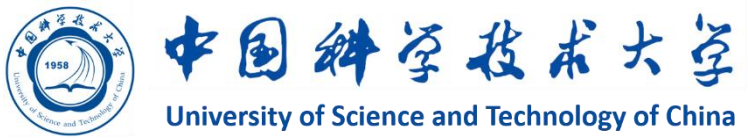


SelectiveEC: Selective Reconstruction in Erasure-coded Storage Systems

Liangliang Xu, Min Lyu, Qiliang Li, Lingjiang Xie, and Yinlong Xu

University of Science and Technology of China

HotStorage 2020

Distributed Storage Systems (DSSes)

- Data is important
  - Large scale
  - Exponential growth
- DSSes are the core infrastructures
  - Thousands of nodes
  - “Fat node”
    - Up to 72 TB of storage (about 1.5M chunks) per node in Pangu[1]
  - Frequent failures

Disk faults

Cluster crushed

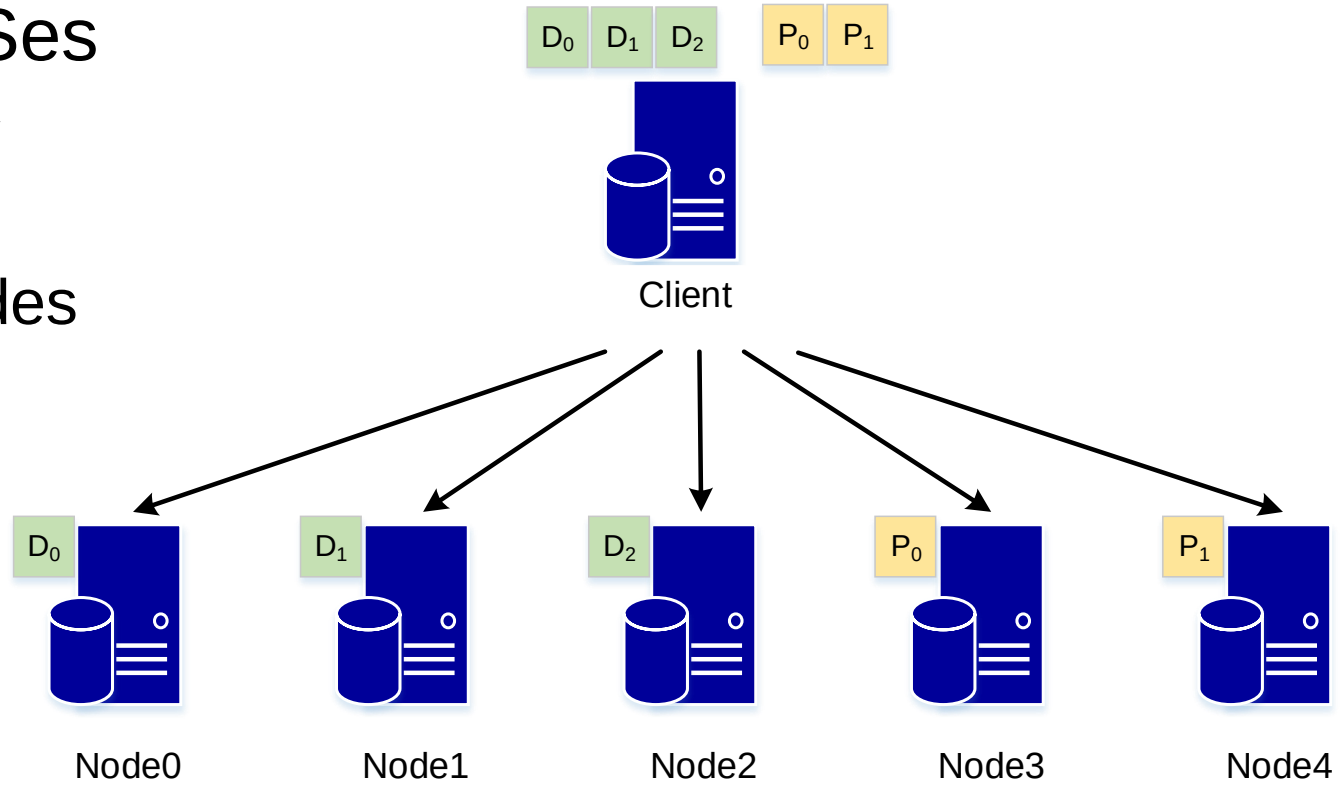
Network failures

Artificial errors

[1] ATC2019: Dayu: Fast and Low-interference Data Recovery in Very-large Storage Systems

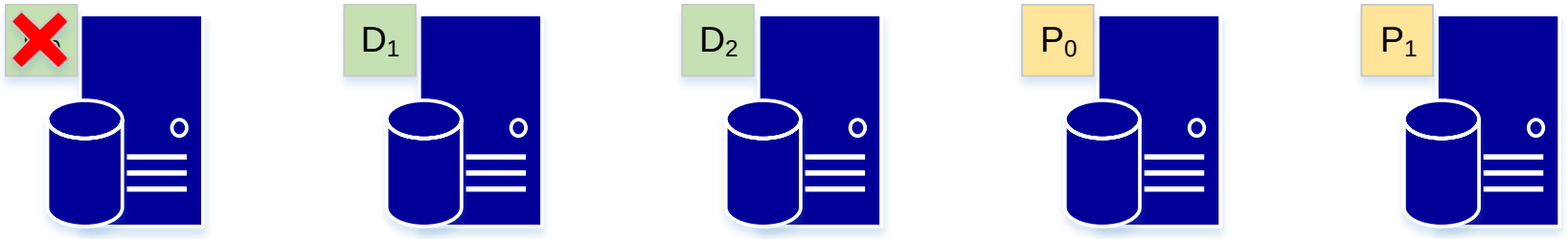
Erasure Coding (EC)

- EC popularly adopted in DSSes
  - Provide high reliability with low storage cost
  - (k, m)-Reed Solomon (RS) codes
    - k data chunks
    - m parity chunks
    - Tolerate any m nodes failures



