OpenAFS Diagnostic Tools

Mark Vitale
Sine Nomine Associates
2021 AFS Technologies Workshop

Objectives

- Provide a brief overview of the in-tree OpenAFS diagnostic debugging tools
- Highlight some lesser known capabilities
- Share some tips on when to use (or not to use!) various tools
- Recommend proactive diagnostics

well known diagnostic tools

SINE NOMINE ASSOCIATES

You may already be familiar with the following:

• fs	cache manager interface
• bos, vos, pts	admin tools with some diagnostics
• tokens	show AFS credentials
• translate_et	convert error codes to messages
• rxdebug	”ping” servers and clients, get stats
• cmdebug	show cache manager debug info
• udebug	check database quorum
• xstat_fs_test	collect fileserver stats
• xstat_cm_test	collect Unix cache manager stats

diagnostics that deserve a closer look

- rxping and rxtraceroute — just what they sound like
- rxstats (aka rpcstats) — native Rx RPC metrics

less familiar tools

- fstrace — low level Unix CM trace
- vldb_check, prdb_check — database file integrity
- volinfo, volscan — fileserver partition raw data
- cbd — fileserver callback debug dumps
- state_analyzer — dafileserver fsstate.dat utility
- fssync-debug, dafssync-debug, salvsync-debug — fileserver debugging
- … and many, many more

logging

Logs are the primary source of preliminary diagnostic info.

Most OpenAFS servers support common logging features:

- retrievable via 'bos getlog'
  - no need to login or know the distro log path
- -logfile override default logfile path/name
- -syslog log to syslog
- -d <n> initial logging (debug) level

Logging levels

- (da)fileserver, (da)volserver, ptserver, vlserver only
- -d <level> set at initialization
  - 0 normal mode; errors & warnings, mostly
  - 1 tracing
  - 5 thread numbers; detail tracing
  - 25 debugging
  - 125 detailed debugging
- runtime control:
  - signal SIGTSTP to increase level (up to 4 times)
  - signal SIGHUP to reset to 0

Logging to syslog

- -syslog[=<facility>]
  - default facility LOG_DAEMON (3)
  - default level LOG_INFO (6)
  - bosserver, (da)fileserver, (da)volserver, ptserver, vlserver
- -syslog [-syslogfacility <facility>]
  - (da)salvager, salvageserver
- useful for debugging ubik servers – unified cell log

translate_et

- Translates raw error codes to human readable messages
- Use for codes in log files, command output, audit logs, packet traces, backtraces, core files, etc.

Example 1:

```
$ grep "marked down" VLLog
Thu Dec  6 17:00:03 2020 [0] Server 172.17.16.250 is marked
down due to DISK_BeginFlags code 5376
$ translate_et 5376
5376 (u).0 = no quorum elected
```

Example 2:

```
(seen in an Rx abort packet in a packet trace)
$ translate_et 49733388
49733388 (uae).12 = Unknown code uae 12 (49733388)

(gotcha: not unix errno 12 ENOMEM, but unix errno 13 EACCES)
```

rxdebug

- Useful for all clients and servers
- Check Rx/UDP connectivity via special Rx DEBUG packets (no RPCs - connectionless)
- Get version information
- Get Rx transport statistics
- Get Rx thread waiting and waited info (busy)

```
rxdebug <host> <port> (options)
```

rxdebug basic info

SINE NOMINE ASSOCIATES

Default is to show basic stats and any "interesting" connections.

```
$ rxdebug afs01.sinenomine.net 7000
Trying 207.89.43.111 (port 7000):
Free packets: 2549, packet reclaims: 6435, calls: 211348,
 used FDs: 64
not waiting for packets.
122 threads are idle
0 calls have waited for a thread
Connection from host 67.78.14.74, port 7001,
Cuid bf47aaaf/3444bd18
  serial 2471,  natMTU 1444, security index 0, client conn
    call 0: # 825, state dally, mode: receiving,
flags: receive_done
    call 1: # 0, state not initialized
    call 2: # 0, state not initialized
    call 3: # 0, state not initialized
```

