Warehouse-Scale Computing and the BDAS Stack

Ion Stoica

UC Berkeley

Overview

Workloads

Hardware trends and implications in modern datacenters

BDAS stack

What is Big Data used For?

Reports, e.g.,
- » Track business processes, transactions

Diagnosis, e.g.,
- » Why is user engagement dropping?
- » Why is the system slow?
- » Detect spam, worms, viruses, DDoS attacks

Decisions, e.g.,
- » Decide what feature to add
- » Decide what ad to show
- » Block worms, viruses, …

Data is as useful as the decisions it enables

Data Processing Goals

Low latency queries on historical data: enable faster decisions

- » E.g., identify why a site is slow and fix it

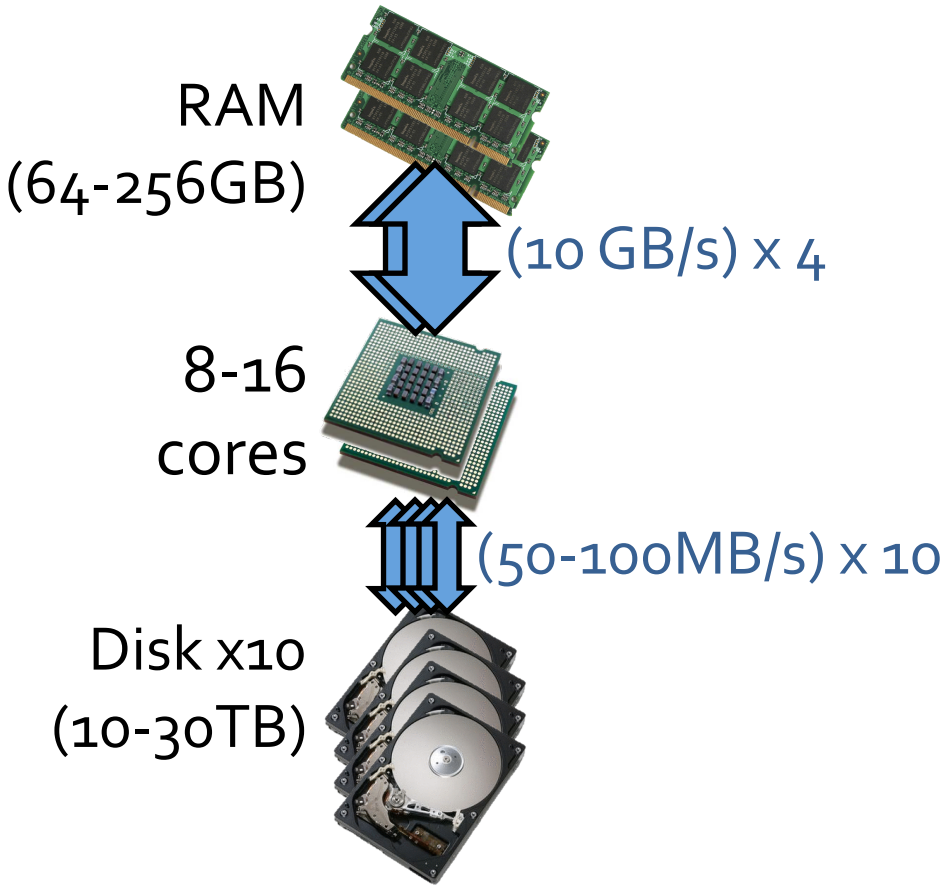
Low latency queries on live data (streaming): enable decisions on real-time data

- » E.g., detect & block worms in real-time (a worm may infect **1mil** hosts in **1.3sec**)

Sophisticated data processing: enable "better" decisions

- » E.g., anomaly detection, trend analysis

Typical Datacenter Node



CPU:

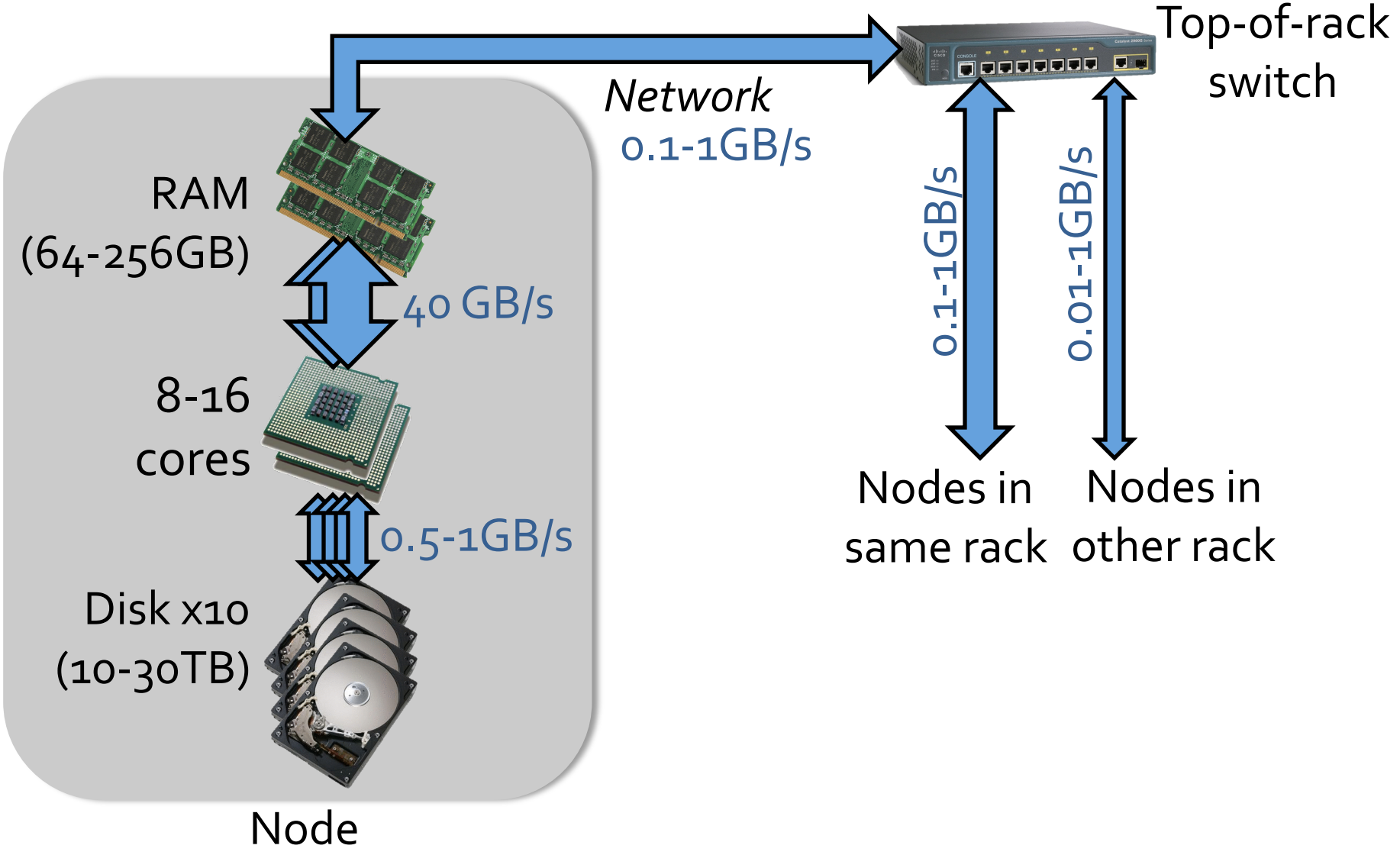
- » 8-16 cores
- » Transfer rate: quad channel DDR3: 4x10GB/s

