RED HAT :: CHICAGO :: 2009
SUMMIT

It's Not your Father's NFS

Steve Dickson
Consulting Software Engineer, Red Hat
steved@redhat.com

presented by

Outline

- Moving to NFS version 4
- NFS version 4.1 and pNFS
- NFS and FS-Cache
- NFS Tracepoints
- NFS and SystemTap scripts
- NFS Metrics
- NFS and IPv6 Support

NFSv4 Advantages

Performance
- Read/Write Delegations

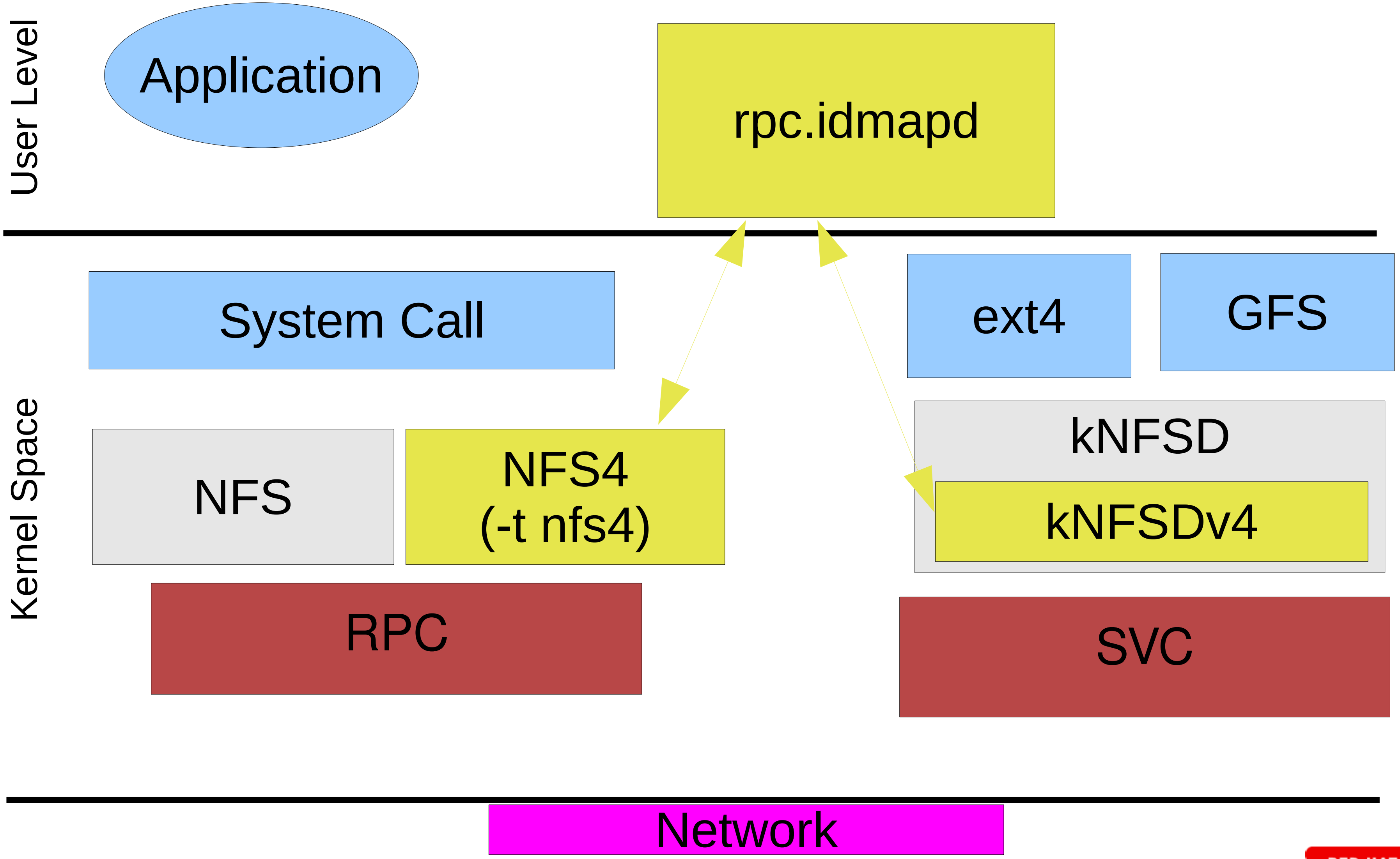
Server maintains client state
- Callbacks to Clients

