中国科学技术大学
University of Science and Technology of China

Update-friendly Encoding from Replication to Erasure Coding in Clustered File Systems

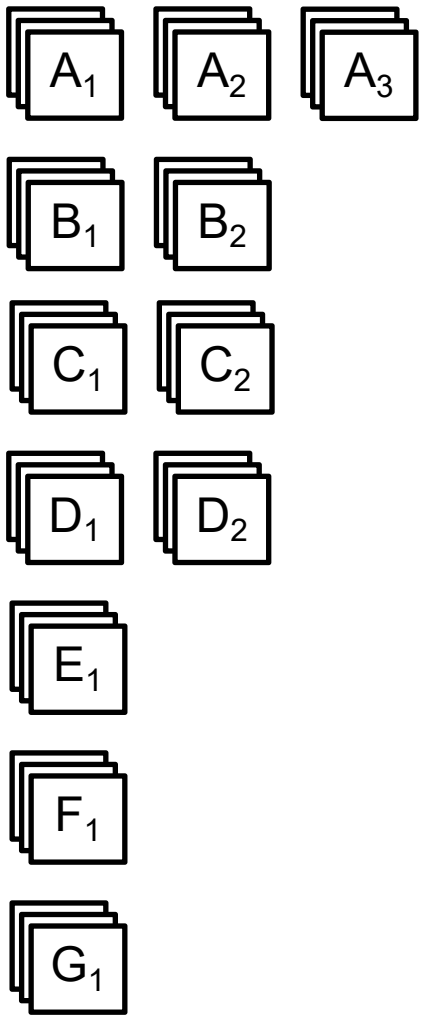
Wei Wang, Min Lyu and Yinlong Xu

University of Science and Technology of China

Motivation

- Hot data-Replication
  - High performance
- Cold data-Erasure Codes
  - Low storage overhead
- Update EC files → Regenerate parity blocks
- Different encoding schemes induce different update costs
  - Existing encoding induce high update costs

