Coding Techniques for Distributed Storage Systems

Jung-Hyun Kim, {jh.kim06}@yonsei.ac.kr, Hong-Yeop Song

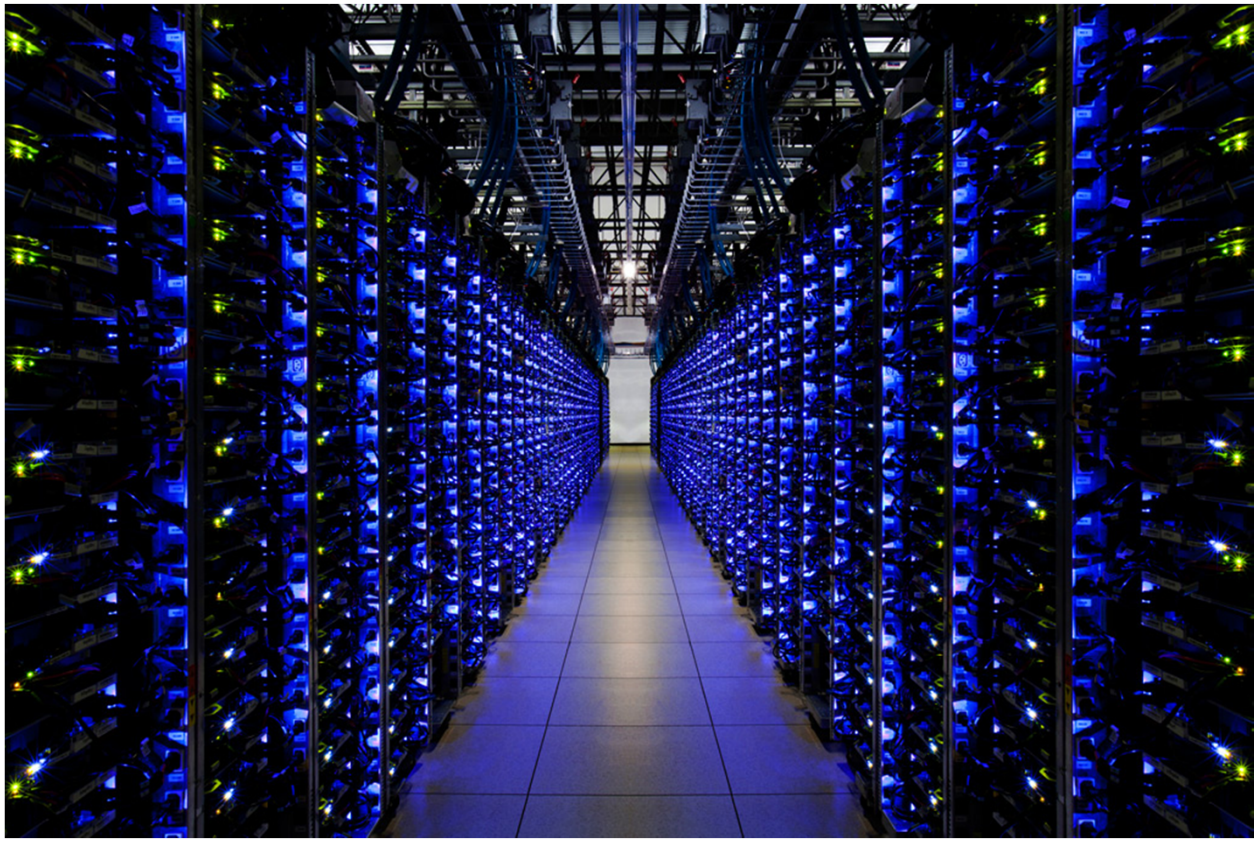
Yonsei Univ. Seoul, KOREA

3rd CITW, Oct. 25. 1, 2013

Contents

What's the Problem? Too big to manage!

- Big data storage
  - Warehouse-scale data center
  - Thousands of servers, Petabytes of disc space

