Cisco Connect

Dubrovnik, Croatia, South East Europe
20-22 May, 2013

Anatomy of Internet Routers

Josef Ungerman

Cisco, CCIE #6167

Agenda

On the Origin of Species

- Router Evolution
- Router Anatomy Basics

Packet Processors

- Lookup, Memories, ASIC, NP, TM, parallelism
- Examples, evolution trends

Switching Fabrics

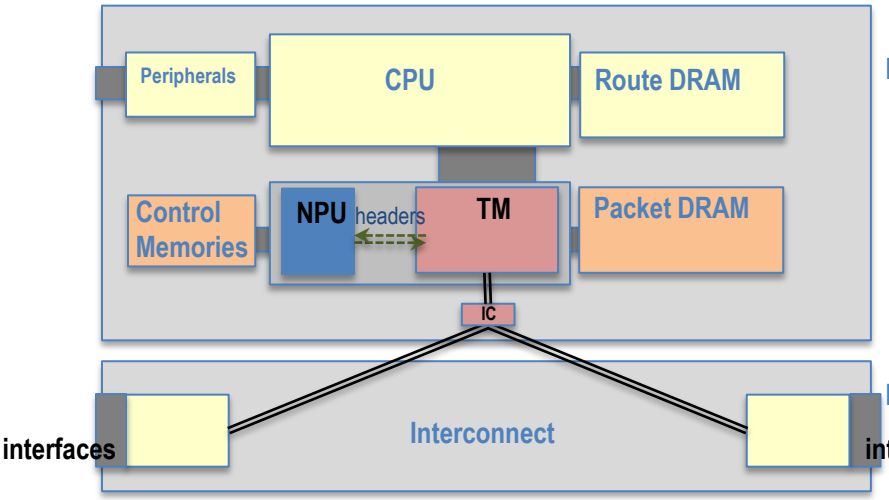
- Interconnects and Crossbars
- Arbitration, Replication, QoS, Speedup, Resiliency

Router Anatomy

- Past, Present, Future – CRS, ASR9000
- 1Tbps per slot?

Hardware Router

Control Plane vs. Data Plane

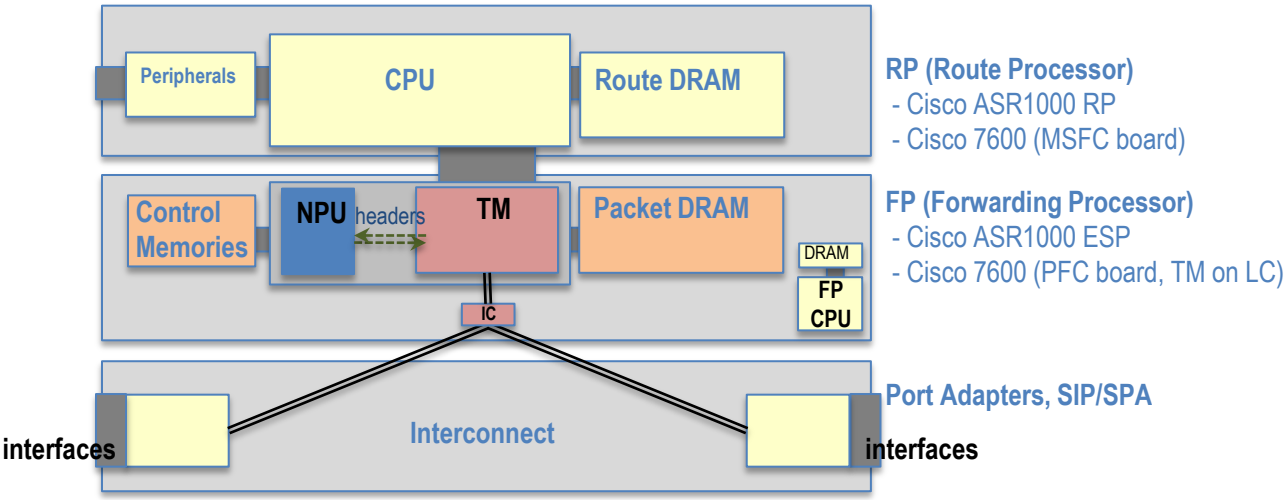


Routing and Forwarding Processor
- Cisco 10000 PRE
- Cisco 7300 NSE
- Cisco ASR903 RSP
- Cisco ASR1001/1002 (fixed)

