

White Paper

Characteristics of Switches and Routers

Mike Shepherd
Technical Partner Manager



Juniper Networks, Inc.
1194 North Mathilda Avenue
Sunnyvale, CA 94089 USA
408 745 2000 or 888 JUNIPER
www.juniper.net

Part Number: 200161-001 Jan 2006

Contents

Characteristics of Switches and Routers	1
Contents.....	2
Overview	3
Executive Summary	3
Introduction	4
Layer 2 Switches.....	5
Transparency	6
Flexibility.....	6
Scalability	7
Efficiency	8
Dependability	8
Security.....	8
Operational Issues.....	9
Layer 3 Platforms	9
IP Backbone Router definition.....	10
Layer 3 Switch definition	11
The significance of Route Aggregation	12
Layer 3 Switches and Routers - The crucial differences	13
Delay-bandwidth buffering.....	14
Congestion buffering.....	16
Local Context Addressing.....	16
Routing Algorithms	18
IP Service Creation.....	19
Interfaces	20
Conclusions.....	21

Overview

The purpose of this document is to explain the differences between Switches and Routers. A key assumption made in this paper is that Switches are Ethernet oriented devices, rather than ATM, Frame Relay or voice, and that they run TCP/IP based protocols rather than legacy enterprise class protocols such as IPX. This document also provides guidance about where to deploy each category of device, based on real operational experiences.

Executive Summary

In order to understand the differences between a Switch and a Router, we must first understand the environment in which they operate. A Switch was designed to operate in a Local Area Network, while a Router was designed to operate in a Wide Area Network. As such, the platforms' fundamental design parameters address different sets of requirements in terms of software capabilities, function optimization, dependability, performance, scale and density. Traffic patterns and demands are very different in each environment, and an approximate rule can be applied: LAN's have 80% local traffic and 20% external traffic, compared to WANs which reverse this trend. In addition, LAN traffic patterns can be flow oriented and show a degree of predictability. For example, FTP servers, Intranet servers and printer resources. This is in contrast to the largest WAN, the public Internet, where networks can be reachable via several different routes, traffic can be asymmetric in nature by returning down a different path, and traffic patterns and quantities constantly change. Another way to view this is that LAN traffic flows tend to be n to 1 , and WAN traffic flows tend to be n to n .

The 80/20 rule should be regarded as an approximate guide because new disruptive technologies, applications and organizations are constantly appearing. For example the peer-to-peer traffic driven by applications such as Skype and e-Donkey. These applications echo the origins of the ARPAnet in the pre-Internet era, which carried peer-to-peer ambitions.

Layer 2 Switches are in effect a set of learning or filtering bridges internally interconnected, which are capable of traffic reductions and loop prevention in an LAN. A Layer 2 Switch has virtually no use in a WAN because of limitations with respect to scalability, flexibility, efficiency and security. Superficially, a Layer 3 Switch looks similar to a Router. However, the Switch was conceived and designed (and architecturally optimized for) for LAN environments, which leads to shortcomings in WAN environments. The key areas of shortfall are: Lack of Delay-bandwidth-buffering, limited congestion buffering and traffic control, inferior routing algorithm robustness, limited IP service creation capability and a smaller range of interfaces.

In this paper, we examine Switch and Router capabilities, and the needs of LAN and WAN environments. By understanding these requirements we can conclude which devices are most suitable in each environment. The lower cost of a Layer 3 Switch can then be compared to the features and functions of a Router, so that a properly informed deployment decision can be made.

Introduction

An industry-wide debate has been in progress for a considerable length of time focusing on the ability of Switches to perform the functions normally associated with Routers in a much cheaper and simpler form. This is potentially true for the Internet and WANs, and also other topological areas such as Metro Area Networks (MANs), and "Last Mile" or access nodes. The wide distribution of Ethernet technology has added fuel to the debate, and in fact many service providers are now choosing Metro Ethernet Switches in their networks.

It is a commonly held belief that by deploying Ethernet, the cost of service provider operating expenses will be driven downwards and will be beneficial to both providers and their customers. This assumption is reasonable, given that history shows Ethernet port costs have consistently been driven down in Enterprise and Local Area Networks (LAN). However, applying LAN technology into a WAN environment is not a trivial matter, especially the issue of scale found in a WAN the size of the Internet. Ethernet has physical limitations such as a maximum of 1024 nodes in a single collision domain, and although an Ethernet network can be expanded far beyond this by using repeaters and Switches, it is impractical on a very large scale. Hierarchical network design is an essential scalability requirement, which is why IP is a suitable technology for global communications. Despite the issue of scale, point-to-point (full duplex) Ethernet access pipes are proving to be an effective means of delivery layer 2 services.

Before Layer 3 routing dominated data networking, Layer 2 bridged WANs were common place. Many Layer 2 technologies became available, such as Token Ring, FDDI and LANE, but over time Ethernet in its various forms emerged as the dominant technology of choice, primarily in LANs. Ethernet has evolved from 10Mbps to the current 10Gbps developments, and is even taking on SONET/SDH like attributes, as demonstrated by the IEEE's 802.17 Resilient Packet Ring (RPR) technology. Meanwhile, the Internet, which is based upon routing IPv4 protocols, has comprehensively out-scaled every Layer 2 based infrastructure. However Layer 2 is still very important to service provider networks, and is trusted to provide multi-service connectivity, for example Voice over Frame Relay and ATM. In the future, Layer 2 based MPLS VPNs will emerge as a key revenue generating technology, extending the benefits of more traditional methods over a converged IP/MPLS infrastructure. These services are also useful when transporting legacy customer's protocols, rather than IP such as SNA, IPX, DECNet and Appletalk, which are still used in specialist applications.

