Building a CIFS/NFS Gateway to Lustre®

Seagate

Chris Gouge
Software Architect
Seagate Cloud Systems and Solutions

LAD 14 Reims France 2014
LUSTRE ADMINS & DEVELOPERS WORKSHOP

Topic Index

- Design Context
- Timeline
- Requirements
  - Goals
  - Workflows
- Software
- Hardware
- CTDB
- User Security
- Manageability & Monitoring

Design Context

ClusterStor Engineered Solution

- Wide applicability of solution
  - Not per site / customer bespoke solutions
  - Use of open source
  - Use of proven, mature HPC tools & IT tools
- Add value
  - Reduced deployment times, Ease of use
  - Service & support
- Participation in the community
  - Significant contributions to Lustre source tree
  - This talk & others on best practices

Timeline

Samba and CTDB Integration with ClusterStor

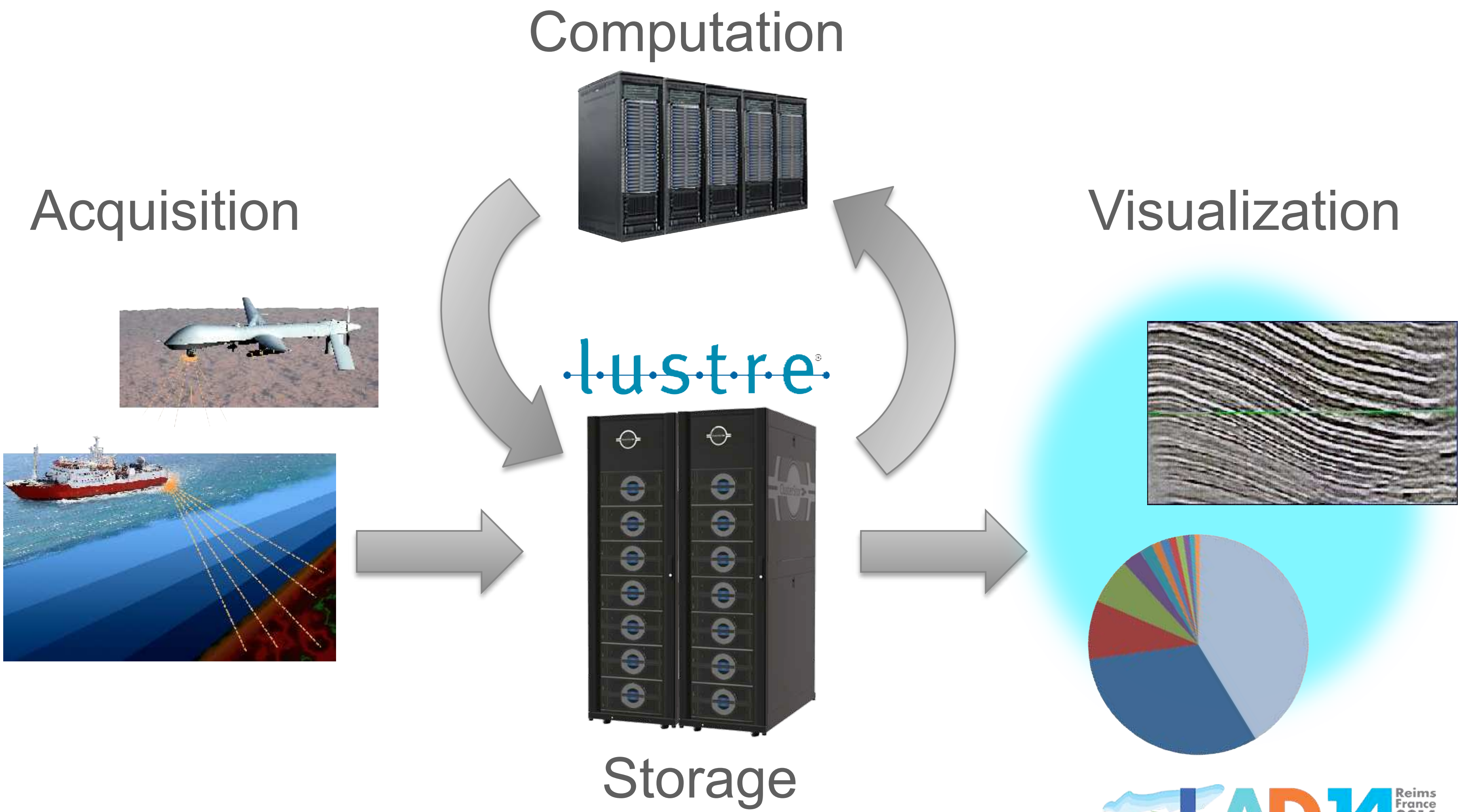
- 2011: Demonstration of Concept, ISC
- 2011-2013: xattr cache in Lustre
  - » LU549 / LU-2869
- 2011: Samba using mandatory locks
  - » LU-1073
- 2012-2014: NFS issues with async flocks
  - » LU-2525 (pending)
- 2012-2014: Integration with Unified Management
- 2013-2014: Updates to Clustering & User Security
- 2014: General Availability