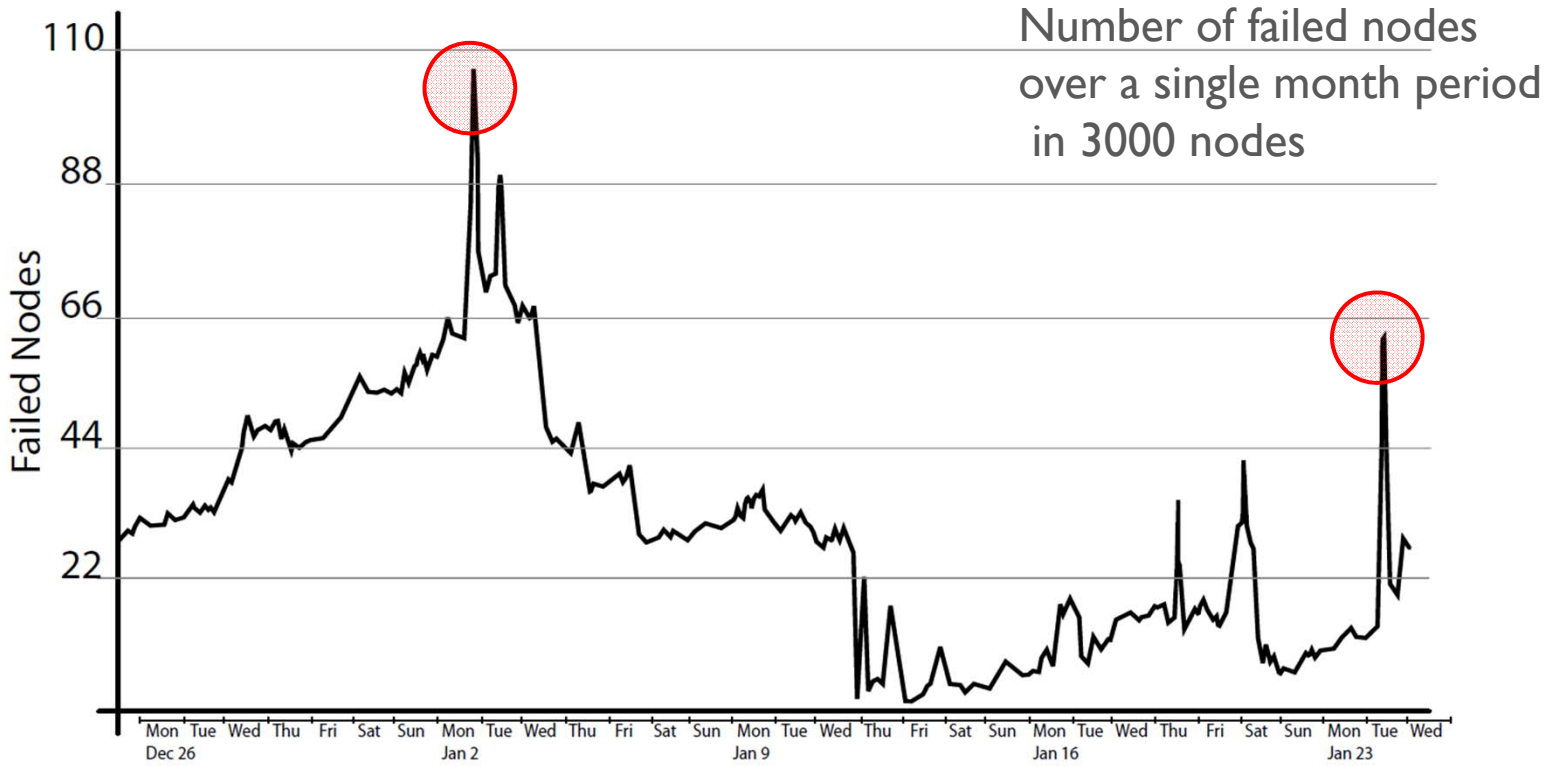


Google
Data Center

http://www.google.com/about/datacenters/

What's the Problem? It needs repairing!

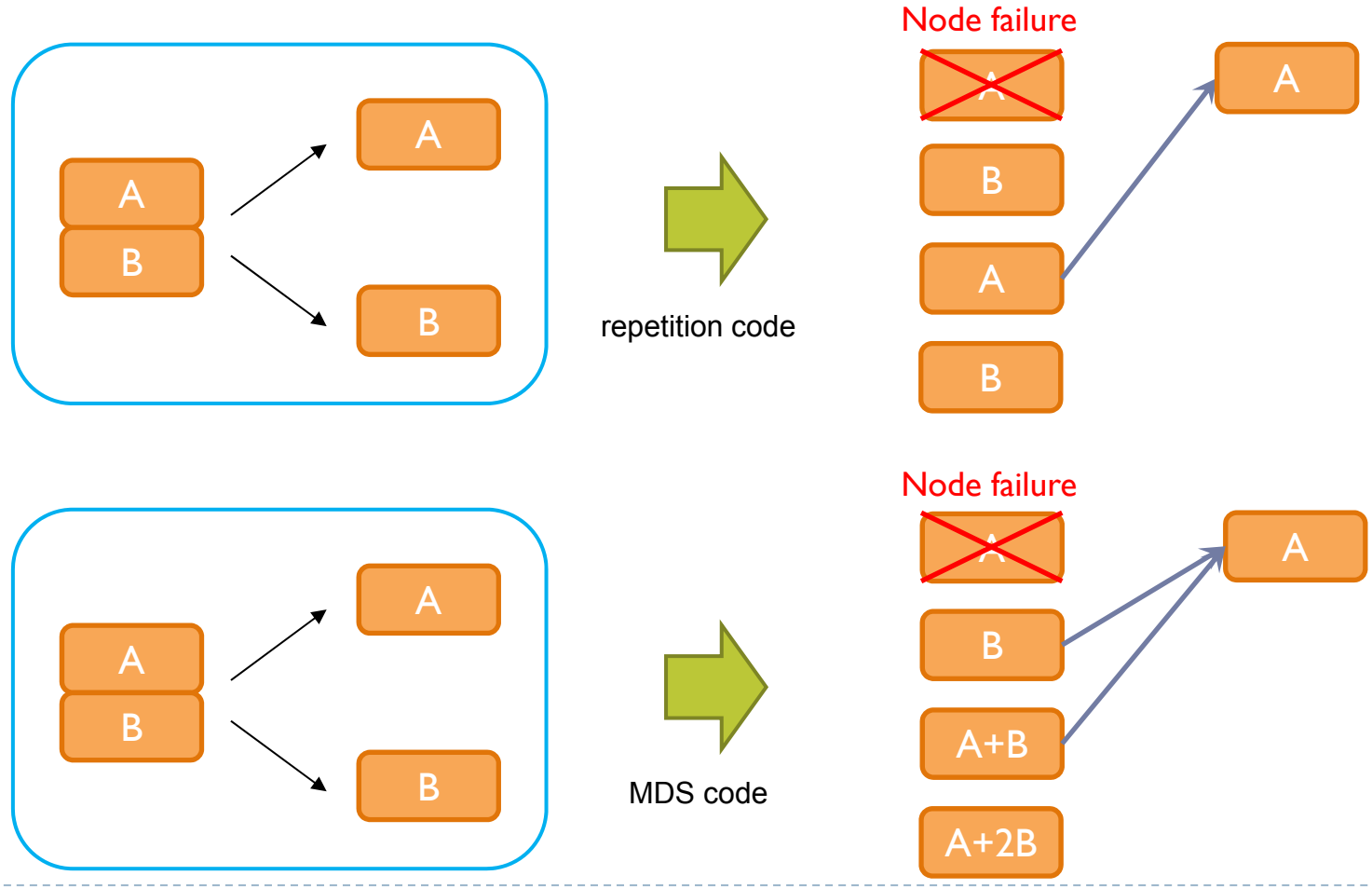
- At ***Facebook***, it is quite typical to have 20 or more node failures per day.



M. Sathiamoorthy, M. Asteris, D. Papailiopoulos, A. G. Dimakis, R. Vadali, S. Chen, D. Borthakur, "XORing Elephants: Novel Erasure Codes for Big Data," in Proc. of the 39th International Conf. on Very Large Data Bases, Aug. 2013.

