xyratex

Advancing Digital Storage Innovation

Improvements in Lustre Data Integrity

Nathan Rutman

Topics

- Lustre Wire Checksum Improvements
  - Cleanup
  - Portability
  - Algorithms
  - Performance
- End-to-End Data Integrity
  - T10DIF/DIX
  - Version Mirroring

Goals

- End user assurance that their data was written to disk accurately
- Protection against all RAID (single) failure modes
- Offload the heavy calculations from the servers
- Support a wider range of client/server hardware

Over-the-Wire Bulk Checksums

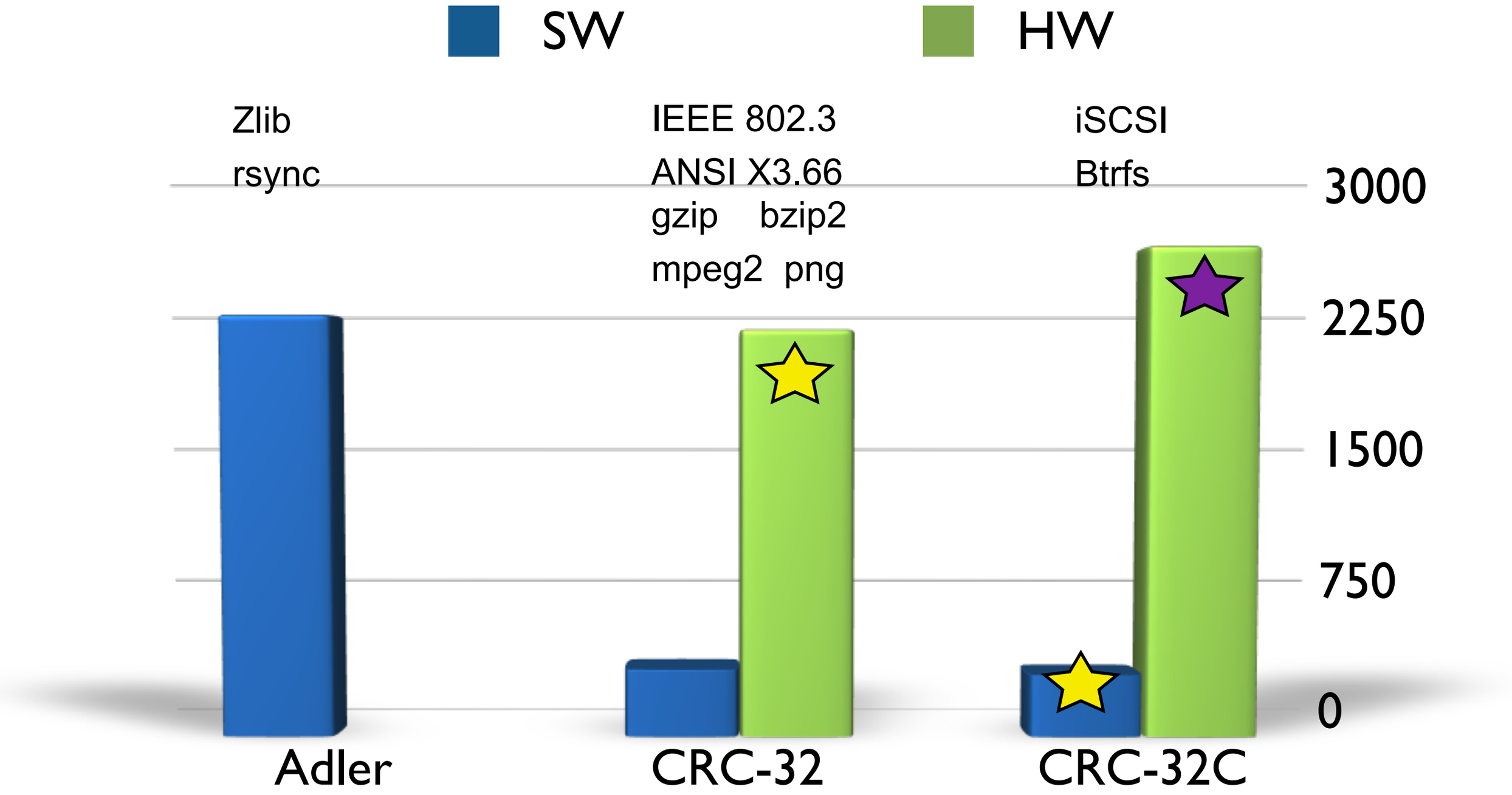
Bulk Checksum History

- Initially only software CRC-32 (IEEE, ANSI) 2007
- Adler-32 added in 1.6.5
  - easy to calculate
  - weak for small message sizes
- Shuichi Ihara added initial support for hardware CRC-32C (Castagnoli)
  - Intel Nehalem
  - bz 23549, landed in Lustre 2.2
- WC added multi-threaded ptlrpcd, and bulk RPC checksums moved into ptlrpcd context: parallelized checksums
  - LU-884, LU-1019, in Lustre 2.2
- sptlrpc implementation used a different set of functions
  - CRC-32, Adler, MD5, SHA1-512

Bulk Checksum Changes

- Cleanup of sptlrpc and bulk checksum algorithms to use the kernel crypto hash library
  - simplifies future additions
  - LU-1201
- Addition of Software CRC-32C support
  - eg. if server has HW support and clients don't
  - LU-1201
- Implementation of Hardware CRC-32 using PCLMULQDQ
  - Intel Westmere
  - MRP-314, still testing

