Lustre Security Isolation

DataDirect Networks, Inc.

Sebastien Buisson sbuisson@ddn.com

2016/04

Background of Security for HPC

- **Security was not first priority on HPC storage systems**
  - Made for Performance and Scalability
  - No direct connection to Internet and physical security

- **Today, HPC storage is NOT just scratch and user home directory use case is commonplace**
  - Same cluster with various use cases
  - Dedicated hardware not efficient
  - Secured data accessible/visible ONLY to people who have credentials and are authorized

© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.

Security Requirements for HPC/LifeScience Storage Systems

- **User/node authentication**
  - Only authenticated users have access
  - Only authenticated nodes are part of Lustre

- **Isolation**
  - Provides isolated namespaces from a single filesystem
  - Limited namespace exposed to clients

© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.

Lustre: User/node Authentication

- **Based on Kerberos**
- **We already have it!**
  - With kerberized Lustre
    - Nodes need Kerberos credentials to be part of the file system
    - Users need their own Kerberos credentials to access Lustre file system

© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.

Lustre: Isolation

- **We combine features of:**
  - Containers
    - Each container mounts Lustre as a client
    - ‘root’ user is allowed inside containers
  - Kerberos
    - Each container authenticates with its own credentials
      - Not that easy because Lustre runs in kernel space
  - Subdirectory mount
    - Each container is allowed to mount only a portion of the namespace
    - Allowance depends on client’s credentials

© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.

What are Containers?

- **We use Docker containers**

- **Containers combine use of:**
  - Linux namespaces for application's view of exec environment
    - Processes, networks, UIDs/GIDs, mounts
  - Linux cgroups for resource control
    - CPU, memory, block I/O, network

- **Containers are lighter than virtual machines**
  - Run with host kernel
  - Dedicated to an application

© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.

Why Containers?

- **We want**
  - Different populations of users on the same file systems
  - Isolation of these different populations of users

- **Containers avoid static distribution of client nodes**
  => dynamic container images instantiation
  - No need to dedicate groups of clients to each population
  - Every client is available for any population
  - Several populations can share same client nodes at the same time

© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.

Lustre Client in Containers

- **Running Lustre client in a Docker container requires:**
  - Being root… and
  - Loading kernel modules
  - Adding new entries under /proc
  - Mounting

⇒ **We run containers in '--privileged' mode**

And simply 'mount'

- **Lustre docker clients are not really independent**
  - More like several mounts on the same node

© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.

Lustre Client in Containers

- **Problem seen when running Lustre client from container**
  - Client does not use NID configured in container
    - Always takes first NID configured on the host

- **Add a check in LNet**
  - Assign high value to NID's `order` if associated network interface is not available

⇒ Work done in patch "LU-7845 lnet: check if address is visible"

