SIAM Conference on Computational Science and Engineering

February 27-March 3, 2017
Hilton Atlanta, Atlanta, Georgia, USA

Enabling In Situ Viz and Data Analysis with Provenance in libMesh

Vítor Silva
Jose J. Camata
Marta Mattoso
Alvaro L. G. A. Coutinho
(Federal university Of Rio de Janeiro/Brazil)

Patrick Valduriez
(INRIA/France)

NACAD
Núcleo Avançado de Computação de Alto Desempenho

HPC

Summary

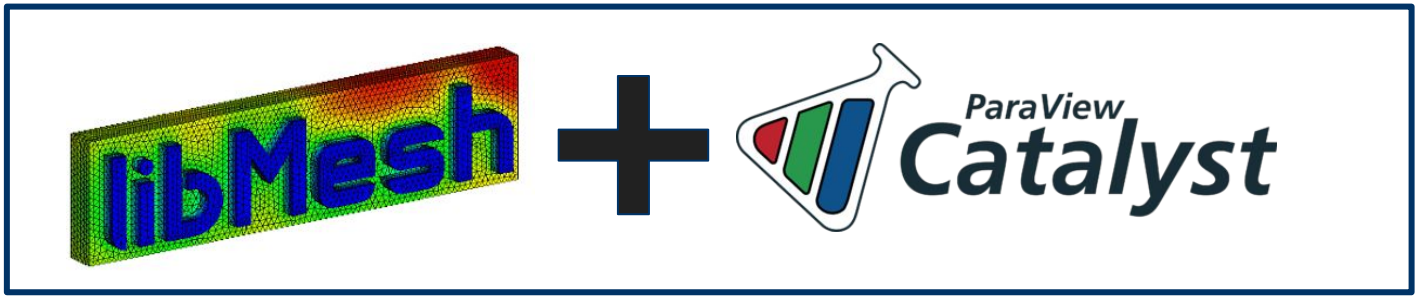
- Introduction
- libMesh Sedimentation Solver
- HPC Simulation Enhancements
  - In-Situ Viz and Raw Data Analysis
  - Provenance support
- Test Cases
  - de Rooij and Dalziel sedimentation tank
  - Real bed bathymetry
- Conclusions

Motivation

- In highly complex simulations, **data is efficiently managed in memory and stored in thousands of "isolated" files** (HDF5, XDMF, Exodus, NetCFD, Ensight, VTK, etc)
  - These data have to be related to enable viz and data analyses at runtime and, very often, after the simulation.
  - Track evolution of QoIs, associate simulation parameters and time steps to identify regions of interest spread among files: **provenance data**
- **Goal**:
  - Online analysis to detect outliers and dispensable input data to dynamically adapt the simulation
- **Challenge:**
  - ***no threat to solver's performance and scalability***

Objectives

- Provide In-Situ visualization capabilities to **libMesh**
  - libMesh is now Paraview-Catalyst enabled



- Introduce Dataflow Analysis tools (**DfAnalyzer**) based on data provenance
  - Tools enable **user steering**
- Show a cost analysis of In-Situ Viz and Dataflow Analysis tools in parallel simulations of sedimentation problems

Context: Sedimentation Simulation using libMesh

- **libMesh-sedimentation:** a sediment transport solver for simulating turbidity currents built upon **libMesh**, an open-source library with parallel adaptive mesh refinement and coarsening (AMR/C) support:
  - Employs a ***residual-based variational multiscale finite element method*** (Guerra et al, IJNMF, 2013)
  - AMR/C is an optimal strategy for large-scale sedimentation simulations

- The sediment transport due to fluid motion is described by a fully-Eulerian framework:
  - incompressible Navier-Stokes equation (fluid)
  - advection-dominated transport equations (sediment concentrations)

