Hierarchical RAID (HRAID): Organization, Operation, Reliability and Performance

Alexander Thomasian and Yujie Tang
Shenzhen Institutes of Advanced Technology (SIAT)
Chinese Academy of Sciences
E-mails: alexthomasian@gmail.com and yj.tang@siat.ac.cn

Outline

1. **Introduction to RAID**
2. HRAID Motivation
3. Performance
4. Reliability Analysis
5. Design Study
6. Future Work
7. Conclusion

RAID Levels

- RAID levels 0, 5, 6, 7 are ℓ=0,1,2,3 disk failure tolerant (ℓDFT), hence RAID($\ell+4$), $\ell \geq 1$
- Only Maximum Distance Separable (MDS) codes considered, i.e., RAID1 (mirrored disks) excluded
- RAID0-0DFT: Data striping — no redundancy
- RAID5-1DFT: Single rotated parity to deal with single disk failures or sector errors
- RAID6-2DFT: Two rotated check blocks with Reed Solomon (RS) coding
- RAID7-3DFT: RS coding

- RAID6 tolerates Latent Sector Errors (LSEs) encountered during rebuild
- EVENODD (EO): Blaum et al. IBM, ISCA'94
- Rotated Diag
  NetApp FAST'(
- X-code: vertica
  Xu and Bruck'99

1. EO and RDP computationally less expensive than RS coding;
2. Both have same disk access pattern as RAID6 with small symbols.

- EO extensions: Blaum et al. 2002
  STAR by Huang and Xu 2008
- RDP extension by Blaum 2006

RAIDℓ Operation in Degraded Mode

- Capability to tolerate $\ell = 1, 2, 3$ disk failures and sector errors
- Reconstruct $n \leq \ell$ blocks on failed disks by XORing $N - \ell$ corresponding blocks
- The disk read load for $n = 1, 2, 3$ disk failures on RAIDℓ, $1 \leq \ell \leq 3$ increases by a factor $n + 1$
- Read response time affected, even if processed at higher priority than writes

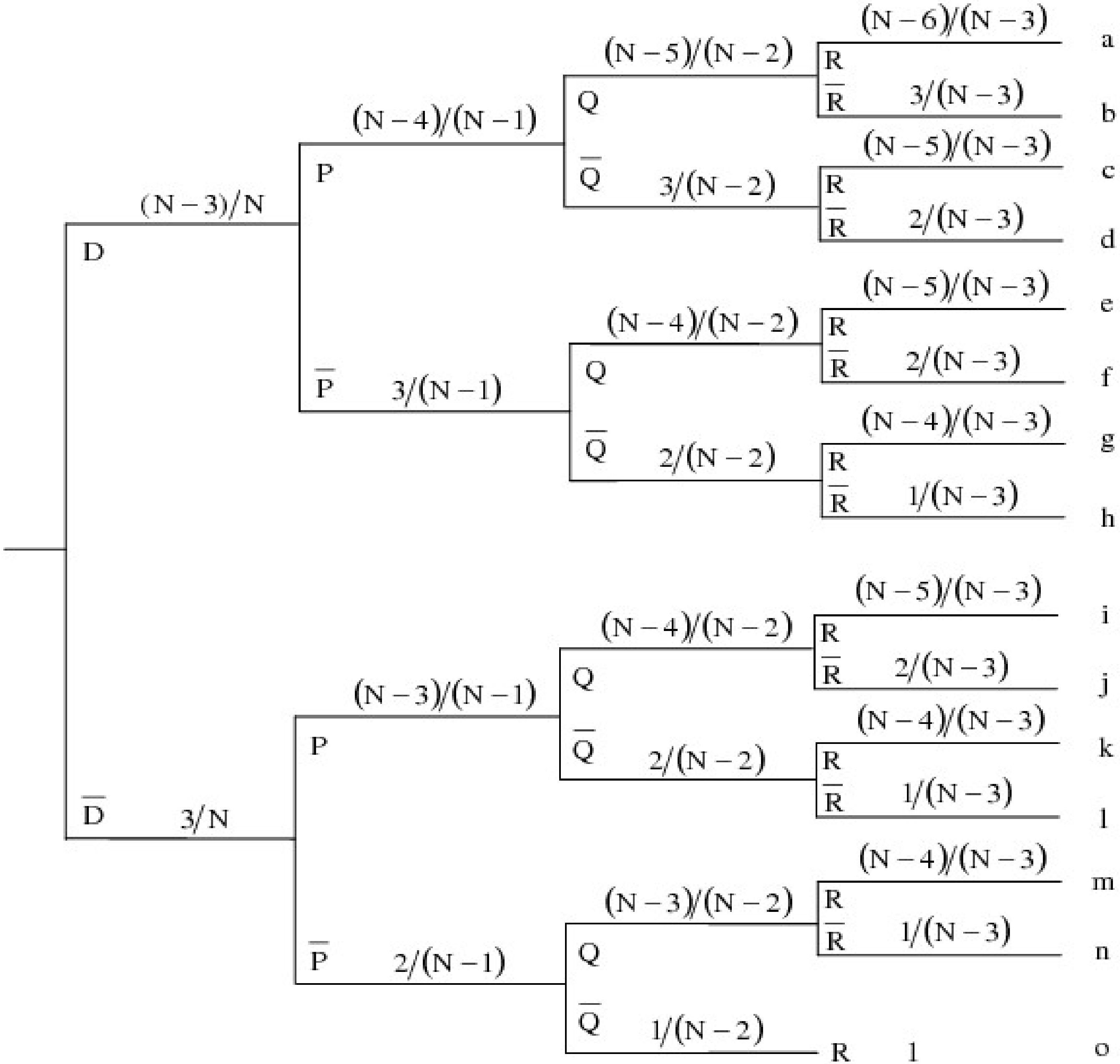


Fig. 1 Decision tree to obtain access costs in RAID7 with N disks and three disk failures.

RAID7 Performance with 3 Failed Disks

- Reconstruction costs with $C = (N-3)D_{SR}$ are:
- (a) no disk failures: $4D_{RMW}$
- (b,c,e) 1 unavailable check block: $3D_{RMW}$
- (d,f,g) two unavailable check blocks: $2D_{RMW}$
- (h) 3 unavailable check blocks: D_{RMW}
- (i) only data block unavailable: $C + 3D_{SW}$
- (j,k,m) data and 1 check block unavailable: $C + 2D_{RMW}$
- (l,n,o) data and 2 check blocks unavailable: $C + D_{RMW}$

Repair Options

- Dedicated sparing: spare disk bandwidth wasted
- Distributed sparing: disk bandwidth not wasted
- Parity sparing or restriping: check blocks used as spare areas
- RAID7 $\longrightarrow$ RAID6 $\longrightarrow$ RAID5 $\longrightarrow$ RAID0

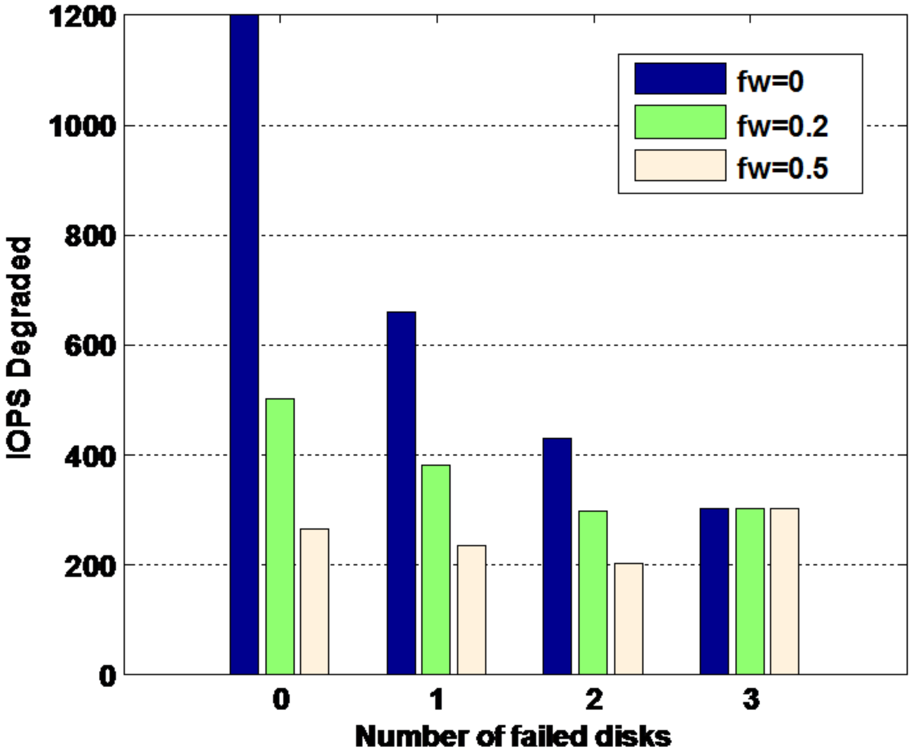
  No disk failures: $\{D_1, D_2, D_3, D_4, \ldots, P, Q, R\}$