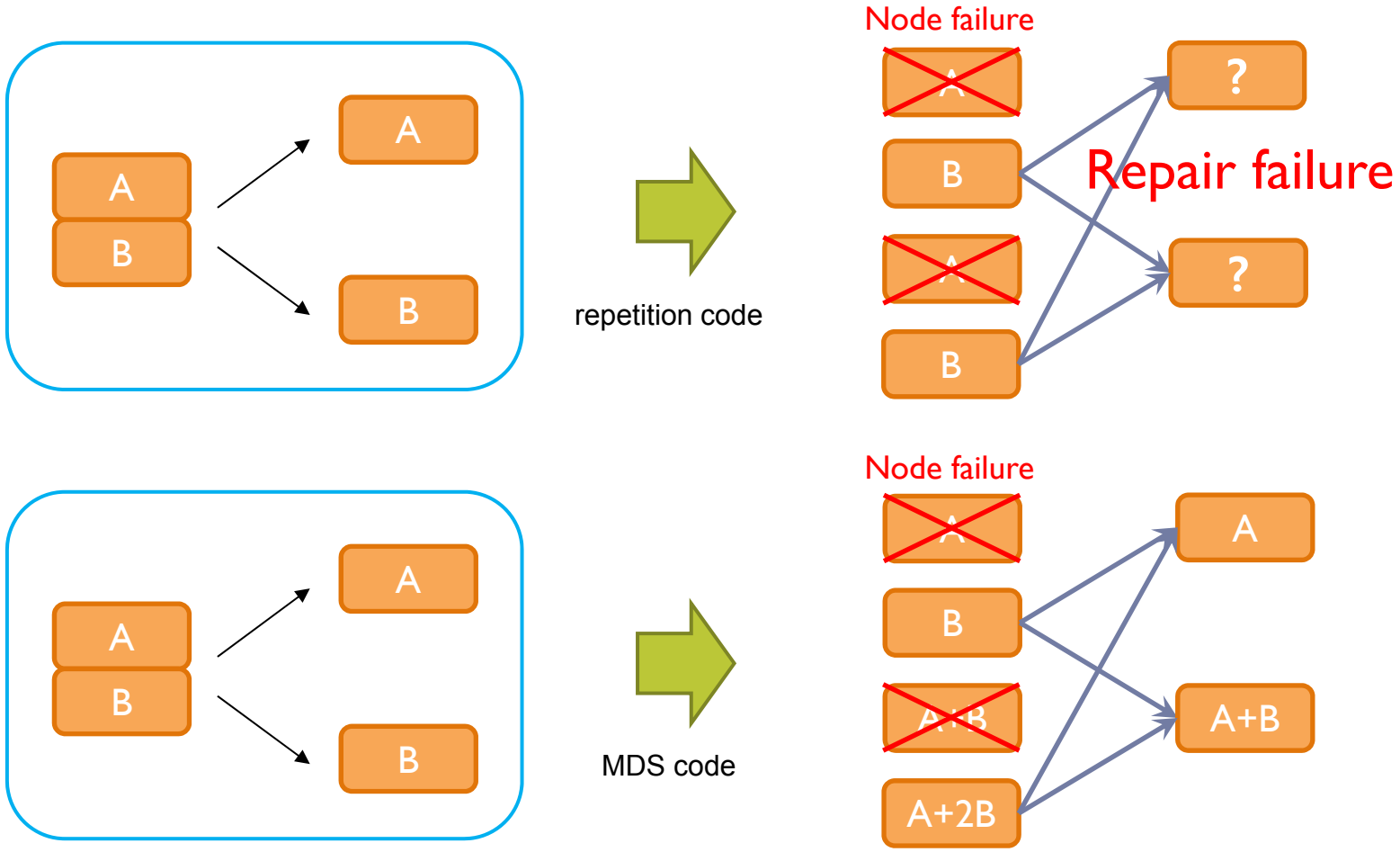
Traditional Codes for DSS

- How to regenerate failed nodes?
  - Repetition codes have higher ***efficiency*** than MDS codes



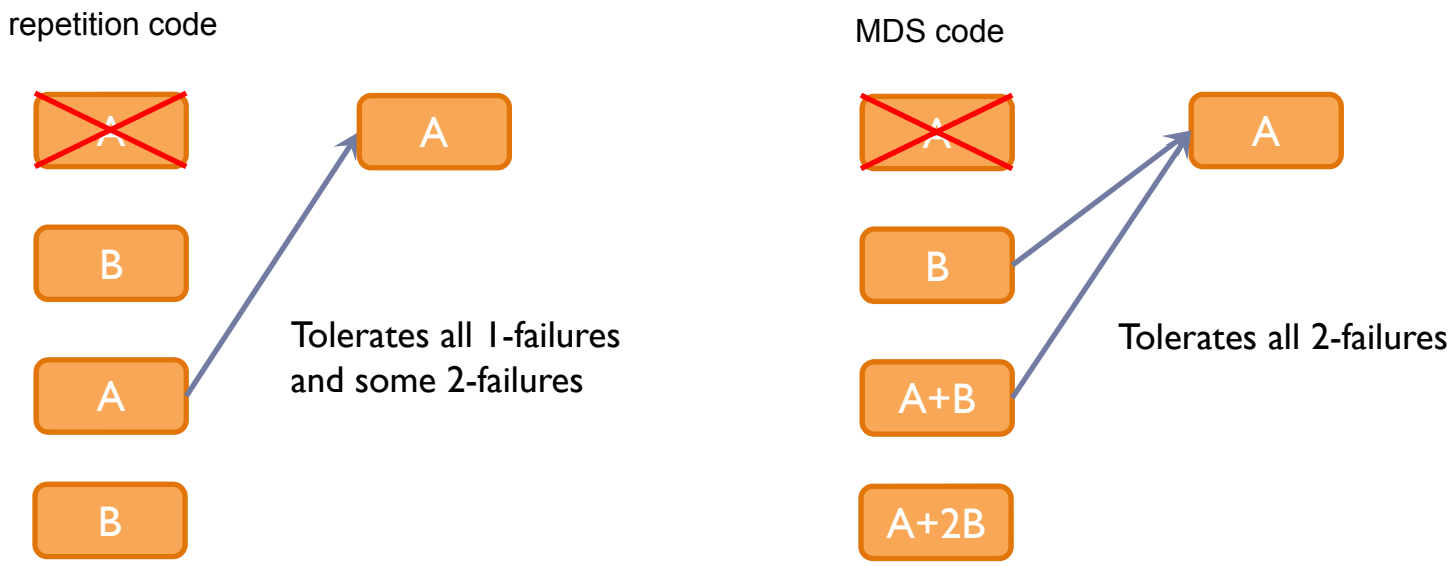
Traditional Codes for DSS

- How to regenerate failed nodes?
  - MDS codes have higher ***reliability*** than repetition codes



Traditional Codes for DSS

- Which one is better?



***Efficient* Repair**
***Low* Reliability**

***Inefficient* Repair**
***High* Reliability**

Q. Tradeoff between "efficiency" and "reliability"?

Regenerating Codes for DSS

- Regenerating Codes?

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE INFOCOM 2007 proceedings.

Network Coding for Distributed Storage Systems

Alexandros G. Dimakis, P. Brighten Godfrey, Martin J. Wainwright and Kannan Ramchandran
Department of Electrical Engineering and Computer Science,
University of California, Berkeley, CA 94704.
Email: {adim, pbg, wainwrig, kannanr}@eecs.berkeley.edu

***Abstract*— Peer-to-peer distributed storage systems provide reliable access to data through redundancy spread over nodes across the Internet. A key goal is to minimize the amount of bandwidth used to maintain that redundancy.**

