HPSS 8 Metadata Services Evolution to Meet the Demands of Extreme Scale Computing

David Boomer, Kayla Broussard, Michael Meseke
IBM

Introduction

- The Extreme Scale era is upon us…..

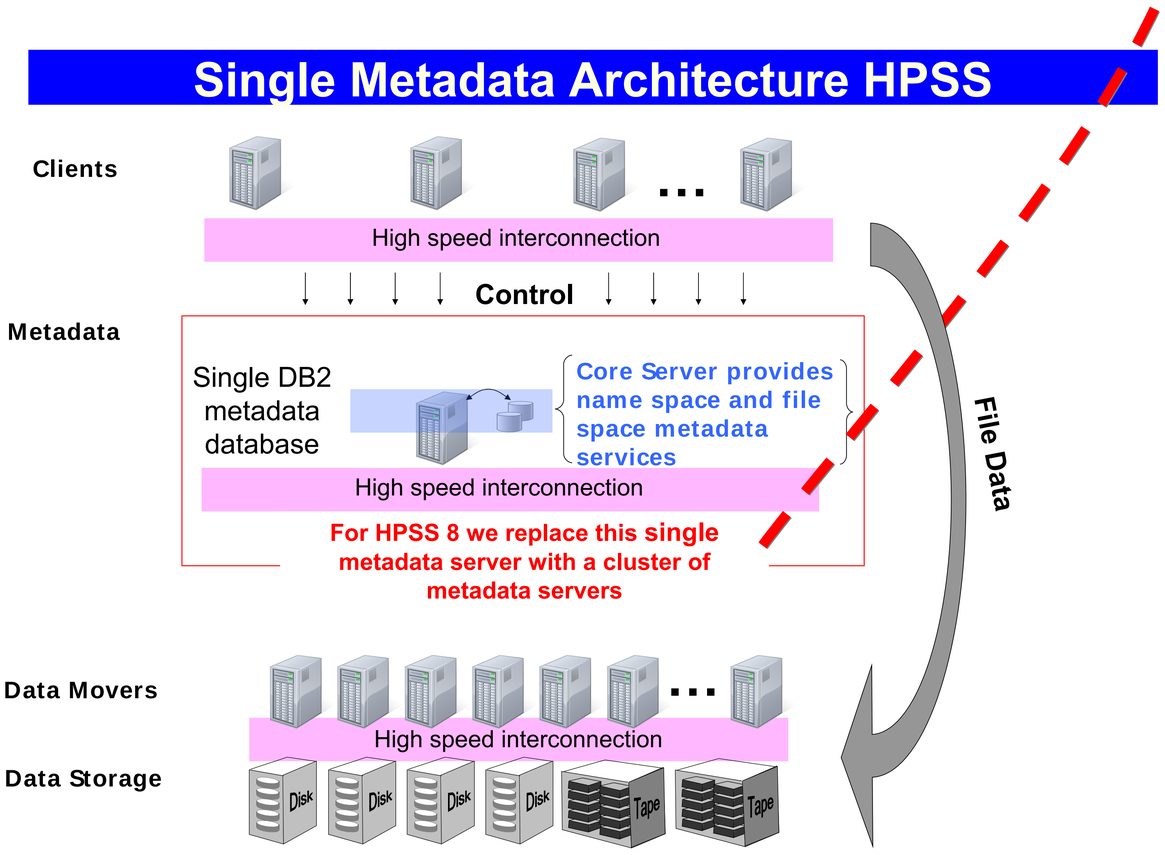
  { petabytes of main memory + hundreds of thousands processors And millions of cores } = { 100s of billions to trillions of files, exabytes of data, **petabytes of metadata** and tens of thousands of storage devices }

- High Performance Storage System (HPSS)
  - manages petabytes of data on hundreds of disks, tapes, and robotic tape libraries
  - provides highly scalable hierarchical storage management that keeps recently used data on disk and less recently used data on tape
  - uses cluster, LAN, WAN and/or SAN technology to aggregate the capacity and performance of many computers, disks, and tape drives into a single virtual file system of exceptional size and versatility
- The HPSS network centric architecture provides an extremely scalable I/O infrastructure necessary to handle the throughput

