WorkOut: I/O Workload Outsourcing for Boosting RAID Reconstruction Performance

Suzhen Wu[1], Hong Jiang[2], Dan Feng[1],
Lei Tian[12], Bo Mao[1]
[1]Huazhong University of Science & Technology
[2]University of Nebraska-Lincoln

Outline

- Background
- Motivation
- WorkOut
- Performance Evaluations
- Conclusion

RAID Reconstruction

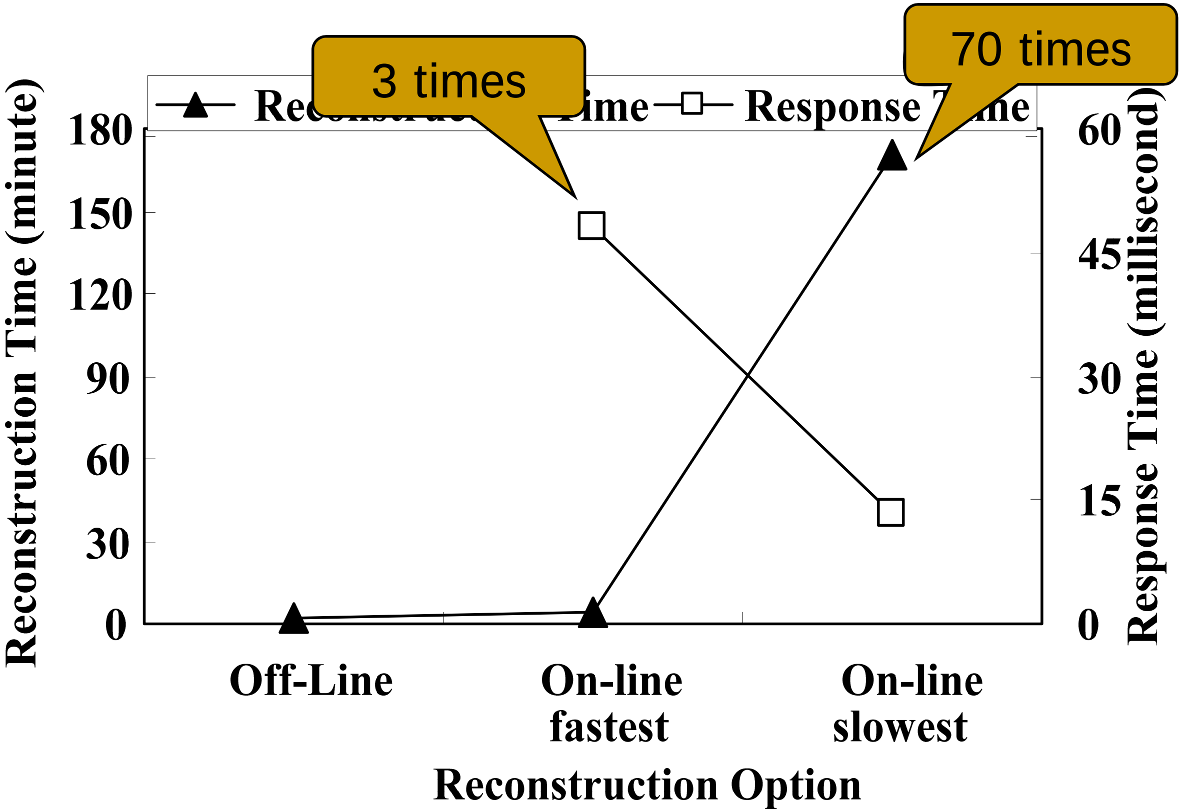
Recovers the data content on a failed disk

- Two metrics
  - Reconstruction time
  - User response time
- Categories
  - Off-line reconstruction
  - On-line reconstruction (*commonly deployed*)

