Networking 7: IP

i.g.batten@bham.ac.uk

Contents

- IP as a concept
- IPv4 addressing
- IPv6 addressing
- Packets and routing

Why IP?

- Far and away the dominant networking protocol of the past thirty years
- A single network layer, over which all transports can run, and…
- …a single network layer which can run over all available lower layers

What is IP?

- A unreliable, unsequenced datagram service
- You put an address on a packet and the network makes an attempt (“best efforts”) to deliver it somewhere, hopefully the right place.
- There is no checksum covering the data, so even protection from corruption is the responsibility of upper layers
- Hard to imagine a network layer which offers less

Why did IP win?

- Easy to implement, and implemented on all “Computer Science Favourites” of the early 1980s (Multics, TOPS-10, TOPS-20, Lisp Machines, Berkeley Unix)
  - Berkeley Unix, in turn, became the operating system of choice for Unix workstations, the dominant computer science environment of the late 80s and early to mid 1990s.
- Works over everything, from long-haul radio to exotic high-speed fibre.
- Has/had the Support of US DoD, (D)ARPA and NSF
- Actually rather good as well

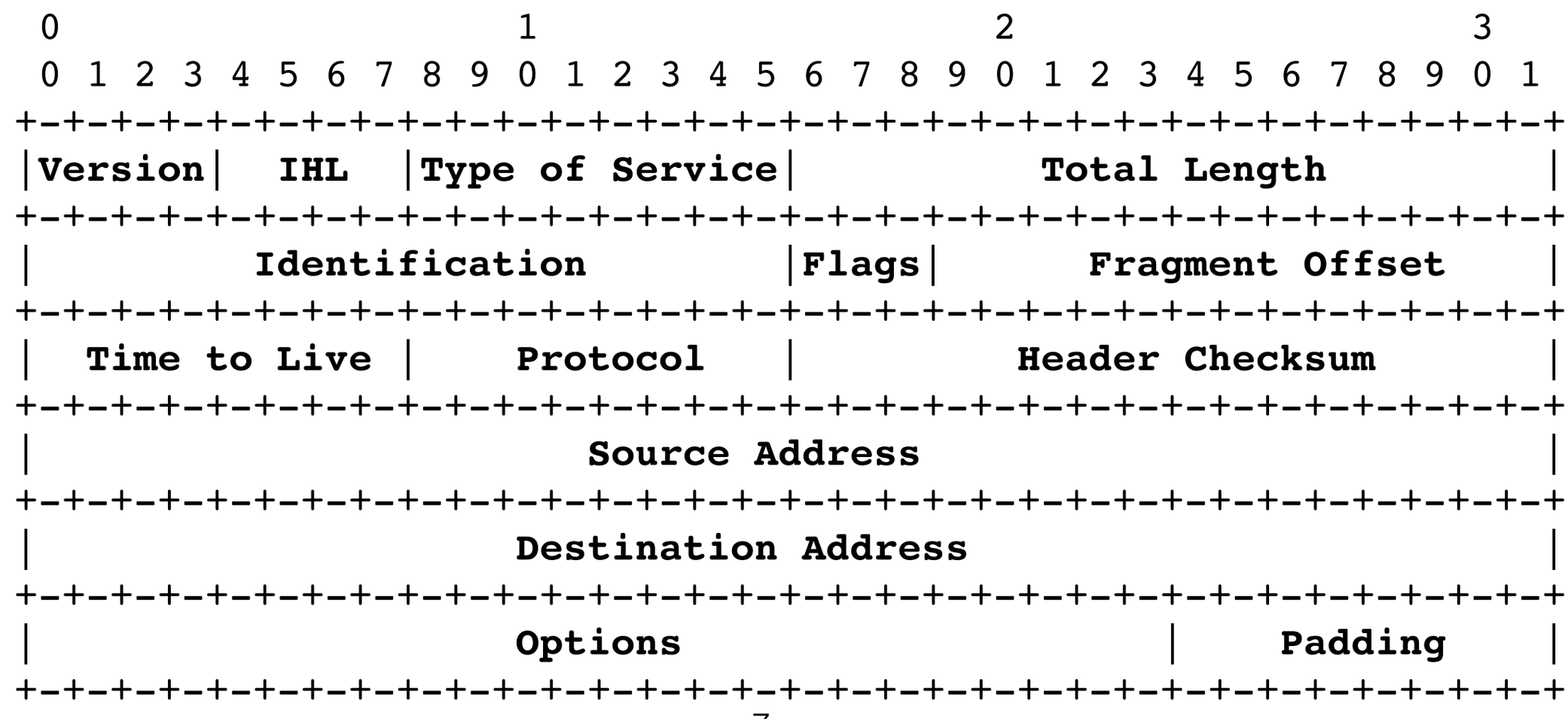
History

- Proposed in 1974 paper [1]
- Experimental versions 0 to 3 described in Internet Experimental Notes (IENs 2, 26, 28, 41, 44, 54)
- IPv4 described in RFC760, January 1981.
- Updated by RFC791, September 1981, which is still current (1349, 6864 and 2472 describe and clarify some little-used extensions).

