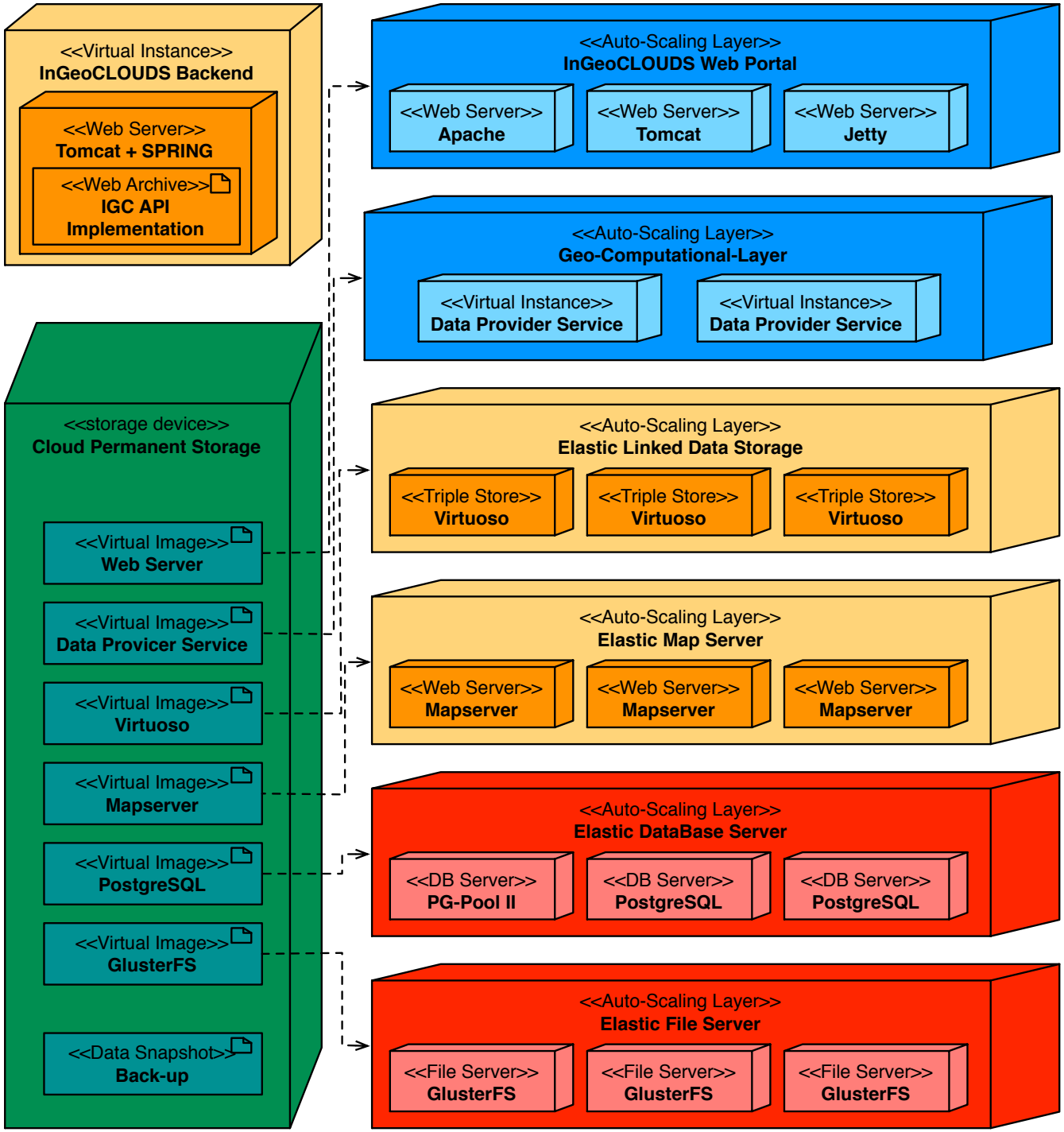
- ***Diverse software requirements***
- ***Diverse resource requirements***
- ***Resource requirements vary over time***
- ***Reduce costs***

InGeoCloudS Challenges and Cloud Computing

- ***Diverse software requirements*** **<->** ***Virtualization***
  - To support a larger number of software requirements
- ***Diverse resource requirements*** **<->** ***Scalability***
  - To support ***large data volumes*** and ***high throughput***
  - To support ***increasing dataset sizes***
- ***Resource requirements vary over time*** **<->** ***Elasticity***
  - To support a ***varying number of users***
  - To support ***on demand computations*** (e.g., shake-map)
- ***Reduce costs*** **<->** ***Pay-as-you-go***
  - To reduce ***infrastructural cost*** during low platform usage

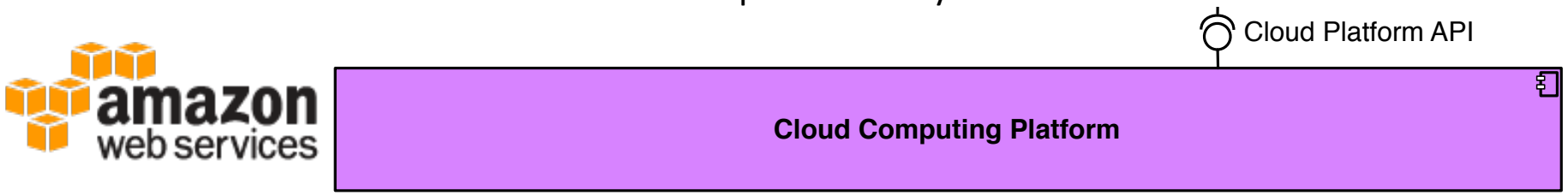
InGeoCloudS
Inspired GEOdata CLOUD Services

