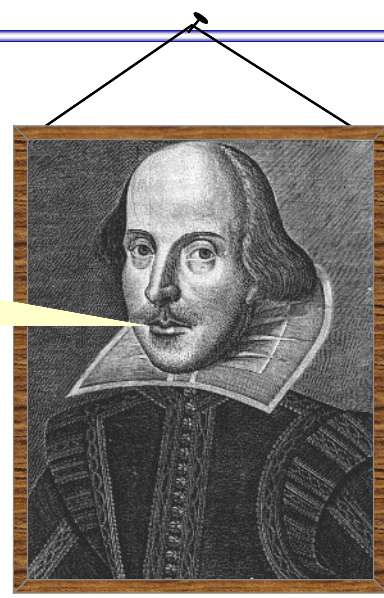
IP Switching and Gigabit Routers

Peter Newman

August 1997

IPSILON
the IP switch company



IPSILON

Contents

- *Why high-speed routing?*
- *Getting up to a gigabit*
- *Gigabit routers*
- *IP switching*
- *A rose by any other name…*
- *Conclusion*

IPSILON

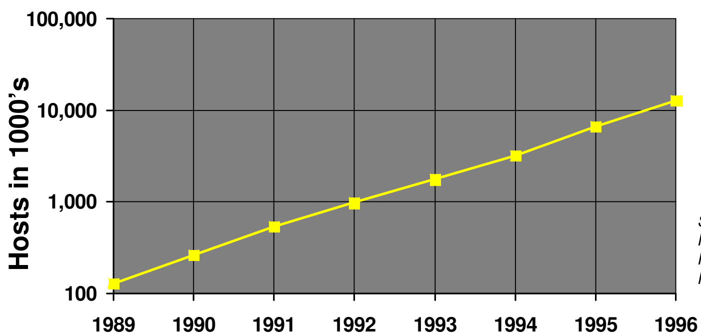
Man shall not live by plain text alone ...

IPSILON

... but by all the delights of multimedia communication

IPSILON

Growth of the Internet



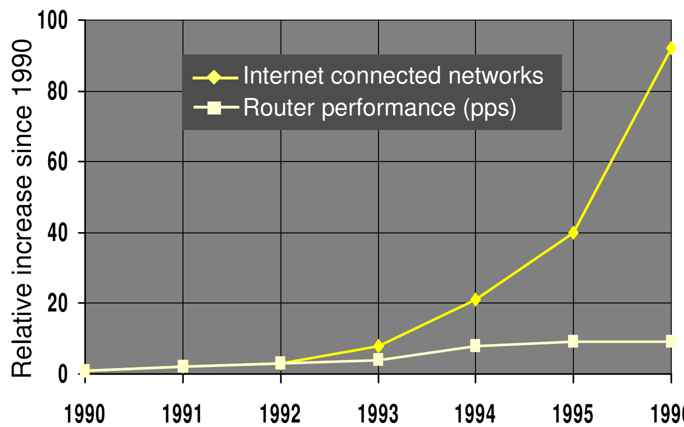
Source: M.Lottor, Network Wizards; Matthew Gray, MIT

- ***Number of hosts has doubled every 55 weeks since 1989***
- ***Number of web sites has doubled every 23 weeks since 1993***
- ***More than 60% of Fortune 1000 companies have intranets***

IPSILON

Routers Are Not Keeping Up

Growth in number of networks connected to Internet vs. growth in router performance

