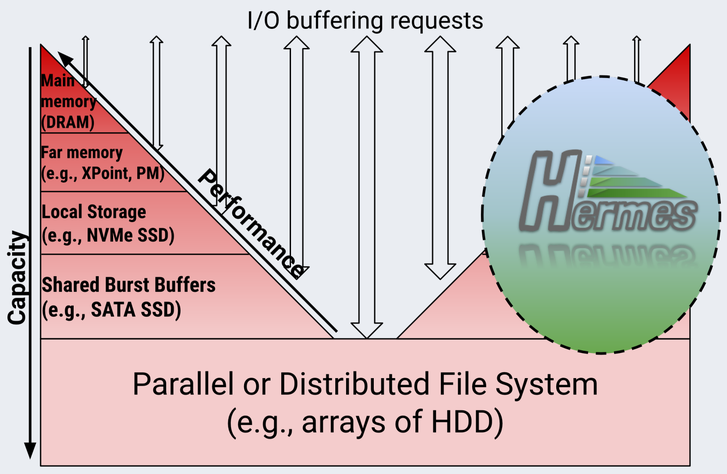
Framework: Software: NSCI: Collaborative Research: Hermes: Extending the HDF Library to Support Intelligent I/O Buffering for Deep Memory and Storage Hierarchy Systems

Xian-He Sun, Illinois Institute of Technology, Elena Pourmal, the HDP Group, Jian Peng, UIUC

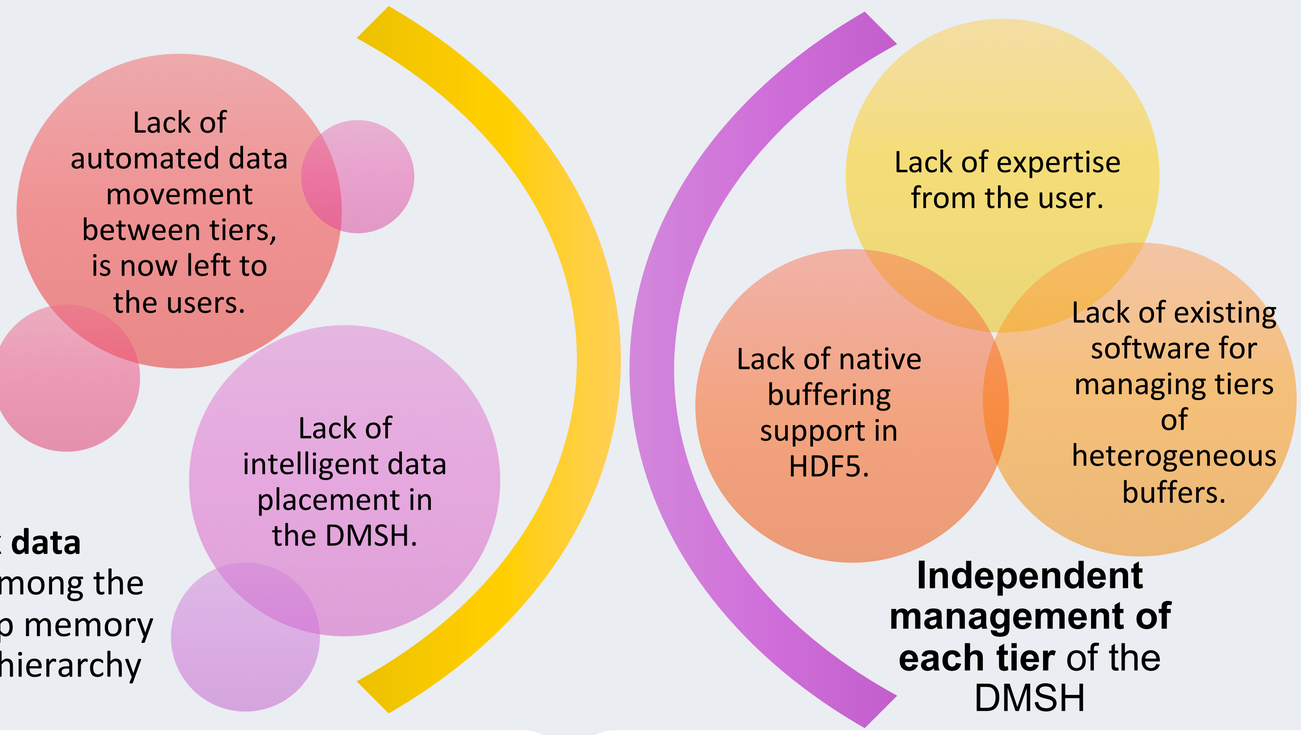
Multi-Tiered Storage



- New storage system designs incorporate non-volatile burst buffers between the main memory and the disks.
- HPC hierarchical storage systems with burst buffers (BB) have been installed at several HPC sites.
- Multiple levels of memory and storage in a hierarchy, called **DMSH**.