Many Ethernet Switches now have Layer 3 capabilities, which feature IP as the dominant protocol. More than ever before, service providers are now under pressure to reduce operating expenses, while offering new IP based services. As Layer 3 Switches are generally cheaper than Routers, many service providers are also asking: Why shouldn't Layer 3 Switches be used in the place of Routers? This issue is addressed in the following sections on Layer 3 platforms, but the key areas of difference between Switches and Routers can be

summarized in five general areas: Scalability, Reliability, Features, Management and provisioning issues, Cost.

Layer 2 Switches

At a point in the 1990's, the term "Layer 2 Switch" began to appear in the data networking industry. It was a term used to describe a LAN platform, which was capable of extremely high performance frame forwarding based on MAC layer addresses. If this sounds familiar, it is because this is the basic operation of a Bridge, but with higher performance.

When most people refer to a Bridge, they really mean a "Filtering Bridge" or "Learning Bridge". For the purposes of this document, the term Bridge (or Layer 2 Switch) is referred to in this context. A Bridge relays every frame received on any port to every other connected LAN. If a frame is received on the same LAN as its destination, then it is not forwarded to other LANs. As they only forward broadcast and non-local traffic, traffic reductions through the LAN topology can be achieved, which also allows reasonable performance over slow WAN links. These devices are designed to examine Layer 2 information such as the Media Access Control (MAC) source and destination address and act upon it. Layer 3 Switches and Routers examine more information further up the ISO 7 layer model, including IP source and destination address, and act upon it.

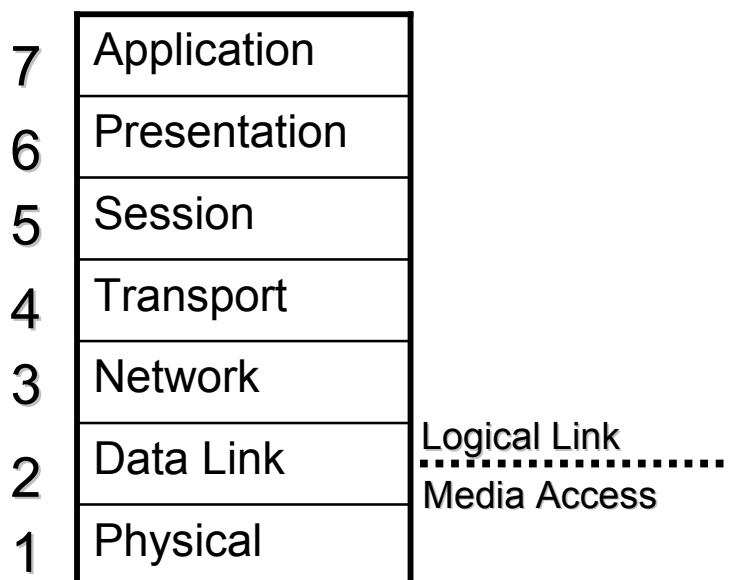


Figure 1: Figure 1. ISO 7 Layer Model

So what is a Layer 2 Switch's relationship to a Bridge? A Switch is actually a group of bridges, usually connected together in a star pattern. Each port on a Switch is really a Bridge, which is then wired to every other port in the Switch. Spanning Tree Protocol associated with Bridging, is applied to each port on a Switch. This means that Bridges and Layer 2 Switches are essentially the same things.

An Ethernet Switch in promiscuous mode listens for all frames sent to its ports. If it sees that a frame has a destination MAC address it has learned from a remote site, the frame is relayed to its destination, keeping the original source MAC address (unlike a Router). Broadcasts frames are relayed to all remote locations, also keeping the original source MAC address intact. The Switch learns and populates its MAC address table by using a discovery and exchange protocol. As only frames with errors are removed, the primary objective of the protocol is to avoid topology loops. This is so that frames do not indefinitely keep circulating, steadily consuming more resources until the network overloads.

Transparency

One of the most appealing features of a Layer 2 Switch is its protocol independence. This means that any Layer 3 protocol can traverse the Switch, such as SNA, IPX, DECNet and Appletalk, since the Switch does not inspect the Layer 3 headers. No modification to the data frame is required, unlike Layer 3 techniques, which operate with independent L2 framing headers on each side of the L3 device.

Flexibility

Flexibility and transparency are mutually exclusive, so a balance must be found between the two extremes. Transparency is associated with ease of operation and good inter-operability, however, flexibility enables features and functions, with the potential to be tailored for specific applications, and the agility to react to varying network conditions. Although a Layer 2 Switch is transparent, it is not flexible. Some examples of how flexibility is necessary in a network include:

Packet classification – This is the ability to assign packets to categories based on information contained in higher layers of protocols, such as applications. These could include differentiating between a real-time Voice over IP (VoIP) packet and File Transfer Protocol (FTP) sessions, or even email, which operates in a store and forward mode. Layer 2 Switches cannot differentiate, as they can only see MAC layer frame content.

Priority queues – This is the ability to service packets out of order, based on their relative importance or even other parameters such as delay or jitter. If a Layer 2 Switch is operating the IEEE 802.1p standard, it could act upon the priority bits contained in the MAC header, but

few devices use this facility, as it is most useful when it can be mapped to the higher application layers. This something a Layer 2 Switch cannot do, as it cannot read or act upon information in Layer 3 or higher.

Congestion avoidance – One of the techniques that could be employed is to deliberately discard packets (head drop) in order to reduce congestion and synchronization effects within end-to-end reliability protocols. Layer 2 Switches cannot participate in this, as they cannot distinguish between the packets subject to congestion avoidance.

