Networking: Lower Layers

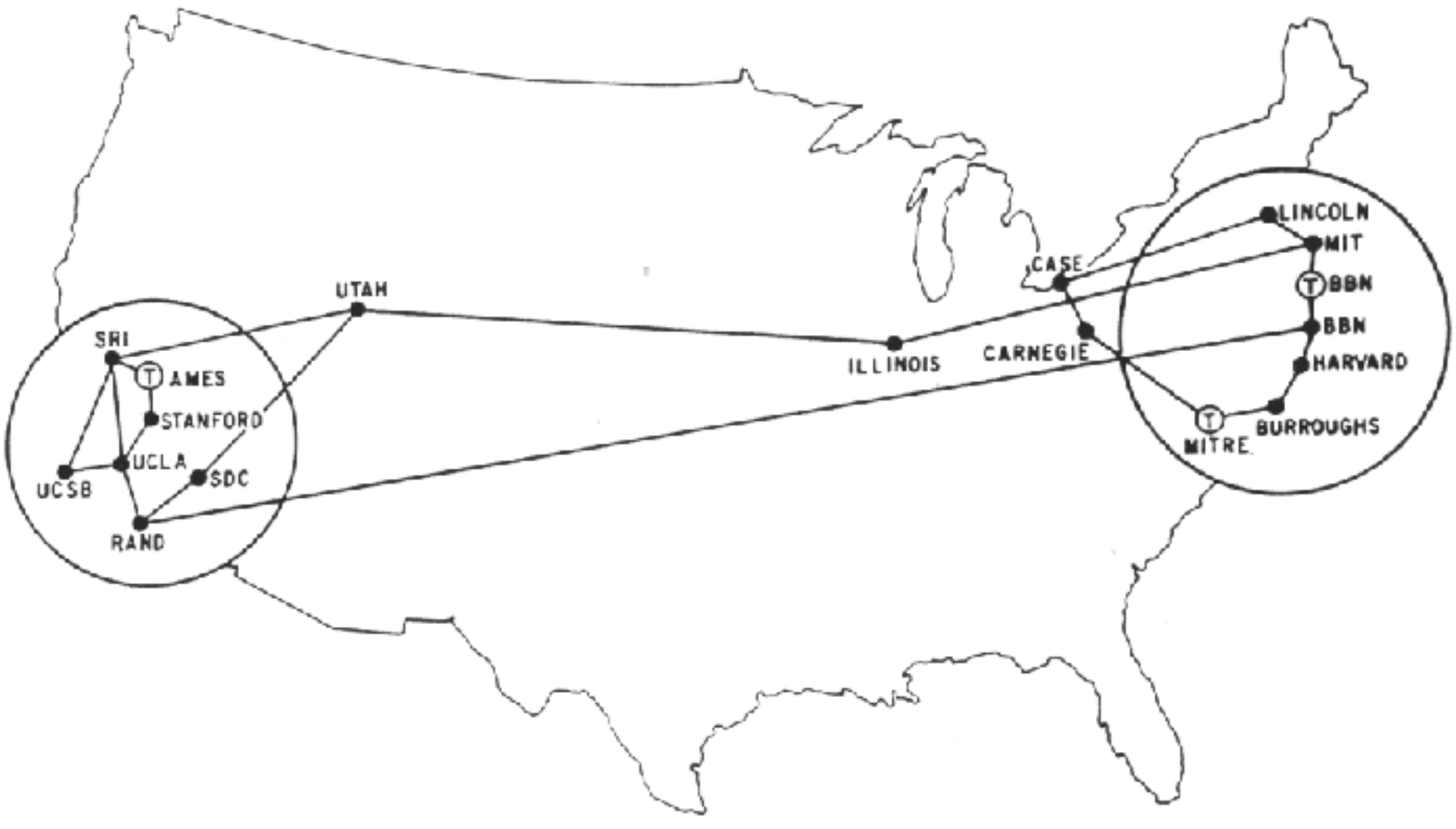
i.g.batten@bham.ac.uk

Panopto: GKAP-LT2: 12 Oct 2018

Check Panopto!

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

ARPANet, Sep 1971



MAP 4 September 1971

LANs, MANs, WANs…

网络空间划分为本地(建筑，校园)，有时是大都市(城市)，然后是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.

而且(不是本课程)有时个人区域网络，PANS，就是现在的蓝牙。

- Also (not this course) sometimes **P**ersonal **A**rea **N**etworks, PANs, mostly today Bluetooth.

- Theoretically, different technical solutions to different engineering problems.

从理论上讲，不同的工程问题有不同的技术解决方案。

- Rapidly converging.

LANs and WANs

从历史上看，广域网络非常缓慢

- Historically, wide-area networks were very slow.
  - 20世纪80年代的快速长途网络可能是每秒几十Kbit
  - 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.

技术的存在是为了走得更快，问题主要是成本和实用性。

尽管局域网在1980年达到了10mbps，在1990年达到了100mbps。

- 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

历史上，局域网和广域网在技术、目的和协议上是不同的。

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

在欧洲和美国(某种程度)，电信垄断限制了广域网能做的事情。

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

在少数研究环境之外，局域网几乎完全是专有的:交互主要是与广域网技术。

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

WANs

用来把电脑连接在一起，比如相隔1公里以上的建筑物之间，或者如果当地法律允许电信公司垄断就可以把电线穿过马路。

- 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: 历史上主要有三种应用:
  - File transfer (lots of problems of format conversion, as even byte-size varied) 文件传输(许多格式转换的问题，甚至字节大小的变化)
  - Job transfer (for use of national facilities for super computers; batch mode) 工作内容转移(用于国家超级计算机设备的使用);批处理模式)
  - Remote login (when you interactive access to remote systems,m which was not always available). 远程登录(当您交互访问远程系统时，但远程系统并不总是可用)。
- UUCP very influential and STILL SHIPPED ON MAC!

  UUCP很有影响力，仍然在MAC上运行!
- ARPANet in the US restricted only to people with government contracts

  在美国，ARPANet仅限于与政府签订合同的人

WAN Technology

- The key point about the WAN is that for most of its history it is slow.

广域网的关键一点是，在其历史的大部分时间里，它是缓慢的。是非常慢！

- 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

这意味着效率非常重要:浪费数十字节是一个重要的性能问题

- This means that efficiency is very important: wasting tens of bytes is a significant performance problem

因此，如果你打算在广域网和局域网上使用相同的协议，局域网中使用的协议必须考虑在低速、损耗的链路和建筑物内的高速网络上工作。

- 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.

这里的关键问题是，当开发时，局域网比广域网快得多;如今，这种情况在很多情况下恰恰相反。

- 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

真正的电路包括从一端到另一端的电气连接

- Real circuits involve electrical connections from end to end

包交换包括在包里放置一个地址，然后将它们分别发送到目的地

- Packet switching involves putting an address on packets and sending them to the destination individually

每个包可以包含完整的目的地信息，也可以与虚拟电路相关联。

- Each packet can contain full destination information, or can be associated instead with a virtual circuit.

