Large Scale Test of a storage solution based on an Industry Standard

Michael Ernst
Brookhaven National Laboratory

ADC Retreat
Naples, Italy
February 2, 2011

Motivation

- Though dCache at BNL supports distributed analysis well (up to 5000 concurrent analysis jobs) we are looking into ways to improve the usability of our environment, e.g. include large-scale interactive analysis.
- We want to explore to what extent we can build on industrial products as far as the facility part is concerned w/o having to rely on community extensions and interfaces.

➢ NFS 4.1 (pNFS) appealing because of performance, simplicity, level of integration w/ OS

➢ BlueArc successfully used by PHENIX and STAR

Areas of Interest

Benefits of Parallel I/O

- Delivers Very High Application Performance
- Allows for Massive Scalability without diminished performance

Benefits of NFS (or most any standard)

- Ensures Interoperability among vendor solutions
- Allows Choice of best-of-breed products
- Eliminates Risks of deploying proprietary technology

Whys is NFS4.1(pNFS) attractive?

Simplicity

- Regular mount-point and real POSIX I/O
- Can be used by unmodified applications (e.g. Mathematica..)

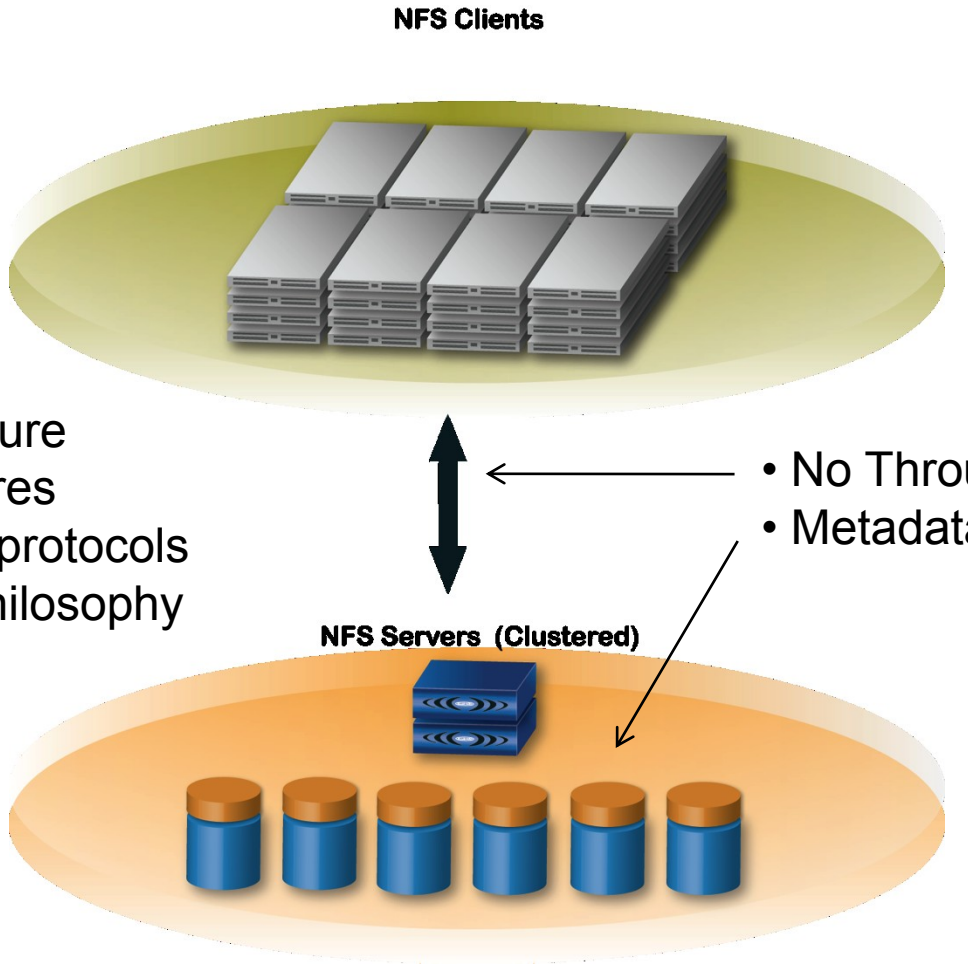
Cost reduction for software providers and Infrastructures

- Less components to maintain and deploy
  - Data clients provided by the OS
  - No additional server needed
- Smart caching (block caching) development done by OS vendors
- Only single FS driver in ROOT

