Computer Networks: Overview & Principles

Ankit Singla

Some slides adapted from Brighten Godfrey, Jennifer Rexford, Scott Shenker, Laurent Vanbever

ETH Zürich
Computer Networks

Part 1: Overview

#1 What is a network made of?

#2 How is it shared?

#3 How does communication happen?

#4 How do we characterize it?
The Internet should allow processes on different hosts to exchange data
everything else is just commentary…
How do you exchange data in a network as complex as this?
To exchange data, Alice and Bob use a set of network protocols
A protocol is like a conversational convention:
who should talk next and how they should respond

Alice

Bob

hello

hello

give me http://ndal.ethz.ch/

here it is
Sometimes implementations are not compliant…
Each protocol is governed by a specific interface

```plaintext
while (...) {
    message = ...;
    send(message, ...);
}
```

```plaintext
while (...) {
    message = receive(...);
}
```

Alice

Bob

Wow client

Wow server

Application Programming Interface
In practice, there exist a lot of network protocols. How does the Internet organize this?
HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)

SITUATION:
There are 14 competing standards.

14?! RIDICULOUS!
We need to develop one universal standard that covers everyone's use cases.
Yeah!

SOON:

SITUATION:
There are 15 competing standards.

https://xkcd.com/927/
Modularity is a key component of any good system

Problem
- can’t build large systems out of spaghetti code
- hard (if not, impossible) to understand, debug, update
- need to bound the scope of changes
- evolve the system without rewriting it from scratch

Solution
- Modularity is how we do it
- …and understand the system at a higher-level
Modularity, based on abstraction, is the way things get done.

— Barbara Liskov, MIT

Photo: Donna Coveney
To provide structure to the design of network protocols, network designers organize protocols in layers, which the network “stack” implements.
Internet communication can be decomposed in 5 independent layers (or 7 layers for the OSI model)

layer

L5  Application
L4  Transport
L3  Network
L2  Link
L1  Physical
Each layer provides a service to the layer above

<table>
<thead>
<tr>
<th>layer</th>
<th>service provided</th>
</tr>
</thead>
<tbody>
<tr>
<td>L5 L4</td>
<td>network access</td>
</tr>
<tr>
<td>L4 L3</td>
<td>end-to-end delivery (reliable or not)</td>
</tr>
<tr>
<td>L3 L2</td>
<td>global best-effort delivery</td>
</tr>
<tr>
<td>L2 L1</td>
<td>local best-effort delivery</td>
</tr>
<tr>
<td>L1</td>
<td>physical transfer of bits</td>
</tr>
</tbody>
</table>
What if no reliable transport is provided?

Every application that needs reliability has to engineer it from scratch

- Programmer burden
- Much higher likelihood of bugs
- Wasteful effort
Each layer provides a service to the layer above by using the services of the layer directly below it.

- **Applications**
  - ...built on...

- **Reliable (or unreliable) transport**
  - ...built on...

- **Best-effort global packet delivery**
  - ...built on...

- **Best-effort local packet delivery**
  - ...built on...

- **Physical transfer of bits**
Each layer has a unit of **data**

<table>
<thead>
<tr>
<th>layer</th>
<th>role</th>
</tr>
</thead>
<tbody>
<tr>
<td>L5 Application</td>
<td>exchanges <strong>messages</strong> between processes</td>
</tr>
<tr>
<td>L4 Transport</td>
<td>transports <strong>segments</strong> between end-systems</td>
</tr>
<tr>
<td>L3 Network</td>
<td>moves <strong>packets</strong> around the network</td>
</tr>
<tr>
<td>L2 Link</td>
<td>moves <strong>frames</strong> across a link</td>
</tr>
<tr>
<td>L1 Physical</td>
<td>moves <strong>bits</strong> across a physical medium</td>
</tr>
</tbody>
</table>
Each layer (except for L3) is implemented with different protocols

