IP NGN Backbone Routers for the Next Decade

Josef Ungerman

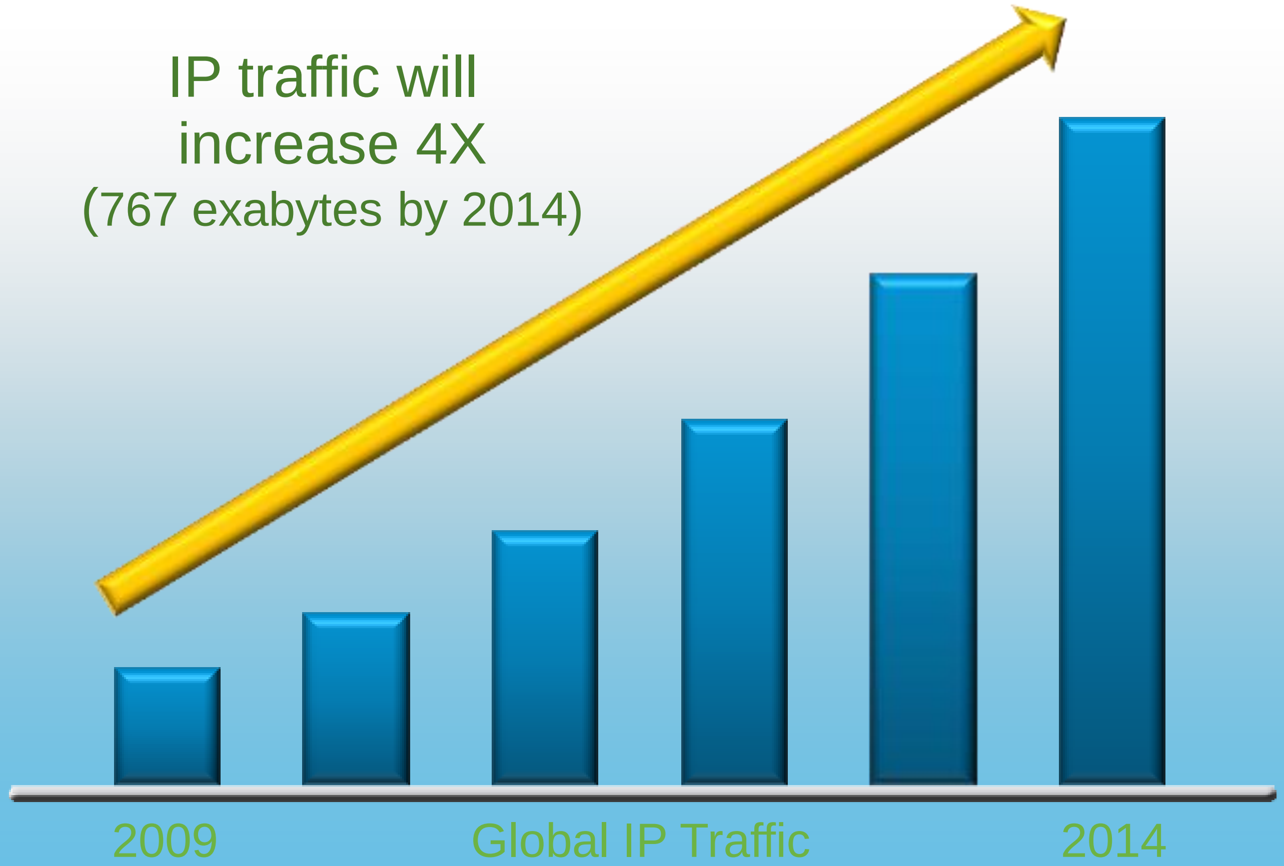
Consulting SE, CCIE #6167

Agenda

- Motivations for IP NGN
- Trends is IP/MPLS Core Design
- Router Anatomy Trends
- Latest Product Updates
- Switching Fabric Technologies
- Network Processor Technologies

SP Infrastructure Problem Definition

Exponential Growth and Evolving Traffic Mix



Source: Cisco Visual Networking Index—Forecast, 2009-2014

Video & Multimedia

90% Consumer Traffic

Mobile Internet

39X Increase in Traffic

IT Services via Cloud

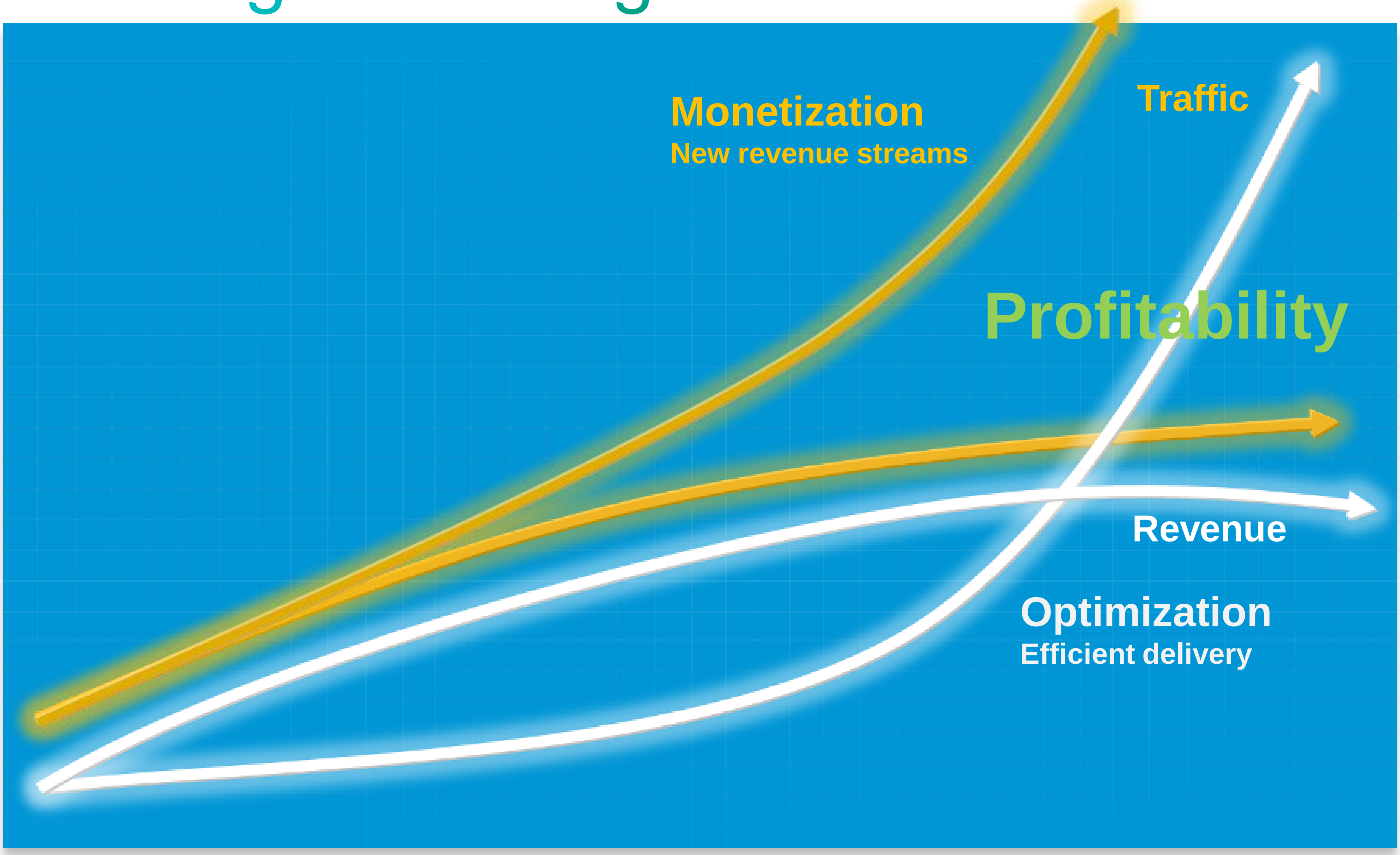
300+% Market Growth

More Issues:

- **IPv6 and IPv4 Address Exhaustion**
- **LTE moving from circuits to packets**
- **new access technologies – WiFi, FTTX**

