



Whamcloud

T10PI End-to-End Data Integrity Protection for Lustre

Shuichi Ihara, Li Xi



Why data integrity is important?

▶ Data corruptions is painful!

- Frequency is low, but cost is very high
- A lot of unusual operations and step by step procedures to recover

▶ What causes data corruptions?

- Facility
- Hardware include network
- Software
- Human errors

Types of data corruption

▶ Two types of data corruption

- Latent sector/block errors
 - Application can't read sector/block and return –EIO
- Silent data corruption
 - Application can read sector/block, but it's NOT expected data and NOT valid data

▶ Silent data corruption causes another corruptions

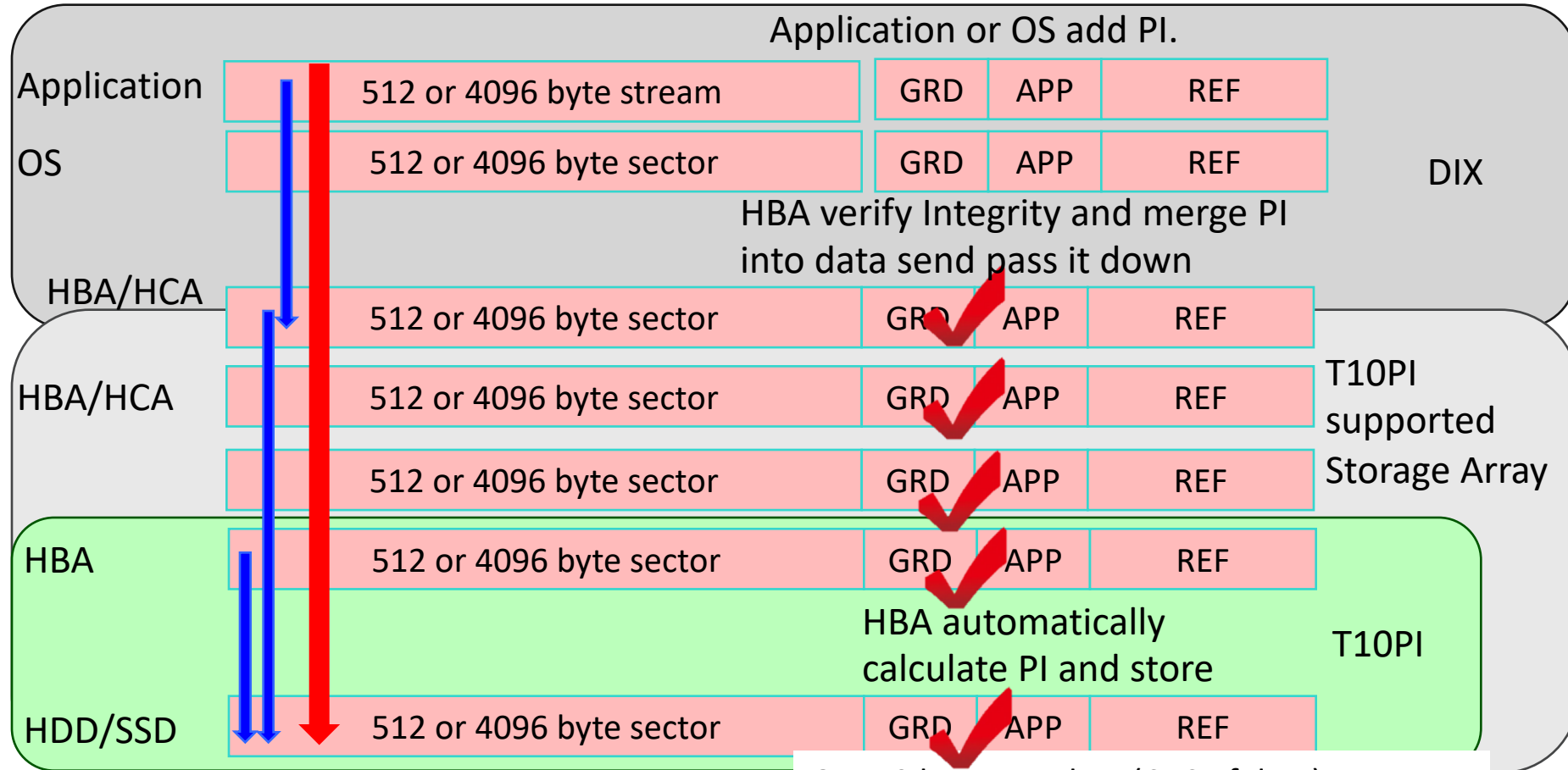
- Application read data as expected and write new data based on it, but it's wrong!

▶ Where/Why this happens?

- All storage stack(App, OS, HBA, Storage Fabric/Array, Disk)
- Lack of integrity check, each storage stack trusts upper/lower component

T10PI(DIF) and DIX(Data Integrity Extensions)

- The standard specifies an additional 8 byte field designated for data integrity/protection for each data block.



