Increasing current Lustre availability to 99.9% with a backup Metadata Server

Lustre is a highly scalable cluster file system consisting of object storage targets, client file systems and a single metadata server. However as there is only one metadata server in the current architecture, it becomes a single point of failure, which may result in an outage. We have introduced another metadata server in the architecture, which continuously monitors the main metadata server. In case of a failure, the backup metadata server takes over all operations increasing the system's availability to 99.9%. In this article we characterize our system by listing down the different types of outages our system may be subject to and the ones it is resistant to. We have calculated the total outage time for one year based on the mean time between failures and mean time to recover for various system components and arrived at Class 3 classification for the system.

1. Outages the system will be subject to

Unplanned Outages:

1. *Physical*
   a. Network Card Failure
   b. Physical Connection (Ethernet) Failure
   c. Switching Element Failure
   d. Processor Failure
   e. RAM/Cache Failure
   f. SCSI Disk Failure/ I/O interconnect
   g. Buses, Battery, Fan
2. *Design*
   a. Lustre Filesystem/OS Error
   b. Software Design Error
   c. Hardware Design Error
3. *Environmental*
   a. Loss of Power
   b. Unfavorable (hot) conditions leading to loss of cooling
4. *Operator*
   a. Inexperienced or Malicious Operator/User action
   b. Accident

Planned Outages:

1. *Upgrading*
   a. Installing/Upgrading Software
   b. Upgrading the hardware
2. *Maintenance*
   a. Backups
   b. Data Reorganization

Disaster:

1. *Natural*
    a. Earthquake/Flood/Hurricane
    b. Fire
2. *Forced*
    a. Terrorist Attack/War

2. The HAMDS system will prevent outages due to the following faults

1. Network card failure of the main MDS
2. Ethernet failure of the main MDS
3. Processor failure for main MDS
4. RAM/Cache failure for main MDS
5. Buses, battery, fan for main MDS
6. Filesystem/OS error
7. Software design error
8. Hardware design error
9. Power failure
10. Loss of cooling due to hot and unfavorable conditions
11. Inexperienced/malicious user/operator causing main MDS to crash
12. Software installation/upgradation
13. Upgrading the Hardware (except SCSI)

3. Specification of the MTBF and the MTTR of system components

MTBF = mean time between failure
MTTR = mean time to recover
MTTF = mean time to failure

MTBF = MTTF + MTTR
% Availability = (MTTF) / (MTTF + MTTR) * 100

Outage	MTBF (hours)	MTTR (minutes)
Realtek NIC card [1]	50000 – 100000	15
CAT5 Ethernet Cabling	100000 – 200000	10
Fast Ehernet Switch [2], [3]	200000 – 500000	60
Processor [8]	70000-100000	500
RAM/Cache [4]	200000 – 500000	60
SCSI disk [5], [6]	100000 – 300000	300
Linux Crash [7]	4360 – 8760	240
Data Organization on SCSI	4380	120

4. Calculation of total down time in 1 year

- If either of the NIC cards fails independently then they can be replaced without the system going down as the backup MDS will take over. However an outage will occur if a NIC card fails and the second one also fails before the first one could be replaced. As seen from above data, the mean time to recover from a failure for a NIC card is 15 minutes, i.e. we are guaranteeing to replace a NIC card within 15 minutes of its failure. Given that the mean time between failures for a NIC card is 100000 hours, the probability for one NIC card not to fail in a time interval of 15 minutes is given by:

  Availability (A') *= (100000 * 60 – 15) / (100000 * 60) * 100 = 99.99975%.*
  (for one NIC card) *……..(1)*

  The system will be down if both the NIC cards are down at the same time. The probability 'P' of both cards being down at the same time is given by:

  P = *0.0000025 * 0.0000025 = 6.25 * 10^ -12* *……..(Using result (1))*

  Actual Availability (A) *= (1 – 6.25* 10^ -12) * 100 = 99.9999999999%.* *……..(2)*

  Therefore using the result obtained in (2), the downtime (D) for the system in one year in seconds is:

  D = *365 * 24 * 60 * 60 * (1- 0.999999999999) = 31.5 microseconds*. ……..(**I**)
  (which is negligible)

- Similarly the downtime due to ethernet cabling, processor and RAM will be negligible as the probability of resources of both the main server and the backup machine at the same time will be extremely small *tending to 0.*

- The downtime due to ethernet switch crash is given by:

  A *= (200000 * 60 – 60) / (200000 * 60) * 100 = 99.9995%.*
  D *= 365 * 24 * 60 * 60 * (1 - 0.999995) = 157.68 seconds* ……..(**II**)

  The two servers will share the SCSI disk in our system. The data is not replicated and this makes the SCSI a single point of failure. In order to provide some resistance against SCSI failure and also disaster recovery, the SCSI disk will be backed up every 24 hours. However in this case, all the metadata will not be recovered and the system can only be restored to the last stable state when the system was backed up.