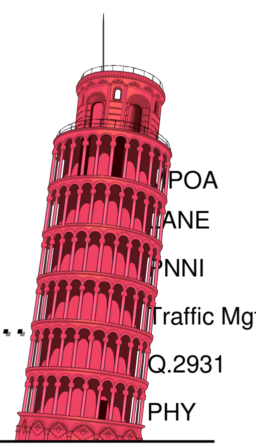


Source:
www.cisco.com
Network Wizards

IPSILON

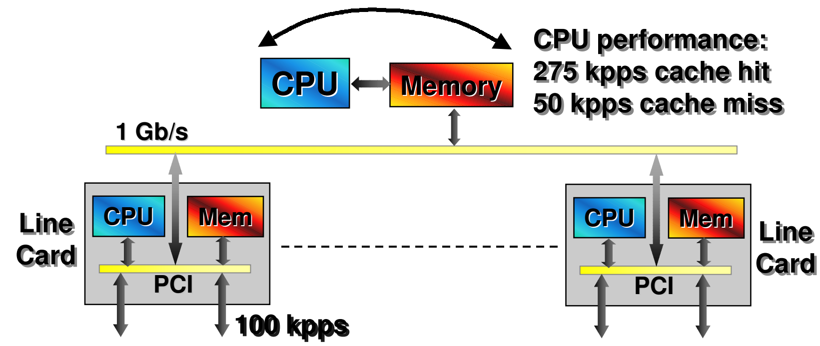
Contents

- *Why high-speed routing?*
- ➜ *Getting up to a gigabit*
- *Gigabit routers*
- *IP switching*
- *A rose by any other name…*
- *Conclusion*



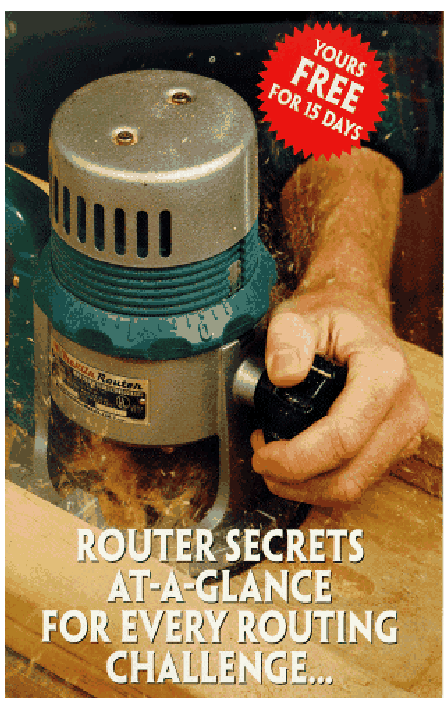
IPSILON

Current Generation Router Architecture



- *Distributed: 11 slots × 100 kpps = 1.1 Mpps*
- *Centralized = 275 k pps*
- *Multicast traffic = 50 k pps*

IPSILON



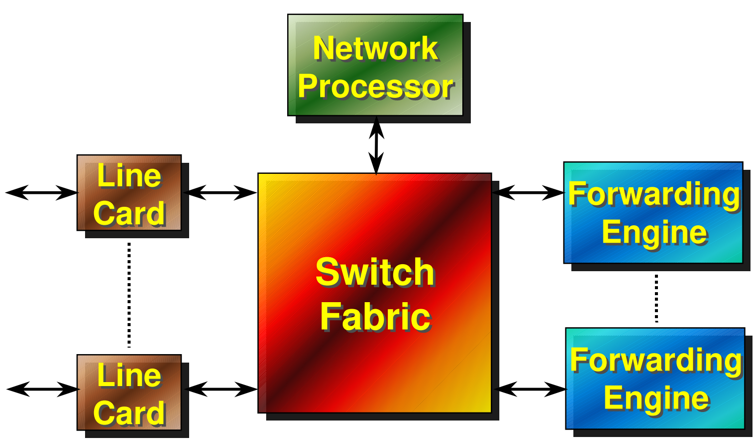
IPSILON

Getting Up to a Gigabit

- *Average packet size on Internet is 2000 bits*
- *Therefore must forward 500 kpps per Gbps of traffic*
- *Replace shared bus by switch fabric*
- *Separate routing and forwarding functions and run in parallel*
- *Network processor performs routing function and scales with number of peers, not with bandwidth*
- *Increase forwarding performance with multiple parallel forwarding engines*

