

# **S<sup>2</sup>-RAID: A New RAID Architecture for Fast Data Recovery**

**Jiguang Wan\*, Jibin Wang\*, Qing Yang+, and Changsheng Xie\***

***\*Huazhong University of Science and Technology, China***

***+University of Rhode Island, USA***

# Overview

- A reconstruction solution-S<sup>2</sup>-RAID
  - Using parallel data layout to boost data construction
- Online reconstruction performance
  - Average user response time
  - Shorten reconstruction time by a factor of 3~6
    - Comparing with the traditional RAID

# Outline

- Reconstruction background
- Data layout strategy
- S<sup>2</sup>-RAID prototype
- Evaluation results
- Performance analyse

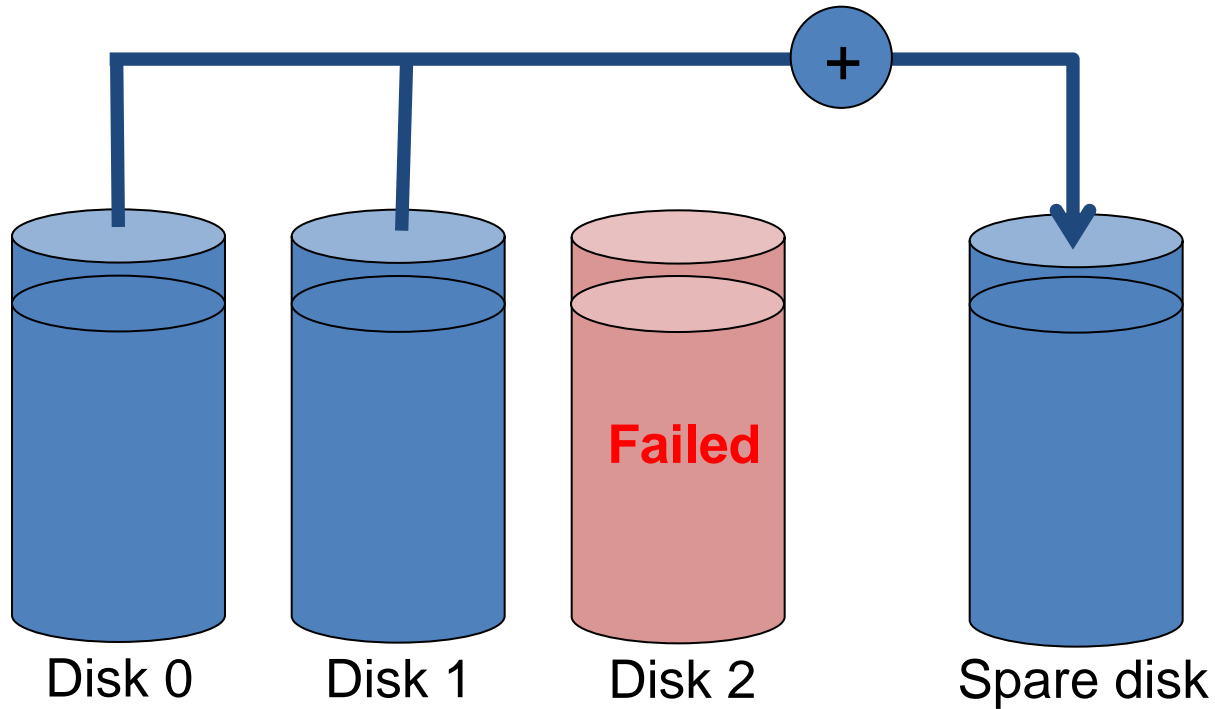
# Background

- High-capacity disk keep increasing.
- Offline reconstruction is result in service down time.
- Existing reconstruction solutions
  - Long reconstruction time and Average user response time

# S<sup>2</sup>-RAID Idea

- Our goals
  - Reducing construction time sharply
  - Maximizing Parallel reconstruction
  - Minimizing the impact on front end performance.
- S<sup>2</sup>-RAID data layout
  - Parallel reconstruction model
  - Using “subRAID” concept
  - Each subRAID uses standard RAID

