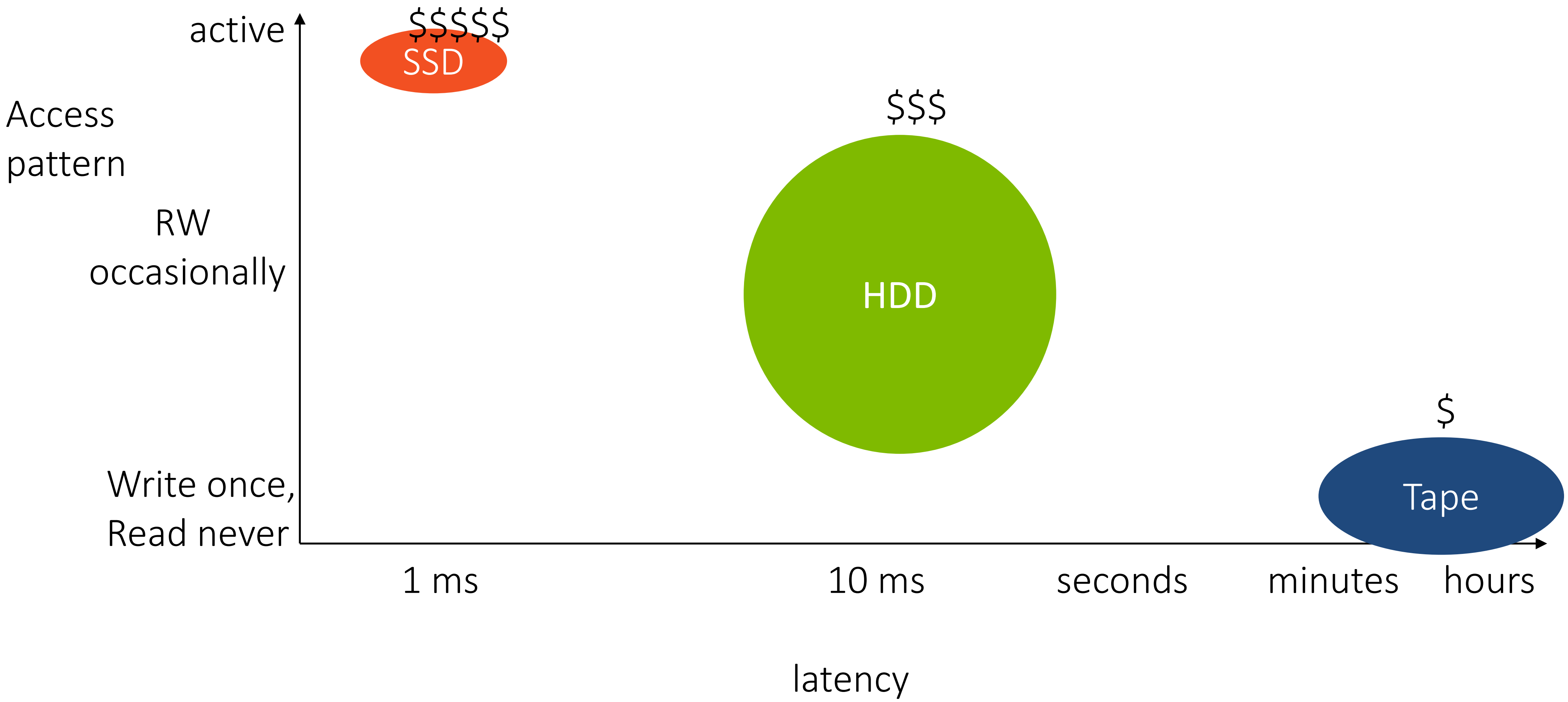
Low cost, highly dense Storage systems

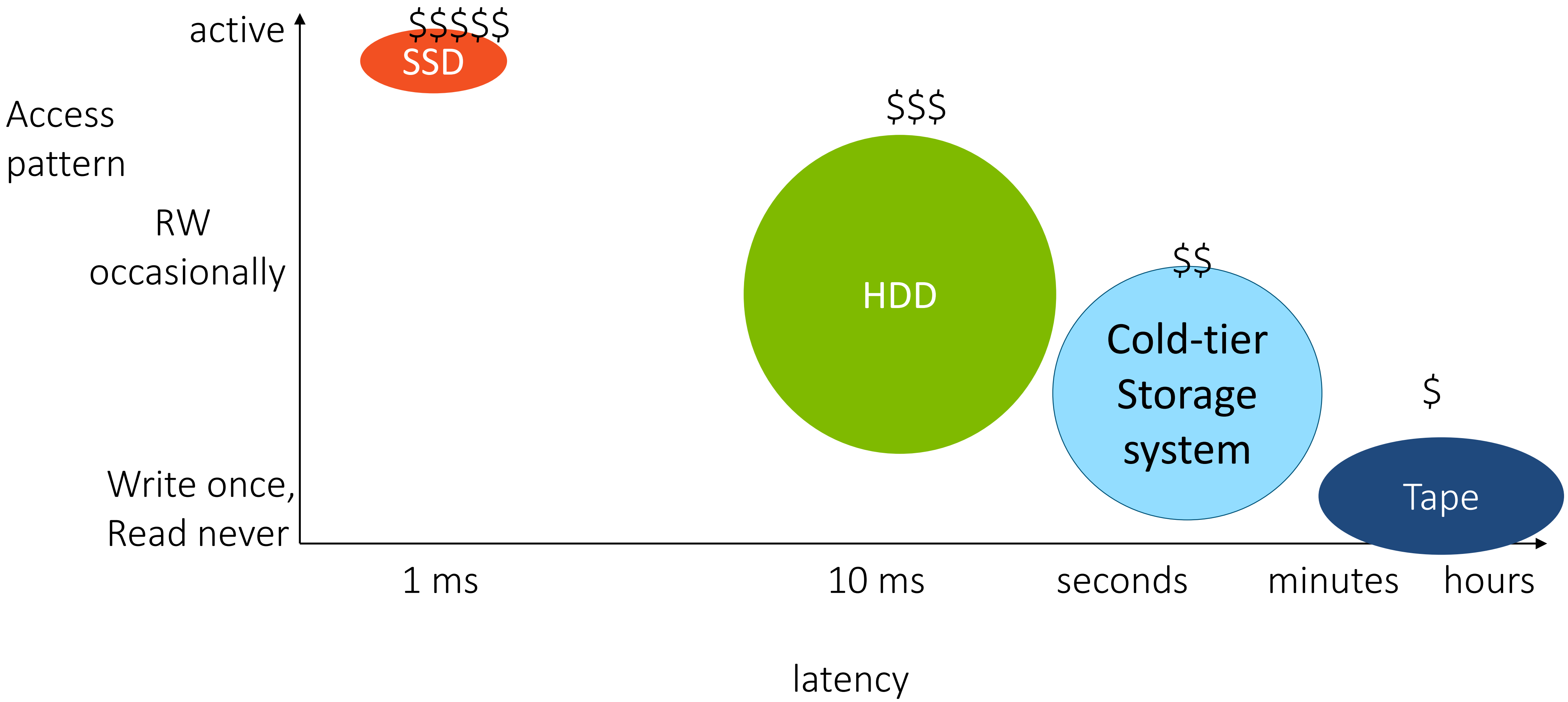
Designing Storage Architectures Meeting
Library of Congress, September 17, 2018

Pashupati Kumar, Principal SW Eng. Manager
Microsoft
Pashupati.kumar@Microsoft.com

Storage Hierarchy and Technologies



Storage Hierarchy and Technologies

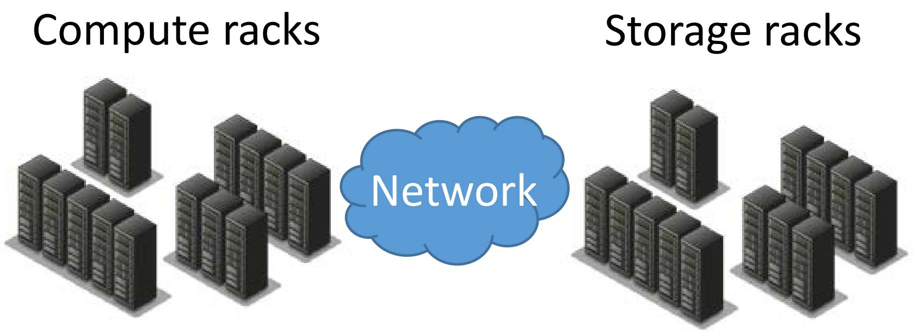


Goal

- Build the lowest cost HDD storage possible
- Deliberately trade performance for lower cost
- Avoid stranded storage
- Flexible performance characteristics
- Use commodity components

Driving storage cost down…

Common in the cloud:



Improves performance/cost:

- Independent resource scaling
- Rack hardware specialization

Reduce overheads in Storage racks!

