GAMES NETWORKING & SECURITY

WEEK 3

LATENCY & DATA TRANSIT DELAY

Jack Ingram & Stephen Selwood

LEARNING OUTCOMES

To understand what latency means for a video game.

To understand the factors that the user can control.

To introduce the factors that are out of the control of the user/player and developer.

To understand the different root causes of latency on the network.

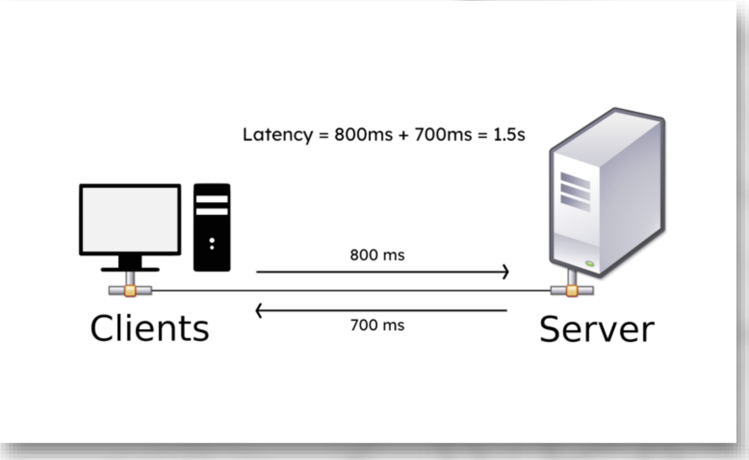
To appreciate that these root causes affect all network traffic, not just game related traffic.

Latency & Data Communication

Latency

Latency refers to the time it takes for a packet of data to be transported from its source to its destination.

All network traffic is affected by latency and there are a number of factors that can contribute to latency.



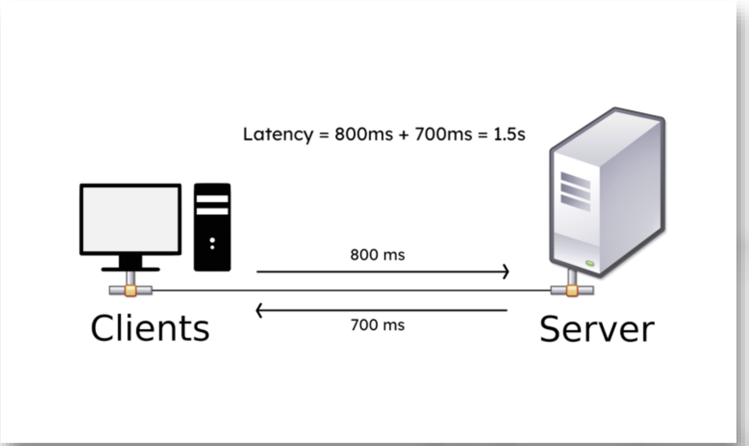
Latency & Data Communication

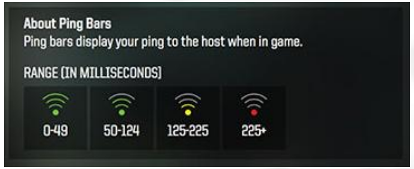
Latency | Round Trip Time

In gaming we often care about the latency of the source-to-destination trip and then the back-to-source trip. This is known as the Round Trip Time (RTT).

In many cases the RTT is twice the latency, although this is not universally true. Some network paths exhibit asymmetric latencies.

In gaming, the RTT from client-to-server is commonly known as *'ping'*.




About Ping Bars
Ping bars display your ping to the host when in game.

RANGE (IN MILLISECONDS)

0-49 50-124 125-225 225+

Dust II - Competitive

PING PLAYER
3 Moving.. W
BOT Kevin

Ping ms 48

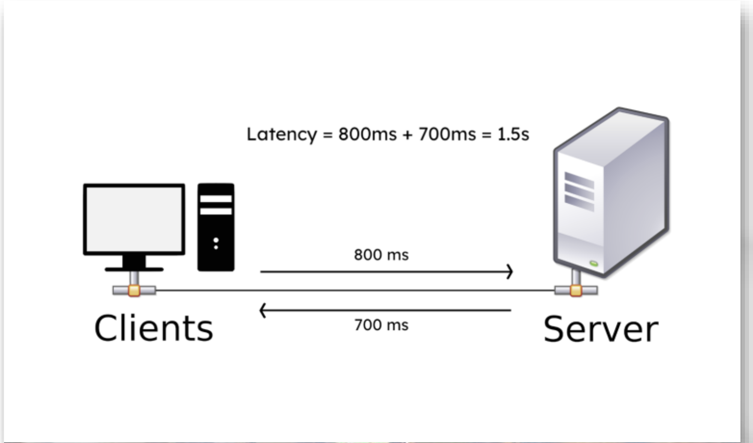
Ping represented in different ways in the UI of *Call of Duty - Black Ops III, Counter-Strike: Global Offensive & speedtest.net*

Latency & Data Communication

Latency | Round Trip Time

An acceptable level of latency for gaming is between 40 to 60 milliseconds.

Anything over 100 milliseconds will often result in noticeable lag in the game and reduce the engagement.



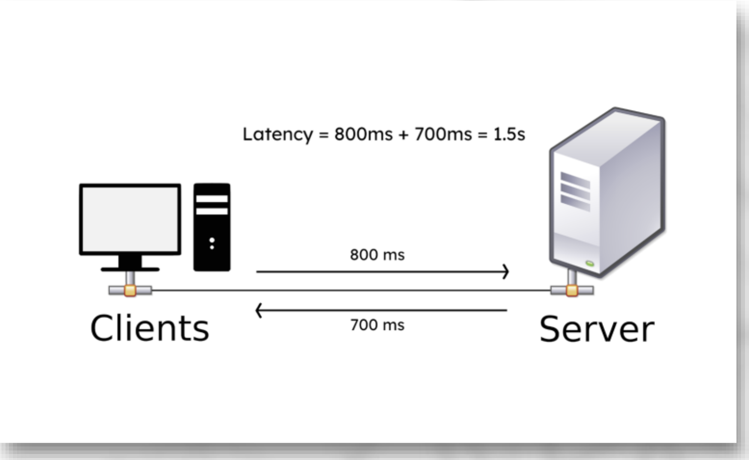
Latency & Data Communication

Latency | Common sources of latency

There are many potential sources of latency across the network.

Of them, some are common issues that most gamers are knowledgeable about and can rectify:

1. Geographical location,
2. Connection speed,
3. Equipment (router, wired or wireless connection etc…)



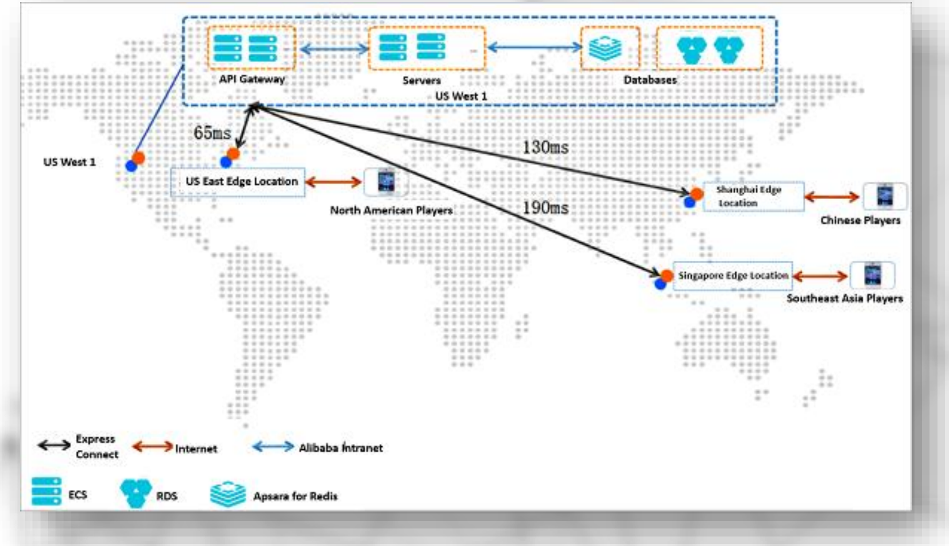
Latency & Data Communication

Latency | Common sources of latency

Geographical Location:

As discussed in the Week 1 lecture, the further the data packet has to travel, the longer the RTT. Players that are located far away from the server will find themselves with a worse connection.

As discussed in the Week 2 lecture, most networked games setup localised servers based on the player-bases' geographical location to solve this issue. Many servers with shorter distances between them are better than a few servers with longer distances.



Latency & Data Communication

Latency | Common sources of latency

Equipment:

Old network equipment (such as network card, router, modem etc.) can increase latency.

Older modems with fewer upstream and downstream channels (bandwidth) will not be able to manage the data transfer speeds required.

This will be further compounded if multiple devices are trying to access the network at the same time.

Latency & Data Communication

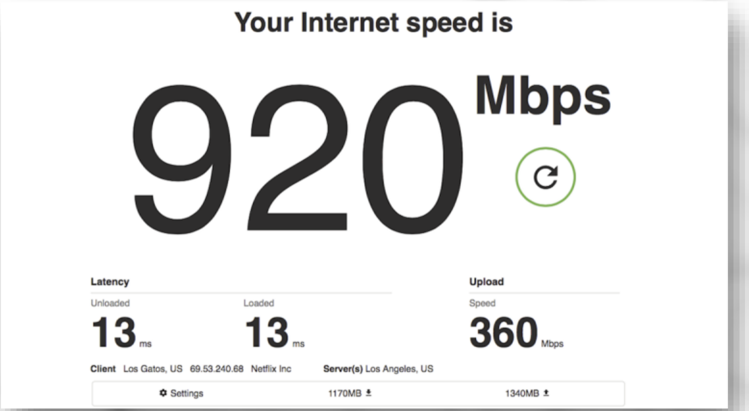
Latency | Common sources of latency

Connection Speed:

The connection speed of your contact with your ISP, and the type of *'wire'* or hard-line connection you have, that connects your location to the overall network can have a significant effect on the latency during a game.