IPSILON

General Structure of a Gigabit Router



IPSILON

Design of the Forwarding Engine

- *Problem: Implementation of longest-prefix match at high-speed*
- *Current approaches:*
  - *Hardware tree-walking*
  - *Special purpose processor with high-speed I/O*
  - *Processor with lookup hardware assist*
  - *Processor with cache*
  - *Content addressable memory (cache)*
  - *Much current research into fast algorithms*
- *Issue of caching routes*
  - *Is there sufficient locality in the traffic for route caching to be useful?*

IPSILON

Design of the Forwarding Engine

- *Full table lookup*
  - *All routes in forwarding table*
  - *Longest prefix match on entire forwarding table for every packet*
- *Route (prefix) cache*
  - *Cache most recently or frequently used routes*
  - *Longest prefix match on cache first, else full table lookup*
  - *Problem of overlapping prefixes complicates cache management*
- *Destination cache*
  - *Cache most recently or frequently used 32-bit destination addresses*
  - *32-bit destination address match on cache first, else full table lookup*
  - *Requires a much larger cache than a route cache*
- *"Layer 3 bridging"*
  - *Does not participate in routing protocols*
  - *Destination cache filled by "learning" from observed traffic*

IPSILON

Hardware Tree-Walking

- *Implementation in ASIC + SRAM*
- *Longest prefix match — protocol independent*
- *Can also handle multicast*
- *Requires 16 bytes per route*
- *For table of 250,000 routes require 4 Mbytes SRAM*
- *Need* 1+log N *memory accesses per route lookup*
- *10 ns SRAM gives one full route lookup in 200 ns*
- *This gives a forwarding rate of 5 Million pkts per sec*
- *Enough on average to serve 10 Gbps of traffic*
- *Worst-case 1.6 Gbps of 40 byte pkts at wire speed*

IPSILON

Problems with Hardware Forwarding

- *Next-hop forwarding no longer sufficient, also require:*
  - *Multicast forwarding*
  - *Quality of service based differentiation (QoS)*
  - *Forwarding policy (packet filtering)*
- *Non-contiguous match required for QoS and policy*
- *Patricia tree cannot perform non-contiguous match*
- *Cecilia tree handles non-contiguous match:*
  - *Can yield multiple matches*
  - *Insertion/deletion can be difficult*
  - *"Never seen one that worked!"*

"We can reduce today's functionality to hardware But will it support tomorrow's requirements?"

IPSILON

More on Hardware Forwarding

- Some policy requires relational operators
  - Difficult to implement in a tree structure (e.g. reject TCP/UDP ports > 1023)
- Combining forwarding and policy may be possible but it is not trivial
- RSVP: Data-driven per flow state
- Logging and accounting functions in hardware?
- How flexible can you make hardware forwarding?

"Hardware is hard not because it is hard to make but because it is hard to change"

IPSILON

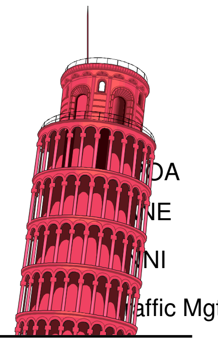
Processor Based Forwarding

- Processor offers flexibility:
  - Permits IP spec to change, e.g:
    - Firewall rules change
    - Changes introduced for RSVP
  - Allows experimentation with new algorithms
- Processor alone too slow — requires hardware assist or cache
- Effectiveness of cache depends on traffic locality
- Processor anywhere in the data path slow things down
- Processor growth curve:
  - Processors double in performance every 1½ years