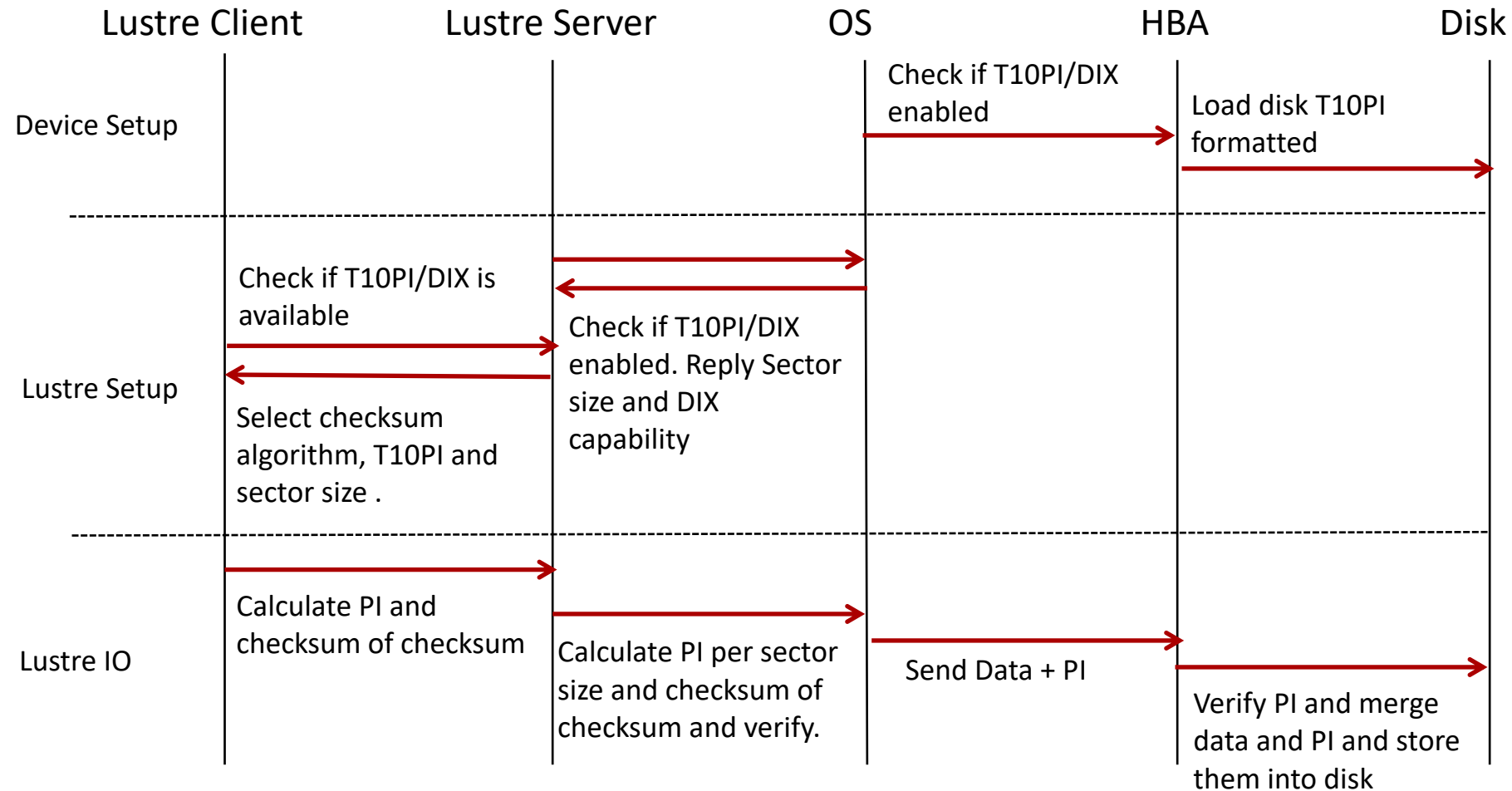
GRD: 2 byte guard tag(CRC of data)
 APP: 2 byte application tag
 REF: 4 byte reference tag

