Advanced Data Placement via Ad-hoc File Systems at Extreme Scales (ADA-FS)

Michael Kluge, Wolfgang E. Nagel, André Brinkmann, Achim Streit,
Sebastian Oeste, Marc-André Vef, Mehmet Soysal

PDSW-DISCS @ SC'16 Salt Lake City, 2016/11/24

Center for Information Services & High Performance Computing

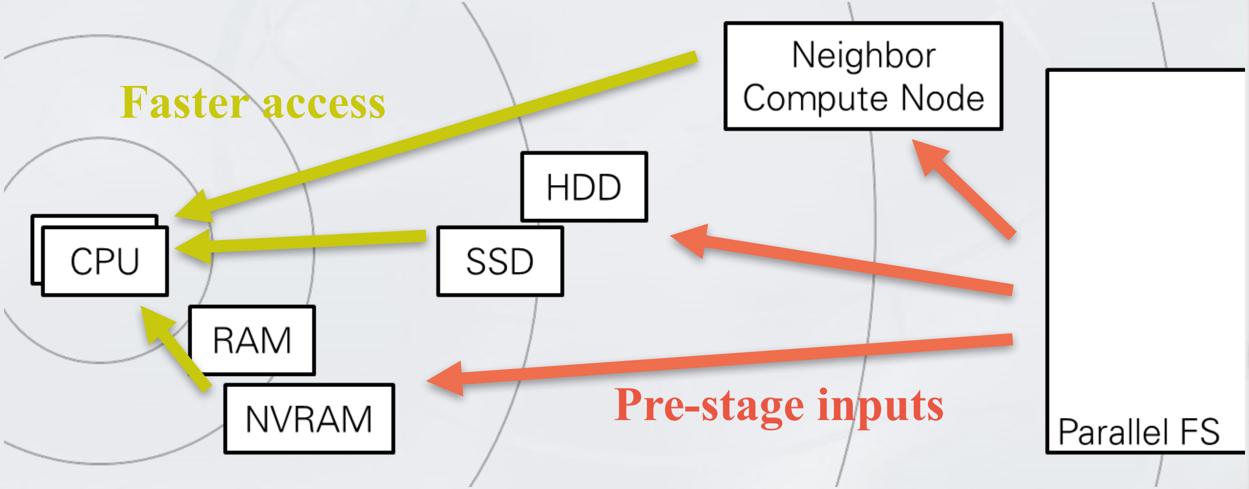
Project Rationale

I/O Challenges at Exascale

- I/O subsystem is the slowest system to access in a HPC machine
- Shared medium: no reliable bandwidth, no good transfer time predictions
- Upcoming architectures with "fat nodes" and intermediate local storages

Goal: optimize I/O

- Using additional storages
- Transparent solution for parallel applications
- Pre-stage inputs early, cache outputs



Proposed Solution

- Ad-hoc overlay file system
  - Separate overlay file system per application run
  - Instantiated on the scheduled compute nodes
  - Lives longer than the users’ job
- Central I/O planner
  - Global Planning of I/O including stage-in/-out of data, for all par. jobs
  - Optimization of data placement in the ad-hoc file system (resp. nodes)
  - Integration with systems batch scheduler
- Application monitoring, resource discovery
  - I/O behavior, machine-specific storage types, sizes, speeds, …

Ad-hoc overlay file system

Research Goals

- Relax POSIX semantics based on access patterns
- No locking
- Distributed Metadata
- Eventual consistency
- Monitoring
- Make applications responsible for their I/O

Related Work

- GPFS, Lustre, BeeGFS,…
- Key-value stores for metadata
- DeltaFS, BurstFS, …

Status

- Design phase for scalable metadata and lock free block storage
- Evaluation of different storage schemata

Central I/O Planner

Research Goals

- Stage in and stage out of data
- Maybe even during job runtime
- Schedule I/O based on estimations from the running/planned jobs

Related Work

- Current batch systems, Data Staging from Grid Environments
- Workpool/Workspace concepts
- I/O scheduling and QoS approaches

Status

- Prototype for a temporary file system based on BeeGFS
- Stage in and stage out based on parallel copy tools
- SLURM integration

Karlsruher Institut für Technologie

Resource Discovery and Monitoring

Research Goals

- Collect available resources
- Monitor FS activities
- Provide planner with estimations about I/O capabilities and current usage
- Learn I/O behavior for standard applications

Related Work

- OpenMPI
- Likwid
- Many data collection tools
- I/O pattern recognition

Status

- Working prototype that discovers node and connection details
- Working on integration into I/O planner

TECHNISCHE
UNIVERSITÄT
DRESDEN