Reliable Ethernet

Aleksandr Movsesyan

December 8, 2009
## Contents

1 Introduction .................................................. 3

   2.1 Current Protocol Stack ................................ 4
      2.1.1 Layer 1 - Physical Layer ....................... 4
      2.1.2 Layer 2 - Link Layer .......................... 5
      2.1.3 Layer 3 - Internet Layer ....................... 5
      2.1.4 Layer 4 - Transmit Layer ....................... 6
      2.1.5 Layer 5 - Applications ......................... 7
   2.2 How Packets are Routed (BGP) ......................... 7

3 Background .................................................... 8
   3.1 UETS/EFR[2] ............................................. 8

4 Bibliography .................................................. 11
1 Introduction

Governments and data centers need to provide correct data almost instantly to allow instant communication and/or vital data to its customers or fellow employees. Our current implementation of Ethernet allows for a decent error recovery, but is limited to speeds of the TCP/IP protocols. As of today, it takes 5 milliseconds (cite) to recover a lost or damaged packet sent over the network. Other solutions such as InfiniBand and fiber optics have a solution that cuts this down from the milliseconds range to the microseconds range which is an order of magnitude faster than Ethernet(cite), but cost XXX times more than Ethernet, which makes it a burden for the target audience to have. This research is important to allow people to have faster connection and error recovery at a lower cost. This solution presented in this paper will lead to three outcomes. First, it will drive further research into making faster cheaper solutions of faster error recover and data transfer. Second, the proposed solution will allow governments and data centers to move to a faster cheaper solution to supply their data. Third and finally, the research that has been done in this area and my contributions will allow for faster connection to all users and companies and will allow better internet solutions that can run on this faster system.

According to current research there is a solution that will decrease error recovery from the current milliseconds to microseconds. The system of UETS (Universal Ethernet Telecommunications Service) with a new switching technique EFR (Ethernet Fabric Routing) will decrease the protocol stack from 5 layers to a 2 layer system. The proposed system will get rid of TCP and IP all together and now we are no longer limited to speeds of these protocols and now are only looking at hardware switching speeds. In addition there will need to be a change in the way switches are made and the way packets are decoded
and checked. Also, this new system can be integrated with the current way of routing by just adding a device on both sides that will transform an incoming packet into the new packet and then back to a regular packet. With end to end TCP/IP in place, this has proven to increase speeds. The lack of the research is the place where my research begins.

With the new system layouts of packet transfer there will need to be a new algorithm that will detect congestion and recover from errors since TCP and IP will no longer be in place, which take care of these problems. Currently to test the new protocol the researchers are using TCP congestion and error recovery algorithms at end nodes. This, I believe, is slowing down the performance of this new system since the packet must be transformed in the old packet type just to make sure there were no errors and to test congestion. I have implemented a new congestion and error recovery algorithms to use the proposed packet structure from end to end and this enables error recovery of the system to go from XXX(time) to XXX(time). It does this by checking for error at each hop and makes sure no hop has a lot of congestion on it.

2 Basics[5]

2.1 Current Protocol Stack

The current protocol stack includes five layers. The following sections describe what each of the layers.

2.1.1 Layer 1 - Physical Layer

This is the actual link between the nodes.
2.1.2 Layer 2 - Link Layer

The link layer is the communication protocol that is used to interface a directly-connected network. Two of the most common layer two protocols are Address Resolution Protocol (ARP) and Trailer Protocol Negotiation (TPN). TPN for link layer encapsulation can only be used if it is known that both systems involved in the communication implement trailers as well. The system should also dynamically negotiate use of the trailer protocol on a per-destination basis, and if not, then the configuration must be disabled for the protocol. The trailer protocol is an encapsulation technique that rearranges the data contents of packets sent on the physical network. These trailers can improve the throughput of higher layer protocols since it reduces the amount of data that needs to be copied. If one system exchanges packets with a system that doesn’t then this can result in strange behavior that drops random packets. The trailer negotiation happens at the time that ARP is used.

