Distributed Systems

COS 418: *Distributed Systems*
Lecture 1, 2017

Mike Freedman

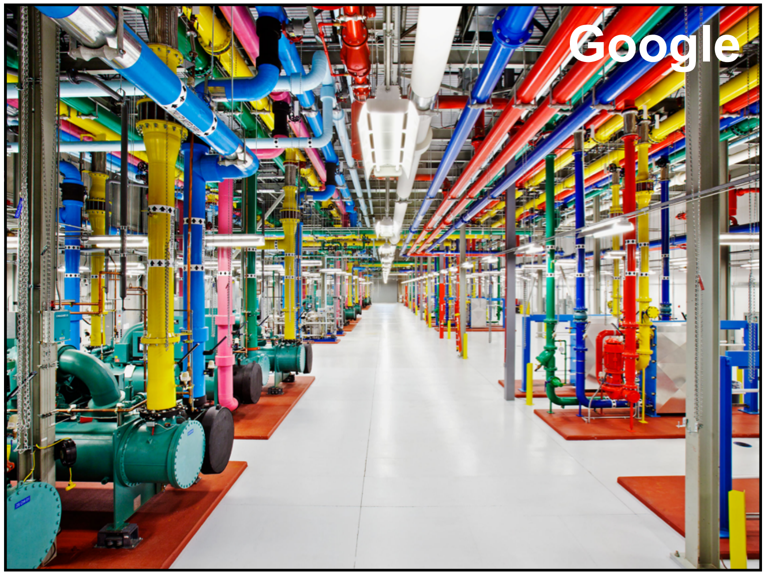
Backrub (Google) 1997

2

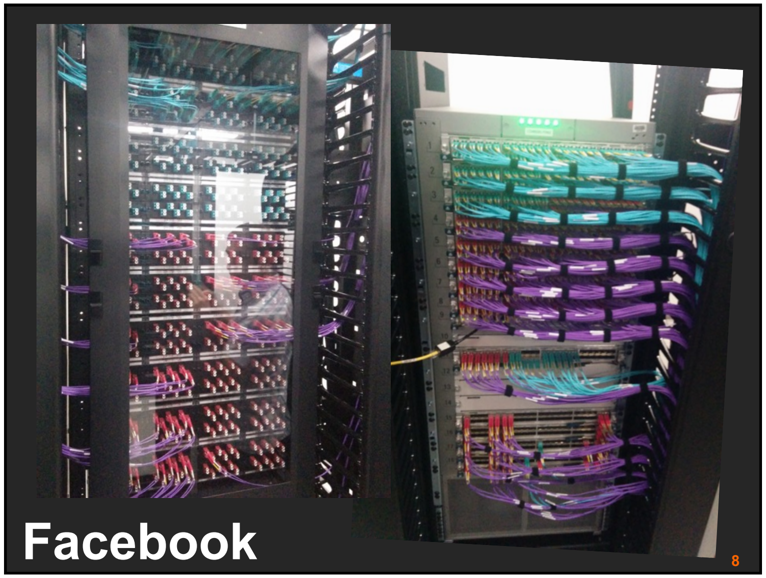
“The Cloud” is not amorphous

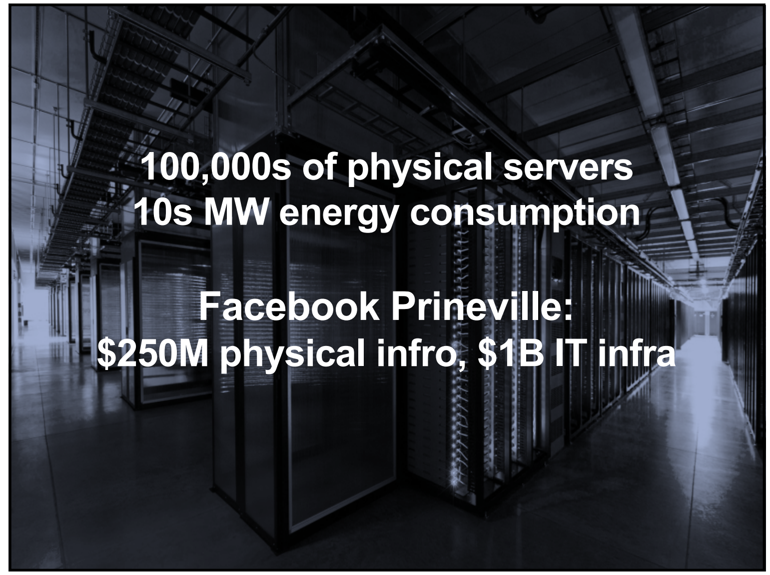
4


Microsoft
5

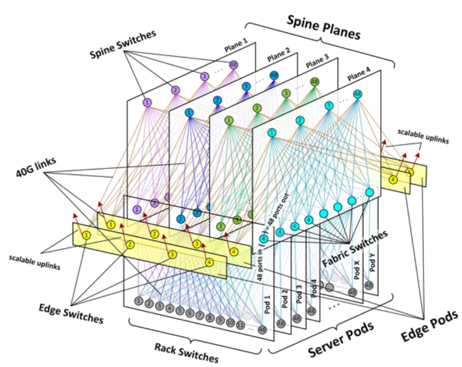

Google


Facebook


Facebook
8



Everything changes at scale



"Pods provide 7.68Tbps to backplane"