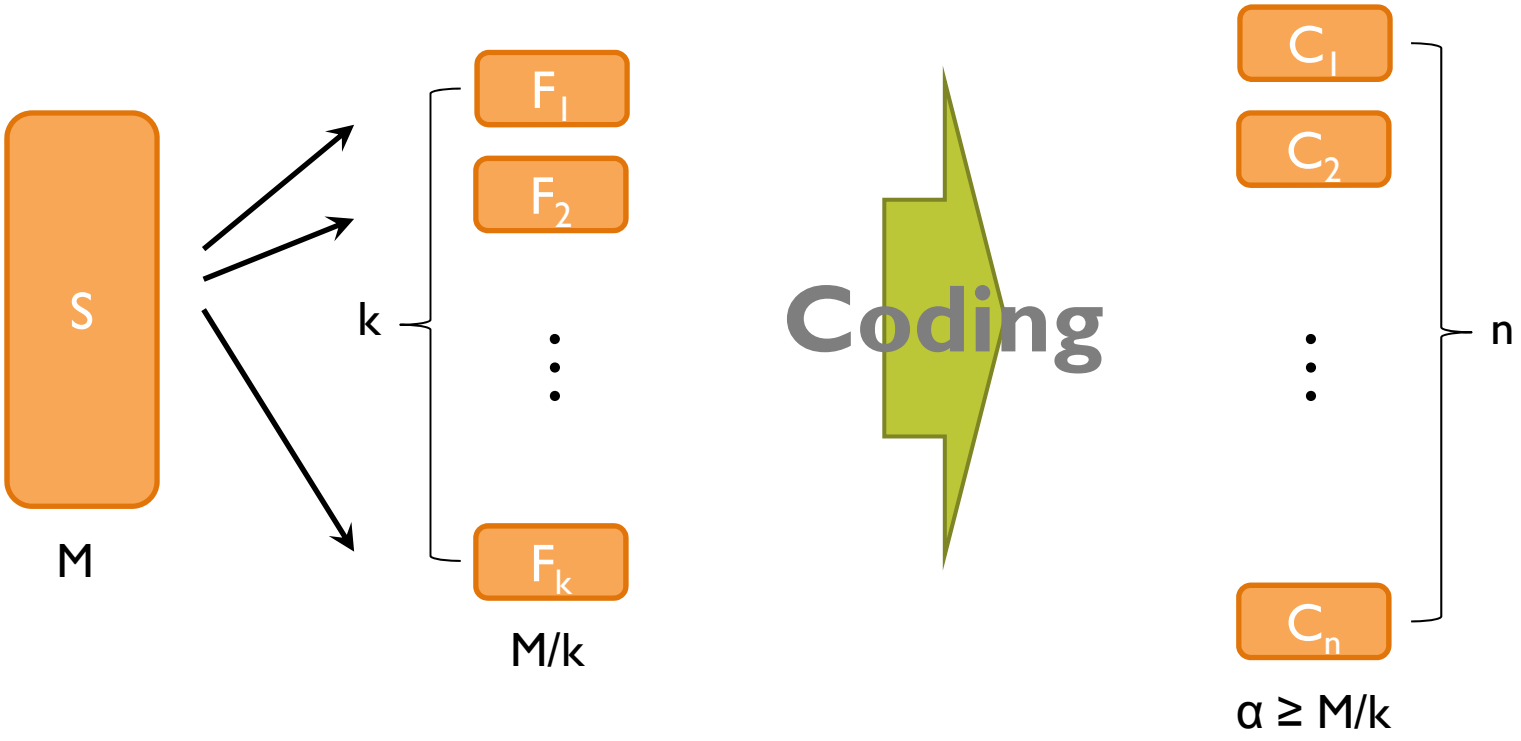
encoded fragment when we only have access to erasure encoded fragments?

In the *naive strategy*, the node which will store the new fragment—which we will call the *newcomer*—

A. G. Dimakis, P. B. Godfrey, Y. Wu, M. Wainwright and K. Ramchandran, "Network Coding for Distributed Storage Systems," IEEE Proc. INFOCOM, (Anchorage, Alaska), May 2007.

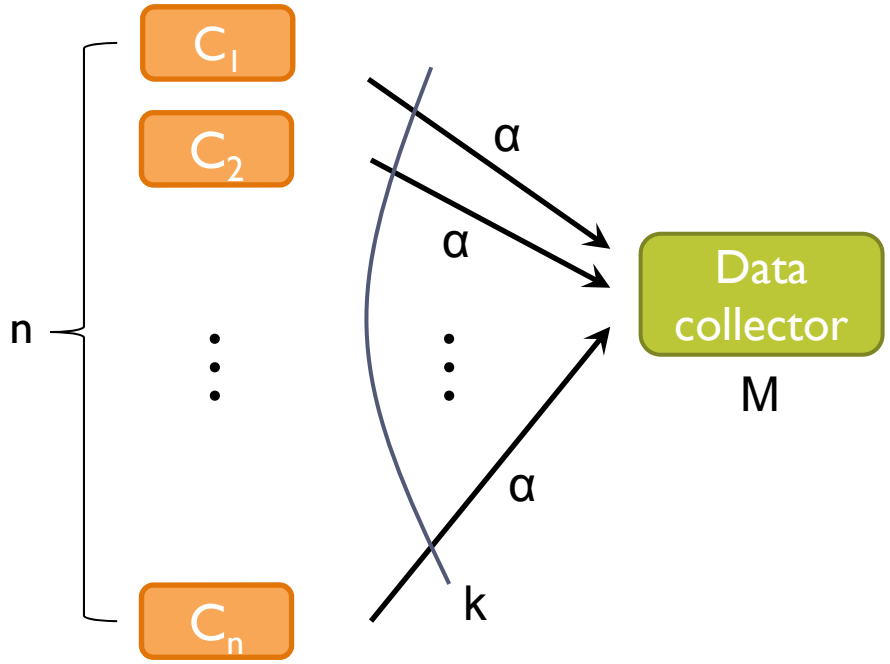
Regenerating Codes for DSS

- Regenerating Codes Framework
  - Storing *n* coded files of *k* original files at *n* distributed nodes



Regenerating Codes for DSS

- Regenerating Codes Framework
  - In aspect of data collection



n : # of storage nodes

k : # of storage nodes for data collection

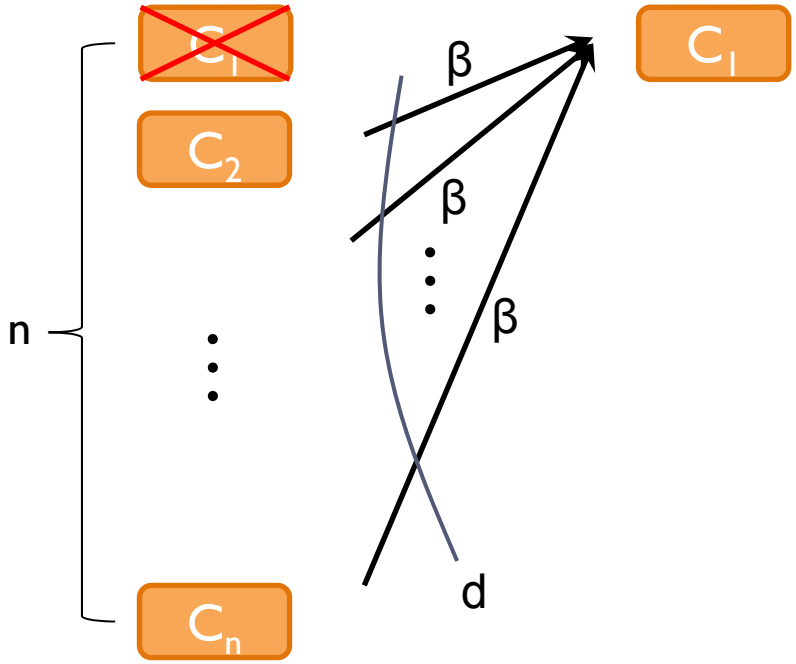
α : storage size

M : data size

A. G. Dimakis, P. B. Godfrey, Y. Wu, M. Wainwright and K. Ramchandran, "Network Coding for Distributed Storage Systems," IEEE Transactions on Information Theory, Vol. 56, Issue 9, Sept. 2010.