虚拟电路使对象看起来像数据包流中的电路

- Virtual circuits make objects that look like circuits out of a stream of packets

假设:网络是由某种媒介连接的路由器(交换机)网络。

- Assumption: network is a mesh of routers (switches) linked by some sort of medium.

Packet Switching

- Each packet has addressing information
  路由器查看传入的信息包，决定将其发送到何处，并在途中将其发送出去
- A router looks at incoming packets, decides where to send it, sends it on its way
  路由器查看传入的信息包，决定将其发送到何处，并在途中将其发送出去
- Router “complexity” scales by the number of packets processed, and possibly other things.
  路由器的“复杂性”取决于处理的数据包的数量，可能还有其他因素。

Connection Orientated / Virtual Circuits

如果底层网络支持虚拟电路，您可以要求网络将数据包流发送到特定的目的地

- If your underlying network supports virtual circuits, you can ask the network to send a stream of packets to a specific destination

网络决定如何路由，并告诉沿途所有的路由器正在干啥，并给你一些token来识别流(“虚拟电路”)

- 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”)

您发送数据时带上token，它则会按顺序到达另一端。假设连接数比端点数少，这个token就更小更容易查找。

- 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.

好处是网络正在做大量的繁重工作，以确保所有的数据到达那里，所以您的网络堆栈被简化了。

- Upside: network is doing a lot of the heavy lifting of ensuring all the data gets there, so your network stack is simplified.

缺点是路由器要复杂得多，因为它们根据带宽、电路数量和电路创建/破坏的速率进行扩展。

- 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).

历史上是，X.25和(在较小程度上)Frame Relay(帧中继), 今天是ATM(即将退出，但仍然存在于许多网络中)和多协议标签交换(MPLS)(大型供应商网络的核心)。

Connectionless / Datagram

底层网络提供“原始”的分组交换

- Underlying network offers packet switching “in the raw”

单个数据包由网络处理，不能保证能到达，不能保证不被破坏。网络是“最好的努力”，没有保证。

- 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.

每个包上都有完整的目的地信息的用户地址包。

- User addresses packets with complete destination information on each and every packet.

对于更复杂的服务，所有的责任都在于端点:网络不会提供帮助(尽管显然它对数据包造成的损害越小越好)

- 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这是IP转移的基本要求

- You can use a virtual circuit as a link in a connectionless network, but not vice versa. 您可以使用虚拟电路作为无连接网络中的链接，但反过来不行。

Lower Layer Technology

- Ethernet
- Token Ring (IBM Token Ring, FDDI)
- Slotted Ring
- ATM
- MPLS Multi Protocol Label Switching
  多协议标签交换
- SDH
- DWM

Ethernet

- Developed by Metcalfe and Boggs at Xerox Palo Alto in the 1970s.

以一种发光的以太命名，据说这种以太可以携带光和无线电，直到迈克尔逊-莫雷实验证明它是错误的

- Named after the luminiferous aether that supposedly carried light and radio until disproved by the Michaelson-Morley experiment

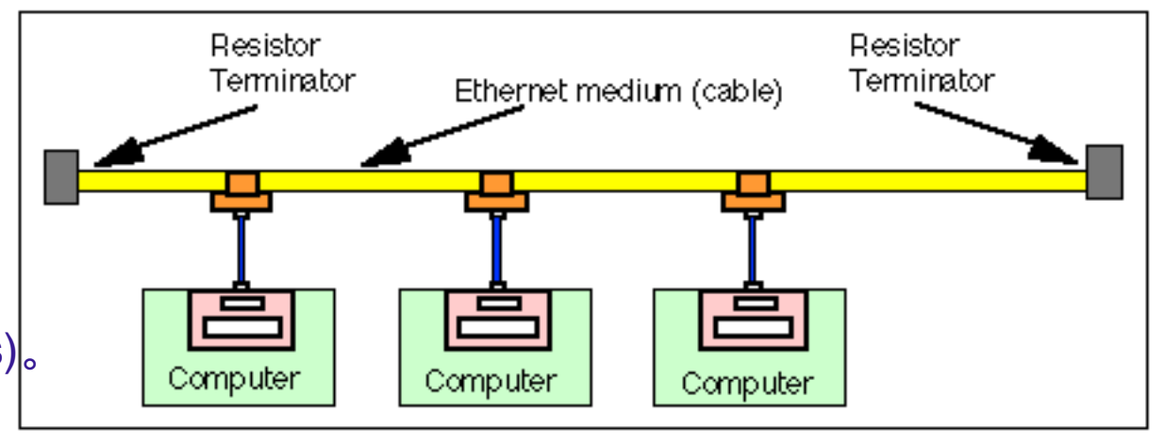
从早期的无线分组网络中获得灵感，特别是夏威夷的AlohaNet。

- Takes inspiration from earlier radio packet networks, notably AlohaNet in Hawaii.

Topology

现在已经没怎么用了，最初的想法就是一整栋楼只用一根线连所有电脑

- The topology of Ethernet was originally a bus: a single cable with computers connected to it.



早期的版本是3Mbps，但实际上“黄色软管”总是10Mbps)。

- (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.

最大长度为500米(由于阻力和时间原因，我们将看到);可放大和再生，最高可达1500米。

最初设计的格式，不幸的是我们仍旧卡在上面。并且要兼容不同的格式

每个以太网package由7个字节开始(0x55)

Format

7字节的电报报头(0x55)，允许接收器同步。

- 7 bytes of **preamble** (0x55) to allow receivers to synchronise.
- 1 byte **start of frame delimiter** (0x5d) 1字节用于开始帧分隔符(0x5d)
- 6 byte **source address** (48 bits) 6字节源地址(48位)
- 6 byte **destination address** 6字节目标地址
- 4 byte **VLAN tag** (optional) 4节的VLAN标记(可选)
  - First two bytes 0x8100 to keep older equipment happy
    前两个字节0x8100以使旧设备满意
- 2 byte **type** or **length** 2字节表示类型或表示长度
  - If ≤ 1500: length. If ≥ 1536: type, with length found by looking for end of the packet 如果≤1500表长度。如果≥1536表类型, 包的末尾写长度
- 42—1500 bytes of **payload** 42 - 1500字节的有效载荷
- 4 byte **CRC**
- 12 byte-time **inter-packet gap**. 12字节时间内的包间间隙

Finding the end without a length

当不知道长度时如何寻找终点?

CheckSum校验和是连续计算的，所以当你有一组字节，其中最后四个字节是整个包的正确校验和时，你知道你已经到达了终点。

- 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的计算(data + CRC)产生了神奇的数字0xC704DD7B -谷歌这个数字为gory GF(2)细节。

- 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 或者等到数据包之间的间隙
- Or both

Basic Logic

每次只有一个终端可以及时发声，其它终端可以看到该终端说了什么，多个传输则会发生碰撞，它用了个很傻的算法就是去听目前有没有终端在发声，如果有 那我就不发声

- 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.

每个电台都在等待，直到没有人说话，然后开始发射。

- Each station waits until no-one else is talking, and then start transmitting.
- What could possibly go wrong?

