Western Digital®

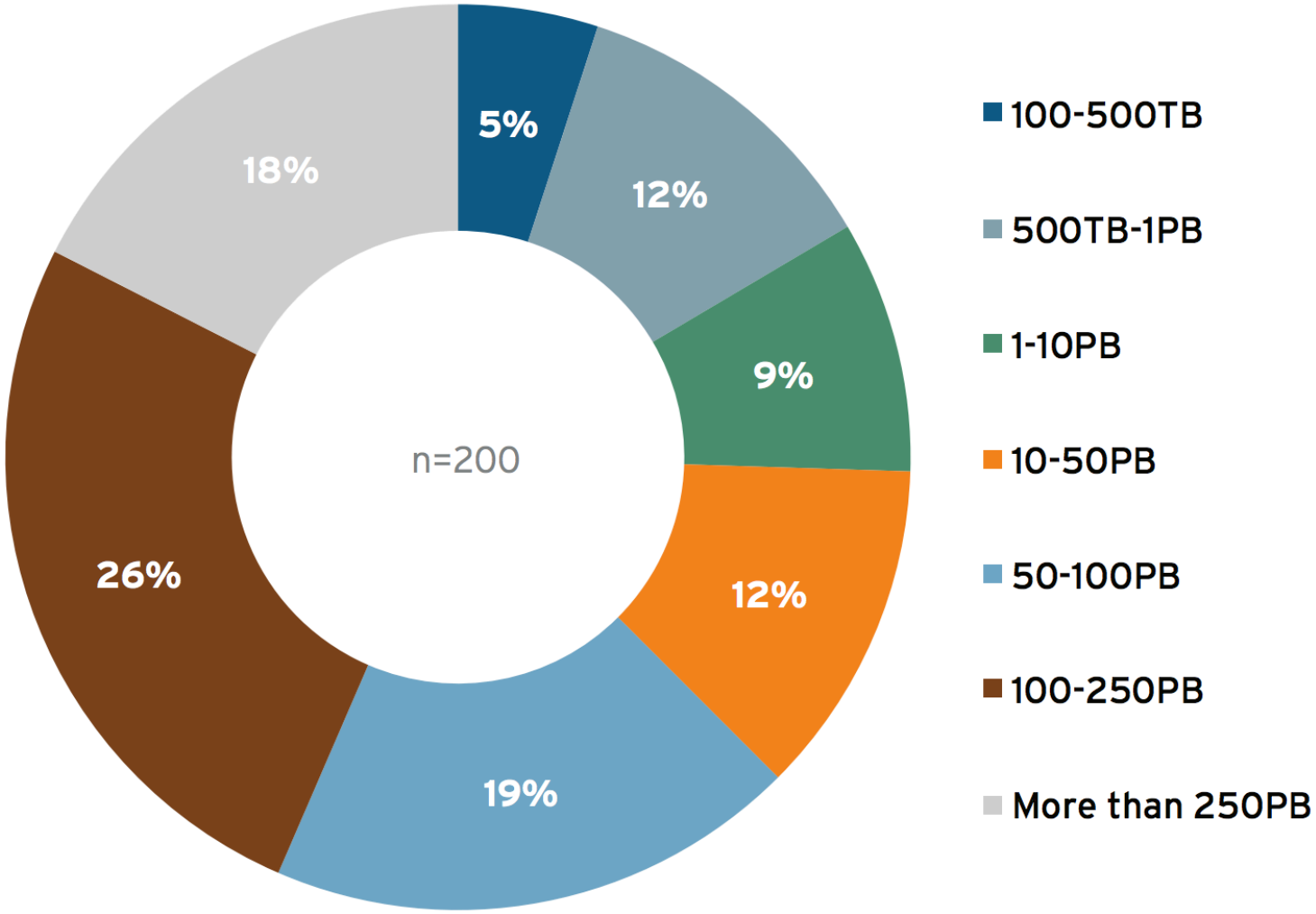
Life After RAID

Mike McWhorter
Senior Technologist

August 8, 2019

Figure 1: Total capacity under management

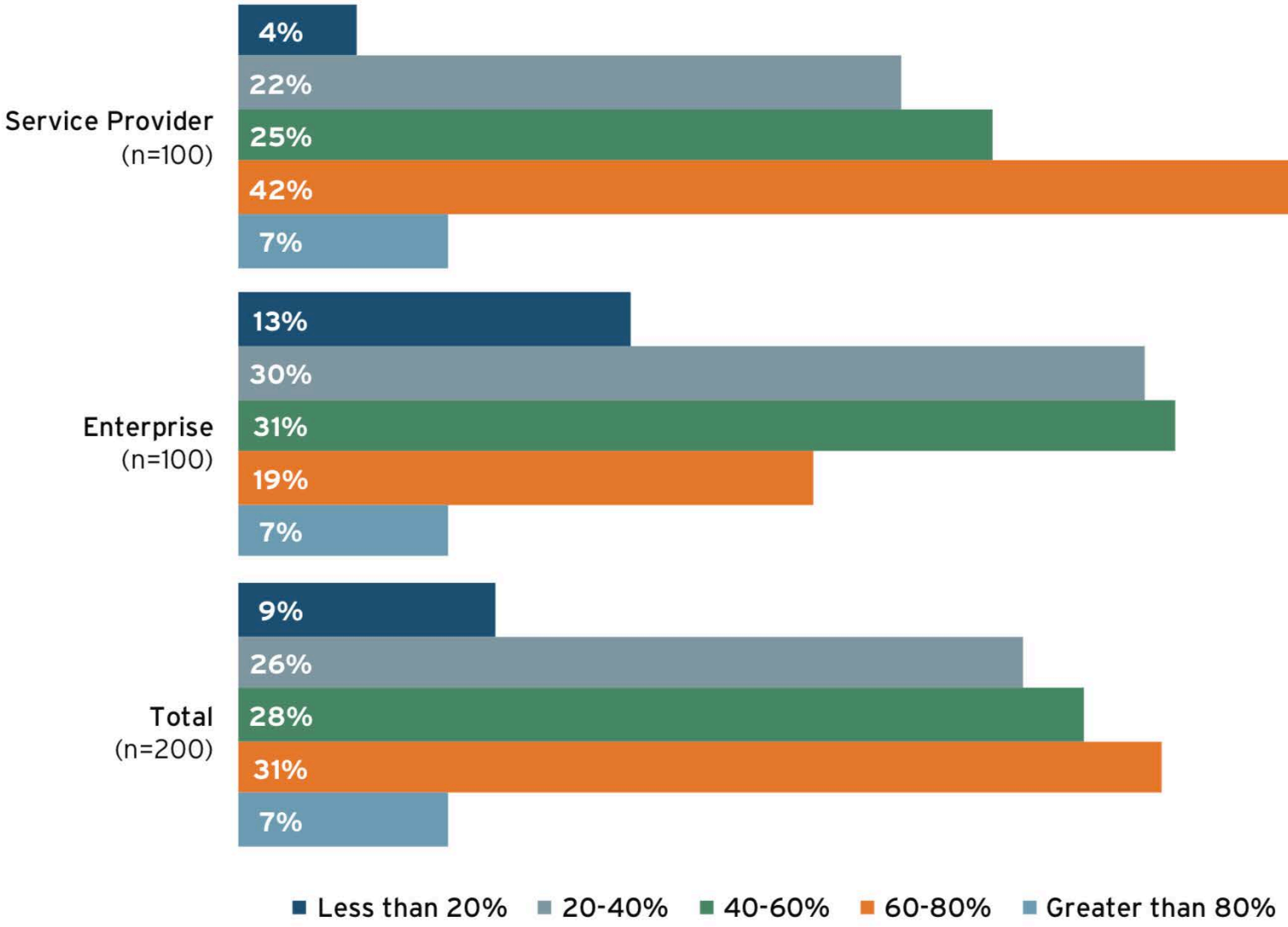
Q. What is your total estimated storage capacity (on-premises and public cloud combined)?



Source: 451 Research and Western Digital/HGST custom survey

Figure 2: Annual storage growth by organization type

Q. How much overall storage growth are you experiencing annually?



Source: 451 Research and Western Digital/HGST custom survey

Protecting Data with Replication

Original

Copy

Copy

Traditional RAID

Rebuild Times

Uncorrectable Bit Error Rate (UBER)

Interface transfer rate (MB/s, max)	600	1200
Sustained transfer rate[5] (MiB/s, typical) / (MB/s, typical)	255 / 267	←
Seek time[6] (read, ms, typical)	7.5	←
Reliability		
Error rate (non-recoverable bits read)	1 in 10^{15}	←
Load/Unload cycles (at 40oC)	600,000	←
Availability (hrs/day x days/wk)	24×7	←
MTBF[2] (M hours)	2.5	←
Annualized Failure Rate[2] (AFR)	0.35%	←
Limited warranty (yrs)	5	←

Weight (g, max)

Environmental

Ambient temperat

Shock (half-sine w

Vibration (G RMS,

Environmental

Ambient temperat

Shock (half-sine w

Vibration (G RMS,

* See "How to Read the

Uncorrectable Bit Error Rate (UBER)

10^{15} = 1,000,000,000,000,000 bits
10^{15} = 125,000,000,000,000 bytes
10^{15} = 125 TB

