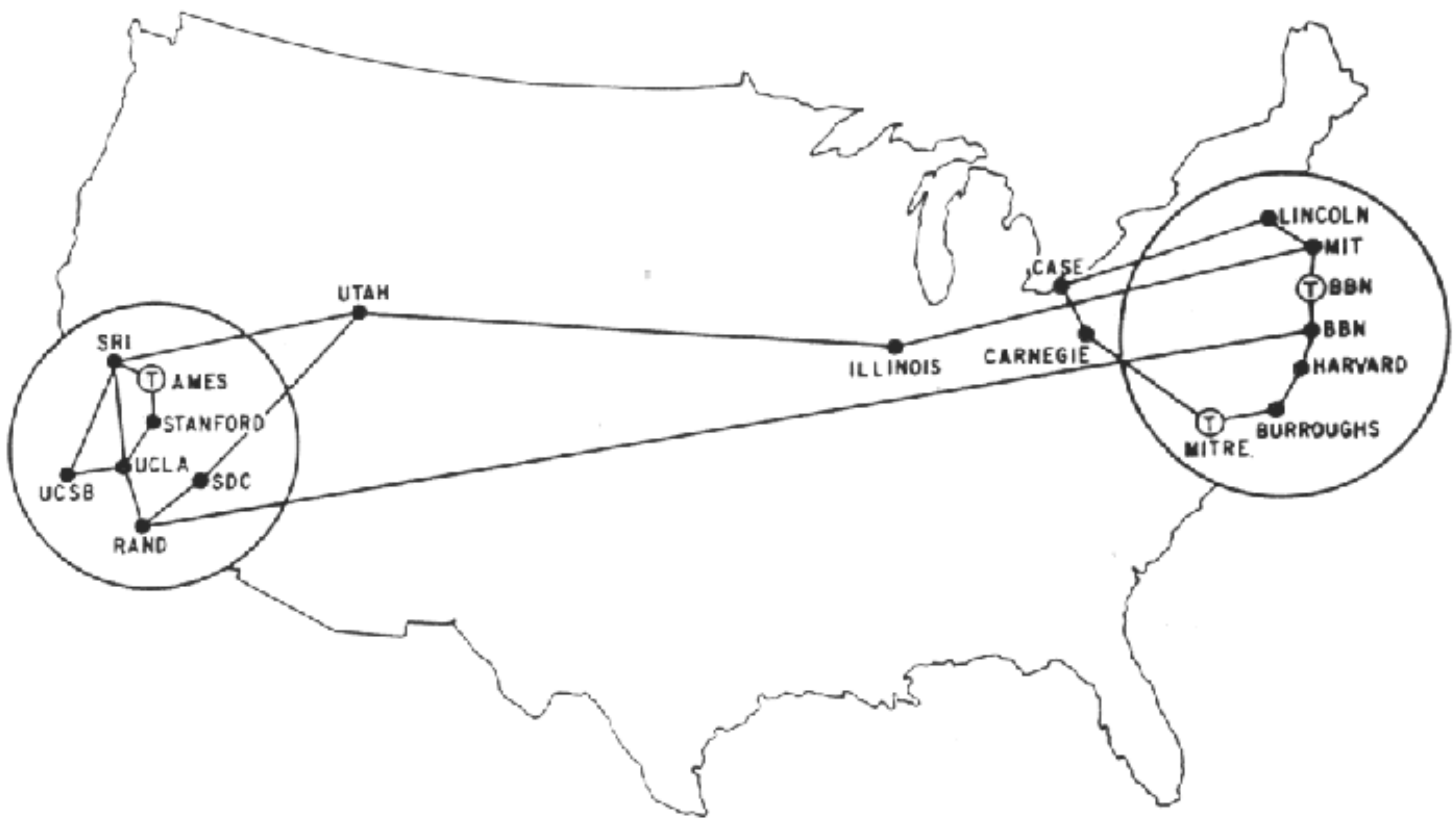
Networking: Lower Layers

i.g.batten@bham.ac.uk

Check Panopto!

- Is it running?
- Is it running?
- Seriously, is it running?

ARPANet, Sep 1971



MAP 4 September 1971

LANs, MANs, WANs…

- Division of networking space into **L**ocal (buildings, campuses), sometimes **M**etropolitan (city) and then **W**ide **A**rea **N**etworks. LANs, MANs, WANs.
- Also (not this course) sometimes **P**ersonal **A**rea **N**etworks, PANs, mostly today Bluetooth.
- Theoretically, different technical solutions to different engineering problems.
- Rapidly converging.

LANs and WANs

- Historically, wide-area networks were very slow.
  - a fast long-haul network of the 1980s might be a few tens of kilobits per second
  - Technology existed to go faster, issue was mostly cost and practical availability.
- Even though local-area networks were hitting 10 Mbps by 1980 and 100 Mbps by 1990.
  - Although many were much slower (X.25, RS232 with Kermit, etc, etc)

LANs and WANs

- Historically, Local Area and Wide Area networks were different in technology, purpose and protocols.

- In Europe and to an extent the US, telco monopolies limited what WANs could do.

- And outside a small number of research environments, LANs were almost entirely proprietary: interworking was mostly with WAN technology.

WANs

- Used to connect computers together, between buildings more than (say) 1km apart, or sometimes just when crossing a road if the local laws grant a monopoly to telcos.
- Historically there were three main applications:
  - File transfer (lots of problems of format conversion, as even byte-size varied)
  - Job transfer (for use of national facilities for super computers; batch mode)
  - Remote login (when you interactive access to remote systems,m which was not always available).
- UUCP very influential and STILL SHIPPED ON MAC!
- ARPANet in the US restricted only to people with government contracts

WAN Technology

- The key point about the WAN is that for most of its history it is slow.
- Very slow.
- UofB JANET connection 1985: 64Kbps.
  cs.bham.ac.uk JANET connection 1987: 9.6Kbps.
  US/UK ARPAnet connection 1986: 2.4Kbps (yes, seriously).
- ARPA/NSFNet backbone 1987: 64Kbps
- 2Mbps links emerge (for most of this) by about 1990.

WAN Technology

- This means that efficiency is very important: wasting tens of bytes is a significant performance problem
- So if you are going to use the same protocols on WAN and LAN, the protocols in use on the LAN has to consider working over slow-speed, lossy links as well as fast networks inside buildings.
- The crucial issue here is when developed, the LANs were much faster than the WANs; today, that is in many cases precisely reversed.

Packets and Circuits

- Real circuits involve electrical connections from end to end
- Packet switching involves putting an address on packets and sending them to the destination individually
- Each packet can contain full destination information, or can be associated instead with a virtual circuit.
- Virtual circuits make objects that look like circuits out of a stream of packets
- Assumption: network is a mesh of routers (switches) linked by some sort of medium.

Packet Switching

- Each packet has addressing information
- A router looks at incoming packets, decides where to send it, sends it on its way
- Router “complexity” scales by the number of packets processed, and possibly other things.

Connection Orientated / Virtual Circuits

- If your underlying network supports virtual circuits, you can ask the network to send a stream of packets to a specific destination
- The network decides a route, tells all the routers along the way what is happening, and give you some sort of token to identify the flow (“virtual circuit”)
- You then send data with that token attached, and it arrives at the other end complete and in order. The assumption is that there are fewer connections than there are endpoints, so this token is smaller and easier to look up.
- Upside: network is doing a lot of the heavy lifting of ensuring all the data gets there, so your network stack is simplified.
- Downside: the routers are much more complex, as they scale by bandwidth **and** number of circuits **and** rate of circuit creation / destruction.
- Historically, X.25 and (to a lesser extent) Frame Relay. Today, ATM (on its way out, but still present in many networks), and Multi Protocol Label Switching (MPLS) (core of big provider networks).