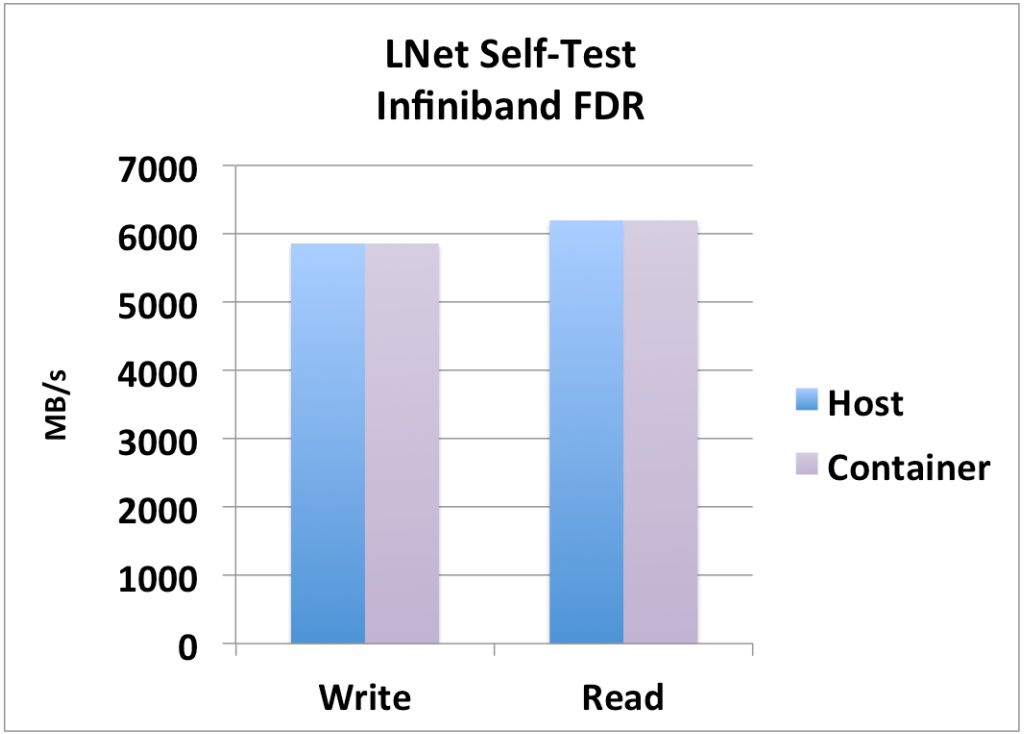
© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.

Lustre Client in Containers

- **What about performance?**
  - R&D test-bed
    - Hardware
      - SFA 7700 delivering 7BG/s+
      - Client node
        - 16 cores
        - 128 GB RAM
        - IB 4X FDR
    - Software
      - CentOS 7 (3.10 kernel)
      - Lustre pre-IEEL3.0 (2.7 + patches)
      - OFED 3.18-1
      - Docker 1.8.2

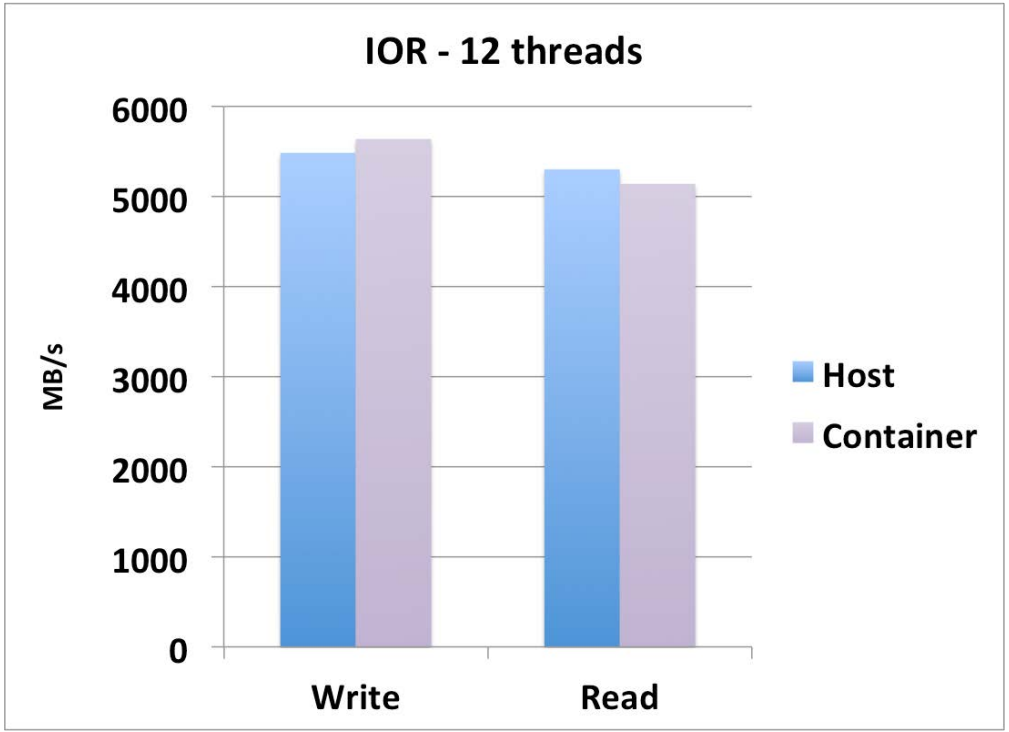
© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.

