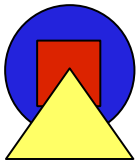
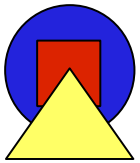


AFS File Servers Dos and Don'ts

Alf Wachsmann

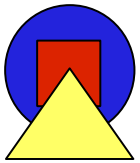
Stanford Linear Accelerator Center

alfw@slac.stanford.edu



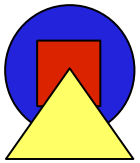
AFS File Server...

- Operating System provides disk space
  `/vicepa, /vicepb, /vicepc, ..., /vicepiu`
  - Should be partitions but can be anything on namei servers
  - Up to 255 vice partitions per server
  - Currently a 2TB size limit per partition
- File server stores AFS volumes inside these partitions
  - No limitation on number of volumes per partition (well, it's 4,294,967,295 and VLDB <2GB)
  - Currently a ~2TB size limit per volume
  - With "namei" file server, "stuff" in partitions is readable
  - With "inode" file server, "stuff" is not easily accessible



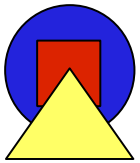
...AFS File Server

- Directories reside in volumes.
- How many files can you have per directory in AFS?
  - You can have 64,000 files in an AFS directory if the filenames are all less than 16 characters long. If the filenames are between 16 and 32 characters than this number decreases. There are 64,000 slots per directory.
    Each file < 16 Characters takes 1 slot.
    Each file > 16 and <32 takes 2 slots, etc...



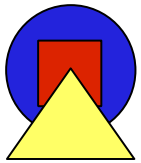
Failure Analysis

- Components that can (and will) fail on an AFS file server:
  - Disk with (parts of) /vicepXY partition(s)
  - Entire RAID array
  - File system with /vicepXY partition(s)
  - File server hardware
- Let's look at each component and understand what AFS can do for you and what you can do for AFS



Single-Disk Failure

- AFS canNOT bundle up a bunch of disks and use them
- Need support from host the disks are connected to (e.g. PVFS (bad idea!))
- Directly connect the disks to computers and make the computers AFS file servers
- Handle disk failures with RAID5 or RAID6
- SAN works well
- Bunch of disks in a computer:
  - Use software RAID5
  - Use LVM
  - Put a journaling file system on the partitions

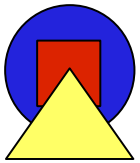


Multi-Disk Failure

- If entire RAID unit fails, AFS clients use RO clones of volumes on other servers
  (doesn't work for RW data)
- Side Note:
  In case you lose a RW volume, a RO volume can easily be converted to a RW volume with:

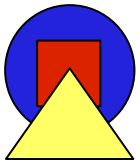
```
vos convertROtoRW -server <machine name> \
                  -partition <partition name> \
                  -id <volume name or ID>
```

  This works only with namei file servers!



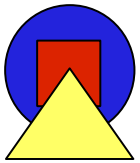
File Systems for /vicepXY...

- /vicepXY should be a **partition** with a file system
- With namei file servers, /vicepXY can be directories or symbolic links
  - Many things are possible:
    - Network File System (NFS, AFS)
    - Clustered File Systems (GFS, Panasas, Ibrix, etc.)
    - Parallel File Systems (PVFS2, GPFS, Luster, etc.)
  - Needs `/vicepXY/AlwaysAttach` to make sure it gets mounted by the file server if it's not a real mountpoint
- These "translations" are NOT A GOOD IDEA
- Use a local file system
- Use a journaled file system



...File Systems for /vicepXY

- There is a 2TB size limit for /vicepXY
- If OS can resize partitions, AFS can use them
- Theoretical limit for AFS volumes is ~2TB
  - Large volumes need more time to move/release
- Files in AFS can be larger than 2GB (with newer AFS server and client software)