Where These Errors Come From

Errors can be introduced by…

- Power Fluctuations
- Electro-Magnetic Interference
- Drive Wear
- Excessive Heat
- Cosmic Rays
- Firmware Bugs
- Manufacturing Defects
- Bit Rot

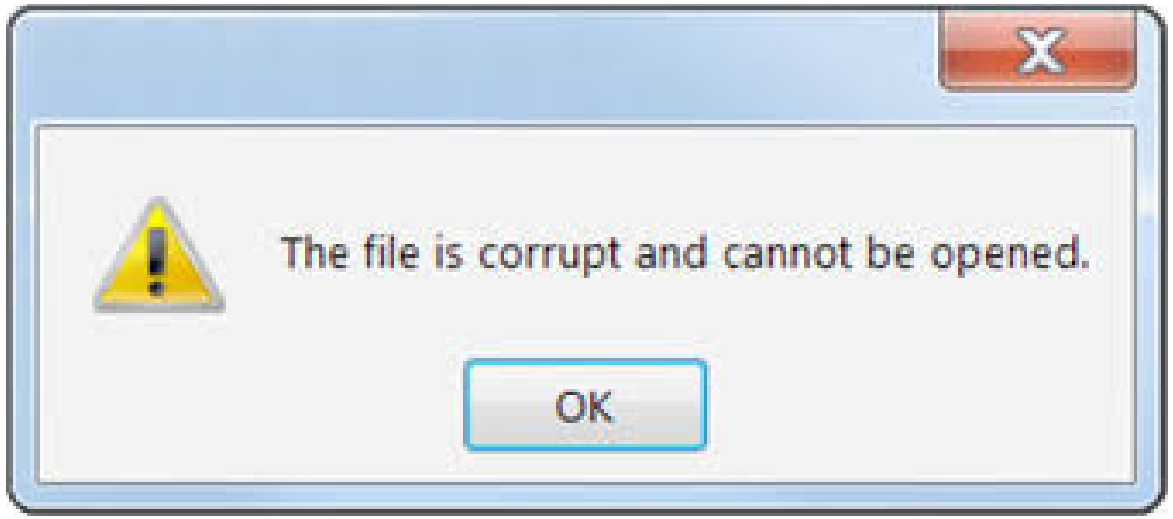
Unrecoverable Read Errors

```
[root@localhost]# dmesg

[  41.0300401] end-request:  I/O error, dev sda, sector 2452445
[  41.0312342] Buffer I/O error on device dm-1, logical block 9245
[  42.5038429] EXT4-fs error (device dm-1):  ext4_wait_block_bitmap:476: comm
 bounce:  Cannot read block bitmap - block_group = 83, block_bitmap = 1326442
```

Silent Errors

Some errors are not caught by the drive controller.



Uncorrectable Bit Error Rate (UBER)

Interface transfer rate (MB/s, max)	600	1200
Sustained transfer rate[5] (MiB/s, typical) / (MB/s, typical)	255 / 267	←
Seek time[6] (read, ms, typical)	7.5	←
Reliability		
Error rate (non-recoverable bits read)	1 in 10^{15}	←
Load/Unload cycles (at 40oC)	600,000	←
Availability (hrs/day x days/wk)	24×7	←
MTBF[2] (M hours)	2.5	←
Annualized Failure Rate[2] (AFR)	0.35%	←
Limited warranty (yrs)	5	←

Weight (g, max)
Environmental
Ambient temperat
Shock (half-sine w
Vibration (G RMS,
Environmental
Ambient temperat
Shock (half-sine w
Vibration (G RMS,

* See "How to Read the

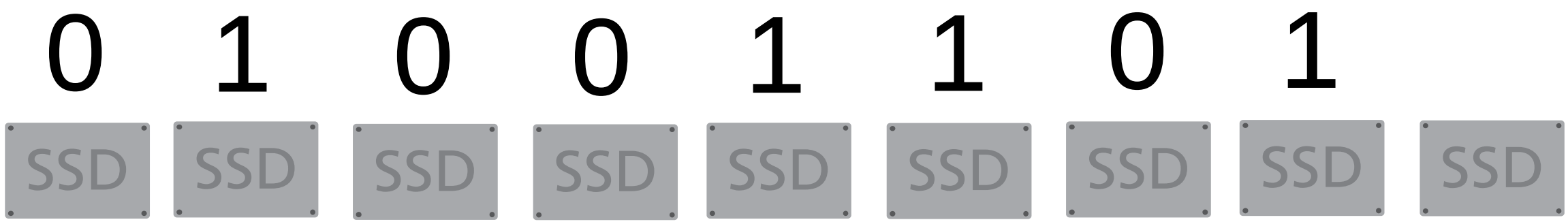
Uncorrectable Bit Error Rate (UBER)

1 in 10^{14} – 1 error per 12.5 TB
1 in 10^{15} – 1 error per 125 TB
1 in 10^{16} – 1 error per 1,250 TB
1 in 10^{17} – 1 error per 12,500 TB

What does this have to do with RAID?