RAM: 64-256 GB

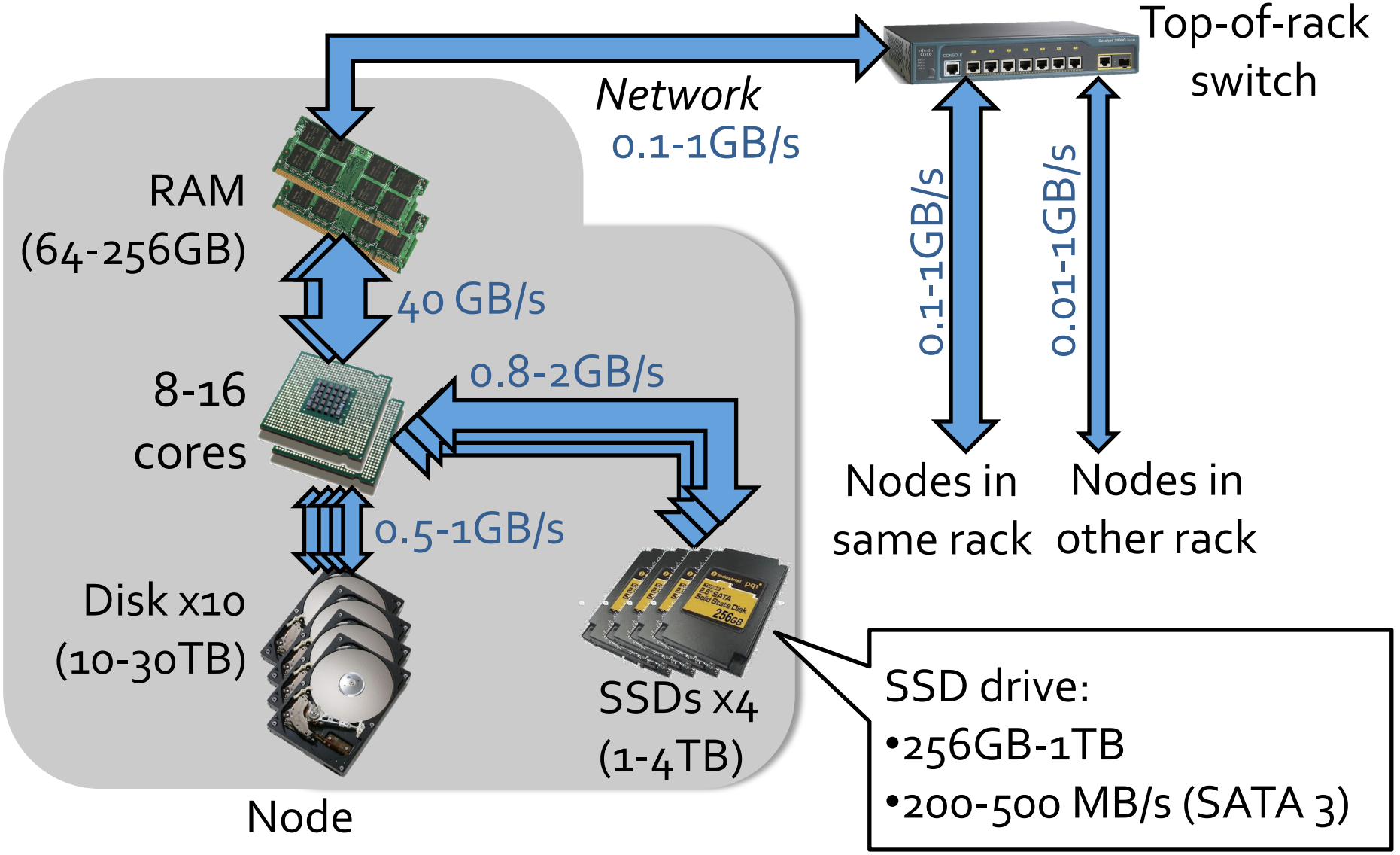
Storage: 8-12 disks, each

- » Capacity: 1-3TB
- » Transfer rate: 50-100MB/s

Typical Datacenter Node



Typical Datacenter Node



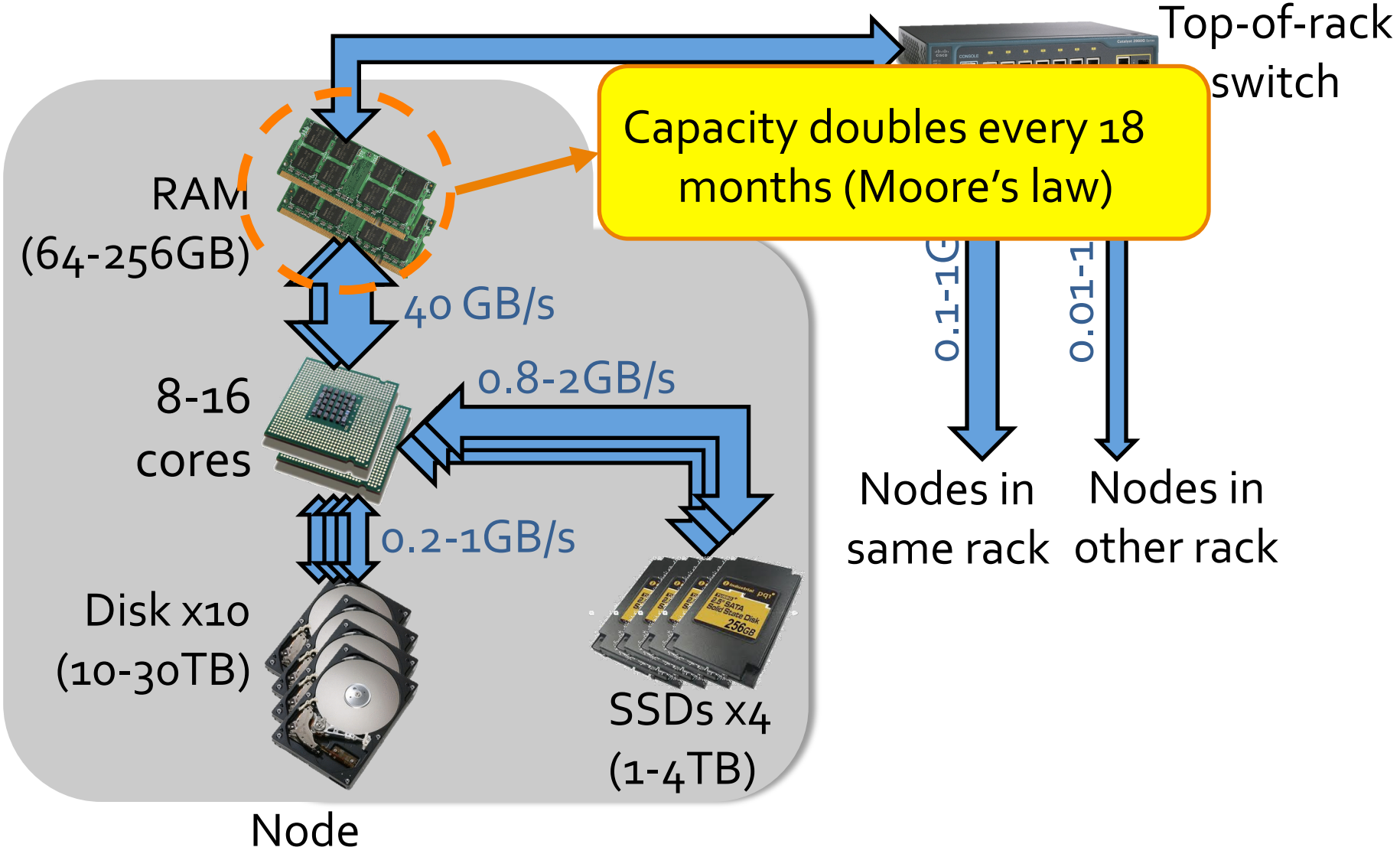
How Much Does $1,000 Buy You Today?

15-20TB disk storage (consumer grade disks)

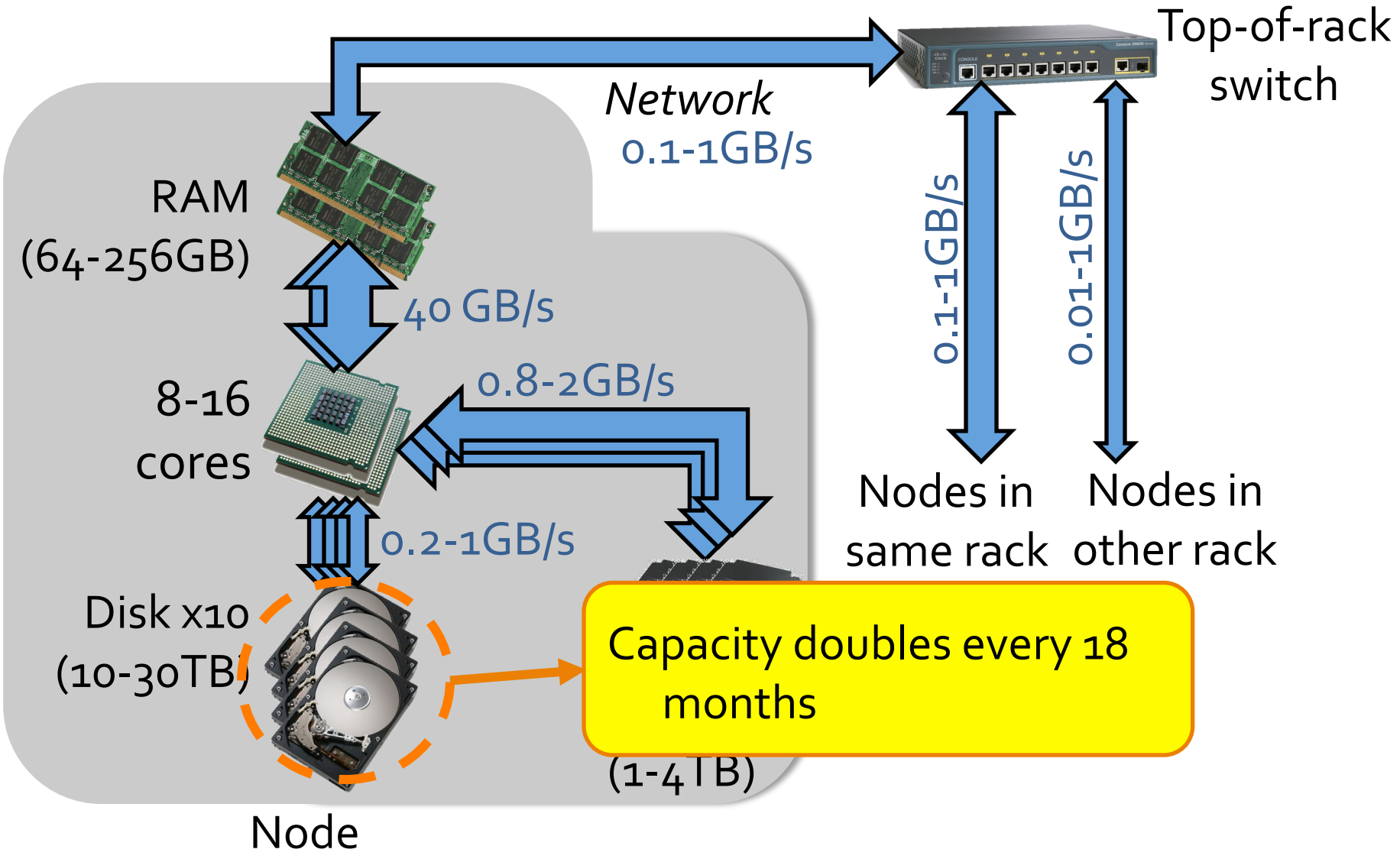
1 TB of SSD storage (SATA-3)

