Your Data is in the Cloud: Who Exactly is Looking After It ?

P Vijay Kumar

Dept of Electrical Communication Engineering
Indian Institute of Science

IISc Open Day

March 4, 2017

Your Data is in the Cloud: Who Exactly is Looking After It ... Exactly ?

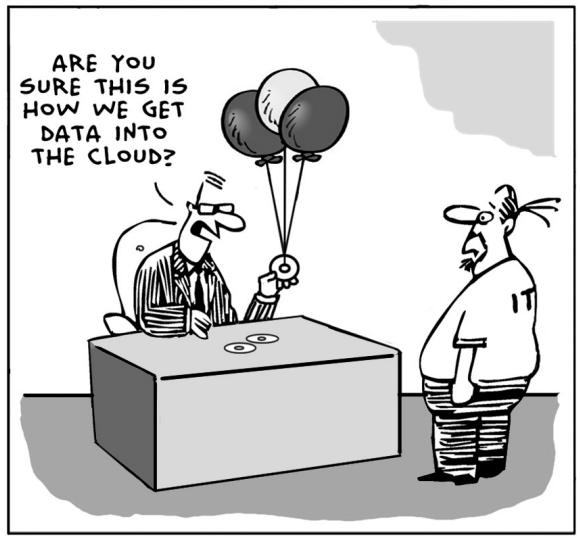
P Vijay Kumar

Dept of Electrical Communication Engineering
Indian Institute of Science

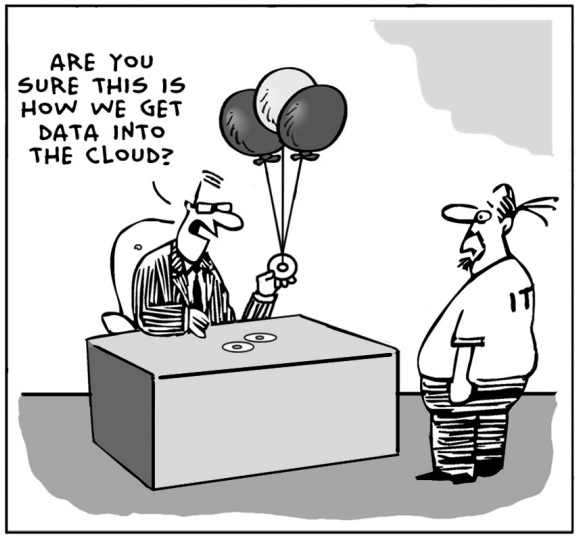
IISc Open Day

March 4, 2017

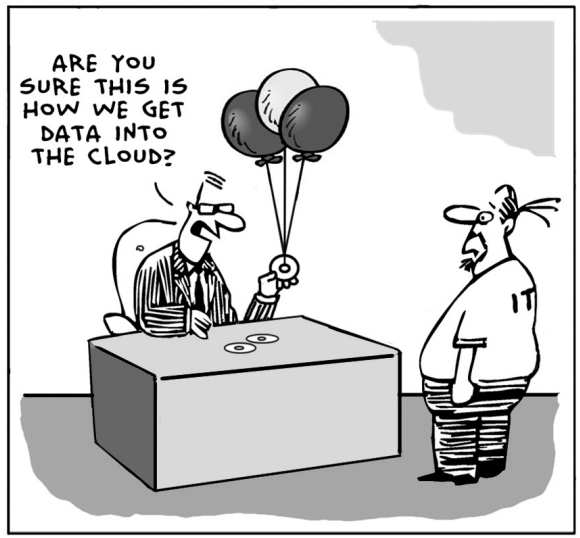
Cloud Storage



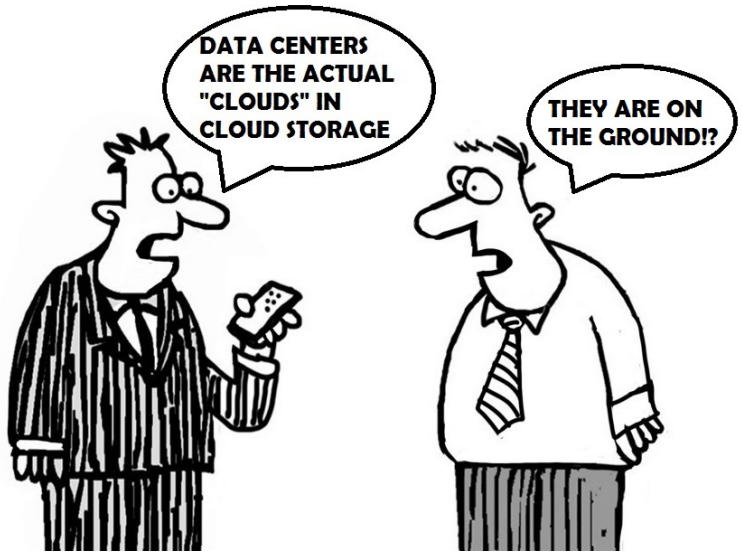
Cloud Storage