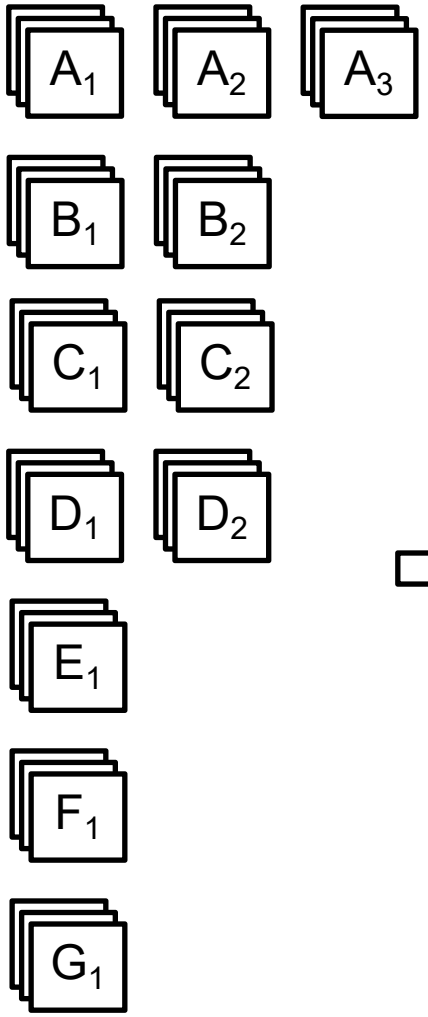
3-Replication to (3,2)-Erasure Code



3-replication

Data block

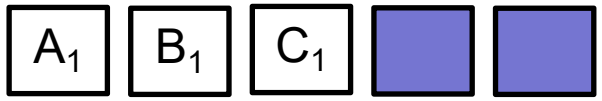
3-Replication to (3,2)-Erasure Code



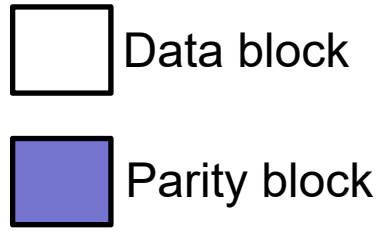
3-replication

➢ Randomly select blocks for a stripe

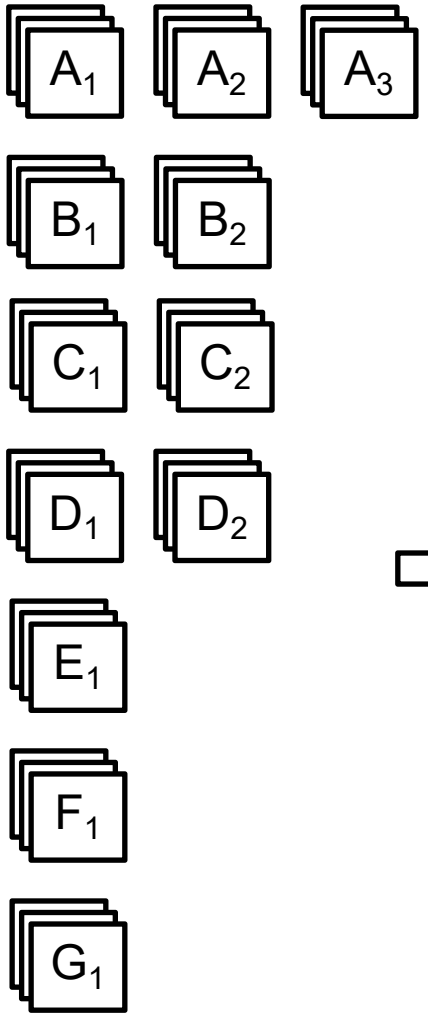
Existing method



(3,2)-EC



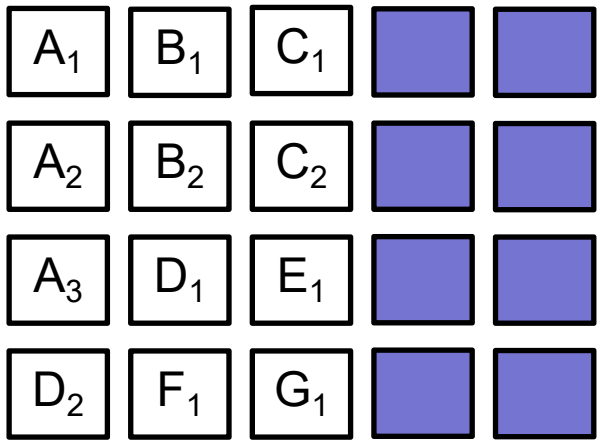
3-Replication to (3,2)-Erasure Code

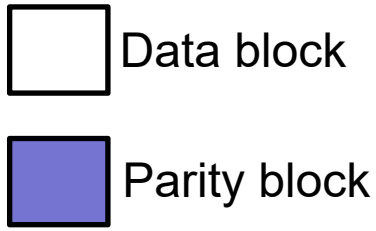


3-replication

- Randomly select blocks for a stripe

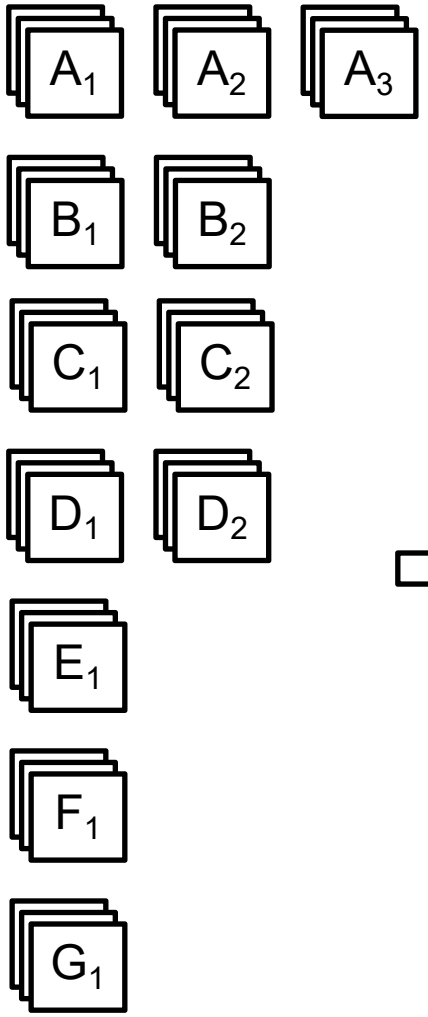
Existing method





(3,2)-EC

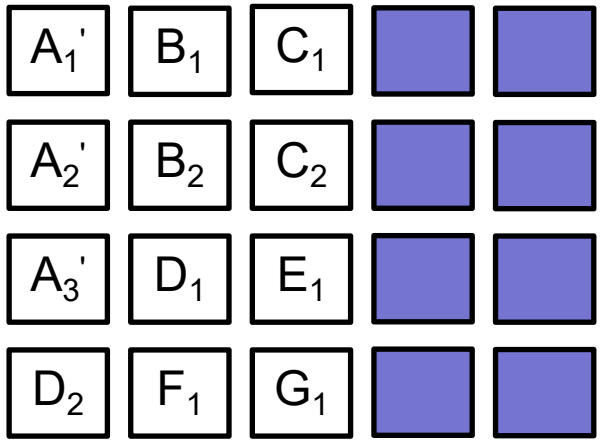
3-Replication to (3,2)-Erasure Code



