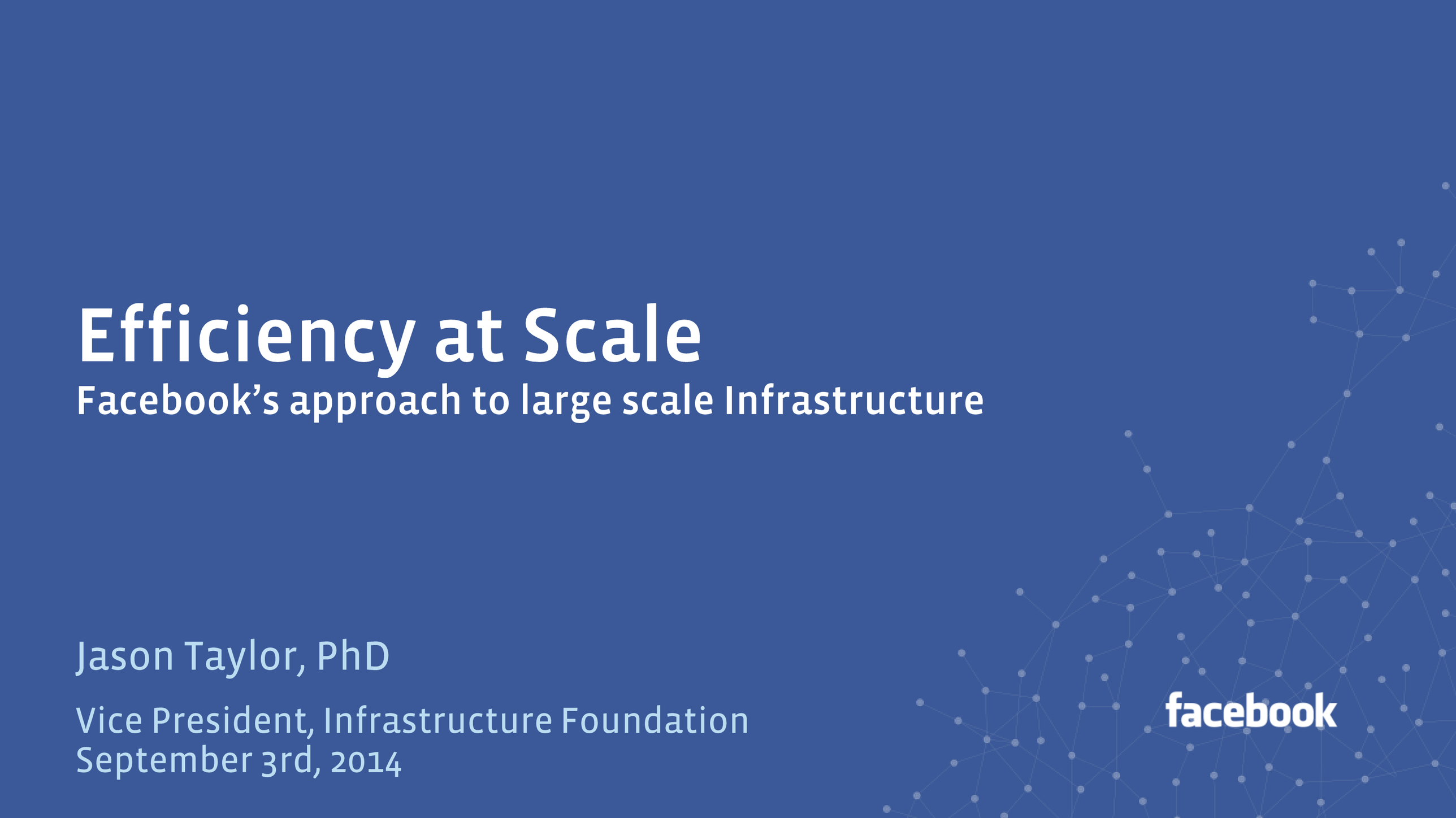

facebook

Efficiency at Scale

Facebook's approach to large scale Infrastructure

Jason Taylor, PhD

Vice President, Infrastructure Foundation
September 3rd, 2014



Agenda

1	Facebook Scale & Infrastructure
2	Efficiency at FB
3	Disaggregated Rack
4	Q & A

Facebook Scale

Data Centers in 5 regions.

