SDC
STORAGE DEVELOPER CONFERENCE
SNIA SANTA CLARA, 2016

Modern Erasure Codes for Distributed Storage Systems

Srinivasan Narayanamurthy (Srini)
NetApp

Everything around us is changing!

- The Data Deluge
  - Disk capacities and densities are increasing faster than the disk transfer rates
  - Increased delay to recover using classical techniques lead to availability exposure
- Changing Storage Technologies
  - Architectures: Scale-out, Distributed Storage, Cloud, Converged
  - Media: Flash, NVM, SMR, Tape, et al.
  - Features: Geo-distribution, Security, Commodity hardware (Failure is a norm!)
- Newer Dimensions of Erasure Codes
  - Optimality tradeoffs redefined
  - More about this inside…

Organization

- Background
  - Erasure Codes Timeline
  - Classical Codes - (n, k) code
- Modern Codes
  - Locally Repairable Codes (Codes on Codes)
  - Regenerating Codes (Network Codes)
- Technical Analysis
  - Optimality Tradeoff and Reliability Analysis
  - System Requirements and Codes
- Literature & Key Players

Background

Timeline – Classical (n, k) codes

Timeline – Overview

Classical Codes

Fountain Codes

Locally Repairable Codes

Regenerating Codes

Tradeoff against "*Storage Overhead*"

