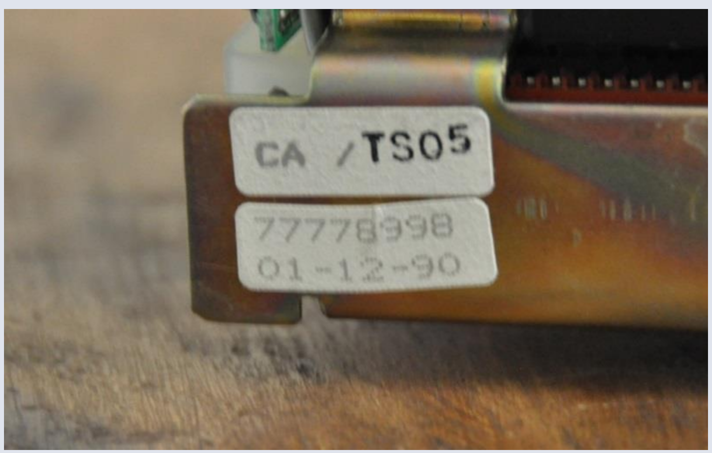
Freezing Exabytes of Data at Facebook’s Cold Storage

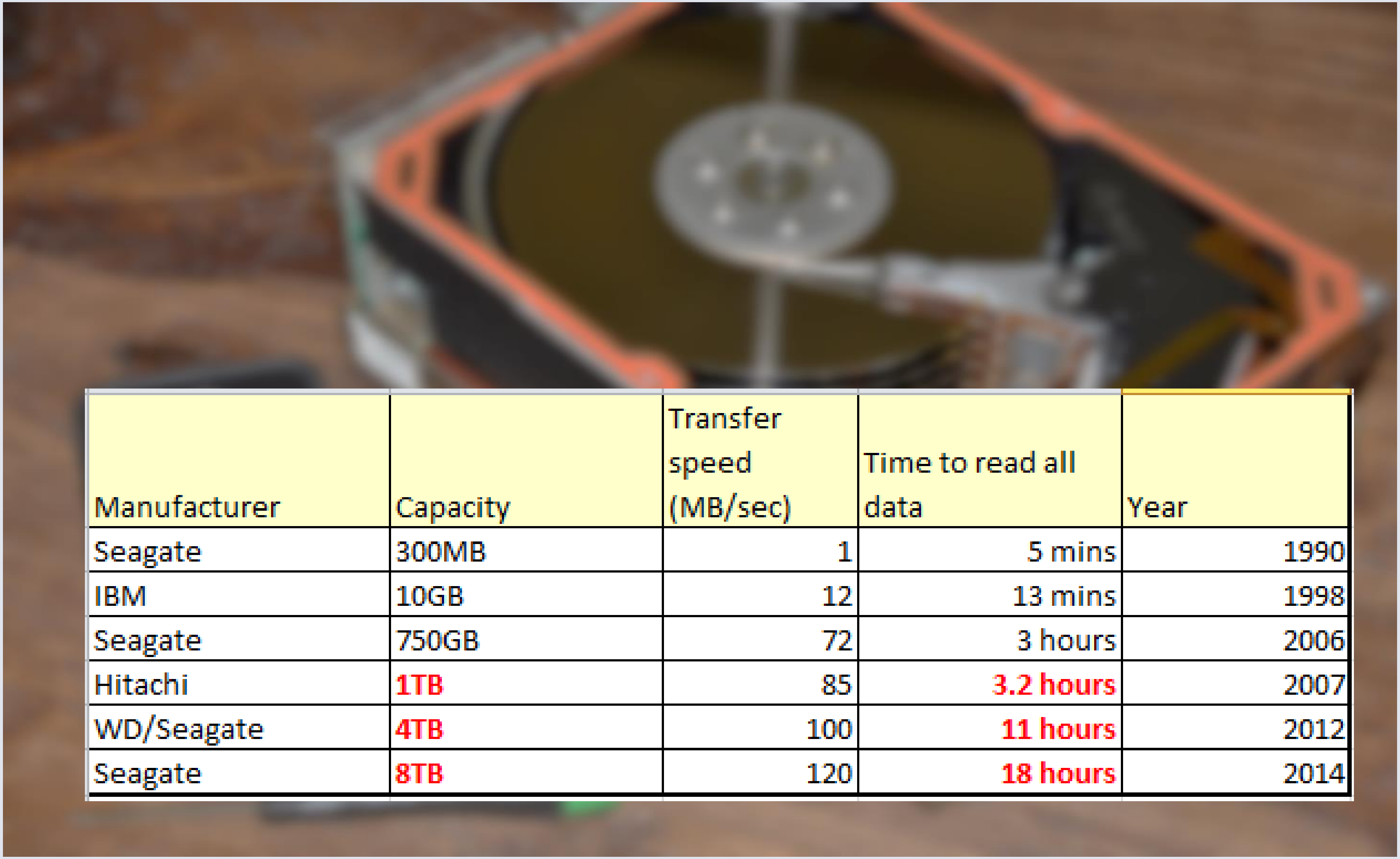
Kestutis Patiejunas (kestutip@fb.com)

1990 vs. 2014



Specs	Value
Form	3.5"
Platters	5
Heads	9
Capacity	300MB
Interface	SCSI
Seek time	17ms
Data transfer rate	1 MB/sec

History of Hard Drive data transfer rates



Manufacturer	Capacity	Transfer speed (MB/sec)	Time to read all data	Year
Seagate	300MB	1	5 mins	1990
IBM	10GB	12	13 mins	1998
Seagate	750GB	72	3 hours	2006
Hitachi	**1TB**	85	**3.2 hours**	2007
WD/Seagate	**4TB**	100	**11 hours**	2012
Seagate	**8TB**	120	**18 hours**	2014

