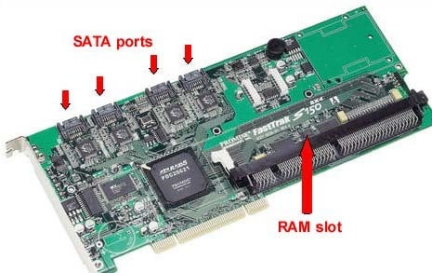
RAID

ICS332
Operating Systems

Henri Casanova (henric@hawaii.edu)

RAID

- Whenever we use a disk we'd like it to be faster, bigger, and/or more reliable
  - Reliability is less of an issue with SSDs, but it does not disappear
- Simple idea: Use a bunch of disks together to store our data
  - Increases reliability
    - If one fails, you have another one (increased perceived MTTF)
  - Increases speed
    - Aggregate disk bandwidth if data is split across disks
  - Increases size
    - Aggregate disk size
- Redundant Array of Independent Disks
  - RAID in software implemented at the OS level, or
  - An intelligent RAID controller in hardware, or
  - A "RAID array" as a stand-alone box
- From the outside, **it just looks like a single disk**
  - This is called transparency
  - (another term for virtualization really)

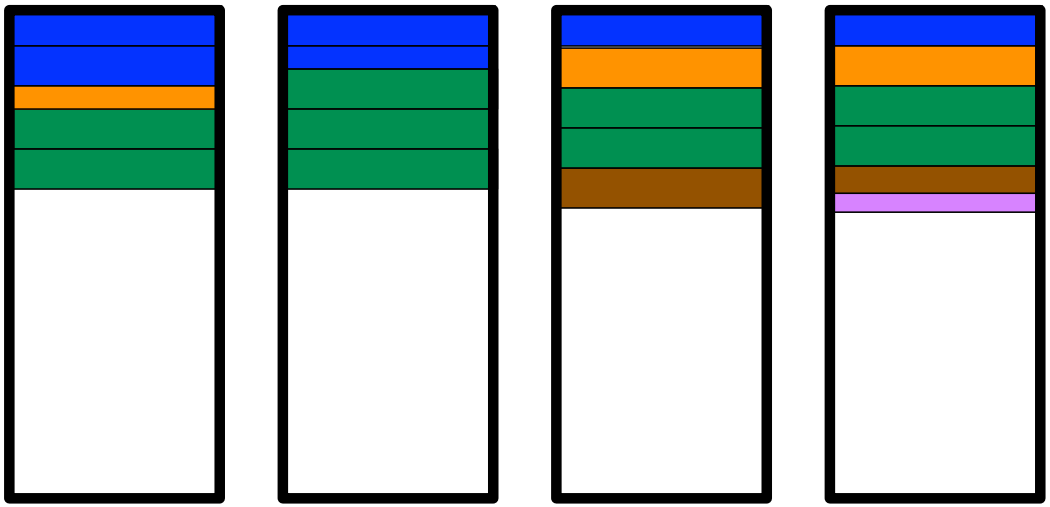


RAID Levels

- Different RAID configuration options or levels
- Each level corresponds to one or more techniques combined together
- Many levels are never used in practice
- But there are some basic levels you should know about: RAID-0, RAID-1, RAID-4, and RAID-5
  - Understanding the other levels if not hard as they're variations/combinations of similar ideas (e.g., different levels of granularity: bit, byte, block)
- Note that level naming is weird
  - RAID-50 doesn't mean there are 50 levels
  - It's really RAID 5 + 0