Connectionless / Datagram

- Underlying network offers packet switching “in the raw”
- Individual packets are processed by the network and may or may not arrive, and may or may not be damaged. Network is “best efforts”, no guarantees.
- User addresses packets with complete destination information on each and every packet.
- All responsibility for more complex services rests with the end-points: the network is not going to help (although obviously the less damage it does to packets the better)
- This is the basic requirement to put IP over
- You can use a virtual circuit as a link in a connectionless network, but not vice versa.

Lower Layer Technology

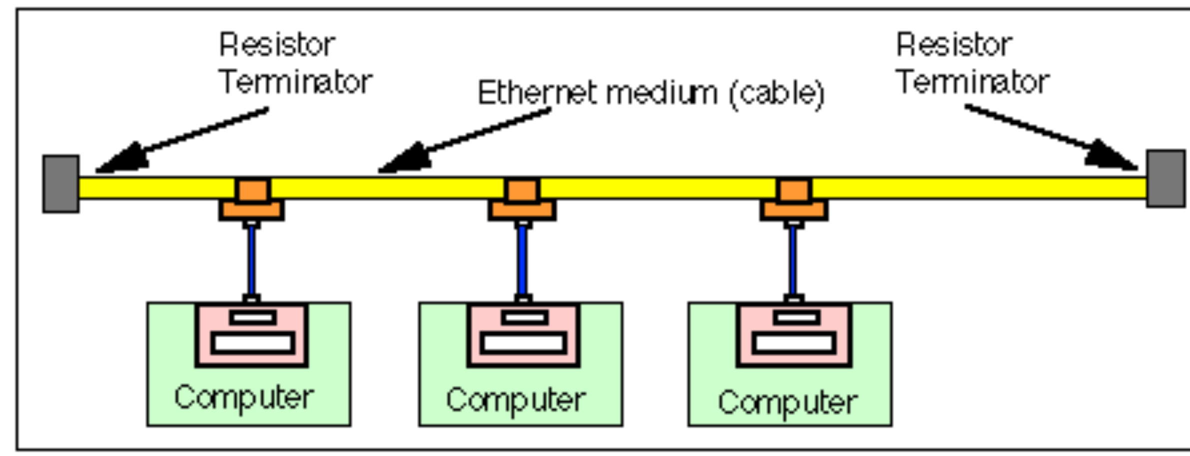
- Ethernet
- Token Ring (IBM Token Ring, FDDI)
- Slotted Ring
- ATM
- MPLS
- SDH
- DWM

Ethernet

- Developed by Metcalfe and Boggs at Xerox Palo Alto in the 1970s.
- Named after the luminiferous aether that supposedly carried light and radio until disproved by the Michaelson-Morley experiment
- Takes inspiration from earlier radio packet networks, notably AlohaNet in Hawaii.

Topology

- The topology of Ethernet was originally a bus: a single cable with computers connected to it.
- (Early versions are 3Mbps, but for practical purposes "yellow hose" is always 10Mbps).
- Maximum length is 500m (both for reasons of resistance and timing as we will see); can be amplified and regenerated to go 1500m max.



Format

- 7 bytes of **preamble** (0x55) to allow receivers to synchronise.
- 1 byte **start of frame delimiter** (0x5d)
- 6 byte **source address** (48 bits)
- 6 byte **destination address**
- 4 byte **VLAN tag** (optional)
  - First two bytes 0x8100 to keep older equipment happy
- 2 byte **type** or **length**
  - If ≤ 1500: length. If ≥ 1536: type, with length found by looking for end of the packet
- 42—1500 bytes of **payload**
- 4 byte **CRC**
- 12 byte-time **inter-packet gap**.

Finding the end without a length

- Checksum is computed continuously, so when you have a set of bytes where the last four bytes are the correct checksum for the whole packet, you know you have reached the end.
- CRC calculation of (data + CRC) generates the magic number 0xC704DD7B - google this number for the gory GF(2) details.
- Or wait until the inter-packet gap
- Or both

Basic Logic

- Only one station can talk effectively at a time, as every station can see what every other station is saying and multiple transmitters will interfere.
- Each station waits until no-one else is talking, and then start transmitting.
- What could possibly go wrong?

Collisions

- Ethernet is formally known as “**CSMA/CD**” — Carrier Sense Multiple Access Collision Detection.
- The magic comes from what happens when there is a collision.

Collision Detection

- As a station transmits, it also listens to the ether and checks the ether only contains the signals that are being sent
    - this has to be done in hardware, as it is mostly an analogue problem.
- If there is a mismatch, someone else is transmitting at the same time.

When Collisions Happen

- First action is to “jam” the network: send a set pattern so everyone knows a collision is in progress.
- Critical that the whole ether knows about the collision before the packet has finished being sent
  - Imposes a minimum packet size (64 octets), which is a function of the maximum diameter of a collision domain (1500m). Jam pattern pads packets to this length at least.

Recovery from Collision

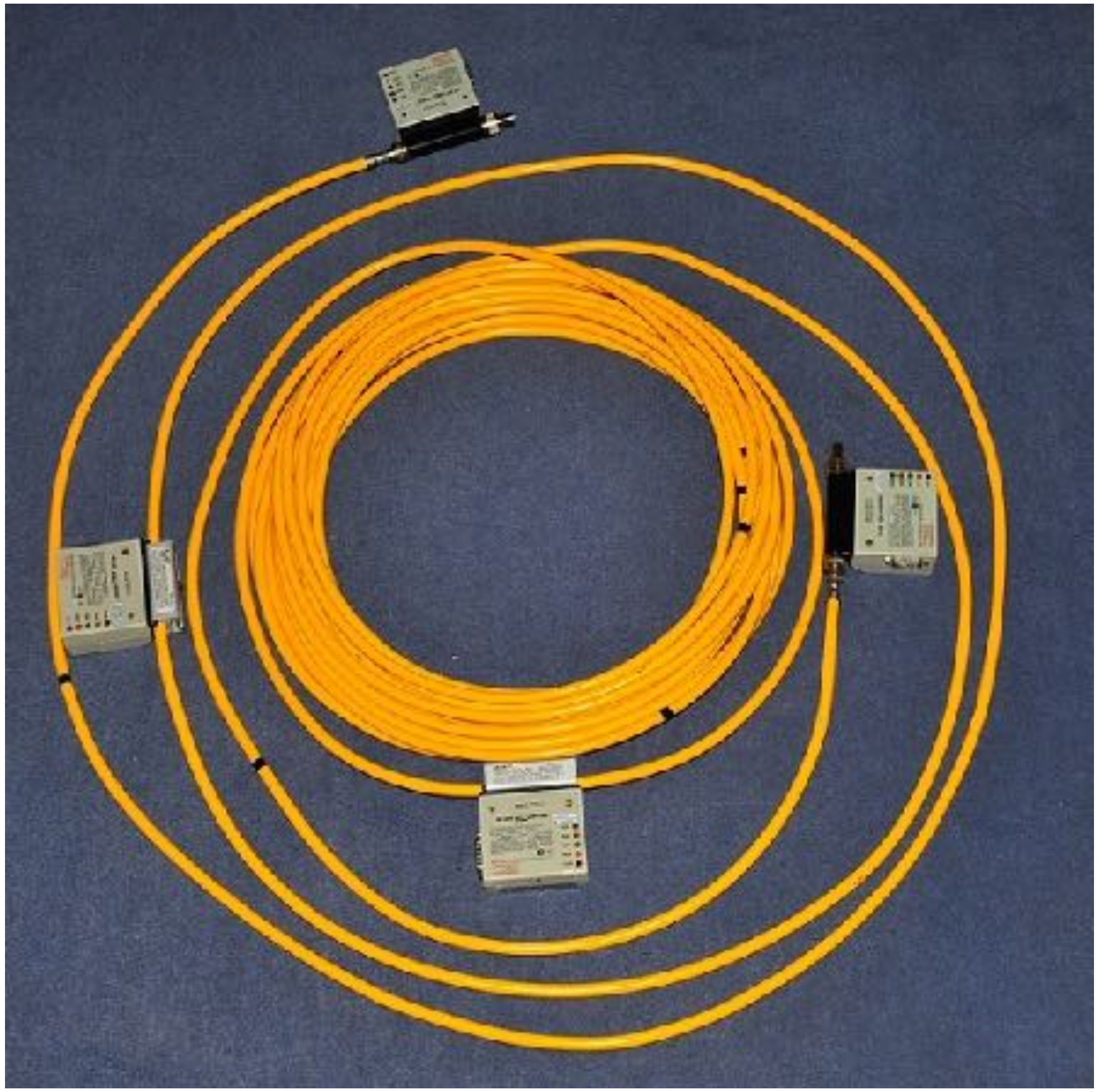
- On the first attempt, choose a random number k from {0,1} and delay k x 512 bit periods before trying again.
- More generally, on the nth attempt, choose a random number k from $\{0..2^n\}$ and delay k x 512 bit periods before trying again.
- After 10 attempts, give up.
- Randoms come from things like serial numbers; they don’t need to be very good quality.