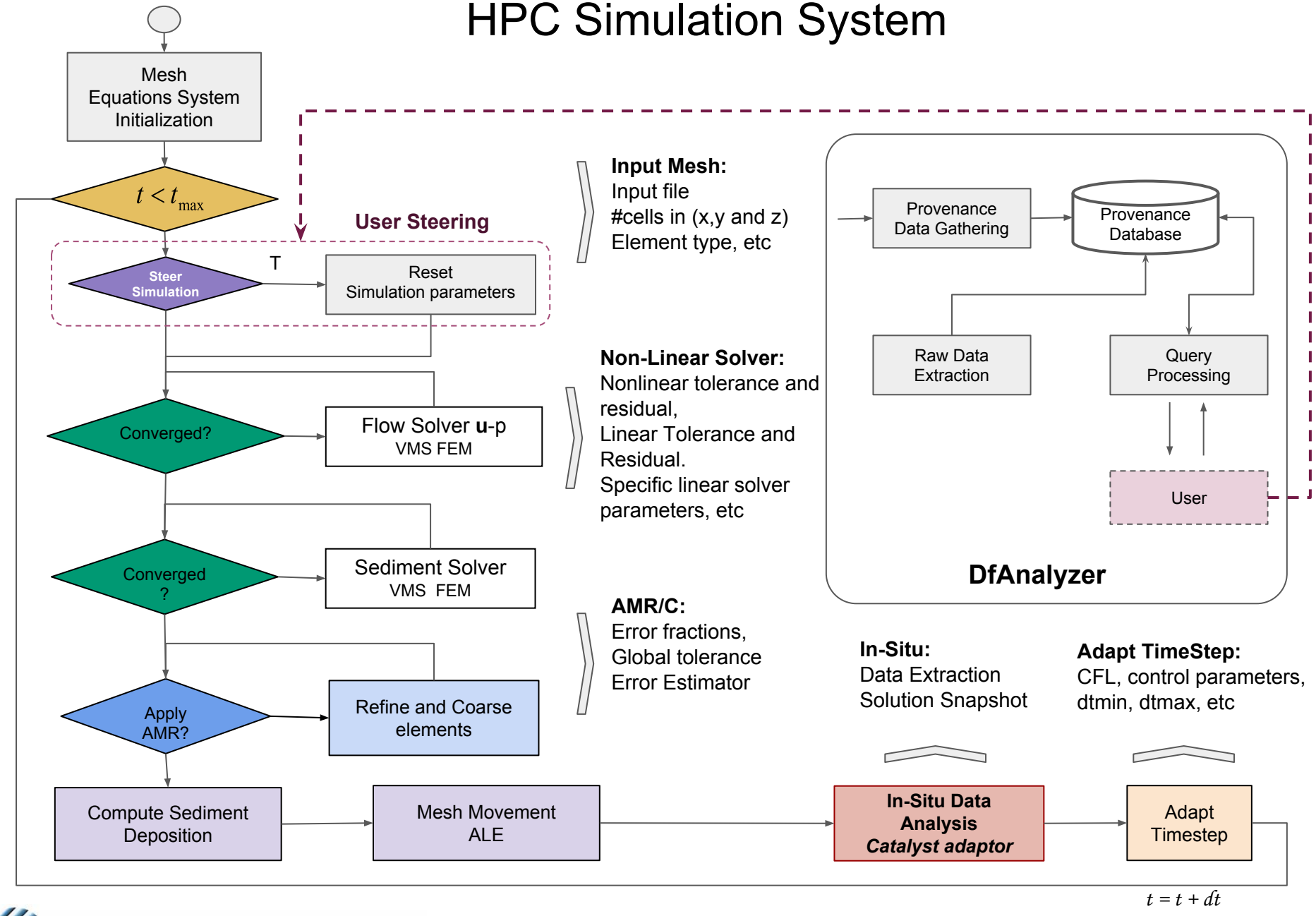
Provenance in Data Analysis

- In data-intensive high performance simulations, provenance data can help users to track all simulations parameters and relate thousands of raw data files
  - Allows reproducibility
  - Data analytical queries in these raw data files is challenging
    - Current solutions are independent and offline

- In our work, data analysis is improved combining provenance support and online queries on raw data file contents
  - Access to raw data files while they are generated
  - Parse raw data files to find useful info and extract relevant subsets of raw data
  - Index over data regions of interest
  - Prepare raw data for queries
  - Runtime queries relating raw data from different files, provenance data, and performance execution data

