S^2-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^2-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^2-RAID prototype
- Evaluation results
- Performance analyse

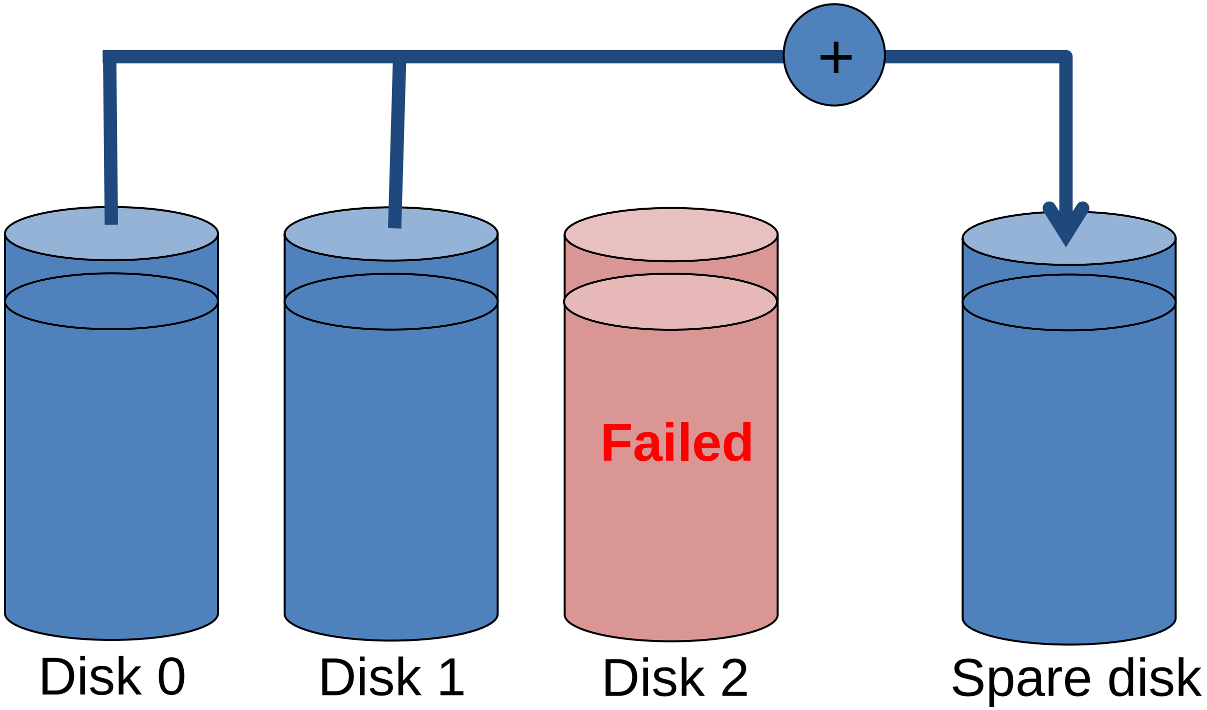
Background

- 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^2-RAID Idea

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

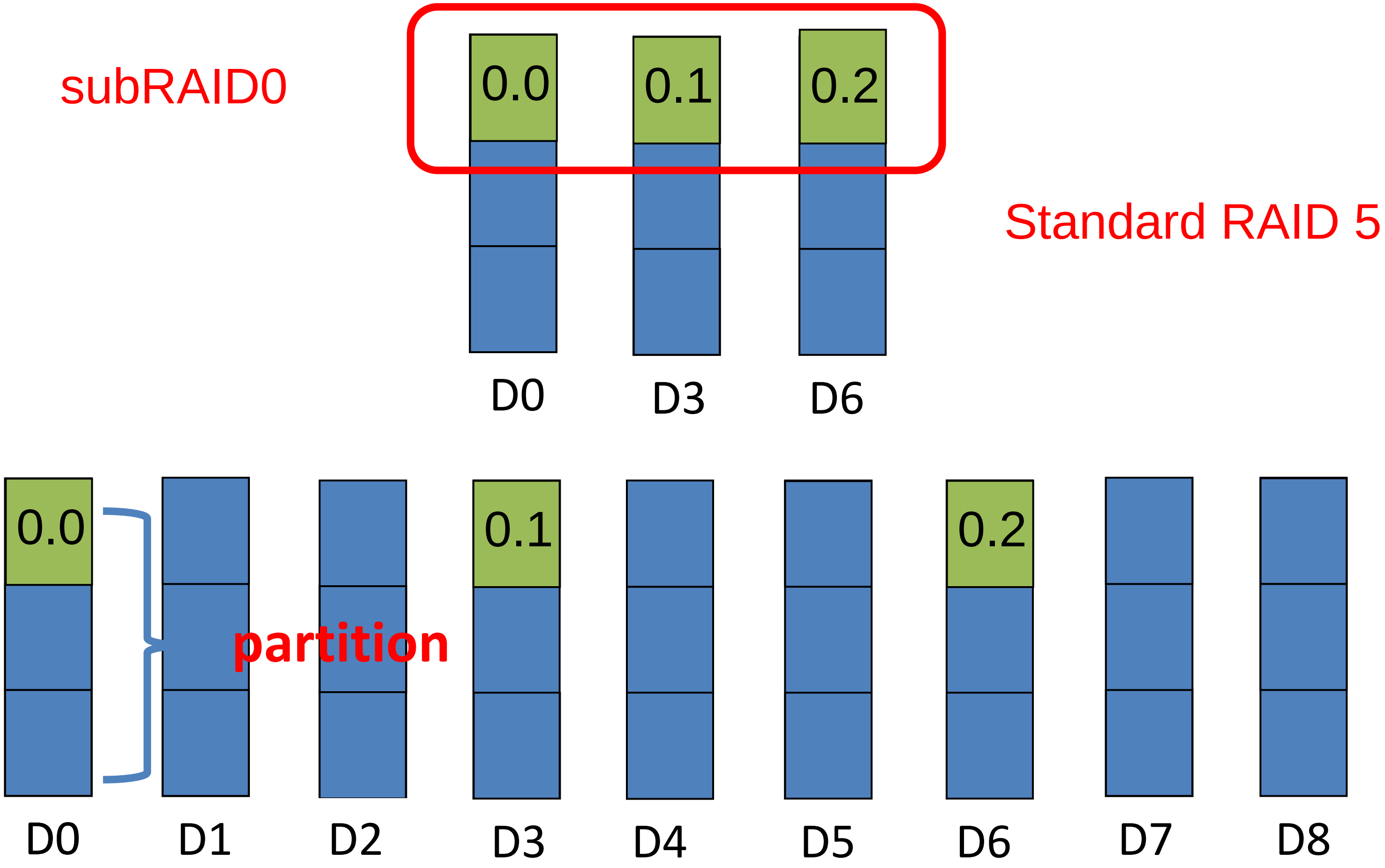
Traditional RAID 5 reconstruction



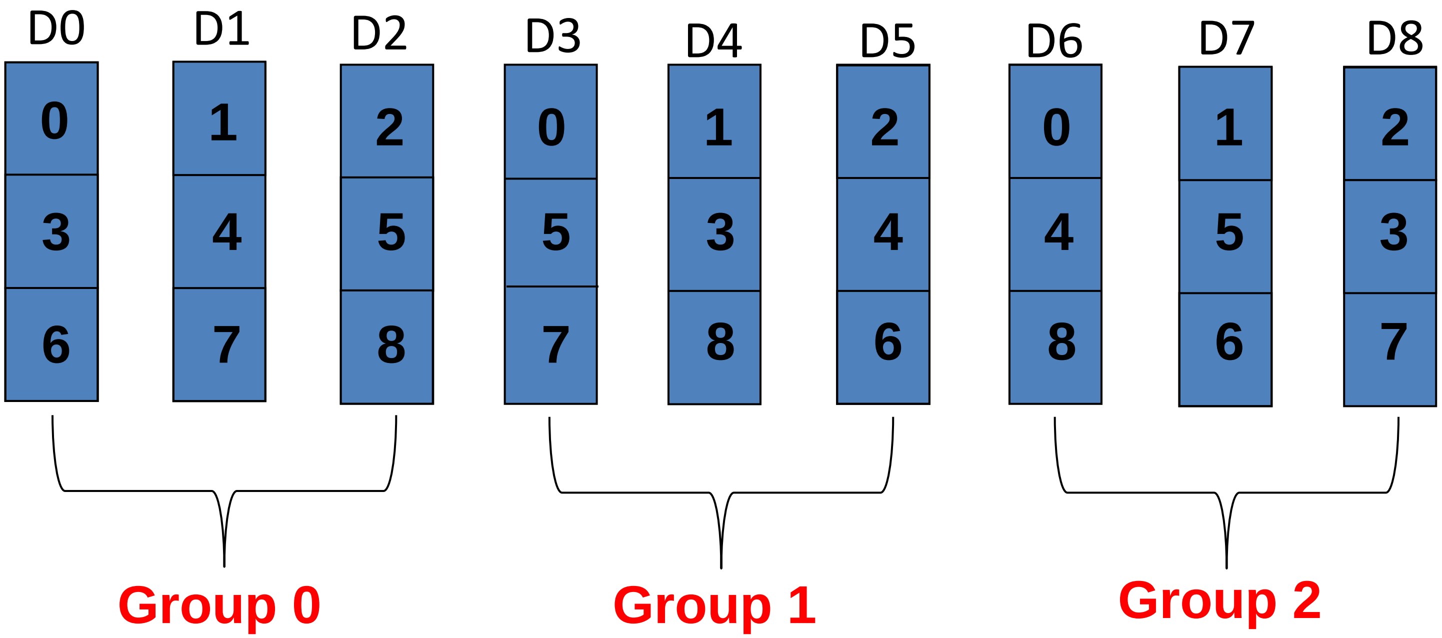
Single reconstruction stream

long reconstruction time

S^2-RAID data layout



S^2-RAID data layout structure