Regenerating Codes for DSS

- Regenerating Codes Framework
  - In aspect of node repair



d : # of storage nodes for node repair

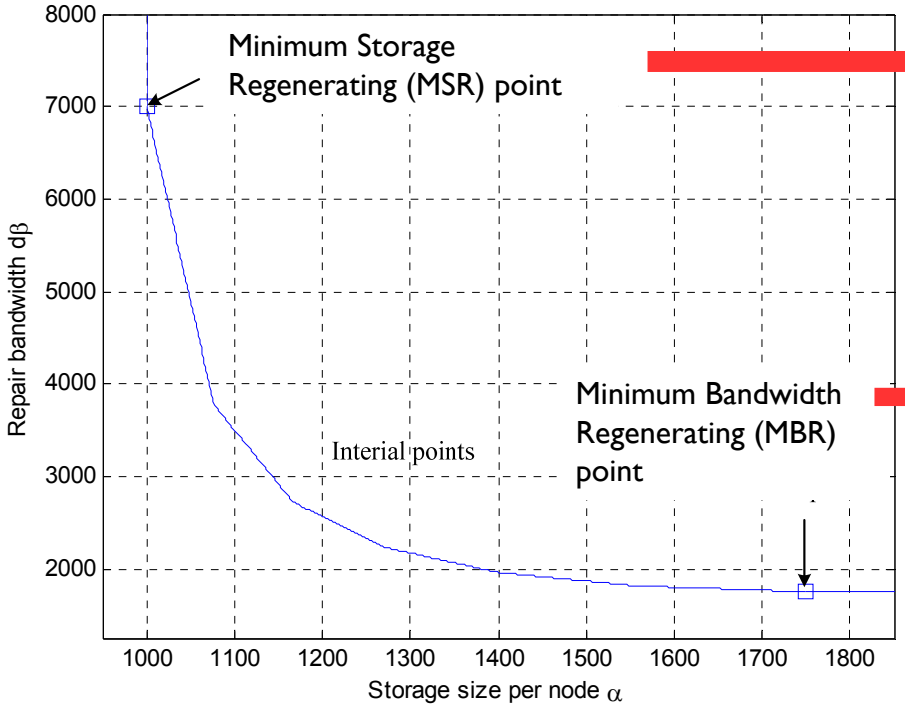
β : download size

dβ (= γ) : repair bandwidth

A. G. Dimakis, P. B. Godfrey, Y. Wu, M. Wainwright and K. Ramchandran, "Network Coding for Distributed Storage Systems," IEEE Transactions on Information Theory, Vol. 56, Issue 9, Sept. 2010.

Regenerating Codes for DSS

- Reducing ***storage size and repair bandwidth***
  - Based on the min-cut bound : $\sum_{i=0}^{k-1} \min\{(d-i)\beta, \alpha\} \geq M$



MSR codes

$$\alpha_{MSR} = \frac{M}{k}$$

$$\gamma_{MSR} = \frac{Md}{k(d-k+1)}$$

MBR codes

$$\alpha_{MBR} = \frac{2Md}{k(2d-k+1)}$$

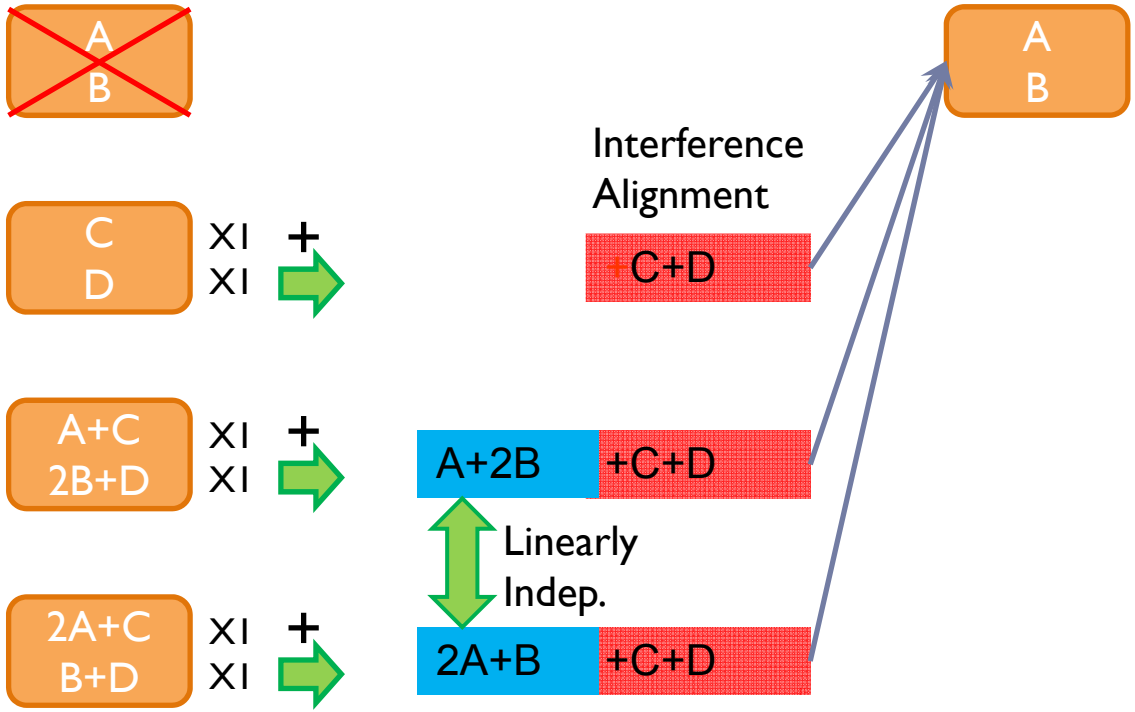
$$\gamma_{MBR} = \frac{2Md}{k(2d-k+1)}$$

< Tradeoff between storage size and repair bandwidth (M=7000, n=15, k=7, d=7) >

N. B. Shah, K.V. Rashmi, P.V. Kumar, and K. Ramchandran, "Distributed Storage Codes With Repair-by-Transfer and Nonachievability of Interior Points on the Storage-Bandwidth Tradeoff," IEEE Trans. Inf. Theory, vol. 58, no. 3, pp. 1837–1852, Mar. 2012.

