TERABIT SWITCHES AND ROUTERS

Amit Singhal

ABSTRACT

Just a few years back, no one would have thought that internet traffic will increase at such a rapid rate that even gigabit capacity routers in the backbone will be insufficient to handle it. Today, routers with terabit switching capacities have become an essential requirement of the core backbone networks and gigabit routers find their place at the mid-core or even the edge. This survey paper explains the issues in designing terabit routers and the solutions for them. It also discusses about some of the key products in this area.

Other Reports on Recent Advances in Networking
Back to Raj Jain's Home Page

Raj Jain is now at
Washington University in Saint Louis
Jain@cse.wustl.edu
http://www.cse.wustl.edu/~jain/

Table of Contents:

---

1. INTRODUCTION

In the present network infrastructure, world's communication service providers are laying fiber at very rapid rates. And most of the fiber connections are now being terminated using DWDM. The combination of fiber and DWDM have made raw bandwidth available in abundance. 64-channel OC-192 capacity fibers are not uncommon these days and OC-768 speeds will be available soon. Terabit routing technologies are required to convert massive amounts of raw bandwidth into usable bandwidth. Present day network infrastructure is shown in Fig 1. Currently, Add/Drop multiplexers are used for spreading a high-speed optical interface across multiple lower-capacity interfaces of traditional routers. But carriers require high-speed router interfaces that can directly connect to the high-speed DWDM equipment to ensure optical inter operability. This will also remove the overhead associated with the extra technologies to enable more economical and efficient wide area communications. As the number of channels transmitted on a single fiber increases with DWDM, routers must also scale port densities to handle all those channels. With increase in the speed of interfaces as well as the port density, next thing which routers need to improve on is the internal switching capacity. 64-channel OC-192 will require over a terabit of switching capacity. Considering an example, a current state-of-the-art gigabit router with 40 Gbps switch capacity can support only a 4-channel OC-48 DWDM connection. Four of these will be required to support a 16-channel OC-48 DWDM connection. And 16 of these are required to support 16-channel OC-192 DWDM connection with a layer of 16 4::1 SONET Add/Drop Multiplexers in between. In comparison to that a single router with terabit switching capacity can support 16-channel OC-192 DWDM connection. With this introduction, we now proceed to understand what is required to build full routers with terabit capacities.

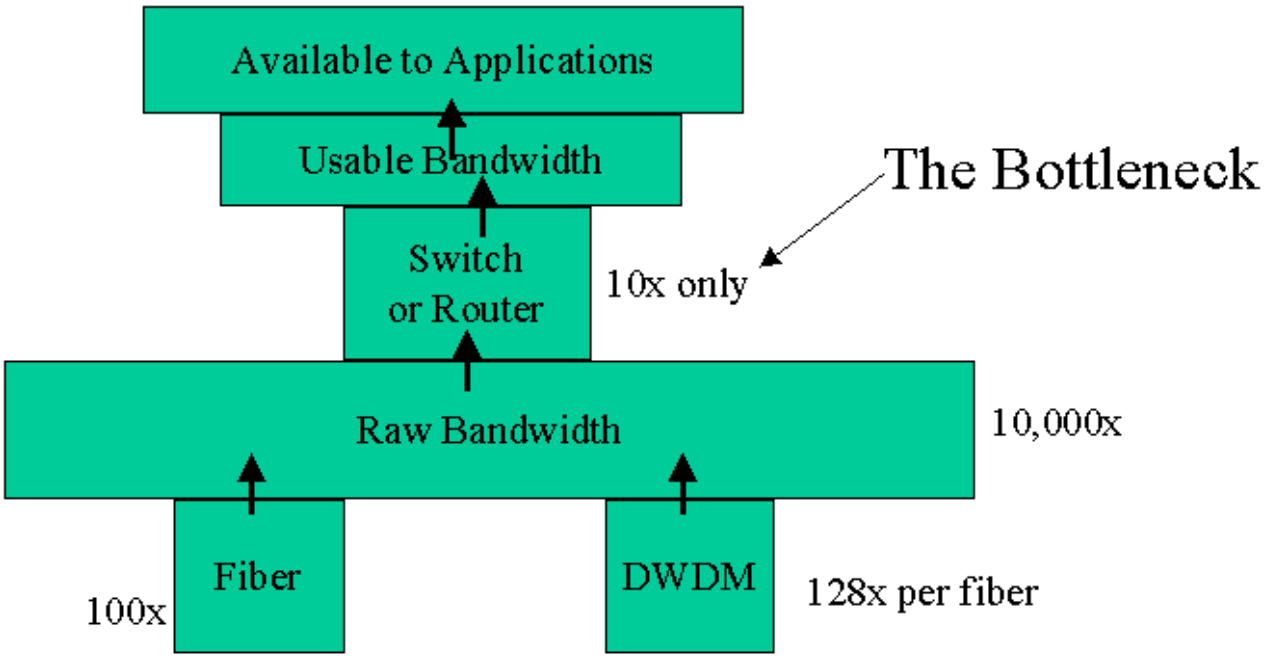

Fig 1. Present Day Network Infrastructure

Back to Table of Contents

---

2. THE ARCHITECTURE OF INTERNET ROUTERS

This section gives a general introduction about the architecture of routers and the functions of its various

components. This is very important for understanding about the bottlenecks in achieving high speed routing and how are these handled in the design of gigabit and even terabit capacity routers available today in the market.

2.1 Router Functions

Functions of a router can be broadly classified into two main categories [Nick97]:

1. Datapath Functions : These functions are applied to every datagram that reaches the router and successfully routed without being dropped at any stage. Main functions included in this category are the forwarding decision, forwarding through the backplane and output link scheduling.

2. Control Functions : These functions include mainly system configuration, management and update of routing table information. These does not apply to every datagram and therefore performed relatively infrequently.

Goal in designing high speed routers is to increase the rate at which datagrams are routed and therefore datapath functions are the ones to be improved to enhance the performance. Here we discuss briefly about the major datapath functions :

- *The Forwarding Decision*: Routing table search is done for each arriving datagram and based on the destination address, output port is determined. Also, a next-hop MAC address is appended to the front of the datagram, the time-to-live(TTL) field of the IP datagram header is decremented, and a new header checksum is calculated.
- *Forwarding through the backplane*: Backplane refers to the physical path between the input port and the output port. Once the forwarding decision is made, the datagram is queued before it can be transferred to the output port across the backplane. If there are not enough space in the queues, then it might even be dropped.
- *Output Link Scheduling*: Once a datagram reaches the output port, it is again queued before it can be transmitted on the output link. In most traditional routers, a single FIFO queue is maintained. But most advanced routers maintain separate queues for different flows, or priority classes and then carefully schedule the departure time of each datagram in order to meet various delay and throughput guarantees.

2.2 Evolution of Present Day Routers

The architecture of earliest routers was based on that of a computer as shown in Fig 2. It has a shared central bus, central CPU, memory and the Line cards for input and output ports. Line cards provide MAC-layer functionality and connects to the external links. Each incoming packet is transferred to the CPU across the shared bus. Forwarding decision is made there and the packet then traverses the shared bus again to the output port. Performance of these routers is limited mainly by two factors : first, processing power of the central CPU since route table search is a highly time-consuming task and second, the fact that every packet has to traverse twice through the shared bus.