RAID-0: Striping

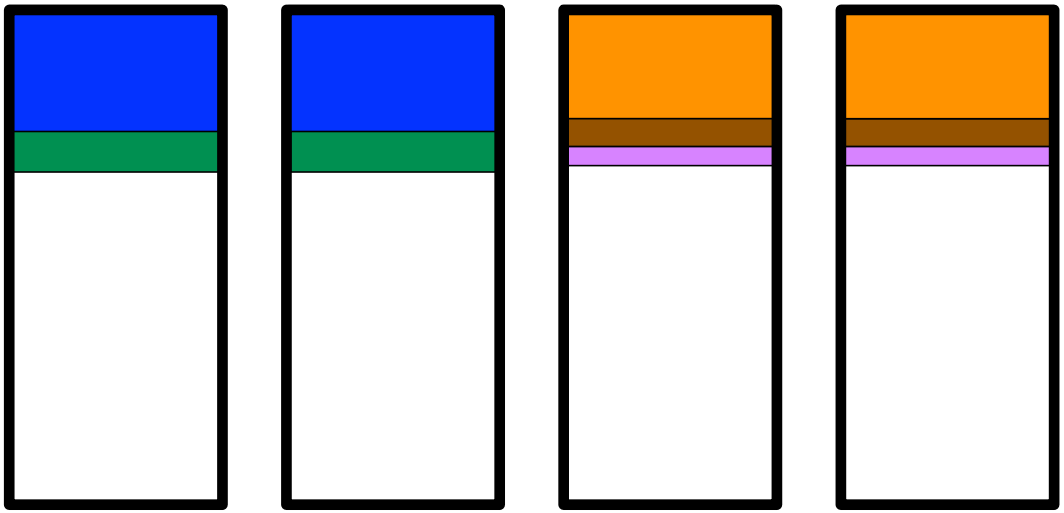
- Data is striped across multiple disks
  - Using a fixed strip size
- Gives the illusion of a larger disk with higher bandwidth when reading/writing a file
  - Accessing a single strip is not any faster
- **Improves performance, but not reliability**
- Useful for high-performance applications



- Fixed stripe size
- 5 files
- 4 disks

RAID-1: Mirroring

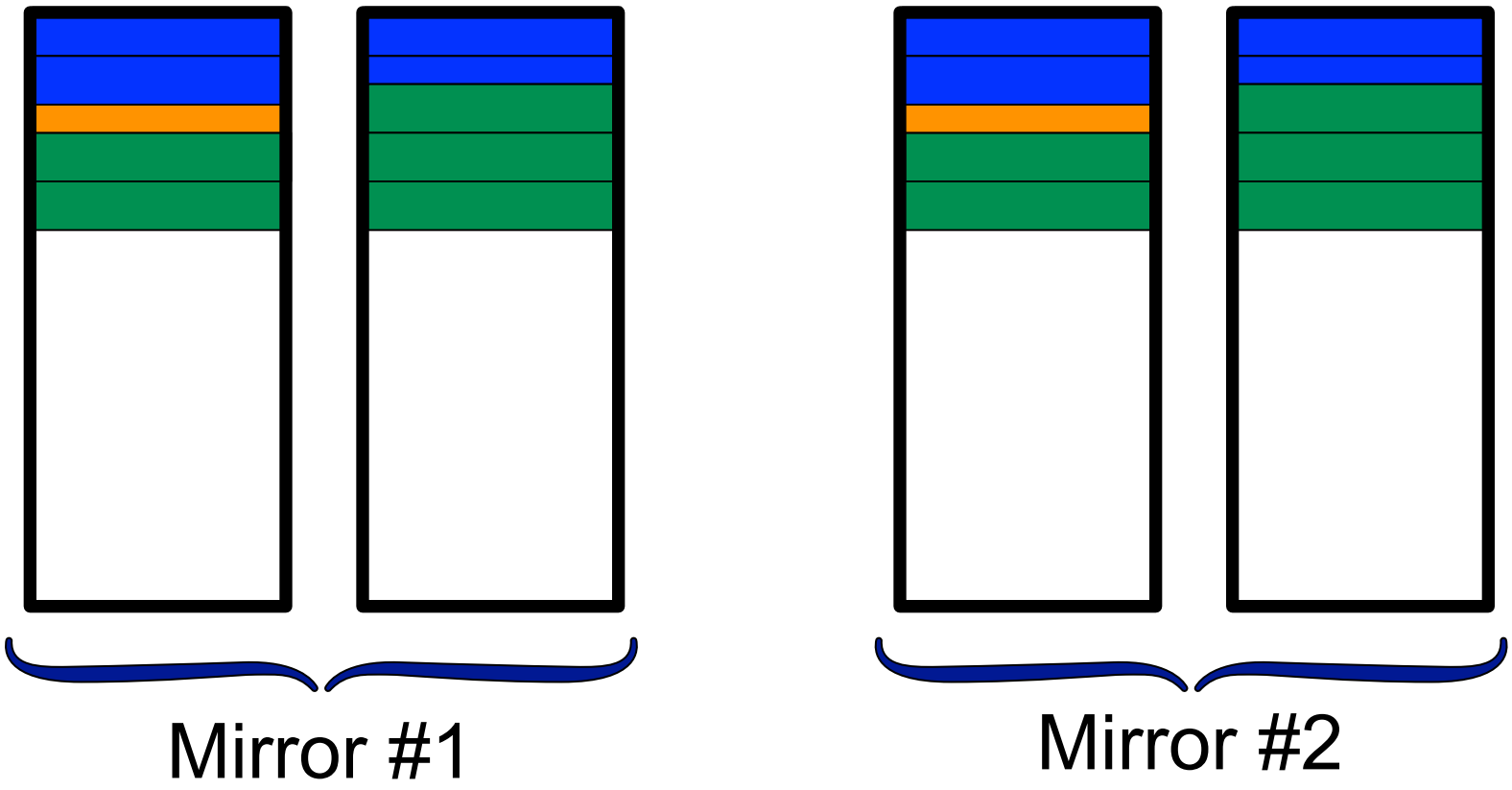
- Mirroring (also called shadowing)
- Write every written byte to 2 disks
  - Uses twice as many disks as RAID-0
- Vastly increases reliability unless you have (extremely unlikely) simultaneous failures
- Performance can be boosted a little bit by reading from the disk with the fastest seek time if using HDDs
  - The one with the arm the closest to the target cylinder



