Redundant Array of Independent Disks

Yashwant K. Malaiya

Redundant Array of Independent Disks (RAID)

- Enables greater levels of performance and/or reliability
- How? By concurrent use of two or more ‘hard disk drives’.
- How Exactly?
  - Striping (of data): data divided and spread across drives
  - Mirroring: identical contents on two disks
  - Error correction techniques: Parity

Hard Disks

- Rotate: one or more platters
- Efficient for blocks of data (sector 512 bytes)
- Inherently error prone
- ECC to check for errors internally
- Need a controller
- Can fail completely

- Modern discs use efficient Low Density Parity codes (LDPC).
- Some errors involving a few bits corrected internally, they are not considered here.

Solid State Drives (SDD)

- Also non-volatile, higher cost, higher reliability, lower power
- Still block oriented
- Controlled by a controller
- Wears out: 1000-6000 cycles
  - Wear levelling by the controller
- Random writes can be slower, caching can help.
- High level discussion applies to both HDD and SDD.

Wearout
- Ex: 64GB SSD, can sustain writes of 64GBx3000 = 192 TB
- Write 40 MB/hour, 260 GB/year. (8760 hours)
- Will last for 192 TB/260 GB = 738 years!

Standard RAID levels

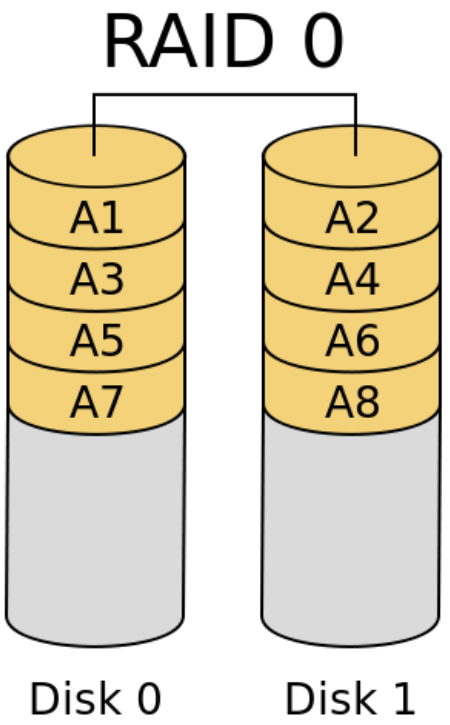
- RAID 0: striping
- RAID 1: mirroring
- RAID 2: bit-level striping, Hamming code for error correction (not used anymore)
- RAID 3: byte-level striping, parity (rare)
- RAID 4: block-level striping, parity
- RAID 5: block-level striping, distributed parity
- RAID 6: block-level striping, distributed double parity

RAID 0

- Data striped across n disks here n = 2
- Read/write in parallel
- No redundancy.

$$R_{sys} = \prod_{i=1}^{n} R_i$$

- Ex: 3 year disk reliability = 0.9 for 100% duty cycle. n = 14
- $R_{sys} = (0.9)^{14} = 0.23$

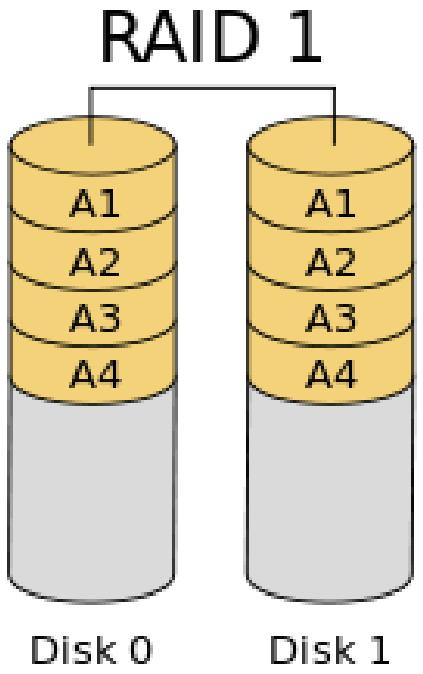


RAID 1

- Disk 1 mirrors Disk 0
- Read/write in parallel
- One of them can be used as a backup.

