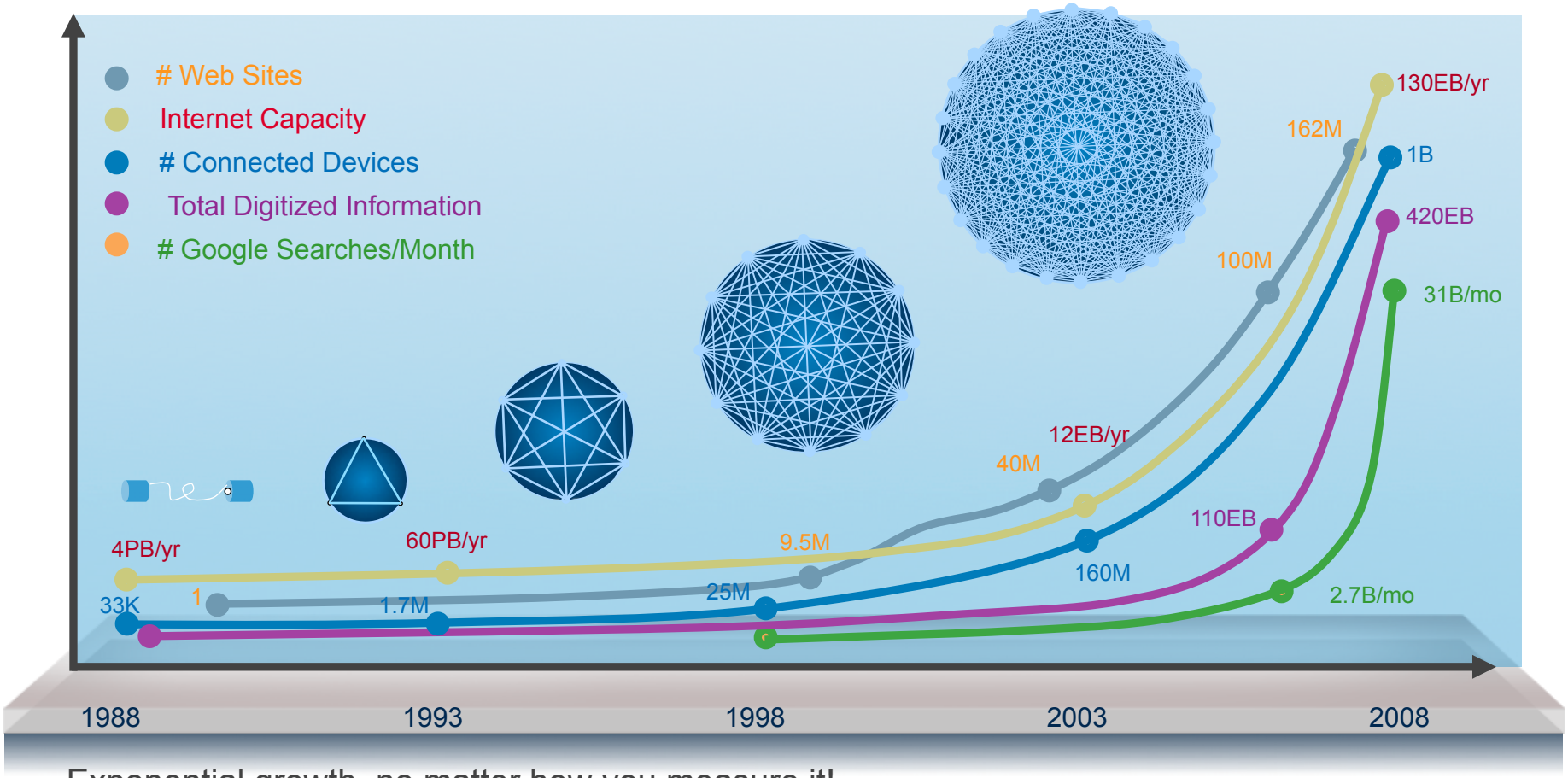
MODERN SERVICE PROVIDER ROUTERS

Guy C Fedorkow
Distinguished Engineer, Juniper Networks
April 2013

THE INTERNET EXPLOSION



Exponential growth, no matter how you measure it!

The clearest indication of value delivered to end-users

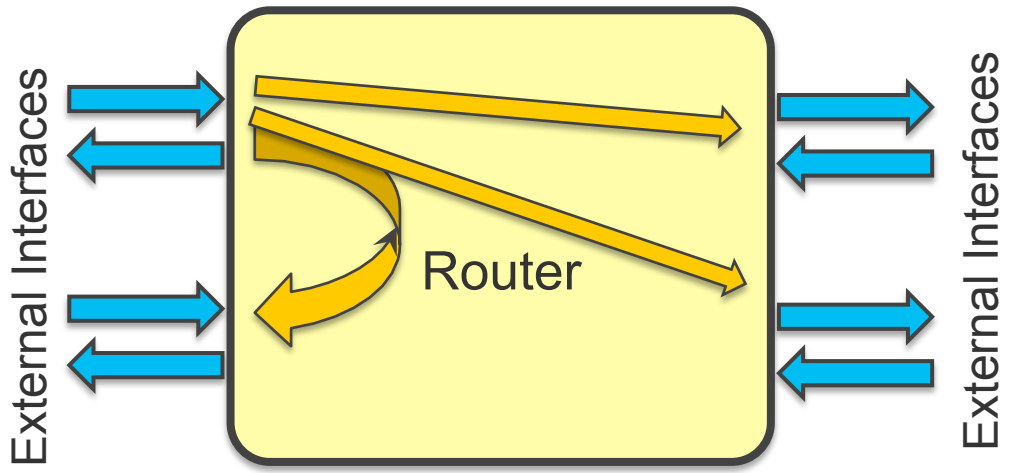
WHAT'S A ROUTER

From RFC-1812

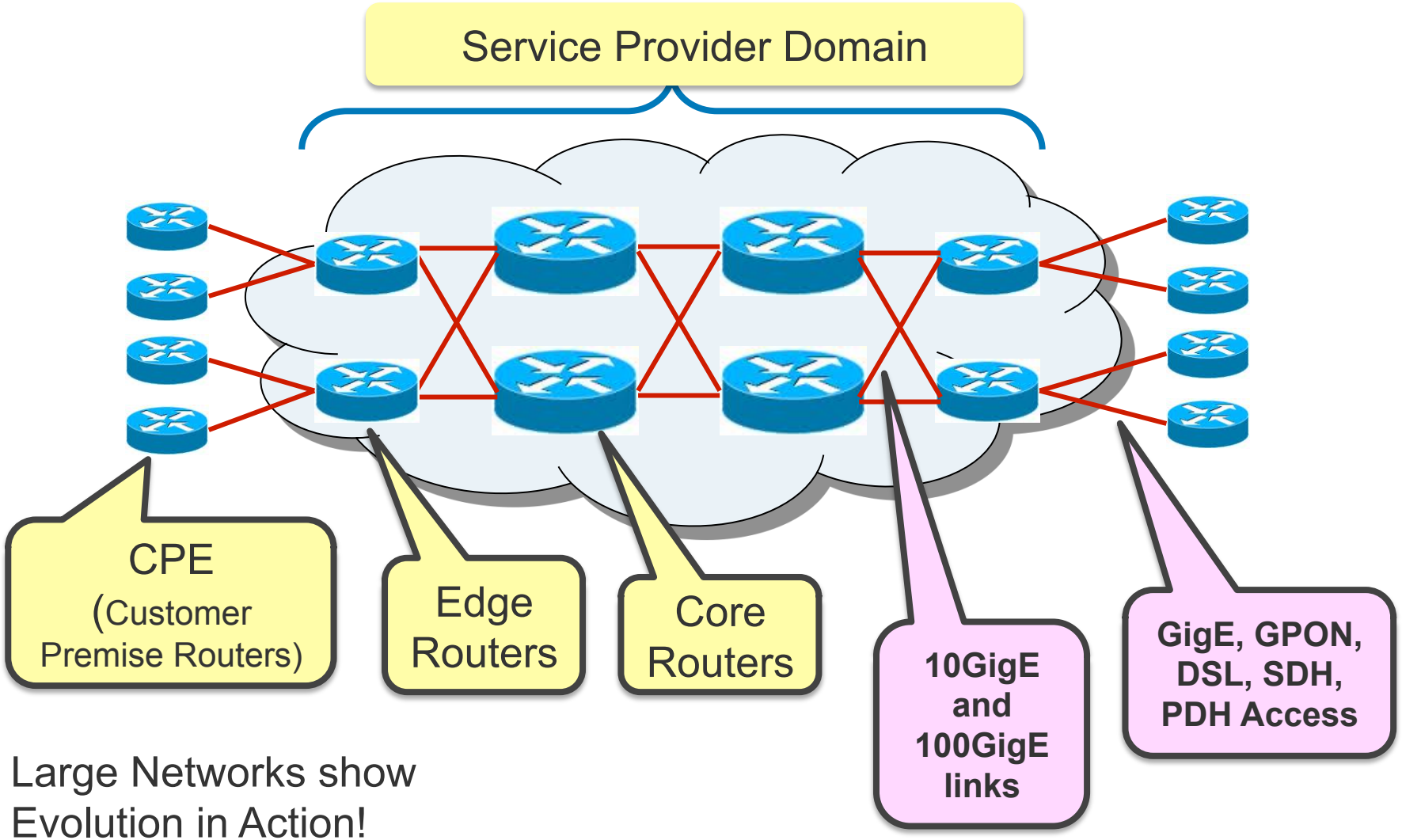
"An IP router can be distinguished from other sorts of packet switching devices in that a router examines the IP protocol header as part of the switching process. It generally removes the Link Layer header a message was received with, modifies the IP header, and replaces the Link Layer header for retransmission."

