Vice Partition Virtualization
VPV

Todd DeSantis
John P Janosik
Rajesh Prasad
Indira V Khopde

VPV: What Is It ?

Think of it as a “vos move”, but instead of moving a single volume to a different fileserver, you are moving all vice partitions to a new fileserver.

VPV: The Problem

- AFS Admins have always asked if we could do something to speed up off-loading all volumes from one fileserver to another so maintenance can be performed.
- It can take many days to off-load the volumes and then move them back, making machine maintenance and other things difficult.
- Corporate policy dictates that OS and AFS updates to fileservers must be done when no volumes are on the machine.
- Can this be done ?

VPV: The Answer

- Yes, it looks like this can be done.
- With a series of AFS and OS commands, we can move the vice partitions from one fileserver to another one within 10 to 15 minutes without client jobs failing.
- The two fileservers share a common data storage for the vice partitions. Also, same OS and same hardware.
- Currently, only for IBM AFS versions at this time !!

VPV: The Obstacles

- Can the fileserver still respond to clients
- Will the clients fail waiting for fileserver response
- Can we umount the vice partitions on fs1
- Can we mount the vice partitions on fs2
- Can fs2 start with the vice partitions
- Can the clients find the new location of the volume
- What AFS and OS commands are needed
- Do this without changing the AFS client

VPV Evolution

- AIX has a feature called LPAR Live Partition Mobility (LPM) which allows an Admin to physically move an LPAR to another set of physical resources while it is still running.
- We were asked to allow the fileserver to survive this move.
- Tried it .... the fileserver process failed. Problems accessing volumes and timeouts.
- We had to make changes to the fileserver code.

VPV Evolution...

- LPAR LPM sends a series of signals to all processes on the machine, and any process can stop the move, ignore the signal, or do things so it can survive the move.
- We found that if we put the fileserver into VBUSY mode, that the fileserver process could survive this migration, and client jobs did not fail.
- With this success, we then looked closer at the VPV tasks.

VPV: Can the fileserver respond

- With the success of LPAR LPM, we use a signal to put the fileserver into VPV VBUSY Mode.
- We use code similar to RX Busy Threshold to have the RX layer return VBUSY.
- We have CallPreamble return VBUSY to client RPCs.

VPV: Can the clients survive

- Yes, with the success of the LPAR LPM, we see that VBUSY returns allow the client to hang on until the fileserver is ready.
- The client's afs_Analyze code will see a VBUSY response from the fileserver and sleep 15 seconds and resend the RPC. It will do this 100 times before failing. This is 25 minutes.
- VPV must finish within this time limit !
- We changed the client code to loop 200 times, thus 50 minutes ! But we do not control client upgrades !

VPV: Can we unmount the vicep(s)

- No – not at first. The fileserver and volserver have open file descriptors to the volumes and files.
- What does “vos move” and “bos shutdown” do ?
- We tested putting the fileserver in VBUSY mode and did “bos shutdown” without stopping the fileserver process ..... we could now unmount the vice partitions.
- We created the new “vos detachpart” command to prepare the source fs for VPV.

VPV: vos detachpart

$ vos detachpart -help

```
Usage: vos detachpart -fromserver <machine name of
server from where to detach> -partition <partition
names to be moved> [-cell <cell name>] [-noauth] [-
localauth] [-verbose] [-timeout <timeout in seconds >]
[-help]
```

- Shutting down the volumes takes TIME !!

VPV: Can fs2 mount the vicep(s)

- Yes – there are OS commands to have the SOURCE fileserver umount and forget about the vicep(s).
- And there are OS commands to have the DESTINATION fileserver discover and mount the vice partitions.
- There are zoning commands to protect the vicep(s).
- AFS commands do NOT do this, the Admin must run OS specific commands to allow the 2 fileservers to play nice with the vice partitions !!

VPV: Can the DEST fs start

- Yes – we prefer that the DESTINATION fs be a new fileserver with no volumes or vice partitions before the start of the VPV.
- After discovering and mounting the vicep(s) on the DEST fs, we can start the fs process and it can see and attach the volumes.
- The SRC and DEST fileservers will be the same OS and same hardware.