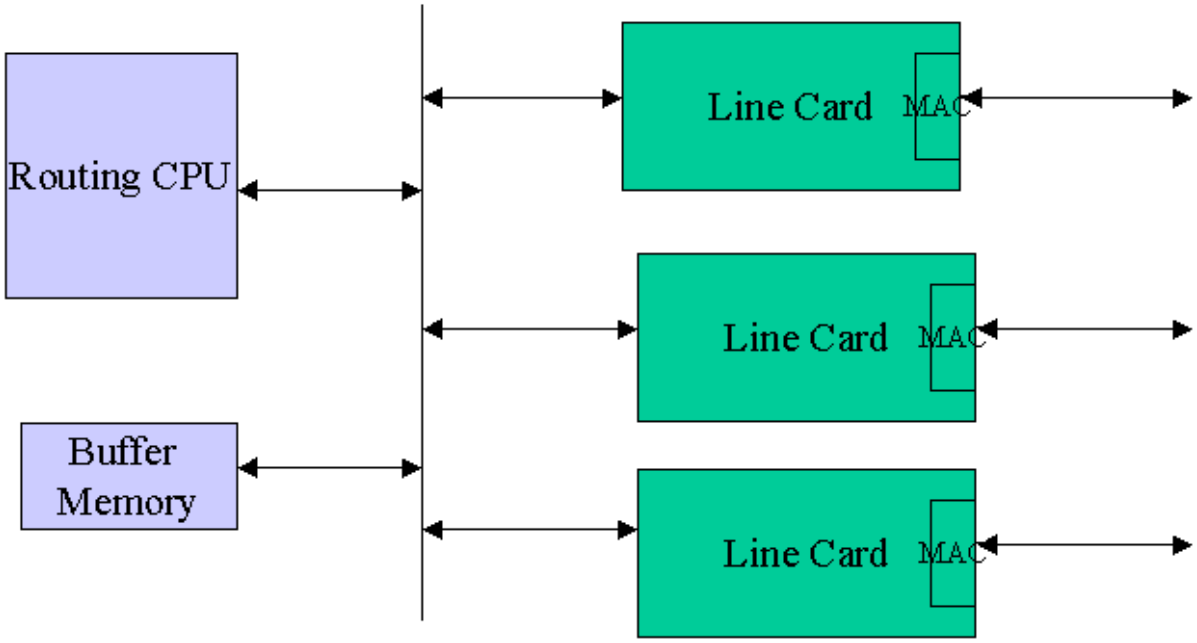

Fig 2. Architecture of Earliest Routers

To remove the first bottleneck, some router vendors introduced parallelism by having multiple CPUs and each CPU now handles a portion of the incoming traffic. But still each packet has to traverse shared bus twice. Very soon, the design of router architecture advanced one step further as shown in Fig 3. Now a route cache and processing power is provided at each interface and forwarding decisions are made locally and each packet now has to traverse the shared bus only once from input port to the output port.

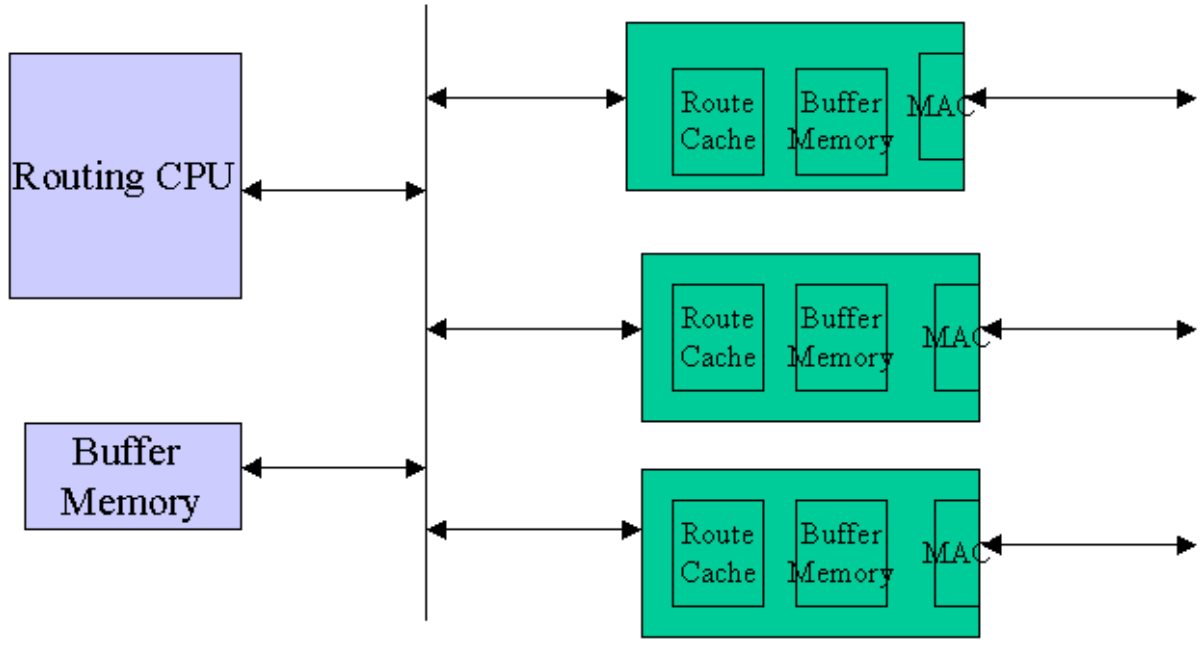

Fig 3. Router Architecture with intelligence on each line card

Even though CPU performance improved with time, it could not keep pace with the increase in line capacity of the physical links and it is not possible to make forwarding decisions for the millions of packets per second coming on each input link. Therefore special purpose ASICs(Application Specific Integrated Circuits) are now placed on each interface which outperform a CPU in making forwarding decisions, managing queues and arbitration access

to the bus.

But use of shared bus still allowed only one packet at a time to move from input port to output port. Finally, this last architectural bottleneck was eliminated by replacing shared bus by a crossbar switch. Multiple line cards can communicate simultaneously with each other now. Fig 4. shows the router architecture with switched backplane.