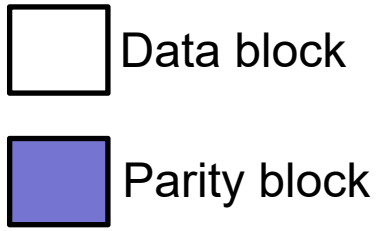
3-replication

➢ Randomly select blocks for a stripe

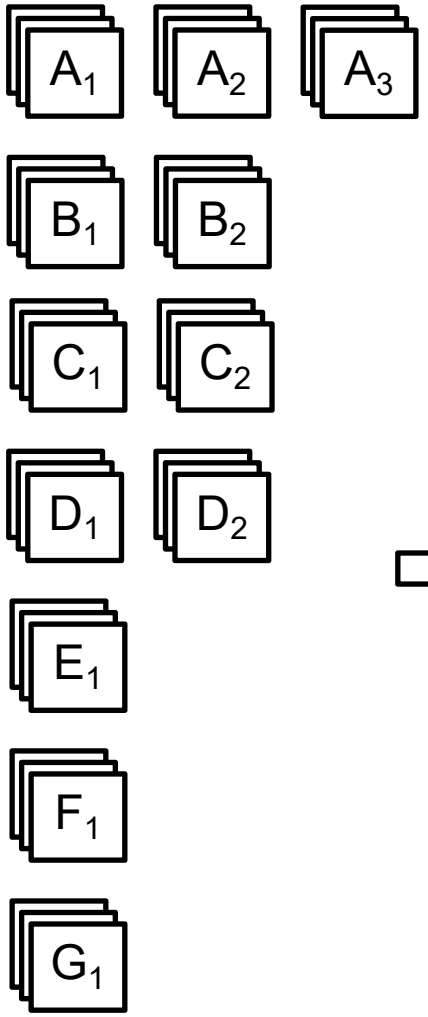
Existing method



(3,2)-EC



3-Replication to (3,2)-Erasure Code

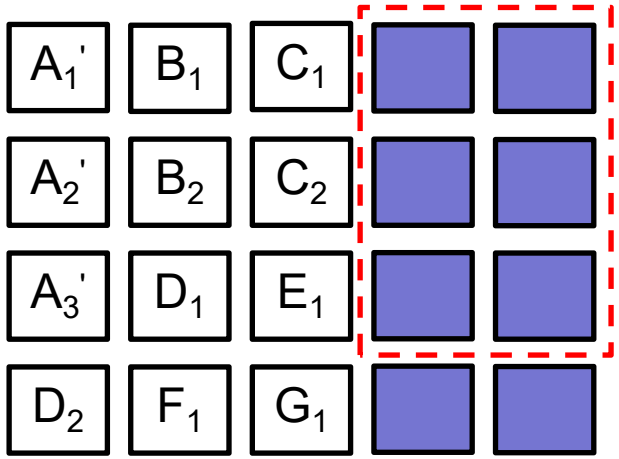


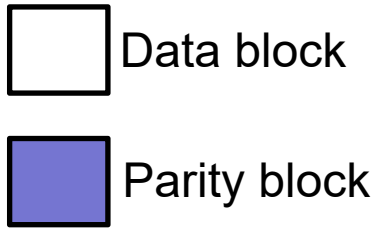
3-replication

- Randomly select blocks for a stripe

Regenerate 6 parity blocks
Disk I/O: 15 blocks
Network transfer: 12 blocks

Existing method

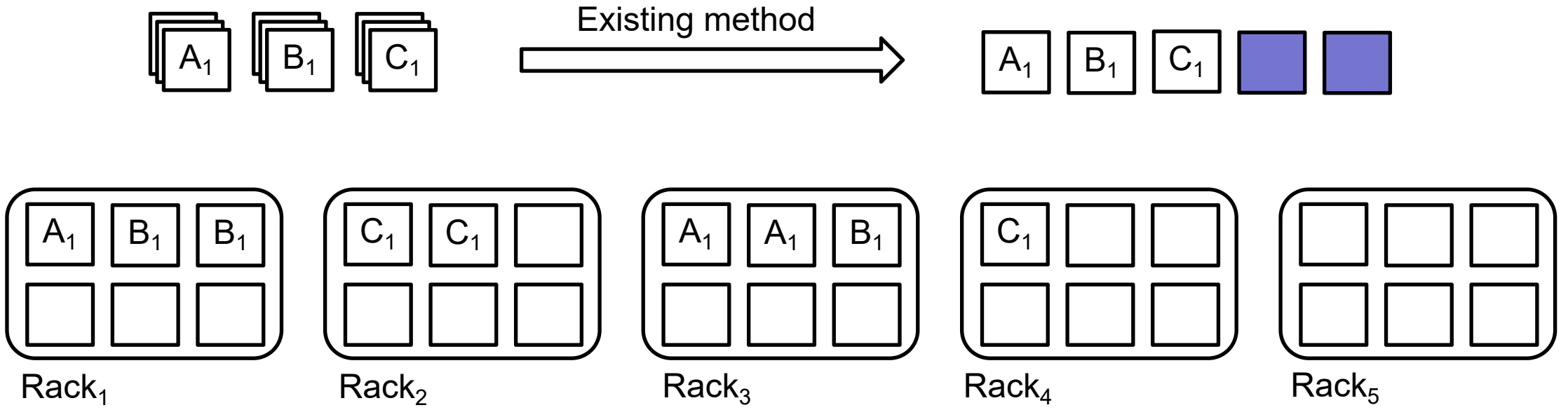




High update cost of erasure-coded files

(3,2)-EC

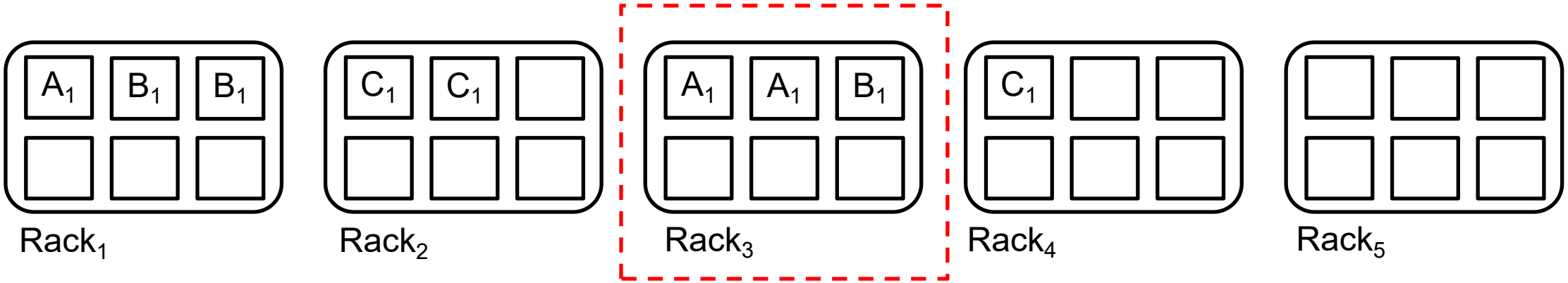
Encoding induces cross-rack traffic



- 5 racks, eack rack consists of 6 nodes
- Encode A_1, B_1, C_1 from 3-Replication to (3,2)-EC

