facebook

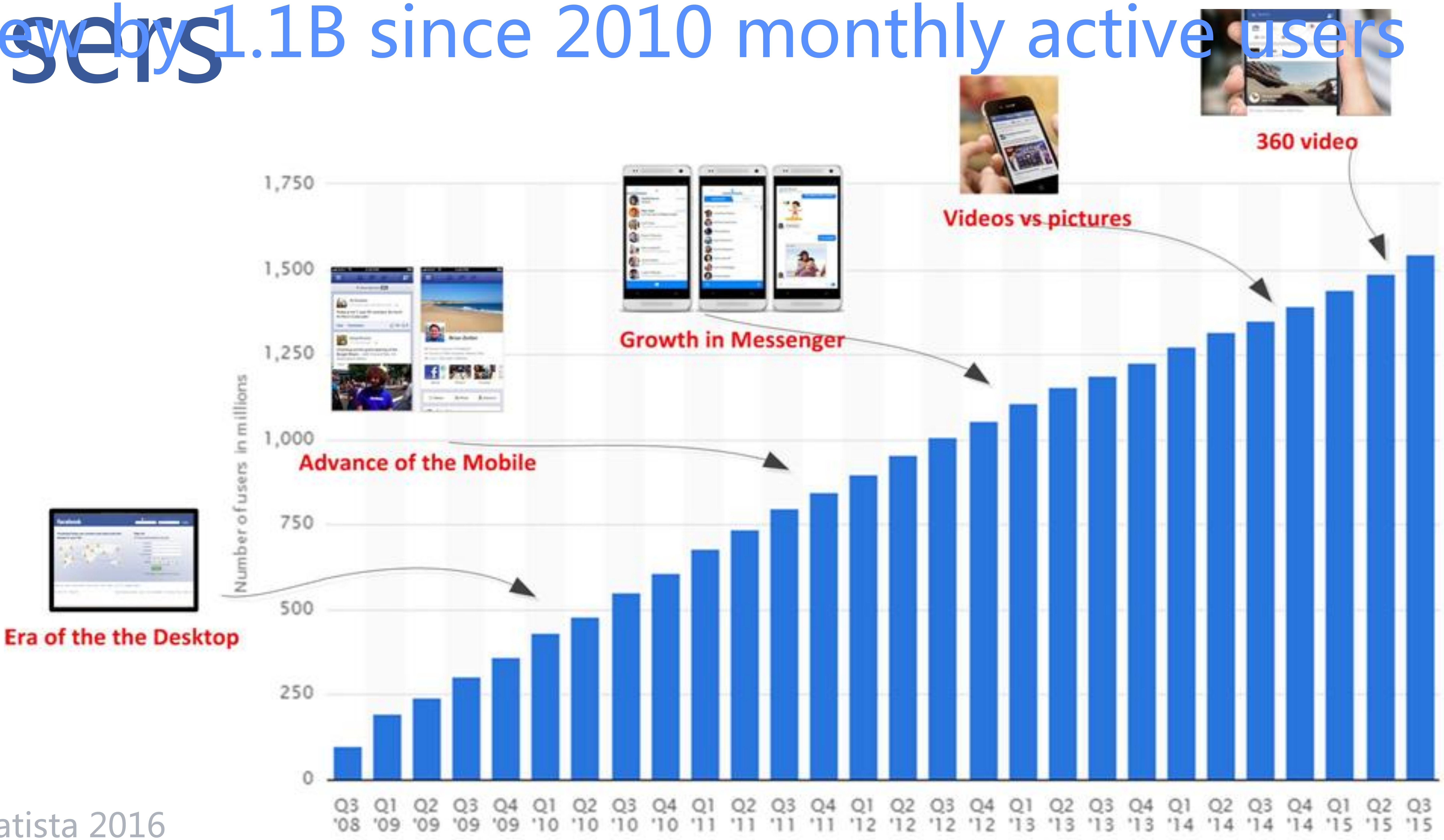
Facebook’s disaggregated storage and compute for Map/Reduce