Multi-Component Messages
- Less Network traffic

Mandates strong security architecture
- Available on **ALL** versions

Elimination of 'side-car' protocols
- No rpc.statd or In-kernel lockd
- Only port 2049

NFSV4 Architecture



NFSv4 Default Protocol

- Current exports will work seamlessly
  - No need for fsid=0 export
- A mount configuration file
  - Options per mount point
  - Options per server
  - Global options
- Mount to negotiate From V4
  - -t nfs4 option no longer needed.

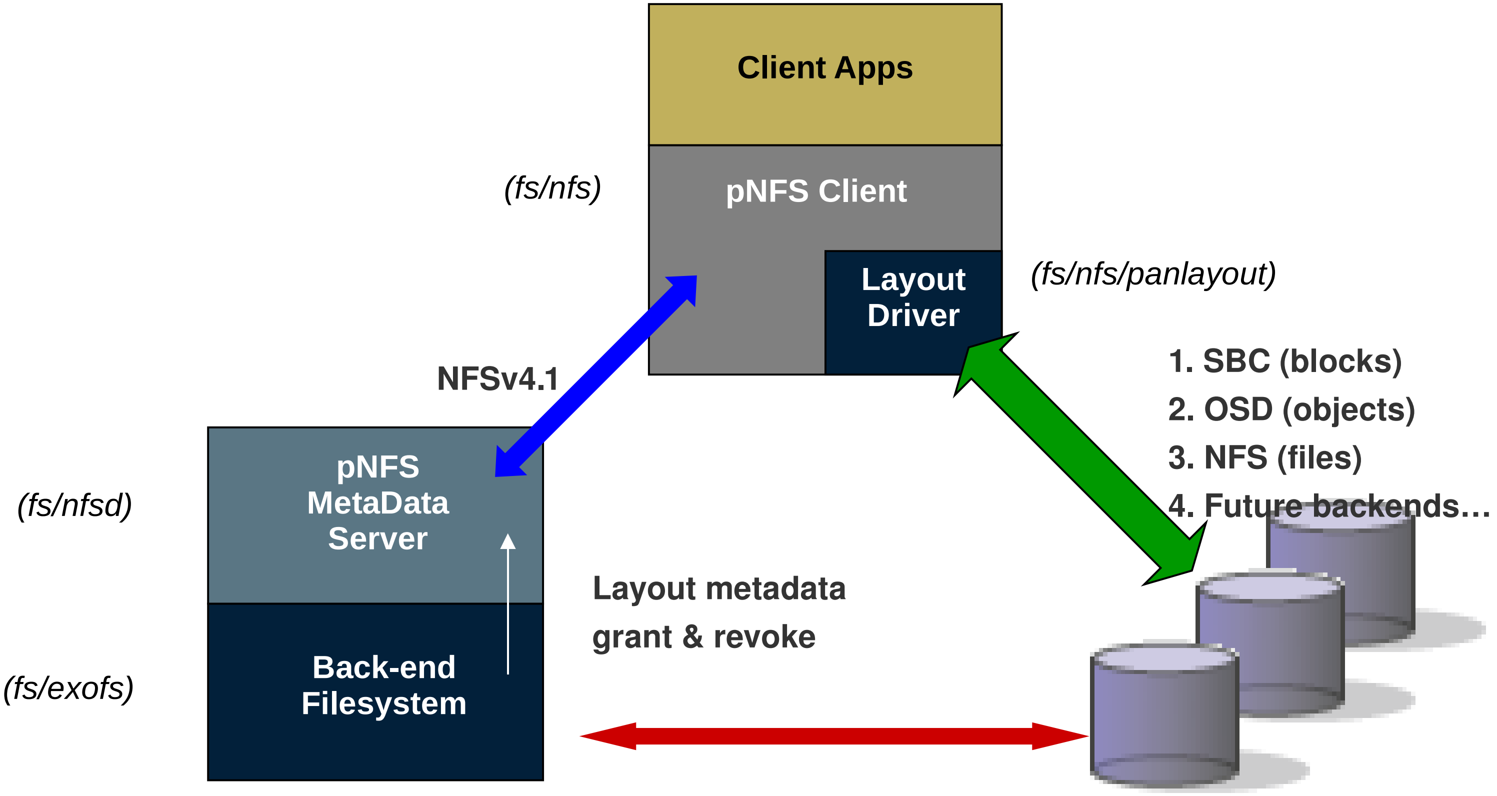
NFS minor version 1 (NFS41)

