Ciência, Tecnologia e Inovação
Ministério da Ciência, Tecnologia e Inovação

GOVERNO FEDERAL BRASIL PAÍS RICO É PAÍS SEM POBREZA

Laboratório Nacional de Computação Científica

DEXL LAB
EXTREME DATA LAB

Workshop LSST Brazil/USA
eScience Infrastructure to support Astronomy Data Management

Fabio Porto
MCTI – LNCC – DEXL Lab
fporto@lncc.br
http://dexl.lncc.br

Outline

- Introduction
- SINAPAD
- DM initiatives @ LNCC - LineA
- Final Remarks.

Laboratório Nacional de Computação Científica
DEXL LAB
EXTREME DATA LAB

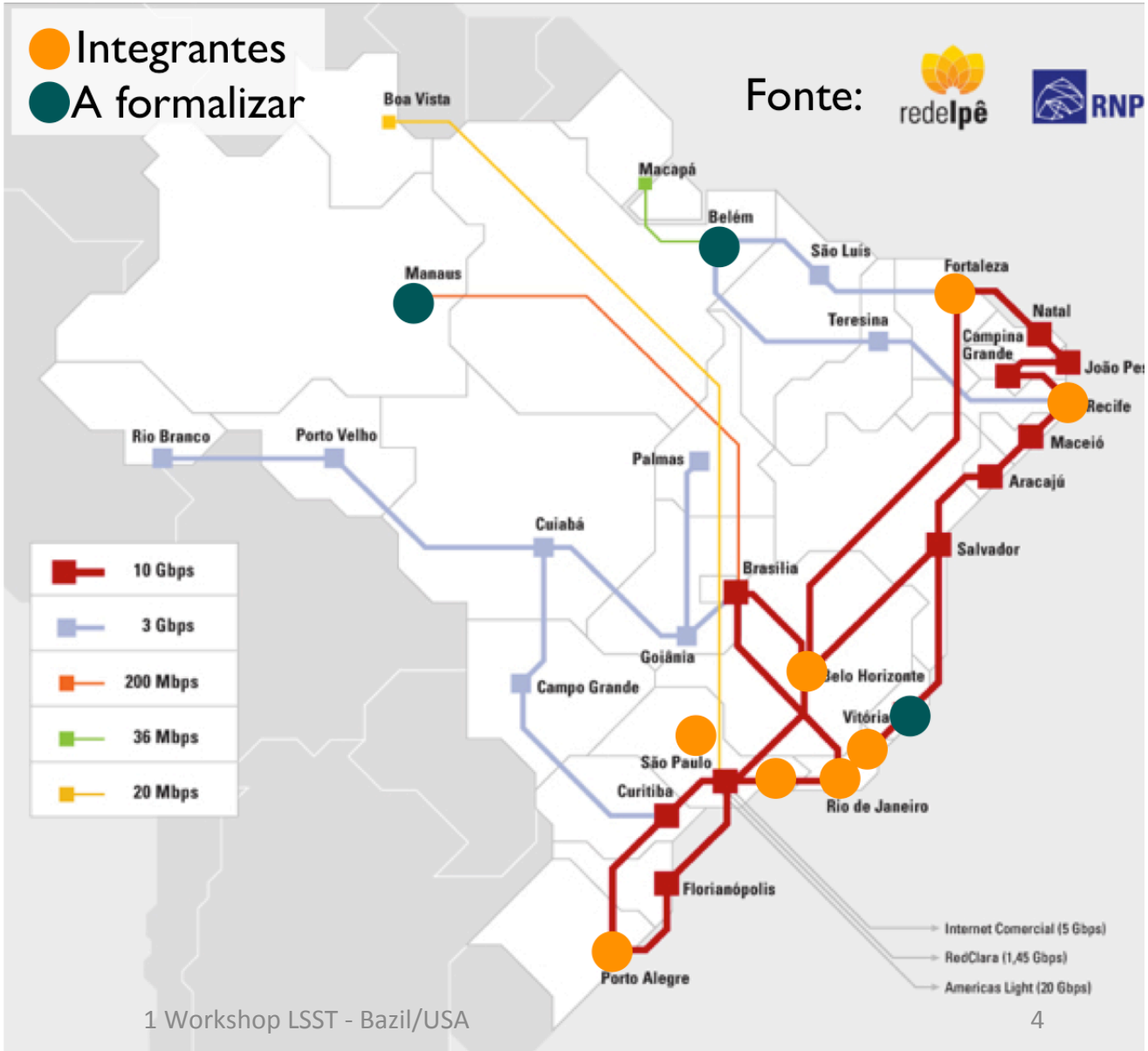
Introduction

- National Laboratory of Scientific Computing
  - Petropolis, "Imperial Town" , 70 km from Rio
  - Institute of the Ministry of Science, Technology and Innovation
  - PhD Program on Computational Modeling
  - One of the three institutes members of LineA
  - Participation on LineA:
    - SINAPAD
    - DM Research and Development
    - DEXL – Data Extreme Lab

Laboratório Nacional de Computação Científica
DEXL LAB EXTREME DATA LAB

SINAPAD National System of High Processing Computing

- Coordinated by LNCC
- Composed of CENAPADs:
  - Universities
  - Research centers
  - Different architectures:
    - Shared-disks
    - Shared-memory
    - GPUs



sinapad.lncc.br

CENAPADS

CENAPAD	Equipamento (#cores)	Desempenho teórico (TFlops)	Memória RAM (TBytes)	Capacidade de disco (TBytes)	Instalação / Última atualização
COPPE	Sun Fire V20Z (4 CPUs) * SGI Altix 450 (64 CPU Cores) SGI Altix ICE 8200/8400LX/UV100 (672 CPU Cores + 896 GPU Cores)	0,017 0,384 8,33	0,004 0,128 1,3	0,072 3 32	2005 2006 / 2007 2008 / 2011
