Using a Shared Storage Class Memory Device to Improve the Reliability of RAID Arrays

S. Chaarawi. U. of Houston
J.-F. Pâris, U. of Houston
A. Amer, Santa Clara U.
T. J. E. Schwarz, U. Católica del Uruguay
D. D. E. Long, U. C. Santa Cruz

CS@UH

The problem

- **Archival storage systems** store
  - Huge amounts of data
  - Over long periods of time
- Must ensure **long-term survival** of these data
  - Disk failure rates
    - Typically exceed 1% per year
    - Can exceed 9–10% per year

CS@UH

Requirements

- Archival storage systems should
  - Be **more reliable** than conventional storage architectures
    - Excludes RAID level 5
  - Be **cost-effective**
    - Excludes mirroring
  - Have **lower power requirements** than conventional storage architectures
    - Not addressed here

CS@UH

Non-Requirements

- Contrary to conventional storage systems
  - Update costs are much less important
  - Access times are less critical

Traditional Solutions

- **Mirroring:**
  - Maintains two copies of all data
  - Safe but costly
- **RAID level 5 arrays:**
  - Use omission correction codes: **parity**
  - Can tolerate one disk failure
  - Cheaper but less safe than mirroring

More Recent Solutions (I)

- <u>**RAID level 6 arrays:**</u>
  - Can tolerate two disk failures
    - Or a single disk failure and bad blocks on several disks
  - Slightly higher storage costs than RAID level 5 arrays
  - More complex update procedures
  - X-Code, EvenOdd, Row-Diagonal Parity

More Recent Solutions (II)

- **Superparity:**
  - Widani et al., MASCOTS 2009
  - Partitions each disk into fixed-size “disklets” used to form conventional RAID stripes
  - Groups these stripes into “supergroups”
  - Adds to each supergroup one or more distinct “superparity” devices

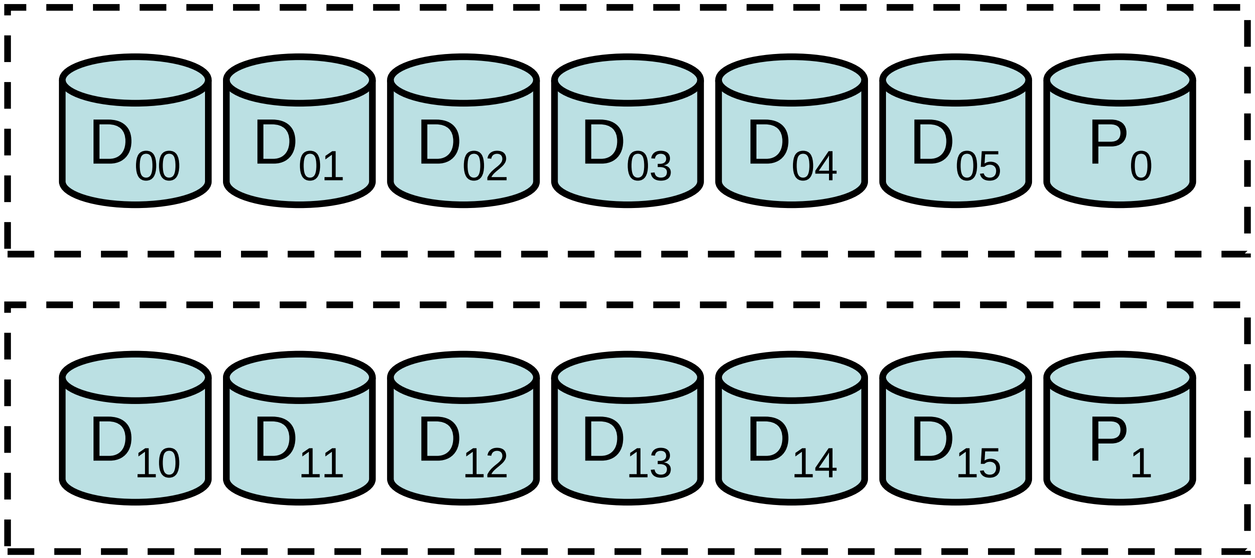
CS@UH

More Recent Solutions (III)

- **Shared Parity Disks**
  - Paris and Amer, IPCCC 2009
  - Does not use disklets
  - Starts with a few RAID level 5 arrays
  - Adds an extra parity disk to these arrays

Example (I)

