A New Router Architecture For Tomorrow's Internet

Dick Kachelmeyer, NetStar, Inc., Eden Prairie, Minnesota USA

Introduction

The Internet has come a long way since the 1970s when a collaboration of universities and government sites first linked their computers in a research experiment.

Today's Internet is a loosely-organized international collaboration of autonomous, interconnected networks. The Internet supports host-to-host communication through voluntary adherence to open protocols and procedures defined as Internet standards. Today these autonomous networks are interconnected by intelligent network nodes using the Internet standard TCP/IP protocol suite.

Early intelligent network nodes were merely small computers configured to act like packet switches. Over time specialization gave birth to a family of network devices each optimized to function at a different network layer. One of these devices came to be known as a router.

Without routers, there would be no Internet. User data transits the Internet inside IP packets. Based in information in a header, IP packets are relayed from router to router (hop by hop) until they arrive at their destination.

A few statistics says it all.

- 1400 Internet Access Providers today, expected to be consolidated to less than 100 by 1998 as the *big boys* get involved.
- Host growth:
  - •08/81 213
  - •12/87 28,174
  - •10/94 3,864,000
  - •01/95 4,852,000
  - •07/95 6,642,000
  - •01/96 9,472,000
- 94,000 networks (up from 40,000 Jan 95)
- 165 Countries
- 75,000 web sites (up from 17,000 Jan 95)
- 20,000 commercial sites (up from 1,700 Jan 95)
- Commercial sites are being added at 73 per day
- Millions of PCs sold *Internet/IP ready*

This phenomenal growth has been fueled primarily by the simple desire to be connected. No other force has been as powerful, and no single application has emerged as the demand driver.

There is just one problem. Traditional routers can't cope with the onslaught of traffic. Concomitant to explosive connectivity there has been an explosive demand for bandwidth and packet forwarding rates as billions of IP packets converge on the routers in the Internet over increasingly fast transmission technologies.

Today's Internet

Before we assess the extent of the problem that traditional routers pose when faced with the inexorable Internet demand, we must understand what the Internet looks like now and where it is likely to be in the near future.

Brief History

In April of 1995, the NSFnet (the main Internet core backbone linking routers with T3 and T1 lines) was decommissioned. The AUP (Acceptable Use Policy), which kept commercial traffic off the NSFnet backbone, no longer made sense. It was time to commercialize the Internet.

In just one year this move has facilitated a veritable gold rush, as large corporations, mostly representing owners of transmission infrastructure, position themselves as Internet service providers (ISPs). They have all made the same strategic determination: data traffic over their infrastructure investment will soon be greater than voice traffic. Their futures lie in data traffic.

The Internet is a complex evolving system governed by a mix of technology, economics, politics and culture. As is true with all complex systems, today's Internet topology is more a function of yesterday's topology than the result of a grand planned course of action. Today's Internet has evolved from yesterday's in response to demand and technological feasibility. It is important to keep in mind that the Internet will continue to evolve; probably in ways we can not foresee.

The main invariant throughout the history of the Internet has been its use of the IP protocol suite as a packet delivery method. IP (Internet Protocol) is the network layer protocol suite used to move packets over the underlying network, it provides a packet delivery service to the layers above and relies on services provided by the physical layers below.

We will review today's Internet topology starting from the inside, moving outward. *Figure 1* will help visualize the concepts under discussion.

