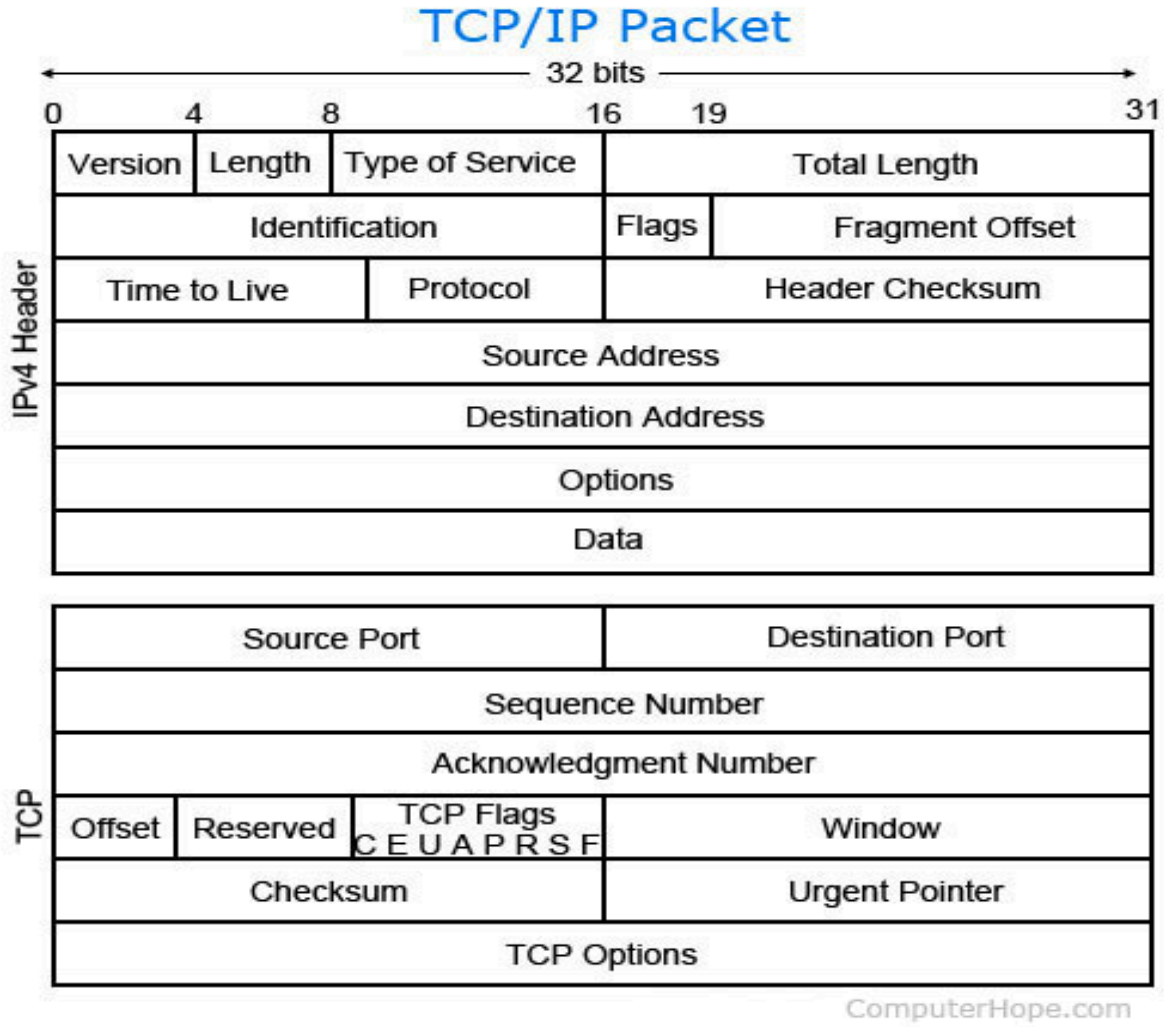
Foundation of Internet Platform Development & Operation