  D_1 fails: $\{-, D_2, D_3, D_4, \ldots, P, Q, D_1\}$

  D_2 fails: $\{-, -, D_3, D_4, \ldots, P, D_2, D_1\}$

  D_3 fails: $\{-, -, -, D_4, \ldots, D_3, D_2, D_1\}$
- Repairs restricted by check strips

Degraded/Restriped RAID7 Performance



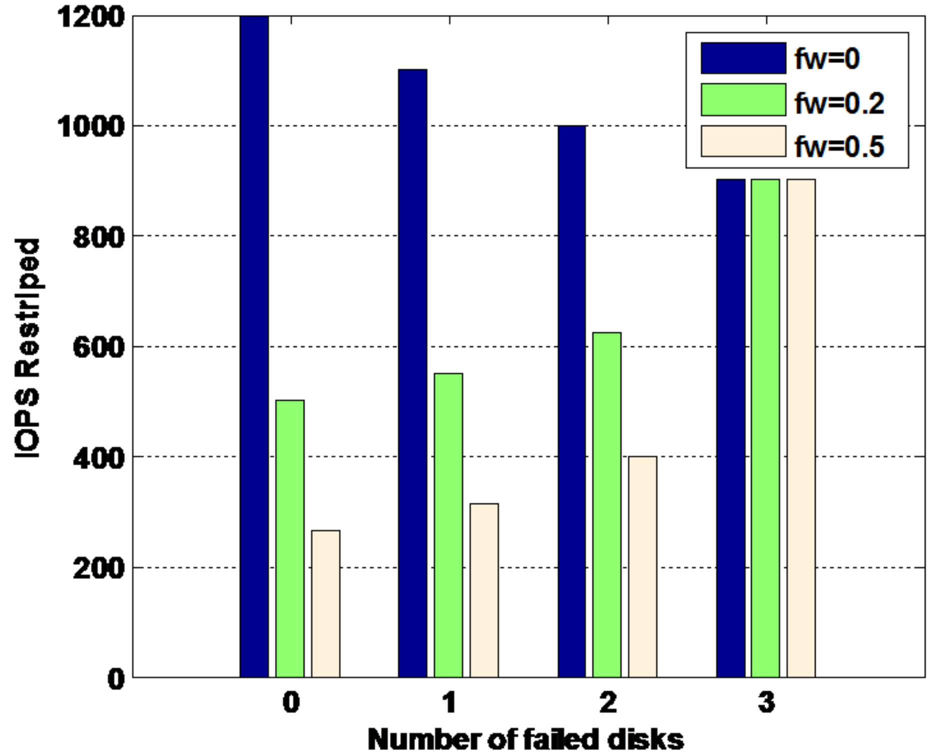


Fig. 2 Max IOPS in degraded mode of operation for varying number of disk failures starting with a fault-free RAID7 with N=12 disks.

Fig. 3 Max IOPS after restriping for varying number of disk failures starting with a fault-free RAID7 with N=12 disks.

Outline

1. Introduction
2. **HRAID Motivation**
3. Performance
4. Reliability Analysis
5. Design Study
6. Future Work
7. Conclusion

Example 1: HRAID k/ℓ with $N = M = 4$

- N: number of nodes/ M: number of disks per node
- kFT protection at inter-node level (Q parities, $k = 1$)
  ℓDFT protection at intra-node level (P parities, $\ell = 1$)

Node 1				Node [illegible]				[illegible]				[illegible]			
$D^1_{1,1}$	$D^1_{1,2}$	$P^1_{1,3}$	$Q^1_{1,4}$	$D^2_{1,1}$	$P^2_{1,2}$	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
$D^1_{2,1}$	$P^1_{2,2}$	$Q^1_{2,3}$	$D^1_{2,4}$	$P^2_{2,1}$	$Q^2_{2,2}$	$D^2_{2,3}$	[illegible]	[illegible]	[illegible]	$D^3_{2,3}$	$P^3_{2,4}$	$D^4_{2,1}$	$D^4_{2,2}$	$P^4_{2,3}$	[illegible]
$P^1_{3,1}$	$Q^1_{3,2}$	$D^1_{3,3}$	$D^1_{3,4}$	$Q^2_{3,1}$	$D^2_{3,2}$	$D^2_{3,3}$	$P^1_{3,4}$	$D^3_{3,1}$	$D^3_{3,2}$	$P^3_{3,3}$	$Q^3_{3,4}$	$D^4_{3,1}$	$P^4_{3,2}$	$Q^4_{3,3}$	$D^4_{3,4}$
$Q^1_{4,1}$	$D^1_{4,2}$	$D^1_{4,3}$	$P^1_{4,4}$	$D^2_{4,1}$	$D^2_{4,2}$	$P^2_{4,3}$	$Q^2_{4,4}$	$D^3_{4,1}$	$P^3_{4,2}$	$Q^3_{4,3}$	$D^3_{4,4}$	$P^4_{4,1}$	$Q^4_{4,2}$	$D^4_{4,3}$	$D^4_{4,4}$

P parities protect Q parities, but not vice-versa

- The storage efficiency for HRAID k/ℓ:

$$u = \frac{(N-k)(M-\ell)}{NM} = 1 - \frac{k}{N} - \frac{\ell}{M} + \frac{k\ell}{NM}$$

Example 2: HRAID2/1 with $N = M = 5$

➢ P intra-node RAID5 parity, Q and S inter-node RS code (only the first row is shown)

Node 1					Node 2					Node 3					Node 4					Node 5				
$D^1_{1,1}$	$D^1_{1,2}$	$P^1_{1,3}$	$Q^1_{1,4}$	$S^1_{1,5}$	$D^2_{1,1}$	$P^2_{1,2}$	$Q^2_{1,3}$	$S^2_{1,4}$	$D^2_{1,5}$	$P^3_{1,1}$	$Q^3_{1,2}$	$S^3_{1,3}$	$D^3_{1,4}$	$D^3_{1,5}$	$Q^4_{1,1}$	$S^4_{1,2}$	$D^4_{1,3}$	$D^4_{1,4}$	$P^4_{1,5}$	$S^5_{1,1}$	$D^5_{1,2}$	$D^5_{1,3}$	$P^5_{1,4}$	$Q^5_{1,5}$

- If node 5 fails, it is reconstructed using S stripes at other nodes

$$(D^1_{1,1}, D^1_{1,2}, P^1_{1,3}, Q^1_{1,4}, D^5_{1,2}), (D^2_{1,1}, P^2_{1,2}, Q^2_{1,3}, D^5_{1,3}, D^5_{1,2}),$$

$$(P^3_{1,1}, Q^3_{1,2}, —, D^3_{1,4}, D^3_{1,5}), (Q^4_{1,1}, —, D^4_{1,3}, D^4_{1,4}, P^5_{1,5})$$

- If nodes 4 and 5 fail, use both Q and S strips

Up to 2 node failures & one disk failure per node can be tolerated

Updating Data Blocks in HRAID1/1

- To update $d^2_{4,1}$ block in strip $D^2_{4,1}$

$d^1_{1,1}$	$d^1_{1,2}$	$p^1_{1,3}$	$q^1_{1,4}$	$d^2_{1,1}$	$p^2_{1,2}$	$q^2_{1,3}$	$d^2_{1,4}$	$p^3_{1,1}$	$q^3_{1,2}$	$d^3_{1,3}$	$d^3_{1,4}$	$q^4_{1,1}$	$d^4_{1,2}$	$d^4_{1,3}$	$p^4_{1,4}$
$d^1_{2,1}$	$p^1_{2,2}$	$q^1_{2,3}$	$d^1_{2,4}$	$p^2_{2,1}$	$q^2_{2,2}$	$d^2_{2,3}$	$d^2_{2,4}$	$q^3_{2,1}$	$d^3_{2,2}$	$p^3_{3,3}$	$p^3_{2,4}$	$d^4_{2,1}$	$d^4_{2,2}$	$p^4_{2,3}$	$q^4_{2,4}$
$p^1_{3,1}$	$q^1_{3,2}$	$d^1_{3,3}$	$d^1_{3,4}$	$q^2_{3,1}$	$d^2_{3,2}$	$d^2_{3,3}$	$p^1_{3,4}$	$d^3_{3,1}$	$d^3_{3,2}$	$d^3_{2,3}$	$q^3_{3,4}$	$d^4_{3,1}$	$p^4_{3,2}$	$q^4_{3,3}$	$d^4_{3,4}$
$q^1_{4,1}$	$d^1_{4,2}$	$d^1_{4,3}$	$p^1_{4,4}$	$d^2_{4,1}$	$d^2_{4,2}$	$p^2_{4,3}$	$q^2_{4,4}$	$d^3_{4,1}$	$p^3_{4,2}$	$q^3_{4,3}$	$d^3_{4,4}$	$p^4_{4,1}$	$q^4_{4,2}$	$d^4_{4,3}$	$d^4_{4,4}$