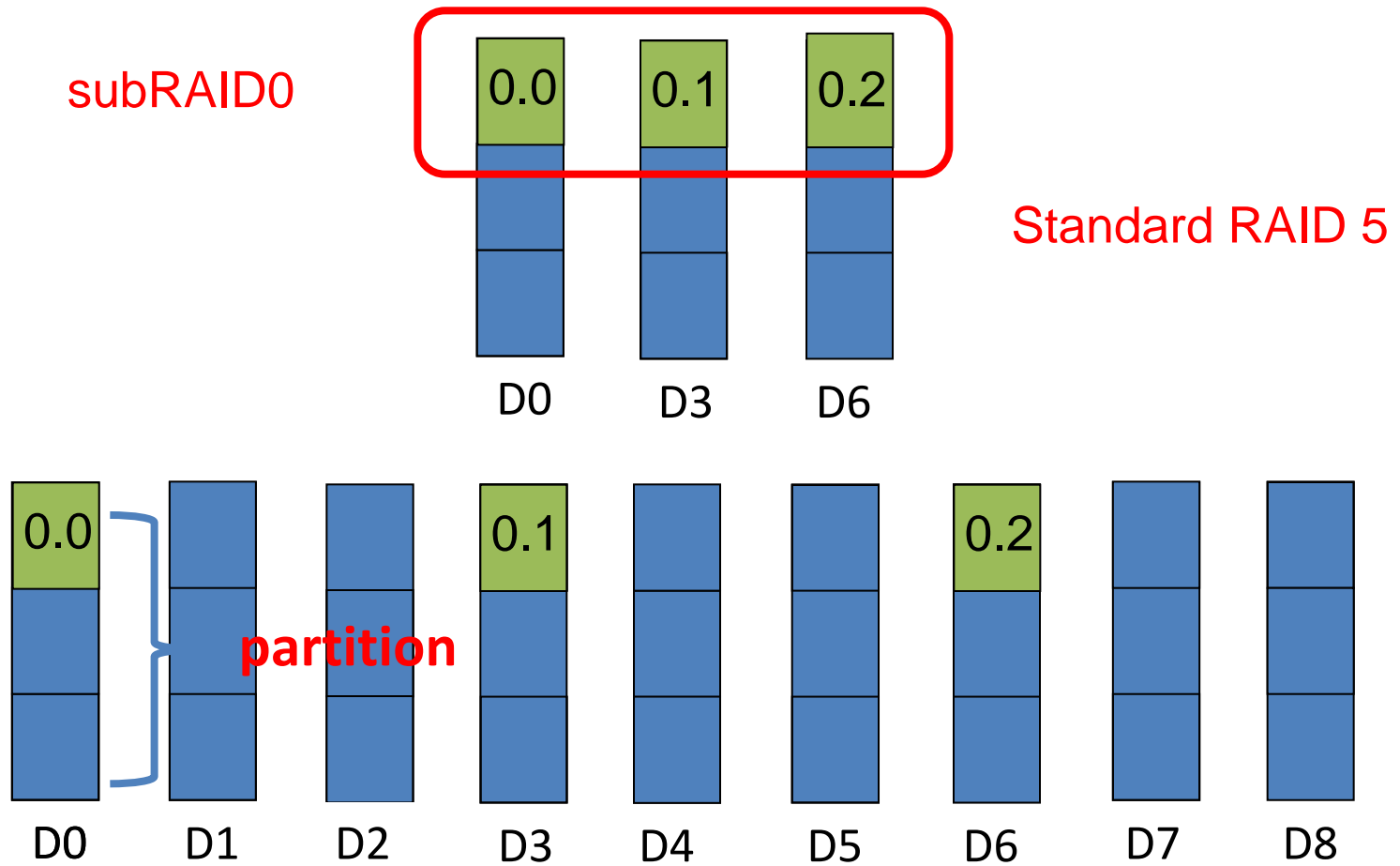
# Traditional RAID 5 reconstruction



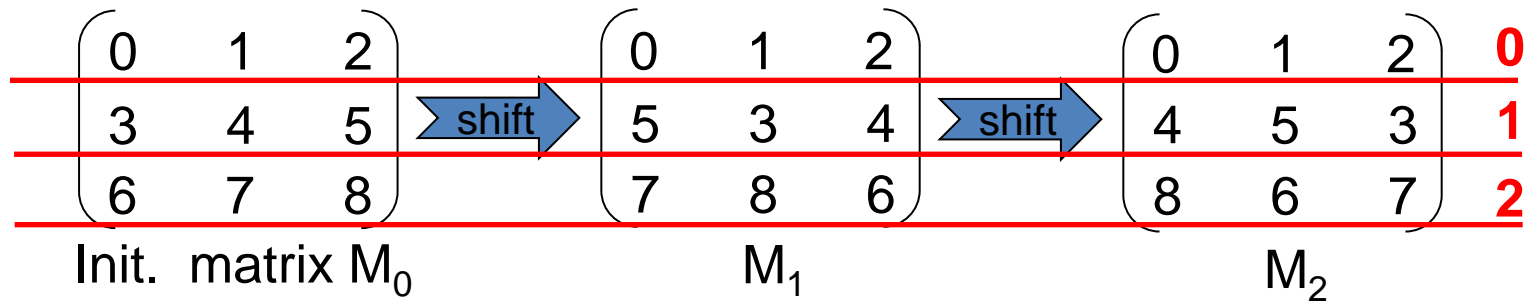
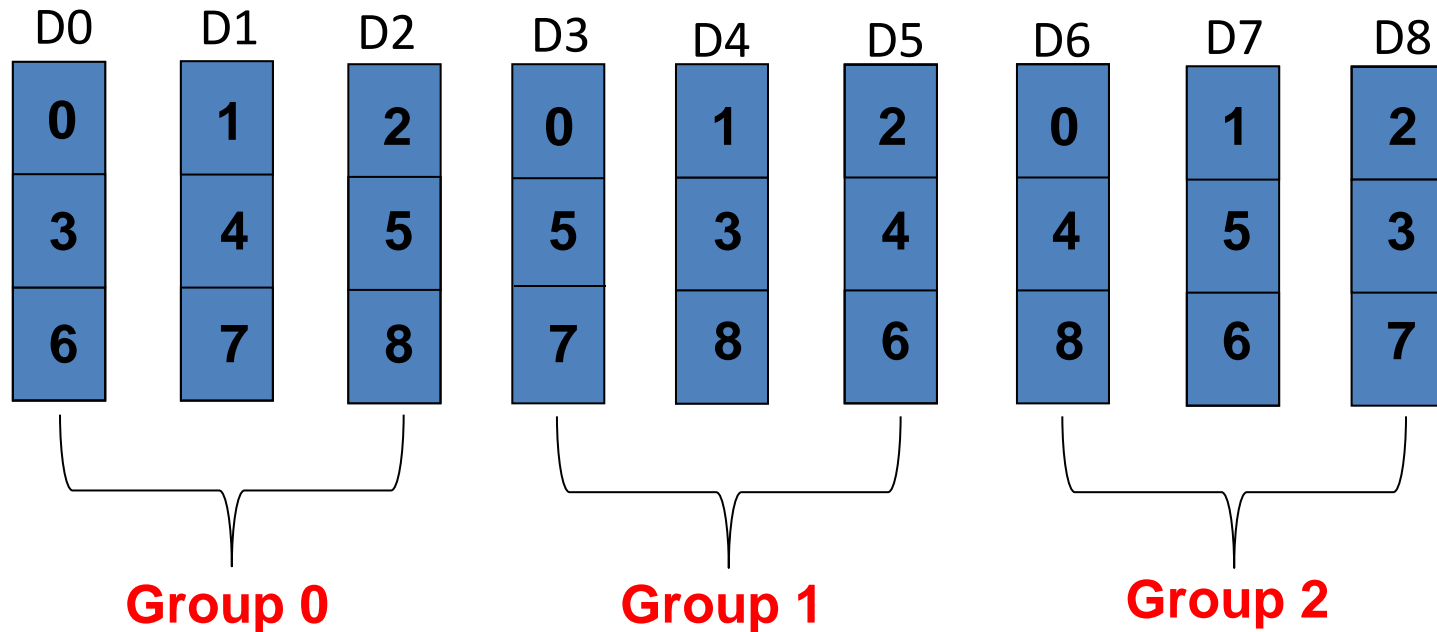
**Single reconstruction stream**