Problems

- Collisions increase non-linearly with load, and the precise curve depends on the exact traffic mix
    - “Ethernet capture effect”
- Latency for a single packet is unpredictable, because some number of collisions may delay it.
    - This can be overstated by advocates of other protocols.

Sizes

- Maximum frame size 1500 bytes payload plus 22 packets of header (larger ends up slowing down stations wanting to exchange small packets)
- Minimum frame size 64 bytes (slightly wasteful for, say, telnet, but making it smaller reduces maximum diameter of network)
- Maximum “diameter” 1500m (from complex rules surrounding number of permissible repeaters).
- 500m and 10Mbps gives name: **10Base5**.

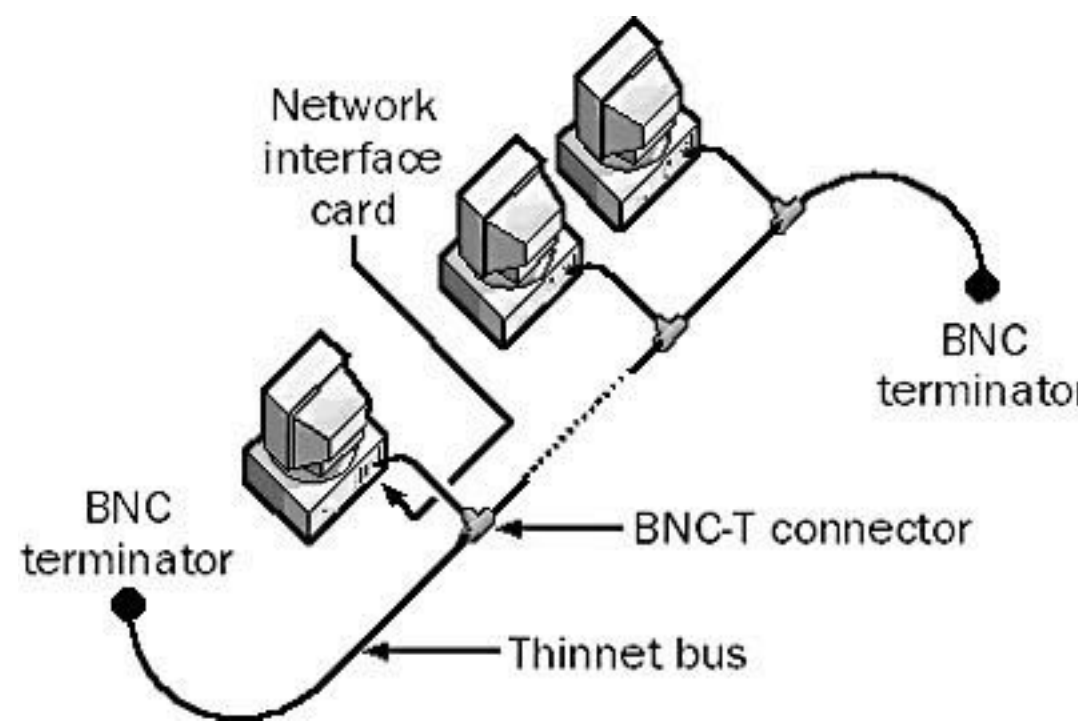


Problems

- Cable is heavy, expensive and difficult to install (tight, or more to the point loose, minimum bend radius requirements).
- Installing taps for transceivers involves drills, and risks damaging the cable.
- Need for transceivers adds cost and complexity.
- Performance issues lurking in the background

An interim: 10base2

- Instead of using thick co-ax, use thin coax. Higher resistance, so limited to 185m: **10base2**.
- Instead of using transceivers, simply bring the coax to the computer and attach it with a tee-piece.
- Cut the cable, rather than drilling into it.



10Base2

- Otherwise it works much the same
  - much smaller maximum diameter of <600m
  - Different terminators
- Can be mixed electrically and logically with 10Base5 (rules are complex and only of historical interest)
- Probably the dominant networking of the 1980s and early 1990s: older buildings still full of it.