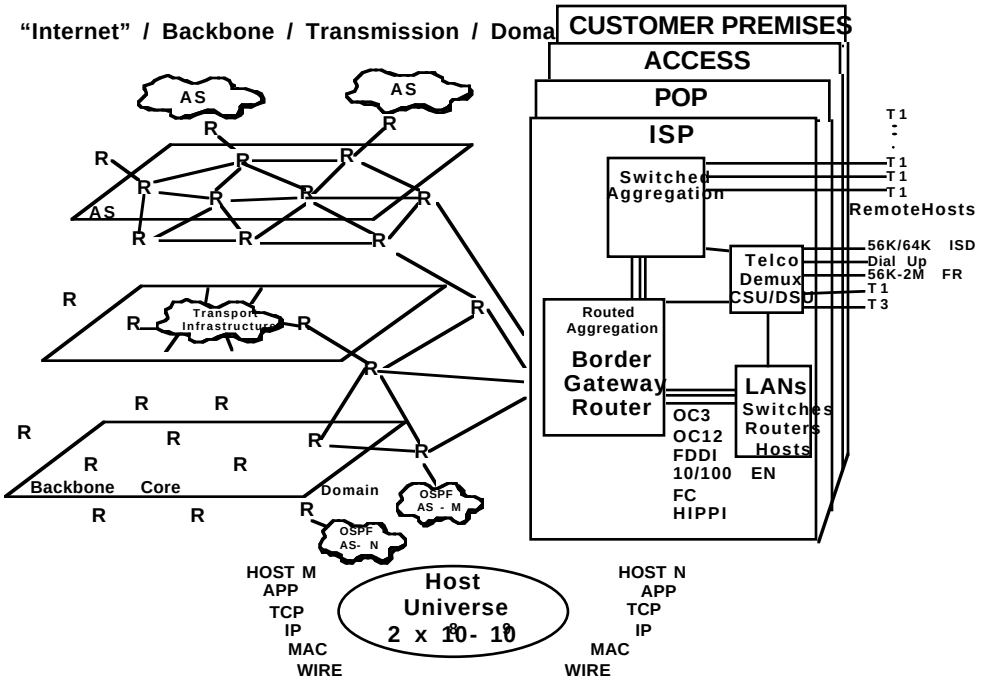

Figure 1

Autonomous Systems

The Internet's core is made up of many autonomous systems. An autonomous system (AS) is a group of IP networks using a single, clearly defined, routing algorithm. ASes primarily reside in the interior of the Internet, but they can be established anywhere. Several ASes are shown in the planes in figure 1. Routers are indicated by the letter "R". The top plane shows that the mesh connecting the routers can be complex. Routers on the border of each plane are usually referred to as "border routers". Border routers interconnect the planes (core router ASes) with smaller regional ASes.

The middle plane shows that each "line" connecting routers is actually a "transmission cloud" implemented over a telecommunications infrastructure built out of switches and fiber/cable in proprietary locations.

In today's "post NSFnet" Internet, core ASes are owned and operated by large organizations which control fiber and copper infrastructures. Smaller organizations own ASes on the core's periphery. Inside an AS domain, IP packets are routed using one or more interior routing protocols (IGP - interior gateway protocol). In most cases interior routing decisions are based on technical parameters like topology, hop "distance", link speed and load.

Within the core ASes, routes are determined on the basis of the network address and a metric such as the type of service a link can provide or hop-count "distance" to the target network. If a core router is connected to five other routers and has 50,000 route entries in its table, then these routes will be distributed across the five connections.

Border Router

Whereas routing information within an AS is exchanged using an interior protocol, routing information is exchanged between ASes using exterior routing protocols (EGP - exterior gateway protocol). Exterior routing decisions are frequently based on a mix of policy rules (on what basis will this IP packet be allowed to traverse *my* AS?) and technical parameters.

It is important to understand that routing information percolates through the Internet from router to router. Routers nearer the interior of the Internet use dynamic routing protocols to periodically modify the entries in their route table by assessing new routing information as it arrives from nearby routers.

NAP

On the far right of the diagram is a depiction of a typical Internet NAP (network access point). Terms such as "Internet Service Provider (ISP)" or "Point of Presence" (POP) or "Customer Premises" are regularly applied to this NAP location. Though characterized by robust variety, each access point typically contains a mix of the following items.

- An aggregation point for remote access to the Internet.

  •Dial-up. Either PPP (Point to Point Protocol) or SLIP (Serial Line Internet Protocol) is typically used. 14.4 Kb/s modem is most common. ISDN (Integrated Services Data Network) is becoming more common.

  •Branch office. Frame Relay over fractional T1 or T1 is common.

  •Depending on the type, these remote lines typically terminate at the NAP in a DSU/CSU (Digital Service

Unit/Channel Service Unit), or rotary dial-up unit, or an aggregation switch (such as Cascade's Frame Relay to LAN switch).

- One or more LANs based on a variety of physical infrastructures. Ethernet is most common. FDDI and HIPPI are often present.
- An aggregation router. This router is a "border" router, a gateway into the Internet. It will aggregate all Internet-bound IP traffic and receive all IP packets destined for locally or remotely connected hosts. This router must support both EGP and IGP routing protocols, support numerous media types and be able to forward an aggregate of millions of packets per second.