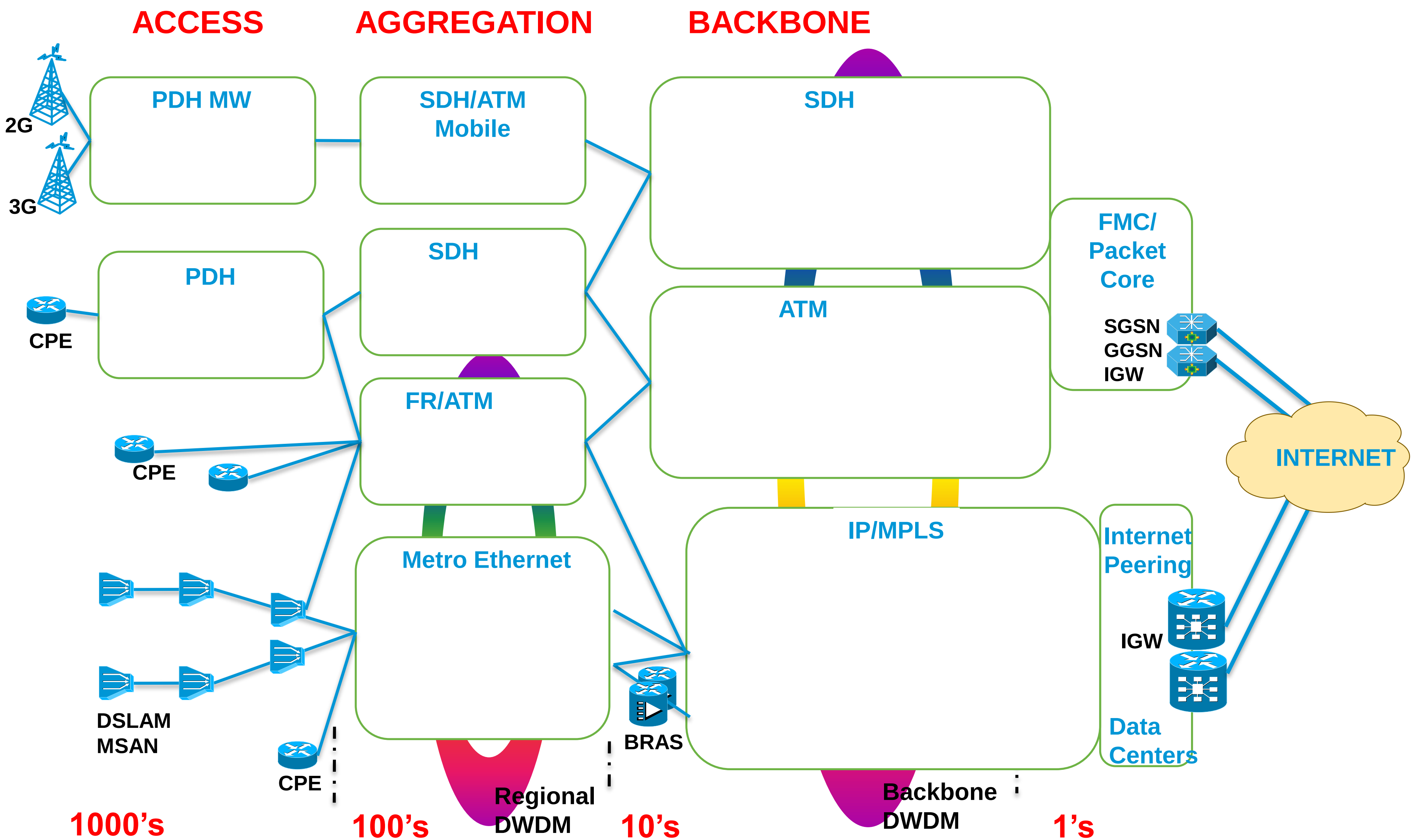
Very Different Traffic Profile

Challenge of Shifting Environment

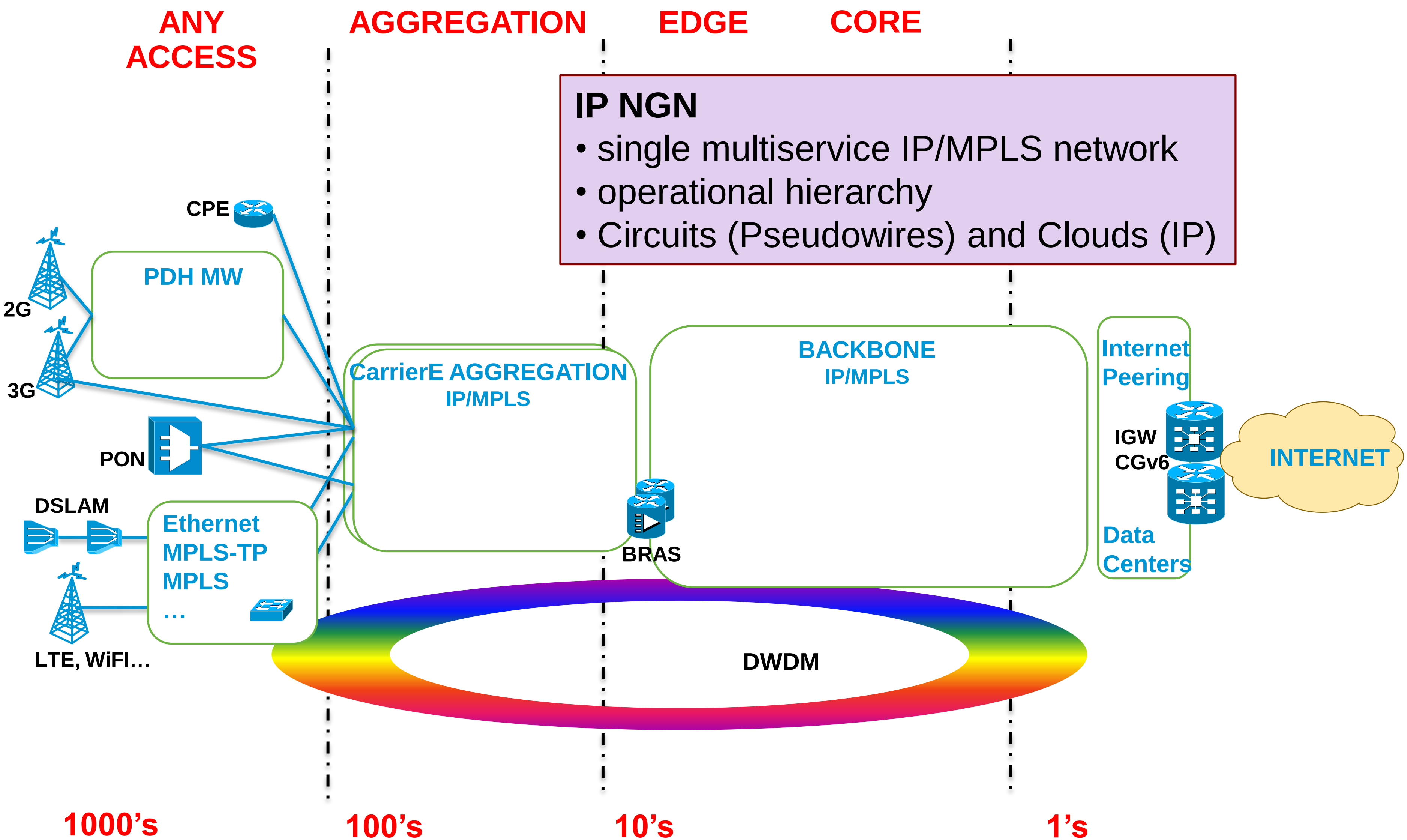


Network Architecture Trends

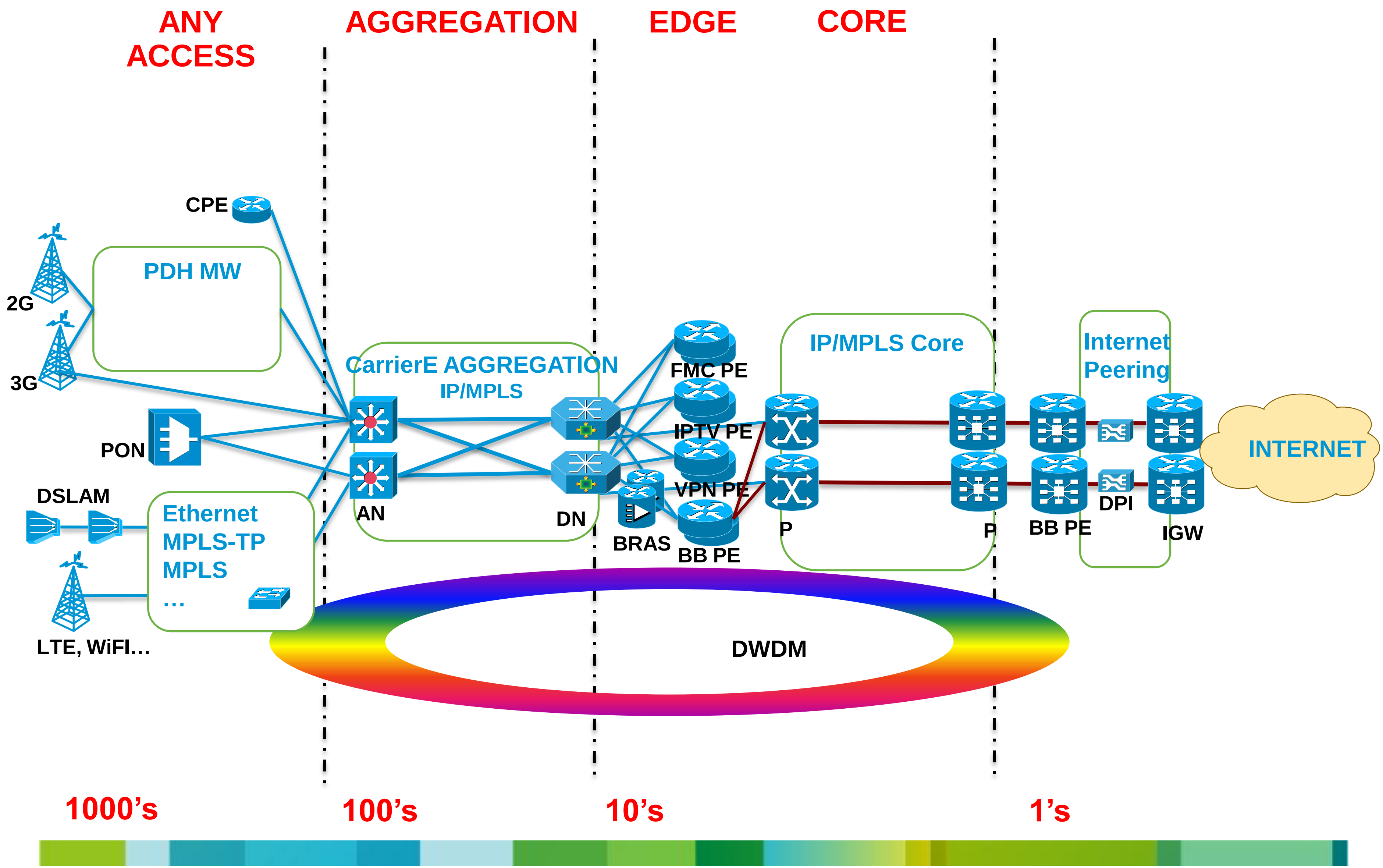
IP NGN = reducing networks and layers



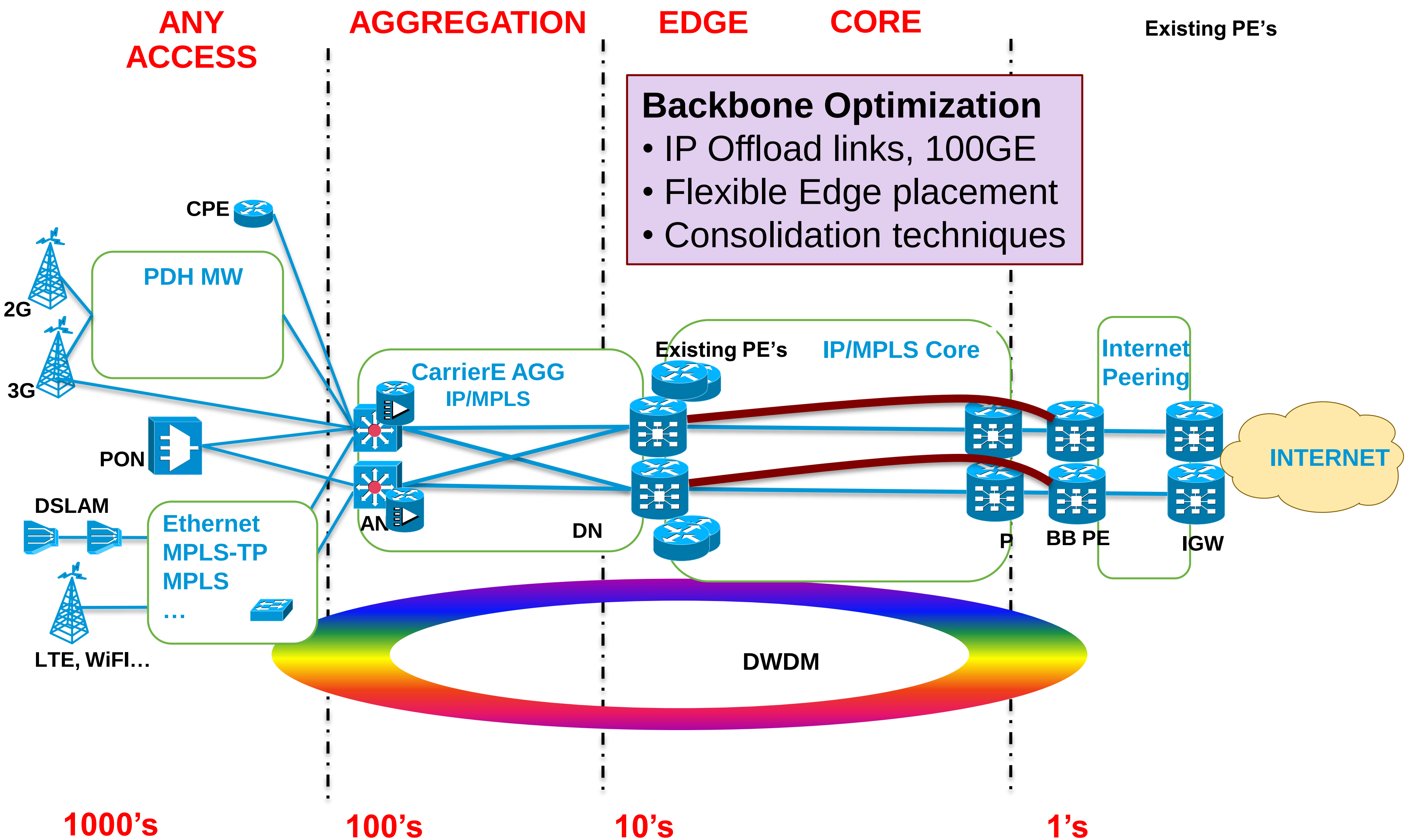
IP NGN = reducing networks and layers



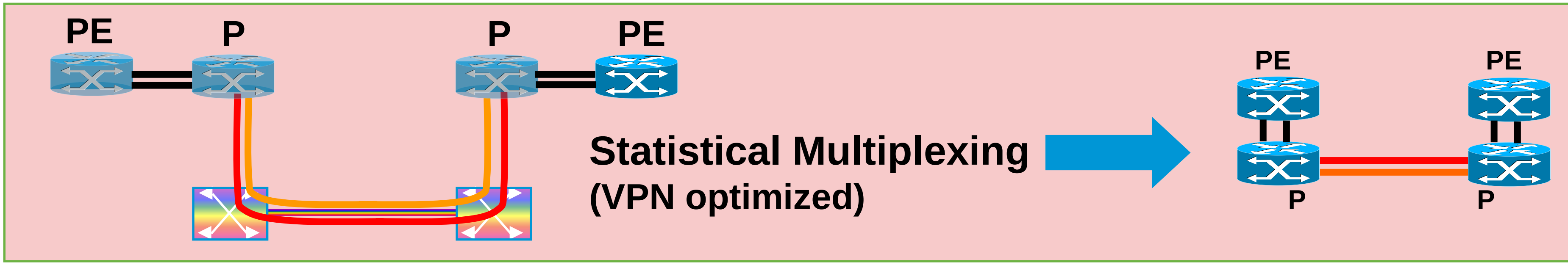
IP NGN – typical status today



IP NGN – optimization trends



Router Bypass Techniques

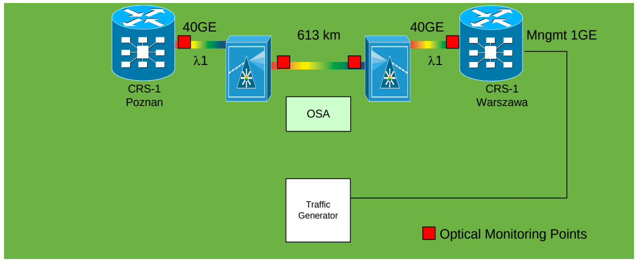


- O-E-O regeneration avoided as much as possible
no need for OTN Switching cross-connects in CEE countries

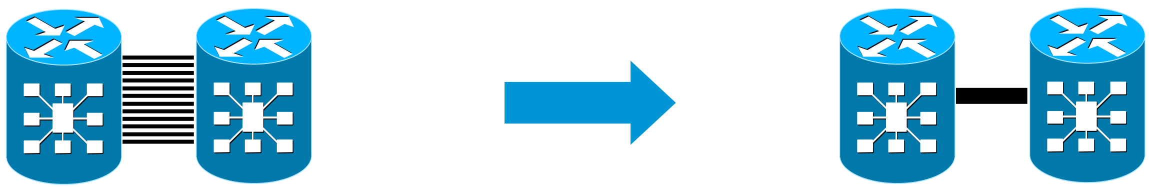
- Static long lambdas are used
no need for dynamic G.MPLS in the static Internet backbone

- Importance of OTN interfaces in routers (IPoDWDM)
STM-256 (OTU3) and 100GE (OTU4)

Real-Life Example:
Warszawa-Poznan, 613km
40G over Siemens 10G WDM



Link Consolidation – 100GE



Link Bundling (LACP)

- up to 64x TGE today (32x deployed)
- dynamic adaptable hash (also 3,5,7,11 links)
- 7-tuple hash for equal load-sharing

100GE

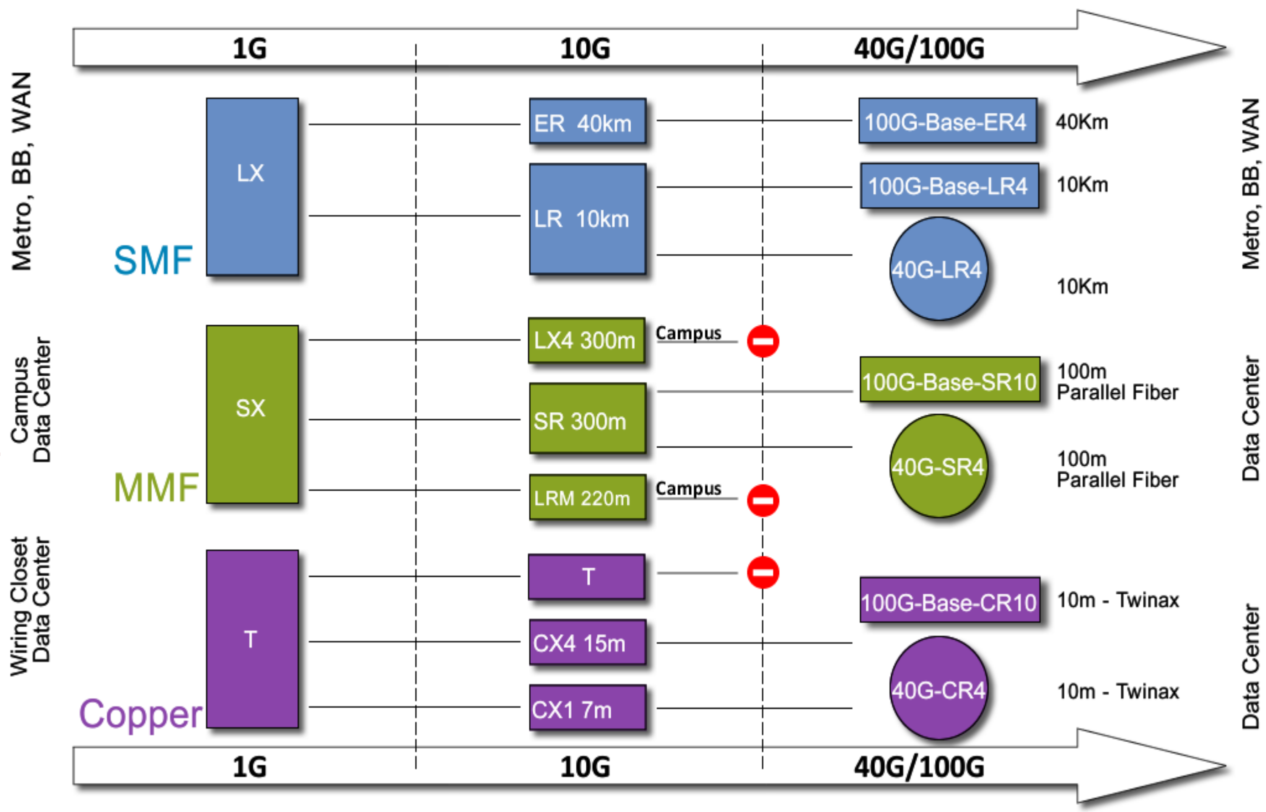
- throughput (100GE is like a bundle 12-14 TGEs)
- no hashing inefficiencies, easy troubleshooting
- contribution HDTV is 1.24Gbps single stream!

