

Ethernet Fabric Routing (UETS/EFR)

A hierarchical, scalable and secure ultrahigh speed switching architecture

José Morales Barroso
L&M Data Communications
jmb@ieee.org

Guillermo Ibáñez Fernández
Dpto. Ingeniería Telemática
Universidad Carlos III Madrid
gibanez@it.uc3m.es

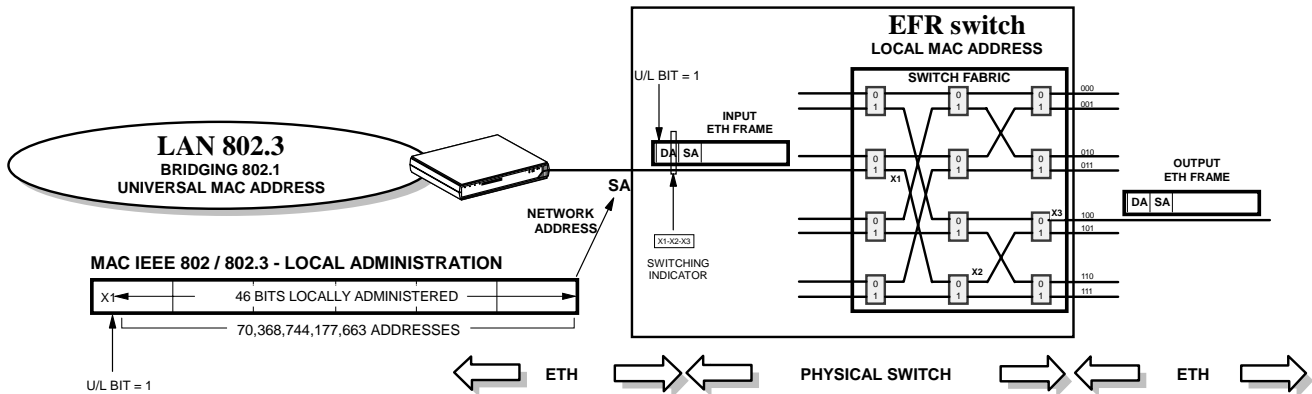


Figure 1. Ethernet Fabric Routing Architecture (EFR). Basic addressing and switching.

Abstract. In this paper we shall try to answer a fundamental question: how to build the Next Generation Network without disruption. The TCP/IP protocol stack has performance limits and the transparent bridge paradigm of Ethernets is complex to scale by lack of addressing hierarchy. To provide both full scalability and high performance, a new Ethernet switching architecture named Universal Ethernet Telecommunications Service (UETS) is proposed, capable of performances in the range of Terabits per second. It is based on hardware switching of Ethernet frames using topological and hierarchically assigned standard local MAC addresses, technique named Ethernet Fabric Routing (EFR). The architecture solves Ethernet's scalability and security problems, has low cost/performance ratio and is compatible with existing Ethernet and IP networks, providing a non disruptive migration path to high performance networks and overcoming TCP/IP performance limitations. The objective is to promote a new standard for UETS/EFR architecture, suitable for high performance specialized island networks that may operate together as an overlay network.

Keywords: bridges, routers, ethernet, hierarchical addressing, internet, broadband, network convergence.

I. INTRODUCTION

IP protocol has been the "glue" to create layer 3 internets connecting dissimilar networks, but today is beginning to show signs of age. In this regard, we should point out that this is the reason for the emergence of MPLS and GMPLS. However, with the universal deployment of Ethernet and its self compatibility (2M, 10M, 100M, 1G, 10G, 100G), the

dissimilarity of layer two subnets has disappeared. The Ethernet switches advantages in cost/performance ratio versus routers is compelling its application to transport over metropolitan and wide area networks, questioning somehow the undisputable role of IP as interconnection layer. More recently, the widespread deployment of broadband connections is creating new markets for services such as scientific applications, HDTV, NAS, 3D visualization, online gaming, etc, with very stringent requirements for packet networks in terms of bandwidth, latency and security. The current paradigms of IP routers and Ethernet bridges were crucial for the present Internet success, but have limitations to satisfy these requirements because neither TCP/IP was designed for terabit performance nor transparent bridges were designed for big size networks.

