SNIA DEVELOPER CONFERENCE

SDC 24

BY Developers FOR Developers

September 16-18, 2024
Santa Clara, CA

Storage @Dropbox

Scaling Magic Pocket An Exabyte Scale Object Storage System

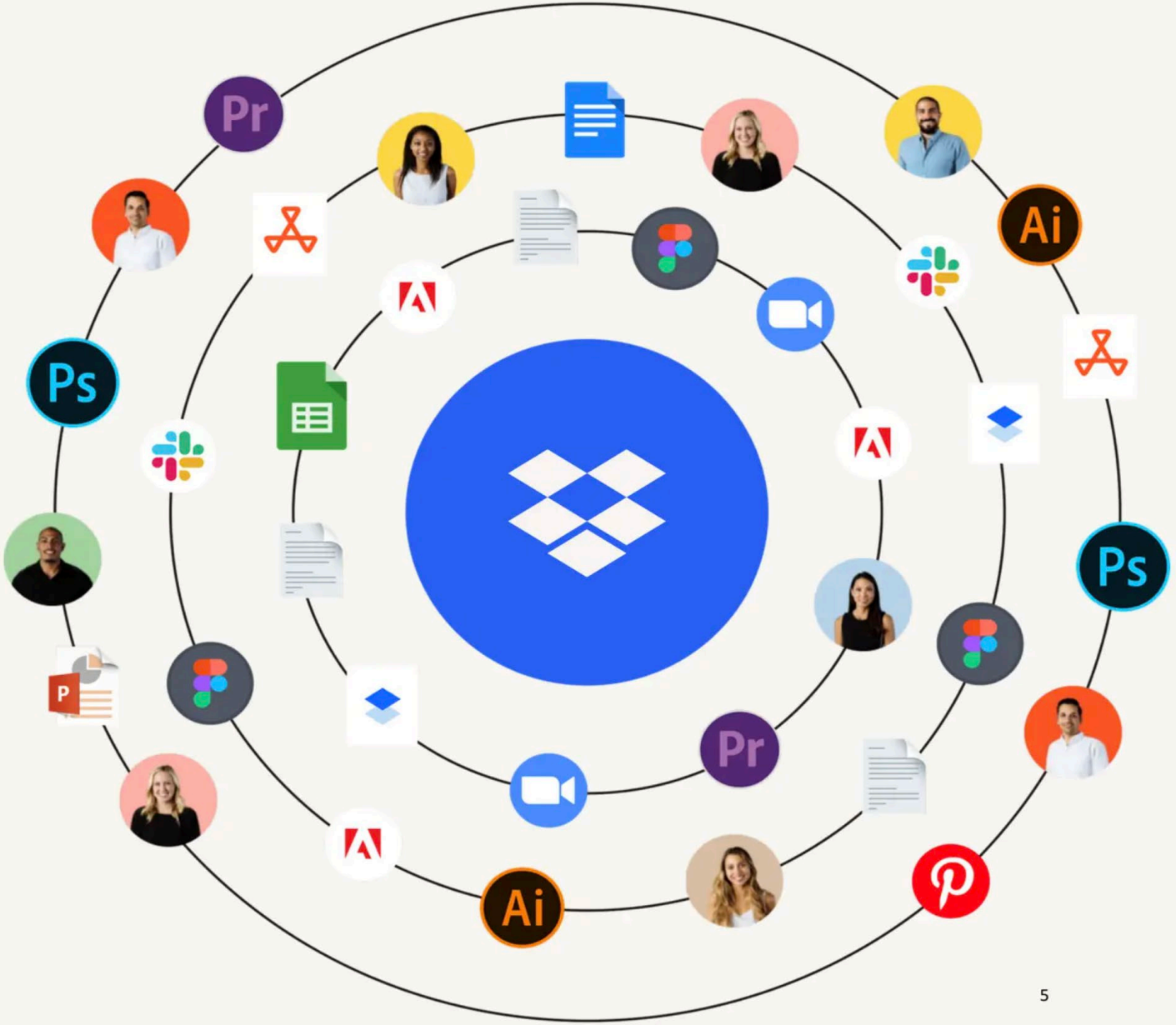
Sandeep Ummadi & Eric Shobe

Global Collaboration Platform at Scale

700M+
Registered Users(1)

18.22M
Paying Users(1)

1T+
Pieces of Content(1)



(1) As of 6/30/2024

Dropbox

Source: https://dropbox.gcs-web.com/news-and-events/presentations

SDC 24

Agenda

- Architecture
- SMR Adoption & Journey
- Operational Insights
- Hardware Evolution
- Storage Platform Insights
- What's next?
- Q&A

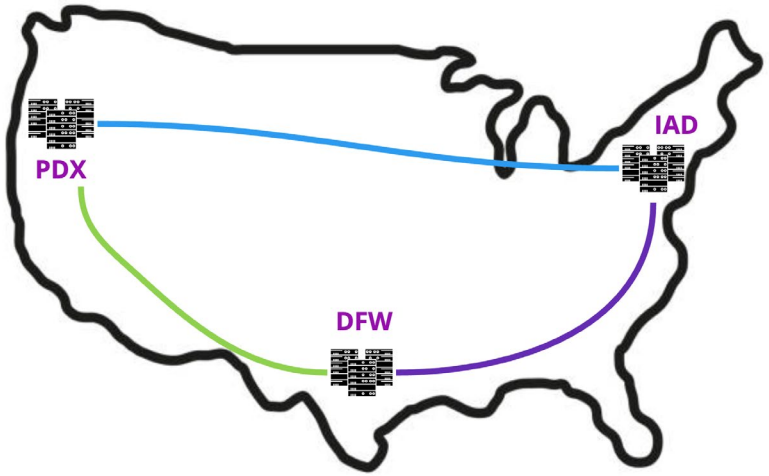
Magic Pocket

Architecture

Overview

Availability	Durability	PUT	GET
4 9s	11+ 9s	Tbps	Tbps

High availability and Durability guarantees



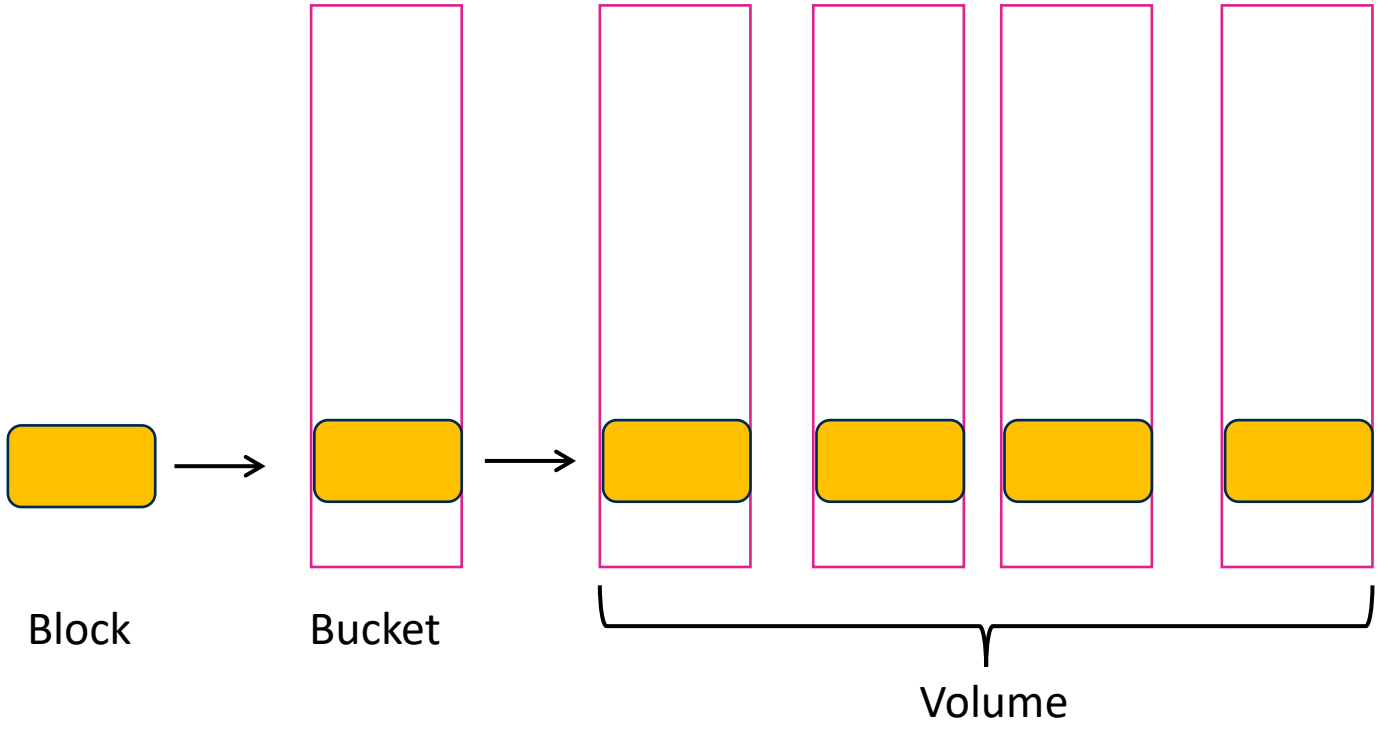
Geo-Replicated Key-Value Blob Storage System