Network II

2019-09-24

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Network Packet

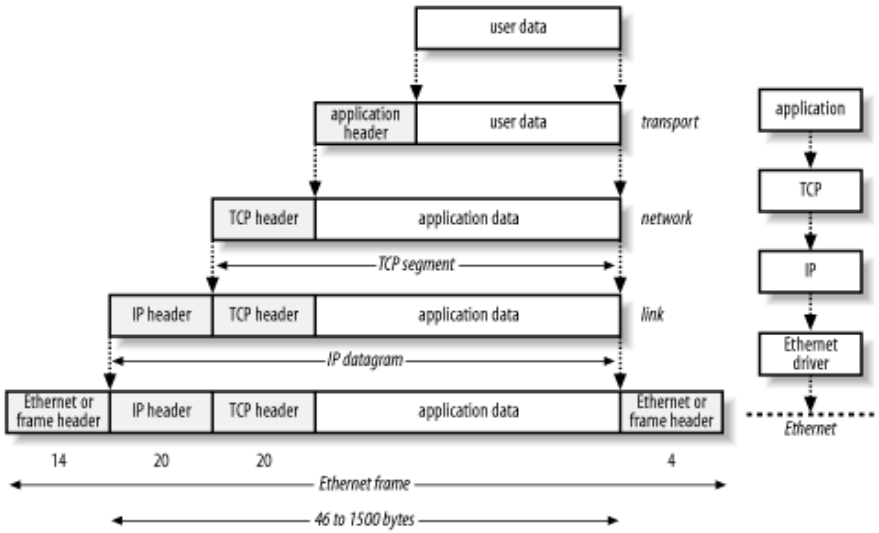


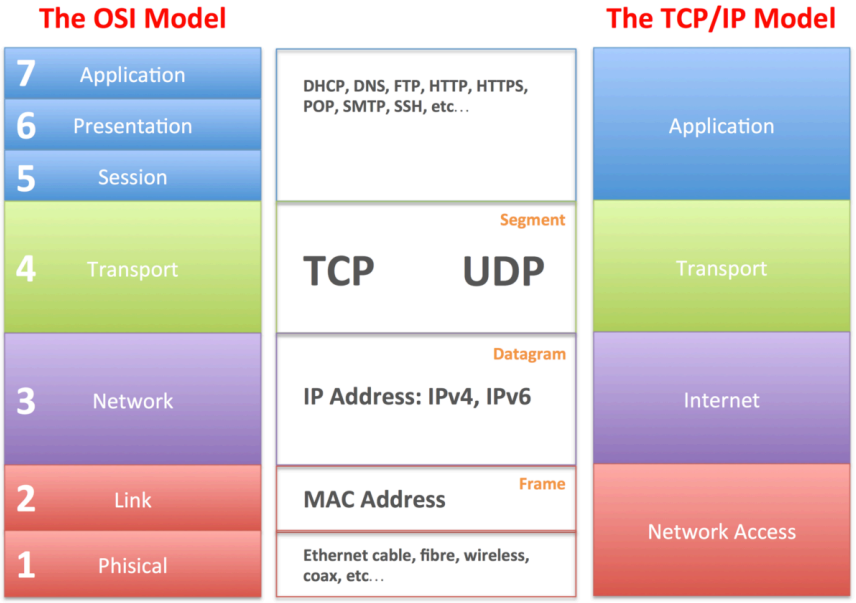
Network Address Translation

- NAT (Network Address Translation) allows a network device (such as a Router, Firewall or a Server running Network Operating Systems to translate addresses between the public internet and a local private network
  - Static NAT
  - Pooled NAT
  - NAPT
    - SNAT
    - DNAT