The goal of "distributed systems"

- Service with higher-level abstractions/interface
  - e.g., file system, database, key-value store, programming model, RESTful web service, …
- Hide complexity
  - Scalable (scale-out)
  - Reliable (fault-tolerant)
  - Well-defined semantics (consistent)
  - Security
- Do "heavy lifting" so app developer doesn't need to

Research results matter: NoSQL

Dynamo: Amazon's Highly Available Key-value Store

Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall and Werner Vogels

Amazon.com

ABSTRACT

Reliability at massive scale is one of the biggest challenges we face at Amazon.com, one of the largest e-commerce operations in the world; even the slightest outage has significant financial consequences and impacts customer trust. The Amazon.com platform, which provides services for many web sites worldwide, is implemented on top of an infrastructure of tens of thousands of servers and network components located in many datacenters around the world. At this scale, small and large components fail continuously and the way persistent state is managed in the face of these failures drives the reliability and scalability of the software systems.

This paper presents the design and implementation of Dynamo, a highly available key-value storage system that some of Amazon's core services use to provide an "always-on" experience. To achieve this level of availability, Dynamo sacrifices consistency under certain failure scenarios. It makes extensive use of object versioning and application-assisted conflict resolution in a manner that provides a novel interface for developers to use.

Categories and Subject Descriptors

D.4.2 [**Operating Systems**]: Storage Management; D.4.5 [**Operating Systems**]: Reliability; D.4.2 [**Operating Systems**]: Performance;

General Terms

Algorithms, Management, Measurement, Performance, Design, Reliability.

One of the lessons our organization has learned from operating Amazon's platform is that the reliability and scalability of a system is dependent on how its application state is managed. Amazon uses a highly decentralized, loosely coupled, service oriented architecture consisting of hundreds of services. In this environment there is a particular need for storage technologies that are always available. For example, customers should be able to view and add items to their shopping cart even if disks are failing, network routes are flapping, or data centers are being destroyed by tornados. Therefore, the service responsible for managing shopping carts requires that it can always write to and read from its data store, and that its data needs to be available across multiple data centers.

Dealing with failures in an infrastructure comprised of millions of components is our standard mode of operation; there are always a small but significant number of server and network components that are failing at any given time. As such Amazon's software systems need to be constructed in a manner that treats failure handling as the normal case without impacting availability or performance.

To meet the reliability and scaling needs, Amazon has developed a number of storage technologies, of which the Amazon Simple Storage Service (also available outside of Amazon and known as Amazon S3), is probably the best known. This paper presents the design and implementation of Dynamo, another highly available and scalable distributed data store built for Amazon's platform. Dynamo is used to manage the state of services that have very high reliability requirements and need tight control over the

Research results matter: Paxos

The Chubby lock service for loosely-coupled distributed systems

Mike Burrows, *Google Inc.*

Abstract

We describe our experiences with the Chubby lock service, which is intended to provide coarse-grained locking as well as reliable (though low-volume) storage for a loosely-coupled distributed system. Chubby provides an interface much like a distributed file system with advisory locks, but the design emphasis is on availability and reliability, as opposed to high performance. Many instances of the service have been used for over a year, with several of them each handling a few tens of thousands of clients concurrently. The paper describes the initial design and expected use, compares it with actual

example, the Google File System [7] uses a Chubby lock to appoint a GFS master server, and Bigtable [3] uses Chubby in several ways: to elect a master, to allow the master to discover the servers it controls, and to permit clients to find the master. In addition, both GFS and Bigtable use Chubby as a well-known and available location to store a small amount of meta-data; in effect they use Chubby as the root of their distributed data structures. Some services use locks to partition work (at a coarse grain) between several servers.

Before Chubby was deployed, most distributed systems at Google used *ad hoc* methods for primary elec-

Research results matter: MapReduce

MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

Google, Inc.

Abstract

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model, as shown in the paper.

Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine communication. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system.

Our implementation of MapReduce runs on a large cluster of commodity machines and is highly scalable: a typical MapReduce computation processes many ter-

given day, etc. Most such computations are conceptually straightforward. However, the input data is usually large and the computations have to be distributed across hundreds or thousands of machines in order to finish in a reasonable amount of time. The issues of how to parallelize the computation, distribute the data, and handle failures conspire to obscure the original simple computation with large amounts of complex code to deal with these issues.