Hardline connections that are fibre optic are much faster than the traditional copper cabling used for a broadband connection.

Slower connection speed = higher latency.
Higher the connection speed = lower latency.



Your Internet speed is

920 Mbps

Latency

Unloaded 13 ms

Loaded 13 ms

Upload

Speed 360 Mbps

Client Los Gatos, US 69.53.240.68 Netflix Inc Server(s) Los Angeles, US

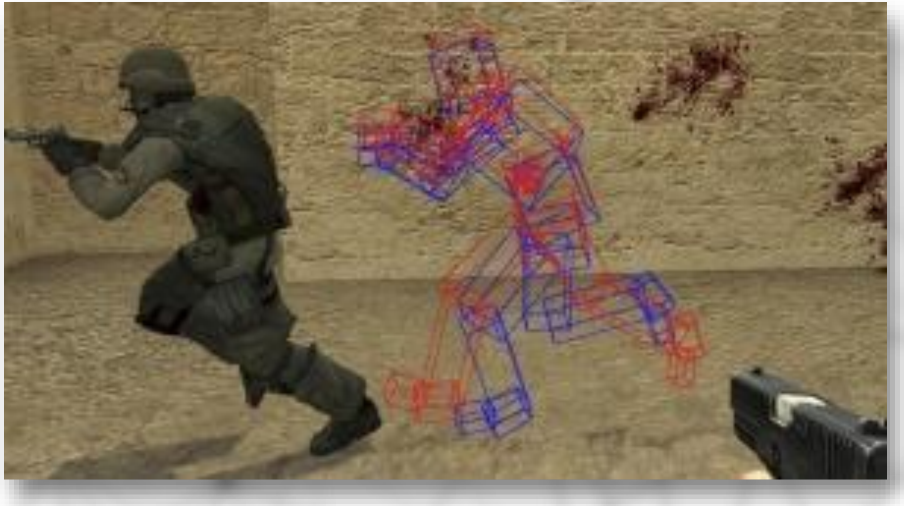
⚙ Settings 1170MB ⬇ 1340MB ⬆

Latency & Data Communication

Latency | Consequences

In games, latency affects the absolute sense of real-time interactivity that can be achieved within the game.

The latency puts a lower bound on how quickly game state information can be exchanged and consequently limits each player’s ability to react to situational changes.



08
3/3
6th

Latency & Data Communication

Latency | Latency on the network

Once on the network, the data packets may experience other forms of latency that might impact the time data is transferred from the client node and the server mode.

These other manifestations of latency come in the form of:

Jitter

Packet Loss

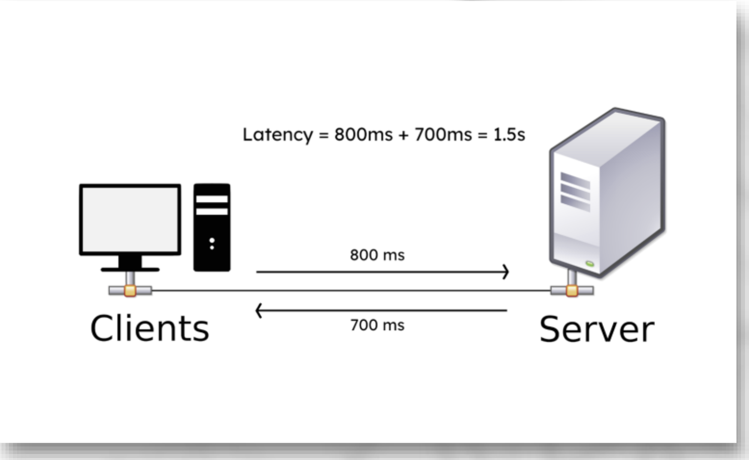
Propagation Delay

Serialisation Errors

Congestion

Prioritisation

Bit Error

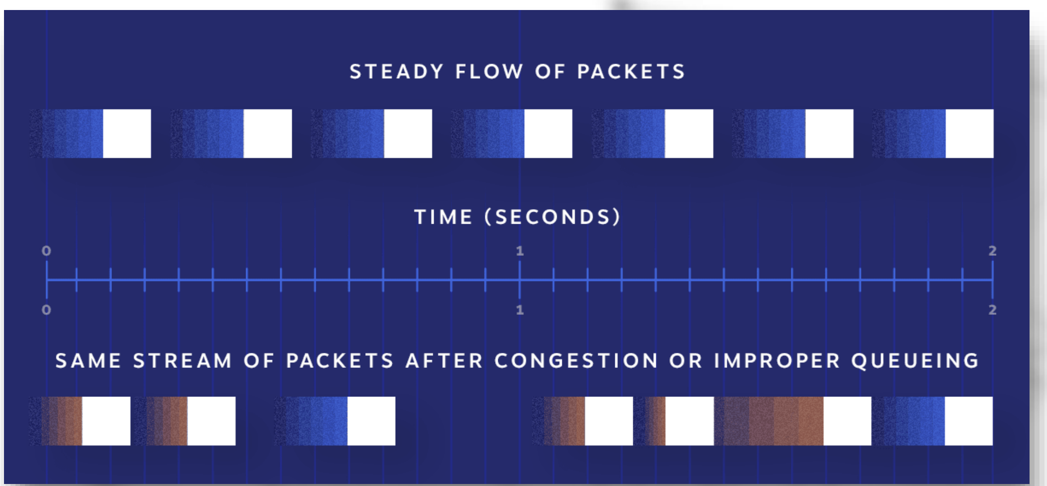


Latency & Data Communication

Sources of Latency | Jitter

Jitter is the variation in the delay times of data packets being received. When the sequence of data packets are transmitted at a regular interval, the average time delay between packet arrival will remain more-or-less constant meaning there is less jitter.

Packets sent with uneven interval transmission delays, jittery packets, can cause latency in the delivery times as they clog up buffers along the way.

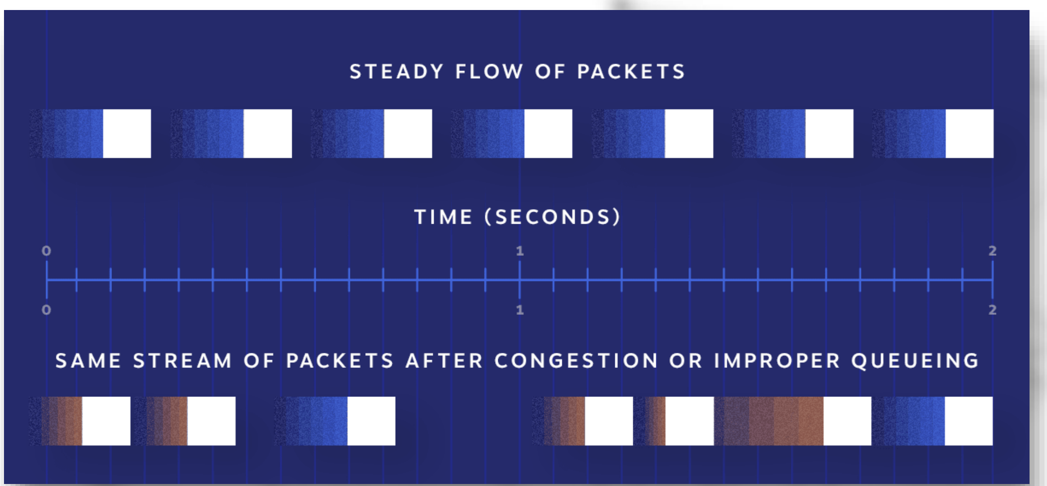

STEADY FLOW OF PACKETS
TIME (SECONDS)
0
1
2
0
1
2
SAME STREAM OF PACKETS AFTER CONGESTION OR IMPROPER QUEUEING

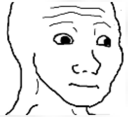
Latency & Data Communication

Sources of Latency | Jitter

It is sufficient to know that latency can fluctuate slowly or rapidly from one packet to the next. One packet can have a low arrival delay, while the next may have a larger delay.

A path that exhibited dynamic latency: 70ms, 120ms, 130ms, 80ms… and so on would be problematic to deal with and would manifest as lag to the client nodes.


