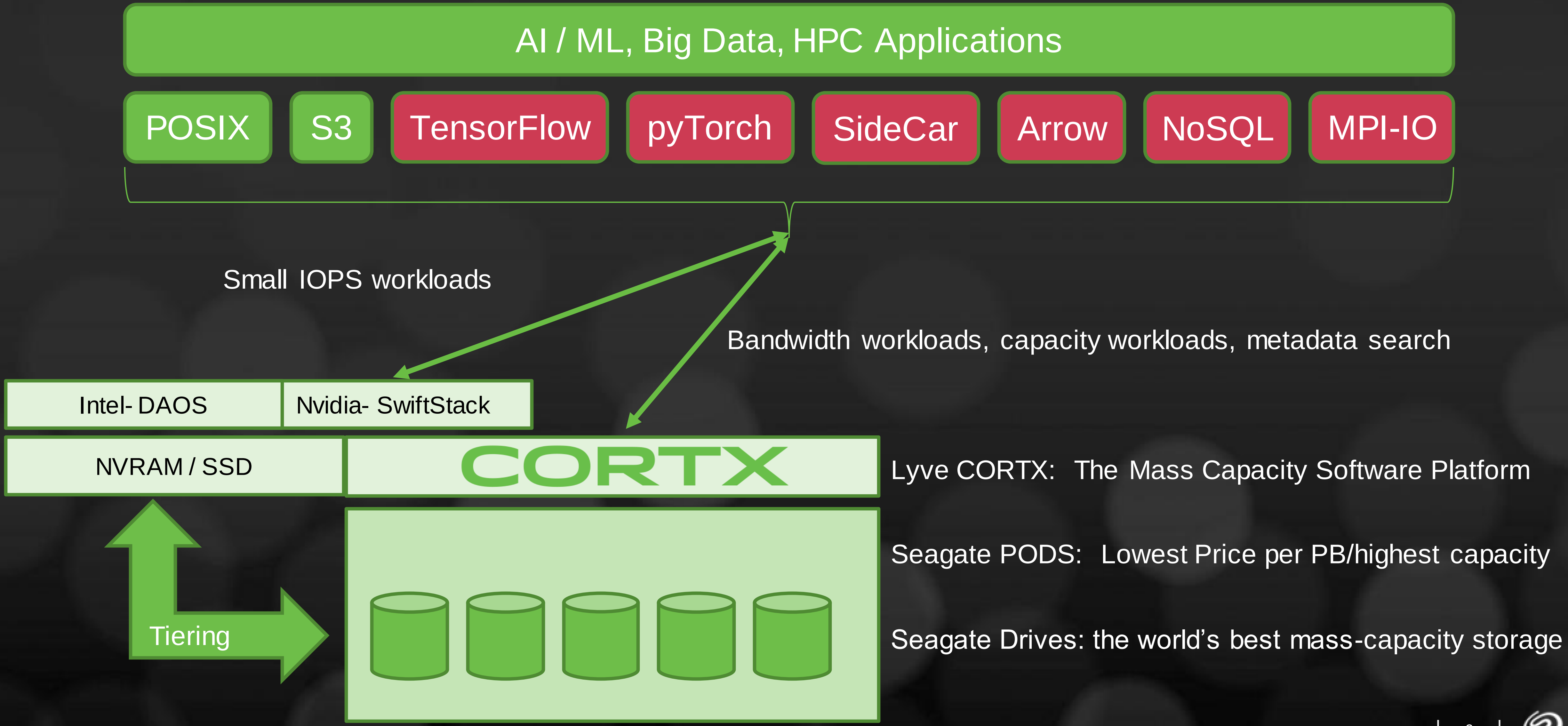
SEAGATE

CORTX

Technical Differentiation

What is the role of CORTX in the IT4.0 Ecosystem?



A Quick Introduction to PODS: Enterprise RAID at JBOD Prices

PODS is a 4U106 running ADAPT firmware

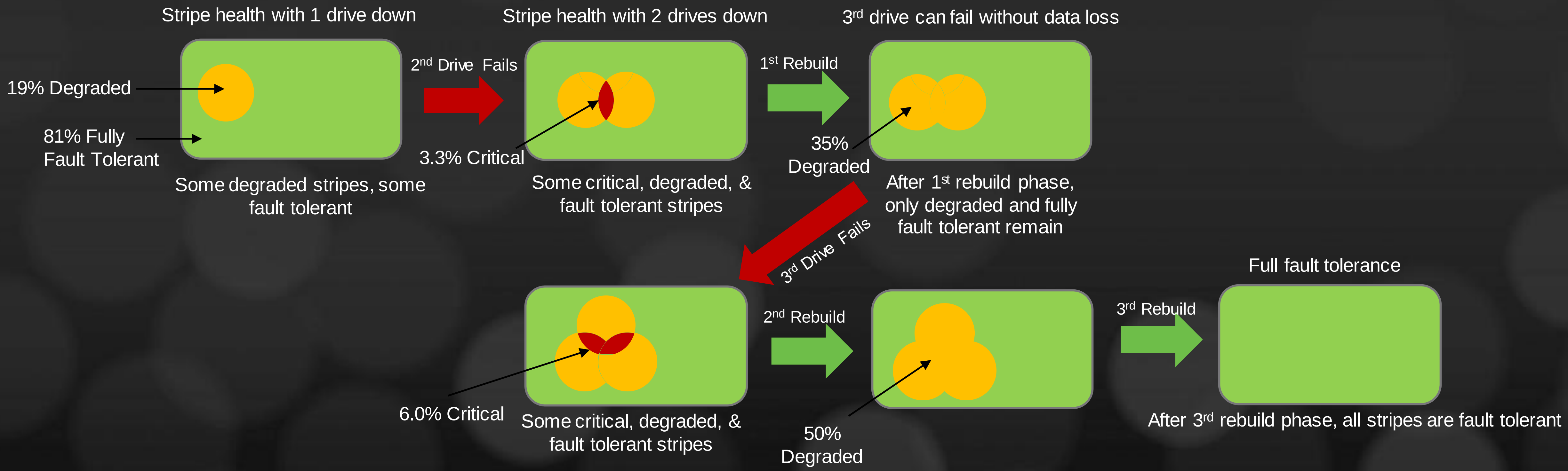
ADAPT converts JBOD into a small number of very large, very reliable disks

- E.g. A 4U106 with 16 TB drives can be exported as 2X 678TB drives

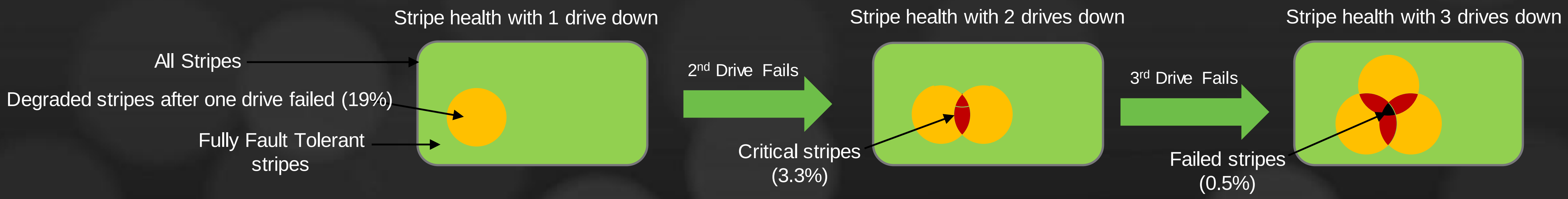
PODS/ADAPT has multiple cool features

- For our purposes here, the most interesting is declustered parity
- https://www.usenix.org/system/files/fastpw13-final14.pdf