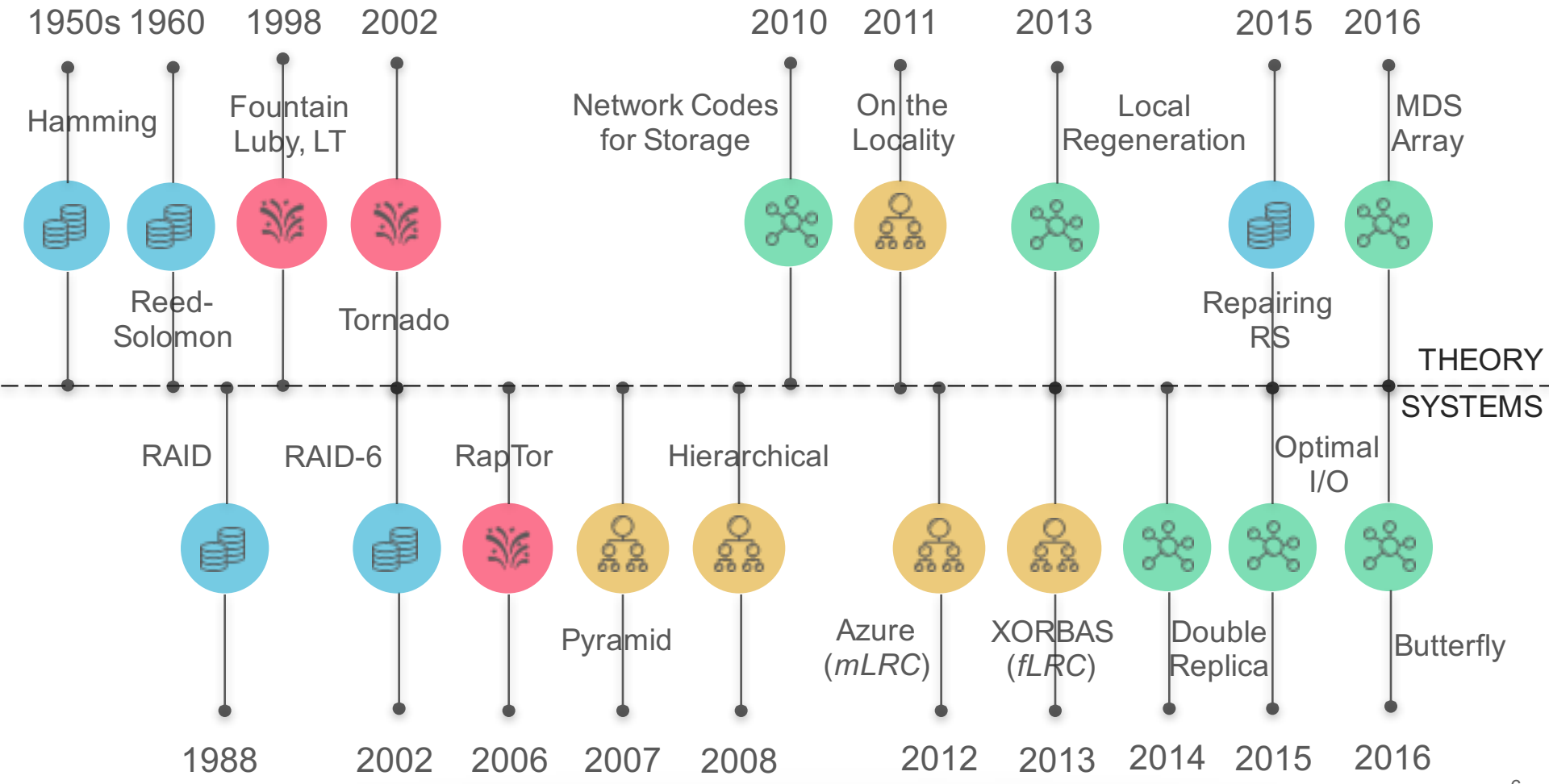
Reliability

Performance

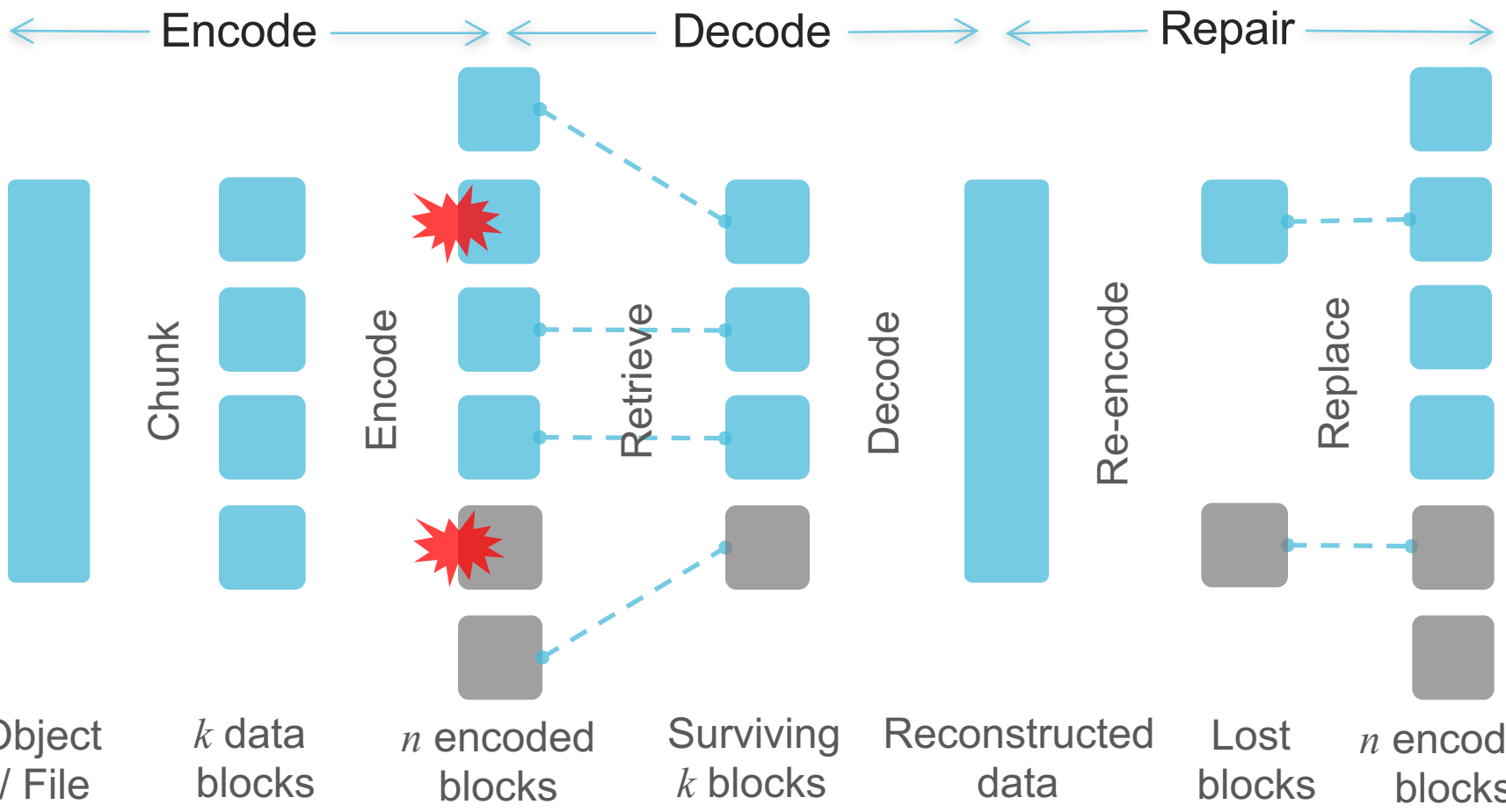
Repair Degree

Repair Bandwidth

Timeline



Classical (n, k) Codes (6,4)