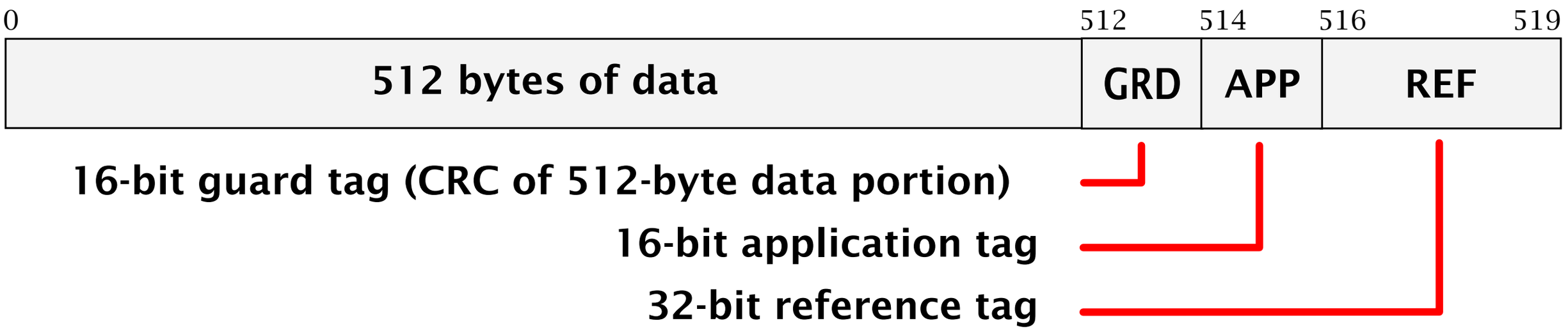
Bulk Checksum Speeds, MB/s



xyratex

End-to-End Data Integrity with T10 and Version Mirroring

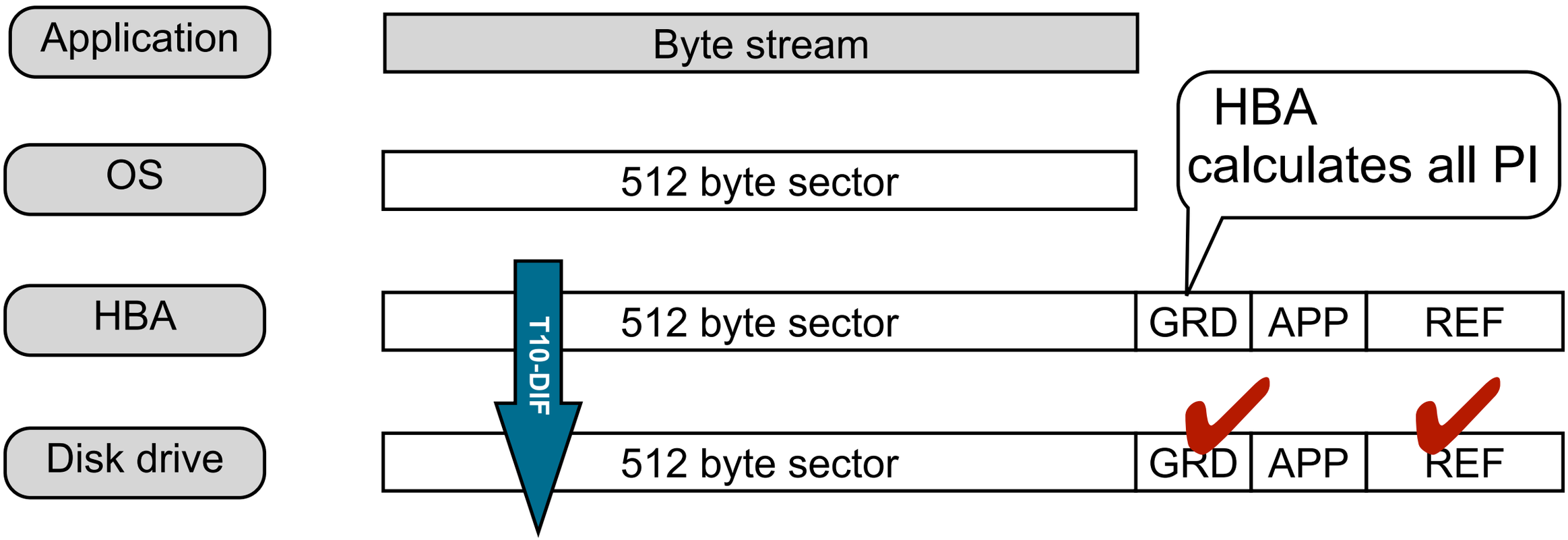
