facebook

An Optical Journey:

Building the largest optical archival data storage system at Facebook

Kestutis Patiejunas (kestutip@fb.com)
Sam Merat (sammerat@fb.com)

Agenda

- Archival at Facebook: context
  - Facebook’s scale
  - Implications for archival & media
- The optical journey at Facebook
  - Why optical
  - History
- Conclusion

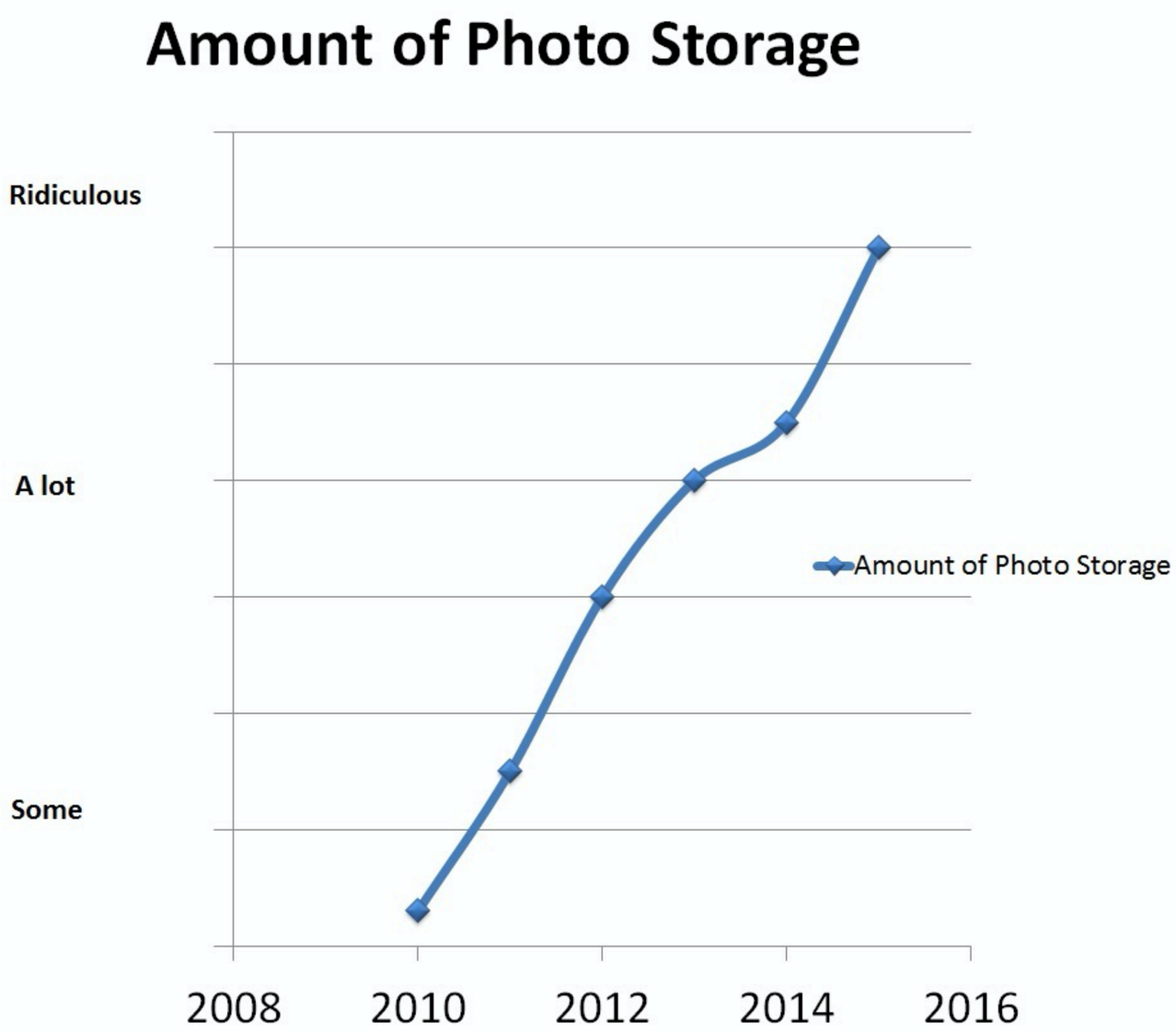
Archival at Facebook: context

Kestutis Patiejunas

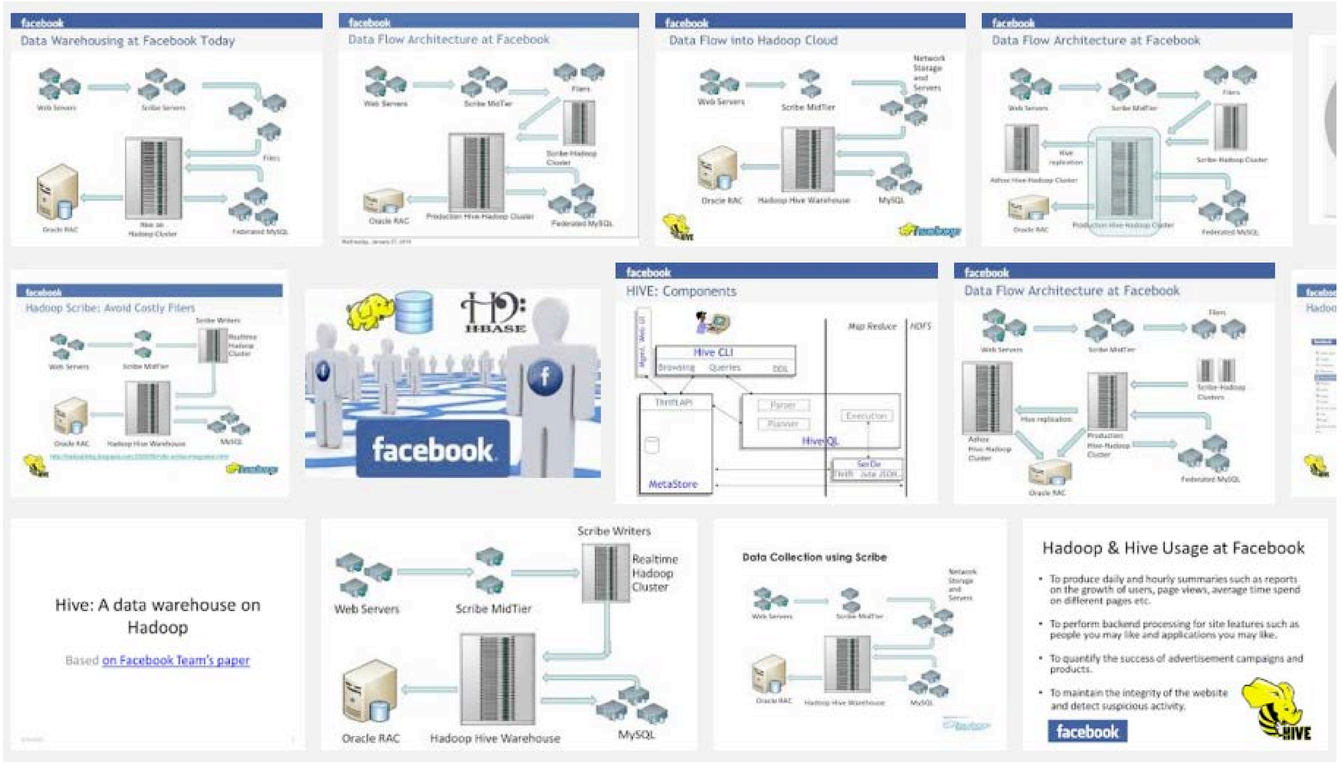
Facebook’s Monthly Active Users

Grew by 1.1B since 2010 monthly active users

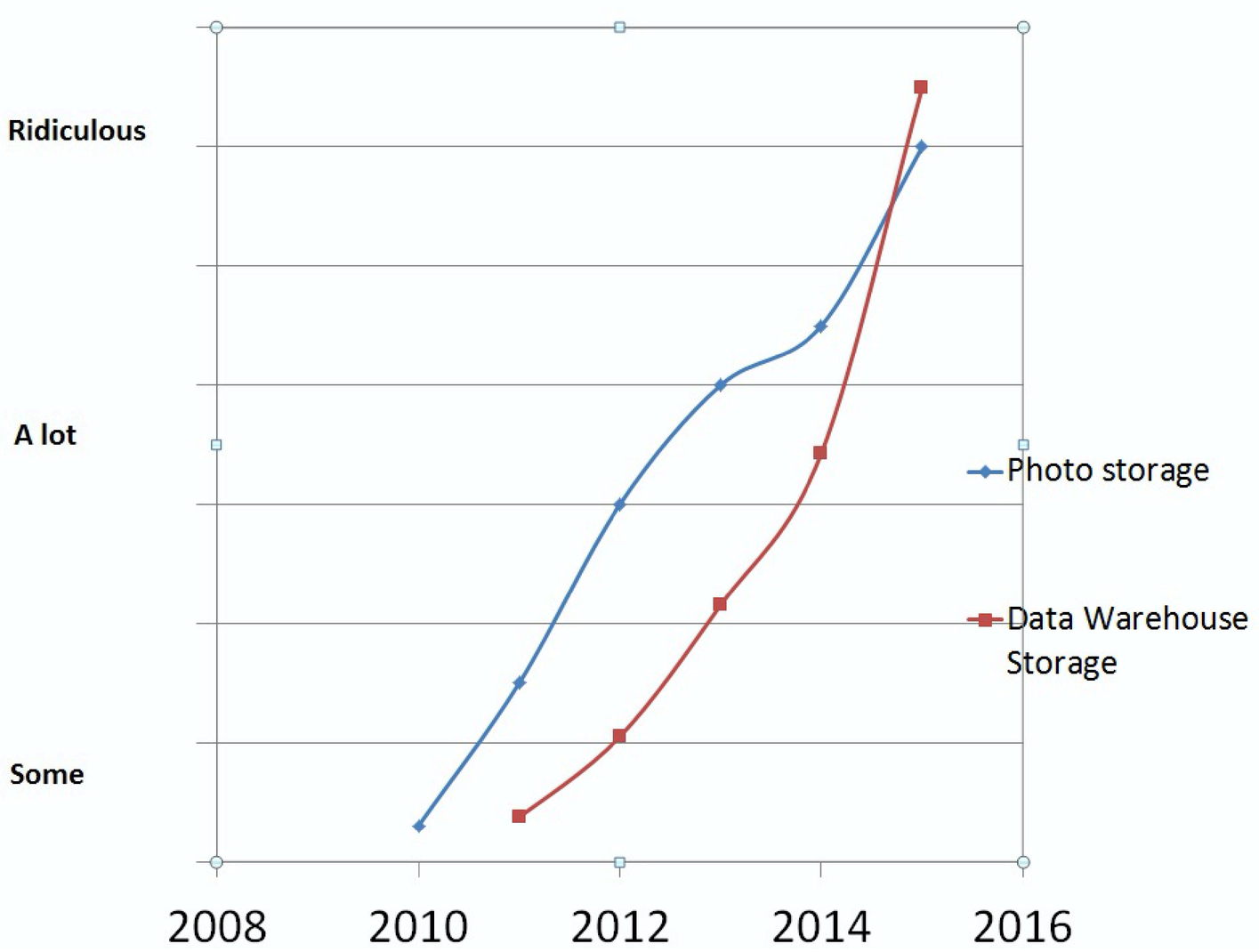

Era of the the Desktop
Advance of the Mobile
Growth in Messenger
Videos vs pictures
360 video
Number of users in millions
1,750
1,500
1,250
1,000
750
500
250
0
Q3 '08 Q1 '09 Q2 '09 Q3 '09 Q4 '09 Q1 '10 Q2 '10 Q3 '10 Q4 '10 Q1 '11 Q2 '11 Q3 '11 Q4 '11 Q1 '12 Q2 '12 Q3 '12 Q4 '12 Q1 '13 Q2 '13 Q3 '13 Q4 '13 Q1 '14 Q2 '14 Q3 '14 Q4 '14 Q1 '15 Q2 '15 Q3 '15
© Statista 2016


Amount of Photo Storage
Ridiculous
A lot
Some
Amount of Photo Storage
2008
2010
2012
2014
2016