[1] Vinton G. Cerf, Robert E. Kahn, "A Protocol for Packet Network Intercommunication", IEEE Transactions on Communications, Vol. 22, No. 5, May 1974 pp. 637–648

Packet Format

- Each row is 32 bits, four bytes, 1 **word**. Options are optional, so a typical header is 20 bytes (5 words)
- Note fields aren't byte aligned: this is a 1970s design.



32 bit addresses

- At the time, insanely large. The Arpanet reached peak of 113 nodes by 1983, split in half by Arpanet / MILNet separation.
  - The question is not “how could they be so short-sighted as not to use 48 or 64 bits?” Rather “isn’t it amazing they didn’t use 16 or 24 bits?”
- Original concept of IP was to have 8 bits indicating the site, and 24 bits to specify a machine at that site.
  - Only proposed users were big US universities and companies, US government and a small number of defence-related operations, all in NATO.
- This was seen as wrong very quickly: 256 sites simply not enough, even in the early 1980s.

Notation

- Conventionally written as four decimal numbers, each encoding one byte (wastefully), separated by dots. Printable form 7..15 bytes long (4 times 1..3 digits plus three dots).
- Typically not zero-padded (usually 192.168.1.1 rather than 192.168.001.001), but sometimes people use %03d.
- Hexadecimal would have been much better (you can encode, decode and mask by eye) but hexadecimal wasn’t used much in the 1970s. We’re lucky it wasn’t octal, which was! IPv6 is hex, as we will see.

1001 0011	1011 1100	1100 0000	1111 1010
147	188	192	250
9 3	b c	c 0	f a

Routing Decisions

- All IP addresses identify a network (the leftmost part of the address) and the host on that network (the rest of the address). We will define “leftmost part” and “rest” soon.
- A router ignores the host on network part for all non-local addresses, looks up the network in a routing table, and chooses which interface to send the packet out through.
  - “Default network” handles all other cases.
- A router close to the centre of the inter-network needs to know about most networks.
- And at the time IP was designed, 64 KILO bytes was a LOT of RAM (maximum address space on a pdp11).

Priority: performance on limited hardware, 1980-style

- With the speeds and feeds of the era, building large distributed routing tables was hard. It was essential to keep the number of networks about which information needed to be exchanged to a minimum.
- Limiting it to 256 networks was unreasonable, but there had to be a low limit.
- 32 bits = ~ 4 billion addresses, in a era where a few thousand computers would be seen as an upper bound. Efficient allocation did not matter and was not a design goal: they anticipated utilisation of a fraction of a percent.
- The priority was being able to make routing decisions quickly on realistic hardware (roughly, a DEC LSI 11/23 “fuzzball”: 64kB per process maximum). Those decisions are with us still today.

“Classful” Addresses

- If the address starts with a 0, the first 8 bits identify the network, the remaining 24 bits identify the host on that network. 1.x.y.z through to 127.x.y.z
- Gives 128 “Class A” addresses to large sites that get 2^{24} (~16 million) hosts each. These went to MIT, BBN, Ford, DEC, Boeing, the UK government.
- HP now has net 15 and DEC’s network 16, via its purchase of Compaq who had bought DEC.
- Net 0 not used, net 127 reserved for loopback (almost exclusively 127.0.0.1). Instantly wastes over 16 million address (2^{25}-1).

“Classful” Addresses

- If the address starts with 10, the first 16 bit identify the network, the remaining 16 bits the host on that network. 128.x.y.z through to 191.x.y.z
- Gives 2^{14} (16384) “Class B” addresses with 2^{16} (65536) hosts each.
- Initially easy to get for smaller universities and companies
  - Birmingham had **two**, one applied for centrally, one applied for to use in CS by Bob Hendley and me, not used after 1990, now sold (we didn’t see a penny of it).

“Classful” Addresses

- If the first three bits are 110, the first 24 bits identify the network, the remaining eight the host on the network. 192.x.y.z through 223.x.y.z
- Gives 2^{21} (2097152) “Class C” addresses with 2^{8} (256) hosts per network.