STEADY FLOW OF PACKETS
TIME (SECONDS)
0
1
2
SAME STREAM OF PACKETS AFTER CONGESTION OR IMPROPER QUEUEING



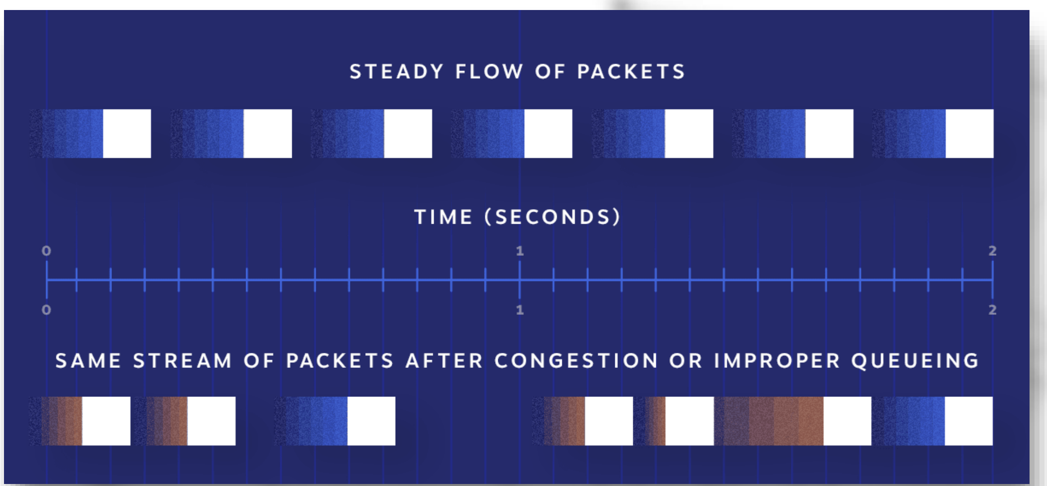
TFW lag to the client nodes

Latency & Data Communication

Sources of Latency | Jitter

Excessive jitter can make it difficult for players (and the game engine) to compensate for long-term average latency from the network.

Jitter must be kept as low as possible.

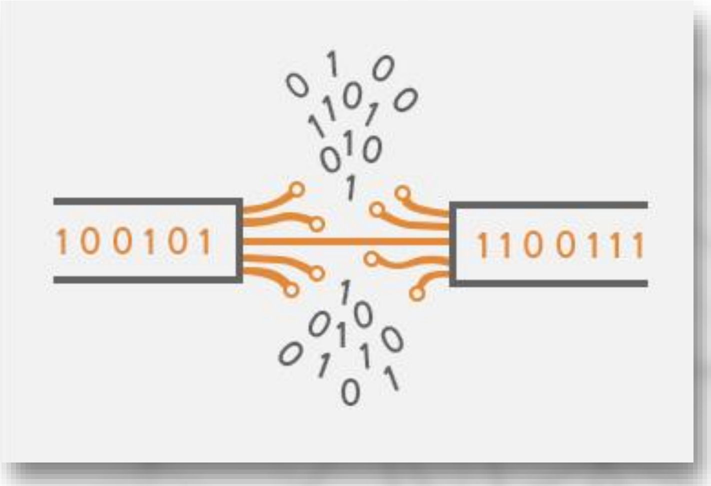

STEADY FLOW OF PACKETS
TIME (SECONDS)
0
1
2
0
1
2
SAME STREAM OF PACKETS AFTER CONGESTION OR IMPROPER QUEUEING

Latency & Data Communication

Sources of Latency | Packet Loss

Packet loss refers to the case when a packet simply never reaches its destination. It is lost somewhere in the network.

An example of loss:
When a packet arrives at a network device, it must be stored in memory prior to being moved along toward its destination. If the device has run out of buffer memory, the packet must be discarded.

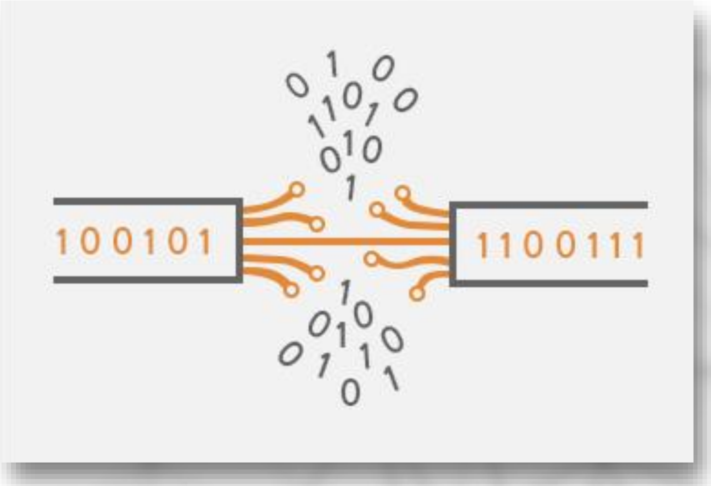


Latency & Data Communication

Sources of Latency | Packet Loss

A path's packet loss characteristic is often described in terms of packet loss rate or packet loss probability (ratio of the number of packets lost per number of packets sent).

In games, the issues of packet loss should be self-evident – game state updates are lost, and the game engine (at client and/or server) must cover up the loss as best it can.

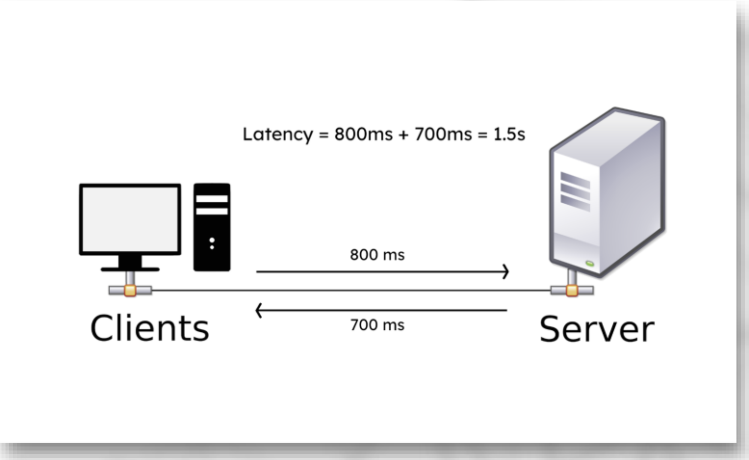


Latency & Data Communication

Sources of Latency

Three main sources of delay add cumulatively to the total latency experienced by a packet:

- Finite propagation delays over large distances
- Serialisation delays
- Congestion-related queuing delays



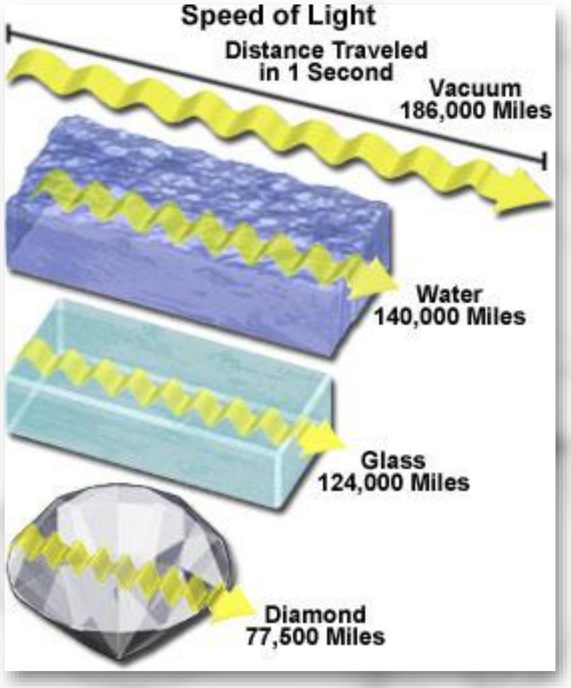
Latency & Data Communication

Sources of Latency | Propagation Delay

The speed of light dictates the top speed with which any information can propagate in a particular medium.

The speed of transmission in fibre optic cables it is about 31% slower. The speed of transmission along wires and cables is usually less than in air due to the physical resistance of the cables.

Since the speed of light is finite, the laws of physics impose a limit on the on latency between geographically distant points. This is propagation delay.



Latency & Data Communication

Sources of Latency | Propagation Delay

The speed of light is 299.8 million meters per second (670.7 million mph or 1.08 billion kph).

Propagation delays become noticeable over links spanning 1000s of km, or where the path hops through a number of routers each thousands of km apart.

For example, a 18,800km path from *London* to *Wellington, NZ* would exhibit at least 63ms latency (or 126ms RTT) simply because of the finite speed of light.

Iceland
Sweden
Norway
Russia
Un King
Poland
Germany
Ukraine
Kazakhstan
France
Italy
Spain
Mongolia
Turkey
China
South Korea
Japan
Iraq
Iran
Afghanistan
Pakistan
Algeria
Libya
Egypt
Saudi Arabia
India
Thailand
Mali
Niger
Chad
Sudan
Nigeria
Ethiopia
Kenya
DRC
Tanzania
Indonesia
Papua New Guinea
Angola
Namibia
Botswana
Madagascar
Indian Ocean
South Atlantic Ocean
Australia
South Africa
N Zeal
No Pac Oc

Latency & Data Communication

Sources of Latency | Propagation Delay

Most game players will come across this issue when they are connected to servers in different countries.

It is also possible to find a high latency network between two geographically close sites if they use different ISPs.