1. Have large number of HDDs for each server
   - ✓Gola is have storage cost same as that of HDD
2. Power off drives that are not currently utilized
   - ✓Put them in lower power mode. E.g. Drives in Standby mode consume 50% less power than in Active Idle state
   - ✓20 – 25% OPEX saving can be realized

HDD – Power Conditions

- Performing HDD Power off/On is not flexible design options
  - Depends on JBOD enclosure implementation
- HDD supports different Power Conditions, that can be controlled via SW

Power Condition	Power (W)	Power Savings (%)	Recovery Time
Idle	2.82	0	
Idle_A	2.82	0	
Idle_B	2.18	23	
Idle_C	1.82	35	
STANDBY_Z	1.29	54	

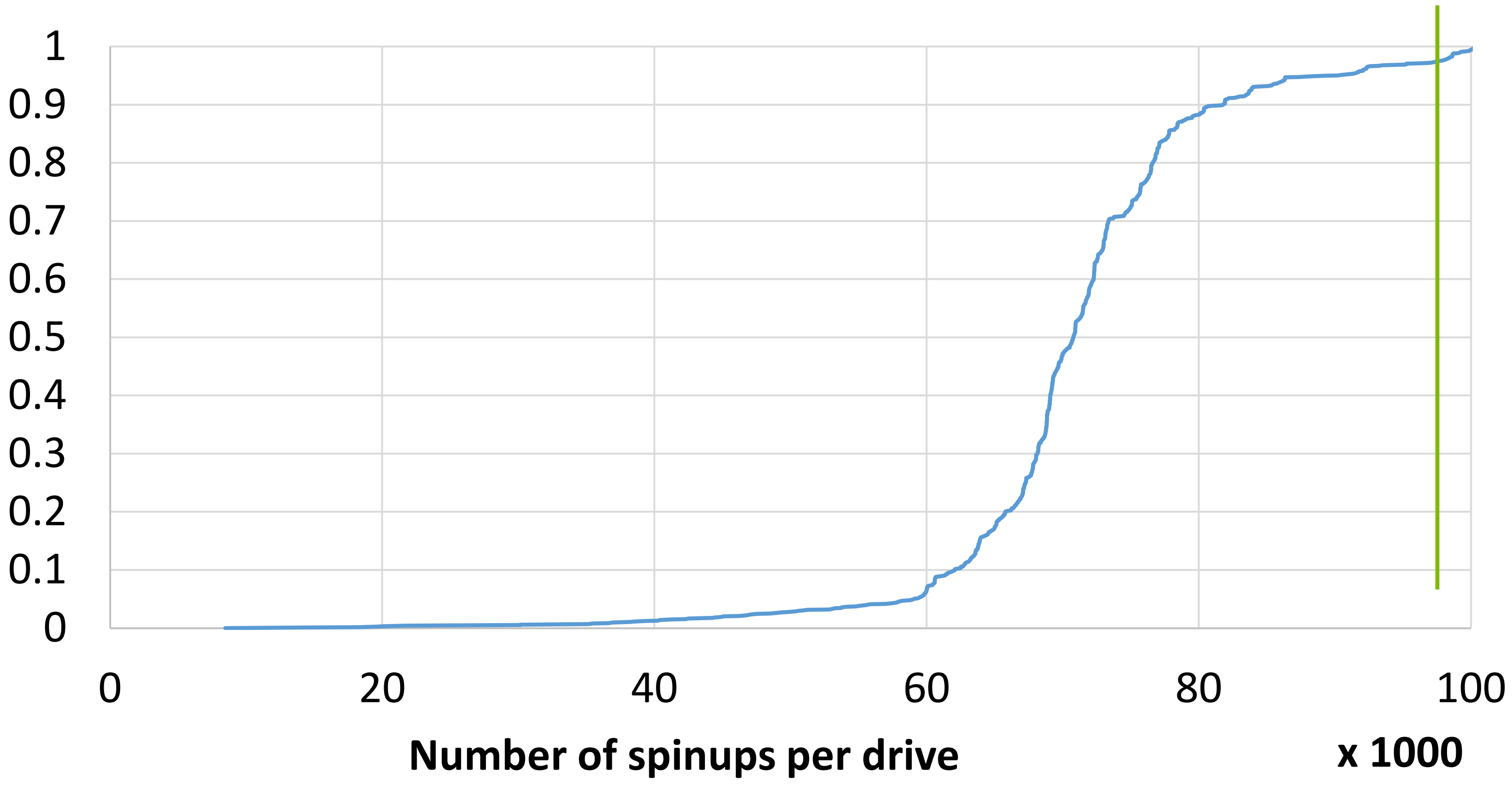
- Use standard SBC (START_STOP_UNIT, 0x1B) to go to desired power condition
- Method to determine current power condition are different for SATA & SAS drives
- SCSI Log pages are available for monitoring power transitions counters
  - Start/Stop Cycles counter Log page
  - Power Condition Transitions Log page