“Classful” Addresses

- Remaining space used for multicast (“Class D”, first four bits 1110, 224.x.y.z through 239.x.y.z) and reserved for experimental use (“Class E”, 1111, 240.x.y.z through 255.x.y.z).
- Multicast never really achieved much outside private networks, so address space that was allocated is wildly excessive.

Classes

	0..7	8..15	16..23	24..31
Class A	Net	Host		
Class B	Net		Host	
Class C	Net			Host

Wasteful Allocation

- Total reachable space is 87% of the available addresses
- But no university, not even MIT, has 16 million hosts, so Class A space very sparsely occupied
- Very few universities or companies have 65536 hosts, so Class B space very sparsely occupied
  - In both cases, even fewer hosts externally accessible
- Class C space was initially available, in bulk, to anyone with an Internet connection (Fulcrum Communications as it then was had 18 Class Cs, 4608 addresses, for <500 employees). This is not untypical.
- Estimates vary, but it's unlikely that today more than 25% of address space is usefully deployed. Huge shortages outside USA, Canada and western Europe (for practical purposes, Cold War era NATO).

But easy for routers

- Most of the early Internet was in fact the holders of the Class A addresses.
- Routers first looked at the address to see if it was “local”: is the destination directly connected to the router via some sort of network connection? If so, send the packet direct to that destination.
- Otherwise, they looked at the first bit of an address. If it was zero, they looked up the first byte in a 128-entry table of “next hops”: the IP numbers of other routers that are directly connected (LAN, WAN) and are believed to be “closer” to the destination.
    - With 32 bit addresses, such a 128 entry table occupies 512 bytes. Easy, even in 1980.
    - The original ARPAnet backbone was “net 10”, later re-purposed as we will see.
- If it’s found, send the packet to that router.
- Otherwise look up remaining two or three bytes in a more complex table (some sort of tree or hash table).
- Otherwise use the default route, if it exists.

Routing vs Memory

- Today you can have a fully populated routing table for every /24 in 128MB (2^{27}) of RAM (ie, a Raspberry Pi or an iPhone can function as a core router)
- There is no current need to do this, but you could have a unique destination for every IP number (all 32 bits) in 32GB of RAM (2^{35}) (a reasonably spec'd desktop PC, a small server)
- Routing now not a big computational problem: you simply index into a sparsely populated array rather than using complex trees and hash tables
- Hardly worth even caching the last eight (or whatever) destinations (a common trick of the past).

Sub-Netting

- People used the large amount of address space to plan their internal networking by structuring the "host" part of the network.
- Users of a Class B could treat their 2^{16} addresses as 256 networks each of 256 hosts. Users of a Class A could treat their 2^{24} addresses as 65536 networks each of 256 hosts, or (with a lot of care and complexity) 256 groups each of 256 networks each of 256 hosts
- Practical limits of the time meant that you didn't want more than ~100 hosts on an Ethernet anyway.
- By late 1980s, you could "subnet" on non-byte boundaries, too
  - Systems that can't subnet, or can't subject on arbitrary boundaries, are now obsolete; the hacks used to work around it are of historical interest only (and vile) — look up "Proxy ARP" if you have a strong stomach.
- Outside world sees one network, internally everyone knows the extra information about the layout of addresses