$$R_{sys} = \prod_{i=1}^{n} [1 - (1 - R_i)^2]$$

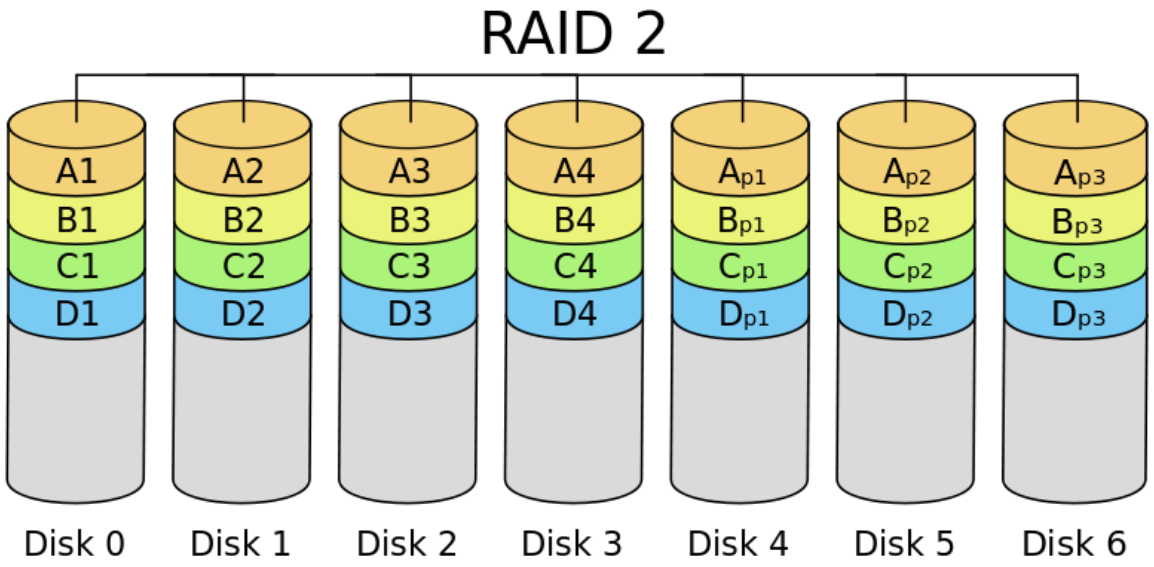
- Ex: 3 year disk reliability = 0.9 for 100% duty cycle. n = 7 pairs
- $R_{sys} = (2x0.9-(0.9)^2)^7 = 0.93$



Failed disk identified using internal ECC

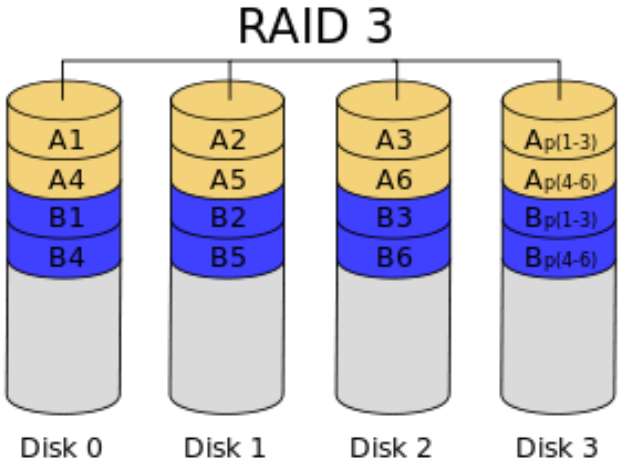
RAID 2

- Used Hamming code check bits as redundancy
- Bit level operations needed. Obsolete.



RAID 3

- Byte level striping not efficient
- Dedicated parity disk
- If one fails, its data can be reconstructed using a spare



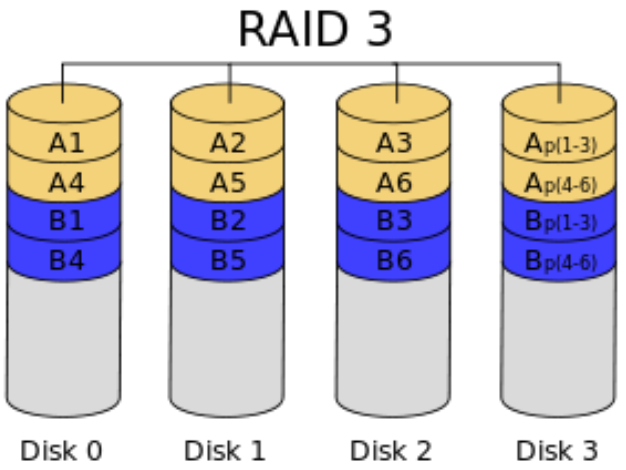
$$R_{sys} = \sum_{j=n-1}^{n} \binom{n}{j} R_j^{\,j} (1 - R_i)^{n-j}$$