CPTEC	Sun Fire V20Z (4 CPUs) *	0,017	0,004	0,072	2005
LNCC	Sun Fire V20Z (4 CPUs) * Sun Blade X6250 (640 CPU Cores) SGI Altix XE340 (240 CPU Cores)	0,017 5 2,4	0,004 1,2 0,72	0,072 48 6	2005 2010 2010
UFC	Sun Fire V20Z (8 CPUs) * Bull bullx (576 CPU Cores)	0,034 6,128	0,008 1,152	0,072 27	2005 2011
UFMG	Sun Fire V20Z (4 CPUs) * Bull NovaScale (878 CPU Cores)	0,017 9	0,004 1,7	0,072 45	2005 2010
UFPE	Bull NovaScale (576 CPU Cores)	6	1,1	11	2010
UFRGS	Sun Fire V20Z (4 CPUs) * Sun Fire X2200M2/X4240/X4600M2 (516 CPU Cores + 1920 GPU Cores) SGI Altix ICE 8400LX (768 CPU cores)	0,017 12,94 7,07	0,004 1,2 2	0,072 60 16	2005 2008 2011
UNICAMP	Sun Fire V20Z (4 CPUs) * SGI Altix 1350 (70 CPUs) SGI Altix 450/ICE 8400LX (560 CPU cores) IBM P750 (1280 CPU cores + 5376 GPU cores)	0,017 0,42 5,96 43,18	0,004 0,274 1,496 5,120	0,072 7 36 224	2005 2005 / 2007 2008 / 2011 2011
SINAPAD (storage)	NetApp - NFS distribuído – FAS3160 (1 unidade central) – FAS3140 (1 unidade de backup) – FAS2050C (8 unidades remotas)			720 302 832	em instalação em instalação 2011
Total	6840 CPU Cores + 8192 GPU Cores	106,684	17,338	2.342	

6840 CPU Cores + 8192 GPU Cores
~106.6 TFlops / ~17.3 TBytes RAM / ~ 2.3 PBytes Storage

DM @ LNCC-LINEA

Motivation

- Large Surveys are placing huge challenges in Data management;
  - Complex data: spatio-temporal, high number of objects matadata (catalog)
  - Queries over the whole sky
  - Catalog analysis using Complex scientific workflows
  - Hard to preview access patterns
    - What analyses would be needed?
- LineA is the Data center for DES!!