Era of the the Desktop
Advance of the Mobile
Growth in Messenger
Videos vs pictures
360 video
Number of users in millions
1,750
1,500
1,250
1,000
750
500
250
0
Q3 '08
Q1 '09
Q2 '09
Q3 '09
Q4 '09
Q1 '10
Q2 '10
Q3 '10
Q4 '10
Q1 '11
Q2 '11
Q3 '11
Q4 '11
Q1 '12
Q2 '12
Q3 '12
Q4 '12
Q1 '13
Q2 '13
Q3 '13
Q4 '13
Q1 '14
Q2 '14
Q3 '14
Q4 '14
Q1 '15
Q2 '15
Q3 '15
© Statista 2016


facebook
Data Warehousing at Facebook Today
facebook
Data Flow Architecture at Facebook
facebook
Data Flow into Hadoop Cloud
facebook
Data Flow Architecture at Facebook
facebook
Hadoop Scribe: Avoid Costly Filers
facebook
HIVE: Components
facebook
Data Flow Architecture at Facebook
Hive: A data warehouse on Hadoop
Based on Facebook Team's paper
Scribe Writers
Realtime Hadoop Cluster
Web Servers
Scribe MidTier
Oracle RAC
Hadoop Hive Warehouse
MySQL
Data Collection using Scribe
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
HIVE


Era of the the Desktop
Advance of the Mobile
Growth in Messenger
Videos vs pictures
360 video
Number of users in millions
1,750
1,500
1,250
1,000
750
500
250
0
Q3 '08 Q1 '09 Q2 '09 Q3 '09 Q4 '09 Q1 '10 Q2 '10 Q3 '10 Q4 '10 Q1 '11 Q2 '11 Q3 '11 Q4 '11 Q1 '12 Q2 '12 Q3 '12 Q4 '12 Q1 '13 Q2 '13 Q3 '13 Q4 '13 Q1 '14 Q2 '14 Q3 '14 Q4 '14 Q1 '15 Q2 '15 Q3 '15
© Statista 2016


Ridiculous
A lot
Some
2008
2010
2012
2014
2016
Photo storage
Data Warehouse Storage

So what is an Exabyte?

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

X 30 times!

- **250k drives stacke**
  >30 times taller th
  Seattle Space Nee

facebook

Freezing Exabytes of Data at Facebook's Cold Storage

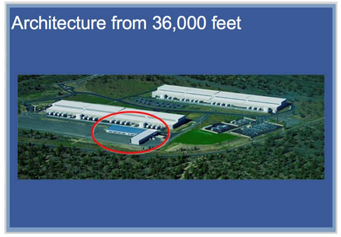
Kestutis Patiejunas (kestutip@fb.com)

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



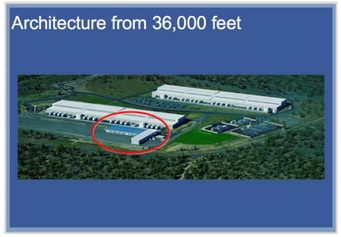
Questions and possibilities for mass storage industry

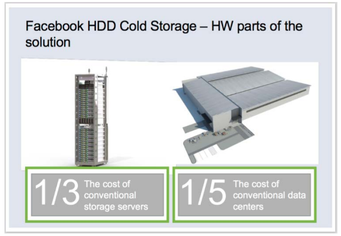
Hard drives:

- 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





The optical journey

Sam Merat

Why optical?

Data matters to Facebook

- Immutability
- Durability
  - Decades, not years
  - Differentiated environmental tolerance
- Efficiency
  - Must be competitive with all other archival media

2014: Assumptions at Facebook

How hard can it be? :)

- Software
  - In-house expertise
  - Looked similar to HDD storage
  - Assumption: low risk
- Robotics
  - Developed by partners
  - Looked similar to tape
  - Assumption: low-medium risk
- Drive & media
  - New application for optical
  - Durability & throughput analysis
  - Assumption: higher risk
- Integration
  - New end-to-end stack
  - Has to fit into datacenters
  - Assumption: higher risk