- Ex: 3 year disk reliability = 0.9 for 100% duty cycle. n = 13, j = 12, 13
- R_{sys} = 0.62

RAID 4

- Block level striping
- Dedicated parity disk
- If one fails, its data can be reconstructed using a spare

$$R_{sys} = \sum_{j=n-1}^{n} \binom{n}{j} R_j^{\,j} (1 - R_i)^{n-j}$$

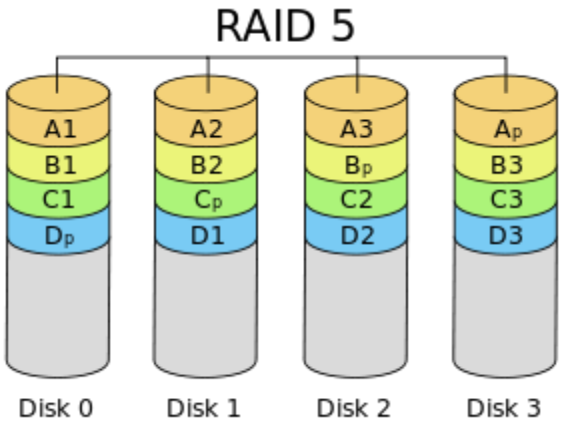


- Ex: 3 year disk reliability = 0.9 for 100% duty cycle. n = 13, j = 12, 13
- R_{sys} = 0.62

Parity disk bottleneck

RAID 5

- Distributed parity
- If one disk fails, its data can be reconstructed using a spare

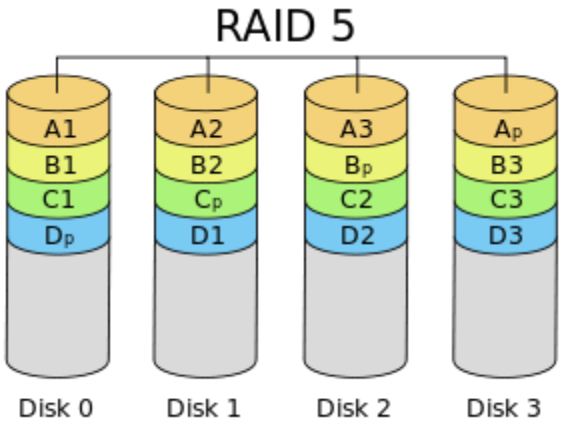


$$R_{sys} = \sum_{j=n-1}^{n} \binom{n}{j} R_j^{\,j} (1 - R_i)^{n-j}$$

- Ex: 3 year disk reliability = 0.9 for 100% duty cycle. n = 13, j = 12, 13
- R_{sys} = 0.62

RAID 5: illustration

- Distributed parity
- If one disk fails, its data can be reconstructed using a spare



Parity block = Block1 ⊕ block2 ⊕ block3

10001101 block1

01101100 block2

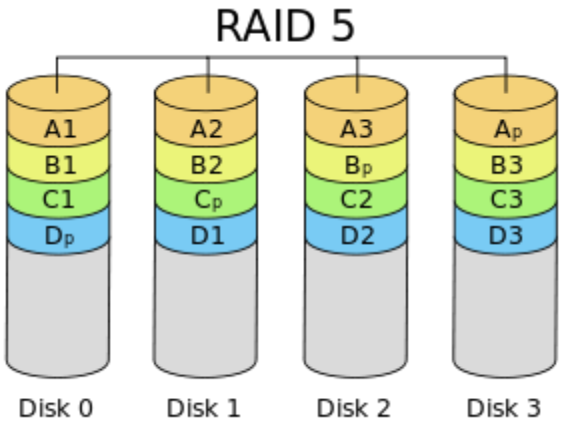
11000110 block3

00100111 parity block (ensures even number of 1s)