Facebook Stats

- 1.28 billion users (3/2014)
- 802 million people use Facebook daily

- 350+ million photos added per day (1/2013)
- 240+ billion photos

- 4.5 billion likes, posts and comments per day (5/2013)

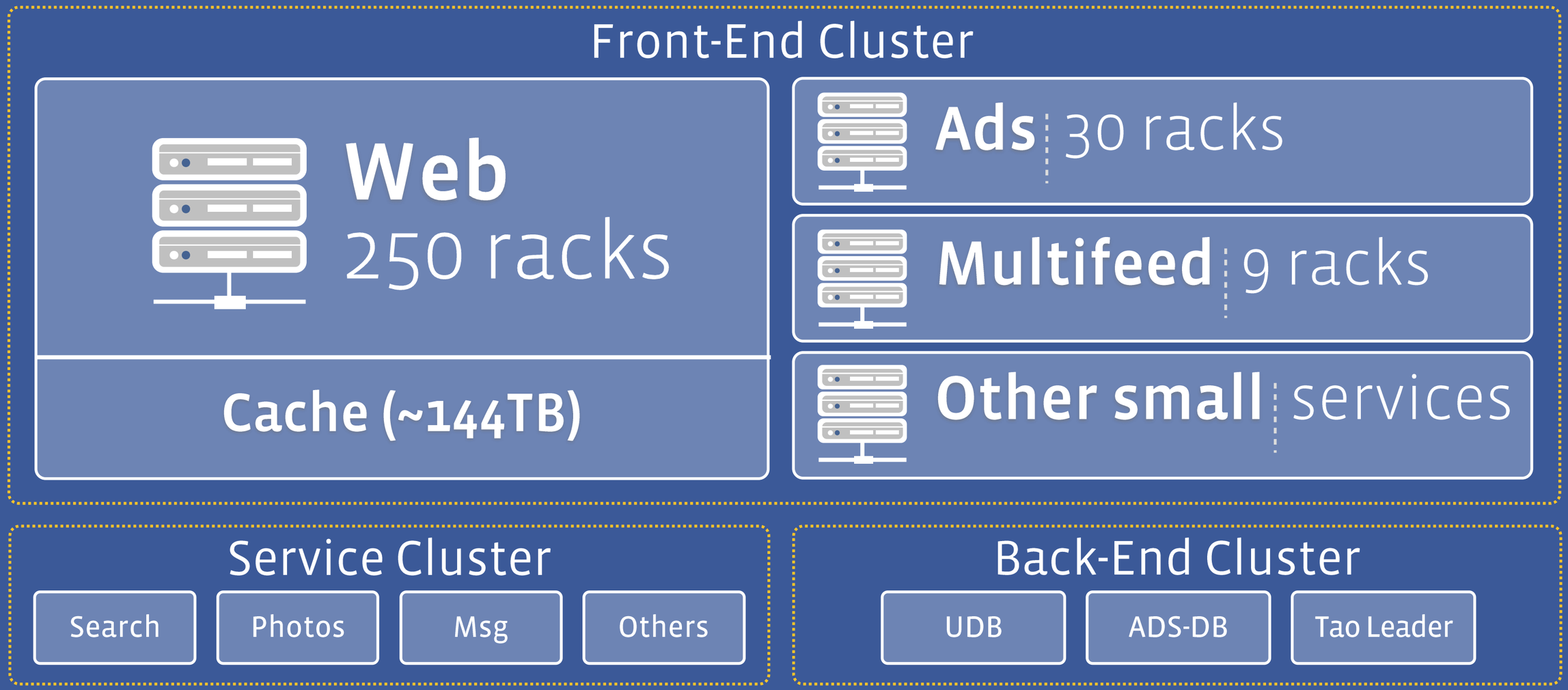
- 300+ PB in our data warehouse (11/2013)

Cost and Efficiency

- Infrastructure spend in 2012 (from our 10-K):
  - "…$1.24 billion for capital expenditures related to the purchase of servers, networking equipment, storage infrastructure, and the construction of data centers."

- Efficiency work has been a top priority for several years
  - $1.2 billion saved over the last three years.