Tape is Dead
Disk is Tape
Flash is Disk
RAM Locality is King

Jim Gray
Microsoft
December 2006

Tape Is Dead
Disk is Tape

- 1TB disks are available
- 10+ TB disks are predicted in 5 years
- Unit disk cost: ~\$400 → ~\$80

- But: ~ 5..15 **hours to read (sequential)**
- ~15..150 **days to read (random)**

- Need to treat **most of disk** as
  Cold-storage archive

Building Facebook HDD Cold Storage

Distinct goals and principles

(otherwise we will get another HDFS)

Goals and non goals

1. **Durable**
2. **High efficiency**
3. **Reasonable availability**
4. **Scale**
5. **Support evolution**
6. **Gets better as it gets bigger**

1. Have low latency for write/read operations
2. Have high availability
3. Be efficient for small objects
4. Be efficient for the objects with short lifetime

Principles

#1. **Durability** comes from eliminating single points of failure and ability to recover full system out of the remaining portions.

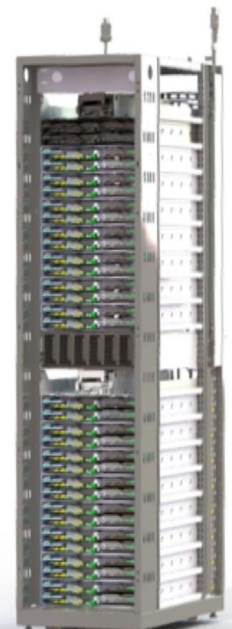
#2. **High efficiency** is from batching and trading latency for the efficiency. We spend mostly on the storing the data and not the metadata.

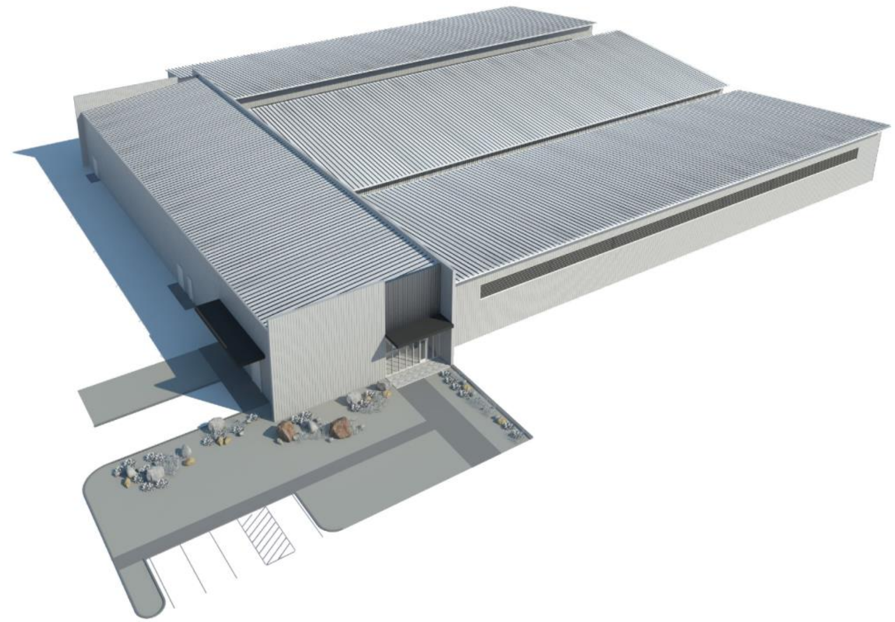
#3. **Simplicity** leads to reliability. Trade complexity and features for simplicity, gain durability and reliability.

#4. **Handle failures** from the day one. Distributed systems fail even on the sunny day, we learn about the mistakes when we find that intended recovery doesn’t work.

