EXPLORING ADAPTIVE TECHNIQUES IN HPC WORKFLOWS

<u>Vítor Silva</u>

Ph.D. Student at COPPE/UFRJ

silva@cos.ufrj.br

Advisor: Marta Mattoso

Co-advisor: Daniel de Oliveira

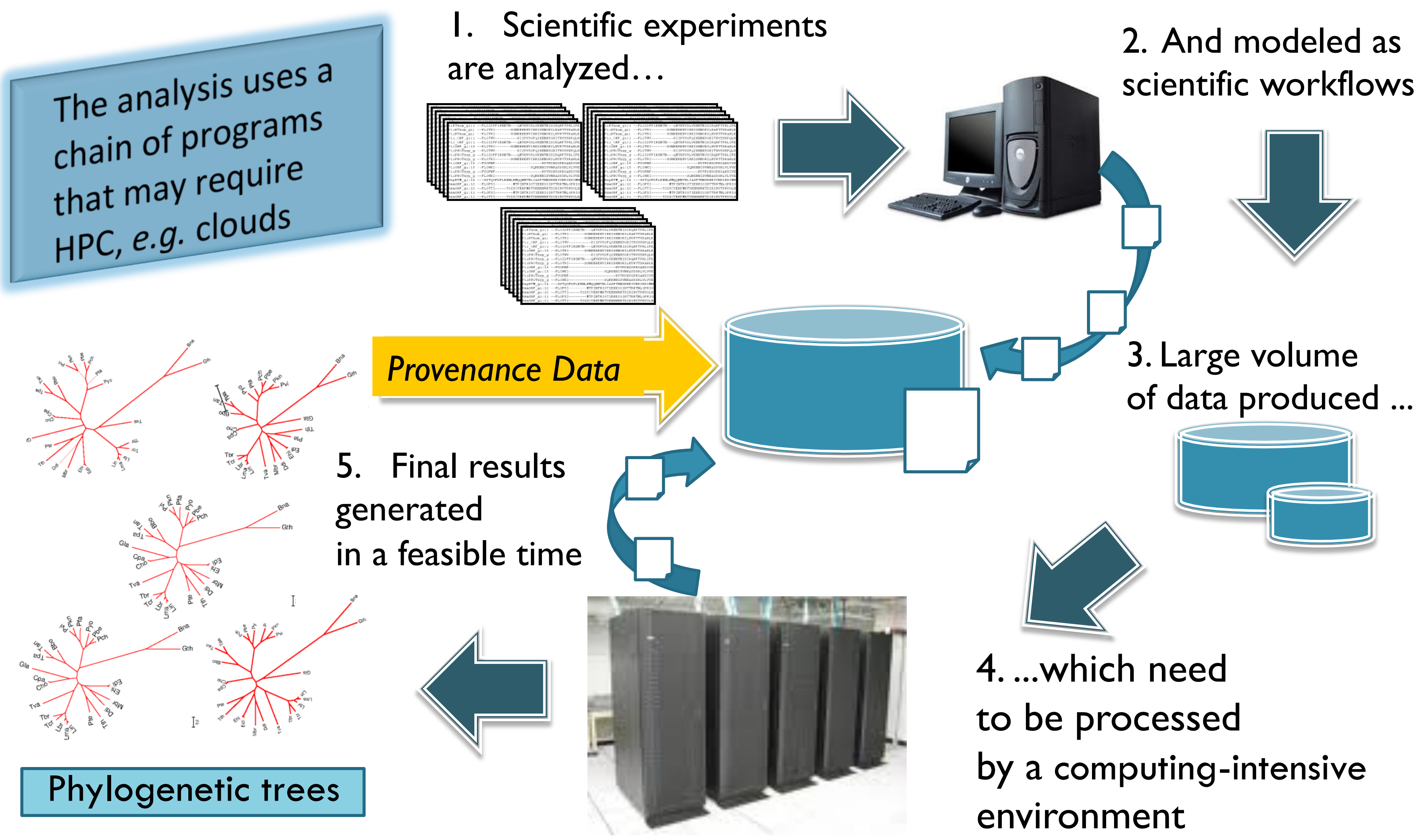
Internship at LIRMM (INRIA)