InGeoCLOUDS Architecture: Auto-Scaling Layers



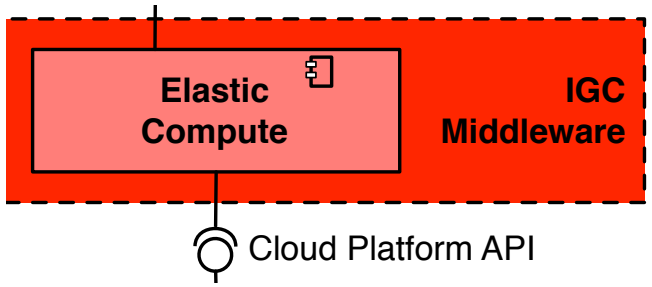
Choice of the Cloud Computing Platform

- Estimated resources:
  - 12 instances, 500GB storage, 35 GB/month network
- We analyzed several Cloud providers:
  - Amazon AWS, SigmaCloud, Atlantic.Net, Flexiant Flexiscale, GoGrid, Google App Engine, Joyent, Microsoft Azure, OpSource, Rackspace, OVH Public Cloud.
- On the basis of several criteria:
  - Functional/Software Requirements, Elasticity Model, As-a-Service Model, Maturity and Diffusion, Migration Cost Model
- Including Monthly Cost:
  - E.g., Amazon AWS €900, Rackspace €1600
    - We observed 15-20% costs drop in the last year



InGeoCloudS Elastic Compute

- This is the ***gateway*** to the Cloud Platform Services
  - ***Transparent access and portability*** to new cloud providers
- Exposed Services:
  - ***Virtual Instances Management***
    - Run a new instance, Stop an instance, attach a storage device, Elastic IP, automatically mount the distributed file system.
  - ***Auto-Scaling Layer Managment***
    - Manage an elastic pool of servers, including load balancing

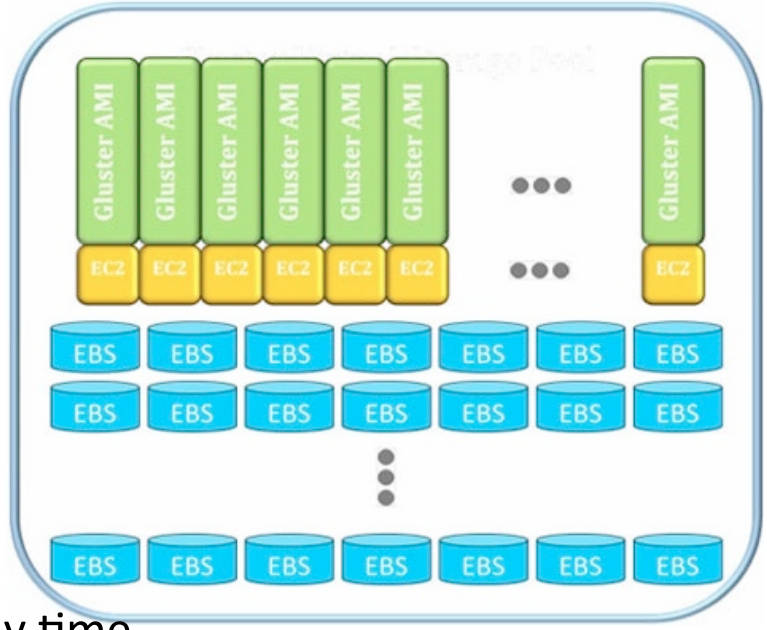


InGeoCloudS Scalable Services

- InGeoCloudS scalable services:
  - Elastic File Server
  - Elastic Database Server
  - Elastic Web Server
  - Elastic Map Server
  - Elastic Linked Data Store

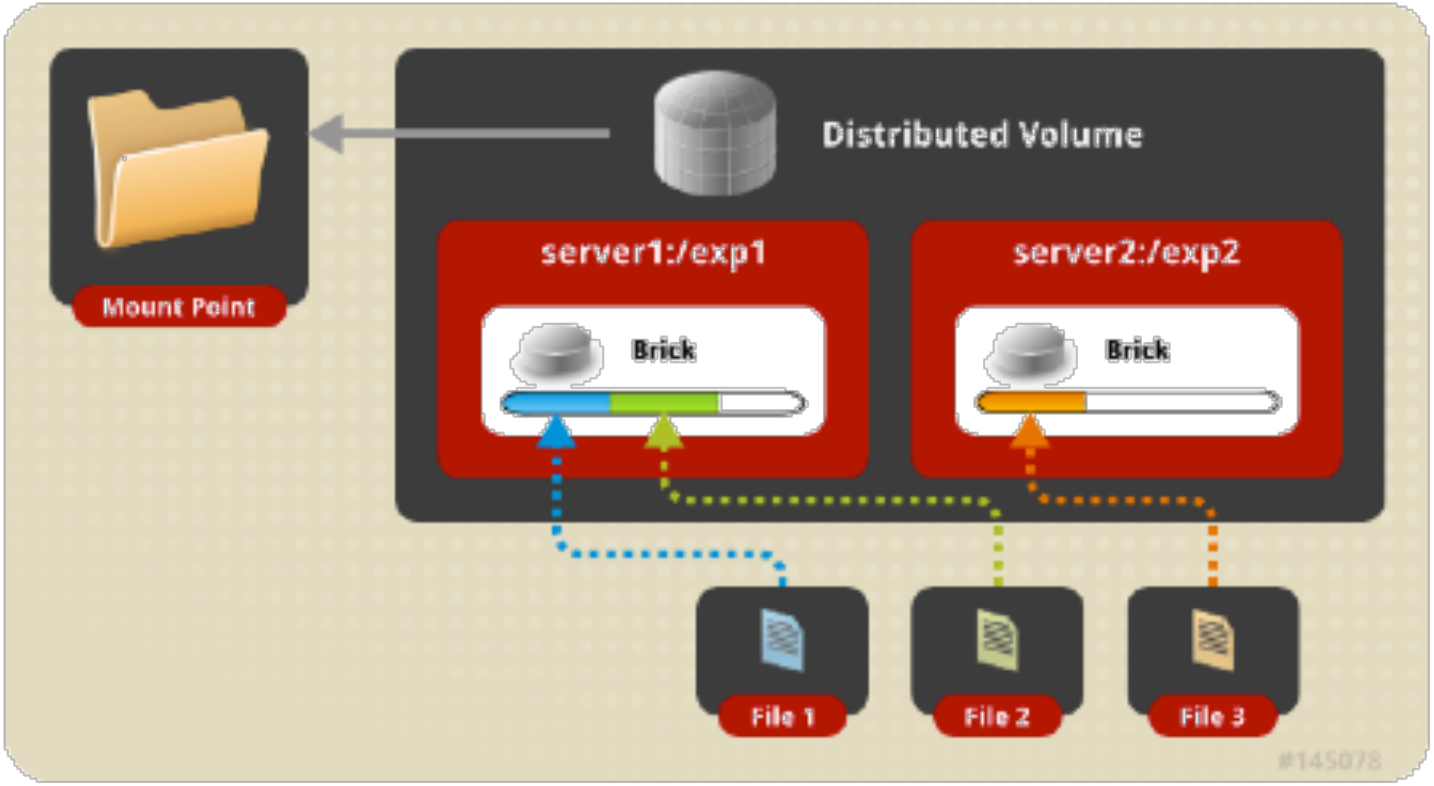
- All of the able are ***hot topics*** from a ***technological and scientific*** point of view.

