ORACLE®

Data Integrity Enhancements & I/O Topology

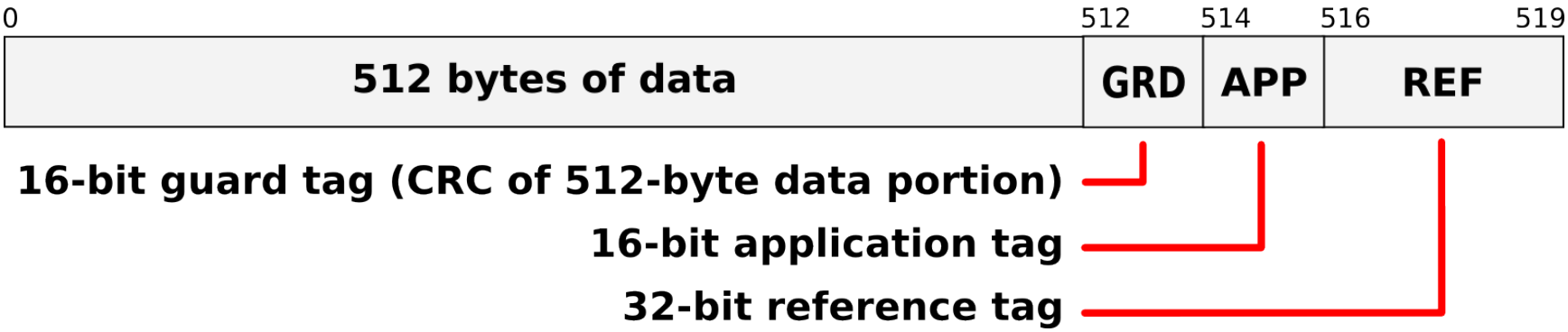
Martin K. Petersen
Consulting Software Developer, Linux Engineering

Data Integrity Enhancements

Linux Data Integrity Extensions

- T10 Protection Information model (also known as DIF) adds integrity checking capabilities to SCSI devices.
- Data Integrity Extensions allow exposing the T10 PI features to the operating system.
- Linux support for T10 PI & DIX:
  - Enables applications & filesystems to add checksums to I/O requests.
  - Preemptive technology that helps prevent corruption while data is *in flight.*
  - Allows us to detect problems at *WRITE* time before the original data is erased from memory, and before corrupted data ends up being stored on media.
  - Orthogonal to btrfs logical block checksumming which allows us to detect latent errors at *READ* time.