A Quick Introduction to PODS: Successful Rebuilding



A Quick Introduction to PODS: Pathological Failure Case



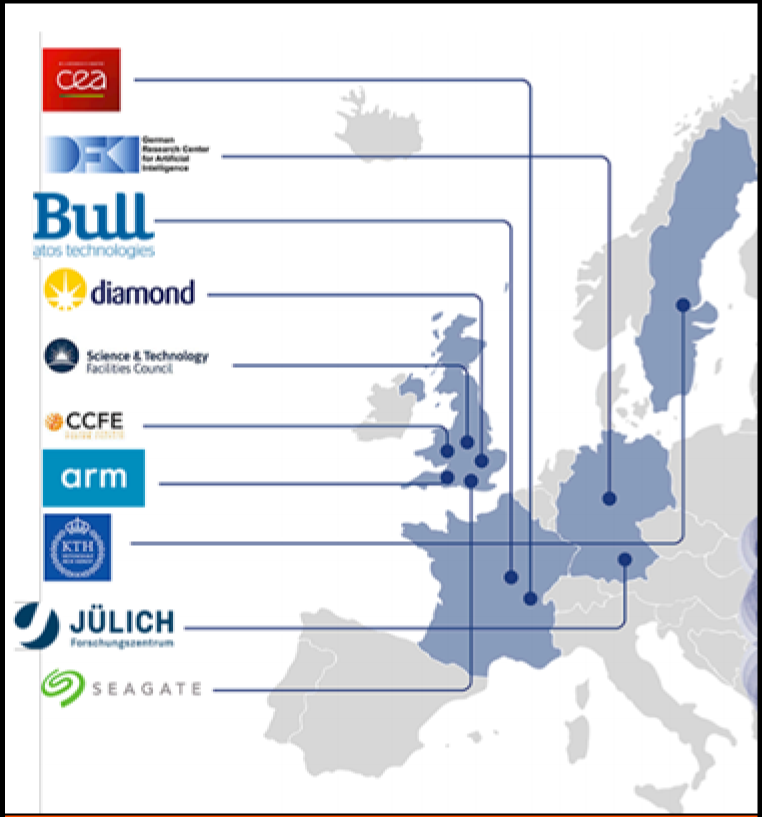
Key CORTX Differentiators

	CORTX	ActiveScale	Ceph	MinIO	OpenIO
Vertically Integrated[1]	Yes	Yes	No	No	No
Community Designed	Yes	No	No	No	No
AI/ML Friendly[2]	Yes	No	Yes	Depends[7]	Depends[7]
Tiered Parity[3]	Yes	No	Maybe	Maybe	Maybe
Lingua Franca	Yes	No	Yes	DANGER[8]	Maybe
Flexible Extensions[4]	Yes	No	No	No	No
Light Weight[5]	Yes	No	TB RAM / PB	No	No
Rich Scalable Labels[6]	Yes	No	No	No	No
Function Shipping	Yes	No	No	No	No
Open Source	Apache	Closed	LGPL	Apache	LPGL/AGPL

1. Quickest Delivery of HW Innovations
2. 4K random common pattern; EC must allow substripe read
3. Protect against all common data center failure classes
4. FDMI architecture allows core functionality to be added modularly
5. Minimum memory footprint, offload EC to Yak ASIC
6. Convert block into rich (scalable) data experience
7. MinIO exports another system; that system determines this
8. Is possible but uncoordinated; data inconsistency can arise

Data and metadata paths designed for HPC (by HPC)

- Exabyte capacity with exascale performance
- Scale-out metadata and integrated user labels
- Peer-to-peer server architecture
- Concurrency without inconsistency
- Processor Agnostic Design
- Machine-based log analysis on highly-structured log records



Core partners in original
EU funded CORTX initiative

Data and metadata paths designed for HPC (by HPC)

Exabyte capacity with exascale performance

- Data structures use very wide variables (e.g. OID is 120 bits)
- High-performance client - server data path (e.g. RDMA and 0-copy)

Scale-out metadata and integrated user labels

- Minimal global metadata (no serialization bottlenecks)
- Highly distributed KVs implemented with streaming b-trees
- Containerized metadata: billions of objects, one common metadata

Peer-to-peer server architecture

- All instances can be clients, data servers *and* metadata servers

Concurrency without inconsistency

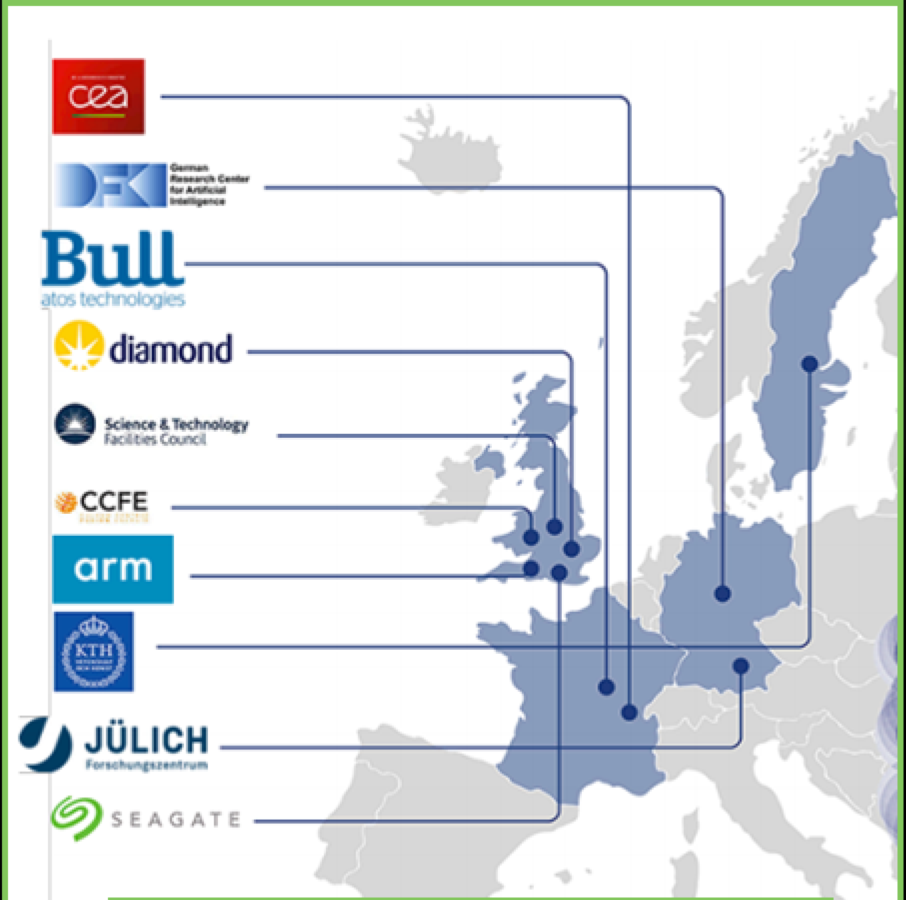
- Redirect-on-write, built-in transactional lockless versioning
- Consistent non-locking concurrent transactional modifications

Processor Agnostic Design

- x86, ARM, RISC-V, OpenTitan, European Processor

Machine-based log analysis on highly-structured log records

- Humans don't scale to exascale



Core partners in original EU funded CORTX initiative

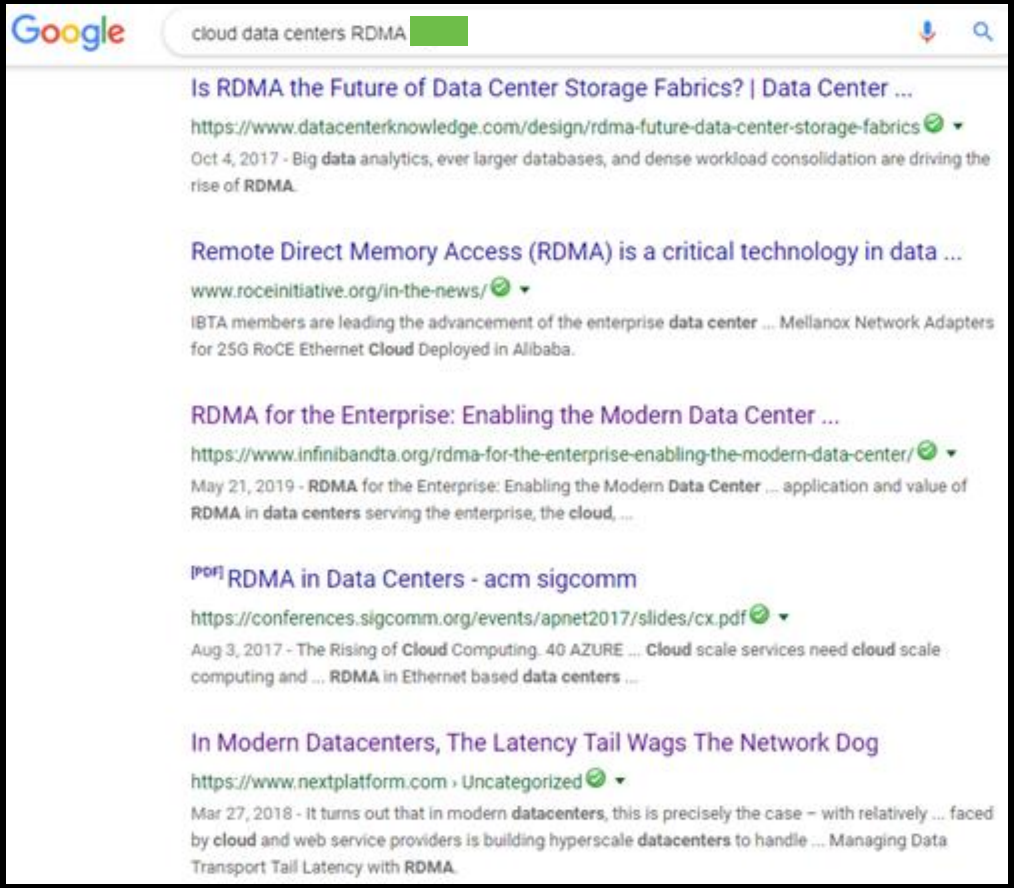
We were ~~lucky~~ smart

Designing for HPC *yesterday* was the prophetic move to better create data center solutions *today*

Storage for Data Centers Before ML	Storage for Data Centers After ML
Diminishing monetization for storing more data	Store everything; insights are accessible
Map-Reduce / Hadoop were compute models	AI/ML and simulation increasingly dominant
Multi-core architectures, embarrassingly parallel	GPU architectures, tightly coupled concurrency
Demands on object store were minimal	Gartner: "Extreme Throughput at Low Latency"
Medium capacity, low performance, get-put	High capacity, high performance, 4K random IO

A GPU is a powerful consumer of parallel streams of data . . .