- Sessions
  - Exactly-Once semantics
    - Duplicate Request Cache
  - Callbacks –
    - More Firewall friendly
      - Made on same connection as requests
      - Client initiated
  - Directory Delegations
  - Enabling pNFS

Linux pNFS Overview

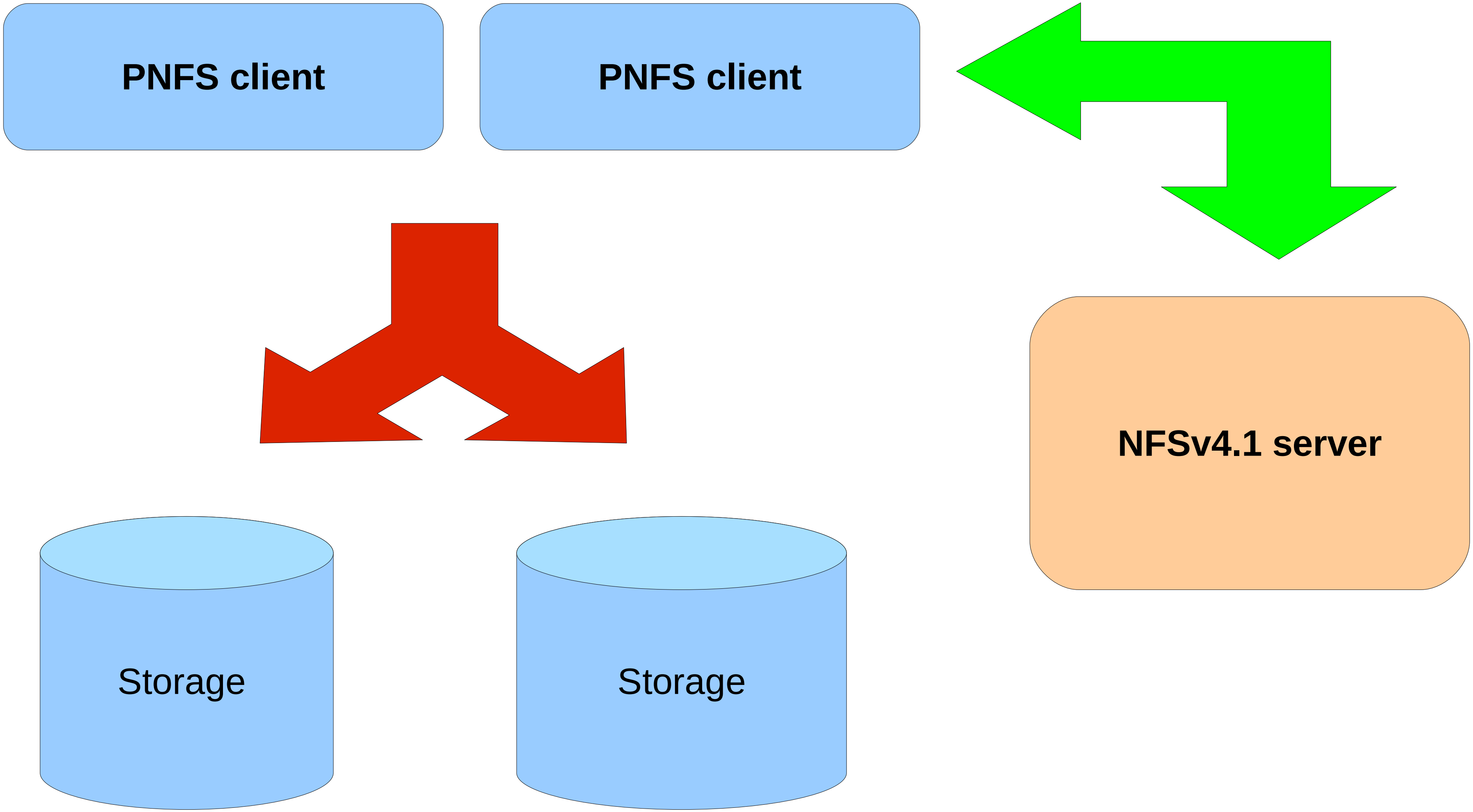
Common client and server code for different back ends