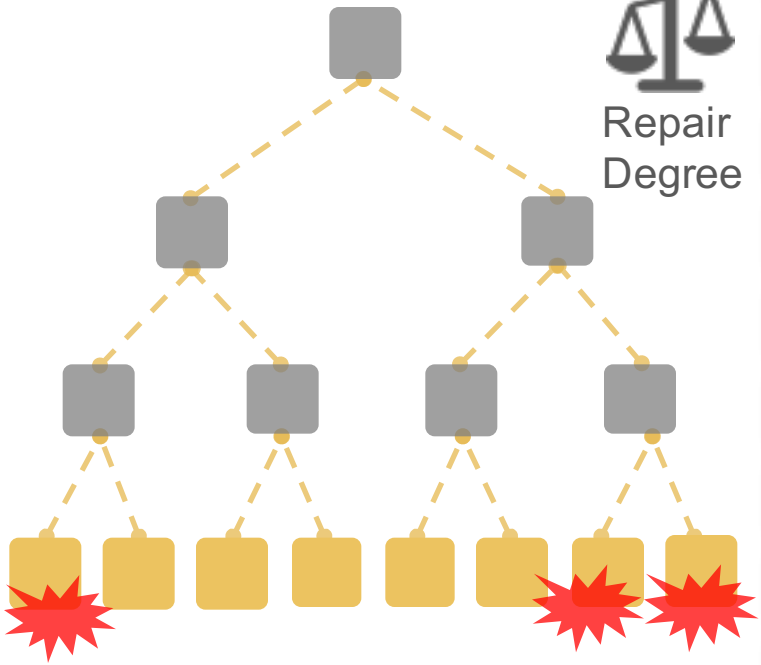
Think distributed systems; repairs are expensive !

Modern Erasure Codes

Locally Repairable Codes – Regenerating codes

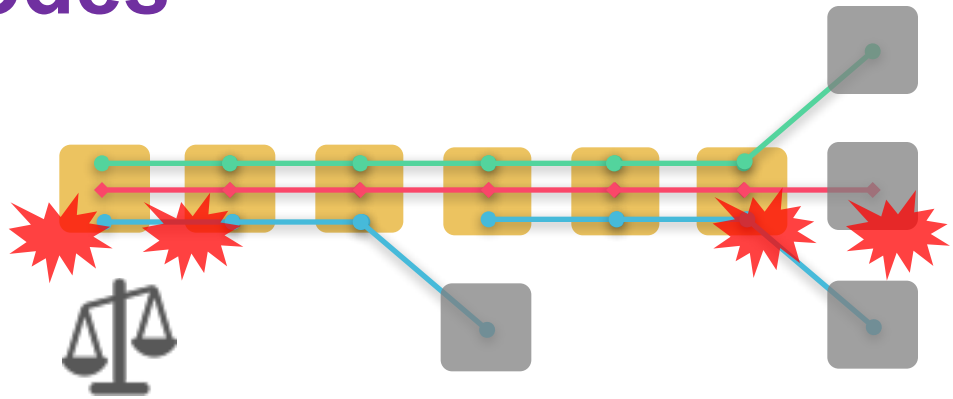
Locally Repairable codes

Hierarchical (Bottom-Up) & Pyramid (Top-Down) Codes



Repair Degree

Microsoft Azure (mLRC)

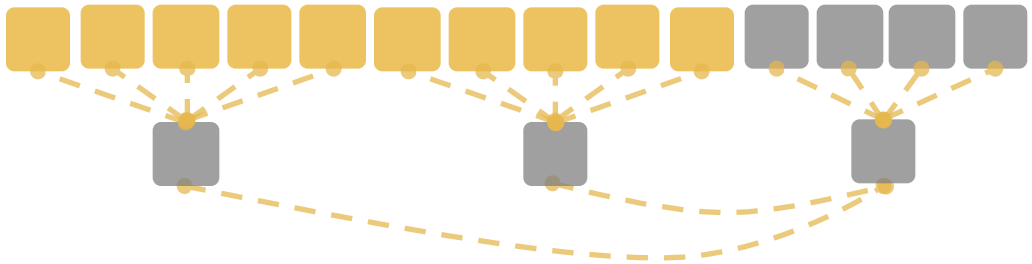


Locality/ Max. Recoverability

(k, l, r) codes : (6, 2, 2)