Lustre Client in Containers



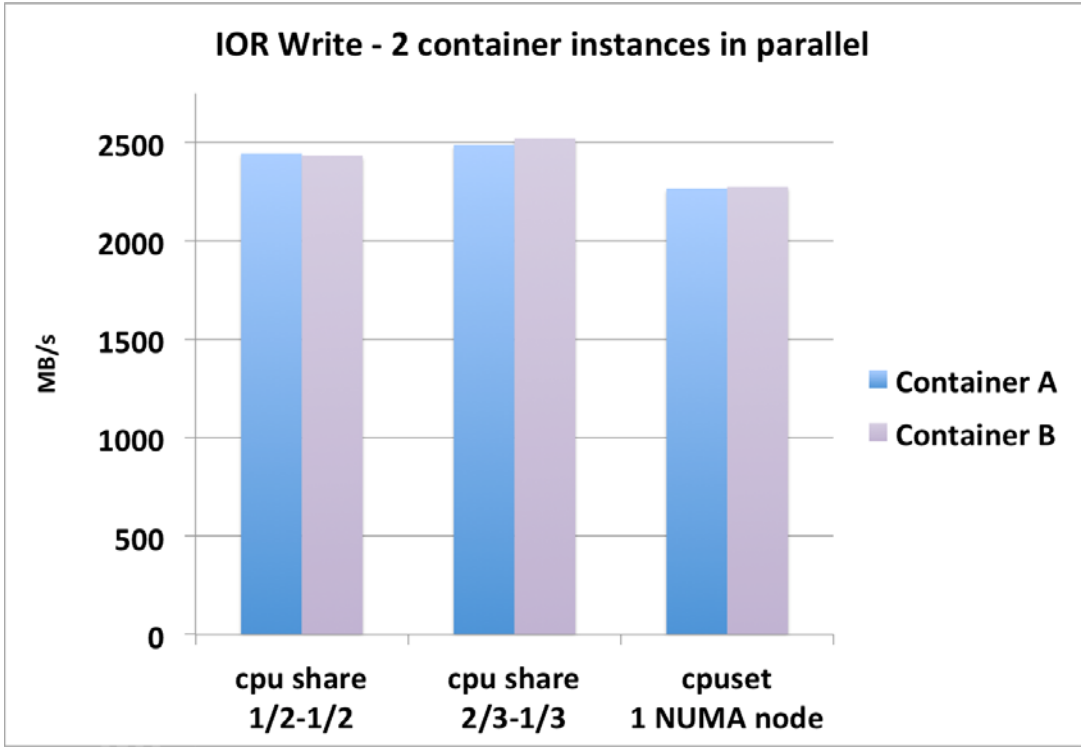
© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.