Member of SBC (Brazilian Computer Society)

COPPE UFRJ

Instituto Alberto Luiz Coimbra de Pós-Graduação e Pesquisa de Engenharia

Scientific Workflow Scenario



Problems about Scientific Workflow Parallel Execution

- Volatility of computational resources
  - *Churn* event
- Failure occurrences at runtime
  - Many computational tasks are processing data in parallel
  - Parameter sweep may present failure in some combinations
- Difficulties in debugging failures
  - Log analysis
  - Difficulties in identifying data-flow involved in a specific failure

Desired Characteristics for the Scientific Workflow Parallel Execution

- Volatility of computational resources
  - *Churn* event
- Failure occurrences
  - Many computation ... allel
  - Parameter sweeper may present failure in some combinations
- Difficulties in debu...
  - Log analysis
  - Difficulties in ident... fic failure

Adaptability

Reliability and reproducibility of experiments

Demeter: an Adaptive Execution Strategy

- An strategy for adaptive parallel execution of scientific workflows
  - Clusters environments
  - Distributed management of the provenance data
  - Queries of provenance data at runtime
  - Adaptations
    - Addition or removal of computational resources
    - Support to failure tolerance

Chiron: Parallel Workflow Execution Engine

- Chiron
  - Its engine is dataflow oriented by a workflow relational algebra
  - Apps: CFD, Risers and Uncertainty Quantification (UQ)
  - Strong provenance support
  - User steering at runtime
  - But, ...
    - Execution is controlled by one master component
    - Do not support adaptations in resource allocation

OGASAWARA, E., DIAS, J., SILVA, V., *et al.*, 2013, "Chiron: A Parallel Engine for Algebraic Scientific Workflows", *Concurrency and Computation*, v. 25, n. 16, pp. 2327–2341.

Chiron: Parallel Workflow Execution Engine

- Chiron
  - Its engine is dataflow oriented by a workflow relational algebra
  - Apps: CFD, Risers and Uncertainty Quantification (UQ)
  - Strong provenance support
  - User steering at runtime
  - But, ...
    - Execution i ... omponent
    - Do not sup ... allocation

