SDC
STORAGE DEVELOPER CONFERENCE
SNIA SANTA CLARA, 2016

Accelerate Finger Printing in Data Deduplication

Xiaodong Liu & Qihua Dai
Intel Corporation

Legal Disclaimer

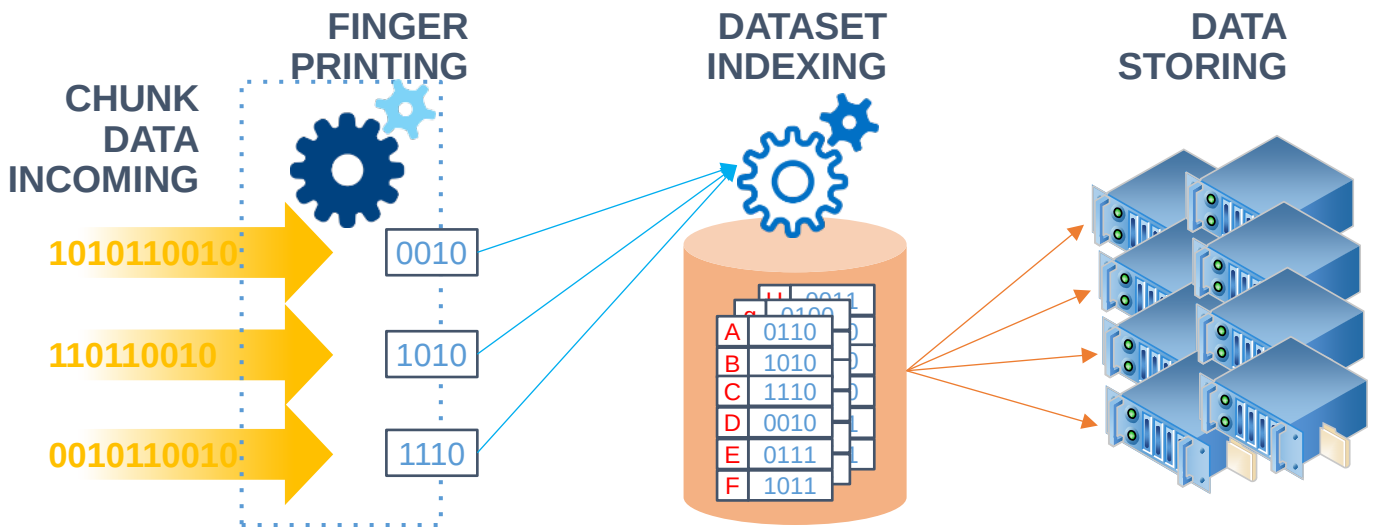
- INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.
- A "Mission Critical Application" is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL'S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS' FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENT IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.
- Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.
- The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.
- Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.
- Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or go to: https://www-ssl.intel.com/content/www/us/en/design/resource-design-center.html
- All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice.
- Intel, Intel logo, Look Inside, Intel Inside, Intel Inside logo, the Look Inside Logo, Intel Atom, and Intel Xeon are trademarks of Intel Corporation in the U.S. and other countries.
- *Other names and brands may be claimed as the property of others.
- Other vendors are listed by Intel as a convenience to Intel's general customer base, but Intel does not make any representations or warranties whatsoever regarding quality, reliability, functionality, or compatibility of these devices. This list and/or these devices may be subject to change without notice.
- 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 computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at http://www.intel.com/content/www/us/en/homepage.html.
- 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.
- 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.
-

Agenda

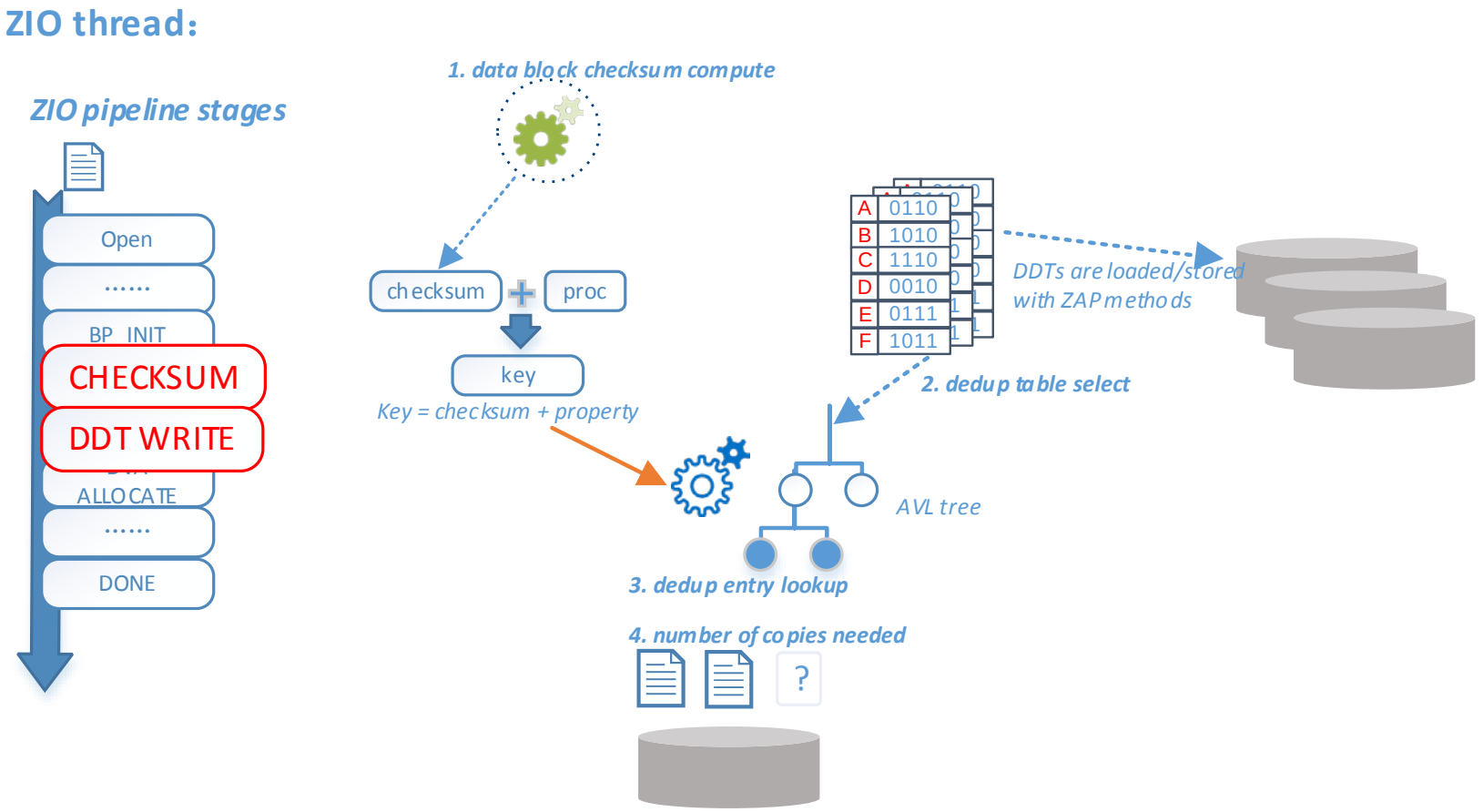
- Data Deduplication in ZFS
- Multi-buffer Technique Based on IA
- Limitations of Multi-buffer Hash in Linux Kernel
- Accelerating finger printing in ZFS Data Deduplication
- Performance Result of ZFS Data Deduplication

