Adjustable flat layouts for Two-Failure Tolerant Storage Systems

Thomas Schwarz, SJ
Marquette University

Motivation

- Storage device batches fail at different rates
  - Example: Backblaze:
    - 1163 Seagate Barracuda 7200.14 disks
      - failed at a rate of 43% per year in 2014,
- Storage devices (sometimes) fail at different rates
  - Bathtub curve seen in about 50% of all HD at Netapp
  - SSD unrecoverable read error rate increases at the end of their lifetime

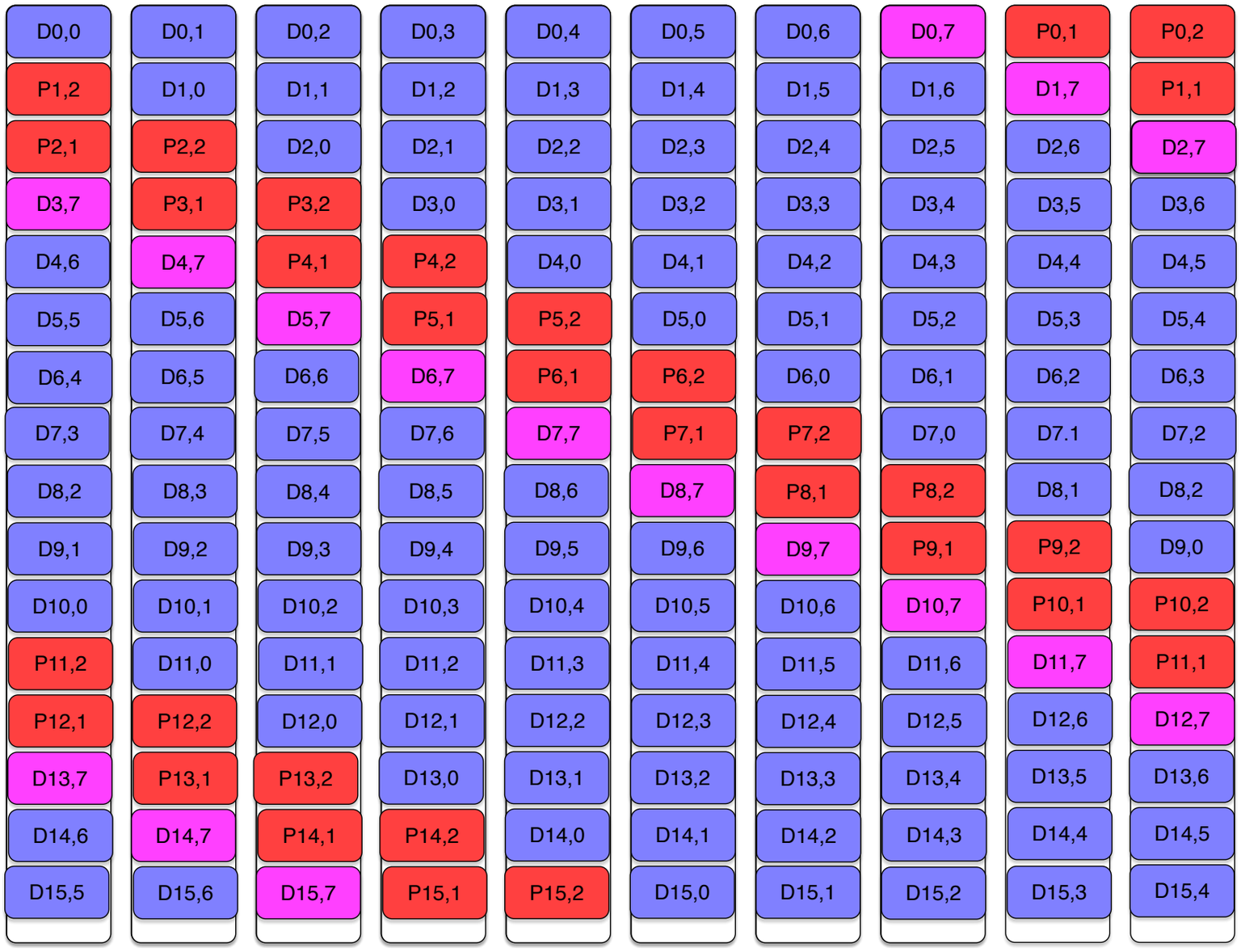
Motivation

- Large storage systems
  - Currently consists of disks or SSDs organized in racks
  - Individual devices are replaced
  - Erasure coding for files, not devices
- My proposal
  - Organize a large number of devices in a storage pod
  - **Level of failure tolerance** in pod **varies** according to prediction of device vulnerability
  - Use a flat layout to increase failure tolerance