100GE

- IEEE 802.3ba, ITU OTU4
- router interface
- router → router or transponder

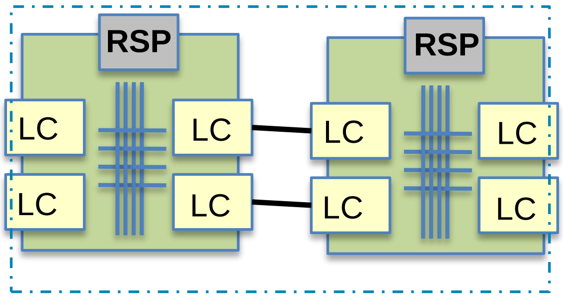
40GE

- IEEE 802.3ba, ITU OTU3e
- data center interface
- router to transponder



Node Consolidation Techniques

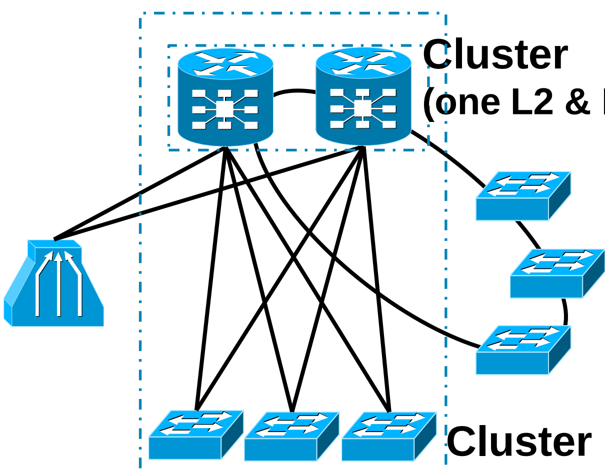
Cluster (ASR9000)



Key motivation is in the Access edge:

Simpler Access Dual-homing

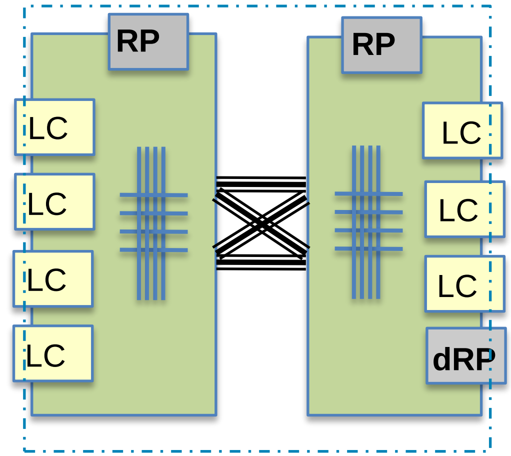
- scaling the L2/L3 control plane (not data plane)



Cluster
(one L2 & IP/MPLS control plane)

Cluster + Satellites
(remote linecards)

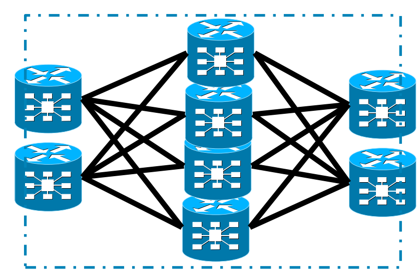
Multi-Chassis (CRS)



Key motivation is in the Core:

Simpler Core PoP

- scaling the non-blocking data plane
- back-to-back, 2+1, 8+2, etc.



Multi-Chassis
(one router)

Optimization: How to move bits cheaper...

...reduce OPEX, CAPEX, and keep reasonable quality?

1) **Reduce the number of networks**
   - IP NGN = single multiservice network

2) **Reduce the number of layers**
   - IP NGN = IP/MPLS + DWDM

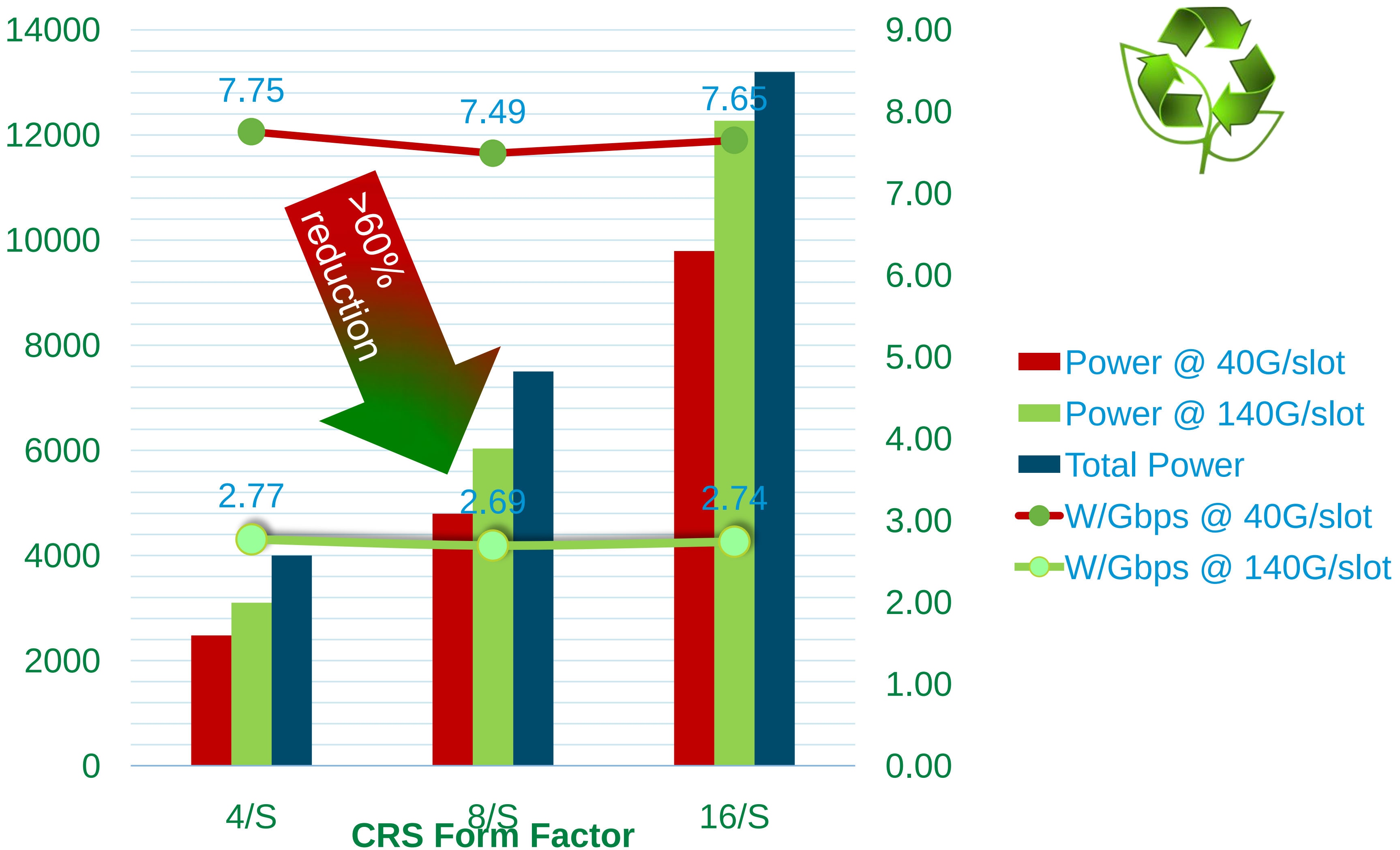
3) **Reduce the number of nodes**
   - Direct Links = huge broadband traffic takes shortest path

4) **Reduce the number of links**
   - MPLS Technology = statistical multiplex and hierarchy

5) **Innovate – make use of modern technologies**
   - Moore's Law = Lower TCO, Price/Gigabit, Watt/Gigabit

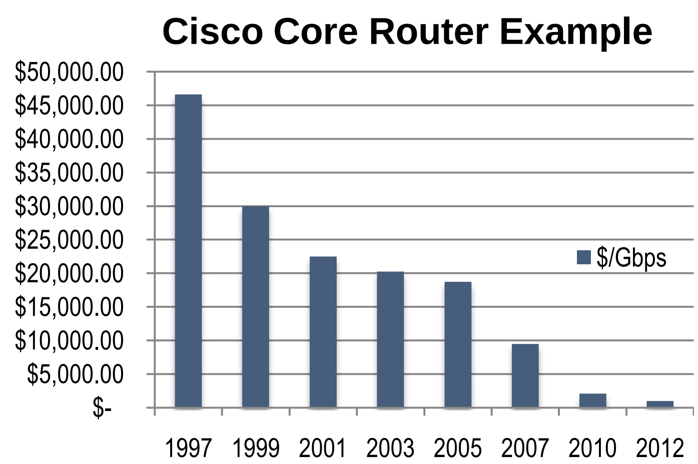
Core Trends – Appeal of Innovations

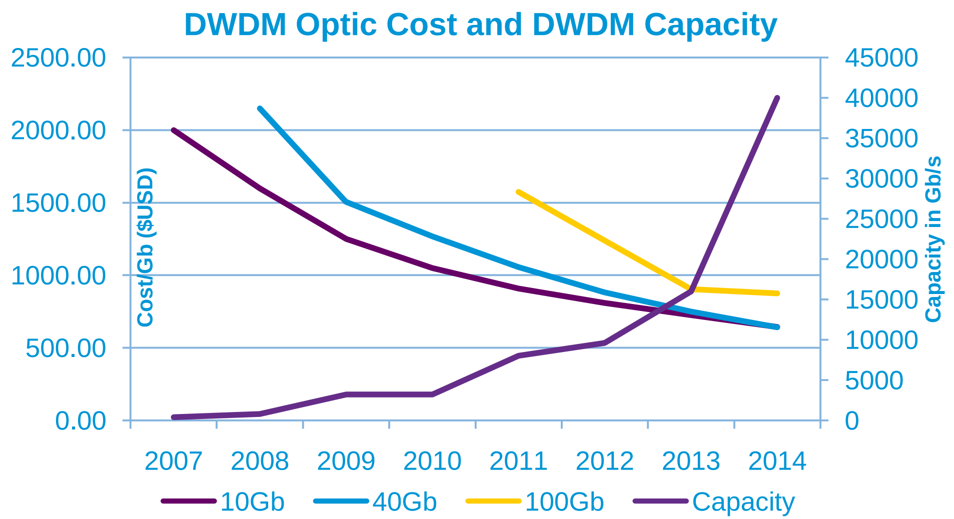
CRS-1 (2005) vs. CRS-3 (2010)



Core Trends – Appeal of Innovations

Routers: 23% Cumulative Average $/Gbps Drop per year / fewer ASICs
Optics: $/G stays flat (best case) or <u>increases</u> from one technology to the next

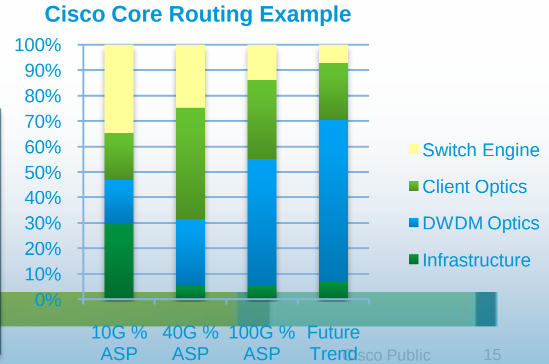




INFONETICS RESEARCH

10G/40G/100G Networking Ports Biannual Worldwide and Regional Market Size and Forecasts
May 2010