$q^{1new}_{4,1} = q^{1old}_{4,1}$ | $p^{1new}_{4,4} = p^{1old}_{4,4} \oplus d^{2diff}_{4,1}$ | $p^{2old}_{4,3} \oplus d^{2diff}_{4,1}$

- For HRAID k/ℓ, $(k+1)(\ell+1)$ blocks need to be read and written per update

Storage Transactions

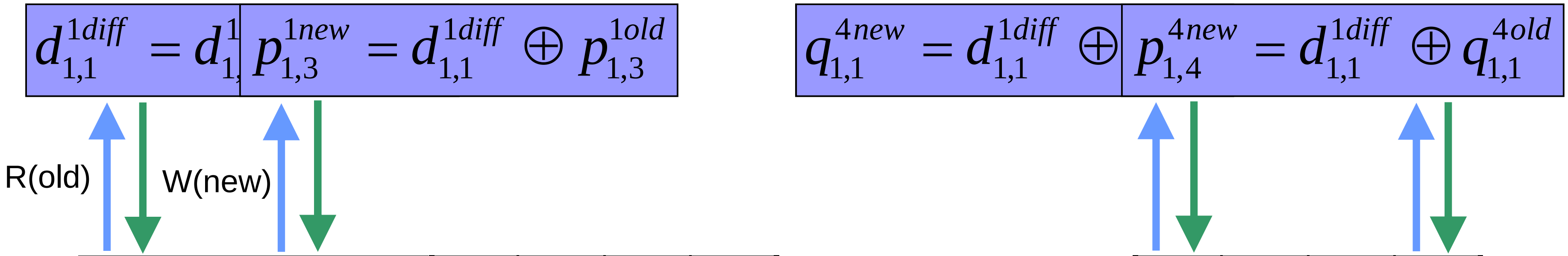
➢ Race conditions arise in updating of two data blocks, such as $d^{1}_{1,1}$ and $d^{1}_{1,2}$, since they affect the same parity block $p^{1}_{1,3}$

$d^{1}_{1,1}$	$d^{1}_{1,2}$	$p^{1}_{1,3}$	$q^{1}_{1,4}$	$d^{2}_{1,1}$	$p^{2}_{1,2}$	$q^{2}_{1,3}$	$d^{2}_{1,4}$	$p^{3}_{1,1}$	$q^{3}_{1,2}$	$d^{3}_{1,3}$	$d^{3}_{1,4}$	$q^{4}_{1,1}$	$d^{4}_{1,2}$	$d^{4}_{1,3}$	$p^{4}_{1,4}$
$d^{1}_{2,1}$	$p^{1}_{2,2}$	$q^{1}_{2,3}$	$d^{1}_{2,4}$	$p^{2}_{2,1}$	$q^{2}_{2,2}$	$d^{2}_{2,3}$	$d^{2}_{2,4}$	$q^{3}_{2,1}$	$d^{3}_{2,2}$	$p^{3}_{3,3}$	$p^{3}_{2,4}$	$d^{4}_{2,1}$	$d^{4}_{2,2}$	$p^{4}_{2,3}$	$q^{4}_{2,4}$
$p^{1}_{3,1}$	$q^{1}_{3,2}$	$d^{1}_{3,3}$	$d^{1}_{3,4}$	$q^{2}_{3,1}$	$d^{2}_{3,2}$	$d^{2}_{3,3}$	$p^{1}_{3,4}$	$d^{3}_{3,1}$	$d^{3}_{3,2}$	$d^{3}_{2,3}$	$q^{3}_{3,4}$	$d^{4}_{3,1}$	$p^{4}_{3,2}$	$q^{4}_{3,3}$	$d^{4}_{3,4}$
$q^{1}_{4,1}$	$d^{1}_{4,2}$	$d^{1}_{4,3}$	$p^{1}_{4,4}$	$d^{2}_{4,1}$	$d^{2}_{4,2}$	$p^{2}_{4,3}$	$q^{2}_{4,4}$	$d^{3}_{4,1}$	$p^{3}_{4,2}$	$q^{3}_{4,3}$	$d^{3}_{4,4}$	$p^{4}_{4,1}$	$q^{4}_{4,2}$	$d^{4}_{4,3}$	$d^{4}_{4,4}$

$$p^{1new}_{1,3} = d^{1diff}_{1,1} \oplus p^{1old}_{1,3} \quad or \quad p^{1new}_{1,3} = d^{1diff}_{1,2} \oplus p^{1old}_{1,3}$$

$$p^{1new}_{1,3} = d^{1diff}_{1,1} \oplus d^{1diff}_{1,2} \oplus p^{1old}_{1,3}$$

Txn to update $d_{1,1}^{1new}$ in HRAID1/1



$d^1_{1,1}$	$d^1_{1,2}$	$p^1_{1,3}$	$q^1_{1,4}$	$d^2_{1,1}$	$p^2_{1,2}$	$q^2_{1,3}$	$d^2_{1,4}$	$p^3_{1,1}$	$q^3_{1,2}$	$d^3_{1,3}$	$d^3_{1,4}$	$q^4_{1,1}$	$d^4_{1,2}$	$d^4_{1,3}$	$p^4_{1,4}$
$d^1_{2,1}$	$p^1_{2,2}$	$q^1_{2,3}$	$d^1_{2,4}$	$p^2_{2,1}$	$q^2_{2,2}$	$d^2_{2,3}$	$d^2_{2,4}$	$q^3_{2,1}$	$d^3_{2,2}$	$p^3_{3,3}$	$p^3_{2,4}$	$d^4_{2,1}$	$d^4_{2,2}$	$p^4_{2,3}$	$q^4_{2,4}$
$p^1_{3,1}$	$q^1_{3,2}$	$d^1_{3,3}$	$d^1_{3,4}$	$q^2_{3,1}$	$d^2_{3,2}$	$d^2_{3,3}$	$p^1_{3,4}$	$d^3_{3,1}$	$d^3_{3,2}$	$d^3_{2,3}$	$q^3_{3,4}$	$d^4_{3,1}$	$p^4_{3,2}$	$q^4_{3,3}$	$d^4_{3,4}$
$q^1_{4,1}$	$d^1_{4,2}$	$d^1_{4,3}$	$p^1_{4,4}$	$d^2_{4,1}$	$d^2_{4,2}$	$p^2_{4,3}$	$q^2_{4,4}$	$d^3_{4,1}$	$p^3_{4,2}$	$q^3_{4,3}$	$d^3_{4,4}$	$p^4_{4,1}$	$q^4_{4,2}$	$d^4_{4,3}$	$d^4_{4,4}$

Reads (resp. writes) preceded by a request for a shared (resp. exclusive) lock.

Since the identity of all locks is known a priori, they can be requested concurrently by a frontend node.

Txn to Update $d_{1,1}^{1}$ in HRAID2/1

- Node 1: $d_{1,1}^{1old}$ and $p_{1,3}^{1old}$ are read, $d_{1,1}^{1diff}$ computed and used to update $p_{1,3}^{1old}$, D and P blocks are written. Send $d_{1,1}^{1diff}$ to nodes 4 and 5.
- Node 4: Read and update Q check block and its parity.
- Node 5: Read and update R check block and its parity.

Txn to Update $d_{1,1}^{1}$ in HRAID2/1

Subtxn at Node 1	Subtxn at Node 4	Subtxn at Node 5
$R(d_{1,1}^{1old}), R(p_{1,3}^{1old})$ $d_{1,1}^{1diff} = d_{1,1}^{1old} \oplus d_{1,1}^{1new}$ $W(d_{1,1}^{1new}), p_{1,3}^{1new} = d_{1,1}^{1diff} \oplus p_{1,3}^{1old}$ $W(p_{1,3}^{1new}), S_{4,5}(d_{1,1}^{1diff})$		
	$R(q_{1,1}^{4old}), R(p_{1,5}^{4old})$ $q_{1,1}^{4new} = q_{1,1}^{4old} \oplus d_{1,1}^{4diff}$ $p_{1,5}^{4new} = p_{1,5}^{4old} \oplus d_{1,1}^{1diff}$ $W(q_{1,1}^{4new}), W(p_{1,5}^{4new})$	$R(r_{1,1}^{5old}), R(p_{1,4}^{5old})$ $q_{1,1}^{4new} = q_{1,1}^{4old} \oplus d_{1,1}^{4diff}$ $p_{1,4}^{5new} = p_{1,4}^{5old} \oplus d_{1,1}^{1diff}$ $W(q_{1,1}^{5new}), W(p_{1,4}^{5new})$

Outline

1. Introduction
2. HRAID Motivation
3. **Performance**
4. Reliability Analysis
5. Design Study
6. Future Work
7. Conclusion

Performance Analysis

- OLTP workload
- Reads/Writes to small randomly placed small blocks of data (high disk arm positioning time)
- Fraction of Reads/Writes: f_R and $f_W = 1 - f_R$
- Mean service time for reads:

$$\bar{x}_R = \bar{x}_{seek} + \bar{x}_{latency} + \bar{x}_{xfer}$$

- Mean service time for writes:

$$\bar{x}_W = \bar{x}_R + T_{HST} \quad \text{(head settling time)}$$

Maximum IOPS for RAID

- RAID0: $\lambda_0 = \left[f_R \overline{x}_{SR} + f_W \overline{x}_{SW} \right]^{-1}$

  RAIDℓ: $\lambda_\ell = \left\{ f_R \overline{x}_{SR} + f_W \left[(l+1)\left(\overline{x}_{SR} + \overline{x}_{SW} \right) \right] \right\}^{-1}$

  HRAIDk/ℓ: $\lambda_{k/\ell} = \left\{ f_R \overline{x}_{SR} + f_W \left[(k+1)(\ell+1)\left(\overline{x}_{SR} + \overline{x}_{SW} \right) \right] \right\}^{-1}$

  Assume HST negligible, so that $\overline{x}_{SW} \approx \overline{x}_{SR}$.

