SCIENTIFIC WORKFLOW EXECUTION ON CLOUD ENVIRONMENT: JOB SCHEDULING AND DATA REPLICATION

Israel Casas

Supervisor: Albert Zomaya Auxiliary Supervisor: Javid Taheri

School of Information Technologies

FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

WORKFLOWS

Clouds and science

- Cloud computing is breaking the barrier to access an immense pool of resources and scientific community depend on computing services to conduct their discoveries. These two ideas enclosed the demand/offer trade in computer science field. On one side, research community demand computing power to conduct their experiments including important areas as genome analysis, weather prediction, seismic analysis. On the other side, cloud providers offer a strong computing power without the initial monetary investment a computer system would require.

The problematic

- Execute computing and data intensive scientific workflows on cloud computing systems.

Objectives

- Obtain the minimum monetary cost and achieve a time constrain dictated by user.

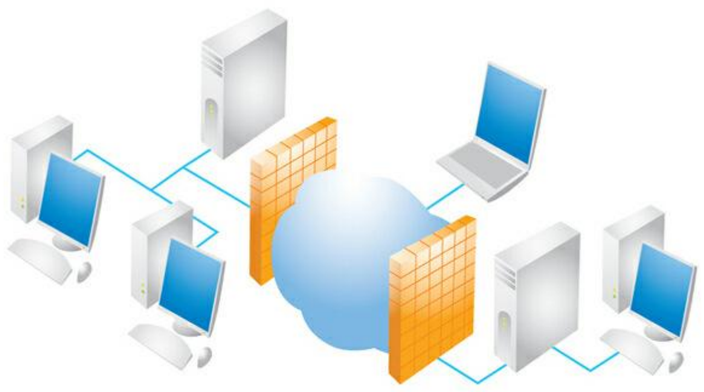
Research proposal

- This research proposes a job scheduling and data replication heuristic to execute scientific workflows on cloud environments. The static scheduler must be able to map tasks in bulk. To complement this, replica algorithm enables scheduler to manage data orientated applications by creating, distributing and deleting files for the execution of workflows. Finally, this work proposes a framework to quote user before execution. Quoting permits scientific to decide for a fast or economical execution for his workflow.

CLOUD ENVIRONMENTS

Computing systems

- Cloud computing is an environment where users have access to a pool of resources on a demand basis .
- These computing resources include network links, servers, storage nodes, applications and services.
- Cloud ownership and operation dictates three deployment models: private, public, and hybrid. In public cloud environments, user is charged for the machine-time factor he consumes. Private clouds are resources own and used by an organization. This clouds are application orientated while public environments give service to a numerous users with different needs. Hybrid clouds are a composition of two or more clouds taking advantage of the principal features of each one.
- This new computing pattern has important features that make it important for science and attractive as a business. In science field, clouds offer a large pool of resources to compute complex applications with characteristics comparable to HPC (High Performance Computing) and (High Through Computing). From a business perspective, cloud offers a profitable platform to cloud providers and clients.
- Cloud environments have five essential characteristics: on-demand self-service, broad network access, resource pooling, rapid elasticity and measured service. Additionally, clouds offer three different service models: Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS). Different research fields already employ cloud services for experiments. These areas include astronomy, high energy physics, molecular biology, earth sciences, gene sequencing, population genetics, machine learning and image processing.



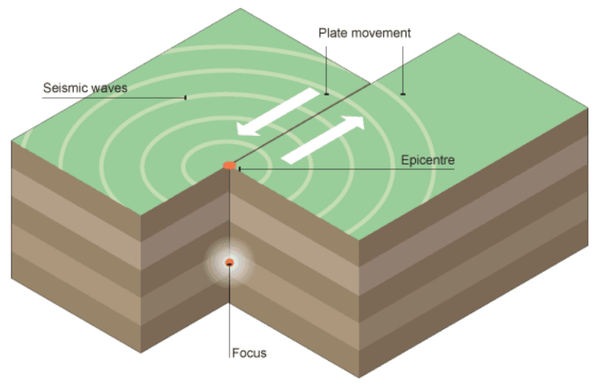
MOTIVATION

Complex applications

- Diverse scientific areas produce massive data as a result of technology advances to collect information. These data demand computing power to perform complex mathematical operations on them.

Genetics

- Genetic investigators agree genome discoveries depend directly on computer science. Genome analysis is divided in two parts: identification of nucleotides and nucleotide sequence overlap discovery. This last stage is the bottleneck for the entire genome analyses. The computing systems that execute the first stage in the genome analysis improve faster than processors of the second stage. While sequencers increase their capacity by three to five times per year (from the last five years), processors growing rate doubles only by two every two years. This difference causes a gap between sequences and analysers. For instance, the HiSeq 2000, one of the most potent genome analysis systems, is capable to process 100 nucleotides within a week. For reference the entire human genome are 3 billion nucleotides.



Seismic analysis

- Seismic analyses produce underground image representation for research purposes. Seismology data comes from a great number of signal reflections called traces. Intensive mathematical calculations are performed on each trace to validate their signal reflection correctness. This number of traces has multiplied during last decades. Recently, Petrobras researchers in Brazil utilized up to one million channels per kilometre square to create earth representation. For reference a mesh of 72 000 traces can produce 1000 GB of data. New challenges in this field demand high resolution images on 3D and even 4D images increasing the number of traces and consequently computing power for analysis.

THIS RESEARCH IS SPONSORED BY

CSIRO CONACYT