# Redundant Arrays of Inexpensive Disks (RAID)

**By:**
**Ramasubramanian K.**
rkmurthy@umich.edu

**University of Michigan**
**21st Sept. 2011**

# Rising CPU and Memory Performance

- Great growth in speed of computers
- Fast CPU alone does not make a system fast
- “Each CPU instruction per second requires one byte of main memory”
- Memory technology has to keep pace with advances in other parts.
- Just increase in capacity not enough
- Speed at which instructions delivered to CPU determines ultimate performance

# Rising CPU and Memory Performance

- **Main memory speed kept pace due to:**

  **→ Invention of caches**
  **→ SRAM technology**

- **Performance of Single Large Expensive magnetic Disks (SLED) had modest improvement**

  **→ Seek and rotation delays**
  **→Seek time improvement by 7% per year**

- **Using large main memories to buffer some of the I/O activity an option only with high locality of reference**

# The pending I/O crisis

- **Impact of improving performance of some parts of a problem leaving others unchanged:**

- **Amdahl's law:**

$$S=1/((1-f)+f/k)$$

**S = the effective speedup**
**f= fraction of work in faster mode**
**k = speedup while in faster mode**

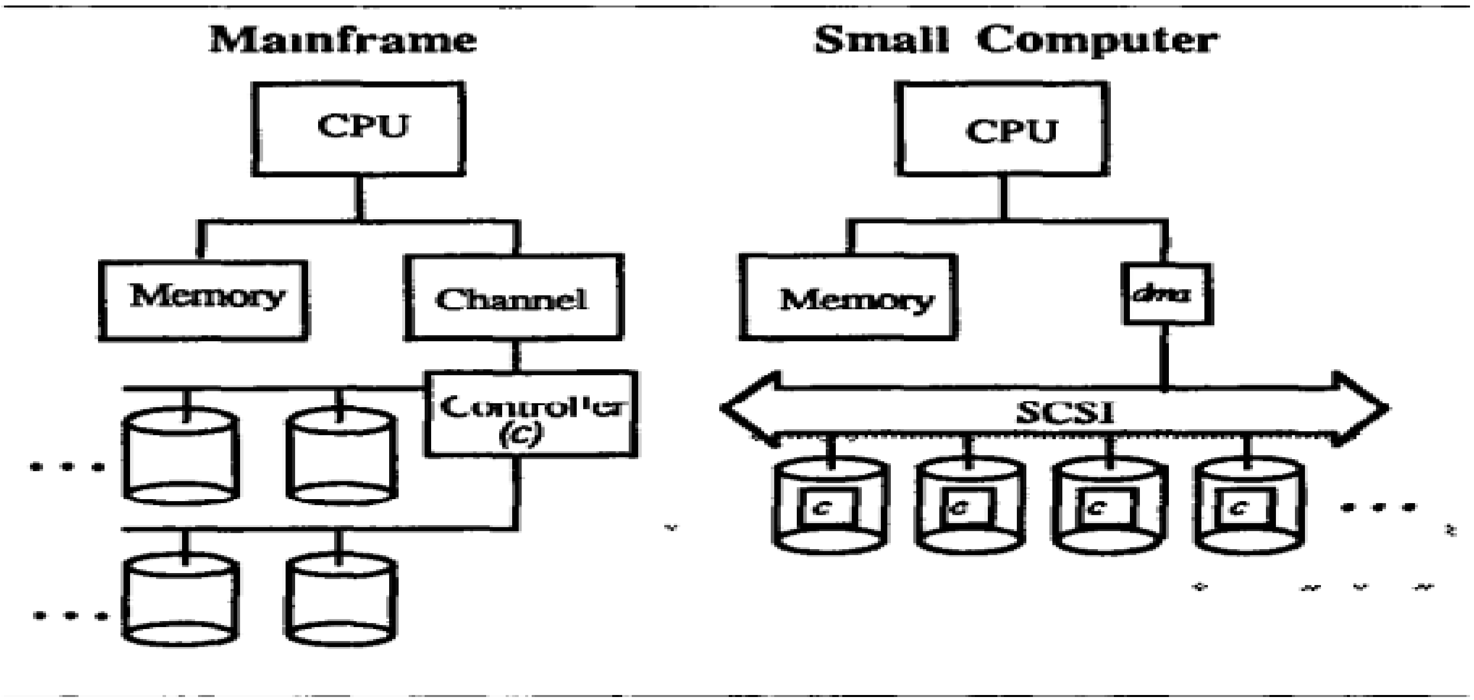
- **Implies that if applications spend 10% time in I/O then when computers are 10 times faster, effective speedup will only be 5%**

