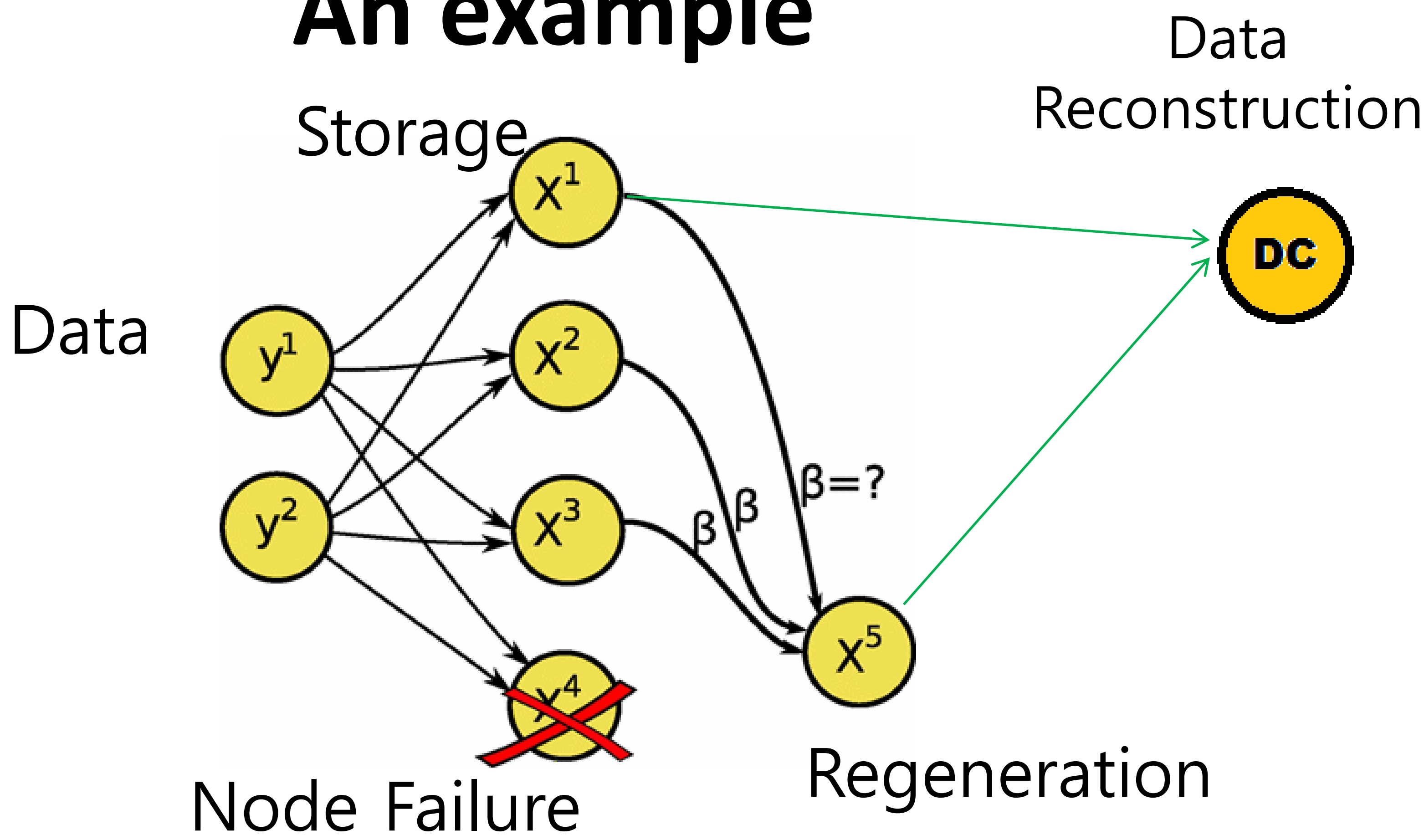
Regenerating Codes for Distributed Storage System

Yongjune Kim and Yaoqing Yang

Contents

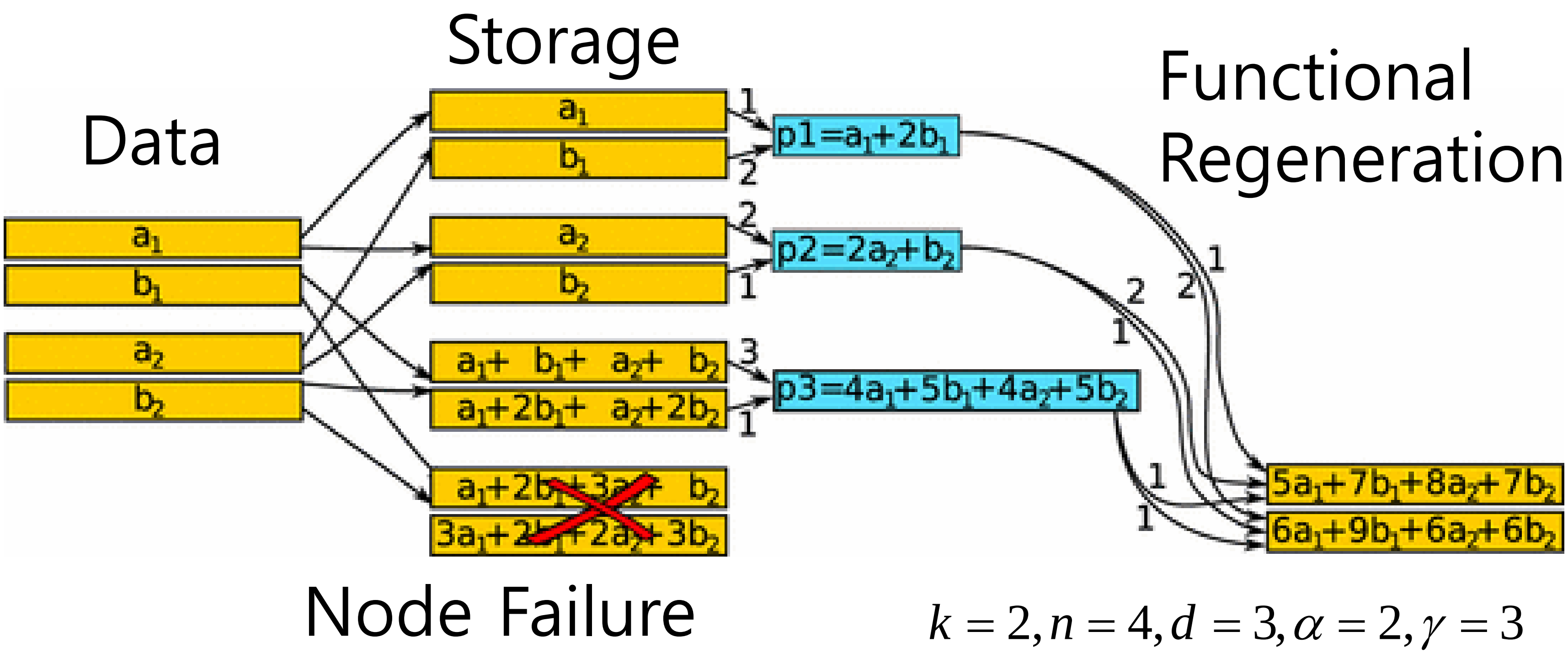
- Tradeoff between storage and communication
  - Dimakis et al. *IEEE Trans. Inf. Theory,* 2010
- Explicit code constructions
  - Rashmi, Shah, and Kumar, *IEEE Trans. Inf. Theory,* 2011.
- Open problems
  - Tian, *IEEE Journal on Selected Areas in Communications*, 2014.
  - Shah, Rashmi, Kumar, and Ramchandran, *IEEE ITW* 2010.