Designed for HPC: Extreme Throughput at Low Latency

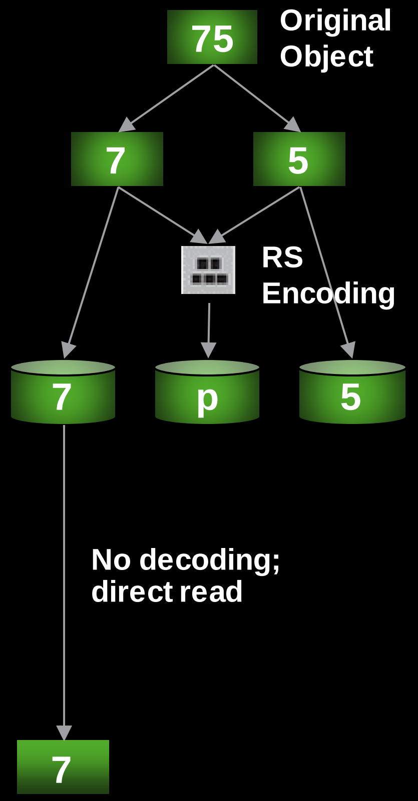


Prime Example

CORTX had RDMA from day one.

Designed for HPC: Small Random Read

CORTX

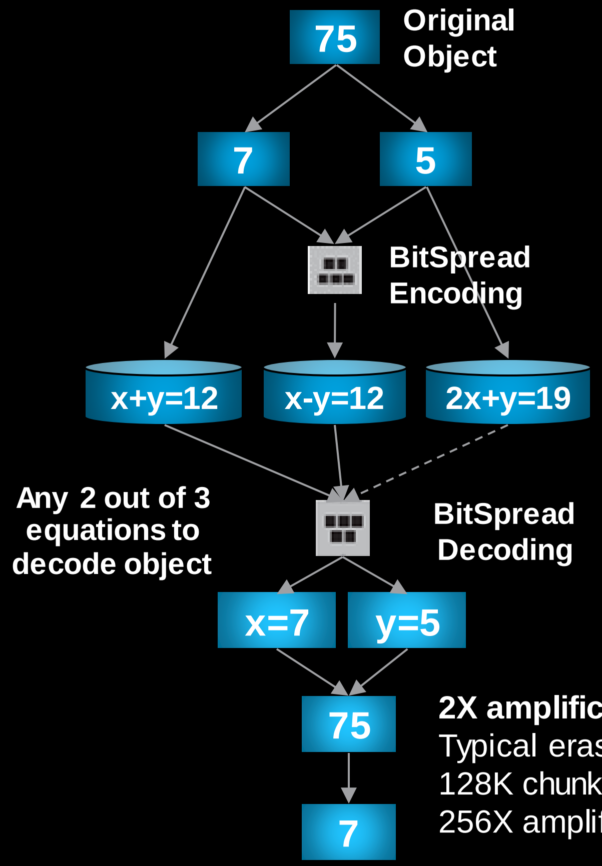


No Read Amplification!

Save an object with erasure 2+1

Partial read of that object

ACTIVESCALE MODULAR OBJECT



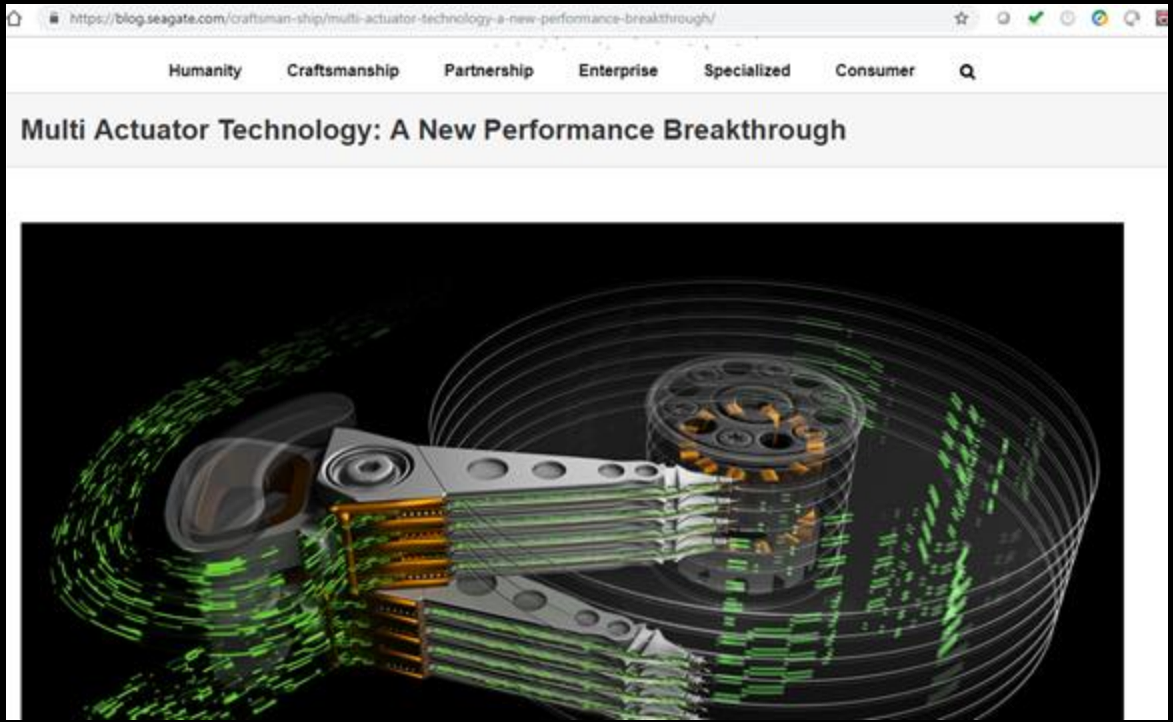
2X amplification as shown. Typical erasure of 8+2 with 128K chunks will result in 256X amplification!

Designed for HPC: Scale-out KV Search

- Renovo
  - “Unprotected left turns in the rain”
- Tesla
  - “Yellow-shirted pedestrians crossing L-R”
- Admins
  - “Files larger than a GB and older than a month”
- MinIO
  - ”Find a blue coffee mug”
- Lyve Pilot Orchestration
  - “Objects accessed by Tom in last 48 hours”

No Value to Save the DataSphere if You Can't Search the DataSphere

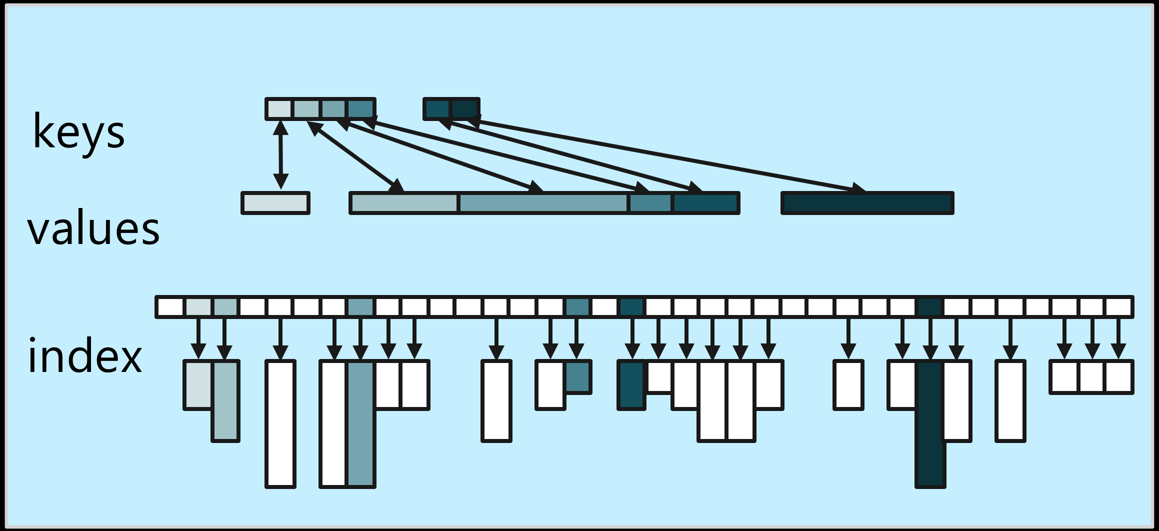
EOS and PODS combine faster hardware with smarter software.


