NFS/RDMA and Sessions Updates

Tom Talpey
Network Appliance

RPC/RDMA Draft Updates

- RPC/RDMA draft updated in February
- http://www.ietf.org/internet-drafts/draft-ietf-nfsv4-rpcrdma-01.txt
- Integrated all comments received, including extensive review from Mallikarjun (thanks!).
- Document has no open issues at this time.

RPC/RDMA Draft Updates

- Protocol itself is unchanged
- Added language on chunking rules, non-restrictions, implementation issues.
- Clarified language around registering memory, removed use of RDDP terms.
- De-ID-nitted (mostly)
- Plus many clarifications and language improvements.

NFSdirect and Problem Statement Draft Updates

- http://www.ietf.org/internet-drafts/draft-ietf-nfsv4-nfsdirect-01.txt
- http://www.ietf.org/internet-drafts/draft-ietf-nfsv4-nfs-rdma-problem-statement-02.txt
- Minor updates only.
- Documents have no open issues at this time.

Sessions Draft Updates

- NFSv4/Sessions draft updated in February
- http://www.ietf.org/internet-drafts/draft-ietf-nfsv4-sess-01.txt
- Integrated all comments received
- Implementation experience
- Document has one clarification issue at this time.

Sessions document issue

- The SEQUENCE operation has a sequence id used for detecting retransmission
- It has been stated that it only needs to be a single bit, this is true only if the transport is reliable and ordered.
- Over unreliable transports (e.g. UDP), old duplicates can cause the server to replay.
  - The field is 32-bit and a client can easily prevent the problem, e.g. using a running counter (or the XID) for the sequence field.
  - Not a major issue - document will be clarified.

Implementation Experience – Connectathon 05

- Sessions Server
  - Protocol, Duplicate Request Cache (DRC) complete
  - Passes Connectathon
  - PyNFS tests, Ethereal working/under development
- Issues
  - Callback channels not supported
  - Needs update to current version of draft protocol
- Jon Bauman's presentation available at www.connectathon.org

Implementation Experience – Connectathon 05

- Sessions Client:
  - Protocol changes complete
  - Callback channel works!
  - Passing Connectathon tests
- Issues
  - Needs update to current version of draft protocol
- Mike Stolarchuk's presentation available at www.connectathon.org

New RPC transport switch

- Updated to Linux 2.6.9
- Supports traditional TCP, UDP over IPv4
- Bull IPv6 (TCP, UDP)
- RDMA over kDAPL
  - Infiniband
  - iWARP

Client Transport Switch Vector

New (further simplified) pointer in the “struct rpc_xprt”:

```
/* abstract functions provided by a transport */
struct rpc_xprt_ops {
    void (*setbufsize)(struct rpc_xprt *);
    void (*connect)(struct rpc_xprt *);
    fastcall int (*send_request)(struct rpc_task *);
    void (*close)(struct rpc_xprt *);
    void (*destroy)(struct rpc_xprt *);
};
```

Transport Hooks

- Each transport registers with switch
- NFS mount (and others) specify transport type and per-transport create data
- Transport gets control via xprt_procs, and network events
- Can unregister/unload

Current RPC switch patchset

- http://troy.citi.umich.edu/~cel/linux-2.6/2.6.9-a/
  - [Many rolled-up NFS improvements in CEL_NFS-ALL.patch]
  - Simplified from earlier versions
- Abstracts transport type, address family, per-xprt parameters, etc.

Mount API extensions

- At a minimum, pass transport type and addresses, NFS generic mount parameters
- Maintain per-transport arguments passed separately, and extensibly
- Logical rearrangement of existing arguments, with tag/bag of new
- Cleans up existing fields (security)

Client RDMA Implementation

- RPC/RDMA module
  - 3000 lines of code, 2 headers, 3 C files
- IB kDAPL providers
  - Available from several IB vendors
  - Also under way within OpenIB

Client RDMA Implementation

- Available as open source
  - BSD-style license
  - www.sourceforge.net/projects/nfs-rdma
- Linux 2.6 supported
  - (to be released to Sourceforge soon)
  - Requires RPC switch patch – 2.6.9
- 2.4 Linuxes (“old” RPC switch):
  - RedHat 7.3 (2.4.18)
  - SuSE 8 Enterprise (2.4.19)
  - RHEL 3.0 (2.4.21)

Server RDMA Implementation

- Will complete the Linux NFS/RDMA end-to-end implementation!
- Under way at UMich CITI
- Linux 2.6 (currently 2.6.9-ish)
- NFSv3 initially
- Full RDMA semantics
- Interoperate with Linux 2.6, 2.4 RDMA clients
- http://www.citi.umich.edu/projects/rdma/

Sessions Implementations

- Full Linux NFSv4/Sessions client and server
- Under way at UMich CITI
- Linux 2.6 (currently 2.6.9-ish)
- Initial release shortly:
- http://www.citi.umich.edu/projects/rdma/