- Query results can help the user to ***steer*** their simulations

HPC Simulation System

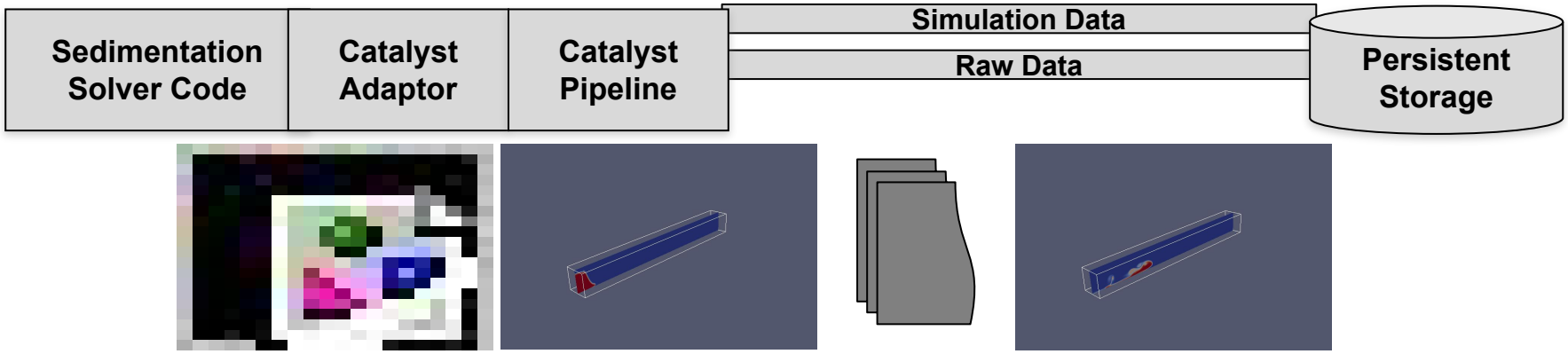


NACAD
Núcleo Avançado de Computação de Alto Desempenho

In-Situ Raw Data Analysis

- Integration between libMesh and Paraview Catalyst is provided by an ***adaptor*** (implemented in one of the Paraview Catalyst APIs)
  - Adaptor invokes Python scripts (exported from ParaView UI) for raw data analysis and visualization
  - Data structures from the simulation code accessed in memory (*in-situ*) are mapped into the adaptor using Paraview Catalyst API
    - Catalyst uses VTK's data model to perform the data structure mapping
    - *e.g.*, we map, mesh, velocity, pressure, sediment appearance, etc. from our sedimentation solver to VTK's data model

***In-situ* raw data analysis**



We can ALSO use this Paraview Catalyst adaptor for *in-transit* data analysis

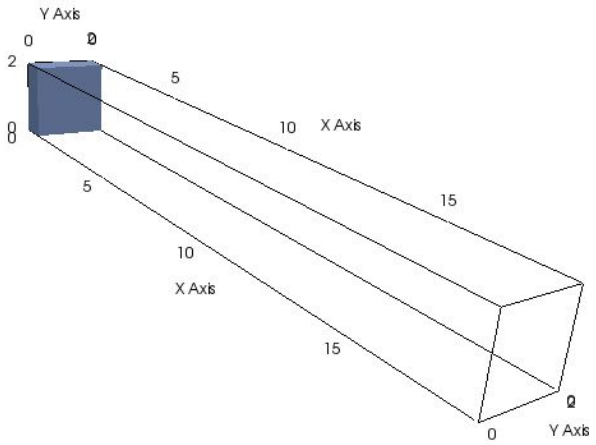
Computational Setup

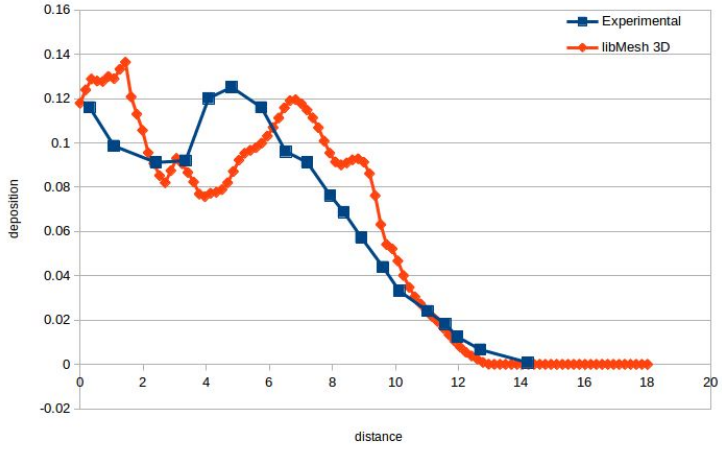
- Experiments were carried out on *Lobo Carneiro* machine at NACAD/COPPE/UFRJ[1]
  - SGI ICE-X 252 computer nodes
  - Each node with two Intel Xeon E5-2670v3 (Haswell)
  - 64GB of memory per node
  - Network: Infiniband FDR - 56Gb/s (Hypercube)
- libMesh-Sedimentation solver:
  - Intel Compilers (version 16.0) with -03 optimization flag
  - MPI Intel
  - Linked with Paraview Catalyst (v. 5.3) / Offscreen Rendering (Mesa 13.0 version)
  - Linked with Dataflow Analyzer (DfAnalyzer)
- Run configuration:
  - Standard nodes for libMesh and Paraview Catalyst
  - A dedicated node for simulation data management
    - Includes database component from DfAnalyzer