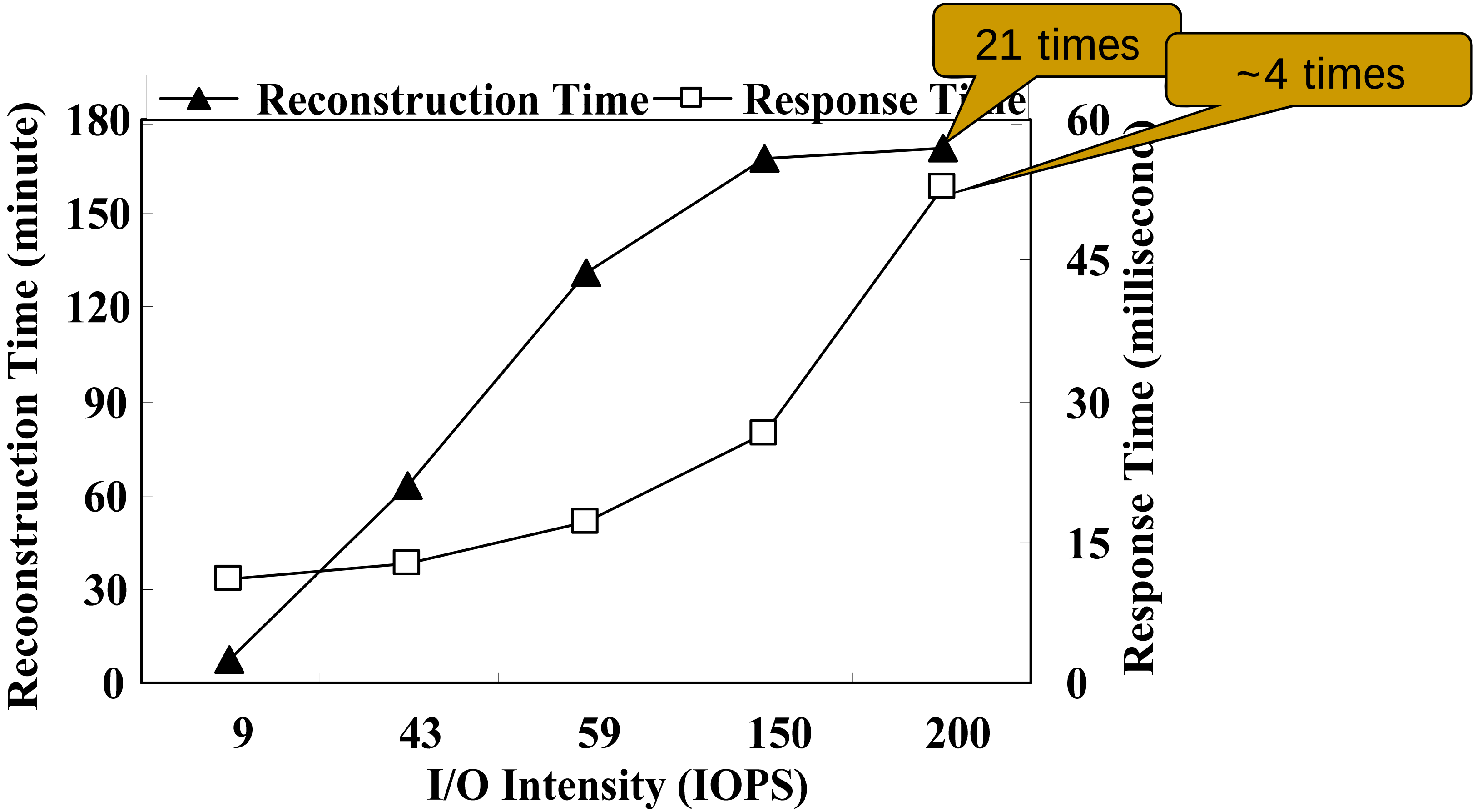
Challenges

- Higher error rates than expected
  - Complete disk failures [Schroeder07, Pinheiro07, Jiang08]
  - Latent sector errors [Bairavasundaram07]
- Correlation in drive failures
  - e.g. after one disk fails, another disk failure will likely occur soon.
- RAID reconstruction might become the common case in large-scale systems.
  - Increasing number of drives

Reconstruction and Its Performance Impact



I/O Intensity Impact on Reconstruction



- Both the reconstruction time and user response time increase with IOPS.

Intuitive Idea

- Observation
  - Performing the rebuild IOs and user IOs simultaneously leads to disk bandwidth contention and frequent long seeks to and from the multiple separate data areas.

- Our intuitive idea
  - To redirect the amount of user IOs that are issued to the degraded RAID set.
  - But, What to redirect? & Where to redirect to?

What To Redirect

- Access locality
  - Existing studies on workload analysis revealed that strong spatial and temporal locality exists even underneath the storage cache.
- Answer to "what to redirect?"
  - Popular read requests
  - All write requests