- **Innovation needed to avoid I/O crisis**

# Why Arrays of Disks??

- **Personal computers created a market for inexpensive magnetic disks.**
- **Such disks had lower cost as well as capacity**
- **Number of I/Os per second per actuator within a factor of two of large disks**
- **For metrics like cost per MB ,inexpensive disk superior or equal to large disks**
- **Small size and low power**
- **Due to creation of standards such as Small Computer System Interface (SCSI) small disk manufacturers provide such functions**

# Why Arrays of disks??




•Same SCSI interface chip embedded as a controller in every disk can be used as the DMA device at the other end of the SCSI bus.

•Hence, arrays of inexpensive disks!

# The bad news: Reliability

- Forces managers to frequently backup information
- Assuming constant failure rate and independent failures,

$$\textit{MTTF of a Disk Array} = \frac{\textit{MTTF of a Single Disk}}{\textit{Number of Disks in the Array}}$$

- MTTF of 100 CP 3100 disks=300 hours
  Scaling to 1000 disks => MTTF=30 hours!!!
- Large arrays of inexpensive disks too unreliable without fault tolerance.

# The solution: RAID

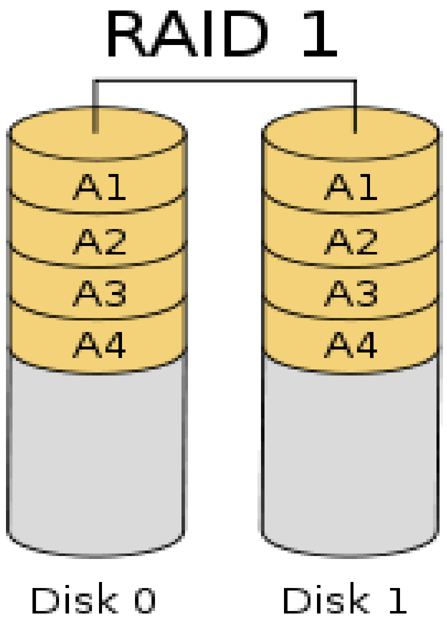
- RAID=***Redundant*** **Array of Independent Disks**
- Use extra disks to store redundant information for recovery in case of disk failure.
- Arrays broken into reliability groups ,each group having extra “check” disks with redundant information.
- Mean Time to Repair (MTTR) reduced by maintaining “hot standby spares” in case a disk fails.
  - Terms used:
    D=Total no. of disks with data
    G=Number of data disks in a group
    C=Number of check disks in a group
    D/G=number of groups

# RAID features

- Reliability Overhead cost decreases from 100% to 4% with RAID level
- Useable storage capacity percentage increases from 50 % to 96%
- Performance metrics:
  - Number of reads
  - Number of writes
  - Read modify writes per second for large as well as small transfers
- Effective Performance per disk