Netmasks

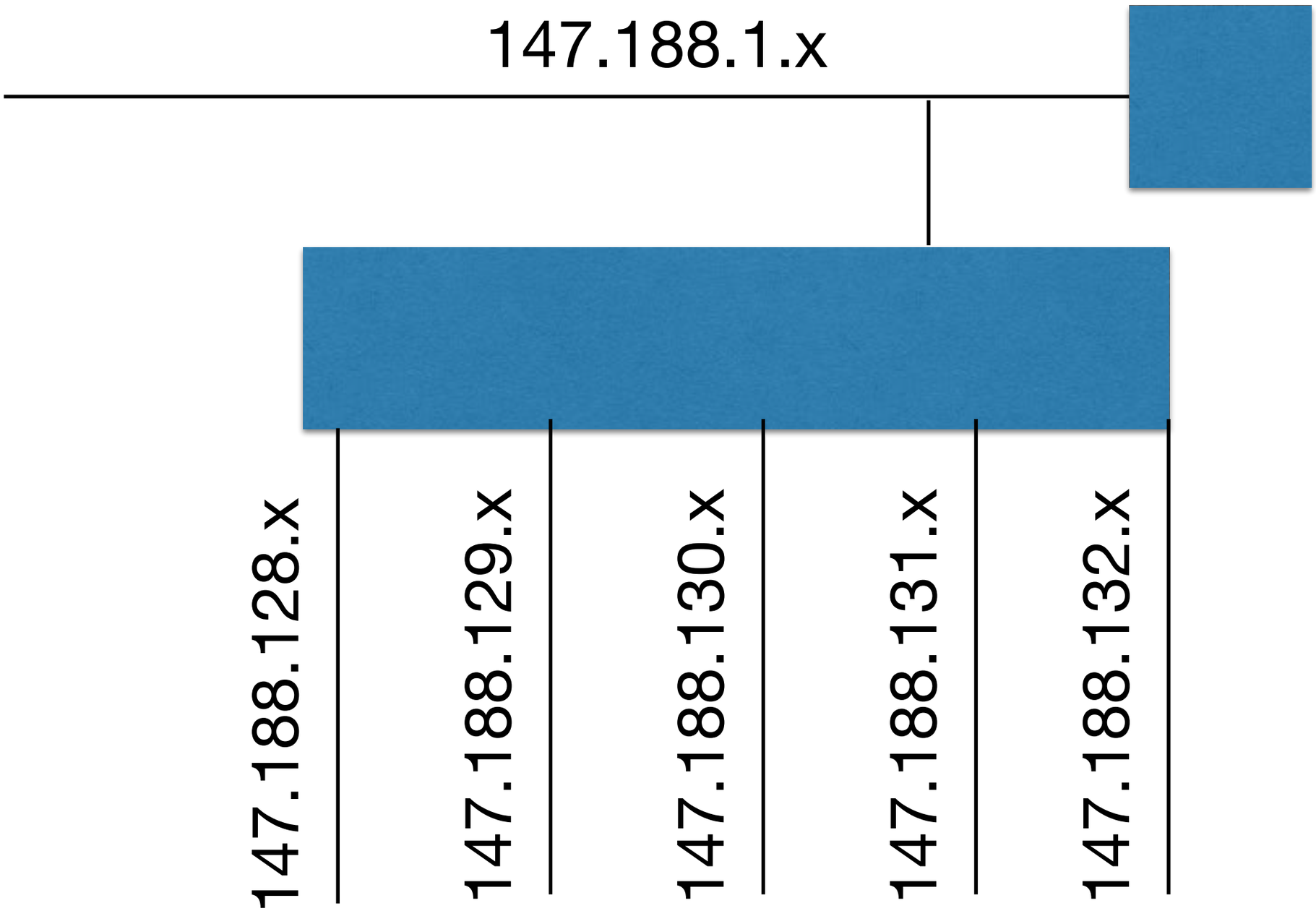
- Originally notated as a netmask: the bit pattern which can be logical-and'd with an address to yield a network number.
- Class A (later /8, as we will see) is 255.0.0.0,
  - 255 = 11111111. 10.1.2.3 & 255.0.0.0 = 10.0.0.0.
- Class B (/16) is 255.255.0.0
- Class C (/24) is 255.255.255.0
- A Class C used as 32 networks each of 8 hosts is 255.255.255.248 (/29): 248 = 11111000.

Subnets

	0..7	8..15	16..23	24..31
Class A	Net	Host		
Class B	Net		Host	
Class C	Net			Host
B, 255.255.255.0	Net		**Subnet**	Host
A, 255.255.255.0	Net	**Subnet**		Host
C, 255.255.255.248	Net			5 bit subnet

3 bit host

Sub-netting


147.188.1.x
147.188.128.x
147.188.129.x
147.188.130.x
147.188.131.x
147.188.132.x

CIDR and Slash Notation

- Problems of waste with Classful networking and need for more flexible sub-netting combined to produce Classless Interdomain Routing, CIDR.
- Every network address has a “netmask” which describes how much of it is network and how much of it is host.
- Class A is now a “slash eight” (MIT is 18.0.0.0/8), Class B is now a “slash sixteen” (Bham is 147.188.0.0/16) and Class C is a “slash twenty four” (FTEL is, amongst others, 192.65.220.0/24).

And in the other direction…

- To make routing tables more compact, a group of eight contiguous Class Cs can be placed together under a /21, or sixteen contiguous Class Bs under a /12.
- This is called “super netting”, but is now less useful as routing table size is not an issue.

Non-Byte Masks

- This extends trivially to boundaries which are not classful.
- So an ISP wanting to give prosumers some IP numbers can hand out a /28, which gives the user 16 IP numbers.
  - My home network is 81.187.150.208/28, giving me 81.187.150.208 through to 81.187.150.223, 14 hosts and two different broadcast addresses.

