(intel®)

Intel® Technologies Unlock Apache Hadoop* Bottlenecks

Unleash Hadoop* potential with intelligent caching to Intel® NVMe* SSDs for the data center


Storage I/O can be a significant performance bottleneck for Hadoop* clusters, especially in hyperscale deployments where a single cluster can have hundreds or even thousands of nodes. Simply adding more, bigger HDDs will not solve scaling challenges and in fact, it can make things worse as the I/O per GB decreases while IT footprint and power consumption increases. The main objective of a scalable Hadoop storage solution is to remove storage I/O bottlenecks in a way that allows businesses to use higher capacity hard drives without a drop in performance.

At a Glance: Hadoop* Acceleration

- Hadoop clusters leverage parallel processing for big data analytics, but storage bottlenecks can limit performance

- By sorting storage I/Os into classes, and targeting classes to specific devices, those bottlenecks can be unlocked

- Example: Directing Hadoop YARN* storage I/O to a fast NVMe*-based Intel® SSD DC P4610, can increase performance up to 2x![1]

- Intel® Cache Acceleration Software (Intel® CAS) then manages that YARN storage device to prevent overloading during heavy traffic

Accelerate with Intel® Cache Acceleration Software to Increase Performance by nearly 2x![1]
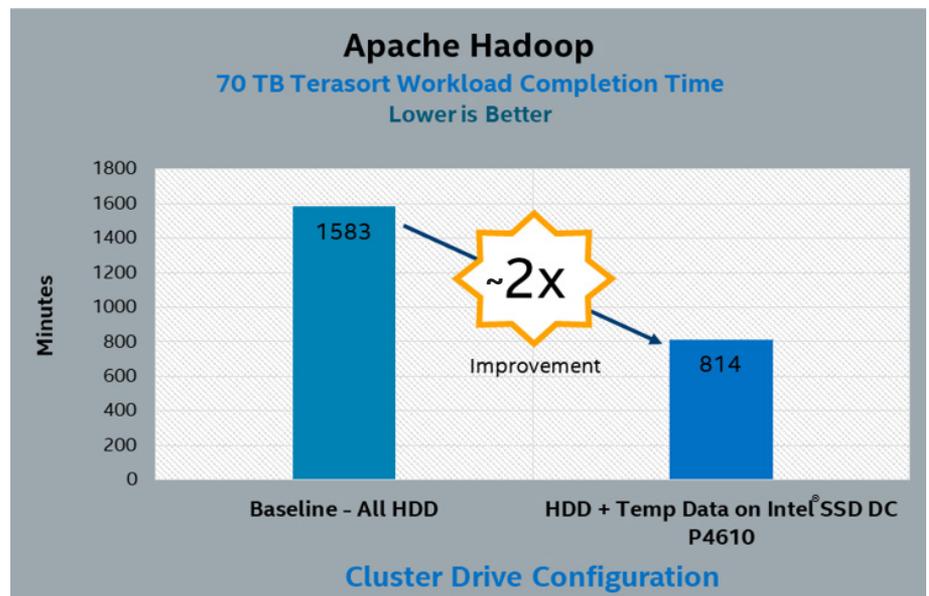
Direct Hadoop's YARN* data to a high-performance Intel® NVMe* cache drive for a 2x performance improvement. Configure and manage the cache with Intel® Cache Acceleration Software (Intel® CAS).

Using an NVMe*-based Intel® SSD to store temporary data managed by YARN* can eliminate contention for HDD throughput and can effectively boost cluster performance. However, this comes with one critical drawback – if the size of the temp data exceeds the size of the SSD, Hadoop jobs will fail. There is no native mechanism in Hadoop to overflow temp space to another drive.

With Intel® CAS, this application gap can be overcome. Intel CAS can manage the NVMe-based SSD as a caching device and prevent job failure. This then allows Hadoop users to gain the performance benefit of placing YARN data on an Intel NVMe SSD plus the flexibility to manage temp data overflow to other storage devices.

A new Intel CAS management feature allows users to select which data or directories to cache. For example, users may place a single directory into cache, to accelerate selected hotspots by caching to an Intel NVMe SSD. If the NVMe cache device becomes full due to workload surges, Intel CAS will smoothly flush data to the backend storage, thus preventing job failures. In this use case, the YARN data is selected as the cacheable directory and all storage I/O related to that class is sent to the Intel CAS managed device, a 6.4TB 3D NAND Intel® SSD DC P4610.

In the end, this Hadoop configuration can allow users to increase performance by up to 2x![1] This can enable users to achieve planned IOPs and capacity targets with half as many spindles/nodes/racks.

Intel Solutions Enable Quicker Business Decisions

Intel NVMe SSDs provide the I/O muscle to handle the heaviest workloads with transformative results. When coordinated with Intel CAS, data can be served, analyzed, and ready for business quicker. This ready-for-the-enterprise solution ensures the speed and data integrity demanded by both organizations of any size, and their customers.

For more information about Intel® CAS please visit: http://www.intel.com/cas

---

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark* and MobileMark*, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit intel.com/benchmarks.

Performance results are based on testing as of January 31, 2019 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.

Terasort 70TB workload. Sized to overflow the NVMe device and confirm Intel® CAS protection against failed jobs due to drive full condition

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No product or component can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.

Printed in USA   0519/JG/HC   ♻Please Recycle   340502-001US