• Can reconstruct any missing block from the others

RAID 5: try

- Distributed parity
- If one disk fails, its data can be reconstructed using a spare



Parity block = Block1 ⊕ block2 ⊕ block3

10001101 block1

01101100 block2

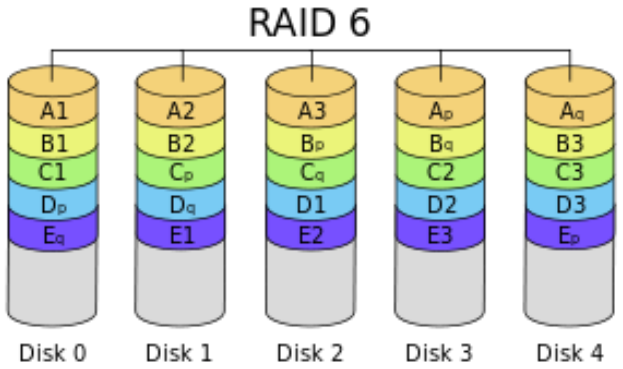
block3 (Go ahead, reconstruct)

00100111 parity block (ensures even number of 1s)

• Can reconstruct any missing block from the others

RAID 6

- Distributed double parity
- If one disk fails, its data can be reconstructed using a spare
- Handles data loss during a rebuild



$$R_{sys} = \sum_{j=n-2}^{n} \binom{n}{j} R_j^{\,j} (1 - R_i)^{n-j}$$

- Ex: 3 year disk reliability = 0.9 for 100% duty cycle. n = 13, j = 11, 12, 13
- R_{sys} = 0.87

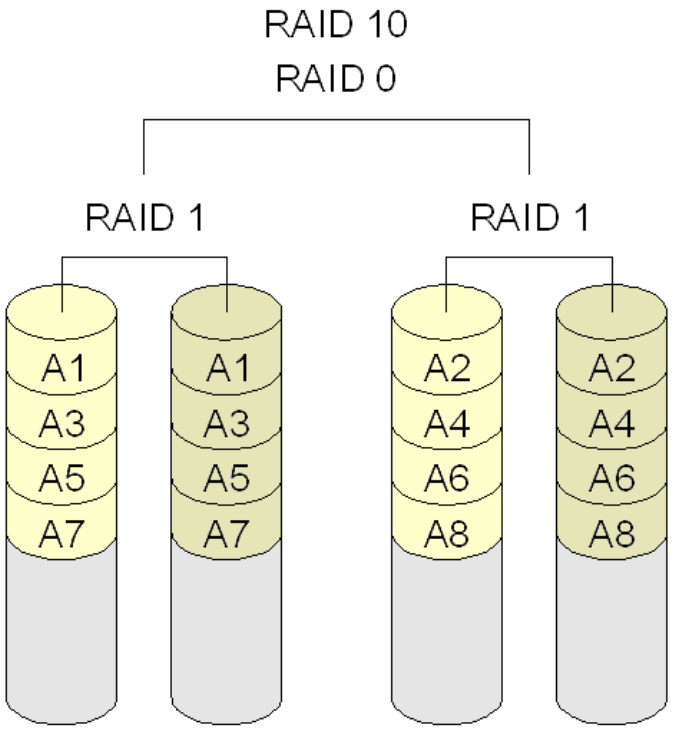
Nested RAID Levels

- RAID 01: mirror (1) of stripes (0)
- RAID 10: stripe of mirrors
- RAID 50: block-level striping of RAID 0 with the distributed parity of RAID 5 for individual subsets
- RAID 51: RAID5 duplicated
- RAID 60: block-level striping of RAID 0 with distributed double parity of RAID 6 for individual subsets.