A rough rule of thumb for propagation delay is:

latency (ms) = (distance of link in km) / 300

Iceland
Sweden
Norway
Russia
Un
Kingd
Poland
Germany
Ukraine
Kazakhstan
Mongolia
France
Italy
Spain
Turkey
China
South Korea
Japan
No
Pac
Oc
Iraq
Iran
Afghanistan
Pakistan
Algeria
Libya
Egypt
Saudi Arabia
India
Thailand
Mali
Niger
Chad
Sudan
Nigeria
Ethiopia
Kenya
DRC
Tanzania
Indonesia
Papua New Guinea
Angola
Namibia
Botswana
Madagascar
Indian Ocean
South Atlantic Ocean
Australia
South Africa
Ne
Zeal

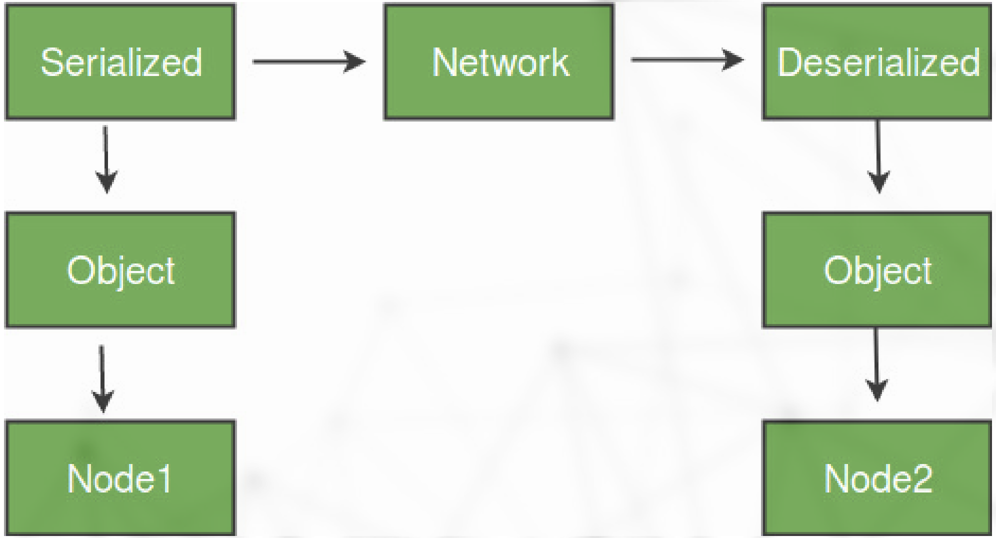
Latency & Data Communication

Sources of Latency | Serialisation

Serialisation occurs in many real-life situations. For example, crowds of people getting on a bus go through the door one at a time.

Serialisation is the time it takes for a unit of data to be serialised for transmission on a narrow, or serial, channel such as a cable.

Serialisation occurs on most link layers, and is another source of latency on networks.



Latency & Data Communication

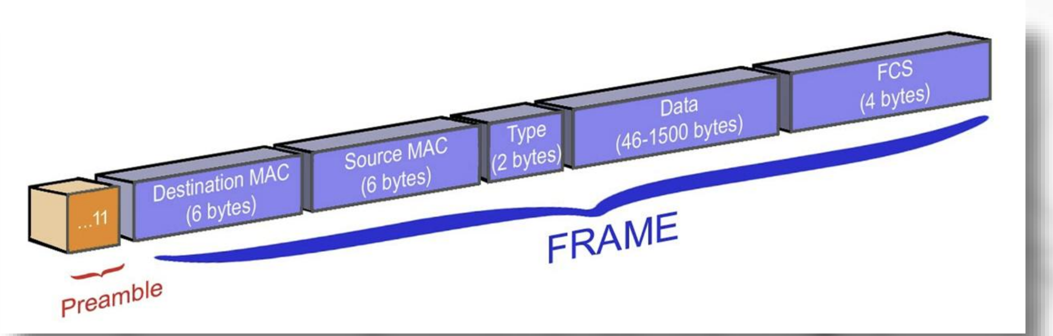
Sources of Latency | Serialisation

A Frame in networking is a digital data transmission unit.

Frames are broken into sequences of bytes, and these bytes are sent one bit at a time.

The finite period taken to transmit an IP packet one bit at a time is referred to as serialisation latency.

This period of time depends on the speed of the link (in bits per second) and the length of the packet being sent.



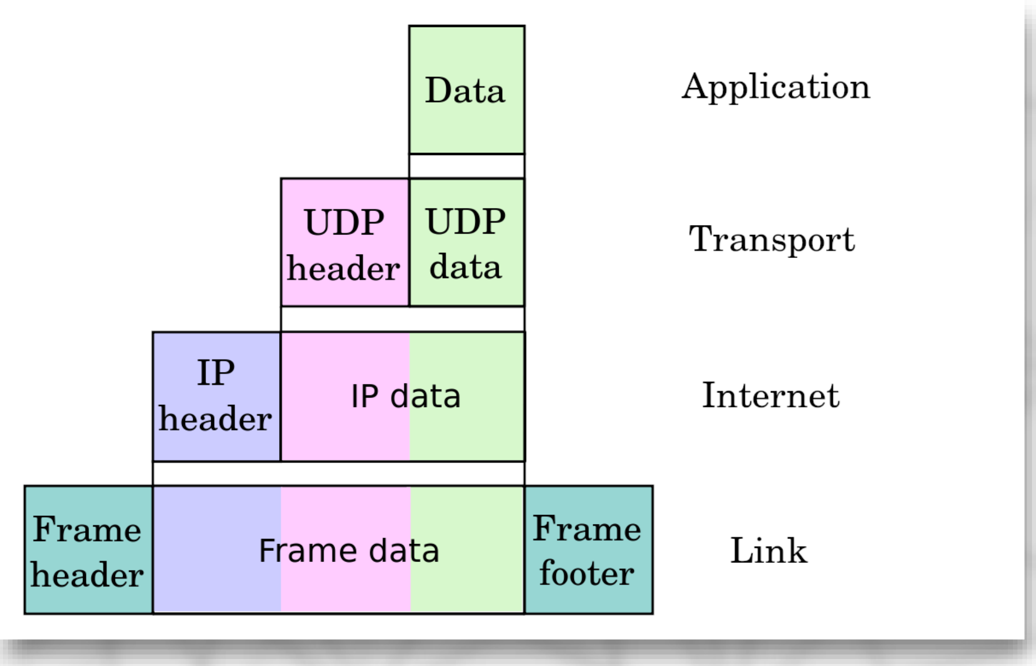
Latency & Data Communication

Sources of Latency | Serialisation

Depending on the link layer technology, there might be extra bits at the beginning and end of each byte or frame.

Thus, the total serialisation latency also depends on the framing protocol used by a particular link layer.

Consider the time taken to transmit a 1500-byte packet on a 100-Mbps Ethernet and a 56-Kbps V.90 dial-up connection.



Latency & Data Communication

Sources of Latency | Serialisation

A 1500 byte packet becomes 1526 bytes long:

8 bytes of ethernet preamble

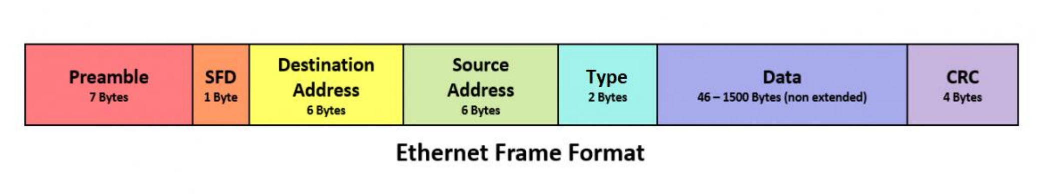
12 bytes for source and destination MAC address

2 bytes for ethernet protocol type

4 bytes trailing Cyclic Redundancy Check (CRC)

This gives us 12,208 bits. At 100Mbps it takes: 12,208 / 100,000,000 = 122µs to transmit the frame containing this packet.

There are 1000 microseconds in 1 millisecond.



Preamble	SFD	Destination Address	Source Address	Type	Data	CRC
7 Bytes	1 Byte	6 Bytes	6 Bytes	2 Bytes	46 – 1500 Bytes (non extended)	4 Bytes

Ethernet Frame Format

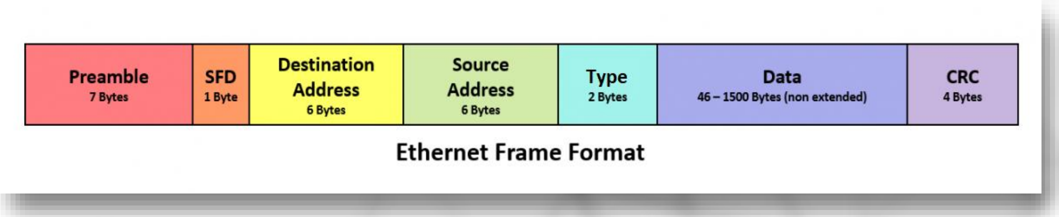
Latency & Data Communication

Sources of Latency | Serialisation

Serialisation latency is primarily an issue with low-speed links.

A similar situation occurs on high-speed links when your ISP imposes temporary rate caps, for example if the download limit has been reached for the month or if the connection is throttled because of suspicious levels of activity.

This is done by limiting the rate of serialisation.



Preamble 7 Bytes	SFD 1 Byte	Destination Address 6 Bytes	Source Address 6 Bytes	Type 2 Bytes	Data 46 – 1500 Bytes (non extended)	CRC 4 Bytes

Ethernet Frame Format

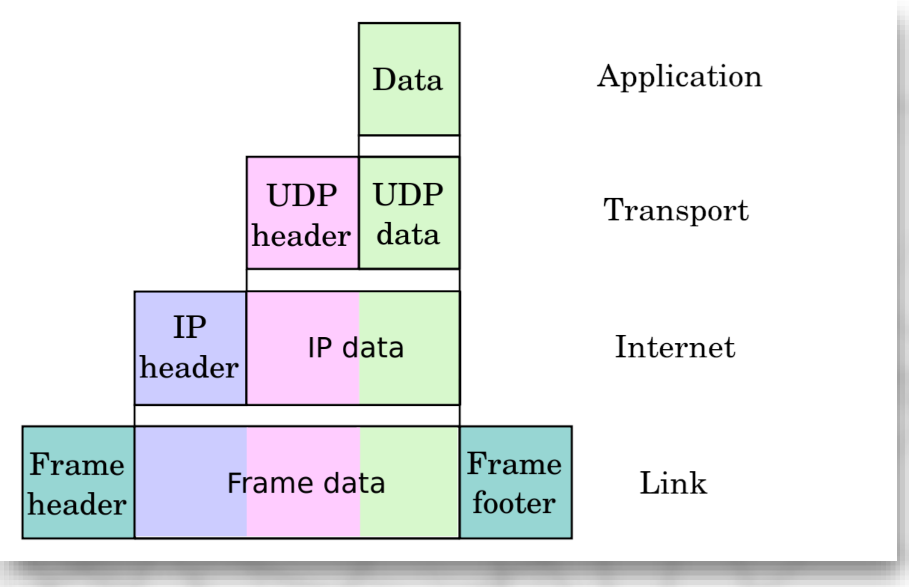
Latency & Data Communication

Sources of Latency | Serialisation

Serialisation latency should only be calculated once since the receiving end is pulling bits off the link at the same rate that the transmitting end is sending them.