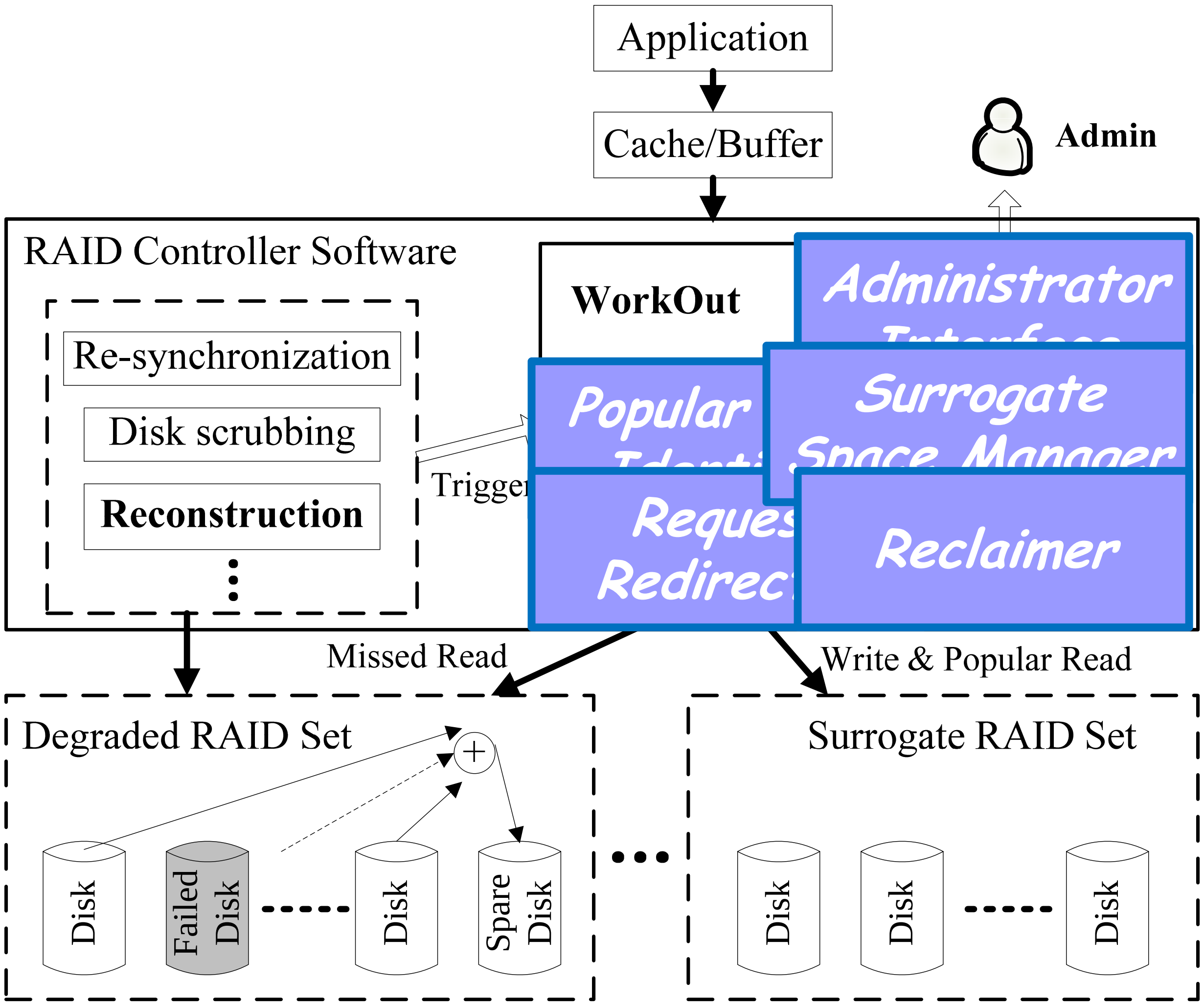
Where To Redirect To

- Availability of spare or free space in data centers
  - A spare pool including a number of disks
  - Free space on other RAID sets
- Answer to "Where to redirect to?"
  - Spare or free space
- Comparison
  - Existing approaches: in the context of a single RAID set
  - Our approach: in the context of data centers with multiple RAID sets

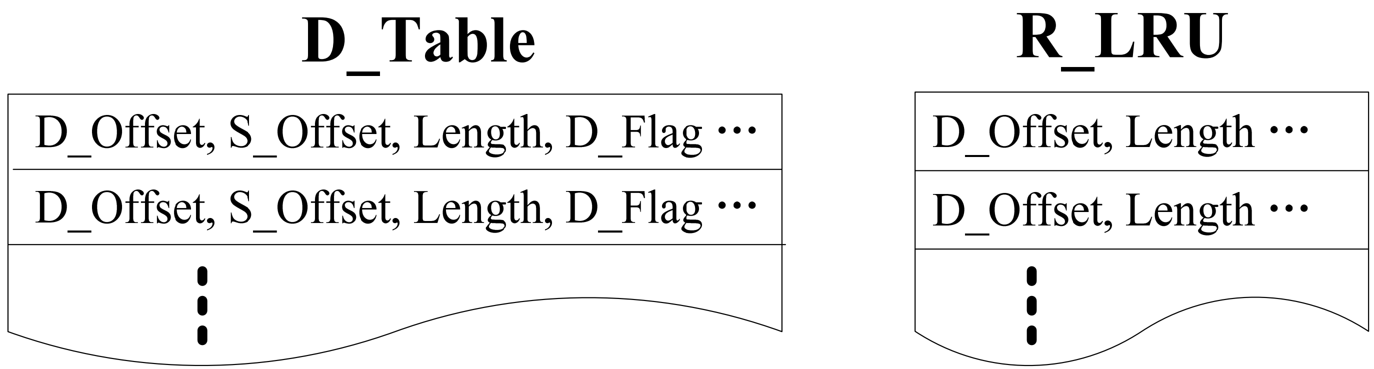
Main Idea of WorkOut

- Workload Outsourcing (Workout)
  - Temporarily redirect all write requests and popular read requests originally targeted at the degraded RAID set to a surrogate RAID set, to significantly improve on-line reconstruction performance.
- Goal
  - Approaches reconstruction-time performance of the off-line reconstruction without affecting user-response-time performance at the same time.

WorkOut Architecture



Data Structure



- D_Table: a log table that manages the redirected data
  - D_Flag=1: Write data from the user application
  - D_Flag=0: Popular read data from D-RAID to S-RAID
- R_LRU: an LRU-style list that identifies the most recent reads

Algorithm During Reconstruction

- Workflow
  - For each write, it will be redirected to its previous location or a new location on the surrogate RAID set according to whether it is an overwrite or not.
  - For each read, Check the D_Table:
    - Whether it hits D_Table or not?
      - If a hit, full hit or partial hit?
      - If a miss, whether it hits R_LRU?