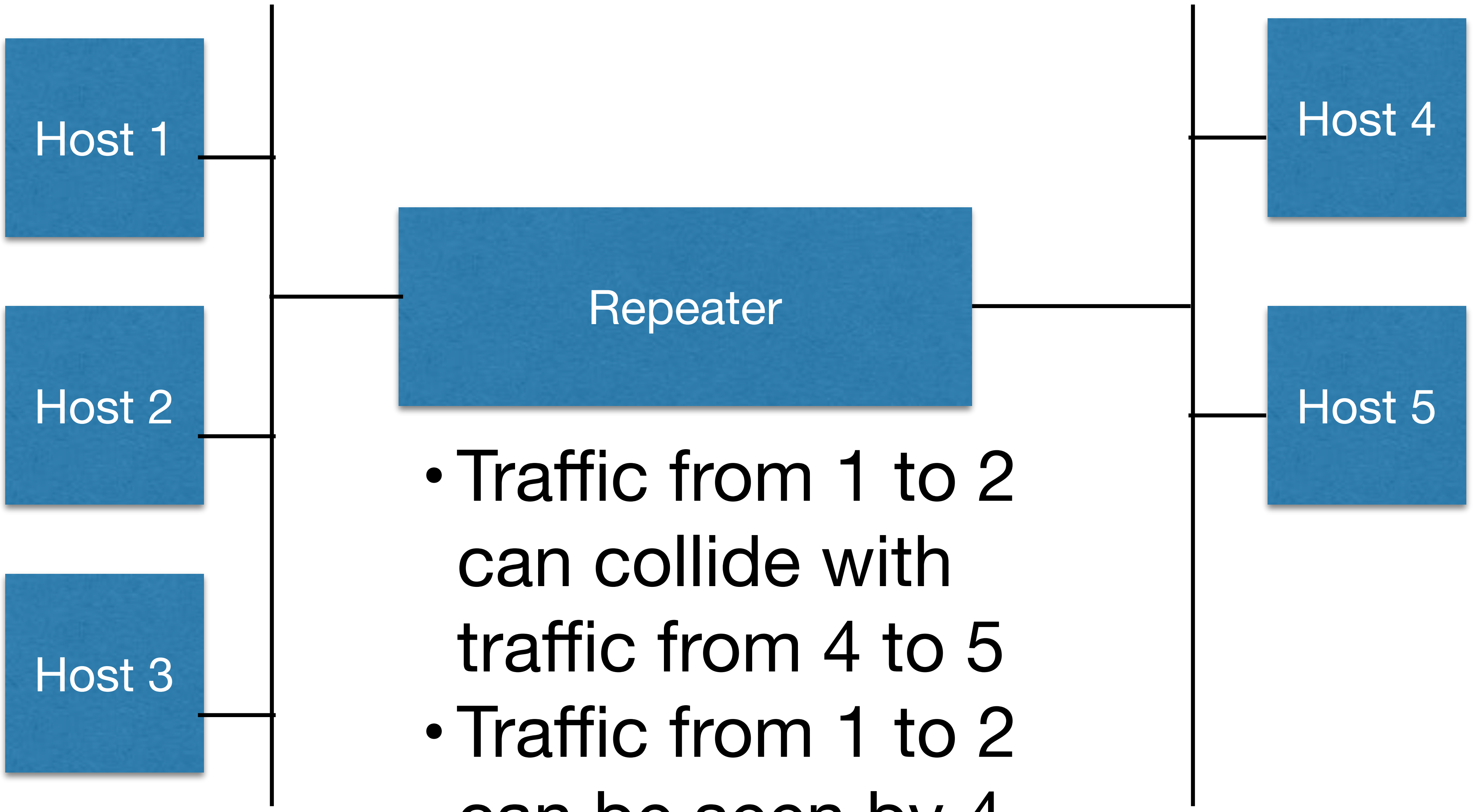
10BaseT

- Coax cable still a pain: expensive, awkward to install, easily damaged.
- 10BaseT looks like “modern” ethernet: Up to 95m of twisted pair (four conductors in two pairs) using RJ45 connectors to a **hub**. Originally “Category 3” cabling, basically voice.

Hubs, Repeaters, etc

- A repeater is just an amplifier: collisions are seen on both sides
- A bridge receives, buffers and transmits frames, so collisions are not propagated
  - “learning” or “filtering” bridges only send frames that belong on the other side; stupid bridges just propagate everything.
- Ether hubs are **repeaters**, not bridges. There are collisions when two stations talk.

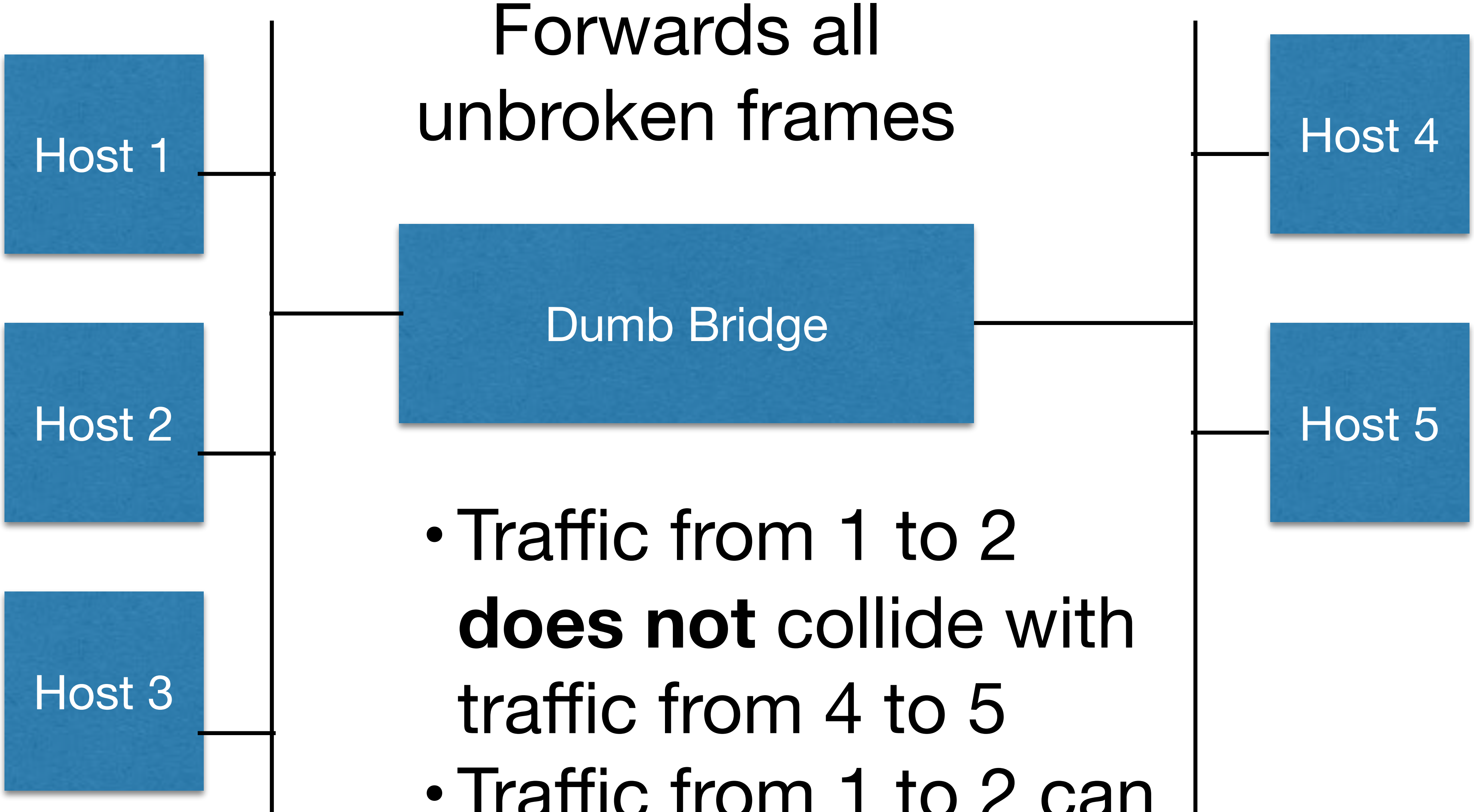
Repeater



- Traffic from 1 to 2 can collide with traffic from 4 to 5
- Traffic from 1 to 2 can be seen by 4 and 5