Yun Jin

Intro

Facebook’ s Monthly Active Users

Grew by 1.1B since 2010 monthly active users



Era of the the Desktop

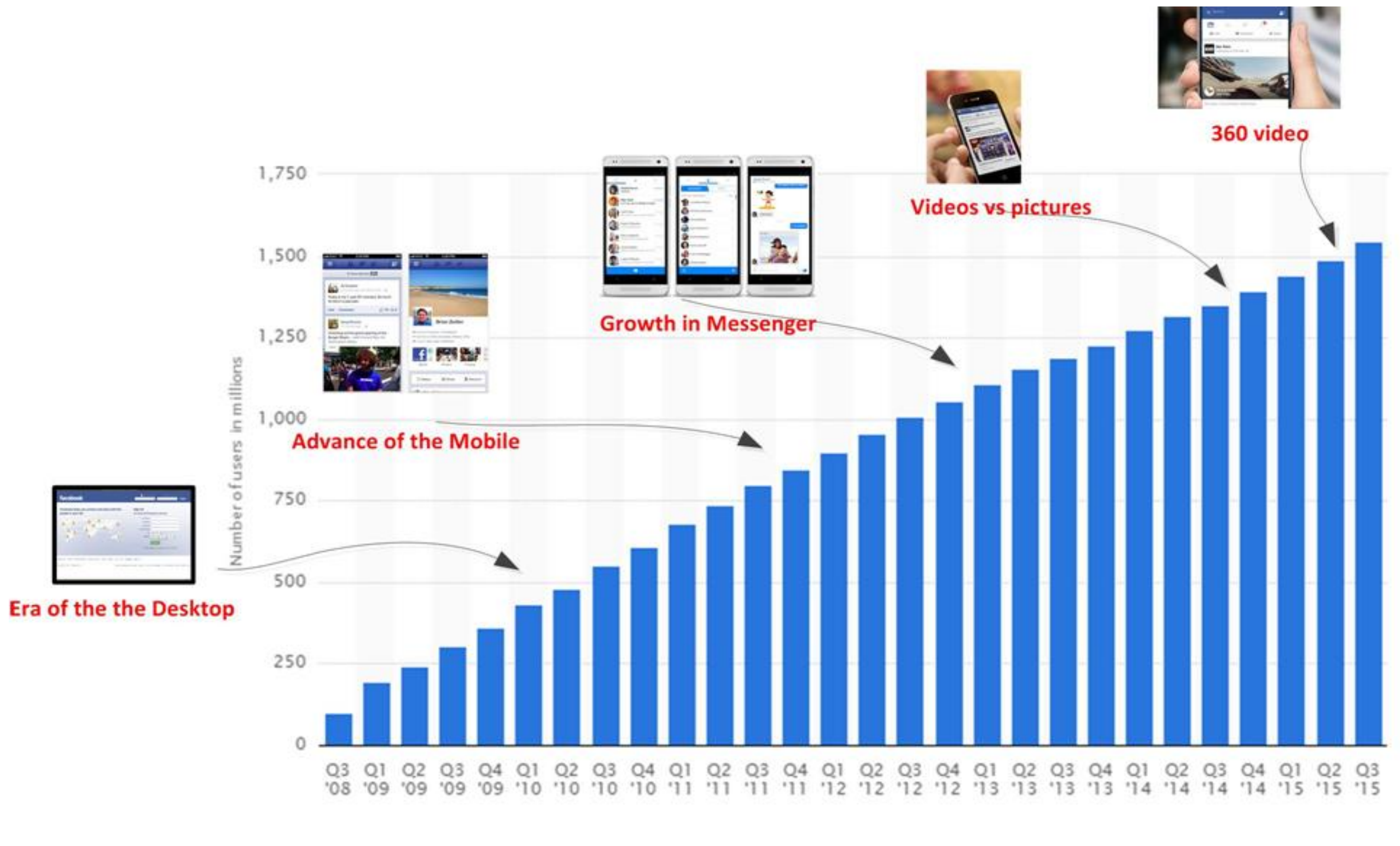
Advance of the Mobile

Growth in Messenger

Videos vs pictures

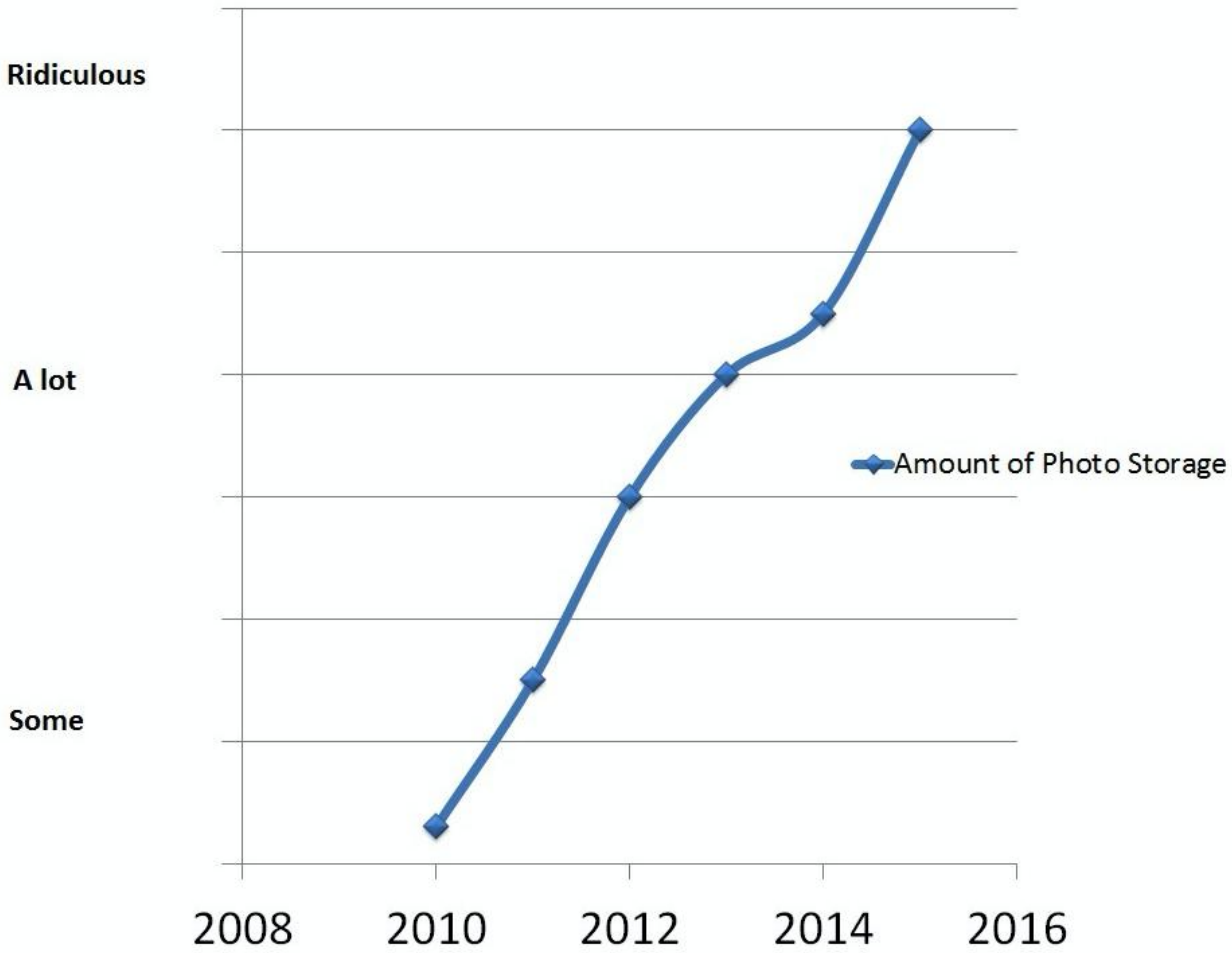
360 video

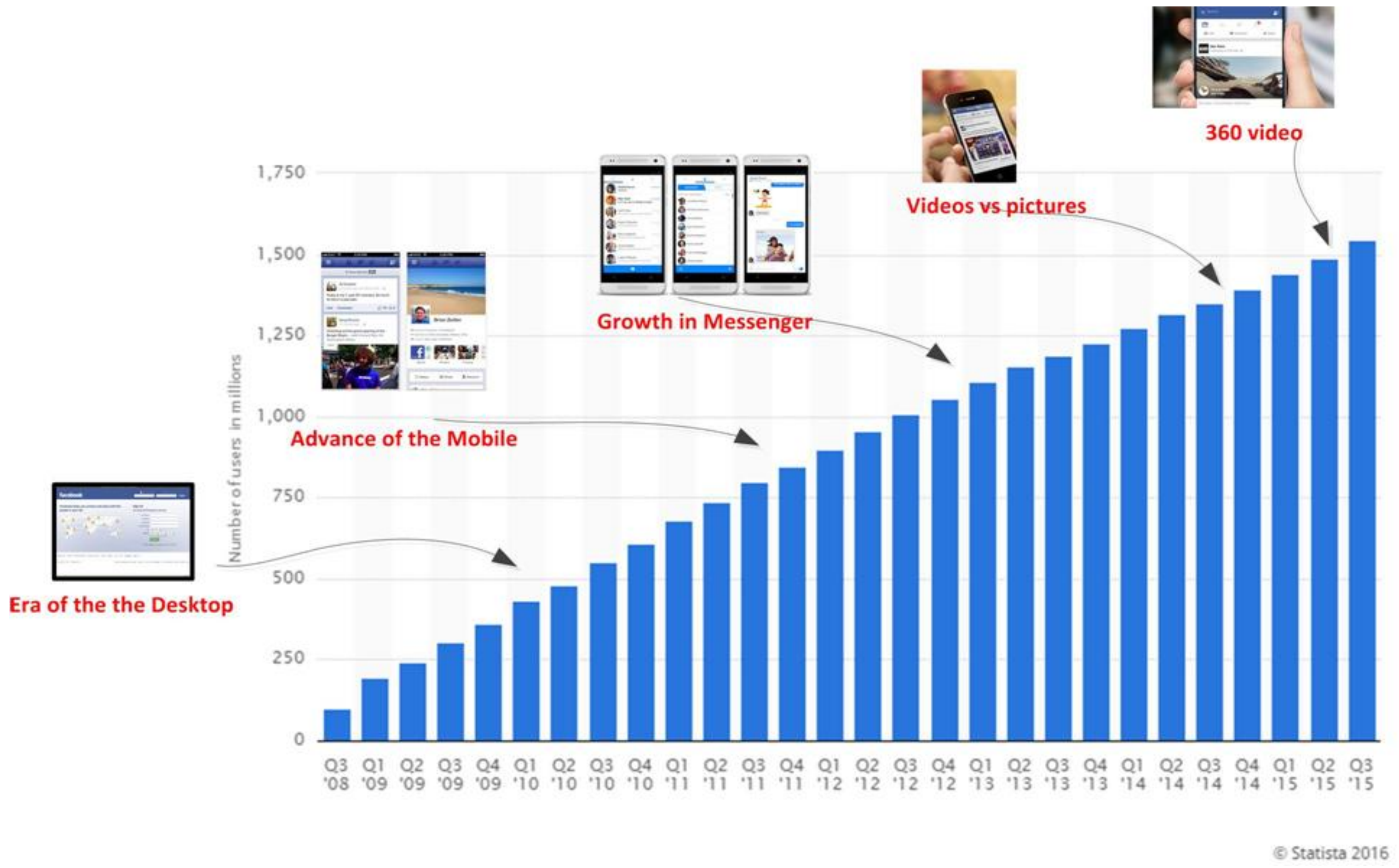
© Statista 2016



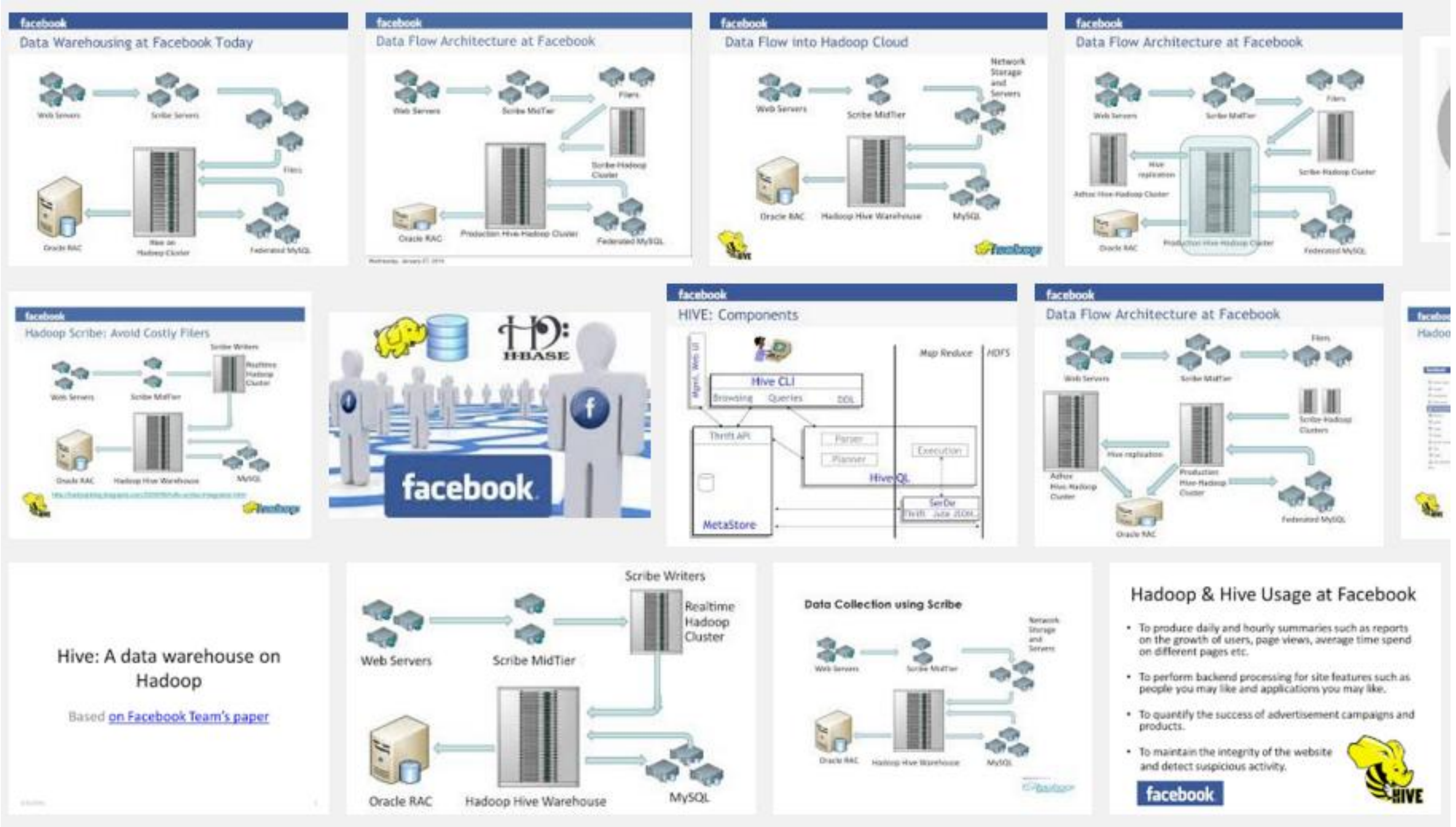
© Statista 2016

Amount of Photo Storage





© Statista 2016

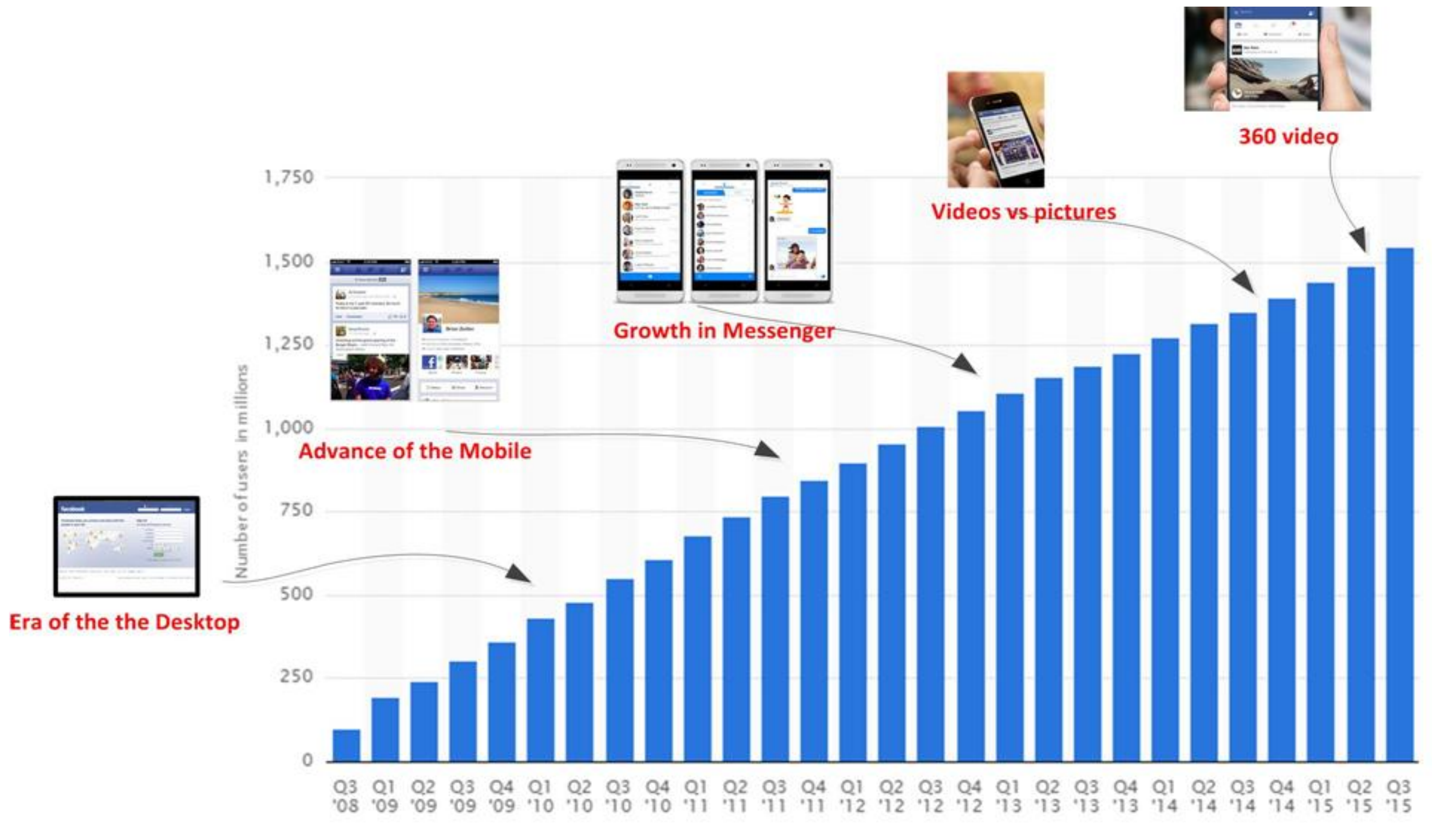

facebook
Data Warehousing at Facebook Today
Data Flow Architecture at Facebook
Data Flow into Hadoop Cloud
Data Flow Architecture at Facebook
Hadoop Scribe: Avoid Costly Filers
HBASE
facebook
HIVE: Components
Hive CLI
MetaStore
Data Flow Architecture at Facebook
Scribe Writers
Realtime Hadoop Cluster
Web Servers
Scribe MidTier
Oracle RAC
Hadoop Hive Warehouse
MySQL
Data Collection using Scribe

Hive: A data warehouse on Hadoop

Based on Facebook Team's paper

Hadoop & Hive Usage at Facebook

- To produce daily and hourly summaries such as reports on the growth of users, page views, average time spend on different pages etc.
- To perform backend processing for site features such as people you may like and applications you may like.
- To quantify the success of advertisement campaigns and products.
- To maintain the integrity of the website and detect suspicious activity.


Era of the the Desktop

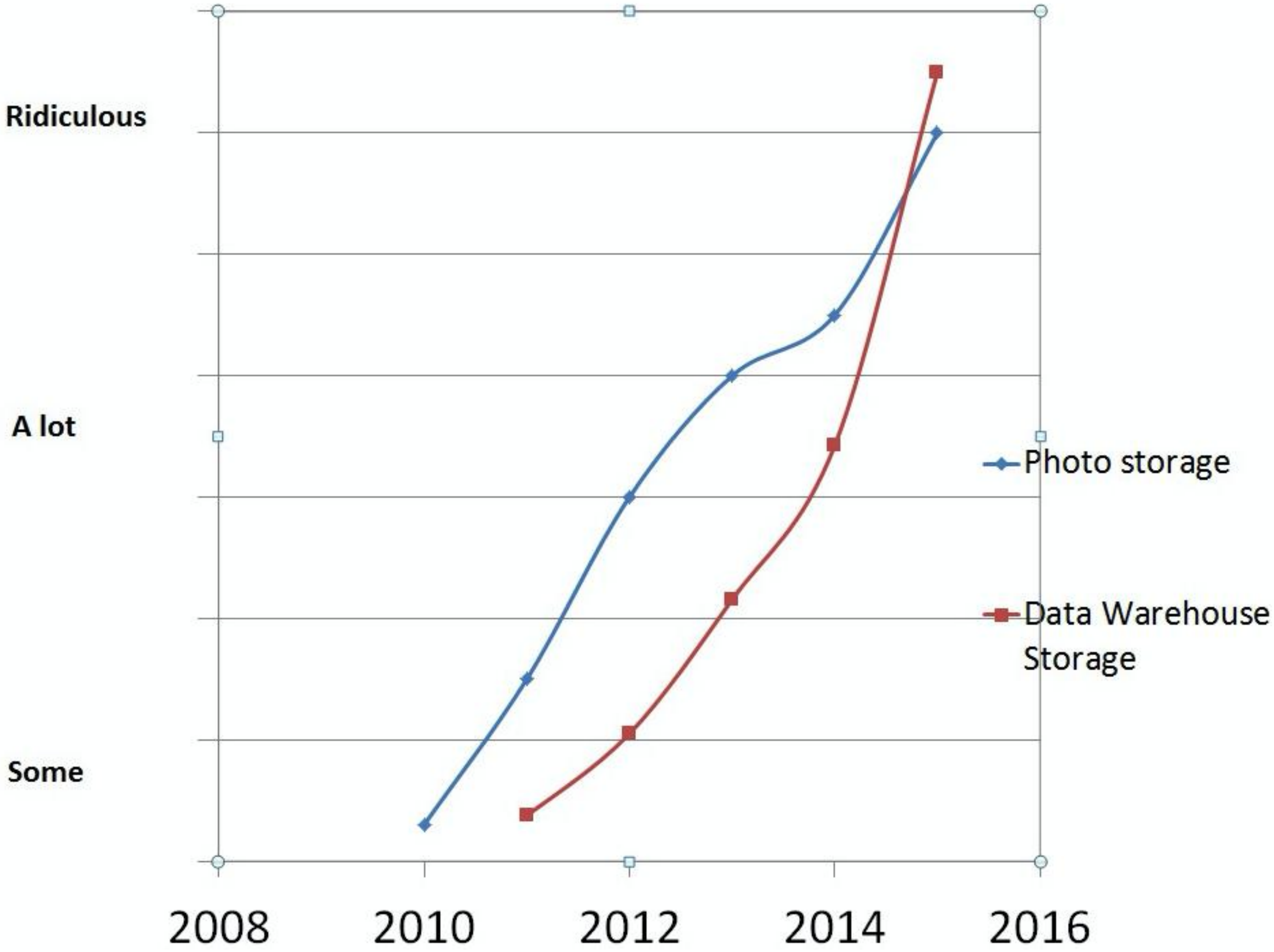
Advance of the Mobile

Growth in Messenger

Videos vs pictures

360 video

© Statista 2016



So what is an Exabyte?

- 1 Exabyte == 1000 Petabytes
- 1 Petabyte == 1000 Terabytes
- 1 Exabyte = ~250,000 4 TB drives

- **250k drives stacked fla >30 times taller than Seattle Space Needle**

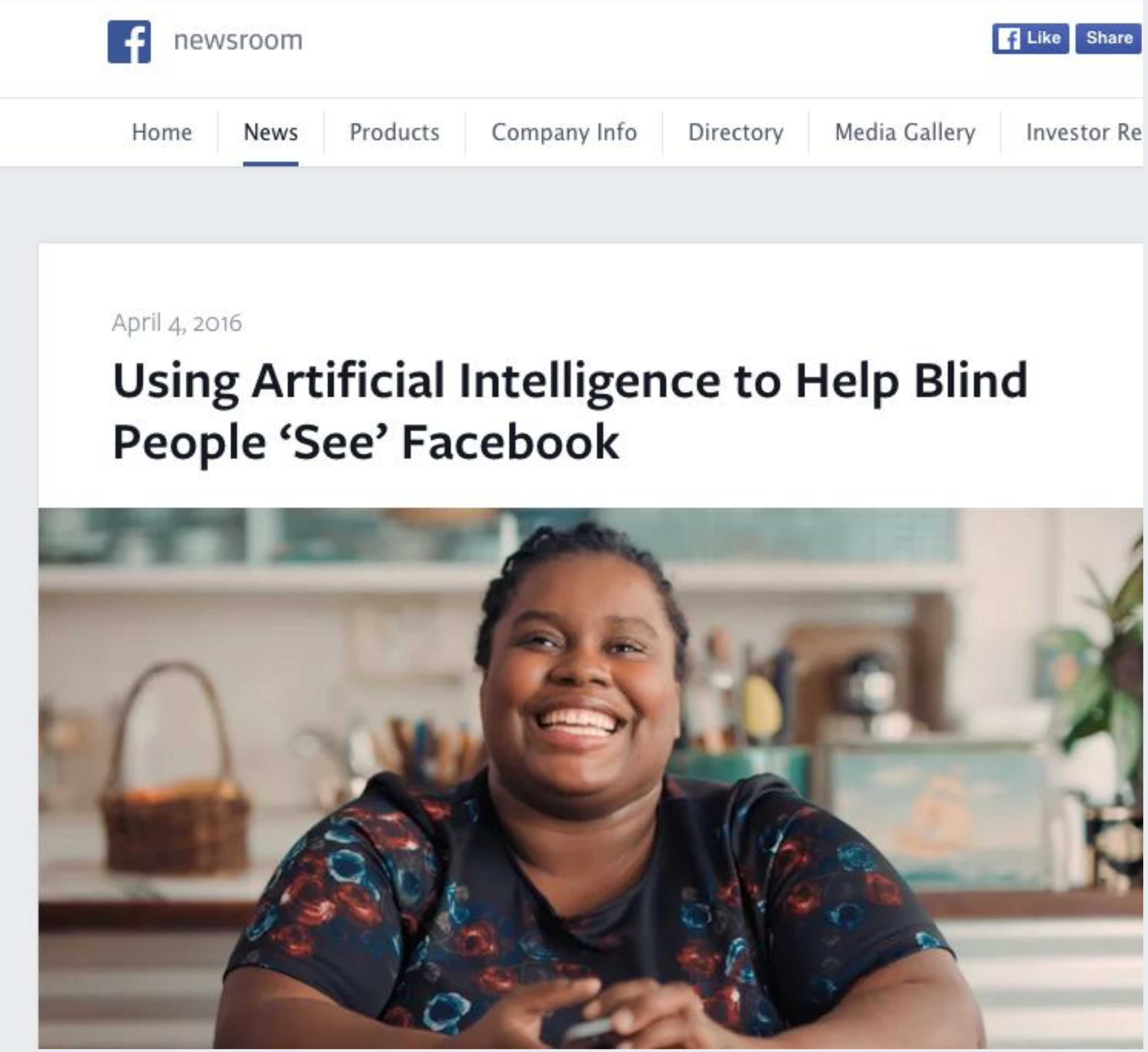
X 30 times!