Elastic File Server



- We evaluated several technologies:
  - S3FS, S3Backer, pNFS, LUSTRE, …
- Our choice was ***GlusterFS***
  - ***No single point of failure***
    - No file metadata server
  - ***Scalable***
    - Can add as many servers as needed at any time.
  - Can use ***standard protocols*** (e.g. NFS)
  - Includes some optimizations, e.g., read ahead, write behind, async I/O, scheduling, caching
- It is currently ***sponsored by RedHat***
- *Other Cloud-based storage solutions are based on the **key-value** access pattern, which is incompatible with every other technology on the Geo-Spatial Software stack*
  - *This is almost a research challenge !*

GlusterFS at work



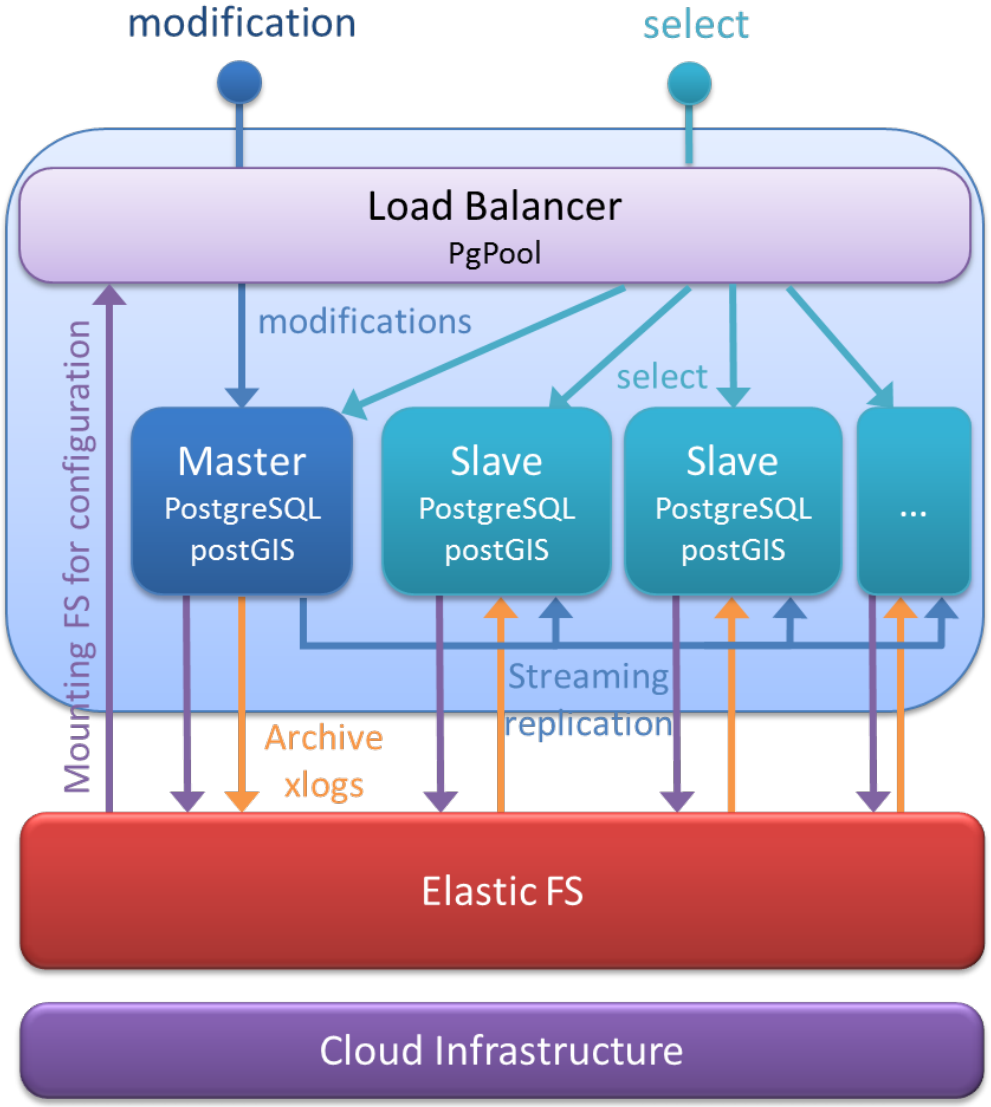
- Transparent access for applications
  - Similar to NFS. Automatic set-up on IGC instances.