API

- PUT
  - Key, Value {Data}, MD5 (data)
- GET
  - Key
- DELETE
  - Key

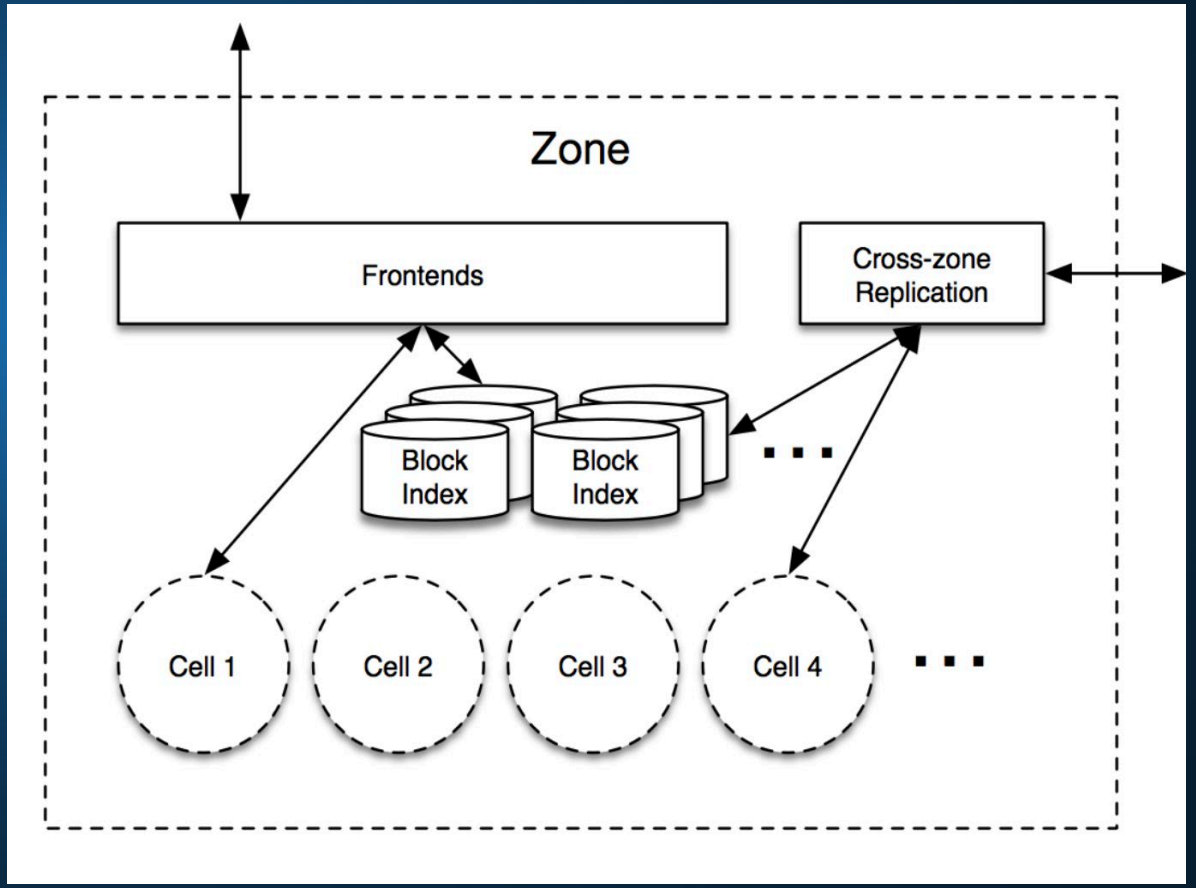
Data Types and Terminology

- Hash (key): Unique identifier to address a block.
- Block (value): One data value
- Bucket: xGB logical append-only block group
- Volume: Replicated group of 1 or more buckets
- Extent: xGB on disk data stored and managed by an OSD (Object Storage Device)



Topology

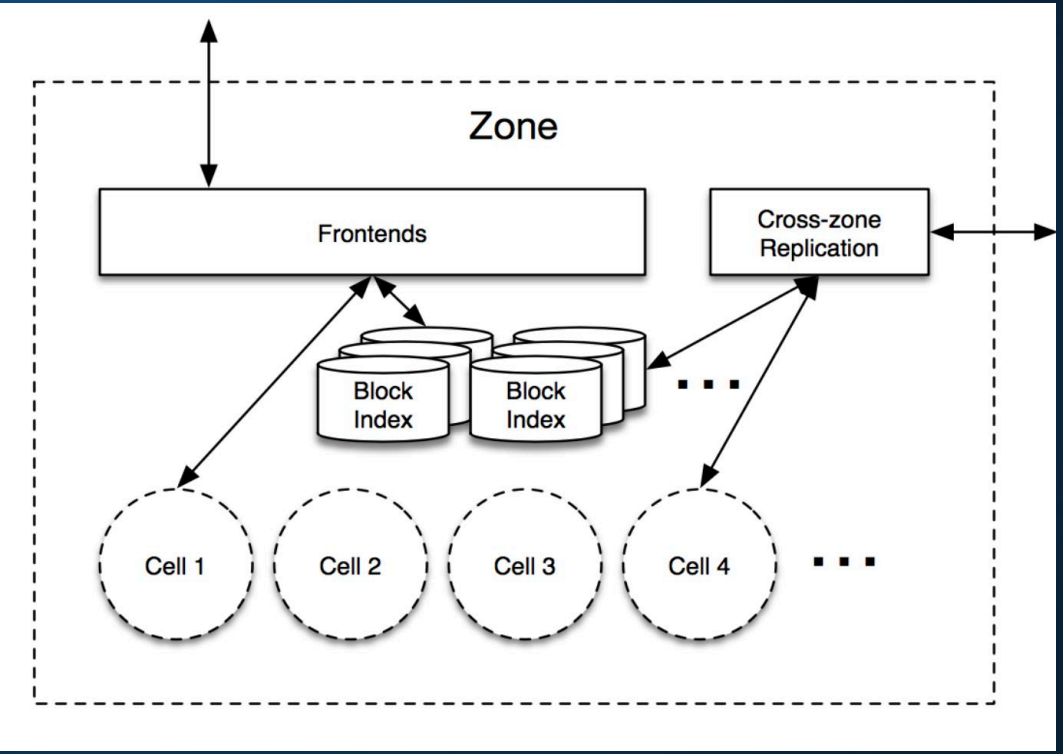
- Pocket – Consists of two or more zones
- Zone - Contains 1..N cells
- Cell contains N storage machines
- Cell is a scaling unit within a zone



Core Components

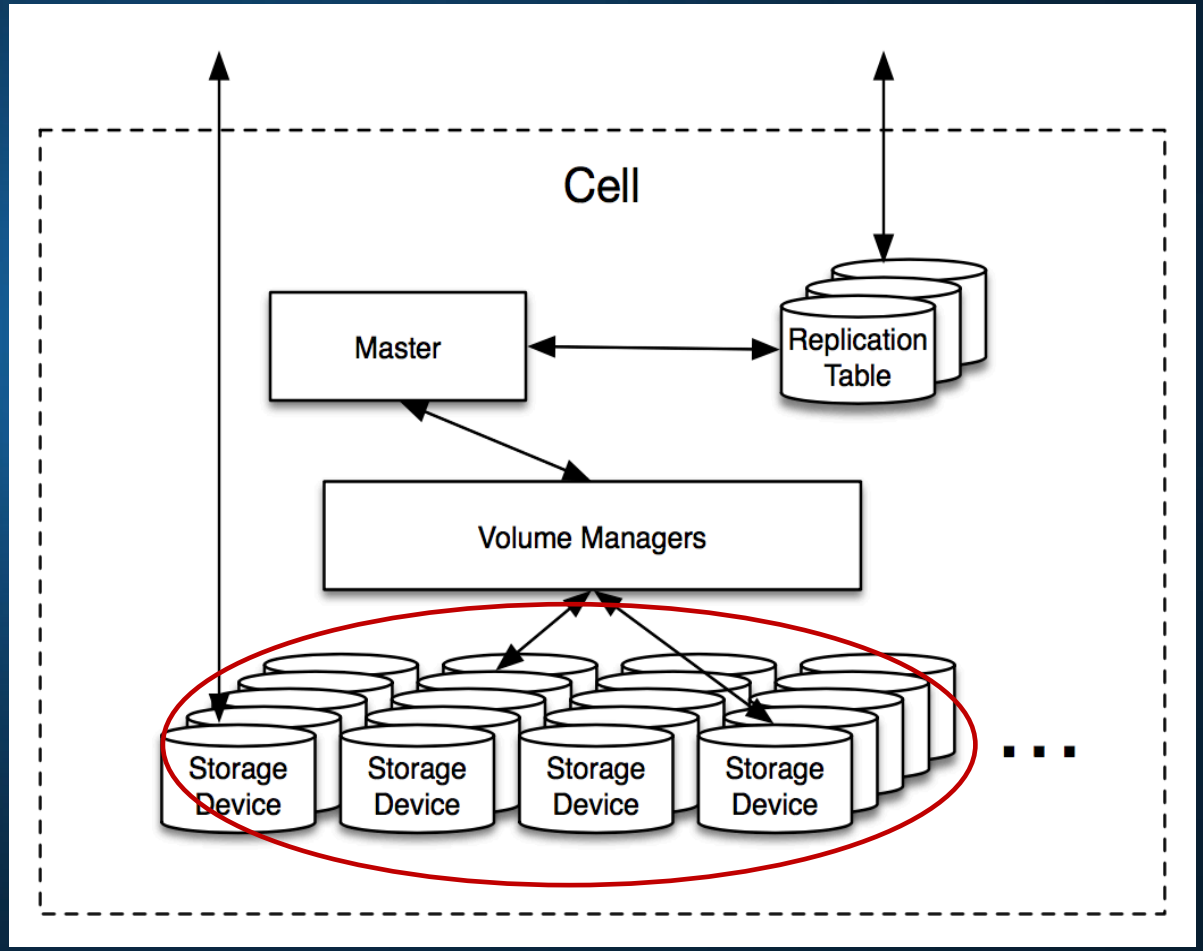
Frontends

- Operates at Zone level
- Horizontally Scalable Stateless Service
- Only public endpoint of the system
- Handles all high-level APIs – PUT, GET, DELETE, etc.,