Port Adapters, SIP/SPA

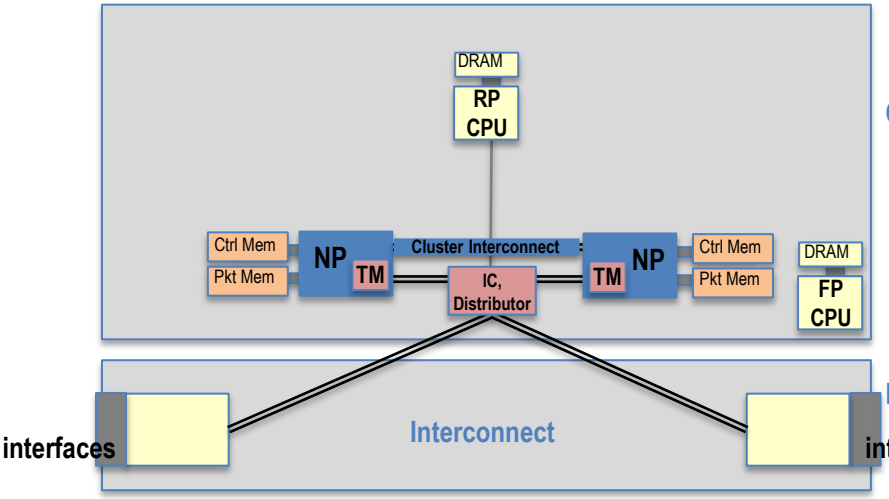
Hardware Router

Centralized Architecture



Scaling the Forwarding Plane

NP Clustering



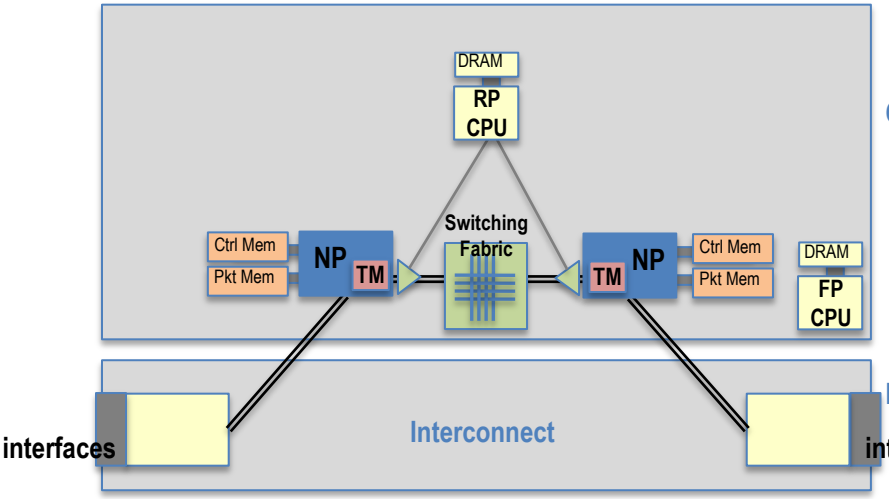
Centralized Router with ASIC Cluster

- ME3600X (fixed)
- ASR903 (modular with RSP)
- ASR1006 (modular with FP and RP)