Synopsis

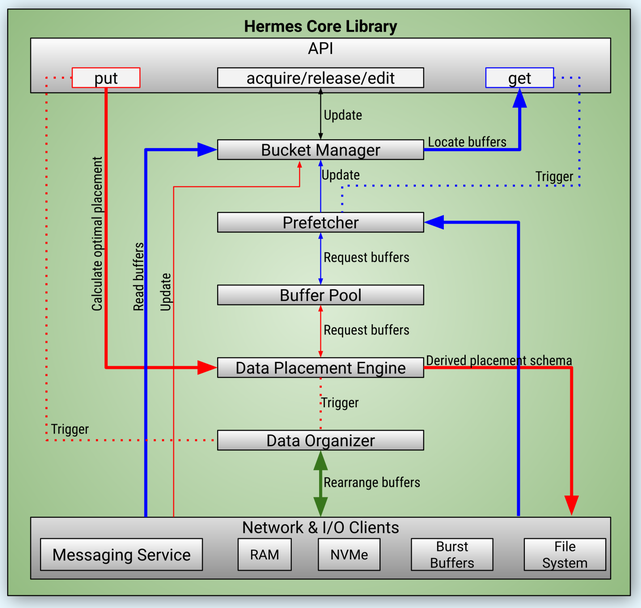
Current Situation



Complex data placement among the tiers of a deep memory and storage hierarchy

Independent management of each tier of the DMSH

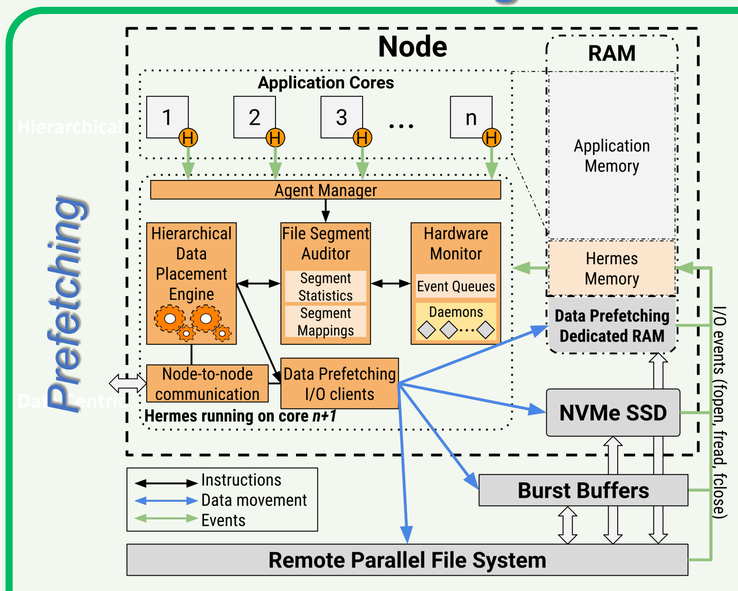
Overview



- **Hermes API:**
  - intercepts all I/O calls from the applications.
  - calculates the operations to be carried out in case of an active buffering scenario.
- **Hermes Data Placement Engine (DPE)**
  - calculates the data destination, i.e., where in the hierarchy should the data be placed.
  - uses various data placement policies.
- **Hermes Data Organizer**
  - event-based component
  - carries out all data movements
    - E.g., for prefetching reasons, evictions, lack of space, or hotness of data etc.
- **Metadata Manager**
  - maintains two types of metadata:
    - user's metadata operations (e.g., files, directories, permissions etc.),
    - Hermes library's internal metadata (e.g., locations of all buffered data and internal temporary files that contain user files).
- **Messaging Service**
  - enables horizontal buffering
  - provides an infrastructure to pass instructions to other nodes to perform operations on data or facilitate its movement
- **Cache Manager**
  - handles all buffers inside Hermes
  - equipped with several data replacement policies