Collisions

这个算法会有什么问题？碰撞

当两个设备同时判断到无人发声时则会一起发声，并且它们不会瞬间知道还有人也在讲话，因为电流传输有一定速度才会接收到

以太网的正式名称是“CSMA/CD”——载波感知多址碰撞检测。

- Ethernet is formally known as “**CSMA/CD**” — Carrier Sense Multiple Access Collision Detection.
- The magic comes from what happens when there is a collision. 魔法来自于碰撞发生时发生的事情。

Collision Detection

当一个终端在发声时，它同时会监听是否有其它终端在发声并且确保你的信号是被正确发出的

如果两个信号被同时发出，它们的内容会被混在一起，如果这种情况发生了则可以判断另一个人同时在传送

当一个电台传输时，它同时也监听以太网并且检查只包含被发送的信号

- 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. 这必须在硬件中完成，因为它主要是一个模拟问题。
- If there is a mismatch, someone else is transmitting at the same time. 如果不匹配，另一个人同时在发送。

the set of all stations whose packets might mutually collide is called a “collision domain”

When Collisions Happen

防止别人再进入使其变得更糟，先发送一个pattern让所有人都知道已经发生碰撞了

首先要做的是“干扰”网络:发送一个固定的模式，让每个人都知道正在进行的碰撞。

- First action is to “jam” the network: send a set pattern so everyone knows a collision is in progress.

关键的是让整个以太知道会碰撞 在包已经完成发送之前

- 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.

强加一个最小包大小(64个八位字节)，这是碰撞域最大直径的函数(1500米)。Jam Pattern Pads至少到这个长度。

从碰撞中恢复

Recovery from Collision

在第一次尝试时，从{0,1}中选择一个随机数k，并在再次尝试之前延迟k x 512位周期。

- On the first attempt, choose a random number k from {0,1} and delay k x 512 bit periods before trying again.

通俗讲：在第n次尝试时，从{0..2^n}中选择一个随机数k，然后在再次传输之前延迟k x 512位周期。

- 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. 超过十次后便放弃
- Randoms come from things like serial numbers; they don’t need to be very good quality.

Randoms就像序列号，他们不需要很好的质量。

Problems

这会让以太网变得不稳定

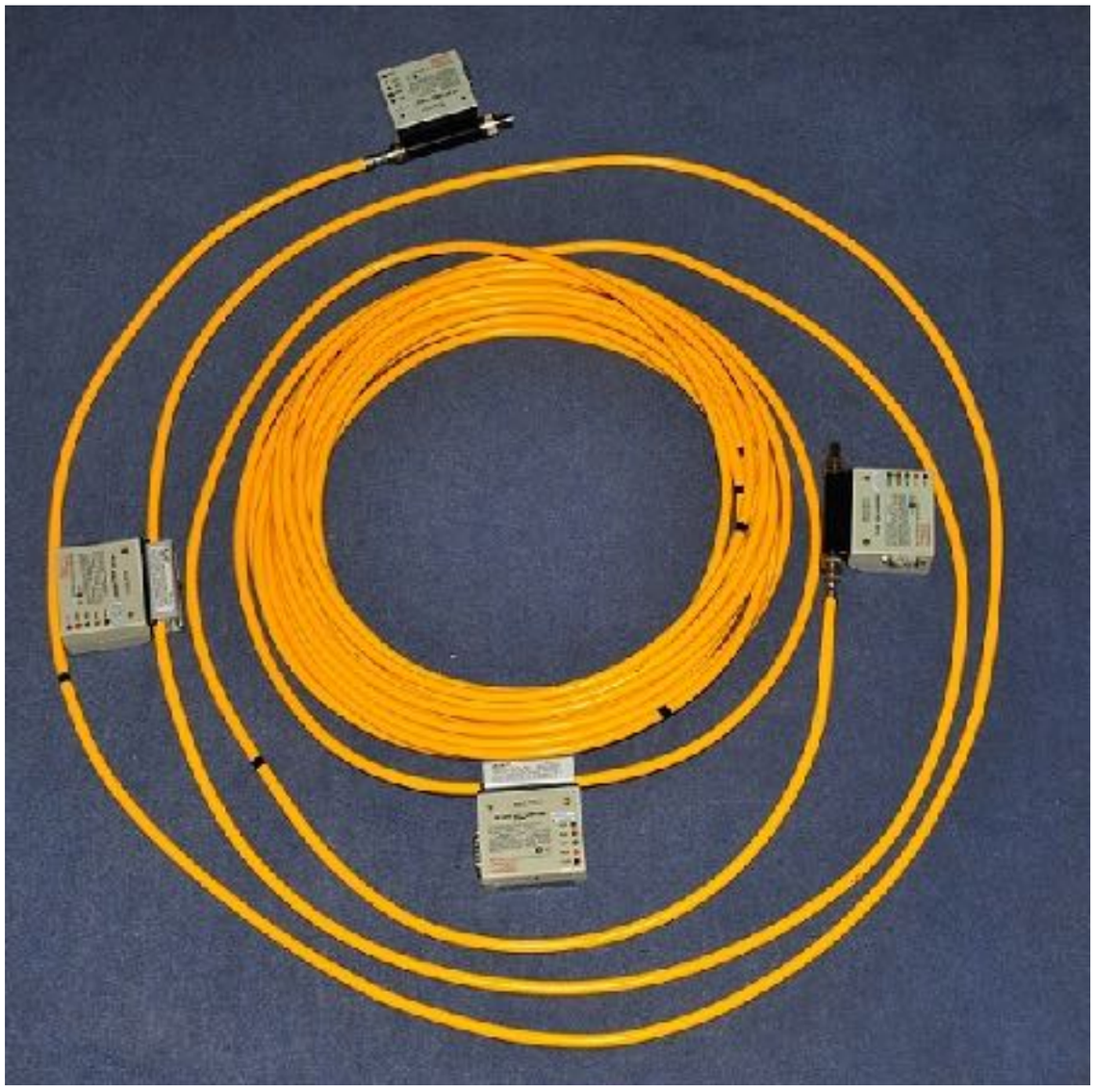
碰撞与负载呈非线性增长，精确的曲线取决于准确的交通混合

- Collisions increase non-linearly with load, and the precise curve depends on the exact traffic mix
  - “Ethernet capture effect” 以太网捕获效应
- Latency for a single packet is unpredictable, because some number of collisions may delay it.
单个数据包的延迟是不可预测的，因为一些碰撞可能会延迟它。
  - This can be overstated by advocates of other protocols. 其他协议的支持者可能夸大了这一点。

Sizes

最大帧大小1500字节有效负载加上22个header(更大的最终会减慢想要交换小包的站点)

- 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

电缆重量大，价格昂贵，安装困难(紧密，或更松一点，最小弯曲半径要求)。

- Cable is heavy, expensive and difficult to install (tight, or more to the point loose, minimum bend radius requirements).

为收发器安装taps需要进行钻孔，而且有损坏电缆的危险。

- Installing taps for transceivers involves drills, and risks damaging the cable.