Aside from a slight offset in time due to propagation delay, transmission and reception occurs concurrently.

A rough rule of thumb for serialisation delay is:



$$\text{latency (ms)} = \frac{8 \times (\text{link layer frame length in bytes})}{}$$

latency (ms) = 8 x (link layer frame length in bytes)

Break.

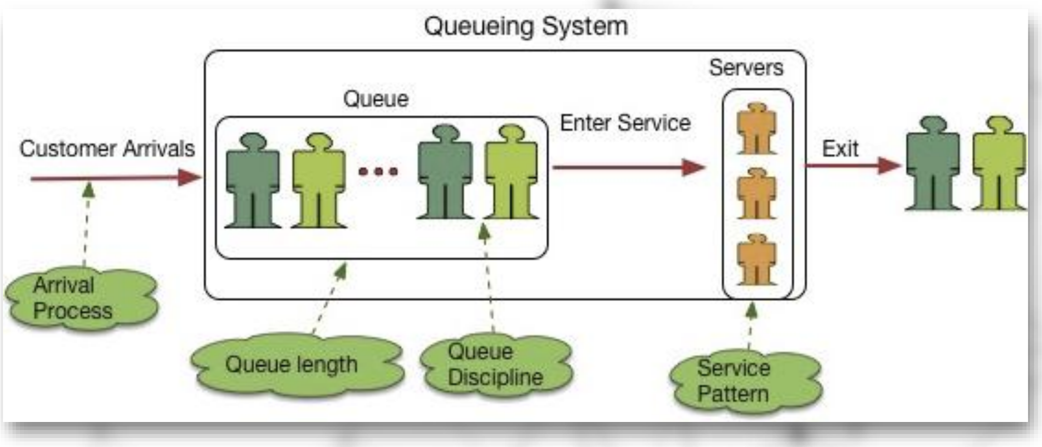
Break.

Latency & Data Communication

Sources of Latency | Queueing Delays

One of the core underlying assumptions of the Internet's 'best effort' philosophy is that everyone's traffic is largely uncorrelated, allowing us to benefit from statistical multiplexing.

Multiplexing occurs when multiple inbound streams of traffic converge on a single outbound link at a particular router or switch. The inbound packets are multiplexed onto the outbound link.



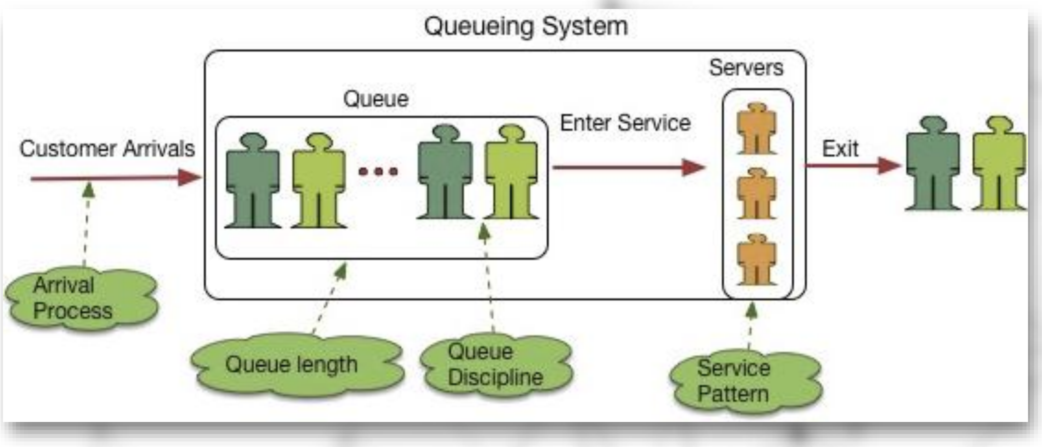
Latency & Data Communication

Sources of Latency | Queueing Delays

However, IP routers do not prearrange guaranteed time slots on the outbound link for the competing inbound packet streams.

Statistical multiplexing assumes that everything will be okay if the average bitrate of all the inbound packet streams does not exceed the capacity of the outbound link.

Most of the time most packets do not simultaneously and therefore will not collide with each other.



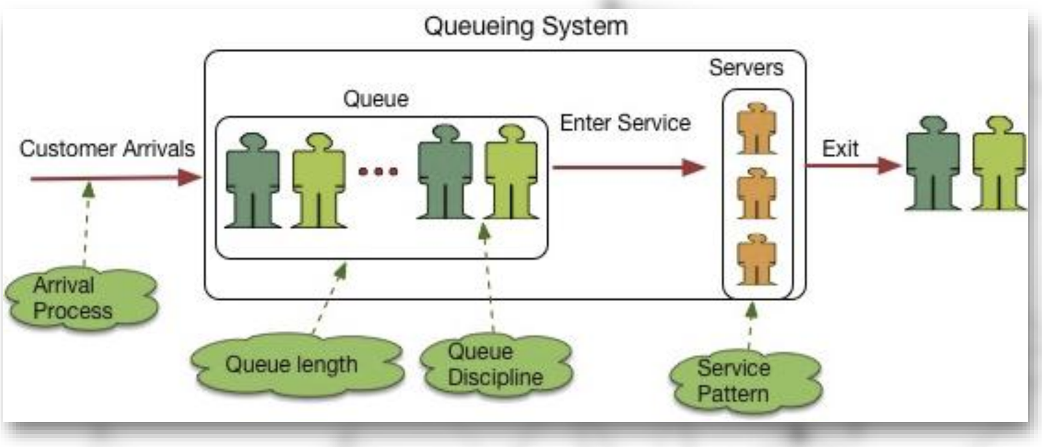
Latency & Data Communication

Sources of Latency | Queueing Delays

Of course, in reality, packet arrivals do *‘collide’*.

When multiple inbound packets arrive at the same instant for the same outbound link, the packets are queued up and transmitted one after the other.

This is known as *Transient Congestion*.

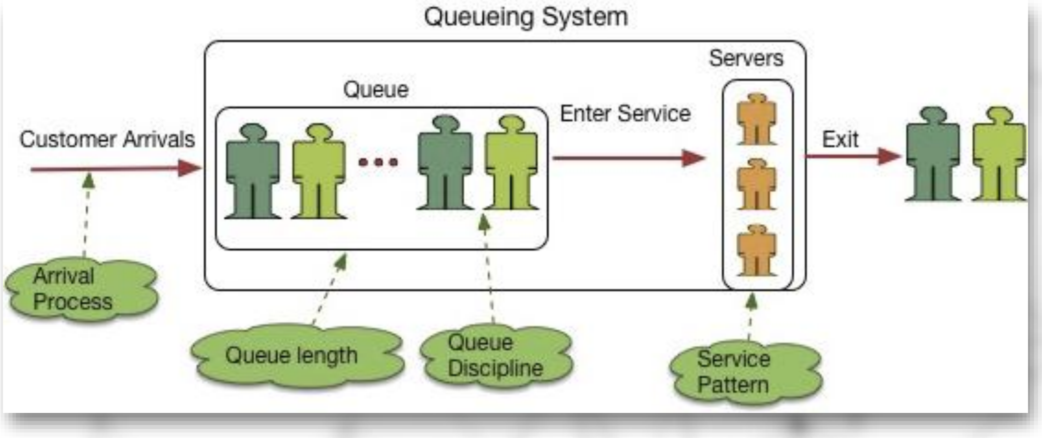


Sources of Latency | Queueing Delays

As we already know, transmitting a single packet on a physical link introduces a finite serialisation delay.

Consequently, any packet queued up for transmission on a particular link will experience additional latency due to the serialisation delays of every packet in the queue ahead of it.

We refer to this as queueing delay.