- The downtime due to SCSI disk failure will be longer because both the servers share the SCSI disk and when it crashes, then there will be an outage.

  Assuming the MTBF for SCSI disk to be around 100000 hours (lower bound) and the time to recover as 300 minutes, the availability for the disk is given by:

  A *= (100000 * 60) – 300 / (100000 * 60) * 100 = 99.995%*
  D *= 365 * 24 * 60 * (1-0.99995) = 26.28 minutes* ……..(**III**)

- The downtime due to operating system(Linux) crash is given by:

  $$\boldsymbol{A'} = (4360 * 60 - 240) / (4360 * 60) * 100 = 99.9\%.$$
  $$\boldsymbol{A} = (1- ((1 - 0.999) \wedge 2)) * 100 = 99.9999\%$$
  $$\boldsymbol{D} = 365 * 24 * 60 * 60 * (1 - 0.999999) = 31.536 \text{ seconds} \qquad \ldots\ldots\ldots (\textbf{IV})$$

- Occasionally the data on the SCSI disk will have to be reorganized. Operations like defragmentation and repartitioning along with backup will have to be performed. The time span between these data maintenance activities will be 6 months. The downtime due to data reorganization is given by:

  $$\boldsymbol{A} = (4380 * 60 - 120) / (4380 * 60) * 100 = 99.95\%.$$
  $$\boldsymbol{D} = 365 * 24 * 60 * (1 - 0.9995) = 262.8 \text{ minutes} \qquad \ldots\ldots\ldots (\textbf{V})$$

- The planned outages (except data reorganization and backups) do not result in an outage as the backup MDS can take control when changes are being done on the main MDS.

- The metadata servers will be provided with a UPS to provide protection against the occasional power failure. A spare generator can be used in case of a power crisis. As an upper bound, we consider *30 minutes* as the time when the system will be down due to power failure. ………(**VI**)

- The metadata servers will be located in Pune. In the last century Maharashtra has experienced only one major earthquake in Koyna, flood in Panchet 1961, plague in 1896-97[9]. Hence the probability of a disaster occurring is remote. We are providing *Tier 1* disaster recovery scheme, i.e. the metadata is going to be backed up every 24 hours and stored at a remote location. The time taken to set up the system again would be around *90 minutes.* ………(**VII**)

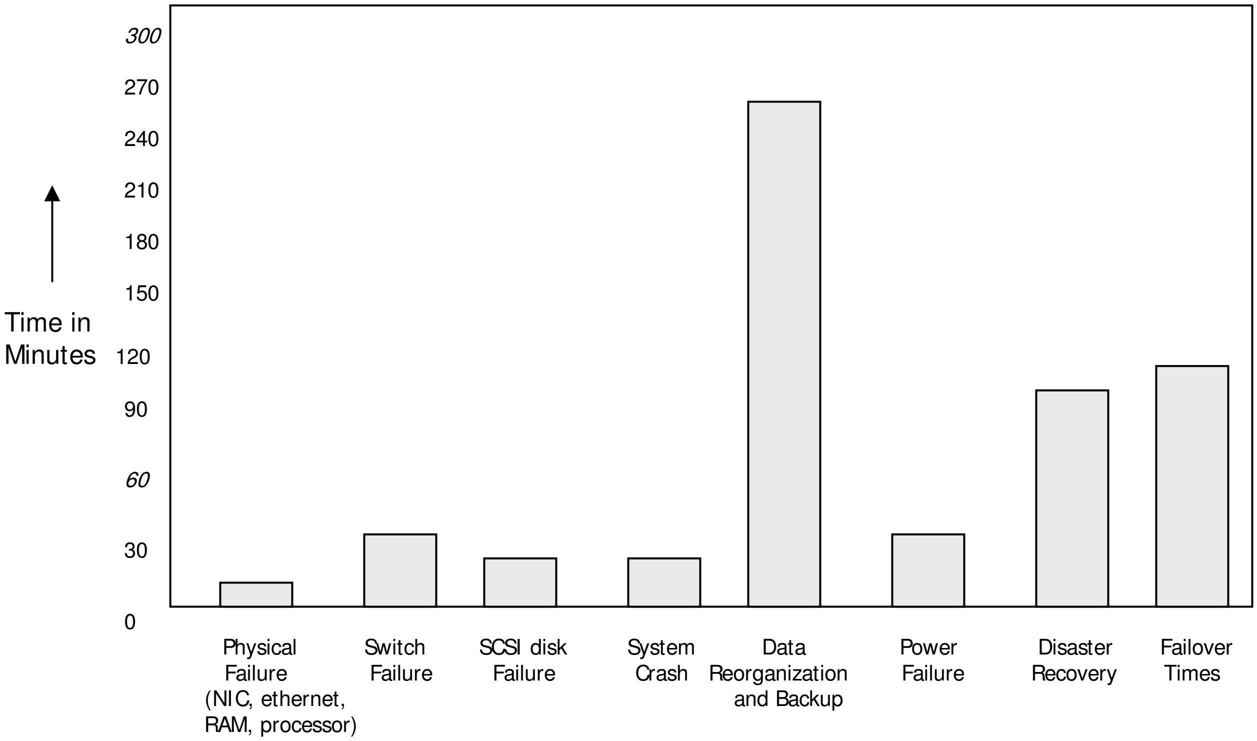
- The failover time will be *30 seconds*. The breakup of this time is *10 seconds* for detecting that the MDS1 has failed, *10 to 12 seconds* for reconstructing the state and the lock data, *10 seconds* for creation of various processes by MDS2 and resumption of execution. This will be a rare occurrence and hence the total outage time due to failovers will not be more than *100 minutes* in a year. ………(**VIII**)