对收发器的需求增加了成本和复杂性

- Need for transceivers adds cost and complexity.

隐藏在后台的性能问题

- Performance issues lurking in the background

An interim: 10base2

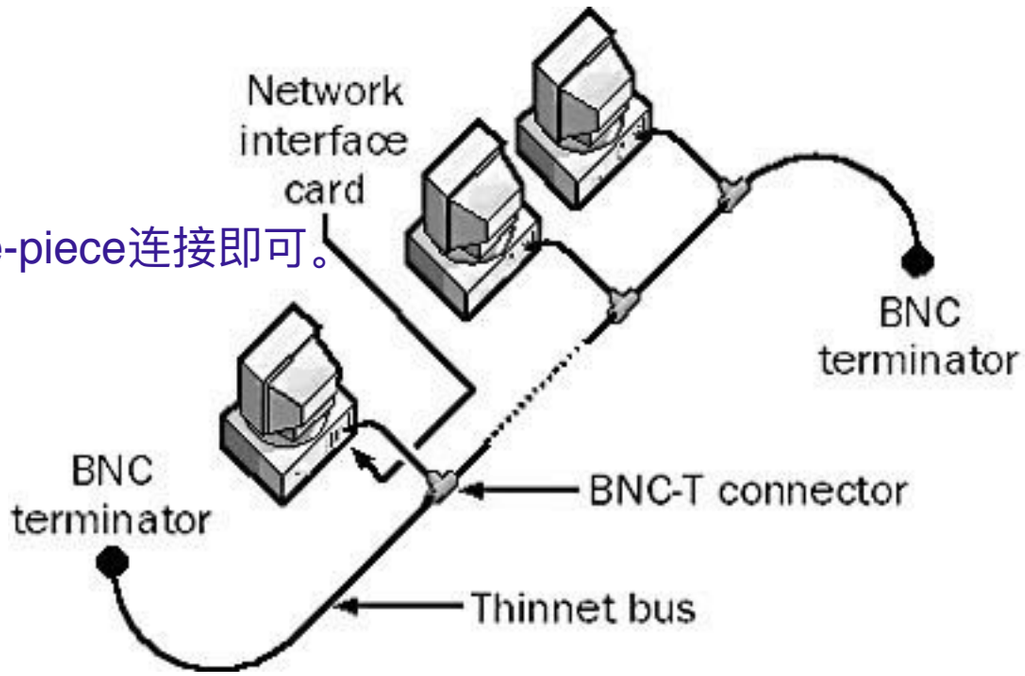
用薄同轴而不是厚同轴。更高的电阻，因此限制在185m: 10base2。

- Instead of using thick co-ax, use thin coax. Higher resistance, so limited to 185m: **10base2**.

不使用收发器，只需将同轴电缆带到计算机上，并将其与tee-piece连接即可。

- 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. 切断电缆，而不是钻进去。



10Base2

- Otherwise it works much the same
  - much smaller maximum diameter of <600m
    最大直径小于600米
  - Different terminators 不同的终端
- Can be mixed electrically and logically with 10Base5 (rules are complex and only of historical interest) 可与10Base5电气和逻辑地混合(规则是复杂的，并且只具有历史意义)
- Probably the dominant networking of the 1980s and early 1990s: older buildings still full of it.
  或许是上世纪80年代和90年代初的主流社交网络:老旧的建筑里仍然充斥着这些东西。

10BaseT

同轴电缆仍然是一个痛苦:昂贵，难以安装，容易损坏。

- 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.

10BaseT看起来像“现代”的以太网:使用RJ45连接器连接到集线器，高达95m的双绞线对(两对四根导线)。最初的“3类”电缆，基本上用于音频。

Hubs, Repeaters, etc

Hubs就是Repeaters，把一端的东西复制到另一端

- A repeater is just an amplifier: collisions are seen on both sides 中继器只是一个放大器:两边都能看到碰撞

ethernet bridge：把建筑连接在一起

- A bridge receives, buffers and transmits frames, so collisions are not propagated 桥接器接收、缓冲和传输帧，因此不会传播冲突

learning bridge 会查看两边的源头，并且学习这两边到底是谁，然后只传输属于另一边的包(看下面英文)

  - “learning” or “filtering” bridges only send frames that belong on the other side; stupid bridges just propagate everything.

“学习”或“过滤”桥只发送属于另一边的帧;愚蠢的桥梁只是传播一切。

- Ether hubs are **repeaters**, not bridges. There are collisions when two stations talk.

