In Situ Data Steering on Sedimentation Simulation with Provenance Data

Vítor Silva
Computer Science - COPPE
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
silva@cos.ufrj.br

José Camata
Department of Civil Engineering
COPPE
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
camata@nacad.ufrj.br

Daniel de Oliveira
Institute of Computing
Fluminense Federal University
Niterói, Brazil
danielcmo@ic.uff.br

Alvaro L.G.A. Coutinho
Department of Civil Engineering
COPPE
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
alvaro@nacad.ufrj.br

Patrick Valduriez
Zenith Team, Inria and LIRMM
Montpellier, France
patrick.valduriez@inria.fr

Marta Mattoso
Computer Science - COPPE
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
marta@cos.ufrj.br

Provenance and simulation data tracing

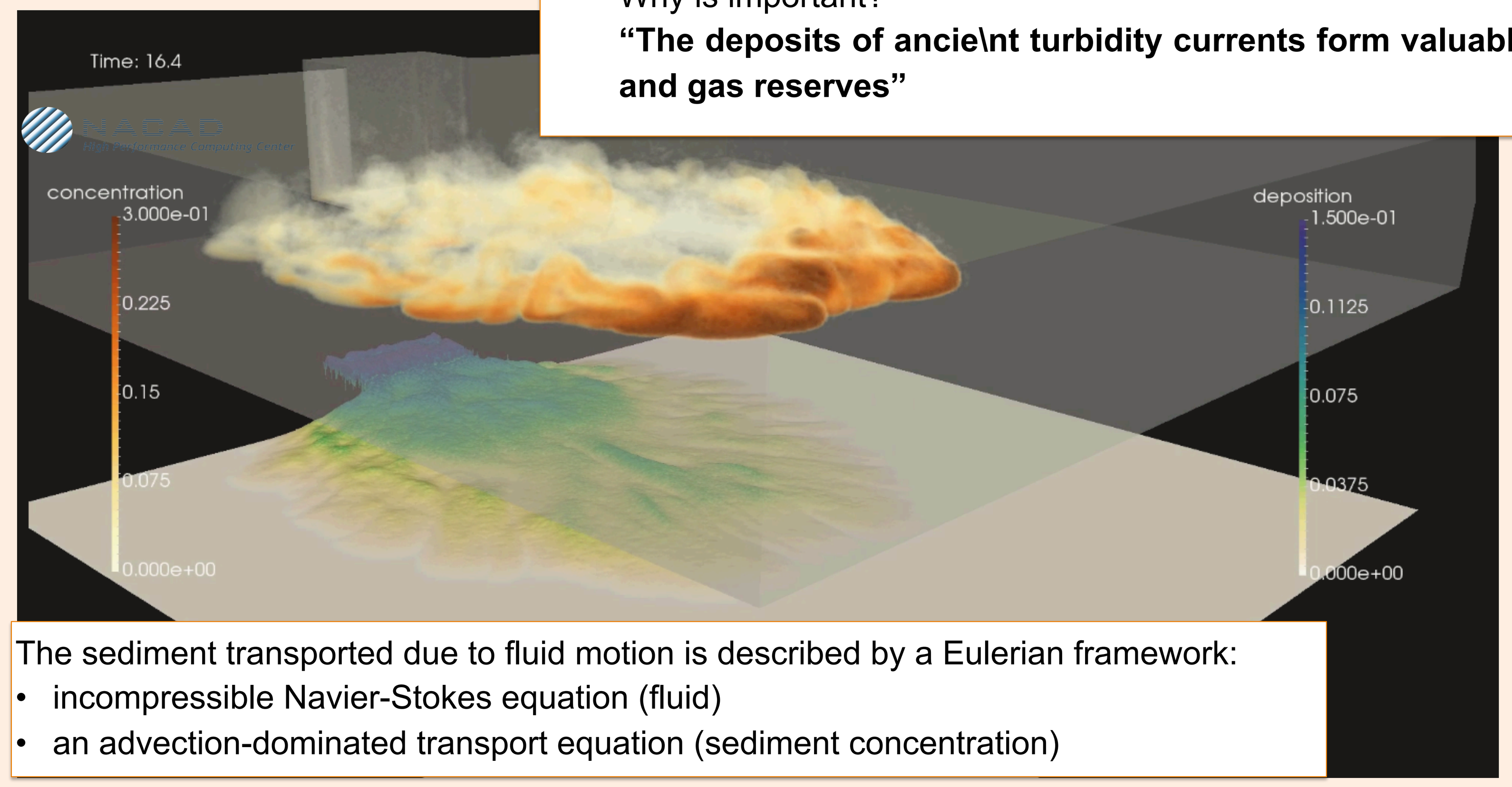
Raw data analysis with ***in situ* visualization** using Catalyst

