IDO: Intelligent Data Outsourcing with Improved RAID Reconstruction Performance in Large-Scale Data Centers

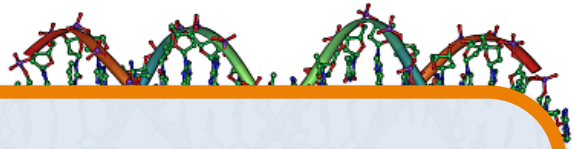
Suzhen Wu§*, Hong Jiang*, **Bo Mao***
§Xiamen University
*University of Nebraska–Lincoln

Data Deluge

Social Network

2,300 tweets per second

Scientific Simulation



How to safely store such a huge data volume proposes a big challenge to the system administrators!

Mobile Apps

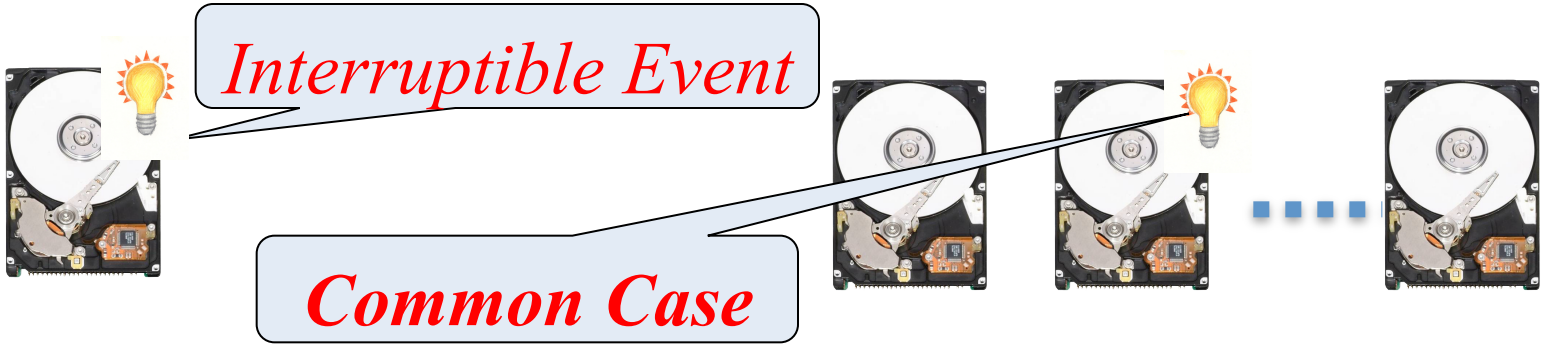
275 EB data flowing per day in 2020

Business Intelligence

Where Are We?

Laptop and Desktop

Data Center



Disk Failure in the Real World

- Higher error rates than expected
  - Complete disk failures, 2%~4% on average;
  - Latent sector errors, 3.45%;
- Correlation in drive failures
  - e.g., after one disk fails, another disk failure will likely occur soon.
- RAID reconstruction becomes an operational state in data centers
  - Increasing disk capacity and number of drives

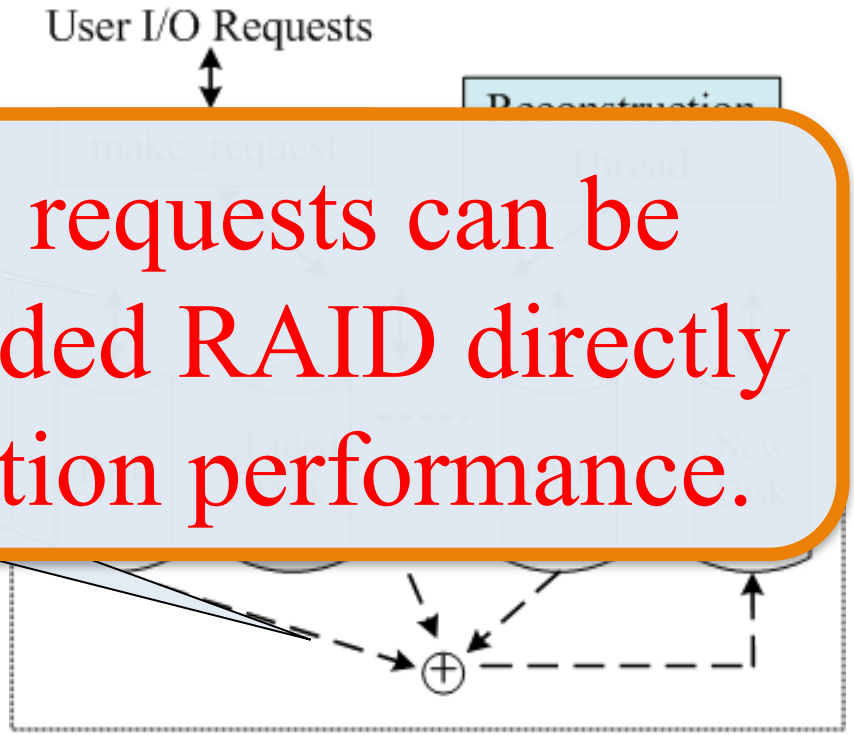
More Observations

- Linux software RAID (MD) mailing list: too many complains about the slow recovery speed.

