SDC

STORAGE DEVELOPER CONFERENCE

SNIA SANTA CLARA, 2017

New Encoding Technique to Reform Erasure Code Data Overwrite

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

- Erasure Code (EC) Data Overwrite
  - The Traditional EC Encoding
  - The New EC Encoding (Refresh EC Encoding)
- EC Creation with the Traditional EC Encoding
- Apply Refresh EC Encoding to EC Creation
- It’s not the end of refresh EC encoding

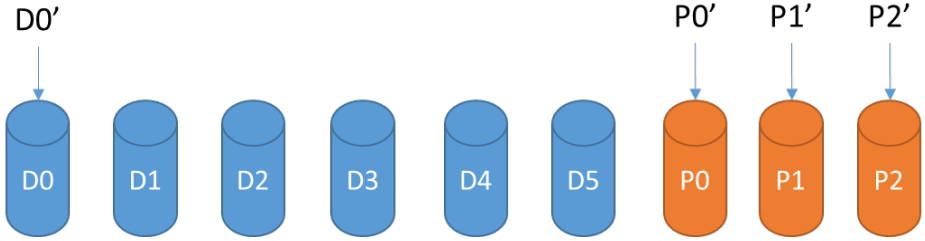
- The workload of Erasure Code
  - Data Creation
  - Data Overwrite
  - Data Recovery

The Traditional EC Encoding

- The Traditional Encoding for EC Data Overwrite

$$[D0 \quad D1 \quad D2 \quad D3 \quad D4 \quad D5] * \begin{bmatrix} \alpha_0 & \alpha_1 & \cdots & \alpha_5 \\ \beta_0 & \beta_1 & \cdots & \beta_5 \\ \gamma_0 & \gamma_1 & \cdots & \gamma_5 \end{bmatrix}^T = \begin{bmatrix} P0 \\ P1 \\ P2 \end{bmatrix}^T$$

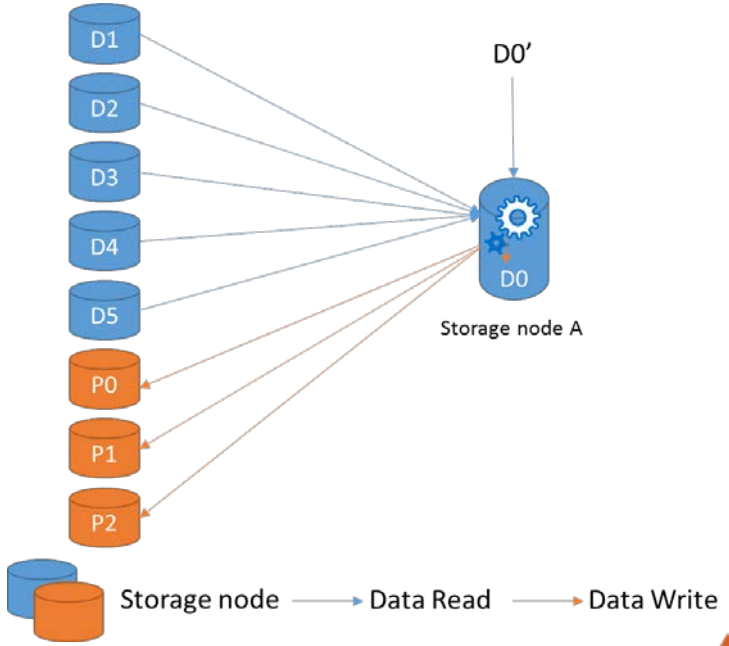
$$[D0' \quad D1 \quad D2 \quad D3 \quad D4 \quad D5] * \begin{bmatrix} \alpha_0 & \alpha_1 & \cdots & \alpha_5 \\ \beta_0 & \beta_1 & \cdots & \beta_5 \\ \gamma_0 & \gamma_1 & \cdots & \gamma_5 \end{bmatrix}^T = \begin{bmatrix} P0' \\ P1' \\ P2' \end{bmatrix}^T$$



