ETSI GR NIN 001 V1.1.1 (2021-03)

Non-IP Networking (NIN);
Problem statement: networking with TCP/IP in the 2020s

Disclaimer

The present document has been produced and approved by the Non-IP Networking ETSI Industry Specification Group (ISG) and represents the views of those members who participated in this ISG.
It does not necessarily represent the views of the entire ETSI membership.

Reference
DGR/NIN-001

Keywords
autonomic networking, core network, fixed networks, intelligence-defined network, internet, layer 3, network monitoring, network performance, network scenarios, next generation protocol, non 3GPP access

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - NAF 742 C
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° 7803/88

Contents

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The information pertaining to these essential IPRs, if any, is publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI IPR Policy, no investigation, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

Foreword

This Group Report (GR) has been produced by ETSI Industry Specification Group (ISG) Non-IP Networking (NIN).

Modal verbs terminology

In the present document "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

Introduction

The TCP/IP suite of network protocols is now over 40 years old and was designed for different requirements than the networking of the 2020s. This raises addressing, mobility, performance, and security issues that have required significant effort, energy, and cost to mitigate, and have been well documented, for instance in ISO/IEC TR 29181-1 [i.1].

Any form of wireless comms, whether they be 2/3/4/5G, satellite or Wi-Fi, needs to go through a 'wired' back-haul at some point, if not multiple points. How do all these technologies converge and connect edges to the fixed 'backbone'?

Connecting across the world on an architecture where ageing network protocols (IP, MPLS and more) are not able to move beyond best effort networking makes it difficult to meet the increasingly challenging guaranteed end-to-end SLAs required for high value, business- and safety-critical applications. National and international core networks need to work in conjunction with future internet protocol paradigms required to satisfy future network demands and deliver the promised 5G benefits, while being secure, robust, trusted, and resilient by design.

With the increasing challenges placed on modern networks to support new use cases (some of which require ultra-low latency) and greater connectivity, Service Providers are looking for candidate technologies that may serve their needs better than the TCP/IP-based networking used in current systems.

1 Scope

The present document describes the challenges of IP-based networking for fixed and mobile networks and ways in which new network protocols can result in improved performance and more efficient operation. Topics covered include:

- efficient use of spectrum;
- efficient forwarding;
- naming and addressing (including addressing lifecycle);
- mobility and multihoming;
- Quality of Service (QoS);
- time-sensitive networking;
- performance;
- authenticity, integrity, confidentiality, access control, and identifiers;
- lawful interception;
- ease of management; and
- migration from current technology.

2 References

2.1 Normative references

Normative references are not applicable in the present document.

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long term validity.

The following referenced documents are not necessary for the application of the present document but they assist the user with regard to a particular subject area.

[i.1] ISO/IEC TR 29181-1:2012: "Information technology -- Future Network -- Problem statement and requirements -- Part 1: Overall aspects".

[i.2] T-Mobile/IoT - Whitepaper, March 019: "The Game Changer for the internet of things".

NOTE: Available at https://www.t-mobile.com/content/dam/tfb/pdf/Whitepaper-Narrow-BandIo-T2019.pdf.

[i.3] IETF RFC 1144: "Compressing TCP/IP Headers for Low-Speed Serial Links".

[i.4] IETF RFC 2508: "Compressing IP/UDP/RTP Headers for Low-Speed Serial Links".

[i.5] ETSI TR 103 369: "CYBER; Design requirements ecosystem".

[i.6] ETSI TS 101 158: "Telecommunications security; Lawful Interception (LI); Requirements for network functions".

[i.7] ETSI TS 101 331: "Lawful Interception (LI); Requirements of Law Enforcement Agencies".

[i.8] IETF RFC 2205: "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification".

[i.9] IETF RFC 8578: "Deterministic Networking Use Cases".

[i.10] IEEE 802.1AS™: "IEEE Standard for Local and Metropolitan Area Networks -- Timing and Synchronization for Time-Sensitive Applications".

[i.11] IEEE 802.1Q™: "IEEE Standard for Local and Metropolitan Area Networks -- Bridges and Bridged Networks".

[i.12] Dr. N. Davies: "The properties and mathematics of data transport quality", 2009.

NOTE: Available at https://www.slideshare.net/mgeddes/intro-dataqualityattenuation.

[i.13] I. Johansson: "Congestion control for 4G and 5G access", Internet Engineering Task Force Internet Draft, July 2016.

NOTE: Available at https://tools.ietf.org/html/draft-johansson-cc-for-4g-5g-02.

[i.14] ETSI TS 133 210: "Digital cellular telecommunications system (Phase 2+) (GSM); Universal Mobile Telecommunications System (UMTS); LTE; 5G; Network Domain Security (NDS); IP network layer security (3GPP TS 33.210)".

NOTE: Available at https://www.etsi.org/deliver/etsi_ts/133200_133299/133210/.

[i.15] "Patterns in network architecture: a return to fundamentals", chapter 6 'Divining Layers', J. Day, Pearson, 2008, ISBN 0-13-225242-2.

[i.16] P. Teymoori, M. Welzly, S. Gjessingz, E. Grasa, R. Riggio, K. Rauschk, D Siracusa: "Congestion Control in the Recursive InterNetworking Architecture (RINA)", IEEE ICC 2016 - Next-Generation Networking and Internet Symposium, 2016.

[i.17] GSMA, June 2019: "NB-IoT Deployment Guide to Basic Feature set Requirements".

NOTE: Available at https://www.gsma.com/iot/wp-content/uploads/2019/07/201906-GSMA-NB-IoT-Deployment-Guide-v3.pdf.

[i.18] IETF RFC 7426: "Software-Defined Networking (SDN): Layers and Architecture Terminology".

[i.19] IEEE Std 802™: "IEEE Standard for Local and Metropolitan Area Networks: Overview and Architecture".

3 Definition of terms, symbols and abbreviations

3.1 Terms

For the purposes of the present document, the following terms apply:

QUIC: UDP-based transport and session-control protocol with claimed performance improvements over TLS/TCP

3.2 Symbols

Void.

3.3 Abbreviations

For the purposes of the present document, the following abbreviations apply:

API	Application Program Interface
ARP	Address Resolution Protocol
AS	Autonomous System
BGP	Border Gateway Protocol
BLER	Block Error Rate
CA	Certificate Authority
CBOR	Concise Binary Object Representation
DHCP	Dynamic Host Configuration Protocol
DNS	Directory Name Service
HTTP	HperText Transfer Protocol
HTTPS	HperText Transfer Protocol Secure
IANA	Internet Assigned Numbers Authority
ICMP	Internet Control Message Protocol
IETF	Internet Engineering Task Force
IoT	Internet of Things
IP	Internet Protocol
IPX	Internetwork Packet Exchange
JSON	JavaScript Object Notation
LAN	Local Area Network
LI	Lawful Interception
LPWAN	Low Power Wide Area Networks
MAC	Media Access Control
MIMO	Multiple Input and Multiple Output
MPLS	Multi-Protocol Label Switching
NAT	Network Address Translation
NR	New Radio
OSI	Open Systems Interconnection
OUI	Organizationally Unique Identifier
PST	Pacific Standard Time
QoS	Quality of Service
RFC	Request For Comment
RIR	Regional Internet Registry
ROA	Route Origin Authorization
ROHC	RObust Header Compression
RPKI	Resource Public Key Infrastructure
RSVP	ReSource Reservation Protocol
RTP	Real-time Transport Protocol
SDN	Software-Defined Networking
SINR	Signal to Interference and Noise Ratio
SIP	Session Initiation Protocol
SYN	SYNchronize (TCP control flag)
TCP	Transmission Control Protocol
TLS	Transport Layer Security
TSN	Time-Sensitive Networking
TTI	Time Transmission Interval
UDP	User Datagram Protocol
USA	United States of America
VPN	Virtual Private Network