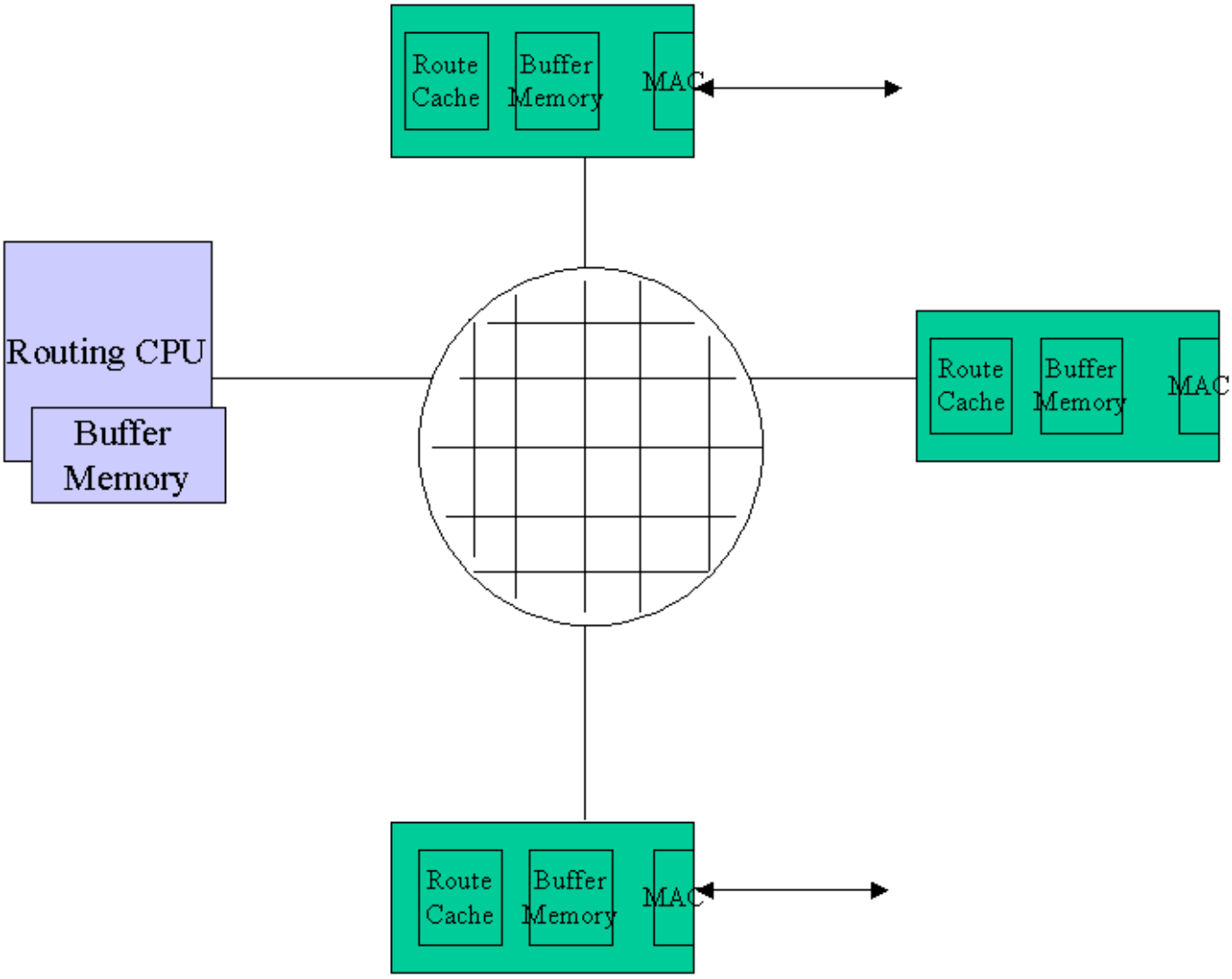

Fig 4. Router Architecture with switched backplane

2.3 Assessing Router Performance

In this section, several parameters are listed which can be used to grade the performance of new generation router architectures [NexSup]. These parameters reflect the exponential traffic growth and the convergence of voice, video and data.

- *High packet transfer rate*: Increasing internet traffic makes the packets per second capacity of a router as the single most important parameter for grading its performance. Further, considering the exponential growth of traffic, the capacity of routers must be scalable.
- *Multi-service support*: Most of the network backbones support both ATM and IP traffic and will continue

to do so as both technologies have their advantages. Therefore routers must support ATM cells, IP frames and other network traffic types in their native modes, delivering full efficiency of the corresponding network type.
- *Guarantee short deterministic delay*: Real time voice and video traffic require short and predictable delay through the system. Unpredictable delay results in a discontinuity which is not acceptable for these applications.
- *Quality of Service*: Routers must be able to support service level agreements, guaranteed line-rate and differential quality of service to different applications, or flows. This quality of service support must be configurable.
- *Multicast Traffic*: Internet traffic is changing from predominantly point-to-point to multicast and therefore routers must support large number of multicast transmissions simultaneously.
- *High Availability*: High speed routers located in the backbones handle huge amounts of data and can not be turned down for upgrades etc. Therefore features such as hot-swappable software tasks- allowing in-service software upgrades are required.

Back to Table of Contents

---

3. SWITCHING Vs ROUTING

The basic difference between switching and routing is that switching uses 'indexing' for determining the next hop for a packet in the address table whereas routing uses 'searching'. Since indexing is O(1) operation, it is much faster than any search technique. Because of this, many people started thinking about replacing routers with switches wherever possible and vendors flooded the market with several products under the name of switches. To differentiate their products, vendors gave different names to them like Layer 3 Switch, IP Switch, Layer 4 Switch, Tag Switch etc. and regardless of what a product does, it is called a switch [Decis97] [Decis96][Torrent]. Therefore it is important to understand the difference between all these different forms of switches.

3.1 Switching Hubs

It operates at Layer 1 of the OSI networking model. Individual ports are assigned to different LAN segments as in a bridge. But while they are useful for managing configuration changes, it must be noted that they still propagate contention among their ports and therefore different from layer 2 bridges.

3.2 Layer 2 Switching

Layer 2 Switches is just another name for multiport bridges. As we know, bridges are used to extend the LANs without extending the contention domain. So Layer 2 switches have been used in some places to replace routers for connecting various LANs to produce one big flat network. But the problem with this approach was the broadcast traffic which is propagated across all ports of a Layer 2 switch. To solve this problem, people soon came up with the concept of "Virtual LAN" or VLAN. Basic feature of VLAN is to divide one large LAN connected by layer 2 switches into many independent and possibly overlapping LANs. This is done by limiting the forwarding of packets in these switches and there are several ways of doing this :

- *Port based grouping*: Packet coming on a certain port may be forwarded to only a subset of all the ports.
- *Layer 2 address based grouping*: Looking at the layer 2 address of the packet, set of output ports is decided.
- *Layer 3 protocol based grouping*: Bridges can also segregate traffic based on the Protocol Type field of the

packet (2 bytes, between the Layer 2 and Layer 3 address fields).
- *Layer 3 subnet based grouping*: For some layer 3 protocols like IP, bridges may only forward traffic to other ports belonging to the same IP subnet. For this they have to look at layer 3 address of the packet.

In brief, VLAN switches modify the forwarding of bridged traffic. Devices referred as Layer 3 VLAN Switches, still operate at layer 2 but they use some layer 3 information.

ATM cell switching is an entirely new form of switching. Even though it is fundamentally different from a traditional LAN bridge, it is important to note that ATM switches fall in the category of Layer 2 products.