https://blog.seagate.com/craftsman-ship/multi-actuator-technology-a-new-performance-breakthrough/

Humanity Craftsmanship Partnership Enterprise Specialized Consumer

Multi Actuator Technology: A New Performance Breakthrough

Solution One: Faster Hardware



Solution Two: Smarter Software

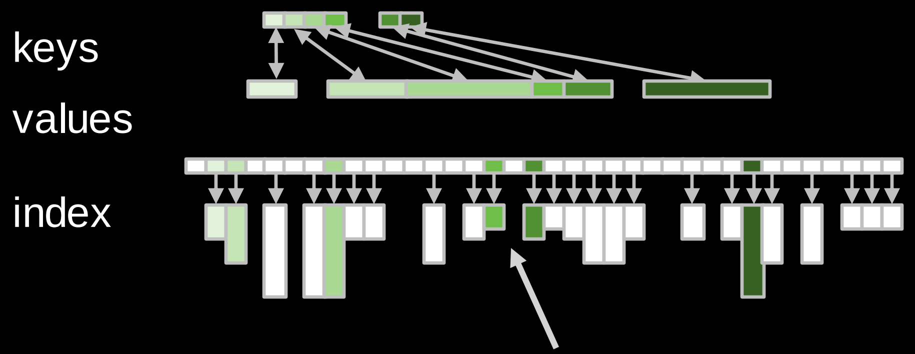
CORTX Key-Value Store

Distributed, in-memory, transactional key-value subsystem

- Unstructured data hard to monetize
  - Too much to be efficiently searched
- Labeled data can be monetized
  - Now the world stores everything
- Will the labels themselves grow too large?
  - Not for CORTX due to scale-out indices
- Integrated KVS (not bolt-on ElasticSearch)
  - Never inconsistent
  - Single system to monitor, scale, provision,etc
  - Single “scrub”

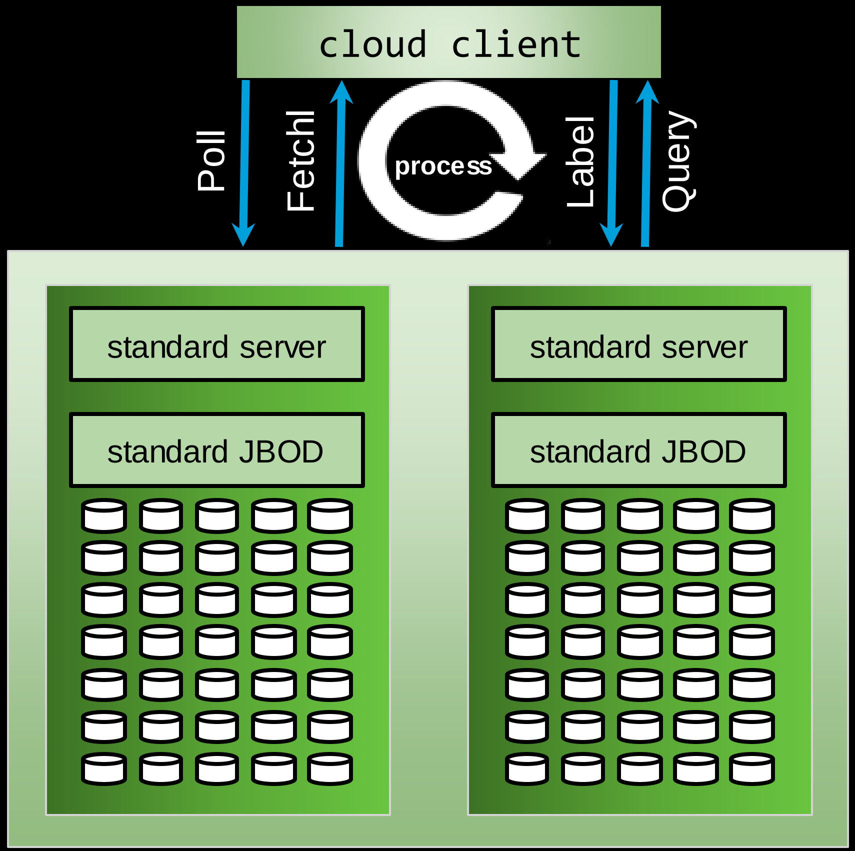
batch operations minimize latency for big data queries

clovis_lookup(op, index, key_vec, val_vec)
clovis_insert(op, index, tx, key_vec, val_vec)
clovis_next(op, index, key_vec, val_vec)

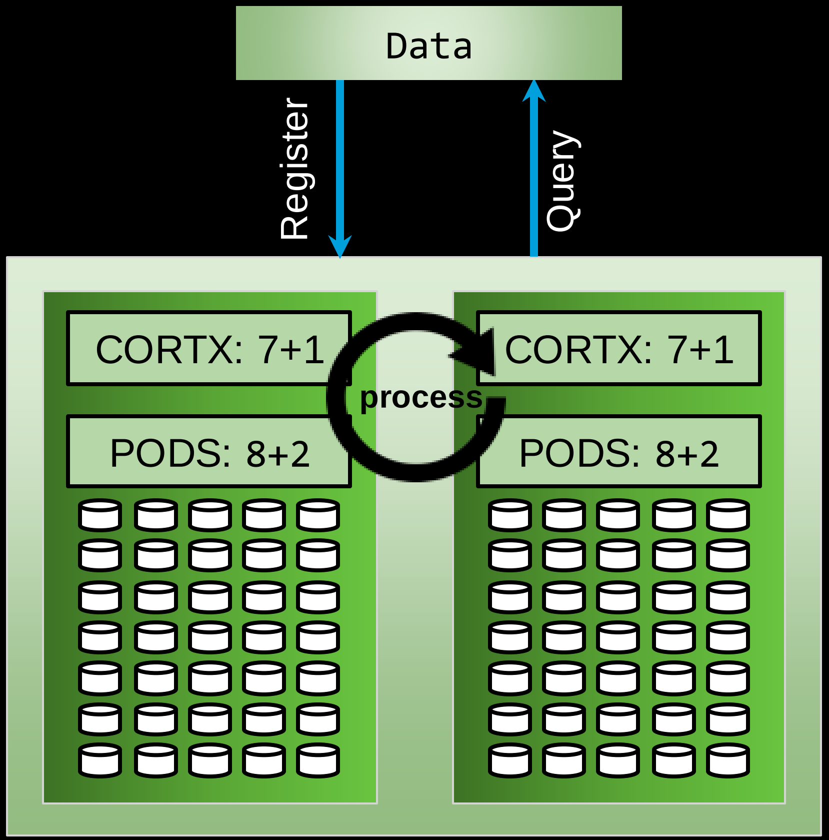


distributed index speeds access to target data

Designed for AI/ML: Automated Label Capture



Yesterday



Tomorrow

Extreme Scale Requires Extreme Protection

Availability in Globally Distributed Storage Systems

Daniel Ford, François Labelle, Florentina I. Popovici, Murray Stokely, Van-Anh Truong*
Luiz Barroso, Carrie Grimes, and Sean Quinlan
{ford,flab,florentina,mstokely}@google.com, vatruong@ieor.columbia.edu
{luiz,cgrimes,sean}@google.com
Google, Inc.

SPATIAL FAILURE BURST
Multiple simultaneous drive failures within a single rack.
Protect against these with erasure across enclosures.
Parity within enclosure is insufficient.

ASPATIAL FAILURE BURST
Multiple simultaneous drive failures across multiple racks.
Protect against these with erasure within enclosures.
Erasure across enclosures is insufficient..

No single tier of parity can protect against all these failures. Google knows about this and presumably had a team of PhD's implement a solution. Private cloud needs our help to solve this; we can do so with PODS & tiered erasure.