Latency & Data Communication

Sources of Latency | Queueing Delays

This is analogous to petrol stations:

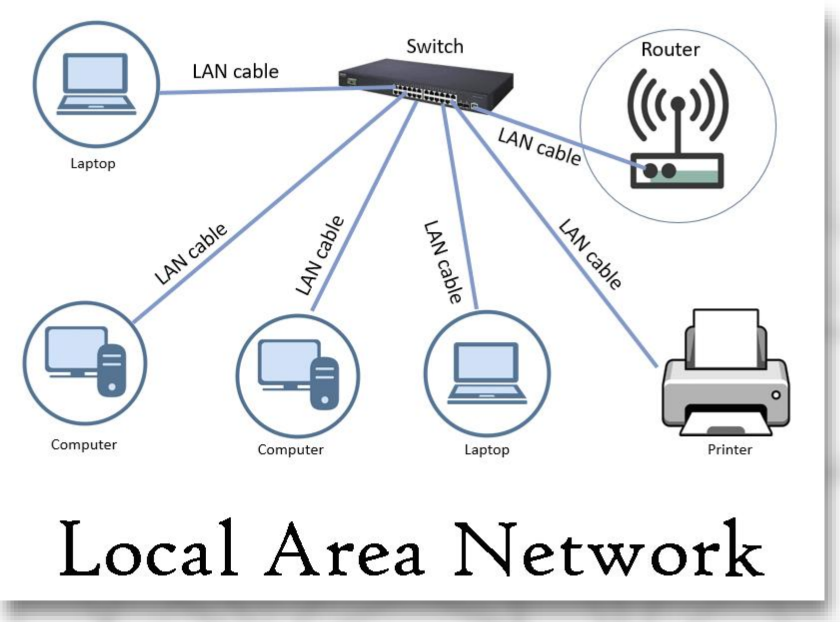
Sometimes there will be no cars on the forecourt and at other times cars will be queuing to get into the petrol stations!

Latency & Data Communication

Sources of Latency | Queueing Delays

In a typical consumer environment, queueing delays are seen when multiple computers on a home LAN try to send packets out through the same router.

When outbound packets converge on the router they will be queued up, waiting their turn to be transmitted on the upstream link.

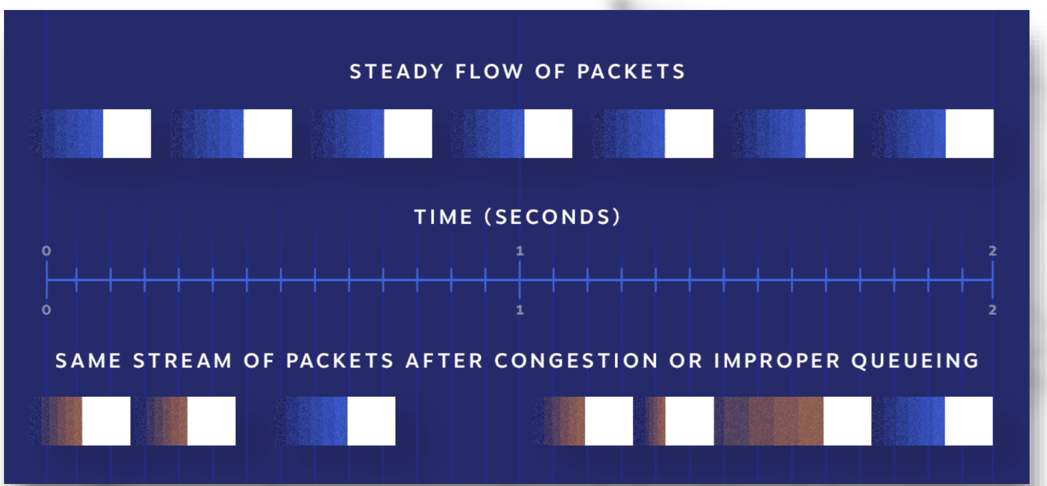


Latency & Data Communication

Sources of Jitter

A number of mechanisms introduce jitter by causing variations in latency from one packet to the next:

- Route length changes
- Packet size variations
- Transient congestion


STEADY FLOW OF PACKETS
TIME (SECONDS)
0
1
2
0
1
2
SAME STREAM OF PACKETS AFTER CONGESTION OR IMPROPER QUEUEING

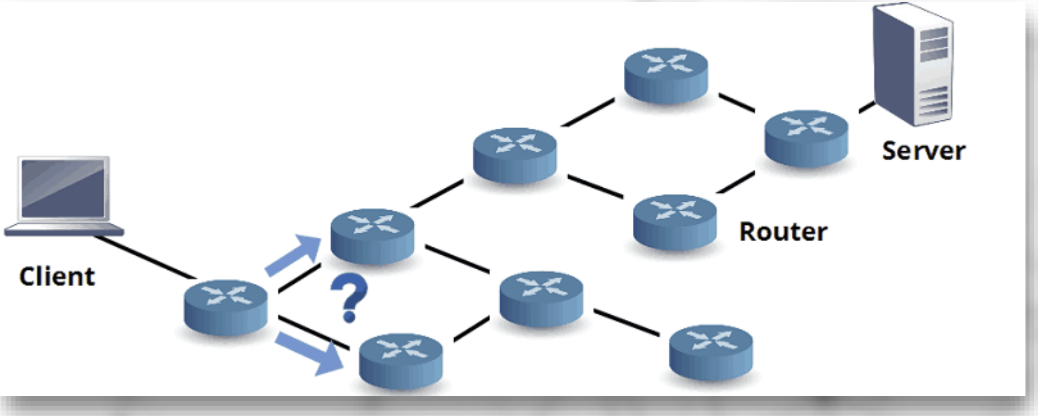
Latency & Data Communication

Sources of Jitter | Route length changes

The actual path taken by a stream of packets can vary over time.

When a route change occurs, the new path may be shorter or longer (in both km and number of hops).

Packets sent immediately after the route change will still get to their destination and yet experience a different latency.

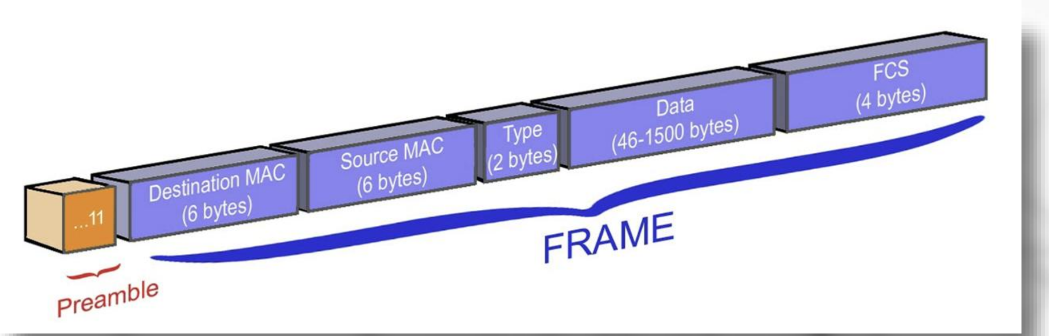


Latency & Data Communication

Sources of Jitter | Packet Size Variation

On links that introduce noticeable serialisation delay, we can experience jitter due simply to the variations in size between consecutive packets sent over the link.

Congestion-induced queueing delay can also be affected by packet sizes. If the size of the packets ahead in the queue vary, then so will the time taken to get to the front of the queue



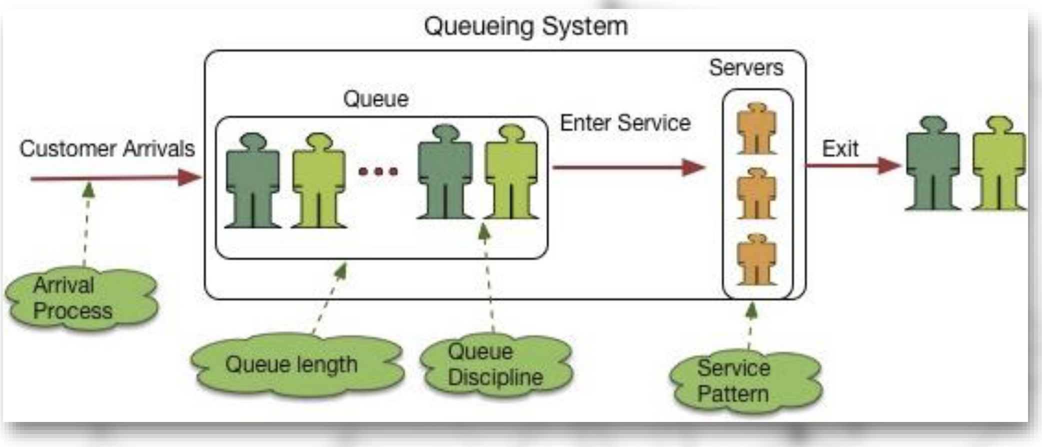
Latency & Data Communication

Sources of Jitter | Transient Congestion

Consider the following case of two home computers on a single 100-Mbps Ethernet, communicating to the outside world over a 128 Kbps link.

Host 1 is generating a stream of 80 byte IP packets, one every 40ms. Assume link layer overheads add a fixed 10 bytes, making the frame 90 bytes long.

These frames take 5.6ms to transmit at 128Kbps.

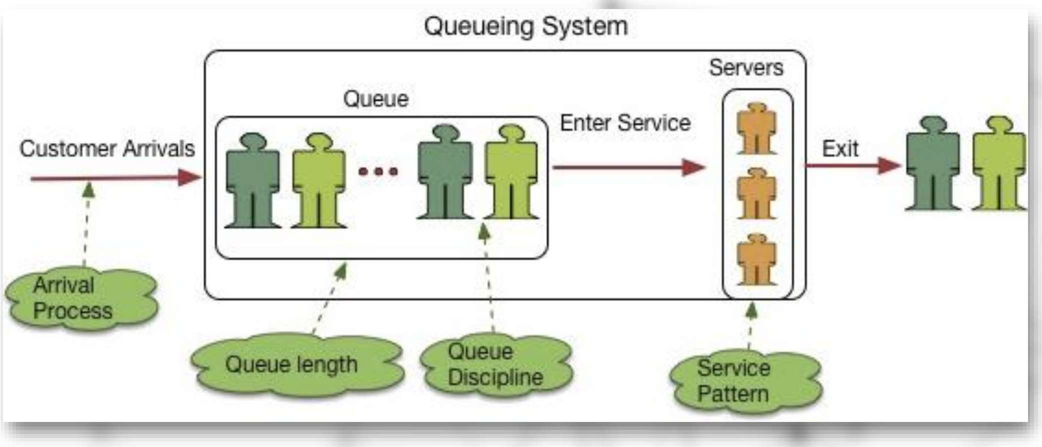


Latency & Data Communication

Sources of Jitter | Transient Congestion

Now host 2 suddenly decides to transmit a random stream of 1500 byte IP packets, with a mean interval of 500ms.