Data Deduplication in ZFS

- Data chunking in block size(512B~1MB) and inline.
- Finger printing algorithms: SHA256, Fletcher2/4

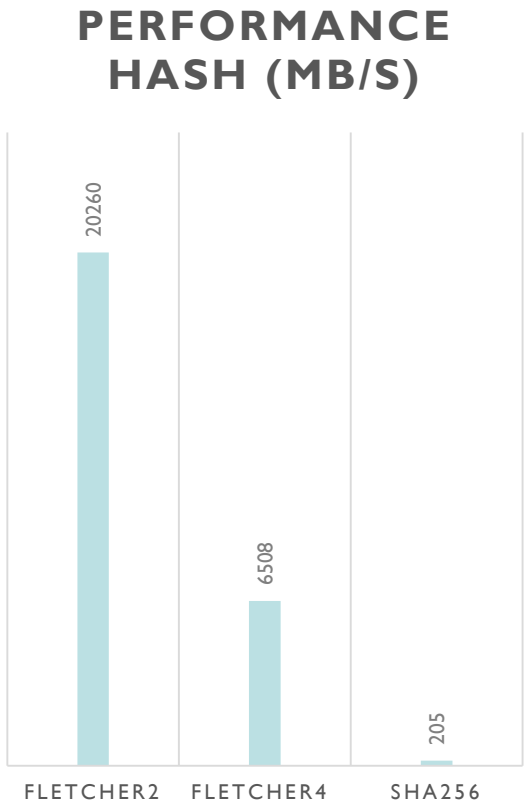


Data Deduplication in ZFS



Finger printing algorithms in ZFS

- Finger printing algorithms:
  - Flecher2/4: good performance, but bad collision rate
  - SHA256: good collision rate, widely used, but low performance and heavy CPU cost.



All results collected by Intel Corporation.

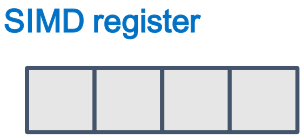
Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, visit Intel Performance Benchmark Limitations (http://www.intel.com/performance/resources/benchmark_limitations.htm).

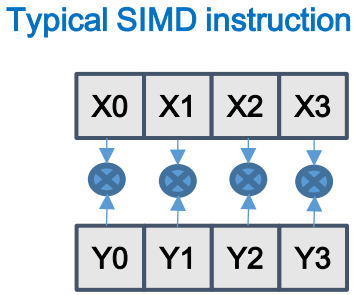
Multi-buffer Technique Based on IA

- Two basic ways that processing multiple buffers in parallel
  - SIMD approach
  - Non-SIMD approach

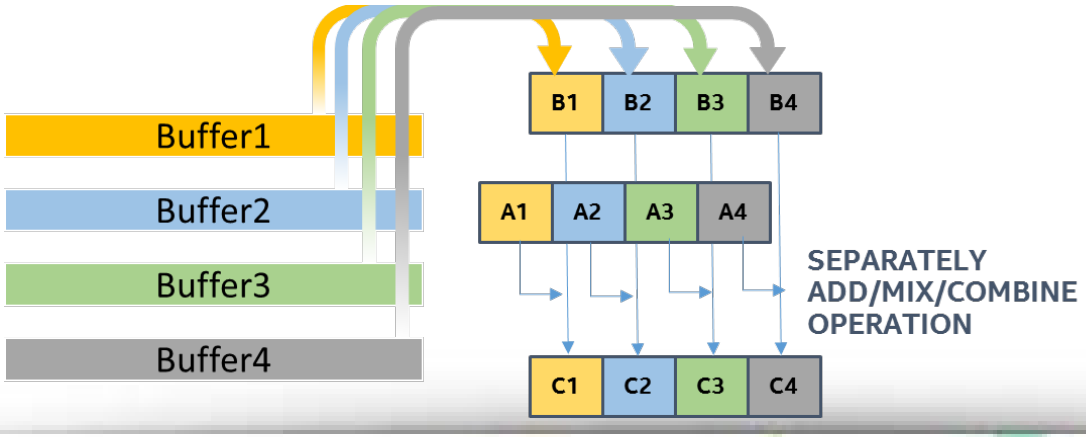
SIMD approach

- Processing multiple buffers in parallel to reduce data dependency limits
  - Based on SIMD registers and instructions