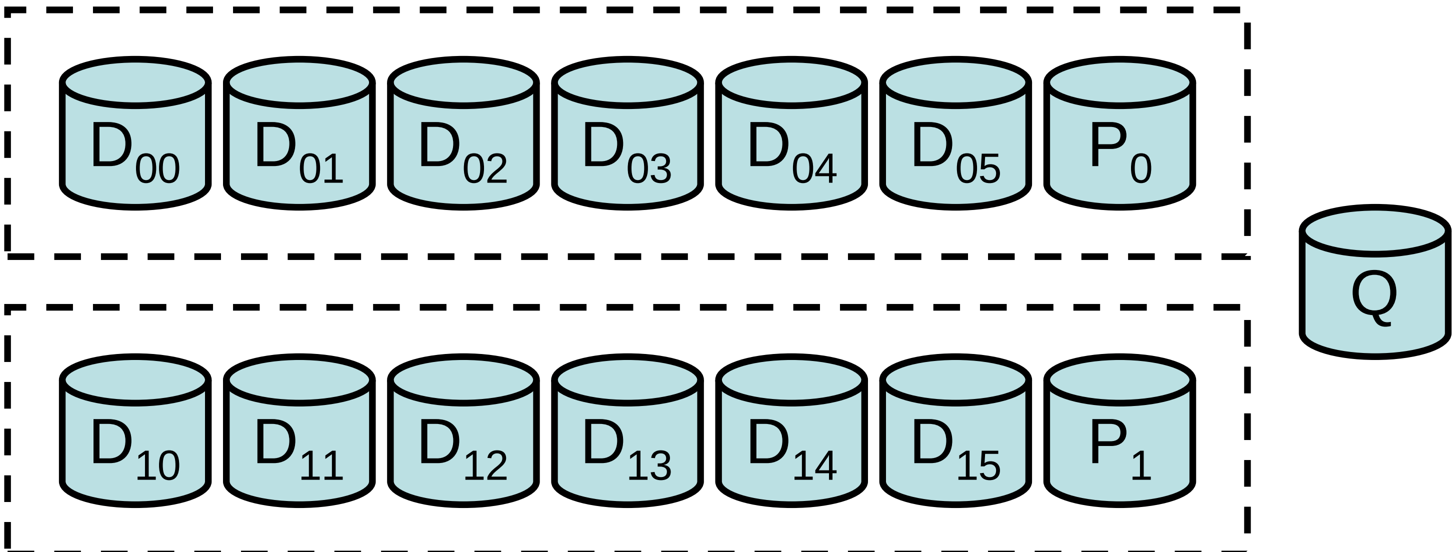
- Start with two RAID arrays:
  - In reality, parity blocks will be distributed among all disks



CS@UH

Example (II)

- Add an extra parity disk



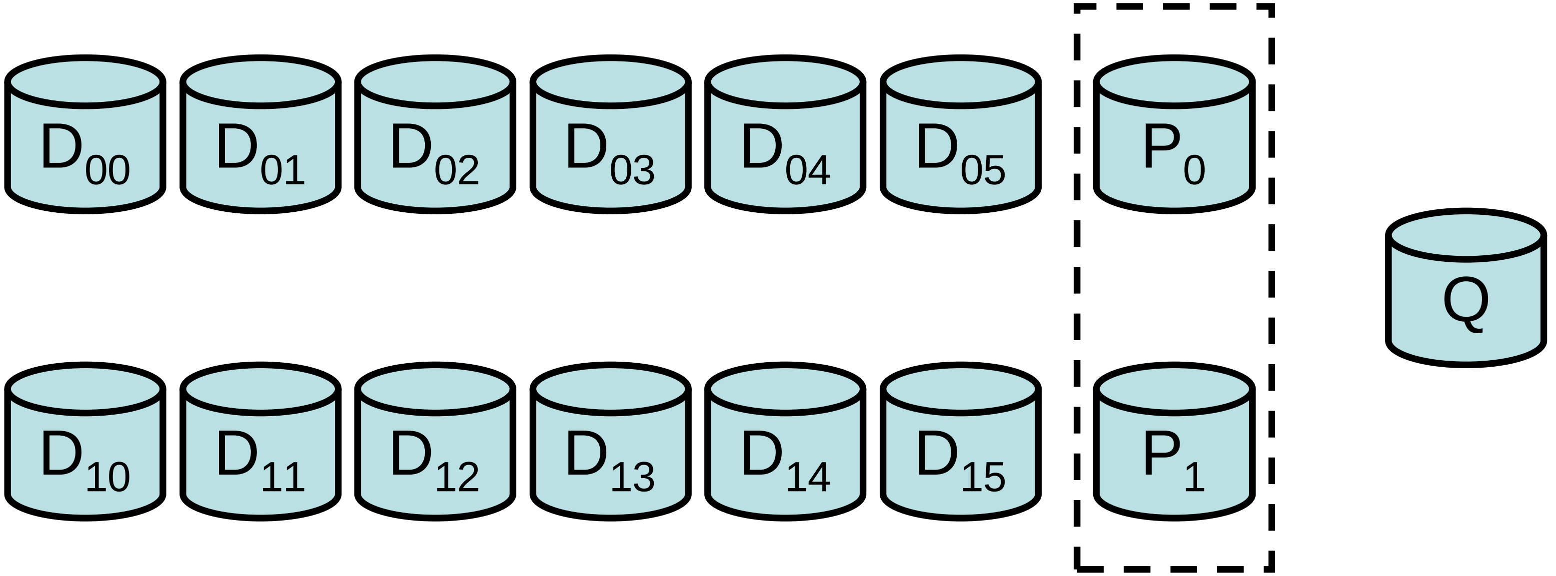
CS@UH

Example (III)