RAID 10

- Stripe of mirrors: each disk in RAID0 is duplicated.

$$R_{sys} = \prod_{i=1}^{ns} [1-(1-R_i)^2]$$



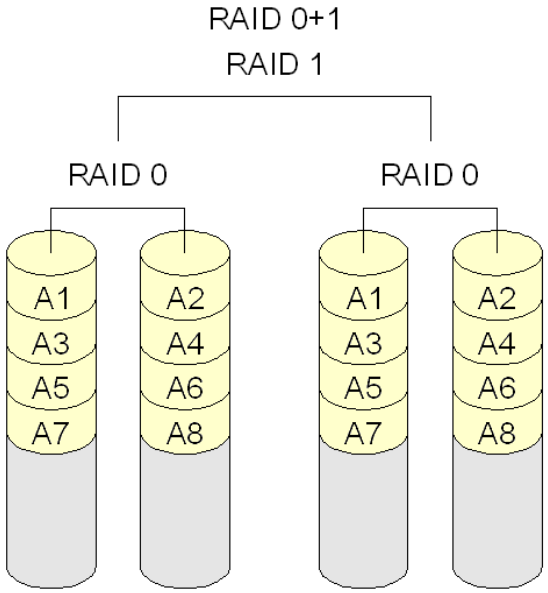
- Ex: 3 year disk reliability = 0.9 for 100% duty cycle. ns = 6 pairs,
- R_{sys} = 0.94

RAID 10: redundancy at lower level

RAID 01

- Mirror of stripes: Complete RAID0 is duplicated.

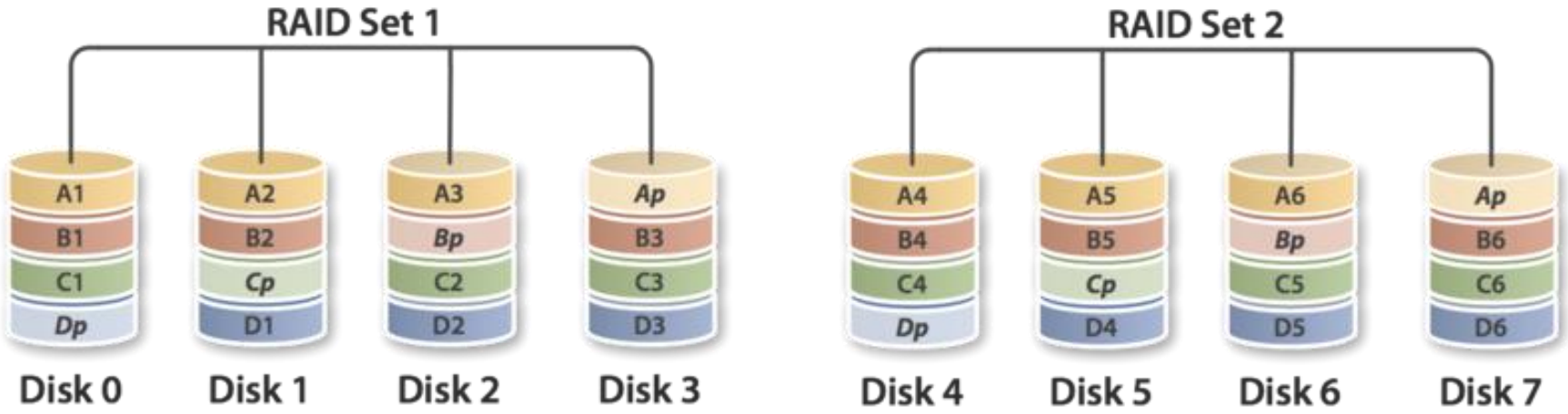
$$R_{sys} = [1-(1-\prod_{i=1}^{ns} R_i)^2]$$



- Ex: 3 year disk reliability = 0.9 for 100% duty cycle. ns = 6 for each of the two sets,
- R_{sys} = 0.78