Architecture from 36,000 feet

Facebook HDD Cold Storage – HW parts of the solution

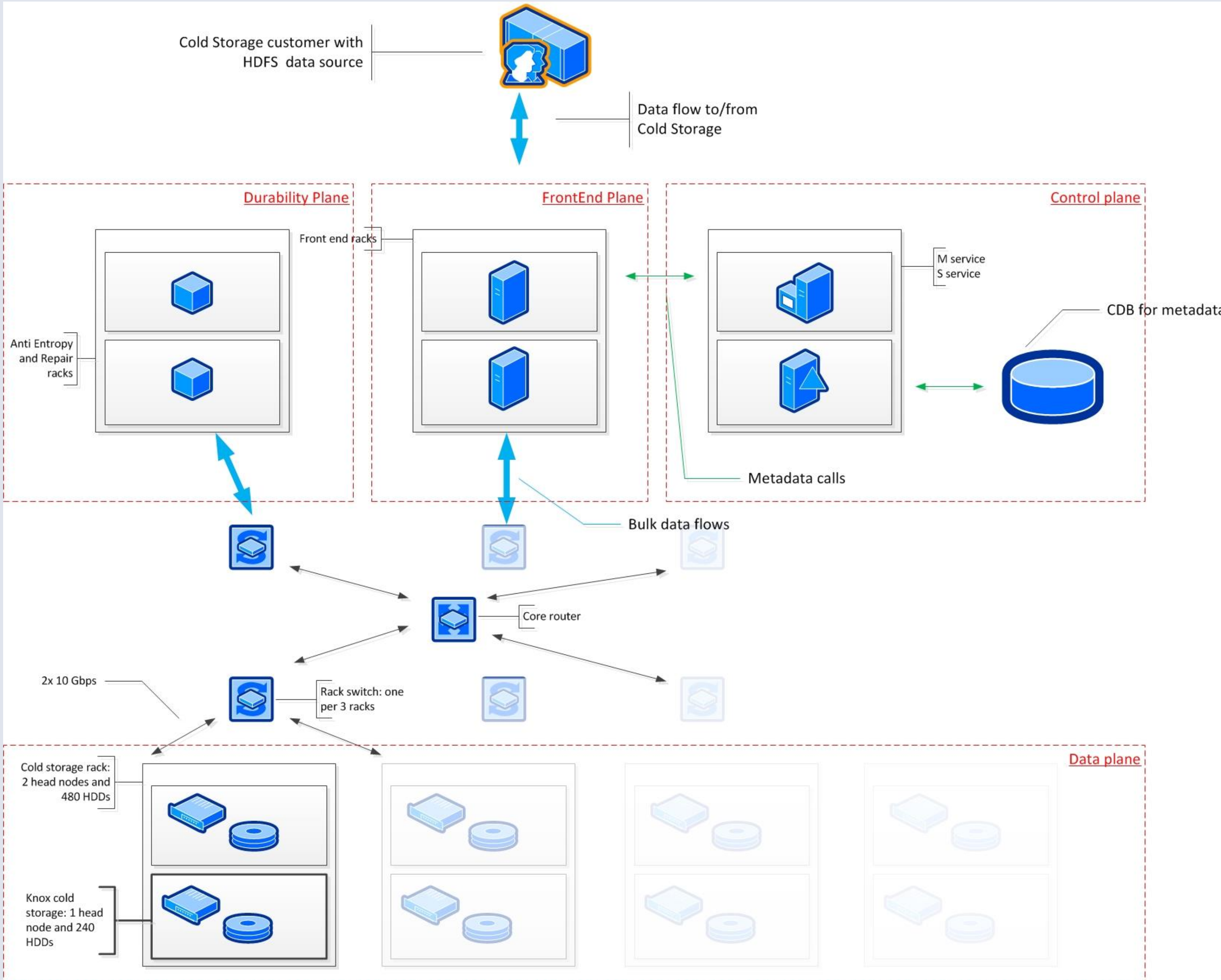




1/3 The cost of conventional storage servers

1/5 The cost of conventional data centers

Software architecture when we started in 2013



Raid rebuild vs Distributed volume rebuild

[redacted]

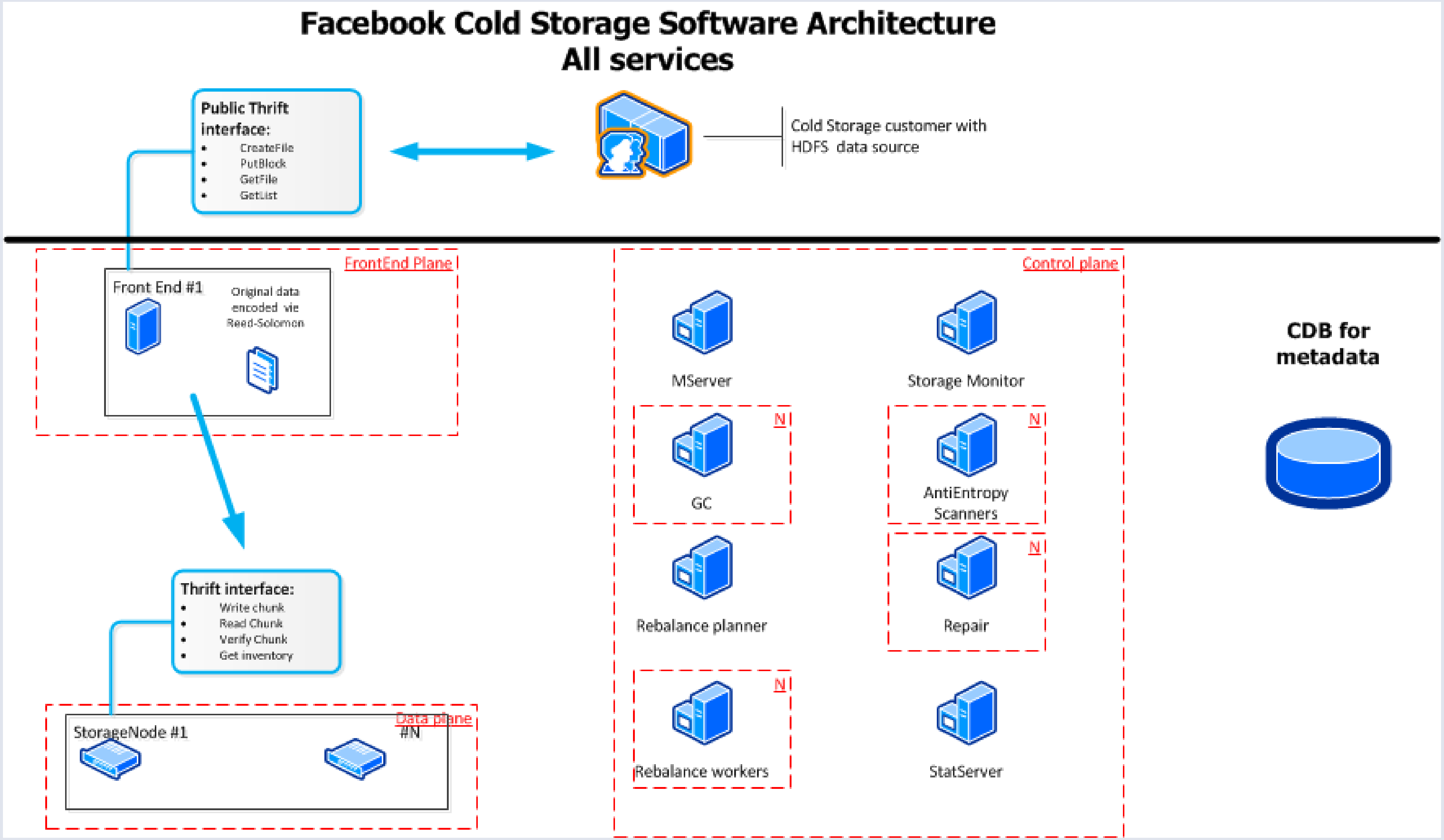
raid rebuild time. Below are some of prevailing stats and options