NAT VS. Proxy

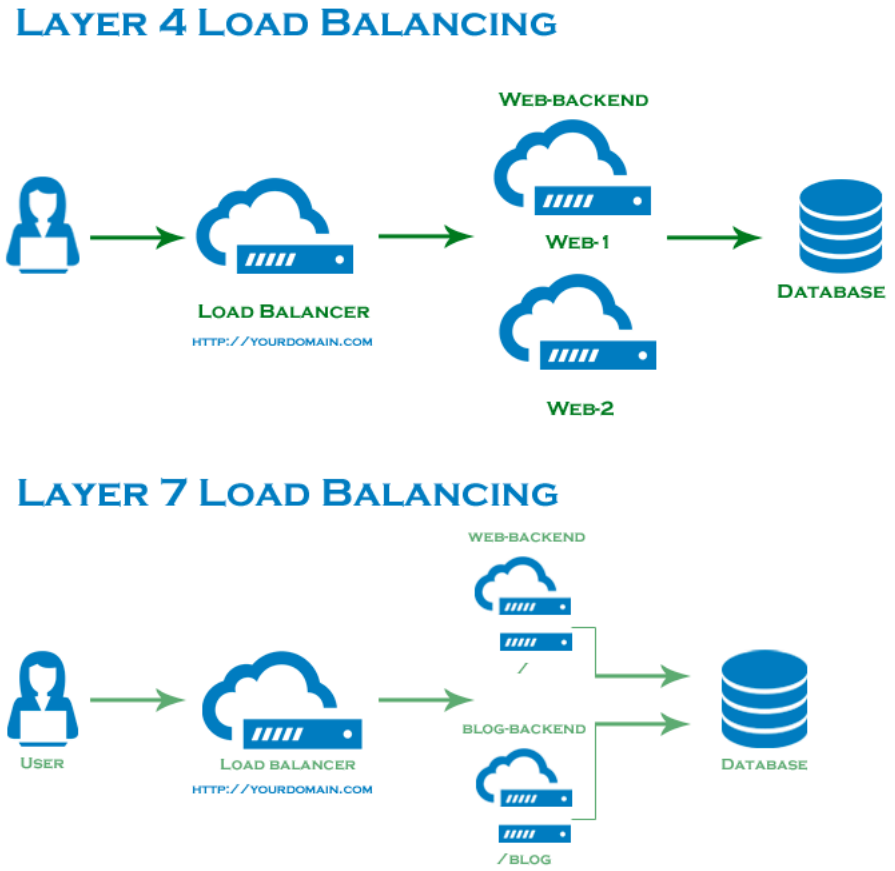
- NAT
  - 4th layer





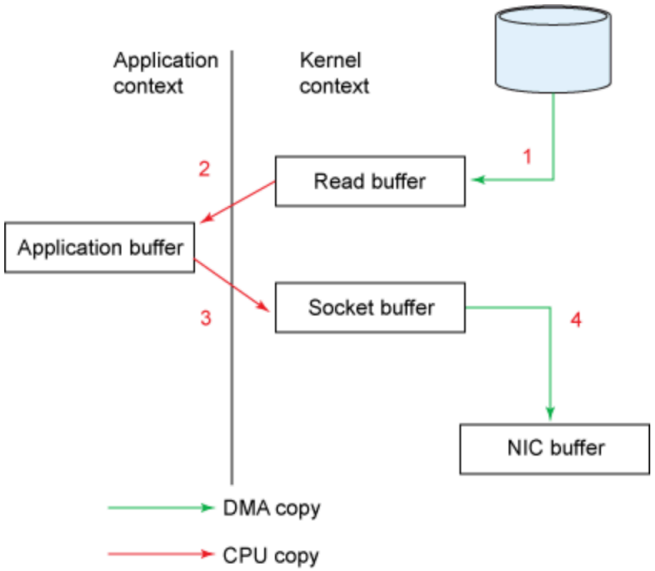
NAT VS. Proxy

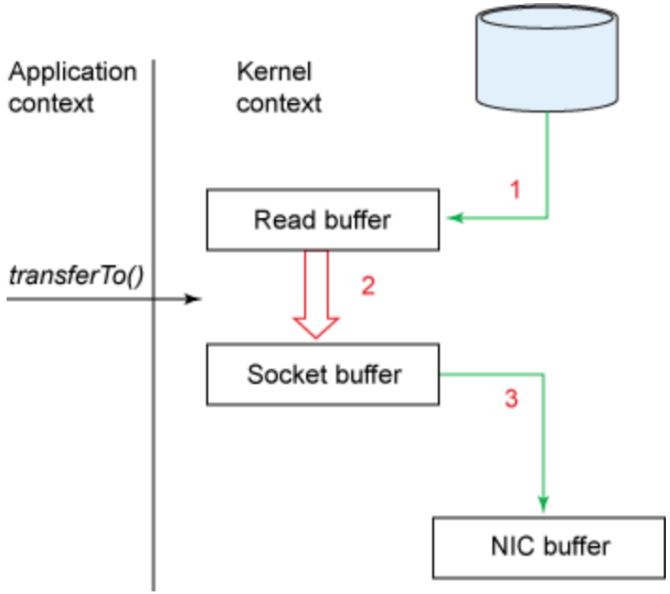
- Proxy
  - 4th layer
    - F5, Array (Hardware)
    - Nginx
    - lvs
    - HAProxy
  - 7th layer
    - Nginx
    - HAProxy



DPDK

- Zero-copy
  - Kernel bypass





DPDK
DATA PLANE DEVELOPMENT KIT

Advanced Network Techs

- InfiniBand
  - InfiniBand is an industry-standard specification that defines an input/output architecture used to interconnect servers, communications infrastructure equipment, storage and embedded systems
- RDMA
  - Remote Direct Memory Access (RDMA) is a technology that allows computers in a network to exchange data in main memory without involving the processor, cache or operating system of either computer

Question

- Which type of load balance should be used for CQRS?
- There are two isolated website (www.a.com, www.b.com) in a server with only one IP. How to configure the server to allow users access the two website using their own urls without port number?