Adjustable Raid 6 Example

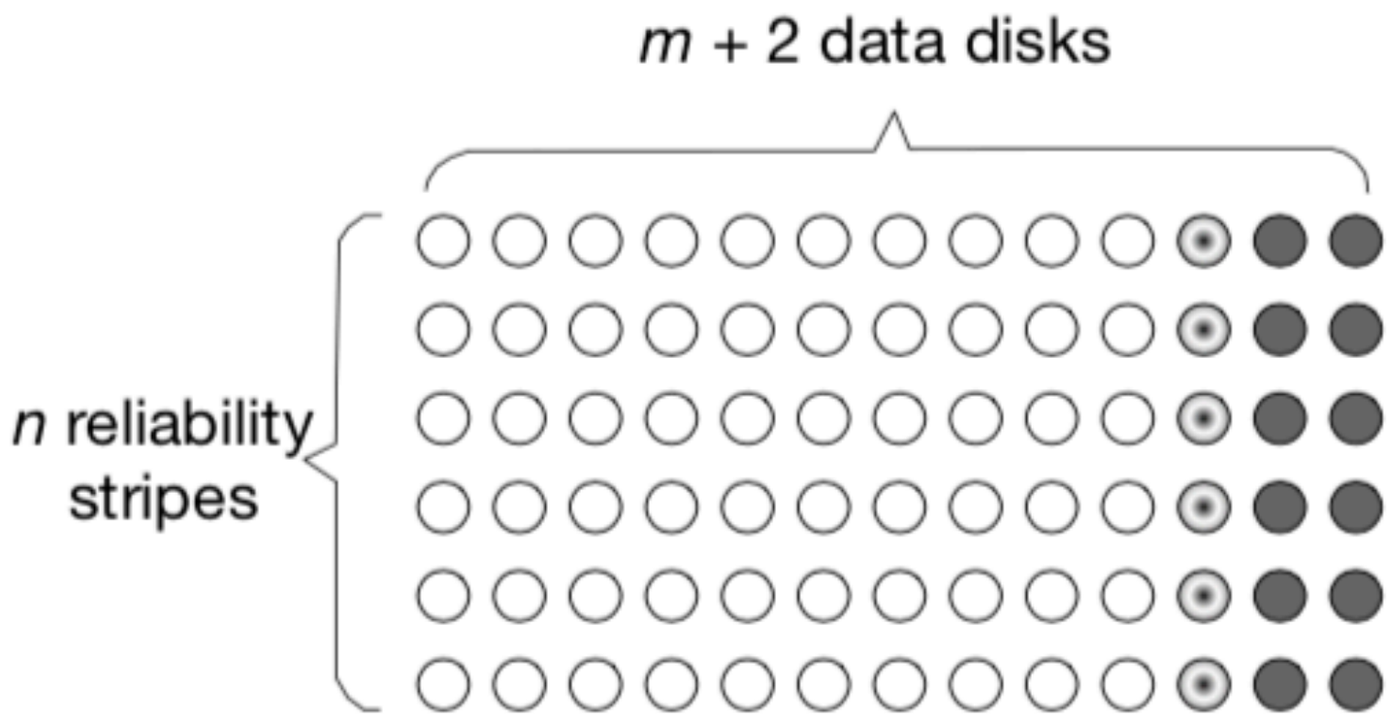
- Group *k* devices into a reliability stripe
  - User data devices
- Add two parity devices to each reliability stripe
- If device failure rate appears to be high:
  - Rededicate a user data device as a parity
- Overall:
  - Trade capacity for additional failure tolerance when needed

Adjustable RAID 6 Example



Adjustable RAID 6

Adjustable RAID 6 Example



Alternative to RAID Stripes

- Use a flat layout:
  - Each user data device is in two or three reliability stripes with one additional parity
- Does not use Galois field arithmetic
- Reconstruction can be done using two or three alternatives
  - Can avoid a single hot spot

