Protocols for Fibre to the Home

by Malcolm Scott

Doctoral Training Centre
Photonic Systems Development

Supervisors: Polina Bayvel and
Benn Thomsen, 2009/2010

I hereby declare that, except where specifically indicated, the work submitted herein is my own original work.

MAScott

Malcolm Scott
31st August, 2010

This report contains 10000 words.

Protocols for Fibre to the Home

Malcolm Scott

Technical Abstract

Fibre to the Home (FTTH), alongside intermediate stages (collectively, FTTx), has recently started to be taken seriously by telecommunication companies around the world, and enabling technologies are being developed rapidly. By far the majority of FTTH deployments in planning and in deployment use a Passive Optical Network (PON) in order to save on fibre costs. In a PON, multiple customers are connected to a single transceiver by means of a branching tree of fibres and passive splitter/combiner units, operating entirely in the optical domain and without power (the exact implementation depending on the variety of PON used).

There are two major current (and competing) PON standards: GPON, based originally on ATM protocols but in its latest incarnation using a custom framing protocol GEM, and EPON (for Ethernet PON), targeting cheaper optical components and native use of Ethernet. Both have seen considerable interest from telcos; meanwhile, telcos have been deploying Ethernet-based next-generation networks (such as BT's 21CN). As a result, PONs used for FTTH also rely on Ethernet, whether this is native in the case of EPON or encapsulated in GEM in the case of GPON. Probable future PON architectures are additionally discussed in Chapter 2, although it is likely that Ethernet will continue to be used through advances at the physical layer.

Ethernet, however, has several barriers to large-scale deployment which start to become an issue at 16,000 nodes—considerably less than the number of customers connected to an average 21CN metro node. A review of these problems and their solutions in a FTTH environment forms the core of this report.

The primary scalability barrier is the forwarding database which must be maintained by every Ethernet switch. This contains a record of every host on the network, and has a limited capacity—usually of the order of 16,000 hosts—in order to maintain performance; this is necessary because the MAC address space is flat, with addresses able to appear anywhere on the network. For comparison, a full deployment of FTTH in the UK could require Ethernet networks of between 800,000 and several million nodes. A further problem is that Ethernet is unable to effectively use a dense mesh topology, making it unsuitable for a telco's core network.

Current network protocol research seeks to address these problems, although most has focused on the latter problem whereas the former is the more important for a FTTH access network. My proposed switch architecture, MOOSE (Multi-level Origin-Organised Scalable Ethernet)—an overview of the operation of which is included in Chapter 3—transparently enforces a hierarchical addressing scheme onto Ethernet, simplifying forwarding databases and providing IP-like shortest-path routing, eliminating both scalability issues.

The report ends with a discussion on current open research problems of Ethernet scalability and potential avenues via which progress could be made.

Contents

List of Figures

CHAPTER 1

Introduction

This report provides a review of current technologies, architectures and standards for Fibre to the Home (FTTH)—that is to say, the provision of data connectivity to homes (and businesses) using optical fibres in place of legacy copper telephone cables. Rather than considering primarily the optical physical-layer characteristics of these systems as several existing reviews have done [Gutierrez et al., 2005; Kramer, 2006; Grobe and Elbers, 2008], I focus on the oft-neglected data link protocol and switching architecture—layer two of the standard Open Systems Interconnection (OSI) model [ITU-T, 1994]—as these play a major part in determining the capability and efficiency of a networked system.

In addition to describing current standards, I present current data link protocol research in the context of its utility in FTTH systems; although FTTH is only recently beginning to be deployed on a large scale, the protocols in use have changed little from those designed in the 1970s, long before the current vast scale of computer networks had been envisaged. Use of the products of modern protocol research would serve to considerably improve the scalability of FTTH systems.

This report forms one half of a joint piece of work on FTTH; my colleague Peter Ogden has written on physical layer optical network scalability [Ogden, 2010].

1.1 Drivers for FTTH

Users' demands for bandwidth are constantly increasing—or more accurately, their demands for bandwidth-heavy services are increasing, whether or not they understand the implications of these demands for access networks. Currently the biggest bandwidth driver is live high definition video, a single channel of which requires approximately 10 Mbit/s after compression—thus high definition Internet television is already beyond the capability of the majority of residential broadband connections: 76% of UK residential broadband connections were found to be slower than 10 Mbit/s in a survey conducted in May 2010 [Ofcom, 2010], although the situation is improving rapidly as this figure is down from 92% a year earlier.

Video is expected to continue in the medium to long term to be a significant bandwidth driver [George, 2006], with higher resolutions and frame rates (Super HD and Ultra HD) expected to gain popularity—requiring up to 200 Mbit/s per channel—along with three-dimensional video with even higher requirements. A number of stereoscopic 3D high-definition television channels have started broadcasting this year; stereoscopic 3D requires a 50% bandwidth overhead with respect to 2D video of the same resolution [Merkle et al., 2007]. However, stereoscopic 3D—filmed using two cameras—is not "true" 3D as it is optimised for a viewer sat in a specific location and with his pupils aligned with the line separating the camera lenses; more realistic 3D video, filmed in ultra-high definition with 20 or more cameras (or rendered from a computer-generated scene) is projected to require 2.5 Gbit/s per channel after compression [George, 2006]!

Although this seems a long way off, one must not forget that this field is moving extremely rapidly. Both high definition television and video distribution via the Internet have only taken off in the past six years (the first European HDTV broadcast started in 2004, and YouTube was founded in 2005). Given the expense involved in deploying FTTH on a large scale, it is not unreasonable to expect the infrastructure to last for multiple decades, by which time video technology—and users' expectations—will have advanced significantly.

It should be noted that because video broadcast has driven bandwidth increases so far, heavily-asymmetric bandwidth provision is common in the residential market, with upstream capacity generally around 3–6 % of the downstream capacity [Virgin Media]. Other applications are likely to start to drive increases in upstream speed, for example peer-to-peer content distribution [Cohen, 2003] and video telephony; FTTH is the only real option where significant upstream speeds are required.

1.1.1 Note regarding FTTx

Other literature frequently uses the more general term "FTTx" when describing fibre access networks; the "x" can stand for any of "home", "business", "premises", "building", "basement", "pole", "last amplifier", "curb", "cabinet" or "node"—in

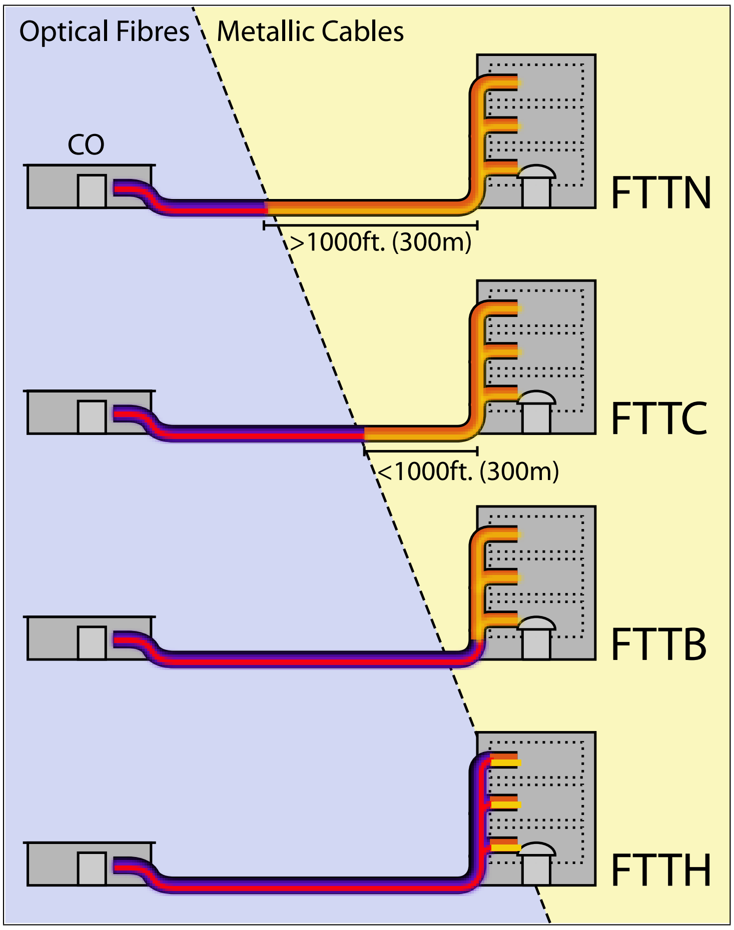


Figure 1.1: Illustration of some of the different degrees of FTTx deployment*

other words, some point between a central office and the end user (listed here in descending order of fibre penetration: see Figure 1.1). Where the fibre does not reach the customer's building, the remainder of the connection is made using legacy copper telephone lines, usually using VDSL [ITU-T, 2004] or VDSL2 [ITU-T, 2006] to achieve a moderately high data throughput over the shortened twisted copper pair cable.

It is likely that over time, fibre will gradually penetrate further into the access network. For greenfield housing developments, there is little reason not to install FTTH from the outset, but for existing buildings some use will be made of the legacy copper cables in order to save money; gradually more sections of copper will be replaced with fibre as demand for broadband data speeds increases. Thus the final objective is FTTH (or fibre to the premises or business, equivalent terms), and the other FTTx technologies in effect refer to partially-completed FTTH deployments.

For this reason, this report uses the term FTTH throughout (which should be read as referring to both residential customers and small businesses which currently use telephone cabling for data connectivity). However, the protocol discussion applies to most degrees of FTTx.

1.2 FTTH Technology

If cost were no object, the ultimate deployment of FTTH would involve one or more point-to-point fibre pairs being installed between every property and a central office (CO), complete with active amplification and signal regeneration along the fibre. Such a configuration would mirror the topology of the current telephone network (in which the central office is the customer's local telephone exchange). However, the cost of such a deployment would be prohibitive—in the UK, for example, a national deployment of point-to-point FTTH would involve the installation of many millions of miles of optical fibre.

The natural way to reduce this cost is to deploy fibres according to a tree structure: instead of laying a cable from each property all the way to the central office, it could just be laid as far as a nearby cabinet on each street or within each large building. This cabinet is then linked to the central office—or another aggregation cabinet—via a single, higher-speed fibre connection. This topology is similar to that used by many smaller-scale, private metropolitan area networks (MANs) and campus networks.

Traditionally, such networks have been built out of many point-to-point links: the cabinets terminate the downstream links to customer premises, and the upstream links from cabinets towards the central office are separate point-to-point links. As a result, the cabinets must contain active optical components and networking equipment—a large number of transceiver modules connected to a network switch. This, unfortunately, offsets a lot of the cost saved in laying less fibre: this equipment is expensive to purchase in the quantities that would be needed, and furthermore managing and maintaining complex network equipment distributed around cabinets all over the country causes high running costs for the provider [Analysys, 2005].

