Accelerating Time to Insight in the Exascale Ecosystem Through the Optimization of Scientific Workflows

James Ahrens

ahrens@lanl.gov

Los Alamos National Laboratory

LA-UR-15-20354

- Our goal is to:
  - identify, automate and accelerate the manual operations that pervade our current workflows
- Discuss trends and examples

Trends: Automation and Functional Abstraction

- Ideally automation of the complete workflow
  - Apply Amdahl's Law
    - Completion time limited by workflow's sequential interactive user time
  - **Manual tasks**
    - **visualization/analysis, meshing, debugging, data movement and scheduling**
  - ***Visualization and analysis example:***
    - ***Deeper change*** than simply running existing interactive analysis operations in a batch-oriented manner

- Functional abstraction and encapsulation
  - At all scales:
    - from tasks defined in a program
    - programs in a workflow
    - workflows in a meta workflows such as those found in ensemble calculations and V&V
  - Benefits - No side effects
    - Parallelism – Clear dependencies and independent units
    - Resilience – make transactional
    - Reproducibility

Trends: “Processors everywhere” and Cost Models

- Opportunity to accelerate our workflow on the multitude of processors in the HPC ecosystem
  - Processors WITH memory, burst buffers, storage and network resources
  - Executing tasks at all computational scales
  - Example:
    - network and storage processors
      - prefetch data for simulation setup
    - floating point processors
      - setup information to calculate ensembles of simulation results
    - memory-associated processors
      - scan results for correlations
- Cost Model
  - Guidance to the scientist about the value/importance of resources
  - Scheduler can automatically concurrently assign components to optimize a workflow

Examples - Automation, Functional Abstraction, Processors, Cost Models

Example 1 – Example Visualization and Analysis

- Traditional Post-processing
  - COST = I/O time for full timesteps + post-processing analysis time
- In situ Image Database
  - COST = analysis time + I/O time for analysis products

Example 2 - Reasoning about optimal computing/storage data representation

- Address how often to save results
  - With notion of data regeneration service
  - User and compute time, storage size