The Traditional EC Encoding

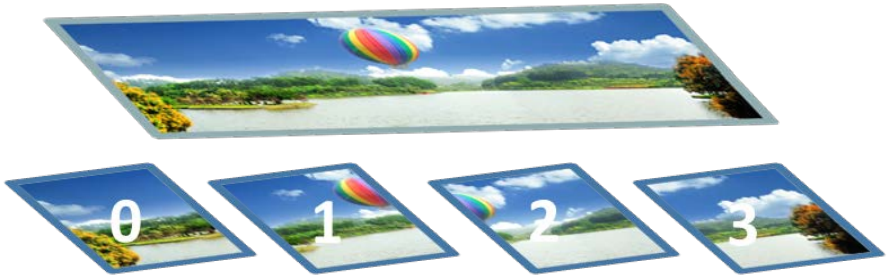
- The Traditional Workflow for EC Data Overwrite

- Number of storage nodes involved ----**9**
- Number of vectors transmitted ---- **9**
- Number of disk operation
  - Disk read ---- **5**
  - Disk write ---- **4**
- Accumulated computation cost
  - XOR ---- **6 * 3**
  - Multiply ---- **5 * 3**

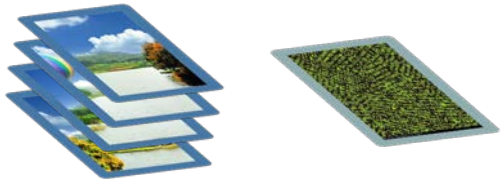


The New EC Encoding (Refresh EC Encoding)

- Refresh EC Encoding Introduction
  - Parity is overlaid by data slices

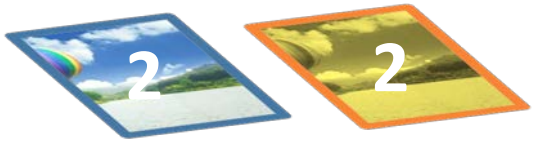


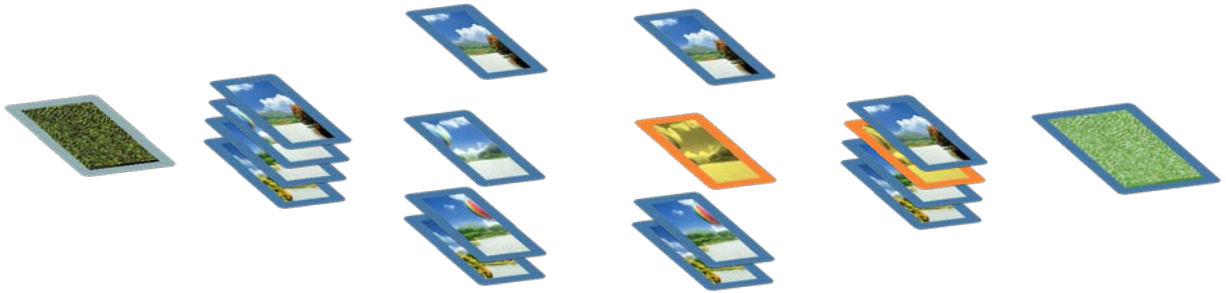
$$P_l = \xi_0 D_0 + \xi_1 D_1 + \ldots + \xi_k D_k$$



The New EC Encoding (Refresh EC Encoding)

- Just replace changed slice from each parity





$$P_l' = P_l - \xi_x D_x + \xi_x D_x'$$
$$= \xi_x (D_x' - D_x) + P_l$$
$$= \xi_x (D_x' \oplus D_x) \oplus P_l$$

The New EC Encoding (Refresh EC Encoding)