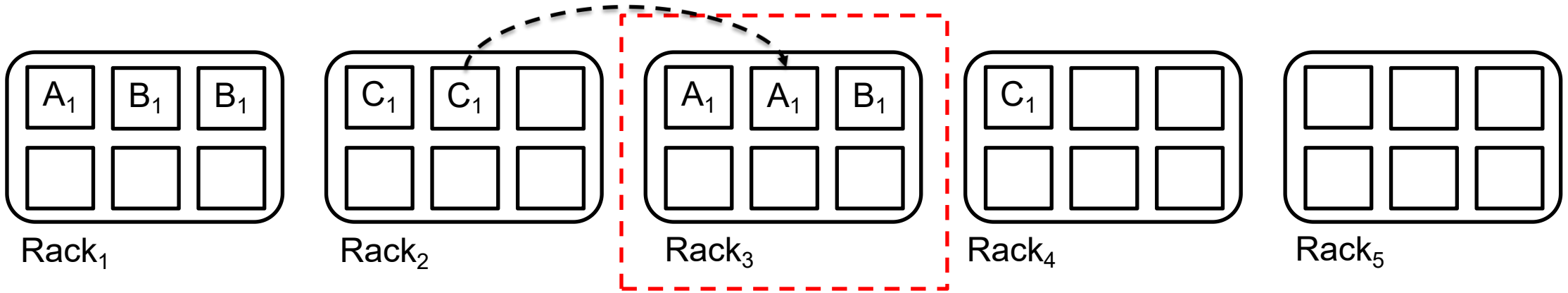
Encoding induces cross-rack traffic

1. Select a rack as the encoding rack randomly



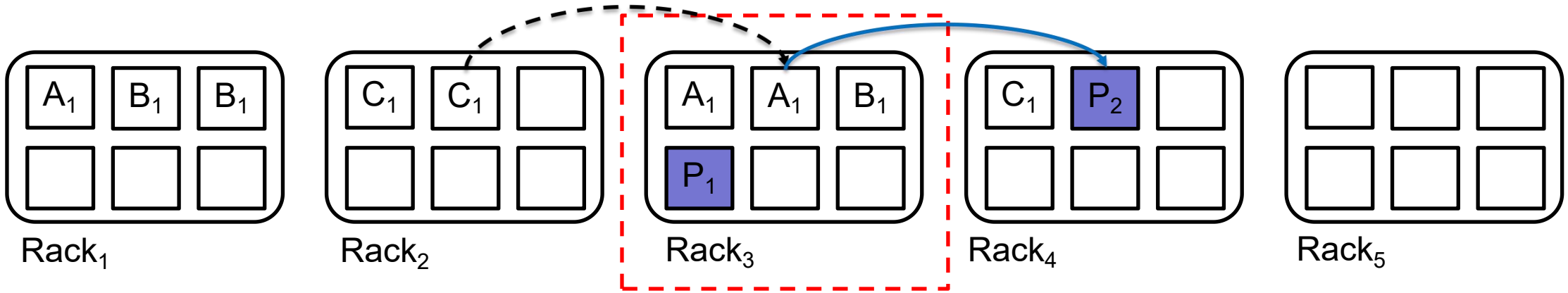
Encoding induces cross-rack traffic

1. Select a rack as the encoding rack randomly
2. Read blocks across racks



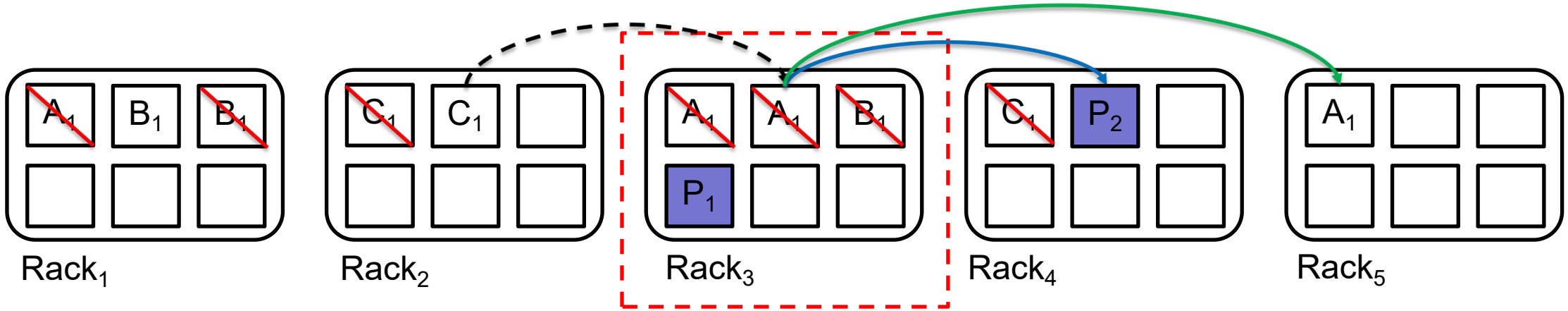
Encoding induces cross-rack traffic

1. Select a rack as the encoding rack randomly
2. Read blocks across racks
3. Encode and write parity blocks



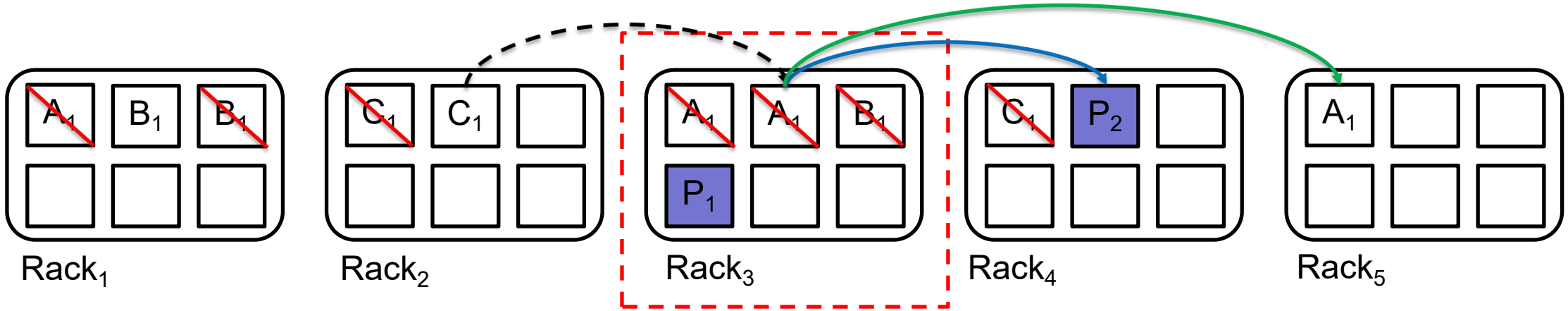
Encoding induces cross-rack traffic

1. Select a rack as the encoding rack randomly
2. Read blocks across racks
3. Encode and write parity blocks
4. Remove redundant replicas and relocate data blocks→tolerate rack failures



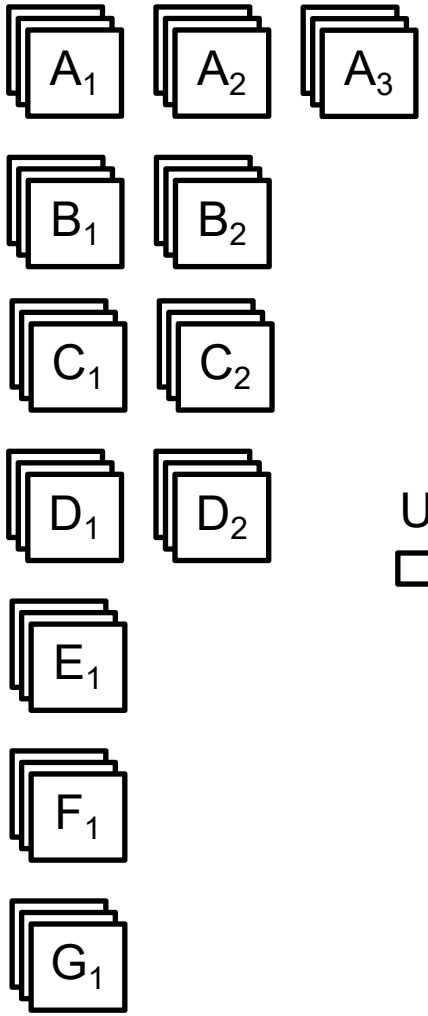
Encoding induces cross-rack traffic

1. Select a rack as the encoding rack randomly
2. Read blocks across racks
3. Encode and write parity blocks
4. Remove redundant replicas and relocate data blocks→tolerate rack failures



Update-friendly Encoding(UFE)