Routers:
- Strip the L2 Header
- Decrement the TTL
- Look up the Next Hop
- Add New L2 header
- Transmit the Packet

And also
Routing topology, Routing policy, Access Control, Filters, Rate Policing, Shaping, Traffic Prioritization, Buffering, **Tunnels**, MPLS, v4/v6 interworking, NAT, Subscriber Authentication, etc, etc.



NATIONAL ISP NETWORK ARCHITECTURE



Large Networks show Evolution in Action!

WHAT'S A SERVICE PROVIDER ROUTER

Core Routers

- Highest Bandwidth Capacity – Nx100GigE
- Emphasis on Route Scaling – Millions of Routes
- Modest Forwarding Feature Set

Common Features

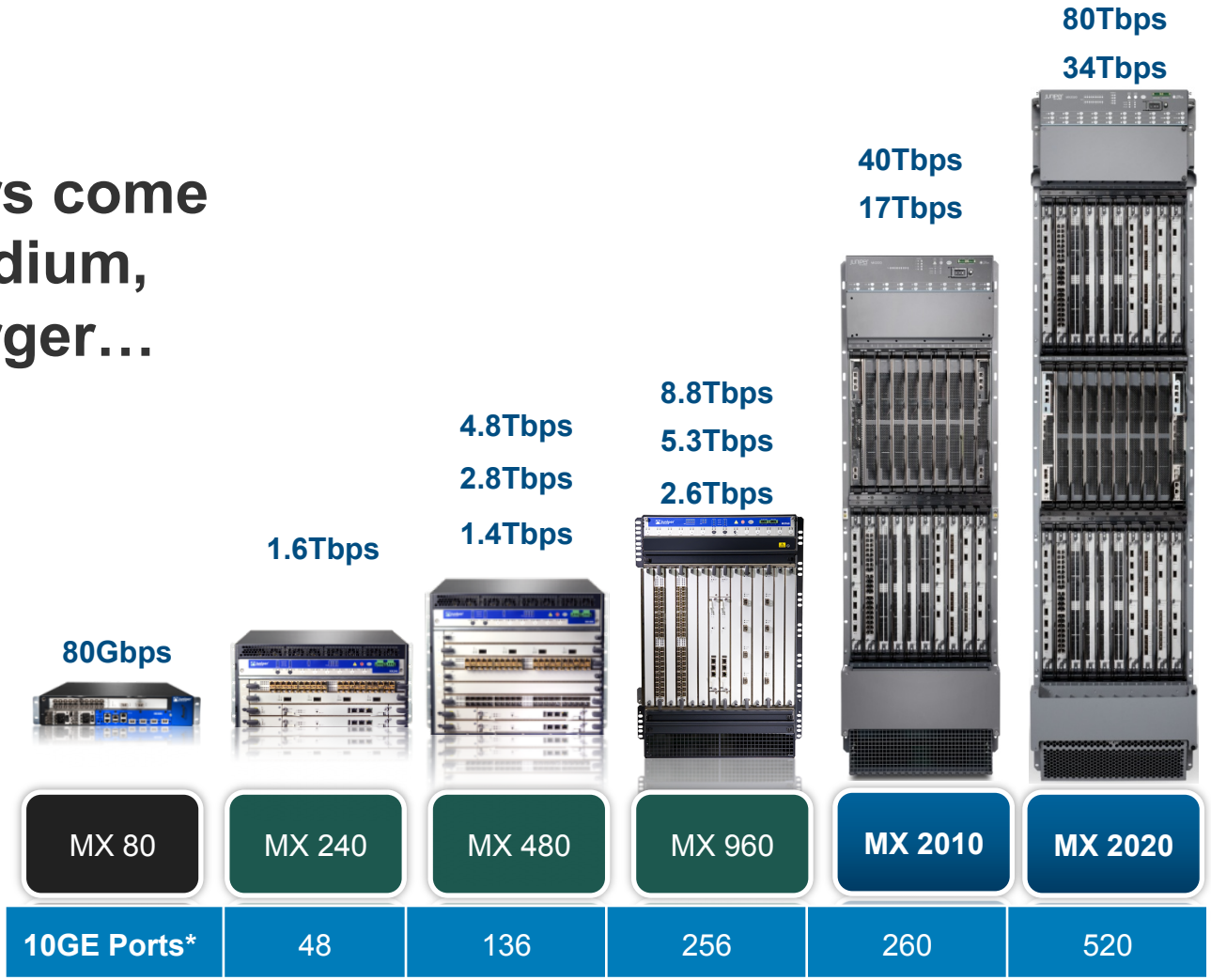
- Common Element Redundancy
- Hot-Swap & ISSU
- BGP Control Plane
- Demanding Environmental (e.g. 55C)
- Redundant DC(!) Power

Edge Routers

- Highest Interface Scaling – Hundreds of GigE and 10GigE interfaces
- Complex Forwarding Feature Set
- Emphasis on Subscriber Scaling

SAMPLE EDGE ROUTER PORTFOLIO

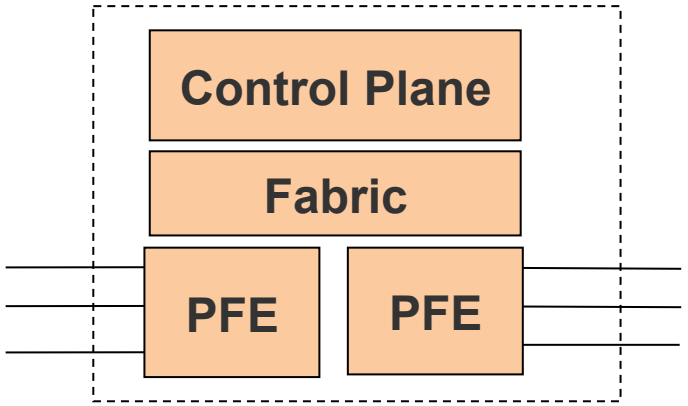
Edge Routers come in small, medium, large and larger…



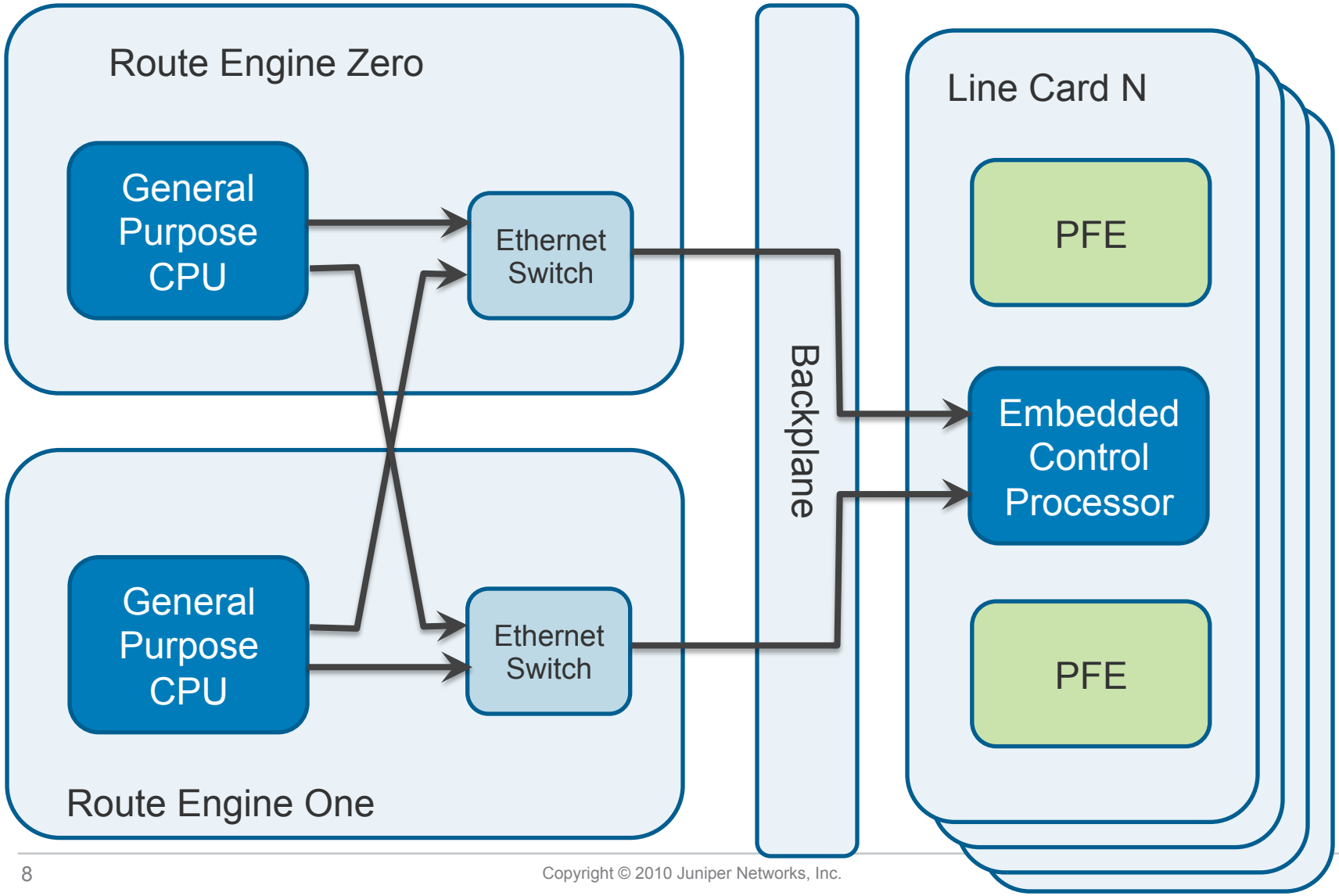
KEY ELEMENTS

High-Scale Routers comprise several key elements:

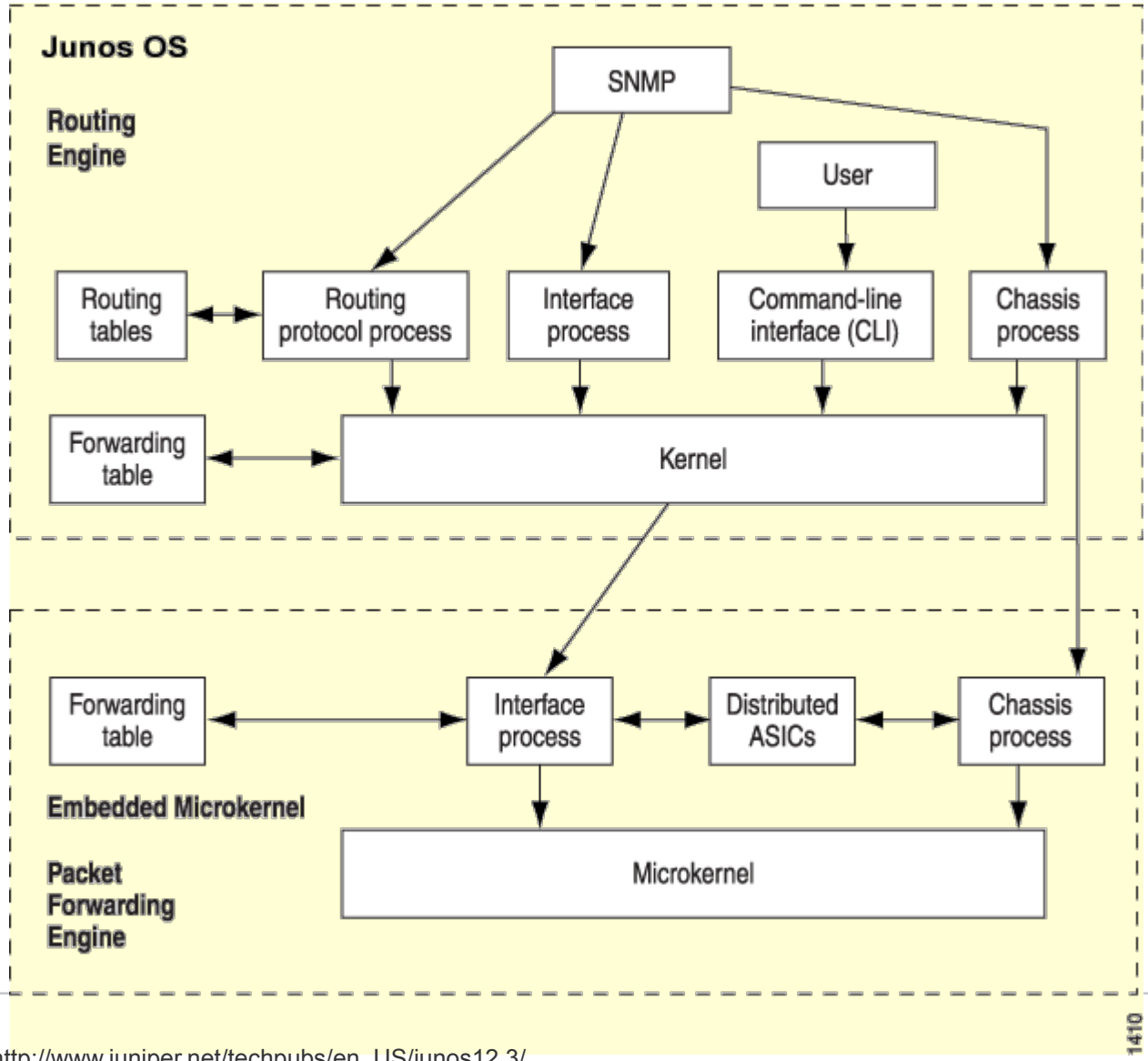
- **Control Plane**
  - Responsible for managing routing tables, authenticating subscribers, configuring interfaces
- **Packet Forwarding Engine(s) (PFE)**
  - Responsible for forwarding each packet (i.e. address lookup, queues, access lists, etc)
- **Fabric**
  - Responsible for moving packets from one line card to another inside the router



ROUTER CONTROL PLANE



JUNOS SOFTWARE ARCHITECTURE



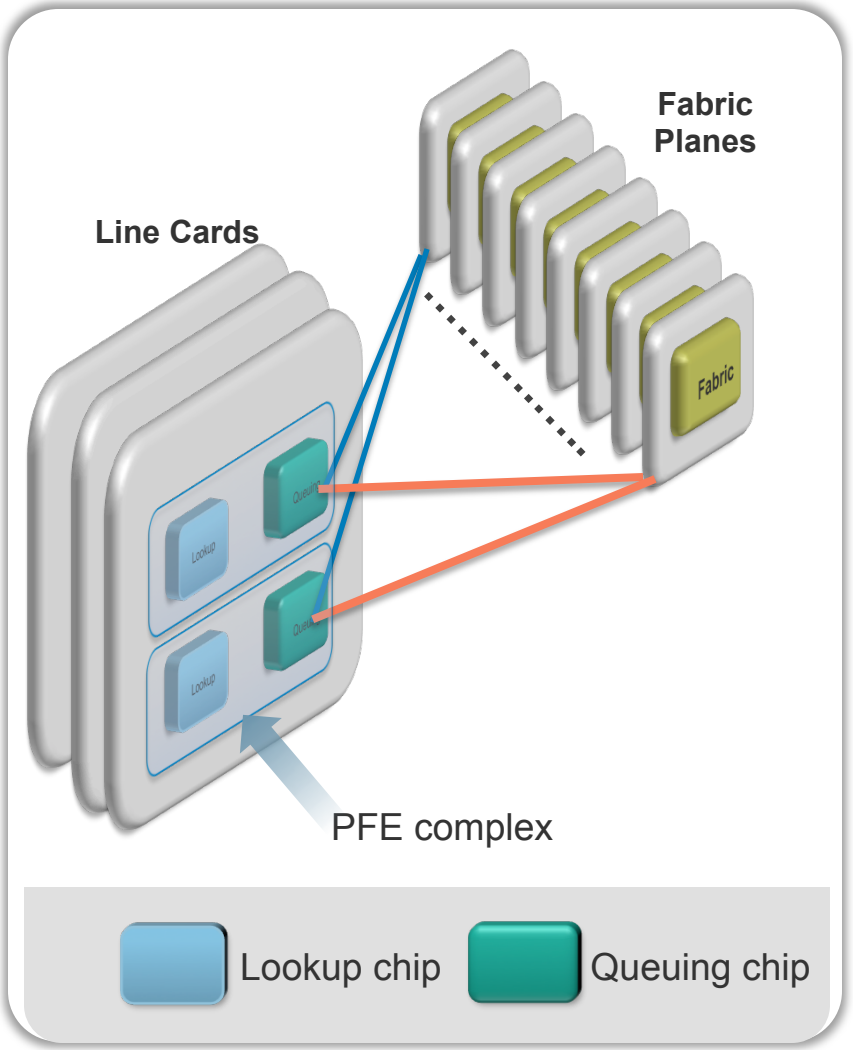
http://www.juniper.net/techpubs/en_US/junos12.3/topics/concept/junos-software-architecture.html