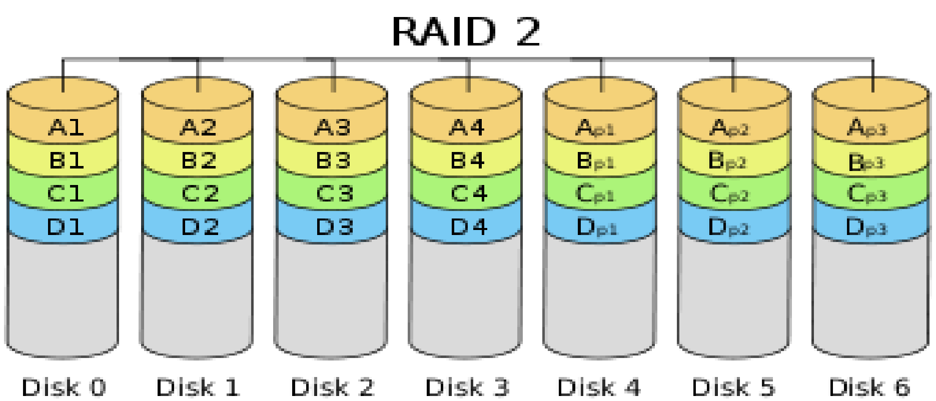
# RAID Level 1: Mirrored Disks

- **Traditional approach for improving reliability of magnetic disks**
- **Most expensive option***
- **Every write to data disk also write to check disk**
- **Doubles the cost of database system**
- **Uses only 50% of disk storage capacity**
- **Largess inspires need for next RAID levels.**



# RAID Level 2: Hamming code for ECC

- Introduction of 4K and 16K DRAM's bought about need for level 2
- Redundant chips added to correct single errors and detect double errors in each group
- Increased no. of memory chips
- Improved reliability




- If data bits in a group are read or written together ,no impact on performance.

# Level 2 :Advantages

- Same performance as level 1 for large writes, but uses fewer check disks
- Since all disks of group accessed for data transfer, higher data rate with increasing group size, desirable for supercomputers
- Single parity disk can detect a single error

# Level 2:Disadvantages

- To correct an error, enough disks needed to identify the disk with error
- Reads of less than group size ➔read whole group
- Writes to portion of disk in 3 steps:

  → Read to get rest of the data
  → Modify to merge new and old information
  → Write to write full group inc. check information

- Reads to smaller amount mean reading a full sector from each of the bit interleaved disks in a group
- Writes of a single unit mean read-modify-write cycle to all disks
- Performance dismal for small transfers for whole system or per disk
- Not suitable for TPS

# RAID Level I

vs.

| | |
|---|---|
| *MTTF* | Exceeds Useful Product Lifetime (4,500,000 hrs or > 500 years) |
| *Total Number of Disks* | 2D |
| *Overhead Cost* | 100% |
| *Useable Storage Capacity* | 50% |

| *Events/Sec vs Single Disk* | *Full RAID* | *Efficiency Per Disk* |
|---|---|---|
| *Large (or Grouped) Reads* | 2D/S | 1 00/S |
| *Large (or Grouped) Writes* | D/S | 50/S |
| *Large (or Grouped) R-M-W* | 4D/3S | 67/S |
| *Small (or Individual) Reads* | 2D | 1 00 |
| *Small (or Individual) Writes* | D | 50 |
| *Small (or Individual) R-M-W* | 4D/3 | 67 |

# RAID Level II

| *MTTF* | Exceeds Useful Lifetime | |
|---|---|---|
| | *G=10* | *G=25* |
| | (494,500 hrs or >50 years) | (103,500 hrs or 12 years) |
| *Total Number of Disks* | 1 40D | 1.20D |
| *Overhead Cost* | 40% | 20% |
| *Useable Storage Capacity* | 71% | 83% |

