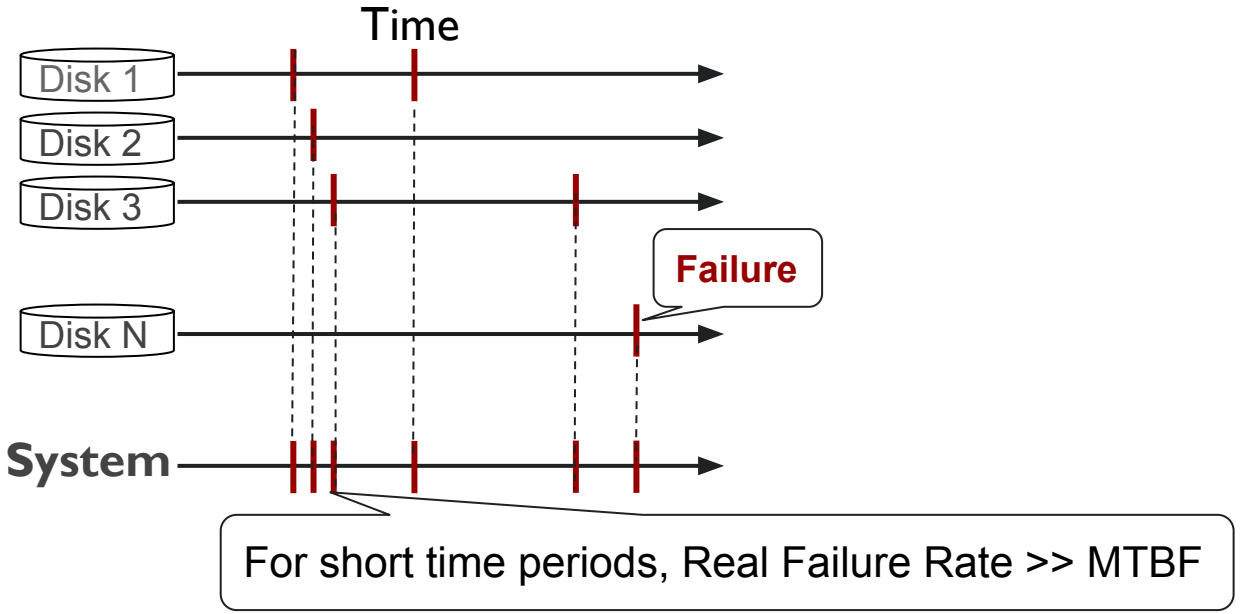
Fractional-Overlap Declustered Parity: Evaluating Reliability for Storage Systems

Huan Ke,
Haryadi S. Gunawi,
Dominic Manno, David Bonnie,
Bradley W. Settlemyer

Correlated Failures

Correlated failures within compressed time windows make storage systems highly vulnerable to data loss.



