ORACLE®

End-To-End Data Integrity Requirements For NFS

Chuck Lever <chuck.lever@oracle.com>
Consulting Member of Technical Staff

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

© 2012 Oracle Corporation

Today’s Data Integrity Solutions

- Application
  - ADB (block poisoning)
  - Filesystem data and metadata checksumming
- Transport integrity and encryption
  - Block CRC
  - Kerberos integrity and privacy
- Storage
  - On-disk checksums
  - RAID N+1
  - Disk scrubbing

ORACLE

© 2012 Oracle Corporation

Domain-Specific Checking Inadequate

- Silent data corruption on storage media
- Storage subsystem complexity increases failure rate
- Corruption often undetected until data is read: too late
- Checksums miss important failures
  - iSCSI can DMA the wrong pages from memory. The checksums will match the bad pages transferred
  - An array may receive incorrect data then use it to generate RAID parity blocks
  - Disks may never write data, or write it to the wrong LBA

© 2012 Oracle Corporation

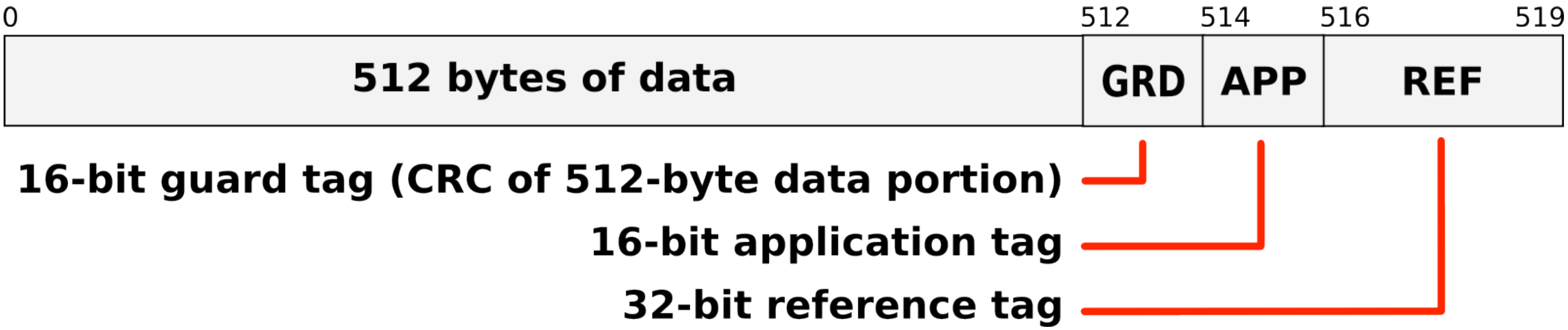
Protection Must Be End-To-End

- Protection should enable integrity check of I/O request at every stage of the I/O path during each write request

- Protected handoffs and conversions as the I/O transitions between domains with different protection schemes may be necessary

© 2012 Oracle Corporation

T10 Protection Information Model



16-bit guard tag (CRC of 512-byte data portion)