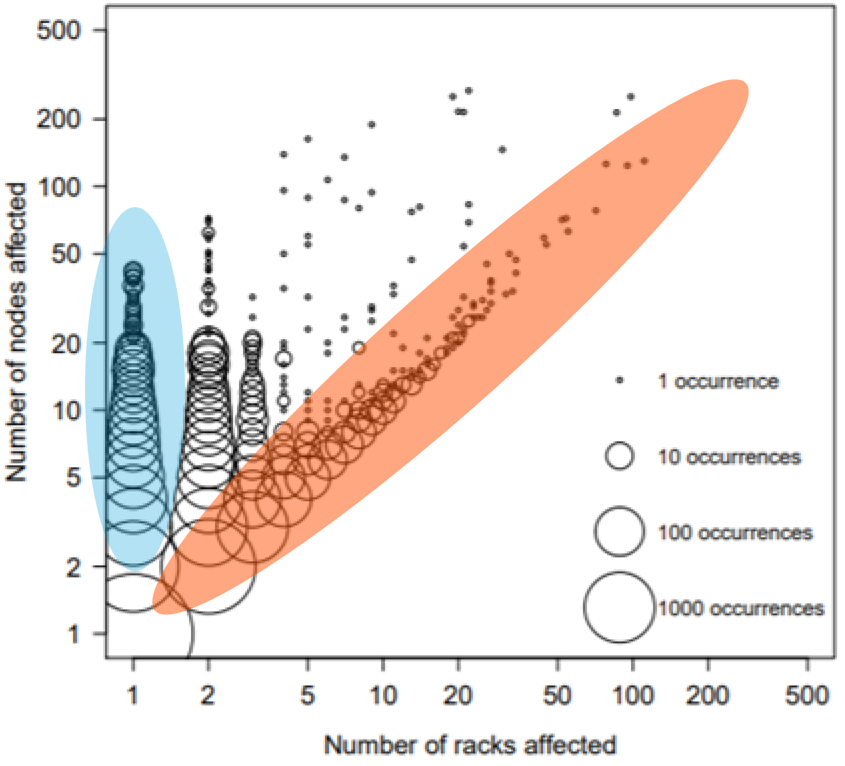
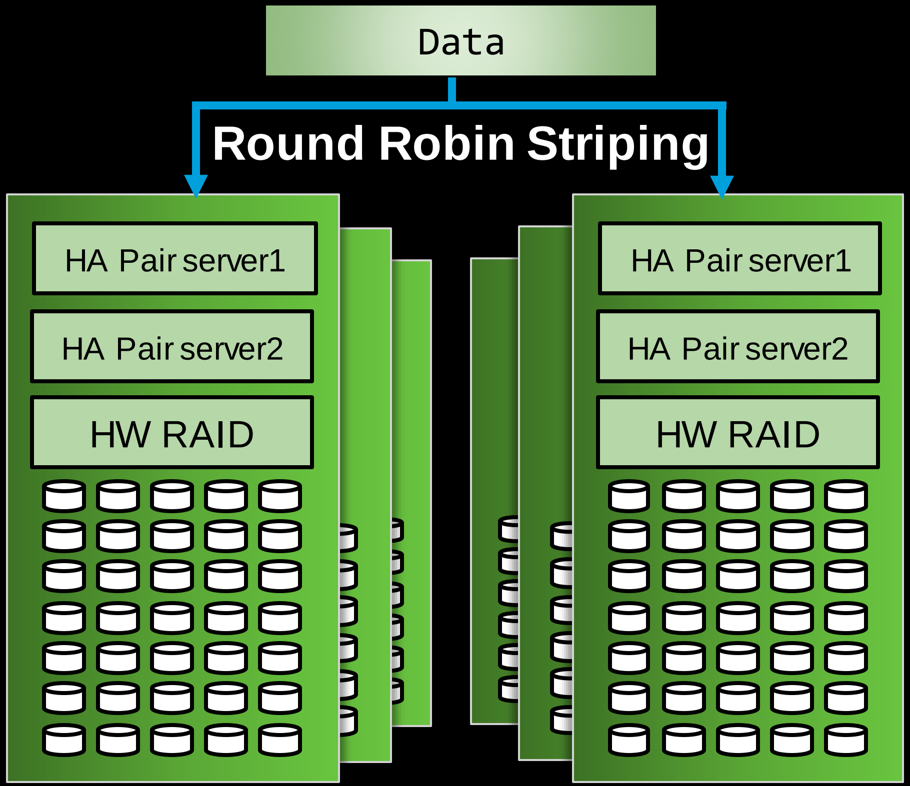


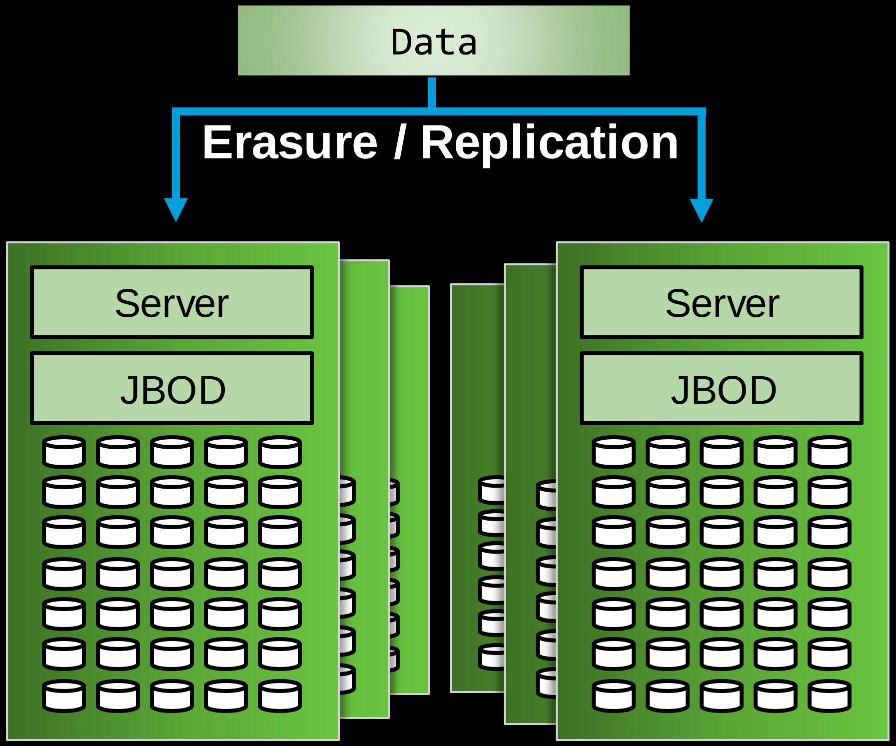
Figure 8: Frequency of failure bursts sorted by racks and nodes affected.

Two Existing Approaches for Data Durability and Availability



Hardware Reliability

Lustre, PVFS, BeeGFS, etc.



Software Reliability

Cloudian, ActiveScale, SwiftStack, Ceph, HDFS, etc.

Two Existing Approaches for Data Durability and Availability

Data

Round Robin Striping

HA Pair server1 HA Pair server1

DURABILITY PROBLEM:
- Cannot protect against Aspatial Failure Bursts

AVAILABILITY PROBLEM:
- Cannot protect against HA server pair failure

COST PROBLEM:
- Hardware RAID is too expensive

Hardware Reliability

Lustre, PVFS, BeeGFS, etc.