Algorithm During Reclaim

- The redirected write data should be reclaimed back to the newly recovered RAID set after the reconstruction process completes.
- All requests must be checked in D_Table:
  - Each write request is served by the recovered RAID set and the corresponding log in D_Table should be deleted if it exists.
  - Read requests can be also handled well, but it is complicated to explain in a short time. More details can be found in our paper.

Design Choices

Optional surrogate RAID set	Device Overhead	Performance	Reliability	Maintainability
A dedicated surrogate RAID1 set	medium	medium	high	simple
A dedicated surrogate RAID5 set	high	high	high	simple
A live surrogate RAID5 set	low	low	medium-high	complicated

Data Consistency

- Data Protection
  - In order to avoid data loss caused by a disk failure in the surrogate RAID set, all redirected write data in the surrogate RAID set should be protected by a redundancy scheme, such as RAID1 or RAID5.
- "Metadata" Protection
  - The content of D_Table should be stored in a NVRAM during the entire period when WorkOut is activated, to prevent data loss in the event of a power supply failure.

Performance Evaluation

- Prototype implementation
  - A built-in module in MD
  - Incorporated into PR & PRO
- Experimental setup
  - Intel Xeon 3.0GHz processor, 1GB DDR memory, 15 Seagate SATA disks (10GB), Linux 2.6.11
- Methodology
  - Open-loop: trace replay
    - Trace: Financial1, Financial2, Websearch2
    - Tool: RAIDmeter
  - Closed-loop: TPC-C-like benchmark

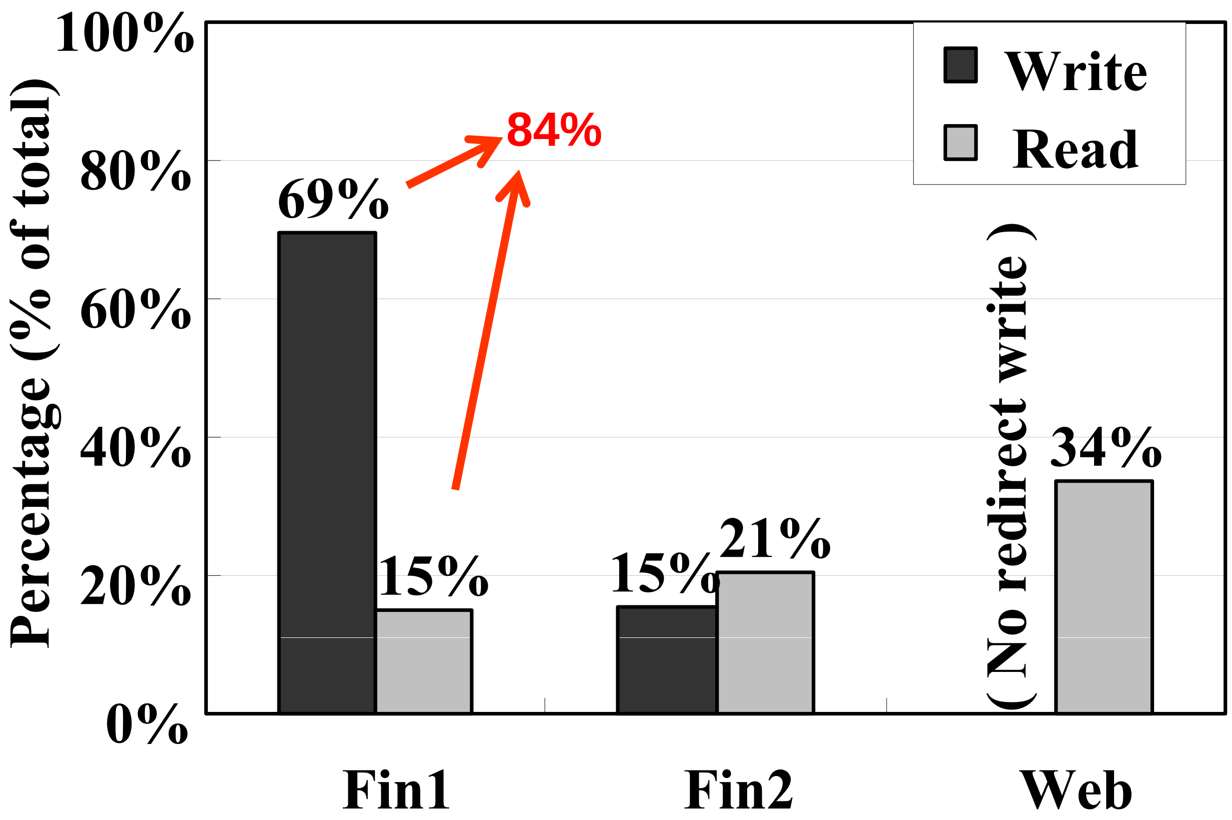
Experimental Results

Trace	Reconstruction Time (second)						
	Off-line	PR	WorkOut+PR	Speedup	PRO	WorkOut+PRO	Speedup
Fin1	136.4	1121.75	203.13	5.52	1109.62	188.26	**5.89**
Fin2		745.19	453.32	1.64	705.79	431.24	1.64
Web		9935.6	7623.22	1.30	9888.27	7851.36	1.26