Data Scientists are DW customers

Data Scientist and DW

Core Data Science

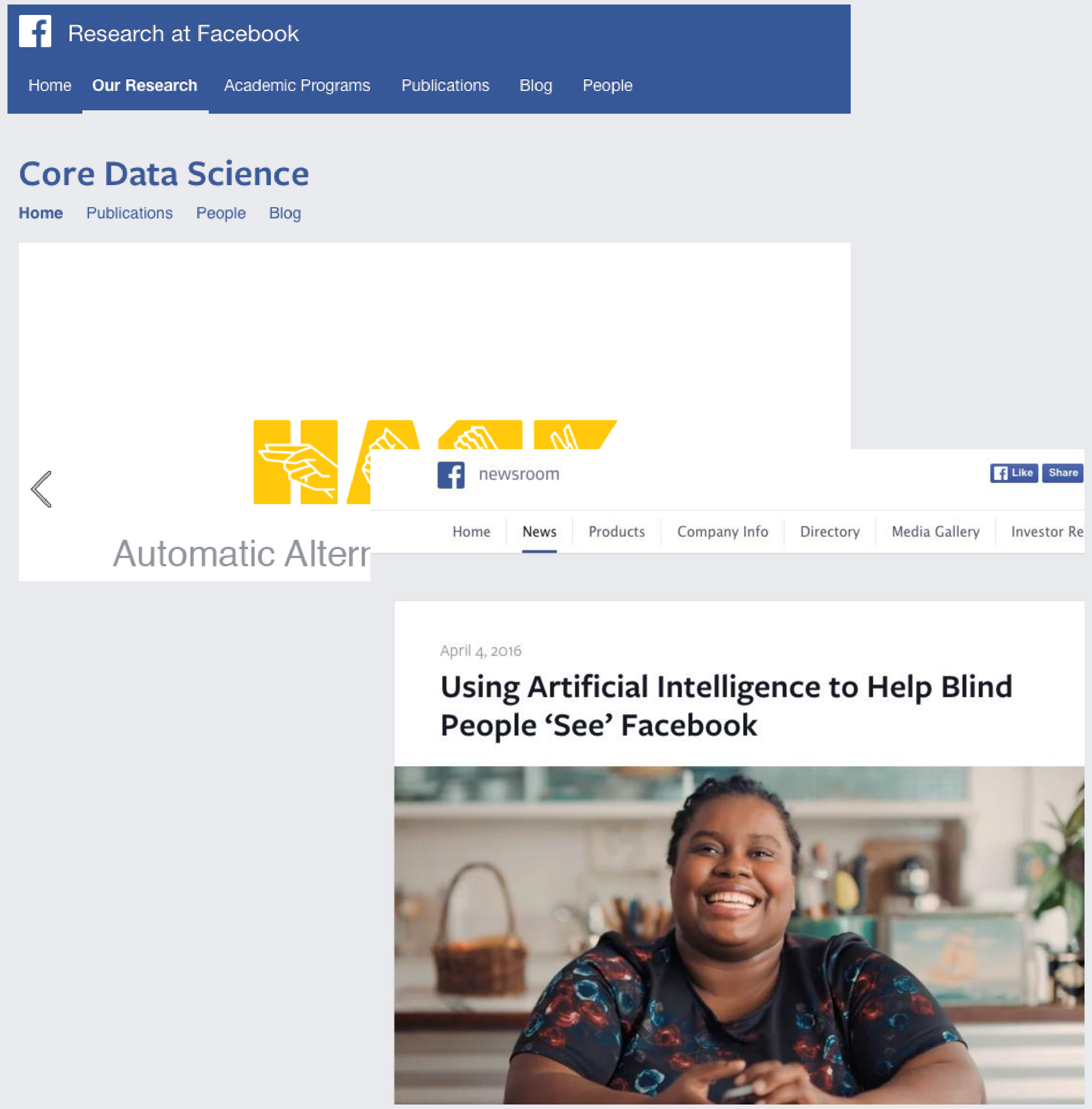
Automatic Alternative Text on Facebook



April 4, 2016

Using Artificial Intelligence to Help Blind People ‘See’ Facebook

Data Scientist and DW

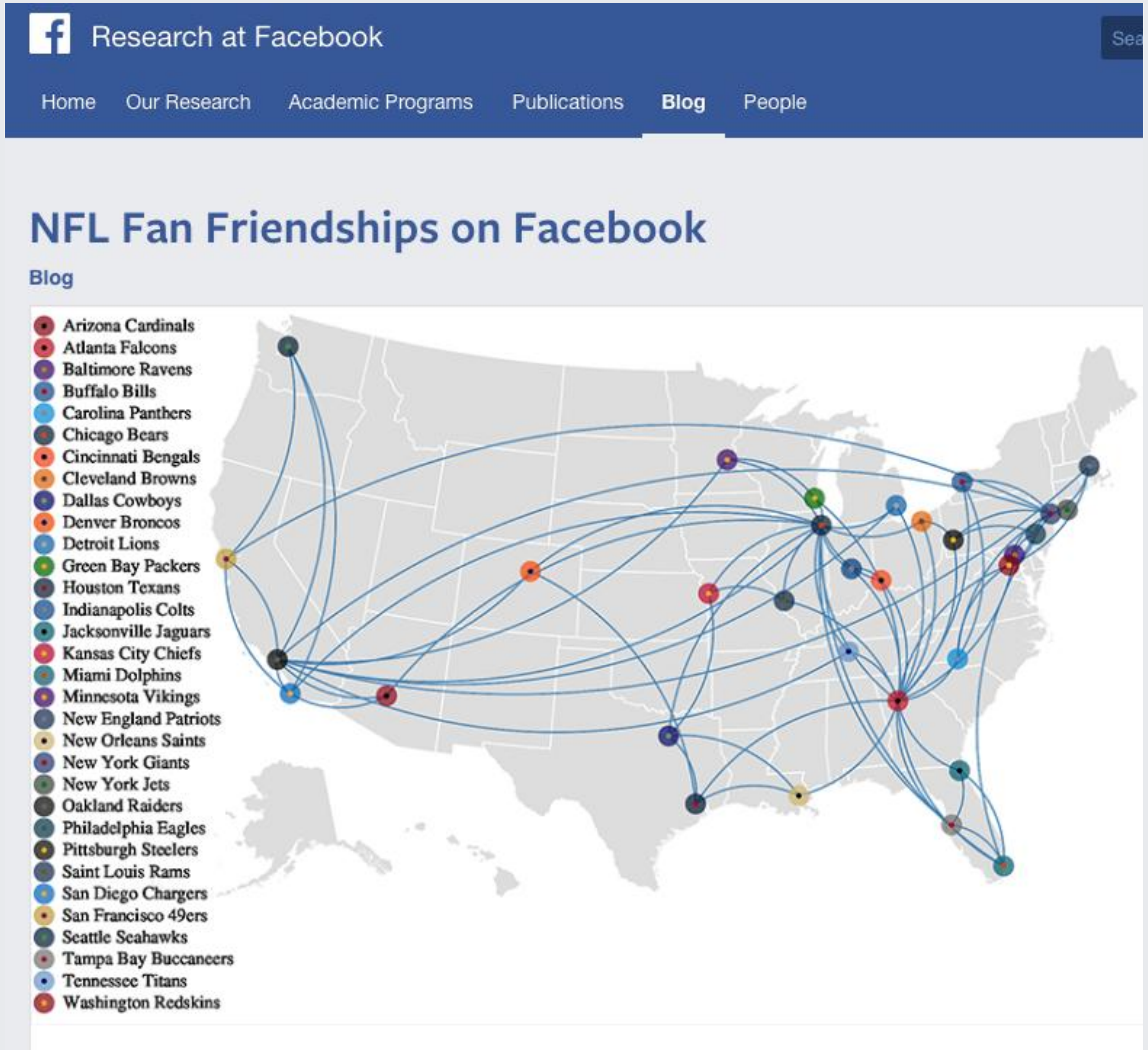


Core Data Science

Automatic Alterr

April 4, 2016

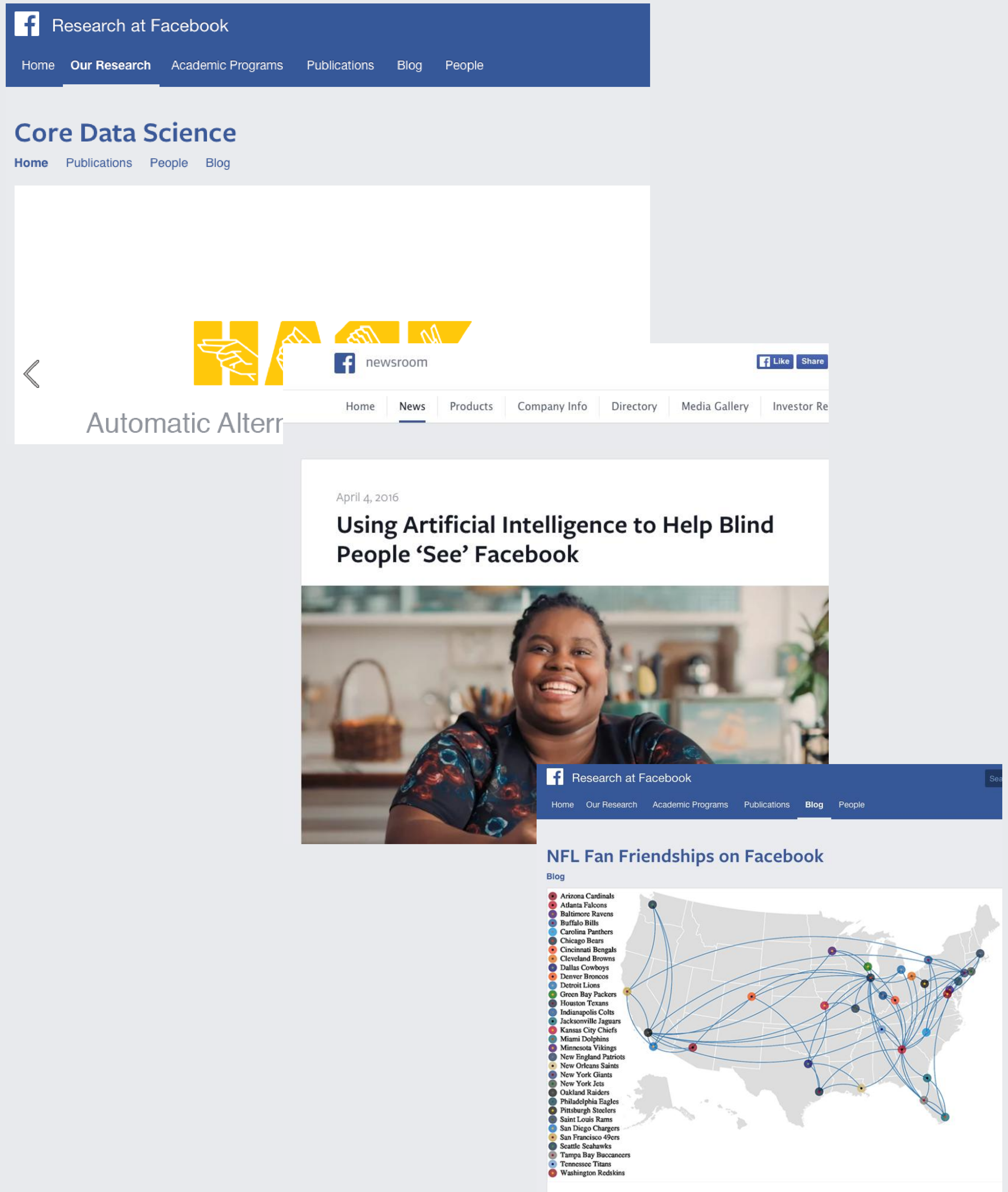
Using Artificial Intelligence to Help Blind People 'See' Facebook



NFL Fan Friendships on Facebook

Blog

Data Scientist and DW



Core Data Science

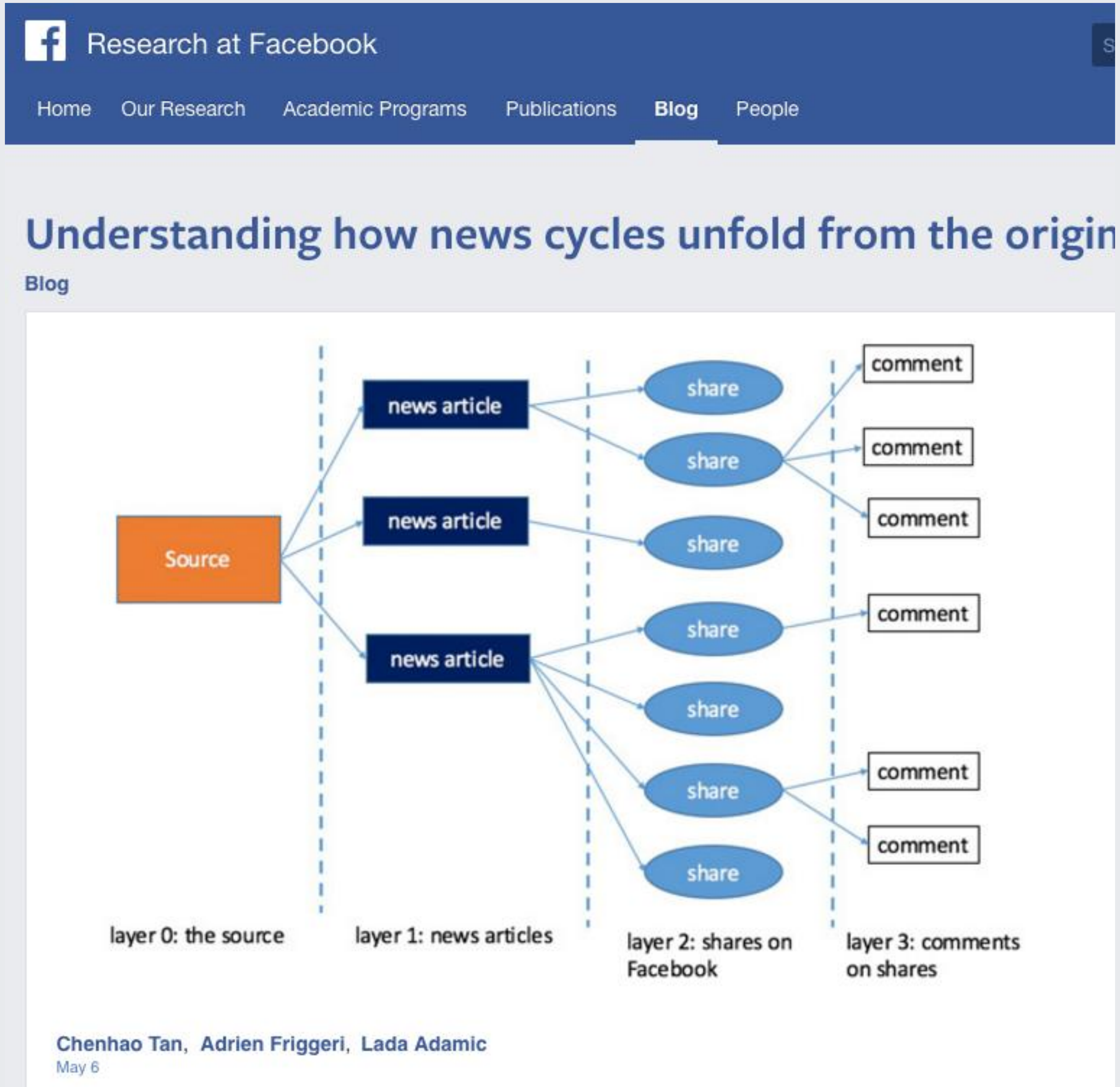
Automatic Altern

April 4, 2016

Using Artificial Intelligence to Help Blind People ‘See’ Facebook

NFL Fan Friendships on Facebook

Blog



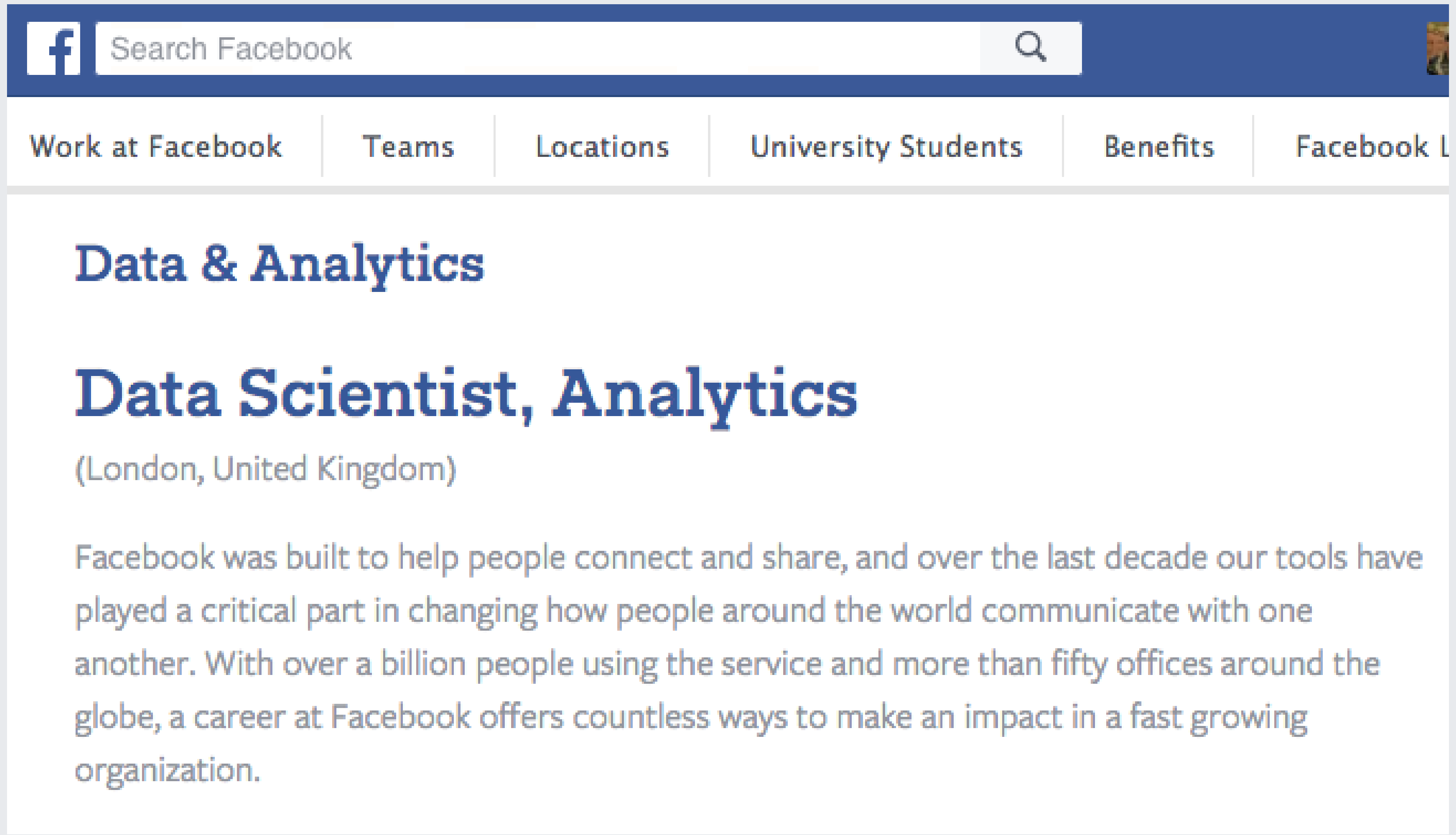
Understanding how news cycles unfold from the origin

Blog

Chenhao Tan, Adrien Friggeri, Lada Adamic

May 6

Data Science@FB: dynamic, growing and hard to predict

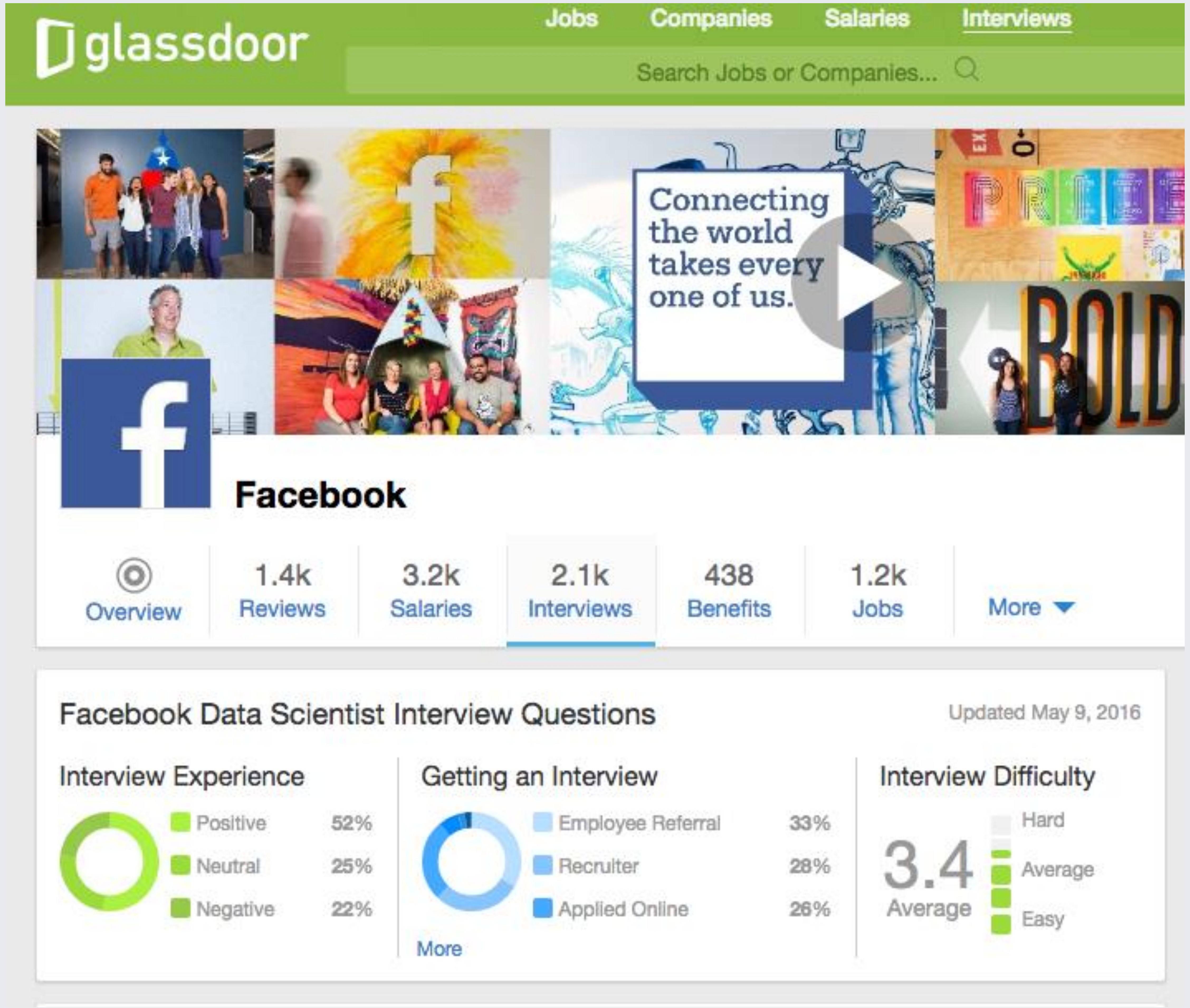


Data & Analytics

Data Scientist, Analytics

(London, United Kingdom)

Facebook was built to help people connect and share, and over the last decade our tools have played a critical part in changing how people around the world communicate with one another. With over a billion people using the service and more than fifty offices around the globe, a career at Facebook offers countless ways to make an impact in a fast growing organization.