Data

Erasure / Replication

DURABILITY PROBLEM:
- Cannot protect against Spatial Failure Bursts

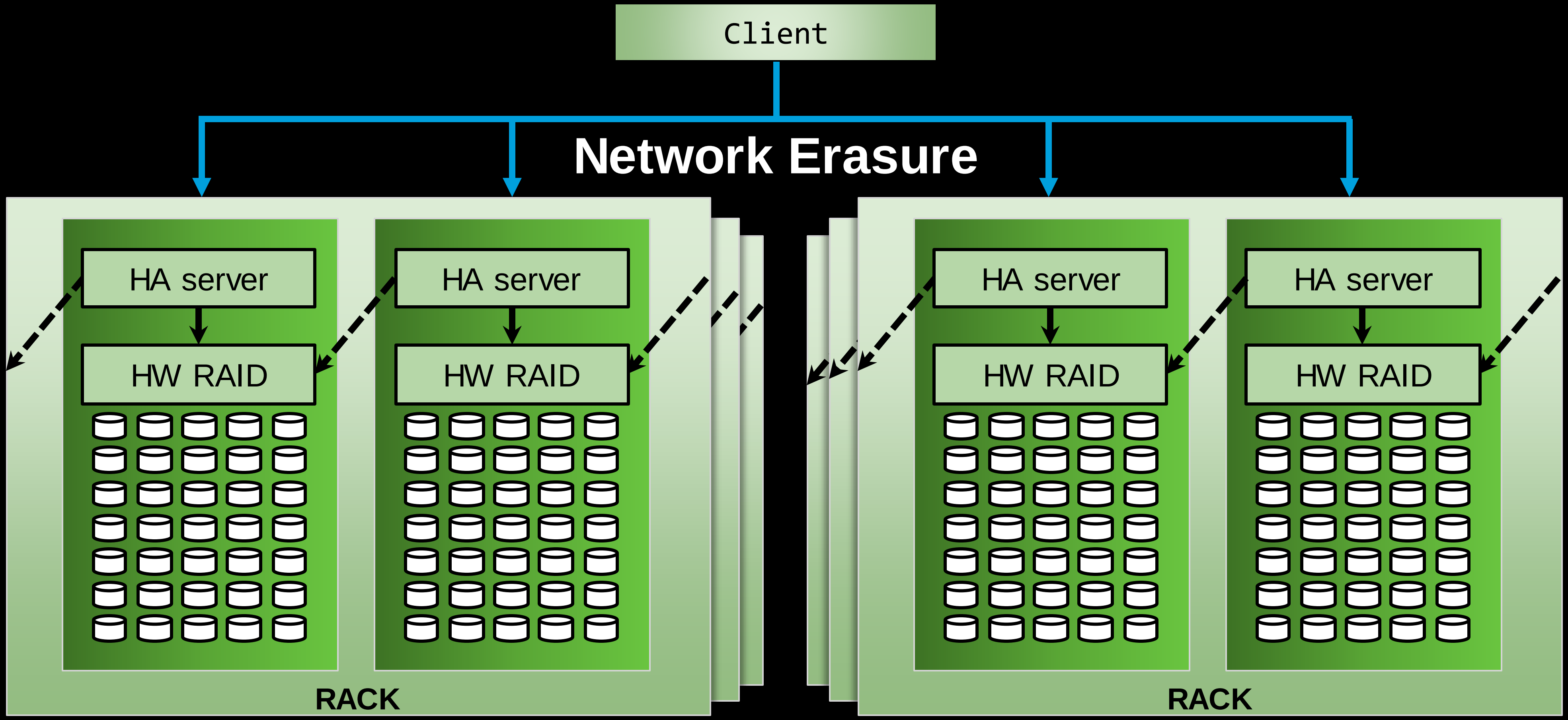
AVAILABILITY PROBLEM:
- Server failure is handled but ‘blast radius’ is concern
- Typical configs recommend 2U12 – highly inefficient

COST PROBLEM:
- Public cloud too expensive
- DIY too dangerous
- On-prem too expensive

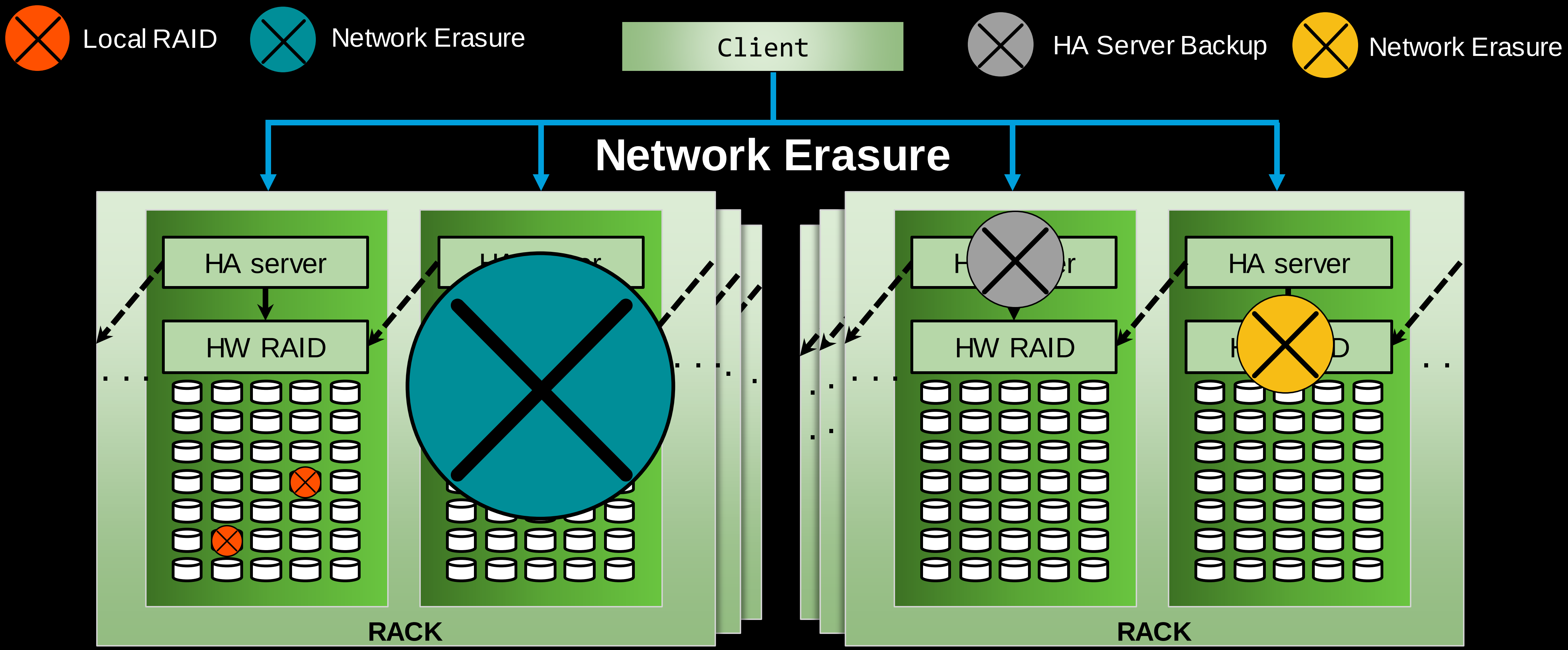
Software Reliability

Cloudian, ActiveScale, SwiftStack, Ceph, HDFS, etc.

CORTX Hybrid Approach for Durability and Availability



CORTX Hybrid Approach
for Durability and Availability



CORTX Hybrid Approach Maximizes Durability, Availability, and Density

Disk Trends – Rebuild Has Been Getting Hard for Many Years

Disk capacity is growing much more quickly than disk performance
Time to rebuild a drive in RAID set is growing quickly