Fragmentation – This is the ability to divide packets into smaller pieces, allowing transmission over slower speed WAN links, which may have higher bit rate errors. This is particularly important to a real-time application such as VoIP, when being transported over a slow speed link. A Layer 2 Switch is permitted to fragment frames; however, they must be re-assembled at the remote end of the WAN link. This lack of flexibility can lead to extra workload on the processor or even reduced throughput.

Sampling - This is the ability to selectively choose packets in transit for statistical analysis. Sampling helps the network planning and design process, to determine if extra peering sessions might be needed, and other uses such as detecting Denial of Service (DoS) attacks. A Layer 2 Switch cannot sample based on packet type, as it cannot read or act upon information in Layer 3 or higher.

Policy based routing – Layer 2 Switches cannot direct packets through a network based on an application or end user. The policy process is useful to allow important traffic, as defined by an administrator, to take a higher quality or cost path, which is something a Router can perform.

Scalability

Layer 2 Switches or Bridges have not become the Internetworking market's platform of choice because of their limited ability to scale. MAC addressing schemes have no hierarchy and cannot be aggregated in the way that IPv4 addresses are deployed.

Consider a scenario if the Internet was to be built with Layer 2 Switches instead of Routers. They would have to handle more than 100million address entries for the USA alone. Each Layer 2 frame would require a 6 byte address lookup in a table of 100million entries just for the USA, consuming large amounts of memory and most importantly, consuming time to perform the lookup. With Internet growth, this limitation would only become worse. To compound this, any topology change would require significant changes to address tables, on a global basis, which would have major impact when the amount of topology churn is considered in today's Internet. It is not possible to see Layer 2 Switches scaling to this level.

Efficiency

Layer 2 Switches do not make the most efficient use of available resources. For example, broadcasts are normally forwarded out of all interfaces, which might be acceptable in a LAN, but is a waste of precious and expensive WAN bandwidth. This is because some applications send broadcasts as part of their mode of operation, which do not necessarily need to traverse WAN links.

As bridging protocols such as Spanning Tree are intended to create loop free topologies, they are not optimized for the most efficient use of MAN or WAN bandwidth. Optimum route calculations or traffic engineering techniques are not defined for Layer 2, although some techniques such as IEEE 802.3ad link sharing do make a difference.

While many Switches lack deep memory buffers, they are not excluded from using them. Deep memory buffers are not usually built into Switches because of manufacturing cost considerations, which is one reason why a Switch is cheaper than a Router. Deep memory buffers are essential to the smooth operation of large bandwidth delayed links, by congestion-avoidance transport protocols such as Transmission Control Protocol (TCP). The need for Delay Bandwidth Buffers is discussed in detail in the Layer 3 Platform section.

Dependability

Layer 2 Switches can offer ways to ensure network availability and reliability. If two Layer 2 Switches are attached to the same LAN, one can act as a redundant backup for the other, using Spanning Tree Protocol calculations. Redundant links may also be used, so that if the primary WAN link fails, a parallel backup can take over.

Security

Networks using Layer 2 Switches can be vulnerable to malicious attacks, because a Switch offers little end-user protection. As broadcasts packets are sent to all remote destinations, malicious or accidental transmissions could render the network unusable, rather like a Denial of Service attack.

If a Denial of Service attack takes place, there is no opportunity for the Switch to prevent this happening, as it cannot read Layer 3 packets and therefore cannot classify traffic. Compare this to a Router which can classify traffic above the Network Layer 3, and could filter out traffic, or apply rate limiting so that the attack is minimized.

If an eavesdropper gains access to network traffic, many things can be learned simply by analyzing MAC addresses. One example of this is that the first 3 bytes of a MAC address identifies a hardware manufacturer, which could even identify the type of hardware.

In order to prevent eavesdropping, it is desirable to encrypt network traffic. However, this option is not commercially available for transparent LAN encryption in a cost effective platform. While Layer 2 Switches do not usually provide cryptographic services, they do offer Virtual LAN (VLAN) technology, to provide logical separation of traffic on common physical links. Switches are also capable of passing encrypted traffic through their ports.

Operational Issues

Layer 2 Switches are relatively simple devices to operate, however, this also means they have some limitations. On initial examination, guaranteed MAC addresses per hardware vendor is an advantage, as no public address space co-ordination is required. Contrast this to the efforts dedicated to address space co-ordination by the Internet authorities. However, this apparent advantage is really a limitation, because the lack of co-ordination means that resources cannot be shared between customers. Sharing resources within a Service Provider, such as Domain Name Servers (DNS), email, World Wide Web, and so on, is the very essence of why the Internet has become an essential business tool.

If a Service Provider deploys Layer 2 Switches, OSS practices must be carefully considered. For example, there is no standard way to share a Network Management station across multiple customers, in order to monitor customer premises equipment. This could impact network reliability, Service Level Agreements and billing.

Layer 3 Platforms

It has long been automatically accepted that Routers are the essential platforms required to build IP and Internet Backbones, and WAN data connections in general. A good generic definition of an IP Router is: A device which moves IP datagrams across an Internetwork from source to destination. However, from a distance, this seems remarkably like a Layer 3 Switch, which operates at a similar level in the ISO 7 Layer model. This could easily cause confusion when determining the purpose of each platform, so a deeper understanding of their respective operation is required. The section attempts to give an explanation, and compare the

two architectures, so that it becomes clear why a Layer 3 Switch is very different to a Router, and therefore each should be deployed in different topological positions.

IP Backbone Router definition

The basic anatomy of a Router is made up of the ability to:

- Switch IP datagrams (Layer 3 forwarding)
- Symmetrical any-port-to-any-port Switching speed,
- Delay-bandwidth buffering plus congestion control,
- Internet scale routing protocols (IS-IS, OSPF, MPLS, BGP),
- Internet scale IP address table handling.