These larger packets arrive at the router and take 94.4ms to transmit.

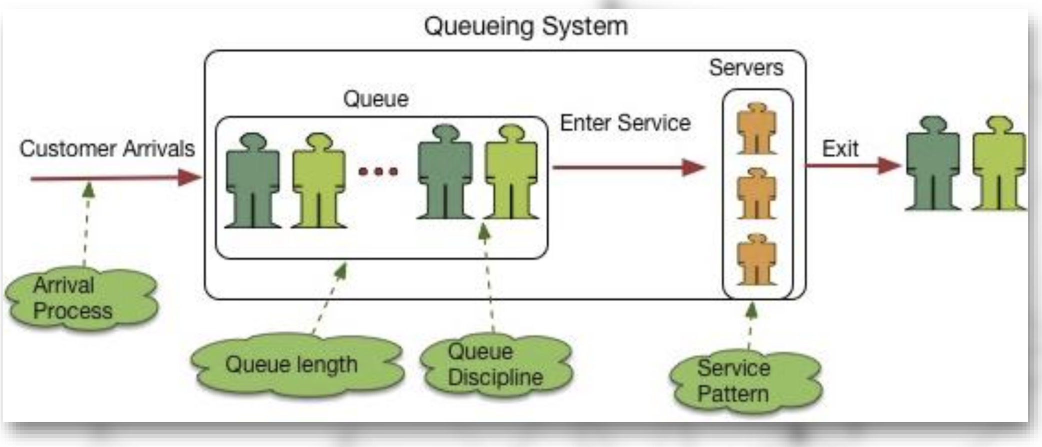


Latency & Data Communication

Sources of Jitter | Transient Congestion

From host 1's perspective, its stream of 80 byte IP packets now experience random jitter.

Much of the time the packets go through immediately, but every so often packets are delayed by up to 94.4ms while the 1500 byte packet is transmitted.

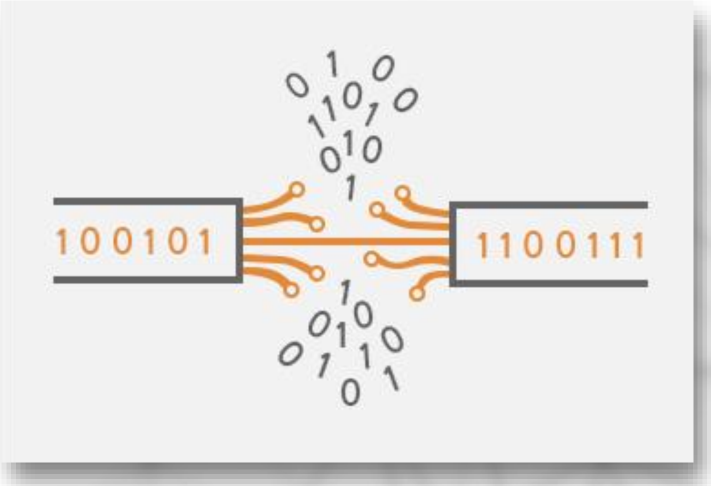


Latency & Data Communication

Sources of Packet Loss

Packet losses are typically due to:

1. Link layer bit errors causing packet corruption
2. Excess transient congestion causing queues to overflow
3. Routing transients temporarily disrupting the path



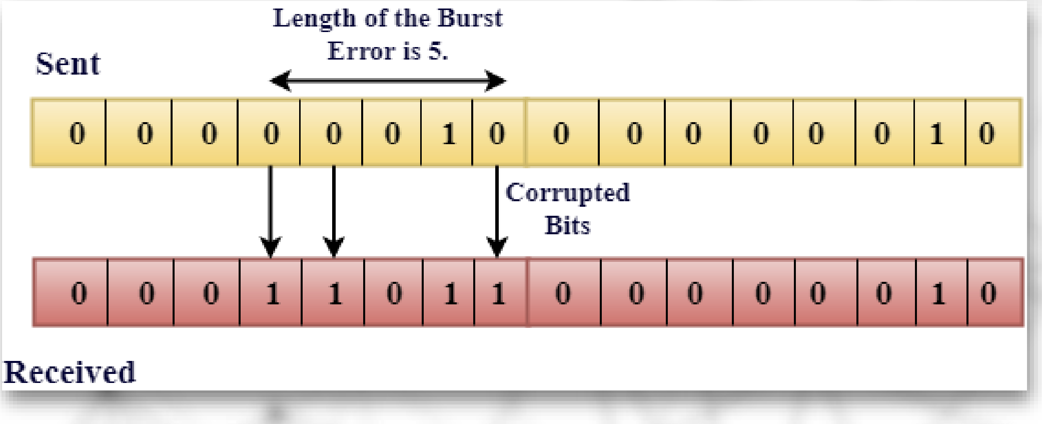
Latency & Data Communication

Sources of Packet Loss | Bit Error

At the physical layer all links experience a finite (usually very low) rate of data corruption – which we refer to as bit errors and characterise by a link’s bit error rate.

Bit errors may be introduced by poor signal-to-noise ratios in the digital-analogue-digital conversion process, resulting in erroneous encoding or decoding of data.

Bit errors can also be caused by electrical glitches in hardware.



0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0

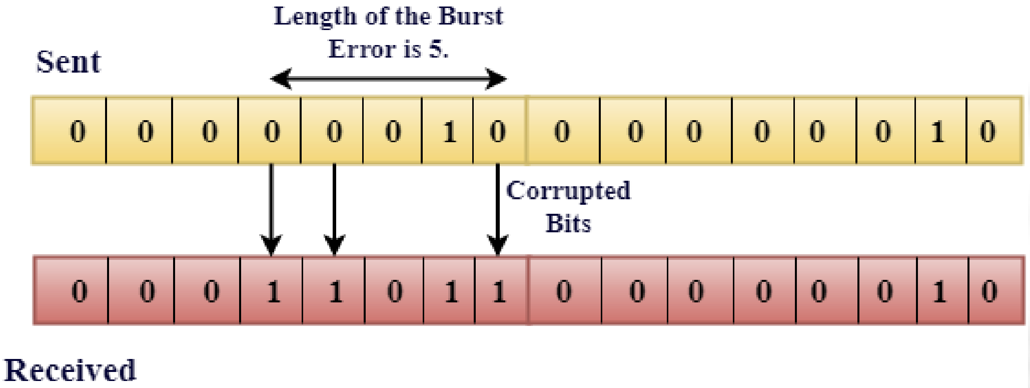
0 0 0 1 1 0 1 1 0 0 0 0 0 0 1 0

Latency & Data Communication

Sources of Packet Loss | Bit Error

Some link technologies encode additional information within each frame to enable limited reconstruction of a frame after one- or two-bit errors. This is known as Forward Error Correction (FEC).

Uncorrected bit errors are usually discovered through Cyclic Redundancy Check (CRC) calculations at both the transmitting and receiving end of the link.



Sent	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0
Received	0	0	0	1	1	0	1	1	0	0	0	0	0	0	1	0

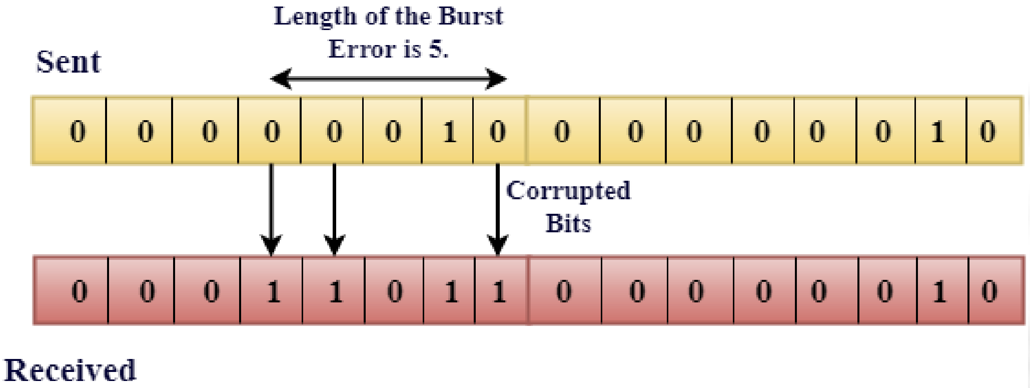
Latency & Data Communication

Sources of Packet Loss | Bit Error

The CRC is a 16-bit or 32-bit value calculated during transmission and sent along with each frame, and then recalculated at the receiver.

If the original and recalculated CRCs differ, the frame is discarded.

Since this can occur anywhere along an IP path, there is no way to inform the end hosts why or how their packet was lost.