<table>
<thead>
<tr>
<th>layer</th>
<th>protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td>L5</td>
<td>Application HTTP, SMTP, FTP, SIP, …</td>
</tr>
<tr>
<td>L4</td>
<td>Transport TCP, UDP, SCTP</td>
</tr>
<tr>
<td>L3</td>
<td>Network IP</td>
</tr>
<tr>
<td>L2</td>
<td>Link Ethernet, Wifi, (A/V)DSL, WiMAX, LTE, …</td>
</tr>
<tr>
<td>L1</td>
<td>Physical Twisted pair, fiber, coaxial cable, …</td>
</tr>
</tbody>
</table>
The Internet Protocol (IP) acts as a unifying network layer

<table>
<thead>
<tr>
<th>layer</th>
<th>protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td>L5 Application</td>
<td>HTTP, SMTP, FTP, SIP, …</td>
</tr>
<tr>
<td>L4 Transport</td>
<td>TCP, UDP, SCTP</td>
</tr>
<tr>
<td>L3 Network</td>
<td>IP</td>
</tr>
<tr>
<td>L2 Link</td>
<td>Ethernet, Wifi, (A/V)DSL, Cable, LTE, …</td>
</tr>
</tbody>
</table>
| L1 Physical        | Twisted pair, fiber, coaxial cable, …
Each layer takes messages from the layer above, and *encapsulates* with its own header and/or trailer.
HTTP(S) → TCP/UDP → IP → Ethernet → your laptop

Header: HA
Message: GET google.ch

Header: HT HA
Message: GET google.ch
GET google.ch

Header

Message

your laptop

Application

HTTP(S)

Transport

TCP/UDP

Network

IP

Link

Ethernet

HTTP(S)

TCP/UDP

IP

Ethernet
In practice, layers are distributed on every network device.
Since when bits arrive they must make it to the application, all the layers exist on a host.
Routers act as **L3 gateway** as such they implement L2 and L3.
Switches act as **L2 gateway**
as such they only implement L2
Let's see how it looks like in practice on a host, using Wireshark

https://www.wireshark.org
Modularity, based on abstraction, is the way things get done.

— Barbara Liskov, MIT

But what is the right modularity?
What is the *right* modularity?

The need for stable interfaces

The end-to-end principle

The fate-sharing principle
Each layer is implemented with different protocols and technologies.

- L5: Application
- L4: Transport
- L3: Network
- L2: Link
- L1: Physical

**Technology:**
- Software
- Hardware
Network stack challenges at increasing speeds

The 100Gbit/s challenge

Jesper Dangaard Brouer
Red Hat Inc.

Linux Conf Au, New Zealand, January 2015
Microsoft supercharges Bing search with programmable chips

Doug Burger called it Project Catapult.

Burger works inside Microsoft Research—the group where the tech giant explores blue-sky ideas—and in November 2012, he pitched a radical new concept to Qi Lu, the man who oversees Microsoft’s Bing web search engine. He wanted to completely change the machines that make Bing run, arming them with a new kind of computer processor.
The interfaces between layers should be durable

Applications
...built on...

Reliable (or unreliable) transport
...built on...

Best-effort global packet delivery
...built on...

Best-effort local packet delivery
...built on...

Physical transfer of bits

Interfaces changing often implies broken layering
The end-end principle
What if no reliable transport is provided?

Every application that needs reliability has to engineer it from scratch

- Programmer burden
- Much higher likelihood of bugs
- Wasteful effort

But what if the network layer tried to provide reliable delivery?
What if the network layer tried to provide reliable delivery?

Reliable (or unreliable) transport
...built on...

Best-effort global packet delivery
Reliable

#1: My voice call wants speedy delivery, even if it’s lossy …
What if the network layer tried to provide reliable delivery?

Reliable (or unreliable) transport

...built on...