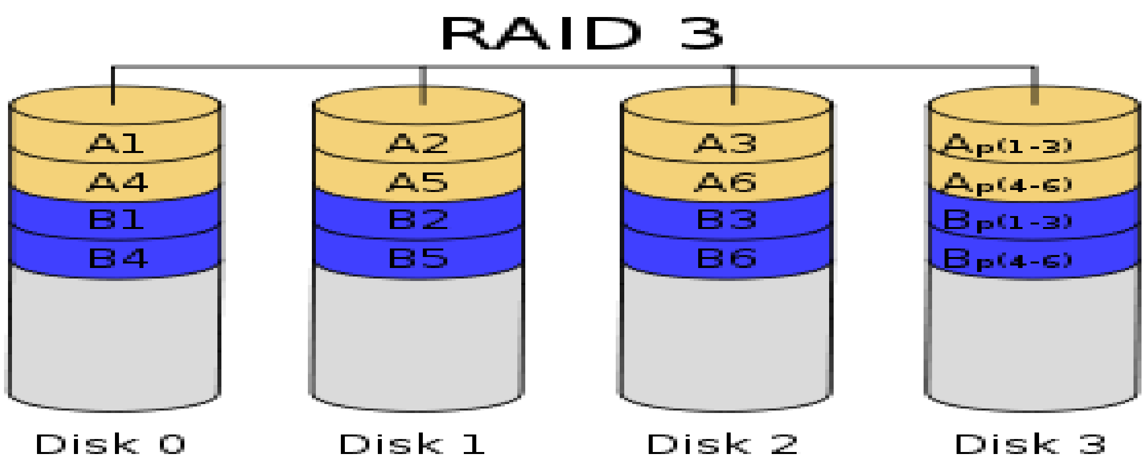
| *Events/Sec (vs Single Disk)* | *Full RAID* | *Efficiency Per Disk L2* | *L2/L1* | *Efficiency Per Disk L2* | *L2/L1* |
|---|---|---|---|---|---|
| *Large Reads* | D/S | 71/S | 71% | 86/S | 86% |
| *Large Writes* | D/S | 71/S | 143% | 86/S | 172% |
| *Large R-M-W* | D/S | 71/S | 107% | 86/S | 129% |
| *Small Reads* | D/SG | 07/S | 6% | 03/S | 3% |
| *Small Writes* | D/2SG | 04/S | 6% | 02/S | 3% |
| *Small R-M-W* | D/SG | 07/S | 9% | 03/S | 4% |

# Need for RAID Level 3

- Most check disks in level 2 RAID used to determine which disk failed
- Only 1 redundant parity disk needed to detect an error
- Extra disks redundant since failure can be detected from special signals provided in the disk interface
- Extra checking information at the end of sector can also be used to detect and correct soft errors

# RAID Level 3: Single Check Disk per Group

- Reduces check disks to one per group(C=1)
- Overhead cost decreases by 4 to 10%
- Effective performance per disk better than level 2 due to fewer check disks
- Reduction in disks ➔ Improved reliability
- Has bought reliability overhead cost to its lowest level



# Level 2 vs. Level 3

**Level 2**

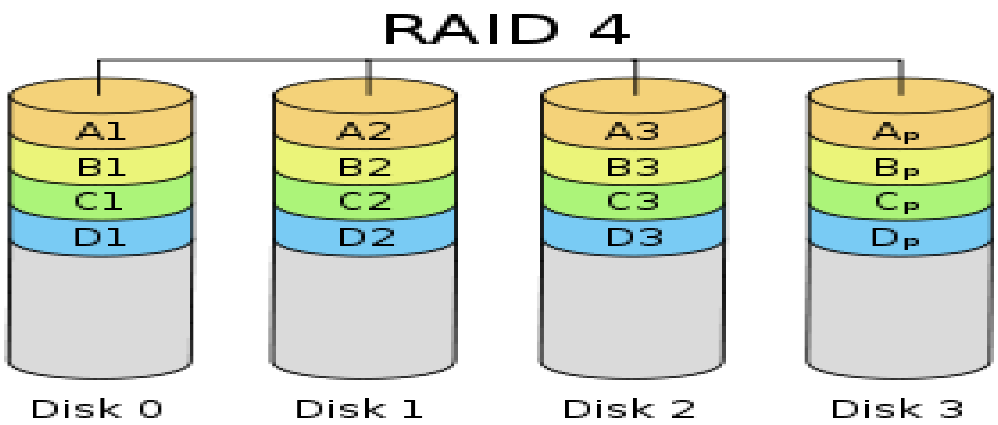
| MTTF | | Exceeds Useful Lifetime | | | |
|---|---|---|---|---|---|
| | | G=10 | | G=25 | |
| | | (494,500 hrs or >50 years) | | (103,500 hrs or 12 years) | |
| Total Number of Disks | | 1 40D | | 1.20D | |
| Overhead Cost | | 40% | | 20% | |
| Useable Storage Capacity | | 71% | | 83% | |
| Events/Sec (vs Single Disk) | Full RAID | Efficiency Per Disk | | Efficiency Per Disk | |
| | | L2 | L2/L1 | L2 | L2/L1 |
| Large Reads | D/S | 71/S | 71% | 86/S | 86% |
| Large Writes | D/S | 71/S | 143% | 86/S | 172% |
| Large R-M-W | D/S | 71/S | 107% | 86/S | 129% |
| Small Reads | D/SG | 07/S | 6% | 03/S | 3% |
| Small Writes | D/2SG | 04/S | 6% | 02/S | 3% |
| Small R-M-W | D/SG | 07/S | 9% | 03/S | 4% |