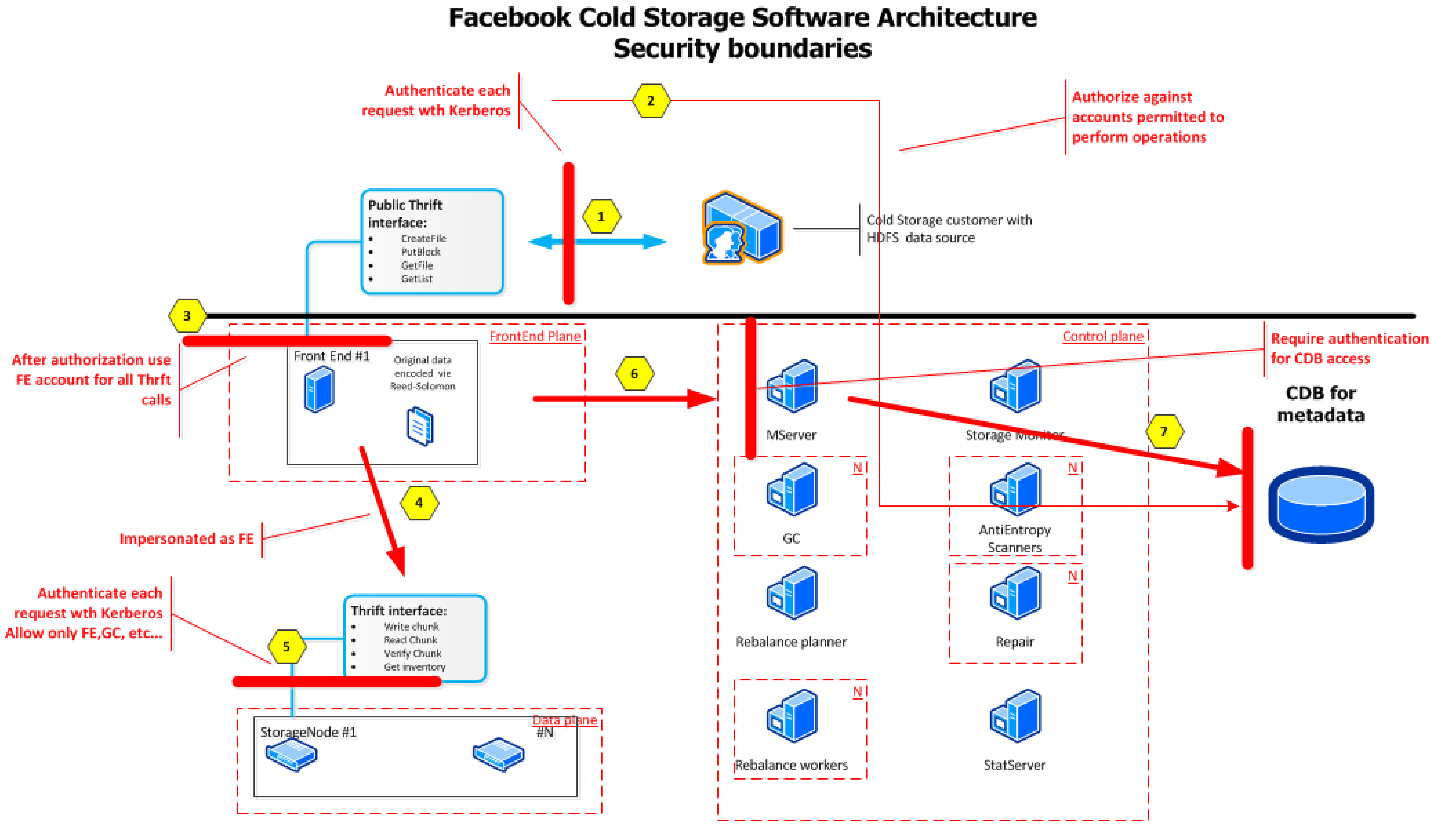
1) Default (Current State) : The [redacted] cards are set at 30% rebuild rate priority, with this a typical 4TB drive take almost 20Days to rebuild.
2) Set the rebuild rate priority to 100% , By doing so , we can possibly increase the rebuild rate and get the volume in healthy state in ~10-12 days.

