Fourth Brazil-France Workshop

On High Performance Computing and Scientific Data Management Driven by Highly Demanding Applications

15-18 September, Gramado, Brazil 2014

User interaction in uncertainty quantification analysis workflows

Jonas Dias, Gabriel M. Guerra, Fernando Rochinha, Alvaro L.G.A. Coutinho, Patrick Valduriez and Marta Mattoso

COPPE

Federal University of Rio de Janeiro

FR-BR HOSCAR Collaboration results

- Students interchange

  - Jonas Dias (PhD, 2013, supervised: M Mattoso & P Valduriez)
  - Vitor Silva (MSc, 2014, internship at INRIA Oct-Dec 2013)

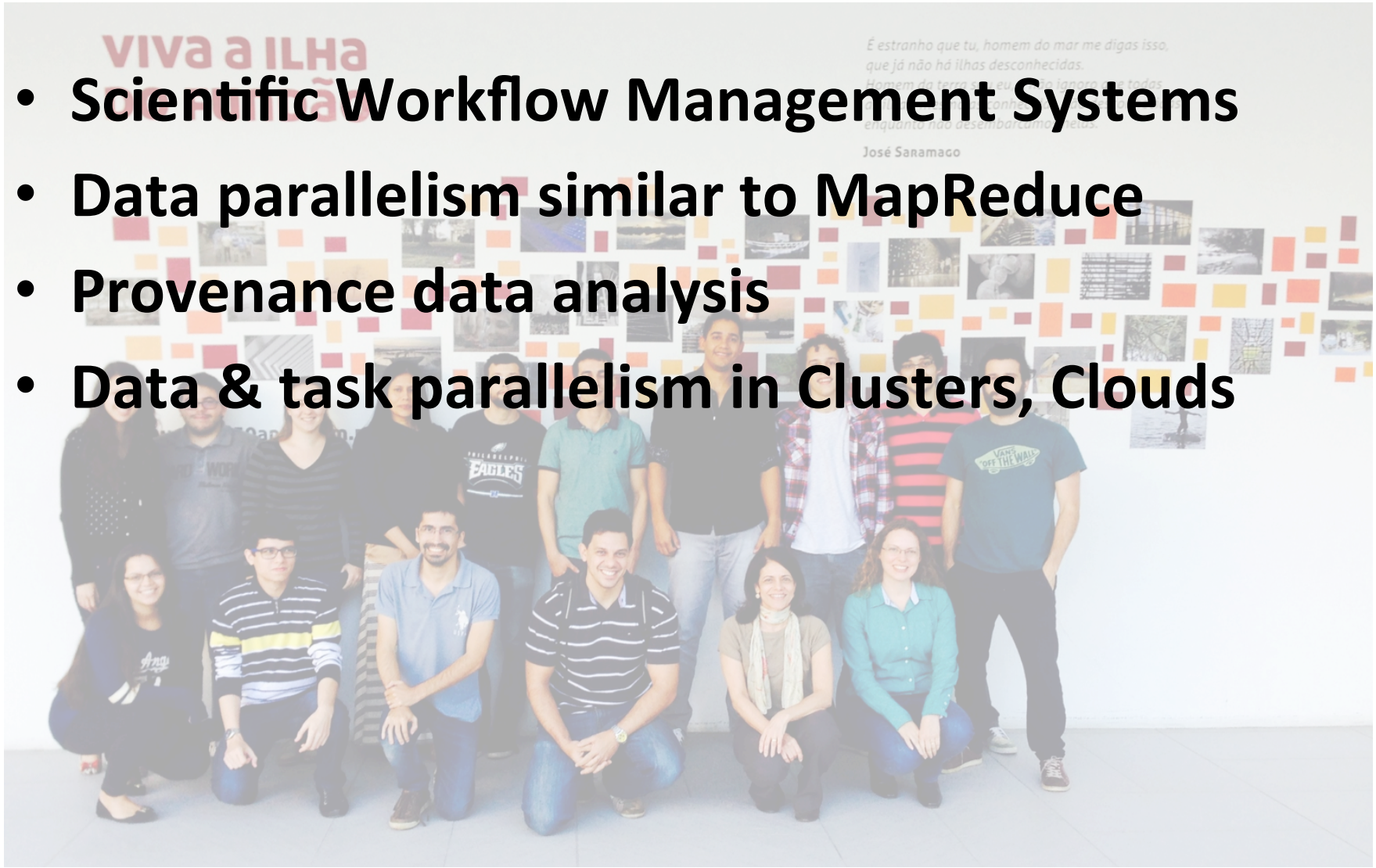
  - Ji Liu (PhD ongoing, supervision: Mattoso, Pacitti, Valduriez)

- Joint papers

- J. Liu, V. Silva, E. Pacitti, P. Valduriez, M. Mattoso, "Scientific Workflow Partitioning in Multisite Cloud". In: **3rd Workshop on Big Data Management in Clouds**, Proc. of the Europar 2014
- J. Dias, E. S. Ogasawara, D. de Oliveira, F. Porto, P. Valduriez, and M. Mattoso. Algebraic Dataflows for Big Data Analysis. **IEEE Bigdata Conference** 2013
- E. S. Ogasawara, J. Dias, V. Silva, F. S. Chirigati, D. de Oliveira, F. Porto, P. Valduriez, and M. Mattoso. Chiron: a parallel engine for algebraic scientific workflows. ***Concurrency and Computation****: Practice and Experience*, 25(16):2327–2341, 2013.
- Chirigati, F S ; Sousa, V. ; Ogasawara, E. ; Oliveira, D. ; Dias, J. ; Porto, F. ; Valduriez, P. ; Mattoso, Marta . Evaluating Parameter Sweep Workflows in High Performance Computing. In**: Int Workshop on Scalable Workflow Enactment Engines and Technologies (SWEET'12**), 2012, Phoenix. SIGMOD.
- J. Dias, G. Guerra, F. Rochinha, A.L.G.A. Coutinho, P. Valduriez, and M. Mattoso, Data-Centric Iteration in Dynamic Workflows, **Sumitted (2nd round review) Future Generation Computer Systems**

Distributed & Parallel Data Mngmt

Distributed & Parallel Data Mngmt



- **Scientific Workflow Management Systems**
- **Data parallelism similar to MapReduce**
- **Provenance data analysis**
- **Data & task parallelism in Clusters, Clouds**

Putting the human in the loop

"In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a difficult time finding."

Exploring the inherent technical challenges in realizing the potential of Big Data.

BY H.V. JAGADISH, JOHANNES GEHRKE, ALEXANDROS LABRINIDIS, YANNIS PAPAKONSTANTINOU, JIGNESH M. PATEL, RAGHU RAMAKRISHNAN, AND CYRUS SHAHABI

Big Data and Its Technical Challenges

Putting the human in the loop

1. Heterogeneity
   - use data provenance
2. Inconsistency & incompleteness
3. Scale
   - parallel data proc; clouds
   - declarative languages
4. Timeliness
   - real-time techniques to summarize and filter data
5. Privacy and data ownership
6. The human perspective: Visualize
   - scale not just for the system but also from the perspective of *humans*.
   - human input at all stages of the analysis pipeline
   - specific parameter values to a given snapshot of an evolving dataset

"In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a difficult time finding."

Exploring the inherent technical challenges in realizing the potential of Big Data.

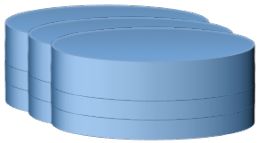
BY H.V. JAGADISH, JOHANNES GEHRKE, ALEXANDROS LABRINIDIS, YANNIS PAPAKONSTANTINOU, JIGNESH M. PATEL, RAGHU RAMAKRISHNAN, AND CYRUS SHAHABI