Requirements Analysis

- Supported Workflows
- Architectural Goals

Workflow #1

CIFS or NFS access required for Data Visualization



Computation

Acquisition

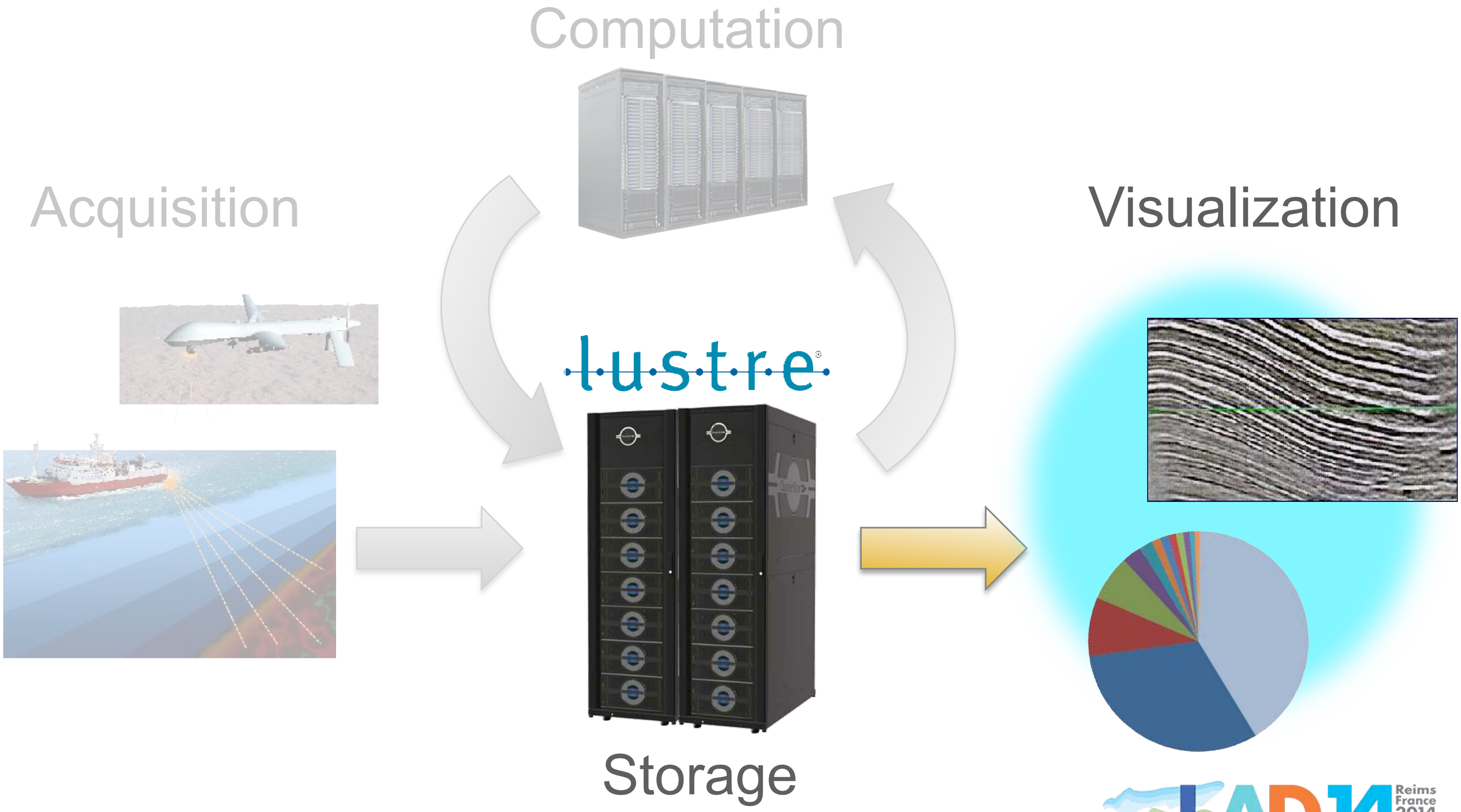
Visualization

lustre

Storage

Workflow #1

CIFS or NFS access required for Data Visualization



Computation

Acquisition

Visualization

lustre

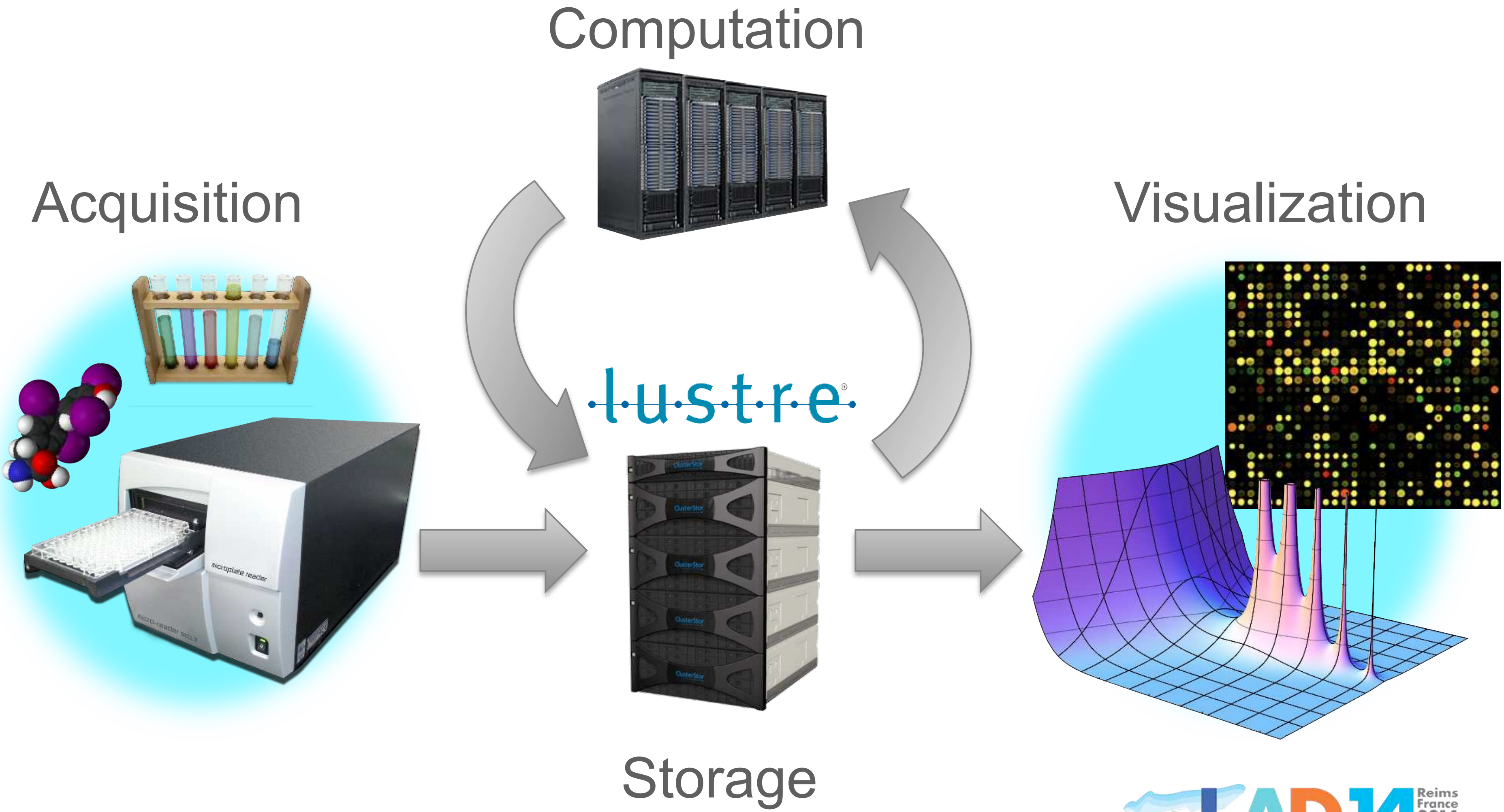
Storage

Workflows

1. Data Visualization
   - Smaller data sets, results of HPC jobs
   - Non-Lustre platform
   - Dozens (not hundreds/thousands) of clients