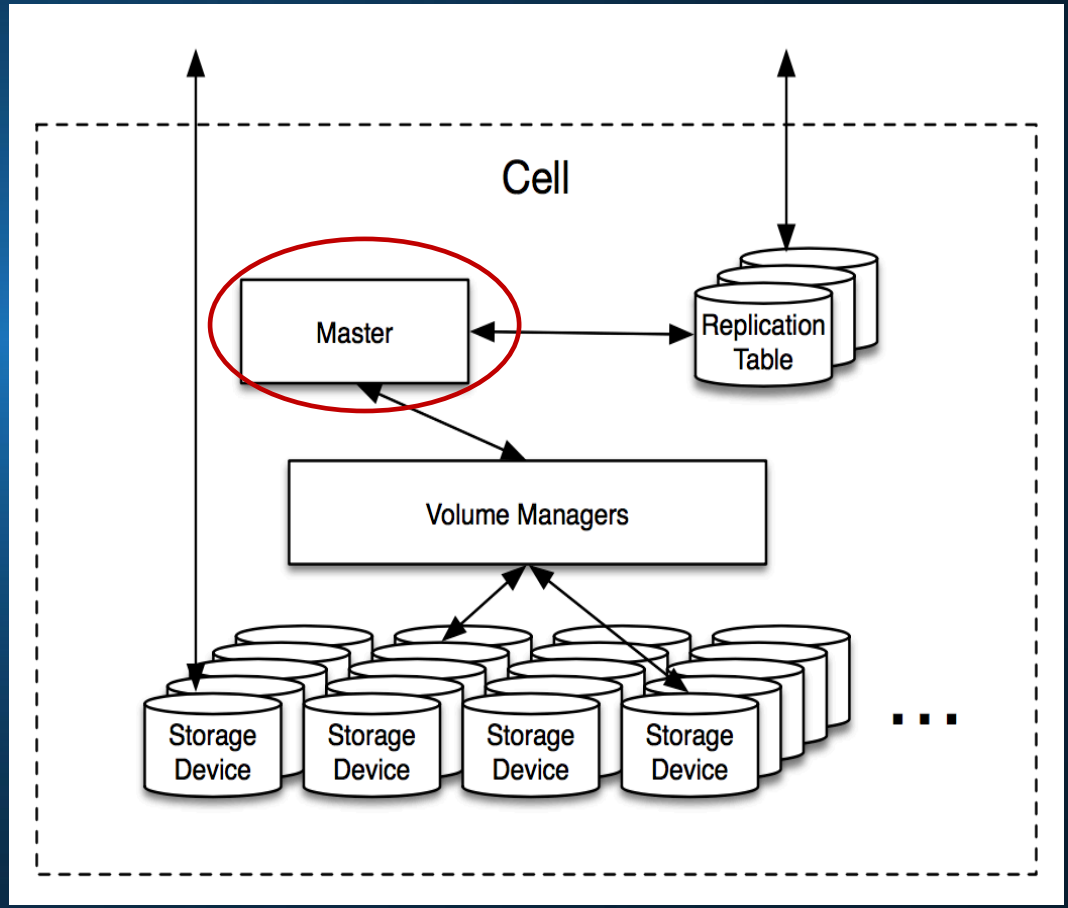
OSD (Object Storage Device)

- Scope: Local to a cell
- OSD Machine: ~100 HDDs
- OSD Daemon: 1 per disk
- Each daemon is a key-value store
- Persistent layer for blocks
- Blocks are maintained in extents



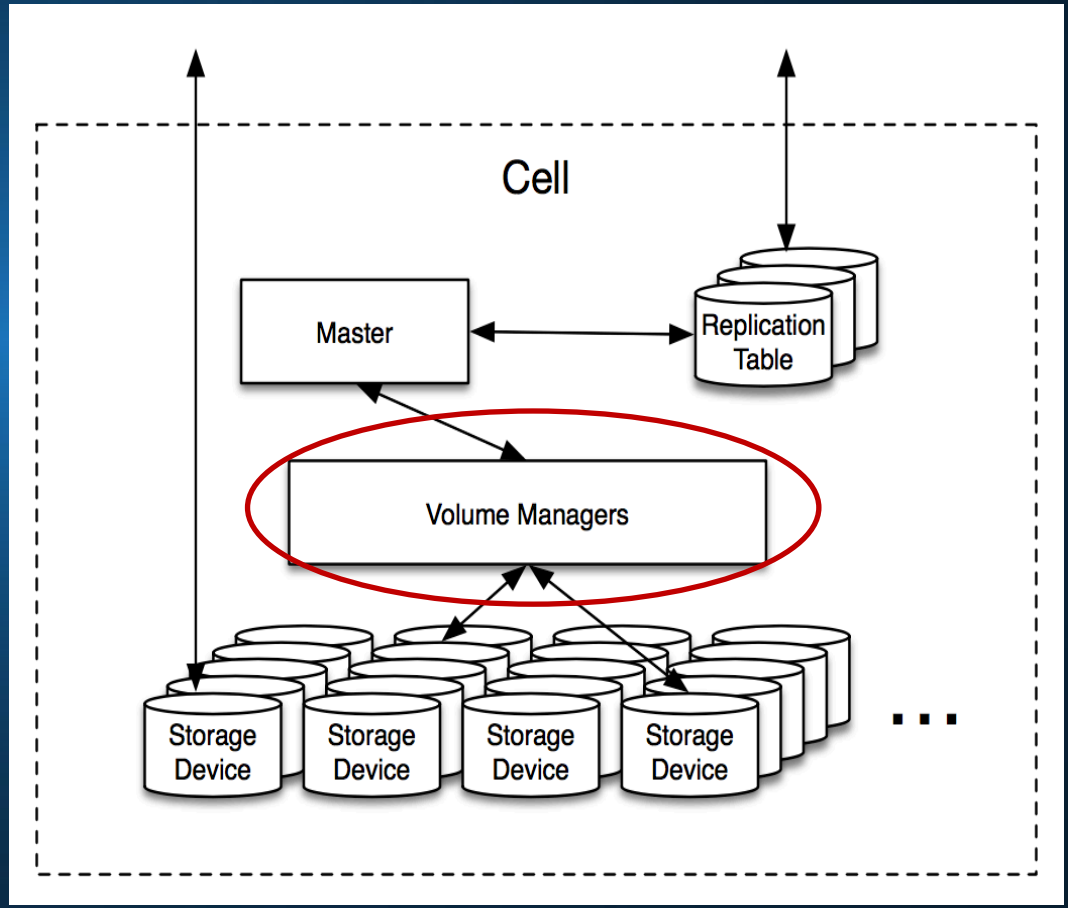
Master

- Scope: Local to a cell, 1 per cell
- It's not in critical I/O path
- Central coordinator in a cell
  - Health checks, failure detections & repairs
  - Balancing OSDs
  - Compaction
  - Erasure Coding
  - Etc.,



Volume Manager

- Scope: Local to a cell, 10s of volume managers
- Executes tasks initiated by master
  - Erasure code volumes
  - Merge & Repack volumes
  - Reconstruct missing extents (OSD failures)
  - Reconstruct Reads (GET I/O path)