An example



"Network Coding for Distributed Storage Systems", A.G.Dimakis et.al. 2010

An example



$$k = 2, n = 4, d = 3, \alpha = 2, \gamma = 3$$

$$(n, k, d, \alpha, \gamma)$$

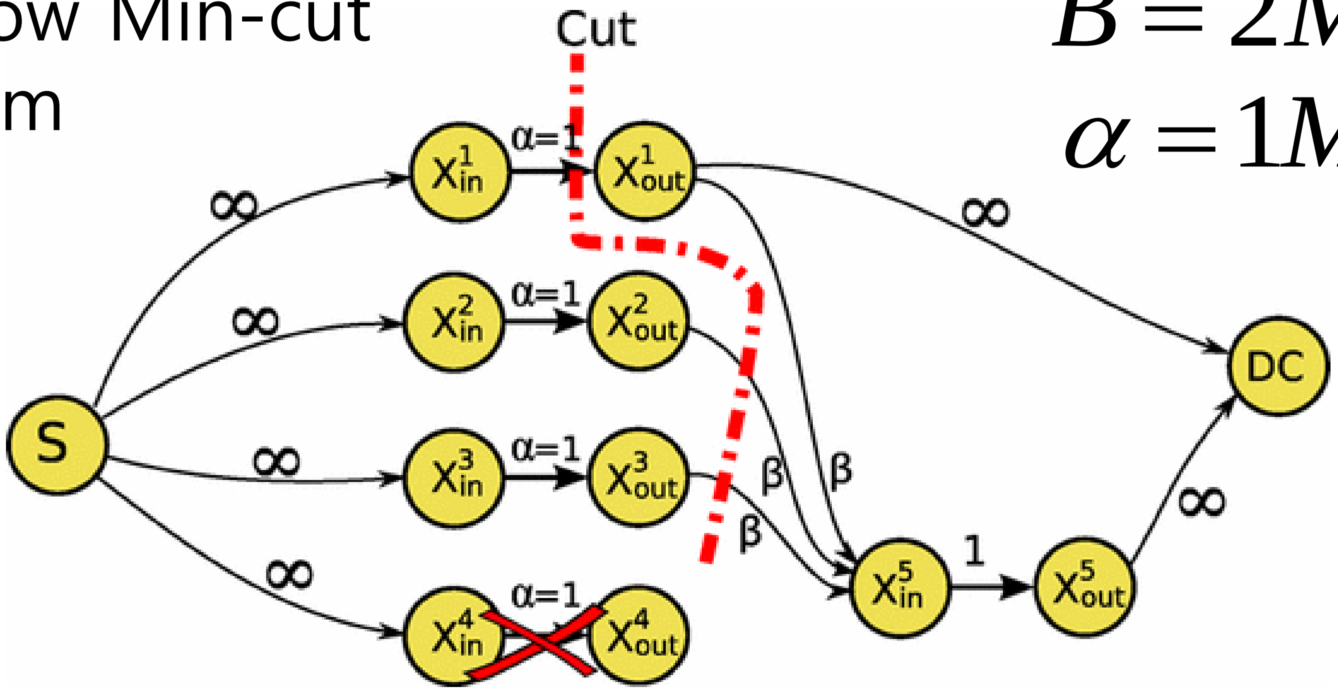
Whole file->k pieces->n fragments
Each fragment: α symbols
Regeneration bandwidth: $\gamma = d\beta$

