CSL373: Operating Systems
Fault Tolerance

Modularity = fault tolerance

- Modularity to control complexity
  - Names are the glue to compose modules
- Strong form of modularity: client/server
  - Limit propagation of errors
- Implementations of client/server:
  - In a single computer using virtualization
  - In a network using protocols
- Compose clients and services using names
  - DNS

How to respond to failures?

- Failures are contained; they don't propagate
  - Benevolent failures
- Can we do better?
  - Keep computing despite failures?
  - Defend against malicious failures (attacks)?
- handle these "failures"
  - Fault-tolerant computing
  - Computer security

Fault-tolerant computing

- General introduction:
  - Replication/Redundancy
- The hard case: transactions
  - updating permanent data in the presence of concurrent actions and failures
- Replication revisited: consistency


Windows
A fatal exception 0E has occurred at 0028:C00068F8 in PPT.EXE<01> +
000059F8. The current application will be terminated.
* Press any key to terminate the application.
* Press CTRL+ALT+DEL to restart your computer. You will
lose any unsaved information in all applications.
Press any key to continue

Availability in practice

- Carrier airlines (2002 FAA fact book)
  - 41 accidents, 6.7M departures
  - ✓ 99.9993% availability
- 911 Phone service (1993 NRIC report)
  - 29 minutes per line per year
  - ✓ 99.994%
- Standard phone service (various sources)
  - 53+ minutes per line per year
  - ✓ 99.99+%
- End-to-end Internet Availability
  - ✓ 95% - 99.6%

PRODUCT OVERVIEW

Cheetah 15K.4

Mainstream enterprise disc drive

Simply the best price/ performance, lowest cost of ownership disc drive ever

KEY FEATURES AND BENEFITS

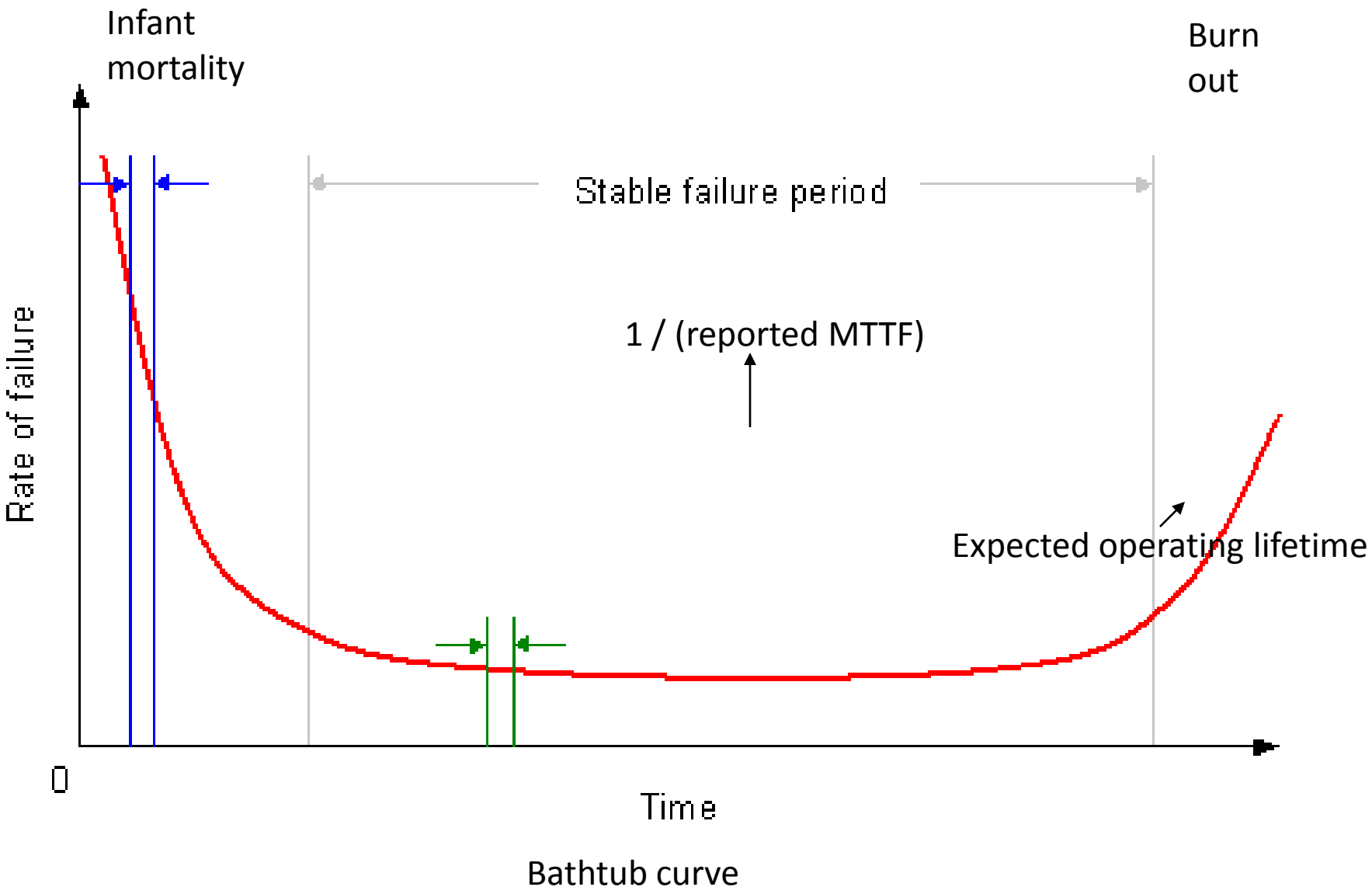
- The Cheetah® 15K.4 is the highest-performance drive ever offered by Seagate®, delivering maximum IOPS with fewer drives to yield lower TCO.
- The Cheetah 15K.4 price-per-performance value united with the breakthrough benefits of serial attached SCSI (SAS) make it the optimal 3.5-inch drive for rock solid enterprise storage.
- Proactive, self-initiated background management functions improve media integrity, increase drive efficiency, reduce incidence of integration failures and improve field reliability.
- The Cheetah 15K.4 shares its electronics architecture and firmware base with Cheetah 10K.7 and Savvio™ to ensure greater factory consistency and reduced time to market.