Results

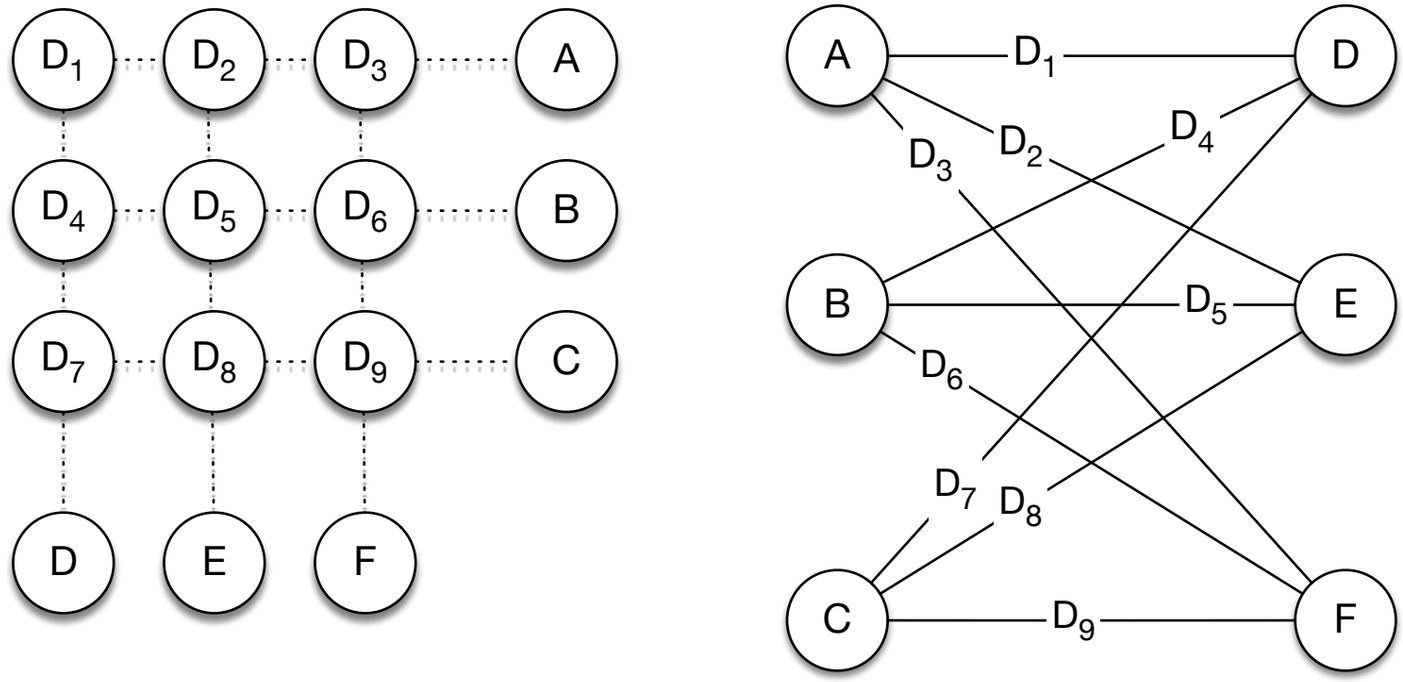
- Adjustable RAID 6
  - Easy to find configurations
- Adjustable flat layouts
- Higher reliability
- No need for Galois field arithmetic
  - Accelerators need extended instruction set
- Flexibility in reconstruction of lost data

Layout Definition

- Flat layouts:
  - Each user data device is part of two reliability stripes
  - Two reliability stripes have one or none data device in common
    - Each reliability stripe contains *k* user data devices
- Therefore:
  - Each data device corresponds to an edge of an undirected graph
  - Each parity device corresponds to a reliability stripe that corresponds to a vertex