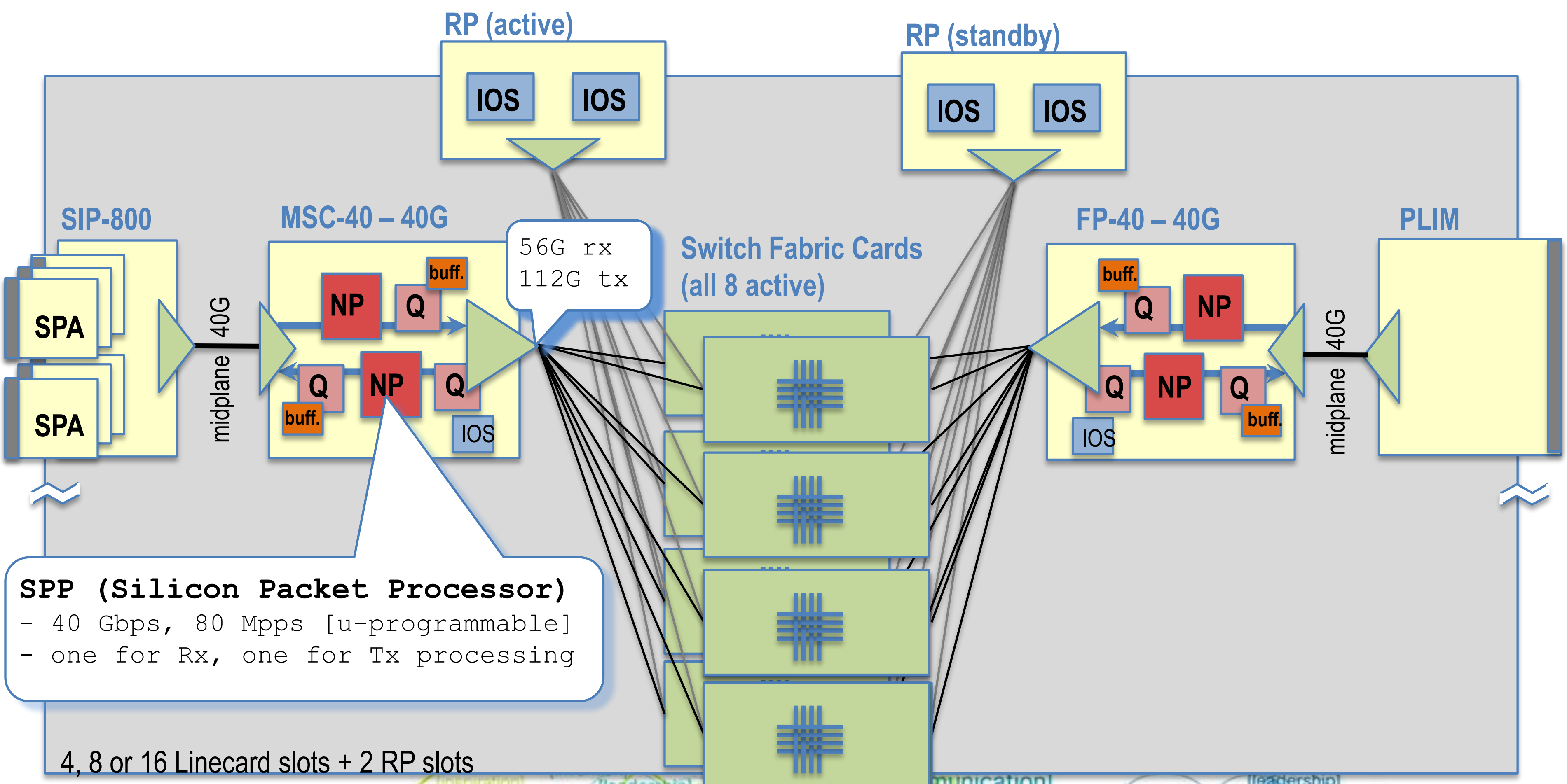
- Silicon has fundamentally followed Moore's law
- Optics is fundamentally an analog problem



Router Anatomy Trends

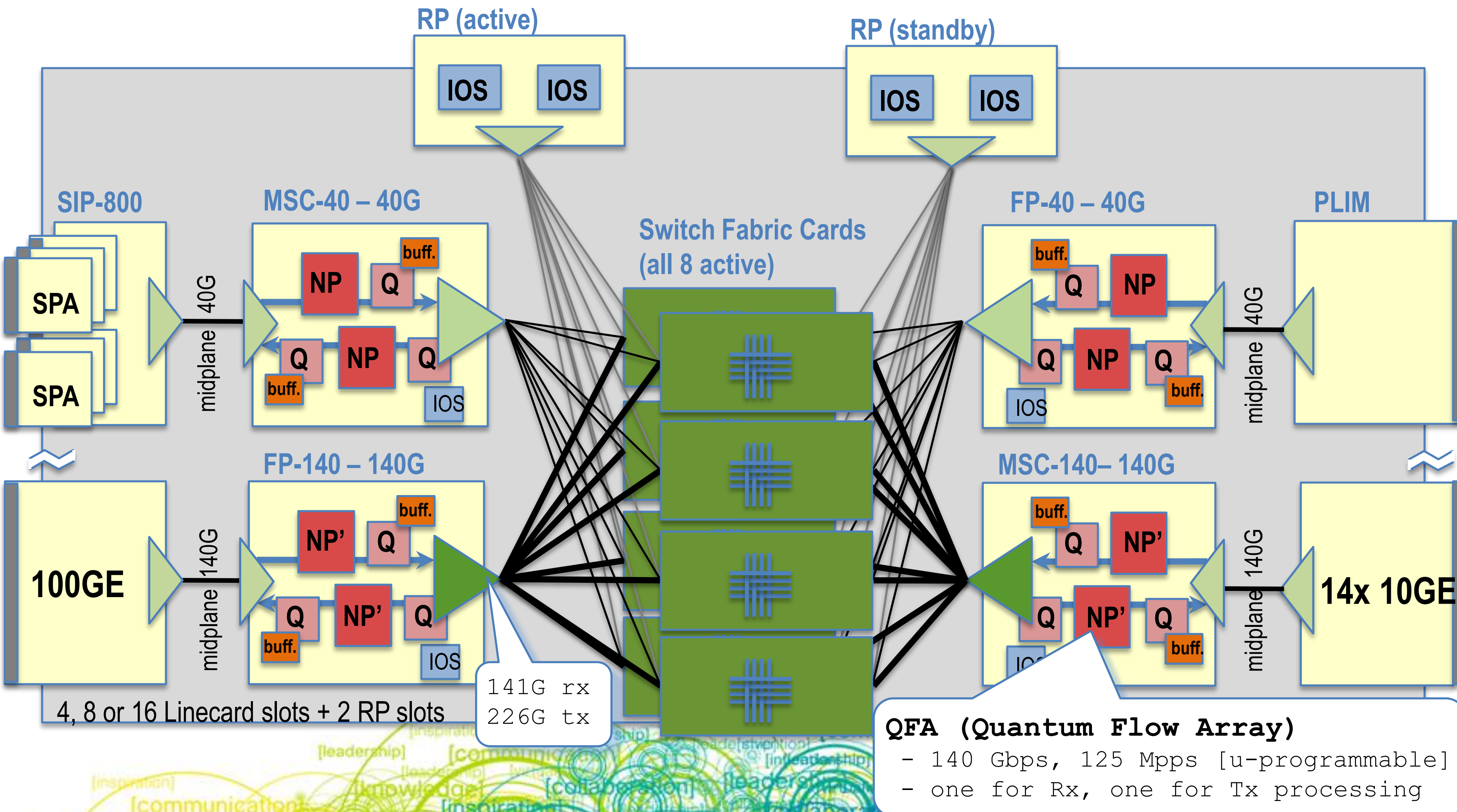
2004: Cisco CRS-1 – 40G (STM-256) per slot

Focus on Quality (scale, modularity, resiliency)



2010: Cisco CRS-3 – 140G per slot

Focus on Quality (scale, modularity, resiliency)

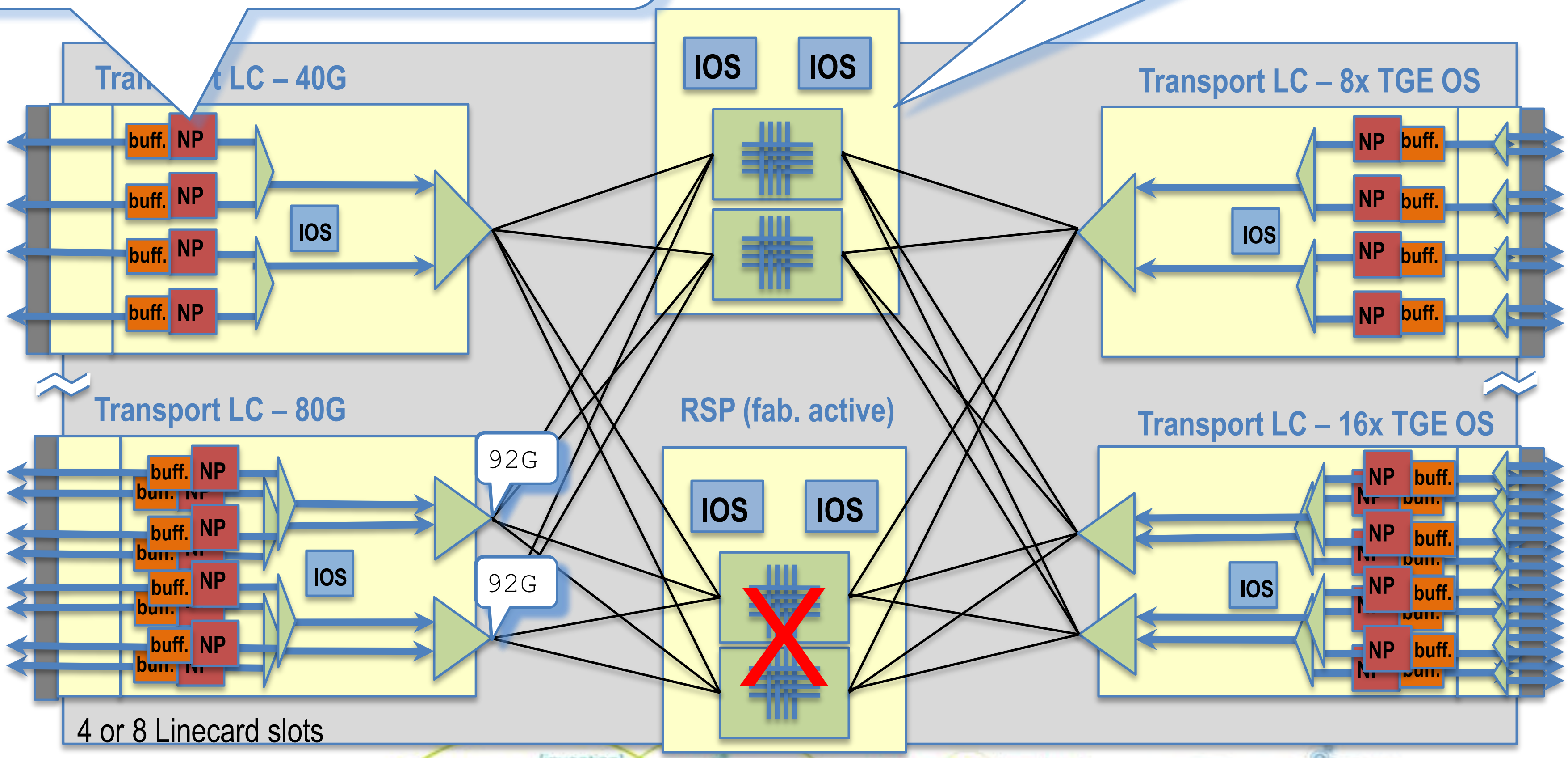


2009: Cisco ASR9000 – 8x 10GE per slot

Compact Router/Switch

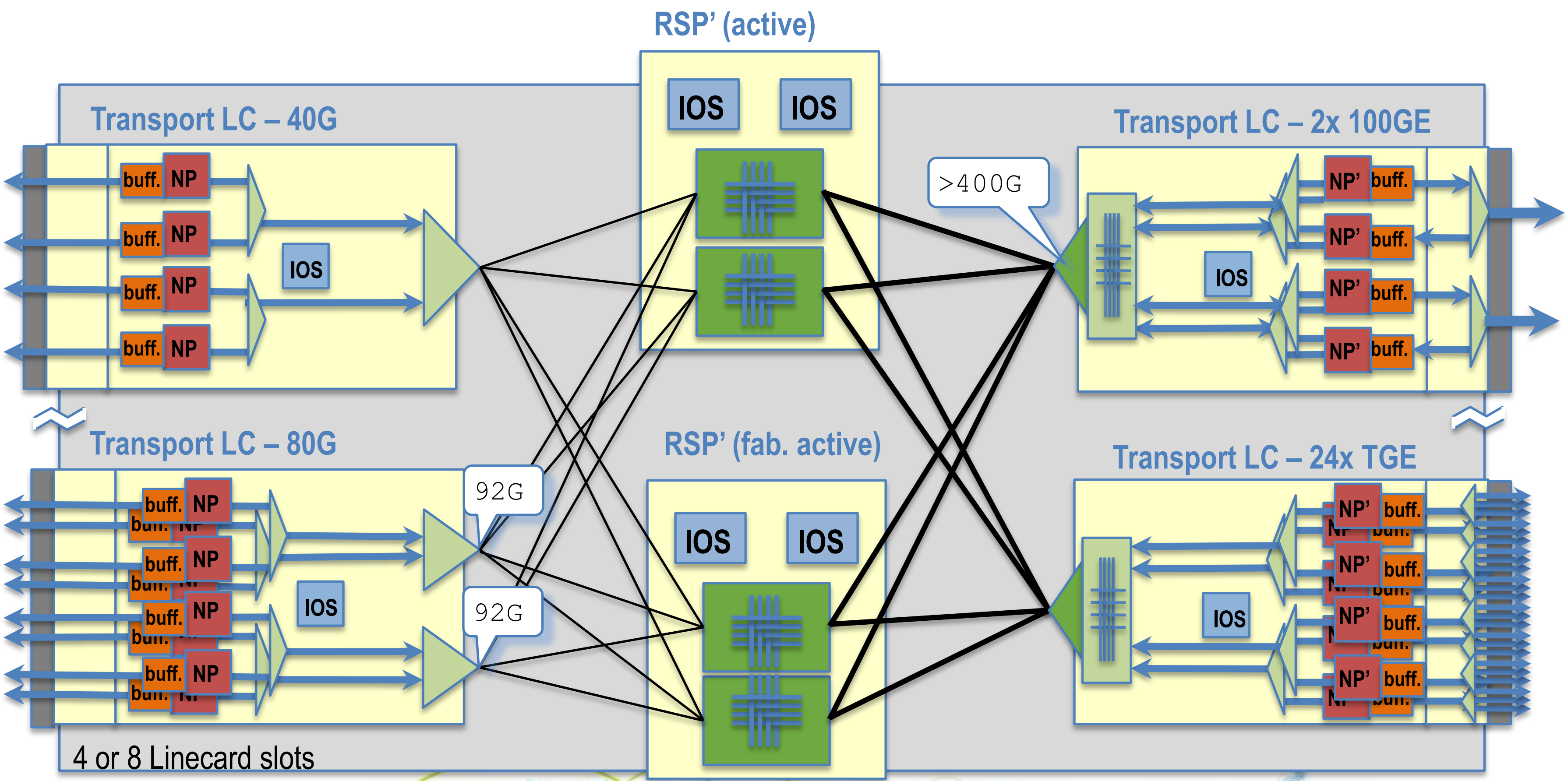
Trident Network Processor
- 30 Gbps, 28 Mpps [u-programmable]
- shared for Rx and Tx processing
- one per 10GE (up to 8 per card)