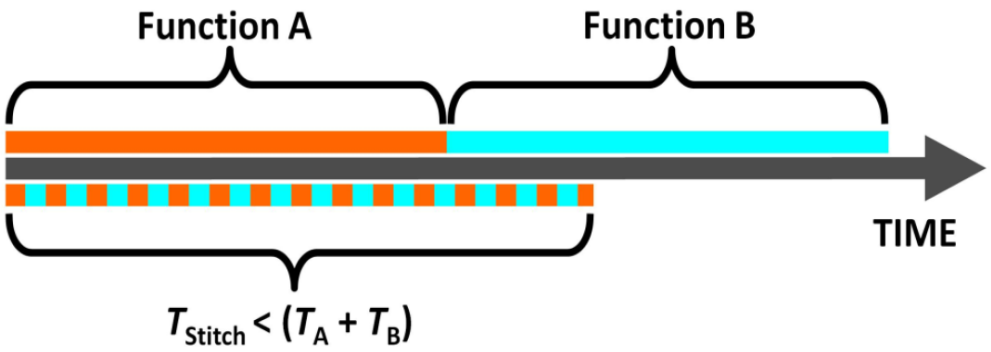


  - Parallel process independent buffers



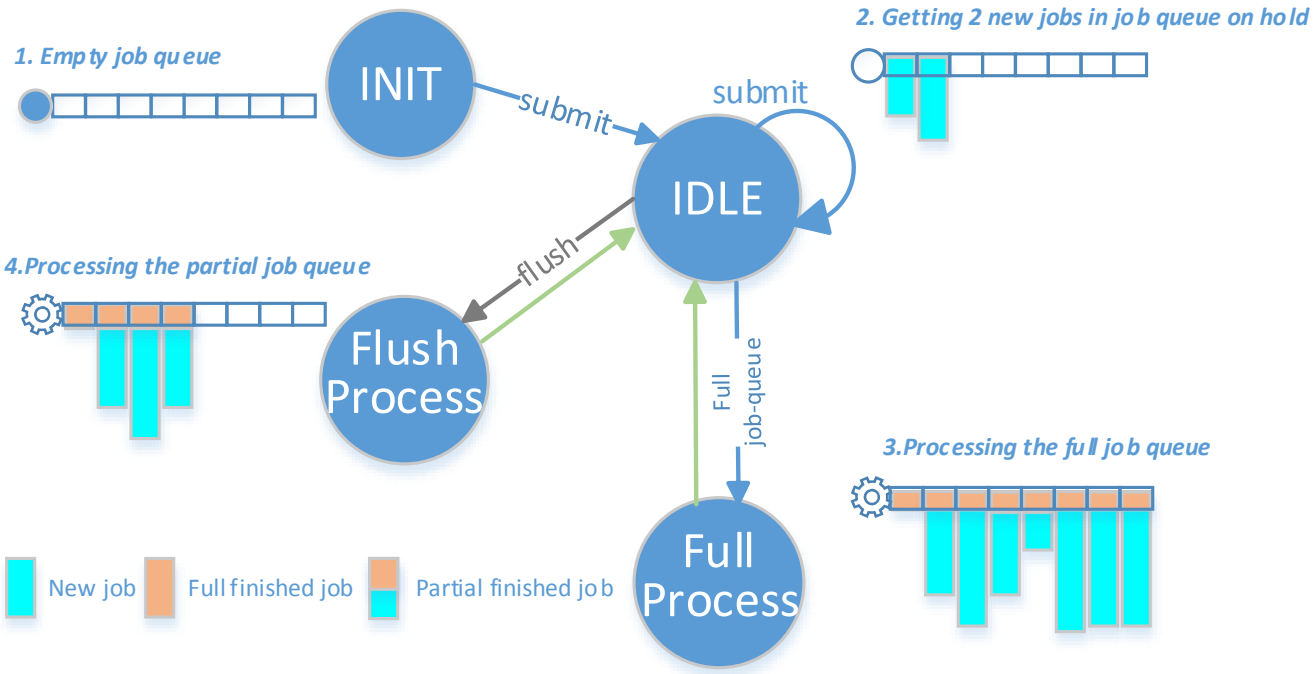
Non-SIMD approach

- Parallel utilization of the processor's execution resources (AES-CBC-Encrypt with AES-NI)
- Taking full advantage of execution unit resources – Function Stitching



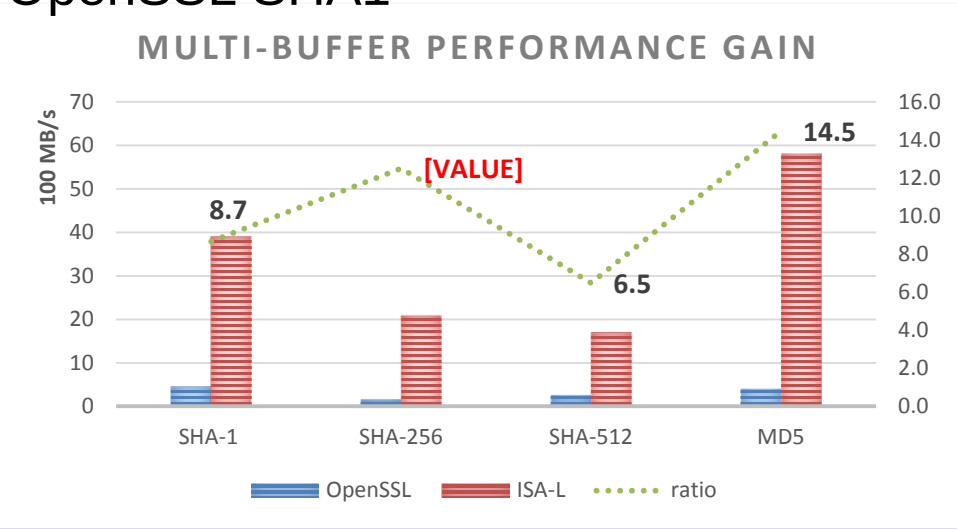
A typical implementation of Multi-Buffer Hash in Intel® Intelligent Storage Acceleration Library (Intel® ISA-L):

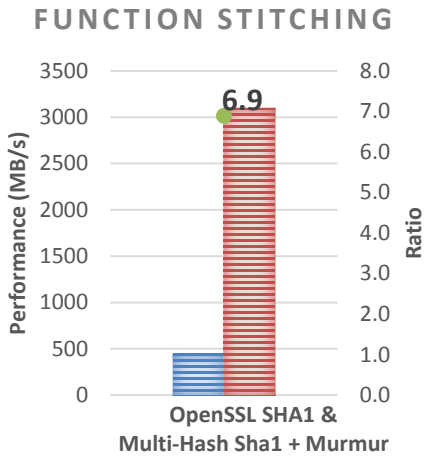
- Scheduler of ISA-L Multi-buffer Hash



Great Performance Gain

- SIMD multi-buffer tech: up to 14.5x performance improvement from Intel® ISA-L multi-buffer hash
- Parallel utilize CPU execution units: Intel® ISA-L Sha1 & murmur3 function stitching with 288bit digest has 6.9x higher throughput than OpenSSL SHA1



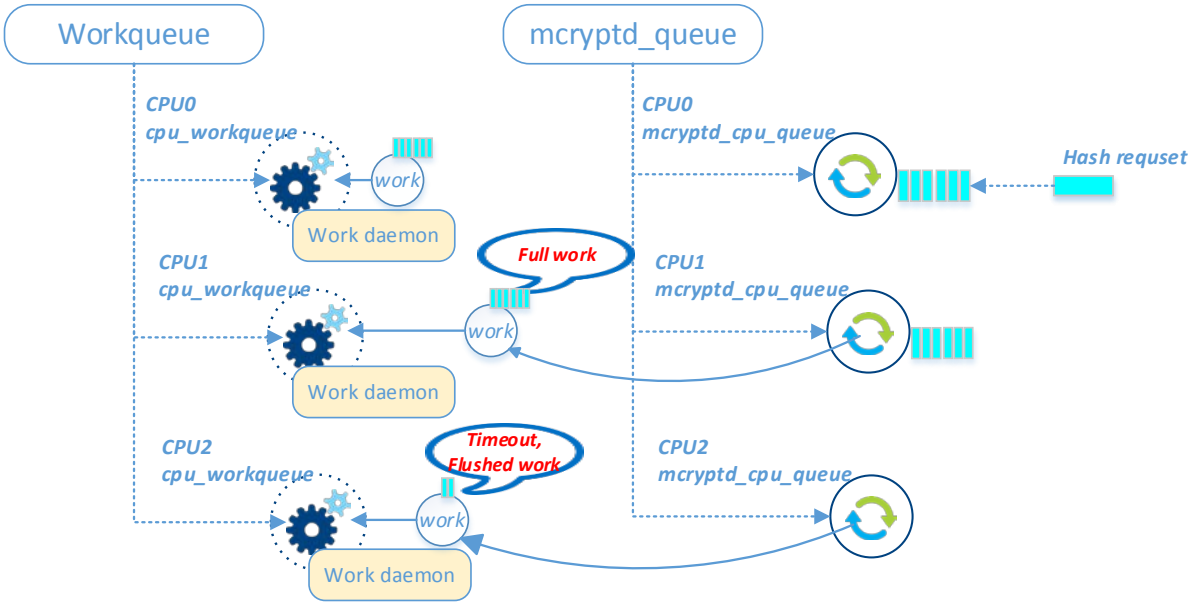


All results collected by Intel Corporation.

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, visit Intel Performance Benchmark Limitations (http://www.intel.com/performance/resources/benchmark_limitations.htm).