3.3 Layer 3 Switching

There is no consistent definition of "Layer 3 Switches" and they refer to wide variety of products. The only common thing between all of these devices is that they use layer 3 information to forward packets. Therefore, as discussed in the previous section, even Layer 2 VLAN switches with protocol/subnet awareness are sometimes referred as Layer 3 VLAN switches. Other products in this category are :

3.3.1 Layer 3 routing functionality in VLAN switches

There are several reasons for doing this. Pattern of network traffic is changing and the 80-20 rule, which says that 80% of all network traffic is intra LAN, is no longer valid. More traffic is crossing the LAN boundaries these days and to forward this kind of traffic, VLAN switches have to use layer 3 routing functionality. Traditionally, these VLAN switches forwarded such traffic to some route servers but as this type of traffic is increasing, it makes more sense to build this functionality within these switches. Many proprietary solutions are available for doing this.

3.3.2 Layer 2 ATM Switching with IP routing

Reason for doing this is that most of the service providers have invested heavily in ATM technology for their backbones. And now they need to map IP traffic on it. There are two very different approaches used in mapping layer 3 traffic to ATM circuits. The first approach aims at improving routing performance by separating the transmission of network control information from the normal data traffic. Control traffic passes through the routers and route servers to initiate call, whereas normal data traffic is switched through already established path. There are proprietary solutions for this like IP Switching, and there are standard techniques like MPOA(Multi Protocol over ATM) as well. The other approach addresses WAN route scalability issues. Routing decisions are performed once at the entry point to the WAN and the remaining forwarding decisions within the WAN are based on switching techniques. Tag Switching is one proprietary solution based on this approach and the IETF is working to develop a standard, MPLS.

3.3.3 Route Caching

Since, number of internet hosts is increasing at an exponential rate, it is not possible to have an entry for each of them in every routing table. Therefore, routers combine many of these entries which have a same next hop. But this worsens already complex task of route search. To improve route lookup time, many products keep a route cache of frequently seen addresses. When addresses are not found in the cache, then the search goes through

traditional software-based slow path. Many products in this category, combine layer 2 switching features with route cache based routing and vendors have named them as Layer 3 Switches, Multilayer Switches, Routing Switches and Switching Routers. Cache sizes range from 2000 to 64,000. Most of these products have a processor based slow-path for looking up routes for cache misses, but few of them take help of external router to perform these functions and they are sometimes referred as "Layer 3 Learning Bridges". Route Cache technique scale poorly with routing table size, and cannot be used for backbone routers that support large routing tables. Frequent topology changes and random traffic pattern also eliminate any benefits from the route cache, and performance is bounded by the speed of the slow path.

3.3.4 Full Routing

Some of the latest products in the market perform full routing at very high speeds. Instead of using a route cache, these products actually perform a complete routing table search for every packet. These products are often called Real Gigabit Routers, Gigabit Switching Routers etc. By eliminating the route cache, these products have a predictable performance for all traffic at all times even in most complex internetworks. Unlike other forms of layer 3 switches, these products improve all aspects of routing to gigabit speeds and not just a subset. These products are suited for deployment in large scale carrier backbones. Some of the techniques used in these products to improve route lookup are discussed later in the paper.

3.4 Switching Above Layer 3

Layerless Switching and Layer 3 Switching are the new buzzwords in the industry. Again there is no consistent definition of what these products do. Vendors are adding the ability to look at layer 4 header information and sometimes more into layer 3 products and marketing them as Layer 4 or Layerless switches. Products operating at layers 2 and 3 handle each packet the same way whether it is part of a long flow between two hosts or one travelling alone. But at layer 4 and higher, there is awareness of the flows and the higher-level applications to which this packet belongs. This information can be used to classify packets into different categories and depending on how the switch is architected, can be used to provide differentiated services and implement service level agreements in the network.

Back to Table of Contents

---

4. EFFICIENT ROUTING TABLE SEARCH

One of the major bottlenecks in backbone routers is the need to compute the longest prefix match for each incoming packet. Data links now operate at gigabits/sec or more and generate nearly 150,000 packets per second at each interface. New protocols, such as RSVP, require route selection based on Protocol Number, Source Address, Destination Port and source Port and therefore make it even more time consuming. The speed of a route lookup algorithm is determined by the number of memory accesses and the speed of the memory. This should be kept in mind while evaluating various route lookup techniques described below.

4.1 Tree-based Algorithms

Each path in the tree from root to leaf corresponds to an entry in the forwarding table and the longest prefix match is the longest path in the tree that matches the destination address of an incoming packet. In the worst case, it takes time proportional to the length of the destination address to find the longest prefix match. The main idea in tree based algorithms is that most nodes require storage for only a few children instead of all possible ones and therefore make frugal use of memory at cost of doing more memory lookups. But as the memory costs are dropping, these algorithms are not the best ones to use. In this category, Patricia-tree algorithm is one of the most common. [Craig99] [Kesh98]

4.2 Techniques to Improve Route Lookup

Various techniques have been proposed to improve route lookup time [Kesh98]. They can be broadly classified into :

Hardware-oriented techniques: Some of these techniques are as simple as using more memory as the costs are dropping and have a separate entry for each internet address. Longest prefix match is not required in this case and complexity of the search is reduced. Other techniques try to reduce the memory access time by combining logic and memory together in a single device.

Table compaction techniques: These techniques make use of the fact that forwarding entries are sparsely distributed in the space of all internet addresses. So they use some complicated compact data structure to store the forwarding table in the primary cache of a processor. This allows route lookup at gigabit speeds.

Hash based techniques: Hashing operates strictly on an exact-match basis and therefore longest prefix match limits the use of hashing for route lookup. The solution to this problem is to try different masks and choosing the one with the longest mask length. Choice of masks can be iterative or hierarchical but none of these solutions scale well with the size of the destination address.

4.3 Route Search at Gigabit Speeds

The solutions described above solve the route lookup problem in most cases. But with media speeds going up, it requires very careful implementation of one or more of the above techniques combined together to have possible advantages from all of them. With 1.5 million packets coming in, a router has only 672 nanoseconds to validate a packet, find an outgoing route and send the packet. Many vendors and research groups in universities have come up with innovative solutions for this. Details of most of the proprietary solutions from vendors have not been disclosed because of patent pending or similar reasons. Some of the well known solutions are mentioned below.

4.3.1 WASHU Algorithm [Craig99]: Developed at Washington University St. Louis., it is a scalable algorithm that uses binary hashing. The algorithm computes a separate hash table for each possible prefix length and therefore maintains 33 hash tables in total. Instead of starting from the longest possible prefix, a binary search on the prefix lengths is performed. Search starts at table 16 and if there is a hit, look for longer match, otherwise look for shorter match. But this scheme has a bug that if the longest match is a 17-bit prefix and there is no entry in table 16 that leads to looking at higher tables. Therefore markers are added, which track best matching shorter prefix. So now the algorithm works as follows. Hash first 16 bits and look in table 16. If find a marker, save best match from marker and look for longer prefix at table 24. If find a prefix, save prefix as best match and look for longer prefix at table 24. If miss, look for shorter prefix. Continue algorithm until tables exhausted.