IPSILON

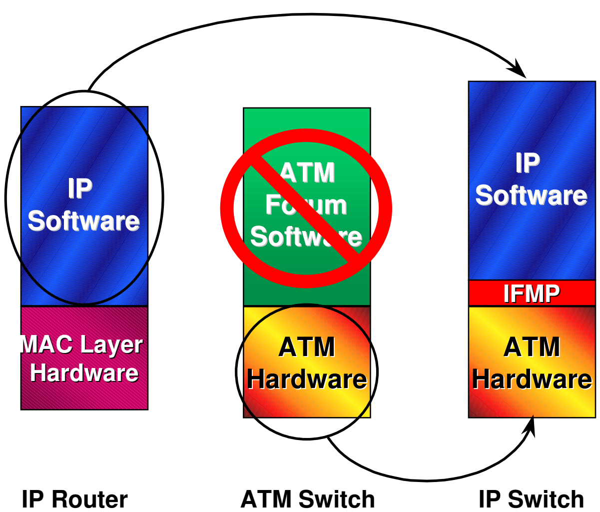
Contents

- Why high-speed routing?
- Getting up to a gigabit
- Gigabit routers
- → IP switching
- A rose by any other name…
- Conclusion



IPSILON

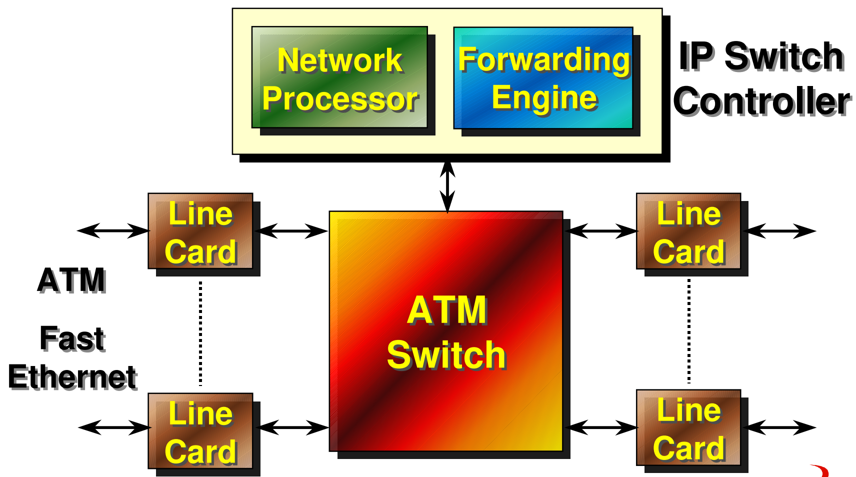
IP Switching: The Marriage of ATM Hardware to IP Control



IFMP:
Ipsilon Flow Management Protocol
A protocol to attach a layer 2 label to a layer 3 flow enabling the flow to be switched at layer 2