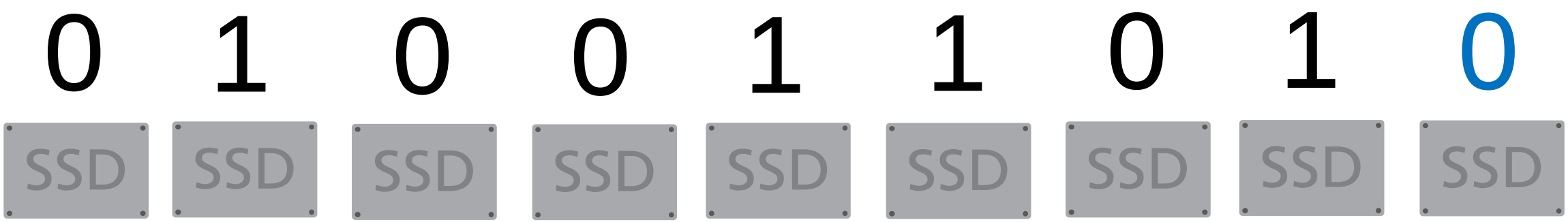
RAID 5 Protection

RAID relies on parity to protect your data.



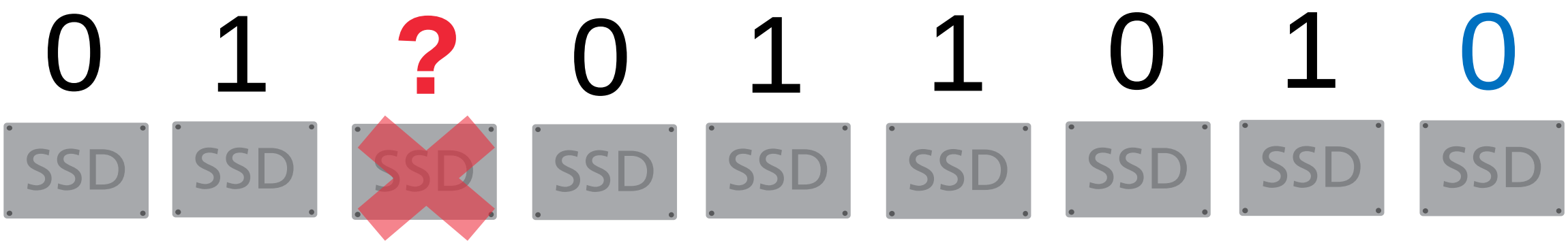
RAID 5 Protection

The parity is either **even** or **odd**.



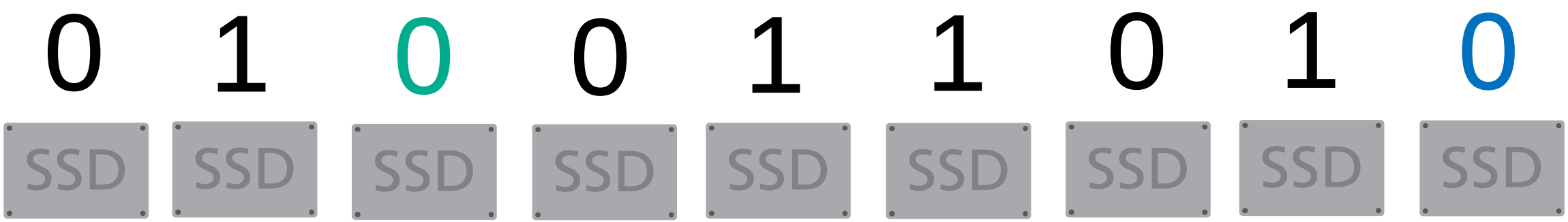
RAID 5 Protection

When a drive fails, we count the number of 1's.