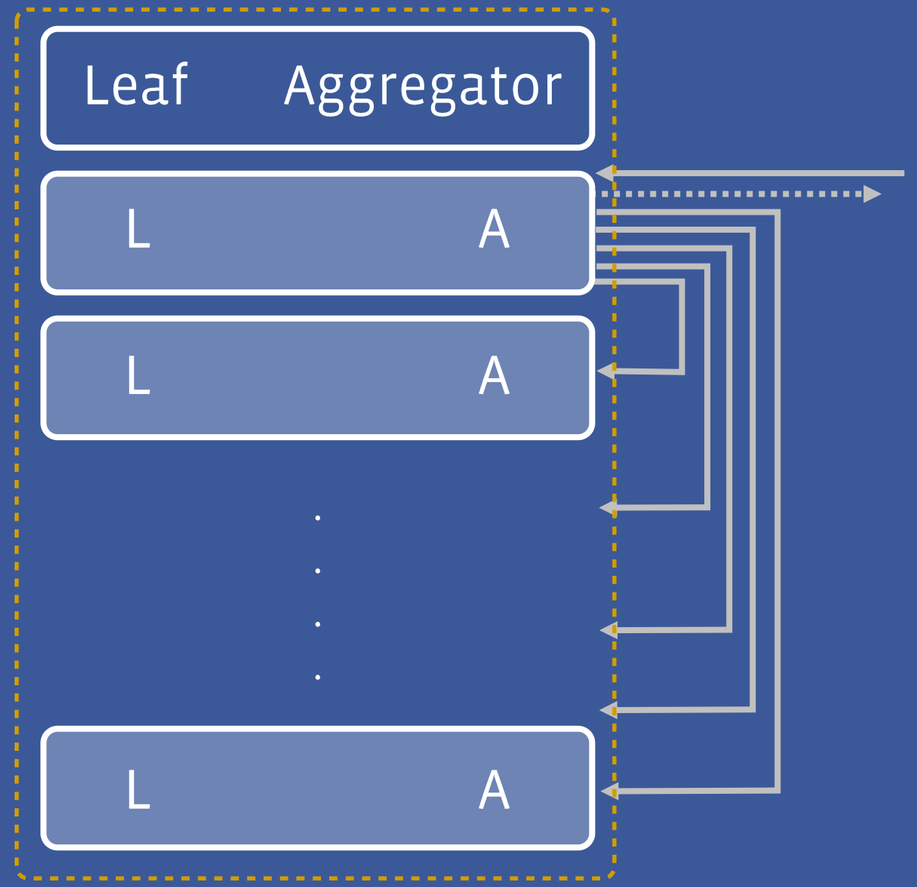
Architecture



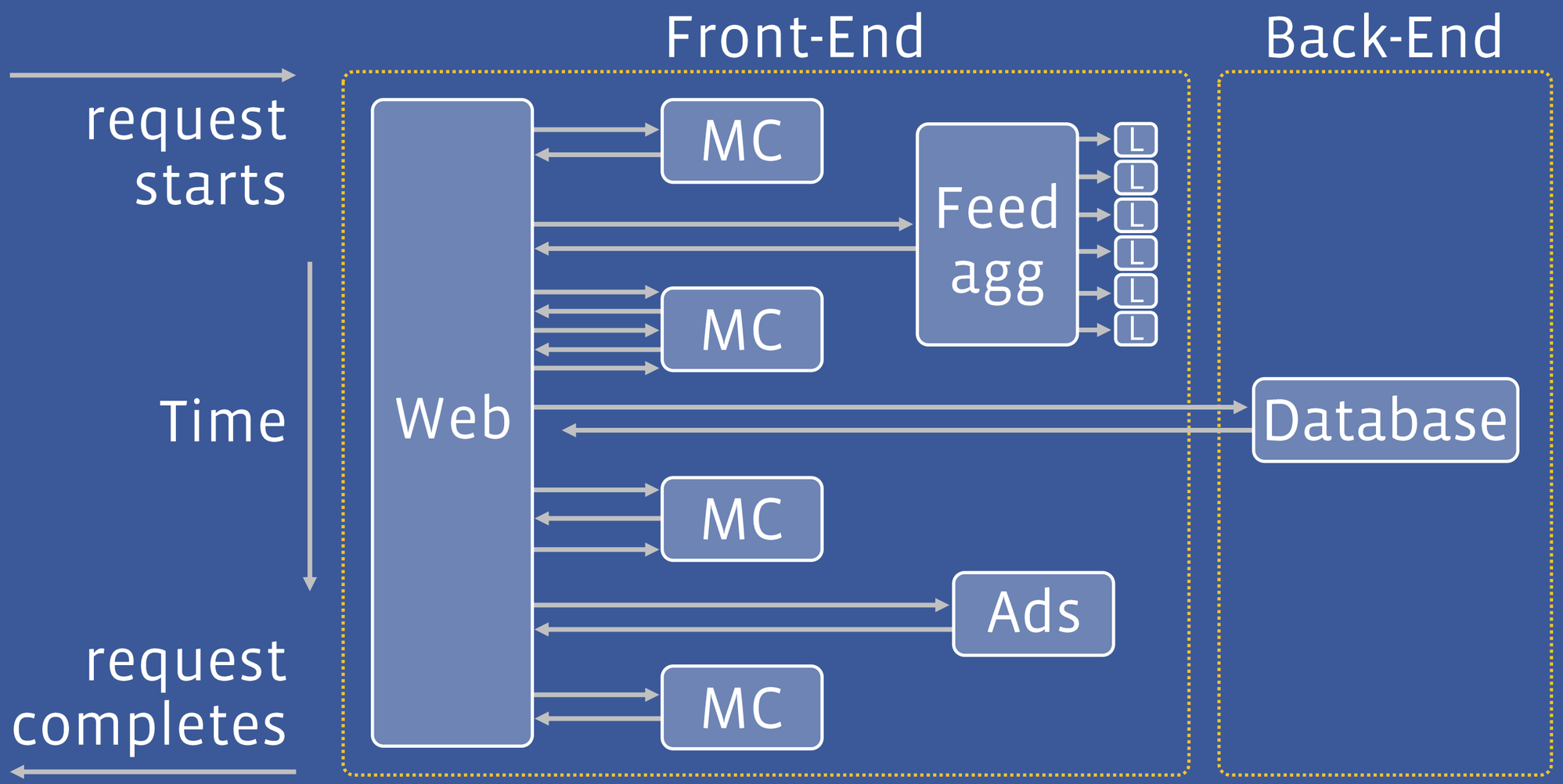
Lots of “vanity free” servers.