KEY SPECIFICATIONS

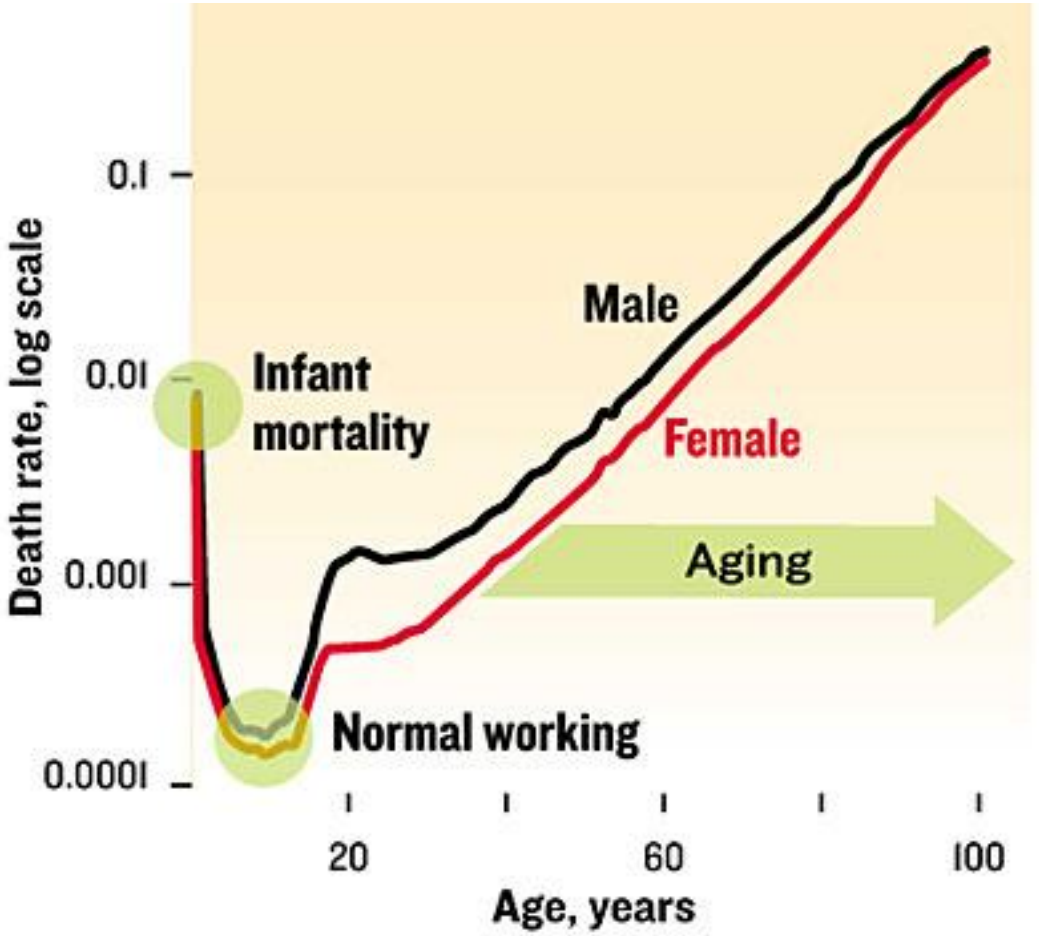
- 146-, 73- and 36-Gbyte capacities
- 3.3-msec average read and 3.8-msec average write seek times
- Up to 96-Mbytes/sec sustained transfer rate
- 1.4 million hours full duty cycle MTBF
- Serial Attached SCSI (SAS), Ultra320 SCSI and 2 Gbits/sec Fibre Channel interfaces
- 5-year warranty