Design of Lustre End-To-End Data Integrity

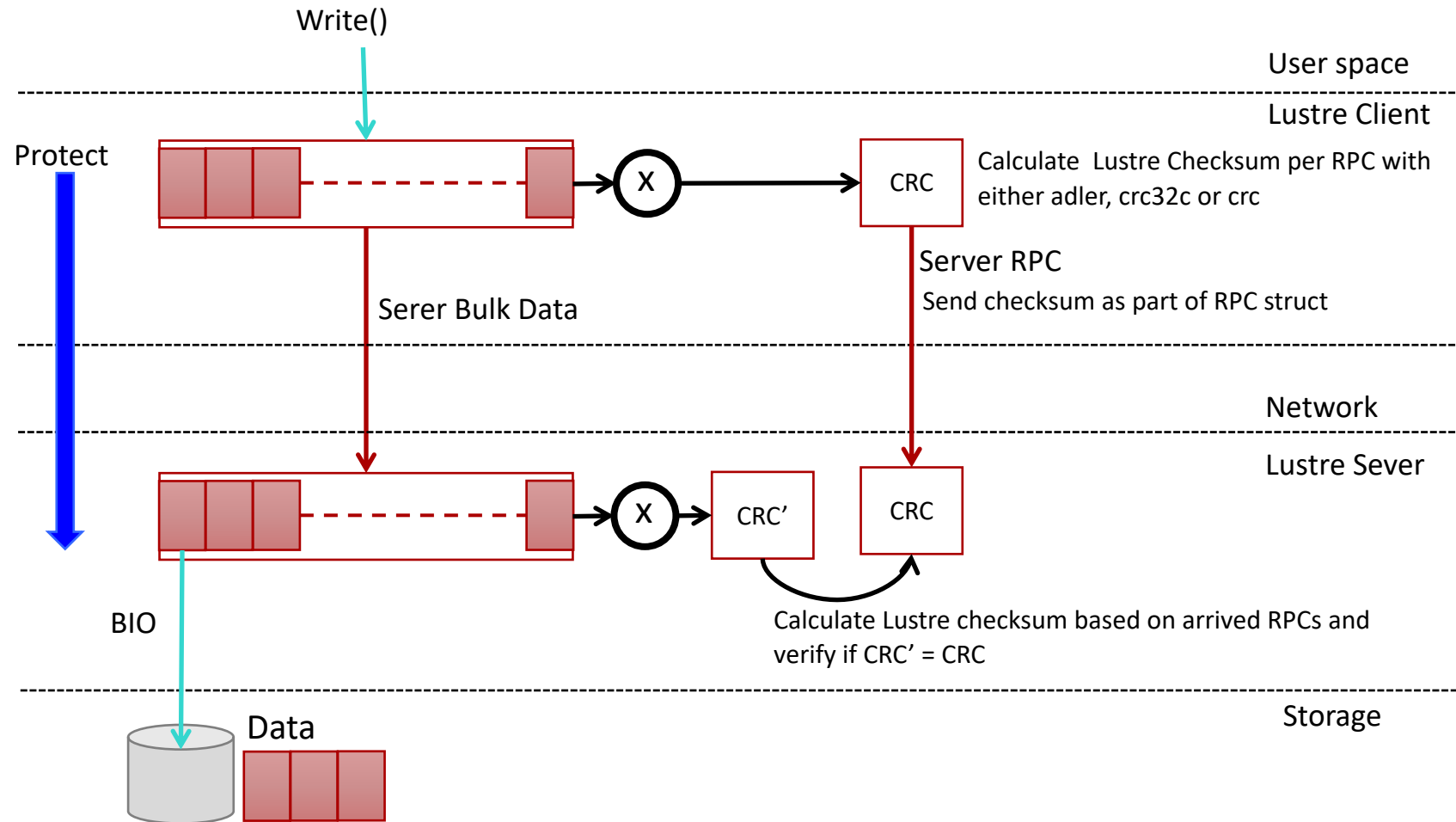


- ▶ Fully transparent End-to-End Data integrity from Lustre client to disk
- ▶ Relies on open standard format T10PI/DIX
- ▶ Any T10PI/DIX supported hardware is usable
- ▶ Consider minimum performance impacts
- ▶ Keep compatibility for old Lustre version or non-T10PI supported hardware

Flow of Lustre End-to-End Data Integrity

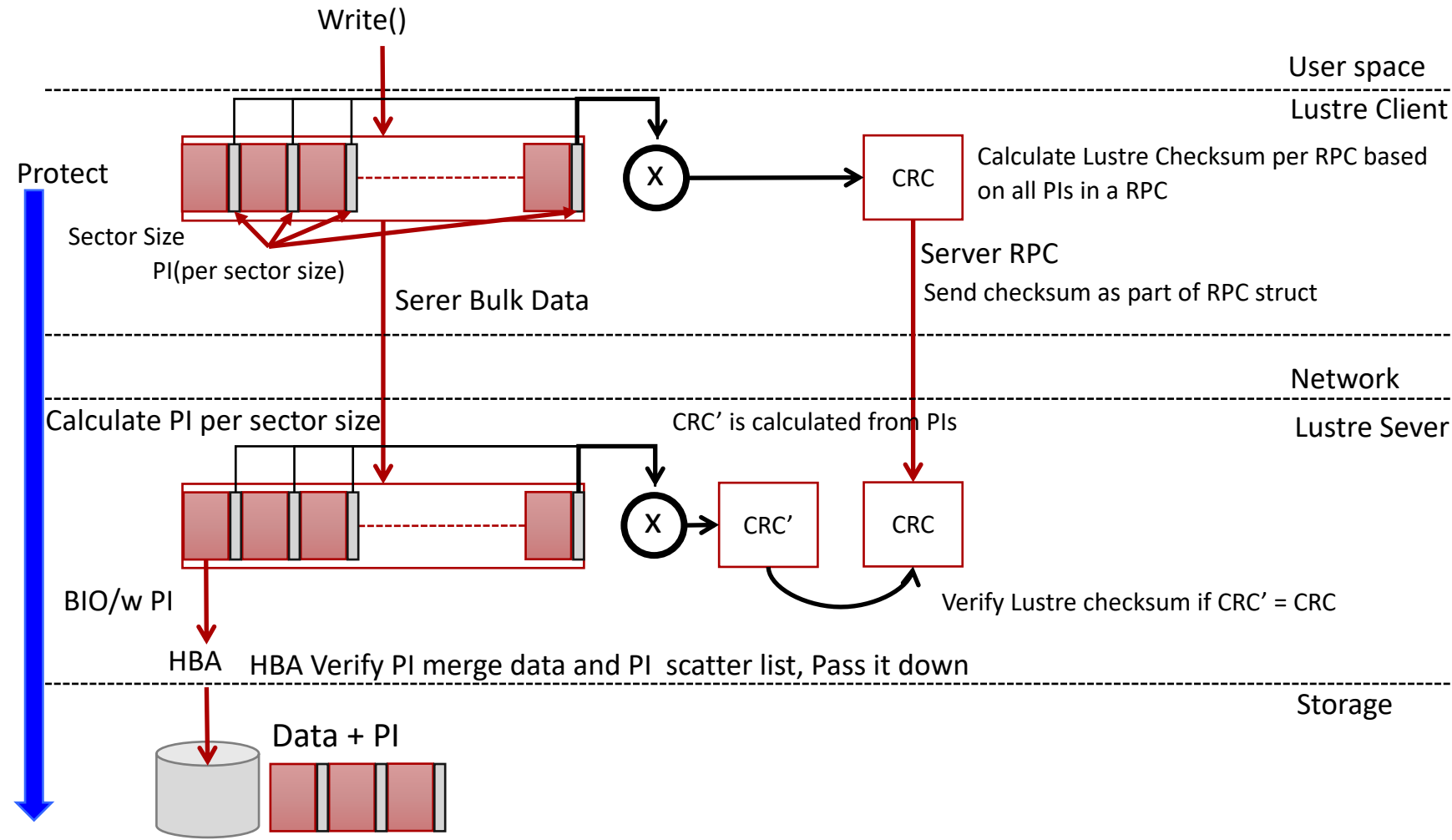


Original Lustre checksum (Write)