[1]http://www.nacad.ufrj.br/en/recursos/sgiicex

Test Case: de Rooij and Dalziel Sedimentation Tank

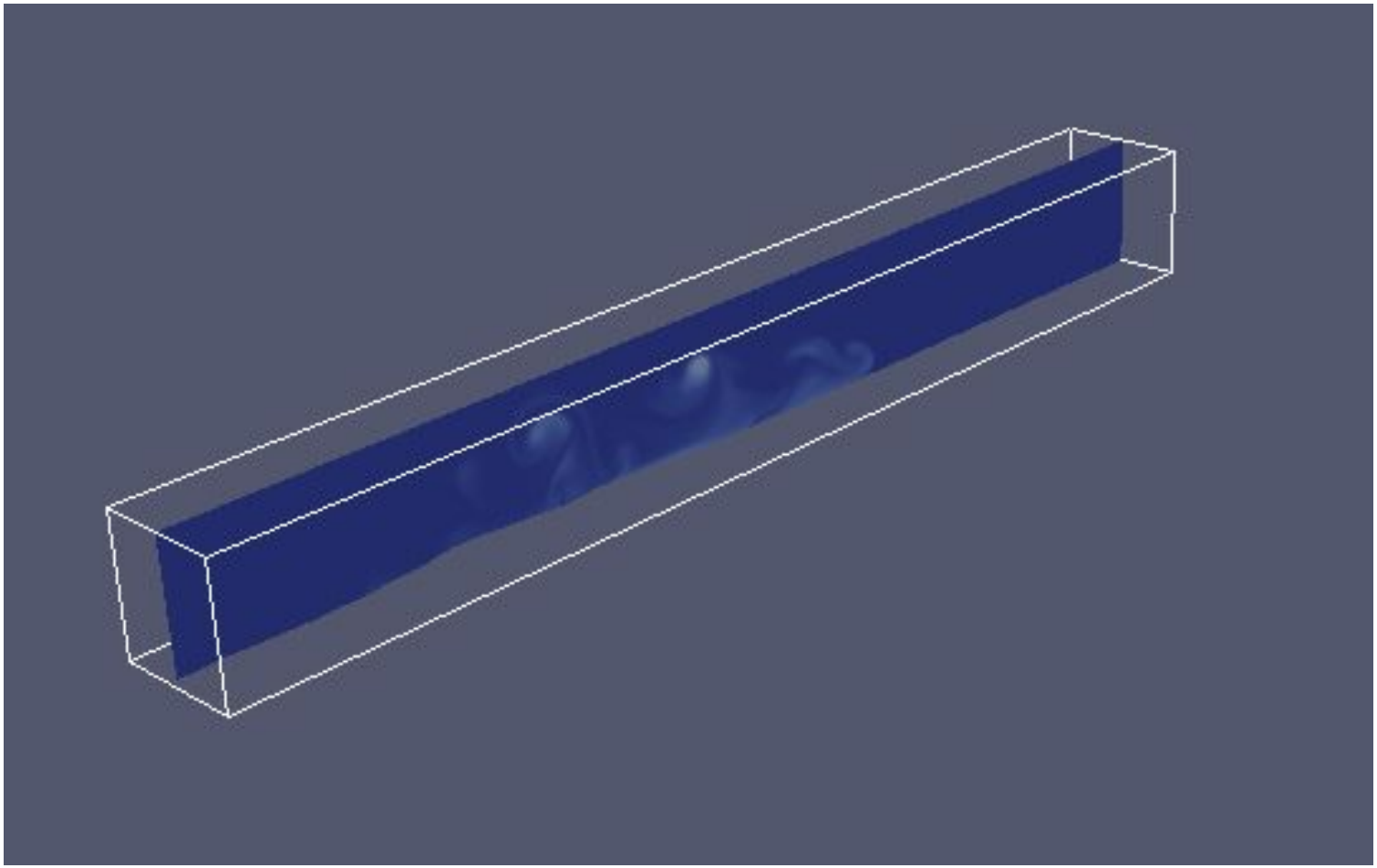
- Simulation Setup
  - Domain size: 20.0 x 2.0 x 2.0
  - Initial uniform grid in all directions, grid space 0.075
  - One initial uniform refinement, solution with max refine level=3
    - 1.5M HEX8 elements
  - Kelley's error estimator for **u** and c
  - The lock, in which the fluid initially is at rest, has dimensions 0.75 x 2.0 x 2.0
  - Reynolds number Re=5000
  - Run with 480 cores at Lobo Carneiro
- Simulation Management
  - Raw data files are written each 50 time steps
  - In-situ data extraction are called each 50 time steps
  - In-situ visualization generates pngs files each 50 timesteps
  - Catalyst data extraction: plot over line filter
  - Catalyst viz: slice filter