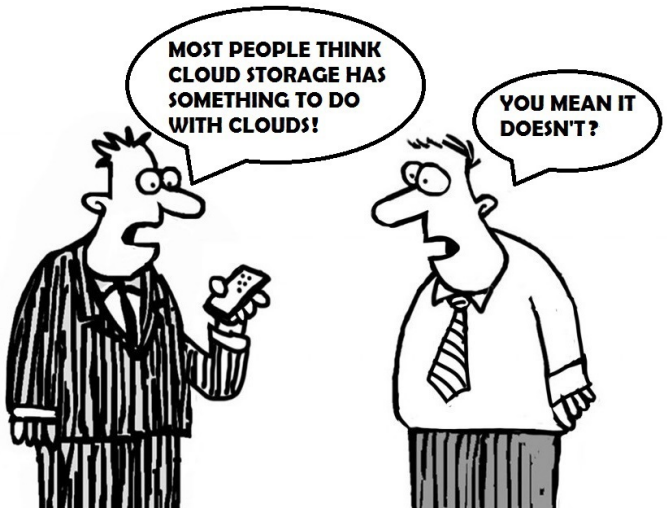
Cloud Storage

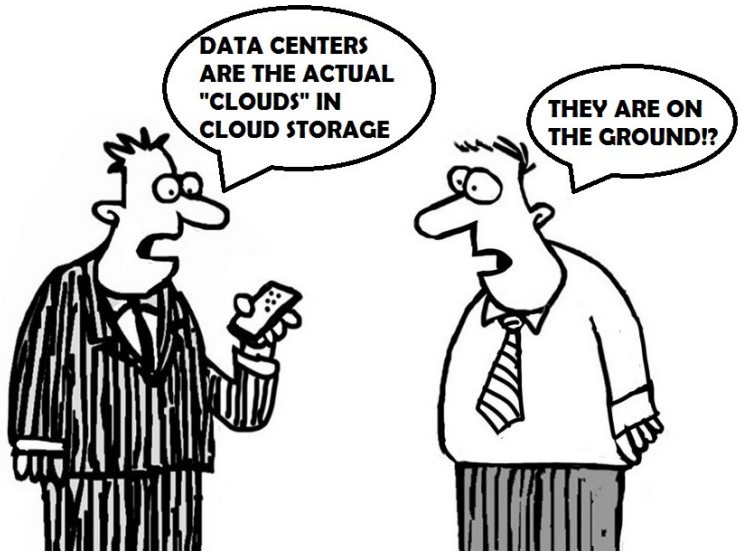


Cloud $\Leftrightarrow$ Data Centre



Cloud $\Leftrightarrow$ Data Centre

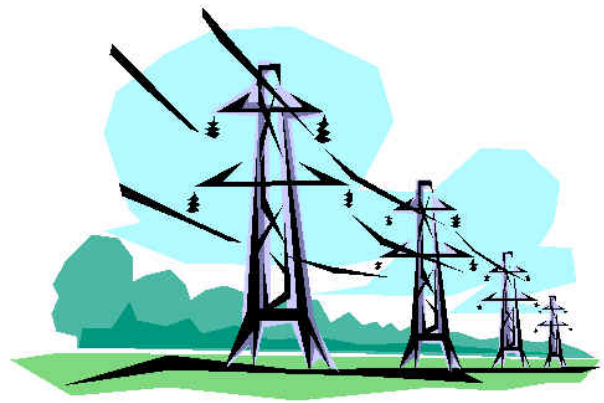




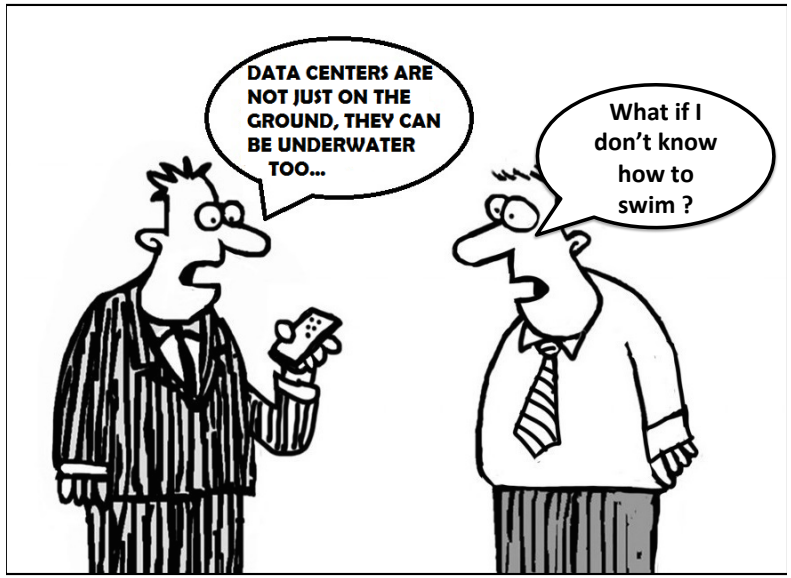
Data Centre in Utah

Data Center

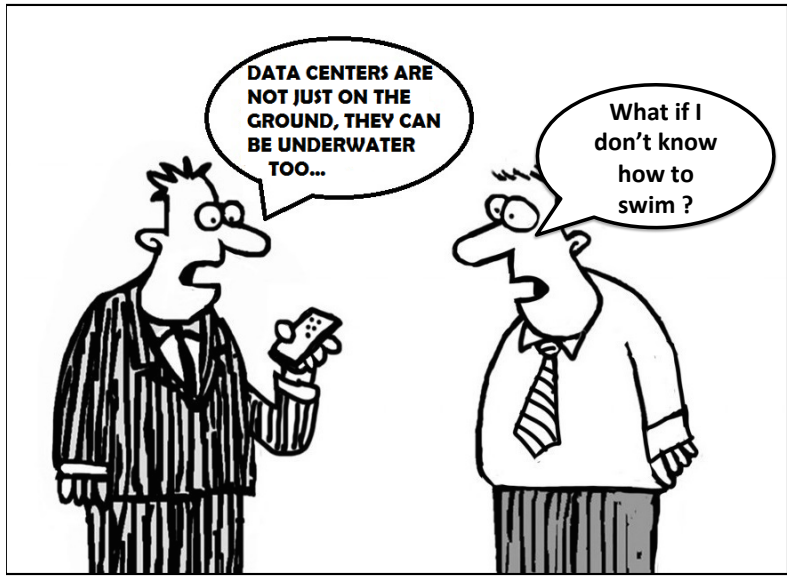
- Some facts about NSA Data Center at Utah
  - Storage capacity is around 3-12 Exabytes (1 Exabyte = 1 billion GB!!!)
  - Estimated maintenance cost of $40 million per year, 65MW of electricity
  - Uses 1.7 million gallons of water per day



