pNFS Lustre layout discussion

IETF 83, nfsv4 WG – Atlanta, November 8, 2012

Sorin Faibish - EMC
Peng Tao - EMC

Motivation

- Lustre uses own RDMA in the data protocol;
  - no need of additional RDMA protocol like NFS or SRP or iSER data protocol is using Lustre client in the kernel
- Lustre client will be in the kernel and all the data path will be based on native Lustre client
  - pNFS will be the layout management
  - all metadata operations will be handled by pNFS/generic NFS layout

Implementation strategy

- Intel and EMC will support the implementation
  - support of Matthew Wilcox and Andy Kleen (conf call agreed) for Lustre client in kernel
- EMC will implement the pNFS client and the pNFS MDS server for Linux
- pNFS will improve MD performance and scalability of Lustre and provide high performance that Lustre has today
  - solution as promised to DoE that started pNFS
  - will relax Lustre POSIX consistency to NFS close-to-open consistency to help performance in shared data access
- Will allow pNFS to be wider adopted in commercial application that use Lustre today: CAD/CAM, O&G, Pharma

Status of the draft – why is needed

Lustre layout is new and different from other layouts

http://www.ietf.org/internet-drafts/draft-faibish-nfsv4-pnfs-lustre-layout-02.txt

1. Take advantage of lnet performance.
2. Built-in mature RDMA support.
3. Take advantage of Lustre OSS asynchronous journal commit mechanism, to improve write performance (http://static.usenix.org/events/fast10/tech/full_papers/oral.pdf)
4. Require new layout and new RFC. Cannot use 5664.
5. pNFS Lustre layout client will use Lustre client modules and assume in kernel and distros

Comments and review (David Black)

- Reference [1] for the protocol spec is outdated like a 3-year-old Lustre, not a current protocol spec.
  - Intel and EMC are committed to put and maintain Lustre client in kernel
  - Intel Lustre team will post an update of the document; in works.
  - Consensus to have source control including the document and only change between major distros releases
  - based on Lustre client that will be updated in major releases; pNFS layout driver unchanged.

Comments on draft (David Black)

- v1 and v3 magic numbers help, but it concerns that the draft is descriptive about what Lustre currently does, as opposed to prescriptive about what a Lustre that supports this pNFS layout MUST do.
  - First draft uses what Lustre does today. The next drafts we will define what MUST do
  - v1 and v3 magic numbers are not for version control but for feature description.
  - We intend to remove them when client in kernel. Will only support last server version at time of client in kernel.
  - Will replace with flags or hints or attributes used at mount time and mount will fail if there is mismatch. Will include in Lustre documents.
- Will require better/longer introduction/overview text
  - Next draft will include more details to already improved draft 02

Comments on draft (Tom)

- Mike Eisler put together a simple way to test XDR file.
  - Requires a lot of changes in the original Makefile and it may not work (according to SteveD and us).
  - Will do it for a later draft. WiP
- Are there other transports other than TCP and IB?
  - In current Lustre implementation, only TCP and IB are supported
  - Can support any RDMA transport
- Need to add a requirements/usecase section or separate draft
  - Will discuss with the list the options a new conf call?
- Why is lmm_magic present?
  - Will remove the magic numbers in future
- Draft 02 also addressed all changes recommended by Jason Glasgow as well as the recommendations from Paris

Discussion

- Next steps:
  - Discussion in the nfsv4 list ; need a call
  - Discussion with Lustre developers-Intel
  - Draft 03 including review from meeting to be posted before next IETF
  - Lustre client to Linux kernel – for next ietf
- Q&A