Workflow #2

CIFS or NFS access required for Data Acquisition



Computation

Acquisition

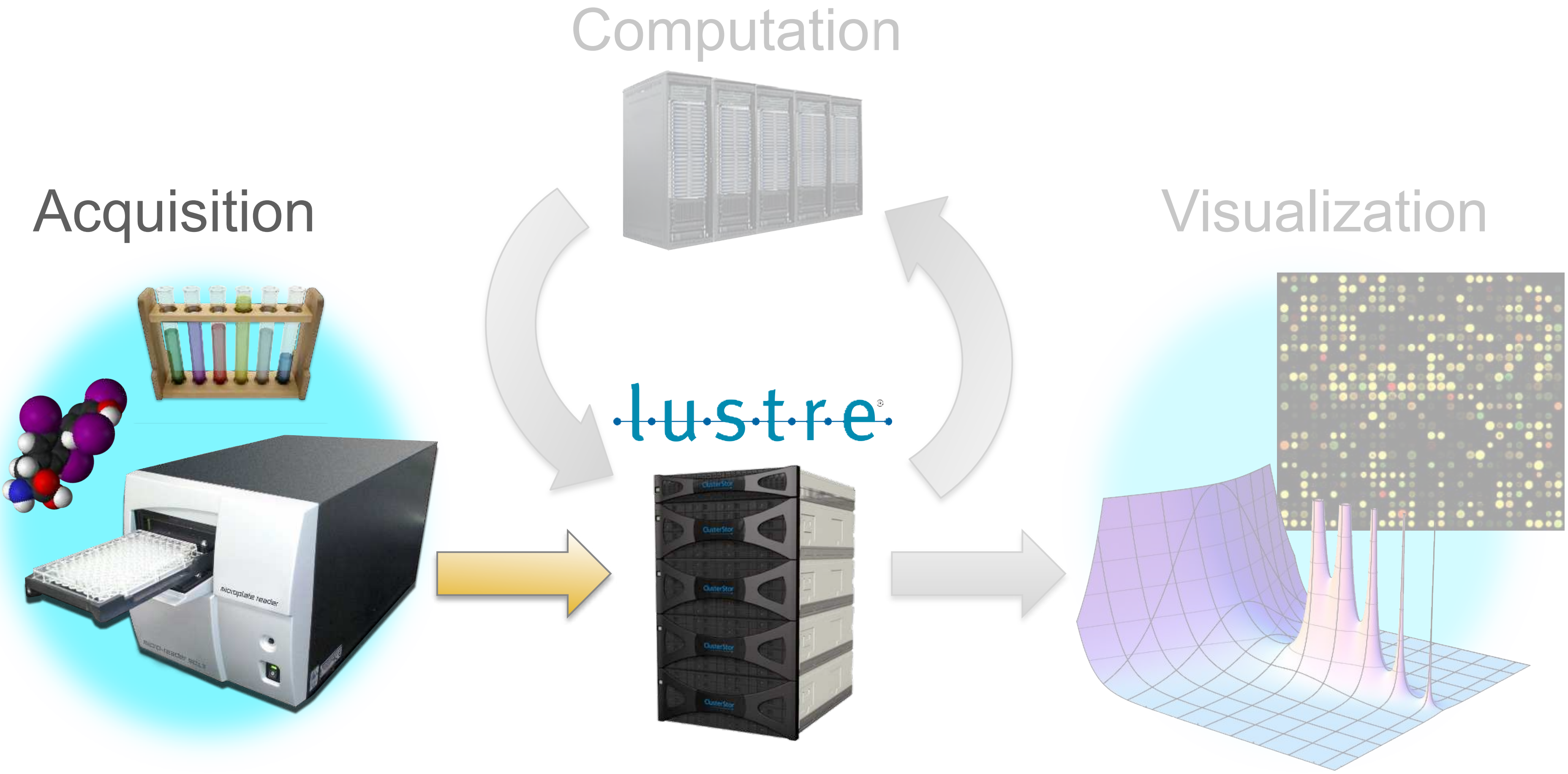
Visualization

lustre

Storage

Workflow #2

CIFS or NFS access required for Data Acquisition



Computation

Acquisition

Visualization

Storage

Workflows

1. Data Visualization
   - Smaller data sets, results of HPC jobs
   - Non-Lustre platform
   - Dozens (not hundreds/thousands) of clients

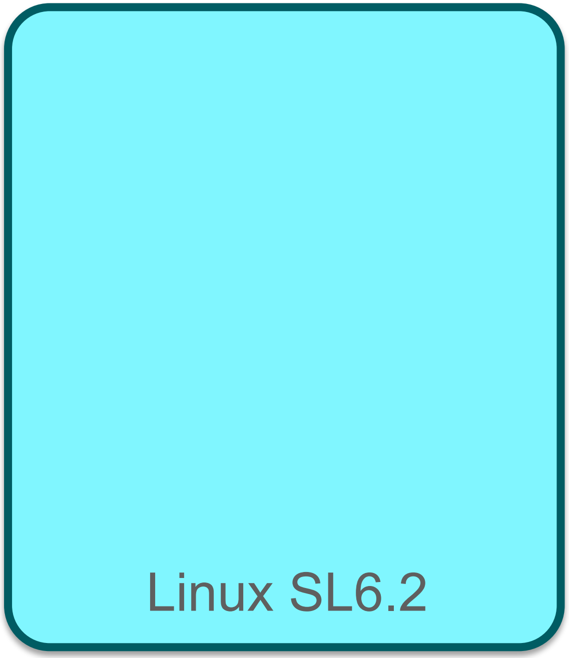
2. Data Acquisition
   - Moderate intake rate, continuous up to many weeks
   - 100s of GB → TB
   - Non-open hardware / non-Lustre platform
   - May run multiple experiments in parallel

Architectural Goals

- Improved Availability
  - Constrained by Windows, SMB, Lustre, Linux HA
- Reliability
  - Consistency of operation
  - Ease of service
- Manageability
  - Unified configuration and monitoring
- Performance
  - Reasonable but not HPC
- Security
  - Conform to site standards