- Relative maxIOPS for RAID0 : RAIDℓ : HRAIDk/ℓ

$$1 : \left[1 + f_W (1 + 2\ell) \right]^{-1} : \left[1 + f_W + 2 f_W (\ell k + \ell + k) \right]^{-1}$$

- For $f_R = 0.8$, the maximum throughput drops to 0.42 for RAID6 and to 0.31 for HRAID1/2

  For $f_R = 0.5$, the drop is 0.40 for RAID6 and 0.15 for HRAID1/2

HRAID Response Time: Preliminaries

- OLTP workload－ accesses to small randomly placed blocks.
- Poisson arrivals, so each disk an M/G/1 queue
- Total arrival rate Λ to D
- Logical requests per dis
- $\overline{x^i}$ denotes the i^{th} mom
- $\rho = \lambda \overline{x}$ is the disk utiliza
- The mean waiting time Khinchine queueing formu

When service time exponential

$\overline{x^2} = 2\overline{x}^2$ and $W = \frac{\rho \overline{x}}{1-\rho}$

W increases rapidly with ρ:

$W = \overline{x}/7$ for $\rho = 0.125$

$W = \overline{x}/3$ for $\rho = 0.25$

$W = \overline{x}$ for $\rho = 0.5$

$W = 9\overline{x}$ for $\rho = 0.9$

$$W = \frac{\lambda \overline{x^2}}{2(1-\rho)}$$

HRAID Response Time

- Consider RAIDℓ and HRAIDk/ℓ response time for read requests, processed as Single Reads (SRs)
- Update requests are processed as an SR followed by a Single Write (SW)
- We prioritize SR requests due to only and simplify discussions: $\bar{x}_{SW}$

We also postulate that service times are exponentially distrbuted:

$$\overline{x^2_{SW}} \approx \overline{x^2_{SR}} = 2(\bar{x})^2$$

$$W_R^{(0)} = \frac{W_0}{1-\rho_R} = \frac{\rho\bar{x}}{1-\rho_R}$$

$$W_0 = \rho_R \frac{\overline{x^2}_{SR}}{2\bar{x}_{SR}} + \rho_W \frac{\overline{x^2}_{SW}}{2\bar{x}_{SW}} = \rho\bar{x}$$

Waiting Times

- RAIDℓ:

$$W_R^{(\ell)} = [f_R + 2f_W(1+\ell)]\frac{\lambda\overline{x}^2}{1-\rho_R}$$

- HRAIDk/ℓ:

$$W_R^{(k/\ell)} = [f_R + 2f_W(1+k)(1+\ell)]\frac{\lambda\overline{x}^2}{1-\rho_R}$$

Normalized Waiting Time Graphs

➢ In Fig. 5, we plot $W_R^{(k/\ell)}$ normalized with respect to $W_R^{(0)}$ versus f_W, which shows that it increases linearly

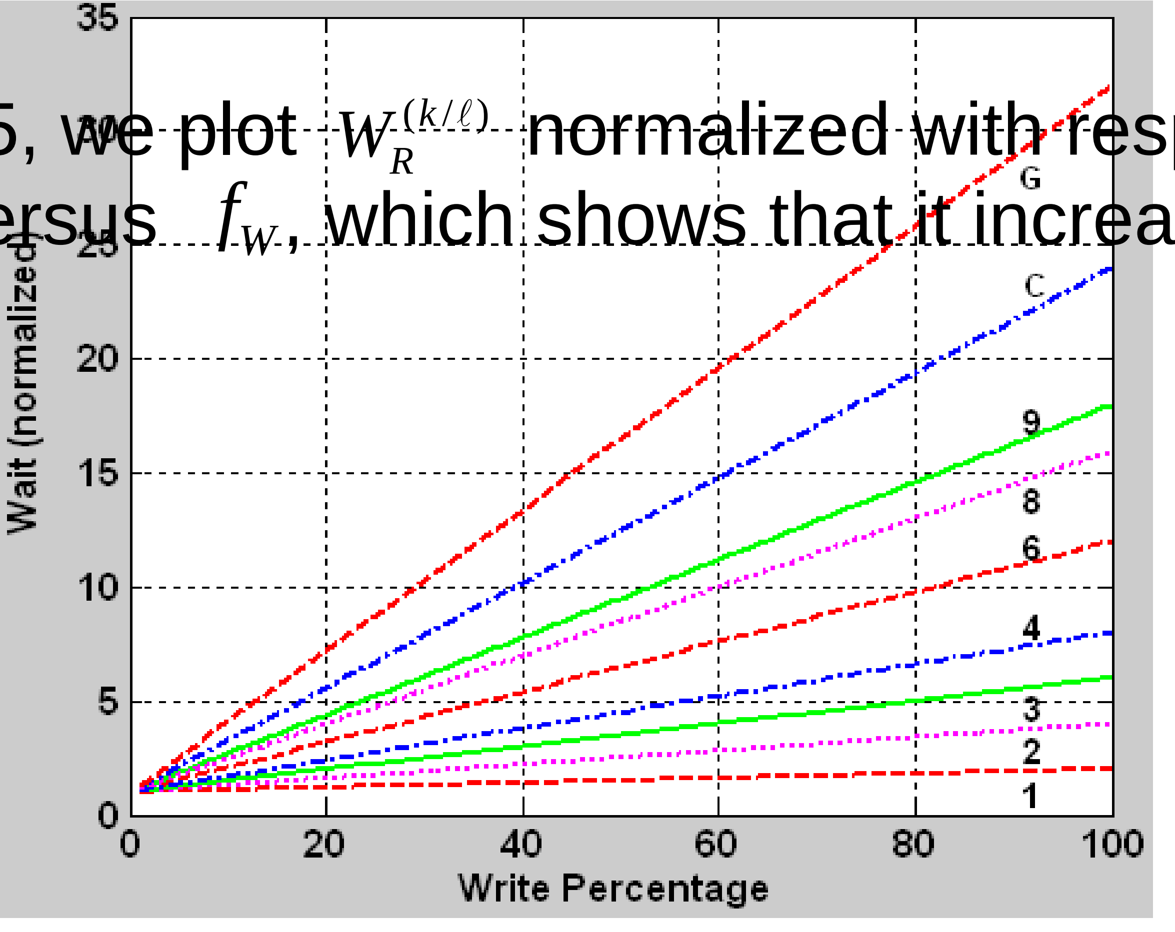


Fig. 5 Normalized mean waiting for different RAID and HRAID configurations, specified as the values of $(k+1)(\ell+1)$ for $0 \le \ell, k \le 3$ given above (we use $C=12$ for $k=3,2$ and $\ell=2,3$ and $G=16$ for $k=\ell=3$)