**long reconstruction time**

# S<sup>2</sup>-RAID data layout



# S<sup>2</sup>-RAID data layout structure



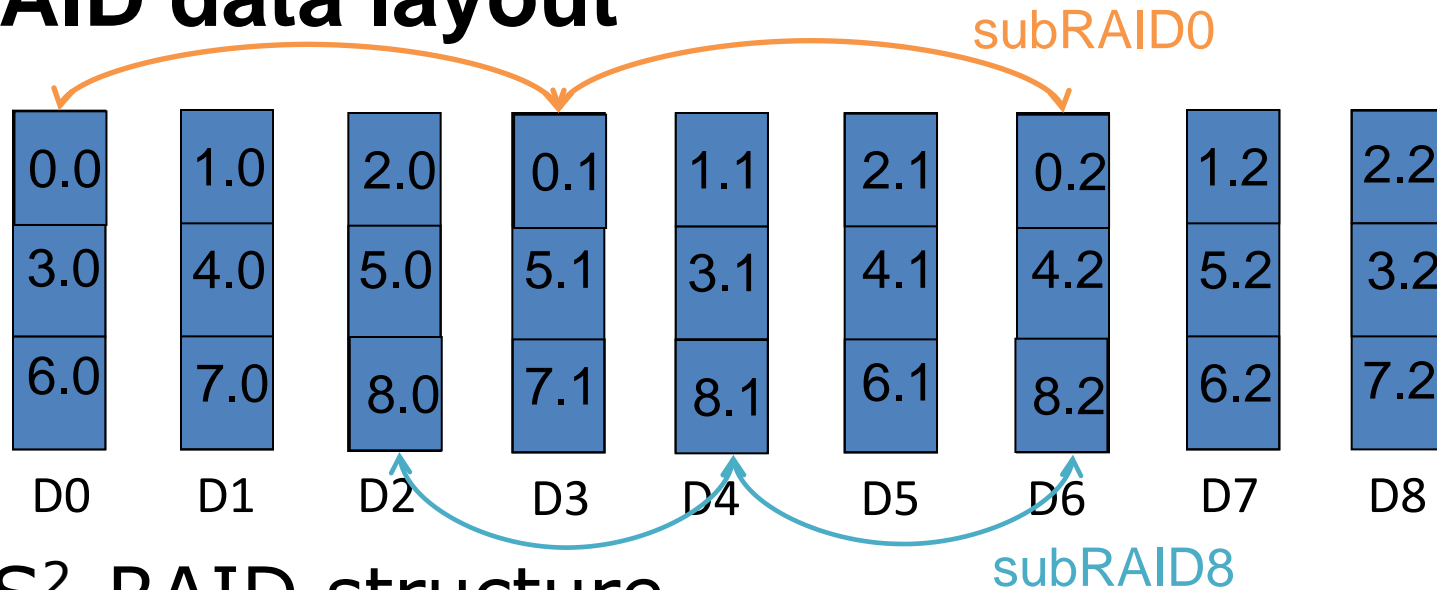
# S<sup>2</sup>-RAID data layout structure

**P<sub>i,j</sub>** : subRAID numbers of the  $(j+1)^{th}$  partition on disks of  $(i+1)^{th}$  group in the RAID  
**K**: the partition number of the disk

$$m_0 = \begin{pmatrix} P_{0,0} \\ P_{0,1} \\ P_{0,2} \\ \dots \\ P_{0,K-1} \end{pmatrix} \quad m_1 = \begin{pmatrix} P_{1,0} \\ P_{1,1} \\ P_{1,2} \\ \dots \\ P_{1,K-1} \end{pmatrix} = \begin{pmatrix} SH_r^0(P_{0,0}) \\ SH_r^1(P_{0,1}) \\ SH_r^2(P_{0,2}) \\ \dots \\ SH_r^{K-1}(P_{0,K-1}) \end{pmatrix} \quad m_i = \begin{pmatrix} P_{i,0} \\ P_{i,1} \\ P_{i,2} \\ \dots \\ P_{i,K-1} \end{pmatrix} = \begin{pmatrix} SH_r^0(P_{i-1,0}) \\ SH_r^1(P_{i-1,1}) \\ SH_r^2(P_{i-1,2}) \\ \dots \\ SH_r^{K-1}(P_{i-1,K-1}) \end{pmatrix}$$

**Note:** the size of the group must be a prime number

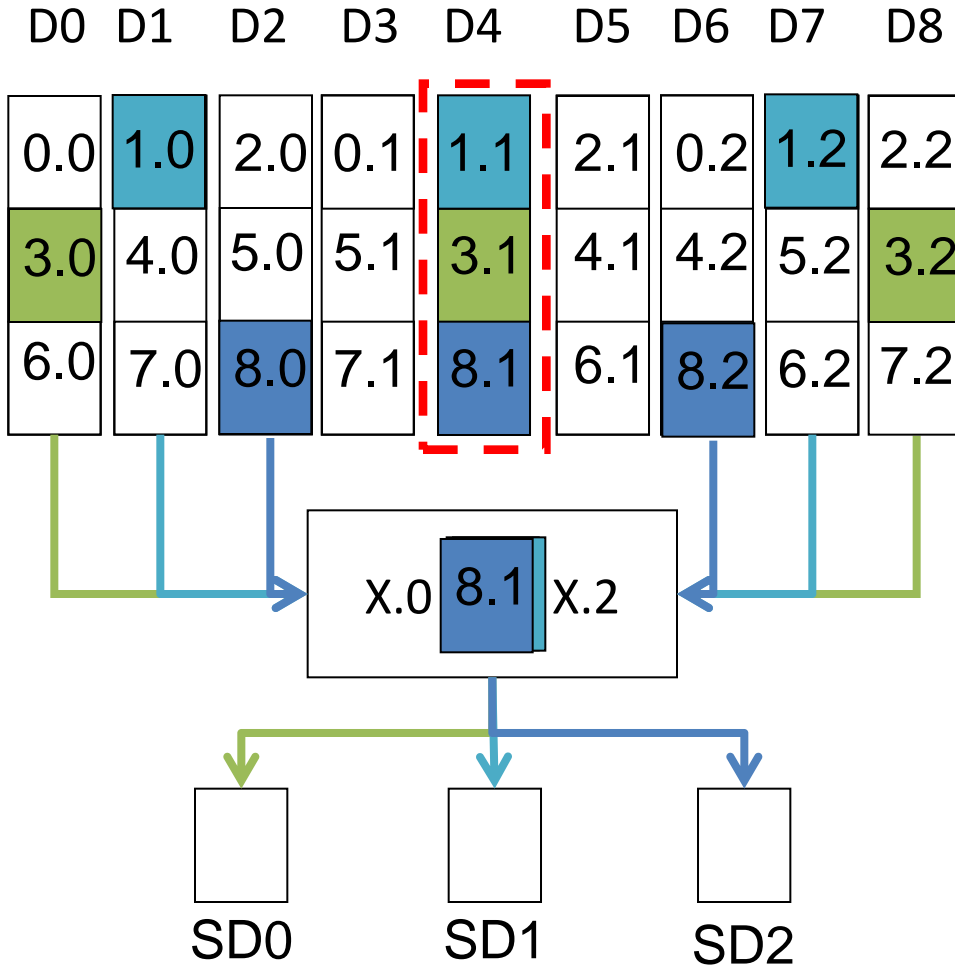
# S<sup>2</sup>-RAID data layout



- S<sup>2</sup>-RAID structure

- 9 disks
- 9 subRAIDs
- RAID type
- RAID 5、RAID 10、RAID 6 etc.