Regenerating Codes for DSS

- Minimum Storage Regenerating (MSR) codes
  - Interference Alignment method, Product-Matrix method, etc.



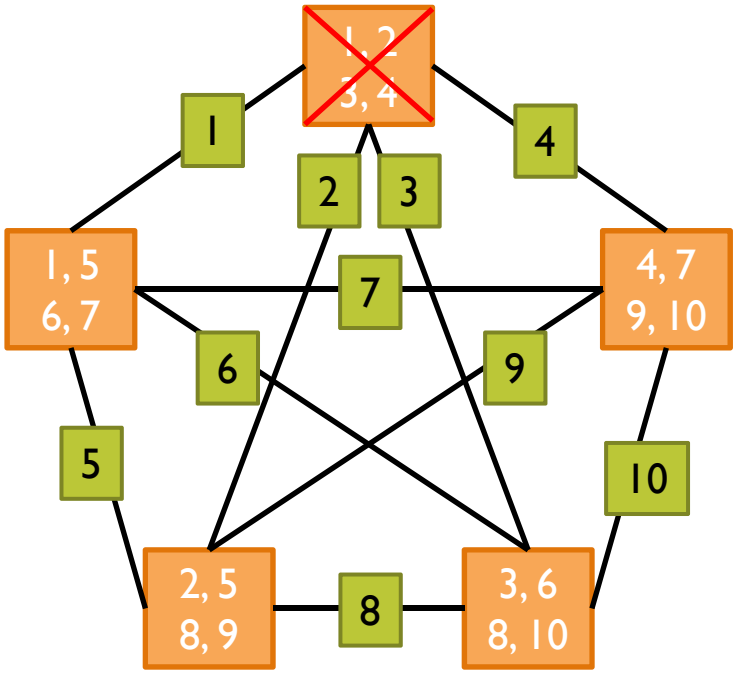
$$\alpha_{MSR} = \frac{M}{k} = \frac{4}{2} = 2$$

$$\gamma_{MSR} = \frac{Md}{k(d-k+1)} = \frac{4 \cdot 3}{2(3-2+1)} = 3$$

N. B. Shah, K.V. Rashmi, P.V. Kumar, and K. Ramchandran, "Interference Alignment in Regenerating Codes for Distributed Storage: Necessity and Code Constructions," IEEE Trans. Inf. Theory, vol. 58, no. 4, pp. 2134–2158, April 2012.
K.V. Rashmi, N. B. Shah, and P.V. Kumar, "Optimal exact-regenerating codes for the MSR and MBR points via a product-matrix construction," IEEE Trans. Inf. Theory, vol. 57, no. 8, pp. 5227-5239, Aug. 2011.

Regenerating Codes for DSS

- Minimum Bandwidth Regenerating (MBR) codes
  - Repair-by-Transfer method, Product-Matrix method, etc.



$$\alpha_{MBR} = \frac{2Md}{k(2d-k+1)} = \frac{2 \cdot 9 \cdot 4}{3(2 \cdot 4 - 3 + 1)} = 4$$

$$\gamma_{MBR} = \frac{2Md}{k(2d-k+1)} = \frac{2 \cdot 9 \cdot 4}{3(2 \cdot 4 - 3 + 1)} = 4$$

K. W. Shum, and Y. Hu, "Functional-Repair-by-Transfer Regenerating Codes," in Proc. of 2012 IEEE International Symposium on Information Theory, Cambridge, MA, July 2012.

K. V. Rashmi, N. B. Shah, and P. V. Kumar, "Optimal exact-regenerating codes for the MSR and MBR points via a product-matrix construction," IEEE Trans. Inf. Theory, vol. 57, no. 8, pp. 5227-5239, Aug. 2011.

Codes with Local Regeneration

- Locality?

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 58, NO. 11, NOVEMBER 2012

On the Locality of Codeword Symbols

Parikshit Gopalan, Cheng Huang, Huseyin Simitci, and Sergey Yekhanin

***Abstract*—Consider a linear $[n, k, d]_q$ code $\mathcal{C}$. We say that the ith coordinate of $\mathcal{C}$ has locality r, if the value at this coordinate can be recovered from accessing some other r coordinates of $\mathcal{C}$. Data storage applications require codes with small redundancy, low locality for information coordinates, large distance, and low locality for parity coordinates. In this paper, we carry out an in-depth study of the relations between these parameters. We establish a tight bound for the redundancy $n - k$ in terms of the message length, the distance, and the locality of information coordinates. We refer to codes attaining the bound as optimal. We prove some structure theorems about optimal codes, which are particularly strong for small distances. This gives a fairly complete picture of the trade**

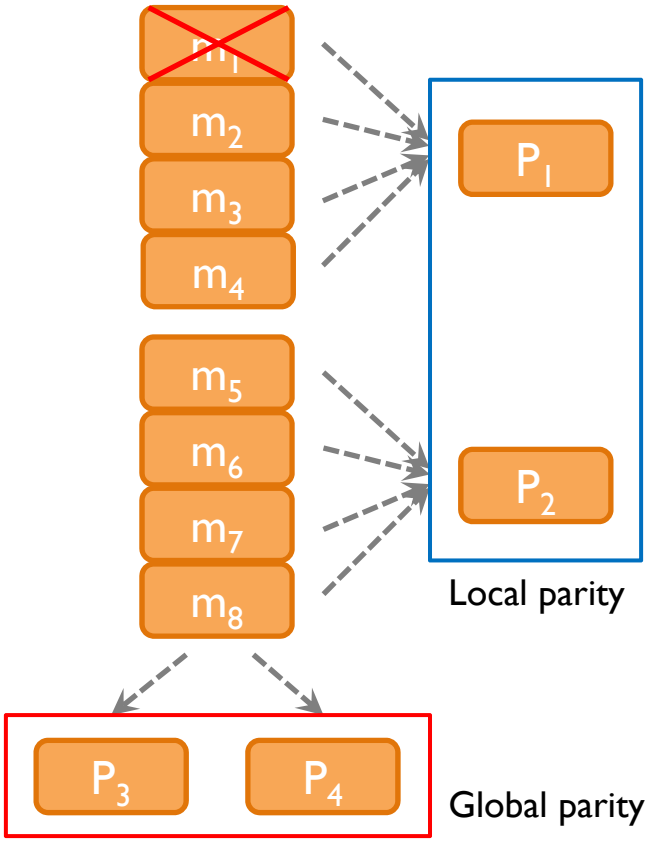
small value of locality is particularly important for information packets.