RAID 5 Protection

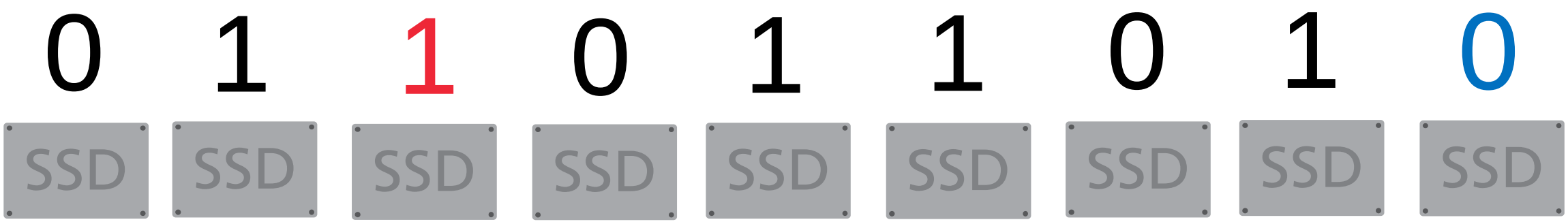
If a drive fails, you can calculate the missing data with an XOR operation.



$$0 \oplus 1 \oplus 0 \oplus 1 \oplus 1 \oplus 0 \oplus 1 \oplus 1 = 0$$

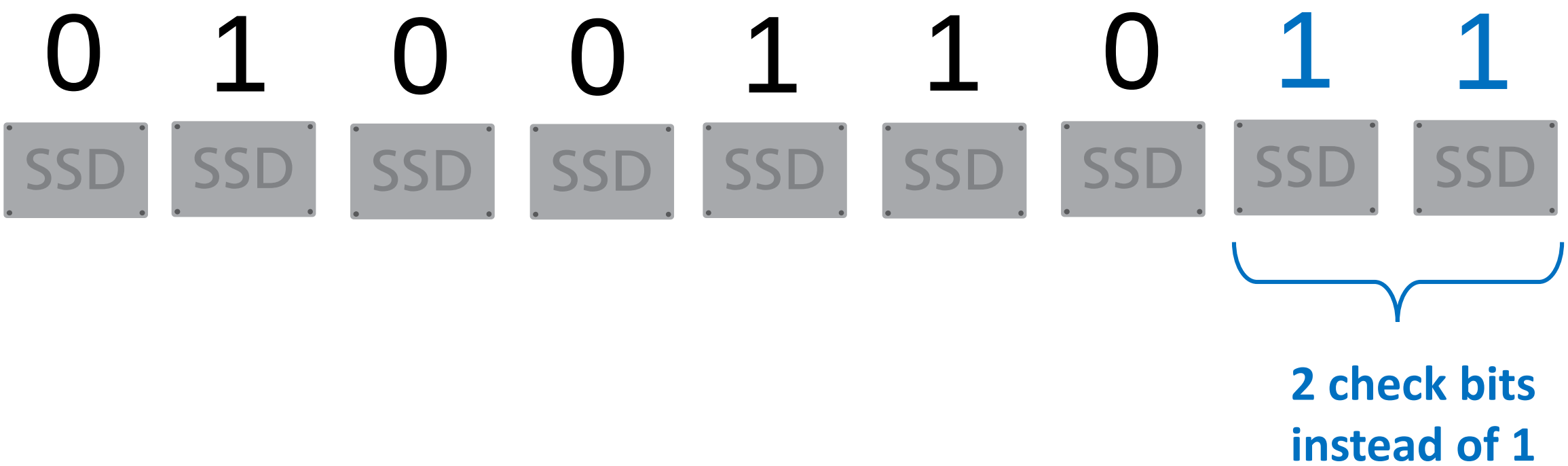
RAID 5 Protection

What do we do about a flipped bit?



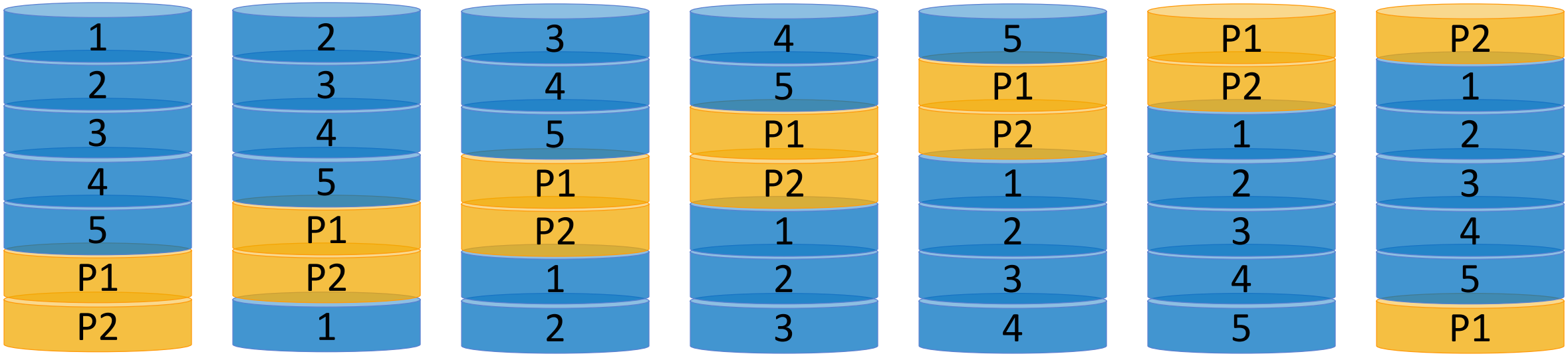
What About RAID 6?