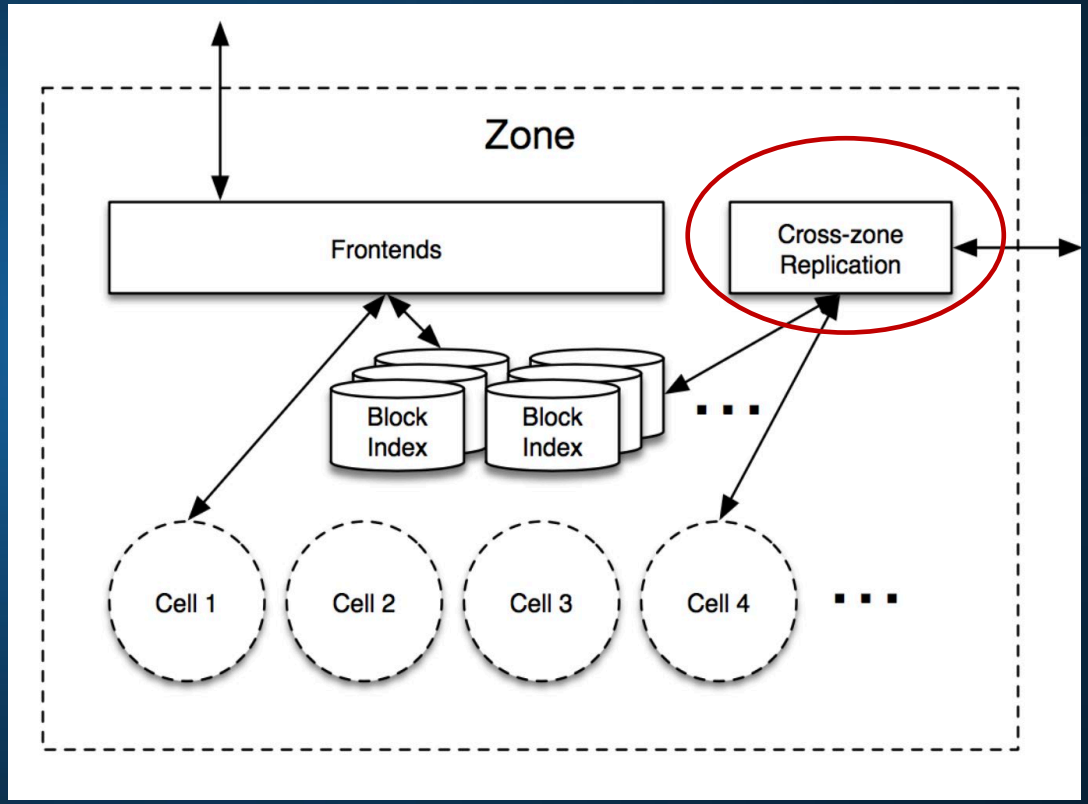


Block Index & Replication Table

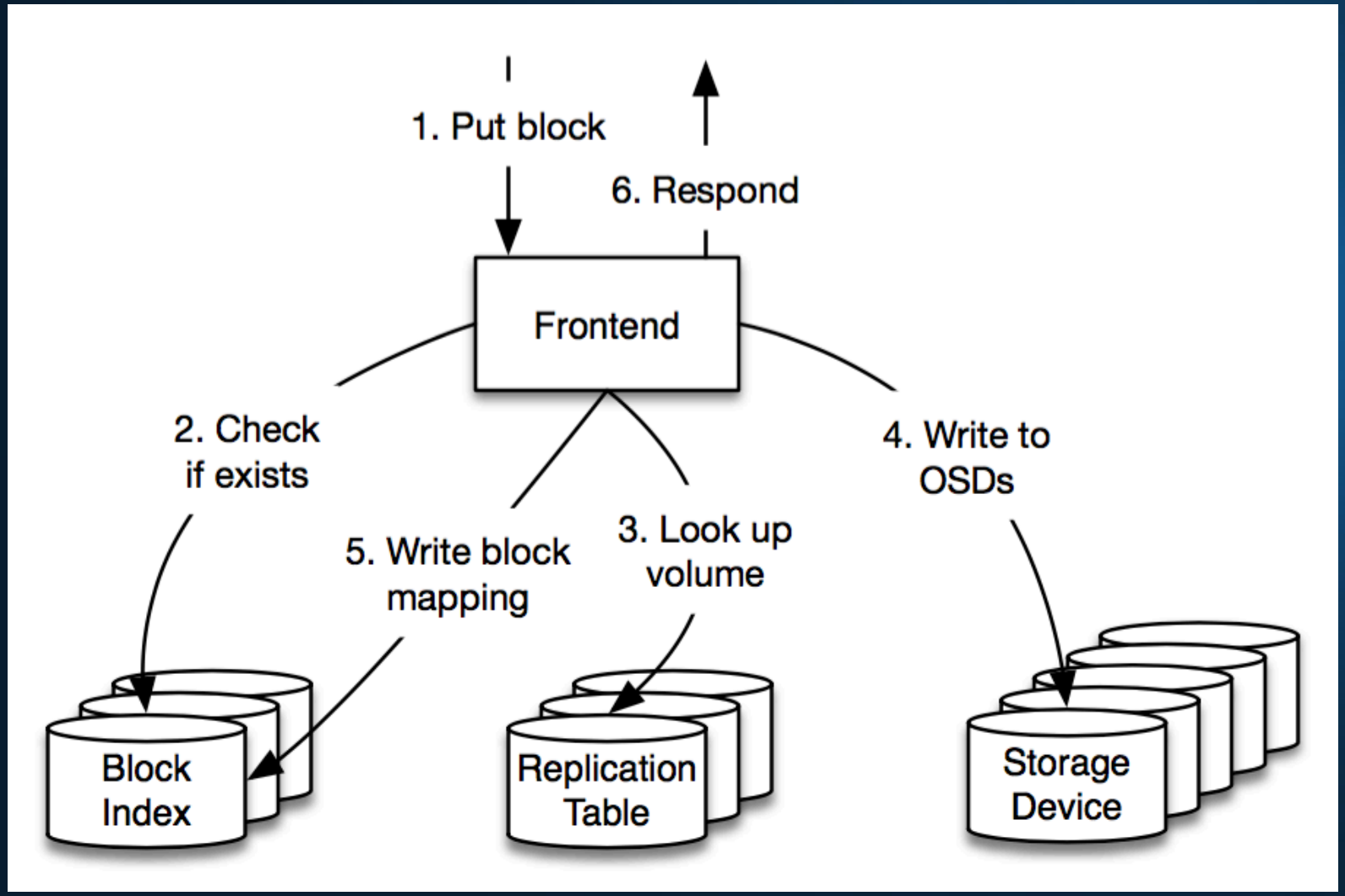
- Block Index
  - Scope: Zone level
  - Sharded MySQL
- Block to bucket where its stored
  - hash → cell, bucket, checksum
- Replication Table
  - Scope: Per Cell
  - MySQL, much smaller compared to Block Index
- Bucket to Volume and OSD Information
  - bucket → volume
  - volume → OSDs, open, type, generation

XZR (Cross Zone Replication)

- Scope: Zone wide
- Replicates block to configured remote zone
  - Gets block from local zone
  - Performs PUT operation on remote Zone.



