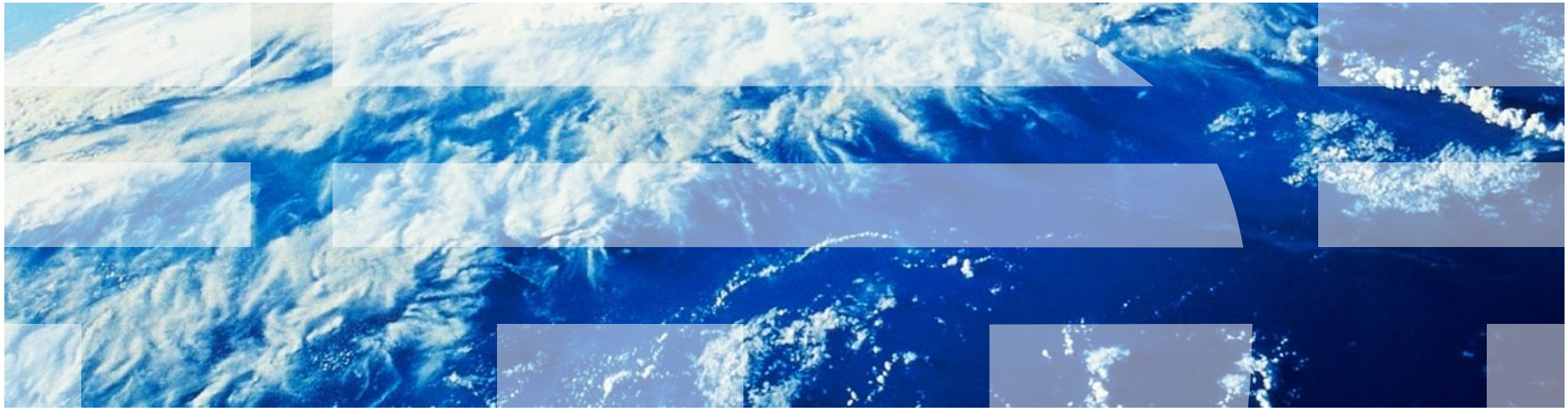
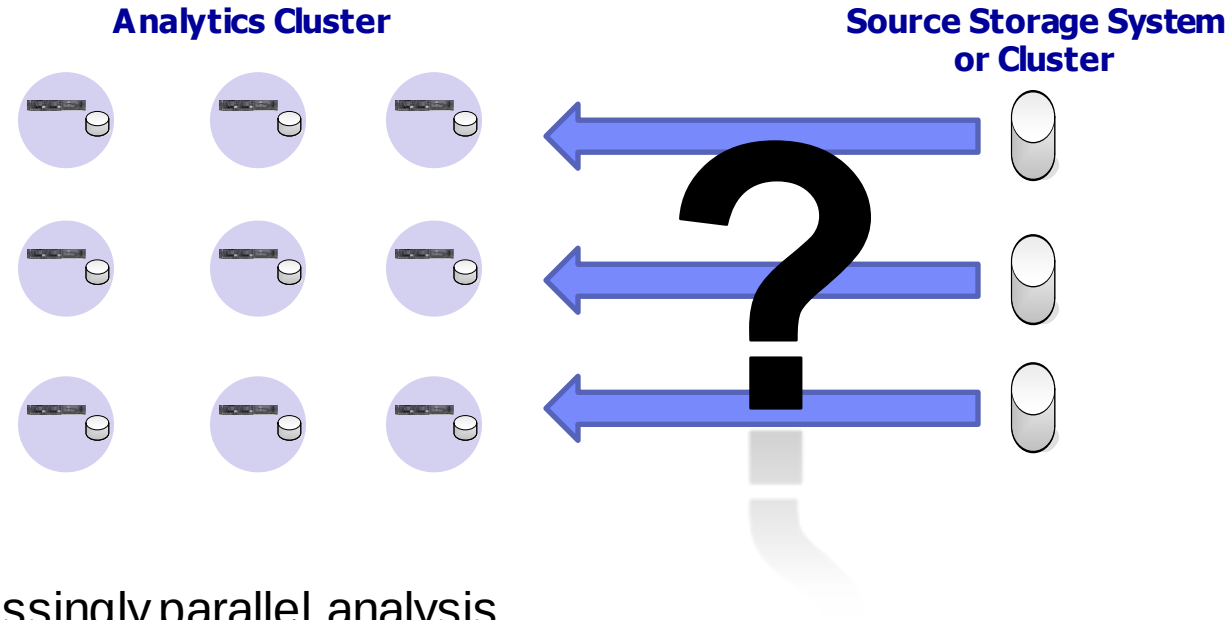
Scalable Data Transfer In and Out of Analytics Clusters

Dean Hildebrand (dhildeb@us.ibm.com)
Prasenjit Sarkar (psarkar@almaden.ibm.com)



GFP03051-USEN-00

Analytics Needs Data



- Embarrassingly parallel analysis
  - Hadoop, etc
- Analytics clusters typically use a local storage file system
  - Datasets may not be stored in analytics cluster
    - Legacy storage system
    - Existing storage cluster
    - Production storage system with backup/disaster recovery support
- Before data can be analyzed, it must be placed on the local storage of each node
  - Typically multiple copies of each data block for better performance