4.3.2 Stanford University's algorithm [Nick98]also has a very good performance and all the details are available in a paper available on their site. A brief description of how it works is given here. This algorithm makes use of the fact that most of the prefixes in route tables of the backbone routers are shorter than 24 bits. The basic

scheme makes use of two tables, both stored in DRAM. The first table (TBL24) stores all possible route prefixes that are up to, and including, 24 bits long. Prefixes shorter than 24 bits are expanded and multiple 24 bit entries are kept for them. Second table (TBLLong) stores all route prefixes in the routing table that are longer than 24-bits. Each entry in TBLLong corresponds to one of the 256 possible longer prefixes that share the single 24-bit prefix in TBL24. The first 24 bits of the address are used as an index into the first table TBL24 and a single memory read is performed, yielding 2 bytes. If the first bit equals zero, then the remaining 15 bits describe the next hop. Otherwise, the remaining 15 bits are multiplied by 256, and the product is added to the last 8 bits of the original destination address, and this value is used as a direct index into TBLLong, which contains the next hop. Two memory accesses in different tables can be pipelined and the algorithm allows 20 million packets per second to be processed. Fig 5. shows how the two tables are accessed to find the next hop.

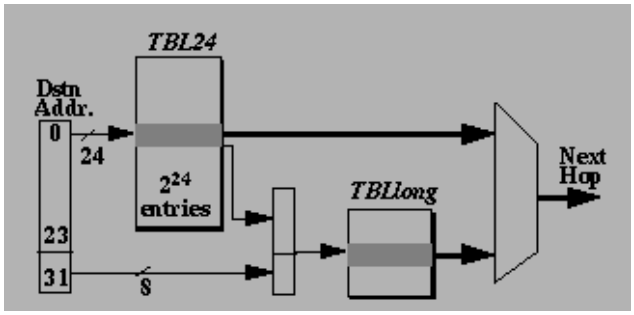

Fig 5. Stanford University's Algorithm for Efficient Route Lookup

4.3.3 Torrent Networking Technologies' ASIK algorithmis widely acclaimed for its performance. Details are not disclosed, but here is some information on its capabilities. It can be easily implemented in hardware and adapted to various link speeds. Memory space required grows with the number of networks rather than the number of hosts. It allows searches based on multiple fields of the IP header and also has a deterministic worst-case time to locate the best route(16 memory accesses).

Other vendors like Pluris and Nexabit also have their own solutions to route lookup problem which have very high performance. But nothing is mentioned about it in their papers. Route lookup is the single most important thing in the design of high speed routers and no vendor wants to share its ideas with anyone else unless the patent is approved.

Back to Table of Contents

---

5. ROUTER ARCHITECTURE FOR THE DIFFERENTIATED SERVICES

Providing any form of differentiated services require the network to keep some state information. The majority of the installed routers use architectures that will experience a degraded performance if they are configured to provide complicated QOS mechanisms. Therefore the traditional approach was that all the sophisticated techniques should be in end systems and network should be kept as simple as possible. But recent research and advances in hardware capabilities have made it possible to make networks more intelligent. [Kesh98] [Vijay98]

5.1 Components of Differentiated Services

Following operations need to be performed at high speed in a router to provide differentiated services :

Packet classification, which can distinguish packets and group them according to different requirements.

Buffer management, which determines how much buffer space should be allocated for different classes of network traffic and in case of congestion, which packets should be dropped.

Packet scheduling, which decides the order in which the packets are serviced to meet different delay and throughput guarantees.

5.2 No Queuing Before Header Processing

The first requirement for differentiated services is that the maximum delay for header processing must be no larger than the delay a packet from the service class with the least delay can experience. Without this constraint, violation of service assurances can be done even before header processing and that is not allowed. Therefore packet header processing must be done at wire speeds and not be traffic-dependent. The implication of this is on the design of forwarding engines. It is the worst-case performance of the forwarding engine which determine the packet processing rate, and not the average case performance. If average case performance is used to determine supported packet processing speeds, buffering will be required before processing.

5.3 Queuing

Once the packet header is processed and next-hop information is known, packet is queued before being transmitted on the output link. Switches can either be input or output queued. Output queued switches require the switch fabric to run at a speed greater than the sum of the speeds of the incoming links and the output queues themselves must run at a speed much faster than the input links. This is often difficult to implement with increasing link speeds. Therefore most of the switch designs are input queued but it suffers from the head-of-line blocking problem which means that a packet at the head of the input queue, while waiting for its turn to be transmitted to a busy output port, can block packets behind it which are destined for an idle output port. This problem is solved by maintaining per-output queues which is also known as virtual output queuing. A centralized scheduling algorithm then examines the contents of all the input queues, and finds a conflict-free match between inputs and outputs. But input queueuing poses another challenge for the scheduling. Most of the packet scheduling algorithms are specified in terms of output queues and this is a non-trivial problem to modify these algorithms based on input queuing.

5.4 Optimized Packet Processing

Increasing link capacities and the need for differentiated services stretch processor based architecture to the limit. Therefore multiprocessor architectures with several forwarding engines are designed. Another efficient solution is described here which is based on functional partitioning of packet processing as done below :

- Buffer and forward packets through some switching fabric.
- Apply filtering and packet classification.
- Determine the next hop of the packet.
- Queue the packet in an appropriate queue based on both the classification decisions and the route table lookup.
- Schedule packet transmission on outgoing links to meet QOS requirements