The advanced anatomy of a modern Router includes:

- Wire-rate forwarding on all ports for any packet size distribution,
- Performance independent of load or external disruption
- Support of CoS queuing, shaping and policing
- Traffic engineering, classification and filtering at wire-rate
- Security
- Carrier-class availability

These elements can be distilled into three functions:

Routing algorithms – Whatever protocol is chosen (BGP, MPLS, OSPF, IS-IS), the optimum route is calculated, so the best and most efficient paths through a topology are selected. The least amount of Router and network resources should be consumed when performing the calculations and the route calculations should be separate from the forwarding decisions. They should be as robust and stable as possible, whatever the environment, including exposure to link failures, traffic loading or topology changes. The calculations should be performed as quickly as possible, so that a common network view is shared among peers, otherwise routing loops could occur. As was previously mentioned in the Layer 2 Switch section, loops should be avoided as they cause networks to overload. Finally, Routers should have the flexibility to adapt to changes in bandwidth, congestion delays, traffic levels and so on.

Switching and Forwarding – Once the best routes through a network have been selected, the Router should forward traffic to its chosen destination at maximum speed. This means that forwarding performance should not be affected by influences such as topology changes. Legacy Router architectures suffered because whenever a topology change occurred, processing resources were consumed by re-calculating routing information, which impacted

forwarding performance. This is far from ideal, especially as the Internet is a hostile environment to stability, where routing and topologies are constantly fluctuating. The answer is to separate routing calculations from the forwarding process, an architecture which has proven to be essential in a modern Internet Router.

Service creation – The most underrated aspect of a Router is its service creation ability, which is directly related to generating Service Provider revenue. The very essence of why IP and the Internet are becoming ubiquitous is their any-to-any nature. Internet scale routing allows anyone to connect to anyone within or outside of their own organizations. However, a Router no longer just provides Internet connectivity; many other revenue generating services can be created and supported by an IP Router. Internet radio, IPTV/video-on-demand and push-web services all rely on IP multicast technology (something a purely optical transport network cannot do). VPNs are also regarded as an area of high revenue opportunity for both Layer 2 and Layer 3 services. Where a layer 2 MPLS based VPN could be very suitable for MAN services, Layer 3 MPLS VPNs could also be very suitable for Small and Medium Enterprises (SME). Content awareness and sensitivity is also a Router service feature. For example the ability to direct web traffic, or complement caching servers in a hosting environment could be combined with security services such as firewalls, access filters or even cryptography. A Router can also differentiate between applications, so that it could assign a different priority to File Transfer Protocol (FTP), than HTML web traffic. This ability means that a Router can operate in the higher layers of the ISO 7 layer model, above Layer 3 (Network).

Layer 3 Switch definition

As has been previously discussed, a Layer 3 Switch shares many of the same attributes associated with a Router. Long before the term Layer 3 Switch was coined, Routers were the primary devices deployed in LANs that could Switch IP datagrams. However, LAN bandwidth was (and still is) relatively cheap, compared to expensive WAN or MAN connections, which lead to the deployment of high performance LANs. Legacy software based Routers could not keep pace with the higher interface speeds or packet processing demands, so Layer 3 Switches were created to solve that problem.

Essentially, Layer 3 Switches, Switch IP datagrams, with most of the forwarding process based in hardware ASICs rather than in software. However, what is notable is that a legacy Router regarded its layer 2 Bridging module as a peer to the Network layer protocols (this gave rise to the term BRouter, which was a combined Bridge and Router). This contrasts with a Layer 3 Switch, which regards the layer 3 traffic as a process above Layer 2 Switching. The advantage of this is that, in a LAN environment where it is easy to keep track of a smaller domain of MAC addresses, frames can be switched on a layer 2 basis extremely quickly. When traffic needs to be sent outside of the MAC domain, then a layer 3 routing type lookup can be performed. This is efficient in a LAN environment, but it means that the Network layer is not involved for all traffic decisions, and therefore all of the disadvantages described with Layer 2 Switches might apply.

As they were conceived to solve a LAN issue, Layer 3 Switches mainly feature LAN related

Interfaces such as Ethernet. However, more recently, they have begun to adopt WAN type interfaces such as SONET/SDH. Many vendors claim that Layer 3 Switches are wire-speed IP Routers, and they are suitable for LAN, MAN and WAN environments. One thing is clear, by simply adding a WAN interface to a Layer 3 Switch does not create a WAN Router. The following section clearly explains why this is the case, and that Layer 3 Switches are optimized for LAN environments and not WAN or MAN environments.

The significance of Route Aggregation

A brief check of the statistics announced by the Reseaux IP Européens (RIPE) organization shows that as of January 2006, there were approximately 177,000 routes being announced on the public Internet. The important point to realize is that this is a summary of every network that is reachable on the global Internet, which could number in the millions. This aggregation conserves precious infrastructure resources, such as CPU processing time, memory, transmission bandwidth and more. This aggregation can only be effectively performed by Routers, and without it, the Internet would not exist at the scale we rely upon today.

The practice of route aggregation was made possible by IETF standards such as RFC1517 – Classless Inter-Domain Routing (CIDR), and a Router’s ability to perform “longest prefix matching” on every packet. The practice of route aggregation with variable length prefixes allows an upstream Router to efficiently advertise reachability to multiple downstream networks, which may be accessible via several different paths. For this reason, a Router must select a path using longest prefix matching, by choosing the most specific advertisement to a given network. For example, if a network is reachable through /24 and /28 advertisements, then the /28 will be chosen as it is the longest prefix match, and therefore the most specific, irrespective of the routing protocol used. This process takes place for every single packet that a Router receives, and is one of the reasons why an Internet backbone Router is so difficult to build.

