The Premier Advanced Recording Technology Forum

THIC Inc.

Los Alamos NATIONAL LABORATORY

Performance Loss and Reliability in a Petabyte RAID-Based Parallel File System

Ian Philp

Los Alamos National Laboratory

Los Alamos, NM 87545

Phone: +1-505-665-1150 FAX: +1-505-665-6333

E-mail: philp@lanl.gov

Presented at the THIC Meeting at the National Center for Atmospheric Research, 1850 Table Mesa Drive, Boulder CO 80305-5602

June 29-30, 2004

System Parameters

TFLOPS	year	MEM (TB)	Storage (TB)	Storage Bandwidth (GB/sec)	Min # disks (capacity)	Min # Disks (bandwidth)
10	2002	10	200	10	2800(72GB)	500(20MB/sec)
100	2005	100	2000	100	8400(250GB) 4200(500GB)	3400(30MB/sec)
200	2005	200	4000	200	16800(250GB) 8400(500GB)	6800(30MB/sec)

Issues with RAID

- Purchase of disks will be driven by minimizing $ per GB
  - we do not yet need to buy extra disks to satisfy the bandwidth requirement (but the gap is closing)
- Is the reliability of RAID-5 good enough
  - do we need more redundancy
- Bandwidth lost due to rebuilds
  - how much is lost and is this acceptable
  - can a RAID system be built that tolerates disk failures without degrading performance

ASCI QA & QB

- 1024 Nodes = 1024x(4P/node)=4096P
- 128 IO Nodes
- 4 RAIDs per node (5+1 RAID-5)
  - 128 nodes x 4 RAIDs/node = 512 RAIDs
  - 512 RAIDs x 6 disks/RAID = 3072 total disks
- Each disk: 72GB Fibre Channel
  - 72GB/disk x 3072 disks = 216TB (180TB data)
- FC disks: 34 failures Jan.-March
  - (3072x2)x(90x24)/34= 390K hrs mttf/disk
- SCSI disks (2700 disks, 30 failures)
  - (2700)x(90x24)/30 = 200K hrs mttf/disk

Parallel File System Usage

- Checkpoint/restart dumps
  - A single appl can use 50% of the machine and dump 50% of its memory
    - 200 TB * 25% = 50 TB
    - 200GB/sec => dump takes 4.5 minutes
    - Requires 6,800 disks @ 30 MB/sec/disk
- Visualization
  - 30 frames/sec * 1.6 GB/frame = 48 GB/sec
  - Needs to be done in real-time
  - Requires 1,640 disks @ 30 MB/sec/disk

(k+1) RAID Fail Rates

$\overline{C} = 216TB$

$C = 72GB$

$b = 20MB/\sec$

$N = 3072$

$k = 5$

DLP = Data Loss Probability

mttf(hrs)	R(hrs)	(k+1) DLP/yr
200,000	1	5.2e-3
400,000	1	1.3e-3
200,000	10	5.0e-2
400,000	10	1.3e-2

(k+1) 1st Disk Rebuild

$mttf = 400{,}000hrs$

$b = 30MB/\sec$

$k = 5$

$\overline{C}$ (PB)	C (GB)	N	R (hrs)	(k+1) DLP/yr	fraction of bandwidth lost
2.4	500	5000	4.7 24	9.7e-3 4.8e-2	0.06 0.26
4.8	500	10000	4.7 24	1.9e-2 9.3e-2	0.11 0.55

More Redundancy

- RAID-5 is (k+1) redundancy
  - if 2 or more disks fail in interval R, the RAID fails (loses data)
  - serious performance degradation during disk rebuilds
- (k+2) redundant RAID
  - requires 3 disks to fail in interval R
  - Reed-Solomon, EVENODD, RM-2, RDP

K+2 Redundancy

- (k+2) redundant RAID
  - can we delay the rebuild of a failed disk until a preventative maintenance (P.M.) period
    - if so, we avoid performance degradation
    - if 2 disks fail, we combine their rebuilds in the P.M. period
    - system performance improves as system degrades!
  - another possibility: initiate rebuilds as soon as the 2nd disk fails

Names

	(k+1)	(k+2)
1st disk rebuild	√	√
2nd disk rebuild	data loss	√
P.M. rebuild	√	√
2nd disk +P.M. rebuild	data loss	√

(k+1) (k+2) Fail Rates

$mttf = 400{,}000hrs$

$\overline{C} = 216TB$

$C = 72GB$

$b = 20MB/\sec$

$N = 3072$

$k = 5$

R	(k+1) DLP/yr	(k+2) DLP/yr	(2k+2) DLP/yr
1hr	1.3e-3	3.3e-8	1.1e-7
10hrs	1.3e-2	3.3e-6	1.0e-5
1mo	2.6e-1	1.3e-3	4.0e-3

(k+1) (k+2) Fail Rates

$mttf = 400{,}000hrs$

$\overline{C} = 2.4PB$

$C = 500GB$

$b = 30MB/\sec$

$N = 5000$

$k = 5$

R	(k+1) DLP/yr	(k+2) DLP/yr	(2k+2) DLP/yr
4.7hrs	9.7e-3	1.2e-6	3.7e-6
24hrs	4.8e-2	2.9e-5	9.1e-5
1mo	3.9e-1	2.1e-3	6.5e-3