As a reaction to this complexity, we designed a new abstraction that allows us to express the simple computations we were trying to perform but hides the messy details of parallelization, fault-tolerance, data distribution and load balancing in a library. Our abstraction is inspired by the *map* and *reduce* primitives present in Lisp and many other functional languages. We realized that most of our computations involved applying a *map* operation to each logical "record" in our input in order to compute a set of intermediate key/value pairs, and then applying a *reduce* operation to all the values that shared the same key, in order to combine the derived data appropriately. Our use of a functional model with user-

hadoop

Spark

kafka

Flink

S4 distributed stream computing platform

STORM

Course Organization

Learning the material: People

- Lecture
  - Professors Mike Freedman, Kyle Jamieson
  - Slides available on course website
  - Office hours immediately after lecture
- Precept:
  - TAs Charlie Murphy, Andrew Or
- Main Q&A forum: www.piazza.com
  - Graded on class participation: so ask & answer!
  - No anonymous posts or questions
  - Can send private messages to instructors

17

Learning the Material: Lectures!

- **Attend lecture / precepts and take notes!**
  - Lecture slides posted day/night before
  - Recommendation: Print slides & take notes
  - Not everything covered in class is on slides
  - You are responsible for everything covered in class
- **Precepts are mandatory (attendance taken)**
- **No required textbooks**
  - Links to Go Programming textbook and two other distributed systems textbooks on website

18

Grading

- Five assignments (10% each)
  - 90% 24 hours late, 80% 2 days late, 50% >5 days late
  - **THREE** free late days (we'll figure which one is best)
  - Only failing grades I've given are for students who don't (try to) do assignments
- Two exams (45% total)
  - Midterm exam before spring break (20%)
  - Final exam during exam period (25%)
- Class participation (5%)
  - In lecture, precept, and Piazza

19

Policies: Write Your Own Code

Programming is an individual creative process. At first, discussions with friends is fine. When writing code, however, the program must be your own work.

Do not copy another person's programs, comments, README description, or any part of submitted assignment. This includes character-by-character transliteration but also derivative works. Cannot use another's code, etc. even while "citing" them.

Writing code for use by another or using another's code is academic fraud in context of coursework.

Do not publish your code e.g., on github, during/after course!

20

Policies: Write Your Own Code

Programming is an individual creative process. At first, discussions with friends is fine. When writing code, however, the program must be your own work.

Do not copy another person's programs, comments, README description, or any part of submitted assignment. This includes character-by-character transliteration but also derivative works. Cannot use another's code, etc. even while "citing" them.

Writing code for use by another or using another's code is academic fraud in context of coursework.

Do not publish your code e.g., on github, during/after course!

Don't Plagiarize!

21

Assignment 1 (in three parts)

- Learn how to program in Go
  - Basic Go assignment (due **Sept 21**)
  - "Sequential" Map Reduce (due Sept 28)
  - Distributed Map Reduce (due Oct 5)

22

Warnings

This is a 400-level course,
with expectations to match.

23

Warning #1: Assignments are a LOT of work

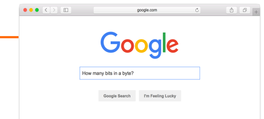
- Assignment 1 is purposely easy to teach Go. Don't be fooled.
- Last year we gave 3-4 weeks for later assignments; many students started 3-4 days before deadline. **Disaster**.
- Distributed systems are hard
  - These aren't simple "CRUD" interfaces
  - Need to understand problem and protocol, carefully design
  - Can take 5x more time to debug than "initially program"
- Assignment #4 builds on your Assignment #3 solution, i.e., you can't do #4 until your own #3 is working! (That's the real world!)

24

Warning #2: Software engineering, not just programming

- COS126, 217, 226 told you how to design & structure your programs. This class doesn't.
- Real software engineering projects don't either.
- You need to learn to do it.
- If your system isn't designed well, can be *significantly* harder to get right.
- Your friend: test-driven development
  - We'll supply tests, bonus points for adding new ones

Warning #3: Don't expect 24x7 answers

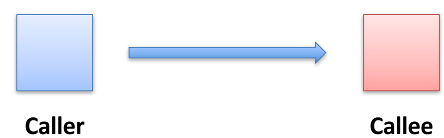


- Try to figure out yourself
- Piazza not designed for debugging
  - Utilize right venue: Go to lab, office hours; sit near friends
  - Send detailed Q's / bug reports, not "no idea what's wrong"
- Instructors are not on pager duty 24 x 7
  - Don't expect response before next business day
  - Questions Friday night @ 11pm should not expect fast responses. Be happy with something before Monday.
- Implications
  - Students should answer each other (+ it's worth credit)
  - Start your assignments early!

