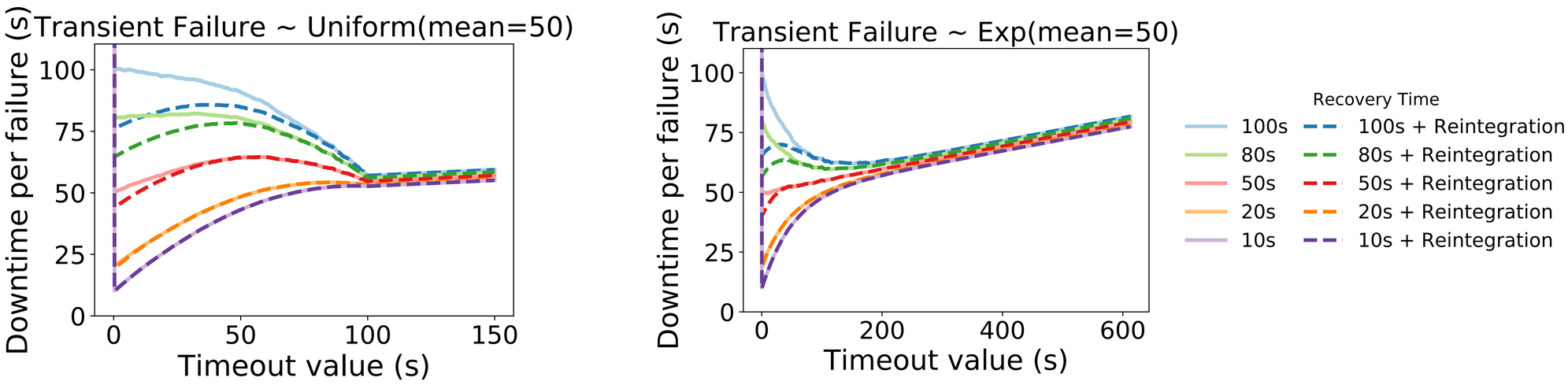
Failure modeling for distributed storage

Daniel Wong, Thomas Kim, Michael Kaminsky, David G. Andersen, Greg Ganger

Motivation: When should we be more aggressive about failure detection?

The optimal timeout depends on the mean transient failure length v.s. the recovery time.



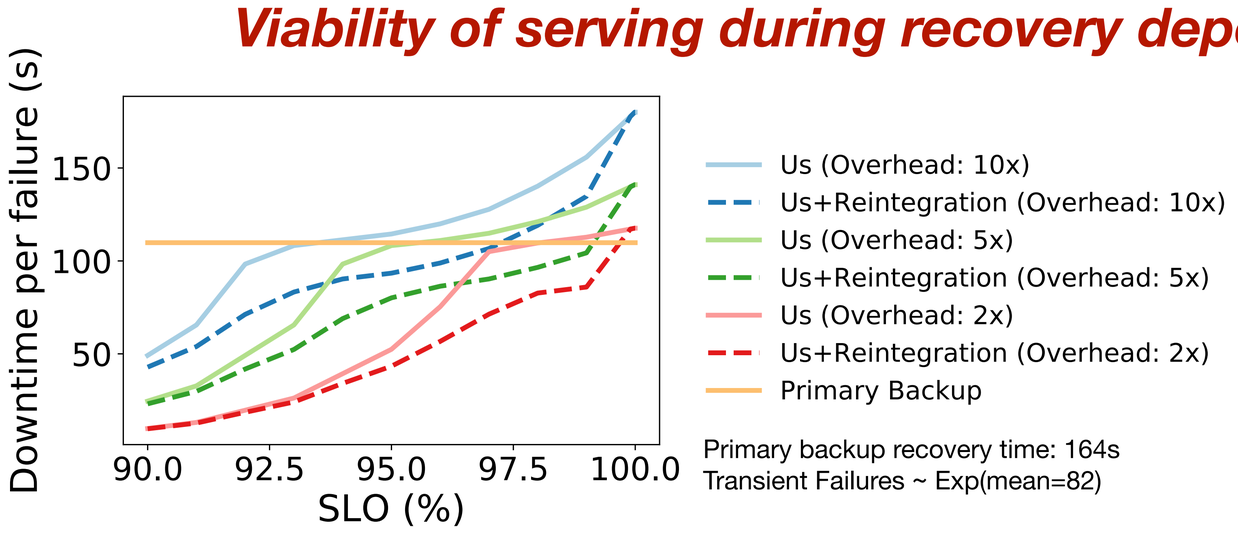
- Ratio of 1 : 20 for permanent : transient failures
- Recovery time: time to achieve desired SLO