BlueArc System Performance

HNAS 3200, AMS 2500	300GB x 144 HDDs, 8D+1P	36 Disks	72 Disks
Sequential Read %	I/O Block Size	Throughput in MB/s	Throughput in MB/s
100	32KB	529	934
100	256KB	557	919
100	1024KB	574	923
75	32KB	572	806
75	256KB	552	764
75	1024KB	545	763
50	32KB	583	819
50	256KB	523	718
50	1024KB	530	722
25	32KB	567	857
25	256KB	528	752
25	1024KB	520	781
0	32KB	611	894
0	256KB	593	818
0	1024KB	566	796

Traditional Network File System

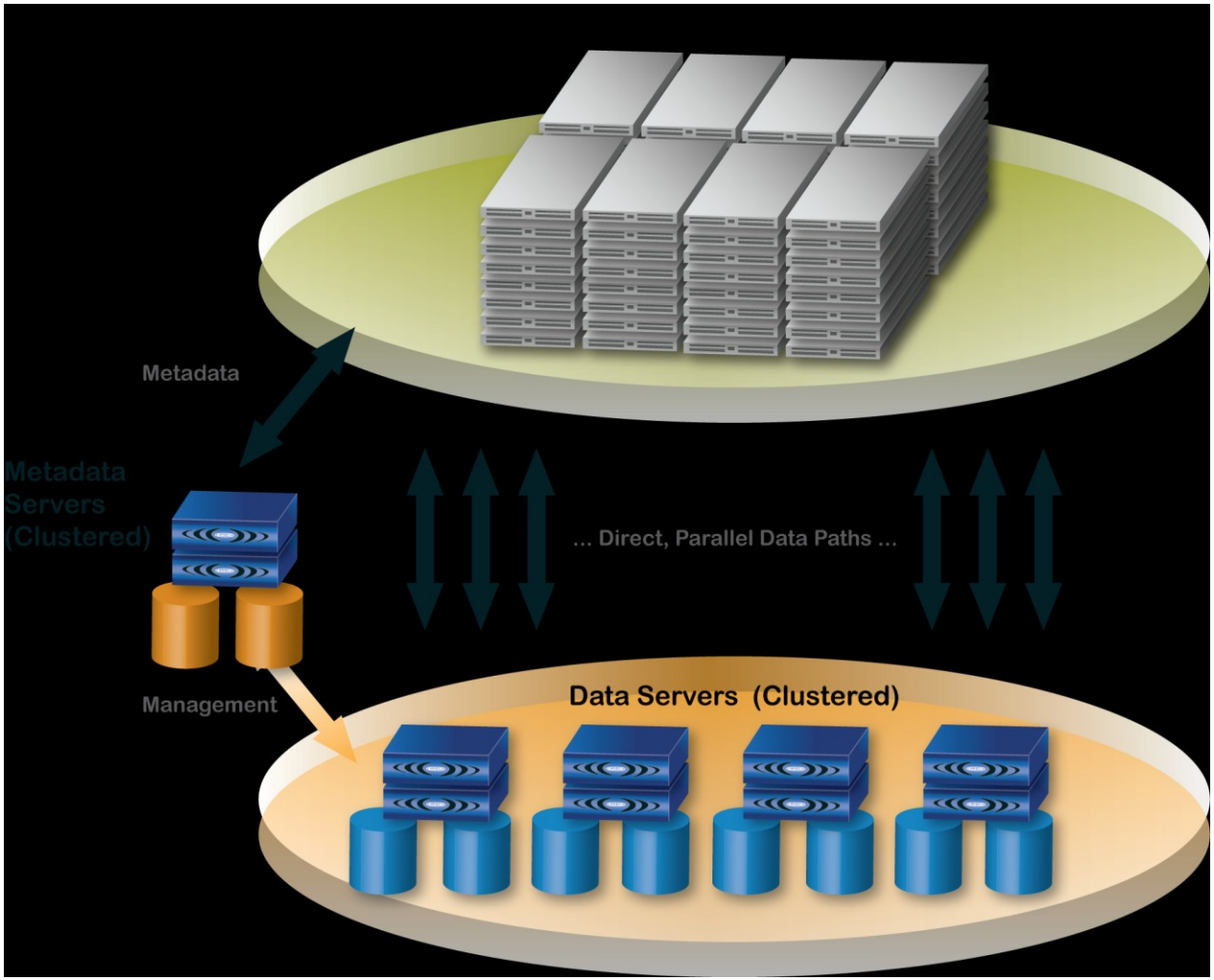


- Proven architecture
- Enterprise features
- Open, standard protocols
- Open storage philosophy