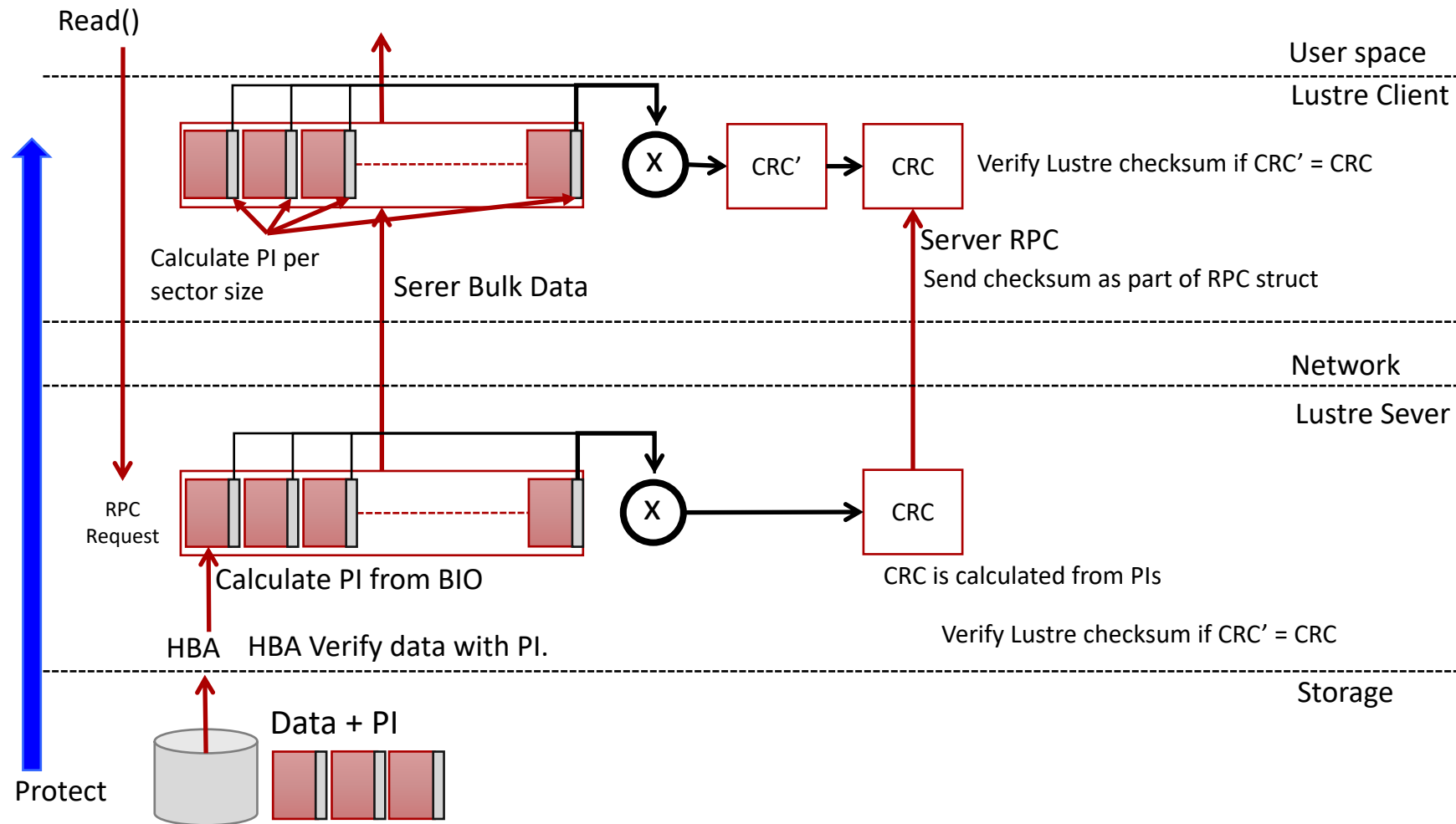




Lustre checksum with T10PI/DIX for Enabling End-to-End Data Integrity(Write)



Lustre checksum with T10PI/DIX for Enabling End-to-End Data Integrity(Read)



Status

► Task is tracked under LU-10472

- Patch being to submit for review
 - T10PI support for BIO (<https://review.whamcloud.com/#/c/30792>): merged ✓
 - T10PI support for Lustre checksum (<https://review.whamcloud.com/#/c/30980>): merged ✓
 - T10PI between RPC and BIO (<https://review.whamcloud.com/#/c/32266>): very close to be merged ✓

► Found T10PI related bugs for RHEL7.x kernel

- Linux-commit: 63573e359d052e506d305c263576499f06355985
- Linux-commit: c611529e7cd3465ec0eada0f44200e8420c38908

► Function test and benchmark started

- Adding test codes
- Fault injection
- Performance compaction against today Lustre checksum

Test Environment

▶ 1 x MDS

- 2 x E5-2640v3, 256GB Memory, 1 x EDR Infiniband
- 1 x LSI SAS3008(Enabled T10PI/DIX)

▶ 1 x OSS

- 2 x E5-2640v3, 256GB Memory, 1 x EDR Infiniband
- 1 x LSI SAS3008(Enabled T10PI/DIX)

