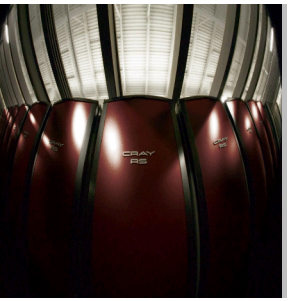
SAND2014-17729PE



Addressing Scientific I/O Needs for Current and Future Architectures

Ron A. Oldfield
Sandia National Laboratories
Albuquerque, NM, USA

Storage Systems and I/O (SSIO) Summit

September 2014

Sandia National Laboratories

Exceptional service in the national interest

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Extreme-Scale Computing

Sandia National Laboratories

- Trends: More FLOPS with comparatively less storage, I/O bandwidth
  - Consequence: A smaller fraction of data can be captured on disk

Oak Ridge National Laboratory

	System Peak	I/O BW
Jaguar (2008)	263 TFLOPS	44 GB/s
Jaguar PF (2009)	1.75 PFLOPS	240 GB/s
Titan (2012)	20 PFLOPS	240 GB/s
Factor Change	**76×**	**5.5×**

Bland, Kendall, Kothe, Rogers, and Shipman. "Jaguar: The World's Most Powerful Computer"
http://archive.hpcwire.com/hpcwire/2012-10-29/titan_sets_high-water_mark_for_gpu_supercomputing.html?featured=top

Argonne National Laboratory

	System Peak	I/O BW
Intrepid (2003)	560 TFLOPS	88 GB/s
Mira (2011)	10 PFLOPS	240 GB/s
Factor Change	**17.8×**	**2.7×**

https://www.alcf.anl.gov/intrepid
https://www.alcf.anl.gov/mira

Lawrence Livermore National Laboratory

	System Peak	I/O BW
ASC Purple (2005)	100 TFLOPS	106 GB/s
Sequoia (2012)	20 PFLOPS	1 TB/s
Factor Change	**200×**	**9.4×**

http://www.sandia.gov/supercomp/sc2002/flyers/SC02ASCIPurplev4.pdf
https://asc.llnl.gov/publications/Sequoia2012.pdf

Sandia National Laboratories

	System Peak	I/O BW
Red Storm (2003)	180 TFLOPS	100 GB/s
Cielo (2011)	1.4 PFLOPS	160 GB/s
Factor Change	**7.8×**	**1.6×**

https://cfwebprod.sandia.gov/cfdocs/CCIM/docs/033768p.pdf
http://www.lanl.gov/orgs/hpc/cielo/

Extreme-Scale Computing

- Trends: More FLOPS with comparatively less storage, I/O bandwidth
  - Consequence: A smaller fraction of data can be captured on disk

If you really need motivation for I/O research, I'm probably in the wrong room!

Oak Ri

Jaguar		
Jaguar		s
Titan (		
Facto		

Bland, Kenda
http://archive

Law

					W
ASC Purple (2005)	100 TFLOPS	106 GB/s			B/s
Sequoia (2012)	20 PFLOPS	1 TB/s	Cielo (2011)	1.4 PFLOPS	GB/s
Factor Change	**200×**	**9.4×**	**Factor Change**	**7.8×**	**1.6×**

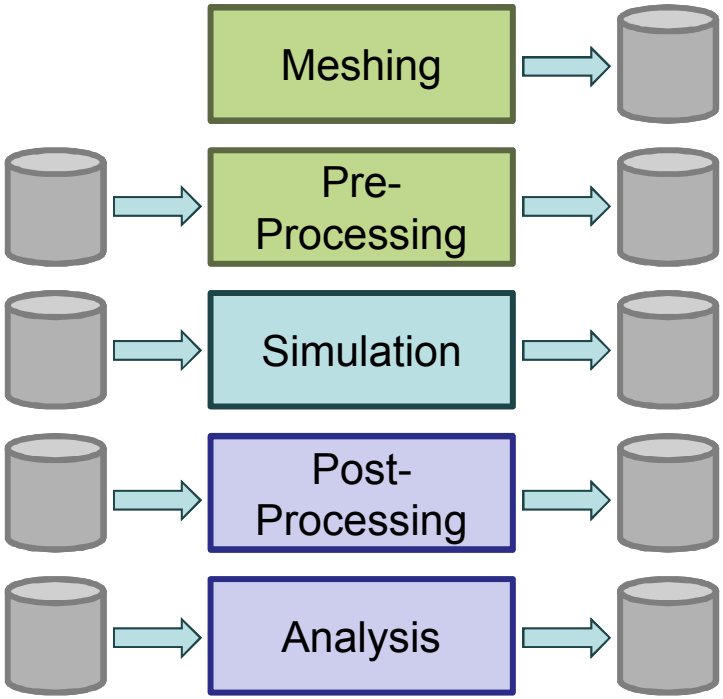
http://www.sandia.gov/supercomp/sc2002/flyers/SC02ASCIPurplev4.pdf
https://asc.llnl.gov/publications/Sequoia2012.pdf

https://cfwebprod.sandia.gov/cfdocs/CCIM/docs/033768p.pdf
http://www.lanl.gov/orgs/hpc/cielo/