The predominant paradigm for layer two networks, based on broadcast of frames through a spanning tree and transparent learning bridges to limit these broadcasts, does not scale because many links are disabled by the spanning tree protocol. The current use of Ethernet universal MAC (UMAC) flat addresses in transparent learning bridges does not scale due to their lack of hierarchy, which precludes network segmentation and any aggregation of routes.

The transparent bridge paradigm is being modified at IEEE 802.1 working groups to provide scalability to big networks, but at the cost of multiple encapsulations and increased complexity. The schemes currently being standardized, such as MAC in MAC, Q-in-Q, Provider Bridges, IP Tunnels or MPLS transport, are both complex and expensive. Frame

overhead is substantially increased with successive encapsulations as a result of the successive protocols adopted (MPLS, Provider Bridges, etc). Ethernet Frame Extension is being standardizing by project IEEE 802.3as to give enough space to labels without reducing the payload. On the other hand, the IP protocol stack, key for the success of Internet, is showing its performance limitations. IP and TCP were designed for maximum interoperability between dissimilar networks, not for extreme performances. Performance of TCP/IP servers has known limits at gigabit speed [1]. Transport time-critical applications through the Internet, such as voice and video, are very different from the typical TCP/IP traffic of data, therefore, it is needed a new standard that could guarantee quality of service to end users. MPLS is supposed to be the solution, but it does not scale, being useful on the core of the network, but not at the edges [2].

This paper describes the UETS/EFR architecture [3], a new paradigm for networks of any size by means of Ethernet as interconnection protocol. The first difference with current technologies is the utilization of hierarchical standard local MAC (LMAC) addresses to perform, in EFR switches, physical frame switching and routing without forwarding tables or label swapping. The hierarchical MAC addressing, linked to physical ports location, is assigned and controlled by an administrator, providing scalable routing and extremely simple hardware based switching. The second difference is the protocol stack simplification, removing layers instead of adding a new sublayer. The network services are offered by the Data Link Layer 2 protocol stack, as shown in figure 2, which provides proven (LAPB/X.25, SDLC/IBM) high performance without the limitations of TCP/IP. The switching principle is shown at figure 1 and dual stack at figure 2.

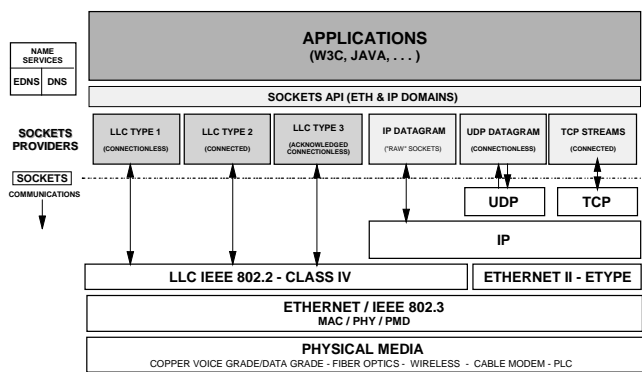


Figure 2. Dual stack: LLC/ETH + TCP/IP/ETH.

The remainder of this paper is structured as follows: Section II describes UETS network architecture, Section III other aspects as scalability, security and performance, Section IV applications and, finally, the conclusions are summarized in Section V.

II. NETWORK ARCHITECTURE

UETS extends the continuity of the Ethernet domain, without IP routers, to an enterprise or service provider's networks. In UETS the hosts at the network may be either standard, operating only with TCP/IP, or dual stack (L2 and L3) terminals, named TUE, which use the LLC (IEEE 802.2) protocols for intra UETS domain and TCP/IP for external and (if desired) internal communications.

A. Network elements

An UETS domain is a network composed of UETS network nodes (CUE), network terminations (NTE/TRUE) and end nodes (TUE). Services such as Ethernet DNS (EDNS) are also used. Inside an UETS domain, end nodes may employ any of two communication stacks: TCP/IP or LLC, in both cases over Ethernet. This domain can be extended over a local, campus, ISP, metropolitan or a world-wide network. EDNS service performs translation from Domain Name addresses or IP addresses to MAC addresses upon request from end nodes.