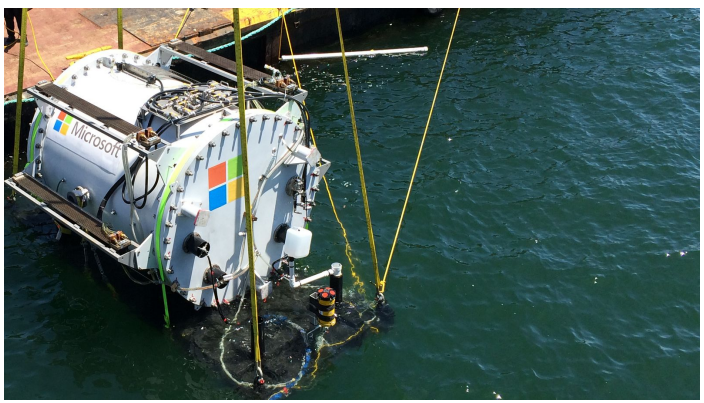
Underwater Data Center

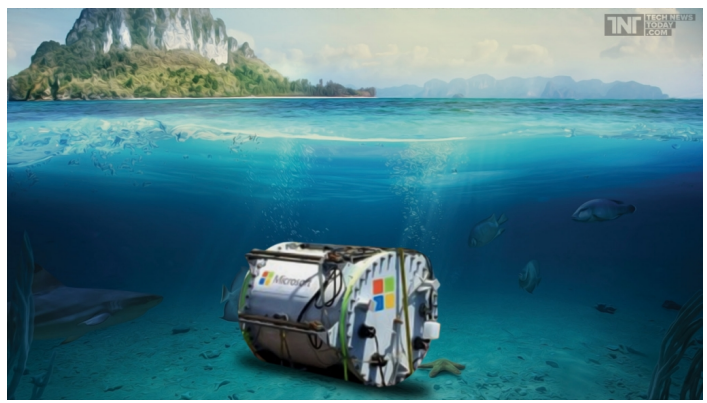


Underwater Data Center



- Microsoft has deployed a data center underwater – easy access to cooling and renewable energy sources!!!





Data Loss

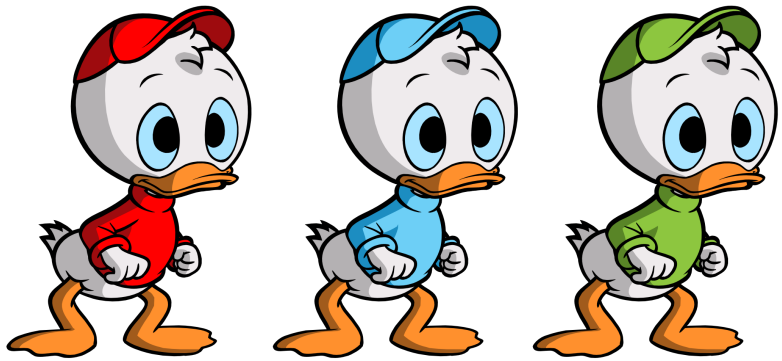
- Data pertaining to a single file is distributed across storage nodes
- Nodes are inexpensive storage devices
- Data may be lost/unavailable because nodes are
  - Prone to failure,
  - Down for maintenance,
  - Busy serving other demands

Data Loss

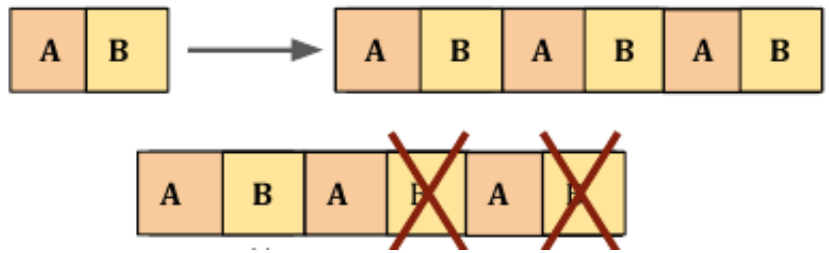
- Protection against data loss or data unavailability is clearly essential

Simple Solution: Data Replication

- In this example, $k = 2$ units of raw data 'A' and 'B' are replicated thrice to get $n = 6$ coded units which are stored across 6 nodes



- To maintain same level of reliability, a failed node should be replaced by a new node



- Here, we can recover from loss of any two nodes

Figure: $[6, 2]$ replication code

Replication Sounds Dumb

Can't we do something better?
... more interesting ?

Erasure Codes to the Rescue!