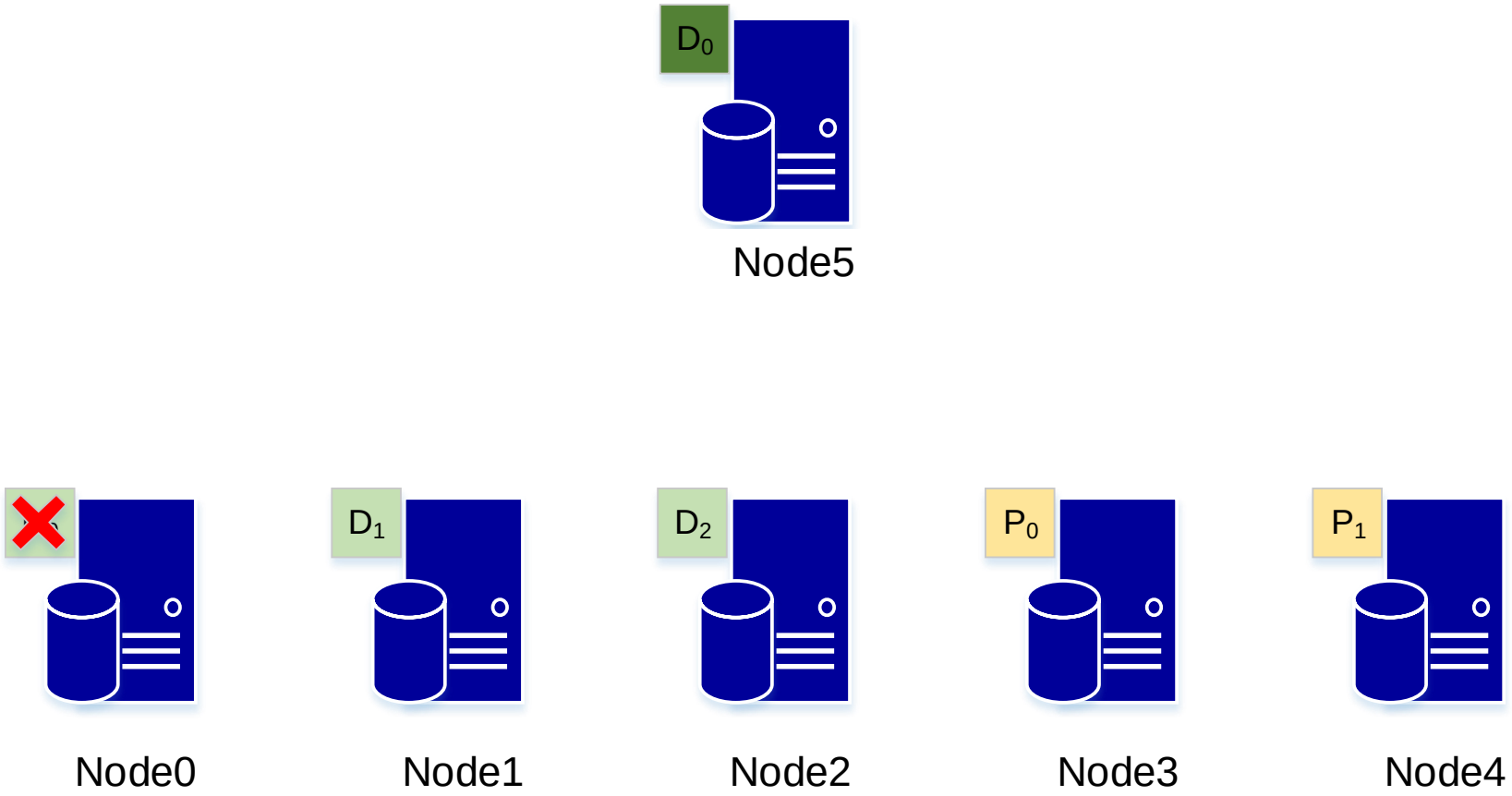
Writing a (3,2)-RS stripe

Reconstruction



Reconstructing a chunk of (3,2)-RS stripe

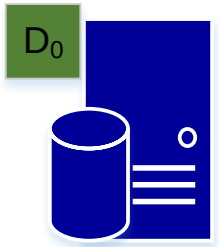
Reconstruction



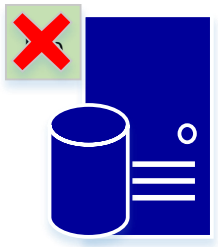
Reconstructing a chunk of (3,2)-RS stripe