RSP (Route/Switch Processor)
- CPU + Switch Fabric
- active/active fabric elements



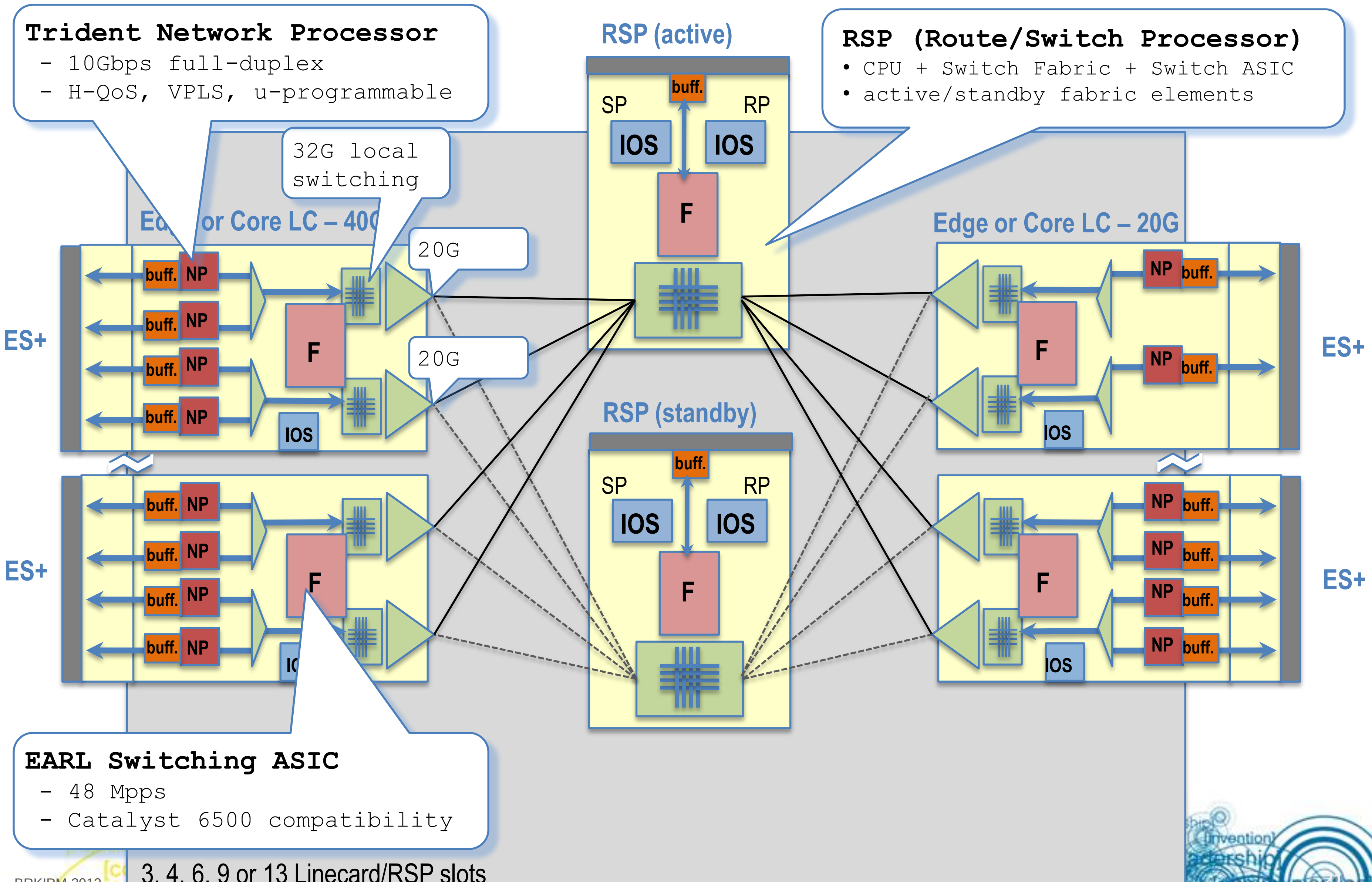
2011: Cisco ASR9000 – 2x 100GE per slot

Compact Router/Switch



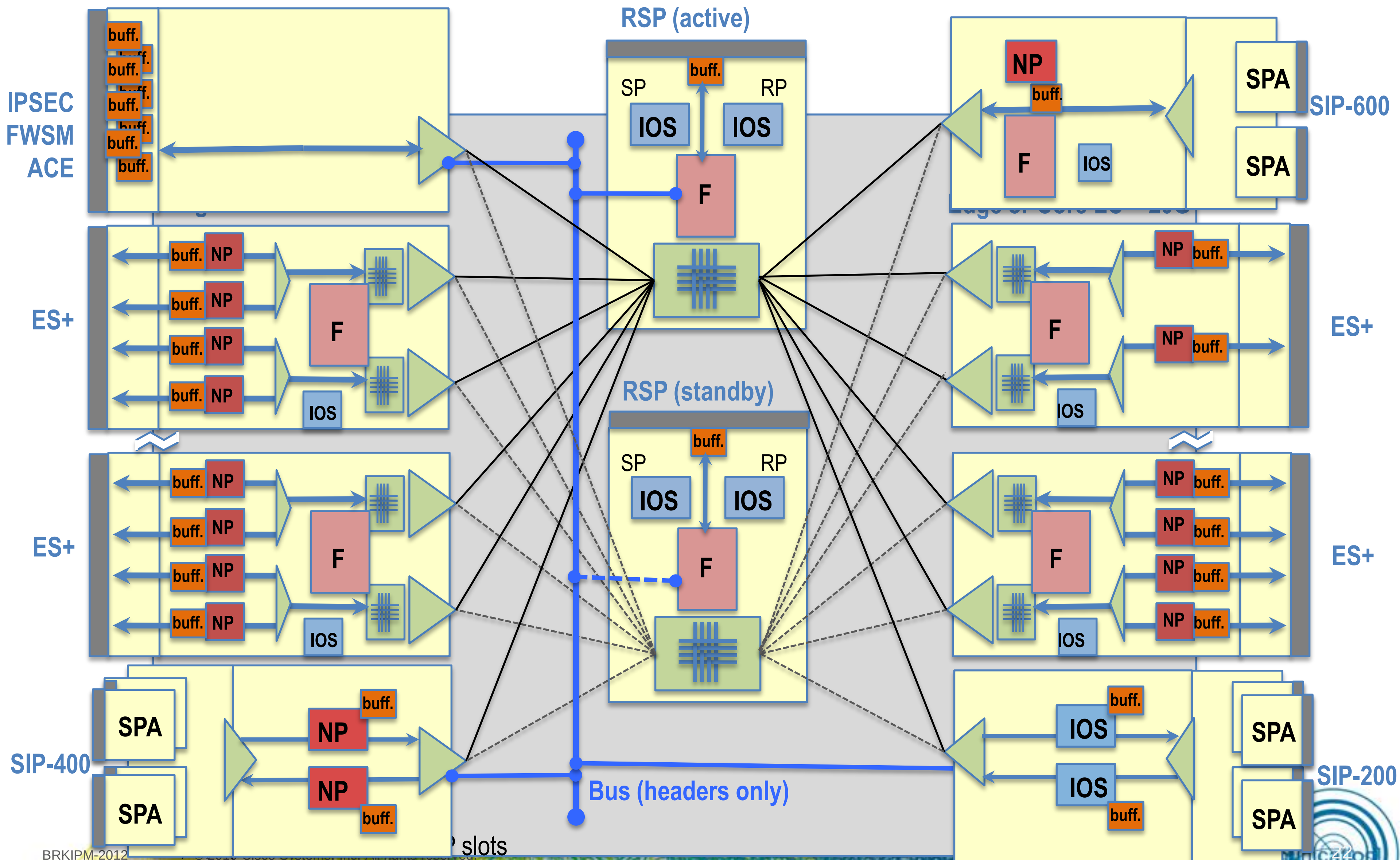
2003: Cisco 7600 – 4x 10GE per slot

The Switch/Router



2003: Cisco 7600 – 4x 10GE per slot

The Switch/Router



How to make a router cheaper...

...and keep a reasonable quality?

1) Compact Anatomy
   - RSP, Route/Switch Processor (instead of RP and FC)
   - Ethernet-oriented Linecard (non-modular, less memory)

2) Linecard Architecture
   - Multiple smaller NP's (eg. 4x 10G instead of 1x 40G)
   - One NP is shared for Rx and Tx (not dedicated NP's per Rx and Tx)
   - Multiple smaller Fabric Ports (eg. 2x 20G instead of 1x 40G)

3) Special Core-facing Linecards
   - 8/16 queues per port (instead of thousands)
   - lower-scale NP (no need for thousands of interfaces)
   - licenses for features that not everybody uses (eg. VPN, OTN, scale)

4) Oversubscribed Cards
   - 2:1 ingress overbooking (eg. PON OLT Aggregation)

IP NGN Routers Update

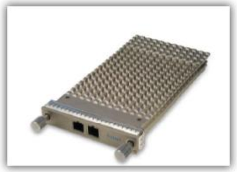
Cisco CRS
Core

~7500 systems
~450 customers
(~80 are CRS-3)

		CRS-4/S	CRS-8/S	CRS-16/S	CRS-MC (8+2)
Chassis	# of Slots Height	4 (+2 RP) 1/3 rack	8 (+2 RP) ½ rack	16 (+2 RP) 1 rack	128 (+2 RP) 10 racks
2004 (CRS-1)	Linecard [Gbps] System [Tbps]	40 .320	40 .640	40 1.28	40 10.24
2010 (CRS-3)	Linecard [Gbps] System [Tbps]	140/105 1.12	140/105 2.24	140/122 4.48	140/122 35.84
Future					

Maximum BW / Protected BW with a failed fabric element

CRS-3
Interface Modules (PLIMs)



1x 100GBE

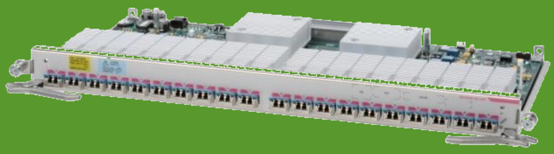
- Line-rate performance (100Gbps)
- CFP optics (LR4, 10km)

14x 10GBE-WL-XFP

- Line-rate performance (140Gbps)
- Configurable LAN/WAN PHY

20x 10GBE-WL-XFP

- Oversubscribed (140Gbps)
- Configurable LAN/WAN PHY



Each PLIM requires FP140 or other forwarding card

CRS-3 and CRS-1
Forwarding Cards

MSC-40 – High-speed edge @ 40Gbps

- H-QoS (8,000 queues), 800 interfaces, WAN

FP-40 – IP/MPLS Core &Peering @ 40Gbps

- Per-port QoS, IP/MPLS, ACL, Netflow…

MSC-140 – High-speed edge

- H-QoS (64,000 queues), scale (12,000 vlans)

FP-140 – IP/MPLS Core & Peering

- Per-port QoS, IP/MPLS, ACL, Netflow…

LSP-140 – MPLS Core P

- Per-port QoS, MPLS, IP Multicast, limited IP

IPv6 Transition: CGSE

Carrier-Grade Services Engine PLIM

Introducing the new engine for **massive** Cisco CGv6 deployments (XR 3.9.1)

Cisco CGSE

- **20+ million** active translations
- **100s of thousands** of subscribers
- **1+ million** connections per second
- **20Gb/s** of throughput
- **XML API** (eg. port-forwarding)
- **Netflow V9** translation logging
- Security

IPv6 Transition solution feature set

- Carrier-Grade NAT44 (3.9.1)
- NAT64 stateless (3.9.3)
- 6rd BR (3.9.3)
- NAT64 stateful (4.1.2)
- DS-Lite, 4rd, dIVI – planned

now: Cisco CRS
2011: XR12K, ASR9K

Cisco ASR9000
Edge and Aggregation

~3500 systems
~500 customers




		ASR-9006	ASR-9010
Chassis	# of Slots	4 (+2 RSP)	8 (+2 RSP)
	Height	¼ rack	½ rack
2009	Linecard [Gbps]	120/80	120/80
	System [Tbps]	.960	1.92
2012	Linecard [Gbps]	240	240
	System [Tbps]	1.92	3.84
Future			

Maximum BW / Protected BW with a failed fabric element

ASR 9000 Line Cards

ASR 9000	#10GE LR	#10GE OS
6-slot	32	64
10-slot	64	128

Fixed Ethernet LCs:

- **Line Rate: 40xGE, 2x10GE+20xGE, 4x10GE, 8x10GE**
- **Oversubscribed: 8x10GE (60G), 16x10GE (90G/120G)**

Ingress/Egress H-QoS, Netflow, IPoDWDM (G.709, FEC, XFP), Video monitoring, SyncE, E-OAM

L2 Scalability: 1MMACs, 8kBDs, 32kPWs
L3 Scalability:1M routes, 4kVRFs, 4kL3intfs

3 LC versions (16x10GE OS "Medium Queue" only):

Line Card		EFPs	Egress Queues	Policers	Buffering
Low Queue	-L	4k	8/port	8k	50ms
Medium Queue	-B	16k	64k	128k	50ms
High Queue	-E	32k	256k	256k	150ms

+licenses: L3VPN (/LC), G.709 (/LC), vidmon (/chassis)

Modular LCs: SIP-700 + max 4x SPA

- **QFP based**
- **20Gbps Full Duplex**
- **Ph1: ChOC12**

SIP-700 + 2x SPA 2-port ChOC12

How to make a router cheaper...