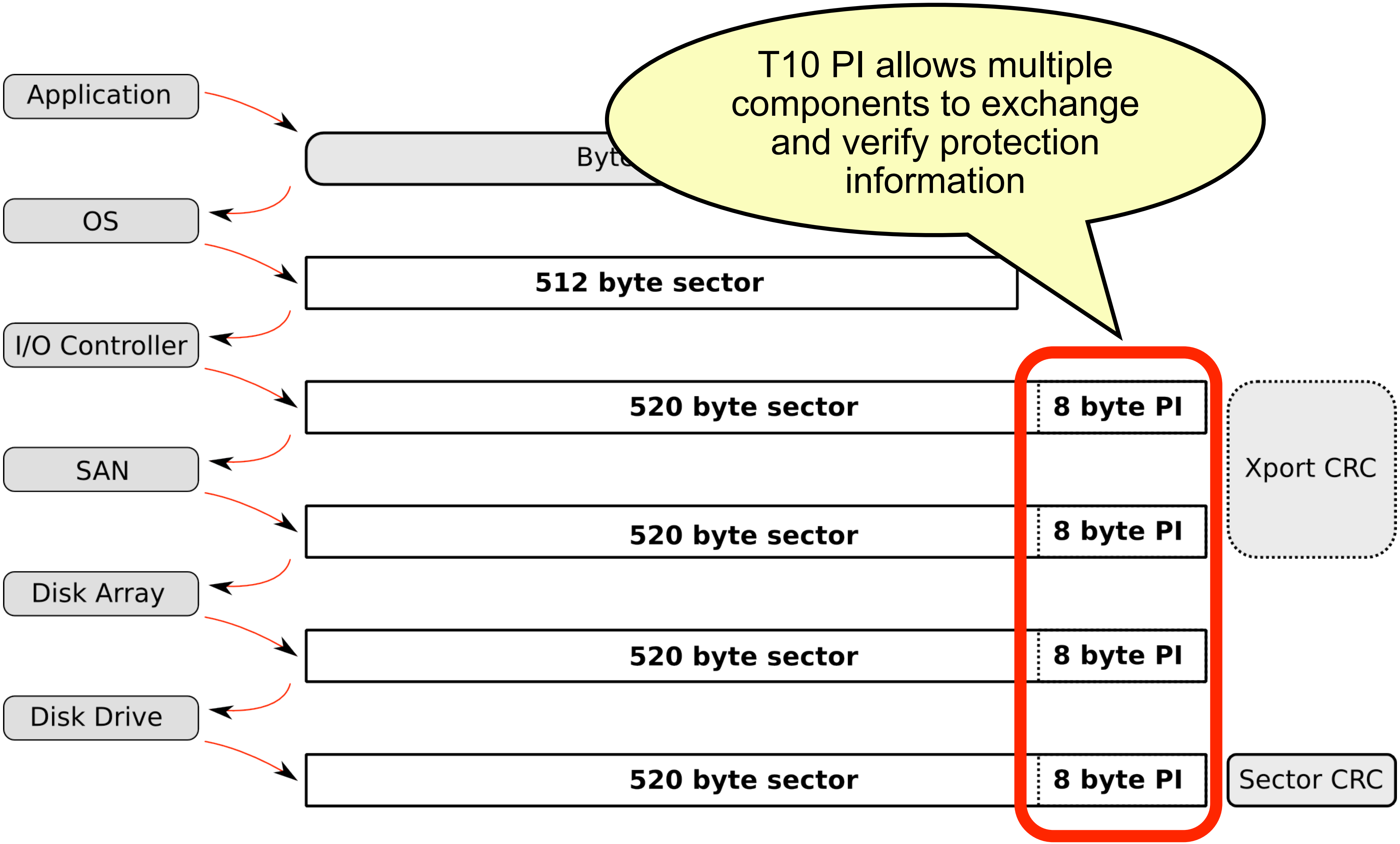
16-bit application tag

32-bit reference tag

- Standard 8 bytes of integrity metadata
  - Any layer of storage subsystem can re-verify the data
- Prevents content corruption and misplacement errors
- Protects path between HBA and storage device
- Protection information is interleaved with data on the wire, i.e. effectively 520-byte logical blocks