▶ 1 x SS8462

- 8 x NL-SAS and 2 x SAS disks connected to OSS/MDS with SAS

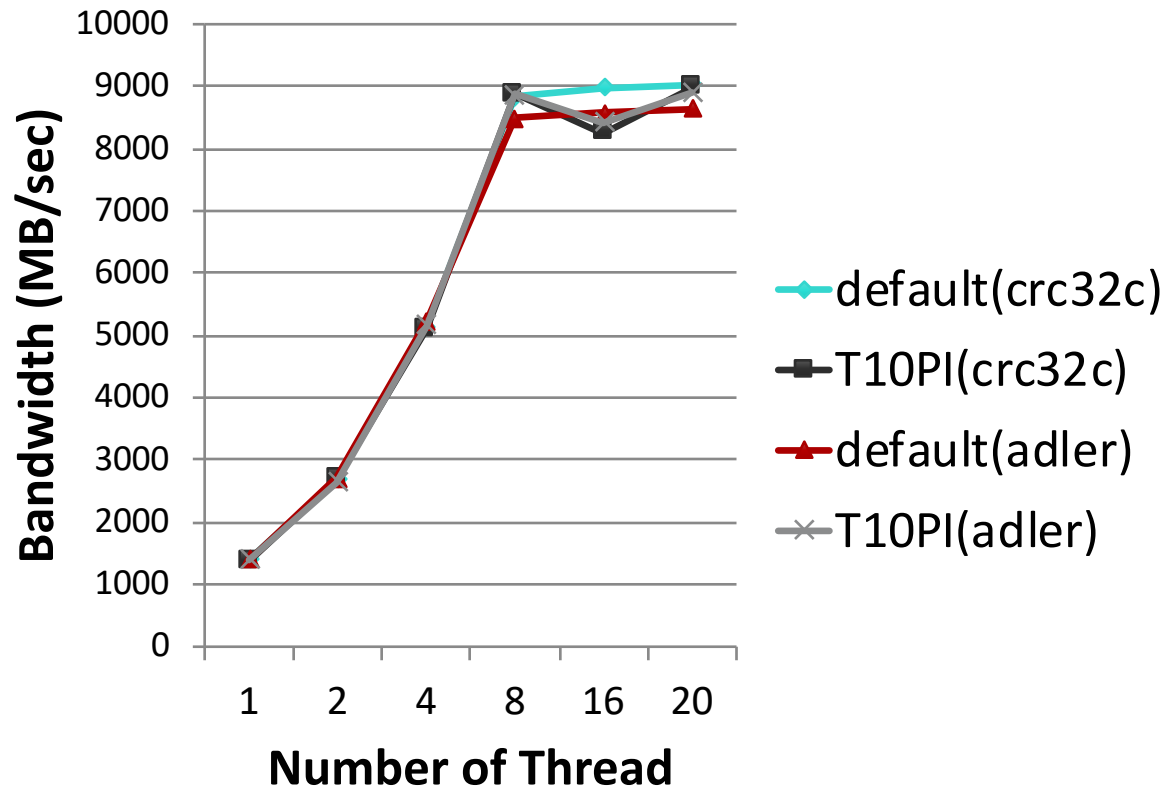
▶ 6 x Client

- 2 x E5-2660v3, 128GB Memory, 1 x EDR Infiniband

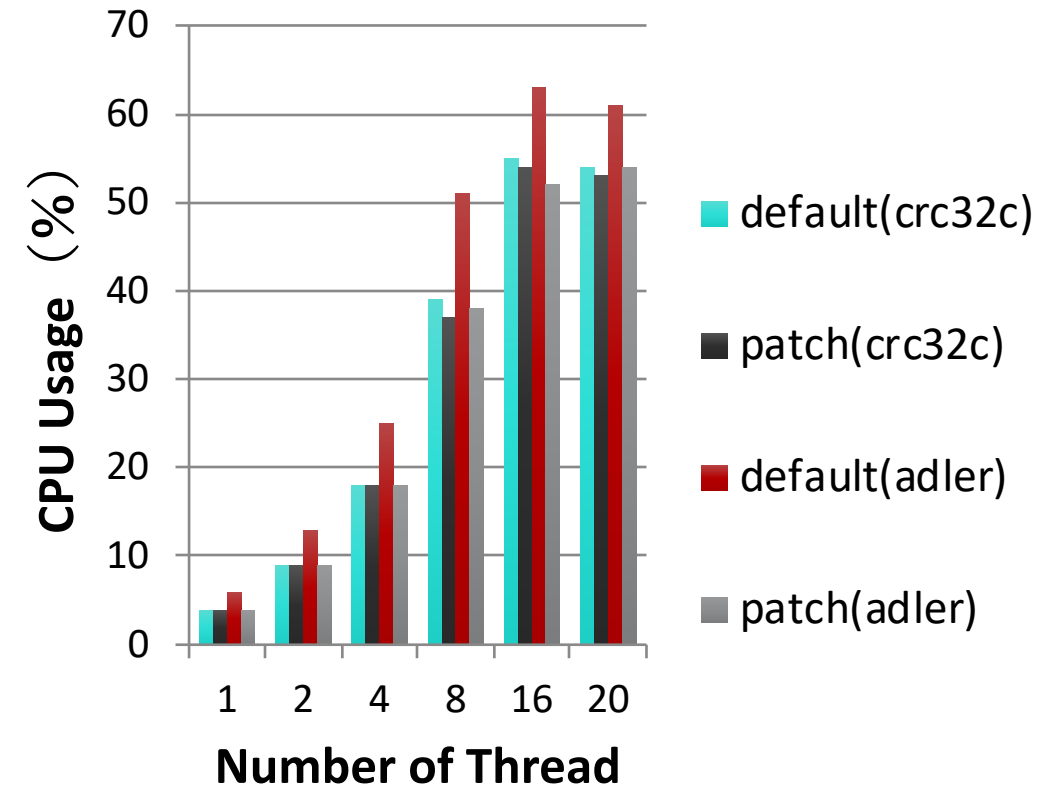
▶ Run IOR with Lustre Fake-IO

Performance Comparison: Single Client (FPP, Sequential, Write)

Single Client Performance(Write)