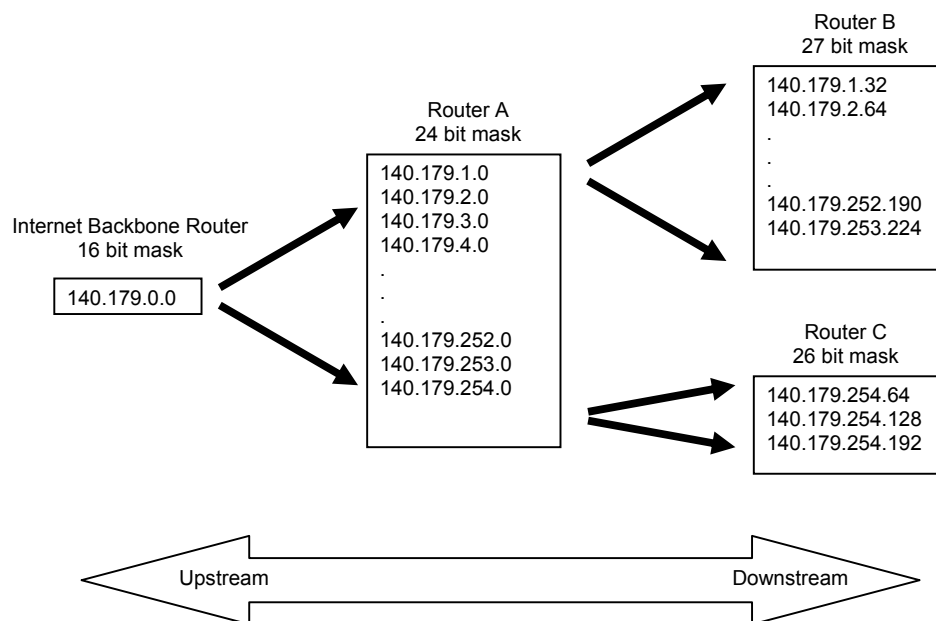


Figure 2: Route aggregation with Variable Length prefixes

Layer 3 Switches and Routers – The crucial differences

It is no accident that large-scale IP networks all rely upon IP Routers. IP Routers have also played a fundamental role in the continuing evolution of communications from The Internet as we currently know it into the next generation of global communications being described by the IPsphere Forum. The key is their capacity to scale on a global level with performance (in a highly dependable manner). The areas of focus are: Delay-bandwidth buffering and congestion buffering, local context addressing, routing algorithm robustness, IP service creation for multiple services using QoS, and interface diversity.

IP Routers themselves have evolved to drive and support the development of 21st Century communications. A key example of innovation within a modern IP Router is the groundbreaking way that route lookups are performed within Juniper Networks Routers. This has made a significant impact in handling the route aggregation demands which were outlined in the previous section of this document, by speeding up prefix matching, route installation, scale, reliability and other carrier class attributes.

JTREE lookups - independent of prefix length looks only at the *differential* bits

Looking at a tree with two prefixes
 208.197.168/23
 208.131.192/18

In binary these prefixes are:

11010000 11000101 1010100
 11010000 10000011 11

The highest order (i.e. first) bit where the prefixes differ is bit number 9. This is the bit JUNOS tests when doing a route lookup.

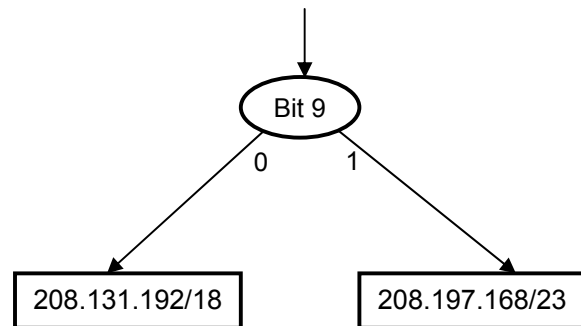


Figure 3: Ground-breaking way of route lookups

Delay-bandwidth buffering

In order to understand the significance of Delay-bandwidth buffering, we must first understand a fundamental mode of operation of Transmission Control Protocol (TCP). Unlike its Layer 4 sister protocol User Datagram Protocol (UDP), TCP is session oriented, maintains connection state and controls the data flow rate to optimize session performance for the characteristics of the transport path.

When a TCP session starts up, it initiates in slow mode. After the first packet is transmitted, it waits for an acknowledgement from the receiver, giving a total time known as the Round Trip Time (RTT). It then immediately sends two packets (double the initial transmission) and waits for another acknowledgement, which again is associated with the total Round Trip Time. This is an exponential increase in the number of packets transmitted each time an acknowledgement is received, so that the next time 4 packets are transmitted, then 8, then 16 and so on. The transmission of packets at a rate of twice the previous rate within each RTT interval is smoothed out by the amount of bandwidth available in the end-to-end path between source host and receiving host. The slow start sequence will continue until packet loss occurs, then increases become linear, not exponential. Packet loss can occur for a number of reasons, but most commonly where congestion takes place because too much traffic is trying to pass through a WAN connection of a fixed size.

If packets are lost, the TCP window size will decrease and so the transmission rate is reduced to match the available bandwidth for the session from end-to-end. Throughput recovery is achieved by using the slow start mechanism. In order to achieve maximum transmission efficiency in times of temporary congestion, the Routers supporting the congested links must be able to buffer sufficient traffic such that the effective RTT is extended to the optimum

before packet loss occurs. The optimum amount of this “Delay Bandwidth Buffering” is a function of the uncongested propagation delay of the transmission path between the source and destination hosts.