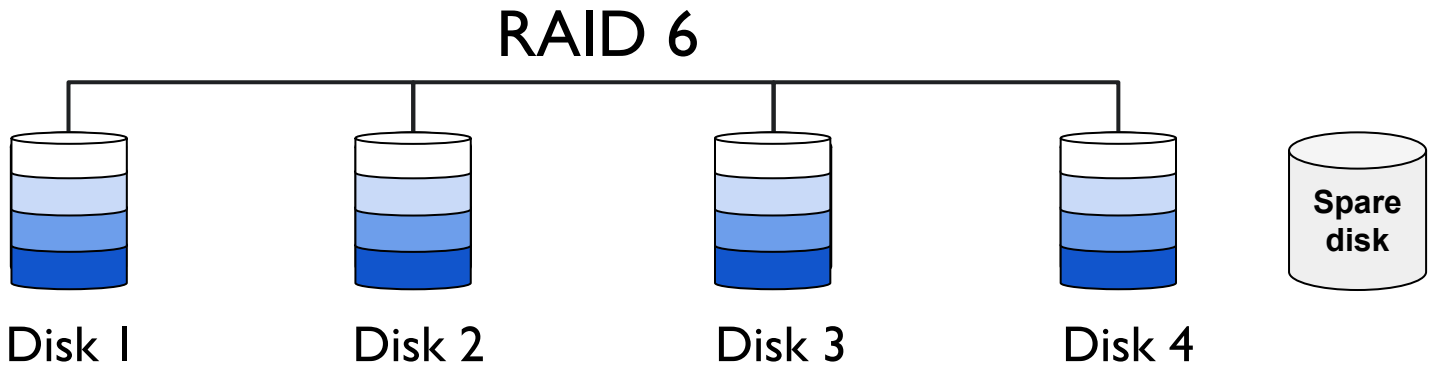
Failure Models

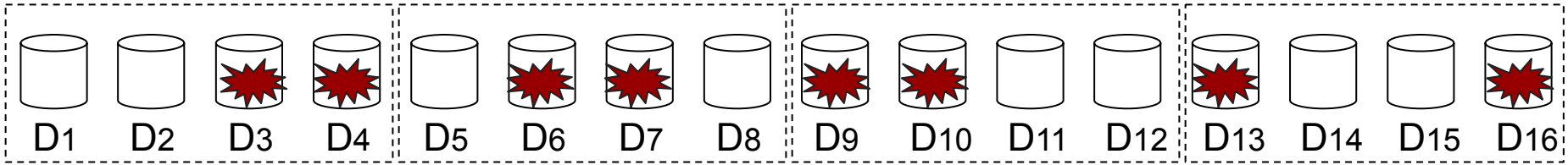
How do we model correlated failures ...

Types	Models
Poisson Failures	$Poisson(\frac{1}{MTBF})$
Exponential Failures	$Exp(\frac{1}{MTBF})$
Batch Failures	$Exp(\frac{1}{MTBF})\&Exp(\frac{1}{0.1MTBF})$

Traditional RAID

RAID (Redundant Array of Inexpensive Disks)

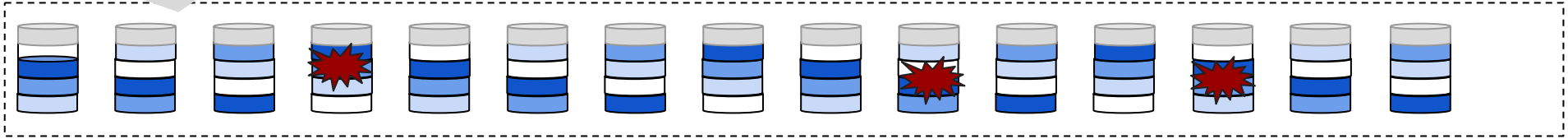




Declustered Parity (DP)

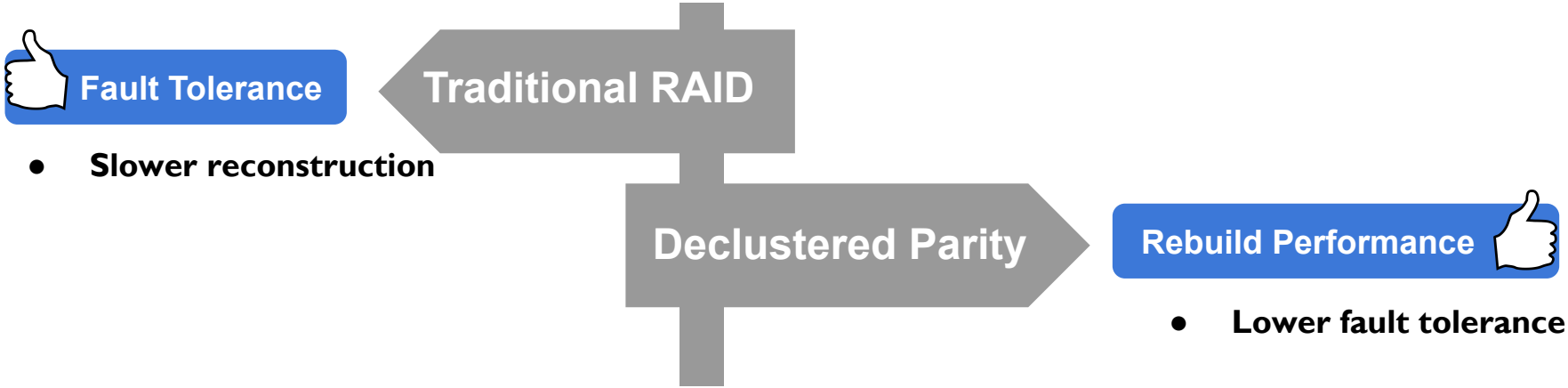
Data/parity are declustered or spread across all disks.