0.2TB of RAM

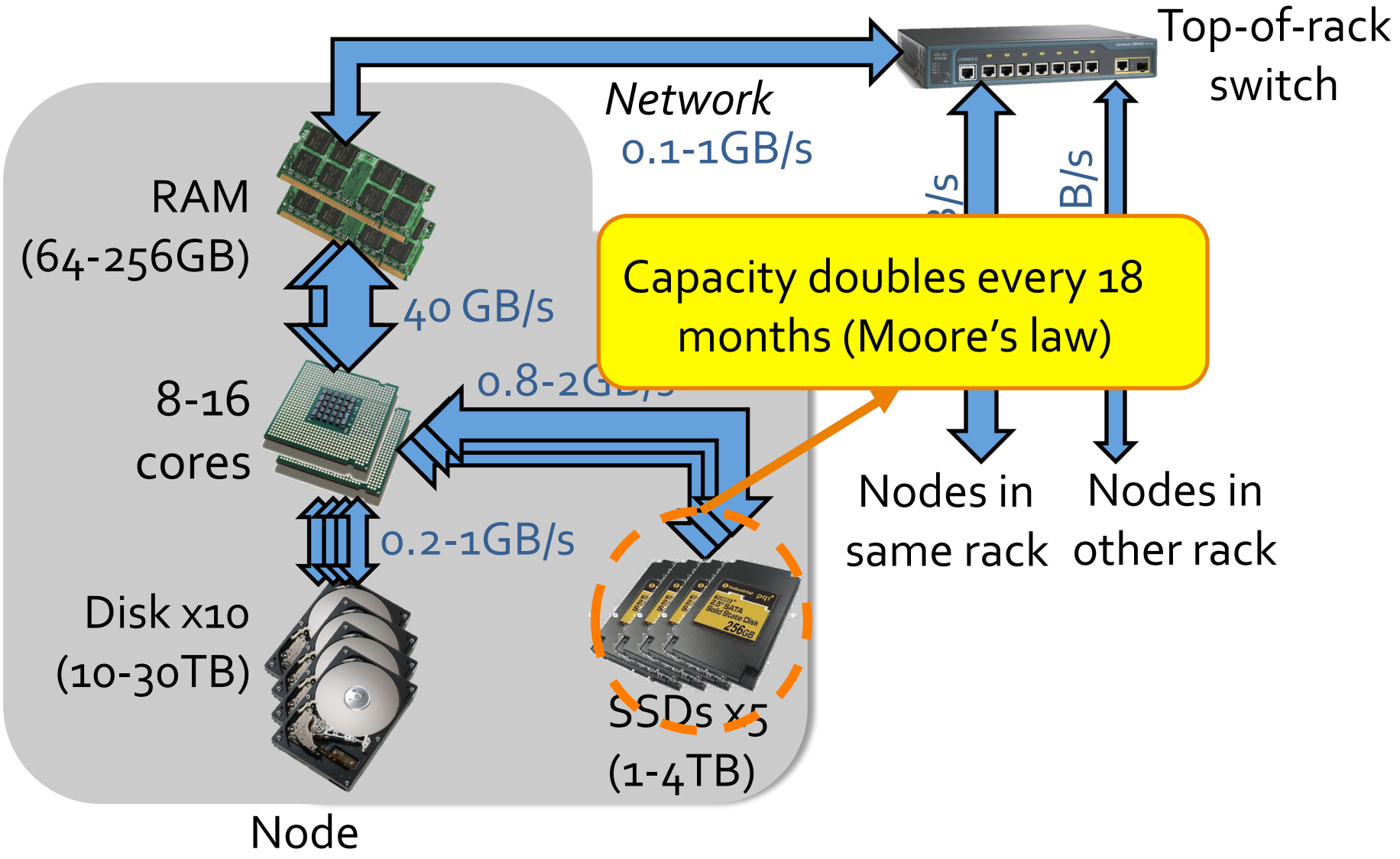
Memory Capacity Trends



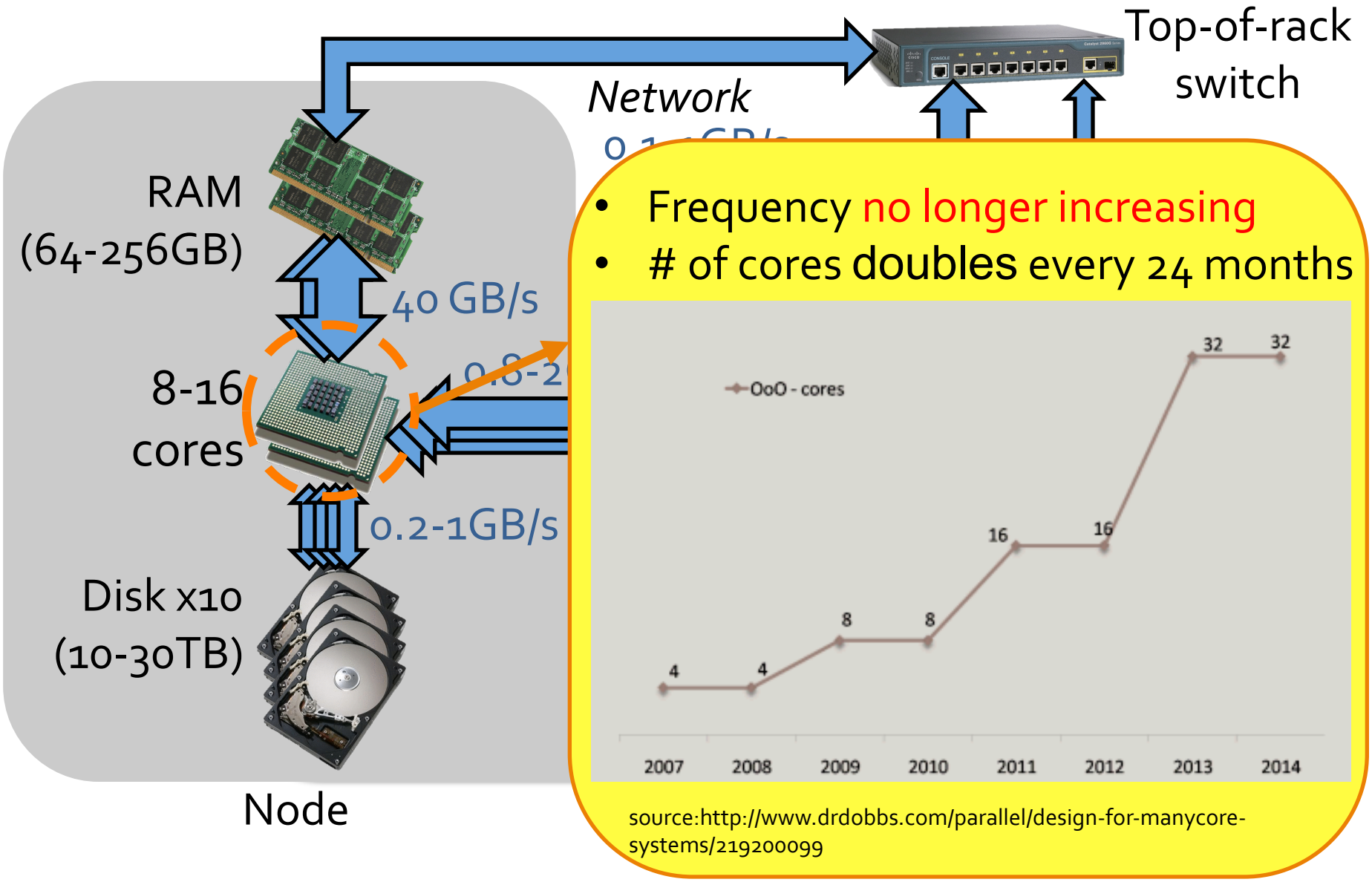
Disk Capacity Trends



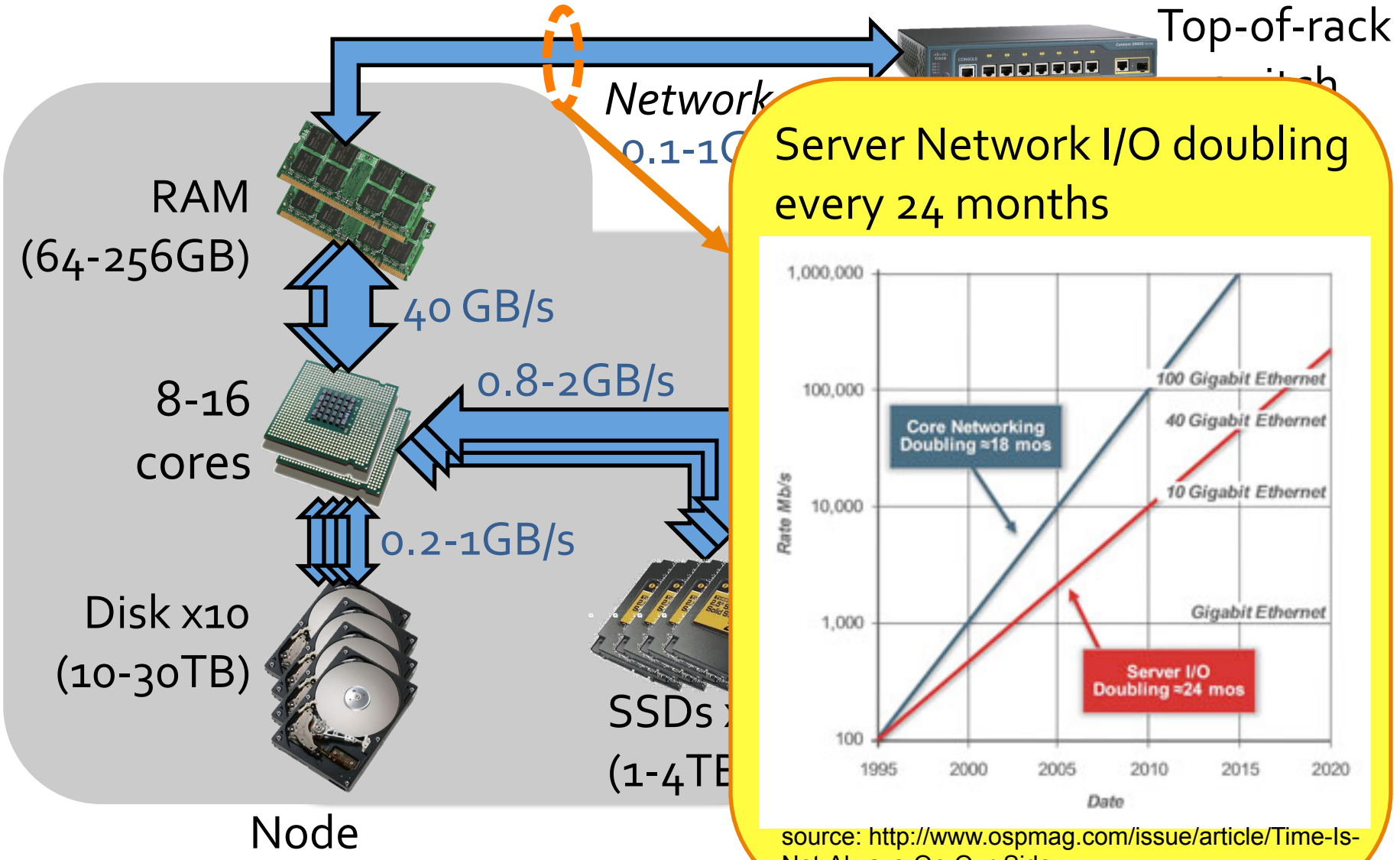
SSD Capacity Trends



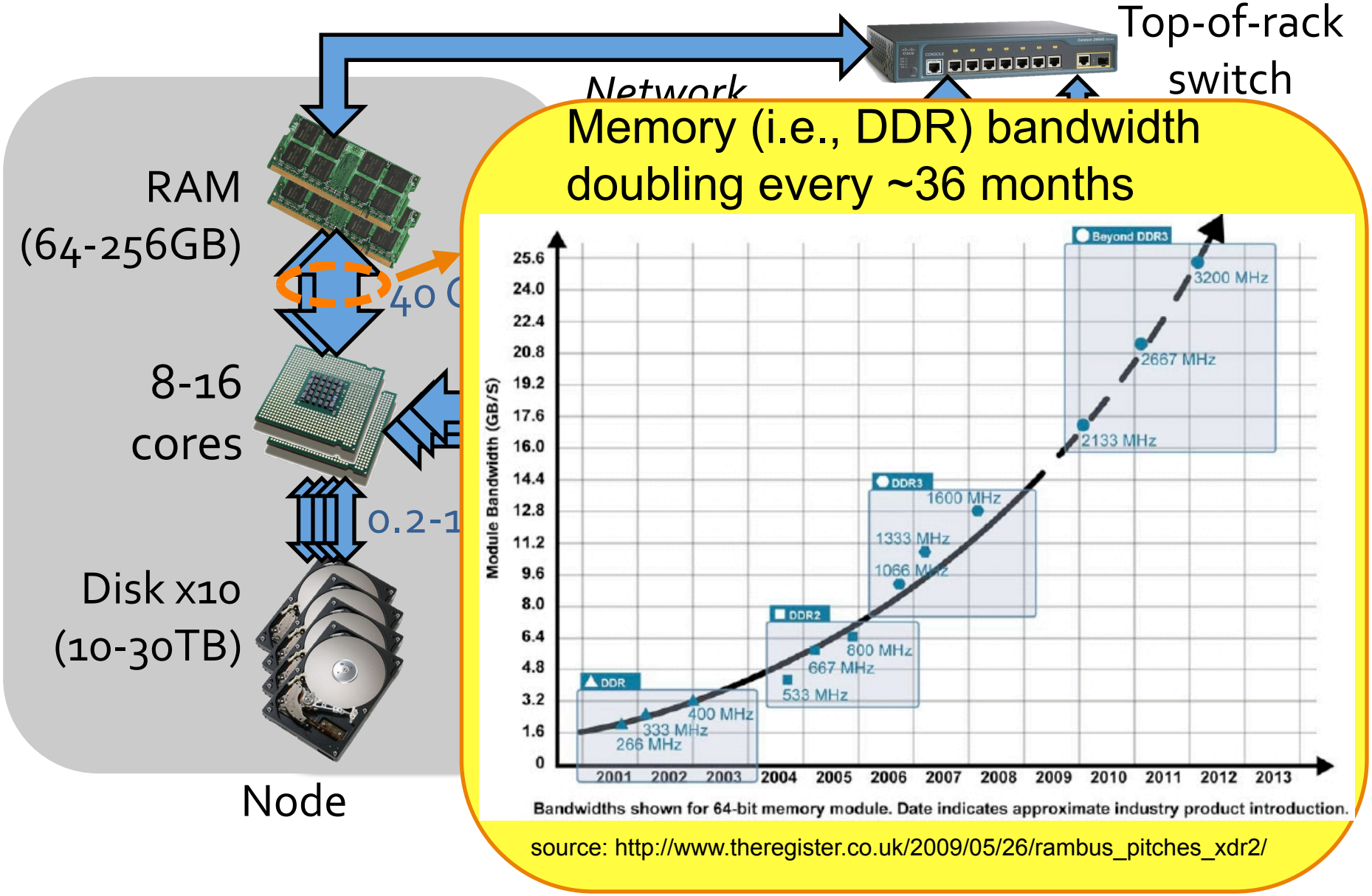
CPU Trends


Top-of-rack switch
Network
RAM (64-256GB)
40 GB/s
8-16 cores
0.2-1GB/s
Disk x10 (10-30TB)
Node
• Frequency no longer increasing
• # of cores doubles every 24 months
OoO - cores
4
4
8
8
16
16
32
32
2007
2008
2009
2010
2011
2012
2013
2014
source:http://www.drdobbs.com/parallel/design-for-manycore-systems/219200099

