SINE NOMINE ASSOCIATES

Deploying DAFS

The OpenAFS Demand Attach File Server

Tom Keiser
Sine Nomine Associates

Agenda

- What is DAFS?
- Why should I deploy DAFS?
- What are the deployment issues?
- Deploying DAFS
- Debugging DAFS-specific issues
- Future Directions
- Conclusions

SINE NOMINE ASSOCIATES

WHAT IS DAFS?

Deploying the OpenAFS Demand Attach File Server

Core Features

SINE NOMINE ASSOCIATES

- Host and Callback state are saved/restored across fileserver restarts
- (Nearly) all I/O is now lock-less
  - Namei linktable still requires file locking
- Volume salvages happen on-demand whenever corruption is detected
- Fast startup and shutdown

Salvageserver

SINE NOMINE ASSOCIATES

- Permits fileserver and volserver to schedule asynchronous volume salvage operations
- Maintains a priority queue of scheduled salvage jobs
- Dispatches jobs as workers become available
- Notifies fileserver of job success/failure

Startup/Shutdown

SINE NOMINE ASSOCIATES

- Fast startup
  - now little more than a directory scan
- Fast shutdown
  - Uses a multi-phased highly parallel shutdown algorithm
- Seldom-used volumes are continuously offlined by a garbage collector to reduce shutdown I/O overhead

Want to know more?

SINE NOMINE ASSOCIATES

- Please see my 2006 AFS BPW talk
- Send email to Andrew or I
- Read the AFSLore wiki page
- Read the updated server man pages

SINE NOMINE ASSOCIATES

WHY DEPLOY DAFS?

Deploying the OpenAFS Demand Attach File Server

Performance

- Dramatically faster restart times
- Client cache state is no longer re-initialized following a server restart
  - Eliminates post-restart InitCallBackState storms
  - Eliminates post-restart FetchStatus storms
- Higher throughput due to lock-less I/O
- Reduced operator intervention due to automation of salvaging process

Reliability

- Several metadata corruption modes have been eliminated
  - DAFS detects the need to salvage a volume when it was owned by an ancillary process (e.g. volserver) at the time of crash
- Safer than “fast-restart” mode
  - metadata corruption is detected, and automated attempt(s) are made to correct the inconsistencies, rather than serving data of unknown integrity

Availability

SINE NOMINE ASSOCIATES

- Salvages happen with the fileserver online
  - Volumes in a consistent state are served immediately; those requiring a salvage do so in the background
- Demand attachment is synchronous; demand salvaging is asynchronous
  - For demand salvage, VBUSY/VRESTARTING error codes are sent to client, thereby freeing precious fileserver threads during the latent operation

Serviceability / Observability

SINE NOMINE ASSOCIATES

- Bouncing servers without significant operational consequences is finally possible
- DAFS maintains significantly more granular statistics in the fileserver process
  - various counters and last-event timestamps (e.g. volume operations, demand attaches, VLRU detaches, demand salvages, hash table lookups, disk header loads)
  - They were necessary to support several DAFS self-tuning algorithms; are also available via fssync-debug for use in support of AFS plant analytics

SINE NOMINE ASSOCIATES

DEPLOYMENT ISSUES (POTENTIAL PIT-FALLS)

Deploying the OpenAFS Demand Attach File Server

Bos Bnode Complications

SINE NOMINE ASSOCIATES

- DAFS requires its own bnode type ‘dafs’
- Because of this complication, simple binary replacement is not a viable upgrade strategy
  - Thus, DAFS server binaries have new names:
    - dafileserver, davolserver, dasalvageserver, dasalvager

Host/CallBack State

SINE NOMINE ASSOCIATES

- State has a 30-minute expiration timer
- If server protocol capabilities change across a restart, then the state must be discarded
  - Currently done by manually deleting fsstate.dat
  - Patch in-progress to make this automatic
  - Won't be required in the future; work underway to introduce a new capability exchange RPC

    (draft-tkeiser-afs3-capability-exchange-00)

Volume State

- DAFS introduces a sophisticated finite state machine for each volume
- The volserver RPC protocol (utilized by vos) is only capable of reporting a binary state for a volume (online/offline)
  - Work is underway to fix this deficiency (draft-tkeiser-afs3-volser-tlv-03)
  - Workaround: locally execute fssync-debug

Salvaging

- Salvaging remains a relatively slow operation
  - We (Deason, Meffie, Derrick, and I) have been working on this issue for several years; a threaded salvager is coming (hopefully 1.10 timeframe?)...
- Inode and DAFS are effectively incompatible
  - Salvaging involves a full inode scan, which makes volume group-level salvages extremely inefficient
- vos listvol output no longer sufficient to verify volumes are ok following a server crash