Current state of DM at LineA

- Focus on:
  - Data transfer and ingestion procedures
    - Support to fault tolerance on data transfer
  - the Catalog metadata management
    - Vertical based partitioning of the sky
      - Criteria based on the number of nodes
      - Use ra,dec to partition the sky
      - Support of Q3C indexing and UDFs
    - Materialized views
      - to restrict data to pipelines
      - To partition data based on sematic criteria
        - » Ra,dec
        - » Photometry, astrometry, spectrometry
    - Adoption of ORDBMS technology
      - PostgreSQL – DES
      - MSSQLServer – SDSS
  - Catalog data input to scientific workflows (pipelines)
  - Data releases form independent views of the sky

ON GOING WORK @ LNCC

TLPP Processing

- Transfer Data
  - From telescope to data storage sites
- Load
  - Data Ingestion Procedures
  - Data Management
  - Data Replication
  - Data Model
  - Query Processing
- Process
  - Scientific Workflows
  - Data locality
  - Provenance
  - Store of workflow results in DB
- Publish
  - Portal Ad-Hoc queries
  - Download data
  - Linked-Data

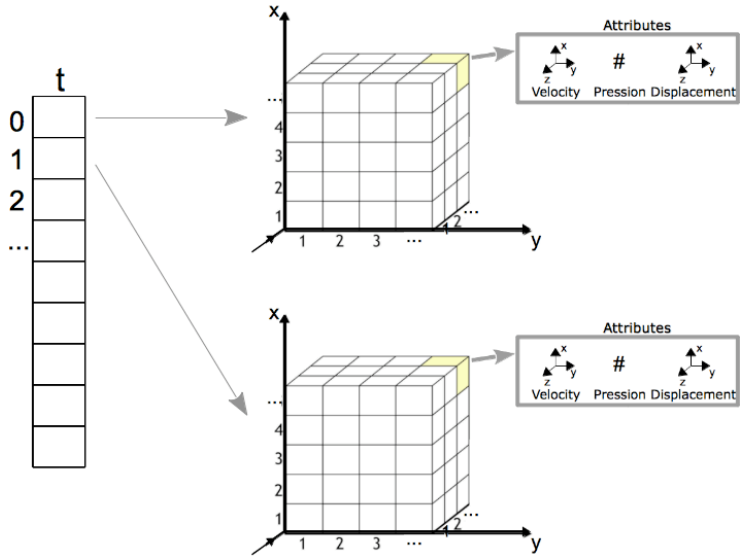
Data Model

- Investigating data models more adequate to store and analyze scientific data
  - Spatio-temporal data
  - Timeseries
- Implications on
  - the easiness of expressing queries
  - Efficiency of processing complex queries

Laboratório Nacional de Computação Científica
DEXL LAB EXTREME DATA LAB

Storage and Access of Multidimensional data

- Use of SciDB multi-array model
  - Support for time-space
  - Multi-scale
  - Array data
    - Time-series values
    - Pre-calculated colors
- Simulation data analysis
  - Ramon Costa, PhD LNCC
  - Costa, R. et al 2012




```
CREATE ARRAY Geometry3D
    <velocity: point3D, pression: double,
displacement: point3D>
    [t=0..500,500,0,
      x=1..7000,1000,0, y=1..7000,1000,0,
z=1..36000,1000,0]

store( slice(Geometry3D, t, 0),  G1);
store( slice(Geometry3D, t, 0),  G2);
```