Reconstruction

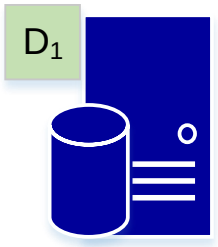
① Reading chunks from source nodes



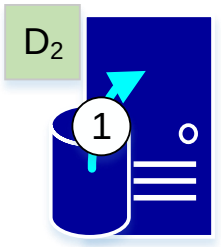
Node5



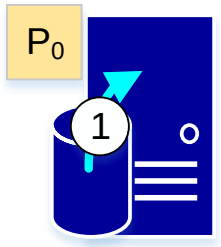
Node0



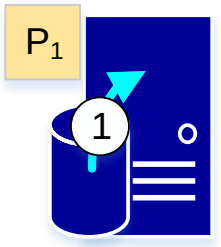
Node1



Node2



Node3

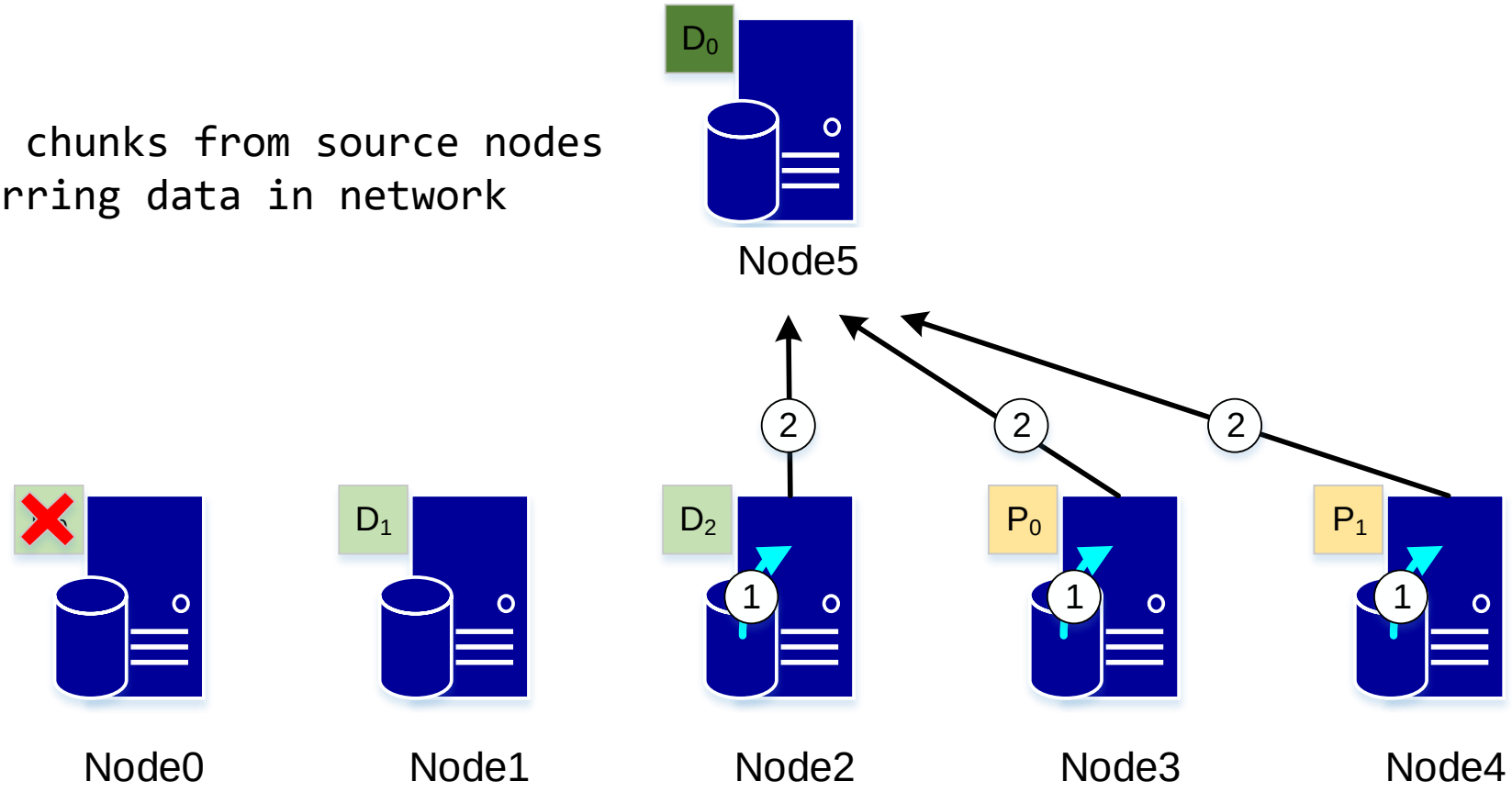


Node4

Reconstructing a chunk of (3,2)-RS stripe

Reconstruction

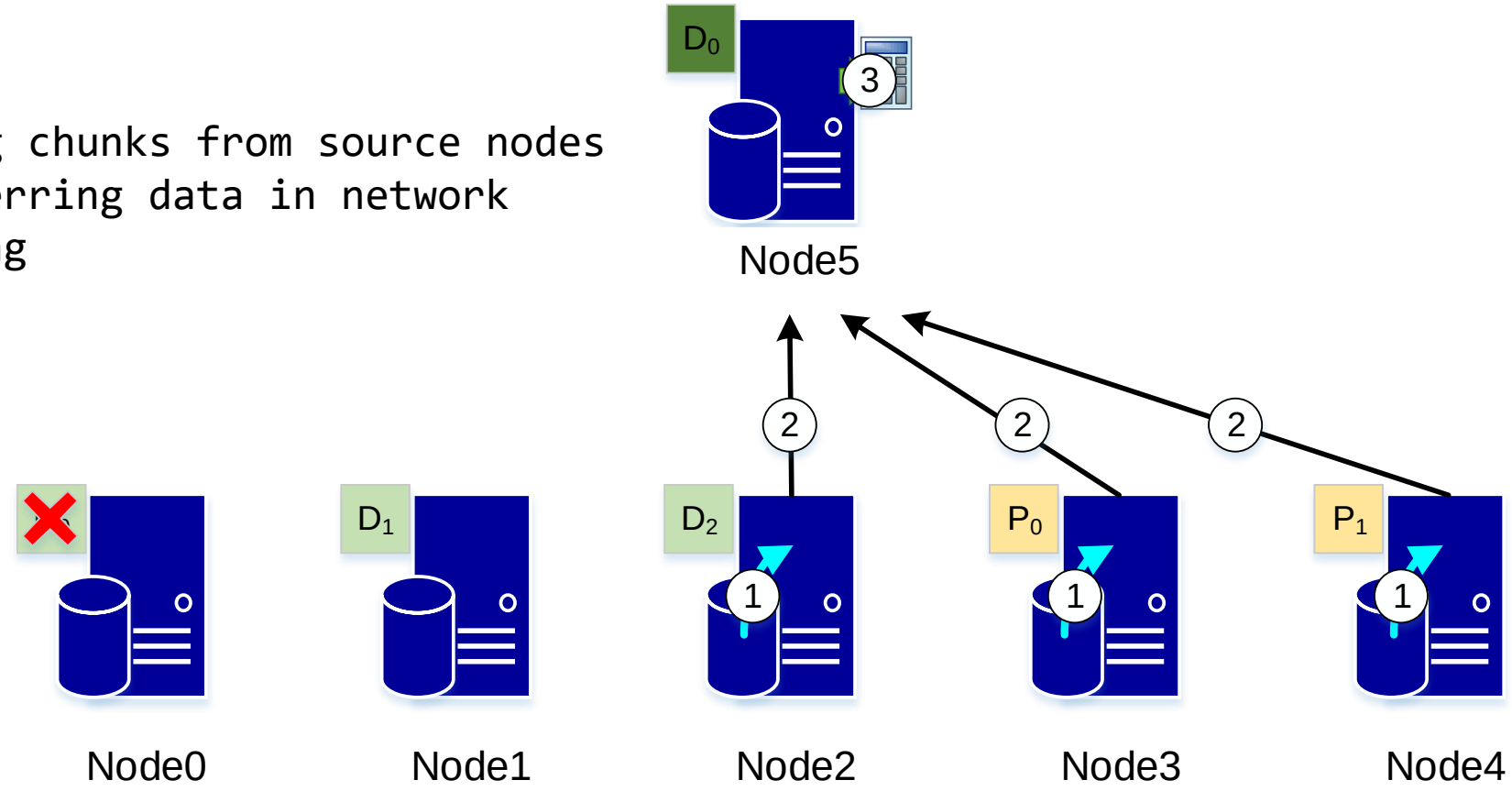
① Reading chunks from source nodes
② Transferring data in network



Reconstructing a chunk of (3,2)-RS stripe

Reconstruction

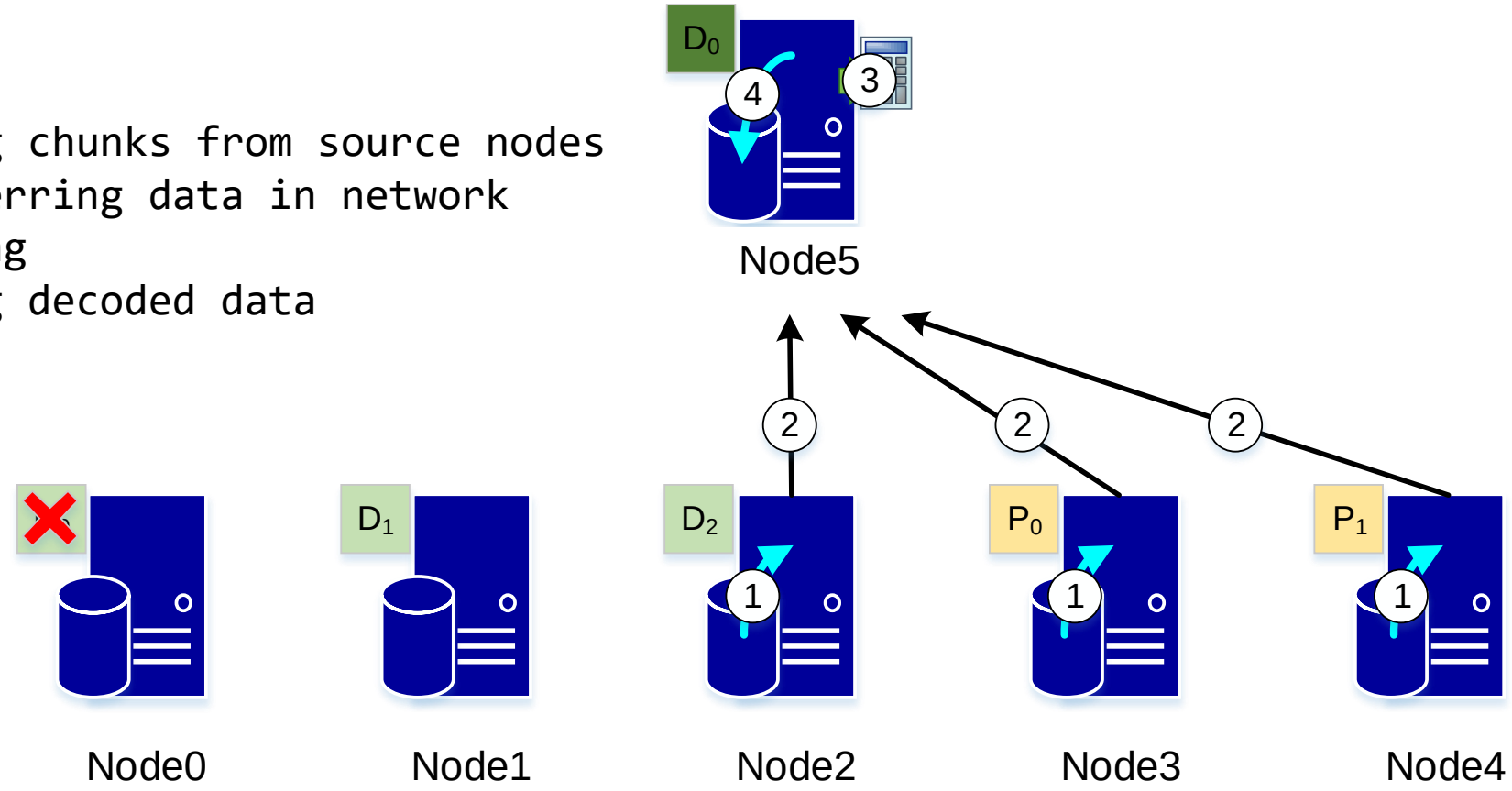
① Reading chunks from source nodes
② Transferring data in network
③ Decoding



Reconstructing a chunk of (3,2)-RS stripe

Reconstruction

① Reading chunks from source nodes
② Transferring data in network
③ Decoding
④ Writing decoded data



Reconstructing a chunk of (3,2)-RS stripe

Breakdown of EC Reconstruction Time

➢ **Settings**

- 28 nodes: 1NN + 27DNs
- quad-core 3.4 GHz Intel Core i5-7500 CPU
- 8GB RAM
- 1T HDD
- 1Gbps switch (30MB/s, 90MB/s or 150MB/s in Pangu[1])
- 128MB chunk size

Reconstructing a (3,2)-RS chunk in 1Gbps network

Stages	Reading chunks from source nodes	Transferring data in network	Decoding	Writing decoded data
Time Ratio	0.68%	85.23%	7.82%	6.27%

