

FreeBSD and Isilon

Zach Loafman, Staff Engineer
zml@freebsd.org



Agenda

- Isilon OneFS
- How Isilon Uses FreeBSD
- Isilon Sponsored Work
- Future Intentions

Updated from MeetBSD '08!

What We Do

- Clustered, scale-out Storage
- Multiple servers operating together to present a single file system image
- Unique, scalable architecture
 - When you need space, just add a node
 - When you need performance, just add a node
- Scales to 144 nodes x 36TB = 5.2PB!
- Ease of use, ease of scaling

What We Do

- Fully clustered architecture
 - No master, every node participates equally
- Data protection is handled by file system
 - RAID handled per file, configurable protection

What We Do

- Each node is accessible via NFS, CIFS, http, ftp, scp, etc.
- Each node is manageable from a web page, ssh, or the console.
- We try to make the cluster appear as a single machine to file access protocols:
 - DNS based load balancing
 - NFS failover

The WOPR



144 nodes, one file system, one namespace!

How breakthroughs begin.™



How We Do It: Hardware

- Multi-core Intel CPUs
- 12-36 SATA/SAS disks per storage node
 - 12 SAS, 12 SATA or 36 SATA options
- Commodity motherboards
- Infiniband for cluster interconnect
- NVRAM for fast journaling
- 10G Chelsio card on accelerators

How We Do It: Open Source

- FreeBSD with custom OneFS module
- Shipping Samba for SMB access (*)
- ... and several other packages: openldap, apache, net-snmp, python, etc.
- Our file system is our core competency. We leverage open source for everything else.

Why did Isilon choose FreeBSD?

- BSD licensing
- Relatively stable APIs
- Optimized, solid network stack
- Friendly developer community

How We Use FreeBSD

- Shipping FreeBSD 6.1 based system
- FreeBSD stable/7 merged
 - In our trunk, not yet shipping
- Many modifications
 - Mostly kernel, some userspace
- Custom VFS module for OneFS

What We've Changed

- SMB support
- NFS improvements
- Infiniband support
- Disk controller improvements

SMB support

- Alternate Data Streams
 - Solaris O_XATTR support
- NTFS ACLs
- Share mode lock support
- Oplocks / Delegations support
- Change notify support
- Atomic open-and-create-ACL call

NFS improvements

- NFS exports changes
 - Transactional export changes
 - More configurable
- File handle affinity
 - Associate each RPC with the nfsd already handling it
 - Integrated as part of sponsored RPCSEC_GSS work
- Fine grained statistics
 - Per-client, per-op stats

Infiniband support

- Mellanox driver
- OpenFabric port
 - 5 years old at this point ☹️
- Sockets Direct Protocol
 - Stream protocol for RDMA fabrics

Work We've Sponsored

- NFS Locking Improvements
 - rpc.lockd rewrite, in-kernel NLM
- RPCSEC_GSS
 - Kerberized NFSv3
- Our general philosophy for sponsorship:
Get it done right, and we'll pick it up later.

NFS Locking

- I thought locking over NFS was fundamentally broken?
 - If everything is working and the stars are aligned, NFS locking is just fine.
 - ... but FreeBSD's rpc.lockd was broken in many ways.
- For a storage appliance, correct protocol locking semantics are critical.

NFS Locking

- Isilon sponsored Doug Rabson to fix it.
- Move the entire service into the kernel
- ONC RPC in the kernel.

- Now present in 6.4, 7.2+, 8.0
- Because of this work, as of OneFS 5.5, Isilon OneFS has fully correct, cluster-coherent NFS locking with failover!

RPCSEC_GSS

- ONC RPC (Sun RPC) has an extensible authentication/authorization scheme
- AUTH_SYS (AUTH_UNIX) is the normal, old style. The RPC contains the uid and a pile of gids.
 - This isn't really authentication.
- RPCSEC_GSS uses GSSAPI for authentication
 - Kerberos is the dominant GSSAPI mechanism, but others are possible
- RPCSEC_GSS is a stepping stone for NFSv4

RPCSEC_GSS

- Isilon sponsored Doug Rabson to build an RPCSEC_GSS implementation for FreeBSD.
- Multithread the new krpc module
- Add RPCSEC_GSS to krpc.
- Make NFS use krpc.
- Now in 8.0, NFSv2-4 use the new RPC

Future Intentions - General

- We need to give back more – FreeBSD is central to our product
- zml@ is now a committer
- Matt Fleming just completed the stable/7 merge
- Now that we're closer, he'll start splitting out pieces to upstream

Future Intentions – SMB

- I'm leading a team working with a new SMB1/SMB2 server
- We chose to keep it in userland, but ..
 - Need something like splice()
 - Need per-thread credentials
- Identity Management
 - We will be Windows SID-pure in our kernel
- See my talk later during EuroBSDCon

Future Intentions – NFSv4

- John Gemingani is leading a team to integrate NFSv4 at Isilon
- Patchwork-quilt to our stable/7 code
- Being very careful to note our differences against current

Future Intentions – NFSv4

- We need to make multi-protocol coherency changes to support NFSv4
- Oplocks/Delegations need a VOP
- Share mode locks in VOP_OPEN
- NFS/SMB byte range locks unified at VOP

Future Intentions – IB

- We intend to start a new OpenFabric port
 - Horizon, not yet in progress
- We're tired of maintaining IB privately..
- ... But too embarrassed with current code

Conclusion

- Hopefully this presentation leaves you with an idea of how at least one vendor has chosen to integrate with FreeBSD.
- Isilon loves FreeBSD
- We wouldn't be here without you
- We intend to give back more

Thanks!



Questions?

How breakthroughs begin.™