Routes, Routing and Forwarding

A router has only one job in life; forward packets as fast as it can. IP is a connectionless network transport service. Packets are either delivered to the final destination or forwarded to routers *nearer* the final destination. When a packet is received by router X, the destination address in the IP header is matched to an entry in a route table the router maintains. A *next hop* address is obtained from this table. This *next hop* address typically points to another router, router Y, connected to one of router X's physical media interfaces. The packet will be forwarded over this interface to router Y. This *next hop* address is the address of the router in the network that is *closer* to the final destination than the current router is. Routers use routing protocols to share information about topology and reachability with each other by exchanging relevant blocks of their route tables.

A *route* entry in a router's route table is not a list of sequential routers through which a packet must flow to find its destination. It is merely a pointer to one of the next hop nodes at the other end of one of the router's media interfaces. The interface pointer is not a complete route in itself, it was placed into the route table by a routing protocol that assessed possible paths to the destination. The determination of *nearer* was made by a routing algorithm from entries in routing tables sent to it by nearby routers. Routers share their tables with each other periodically (dynamic routing protocol) and each router uses this information to determine which router to forward packets for a given network destination.

It is important to distinguish between the two independent processes (just mentioned above) that all routers support.

- The forwarding of packets based on an IP address in the packet header and the associated next hop address in the current image of the route table.
- The ancillary maintenance of up-to-date route tables by periodically analyzing route tables from nearby routers, making modifications to the local table as needed.

Tomorrow's Internet

Though the exact nature of the Internet of two or three years hence is unclear, what is clear (take another look at the statistics listed at the beginning of this paper) is that several thresholds will soon be crossed which will severely challenge the ability of the traditional router to cope.

Table 1 illustrates the key issues facing routers.

Table 1.

Dimension	Today	1998 Estimate	Implication
Number of Networks	100,000	500,000	♦ Core-router's route table size increases to megaby lengths ♦ Route-table lookup must not limit forwarding rates ♦ Hardware assisted Patricia Tree traversal needed (Practical Algorithm To Retrieve Information Coded Alphanumeric) ♦ Time to add/delete routes increases dramatically ♦ Routing Algorithms must deal with megabyte table
Number of Hosts	10,000,000	50,000,000	♦ IP-address space ♦ CIDR (Classless Inter-Domain Routing) ♦ IPv6 128-bit IP address
Route-Table Updates	2 per second	50 per second	♦ Time to analyze routes on very large tables (hardw assisted) ♦ Time to add/delete routes increases dramatically ♦ Routing Algorithms must deal with large tables
Packet Forwarding Rates on High-speed Media (200 Byte Packets)	130,000 packets per second (PPS) SONET OC3c/STM1 xmit/rcv simultaneous	8,000,000 packets per second (PPS) SONET OC192/STM12 xmit/rcv simultaneous	♦ Near 100ns per packet each direction ♦ Route-table lookup must occur in 100ns ♦ 2ns processor CPU clock ♦ 50 clocks to forward
IPv6	Experimental Testbed	Preferred to IPv4	♦ New header ♦ 128-bit address
Integrated Services	Experimental	Common	♦ RSVP/CBQ/QoS ♦ More processing to make forwarding decision
IP Over SONET	Desired	Required	♦ Economics ♦ ATM SAR chip unlikely for OC48/192
IP Multicast	Experimental	Required	♦ New applications require IP Multicast ♦ Video conference ♦ Distance learning ♦ Entertainment

Increased Demand From Many Sources

The source of the increased demand is manifesting itself in numerous ways:

- An increased use of switching technology in the LAN.
- The migration to higher speed (near gigabit) media in the LAN.
- The migration to higher speed (near gigabit) media on the WAN, i.e., over the Internet backbones.
- A tidal wave of relatively small IP packets ("averaging" around 200 bytes each) converge on the boarder and backbone routers as packets are aggregated from hundreds of thousands of hosts and fed into the interior of the network.
- The explosion in the number of hosts attached to the Internet has caused the route tables in the interior routers to grow beyond anything ever envisioned. Instead of route tables with a few hundred routes, ISPs are now talking about route tables with hundreds of thousands of routes.
- The stochastic nature of millions of hosts sending packets to each other has increased the rate at which route tables must be modified; add new routes, delete old routes. Instead of a route table update occurring once every few seconds, ISPs are saying route tables will require updating 20 to 50 times a second.
- The number of individual network is 94,000 today and climbing fast.
- IP-ready PCs. It has become very easy for the PC user to make use of the Internet. Windows 95 for example comes with a ready-to-use IP stack.

