Network Architectural Implications of Fibre to the Home

Malcolm Scott

University of Cambridge Computer Laboratory and University College London

Abstract: When designing a new access network technology, it is natural to primarily consider the physical layer—the photonic and electronic developments which allow for ever-faster link speeds, and to bring more and more customers and services onto a single converged infrastructure. However the oft-ignored protocols operating on top of this infrastructure, in particular Ethernet, are suffering scalability problems. In this paper I describe some of the protocol and architectural challenges which must be considered in any deployment of Fibre to the Home.

1 Introduction

Fibre to the Home (FTTH) is being taken increasingly 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 at the time of writing use a Passive Optical Network (PON) in order to dramatically save on fibre costs by providing aggregation within the "last mile" between the central office (CO) and customer premises; multiple customers (each with an Optical Network Unit, ONU) are connected to a single transceiver (Optical Line Termination, OLT) by means of a branching tree of fibres and passive splitter/combiner units. These invariably operate entirely in the optical domain, or in OSI networking terminology at the physical layer.

There are two current PON standards: GPON (ITU-T G.984 [1]) and EPON (IEEE 802.3ah [2]. Both specify the physical layer and multiple-access scheme, but omit any issues of network architecture.

Regardless of the PON standard in use, any modern FTTH deployment is highly likely to use Ethernet, either directly in the case of EPON or encapsulated in GEM frames in the case of GPON. It is the natural choice for transmission of IP traffic, and has become commonplace in almost every variety of modern data network. As a result it is also a low-cost option, and has sufficient market penetration to remain ubiquitous for the foreseeable future.

However, an important matter which is covered by neither standard is the manner in which an Ethernet switch and higher-layer protocols might operate on a PON; as I will discuss, the network architecture of a PON is significantly different from that of any prior use of Ethernet.

2 Ethernet Switching Domain Size

Almost every traditional use of Ethernet operates using small switching domains joined together at a higher layer by IP routers (usually one switching domain corresponds to one IP subnet). It is well-known that running Ethernet and IP with large switching domains exhibits a variety of scalability problems; I will discuss these problems in more detail with possible solutions in Section 3. The current industrial recommendation is that a switching domain should contain no more than 500 hosts, reduced to 200 if non-IP protocols are also in use [3].

Any Ethernet-based PON is logically a collection of point-to-point links between small Ethernet bridges in each customer's ONU and several OLTs in the CO. The branching structure of the PON is transparent to Ethernet and acts merely as a way to multiplex virtual point-to-point links onto one transceiver.

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 [4]. Thus if each OLT were a separate Ethernet containing one device per customer there would be no scalability issue as far as Ethernet is concerned. However the OLTs must be

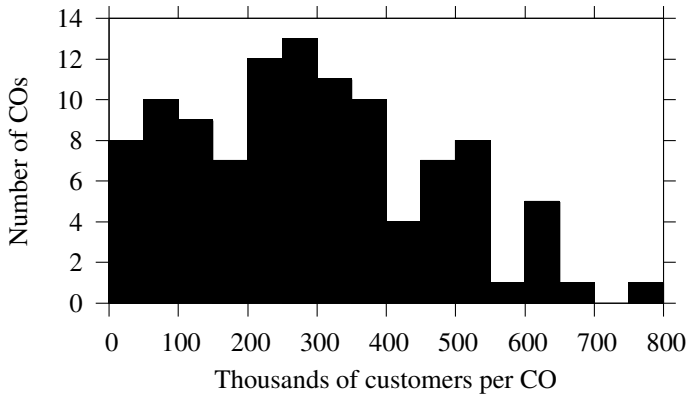


Figure 1: Histogram of estimated number of customers per BT 21CN metro node (data from Ogden [4], derived from the number of distinct postcodes served by each node, assuming the average number of customers per postcode to be 16)