ARP will translate a MAC address to a specific IP so the packet can be routed. This is the way that the switches make their MAC address and learn about all their local links. The way this is done is the switch will send out an ARP broadcast asking each of its local nodes “who has ___ (IP address)?” The correct node will respond and then the switch will make an entry into the table stating what link that address is to go off on when a packet arrives with the ____ IP address.

2.1.3 Layer 3 - Internet Layer

The internet layer is routing layer. This is responsible for figuring out the route by looking at the destination IP address and its own subnet mask. This lets the layer 3 protocol to determine if the current destination IP is on the same network as it is using. If so, then the packet is routed to the correct port to
get to its destination. If the destination IP is on a different network, then the router sends the packet to its default gateway, or gateway of last resort. This is usually another router that handles a bigger range of networks.

The two main layer 3 headers are ICMP and IP. IP has many features that allow for error checking and special case packets. Some examples of these features include checksums and different classes for addresses. The checksum is added to the IP header to ensure that when the packet reaches the other side the packet has not been changed. If it has the checksum will not be the same and so the packet will need to be resent. The different classes in IP have purposes of their own. For instance class D is used for multicasting and class E is mostly used for testing.

2.1.4 Layer 4 - Transmit Layer

One of the transport protocols is the User Defined Protocol (UDP). This protocol offers a very minimal transport service that is not guaranteed, but allows users direct access to the datagram service of the IP layer. “UDP is used by applications that do not require the level of service of TCP or that wish to use communications services (e.g., multicast or broadcast delivery) not available from TCP.” The only features that UDP adds to the stack are checksumming of data and multiplexing by port numbers. So, any application running over UDP must deal with end-to-end communication problems like retransmission for reliable delivery, packetization and reassembly, flow control, and congestion avoidance when necessary.

The other transport protocol is the Transport Control Protocol (TCP) which is the primary virtual-circuit transport protocol for the Internet suite. “TCP provides reliable, in-sequence delivery of a full-duplex stream of octets (8-bit
TCP is used by those applications needing reliable, connection-oriented transport service.” TCP reserves specific port numbers (0 - 255) for “well known” services that are standardized across the Internet. The rest of the ports are free to use by any application that needs them. Other ports are then usually reserved to ensure that processes in the operating system that need ports get them and that they are always the same. Unlike UDP, TCP’s checksum is not optional. TCP is also responsible to handle congestion in the system. If congestion occurs, TCP has an algorithm that is called additive increase multiplicative decrease. This algorithm just cuts the transmission in half to a host when congestion occurs on their line. To get back to regular transmission rate, once congestion clears the rate is slowly increased until the regular rate is obtained or congestion occurs again and the rate is decreased in half.

2.1.5 Layer 5 - Applications

These are the specific applications that interact with the rest of the layers in the TCP suit.

2.2 How Packets are Routed (BGP)

Packets are routed using IP addresses in the current system of the internet. The first thing that happens is the source computer figures out the destination IP address and stores it in the packet. Then the source computer sends the packet to its router. Each router has a subnet mask and an IP address that give it its network address. The subnet mask is then applied to the destination IP address to see if the destination is on the same network. If so then the packet is routed to the correct port on the router. Along with the subnet mask and IP address, each router has a default gateway, which tells the router where to send packets if it isn't on the same network. This default gateway is another router
that takes care of a different network. If the destination IP’s network is not
the same as the current router’s, then the router sends the packet to its default
gateway. These steps are then repeated until the packet reaches its destination
or the packet takes 16 hops. It has been proven that any packet taking more
than 16 hops means that the destination is unreachable or does not exist.