Response Time Graphs

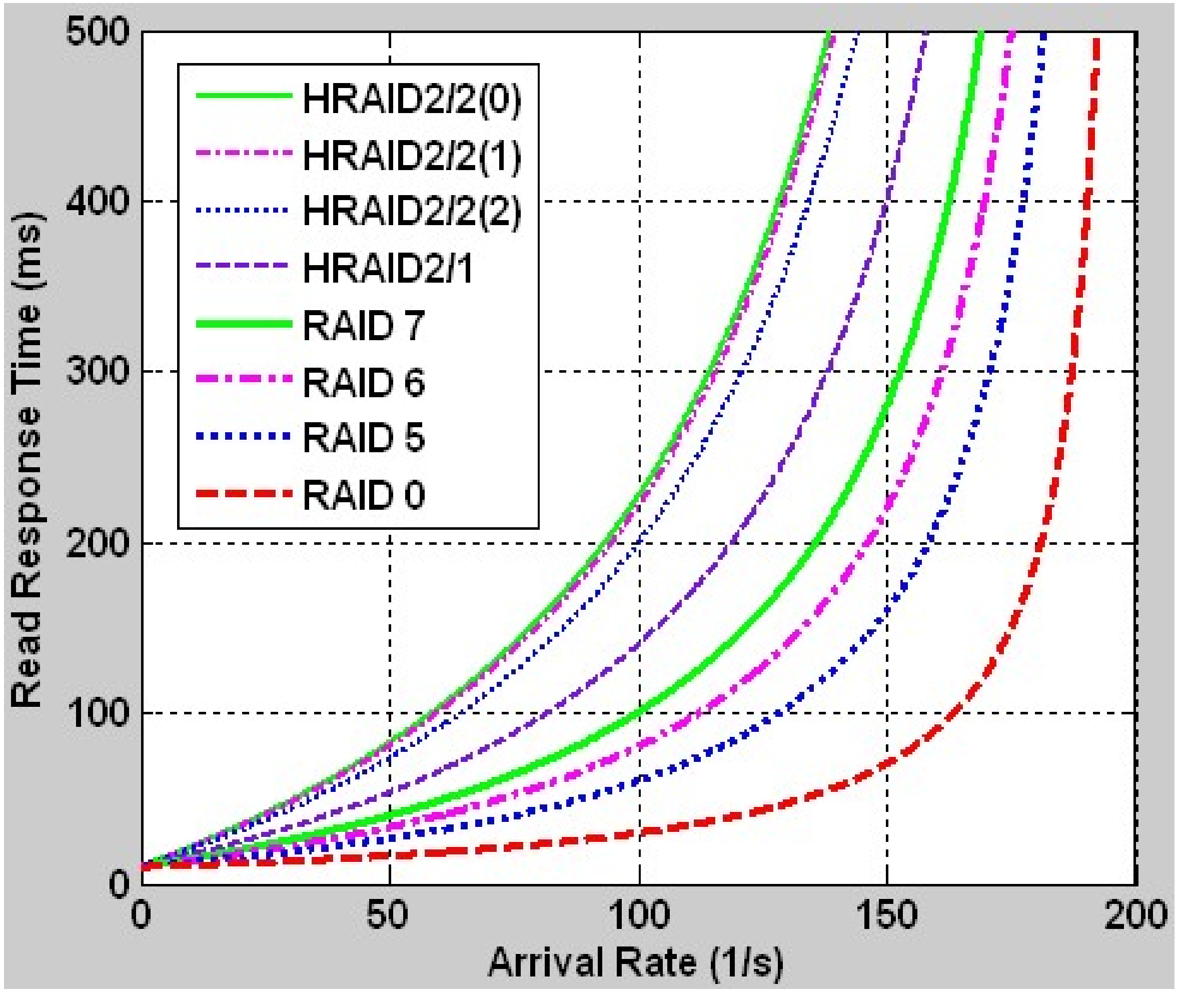


Fig. 4 Prioritized read response time with 0, 1, 2 node failures for HRAID2/2 (in parentheses) and no failures for HRAID1/1, RAIDℓ, $0 \le \ell \le 3$ for R:W=1:1

HRAID2/2 with Node Failures

Table 3: Cost of Operation for HRAID2/2 with N nodes.
Mode settings 0F, 1F and 2F indicate number of failed nodes

Mode
0F
1F
2F

$\bar{x}_{RMW}$ shorthand for $\bar{x}_{SR} + \bar{x}_{SW}$:

$$T_{0F} = \frac{NM}{f_R \bar{x}_{SR} + 9 f_W \bar{x}_{RMW}}$$

$$T_{1F} = \frac{(N-1)M}{\frac{2N-5}{N} f_R \bar{x}_{SR} + f_W \left[\frac{2(N-4)}{N} \bar{x}_{SR} + \frac{9(N-1)}{N} \bar{x}_{RMW} \right]}$$

$$T_{2F} = \frac{(N-2)M}{\frac{3N-10}{N} f_R \bar{x}_{SR} + f_W \left[\frac{2(N-3)(N-5)}{N(N-1)} \bar{x}_{SR} + \frac{9(N-2)}{N} \bar{x}_{RMW} \right]}$$

- The maximum throughput of HRAID2/2 with zero, one, and two node failures are obtained using the cost functions given in Table 3

Outline

1. Introduction
2. HRAID Motivation
3. Performance
4. **Reliability Analysis**
5. Design Study
6. Future Work
7. Conclusion

Number of Failed Disks Tolerated

- HRAID1/1 cannot recover from four disk failures, when their coordinates constitute a rectangle.

  For $N = M = 12$, 2.5 data losses in 10,000 cases.

- Regardless of N and M, HRAID k / ℓ can tolerate all $(k+1)(\ell+1)-1$ disk failures

- The probability that $(k+1)(\ell+1)$ disk failures result in data loss is given as:

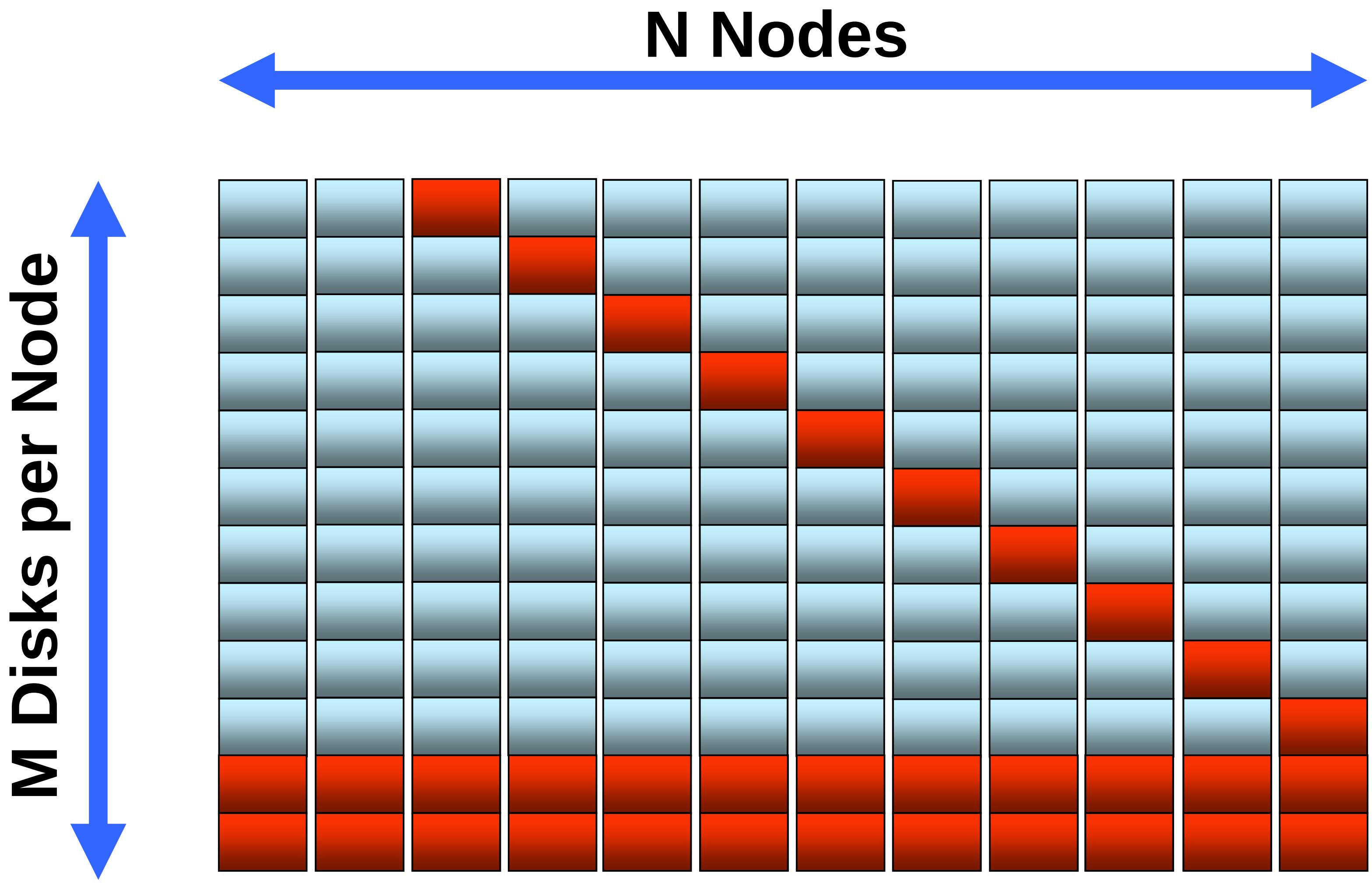
$$P\left[Data\_Loss\right] = \frac{\binom{N}{k+1}\binom{M}{\ell+1}}{\binom{N \times M}{(k+1)(\ell+1)}}$$

