Exploiting the Heterogeneous Storage Landscape in a Data Center

Julian M. Kunkel

Julian.Kunkel@googlemail.com

CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

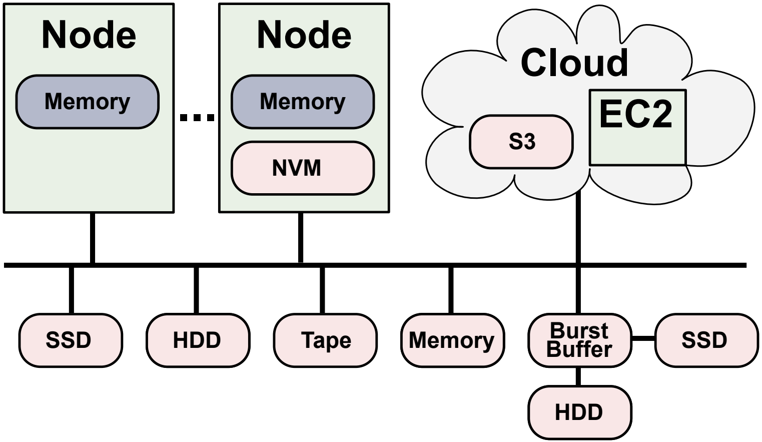
Per3S Workshop

2018-01-12

Outline

Disclaimer: This material reflects only the author's view and the EU-Commission is not responsible for any use that may be made of the information it contains

Heterogeneous Storage Landscape in Future Data Centers



HPC system with compute nodes and storage

Status Quo

Storage Systems for HPC

- Data (Files) are transferred to/from compute nodes
- Naive data management with tiering $\Rightarrow$ copy data between tiers
- Data life cycle and workflow management with simple methods
- Fault tolerance is an issue in most programming model

Big Data

- Compute and storage capabilities are tightly coupled
- Move compute to data (efficient due to lightweight compute) $\Rightarrow$ Active storage
- Programming models are fault tolerant
- Tools/Programms support different file formats interchangeably

Performance Obstacles to Exploit Heterogeneous Storage

Semantical Gap of Data Access

- Access of files and objects that are just an array of Bytes
- Hierarchical namespace
- Consistency semantics
- Applications work with (semi)structured data
- Storage system does not understand data structures and usage patters

Strict Separation of Compute and I/O

...

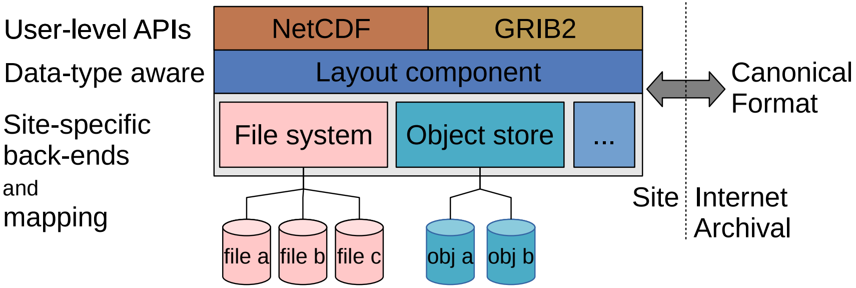
Storage Stack Lacks Performance Understanding

...

Approach of the Earth-System Data Middleware (in ESiWACE)

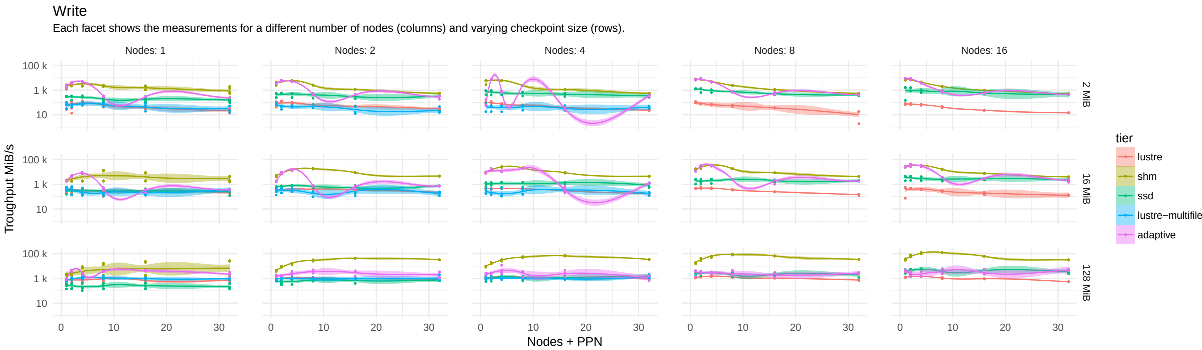
One Key Concepts: Storage layout is optimized to data center storage

- Site-specific (optimized) data layout schemes
  - Based on site-configuration & *limited* performance model
  - Flexible mapping of data to multiple storage backends / storage systems



First Results with POSIX Backend using Multiple Storage Systems

- Depending on data volume, it chooses the storage system dynamically



Adaptive Tier Selection for HDF5/NetCDF without requiring changes to existing applications. (SC17 Research Poster).

Proposed Approach

- The **standardization** of a high-level *data model & interface*
  - Targeting data intensive and HPC workloads
  - Lifting semantic access to a new level
- Development of a reference implementation of a **smart runtime system**
  - Implementing key features
- Demonstration of benefits on relevant data-intense scientific applications

The Structured Data Model (Interface) SDMI: Key features

- High-level data model for HPC
  - Storage understands data structures vs. byte array
  - Relaxed consistency
- Semantic namespace
  - Organize based on domain-specific metadata (instead of hierarchical)
  - Support domain-specific operations and addressing schemes
- Integrated processing capabilities
  - Offload data-intensive compute to storage system
  - In-situ/In-transit workflows
- Workflow management
  - Manage data-driven workflows, support services
- Performance-portability
  - Intents vs. technical hints
  - Guided interfaces
- Enhanced data management features
  - Embedded performance analysis
  - Resilience, import/export, ...

Smart Runtime Prototype Key Features

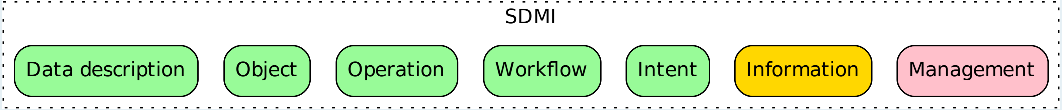
- Semantic access
  - Search and access based on metadata
- Self-aware
  - Understand performance characteristics
- Automatic layouting + smart data replication
  - Across multiple storage systems
  - Adapt data layout during runtime
- Managed workflows
  - Offloading of I/O intense kernels to storage
  - Scheduler considers compute and I/O requirements
- Compatibility
  - Enable access to legacy applications (with performance loss)

Towards a Governance Body

Development of the data model and interfaces

- Establishing a Forum similarly to MPI
- Define data model for HPC
  - Must be beneficial for Big Data + Desktop, too
- Open board: encourage community collaboration
- **You are welcome to participate, just contact me**

Simplified Draft APIs



The ESiWACE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No **675191**

Disclaimer: This material reflects only the author's view and the EU-Commission is not responsible for any use that may be made of the information it contains

ESDM Architecture: Detailed View