distributed spare space | parallel reads/writes | ZFS dRAID | GridRAID



The probability of data loss is 100%

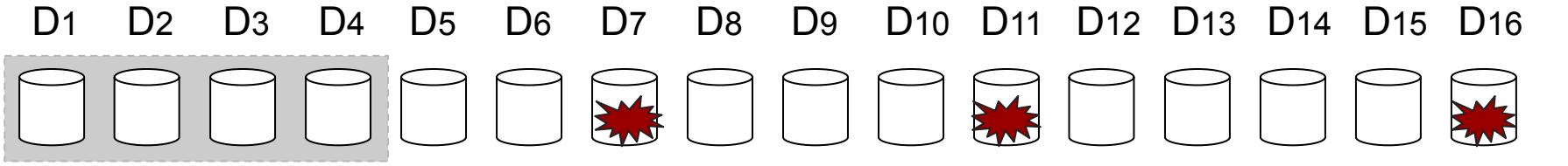
Motivations



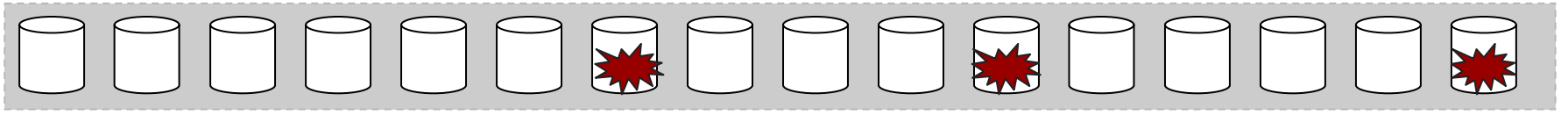
How the interactions between fault tolerance and rebuild performance together impact system reliability is still unclear.

Fractional Overlap Declustered Parity

FODP, a flexible tool to explore the middle space between fault tolerance and rebuild performance.



- Flexible rebuild performance
- Uniform data distribution
- Adjustable failure domains
- Higher fault tolerance

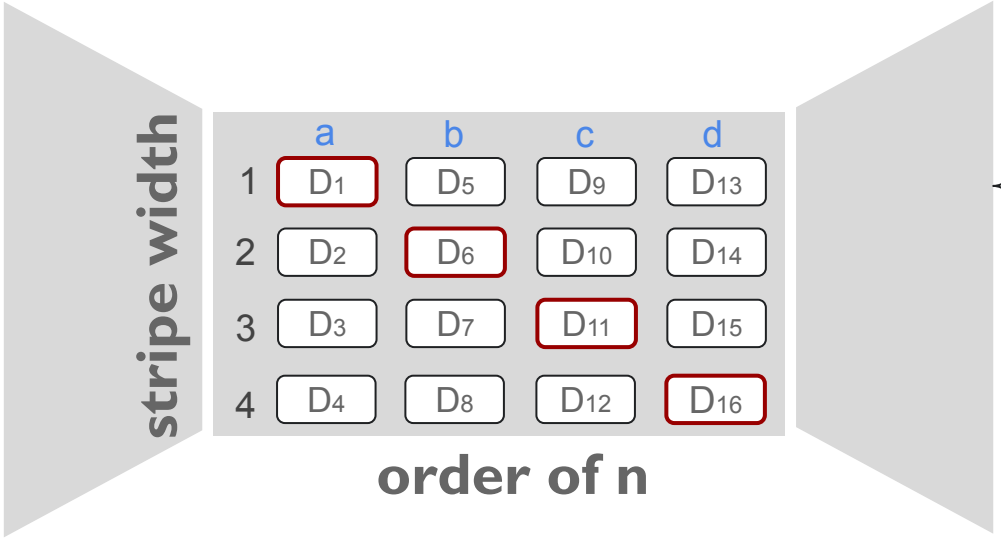


FODP Construction

Latin square of order n:

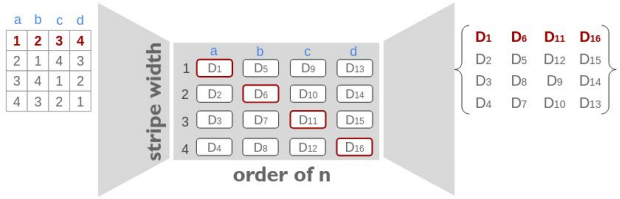
- a n×n array over n elements and each element appears once in each row and column.

a	b	c	d
1	**2**	**3**	**4**
2	1	4	3
3	4	1	2
4	3	2	1



D1	**D6**	**D11**	**D16**
D2	D5	D12	D15
D3	D8	D9	D14
D4	D7	D10	D13

Overlap fraction



Each latin square corresponds to n disk subsets that cover the whole disk matrix.

