Redundant Array of Independent Disks RAID

- ## Disk Data Model

  a sequence of r same sized stripe units ("scsi" interface) with indicies 0,1,···,r-1. Stripe units may be (groups of) sectors, tracks, cylinders, or ···.

- ## User Data Model

  a sequence of same sized data stripe units with indicies 0,1,2,··· . Data stripe units typically created by file system, database management system, or other (sophisticated) user program.

- ## Reliability Groups a.k.a. Stripes

  user data is partitioned into fixed sized groups with each group containing additional redundancy information.

RAID I

- Disk Array Data Layouts

Data Layouts: map stripe unit indicies to disk sectors, tracks, or cylinders ⋯

n disks

r stripe units/disk

k stripe width

b number of stripes

g groups

$n = k \cdot g$

m user data stripe units/stripe

$c = k - m$ redundant stripe units/stripe

Total number of data stripe units $b(k - c)$.

Total number of parity stripe units $b \cdot c$.

The total number of stripe units within for a completely filled/utilized disk array is $bk = nr$.

RAID II

$d_0, d_1, d_2, d_3, \cdots$ user data

$\{ d_0, d_1, d_2, c_0 \}$ stripes

$\{ d_3, d_4, d_5, c_1 \}$

$\{ d_6, d_7, d_8, c_2 \}$

$\ddots$ c_i redundant data

data stripe unit indicies $DSUI = \{ 0, 1, 2, \ldots b(k\text{-}c)\text{-}1 \}$

disk: $DSUI \mapsto \{ 0, 1, 2, \ldots n\text{-}1 \}$

offset: $DSUI \mapsto \{ 0, 1, 2, \ldots r\text{-}1 \}$

checkDisk: $DSUI \times \{ 0, 1, \ldots, c\text{-}1 \} \mapsto \{ 0, 1, 2, \ldots, n\text{-}1 \}$

checkOffset: $DSUI \times \{ 0, 1, \ldots, c\text{-}1 \} \mapsto \{ 0, 1, 2, \ldots, r\text{-}1 \}$

RAID III

- Layout Taxonomy

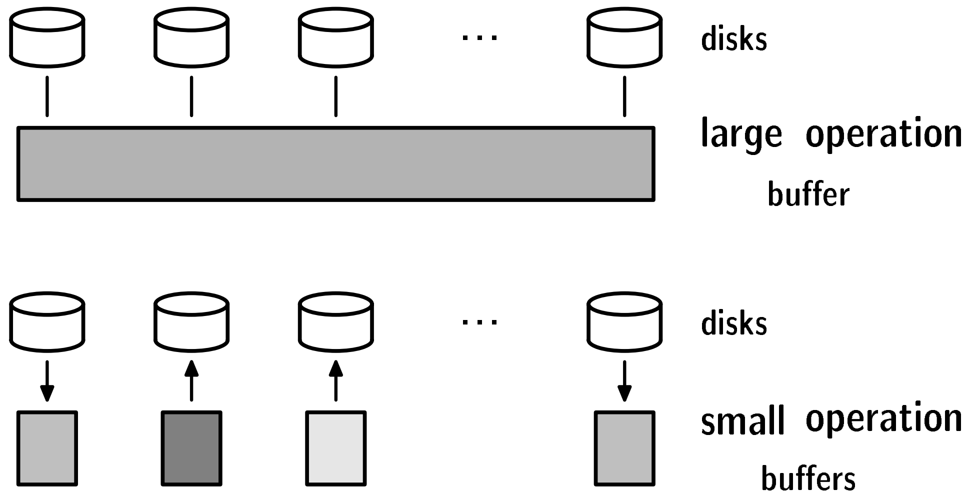
		typically
Level 0	just a bunch of disks JBOD, no redundancy	$m = k = 1$
Level 1	mirroring	$m = 1, k = 2$
Level 2	fine-grained interleaving with ECC error correction	$m = 10, k = 14$ $m = 20, k = 25$
Level 3	fine-grained interleaving with dedicated parity disk	$k = m+1$
Level 4	stripe unit interleaving; dedicated parity disk	$k = m+1$
Level 5	stripe-unit interleaving; distributed parity	$k = m+1$
Level 6	Level 5 with additional redundant stripe-units; typically one more	$k = m+2$

RAID IV

- Workloads

Large Operations Parallel read or write accesses one stripe unit from each disk; high data transfer rates obtained.