**Level 3**

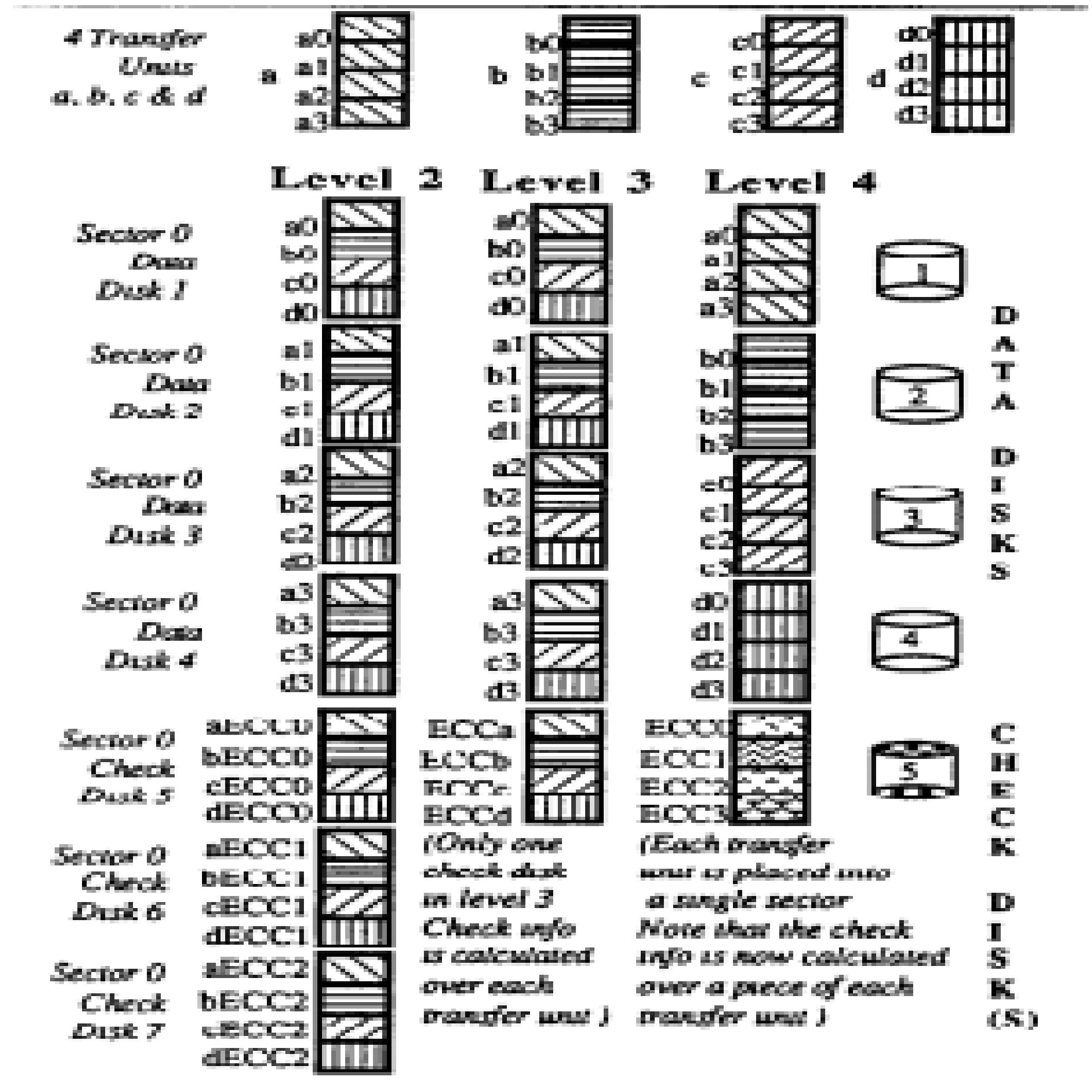
| MTTF | | Exceeds Useful Lifetime | | | | | |
|---|---|---|---|---|---|---|---|
| | | G=10 | | | G=25 | | |
| | | (820,000 hrs or >90 years) | | | (346,000 hrs or 40 years) | | |
| Total Number of Disks | | 1 10D | | | 1 04D | | |
| Overhead Cost | | 10% | | | 4% | | |
| Useable Storage Capacity | | 91% | | | 96% | | |
| Events/Sec (vs Single Disk) | Full RAID | Efficiency Per Disk | | | Efficiency Per Disk | | |
| | | L3 | L3/L2 | L3/L1 | L3 | L3/L2 | L3/L1 |
| Large Reads | D/S | 91/S | 127% | 91% | 96/S | 112% | 96% |
| Large Writes | D/S | 91/S | 127% | 182% | 96/S | 112% | 192% |
| Large R-M-W | D/S | 91/S | 127% | 136% | 96/S | 112% | 142% |
| Small Reads | D/SG | 09/S | 127% | 8% | 04/S | 112% | 3% |
| Small Writes | D/2SG | 05/S | 127% | 8% | 02/S | 112% | 3% |
| Small R-M-W | D/SG | 09/S | 127% | 11% | 04/S | 112% | 5% |