UFE - Main Idea



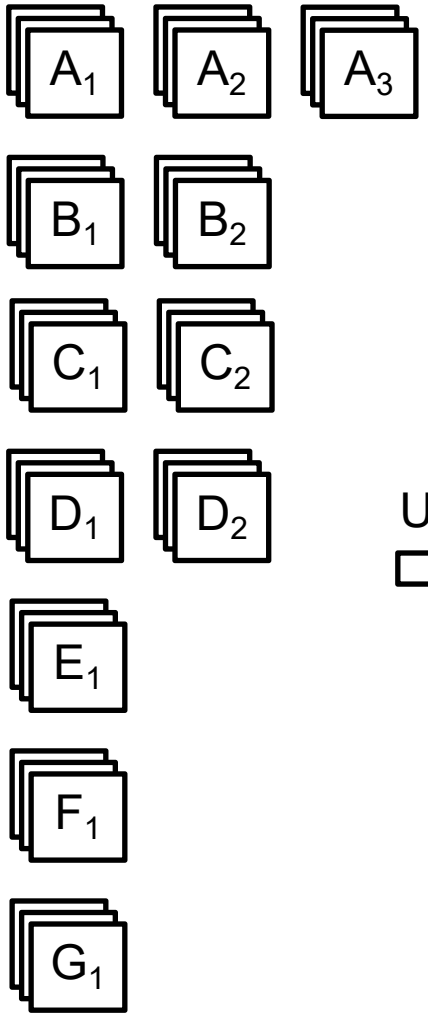
- Select data blocks of a stripe from as few files as possible