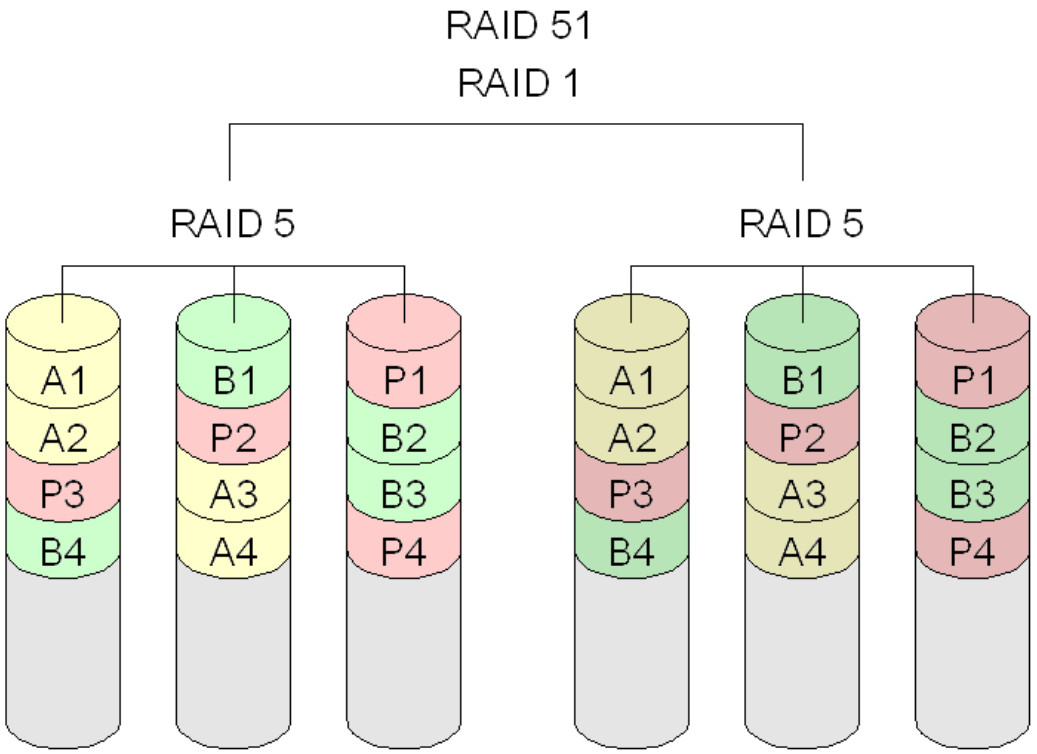
RAID 01: redundancy at higher level

RAID 50



- Multiple RAID 5 for higher capacity

RAID 51



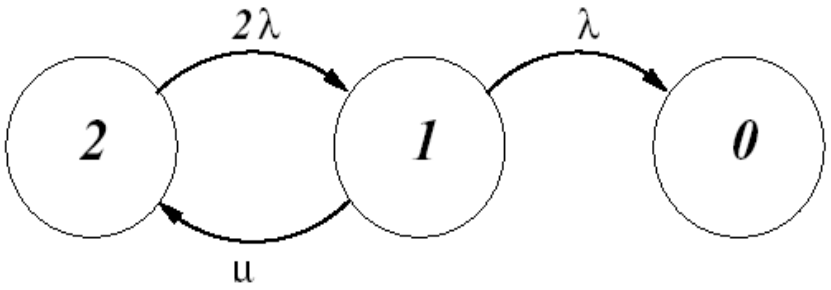
- Multiple RAID 5 for higher reliability (not capacity)

RAIDS Comparison

Level	Space efficiency	Fault tolerance	Read performance	Write performance
0	1	none	nx	nx
1 mirror	1/2	1 drive	2x	x
2	<1	1	var	var
3	<1	1	(n-1)x	(n-1)x
4 parity	<1	1	(n-1)x	(n-1)x
5 Dist Parity	<1	1	(n-1)x	(n-1)x
6 Dist Double P	<1	2	(n-2)x	(n-2)x
10 st of mirrors	1/2	1/set	nx	(n/2)x

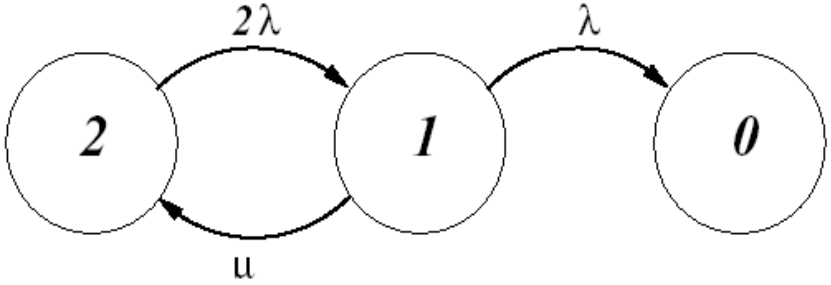
Markov modeling

- We have computed reliability using combinatorial modeling.
- Time dependent modeling can be done failure / repair rates.
- Repair can be done using rebuilding.
- MTTDL: mean time to data loss
- RAID 1: data is lost if the second disk fails before the first could be rebuilt.