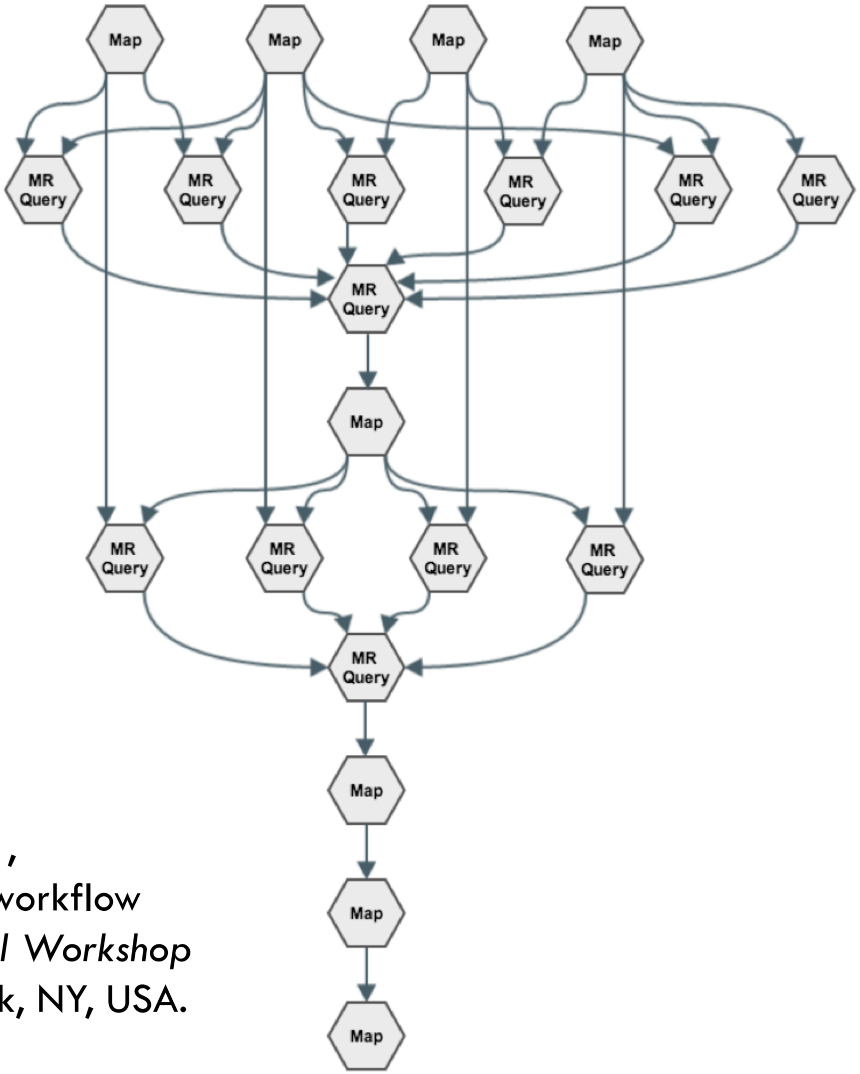
Demeter

OGASAWARA, E., DIAS, J., SILVA, V., *et al.*, 2013, "Chiron: A Parallel Engine for Algebraic Scientific Workflows", *Concurrency and Computation*, v. 25, n. 16, pp. 2327–2341.

Experiment: Montage workflow

- Synthetic Montage workflow
  - Astronomy application to blend astronomical images
  - A kind of benchmark in parallel workflow execution evaluations

- Our input dataset was based on the experiments of Vöckler *et al.* (2011)

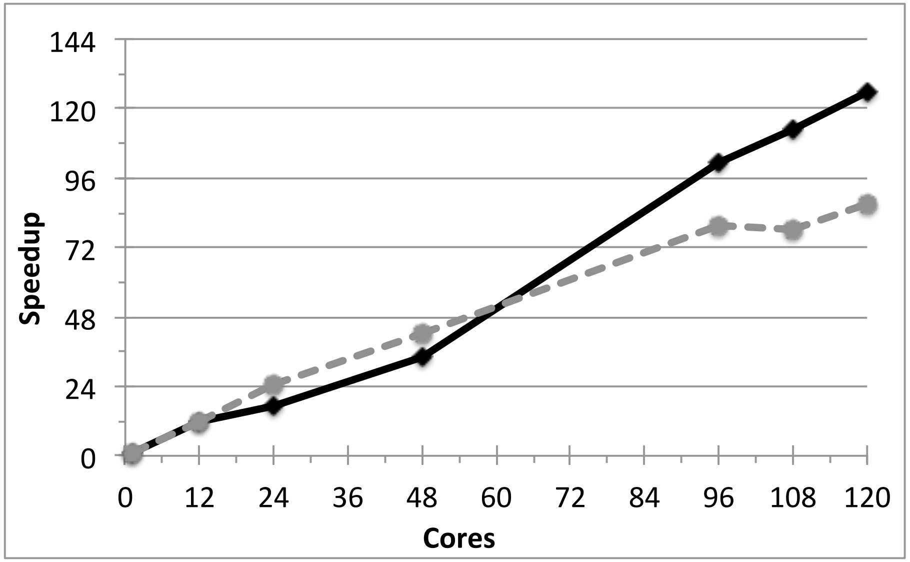


VÖCKLER, J.-S., JUVE, G., DEELMAN, E., *et al.*, 2011, "Experiences using cloud computing for a scientific workflow application". In: *Proceedings of the 2nd International Workshop on Scientific Cloud Computing*, pp. 15–24, New York, NY, USA.