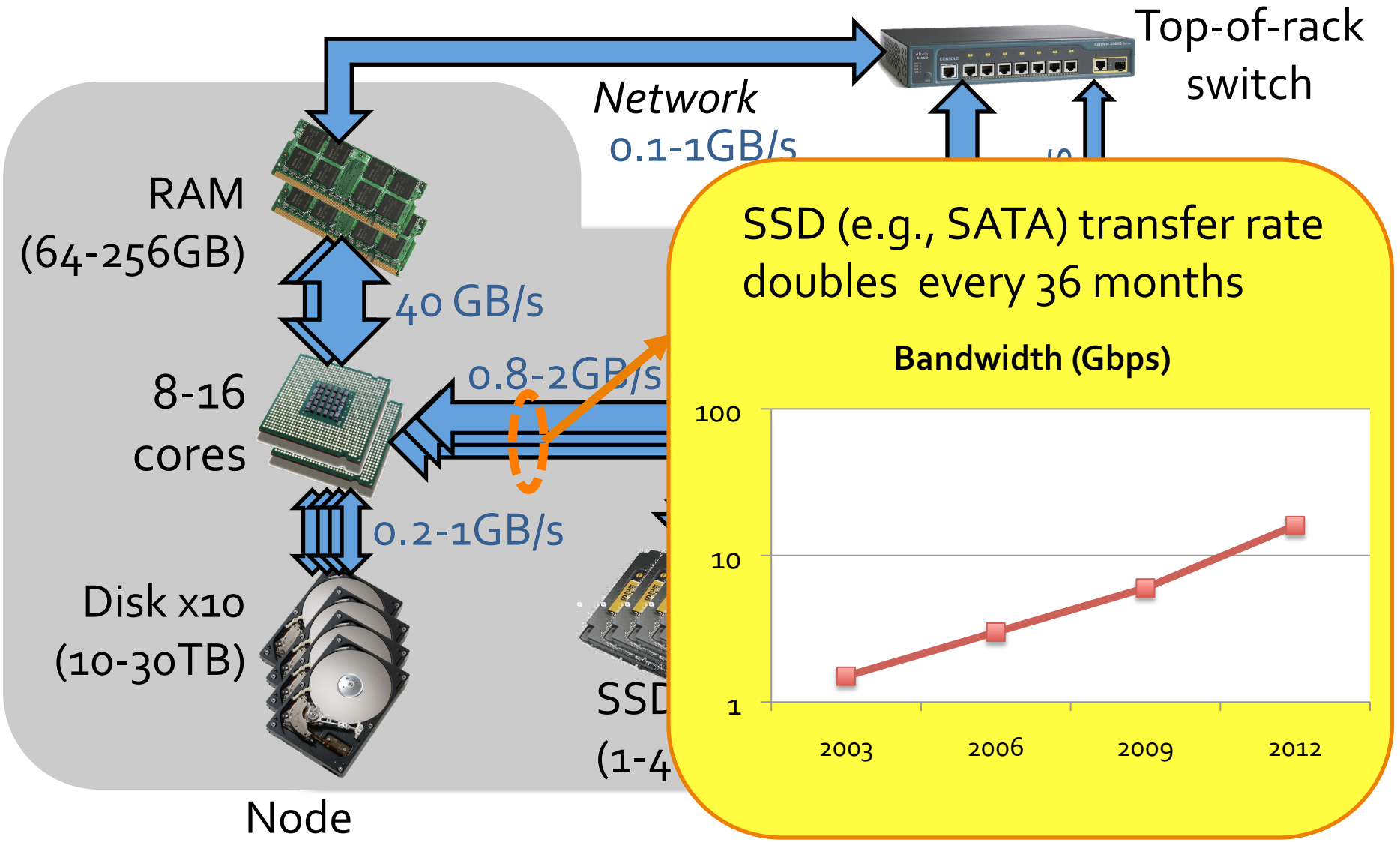
Network I/O Trends



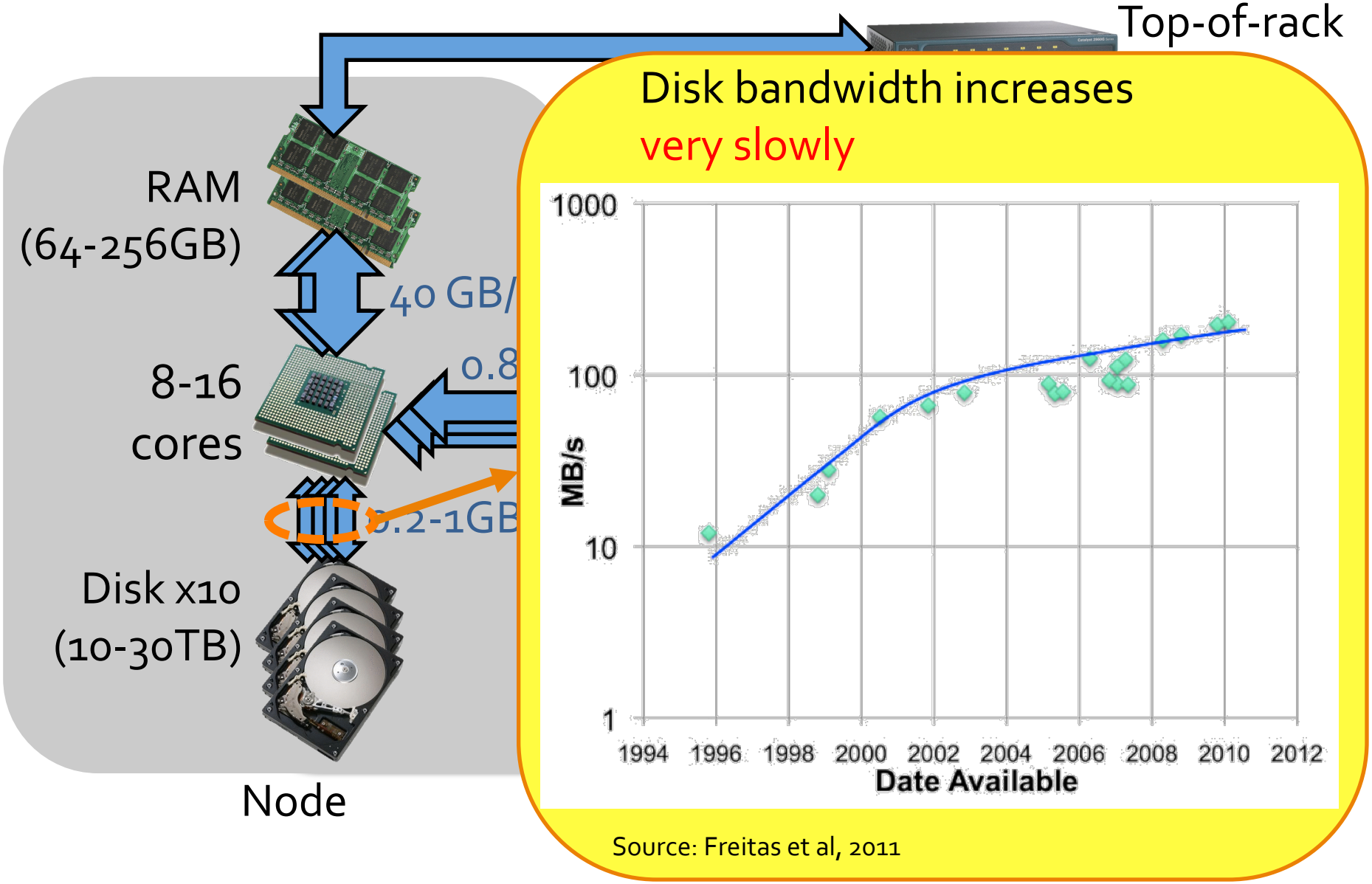
Network I/O Trends



SSD Throughput Trends



Disk Throughput Trends



Source: Freitas et al, 2011

Trend Summary

Rate doubles every 24 months

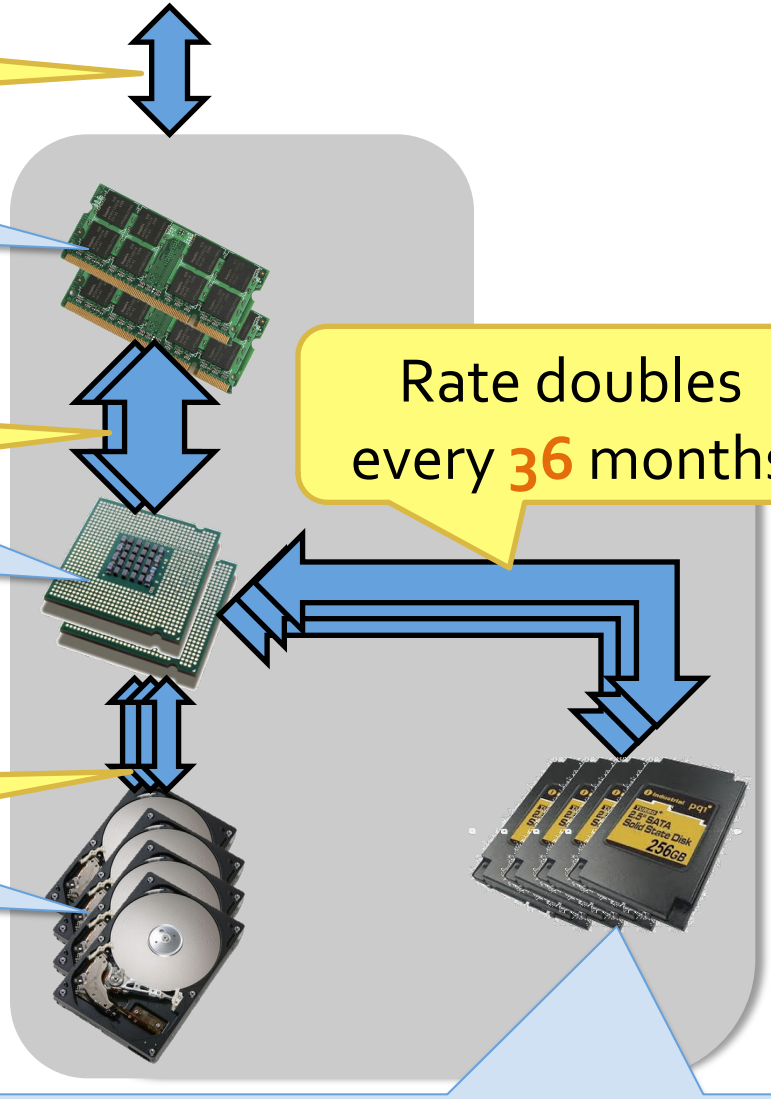
Capacity doubles every 18 months

Rate doubles every 36 months

Freq. no longer increasing; # of cores doubles every 24 months

Rate increases very slowly

Capacity doubles every 18 months



Rate doubles every 36 months

Capacity doubles every 18 months

Trend Summary (cont'd)

