Input/Output APIs and Data Organization for High Performance Scientific Computing

November 17, 2008

Petascale Data Storage Institute Workshop at Supercomputing 2008

Jay Lofstead (GT), Fang Zheng (GT), Scott Klasky (ORNL), Karsten Schwan (GT)

Overview

- Motivation
- Architecture
- Performance

Motivation

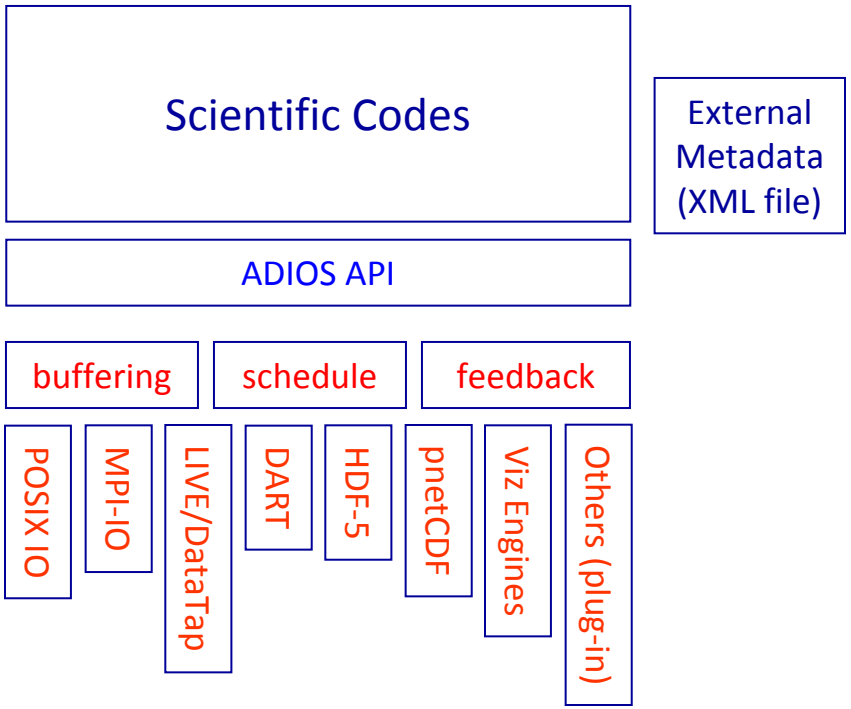
- Many codes write lots of data, but rarely read
  - TBs of data
  - different types and sizes
- HDF-5 and pNetCDF used
  - convenient
  - tool integration
  - portable format

Performance/Resilience Challenges

- pNetCDF
  - “right sized” header
  - coordination for each data declaration
  - data stored as logically described
- HDF-5
  - b-tree format
  - coordination for each data declaration
  - single metadata store vulnerable to corruption.

Architecture (ADIOS)

- Change IO method by changing XML file
- Switch between synchronous and asynchronous
- Hook into other systems like visualization and workflow



Architecture (BP)

- Individual outputs into “process group” segments
- Metadata indices next
- Index offsets and version flag at end

Process Group 1	Process Group 2	...	Process Group n	Process Group Index	Vars Index	Attributes Index	Index Offsets and Version #

Resilience Features

- Random node failure
  - timeouts
  - mark index entries as suspect
- Root node failure
  - scan file to rebuild index
  - use local size values to find offsets

Data Characteristics

- Identify file contents efficiently
  - min/max
  - local array sizes
- Local-only makes it “free”
  - no communication
- Indices for summaries/direct access
  - copies for resilience

Architecture (Strategy)

- ADIOS API for flexibility
  - Use PHDF-5/PNetCDF during development for “correctness”
  - Use POSIX/MPI-IO methods (BP output format) during production runs for performance

Performance Overview

- Chimera (supernova) (8192 processes)
  - relatively small writes (~1 MB per process)
  - 1400 seconds pHDF-5 vs. 1.4 seconds POSIX (or 10 seconds MPI-IO independent)
- GTC (fusion) (29,000 processes)
  - 25 GB/sec (out of 40 GB/sec theoretical max) writing restarts
  - 3% of wall clock time spent on IO
  - > 60 TB of total output

Performance Overview

- Collecting characteristics unmeasurable
  - 10, 50, 100 million entry arrays per processes
  - 128-2048 processes
  - weak scaling

Performance Overview

- Data conversion
  - Chimera 8192 process run took 117 seconds to convert to HDF-5 (compare 1400 seconds to write directly) on a single process
  - Other tests have shown linear conversion performance with size

Parallel conversion will be faster...

Summary

- Use ADIOS API
  - selectively choose consistency
- BP intermediate format
  - performance
  - resilience
  - later convert to HDF-5/NetCDF

Questions?