Small Operations Independent read or write accesses one data stripe unit from each disk; high numbers of i/o operations obtained.



disks

large operation buffer

disks

small operation buffers

- Operations small reads, writes, read-modify-writes (r m w) & large reads, writes, read-modify-writes.

- Relative Efficiency $$\frac{\text{RAID operations / sec.}}{\text{single disk operations / sec.}}$$

RAID V

- RAID Level 1 mirroring

stripe units	disk 0	disk 1
0:	d0	d0
1:	d1	d1
2:	d2	d2

$\ddots$

$$disk(a) = 0$$

$$offset(a) = \mathbf{a}$$

$$checkDisk(a) = 1$$

$$checkOffset(a) = \mathbf{a}$$

RAID VI

- **RAID Level 1** mirroring

g stripes $n = 2 \cdot g$ disks

$k = 2\,,\ m = 1$

		Relative Efficiency
small	read:	$2g$
	write:	g
	r m w:	$9g/8$
large	read:	$2g/s$
	write:	g/s
	r m w:	$9g/8s$

s slowdown: Within large operations, all disks complete their individual tasks to finish the operation; s is less than 2.

RAID VII

- RAID Level 2 fine-grain interleaving with ECC

disk 0	disk 1	disk 2	disk 3	disk 4	disk 5	disk 6
d0,3:0	d0,3:1	d0,3:2	d0,3:3	c0,3:4	c0,3:5	c0,3:6
d4,7:0	d4,7:1	d4,7:2	d4,7:3	c4,7:4	c4,7:5	c4,7:6
d8,11:0	d8,11:1	d8,11:2	d8,11:3	c8,11:4	c8,11:5	c8,11:6

$\ddots$

$$disk(a) = \{\, \mathbf{0, 1, \ldots, m\text{-}1} \,\}$$

$$offset(a) = \mathbf{a/m}$$

$$checkDisk(a) = \{\, \mathbf{m, m+1, \ldots, k\text{-}1} \,\}$$

$$checkOffset(a) = \mathbf{a/m}$$

RAID VIII

- RAID Level 2 fine-grain interleaving with ECC

g stripes $n = k \cdot g$ disks

$k = m + c$ stripe width

This level attempts to provide good performance with less redundant data.

		Relative Efficiency
small	read:	g/s
	write:	$g/2s$
	r m w:	g/s
large	read:	$g \cdot m/s$
	write:	$g \cdot m/s$
	r m w:	$g \cdot m/s$

RAID IX

Level 2 redundancy not needed; disk internal ecc guarantees one bit error in 10^{14}.

- RAID Level 3 fine-grained interleaving, dedicated parity disk

	disk 0	disk 1	disk 2	disk 3	disk 4
0:	d0,3:0	d0,3:1	d0,3:2	d0,3:3	c0,3
1:	d4,7:0	d4,7:1	d4,7:2	d4,7:3	c4,7
2:	d8,11:0	d8,11:1	d8,11:2	d8,11:3	c8,11

$\ddots$

$$disk(a) = \{\, 0, 1, \ldots, m\text{-}1 \,\}$$

$$offset(a) = a/m$$

$$checkDisk(a) = k\text{-}1 = m$$

$$checkOffset(a) = a/m$$

RAID X

- **RAID Level 3** fine-grained interleaving, dedicated parity disk

g stripes $n = k \cdot g$ disks

$k = m + 1$ stripe width

This level provides good performance with less redundant data. Level 2 and Level 3 performances are identical with one redundant disk.

		Relative Efficiency
small	read:	g/s
	write:	$g/2s$
	r m w:	g/s
large	read:	$g \cdot m \,/\, s$
	write:	$g \cdot m \,/\, s$
	r m w:	$g \cdot m \,/\, s$

RAID XI

Un-interleave data to improve small operations.

- RAID Level 4 stripe-unit interleaving,
  dedicated parity disk

	disk 0	disk 1	disk 2	disk 3	disk 4
0:	d0	d1	d2	d3	c0,3
1:	d4	d5	d6	d7	c4,7
2:	d8	d9	d10	d11	c8,11

$\ddots$

$$disk(a) = \mathrm{a\%m}$$

$$offset(a) = \mathrm{a/m}$$

$$checkDisk(a) = \mathrm{k\text{-}1} = \mathrm{m}$$

$$checkOffset(a) = \mathrm{a/m}$$

RAID XII

- **RAID Level 4** stripe-unit interleaving, dedicated parity disk

g stripes $n = k \cdot g$ disks

$k = m + 1$ stripe width

Level 4 provides better small operation performance than Level 3.