InGeoCloudS
Inspired GEOdata CLOUD Services

Elastic File Server Scalability

Elastic DataBase Server

- PostgreSQL (+PostGIS)
- PgPool ***Load balancer***
  - Master/Slave architecture
  - Streaming replication
- ***Scalability***
  - Parallel read operations
  - Can add as many servers as needed at any time.
- ***Reliability***
  - Automatic fail-over
  - A slave replaces the Master

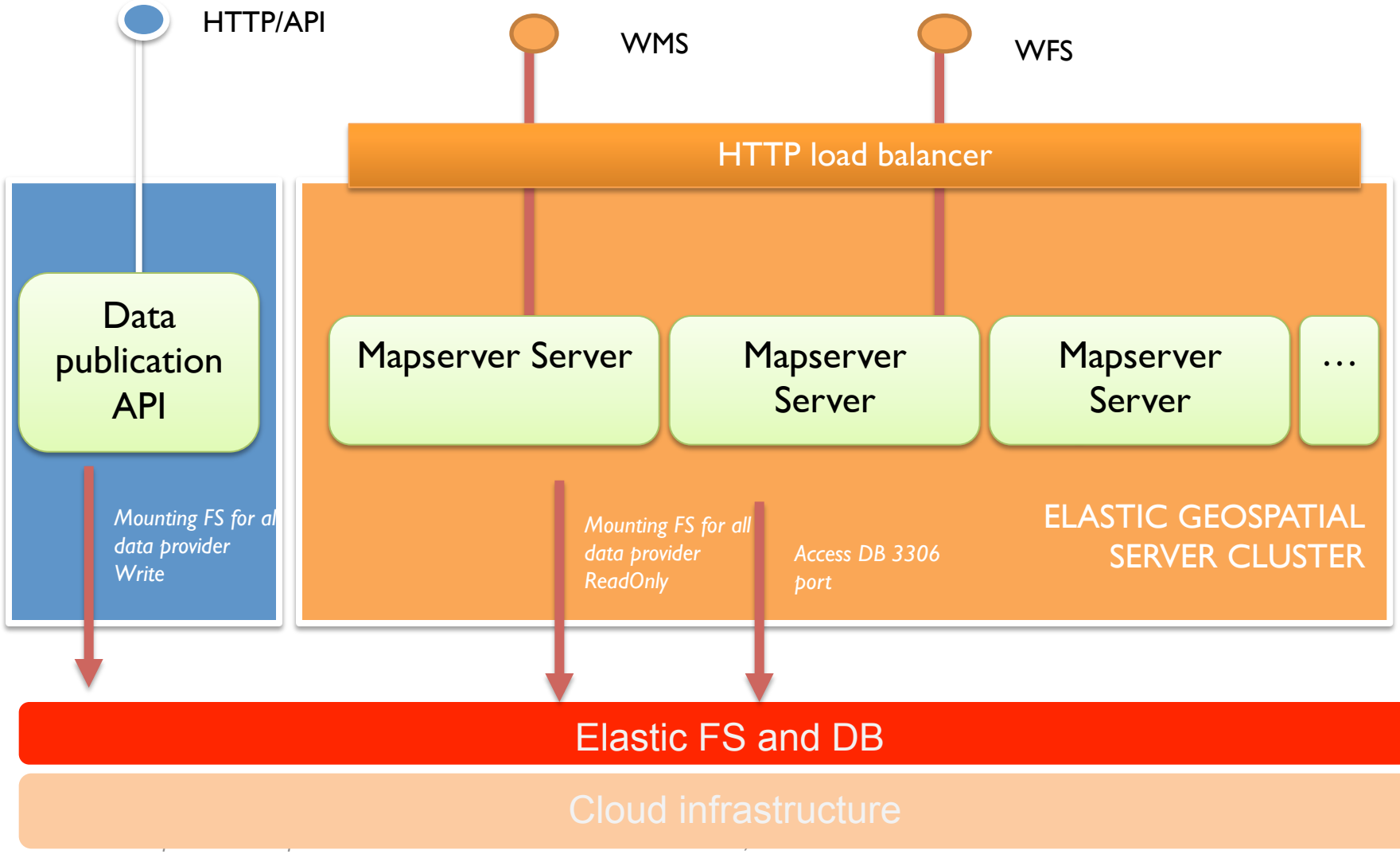


Data Publication Objectives

- Simplify the process of "transforming" ***geo-data as geo-services***
- Guarantee the geo-service compliance with ***OGC*** standards and ***INSPIRE*** requirements
- 3 components in the Data Publication :
  - Read Only services with OGC:WMS (image) and OGC:WFS (data)
  - CRUD API to manage the configuration of each service by data-provider
  - Metadata management (ISO 1911 + OGC:CSW)