# S<sup>2</sup>-RAID 5 reconstruction



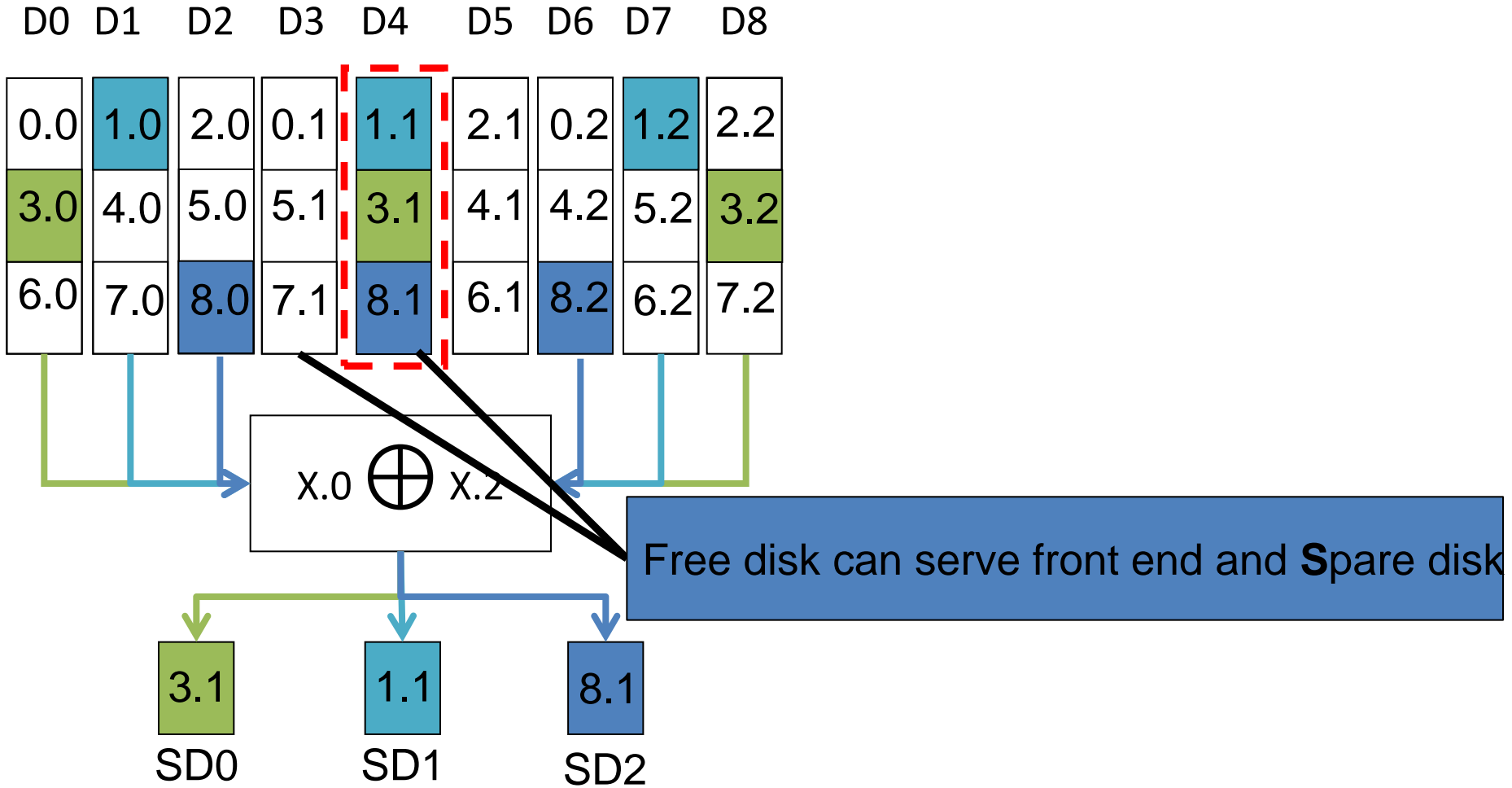
D4 was divided into 3 partitions

Reconstruction speed!

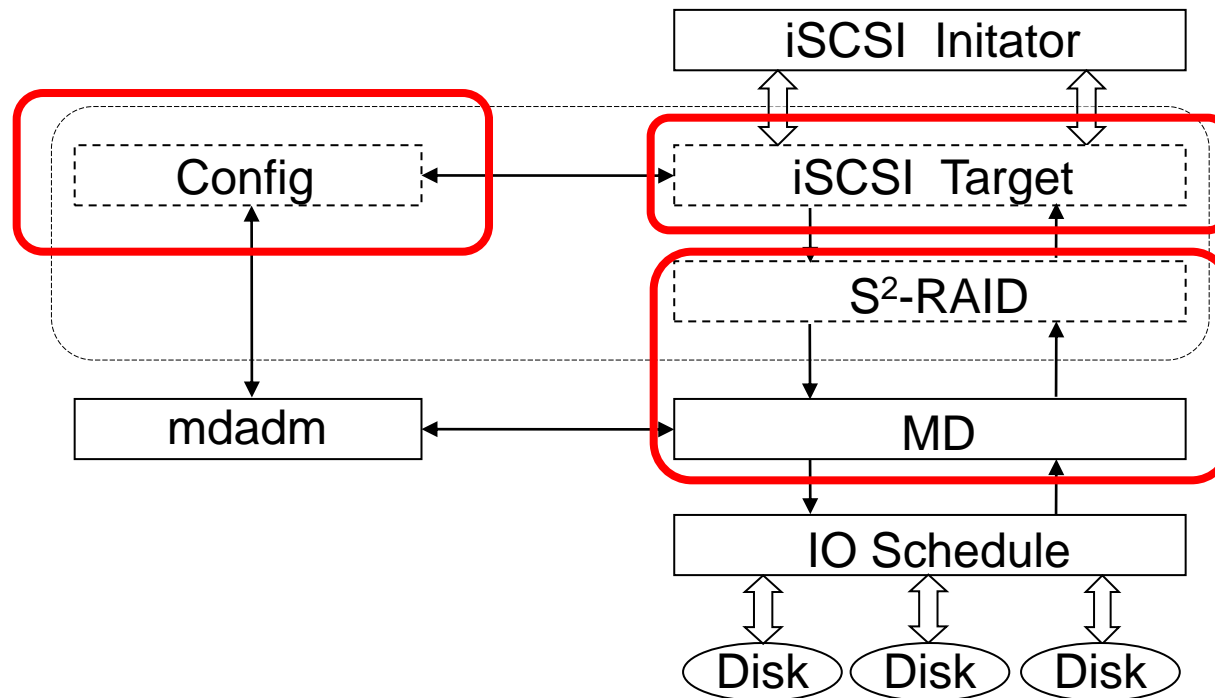
No bottleneck in reconstruction

No operation conflict(write or read)

# S<sup>2</sup>-RAID 5 reconstruction



# S<sup>2</sup>-RAID prototype structure



- S<sup>2</sup>-RAID prototype based on MD, are using the open source
- The *iSCSI target* module modifies the IET SCSI command handling and disk IO parts.
- The *Config* module provides RAID setup and configuration functions using mdadm commands to realize different S<sup>2</sup>-RAID subRAID functions.
- The *S<sup>2</sup>-RAID* module realizes the basic functions of RAID10 and RAID5 including RAID rebuilder based on MD.