Volume Operations

SINE NOMINE ASSOCIATES

- All volume operations (including salvages) may not begin until the Volume Group Cache (VGC) has been fully populated
- Volume operations which are expected to either be fast, or to fail, may now be highly latent due to demand salvage capability

SINE NOMINE ASSOCIATES

DEPLOYING DAFS

Deploying the OpenAFS Demand Attach File Server

dafs bnode

SINE NOMINE ASSOCIATES

- This new bnode takes four executables:
  - dafileserver
  - davolserver
  - dasalvageserver
  - dasalvager
- This permits operators to seamlessly switch between dafs and non-dafs servers
  - with one important caveat: fsstate.dat consistency must be maintained manually; fix on the way...

Volume Package Tunables

- -vhashsize <log2(hash table size)>
  - Volume hash table size plays a significant role in fileserver performance (especially during startup)
- -vattachpar <number of threads>
  - Controls (max) parallelism of startup and shutdown processes
- -unsafe-nosalvage
  - Provide “fast-restart” semantics with DAFS

VLRU (GC) Tunables

SINE NOMINE ASSOCIATES

- -vlrudisable
  - Disables the VLRU inactive volume detachment thread
- -vlruthresh <minutes>
  - Inactivity timer for VLRU garbage collector
  - Default: 120 minutes
- -vlruinterval <seconds>
  - Determines periodicity of background offline batch jobs
  - Default: 120 seconds
- -vlrumax <number of volumes>
  - Sets maximum number of volumes to offline per batch job
  - Default: 8 volumes

Fileserver State Tunables

SINE NOMINE ASSOCIATES

- -fs-state-dont-save
  - Turns off the saving of fileserver (host/callback) state during shutdown
- -fs-state-dont-restore
  - Turns off the restoration of fileserver state during startup
- -fs-state-verify <none|save|restore|both>
  - Control whether/when state dump integrity checking occurs

SINE NOMINE ASSOCIATES

DEBUGGING DAFS

Deploying the OpenAFS Demand Attach File Server

Debugging Utilities

SINE NOMINE ASSOCIATES

- state_analyzer:
  - debug fsstate.dat fileserver state dumps
- dafssync-debug:
  - communicate with the fileserver via the FSSYNC protocol
- dasalvsync-debug:
  - communicate with the salvageserver via the SALVSYNC protocol

state_analyzer

SINE NOMINE ASSOCIATES

- Provides a means to query/analyze the hosts, authenticated users, files (by fid), and call backs
- Interactive debugger-like command line interface

fssync-debug

SINE NOMINE ASSOCIATES

- This is a deep internal tool, however there are a few subcommands that are benign
  - "query" (dumps the state of a volume)
    - fssync-debug query <volume id> <partition>
  - "header" (dumps the cached volume disk header, if any)
    - fssync-debug header <volume id> <partition>

fssync-debug

SINE NOMINE ASSOCIATES

- "vgcquery" (queries the volume group cache)
  - fssync-debug vgcquery <volume id> <partition>
- "vnode" (dumps the state of a vnode)
  - fssync-debug vnode <volume id> <vnode id> <uniquifier> <partition>
- "volop" (dumps the state of any running volume operation)
  - fssync-debug volop <volume id> <partition>

fssync-debug

- "stats pkg" (queries global volume package statistics)
  - fssync-debug stats pkg

SINE NOMINE ASSOCIATES

FUTURE DIRECTIONS

Deploying the OpenAFS Demand Attach File Server

Potential DAFS Futures

SINE NOMINE ASSOCIATES

- Threaded salvageserver (in-progress)
  - Leverage I/O parallelism by salvaging multiple vnodes at once
  - Potentially eliminate fork() overhead
- Skip marking RO volumes in-use during demand attachment
  - Reduction in attachment I/O overhead
- Live partition attachment/detachment

SINE NOMINE ASSOCIATES

CONCLUSIONS

Deploying the OpenAFS Demand Attach File Server

Conclusions

SINE NOMINE ASSOCIATES

- DAFS provides a significant suite of improvements to the OpenAFS fileserver
- Deployment is relatively straightforward, so long as the migration issues are understood, and appropriate procedures are followed

SINE NOMINE ASSOCIATES

QUESTIONS?

Deploying the OpenAFS Demand Attach File Server

THANKS!

DAFS Development Contacts:

Tom Keiser
Sine Nomine Associates
tkeiser@sinenomine.net

Andrew Deason
Sine Nomine Associates
adeason@sinenomine.net

Michael Meffie
Sine Nomine Associates
mmefie@sinenomine.net