Facebook Data Scientist Interview Questions

Updated May 9, 2016

Interview Experience

Positive	52%
Neutral	25%
Negative	22%

Getting an Interview

Employee Referral	33%
Recruiter	28%
Applied Online	26%

More

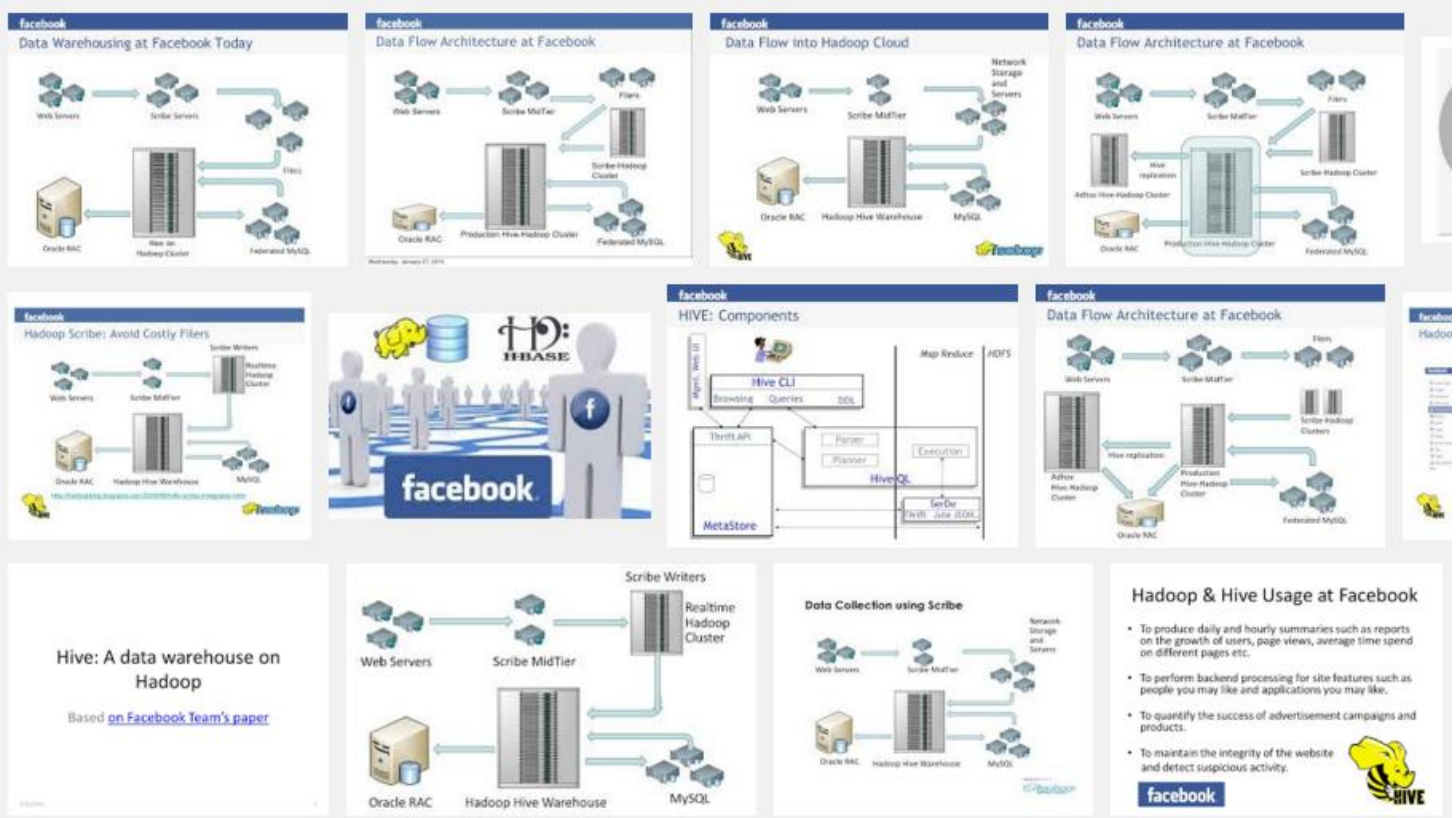
Interview Difficulty

3.4 Average

Hard
Average
Easy

Data Warehouse at Facebook

Data Warehouse at Facebook



Hive: A data warehouse on Hadoop

Based on Facebook Team's paper

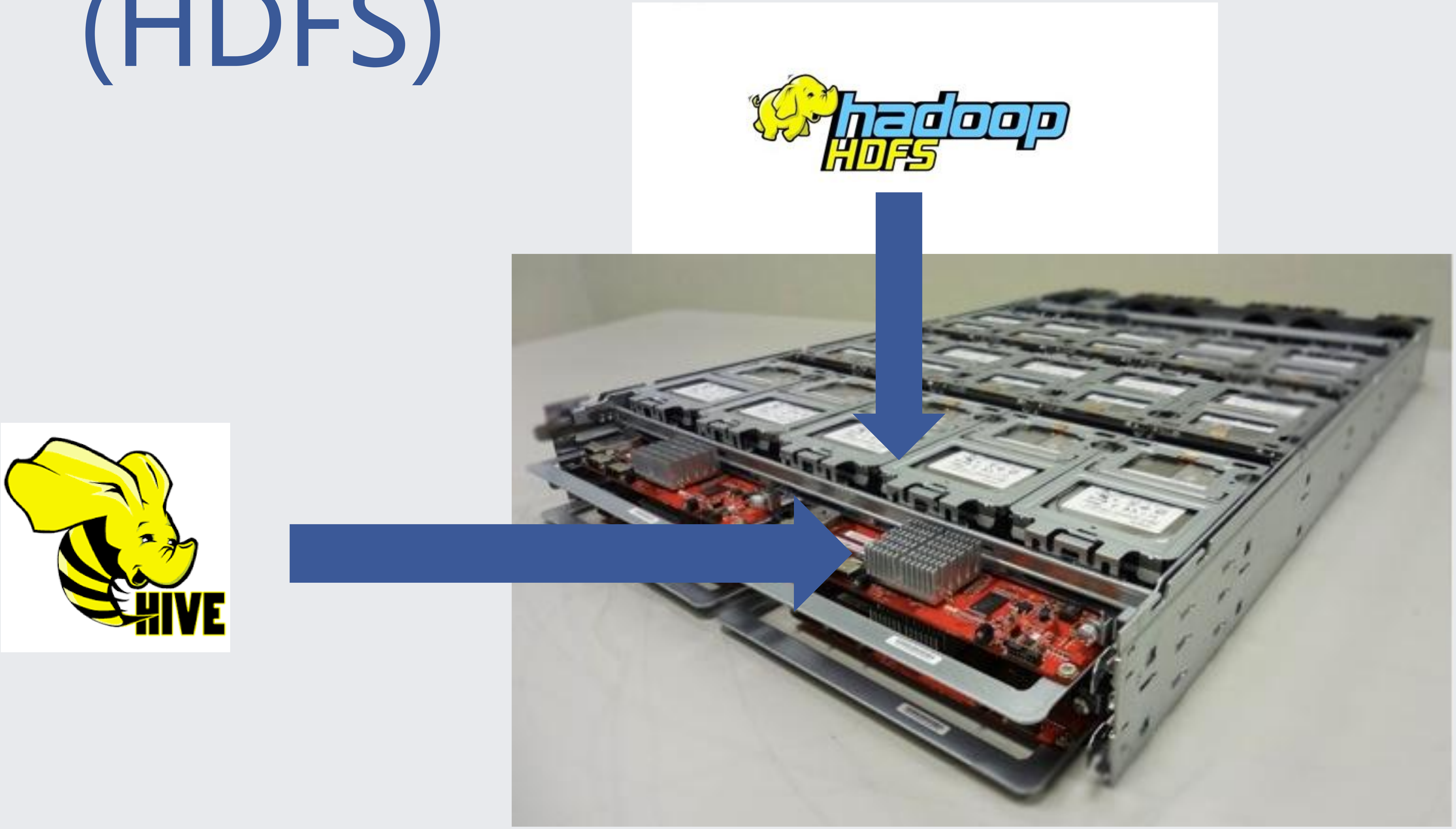
Hadoop & Hive Usage at Facebook

- To produce daily and hourly summaries such as reports on the growth of users, page views, average time spend on different pages etc.
- To perform backend processing for site features such as people you may like and applications you may like.
- To quantify the success of advertisement campaigns and products.
- To maintain the integrity of the website and detect suspicious activity.

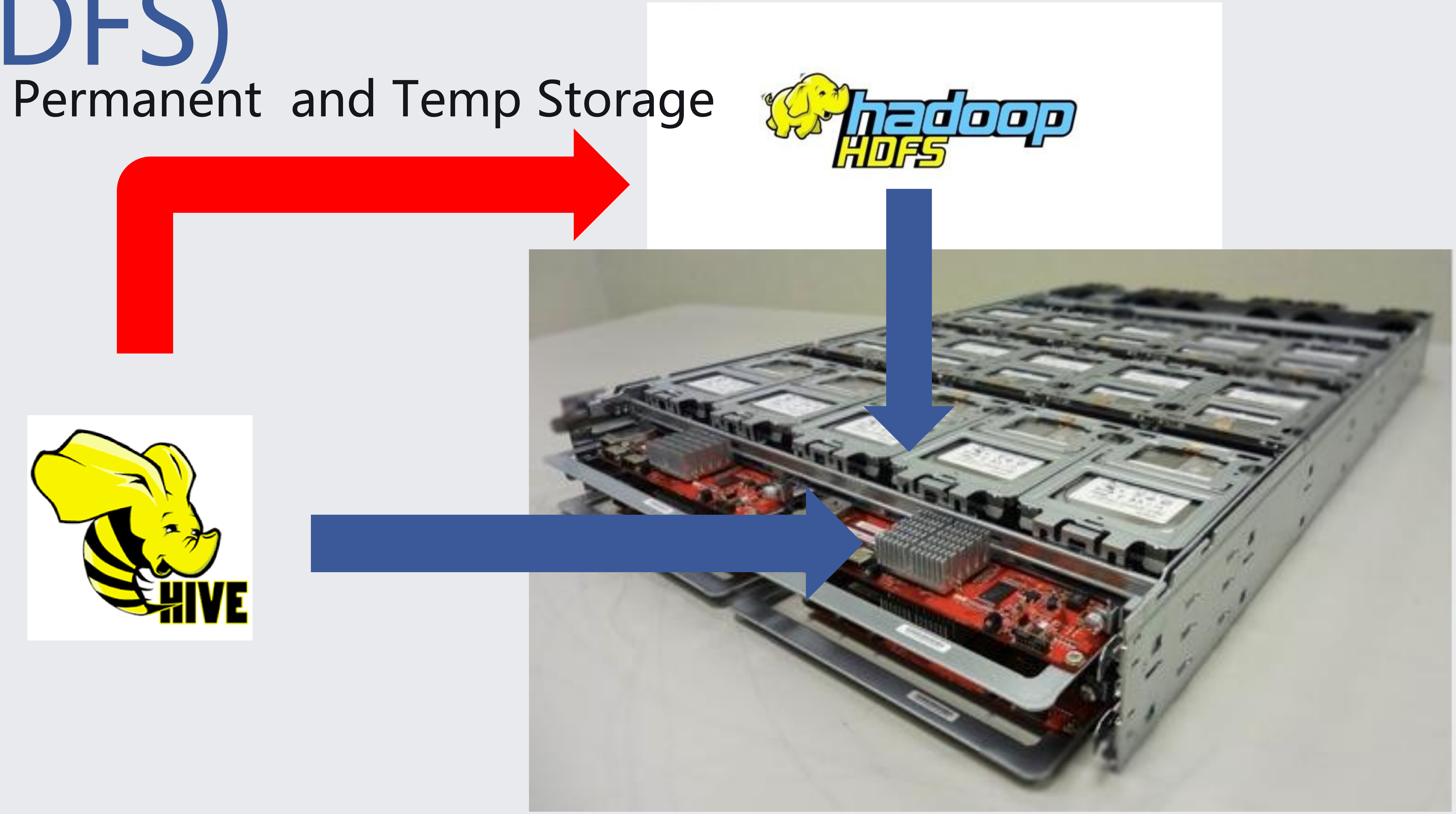
facebook

DW @FB: Hive on Hadoop (HDFS)

DW @FB: Hive on Hadoop (HDFS)

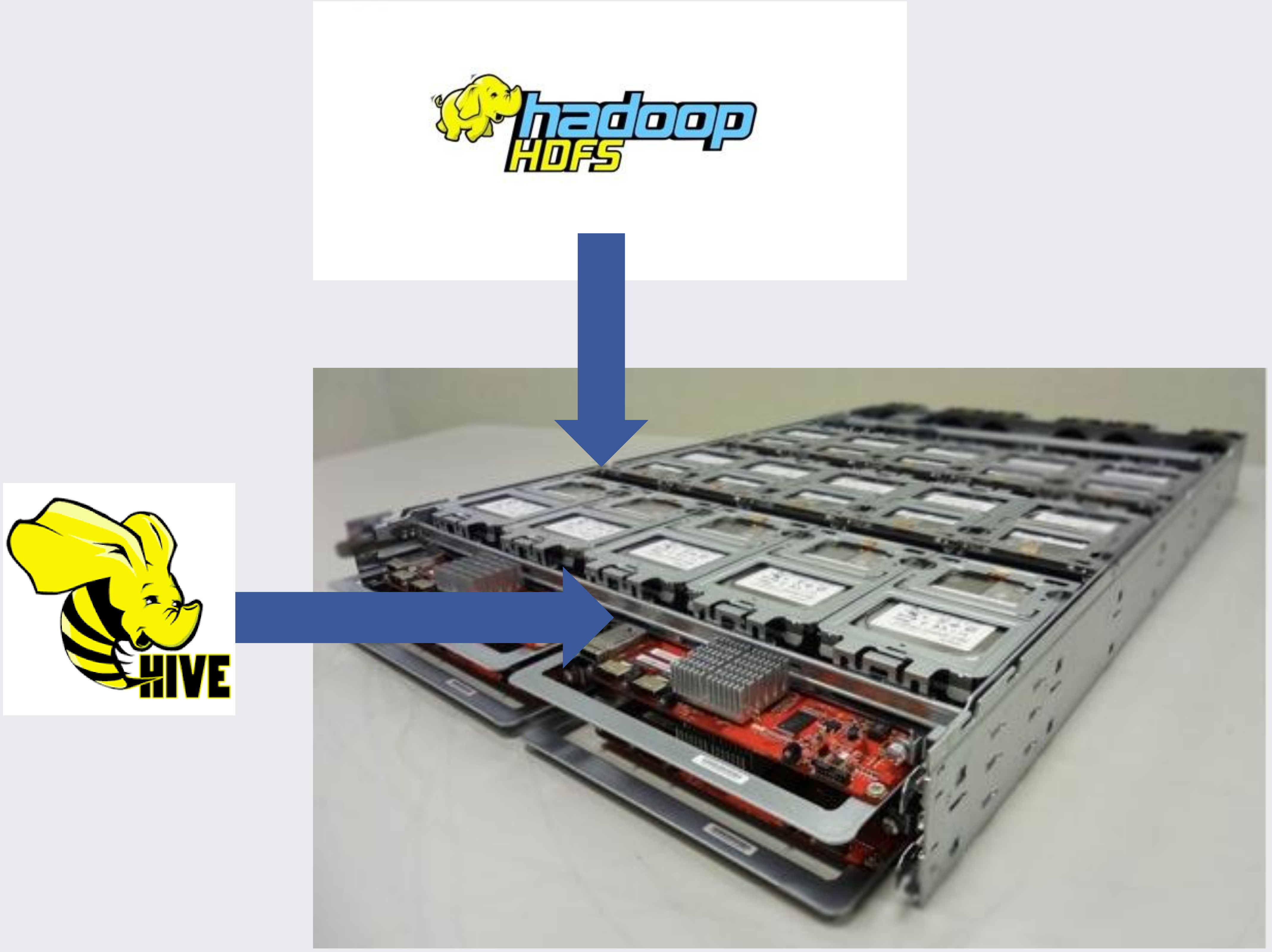


DW @FB: Hive on Hadoop (HDFS)



Permanent and Temp Storage

Dynamic demand vs. static resource allocation



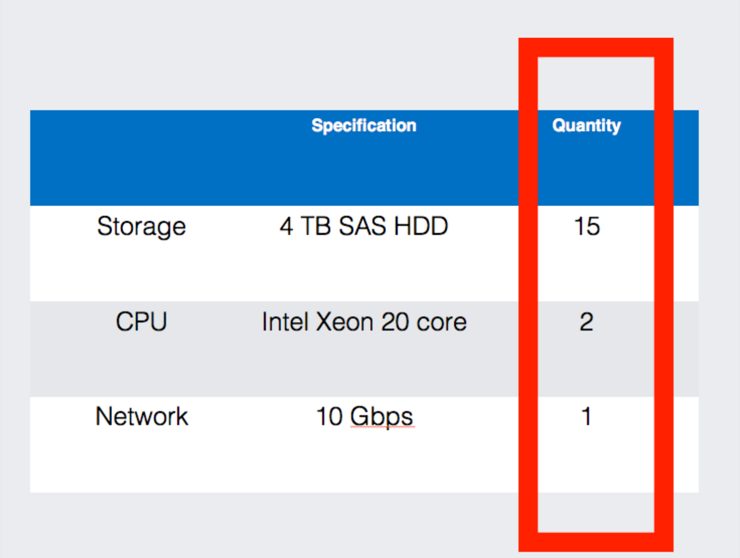
	Specification	Quantity
Storage	4 TB SAS HDD	15
CPU	Intel Xeon 20 core	2
Network	10 Gbps	1

Efficiency at hyper scale

	Specification	Quantity
Storage	4 TB SAS HDD	15
CPU	Intel Xeon 20 core	2
Network	10 Gbps	1

Elasticity

Splitting DC for Storage and Compute



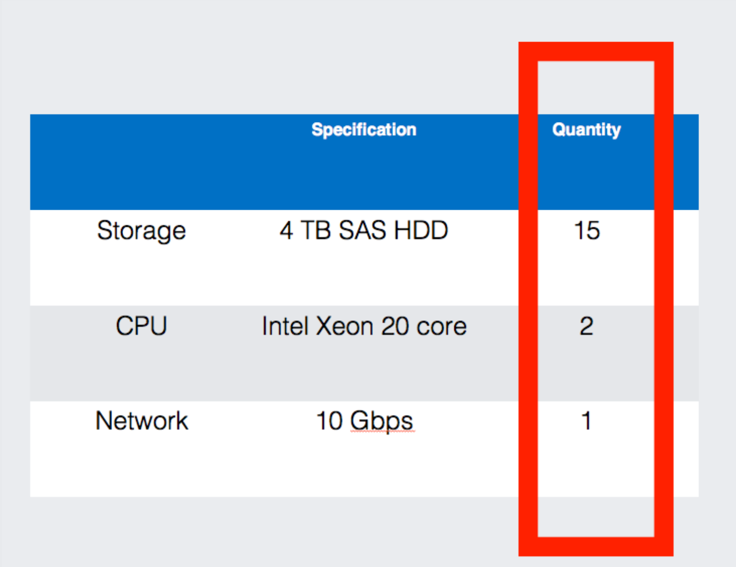
	Specification	Quantity
Storage	4 TB SAS HDD	15
CPU	Intel Xeon 20 core	2
Network	10 Gbps	1