FABRIC-BASED ARCHITECTURE

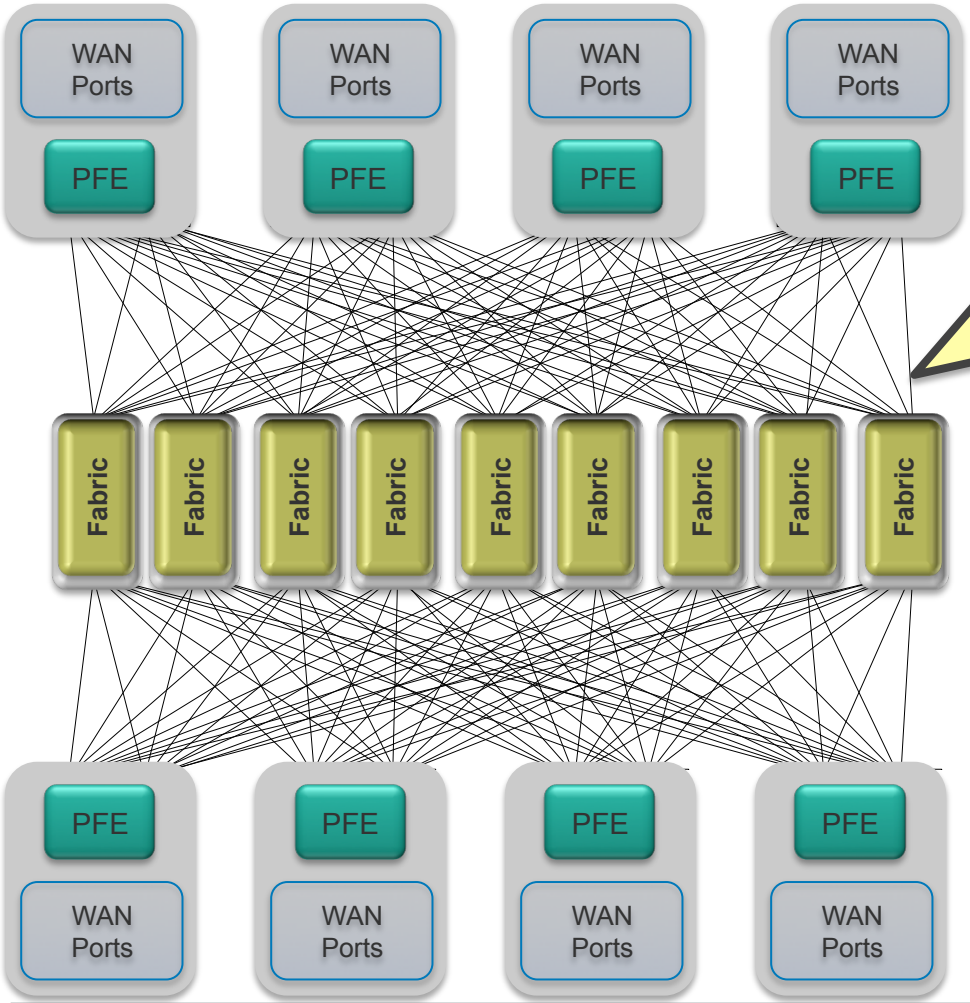
Most high-scale routers are Fabric Based

- Multiple Line Cards, each containing PFEs

- A chassis-wide Interconnect Fabric transfers traffic from ingress to egress line cards



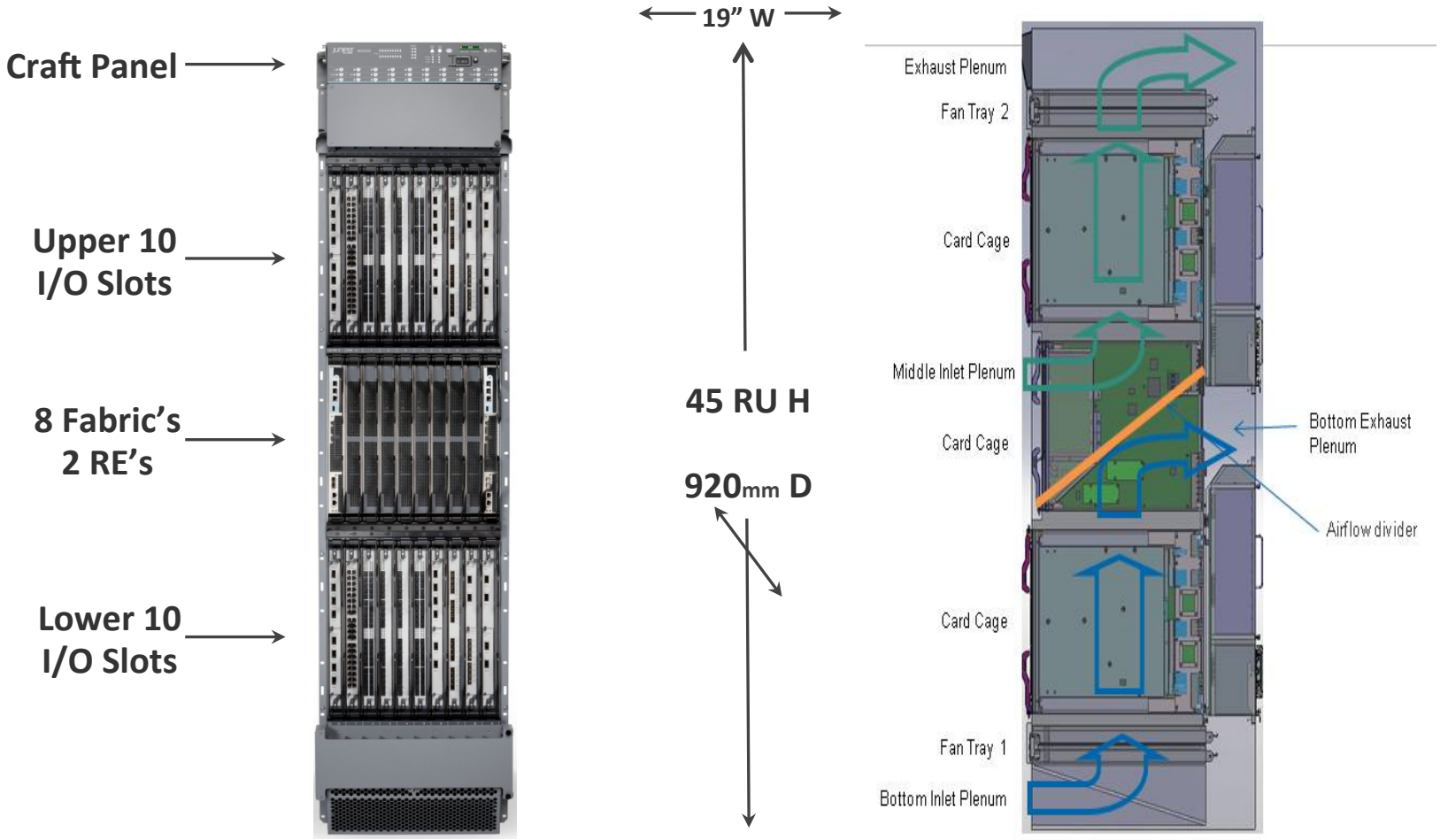
FABRIC-BASED ARCHITECTURE



Each line card connects to all Fabric cards.

There are a lot of wires on the backplane!

MX2020 PACKAGING

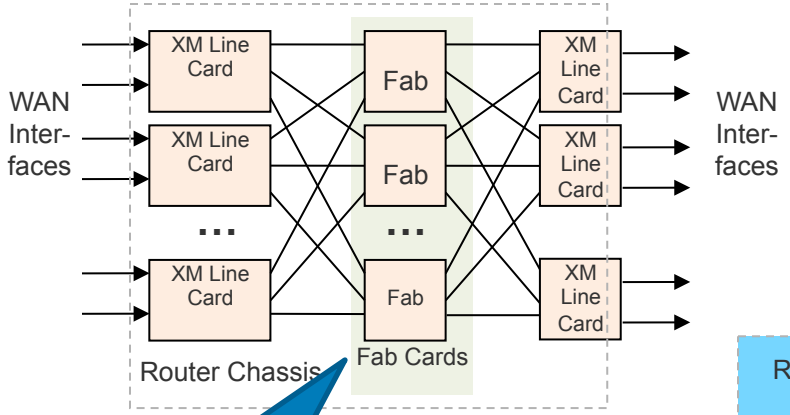


CHASSIS FRONT

CHASSIS SIDE

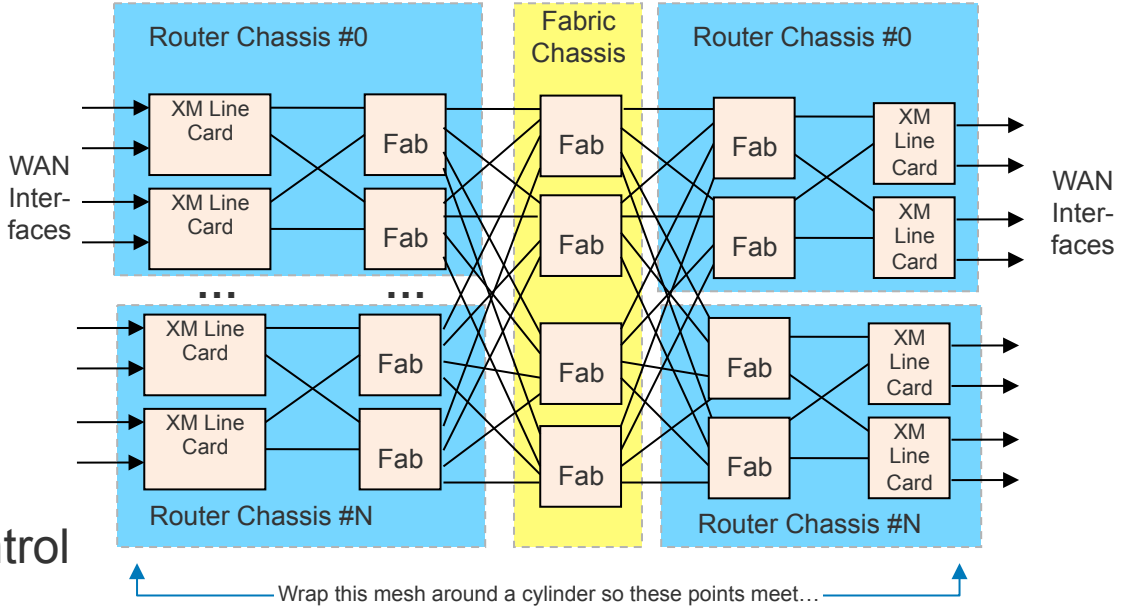
IF ONE CHASSIS IS TOO SMALL… MULTICHASSIS ROUTERS

What to do when one chassis isn't big enough?



Single Stage fabric only scales so far…

…So if that's not enough, go to multiple fabric stages using Benes or Butterfly networks



Issues to Solve:

- Expensive optical cables
- SW Scaling & Distributed Control
- Proto costs(!)