4 Efficient use of cellular radio spectrum

4.1 Introduction

4.1.0 Requirement for efficiency

Radio spectrum is regulated, finite, and expensive for a cellular radio network to acquire and use. Spectrum is shared among all devices attached to the radio network to enable download (from the network to the device, the 'downlink') and upload (from the device to the network, the 'uplink'). If the operator can share the spectrum efficiently, among users and their applications, then the operator may reduce their costs, transmit more data in shorter time intervals, and enable more devices to communicate simultaneously. This clause summarizes existing techniques to use spectrum efficiently at the radio and other layers, and how non-IP network protocols could further improve spectral efficiency.

4.1.1 Definition of efficiency

In the present document, networking efficiency is defined as the number of application bits per Hz per second. Application bits are the data communicated between the client and server applications once all network headers have been removed.

4.2 Radio techniques for spectral efficiency

4.2.1 Space division

Radio frequencies experience path loss as they travel through space, proportional to the distance they travel and affected by environmental conditions (absorption losses) along the path. Eventually the signal is attenuated to the point where the Signal to Interference and Noise Ratio (SINR) is too low to carry information reliably. Cellular networks calculate the path loss to determine the boundaries of each cell - after which the frequency ceases to be reliable, assuming a fixed power level at the transmitter. This allows networks to reuse that same frequency in other cells, although to avoid interference at the cell boundary, the same frequency is typically not used in adjacent cells that abut that boundary.

Whilst traditional cellular antennae transmit signals in an arc to fill a portion of the cell, Massive Multiple Input and Multiple Output (MIMO) antenna systems allow the narrow targeting of signals - 'beamforming' - which provides higher throughput and reduced latency between the MIMO antenna and receiving device.

Spatial multiplexing allows more than one data signal to be transmitted and received simultaneously on the same channel. MIMO achieves this by utilizing multipath-propagation to increase the number of paths a signal can take from transmitter to receiver.

4.2.2 Time division

The same frequency may be reused to communicate information if the frequency is divided into timeslots. Attached devices are allocated a timeslot by the radio network and transmit or receive within that slot. The longer the timeslot, the more information can be communicated within it - but at the cost of reducing the number of devices that can be served on that frequency in a given time.

4.2.3 Efficient modulation schemes

Cellular networks from 2G onwards transmit digital data. The transmitter takes digital data as an input, and transmits analogue radio in a way that allows the receiver to parse and reconstruct the digital data. This is achieved through modulation - wherein characteristics of the analogue radio wave are controlled and adapted to signal information to the receiver - with both sender and receiver utilizing modulators-demodulators (modems).

Given that the speed of a radio wave is fixed - defined by the speed of light through air - there are three characteristics of a radio wave that can be modulated: amplitude, frequency and phase.

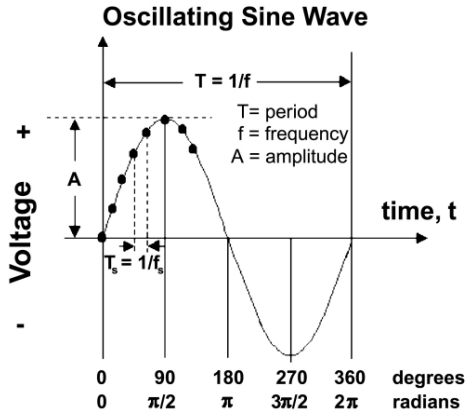

Figure 4.1: Characteristics of a sinusoidal wave
(source: National Institute of Standards & Technology, 2010
Public Domain as per https://commons.wikimedia.org/wiki/File:Oscillating_sine_wave.gif)

When a device attaches to the cellular network, it is allocated a frequency from the operator's spectrum. Data to be communicated to/from the device is encoded into an analogue signal. One of the signal characteristics - amplitude, frequency, phase - is selected, and used to modulate a number of 'sub-carriers' (sub-divisions of the operator spectrum).

The resulting information is transmitted to the receiver, which demodulates the sub-carriers to retrieve the information.

The goal of a modulation scheme is to balance data throughput - the amount of data that can be modulated per second per Hz - with resilience from interference between the tightly-spaced frequency sub-carriers.

4.2.4 Radio resource block allocation

Devices continuously signal their received signal strength to the network. The network uses these values, as well as the size of packet queue for that user, to optimize the amount of radio resource to dedicate to that user, in the form of radio resource blocks. 5G New Radio (NR) introduces the concept of bandwidth parts, which allows the spectrum to be flexibly sliced up into groups of resource blocks for different users depending upon their needs: for example small groups for NB-IoT and large groups for enhanced mobile broadband.

4.2.5 Transmission interval

The Time Transmission Interval (TTI) is the duration of a transmission. A shorter TTI allows more transmissions per second, but with a reduced data payload (i.e. fewer radio resource blocks). The TTI can be adapted based on the application, for example a short TTI for Ultra-Reliable Low-Latency Communications with small. frequent payloads.

4.2.6 Narrowband IoT and Non-IP Data Delivery

Narrowband-IoT (NB-IoT) [i.17] is used for communication with low-power, low-throughput devices. The use of a narrow frequency band (200 kHz) at a lower carrier frequency reduces maximum throughput but also allows for signal penetration indoors.

NB-IoT may operate in a 'Non-IP Data Delivery' mode to remove the IP header from the transmitted payload, improving efficiency (see [i.2]). This is important as it minimizes radio transmissions, which apply a significant drain to the batteries of low-power devices.

4.3 Non-radio techniques

4.3.1 RObust Header Compression (ROHC)

IP encapsulation results in a per-packet header overhead: 20 bytes for IPv4, 40 bytes for IPv6 due to the increased address lengths. Transport protocols contribute an additional 20 bytes (TCP) or 8 bytes (UDP), and for "live" media such as audio there will also be at least 12 bytes of RTP header.

To tackle this problem, RObust Header Compression (ROHC, IETF RFC 1144 [i.3] and IETF RFC 2508 [i.4]) identifies redundant information (that which will be sent in every packet of a flow) and only transmits it in the first packet. Subsequent variable information (segment numbers, etc.) are transmitted in compressed form, and reassembled by the receiver.

Whilst ROHC may seem ideal to solve the problem of header overheads - reducing them to around 2 or 3 bytes - there are caveats:

- It incurs a cost to operators if they activate ROHC in a software licence.
- It incurs compute energy and latency.
- It requires a tuning of the Block Error Rate (BLER) used in radio transmission to ensure that the important information (the flow metadata) is not lost, since that would affect the following packets in the flow.

It may be for these reasons that operators typically only apply ROHC to VoLTE (Voice over LTE) flows. There are two reasons for this:

- the (IP, UDP, RTP) headers are almost the same size as the VoLTE payload (~60 bytes) , meaning header compression is more beneficial than e.g. video streams with per-packet payloads of 1 340 bytes;
- operators may be penalized by regulators if their voice call completion rate drops below a certain threshold, hence operators are more likely to invest in resources to ensure that does not happen.