***In situ* data steering** with strategic simulation data

Context

- **Sedimentation process:** suspended solid particles have large enough mass that they settle under their own weight
  - Turbidity current – sediments are suspended by fluid turbulence within the flow

Why is important?
"The deposits of ancie\nt turbidity currents form valuable oil and gas reserves"



The sediment transported due to fluid motion is described by a Eulerian framework:
- incompressible Navier-Stokes equation (fluid)
- an advection-dominated transport equation (sediment concentration)

- **libMesh-sedimentation**: a turbidity solver built upon libMesh [1], an open-source library with **parallel adaptive mesh refinement (AMR) and coarsening** support:
  - Employs a variational multiscale finite element method [2]
  - AMR is an optimal strategy for tacking large-scale sedimentation simulations

Motivation

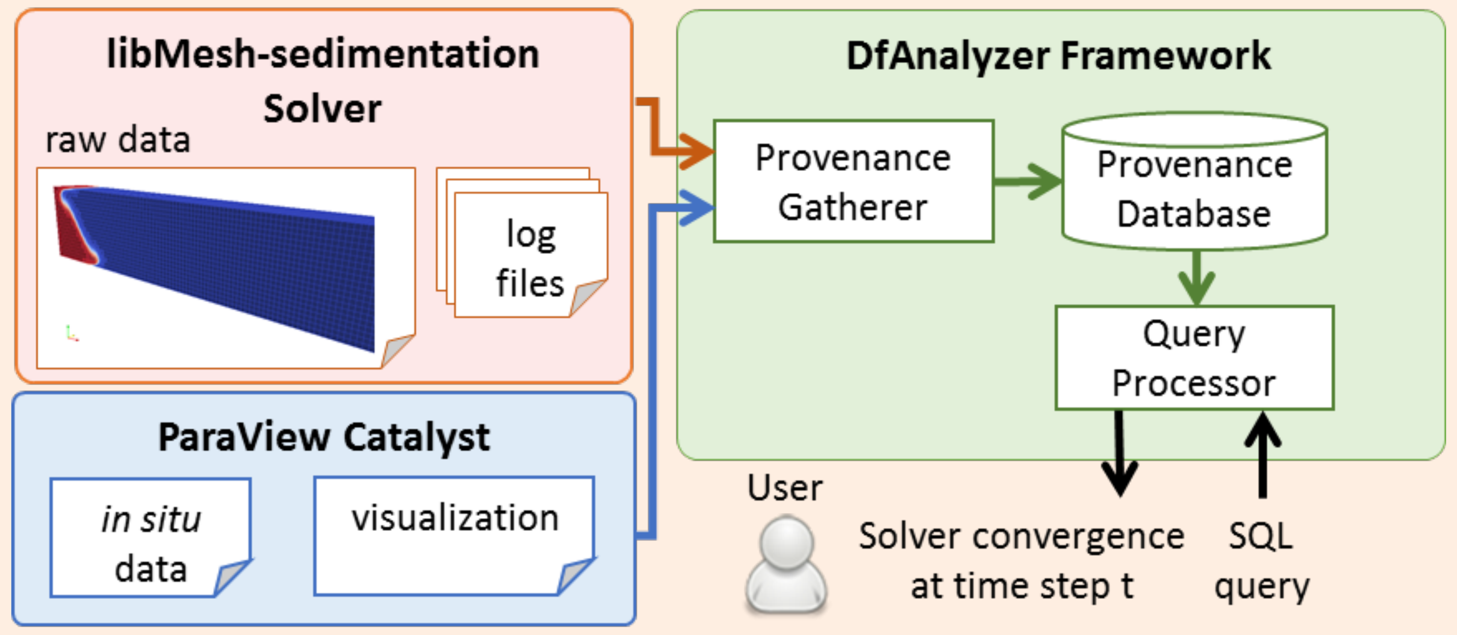
- Existing solutions only perform raw data analysis after the simulation ends, such as FastQuery [3] and RAW [4]
- In complex simulation scenarios, users need…
  - **to track quantities of interest** such as residuals, errors estimates, etc. to control as much as possible the execution
  - **to analyze raw data from multiple related files** generated in different solver steps at runtime
- Due to these issues, we propose DfAnalyzer, a solution based on provenance data to extract and related strategic simulation data for online queries