For more information on why 15K is the industry's best price/performance disc drive for use in mainstream storage applications, visit ***http://specials.seagate.com/15k***

Disk failure conditional probability distribution



Bathtub curve



Human Mortality Rates (US, 1999)

From: L. Gavrilov & N. Gavrilova, "Why We Fall Apart," IEEE Spectrum, Sep. 2004.
Data from http://www.mortality.org

Disk Performance

- Throughput: 125 requests/second
- Bandwidth: 20-200MB/s (max) 15-30MB/s(sustained)

- Speed gap between disks and CPU/Memory is widening
  - CPU speed increases @ 60%/year
  - Disks speed increas @ 10-15%/year

- Improvement in disk technologies impressive in capacity/cost area

- Single Large Expensive Disk (SLED)

Fail-fast disk

```
failfast_get (data, sn) {
        get (s, sn);
        if (checksum(s.data) = s.cksum) {
                data ← s.data;
                return OK;
        } else {
                return BAD;
        }
}
```

Careful disk

```
careful_get (data, sn) {
        r ← 0;
        while (r < 10) {
                r ← failfast_get (data, sn);
                if (r = OK) return OK;
                r++;
        }
        return BAD;
}
```

Durable disk (RAID 1)

```
durable_get (data, sn) {
        r ← disk1.careful_get (data, sn);
        if (r = OK) return OK;
        r ← disk2.careful_get (data, sn);
        signal(repair disk1);
        return r;
}
```

Improvement of Reliability via Redundancy

- As the number of disks per component increases, the probability of failure also increases
  - Suppose a (reliable) disk fails every 100,000 hours. Reliability of a disk in an array of N disks = 100,000/N.
  - 100,000/100 = 1000 hours = 41.66 days!

- Solution?
  - Redundancy

Redundancy

- Mirroring

- Data Striping

Reliability in Mirroring

- Suppose mean time to repair is 10 hours, the mean time to data loss of a mirrored disk system is:

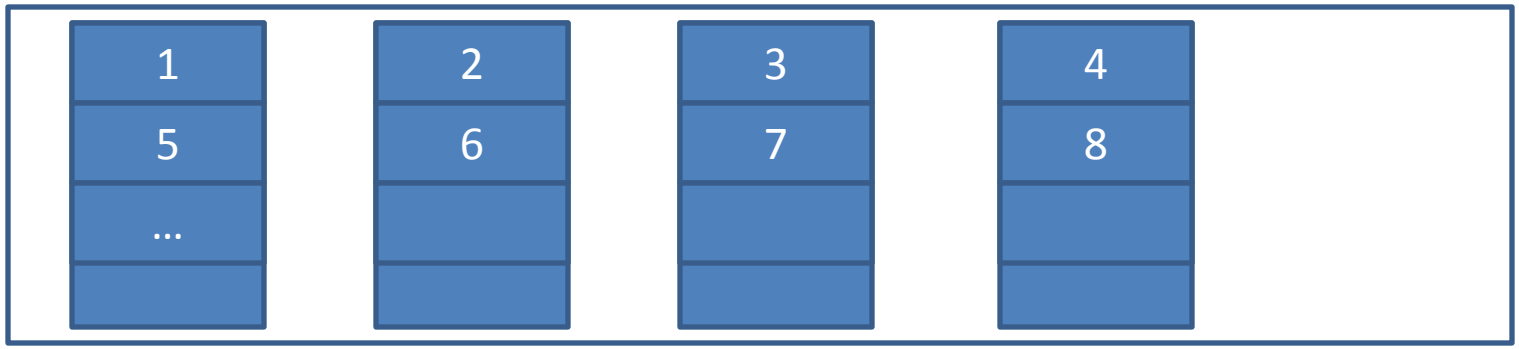
  (100,000^2)/(2*10) hrs ~ 57,000 years!

- Main disadvantage: most expensive approach

Parallel Disk Systems

- We cannot improve disk performance significantly as a single drive. But, could we combine the power of many drives?

- Solutions:
  - Parallel Disk Systems
  - Higher Reliability and Higher data-transfer rate