- Each disk has (stripe-width-1) **overlaps** within a disk subset.

Overlap fraction λ for each disk:

$$\lambda = \frac{overlaps}{N-1}$$

	RAID	FODP	SODP	DP
	λ_{min}	$\lambda < 1$	$\lambda = 1$	$\lambda > 1$
Rebuild Perf	L	M	H	H
Fault Tolerance	H	H	M	L

Mutually Orthogonal Latin Squares

Two latin squares are **mutually orthogonal**:

- Any order pair of entries from each latin square in the same row and column occurs exactly once.

1	2	3	4
2	1	4	3
3	4	1	2
4	3	2	1

+

1	3	4	2
2	4	3	1
3	1	2	4
4	2	1	3

→

1,1	2,3	3,4	4,2
2,2	1,4	4,3	3,1
3,3	4,1	1,2	2,4
4,4	3,2	2,1	1,3

- With any given order of n, there can be at most (n-1) mutually orthogonal latin squares (MOLS).

MOLS in FODP

L_1

	a	b	c	d
1	1	2	3	4
2	2	1	4	3
3	3	4	1	2
4	4	3	2	1

L_2

1	3	4	2	
2	4	3	1	
3	1	2	4	
4	2	1	3	

L_3

1	4	2	3	
2	3	1	4	
3	2	4	1	
4	1	3	2	

	a	b	c	d
1	D1	D5	D9	D13
2	D2	D6	D10	D14
3	D3	D7	D11	D15
4	D4	D8	D12	D16

D1	D6	D11	D16	$\lambda = \frac{3}{15}$
D2	D5	D12	D15	
D3	D8	D9	D14	
D4	D7	D10	D13	

D1	D7	D12	D14	$\lambda = \frac{6}{15}$
D2	D8	D11	D13	
D3	D5	D10	D16	
D4	D6	D9	D15	

D1	D8	D10	D15	$\lambda = \frac{9}{15}$
D2	D7	D9	D16	
D3	D6	D12	D13	
D4	D5	D11	D14	

Trade-offs in FODP

FODP gives us the flexibility to explore the trade-offs between fault tolerance and rebuild performance.

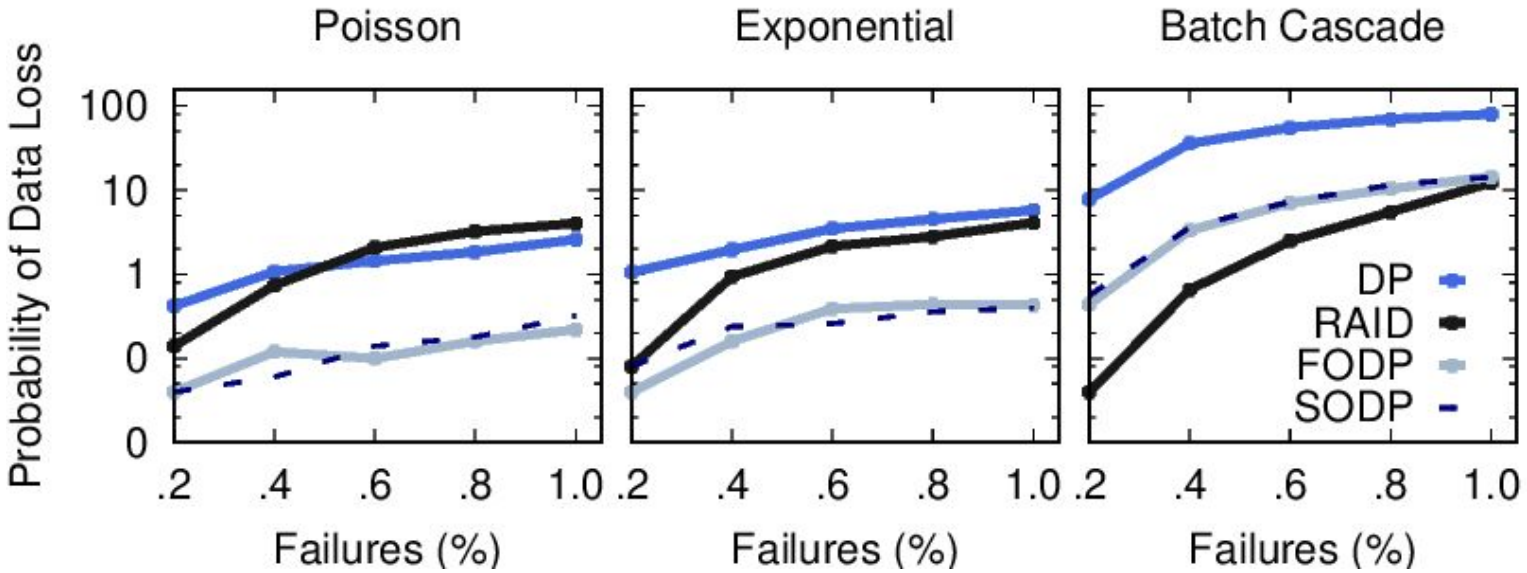
- The lower λ is, the more failures that can be tolerated.
- The larger λ is, the more overlaps can be used for rebuilds.