Best-effort global packet delivery

Reliable

#2: Can the network even achieve this at all?
Example: reliably transfer file from host A to B

**Solution 1:**
Check reliability at every step (involving network layer)

**Solution 2:**
Allow unreliable steps (network layer is best-effort)
B checks and tells A to retry on failure
Example: reliably transfer file from host A to B

Solution 1:

Check reliability at every step (involving network layer)

Problem: Bugs, failures are a truth of life
Example: reliably transfer file from host A to B

Solution 1:
Check reliability at every step (involving network layer)

Problem:
Bugs, failures are a truth of life

It's not reliable transfer, if A-B depend on network that can fail
Example: reliably transfer file from host A to B

Solution 2:

Allow unreliable steps (network layer is best-effort).
B checks correctness. On failure, B tells A to retry.

Can still fail, but only if A / B themselves fail.

Solution 2 depends only on what end-points themselves control
Example: reliably transfer file from host A to B

“the end-to-end check of the file transfer application must still be implemented no matter how reliable the communication system becomes”
Question: should we ever implement reliability in the network?
Question: should we ever implement reliability in the network?

Perhaps, to reduce the number of end-end retries needed.

\[ P \text{ (retry)} = 1 - 0.90^{10} = 0.65 \]

\[ P \text{ (retry)} = 1 - 0.99^{10} = 0.10 \]
Implementing reliability in the network …

… does not reduce end-host complexity

… does increase network complexity

… often imposes overhead for apps that don’t need it

… but can enhance performance in some cases
Beyond a host-centric view …

Hosts are not the only stakeholders!

Networks want protection from malicious hosts

Things like firewalls are easier to implement in-network
The fate-sharing principle
A distributed system is one in which the failure of a computer you didn't even know existed can render your own computer unusable.

— Leslie Lamport, Microsoft Research

How do we prevent this?
The fate-sharing principle

When storing state in a distributed system, co-locate it with entities that rely on that state.

State is lost only if those entities fail; then it doesn’t matter.

Example: network connection state at end hosts.
Food for thought ...

• When do we break layering?
• How do we define “end”?
  • TCP vs. application?
  • iPhone vs. iWatch?
  • Distributed applications?
• When do we break the end-to-end argument?
Computer Networks

Part 1: Overview

#1 What is a network made of?

#2 How is it shared?

#3 How does communication happen?

#4 How do we characterize it?
Performance = bandwidth?
To get many megabits-per-second ...
To get many megabits-per-second...

4400 km
80 Km / hour
1 TB USB stick
= 40 Mbps
... but 55 hours
A network *connection* is characterized by its delay, loss rate, and throughput.

How long does it take for a packet to reach the destination?

What fraction of packets sent to a destination are dropped?

At what rate is the destination receiving data from the source?
A network *connection* is characterized by its delay, loss rate, and throughput.
Sources of network delays
Each packet suffers from several types of delays at each node along the path.

$\text{transmission delay} \oplus \text{propagation delay} \oplus \text{processing delay} \oplus \text{queuing delay} = \text{total delay}$

due to link properties
due to traffic mix & switch internals
Overall, the main culprits for the overall delay are the transmission, propagation and queuing delays.

\[
\text{transmission delay} + \text{propagation delay} + \text{processing delay} + \text{queuing delay} = \text{total delay}
\]

and tend to be tiny.
The transmission delay is the amount of time required to push all of the bits onto the link.

\[
\text{Transmission delay} = \frac{\text{packet size}}{\text{link bandwidth}}
\]

Example:

\[
\frac{1000 \text{ bits}}{10 \text{ μsec}} = 100 \text{ Mbps}
\]
The propagation delay is the amount of time required for a bit to travel to the end of the link. The propagation delay can be calculated using the formula:

\[
\text{Propagation delay} = \frac{\text{link length}}{\text{propagation speed}}
\]

\[\text{[m]} \quad \text{[m/sec]} \quad \text{(fraction of speed of light)}\]

Example:

\[
\frac{30000 \text{ m}}{2 \times 10^8 \text{ m/sec}} = 150 \mu\text{sec}
\]

(speed of light in fiber)
How long does it take for a 100 Byte packet to travel from A to B? (not considering queuing for now)

Time to transmit one bit = $10^{-6}$s

Time to transmit 800 bits = $800 \times 10^{-6}$s

Time when that bit reaches B: $10^{-6} + 10^{-3}$s

The last bit reaches B at $(800 \times 10^{-6}) + 10^{-3}$s = 1.8ms
If we have a 1 Gbps link, the total time decreases to 1.08ms.