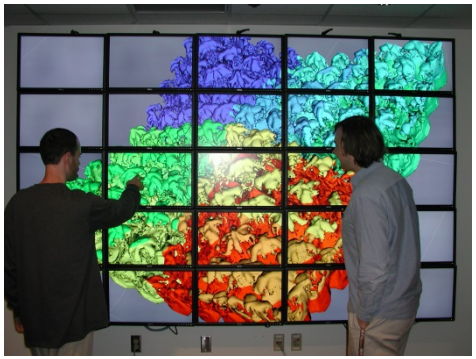
Big Data and Its Technical Challenges

JULY 2014 | VOL. 57 | NO. 7 | **COMMUNICATIONS OF THE ACM**

Tracking Data Transformations

- data transformations – *ad-hoc*
- files generated independently
- parallel processing unaware of data-flow
- analysts need to manually manage the larger life cycle of big data flow analysis





xterm

```
Jul  6 13:20 OSA_May2011/QQ_pngs/GREY.GREY,rel0.25,maf0.05,young_v_old.cov1.ld5,5000kb,P,wide,p5e-05.png
Jul  6 13:22 OSA_May2011/QQ_pngs/GREY.GREY,rel0.25,maf0.05,young_v_old.cov1.ld8,1000kb,P,p0001.png
Jul  6 13:21 OSA_May2011/QQ_pngs/GREY.GREY,rel0.25,maf0.05,young_v_old.cov1.ld8,1000kb,P,p0005.png
Jul  6 13:21 OSA_May2011/QQ_pngs/GREY.GREY,rel0.25,maf0.05,young_v_old.cov1.ld8,1000kb,P,p1e-05.png
Jul  6 13:21 OSA_May2011/QQ_pngs/GREY.GREY,rel0.25,maf0.05,young_v_old.cov1.ld8,1000kb,P,p5e-05.png
Jul  6 13:22 OSA_May2011/QQ_pngs/GREY.GREY,rel0.25,maf0.05,young_v_old.cov1.ld8,5000kb,P,wide,p0001.png
Jul  6 13:21 OSA_May2011/QQ_pngs/GREY.GREY,rel0.25,maf0.05,young_v_old.cov1.ld8,5000kb,P,wide,p0005.png
Jul  6 13:22 OSA_May2011/QQ_pngs/GREY.GREY,rel0.25,maf0.05,young_v_old.cov1.ld8,5000kb,P,wide,p1e-05.png
Jul  6 13:21 OSA_May2011/QQ_pngs/GREY.GREY,rel0.25,maf0.05,young_v_old.cov1.ld8,5000kb,P,wide,p5e-05.png
Jul  6 13:14 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,cc,cov1.ld2,2500kb,P,p0001.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,cc,cov1.ld2,2500kb,P,p0005.png
Jul  6 13:14 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,cc,cov1.ld2,2500kb,P,p1e-05.png
Jul  6 13:14 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,cc,cov1.ld2,2500kb,P,p5e-05.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,cc,cov1.ld2,5000kb,P,wide,p0001.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,cc,cov1.ld2,5000kb,P,wide,p0005.png
Jul  6 13:16 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,cc,cov1.ld2,5000kb,P,wide,p1e-05.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,cc,cov1.ld2,5000kb,P,wide,p5e-05.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,cc,cov1.ld5,2500kb,P,p0001.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,cc,cov1.ld5,2500kb,P,p0005.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,cc,cov1.ld5,2500kb,P,p1e-05.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,cc,cov1.ld5,2500kb,P,p5e-05.png
Jul  6 13:16 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,cc,cov1.ld5,5000kb,P,wide,p0001.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,cc,cov1.ld5,5000kb,P,wide,p0005.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,cc,cov1.ld5,5000kb,P,wide,p1e-05.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,cc,cov1.ld5,5000kb,P,wide,p5e-05.png
Jul  6 13:16 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,cc,cov1.ld8,2500kb,P,p0001.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,cc,cov1.ld8,2500kb,P,p0005.png
Jul  6 13:16 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,cc,cov1.ld8,2500kb,P,p1e-05.png
Jul  6 13:17 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,cc,cov1.ld8,2500kb,P,p5e-05.png
Jul  6 13:16 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,cc,cov1.ld8,5000kb,P,wide,p0001.png
Jul  6 13:16 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,cc,cov1.ld8,5000kb,P,wide,p0005.png
Jul  6 13:16 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,cc,cov1.ld8,5000kb,P,wide,p1e-05.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,cc,cov1.ld8,5000kb,P,wide,p5e-05.png
Jul  6 13:16 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,young_v_old.cov1.ld2,2500kb,P,p0001.png
Jul  6 13:15 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,young_v_old.cov1.ld2,2500kb,P,p0005.png
Jul  6 13:16 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,young_v_old.cov1.ld2,2500kb,P,p5e-05.png
Jul  6 13:16 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,young_v_old.cov1.ld2,5000kb,P,wide,p0001.png
Jul  6 13:16 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,young_v_old.cov1.ld2,5000kb,P,wide,p0005.png
Jul  6 13:16 OSA_May2011/QQ_pngs/IWH.IWH,rel0.25,maf0.05,young_v_old.cov1.ld2,5000kb,P,wide,p5e-05.png
```

BLOG@CACM

Data Science Workflow: Overview and Challenges, by Philip Guo

User interaction in uncertainty quantification analysis workflows

Interdisciplinary work (CS, CFD, UQ)

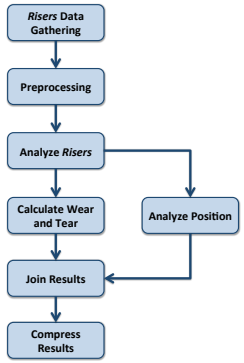
CS

- **Vitor Silva**
- **Jonas Dias**
- **Marta Mattoso**
- **Patrick Valduriez (INRIA)**

CFD, UQ

- **Renato Elias**
- **Gabriel Guerra**
- **Fernando Rochinha**
- **Alvaro Coutinho**

Scientific Workflows



- Powerful paradigm for formalizing and automating complex and data intensive scientific processes

- Focused on large scientific **Data-flows** unlike Business Workflows

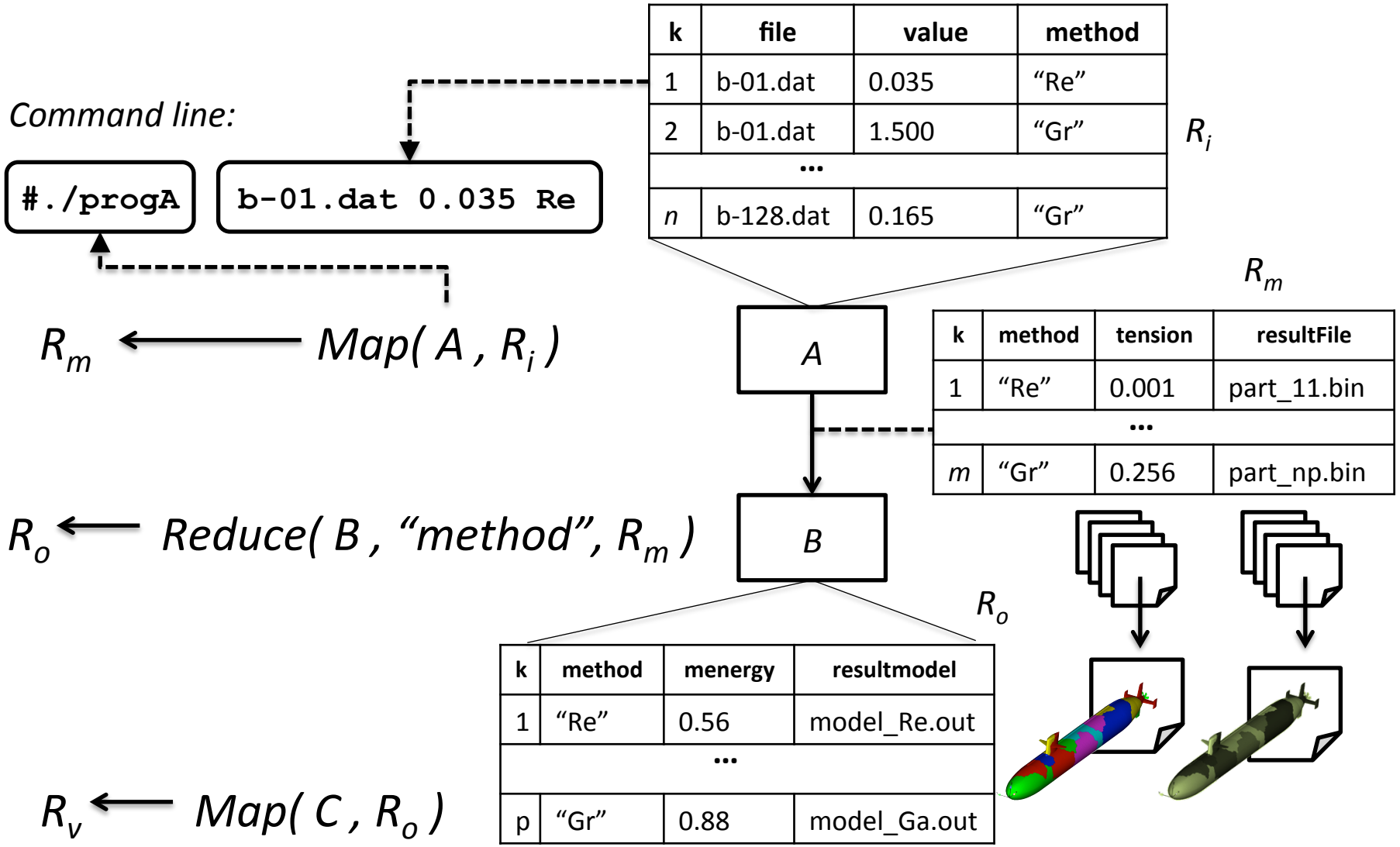
SWfMS & Provenance

- Scientific Workflow Management Systems (SWfMS)
- Efficient execution of scientific workflows
- Tracing the execution through provenance
- **Provenance** data (W3C PROV working group)
  - to enable scientific discovery
  - reproducibility,
  - result interpretation, and
  - problem diagnosis in scientific experiments