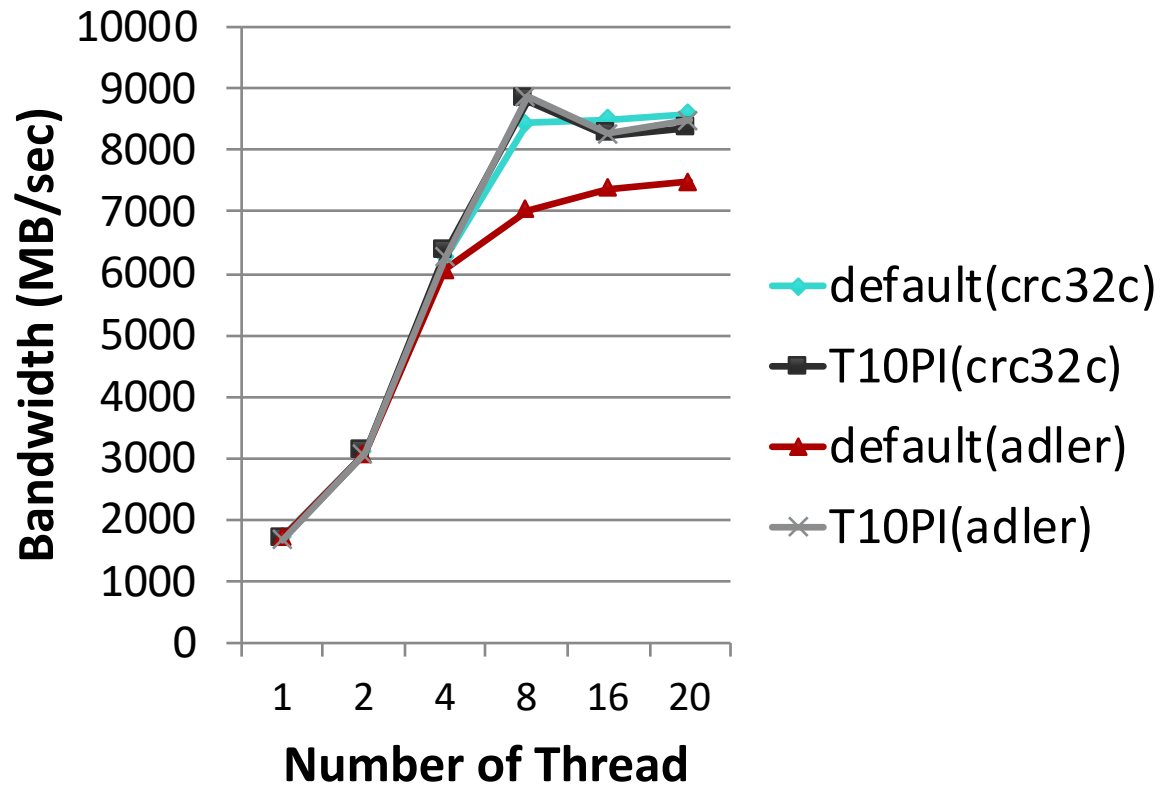


Client CPU Usages

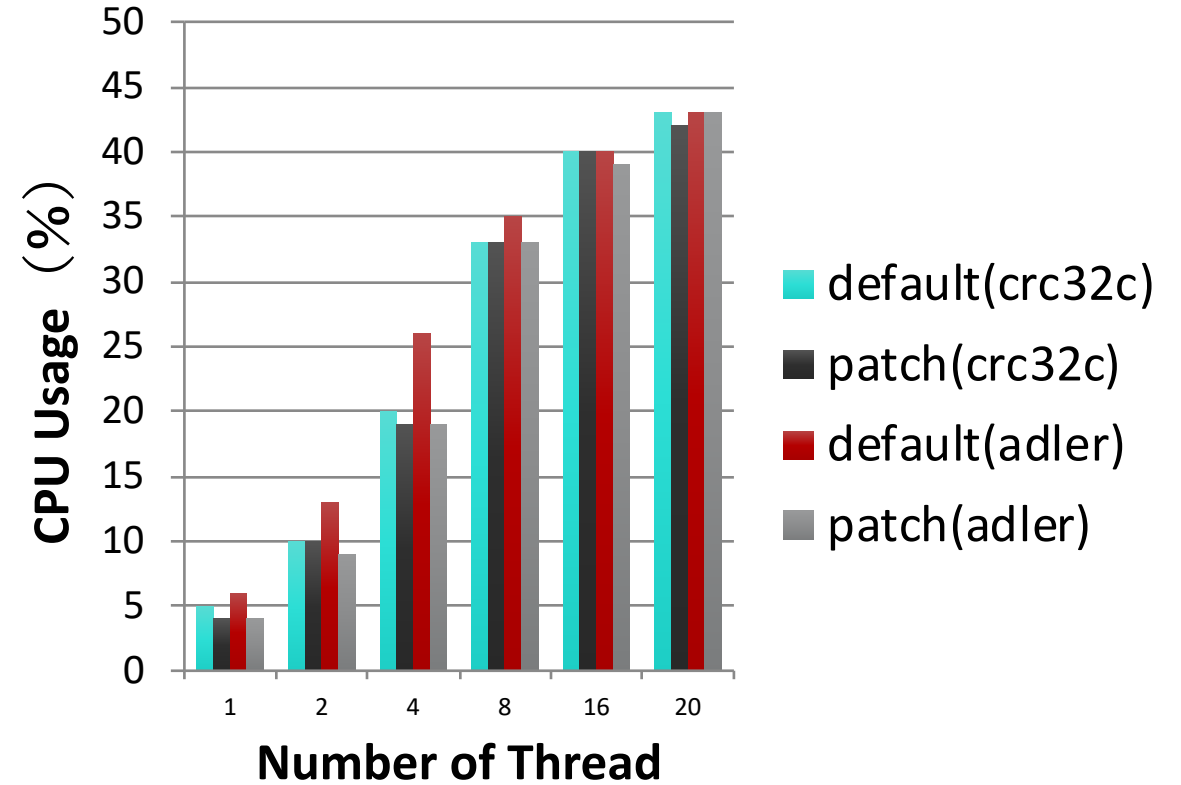


Performance Comparison- Single Client: (FPP, Sequential, Read)

Single Client Performance(Read)