T10 DIF



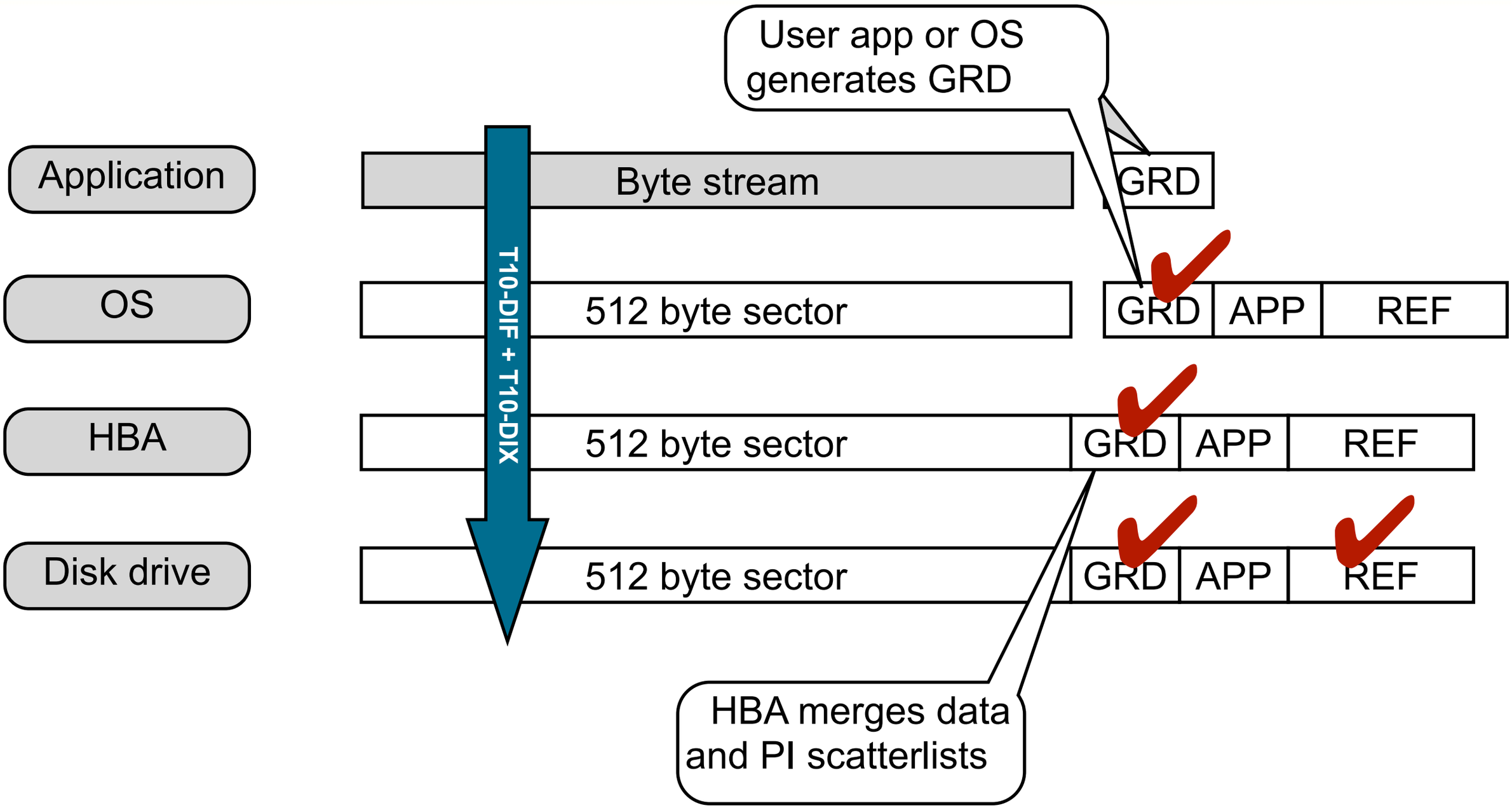
- 8 bytes of PI appended to 512 byte sectors
- HBA and disk drives must support T10-DIF in hardware