# Experimental Setup

- Hardware of server and client
- Evaluation tools of the storage server and client

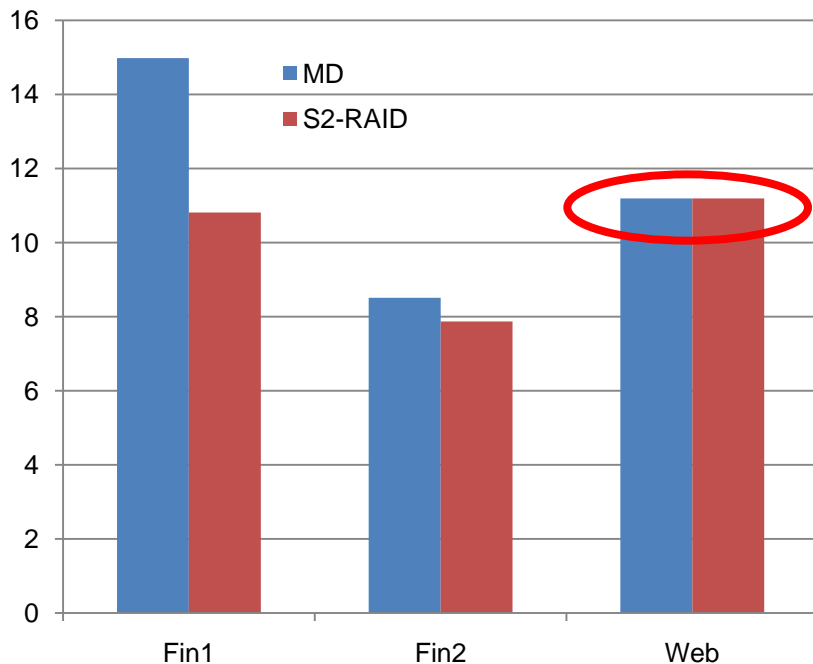
OS	Fedora Core 8.0
blktrace	blktrace 1.0
postmark	1 Spinnaker 16013AS, 160GB, 7200RPM.
disks	12 ST3080820002308GB6000RPM.
TPC-C	12 ST3080820002308GB6000RPM.
Disks	12 ST3080820002308GB6000RPM.
postgresql	postgresql 8.1.19
gnuplot	gnuplot 4.2.5
mainboard	SUPER X7DVL-I
TPC-W	TPC-W 1.5
	GA-945GCMX-S2
Jdk	jdk 1.5.0_06
CPU	Intel(R) Celeron(R) CPU 2.80GHz
Tomcat	tomcat 5.5
	Intel(R) Xeon(R) CPU 5110 @ 1.60GHz
Mysql	mysql 5.0.45
NIC	Tigon3
	Intel® PRO/1000
iscsi-initiator	iscsi-initiator 5.1.0.0865
	IBM® 6210
HBA	Highpoint 2240 RAID,

edinet r

# S<sup>2</sup>-RAID 5 reconstruction performance

■ Two evaluation parameters

- Average User Response Time
- Reconstruction Time

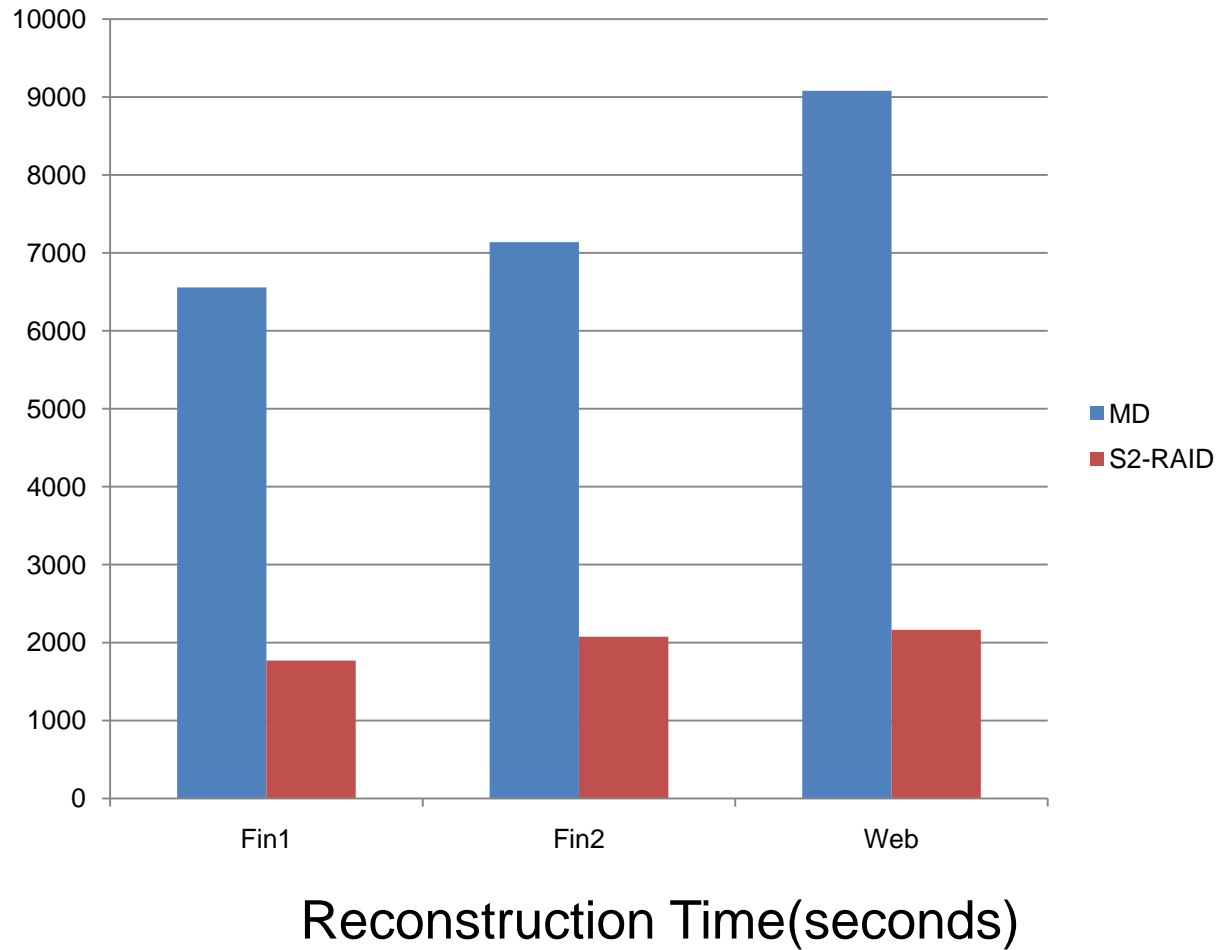


Average User Response Time(ms)