FODP+1

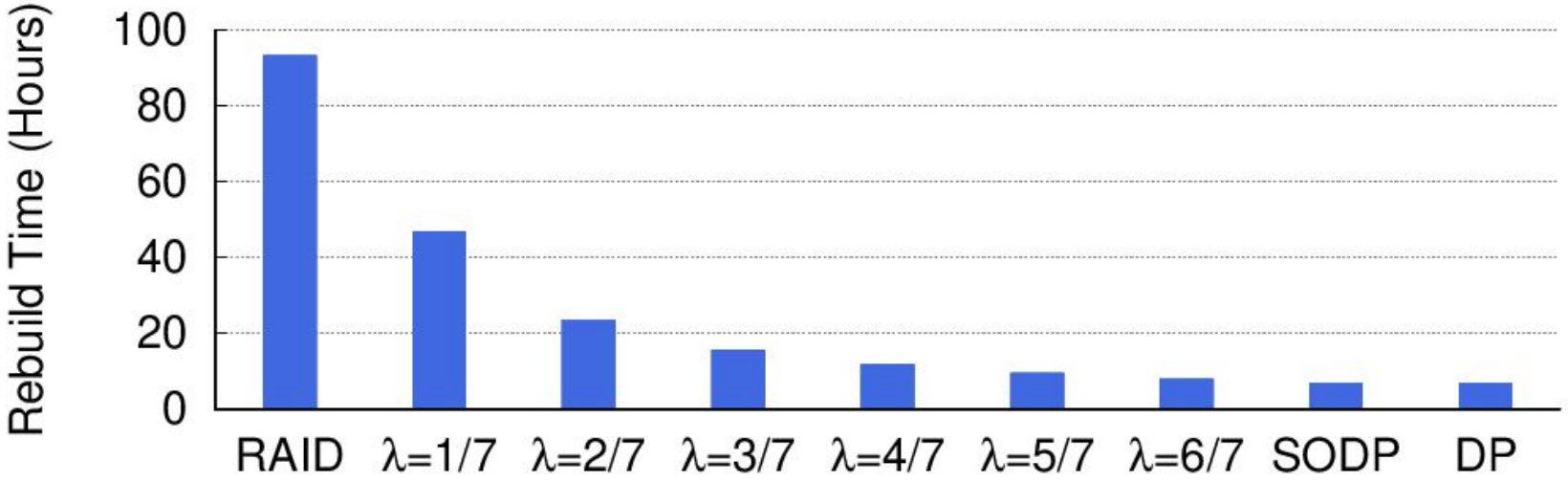
If data loss occurs, FODP loses more data than DP