Challenges

- Spin-up and down cycle
  - current limitation and future progression
- Disk AFR
  - Need to characterize disk failure rates for this “new workload”
- Drive technology for cold/Archive use cases
- Power surge during standby to Active idle state

Is it ok to do all these spin-ups?

datasheet spec: 50K per year.



Avoid Stranded Storage

- Software can cope with loss of a server
  - But how much work does that cause?
    - Aggressive re-replication of data consumes lots of resources
    - Gets really worse, as storage server has 10-12 x HDDs

- Suppose data is still accessible
  - Even at a lower performance
  - Software can adjust load balancing
  - Much easier to handle, fewer resources used, lower COGS

Traditional SAS redundancy is expensive

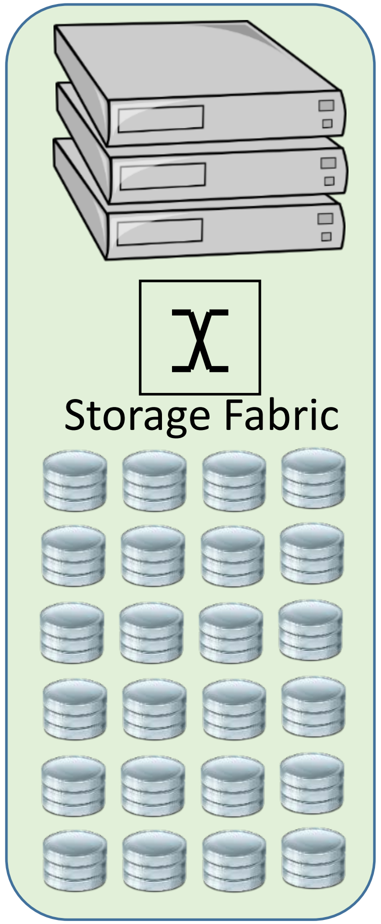
- Traditional method was SAS dual attached disks
  - More expensive disks
  - Dual links to the disks
  - Dual expander hierarchy
  - Dual everything
  - Massively wasteful and expensive
- Not actually what we want

Rack-Scale HDD Storage Disaggregation

- Relaxing the HDD Ownership Principle
  - At a given time, a HDD is managed by one server…
  - …but it is possible to reconfigure which server it is.
- Enables 4 types of disaggregation:
  - Configuration Disaggregation
  - Failure Disaggregation
  - Dynamic Elastic Disaggregation
  - Complete Disaggregation