The CUE (Central Universal Ethernet) are a new network node concept, which uses hierarchical local (U/L bit = 1) addresses linked to physical port ids. Ethernet frame switching and routing are based exclusively on the local MAC destination address. They get assigned an address prefix and a bit mask of variable bit length, as described in the example of figure 3. The destination MAC addresses that match with the address prefix are switched according to the bits of the bit mask, selecting the output port according to the value of the bit group. In figure 1, the switch mask correspond to last three address bits (switching indicator), any frame with destination MAC that matches the prefix and ends in '100' is always switched to physical output port number four.

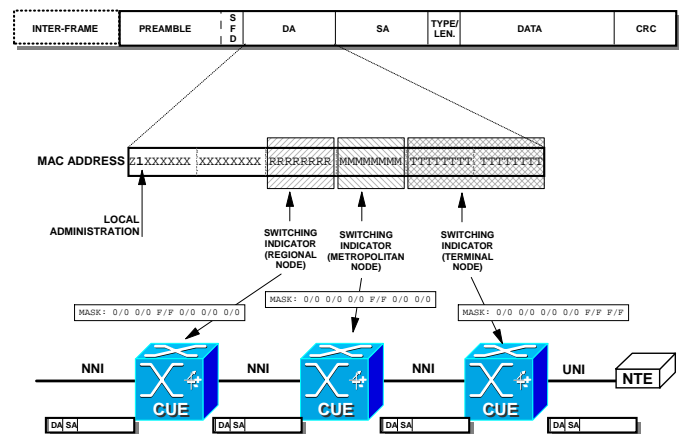


Figure 3. Addressing and switching principle. Example.

NTE is the basic access device that allows remote 802.1 networks and data terminals connections, acting as a gateway, at the edge of UETS network, between an Ethernet 802.1 LAN and a UETS domains, as shown in figure 4. NTE performs address translation (ENAT) or encapsulation between 802.1 universal (UMAC) and UETS local (LMAC)

addresses. Also, may intercept DNS requests and convey it to the EDNS service. At network terminations interfacing with 802.1 LANs, standard MAC addresses encapsulation (tunneling) or translation is required to allow interoperability. In the UETS domain, MAC addresses are physically dependent and controlled by the ISP or network owner, allowing hardware based switching and enhancing network security by preventing layer two address spoofing. UETS addresses may be hidden to service users for enhanced security.

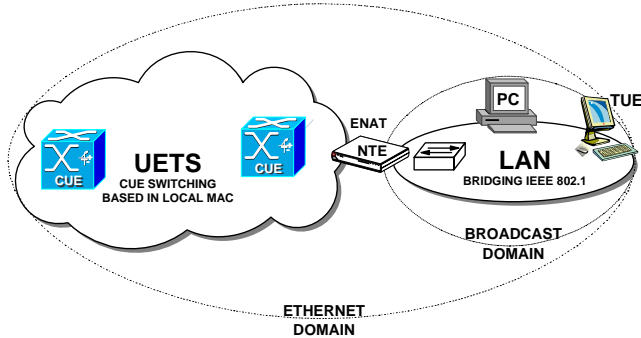


Figure 4. UETS domain and standard Ethernet domain.

B. Protocol stack

Native UETS protocol stack is very simple: Applications use sockets to interface with layer two LLC protocol connectionless and connection oriented variants. Some transport layer functions such as connection identification (ports) are implemented at layer two in the sockets providers. In order to guarantee the interoperability with TCP/IP hosts and networks, UETS uses the dual stack communication architecture shown in figure 2.

1) LLC over Ethernet

Current applications use ports at transport (TCP/UDP) protocol layer to identify points of access in a machine and to define connections using endpoint service identifiers (pair protocol, port). At UETS stack (LLC over Ethernet) this function is performed using link service access points (LSAP), that provide interface ports for users above LLC sublayer, see figure 5.

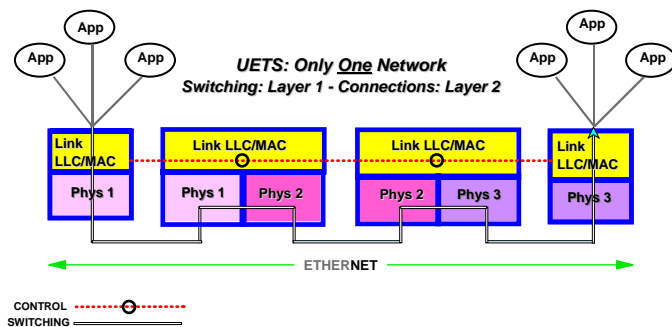


Figure 5. LLC over Ethernet: control and switching.