Naming and layering

(Data-parallel programming at scale)

Naming and system components



- How to design interface between components?
- Many interactions involve naming things
  - Naming objects that caller asks callee to manipulate
  - Naming caller and callee together

Potential Name Syntax

- Human readable?
  - If users interact with the names
- Fixed length?
  - If equipment processes at high speed
- Large name space?
  - If many nodes need unique names
- Hierarchical names?
  - If the system is very large and/or federated
- Self-certifying?
  - If preventing "spoofing" is important

Properties of Naming

- Enabling sharing in applications
  - Multiple components or users can name a shared object.
  - Without names, client-server interface pass entire object by value
- Retrieval
  - Accessing same object later on, just by remembering name
- Indirection mechanism
  - Component A knows about name N
  - Interposition: can change what N refers to without changing A
- Hiding
  - Hides impl. details, don't know where google.com located
  - For security purposes, might only access resource if know name (e.g., dropbox or Google docs URL –> knowledge gives access)

Names all around…

- Registers: LD R0, 0x1234
- IP addresses: 128.112.132.86
- Host names: www.cs.princeton.edu
- Path names: /courses/archive/spring17/cos518/syllabus.html vs. "syllabus.html"
- ".." (to parent directory)
- URLs: http://www.cs.princeton.edu/courses/archive/spring17/cos518/
- Email addresses
- Function names: ls
- Phone numbers: 609-258-9169 vs. x8-9179
- SSNs

High-level view of naming

- Set of possible names
  - Syntax and semantics?
- Set of possible values that names map to
- Lookup algorithm that translates name to value
  - What is context used to resolve (if any)?
  - Who supplies context?

Different Kinds of Names

- **Host names:** www.cs.princeton.edu
  - Mnemonic, variable-length, appreciated *by humans*
  - Hierarchical, based on organizations
- **IP addresses**: 128.112.7.156
  - Numerical 32-bit address appreciated *by routers*
  - Hierarchical, based on organizations and topology
- **MAC addresses** : 00-15-C5-49-04-A9
  - Numerical 48-bit address appreciated *by adapters*
  - Non-hierarchical, unrelated to network topology

Hierarchical Assignment Processes

- **Host names:** www.cs.princeton.edu
  - Domain: registrar for each top-level domain (eg, .edu)
  - Host name: local administrator assigns to each host
- **IP addresses:** 128.112.7.156
  - Prefixes: ICANN, regional Internet registries, and ISPs
  - Hosts: static configuration, or dynamic using DHCP
- **MAC addresses:** 00-15-C5-49-04-A9
  - Blocks: assigned to vendors by the IEEE
  - Adapters: assigned by the vendor from its block

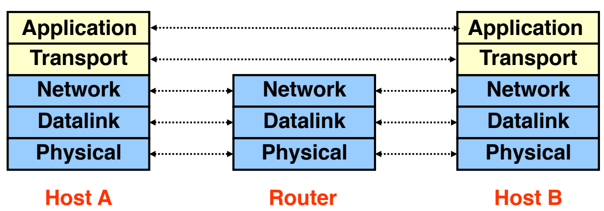
Layering

Layering: abstractions, abstractions, abstractions …

- Partition the system
  - Each layer solely relies on services from layer below
  - Each layer solely exports services to layer above
- Interface between layers defines interaction
  - Hides implementation details
  - Layers can change without disturbing other layers

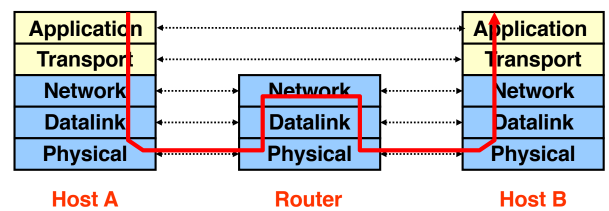
Five Layers Summary

- Lower three layers implemented everywhere
- Top two layers implemented only at hosts
- Logically, layers interacts with peer's corresponding layer



Physical Communication

- Communication goes down to physical network
- Then from network peer to peer
- Then up to relevant layer



Drawbacks of Layering

- Layer N may duplicate layer N-1 functionality
  - E.g., error recovery to retransmit lost data
- Layers may need same information
  - E.g., timestamps, maximum transmission unit size
- Layering can hurt performance
  - E.g., hiding details about what is really going on
- Some layers are not always cleanly separated
  - Inter-layer dependencies for performance reasons
  - Some dependencies in standards (header checksums)
- Headers start to get really big
  - Sometimes header bytes >> actual content

Placing Network Functionality

- Hugely influential paper: "End-to-End Arguments in System Design" by Saltzer, Reed, and Clark (1984)
- "Sacred Text" of the Internet
  - Endless disputes about what it means
  - Everyone cites it as supporting their position

Monday:

Network communication and RPC

41