➢ Network transferring contributes most to the reconstruction time

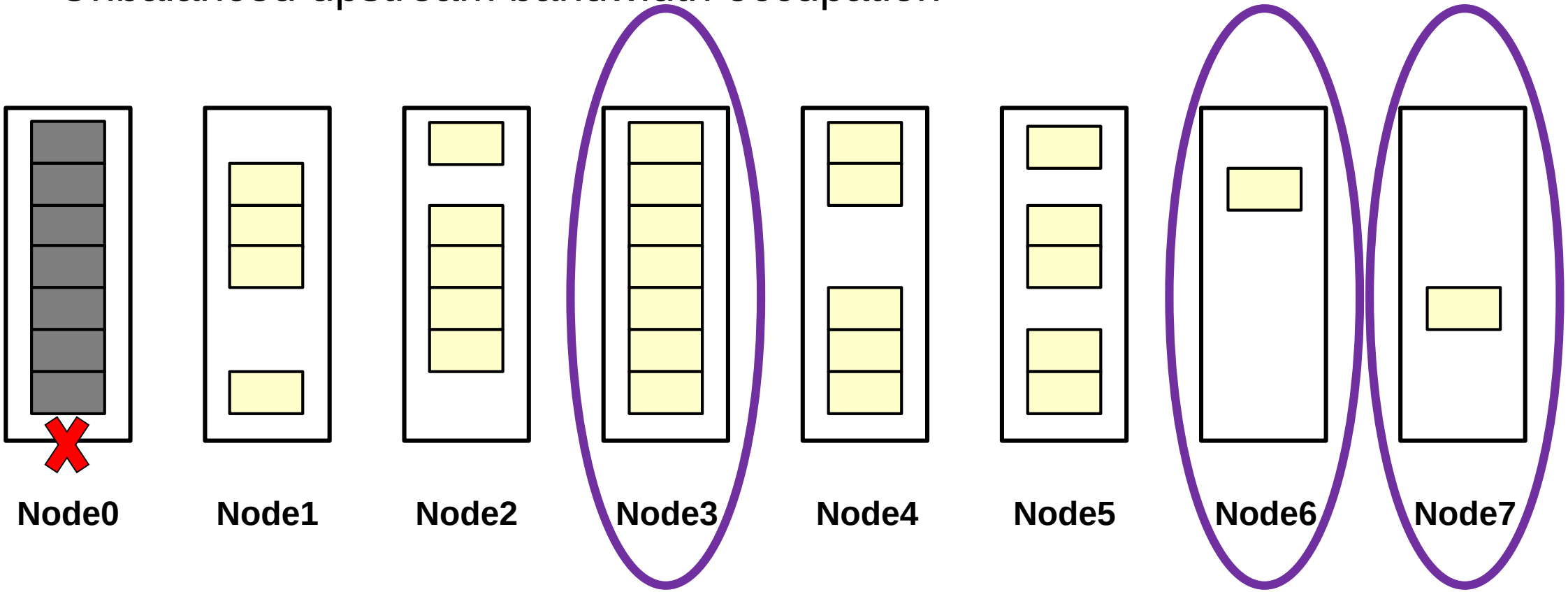
[1] ATC2019: Dayu: Fast and Low-interference Data Recovery in Very-large Storage Systems

Random Data Layout

- Random distribution
  - Load balance in a large amount of stripes
- Reconstruction batch by batch
  - Limited network, disk I/O, CPU and memory resource
  - Optimal batch size
    - # of live nodes
    - Detailed analysis in the paper

Random Data Layout

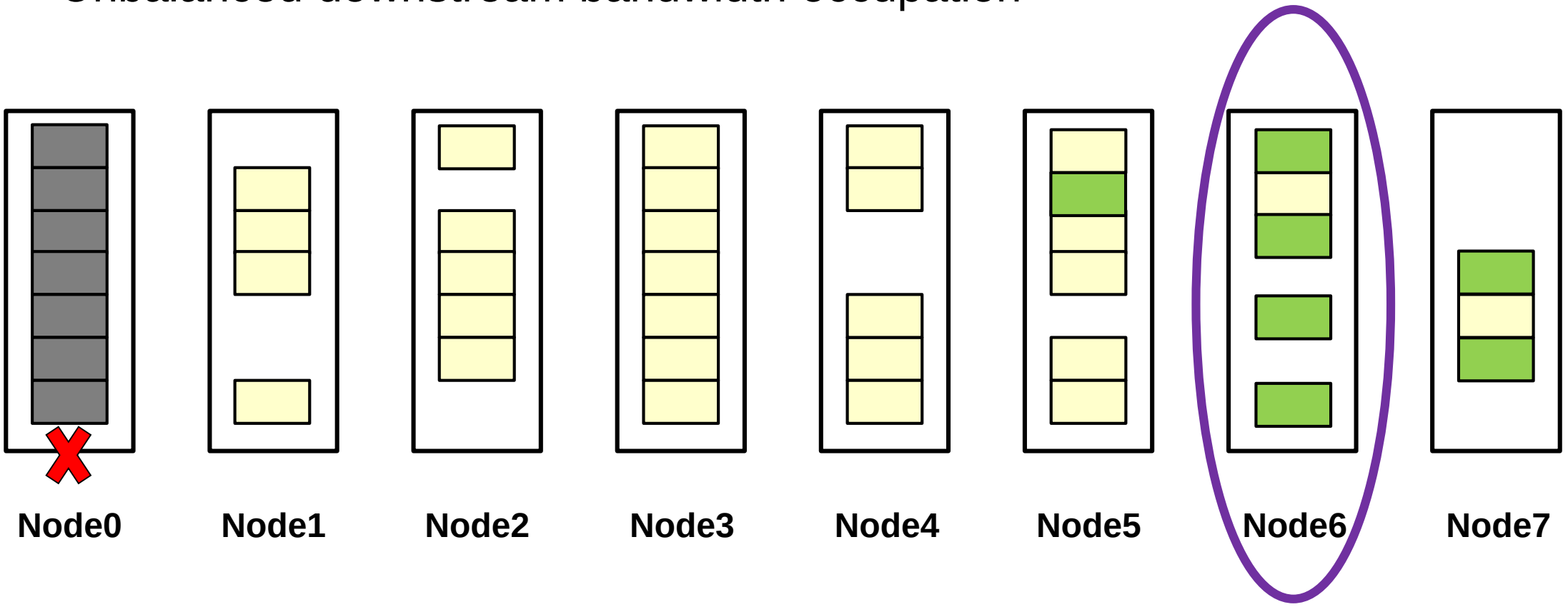
- Nonuniform data layout in a batch
  - Unbalanced upstream bandwidth occupation



Random data layout of (3,2)-RS stripes

Random Data Layout

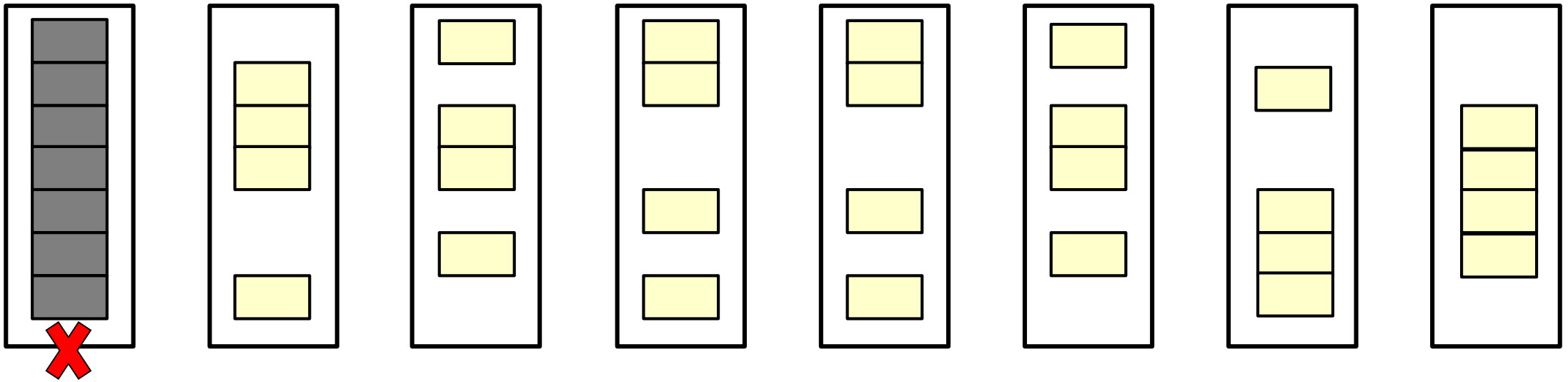
- Nonuniform choices of replacement nodes
  - Unbalanced downstream bandwidth occupation