Software

CIFS/NFS Gateway for Lustre

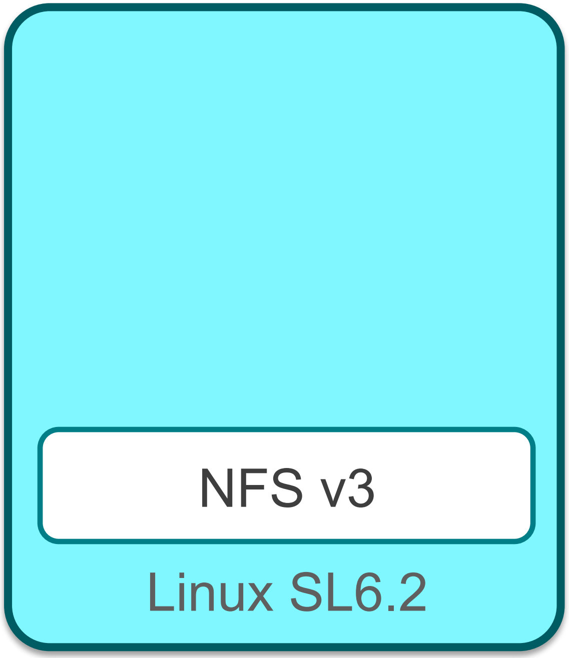


Linux SL6.2

Base operating system

Software

CIFS/NFS Gateway for Lustre

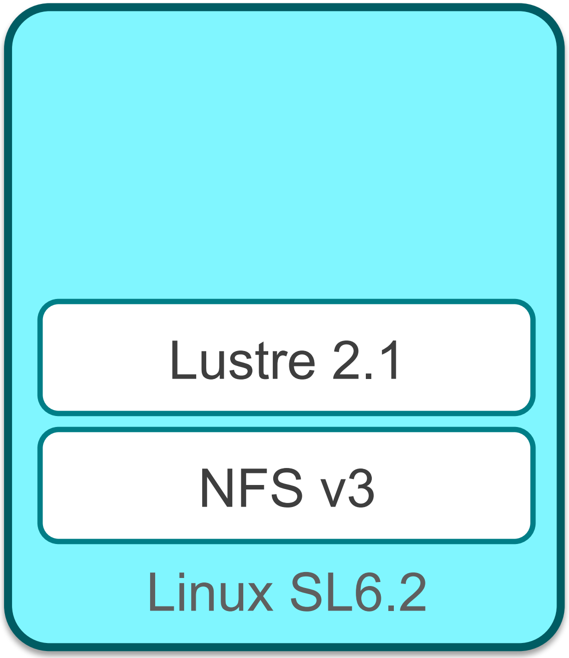


NFS:

Included in base OS with kernel support

Software

CIFS/NFS Gateway for Lustre



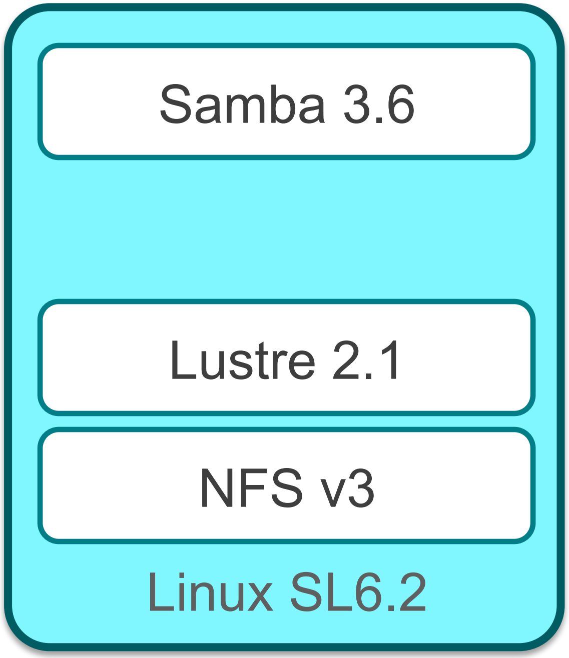
Lustre:

Kernel patches, used here to access Lustre servers

(Typically would use Lustre client module)

Software

CIFS/NFS Gateway for Lustre

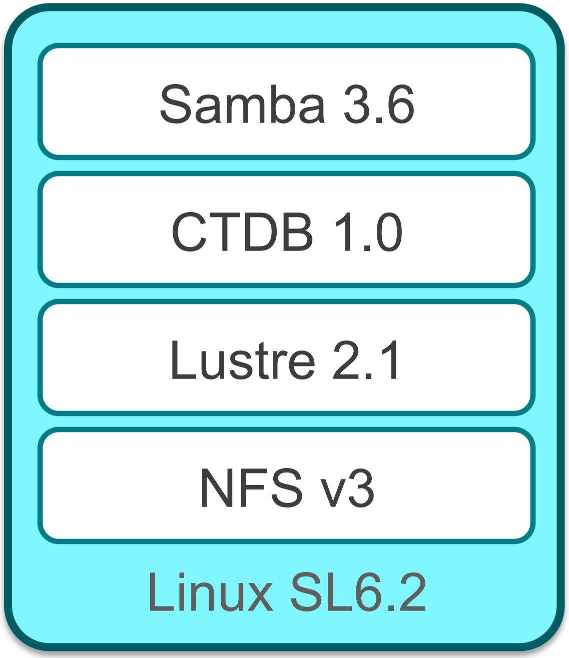


Samba:

Server to enable Windows client support

Software

CIFS/NFS Gateway for Lustre


Samba 3.6

CTDB 1.0

Lustre 2.1

NFS v3

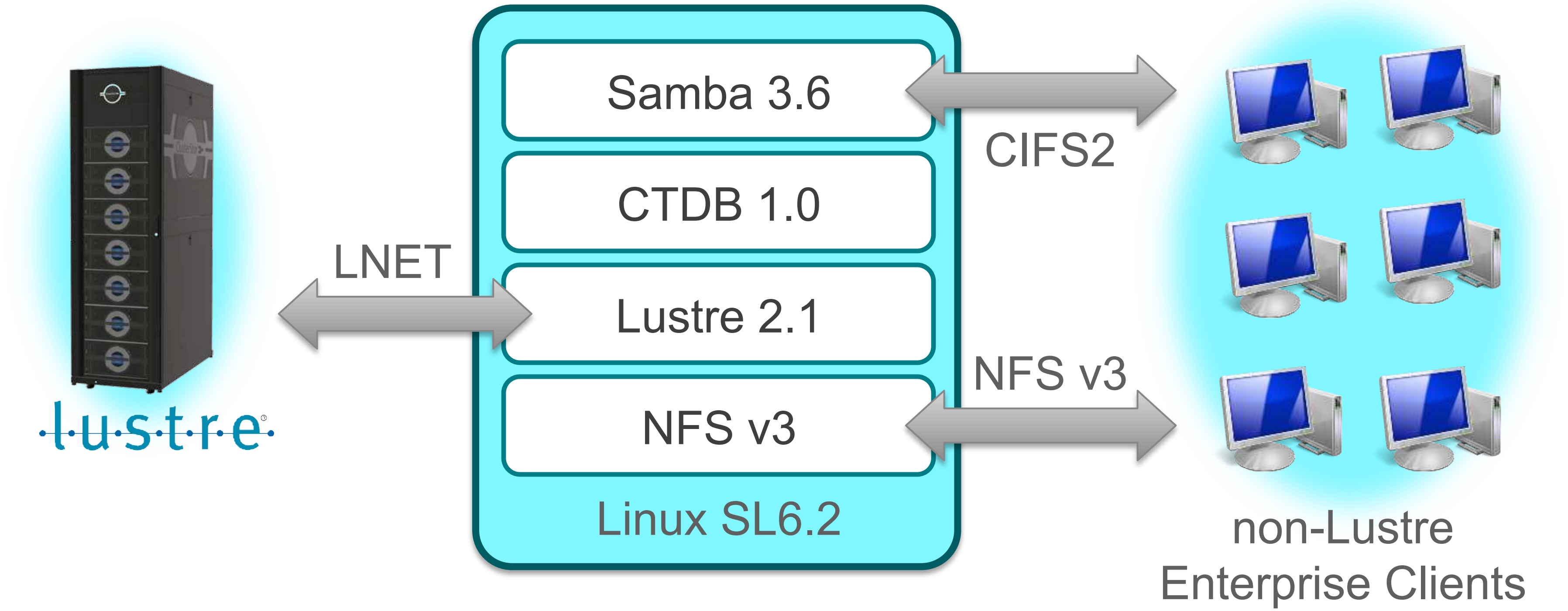
Linux SL6.2

CTDB:

Clustering support

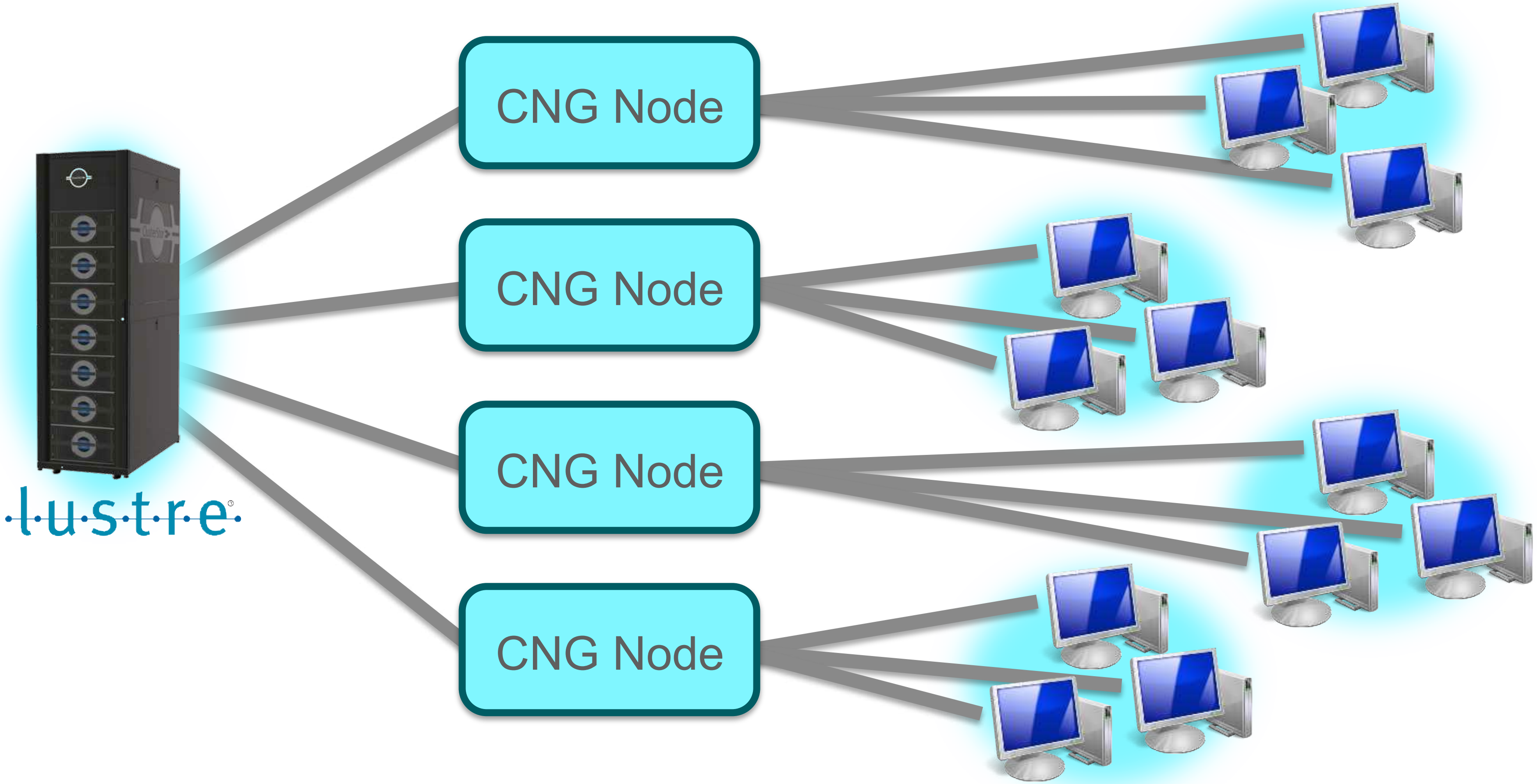
Software

CIFS/NFS Gateway for Lustre



Conceptual Model

2 or 4 CIFS/NFS Gateway Nodes for 1 Lustre filesystem



Conceptual Model

Networking

- Each Gateway node is a Lustre client
  - Connects to Lustre through same core switch as any other Lustre client
  - Typically not the only Lustre clients
- Connects to separate network for sharing to enterprise clients
- Enterprise authentication occurs on enterprise network

Hardware

- 4-node rack-mounted unit
  - Redundant power supplies
  - 2-node option
- Each node
  - 1 x E5 2680 CPU
    - 8 cores, 2.7GHz
  - 32 GB DDR 1600 DRAM
  - diskless
  - BMC / IPMI
    - power control
    - environment monitoring

Hardware

Networking

- Lustre servers
  - InfiniBand FDR or 40 Gb Ethernet
- Enterprise clients
  - 10/40 Gb Ethernet
- Manageability
  - 2 x 1 Gb Ethernet