Therefore, in seeking to remove active components from the network between the central office and the customer premises, the concept of a Passive Optical Network (PON) was born [Lin and Spears, 1989]. This involves the use of passive fibre splitters which allow multiple customers (typically 32–128) to share a single fibre pair, or indeed a single fibre for both upstream and downstream communications. The disadvantage is that the customers must share the capacity of the fibre, but since the fibre can be run at a gigabit per second in current standards, with 10 Gbit/s standards on the horizon,

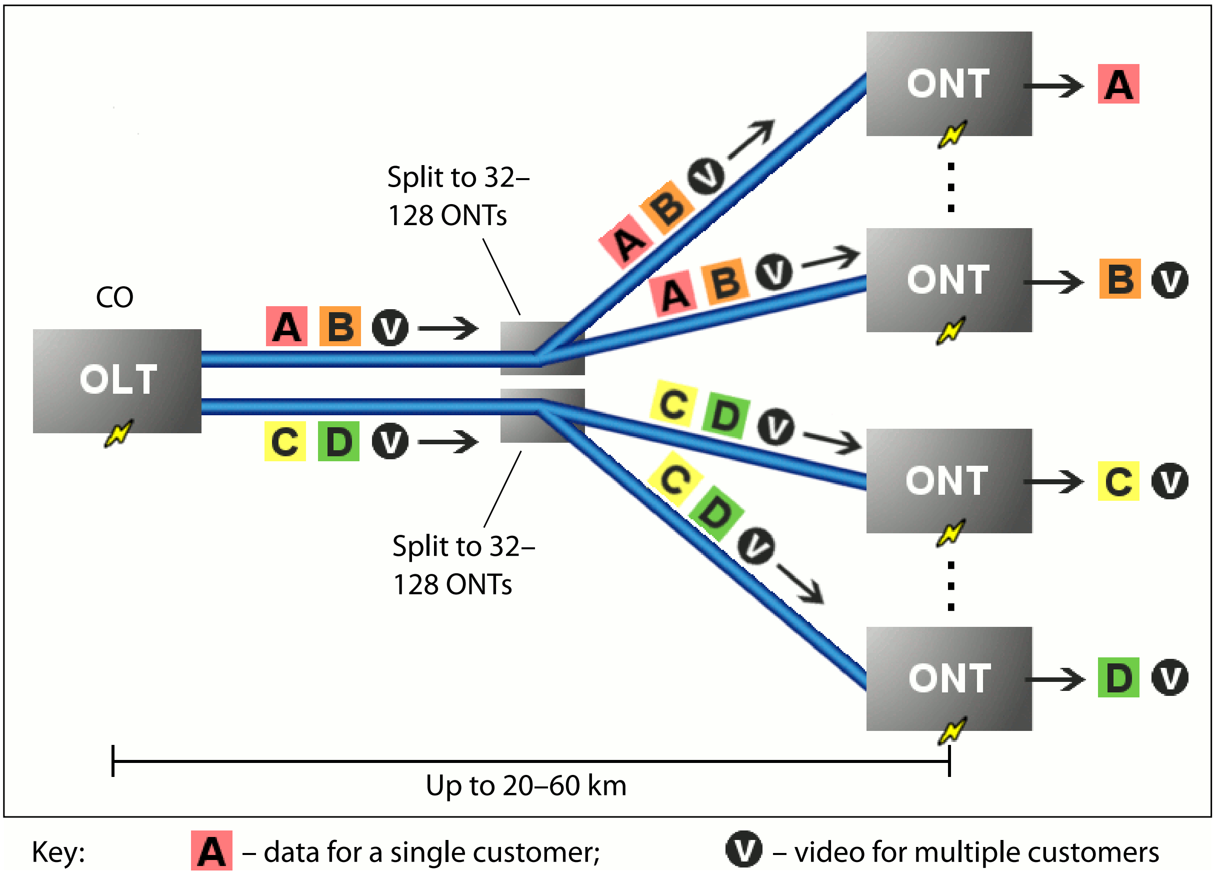


Figure 1.2: Overview of a PON (upstream communications, i.e. ONT to OLT, are omitted for clarity)*

this is not considered to be a problem. Almost all recent FTTH deployments, as well as those currently being planned, use passive optical networking.

Besides the crucial passive splitter, the key components of a PON are the Optical Line Termination (OLT) situated in the central office, at the root of the tree of fibres and controlling use of those fibres, and the Optical Network Units (ONUs)—also known as Optical Network Terminals (ONTs), with the two terms used interchangeably—situated at the leaves of the tree, in each customer's premises (or, in the case of other FTTx varieties, in the cabinet from which lead copper cables to multiple customers' premises). An overview of a simple PON is shown in Figure 1.2.

In the next chapter I describe current PON standards, their similarities and differences. The protocol issues I discuss in Chapter 3, whilst significant in any large access network, are especially serious in passive optical networks since switches are more centralised and hence larger; addressing these problems is the focus of this report.

CHAPTER 2

PON Technologies and Standards

Although the overall layout of a passive optical network is well-specified, there are several options for system design, in terms of optical components, multiplexing schemes, physical layer protocols and data link layer protocols. As previously mentioned, the latter aspect is the focus of this report; however to put this in context I will give an overview of current and future PON standards and the design choices they have made.

Broadly speaking, there are two competing families of standard currently in use—those recommended by the International Telecommunications Union standardisation sector (ITU-T) and those of the Institute of Electrical and Electronics Engineers (IEEE) Computer Society.

2.1 Brief History of PON Standards

Early PON systems [Du Chaffaut et al., 1990] were designed around (and heavily tied to) the Asynchronous Transfer Mode (ATM) protocol stack [Minzer, 1989], which was very much in favour with telecommunication companies at the time and was considered by many to be the future of all computer networking [McAuley, 1990; Leslie et al., 1993]. Such systems came to be known as A-PON, for ATM PON, and were adopted and standardised by ITU-T as Broadband PON (BPON) in the G.983 series of recommendations [ITU-T, 1998].

BPON still used ATM at its core; the change of name served only to emphasise its ability to carry non-ATM traffic by encapsulating it within ATM cells. It provided for a maximum bandwidth of 622 Mbit/s in the downstream direction and 155 Mbit/s upstream, supporting 32–64 users per PON; each user received an average downstream bandwidth of 10–20 Mbit/s. BPON saw moderate deployment, particularly in Japan (NTT had about 100k customers on BPON by 2004 [Gutierrez et al., 2005]).

ITU-T later evolved the BPON standard to form a gigabit-capable PON, GPON (G.984) [ITU-T, 2003–2008]. As well as increasing the maximum bandwidth (to 2.488 Gbit/s) and the branching factor (to 64–128 users per PON), the GPON standards recognised that ATM was losing popularity—for reasons which I discuss below in Section 2.3—and replaced it with a more flexible and efficient, albeit custom, framing protocol: GEM (GPON Encapsulation Method).

The IEEE too observed that whilst PONs in general were popular, ATM was not, and that in fact many deployments simply used ATM to encapsulate the IEEE's own data link protocol Ethernet (see Section 3.1). Furthermore, BPON and the upcoming GPON were designed with very strict timing requirements—in particular BPON allows only 154 ns to shut down the laser in one ONU which has finished transmission, to power up the laser in another ONU which is about to transmit, and for the OLT to perform gain adjustment and clock synchronisation [Kramer, 2006], and this situation only became worse with GPON which allows less than 49 ns. Whilst these timings could be achieved with modern components, they mandated the use of more expensive components in ONUs, causing a significantly higher cost per customer than would a system with greater tolerances.

As a result, the IEEE chartered the Ethernet in the First Mile Task Force to design their own PON system, EPON (Ethernet PON), which was specified in the 802.3ah standard [IEEE EFM Task Force, 2004] shortly before GPON was finalised. EPON (sometimes also called GEPON due to its capability to run at gigabit speeds) did two things differently to the ITU-T series of PON standards:

- Not only would EPON use Ethernet as its native protocol, but it would *be* Ethernet: EPON was an amendment to the Ethernet specification itself, making as few changes from the current spec as possible (which already had provision for operation over optical fibre as a native part of the physical layer specification, although not in the unusual point-to-multipoint topology used in PONs). EPON adds to the Ethernet standard two new physical layer specifications adapted for FTTH use and extends the existing medium access control (MAC) protocol.

- EPON would allow for smaller and cheaper ONU implementations by using less strict timing tolerances. Rather than GPON's 49 ns allowed for switching from one ONU to another, EPON permits up to 1.4 μs (which can be reduced by the OLT to match the timings required by the components in use).

EPON has been very successful, with 3 million lines deployed within two years of the standard's ratification [Kramer, 2006]. This success is likely to be predominantly due to the lower cost of EPON, both in terms of hardware (cheaper ONUs) and ease of deployment in an existing Ethernet environment, but timing also played a part: since the EPON standard was ready before GPON, some operators seeking to upgrade from their older BPON deployments chose EPON in order to make faster progress.

GPON has also seen trials and initial deployments by several large telcos, but it should be noted that these are largely used as a basis for transmitting Ethernet via encapsulation within GEM frames.

An update to the EPON standard to support 10 Gbit/s operation is expected to be published next month. Furthermore there are other advances in PON technology still largely at the research stage which I will describe at the end of the next section.

2.2 Multiplexing

Multiplexing is the technique of transmitting multiple independent streams of data on the same medium. On a PON this is crucial, since fibre is shared by multiple customers.

Indeed in a PON there are multiple levels of multiplexing: downstream and upstream channels must also be separated. The simplest way to achieve this is to install separate downstream and upstream fibres—sometimes referred to as Space Division Multiplexing (SDM)—but in order to save on fibre costs modern PONs support the transmission of data in both directions along a single fibre using light of different wavelengths—Wavelength Division Multiplexing (WDM).

Within these channels the data for multiple customers must be multiplexed together; current PONs typically place between 32 and 128 customers on the same branched fibre. Unusually, due to the unique asymmetry inherent in a PON, it is beneficial to consider the multiplexing schemes on the upstream and downstream channels separately; they have rather different characteristics despite superficial similarity.

2.2.1 Upstream Channel

Due to the unidirectional nature of the passive splitters/combiners used in almost all PONs, upstream transmissions from an ONU can be received only by the OLT, and not by other customers' ONUs; this arrangement is referred to as "point-to-multipoint" (P2MP). As a result, whilst the upstream connection from each ONU can be considered logically to be a point-to-point link to the OLT, ONUs' transmissions must be moderated using a medium access control (MAC) protocol to ensure that no two ONUs transmit simultaneously—or at least that if two ONUs' transmissions do collide, they notice that this has happened and recover from the situation [Zheng and Mouftah, 2005]. Ethernet was originally designed for shared-medium (broadcast) operation and has traditionally used this latter method of collision detection (CSMA/CD), but that is not possible on a PON without significant—and performance-affecting—modifications due to the inability of ONUs to detect collisions themselves [Kramer and Pesavento, 2002]. (This is one of the reasons why EPON could not simply use the previous version of the Ethernet standard *per se* and required slight protocol extensions. Furthermore, only in a P2MP network must a switch repeat a transmission it has just received back to the same physical port, in order to allow two ONUs to communicate with each other; in EPON, the Logical Link ID, or LLID, exists for this purpose [Beck, 2005].)