1.	VERY slow mdadm recovery speed 12KB/s	linux-raid
2.	Why is kaha recovery very slow,can I speed up it?	activemq-users
3.	[HACKERS] [COMMITTERS] Re: pgsql: Add URLs for : * Speed WAL	pgsql-hackers
4.	[COMMITTERS] pgsql: Update TODO description: * Speed WAL re	pgsql-committe
5.	[COMMITTERS] Re: pgsql: Add URLs for : * Speed WAL recovery	pgsql-committe
6.	[COMMITTERS] pgsql: Add URLs for : * Speed WAL recovery	pgsql-committe
7.	[COMMITTERS] pgsql: Add URLs for : * Speed WAL recovery by	pgsql-committe
8.	[COMMITTERS] pgsql: Add: > * Speed WAL recovery by allowing	pgsql-committe
9.	Journal recovery speed	activemq-users
10.	recovery speed on many-disk RAID 1	linux-raid
11.	Recovery speed at 1MB/s/device, unable to change	linux-raid
12.	[Evms-devel] Raid1 recovery speed	evms-devel
13.	[Evms-devel] recovery speed	evms-devel
14.	Recovery/Import Loading Speed	mysql

RAID Reconstruction Challenges

- Online RAID Reconstruction:



How many user I/O requests can be eliminated from degraded RAID directly affects the reconstruction performance.

- Two challenges:
  - Real-time user performance;
  - Window of vulnerability.

The State of the arts

- Optimizing the reconstruction workflow:
  - DOR (CMU PDL)
  - Live-block recovery (USENIX FAST'04)
  - PRO (USENIX FAST'07)
- Optimizing the user I/O requests:
  - MICRO (IEEE TC'08)
  - WorkOut (USENIX FAST'09)
  - VDF (USENIX ATC'11)

Compare with State of the arts

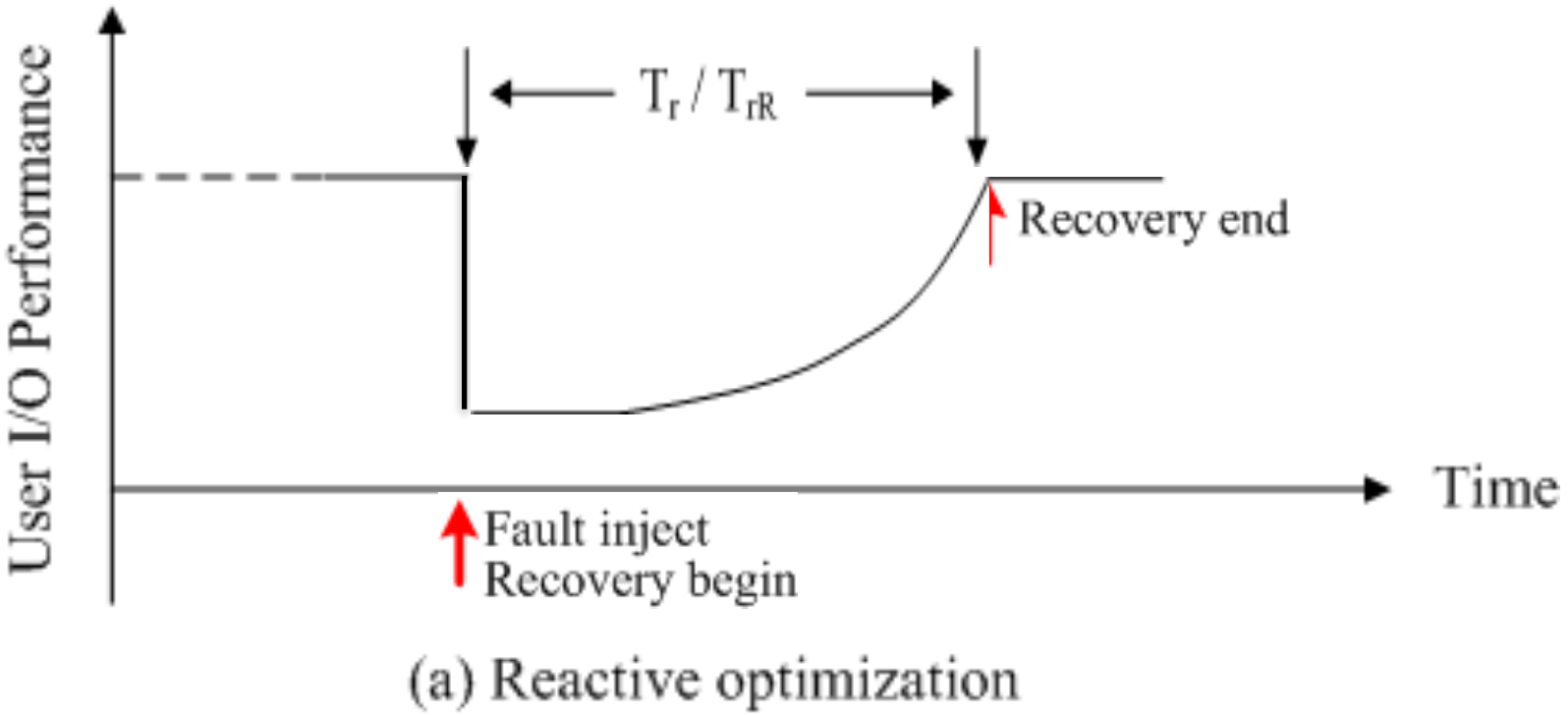
Characteristics	PRO (FAST'07)	WorkOut (FAST'09)	VDF (USENIX'11)	IDO (LISA'12)
Proactive				✓
Temporal Locality	✓	✓	✓	✓
Spatial Locality	✓			✓
User I/O		✓	✓	✓
Reconstruction I/O	✓			✓