An example

Max-flow Min-cut Theorem

$$B = 2Mb$$

$$\alpha = 1Mb$$



$$k = 2, n = 4, d = 3$$

$$\alpha + 2\beta \geq 2 \Rightarrow \gamma = d\beta \geq 1.5Mb$$

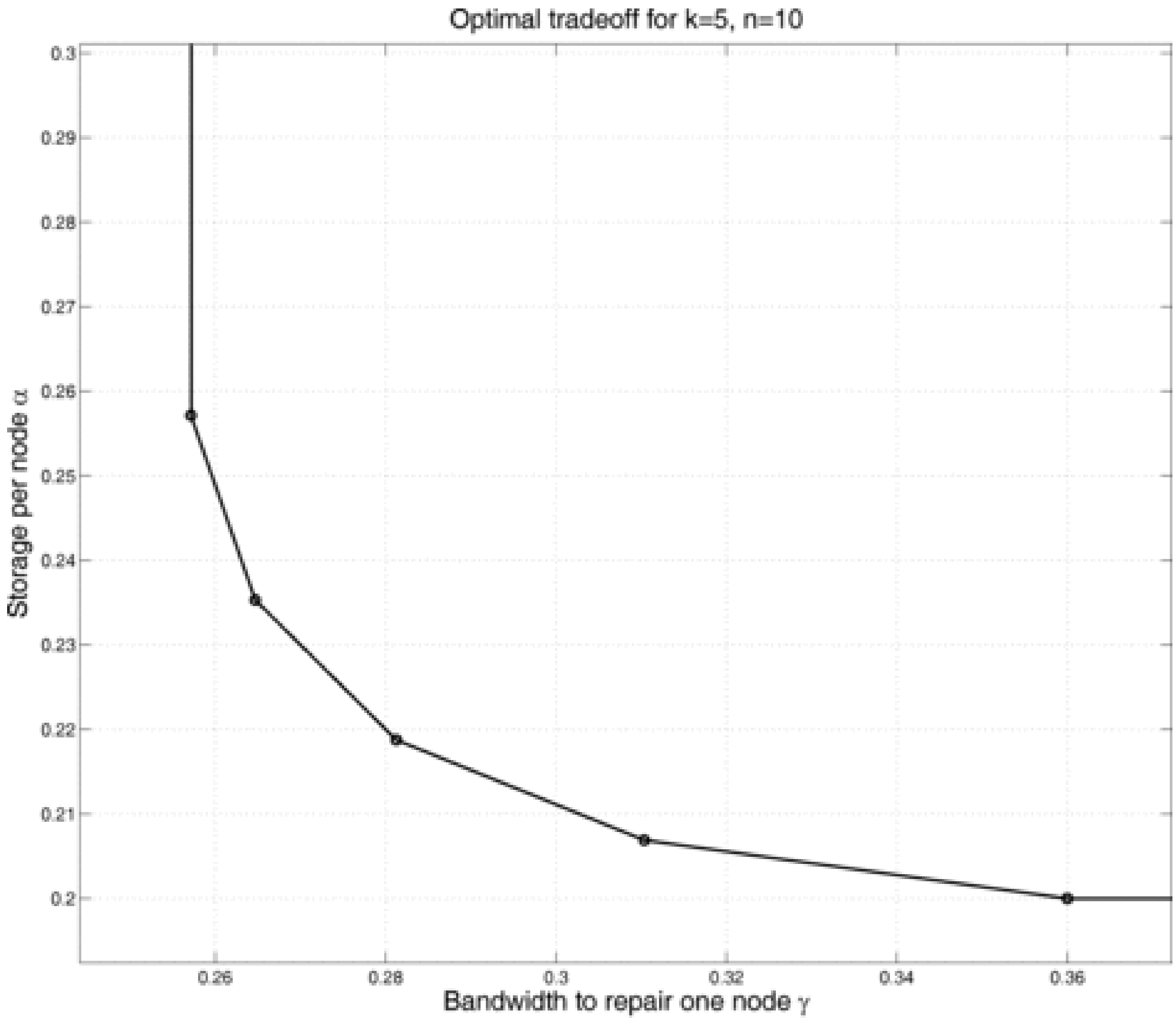
Minimum Storage Size

$$\alpha^*(n,k,d,\gamma)=\begin{cases}\dfrac{B}{k}, & \gamma\in[f(0),+\infty)\\[2ex] \dfrac{B-g(i)\gamma}{k-i}, & \gamma\in[f(i),f(i-1))\end{cases}$$

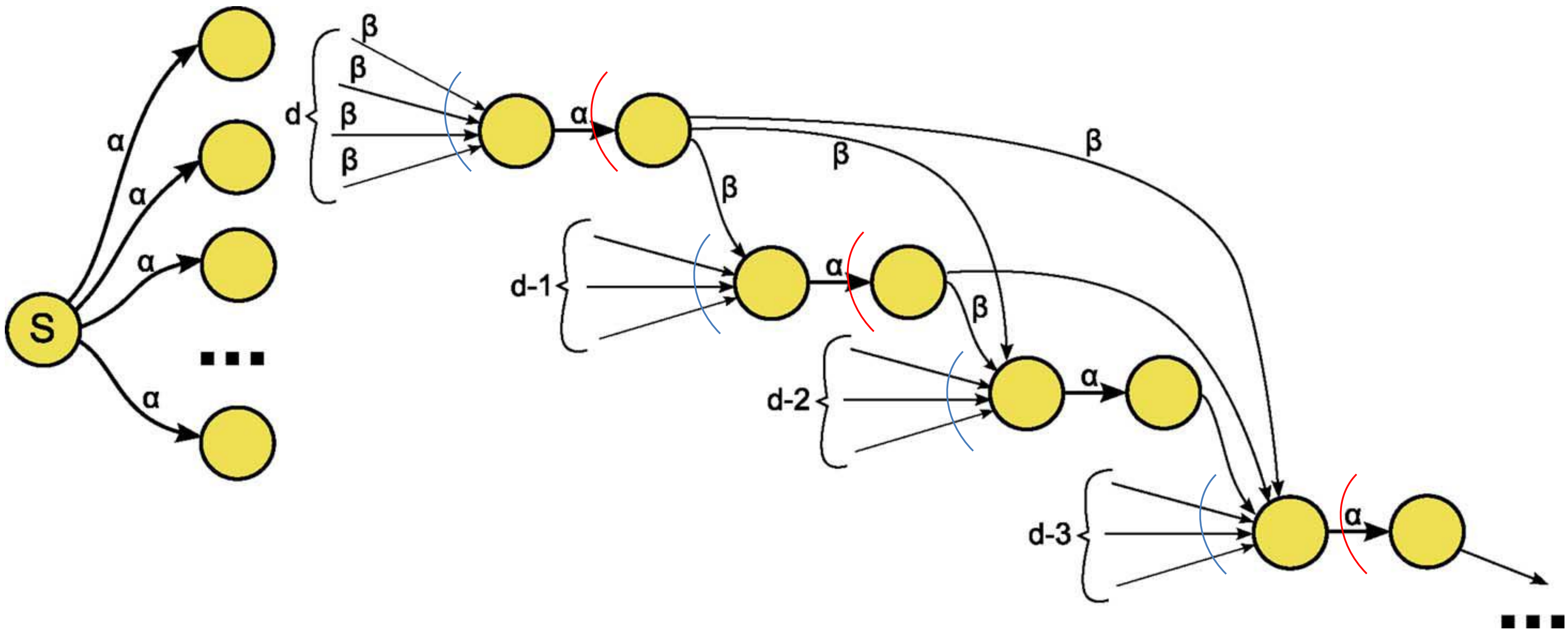
$$f(i)\triangleq\frac{2Bd}{(2k-i-1)i+2k(d-k+1)}$$

$$g(i)\triangleq\frac{(2d-2k+i+1)i}{2d}$$

Minimum Storage Size



Sketch of the Proof (Lower bound)

