How scalable is the capacity of (electronic) IP routers?

Nick McKeown
Professor of Electrical Engineering and Computer Science, Stanford University
nickm@stanford.edu
http://www.stanford.edu/~nickm
Why ask the question?

Widely held assumption: Electronic IP routers will not keep up with link capacity.

Background:

\[
\text{Router Capacity} = (\text{number of lines}) \times (\text{line-rate})
\]

- Biggest router capacity 4 years ago \(\approx 10\text{Gb/s}\)
- Biggest router capacity 2 years ago \(\approx 40\text{Gb/s}\)
- Biggest router capacity today \(\approx 160\text{Gb/s}\)
- Next couple of generations: \(\approx 1-40\text{Tb/s}\)
Why it’s hard for capacity to keep up with link rates

Packet processing power

Link rate
Why it’s hard for capacity to keep up with link rates

Packet processing Power

Link Speed

Source: SPEC95Int & David Miller, Stanford.
Instructions per packet

What will happen

What we’d like: (more features) QoS, Multicast, Security, ...

Instructions per packet

time
What limits a router’s capacity?

- **Limited by memory random access time**
- **It’s a packet switch:**
  Must be able to buffer every packet for an unpredictable amount of time.
- **Hop-by-hop routing:**
  Once per ~1000 bits it must index into a forwarding table with ~100k entries.
- **[Optional QoS support**
  - Very complex per-packet processing]
What really limits the capacity?

• At first glance: the random access time to memory.

• In fact, this can be solved by more parallelism (replication and pipelining).

• Dilemma: But parallelism requires more power and space.
What *really* limits the capacity?

**Suggestion:**
- Don’t assume optics will oust CMOS in IP routers because of increased system capacity.
- It *might* oust CMOS because of reduced *(power x space)* for a given capacity.
Outline

• A brief history of IP routers
• Where they will go next
  - Incorporating optics into routers
  - More parallelism (with or without optics)
First Generation Routers

Typically <0.5Gb/s aggregate capacity
First Generation Routers

Queueing Structure: Centralized Shared Memory

Large, single dynamically allocated memory buffer:
N writes per “cell” time
N reads per “cell” time.
Limited by memory bandwidth.

Numerous work has proven and made possible QoS:
- Fairness
- Delay Guarantees
- Delay Variation Control
- Loss Guarantees
- Statistical Guarantees
Second Generation Routers

Typically <5Gb/s aggregate capacity
Second Generation Routers

Queueing Structure: Combined Input and Output Queueing

1 write per “cell” time

Rate of writes/reads determined by bus speed

1 read per “cell” time
Third Generation Routers

Switched Backplane

Typically <50Gb/s aggregate capacity
Third Generation Routers
Queueing Structure

1 write per “cell” time

Rate of writes/reads determined by switch fabric speedup

1 read per “cell” time

Switch

Arbiter

Flow-control backpressure

Per-flow/class or per-output queues (VOQs)

Per-flow/class or per-input queues
Third Generation Routers

- Size-constrained: 19” or 23” wide.
- Power-constrained.
Complex linecards

Typical IP Router Linecard

OC192c linecard:
~10-30M gates
~2Gbits of memory
~2 square feet
>$10k cost
100’s of Watts

"Backplane"
Fourth Generation Routers/Switches

Optics inside a router for the first time

Switch Core

The LCS Protocol

Optical links

1000's of feet

Linecards

0.3 - 10Tb/s routers in development
Where next?

- Incorporating (more) optics into a router.
- More parallelism (with or without optics).
Incorporating optics into a router

- Replacing the switch fabric with an optical datapath.
- Increasing the internal “cell” size to reduce rate of arbitration and reconfiguration.
Replacing the switch fabric with optics

Candidate technologies
1. MEMs
2. Fast tunable lasers + passive optical couplers
Replacing the switch fabric with optics

- Most common internal “cell” size is 64 bytes (50ns @ OC192, 12ns @ OC768)
- Too fast for arbitration
- Too fast for reconfiguration
- What we’ll see:
  - Increased cell length
  - E.g. switch bursts of cells
  - But less efficient.
More parallelism

- Parallel packet buffers
- Parallel lookup tables

Multiple parallel routers
Multiple parallel routers

What we’d like:

IP Router capacity
100s of Tb/s

The building blocks we’d like to use:
Why this might be a good idea

- Larger overall capacity
- Faster line rates
- Redundancy
- Familiarity
  - “After all, this is how the Internet is built”
Multiple parallel routers
Load Balancing architectures
**Method #1: Random packet load-balancing**

Method: As packets arrive they are randomly distributed, packet by packet over each router.

---

**Advantages:**
- Almost unlimited capacity
- Load-balancer is simple
- Load-balancer needs no packet buffering

**Disadvantages:**
- Random fluctuations in traffic $\Rightarrow$ each router is loaded differently
  - Packets within a flow may become mis-sequenced
  - It is not possible to predict the system performance
Method #2: Random flow load-balancing

Method: Each new flow (e.g. TCP connection) is randomly assigned to a router. All packets in a flow follow the same path.

Advantages:
- Almost unlimited capacity
- Load-balancer is simple (e.g. hashing of flow ID).
- Load-balancer needs no packet buffering.
- No mis-sequencing of packets within a flow.

Disadvantages:
- Random fluctuations in traffic ⇒ each router is loaded differently

It is not possible to predict the system performance
Observations

- Random load-balancing: It’s hard to predict system performance.

If designers, system builders, network operators etc. need to know the worst case performance, random load-balancing will not suffice.
Method #3: Intelligent packet load-balancing

Goal: Each new packet is carefully assigned to a router so that:

- Packets are not mis-sequenced.
- The throughput is maximized and understood.
- Delay of each packet can be controlled.

We call this “Parallel Packet Switching”
Method #3: Intelligent packet load-balancing

Parallel Packet Switching

![Diagram of parallel packet switching with routers and bufferless connections]
Parallel Packet Switching

**Advantages**
- Single-stage of buffering
- No excess link capacity
- $k \uparrow \Rightarrow$ power per subsystem $\downarrow$
- $k \uparrow \Rightarrow$ memory bandwidth $\downarrow$
- $k \uparrow \Rightarrow$ lookup rate $\downarrow$
Precise Emulation of a Shared Memory Switch

Shared Memory Switch

Parallel Packet Switch

N

= ?

N
Parallel Packet Switch

Theorem

1. If $S > 2k/(k+2) \approx 2$ then a parallel packet switch can precisely emulate a FCFS shared memory switch for all traffic.
Example of an IP Router with Parallel Packet Switching

Overall capacity 160Tb/s
My conclusions

• The capacity of electronic IP routers will scale a long way yet.

• The opportunity of optics is to reduce power and space
  - By using optics within the router.
  - By replacing routers with circuit switches.