Port Adapters, SIP/SPA

Scaling the Forwarding Plane

Switching Fabric



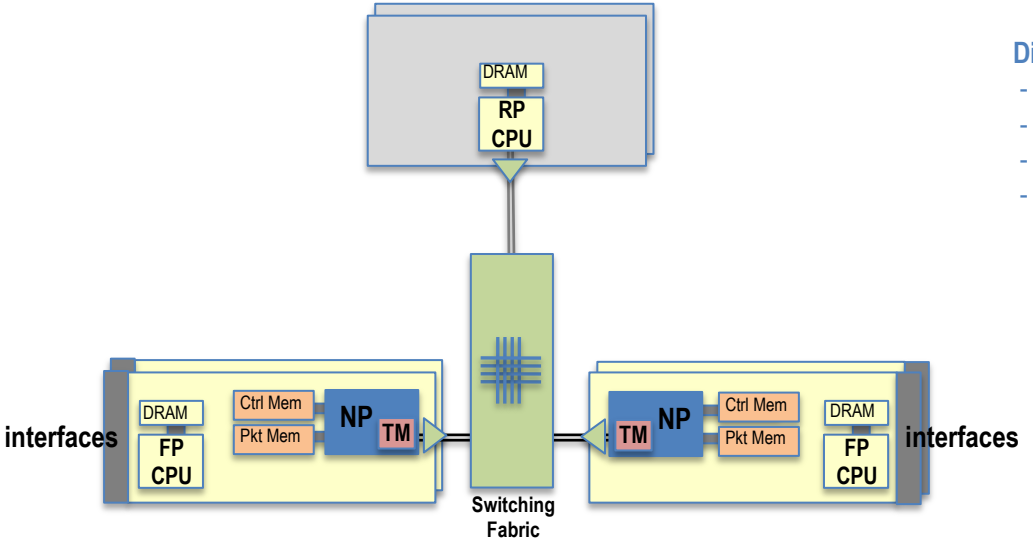
Centralized Router with Switching Fabric

- Cisco 7600 without DFC (TM on LC)
- Cisco ASR9001

Port Adapters – MPA

Scaling the Forwarding Plane

Distributed Architecture



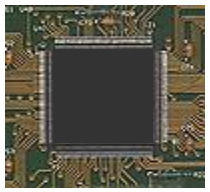
Distributed Router

- Cisco 7600 with DFC (ES+)
- Cisco 12000
- Cisco CRS
- Cisco ASR9000

Packet Processors

Packet Processing Trade-offs

Performance vs. Flexibility



CPU (Central Processing Unit)

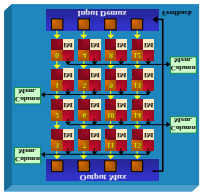
CPU

- multi-purpose processors
- high s/w flexibility [weeks], but low performance [1's of Mpps]
- high power, low cost
- **usage example:** access routers (ISR's)

ASIC (Application Specific Integrated Circuit)

ASIC

- mono-purpose hard-wired functionality
- complex design process [years]
- high performance [100's of Mpps]
- high development cost (but cheap production)
- **usage example:** switches (Catalysts)



NP (Network Processor) = "something in between"

NP

- performance [10's of Mpps] + programmability [months]
- cost vs. performance vs. flexibility vs. latency vs. power
- high development cost
- **usage example**: core→edge, aggregation routers