Integrated with existing NFSv4 codebase



(Slide Courtesy of: Benny Halevy, Panasas)

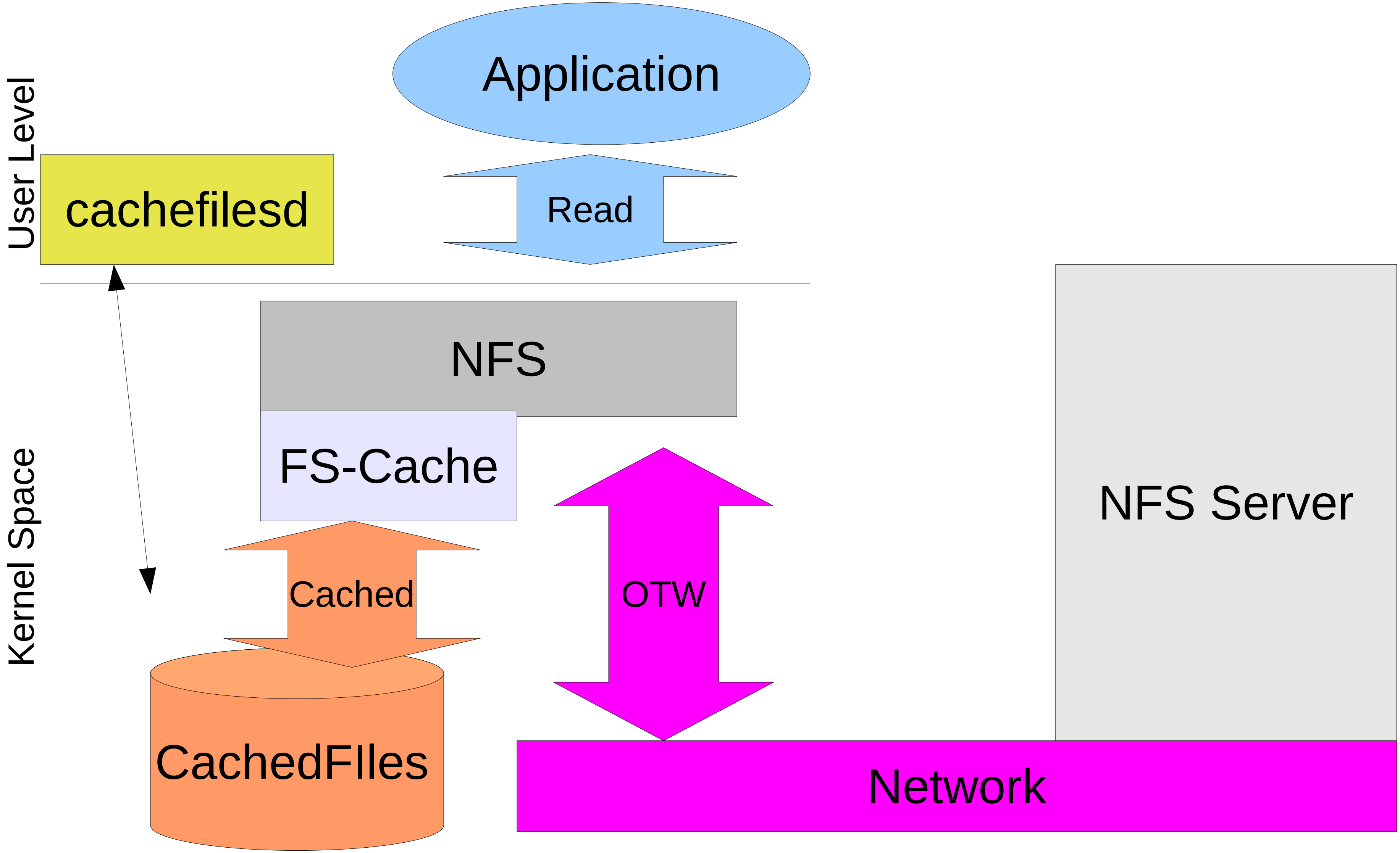
PNFS
Allows Clients to access storage directly