2) TCP/IP over Ethernet

End nodes use TCP/IP in the standard way, because there is no problem to transport IP datagrams over Ethernet. In UETS they are isolated Ethernet and IP domains; therefore, it is possible to offer simultaneously services in the Ethernet domain with the IEEE 802.2 LLC protocols, and in the IP domain with the IP, UDP and TCP protocols. This stack is used to Internet access employing the not secure IP domain.

C. MAC addressing

1) MAC addresses structure

UETS addressing is based in standard IEEE 802 format, by means of locally administered Ethernet MAC (LMAC) addresses. The universal/local (U/L) bit, adjacent to the individual/group (I/G) address bit, is set to Local value. According to the definition of IEEE STD 802-2001, if this bit set to 1, it means that the entire 48-bit address has been locally administered. Universal Addresses (UMAC) are not allowed inside an EFR domain.

2) MAC addresses configuration

EFR switching is based on the correlation of MAC addresses to the physical hardware architecture. Due to its locally controlled address structure, configuration of addresses may be coordinated and arranged hierarchically. Configuration of MAC addresses may be performed either as a standard network management activity via SNMP over Ethernet (RFC1089), or via automatic address assignment protocols, executed by EFR switches through information exchange with root switch and address server.

3) Local MAC addresses assignment

The LMAC addresses may be assigned hierarchically according to the network topology of switches and may also correspond to the physical design of switches, thus allowing switching of frames based only on UETS LMAC destination address. Hierarchical assignment of local addresses makes possible hierarchical routing via simple switching up to global size domains, something impossible with universal MAC (UMAC) addresses.

4) Pseudo-IP mode.

The 6 byte MAC address structure permits the use of the standard 4 byte IPv4 addresses as layer two addresses, embedded in the 6 byte address.

5) Mobile and fixed addresses.

A possible assignment of LMAC addresses for mobile and fixed addresses usage consists of using a bit, next to the U/L bit, to indicate mobile or fixed UETS address. A mobile address may be assigned upon initialization to a fixed UETS address.

6) Multicast addresses

Multicast group addressing can be used in UETS domains. UETS switches may implement multicast addressing through snooping of subscriptions to multicast groups in a way similar to IGMP snooping and GMRP in routers. Multicast is established at UETS as a complementary switching plane through subscriptions to multicast groups.

D. Flow and congestion control, QoS and CoS

Flow and congestion control is currently performed in an end-to-end basis by TCP. But TCP/IP mechanisms were designed to obtain reliability and controlled flow through a best effort IP network. In the context of increased performance and reliability of today's networks, it begins to show its limitations. Congestion control performed at layer two [4] is currently under discussion at IEEE Congestion Management Study Group.

The requirement for end-to-end QoS guarantee is important in some services. The proposed LLC/ETH protocol stack allows service differentiation to four basic classes of traffic, premium service, mission critical, stop and go and best effort:

- "PREMIUM" real time traffic (both rt-CBR and rt-VBR), such as circuit emulation, voice or video, critically sensitive to delay, uses LLC1 protocol, connectionless unacknowledged.
- "MISSION CRITICAL" non real time traffic, such as network file services, uses LLC2 protocol, connection oriented with end-to-end flow control.
- "STOP AND GO" services, such as Inter Process Communications (IPC) use the acknowledged connectionless LLC3 protocol.
- "BEST EFFORT" services, such as TCP/IP are encapsulated over Ethernet frames with Ethertypes.

IEEE 802.1 VLANs currently provides only 3 priority bits that allow distinctive class of service assignment to different traffics. EFR switches use the standard Ethernet frame structure; the only difference with 802.1 bridges is the U/L bit usage for addresses. Thus they can implement the same priority mechanisms as standard 802.1Q switches regarding priority levels in order to provide different classes of service inside the switch, according to the explicit priority tag (802.1p) in the frame. If traffic prioritization is performed per frame by traffic classification instead of by explicit tagging, the same mechanisms can be applied as those used by standard Ethernet switches.