D0	D1	D2	D3	D4	D5	D6	D7	D8
0	1	2	0	1	2	0	1	2
3	4	5	5	3	4	4	5	3
6	7	8	7	8	6	8	6	7

$$\underset{\text{Init. matrix } M_0}{\begin{pmatrix}0&1&2\\3&4&5\\6&7&8\end{pmatrix}} \xrightarrow{\text{shift}} \underset{M_1}{\begin{pmatrix}0&1&2\\5&3&4\\7&8&6\end{pmatrix}} \xrightarrow{\text{shift}} \underset{M_2}{\begin{pmatrix}0&1&2\\4&5&3\\8&6&7\end{pmatrix}} \begin{array}{c}0\\1\\2\end{array}$$

S²-RAID data layout structure

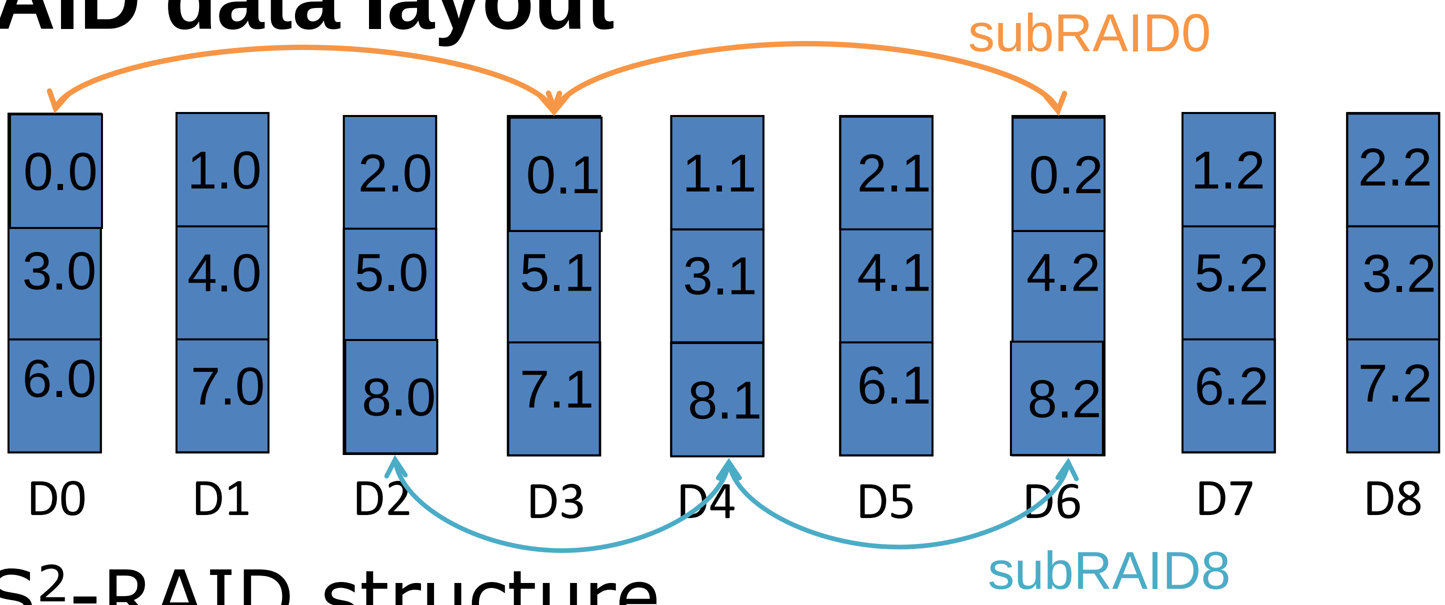
$P_{i,j}$:subRAID numbers of the ***(j+1)*** *(j+1)*th partition on disks of *(i+1)*th group in the RAID