Operators are less likely to apply ROHC to general Internet traffic due to the processing and licence costs. This is also the case with Low-Power IoT traffic, where the BLER issue is exacerbated. This has motivated new workarounds, such as Narrowband IoT (NB-IoT) omitting IP from the transmission entirely.

Rather than compressing and inflating IP and transport layer headers, ISG NIN proposes to tackle the problem at source through an efficient protocol design. This can reduce costs and latency, and avoid issues with Block Error Rates at transmission.

4.3.2 Payload compression

Payload - the application data carried in packets - is sized according to the amount of data and its serialization. Efficient serialization schemes reduce the overhead in representing data for interpretation by the receiving application: for example, JSON or CBOR (Concise Binary Object Representation). The entire data payload can be further compressed using gzip or similar for transmission with decompression at the client.

4.4 Areas that non-IP networking can help improve

4.4.1 Transmission overheads

Whilst ROHC (see clause 4.3.1) reduces IP header overheads significantly, it incurs financial cost, and requires compute and energy expenditure at both network and user equipment. Hence today it is typically used only for operator VoLTE (Voice over LTE) services, which have a significantly larger header-size to payload-size ratio than e.g. video streaming.

IoT services - especially Low Power Wide Area Networks (LPWAN) - are similar to VoLTE in that they have a large header-size to payload-size ratio (with a maximum throughput of 250 Kb/s on downlink). Removing redundant encapsulations and reducing header size will intrinsically improve networking efficiency for such services without a requirement for ROHC.

4.4.2 The propagation range of ultra-low latency services

The Transmission Time Interval (TTI) of the 5G radio air interface is ~140 μs. To realize this at cell edge in a network configured as described in clause 4.4.4, the radio payload per TTI can be no larger than ~18 bytes. ~100 byte payloads (TCP/IP + IPsec) will take up approximately 5 TTIs, which places a latency constraint on both the application and other applications competing for transmission blocks. Lower networking payloads means more efficient use of TTI.

4.4.3 Mobility performance

Smaller transmissions can be communicated with reduced latency, proportional to the reduction in the ΔQ/S metric (see [i.12]). This is especially important during mobility handover of a client terminal between two cells. "*Packet retransmission typically means that the amount of data to transmit increases immediately after the handover, (hence) is a good practice to keep the amount of data in flight as small as possible, without sacrificing throughput*" [i.13].

4.4.4 Energy expenditure

Cellular networks are implemented as neighbouring cells, with antennae transmitting and receiving as far as the 'cell edge'. Past this edge the Signal to Interference and Noise Ratio (SINR) will degrade to the point where the client device will need to attach to the new cell it has now entered.

The key factor in this radio degradation is propagation, inversely proportional to the square of the distance between the user device and antenna: the power required to transmit information to cell edge is approximately ten times that required to transmit at the cell centre. For example, consider a radio cell configured for 1 Mb/s uplink at the cell edge. To support low-latency services, the first packet needs to arrive with high reliability. This requires a Packet Error Rate limit of 10 %, which increases the link budget for the radio air interface by 10 dB to 20 dB. This leaves the following choices:

- Increase the power by a factor of 10 or greater. This is not feasible at the battery-powered terminal, nor sustainable at the radio base station.
- Halve the distance to the terminal.
- Reduce the transmission payload.

Of these, only the last option is realistic. However the TCP/IPv4 header alone is 40 bytes, and 3GPP recommend use of IPsec as described in ETSI TS 133 210 [i.14] which adds a further 50 bytes to 60 bytes if used on User Plane traffic. Header compression is expensive, increases ΔQ/S at both network and client, and uses more energy. Whereas the problem can be more efficiently and effectively solved by having greatly reduced header sizes, with scope-specific shorter addresses and an overall simplification of the radio protocol stack.

Energy is also a concern at the terminal. Network antennae are hosted by a fixed-power base station; terminal devices, however, are typically battery powered; and have several orders of magnitude less transception power. Reduced header overheads will allow these devices to transceive application data in shorter radio bursts, saving battery power.

4.4.5 Resilience

Radio signals are volatile in uncontrolled environments, especially where the client device is moving. This leads to signal reflections, channel fading and other factors that reduce SINR. Where application data is inflated with headers, it will require more radio resource blocks to transmit. This increases the probability of one of the transmitted resource blocks being corrupted in transit or not received, requiring retransmission. Conversely, a more efficient application data networking protocol requires fewer resource blocks and hence less risk of one needing to be retransmitted.

4.4.6 Congestion control

A cellular client's traffic traverses the operator's radio network and core network to access the Internet. Of these networks, the radio link is the most susceptible to volatile bandwidth and jitter, due to unpredictable signal strength at the client and radio cell capacity at the operator.

End-to-end congestion controls (as used in TCP) operate at the application layer and require round trips to indicate congestion. This creates two problems: the reaction to congestion occurs at the farthest point from the likeliest source (the radio link), and, since congestion control is not aware of radio state, it risks server retransmissions of data already queued at the radio for retransmission, thus compounding the problem.

A related issue is that the radio resource scheduler - responsible for allocating the appropriate portion of radio frequency for transmission - may add 10 ms - 20 ms delay. This may be wrongly perceived as congestion by end-to-end clients that use jitter as a congestion signal [i.13], [i.15].

Shorter congestion control loops between the local sender and receiver mean a faster reaction to congestion [i.13], [i.16], and avoiding propagation of the issue to endpoints not aware of the local resource management state.

5 Naming and addressing

5.1 Introduction

There used to be many alternatives to the TCP/IP communications stack; a few still exist (in silos) and many more existed in the past.

But with the evolution of the Internet, a single unifying set of protocols was required to interconnect networks.

The Internet as it is today is purely and solely based on TCP/IP (IPv4 and IPv6 addresses).

The following features are common to the TCP/IP addressing as well as all other stacks:

- Allocation.
- Abundance.
- Assignment mode.
- End system Configuration.
- Resolution.
- Hierarchy.
- Advertisements.

5.2 Allocation

Which entity(ies) can globally assign unique addresses?

The current address allocation model for IPv4 and IPv6 addresses is a **centralized-model**, it is characterized by:

- Monopoly: where few entities govern the allocation of addresses, (i.e. IANA and the 5 Regional Internet Registries (RIRs) for IPv4 and IPv6 addresses allocation).
- Unfairness: back in the 90's, Class-B IPv4 subnets (/16) were assigned to many hospitals and small businesses in the USA.
- Justification: requesting entities need to justify their needs of addresses. Also, the requesting entities can only be Internet service providers or telco providers.
- Renewal/maintenance: addresses' owners need to pay annual recurring fees for renewal of their allocated addresses (cost varies per RIR per subnet IPv4/IPv6, Figure 5.1).

5.3 Abundance

With only 32-bits available for the IPv4 address, depletion/exhaustion of IPv4 global addresses started to happen years ago. Theoretically there are less than 4,3 billion IPv4 addresses that could be globally assigned and many of them are reserved (i.e. multicast, reserved, etc.).

Back in mid 90's, during the discussions of the next generation Internet protocol IPng, the view that 8 bytes of address were enough to meet the current and future needs of the Internet (squaring the size of the IPv4 address space).