Enabling technologies

- Scientific Workflow Management Systems
- Data parallelism similar to MapReduce
- Provenance data analysis
- Chiron's Dataflow Algebraic Approach
  - non intrusive wrt parallel numerical solution
  - online data analysis
  - convergence tracking
  - visualization of partial results
  - dynamic interference on loop parameters

Dataflow Algebraic Workflow Engine

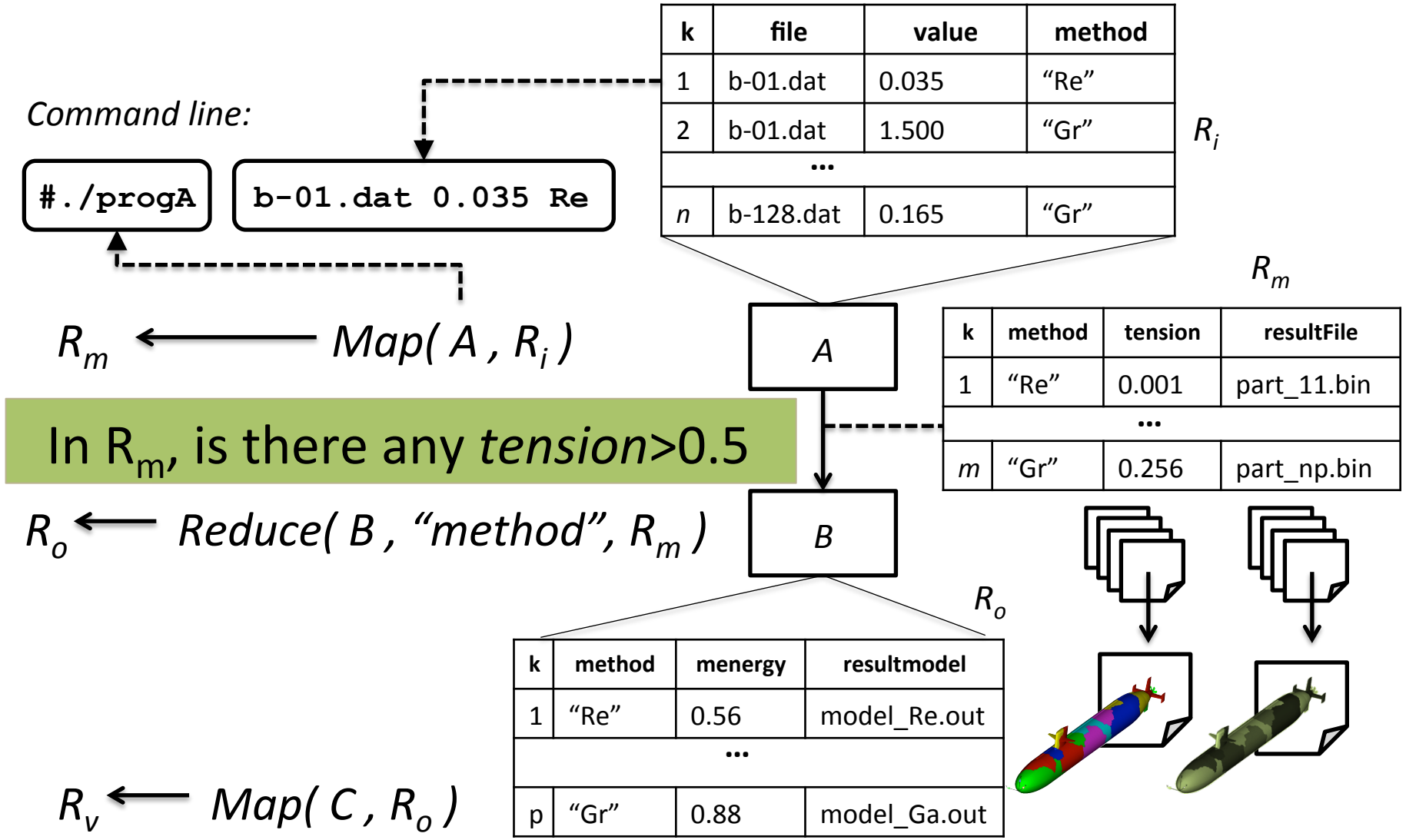


Algebraic workflow execution engine is driven by provenance DB

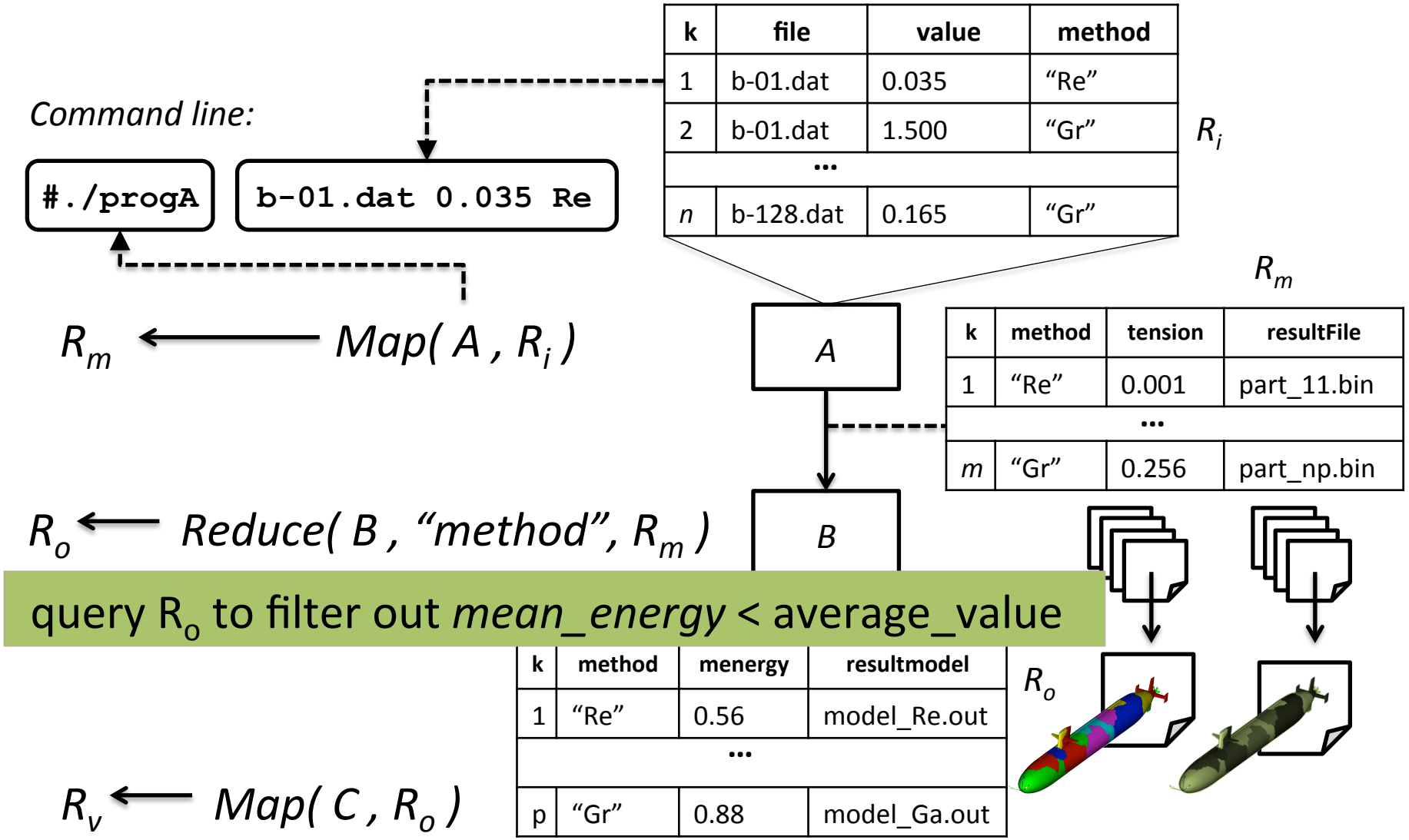
- Algebraic **workflow definition** is stored in relations inside the provenance database (ProvDB)
- Engine reads wf definition and develops an **optimized** execution plan also stored at the ProvDB
- Data from ProvDB tuples are mapped to tasks as a **Map & Reduce** data parallel execution
- Workflow engine is aware of the complete **dataflow**
- As the workflow is executed, ProvDB is augmented with **runtime information**
- ProvDB becomes an important **statistics** catalog that can be queried and analyzed

