- 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 mesh fan-out and no-fan-out split

1. Router scaling:

  traffic volumes grow over time
  
  \[ \text{roughly doubles every 15 months or so.} \]
  
  \[ \text{link capacities of edges increase, allowing more traffic to be} \]
  
  \[ \text{pumped into core} \]
  
  \[ \text{network switching capacity must also increase} \]
  
  \[ \text{correspondingly.} \]

  \[ \text{scale up core routers} \]
  
  \[ \text{add a core router hierarchy, more routers and} \]
  
  \[ \text{increase switching capacity.} \]

  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?

\[ \text{central offices are full} \]

\[ \text{building central office expensive} \]

\[ \text{increasing switching capacity and replacing older} \]

routers is a more viable option.
1 Tbps → 10 Tbps → 1000 …

But what is the challenge for the switching fabric?

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

At 10 kW → can 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.

And this is exactly how things have progressed over time & for.

The higher and higher capacity switches consume more and more power.

→ more and more racks to spread power directly.

One trend: distributed multi-stage switch → unpredictable performance.

A single stage switch that is central → good performance.

→ can fit on a single rack.

But scalability is an issue.

Switch hardware has to even $x^2$ faster where

$x$ is the switch increase in link capacity.

Today's racks → a max of 10 kW per rack. (cooling is hard)

→ 2.5 Tbps → due to the limitation of central switches.
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 designed, what memories to use.

How much buffer to provision?

→ Rule of thumb to ensure good utilization

40Gbps @ 0.5s → lot of memory.

SEAM ... 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 huge buffers.

3) Power aware design and routing.