No reconfiguration during normal operation

Reconfiguration part of normal operation



Rack Scale HDD Disaggregation

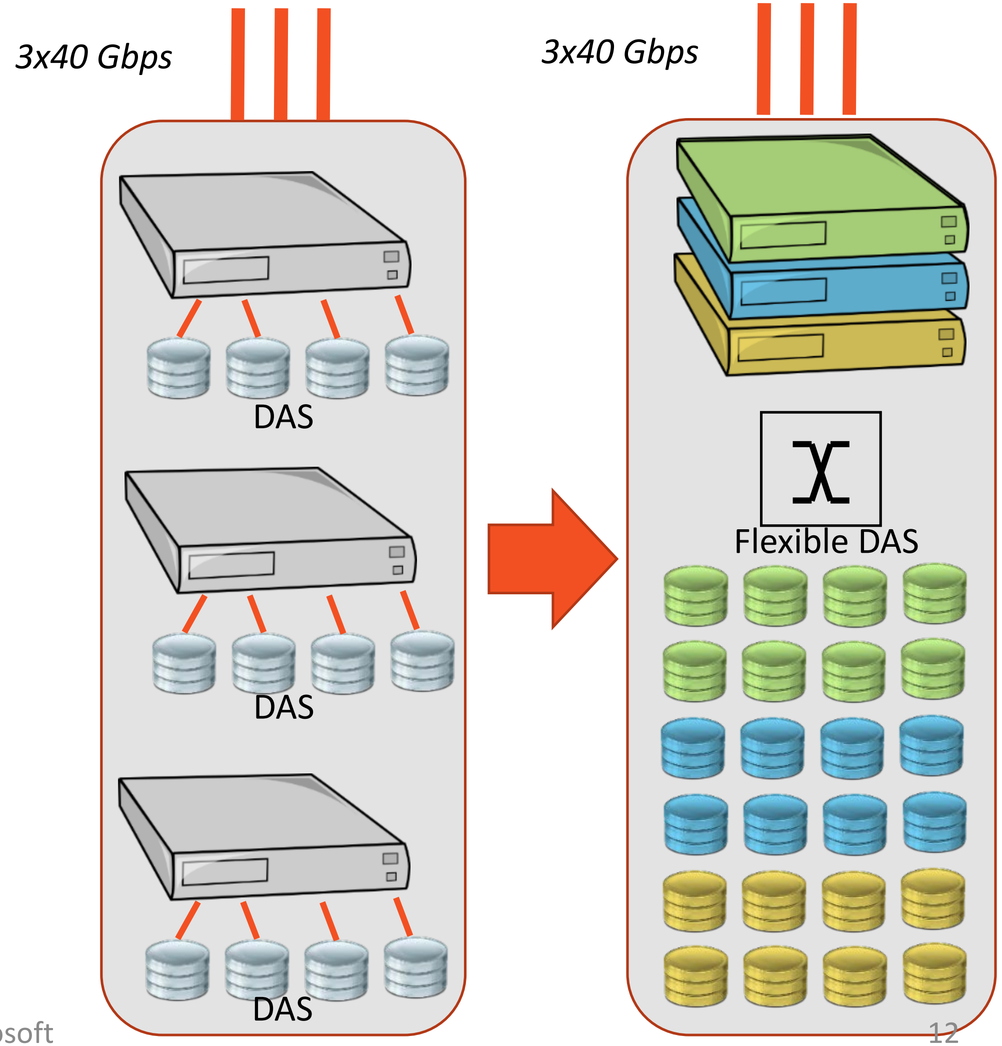
Rack bandwidth for storage:

For the Cloud: low cost components

- Commodity servers
- SATA HDDs

Any HDD connected to *any* server

- Server elasticity

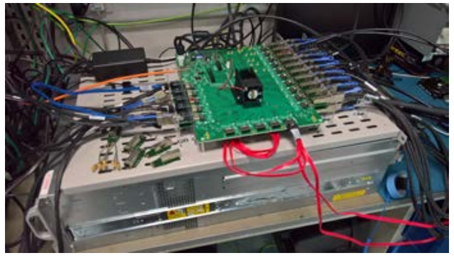


Experience with Failure Disaggregation

- Hardware trends impact data availability:
  - HDD and SSD capacities grow
  - Servers can have a LOT of direct-attached storage
  - e.g.: **Petabytes** of data per Pelican (cold storage) server
  - On failure, amount of data and time to recover increases
- Failure disaggregation improves availability
  - Reduces data unavailability to tens of seconds or less
  - No resources used to rebuild data
  - No reconfiguration overhead for normal operation

Pelican prototype has:
- **1152 HDDs/rack**
- **2 servers**

Conclusion



- In the cloud today: no disaggregation in storage racks
  - Fixed drive-to-server mapping
- We designed a storage fabric to explore in-rack disaggregation
- Rack-scale storage disaggregation can be useful and affordable
  - Configuration disaggregation
  - Failure disaggregation
  - Dynamic elastic disaggregation
    - Substantial benefits
    - No/small reconfiguration overheads
    - Little or no software/hardware changes

