**SYSTEM ANALYSIS**

**EXISTING SYSTEM:**

* The existing data deduplication schemes for primary storage, such as iDedup and Offline-Dedupe, are capacity oriented in that they focus on storage capacity savings and only select the large requests to deduplicate and bypass all the small requests (*e.g.*, 4KB, 8KB or less).
* The rationale is that the small I/O requests only account for a tiny fraction of the storage capacity requirement, making deduplication on them unprofitable and potentially counterproductive considering the substantial deduplication overhead involved.

**DISADVANTAGES OF EXISTING SYSTEM:**

* The existing data deduplication schemes fail to consider these workload characteristics in primary storage systems, missing the opportunity to address one of the most important issues in primary storage, that of performance.
* Existing scheme focuses on improving the read performance by exploiting and creating multiple duplications on disks to reduce the diskseek delay, but does not optimize the write requests. That is, it uses the data deduplication technique to detect the redundant content on disks but does not eliminate them on the I/O path. This allows the disk head to service the read requests by prefetching the nearest blocks from all the redundant data blocks on disk to reduce the seek latency.
* They only select the large requests to deduplicate and ignore all small requests (*e.g.*, 4KB, 8KB or less) because the latter only occupy a tiny fraction of the storage capacity.
* Moreover, none of the existing studies has considered the problem of space allocation between the read cache and the index cache. Most of them only use an index cache to keep the hot index in memory, leaving the memory contention problem unsolved.

**PROPOSED SYSTEM:**

* To address the important performance issue of primary storage in the Cloud, and the above deduplication-induced problems, we propose a Performance-Oriented data Deduplication scheme, called POD, rather than a capacity-oriented one (*e.g.*, iDedup), to improve the I/O performance of primary storage systems in the Cloud by considering the workload characteristics.
* POD takes a two-pronged approach to improving the performance of primary storage systems and minimizing performance overhead of deduplication, namely, a request-based selective deduplication technique, called Select-Dedupe, to alleviate the data fragmentation and an adaptive memory management scheme, called iCache, to ease the memory contention between the bursty read traffic and the bursty write traffic.
* More specifically, Select-Dedupe takes the workload characteristics of small-I/O-request domination into the design considerations. It deduplicates all the write requests if their write data is already stored sequentially on disks, including the small write requests that would otherwise be bypassed from by the capacity-oriented deduplication schemes.
* For other write requests, Select-Dedupe does not deduplicate their redundant write data to maintain the performance of the subsequent read requests to these data. iCache dynamically adjusts the cache space partition between the index cache and the read cache according to the workload characteristics, and swaps these data between memory and back-end storage devices accordingly. During the write-intensive bursty periods, iCache enlarges the index cache size and shrinks the read cache size to detect much more redundant write requests, thus improving the write performance.

**ADVANTAGES OF PROPOSED SYSTEM:**

* The extensive trace-driven experiments conducted on our lightweight prototype implementation of POD show that POD significantly outperforms iDedup in the I/O performance measure of primary storage systems without sacrificing the space savings of the latter.
* Moreover, as an application of the POD technology to a background I/O task in primary cloud storage, it is shown to significantly improve the online RAID reconstruction performance by reducing the user I/O intensity during recovery.
* *Reducing small write traffic* - By calculating and comparing the hash values of the incoming small write data, POD is designed to detect and remove a significant amount of redundant write data, thus effectively filtering out small write requests and improving I/O performance of primary storage systems in the Cloud.
* *Improving cache efficiency* - By dynamically adjusting the storage cache space partition between the index cache and the read cache, POD efficiently utilizes the storage cache adapting to the primary storage workload characteristics.
* *Guaranteeing read performance* - To avoid the negative read-performance impact of the deduplication-induced read amplification problem, POD is designed to judiciously and selectively, instead of blindly, deduplicate write data and effectively utilize the storage cache.