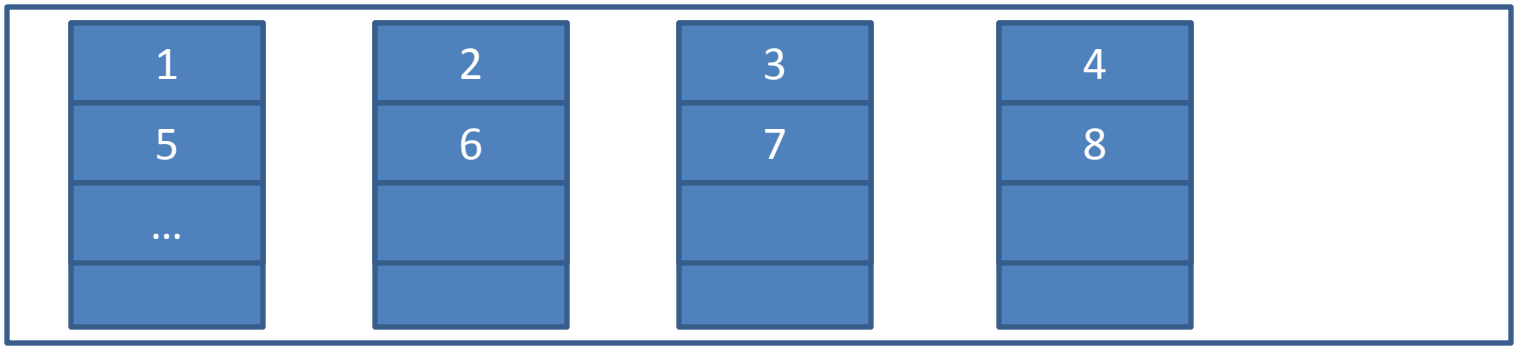


Data Striping

- Fundamental to RAID
- A method of concatenating multiple drives into one logical storage unit
- Splitting the bits of each byte across multiple disks: *bit-level striping*
  - E.g., an array of eight disks, write bit *i* of each byte to disk *i*
- Sectors are eight times the normal size
- Eight times the access rate
- Similarly for blocks of file, *block-level striping*

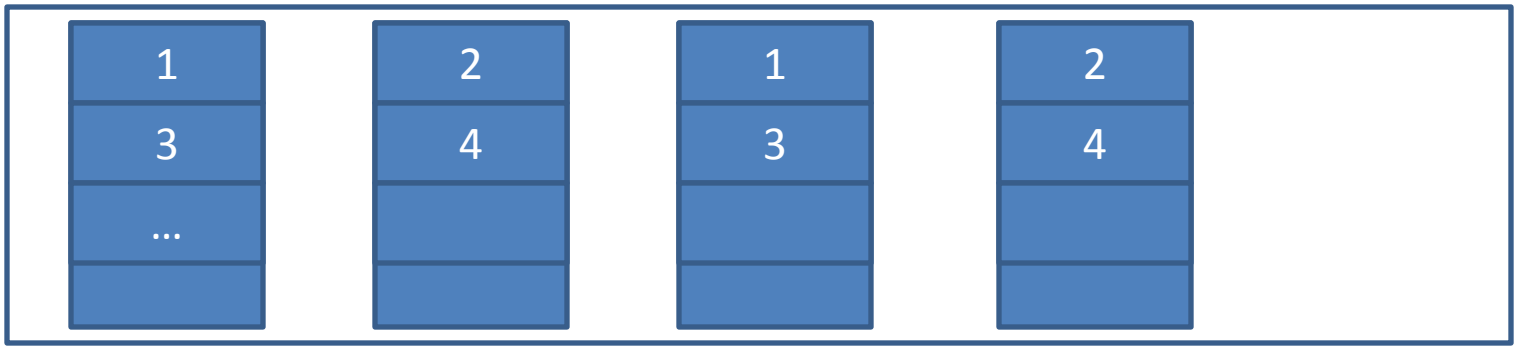
RAID 0

- Striping at the level of blocks
- No redundancy, hence reliability problems



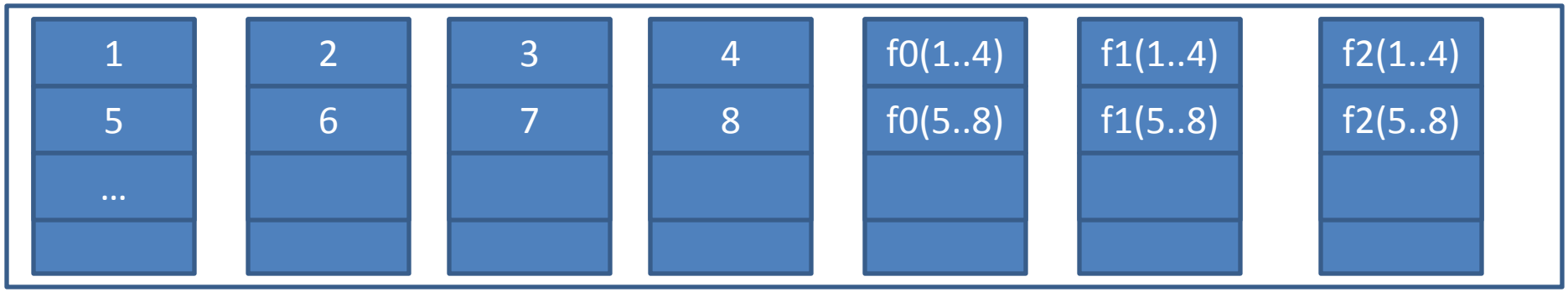
RAID 1 (Mirroring)