- Storage Overhead ($\frac{n}{k}$) determines the efficiency of storage
- It is possible to have Erasure Codes having same reliability as replication, but with arbitrarily low storage overhead

Erasure Code

- In an $[n, k]$ erasure code, k units of data are encoded to get n coded units
- An maximum distance separable (MDS) code can recover in the presence of any $n - k$ reassured units

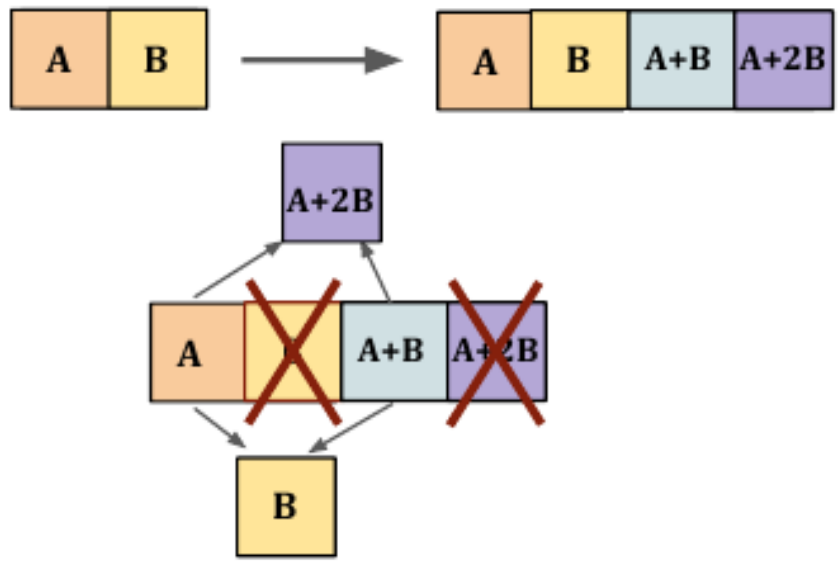


Figure: $[4, 2]$ MDS code

Replication vs Erasure Code

Code	Storage O/h	Reliability
[6,2] 3-rep	3x	2
[4,2] RS	2x	2
[6,4] RS	1.5x	2

- Low storage overhead implies
  - Lesser hardware requirements in data centers
  - Lower power consumption
  - Lower water consumption
  - Millions of dollars saved $$$

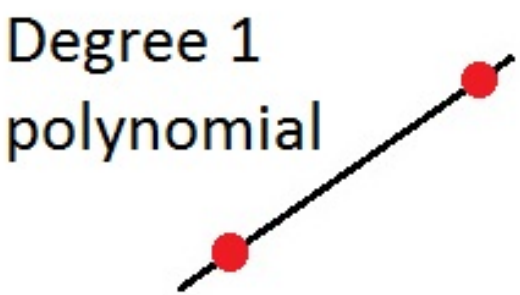
The Prototype MDS Code

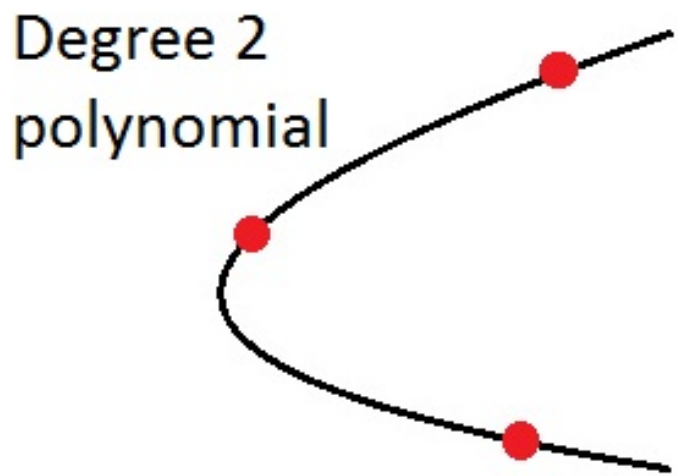
The Reed-Solomon Code

- RS codes – most widely used example of error-correcting codes – widely used as erasure codes in CD/DVDs
- well before the application to cloud storage