Trace	Average User Response Time during Reconstruction (millisecond)							
	Normal	Degraded	PR	WorkOut+PR	Speedup	PRO	WorkOut+PRO	Speedup
Fin1	7.92	9.52	12.71	4.43	**2.87**	9.83	4.58	2.15
Fin2	8.13	13.36	25.8	9.69	2.66	22.97	10.19	2.25
Web	18.46	26.95	38.57	28.35	1.36	35.58	29.12	1.22

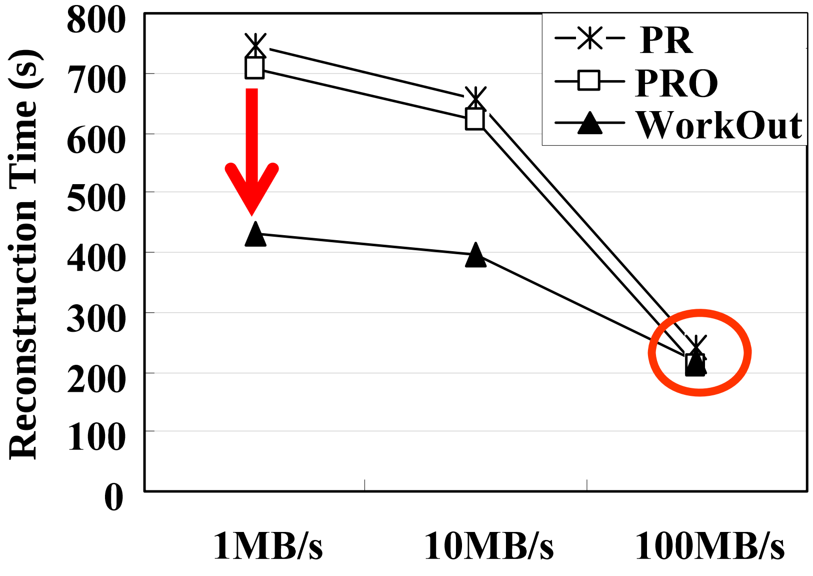
- Degraded RAID set: RAID5, 8 disks, 64KB stripe unit size
- Surrogate RAID set: RAID5, 4 disks, 64KB stripe unit size
- Minimum reconstruction bandwidth: 1MB/s

Percentage of Redirected Requests

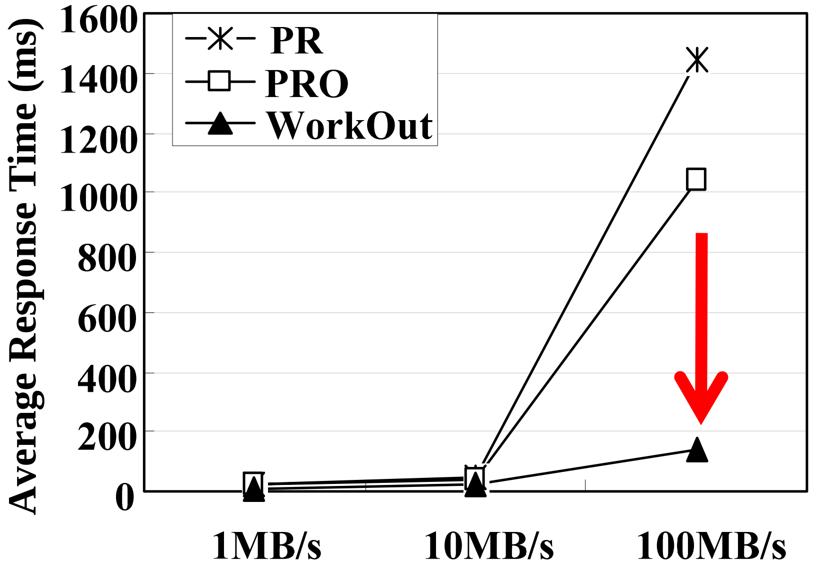


- Minimum reconstruction bandwidth of 1MB/s

Sensitivity Study (1)