- Single disk failures handled within each individual RAID array
- Double disk failures handled by whole structure

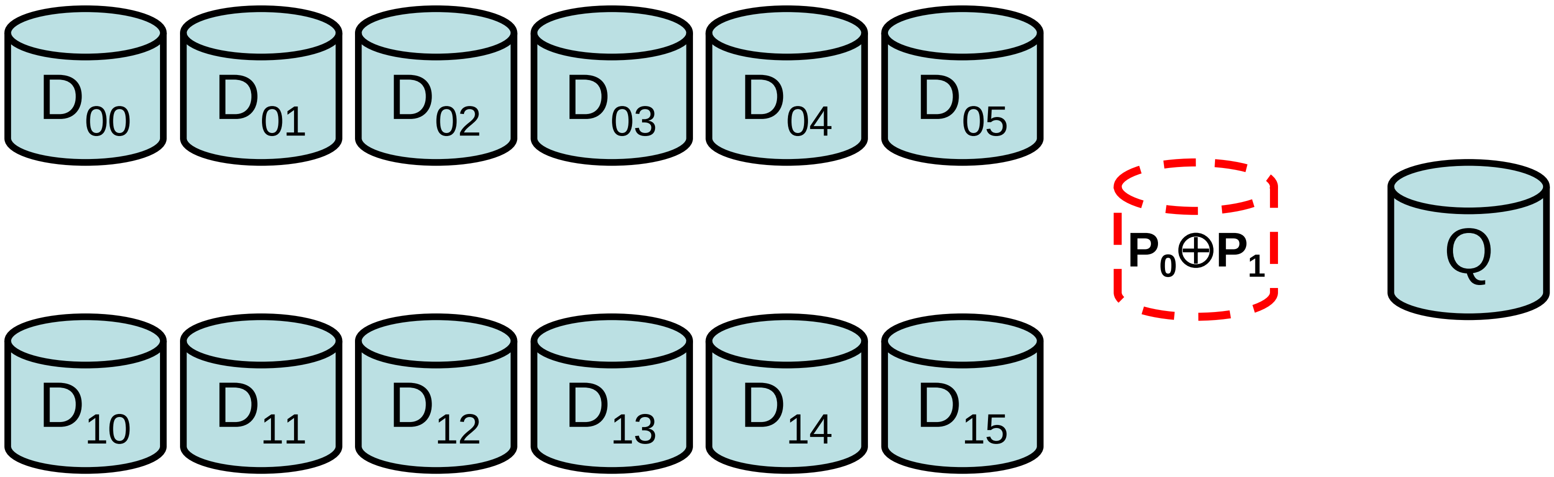
Example (IV)

- We XOR the two parity disks to form a single virtual drive



Example (V)