Gets better as it gets bigger

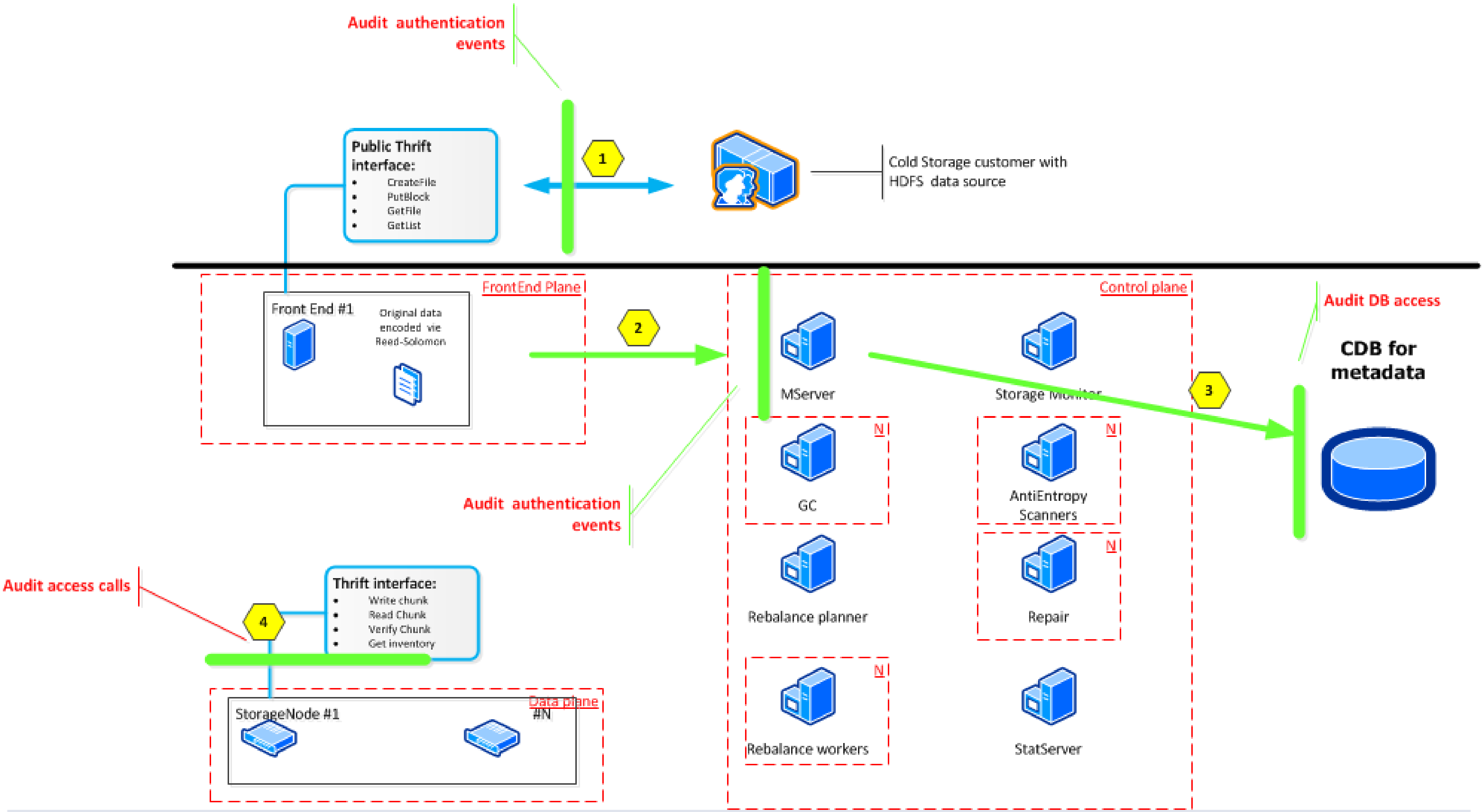
Number of racks	Capacity (PB)	Amount of data to read(write) in 1h at 50%	PB in 24 hours
30	52	0.1	1.2
90	156	0.2	3.7
200	346	0.3	8.2
500	865	0.9	20.6
1000	1730	1.7	41.2
2000	3460	3.4	82.4

The Real Software architecture after 9 months in production




Facebook Cold Storage Software Architecture
Security boundaries
Authenticate each
request wth Kerberos
2
Authorize against
accounts permitted to
perform operations
Public Thrift
interface:
CreateFile
PutBlock
GetFile
GetList
1
Cold Storage customer with
HDFS data source
3
FrontEnd Plane
Control plane
Require authentication
for CDB access
After authorization use
FE account for all Thrft
calls
Front End #1
Original data
encoded vie
Reed-Solomon
6
MServer
Storage Monitor
7
CDB for
metadata
GC
AntiEntropy
Scanners
4
Impersonated as FE
Authenticate each
request wth Kerberos
Allow only FE,GC, etc...
Thrift interface:
Write chunk
Read Chunk
Verify Chunk
Get inventory
5
Rebalance planner
Repair
StorageNode #1
Data plane
#N
Rebalance workers
StatServer
N

Facebook Cold Storage Software Architecture
Audit points



Raw disk storage

Problem:

- Takes 12h to fill 4TB HDD
- XFS can be formatted/erased in ~1sec

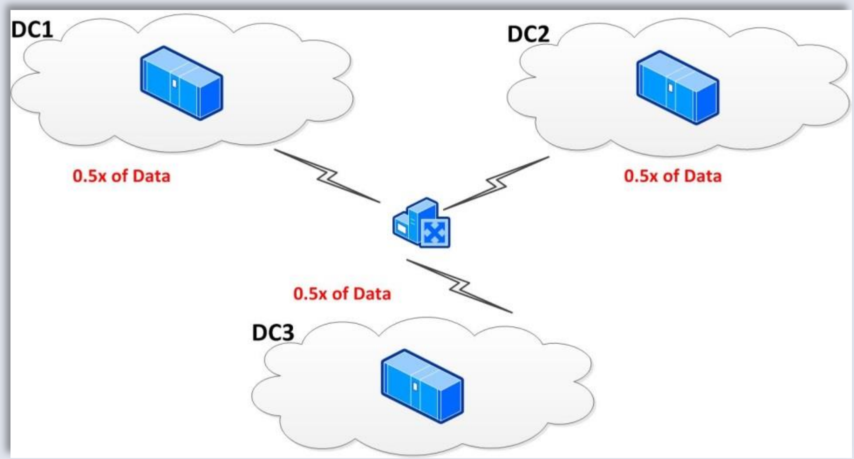
Solution

- Custom Raw disk storage format
- Metadata stored in 3 places
- Metadata is distributed
- Have to do full disk overwrite to erase data

Is this good enough?