IPSILON

General Structure of an IP Switch

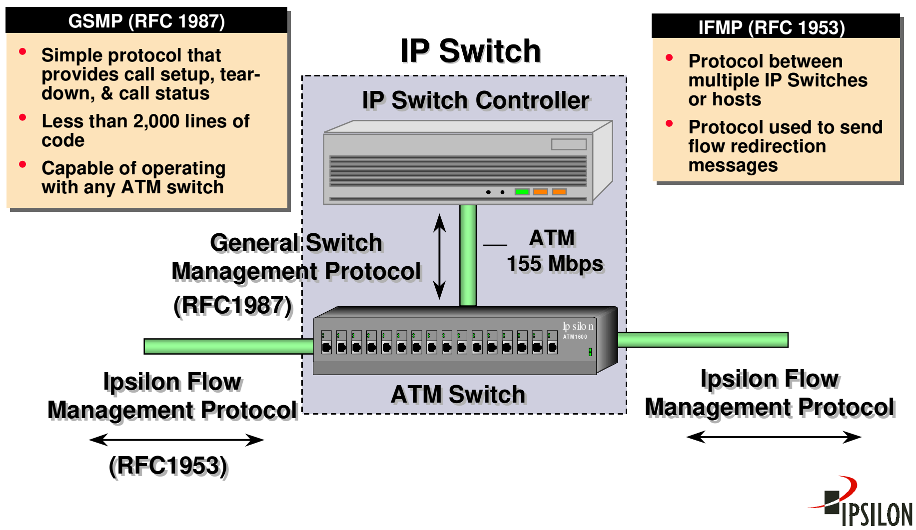


Operation of an IP Switch

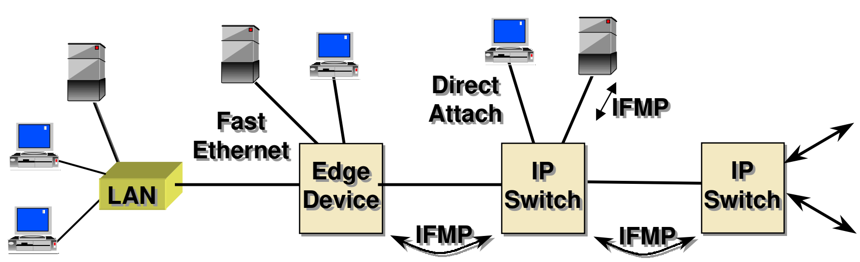
- *Default: Layer 3 forwarding in IP switch controller*
- *Flow classification for each new flow:*
  - *Quality of service differentiation*
  - *Packet filtering for forwarding policy*
  - *Select long duration flows*
- *Labelling: Assign layer 2 label (VCI) to selected flows*
- *Switching: switch labelled flows directly at layer 2*

"IP switching exposes access to the switching fabric"

IP Switch Implementation



IP Switching Network

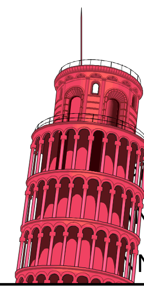


- *Propagate label toward edge of network*
- *IFMP: forwarding cache distribution protocol*
- *Packets labelled at edge of network*
  - *Packet arrival rate is lower than in center of network*
  - *Layer 3 processing at edge of network*
  - *Layer 2 switching within network*

IPSILON

Contents

- Why high-speed routing?
- Getting up to a gigabit
- Gigabit routers
- IP switching
- → A rose by any other name…
- Conclusion



DA
NE
NI

IPSILON

A Rose By Any Other Name Would Smell

Slightly different mechanisms to bind layer 2 labels to layer 3 flows

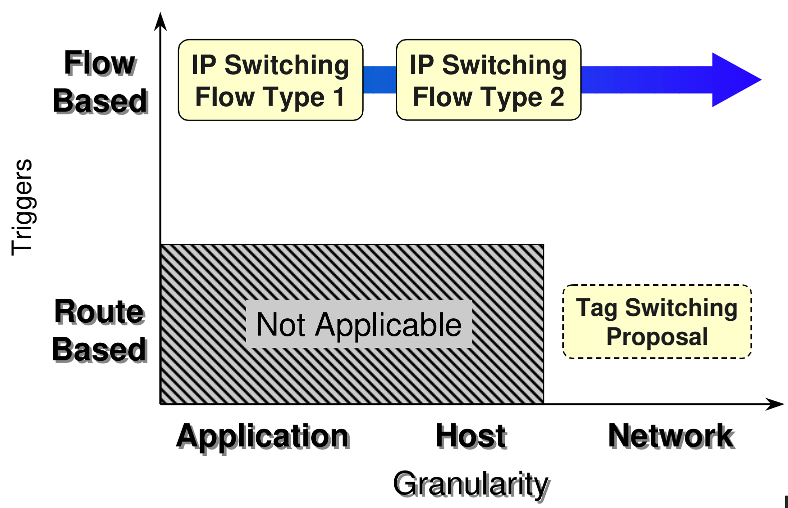
- *IP/ATM — Guru Parulkar, Washington University*
  - *Pool of pre-established PVCs*
  - *New incoming flow takes a PVC*
  - *Packets not forwarded until PVC activated*
- *Cell Switch Router — Toshiba*
  - *Permits entire classical IP over ATM subnets between CSRs*
  - *Requires ATM signalling to establish VC across subnet.*
  - *Proposes use of RSVP to propagate mapping of flow and VCI*
- *Multiprotocol Label Switching (MPLS)*
  - *Specification for operation over ATM planned for Dec 1997*
- MPOA