Observation 1

- RAID reconstruction is an operational state in large-scale data centers which means reactive scheme is inefficient.
  - Reactive vs. Proactive?

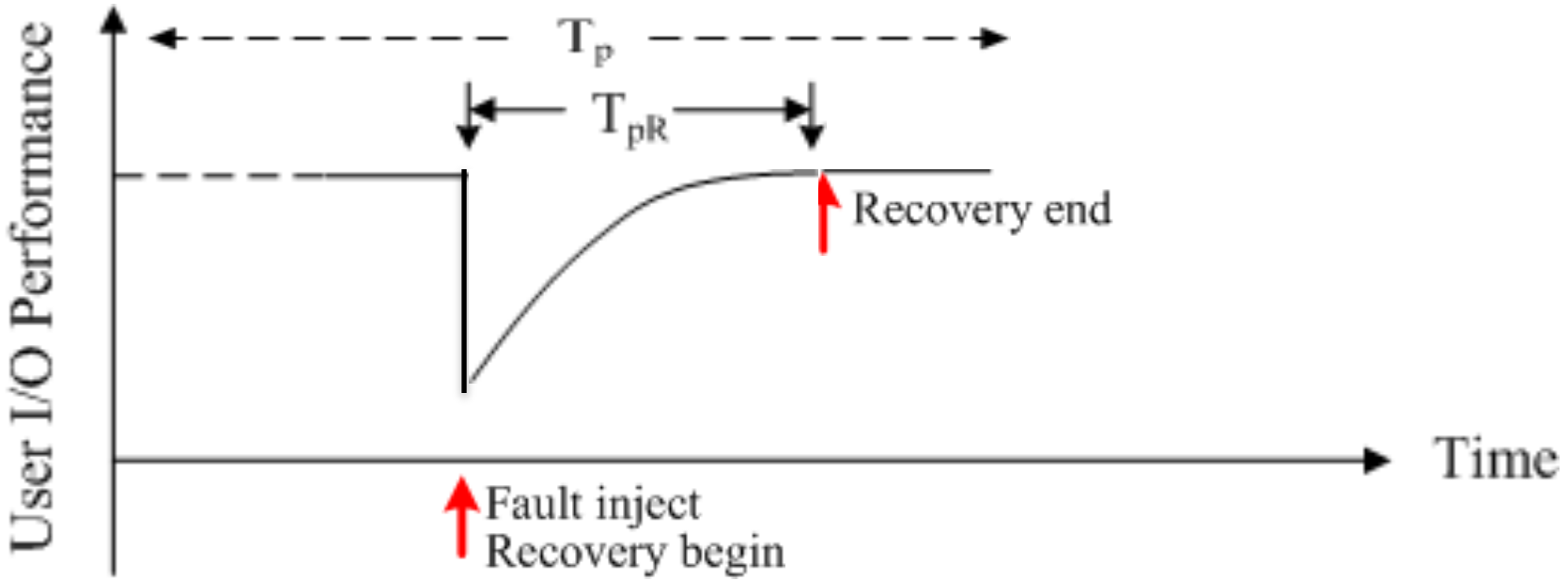
- *Existing studies are all reactive schemes.*

Example 1: Reactive vs. Proactive



(a) Reactive optimization

Example 1: Reactive vs. Proactive



(b) Proactive optimization

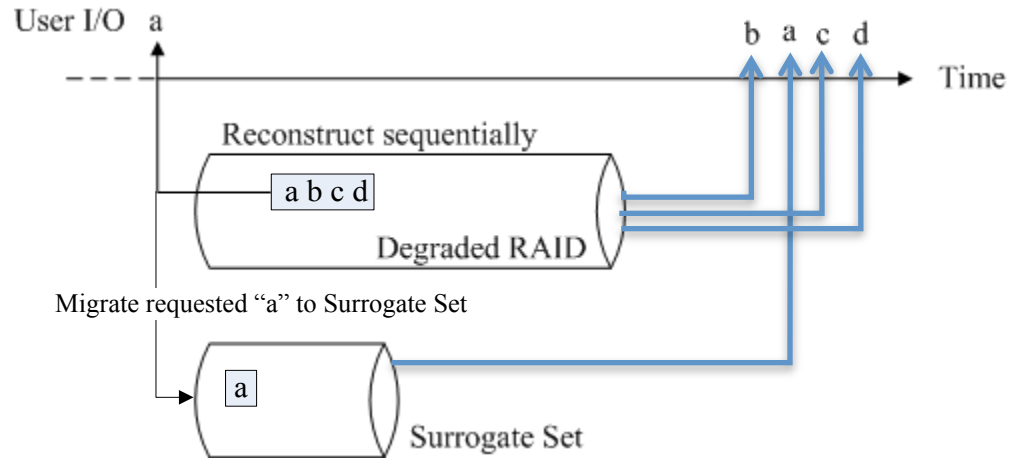
Observation 2

- With large RAM and SSDs, the temporary locality is poor at HDD level. However, the spatial locality is good due to the sequential accesses of HDDs.
  - Temporal locality vs. Spatial locality?