Both EPON and GPON therefore use Time Division Multiple Access (TDMA), informally known as "time-sharing": time is divided into slots, of either fixed or variable length and long enough to contain one or more data frames (usually around 100–1000 μs). During a given slot, one particular ONU is permitted to transmit and all others must have turned off their lasers. The OLT is responsible for determining a transmission schedule and sending that to the ONUs—this is sometimes considered to be a form of batch polling by the OLT—and the ONUs must maintain an accurate clock which is synchronised to that of the OLT in order to transmit at exactly the right time.

The number of time slots allocated to each ONU need not remain fixed; both EPON and GPON provide flexible mechanisms to allow the OLT to dynamically allocate bandwidth to ONUs according to demand and the network operator's policy [Skubic et al., 2009]. These mechanisms are nonspecific as to the algorithms employed, particularly in the case of EPON where the extremely simple request-based protocol leaves a lot of scope for interesting dynamic bandwidth allocation algorithms [Kramer, 2002; Kramer and Pesavento, 2002; Choi and Huh, 2002; Assi et al., 2003; McGarry et al., 2002; Chen et al., 2006].

Future Multiplexing Schemes

TDMA was chosen for EPON and GPON due to its simplicity and low cost; however it is not efficient as the capacity of a single channel must be shared between multiple users. Falling component prices coupled with ever-increasing bandwidth demands mean that WDM-based schemes, providing multiple channels in each direction using multiple wavelengths of light, are widely considered to be the next step in the evolution of PONs [Gutierrez et al., 2005; McGarry et al., 2006; Davey et al., 2006; Grobe and Elbers, 2008]—these were demonstrated to be possible in the early days of PON research [Monnard et al., 1997] but were at that stage prohibitively expensive for mass deployment. Largely they are still at the research stage, although KT have already deployed a WDM PON fibre-to-the-pole system in South Korea [Lee et al., 2007].

Using Coarse WDM (CWDM), it is possible to provide approximately 8–20 channels on a single fibre; Dense WDM (DWDM) can increase this to more than 100 by using more expensive components capable of narrower line widths and higher frequency stability. It is envisaged that—at least initially—the best value will be achieved by using a hybrid system using TDMA on multiple CWDM channels [Gutierrez et al., 2005]; Ogden [2010] recommends using this approach to provide 625 Mbit/s per customer by provisioning eight CWDM channels of 10 Gbit/s apiece, each shared by 16 customers.

It should be noted that whilst a traditional WDM system would require either several different types of ONU each containing a different fixed frequency laser capable only of operating on one particular channel, or expensive tunable lasers in every ONU, techniques have been devised to avoid this issue and produce relatively inexpensive "colourless" ONUs for use in WDM PONs. For example, SUCCESS-HPON [An et al., 2005] uses tunable lasers in the OLT only, transmitting continuous wave bursts which are modulated by the ONU and reflected back towards the OLT.

Beyond DWDM, there is increasing interest in the use of coherent receivers with advanced modulation schemes such as Orthogonal Frequency Division Multiple Access (OFDMA) [Narikawa et al., 2006; Cvijetic et al., 2010] in order to further increase the spectral efficiency of passive optical networks. These are still some way from being suitable for use, with potentially-major problems still to overcome—such as a significantly greater energy cost [Ogden, 2010]—but thankfully these systems are being designed with physical compatibility with existing PONs in mind, for ease of upgrading when the technology becomes mature.

2.2.2 Downstream Channel

Since the downstream channel has only one transmitter—the OLT—multiplexing is a simpler task. Downstream transmissions from the OLT can physically be received by any or all ONUs, and no collisions can ever occur.

The same basic modulation techniques are available as for the upstream channel, but they are used in a different way. In the case of EPON and GPON, the downstream channel is broadcast to all ONUs, and each frame is labelled with the address of its target ONU. That ONU will forward the frame onto its end user's LAN, and all other ONUs will discard the frame. This is—loosely speaking—a form of TDMA, with the OLT determining its own transmission schedule and each time slot lasting the duration of a frame. (The control signalling frames for the upstream TDMA—stating the required transmission schedule, etc.—will be present on the downstream channel, but is treated similarly to other frames queued for downstream transmission by the OLT.)

It should be noted that since all frames are broadcast to all ONUs, it is easy for an ONU under the control of an end user to intercept data intended for other customers. This problem and possible solutions are discussed further in Section 4.1.

As above, hybrid CWDM+TDMA and pure DWDM (and, perhaps, coherent reception) present possible upgrade paths in order to divide the bandwidth load between multiple downstream data channels.

2.3 Use of Ethernet

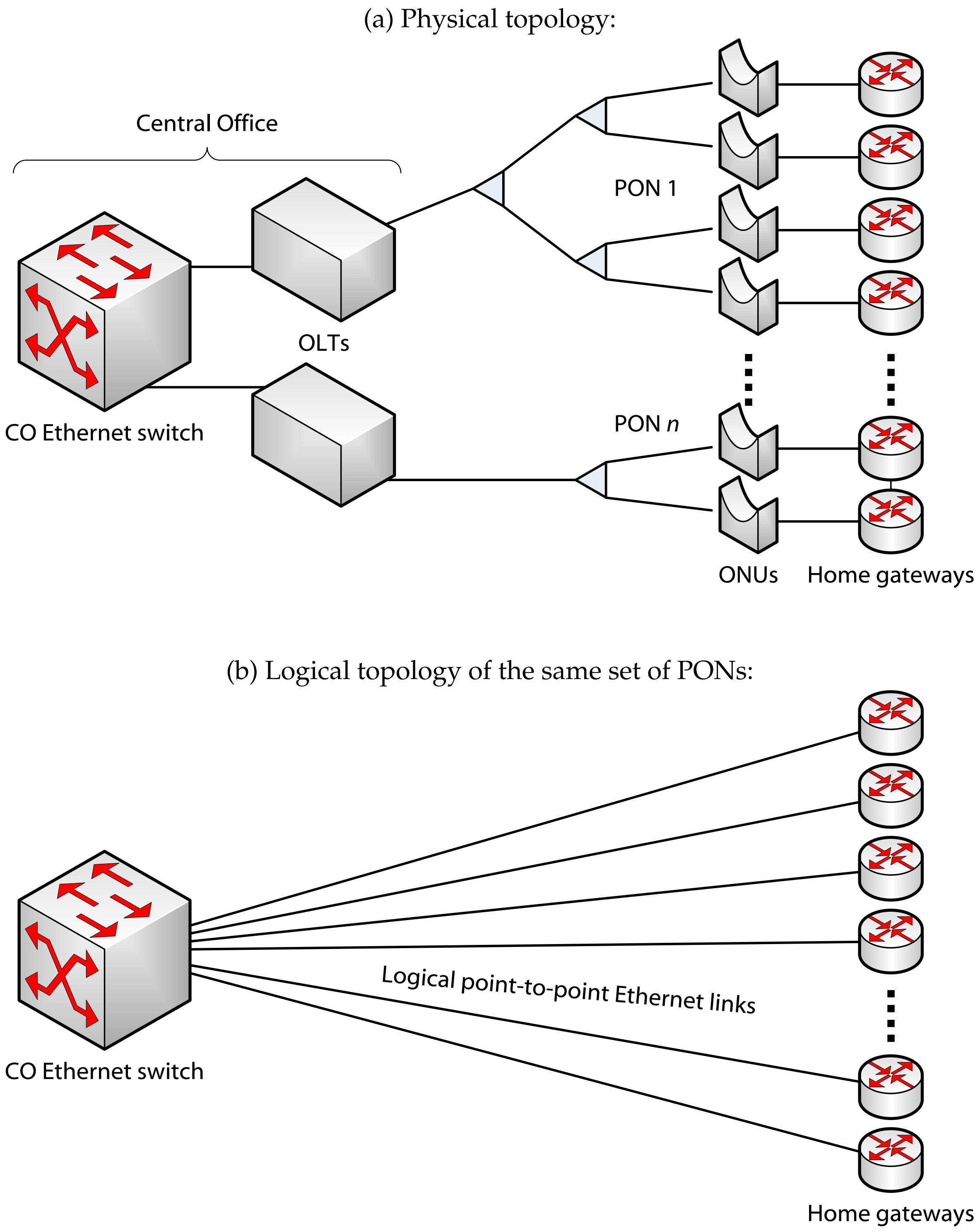
Ethernet has become ubiquitous in almost every variety of modern computer network, from small LANs within the home to large datacentres and wide-area networks (WANs). Because of this ubiquity and the protocol's simplicity, Ethernet components are available at low cost and software drivers are integrated into almost every operating system, including those targeted at embedded controllers. An overview of Ethernet is given in Section 3.1.

Telcos have historically been reluctant to deploy Ethernet in their networks, mistrusting its general-purpose, best-effort nature and preferring more complex protocols such as ATM which were designed specifically for the telecommunications sector with their demands for explicit support of real-time traffic (as needed by voice telephony) forming an integral part of the protocol [Minzer, 1989]. However Ethernet permits transmitting nodes to schedule packets according to any policy, and modern switches can enforce quality-of-service guarantees; furthermore telcos now recognise that the majority of traffic in their networks is data as opposed to voice.

Thus Ethernet has at last started to be considered a mature basis for building telcos' next-generation networks—only thirty years after its inception!—and is seeing significant deployment in this sector. BT's 21st Century Network (21CN), currently undergoing national roll-out to replace their older ATM-based network, makes heavy use of Ethernet; the project aims as one of its primary objectives to simplify both BT's core and access networks.

Therefore, as a result of its increasing uptake within the telecommunications industry, almost all modern PONs run on Ethernet at some level—whether used as the native protocol on an EPON, or encapsulated in GEM on a GPON, or even fragmented across ATM cells on a legacy BPON. Any Ethernet PON is logically a collection of point-to-point links between small Ethernet bridges in each ONU, providing connectivity into a home or business LAN, and a larger-scale Ethernet switch situated in the CO which aggregates connections from several PONs. The branching structure of the PON is transparent to Ethernet and acts merely as a way to multiplex many point-to-point links onto fewer optical fibres. The important distinction between physical and logical topologies in an Ethernet PON is illustrated in Figure 2.1.

It is clear that an Ethernet PON will lead to a very large number of logical links connected to one Ethernet switch. The implications of this are explored in the next chapter.



(a) Physical topology:

(b) Logical topology of the same set of PONs:

Figure 2.1: Physical and logical topology of a simple Ethernet PON deployment

CHAPTER 3

Scaling Ethernet

This chapter draws upon—and elaborates, with additional background material and FTTH-specific detail—work I have previously published in a workshop paper [Scott et al., 2009].