Foundation of Internet Platform Development & Operation

Distributed Storage

2019-09-24

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Distributed Storage

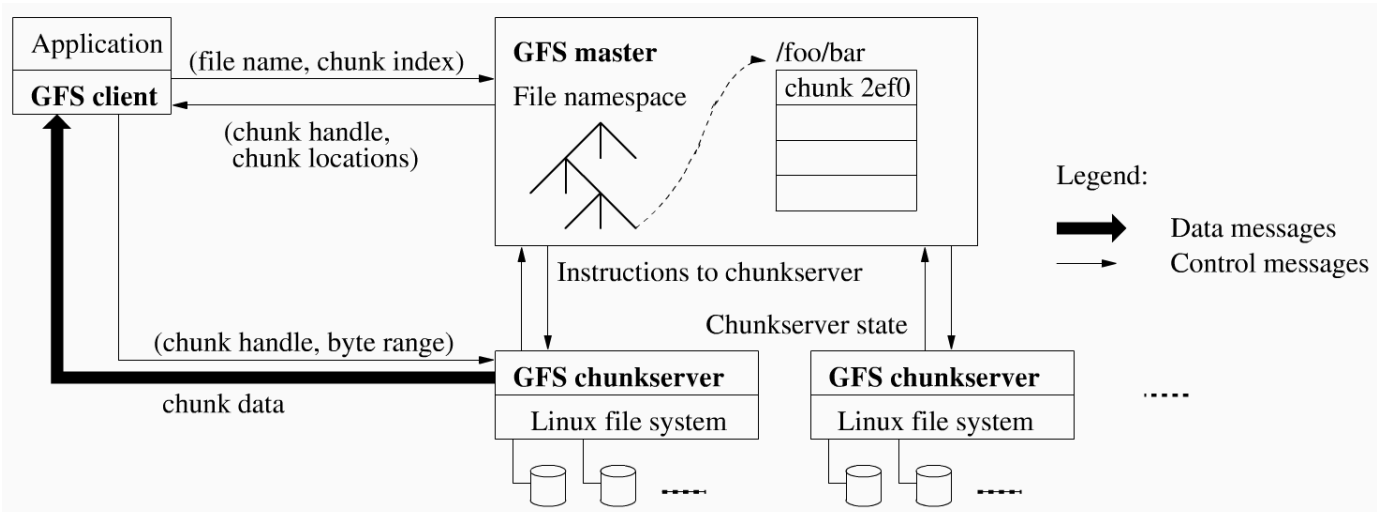
- Challenges:
  - Need to store/access massive data sets efficiently
  - Want to use really cheap (i.e., unreliable) hardware
  - Spread data over many unreliable machines => failure is the norm
- Solutions
  - GFS
    - SOSP 2003
  - Ceph
    - OSDI 2006

GFS

- Objectives
  - Cheap hardware
    - Unstable
  - Workload
    - Write once, read many times

GFS

- Architecture
  - One master server (with replicated state)
  - Many chunk servers (hundreds, maybe thousands)
    - Broken into different racks; less bandwidth between racks
    - Store 64MB chunks of files, identified by globally unique ID
  - Potentially many clients accessing same or different files on cluster



GFS

- Challenges
  - Not really a FS--just a library applications can use to access storage
  - Mutation:
    - Consistent - read same data from all replicas of chunk
    - Defined - state equivalent to serial application of client operations

GFS

- Master
  - File names stored in prefix-compressed form
  - Ownership, permission, etc.
  - Mappings:
    - File -> Chunks (*)
    - Chunk -> Version (*), Replicas, Reference count (*), Leases (*) = stored stably in a log
  - Average: 100 bytes per file