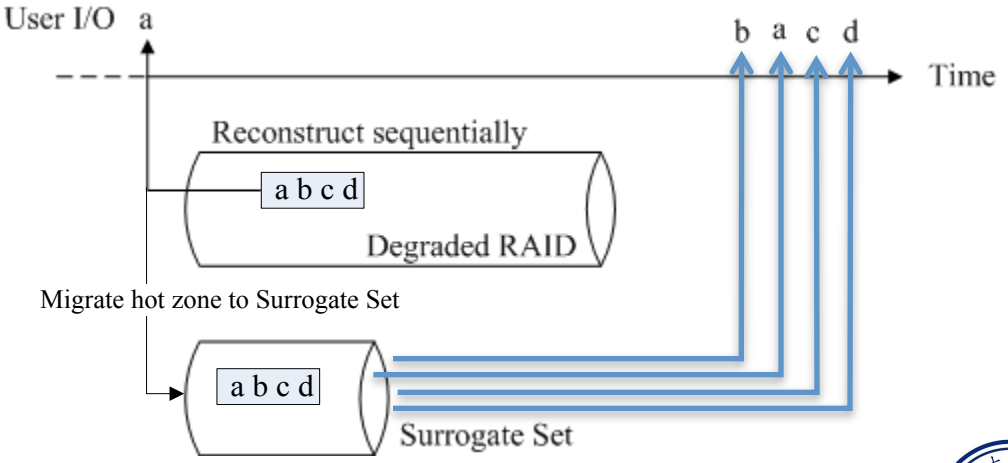
- *Existing studies mostly focus on temporal locality and ignore spatial locality.*