What if we had a simplest roof water leak?
Disgruntled employee?
Software bug?
Fire?
Storm?
Earthquake?

What if we use Reed-Solomon across datacenters?



Metric	2 replicas	10/15 Reed solomon (3 datacenters)	Savings (percentage)
Storage	2.8	1.5	187%
Network required per DC	100%	50%	200%
Availability	99.998910%	99.99674	1%
Downtime per year (minutes)	5.7	17.1	33%

Conclusion: trade between storage, network and CPU

Like RAID systems do

Like HDFS and similar systems do

Just do this at the datacenter level

(can mix Cold and Regular datacenters)

So was Jim Gray 100% right about the future?

Tape is Dead
Disk is Tape
Flash is Disk
RAM Locality is King

Jim Gray
Microsoft
December 2006

Tape Is Dead
Disk is Tape

- 1TB disks are available
- 10+ TB disks are predicted in 5 years
- Unit disk cost: ~$400 → ~$80

- But: ~ 5..15 **hours to read (sequential)**
- ~15..150 **days to read (random)**

- Need to treat **most of disk** as **Cold-storage archive**

Questions and possibilities for mass storage industry

Hard drives:

- hit density wall with PMR – 1TB/platter
- adding more platters – 4-8TB
- adding SMR (only 15-20% increase)
- waiting for HAMR!
- going back to 5” factor?

Optical:

- 100GB/disc is cheap
- 300GB within 18 months
- 500GB 2-3 years
- Sony and Panasonic has 1TB/disc on the roadmap

Questions and possibilities for mass storage industry

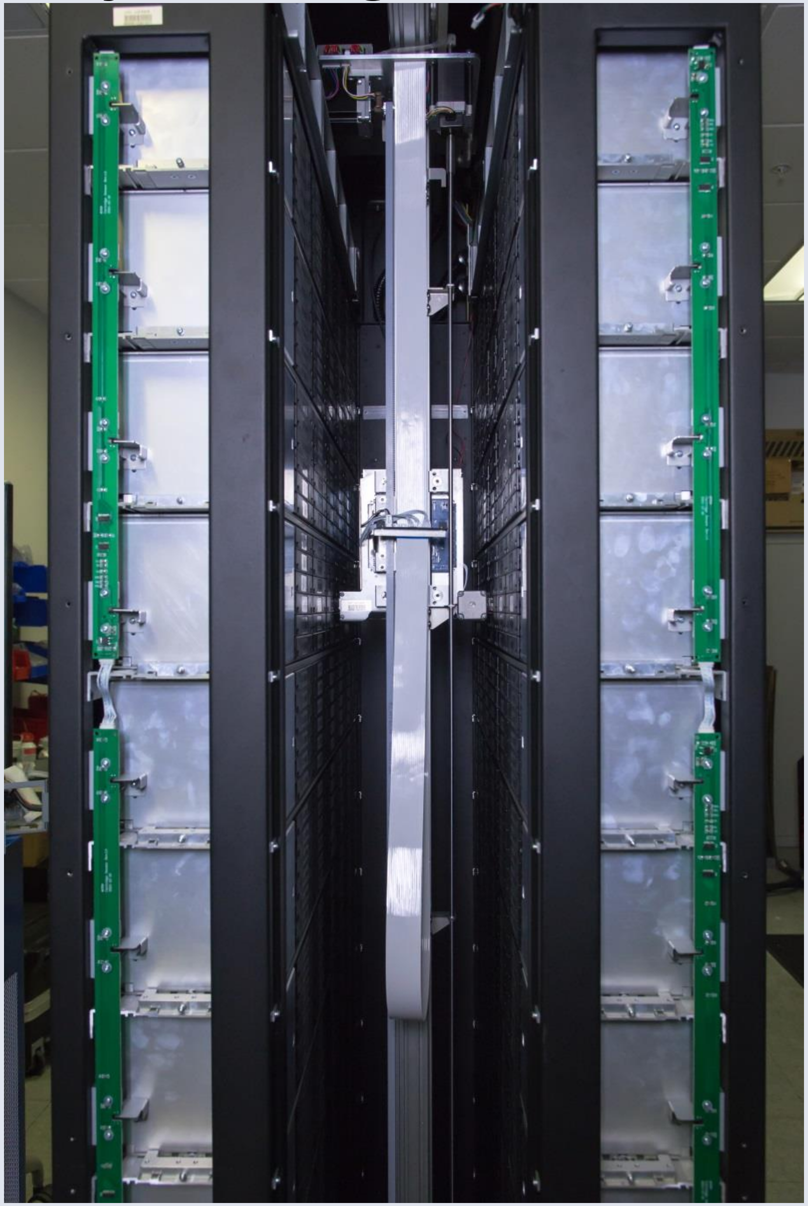
Hard drives:

- Less demand from IOPS intensive workload (either shifting to SSD, or can’t use all of the capacity)
- Small demand from consumers for large capacity