- Introduce redundancy through mirroring
- Expensive (cost/MB)
- Performance Issues



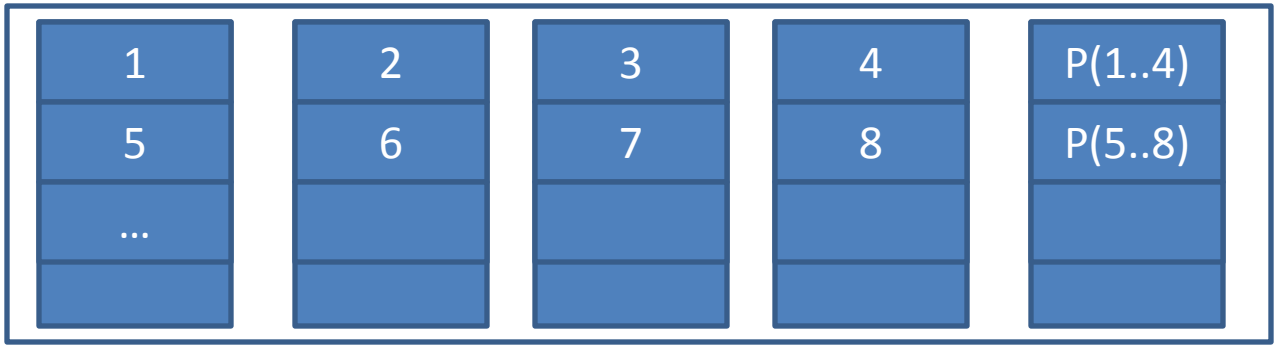
RAID 2

- Uses Hamming (or any other) error-correcting code (ECC)
- Intended for use in drives which do not have in-built error detection
- Central Idea: If one of the disks fail, the remaining bits of the byte and the associated ECC bits can be used to reconstruct the data



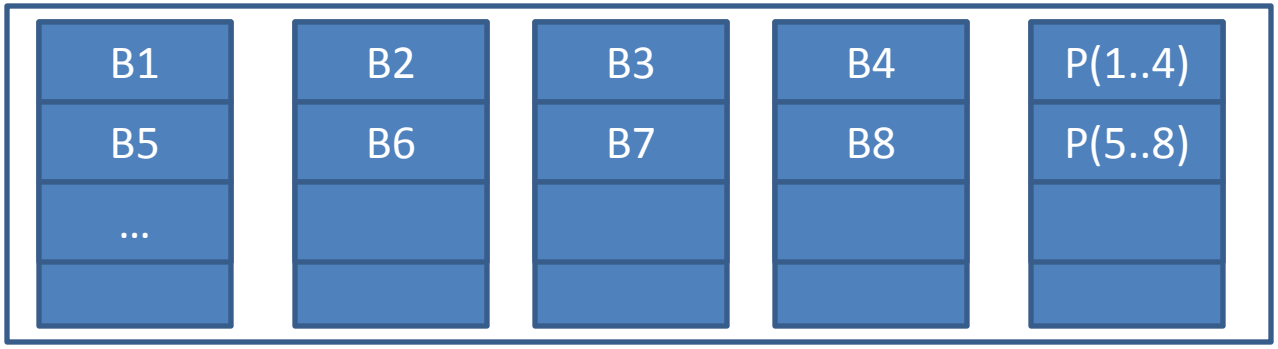
RAID 3 (Bit-interleaved parity)

- Disk Controllers can detect whether a sector has been read correctly
- Storage overhead reduced – only 1 parity disk
- Expense of computing and writing parity
- Need to include a dedicated parity hardware