Enabling *In Situ* Data Analysis

- A **ParaView Catalyst adaptor** embedded into our parallel libMesh-sedimentation solver to **support *in situ* data analysis and visualization**
  - It accesses libMesh data structures extracting additional data information from simulation field
  - Python scripts are generated on ParaView UI and invoked by the Catalyst adaptor

DfAnalyzer Framework

- **Dataflow analysis** [7] based on two components:
  Provenance Gatherer (PG) and **Query Processor (QP)**
  - **PG** → Raw data & Provenance data capture
  - Provenance and Raw data loading into the provenance database that follows the W3C PROV recommendations [8]
  - **QP** → Interfaces SQL queries submitted by the users to MonetDB
- The raw data aware provenance database is ready for queries at runtime
- DfAnalyzer is an instantiation of ARMFUL, a component-based architecture for analyzing raw data from multiple files, to the sedimentation solver



Experimental Evaluation

- A proof-of-concept experiment executes a sedimentation simulation on 1040 cores in Stampede cluster, at the Texas Advanced Computing Center (TACC)
- Sedimentation solver runs 200 time steps
  - Adaptive mesh refinement applied every 10 time steps
  - Initial input mesh with 480 x 80 x 80 hexahedra

Elapsed time in minutes

solver	*in situ data analysis*	*provenance data capture*	*total*
136.80 (99.31%)	0.60 (0.44%)	0.35 (0.25%)	137.75

Monitoring: analysis of the sediment concentration in a horizontal line in the mesh

time step	x	y	z	d
1	0.000	1.000	0.000	2.00E-04
1	0.180	1.000	0.000	2.00E-04
1	0.360	1.000	0.000	2.00E-04
1	0.540	1.000	0.000	2.00E-04
1	0.720	1.000	0.000	1.99E-04
1	0.900	1.000	0.000	1.19E-04
1	1.080	1.000	0.000	3.04E-08
…	…	…	…	…

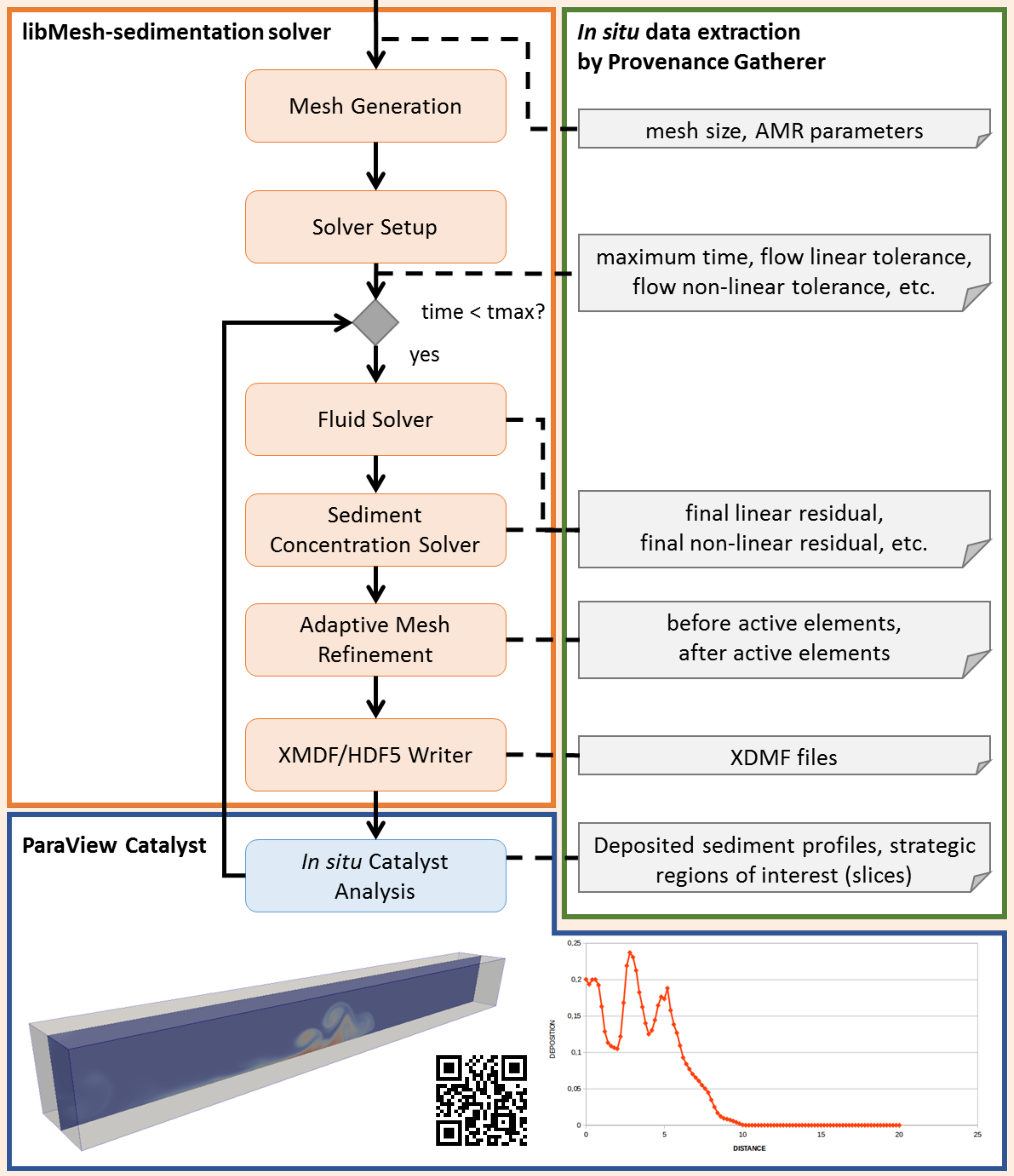
Debugging and user steering: analysis of the algorithm output parameters after the simulation solver convergence