TCP allows for outstanding, or unacknowledged, data up to a total of its window size. This can typically be up to 64 Kilobytes, and is derived from the feedback loop of available bandwidth multiplied by Round Trip Time ($BW \times RTT$). This value determines how long the source will take to receive an acknowledgement from the destination. When a packet is lost due to an event such as network congestion, the source will not be aware of the event and so will be unable to react until $BW \times RTT$ bits has been transmitted.

This means that the TCP throughput data rate is based upon the amount of data that can be loaded into the end-to-end path, divided by the propagation delay. If the buffer of an intermediate system in the end-to-end path is less than the propagation delay multiplied by link bandwidth ($Delay \times BW$), then no more data can be sent until an acknowledgement is received. Therefore, the buffer size in an intermediate system, in other words a Router, must be greater than 2 times the $RTT \times BW$ in order to optimize data transmission performance.

So, $RTT \times BW =$ total number of bits in transit between sender and receiver

TCP throughput will degrade based upon the formula $1 / (1 + 2pw)$

p = packet loss probability

w = bandwidth \times delay product

This means that there is a direct relationship between the bandwidth of a link, and the amount of buffering needed to achieve maximum performance by avoiding packet loss.

A more complete description of these mechanisms can be found in the following papers:

http://www.juniper.net/solutions/literature/white_papers/200019.pdf

http://www.juniper.net/solutions/literature/white_papers/200022.pdf

Buffering packets should be performed throughout the complete end-to-end path, and not just the edge of a service providers’ network. However, the edge of a network is the most critical place to buffer packets, as this is where congestion is most likely to occur due to speed mismatches or uplink oversubscription. Consider the case where many Gigabit Ethernet connections are feeding into an OC-48c/STM-16 connection. This could be a typical scenario when connecting a MAN to a WAN, which could lead to a bottleneck and therefore packet loss if devices with insufficient buffering are deployed.

In summary, the need for Delay-bandwidth buffering is directly related to the fundamental operation of TCP and the overall propagation delay of the session from end to end. Optimal

throughput cannot be successfully achieved without the presence of Delay-bandwidth buffers. If a link speed of 10Gbps, OC-192c/STM-64, is considered, then a huge amount of buffering must be associated with this link to in order avoid sub-optimal TCP transmission, of the order of greater than twice $RTT \times 10Gb$. As almost no Delay-bandwidth buffer memory is required in a LAN environment, it follows that Layer 3 Switches usually have no Delay-bandwidth buffer memory. This makes Layers 3 Switches cheaper than Routers, but more importantly, they are not optimized to run in WAN environments.

Congestion buffering

This type of buffering is similar to Delay-bandwidth buffering. When outbound interfaces become congested, sufficient memory must be available to buffer packets into one or many queues while they are prioritized for transmission. This is related to Class of Service (CoS) features, where a platform should be capable of classifying traffic according to its importance.

Once again, a general assumption with Layer 3 Switches is that in a LAN environment, lots of bandwidth is available and therefore congestion does not often occur. In a LAN, traffic tends to flow in any-port-to-any-port patterns, where port speeds are relatively equally matched. This could also be regarded as n to n connectivity. This is in contrast to a WAN environment, where expensive bandwidth tends to cause speed mismatches and therefore traffic congestion. Traffic flows tend to head from many high-speed LAN ports to a single lower speed WAN port when an Enterprise connects to the Internet. This could be regarded as n to 1 connectivity. A specific example would be a Fast Ethernet based LAN connection with layer 3 traffic flowing out to a T1/E1 WAN connection. The mismatch in speed can result in congestion, causing delayed and dropped packets. The correct device to handle these types of traffic flows is a WAN Router.

Even in a MAN environment with Gigabit Ethernet connection uplinks to the wide area backbone, multiple fast Ethernet connections will be sharing the uplink in an oversubscribed manner so temporary congestion conditions can occur.

As less queue buffer memory is required in a LAN environment, it follows that Layer 3 Switches usually have far less queue buffer memory than a Router. This makes Layers 3 Switches cheaper than Routers, but most importantly, they are not optimized to run in WAN environments.

Local Context Addressing

Consider the design criteria for a typical Layer 3 Switch. Incoming packets are examined and initially, route lookups and forwarding are performed in the "slow path". The slow-path is a normal software based Router engine, which performs a longest prefix match operation.

These new addresses are used to populate the fast-cache, which takes into account the most popular traffic patterns, by ranking the most used flows at the top of the table (therefore allowing the most used addresses the fastest lookups). It is normal for the cache table to be capable of up to 32,000 entries, containing the necessary forwarding information for packet lookup and forwarding. When a packet arrives at a Switch's port, the cache is checked for a match, by performing a sequential search and if found, is forwarded in hardware. If a cache-table match is not found, then a lookup is performed in the slow-path, and the entry at the bottom of the cache (least used) replaced with the new entry.

