Kerberized Lustre 2.0 over the WAN

Josephine Palencia
Robert Budden
Kevin Sullivan
Content

- Kerberos V5 primer
  - authenticated lustre components

- Setup
  - PSC, SDSC, NRL, UF/orida (OSG)

- Distributed OST/OST pools
  - arranging user directories

- Increasing Lustre 2.0 accessibility
  - test systems: kerberos, Quota, ACL
  - tools enabling kerberized data transfer (konFUSE)
  - integration with NFS4 (non-/kerberized)
  - enable kerberos authentication by LNET

- Current constraints/problems

- Summary

- References
Kerberos V5 Primer

- Client
  - authenticates itself to KAS,
  - demonstrates to TGS that it's authorized to receive a ticket for a service
  - demonstrates to SS that it has been approved to receive service

U = user, CL = client
KAS = Kerberos authentication server
TGS = Ticket Granting Server
SS = Service server
**Authenticated Lustre components**

```
lctl conf_param jwan.srpc.flavor.default=krb5p
  *jwan.srpc.flavor.tcp0=krb5n
  *jwan.srpc.flavor.tcp1=null
jwan.srpc.flavor.default.cli2ost=krb5i
jwan.srpc.flavor.default.mdt2mdt=null
jwan.srpc.flavor.default.mdt2ost=krb5i
_mgs.srpc.flavor.default=krb5p
```

<table>
<thead>
<tr>
<th>Flavor</th>
<th>Auth</th>
<th>RPC message</th>
<th>Bulk Data Prot</th>
</tr>
</thead>
<tbody>
<tr>
<td>null</td>
<td>null</td>
<td>null</td>
<td>checksum(adler32)</td>
</tr>
<tr>
<td>plain</td>
<td>null</td>
<td>null</td>
<td>checksum(adler32)</td>
</tr>
<tr>
<td>krb5n</td>
<td>GSS/krb5</td>
<td>null</td>
<td>checksum(adler32)</td>
</tr>
<tr>
<td>krb5a</td>
<td>GSS/krb5</td>
<td>partly integrity</td>
<td>checksum(adler32)</td>
</tr>
<tr>
<td>krb5i</td>
<td>GSS/krb5</td>
<td>integrity</td>
<td>integrity(sha1)</td>
</tr>
<tr>
<td>krb5p</td>
<td>GSS/krb5</td>
<td>privacy</td>
<td>privacy(sha1/aes128)</td>
</tr>
</tbody>
</table>

lustre_(root, mgs, mds, oss)/goldeneye.jwan.teragrid.org@TERAGRID.ORG(AES-256 CTS mode with 96-bit SHA-1 HMAC)

- Shift among kerberos flavors during mounts
- Prelim IO benchmark compare different flavors to check performance overhead by kerberos

*Kerb auth by LNET*
Setup

- Lustre 2.0 Alpha 5 (CentOS5)
- Release: 1.9.280
- Kernel: 2.6.18-128.7.1
- Kerberos auth: krb5p
Cross-realm kerberos authentication work
NSF funding to set up Kerberized Lustre 2.0 for OSG
- Secure Lustre 2.* based at University of Florida
- Project entails setting up kerberos infrastructure, creating kerberos realms
ExTENCI: Extending Science Through Enhanced National CyberInfrastructure

- consortium similar to Teragrid with funding from NSF/DOE
- provides/uses middleware called Virtual Data Toolkit (VDT)
- established worldwide interoperable systems - World Wide LHC Computing Grid for CERN LHC experiments

OSG Collaborators: www.openscience.org

- University of Florida
- Pittsburgh Supercomputing Center
- University of Chicago
- Clemson University
- Louisiana State University
- Purdue University
- University of Wisconsin, Madison
- Fermi National Accelerator Laboratory
- Brookhaven National Laboratory
- Florida State University
- Florida International University
Distributed OSTs/OST pools

- Decentralize/distribute secure Lustre WAN storage across several organizations/different sites
- Sites write to local OST pools but filesystem visible on WAN
- Data striped
  - across PSC dist OSTs/pool
  - across SDSC dist OSTs/pool
  - across both sites OSTs
Distributed OST/OST pool

Best way to arrange TeraGrid user directories in a lustre WAN with distributed OST/OST pool?

1. Sysadmins: OSS-centric directory arrangement
   Data is 'striped' to local OSS contributed by each site
   so /jwan/$SITE/users

   For the 6 sites and user joe (N-> 0-9),
   /jwan/psc/N/joe
   /jwan/sdsc
   /jwan/taccs
   /jwan/iu
   /jwan/nics
   /jwan/ncsa...

2. Users: user-centric way of seeing their data in specific machines
   so /jwan/$USER/$SITE/$MACHINE

   /jwan/joe/psc
   /psc/pople
   /jwan/joe/tacc
   /tacc/ranger
   /tacc/lonestar
   /jwan/joe/nics
   /nics/kraken...
Distributed OST/OST pool

Best way to arrange TeraGrid user directories in a lustre WAN with distributed OST pool?