The anticipation, back then, for the new protocol was its ability to uniquely identify and address at least 1 000 000 000 (10^9) leaf networks.

More address length would waste bandwidth, promote inefficient assignment, and cause problems in some networks (such as mobiles and other low speed links).

But then it was agreed that 16 bytes for the IPng address was about right (that resulted in the IPv6 128-bits address length). This length supports auto-configuration as will be shown later.

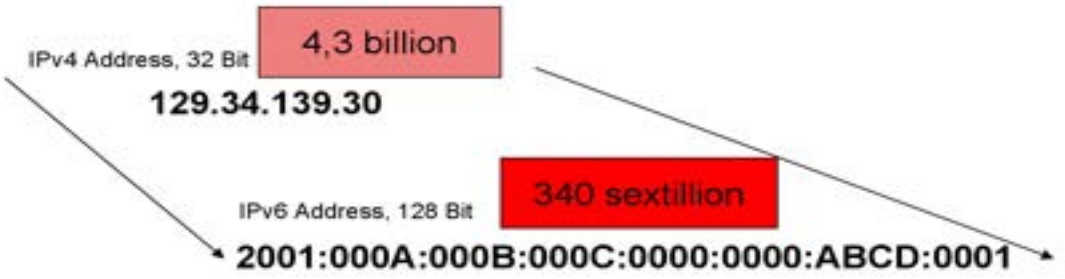

Figure 5.1

5.4 Assignment mode

Are the addresses, allocated to the end systems, assigned per device or per interface?

Per-interface is the assignment mode for IP addresses (same as IPX, AppleTalk), although in this mode, the consumption of addresses is high and could lead to depletion (i.e. IPv4 depletion), but it helps a lot in end systems discovery.

Figure 5.2 shows that for a node (a router for instance), the IP addresses are different on the different interfaces.

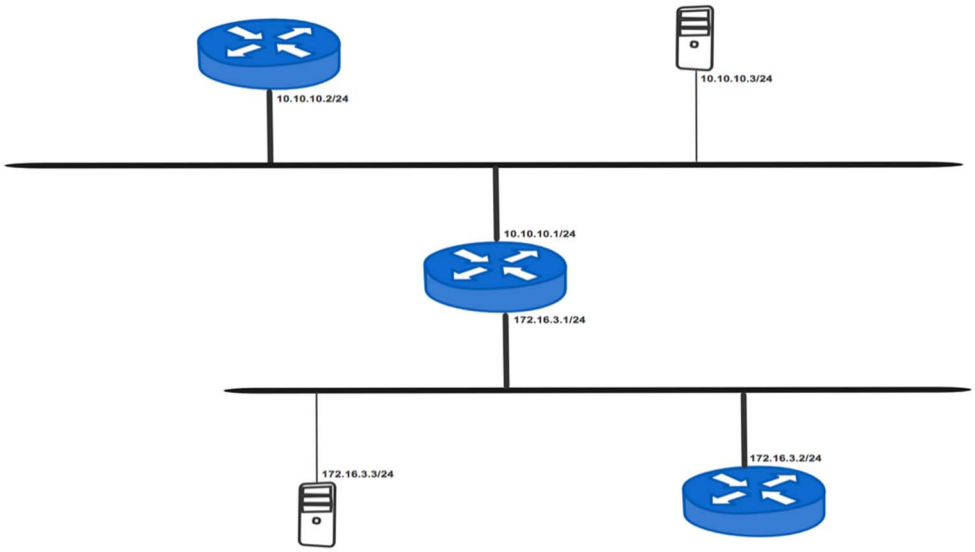

Figure 5.2

This 'Per-interface' assignment mode does not allow for mobility within an area/domain without additional mechanisms.

Another assignment mode - though irrelevant to IP addresses - is 'Per-device'. This mode allows for better mobility within an area or domain as the address globally identifies the node itself rather than one of its interfaces or links.

The 'AA.1.2.3.5' node can 'roam' easily within the boundary of its area (AA) and will still be reachable (Figures 5.3 and 5.4).