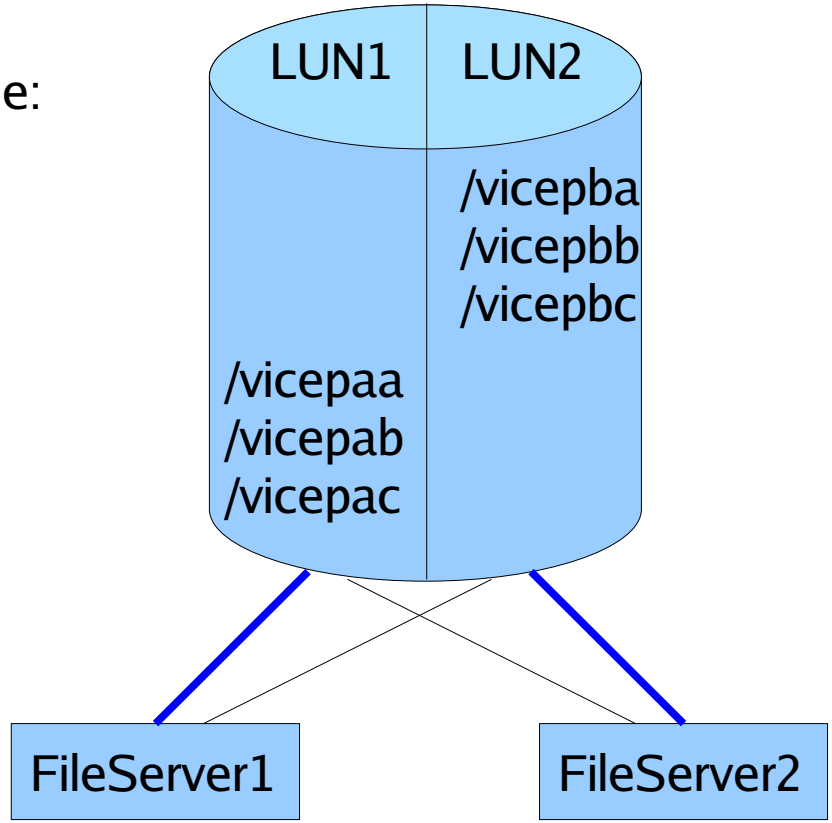


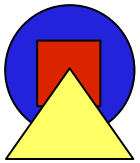
Server Failure

- General advise: use more and smaller AFS file servers instead of few big ones
  - Big in terms of size of the computer and
  - Big in terms of amount of disk space
- Old Transarc recommendation:
  200 AFS clients per AFS file server
- SLAC sees no problem with 2000-3000 clients per server (2CPU Sun V240s, E280Rs; up to 8 partitions with ~60GB each)
- Support for maximum of 104,000 clients per server
- Problem: If server fails, direct attached partitions become unaccessible

Server Fail-Over

Normal Mode:

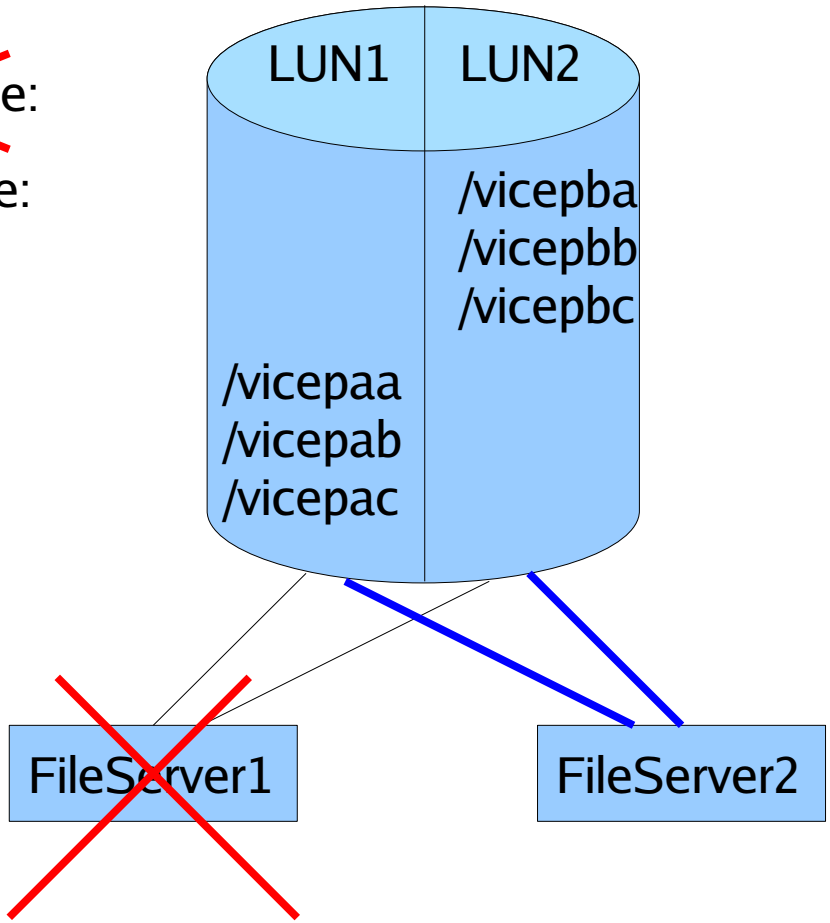


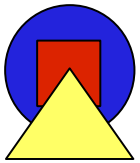


Server Fail-Over

~~Normal Mode:~~

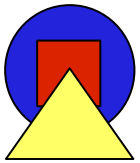
Failure Mode:





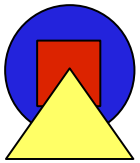
Tolerating Server Failures

- Disk array attached to two servers
- 2 LUNs zoned to be visible to both servers
- First server uses first LUN for its `/vicepaa` - `/vicepaX` partitions
- Second server uses second LUN for its `/vicepba` - `/vicepbY` partitions
- In case first server fails: second server mounts partitions in first LUN
- Run "`vos salvage`" on these "new" partitions
- Run "`vos syncvldb`" on these "new" partitions
- Run "`vos syncserv`" on these "new" partitions



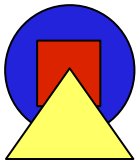
Server Failure Gotcha

- If the file server goes down completely, clients will switch to other replicas and mark RW volumes as unavailable properly
- If the file server is up and answering Rx pings but not serving data, clients will not switch over and instead will just hang
  - Make sure a AFS file server goes **completely** down!



File Server Performance

- Many tweaks possible to improve performance
  - Complicated and very specific to your situation
- Costly operations at AFS file server startup:
  - Running salvage
    - Shut down your file servers cleanly (if possible)
  - Attaching volumes
    - Don't have too many volumes per single file server



Monitoring

- Use the monitoring tools that come with AFS to check on your file servers
  - `xstat_fs_test`
  - `afsmonitor`
  - `scout`
  - `udebug`
  - `AFS:::Monitor` Perl module
- Watch out for bug with AFS on 64 bit Linux systems:
  - Some counters are always "0"
- Use general purpose monitoring tools like Nagios