PACKET FORWARDING ENGINE (PFE)

PFE's do the work to move packets from Ingress to Egress

Key Functions:

- L2 & L3 Analysis & Features

  Figure out who's packet it is, what should happen to it, and where it should go.

- Packet Buffering

  Store the Packet in DRAM until there's room to transmit it

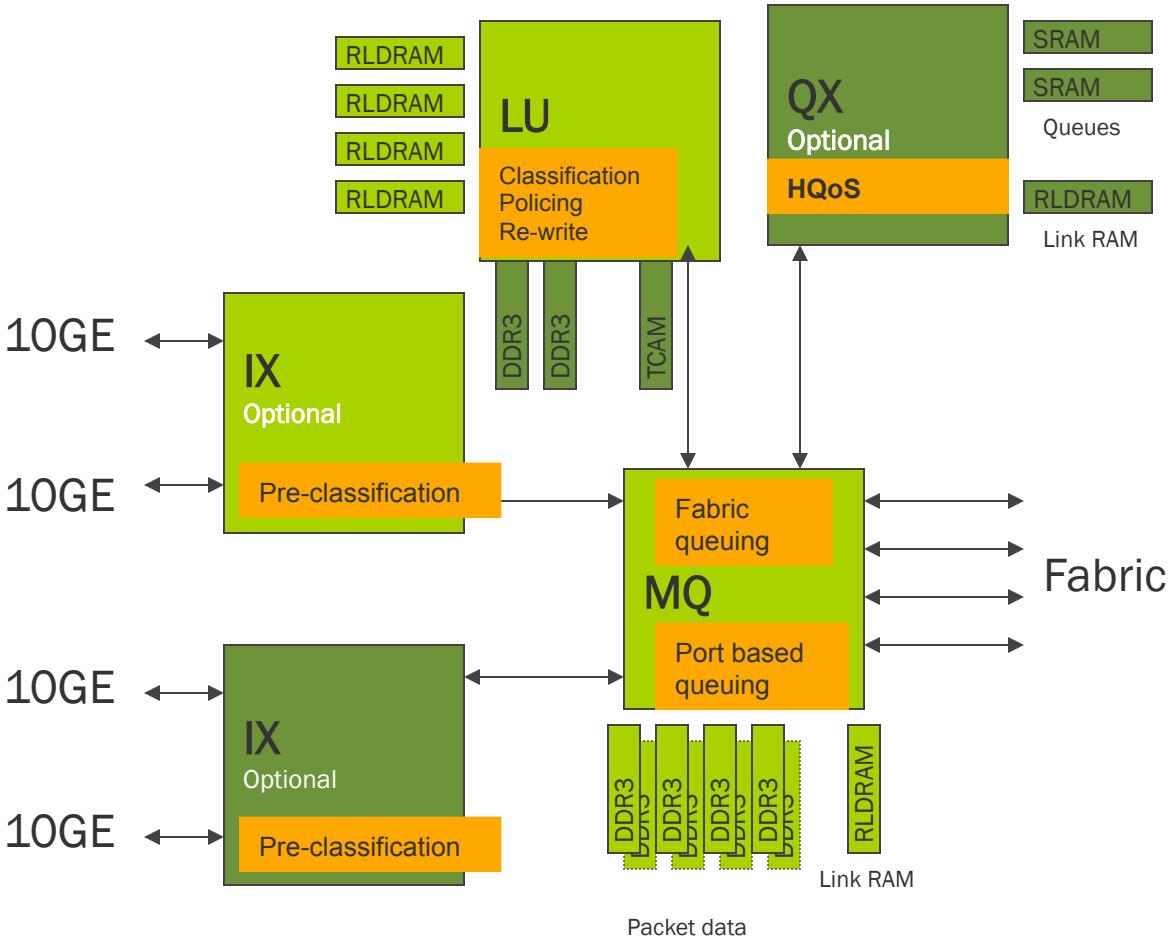
- Queuing & Scheduling

  Decide which packets should go in what order to achieve fairness and real-time delivery guarantees.

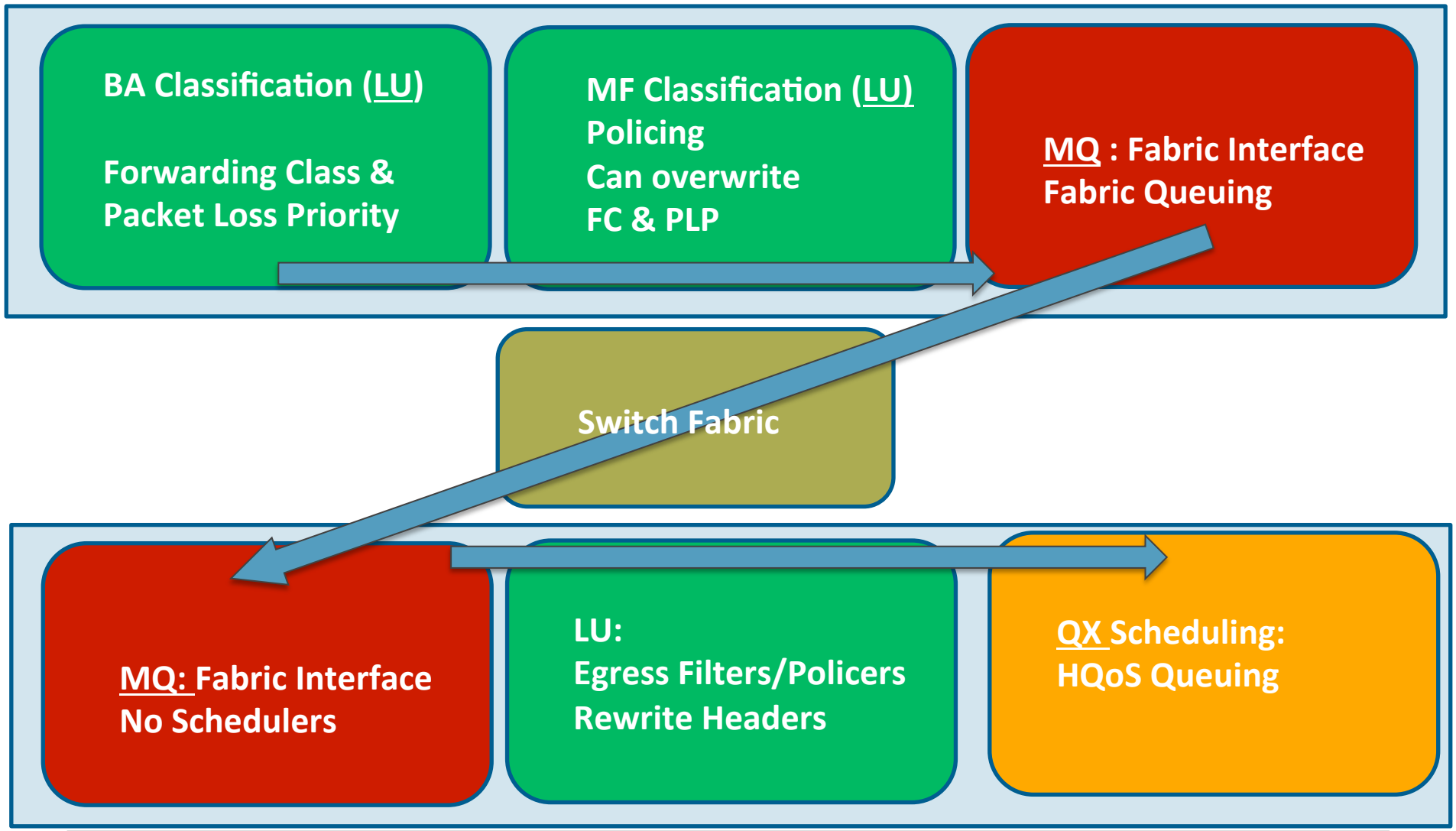
PFEs may be micro-programmable, table-driven or hard-coded

[It's the old Cost/Performance/Flexibility Tradeoff Matrix…]