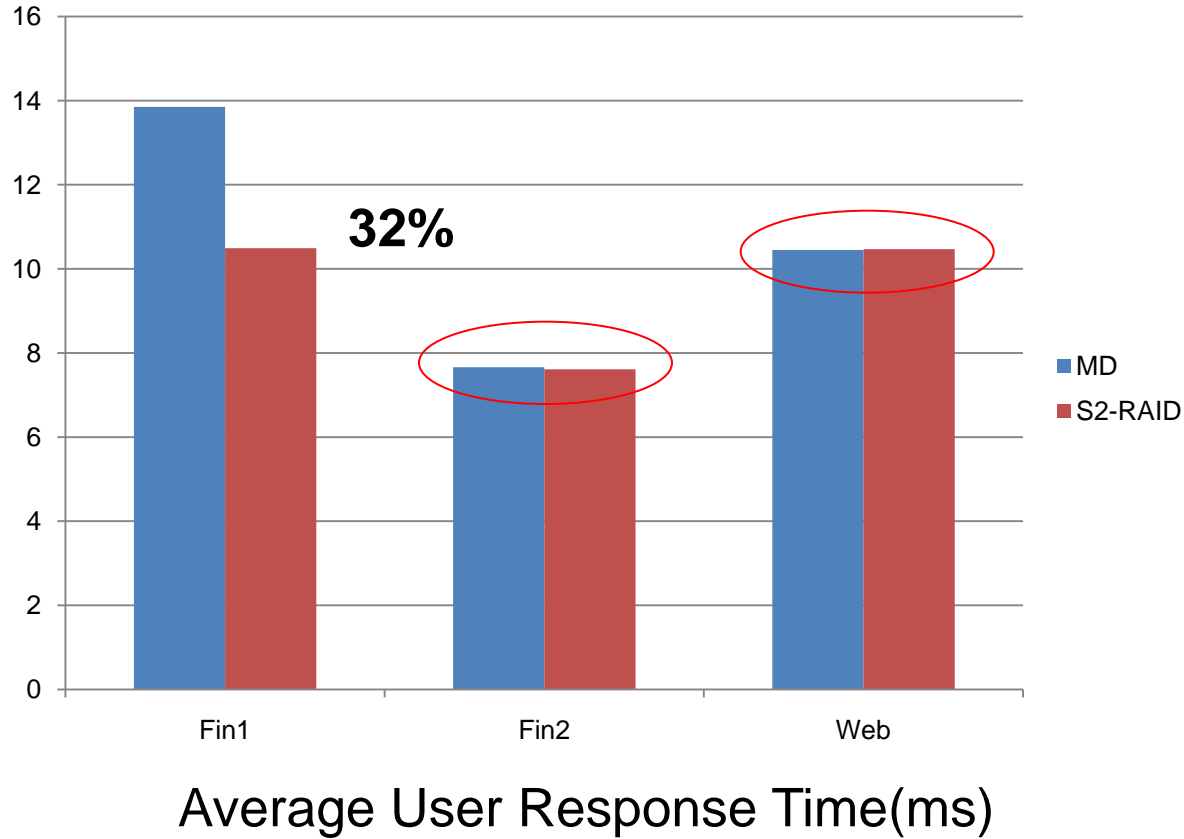
<i>Trace File</i>	<i>Write Ratio</i>	<i>Ave Req Size: KB</i>	<i>Total Req</i>
<i>Financial-1</i>	<i>76.84%</i>	<i>3.38</i>	<i>5,334,987</i>
<i>Financial-2</i>	<i>17.65%</i>	<i>2.39</i>	<i>3,699,195</i>
<i>Websearch</i>	<i>0%</i>	<i>15.07</i>	<i>4,579,809</i>

SPC trace characteristics

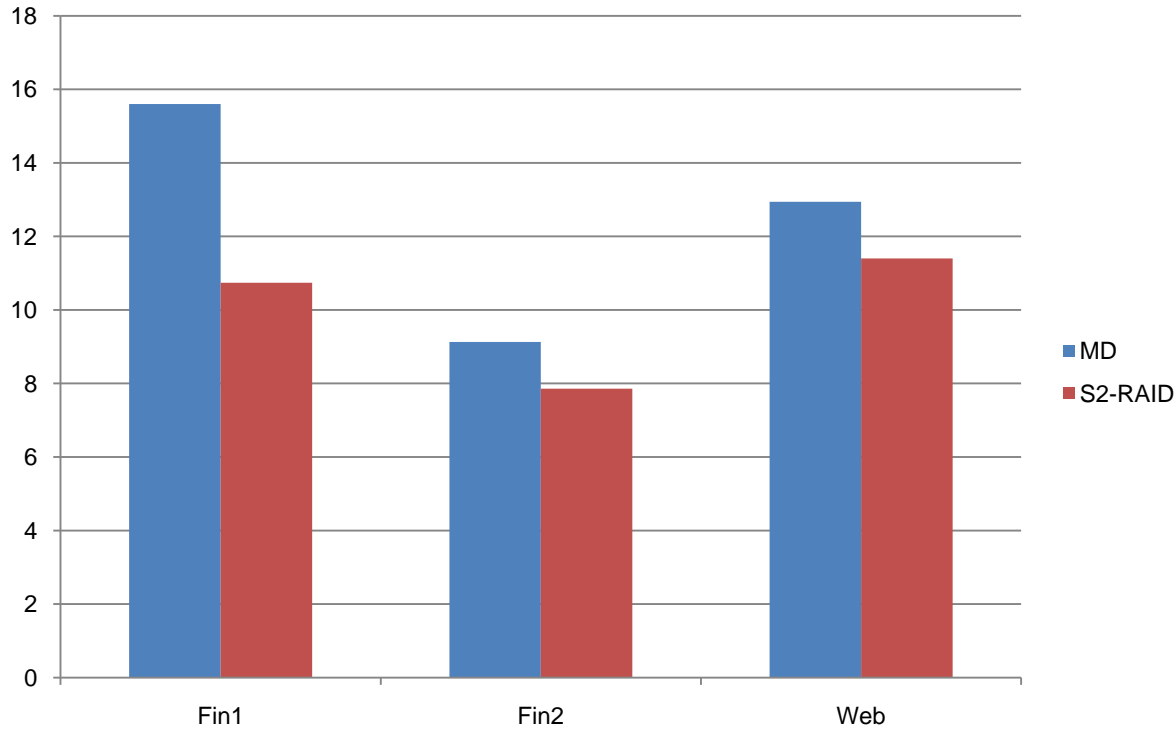
# S<sup>2</sup>-RAID 5 reconstruction performance