***E[T]* < recovery time**	***E[T]* > recovery time**
Low timeout better	High timeout better

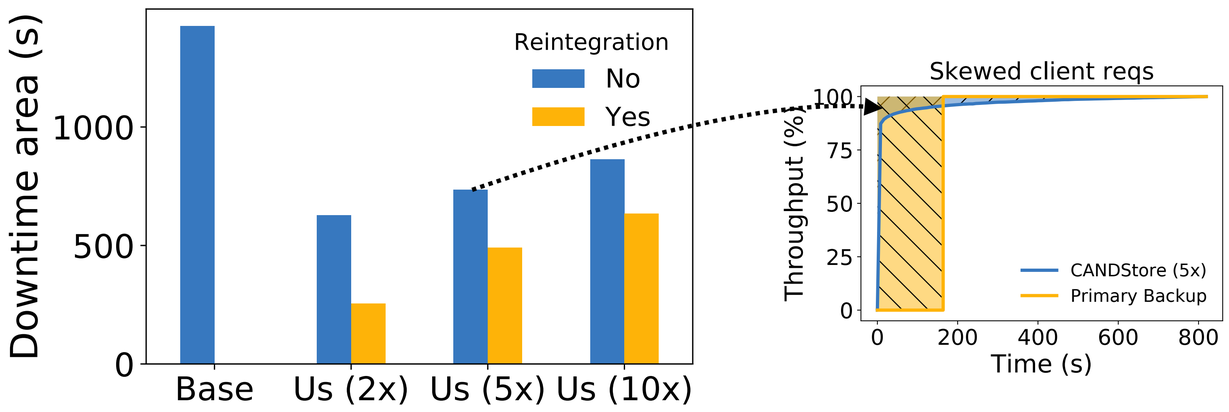
Takeaways

- Reintegration applies to failures of length *T*, where timeout < *T* < recovery time
- Reintegration enables low timeouts even for long recovery times

Viability of serving during recovery depends on desired SLO v.s. overhead.



- Overhead: multiple of time for primary backup to recover fully
- SLO: % of throughput
  - With 100 shards, 1 shard's failure drops availability to 99%
  - Time to 99.9% = time for 1 shard to 90%

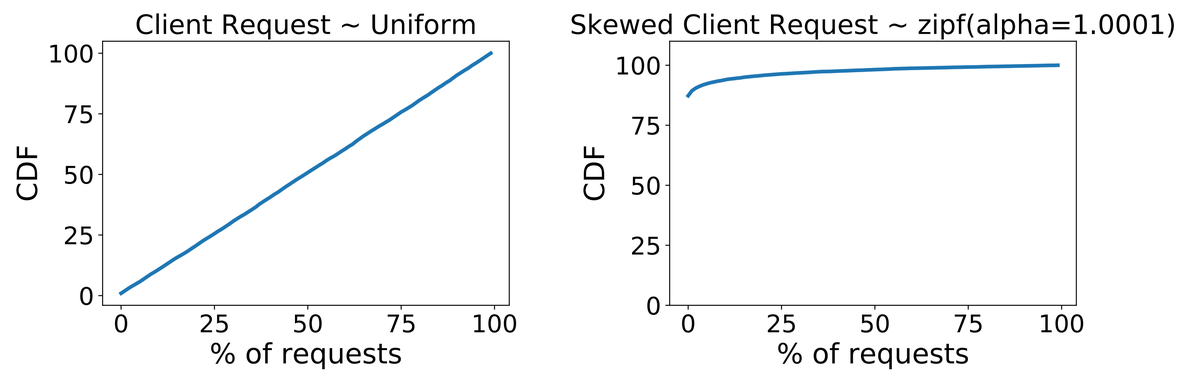


CANDStore

- Serve while recovering (instead of shutting down the world)
- Reintegration after transient failures

Benefits depend on

- Ratio of SLO to overhead
  - Don't need high throughputs
  - Low recovery time overhead
- Skewed client request distribution
  - serve most requests with small % of keys
- Value of partial availability



What determines your operating regime?

Optimal timeout and viability of reintegration depend on:

1. Recovery time of system
2. Expected length of transient failure
3. Ratio of transient to permanent failures

No observed relationship between operating regime and:

- Number of failures
- Distribution of transient failures

Viability of serving during recovery depends on

1. Recovery time of system
2. Desired SLO
3. Skew in client request distribution

Things we would like to know

- Real world data about:
  - Distributions of causes of failures
  - Distributions of lengths of failures
  - Large groups of correlated failures