GFS

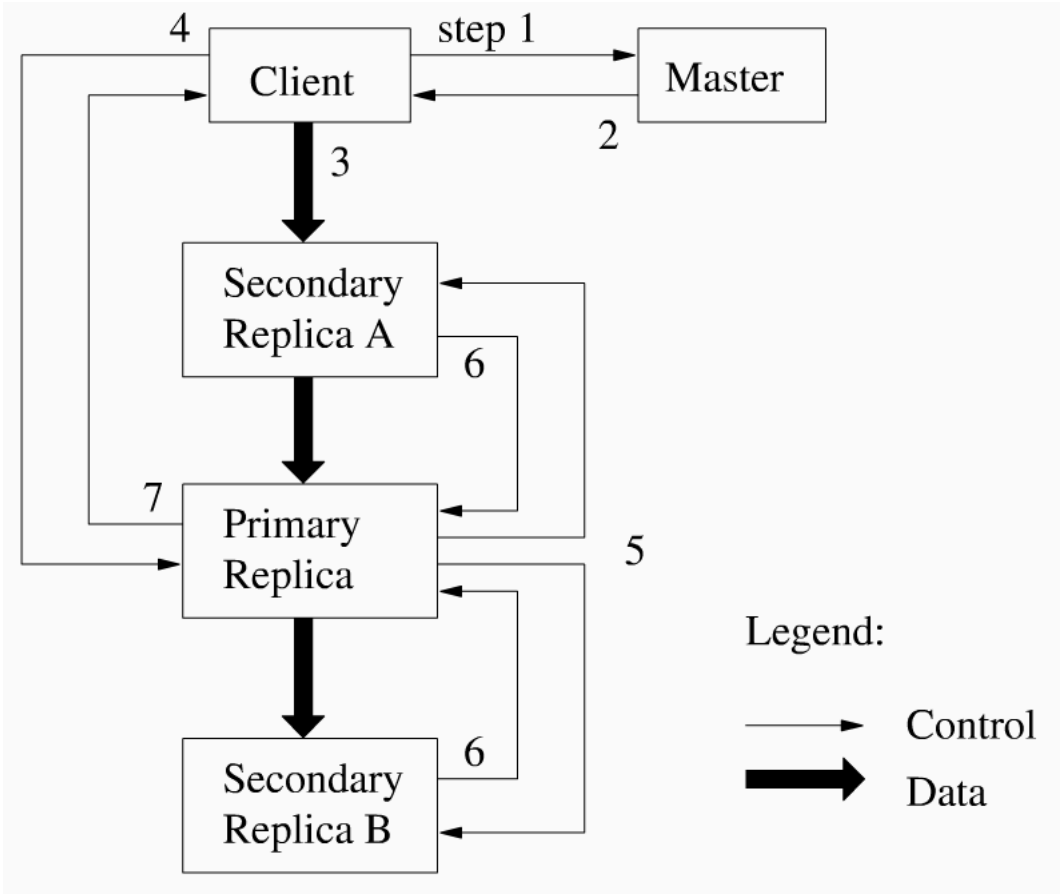
- Chunk server
  - Chunks (one Linux file per chunk)
  - Chunk Metadata: Version number, checksum

GFS

- Read a file
  - Ask server for (file, chunk index)
    - Server returns chunk ID, version, locations of replicas
    - Server may return info for subsequent chunks; client can cache
  - Read chunk from nearest chunk server
    - Nearest easy to determine because of Google's simple network topology
      - ***How to define nearest?***

GFS

- Write a file



GFS

- Write a file
  - One of the chunk servers is the *primary* for the chunk
    - This is determined by having a lease from the master server
    - Master increments version number every time it grants a lease
    - Leases are renewed through periodic heartbeats
  - Client asks master who the primary and secondary replicas are for chunk
  - Client sends data to all replicas, in chain fashion
    - Why? To make best use of topology and max out upstream bandwidth
    - Trade-off is high latency... Could you have lower latency?
    - Maybe client stripes chunks across replicas and have them reconcile?

GFS

- Replication
  - Why to replicate chunks?
  - Creation
    - ***Decided by master according to?***
  - Re-replication
    - Prioritized
      - ***Why?***
  - Re-balance
    - ***Why periodically?***

GFS

- Lease
  - Revoking
    - Rename a file
    - Snapshot

GFS

- Master reboots
  - Reconstruct state: Ask chunk servers what chunks they have
  - What if chunk server has lower version number?
    - Consider it stale--that server no longer has a replica
  - What if chunk server has higher version number?
    - Master updates it's own version number; must have crashed before logging it