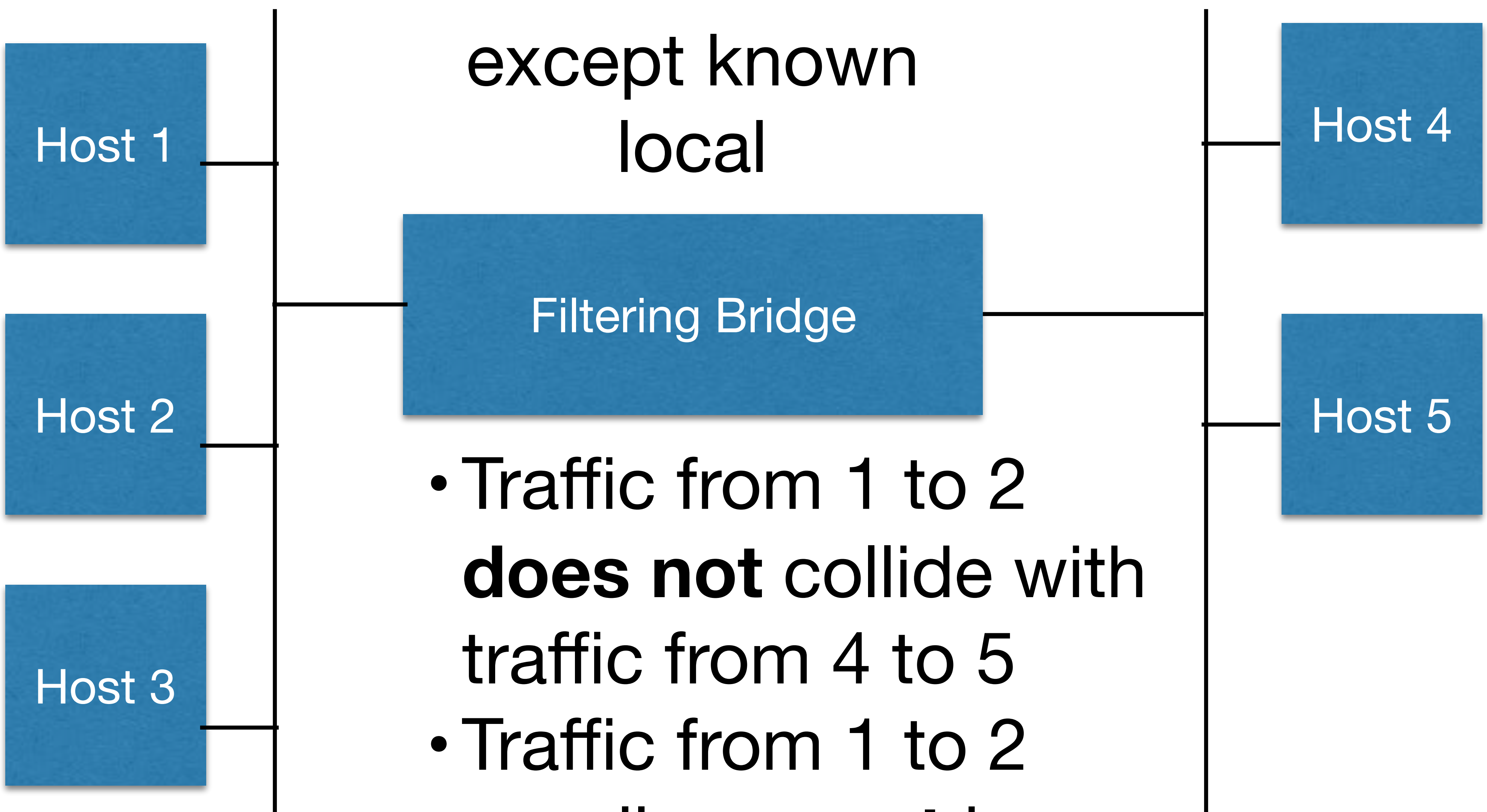
Dumb Bridge

Forwards all unbroken frames



- Traffic from 1 to 2 **does not** collide with traffic from 4 to 5
- Traffic from 1 to 2 can be seen by 4 and 5

Filtering/Learning Bridge



Forwards all unbroken frames except known local

- Traffic from 1 to 2 **does not** collide with traffic from 4 to 5
- Traffic from 1 to 2 usually **cannot** be seen by 4 and 5

Faster and Faster

- 10BaseT is no faster than 10Base2, but cheaper and more flexible to install.
- 100BaseT raised the speed, but still had potential for collisions
- Full duplex and switching made 100BaseT much faster, following by 1000BaseT (GigE) and then 10GigE, 40GigE and the nascent 100GigE.
- Technology similar, but stricter wiring rules (“Cat5” for 100BaseT, “Cat5e” or “Cat6” for faster).

Ethernet Switches

- A switch is a set of learning bridges in a box.
- Each interface is its own collision domain.
- Packets to unknown destinations are sent out of all ports, otherwise only traffic for devices plugged in to the port is sent.
- “Full Duplex” means traffic goes in and out without colliding as **each direction** is a separate collision domain.
- Large buffers internally deal with congestion.
- Result: no collisions.

Cut-Through Switches

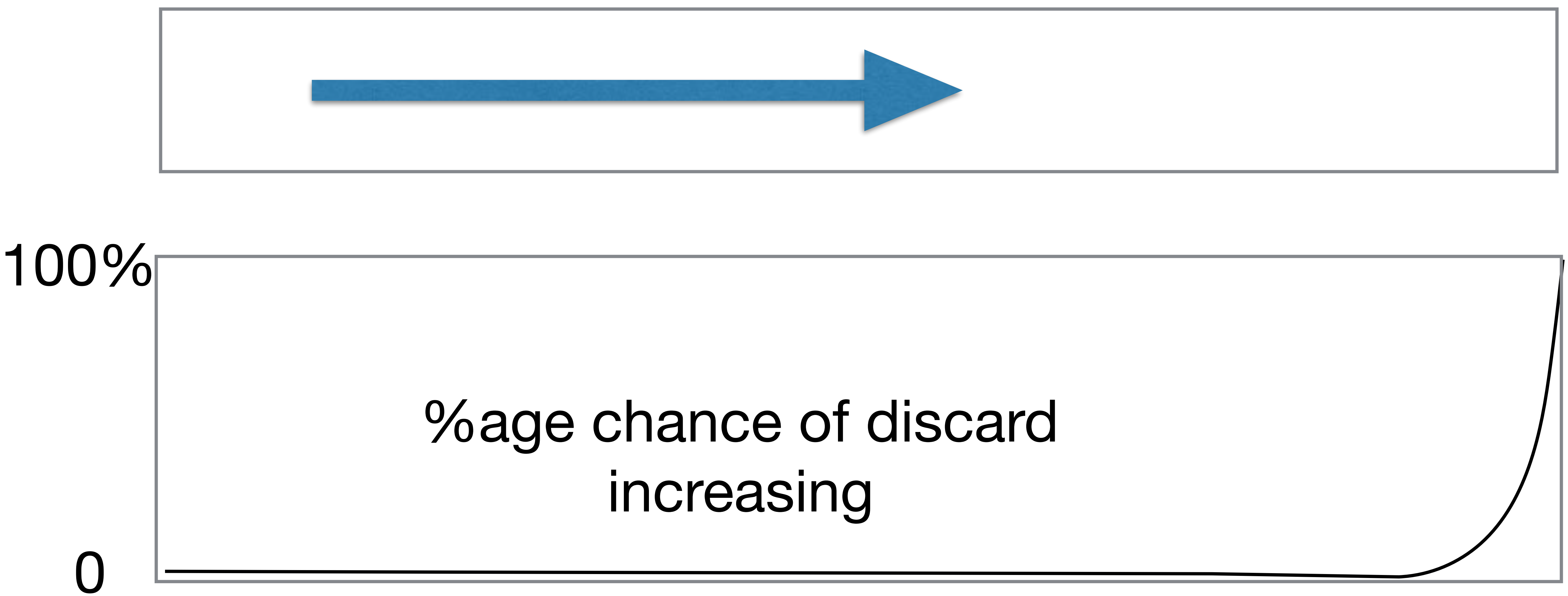
- Conservative switches accept frames in their entirety, check the checksum, then transmit them to other interfaces
  - Introduces additional latency compared to a straight piece of wire (For GigE, 1 bit period is 1ns, full packet is 1500*8ns = 12µs, equivalent to ~3.6km of copper; for 10BaseT it's 1.2ms, or 360km of copper).
- Aggressive switches look at the header, and immediately start transmitting on the correct interface ("cut through).
  - Latency is just the 160–192 bits of the header, so <2% of a full packet: ~60m of copper for GigE, 6km for 10BaseT.
- This propagates broken frames if there are any to be propagated, as it can't check the checksum