These considerations lead us to introduce the concept of an (r, d)-*code*, i.e., a linear code of distance d, where all information symbols have locality at most r. Storage systems based on (r, d)-codes provide fast recovery of information packets from a single node failure (typical scenario), and ensure that no data are lost even if up to $d - 1$ nodes fail simultaneously [9]. One specific class of (r, d)-codes called Pyramid Codes has been considered in [8].

P. Gopalan, C. Huang, H. Simitci, and S. Yekhanin, "On the Locality of Codeword Symbols," IEEE Trans. Inf. Theory, vol. 58, no. 11, pp. 6925–6934, Nov. 2012.

Codes with Local Regeneration

- Local Reconstruction Codes (LRC)
  - High ***access efficiency*** for node repair



(r, δ) locality
: given that at the most $(\delta-1)$ symbols are erased – can be deduced by reading at most r other surviving symbols.

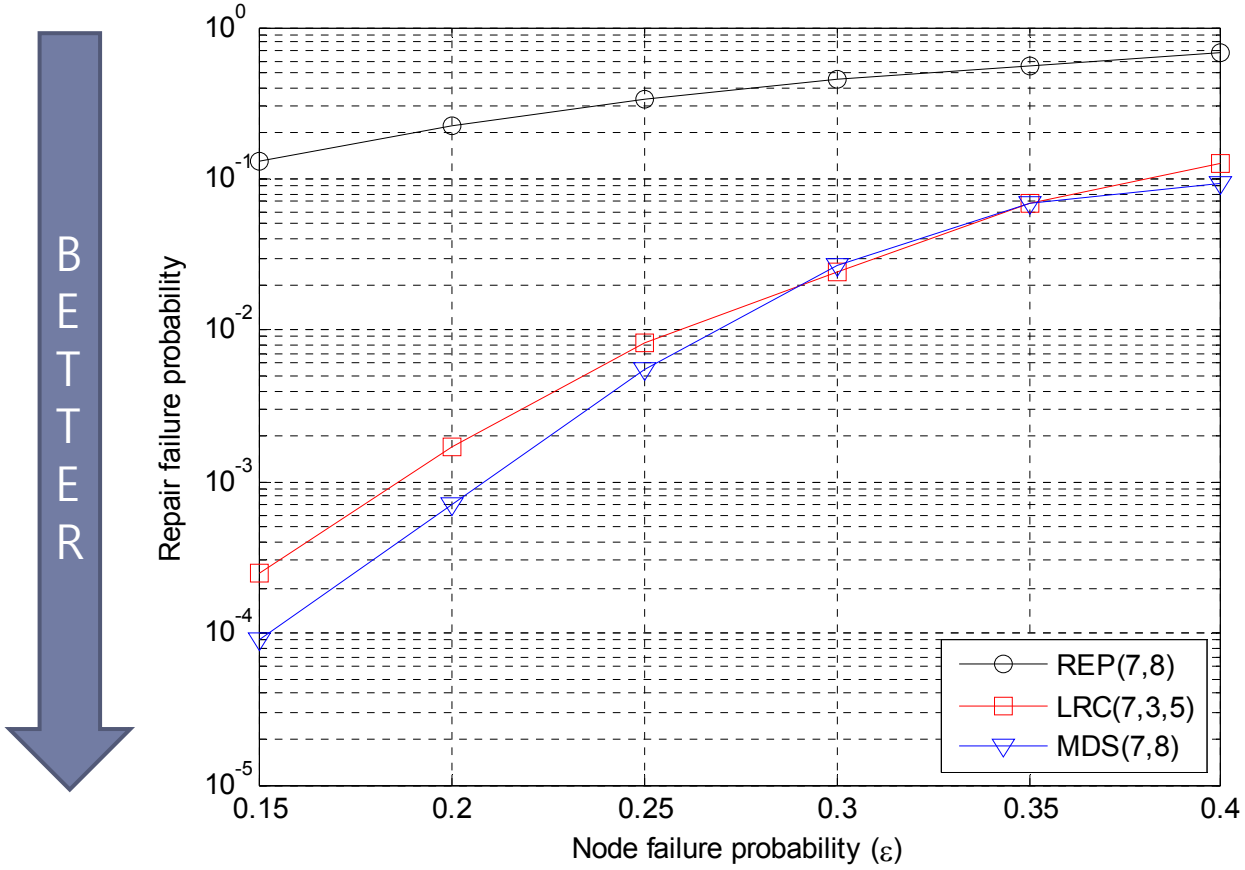
$$d_{min}(C_{LRC}) \leq n - \left\lceil \frac{M}{\alpha} \right\rceil + 1 - \left(\left\lceil \frac{M}{r\alpha} - 1 \right\rceil \right) (\delta - 1)$$

C. Huang, H. Simitci, Y. Xu, A. Ogus, B. Calder, P. Gopalan, J. Li, and S. Yekhanin, "Erasure Coding in Windows Azure Storage," 2012 USENIX Annual Technical Conference, 2012.

A. S. Rawat, N. Silberstein, O. O. Koyluoglu, and S. Vishwanath, "Optimal locally repairable codes with local minimum storage regeneration via rank-metric codes," in Information Theory and Applications Workshop (ITA), San Diego, CA, Feb. 2013.