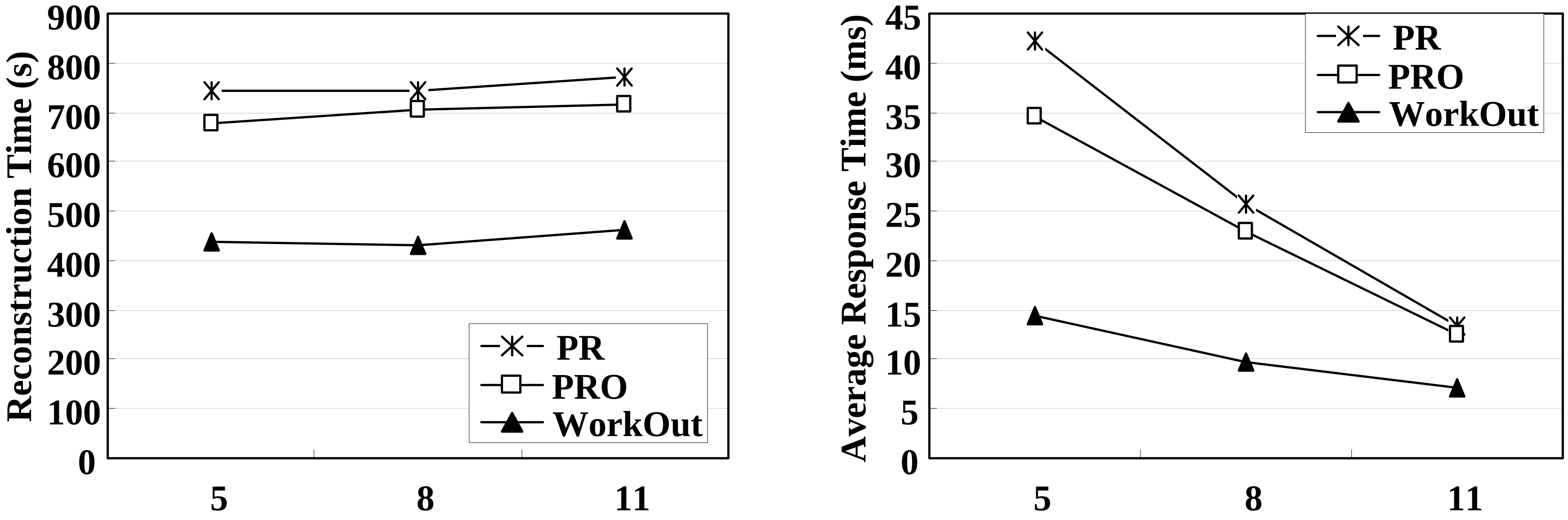
(a)



(b)

- Different minimum reconstruction bandwidth: 1MB/s, 10MB/s, 100MB/s

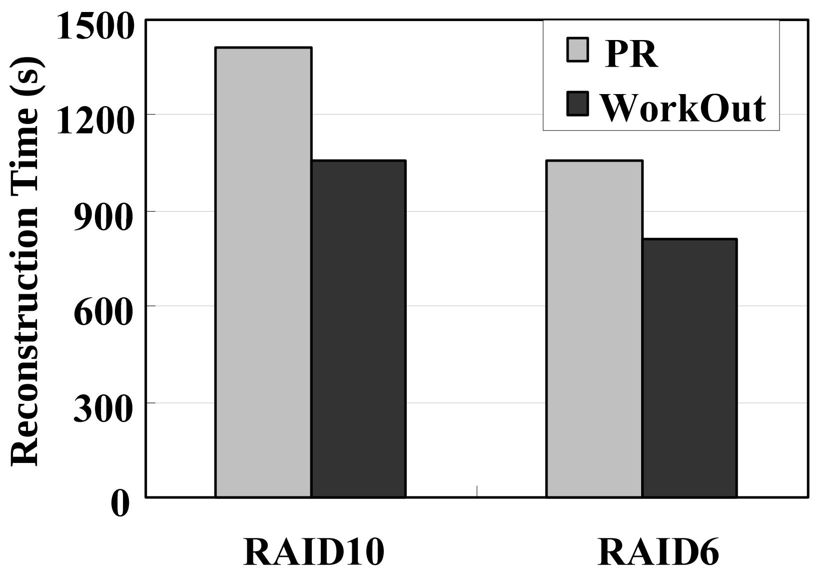
Sensitivity Study (2)



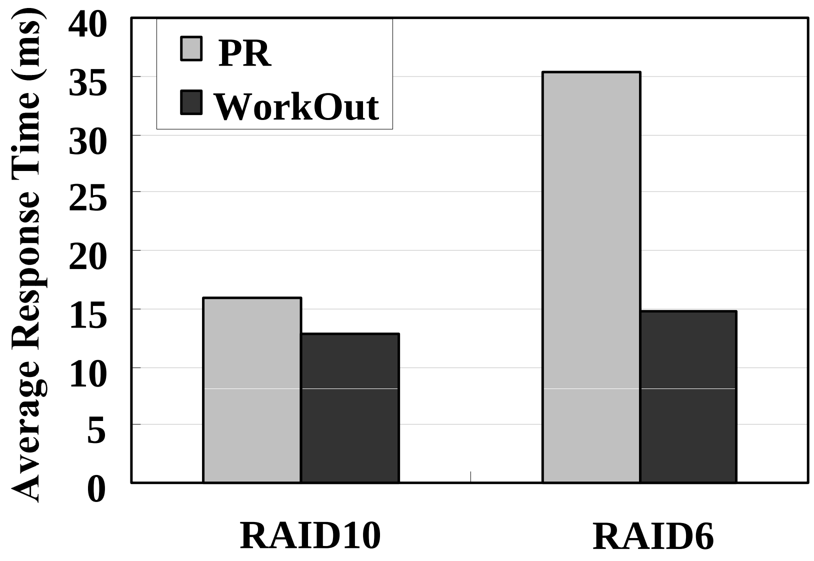
(a) (b)