Lustre Client in Containers



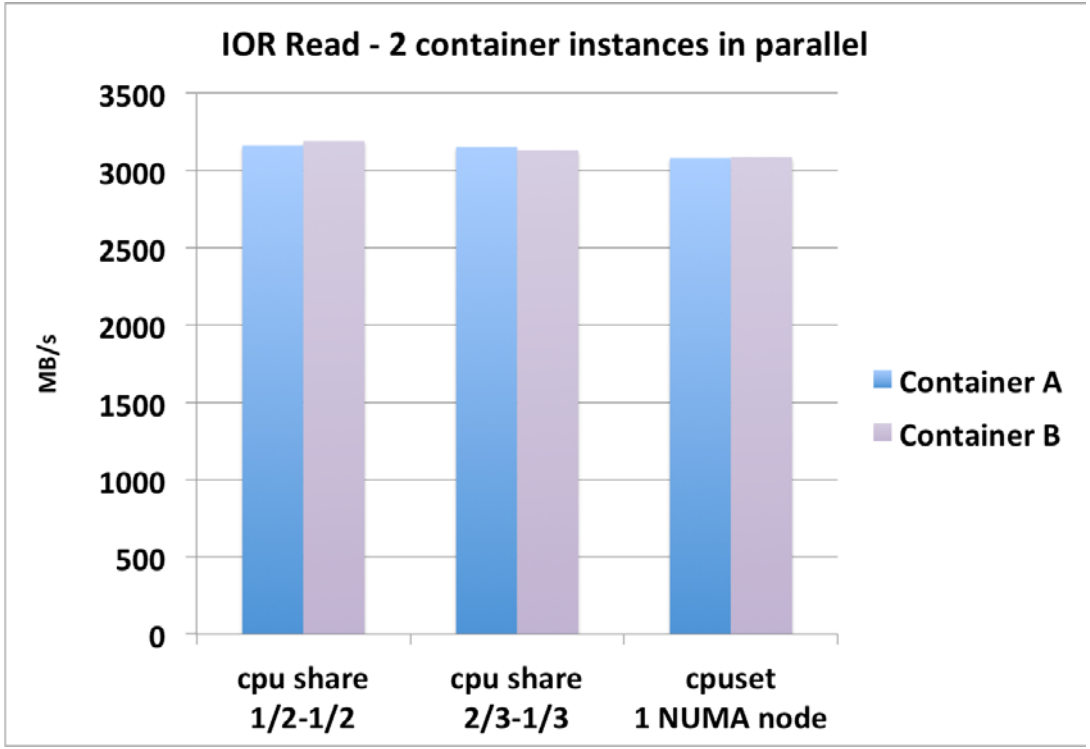
© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.

Lustre Client in Containers



© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.

Lustre Client in Containers



© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.

Lustre Client in Containers

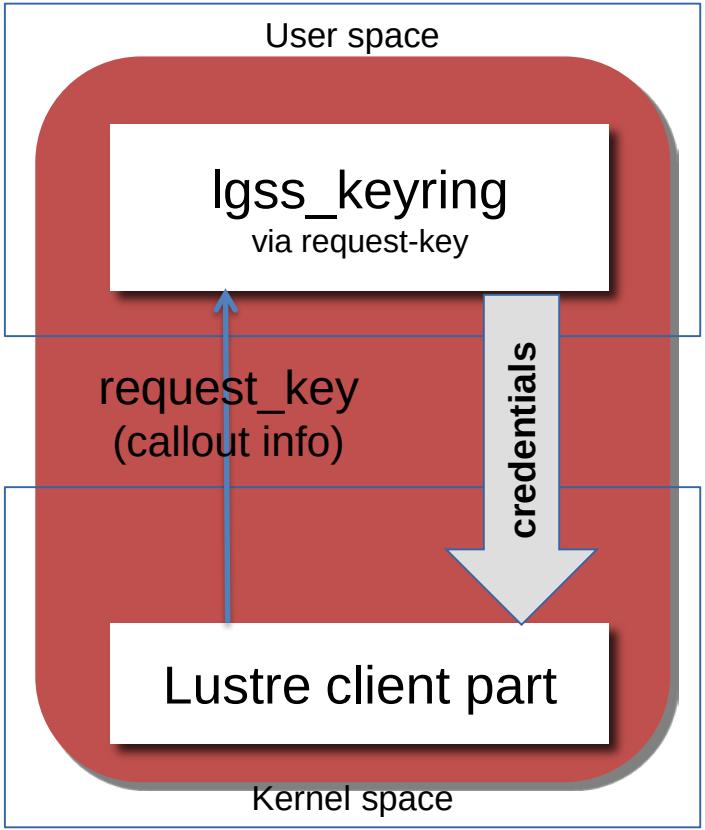
- **Docker containers are considered sandboxes for users**
  - ‘root’ can do anything

- **But still: need to know in which container Lustre client is running**
  - ⇒ Kerberos authentication
    - Reliable because only security admin can install new credentials

© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.