Traditional Routers Can't Cope

As indicated above, a router has only one job in life; forward packets as fast as it can. But traditional shared-bus routers, faced with the issues indicated in the table above, simply can't cope.

The flood of packets is overwhelming the current routers. Router brownouts (host unreachable message) are frequent.

Meltdowns (backbones lost for periods of hours) are being predicted. Stopgap measures are only expected to stem the tide for a period of months. What is needed is a leap of at least 10x or more in router performance. Incremental optimizations to traditional routers have been made over time as new protocols and modified hardware were introduced, but the overall architecture has remained the same and has run out of gas.

Table 2 shows how much time a router has to forward a 200 byte IP packet if it tries to keep up with the media speeds being placed in the interior of the Internet. When coupled with the predicted growth in the route table size (500,000 route entries in a year or two) it becomes readily apparent that today's routers will not scale to meet those needs.

A New Router Architecture Is Needed

Searching a large table and "getting rid" of a packet in 1µs will require a new router architecture.

An Internet router which will meet the demands discussed above must be able to:

- Interconnect multiple media types including new high-speed media types.
- Provide media access on cards with high port density and a high card slot-count in a low profile chassis.
- Forward small IP packets (200 bytes or so) at the full line rates of new high-speed media.

Table 2

Media	Line Speed (Mb/s)	Forwarding 200 Byte packets line speed (KPPS) *	Single packet forwarding time µ(s)
HSSI / DS3	52	32.5	30
FDDI	100	62.5	16
Ethernet	100	62.5	16
OC-3c (STS-3c) / SDH-1 (STM-1)	155	96.875	10
OC-12c (STS12c) / SDH-4 (STM-4)	622	388.75	2.6
HIPPI	800	500	2
Fibre Channel	1000	625	1.6
Ethernet	1000	625	1.6
OC-48c (STS-48c) / SDH-16 (STM-16)	2488	1,555	.64
OC-192 (STS-192) / SDH-64 (STM-64)	9953	6,221	.16

* Note, different media present different levels of non-data overheads. The actual packet forwarding rates will be somewhat less than shown above.

- Use full route table lookup schemes on a very large route tables (300,000 or more route entries).
- Support dynamic routing protocols in an increasingly dynamic environment where routing topologies change as often as 50 times per second.

Such a router must be able to determine the disposition of an IP packet in a matter of microseconds (a 200 byte packet at 1 gigabit per second is 1.6 µs) over multiple high-speed media.

A Simple Packet Forwarding Engine

A device that forwards packets (router) can be segmented into three architectural components:

1. A collection of various network media interfaces which connect independent networks (or hosts) to the device.
2. A packet forwarding engine that moves packets between interfaces based on the content of the packet headers.
3. A connection fabric that interconnects media attached interfaces, and forwarding engines.

Figure 2 below graphically illustrates the main differences between the traditional router architecture and a new switch-based router architecture implemented by NetStar in its GigaRouter. *Table 3* summarizes the differences and lists the benefits of the new architecture over the traditional architecture.

A Closer Look

An example of a router based on the new architecture is shown in *Figure 3*. Packet forwarding under this architecture is described below.

It is important to note that each media card in this architecture acts as an independent IP router, receiving and sending packets between the media and the switch. Each card has its own instantiation of the IP forwarding engine, its own dedicated one Gbps connection to the switch, and its own complete route table.

Forwarding Process In A Switch-based Router

NetStar's switch-based routing architecture forwards packets in the following manner:

1. IP packets arrive at the input side of one of the ports of a media card.
2. The frame header is stripped off and the remaining IP packet is added to a 4 Mbytes input buffer (a separate 4 Mbytes output buffer also exists for each card).

    2.1. The primary function of this large buffer is to decouple media cards across the switch and to help "speed match" media cards of differing speeds.

        2.1.1. Aside from the obvious speed matching between, say FDDI to HIPPI, this architecture will readily match the slight differences between OC-3c SONET and STM-1 SDH.