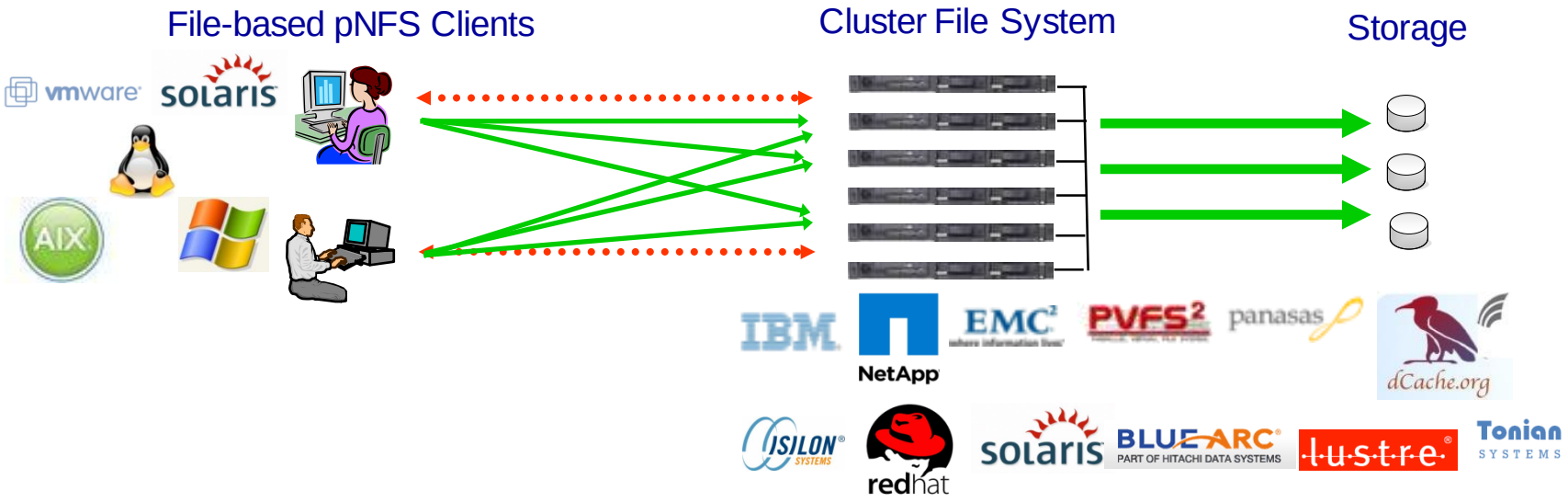
Analytics Needs Efficient Access to Data

- Data ingest must be done to optimize I/O performance during analysis
  - Efficient data placement on analytic cluster nodes
  - Efficient data access from source storage system
- Variety of data placement considerations:
  - Application/Workload – read/write, block size, etc.
  - Bandwidth between nodes
  - Number of nodes for analytics job
  - Capabilities of nodes, e.g., GPUs, etc
  - Tradeoff between time to ingest and ideal layout
- Heterogeneous storage systems
- Overview of rest of talk
  - Survey of ingest mechanisms and local storage cluster file systems
  - Outline of research challenges

Standard Ingest Mechanisms

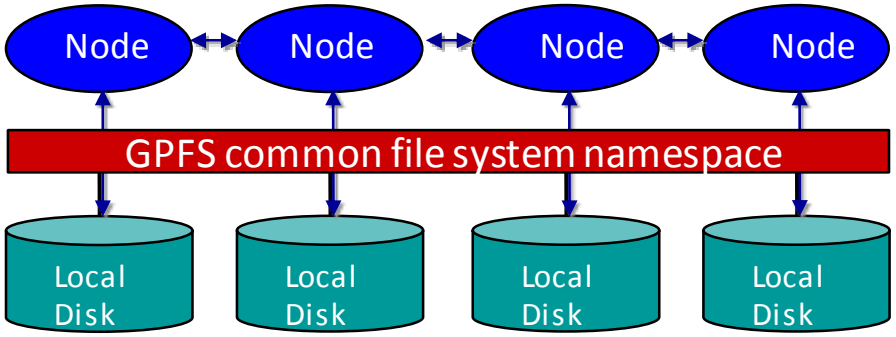
- FTP/HTTP/ NFS/CIFS/SMB2
  - No data location awareness
  - Serial protocols
- GridFTP
  - Parallel protocol
  - Complex
  - Not universally supported
- pNFS
  - Standard
  - Location awareness
    - Can perform direct read and write to specific storage nodes
    - Targeted placement of data on target cluster
  - Parallel protocol

pNFS: In Depth



- Standard, secure, and scalable access to data
  - Separates namespace (metadata) from data
  - Direct and parallel data access
  - Scale with underlying file system
  - Improve individual and aggregate client performance
- pNFS client (file layout) available in RHEL 6.2 and SLES 11 SP2
- pNFS server implementations
  - Linux Ganesha (user-level), Linux kernel, Solaris