News Feed rack

- The rack is our unit of capacity
  - All 40 servers work together
- Leaf + agg code runs on all servers
  - Leaf has most of the RAM
  - Aggregator uses most of the CPU
- Lots of network BW within the rack



Life of a “hit”


Front-End
Back-End
request starts
Time
request completes
Web
MC
MC
MC
MC
Feed agg
L
L
L
L
L
L
Database
Ads

Five Standard Servers

Standard Systems	I Web	III Database	IV Hadoop	V Photos	VI Feed
CPU	High 2 x E5-2670	High 2 x E5-2660	High 2 x E5-2660	Low	High 2 x E5-2660
Memory	Low	High 144GB	Medium 64GB	Low	High 144GB
Disk	Low	High IOPS 3.2 TB Flash	High 15 x 4TB SAS	High 15 x 4TB SAS	Medium
Services	Web, Chat	Database	Hadoop (big data)	Photos, Video	Multifeed, Search, Ads

Five Server Types

- **Advantages**:
  - Volume pricing
  - Re-purposing
  - Easier operations - simpler repairs, drivers, DC headcount
  - New servers allocated in hours rather than months

- **Drawbacks**:
  - 40 major services; 200 minor ones - not all fit perfectly
  - The needs of the service change over time.

Agenda

1	Facebook Scale & Infrastructure
2	**Efficiency at FB**
3	Disaggregated Rack
4	Q & A

Efficiency at FB

Data Centers

- Heat management, electrical efficiency & operations

Servers

- “Vanity free” design & supply chain optimization

Software

- Horizontal wins like HPHP/HHVM, cache, db, web & service optimizations

Next Opportunities?

Disaggregated Rack

- Better component/service fit
- Extending component useful life

Developing New Components

- CPU, RAM, Disk & Flash

Agenda

1	Facebook Scale & Infrastructure
2	Efficiency at FB
3	**Disaggregated Rack**
4	Q & A

A rack of news feed servers...