Layout Definition

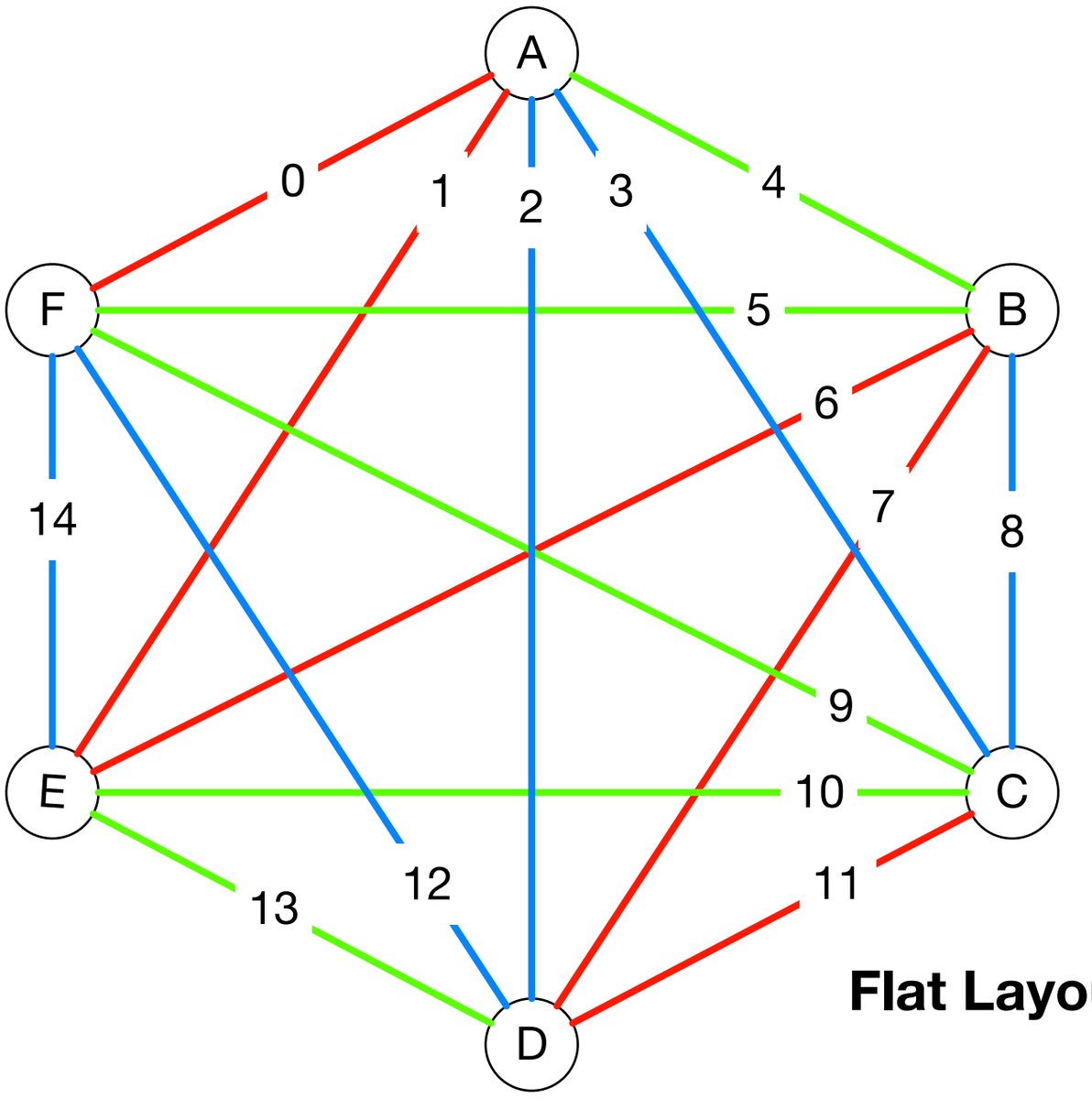
- Use graph view:
  - Edges are user data devices
  - Vertices are parity data devices