K: the partition number of the disk

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

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

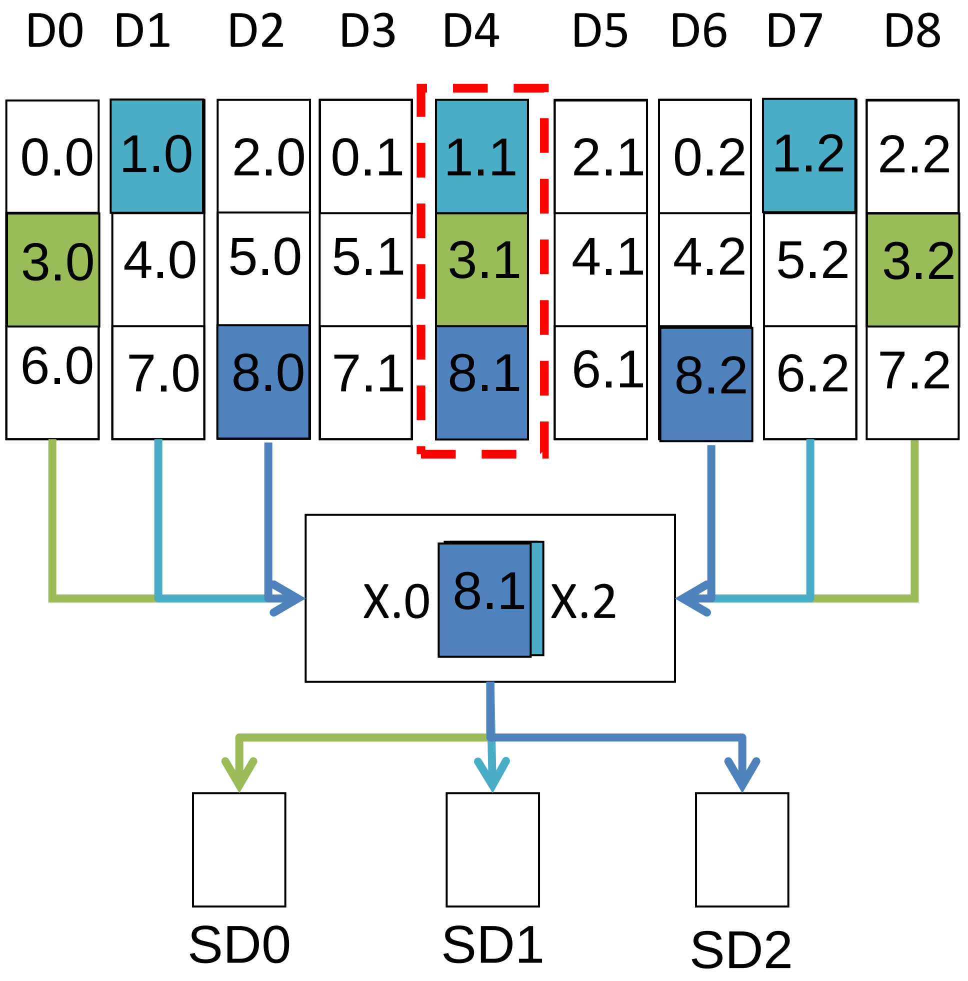
S²-RAID data layout



D0	D1	D2	D3	D4	D5	D6	D7	D8
0.0	1.0	2.0	0.1	1.1	2.1	0.2	1.2	2.2
3.0	4.0	5.0	5.1	3.1	4.1	4.2	5.2	3.2
6.0	7.0	8.0	7.1	8.1	6.1	8.2	6.2	7.2

- S²-RAID structure
  - 9 disks
  - 9 subRAIDs
  - RAID type
  - RAID 5、RAID 10、RAID 6 etc.

S^2-RAID 5 reconstruction



D0	D1	D2	D3	D4	D5	D6	D7	D8
0.0	1.0	2.0	0.1	1.1	2.1	0.2	1.2	2.2
3.0	4.0	5.0	5.1	3.1	4.1	4.2	5.2	3.2
6.0	7.0	8.0	7.1	8.1	6.1	8.2	6.2	7.2