Lustre Kerberos

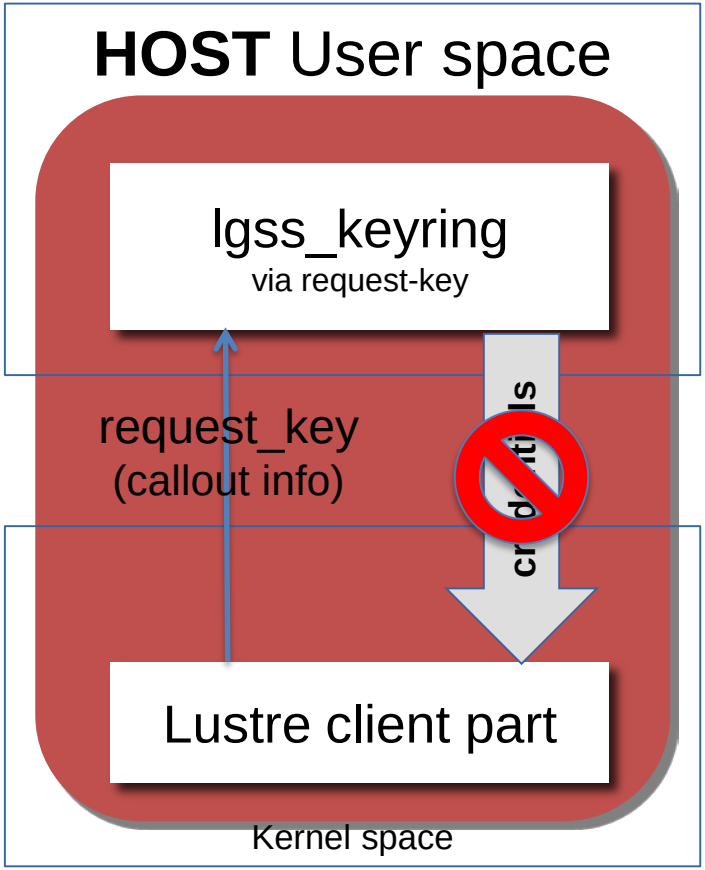
- **Standard credentials retrieval on client side**



DDN STORAGE
© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.
ddn.com

Lustre Kerberos

- **Credentials retrieval from container**



© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.