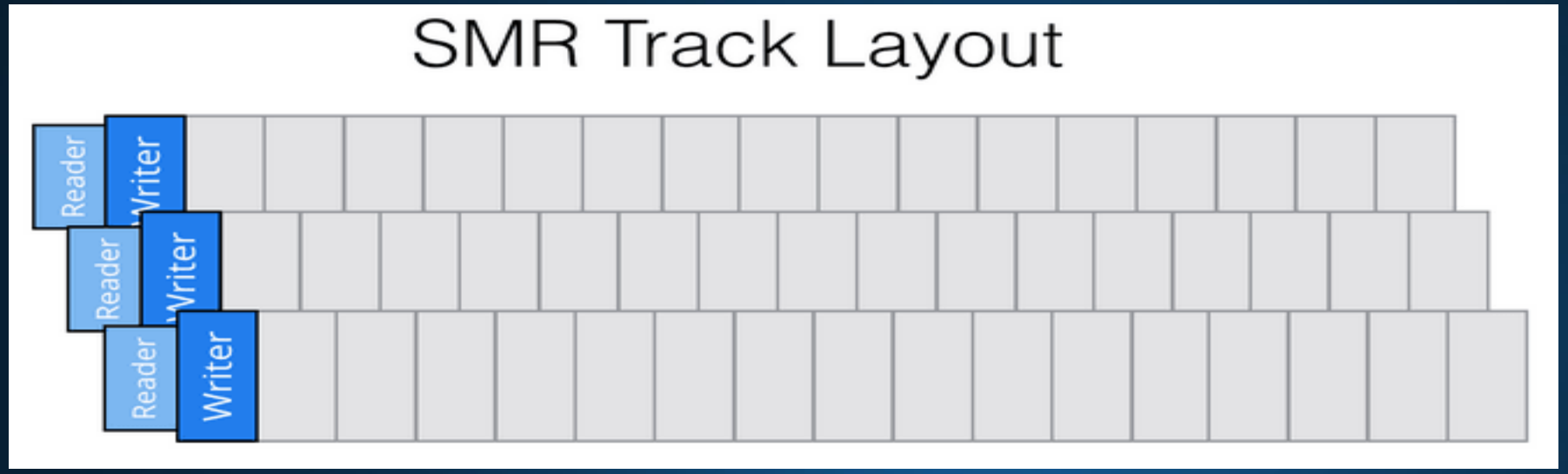
PUT



SMR

Adoption & Scaling

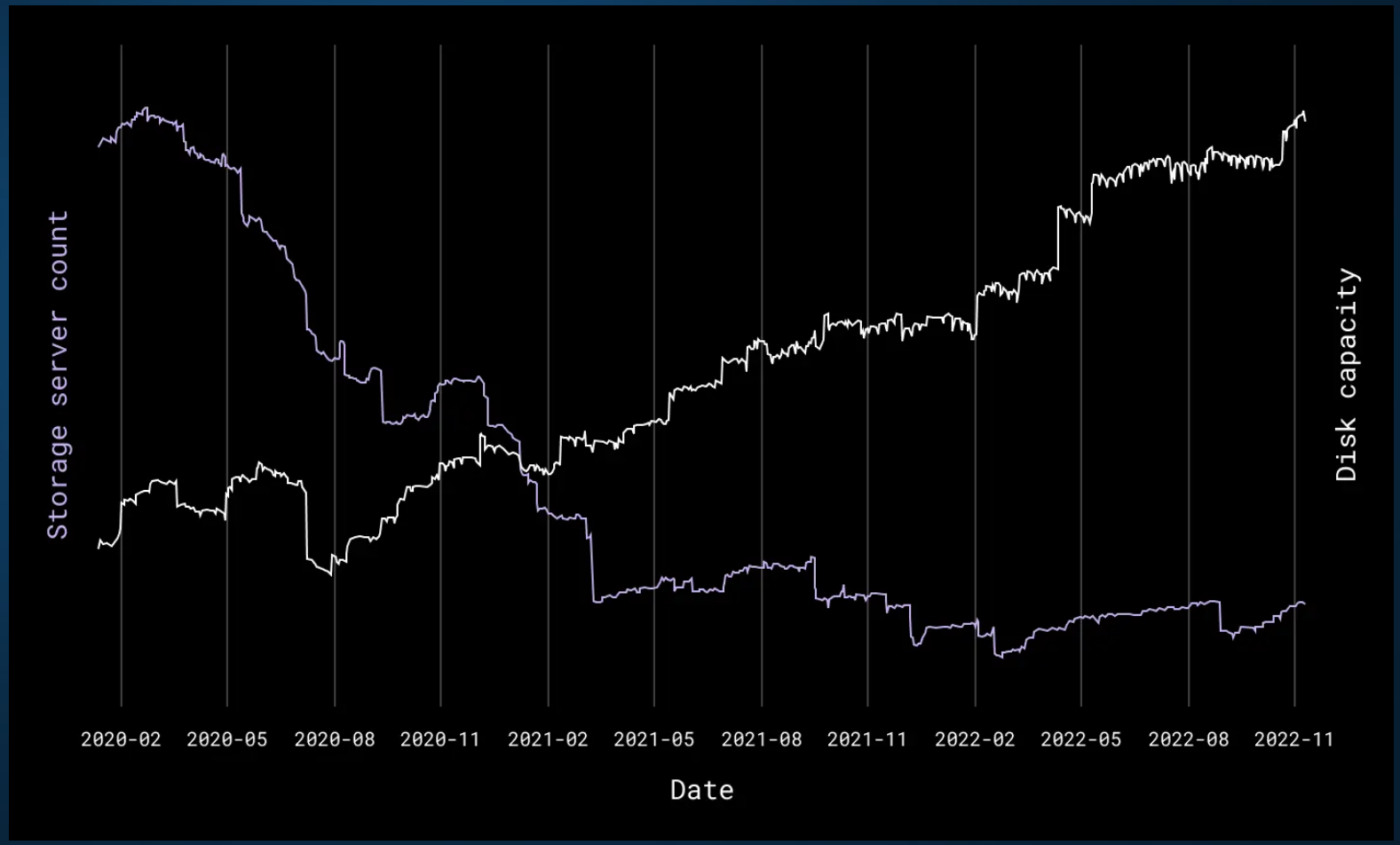
SMR V.0



- Increased Density & Cost Savings
- Go -> Rust (OSD & Volume Manager)
- Libzbc + Custom Disk Format
- SSD Cache for staging writes
- Metadata in Cconventional Zone
- 14TB Drives
- https://github.com/dropbox/pb-jelly - Rust protobuf code gen framework

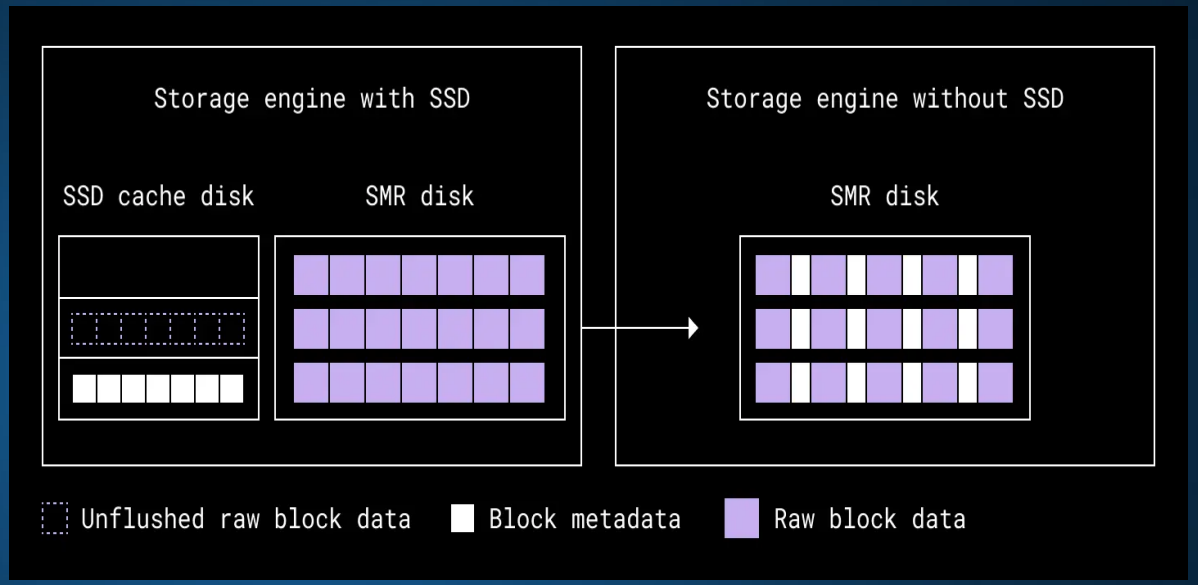
SSD Cache Bottleneck

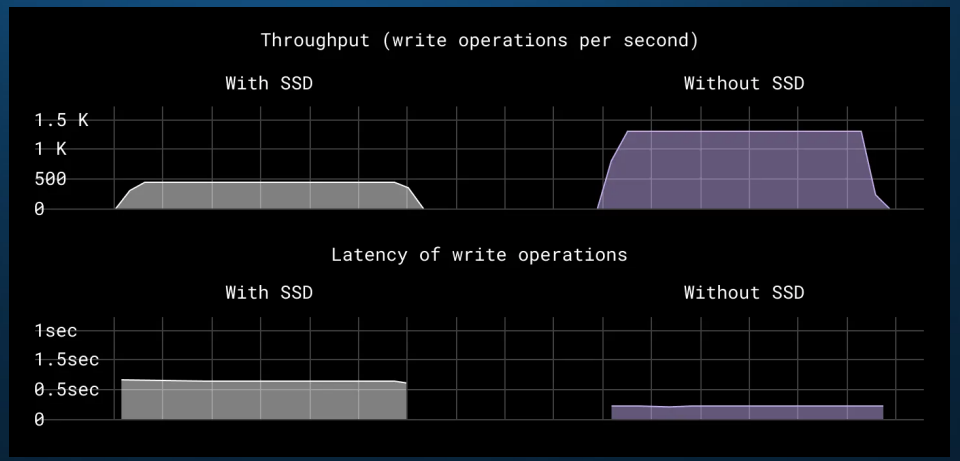
- Increased Density per Server
- Number of Servers Decreased
- Aggregated Disks Throughput started to be higher than one or two SSDs throughput per server



OSD Current Version

- Removed SSD Cache
- A simpler Software Stack
- One less hardware component in our storage servers
- Improved performance
- Our fleet is 95%+ SMR





Operations & Lessons

Background

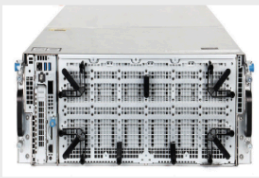
- 8-10 Engineers (SWEs & SREs)
- Interface with many teams
  - Hardware Engineering
  - Capacity Planning
  - Data Center Operations
  - Fleet orchestration
  - Security
  - Etc.,

Lessons from Operations

- Verifiers
- DRTs
- SEVs and Blameless postmortems
- Strict SLAs with Dependency teams
- Good hardware diversity
- Bias for simplicity
- “Invest in preparedness, not predication” – Nassim Taleb
- Culture & Discipline