# RAID Level 4:Independent Reads/Writes

- Improves performance of small transfers through parallelism
- Each individual transfer unit of data kept in a single disk
- Data between disks is interleaved at the sector level rather than bit level
- Parity calculation simpler than level 3:
  new parity=(old data xor new data) xor old parity
- Small read involves only one read on one disk



# Comparing location of data and check information in sectors of levels 2,3 and 4



# Level 3 vs. Level 4

| *MTTF* | | Exceeds Useful Lifetime | |
|---|---|---|---|
| | | G=10 (820,000 hrs or >90 years) | G=25 (346,000 hrs or 40 years) |
| *Total Number of Disks* | | 1 10D | 1 04D |
| *Overhead Cost* | | 10% | 4% |
| *Useable Storage Capacity* | | 91% | 96% |

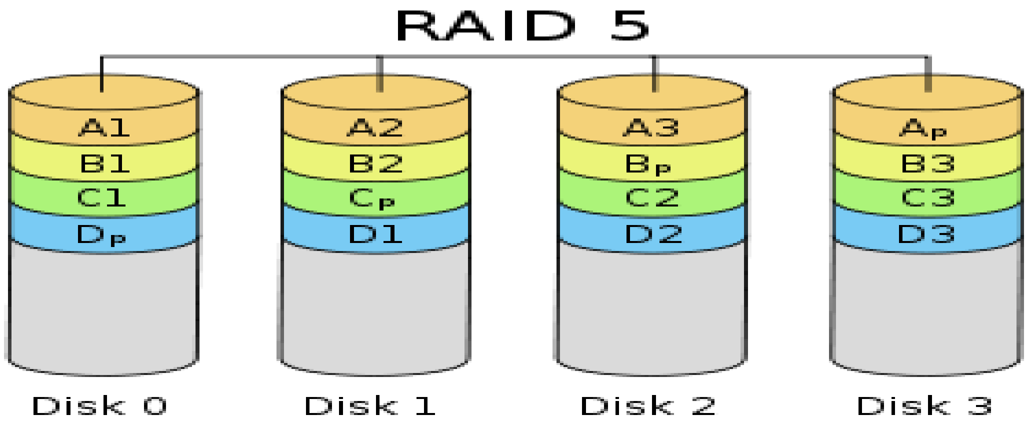
| *Events/Sec (vs Single Disk)* | *Full RAID* | *Efficiency Per Disk* L3 | L3/L2 | L3/L1 | *Efficiency Per Disk* L3 | L3/L2 | L3/L1 |
|---|---|---|---|---|---|---|---|
| *Large Reads* | D/S | 91/S | 127% | 91% | 96/S | 112% | 96% |
| *Large Writes* | D/S | 91/S | 127% | 182% | 96/S | 112% | 192% |
| *Large R-M-W* | D/S | 91/S | 127% | 136% | 96/S | 112% | 142% |
| *Small Reads* | D/SG | 09/S | 127% | 8% | 04/S | 112% | 3% |
| *Small Writes* | D/2SG | 05/S | 127% | 8% | 02/S | 112% | 3% |
| *Small R-M-W* | D/SG | 09/S | 127% | 11% | 04/S | 112% | 5% |