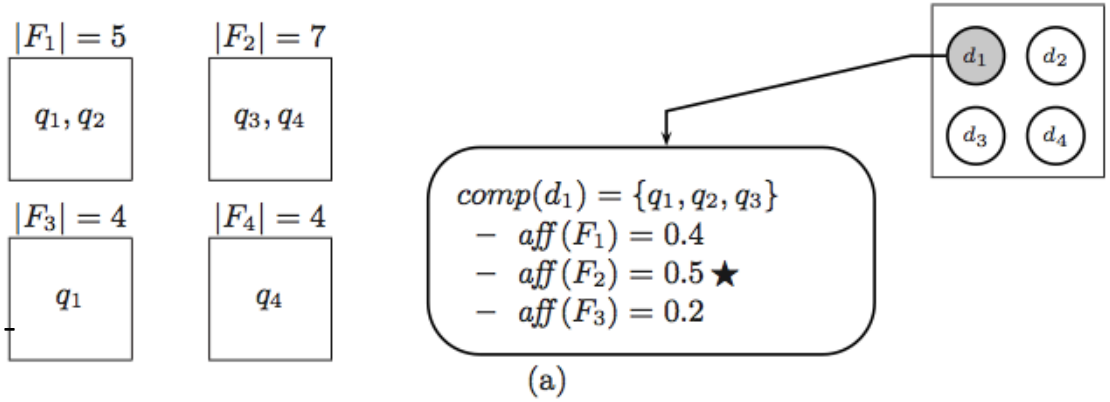
Data partitioning

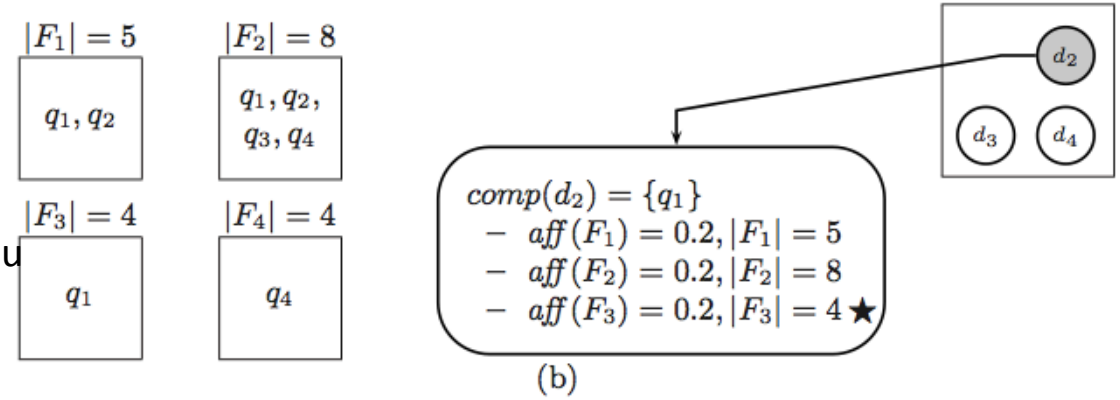
- Evident space clustering criteria
- What about the remaining object/source attributes that could provide more fine grained partitioning?
- What about updating with new data and new queries?

Dynamic Data Partitioning

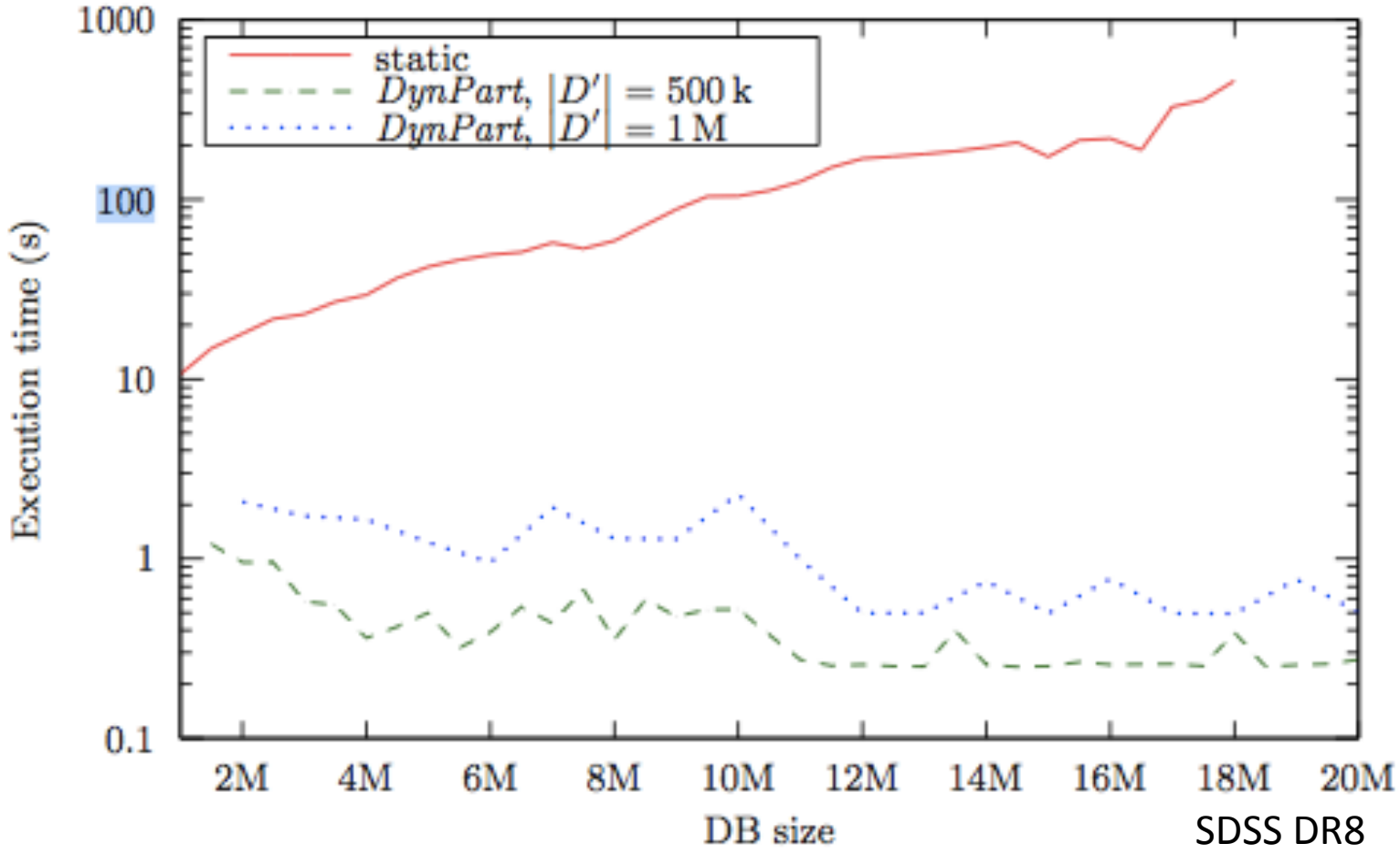
- Devise a strategy for data partitioning based on:
  - Query Workload (SDSS + DES)
  - Extensible for high number of partitions
  - Workload graph based strategy (Schism)
    - Curino, C., Jones, E., Zhang, Y., Madden, S.: Schism: a workload-driven approach to database replication and partitioning. VLDB2010
  - Dynamic strategy (new queries and new data)
    - Collaboration with INRIA, Univ. Montpellier
    - PhD Student Miguel Liroz-Gistau
  - Liroz-Gistau et al 2012