Recovering the Class As

- Some of the Class As were recovered, by a variety of threats, cajoling and bankruptcy (and Stanford being cool and stand-up and giving it back voluntarily in exchange for a bunch of Class Bs)
- Widespread allocation had stopped at Net 57 anyway (Societe Internationale de Telecommunications Aeronautiques S.C.R.L.)
- Remainder and returned networks were broken up and issued in varying size allocations
- http://www.iana.org/assignments/ipv4-address-space/ipv4-address-space.xhtml

RFC1918

- Class A Network 10 became free when the Arpanet backbone was closed down.
- So 10.0.0.0/8, along with the available 172.16.0.0/12 (172.16.x.y through to 172.31.x.y) and 192.168.0.0/16 (192.168.x.y) were allocated for private use. These **must not** be routed outside private domains.
- This usually requires “address translation”, which we will cover later.

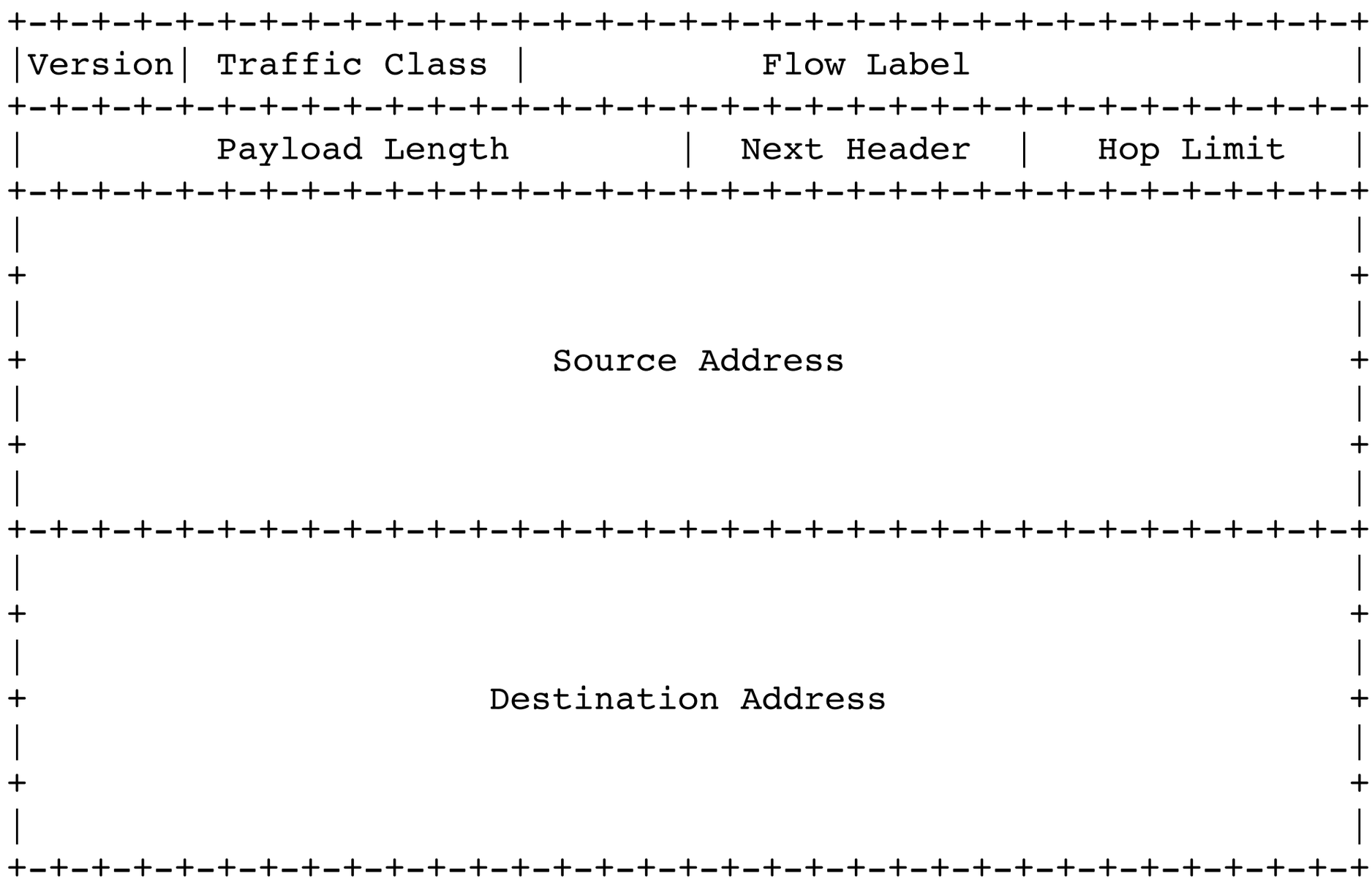
IPv6: IP with big addresses

- 32 bit addresses have run out (last /8s allocated in February 2011).
- 128 bit addresses gives 2^{96} more addresses.
- Population of planet is $<2^{34}$ (16 billion) and likely to remain so
- 2^{34} people, 2^{128} addresses, 2^{94} addresses per person.
- 2^{94} = 19807040628566084398385987584 (~ 2 x 10^{28})