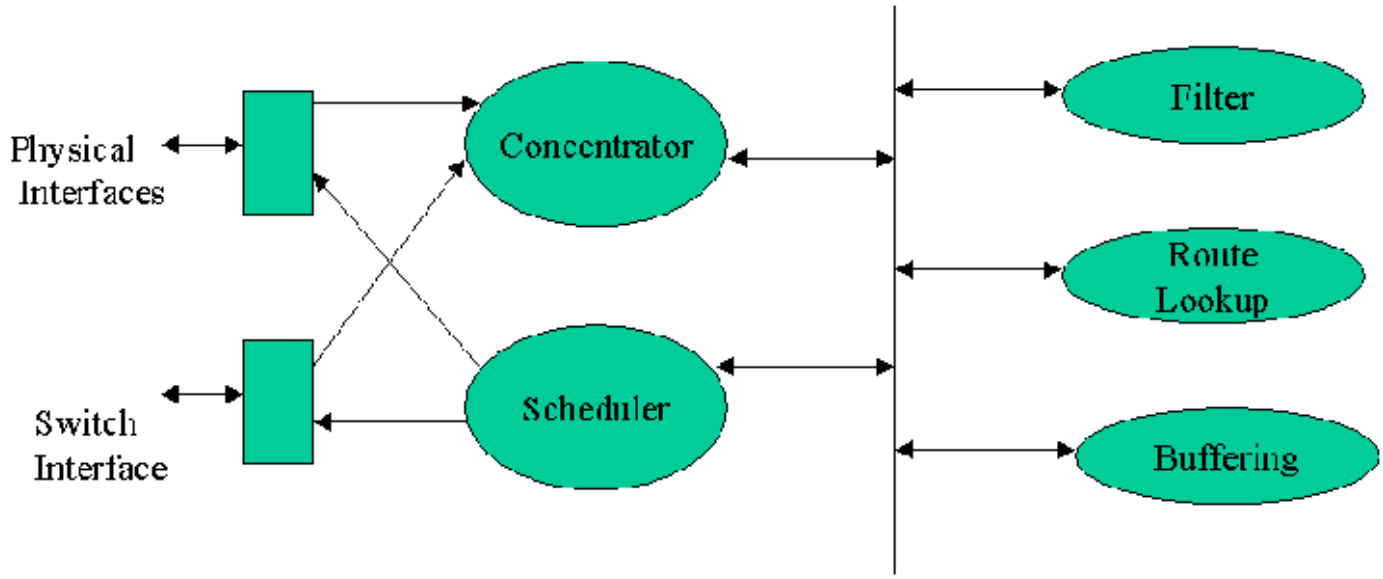

Fig 6. Router Architecture with Partioned Forwarding Engine

Processing elements optimized for each task are used in sequence and pipelining is done between these stages. Specializing the task each processor performs makes it possible to get a high locality of reference for memory accesses also which further improves the performance. Fig 6. shows the router architecture based on this approach. Further, combination of this design with the multiple shared processor architecture is also possible to provide very high packet forwarding rates.

Back to Table of Contents

---

6. SURVEY OF PRODUCTS

This section provides a survey of the terabit and gigabit capacity routers available in the market. Comparative analysis of all the major products classifies them into various categories based on architecture design as well as performance. Later, a slightly detailed description of two state-of-the-art products is also given.

6.1 Competitive Study of Leading Market Products

This competitive study identifies the key router vendors and maps each of them into the landscape of edge, mid-core, and core routing requirements. In addition, the study provides an overview of the critical line capacity and total switching capacity requirements for edge and core environments and compares the various architectural approaches being used to address these performance needs. Many of the data and ideas in this section are borrowed from a white paper at the site of Pluris corporation which has one of the leading products in this category. [PluComp] [NexProd]

6.1.1 Single Box Vs Multi-Chassis Architectures

Architectures from leading vendors can be divided into two broad categories based on how they scale to the increasing network demands :

Single Box Architectures: Traditional single-box designs have high-capacity switching fabrics but they use blocking LAN interconnects to link multiple boxes to increase the overall network switching capacity. Because of

the inherent limitations of using external line cards to handle the LAN interconnects, such single-box architectures cannot seamlessly grow their capacities to meet ever-higher traffic requirements. Currently, the leading router vendors offering high-end, single-box solutions include Lucent/Ascend, Cisco, Juniper, NEO Networks, Nexabit/Lucent, and Torrent/Ericsson. These routers tend to be most appropriate for edge to mid-core and core deployments with maximum line capacities between 25 Gbps and 160 Gbps.

Multi-Chassis Integrated Architectures: Distributed multi-chassis designs make use of an integrated, expandable switching fabric to provide non-blocking interconnection between multiple expansion chassis. By delivering seamless non-blocking connections between all elements of the system, these integrated architectures can provide smooth scaling to terabit levels and beyond. Most of the integrated multi-chassis solutions range from edge to core applications with maximum line capacities topping out at 160 Gbps for NetCore/Tellabs and Argon/Siemens , 1.4 Tbps for Avici and as high as 19.2 Tbps for Pluris. However, specific architectural and implementation choices can dramatically impact the overall scalability and deployment flexibility of such multi-chassis systems. The current multi-chassis architectures fall into the following categories:

* Star Architecture: Such architectures expand by using a central switch to aggregate multiple smaller leaf nodes. Examples for this architecture are NetCore/Tellabs and Argon/Siemens. The star architecture has limited scalability and reliability since it relies on a centralized bottleneck and a single point of failure.

* Matrix Architecture: Such architectures expand by building a scalable matrix of switching elements. Examples for this architecture are Avici, with a three-dimensional switching matrix that expands electrically using copper and Pluris, with a multi-dimensional switching matrix that expands optically via fiber-optic interconnects.

6.1.2 Line Capacity and Total Switching Capacity

To get into more detailed architectural comparisons, it is important to further define the differences between line capacity and total switching capacity and to know what are these values for various types of scalable gigabit and terabit systems available in the market.

*Line Capacity :*Line capacity refers to the effective input/output bandwidth that is available to a subscriber via the line-card ports. For example, a line card that has four OC-48 ports at 2.5 Gbps each would deliver 10 Gbps of line capacity. Invariably line capacity represents only a percentage of overall switching capacity. Gigabit routing devices typically can provide total line capacity of up to tens of gigabits per second, and are able to support multiple port interface speeds up to OC-48 (2.5 Gbps) or OC-192 (10 Gbps). Leading gigabit routing vendors include Cisco, Lucent/Ascend, Argon/Siemens, NetCore/Tellabs, Juniper, Nexabit/Lucent, and Torrent/Ericsson. Terabit routing devices are designed with the aggregate line capacity to handle thousands of gigabits per second and to provide ultra-scalable performance and high port density. These routers can support port interface speeds as high as OC-192 (10 Gbps) and beyond. Currently the leading terabit routing vendors include Pluris and Avici.

*Switching Capacity :*The switching capacity of a system consists of the total bandwidth for all line-card connections and internal switching connections throughout the system. The switching capacity should be substantially higher than the line capacity to ensure non-blocking switching between any two ports. Additional switching capacity is also needed to provide active redundancy and a higher level of faulttolerance. Therefore switching capacity includes: Bandwidth used for line card connections, Bandwidth available to modular expansion of line card connections, Bandwidth for non-blocking intra-chassis switching, Bandwidth for non-blocking inter-chassis switching and for modular multi-chassis expansion, Aggregate bandwidth needed to support redundancy and fault tolerance.

6.1.3 Comparative Product Positioning

Table 1. shows various single-box and multi-chassis architectures. For comparison, Table 1. compares only "single-chassis" versions of the multi-chassis systems to better illustrate relative throughputs for their basic configurations. Key factors to consider when comparing single-box and multi-chassis systems are the switch fabric capacity, line card capacity, number of cards supported, WAN interfaces supported (e.g., OC-3, OC-12,

OC-48,OC-192), and line card performance in packets per second.