		Relative Efficiency
small	read:	$g \cdot m$
	write:	$g / 2$
	r m w:	g
large	read:	$g \cdot m / s$
	write:	$g \cdot m / s$
	r m w:	$g \cdot m / s$

RAID XIII

- **RAID Level 5** stripe-unit interleaving, distributed parity stripe units

Level 4 with distributed parity stripe units.

	disk 0	disk 1	disk 2	disk 3	disk 4
0:	d0	d1	d2	d3	p0,3
1:	d4	d5	d6	p4,7	d7
2:	d8	d9	p8,11	d10	d11
3:	d12	p12,15	d13	d14	d15
4:	p16,19	d16	d17	d18	d19

$\ddots$

$P = m - Q\%k \quad R = a\%m \quad Q = a/m$

$$disk(a) = \begin{cases} R & \text{if } R < P \\ R + 1 & \text{otherwise} \end{cases}$$

$$offset(a) = Q$$

$$checkDisk(a) = P$$

$$checkOffset(a) = Q$$

RAID XIV

- **RAID Level 5** stripe-unit interleaving, distributed parity stripe units

g stripes $n = k \cdot g$ disks

$k = m + 1$ stripe width

Level 5 provides better than Level 4 small operation performance.

		Relative Efficiency
small	read:	$g \cdot k$
	write:	$g \cdot k / 4$
	r m w:	$g \cdot k / 2$
large	read:	$g \cdot m / s$
	write:	$g \cdot m / s$
	r m w:	$g \cdot m / s$

RAID XV

- RAID Level 5 Left Symmetric layout

	disk 0	disk 1	disk 2	disk 3	disk 4
0:	d0	d1	d2	d3	p0,3
1:	d5	d6	d7	p4,7	d4
2:	d10	d11	p8,11	d8	d9
3:	d15	p12,15	d12	d13	d14
4:	p16,19	d16	d17	d18	d19

$\ddots$

$P = m - Q\%k \quad Q = a/m$

$disk(a) = a\%k$

$offset(a) = Q$

$checkDisk(a) = P$

$checkOffset(a) = Q$

RAID XVI

- RAID Level 6 stripe-unit interleaving, distributed check units

Level 5 with additional check stripe units.

	disk 0	disk 1	disk 2	disk 3	disk 4	disk 5
0:	d0	d1	d2	d3	c0,3:0	c0,3:1
1:	d4	d5	d6	c4,7:0	c4,7:1	d7
2:	d8	d9	c8,11:0	c8,11:1	d10	d11
3:	d12	c12,15:0	c12,15:1	d13	d14	d15
4:	c16,19:0	c16,19:1	d16	d17	d18	d19

$\ddots$

$P = m - Q\%k \quad R = a\%m \quad Q = a/m$

$$disk(a) = \begin{cases} R & \text{if } R < P \\ R + c & \text{otherwise} \end{cases}$$

$offset(a) = Q$

$checkDisk(a) = \{\, P, P{+}1, \ldots, P{+}(c{-}1) \,\}$

$checkOffset(a) = Q$

RAID XVII

- RAID Level 6 stripe-unit interleaving,
  distributed check stripe units

g stripes

$n = k \cdot g$ disks

$k = m+2$ typical stripe width

Level 6 provides better than Level 4 small operation performance as well as better reliability.

		Relative Efficiency
small	read:	$g \cdot k$
	write:	$g \cdot k / 4$
	r m w:	$g \cdot k / 2$
large	read:	$g \cdot m / s$
	write:	$g \cdot m / s$
	r m w:	$g \cdot m / s$

RAID XVIII

- RAID Level 6 “Left Symmetric” layout

	disk 0	disk 1	disk 2	disk 3	disk 4	disk 5
0:	d0	d1	d2	d3	c0,3:0	c0,3:1
1:	d6	d7	c4,7:0	c4,7:1	d4	d5
2:	c8,11:0	c8,11:1	d8	d9	d10	d11

$$\ddots$$

assume $m = c\lambda$; $P = m - (Q\%k)\lambda$ $Q = a/m$

$disk(a) = a\%k$

$offset(a) = Q$

$checkDisk(a) = \{ P, P+1, \ldots, P+(c-1) \}$

$checkOffset(a) = Q$