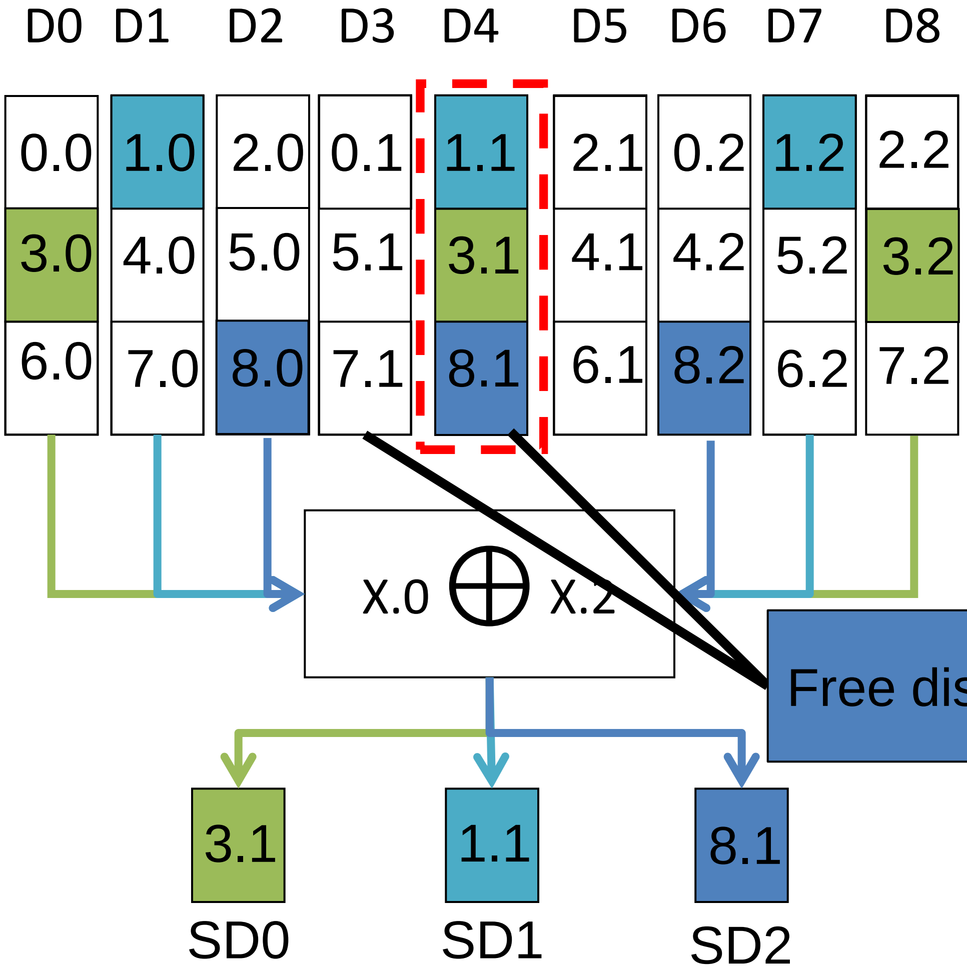
D4 was divided into 3 partitions

Reconstruction speed!

No bottleneck in reconstruction

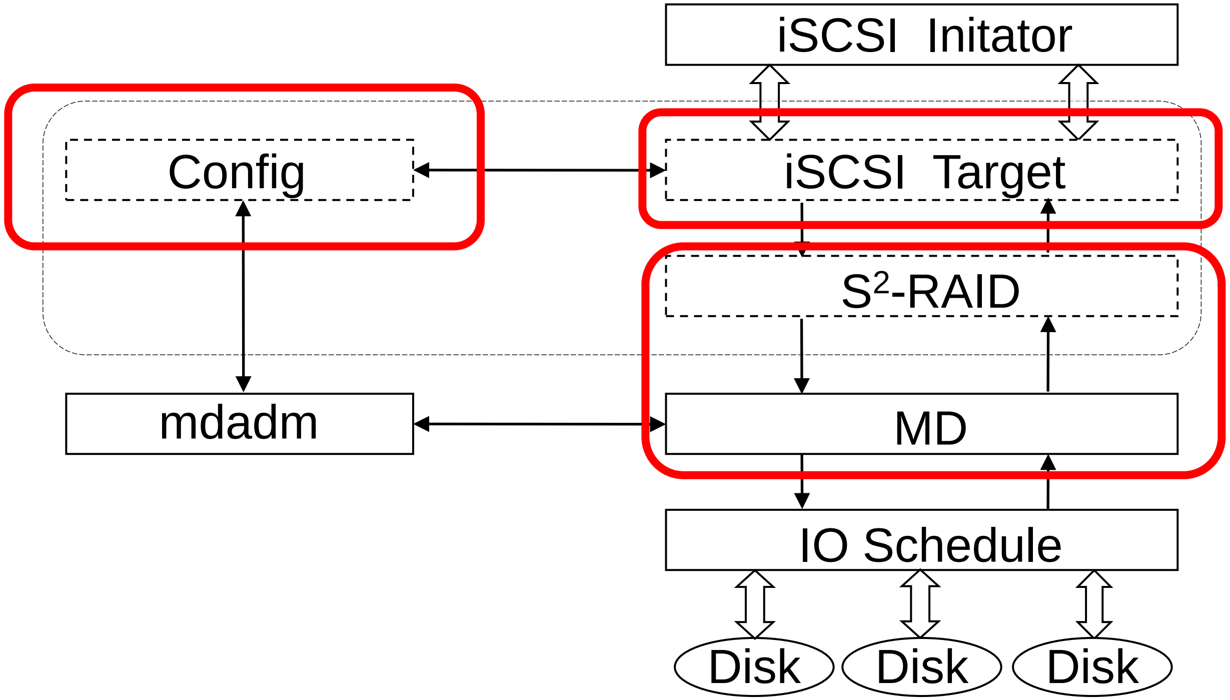
No operation conflict(write or read)

S^2-RAID 5 reconstruction



Free disk can serve front end and **S**pare disk

S²-RAID prototype structure



- S²-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 S²-RAID subRAID functions.
- The *S²-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
disks	1 Seagate ST3160023AS, 160GB, 7200RPM. 12 Seagate ST3160023AS, 160GB, 7200RPM.
mainboard	SUPER X7DVL-I
CPU	Intel(R) Xeon(R) CPU 5110 @ 1.60GHz
NIC	Intel® PRO/1000
memory	1GB DDR2
HBA	Highpoint 2240 RAID,

OS	Fedora Core 8.0
Disks	Seagate ST3160023AS, 160GB, 7200RPM
Mainboard	GA-945GCMX-S2
CPU	Intel(R) Celeron(R) CPU 2.80GHz
NIC	Tigon3
Memory	512MB DDR2

blktrace	blktrace 1.0
postmark	postmark 1.51
TPC-C	TPCC-Uva-1.2.3
postgresql	postgresql 8.1.15
gnuplot	gnuplot 4.2.5
TPC-W	TPC-W 1.5
Jdk	jdk 1.5.0.06
Tomcat	tomcat 5.5
Mysql	mysql 5.0.45
iscsi-initiator	iscsi-initiator-2.0865