Layout and corresponding graph

Layout Definition

- Densest layouts correspond to a complete graph

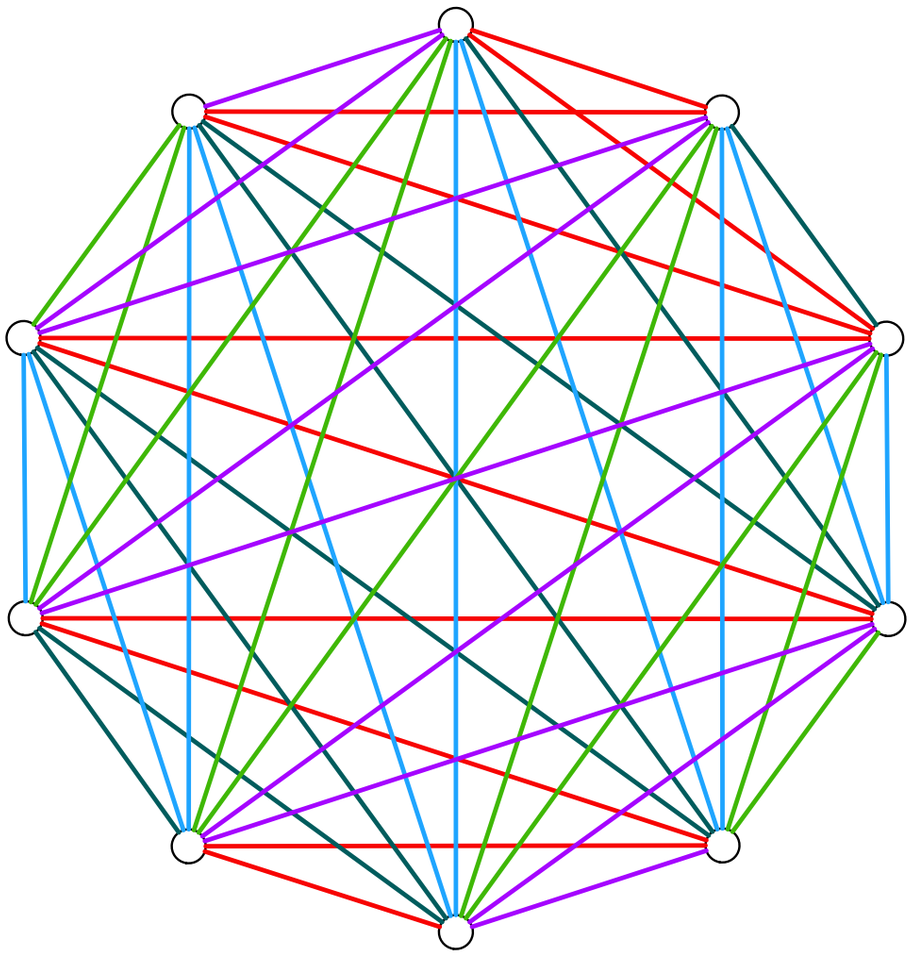


A: 0, 1, 2, 3, 4
B: 4, 5, 6, 7, 8
C: 8, 3, **9**, 10, 11
D: 11, 7, 12, 13
E: 13, 10, 6, 1, 14
F: 14, 12, **9**, 5, 0

Flat Layout with 6 stripes and 14 user data devices

Layout Definition

- If we want to create additional reliability stripes, we can use a graph factorization



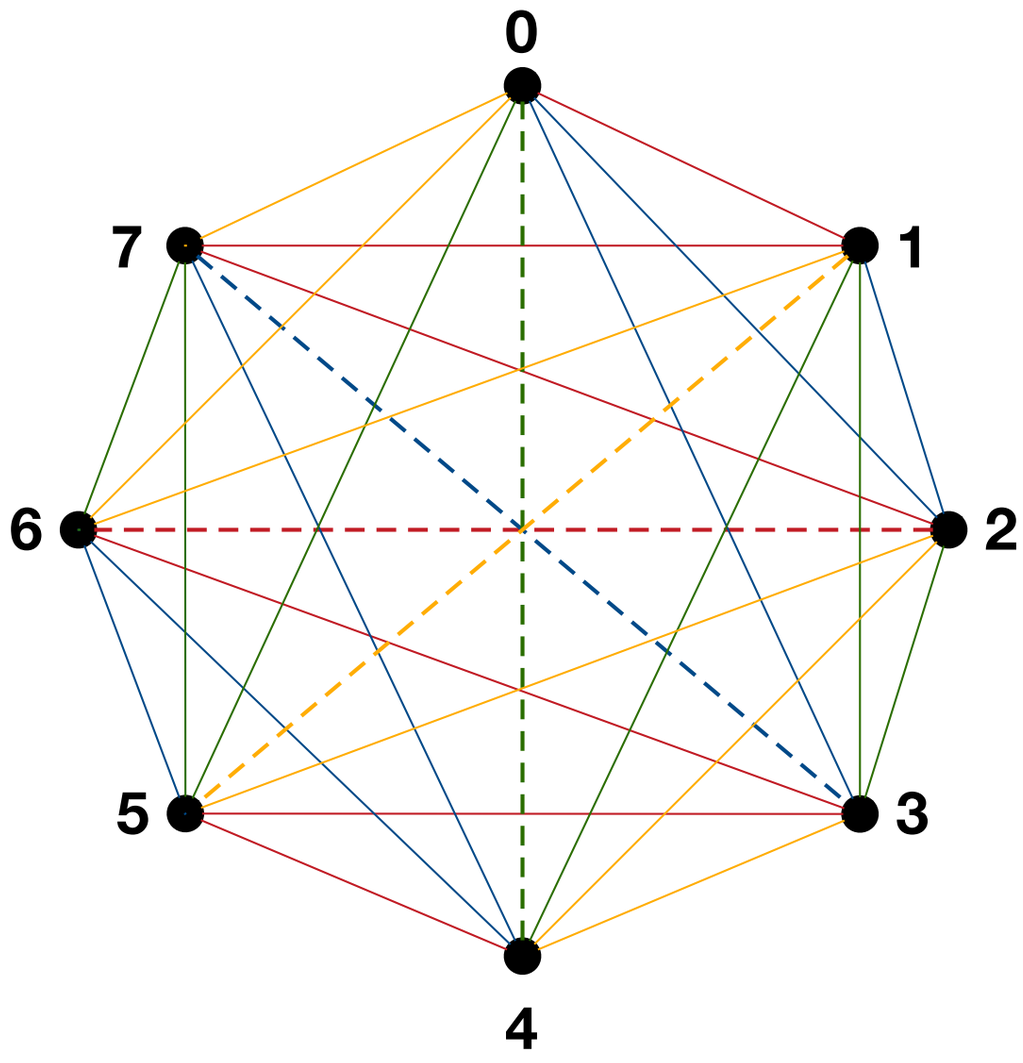
- Each user data device is in three reliability stripes
- Any two stripes intersect in one or none user data devices
- This factorization invented by Lawless 1974