...and keep a reasonable quality?

1) Compact Anatomy
   - RSP, Route/Switch Processor (instead of RP and F…
   - Ethernet-oriented Linecard (non-modular, less…

2) Linecard Architecture
   - Multiple smaller NP's (eg. 4x 10G … 1x 40G)
   - One NP is shared for Rx and … dicated NP's per Rx and Tx)
   - Multiple smaller Fabric … x 20G instead of 1x 40G)

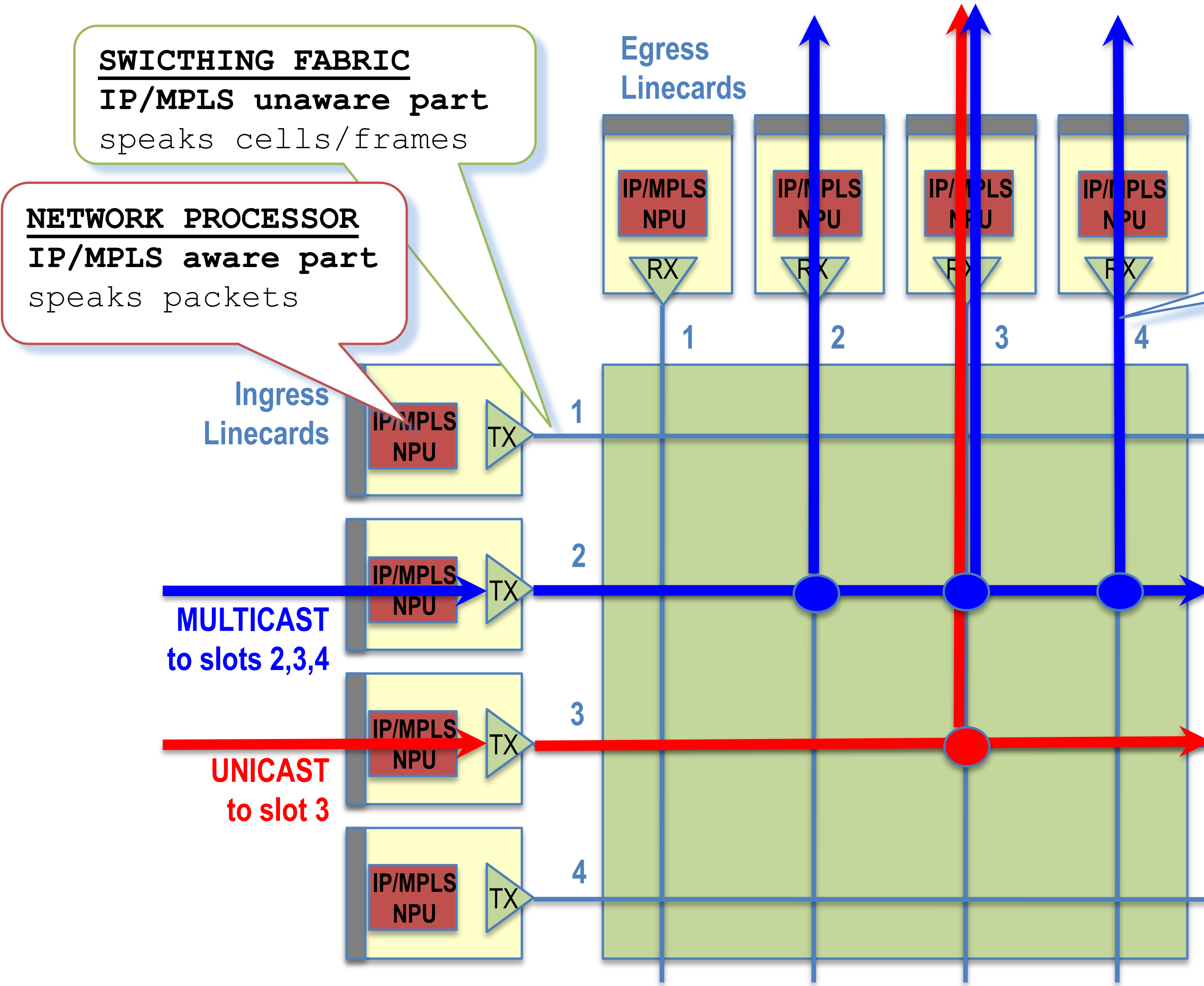
3) Special Core-faci… ds
   - 8/16 queu… (instead of thousands)
   - lower … o need for thousands of interfaces)
   - lice… eatures that not everybody uses (eg. VPN, OTN, scale)

4) Oversubscribed Cards
   - 2:1 ingress overbooking (eg. PON OLT Aggregation)

How to define QUALITY?

Architecture Matters!

Forwarding Architecture 101



Fabric Port
- addressable entity
- singleduplex pipe

What's the capacity?

4 fab.ports @10Gbps

non-blocking

ENGINEERING:

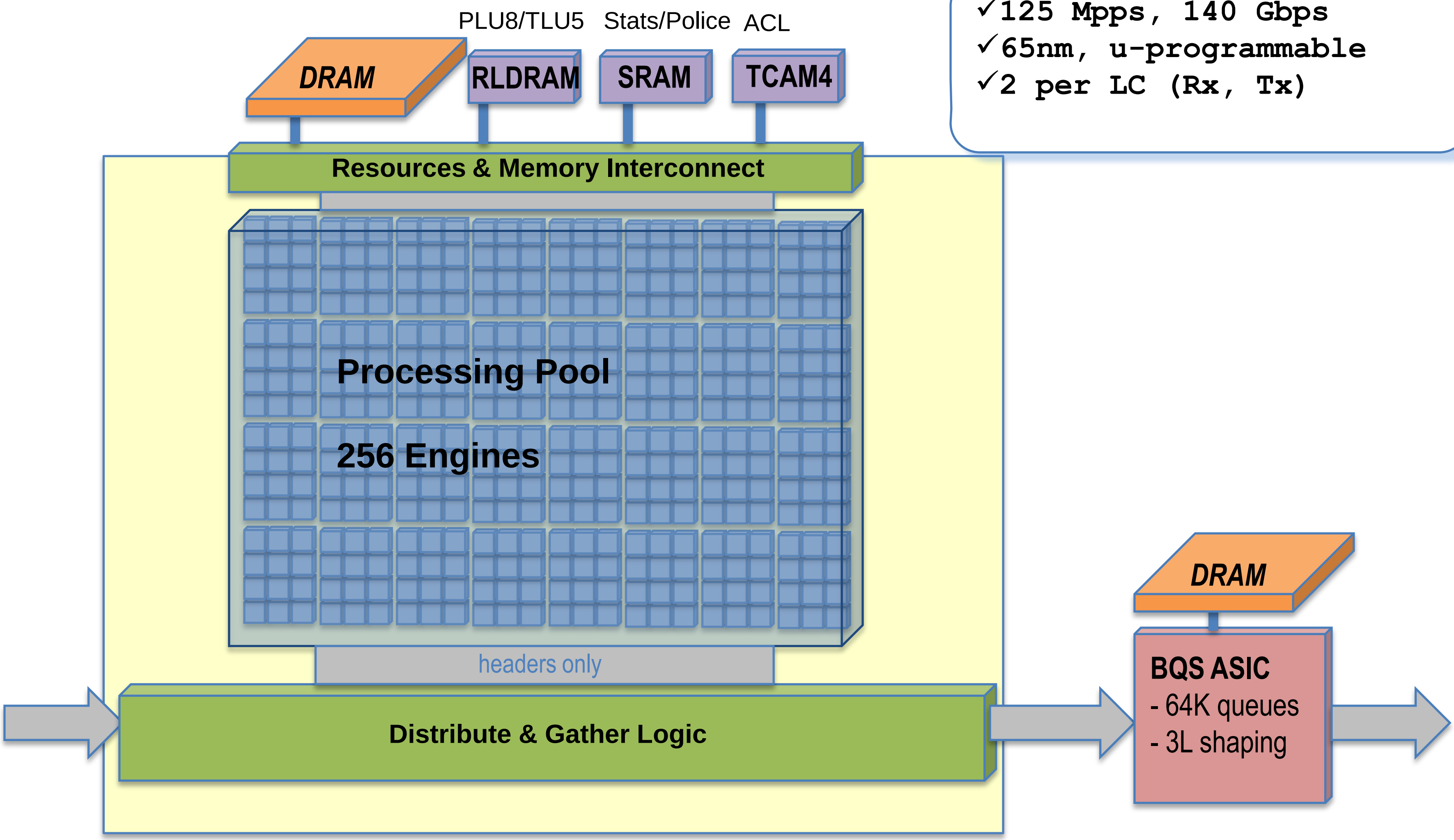
4 * 10 = 40Gbps fdx

MARKETING:

4 * 10 *2 = 80Gbps

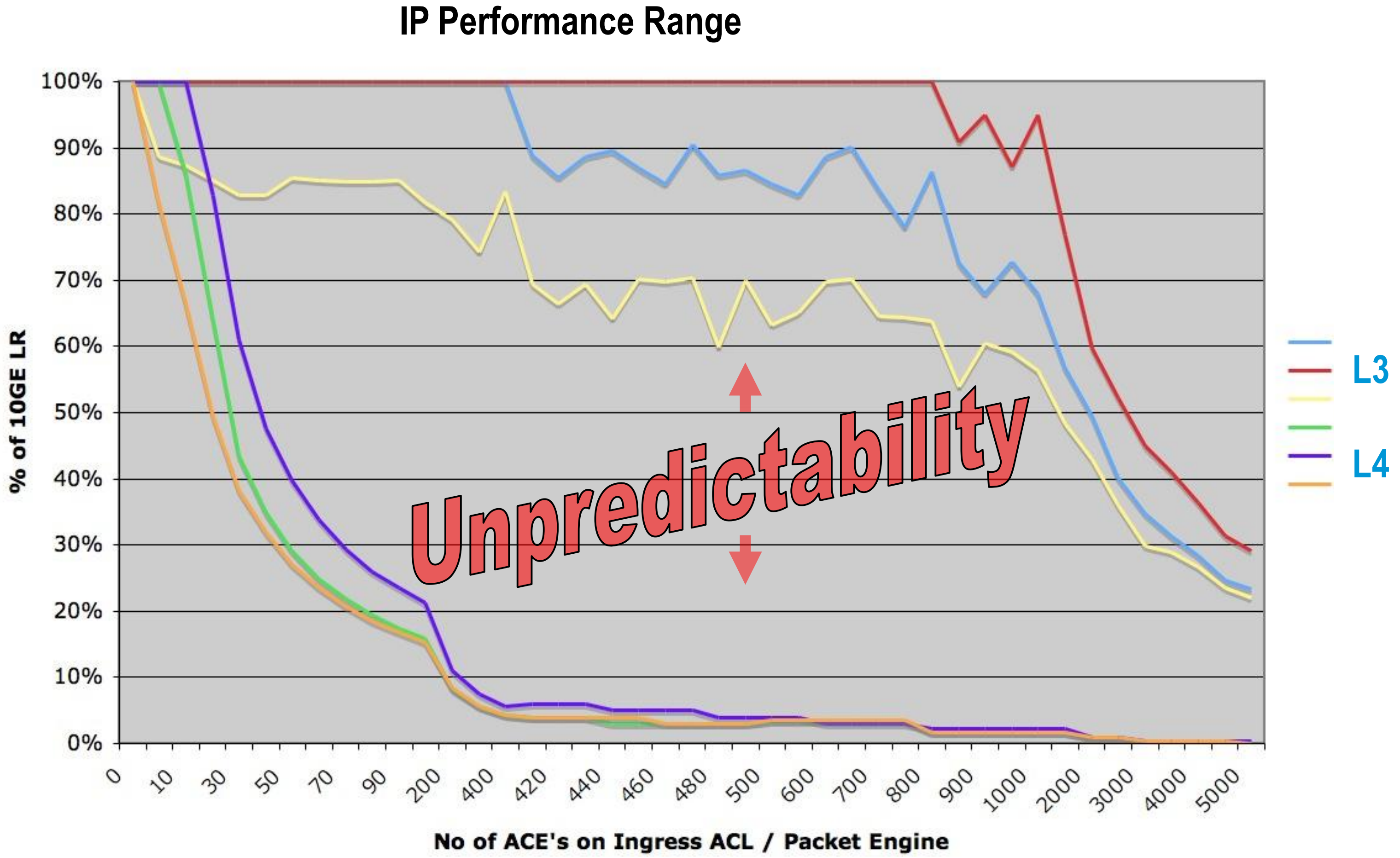
SMP Network Processor Example

2010: CRS QFA (Quantum Flow Array)



- ✓125 Mpps, 140 Gbps
- ✓65nm, u-programmable
- ✓2 per LC (Rx, Tx)