- Different number of disks (5, 8, 11)

Sensitivity Study (3)



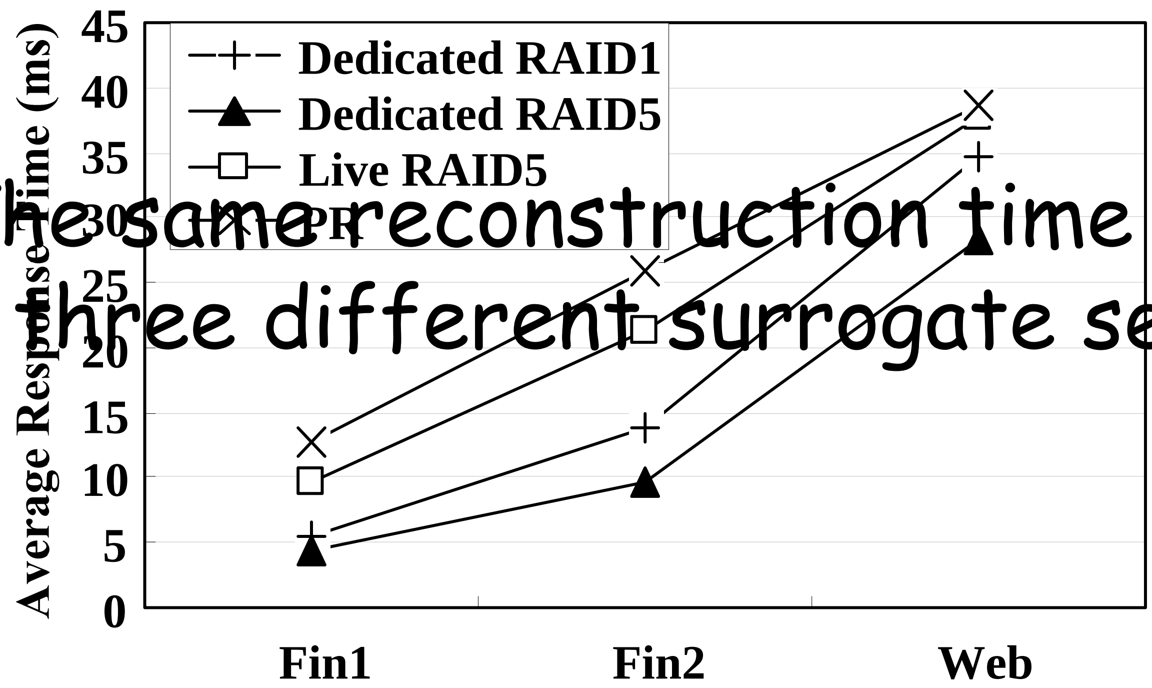
(a)



(b)

- Different RAID level: RAID10 (4 disks), RAID6 (8 disks)