NFS and FS-Cache

- Main Goal: Improve Server Scalability
  - Some short term performance degradation on client
- Only Reads are Cached.
  - Opening the file for writes flushes and disables cache.
- Cache is maintained through umounts and reboots
- User level daemon used to maintain cache
  - Cachefilesd
- Possible Tech Preview in RHEL6

FS-Cache Architecture



NFS TracePoints

Trace Points are availability in RHEL5.3 and beyond.

- 3 tracepoints used for diagnostics

rpc_call_status

- Shows all errors that occur during NFS operations

rpc_connect_status

- Shows errors that occur during network connections

rpc_bind_status

- Show errors that occur during the binding of network connections

NFS and TracePoints (con't)

Need to install kernel-devel rpm

yum install kernel-devel

stap -L 'kernel.trace("*")'

Show all the available tracepoint

The tracepoints can be access by systemtap script:

```
probe kernel.trace(“rpc_call_status”)
{
     terror =  task_status($task);
     If (terror) {
            printf(“%s[%d]:call_status::%s:%s: error %d(%s)\n”,
               execname(), pid(), cl_server($task), cl_prog($task),
               terror, errno_str(terror));
     }
}
```

NFS and Systemtap

- Systemtap home page: **http://sourceware.org/systemtap/wiki/HomePage**
- Need to install **both** kernel-devel and kernel-debuginfo rpms
- man -k stap – shows all the 'built in' tap scripts which live in **/usr/share/systemtap/tapset** directory
  - man stapprobes.nfs – shows NFS scripts
- “Home grown” NFS tap scripts
  - git://fedorapeople.org/~steved/systemtap.git

NFS Metrics

- iostat -n
  - New '-n' flag to iostat command
    - yum install sysstat
    - Operations per sec
    - Reads and Writes per sec

```
Filesystem:  rBlk_nor/s   wBlk_nor/s   rBlk_dir/s   wBlk_dir/s   rBlk_svr/s   wBlk_svr/s   ops/s
rops/s    wops/s

tophat:/home   0.50        0.01        0.00        0.00        0.00        0.01      0.00      0.00       0.00
tophat:/home   15.71        0.01        0.00        0.00        0.00        0.01      0.00      0.00     0.00
```