3.1 Overview of Ethernet

Ethernet was originally designed in the 1970s for use on shared-medium networks using coaxial copper cable [Metcalfe and Boggs, 1976] but it has been adapted, extended and standardised over the decades since [IEEE Computer Society, 2001]. It is now predominantly used as a basis for the creation of Internet Protocol (IP) subnets, whether or not the network concerned is actually connected to the Internet—similar to Ethernet at the data link layer, IP has become the ubiquitous network layer protocol.

Applications do not need to be aware of the underlying operation of Ethernet; as far as the application developer is concerned, communication takes place between IP addresses—or higher-layer identifiers, such as DNS host names, which are resolved to IP addresses—and not Ethernet's MAC addresses. As a result there is a need for automatic conversion from IP to MAC addresses. Each higher-layer protocol suite has its own protocol to bridge this gap; in the case of IP version 4, that protocol is the Address Resolution Protocol, ARP [Plummer, 1982]. Although this is not part of Ethernet, it is instructive to bear in mind how it operates.

In order to determine the MAC address which should be used for a given IP address, ARP sends a broadcast query to all hosts on the relevant IP subnet—corresponding, usually, to a single Ethernet. In cases where communication is taking place between hosts in multiple subnets via a router, the router is responsible for performing ARP queries as the querier must be in the same subnet as the target host. The host claiming the queried IP address will reply stating its MAC address. The querying host will then store this mapping in its ARP cache. Entries in the ARP cache are typically held for a number of minutes (the exact cache lifetime varies between operating systems); it is expected that the MAC address will not change and cache entries expire mainly in order to free up memory.

Despite its shared-medium roots, Ethernet is now predominantly used to build packet-switched networks comprising point-to-point links between two switches—strictly, IEEE 802.1D Media Access Control (MAC) bridges [IEEE Computer Society, 2004a]—or between a switch and a single host. Ethernet networks can form a variety of topologies, the most common being a tree (although as it stands, non-tree topologies are not used efficiently: see Section 3.3.2).

Although Ethernet is usually referred to as a protocol, it is more accurate to describe it as a family of closely-linked protocols; different parts of Ethernet operate at the physical and data link layers of the OSI network model [ITU-T, 1994].

The physical (PHY) layer is specified separately for different types of medium operating at a variety of speeds; common Ethernet PHY standards in use today include 1000BASE-T (Gigabit Ethernet over Category 5/5e copper cabling), 1000BASE-SX (Gigabit Ethernet over a pair of short multi-mode optical fibres at 770–860 nm), 1000BASE-LX (Gigabit Ethernet over multi- or single-mode fibres at 1270–1355 nm, with a longer range) and 10GBASE-LR (10 Gigabit Ethernet over single-mode fibres at 1310 nm), with 40- and 100-Gigabit versions standardised recently. During the development of EPON, the Ethernet in the First Mile Task Force specified two additional PHY variants specifically targeting the requirements of PONs (but also available for conventional, non-PON use): 1000BASE-LX10 (a variant of 1000BASE-LX which is capable of longer-distance transmission, but which still requires separate fibres for transmission in each direction) and 1000BASE-BX10 (using a single fibre and two different wavelengths of light—1310 nm and 1490 nm—for bidirectional use).

The Ethernet PHY is responsible for providing a serialised bitstream facility (only) to the Medium Access Control (MAC) layer.

3.1.1 MAC: Addressing and Switching

The MAC is responsible for dividing the bitstream into frames; frames are labelled with a header containing, amongst other attributes, source and destination MAC addresses—thus enabling the statistical multiplexing of multiple hosts' frames on a single link.

Every Ethernet interface is assigned a unique, 6-byte MAC address at the time of manufacture. This address is formed of three bytes identifying the device's manufacturer—using an Organisationally Unique Identifier (OUI) assigned by the IEEE Computer Society [2001, §9]—with the remainder assigned by the manufacturer. It is also possible to override the manufacturer-assigned MAC address according to some local scheme; one bit in the first byte acts as a flag to indicate such a locally-administered address, or LAA (this bit is set to zero in every manufacturer-assigned address).

Switches make use of MAC addresses in order to bridge together multiple point-to-point or shared-medium Ethernet segments. When a frame passes through a switch, the switch learns the location of the sender; the source address of the frame is stored in a *forwarding database* in the switch's memory together with the interface on which the frame arrived [IEEE Computer Society, 2004a, §7.8–7.9]. This is used to direct subsequent frames: the switch looks for frames' destination addresses in the database in order to determine the interface to which the frame should be forwarded. As a fall-back mechanism, if the switch has no record of the location of a particular address, the frame can be flooded to all interfaces—but since this is very wasteful of link capacity, the intention is that this will happen seldom or never for unicast frames.

MAC addresses can also refer to—using another flag bit—groups of multiple hosts; currently Ethernet does not natively provide multicast routing, generally using broadcast for all group addresses, but some switches can use a technique known as IGMP snooping [Christensen et al., 2006] to hook into IP multicast and infer Ethernet multicast groups.

3.2 Ethernet Size in an Access Network

As discussed in the previous chapter, a single PON will be seen by an Ethernet switch as a collection of point-to-point links, one per ONU. A modern PON will typically connect up to 128 ONUs to each OLT: higher branching factors do not significantly reduce the amount of fibre which must be installed but nevertheless reduce the bandwidth available to each customer, so 128 is the optimum [Ogden, 2010]. Building an Ethernet switch with 128 downstream links presents no difficulty; conventional access switches built from modular components frequently have more than 128 physical ports today.

However the OLTs must be interconnected with each other and with the core network, and it is convenient to use an Ethernet switch in the Central Office for this pur-

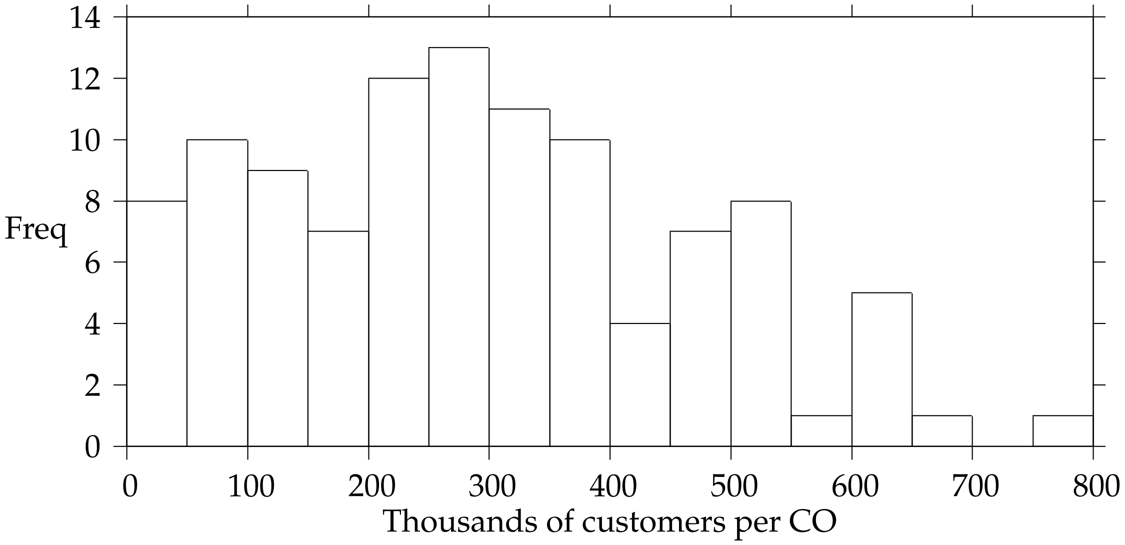
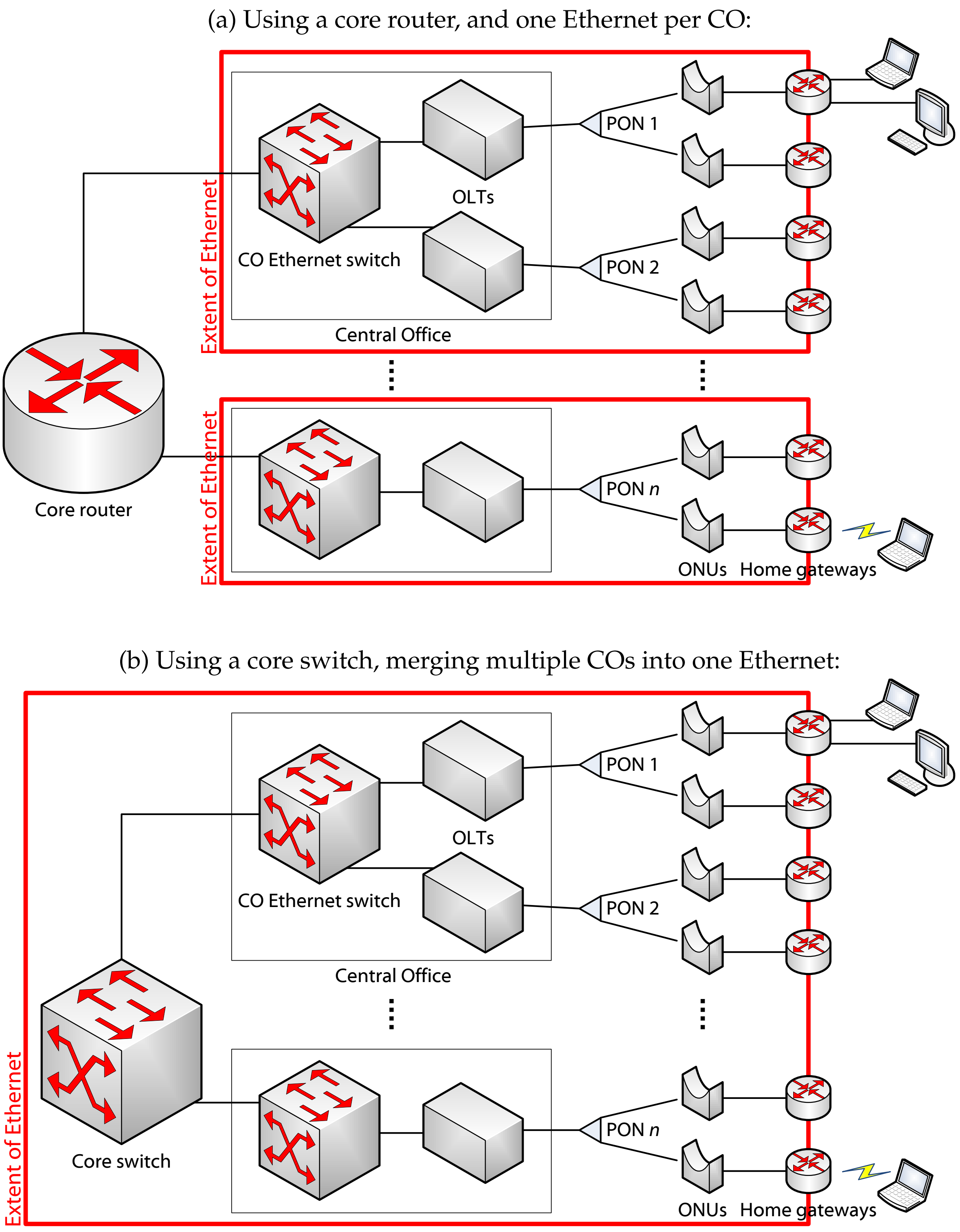


Figure 3.1: Histogram of estimated number of customers per BT 21CN metro node (data from [Ogden, 2010], derived from the number of distinct postcodes served by each node, assuming the average number of customers per postcode to be 16)

pose. The alternative—installing large IP routers in COs in order to interconnect the hundreds or thousands of per-OLT Ethernet networks—is considerably more expensive in terms of equipment cost, power usage and management complexity. Modern OLTs are designed specifically for this mode of operation, with OLTs available in the form of a GBIC module suitable for insertion into a modular switch chassis [Kramer, 2006].