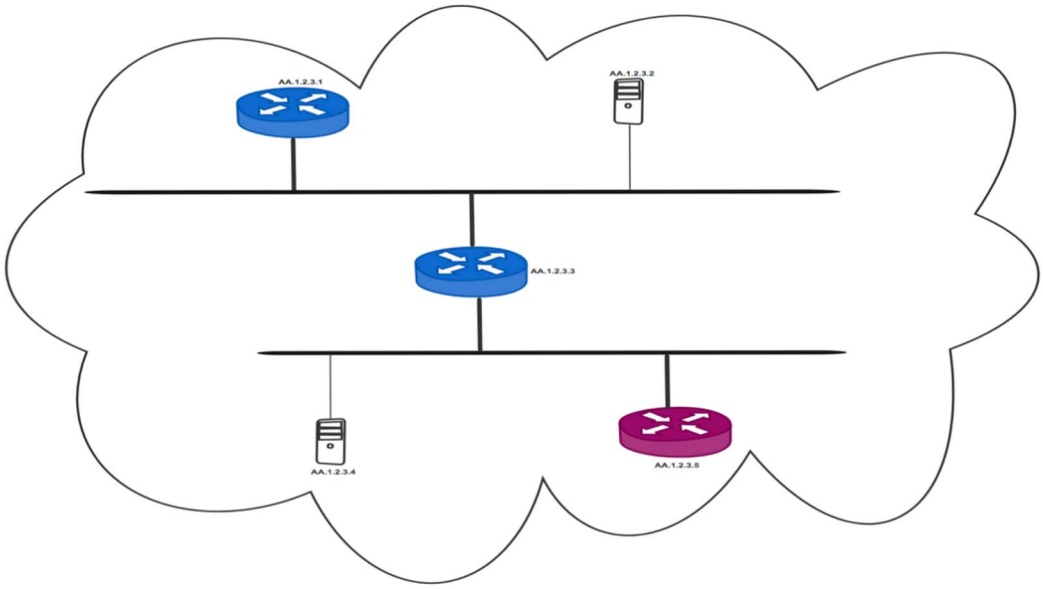

Figure 5.3

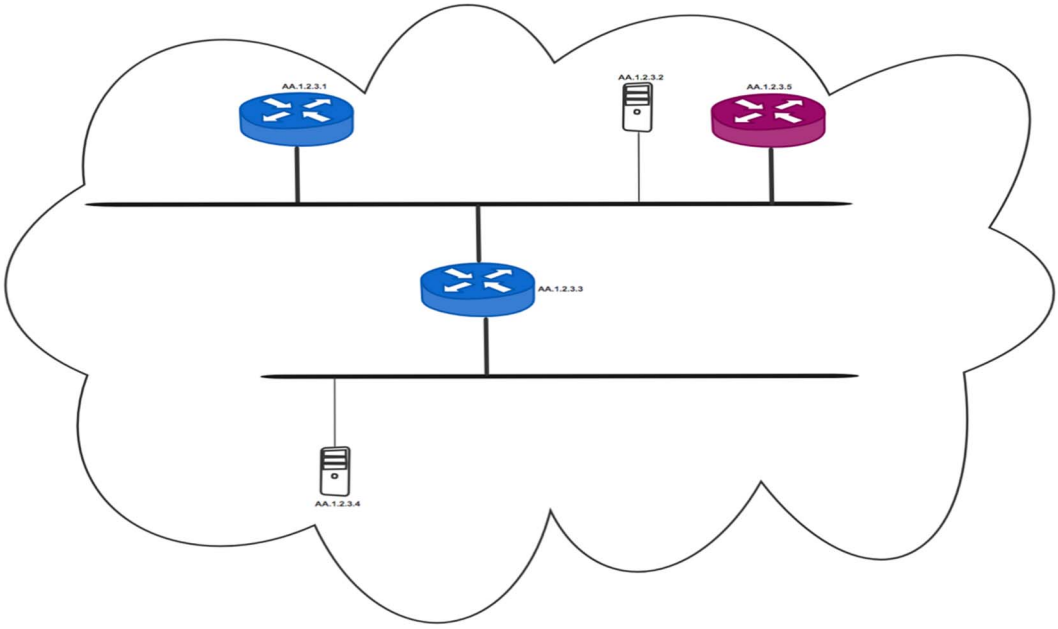

Figure 5.4

5.5 End system configuration

The current mode of IP address assignment is **Server-based** (i.e. DHCP/BOOTP). That means, upon booting up; an end system needs to communicate with a server to retrieve their IP address, subnet mask, name, default gateway and other information for proper operation.

For auto-configuration to work, the Internet address needs to be large enough to accommodate/embed the IEEE MAC (Media Access Control) 48-bits address defined in clause 8 of IEEE Std 802 [i.19].

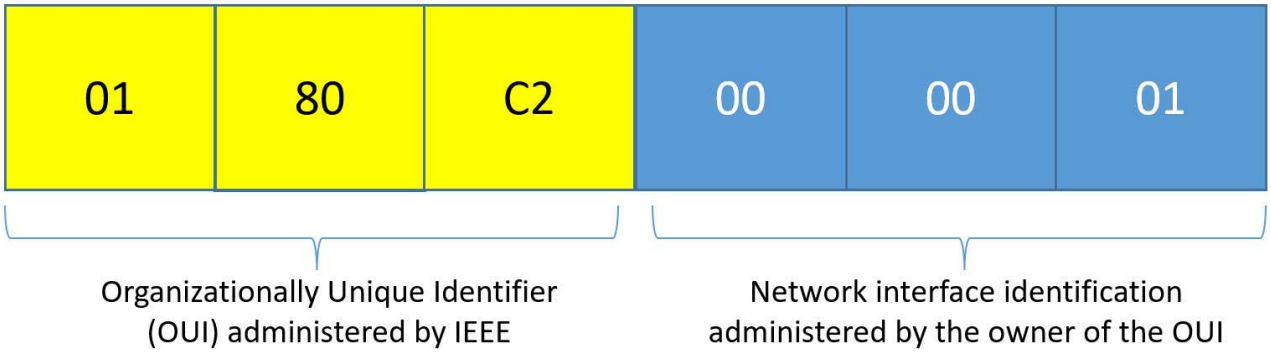

Figure 5.5: Structure of a MAC address

With only 32-bits for its address, IPv4 address is not big enough to accommodate the MAC address, hence a server-based auto-configuration technique is always required (unless manual configuration for each new host is an easy task for the network team).

With its 128-bits address, IPv6 can easily support '**Server-less**' auto-configuration by embedding the IEEE MAC 48-bits addresses as the Host ID while depending on some other technique to solicit (or listen to) the local router for the Network ID defined in IEEE Std 802 [i.19].

Along with auto-configuration always come security concerns. These may restrict the ability to offer this level of address autoconfiguration in some environments but there need to be mechanisms in place to support whatever level of automation which the local environment feels comfortable with.

5.6 Resolution

Upon power-up, an end system retrieves all necessary information, for its data communications, from the DHCP server (IP address, subnet mask, name, default gateway), the DHCP requests and responses are shown in dotted red lines in Figure 5.6.

The router knows the location of an end system (on which segment) by virtue of the Network ID of their layer-3 address, that matches the same network ID of one of the router's interfaces' addresses.

The router can find the Layer-2 MAC address of an end system using Address Resolution Protocol (ARP).

An end system can discover other end systems attached to the same LAN, in order for direct communications between them, by virtue of the same Network ID of their layer-3 addresses along with the result of the ARP protocol (as they need to communicate directly using their MAC addresses).

In case of multiple gateway routers, an end system can be redirected to the best exit gateway via ICMP redirects, shown in blue arrow in Figure 5.6.

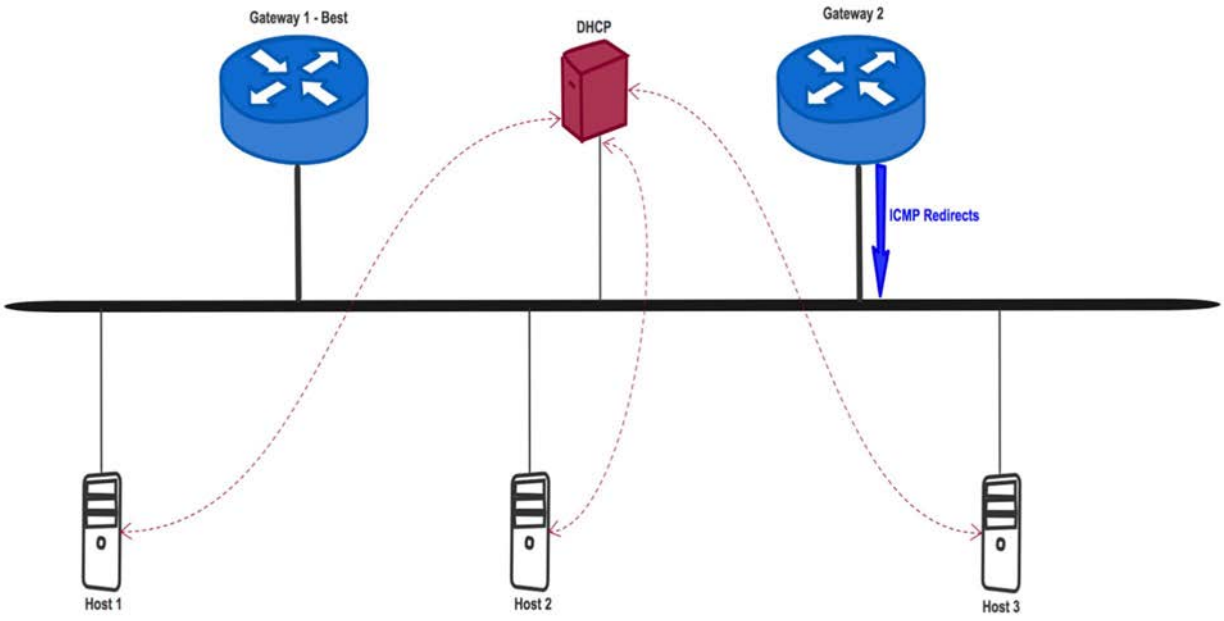

Figure 5.6

5.7 Hierarchy

Whether the Internet address should be fixed in length or variable was a debate since a long time.

IPv4 has its address length fixed (4 bytes) since its inception. In the mid 90's, when people were thinking about the IPng that would replace the existing IPv4, people pro the variable length address viewed that the size of the address could be adjusted to the demands of a particular environment, and to ensure the ability to meet any future networking requirements, but despite all that, IPv6 was born with fixed length address!