	Capacity in Gbps		Number of line cards	Wan Interface Support	Number of OC-48 ports	Line Card Performance (million pps)
Product	**Switch Fabric**	**Line Card**				
Single Box Edge to Mid-Core Devices						
Cisco 12012	60	27	11	OC-3/12/48	8	1
Juniper M40	40	20	8	OC-3/12/48	8	2.5
Lucent PacketStar 6416	60	40	16	OC-3/12/48	16	NA
Torrent IP9000	20	20	16	OC-3/12	NA	NA
Single Box Mid-Core to Core Devices						
Nexabit NX64000	6,400	160	16	OC-3/12 /48/192	64	NA
Integrated Multi-Chassis Edge to Mid-Core Devices						
Argon GPN	40	20	8	OC-3/12/48	8	NA
NetCore Everest	20	10	4	OC-3/12/48	4	NA
Integrated Multi-Chassis Mid-Core to Core Devices						
Avici Systems TSR	640	100	10	OC-3/12 /48/192	10	7
Pluris TNR	1,440	150	15	OC-3/12 /48/192	60	33

Table 1. Single Chassis Configurations

Table 2. provides a relative comparison of fully expanded systems to show the maximum scalability of each type of architecture. It illustrates that single-box systems cannot expand beyond their previous capacities, whereas the multi-chassis architectures are able to deliver significantly more performance than in their single-chassis versions. Among the multi-chassis architectures, systems from vendors such as Argon/Siemens and NetCore/Tellabs provide switching capacities in the 320 Gbps to 1.2 Tbps range with line capacities of 160 to 640 Gbps, which can provide adequate performance to address mid-core routing environments. Systems from Avici and Pluris sit at the next performance level, delivering the terabit and greater switching capacities required for core routing requirements.

	Capacity in Gbps		Number of line cards	Wan Interface Support	Number of OC-48 ports	Line Card Performance (million pps)
Product	**Switch Fabric**	**Line Card**				
Single Box Edge to Mid-Core Devices						
Cisco 12012	60	27	11	OC-3/12/48	8	1
Juniper M40	40	20	8	OC-3/12/48	8	2.5
Lucent PacketStar 6416	60	40	16	OC-3/12/48	16	NA
Torrent IP9000	20	20	16	OC-3/12	NA	NA
Single Box Mid-Core to Core Devices						
Nexabit NX64000	6,400	160	16	OC-3/12 /48/192	64	NA
Integrated Multi-Chassis Edge to Mid-Core Devices						
Argon GPN	320	160	64	OC-3/12/48	64	NA
NetCore Everest	1,200	640	256	OC-3/12/48	256	NA
Integrated Multi-Chassis Mid-Core to Core Devices						
Avici Systems TSR	36,000	1,400	560	OC-3/12 /48/192	560	7
Pluris TNR	184,000	19,200	1,920	OC-3/12 /48/192	7,680	33

Table 2. Fully Expanded Configurations

6.2 Individual Case Studies of Some Leading Products

Study of high speed switches and routers can not be complete without understanding the key design features and capabilities of some leading products. Brief case studies of two products is given below. The Tiny Tera is a state-of-the-art and a result of acadamic research whereas Nexabit's NX64000 is a commercially acclaimed product in this area. For a more detailed description of them, reader is advised to visit the website of these products.

6.2.1 The Tiny Tera [Nick96]

Tiny Tera is a Stanford University research project, the goal of which is to design a small, 1 Tbps packet switch using normal CMOS technology. The system is suited for an ATM switch or Internet core router. It efficiently routes both unicast and multicast traffic. The current version has 32 ports each operating at a 10 Gbps (Sonet OC-192 rate) speed. The switch is a small stack composed of a pile of round shaped crossbar slices and a scheduler. See Fig 7. Each slice (6 cm diameter) contains a single 32*32 1-bit crossbar chip. A port is connected to the slices radially. The port design is scalable in data rate and packet size. The basic switching unit is 64 bits, called a chunk.

Unicast traffic use a buffering scheme called "Virtual Output Queuing" described earlier. When the 64-bit data chunks are transferred over the 32*32 switch, the scheduler uses a heuristic algorithm called iSLIP. It achieves fairness using independent round-robin arbiters at each input and output. This leads to maximum throughput of just 63%, but slight modifications give 100% throughput. If the iSLIP algorithm is implemented in hardware it can make decision in less than 40 nanoseconds. The switch also has special input queues for multicast. A multicast input can deliver simultaneously to many outputs. The switch uses fan-out splitting, which means that the crossbar may deliver packets to the output over a number of transfer slots. Developing good multicast scheduling algorithms was an important part of the Tiny Tera Project.