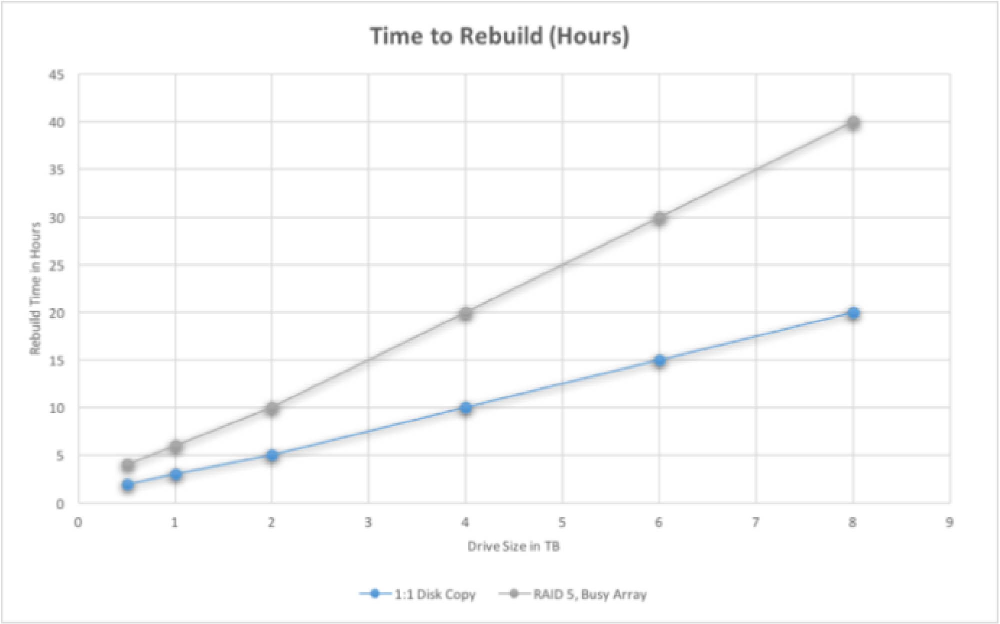


Figure 1: Disk rebuild times, given 115 MBps throughput

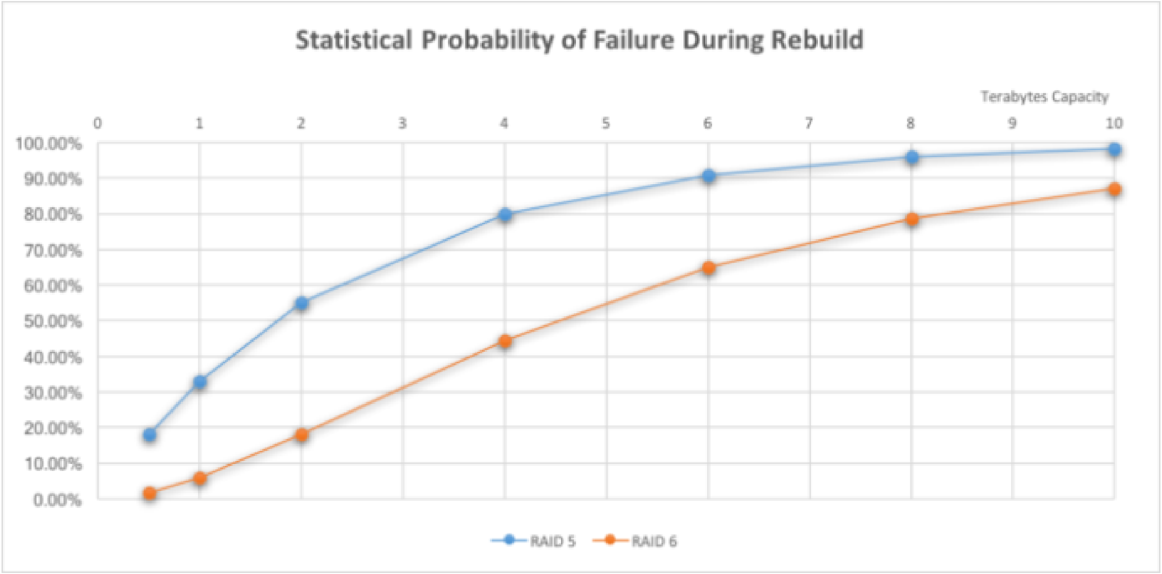
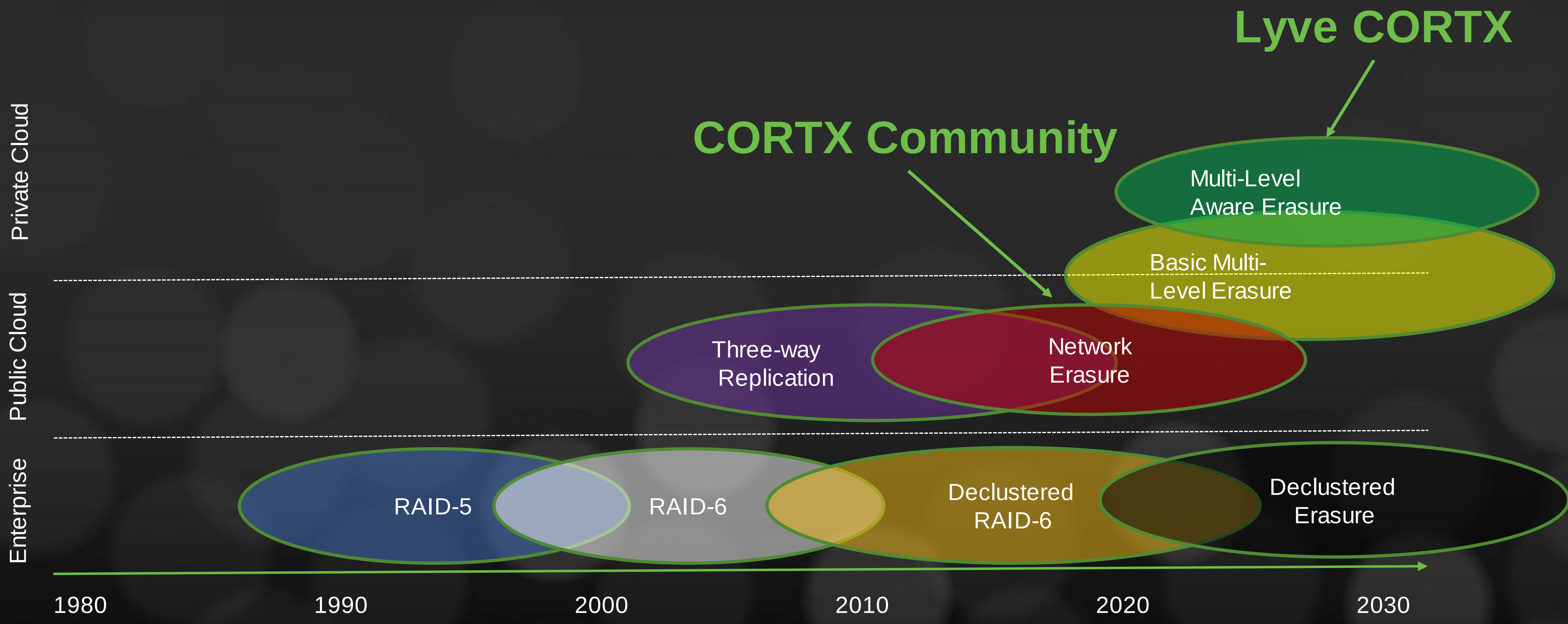


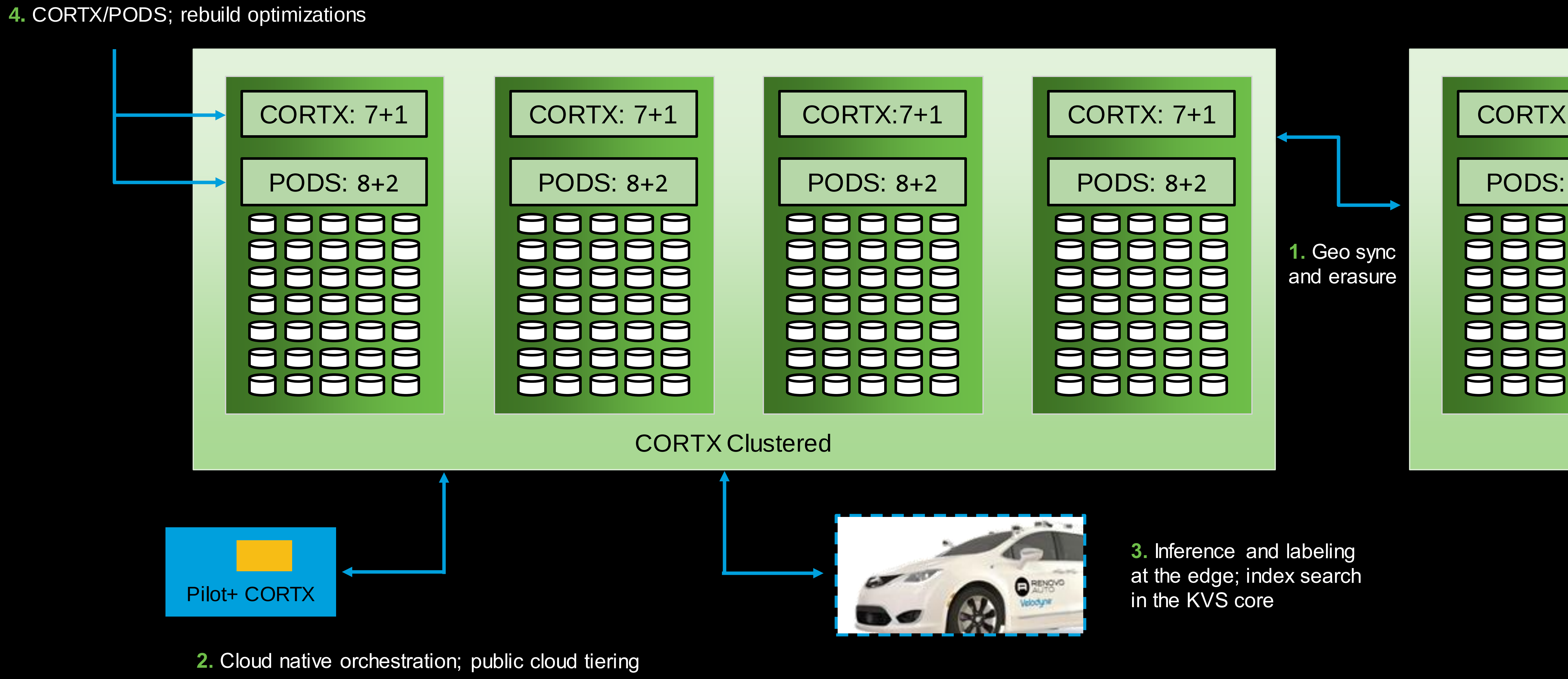
Figure 2 – In a 6 drive set, with a manufacturers stated error rate of 1.0E-14

RAID5 and RAID6 now almost deterministic to encounter additional failure during rebuild and therefore lose data. New methods needed.

Dominant Data Reliability Mechanisms



CORTX: Four Key Initiatives



The Power of Codesign

Cooperative Optimizations in Multi-Level Erasure Systems

Ceph + PODS <<< CORTX + PODS >>> CORTX + JBOD

Improved efficiencies by breaking the standard server-disk API.