E. Ethernet Fabric Routing and Switching.

The terms *routing* and *switching* are sometimes used without clear distinction. In the UETS/EFR network they are clearly differentiated.

1) Switching

Switching of frames is performed hierarchically downwards across the switches of a common UETS address domain (i.e. under the same UETS address administrator), through successive decoding of destination address bits via successive masks at corresponding switches as shown in figure 3. In the upward direction, the frames can be hierarchically switched up through the successive switches till a switch with common prefix to origin and destination MAC addresses is reached. From this switch, the frame can be switched downward till destination.

Switching is performed via simple hardware mechanisms that decode MAC addresses for output port selection and

does not need any kind of forwarding table, thus avoiding memory access time and processing delay.

2) Routing

Routing is performed across switches for transit of frames toward their destination area (i.e. frames that are outside the UETS address domain of the switch). This is named *micro-routing*, and uses small size (micro) tables for inter UETS domain frame routing. It is also worth to note an important advantage: both switching and routing are performed *without* label swapping at nodes.

F. Mesh and trees

UETS architecture leads to hierarchical tree topologies that enable the addresses decoding till destination. Whereas the dominant structure of campus or building's Ethernet networks is hierarchical with aggregations at horizontal backbones, vertical backbone and core, UETS networks may have arbitrary structure. The rule is that the links belonging to the hierarchical tree forward via switching and the cross or transversal links, additional to the hierarchical tree, forward the frames using micro-routing.

G. Switch designs

EFR switching is designed to perform physical (hardware) frame switching. So it is well suited to Banyan type switches where selection of the switching path is performed by successive decoding of the destination address field. However other switch designs can be used, such as those oriented to tag switching or label switching such as ATM based cell switches or MPLS switches. Figure 6 illustrates this option. Label swapping is not needed, eliminating memory accesses and frame modification. The switch does neither need internal tables to map tags nor labels to output ports. Fibre Channel fabric can also be used directly, applying the 24 bits mask to the MAC destination addresses.

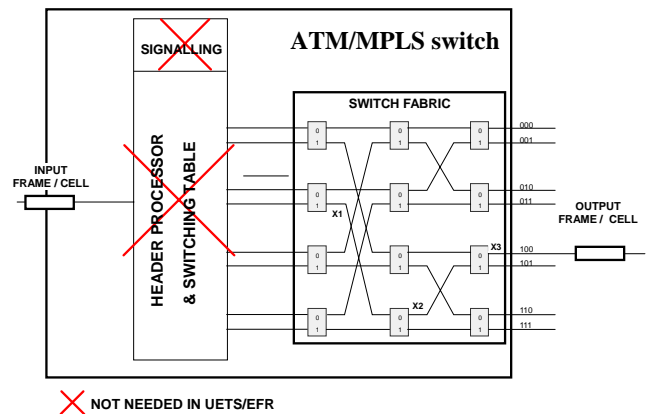


Figure 6. Alternative EFR switch

III. OTHER ASPECTS

A. Scalability and domain size

EFR switching architecture is able to scale from small domains up to international world-wide domains. Scalability of UETS and EFR derives from two facts: its inherent

hierarchy and the 48 bit MAC, which can be used for up to 2^{46} (70 trillion) unicast addresses. If even more addresses are needed, using UETS with MAC-in-MAC and VLANs provides the equivalent to IPv6 addressing capability.

B. VLANs and encapsulation

The architecture is fully compatible with IEEE 802.1Q VLANs and other Ethernet encapsulations such as MAC in MAC or SNAP. Those encapsulations are used at UETS border switches to connect standard 802.1D bridged areas.

C. Security

The UETS/EFR architecture provides inherent layer two security. Due to the direct binding of MAC addresses to the network point of attachment, they can not be spoofed, as they are controlled by the switch owner (ISP or UETS domain authority), and physical access to the point of attachment is required for supplantation. This security may be complemented with 802.1X. No attacks based on manipulation of the spanning tree protocol are possible, as there is no spanning tree protocol in the switched network. Since the switches do not learn addresses, no attacks based on table overflow are possible to the switches caches.

