DAOS Feature Update

DAOS User Group – SC 2021

Liang Zhen

intel®

Notices and Disclaimers

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.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. For more complete information about performance and benchmark results, visit http://www.intel.com/benchmarks .

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 http://www.intel.com/benchmarks .

Intel Advanced Vector Extensions (Intel AVX) provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at http://www.intel.com/go/turbo.

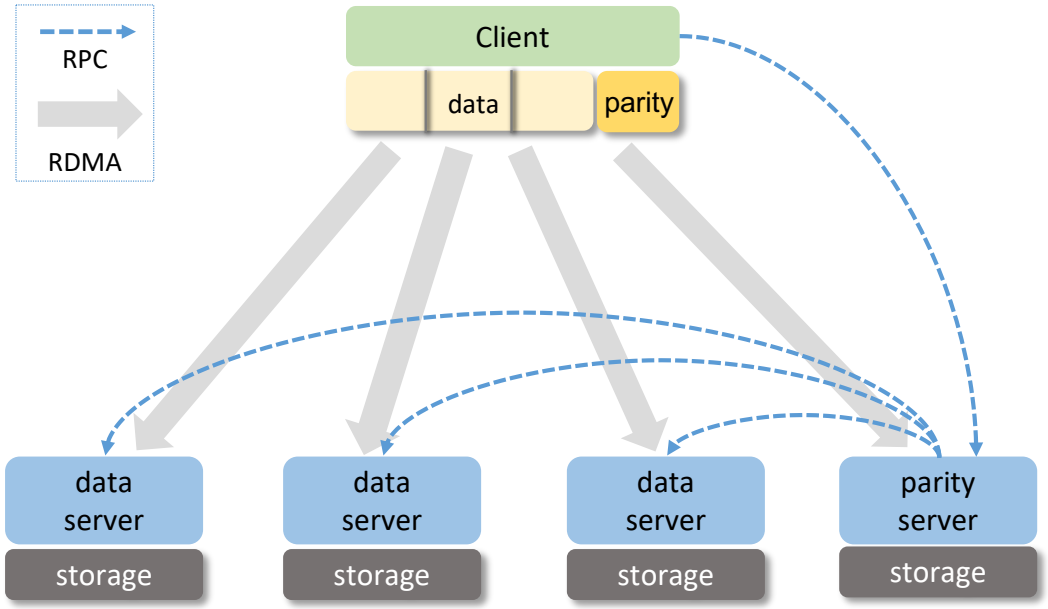
Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

Erasure Code

- Support both array and single value
  - Client-side encoding
- Aggregation
  - Local aggregation for incremental writes
  - Coordinated aggregation for overwrites
- Degraded mode
- Rebuild
  - Reconstruct data or parity on servers
- User interface
  - DAOS API: explicitly select EC object class, customized cell size
  - DFS auto-selection

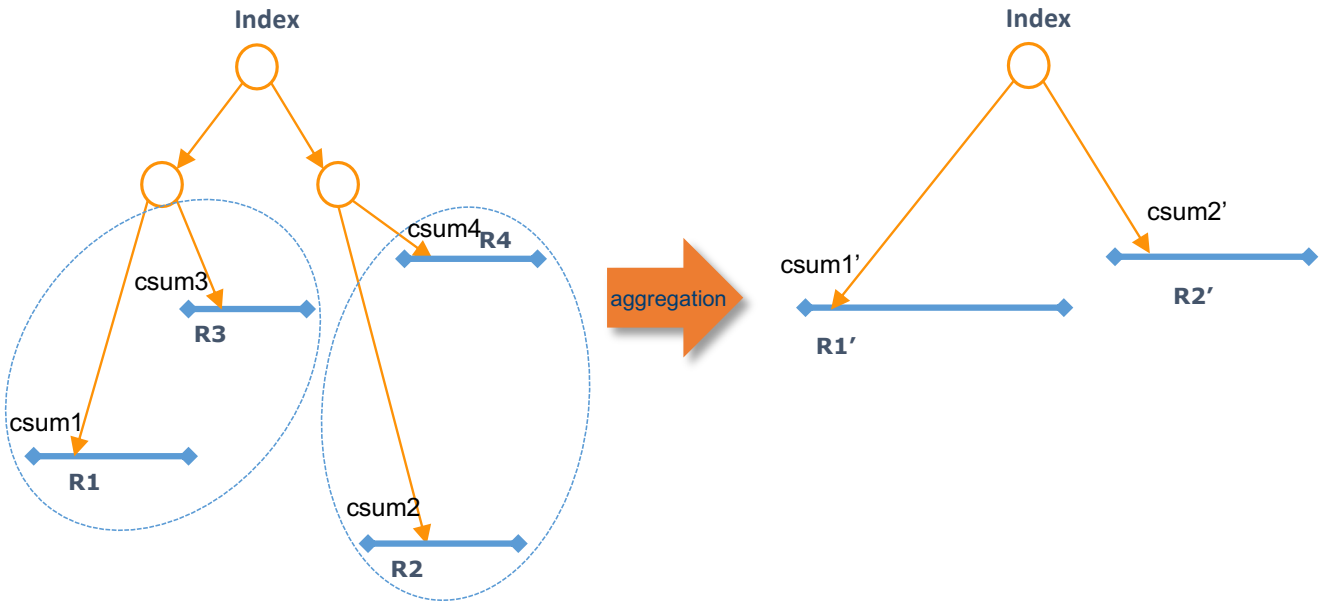


Distributed transaction

- Variant of two-phase commit protocol
- Client cached transaction
  - Submit reads
  - Cache all the writes
  - Send writes to a leader server on commit()
    - The leader is algorithmically selected from the servers involved in the transaction
    - The Leader server is the TX coordinator and runs the full two-phase commit protocol
- Transaction ID based resend check
- Transaction table aggregation

End-to-end data integrity (Checksum)

- Integrate checksum with other services
  - Rebuild, aggregation
- Checksum error handling
  - Report checksum errors
  - Degraded mode fetch
- Checksum scrubbing
  - Report errors through RAS event
  - Admin excludes faulty device



DAOS I/O stack

- VMD device support
  - Control SSD's status LEDs
- Object ID format change
  - Automatically select number of shards
  - Embed number of shards in object ID
- Object table
  - Collect all object IDs of container snapshot
  - Export object IDs through enumeration API

DAOS performance

- Pre-registered DMA buffer
  - MR registration of RDMA is expensive
  - Data flow: client memory -> DMA buffer -> SSD
- Resource throttling of rebuild
  - Throttle rebuild resource consumption based on overall DMA buffer size
- Performance throttling of aggregation
  - Track space consumption
  - Throttle I/O ULTs under space pressure

Version compatibility

- Forward compatible format
  - Stabilize durable format in VOS
  - Stabilize RPC format
- Unified PMEM data structure
  - SMD->VOS
  - One tool to do consistency check

Usability

- Redundancy factor
- RAS events
  - Structured RAS events for rebuild, SPDK I/O errors, checksum scrubbing errors…
  - Write RAS events to syslog
- Telemetry
  - Server-side metrics with Prometheus/Grafana integration
- Tools update
  - daos_perf, obj_ctl
  - object consistency checker

Control Plane/Tools Updates

- Usability Improvements
  - Tools updated to support identification of pools/containers by label instead of UUID
  - Automatic server config generator added to dmg
  - Pool create now has a single –size parameter to specify total pool size
  - Better alignment between dmg/daos tools
  - Tools accept positional arguments for required parameters


intel