Great potential for optimizations in parallel execution

As the workflow executes … user steering (HIL)



As the workflow executes … user steering (HIL)

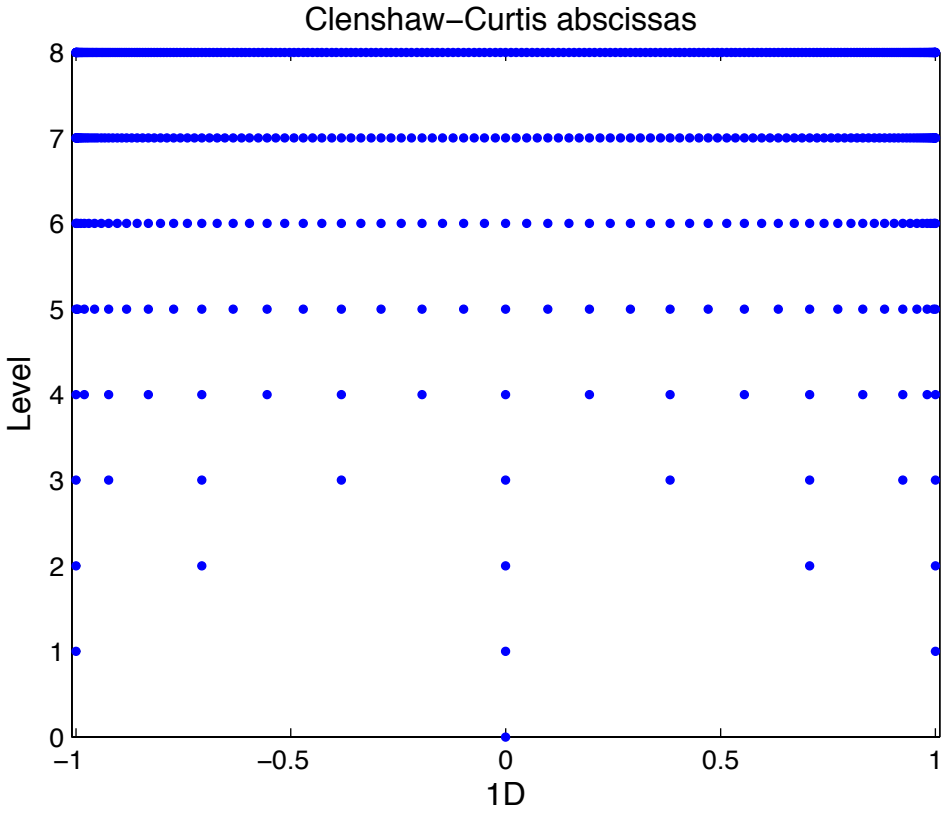


Uncertainty Quantification (UQ)

- Measures confidence in numerical simulation
- Explores a model using distributed input in a stochastic way
  - Combines average result and standard deviation
- Exploration size is associated to a pre-defined precision
  - Stochastic collocation method
  - Interpolation levels

Guerra et al. (2012) Uncertainty Quantification in Computational Predictive Models for Fluid Dynamics Using Workflow Management Engine. *International Journal for Uncertainty Quantification*, 2(1):53–71

Uncertainty Quantification using Adaptive Sparse Grid Collocation

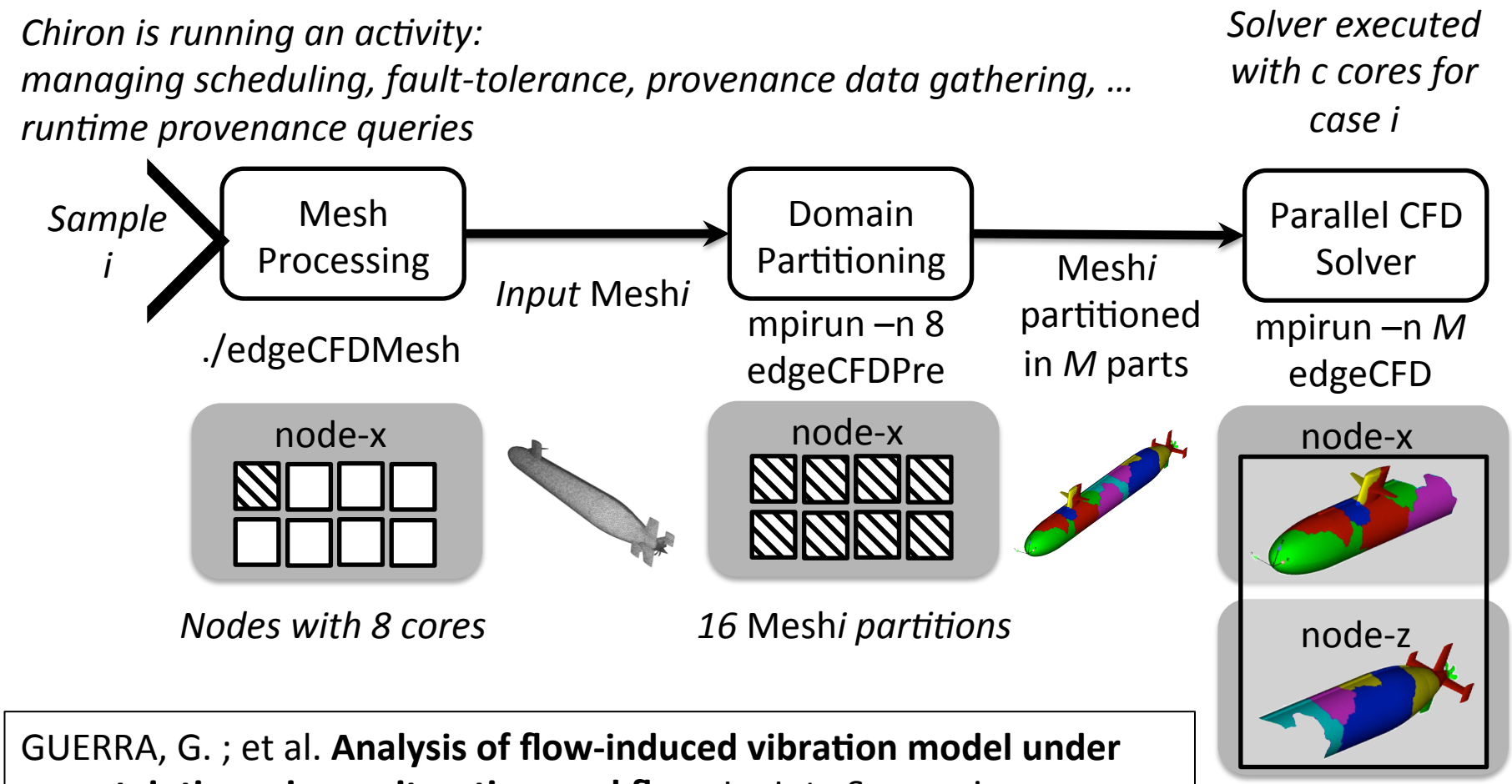


1D sparse grid for increasing interpolation level

- MC reference solution with 1,000 samples, 126hs in 32 cores
- ASGC with 8 approximation levels, 257 support nodes, 39hs (32 cores)