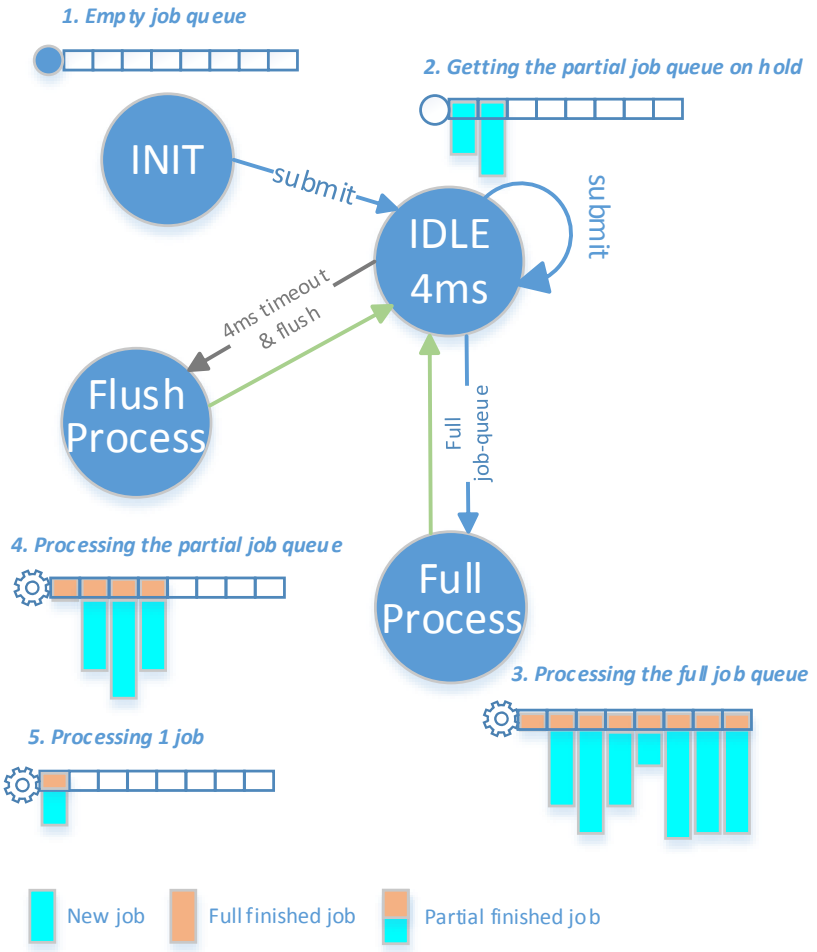
Multi-buffer hash in Linux Kernel Crypto Framework (LKCF)

- Intel® ISA-L Multi-buffer SHA1, SHA256, SHA512 are integrated
- The unified LKCF hash APIs
- Multi-buffer framework in LKCF:
  - The jobs in mcryptd_queue are assigned to the work daemon in workqueue
  - For sha256 AVX2 version, work daemon is hold until 8 jobs submitted, or timeout (usually 4ms)



The limitations of multi-buffer hash in LKCF

- partial job queue (< 8 jobs) always holds until 4ms timeout and flush.
- full job queue (= 8 jobs) gets processed immediately



The limitations of multi-buffer hash in LKCF

- Multi-buffer sha256 never gets the better performance than single-buffer sha256 for <8 jobs.

Intel® Xeon® Processor E5-2650v3 @ 2.3 GHz 1 Socket		
	Single-buffer sha256 (single core)	Multi-buffer sha256 in LKCF (single core)
1 job	183MB/s	24.3MB/s
2 jobs	183MB/s	48.6MB/s
3 jobs	183MB/s	72.9MB/s
4 jobs	183MB/s	97.2MB/s
5 jobs	183MB/s	121.5MB/s
6 jobs	183MB/s	145.8MB/s
7 jobs	183MB/s	170.1MB/s
8 jobs	183MB/s	**877MB/s**

All results collected by Intel Corporation.

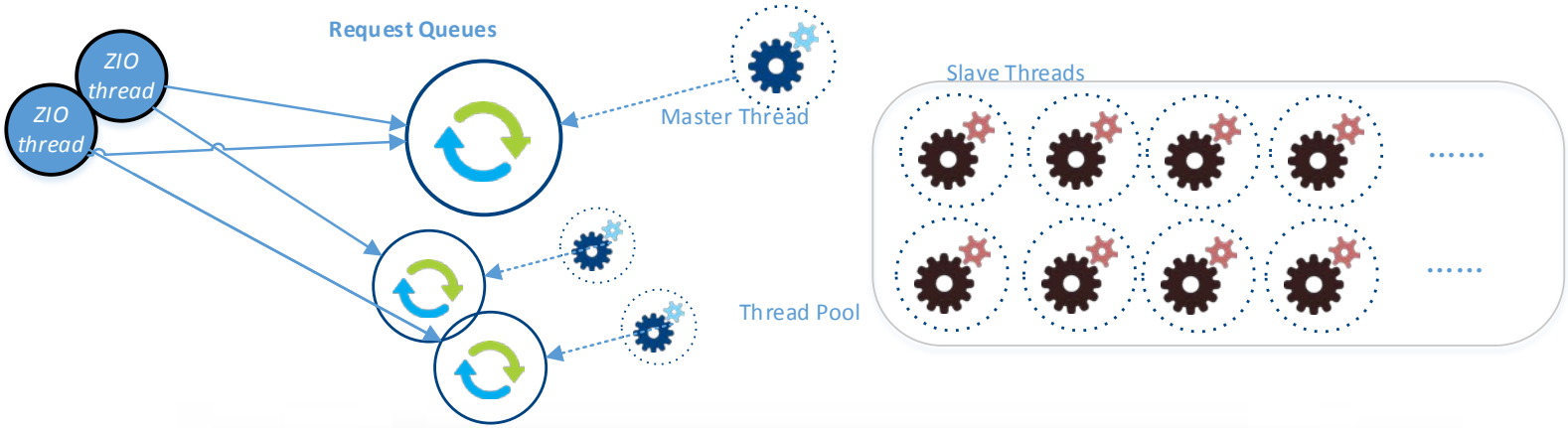
Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, visit Intel Performance Benchmark Limitations (http://www.intel.com/performance/resources/benchmark_limitations.htm).

Finger Printing Accelerating in ZFS Data Deduplication

- SHA256 in ZFS data deduplication is traditional Single-buffer
  - Suitable for light workload
  - High compute resource cost on heavy workload
- Using Multi-buffer sha256 in LKCF for ZFS data deduplication
  - Always has the timeout overhead for < 8 jobs (AVX2), and <16 jobs (AVX512)
  - Not suitable for file system
- Our acceleration solution for finger printing accelerating for ZFS data deduplication
  - Basically leverage multi-buffer and single buffer functions adaptively
  - Break through advanced usage considerations.