Random data layout of (3,2)-RS stripes

Goals

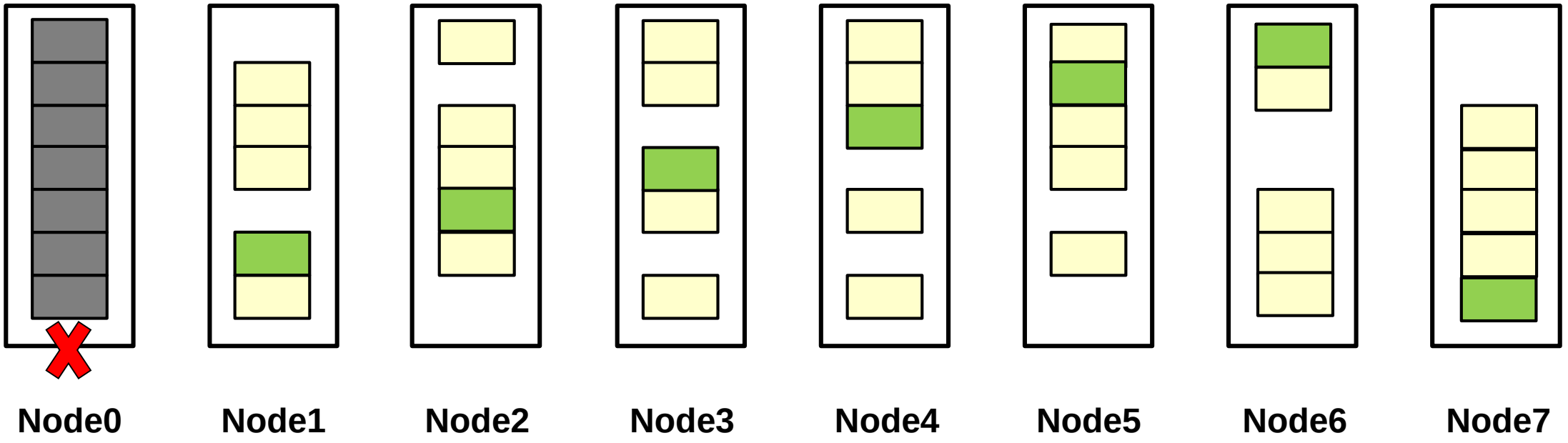
- Balanced distribution of source nodes



Random data layout of (3,2)-RS stripes

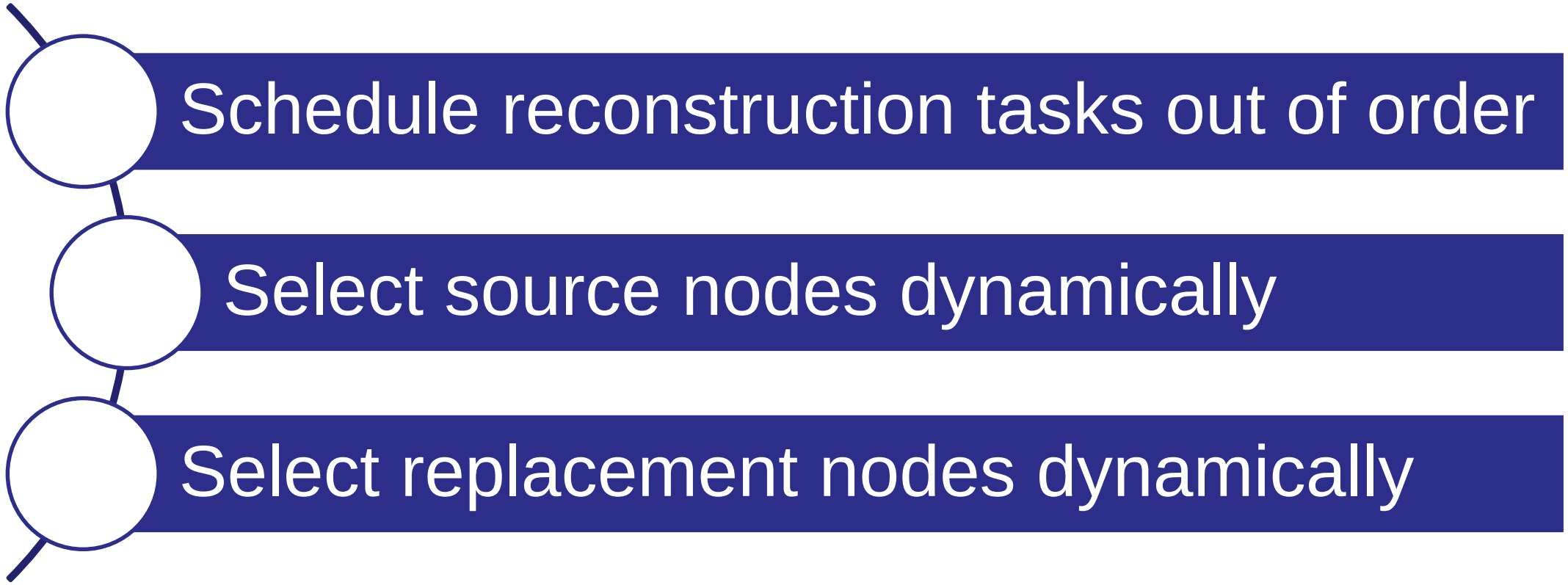
Goals

- Balanced distribution of source nodes
- Balanced distribution of replacement nodes



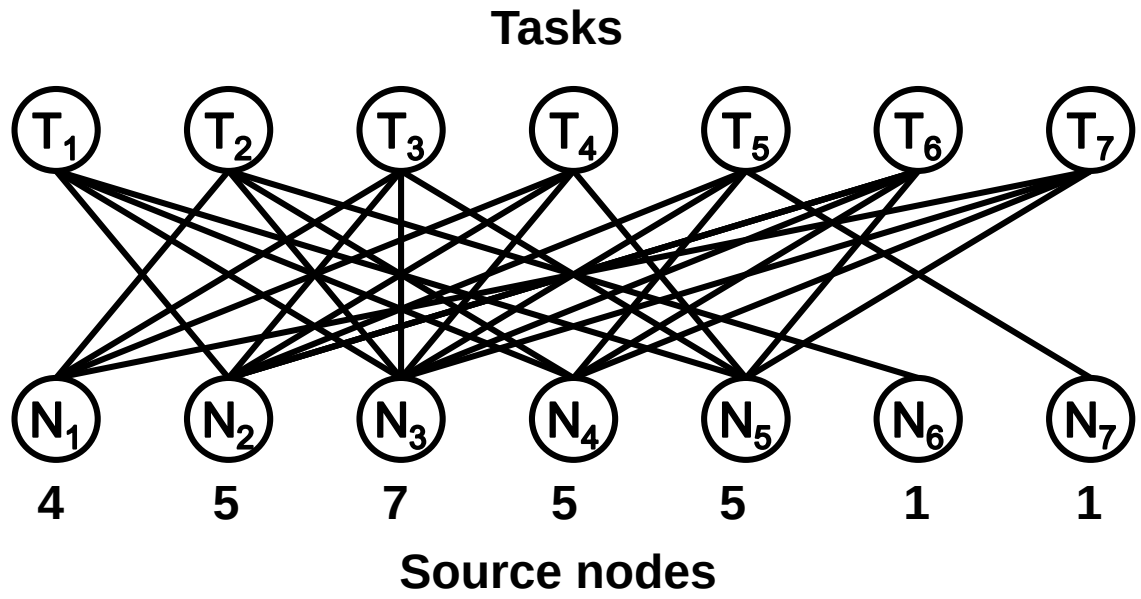
Random data layout of (3,2)-RS stripes

SelectiveEC



Graph Model

- Bipartite graph $G_s = (T \cup N, E)$ for the selection of source nodes
  - T: tasks, i.e. each having k+m-1 source nodes
  - N: source nodes, i.e. all of live nodes
  - $(T_i, N_j) \in E$ iff there is a chunk of stripe T_i in source node N_j