"It is always something

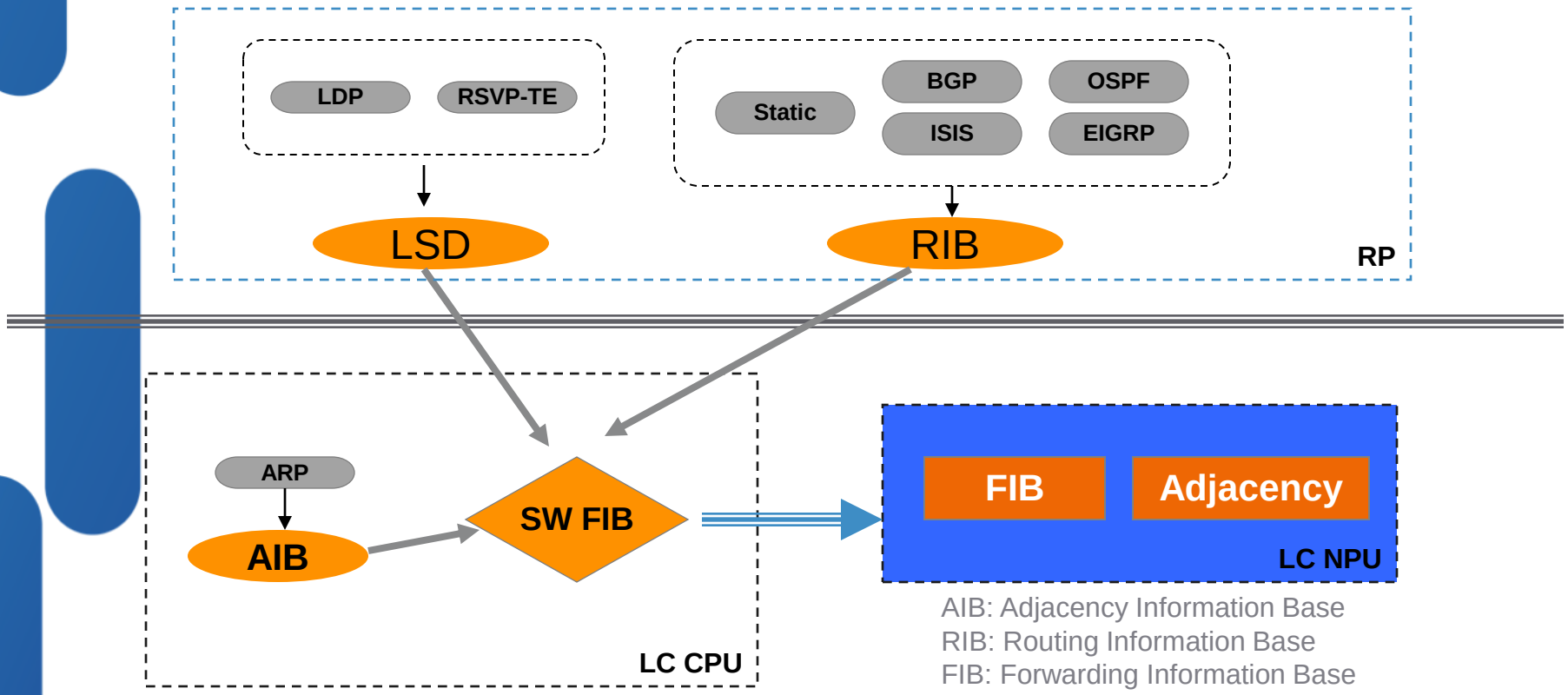
(corollary). Good, Fast, Cheap:

Pick any two (you can't have all three)."

RFC 1925

"The Twelve Networking Truths"

Hardware Routing Terminology



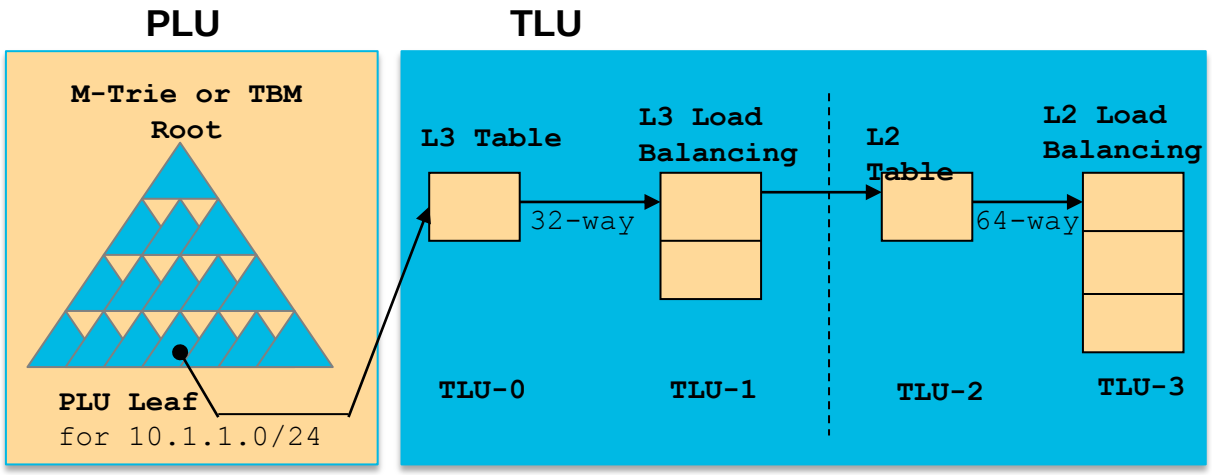
AIB: Adjacency Information Base
RIB: Routing Information Base
FIB: Forwarding Information Base
LSD: Label Switch Database