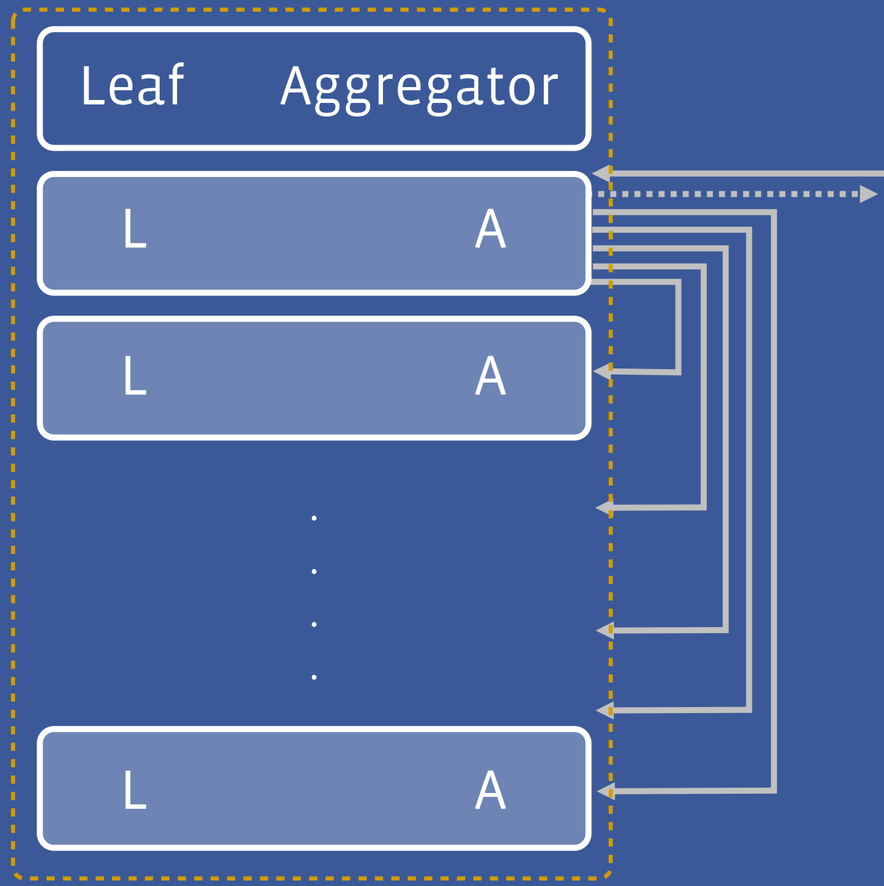
=>

80 processors
640 cores COMPUTE

5.8 TB RAM

80 TB STORAGE

30 TB FLASH



The application lives on a rack of equipment--not a single server.

Compute

- Standard Server
  - 2 processors (or many)
  - 8 or 16 DIMM slots
  - no hard drive - small flash boot partition.
  - big NIC - 10 Gbps or more

Ram Sled

- Hardware
  - 128GB to 512GB
  - compute: FPGA, ASIC, mobile processor or desktop processor
- Performance
  - 450k to 1 million key/value gets/sec
- Cost
  - Excluding RAM cost: $500 to $700 or a few dollars per GB

Storage Sled (Knox)

- Hardware
  - 15 drives
  - Replace SAS expander w/ small server
- Performance
  - 3k IOPS
- Cost
  - Excluding drives: $500 to $700 or less than $0.01 per GB

Flash Sled

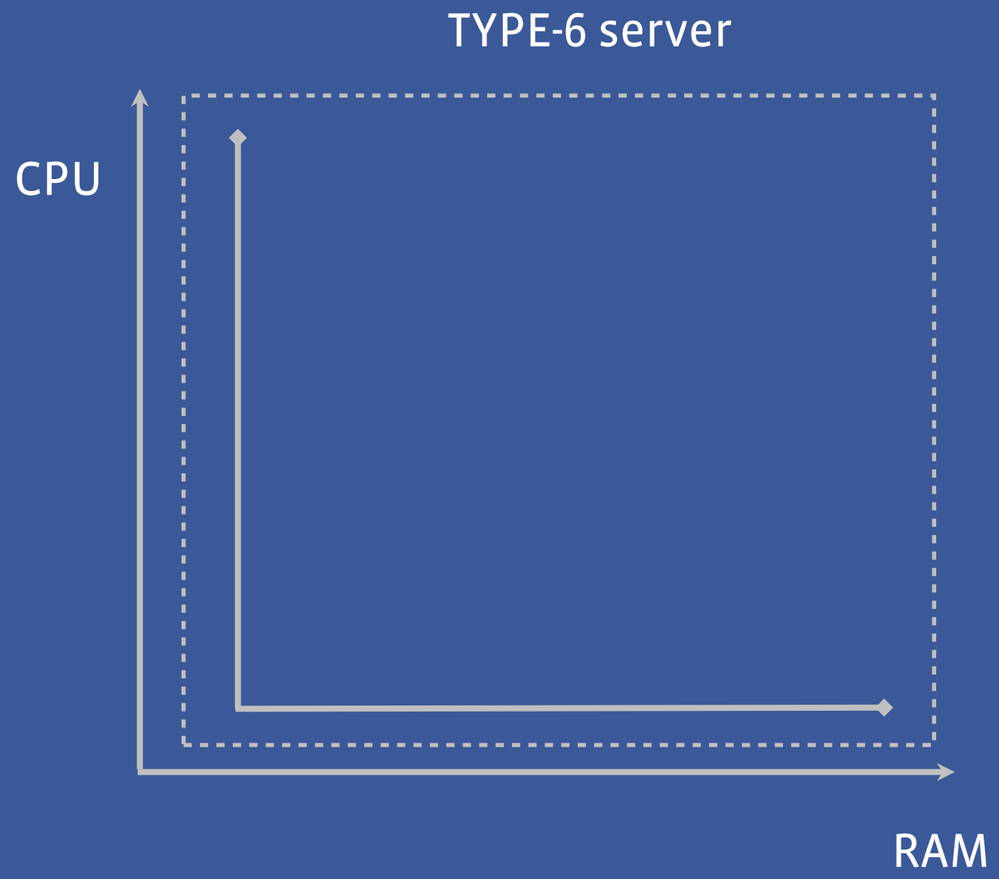
- Hardware
  - 175GB to 18TB of flash
- Performance
  - 600k IOPS
- Cost
  - Excluding flash cost: $500 to $700