Hardware Evolution

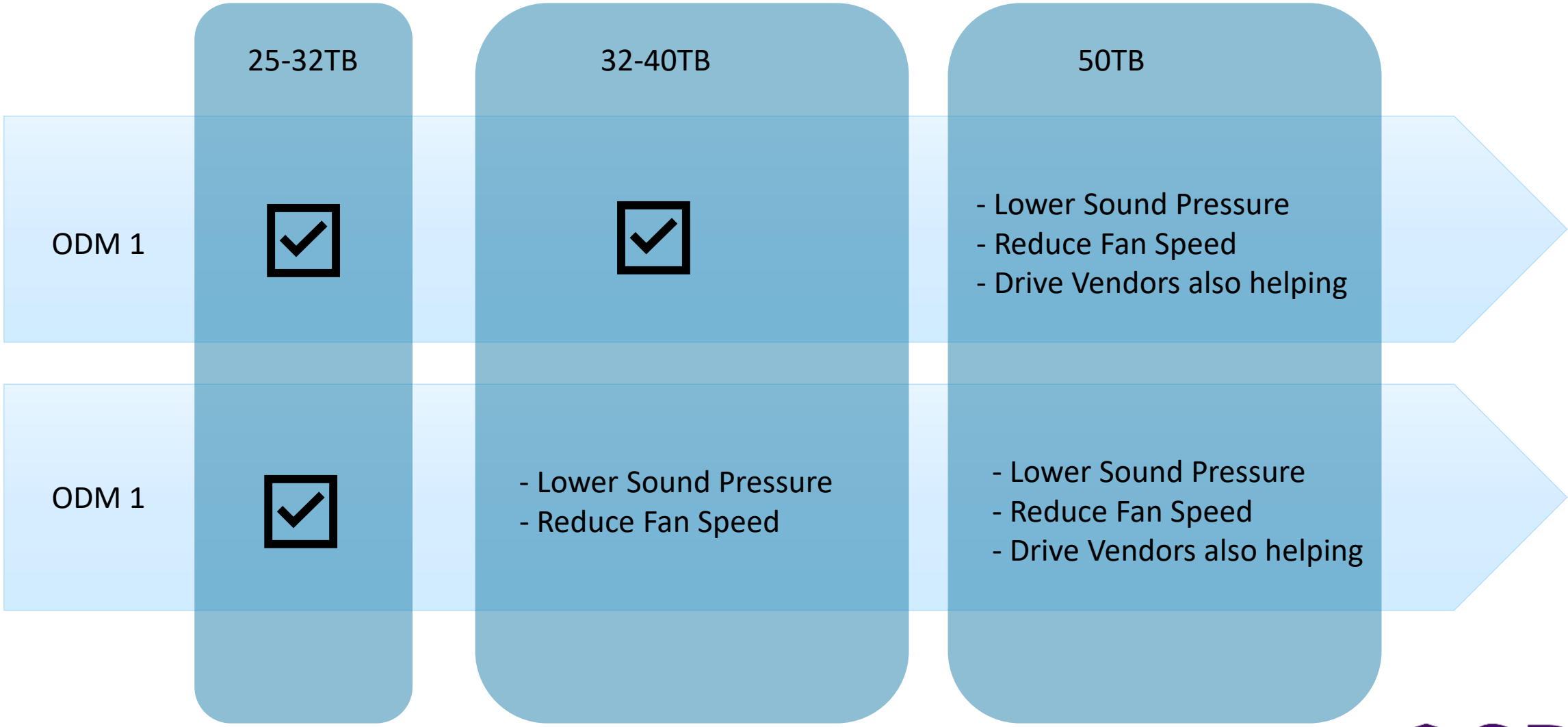
Storage Hardware Evolution

Generation 1	Generation 2	Generation 3	Generation 4	Generation 5	Generation 6	Generation 7
144 TB	192-240 TB	270 TB	624-720 TB	1.4-1.6 PB	1.8-2.6 PB	2.7 PB+
 Big Show	 Andre	 Atlas	 Axel	 Stimpy	 Scooby	 Sonic
HDD: 4TB Drive: 36	HDD: 4TB Drive: 48-60	HDD: 6-8TB Drive: 35-45	HDD: 8TB Drive: 78-90	HDD: 14 - 16TB Drive: 100-102	HDD: 18- 25TB Drive: 100-102	HDD: 28+TB Drive: 96
2013	2014	2016	2017	2018	2023	2024

Storage Platform Challenges

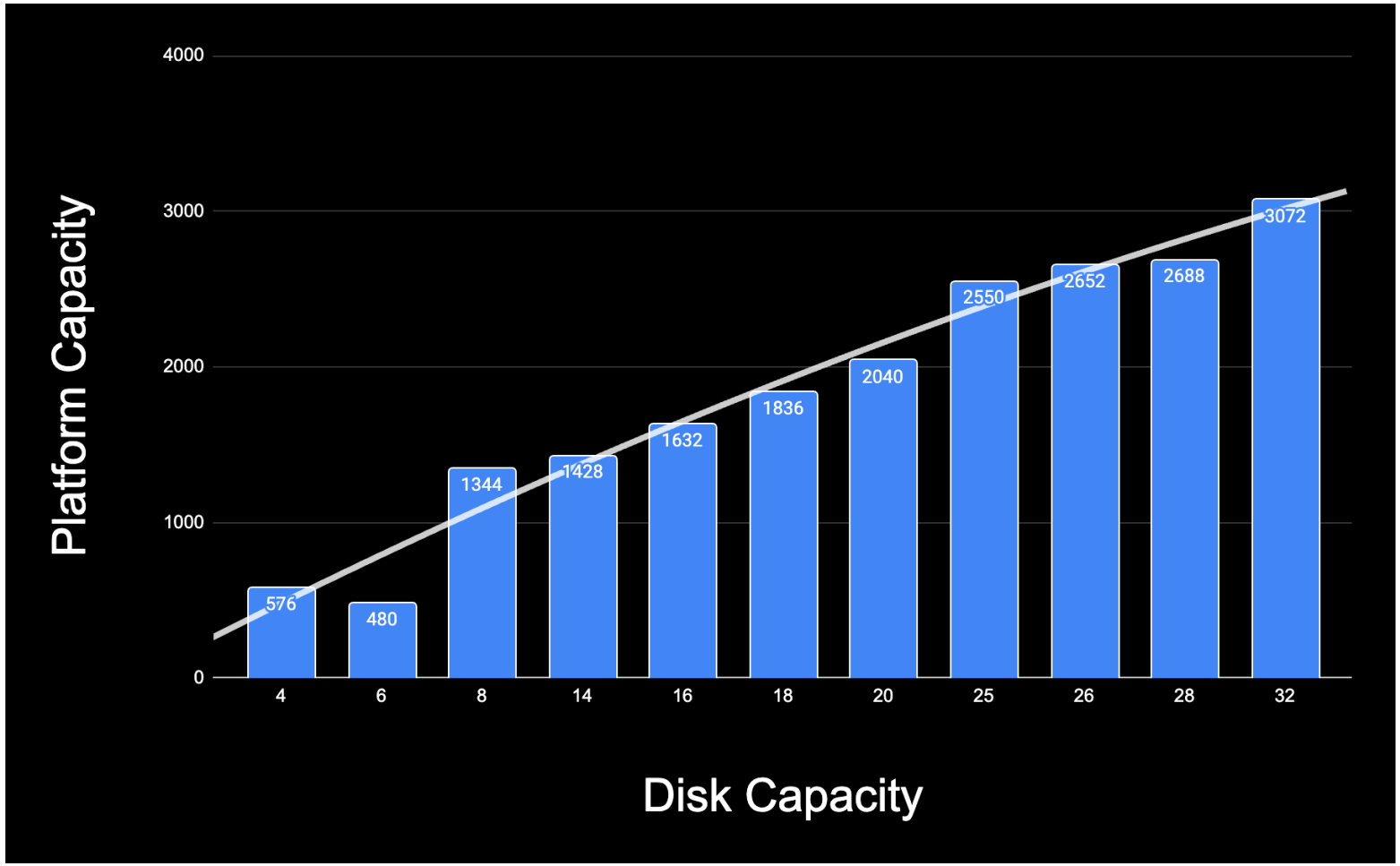
- Vibration Margin
  - I/O degradation
  - write/read failures
- Thermals
  - Impacts AFR
- Weight and Power
  - Be Aware
- SLAs
  - Low AFR
  - Fill and Drain

Storage Platform – Vibration



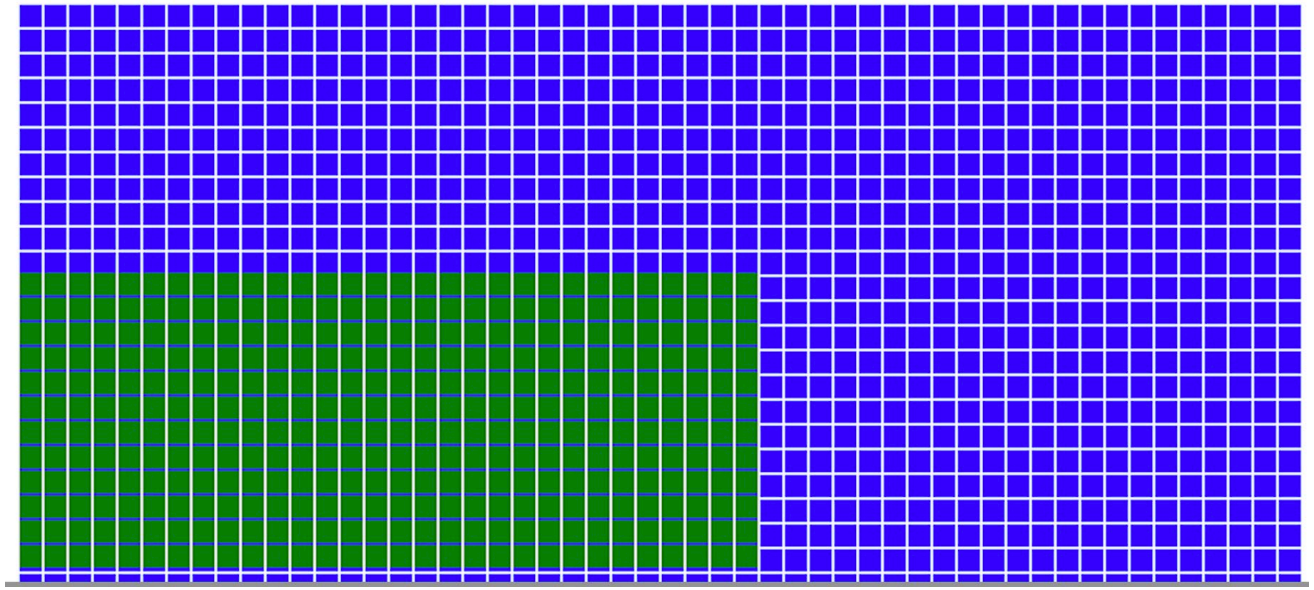
Disk Evolution

- SMR enabled at 14TB
  - Year 2018, 7y experience
  - 10-20%+ extra capacity
- 95+% of the fleet SMR
- Trends
  - Platter Count Increasing
  - Areal Density
  - SMR