- 5 files
- 4 disks

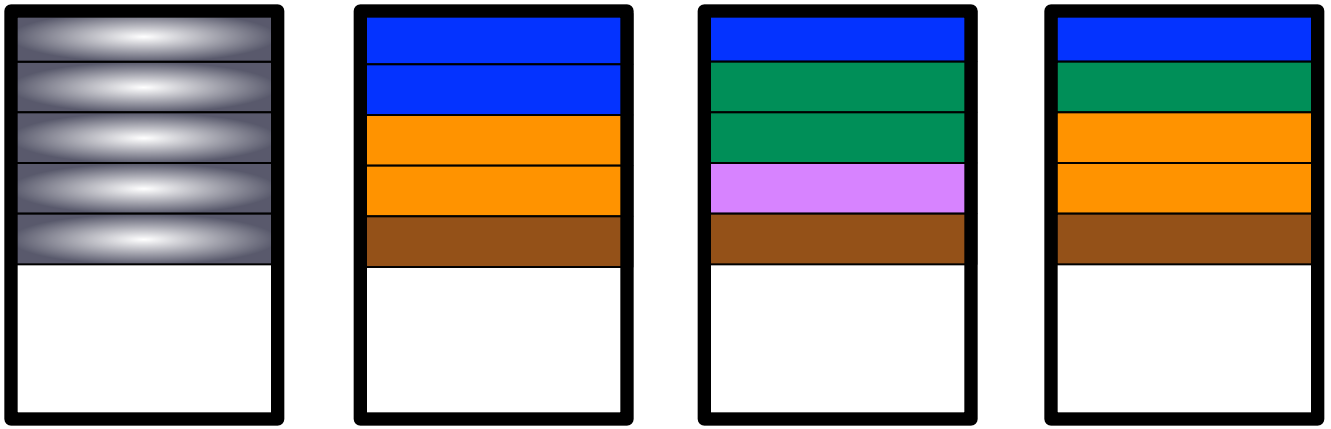
RAID-10 = 1 + 0

- One can do mirroring and striping within each mirror!



RAID-4: Parity

- Stores parity bits for each block/stripe so that lost data when one disk fails can be reconstructed
  - RAID-2: bit-level (rarely used), RAID-3: byte-level (rarely used)
- One disk stores all the parity bits