Storage is cheap and capacity increases exponentially

Most transfer rates increase exponentially but slower
→ gap between capacity and transfer rate increases

Multiple channels/disks alleviate problem but don't come for free

» E.g., disk striping increases block size

Datacenter apps must carefully select where to place computations & data

Challenge and Opportunity

Accessing disk very slow: 1,000s to read/write 100GB from/to disk

- » Transfer rate not increasing
- » Will get worse: 512 GB per node in one-two years
- » Faster to access remote memory!
- » SSDs not widely deployed in datacenters

Challenge and Opportunity (cont'd)

A few node cluster = 1TB RAM: enough to handle many large datasets

» E.g., 1 billion users, 1KB metadata (on average)

# of cores doubles only every 24 months → memory/core increases exponentially

Judiciously using RAM is key

Existing Open Stack…

..mostly focused on large on-disk datasets

- » Massive data and sophisticated data processing, but slow

We add RAM to the mix

- » Enable interactive queries and data streaming: speedup queries and iterative algorithms by up to 30X
- » Dramatically increase ability to explore / mine the data

(SSDs in the future)

BDAS Software Stack

Leverage open source ecosystem (e.g., HDFS, Hadoop)

Abstractions to take advantage of storage hierarchy
- » Many real-world working sets fit in memory → RDDs
- » Controllable data placement to minimize communication

Virtualize cluster resources (Mesos)
- » Allow multiple frameworks to share cluster and data → Resource offers

Simplify parallel *programming*
- » Scala interface to Spark
- » Shark distributed SQL engine

Project History

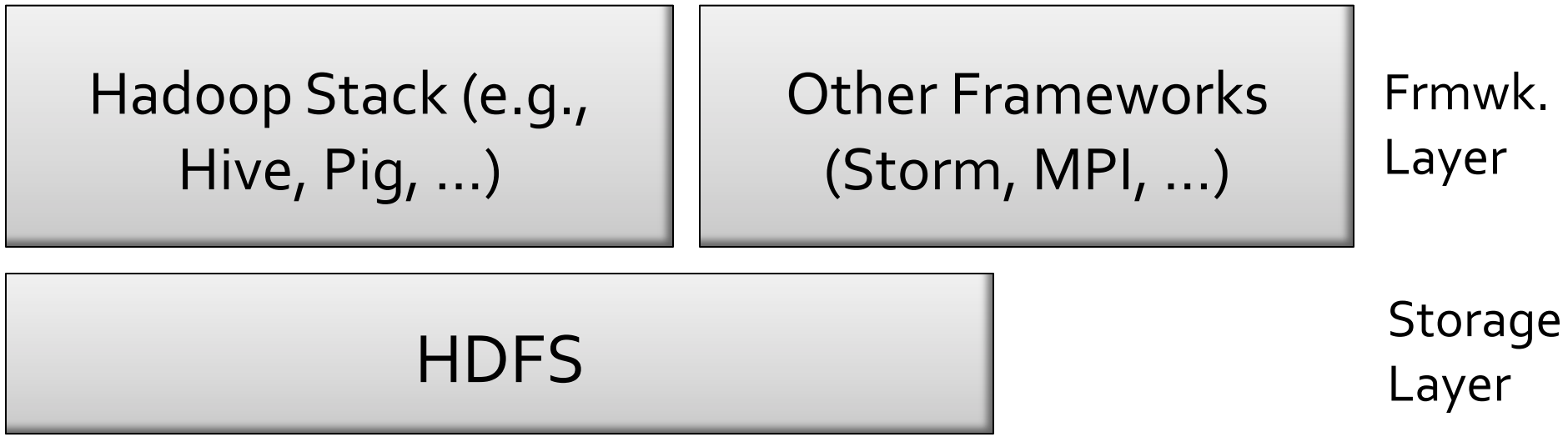
Mesos started in early 2009, open sourced 2010

Spark started in late 2009, open sourced 2010

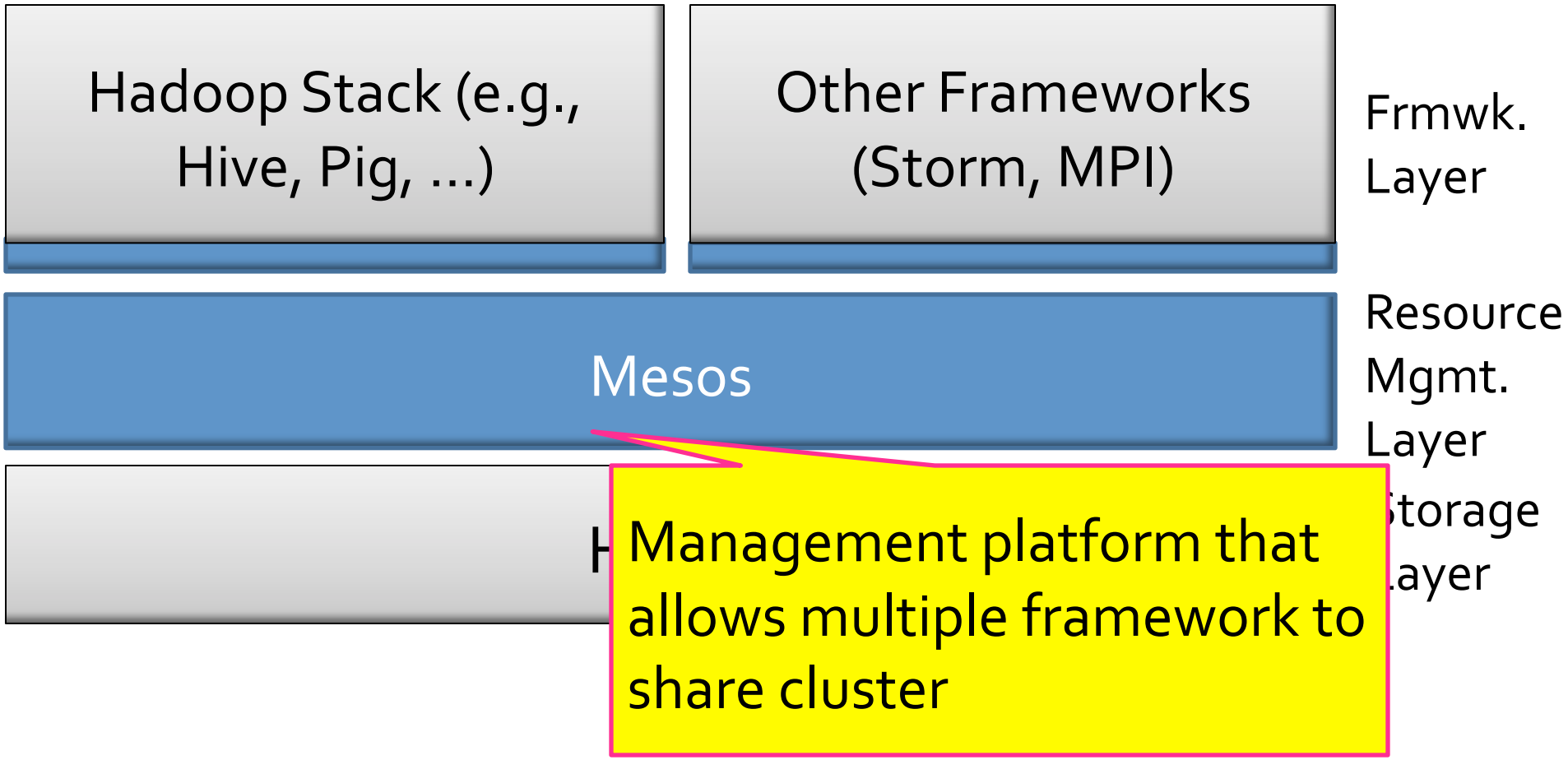
Shark started summer 2011, alpha April 2012

Used at Twitter, Foursquare, Klout, Quantifind, Conviva, Yahoo! Research, Airbnb & others