In reality, more like 2^{64}

- Minimum allocation unit is a /64; even your mobile phone will probably get a /64. IPv6 is “really” a 64 bit protocol.
- Intention is that with 64 bits available after the routed “prefix”, allocation of addresses on local networks is much easier
- 2^{64} still gives 2^{30} addresses each (~1bn per person).
  - At the moment, only 2^{61} addresses available to allocate, so 2^{27} each (~120million per person)

IPv6


Version
Traffic Class
Flow Label
Payload Length
Next Header
Hop Limit
Source Address
Destination Address

IPv6 Addresses

- Standard format is a hex string, broken into 16 bit chunks with colons, leading zeros suppressed
  - 2001:8b0:129f:a90f:60c:ceff:fedd:f68
- “::” means “as many zeros as fit here”
  - 2001:8b0:129f:a90f:: = 2001:8b0:129f:a90f:0:0:0:0

IPv6 reservations

- ::/128 “uninitialised”, ::1/128 “loopback” (note only one address, not 16 million!)
- ::ffff/96 IPv4 mapping (ie ::ffff:1.2.3.4)
- fc00::/7 Private address space
- fe80::/8 link local,
- 2001:db8::/32 documentation etc (and a few others for similar purposes)
- 2000::/3 allocated as single block for normal use.
  - 2^{61} address blocks for end-users.