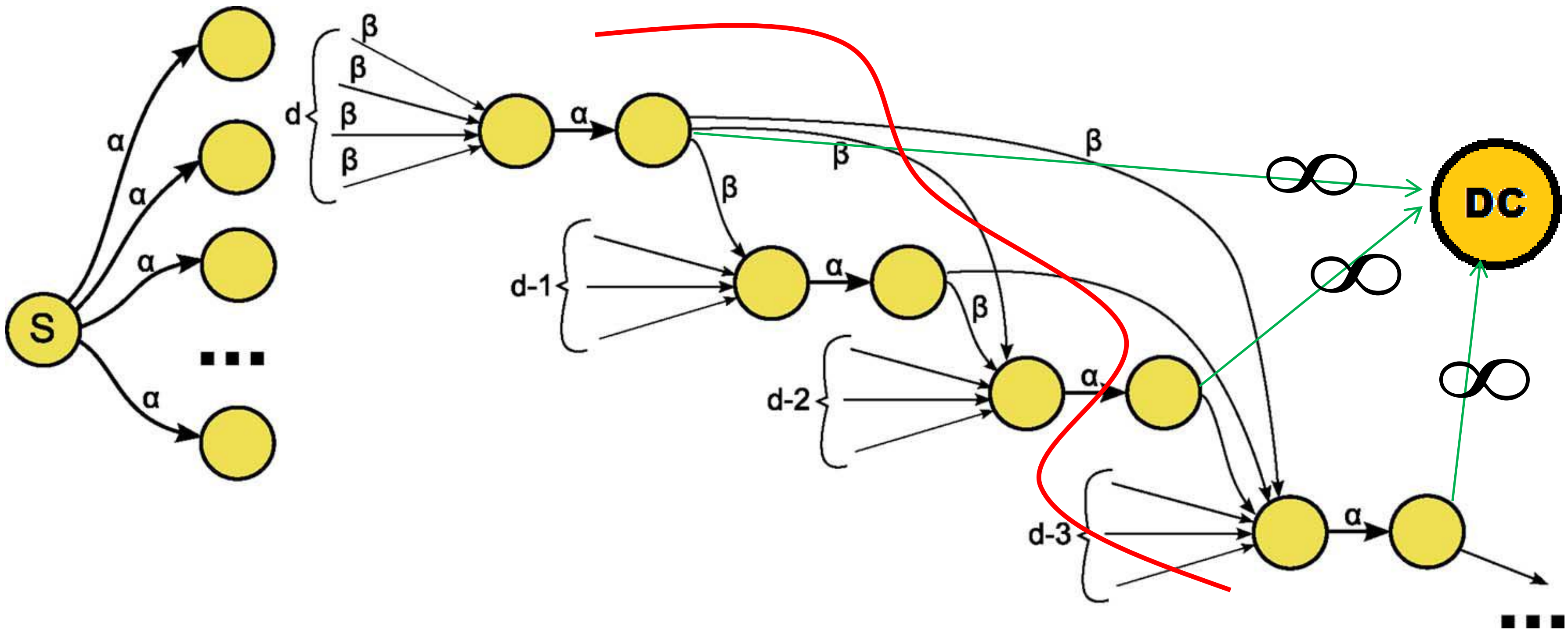


∃ Information flow

∃ Cut S

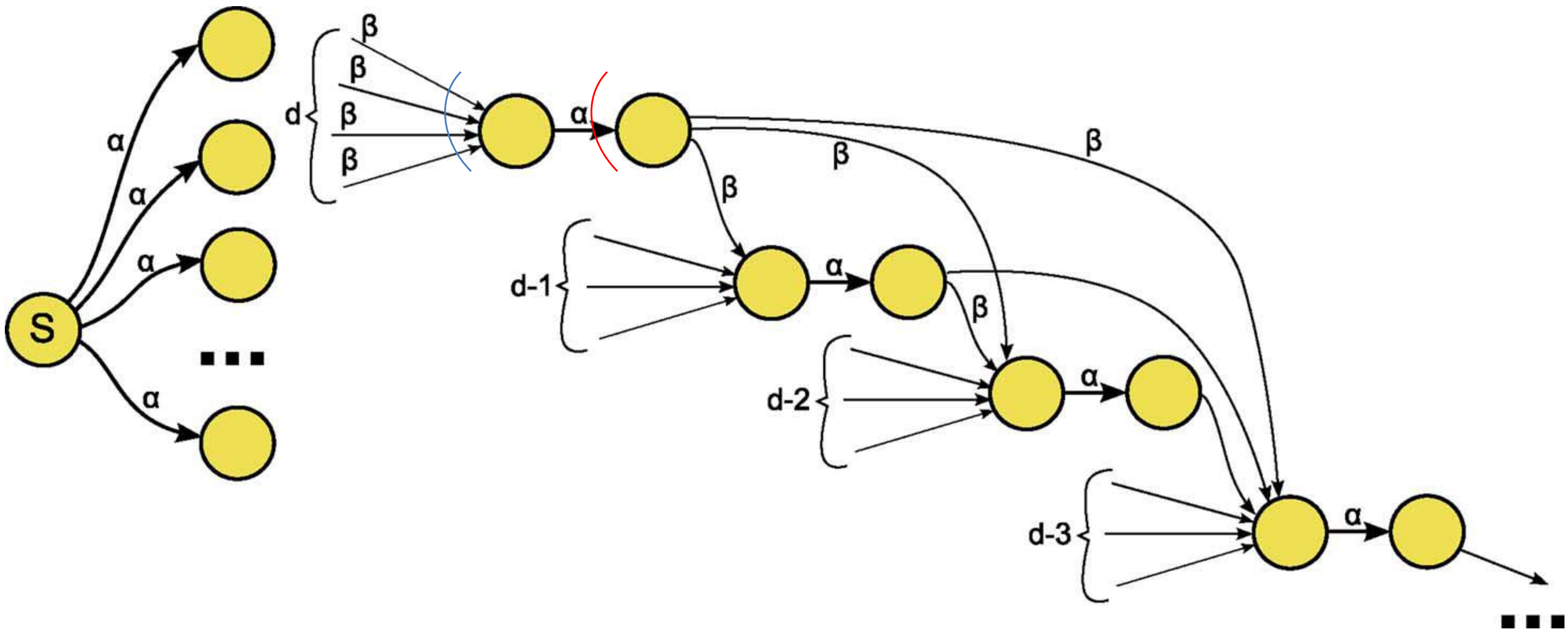
$$C = \sum_{e \in S} c_e = \sum_{i=0}^{\min\{d,k\}-1} \min\{(d-i)\beta, \alpha\}$$

Sketch of the Proof (Lower bound)



$\forall$ Information flow $\quad \text{mincut}(s,t) \geq \sum_{i=0}^{\min\{d,k\}-1} \min\{(d-i)\beta, \alpha\}$