Example 2: Temporal vs. Spatial

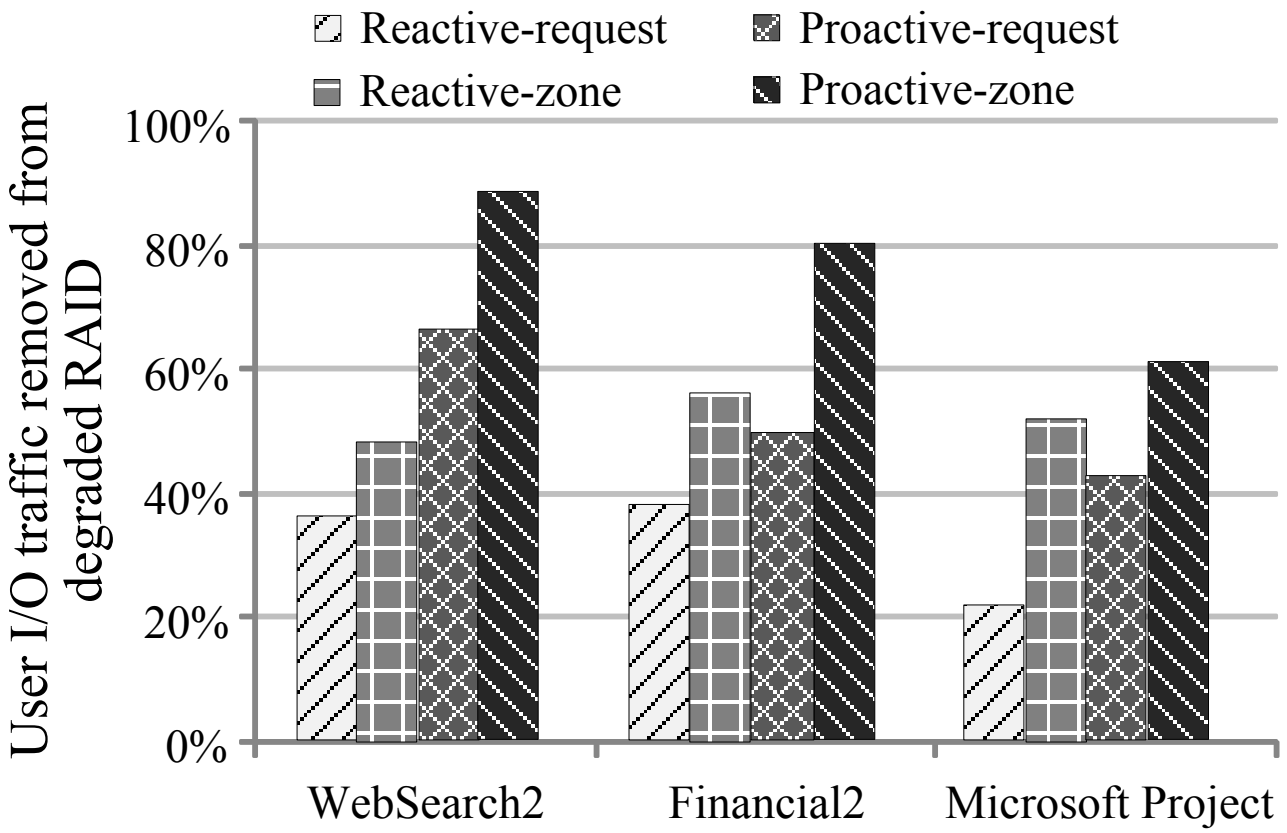
(a) Request-based approach



(b) Zone-based approach



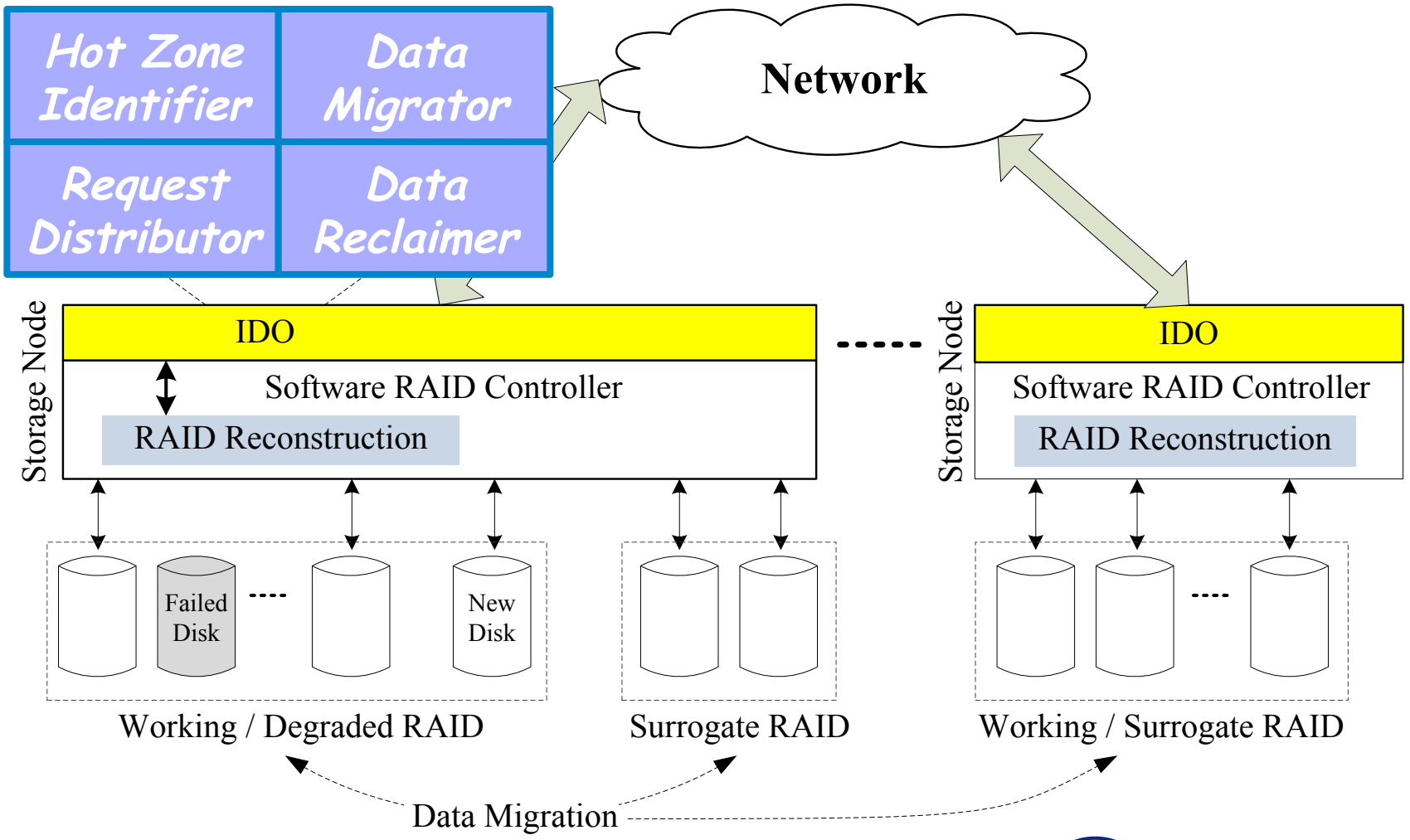
The Motivation



IDO: Intelligent Data Outsourcing

- The main idea:
  - Proactively identify the hot data zones;
  - Upon disk failure,
    - Recovery the hot data zones first;
    - Migrate the hot data zones to surrogate set;
    - Redirect the user I/O requests.
- The design objectives
  - Reducing reconstruction time;
  - Improving the user I/O performance;
  - Applicable to other background tasks.