| *MTTF* | | Exceeds Useful Lifetime | |
|---|---|---|---|
| | | G=10 (820,000 hrs or >90 years) | G=25 (346,000 hrs or 40 years) |
| *Total Number of Disks* | | 1 10D | 1 04D |
| *Overhead Cost* | | 10% | 4% |
| *Useable Storage Capacity* | | 91% | 96% |

| *Events/Sec (vs Single Disk)* | *Full RAID* | *Efficiency Per Disk* L4 | L4/L3 | L4/L1 | *Efficiency Per Disk* L4 | L4/L3 | L4/L1 |
|---|---|---|---|---|---|---|---|
| *Large Reads* | D/S | 91/S | 100% | 91% | 96/S | 100% | 96% |
| *Large Writes* | D/S | 91/S | 100% | 182% | 96/S | 100% | 192% |
| *Large R-M-W* | D/S | 91/S | 100% | 136% | 96/S | 100% | 146% |
| *Small Reads* | D | 91 | 1200% | 91% | 96 | 3000% | 96% |
| *Small Writes* | D/2G | 05 | 120% | 9% | 02 | 120% | 4% |
| *Small R-M-W* | D/G | 09 | 120% | 14% | 04 | 120% | 6% |

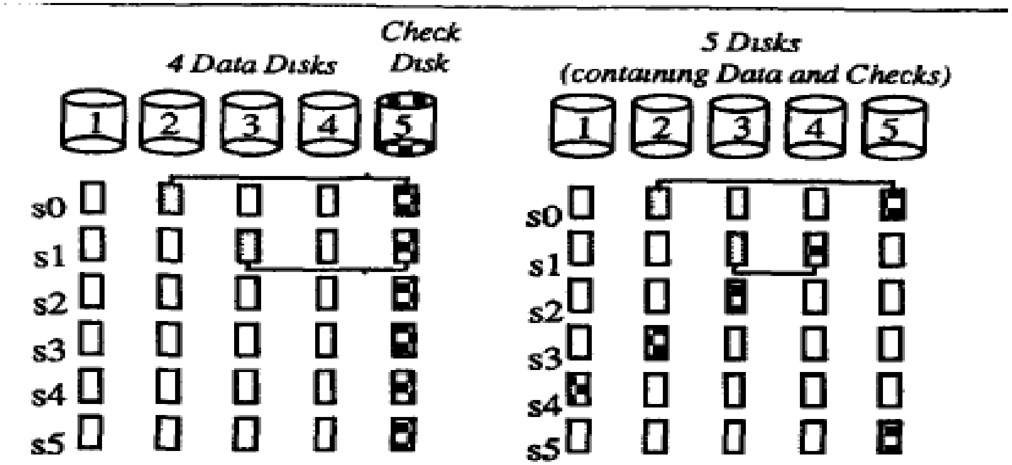
# RAID Level 5: No Single Check Disk

- Level 4 small write uses 2 disks to perform 4 accesses-2 reads,2 writes
- Writes still limited to one per group since every write must read and write the check disk
- Level 5 distributes data and check information across all disks-inc. check disks
- Can support multiple individual writes per group



# RAID Level 5: features

- Small read-modify-writes perform close to the speed per disk of a level 1 RAID
- Has large transfer performance per disk and high useful storage capacity percentage like levels 3 and 4
- Improves performance of small reads since one more disk per group contains data.