3 Background

3.1 UETS/EFR[2]

UETS or Universal Ethernet Telecommunications Service is a new Ethernet
switching architecture that was the start of working toward a more reliable
Ethernet. This system is based on hardware switching of Ethernet frames us-
ing topological and hierarchically assigned standard local MAC addresses. This
technique is called EFR or Ethernet Fabric Routing. This new system gives
four main benefits of security, scalability, a low cost/performance ratio, and is
compatible with existing Ethernet and IP networks. A UETS domain is a net-
work composed of UETS network nodes called a CUE, network terminals called
NTE or TRUE, and end nodes called a TUE. The system also provides EDNS.
In EDNS, IP addresses and hostnames are converted to local MAC addresses or
LMAC upon request at end nodes. The end nodes can have TCP/IP and/or the
new LLC stacks. The CUE or Central Universal Ethernet is a new network con-
cept. It uses hierarchical local (U/L bit = 1) address linked to physical port ids.
Frame switching and routing is based on the LMAC destination address. Each
switch in the UETS domain has a bit has to apply to each Ethernet destination
address to ensure that packets get routed on the correct port. Just as in the
routing of IP packets this mask is applied and then the switch uses a multiplex
with the related bits to send off the packet on the correct port. The NTE is
the gateway between an Ethernet 802.1 LAN and a UETS network. This node in the system performs address translation (ENAT) or encapsulation between 802.1 universal (UMAC) and UETS LMAC addresses. This is the node that actually intercepts DNS requests and conveys it to the EDNS service. In this new system, MAC addresses are physically dependent and controlled by ISP or network owner. In other words, MAC addresses are assigned by the owner to the switch and these addresses may be hidden to service users to enhance the security even more. This new system uses a dual stack to allow for it to be compatible with the current stack. Applications use sockets to interface with layer two LLC protocol connectionless and connection oriented variants. The stack contains the OSI model and LLC on top. LLC interacts with the different layers of the OSI stack and so things can go back and forth. Unlike the current application that use ports at transport (TCP/UDP) protocol layer to id points of access in a machine and to define connection using endpoints service id, the UETS stack does this by using link service access points (LSAP) that provides interface ports for users above LLC sub layer. In the current implementation if UETS, end nodes use TCP/IP in the standard way for error correction, congestion control, and transforming packets from the current OSI model to the UETS domain. It is possible to offer simultaneous services in the Ethernet domain with the IEE 802.2 LLC protocols and in the IP domain with the IP, UDP, and TCP protocols. MAC addressing is very important in this system. The structure of MAC addressing is that UETS addressing is based on locally administered Ethernet MAC (LMAC) and set the U/L bit the packets to 1. Universal addresses (UMAC) are not allowed inside an EFR domain. EFR switching is based on the correlation of MAC addresses to the physical hardware architecture. This can be done as a standard network management activity via SNMP over Ethernet or via automatic address assignment protocols executed by EFR switches through
info exchange with the root switch and the address server. The LMAC address assignment may hierarchically according to the network topology. UETS also allows for a pseudo-IP mode since the 6 byte MAC address structure permits use of IPv4 addresses as layer two addresses embedded in the 6 byte address. Mobile and fixed addresses are assigned and a few bits in the packet are set to tell the system that these are fixed addresses. All of this was designed with the four main goals in mind. First, scalability and domains size can be whatever the administrator wants it to be from small domains up to an international world-wide domains due to it hierarchy and that it uses a 48 bit MAC address which allows for about 70 trillion unicast addresses. UETS is compatible with IEEE 802.1Q VLANs and other Ethernet encapsulations such as MAC in MAC or SNAP. For security, this system has many improvements that the current system doesn’t have. First address spoofing is not possible since the addresses are connected to physically linked ports and these are assigned and controlled by the owner of the switch. Next, there can be no attacks on the spanning tree algorithm as in the current system since there is no more spanning tree algorithm. Finally, the switches do not have address learning so no attacks can happen that overflow the buffer of the switch. UETS has a performance benefit over the OSI model. Fastest IP routers today have reached a per-chassis capacity of 640 Gbps. The UETS system has been demonstrated switching up to 10 Tbps in a single chassis with an estimated scalability of up to 50 Tbps per chassis.
4 Bibliography

References