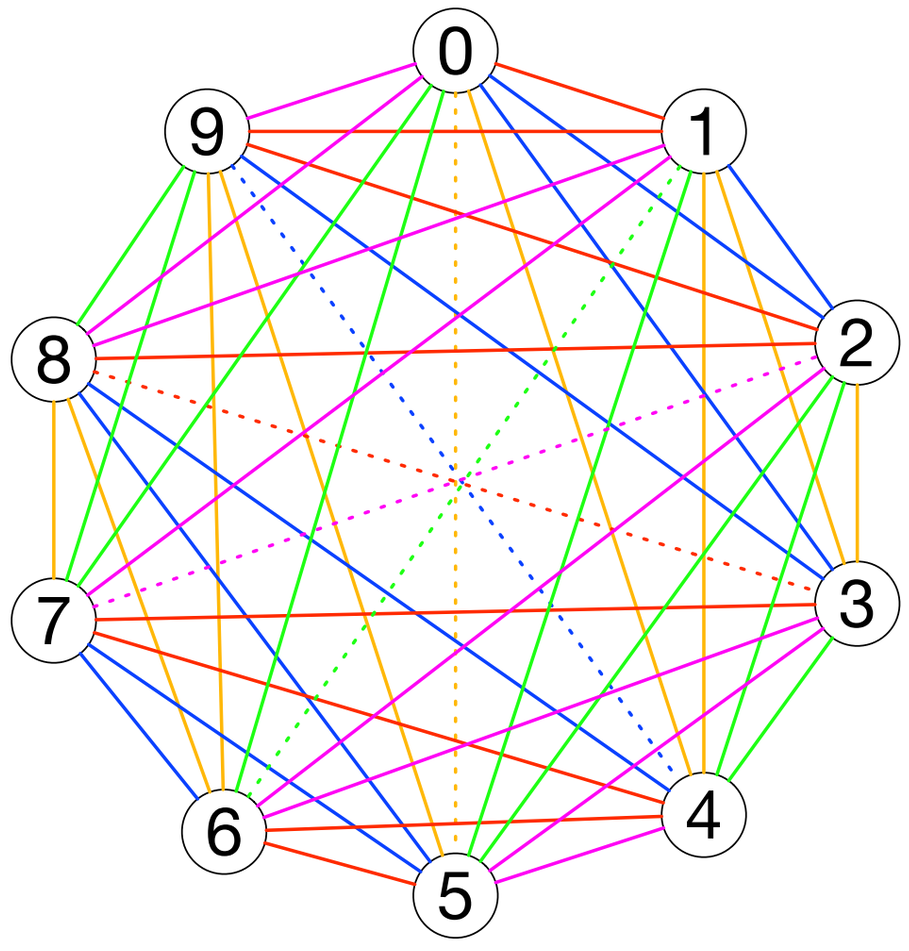
Layout Definition

- Can add additional parity devices to an ensemble in case of need
- How about switching some user data devices to parity?
  - Cannot be done instantaneously because those data devices need to emptied
  - But it can be done