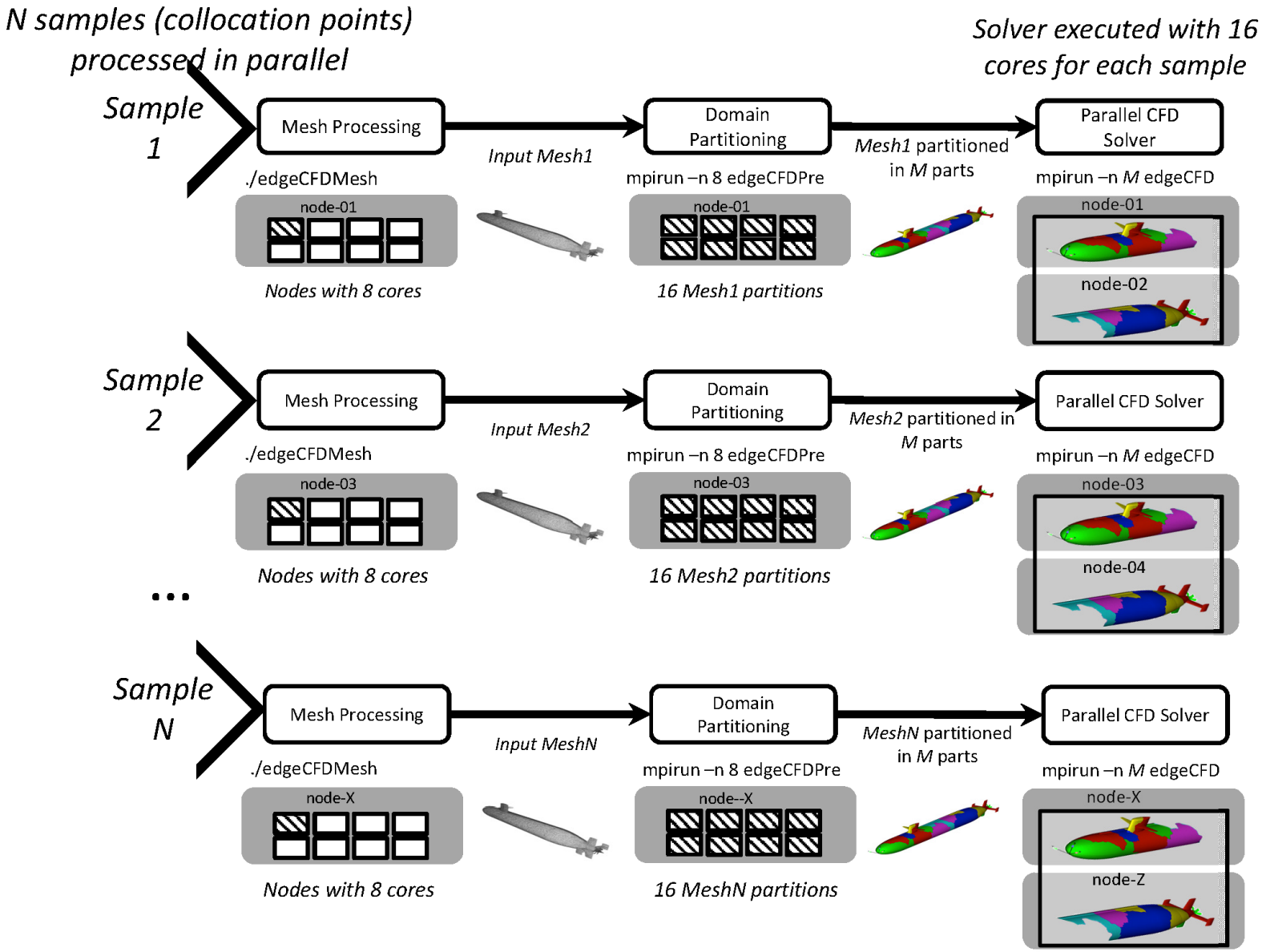
Guerra et al. (2012) Uncertainty Quantification in Computational Predictive Models for Fluid Dynamics Using Workflow Management Engine. *International Journal for Uncertainty Quantification*, 2(1):53–71

Turbulence UQ analysis



GUERRA, G. ; et al. **Analysis of flow-induced vibration model under uncertainties using an iterative workflow**. In: Int. Symposium on Uncertainty Quantification and Stochastic Modeling 2012

Two-level Parallel Strategy

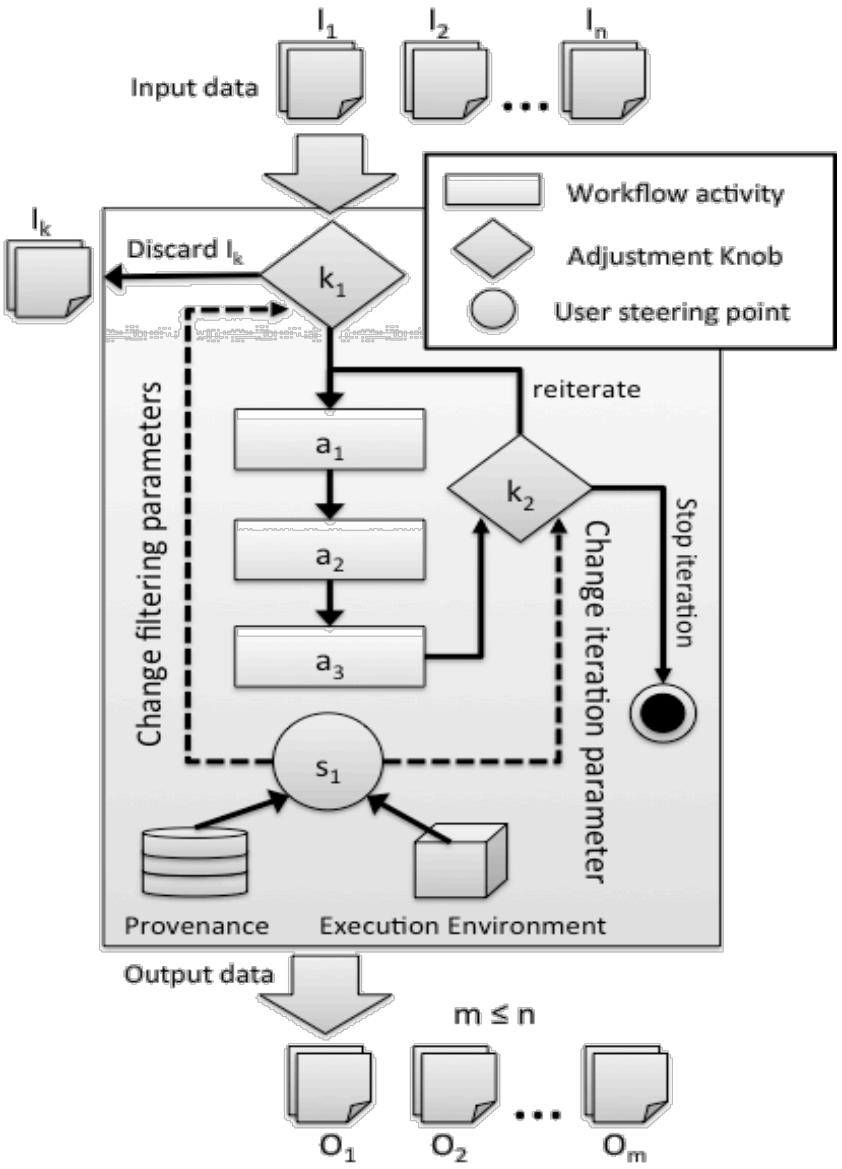


Adjusting the interpolation levels

- Execution restarts over and over
- UQ analyst may loose track of what has already been explored and how the UQ workflow evolved
- The user should be able to analyze partial results during execution
  - to dynamically interfere in the next steps of the workflow
  - instead of interrupting and resubmitting the workflow

Dias, J., et al. (2011) **Supporting Dynamic Parameter Sweep in Adaptive and User-Steered Workflow**. WORKS Workshop at IEEE Supercomputing

Dynamic Workflow example



- User steering points
  - Select provenance data
  - Trigger adjustments
  - Similar to checkpoints in Taverna
- Adjustable knobs
  - Store adjustable parameter
  - Change iteration
  - May affect the dataflow

Workflow execution
Off-line (black-box) X On-line (steering)

- Only after the whole workflow execution :
  - Check on data derivation & results
  - Change # interpolation levels
- Interrupt the execution

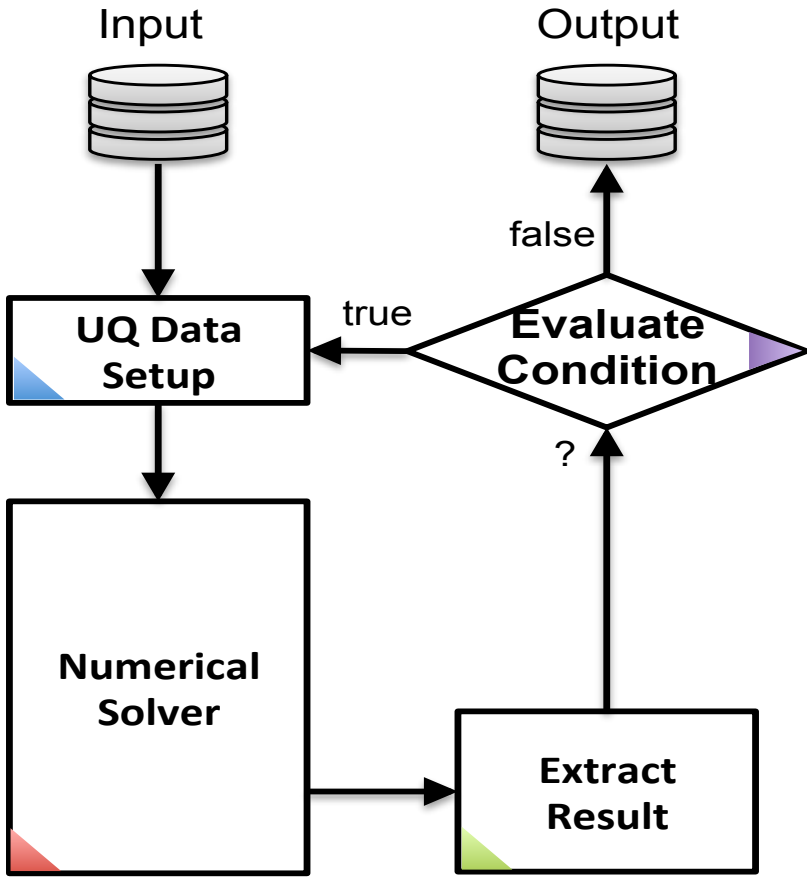
"Off-line"