以太集线器是中继器，不是桥接器。两个电台通话时会发生碰撞。

Repeater 中继器

从1到2的传输可能与从4到5的传输发生冲突

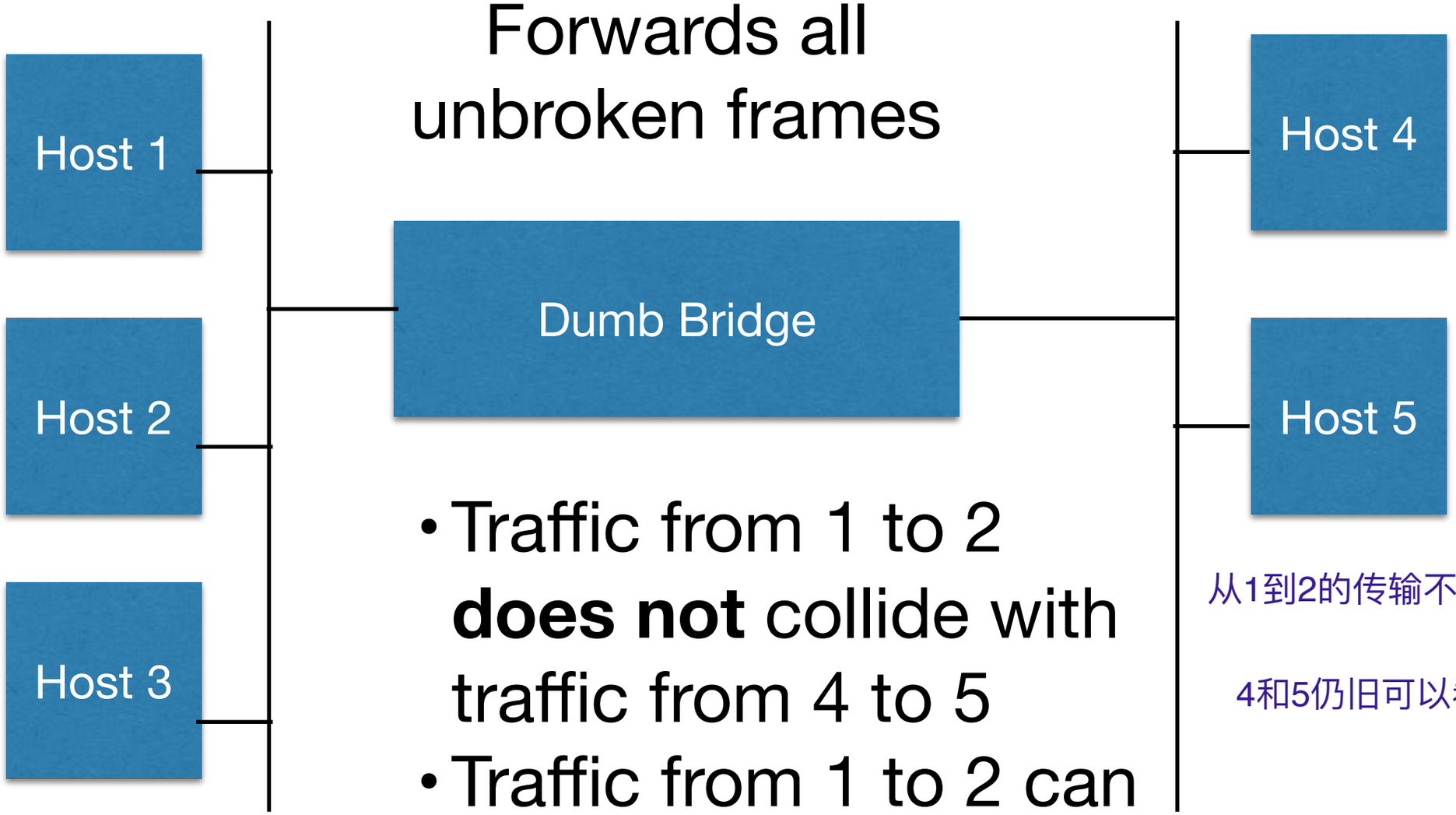
并且4和5可以看到从1到2的传输

- 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

所有完整的框架向前推进

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

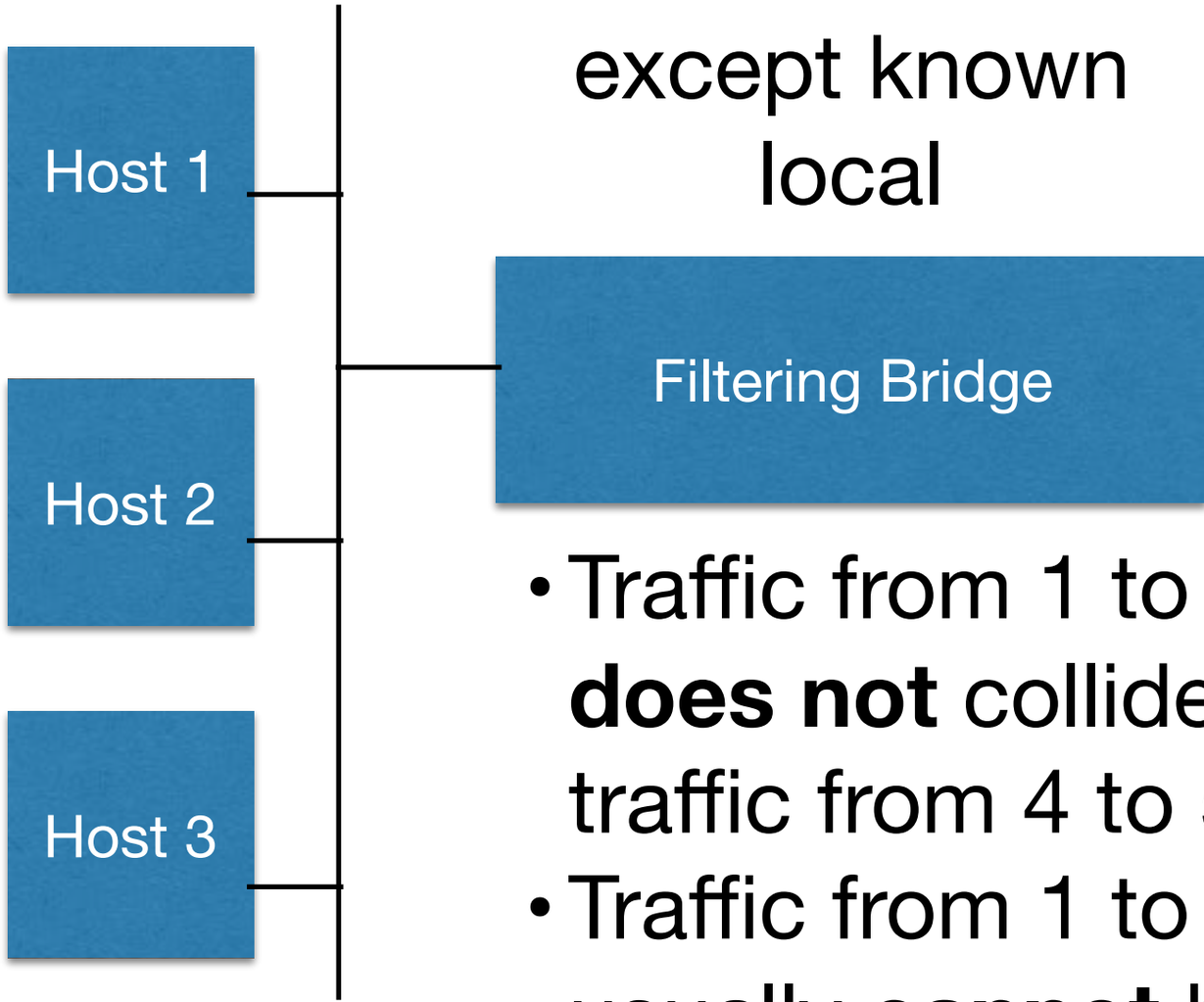
从1到2的传输不会与从4到5的流量发生冲突

4和5仍旧可以看到从1到2的传输

Filtering/Learning Bridge

Forwards all unbroken frames except known local

除了已知的局部帧之外，所有的帧都是完整的



- 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

从1到2的传输不会与从4到5的流量发生冲突

4和5无法看到从1到2的传输

Faster and Faster

10BaseT的速度没10Base2快，但更便宜，安装更灵活。

- 10BaseT is no faster than 10Base2, but cheaper and more flexible to install.

100BaseT提高了速度，但仍有可能发生碰撞

- 100BaseT raised the speed, but still had potential for collisions

全双工和开关使100BaseT的速度快得多，其次是1000BaseT (GigE)，然后是10GigE、40GigE和新生的100GigE。

- Full duplex and switching made 100BaseT much faster, following by 1000BaseT (GigE) and then 10GigE, 40GigE and the nascent 100GigE.

技术类似，但更严格的布线规则(“Cat5”用于100BaseT，“Cat5e”或“Cat6”用于更快)。

- Technology similar, but stricter wiring rules (“Cat5” for 100BaseT, “Cat5e” or “Cat6” for faster).

五类线 超五类 六类

尽管我们把网速提得越来越快，但还是有冲突的情况发生，
所以我们有了交换机

Ethernet Switches

交换机本质上就是一组封装在盒子内的learning bridges

- A switch is a set of learning bridges in a box.

