whamcloud

Scalability Test Plan on Hyperion
For the
Imperative Recovery Project
On the
ORNL Scalable File System Development Contract

Revision History

Date	Revision	Author
07/19/2011	Original	J. Xiong
08/16/2011	Rev A	J. Xiong I. Colle (minor formatting edits)

Table of Contents

whamcloud

I. Introduction

The following test plan applies to the Imperative Recovery (IR) project within the ORNL Scalable File System Development contract signed 11/23/2010 and modified on 03/18/2011. ORNL is experiencing exceedingly long service outages during server failover and recovery. Currently this process takes unacceptably long because it depends on timeouts scaled to accommodate congested service latency.

Large-scale lustre implementations have historically experienced problems recovering in a timely manner after a server failure. This is due to the way that clients detect the server failure and how the servers perform their recovery.

From the client side, the only way it can detect the failure of a server is by the timeout of Remote Procedure Calls (RPCs). If a server has crashed, the RPCs to the targets associated with that server will time out and the client will attempt to reconnect to the failed server, or to a failover server if configured. After reconnection, the client will participate in recovery and then continue in normal service.

A restarting target will try to recover its state by replaying RPCs executed but not committed by its previous incarnation. These RPCs must be executed in the original order to ensure the target reconstructs its state consistently. The target must therefore wait for all clients to reconnect before recovery completes and new requests can be serviced because a missed replay transactions could result in recovery failure and a late replay transaction could be invalidated by new requests. This wait, called the “recovery window” is bounded to ensure a failing client cannot delay resumption of service indefinitely.

The above processes are time consuming, as most are driven by the RPC timeout which must be scaled with system size to prevent false diagnosis of server death. This is especially difficult in an environment of ORNL’s scale.

The purpose of imperative recovery is to reduce the recovery window by actively informing clients of server failure. The resulting reduction in the recovery window will minimize target downtime and therefore increase overall system availability.

II. Test cases

Right now there are two recovery types in the system: standard recovery and imperative recovery. We’re going to compare the performance between them under different situations.

There are several control entities under procfs about imperative recovery. This is a list of proc entries we will use during the test.

Procfs entries to be used

NODE	PATH	SETTINGS
MGS	mgs.MGS.live.lustre	Read: print out stats and state of IR [root@wolf6 tests]# lctl get_param mgs.MGS.live.lustre ... Imperative Recovery Status: state: Full, nonir clients: 0 nidtbl version: 5 notify total/max/count: 0/0/3 Write: 0 Clear stats; Write: "status=Disabled" Disable IR; Write: "status=Full" Set IR to full state.
CLIENT	Osc.lustre-OST0000-osc-ffff*.pinger_recov	This can control if standard recovery will be used. If this is disabled, the corresponding OST can be recovered by only imperative recovery. Read: Write: 0 – disable pinger recover; 1 – enable pinger recover.
CLIENT	Osc.lustre-OST0000-osc-ffff*.import	Read: target: lustre-OST0000_UUID state: FULL instance: 1859057999 Check state: FULL to make sure the osc has been reconnected to the ost successfully.

jay 8/22/11 5:52 PM
Deleted: cat lctl lctl lctl

Test cases

Test Case Name	**ir.scalability.ost**
Purpose	To measure and compare how quick the IR can recover a failing OST
Actors	OST, MGS and client
Description	Shutdown and restart an OST in a running cluster, investigate the time for the OST to get recovered. This test case will run with the cases of IR and SR specifically

Sarp Oral 8/22/11 7:26 PM
Comment [1]: 1.For this and the other cases: Does make any sense to test all with and without failover pairs declared?

Replied by Jinshan: No, we don't have such test env. But we have done this test as unit test. The test cases here are focusing on scalability test.

Environment Settings	1. Must have workload on the failing OST. I recommend running IOR on each client node to write objects on this OST; 2. When the system is up, make sure the MGS is on Full state by checking: at the MGS: grep state: /proc/fs/lustre/mgs/MGS/live/{fsname}
Trigger	OST shutdown and restarted
Preconditions	1. MGS is running; 2. all clients have connected; 3. IOR are running on clients
Postconditions	
Special Requirements	Disable SR by: On all clients: lctl set_param osc.{ost import name}.pinger_recov=0 Disable IR by: On the MGS: lctl set_param mgs.MGS.live.{fsname}="status=Disabled"
Assumptions	
Expected Results	Clients should be able to reconnect to the failing target; Investigate the time for all clients to get reconnected to the failing target. We shall calculate the maximum and average recovery time from each client
Notes and Issues	Compare maximum and average recovery time under SR and IR case

jay 8/22/11 5:53 PM
Deleted: and

jay 8/22/11 5:55 PM
Deleted: .