I. S. Reed and G. Solomon. Polynomial codes over certain finite fields. J. SIAM, 1960.

The Underlying Principle of RS Codes



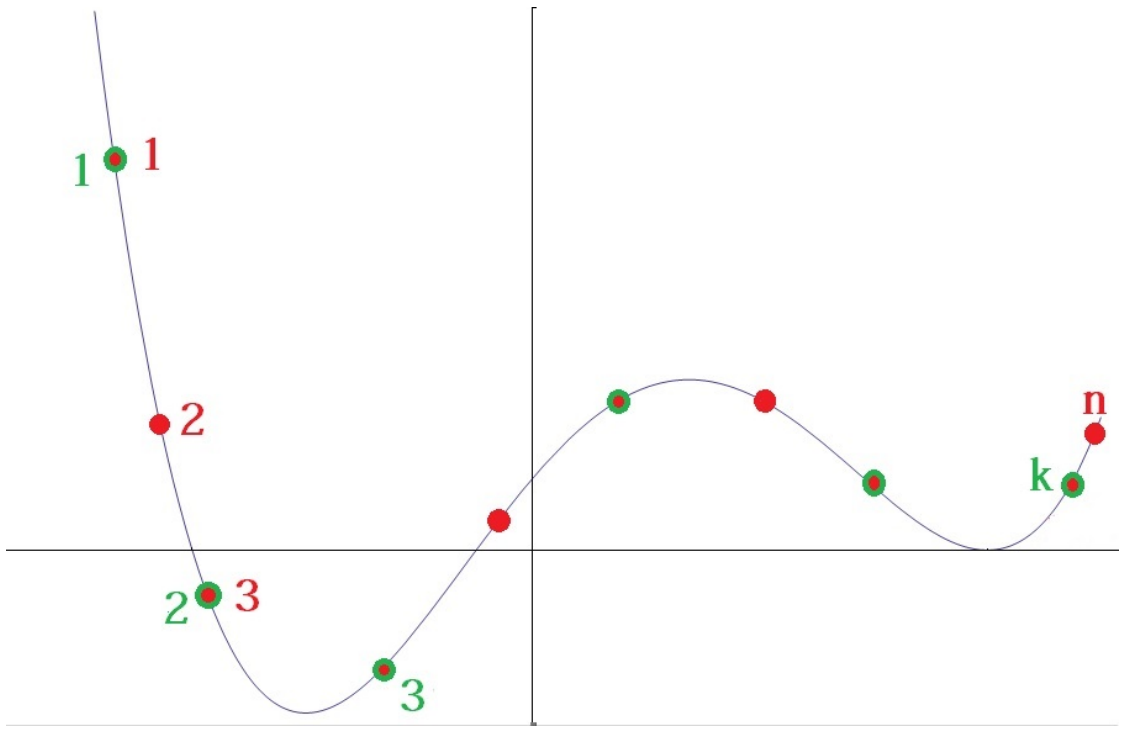


$$f(x) = f_0 + f_1 x$$

$$f(x) = f_0 + f_1 x + f_2 x^2$$

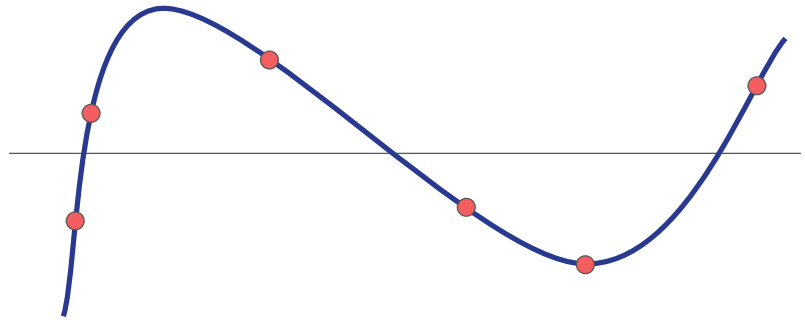
The Underlying Principle of RS Codes

- Polynomial $f(x) = f_0 + \cdots f_{k-1}x^{k-1}$ of degree k-1, evaluated at n points.
- Coefficients of the polynomial are information (message) symbols and the n evaluations are the code symbols
- The values at any k points are sufficient to describe a degree k-1 polynomial.