$$W = \{q_1, q_2, q_3, q_4\},\ f(q_1) = 0.2, f(q_2) = 0.2, f(q_3) = 0.5, f(q_4) = 0.1$$





Efficiency of Dynamic Partitioning

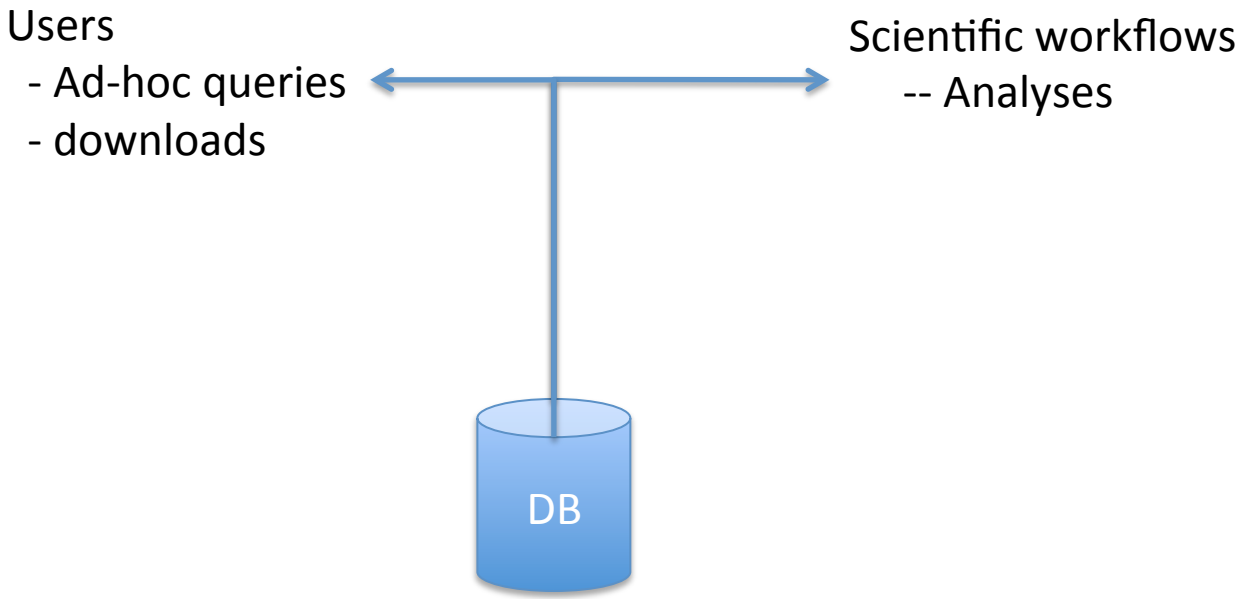


SDSS DR8
PaToH HG partitioner

TLPP Processing

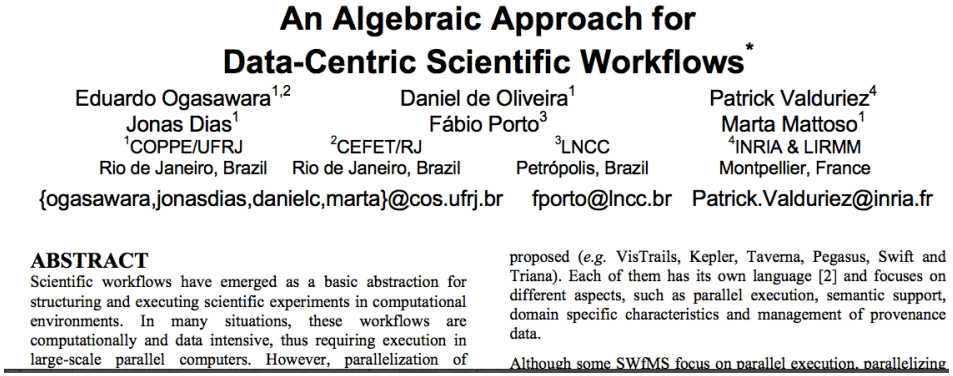
- Transfer Data
  - From telescope to data storage sites
- Load
  - Data Ingestion Procedures
  - Data Management
  - Data Replication
  - Data Model
  - Query Processing
- Process
  - Scientific Workflows
  - Data locality
  - Provenance
  - Store of workflow results in DB
- Publish
  - Portal Ad-Hoc queries
  - Linked-Data

Processing Scientific workflows on Database data



Parallelization of Scientific Workflows

- I- Algebraic Approach for SW
  - Inspired by query processing techniques
  - Data-oriented model
  - Activities classified as:
    - Map, Reduce, Split, Merge,etc..
  - Automatic Optimization of SW
  - Cooperation with UFRJ/COPPE, INRIA-Montpellier
  - Two systems:
    - QEF (Query Engine Framework – LNCC)
    - Chiron (SW with Provenance – CEFET-COPPE/UFRJ



An Algebraic Approach for Data-Centric Scientific Workflows*

Eduardo Ogasawara[1,2] Daniel de Oliveira[1] Patrick Valduriez[4]
Jonas Dias[1] Fábio Porto[3] Marta Mattoso[1]
[1]COPPE/UFRJ [2]CEFET/RJ [3]LNCC [4]INRIA & LIRMM
Rio de Janeiro, Brazil Rio de Janeiro, Brazil Petrópolis, Brazil Montpellier, France
{ogasawara,jonasdias,danielc,marta}@cos.ufrj.br fporto@lncc.br Patrick.Valduriez@inria.fr

ABSTRACT
Scientific workflows have emerged as a basic abstraction for structuring and executing scientific experiments in computational environments. In many situations, these workflows are computationally and data intensive, thus requiring execution in large-scale parallel computers. However, parallelization of

proposed (*e.g.* VisTrails, Kepler, Taverna, Pegasus, Swift and Triana). Each of them has its own language [2] and focuses on different aspects, such as parallel execution, semantic support, domain specific characteristics and management of provenance data.

