Enabling Data Processing under Erasure Coding in the Fog
Jad Darrous, Shadi Ibrahim

To cite this version:
Jad Darrous, Shadi Ibrahim. Enabling Data Processing under Erasure Coding in the Fog. ICPP 2019 - 48th International Conference on Parallel Processing, Aug 2019, Kyoto, Japan. pp.1. hal-02388835
Erasure Coding as alternative to replication

- **Half** the storage overhead of replication, for RS(6, 3)
- **Low** CPU overhead for encoding/decoding (5.3 GB/s)[5]
  - EC has been deployed in storage and caching systems[7]
  - HDFS is now equipped with EC since the 3.0.0 release[6]
- When performing MapReduce applications in data-intensive clusters
  - Unlink replication, most of the (map) task input data is transferred
  - **A reduction** by half of the network traffic and disk accesses when writing the output data

How to effectively realize EC for big data processing in Fog?

Quantitatively analyze the impact of heterogeneity

A Fog infrastructure is emulated on 10 machines each representing a Fog site

- Sort application performs **35% faster** under Replication compared to EC
- The main reason behind the low performance of EC is the **heterogeneity of the network**
- Map tasks **wait for the last chunk** to process the current piece of data
- This leads to **high variation** in map runtimes under EC (60%)