Test Case Name	**ir.scalability.mdt**
Purpose	To measure how quickly the IR can recover a failing MDT
Actors	MDT, MGS and client
Description	Shutdown and restart an MDT in a running cluster, investigate the time for the MDT to get recovered. This test case will run with the cases of IR and SR specifically
Environment Settings	1. Configure the MGS and MDT separately; 2. There must be workload on the failing MDT. I recommend running mdsrate on each client node to create objects on this MDT; 3. When the system is up, make sure the MGS is on Full state by checking: at the MGS: grep state: /proc/fs/lustre/mgs/MGS/live/{fsname}
Trigger	MDT shutdown and restarted

Preconditions	1. MGS is running; 2. all clients have connected; 3. mdsrate is running for a while to make sure there is enough dirty data unflushed on the MDT
Postconditions	
Special Requirements	Disable SR by: On all clients: lctl set_param osc.{ost import name}.pinger_recov=0 Disable IR by: On the MGS: lctl set_param mgs.MGS.live.{fsname}="status=Disabled"
Assumptions	
Expected Results	Clients should be able to reconnect to the failing target; Investigate the time for all clients to get reconnected to the failing target. We shall calculate the maximum and average recovery time from each client.
Notes and Issues	Compare maximum and average recovery time under SR and IR case

Test Case Name	**ir.scalability.multiple_targets**
Purpose	To make sure imperative recovery works well if multiple targets fail meanwhile
Actors	OST, MGS and client
Description	Suppose there are tens of OSTs on each OSS, and restart as many OSSes as possible on the same time, to investigate if IR can handle this case smoothly. This test case will run with the cases of IR and SR specifically
Environment Settings	N/A
Trigger	OSSes shutdown and restarted
Preconditions	MGS is running and all clients have connected
Postconditions	
Special Requirements	Disable SR by: On all clients: lctl set_param osc.{ost import name}.pinger_recov=0 Disable IR by: On the MGS: lctl set_param mgs.MGS.live.{fsname}="status=Disabled"
Assumptions	
Expected Results	Clients should be able to reconnect to all restarting OSTs in a short time
Notes and Issues	Compare maximum and average recovery time under SR

	and IR case

Test Case Name	**ir.scalability.scalability**
Purpose	To verify that imperative recovery can notify many clients in a reasonable time
Actors	OST, MGS and client
Description	Mount thousands of mount points on each client node, and then restart an OST; This test case will run with the cases of IR and SR specifically
Environment Settings	N/A
Trigger	OSSes shutdown and restarted
Preconditions	1. MGS is running and all clients have connected 2. Make sure you have this line in your modprobe.conf file: options obdclass lu_cache_percent=1, otherwise, you can't mount over 200 mountpoints on one node 3. Clear the IR stats by: at the MGS node: lctl set_param mgs.MGS.live.{fsname}=0
Postconditions	
Special Requirements	Disable SR by: On all clients: lctl set_param osc.{ost import name}.pinger_recov=0 Disable IR by: On the MGS: lctl set_param mgs.MGS.live.{fsname}="status=Disabled"
Assumptions	
Expected Results	Clients should be able to reconnect to the restarting OST in a short time. Collect the output of lctl get_param mgs.MGS.live.{fsname}
Notes and Issues	One of the IR's problems is that the MGS has to notify each clients for the update of target register, so it may have the scalability problem if there are two many clients in the cluster. This test is to verify IR works well with at least 100K clients. Compare maximum and average recovery time under SR and IR case

Sarp Oral 8/22/11 7:29 PM

Comment [2]: A version of this test case with ongoing I/O to other OSTs while the tested OST is restarted/tested, would be nice to have.

Replied by Jinshan: Since we don't have so many nodes, so we have simulate 75K clients by mounting 600 mountpoints on each client node. In this case, I don't think it will make much sense to add workload. Actually serious scalability test has still to be done at ORNL.

Sarp Oral 8/22/11 7:33 PM

Comment [3]: Is the CPU load on the MGS measured during this test?

Replied by Jinshan: in our test, the MGS can notify 75K clients in a really short time, 8 seconds on average(no workload in cluster), so we didn't catch CPU load on the MGS. What I can verify is that the CPU load on the MGS is high by a glimpse.