Update-friendly Encode

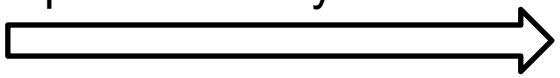
Data block

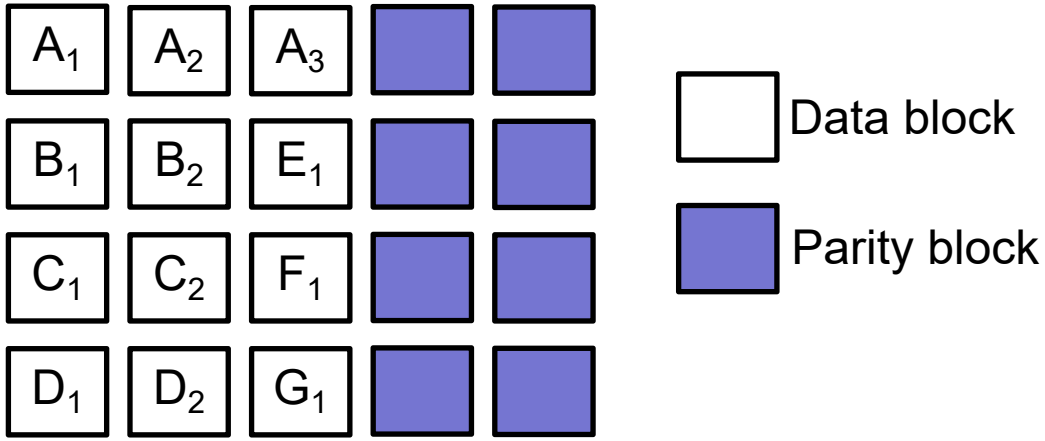
Parity block

UFE - Main Idea

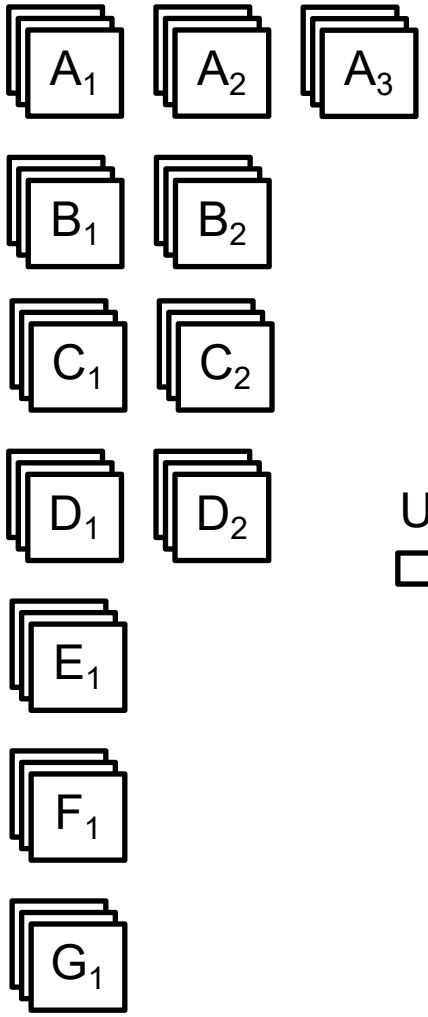


- Select data blocks of a stripe from as few files as possible





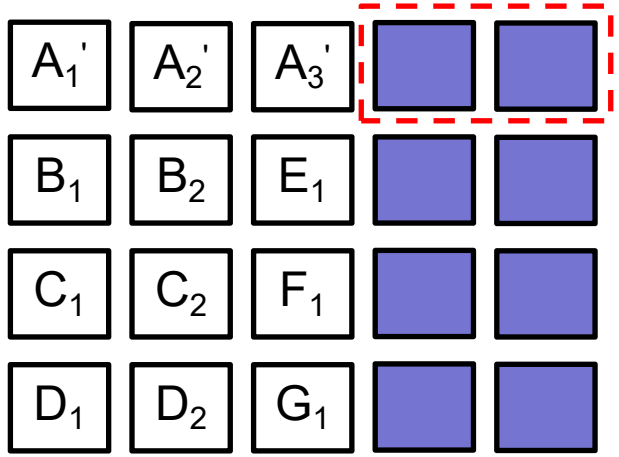
UFE - Main Idea

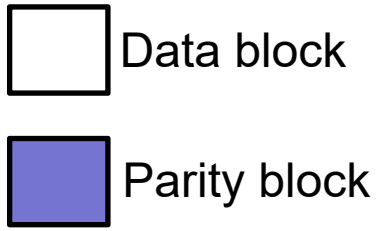


- Select data blocks of a stripe from as few files as possible

Regenerate 2 parity blocks
Disk I/O: 5 blocks