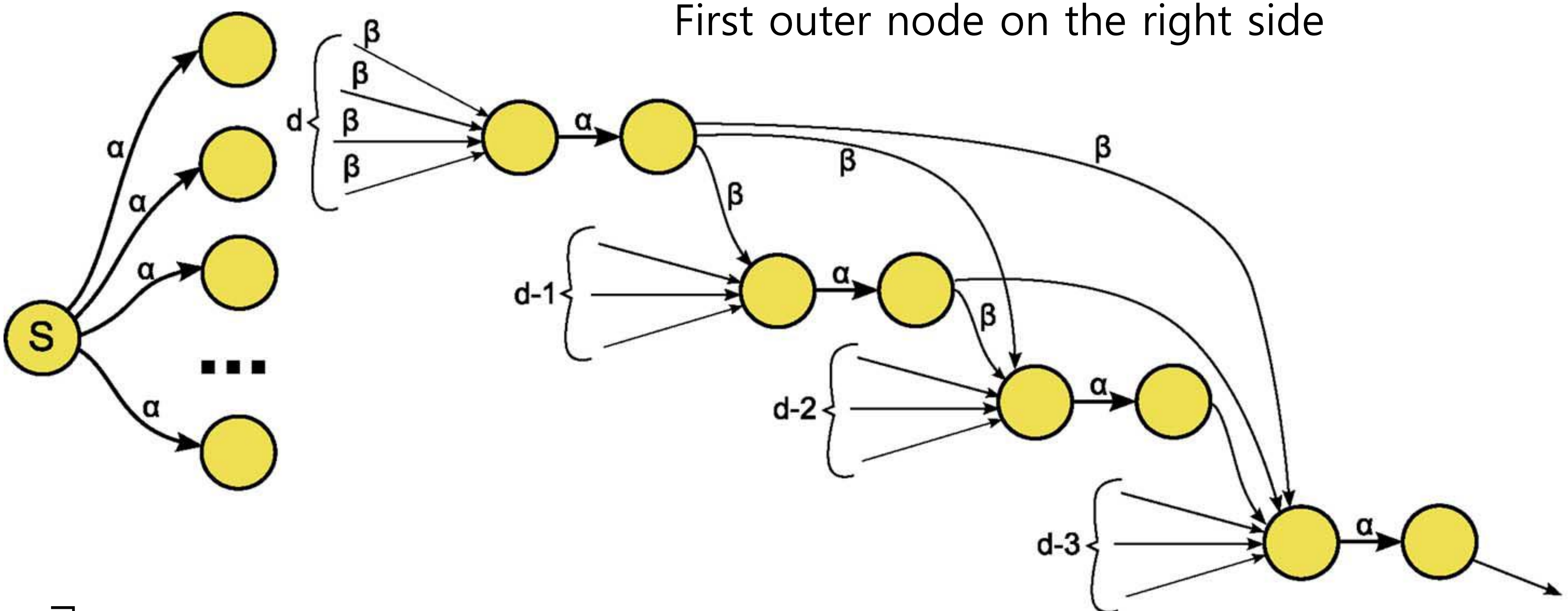
Sketch of the Proof (Lower bound)



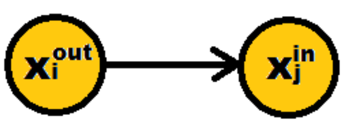
∀ Information flow $\quad \text{mincut}(s,t) \geq \sum_{i=0}^{\min\{d,k\}-1} \min\{(d-i)\beta, \alpha\}$

Sketch of the Proof (Lower bound)

First outer node on the right side



∃ Topological Order (any acyclic graph)



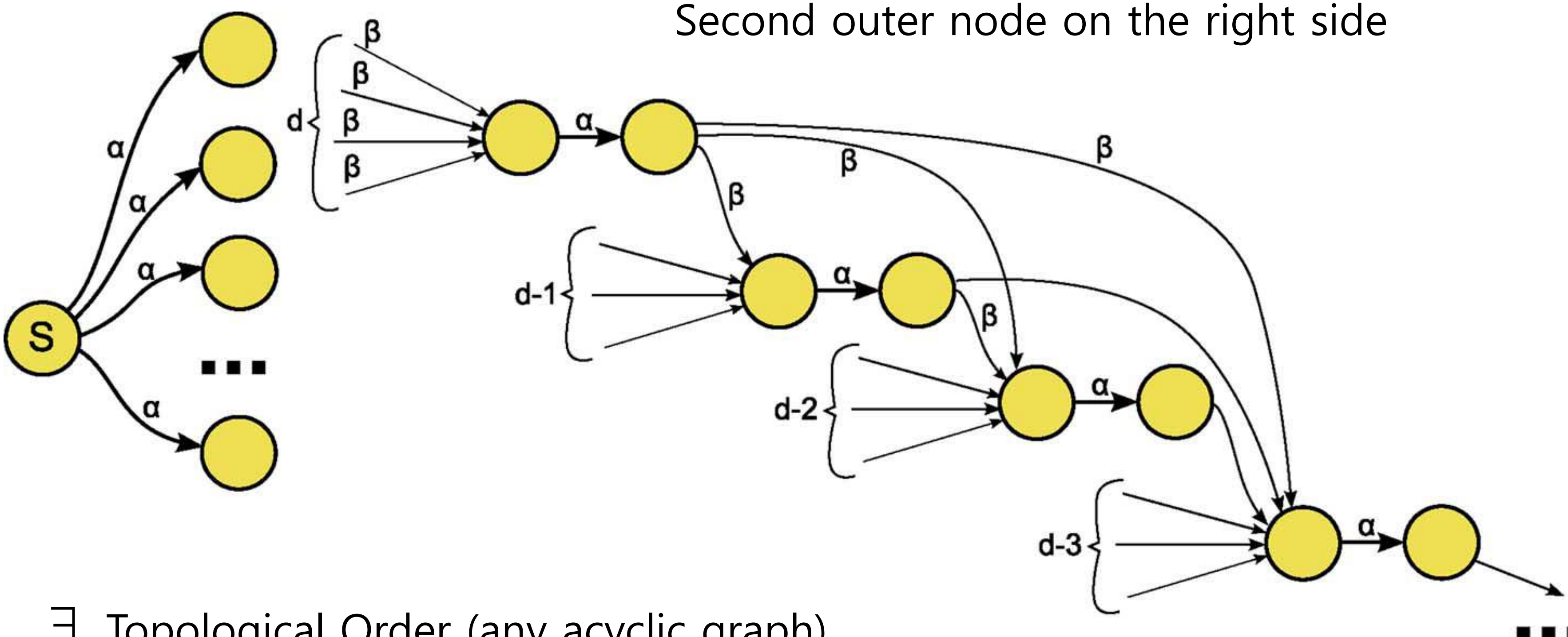
$\Rightarrow i < j$

$$c_1 \geq \min\{d\beta, \alpha\}$$

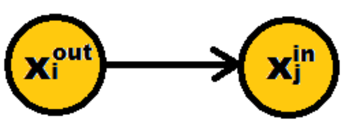
Time order is feasible

Sketch of the Proof (Lower bound)

