[illegible]:　　　Lecture 3: Router Design

- Router lay-out in an ISP　10
- Router design in GSR　30
  and Challenges (McK97)
- Scaling and moving routers — degree / why not more routers.
- Research — buffers
  - centralize switch vs. distributed switch.
  - power-aware design and algorithms.
- Takeaway.

- Access trees, Border routers, backbone routers | show switching capacities.
  PoPs or central offices, racks based organization
- Observations about link speeds and no degrees.
- functions of various routers.
- Router design discussions → backbone routers.

Any router:
- Control: coarse, but important for overall correctness
- Data: fine-grained, fast, simple
- Management: coarse, but critical to diagnose faulty failures

Router design: fast data plane implementation in backbone routers.
Data plane functionality: forwarding, packet classification, QoS marks, filtering, deep packet inspection, route/address lookup

~~Forwarding~~ line cards and backplane organization, CPU location

Switching capacity needed.
　cross bar → parallelism.
　N×N
　heirarchical switching systems...

(1) Scheduler role:
(2) Packet length: throughput
(3) Blocking: HOL → VOQ.
　input, output ~ delay.

- Simple, high throughput, fast and Starvation free.

- describe the arbiter → implementation.

- throughput is good, but delay is unpredictable
  - priority,
  - speedup. output port memory and link speeds.

- Multicast. - different queue.
  leverage natural mcast
  fan-out and no-fanout split

10

(1) Router scaling:

traffic volumes grow over time
→ roughly doubles every 18 months ~~or~~ or so.
→ link capacities at edges increases allowing more traffic to be pumped into core

⇒ Network switching capacity must also increase correspondingly. → scale up core routers
→ add a core router hierarchy, more routers and increase switching capacity.

But ~~do~~ we necessarily need routers which can switch more ~~and~~ more traffic.

Second option means we don't necessarily need to build faster and faster high-end routers

But is this feasible

⇒ central offices are full

⇒ building central offices expensive

⇒ increasing switching capacity and replacing older routers is a more viable option.

1 Tbs → 10 Tbps → 1000 . . .

But what is the challenge for the switching fabric.

Racks today can provide 10KW per rack (cooling and placement becomes hard)

at 10KW → can't switch more than 2.5 Tbps today

Switching challenge: centralized switch is limited ~~in throughput.~~
has to scale x^2 where x is scaling factor for link capacity

10 And thus is exactly how things have progressed over time so far

The higher and higher capacity routers consume more and more power.
↳ more and more RACKs. to spread power density.
↓
one trend: distributed multi-stage switch ⇒ unpredictable performance.

A single stage switch fabric that is central ⇒ good performance
∟ can fit on a single rack.
but scalability is an issue
Switch hardware has to run x^2 faster where x is the ~~switch~~ increase in link capacity

Todays Racks ← a maximum of 10KW per rack. (cooling is hard)
← 2.5 Tbps. due to the limitation of central arbiter.

How do you build a single rack switch fabric that can switch Tbps under 10 kW.

(2) A second challenge is in the design of buffers and how these are design, what memories to use.

How much buffer to provision?

→ Rule of thumb to ensure good utilization

40Gbps @ 0.5s → lot of memory.

SRAM . . . . heat and cost — 5ns.

DRAM lot fewer chips, less power-consuming but slow — 50ns

Multiple DRAM chips with a wide bus.
- Power and expensive.

Another issue is do we really need such large buffers.

(3) Power aware design and routing