# S<sup>2</sup>-RAID 5 Normal Performance

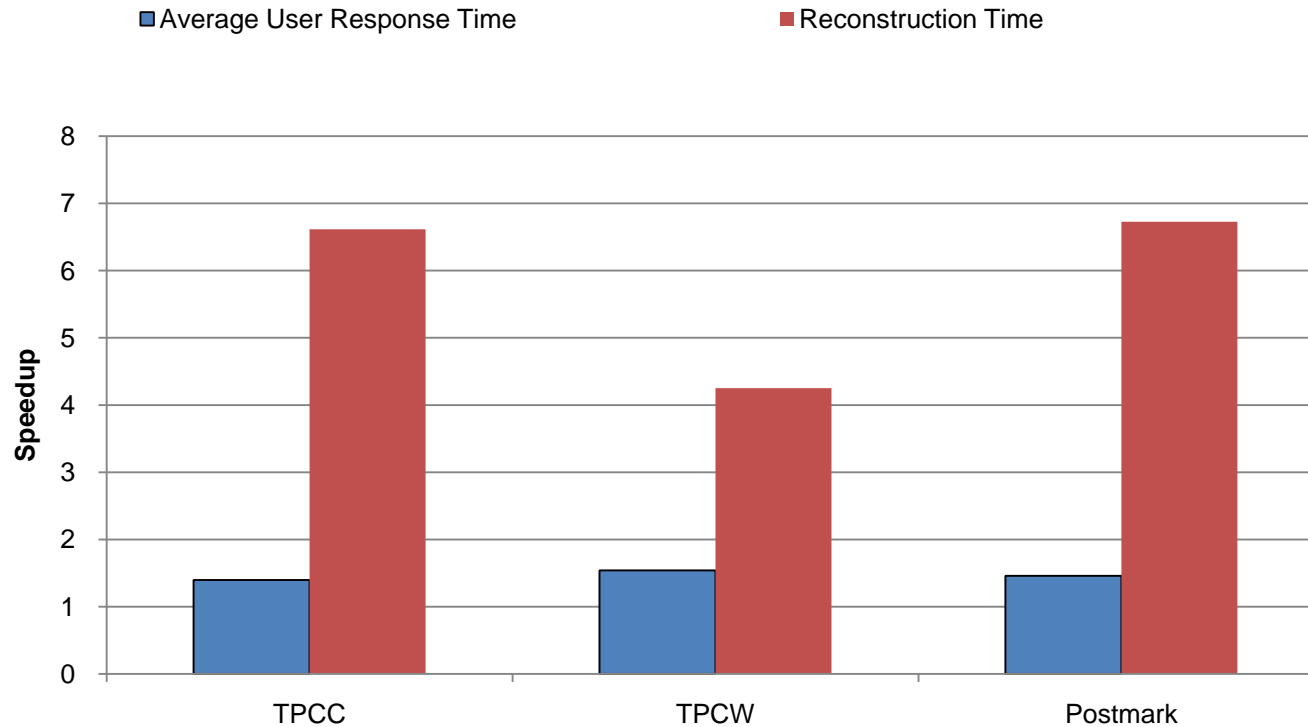


# S<sup>2</sup>-RAID 5 Degraded Performance



Average User Response Time(ms)