Reference: Koren and Krishna

RAID1 - Reliability Calculation



- **Assumptions:**
  - **disks fail independently**
  - **failure process - Poisson process with rate λ**
  - **repair time - exponential with mean time 1/μ**
- **Markov chain: state - number of good disks**

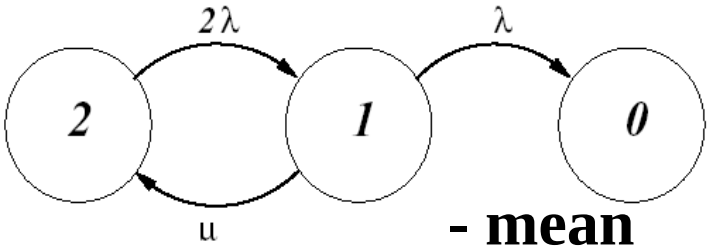
$$\frac{dP_2(t)}{dt} = -2\lambda P_2(t) + \mu P_1(t) \qquad \frac{dP_1(t)}{dt} = -(\lambda+\mu)P_1(t) + 2\lambda P_2(t)$$

$$P_0(t) = 1 - P_1(t) - P_2(t) \qquad P_2(0) = 1; \qquad P_0(0) = P_1(0) = 0$$

- **Reliability at time t -**

$$R(t) = P_1(t) + P_2(t) = 1 - P_0(t)$$

RAID1 - MTTDL Calculation



- mean

- Starting in state 2 at t=0
  time before entering state 1 = 1/(2λ)
- Mean time spent in state 1 is 1/(λ + μ)
- Go back to state 2 with probability q = μ /(μ + λ)
  or to state 0 with probability p = λ /(μ + λ)
- Probability of n visits to state 1 before transition to state 0 is

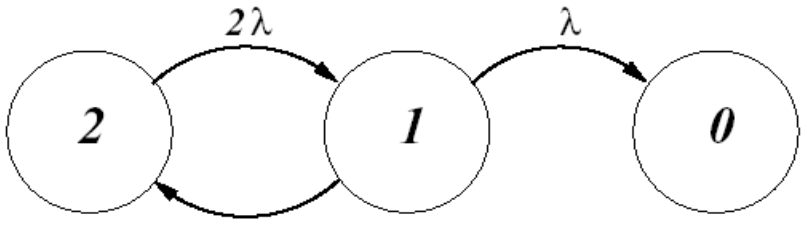
$$q^{n-1}p$$

- Mean time to enter state 0 with n visits to state 1:

$$T_{2\to 0}(n) = n\left(\frac{1}{2\lambda} + \frac{1}{\lambda+\mu}\right) = n\frac{3\lambda+\mu}{2\lambda(\lambda+\mu)}$$

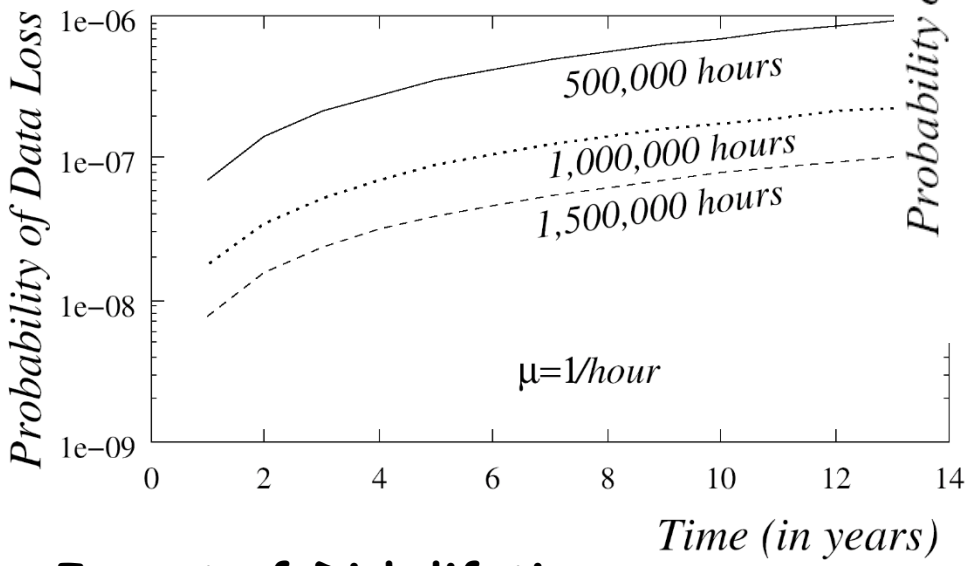
$$MTTDL = \sum_{n=1}^{\infty} q^{n-1}pT_{2\to 0}(n) = \sum_{n=1}^{\infty} nq^{n-1}pT_{2\to 0}(1) = \frac{T_{2\to 0}(1)}{p} = \frac{3\lambda+\mu}{2\lambda^2}$$