Clustering

Samba CTDB

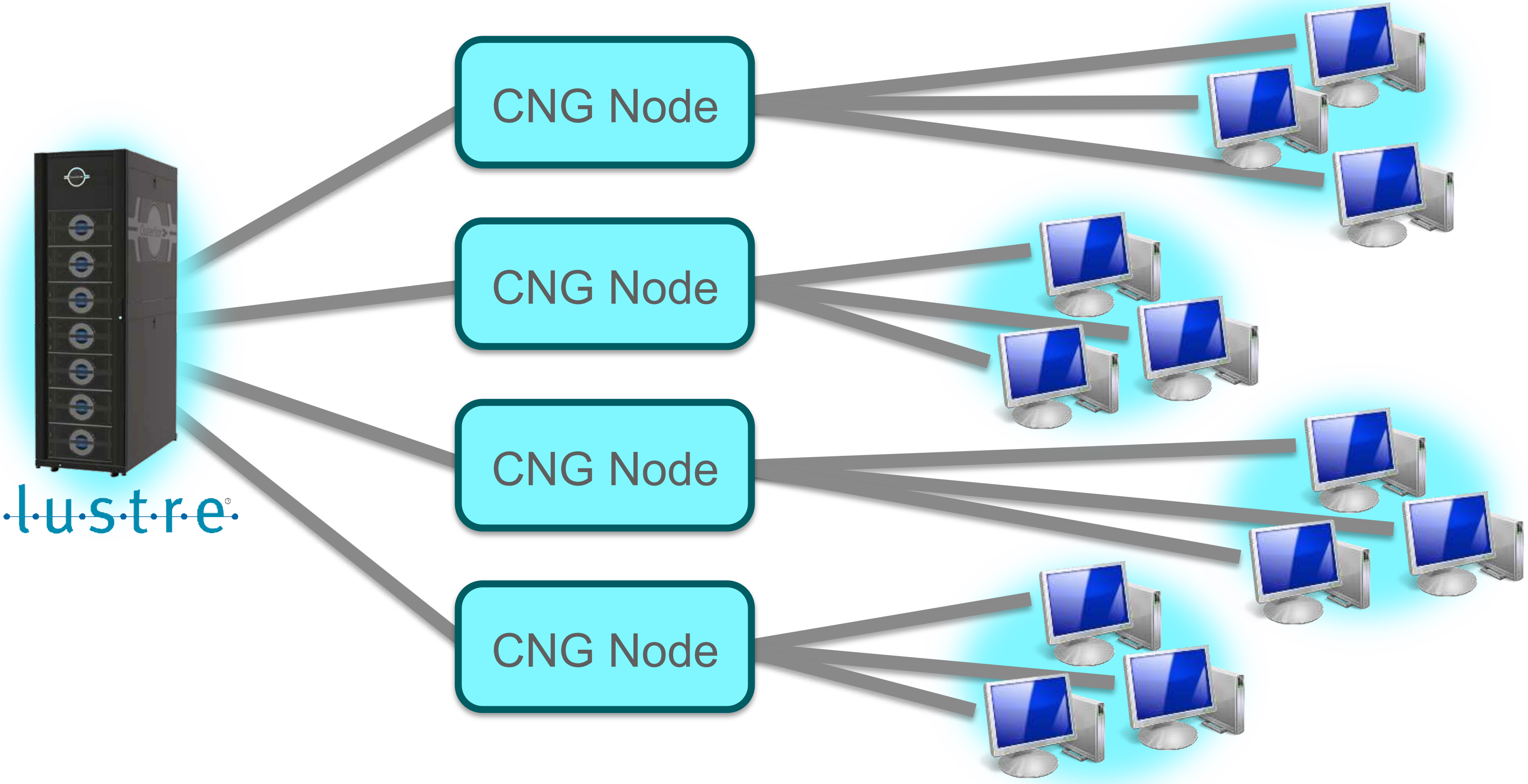
- "Clustered Trivial Database"
- Optional part of Samba
  - Also supports NFS
- Provides IP failover for clustered nodes
- Limited load balancing
  - Best used with Round-Robin DNS
- Implements a clustered version of Samba's Trivial Database
  - Replicated to all nodes

CTDB Limitations

- Cluster size max ~ 6 nodes
- Availability is improved but not always seamless
  - Mismatch of expectations / timeouts:
    - Windows applications vs. Lustre & HA configuration
  - May have to reconnect Windows clients manually
- Load balancing can become unbalanced
- Requires storage for TDB files
  - Challenge: Store local data on diskless nodes!
  - Solution: Add a “replicate only” node to CTDB cluster
  - Don’t forget to backup the TDB files