- And obtain a single RAID level 6 array

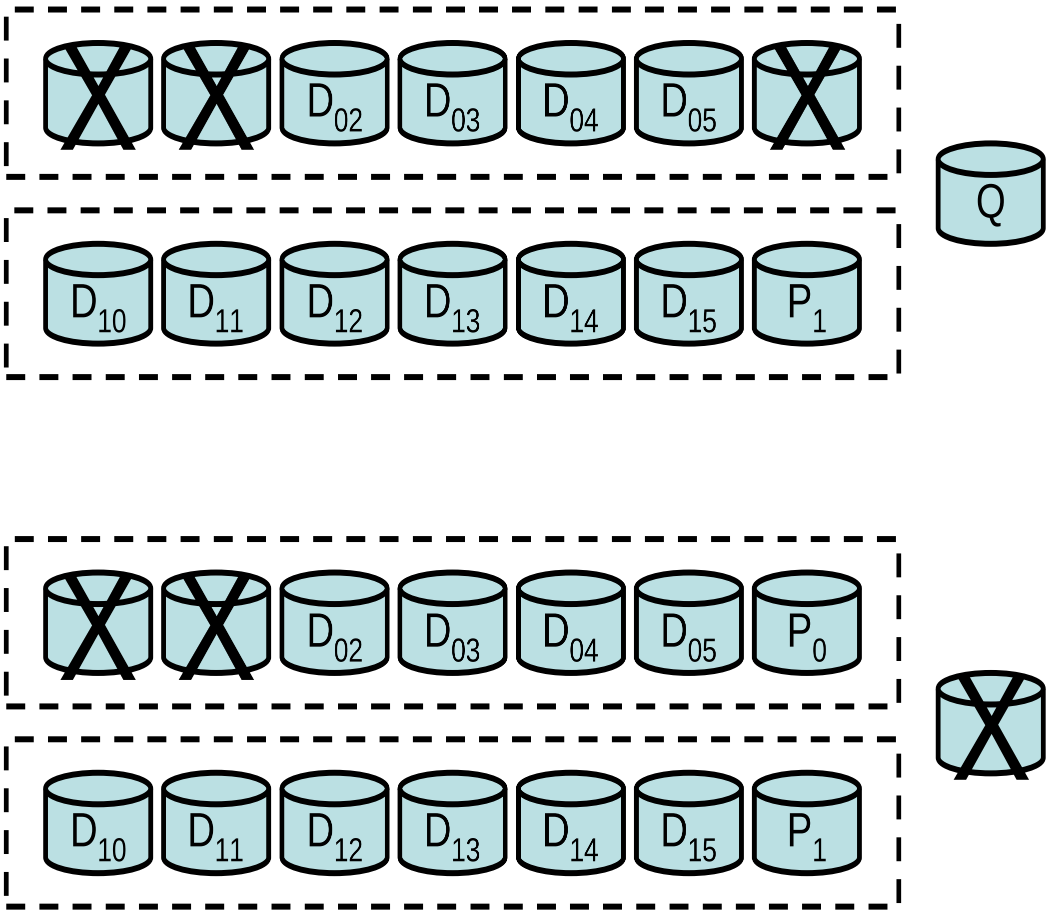


Example (VI)

- Our array tolerates **all double failures**
- Also tolerates **most triple failures**
  - Triple failures causing a data loss include failures of:
    - Three disks in same RAID array
    - Two disks in same RAID array plus shared parity disk Q

CS@UH

Triple Failures Causing a Data Loss



Our Idea

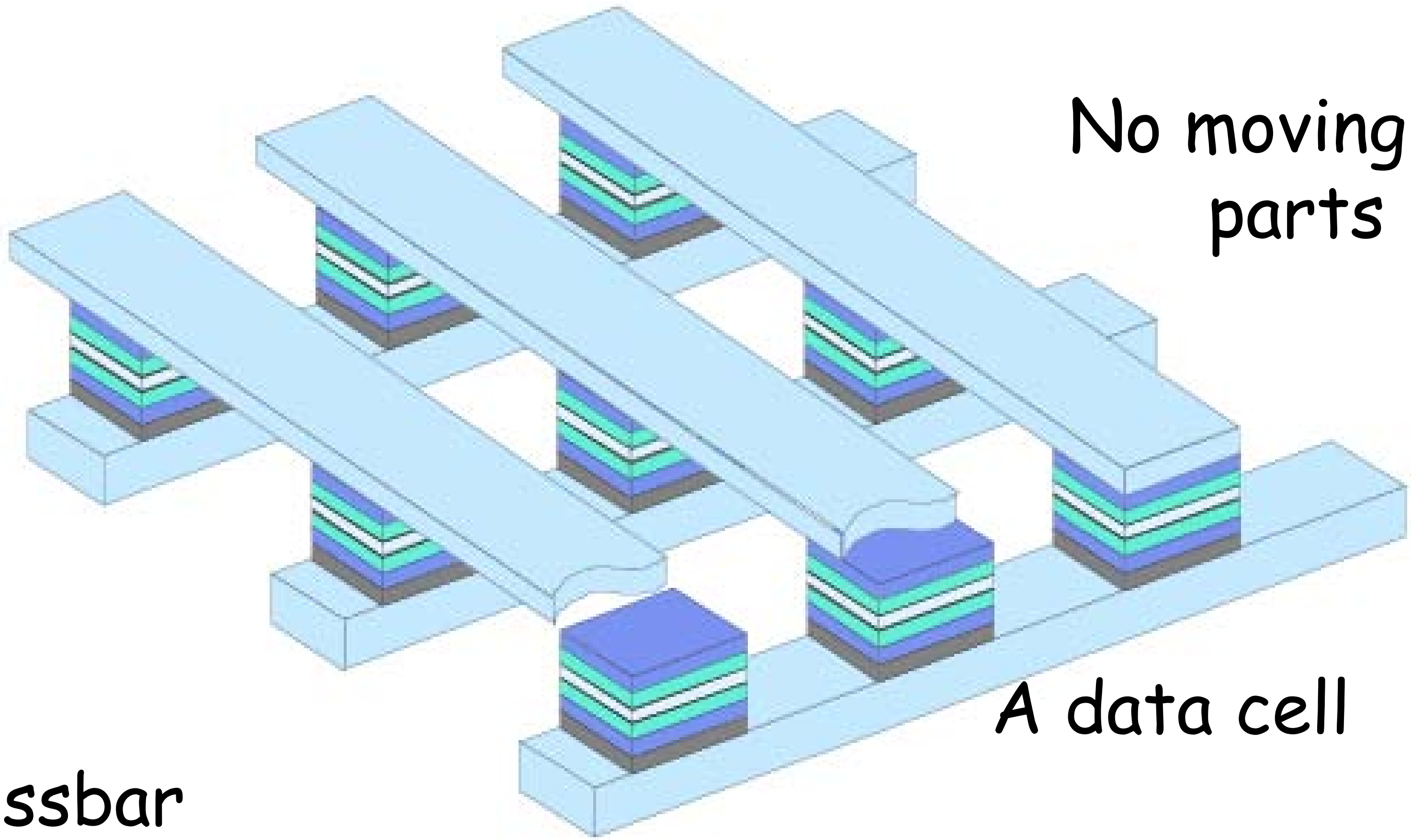
- Replace the shared parity disk by a **much more reliable** device
  - A **Storage Class Memory (SCM)** device
- Will reduce the risk of data loss