Thus, by bridging together all of a CO's PONs at the Ethernet layer, every customer served by these PONs will be on a single large Ethernet. In order to gain a picture of the scale of these Ethernet networks, data available on BT's 21CN topology can be used: Figure 3.1 shows an estimate of the distribution of customers per "metro node"—BT's term for a central office linking the access network to the core. The range of metro node sizes is high: the smallest handles approximately 15,000 customers whereas the largest handles almost 800,000.

Ethernet could indeed be pushed further into the core of the network: even large scale Ethernet switches remain more cost-effective than high-speed IP routers. Figure 3.2 summarises the two main options for deploying Ethernet switches and IP routers on a FTTH network, and illustrates the extent of the resulting Ethernet switching domain. If the core network as well as the access network were Ethernet-based, the constituent Ethernet switches would be responsible for managing communication between several million devices.



(a) Using a core router, and one Ethernet per CO:

(b) Using a core switch, merging multiple COs into one Ethernet:

Figure 3.2: Network topology options affecting Ethernet bridging domain size

It should be noted that although Ethernet is also commonplace within the home, my assumption here is that home LANs will not be bridged into the access network, with a simple router appliance (a "home gateway") separating the two Ethernets and forwarding packets across the boundary at the IP layer (with the ONU and/or OLT enforcing this by permitting only one Ethernet device—the gateway—to connect). This has security and manageability benefits both for the telco and for the home user and is the approach used today by most existing layer-2 access networks such as cable broadband ISPs. If this were not done, for whatever reason, the resulting Ethernet would be larger still.

3.3 Scalability Problems

Ethernet has a few barriers currently preventing its operation on networks of this scale. These problems are usually described in terms of their applicability to very large datacentre networks but the issues apply similarly to FTTH.

3.3.1 Forwarding Database

One of the more significant problems is the limited capacity of switches' forwarding databases. This database must be stored in very fast memory in each switch as it must be referred to for every forwarding decision: on a gigabit or ten-gigabit switch containing hundreds of interfaces, this could require millions or even billions of lookups from this database per second. On high-speed switches the database is stored in a content-addressable memory (CAM) as this is the only way to provide the performance needed; increasing the capacity of a CAM whilst constraining energy consumption without sacrificing speed is proving to be challenging [Yu et al., 2005; Pagiamtzis and Sheikholeslami, 2006].

Consequently, in modern switches the capacity of the address database is generally of the order of 16,000 entries [3Com Corporation]. Due to hosts' frequent use of broadcast frames for signalling, it is very likely that every host will appear in the address databases of most switches, therefore one should consider the minimum address database capacity of any switch on an Ethernet network to be the maximum number of nodes possible on the network. An Ethernet with more nodes than this will perform very poorly—at best, frames will be flooded to all links, which may saturate the capacity of links at the edge of the network with traffic not intended for that link at all. In extreme cases, the network could completely fail to provide any useful throughput.

Recall that on BT 21CN, metro nodes serve between 15,000 and 800,000 customers (Figure 3.1). Consequently, for the moment assuming the core is routed rather than switched, an Ethernet switch in even the smallest metro node would be close to filling its forwarding database if FTTH were to be deployed to all customers; the largest node would be several times over the limit for a single Ethernet switch.

3.3.2 Non-tree Topologies

Ethernet is unable to make efficient use of networks containing loops, which also presents a scalability problem. If an Ethernet network were to contain a loop, frames would be forwarded around this loop indefinitely, using up all available capacity, since frames do not keep track of the number of times they have been forwarded (as IP packets do, using a Time-to-Live counter which is decremented by each router). Ethernet handles this situation by invoking the Rapid Spanning Tree Protocol, RSTP [IEEE Computer Society, 2004a, §17], which removes loops by disabling any redundant links and converting any topology into a tree. Dense meshes with a high degree of interconnection—like those found in telcos' core networks—would find a large proportion of links disabled entirely if they ran Ethernet; this constrains frames to suboptimal routes and may introduce bottlenecks, particularly around the root of the spanning tree.

3.3.3 How Does BT Cope?

BT chose Ethernet for 21CN, which will be used for all of their voice and data services, and have therefore had to work around these scalability problems. In order to achieve this, they have used two key technologies. Firstly, in order to avoid the spanning tree problem in the core, they have deployed MPLS label edge routers (LERs) in each metro node, making the core of 21CN a MPLS cloud which provides a Virtual Private LAN Service (VPLS) [Rosen et al., 2001]. VPLS connects the Ethernet islands of the access network together through tunnels across the MPLS core.

MPLS LERs are power-hungry and expensive, and since MPLS and VPLS have scalability problems of their own, this approach likely only postpones the problem. MPLS works by adding one or more labels to the start of every frame, i.e. encapsulating the frame inside its own protocol. To provide VPLS, the LERs must determine each frame's initial label(s) based upon its destination address via a lookup table. Frames follow prenegotiated label-switched paths (LSPs) which are precomputed at connection setup time and the relevant next hop for each LSP is stored in a lookup table on each intermediate MPLS switch. Each switch must hence use each frame's label to index into this lookup table to determine how to switch the frame.

The effect is to provide a large Ethernet network transparently overlaid on the MPLS cloud. Whilst this solves the problem of shortest-path routing across the core, the overlay network is still susceptible to most of the usual Ethernet scalability problems—and in fact VPLS adds further large lookup tables on every core label switch router (LSR) that can in some configurations scale even more badly than Ethernet's forwarding databases: LSRs must store the next hop for every LSP in which they participate, which in the core of an unrestricted network could scale as $O(\text{hosts}^2)$.

Secondly, BT have deployed Provider Backbone Bridge Traffic Engineering, PBB-TE (standardised by the IEEE Computer Society as 802.1Qay [2009]). This standard aims to make Ethernet more deterministic and to avoid address database scalability problems—but at the cost of significantly crippling the self-managing nature of Ethernet. PBB-TE does away with RSTP; the network must be manually constrained to follow a tree topology. It also disables switches' ability to learn MAC addresses; address databases must be centrally provisioned. In effect, Ethernet switches in PBB-TE are reduced to dumb frame relays following centrally-managed rules; switches must be reconfigured from a central management system every time the network topology changes, for example in the event of a cut fibre. PBB-TE also adds the VPLS-like capability to encapsulate Ethernet inside another layer of Ethernet (MAC-in-MAC), which suffers the same problems as VPLS regarding the lookup tables required for encapsulation that I described above.

In short, the industry-standard solutions to the poor scalability of Ethernet are clumsy workarounds which leave little of Ethernet's decentralised operation. Current research aims to do better to increase the scalability of Ethernet without adversely affecting its utility.

3.4 Ethernet's Underlying Problem

Ethernet's poor scalability arises in various guises, as outlined above. It would seem at first glance that these are entirely distinct and unrelated. However, there is a common underlying cause: that ***MAC addresses provide no location information***.

Although globally-unique MAC addresses have a hierarchy of sorts—in that they start with an OUI—this exists solely for the purpose of allocating unique addresses in a decentralised fashion, and is of no use to Ethernet switches which must treat the unicast address space as flat.

A flat address space has the advantage that no configuration of devices is required; a device can use its unique, manufacturer-assigned MAC address anywhere on any network. However, this leaves each switch with the task of discovering and storing the location of every addressable device. If the MAC address space were not flat, but instead contained enough information to locate the device possessing the address, several advantages would be gained.

Firstly, large forwarding databases would no longer have to be maintained on every switch. This location information could instead be distributed across the network so that frames are directed towards their destinations according to successive stages of a hierarchy.

Secondly, a hierarchical MAC address space would make the addition of prefix-based shortest-path routing possible. Flat addressing does not lend itself to easy routing: any address can be located anywhere on the network. The use of hierarchical addresses, with each switch handling a block of sequential addresses akin to an IP subnet, would reduce the routing problem to the one that existing routing protocols already solve for IP.

It is possible for network administrators to assign hierarchical addresses to devices manually, as a LAA (see Section 3.1.1). However, configuring and maintaining a LAA on every device based upon where it is connected would be a considerable and unwelcome administrative overhead.

3.5 Addressing Scalability with MOOSE

I have developed MOOSE (Multi-level Origin-Organised Scalable Ethernet), a novel system for applying hierarchical addressing to an Ethernet transparently and without any configuration to edge devices.

The basic operation of MOOSE is to assign a new hierarchical MAC address to each host on the network; in a PON, the "hosts" will in fact be home gateway appliances acting in their capacity as simple IP routers, as well as the large core IP routers providing connectivity to the Internet. This address is assigned automatically and dynamically from the space of unicast locally-administered MAC addresses, and is referred to as a *MOOSE address* to avoid confusion with hosts' static, manufacturer-assigned MAC addresses.

Every frame entering a MOOSE-enabled Ethernet network has its source address rewritten in-place to the sending host's MOOSE address by the first MOOSE-aware switch it traverses. The switch that performs address rewriting for a host—i.e. the closest MOOSE switch to that host—is the host's *home switch* and is responsible for assigning a MOOSE address to that host. (If non-MOOSE switches or hubs are in use, a host may have more than one "closest" MOOSE switch, in which case an RSTP-like protocol is used to elect a switch to handle the hosts on each segment.)

The destination address is left intact in the expectation that it already is a MOOSE address. This assumption is valid since hosts' ARP caches will already contain the MOOSE addresses of any hosts being communicated with; any frame received by any host will already have had its source address rewritten. A host's manufacturer-assigned MAC address is never seen beyond that host's home switch. This is a crucial point since encapsulation-based technologies such as MPLS do not reveal to the destination host the address used for routing; as a result, switches must also convert destination as well as source addresses of frames entering the network. In other words, switches still need to maintain large databases of remote hosts on the network. The only destination rewriting that MOOSE switches perform, however, is of frames destined for local hosts, setting the destination address back to the host's manufacturer-assigned MAC address; this is simple as the required information is already known by that switch, and necessary because otherwise that host's network interface card would discard the frame as misaddressed.