System Overview

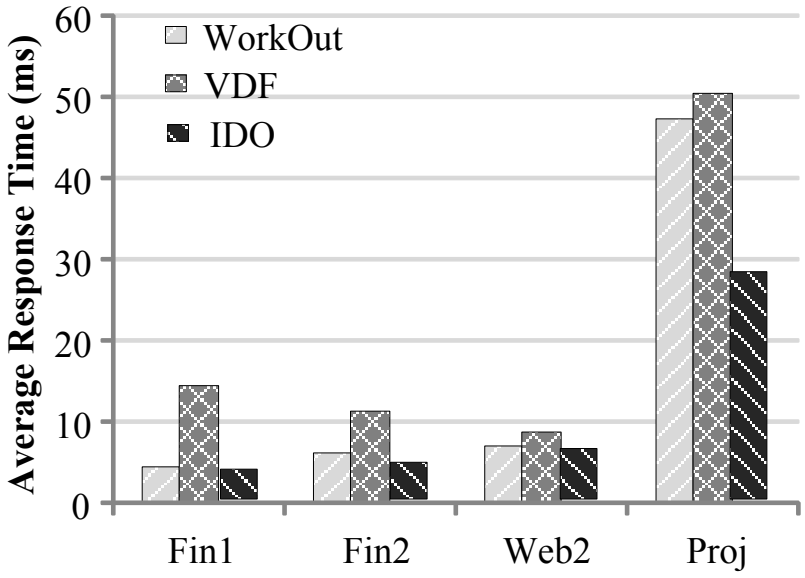


Performance Evaluation

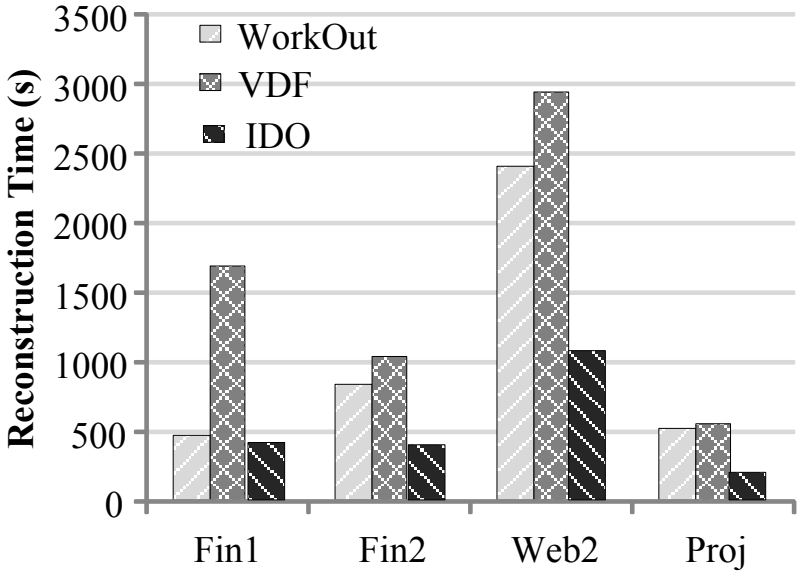
- IDO prototype is a built-in module in Linux MD, compared with WorkOut and VDF.
- Intel Xeon 3440 processor, 8GB DDR memory, WDC WD1600AAJS SATA disks.
- Trace-driven evaluations

Trace	Trace Characteristic		
	Read Ratio	IOPS	Aver. Req. Size(KB)
Fin1	32.8%	69	6.2
Fin2	82.4%	125	2.2
Web2	100%	113	15.1
Proj	97.6%	29	57.8