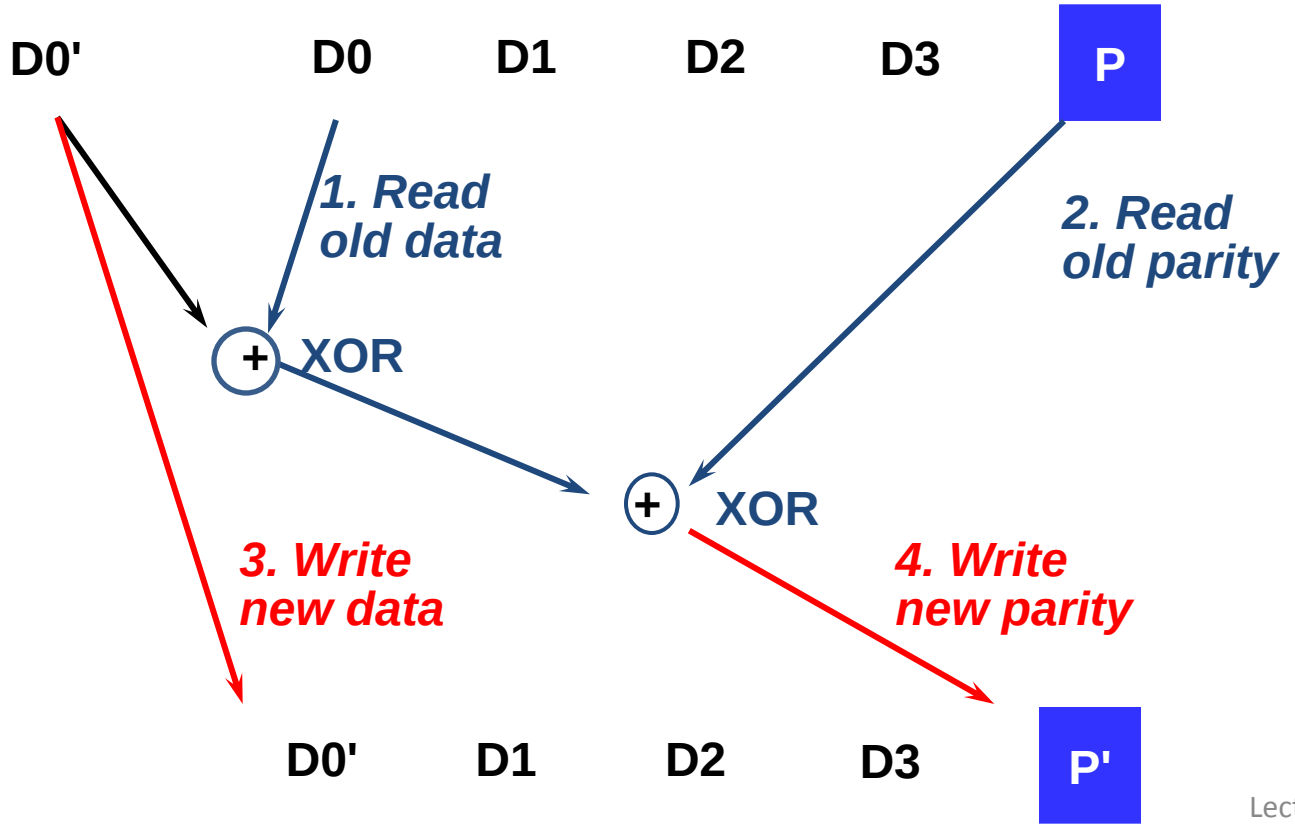
RAID 4 (block-interleaved parity)

- Stripes data at a block level across several drives with parity stored on one drive
- Allows recovery from the failure of any of the disks
- Performance is very good for reads
- Writes require that parity data be updated each time. Slows small random writes, but large writes are fairly fast



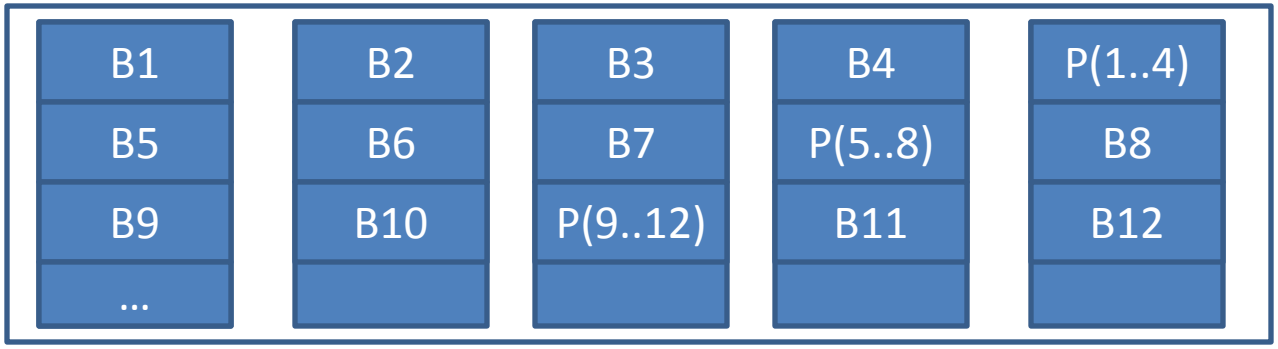
Problem of Disk Arrays: Small Writes

RAID-5: Small Write Algorithm



RAID 5 (Block-Interleaved distributed parity)

- Spreads data and parity among N+1 disks, rather than storing data in N disks, and parity in 1 disk
- Avoids potential overuse of single parity disk
- Most common parity RAID system