NIC at 70% utilization	IOPS	Capacity
1 Gbps	21k	175 GB
10 Gb	210k	1.75 TB
25 Gb	525k	4.4 TB
40 Gb	840k	7.7 TB
50 Gb	1.05M	8.8 TB
100 Gb	2.1M	17.5 TB

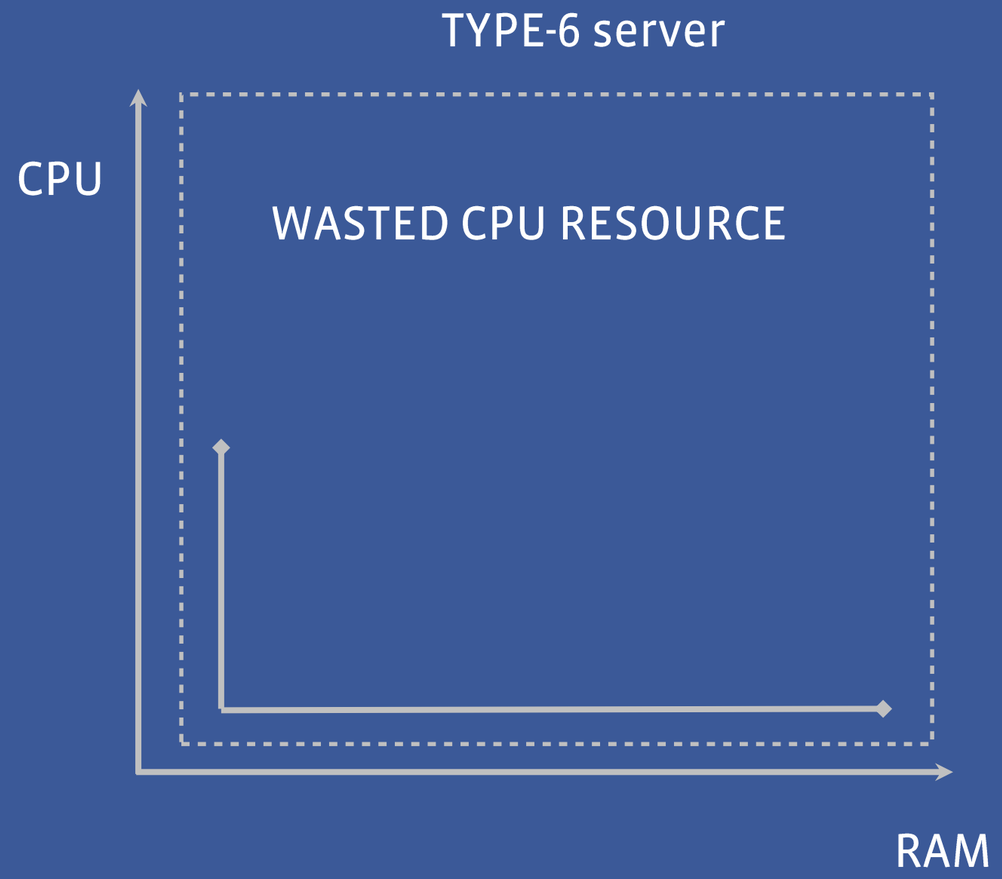
Three Disaggregated Rack Wins

- Server/Service Fit - across services
- Server/Service Fit - over time
- Longer useful life through smarter hardware refreshes.