3. The destination IP address is used in a full route table lookup to locate the output media card. The buffer is multiported. Packets can be simultaneously entering and exiting the buffer. Packets remain in the buffer only during the IP lookup process.
4. The packet is sent across the switch on its own dedicated 1 Gbps connection to the selected output media card.
5. The packet is added to the outbound 4 Mbytes buffer on the outbound media card. Outbound and inbound paths coexist on the same media board, but are entirely independent and can operate entirely in parallel.
6. In most cases the media address associated with the outbound IP address will be in the ARP cache. If not, the outbound card sends out an ARP request. RFC1577 is used for ATM, RFC 1374 for HIPPI.

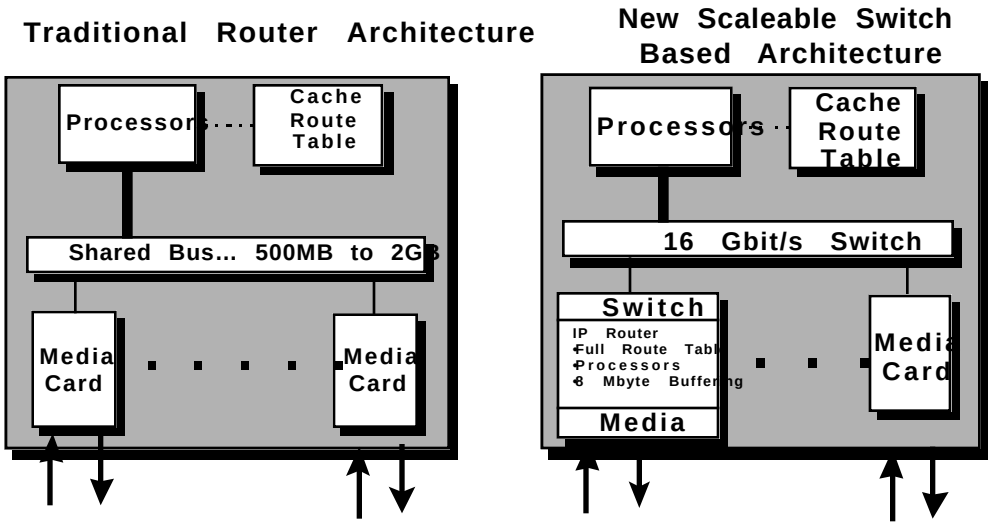

Figure 2

6.1. If fragmentation is required to fit into the PDU of outbound media, the outbound card fragments. Fragmentation is a network issue, MTU discovery is highly recommended by the Internet community.

7. The framing hardware frames the IP packet in the appropriate media PDU frame and sends it out on the "wire".

A special note for ATM. The GigaRouter is not an ATM switch, it is an IP router. The GigaRouter's ATM media card supports UNI 3.0/3.1 and SVCs. In an ATM environment, this card becomes a VC termination point over UNI (or an initiation point). In the receive mode, the media card accepts all cells, the objective is to retrieve the IP packet from the AAL5 cell flow. Cells are not forwarded, IP packets are. Traffic shaping is supported, but is used only to control the burst, peak and average cell rates on cells being injected into a VC by the GigaRouter's ATM card(s).

Additional Modes Supported

The NetStar GigaRouter supports two additional modes for HIPPI media.

HIPPI Switch Emulation

If the HIPPI PDU does not contain an IP address but instead a HIPPI I-field address, then the receiving HIPPI media card will forward the HIPPI PDU to a different HIPPI media card, thus emulating a HIPPI switch. The GigaRouter HIPPI media card will forward IP or emulate a HIPPI switch on a per-frame basis.

HIPPI Tunneling Over ATM

In an altogether different mode, two GigaRouters on either end of an ATM PVC over an OC3c/STM1 cloud can set up a HIPPI connection, thus tunneling HIPPI PDUs through an ATM cloud. This in affect extends HIPPI beyond the 1 to 10 Km serial HIPPI range.

ATM And The IP-World

ATM presents a unique set of challenges to the IP-based Internet community. Though this paper is not intended to be an ATM tutorial, a few IP related items are briefly addressed below. If more information is needed, the Internet itself can be queried.

ATM

ATM is a connection-based non-broadcast method for delivery of fixed-length cells through a public or private mesh interconnected by cell switches. This technology is very different from shared-media based broadcast oriented networks. ATM is only able to co-exist with other networking technologies through the use of several mappings or emulations implemented in services at several different OSI model layers. The effective use of ATM technologies in the IP-based Internet is still evolving.