rxdebug basic stats

- Free packets — *number of free packets available for use*
- Packet reclaims — *number of times packets were reclaimed*
- calls — *number of RPC calls successfully made*
- used FDs — *user level file descriptors*
- "not waiting for packets" is normal
  - waiting for packets occurs when program is out of buffers for packets.
- Thread stats
  - ”calls waiting for a thread” — *current value*
  - “threads are idle” — *current value*
  - “calls have waited for a thread” — *running total since init*

rxdebug options

- -version *show AFS version id*
- -rxstats *show Rx statistics*
- -peers *show peers*
- -long *additional info for –peers only*
- filters for rx_connections (and their calls):
  - <default> *"interesting" rx_connections*
  - -noconns *show no connections*
  - -allconnections *show all conns*
  - -nodally *skip conns w/ all calls dally or not init*
  - -onlyserver
  - -onlyclient
  - -onlyhost <host>
  - -onlyport <port>
  - -onlyauth [clear | auth | crypt | null | none | noauth | unauth]

rxdebug -rxstats

SINE NOMINE ASSOCIATES

Displays additional rx stats after the basic info:

```
rxdebug afs01.sinenomine.net 7000 -rxstats -noconns
…
0 calls have waited for a thread    (end basic)
rx stats: free packets 1580, allocs 111141972, alloc-
failures(rcv 0/0,send 0/0,ack 0)
    greedy 0, bogusReads 1 (last from host 268a4b6a),
noPackets 0, noBuffers 0, selects 0, sendSelects 0
    packets read: data 32044362 ack 39349314 busy 0 abort 5319
ackall 0 challenge 1958 response 699 debug 273137 params 0
unused 0 unused 0 unused 0 version 0
    other read counters: data 32044305, ack 39340689, dup
400746 spurious 8132 dally 2
    packets sent: data 71105089 ack 15803096 busy 56 abort 509
ackall 0 challenge 873 response 1958 debug 0 params 0 unused
0 unused 0 unused 0 version 0
    other send counters: ack 15803096, data 71093723 (not
resends), resends 11366, pushed 0, acked&ignored 39482576
        (these should be small) sendFailed 0, fatalErrors 83
    Average rtt is 0.001, with 66693790 samples
    Minimum rtt is 0.000, maximum is 1.311
    5 server connections, 26 client connections, 13 peer
structs, 23 call structs, 22 free call structs
    0 clock updates
```

rxdebug connection info

- Connection from host x *IP address of remote host*
- port *client or server port*
- Cuid *Rx epoch/connection id*
- serial *last packet sent*
- security index - authentication used for this connection
  - 0 no authentication (rxnull)
  - 2 kerberos style authentication (rxkad)
  - 5 rxgk (new in 1.9.x)
- client or server conn
- details for each of the 4 call slots

rxdebug call info

For each of a connection's four call channels:

- call number — *most recent call*
- state
  - not initialized — *available for work*
  - precall — *queued, waiting for a thread*
  - active — *call is running on a thread*
  - dally — *call is idle/waiting for end*
- mode
  - eof — *completed*
  - error — *an error has been received*
  - receiving — *actively receiving data*
  - sending — *actively sending data*
  - unknown — *none of the above*
- flags
  - (see next slide)

rxdebug call info

Call flag values:

- call cleared — *cleared in precall state*
- reader wait — *waiting for next packet from the sender*
- receive done — *finished receiving packets*
- wait packets — *program is out of buffers for packets*
- waiting for process — *RPC is queued*
- window alloc — *RPC waiting for first window on a conn*
- window send — *transmit window full; waiting on an ack*

rxping and rxtraceroute

Ping and traceroute, but for Rx/UDP

- both use slightly modified VERSION packets to probe any AFS component.
- both require root or at least the capability to receive ICMP traffic
- https://gerrit.openafs.org/11907 "Introduce rxping and rxtraceroute" under review
- common options:
  - -host <rx host to ping> *required*
  - -port <UDP port> *default 7000*
  - -n *no reverse DNS lookup*
  - -v4 | -v6
- `rxping -host <host> [-port <port>] [<options>]`
  - -i <secs> *ping interval default 1s*
  - -count | -c <count> *how many ping requests to send*
- `rxtraceroute -host <host> [-port <port>] [options]`
  - -size <nnnn> *default sizeof(rx_header) 28 bytes*
  - -max-hops <n> *default 30*
  - -queries <n> *# of queries per hop (default 3)*
  - -dont-fragment *set dont-frag on outgoing*

cmdebug

- Allows remote debugging of cache manager problems
- With no options, shows contended global locks (rare) and active cache entries (vcaches)
- CAUTION: avoid displaying vcaches for a cache manager that has a large, active working set of AFS files:
  - afsd –stat <nnn> where n> 10000
  - afsd with dynamic vcaches (Linux only, default)

```
cmdebug <cm> [<port>] [<options>]
   • port          defaults to 7001
   • -cellservdb   show current known cells
   • -cache        show (static) cache config
   • -ctime        human readable CB times
   • -callbacks    vcaches with callbacks
   • -refcounts    vcaches with positive refs
   • -long         all locks, all vcaches
```

udebug

- Reports the database server state
  - Which database server is the current sync site
  - Database version number
  - Database quorum state bitmap:
    - 0x01 This machine is the coordinator
    - 0x02 site with the highest (latest) DB version is known
    - 0x04 has a copy of the highest (latest) DB
    - 0x08 DB version number has been updated correctly
    - 0x10 All sites have the highest (latest) DB version
    - normal states: 0x1F or 0x17 (if first write hasn't occurred yet)
  - Voting/election details
  - `udebug <dbserver> <port>`
    - `-long` *additional info for each site (default for syncsite)*

xstat_fs_test

- Lightweight collection of fileserver metrics via RXAFS_GetXStats
- xstat_fs_test <fileserver> [<options>]
  - -collID <n> *metric collection set*
  - -onceonly
  - -frequency <ss>
  - -period <mm>
  - -debug *adds raw output*
    - *useful in case of version mismatch between fileserver and xstat_fs_test*

xstat_fs_test collection ids

- 0 "call info" – not implemented
- 1 "perf info" – extensive counters
- 2 "full perf" – all of collID 1 *plus*
    - RXAFS_* RPC metrics, by RPC:
        - total and "ok" counts
        - service times: min, max, sum, sum of squares
- 3 callback system metrics
    - critical for monitoring and avoiding callback memory exhaustion.

this gets everything:

```
xstat_fs_test <fileserver> -collID 2 3
```

xstat_cm_test

- Lightweight collection of cache manger (Unix only) metrics via RXAFSCB_GetXStats
- xstat_cm_test <cachemgr> [<options>]
  - -collID <n> *metric collection set*
  - -onceonly
  - -frequency <ss>
  - -period <mm>
  - -debug *adds raw output*
    - *useful in case of version mismatch between cm and xstat_cm_test*

xstat_cm_test collection ids

- 0 "call info" internal function call counters
- 1 "perf info" performance counters
  - fileserver and vlserver up/down stats
  - enabled on CM, but disabled in xstat_cm_test!
- 2 "full perf" all of collID 1 *plus*
  - RXAFSCB_* and RXAFS_* RPC metrics, by RPC:
    - total and "ok" counts
    - service times: min, max, sum, sum of squares
  - bonus: RXAFS_* RPC error counts by category:
    - server/network/prot/vol/busies/other
- 3 cache eviction metrics
  - https://gerrit.openafs.org/14200 "afs: provide cache eviction statistics" currently under review