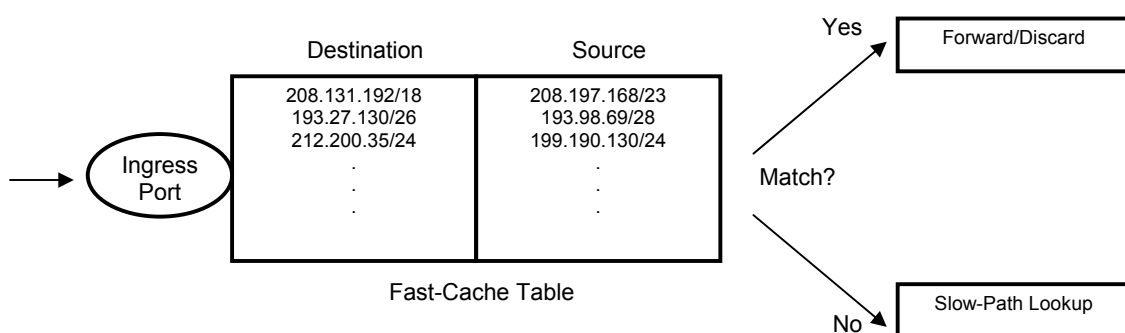


Figure 4: Fast-cache operation

The fast cache means that additional time is not wasted making a lookup in the routing engine for any address contained in the caching space. The fast-path caching mechanism is also much easier to develop and implement than the "longest prefix match for every packet" operation used on every single packet by Routers. Modern Internet Routers use ASIC's in the forwarding path which does not affect their performance, irrespective of external conditions. Because the Layer 3 Switch cache contains discrete IP addresses, an "exact match lookup algorithm" can be used. The first problem with this technique is scalability. Caching discrete IP addresses supports 10's of thousands of hosts, while the Router method of supporting IP prefixes supports millions of networks and hosts (see route aggregation section). The second problem is of predictable performance. The caching method creates fast and slow paths through the device, which are dependent on traffic patterns and the number of destinations to be supported. Therefore external conditions do affect Switch performance, particularly when the cache size does not contain the necessary forwarding information and the slow path must be used. This makes a Layer 3 Switch particularly vulnerable when connected to the Internet.

In a LAN topology, a Layer 3 Switch works very well, as it rarely uses more than 32,000 discrete IP addresses (size of the fast-cache). The Layer 3 Switch is designed for an environment where 80% of the traffic stays in the LAN, and 20% goes over the WAN or foreign destinations. With this assumption, a Layer 3 Switch can operate with fewer amounts of memory and other expensive resources for forwarding table address storage. It can also be much less sophisticated in how forwarding tables are created, which are reference by the hardware. However, this rule is reversed in a WAN environment, where 80% of traffic goes to a foreign destination and 20%, or even less, stays within a local area. By the start of 2006, the Internet routing table contained nearly 178,000 routes; therefore by applying the WAN 80/20, more than 142,000 aggregated routes should be maintained in an address table. This

immediately breaks the limits of the 32000 spaces in a Layer 3 Switch fast cache. The result is that it would be very common for a forwarding decision to be made by the Layer 3 Switch routing engine, rather than the fast cache if deployed in a WAN context. This whole process impacts forwarding performance.

It should be concluded that while Layer 3 Switches are cheaper than Routers as they use simpler traffic forwarding techniques, and less sophisticated mechanisms for populating the forwarding tables, they under-perform when deployed in a WAN and particularly the Internet, and large scale complex environments.

Routing Algorithms

The role and significance of routing algorithms has previously been discussed in the section "IP Backbone Router definition". Layer 3 Switches now support BGP, OSPF, MPLS, and in some instances ISIS too, however, scalability of these protocols is a fundamental requirement in a WAN, especially the Internet. Protocol scalability is not simply about building a platform with enough address space. Robustness and stability is essential, which is exactly what a modern Internet Router is designed for. An example of this is the way that a forwarding table should be updated. The Internet's topology is constantly fluctuating, and triggering re-calculations of routing table information and therefore changes to forwarding tables. If the routing engine function is separated from the forwarding engine function, then route re-calculations will not impact forwarding performance, so that maximum performance can be achieved regardless of external topology influences. Furthermore, once the re-calculation has taken place, only the changes should be inserted into the forwarding table, and all other entries maintained. These changes are "atomically" stitched into the forwarding table in a single clock cycle, so that no forwarding interruptions take place, with no impact on maximum performance.

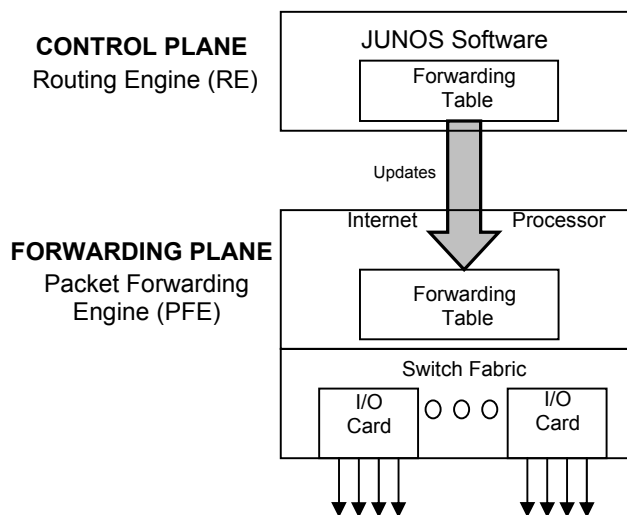


Figure 5: Modern IP Router Architecture

If the IETF's list of Request For Comment (RFC) pages is studied, it is noticeable how few authors have been involved in the creation of routing protocols. As this level of expertise and experience is extremely rare, any platform manufacturer with this pool of expertise has a distinct advantage over those that do not.

IP Service Creation

As previously mentioned in the section "IP Backbone Router definition", IP service creation is one of three fundamental functions of a Router. Layer 3 Switches are increasingly adopting these capabilities, but once again they begin to fall short when they are deployed in WAN environments. Typically, IP service creation happens at the first layer 3 platform that the traffic meets within the service provider's network, normally the service provider access Router. At that point in the network, it is essential that the platform has a rich set of service related features, consistent with the features offered by WAN backbone itself while maintaining consistent (dependability, QoS) high performance to satisfy the user experience. A hardware based Router meets this need at scale, while providing performance without compromise.