Today, IP datagrams can be forwarded over ATM links by routers supporting UNI 3.0/3.1 (user network interface). It is currently easier for routers to use ATM links than it is for ATM switches to perform a routing function.

Table 3

Architecture Component	Traditional	New	Internet Benefit
Connection fabric bandwidth	♦ Limited bandwidth ♦ 500 Mbps up to 2 Gbps ♦ Shared by all media	♦ 16-Gbps non-blocking switch ♦ Not shared ♦ 1 Gbps dedicated bandwidth per media card	♦ Able to sustain multi-million packet per second forwarding rates ♦ Bandwidth scaleable ♦ New dedicated bandwidth added when add media ♦ Supports full complement of high speed media
Connection fabric latency	♦ Shared bus presents high/variable latency ♦ Sustained packet forwarding is far short of theoretical	♦ Low switch latency	♦ Able to sustain multi-million aggregate packet per second forwarding rates
Route table lookup	♦ Cached ♦ Cache misses and flushing impact performance ♦ Limits size of route table	♦ Full route table lookup ♦ Support very large tables	♦ Accommodate expected growth in both route tabl and route table updates
Slot count	♦ Low ♦ Limited by shared bus bandwidth	♦ High	♦ Able to connect multiple high-speed media types
Port count	♦ Low ♦ Limited by shared bus bandwidth	♦ High	♦ Able to aggregate multiple slower media on single ♦ Cost per port is low
Packet forwarding engine	♦ Single dedicated	♦ Multiple instantiations ♦ One for each media card	♦ Can tune design parameters to maximize forwardir varying media types ♦ Maintain forwarding at full line rates with small pac regardless of media speeds

Adaptation Layer - "Framing IP In ATM"

The function of mapping user Protocol Data Units (PDUs) into the information field of the ATM cell and vice versa is performed in the ATM Adaptation Layer (AAL). When a VC (Virtual Channel) is created, a specific AAL type is associated with the VC. For PVCs (Permanent Virtual Circuit) the AAL type is administratively configured at the end points when the Connection (circuit) is set up. For SVCs (Switched Virtual Circuits), the AAL type is communicated along the VC path via Q.93B as part of call setup establishment and the end points use the signaled information for configuration. ATM switches generally do not care about the AAL type of VCs.

The Internet (IP routers) make use of the AAL5 format for moving IP packets over an ATM infrastructure. A very simplistic way of viewing this is to think of AAL5 as a framing protocol used to carry IP packets over an ATM link. AAL5 specifies a packet format with a maximum size of (64K - 1) octets of user data. Cells for an AAL5 PDU are transmitted by the router's ATM interface first to last, the last cell indicating the end of the PDU. ATM standards guarantee that on a given VC, cell ordering is preserved end-to-end.

Since the use of ATM endpoint addresses and E.164 public UNI (User Network Interface) addresses by ATMARP are analogous to the use of Ethernet addresses, the notion of "hardware address" is extended to encompass ATM addresses in the context of ATMARP, even though ATM addresses need not have hardware significance.

RFC1577 describes the initial deployment of ATM within "classical" IP networks as a direct replacement for local area networks (Ethernet) and for IP links which interconnect routers, either within or between administrative domains. The "classical" model here refers to the treatment of the ATM host adapter as a networking interface to the IP protocol stack operating in a LAN-based paradigm.

IP Routing In An ATM-Based LIS (Logical IP Subnets)

Each VC directly connects two IP members within the same LIS. In the LIS scenario, each separate administrative entity configures its hosts and routers within a closed logical IP subnetwork. Each LIS operates and communicates independently of other LISes on the same ATM network. Hosts connected to ATM communicate directly to other hosts within the same LIS.

Communication to hosts outside of the local LIS is provided via an IP router. This router is an ATM endpoint attached to the ATM network that is configured as a member of one or more LISes. This configuration may result in a number of disjoint LISes operating over the same ATM network. Hosts of differing IP subnets MUST communicate via an intermediate IP router even though it may be possible to open a direct VC between the two IP members over the ATM network. The requirements for IP members (hosts, routers) operating in an ATM LIS configuration are:

- All members have the same IP network/subnet number and address mask.