Lustre Kerberos in Containers

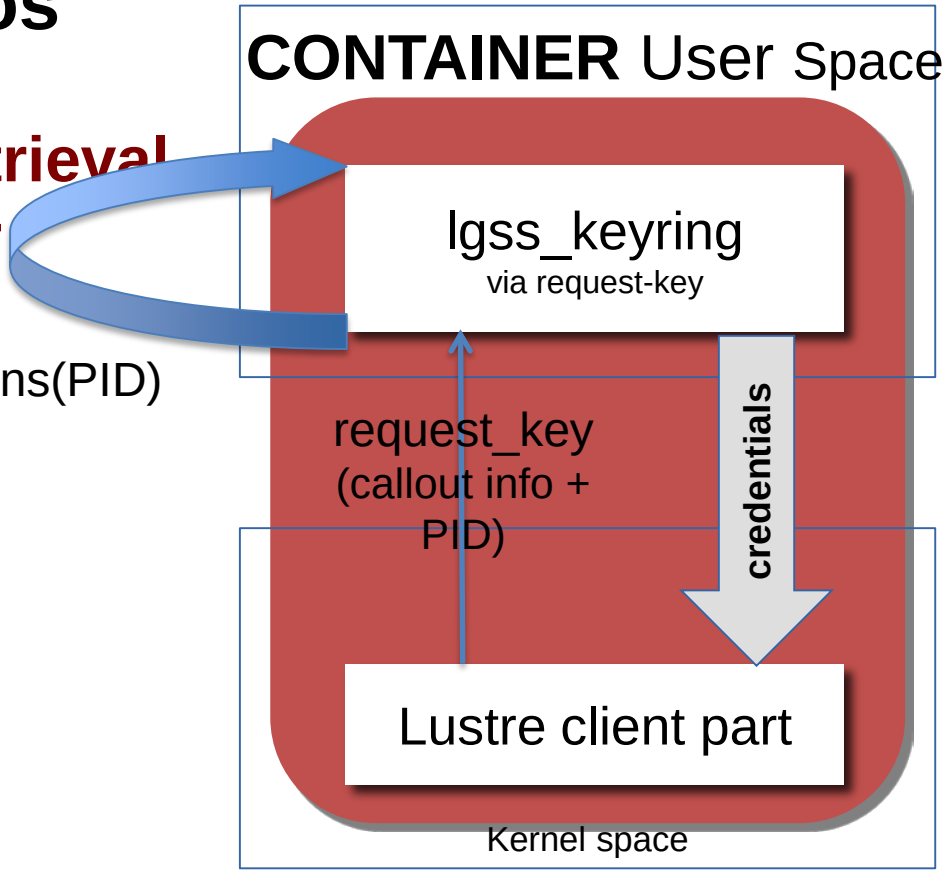
- **Credentials retrieval from container**
  - At client startup, get PID of container's init process
  - Store this PID info into `struct obd_import`'s `imp_sec_refpid` new field
  - When calling `request_key`, add `imp_sec_refpid` to callout info

⇒ Work done in patch "LU-7845 gss: support namespace in lgss_keyring"

© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.

Lustre Kerberos

- **Credentials retrieval from container**



DDN STORAGE
© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.
ddn.com

Lustre Access from Containers

- **Execution environment is isolated**
- **Lustre clients in containers are identified**

- **But still: isolation of data on Lustre is not guaranteed**

  ⇒ Enforce subdirectory mount

© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.

Lustre Access from Containers

- **Subdirectory mount (future Lustre feature)**
  - Aka fileset mount

```
mount -t lustre mgsname:/fsname/subdir /mount_point
```

- **Each container is assigned a dedicated subdirectory on Lustre**
- **But for complete isolation, we do not want containers to be aware of this**
  - Automatically return specific subdir on mount
  - Based on client NID

© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.

Lustre Access from Containers

- **Definition of a *nodemap* (foundation for UID/GID mapping feature):**

When an operation is made from a NID, Lustre decides if that NID is part of a ***nodemap***, a policy group consisting of one or more NID ranges.

- **So why not using *nodemap* feature?**

© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.

Lustre Access from Containers

- **Enhance *nodemap* feature**
  - Commands:

    ```
    lctl nodemap_set_fileset
    lctl set_param nodemap.set_nodemap_fileset
    ```

  - Sets a fileset on a nodemap

  ⇒ Work done in patch “LU-7846 nodemap: add fileset info to nodemap”

© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.