- Refresh EC Encoding for EC (k, w)
  - EC Creation with the EC encoding

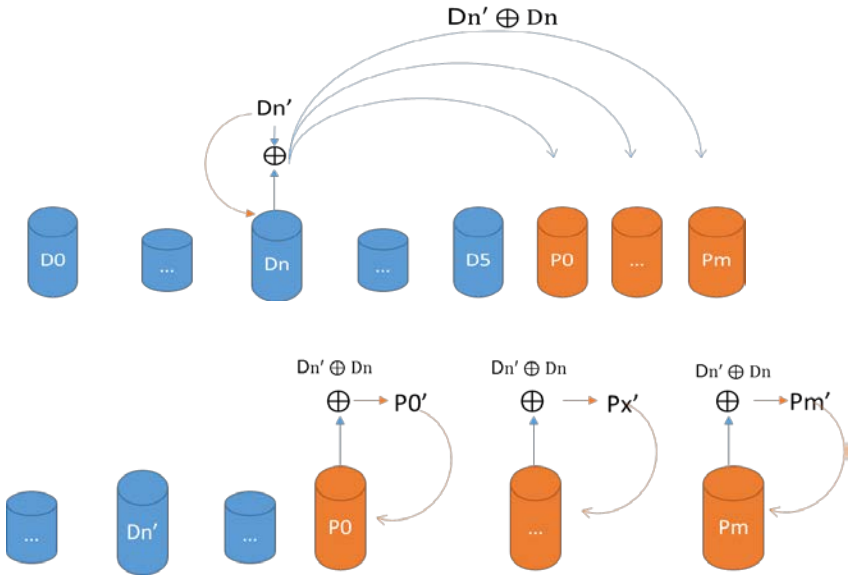
$$[D_0 \quad D_1 \quad \dots \quad D_m \quad \dots \quad D_{k-1}] * \begin{bmatrix} \alpha_0 & \alpha_1 & \dots & \alpha_{k-1} \\ \dots & \dots & \dots & \dots \\ \omega_0 & \omega_1 & \dots & \omega_{k-1} \end{bmatrix}^T = \begin{bmatrix} P_0 \\ \dots \\ P_{w-1} \end{bmatrix}^T$$

  - EC data overwrite with the refresh EC encoding

$$P_0' = \alpha_m(D'm \oplus Dm) \quad \oplus P_0$$
$$P_1' = \beta_m(D'm \oplus Dm) \quad \oplus P_1$$
$$\dots\dots$$
$$P_{k-1}' = \omega_m(D'm \oplus Dm) \quad \oplus P_{k-1}$$

The New EC Encoding (Refresh EC Encoding)

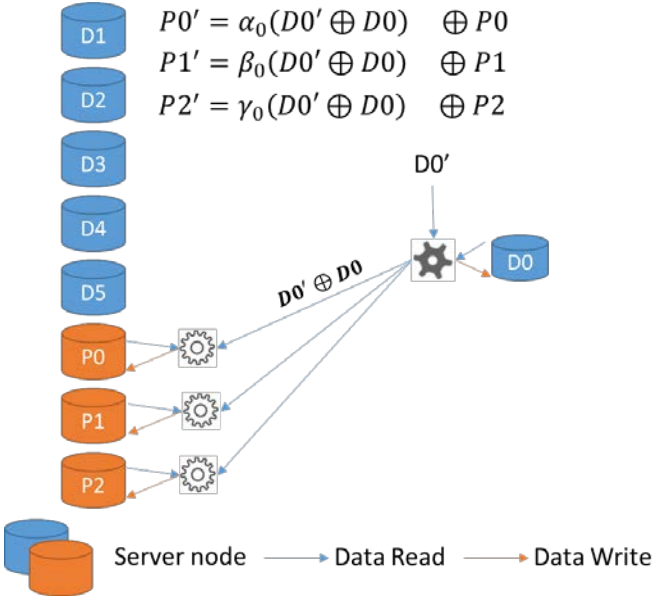
- The New Workflow for EC Data Overwrite
  - Get difference by XOR on changer
  - Replace source data vector
  - Spread difference
  - Refresh difference into parity



The New EC Encoding (Refresh EC Encoding)

- The New Workflow for EC Data Overwrite – cont.