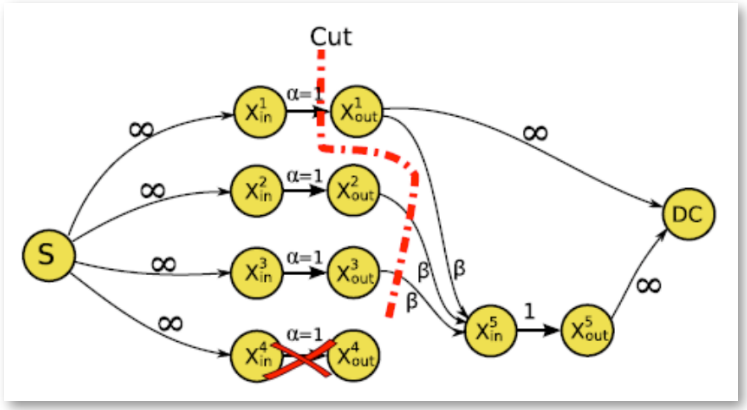
Facebook XORBAS (fLRC)

Locality/Min. Dist.



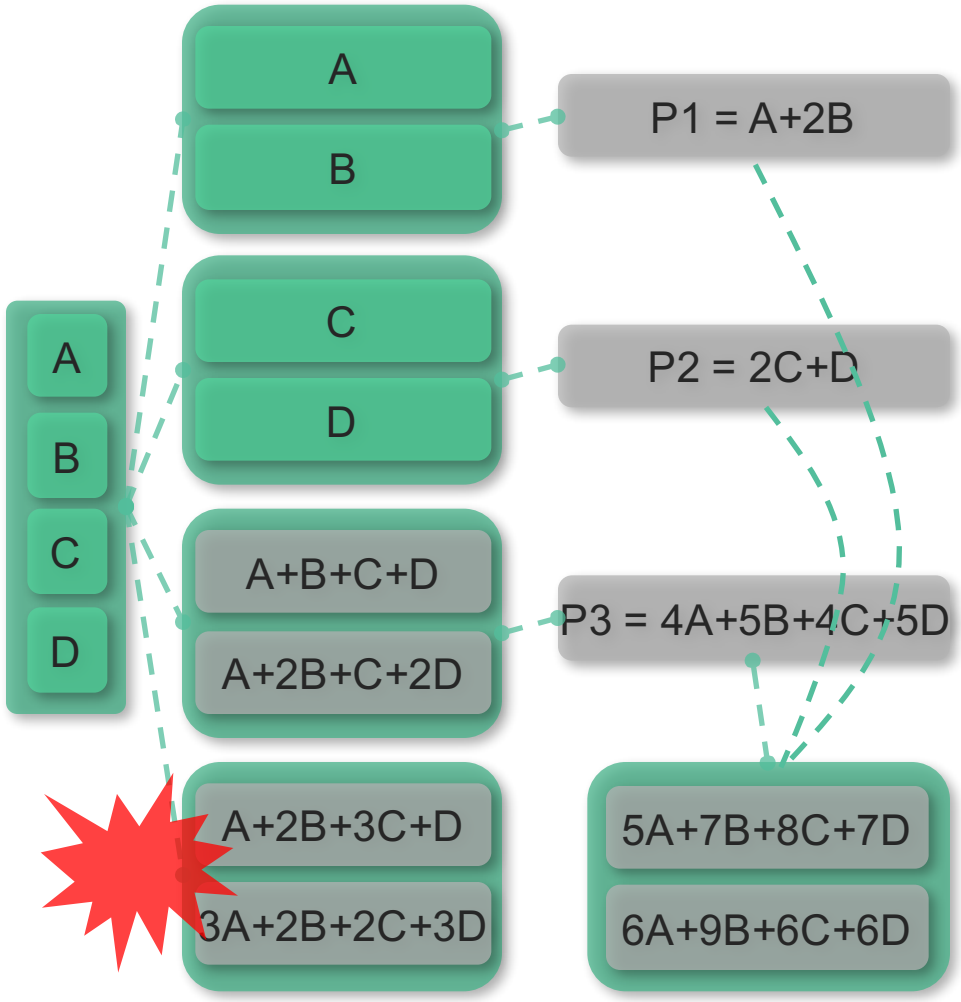
RS (14,10) & optimizing single block failures