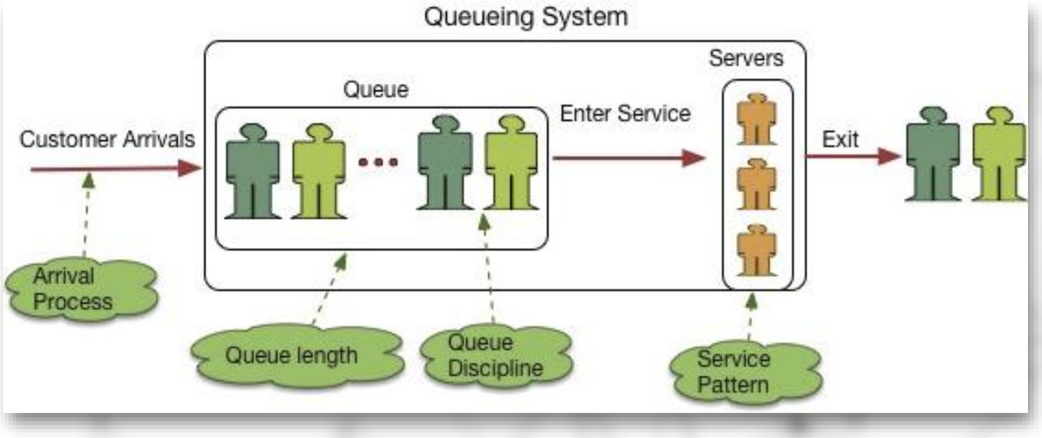
Sent	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0
Received	0	0	0	1	1	0	1	1	0	0	0	0	0	0	1	0

Sources of Packet Loss | Transient Congestion

Transient congestion can become so severe that queueing points along the path simply run out of space to hold new packets.

When this happens, new packets are simply dropped until the queue has emptied enough to take new packets.

This is the network’s most aggressive form of self protection against too much traffic.



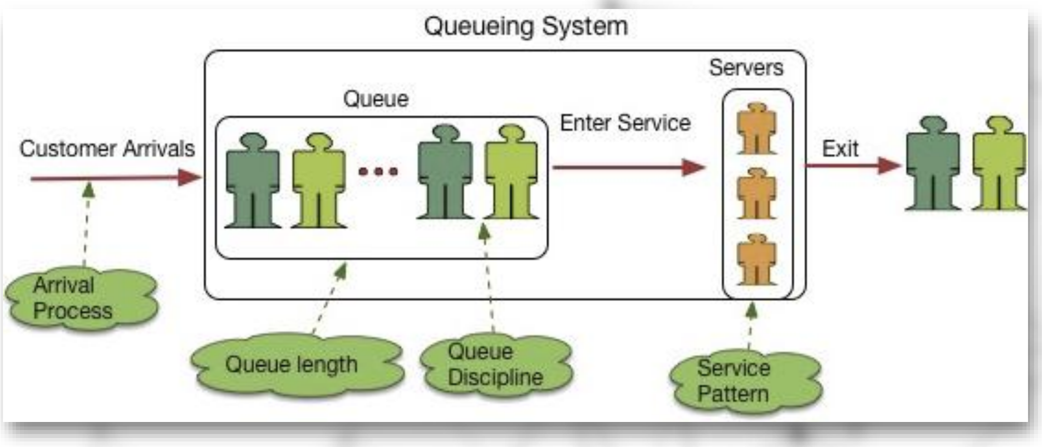
Latency & Data Communication

Sources of Packet Loss | Transient Congestion

Some networks even employ proactive packet drop mechanisms that introduce a random, non-zero loss probability well before the queue is full.

This is referred to as Active Queue Management, with the most well-known variant known as Random Early Detection (RED).

Proactive packet dropping is intended to force TCP-based applications to slow down before congestion becomes too serious.



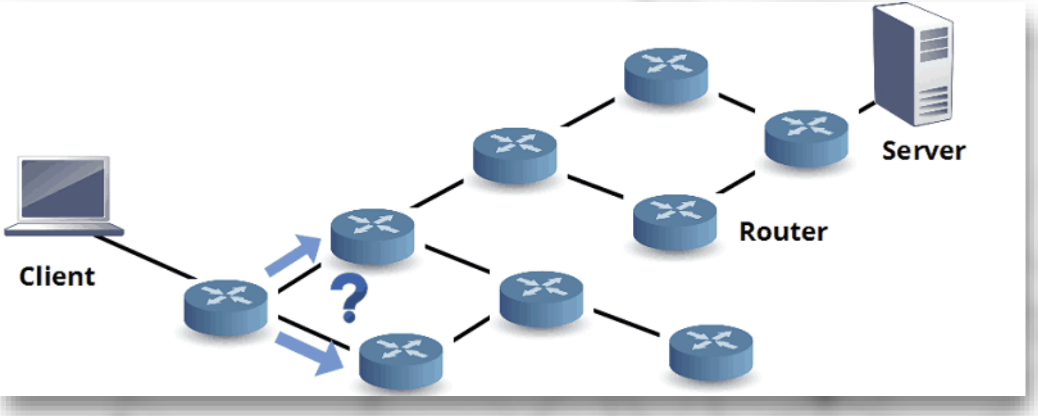
Latency & Data Communication

Sources of Packet Loss | Routing Changes

Dynamic Routing Changes do not always converge immediately on a fully functional and complete end-to-end path.

When route changes occur, there can be periods of time (from tens of seconds to minutes) where no valid shortest path exists.

This manifests itself as unexpected packet loss affecting tens, hundreds or thousands of packets.



Latency & Data Communication

Solving Transit Delays

Online games need greater control over network latency, jitter and packet loss than more traditional email, online 'chat' and web surfing applications.

There are several mechanisms that ISPs can deploy to control network conditions on behalf of gamers.

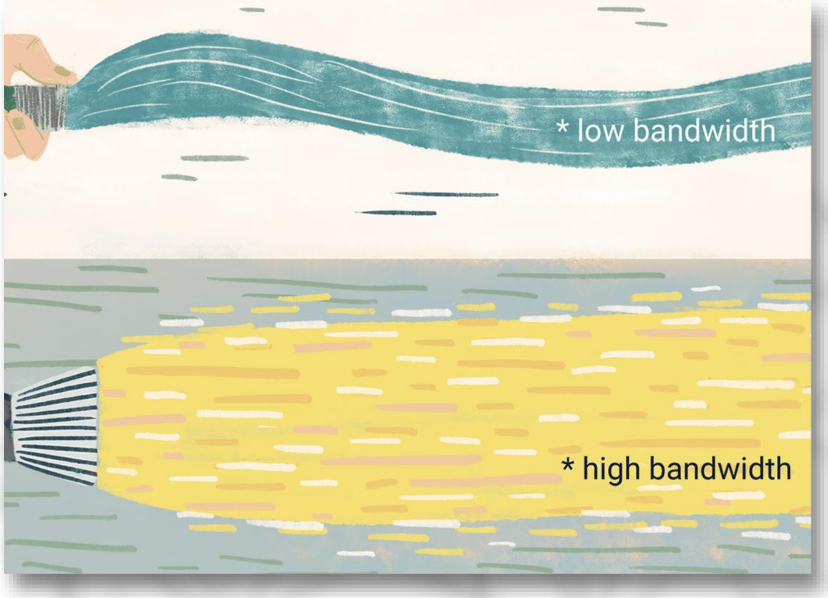
However, utilising these effectively comes with its own difficulties.

Latency & Data Communication

Solving Transit Delays | More Bandwidth

One approach is for ISPs to ensure that their link and router capacities far exceed the offered traffic loads, and to utilise creative routing of traffic to ensure that no single router becomes a point of serious congestion.

This approach tends to be practical only for large or core network operators who have flexible access to physical link infrastructures.



Latency & Data Communication

Solving Transit Delays | Privilege Access

This approach tends not to work where high-speed technologies cannot be deployed in a cost-effective manner.

ISPs and consumers must contemplate ways of prioritising access rather than hoping for the best.

This means identifying some IP packets as worthy of ‘better’ service than others.

Latency & Data Communication

Solving Transit Delays | Privilege Access

Central to these schemes is the goal of letting some people ‘jump the queue’ and get ahead of others.

This requires three steps:

1. Classification (to identify who or what deserves preferential treatment)
2. Separate queueing (to isolate those getting preferential treatment)
3. Scheduling (to actually provide the preferential treatment)

Latency & Data Communication

Solving Transit Delays | Privilege Access

The same applies in IP networks.

Congested routers need to classify, separately queue and preferentially schedule some packets over others.

Output ports that previously used only one queue will now have two (or more) queues – one for normal best effort service, and another for high-priority packets.

Latency & Data Communication

Solving Transit Delays | Privilege Access

Many network routers can classify IP packets using five pieces of information:

- The source IP address
- The destination IP address
- The protocol type
- The source port number
- The destination port number

This is called Flow Classification.

Latency & Data Communication

Solving Transit Delays | Privilege Access

A set of rules in each router dictates which combinations of IP address and port numbers are considered priority.

Based on these rules, every packet can be classified into a high-priority or normal-priority queue.

Ultimately, the scheduler decides when to transmit packets from each queue.

Latency & Data Communication

Solving Transit Delays | Privilege Access

Serialisation delays still apply to each packet transmission, and congestion-induced transmission delays and the potential for packet loss still exist.

However, these events now occur on a per-queue basis.

By splitting traffic into separate queues, we isolate the game traffic from much of the queueing and serialisation delays that afflict regular best effort traffic.

Latency & Data Communication

Solving Transit Delays | Privilege Access

This is fundamentally the principle behind *Net Neutrality*, the principle that if a service provider pays more to the ISP their data packets will receive preferential treatment and creating a two-tier system.

This two-tier system raises a number of ethical issues.

A short video on the subject can be found here.

REVIEW

Latency on the network can be caused my many different normal processes on the network

Latency can also be caused by the sheer volume of packets in transit on the network

There are solutions that the service providers can implement, both good and bad

QUESTIONS?