3.5.1 Hierarchical Address Structure

A MOOSE address consists of a *switch identifier* followed by a *host identifier*. Since these two identifiers when concatenated must form a valid unicast LAA MAC address, in order to remain compatible with Ethernet and avoid any conflict with manufacturer-assigned or group addresses, the settings of two bits in the first byte of the switch identifier are fixed: the least significant bit must be 0 to indicate a unicast address, and the second-least significant bit must be 1 to indicate a LAA.

For the examples given here, I use the simplest case of a three-byte switch identifier followed by a three-byte host identifier. The switch ID could however have a variable length, and/or an internal hierarchy of its own—for example six bits to identify a network area followed by two bytes to identify a switch within that area—which could then be used to further aid routing decisions.

Each host (or router) is assigned a host identifier by its home switch from the pool

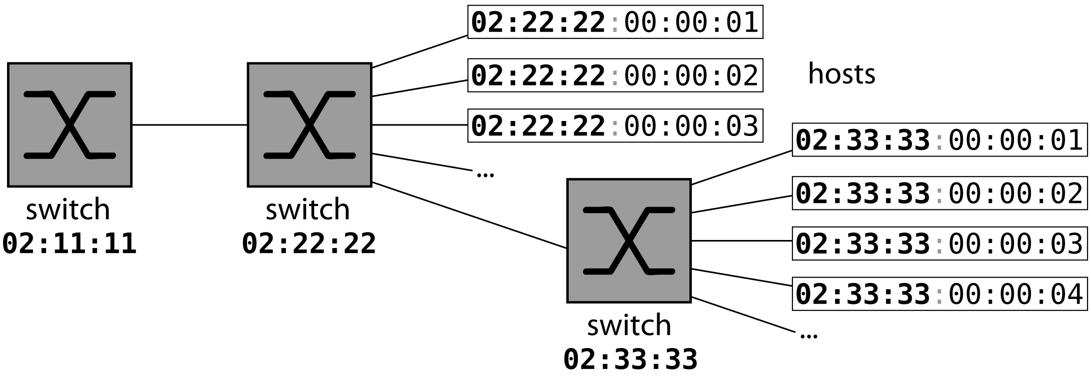


Figure 3.3: Assignment of MOOSE addresses by switches

of identifiers available to that switch; see Figure 3.3. Only a host's home switch ever bases a forwarding decision on the host identifier, so the detail of how these are allocated can vary from switch to switch—either chosen by the manufacturer according to implementation-specific optimisation, or by the network operator according to local policy. Suitable schemes for an EPON include:

- sequential assignment;
- an identifier for the OLT followed by a condensed form of the logical link ID (LLID) which identifies a particular endpoint on the PON;
- a hash of the host's real MAC address;
- an ONU serial number;
- an identifier for the customer with whom the ONU is located.

Sequential assignment has the disadvantage that it presents a potential denial-of-service attack in which a malicious host attempts to use up all available host identifiers on a switch; in the case of an EPON it is likely that a limit of one host identifier per logical link will be enforced to mitigate this possibility, but avoiding sequential assignment will provide an extra level of security. Deterministic schemes such as the latter three have the added advantage that they do not require dynamic state and hence can be recovered easily in the event of a switch reboot.

It is hence possible to route frames through the network to remote hosts by simply inspecting the switch identifier in the frame's destination address, and ignoring the host identifier until the frame reaches the destination host's home switch. Switches no longer need to keep a database of all hosts' MAC addresses; they only need store the locations of other switches and of any directly-connected hosts.

3.5.2 Shortest Path Routing

As described so far, MOOSE switches must still forward frames along a spanning tree, disabling any redundant links. As discussed in Section 3.3.2, this is generally an undesirable property. An EPON is conveniently already a tree structure, so the spanning tree problem does not present itself at the edge of the network, but it is likely to severely affect the highly-interconnected core. With MOOSE, the foundations are in place to do much better than this using shortest-path routing.

For the purpose of frame forwarding, a MOOSE switch can be considered akin to an IP router. A router's forwarding database lists the location of each subnet—i.e. address prefix—in order to direct frames onwards. Each MOOSE switch has one local "subnet", containing all addresses starting with its switch identifier, and handles frames for remote subnets by passing them to the most appropriate neighbour. Bearing this in mind, switches can run a routing protocol in order to distribute subnet information to other switches. The protocol ensures that switches always have up-to-date information about the shortest path to every host, and can therefore route frames directly towards their destinations, rather than constraining them to a spanning tree.

3.5.3 Broadcast and Multicast

MOOSE must still support arbitrary broadcast frames for compatibility; these need to be forwarded along a spanning tree in order that they reach each host exactly once. An explicit spanning tree protocol is not required however, as the tree can be deduced in a distributed manner from the routing table using the technique of reverse-path forwarding; multicast routing protocols such as PIM [Adams et al., 2005] use this technique today on IP networks.

Indeed, MOOSE switches can make use of multicast features of their routing protocol in order to provide a native Ethernet multicast facility. This could be used to great advantage on fibre access network—for example, if a live television service is provided, multicast would eliminate the need to separately stream the channel to every customer by providing the ability for a single packet to have multiple destinations, thus freeing up a significant quantity of capacity upstream of the OLT.

3.5.4 Example

To illustrate the basic behaviour of MOOSE switches, a simple example will be used which will describe the steps involved in forwarding a broadcast frame containing

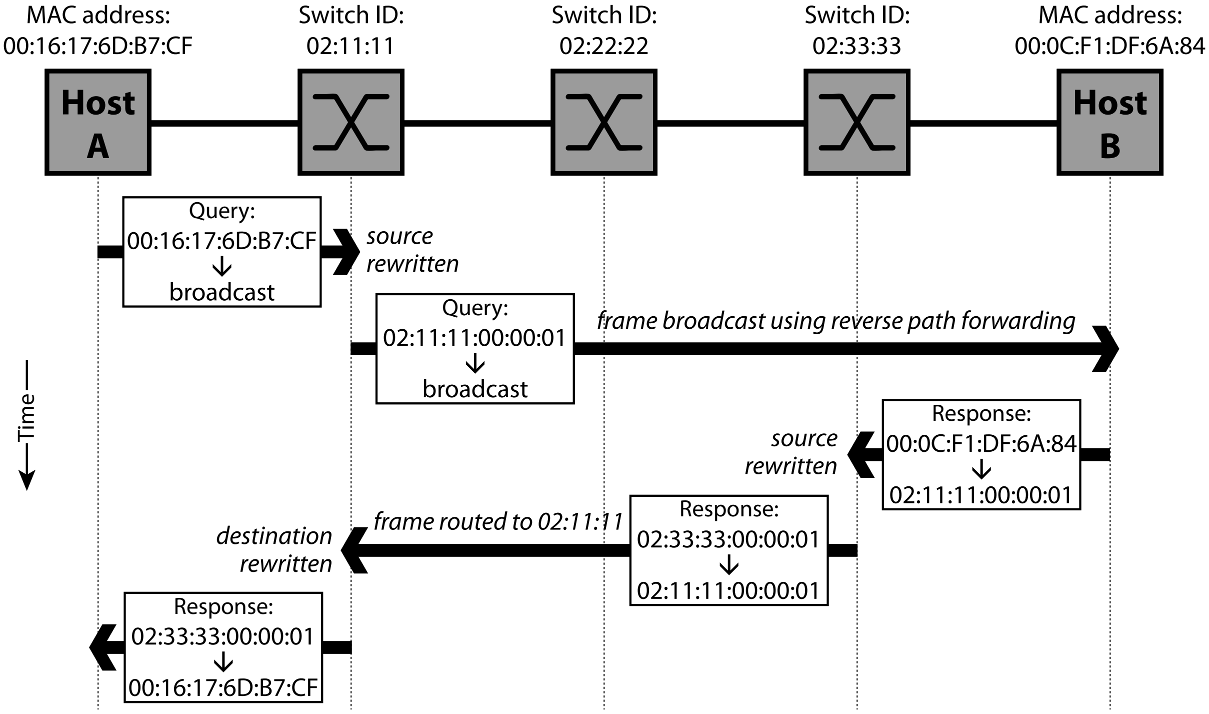


Figure 3.4: Sequence diagram of a broadcast query and subsequent unicast response

a query in some higher-layer protocol, and subsequent unicast frame containing the response, between two hosts A and B via three MOOSE switches 02:11:11, 02:22:22 and 02:33:33. This is summarised in Figure 3.4.

Query

1. Host A transmits the broadcast query frame as it would on any Ethernet network, with its own manufacturer-assigned MAC address in the Ethernet header's source field and the standard broadcast address in the destination field.

2. The frame is received by switch 02:11:11, which observes the non-MOOSE address in the frame's source field, and rewrites the source field into a MOOSE address containing the switch identifier and the appropriate host identifier. As this is Host A's first frame, the switch must allocate a host identifier (in this case 00:00:01, making Host A's complete MOOSE address 02:11:11:00:00:01).

3. The three switches broadcast the frame using reverse path forwarding away from Host A.

4. The frame is received by Host B (and any other hosts on the network) in its current form; no further rewriting is performed.

Response

1. Host B looks up Host A's IP address in its ARP cache to determine a suitable destination address for the response frame. Since the rewritten query frame arrived at Host B with the source field containing the MOOSE address 02:11:11:00:00:01, this is the address returned by the cache lookup.
2. As above, switch 02:33:33 assigns a MOOSE address to Host B (02:33:33:00:00:01) and rewrites the source address of the frame.
3. The frame is now routed through the network based solely on the destination switch identifier—the host identifier is ignored for now. The routing table is consulted for the location of switch 02:11:11 and the frame is forwarded accordingly.
4. On receiving the frame, switch 02:11:11 observes that it is destined for a directly-connected host (02:11:11:00:00:01). It prepares the frame for transmission along its final hop by rewriting the destination address to Host A's manufacturer-assigned MAC address. The source field of the frame is again left as the MOOSE address of Host B in order that this address is used for any further communication with Host B.

3.5.5 Mobility

A consequence of introducing location-based hierarchy into MAC addresses is the need to explicitly handle hosts which move from one location in the network to another. In a traditional Ethernet, hosts can migrate between switches and the host's new location will be learned as soon as it sends a frame. With MOOSE, if a host relocates to a new switch its address changes and any ARP cache entries on other hosts pertaining to the migrated host become incorrect; frames will continue to be sent to the host's old location for a while.

In the FTTH network topologies proposed thus far, mobility is unlikely to ever take place without a home gateway being powered down, unplugged, and replugged in a new location; in this case, it can simply be treated as a new device when it starts up in its new location. However one of the benefits of an Ethernet PON is the flexibility to deploy new services in the future, and mobility may start to become an issue at that point—in particular if wireless access points are ever directly connected to ONUs.