Environment to Execute Montage Workflow

- Uranus – SGI Altix ICE 8400
  - Cluster environment
  - NACAD / COPPE / UFRJ
    - One of the current Intel® Parallel Computing Centers
  - 128 CPUs Intel Xeon: 640 Cores (64 nodes)
  - 1.28 TBytes of RAM memory (distributed)
  - Storage: SGI InfiniteStorage NAS (72 TBytes)

Experimental Results: Montage workflow



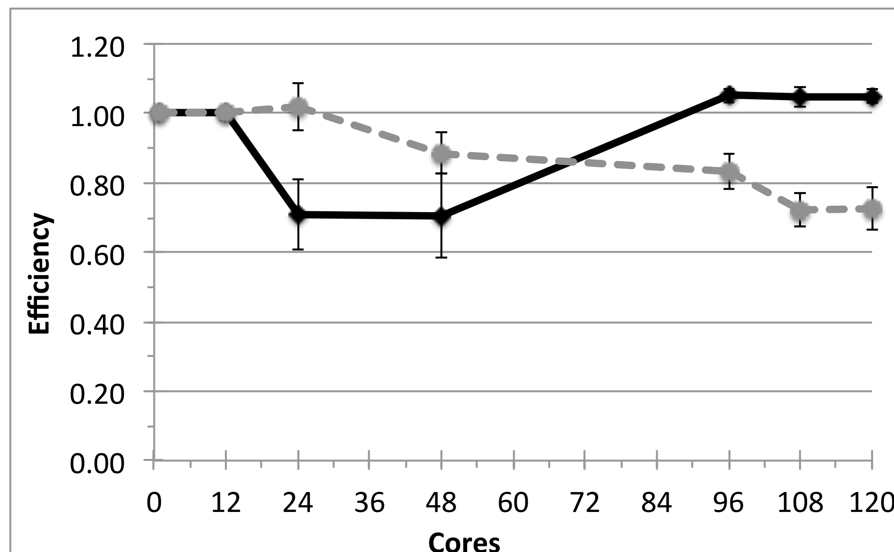
Demeter had a better performance (almost 27%) from 96 cores

Chiron

Cores	Time (min)
1	4501.20
12	375.10
24	184.00
48	106.00
96	56.30
108	57.70
120	51.70

Demeter

Cores	Time (min)
1	5364.00
12	447.00
24	315.50
48	158.50
96	53.10
108	47.40
120	42.60



Exploring Workflow Fragments during Internship at INRIA

- Development during internship (Oct– Dec/2013)

  joint work with Ji, Pacitti, Valduriez (INRIA); Oliveira (UFF) and Mattoso (UFRJ)

  - Partition workflow into several **fragments** in order to execute in multiple cloud sites
  - Publication of a paper **[1]**

- Opportunities to explore workflow fragments
  - Use different parallel execution strategies for each fragment
  - Generate optimized execution plan, based on workflow fragments

[1] Liu, J.; Silva, V.; Pacitti, E.; Valduriez, P.; Mattoso, M., "Scientific Workflow Partitioning in Multisite Cloud". In: 3rd Workshop on Big Data Management in Clouds, Proc. of the Europar 2014.

Next steps in PhD

- Profiling Chiron with Demeter
- More adaptive support
- Experiments in new Oil & Gas applications with UQ
- Workflow big data distributed management
- Supporting index and queries in binary data

THANK YOU!