  - Number of storage nodes involved ---- **4**
  - Number of vectors transmitted ---- **3**
  - Number of disk operation
    - Disk read ---- **4**
    - Disk write ---- **4**
  - Distributed computation cost
    - XOR ---- **1 + 1 * 3**
    - Multiply ---- **1 * 3**



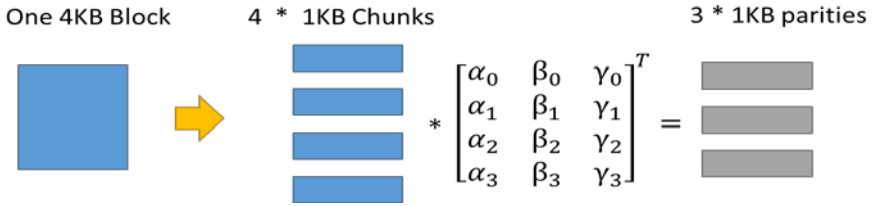
The Comparison for Typical EC Data Overwrite

EC (k, m)	6, 3		9, 3		12, 4	
mode	Refresh EC Encoding	Traditional EC Encoding	Refresh EC Encoding	Traditional EC Encoding	Refresh EC Encoding	Traditional EC Encoding
servers involved	4	9	4	12	5	16
vector transmit	4	9	4	12	5	16
GF XOR	1 + 1 * 3	5 * 3	1 + 1 * 3	8 * 3	1 + 1 * 4	11 * 4
GF multiplication	1 * 3	6 * 3	1 * 3	9 * 3	1 * 4	12 * 4
disk read	4	5	4	8	5	11
disk write	4	4	4	4	5	5

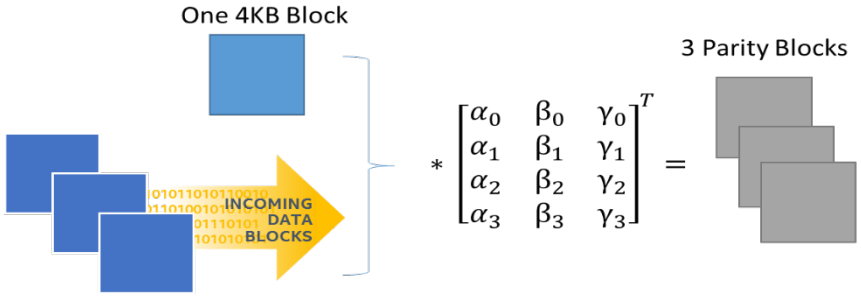
EC (k, m) data overwrite for one source vector

- The workload of Erasure Code
  - Data Creation – The Traditional EC Encoding
  - Data Overwrite – The Refresh EC Encoding
  - Data Recovery – The Traditional EC Encoding