Second outer node on the right side



∃ Topological Order (any acyclic graph)



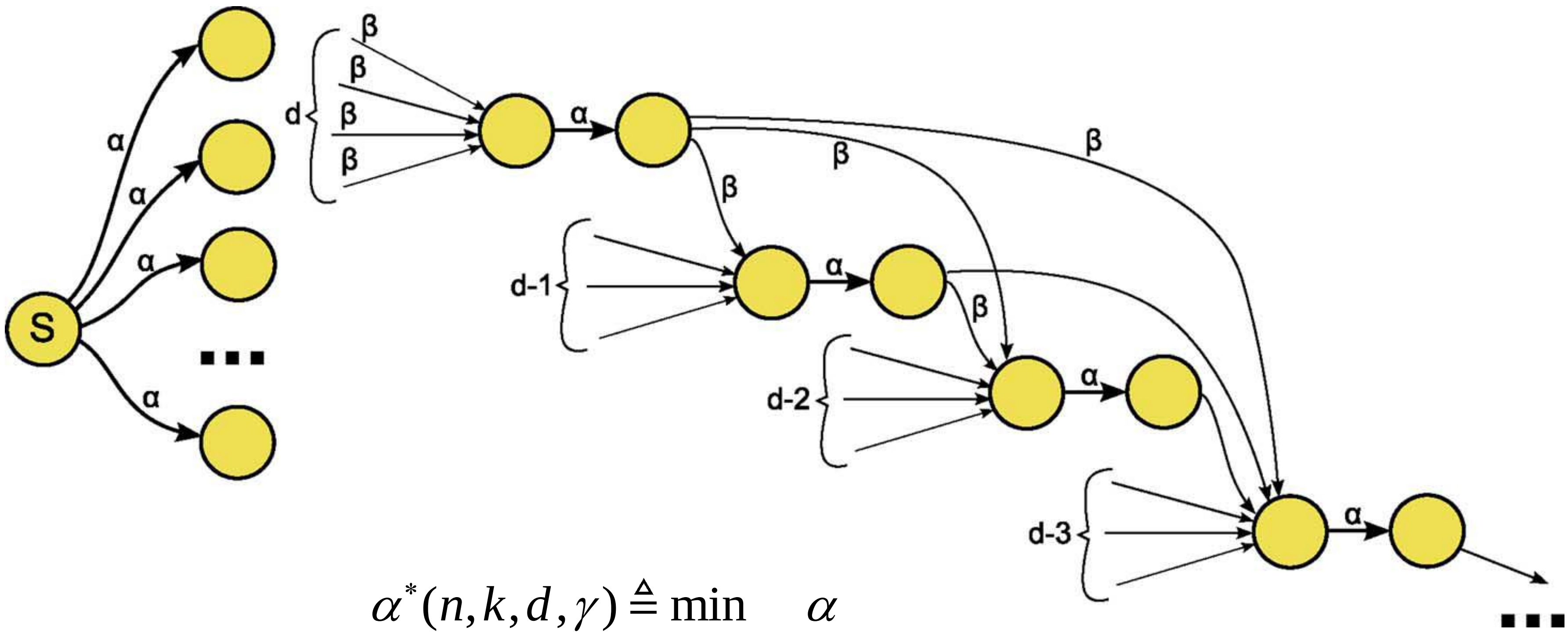
$$\Rightarrow i < j$$

Time order is feasible

$$c_2 \geq \min\{(d-1)\beta, \alpha\}$$

$$C = \sum_{e \in S} c_e = \sum_{i=0}^{\min\{d,k\}-1} \min\{(d-i)\beta, \alpha\}$$

Sketch of the Proof (Lower bound)

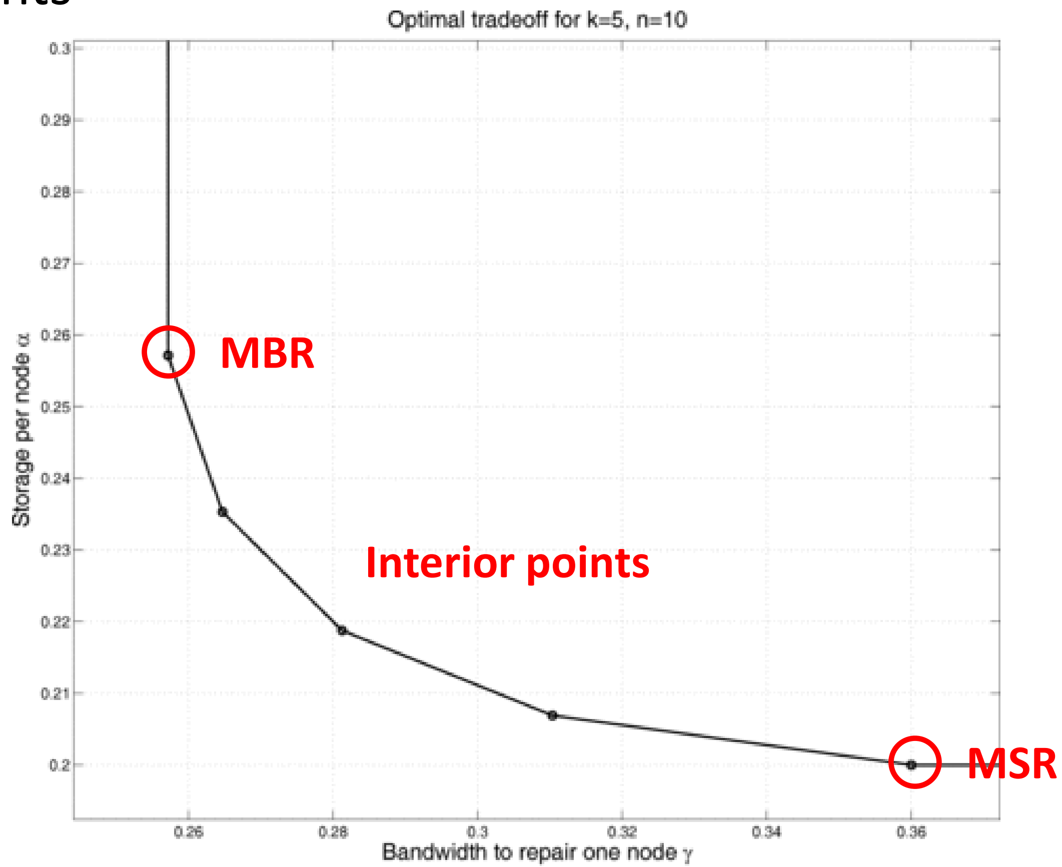


$$\alpha^*(n,k,d,\gamma) \triangleq \min \quad \alpha$$

$$\text{subject to:} \quad \sum_{i=0}^{\min\{k-1,d\}} \min\left\{(d-i)\beta, \alpha\right\} \geq B$$