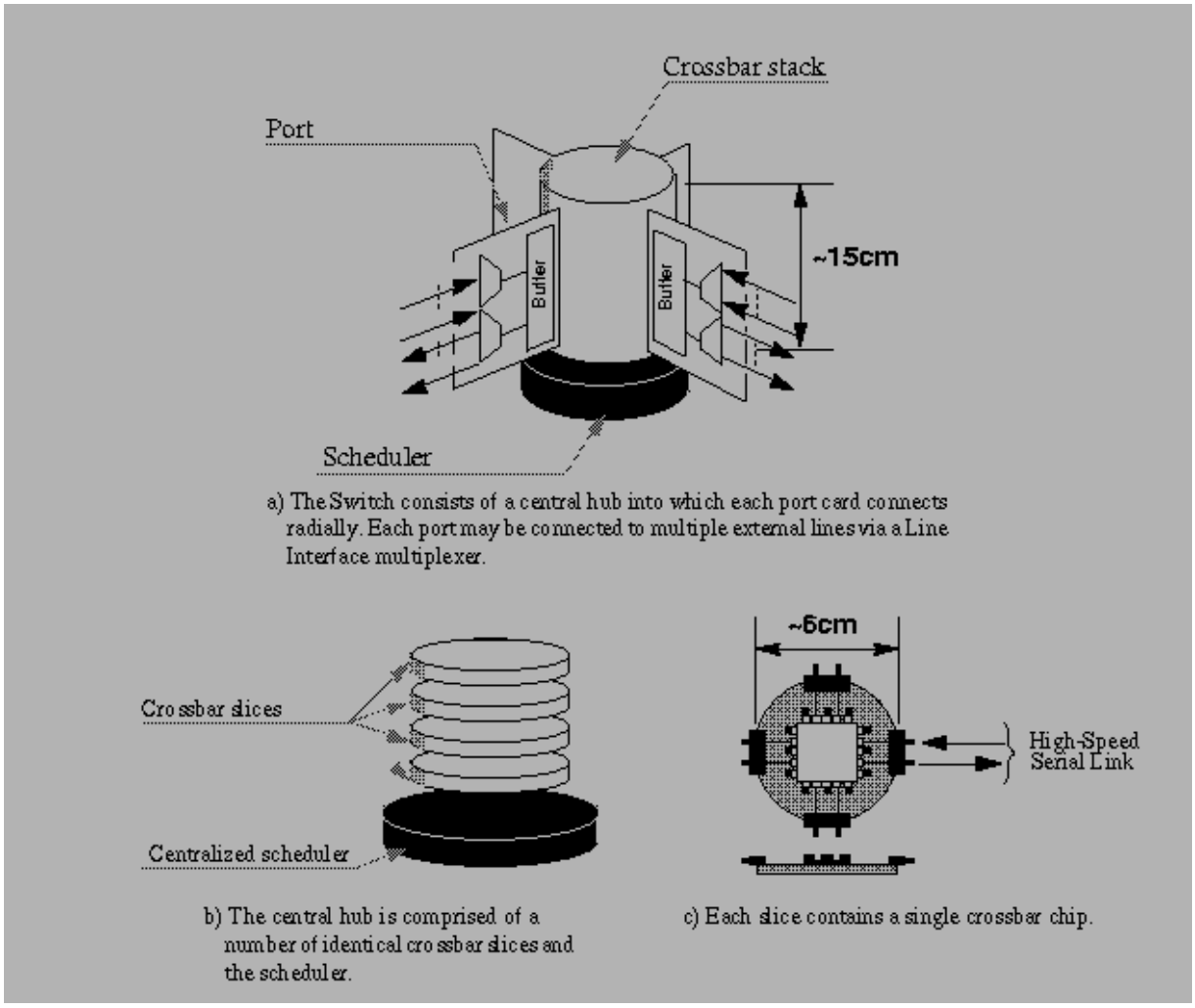

Fig 7. Architecture of Tiny Tera

6.2.2 Nexabit NX64000 [NexProd]

The NX64000?s innovative switch fabric delivers 6.4 Tbps switch capacity per chassis. The NX64000 supports up to 192 OC-3, 96 OC-12, 64 OC-48 and 16 OC-192 lines. The NX64000 allows Service Providers to even scale to higher speed like OC-768 and OC-3072 and port densities can also be increased further by interconnecting multiple chassis.

The NX64000 implements a distributed programmable hardware forwarding engine on each line-card. This approach facilitates wire-speed route lookup for full, 32-bit network prefixes-even at OC-192 rates. The distributed model enables support for 128 independent forwarding tables on each line-card. Each forwarding engine is capable of storing over one million entries.

One of the unique features of the NX64000 is its ability to support IP-IP tunnel encapsulation and de-capsulation at line-rates. The NX64000 is the only product in the market that can support a guaranteed delay of 40 microseconds to variable sized packets independent of packet size and type and therefore is capable of providing ATM-comaparable QOS in an IP world. It provides IP-CBR for which Nexabit also has a patent.

The NX64000 has been architected to support line-rate forwarding of any mix of unicast/multicast traffic. It provides direct connectivity to DWDM for optical internetworking and there multi-service platform supports ATM, IP, Frame Relay and MPLS.

Additionally, it delivers extensive performance-related statistics at the connection and packet level without compromising line-rate performance. These statistics facilitate network analysis and planning, and are also useful for accounting and billing purposes.

Back to Table of Contents

Summary

It is very clear now that with deployment of more and more fiber and improvements in DWDM technology, terabit capacity routers are required to convert the abundant raw bandwidth into useful bandwidth. These routers require fast switched backplanes and multiple forwarding engines to eliminate the bottlenecks provided in traditional routers. Ability to efficiently support differentiated services is another feature which will be used along with total switching capacity to evaluate these routers. Switching is faster than routing, and many products in the market combine some sort of switching with routing functionality to improve the performance and it is important to understand what the product actually does. But the products which scale up all aspects of routing rather than the subset of them, are bound to perform better with arbitrary traffic patterns. Route lookup is the major bottleneck in the performance of routers and many efficient solutions are being proposed to improve it. Supporting differentiated services at such high interface speeds poses some new challenges for the design of router architecture and some solutions are discussed here. Finally a survey of leading market products is presented.

Back to Table of Contents

References

[Decis97] Decisys, "Route Once, Switch Many, " July 1999, 23 pages, http://www.netreference.com /PublishedArchive/WhitePapers/WPIndex.html

[Decis96] Decisys, "The Evolution of Routing, " Sep 1996, 6 pages, http://www.netreference.com /PublishedArchive/WhitePapers/WPIndex.html

[NexNeed] Nexabit, "The New Network Infrastructure : The Need for Terabit Switch/Routers, " 1999, 11 pages, http://www.nexabit.com/need.html

[NexProd] Nexabit, "NX64000 Multi-Terabit Switch/Router Product Description," 1999, 18 pages, http://www.nexabit.com/proddescr.html

[NexSup] Nexabit, "Will The New Super Routers Have What it Takes, " 1999, 12 pages, http://www.nexabit.com /architecture.pdf

[PluComp] Pluris, "Competitive Study, ", April 1999, 10 pages, http://www.pluris.com/html/coretech /whitepaper4.htm

[PluPrac] Pluris, "Practical Implementation of Terabit Routing Scenarios", April 1999, 14 pages, http://www.pluris.com/html/coretech/whitepaper5.htm

[Klaus98] Klaus Lindberg, "Multi-gigabit Routers", May 1998, http://www.csc.fi/lindberg/tik/paper.html

[Craig99] Craig Partridge, "Designing and Building Gigabit and Terabit Internet Routers," Networld+Interop99, May1999

[Nick97] Nick McKeown, "A Fast Switched Backplane for a Gigabit Switched Router", Business Communications Review, Dec 1997, Vol. 27, No. 12, http://www.bcr.com/bcrmag/12/mckeown.htm

[Nick96] N. McKeown, M. Izzard, A. Mekkittikul, W. Ellersick, M. Horowitz, " The Tiny Tera : A Packet Switch core, " Hot Interconnects V, Aug 1996, http://tiny-tera.stanford.edu/~nickm/papers/HOTI_96.pdf

[Nick98] P. Gupta, S. Lin, N. McKeown, "Routing Lookups in Hardware at Memory Access Speeds, " IEEE Infocom, April 1998, http://tiny-tera.stanford.edu/~nickm/papers/Infocom98_lookup.pdf

[NickReal] N. McKeown, "High Performance Routing and Switching, " Stanford University Telecom Center Workshop on Routing and Switching, Sep 1997, http://tiny-tera.stanford.edu/~nickm/talks /Telecom_Center_Workshop_Sept1997.pdf

[CiscoOpt] Cisco, "Optical Internetworking : A Roadmap to the Data Network of the Future, " 10 pages, http://www.cisco.com/warp/public/cc/cisco/mkt/servprod/opt/tech/coint_wp.htm

[Avici99] Avici, "The World of Terabit Switch/Router Technology, " 1999, 3 pages, http://www.avici.com /white_papers/a_new_world_1.html

[Tel99] Doug Allen, "Terabit Routing : Simplifying the Core, " Telecommunications Online, May 1999, http://www.telecoms-mag.com/issues/199905/tcs/terabit.html

[Kesh98] S. Keshav, R. Sharma, "Issues and Trends in Router Design, ", IEEE Communications Magazine, May 1998, pp.144-151, http://www.cs.cornell.edu/skeshav/papers/routertrends.pdf

[Vijay98] V. Kumar, T. Lakshman, D. Stiliadis, "Beyond Best Effort : Router Architectures for the Differentiated Services of Tomorrow's Internet, " IEEE Communications Magazine, May1998, pp. 152-163, http://www.bell-labs.com/~stiliadi/router/router.html

Back to Table of Contents

List of Acronyms

DWDM : Dense Wave Division Multiplexing
SONET : Synchronous Optical Network
MAC : Media Access Control
MPLS : Multiple Protocol Label Switching

Back to Table of Contents

Last Modified: November 16,1999

Note: This paper is available on-line at http://www.cse.wustl.edu/~jain/cis788-99/terabit_routing/index.html