Fluid		Sediments	
linear residual norm	nonlinear residual norm	linear residual norm	nonlinear residual norm
3.98E-06	13.54823207	8.66E-06	0.004445721
4.30E-06	0.390224835	2.00E-09	0.002435432
4.30E-06	0.390224835	1.31E-05	0.016144017
7.00E-09	0.002712742	2.00E-09	0.002435432
7.00E-09	0.002712742	1.31E-05	0.016144017
…	…	…	…

Acknowledgements

This work was partially funded by CNPq, CAPES, FAPERJ, Inria, and the European Commission (HPC4E H2020 project). These research results have received funding from the EU H2020 Programme and from MCTI/RNP-Brazil under the HPC4E grant agreement n° 689772. Computer time on Stampede is provided by the Texas Advanced Computer Center (TACC) at University of Texas at Austin. Stampede is supported by the US National Science Foundation under award ACI-1134872

In situ Data Steering on Sedimentation Solver



References

[1] B.S. Kirk, J.W. Peterson, R.H. Stogner, and G.F. Carey. libMesh : a C++ library for parallel adaptive mesh refinement/coarsening simulations. *Engineering with Computers*, 22(3-4):237–254, 2006.

[2] G.M. Guerra, S. Zio, J.J. Camata, J. Dias, R.N. Elias, M. Mattoso, P.L. B. Paraizo, A.L. G. A. Coutinho, and F.A. Rochinha. Uncertainty quantification in numerical simulation of particle-laden flows. *Computational Geosciences*, 20(1):265–281, 2016.

[3] J. Chou, K. Wu, and Prabhat. FastQuery: A Parallel Indexing System for Scientific Data. *IEEE International Conference on Cluster Computing*, 455–464, 2011.

[4] M. Karpathiotakis, M. Branco, I. Alagiannis, and A. Ailamaki, 2014, Adaptive Query Processing on RAW Data, VLDB Endowment, v. 7, n. 12, p. 1119–1130.

[5] S.B. Davidson and J. Freire. Provenance and scientific workflows: challenges and opportunities. *ACM SIGMOD international conference on Management of data*, 1345–1350, 2008.

[6] U. Ayachit, A. Bauer, B. Geveci, P. O'Leary, K. Moreland, N. Fabian, and J. Mauldin. ParaView Catalyst: Enabling In Situ Data Analysis and Visualization. *In Situ Infrastructures for Enabling Extreme-Scale Analysis and Visualization*, 25–29, 2015.

[7] V. Silva, D. de Oliveira, P. Valduriez, and M. Mattoso. Analyzing related raw data files through dataflows. *CCPE*, 28(8):2528–2545, 2016.

[8] P. Groth and L. Moreau. W3C PROV - An Overview of the PROV Family of Documents. Available at: https://www.w3.org/TR/prov-overview, 2013.