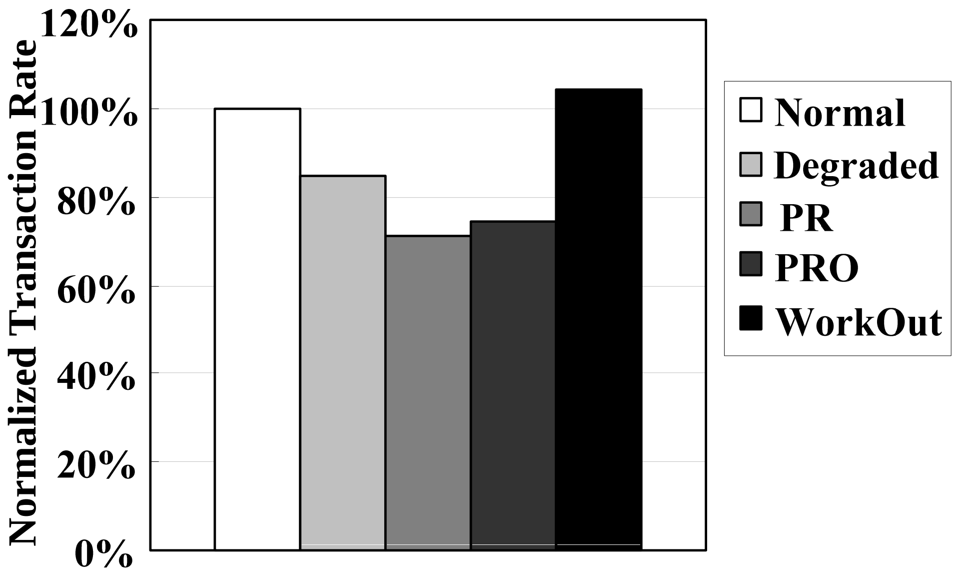
Different Surrogate Set



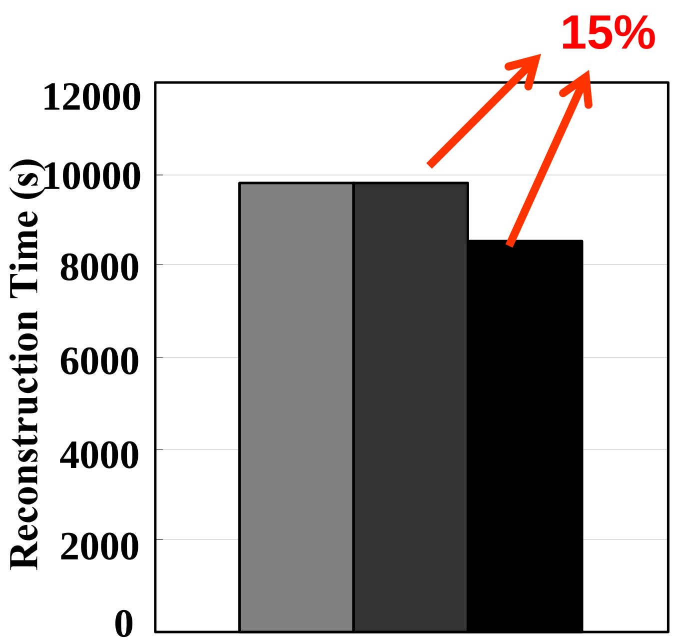
The same reconstruction time for the three different surrogate sets

- Dedicated RAID1: 2 disks
- Dedicated RAID5: 4 disks
- Live RAID5: 4 disks (Replaying the Fin1 workload on it)

TPC-C-like Benchmark



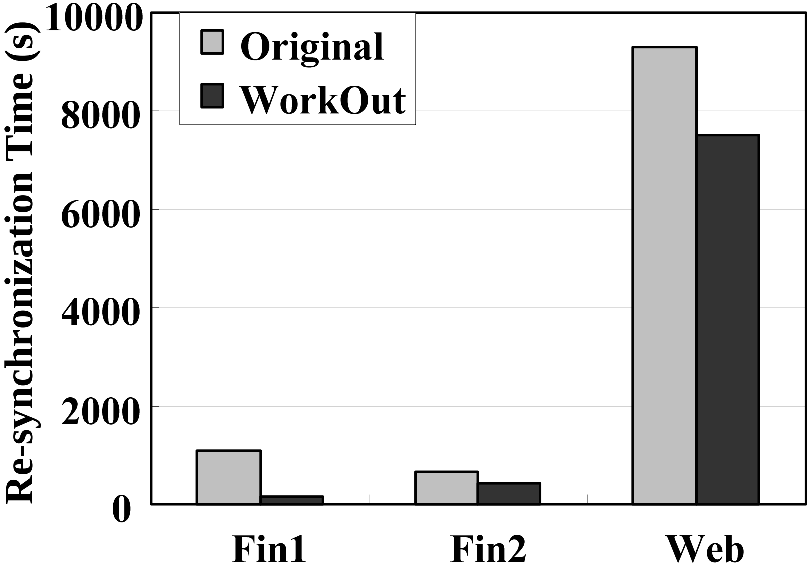
(a) Transaction rate



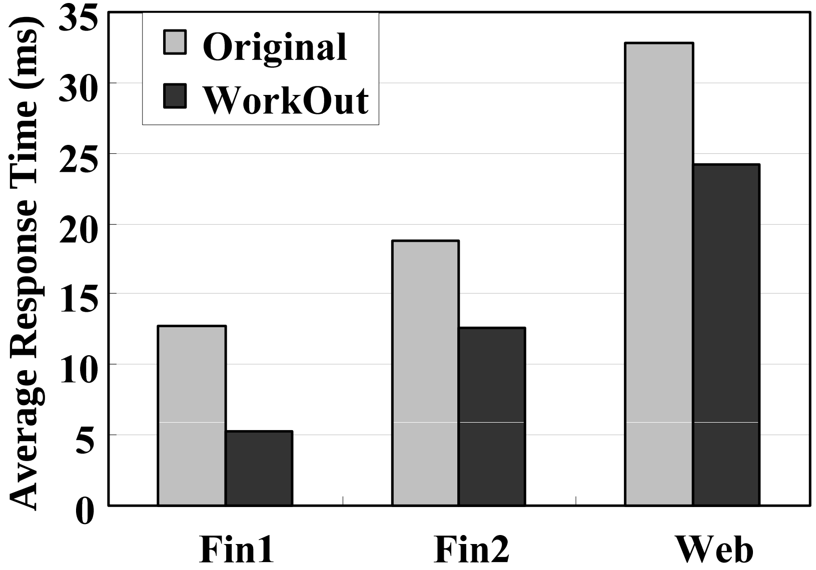
(b) Reconstruction time

- Minimum reconstruction bandwidth of 1MB/s

Extendibility—Re-synchronization



(a)



(b)

- Re-synchronization: RAID5, 8 disks, 64KB stripe unit size
- Surrogate RAID set: RAID5, 4 disks, 64KB stripe unit size
- Minimum Re-synchronization bandwidth: 1MB/s

Conclusion

- WorkOut outsources a significant amount of user I/O requests away from the degraded RAID set to a surrogate RAID set, thus improving RAID reconstruction performance;
- Insights and guidance for storage system designers and administrators by exploiting three design options;
- WorkOut can improve the performance of other background support RAID tasks such as re-synchronization.

Q & A ?

Thanks!