Impact of Failures

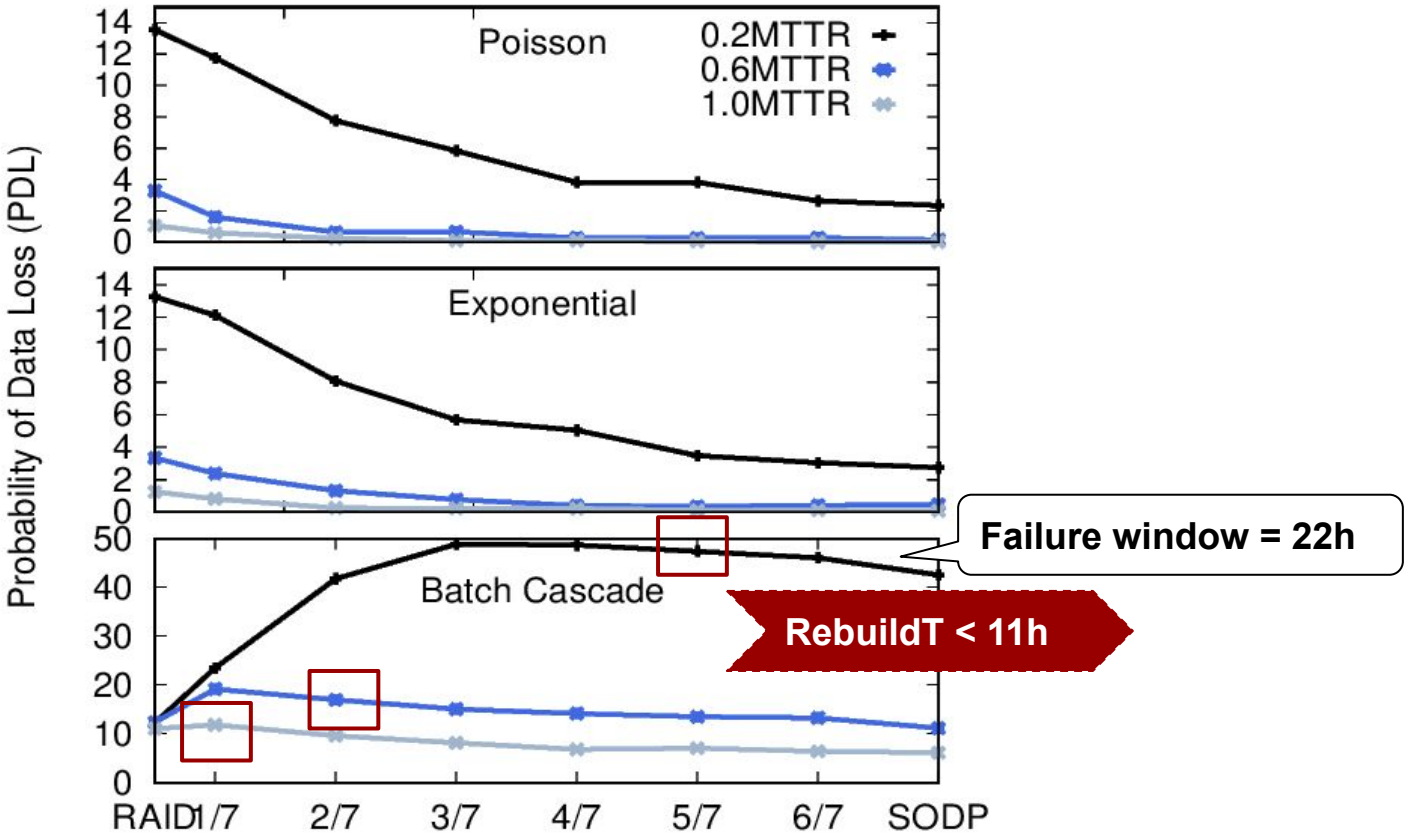
Assume MTBF = 0.5 MTTR in Campaign system with 11+2 configurations within each server.



Impact of Overlap Fraction



Impact of Overlap Fraction



FODP Conclusion

"*Why should we address correlated failures?*"

Storage systems are becoming **larger** and **denser** and failures are increasingly **correlated in time**!

FODP, a flexible tool to study and explore rebuild performance and failure domains in systems.

FODP-Plus-One, reducing the magnitude of data loss by adding a layer of parity on top of FODP stripes.

CERES

Thank you!
Questions?

THE UNIVERSITY OF CHICAGO

Los Alamos NATIONAL LABORATORY EST. 1943

http://ucare.cs.uchicago.edu