EXPLORING ADAPTIVE TECHNIQUES IN HPC WORKFLOWS

Vítor Silva

Scientific Workflows in HPC

- The same problems of parallel programs
  - Volatility of computational resources
  - Failure occurrences at runtime
  - Difficulties in debugging failures

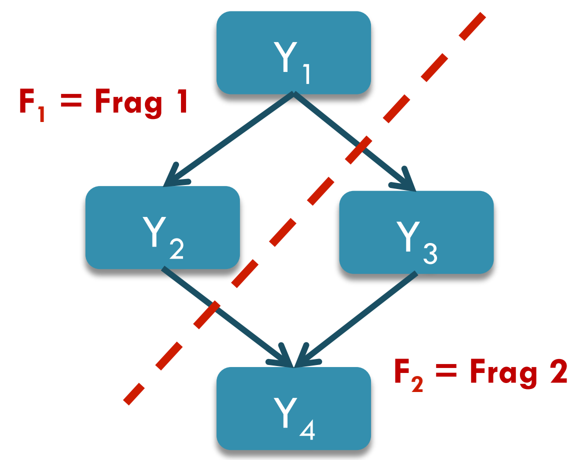
- However, …
  - SWfMS can take over debugging by taking advantage from "knowing" what is behind workflow tasks and data-flow
    - Queries about activities that presented failures
    - Adjustments in parameters or workflow modeling

Background (1)

- Workflow fragment (proposed by Ogasawara et al.)

A workflow W includes a set of activities $Y = \{Y_1, \ldots, Y_n\}$. Given $Y_i \mid (1 \leq i \leq n)$, let $R = \{R_1, \ldots, R_m\}$ be the input relation set for activity Y_i, then $Input(Y_i) \supseteq R$. Also, let T be the output relation set produced by activity Y_i, then $Output(Y_i) \supseteq T$. We denote the dependency between two activities as $Dep(Y_j, Y_i) \leftrightarrow \exists\ R_{k'} \in Input(Y_j) \mid R_{k'} \in Output(Y_i)$. Additionally, a fragment of a workflow, *fragment* for short, is a subset F of the activities of a workflow W, such that either F is an unitary set or $\forall Y_j \in F,\ \exists\ Y_i \in F \mid (Dep(Y_i, Y_j)) \vee (Dep(Y_j, Y_i))$.

Workflow W



$F_1 = \{Y_1, Y_2\}$; $F_2 = \{Y_3, Y_4\}$; $W = F_1 \cup F_2$

- Activation

Given a workflow W, a set $X = \{x_1, \ldots, x_k\}$ of activations is created for its execution. Each activation x_i belongs to a particular activity Y_j, which is represented as $Act(x_i) = Y_j$.

E. Ogasawara, J. Dias, D. Oliveira, F. Porto, P. Valduriez, and M. Mattoso, "An Algebraic Approach for Data-Centric Scientific Workflows," *Proc. of VLDB Endowment*, vol. 4, no. 12, pp. 1328–1339, 2011.

Background (2)

- Dataflow strategy
  - First Activity First (FAF) x First Tuple First (FTF)
- Dispatching strategy
  - Static x Dynamic

given a workflow W, an associated workflow activations set $X=\{x_1, \ldots, x_k\}$ is evaluated according to a schedule. The schedule of activations depends on the dataflow strategy assigned to the corresponding workflow fragment. Thus, given a fragment F_i and a dataflow strategy DS_i, a mapping function $DSF(F_i, DS_i)$ assigns a dataflow strategy to a fragment of the workflow. In this context, given a set of activations $X'=\{x_1, \ldots, x_m\}$ associated to a fragment F_i, a dataflow strategy (DS_i) imposes a partial activation order among activations of X'

E. Ogasawara, J. Dias, D. Oliveira, F. Porto, P. Valduriez, and M. Mattoso, "An Algebraic Approach for Data-Centric Scientific Workflows," *Proc. of VLDB Endowment*, vol. 4, no. 12, pp. 1328–1339, 2011.

Chiron - Algebraic Operators

- Program invocation
  - Map (1:1)
  - SplitMap (1: n)
  - Reduce (n : 1)
  - Filter (1: 0-1)
- Relational algebra expressions
  - SRQuery → Single Relation Query
  - MRQuery → Multiple Relation Query