Network transfer: 4 blocks

Update-friendly Encode

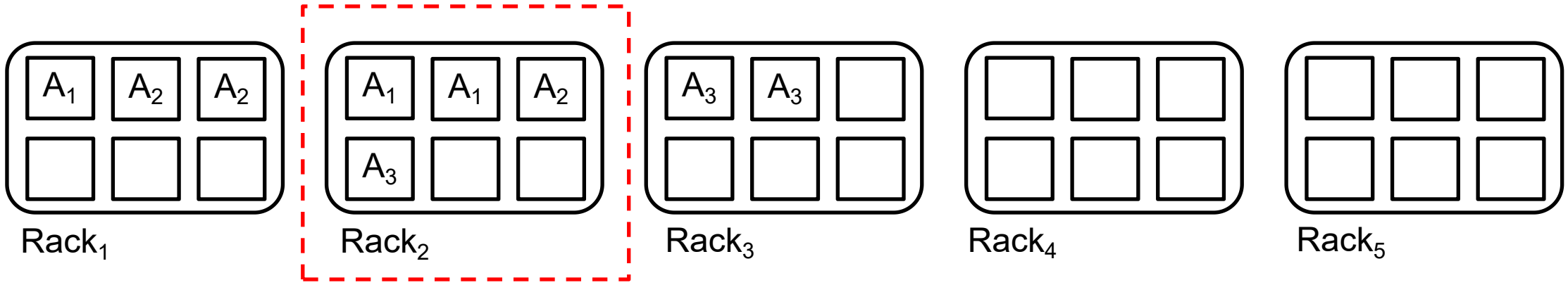




UFE - Main Idea

Encoding steps (3-replication to (n,m)-EC):

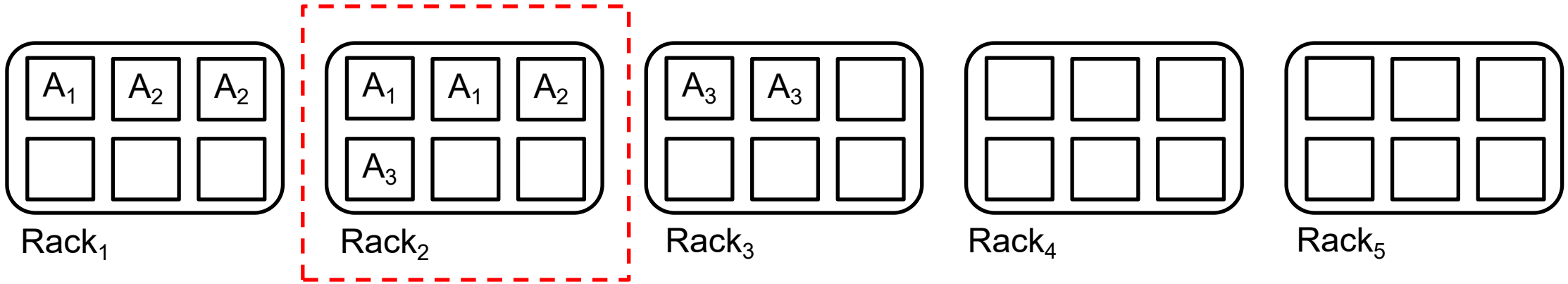
- Select the rack holding the most blocks of a stripe as the encoding rack



UFE - Main Idea

Encoding steps (3-replication to (n,m)-EC):

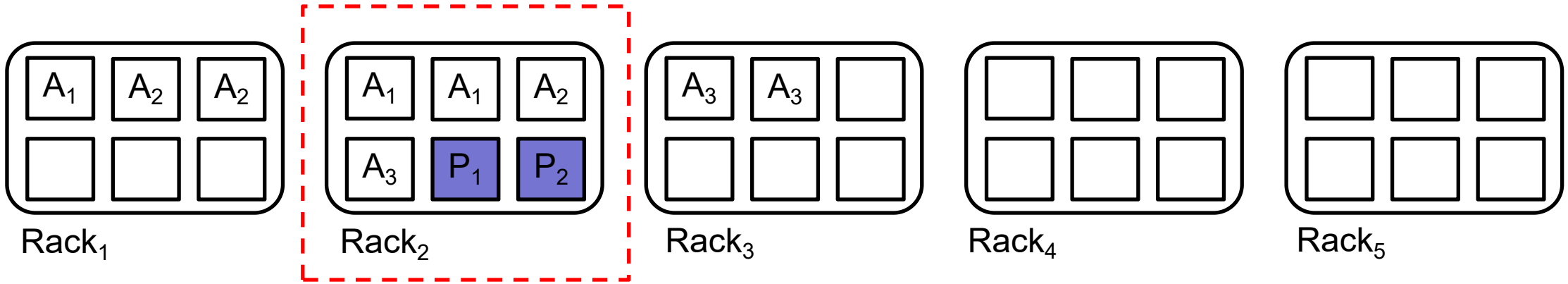
- Select the rack holding the most blocks of a stripe as the encoding rack
  - Minimize cross-rack read