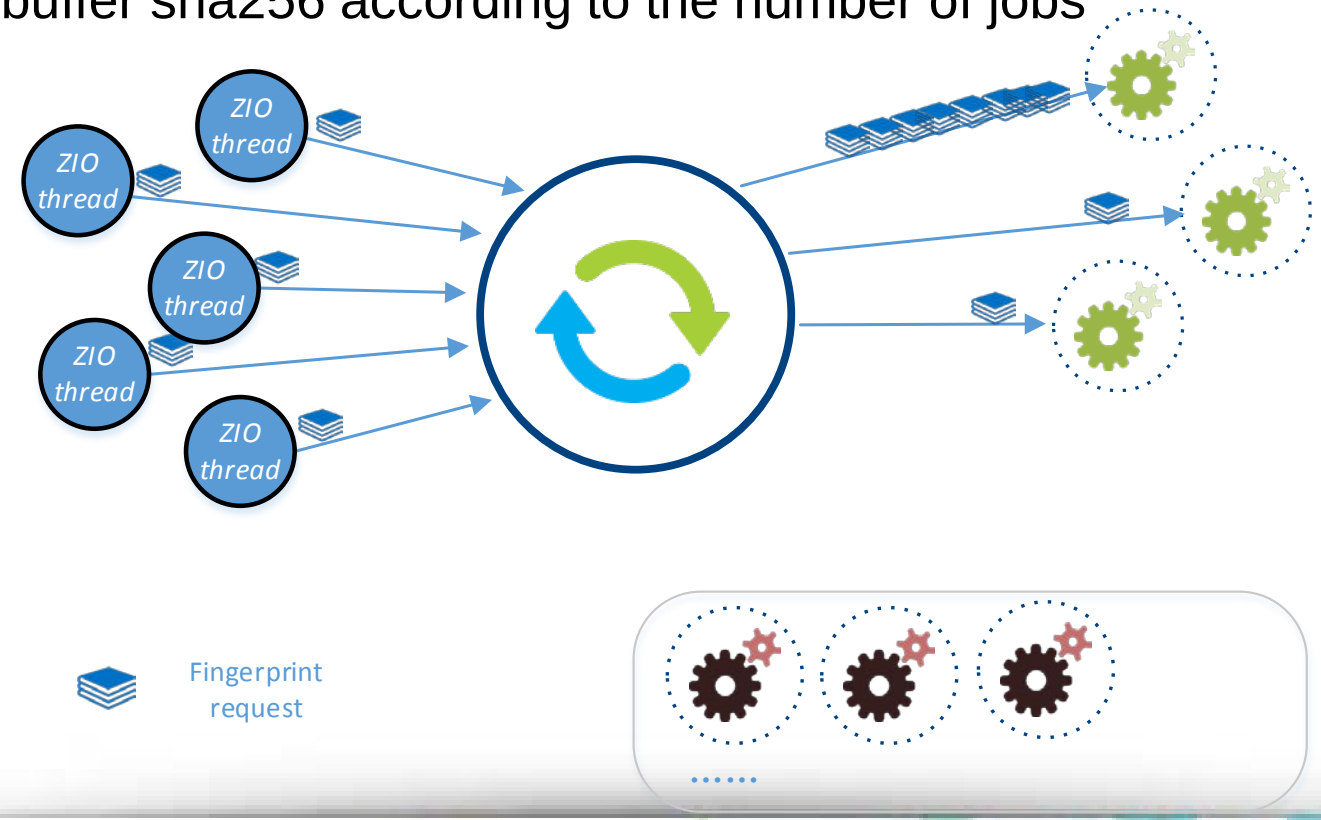
Finger Printing Accelerating in ZFS Data Deduplication

- Totally changed to asynchronize mode internally.
- Components:
  - Hash thread pool for the slave threads
  - Hash request queues with master thread attached
  - Different hash request queues for the different lengths of the job.
- Consumer-producer system:
  - Producer: ZIO threads
  - Consumer: master and slave hash threads

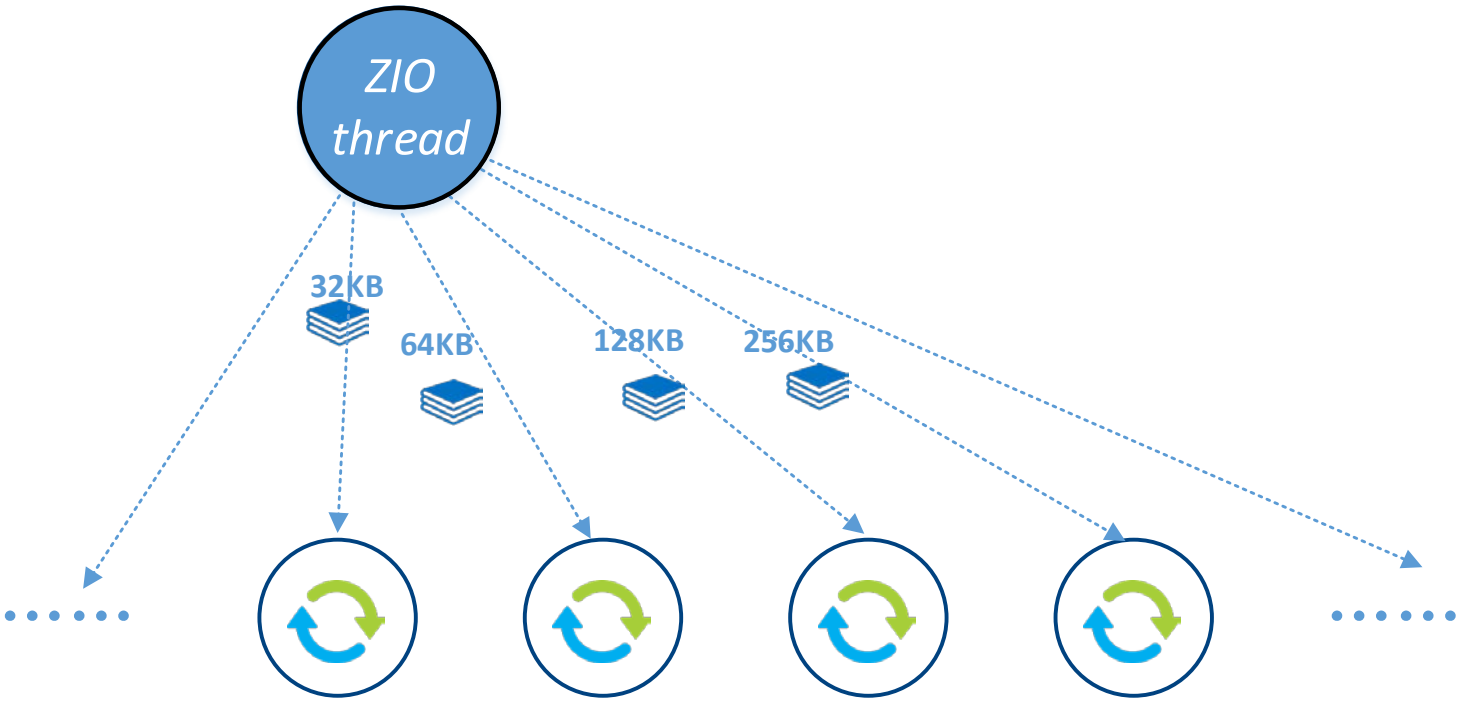


Finger Printing Accelerating in ZFS Data Deduplication

- ZIO threads submit sha256 jobs to the request queues.
- Once the jobs submitted, several threads (master and/or slave threads) will fetch and process the jobs using single-buffer and/or multi-buffer sha256 according to the number of jobs