Bad Network Processor Example (non-Cisco) ACL performance impact

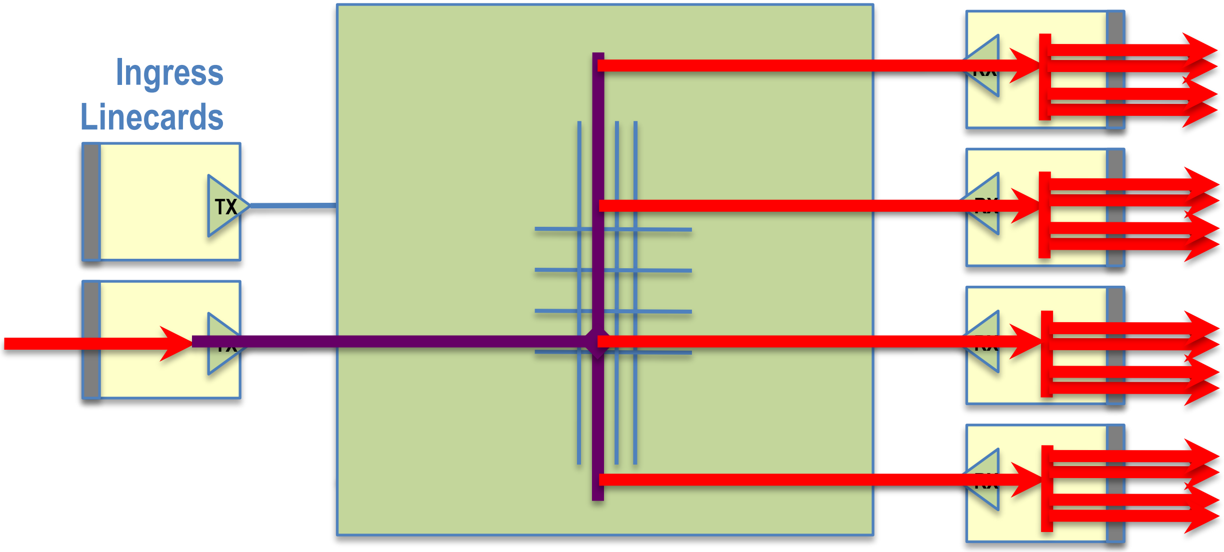


Vendor is Saving on Memory – ACL memory is shared with Route memory

- Effect #1: ACL drastically impacts forwarding performance
- Effect #2: FIB cannot be hierarchical → slow BGP convergence

Switching Fabric and Multicast

Good vs. Bad IPTV Experience

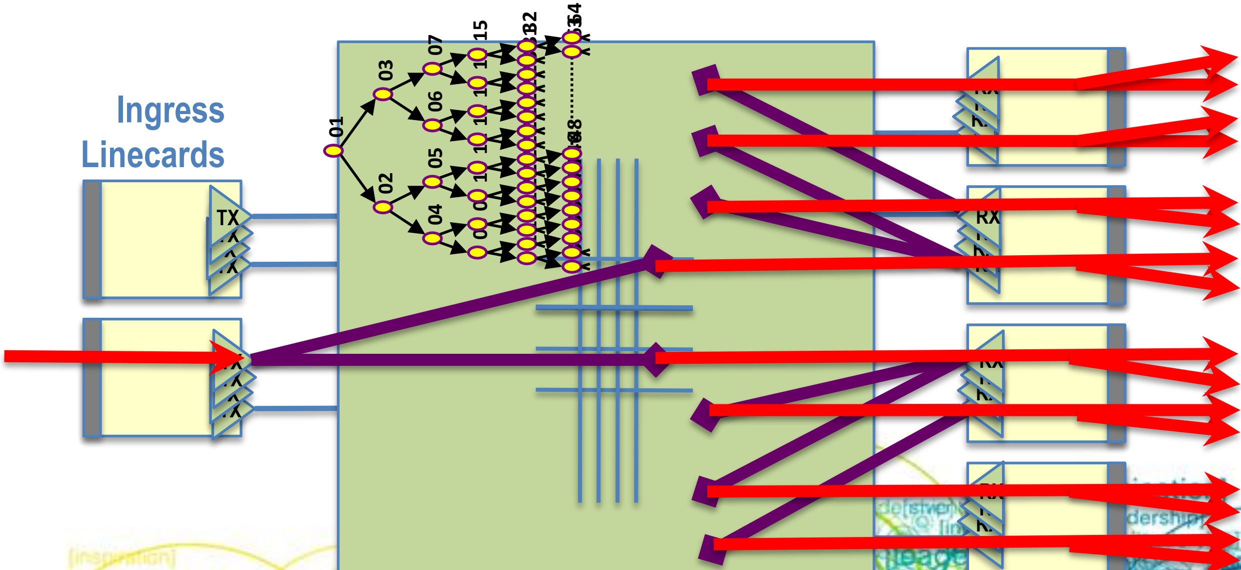


<u>Good:</u>

Egress Replication

- Cisco CRS, 12000
- Cisco ASR9K, 7600

10Gbps of multicast eats 10Gbps fabric bw!



<u>Bad:</u> Binary Ingress Replication

- dumb switch fabric
- non-Cisco

10Gbps of multicast eats 80Gbps fabric bw! (10G multicast impossible)

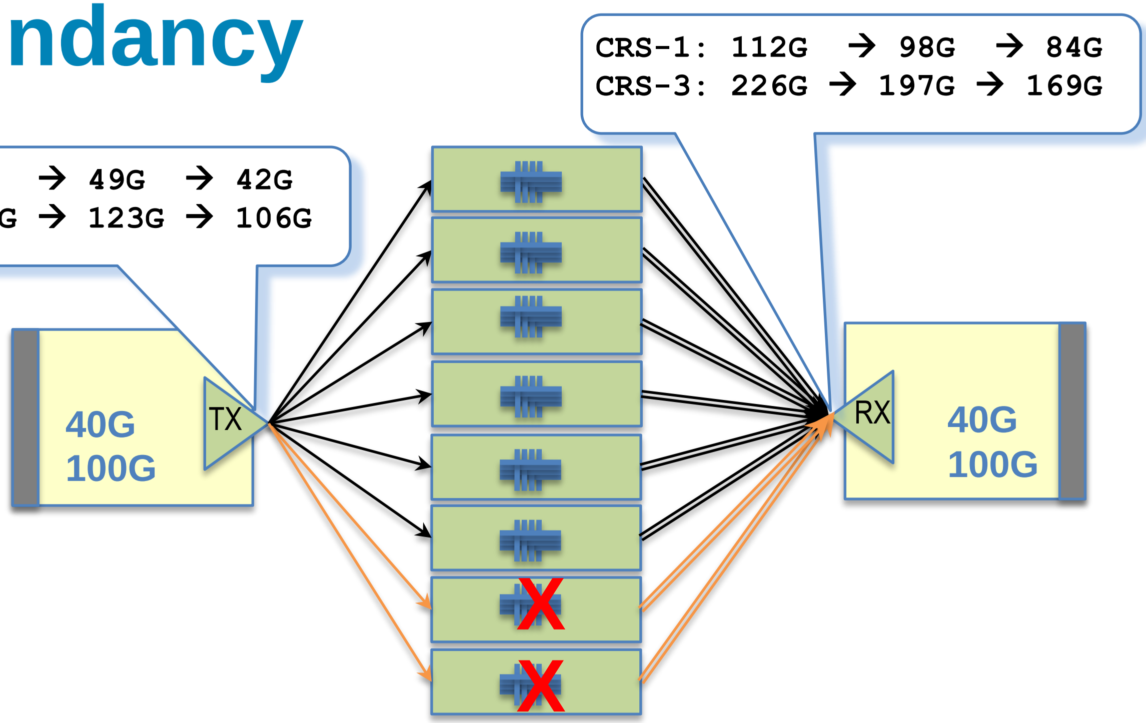
Good Fabric Redundancy

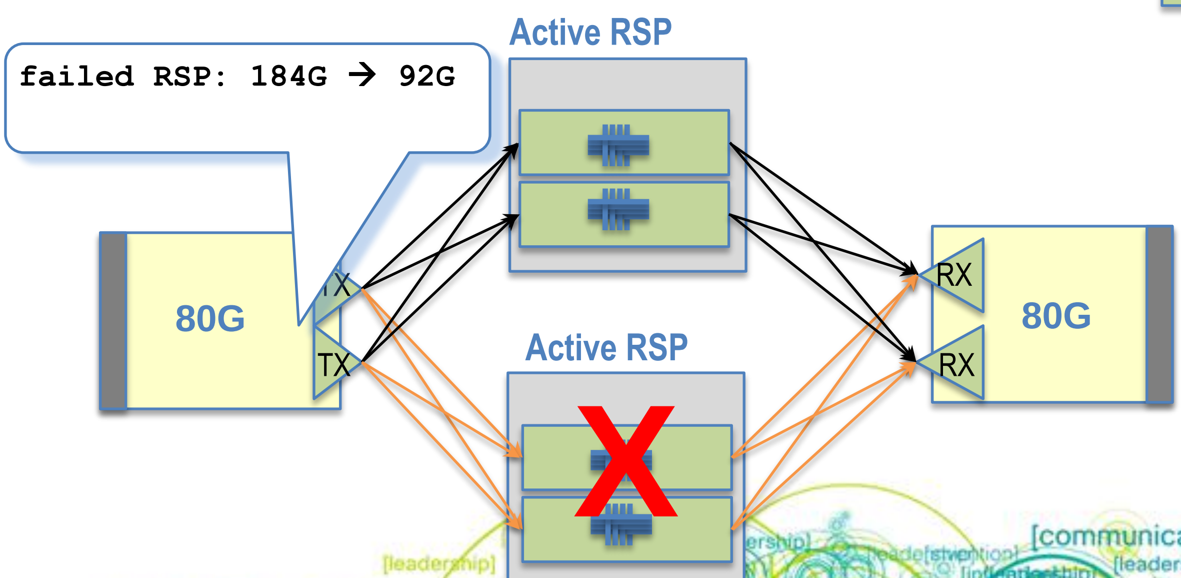
CRS-1

- 40G eth. non-blocking with 1 or 2 failed fabric cards

CRS-3

- 100G eth. non-blocking with 1 or 2 failed fabric cards





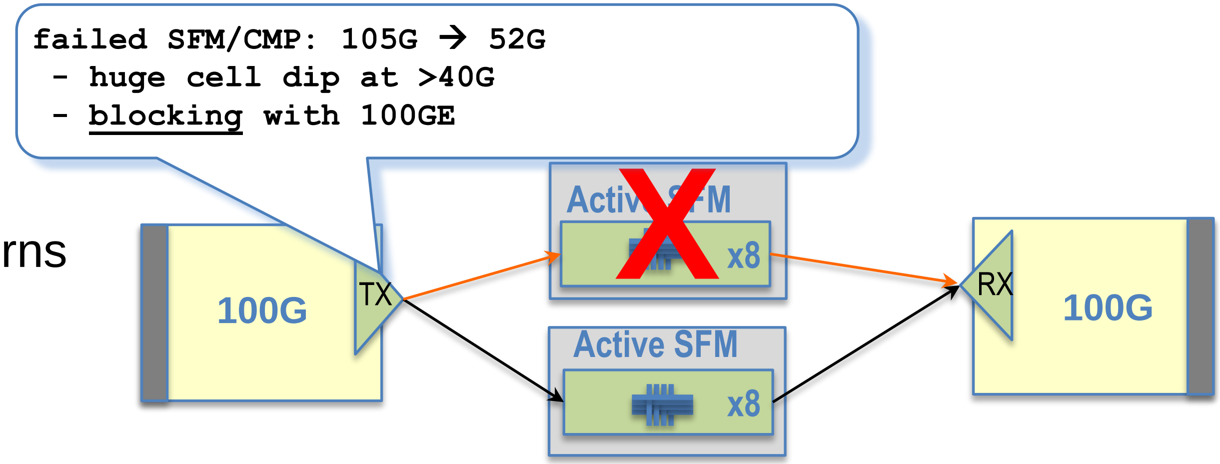
ASR9000

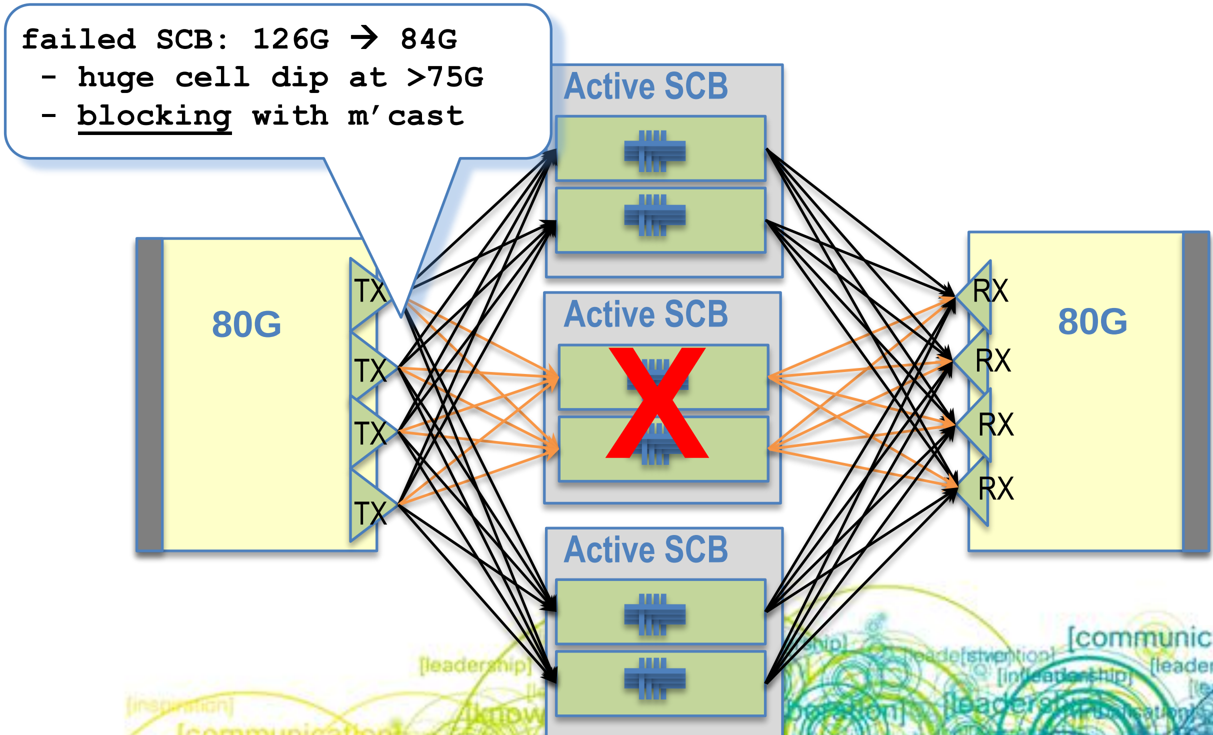
- 80G eth. non-blocking with failed RSP
- cell dip is not an issue (super-frame based fabric, not cell based)