# Level 4 vs. Level 5

| MTTF | Exceeds Useful Lifetime | |
|---|---|---|
| | G=10 | G=25 |
| | (820,000 hrs or >90 years) | (346,000 hrs or 40 years) |
| Total Number of Disks | 1 10D | 1 04D |
| Overhead Cost | 10% | 4% |
| Useable Storage Capacity | 91% | 96% |

| Events/Sec (vs Single Disk) | Full RAID | Efficiency Per Disk | | | Efficiency Per Disk | | |
|---|---|---|---|---|---|---|---|
| | | L4 | L4/L3 | L4/L1 | L4 | L4/L3 | L4/L1 |
| Large Reads | D/S | 91/S | 100% | 91% | 96/S | 100% | 96% |
| Large Writes | D/S | 91/S | 100% | 182% | 96/S | 100% | 192% |
| Large R-M-W | D/S | 91/S | 100% | 136% | 96/S | 100% | 146% |
| Small Reads | D | 91 | 1200% | 91% | 96 | 3000% | 96% |
| Small Writes | D/2G | 05 | 120% | 9% | 02 | 120% | 4% |
| Small R-M-W | D/G | 09 | 120% | 14% | 04 | 120% | 6% |

| MTTF | Exceeds Useful Lifetime | |
|---|---|---|
| | G=10 | G=25 |
| | (820,000 hrs or >90 years) | (346,000 hrs or 40 years) |
| Total Number of Disks | 1 10D | 1 04D |
| Overhead Cost | 10% | 4% |
| Useable Storage Capacity | 91% | 96% |

| Events/Sec (vs Single Disk) | Full RAID | Efficiency Per Disk | | | Efficiency Per Disk | | |
|---|---|---|---|---|---|---|---|
| | | L5 | L5/L4 | L5/L1 | L5 | L5/L4 | L5/L1 |
| Large Reads | D/S | 91/S | 100% | 91% | 96/S | 100% | 96% |
| Large Writes | D/S | 91/S | 100% | 182% | 96/S | 100% | 192% |
| Large R-M-W | D/S | 91/S | 100% | 136% | 96/S | 100% | 144% |
| Small Reads | (1+C/G)D | 1 00 | 110% | 100% | 1 00 | 104% | 100% |
| Small Writes | (1+C/G)D/4 | 25 | 550% | 50% | 25 | 1300% | 50% |
| Small R-M-W | (1+C/G)D/2 | 50 | 550% | 75% | 50 | 1300% | 75% |

# Observations

- Decision between hardware and software solutions for disk striping and parity support is strictly one of cost and benefit
- Performance of RAID improves as size of smallest transfer unit increases
- Performance improves significantly with full track buffer in every disk

# Things to remember

- Level 5 can be used for supercomputing and transaction processing applications
- RAID offers significant advantage over SLED for the same cost*
- RAID level 5 offers factor of 10 improvement in performance, reliability and power consumption while reducing size
- RAID offers advantage of modular growth
- Due to low power consumption, battery backup for whole disk array can be considered

# Conclusion

- RAID :Cost effective option to meet challenge of exponential growth in processor and memory speeds
- Smaller size simplifies interconnection of many components, packaging and labeling
- RAIDs expected to replace SLEDs completely in the future I/O systems

# References


- "A Case for Redundant Arrays of Inexpensive Disks" by David A Patterson, Garth Gibson, and Randy H Katz
- "RAID: A personal recollection of how storage became a system" by Randy H. Katz