3. Combining both and adopting something similar to AFS

so /jwan/users/$USER/$PSC/$machine symlink to /jwan/$SITE/N/$USER/$MACHINE

More specifically,

/jwan/users/joe/psc/pople --> /jwan/psc/users/N/joe/pople

where

/jwan/users/$USER and /jwan/users/$USER/$SITE level directories are both read-only and highly-available to users
Constraints: Part I

- Lustre 2.0 – only interoperable with (non-kerberized) Lustre 1.8
  - Mount Lustre 1.8.* on a kerberized Lustre 2.0 client without disabling kerberos on the Lustre 2.0 servers

- Native clients, OSTs- must be running lustre 2.0
  - same release (kernel/lustre rpms- available)

- Other sites/partners don't have (will not have) the kerberos infrastructure

How to make kerberized lustre 2.* accessible and usable in the current framework?
Made available resources @ PSC

- 11,115 Teragrid Accounts auto-synced from TGCDB (Teragrid Central Database)
- Teragrid users can login to VM kerberized lustre 2.0 client `goldeneye.jwan.teragrid.org` and
  - access the filesystem `/jwan`
  - run tests: kerberos, quota, ACLs
  - request writes to local PSC OSTs (pool), remote SDSC osts (pool), combination of local and remote OSTs (pools)
- perform data transfer on Lustre 2.0 (gridftp/gsiscp/kftp) with VM data target `thunderball.jwan.teragrid.org`
Kerberized scp/kftp/gridftp: konFUSEd

pople.psc.edu

Login Node

File System

/boot
/usr
/home
/tmp

kinit joe@TERAGRID.ORG

krb5cc_1000 written to /tmp

“konfused” (fuse interceptor)
ticket propagated to /tmp

thunderball.jwan.teragrid.org

Lustre WAN Client

• GridFTP
• HPN SSH Server
• KFTP Server

File System

/boot
/jwan
/tmp
Kerberized scp/kftp/gridftp: konFUSE

```
joeuser@tg-login3.pople.psc.teragrid.org:~]$ scp file thunderball.jwan.teragrid.org:/jwan/users/joeuser/test
joeuser@thunderball.jwan.teragrid.org's password:
scp: /jwan/users/joeuser/test: Operation not permitted

[joeuser@tg-login3.pople.psc.teragrid.org:~]$ module load konfuse

[joeuser@tg-login3.pople.psc.teragrid.org:~]$ env | grep KRB
KRB5CCNAME=/konfusefs/krb5cc_20033

[joeuser@tg-login3.pople.psc.teragrid.org:~]$ kinit joeuser@TERAGRID.ORG
joeuser@TERAGRID.ORG's Password:

[joeuser@tg-login3.pople.psc.teragrid.org:~]$ scp file thunderball.jwan.teragrid.org:/jwan/users/joeuser/test
file 0% 0 0.0KB/s 0.0KB/s ---:--- ETA file 100% 611KB 611.0KB/s 611.0KB/s 00:00
Max throughput: 611.0KB/s

[joeuser@tg-login3.pople.psc.teragrid.org:~]$ exit
```
Kerberized NFS4

<table>
<thead>
<tr>
<th>Lustre2.0</th>
<th>NFS4</th>
<th>User Access to /lwan-nfs4</th>
</tr>
</thead>
<tbody>
<tr>
<td>null</td>
<td>sec, krb5 (i, p)</td>
<td>✔</td>
</tr>
<tr>
<td>krb5 (n, a, i, p)</td>
<td>-</td>
<td>✗ /lwan-nfs4 remains mounted but inaccessible to users</td>
</tr>
</tbody>
</table>

Exports /lwan2.0 via NFS4 kerberos - enabled

mount /lwan2.0 via Lustre with auth krb5p
lwan.psc.flavor.default = krb5p

mount /lwan2.0-nfs4 via NFS4 (sec = krb5p)
Kerberized NFS4

- Keyring
- Pipefs (remnant)

Keyring (keyctl)

Lustre

+ mechanism for key forwarding

NFS4

1 keyring

% keyctl show
Session Keyring
-3 --alswrv 0 0 keyring: session.4098
138187962 --alswrv 0 0 \_ key: key type
604914970 --alswrv 0 0 \_ user: mykey
603447101 --alswrv 0 0 \_ user: nfs
882237880 --alswrv 0 0 \_ user: lgss

Keys
- TGT
- lgss (lustre)
- nfs (NFS4)

jwan.srpc.flavor.default =

sec = sec
krb5
krb5i
krb5p

null
krb5n
krb5a

*Univ of Michigan: kerb NFS4
*Group doing NFS Keyring port

© 2010 Pittsburgh Supercomputing Center
Constraint: Part II

- Lustre kerberos in Single Kerberos Realm
  - necessary to sync all accounts on MDS and all clients
  - user-land: 3-fold check (uid, gid, kerberos principal)
  - lustre ID mapping turned off for single kerberos realm