On the other hand, when it comes to the boundary between the Network part of the address (Network ID) and the host part (Host ID); this boundary could be fixed or variable.

IPv4 addresses originally had fixed boundary between the Network ID and Host ID (classful networks, Figure 5.7).

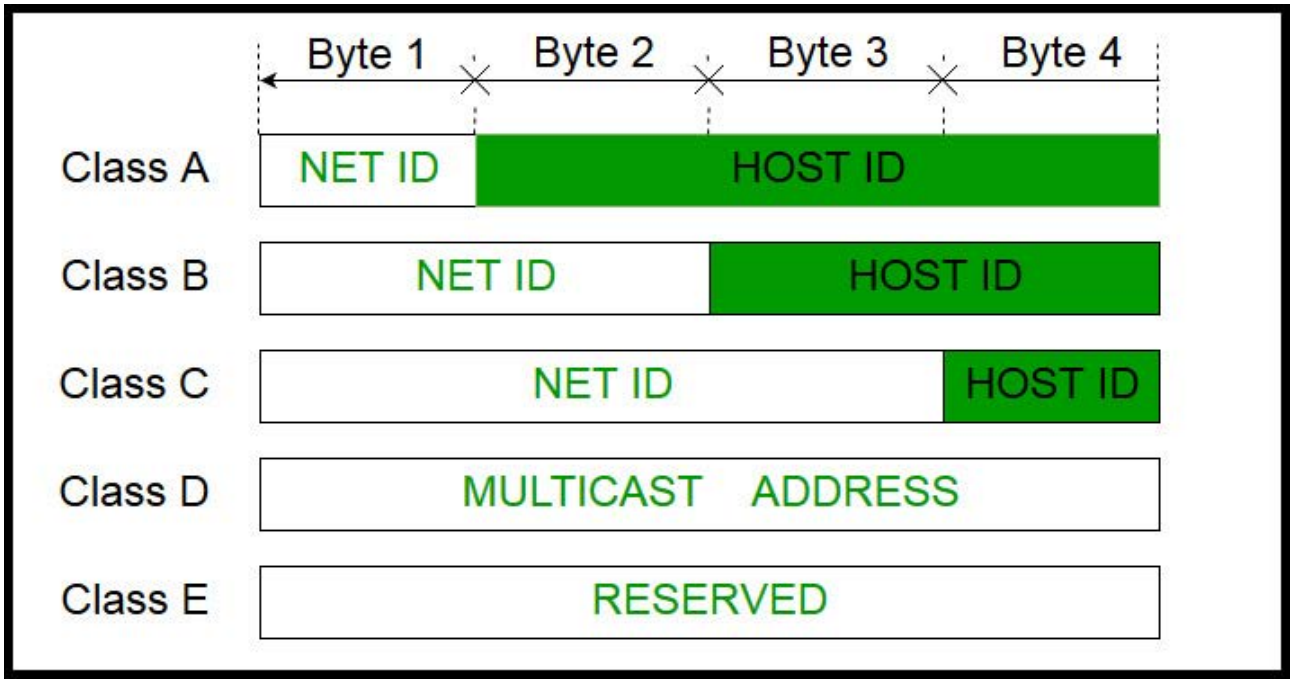

Figure 5.7

Then IPv4 addresses started to accommodate variable boundaries between Network ID and Host ID (classless subnets, Figure 5.8), where the notion of prefix (network portion) and suffix (host portion) arose.

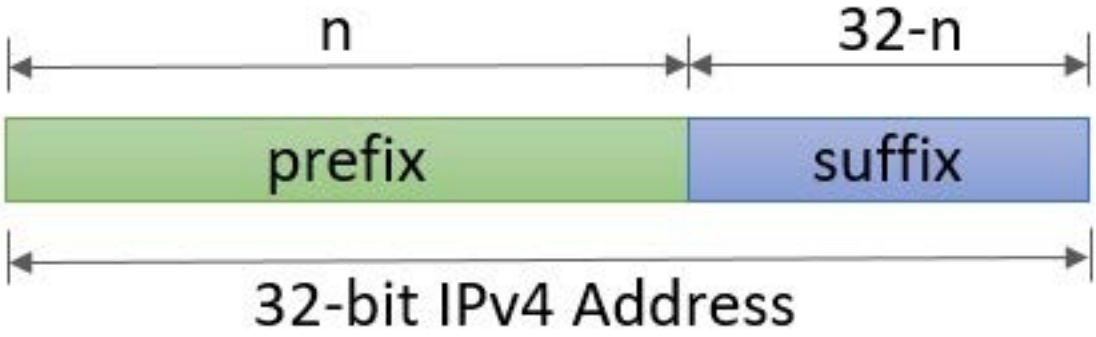

Figure 5.8

5.8 Advertisements

IP address advertisements have impact on both the Internet routing security as well as the size of the global Internet routing table.

In order to secure the Internet routing architecture and prevent against repetitive BGP hijacking incidents; authenticity and verification of Route Origin Authorization (ROA) are deemed necessary. Currently less than 10 % of the publicly advertised subnets are verified against Resource Public Key Infrastructure (RPKI) framework.

As for the Internet scaling and the increasing size of the routing tables held in the Internet backbone routers; providers need to more aggressively advertise their routes only in aggregates. Currently no subnet mask greater than 24 (Class-C subnet) should be advertised and accepted into the Internet routing system.

In that regard, providers also advise their new customers to renumber their networks in the best interest of the entire Internet community.

NOTE: Even if future Internet addresses are to be designed with aggregation in mind, switching to these new addresses will not solve the routing table size problem unless these new addresses are assigned rigorously to maximize the effect of such aggregation. This efficient advertising of routes can be maintained if the new address architecture allows autoconfiguration mechanisms to allow easy renumbering if a customer decides to switch providers. Customers who receive service from more than one provider may limit the ultimate efficiency of any route aggregation.

6 Security

6.1 IPsec

IPsec, also known as the Internet Protocol Security or IP Security protocol, defines the architecture for security services for IP network traffic. IPsec describes the framework for providing security at the IP layer, as well as the suite of protocols designed to provide that security, through authentication and encryption of IP network packets. Also included in IPsec are protocols that define the cryptographic algorithms used to encrypt, decrypt, and authenticate packets, as well as the protocols needed for secure key exchange and key management.

IPsec can be used to protect network data, for example, by setting up circuits using IPsec tunnelling, in which all data being sent between two endpoints is encrypted, as with a Virtual Private Network (VPN) connection; for encrypting application layer data; and for providing security for routers sending routing data across the public internet. IPsec can also be used to provide authentication without encryption, for example to authenticate that data originates from a known sender.

Internet traffic can be secured from host to host without the use of IPsec, for example by encryption at the application layer (Layer 7 of the OSI model) with HTTP Secure (HTTPS) or at the transport layer (Layer 4 of the OSI model) with the Transport Layer Security (TLS) protocol. However, when traffic uses encryption or authentication at these higher layers, threat actors may still be able to intercept protocol information that may expose data that should be encrypted.

6.2 Internet routing security and BGP hijacking

Internet routing security is one of the pressing issues in the Internet. It encompasses the correct announcement and propagation of IP prefixes between the domains or - using the Internet terminology - Autonomous Systems (AS).

BGP (Border Gateway Protocol) is the protocol that manages the advertisement and propagation of prefixes between the different domains.

