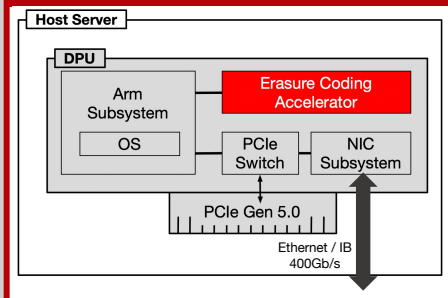
Evaluation of Erasure Coding and Opportunistic Offloading Algorithms using DPU in Distributed Storage Systems

Junghyun Ryu †, Hongsu Byun †, Myungcheol Lee ‡, Jinchun Choi ‡, Youngjae Kim †
†Department of Computer Science and Engineering, Sogang University, Seoul, South Korea
‡Electronics and Telecommunications Research Institute, Daejeon, South Korea

IEEE NVMSA 2024

NVIDIA BlueField-3 DPU & Erasure Coding



(a) An architecture of BlueField-3 DPU

(b) NVIDIA BF-3 DPU

Processor	ARM Cortex-A78 (16-core, 2.0GHz)
Memory	DDR5, 32GB
Network Interface	Ethernet / ConnectX-7
OS	Ubuntu 22.04

(c) BF-3 DPU's Specification

Figure 1. < Details of NVIDIA BlueField-3 DPU >

- The Data Processing Unit (DPU) enhances overall system performance by performing data processing tasks independently from the CPU or GPU[1]. DPUs can accelerate specific tasks using its hardware accelerators. such as erasure coding, encryption, and compression.

1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1
x1	x2	x3	x4
x5	x6	x7	x8

Generator Matrix

X

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

Original Data

=

1	2	3	4	→ Data 1
5	6	7	8	→ Data 2
9	10	11	12	→ Data 3
13	14	15	16	→ Data 4
c1	c2	c3	c4	→ Parity 1
c5	c6	c7	c8	→ Parity 2

Encoded Blocks

Figure 2. < Erasure Coding with matrix multiplication >

Figure 3. < Comparing execution time of Erasure Coding and replication method [2] >

Figure 4.< Storage systems that implementing Erasure Coding >

- Erasure Coding is a computation-intensive task based on matrix multiplication. (Figure3, 4)
- Given little research on using DPUs for Erasure Coding, our goal is to reduce the workload on the CPU by offloading Erasure Coding to DPUs.

Preliminary Result

We compared the execution time of Erasure Coding on a **CPU** and a **DPU** as a preliminary result.

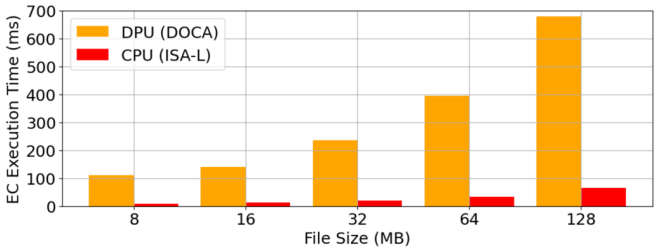


Figure 5. < Erasure Coding (k=128, p=32) Host CPU vs. DPU >

- We used an Intel Xeon Silver 4410y (48-core, 3.9 GHz) as the host server's CPU and compared it with the NVIDIA BuleField-3 DPU.
- The DPU was 13.3× slower than the CPU for an 8MB file, and 10.5× slower for a 128MB file.
- This finding indicates that unconditional offloading is not beneficial.

Resource-aware Offloading Algorithm

Unconditional offloading is not beneficial; instead, offloading should be considered opportunistically when the CPU is busy. Therefore, we propose a deterministic algorithm that is resource-aware of the host.

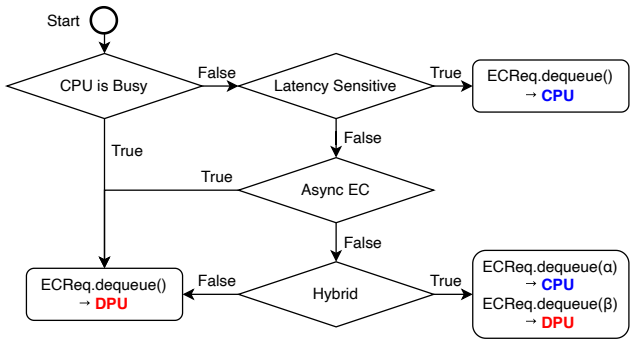


Figure 6. < Offloading algorithm execution flow >

The algorithm flowchart in Figure 6 determines which location to dequeue a job from the queue of Erasure Coding requests.

① In latency-sensitive scenarios, Erasure Coding is executed on the CPU to achieve the fastest response.
② If not, the DPU is asynchronously requested to perform Erasure Coding, allowing the Host Server to focus on other tasks.
③ Lastly, hybrid execution can be requested, where α and β represent the number of Erasure Coding tasks executed by the CPU and DPU, respectively, with their values dependent on system performance.

Future Work

Based on the insights from the preliminary results and the algorithm design, we plan to enhance this work with the following methodologies.

1. Apply DPUs and the Erasure Coding offloading algorithm to a working distributed storage system and improve the whole system performance. Figure 7 illustrates HDFS leveraging the DPU.

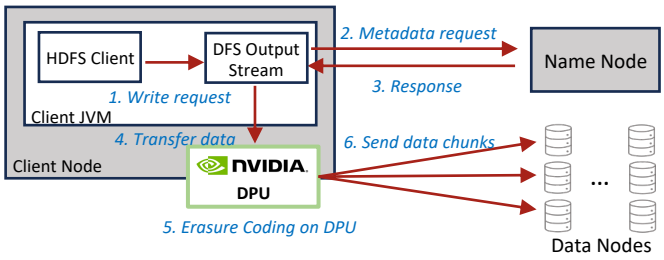


Figure 7. < HDFS write & EC scenario with DPU >

2. Break down the execution steps of Erasure Coding to fully utilize both ARM processor and hardware accelerators of DPUs. The design is as follows: the ARM processor performs tasks such as initializing DOCA and message buffer, splitting data chunks, and creating the generator matrix, and the hardware accelerator performs matrix multiplication.

[1]"NVIDIA DOCA Overview." https://docs.nvidia.com/doca/sdk/nvidia+doca+overview/index.html, 2024.
[2] "Evaluation erasure coding hadoop 3." https://db-blog.web.cern.ch/blog/emil- kleszcz/2019- 10- evaluation- erasure- coding- hadoop- 3, 2019.

Acknowledgement

This work was supported by Institute for Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No. 2021-0-00136, Development of Big Blockchain Data Highly Scalable Distributed Storage Technology for Increased Applications in Various Industries).