2014: Development starts

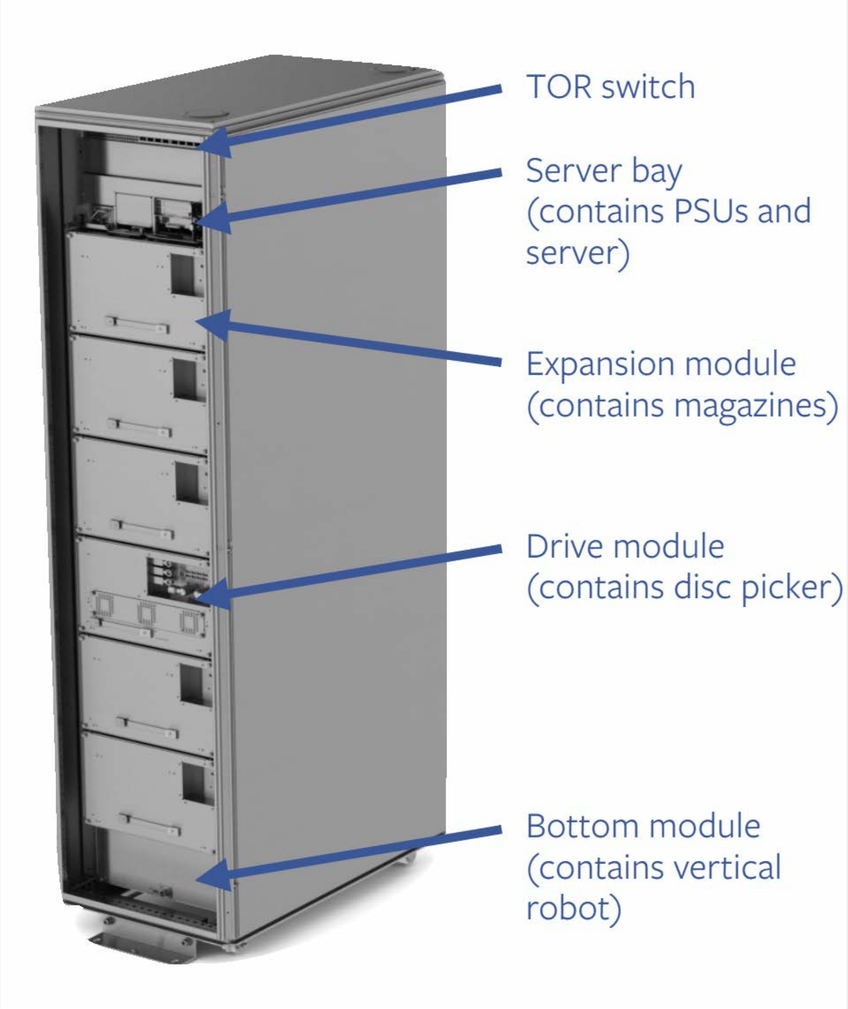
Optical in the lab

- Facebook starts on software & datacenter integration
  - Rack dimensions, power consumption, throughput
  - Validation
- Early partners
  - Design & manufacturing of datacenter rack
  - Media, drive evaluation & development

2014: First set of lessons learned

- Robotics design
  - More challenging than anticipated
  - Movement; calibration; misalignment; scratching media; datacenter cabling; operator access to unit; …
- Media durability
  - Verification is time-consuming, subtle
  - Fault identification; write patterns; scrubbing; monitoring; environment conditions; …

2015 & 2016: Gen1 Spec



- 1 server used for robotics control and IO.
- Accessed through Ethernet.
- 12 burners per rack ~ 216MB/s
- 5k discs ~ 0.5 PB

2015 & 2016: Gen 1 deployment

Optical enters the Facebook datacenters

- **Deployed 10s of petabytes of Panasonic optical system**
  - 100 GB discs
- Studied disc reliability
  - 0.015% disc failure rate
- Deployment deepened collaboration with Panasonic
  - Identified, exposed end-to-end diagnostics
  - Iterated on validation process

2017: Gen 2 deployment

Optical expands in the Facebook datacenters

- Deploying 300 GB disks (3x density increase) ~ 1.5PB
- Increase burner speed by 3x ~ 720MB/s
- Scaling deployment to 100s of petabytes (10x increase)
- Adding more Facebook applications to optical
- Datacenter integration improvements

2018: Gen 3 deployment

Next generation optical at Facebook

- 500 GB discs (5x improvement over first generation)
- EB scale
- Applying learning from Gen 1 & Gen 2 deployments
  - Close collaboration with Panasonic
    - Improvements to entire stack

Conclusions

- Facebook is committed to have massive archival storage based on alternative non magnetic technology
- In cooperation with Panasonic Facebook has many tens of PB optical storage operational in datacenter environment
- 2016/2017 Gen2 deployment of optical storage will be order of magnitude bigger
- Finally: optical storage is reaching maturity and is a viable and competitive alternative for a massive data storage

facebook