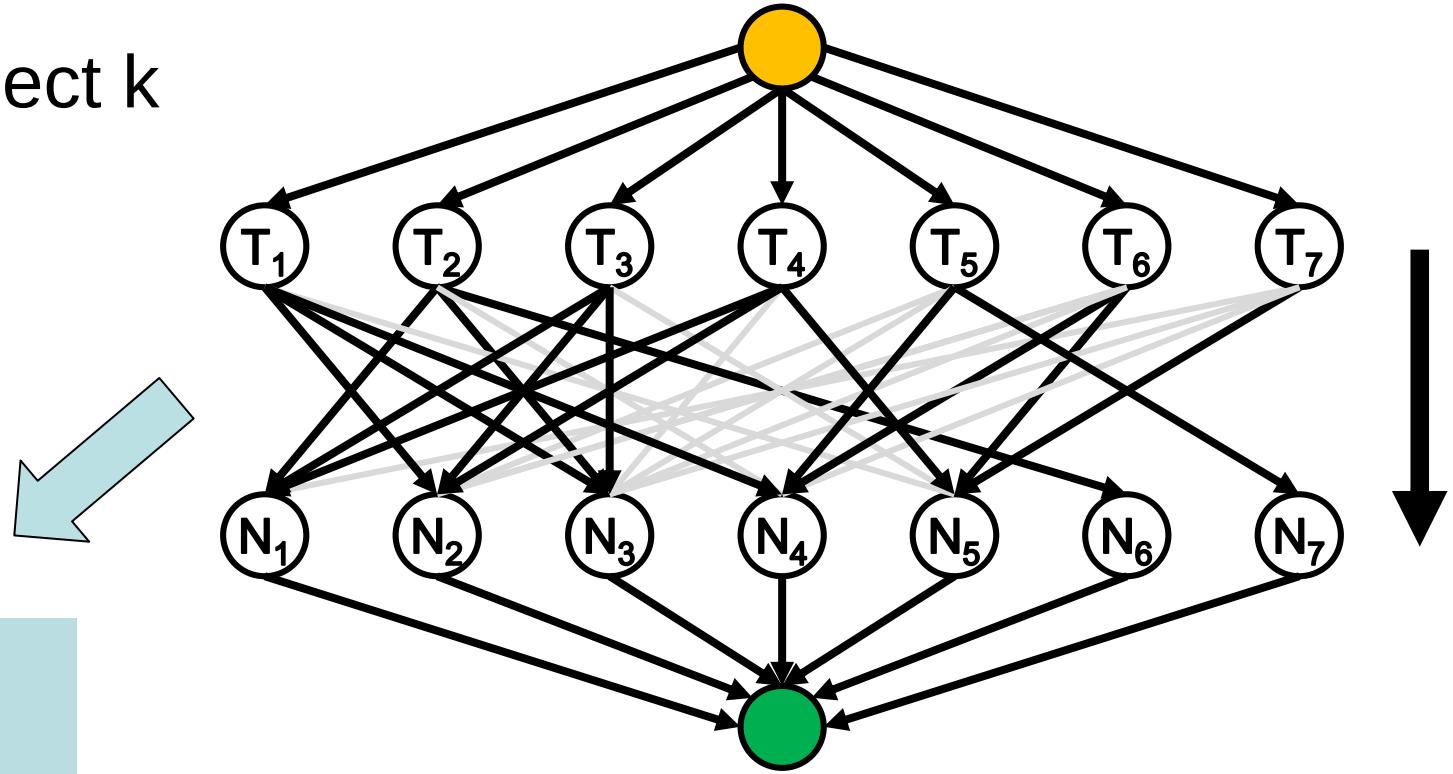


$G_s = (T \cup N, E)$ for (3, 2)-RS

- Connections of tasks and live nodes
- Nonuniform distribution of chunks

Select k Source Nodes Dynamically

- Goal: balance upstream bandwidth occupation
- Using maximum flow to select k source nodes
  - Construct a flow graph FG_s
  - Find a maximum flow

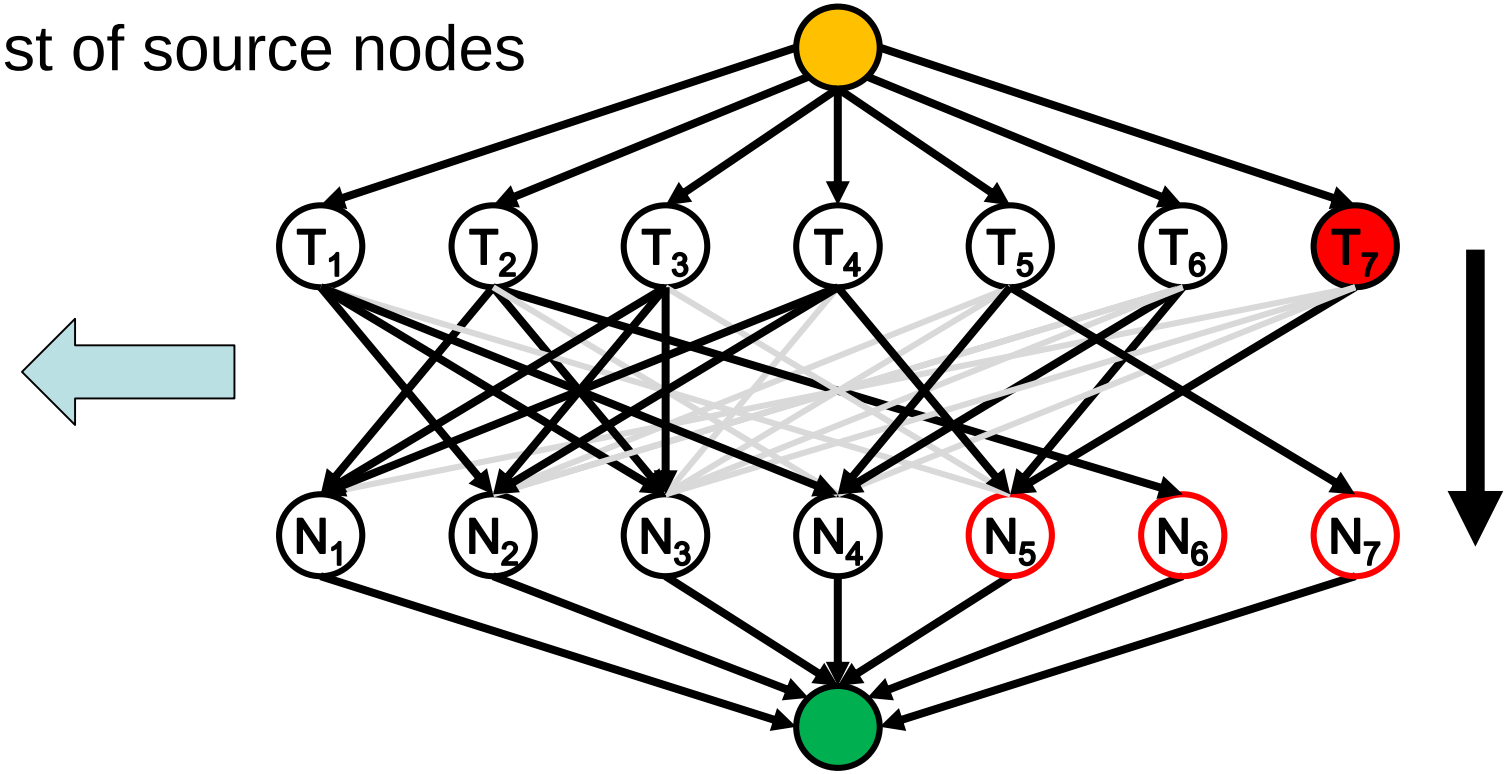


- Maximum flow value = 17
- No conflict in the chosen source connections

Schedule Reconstruction Tasks Out of Order

- Preparation work
  - Find the most unsaturated task
  - Compute an unsaturated list of source nodes