Performance Comparison

- Repair failure probability for different node failure probability



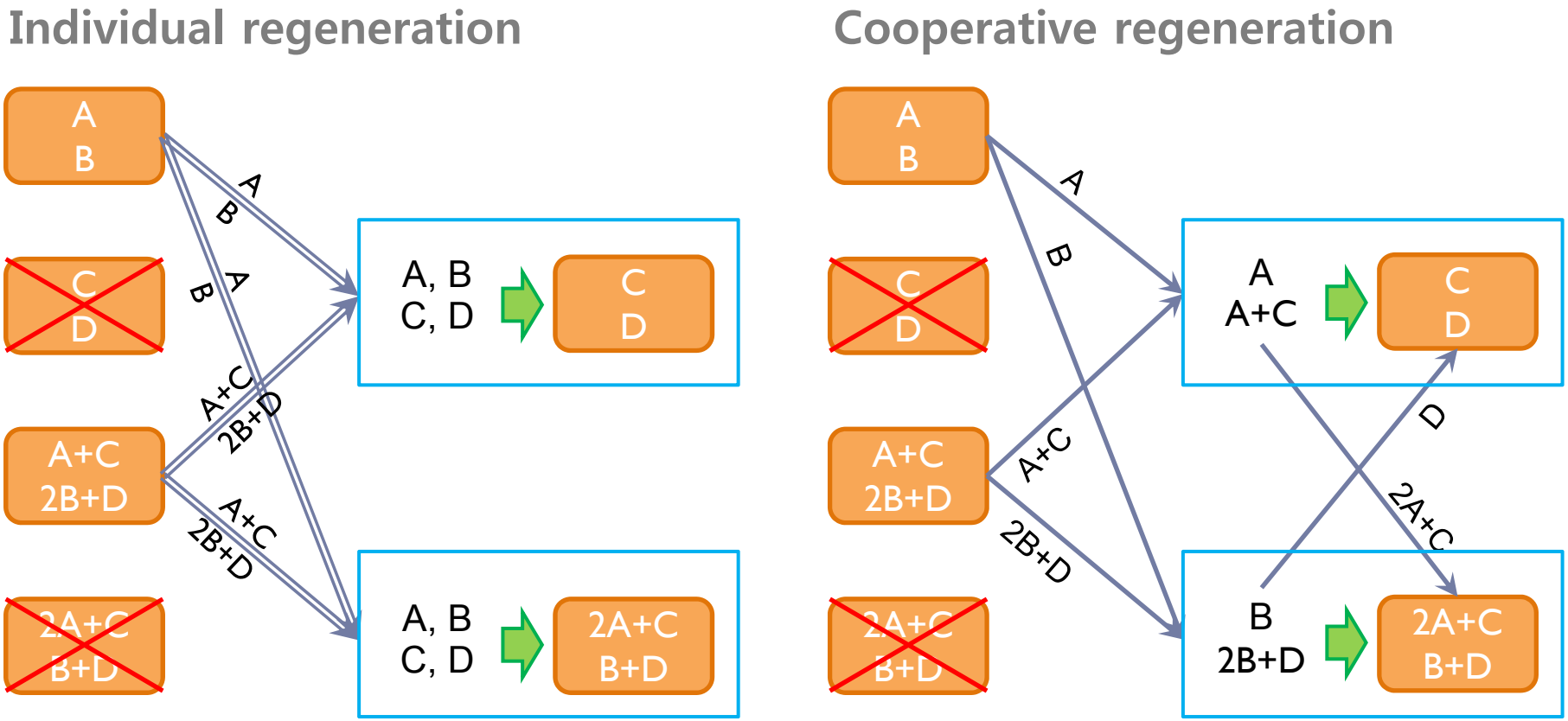
Node failure prob. : the probability that a node is unavailable

Repair failure prob. : the probability that any newcomer nodes can not repair the original data symbol from coded data symbols of surviving storage nodes

Jung-Hyun Kim, Jin Soo Park, Ki-Hyeon Park, Inseon Kim, Mi-Young Nam, Hong-Yeop Song, "Reliability Comparison of Various Regenerating Codes for Cloud Services", International Conference on ICT Convergence 2013 (ICTC2013), Jeju Island, Korea, October 13-16, 2013.

Advanced Techniques

- Reducing ***repair bandwidth*** using cooperation

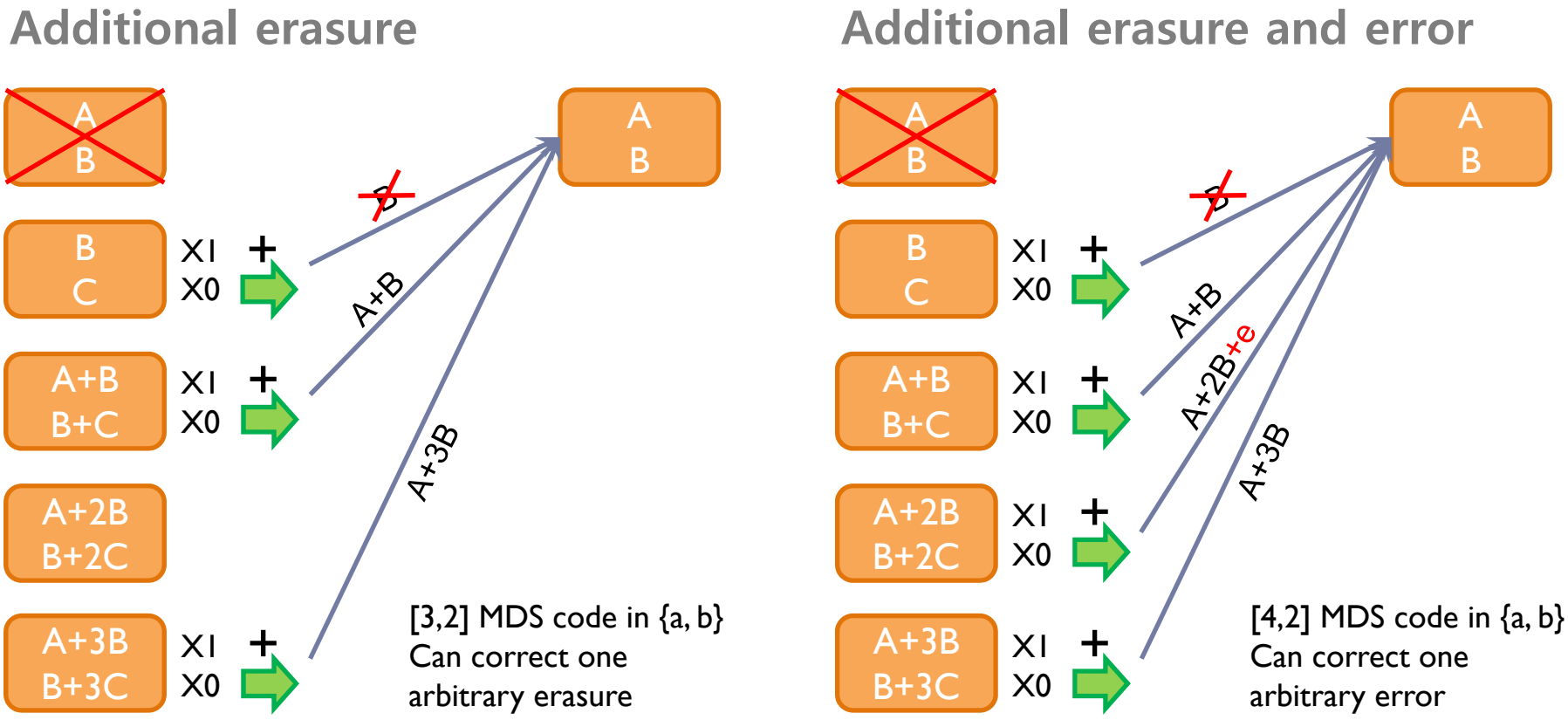


K. W. Shum, "Cooperative Regenerating Codes for Distributed Storage Systems," in Proc. of 2011 International Conference on Communications, Kyoto, June 2011.

K. W. Shum and Y. Hu, "Cooperative Regenerating Codes," IEEE Trans. Inf. Theory, vol. 59, no. 11, pp. 7229–7258, Nov. 2013.

Advanced Techniques

- Handling ***erasures and errors*** during the data construction and node repair operations



K.V. Rashmi, N. Shah, K. Ramchandran, and P. Kumar, "Regenerating codes for errors and erasures in distributed storage," in Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on, 2012, pp. 1202–1206.

Thank You!

Questions?