Server/Service Fit - across services

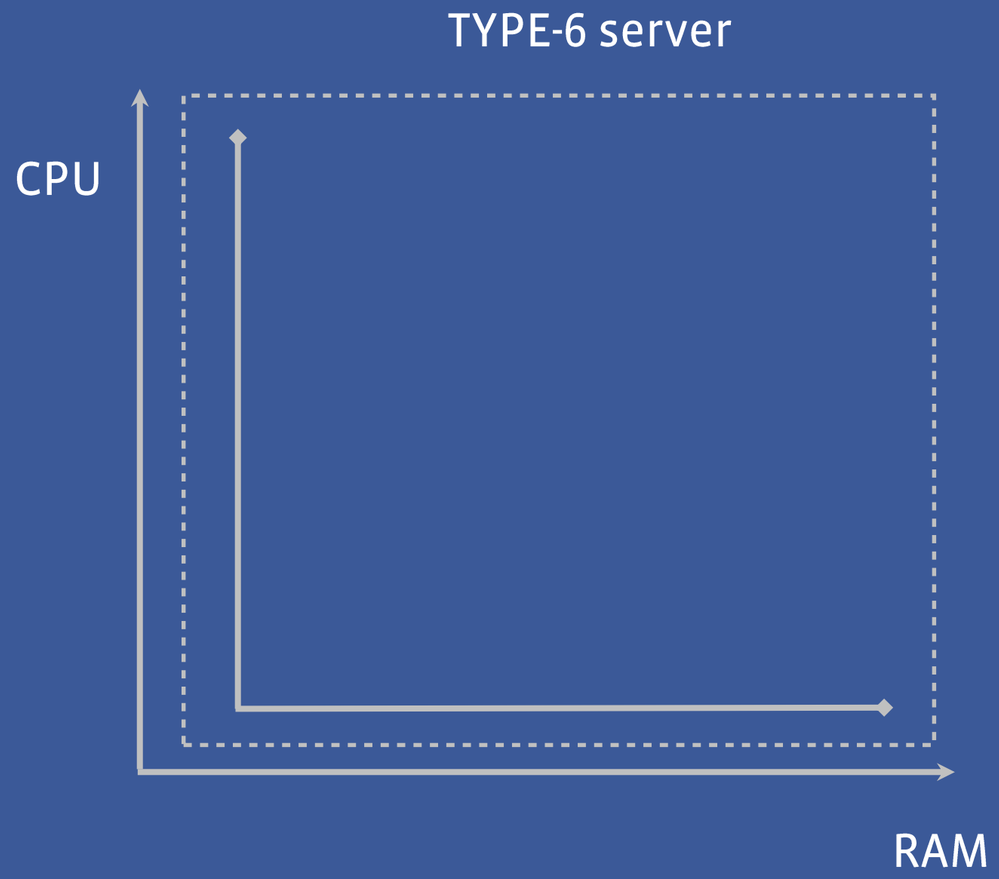


MultiFeed
(news feed)