Time to transmit one bit = $10^{-9}$s

Time to transmit 800 bits = $800 \times 10^{-9}$s

Time when that bit reaches B: $10^{-9} + 10^{-3}$s

The last bit reaches B at $(800 \times 10^{-9}) + 10^{-3}$s = 1.08ms
If we now exchange a 1GB file split in 100B packets

\[ 10^7 \times 100B \text{ packets} \]

The last bit reaches B at
\[ (10^7 \times 800 \times 10^{-9}) + 10^{-3}s \]
\[ = 8001\text{ms} \]
Different transmission characteristics imply different tradeoffs in terms of which delay dominates.

- $10^7 \times 100\text{B}$ pkt 1Gbps link: transmission delay dominates.
- $1 \times 100\text{B}$ pkt 1Gbps link: propagation delay dominates.
- $1 \times 100\text{B}$ pkt 1Mbps link: both matter.

Often, we can’t know in advance which one matters!
The queuing delay is the amount of time a packet waits (in a buffer) to be transmitted on a link.

Queuing delay is the hardest to evaluate as it varies from packet to packet.

It is characterized with statistical measures e.g., average delay & variance, probability of exceeding $x$. 
Queuing delay depends on the traffic pattern
Queuing delay depends on the traffic pattern

Transient overload!
Queues absorb transient bursts, but introduce queueing delays
The time a packet has to sit in a buffer before being processed depends on the traffic pattern

- arrival rate at the queue
- transmission rate of the outgoing link
- traffic burstiness
average packet arrival rate $a$ [packet/sec]

transmission rate of outgoing link $R$ [bit/sec]

fixed packet length $L$ [bit]

average bits arrival rate $La$ [bit/sec]

traffic intensity $La/R$
When the traffic intensity is $>1$, the queue will increase without bound, and so does the queuing delay

Golden rule

Design your queuing system, so that it operates far from that point
When the traffic intensity is $\leq 1$, queueing delay depends on the burst size.
Why does delay matter?
Can you perceive…

1 second?

500 milliseconds?

100 milliseconds?

10 milliseconds?

spoken syllable

~150 - 200 ms
Can you perceive…

1 second?

500 milliseconds?

100 milliseconds?

10 milliseconds?

spoken syllable

50 - 200 ms

Hiromi Uehara

“Kung Fu World Champion”
Can you perceive…

1 second?

500 milliseconds?

100 milliseconds?

10 milliseconds?

spoken syllable ~150 - 200 ms

Hiromi Uehara 88 ms per note
We care about very small delays!

In online services

<table>
<thead>
<tr>
<th></th>
<th>Delay</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon</td>
<td>+100ms</td>
<td>-1% revenue</td>
</tr>
<tr>
<td>Bing</td>
<td>+500ms</td>
<td>-1.2% revenue</td>
</tr>
<tr>
<td>Google</td>
<td>+400ms</td>
<td>-0.7% searches</td>
</tr>
</tbody>
</table>

[In gaming](#)

What is a nanosecond, a microsecond, … ?

Via Colin Scott (https://people.eecs.berkeley.edu/~rcs/research/interactive_latency.html)
What is a nanosecond, a microsecond, ...?
What is a nanosecond, a microsecond, … ?
What are Internet delays like?
What do Internet delays look like in practice?

[~]$ ping google.ch
What do Internet delays look like in practice?