Bad Fabric Redundancy

Non-Cisco

- 40G system sold as 100G
- if one CMP fails, system turns blocking
- unreliable 100GE





Non-Cisco

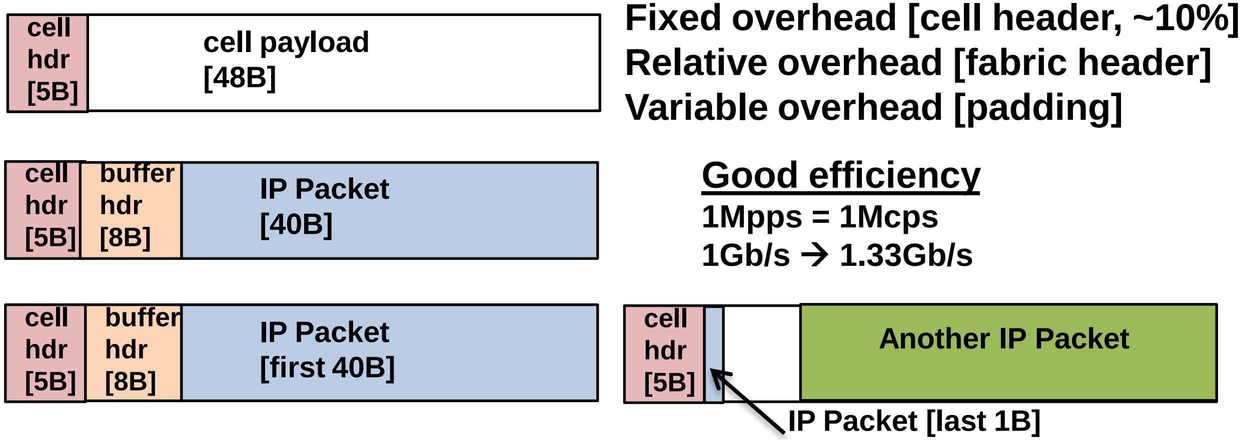
- 70G system sold as 80G
- minimum speedup
- huge cell dip
- multicast impacts unicast

Cell dip and Speedup

cell format example:

40B IP Packet:

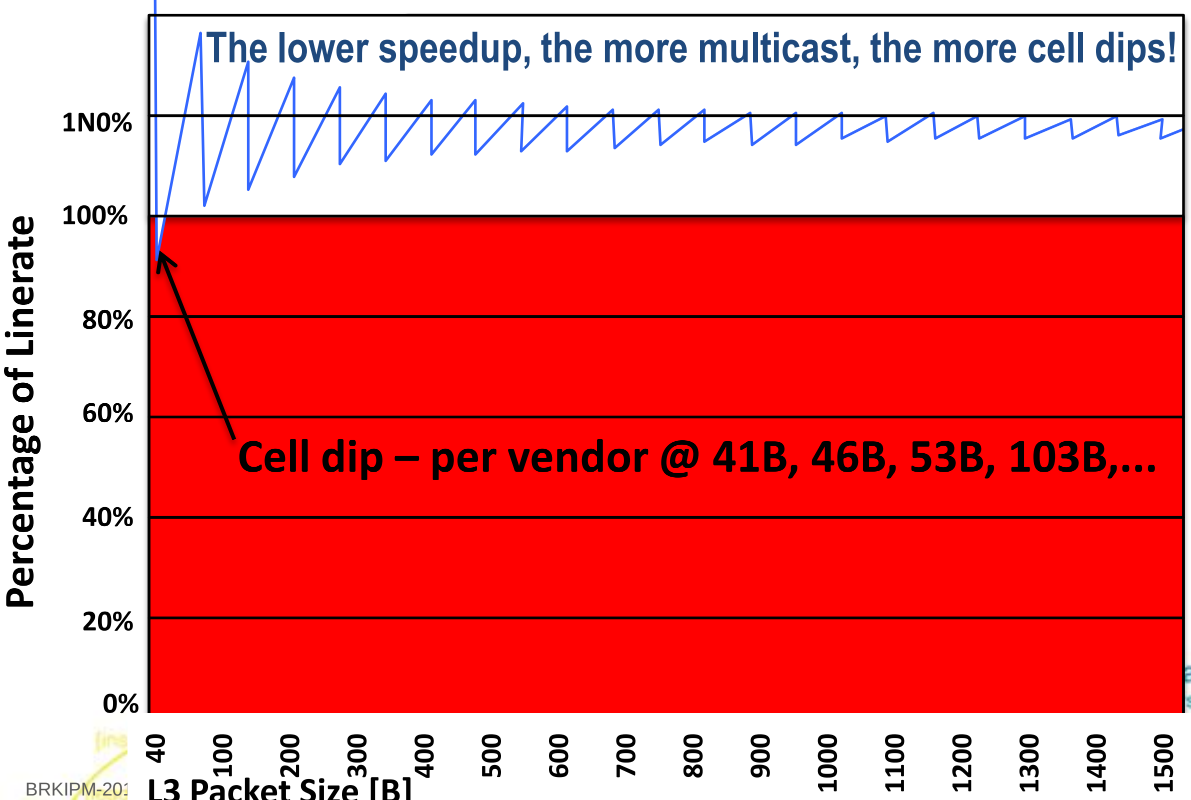
41B IP Packet:



Fixed overhead [cell header, ~10%]
Relative overhead [fabric header]
Variable overhead [padding]

Good efficiency

1Mpps = 1Mcps
1Gb/s → 1.33Gb/s



Cell Dip mitigation:

1. Enough Speedup
2. Packet Packing (CRS)
3. Super-framing (ASR9K)

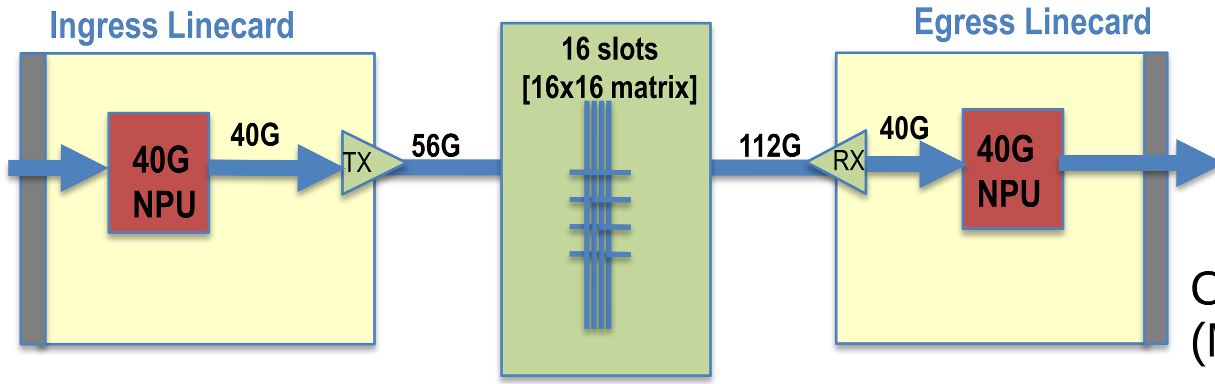
Quality Differences in 40G Solutions

40 ≠ 40

Good:

40 Gbps per slot

non-blocking

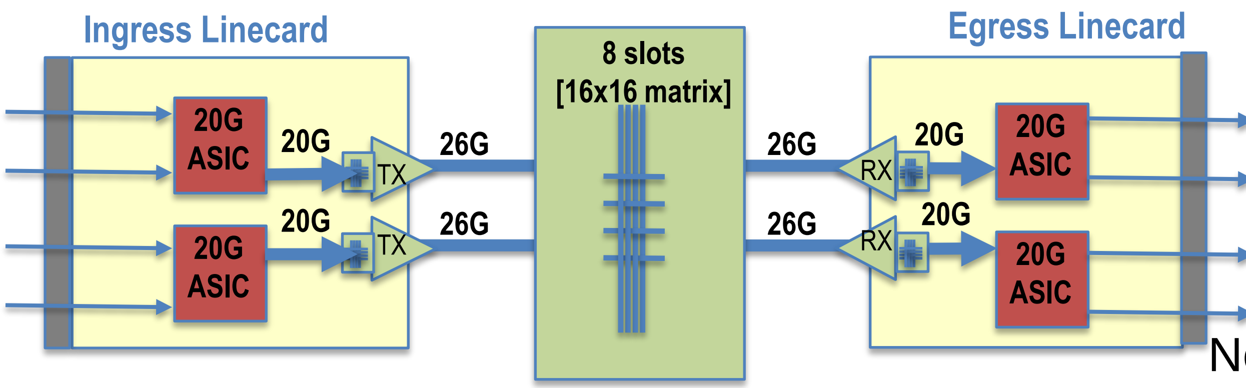


Cisco CRS-1 (MSC-40)

Good-enough:

40 Gbps per slot

non-blocking

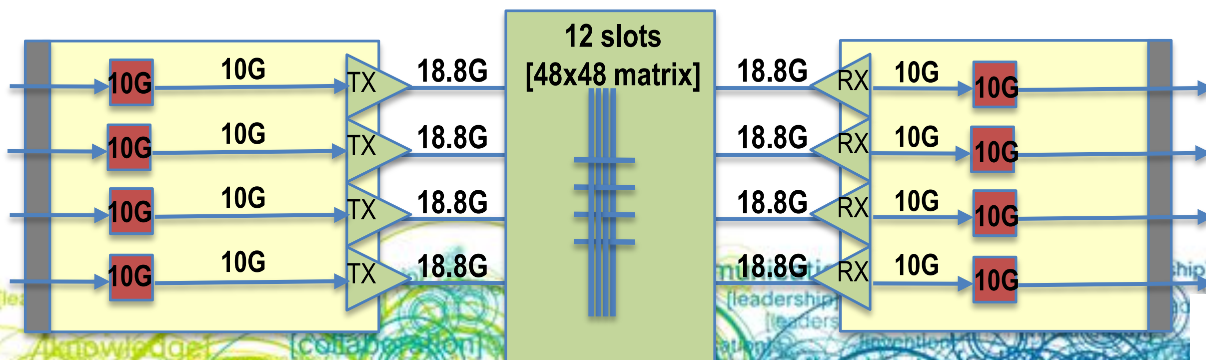


Non-Cisco

Good-enough?:

40 Gbps per slot

non-blocking!



Non-Cisco

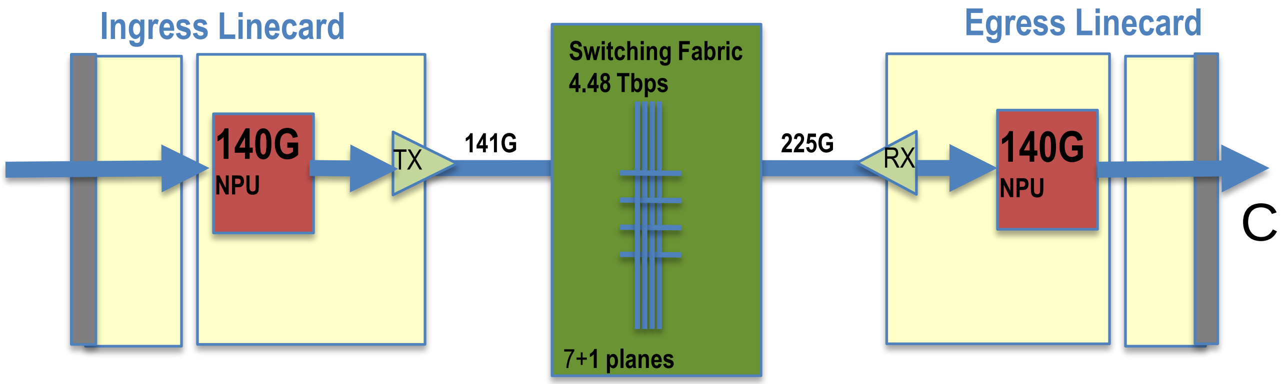
Impact of too many fabric connections

"How to do 100GE?"

Good 100GE

100GE non-blocking even with failed fabric card

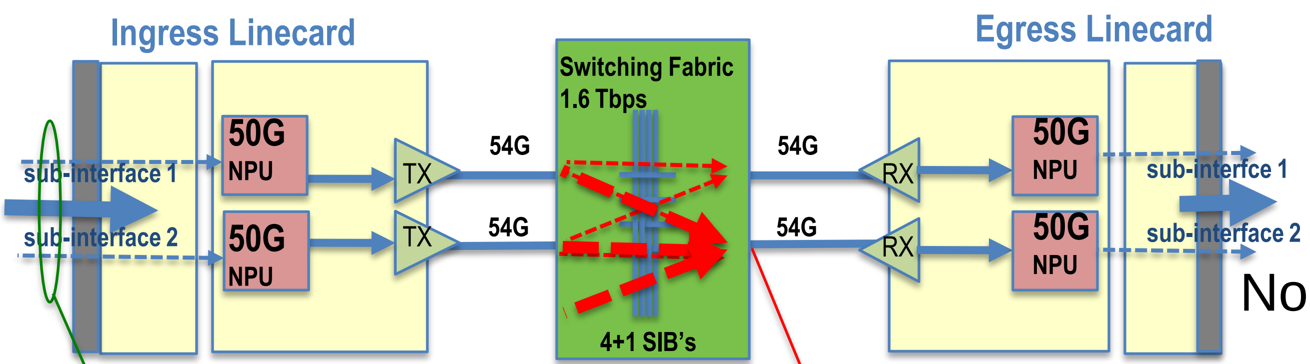
1.4x → 1.2x speedup



Bad 100GE

ECMP/Link-Bundling

No 100G processor available, per-destination IP load-sharing across two old 50G processors with two old 54G fabric ports



More than 54G → Fabric Port gets overloaded!

Head of Line Blocking?

Port-Channel across 2 vlans on the same physical port

(to keep 1 IP and 1 MAC address per port)

→ no multi-vendor interoperability

Summary

- Motivation for IP NGN = traffic growth
- How to make the Network cheaper
- How to make the Router cheaper
- Quality differences

There are Good, Good-enough or Bad Solutions.

Thank you.