Random Early Drop

- Naively, when a buffer fills up, you start to drop packets as you can't put them anywhere
- We will come on to transport connections in detail, but in general, packet loss results in a timeout followed by a retransmission, which net slows things down after some interval
- A new strategy is to randomly drop packets with a probability which increases as the buffer fills, so the dropping starts earlier but more gently, hopefully reducing speed before real loss starts to happen.
- The loss of packets is seen by the sender when the acknowledgements stop, and is a signal to the sender to slow down. You hope.

Random Early Drop

Buffer filling…



Token Ring/Bus

- Ethernet was argued to behave badly under high load, although limited evidence was available.
- Token Rings and Token Buses pass a “token” from station to station.
- The station that holds the token can transmit, and then passes the token on when it has finished.

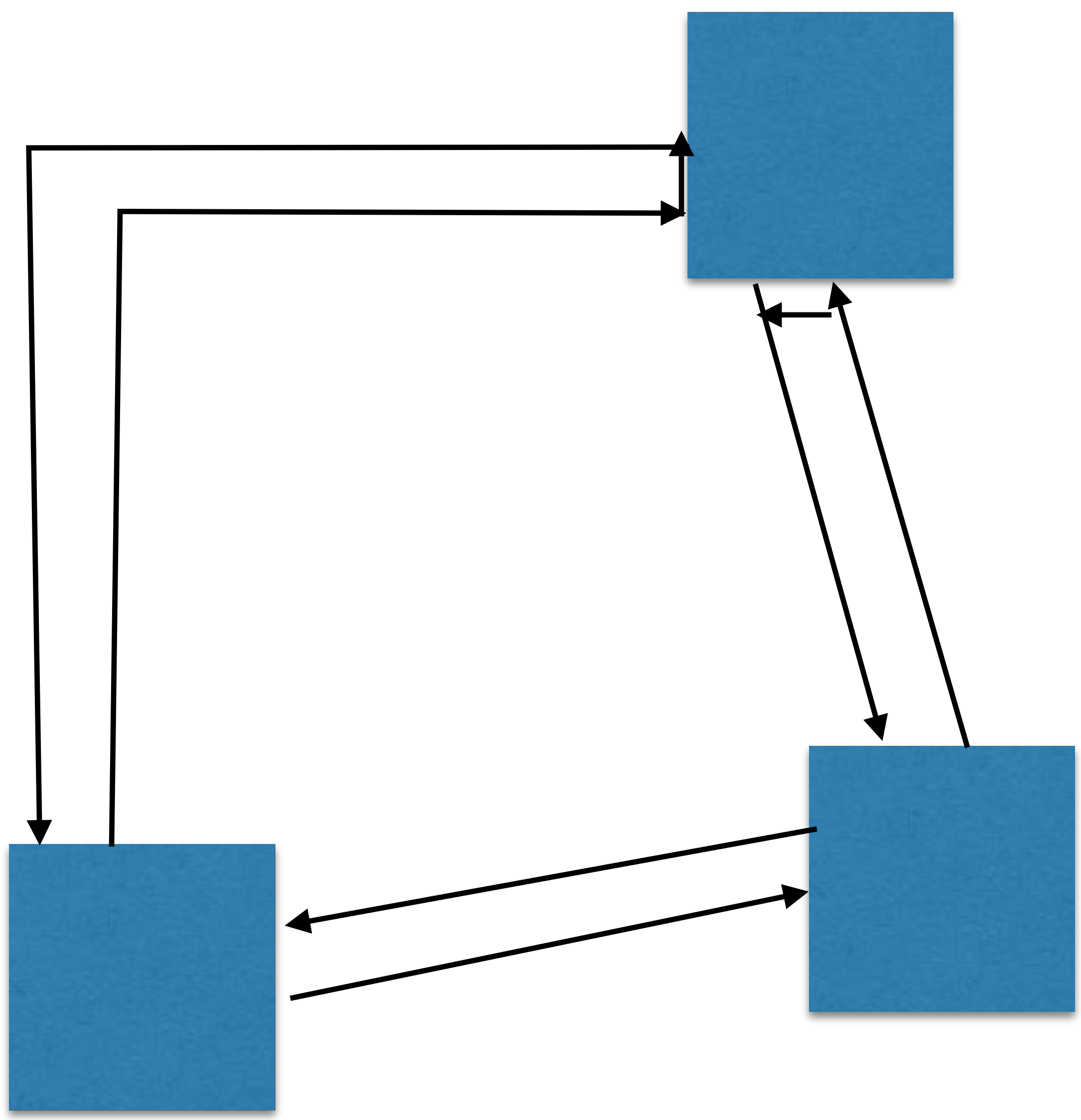
Problems

- In theory, offers bounded latency: the token will always circulate in n_stations * max_packet_period * fudge.
- In practice, very complicated to get right
  - Token loss/creation
  - Station failure

Examples

- IBM Token Ring (4Mbps, later 16Mbps)
  - Still occasionally encountered
  - Uses star topology for wiring
- FDDI Fibre (100Mbps, fastest game in town until switched full-duplex 100BaseT with cut-through switches).
  - Genuine dual ring, with complex passthrough and loop reversal algorithms
  - Still in use in interconnects and data centres, although not in new installations
  - Extraordinarily robust and stable in performance

Dealing with Failure



What happens if two nodes fail in a large ring?

CDDI

- There is also a variant called CDDI, FDDI over copper, using very specialised hubs with multiple paths.
- It works well and can survive multiple failures; it was also staggeringly expensive until supplanted by switched 100BaseT.

Slotted Rings

- Known as “Cambridge Rings” from their place of development (Cambridge in East Anglia, not Cambridge Mass).
- Instead of circulating a token, empty data frames circulate, in the manner of the conveyor belt in a Sushi restaurant, or other alternatives