Lustre Access from Containers

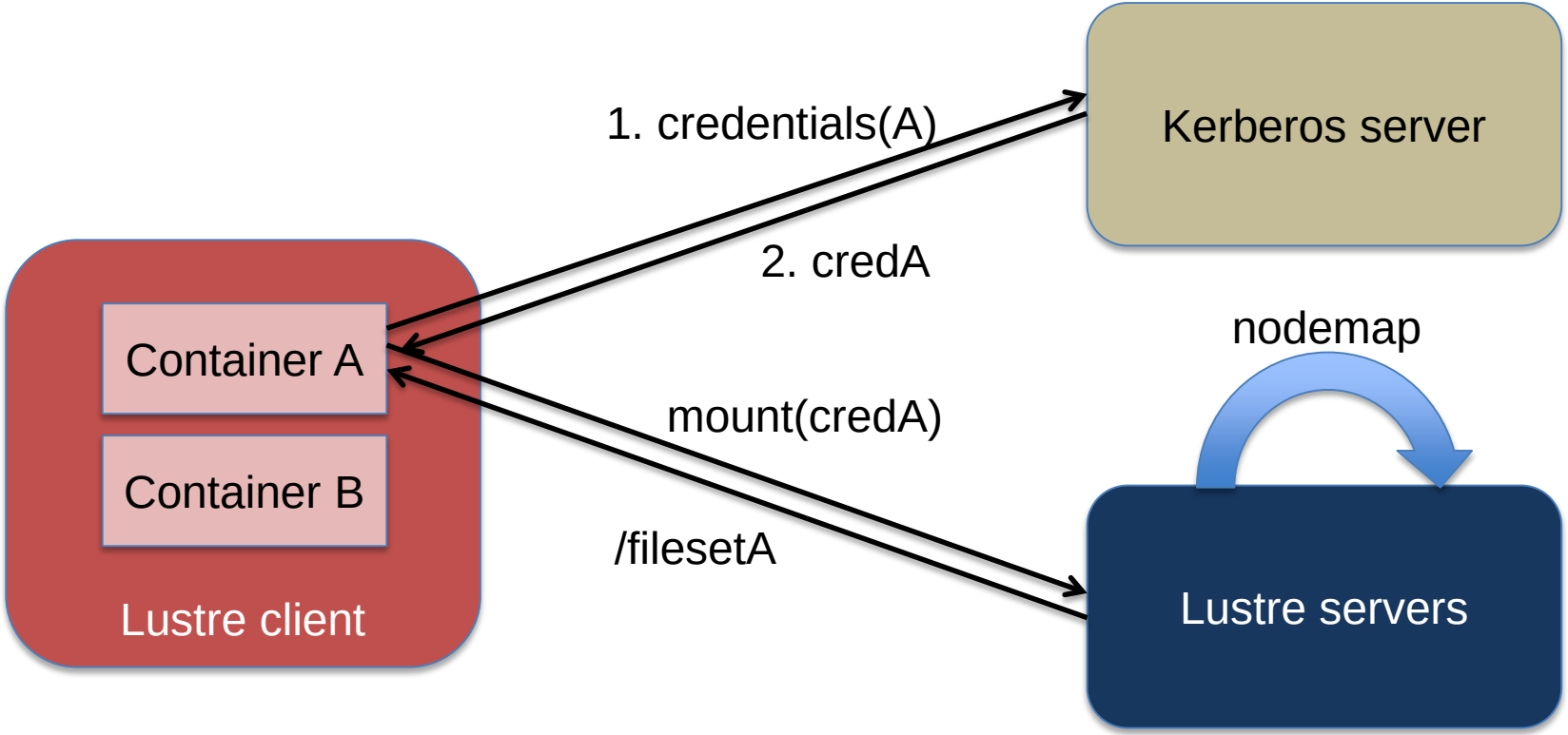
► Enhance *nodemap* feature

- When fileset info is set
  - All clients pertaining to this nodemap will be automatically returned this fileset at mount
  - If explicit ask for subdirectory: do a sub-subdir mount

⇒ Work done in patch “LU-7846 mdt: mount with fileset info from nodemap”

© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.

Lustre: Isolation



© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.

Lustre: Isolation

- **We are able to provide isolation feature to Lustre**

- **By combining:**
  - Linux Containers
  - Kerberos authentication
  - Subdirectory mount
  - Nodemap

© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.

Thank You!

Keep in touch with us

Team-jpsales@ddn.com

@ddn_limitless

company/datadirect-networks

102-0081
東京都千代田区四番町6-2
東急番町ビル 8F

TEL:03-3261-9101
FAX：03-3261-9140

© 2016 DataDirect Networks, Inc. * Other names and brands may be claimed as the property of others.
Any statements or representations around future events are subject to change.