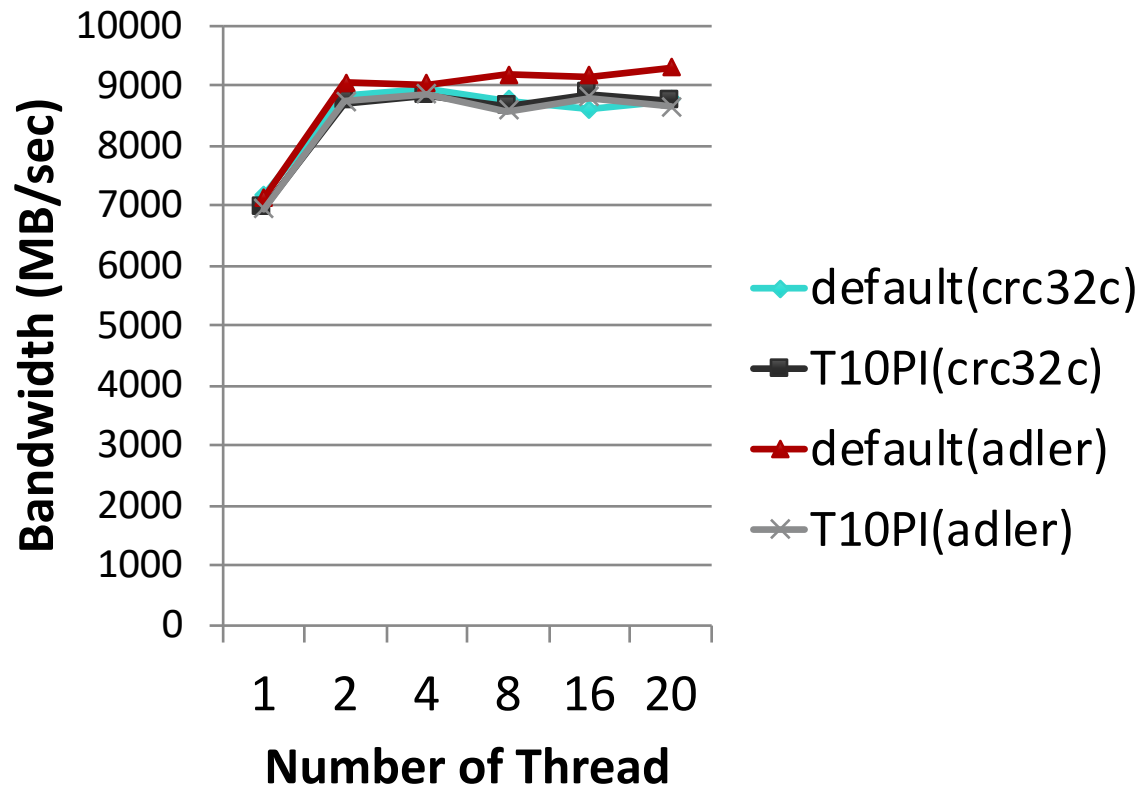


Client CPU Usages

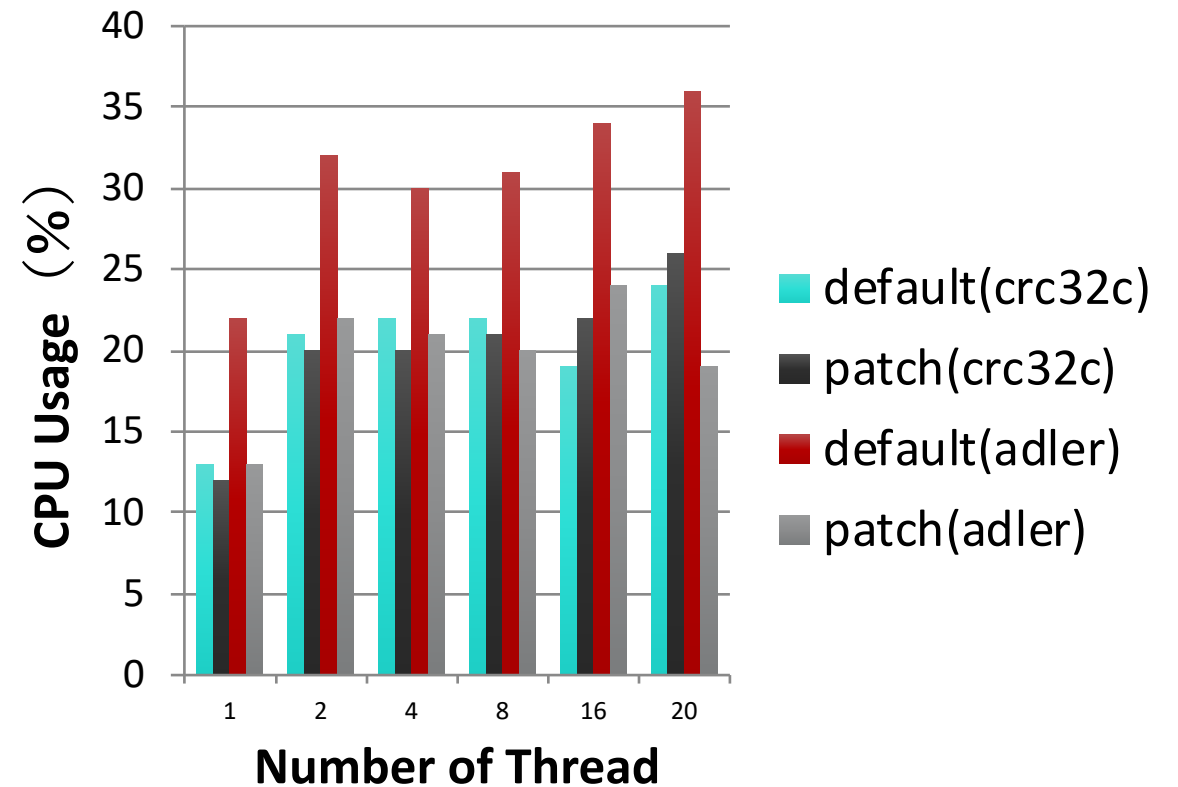


Performance Comparison: Multi Client/Single Server(FPP, Sequential, Write)

Single Server Performance(Write)