Slotted Ring

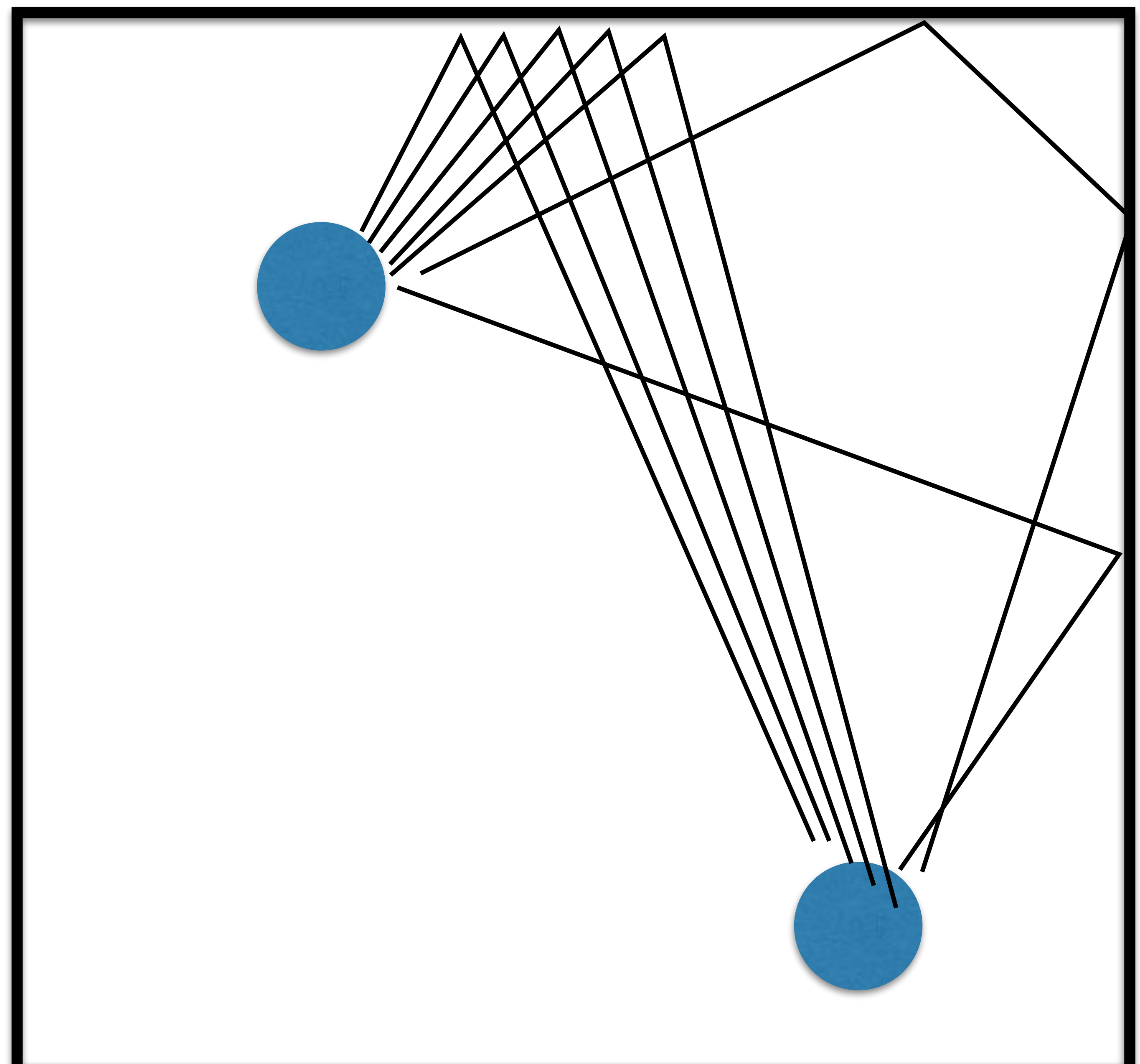
- Requires a minimum length of network, so that there are a sufficient number of empty packets circulating
  - Hence long lengths of cable coiled under the floor
- Popular in UK universities as boards were cheap and easy to build and drivers were available for common Unix variants; never achieved significant traction elsewhere.
- Probably lurking in floor voids of cl.cam.ac.uk, ukc.ac.uk and elsewhere.

ATM: The Telco Strikes Back!

- ATM: Asynchronous Transfer Mode
- Proposed by Telcos as part of the broadband unified services architectures of the 1990s.
- For reasons of nasty politics, breaks data into a stream of 48-byte packets.
  - Americans ~~and everyone remotely sensible~~ wanted 64, French wanted 32 because then they could run voice without needing echo cancellation, compromise of 48 suited no-one.
- Virtual circuits, so only needs a 5-byte header (but again political, as 5 is ~10% of 48 which was seen as “acceptable”)

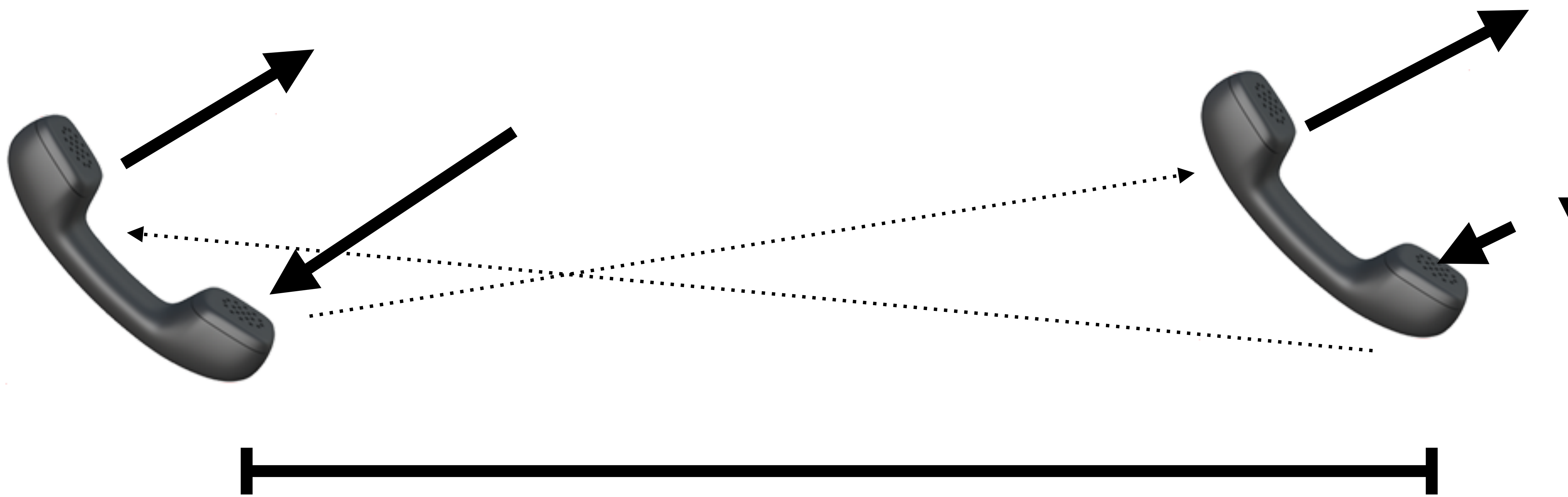
In passing…echo cancellation

- If you are speaking in a room, the echo from your voice is a diffuse field of noise, as the many possible paths all have slightly different lengths.
- Your brain is very good at dealing with this, and you aren't normally aware of the reverberation of a small room (but wait until you get older!)
- Your brain rejects any stronger echoes arriving within ~50ms ("Haas effect")



Telephones aren't rooms

Any echo is a sharp, single event that your brain struggles to reject



Target: 35ms RTT, equal to ~10m of air
Light travels 10000km
Reality in digital systems…?