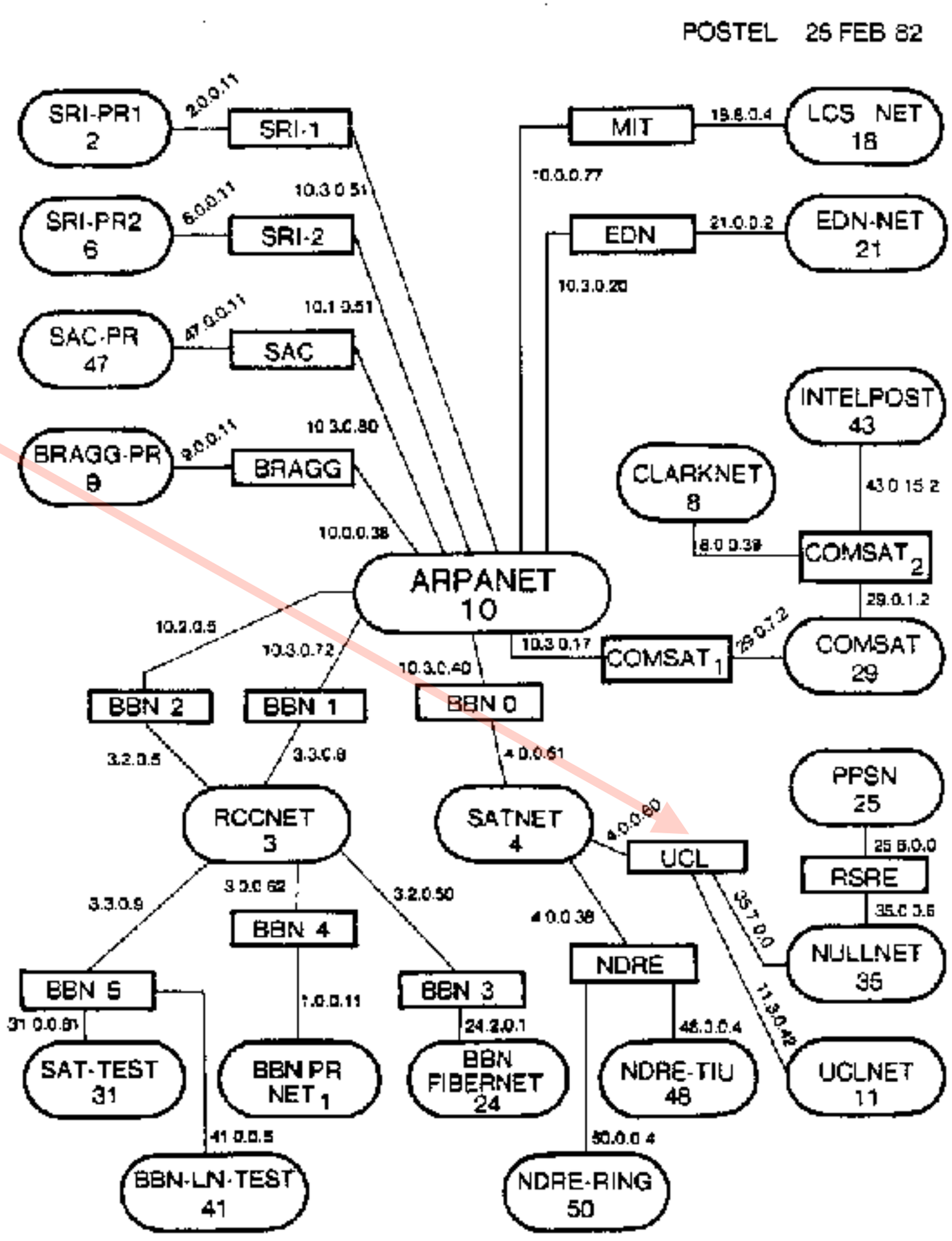
IPv6 Routing Tables

- Will in the long-term require the complexity originally used for IPv4
- At the moment, very sparse (sadly)
- 1TB is 2^40 bytes, so 2^64 is 2^24 TB, which isn't going to happen any time soon. But nor is 2^64 allocated networks.

IP in operation

- Sender: choose an interface we believe to be closer to the destination, and send the packet
- Recipient: if the packet is for us, process locally.
  - Otherwise, send it on.
- We will cover routing in more detail later, so we are just trying to get the general flavour here

- Consider packet at router at UCL
- Network 11? It's local, send via local ethernet (or Cambridge Ring, I seem to recall).
- Network 35 or 25? Send via Nullnet, whatever that might be.
- Otherwise, send via Satnet to main Arpanet.



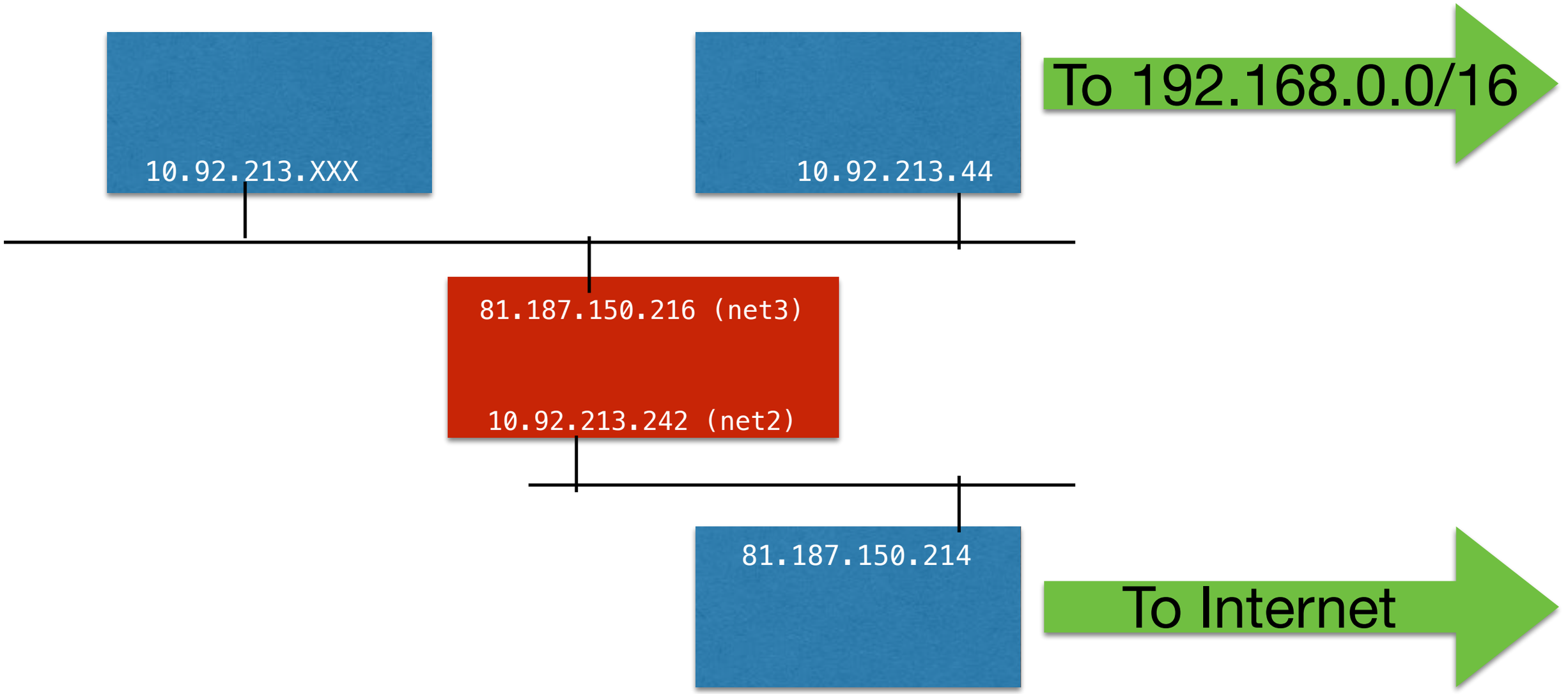
Simple Example

```
igb@ossec-sol:~$ netstat -nrv

IRE Table: IPv4
  Destination             Mask           Gateway           Device  MTU  Ref Flg  Out  In/Fwd
------------------- --------------- ------------------- ------ ----- --- --- ----- ------
default               0.0.0.0         81.187.150.214                0   2 UG     8426       0
10.92.213.0           255.255.255.0   10.92.213.242       net2    1500   6 U    757664       0
81.187.150.208        255.255.255.240 81.187.150.216      net3    1500   5 U      1854       0
127.0.0.1             255.255.255.255 127.0.0.1           lo0     8232   2 UH        0       0
192.168.0.0           255.255.0.0     10.92.213.44                  0   1 UG        0       0
```