[~]$ ping google.ch
PING google.ch (172.217.16.131): 56 data bytes
64 bytes from 172.217.16.131: icmp_seq=0 ttl=56 time=6.428 ms
64 bytes from 172.217.16.131: icmp_seq=1 ttl=56 time=1.215 ms
64 bytes from 172.217.16.131: icmp_seq=2 ttl=56 time=1.200 ms
64 bytes from 172.217.16.131: icmp_seq=3 ttl=56 time=1.180 ms
64 bytes from 172.217.16.131: icmp_seq=4 ttl=56 time=1.195 ms
64 bytes from 172.217.16.131: icmp_seq=5 ttl=56 time=1.144 ms
64 bytes from 172.217.16.131: icmp_seq=6 ttl=56 time=1.156 ms
64 bytes from 172.217.16.131: icmp_seq=7 ttl=56 time=1.096 ms
64 bytes from 172.217.16.131: icmp_seq=8 ttl=56 time=1.175 ms
64 bytes from 172.217.16.131: icmp_seq=9 ttl=56 time=1.181 ms
64 bytes from 172.217.16.131: icmp_seq=10 ttl=56 time=1.138 ms
64 bytes from 172.217.16.131: icmp_seq=11 ttl=56 time=1.130 ms
64 bytes from 172.217.16.131: icmp_seq=12 ttl=56 time=1.227 ms
64 bytes from 172.217.16.131: icmp_seq=13 ttl=56 time=1.162 ms
^C
--- google.ch ping statistics ---
14 packets transmitted, 14 packets received, 0.0% packet loss
round-trip min/avg/max/stddev = 1.096/1.545/6.428/1.355 ms
we describe QUIC's impact on Google Search and YouTube. These Amazon estimates that every 100 ms increase in latency cuts profits well under 20% of Search Latency and Video Latency. An almost whether a transport change is working correctly, but application and
The impact of QUIC's improvements on different application metrics
9
top achieve a 0-RTT handshake, which is at least a 2-RTT latency
the 0-RTT handshake: about 88% of QUIC connections from desk-
of all connections have a minimum RTT larger than 300ms. Of the
all connections have a minimum RTT larger than 150ms, and 10%
that users with high RTTs are in a significant tail: more than 20% of
the user's RTT increases, the impact of saving handshake round trips
QUIC
in Figure 9 which shows desktop latency reduction for users in
Search Latency. The percentile data shows that QUIC
[dynamically-generated payloads.
6.3 Search Latency
represents a low-load latency-sensitive, performed by a user results in a total response load of 100 KB
delivered to the client by Google Search, including all corresponding
search term and when all the search-result content is generated and
Recall that Search Latency is the delay between when a user enters a
6
24
24
36
190
[The QUIC Transport Protocol: Design and Internet-Scale Deployment, Langley et. al, SIGCOMM 2017]
We describe QUIC’s impact on Google Search and YouTube. These Amazon estimates that every 100 ms increase in latency cuts profits well under 20% of Search Latency and Video Latency. An almost whether a transport change is working correctly, but application and The impact of QUIC’s improvements on different application metrics trends are similar. 

9 top achieve a 0-RTT handshake, which is at least a 2-RTT latency handshake improvements, most of the latency reduction comes from 
of all connections have a minimum RTT larger than 300ms. Of the 
of all connections have a minimum RTT larger than 150ms, and 10% 
that users with high RTTs are in a significant tail: more than 20% of 
is higher, leading to larger gains in QUIC 
the user’s RTT increases, the impact of saving handshake round trips comes primarily from reducing handshake latency, as demonstrated 
ments increase as base Search Latency increases. This improvement 
Search Latency. The percentile data shows that QUIC 
dynamically-generated payloads. 

Search Latency represents delivery latency for small, delay-sensitive, 
images and embedded content. On average, an individual search 
delivered to the client by Google Search, including all corresponding 
search term and when all the search-result content is generated and 
Recall that Search Latency is the delay between when a user enters a 
6.3 Search Latency 
represents a heavy-load bandwidth-sensitive application. 
represents a low-load latency-sensitive application, and YouTube 
impact. Second, they represent diverse transport use-cases: Search 
formance for these highly optimized applications has direct revenue 

We chose these applications for two reasons. First, improving per-
applications. For instance, handshake latency contributes to 
impact of networking changes on applications. 
user-relevant metrics are a measure of the usefulness of the change. 