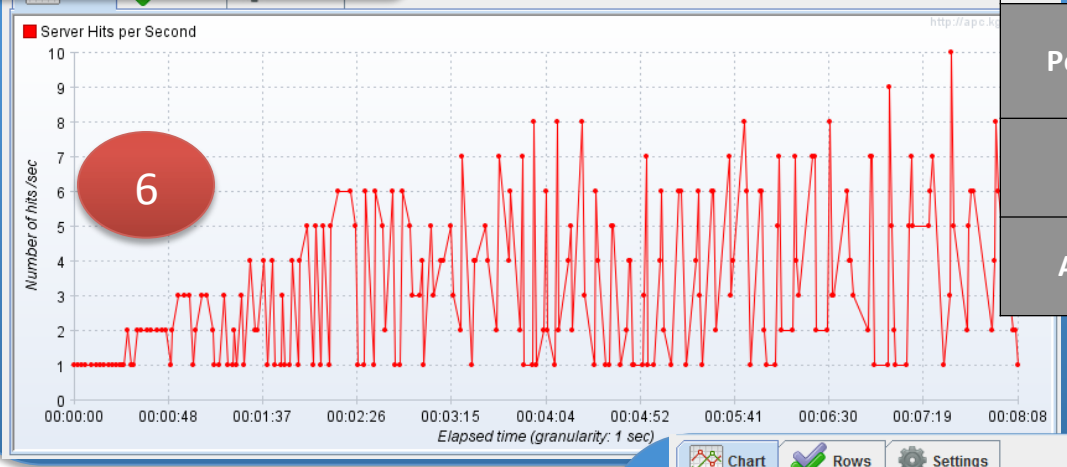
Data Publication Component Architecture

InGeoCloudS
Inspired GEOdata CLOUD Services



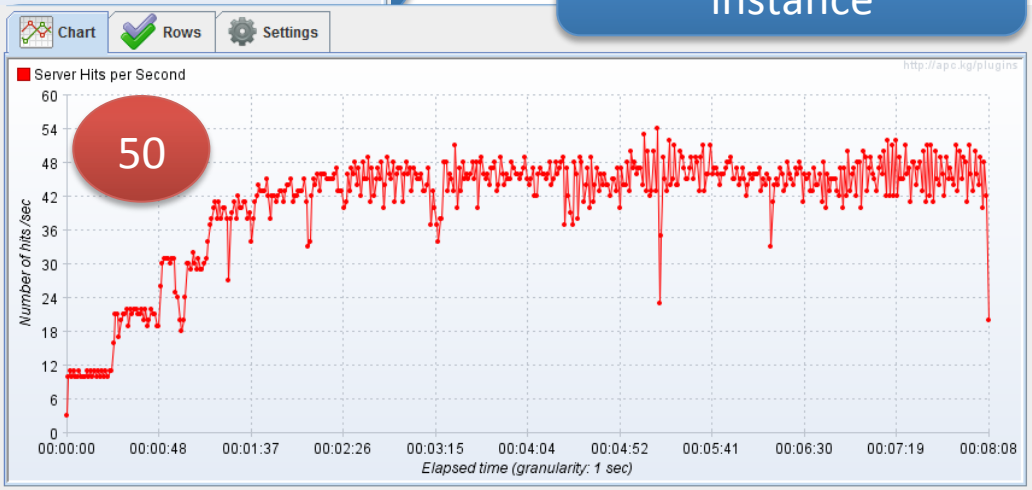
Example with the number of requests with a WMS GetMap