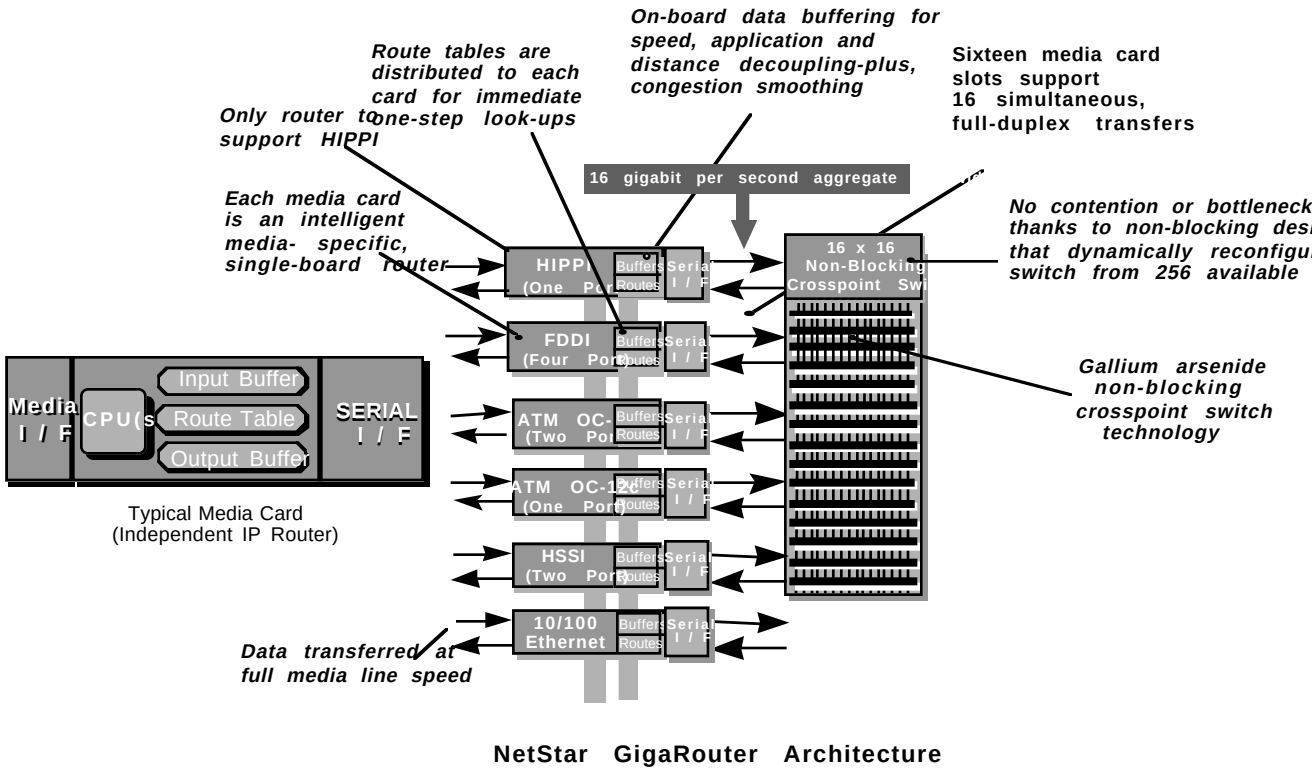

Figure 3

- All members within a LIS are directly connected to the ATM network.
- All members outside of the LIS are accessed via a router.
- All members of a LIS MUST have a mechanism for resolving IP addresses to ATM addresses via ATMARP.
- All members of a LIS MUST have a mechanism for resolving VCs to IP addresses via ATMARP.
- All members within a LIS MUST be able to communicate via ATM with all other members in the same LIS.
- The default MTU size for IP members operating over the ATM network SHALL be 9180 octets. The LLC/SNAP header is 8 octets, therefore the default ATM AAL5 protocol data unit size is 9188 octets.
- In classical IP subnets, values other than the default can be used if and only if all members in the LIS have been configured to use the non-default value.

Summary

The Internet has been successful beyond anyone's wildest dreams. But success has stressed the current infrastructure to the breaking point. Even conservative predictions will require a new router architecture to cope. A switch-based router, such as the NetStar GigaRouter, meets today's needs. Its architecture is scaleable and will also meet the demands of two years out. Beyond that, it is evident that routers must make even more creative use of high-speed switches, parallel functionality, (e.g., separate header processing from data movement) and design hardware assisting engines as integral parts of new software algorithms.

NetStar has already taken the first steps with its 16 Gbps switch-based GigaRouter. We look forward to meeting the challenges ahead as we develop the next generation router.