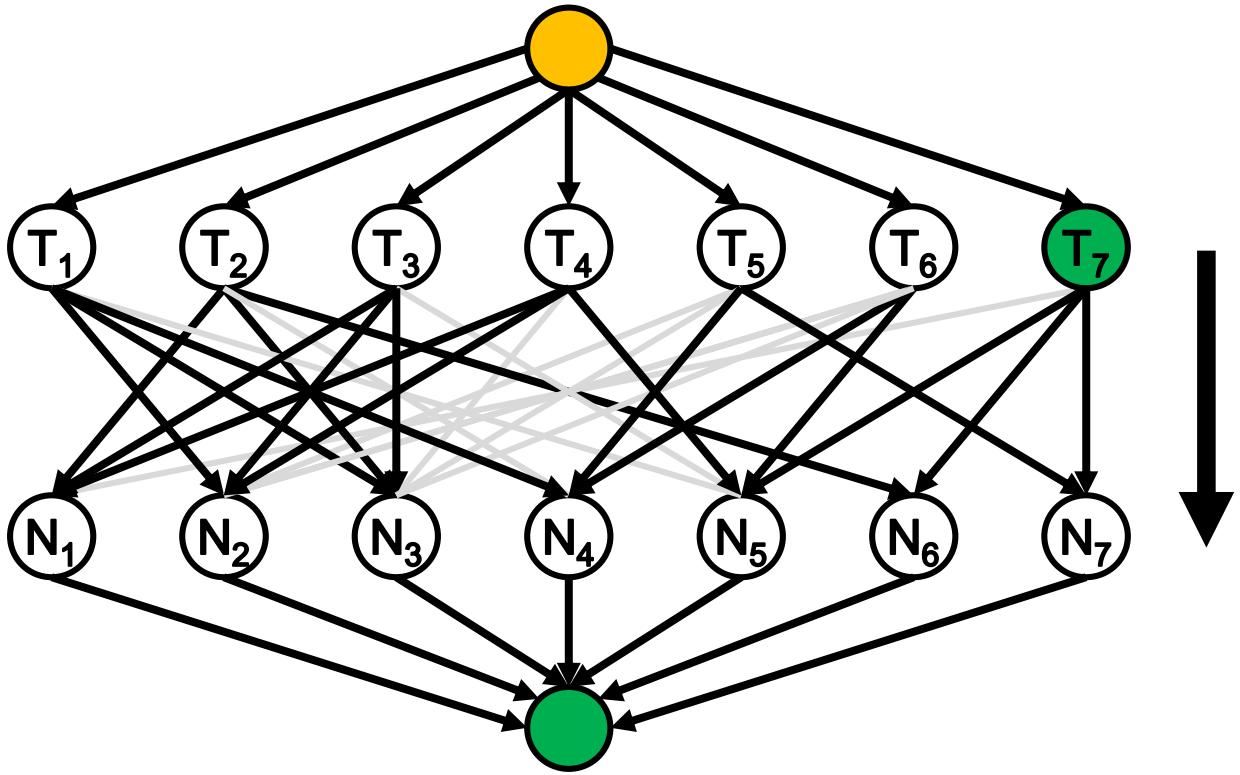
- Task to be replaced: T_7
- Unsaturated list: N_5, N_6, N_7



Schedule Reconstruction Tasks Out of Order

- Schedule reconstruction tasks
  - Scan the reconstruction queue
  - Find a new task
    - More connections with unsaturated list
  - Update FG_s
    - Find a maximum flow

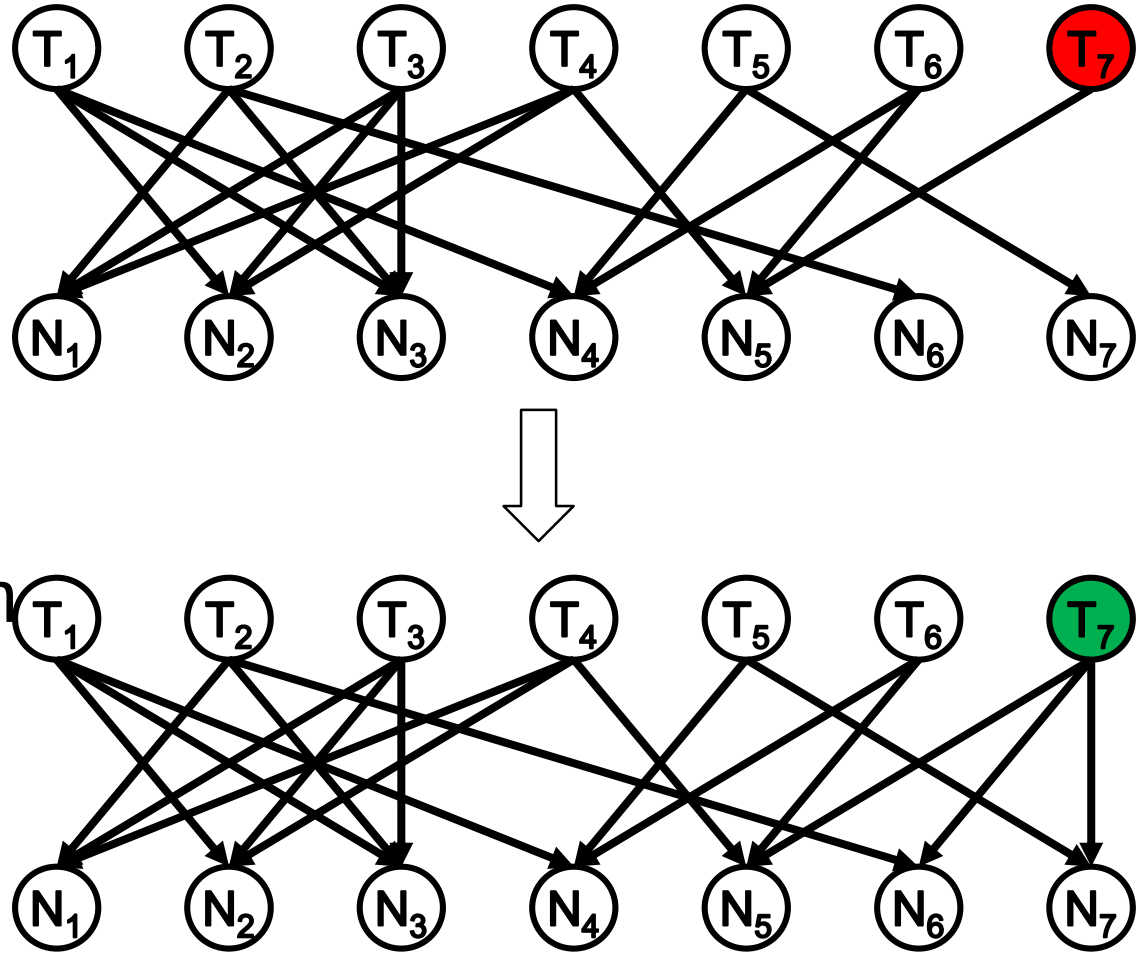
Replace a new task: T_7



Maximum flow value = 19

Schedule Reconstruction Tasks Out of Order

- Schedule reconstruction tasks
  - Scan the reconstruction queue
  - Find a new task
    - More connections with unsaturated list
  - Update FG_s
    - Find a maximum flow
- Achieve more balanced upstream bandwidth occupation



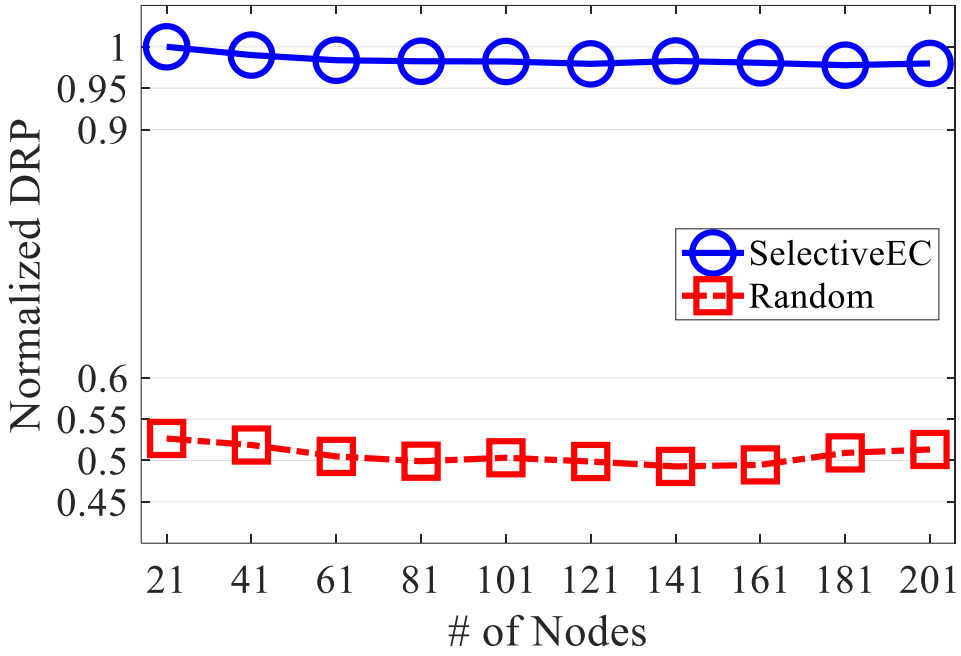
Select Replacement Nodes Dynamically

- Construct bipartite graph G_r for the selection of replacement nodes
  - Complement of G_s
  - Find a perfect matching
    - Easy to find in large-scale DSSes
- Achieve load balance of replacement nodes
  - Balanced downstream bandwidth occupation
  - Balanced disk I/O, CPU and memory usage