Max Number of Disk Failures Tolerated

- This number is maximized with no controller failures: $D_{tol} = N \times \ell + (M - \ell)k$
- For $M = N$, $D_{tol} = N(k+\ell) - k\ell$
- For $N = M = 12$, and $k = \ell = 2$ up to 44 disks failures can be tolerated, while $D_{red} = N \times (k+\ell) = 48$ disks out of 144 disks are check disks
- N increases

$$\frac{D_{tol}}{D_{red}} = 1 - \frac{k\ell}{N(k+\ell)} \rightarrow 1$$

D_{tol} for HRAID1/2 with N=M=12



Approximate Reliability Analysis

- Let $r = 1 - \varepsilon$ denote disk reliability where $\varepsilon \ll 1$
- The approximate reliability expressions for RAID5 with N disks, which can tolerate one disk failure is:

$$R_1 = r^N + N(1-r)r^{N-1} \approx 1 - \binom{N}{2}\varepsilon^2 + 2\binom{N}{3}\varepsilon^3 - \ldots$$

We have subtracted the probability of data loss due to 2 disk failures

- For ℓDFT

$$R_\ell \approx 1 - \binom{N}{\ell+1}\varepsilon^{\ell+1} + (\ell+1)\binom{N}{\ell+1}\varepsilon^{\ell+2} - \ldots \qquad \ell \geq 1$$

HRAID1/2 vs HRAID2/1 Reliability

- $R_{1/2} = R_2^N + N(1 - R_2)R_2^{N-1}$

$$\approx 1 - \frac{N(N-1)M^2(M-1)^2(M-2)^2}{72}\varepsilon^6 + \ldots$$

$$R_{2/1} = R_1^N + N(1 - R_1)R_1^{N-1} + \binom{N}{2}(1 - R_1)^2 R_1^{N-2}$$

$$\approx 1 - \frac{N(N-1)(N-2)M^3(M-1)^3}{24}\varepsilon^6 + \ldots$$

- From disk reliability viewpoint $R_{1/2} > R_{2/1}$, since:

$$N > 2 + \frac{2(M-2)^2}{3M(M-1)}$$

Performability Analysis

- P = number of I/Os processed until data loss.
- Performability combines systems performance, reliability, and availability
- Defined by John Meyer (FTCS'78)
- Used in mathematical studies of system availability
- The Storage Group at SUN (now Oracle) used a Markov chain model to obtain the probability of normal and faulty states ($P_{S_i}, 1 \le i \le N$)

Performability Analysis (Continued)

- Fault injection was used to obtain the number of I/Os per second (IOPS)

$$P = \sum_{i=1}^{N} P_{S_i} IOPS_{S_i}$$

- In a system with no repair, the time in state $S_i(T_{S_i})$ is determined by the simulator (in our study)

$$P = \sum_{i \geq 1} T_{S_i} IOPS_i$$

- $IOPS_{S_i}$ obtained using previous formulas

HRAIDk/ℓ MTTDL and Performability

- Simulation assumptions

  1) All failure times are exponentially distributed;

  2) The disk failure rate is $\delta = 1$ per million hours, so that Mean Time to Failure (MTTF) = 10^6 hours;

  3) The controller failure rate γ is varied wrt δ;

  4) The number of controller failures (N_c) and disk failures (N_d) determine the state of the system.

- Simulation procedure outlined on next page

Initializations: $Clock = 0, P = 0$

- Given the total failure rate: $\Lambda = (N - N_c)\delta + (D - N_d)\gamma, (D = N \times M)$ determine the Time to Next Failure: TNF $= -[\Lambda]^{-1} \ln(u_1)$, where u_1 uniform r.v. in (0,1) and ln is the natural log.
- Increment the time: $Clock = Clock + TNF$ and performability: $P = P + TNF \times IOPS$.
- The probability of a controller failure is $p_c = (N - N_c)\gamma / \Lambda$.
- The identity of failed controllers and disks is determined probabilistically using uniform distributions.
- When a controller fails, all nonfailed disks attached to it are considered failed and N_c is incremented accordingly.
- When the number of failed disks at a node exceeds ℓ, the node is considered to be failed (in this study).
- The simulation is stopped when the number of failed nodes exceeds k, i.e., data loss occurs.

Effect of Controller Failure Rate

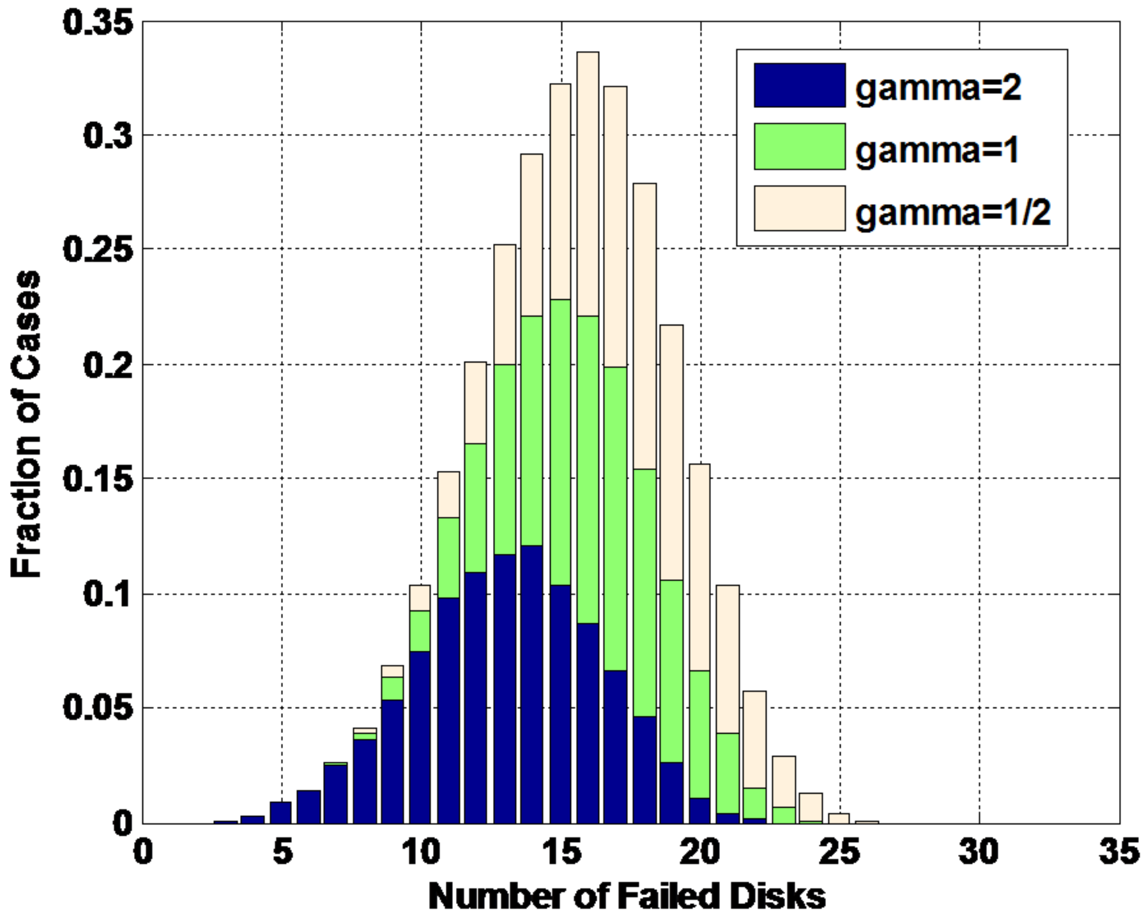


Fig. 6 The distribution of the number of failed disks for varying values of γ for HRAID2/2 when data loss occurs.