BGP configuration is mostly done via out-of-band mechanisms where network operators tell each other which prefixes to announce among themselves. Hence, an accidental misconfiguration or a malicious attacker controlling a BGP router can divert traffic to networks which should not receive it or make ranges of IP addresses unavailable (thus effectively denying global services). This attack is commonly known as BGP hijacking and can be accomplished by forging BGP announcements and propagating them to neighbouring ASs.

There have been many incidents of BGP hijacking. For example, on November 12th, 2018, between 1:00pm and 2:23pm PST, some customers in the USA noticed issues connecting to certain services. The reason was that the traffic (addressed to these services) was getting routed to a different network and dropped there.

That was a severe denial of service. Some analysis indicated that the origin of this leak was the BGP peering relationship between two provider, from which routes leaked via some transit ISPs.

Numerous similar incidents taking place in 2017 were reported here:
https://www.internetsociety.org/blog/2018/01/14000-incidents-2017-routing-security-year-review/.

One solution to mitigate the BGP hijacking is the use of Resource Public Key Infrastructure (RPKI), that is a repository where the legitimate owners of IP prefixes, AS numbers and Route Origin Authorizations (ROAs), a certificate to allow an AS to announce an IP prefix) are recorded.

Unfortunately, the global deployment of the RPKI is slower than expected with only 10 % of the total Class-C (/24) subnets owned by the five RIRs being protected by the RPKI. The reasons of this have been mainly:

- Centralized operations: Certification Authorities (CAs) hold ultimate control of resources in the RPKI. Since IP addresses are a critical asset of RPKI's participants (especially ISPs), they would like to have a higher degree of control over them, but without losing the security of being certified by a CA, i.e. balanced power between users and CAs.
- Management complexity: PKIs are cumbersome to manage, like the case of key rollover. In addition, deploying these extensions is not trivial and requires trained staff and investment.
- Exposure of business relationships through peering agreements in the RPKI.

In addition, the RPKI faces implementation and transparency challenges.

6.3 BGP instability

Routing instability is one of the most important and pathological problems of the Internet. This kind of instability can cause loss of service, waste of network resources and service degradation for QoS demanding applications.

Border Gateway Protocol (BGP) is the most important and widely used inter-domain routing protocol, and so far it has been very successful in accommodating the fast growing demands of the Internet. It has been successful in keeping the size of routing tables reasonable regardless of the huge increase in the number of networks advertised onto the Internet.

But like other routing protocols, BGP suffers from routing instability that can be very expensive because it incurs high costs by making major changes in traffic paths at high-speed inter-domain links.

A common case of BGP instability is triggered by congestion inside a routing domain causing loss of interior BGP (iBGP) sessions between BGP routers inside that domain, that would cause route flapping and hence instability in the Internet routing system.

The same happens with BGP Flapping that can occur when there is an unstable peer, so the BGP routes advertised by that peer keep on disappearing and reappearing in the routing table.

Another important vulnerability of BGP comes from its underlying transport protocol, TCP, that is itself vulnerable to several kinds of denial of service attacks (for example SYN flooding attacks).

Also, a common problem of BGP is 'black holing', where a bug, hacker or wrong manual configuration can cause a BGP speaker to announce some networks back to the AS to which they originally belong. Without proper precautions or loop avoidance mechanisms, this can cause serious instability of the inter-domain routing system.

6.4 Control plane security

In an IP network control messages are routed in the same way as user data. This provides opportunities for user programs to impersonate control plane entities and subvert control plane protocols.

More generally, because packets are routed purely on the basis of addresses in the packet headers there is no authentication of the sender of a packet, so a malicious botnet can flood the network with packets addressed to some distant location.

6.5 Lawful interception

Lawful interception consists of the real-time "handover" of copies of network metadata, including signalling, as well as content. For many designated networks and services, including those available to the public, the requirement to support interception provisioning capabilities "by design" is a condition of licensing or required by law. Acquisition can occur anywhere in the communications path. Virtualization is producing new design challenges to meet these requirements. For more detail of the requirements placed on network operators for LI see clause 4.3 of ETSI TR 103 369 [i.5], also ETSI TS 101 158 [i.6] and ETSI TS 101 331 [i.7].

The broad consequence of these requirements on the network provision is that the network operator has to give assurance of the unobservability of any measure implemented and activated against any network element or network user in support of a lawful request to intercept the content or signalling associated to the element or user. Of itself IP does not inhibit LI but ensuring that any points of interception cannot be bypassed by the user requires careful design of the network.

7 Quality of Service and time-sensitive traffic

The Internet was designed to carry messages between computers. In computation, the priority is to process data accurately and in the right order; how long any part of the process takes is of minor importance. Moreover, the size of these messages and the rate at which they are generated are in general unpredictable, as they will depend on details of the communicating processes including whether they need to wait for some external event. Thus a user surfing the Web may click on a link, causing the browser to fetch a large amount of text and images from a server, and then spend some time reading before clicking on another link.

The demand on any part of the network is therefore variable, in much the same way as demand on a road system, and just as on a road system queues are liable to build up at busy times. Unlike roads, queuing only occurs inside switches, where the queues occupy buffers that are of finite size, and when a buffer is full incoming packets are simply dropped. Transport protocols such as TCP make the assumption that delays are an indication of congestion and reduce the rate at which they send data accordingly, in an attempt to prevent buffers filling up. On a mobile network performance can be compromised by the interaction between this mechanism and lower-layer retransmissions occasioned by radio interference. However, the main effect is that the data rate and the latency (the time from sending a packet until it arrives at its destination) depend on conditions in the network in a way that is neither predictable nor controllable by the application.

Much of the traffic on the Internet now consists of media such as audio and video, where a continuous stream of data is sent, and where it needs to be consumed at a steady rate to avoid jerkiness in video and clicks and pops in audio. For consuming streamed content such as podcasts this can be ensured by downloading the data sufficiently far ahead of it being required to smooth over any gaps in transmission, but for two-way live transmissions such as telephone calls or video conferences the delays introduced can make it difficult to have a natural conversation. Minimizing the delay makes it more likely that data will not arrive in time, resulting in audio artefacts and drop-outs. Some of the new applications proposed for 5G, such as those involving tactile feedback, will need even shorter delays.

Various schemes have been proposed to mitigate the effects on live media. Using UDP instead of TCP allows packets to be sent at a constant rate but does not prevent them being dropped before reaching their destination. RSVP (IETF RFC 2205 [i.8]) allows capacity to be reserved for a media stream but does not guarantee that the packets will be sent over the route on which the reservation has been made, and is not widely used. DetNet (IETF RFC 8578 [i.9]) provides for sending multiple copies of a packet by different routes in the hope that one of them will arrive on time, but to be able to offer guarantees it needs to invoke other protocols such as TSN (IEEE 802.1AS [i.10], IEEE 802.1Q [i.11]), which is a complex lower-layer protocol that has evolved from Residential Ethernet and Audio-Video Bridging and requires network-wide synchronization. SDN [i.18] allows media streams to be given priority over other traffic.

IP is a "connectionless" service with each packet being self-contained, so does not include any provision for applications to signal that they will be sending a regular stream of packets, nor to negotiate any kind of latency guarantees. This results in the poor service that is experienced by users of videoconferencing and voice-over-IP applications.