每一个交换机上的接口port都有自己的冲突域

- Each interface is its own collision domain. 每个接口都是自己的冲突域。
- Packets to unknown destinations are sent out of all ports, otherwise only traffic for devices plugged in to the port is sent. 对于未知目的地的信息包则从所有端口发出去，
否则只发送到连接到端口的设备。
- “Full Duplex” means traffic goes in and out without colliding as **each direction** is a separate collision domain. “全双工”是指交通进出没有碰撞，因为每个方向是一个单独的碰撞域。
- Large buffers internally deal with congestion. 大缓冲区内部处理拥塞。
- Result: no collisions.

直通式交换机

Cut-Through Switches

保守开关接受整个帧，检查校验和，然后将它们传输到其他接口

- 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).

相较于一段直电线来说增加了延迟(对于GigE来说,1位周期是1 ns,整个package就是1500 * 8 ns = 12µs,相当于~ 3.6公里的铜;对于10BaseT，它是1.2ms，即360公里的铜)。

- Aggressive switches look at the header, and immediately start transmitting on the correct interface ("cut through).

激进的开关取决于header，然后立即开始在正确的接口传输("切断")。

  - Latency is just the 160–192 bits of the header, so <2% of a full packet: ~60m of copper for GigE, 6km for 10BaseT.

延迟只是信息开始的160-192位，小于整个包的2%: 相当于60m铜的GigE, 6km铜的10BaseT。

- This propagates broken frames if there are any to be propagated, as it can't check the checksum

如果有需要传播的帧，将会传播破碎的帧，因为它不能检查校验和

对于这些设备你可能会遇到的问题是buffering

Random Early Drop

当缓冲区被填满时，你就会开始丢弃数据包，因为你不能把它们放在任何地方

- Naively, when a buffer fills up, you start to drop packets as you can’t put them anywhere

通常情况下，丢包会导致超时，然后是重传，经过一段时间后，net会减慢传输速度

- 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. 一种新的策略是随机地丢包，其概率随着缓冲区的填充而增加，因此丢包开始得更早，但更轻柔，希望在真正的丢失发生之前降低速度。

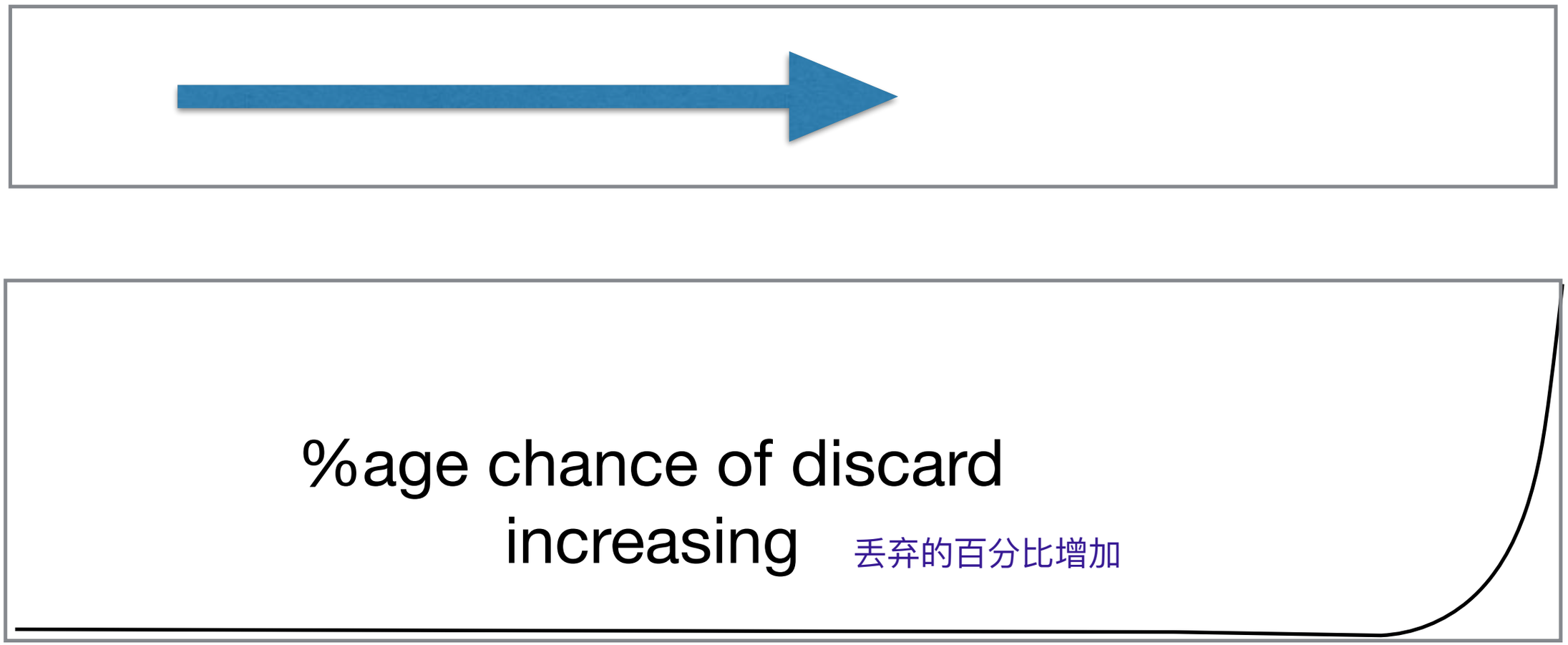
Random Early Drop是一种技术

- 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.

当确认信息停止时，发送方会看到报文丢失，这是让发送方减慢速度的信号。也就是你希望的。

Random Early Drop

Buffer filling…



Token Ring/Bus

人们认为以太网在高负载下表现不佳，尽管证据有限。

- 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. Token Rings和Token Buses通过一个“Token”穿过一个个站点。

- The station that holds the token can transmit, and then passes the token on when it has finished.