Compute

Storage

Elasticity

Splitting DC for Storage and Compute



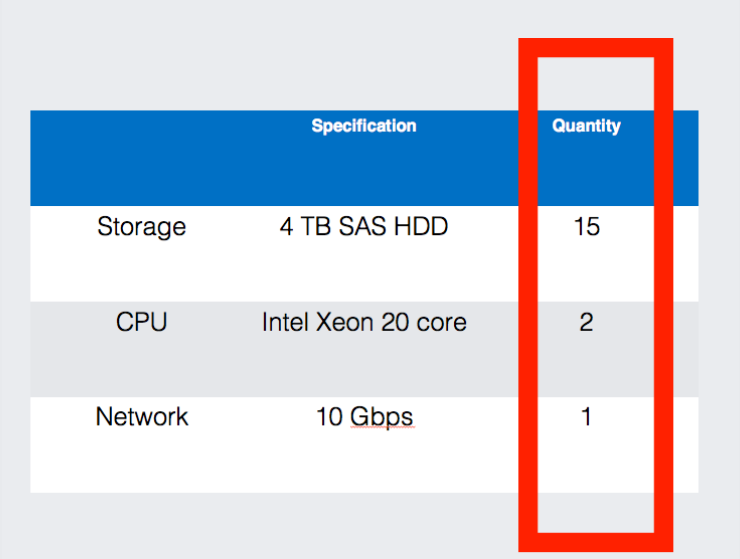
	Specification	Quantity
Storage	4 TB SAS HDD	15
CPU	Intel Xeon 20 core	2
Network	10 Gbps	1

Compute

Storage

Elasticity

Splitting DC for Storage and Compute

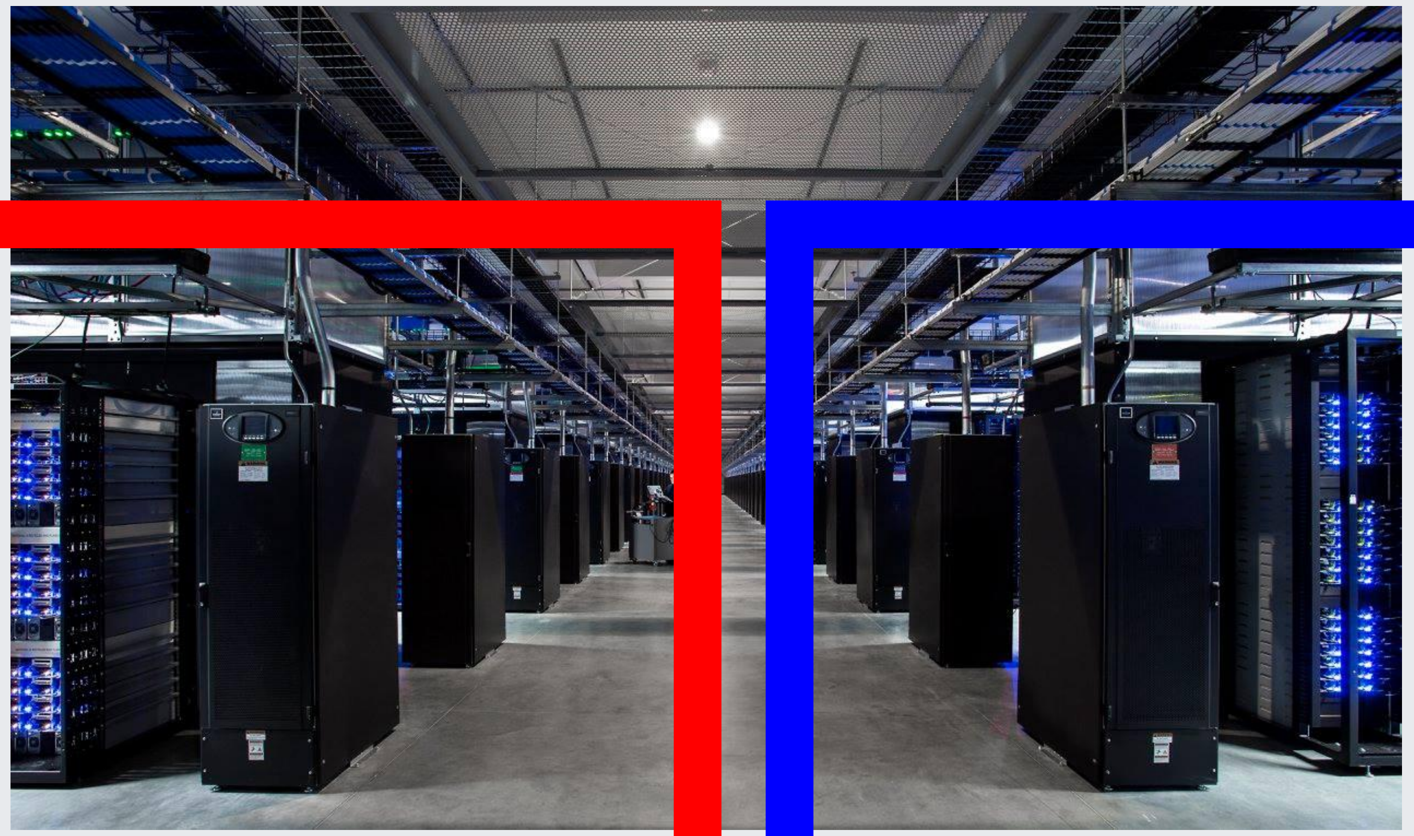


	Specification	Quantity
Storage	4 TB SAS HDD	15
CPU	Intel Xeon 20 core	2
Network	10 Gbps	1

Compute

Storage

Elasticity



Compute

- Add more compute if/when needed
- Upgrade to latest Intel’ s CPUs
- Replace compute every 3 years or sooner

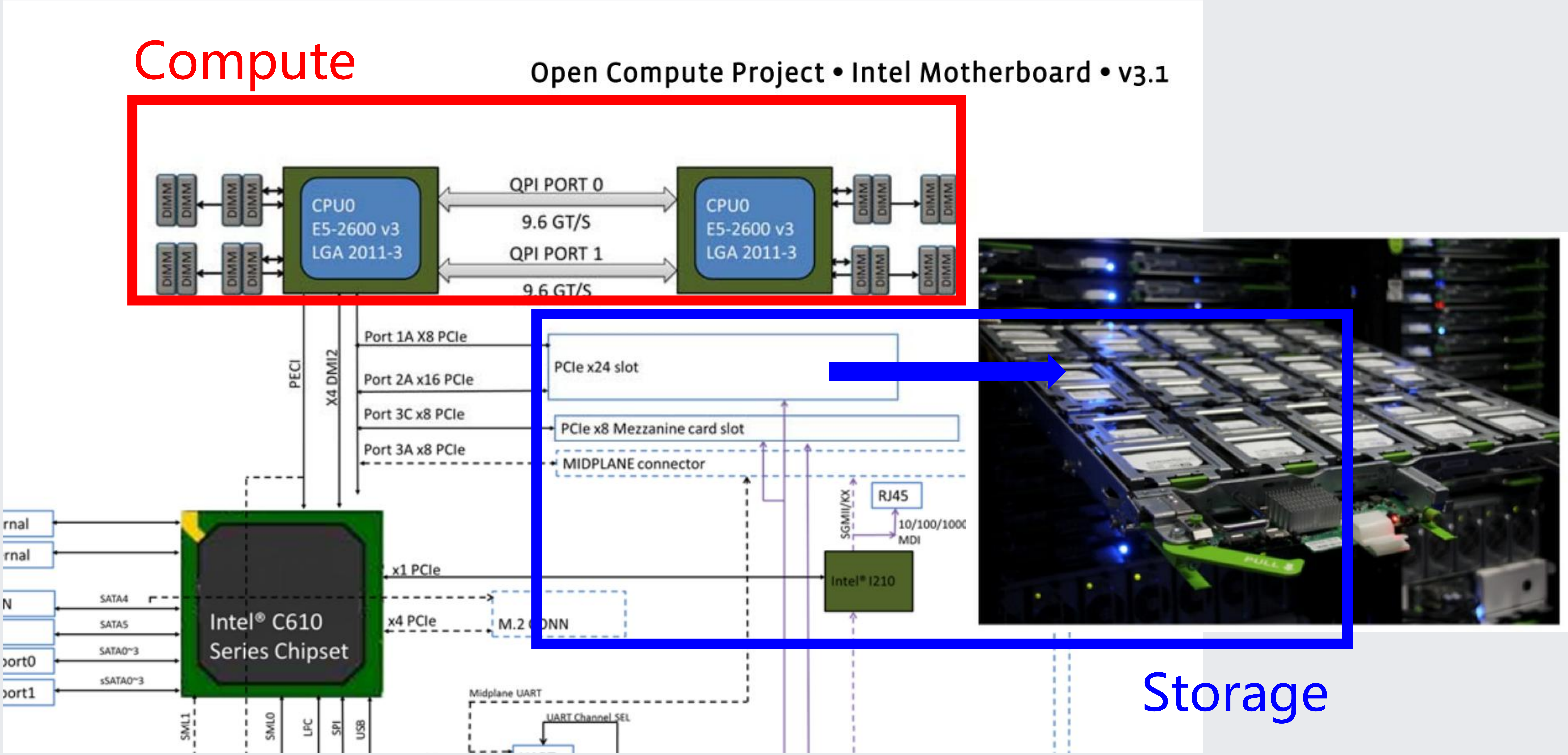
Storage

- Keep HDDs longer than 3 years
- Grow storage capacity independent from compute
- Migrate to 8-10 TB HDDs

Storage and Compute separation

Is that possible?