Exact-Regeneration Code Constructions

- MBR (Minimum bandwidth regeneration) points
- MSR (Minimum storage regeneration) points
- Interior points



Exact-Regeneration Code Constructions

- **MBR point***: Explicit code constructions for all $[n, k, d]$ parameters are possible
- **MSR point***: Explicit code constructions for $[n, k, d \geq 2k - 2]$ parameters are possible
  - Low rate: $n - 1 \geq d \geq 2k - 2, \frac{k}{n} \leq \frac{k}{2k-1} \approx \frac{1}{2}$
- **Interior points****: Non-existence of exact regeneration codes

* Rashmi, Shah, and Kumar, *IEEE Trans. Inf. Theory,* 2011.
** Shah, Rashimi, Kumar and Ramchandram, *IEEE Trans. Inf. Theory,* 2012.

Product-Matrix Framework

$$C = \Psi M$$

$$\begin{bmatrix} c_1^t \\ \vdots \\ c_i^t \\ \vdots \\ c_n^t \end{bmatrix} = \begin{bmatrix} \psi_1^t \\ \vdots \\ \psi_i^t \\ \vdots \\ \psi_n^t \end{bmatrix} M$$

- **Code matrix** $C: n \times \alpha$
- **Encoding matrix** $\Psi: n \times d$
- **Message matrix** $M: d \times \alpha$
  - $d\alpha > B$, M contains the B message symbols and some redundancy

➔ $c_i^t = \psi_i^t M$ (α symbols) is stored in i-th node

Regeneration and Reconstruction

$$C = \Psi M$$

$$\begin{bmatrix} c_1^t \\ \vdots \\ c_f^t \\ \vdots \\ c_n^t \end{bmatrix} = \begin{bmatrix} \psi_1^t \\ \vdots \\ \psi_f^t \\ \vdots \\ \psi_n^t \end{bmatrix} M$$

Regeneration: Repair $c_f^t = \psi_f^t M$ of the failed node from $d\beta$ symbols (d nodes)

Reconstruction: Recovering M from $k\alpha$ symbols (k nodes)

MBR Code Construction

Parameter set: $\left(\alpha = d, \beta = 1, B = \binom{k+1}{2} + k(d-k)\right)$

Message matrix M (symmetric): Independent B symbols

$$M = \begin{bmatrix} S & T \\ T^t & 0 \end{bmatrix}$$

- S: $(k \times k)$ symmetric matrix with $\binom{k+1}{2}$ symbols
- T: $\left(k \times (d-k)\right)$ matrix with $k \times (d-k)$ symbols

Encoding matrix Ψ:

$$\Psi = [\Phi \;\; \Delta]$$

- Any d rows of Ψ are linearly independent (Ψ: $n \times d$)
- Any k rows of Φ are linearly independent (Φ: $n \times k$)

MBR Exact-Regeneration

Theorem: Exact-regeneration of any failed node can be achieved by downloading one symbol each from any d nodes

Proof:

- Want to repair $c_f^t = \psi_f^t M$ in the failed node.
- Get the following d symbols from d helper nodes.

$$\begin{bmatrix} c_{i_1}^t \psi_f \\ \vdots \\ c_{i_d}^t \psi_f \end{bmatrix} = \begin{bmatrix} \psi_{i_1}^t \\ \vdots \\ \psi_{i_d}^t \end{bmatrix} M\psi_f = \Psi_{\text{repair}} M\psi_f$$

- Since Ψ_{repair} is invertible, we can obtain $M\psi_f$.
- Since M is symmetric, $\left(M\psi_f\right)^t = \psi_f^t M$, which is the data stored in the failed node.

MBR Data-Reconstruction

Theorem: All the B message symbols can be recovered by connecting to any k nodes

Proof:

- Want to recover M (B message symbols).
- Get the following $k\alpha$ symbols from k helper nodes.

$$\begin{bmatrix} c_{i_1}^t \\ \vdots \\ c_{i_k}^t \end{bmatrix} = \begin{bmatrix} \psi_{i_1}^t \\ \vdots \\ \psi_{i_k}^t \end{bmatrix} M = \Psi_{\mathrm{DC}} M = [\Phi_{\mathrm{DC}} \quad \Delta_{\mathrm{DC}}] \begin{bmatrix} S & T \\ T^t & 0 \end{bmatrix}$$

$$= [\Phi_{\mathrm{DC}} S + \Delta_{\mathrm{DC}} T^t \quad \Phi_{\mathrm{DC}} T]$$

- Since Φ_{DC} is invertible, we can obtain T from $\Phi_{\mathrm{DC}} T$.
- Afterwards, we can obtain S from $\Phi_{\mathrm{DC}} S + \Delta_{\mathrm{DC}} T^t$.
- From S and T, we know the B message symbols.

MSR Code Construction

Parameter set: $\left(\alpha = k-1, \beta = 1, B = k\alpha = \alpha(\alpha+1)\right)$ where $d = 2k - 2 = 2\alpha$

Message matrix M **(symmetric)**

$$M = \begin{bmatrix} S_1 \\ S_2 \end{bmatrix}$$

- S_1 and S_2: $(\alpha \times \alpha)$ symmetric matrices with $\binom{\alpha+1}{2}$ symbols
  - M has the $\alpha \times (\alpha + 1) = B$ symbols

Encoding matrix Ψ

$$\Psi = [\Phi \;\; \Gamma\Phi]$$

- Any d rows of Ψ are linearly independent (Ψ: $n \times d$)
- Any α rows of Φ are linearly independent (Φ: $n \times \alpha$)
- The n diagonal elements of the diagonal matrix Γ are distinct

MSR Exact-Regeneration

Theorem: Exact-regeneration of any failed node can be achieved by downloading one symbol each from any $d = 2k - 2 = 2\alpha$ nodes

Proof:

- Want to repair c_f^t of the failed node.

$$c_f^t = \psi_f^t M = \begin{bmatrix} \phi_f^t & \lambda_f \phi_f^t \end{bmatrix} \begin{bmatrix} S_1 \\ S_2 \end{bmatrix} = \phi_f^t S_1 + \lambda_f \phi_f^t S_2$$

- Get the following d symbols from d helper nodes.

$$\begin{bmatrix} c_{i_1}^t \phi_f \\ \vdots \\ c_{i_d}^t \phi_f \end{bmatrix} = \begin{bmatrix} \psi_{i_1}^t \\ \vdots \\ \psi_{i_d}^t \end{bmatrix} M \phi_f = \Psi_{\mathrm{repair}} M \phi_f$$

- Since Ψ_{repair} is invertible, we can obtain $M\phi_f = \begin{bmatrix} S_1 \phi_f \\ S_2 \phi_f \end{bmatrix}$.