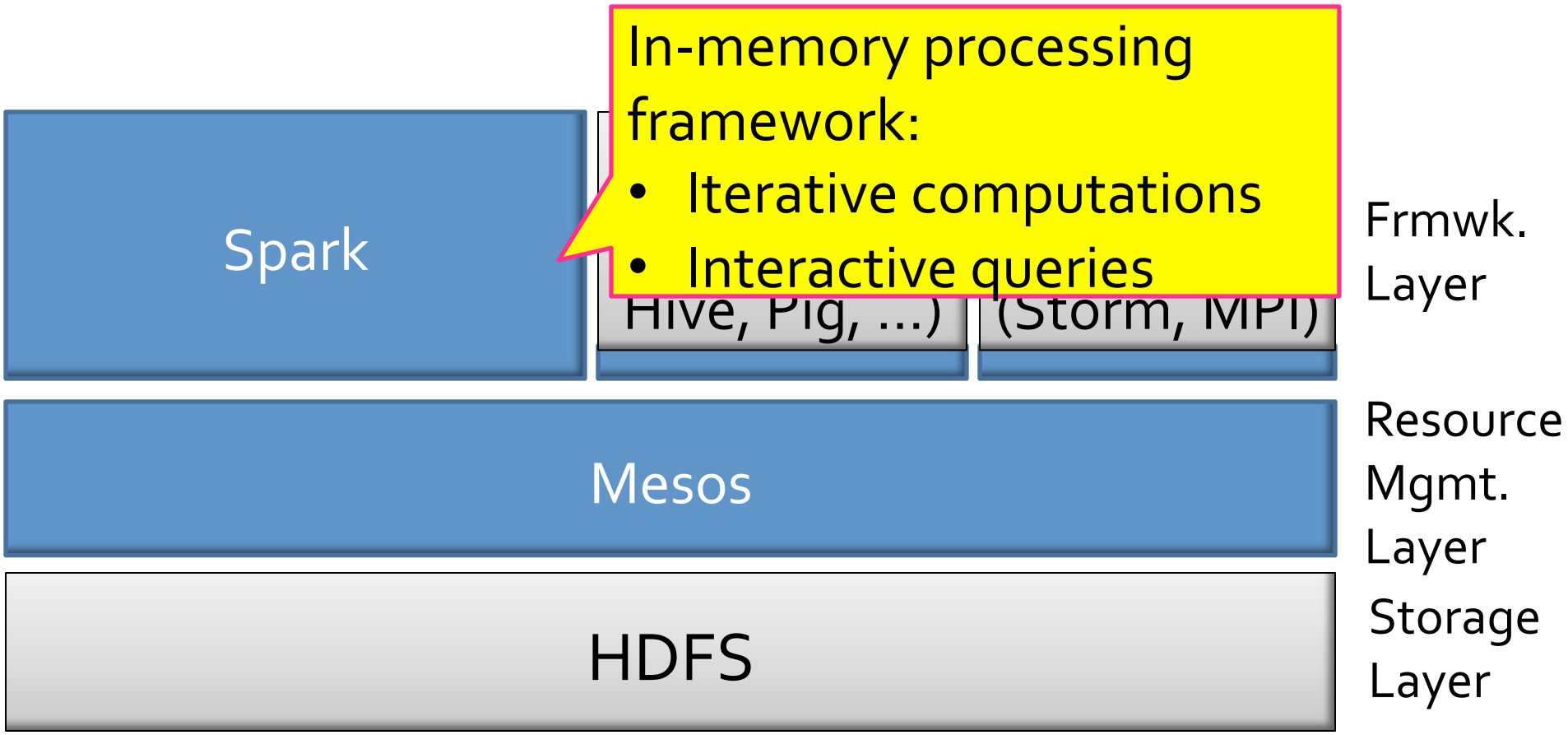
Today's Open Analytics Stack



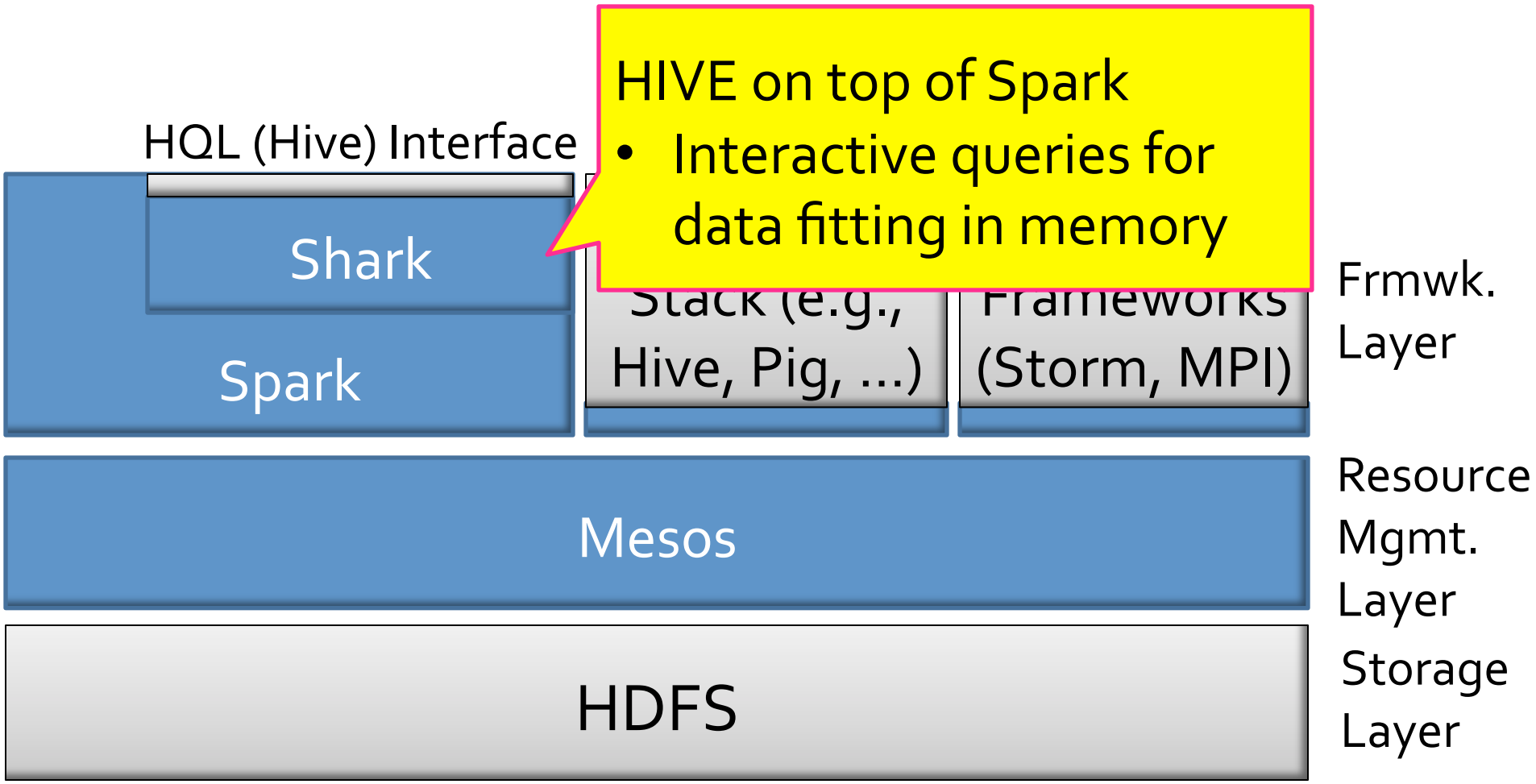
BDAS Software Stack



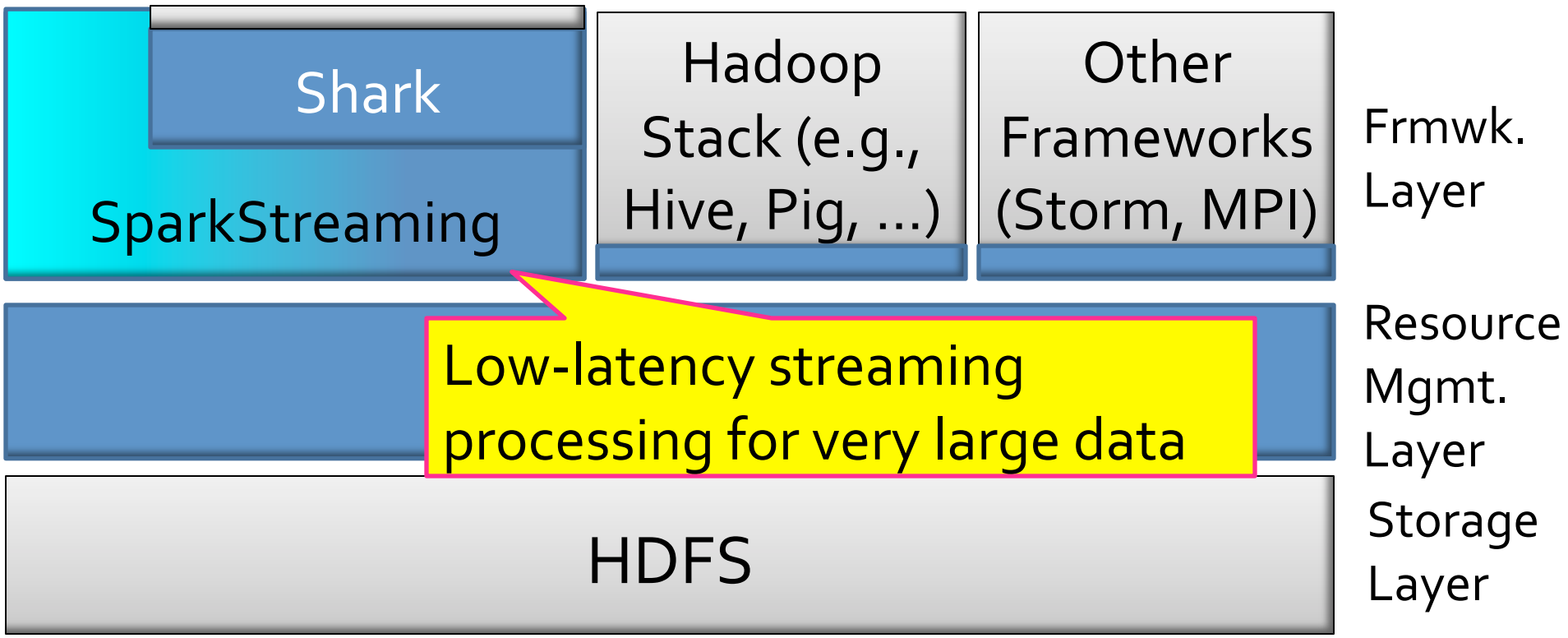
BDAS Software Stack



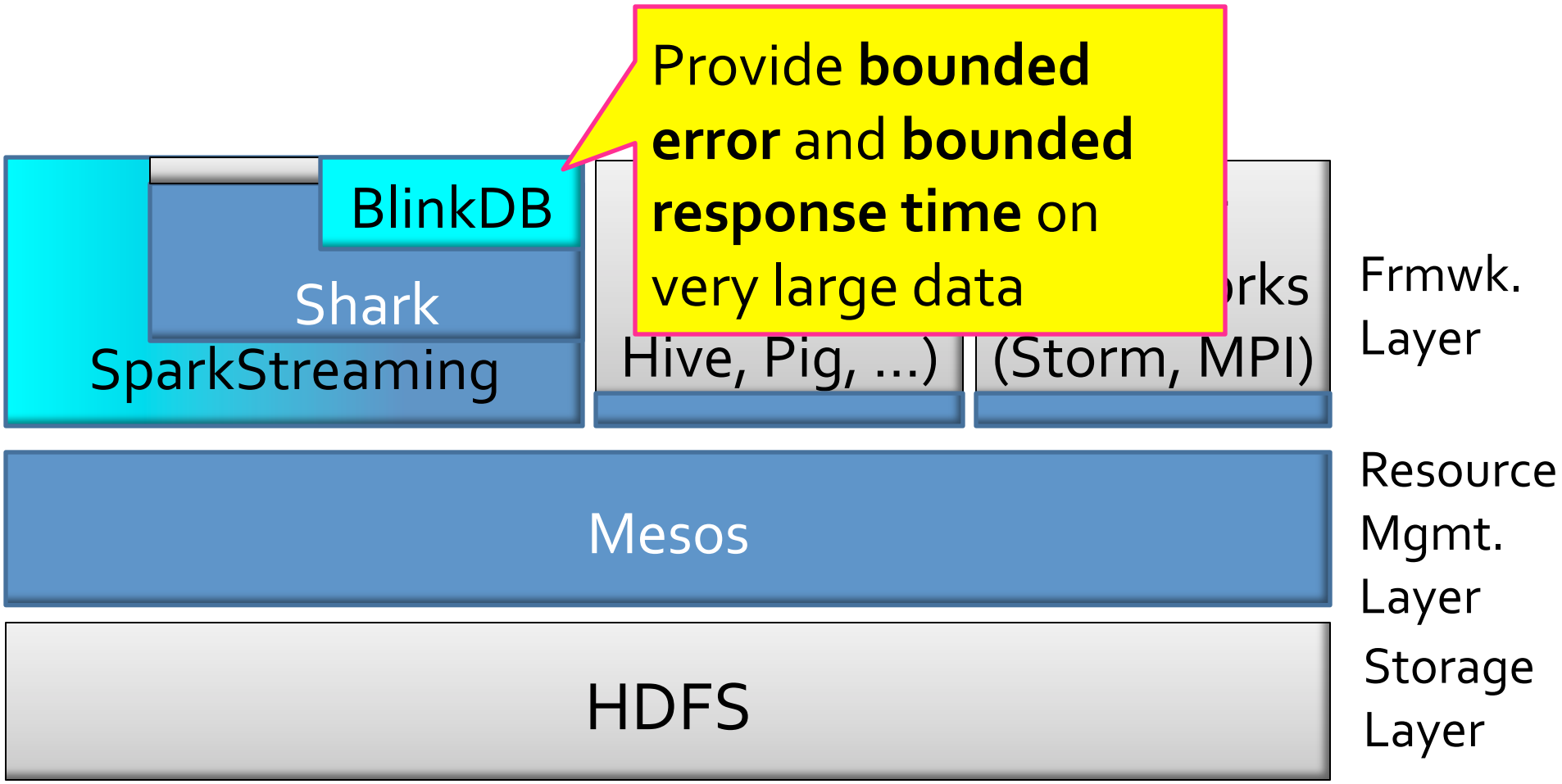
BDAS Software Stack



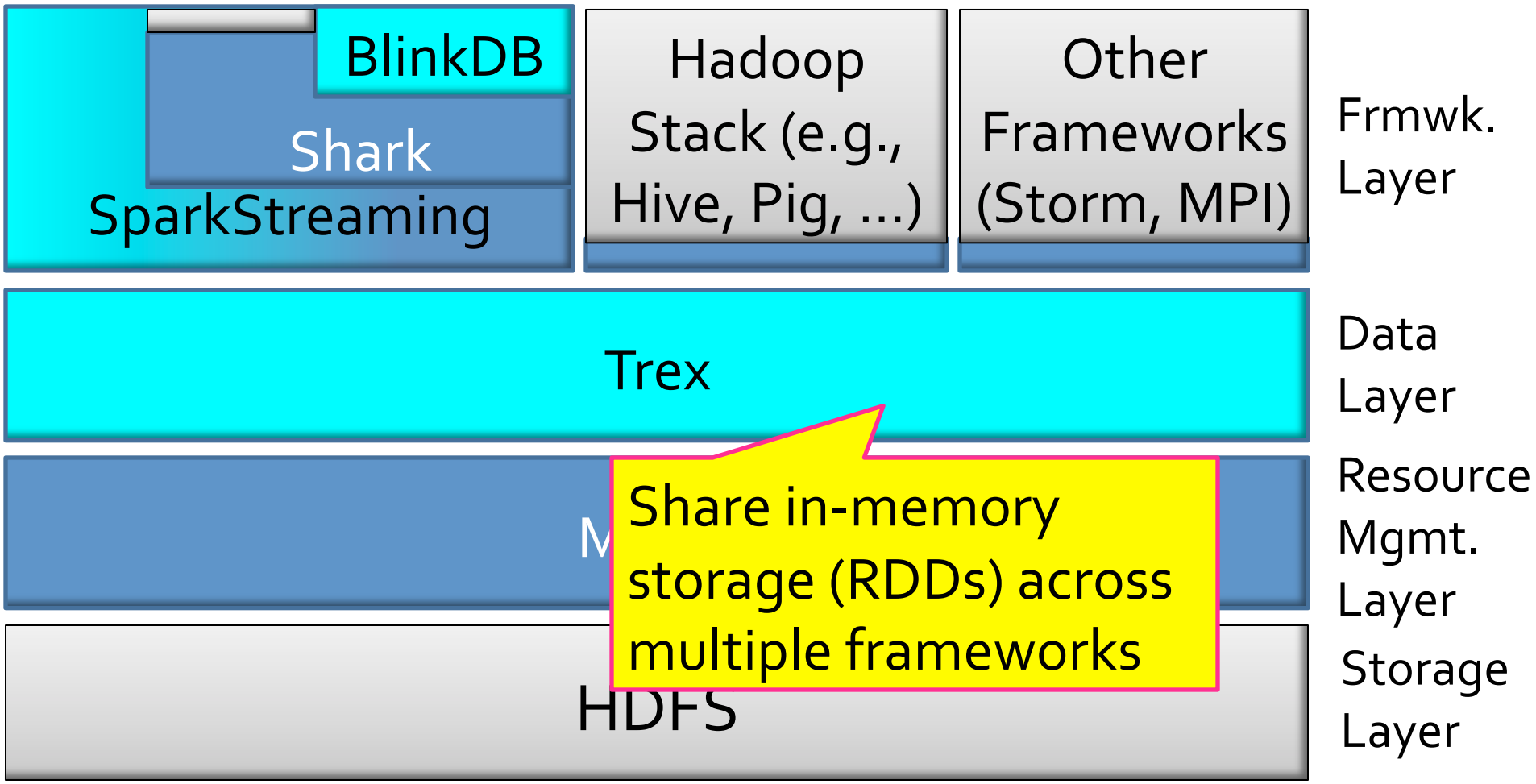
BDAS Software Stack



BDAS Software Stack



BDAS Software Stack



Summary

Today: first iteration of BDAS Software Stack

- » Mesos: enable multiple frameworks to share cluster resources and data
- » Spark: enable interactive and iterative computations through the use of RDDs
- » Shark: enable interactive Hive queries

Next: full stack to allow users to trade between answer (1) quality, (2) response time, and (3) cost

Fully compatible with open standards