CS@UH

Storage Class Memories

- Solid-state storage
  - Non-volatile
  - Much faster than conventional disks
- Numerous proposals:
  - *Ferro-electric RAM* (FRAM)
  - *Magneto-resistive RAM* (MRAM)
  - *Phase-change memories* (PCM)
- We focus on PCMs as exemplar of these technologies

Phase-Change Memories



No moving parts

A data cell

Crossbar organization

CS@UH

Phase-Change Memories

- Cells contain a ***chalcogenide*** material that has ***two states***
  - ***Amorphous*** with high electrical resistivity
  - ***Crystalline*** with low electrical resistivity
- ***Quickly cooling*** material from above fusion point leaves it in ***amorphous state***
- ***Slowly cooling*** material leaves it in ***crystalline state***

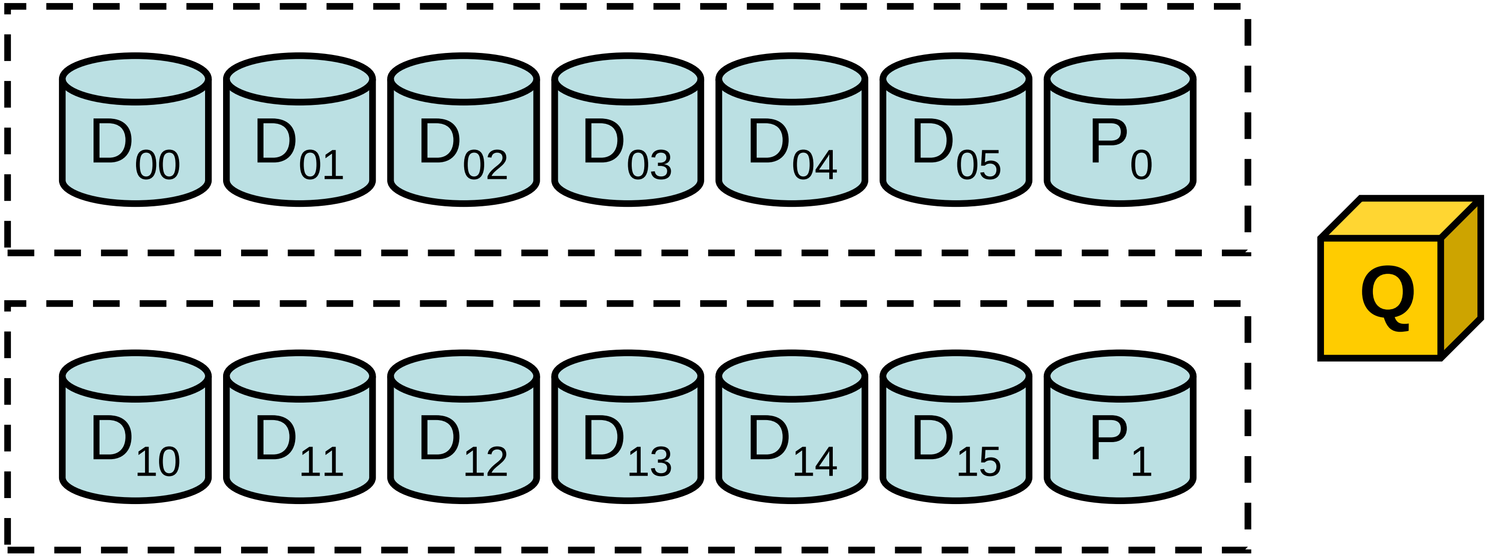
Key Parameters of Future PCMs

Target date	2012
Access time	100 ns
Data Rate	200–1000 MB/s
Write Endurance	10^9 write cycles
Read Endurance	no upper limit
Capacity	16 GB
Capacity growth	> 40% per year
MTTF	**10–50 million hours**
Cost	**< $2/GB**

CS@UH

New Array Organization

- Use SCM device as shared parity device



Reliability Analysis

- **Reliability $R(t)$:**
  - Probability that system will operate correctly over the time interval [0, t] given that it operated correctly at time $t = 0$
  - Hard to estimate
- **Mean Time To Data Loss (MTTDL):**
  - Single value
  - Much easier to compute