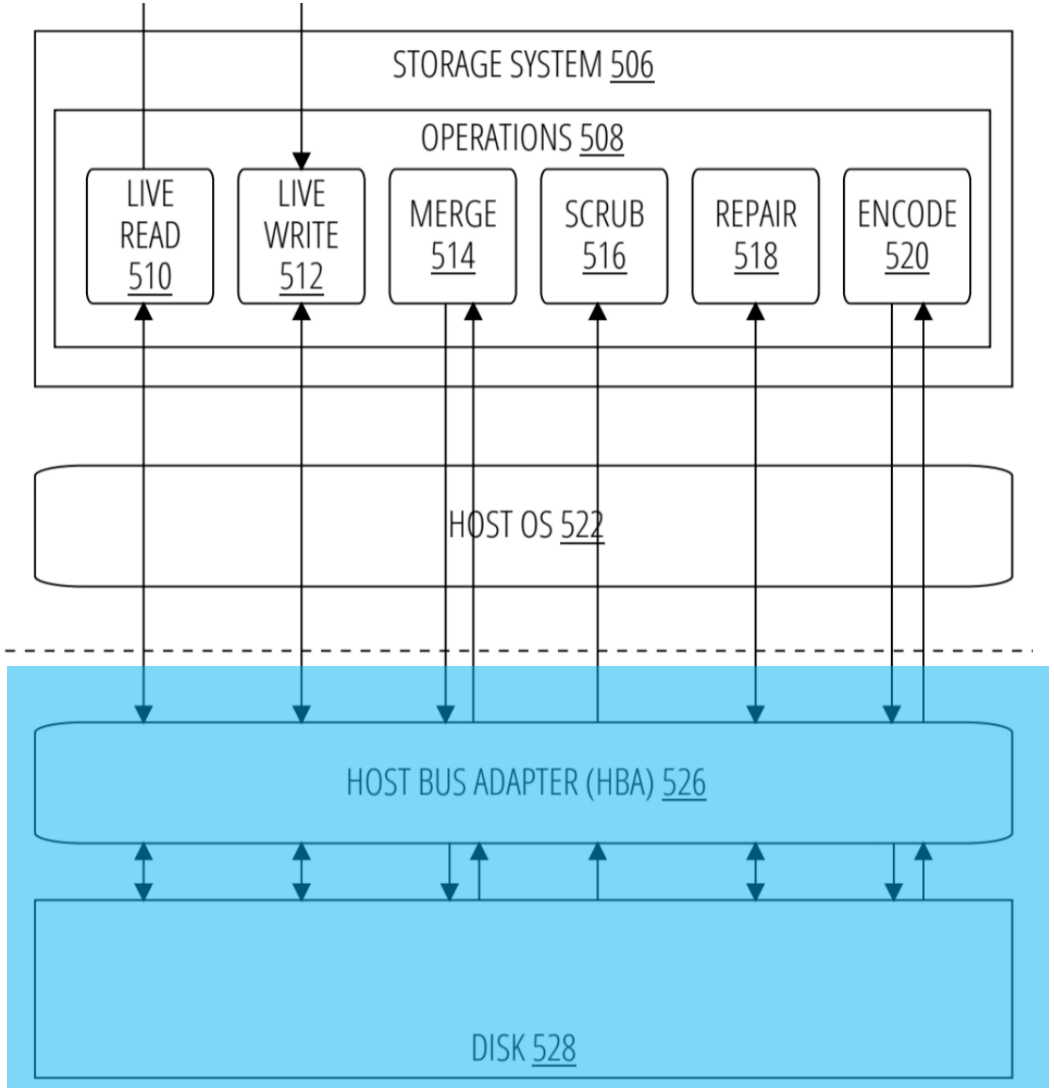
Power Impact

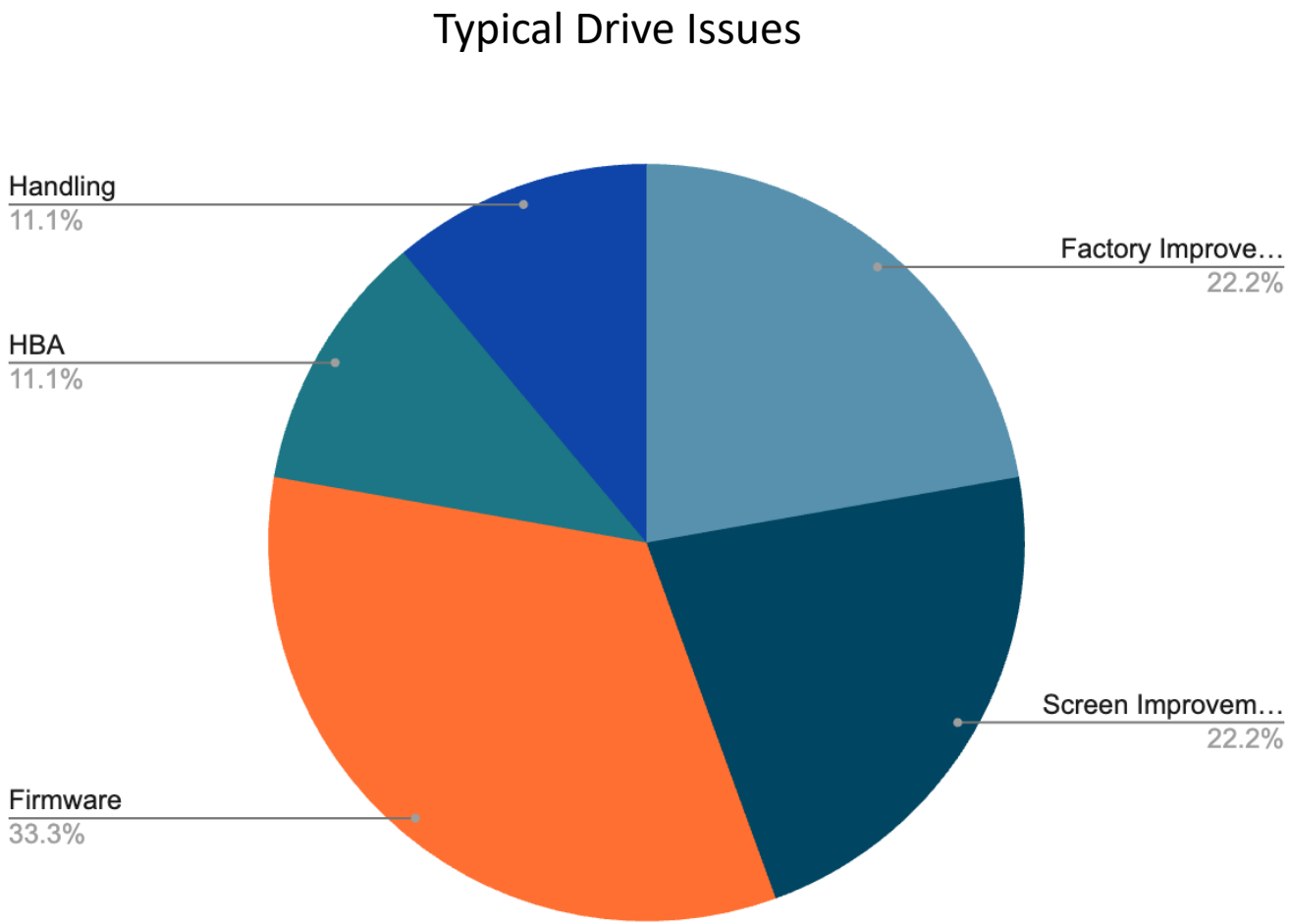
Generation	Number of racks per exabyte	p90 power (kW) per rack	Total power(kW) per exabyte
4th	1456	6.46	9408 kW
5th	642	7.6	4883 kW
6th	411	8.2	3371 kW
7th	360	10	4000 kw