- Method A: block self-encode



- Method B: waiting block companion encode



Apply Refresh EC Encoding to EC Creation

- Step 1

$$P_J = \alpha_0 * D_A$$

$$P_K = \beta_0 * D_A$$

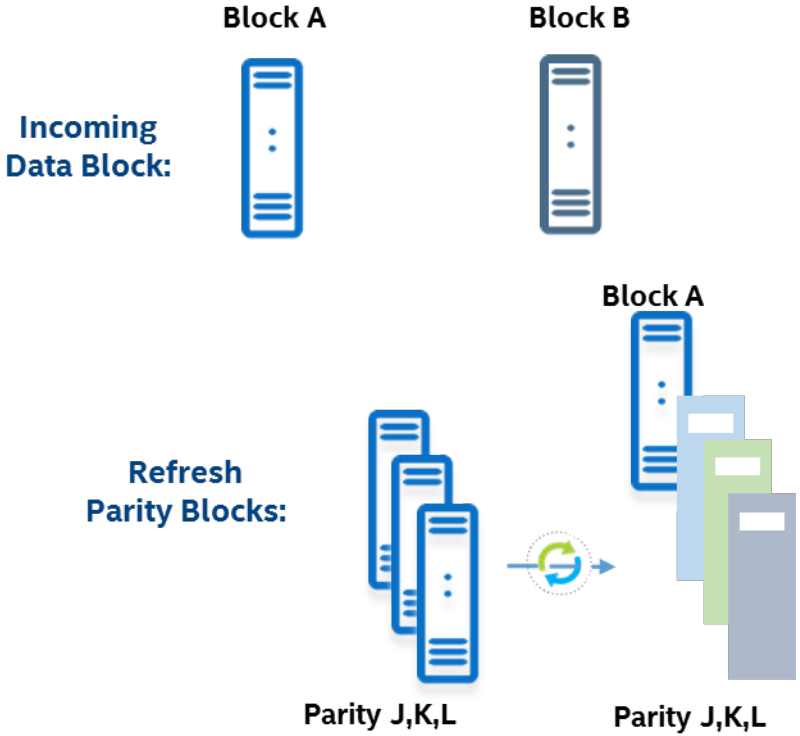
$$P_L = \gamma_0 * D_A$$

- Step 2

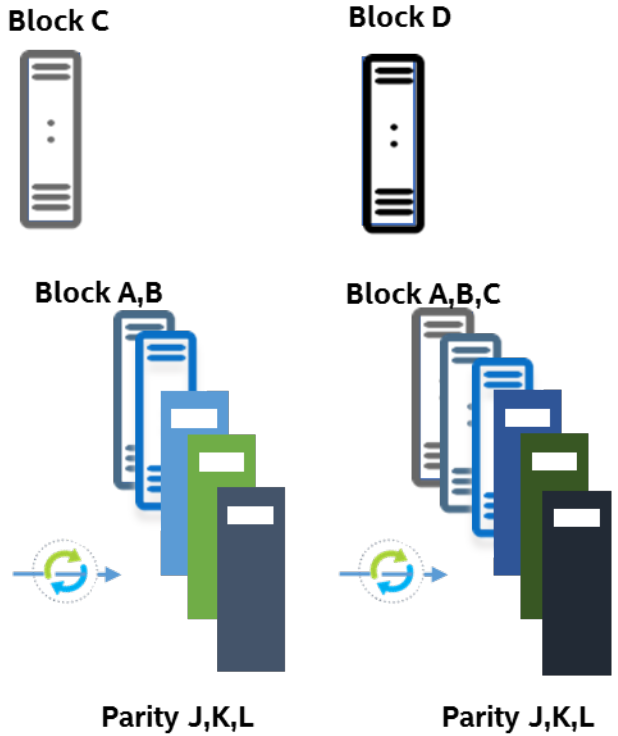
$$P_J' = P_J + \alpha_1 * D_B$$

$$P_K' = P_K + \beta_1 * D_B$$

$$P_L' = P_L + \gamma_1 * D_B$$



Apply Refresh EC Encoding to EC Creation



- Step 3

$$P_J'' = P_J' + \alpha_2 * D_C$$

$$P_K'' = P_K' + \beta_2 * D_C$$

$$P_L'' = P_L' + \gamma_2 * D_C$$

- Step 4

$$P_J''' = P_J'' + \alpha_3 * D_D$$

$$P_K''' = P_K'' + \beta_3 * D_D$$

$$P_L''' = P_L'' + \gamma_3 * D_D$$

Apply Refresh EC Encoding to EC Creation

- Benefit
  - Aligned Block Size
  - Stable Latency
  - Timely Reliability

- The workload of Erasure Code
  - Data Creation – The Refresh EC Encoding
  - Data Overwrite – The Refresh EC Encoding
  - Data Recovery – The Traditional EC Encoding

It’s not the end of refresh EC encoding

- Contradictions of EC
  - Space vs redundancy
  - Replica vs Parity
  - Gradualness vs immediacy

Thank You