xyratex

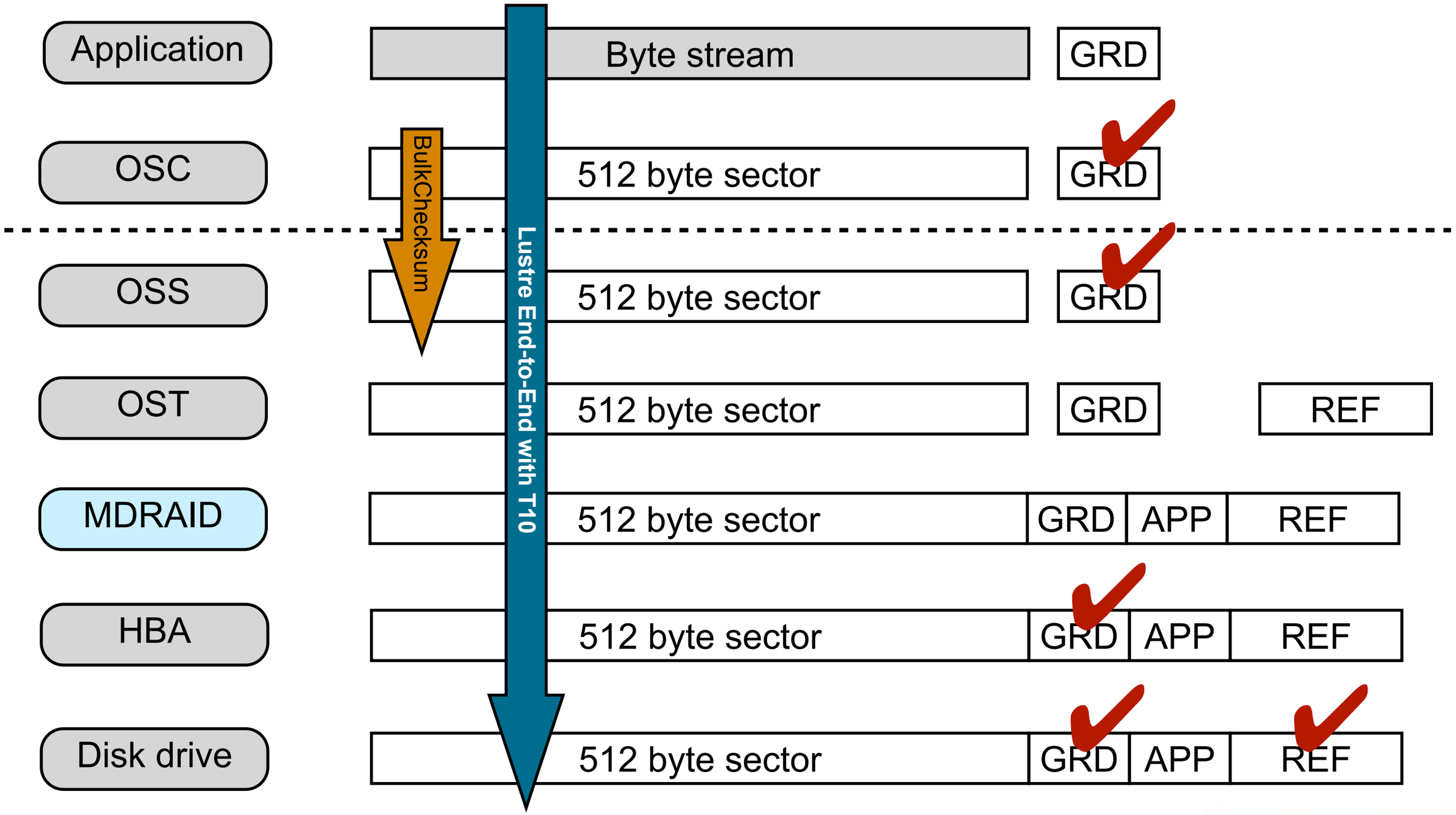
T10 DIF


Application
Byte stream
OS
512 byte sector
HBA calculates all PI
HBA
512 byte sector
GRD
APP
REF
T10-DIF
Disk drive
512 byte sector
GRD
APP
REF

T10 DIX


User app or OS generates GRD
Application
Byte stream
GRD
OS
512 byte sector
GRD
APP
REF
HBA
512 byte sector
GRD
APP
REF
T10-DIF + T10-DIX
Disk drive
512 byte sector
GRD
APP
REF
HBA merges data and PI scatterlists

T10 DIX with Lustre



Changes to Lustre