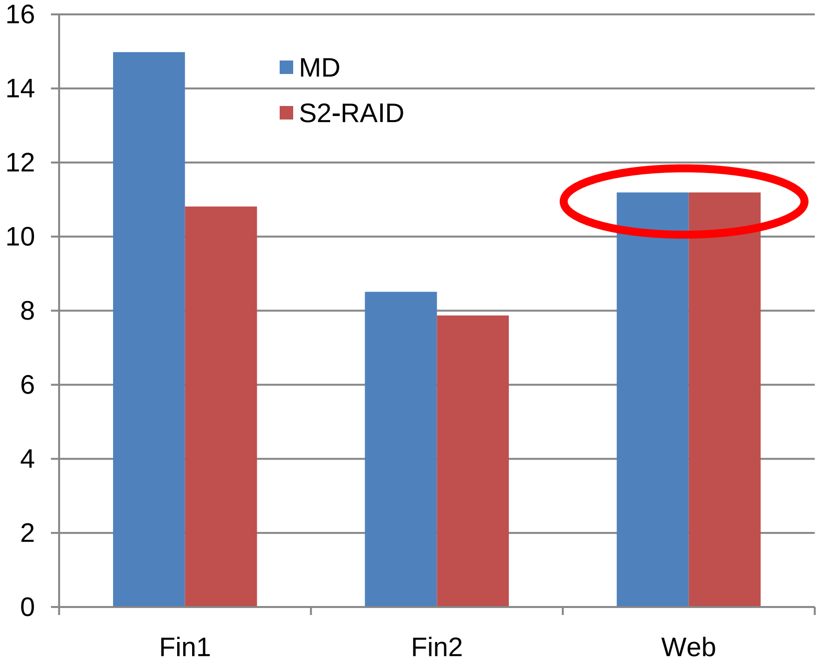
server

client

tools

S^2-RAID 5 reconstruction performance

- Two evaluation parameters
  - Average User Response Time
  - Reconstruction Time

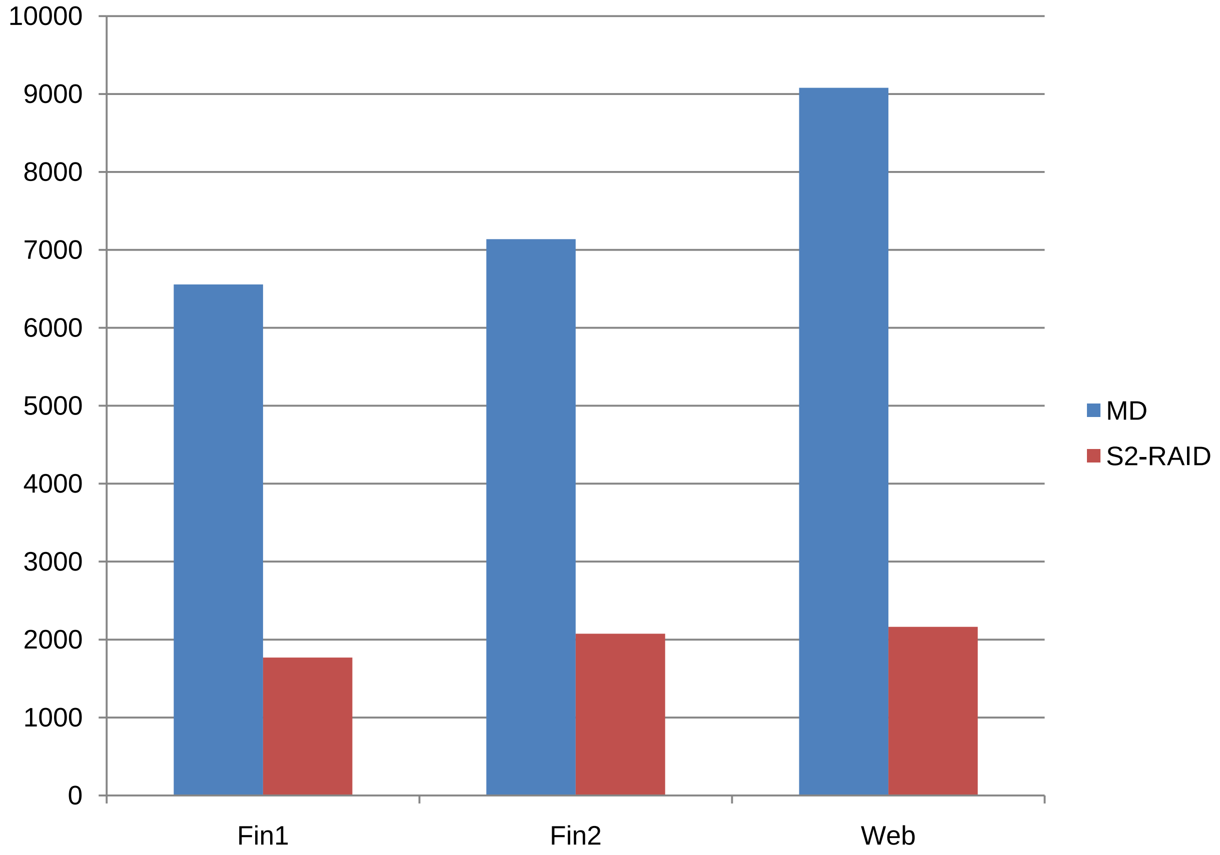


Average User Response Time(ms)