Approximate Reliability of RAID1

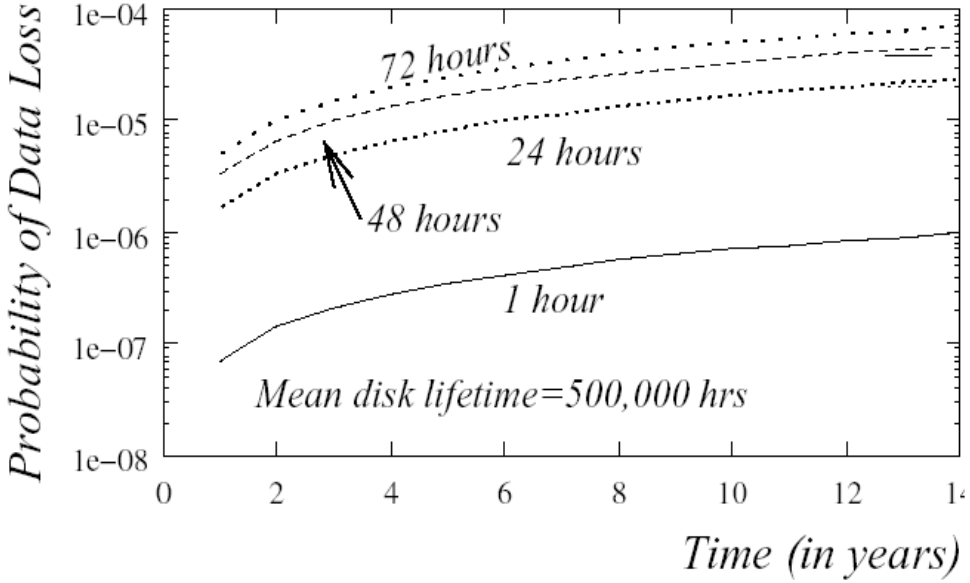


- If $\mu >> \lambda$, the transition rate into state 0 from the aggregate of states 1 and 2 is 1/MTTDL
- Approximate reliability:

$$R(t) = e^{-t/MTTDL}$$

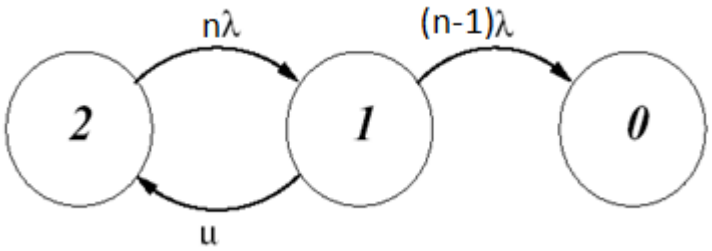


Impact of Disk lifetime



Impact of Disk Repair time

RAID4 - MTTDL Calculation



- RAID 4/5: data is lost if the second disk fails before the first failed (any one of n) could be rebuilt.

$$MTTDL = \frac{(2n-1)\lambda + \mu}{n(n-1)\lambda^2} \approx \frac{\mu}{n(n-1)\lambda^2}$$

- Detailed MTTDL calculators are available on the web:
  - https://www.servethehome.com/raid-calculator/raid-reliability-calculator-simple-mttdl-model/
  - https://wintelguy.com/raidmttdl.pl

Additional Reading

- Modeling the Reliability of Raid SetS – Dell
- Triple-Parity RAID and Beyond, Adam Leventhal, Sun Microsystems
- Estimation of RAID Reliability
- Enhanced Reliability Modeling of RAID Storage Systems

Notes