RAID 6 uses dual “parity”.



Real World Data Layout

The parity blocks are distributed.

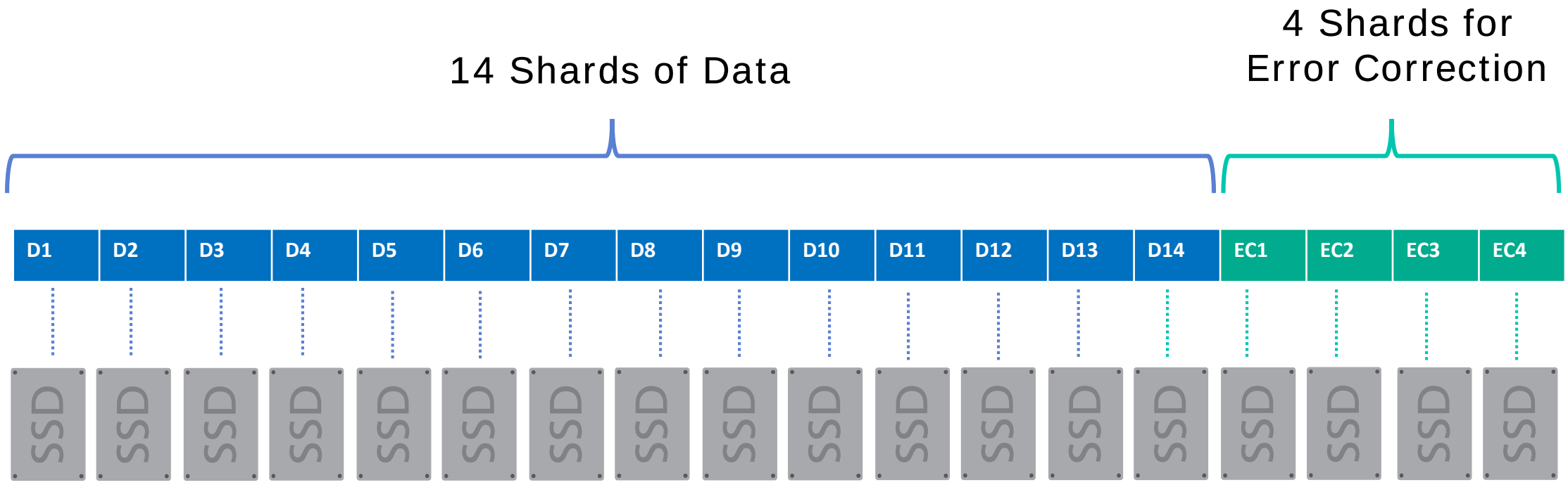


So what do we use?