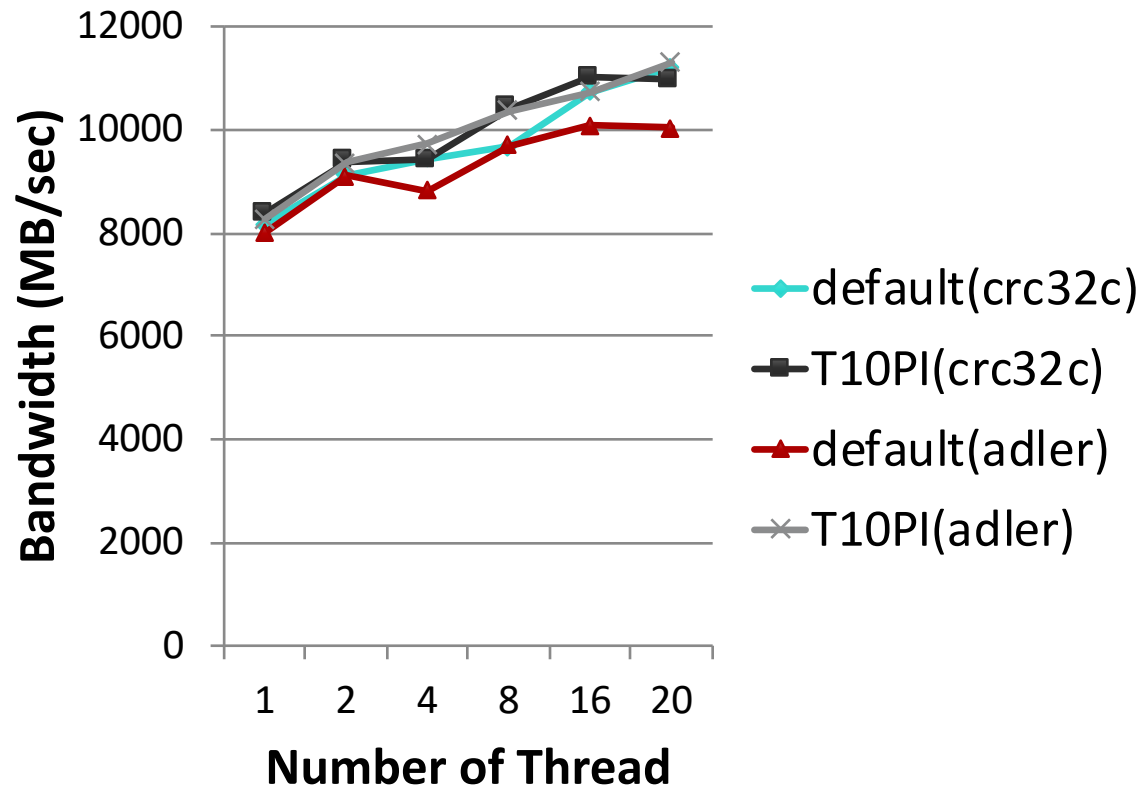


Server CPU Usages

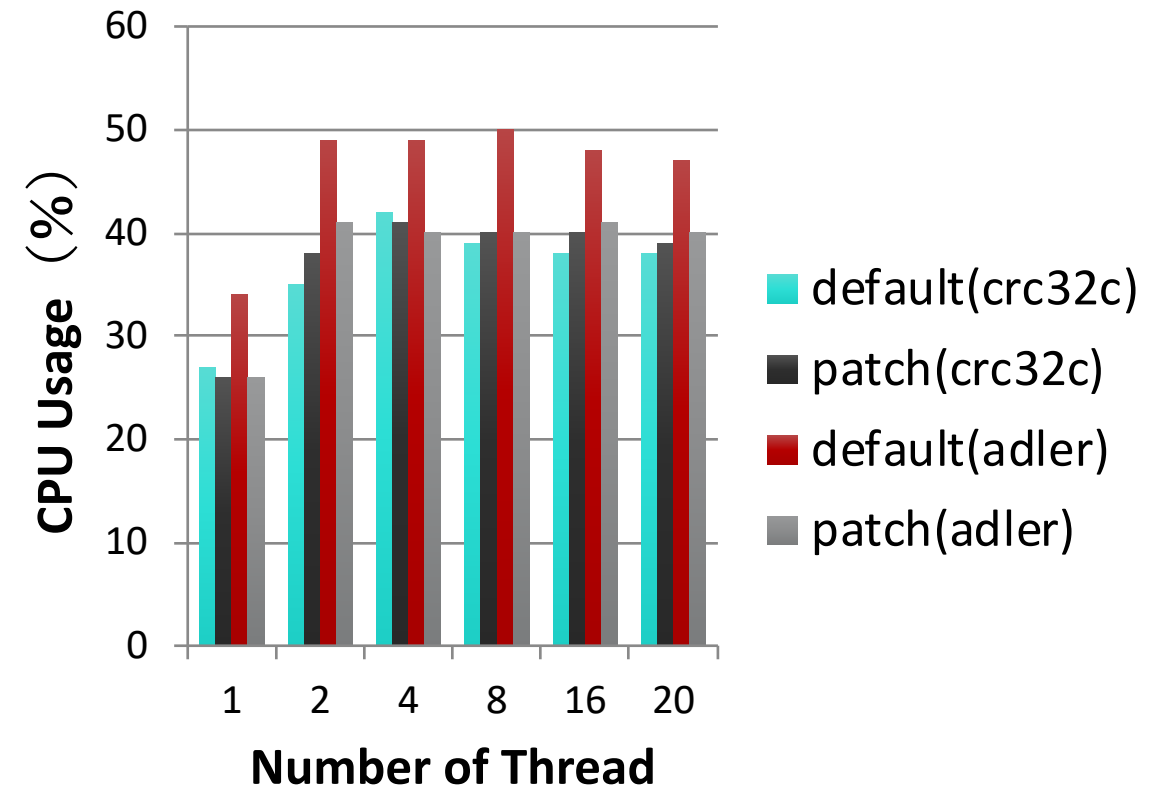


Performance Comparison: Multi Client/Single Server(FPP, Sequential, Read)

Single Server Performance(Read)



Server CPU Usages



Conclusions

► Designed Lustre End-to-End Data integrity

- Reused current Lustre checksum design and expended with T10PI/DIX.
- Flexible and adaptable to any T10PI/DIX supported hardware and software.
- Minimum performance impacts.

► Further Work

- Client side T10PI support to protect data corruption because of client side problems
- Continue benchmark and test not only between client and server, but also entire End-to-End comportment.



Whamcloud

Thank You!