The following list is an example of IP based services which could be deployed and provide competitive differentiation:

- Triple Play
- VoIP
- IPTV
- IMS
- Fixed-Mobile Convergence
- Real-time gaming & gambling
- Broadband Services
- Value Added Internet Transit
- IP VPNs (L2, L3, VPLS, IPSec)
- Flexible Bandwidth
- Usage Billing
- Small Medium Enterprise packaged connectivity
- Virtual Leased Line
- Asymmetric Access
- Frame Relay/ATM Gateway
- Differential QoS

- Security and Denial of Service (DoS) protection Services
- Hosting Packages
- Destination based usage Billing
- Lawful Intercept

Supporting Router (or Layer 3 Switch) features include:

- Billing & Accounting
- Rate Limiting
- Securing Infrastructures
- Routing Protocols
- Sampling
- Filtering
- Flexible counting
- MPLS Traffic Engineering & DS-TE
- Scalability
- Layer 2 tunneling
- Performance

Interfaces

As Layer 3 Switches were designed for LAN's, their primary focus is on Ethernet as the framing technology. Layer 3 Switches have begun to adopt many WAN interfaces such as OC-12c/STM-4 and OC-3c/STM-1, but less common are many others such as Channelized interfaces. Their internal architecture is optimized towards Ethernet technology, and so is the cost model of the platforms. For this reason, many Layer 3 Switch vendors are proposing Ethernet as the best technology in a Metro environment. The Layer 2 framing layer can be considered as largely irrelevant compared to the qualities needed for MANs and WANs. As previously discussed, these qualities are oriented around scalability, robustness and performance without compromise.

It is also noteworthy that Interfaces tend to be the most expensive components in a Router or Switch. This is particularly true for the highest speed interfaces, which for industry-leading routers is currently 40Gbps with OC768c/STM-256. This is because of the advanced LASER optics used and framer ASICs used, and is just as applicable to a 10Gigabit Ethernet interface, as it is for OC-192c/STM-64. Therefore expensive interfaces can apply to LAN, MAN and WAN environments, so it is even more important that a Layer 3 platform can drive these interfaces at full speed in any conditions. The Layer 3 platform that does this best is a Router, and gives a better Return On Investment (ROI) over the active life of its deployment than a Layer 3 Switch.

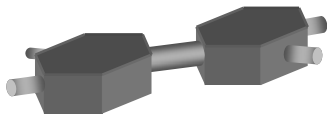
Conclusions

The Global rise of IP infrastructures in recent years has illustrated an important shift to TCP/IP connectivity, with IP/MPLS, fast becoming the underlying technology of choice. The move towards IP as the communications technology of choice is being used to consolidate service provider infrastructures onto a common technology platform, and provide a rich set of new services to consumers and business. The key component to this radical shift is the modern IP Router, rather than other classes of devices such as a Layer 2 Switch, which has no IP intelligence. Layer 3 Switches also have a role to play, but have limitations with respect to the next generation of public networks being built for the 21st century.

WANs and the Internet require high performance IP routing in large network topologies. In the same environment, Layer 3 Switches frequently resort to slow –path routing engine look-ups, which result in poor and unpredictable performance, coupled with inferior scalability.

Value added services such as IP-VPNs and high performance multicast delivering streaming media services impose a tax on the performance of a Provider Edge node in an access Point of Presence (PoP). A modern hardware based IP Router can perform these tasks without compromising performance, scalability or other services which potentially run on the same platform. This is not the case with Layer 3 Switches, which tend to exhibit performance degradation when more services are introduced, especially in large network topologies. Only a Carrier Grade Router can offer the necessary solutions to bandwidth congestion, stability under extreme stress and scalability on a Global level. Classification and prioritization of different traffic types will become pivotal as networks are designed and built to handle a wider variety of traffic. This could mean identifying a voice over IP stream from other data traffic and treating it appropriately, to meet Service Level Agreements and billing requirements.

Layer 2 and Layer 3 Switches are essential platforms, which provide great value when deployed in the correct environments. However, a Router provides superior value and ROI for both WAN and Internet and now IPsphere defined environments. This is due to superior features, scalability, robustness and performance. Routers are necessary to build the Next Generation Networks described by the IPsphere Forum!



www.IPSphereforum.org

A **network** that combines the **reach** of the Internet with the **performance and security** of private networks to **support ALL communications**

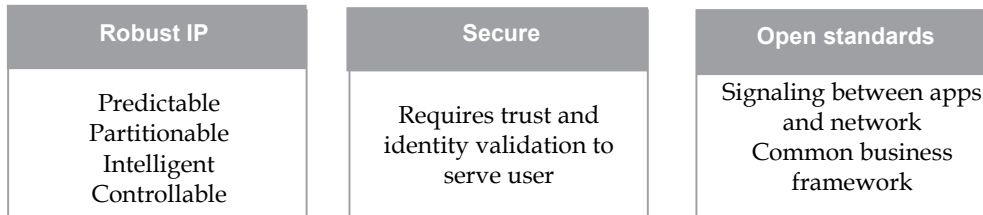


Figure 6: Goals of the IPSphere Forum

Copyright © 2006, Juniper Networks, Inc. All rights reserved. Juniper Networks and the Juniper Networks logo are registered trademarks of Juniper Networks, Inc. in the United States and other countries. All other trademarks, service marks, registered trademarks, or registered service marks in this document are the property of Juniper Networks or their respective owners. All specifications are subject to change without notice. Juniper Networks assumes no responsibility for any inaccuracies in this document or for any obligation to update information in this document. Juniper Networks reserves the right to change, modify, transfer, or otherwise revise this publication without notice.