VPV: Can the clients find the volumes

- Not yet – we have to update the VLDB
- We created a new vos command: “vos attachpart” which will call BreakVolumeCallback and update the volume location in the VLDB.
- The process of updating the VLDB entries for all volumes and calling BreakVolumeCallback for all volumes will take a lot of time !!!

VPV: vos attachpart

$ vos attachpart -help

```
Usage: vos attachpart -fromserver <machine name of
server from where to detach> -toserver <machine name of
server where to attach> -partition <partition names to
be moved> [-cell <cell name>] [-noauth] [-localauth] [-
verbose] [-timeout <timeout in seconds >] [-help]
```

- Calling BreakVolumeCallback for all volumes takes TIME !
- Updating the VLDB entry for all volumes takes TIME !

VPV: What AFS and OS commands

- Two new AFS commands
  - vos detachpart
  - vos attachpart
- AIX Related commands to move the vicep(s)
  - umount / mount
  - exportvg / importvg
  - varyoffvg / varyonvg
  - Maybe other OS hocus pocus !!
- Cabling and zoning on the shared storage.

VPV: Steps Involved

1) Mark fileservser1 as VBUSY manually by sending a signal SIGUSR1 to the file server.

   kill –SIGUSR1 <fileserver1 PID>

2) Run “vos detach” command to detach the partition from fileserver1.

   vos detachpart –fromserver fileserver1 /

   partition <all | partition name> -cell <cell>

3) Unmount all/subset of vice partitions from fileserver1.

   umount / exportvg / varyoffvg ...

4) Mount all/subset of vice partitions on fileserver2.

   varyonvg / importvg / mount

VPV: Steps Involved ....

5) Restart the fs instance on fileserver2

   bos restart fileserver2 -instance fs –cell <cell>

6) Run “vos attach” command to attach partitions to fileserver2.

   vos attachpart –fromserver fileserver1 –toserver fileserver2 \

   -partition <all | partition name> –cell <cell>

7) Unmark fileserver1 from VBUSY manually by sending a signal SIGUSR2 to the file server.

   kill –SIGUSR2 <fileserver1 PID>

8) Shut down the fs instance on fileserver1.

   bos shutdown -server fileserver1 -instance fs –cell <cell>

VPV: Son of a gun – It Works !

- We have used this in several AFS Labs to this point and hope to start using it more as we train the AFS Admins at other sites.
- Our first site uses VPV to do Maintenance on fileserver machines. They create a pod of 4 fileserver machines, 3 active and 1 spare. (fsa fsb fsc fsm). All have unique vicep names.
  - VPV fsa to fsm .... update fsa … VPV fsm to fsa
  - VPV fsb to fsm … update fsb … VPV fsm to fsb
  - VPV fsc to fsm … update fsc … VPV fsm to fsc

VPV: In Production

- The AFS Admins created scripts to do the AFS VPV and OS commands to do the VPV tasks.
- VPV used for maintenance.
- VPV used for retiring machines.
- VPV used for load balancing.
  - We can VPV all or a subset of the vicep(s)
- We can use VPV VBUSY mode for other work on the fs machines

VPV: Limitations

- VPV will take some time to complete
  - Shutting down volumes.
  - umount on SRC, mount on DEST.
  - BreakVolumeCallbacks and update VLDB.
  - 25 minutes before clients will FAIL !
  - New client code will extend to 50 minutes !
  - Don't VPV during normal business hours, users complain waiting for VBUSY volumes !

VPV: Limitations ....

- Suspend all volume activity
  - Move, release, backup, etc.
- If a volume is really busy .... lots of access, maybe move it to another fileserver before starting VPV.
- If a fileserver has many volumes and many vice partitions, you may need to move a subset of vice partitions at a time to stay within the 25 minute limit.

VPV: What's Next

- Try to improve the speed of shutting down volumes, breaking callbacks and updating the VLDB.
- Currently in IBM AFS code, not OpenAFS, merge it into OpenAFS code.
- Can it work with Demand Attached fileserver.

VPV: Questions ?

Thank You