The Challenge

- Exascale computing presents data throughput and metadata management challenges
- HPSS metadata management services are limited to the capability of a single non-partitioned database engine running on a single computer system
  - Significant bottleneck when dealing with trillions of files
  - Each file HPSS manages requires a distinct set of database transactions, regardless of size (more files = more database transactions)
- Eventually the pace of ingest will surpass the single database engine capacity
  - Fixed per file metadata transaction overhead directly impacts the aggregate throughput

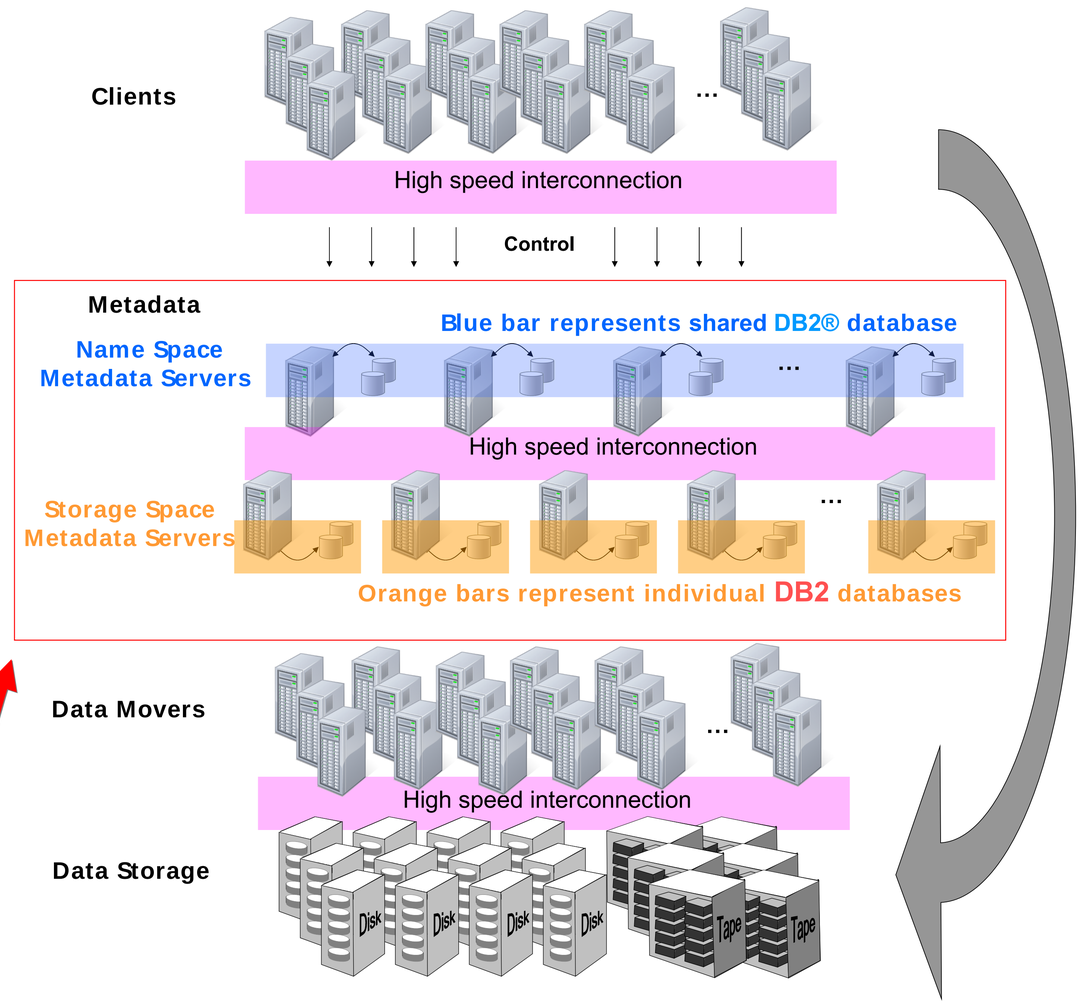
Single Metadata Architecture HPSS



The Solution

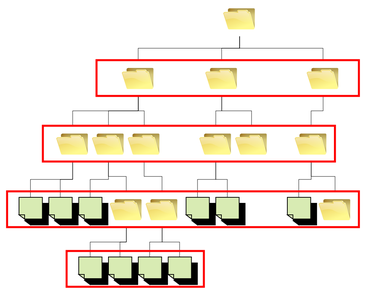
- Extreme Scale HPSS includes
  - A distributed POSIX namespace via multiple metadata servers (NS MDS)
  - Multiple storage space metadata servers (SS MDS)
- DB2 - IBM's enterprise class relational database – is the central component
  - DB2's Data Partitioning Feature provides the necessary infrastructure to distribute HPSS metadata
  - The partitioning feature is based on a share nothing architecture, where each system manages the local partition, but has access to all partitions transparently.
  - **DB2 is extremely well tested and supported by IBM; HPSS development takes advantage of this mature and robust capability**
- Provides a scalable metadata services infrastructure allowing additional NS MDS and SS MDS nodes to be added with minimal downtime and restructuring