parity bits

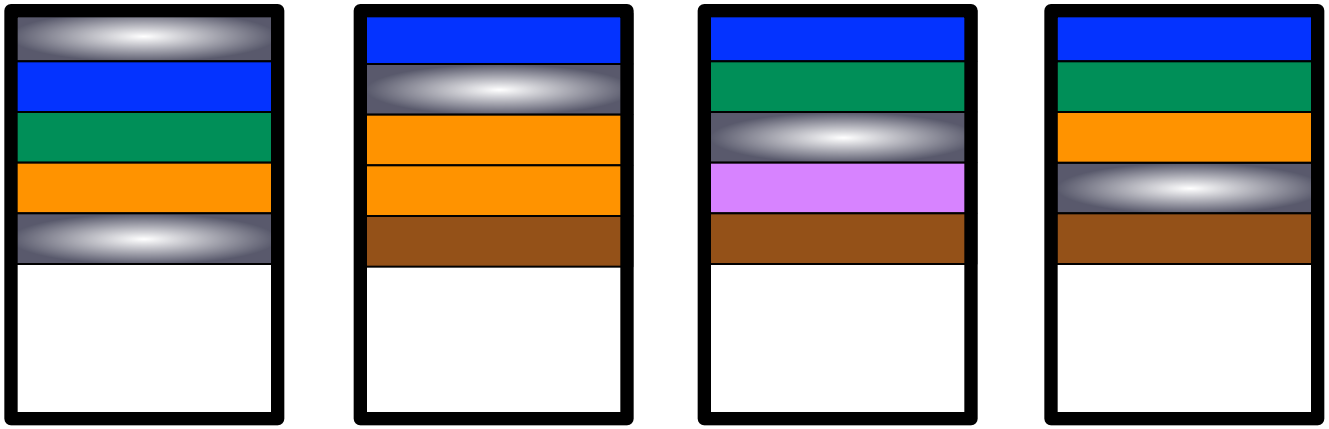
- What are the parity bits?…

RAID And Parity Bits

- Say you store 4 bits on a disk: **0 1 1 0**
- You compute a 5th bit (the parity bit) as the XOR of those bits: (((0 xor 1) xor 1) xor 0) =0
- You store that bit somewhere on another disk
  - So to store 4 bits, you use 5 bits
- Say you lose one bit: **0 ? 1 0**
- You can XOR the remaining bits with the parity bit to recover the lost bit: (((0 xor 0) xor 1) xor 0) = 1
- Say you lose a different bit: **0 1 1 ?**
- The XOR still works: (((0 xor 1) xor 1) xor 0) = 0

RAID-5: Parity Spread out

- All disks store some of the parity bits



- This is better for random writes because writes of parity bits can happen in parallel
- RAID-4 is almost never used and RAID-5 is preferred

RAID Levels: Comparison

- OSTEP has detailed analysis of all these RAID levels, and shows this summary table (*N*: # of disks, *B*: number of blocks or stripes per disk, *S*: sequential bandwidth, *R*: random bandwidth)

	RAID-0	RAID-1	RAID-4	RAID-5
Capacity	$N \cdot B$	$(N \cdot B)/2$	$(N-1) \cdot B$	$(N-1) \cdot B$
Reliability	0	1 (for sure) $\frac{N}{2}$ (if lucky)	1	1
Throughput				
Sequential Read	$N \cdot S$	$(N/2) \cdot S$	$(N-1) \cdot S$	$(N-1) \cdot S$
Sequential Write	$N \cdot S$	$(N/2) \cdot S$	$(N-1) \cdot S$	$(N-1) \cdot S$
Random Read	$N \cdot R$	$N \cdot R$	$(N-1) \cdot R$	$N \cdot R$
Random Write	$N \cdot R$	$(N/2) \cdot R$	$\frac{1}{2} \cdot R$	$\frac{N}{4} R$
Latency				
Read	T	T	T	T
Write	T	T	$2T$	$2T$

Let's explain this table…

RAID Levels: Comparison

- *N*: # of disks, *B*: number of blocks or stripes per disk, *S*: sequential bandwidth, *R*: random bandwidth

	RAID-0	RAID-1	RAID-4	RAID-5
Capacity	$N \cdot B$	$(N \cdot B)/2$	$(N-1) \cdot B$	$(N-1) \cdot B$

Number of blocks of real data that can be stored

"wastes" 1/2 the capacity

"wastes" only one parity disk capacity

RAID Levels: Comparison

- *N*: # of disks, *B*: number of blocks or stripes per disk, *S*: sequential bandwidth, *R*: random bandwidth