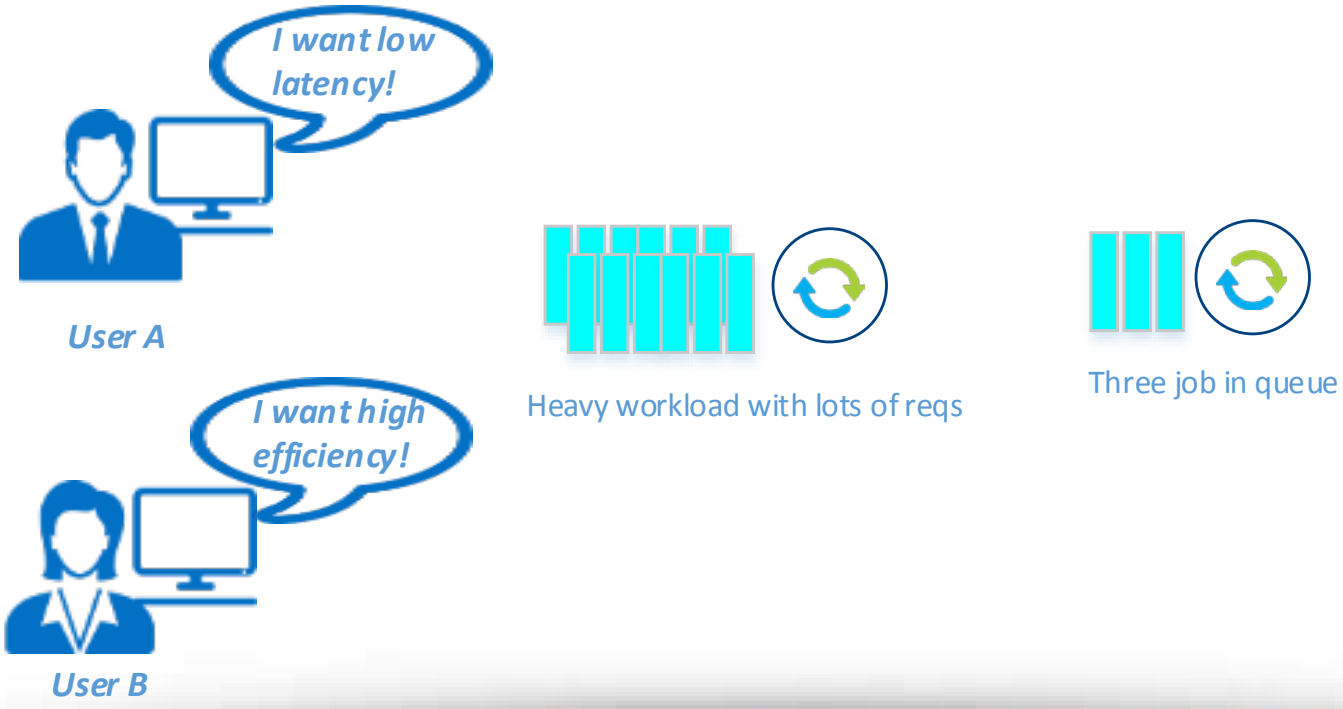


The different length of jobs



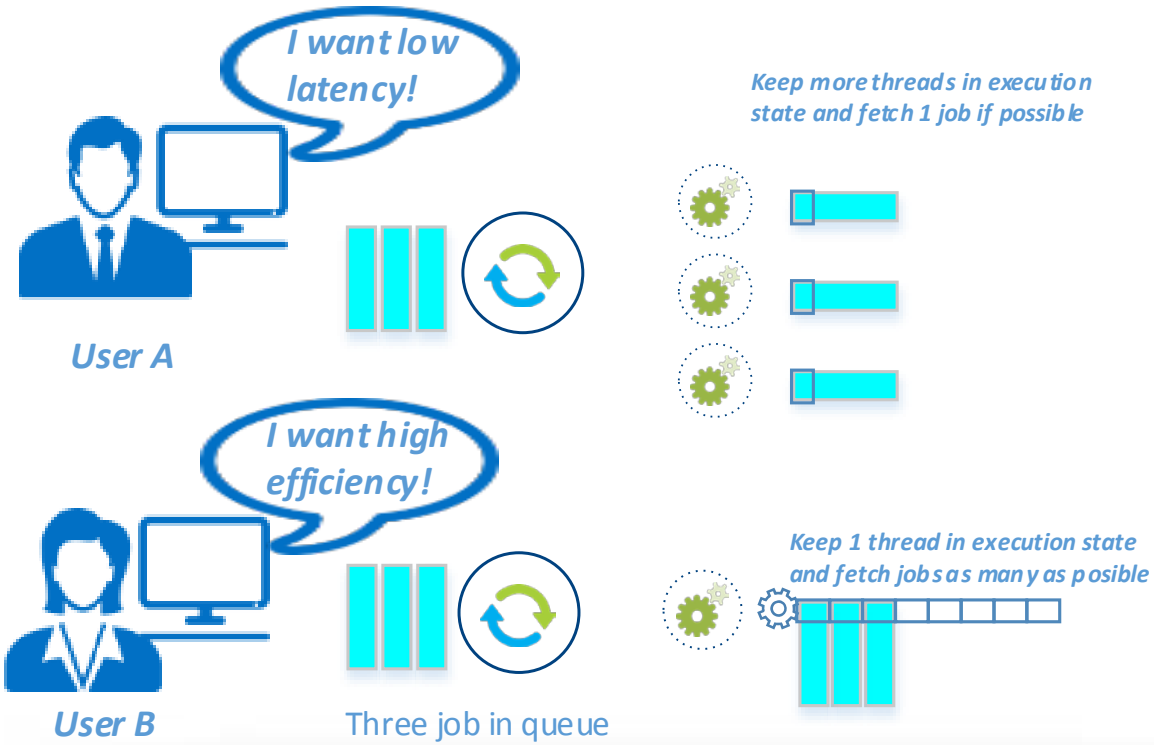
Users' Different Preference

- User A prefers each IO has low latency
- User B prefers high resource usage efficiency (low CPU utilization)



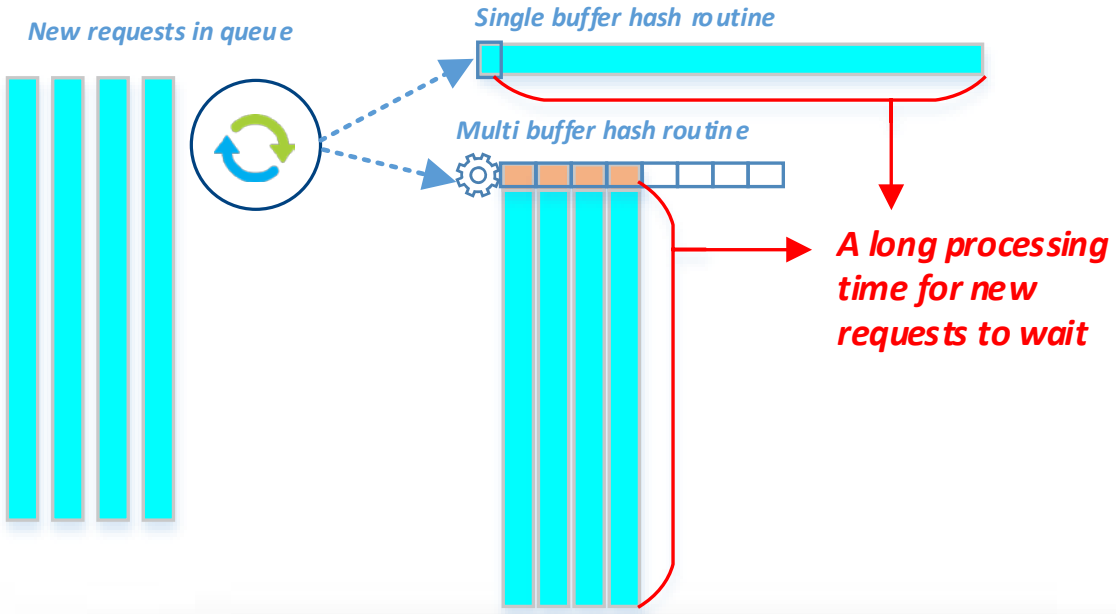
Latency Or Efficiency

- Parameter M to be used to adjust the system to get better latency or better efficiency.
- M is the number of primary threads



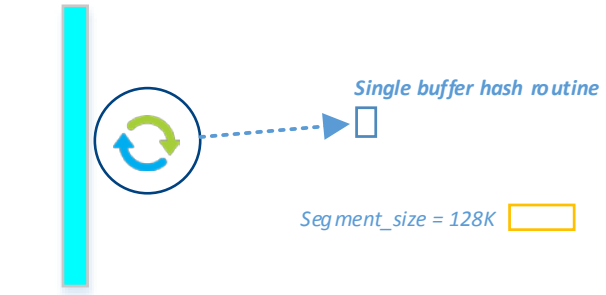
Tolerable waiting time for long length buffer.