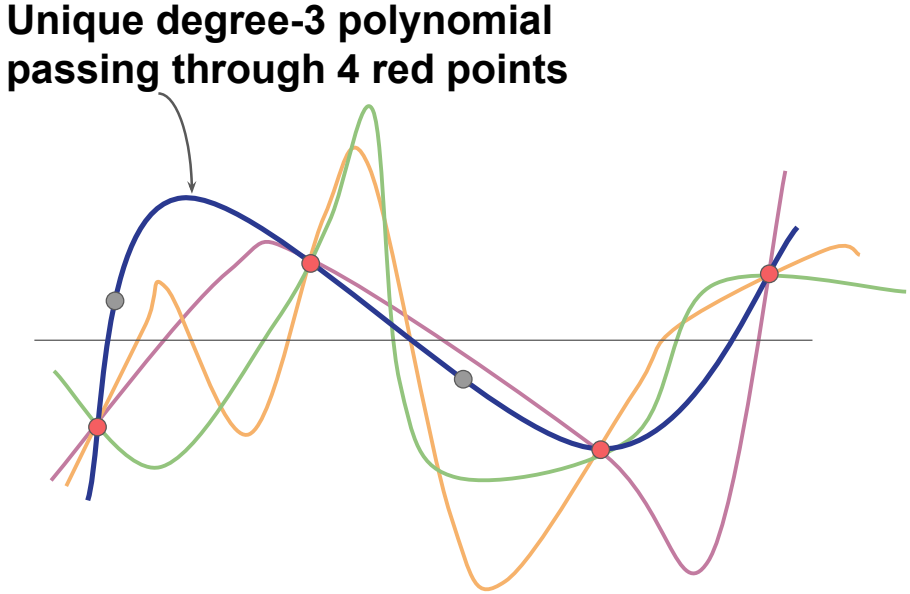


An example: $[6, 4]$ RS code

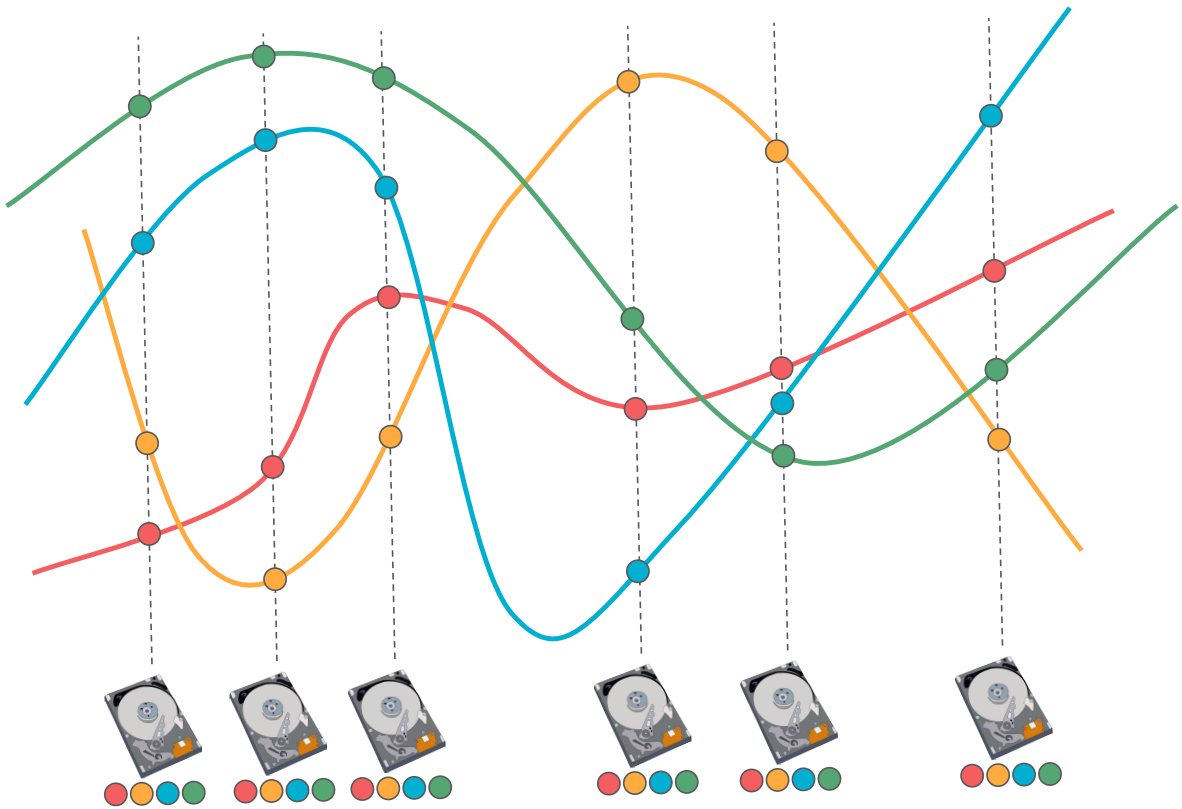
- The first figure shows a degree-3 (cubic) polynomial $f(x) = f_0 + \cdots + f_3 x^3$ evaluated at 6 points



- Using any 4 points, we can uniquely determine the polynomial and thus decode information symbols



How Exactly is Data Stored?



- Different information symbols result in different polynomials
- All these polynomials are evaluated at the same n points to obtain code symbols
- All code symbols corresponding to a particular point is stored in the same node

Node Repair

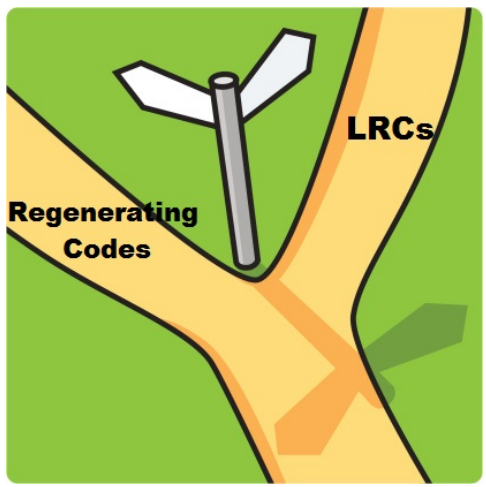
- Single node failure - the most likely failure event
- Node repair - the process of reconstructing lost node data from remaining nodes
- For RS codes repair requires downloading entire contents of k other nodes
- definitely not optimal

Node Repair: Can we do better?

As it turns out, yes, quite a bit better..

Codes for Efficient Node Repair

- Two philosophies for efficient node repair:
  - Locally Recoverable Codes (LRC) : reduce number of helper nodes accessed for repair of a failed node
  - Regenerating Codes : reduce the amount of data downloaded from each node for repair



P. Gopalan, C. Huang, H. Simitci, and S. Yekhanin, On the Locality of Codeword Symbols, IEEE Trans. Inf. Theory, vol. 58, no. 11, pp. 69256934,Nov. 2012.

A. G. Dimakis, P. B. Godfrey, Y. Wu, M. Wainwright, and K. Ramchandran, Network Coding for Distributed Storage Systems, IEEE Trans. Inform. Theory ,Sep.2010.