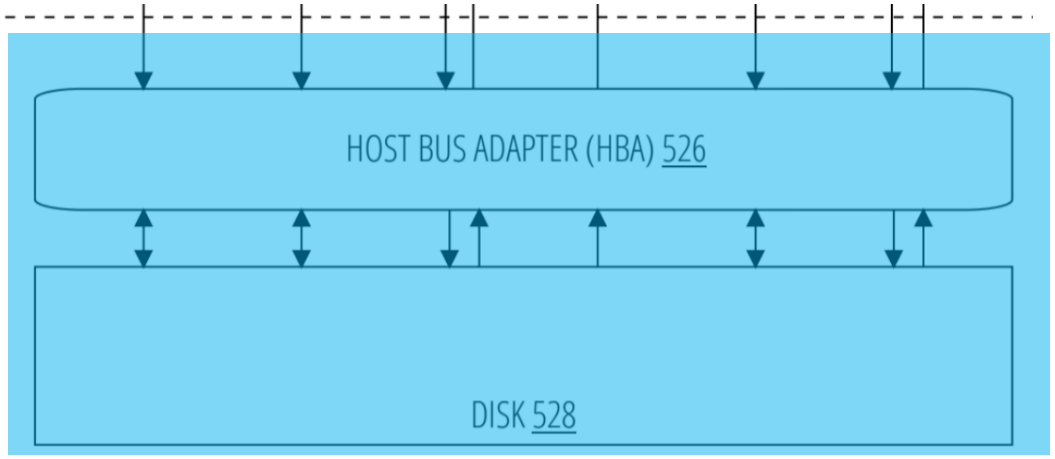
1 Exabyte for 4th gen vs 7th gen comparison
½ the Power and 75% less space

Focus Areas





Deep Collaboration



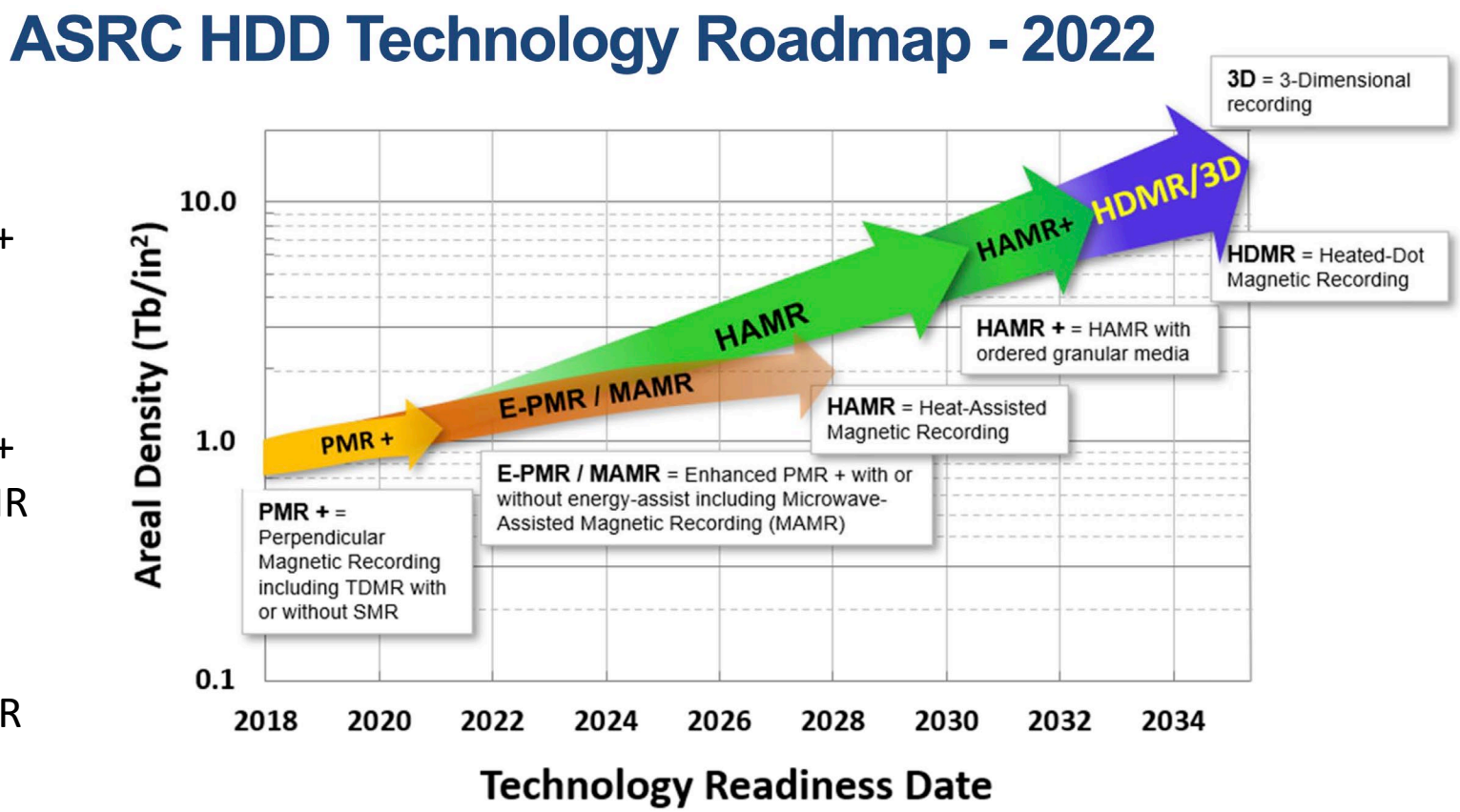
Add pic of drives/racks

- Integrate Partners
  - Partners test DBX workloads at-scale
- Magic Pocket Simulator
  - Allows for us to run safely
- At-Scale testing
  - 5,500+ drives
    - ~4,000 on site
    - ~1,500 at vendors

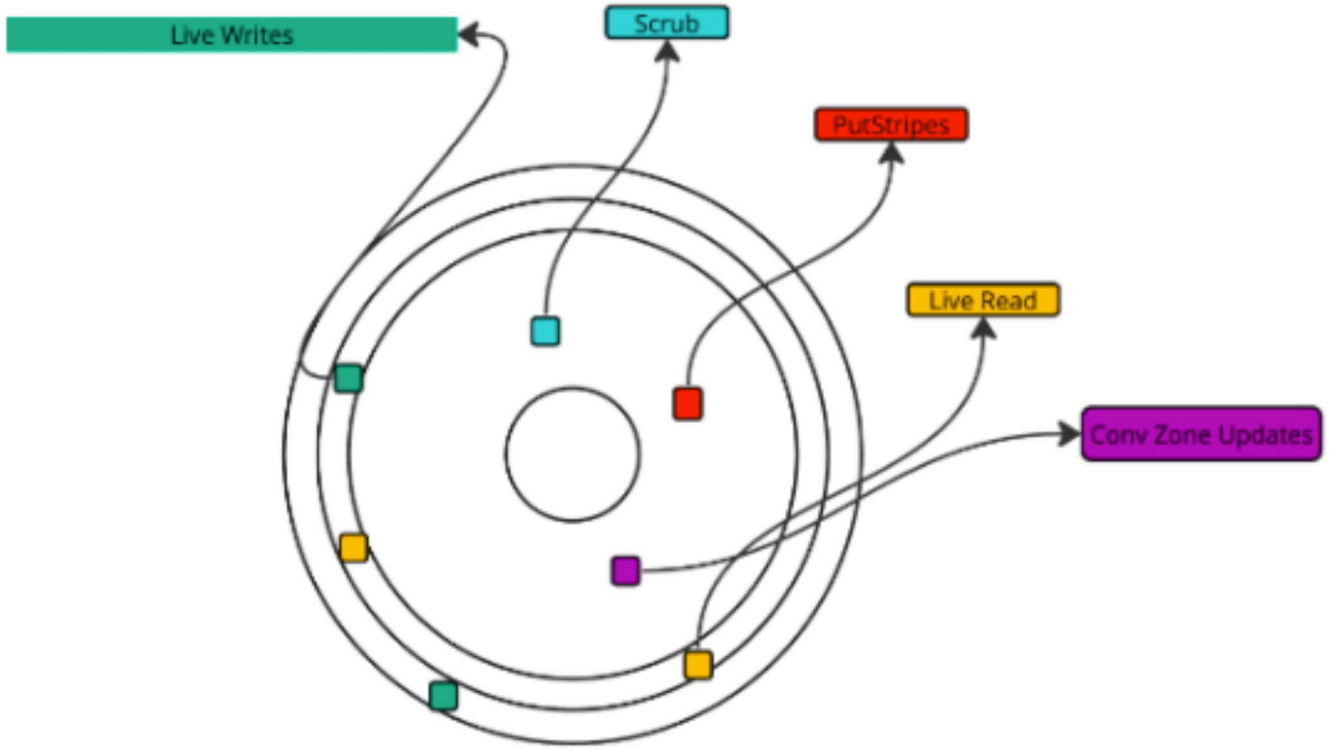
What’s Next?

Climbing the Density Trend

DBX Model	Disks	Gb/in^2	
14TB	8	1074	PMR+
18TB	9	1058	
20TB	9	1160	
25TB	10	1260	PMR+ E-PMR
26TB	10	1322	
28TB	10	1430	
30TB+	**11**	**1439***	
32TB+	**10**	**1857**	HAMR



Challenges



- Very Mixed Workload
  - Mostly random IOPS
  - Average payload is 1MiB

https://docs.google.com/presentation/d/1lexXyQj8W-PibLbK4WwAUa0zsKFLiYrdAhOW_gokFVQ/edit#slide=id.g2b9dc

Challenges

- Increasing Disk Densities
- Decreasing IOPS/TB
- Limited Software Levers

Thank you!

Please take a moment to rate this session.

Your feedback is important to us.