UFE - Main Idea

Encoding steps (3-replication to (n,m)-EC):

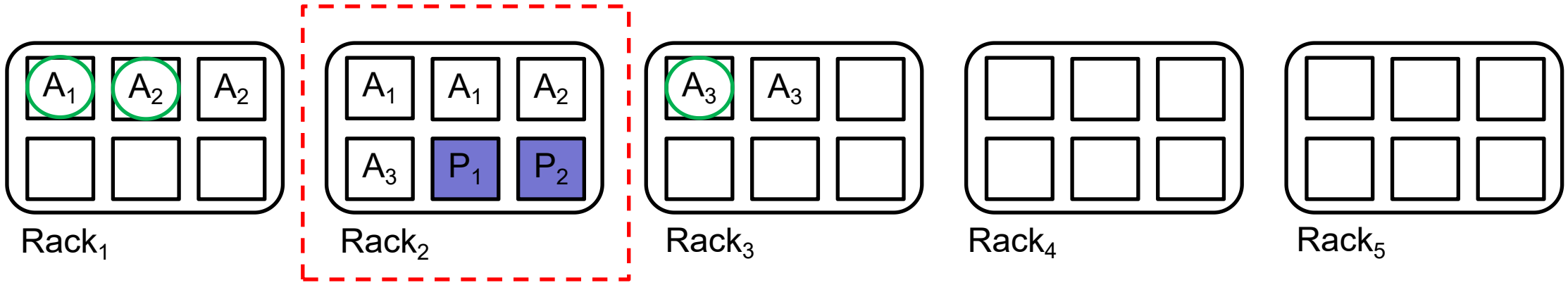
- Select the rack holding the most blocks of a stripe as the encoding rack
  - Minimize cross-rack read
- Generate and store parity blocks in the encoding rack



UFE - Main Idea

Encoding steps (3-replication to (n,m)-EC):

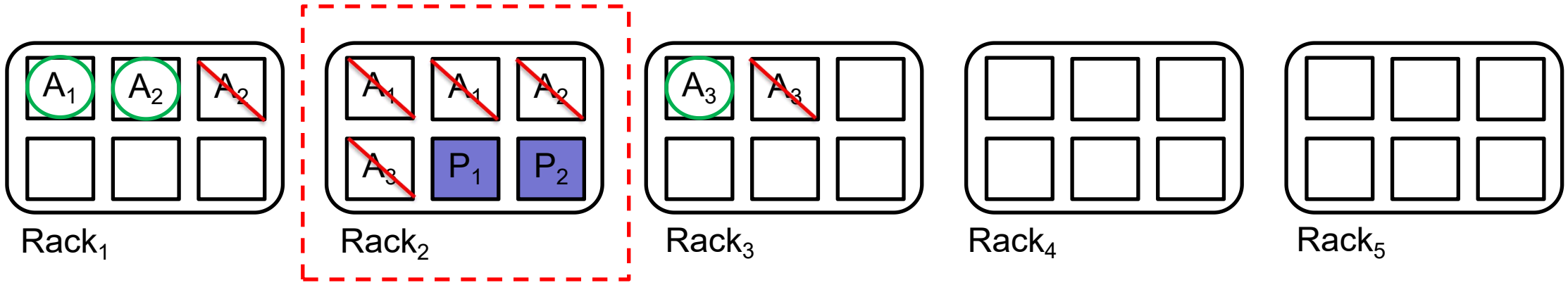
- Select the rack holding the most blocks of a stripe as the encoding rack
  - Minimize cross-rack read
- Generate and store parity blocks in the encoding rack
- Select racks to store the stripe
  - Tolerate single-rack failures: each rack stores no more than m blocks of a stripe



UFE - Main Idea

Encoding steps (3-replication to (n,m)-EC):

- Select the rack holding the most blocks of a stripe as the encoding rack
  - Minimize cross-rack read
- Generate and store parity blocks in the encoding rack
- Select racks to store the stripe
  - Tolerate single-rack failures: each rack stores no more than m blocks of a stripe
- Remove other replicas



Preliminary Experiment Results

- Configrations
  - 10 racks, each rack consists of 3 nodes
  - Encoding 6000 files from 3-Replication to (6,3)-EC
- Results
  - UFE reduces the deletion cost by 77% and cross-rack traffic by 76%

	Deletion cost(blocks)		**Encoding cost(blocks)**
	Read	Write	Cross-rack traffic
UFE	3.2	2.7	1.3
HDFS	14.0	8.4	5.5

Following Work

- Deploy UFE in real Clustered File Systems
- Metadata management
- Experiment

That's all! Thank you!