Windows Azure LRC

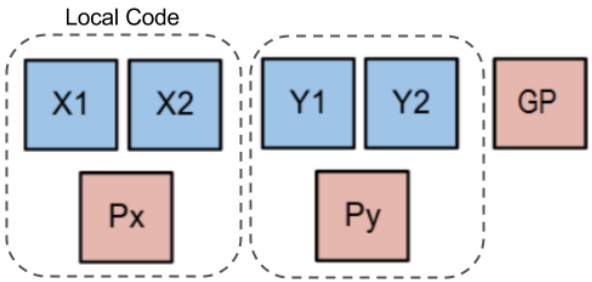


Figure: $(7, 4, 2)$ LRC

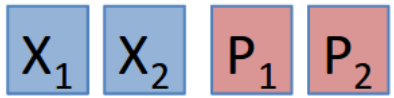


Figure: $(4, 2)$ RS

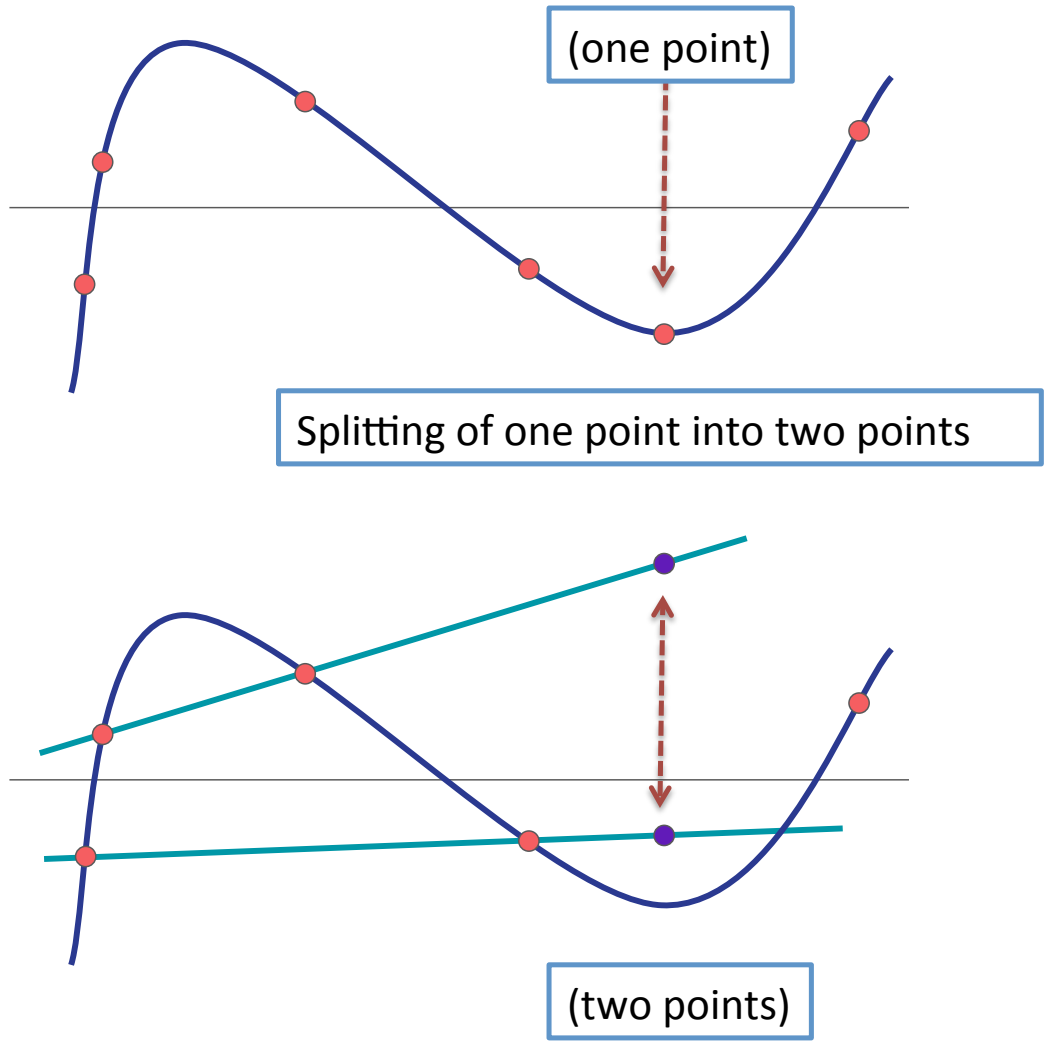
- Windows Azure Storage system uses LRCs
- We show an example LRC where:
  - Number of helper nodes contacted for node repair are the same for both the codes
  - They also provide the same reliability
  - Overheads however are quite different, 1.75 for the LRC versus 2 for the RS code

- Usage of LRCs has saved Microsoft millions of dollars $$$

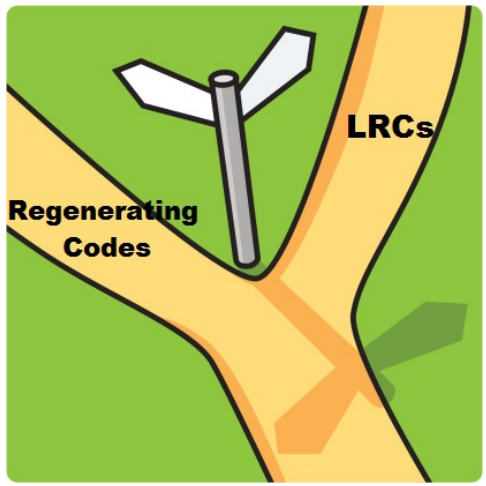
Huang, Simitci, Xu, Ogus, Calder, Gopalan, Li, Yekhanin, Erasure Coding in Windows Azure Storage, USENIX, Boston, MA, 2012.

Polynomial Viewpoint of an LRC

- The first image shows a degree 3 polynomial evaluated at 6 points
- In the second image depicts $(7, 4, 2)$ LRC
- The two straight lines represent local codes



Back to the Fork in the Road



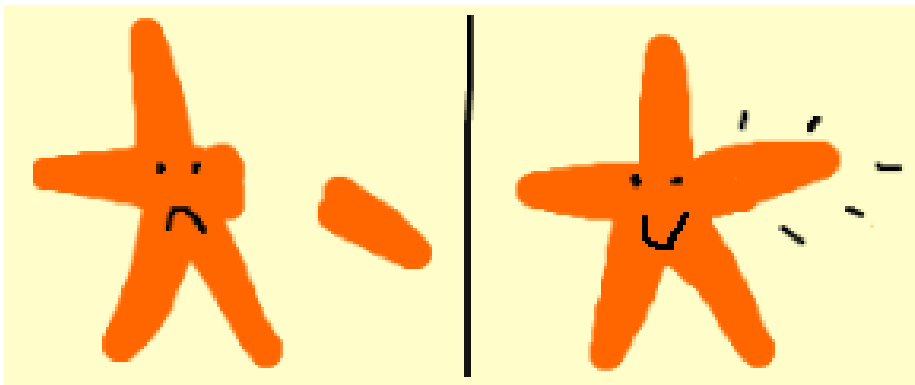
Regenerating Codes

- In a regenerating code, a file is encoded and stored across n nodes and it can be reconstructed by contacting any k nodes. Every node stores α symbols.