With this in mind there are two strategies for dealing with mobility, as illustrated in Figure 3.5, which can be used separately or in conjunction:

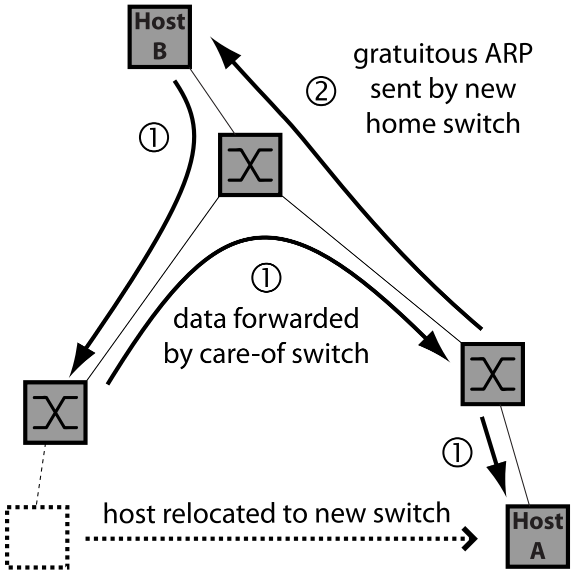

Host B
gratuitous ARP sent by new home switch
②
①
①
data forwarded by care-of switch
①
host relocated to new switch
Host A

Figure 3.5: Two ways to handle a host A roaming onto another switch whilst maintaining communication with another host B

1. The previous home switch of the migrated host can forward frames sent to the host's old address until outdated ARP cache entries expire.

   This is similar to IP Mobility [Perkins, 2002]; however, unlike IP Mobility it requires no host support. A handover protocol is necessary for the old and new home switches to set up such forwarding: on arrival of a new host at a switch, that switch would ask all other switches (via multicast) whether any had seen this host before, identifying it using its manufacturer-assigned MAC address, and it would instruct such switches to redirect frames.

2. A broadcast ARP announcement (or "gratuitous ARP") can be sent by the new home switch to immediately update remote ARP caches with the new MOOSE address.

   This is the technique used by Xen when migrating live virtual machines [Clark et al., 2005]. This is a simple approach as a handover protocol is not required, and as a result this works even if the previous switch is no longer reachable—for example if this host migration happened as a result of a switch failure. However it does results in additional broadcast traffic.

Unless the frequency of host migrations is very high—which, as stated above, is unlikely on a PON—the additional load introduced by either mobility approach is expected to be negligible.

3.6 Other Solutions

There has been other recent work on improving the scalability of Ethernet, usually focusing on one particular aspect of the problem. So far only MPLS-VPLS has been widely adopted, and this has significant problems of its own as described above.

SmartBridge [Rodeheffer et al., 2000] and Rbridges [Perlman, 2004] both encapsulate Ethernet frames in a new inter-switch protocol, and run a routing protocol between switches. In this regard they are similar to MOOSE, but they do not impose an addressing hierarchy; the routing protocol therefore advertises every individual MAC address to other switches, which must again build up a database containing all hosts on the network. Rbridges is currently undergoing standardisation by the Internet Engineering Task Force TRILL working group, aiming explicitly to deal only with the problem of shortest-path routing—the authors acknowledge that this does not increase the maximum size of an Ethernet network [Touch and Perlman, 2009].

Myers et al. [2004] suggested that Ethernet's main failing is its broadcast service, and propose a new architecture in which hosts make explicit use of directory services operated by switches rather than broadcasting queries. Whilst this is a valid point, and the problem of generalised broadcast traffic reduction is not one which is addressed by MOOSE, the modifications to Ethernet suggested by Myers et al. are not backwards-compatible and would require at least software modifications to all connected devices. Ethernet is, perhaps unfortunately, too widespread for this to be practical. This difficult issue is discussed further in Section 4.2.

SEATTLE [Kim et al., 2008] takes a more scalable approach. A routing protocol is again operated between switches, but in contrast to the approaches described above and in common with MOOSE, the routing protocol only propagates switch location information rather than every MAC address on the network. Manufacturer-assigned MAC addresses are still used, and thus a mechanism is required to determine the switch to which a given address is connected. This is achieved in SEATTLE using a distributed hash table—an algorithm which distributes part of the database to each participating switch in a deterministic fashion, so that each switch knows which other switch to ask for the location of a given host. Unfortunately the algorithm introduces considerable complexity to switches, and it is likely that a high-speed SEATTLE switch would be difficult to implement in hardware.

Chapter 4

Conclusion and Future Directions

In this report I have reviewed the current state of FTTH and PON research and development, noting in particular the convergence on Ethernet as a data link layer protocol and switching technology. I have identified the scalability barriers which will affect telcos undergoing large-scale deployment of FTTH, and present some current computer science research—including my own Ethernet extension, MOOSE—which could be used to address these problems.

PON deployment has already begun worldwide, working around some of these scalability issues by combining a variety of legacy technologies, but these just postpone the problems; current FTTH deployments are still small compared to POTS networks, and protocol changes will be required in order to deploy fibre to everywhere currently served by a copper telephone line. However a benefit of a well-designed PON is upgradability: some of the early PONs in Japan and the US have already been upgraded from BPON to GPON or EPON. Roll-out of MOOSE to an existing PON could simply be a matter of applying software upgrades to ONUs and OLTs.

It is impossible to address every potential problem which might arise on an Ethernet PON. There are a few known problems yet to be solved, two of which I will outline here—although one can only conjecture at this stage on the issues which may reveal themselves after another decade of Internet application evolution when, perhaps, every home will have thousands of networked devices and multi-gigabit connectivity.

4.1 Ethernet Security

Ethernet is a simple protocol, designed at a time when networks were used almost solely for research and security issues were not important. The standard behaviour of an Ethernet switch when faced with a situation in which it cannot switch a frame correctly is to flood the frame to all ports, thus allowing a user to obtain frames intended for other users, and perhaps also saturating slower links. It is relatively easy for a single host to trigger this failure mode; one common method is to fill up switches' address databases by sending frames purporting to be from a vast number of imaginary hosts with randomised source MAC addresses.

Modern switches can mitigate this by limiting the number of MAC addresses which may be associated with a particular interface, or by requiring each MAC address to authenticate [IEEE Computer Society, 2004b] before it can start transmitting. (In the case of a PON, the ONU is managed by the network operator, but restrictions should still be imposed by the OLT switch as nothing prevents the customer from plugging the fibre into his own custom PON.) A protocol like MOOSE is fundamentally invulnerable to this kind of attack, though, as the set of switch identifiers participating in MOOSE switching is controlled by the network operator.

However, receiving some other users' frames is a by-product of normal operation of some types of network—for example on a non-WDM PON, all downstream frames are received by every ONU anyway. It is generally considered advisable for hosts to encrypt any data that they do not wish for hosts other than the intended target to read. This is usually done in a higher-layer protocol such as HTTPS or IPsec, but there are efforts to bring security into the MAC layer. The core cryptography elements of a MAC Security standard have been produced by the IEEE Computer Society as 802.1ae [2006] but on its own this is of little use; work is ongoing to develop the associated OA&M protocols such as those for key management (802.1af).

Most EPON deployments currently use proprietary security solutions due to unavailability of a suitable standardised solution [Kramer, 2006]—this is suboptimal and standardisation of a sufficiently flexible security platform will lead to lower costs and better security.

4.2 Use of Broadcast

A problem whose solution has seen rather less progress is that of broadcast traffic. Not only does Ethernet flood frames destined for unknown hosts, but it also uses—and encourages higher-layer protocols to use—broadcast for control messages. Two of the most significant participants are ARP [Plummer, 1982], which performs address resolution via broadcast queries, and DHCP [Droms, 1997] which uses broadcast messages for automatic configuration, but there are many other more minor protocols which make use of Ethernet's broadcast facility. The amount of broadcast traffic on modern Ethernet-based networks is rising, and the future looks bleak: it has been shown that ARP traffic alone could reach hundreds of megabits per second on a million-node Ethernet network [Myers et al., 2004].

Resolving this problem in the general case—by converting every conceivable broadcast-based protocol on-the-fly to a non-broadcast mechanism—is extremely difficult, if not impossible. One potential approach could be to have switches inspect broadcast packets and imply in a generalised manner the host or hosts which are likely to be interested in receiving the packet, for example by tracking which hosts are participating in a particular (unknown) protocol. The addition of a native Ethernet multicast facility such as that provided by MOOSE, or genericised VLAN (virtual LAN) techniques as proposed for SEATTLE by Kim et al. [2008], could be used as a stepping-stone leading towards this goal by redirecting broadcast packets to implied MAC multicast groups.

A less ambitious—yet rather more realistic—approach would be to reengineer the most common sources of broadcast traffic, such as ARP (and its IP version 6 equivalent, Neighbour Discovery). There has been very recent interest within the Internet Engineering Task Force in starting a working group to bring together possible ways to do this and to develop them further. The problem is still hard, since it is unreasonable to expect all devices to switch to a newer version of ARP or ND; it is likely that switches' participation will be required in order to convert queries to a newer, non-broadcast protocol, likely involving a distributed directory service which would allow switches to answer ARP queries themselves.

A MOOSE-specific alternative would be to take advantage of the new, hierarchical MAC addresses: these could be assigned in such a way as to map directly and deterministically onto the IP address space, allowing trivial conversion by any MOOSE-aware device between an IP address and the corresponding MAC address, doing away with the need for ARP or an ARP-like protocol entirely. This may impose excessive restrictions on how a MOOSE network is deployed, however.

4.3 Acknowledgements

I am grateful to my colleague Peter Ogden for providing the data I used in Section 3.2 and for useful discussions on PON scalability at the physical layer.

Finally, my thanks go to Prof. Jon Crowcroft for first inspiring me to work on Ethernet, and to my supervisors Prof. Polina Bayvel and Dr. Benn Thomsen for giving me the opportunity to continue this work in a new context.

Bibliography

3Com Corporation. Switch 5500G 10/100/1000 family data sheet. URL `http://web.archive.org/web/20070709222433/http://www.3com.com/other/pdfs/products/en_US/400908.pdf`.

A. Adams, J. Nicholas, and W. Siadak. Protocol Independent Multicast - Dense Mode (PIM-DM): Protocol Specification (Revised). RFC 3973 (Experimental), Jan. 2005. URL `http://www.ietf.org/rfc/rfc3973.txt`.

F.-T. An, D. Gutierrez, K. S. Kim, J. W. Lee, and L. G. Kazovsky. SUCCESS-HPON: a next-generation optical access architecture for smooth migration from TDM-PON to WDM-PON. *IEEE Communications Magazine*, 43(11):S40–S47, Nov. 2005. ISSN 0163-6804. doi: 10.1109/MCOM.2005.1541698.

Analysys. Cost of the BT UK local loop network. Report for Ofcom, Feb. 2005. URL `http://web.archive.org/web/20060406082036/http://www.ofcom.org.uk/consult/condocs/copper/loop.pdf`.

C. M. Assi, Y. Ye, S. Dixit, and M. A. Ali. Dynamic bandwidth allocation for quality-of-service over Ethernet PONs. *IEEE Journal on Selected Areas in Communications*, 21(9):1467–1477, Nov. 2003.