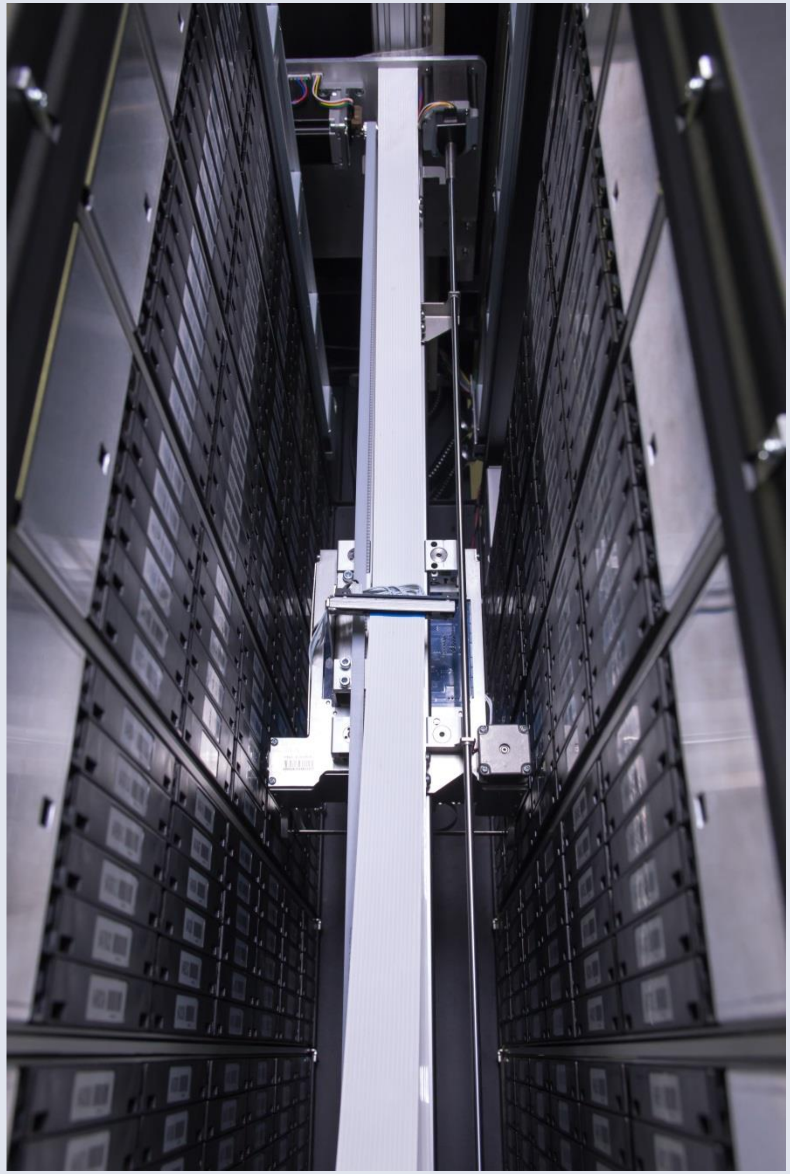
Optical:

- 4k or 8k movies will need lots of storage

Facebook Blu-Ray storage rack

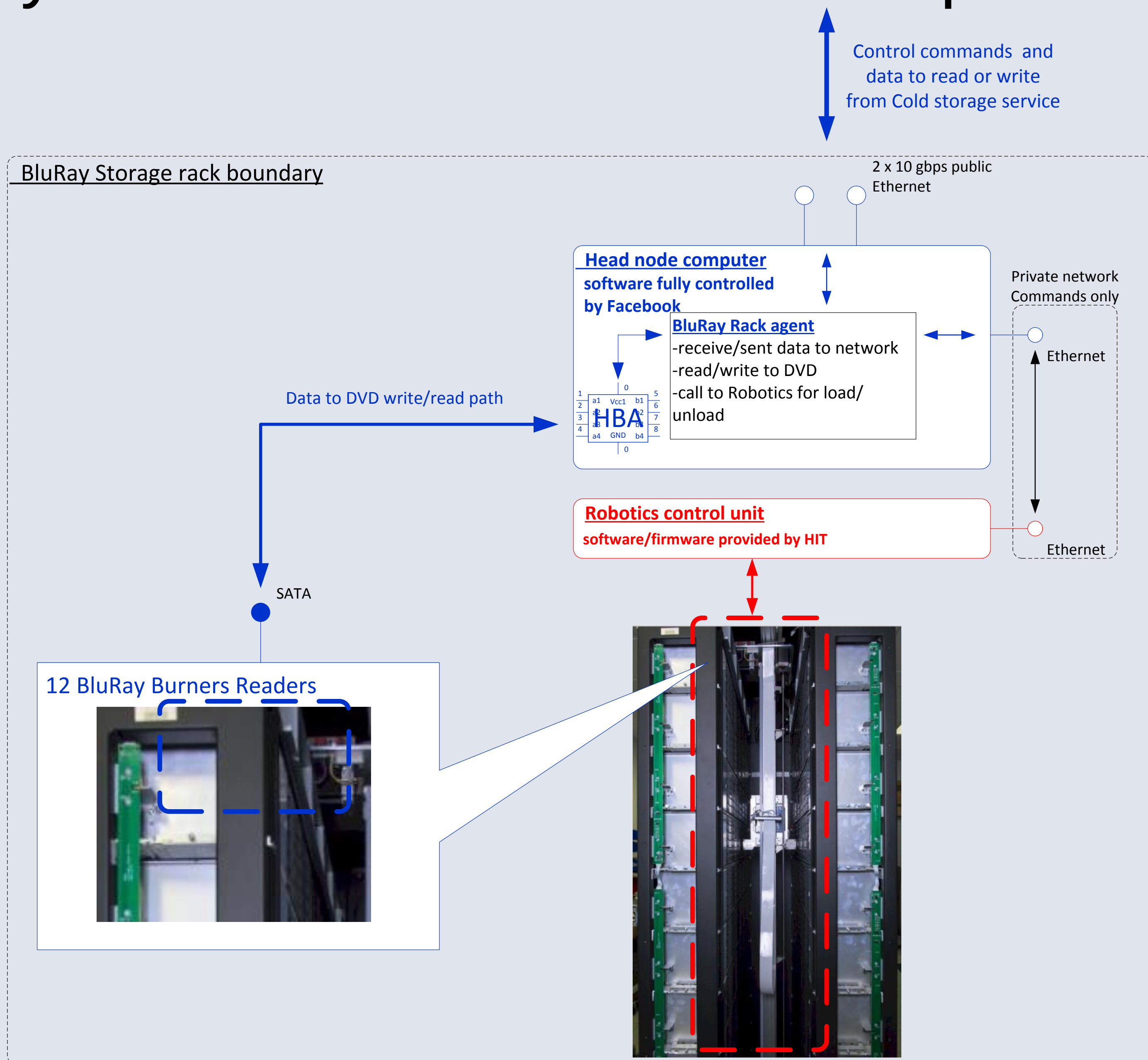


Facebook Blu-Ray storage rack

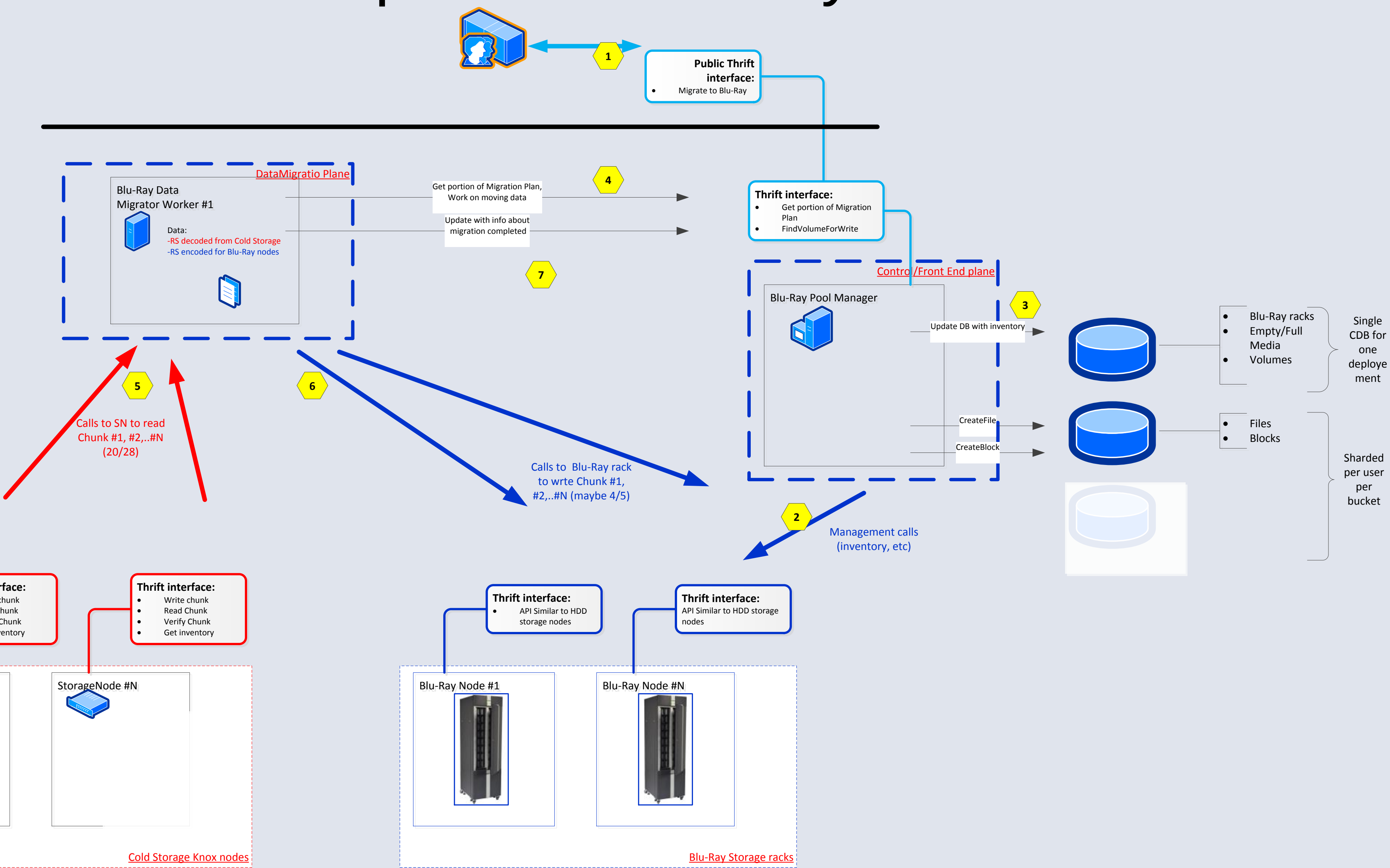


Facebook Blu-Ray storage rack

BluRay rack software/hardware components



Details data write path into BluRay racks



Limits of BluRay storage, pros and cons vs HDD storage

1. **Load time**
   Time to load media for read/write ~30s – loadings should be amortized for reading/writing big data sets (full discs).

2. **IO speed**
   Current time to read/write 100GB disc is 1.3 hours (about 5-6x longer than on HDDs).

3. **Small fraction of active storage**
   BluRay rack has only 12 burners or 1.2TB of active storage. HDD rack has 32 active HDDs or 128TB of active storage. Write/read strategy is to cluster the data across the time and spread across multiple racks.

4. **Efficiency and longevity**
   Optical has big edge vs. HDD

Conclusions on Cold Storage

When data amounts are massive – efficiency is important

Specialization allows to achieve efficiency

If we are approaching the end of Moore’s and Kryder’s laws which of the storage media has more iterations left: silicon, magnetic or optical?

If we can’t see the future can we hedge our bets and how far we can push unexplored technologies to extract extra efficiency?