In-Situ Catalyst data extraction plot over line filter

Test Case: de Rooij and Dalziel Sedimentation Tank



In-Situ Catalyst viz: slice filter

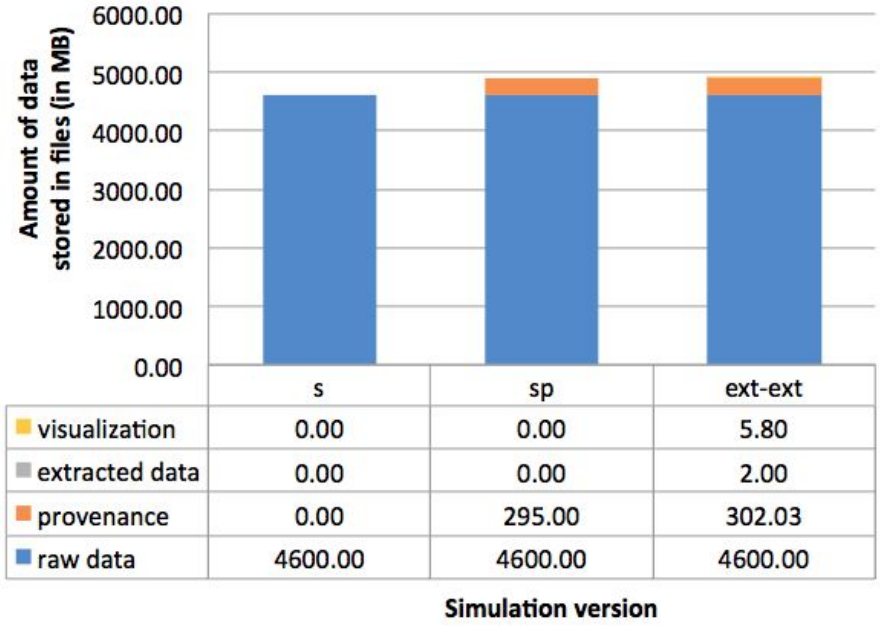
Sedimentation Tank: Computation Cost - CPU Time

TABLE 1: Elapsed time for different stages on libMesh-Sedimentation - Sedimentation Tank

Time Contribution	CPU Time (in s)	cost/call	%cost
Flow Solver	23533.21	1.26	**37.85%**
AMR/C	13122.20	93.73	**21.11%**
Sediment Solver	4941.66	0.27	**7.95%**
InSitu Catalyst Viz+Extraction	2065.08	22.45	**3.32%**
XDMF/HDF5 Raw Data	1329.21	14.45	**2.14%**
Provenance (DfAnalyzer)	83.38	0.01	**0.13%**
Others (libMesh)	17096.26		**27.5%**
Total	62171.00		