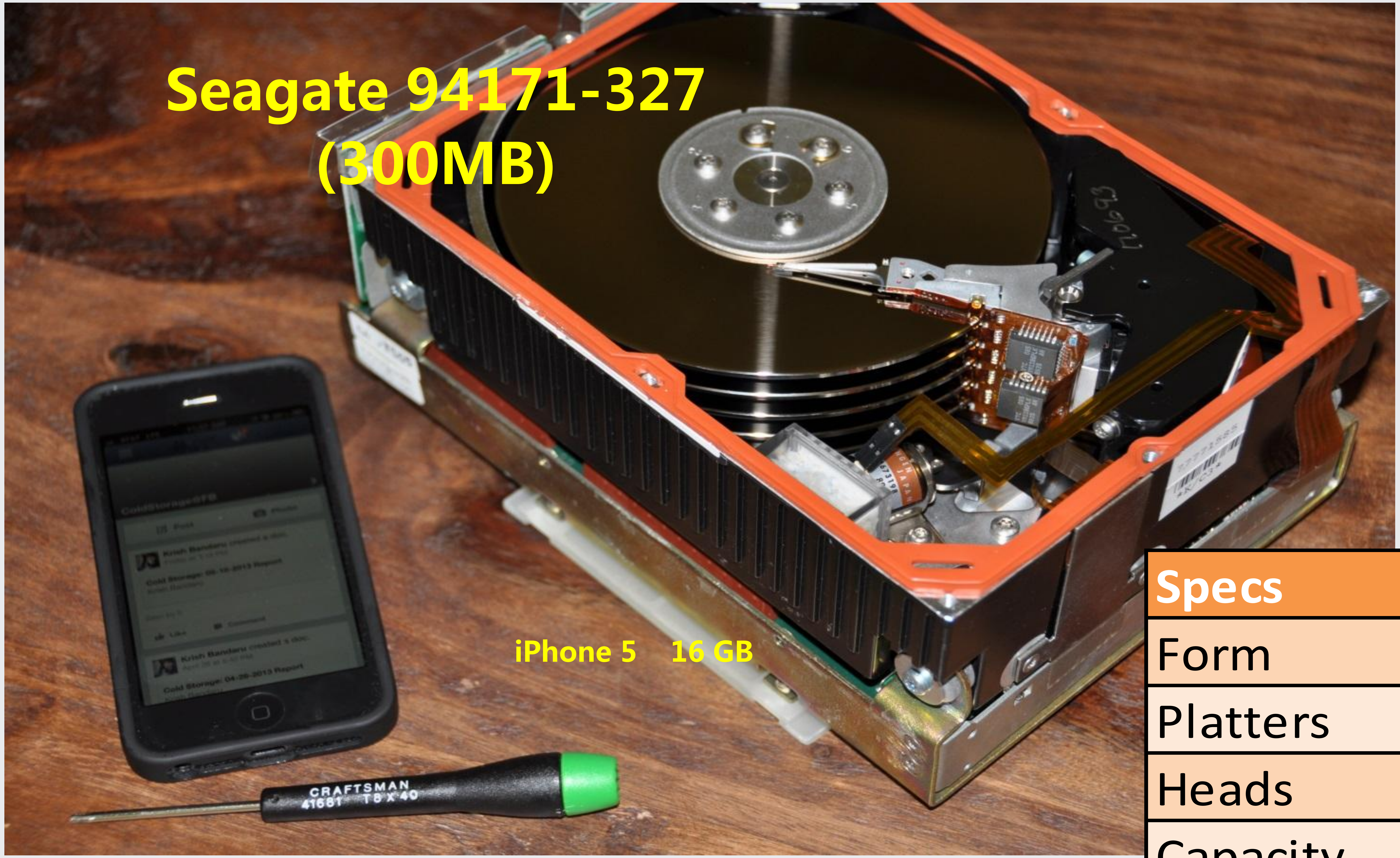
Compute and storage: local vs. remote



Compute

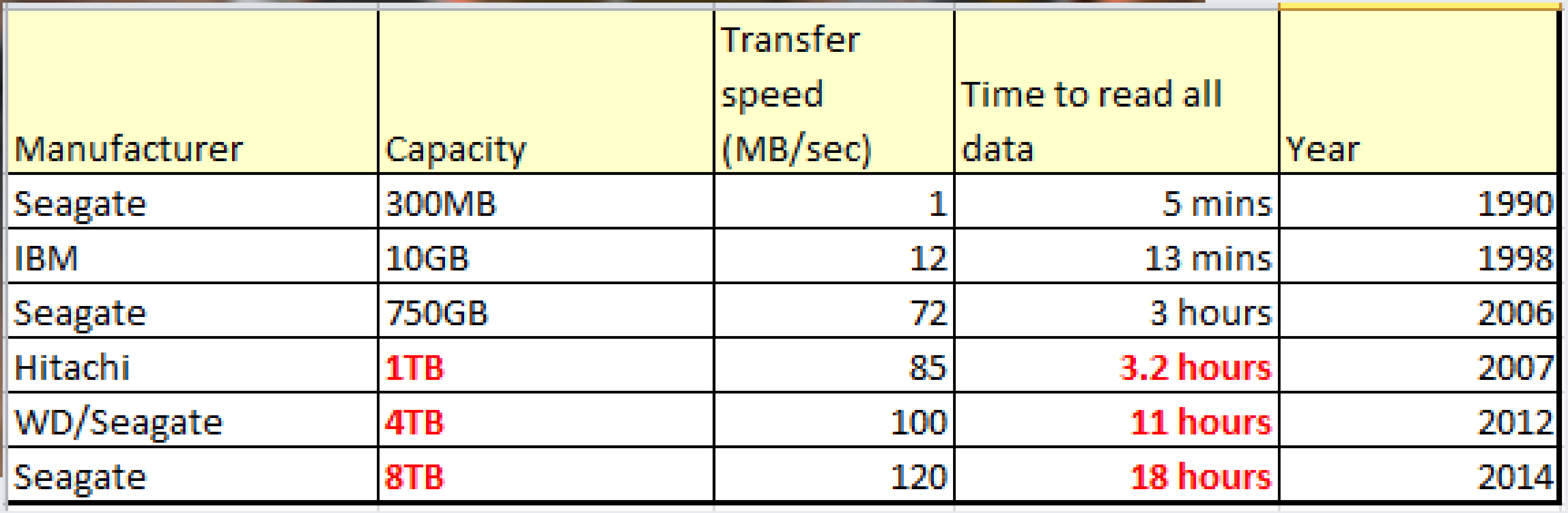
Storage

Last ~25 years of HDD evolution



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

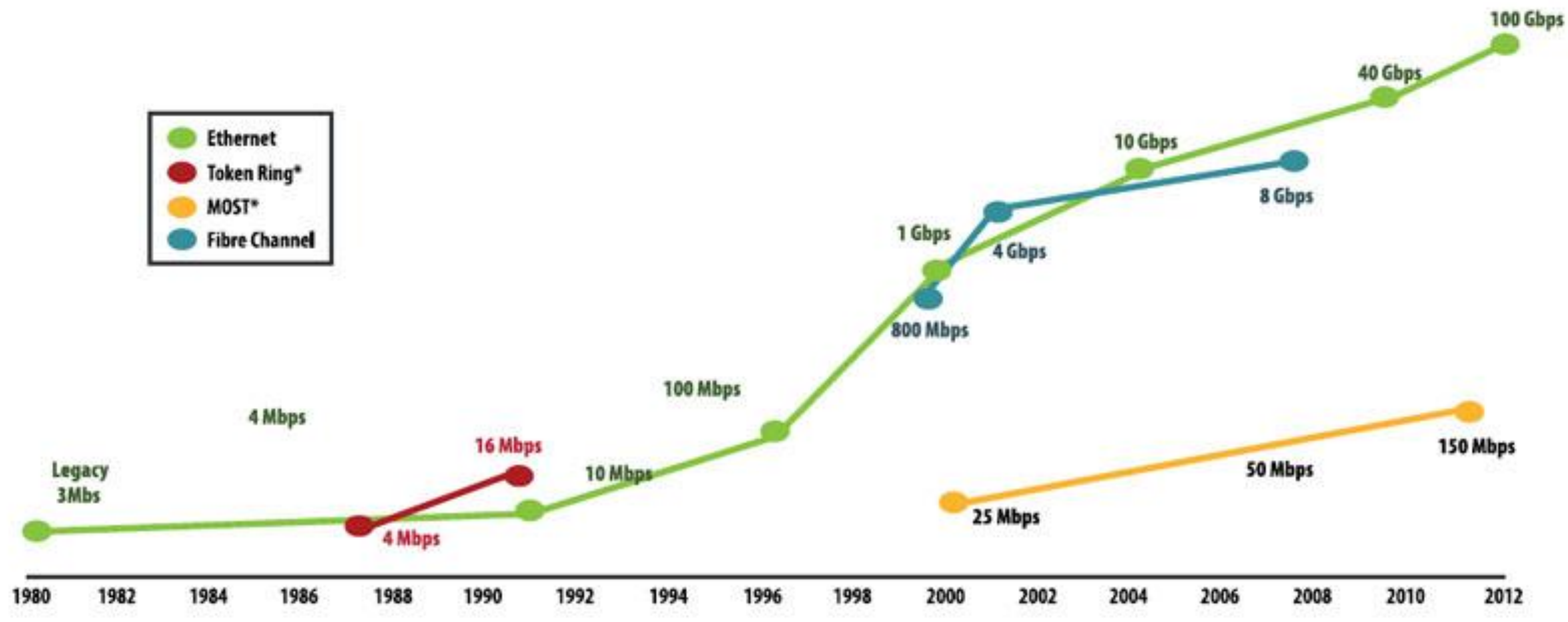
Last ~25 years of HDD evolution



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

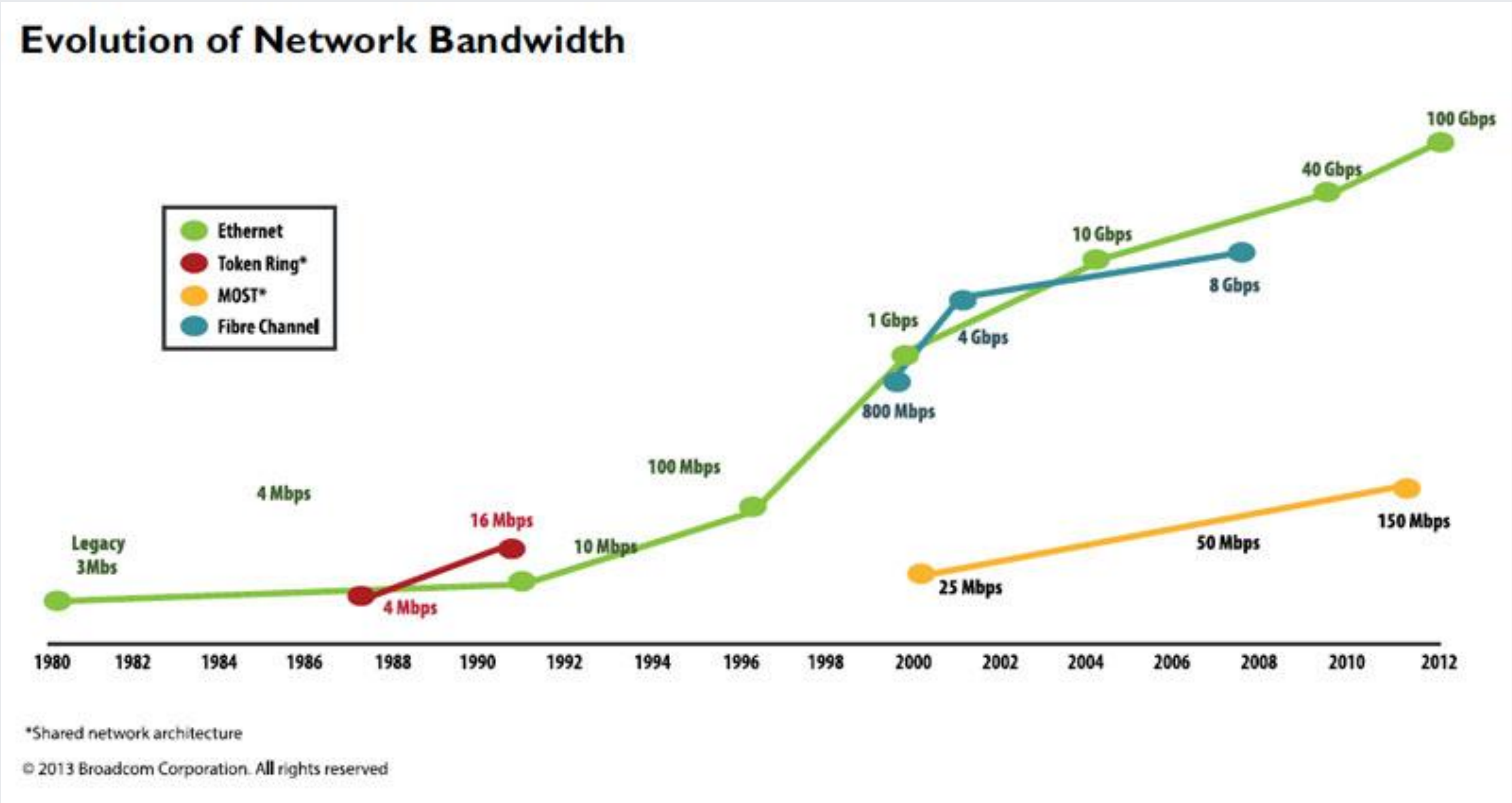
Last ~25 years of Ethernet

Evolution of Network Bandwidth



*Shared network architecture

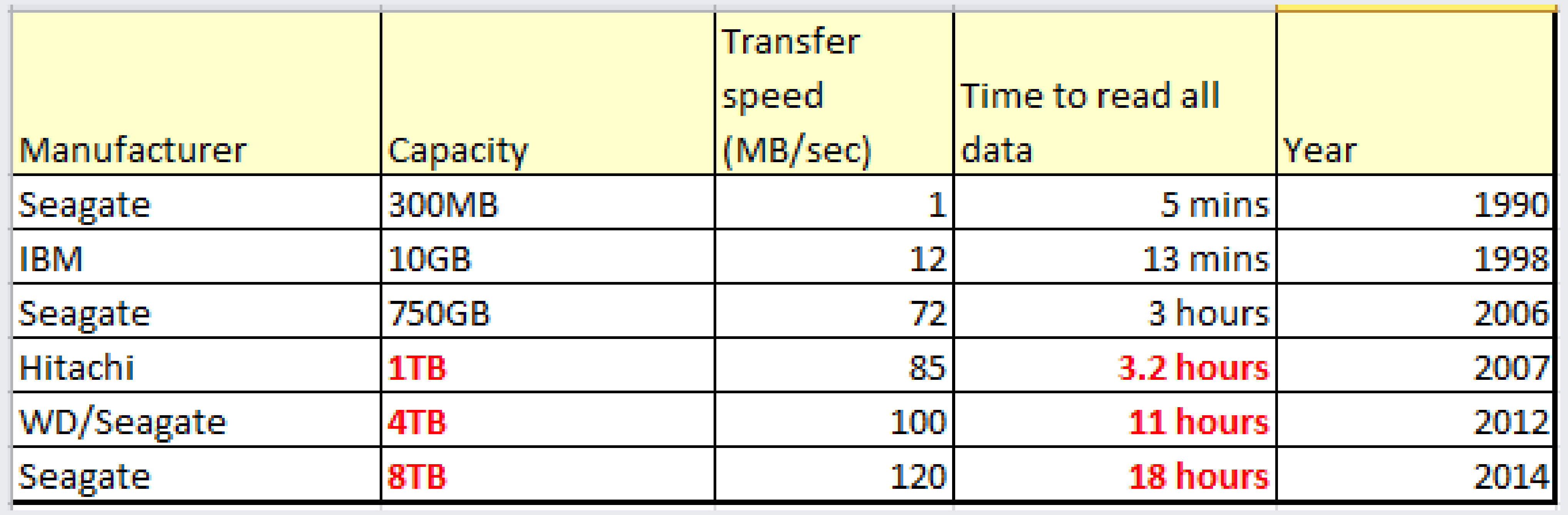
Last ~25 years of HDD vs. Ethernet



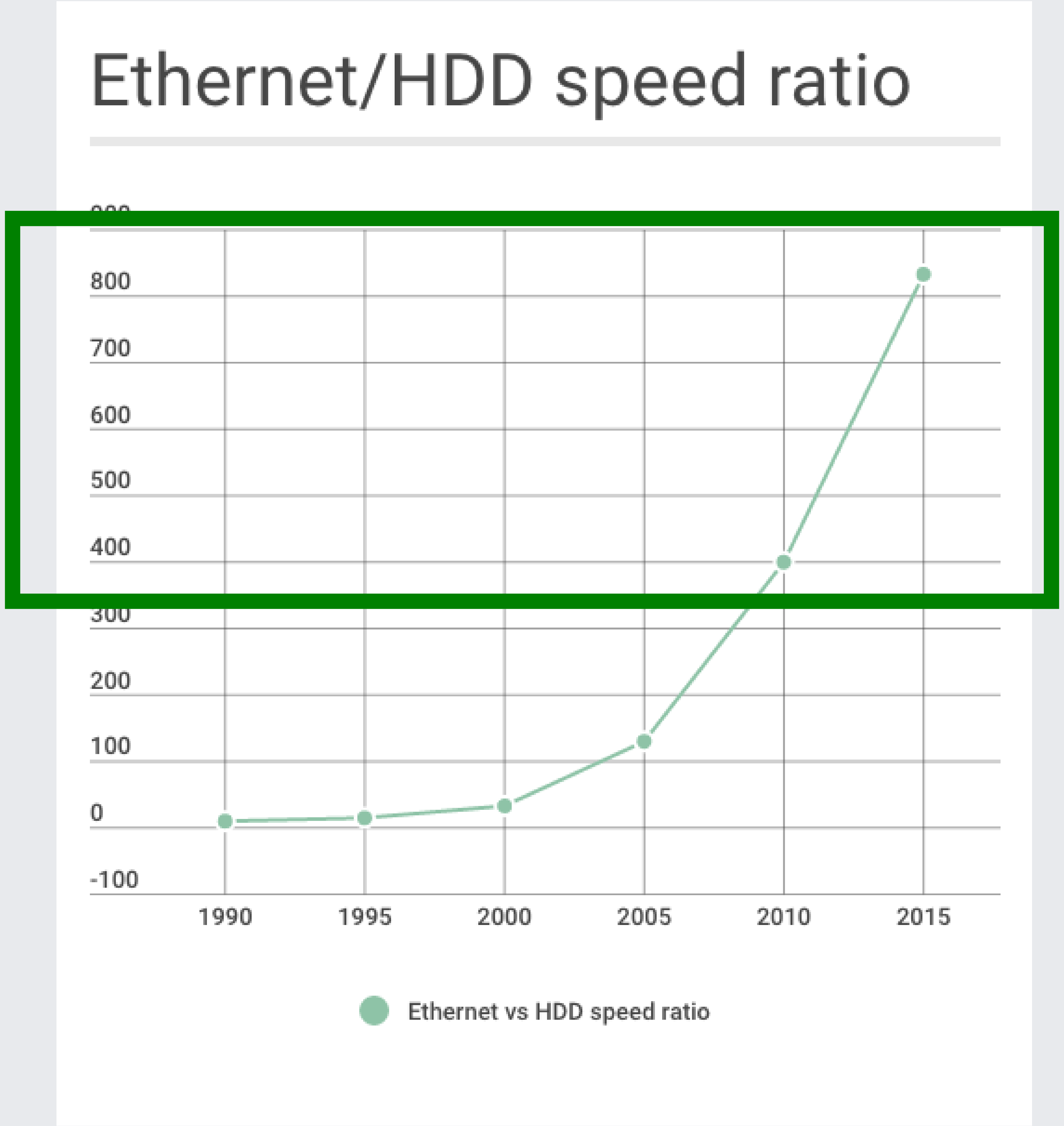
Evolution of Network Bandwidth

*Shared network architecture

© 2013 Broadcom Corporation. All rights reserved



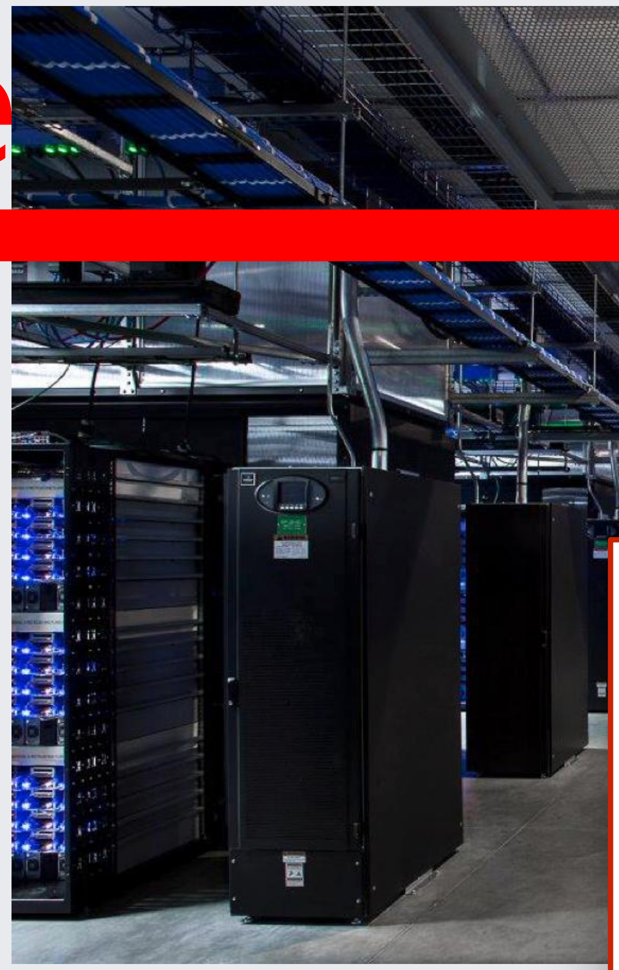
Manufacturer	Capacity	Transfer speed (MB/sec)	Time to read all data	Year
Seagate	300MB	1	5 mins	1990
IBM	10GB	12	13 mins	1998
Seagate	750GB	72	3 hours	2006
Hitachi	1TB	85	3.2 hours	2007
WD/Seagate	4TB	100	11 hours	2012
Seagate	8TB	120	18 hours	2014

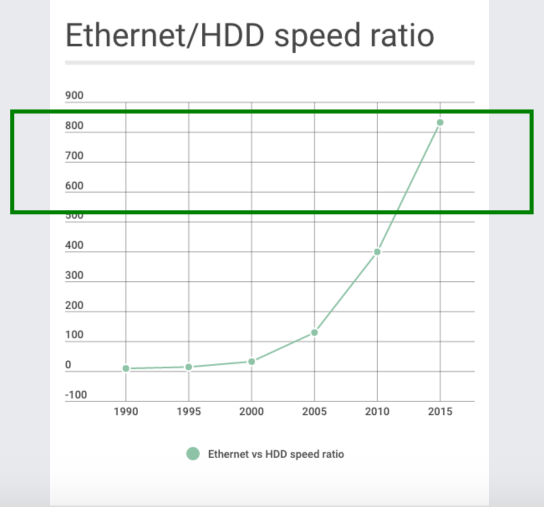


Ethernet/HDD speed ratio

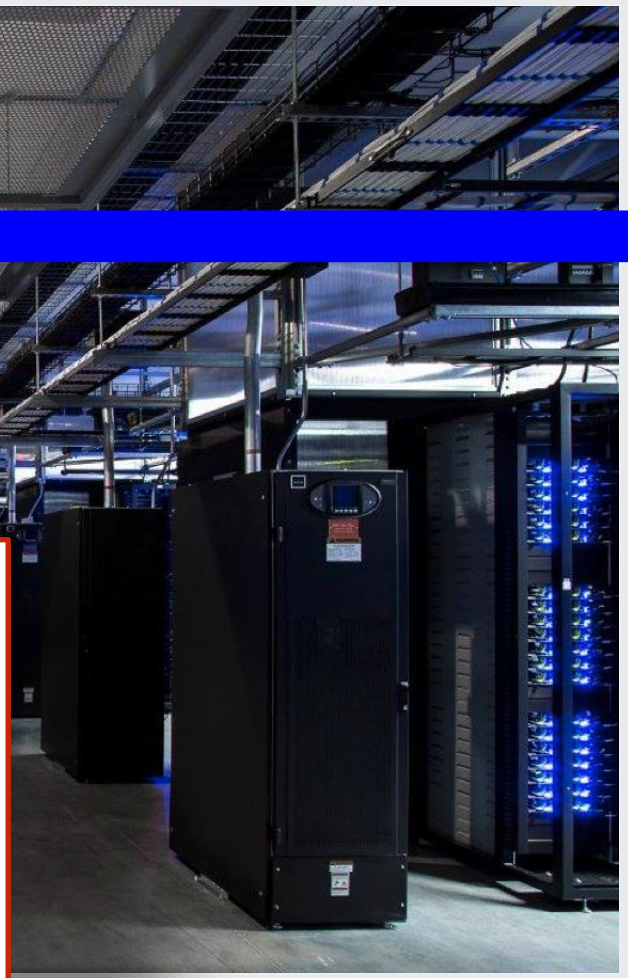
When Ethernet is faster than HDDs And Data Center network is ‘flat’

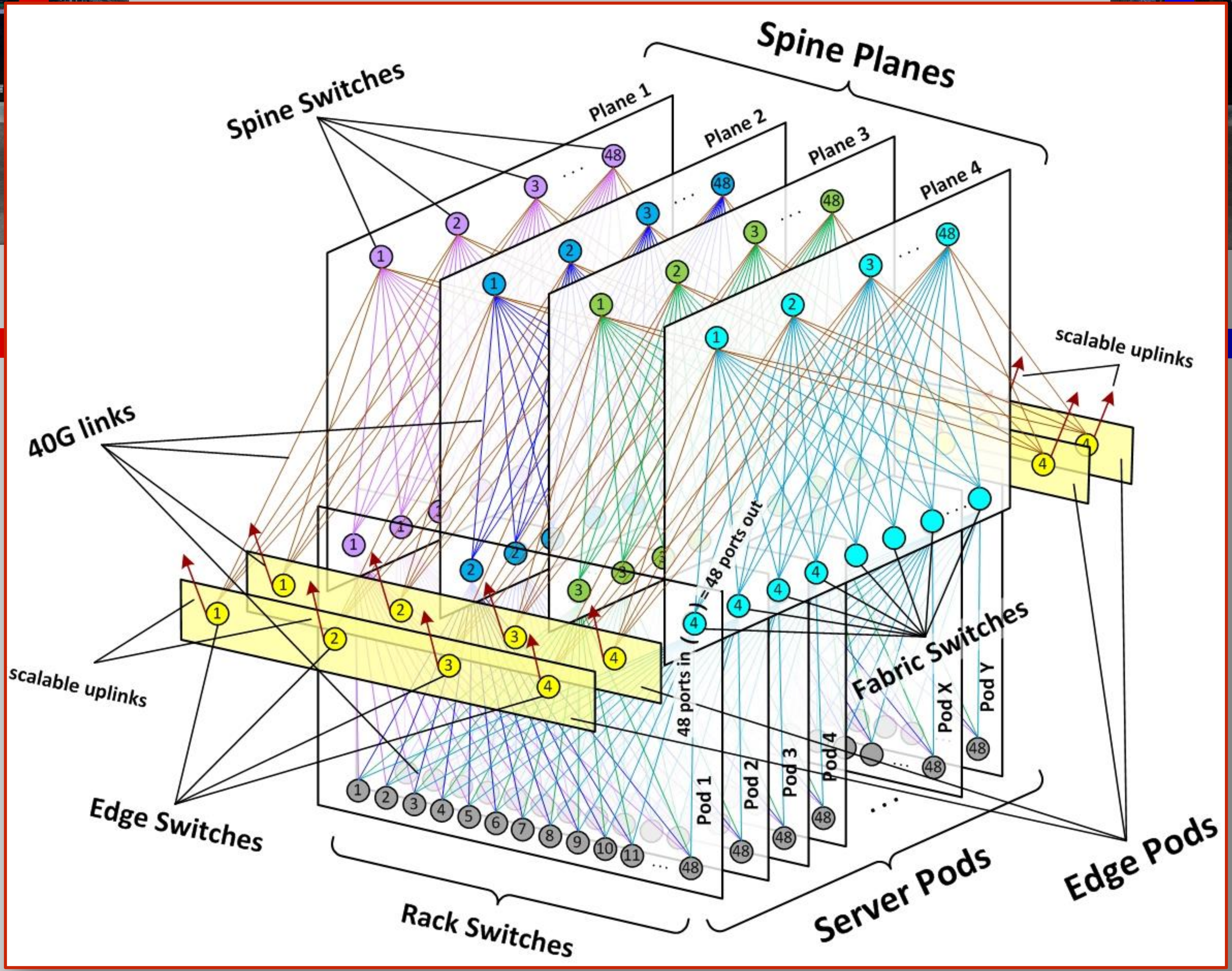
Compute





Storage





facebook

Warm Storage

Architecture

Metadata calls
Metadata plane
Control plane
1 +{N}
Directory Service
MySQL for Directories
N
File Service
Sharded MySQL or ZippyDB
N
Block Pool Service
Sharded MySQL or ZippyDB
Client
Original data encoded vie Reed-Solomon
StatServer
N
Rebalance workers
Rebalance planner
N
AntiEntropy Scanners
CDB for metadata
N
GC
N
Repair
N
Volume Allocator
N
Storage Monitor
Thrift: Write/Read/Scan Chunk, Get inventory
N data streams
Data plan
Storage Node #1
Storage Service
Block GC
Storage Node #N
Storage Service
Block GC
. . .