- Lustre kerberos- Cross-realm
  - lustre ID mapping enabled
Summary

➢ Workable secure Lustre 2.0 system with constraints
➢ Continue to increase accessibility
  • tool for kerberized data transfers- konfuse
  • integration with NFS4
  • kerberos authentication by LNET
➢ Lustre kerberos realm: Single->Cross realm
  • both user-land and systems
  • key management
References

TeraGrid Kerberized Lustre 2.0 Wiki

References

PSC Kerblustre Wiki
wiki.psc.edu/twiki/view/KerbLustre/WebHome

Welcome to the KerbLustre web

Introduction

KerbLustre is a resource guide to the kerberos-enabled Lustre 2.0 file system we have established over the WAN. Lustre Kerberos authentication is enabled by default.

- Lustre version: 2.0.5 Alpha
- Lustre release: 2.0.306
- Kernel: 2.6.19-120.7.1
- Kerberos flavor: krb5p (default)
- Network id: 112.58 Subnets
- Infiniband: 1.5 (being worked on)
- Storage: 1.4TB (Single-realm) and 1.3TB (Cross-realm)

Systems in the TERAGRID.ORG Kerberos realm

<table>
<thead>
<tr>
<th>FS mounted</th>
<th>Hostname</th>
<th>Server</th>
<th>OS</th>
<th>Production Status</th>
<th>Lustre Purpose</th>
<th>Disk Space</th>
<th>Kerberos Realm</th>
<th>Network</th>
<th>Notes</th>
<th>Life Cycle DB</th>
</tr>
</thead>
<tbody>
<tr>
<td>jwan</td>
<td>mgs.jwan.teragrid.org</td>
<td>Physical Machine</td>
<td>CentOS 5</td>
<td>Production</td>
<td>MGS</td>
<td></td>
<td>TERAGRID.ORG</td>
<td>112</td>
<td>rcons4</td>
<td>system info</td>
</tr>
<tr>
<td>mds00w.psc.jwan.teragrid.org</td>
<td>Physical Machine</td>
<td>CentOS 5</td>
<td>Production</td>
<td>MDS</td>
<td></td>
<td>TERAGRID.ORG</td>
<td>112</td>
<td>rcons5</td>
<td>system info</td>
<td></td>
</tr>
<tr>
<td>oss01w.psc.jwan.teragrid.org</td>
<td>Physical Machine</td>
<td>CentOS 5</td>
<td>Production</td>
<td>OSS</td>
<td>1.4 TB</td>
<td>TERAGRID.ORG</td>
<td>112</td>
<td>rcons5</td>
<td>system info</td>
<td></td>
</tr>
<tr>
<td>jwan1</td>
<td>mgs.jwan.teragrid.org</td>
<td>Physical Machine</td>
<td>CentOS 5</td>
<td>Production</td>
<td>MGS</td>
<td></td>
<td>TERAGRID.ORG</td>
<td>112</td>
<td>rcons5,power reset</td>
<td>system info</td>
</tr>
<tr>
<td>mds01w.psc.jwan.teragrid.org</td>
<td>Physical Machine</td>
<td>CentOS 5</td>
<td>Production</td>
<td>MDS</td>
<td></td>
<td>TERAGRID.ORG</td>
<td>112</td>
<td>rcons5,power reset</td>
<td>system info</td>
<td></td>
</tr>
<tr>
<td>oss00w.psc.jwan.teragrid.org</td>
<td>Physical Machine</td>
<td>CentOS 5</td>
<td>Production</td>
<td>OSS</td>
<td>1.4 TB</td>
<td>TERAGRID.ORG</td>
<td>112</td>
<td>rcons6</td>
<td>system info</td>
<td></td>
</tr>
<tr>
<td>attractor.psc.edu</td>
<td>Physical Machine</td>
<td>CentOS 5</td>
<td>Development</td>
<td>Other</td>
<td></td>
<td>TERAGRID.ORG</td>
<td>112</td>
<td>rcons6</td>
<td>system info</td>
<td></td>
</tr>
</tbody>
</table>

Add or Change Machines

PSC: Virtual machines

<table>
<thead>
<tr>
<th>VM Name</th>
<th>Hostname</th>
<th>Server</th>
<th>OS</th>
<th>Production Status</th>
<th>Lustre Purpose</th>
<th>Disk Space</th>
<th>Kerberos Realm</th>
<th>Network</th>
<th>Notes</th>
<th>Life Cycle DB</th>
</tr>
</thead>
</table>

https://wiki.psc.edu/twiki/edit/KerbLustre/WebHome?c=1271008331
RPM packaging

Index of ftp://ftp.psc.edu/pub/jwan/Lustre-2.0-alpha/

- 1.10.0.40  Lustre 2.0 Beta 1
- VM client/server rpms