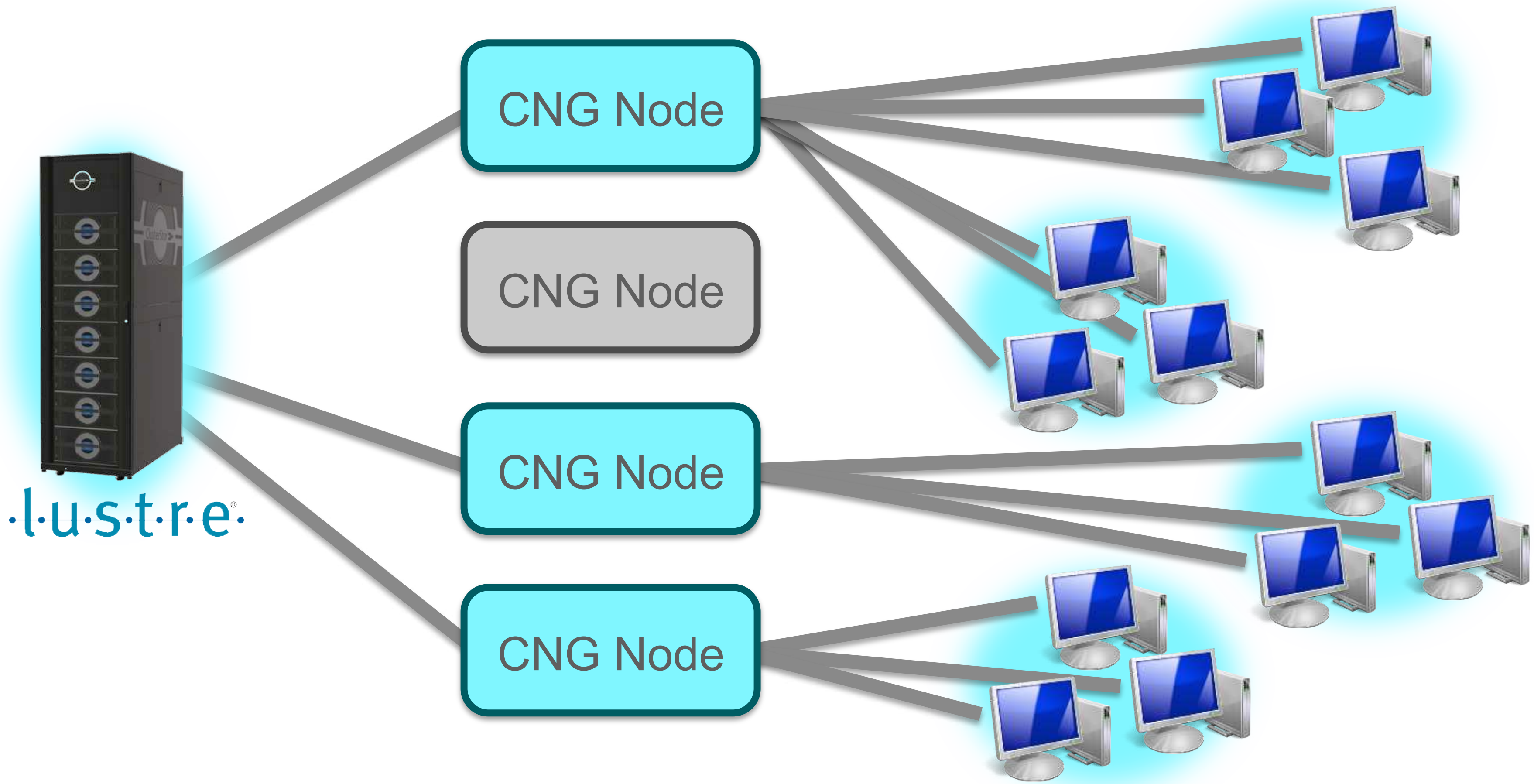
Improved Availability

IP Failover



Improved Availability

IP Failover



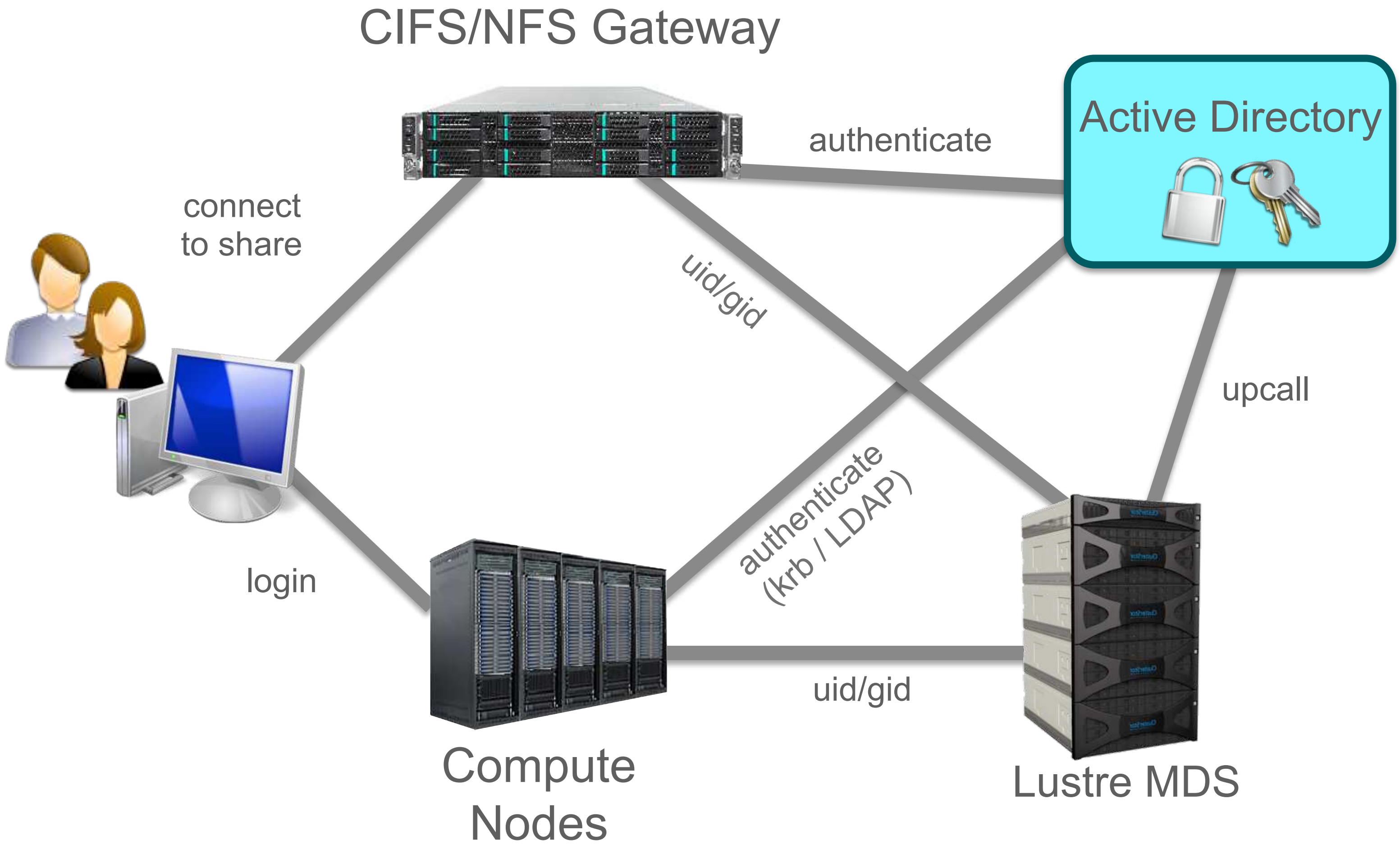
LAD14 Reims France 2014 LUSTRE ADMINS & DEVELOPERS WORKSHOP

User Security Challenges

- Host vs. User paradigm
  - » Lustre and NFS use host-level kernel/root connection
  - » SMB/CIFS requires user credential to connect
- User accounts are typically external
  - » Active Directory, etc.
- Lustre MDS upcall may not always match
  - » Lustre upcall: NIS
  - » Windows clients: AD
- File/Directory permissions follow POSIX
  - » ACLs might cascade through tree differently