NFS Metrics

- nfs-iostat
  - NFS client per-mount I/0 statistics
    - Statistic per memory page
    - Statistics per directory operations
    - Statistics per file access

```
rawhide:/home mounted on /mnt/home:

 op/s            rpc bklog
 233.00     2.10
read:           ops/s             kB/s          kB/op        retrans        avg RTT (ms)       avg exe
(ms)
           232.000        14908.719   64.262       0 (0.0%)        61.875                83.925
```

NFS Metrics

mountstats

Overall NFS client per-mount statistics

```
GETATTR:
        3 ops (0%)    0 retrans (0%)      0 major timeouts
        avg bytes sent per op: 138 avg bytes received per op: 112
        backlog wait: 0.000000      RTT: 0.333333      total execute time: 0.333333 (milliseconds)
LOOKUP:
        4 ops (0%)    0 retrans (0%)      0 major timeouts
        avg bytes sent per op: 144 avg bytes received per op: 176
        backlog wait: 0.000000      RTT: 0.750000      total execute time: 0.750000 (milliseconds)
READ:
        8001 ops (20%)    0 retrans (0%)      0 major timeouts
        avg bytes sent per op: 140 avg bytes received per op: 65655
        backlog wait: 22.235471     RTT: 58.915511    total execute time: 81.165479 (milliseconds)
WRITE:
        14997 ops (37%) 0 retrans (0%)      0 major timeouts
        avg bytes sent per op: 35107    avg bytes received per op: 136
        backlog wait: 1892.769887       RTT: 51.310862    total execute time: 1944.124225 (millisecon
```

NFS and IPv6 Support

Client side (almost done):

Goal is to have it “just work” when a hostname resolves to IPv6 address.

NFSv4 support is complete. NFSv2/3 is done except for rpc.statd, which is being rewritten.

Current release target is Fedora 13.

Server side (still experimental):

Kernel pieces are mostly in-place, rpc.nfsd is finished

IPv6-capable mountd/exportfs is still work-in-progress

Acknowledgments

NFSv4.1: An update
Mike Eisler, Network Appliance February 23, 2009
http://www.connectathon.org/talks09/eisler_cthon_2009.pdf

Progress on NFSv4.1: Definition and a review of changes from NFSv4.
Dave Noveck, Network Appliance, February 5, 2007
http://www.connectathon.org/talks07/NFSv41update.pdf

NFSv4 Sessions Linux Implementation Experience
Jon Bauman & Mike Stolarchuk, CITI, U of Michigan
Center for Information Technology Integration
University of Michigan, Ann Arbor
http://www.connectathon.org/talks05/bauman.pdf

Parallel NFS (pNFS)
Garth Goodson, Network Appliance, February 28, 2006
http://www.connectathon.org/talks06/goodson.pdf

Acknowledgments

NFS Version 4 Minor Version 1

draft-ietf-nfsv4-minorversion1-25.txt

http://www.nfsv4-editor.org/draft-25/draft-ietf-nfsv4-minorversion1-25.html

Object-based pNFS in linux

Benny Halevy, Panasas, May 4, 2009

Questions?

Slides:
http://people.redhat.com/steved/Summit09/Summit2009.odp

Email: steved@redhat.com

Tells what you think at:
http://redhat.com/sumit-survery

QUESTIONS?

TELL US WHAT YOU THINK:
REDHAT.COM/SUMMIT-SURVEY

FOLLOW US:

TWITTER.COM/REDHATSUMMIT

TWEET ABOUT US:

ADD #SUMMIT AND/OR #JBOSSWORLD TO THE END OF YOUR EVENT-RELATED TWEET

Introduce Red Hat

Create an agenda slide for every presentation.

Outline for the audience what you're going to tell them, and prepare them for a call to action after the presentation.

If this is an internal only presenation, please put **INTERNAL USE ONLY** at the bottom of the master slide.