- Can become a challenge
  - Complete disaggregation
    - High reconfiguration overhead
    - Hard to implement and maintain

Performance

- Design biased for throughput
- User data is striped across many drives in a group
- Drive is assigned to a group with following consideration
  - Across multiple components
  - Minimal contention for storage bandwidth
  - Minimize overall rack vibration and cooling requirement

Configuration

- Breakdown

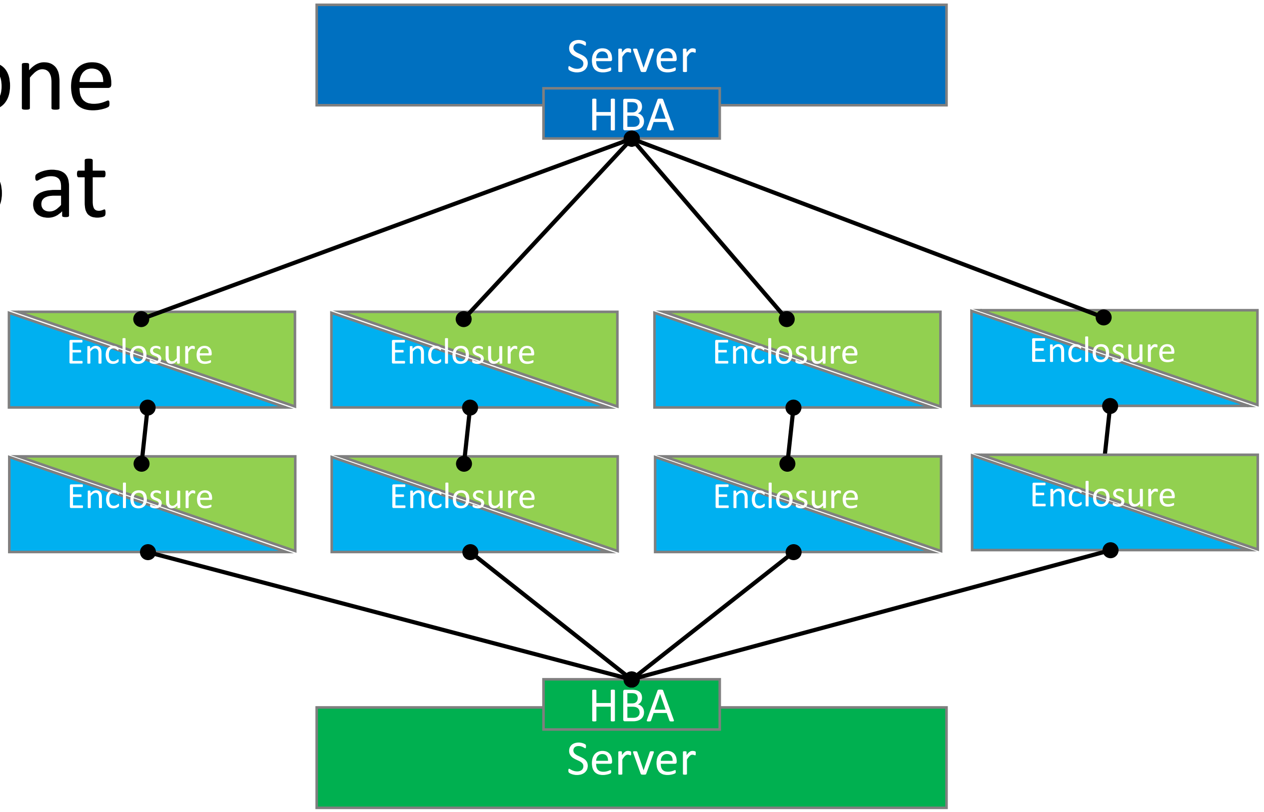
Servers	2
Leases	2
Classes/Lease	2
Groups/class	11 (Only 1 Group/Class can be Active
Disks/group	20
Total disks	2*2*11*20 = 880
Erasure coding scheme	15+3 (Overhead = 18/15 = 1.2)
% of disk in Active (on loaded system)	80/880 = 9% (72 / 880 = 8.2 %)

- HDD labelling in an enclosure

5	8	9	7	3	4	10	0	0	9	6
4	10	8	6	2	5	8	1	1	10	7
3	9	5	5	1	6	9	2	2	6	0
2	8	4	4	0	7	10	3	3	7	1
1	7	3	3	10	7	0	4	4	8	2
0	6	2	2	9	6	1	5	5	9	3
7	10	1	1	8	5	2	6	8	10	4
6	9	0	0	10	4	3	7	9	8	5