Regenerating Codes

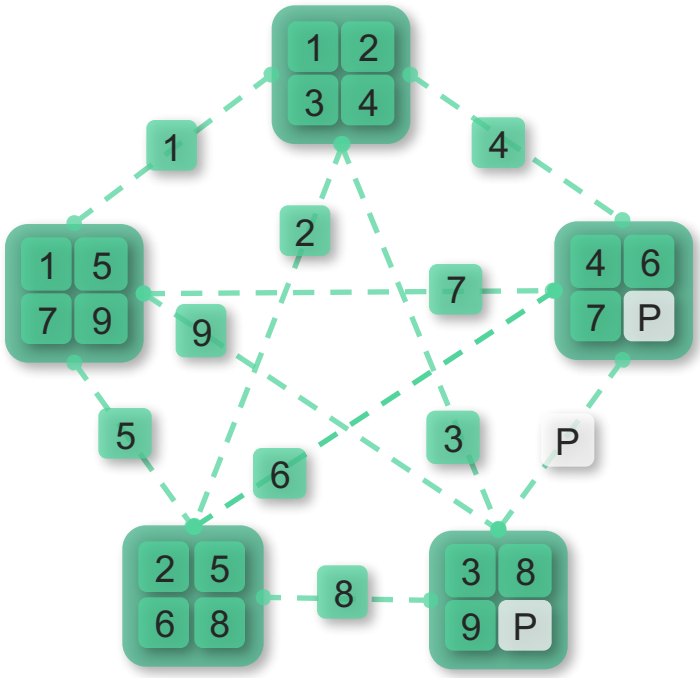


An Information Flow Graph & Min-Cut Bound

Functional Repair

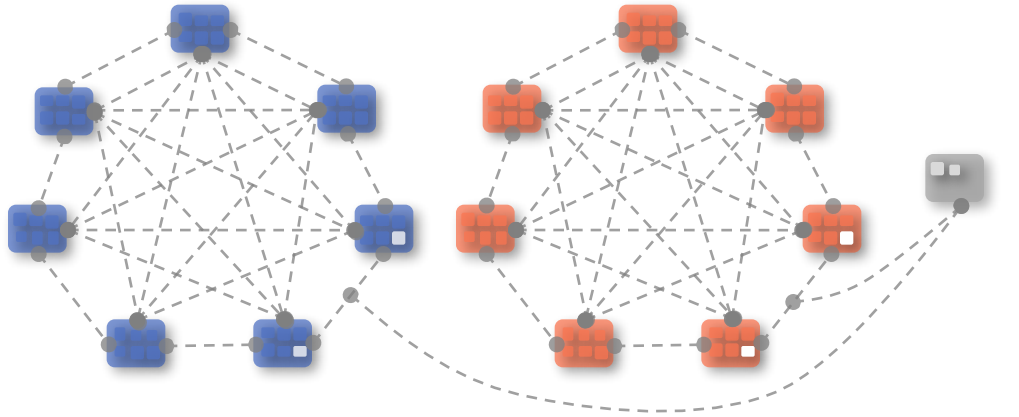


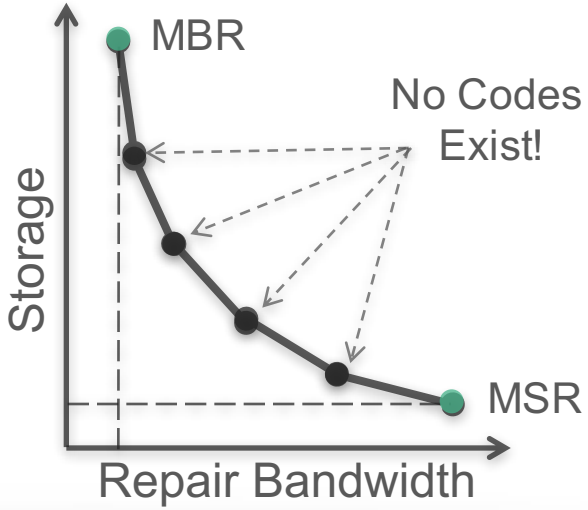
Repair By Transfer (MBR) Codes



Pentagon Code

Local Regenerating Codes





Storage / Repair BW

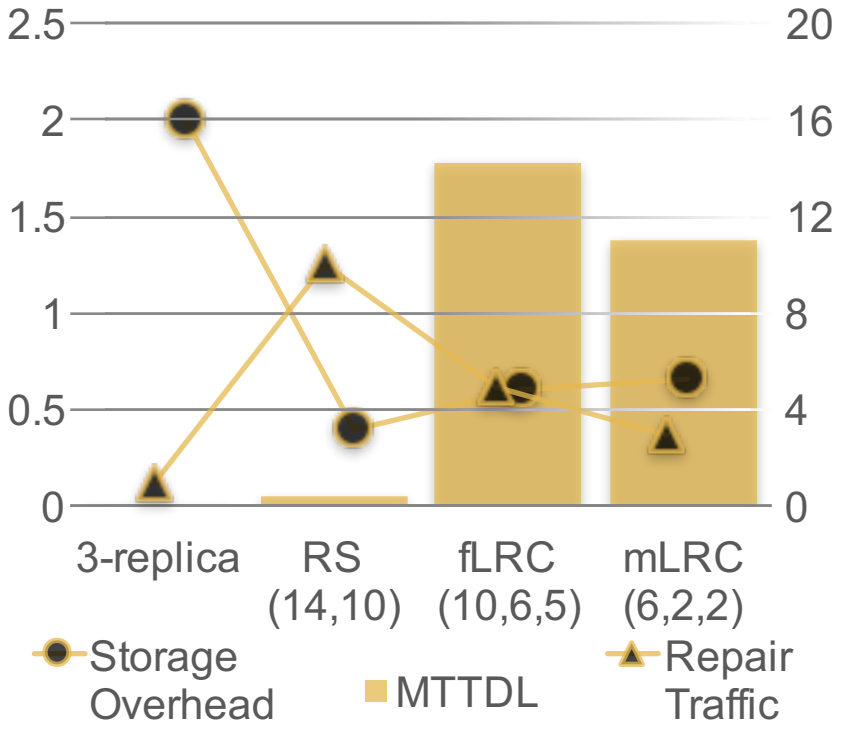
Technical Analysis

Optimality Tradeoffs – Reliability Analysis – System Requirements

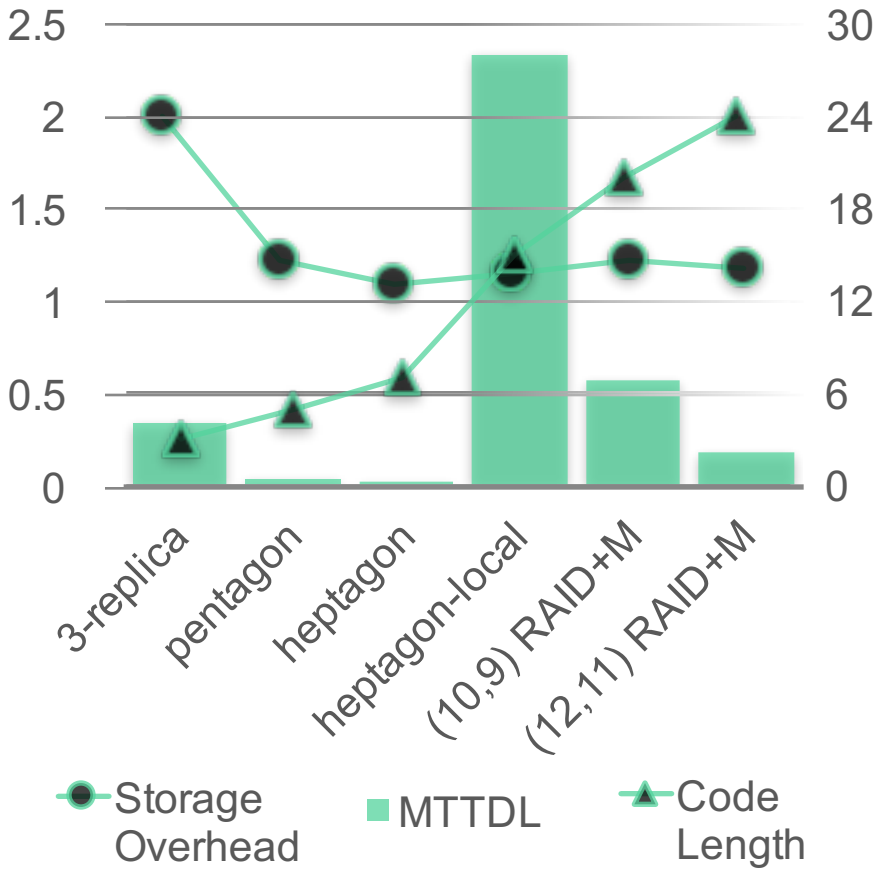
Summary of Codes and their Tradeoffs

Codes/Family	Tradeoff	
MDS	Storage overhead	Reliability
Replication & Parity	Storage overhead	Reliability
Reed-Solomon	Storage overhead	Reliability
Near-Optimal	Correction capability	Computational Complexity
Fountain	Rate	Probability of Correction
Codes on Codes	Storage overhead	Repair Degree (Fan-in)
Azure (mLRC)	MDS	Maximum Recoverability
XORBAS (fLRC)	Locality	Minimum Distance
Regenerating	Storage overhead	Repair Bandwidth
Local Regenerating	Storage overhead	Reconstruction Cost