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

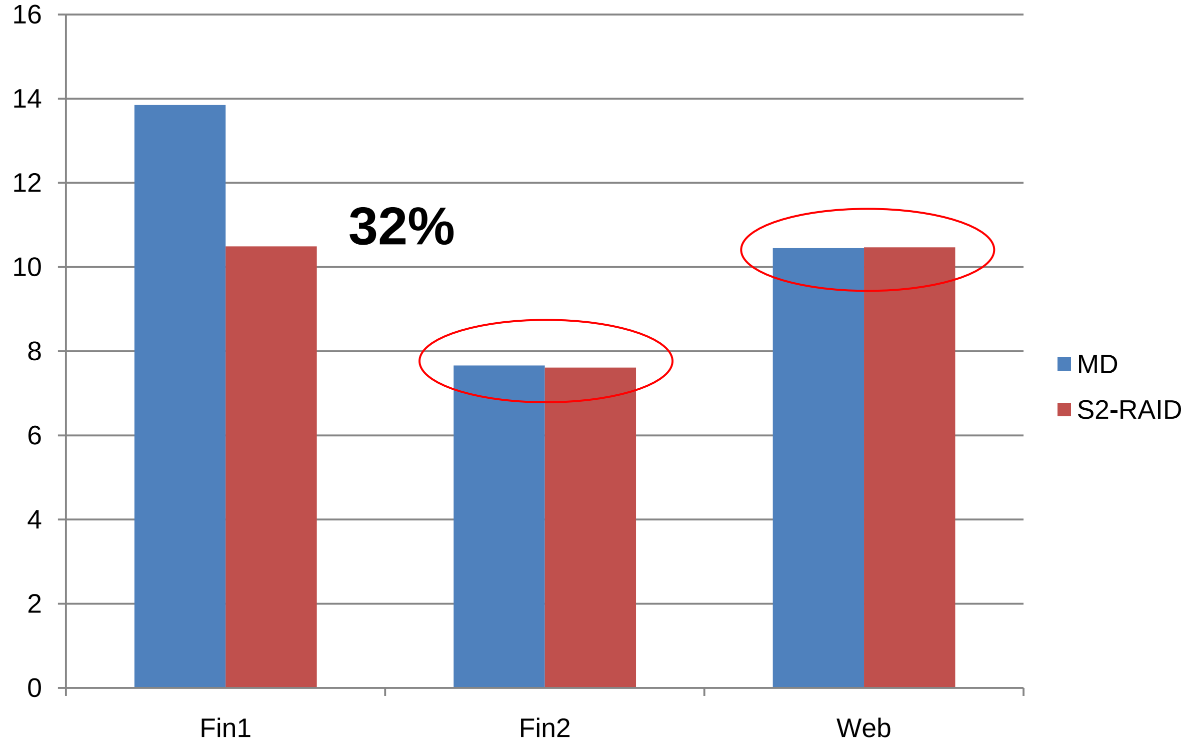
SPC trace characteristics

S^2-RAID 5 reconstruction performance



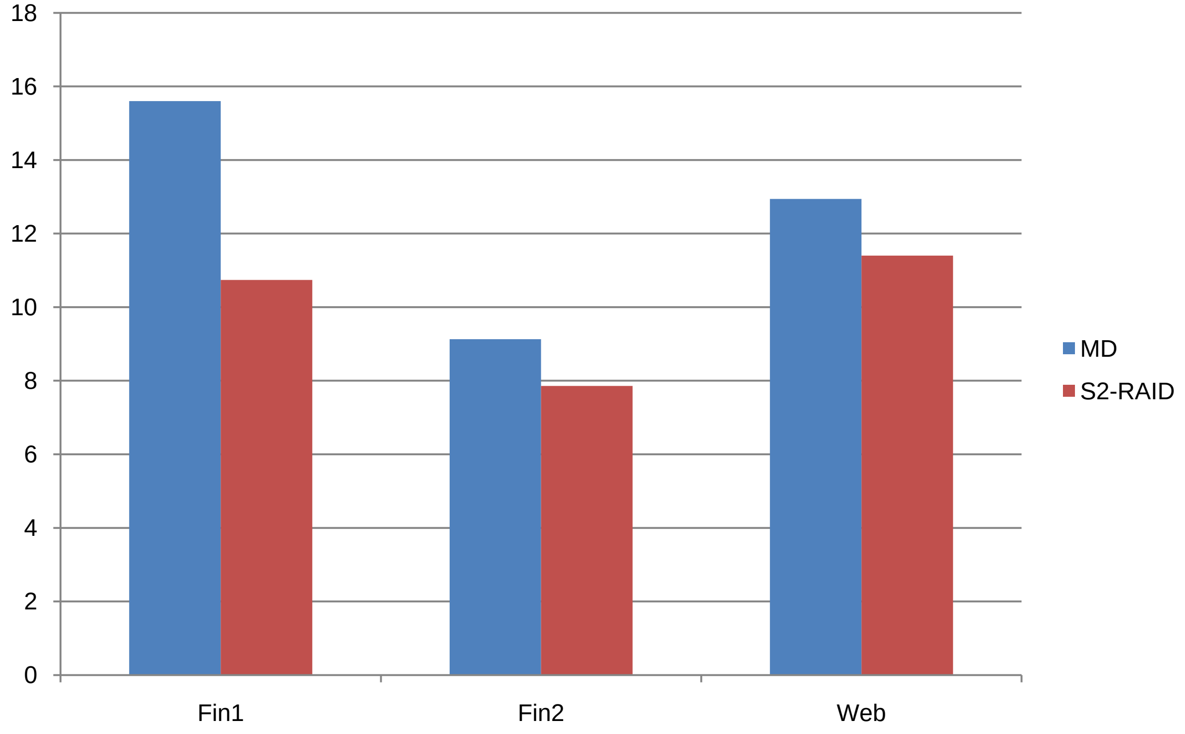
Reconstruction Time(seconds)

S^2-RAID 5 Normal Performance



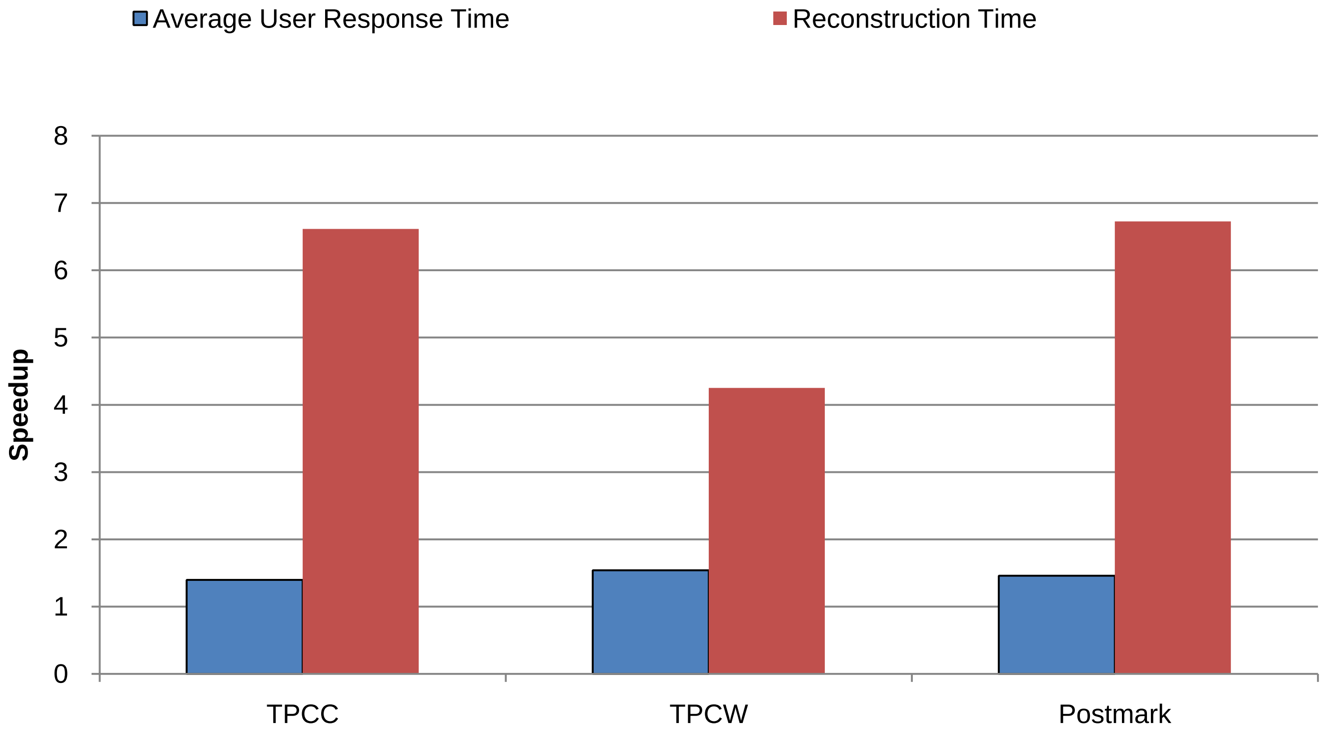
Average User Response Time(ms)