The failure detection module will take 10 seconds to detect that that MDS1 has failed. The MDS1 will send RPC packets (heartbeat) to the MDS2 after 1 second interval. If three timeouts occur then MDS2 will send liveness check packets to MDS1 after a gap of 1 second checking whether it has gone down. If MDS1 does not reply even after 3 timeouts, then MDS2 will start sending liveness check messages to MDS1 over the SCSI port connection between the two servers after every 1 second. If 3 timeouts occur then MDS2 concludes that MDS1 has failed and initiates the failover process. During this interval of 10 seconds, MDS2 may also detect two other types of failures apart from the MDS1 failure. If during the first 3 seconds, no heartbeat messages reach MDS2 then it starts the liveness check module. If during the next 3 seconds, MDS2 realizes that it is unable to send RPC messages to the MDS1 then there is some problem in MDS2's NIC card or the ethernet cabling. If MDS2 is able to send RPC messages to MDS1 but still does not get any response from MDS1 then it proceeds to sending messages on the SCSI port. If communication is possible over this cable and MDS2 gets a reply from MDS1 then it concludes that the network is down otherwise if MDS2 does not get any reply from MDS1 over the SCSI port even after 3 seconds then it concludes that the MDS1 is down.

The MDS2 now starts of the failover process. It starts reading all the state data and lock data logged by MDS1 on the shared SCSI and reconstructing the state. It brings itself to the same state with exactly the same processes in memory in which the MDS1 was before failure. This procedure would take around 10 to 12 seconds. In the last 10 seconds it would reestablish connections with the various clients and resume execution of various operations.

5. The distribution of outage times in a year

The following bar diagram illustrates the results obtained from the calculations in Steps I to VIII. The time in minutes is plotted on the Y - axis while the different failures are plotted on the X - axis.



6. Class to which the system belongs

Availability	Total Outage per Year	Class No. of # 9
90%	36.5 days	1
99%	87.6 hours	2
99.9%	8.76 hours	3
99.99%	52.56 minutes	4
99.999%	5. 256 minutes	5
99.9999%	31.536 seconds	6
99.99999%	3.1536 seconds	7

The total outage time in minutes as per the above bar diagram is given by:

Outage *= (1) + (2.628) + (26.28) + (0.5) + (262.8) + (30) + (90) + (100)*

Thus the downtime in a year is *513.208 minutes*, *i.e.* ***8.5534 hours***.

From the above table [10] we can see that the HAMDS for Lustre filesystem belongs to ***Class 3.***

References:

[1] http://adaptec.com, Adaptec Duo64, "ANA™-62022 Two-Port, 64-Bit PCI Network Interface Card for Fast Ethernet Environments."

[2] http://www.sixnetio.com/html_files/products_and_groups/mtbf.htm, SIXNET MTBF for fast ethernet switches.

[3] http://www.starmicrotech.com/index.cfm, Cisco Catalyst 2950G-24 24 port Intelligent Ethernet Switch.

[4] http://www.synchrotech.com/

[5] RK05/DIABLO/PERTEC DRIVES UPGRADED TO SCSI AEM-5C Cartridge Disk Replacement

[6] http://www.hardwareanalysis.com/

[7] http://gnet.dhs.org/stories/bloor.php3, Bloor Research, "Why Linux is better than Windows".

[8] http://www.itox.com/pages/products/mothers/370/gcs15.cfm

[9] Nanda Dabhole Kasabe, "A record of growth and triumph over disasters". November 8, 1999, The Indian Express.

[10] Gregory F. Pfister, "In Search of Clusters" 2nd Edition, 1998, Prentice Hall.