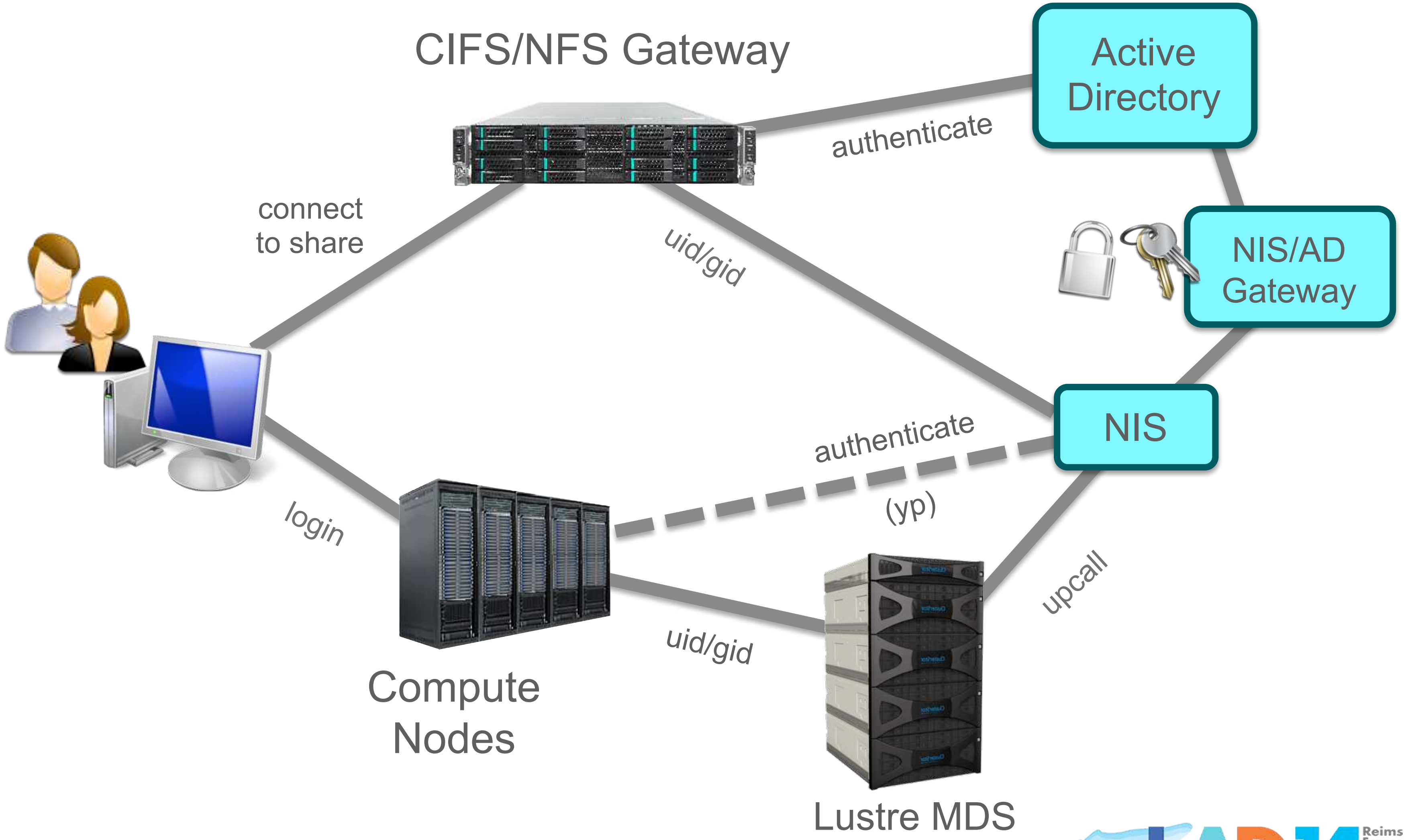
User Security Site Integration

AD Example



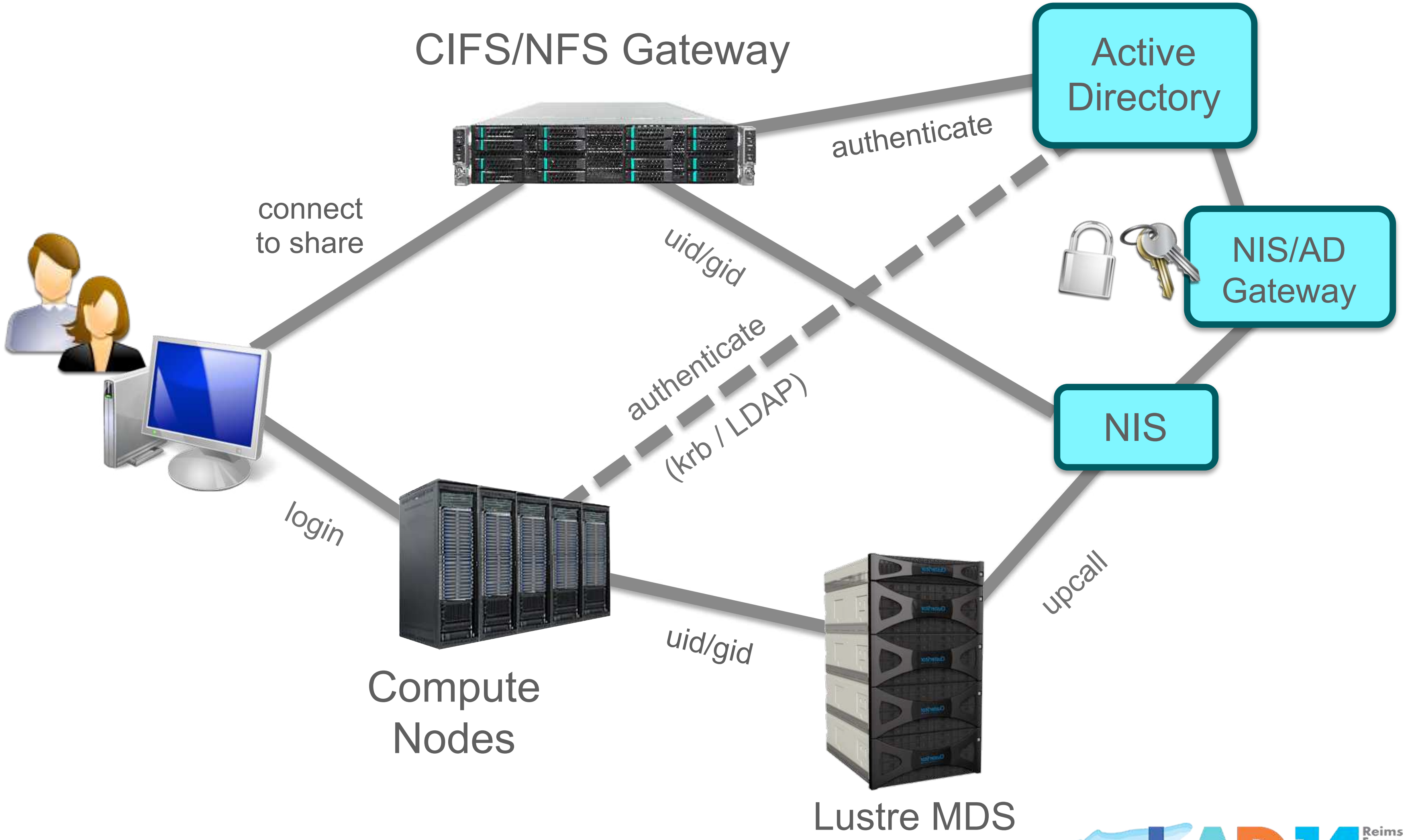
User Security Site Integration

AD+NIS Example 1



User Security Site Integration

AD+NIS Example 2



Takeaways from User Security Variations

- Understand how Windows/SMB clients will authenticate during user connection to shares
  - 100% independent of Lustre configuration
- Understand where POSIX uid/gid will come from
  - Must correspond to upcall configuration
- Make sure that Samba configuration matches
  - If careless, Samba might allocate its own uids
  - id mapping in Samba can be subtle/tricky

Manageability & Monitoring

- Icinga / Nagios checks
  - Hardware components
  - Key software layers
  - Network performance
- Power control / Bringup
  - Follow best-practice mount/unmount sequences
- Unified CIFS/NFS sharing setup
  - Entire filesystem or separate shares
  - Security configuration for each sharing protocol
  - May divide clients into groups

Thank You

Chris Gouge
Seagate Cloud Systems and Solutions