Our Model

- Device failures are mutually independent and follow a Poisson law
  - *A reasonable approximation*
- Device repairs can be performed in parallel
- Device repair times follow an exponential law
  - *Not true but required to make the model tractable*

Scope of Investigation

- We computed the MTTDL of
  - A pair of RAID 5 arrays with 7 disks each **plus** a shared parity SCM
  - A pair of RAID 5 arrays with 7 disks each **plus** a shared parity disk

and compare it with the MTTDLs of

  - A pair of RAID 5 arrays with 7 disks each
  - A pair of RAID 6 arrays with 8 disks each

CS@UH

System Parameters (I)

- Disk mean time to fail was assumed to be 100,000 hours (11 years and 5 months)
  - Corresponds to a failure rate λ of 8 to 9% per year
  - High end of failure rates observed by Schroeder + Gibson and Pinheiro et al.
- SCM device MTTF was assumed to be a multiple of disk MTTF

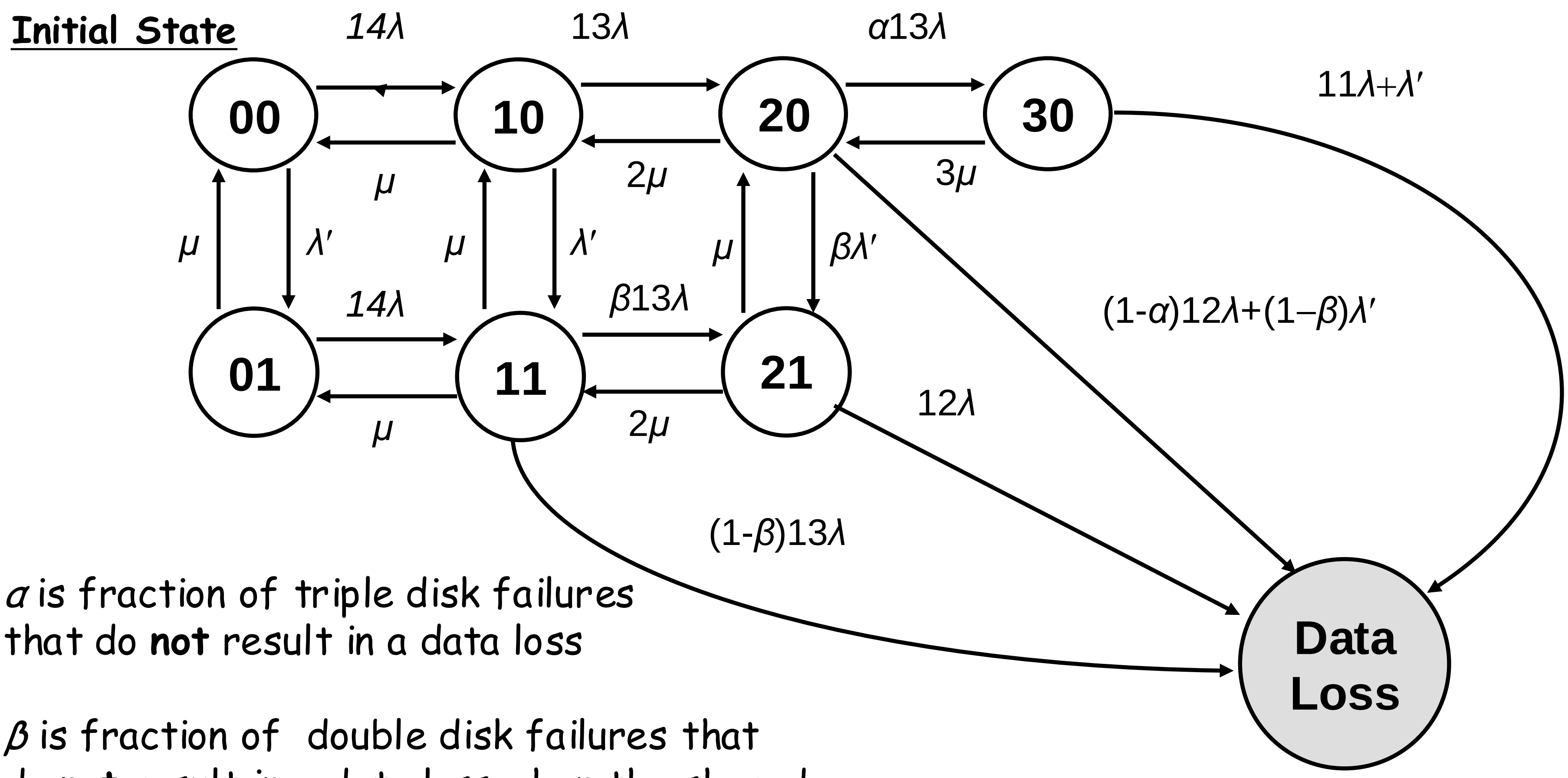
CS@UH

System Parameters

- Disk and SCM device repair times varied between 12 hours and one week
  - Corresponds to repair rates μ varying between 2 and 0.141 repairs/day