- For longer block size, processing time is longer
- New requests in queue have to tolerate a high waiting time.

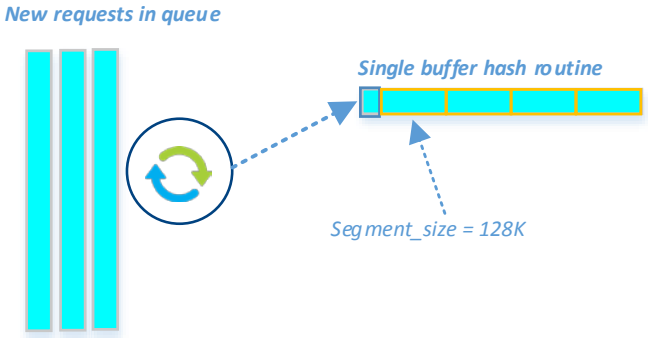


Tolerable waiting time for long length buffer.

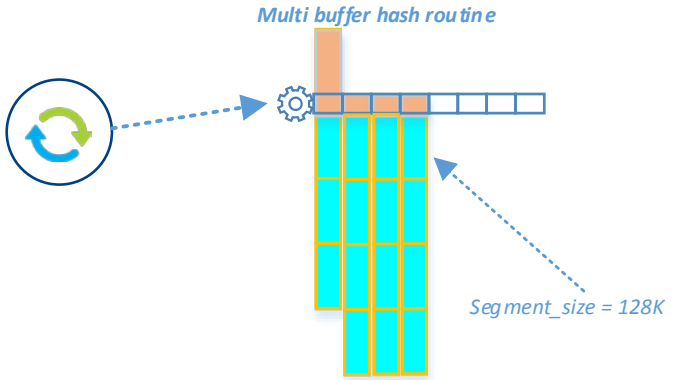
- Enable snoop mechanism and add parameter of segment size



1. snoop and single buffer process



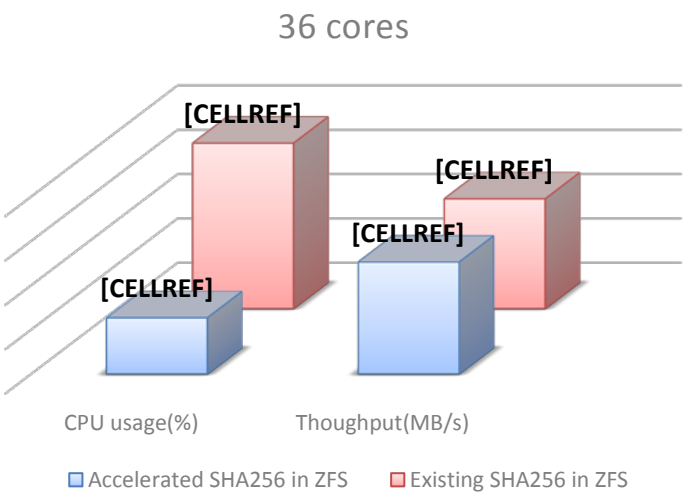
2. snoop after segment size data is processed