IPSILON

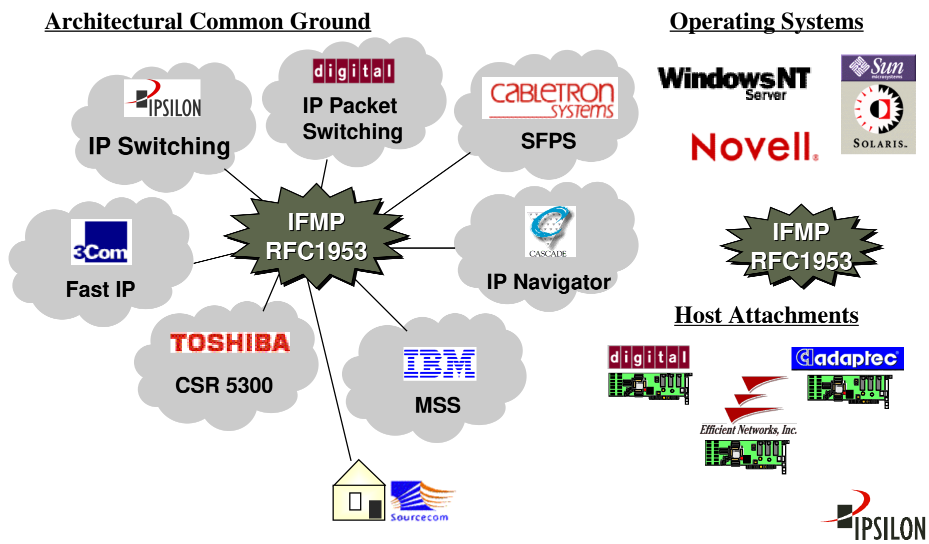
Imitation is the Sincerest Form of Flattery

- *Tag Switching — Cisco*
  - *Tag distribution protocol to propagate flow/VCI mapping*
  - *Mapping based on routing protocol, not data driven*
  - *Higher level of aggregation than current IP switching implementation*
  - *Ties the switching to the routing*
  - *External route changes require route redistribution*
  - *Does not support QoS differentiation for specific flows*
- *ARIS:* Some improvements over tag switching — *IBM*
  - *Establish shared tree to egress node*
  - *Order* N *work, order* N^2 *for tag switching*
  - *Not affected by external route changes*
  - *Uses virtual paths to avoid cell interleaving problem*
- *IPSOFACTO — NEC*
  - *Transmitter selects unused VC to reduce latency*

Joy of Flows: Flow Granularity

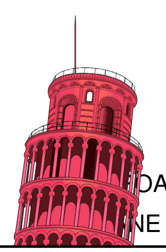


IP Switching Interoperability
RFC1953 IFMP



Contents

- *Why high-speed routing?*
- *Getting up to a gigabit*
- *Gigabit routers*
- *IP switching*
- *A rose by any other name…*
- ➜ *Conclusion*



IPSILON

Conclusion

- *Cost: IP switching utilizes volume hardware*
  - *Forwarding performance tracks processor growth curve*
  - *Switching performance tracks switch fabric growth curve*
- *Flow classification offers quality of service selection and sophisticated packet filtering policy for every flow*
- *IP multicast uses switch multicast hardware*
- *IP switching uses caching, Gigabit routers with hardware forwarding do not*
- *IP switching pushes layer 3 processing to edge of network thus requires lower forwarding performance*

DA

www.ipsilon.com/~pn

IPSILON