Facebook Warm Storage Architecture
All services

Meta a calls
Metadata plane
1 +{N}
Directory Service
MySQL for Directories
N
File Service
Sharded MySQL or ZippyDB
N
Block Pool Service
Sharded MySQL or ZippyDB
Client
Original data encoded vie Reed-Solomon
Control plane
StatServer
N
Rebalance workers
Rebalance planner
N
AntiEntropy Scanners
CDB for metadata
N
GC
N
Repair
N
Volume Allocator
N
Storage Monitor
Thrift: Write/Read/Scan Chunk, Get inventory
N data streams
Data plane
Storage Node #1
Storage Service
Block GC
. . .
Storage Node #N
Storage Service
Block GC

Facebook Warm Storage Architecture
All services

Facebook Warm Storage Architecture
All services

Metadata calls

Client

Original data encoded vie Reed-Solomon

Metadata plane

Directory Service 1 +{N}

MySQL for Directories

File Service N

Sharded MySQL or ZippyDB

Block Pool Service N

Sharded MySQL or ZippyDB

Control pla

StatServer

Rebalance workers N

Rebalance planner

AntiEntropy Scanners N

CDB for metadata

GC N

Repair N

Volume Allocator N

Storage Monitor N

Thrift: Write/Read/Scan Chunk, Get inventory

N data streams

Data plane

Storage Node #1

Storage Service

Block GC

. . .

Storage Node #N

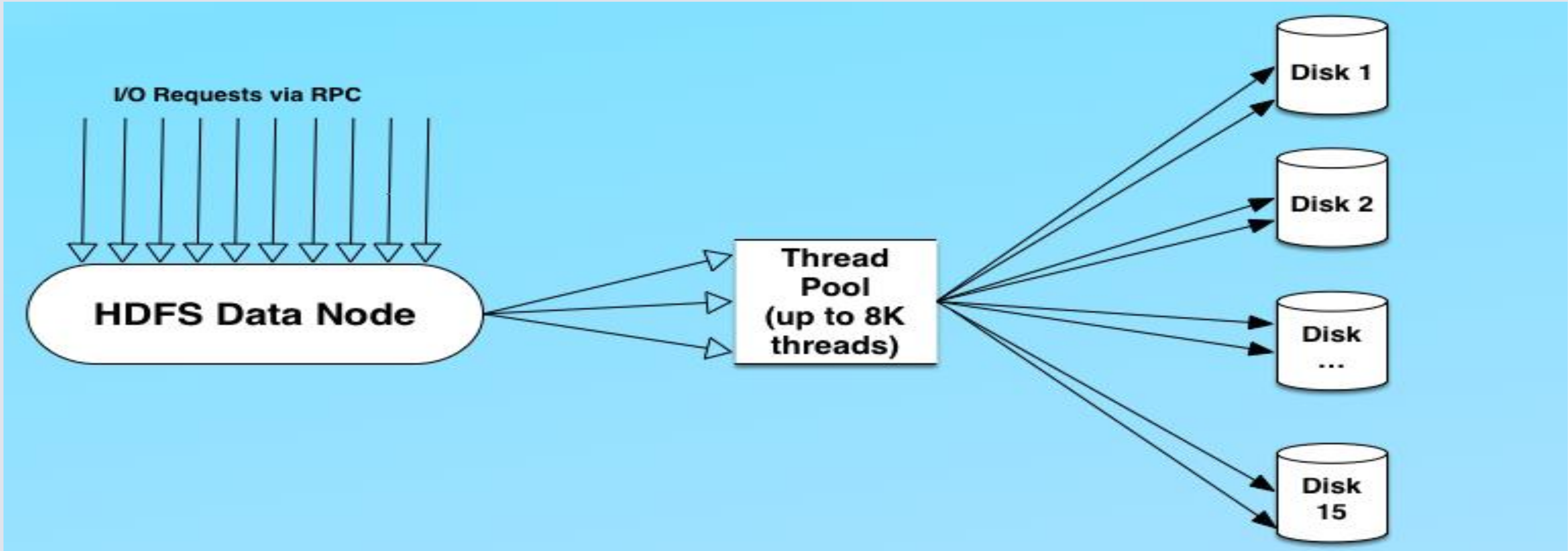
Storage Service

Block GC

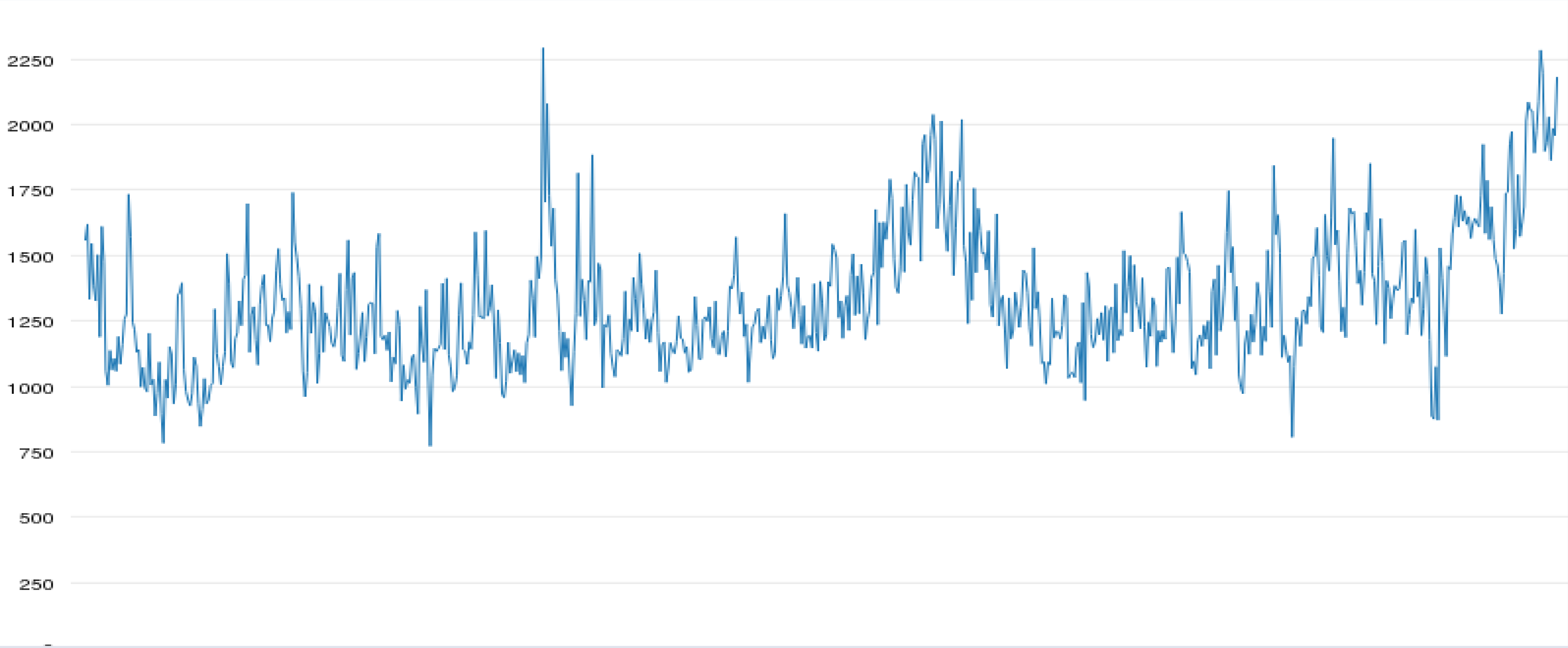
Storage Service IO Model

HDFS Data Node I/O Model

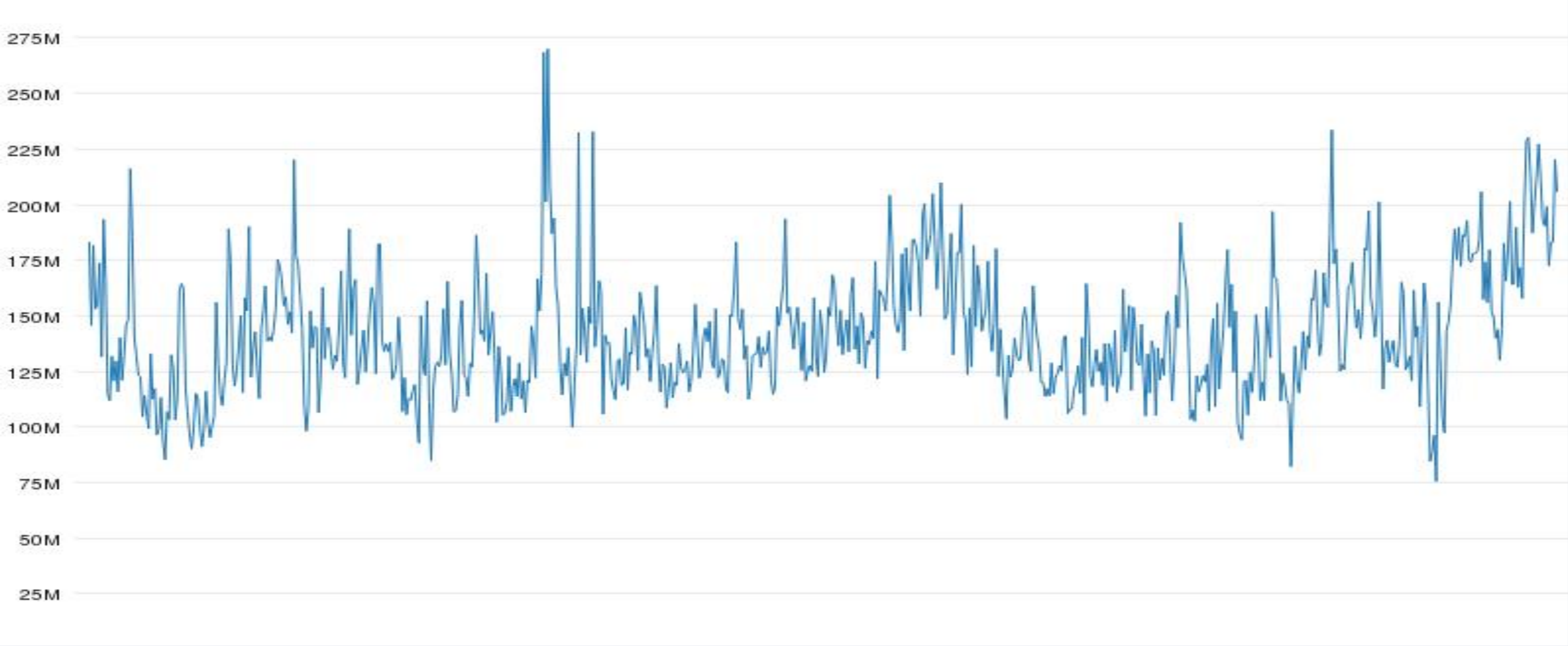
- HDFS Disk Thrashing
  - 4k – 8K I/O threads



HDFS IOPS/Node – very large cluster

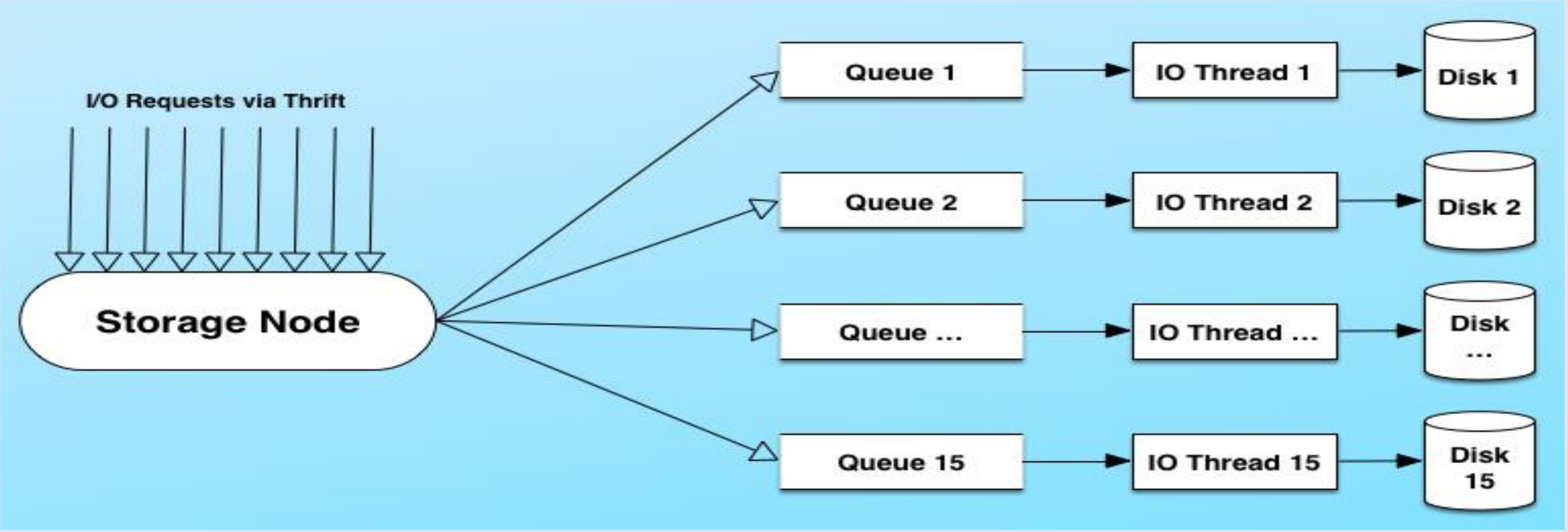


HDFS (MB/s per Node – 15 disks)



Storage Node Single Thread I/O Model

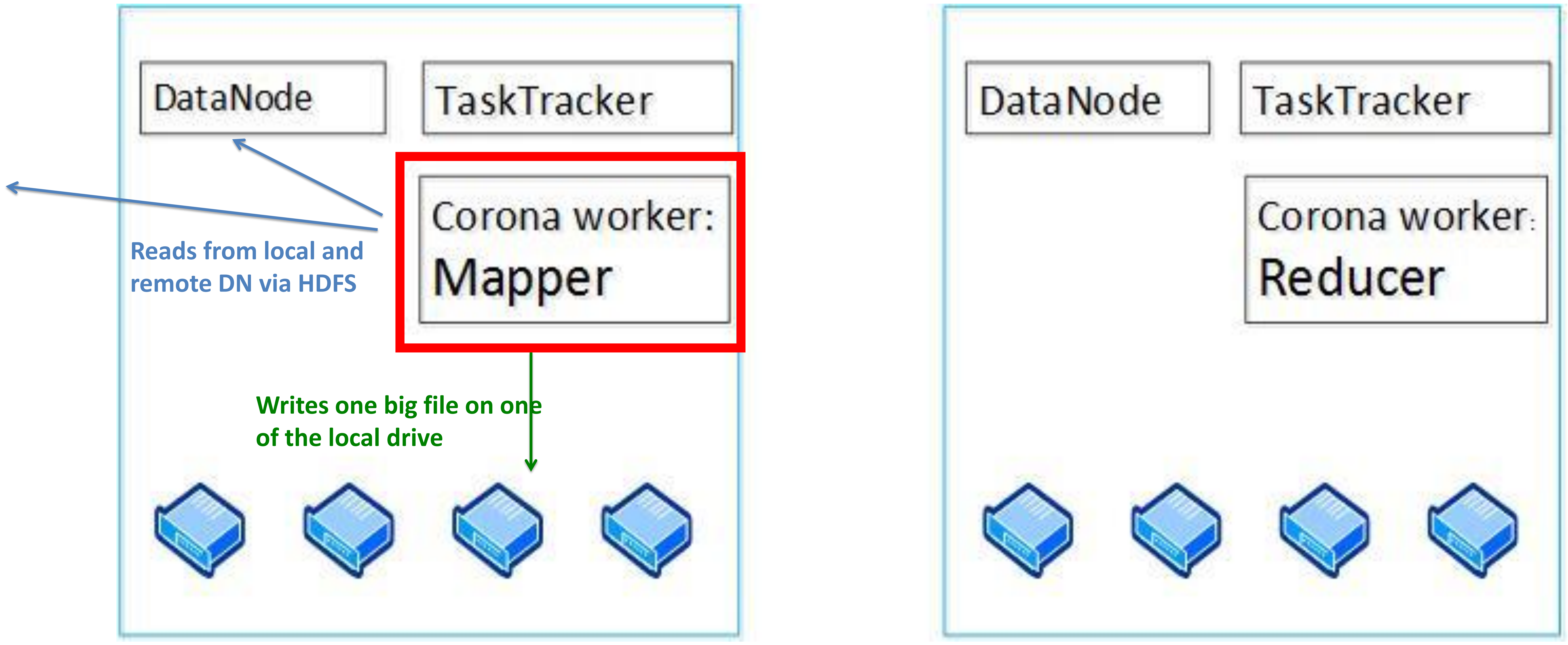
- 1 thread per disk
- I/O operations are pushed to a priority queue



facebook

TempFS and Dis Aggregated storage

Co-located HDFS Map Reduce



DataNode

TaskTracker

Corona worker:

Mapper

Reads from local and remote DN via HDFS

Writes one big file on one of the local drive

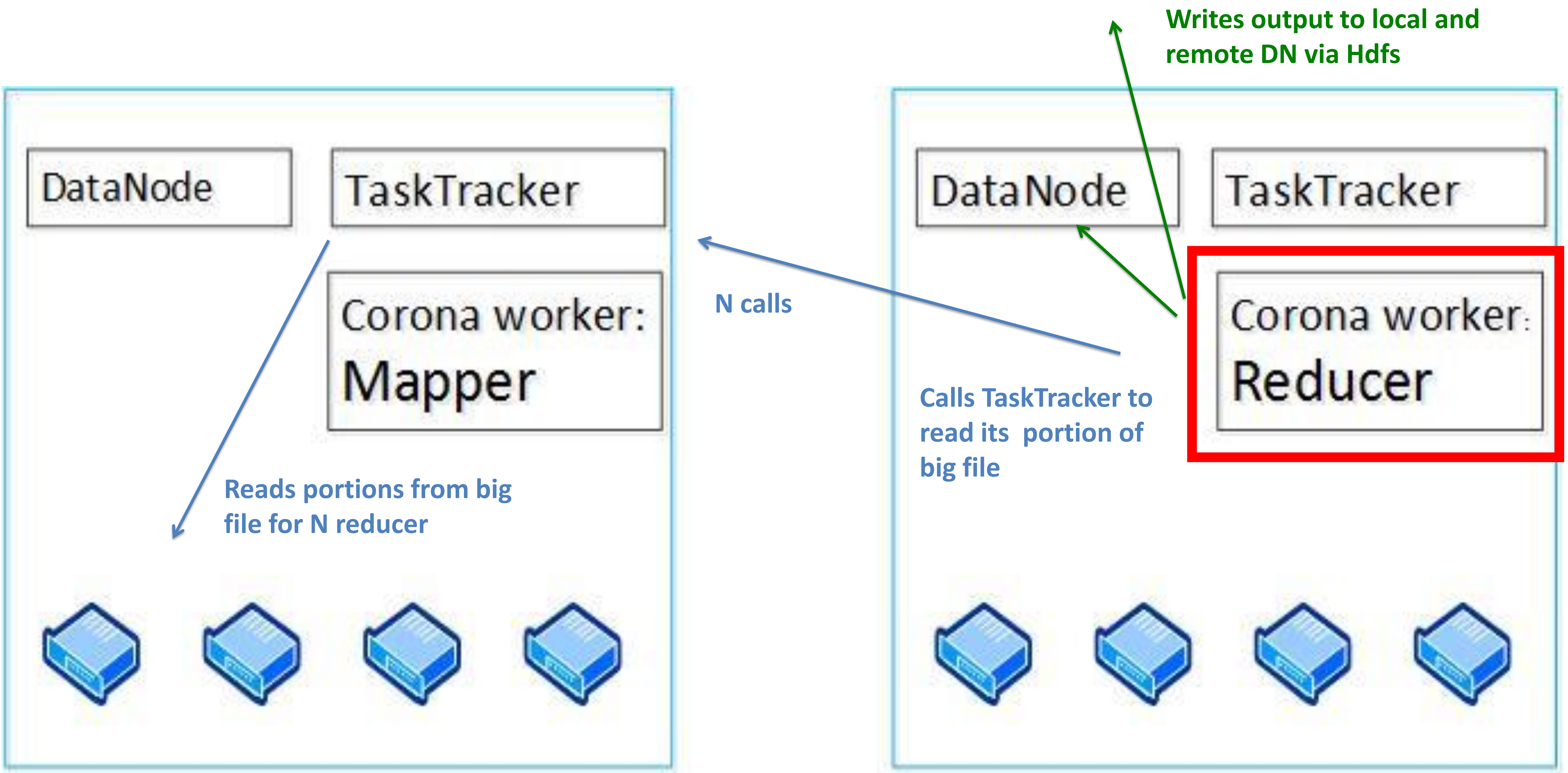
DataNode

TaskTracker

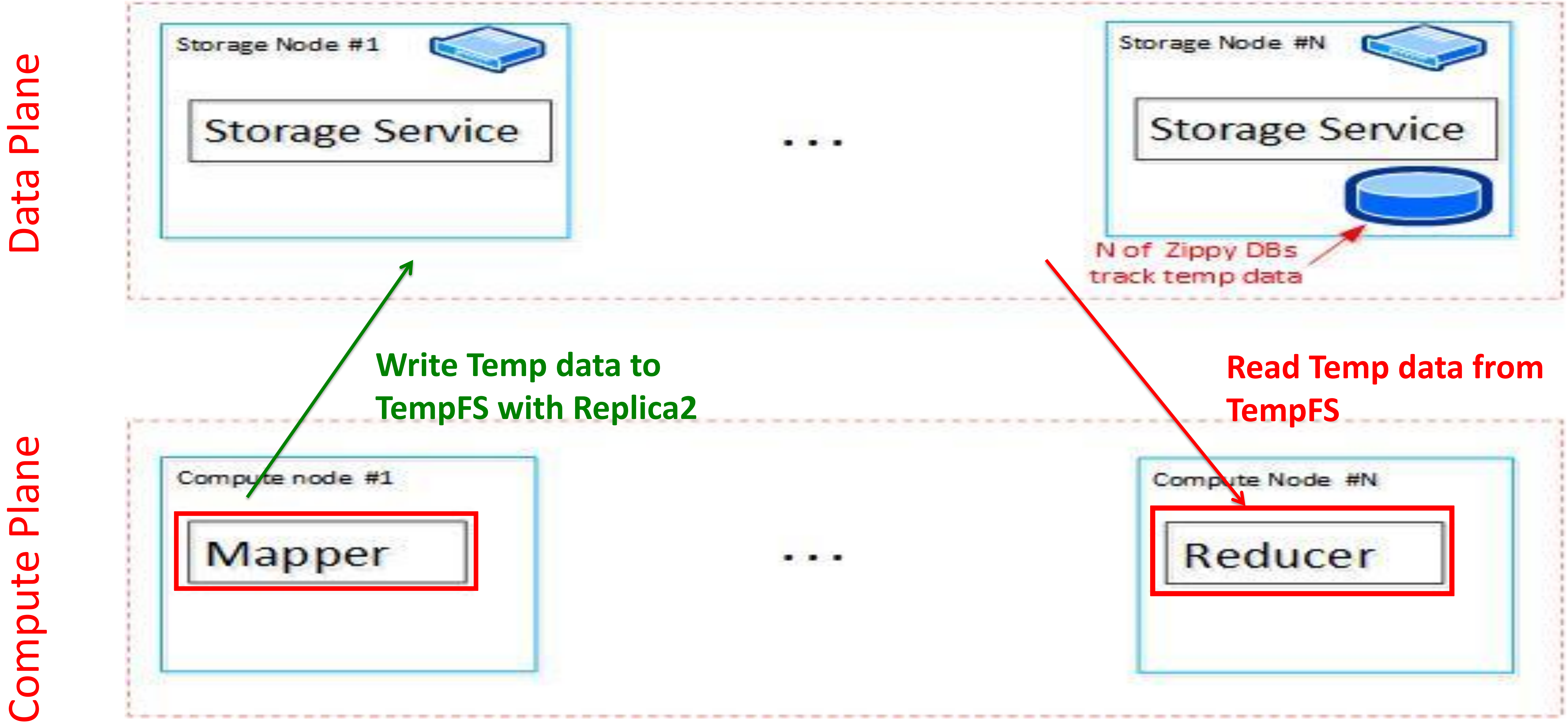
Corona worker:

Reducer

Co-located HDFS Map Reduce


Writes output to local and remote DN via Hdfs
DataNode
TaskTracker
Corona worker:
Mapper
Reads portions from big file for N reducer
N calls
DataNode
TaskTracker
Corona worker:
Reducer
Calls TaskTracker to read its portion of big file

Disaggregated Map Reduce



Data Plane

Storage Node #1

Storage Service

. . .

Storage Node #N

Storage Service

N of Zippy DBs track temp data

Write Temp data to TempFS with Replica2

Read Temp data from TempFS

Compute Plane

Compute node #1

Mapper

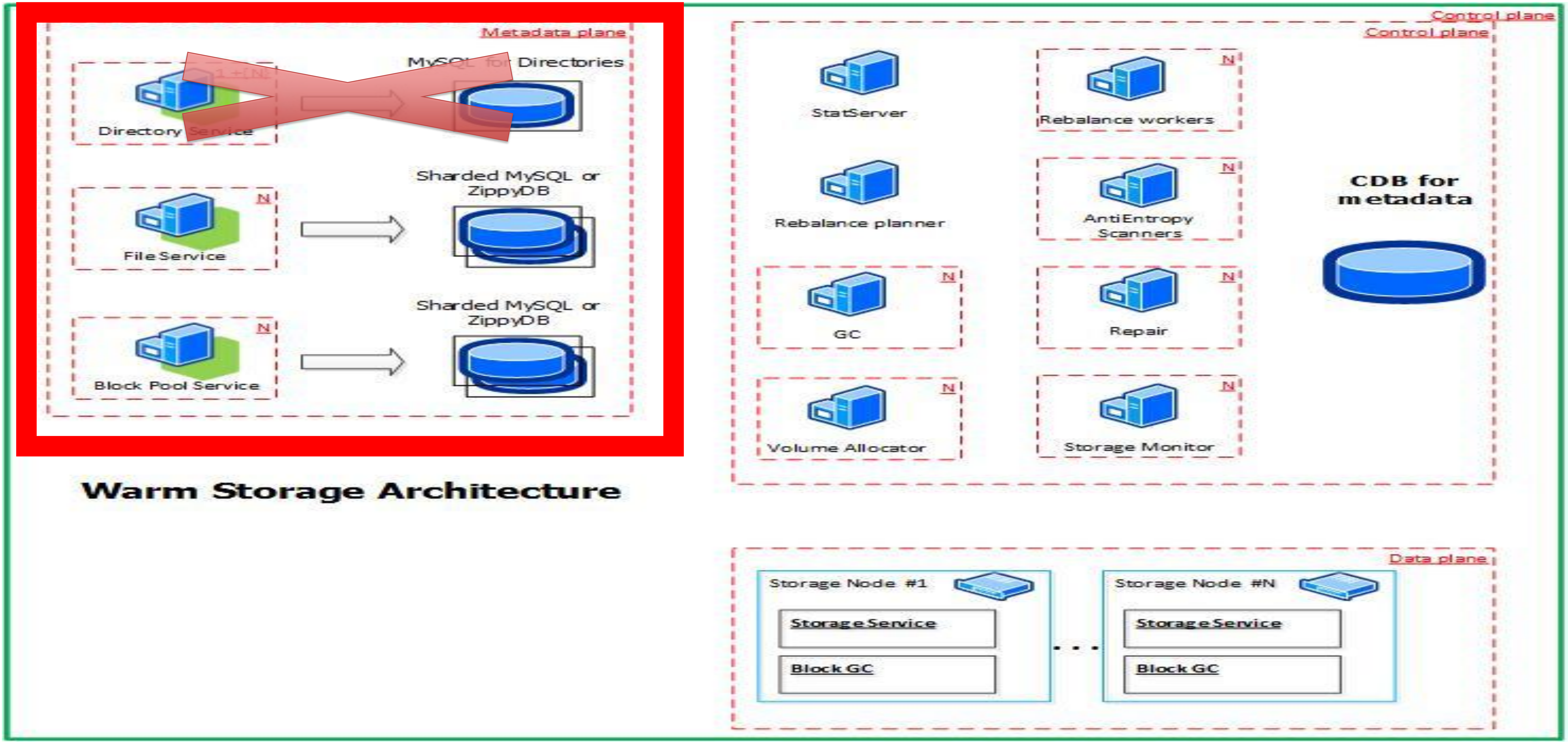
. . .

Compute Node #N

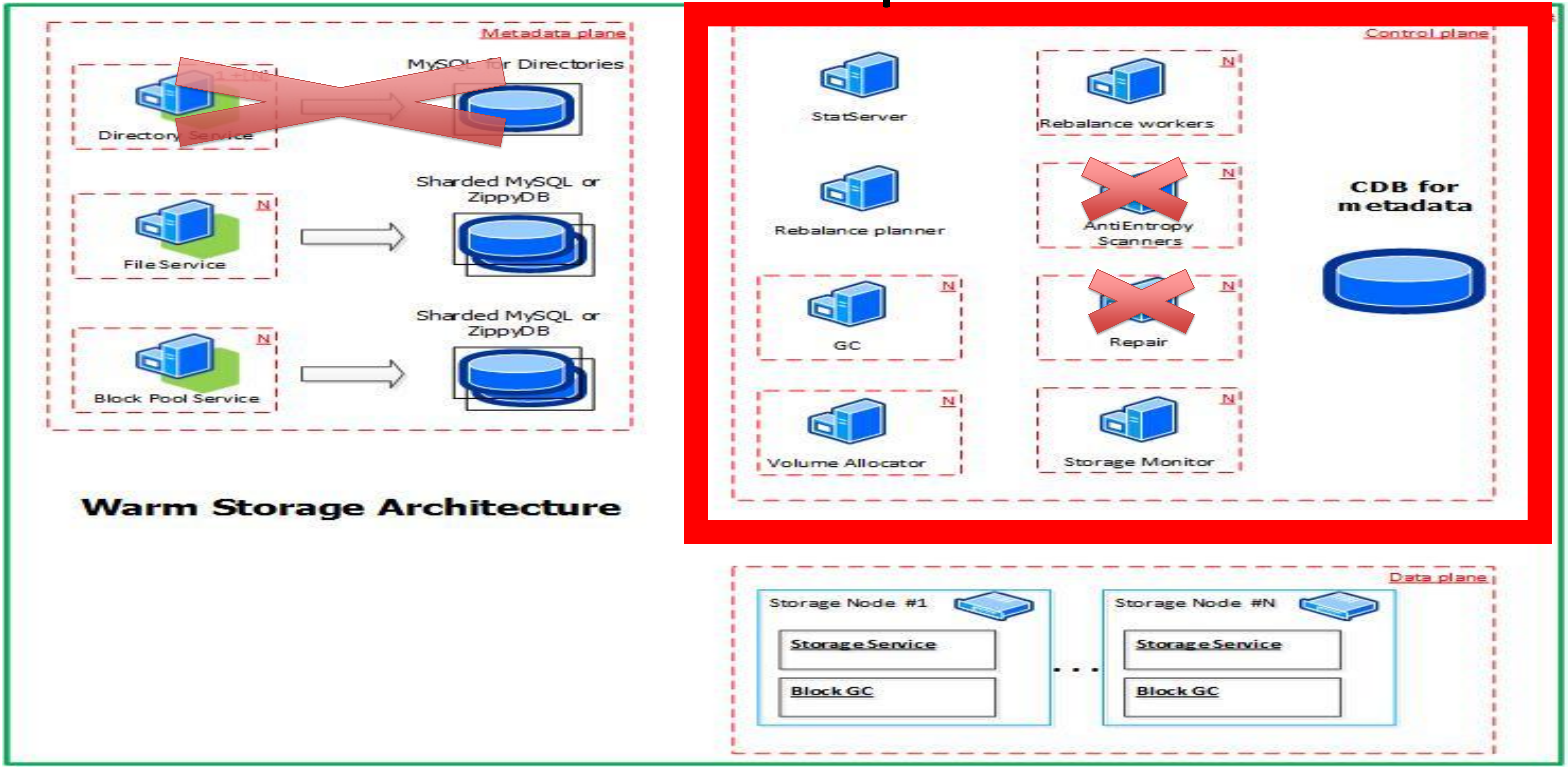
Reducer

TempFS Architecture

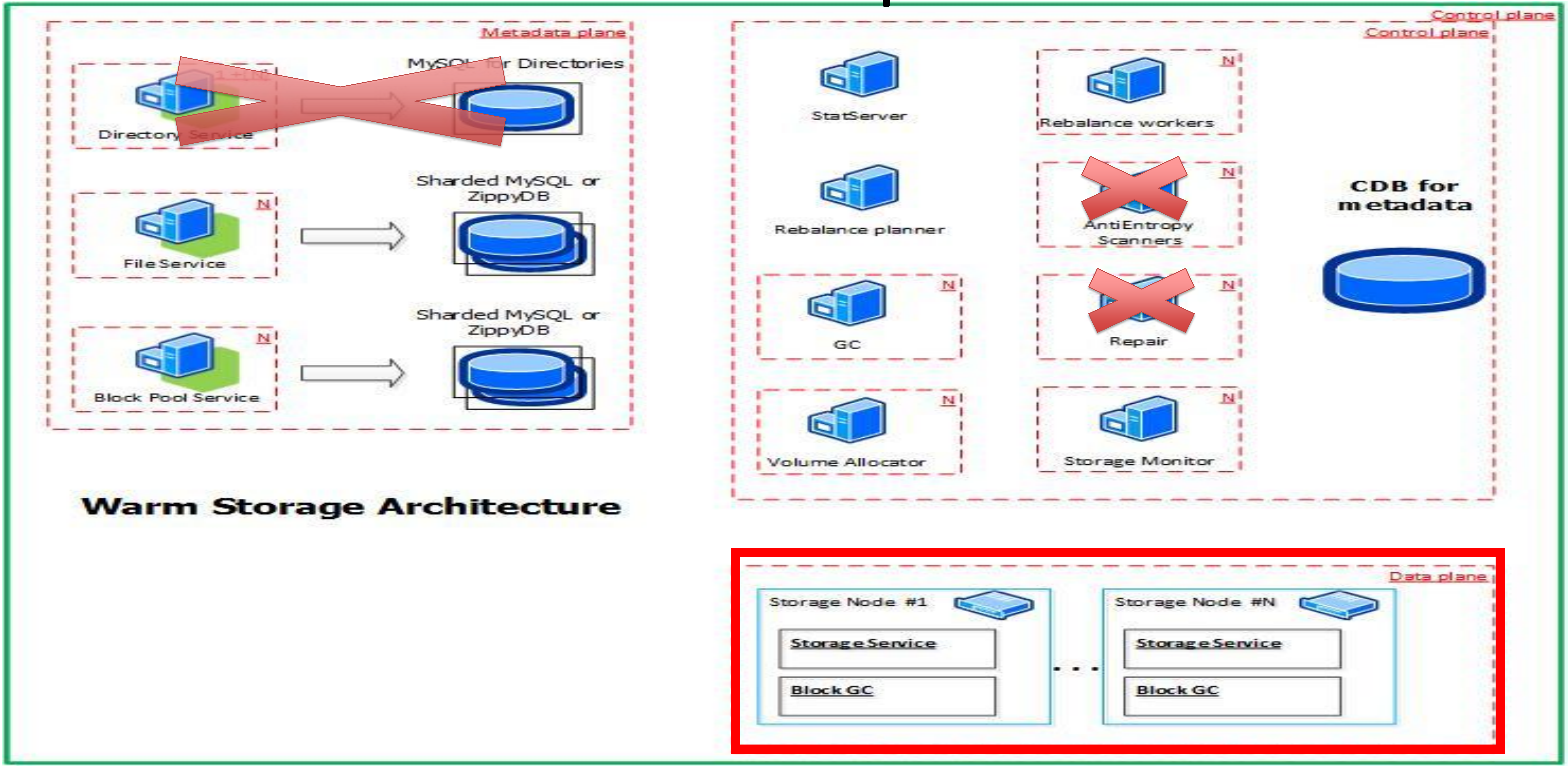
Changes in Warm Storage Architecture for TempFS



Changes in Warm Storage Architecture for TempFS



Changes in Warm Storage Architecture for TempFS



Future Work

- Streaming protocol
- Optimizing corona for Disagg

Conclusions

• Conclusions

1. The end is close for Moore’s and Kryder’s laws
2. But networking is still improving
3. Efficiency at hyper scale is hard

4. Storage and Compute separation gives better choices and helps with efficiency