Layout Definition

- Punctured Layouts: Remove the middle edge from each factor





Layout Definition

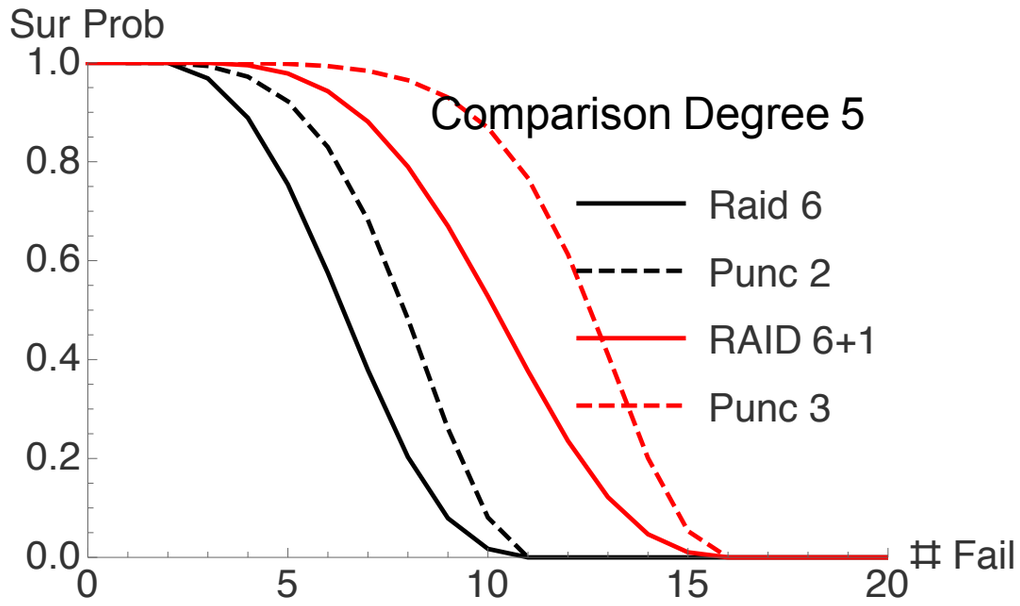
- Available only for certain parity - data device numbers

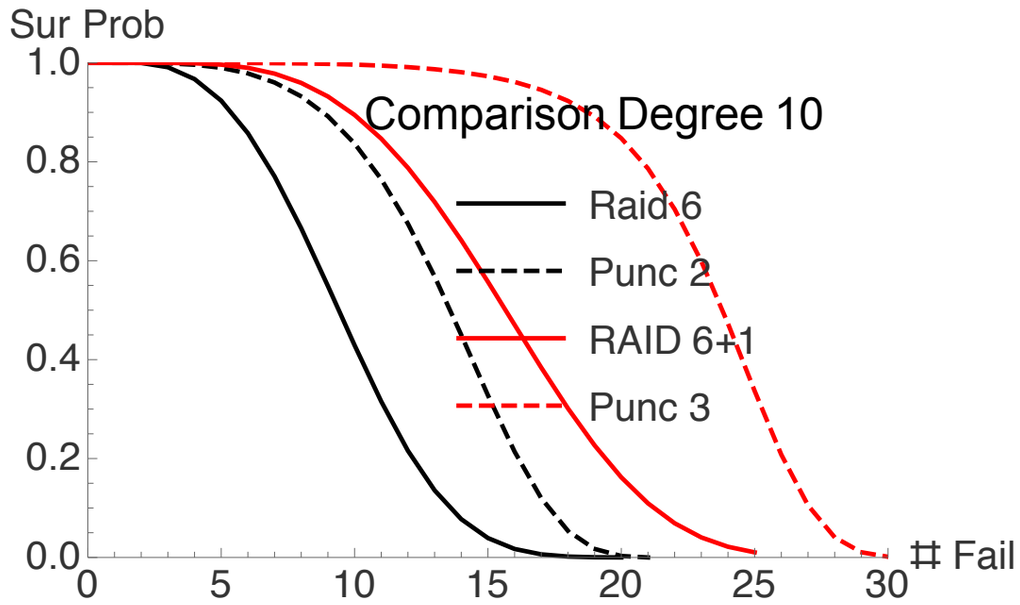
DIMENSIONS OF PUNCTURED LAYOUTS. ON THE LEFT, WE GIVE THE NUMBERS FOR THE TWO-FAILURE TOLERANT AND ON THE RIGHT FOR THE THREE-FAILURE TOLERANT LAYOUT.

d	# Data	# Parity	# Total Disks	Stripe Sizes
3	15/12	6/9	21	5/4
4	28/24	8/12	36	7/6
5	45/40	10/15	55	9/8
6	66/60	12/18	78	11/10
7	91/84	14/21	105	13/12
8	120/112	16/24	136	15/14
9	153/144	18/27	171	17/16
10	190/180	20/30	210	19/18
11	231/220	22/33	253	21/20