  持有token的站点可以进行传输，然后在完成后将token其传递下去。

Problems

从理论上讲，它提供了有限的延迟:令牌将始终在n_stations * max_packet_period *中循环。

- 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

在实践中，非常复杂：如token丢失/新建，站点错误

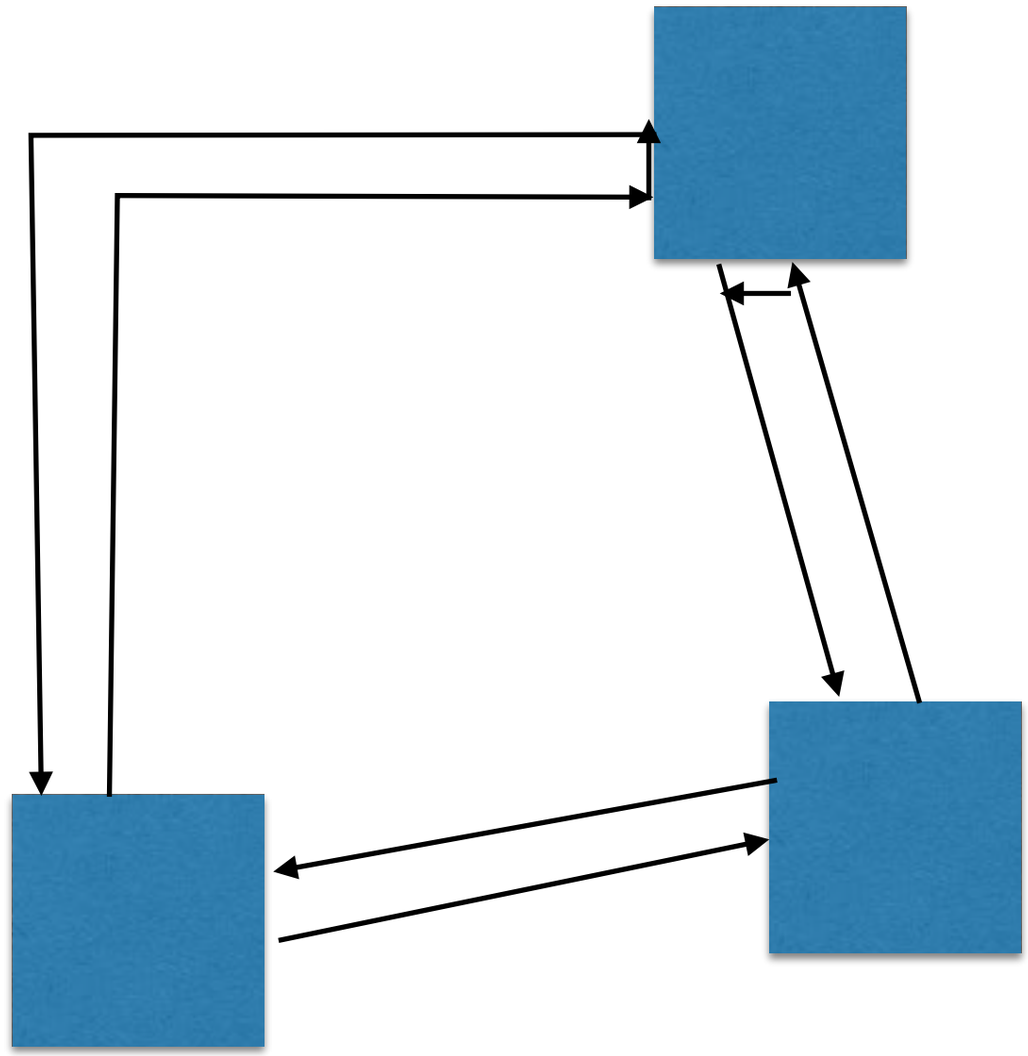
Examples

- IBM Token Ring (4Mbps, later 16Mbps)
  - Still occasionally encountered
  - Uses star topology for wiring

IBM Token Ring使用星形拓扑布线
星形拓扑是局域网络(LAN)的拓扑，其中所有节点单独连接到一个中心连接点，比如集线器或交换机。星形比Bus需要更多的电缆，但好处是如果电缆失败，只有一个节点会被关闭。

- FDDI Fibre (100Mbps, fastest game in town until switched full-duplex 100BaseT with cut-through switches).
  直到直通开关的全双工100BaseT出现之前，100Mbps的FDDI光线就是最快的
  - Genuine dual ring, with complex passthrough and loop reversal algorithms 真正的双环，具有复杂的穿透和循环反转算法
  - Still in use in interconnects and data centres, although not in new installations 目前仍在连接器和数据中心使用，不过在新安装的设备中没有
  - Extraordinarily robust and stable in performance 性能异常稳定

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

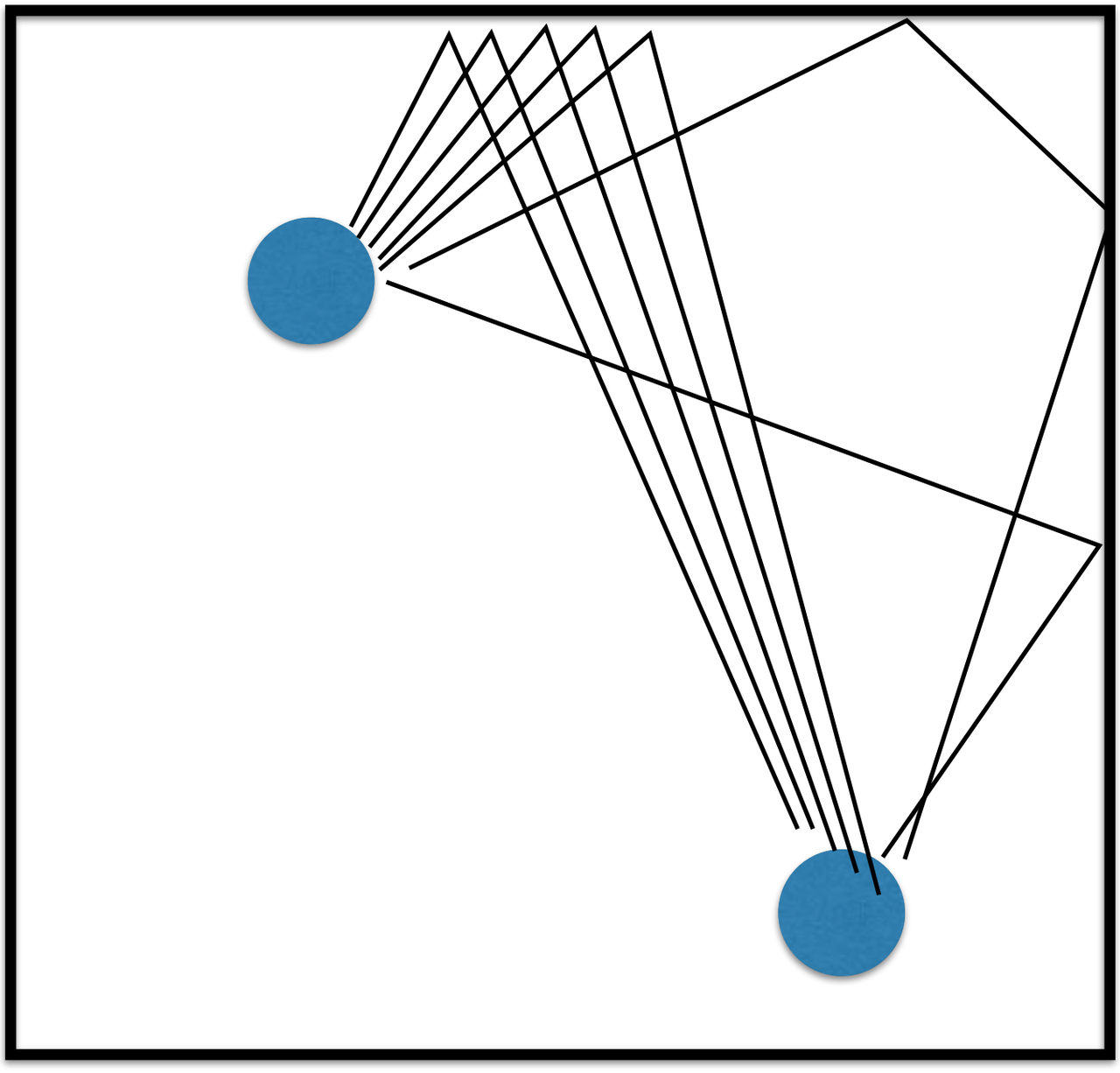
- Requires a minimum length of network, so that there are a sufficient number of empty packets circulating
  - Hence long lengths of cable coiled under the floor
- Popular in UK universities as boards were cheap and easy to build and drivers were available for common Unix variants; never achieved significant 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”)