S^2-RAID 5 Degraded Performance



Average User Response Time(ms)

Other Benchmark Performance (MD vs S^2-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

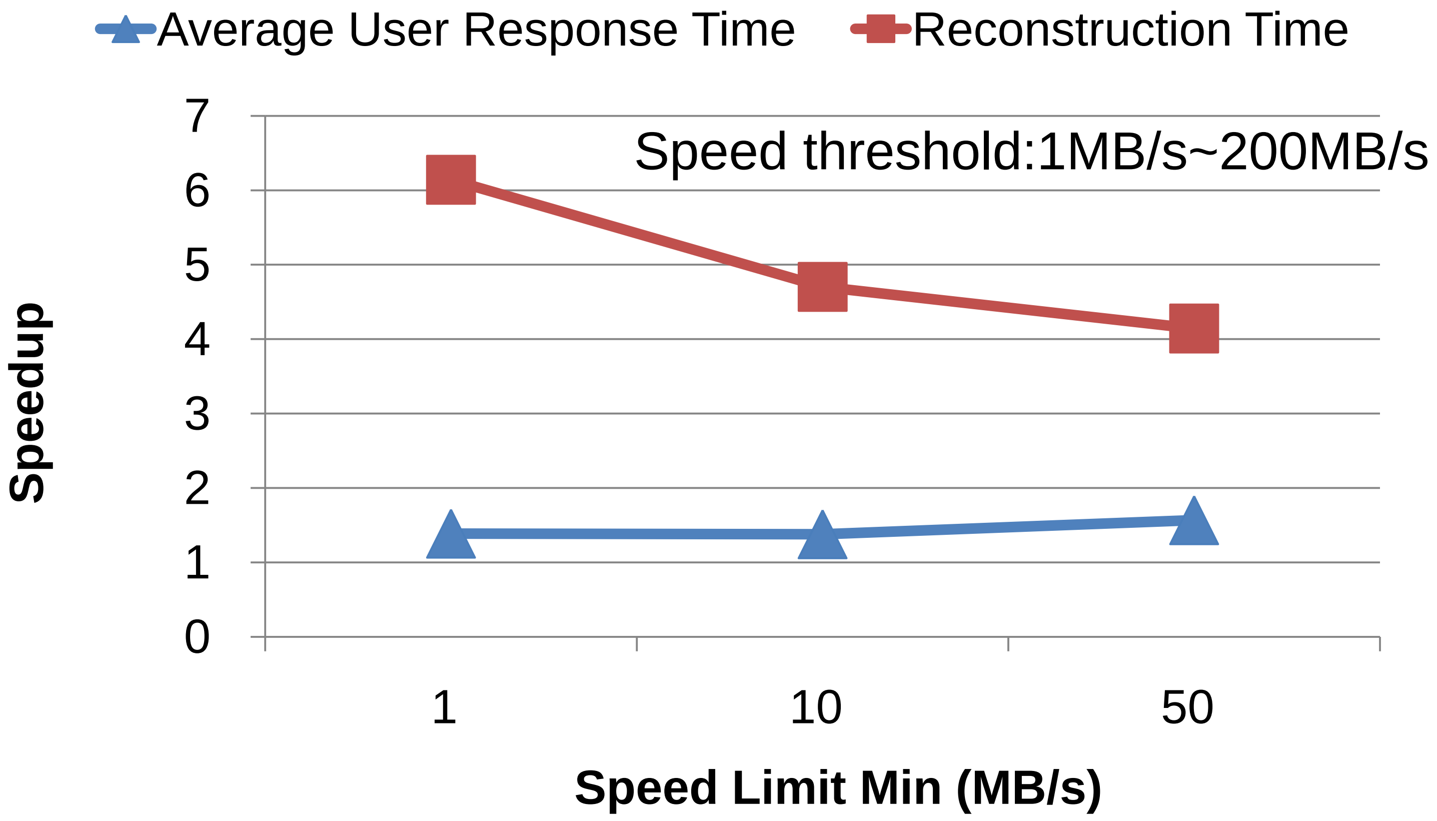
- Some sensitivity parameters

  Reconstruction speed bandwidth

  I/O request block size

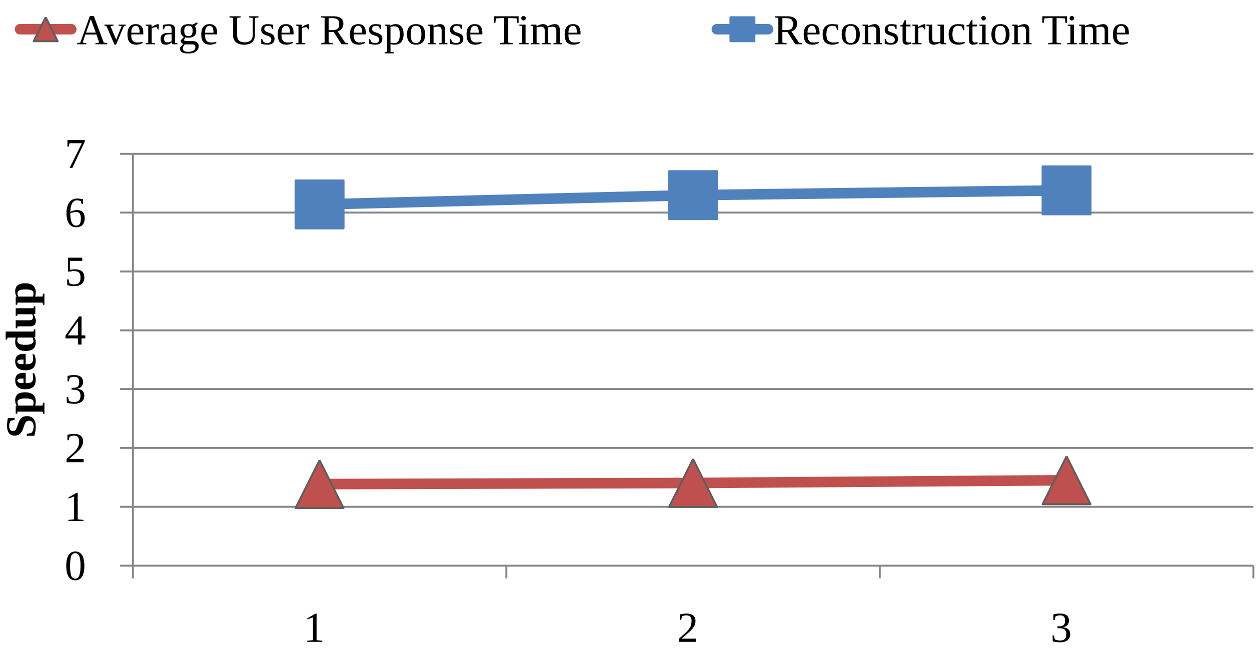
  Number of spare disk(additional disk not system disk)

Reconstruction speed bandwidth



The result is based on Financial-1 traces