Erasure Coding – Error Detection and Correction

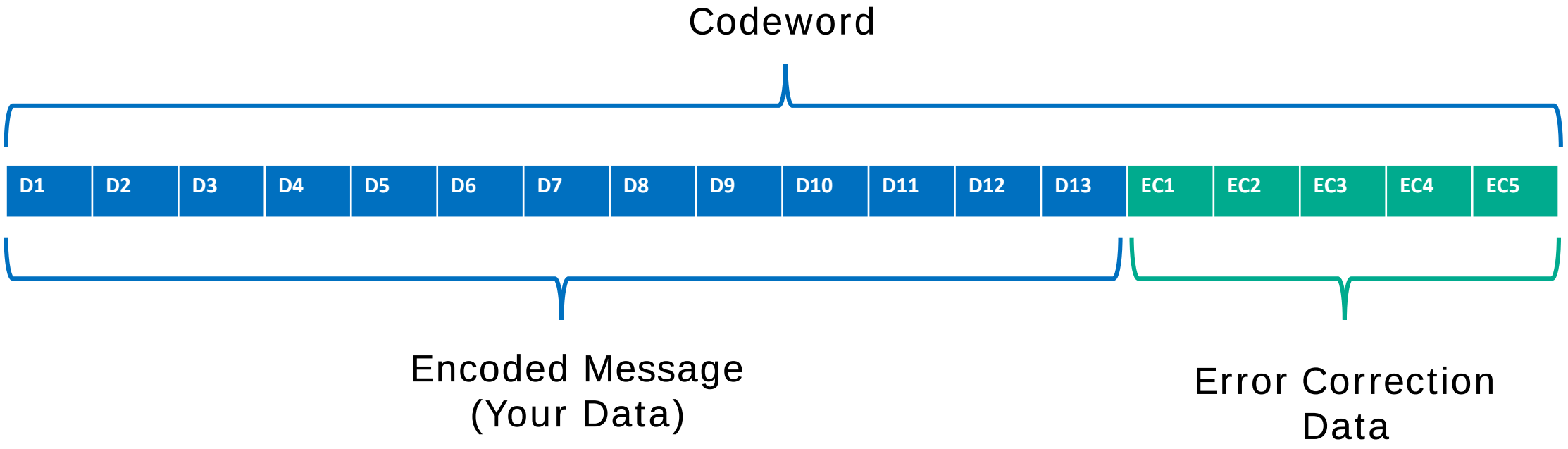
Reed-Solomon Erasure Coding

The Next Generation of RAID



Reed-Solomon EC Is Tunable

You can optimize for resiliency or capacity by adjusting the amount of error correction data.



Protection Levels

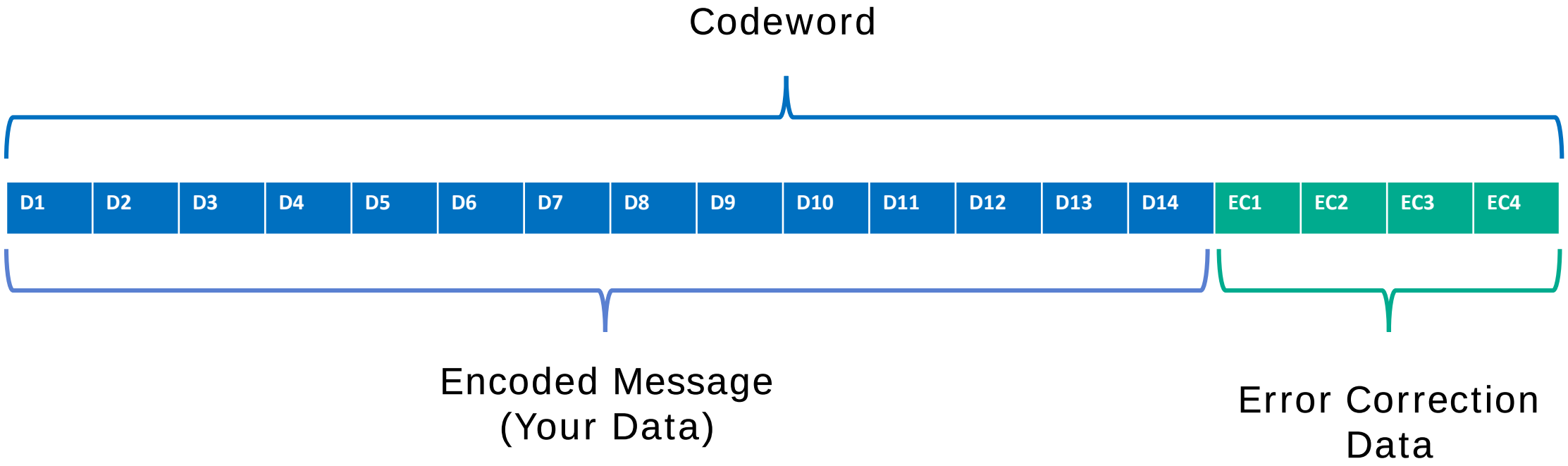
You can tune the codeword for capacity or redundancy.

(Data Shards + Error Correction Shards)

Correcting Erasures
(Failed Drives)

Reed-Solomon EC can correct n-k erasures.