Reliability Evaluation

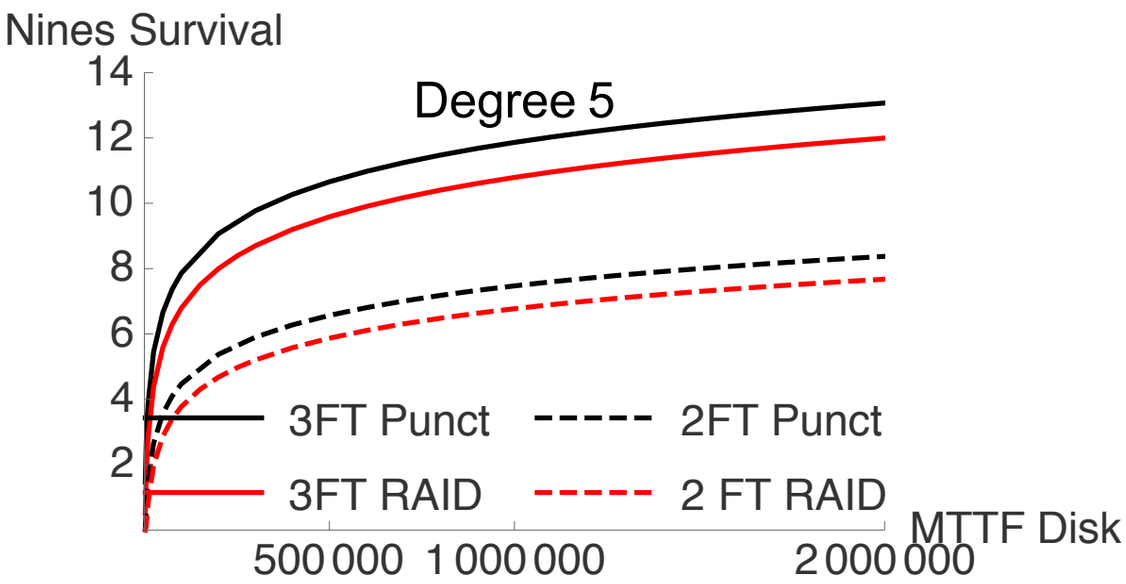
- We compare with an adjustable RAID Level 6 configuration
- Robustness: Probability that *f* device failures have let to data loss

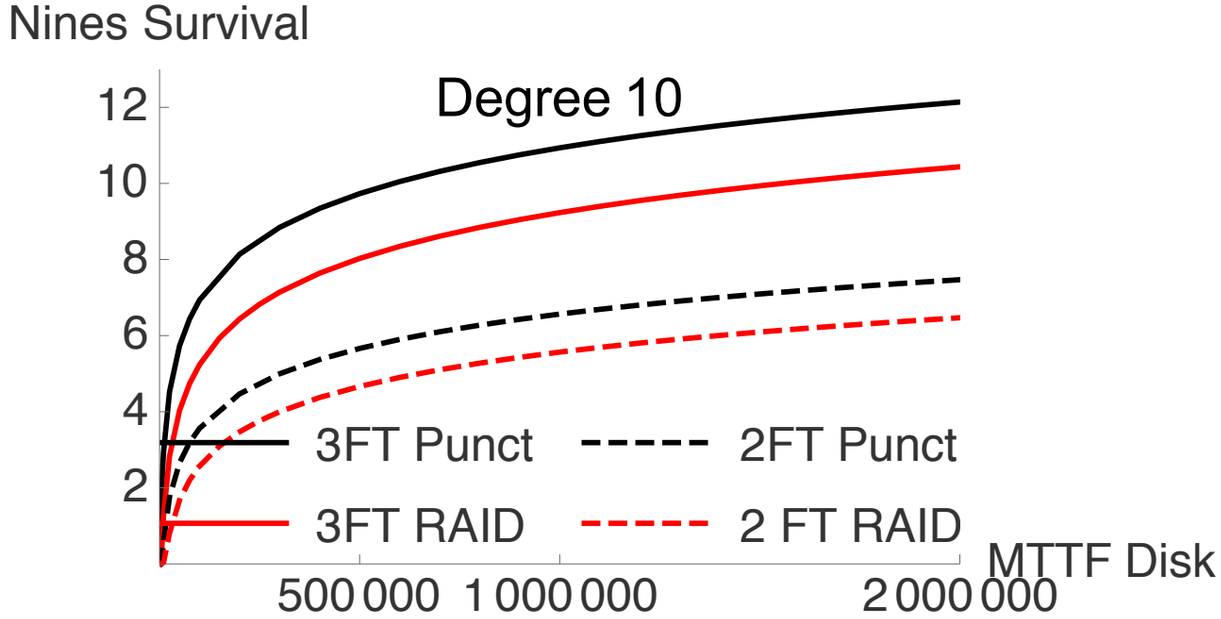




Reliability Evaluation

- Calculation of five and six year survival probabilities:





Results

- Adjustable RAID 6
  - Easy to find configurations
- Adjustable flat layouts
- Higher reliability
- No need for Galois field arithmetic
  - Accelerators need extended instruction set
- Flexibility in reconstruction of lost data