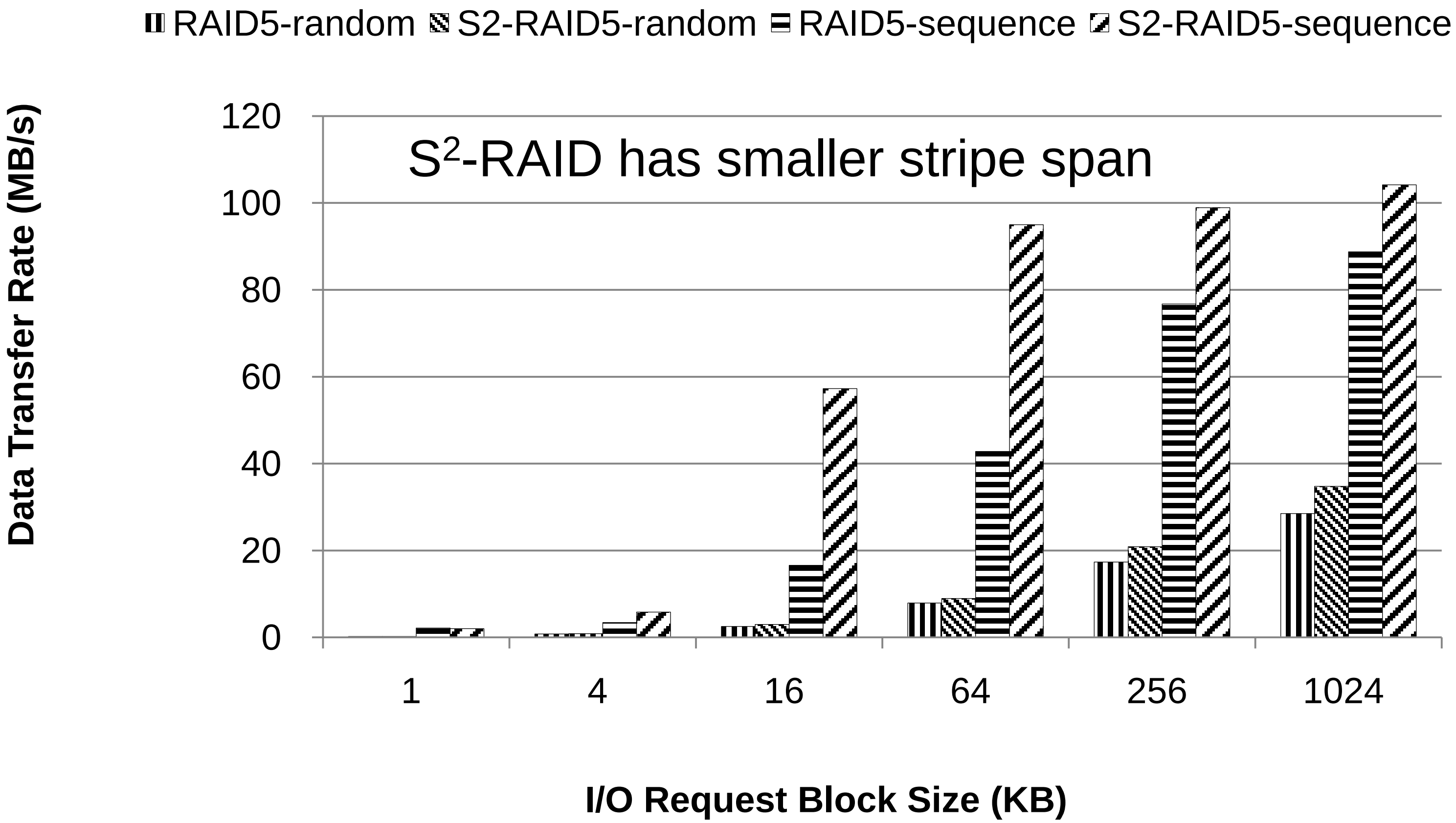
Number of spare disk



No. of Spare Disks

Speed bandwidth and No. of spare disk is insensitive to s^2-RAID

I/O request block size



Conclusion

- A parallel reconstruction data layout
- Implement the s^2-RAID prototype and evaluation of this structure

- S^2-Raid reduces the reconstruction time greatly.
- User response time of S^2-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?