RAID5 Results

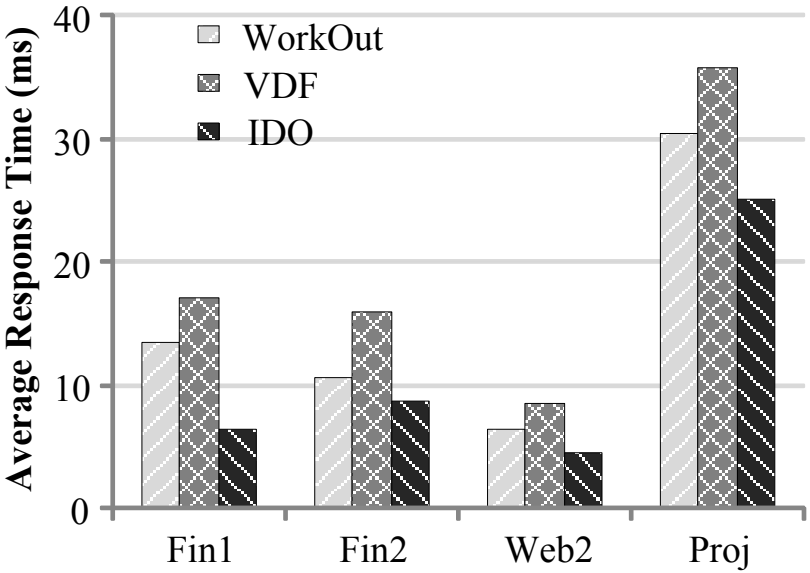


(a) Average Response Time during Recovery

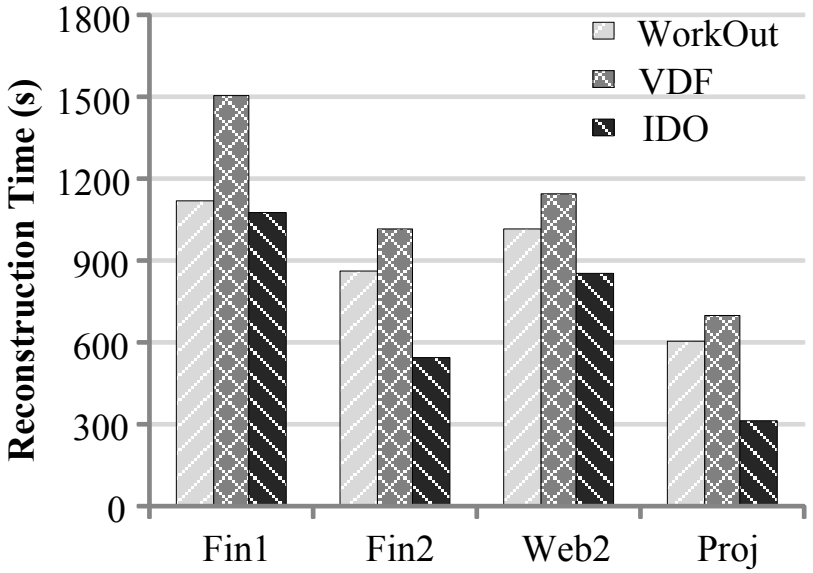


(b) Reconstruction Time

RAID6 Results



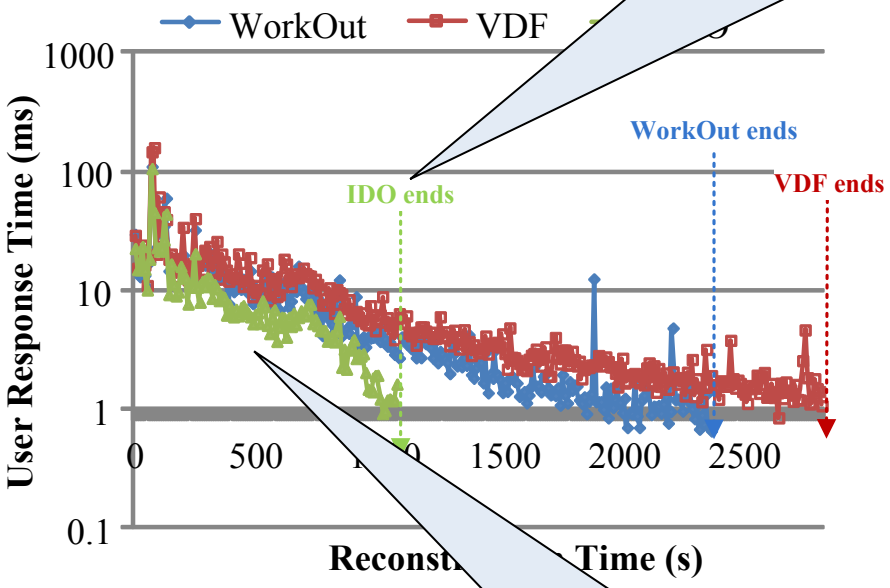
(a) Average Response Time during Recovery