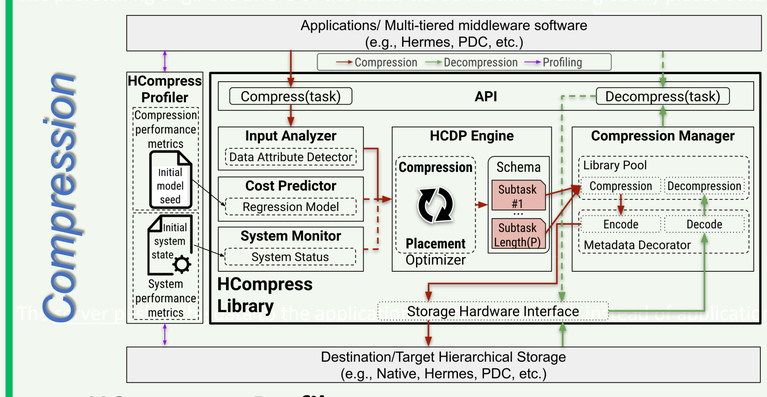
Current Progress

Prefetching



- **Server-Push**
  - Event are captured by kernel's inotify utility
  - Prefetched data is push to the hierarchy
- **Data Centric (Score Incorporates)**
  - Recency, Frequency, and Sequencing
- **Hierarchical Placement**
  - The engine calculates placement of prefetch data based on multi-tiered storage and data characteristics.

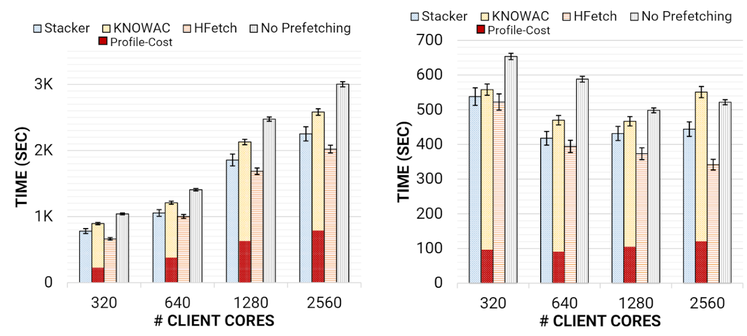
Compression



- **HCompress Profiler**
  - Runs a exhaustive benchmark to capture system and compression characteristics for predefined and user-defined inputs.
- **Compression Cost Predictor**
  - Uses linear regression model to predict the Expected Compression Cost for the engine.
  - Uses reinforcement learning to improve accuracy for dynamic inputs.
- **HCDP Engine**
  - Employs a dynamic programming (DP) algorithm against three dimensions
  - Data characteristics, Compression libraries, and Storage tiers

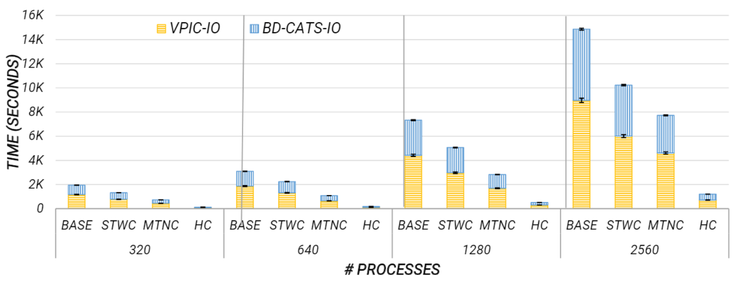
Results

Hierarchical Prefetching



(a) Montage (weak scaling) (b) WRF (strong scaling)

Hierarchical Compression



VPIC-IO plus BD-CATS-IO workflow with 10 timesteps.

Ongoing

Hermes Container Library (HCL)

- **High Performance**
  - Utilizes RPC over RDMA design to build high-level functions.
  - Utilizes network aggregation and caching techniques.
- **Flexible**
  - STL like interface is easy to use and program.
- **Persistent**
  - Utilizes modern NVRAM and NVMe to build persistent data structures.
- **Open Source**
  - https://github.com/HDFGroup/hcl

Contact

Xian-He Sun, PI
sun@iit.edu
www.cs.iit.edu/~scs

Anthony Kougkas, Lead
akougkas@iit.edu

Find more