- During a node failure, repair is done by contacting $d(< n)$ helper nodes where each node sends β symbols

- Repair bandwidth ($d\beta$) is the amount of data downloaded during repair

- We wish to minimize both storage overhead and repair bandwidth



- Minimum Storage Regenerating (MSR) Codes
  - Require minimum storage overhead (They are MDS codes)
  - For this amount of storage they are optimal w.r.t repair bandwidth

A Bright Idea ?

- We have seen that replicating symbols helps
- why not replicate RS codes ?

A Bright Idea ?

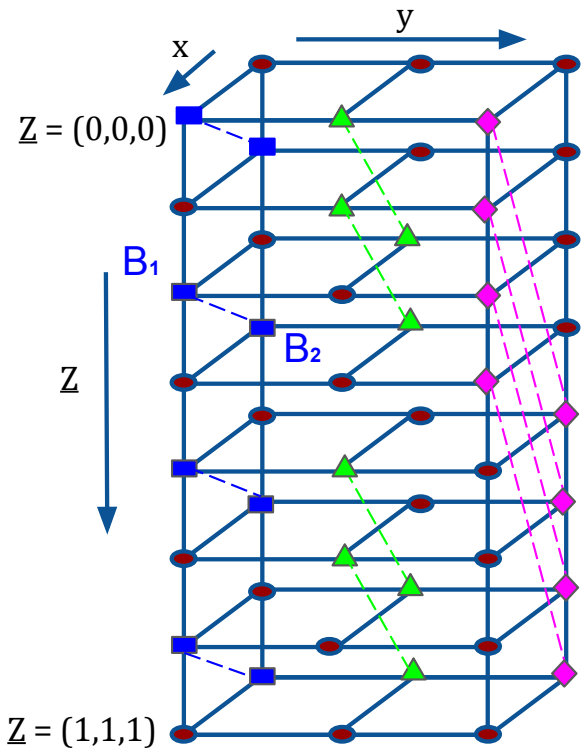
- We have seen that replicating symbols helps
- why not replicate RS codes ?

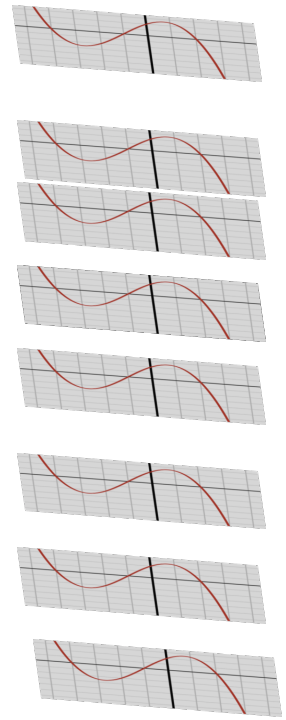
(turns out to be a really dumb idea)

Wait a Minute!

- Not a dumb idea...

Coupled Layer Regenerating Code





(each layer is a Reed-Solomon code
but the different layers are coupled!)

---

1. M.Ye, A. Barg, Explicit constructions of optimal-access MDS codes with nearly optimal sub-packetization, arXiv:1607.07335v3 [cs.IT] May 27, 2016.

2. B Sasidharan, M. Vajha, PVK, "An explicit, coupled-layer construction of a high-rate MSR code with low sub-packetization level, small field size and all-node repair. arXiv:1607.07335v1 [cs.IT], July 25, 2016.
3. B Sasidharan, M. Vajha, PVK, "An explicit, coupled-layer construction of a high-rate MSR code with low sub-packetization level, small field size and all-node repair with ($d < (n-1)$). arXiv:1607.07335v1 [cs.IT], July 25, 2016.

Ceph Implementation of CL-MSR code

- Implemented the CL-MSR code in Ceph, an open source distributed storage system.
- Evaluated using Amazon Web Services cluster (the cloud!)
- Significantly fast repair in comparison with RS codes
- Repair bandwidth savings upto 75% in comparison with RS code was observed

- Myna Vajha, **Ganesh Kini, Bhagyashree Puranik, Vinayak Ramkumar**, Elita Lobo, Birenjith Sasidharan, PVK, Min Ye, Alexander Barg, Syed Hussain, Srinivasan Narayanamurthy, Siddhartha Nandi "Pairing up for Regeneration: The Mantra for Fast and Efficient Node Repair in Distributed Storage," submitted to 2017 USENIX Annual Technical Conference.

Erasure Codes in Practice

Google File System: 3-replication
GFS II (Colossus) : $[9, 6]$ RS Code

Facebook HDFS: $[14, 10]$ RS Code

Yahoo Object Store: $[11, 8]$ RS Code

Microsoft Windows Azure: $(16, 12, 6)$ LRC

Coding for Distributed Storage: New Branches Have Sprouted!

- Regenerating Codes
- Locally Recoverable Codes

Other References

- Cloud cartoons : cloudtweaks.com/humor/. pinterest.com/monitorus/cloud-humor/,pinterest.com/compassitesinc/cloud-enablement/, igvita.com/posts/
- Data Center Images : chronushosting.net/images/slider/,techcentricnews.files.wordpress.com/,technewstoday.com/2 microsoft-dives-underwater-to-build-a-cool-data-center/,extremetech.com/
- Other images : patreon.com/, pre12.deviantart.net/, moonshinenews.com/tag/disaster-recovery/, winterbluemusic.com/superman-drawing/, prinklerdon.com/, clipartfest.com/, disney.wikia.com/, clipartvision.com/,cdn.drawception.com/
- pritishankarhomage.wordpress.com/photo-gallery/
- http://vignette4.wikia.nocookie.net/poohadventures/images/8/8c/Genie.png/

Thanks!