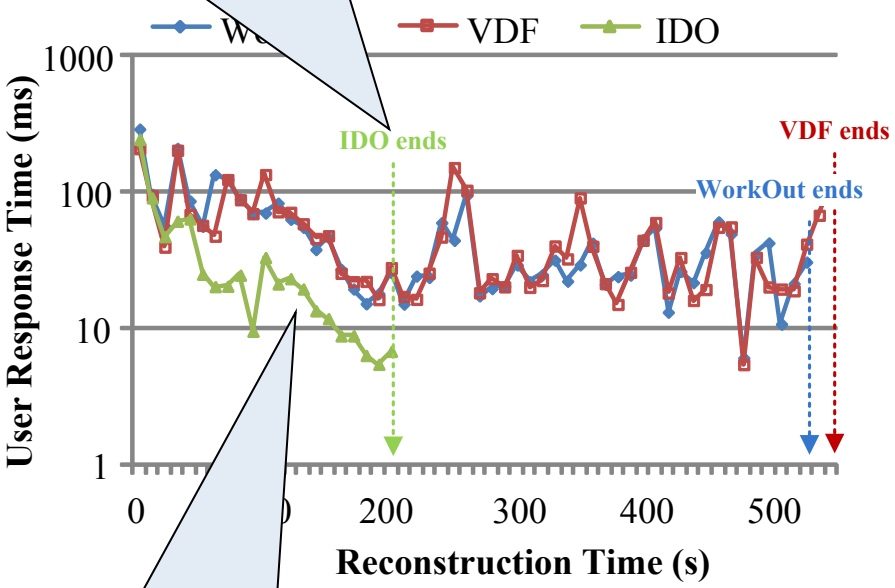


(b) Reconstruction Time

Detailed Real-time Results

Shorter Reconstruction Times





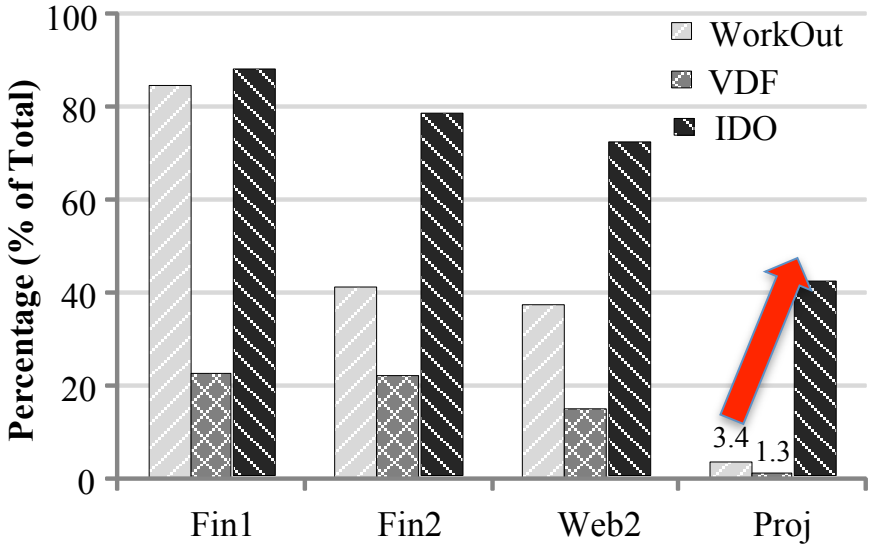
(a) WebSearch2.s

Microsoft Project

Lower user response times

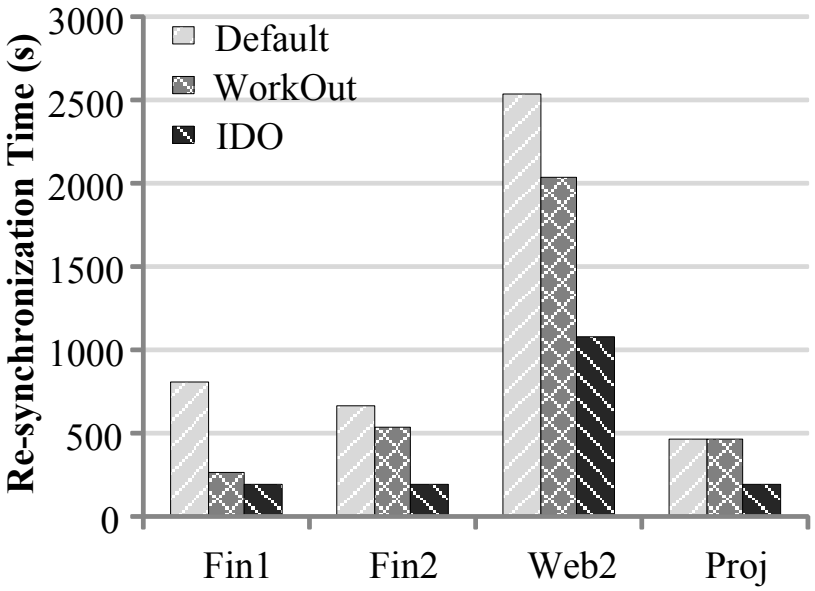
Reduce I/Os and Sensitivity Study

- Reduced I/Os:

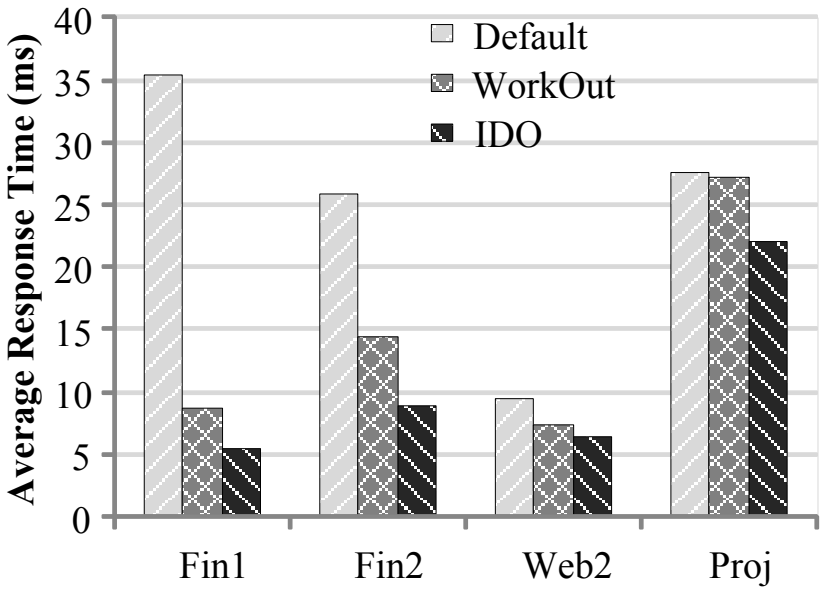


- Sensitivity & overhead analysis (in the paper).

Extendibility Evaluation



(a) Re-synchronization Time



(b) Average Response Time

Summary of IDO

- RAID reconstruction is an operational state in large-scale data centers!
- Salient features of IDO:
  - Proactive;
  - Exploit both temporal and spatial localities;
  - Optimize both user and reconstruction IOs;
  - Portability and extendibility.

Thanks!