America is big

- Speed of light means that for a phone call from New York to San Francisco you are not realistically going to be able to get it under 35ms whatever you do
- Hence you need to use complex electronics to filter out the echo (“echo cancellation”) to get decent “toll quality” audio.
- France is a lot smaller, and you can get away without the complexity

Latency caused by filling packets

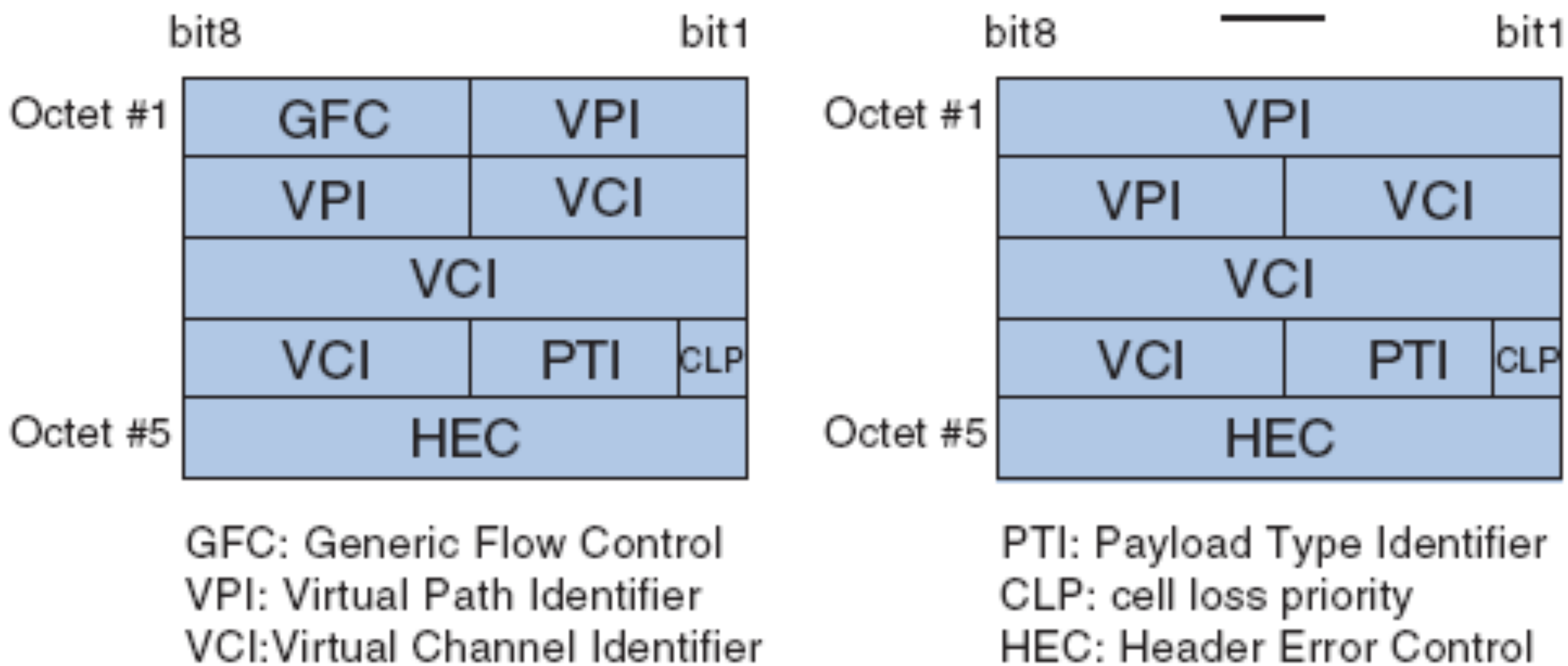
- Filling a 64 byte packet when you are sending 8KHz, 8 bit samples (ie, 64Kbps): 8ms
  - Note: filling a 1280 byte packet (20x bigger) is 160ms!
- Receiving it at the other end: 8ms
- That's 32ms round trip: almost all your budget gone
- With 32 byte packets, 16ms: you've got time to switch the packet
- 35ms - 16ms = 19ms, 5700km at speed of light
- Americans were running echo cancellation already so didn't care, and wanted larger packets for efficiency
- French wanted smaller packets to avoid the problem.
- Everyone lost, as 48 byte packets satisfied no-one (and made the standard look a bit mad)

ATM Justification

- Smaller packets gives lower latency (but not low enough, as we saw)
- Switching a stream of small datagrams is allegedly very inefficient (large headers, lots of routing decisions)
- ATM is therefore virtual circuit orientated
- Also incorporates extensive traffic shaping and policing options (more later)

ATM Headers

User-Network (UNI) Network-Network (NNI)



Note: for extra fun, addressing information is not byte-aligned

ATM25

- 25Mbps
- Can be built using adaptations of IBM 16Mbps Token Ring hardware; easy to encapsulate into USB 1.1 or USB 2.0.
- Was the dominant interface for ADSL modems during the late 1990s, and is the internal switching format for ADSL exchange equipment
- Still very influential in the form of PPPoA.

ATM155, 622…

- Faster variants used (mostly) within telco core networks, although enjoyed a brief period of use in data centres prior to being killed by cheap GigE.
- Can be used to carry IP in various forms
  - “classical” uses a virtual circuit as a two-station network,
  - “LAN Emulation”, aka “LANE”, tries to emulate a larger ethernet with lots of switching: scales very badly
- Further breaking ~1500 byte IP/Ethernet up into 48 byte cells (“AAL5”) appalling for performance and reliability
- But is a good way to mix “toll quality” voice with data for multi-service networks.
- Proved too complex, too expensive, and switch vendors were acquired and progressively run down
- Still in use in carrier networks, but being pushed out by ethernet.

Nailed Up Circuits

- ATM is virtual circuit orientated: you ask the network to establish a circuit, and once set up the packets just have to say which circuit they are on.
- Original idea for UK ADSL broadband was switched virtual circuits (SVC): you could choose your ISP dynamically, and a visitor could plug into your line and use their ISP (think dial-up, if you are old enough).
- Unfortunately…

Performance Hopeless

- ATM switches couldn't handle volume of circuit establishment required, even in early trials ("Project Ascot" in Ealing, a few thousand houses)
- Solution was "permanent virtual circuits" (PVCs), nailed up at the point at which the service is commissioned. Hence the "0.38" or "0.101" you may be familiar with: that's the identity of the PVC from your house to your ISP.
- Messy.

A bit of transmission

- SDH: Synchronous Digital Hierarchy
  - aka SONET (synchronous optical networking) in US., which has detailed differences.
- Multiplexes “trails” of 2Mbps upwards into STM1 (155Mbps), STM4 (622Mbps), STM16 (2.4Gbps) and STM64 (10Gbps).
- You can extract and insert individual 2Mbps trails from a passing 10Gbps stream (“add/drop multiplexor”)
- “Packet over SONET” aka PoS still regularly used for long-haul Internet traffic. Most telco transmission equipment up until five year ago was SDH.

Wave Division Multiplexing

- (D|C) WDM
  - Dense/Coarse Wave Division Multiplexing
- Use different colour light to transmit multiple streams down a single fibre. For “colour” say “lambda” if you want to hang with the cool kids.
- Coarse: 20nm difference between adjacent channels
- Dense: originally 0.8nm difference between adjacent channels (100GHz channels based around 193.1THz reference).
  - now can be 0.4nm or 0.2nm differences in wavelength.
- Commercial systems go up to 10Tbps and beyond˜

Using WDM

- Each channel can carry different traffic (including ATM, ethernet, SDH, whatever)
- Increasingly, ethernet straight over WDM is the way telcos are going, with the assumption that most ethernet will just be carrying IP (what else is there?)

Summary

- Ethernet works for getting data between computers that have cables between them. It won the battle.
- Other things can be made to work, but were more expensive/more complicated/harder/more political, and lost
- In 2017:
  - Short range: ether over copper
  - Medium range and/or hostile environments: ether over multimode fibre
  - Long range: ether over WDM