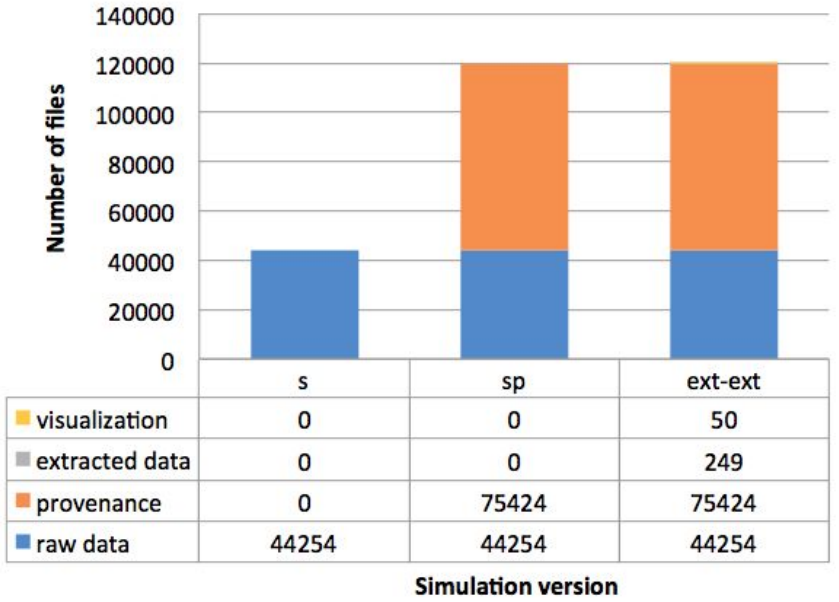
- **Remarks:**
  - Provenance adds low overhead to overall simulation costs.
  - In-Situ Viz + extraction cost in relation to Raw data Writer could be offset by disk bandwidth constraints as well as limited disk capacity.
  - CPU time spent in In-Situ Catalyst depends how many filters are applied.

Sedimentation Tank Computation Cost: Raw Data vs. In-Situ and Provenance Data Storage



	s	sp	ext-ext
visualization	0.00	0.00	5.80
extracted data	0.00	0.00	2.00
provenance	0.00	295.00	302.03
raw data	4600.00	4600.00	4600.00

Simulation version



	s	sp	ext-ext
visualization	0	0	50
extracted data	0	0	249
provenance	0	75424	75424
raw data	44254	44254	44254

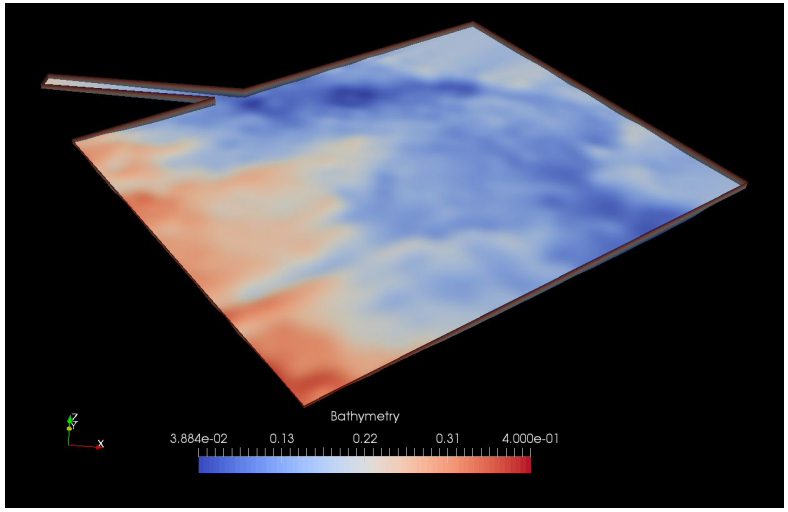
Simulation version

Legend:
s:: Solver
sp: Solver with Provenance
ext-ext: Solver + Provenance + In-situ

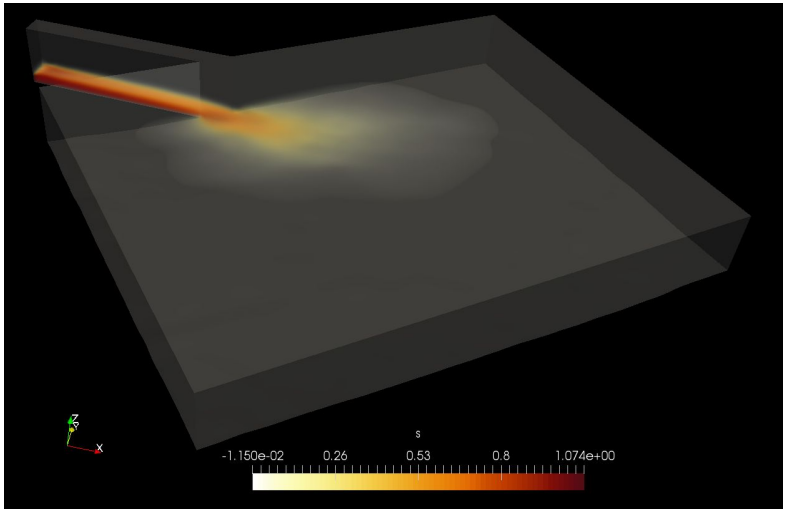
Provenance Data: 6%
In-Situ Visualization : 0.12%
Data Extraction: 0.04%
% in relation of raw data stored

