Partial Parity Log

Partial Parity Log (PPL) is a feature available for RAID5 arrays. The issue addressed by PPL is that after a dirty shutdown, parity of a particular stripe may become inconsistent with data on other member disks. If the array is also in degraded state, there is no way to recalculate parity, because one of the disks is missing. This can lead to silent data corruption when rebuilding the array or using it is as degraded - data calculated from parity for array blocks that have not been touched by a write request during the unclean shutdown can be incorrect. Such condition is known as the RAID5 Write Hole. Because of this, md by default does not allow starting a dirty degraded array.

Partial parity for a write operation is the XOR of stripe data chunks not modified by this write. It is just enough data needed for recovering from the write hole. XORing partial parity with the modified chunks produces parity for the stripe, consistent with its state before the write operation, regardless of which chunk writes have completed. If one of the not modified data disks of this stripe is missing, this updated parity can be used to recover its contents. PPL recovery is also performed when starting an array after an unclean shutdown and all disks are available, eliminating the need to resync the array. Because of this, using write-intent bitmap and PPL together is not supported.

When handling a write request PPL writes partial parity before new data and parity are dispatched to disks. PPL is a distributed log - it is stored on array member drives in the metadata area, on the parity drive of a particular stripe. It does not require a dedicated journaling drive. Write performance is reduced by up to 30%-40% but it scales with the number of drives in the array and the journaling drive does not become a bottleneck or a single point of failure.

Unlike raid5-cache, the other solution in md for closing the write hole, PPL is not a true journal. It does not protect from losing in-flight data, only from silent data corruption. If a dirty disk of a stripe is lost, no PPL recovery is performed for this stripe (parity is not updated). So it is possible to have arbitrary data in the written part of a stripe if that disk is lost. In such case the behavior is the same as in plain raid5.

PPL is available for md version-1 metadata and external (specifically IMSM) metadata arrays. It can be enabled using mdadm option --consistency-policy=ppl.

There is a limitation of maximum 64 disks in the array for PPL. It allows to keep data structures and implementation simple. RAID5 arrays with so many disks are not likely due to high risk of multiple disks failure. Such restriction should not be a real life limitation.