Machine with two network interfaces
Note net2 is RFC1918, probably “internal”, and net3 is not, probably “external”
192.168/16 and default are additional routes

In Pictures


10.92.213.XXX
10.92.213.44
To 192.168.0.0/16
81.187.150.216 (net3)
10.92.213.242 (net2)
81.187.150.214
To Internet

Routing Decisions

- Decrement the TTL (and do so each time you've held on to the packet for a second, not that that happens).
- Look at each destination we know about, starting with the longest mask and working to the shortest
- Here we have locally connected ethernets (net2 and net 3)
- Traffic to 10.92.213.0/24 and 81.187.150.208.28 is "local" and goes direct over ethernet
- Traffic to 192.168.0.0 is sent over the ethernet to 10.92.213.44 (a **gateway**)
- Traffic to anywhere else is sent over the ethernet to 81.187.150.214 (again, a gateway)
- Machine might itself be a router: whether it forwards packets that arrive on one interface with addresses on the other side is a policy decision.

IP on Ethernet

- On an ethernet, or other “point to multi-point” network, how do we find the MAC address of the next hop?
- IPv4 uses ARP: Address Resolution Protocol
  - Simple, old and frighteningly insecure
  - Ask “WHO HAS” a particular IP number
  - Station with that IP number, or someone who claims to know about it, tells us. Ripe for exploitation, as we will learn in network security lectures
- IPv6 uses “Neighbo(u)r Discovery Protocol” to do similar job, with ICMPv6 messages 135 (solicitation) and 136 (advertisement)

IP on point-to-point links

- If a link is point to point, you just send the packet down it.
- Might be a physical point-to-point link (a serial line of some sort, perhaps) or might be a tunnel (packets encapsulated in other packets). We will talk about tunnels later.
- “The internet is a large open field, with many tunnels running underneath it”.

IP to Gateways

- When sending a packet to a gateway, you look up the gateway address using ARP if necessary, but the IP destination remains the ultimate destination. Gateways don’t appear in the IP header.
- Gateways at other end of point to point links just involve sending the packet.

Hop Counts

- Each packet has a Time To Live (TTL).
- Decremented each time the packet is processed, or whenever a router holds on to it for more than one second (rare today).
- When it hits zero, packet is discarded and an error report (“ICMP Time Exceeded”) is generated.
- Typical initial value 30–60, depending on current estimates of “diameter” of internet
- Prevents packets circulating endlessly in case of routing loops or similar.
- Requires re-computation of the header checksum in IPv4
  - Fast algorithms used to recompute it based on knowing what has changed: this isn’t a secure hash
- IPv6 doesn’t have a header checksum, relying instead on lower layers being reliable and upper layers being sensible
  - Good reason for routers and switches to have ECC RAM. 1 flipped bit in 10^{12} = 1 silently broken packet per second at 1Tb/sec.