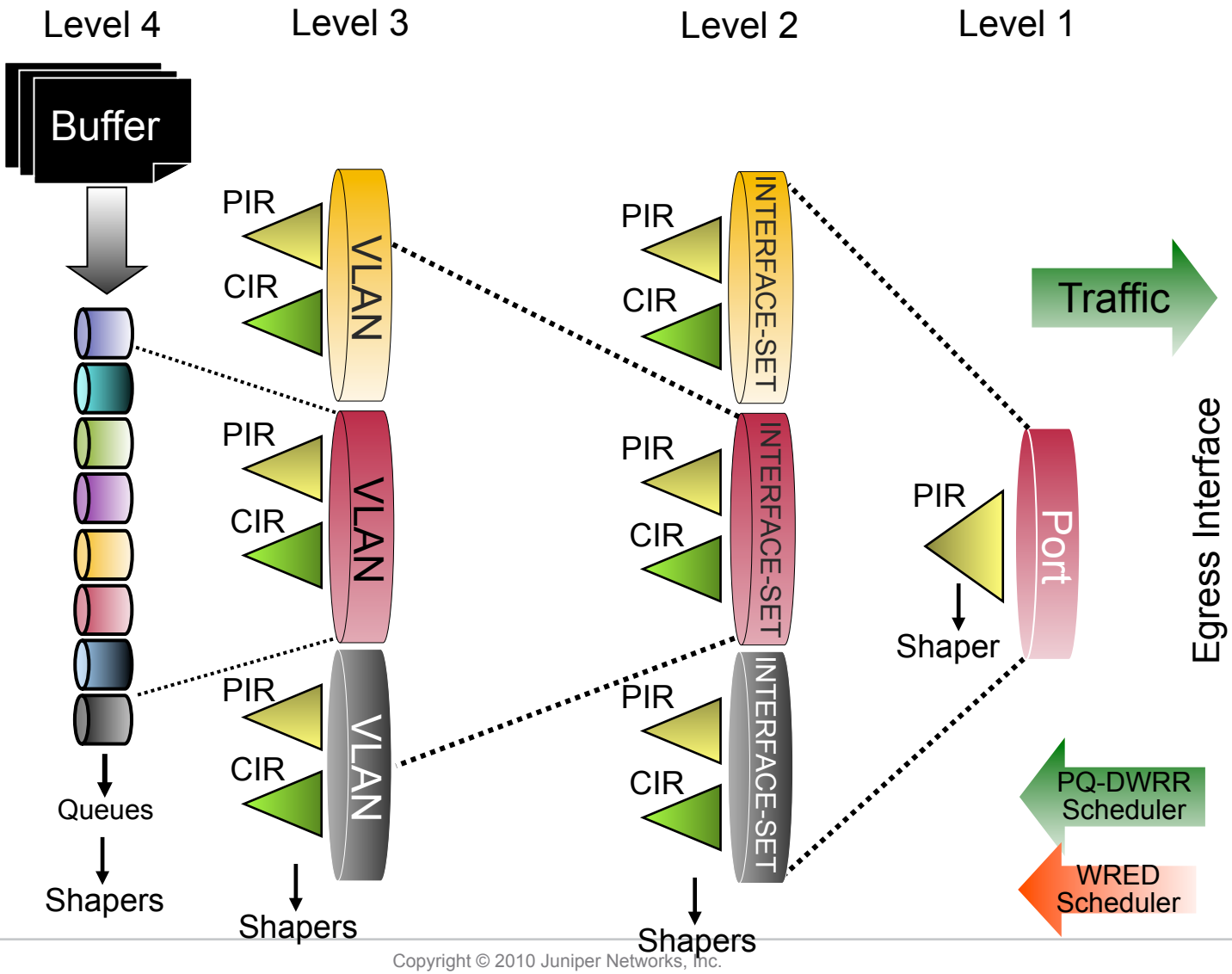
TRIO PFE ARCHITECTURE



SAMPLE TRIO PACKET PATH



FANCY QUEUING

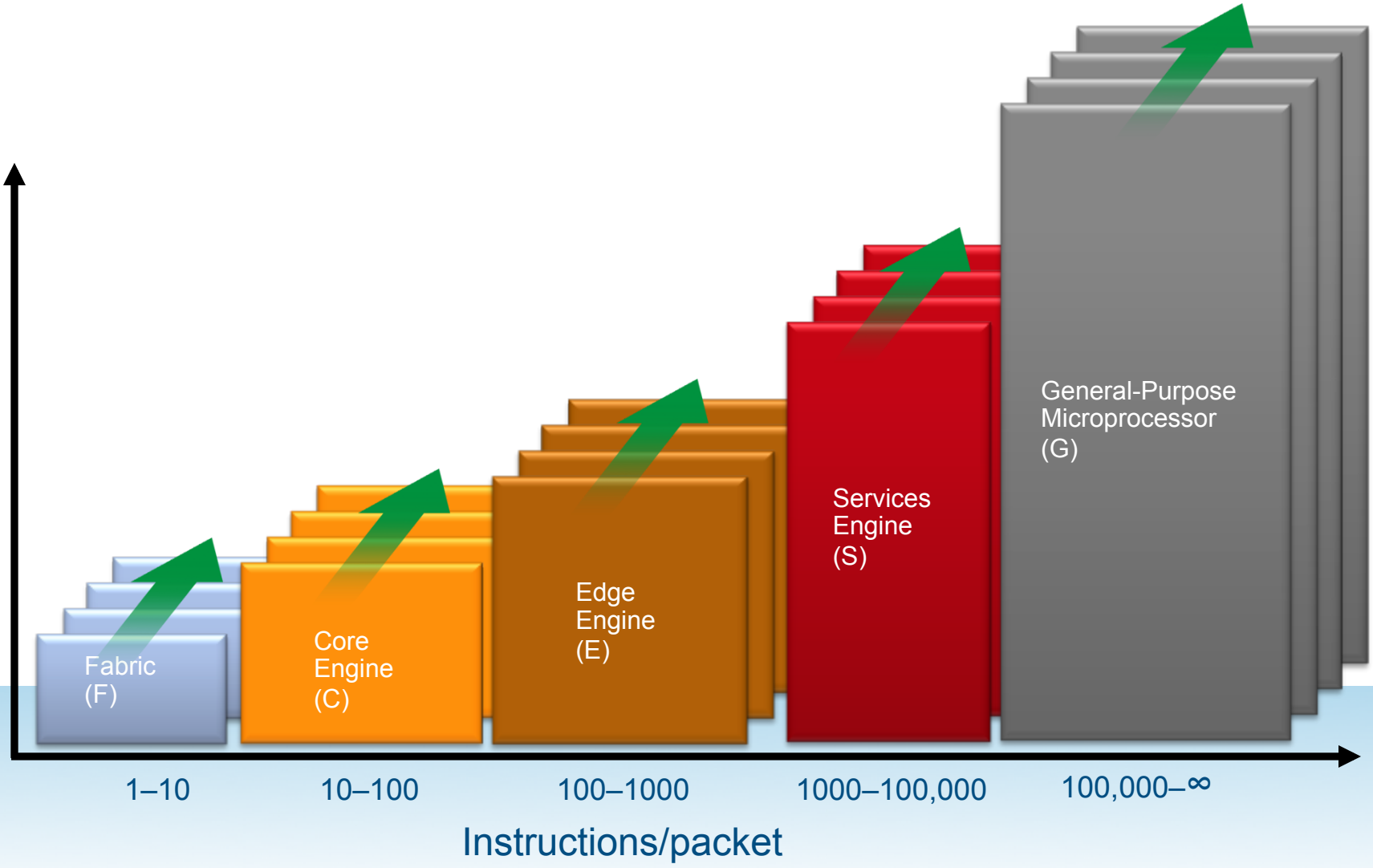


WHAT'S HARD ABOUT HIGH-SCALE ROUTERS?

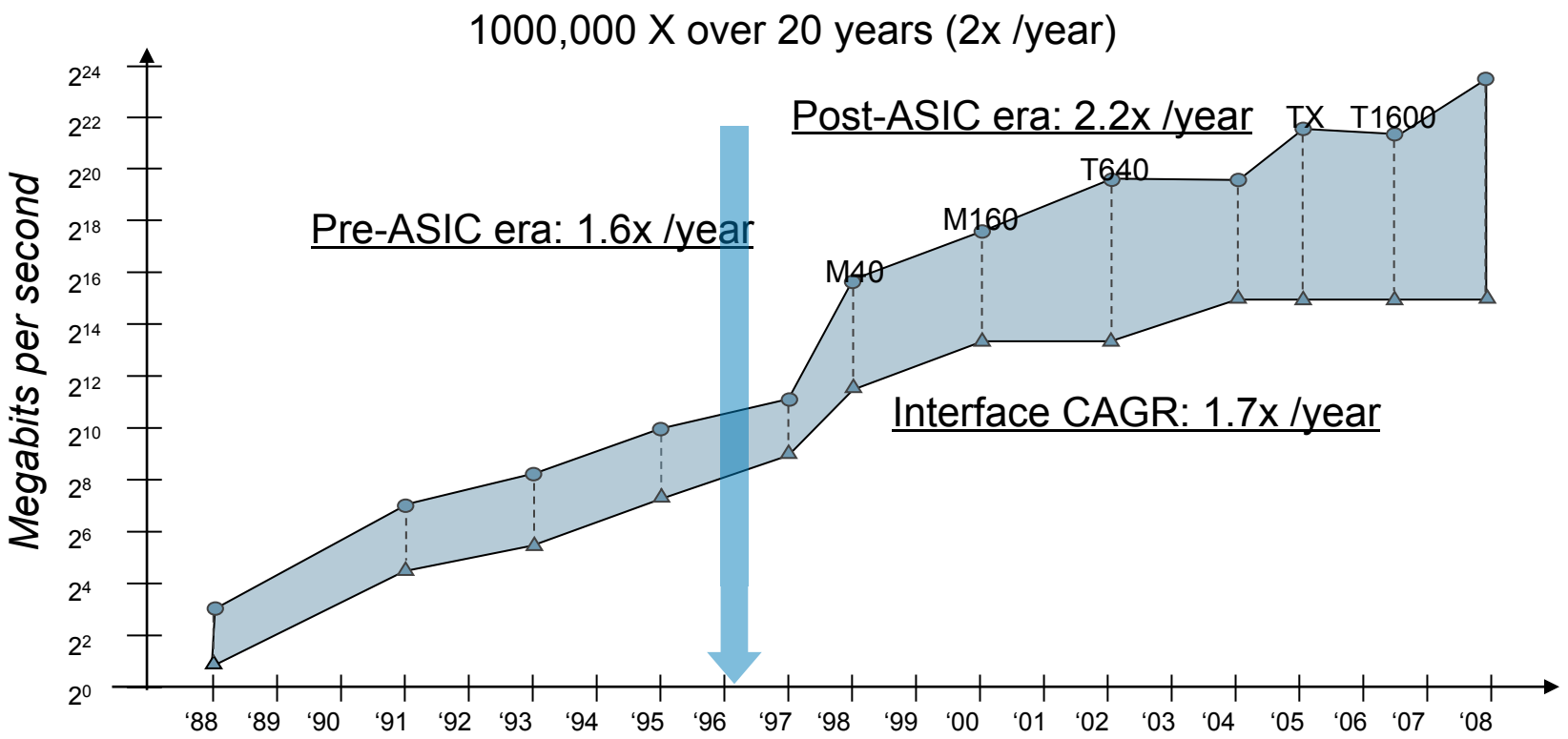
- Power Management
  - Keeping power dissipation down
  - Getting the heat out
- Signal Integrity
  - Crosstalk from a zillion wires on the backplane
- Features
  - QoS
  - Multicast
  - Tunnels
  - Link Aggregation & ECMP Load Balancing