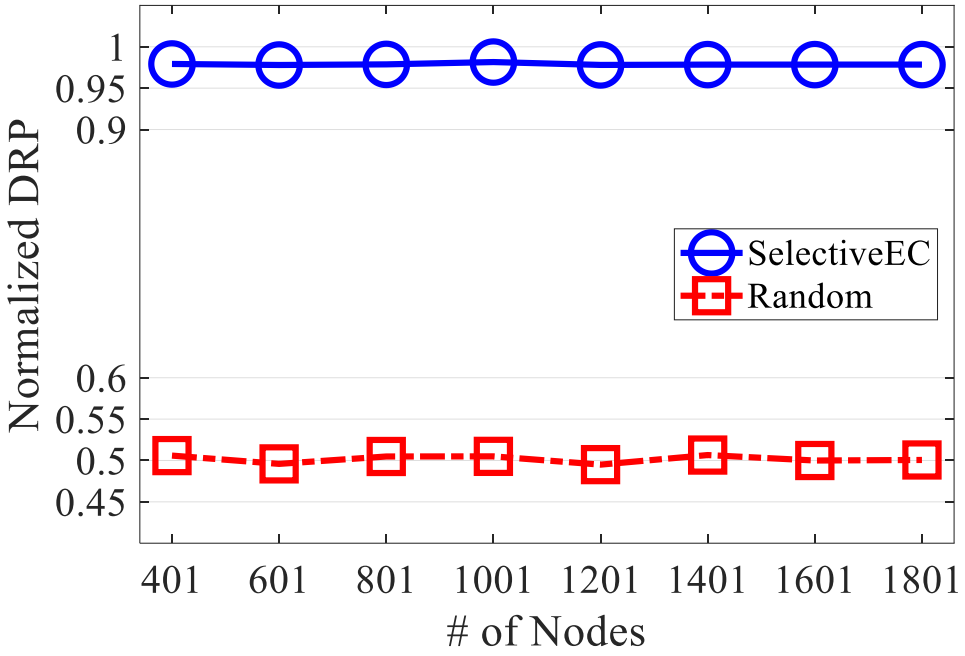
Evaluation

- Implement simulative prototype of SeletiveEC
- The simulations run in a server with
  - Two 12-core Intel Xeon E5-2650 processors
  - 64GB DDR4 memory
  - Linux 3.10.0
- (3,2)-RS stripes
- # of chunks in a “fat node”
  - 100 times of the number of live nodes
- DRP: the degree of recovery parallelism

The First Batch



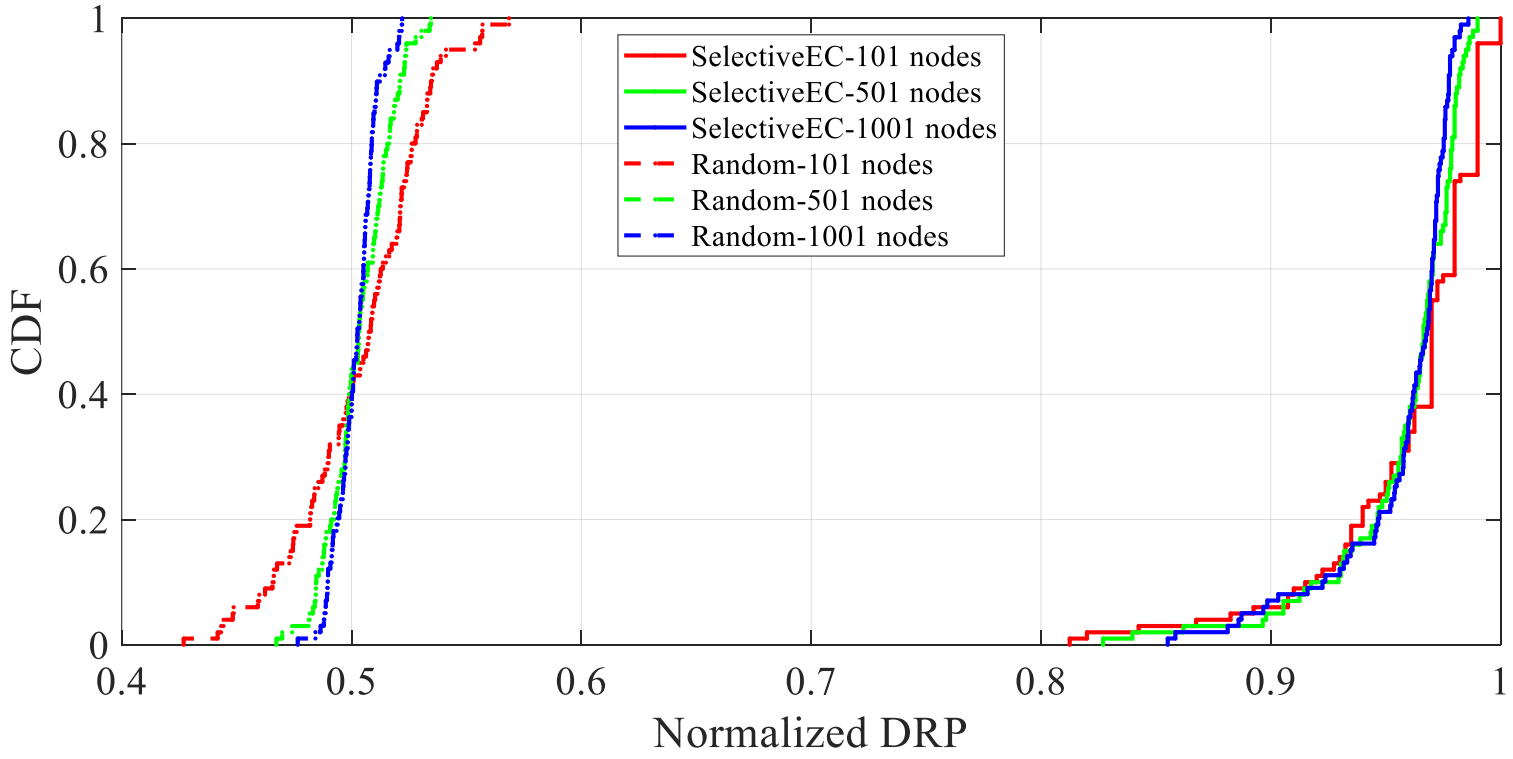
Small scale



Large scale

- For small scale, DRP of SelectiveEC are all bigger than 0.975
- For large scale, DRP of SelectiveEC improves the DRP up to 97.6%

Full Batches



- Around 0.97 for SelectiveEC
- Around 0.50 for random reconstruction

Summary

- SelectiveEC, a balanced scheduling module
  - Schedule reconstruction tasks out of order
  - Select source nodes dynamically
  - Select replacement nodes dynamically
  - Improve the load balance for single failure recovery effectively
- Simulation results
  - Improve the degree of recovery parallelism significantly
- Future work
  - Deploy in practical systems
  - Optimize the algorithms to support multiple failures

Thanks for your attention!

Q&A

Liangliang Xu@USTC

llxu@mail.ustc.edu.cn