Usage Models Conflict with Trends

Sandia National Laboratories

App workflows historically use parallel file system for communication

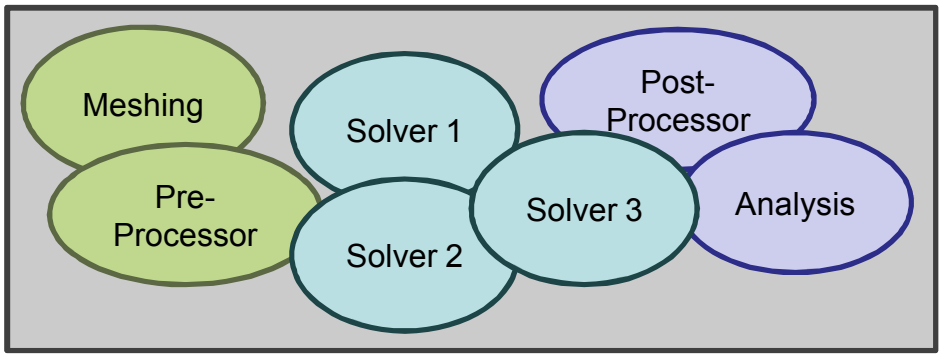


One way to relieve I/O pressure is to integrate components

Two Existing Approaches to Integration

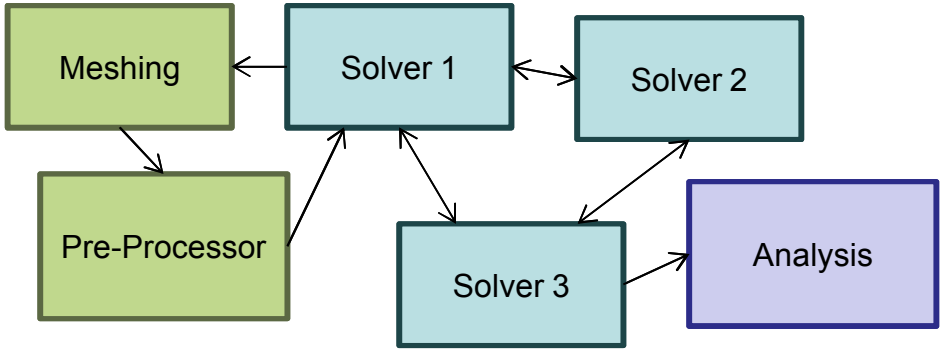
Sandia National Laboratories

Tightly Coupled (In Situ)



- Pros
  - Standard communication (MPI)
  - Supported by HPC runtimes
- Implementation Challenges
  - Configuration/build (lib conflicts)
  - Data structure mismatches
  - Resilience (one fails, they all fail)

Loosely Coupled (In Transit)



- Pros
  - Configuration/build is easy
  - Resilience is easier to manage
- Implementation Challenges
  - Not well supported by runtimes
  - No dynamic scheduling, placement, load balancing, …
  - No standard comm interface

Observations

Sandia National Laboratories

- Tightly coupled and loosely coupled approaches will co-exist
- Gaps remain before these approaches become “productive”
  - Need portable, fast, memory-efficient mechanisms and interfaces for sharing data
    - POSIX file system is not sufficient
    - Need the right “hooks” into in-memory data structures (avoid copies)
    - Need to deal with data structure mismatches in coupled codes
    - Need to deal with multi-resolution/multi-scale issues
  - Need new definitions for “persistence” of transient data
    - E.g., time windows, data set versioning, …
  - Need new system software that supports integrated workflows
    - Scheduling, load balancing, node and data placement
    - Runtime requirements may differ for coupled components
  - Need resilience…everywhere… nuff said

We've been addressing some of the gaps

Sandia National Laboratories

- Capabilities for Integrated Workflows (Nessie, NNTI – ASC)
  - RPC-based framework for developing data services
  - Portable RDMA abstraction over HPC interconnects (Cray XT/XE, IBM BG, IB)
- Capabilities for data sharing (Kelpie, Sirocco – ASC)
  - Kelpie: In-memory, high-performance key-value store
  - Sirocco: Peer-to-peer like storage system. Supports many media, adaptable and resilient.
- Capabilities for In-situ Analysis and Visualization (ASC)
  - ParaView/Catalyst (SNL/Kitware)
  - Current focus on modularity, low memory footprint, scalability
- Resilient integrated workflows (D2T – LDRD)
  - D2T (Lofstead) – distributed transaction-based approaches
- OS and Runtime changes to support integrated workflows (ASC and ASCR)
  - Hobbes and Argo – Both ASCR projects
  - Resource management, data sharing, application composition, prog models.