FIB Memory & Forwarding Chain

TLU/PLU

- memories storing Trie Data (today typically RLDRAM)
- Typically multiple channels for parallel/pipelined lookup

- **PLU (Packet Lookup Unit)** – L3 lookup data (FIB itself)
- **TLU (Table Lookup Unit)** – L2 adjacencies data (hierarchy, load-sharing)



Pointer chain allowing In-place Modify (PIC)

CAM

CAM (Content Addressable Memory)

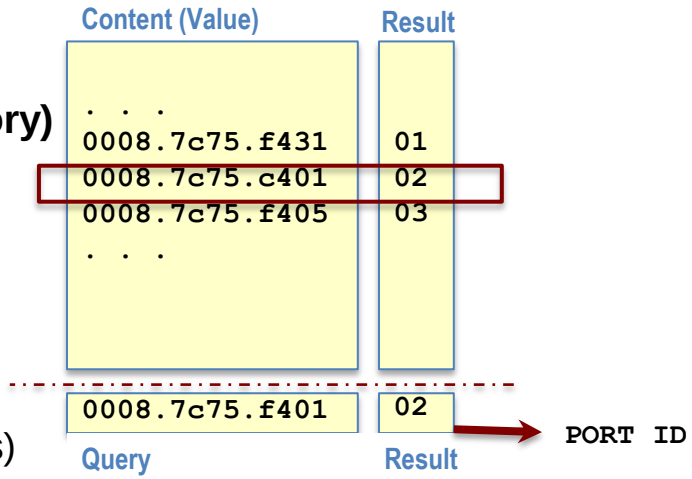
"Associative Memory"

SRAM with a Comparator at each cell

Stable O(1) lookup performance

Is expensive & power-hungry

usage: L2 switching (MAC addresses)



L2 Switching (also VPLS)

Destination MAC address lookup → Find the egress port (**Forwarding**)

- Read @ Line-rate = Wire-speed Switching

Source MAC address lookup → Find the ingress port (**Learning**)

- Write @ Line-rate = Wire-speed Learning

TCAM

TCAM (Ternary CAM)

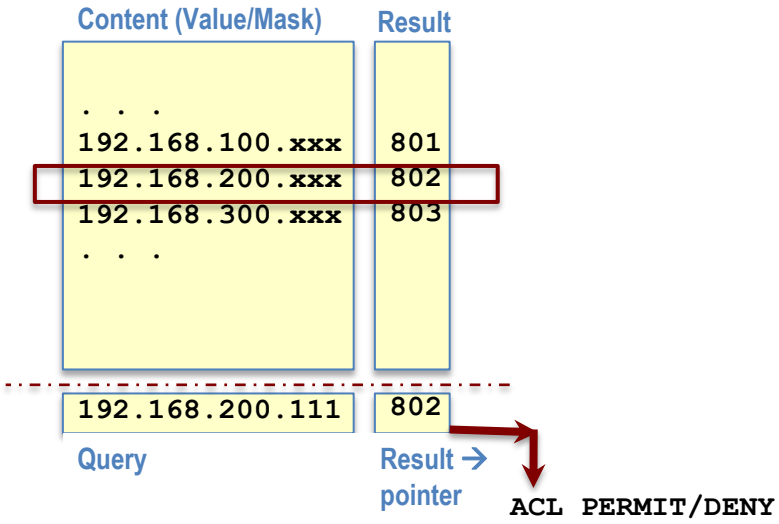
"CAM with a wildcard" (VMR)