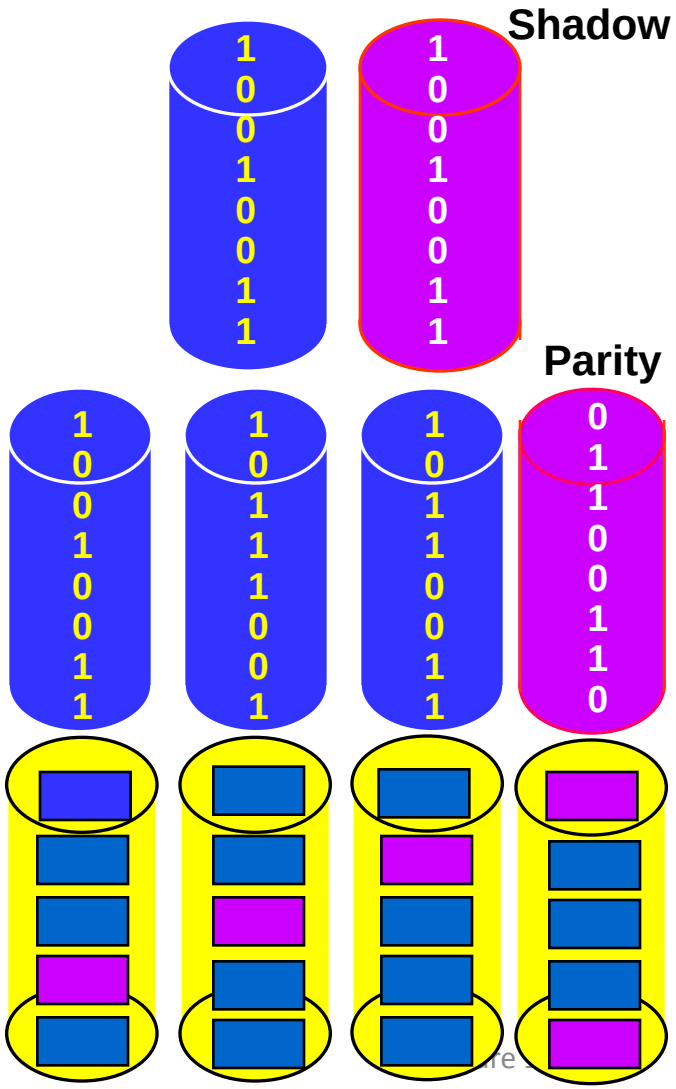


Redundant Array of Inexpensive Disks (RAID)