Reliability Analysis



Locally Repairable Codes



Regenerating Codes

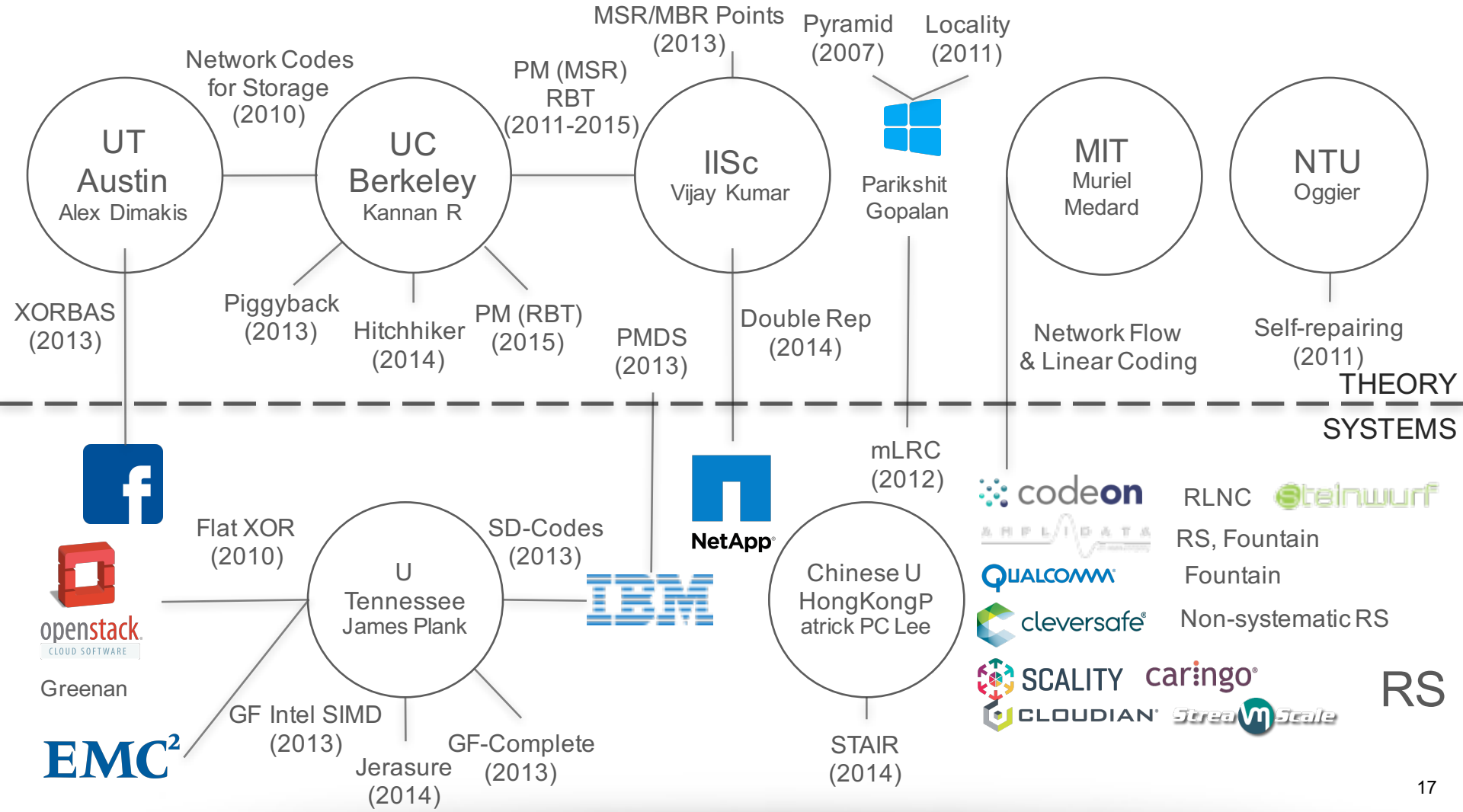
System Requirements and Example Codes

	System	Properties of the System		Requirements for a Code		Example family/code
		Most Important	Least Important	Most Important	Least Important	
Architecture	General-purpose storage array	Reliability & Performance	Cost	Reliability	Complexity	MSR, SD/STAIR Codes
	Geo-distributed storage	Repair over WAN is expensive	Storage overhead across DR sites	Local repair	Storage overhead	LRC
	Secure Storage	Security	Storage overhead	Faster degraded reads	Repair time	Non-systematic codes; MBR
	Distributed Systems	Parallelism & Availability	Storage overhead	Systematic	Storage overhead	Replication