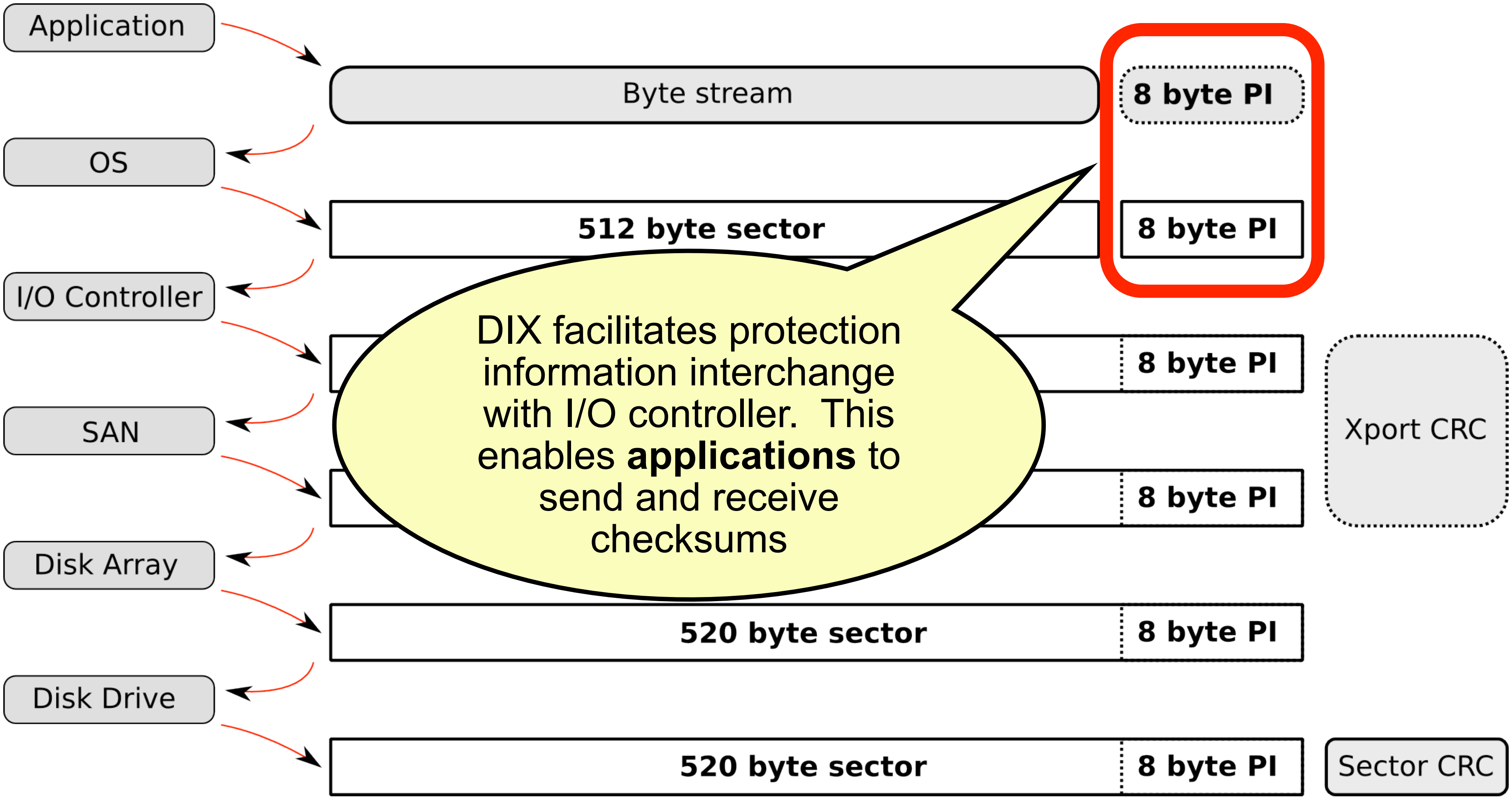
© 2012 Oracle Corporation

T10 Protection Information Model



© 2012 Oracle Corporation

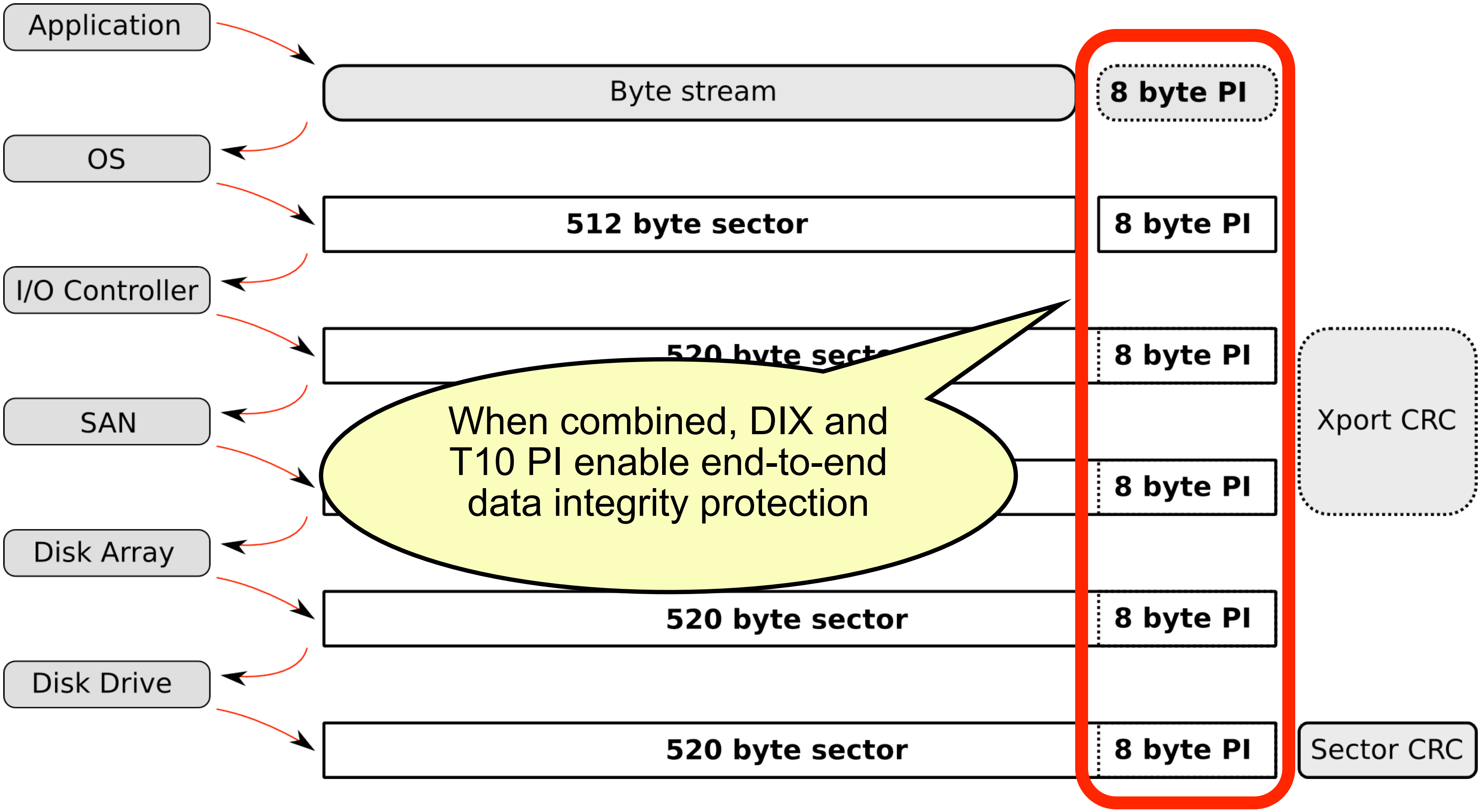
T10 Data Integrity Extensions



DIX facilitates protection information interchange with I/O controller. This enables **applications** to send and receive checksums

© 2012 Oracle Corporation

Data Integrity Extensions + T10 PI



When combined, DIX and T10 PI enable end-to-end data integrity protection

© 2012 Oracle Corporation

Marrying NFS with T10 PI + DIX

- When PI-capable storage is available on an NFS server, enable applications on NFS clients to access and update PI
  - Use of a maturing standard enables choice of access
  - Proposed optional feature of NFSv4.(n+1)

- Challenges to make this work for byte-stream access model
  - Obvious fit with pNFS block layout
  - PI-enabled NFS I/O and locking must be block aligned
  - Initial use mostly by non-POSIX NFS clients

© 2012 Oracle Corporation

High-Level Requirements

- Applications can read or write PI-protected data via SCSI, a pNFS layout, or non-pNFS NFS
- NFS servers can indicate that a particular share’s underlying storage supports PI
  - Not all classes of NFS servers may support PI, and not all FSIDs on one server may support it
  - Report supported T10 PI classes
  - Report data-to-PI size ratio of underlying storage
- NFS clients access PI data via an NFSv4 protocol extension
  - Must allow concurrent READ_PLUS variations (*e.g.* ADB)
  - Each READ contains one or more 8-byte PI values

© 2012 Oracle Corporation

Additional Considerations

- T10 PI + DIX is not owned by IETF
  - How to reference T10 work items in an IETF standard

- Security considerations
  - Ignore, recommend or require integrity-capable transport

© 2012 Oracle Corporation

The preceding is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

© 2012 Oracle Corporation

ORACLE IS THE INFORMATION COMPANY