8 Network management

The original intention for IP was that all information concerning the service a packet should receive from the network would be carried in the packet header, and each packet would be treated in isolation, without considering any of the kind of context that would require network elements to store "state" information. However, in practice an increasing amount of information that is not included in packet headers has been found to be necessary, and a plethora of additional protocols have been created to convey it.

Most transmission is not of an isolated packet but of a flow, i.e. a sequence of packets, for instance to carry continuous media such as audio and video or as part of a session transferring data via a protocol such as TCP, and it is the service experienced by the flow as a whole rather than individual packets that is important. Moreover, in the case of media flows, where packets are sent at regular intervals, resources can be reserved for a flow to ensure its packets arrive at their destination in a timely manner. In IP networks, either additional protocols (such as RSVP) are used to provide the network with the necessary information, or the network can only guess what the application's requirements are. Also, identifying which flow a packet belongs to is not straightforward, typically requiring inspection of five separate fields in the packet header.

Whereas originally applications would identify the entities with which they wished to communicate by IP addresses, which were permanently assigned, now they use other forms such as domain names, requiring another protocol (DNS) to discover the address that needs to go in the packet headers. Most client devices no longer have permanently assigned addresses, but need to acquire one using DHCP. Interfaces also have layer 2 (MAC) addresses, which are permanently assigned; a further protocol, ARP, is used to discover the MAC address associated with each IP address.

Another protocol which is used in session establishment is SIP, which is text-based. While parsing the messages is easy for a laptop or smartphone, for simpler IoT devices it would represent significant effort. Also, the semantics are somewhat untidy, leading to interoperability problems.

9 Efficient forwarding

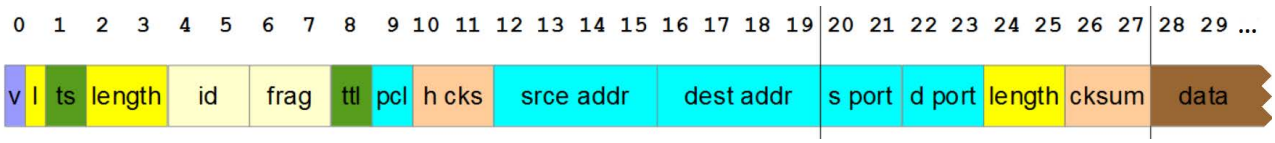

Figure 9.1: UDP/IPv4 packet

Figure 9.1 shows the structure of a packet carrying a UDP datagram over IPv4; the IP header occupies the first 20 octets and the UDP header the next 8. This format was defined in an age when packets carried messages between computers. Now, many packets carry digital audio or video, but they still need to be assembled in memory, for instance the checksum needs to be calculated as the packet is assembled so is the last field to be written, but it is located near the front of the packet.

Similarly, at each network node the packet would be read into a computer's memory and a software routine run to process the packet and forward it on. Contrast this with Ethernet packets, in which the destination address is the first field and the Frame Check Sequence is the last, so that packets can be processed "on the fly" by hardware.

Originally, the destination address was sufficient to define how a packet should be routed. Now, however, time-critical traffic shares the network with less urgent packets, and switches have multiple queues so that each packet can be given the level of service it requires. Most packets are part of a "flow" such as an audio or video stream, or a TCP (or QUIC) session; flows are typically identified using all of the fields coloured blue in Figure 9.1, and routing decisions are made per flow rather than for every packet.

A switch will typically contain a "routing table" with an entry for each flow, in which routing decisions are stored, so that a packet that is part of a flow can be forwarded without needing to be inspected by software. However, the forwarding hardware still needs to inspect all the relevant fields to identify the flow. In the case of flows using Berkeley Sockets or a similar API, this is illustrated in Figure 9.2.

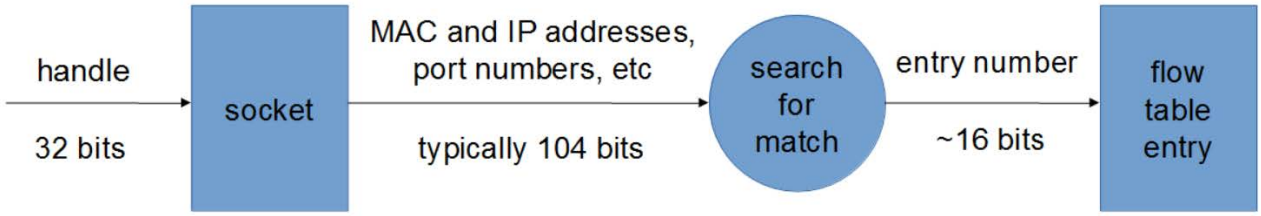

Figure 9.2: Flow identification in current networks

The application requests a socket on which it can exchange information with a remote entity, and is supplied with a "handle" value that identifies the flow. To send a packet, it supplies the handle value and the data; the driver software adds the headers including the information shown in Figure 9.1 and transmits the packet. When the packet arrives at a switch, the switch has to search the table for a match in order to identify the flow.

Much of that work would be eliminated if the packet header identified the flow table entry directly, as in Figure 9.3.

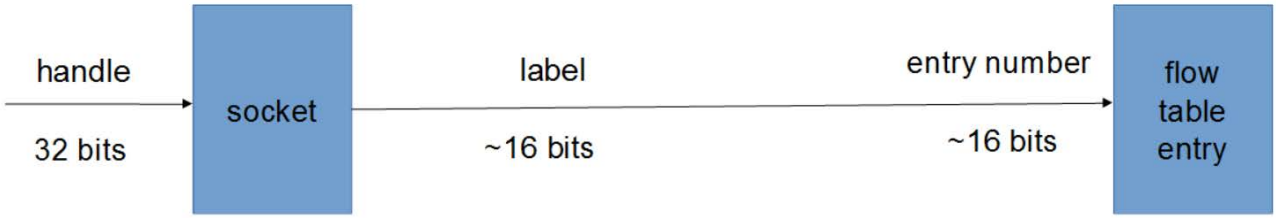

Figure 9.3: More efficient flow identification

10 Migration

New technologies can fail to be adopted because the pain of migrating is greater than the pain of living with the drawbacks of the old technology. For instance, when the world was running out of IPv4 addresses it was easier to introduce NAT than to switch to IPv6. It is instructive to consider where the pain points were in that.

When Internet Protocol was first developed, processing of packet headers was entirely by software; to add support for IPv6 to a switch would simply require a software upgrade, provided of course that there was enough program memory space. Since then, line speeds have increased faster than processor speeds and hardware has become cheaper (as measured by the cost per transistor), so headers are processed by hardware, or at least with hardware assistance, and supporting a new header format requires new hardware, potentially replacing existing hardware which has not reached end-of-life.

Carrying IPv6 packets requires some measure of support from end to end. If a server has an IPv6 address the client needs to send IPv6 packets, and switches and middleboxes need to be able to process them. In an interconnected public system such as the Internet, this requires support from multiple independent organizations, and IPv6 packets cannot be exchanged until all have implemented it. The IPv6 packets can, of course, be encapsulated in IPv4 for carriage over an IPv4 (sub-)network, but this adds complexity to the system and requires knowledge of the IPv4 address to use to reach a particular IPv6 address. However, that may become less of a problem as technologies such as SDN, which decouple routing decisions from the forwarding process, are introduced.

History

Document history		
V1.1.1	March 2021	Publication