D. Expected UETS performances

The simplicity of UETS/EFR physical switching, without the limitations imposed by the memory access and process related with routing or look-up tables, allows operating at full speed over the switch fabrics. Fastest IP routers today has reached a per-chassis capacity of 640 Gbps, increased by a factor of two in the last five years. In the same period of time, Ethernet has increased its speed by a factor of ten, and it is well known that "the challenges for realizing 100 GbE are related to Ethernet *switching* and not to Ethernet *transport*" [5]. With simple commercial Banyan type crosspoint switches, especially well suited to build EFR network nodes, has been demonstrated switching up to 10 Tbps in a single chassis, with an estimated scalability up to 50 Tbps per chassis [6].

E. EDNS performance requirements

Each tuple at EDNS contains at least a Universal MAC address and Local MAC address, IP address(es) and host ID. The operation and requirements for the EDNS service is in the worst case similar to the conventional DNS service. But, if there is a numerical correspondence between UETS addresses and geographical domain names (e.g.: *.es .us .it*), EDNS load is reduced drastically.

F. Future work/open issues/IANA

The main aspects of UETS architecture need to be discussed to obtain maximum interoperability and performance are: detailed LLC protocol extensions and alternative Data Link Control protocol specification, congestion control mechanisms at UETS switches, interswitches routing protocol and UETS addresses assignment and coordination. UETS architecture uses local Ethernet addresses that can be administered in a global way. In this case, hierarchical assignment of UETS addresses should be performed with

similar procedures to current assignment of IP addresses. A specific layer 2 multicast group addresses is needed to address to "all UETS switches" to be used by switches configuration protocol. For global interoperation, a coordination of UETS addresses (47 bit) is needed to assign UETS prefixes.

IV. APPLICATIONS

UETS can be implemented as separate islands embedded inside an IP "cloud", which can be interconnected. The main advantages of the architecture are its inherent security, wire speed performance, compatibility and interoperability with existing IP networks and applications, and lower cost/performance ratios. Applications specially suited to this architecture are: high performance provider networks of any size (LAN/MAN/WAN), quadruple play access networks, scientific networks, distance learning, computing on net, HDTV distribution, home networking, storage networks (SAN/NAS), military secure networks, networks of workstations (NOW) and Layer 2 VPNs.

V. CONCLUSIONS

Today, performance of networks is limited by TCP/IP throughput limits and lack of scalability of layer 2 bridged networks. The new and simple UETS/EFR Ethernet based switching architecture described in this paper, addresses these problems from the root. It is compatible with TCP/IP networks and exhibits outstanding advantages, allowing: performances only limited by hardware switching speeds; inherent security at layer two; lower latencies than TCP/IP protocols; full link utilization by flow and congestion control; multipoint to multipoint communications and scalability to all network sizes. In UETS, Ethernet support, simultaneously, link (Layer 2 LLC) and network (Layer 3 IP) services. For that reason, this technology can co-exist indefinitely and harmoniously with current Internet over TCP/IP, working in a complementary way. From a technological point of view, no migration is needed, as opposed to what happens with IPv4 and IPv6.

VI. REFERENCES

- [1] G. Regnier et al., "TCP Onloading for Data Center Servers", Computer, November 2004, pp. 48-58.
- [2] S. Cherry, "Ethernet's High-Wire Act", IEEE Spectrum, April 2005, pp. 53-55.
- [3] J. Morales Barroso, "From Computer Networks to the Computer on Net", IEEE Communications Magazine / Global Communications Newsletter, October 2005, pp. 2-4.
- [4] Tanmay Gupta, Manoj Wadekar, Jeff Wise, "Congestion Management Capabilities of Various Fabrics", IEEE 802.1 Congestion Management Interim. January 2006
- [5] M. Duell, M. Zirngibl, "100 Gigabit Ethernet - Applications, Features, Challenges". IEEE INFOCOM 2006 - The Terabit Challenge. April 2006.
- [6] M. Baldi, Y. Ofek., "Multi-Terabit/s IP Switching with Guaranteed Service for Streaming Traffic". IEEE INFOCOM 2006 - The Terabit Challenge. April 2006.