First, networking remains just one constituent of end-to-end ap-
Microbenchmarks such as the ones above are a useful measure of 

For the sake of brevity we show only desktop data for these supporting graphs. Mobile 
As shown in Table 1, users in QUIC 
experienced reduced mean 

Table 1: Percent reduction in global Video Rebuffer Rate for users in 
and mobile environments and usage, the lower gains are explained 
and mobile environments and usage, the lower gains are explained 
and mobile environments and usage, the lower gains are explained 

QUIC includes richer signaling than TCP, which enables QUIC 
a role in decreasing Search latency at higher RTTs. Recall that 
tions to our servers. These results were gathered from video playbacks. 

Figure 8: Distribution of connection minimum RTTs for TCP connec-
Figure 8 shows the relationship between TCP 
Figure 8: Distribution of connection minimum RTTs for TCP connec-
Figure 8: Distribution of connection minimum RTTs for TCP connec-
Figure 8: Distribution of connection minimum RTTs for TCP connec-

Table 2: Percent reduction in global Video Rebuffer Rate for users in 
Table 2: Percent reduction in global Video Rebuffer Rate for users in 
Table 2: Percent reduction in global Video Rebuffer Rate for users in 
Table 2: Percent reduction in global Video Rebuffer Rate for users in 

36% of connections over 100ms 

[The QUIC Transport Protocol: Design and Internet-Scale Deployment, Langley et. al, SIGCOMM 2017]
Caveat: There are some obvious geolocation errors
A network connection is characterized by its delay, loss rate, and throughput.
In practice, queues are not infinite. There is an upper bound on queuing delay.

queue

N+1 packets

packet size \( L \)  
transmission rate: \( R \)

queuing delay upper bound: \( N \times \frac{L}{R} \)
If the queue is persistently overloaded, it will eventually drop packets (loss)
Why does loss matter?

Sometimes, it doesn’t!
e.g., losing a negligible part of conversation

Sometimes, translates into service delays
“This packet didn’t arrive? I’ll try again.”

Can also send redundant information in advance
A, B, A ⊕ B — incurs a capacity penalty
What does loss on the Internet look like?

[The QUIC Transport Protocol: Design and Internet-Scale Deployment, Langley et. al, SIGCOMM 2017]
A network connection is characterized by its delay, loss rate, and throughput.
The throughput is the rate at which a host receives data

Average throughput = \( \frac{\text{data size}}{\text{transfer time}} \)

[#bits/sec] = [#bits] / [sec]
To compute throughput, one has to consider the bottleneck link

Average throughput $= \min(R_S, R_L)$

= transmission rate of the bottleneck link
To compute throughput, one has to consider the bottleneck link and the intervening traffic.

If $4 \times \min(R_S, R_L) > R$, the bottleneck is now in the core, providing each download $R/4$ of throughput.
Why does throughput matter?

**Some applications need little bandwidth**

* e.g., gaming typically needs few kbps

**Some need a lot more**

* e.g., HD video streaming often needs 5+ Mbps

(c) copyright 2008, Blender Foundation / www.bigbuckbunny.org, CC-BY-3.0
What does throughput on the Internet look like?
What does throughput on the Internet look like?

[Map showing average Mbps across different regions, with the title: Akamai's State of the Internet, 2017]
What does throughput on the Internet look like?

[Map showing the percentage of users with 10+ Mbps Internet speed worldwide, based on Akamai’s State of the Internet, 2017]
A network *connection* is characterized by its delay, loss rate, and throughput.
As technology improves, throughput increases & delay decreases, except for propagation

(speed of light)

Data in fiber moves at $\approx \frac{2c}{3}$

Significant room for improvement here!
Because of propagation delays, Content Delivery Networks move content closer to you

http://wwwnui.akamai.com/gnet/globe/index.html
Readings for this week

Kurose-Ross (KR) sections

1.5 Protocol layers and their service models

END-TO-END ARGUMENTS IN SYSTEM DESIGN

J.H. Saltzer, D.P. Reed and D.D. Clark*

M.I.T. Laboratory for Computer Science

IEEE ICDCS, 1981