Real Case: Sedimentation and Deposition with Real Bed Bathymetry

- **Run on 480 cores at Lobo Carneiro**
- **Domain Size:**
  - Tank: 14 x 12 x 2
- **Fixed Unstructured mesh:**
  - 7.6M linear tetrahedral elements
  - 1.4M nodes
- **Dimensionless Parameters:**
  - Grashof: 10^6 (Reynolds approx. 2000)
  - Monodisperse:
    - Settling velocity: 5.6651E-03
- **Boundaries conditions:**
  - **Flow:**
    - no slip is applied at bottom and channel walls
    - Prescribed velocity at front wall channel
    - Free slip at top
  - **Sediment:**
    - Concentration prescribed at front channel wall
- **Data Analysis**
  - In-Situ visualization of sediment deposition
  - Data Extraction: sediments profile over lines



Bed bathymetry



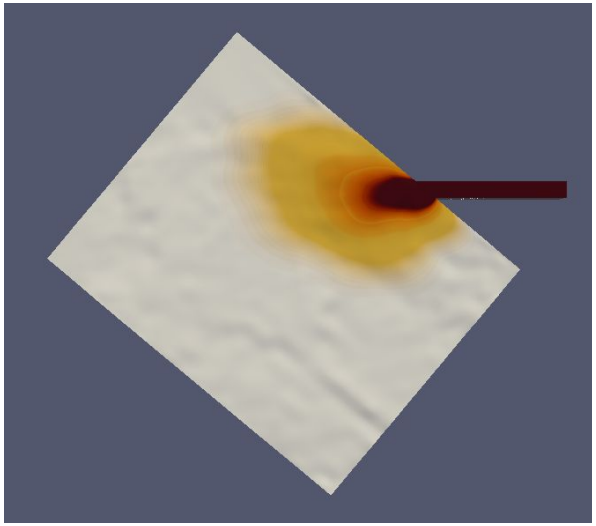
Sediment concentration at t=200

Real Bed Bathymetry Case: Computation Cost

TABLE 2: Elapsed time for different stages - Real Bed Bathymetry Case - Simulation final time t=100

Time Contribution	Elapsed Time	%Cost
Flow Solver	72523.49	50.71%
Concentration Solver	28000.50	19.58%
In-Situ Viz + Extraction	2175.16	1.52%
XDMF/HDF5 Raw data	421.23	0.29%
Provenance	451.70	0.32%
Total	143029.00	

Storage Requirements	Size (in GB)	% Raw Data
XDMF/HDF5 Raw data	23.44	
Provenance + Data Extraction	0.38	1.60%
In-Situ Viz + Extraction	0.28	1.21%



Sedimentation map generated by In-Situ Catalyst ⇨

Conclusions

- In this work we have shown the integration of In-Situ Visualization and Dataflow Analysis in libMesh
- We have used ParaView Catalyst and DfAnalyzer, openly available
- Performance measurements in representative sedimentation test cases show low overhead for both In-Situ Visualization and Dataflow Analysis tools
- What we have gained?
  - Parallel runtime interaction with simulation parameters: reset tolerances, solvers, etc, querying the provenance database from DfAnalyzer
  - Tracking QoIs at runtime through visual information provided by Paraview Catalyst/DfAnalyzer
  - Registry of computational set-up for complex multiphysics analysis for further simulations: increases reproducibility
  - Possible reduction of data stored in persistent file systems
- Want to see more?
  - Steering UQ: Dias *et al.*, FGCS, 2015
  - UQ-support in sedimentation simulations: Guerra *et al.*, Computational Geosciences, 2016
  - Dataflow Analysis in CFD, Silva *et al.*, FGCS, 2017
- Future work: scalability at higher core counts

Thanks