ORACLE®

4KB Sectors, I/O Topology

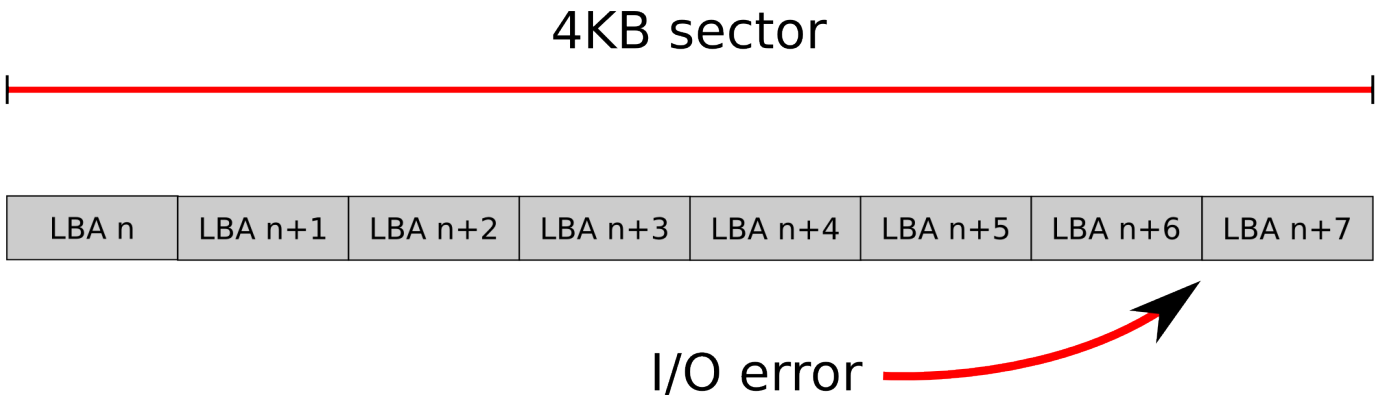
Martin K. Petersen
Consulting Software Developer, Linux Engineering

4KB Hardware Sector Update

- Disk drive vendors are switching to 4KB hardware sectors to increase yield.
- Desktop drives will emulate 512-byte sectors for backwards compatibility and booting.
- Enterprise drives will be 4KB native.
- For legacy reasons the DOS partition table scheme used by Windows and Linux is not 4KB-aligned.
- Windows Vista+ aligns 1st partition to 1MB+*e* boundary.
- For XP compatibility, desktop drives will ship with sector 63 aligned on a 4KB boundary.
- Enterprise and NL drives will be naturally aligned.
- Enterprise arrays align behind our backs.

4KB Hardware Sector Caveats

- Tendency to focus on alignment for performance reasons but fractured writes are a correctness problem.
- Filesystem journal padding:



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.

Linux I/O Topology

- List of regions:
  - Offset
  - Lengh
  - Minimum I/O size without incurring penalty
  - Optimal I/O size without incurring penalty
  - Maximum I/O size
  - Alignment
  - Consistency flag
- Functions for layering, combining and appending these lists
- /sys/block/foo/topology