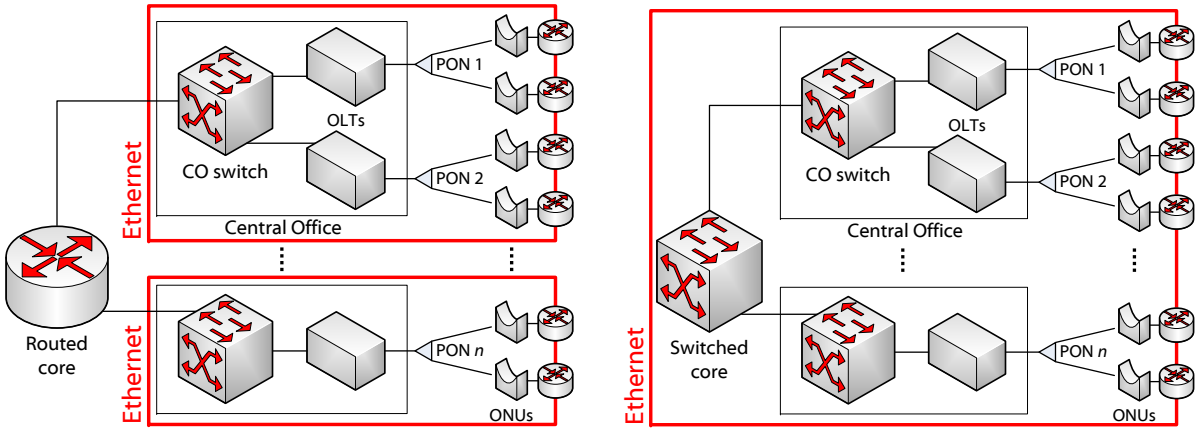
interconnected with each other and with the core network, and it is convenient to use Ethernet switches for this purpose, rather than costly and power-hungry routers. Modern OLTs are designed specifically for this mode of operation, with OLTs available in the form of a GBIC module suitable for insertion into a larger switch chassis [5]. Thus, by bridging together all of a CO's PONs using Ethernet switches, every customer served by these PONs will be on a single large Ethernet network.

In order to gain a sense of the scale of the resulting Ethernet deployments, we can refer to data on BT's 21CN deployment. Figure 1 shows an estimate of the distribution of customers per metro node (BT's term for a CO linking the access network to the core). The range of metro node sizes is high, but even the smallest COs serve well in excess of the recommended 500 devices, and the largest are several orders of magnitude beyond Ethernet's capabilities as it stands.

Furthermore, it can be desirable to push Ethernet into the core network interconnecting COs, as 21CN does; large scale Ethernet switches remain more cost-effective than high-speed IP routers. Figure 2 summarises the two main options for deploying Ethernet switches and IP routers on a FTTH network, and illustrates the extent of the resulting switching domain. If the core network as well as the access network were Ethernet-based, the constituent switches would be responsible for managing communication between several million devices.

3 Scalability Challenges and Solutions

It is clear that any Ethernet PON will lead to switched Ethernet networks with a very large number of logical links. Here I will briefly describe three examples of issues which are likely to arise when running a FTTH-scale Ethernet and IP network, along with some of the ongoing efforts to mitigate against each.



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

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

Figure 2: Network topology options affecting Ethernet switching domain size

3.1 Ethernet Forwarding Database

Hosts are identified by manufacturer-assigned MAC addresses, which have no structure of use in address aggregation. One major problem is that each Ethernet switch must separately learn the location of every host. The database must then be consulted rapidly for every frame forwarded and in large switches is usually implemented as a content-addressable memory (CAM); due to speed and power constraints [6] the capacity is usually limited to 16000–64000 entries. An Ethernet with more hosts or routers than this will perform very poorly—at best, frames will be flooded to all links, which may saturate links with unnecessary traffic. In extreme cases, the network could completely fail to provide any useful throughput.