    -- and -- "Feature Velocity"
- Scaling the Control Plane -- FIB and Subscribers
- Service Delivery (NAT&Firewall, Content Delivery, DPI, & who knows what else)

SILICON
THE FOUNDATION OF PERFORMANCE



ROUTER PERFORMANCE 1988 – 2008



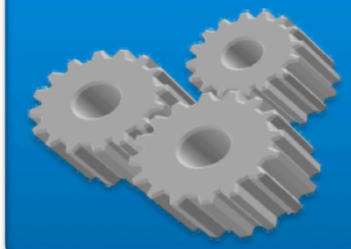
MEMORY CHOICES WITH NETWORKING ASICS

Packet buffering

- Need high throughput, high density
- Long bursts ok
- SDRAM or RLDRAM (Reduced Latency DRAM)

Queuing/Link memory

- Need high throughput, low latency
- Shorter bursts
- SRAM, RLDRAM, or SDRAM

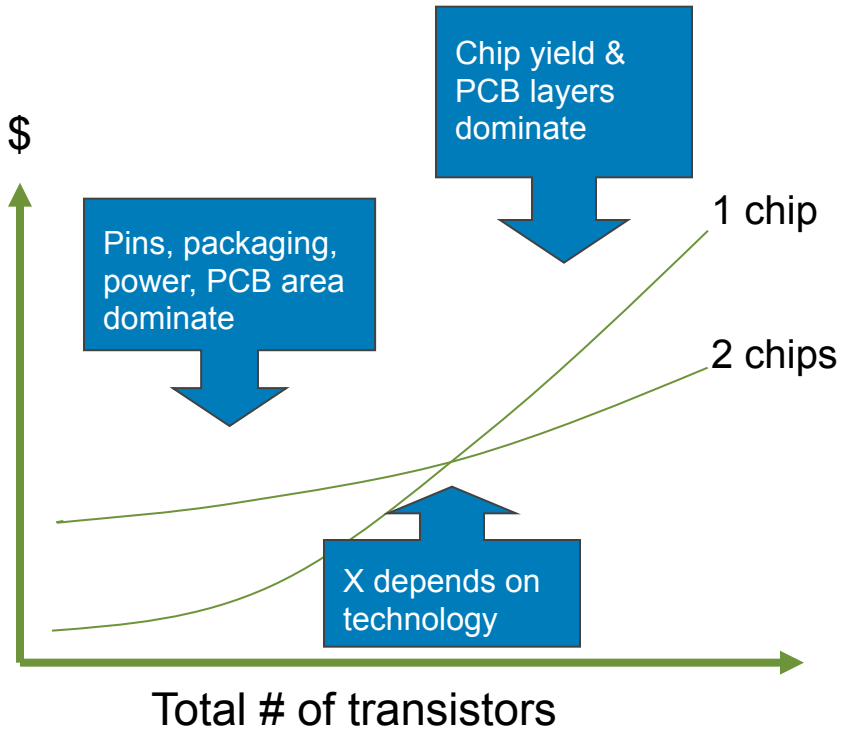


Control memory

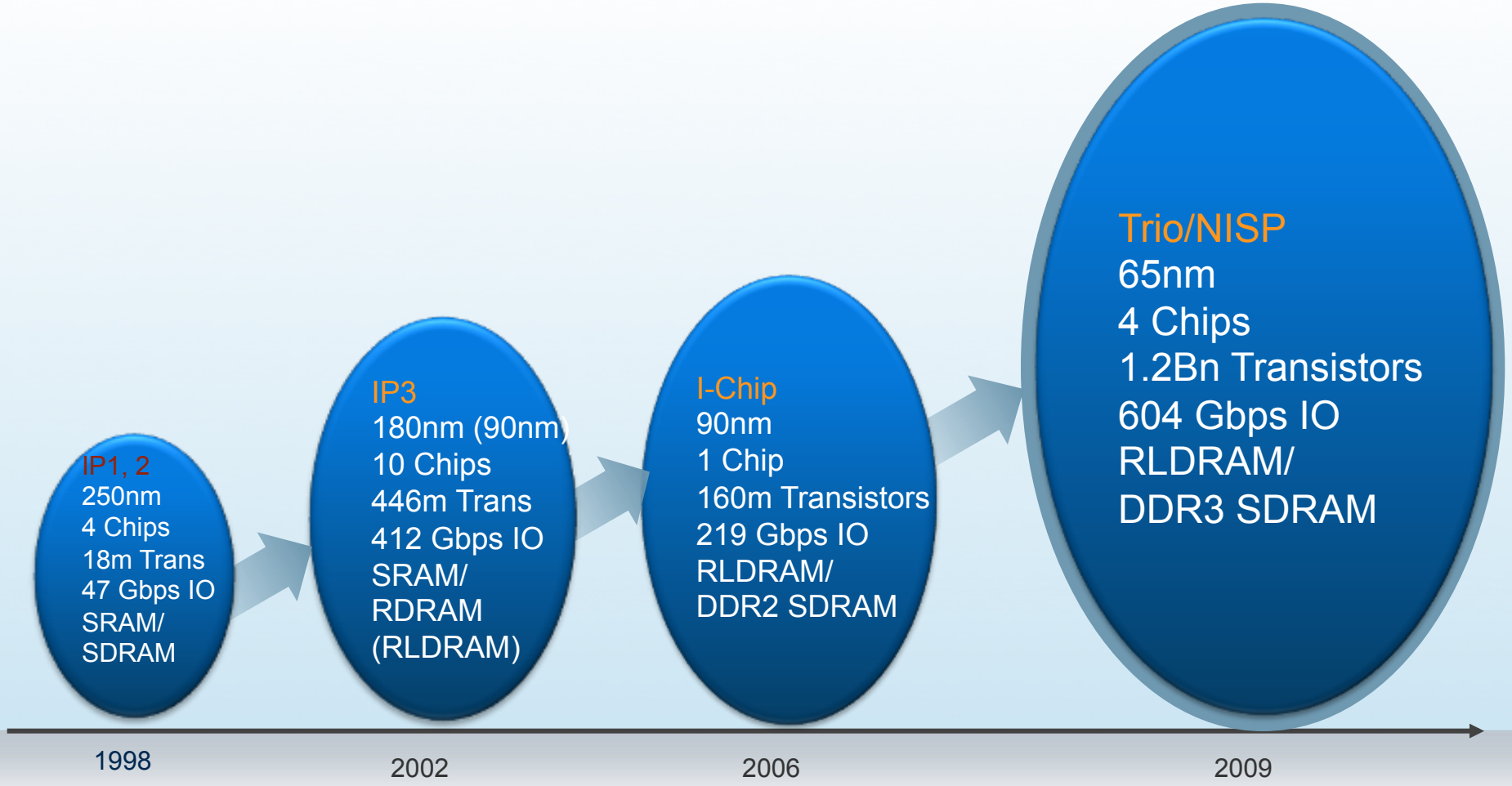
- Need high throughput, low latency
- Even smaller access quantum
- SRAM, TCAM, or RLDRAM

ARCHITECTURE – CHIP PARTITIONING

- Fewer chips does not necessarily mean less overall cost
  - Chips get very expensive once they cross a certain die size
  - Economics of silicon is all about fabrication yield
- Goals
  - Balance size of each chip within packet forwarding engine
  - Minimize pin-count on each chip
  - Minimize overall component cost
  - Flexibility of support different configs with the same chipset



EXAMPLES OF SILICON PROCESS IMPROVEMENT, CHIP PARTITIONING, AND MEMORY USAGE



Questions?

everywhere

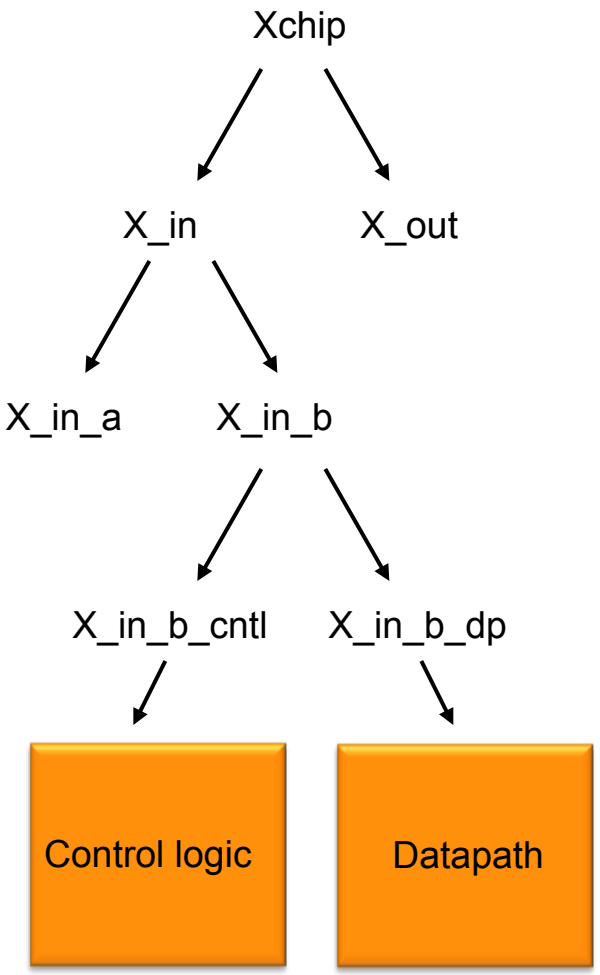
For further information on router functionality:
Juniper MX Series, Douglas Richard Hanks Jr and Harry Reynolds, published by O'Reilly

