6. CONCLUSION AND FUTURE WORK

AppDedupe is an application aware scalable inline distributed deduplication framework for big data management which achieves a tradeoff between scalable performance and distributed deduplication effectiveness by exploiting application awareness, data similarity and locality. It adopts a two-tiered data routing scheme to route data at the super-chunk granularity to reduce cross-node data redundancy with good load balance and low communication overhead and employs application aware similarity index based optimization to improve deduplication efficiency in each node with very low RAM usage.

The evaluation clearly demonstrates the AppDedupe's significant advantages over the state-of-the-art distributed deduplication schemes for large clusters in the following important two ways. First, it outperforms the extremely costly and poorly scalable stateful tight coupling scheme in the cluster wide deduplication ratio but only at a slightly higher system overhead than the highly scalable loose coupling schemes. Second, it significantly improves the stateless loose coupling schemes in the cluster-wide effective deduplication ratio while retaining the latter's high system scalability with low overhead.

The scope of deduplication is broadening as all storage tiers, bandwidth reduction, regularity data, cost reduction for cloud backup services and so on. As a direction of future work, to further optimize our scheme for other resource constrained mobile devices like smartphone or tablet and investigate the secure deduplication issue in cloud backup services of the personal computing environment.