Small Amazon instance

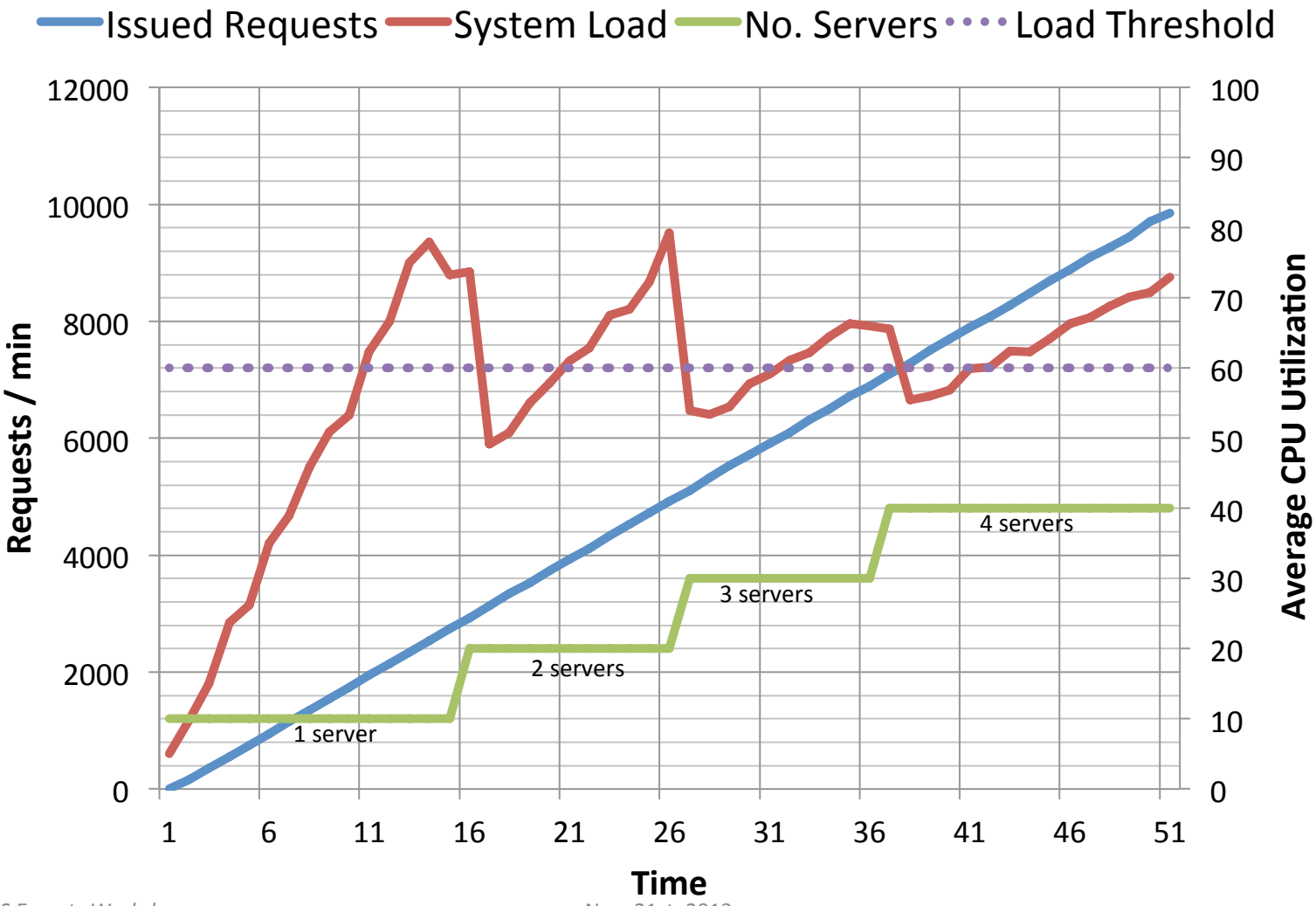


	WMS
Performance	GetMap 800x600 <5 s
Capacity	simultaneaus requests > 20/s
Availability	99%

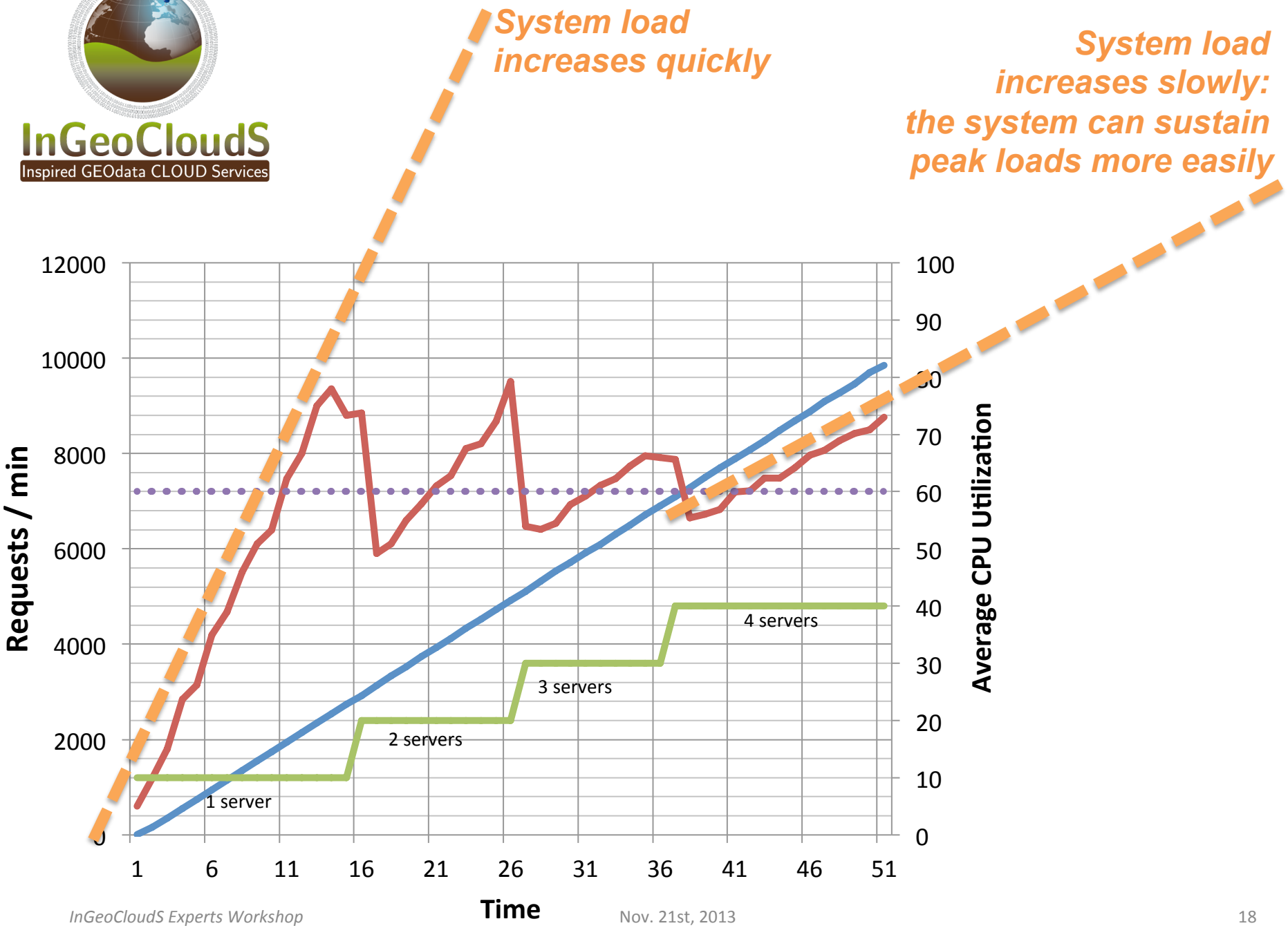
Large Amazon instance



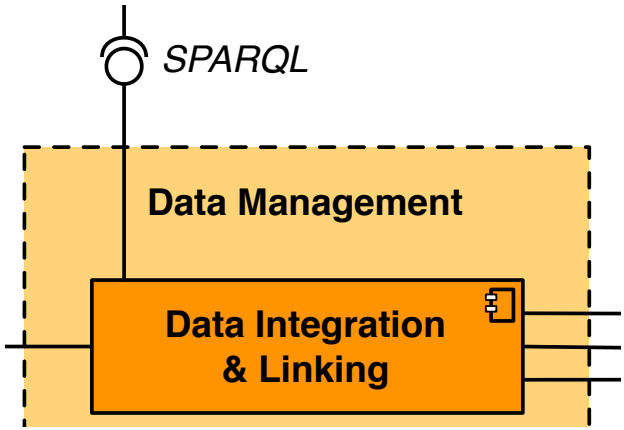
Elasticity Experiment: Elastic Web Server