MICRO ARCHITECTURE

Take each subsystem, divide into blocks, divide each block into sub-blocks, design down to the basic logic elements

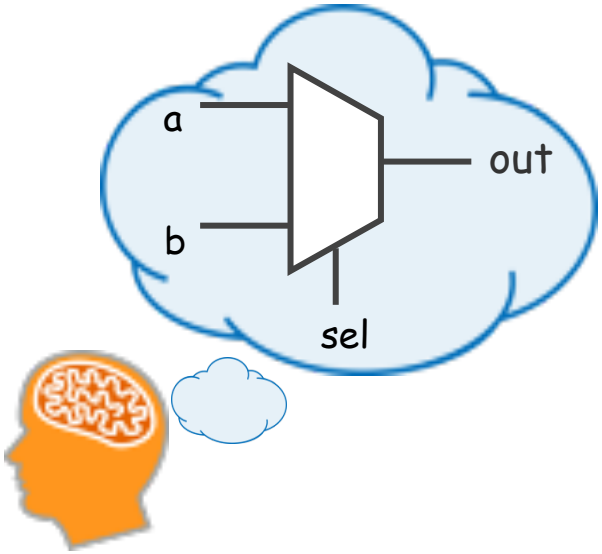
Document both functionality and architecture

- Rigorous peer reviews of all documents



REGISTER TRANSFER LEVEL CODING

Translate micro architecture for all blocks to “Register Transfer Level” code.




```
always @ (sel or a or b)
   begin
            if (sel == 1)
                        out = a;
            else
                        out = b;
end
                  OR
assign out = sel ? a : b;
```

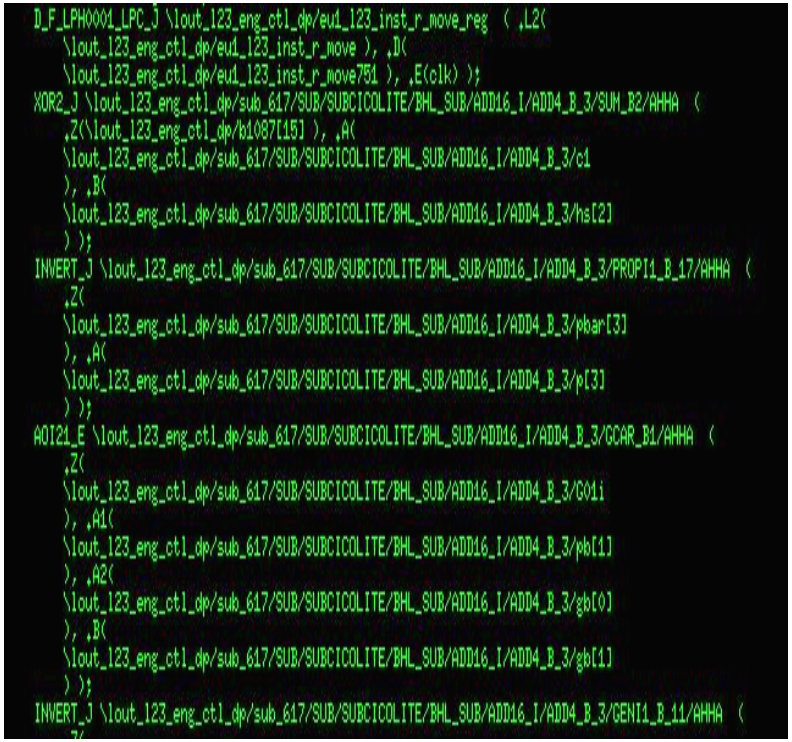
- A large chip will have hundreds of thousands of lines of RTL code
- Must always keep in mind physical placement and timing during the micro architecture phase
  - You pay now or you pay more later

SYNTHESIS & TIMING

Synthesis is the exercise of mapping RTL to GATES in the technology of choice

INPUT

- RTL code
- Specification of clocks and cycle-time (frequency)
- Input and output constraints for module being synthesized
- Wire-load models as basis to model interconnect effects on gates
- Recent trends: physical synthesis



VERIFICATION

Goal: First-time-right silicon

- Avoid expensive ASIC respins
- Simulations are far easier to debug than real chips

Recipe: At least as many verification engineers as design engineers per chip

Performed at multiple levels

- Block level
- Chip level
- Sub-system level
- System level
- Software/hardware co-simulation

TOOLS

Test-bench tool
SystemVerilog
C/C++, Verilog
Coverage tools
Equivalency checkers
Simulators
Waveform viewers

PHYSICAL DESIGN

Power and clock planning

Perform high-level floor-planning

Place I/O, SRAMs, & Register Arrays

Random logic placements

Perform congestion analysis

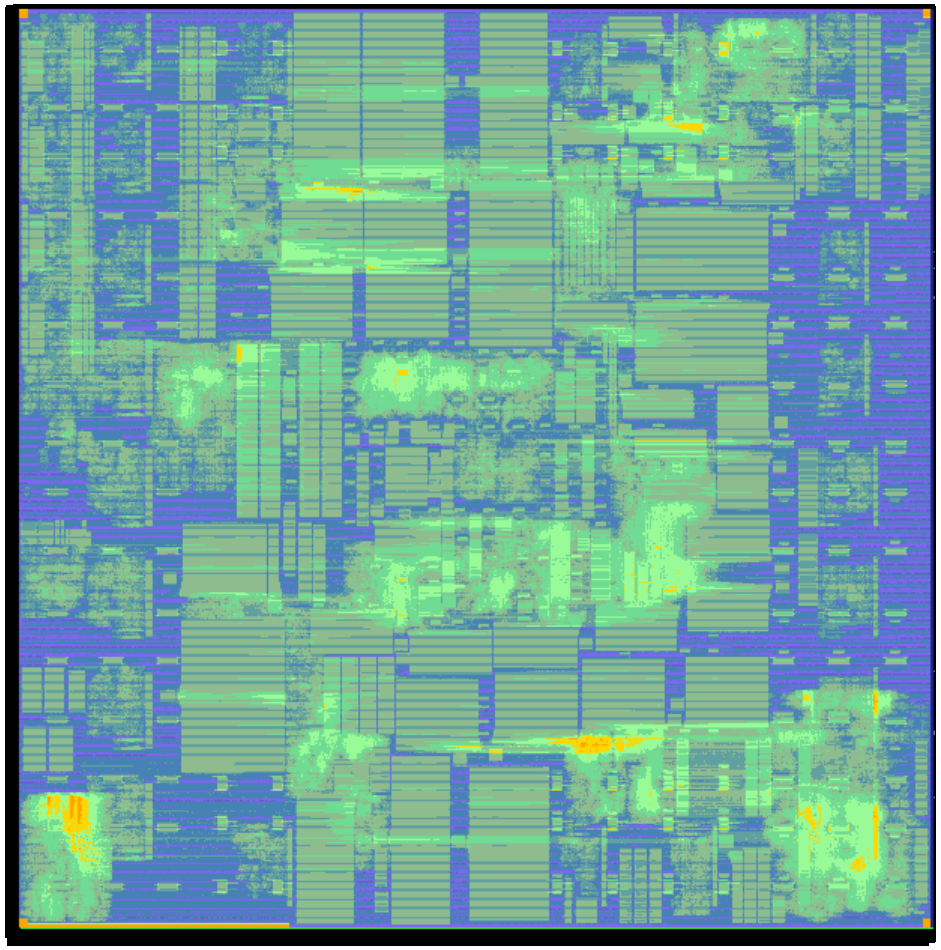
Wire up all the logic and IOs

Run timing with physical placement

Many iterations of all of the above

PHYSICAL DESIGN EXAMPLE

1) Memory placement
2) Logic placement & clocks
3) M1 routing
4) M2 routing
5) M3 routing
6) M4 routing
7) M5 routing
8) M6 routing
9) M2/M4/M6 routing
10) M1/M3/M5 routing



ASIC TAPEOUT

Criteria for ASIC Tapeout

- All functionality complete
- All verification complete
- Performance simulations meet goals
- Chip is error free from a testability perspective
- Chip meets timing under all process, temperature and voltage conditions
- Design and verification database is archived

MANUFACTURING

After the ASIC is taped out

- Masks are generated for photolithography
- ASICs are then built layer-by-layer on a silicon substrate wafer

Once the ASIC wafer is complete

- Each die is tested in wafer test
- Only good die are laser cut for packaging

Once cut die are available

- They are put in a package
- The packaged devices are then tested again

Tested packaged parts are put on system boards

- Test with other hardware and software