Effect of N/M on Number Failed Disk

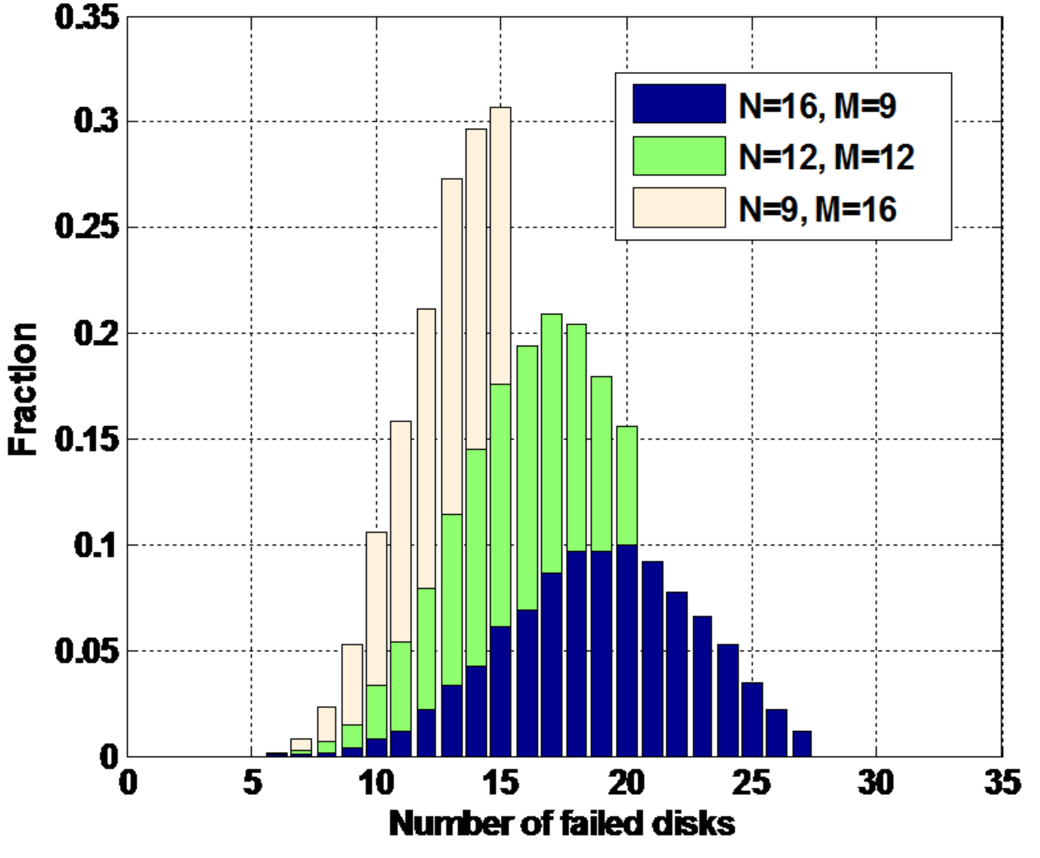


Fig. 7 Number of failed disks leading to data loss for three HRAID1/2 configurations with 9×16, 12×12, and 16×9 disks.
The heights of frequencies are not cumulative.

MTTDL & Cause of Data Loss

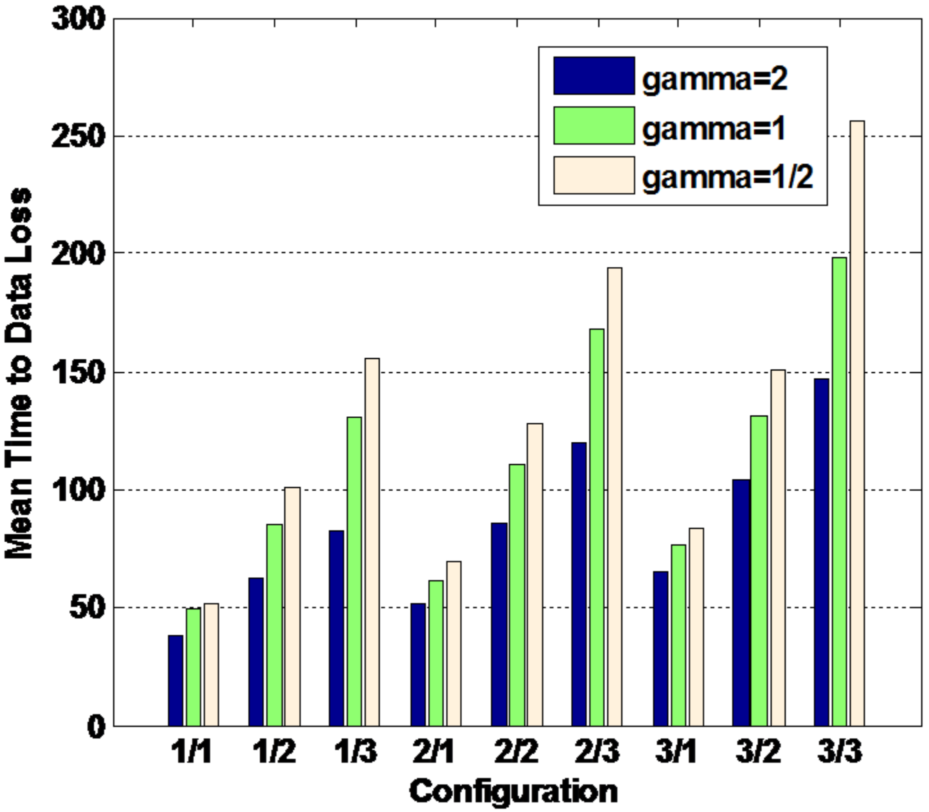


Fig. 8 The MTTDL for different HRAID configurations (1000s of hours)

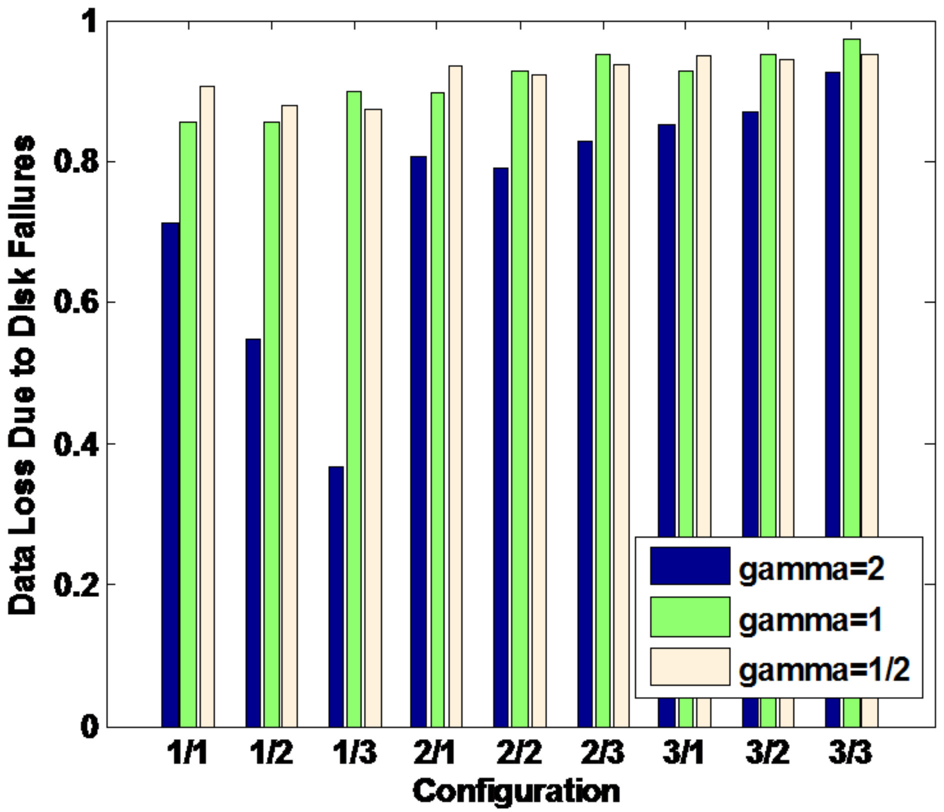


Fig. 9 The percentage of cases data loss due to disk failures

Effect of Controller Failure Rate

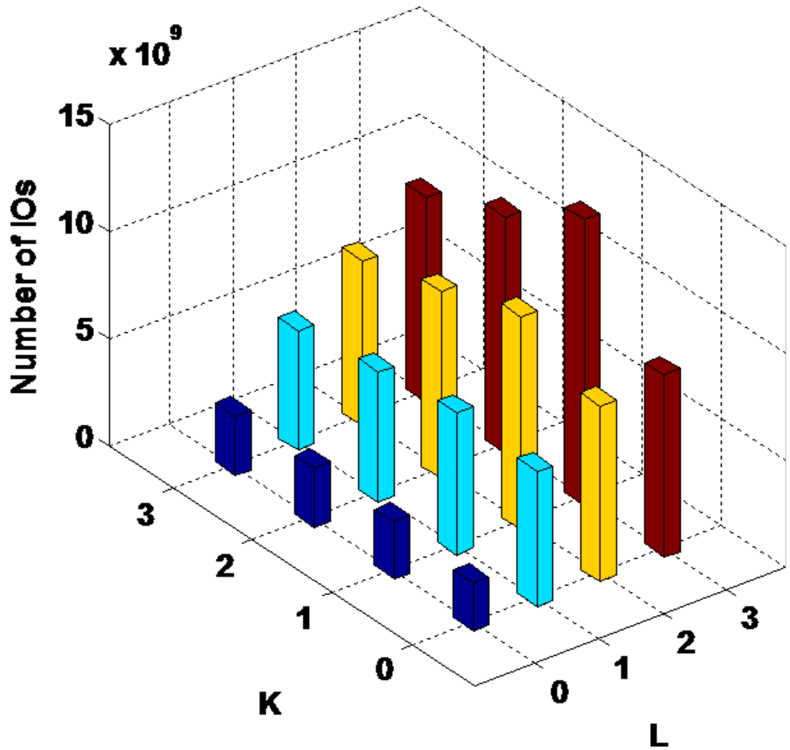


Fig. 10 Performability for different HRAID k / ℓ configurations with N = 12 nodes and M = 12 disks per node , $f_w = 0.5$ and the controller failure rate set to $\gamma = 1$. K and L stand for k and ℓ used in the text