Cluster File Systems: GPFS-SNC



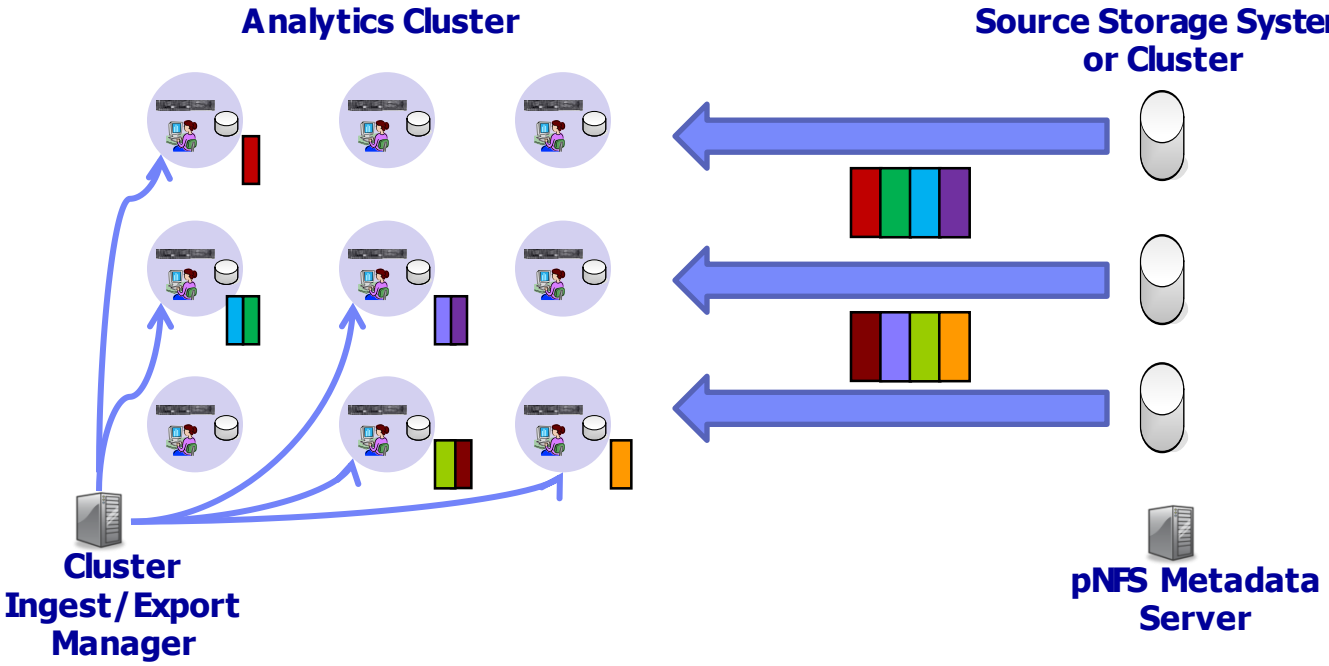
- Requirements on cluster file systems:
  - Scalable data and metadata
    - pNFS client can mount and retrieve layout from any node
    - Metadata requests load balanced across cluster
    - Direct data access from any node
- Candidate File System: GPFS
  - *Cluster*: thousands of nodes, fast reliable communication, common admin domain.
  - *Shared disk*: all data and metadata on disk accessible from any node, coordinated by distributed lock service.
  - *Parallel*: data *and metadata* flow to/from all nodes from/to all disks in parallel; files striped across all disks
  - *Policy Engine*: allows programmatic (SQL) access to file system attributes and configuration actions
- Extensions for Analytics: GPFS-SNC
  - *Locality:* Expose data layout
  - *Write Affinity:* Allow applications to determine file placement
  - *Metablocks*: Tune block size for applications
  - *Pipelined replication*: Effective use of network b/w particularly cross-rack
  - *Distributed recovery:* Failure should not affect performance

Useful for ingest and export

Analytic Cluster Ingest Research Challenges

- Data ingest and export considerations
  - Application's desired distribution
    - How can the application specify its desired distribution
    - How can any type of distribution be supported
    - How to translate large directory trees into desired distribution
  - Network bandwidth between analytics cluster and source storage may be asymmetrical
    - Specific nodes may have higher bandwidth than others (10gigE gateway servers)
    - Need strategy to distribute data assuming some nodes have higher bandwidth
    - Within analytics cluster there may exist a hierarchical network topology
  - Number of nodes that should participate in data transfer
    - May change due to failures
- Who should instigate data transfer
  - Cluster ingest/export manager using cluster pNFS clients
  - External manager and external pNFS nodes

Putting it All Together: Analytic Cluster Data Ingest Example



- Example Analytics cluster needs to ingest a 1TB dataset
  - 1000s nodes
  - Each node contains
    - pNFS client and server
    - GPFS-SNC file system client
    - Locally attached storage
- Cluster Ingest/Export Manager directs each cluster node to read in portion of dataset
- Each pNFS client requests a layout for each file it will ingest and writes data locally
  - Direct source data access