- Since S_1 and S_2 are symmetric, $\left(S_i \phi_f\right)^t = \phi_f^t S_i$ for $i = 1, 2$, where we can repair $\phi_f^t S_1 + \lambda_f \phi_f^t S_2$.

MSR Exact-Reconstruction

Theorem: All the B message symbols can be recovered by connecting to any k nodes

Proof:

- Want to recover M.
- Get the following $k\alpha$ symbols from k helper nodes.

$$\begin{bmatrix} c_{i_1}^t \\ \vdots \\ c_{i_k}^t \end{bmatrix} = \begin{bmatrix} \psi_{i_1}^t \\ \vdots \\ \psi_{i_k}^t \end{bmatrix} M = \Psi_{\mathrm{DC}} M = [\Phi_{\mathrm{DC}} \;\; \Gamma_{\mathrm{DC}}\Phi_{\mathrm{DC}}] \begin{bmatrix} S_1 \\ S_2 \end{bmatrix}$$
$$= [\Phi_{\mathrm{DC}} S_1 + \Gamma_{\mathrm{DC}}\Phi_{\mathrm{DC}} S_2]$$

- Post multiply with Φ_{DC}^t,

$$[\Phi_{\mathrm{DC}} S_1 + \Gamma_{\mathrm{DC}}\Phi_{\mathrm{DC}} S_2]\Phi_{\mathrm{DC}}^t = \Phi_{\mathrm{DC}} S_1 \Phi_{\mathrm{DC}}^t + \Gamma_{\mathrm{DC}}\Phi_{\mathrm{DC}} S_2 \Phi_{\mathrm{DC}}^t$$
$$= P + \Gamma_{\mathrm{DC}} Q$$

  - $P = \Phi_{\mathrm{DC}} S_1 \Phi_{\mathrm{DC}}^t$ and $Q = \Phi_{\mathrm{DC}} S_2 \Phi_{\mathrm{DC}}^t$ are symmetric.

MSR Exact-Reconstruction

Proof (continued):

- Know $A = P + \Gamma_{\mathrm{DC}} Q$ where P and Q are symmetric.
- By comparing (i, j)-th element and (j, i)-th element of $A = P + \Gamma_{\mathrm{DC}} Q$, all the nondiagonal elements of P and Q are known by
  - $A_{i,j} = P_{i,j} + \lambda_i Q_{i,j}$ and $A_{j,i} = P_{j,i} + \lambda_j Q_{j,i} = P_{i,j} + \lambda_j Q_{i,j}$
  - $Q_{i,j} = \frac{A_{i,j} - A_{j,i}}{\lambda_i - \lambda_j}$
- The elements in the i-th row of $P = \Phi_{\mathrm{DC}} S_1 \Phi_{\mathrm{DC}}^t$ except $P_{i,i}$ are given by

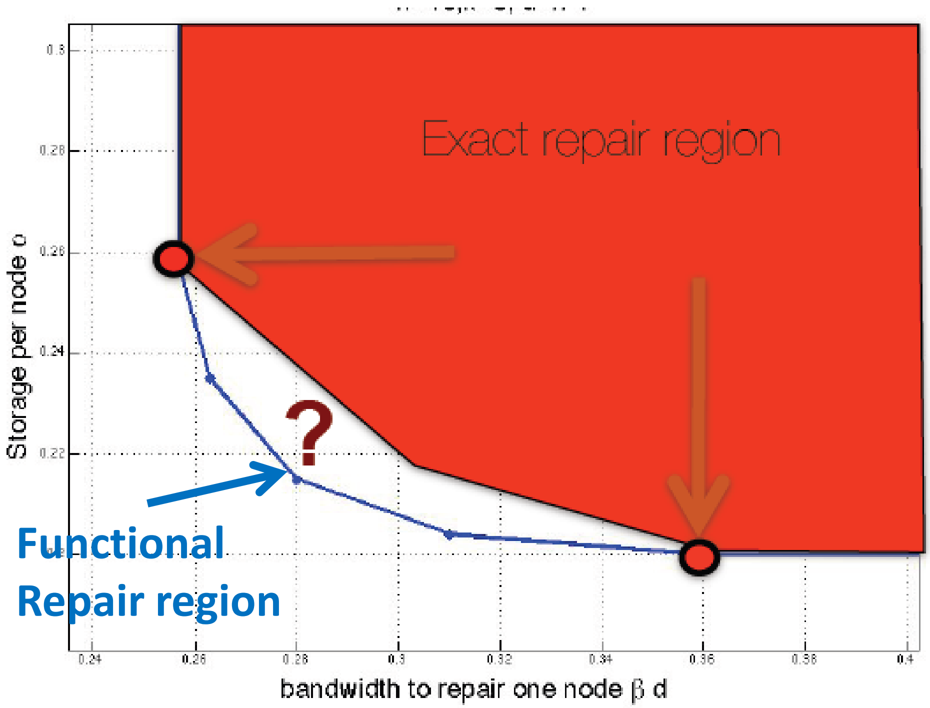
$$\phi_i^t S_1 [\phi_1 \cdots \phi_{i-1} \; \phi_{i+1} \cdots \phi_{\alpha+1}]$$

- Since $[\phi_1 \cdots \phi_{i-1} \; \phi_{i+1} \cdots \phi_{\alpha+1}]$ is invertible, know $\phi_i^t S_1$ for $1 \le i \le k$.
- Selecting the first α of these, know $\begin{bmatrix} \phi_1^t \\ \vdots \\ \phi_\alpha^t \end{bmatrix} S_1 = \Phi'_{\mathrm{DC}} S_1$.
- Since Φ'_{DC} is invertible, reconstruct S_1.
- Similarly, we can reconstruct S_2 from Q.

Open Problems

- Gap between functional regenerating codes and exact regenerating codes at interior points (*)

- Explicit code constructions for $[n, k, d < 2k - 2]$ at the MSR point
  - For high rate code
  - Not achievable if $\beta = 1$ (**)



* Tian, *IEEE Journal on Selected Areas in Communications*, 2014.
** Shah, Rashmi, Kumar, and Ramchandran, *IEEE ITW* 2010.