InGeoCloudS
Inspired GEOdata CLOUD Services

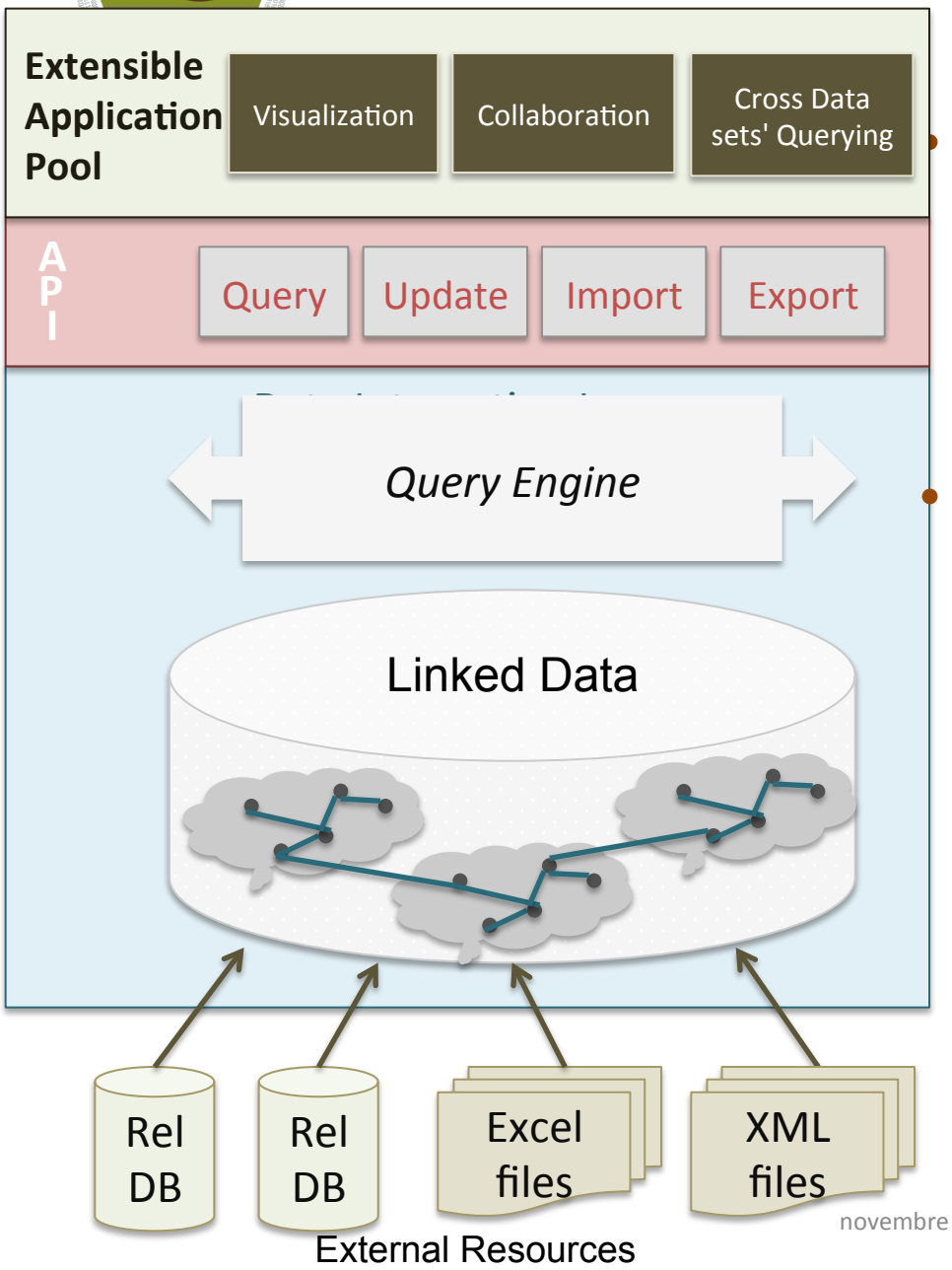


Data Integration and Linking



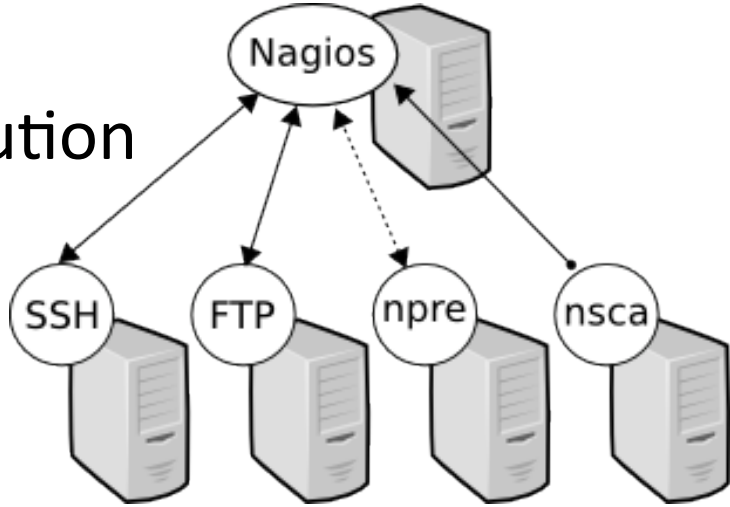
- Purpose:
  - integrate, describe and query heterogeneous data in a uniform way
- Approach:
  - Creation of a Conceptual Model to integrate and cover all the thematic fields
  - Map the source relational data into RDF data compliant to the Conceptual Model
  - Rely on a scalable RDF Triple Store (Virtuoso) to enforce the mappings and enable the storage and query of the RDF data