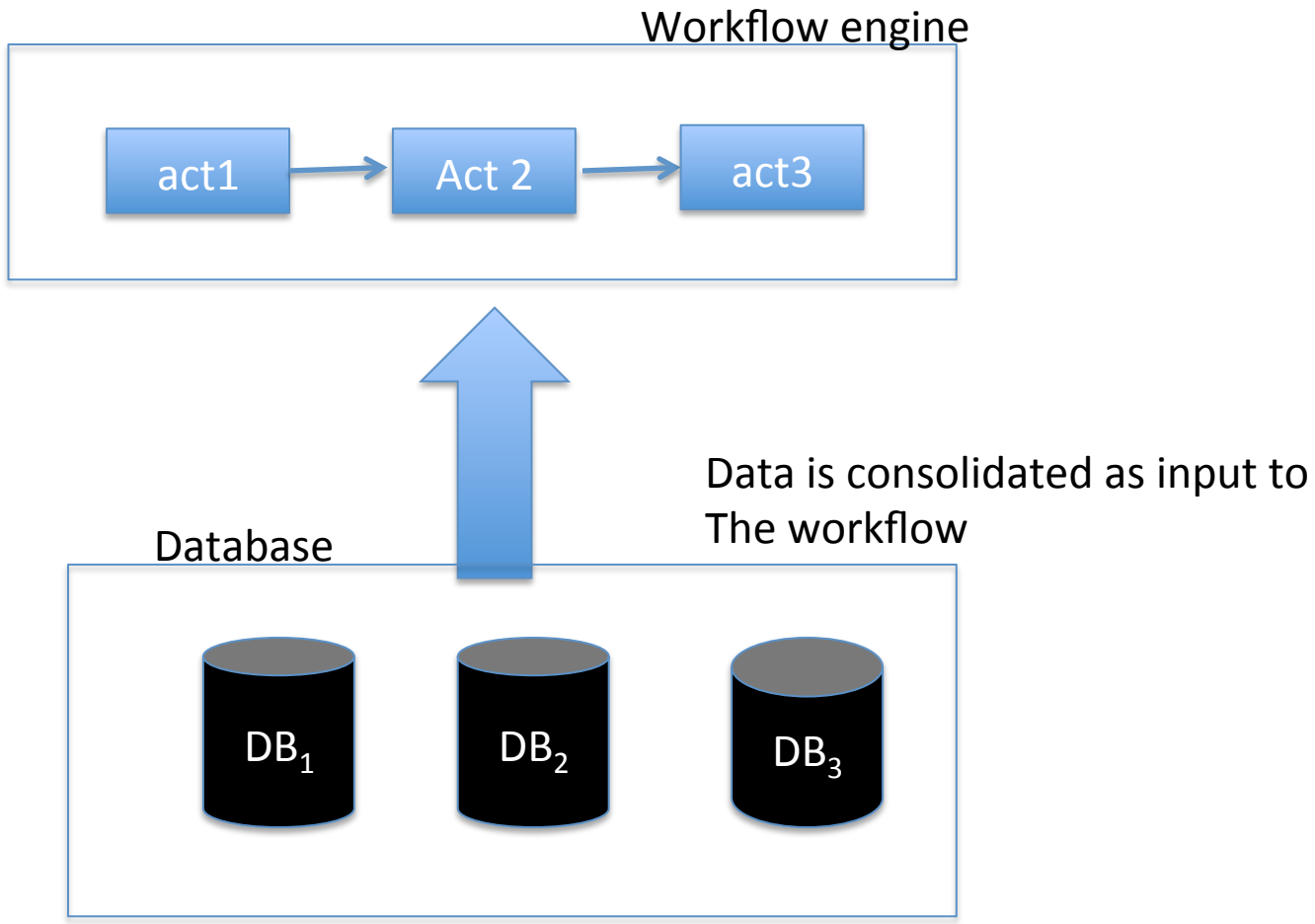
Although some SWfMS focus on parallel execution, parallelizing