CS@UH

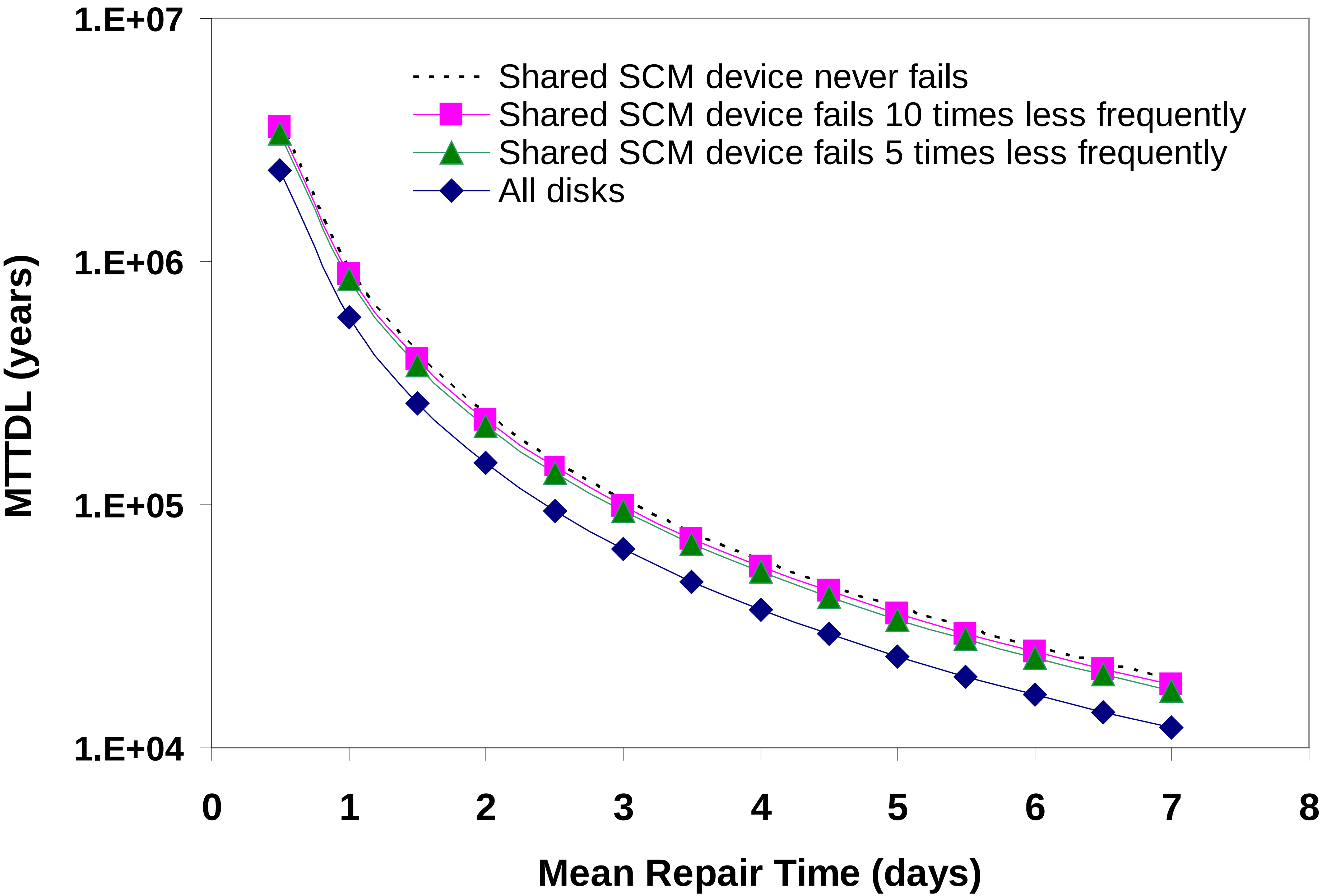
State Diagram



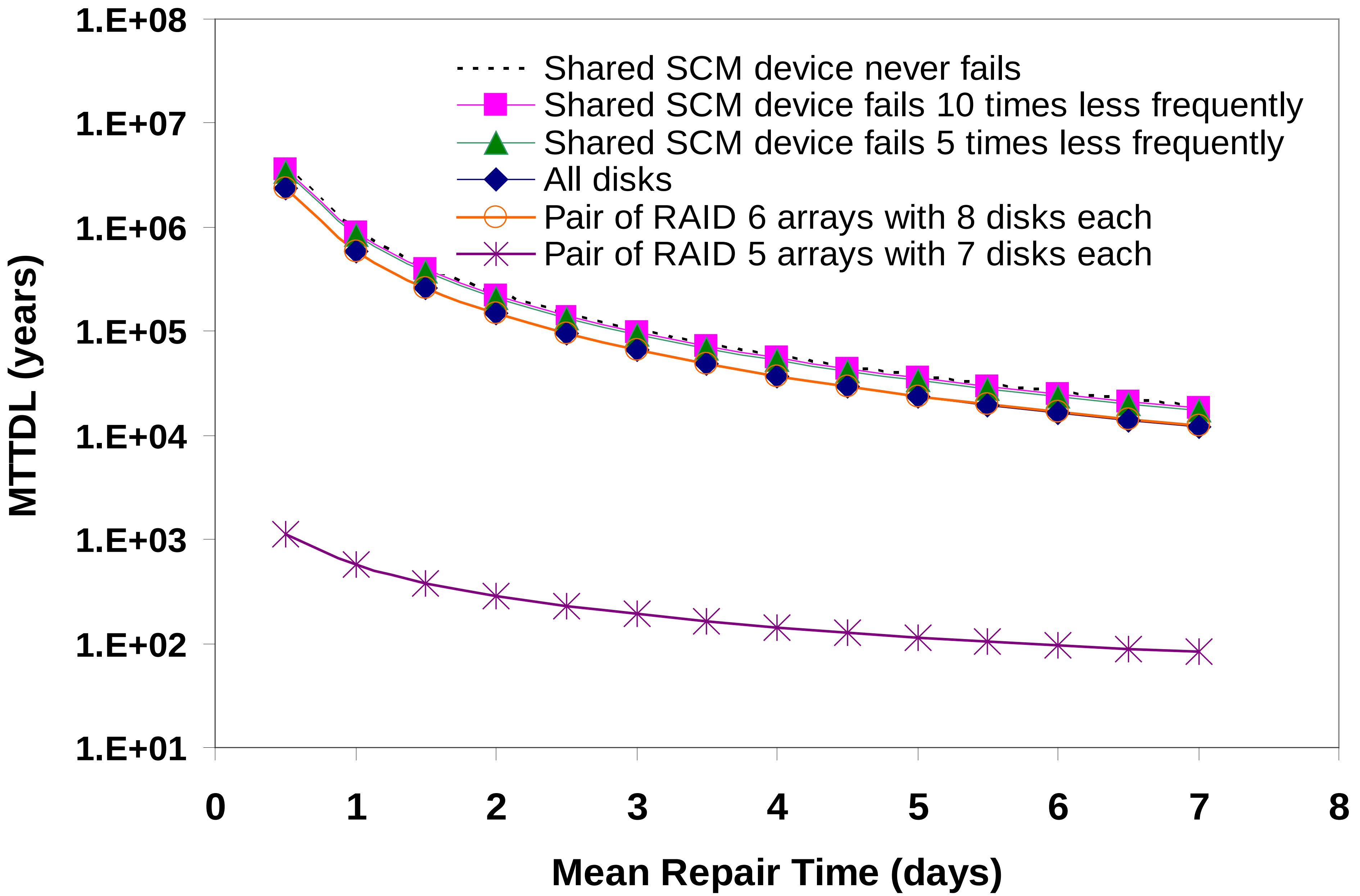
α is fraction of triple disk failures that do **not** result in a data loss

β is fraction of double disk failures that do **not** result in a data loss when the shared parity device is down

Impact of SCM Reliability



Comparison with other solutions



Main Conclusions

- Replacing the shared parity disk by a shared parity device increases the MTTDL of the array by 40 to 59 percent
- Adding a shared parity device that is 10 times more reliable than a regular disk to a pair of RAID 5 arrays increases the MTTDL of the array by at least 21,000 and up to 31,000 percent
- Shared parity organizations always outperform RAID level 6 organization

Cost Considerations

- SCM devices are still much more expensive that magnetic disks

- Replacing shared parity disk by
  a **pair of mirrored disks**
  would have achieved
  same performance improvements
  at a **much lower cost**

Additional Slides

Organization	*Relative MTTDL*
Two RAID 5 arrays	0.00096
All Disks	**1.0**
Two RAID 6 arrays	1.0012
SCM 5 $\times$ better	1.4274
SCM 10 $\times$ better	1.5080
SCN 100 $\times$ better	1.5887
SSD never fails	1.5982

CS@UH

Why we selected MTTDLs

- Much easier to compute than other reliability indices
- Data survival rates computed from MTTDL are a good approximation of actual data survival rates as long as disk MTTRs are at least one thousand times faster than disk MTTFs:
  - J.-F. Pâris, T. J. E. Schwarz, D. D. E. Long and A. Amer, When MTTDLs Are Not Good Enough: Providing Better Estimates of Disk Array Reliability, *Proc. 7th I2TS '08 Symp.*, Foz do Iguaçu, PR, Brazil, Dec. 2008.