Any backwards-compatible solution to this problem must introduce a means of locating a host without storing this complete database on every switch. Kim et al. propose SEATTLE [7], which distributes MAC address location data between the complete network of switches in a distributed hash table. However the implications of deploying this scheme on FTTH need careful consideration: FTTH deployments as described would have relatively few switches, and furthermore relying on a remote switch for local forwarding decisions may cause a minor local outage to have far-reaching effects. As an alternative I propose MOOSE [8] in which switches transparently rewrite MAC addresses to form a hierarchy whereby the location of a host can be determined immediately by inspecting its hierarchical address.

3.2 Non-Tree Topologies

If an Ethernet network contains a redundant path, a single broadcast frame can loop indefinitely, causing a broadcast storm which uses all the capacity of all links. Ethernet handles this situation by invoking the Rapid Spanning Tree Protocol, RSTP [9, §17], which breaks loops by disabling any redundant links and converting any topology into a tree. Any redundant connectivity cannot therefore be used for increased capacity. Networks with a high degree of interconnectivity, such as a telco's core network, would find a large proportion of links disabled; this constrains forwarding to suboptimal paths and may introduce bottlenecks, particularly around the root of the spanning tree.

IP solved this problem from the outset by using routers to direct packets along the best possible path to their destination; Perlman [10] has proposed Rbridges to bring routing into Ethernet, but not with the aim of benefitting large networks. MOOSE lends itself to a more-scalable routed Ethernet; by introducing a hierarchy, IP-like address aggregation can occur.

3.3 Broadcast Traffic

Many protocols implemented on top of Ethernet and IP make use of the broadcast feature to locate hosts and advertise services within their subnet. In particular, ARP [11] uses broadcast queries to convert IP addresses into MAC addresses, and is used by every IP host. If all devices on the network are under the telco's control, including the routers in customers' premises, the use of unnecessary broadcast protocols can be minimised but it is likely that ARP must remain. In a large subnet, ARP traffic alone could consume a significant proportion of the capacity on all links [12].

Since all traffic must pass through the CO switch, there is scope for this switch to intercept broadcast ARP messages and handle them more efficiently, using the techniques proposed by Elmeleegy and Cox [13] or my proposed ELK directory service [8].

The next version of IP, IPv6, replaces ARP with a multicast-based protocol, ND [14], but it remains to be seen whether this provides any real advantage with large numbers of hosts.

4 Existing Workarounds and Conclusions

BT chose Ethernet for 21CN, and have therefore had to work around some of these scalability problems. 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 interconnecting the traditional Ethernet networks of the access network. Although this fixes the routing problem, MPLS LERs must still perform expensive lookup operations when encapsulating each frame and are therefore power-hungry and expensive.

Secondly, BT have deployed Provider Backbone Bridge Traffic Engineering, PBB-TE [15]. This technology 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 strictly adhere to 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 switches also suffer from the same problems as MPLS LERs regarding 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 and low-cost nature. Ongoing research is gradually solving these problems, but it is of utmost importance when designing a new access network technology such as FTTH to be aware that a seemingly-simple change to the network topology can cause significant challenges in protocols and architecture.

References

[1] ITU-T. Gigabit-capable passive optical networks (GPON). ITU-T Rec. G.984.1 thru 4, 2003–2008.

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

[3] P. Oppenheimer. *Top-Down Network Design*. Cisco Press, 3 edition, 2010. ISBN 978-1587202834.

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

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

[6] 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.

[7] 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.

[8] M. Scott, A. 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)*, September 2009.

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

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

[11] D. C. Plummer. Ethernet Address Resolution Protocol. RFC 826, November 1982. http://www.ietf.org/rfc/rfc826.

[12] 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*, November 2004.

[13] K. Elmeleegy and A. L. Cox. Etherproxy: scaling Ethernet by suppressing broadcast traffic. In *Proc. IEEE INFOCOM*, pages 1584–1592, 2009.

[14] T. Narten, E. Nordmark, W. Simpson, and H. Soliman. Neighbour Discovery for IP version 6 (IPv6). RFC 4861, September 2007. URL http://www.ietf.org/rfc/rfc4861.txt.

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