BUT

- No Throughput Aggregation
- Metadata and data together

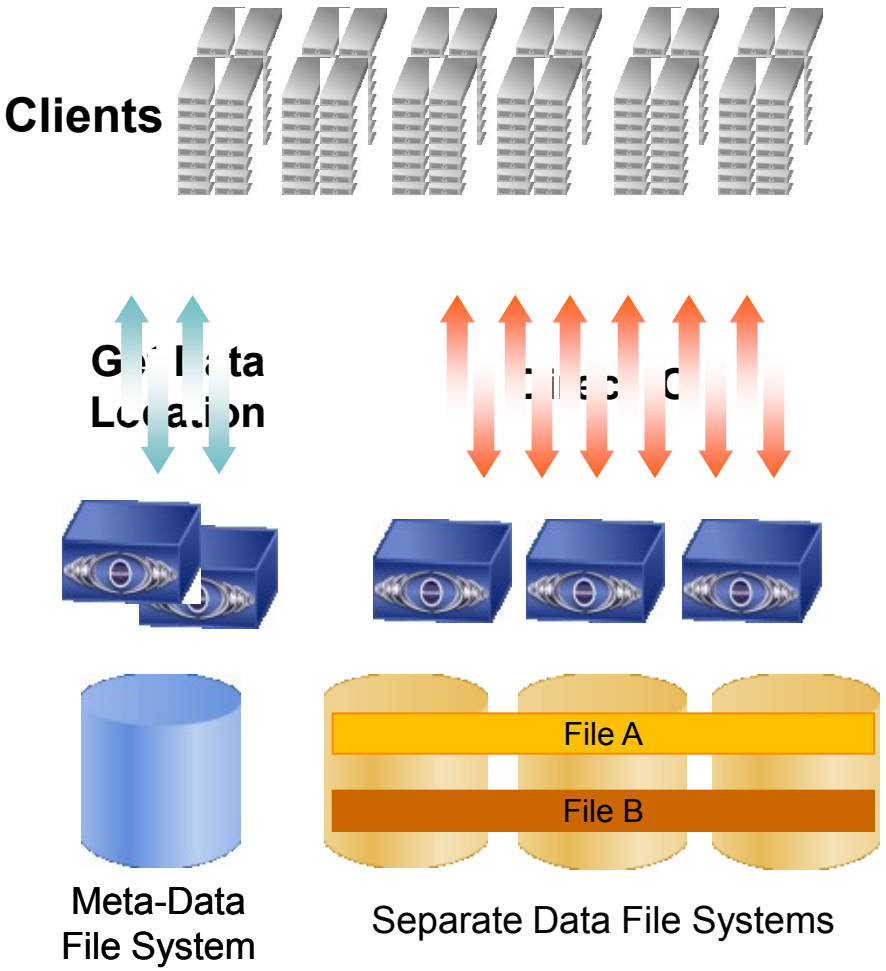
pNFS Architecture



BlueArc pNFS Based NAS Platform

- Highly scalable single metadata server
  - Clustered for HA
  - Architecture supports multiple clustered metadata servers
- Support for heterogeneous data servers
- Built on Enterprise Class Platform
  - Reliability
  - Full Featured NAS – Quota, Snapshots etc.

pNFS for HPC Performance Scaling



- Data files automatically spread across multiple data FS for load balancing and performance
- Individual data files optionally striped across multiple data FS for performance
- Extreme Flexibility: A single server can run the entire cluster (acting as both metadata server and data mover)
- Can scale performance by adding data movers (and relocate the FS)
- Can scale capacity by adding data file systems

Product Availability and Test Bed at BNL

- BA's pNFS – Mid 2011
  - Up to 40 data movers with up to 8PB capacity each – growing to 128 data movers
  - Mercury - High performance metadata server
  - BlueArc Mercury or BlueArc Linux-based data mover appliances
- Performance Goals for Test Bed at BNL
  - Will become DDM endpoint w/ SRM, GridFTP, Xrootd
  - Up to 2k concurrent Analysis Jobs
  - Up to 4 GB/s throughput between storage system and clients
  - Aiming at having ~500 TB of disk space

Test Bed at BNL