	RAID-0	RAID-1	RAID-4	RAID-5
Capacity	$N \cdot B$	$(N \cdot B)/2$	$(N-1) \cdot B$	$(N-1) \cdot B$
Reliability	0	1 (for sure) $\frac{N}{2}$ (if lucky)	1	1

Number of disk failures tolerated

Striping does nothing for reliability

Easy to understand on the example picture earlier

Thanks to parity bits

RAID Levels: Comparison

- *N*: # of disks, *B*: number of blocks or stripes per disk, *S*: sequential bandwidth, *R*: random bandwidth

	RAID-0	RAID-1	RAID-4	RAID-5
Capacity	$N \cdot B$	$(N \cdot B)/2$	$(N-1) \cdot B$	$(N-1) \cdot B$
Reliability	0	1 (for sure) $\frac{N}{2}$ (if lucky)	1	1
Throughput				
Sequential Read	$N \cdot S$	$(N/2) \cdot S$	$(N-1) \cdot S$	$(N-1) \cdot S$
Sequential Write	$N \cdot S$	$(N/2) \cdot S$	$(N-1) \cdot S$	$(N-1) \cdot S$

Number of blocks accessed per time unit

Full bandwidth

Half bandwidth

One disk is "wasted"

RAID Levels: Comparison

- *N*: # of disks, *B*: number of blocks or stripes per disk, *S*: sequential bandwidth, *R*: random bandwidth

	RAID-0	RAID-1	RAID-4	RAID-5
Capacity	$N \cdot B$	$(N \cdot B)/2$	$(N-1) \cdot B$	$(N-1) \cdot B$
Reliability	0	1 (for sure) $\frac{N}{2}$ (if lucky)	1	1
Throughput				
Sequential Read	$N \cdot S$	$(N/2) \cdot S$	$(N-1) \cdot S$	$(N-1) \cdot S$
Sequential Write	$N \cdot S$	$(N/2) \cdot S$	$(N-1) \cdot S$	$(N-1) \cdot S$
Random Read	$N \cdot R$	$N \cdot R$	$(N-1) \cdot R$	$N \cdot R$
Random Write	$N \cdot R$	$(N/2) \cdot R$	$\frac{1}{2} \cdot R$	$\frac{N}{4} R$

Full bandwidth

Half Bandwidth

The parity disk is the bottleneck: 1 read + 1 write to update the parity bits

We can keep all disks busy, and each performs 2 reads and 2 writes

RAID Levels: Comparison

- *N*: # of disks, *B*: number of blocks or stripes per disk, *S*: sequential bandwidth, *R*: random bandwidth

	RAID-0	RAID-1	RAID-4	RAID-5
Capacity	$N \cdot B$	$(N \cdot B)/2$	$(N-1) \cdot B$	$(N-1) \cdot B$
Reliability		r sure)	1	1
		f lucky)		
Throughput				
Sequential		$2) \cdot S$	$(N-1) \cdot S$	$(N-1) \cdot S$
Sequential		$2) \cdot S$	$(N-1) \cdot S$	$(N-1) \cdot S$
Random Read	$N \cdot R$	$N \cdot R$	$(N-1) \cdot R$	$N \cdot R$
Random Write	$N \cdot R$	$(N/2) \cdot R$	$\frac{1}{2} \cdot R$	$\frac{N}{4} R$
Latency				
Read	T	T	T	T
Write	T	T	$2T$	$2T$

One read + one write to update parity bits

Main Takeaways

- RAID can be used to:
  - Boost performance
  - Boost resilience
  - Boost both
- Different RAID levels have different properties for performance and resilience
- RAID can be implemented in software, but typically it's part of a hardware controller
- A key technique is to use parity bits computed via XOR to recover lost data

Conclusion

- RAID is used widely (nobody wants to lose data)
- Picking the level (and the parameters for each level) is a bit of a dark art
  - Done based on the intended workload, and often on hunches
- RAID-6 (which we haven't talked about) is often mentioned
  - It uses more parity data
    - So it has slower writes, and "wastes" more capacity
  - It allows for a drive to fail while another is being rebuilt
    - Often said to tolerate simultaneous failures (not the case strictly speaking)
  - Useful when drives are slow and large (HDDs), when reading happens more often than writing, and when losing data would be an utter catastrophe (isn't it always though?)