- Additional checksum data described or carried in brw RPC
- Add PI and checking to data path
- For mmap'ed pages, early GRD failure implies data has changed, recompute from OSC
- Disable bulk checksums
- Optional GRD checking on OSS can push all checksum load to HBA/disk hardware

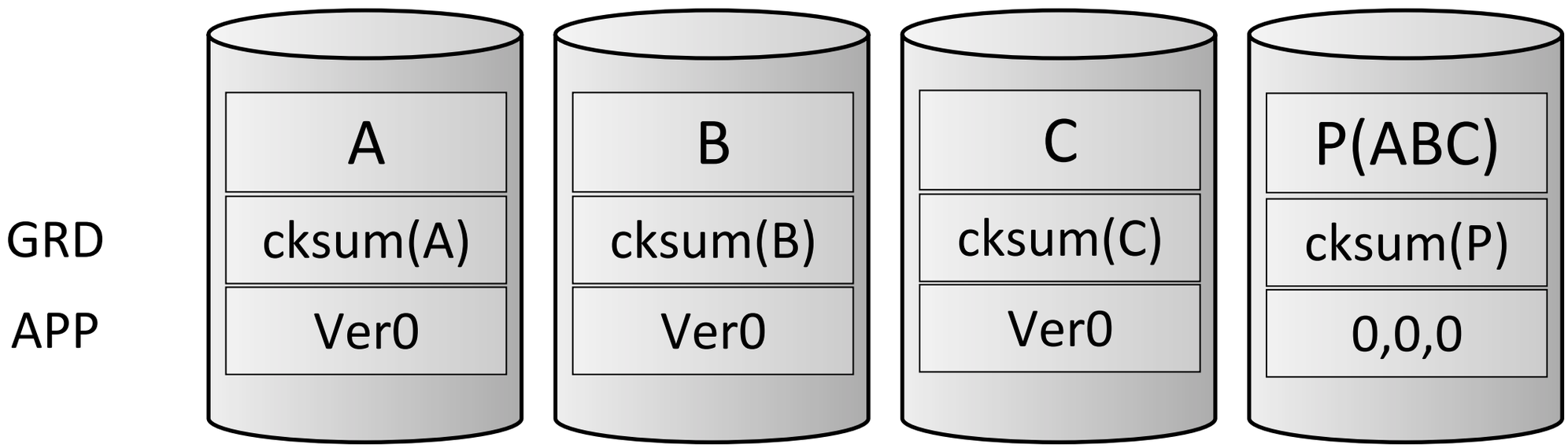
RAID failure modes

- *Parity Lost and Parity Regained* - Andrew Krioukov
- Latent Sector (reliable read) errors
- Data Corruption
- Misdirected Writes
- Lost Writes
- Torn Writes
- Parity Pollution