- The end result creates a metadata services layer that aggregates the capacity, resources and performance of many computers into a single logical metadata repository

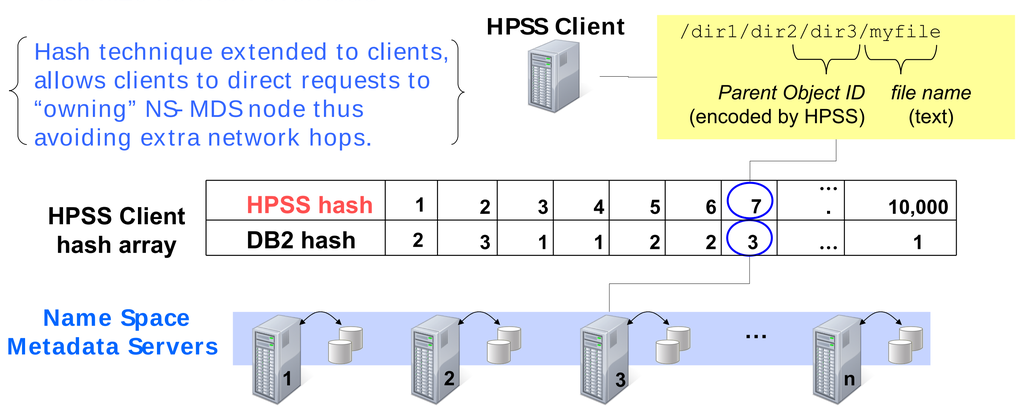
Distributed Metadata Architecture HPSS 8



Distributed Namespace Metadata Challenges

- Distributed namespaces are well researched with many techniques, each with various strengths and weaknesses
- Distribution of HPSS file metadata and client references over multiple metadata servers is accomplished using a hash technique based on directory id and file name
  - Hash results in a predictable range of values
  - Easily integrated with DB2 partitioning feature
  - Easily integrated with HPSS client
  - Consistent interface for determining metadata location
- HPSS V8 hash technique and DB2 functionality combine to:
  - Maintain POSIX naming rules
  - Provide balanced and self-leveling metadata distribution
  - Co-locate related metadata items
  - Minimize directory hotspots
  - Minimize network overhead





Prototype Results

- Extreme Scale HPSS 8 distributed metadata services
  - Single NS-MDS node and 8 SS-MDS nodes provides **10x** performance of HPSS V7 single metadata server architecture
  - Additional metadata nodes provide additional metadata transaction capacity
- Significant growth capability expected beyond prototype configuration

Measurement / HPSS Version	File Creates per Second	Database Transactions per Second	Scalability
HPSS V7 1 Metadata Server Node	800	3,200	1x
HPSS V8 1 NS-MDS Node 8 SS-MDS Nodes	8,000	32,000	10x
HPSS V8 2 NS-MDS Nodes 16 SS-MDS Nodes	16,000	64,000	20x