Laboratório Nacional de Computação Científica
DEXL LAB EXTREME DATA LAB

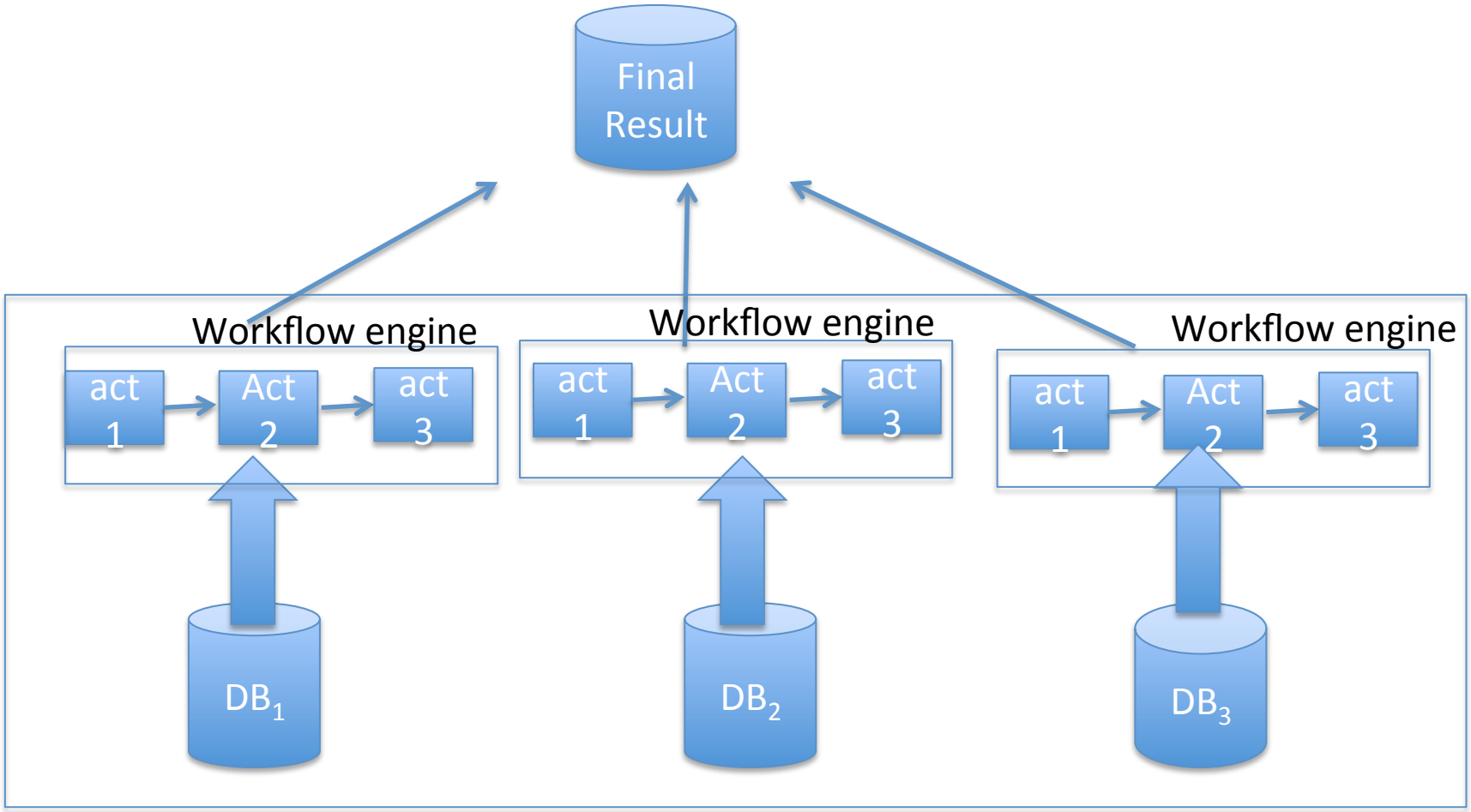
Parallelization of Scientific Workflows

- II – Parallelization of SW with Hadoop and Column-Oriented Database
  - Integration of Hadoop with SW engine (QEF)
    - An activity of the SW becomes a Map/Reduce task
    - The SW engine becomes a Map task
  - Distributed column-store
    - Queries refer to a small number of catalog attributes
    - Column store reduce the number of access to DB pages
    - Use the results on data partitioning
    - MonetDB
  - 1 Master student + 1 Engineer (LNCC)

Traditional WF – Database architecture



Integrated architecture



TLPP Processing

- Transfer Data
  - From telescope to data storage sites
- Load
  - Data Ingestion Procedures
  - Data Management
  - Data Replication
  - Data Model
  - Query Processing
- Process
  - Scientific Workflows
  - Data locality
  - Provenance
  - Store of workflow results in DB
- Publish
  - Portal Ad-Hoc queries
  - Linked-Data

Publish Results

- Investigating the use of Linked-data standard for:
  - Publishing metadata and data
  - Linking data among collaboration sites
  - Searching for data enriched with analysis results
  - CNPq Universal Project with PUC-Rio, UFF
  - Integration of autonomous databases
  - Using QEF as query engine for Integration queries

Final Remarks

- Large Astronomy surveys will require new results from database research
- The LNCC-DEXL Lab together with LineA and collaborators are working on some of the challenges
- Some initiatives going on
  - Brazilian eScience Workshop July 16-19 2012
    - One day dedicated to Astronomy
  - LNCC Meeting, July 19
    - Massive Data Management Workshop
      - http://www.lncc.br/meeting2012