- **Use YAK ASIC for both levels of erasure**

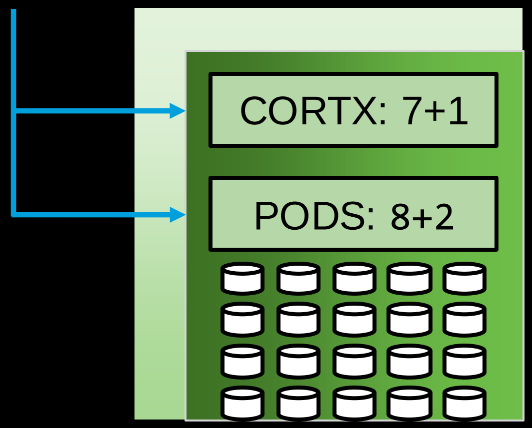
  CORTX: “Hey PODS, compute this for me please.”
- **Only rebuild live data on device failure**

  PODS: “Hey CORTX, which blocks are live?”
- **Temporarily boost local parity when network lost**

  CORTX: “Hey PODS, amplify from 8+2 to 8+3 please”
- **Retrieve lost PODS data over network**

  PODS: “Hey CORTX, I lost some block ranges. Restore please”

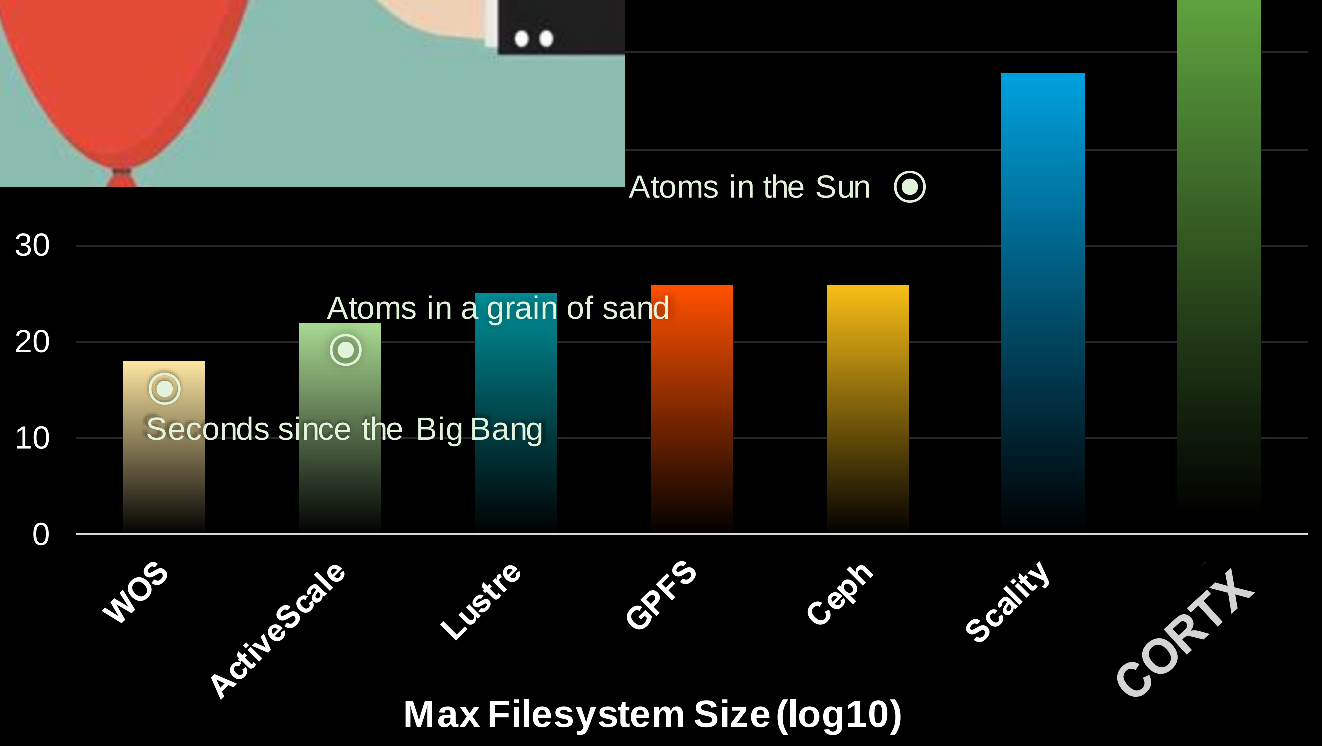
4. CORTX/PODS; rebuild optimizations



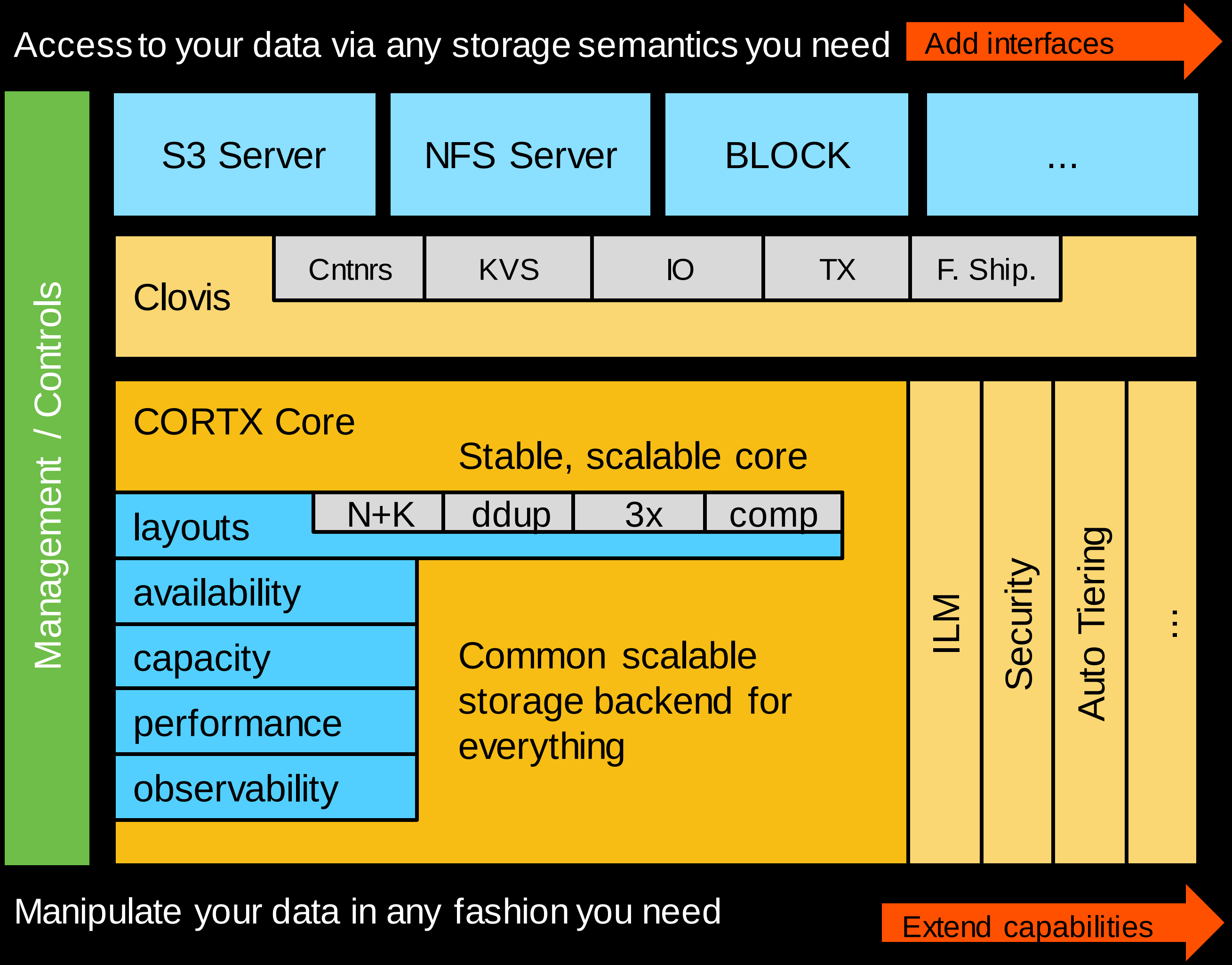
Designed for IT4.0: Massively Scalable

CORTX capacity limits:

- Billion billion billion billion billion exabytes (2^20
- 1.3 billion billion billion billion (2^120) objects
- Unlimited Object sizes



Lyve CORTX Extensibility



E.g. pNFS and Apache Flink prototypes added by community. TensorFlow integration added in hackathon

E.g. HSM added by community

For IT4.0’s zettabyte data growth needs, CORTX enables customers to

CORTX

Scale without limits

Scale without pain

Scale without performance loss

This is unlike options available today, because of the lowest cost per byte economics delivered by Seagate’s unique end to end innovation from the drives to the software

Thank You!

https://github.com/Seagate/cortx
https://seagate.com/cortx/innersource

mailto:john.bent@seagate.com