- Outcomes
  - Data recovery
  - Data loss (detected)
  - Data corruption (silent)

xyratex

RAID with Version Mirroring

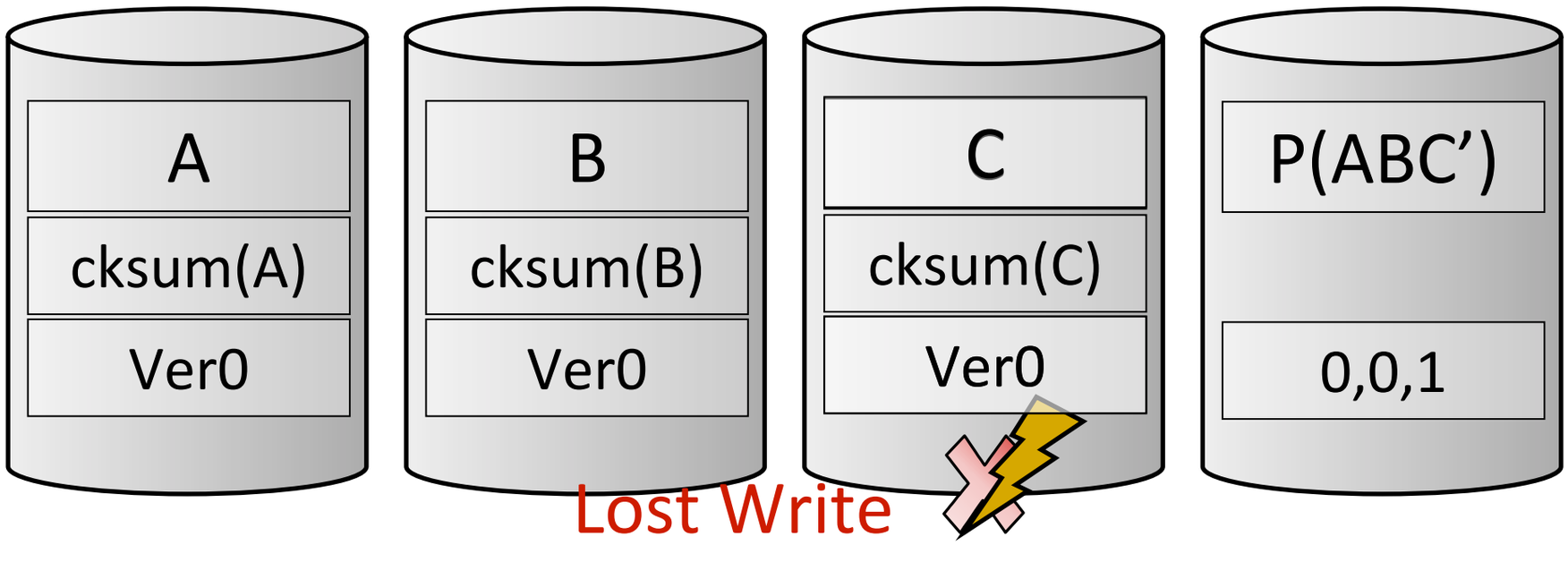


- RAID stripe across disks
- Block (chunk) on one disk
- Multiple sectors per block
- Store block version in T10 APP field
  - sectors within a chunk store the same version
  - parity block contains version vector

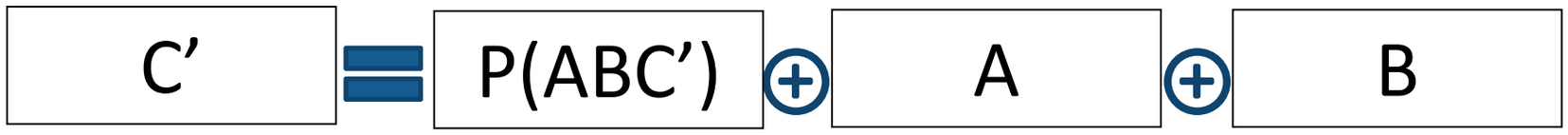
xyratex

Rebuild with Version Mirroring

Update C to C', write C' and P(ABC')



Later, attempt to update A to A'
First, read B & C to prepare for constructing new P
But first read P to verify versions before writing P(A'BC')
Version mismatch, latest is in P, so reconstruct C' from P



Write C'
Calculate P(A'BC')
Write A' and P(A'BC')

xyratex

RAID failure modes

- Latent Sector (reliable read) errors - drive detects
- Data Corruption - GRD, lightweight, partial reads
- Misdirected Writes - REF
- Lost Writes - parity block version vector
- Torn Writes - sector versions
- Parity Pollution - GRD + versions allow safe reconstruction

Fin