Hybrid Hierarchical Data Management System: Accelerating Data Processing on HPC Systems

HPC Systems : Compute-intensive Applications

Big Data: More Data-intensive Applications

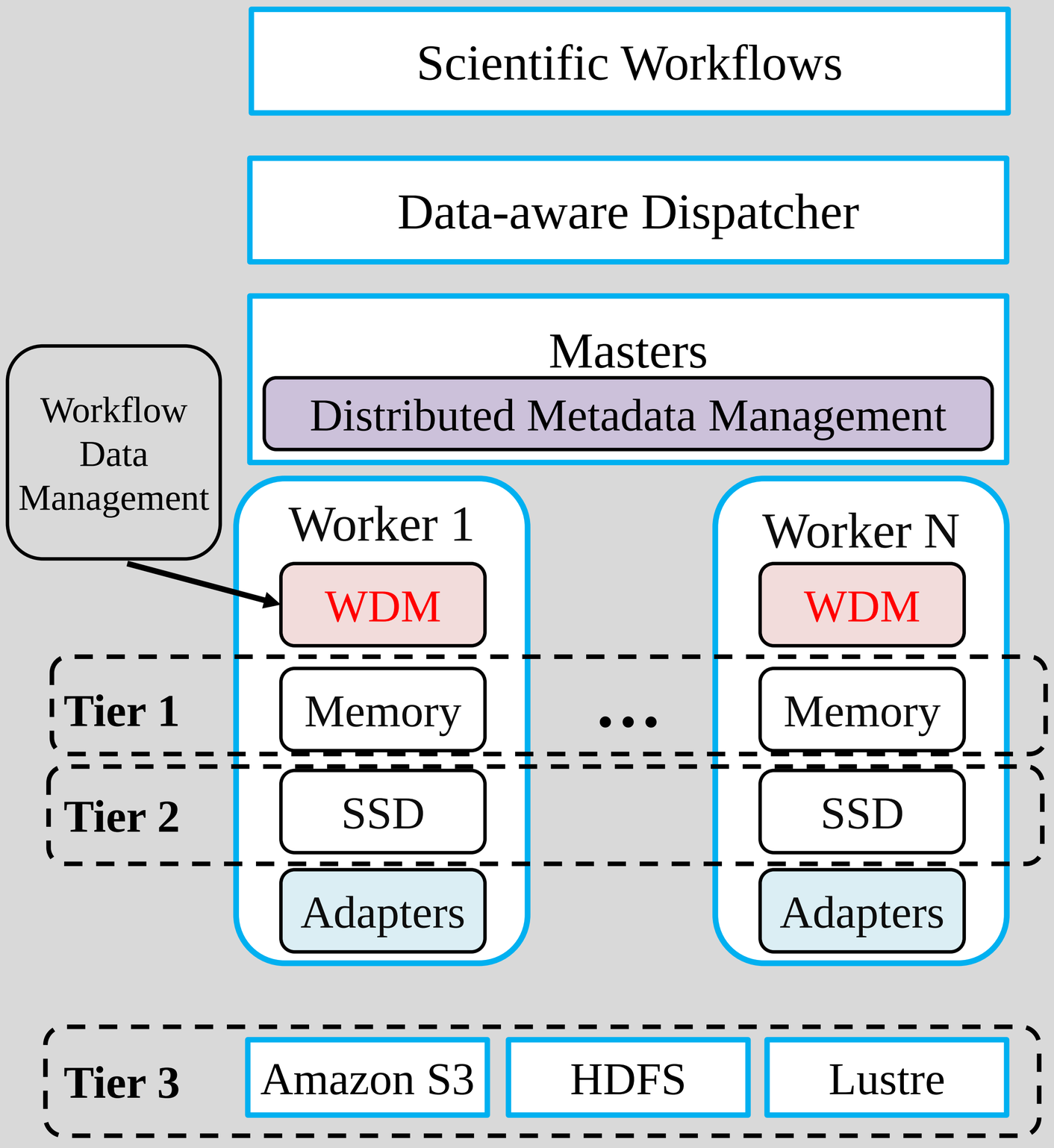
HPC + Big Data: Challenges in Data Management

- Multiple Data Sources
- Heterogeneous Storage Devices
- Bottleneck in Metadata Management

Existing Data Management Systems :

SRB, iRODS: Data Sharing
Qserv, FastQuery: Data Querying
SciDB: Data Processing(Array Data)

How to Accelerate Data Processing on HPC Systems ?



Hybrid Hierarchical Data Management System Overview

Tiered Data Management

- Cross-Tier Data Sharing
- Customizing Data Management Strategies Based on Workflow Data Access Patterns
- Unified Data Access Interface

Data-aware Job Scheduling

- Move the Compute to the Data
- Dynamic Task Scheduling
- Storage Tier Aware + Data Access Pattern

Distributed Metadata Management

- KV-Based Metadata Management
- Scalable Metadata Server
- Hash+SubTree Namespace Management

Domain Specific Optimizations

- Optimizations for Hadoop/Spark shuffle
- Optimizations for lots of small files
- …