Linked Open Data as Service



- Abstraction layer for data access
  abstract the applications from the specific setup of the data management service (such as local vs. remote, federation, and distribution)
- Beyond Data Access
  - Enabling automation of discovery, composition, and use of datasets
  - Data Markets
  - Online Visualization Services
  - Data Publishing Solutions
  - Data Aggregators
  - BI / Analytics as a Service

Monitoring



- We are using a ***Nagios***-based solution
  - Every instance has specific *Nagios clients* generating the indicators to be monitored
  - The information received by Nagios is then stored in a *Amazon RDS*
  - We can analyze the monitoring indicators at any point in time, even when the platform is not running
  - Indicators include:
    - Avg. CPU load, memory, disk usage, response time, etc.
  - We developed a dedicated interface
    - Which is intended for admin use

InGeoCloudS
Inspired GEOdata CLOUD Services

Monitoring

Inspired Geodata Cloud Services

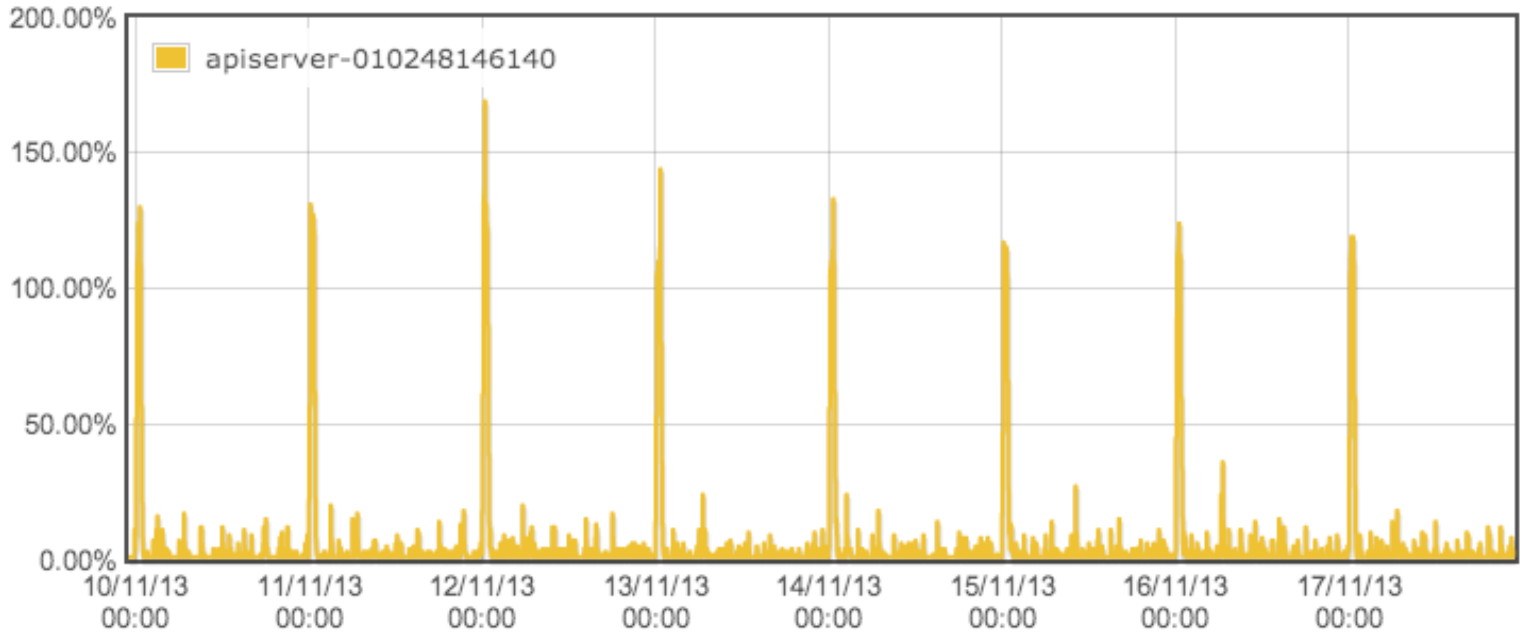
Tables

API
- cpu
- **load average**
- memory

DB
- archive xlogs
- postgresql client proc nr
- cpu
- load average
- memory
- response time

MONITORING
- cpu
- load average
- memory

I_API_CPU_USED Table



Accounting Service

- We can have per-service cost from Amazon billing

- *Elastic Database Server* cost:
  - Compute hours/month …….. XXX $
  - Storage GB/month …….. XXX $
  - Data transfer …….. XXX $

- This allows to estimate the cost of the IGC platform components
  - Also useful for you own private IGC platform deployment
- We need more:
  - Per-user split of costs

Accounting Service

- IGC provides Accounting APIs
  - They provide a detailed user's share of cost

- For each Data Provider:
  - *Elastic Web Server* …… XXX $
  - *Elastic Map Server* …… XXX $
  - Other ……
  - GRAND TOTAL …… *$ not a lot $*

- This is computed:
  - By measuring directly storage occupancy (both DB and FS)
  - By application logs to estimate usage shares of indivisible services (e.g., compute hours of Map Server)

Accounting Service

- So… how much does it cost ?

- We will this discuss later in the session "***InGeoCloudS Sustainability, Costs, and Opportunities for Cooperation and Trials***"

Conclusions

- InGeoCloudS is an interesting and evolving ***cloud-based platform for geo-data providers***
- The IGC platform was designed on the basis of actual data providers use cases:
  - To support multiple applications
  - To enable fast porting to the cloud
- It provides ***scalable services and on-demand computation***, by taking advantage of:
  - Cloud "infinite" resources
  - Pay-as-you-go cost model
- The platform can support a much larger number of users than the project consortium size
  - ***The more users, the smaller the cost !***

Thanks for your attention