During workflow execution

- Partial results & provenance are analyzed
- Snapshots of current simulation results to refine the model (iterations) during runtime
- **Fine tuning of parameters**
- **Interfere on loop specification**

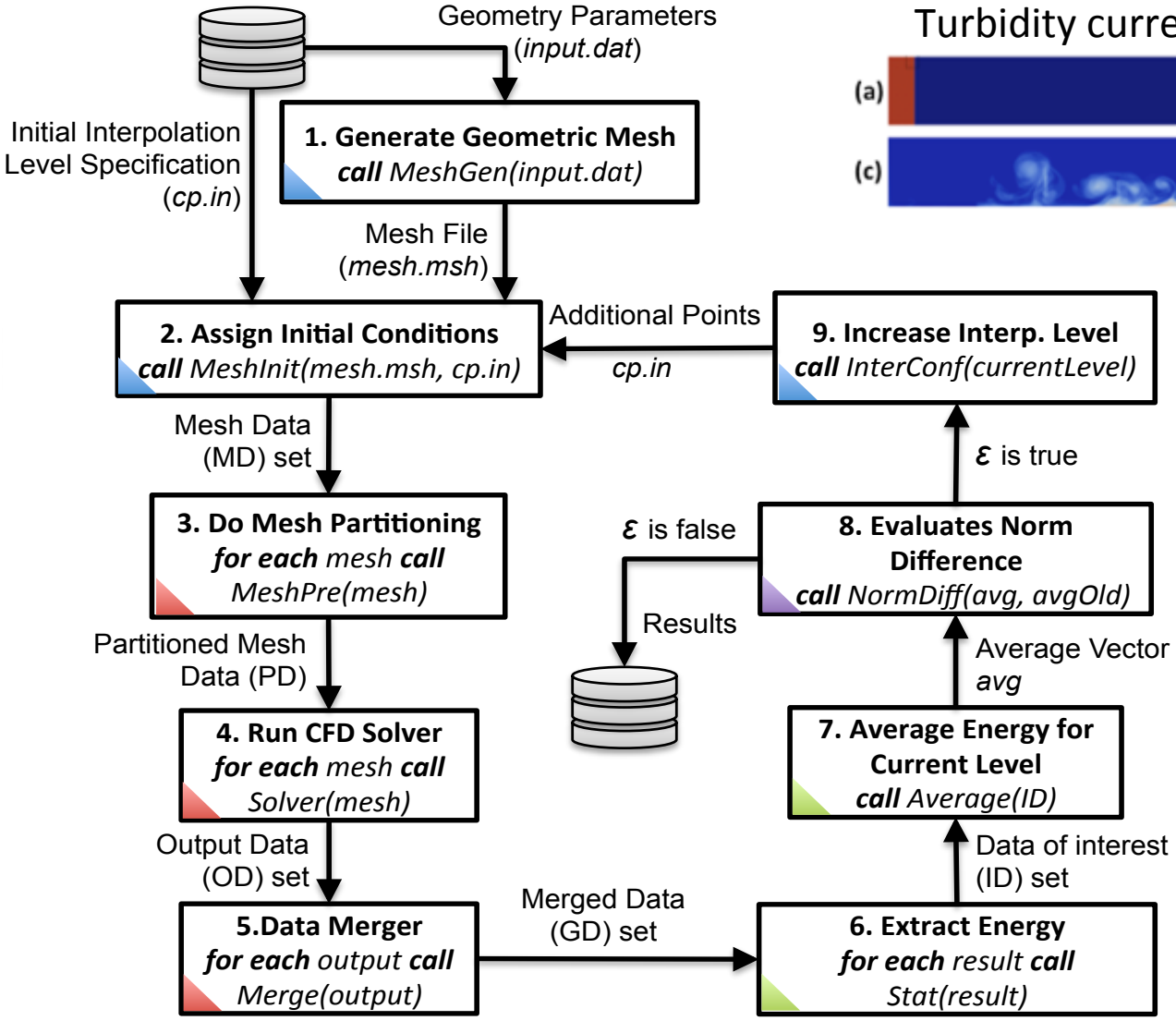
"On-line"

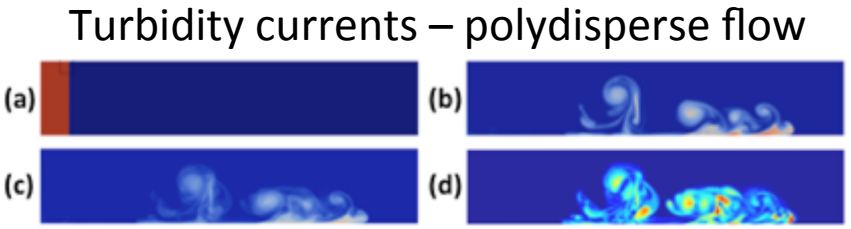
Executes for predefined levels with predefined condition



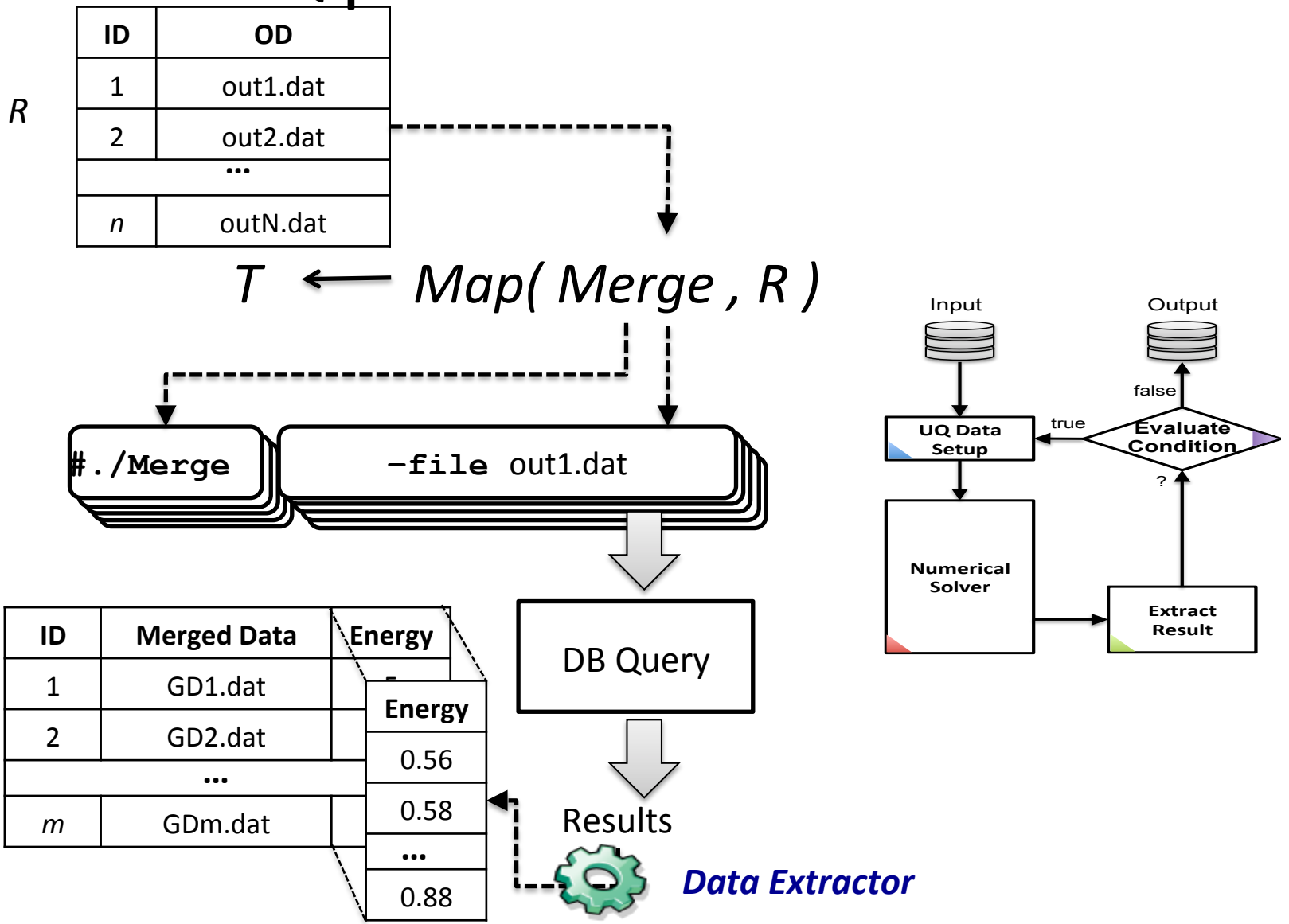
criterion is the difference between the vector norms and should be below a given **threshold**, initially defined as 0.001