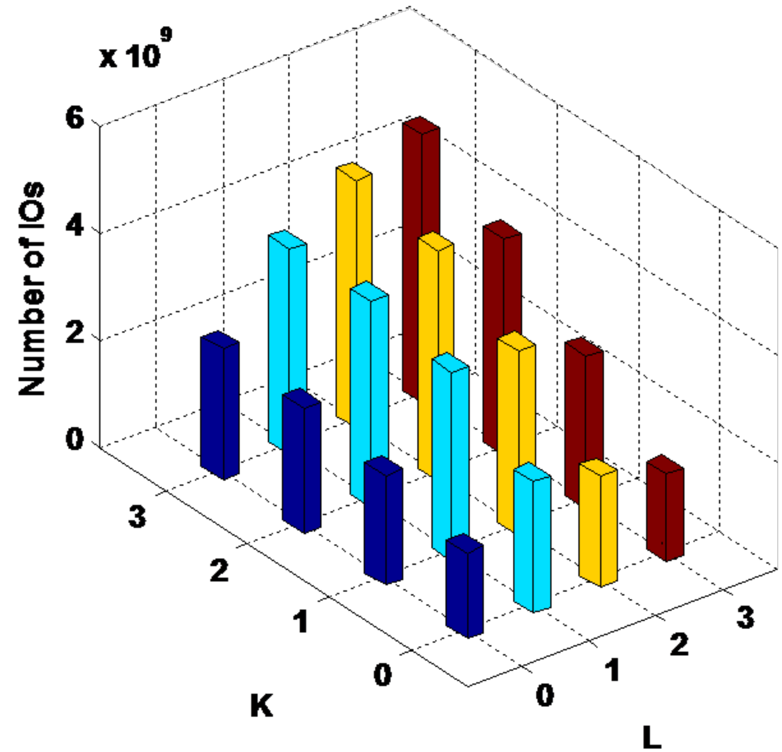


Fig. 11 Performability for different HRAID k / ℓ configurations with N = 12 nodes and M = 12 disks per node, $f_w = 0.5$ and the controller failure rate set to $\gamma = 6$. K and L stand for k and ℓ used in the text

Outline

1. Introduction
2. HRAID Motivation
3. Performance
4. Reliability Analysis
5. **Design Study**
6. Future Work
7. Conclusion

System Design Considerations

- Postulating controllers do not fail we estimate the mean number of disk failures leading to data loss ($\bar{d}$)

$$T_{data\_loss} = \sum_{i \geq 1}^{\bar{d}} \left[(D - i)\delta \right]^{-1}$$

- Given T_{data_loss}, we determine sufficiently reliable controllers, so that the time to controller failure exceeds T_{data_loss}:

$$T_{controller\_failure} = \sum_{j=0}^{k} \left[(N - j)\gamma \right]^{-1}$$

- To further ensure that HRAID failures are due to disk rather than controller failures, we obtain the 95^{th} percentile of number of disk failures that it takes for data loss to occur

Design Study

Table 4: The mean and 95^{th} percentile of the number of disk failures leading to data loss, the time to data loss, and the minimum controller failure rates required. The disk MTTF is set to one million hours, which is also the time unit

HRAID	$\bar{d}$	$T(\bar{d})$	$\gamma(\bar{d})$	$d^{95\%}$	$T(d^{95\%})$	$\gamma(d^{95\%})$
1,1	8	0.06	2.9	11	0.08	6
1,2	16	0.11	1.6	21	0.16	3
1,3	25	0.18	1.0	31	0.24	2
2,1	10	0.07	3.9	13	0.10	6
2,2	19	0.14	2.0	23	0.17	4
2,3	29	0.22	1.2	34	0.27	3
3,1	12	0.09	4.3	15	0.11	7
3,2	22	0.16	2.4	26	0.20	4
3,3	32	0.25	1.5	37	0.30	3

Considerations on Design Study

- For $1 \le k \le 3$ the mean number of disks to data loss increases linearly with ℓ: $\bar{d}_\ell \approx \ell \times \bar{d}_1$, $1 \le \ell \le 3$. The same can be said for percentiles
- $d_\ell^{95\%} = \bar{d}_\ell + I\sigma$, where the increment $3 \le I \le 5$ is a multiple of the standard deviation σ (not shown)
- Controller failure rates with the 95^{th} percentile requirement are twice as high as the mean. Controller failures rates with $k > 1$ are roughly equal to controller failure rates with level $k = 1$divided by k

Related Work

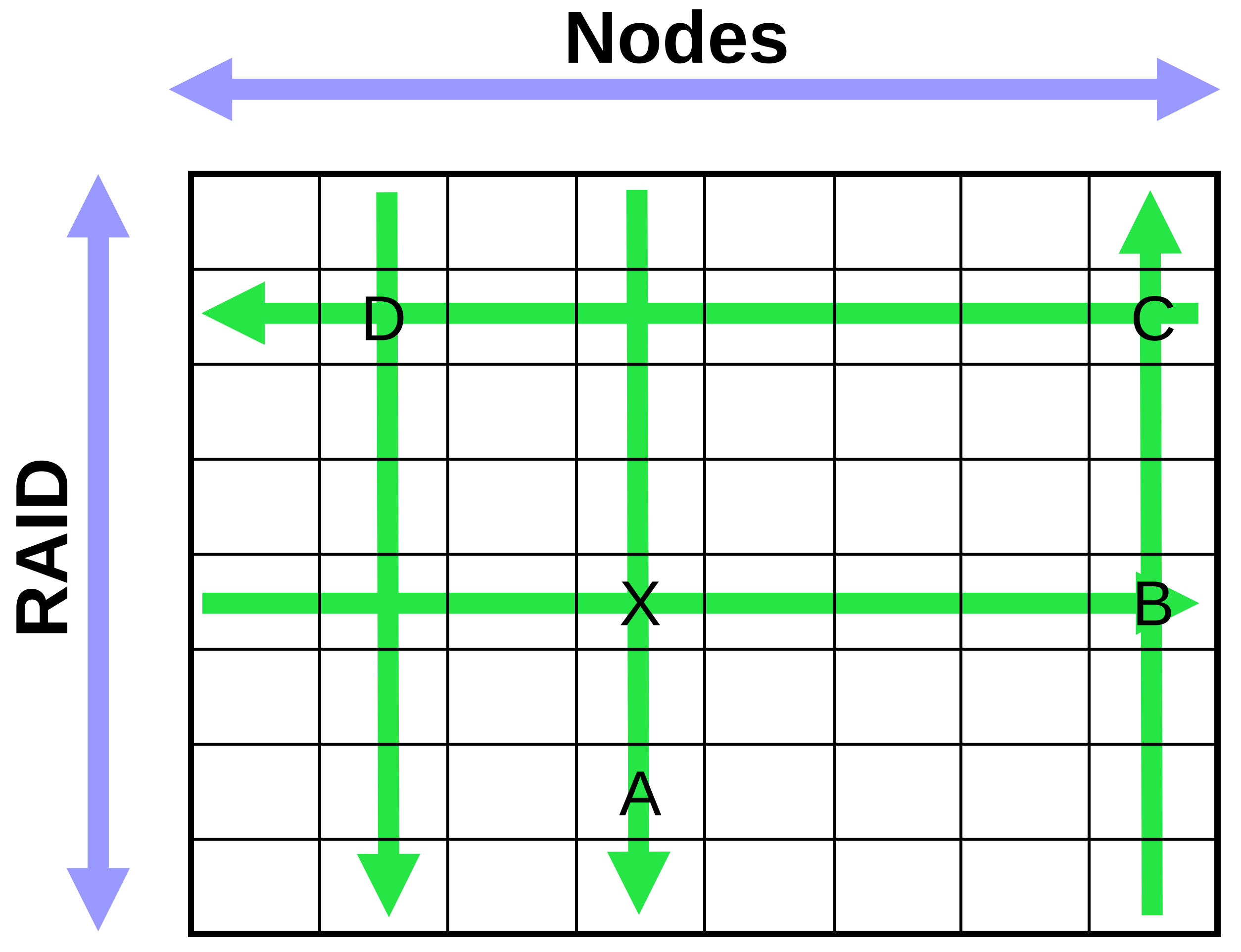
- Hierarchical RAID - Baek et al. in PODC'01
- Storage bricks projects:
- Dealing with controller/n failures
- Self-managed storage to e calls
- Non disk failures constitu of all failures - Schroeder and
- Storage clouds: Simple Storage Service (S3) Amazon

• IBM's Intelligent Bricks Project - Wilcke et. al. IBM J. R. & D.'06
• HP's Federated Array of Bricks (FAB) - Saito et al. ASPLOS'04
• RepStore: Microsoft China - Lin and Jin, ICAC'04

Future Work

- Assess the cost effectiveness of redundancy on MTTDL and performability.
- We consider a node to be failed when the number of failed disks at a node exceeds ℓ. This implies failed disks cannot be recovered locally, but recovery using inter-node check codes is possible.
- Figure on next page shows chained recovery.
- Study the effect of this assumption on MTTDL.

Chained Recovery of HRAID1/1



Thank You !

Q & A