(k+1) (k+2) Fail Rates

$mttf = 400{,}000hrs$

$b = 30MB/\sec$

$k = 5$

$\overline{C}$ (PB)	C (GB)	N	R	(k+1) DLP/yr	(k+2) DLP/yr	(2k+2) DLP/yr
2.4	500	5000	4.7hrs 1mo	9.7e-3 3.9e-1	1.2e-6 2.1e-3	3.7e-6 6.5e-3
4.8	500	10000	4.7hrs 1mo	1.9e-2 6.3e-1	2.3e-6 4.2e-3	7.4e-6 1.3e-2

(k+2) With 2nd Disk Rebuild

- What happens in a (k+2) scheme
  - ignore 1st disk failure
  - initiate dual rebuild immediately upon 2nd failure
  - Rebuild all failed disks during P.M.
  - R_{PM}: P.M. rebuild (P.M. interval)
- How much bandwidth is lost ?
  - depends upon # of double fails within a RAID between maintenance periods, i.e., the (k+1) P.M. rebuild fail rate

2nd Disk Rebuild

$\overline{C} = 2.4PB$

$C = 500GB$

$b = 30MB/\sec$

$R = 4.7hrs$

$R_D = 1mo$

$N = 5000$

$k = 5$

$mttf = 400{,}000hrs$

Data Loss Probability/year

(k+1) 1st disk	(k+2) 1st disk	(k+2) delayed (1mo)	(k+2) 2nd disk	(k+2) 2nd+P.M.
9.7e-3	1.2e-6	2.1e-3	6.8e-4	6.1e-5

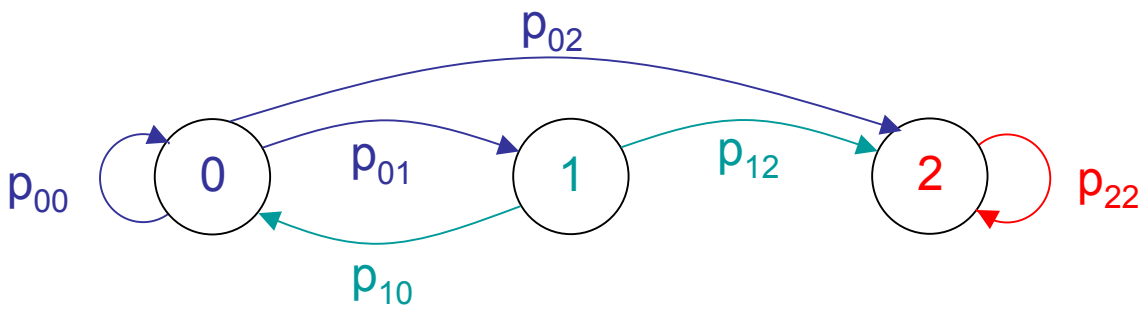
Fraction of Bandwidth Lost

$\overline{C}$ (PB)	C (GB)	N	R (hrs)	(k+1) 1st rebuild	(k+2) 2nd rebuild	(k+2)2nd+P.M.
2.4	500	5000	4.7	0.06	0.007	0.00044
4.8	500	10000	4.7	0.11	0.013	0.00088

Summary

- RAID-5 (k+1) might be reliable enough
  - But only if application is blocked out during rebuilds
- Bandwidth loss (due to rebuilds) is likely to be a problem
- Can we avoid bandwidth degradation with (k+2) redundancy and scheduled rebuilds
- Overhead of 2-D parity vs. overhead of (k+1) disk rebuilds

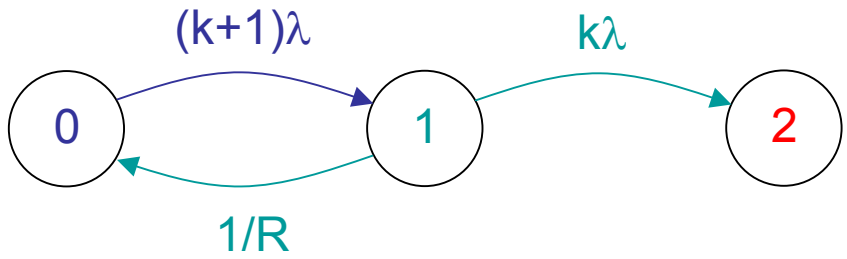
Discrete-Time Model



$$\begin{bmatrix}\pi_0 & \pi_1 & \pi_2\end{bmatrix}^{(t+1)} = \begin{bmatrix}\pi_0 & \pi_1 & \pi_2\end{bmatrix}^{(t)} \begin{bmatrix} p_{00} & p_{01} & p_{02} \\ p_{10} & p_{11} & p_{12} \\ p_{20} & p_{21} & p_{22} \end{bmatrix}$$

$$\begin{bmatrix}\pi_0 & \pi_1 & \pi_2\end{bmatrix}^{(0)} = \begin{bmatrix}1 & 0 & 0\end{bmatrix}$$

Continuous-Time Model



$$\frac{d}{dt}X_0(t) = X_1(t)\frac{1}{R} - X_0(t)(k+1)\lambda$$

$$\frac{d}{dt}X_1(t) = X_0(t)(k+1)\lambda - X_1(t)\left(k\lambda + \frac{1}{R}\right)$$

$$\frac{d}{dt}X_2(t) = X_1(t)k\lambda$$

Continuous-Time Model