UQ Executes for all predefined levels





Tuple generation of Activity 5
UQ parallel execution



Tuple generation of Activity 5
UQ parallel execution

R

ID	OD
1	out1.dat
2	out2.dat
...	
n	outN.dat

T ← *Map(Merge , R)*

`#./Merge` `-file` out1.dat

DB Query

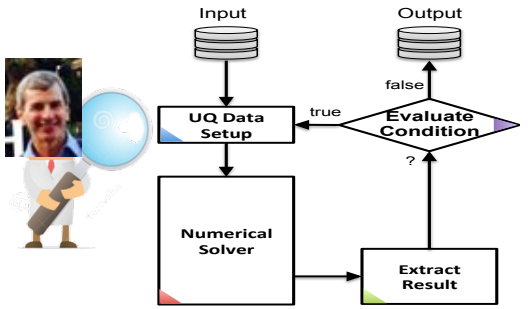
Results

T

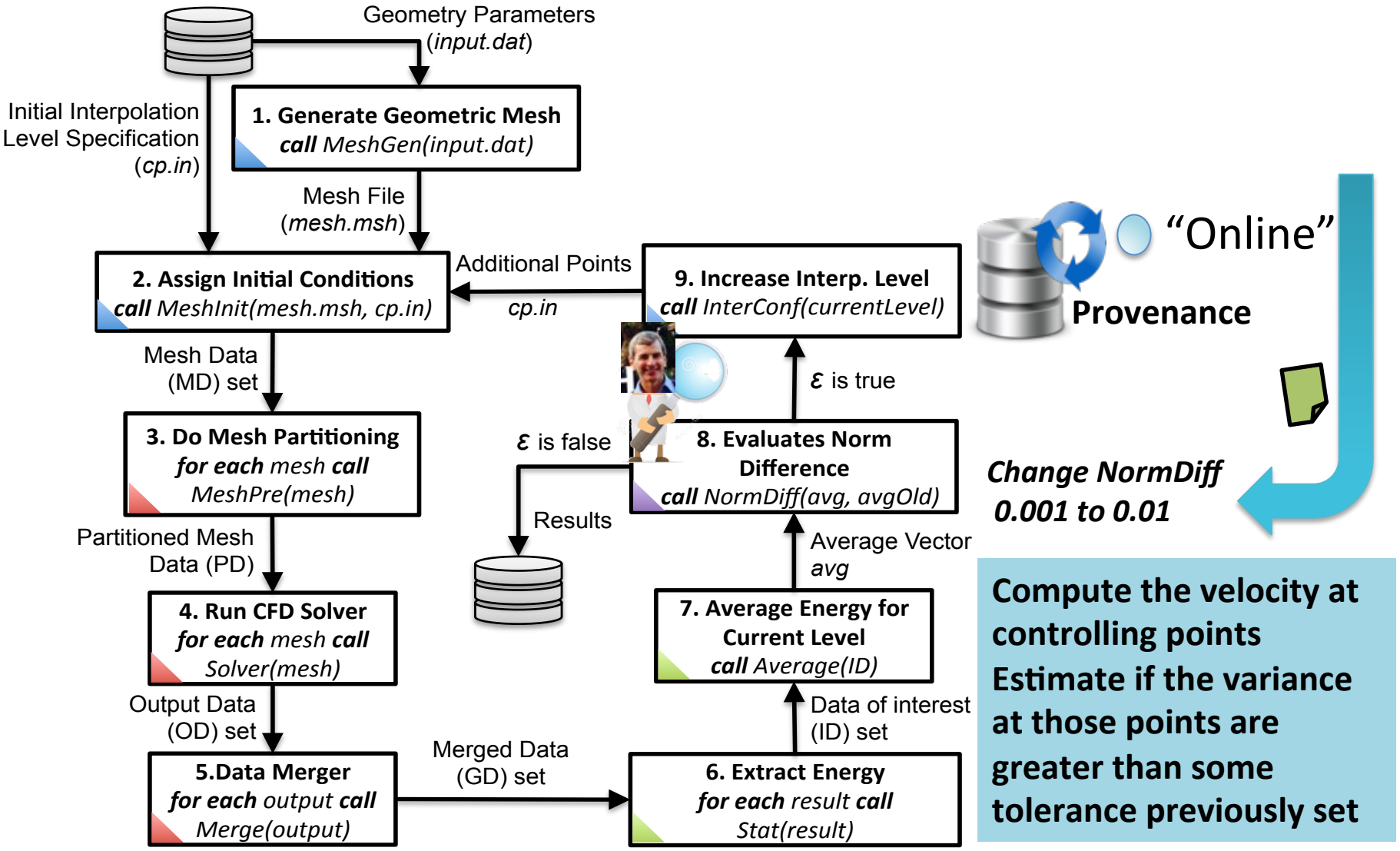
ID	Merged Data	Energy
1	GD1.dat	
2	GD2.dat	
...		
m	GDm.dat	