This gets everything currently available:

```
xstat_cm_test <cachemgr> -collID 0 2
```

RXSTATS facility

SINE NOMINE ASSOCIATES

- An extensive system for collecting RPC statistics at the Rx level; implemented automatically in all Rx components at build time by rxgen
- not to be confused with rxdebug –rxstats, or with xstat_cm_test and xstat_fs_test RPC stats
- Now installed by default starting with 1.9.0.
- uses a dedicated RXSTATS service with its own RPCs and thread pool
- multiple dimensions of data granularity:
  - for each Rx service and each RPC
  - by client (issued by component) and by server (received by component)
    - total invocation count; bytes sent, bytes received
    - service times, broken out by application (“exec”) and Rx (“queue”) portions: min/max/sum/sqr
- may be accumulated and reported both for entire server (“process”) and/or each peer (“peer”)

RXSTATS usage

- all "process" options and commands have "peer" equivalents
- query state: rxstat_query_process <host> <port>
- enable at initialization via:
  - -enable_process_stats
  - peer stats enabled by default only in Windows CM
- enable on demand via:
  - rxstat_enable_process <cell> <host> <port> (except cache managers)
  - fs rxstatproc -enable (Unix and Windows CMs)
- clear:
  - rxstat_clear_process <cell> <host> <port> (except cache managers)
  - fs rxstatproc -clear (Unix and Windows CMs)
- collect: rxstat_get_process <host> <port>

rxstat_get_process example

```
$ rxstat_get_process localhost 7001
Process RPC stats for fileserver interface accessed as a client

    RXAFS_FetchData
        Never invoked
    RXAFS_FetchACL
        Never invoked
    RXAFS_FetchStatus
        invoc 14 bytes_sent 224 bytes_rcvd 1680
        qsum 0.000085 qsqr 0.000000 qmin 0.000004 qmax 0.000008
        xsum 0.470860 xsqr 0.023967 xmin 0.023997 xmax 0.120193
    RXAFS_StoreData
        Never invoked
…
Process RPC stats for callback interface accessed as a server
…
    RXAFSCB_InitCallBackState3
        invoc 1 bytes_sent 0 bytes_rcvd 48
        qsum 0.000031 qsqr 0.000000 qmin 0.000031 qmax 0.000031
        xsum 0.000651 xsqr 0.000000 xmin 0.000651 xmax 0.000651
    RXAFSCB_ProbeUuid
        Never invoked

Process RPC stats for volserver interface accessed as a client

    VL_ProbeServer
        invoc 5 bytes_sent 20 bytes_rcvd 0
        qsum 0.000025 qsqr 0.000000 qmin 0.000003 qmax 0.000007
        xsum 51.007456 xsqr 522.372842 xmin 9.421475 xmax 11.016040
…
```

SINE NOMINE ASSOCIATES

proactive diagnostics

- curate logs
- prepare for crashes
  - ensure debug symbols are available
  - enable core processing
- implement realtime monitoring
  - collect AFS metrics periodically and store them in a performance database for graphical reports, forensic research, capacity planning, event monitoring, etc.

monitoring recommendations

- `rxdebug <host> <port> -rxstats -noconn`
  - "calls have waited for a thread", "idle threads"
  - resends *packets resent due to a timeout;*
    - » *ratio of resends to data shouid be low*
  - sendFailed, fatalErrors *could indicate a network error or rx bug*
  - https://gerrit.openafs.org/14358 "rxdebug: Add rxdebug -raw option" under review
- `xstat_fs_test <fileserver> --onceonly -collID 2 3`
  - callback space: GotSomeSpaces, nFEs, nCBs
  - RPC error counts
  - https://gerrit.openafs.org/14359 "xstat: Add the xstat_fs_test -format option" under review
- `rxstat_get_process <server> <port>`
  - client and server RPC metrics for all services: counts, execution and queue times (min, max, total, sum-squared)

Questions and discussion