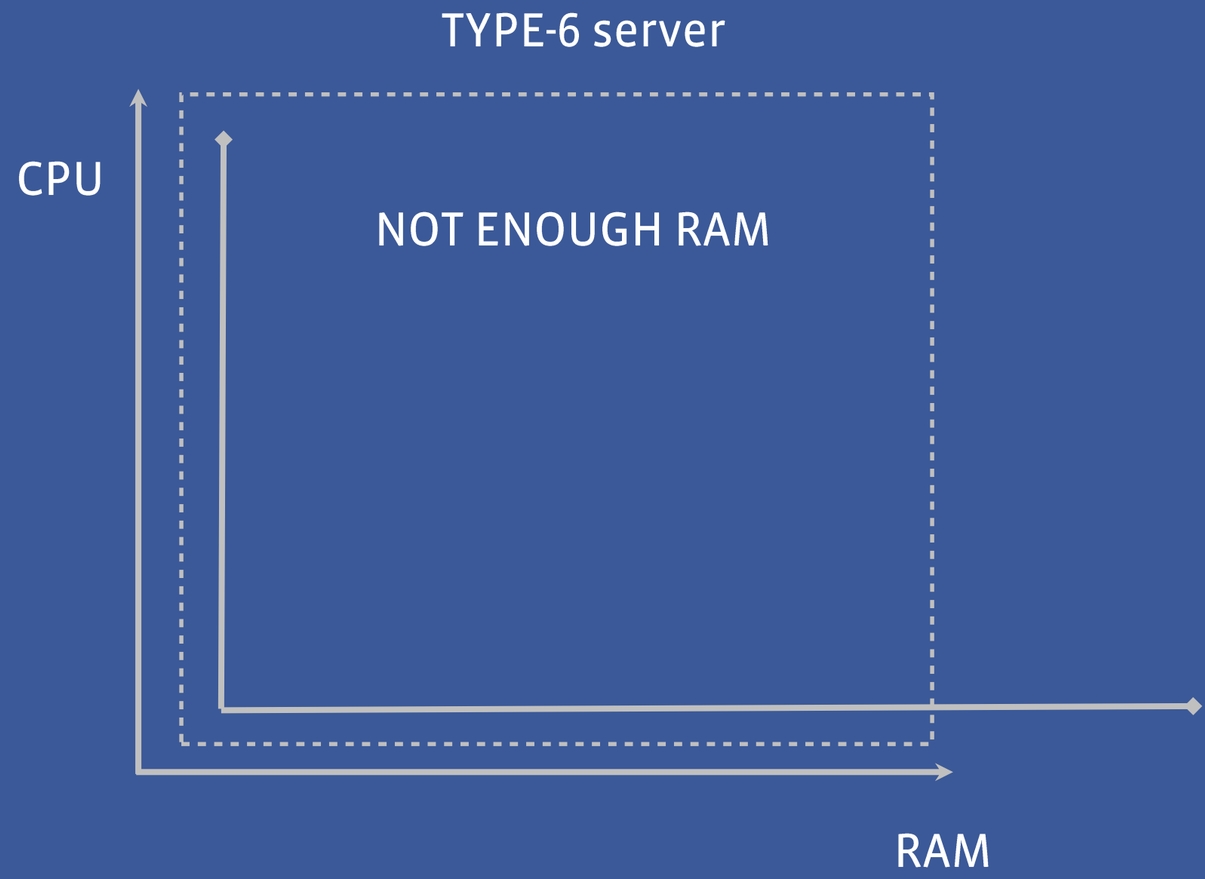


Other Service A

Server/Service Fit - over time



Year 1



Year 2 - more RAM needed

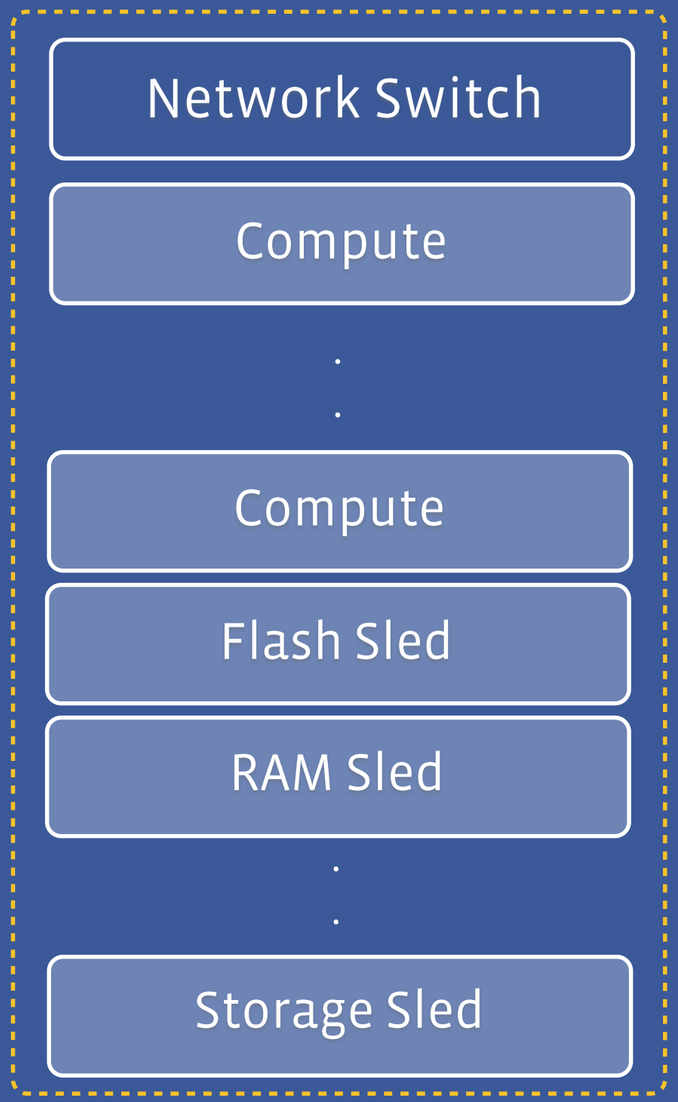
Longer Useful Life

Today servers are typically kept in production for about 3 years.

With disaggregated rack:

- Compute - 3 to 6 years
- RAM sled - 5 years or more
- Disk sled - 4 to 5 years depending on usage
- Flash sled - 6 years depending on write volume

A Disaggregated Rack for Graph Search...



=>

40 processors
320 cores COMPUTE

3.1 TB RAM

60 TB STORAGE

30 TB FLASH

20 Compute Servers

8 Flash Sleds

2 RAM Sleds

1 Storage Sled

=> 1:10 RAM:Flash ratio

* Add 4 more flash sleds in 2014 to get to a 1:15 RAM:Flash ratio *

Disaggregated Rack

- **Strengths**:
  - Volume pricing, serviceability, etc.
  - Custom Configurations
  - Hardware evolves with service
  - Smarter Technology Refreshes
  - Speed of Innovation

- **Potential issues**:
  - Physical changes required
  - Interface overhead

Approximate Win Estimates

Conservative assumptions show a 12% to 20% opex savings.

More aggressive assumptions promise between 14% and 30% opex savings.

* These are reasonable savings estimates of what may be possible across several use cases.


facebook