Energy
0.56
0.58
...
0.88

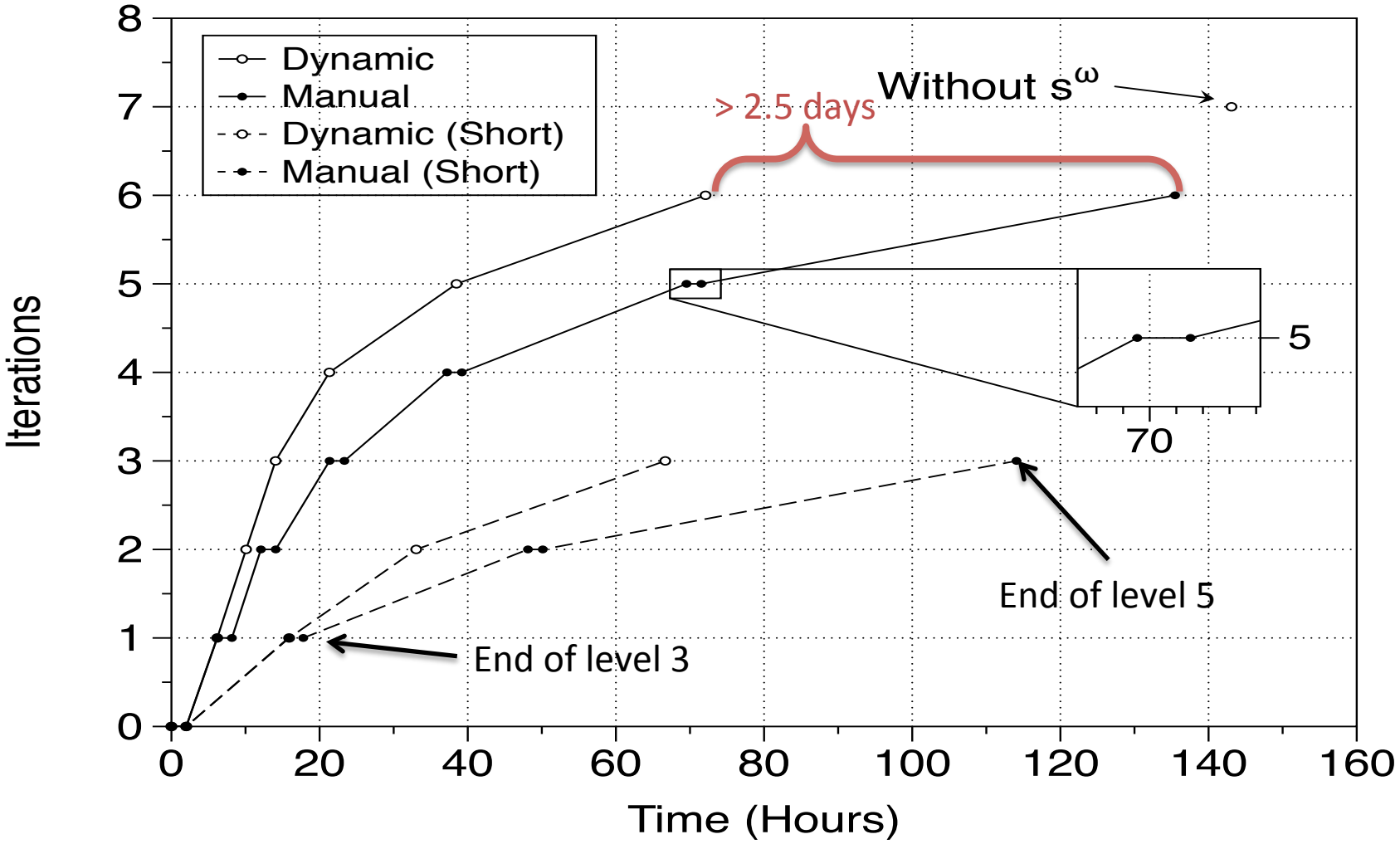
Data Extractor



UQ With User Steering & Intervention

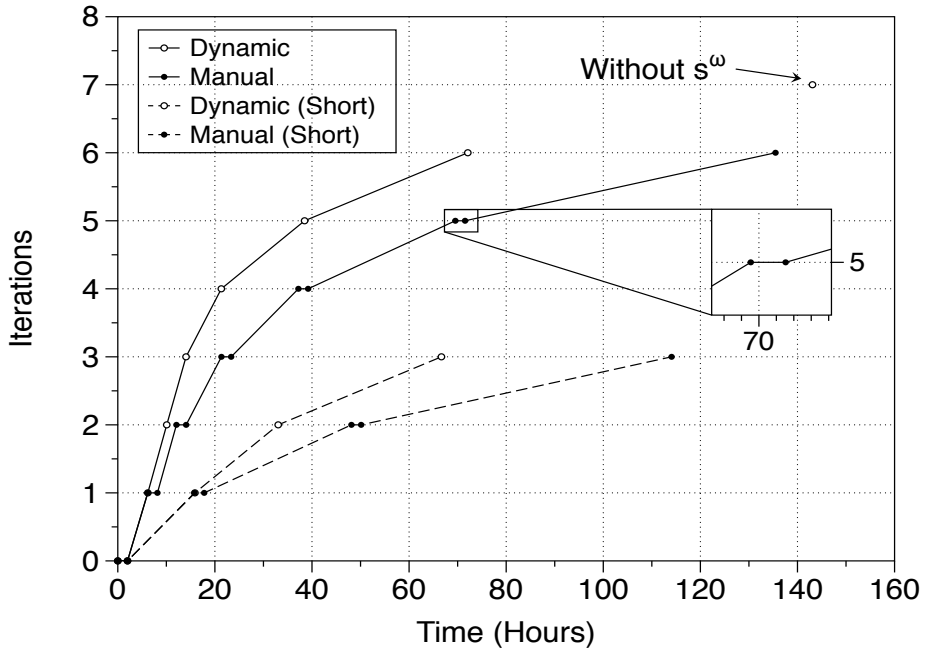


Execution time results manual and dynamic workflows

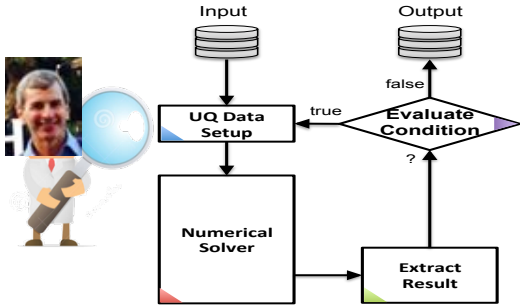


Execution time savings

Manual Choice	Correct Level			
	5	*6*	*7*	*8*
5	**-3.63%**	-	-	-
6	51.70%	**-2.05%**	-	-
7	77.67%	52.83%	**-1.11%**	-
8	89.69%	78.22%	53.32%	**-0.59%**

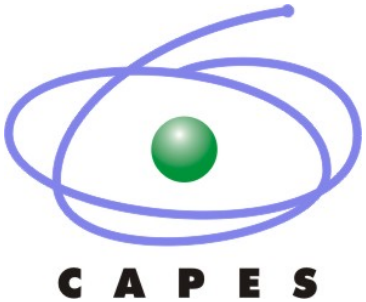


Concluding…



- Parallel Scientific Workflows & Provenance:
  - is aware of the dependencies and the data-flow
  - data-flow parallel execution
  - integration (indexing) of scientific resources (files)
  - big data analytics through provenance
- **User steering & dynamic intervention**
  - patterns that humans can easily detect but computer algorithms have a hard time finding
  - dynamic loops
- Provenance DB acts as a statistics catalog

Acknowledgements



Fourth Brazil-France Workshop

On High Performance Computing and Scientific Data Management Driven by Highly Demanding Applications

15-18 September, Gramado, Brazil 2014

Thank you!

Marta Mattoso
Federal Univ Rio de Janeiro

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

- Chiron workflow engine
  - http://sourceforge.net/projects/chironengine/
- SciCumulus: Chiron with Cloud services
  - http://sourceforge.net/projects/scicumulus/

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

Some of our papers in user steering

- Goncalves, J. ; Oliveira, D. C. M. ; Oliveira, D. ; Ocana, K. ; Ogasawara, E. ; Dias, J. ; Mattoso, M. **Performance Analysis of Data Filtering in Scientific Workflows**. Journal of Information and Data Management - JIDM, v. 4, p. 17-26, 2013.
- F. Costa, V. Silva, D. Oliveira, K. Ocana, J. Dias, E. Ogasawara, and M. Mattoso, 2013, **Capturing and Querying Workflow Runtime Provenance with PROV: a Practical Approach**, In: *Proc. of the International Workshop on Managing and Querying Provenance Data at Scale/ EDBT*
- K.A.C.S. Ocaña, D. Oliveira, J. Dias, E. Ogasawara, and M. Mattoso, 2011, **Optimizing Phylogenetic Analysis Using SciHmm Cloud-based Scientific Workflow**, In: *2011 IEEE Seventh International Conference on e-Science (e-Science)*, p. 190–197
- G. Guerra, F. Rochinha, R. Elias, D. Oliveira, E. Ogasawara, J. Dias, M. Mattoso, and A.L.G.A. Coutinho, 2012, **Uncertainty Quantification in Computational Predictive Models for Fluid Dynamics Using Workflow Management Engine**, *International Journal for Uncertainty Quantification*, v. 2, n. 1, p. 53–71
- J. Dias, E. Ogasawara, D. Oliveira, F. Porto, A. Coutinho, and M. Mattoso, 2011, **Supporting Dynamic Parameter Sweep in Adaptive and User-Steered Workflow**, In: *6th Workshop on Workflows in Support of Large-Scale Science*, *ACM/IEEE* Supercomputing'11; p. 31–36
- F. Horta, J. Dias, K. Ocaña, D. Oliveira, E. Ogasaw, and M. Mattoso, 2012, Poster: **Using Provenance to Visualize Data from Large-Scale Experiments**, In: *Poster of the ACM/IEEE* Supercomputing'13

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

Some of our papers on workflows

- J. Dias, E. S. Ogasawara, D. de Oliveira, F. Porto, P. Valduriez, and M. Mattoso. Algebraic Dataflows for Big Data Analysis. **IEEE Bigdata Conference** 2013
- E. S. Ogasawara, D. de Oliveira, P. Valduriez, J. Dias, F. Porto, and M. Mattoso. An algebraic approach for data-centric scientific workflows. Proceedings of the VLDB Endowment (**PVLDB**), 4(12):1328–1339, 2011.
- E. S. Ogasawara, J. Dias, V. Silva, F. S. Chirigati, D. de Oliveira, F. Porto, P. Valduriez, and M. Mattoso. Chiron: a parallel engine for algebraic scientific workflows. *Concurrency and Computation: Practice and Experience*, 25(16):2327–2341, 2013.
- Mattoso, M., Werner, C., Travassos, G. H., Braganholo, V., Murta, L., Ogasawara, E., Oliveira, D., Cruz, S. M. S. da, Martinho, W., (2010), "Towards Supporting the Life Cycle of Large-scale Scientific Experiments", *International Journal of Business Process Integration and Management*, v. 5, n. 1, p. 79–92.
- Mattoso, Marta ; Dias, J. ; Oliveira, D. ; Ocana, K. ; Ogasawara, E. ; Costa, F. ; Horta, F. ; Sousa, V. ; Araujo, I. . User-Steering on HPC Workflows: State-of-the-art and Future Directions. In: Scalable Workflow Enactment Engines and Technologies, 2013, Nova Iorque. **SWEET'13, SIGMOD Workshop**, 2013.
- Chirigati, F S ; Sousa, V. ; Ogasawara, E. ; Oliveira, D. ; Dias, J. ; Porto, F. ; Valduriez, P. ; Mattoso, Marta . Evaluating Parameter Sweep Workflows in High Performance Computing. In: international workshop on Scalable Workflow Enactment Engines and Technologies (SWEET'12), 2012, Phoenix. SIGMOD.
- Oliveira, D., Ocaña, K., Baião, F., Mattoso, M., (2012b), "A Provenance-based Adaptive Scheduling Heuristic for Parallel Scientific Workflows in Clouds", *Journal of Grid Computing*, v. 10, n. 3, p. 521–552.