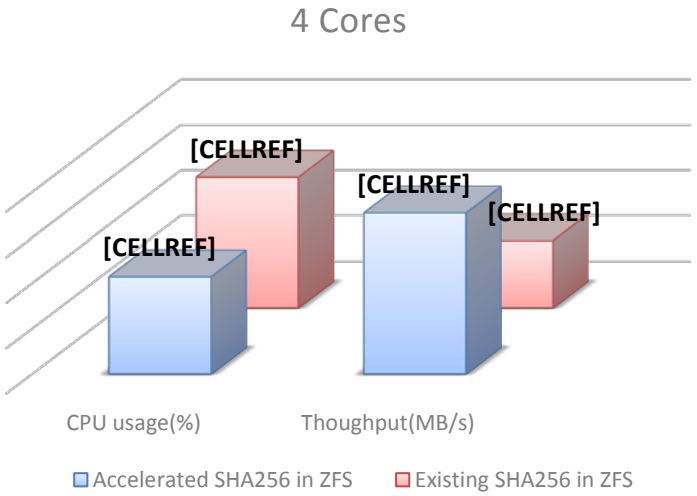


3. transform to multi-buffer hash, snoop after segment size data is processed

Performance Result of ZFS Data Deduplication

- Running Dedup workload by DEDISbench



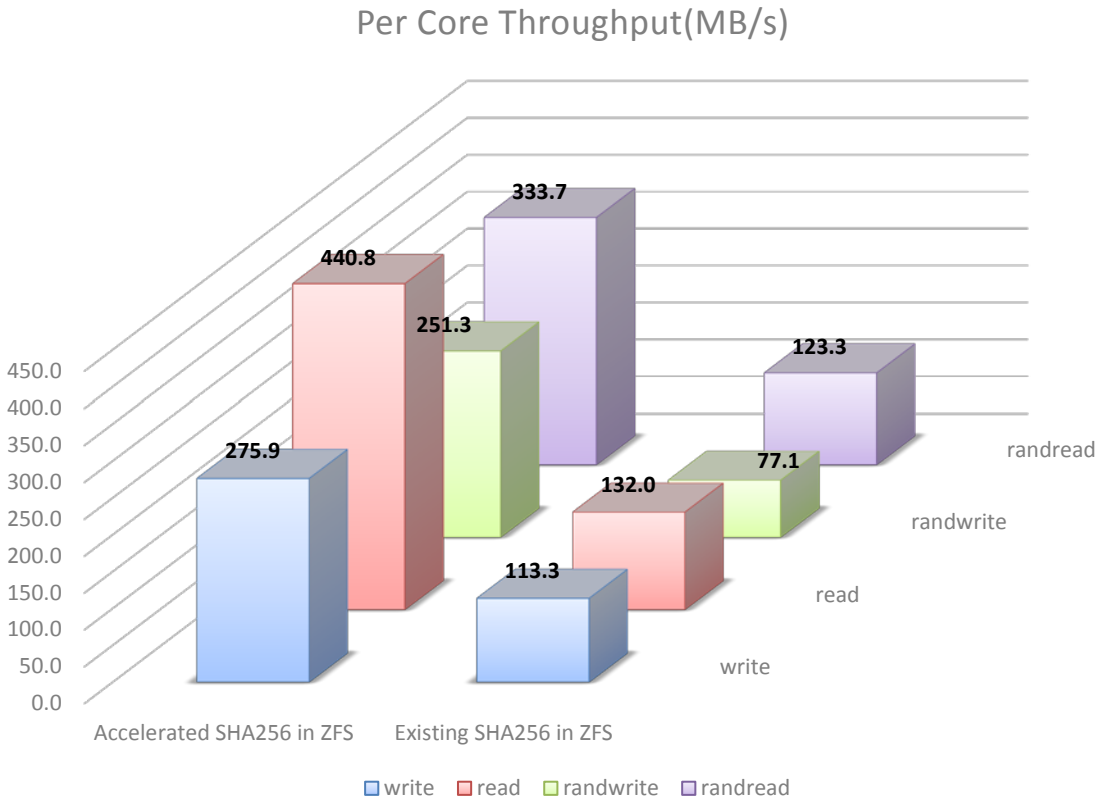


- With enough computation resource, the accelerated Sha256 in ZFS uses **1/3** CPU resource, and has same throughput. (Achieved disk boundary)
- With 4 CPU cores, the accelerated Sha256 in ZFS gets **2.4X** throughput and **3/4** CPU resource. (Achieved CPU boundary)

All results collected by Intel Corporation.

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, visit Intel Performance Benchmark Limitations (http://www.intel.com/performance/resources/benchmark_limitations.htm).

Running FIO workload



- **ZFS with the accelerated Sha256 gets 3X per core throughput**

All results collected by Intel Corporation.

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, visit Intel Performance Benchmark Limitations (http://www.intel.com/performance/resources/benchmark_limitations.htm).

Conclusion:

- Great performance improvement:
  - Saved Computation Resource x Throughput Gain >= 2.5X in ZFS data deduplication
- Our Design is a proper solution for ZFS to do finger printing in data deduplication
- Plan to upstream this design to ZFS
- It's a general framework to benefit other data deduplication applications

Thank you

Appendix

- Multi-buffer hash code:
  - Intel® ISA-L(both user mode and kernel mode)
    - https://github.com/01org/isa-l
    - https://github.com/01org/isa-l_crypto
  - LKCF*(only for kernel mode)
    - Kernel_top_path/arch/x86/crypto/sha1_mb
    - Kernel_top_path/arch/x86/crypto/sha256_mb
    - Kernel_top_path/arch/x86/crypto/sha512_mb

Intel® ISA-L performance table

ISA-L Function	Next Generation Intel® Xeon® Processor (Codenamed Skylake) @ 2.0 GHz 1 Socket			
	ISA-L		OpenSSL 1.0.2g	
	Cycle/Byte Performance (lower is better)	Single Core Throughput (higher is better)	Cycle/Byte Performance (lower is better)	Single Core Throughput (higher is better)
Multihash SHA-1**	0.46	4.2 GB/s	-	-
Multihash SHA-1 Murmur**	0.63	3.1 GB/s	-	-
Multibuffer SHA-1**	0.50	3.9 GB/s	4.45	450 MB/s
Multibuffer SHA-256**	0.93	2.1 GB/s	11.89	168 MB/s
Multibuffer SHA-512**	1.15	1.7 GB/s	7.64	262 MB/s
Multibuffer MD5**	0.34	5.8 GB/s	4.99	401 MB/s

Up to 15X bandwidth boost

Up to 13X bandwidth boost

** ISA-L function uses AVX512 instructions

All results collected by Intel Corporation.

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, visit Intel Performance Benchmark Limitations (http://www.intel.com/performance/resources/benchmark_limitations.htm).

BIOS Configuration

- P-States: Disabled
- Turbo: Disabled
- Speed Step: Disabled
- C-States: Disabled
- Power Performance Tuning: Disabled
- ENERGY_PERF_BIAS_CFG: PERF
- Isochronous: Disabled
- Memory Power Savings: Disabled

Performance test configuration for ZFS data deduplication

- ZFS version: SPL-0.6.5.6, ZFS-0.6.5.6
- Linux version: Ubuntu 14.04.4 LTS, kernel 4.4.11
- Test tool version: fio-2.2.10, DEDISbench-v1.0.0
- Hardware:
  - Intel(R) Xeon(R) CPU E5-2699 v3 @ 2.30GHz
  - MEM: DDR4, 2133MHz, 16GB * 8
  - DISK: INTEL SSD S3500 480GB * 7
- ZFS configuration:
  - zpool create pool sdb sdc sdd sde sda sdg sdh -f
  - zfs create pool/fs
  - zfs set recordsize=128k pool/fs
  - zfs set primarycache=metadata pool/fs
  - zfs set secondarycache=metadata pool/fs
  - zfs set checksum=sha256 pool/fs
  - zfs set dedup=sha256 pool/fs
- ZFS modification:
  - Change zio_taskq_batch_pc from 75 to 100 for ZFS original SHA256 performance test
  - Change zio_taskqs[WRITE][ISSUE] from ZTI_BATCH to ZTI_N(40) for ZFS sha256-mb performance test

- DEDISbench command:
  - time DEDISbench -g/etc/dedisbench/dist_highperf -w -c4 -d/pool/fs/ -b131072 -z -o./act_distrib -p -s256000 -f2560000
  - time DEDISbench -g/etc/dedisbench/dist_highperf -w -c10 -d/pool/fs/ -b131072 -z -o./act_distrib -p -s256000 -f2560000
- Fio command:

```
[global]
runtime=100
time_based
group_reporting
end_fsync=1
directory=/pool/fs
rw=write ? randwrite ? read ? randread
bs=128k
ioengine=libaio
size=2G
iodepth=64
numjobs=64
[zfstest]
thread=1
```

Referenced paper:

- Processing Multiple Buffers in Parallel to Increase Performance on Intel ® Architecture Processors
  https://www.baidu.com/link?url=W1Txipv0F-MEYbY31p9YkTDUuenJznKOzpBZi2ByRn1belmhPgodxMmE2EzZrL4uYDLkbwbyTQhSkIFrSn-EXoaGbn8K9LMKBc-FvWCFkqt5BQBS7Q0wDaQb8AxeNHbnR8Slc9FiGEST6dOk1sfEZ-eMAAx-M-G9uJaMmoBk3-C&wd=&eqid=904bfa1900030a870000000257c3bda2
- Fast Cryptographic Computation on Intel® Architecture Processors Via Function Stitching
  https://www.baidu.com/link?url=idGx1TveIgzKkXsVVRJmgaFut08N1nL6UcvpyLvT-qne1mXcptrhU88xXhXeiXIzUodQxOPa9WV77QKHy1R1NQMRMz6Kvi5lHAsSh87-iRUv5WqYNUZ93opRWNI_BQacCj8l7d-dh80B6v6ZTeOpSEGT23SVoxMLse_wwg3fwOK&wd=&eqid=fdced91e0003730f0000000257c3bdd9
- Multi-Hash A Family of Cryptographic Hash Algorithm Extensions
  https://www.baidu.com/link?url=UbNFRBE9pqrJARIXNNaInUTYS1LOZsHYqnxYJlbPXC56gozPVo_ok2c0fayCU_yj2SHWN0BLeOeLZisaGNvvhyj-3T86p7X5E85UcWBjUeEeNkK6_3hvChQvDiS7iPkxAoeTDOq4YxQX0BsLiy7tr_&wd=&eqid=b07ae58200036f170000000257c3bdfa