Workload	Big Data (say, Hadoop)	Large volumes of data	Write latency	Storage overhead	Repair bandwidth	Regenerating (MSR/MBR), systematic

Literature & Key Players

Theory & Systems

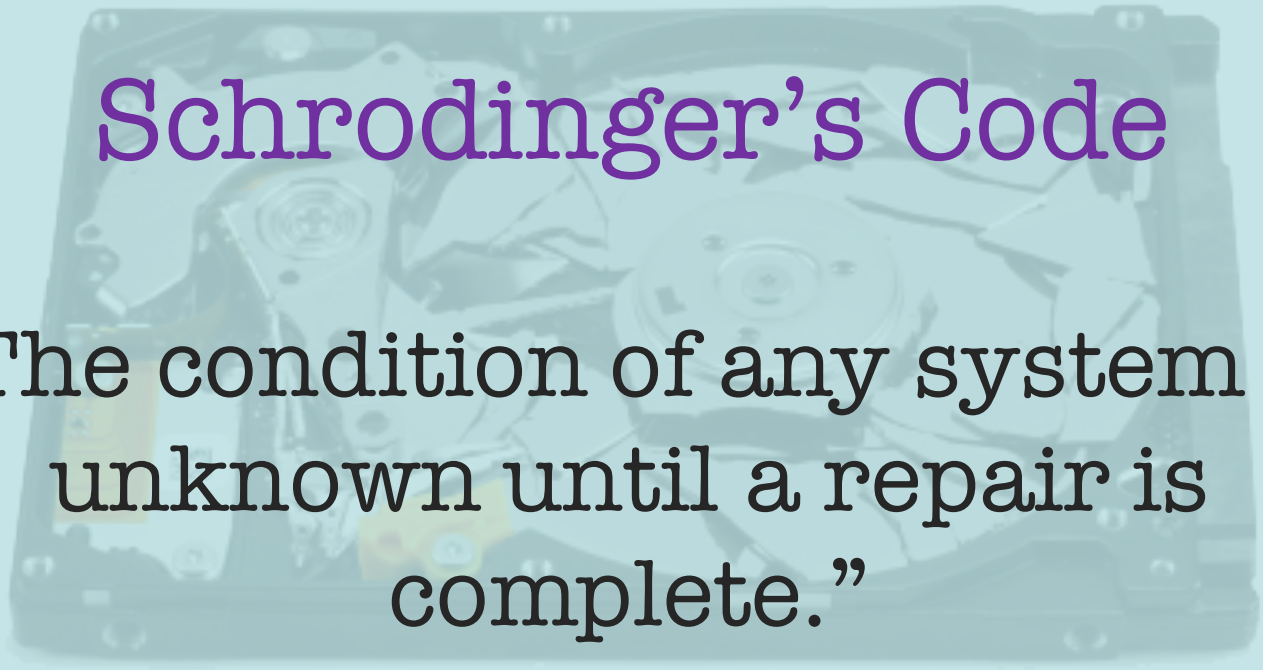
Researchers, Big Players & Startups



Related Areas

- Cross-object Coding
  - Sector & Disk failures – PMDS, SD, STAIR Codes
- Other media
  - Flash: LDPC, WOM, Multi-write codes; NVM
- Security
  - Dispersal, AONT-RS
- Cloud
  - NC-Cloud
- Transformational Codes:
  - Transform encoded data to different parameters as they become hot/cold without decoding and re-encoding

Schrodinger's Code



"The condition of any system is unknown until a repair is complete."