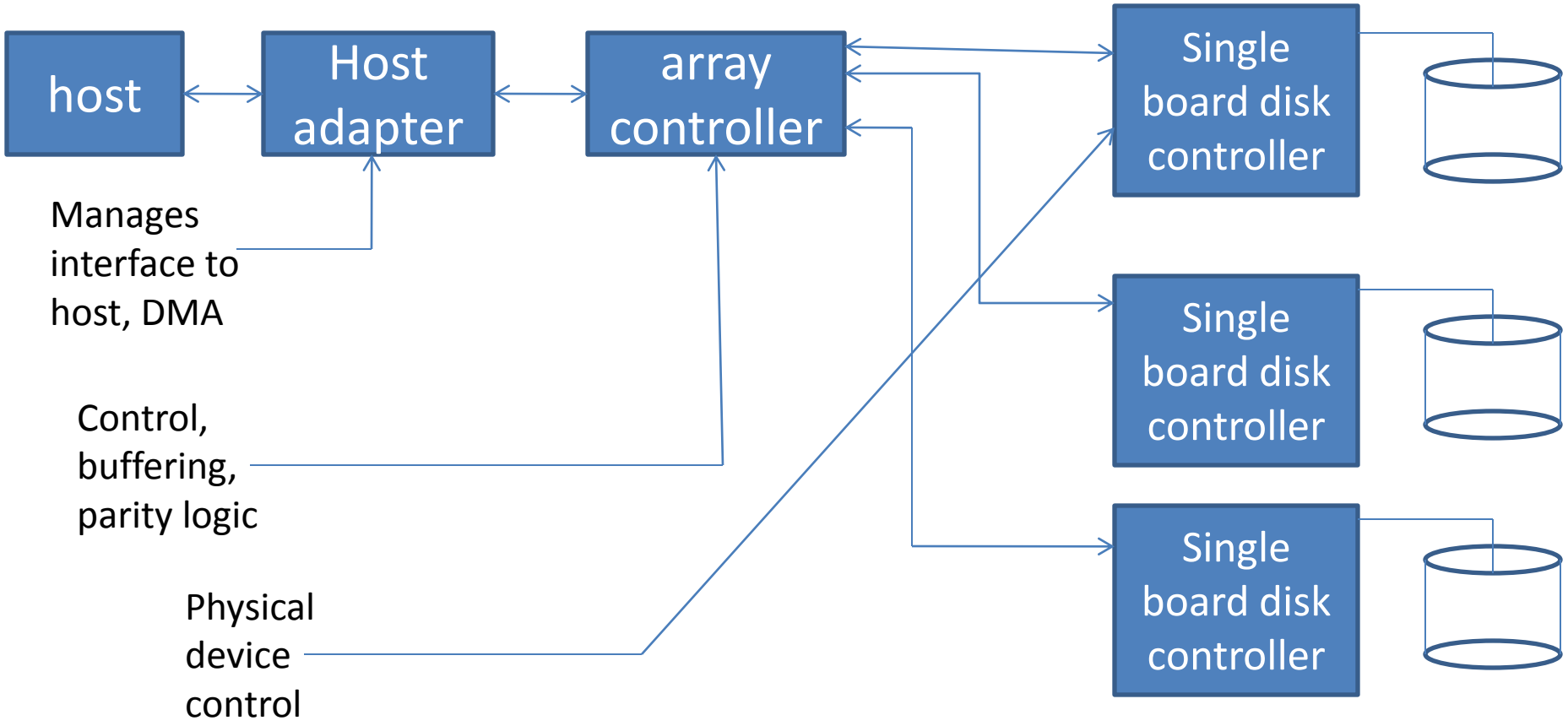
Disk Mirroring, Shadowing

Parity Data Bandwidth Array

High I/O Rate Parity Array

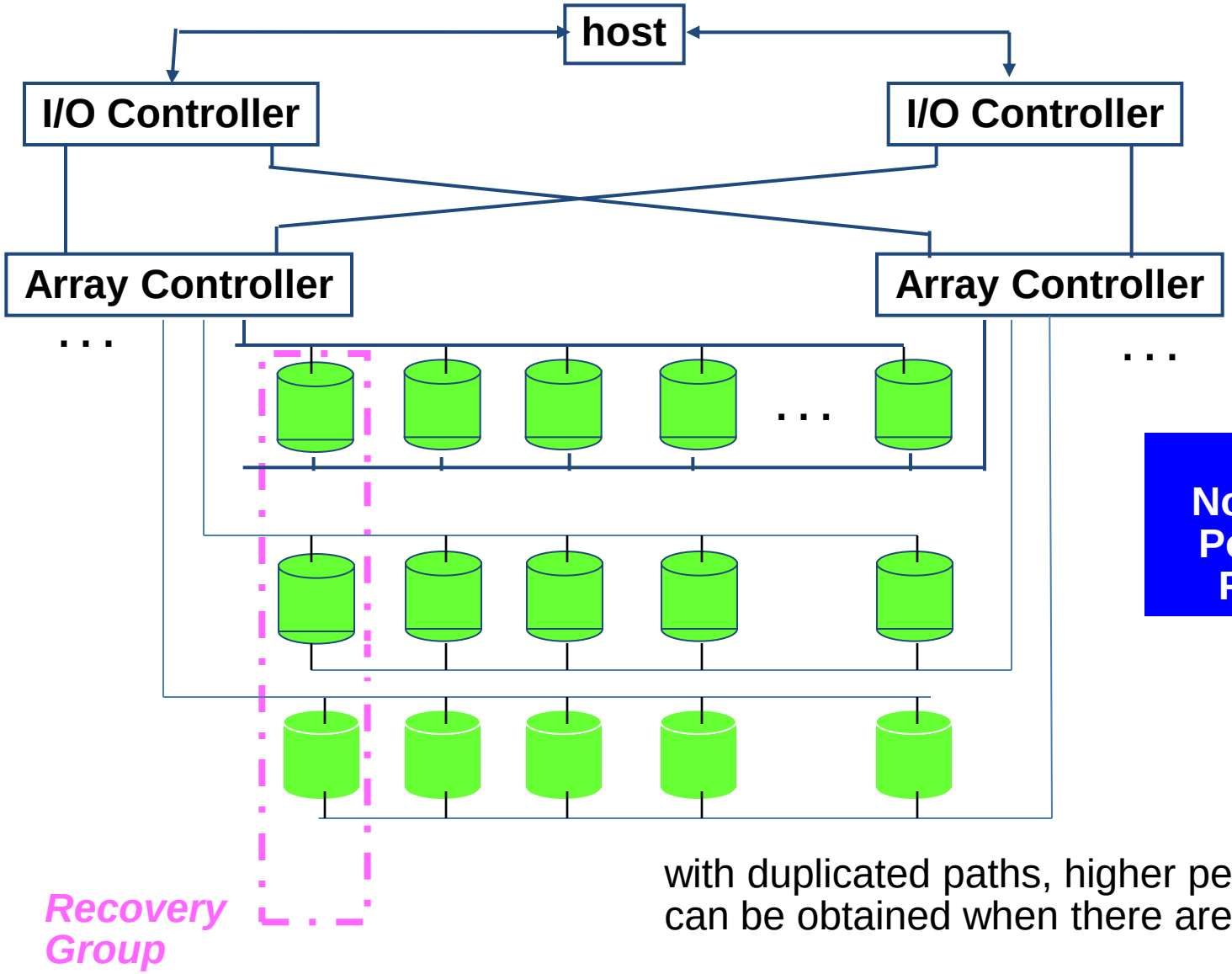


Subsystem Organization



- Striping software off-loaded from host to array controller
- No applications modification
- No reduction to host performance

System-level Availability



Goal: No Single Points of Failure

with duplicated paths, higher performance can be obtained when there are no failures