M. Beck. *Ethernet in the First Mile: The IEEE802.3ah EFM Standard*. McGraw-Hill Professional Publishing, 2005. ISBN 978-0-07-145506-0.

B. Chen, J. Chen, and S. He. Efficient and fine scheduling algorithm for bandwidth allocation in ethernet passive optical networks. *IEEE Journal of Selected Topics in Quantum Electronics*, 12(4):653–660, July 2006.

S. Choi and J. Huh. Dynamic bandwidth allocation algorithm for multimedia services over Ethernet PONs. *ETRI Journal*, 24(6):465–468, 2002.

M. Christensen, K. Kimball, and F. Solensky. Considerations for Internet Group Management Protocol (IGMP) and Multicast Listener Discovery (MLD) Snooping Switches. RFC 4541 (Informational), May 2006. URL `http://www.ietf.org/rfc/rfc4541.txt`.

C. Clark, K. Fraser, S. Hand, J. G. Hansen, E. Jul, C. Limpach, I. Pratt, and A. Warfield. Live migration of virtual machines. In *Proc. USENIX NSDI*, 2005.

B. Cohen. Incentives build robustness in BitTorrent. In *Workshop on Economics of Peer-to-Peer systems*, volume 6, 2003.

N. Cvijetic, D. Qian, J. Hu, and T. Wang. Orthogonal frequency division multiple access PON (OFDMA-PON) for colorless upstream transmission beyond 10 Gb/s. *IEEE Journal on Selected Areas in Communications*, 28(6):781–790, Aug. 2010. ISSN 0733-8716. doi: 10.1109/JSAC.2010.100803.

R. Davey, J. Kani, F. Bourgart, and K. McCammon. Options for future optical access networks. *IEEE Communications Magazine*, 44(10):50–56, 2006.

R. Droms. Dynamic Host Configuration Protocol. RFC 2131 (Draft Standard), Mar. 1997. URL `http://www.ietf.org/rfc/rfc2131.txt`. Updated by RFCs 3396, 4361.

G. Du Chaffaut, D. Chapelain, A. Madani, and S. Carpentier. An ATM cell based transmission system on a PON structure. In *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, volume 1, pages 120–124, dec 1990. doi: 10.1109/GLOCOM.1990.116491.

J. George. Applications compelling fiber to the home. *The FTTH Prism*, 3(2):47–55, Oct. 2006.

K. Grobe and J.-P. Elbers. PON in adolescence: from TDMA to WDM-PON. *IEEE Communications Magazine*, 46(1):26–34, Jan. 2008. ISSN 0163-6804. doi: 10.1109/MCOM.2008.4427227.

D. Gutierrez, K. S. Kim, S. Rotolo, F. T. An, and L. G. Kazovsky. FTTH standards, deployments and research issues. In *Proc. 8th Joint Conference on Information Sciences (JCIS)*, pages 1358–1361, 2005.

IEEE Computer Society. Std 802: Standard for local and metropolitan area networks: Overview and architecture, 2001.

IEEE Computer Society. Std 802.1D: Standard for local and metropolitan area networks: Media access control (MAC) bridges, 2004a.

IEEE Computer Society. Std 802.1X: Port based network access control, 2004b.

IEEE Computer Society. Std 802.1ae: Standard for local and metropolitan area networks: Media access control (MAC) security, 2006.

IEEE Computer Society. Std 802.1Qay: Virtual bridged local area networks—amendment 10: Provider backbone bridge traffic engineering, 2009.

IEEE EFM Task Force. Std 802.3ah-2004: Media access control parameters, physical layers, and management parameters for subscriber access networks, 2004.

ITU-T. Information technology – Open Systems Interconnection – basic reference model: the basic model. ITU-T Recommendation X.200, July 1994.

ITU-T. Broadband optical access systems based on passive optical networks (PON). ITU-T Recommendation G.983.1 (superceded), Oct. 1998.

ITU-T. Gigabit-capable passive optical networks (GPON). ITU-T Recommendations G.983.1 thru 4, 2003–2008.

ITU-T. Very high speed digital subscriber line transceivers. ITU-T Recommendation G.993.1, June 2004.

ITU-T. Very high speed digital subscriber line transceivers 2 (VDSL2). ITU-T Recommendation G.993.2, Feb. 2006.

C. Kim, M. Caesar, and J. Rexford. Floodless in SEATTLE: a scalable Ethernet architecture for large enterprises. In *Proc. SIGCOMM*, pages 3–14, 2008. doi: 10.1145/1402958.1402961.

G. Kramer. Interleaved polling with adaptive cycle time (IPACT): a dynamic bandwidth distribution scheme in an optical access network. *Photonic Network Communications*, 4(1):89–107, Jan. 2002.

G. Kramer. What is next for Ethernet PON? In *Proc. 5th International Conference on Optical Internet (COIN)*, July 2006.

G. Kramer and G. Pesavento. Ethernet passive optical network (EPON): building a next-generation optical access network. *IEEE Communications Magazine*, 40(2):66–73, Feb. 2002. ISSN 0163-6804. doi: 10.1109/35.983910.

C. H. Lee, S. M. Lee, K. M. Choi, J. H. Moon, S. G. Mun, K. T. Jeong, J. H. Kim, and B. Kim. WDM-PON experiences in Korea. *Journal of Optical Networking*, 6(5): 451–464, May 2007.

I. M. Leslie, D. R. McAuley, and D. L. Tennenhouse. ATM everywhere? *IEEE Network*, 7(2):40–46, Mar. 1993. ISSN 0890-8044. doi: 10.1109/65.216903.

Y.-K. M. Lin and D. R. Spears. Passive optical subscriber loops with multiaccess. *Journal of Lightwave Technology*, 7(11):1769–1777, Nov. 1989. ISSN 0733-8724. doi: 10.1109/50.45900.

D. McAuley. Protocol design for high speed networks. Technical Report UCAM-CL-TR-186, Jan. 1990.

M. P. McGarry, M. Maier, and M. Reisslein. Ethernet PONs: a survey of dynamic bandwidth allocation (DBA) algorithms. *IEEE Communications Magazine*, 42(8): 66–73, Feb. 2002.

M. P. McGarry, M. Reisslein, and M. Maier. WDM Ethernet passive optical networks. *IEEE Communications Magazine*, 44(2):15–22, Feb. 2006. ISSN 0163-6804. doi: 10.1109/MCOM.2006.1593545.

P. Merkle, A. Smolic, K. Müller, and T. Wiegand. Coding efficiency and complexity analysis of MVC prediction structures. In *Proc. 15th European Signal Processing Conference (EUSIPCO)*, Sept. 2007.

R. M. Metcalfe and D. R. Boggs. Ethernet: distributed packet switching for local computer networks. *Commun. ACM*, 19(7):395–404, 1976. ISSN 0001-0782. doi: 10.1145/360248.360253.

S. E. Minzer. Broadband ISDN and asynchronous transfer mode (ATM). *IEEE Communications Magazine*, 27(9):17–24, 57, Sept. 1989. ISSN 0163-6804. doi: 10.1109/35.35508.

R. Monnard, M. Zirngibl, C. Doerr, C. Joyner, and L. Stulz. Demonstration of a 12 x 155 Mb/s WDM PON under outside plant temperature conditions. *IEEE Photonics Technology Letters*, 9(12):1655–1657, 1997.

A. Myers, E. Ng, and H. Zhang. Rethinking the service model: Scaling Ethernet to a million nodes. In *Proc. ACM SIGCOMM Workshop on Hot Topics in Networking*, Nov. 2004.

S. Narikawa, N. Sakurai, K. Kumozaki, and T. Imai. Coherent WDM-PON based on heterodyne detection with digital signal processing for simple ONU structure. In *Proc. European Conference on Optical Communications (ECOC)*, pages 1–2, Sept. 2006. doi: 10.1109/ECOC.2006.4801085.

Ofcom. UK broadband speeds, May 2010: the performance of fixed-line broadband delivered to UK residential customers, July 2010. URL `http://stakeholders.ofcom.org.uk/binaries/research/telecoms-research/bbspeeds2010/bbspeeds2010.pdf`. Retrieved 27th August 2010.

P. Ogden. Fibre to the home—is it a reality? Master's thesis, University of Cambridge Department of Engineering, Aug. 2010.

K. Pagiamtzis and A. Sheikholeslami. Content-Addressable Memory (CAM) circuits and architectures: a tutorial and survey. *IEEE Journal of Solid-State Circuits*, 41: 712–727, 2006.

C. Perkins. IP Mobility Support for IPv4. RFC 3344 (Proposed Standard), Aug. 2002. URL `http://www.ietf.org/rfc/rfc3344.txt`. Updated by RFC 4721.

R. Perlman. Rbridges: transparent routing. In *Proc. IEEE INFOCOM*, volume 2, 2004.

D. C. Plummer. Ethernet Address Resolution Protocol. RFC 826 (Standard), Nov. 1982. URL `http://www.ietf.org/rfc/rfc826.txt`.

T. L. Rodeheffer, C. A. Thekkath, and D. C. Anderson. SmartBridge: a scalable bridge architecture. In *Proc. SIGCOMM*, 2000.

E. Rosen, A. Viswanathan, and R. Callon. Multiprotocol Label Switching Architecture. RFC 3031 (Proposed Standard), Jan. 2001. URL `http://www.ietf.org/rfc/rfc3031.txt`.

M. A. Scott, A. W. Moore, and J. Crowcroft. Addressing the scalability of Ethernet with MOOSE. In *ITC 21 First Workshop on Data Center – Converged and Virtual Ethernet Switching (DC CAVES)*, Sept. 2009.

B. Skubic, J. Chen, J. Ahmed, L. Wosinska, and B. Mukherjee. A comparison of dynamic bandwidth allocation for EPON, GPON, and next-generation TDM PON. *IEEE Communications Magazine*, 47(3):40–48, 2009.

J. Touch and R. Perlman. Transparent interconnection of lots of links (TRILL): problem and applicability statement. RFC 5556 (Informational), May 2009. URL `http://www.ietf.org/rfc/rfc5556.txt`.

Virgin Media. Broadband traffic management. URL `http://allyours.virginmedia.com/html/internet/traffic.html`.

F. Yu, R. H. Katz, and T. V. Lakshman. Efficient multimatch packet classification and lookup with tcam. *IEEE Micro*, 25(1):50–59, Jan. 2005. ISSN 0272-1732. doi: 10.1109/MM.2005.8.

J. Zheng and H. T. Mouftah. Media access control for Ethernet passive optical networks: an overview. *IEEE Communications Magazine*, 43(2):145–150, Feb. 2005. ISSN 0163-6804. doi: 10.1109/MCOM.2005.1391515.

Attributions

[*]Figure adapted from work originally by “Riick”, used under the terms of the Creative Commons Attribution–ShareAlike 3.0 Unported licence

Appendix A

Risk Assessment

Since this project was entirely computer-based, I was subjected to no unusual risks beyond those related to long-term computer use, which—having spent a considerable proportion of my working life using computers—I am used to mitigating against as a matter of habit.