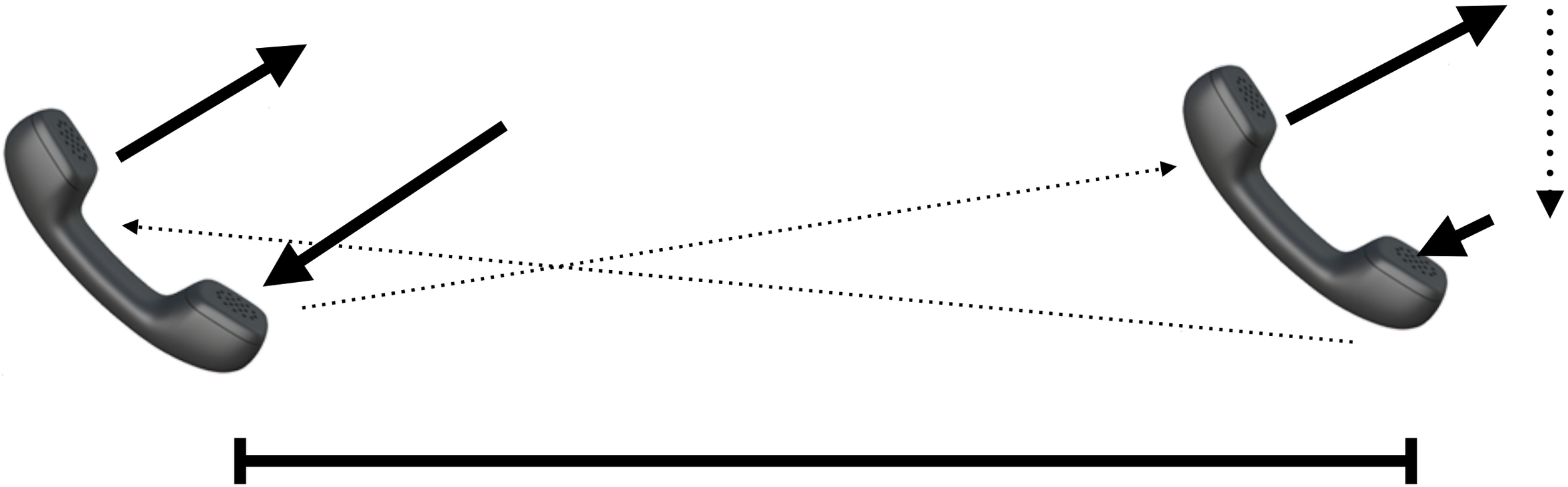
In passing…echo cancellation

- 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

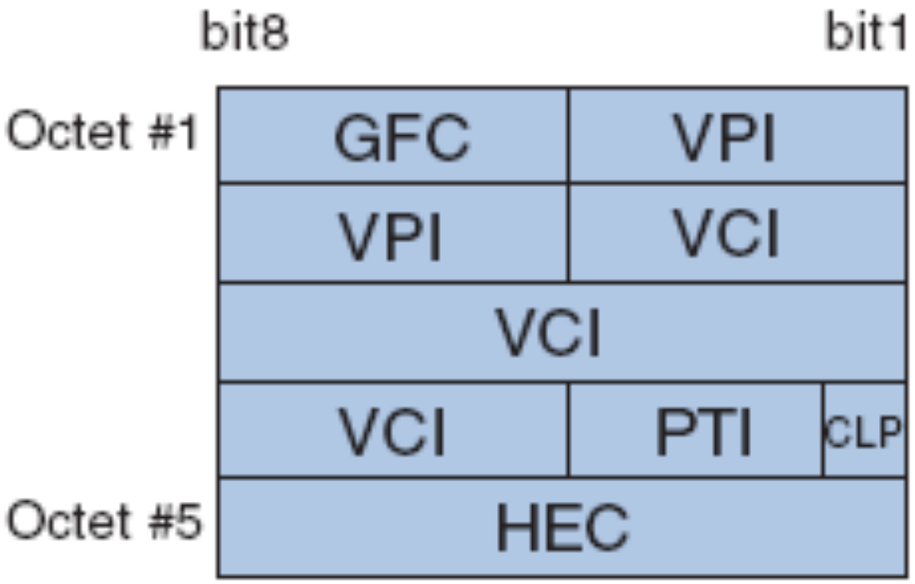
- Filling a 64 byte packet when you are sending 8KHz, 8 bit samples (ie, 64Kbps): 8ms
  - Note: filling a 1280 byte packet (20x bigger) is 160ms!
- Receiving it at the other end: 8ms
- That’s 32ms round trip: almost all your budget gone
- With 32 byte packets, 16ms: you’ve got time to switch the packet
- 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)

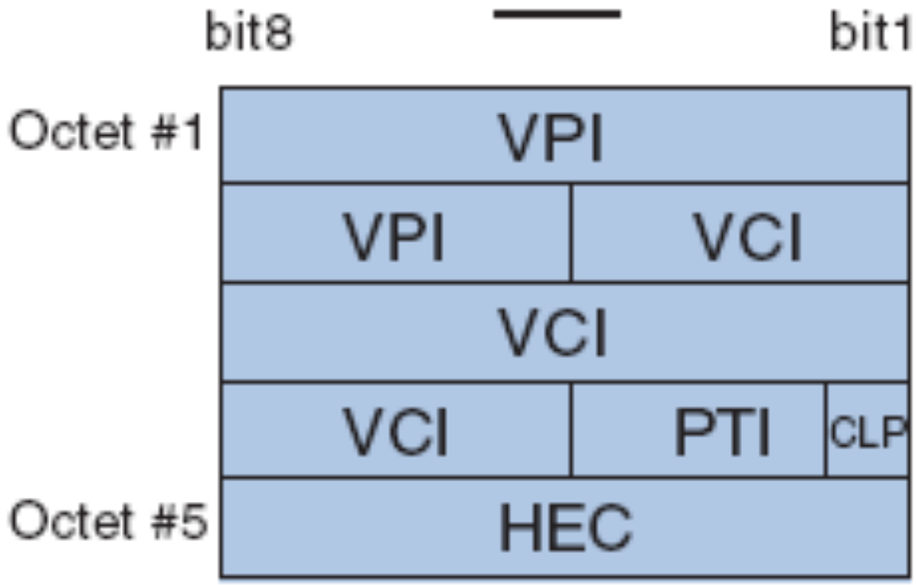
ATM Headers

User-Network (UNI)



GFC: Generic Flow Control
VPI: Virtual Path Identifier
VCI:Virtual Channel Identifier

Network-Network (NNI)



PTI: Payload Type Identifier
CLP: cell loss priority
HEC: Header Error Control

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 beyond̃

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