# Other Benchmark Performance (MD vs S<sup>2</sup>-RAID)



TPCC:20 warehouses with 10 terminals per warehouse interval of 120 minutes

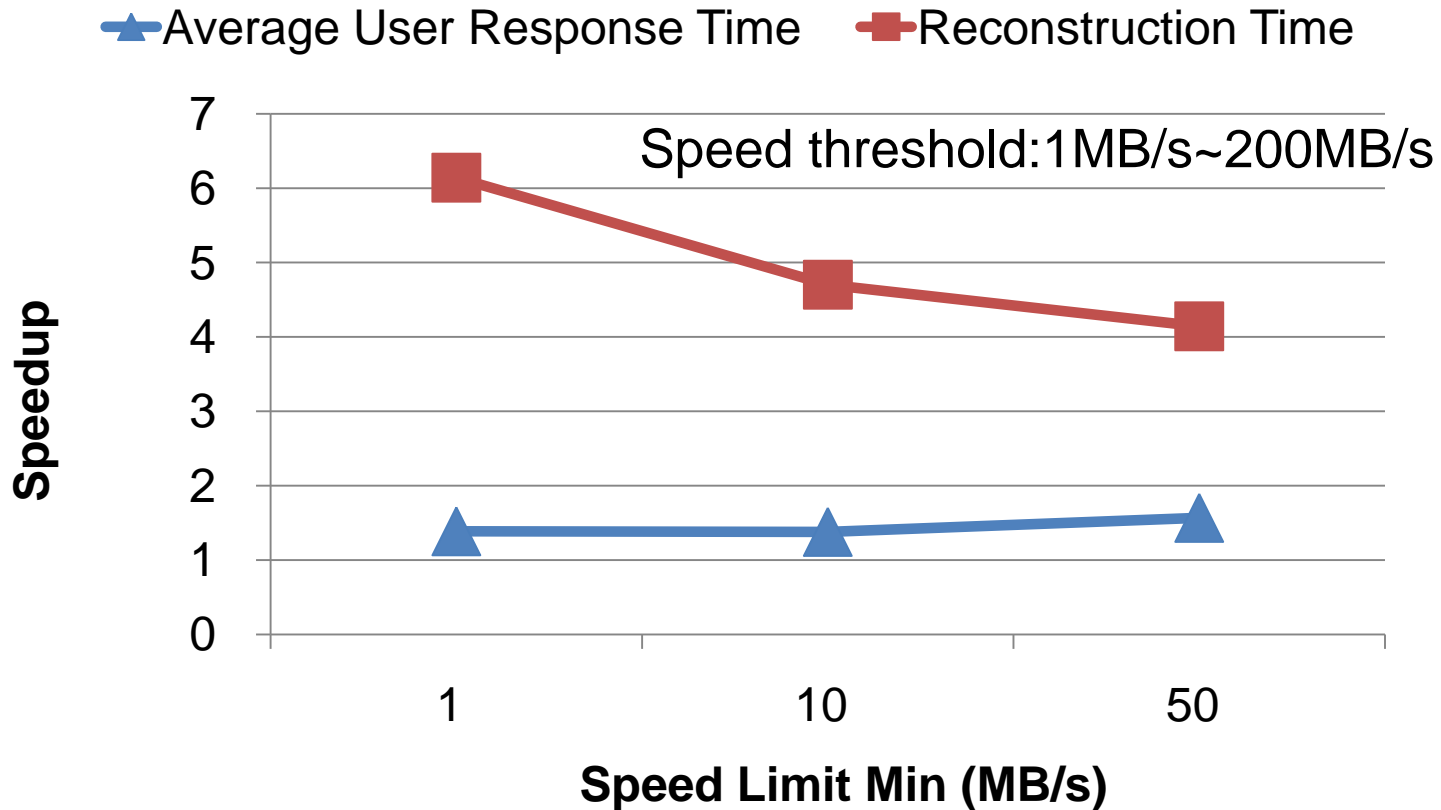
TPCW:150 emulated browsers

Postmark:20,000 files of size 4KB to 500KB and to perform 100,000 transactions

# Sensitivity Parameters for Reconstruction

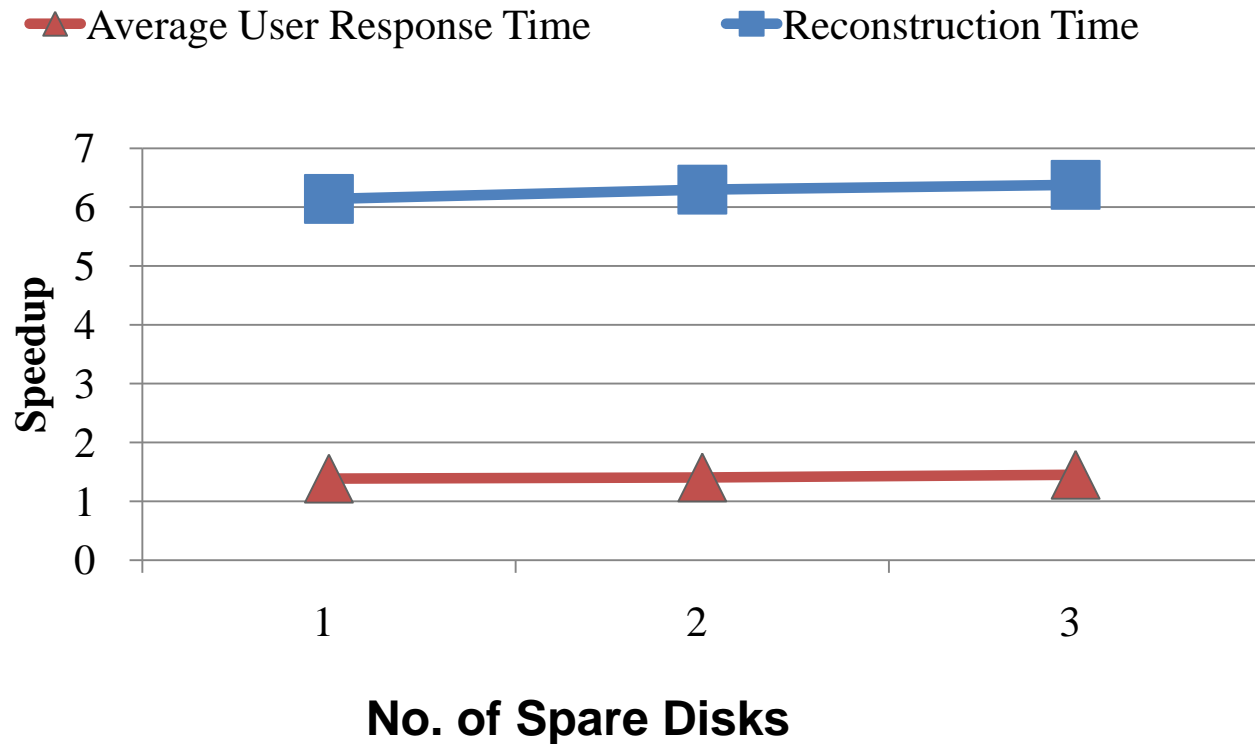
- Some sensitivity parameters
  - Reconstruction speed bandwidth
  - I/O request block size
  - Number of spare disk(additional disk not system disk)

# Reconstruction speed bandwidth



The result is based on Financial-1 traces

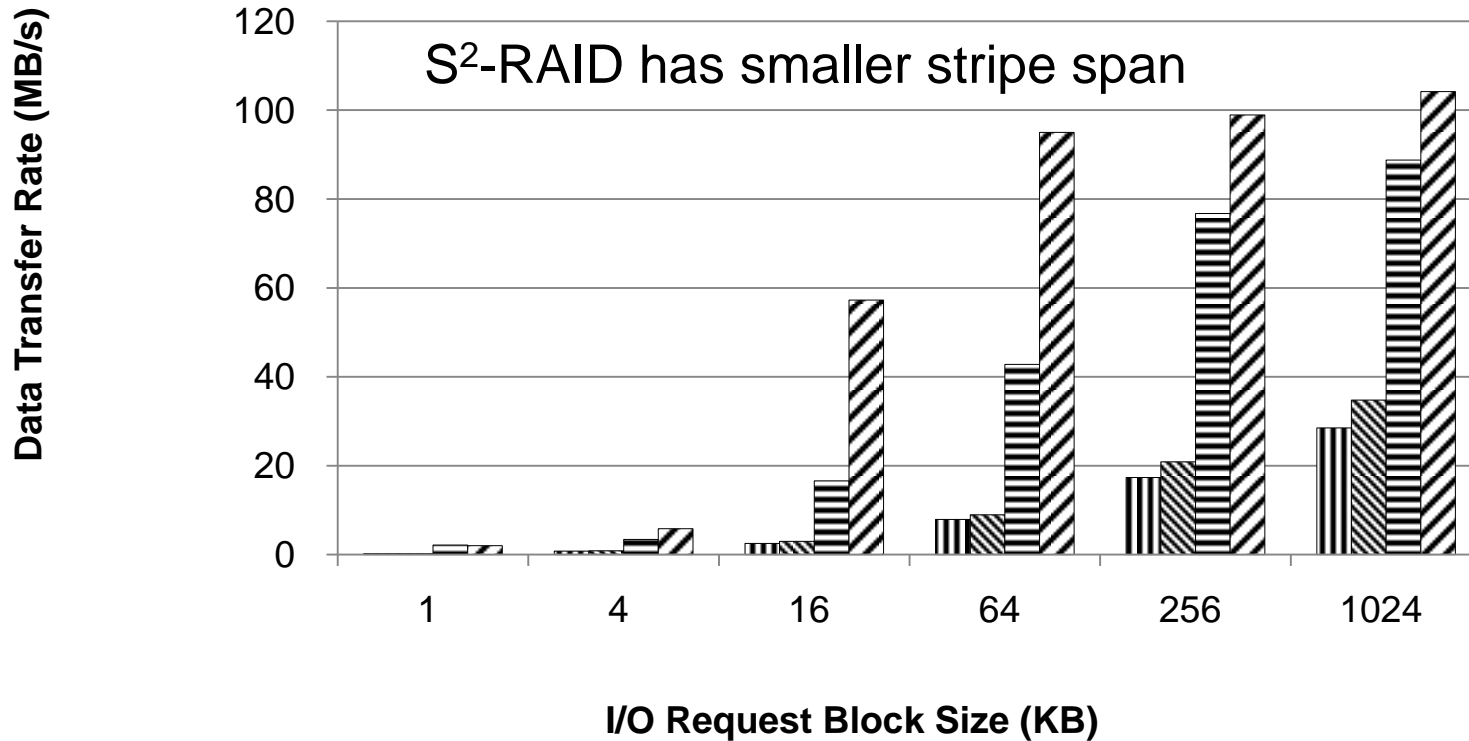
# Number of spare disk



**Speed bandwidth and No. of spare disk is insensitive to s<sup>2</sup>-RAID**

# I/O request block size

▣ RAID5-random   ▤ S2-RAID5-random   ▥ RAID5-sequence   ▦ S2-RAID5-sequence



# Conclusion

- A parallel reconstruction data layout
- Implement the s<sup>2</sup>-RAID prototype and evaluation of this structure
- S<sup>2</sup>-Raid reduces the reconstruction time greatly.
- User response time of S<sup>2</sup>-Raid is comparable to that of MD.
- Optimization?
  - Embedding existing rebuilding process (distributed sparing)-----Reduce the number of disks
  - Tolerate the mulit-disk failures.

**Thank you for your attention!**

**Questions?**