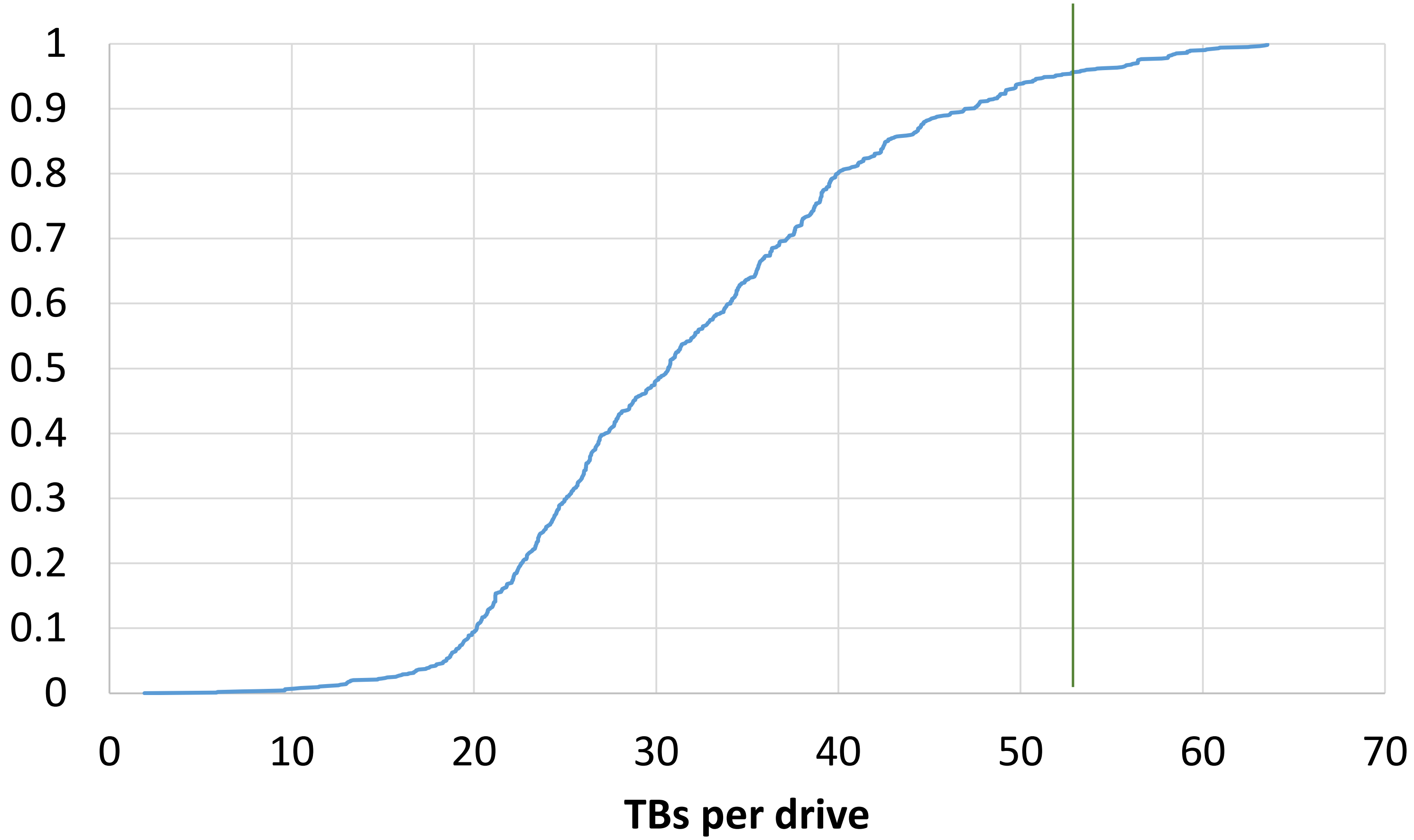
Another example

- Enclosures with dual cables
- With any single failure one server still has access to at least 7/8 of the disks



TBs transferred

datasheet spec: 60TB/year



Power On Hours

datasheet spec: 3120 POH/year (about 1/3rd of a year)