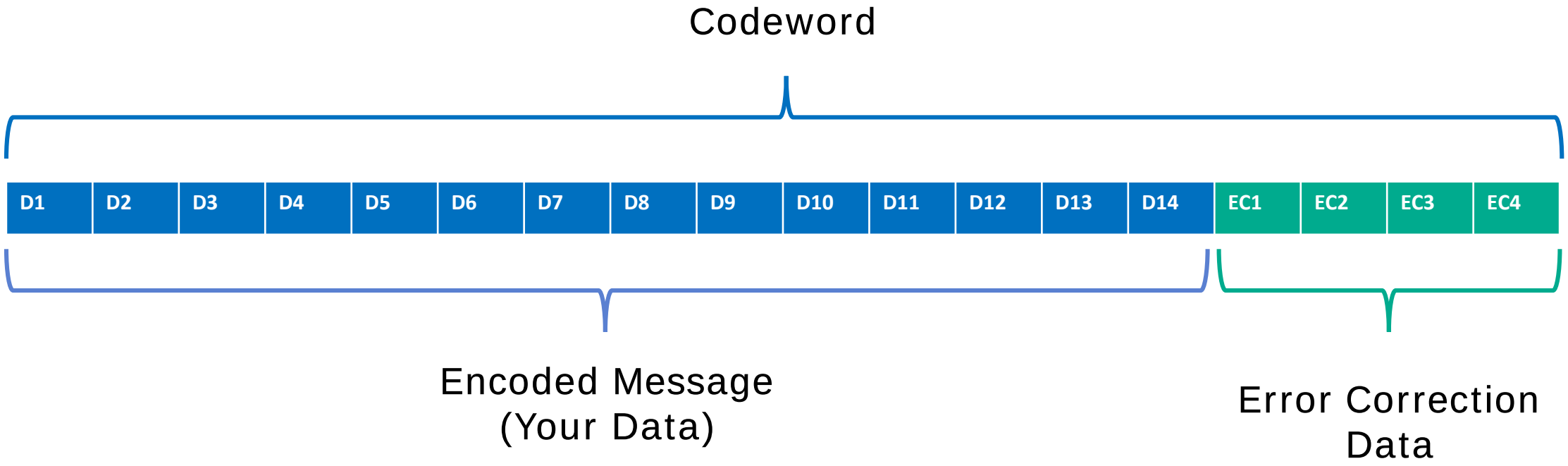
n=length of codeword
k=length of the encoded message



Correcting Bit Errors

Reed-Solomon EC can correct (n-k)/2 flipped bits.

n=length of codeword
k=length of the encoded message



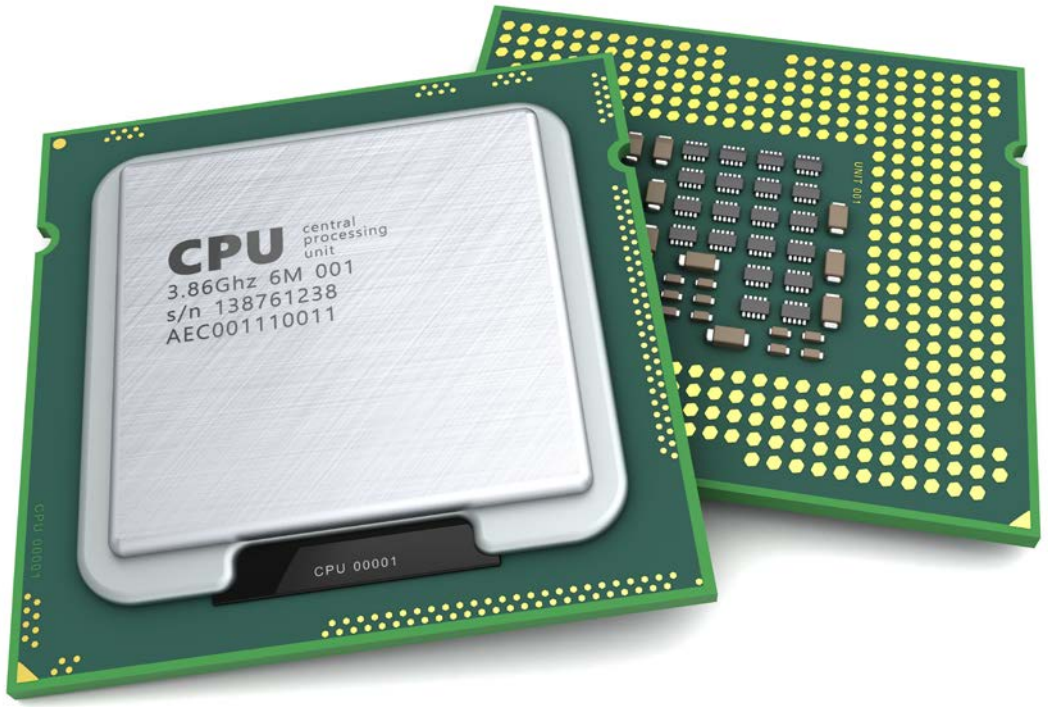
Protection Levels

You can tune the codeword for capacity or redundancy.

(Data Shards + Error Correction Shards)

Intel® Storage Acceleration Library (ISA-L)

Includes Native CPU instructions for Reed-Solomon Erasure Coding



Where Is Erasure Coding Used?

- Object Storage Systems
- Cloud Hosting Providers
- HDFS
- Optical Drives
- RAID 6 (sort of)

Using RAID Safely

- Avoid RAID 5.
- Use disks with a low error rate.
- Run consistency checks!
- Use a file system with ECC.
  - BTRFS
  - ZFS

Western Digital®

Architecting Data Infrastructure for the Zettabyte Age

Backup Slides

RAID 6 with RS Codes

Some versions of RAID 6 use Reed-Solomon Codes.