T10 Protection Information Model

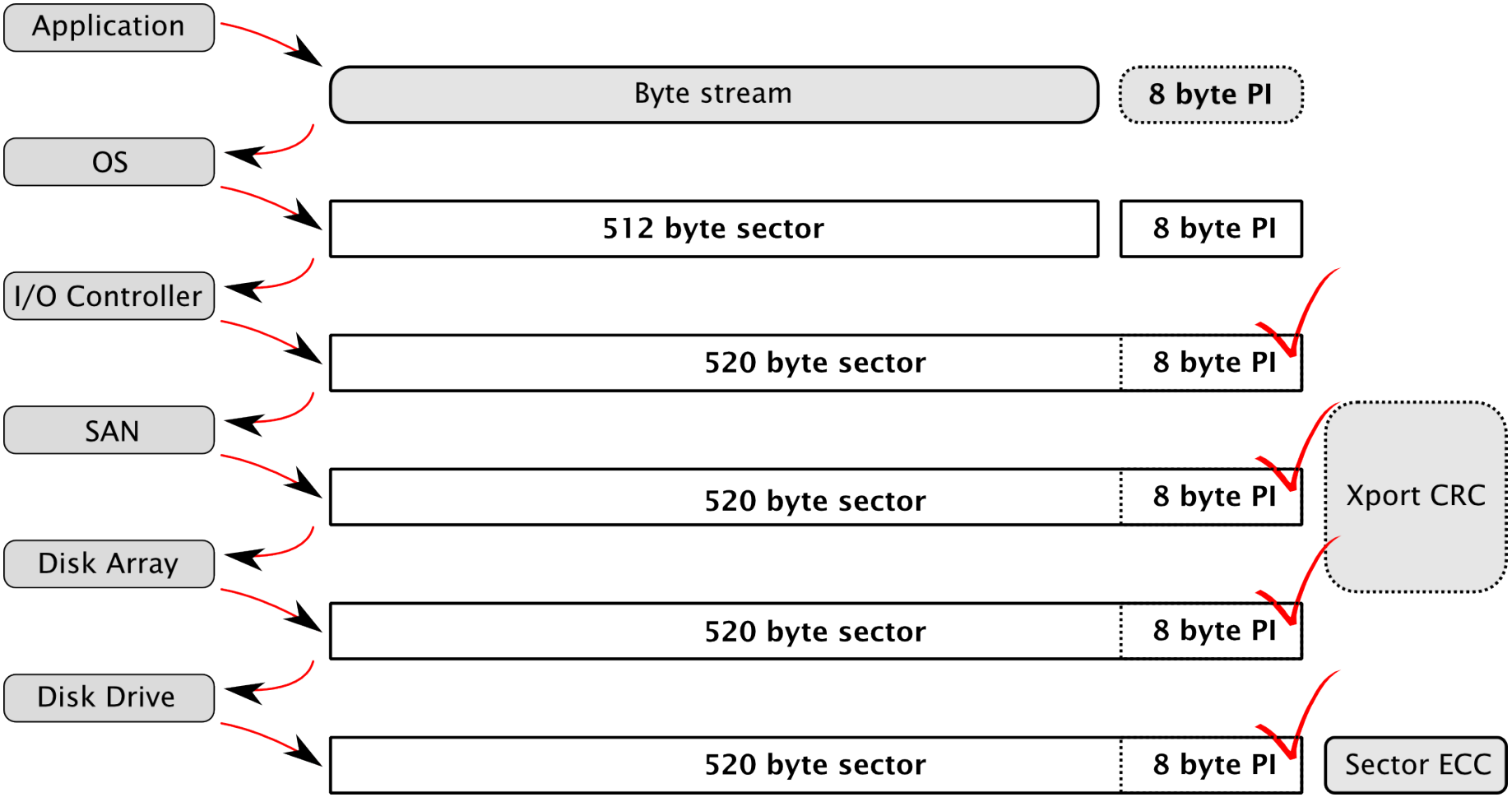


- Part of the SCSI Block Commands standard.
- Prevents data corruption and misplacement errors.
- Protects path between host adapter and storage device.
- Protection information is interleaved with data on the wire, i.e. effectively 520-byte sectors.
- Disk drives available now.
- RAID array support coming.

Data Integrity Extensions

- Developed by Oracle in collaboration with Emulex.
- DIX is a set of host adapter extensions that allow the operating system to send and receive T10 Protection Information instead of letting the controller generate it transparently.
- Enables true end-to-end data integrity protection.
- Specified for SCSI/SAS/FC class controllers.
- Hardware availability: 2009.
- Easy migration: Does not require T10 PI-capable storage.
- SATA support has been proposed in T13.

Data Integrity Extensions + T10 PI I/O



I/O Topology

4KB Hardware Sectors & Alignment

- Disk drive vendors are switching to 4KB hardware sectors to increase yield.
- For legacy reasons the DOS partition table scheme used by Windows and Linux is not 4KB-aligned.
- SATA drives will emulate 512-byte sectors and work with existing partitioning schemes and BIOS. However, the emulation comes with a read-modify-write performance penalty.
- Enterprise class drives will switch to 4KB sectors wholesale and will not emulate 512-byte sectors.
- Solid State Drives and RAID arrays also have similar alignment requirements. Both for performance reasons and for preventing fractured writes.

Linux I/O Topology

- The Linux I/O Topology patches allow us to extract correct alignment information from storage devices that support it.
- For RAID arrays we can also retrieve information about stripe size and internal block size.
- Using these parameters partitions and filesystems can be laid out in accordance with the characteristics of the underlying storage.
- The topology parameters are adjusted when block devices are partitioned, stacked, or combined using LVM or software RAID.
- Aiming for inclusion in 2.6.31.