CAM with a Selector at some cells

Stable O(1) lookup performance

3rd state – Don't Care bit (mask)

usage: IP lookup (addr/mask)



IP Lookup Applications

L3 Switching (Dst Lookup) & **RPF** (Src Lookup)

Netflow Implementation (flow lookup)

ACL Implementation (Filters, QoS, Policers…)

various other lookups

TCAM Evolution
CAM2 – 180nm, 80Msps, 4Mb, 72/144/288b wide
CAM3 – 130nm, 125 Msps, 18Mb, 72/144/288b wide
CAM4 – 90nm, 250Msps, 40Mb, 80/160/320b wide

Pipelining Programmable ASIC

2002: Engine3 (ISE) – Cisco 12000

✓4 Mpps, 3 Gbps
✓u-programmable stages
✓2 per LC (Rx, Tx)

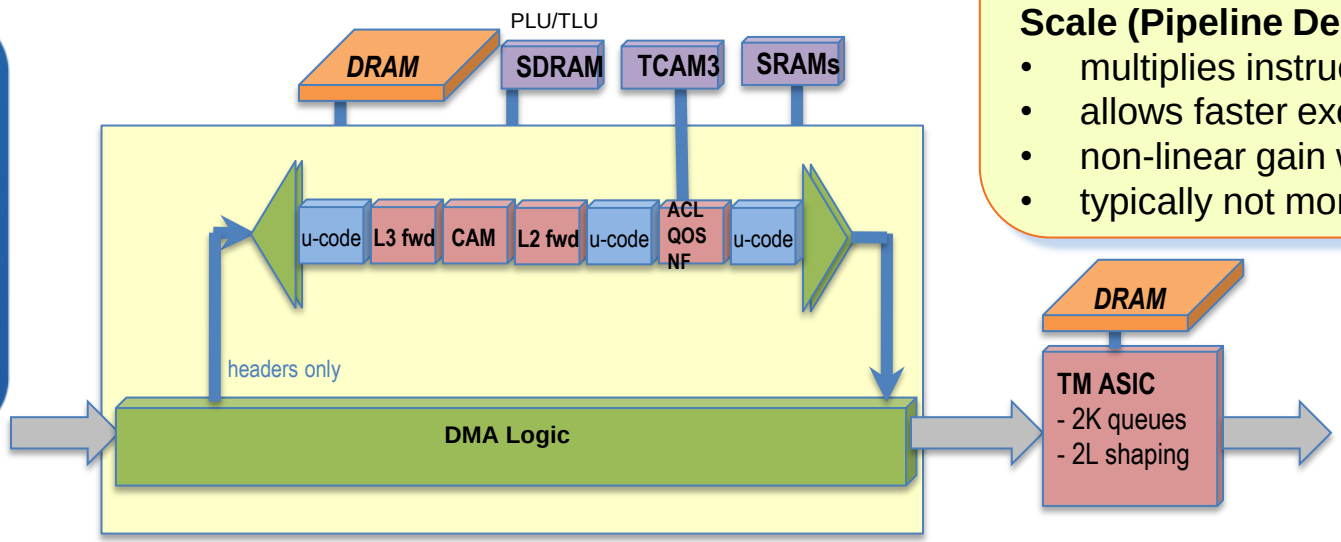
Parallelism Principle #1

Pipeline

- Systolic Array, 1-D Array

Scale (Pipeline Depth):

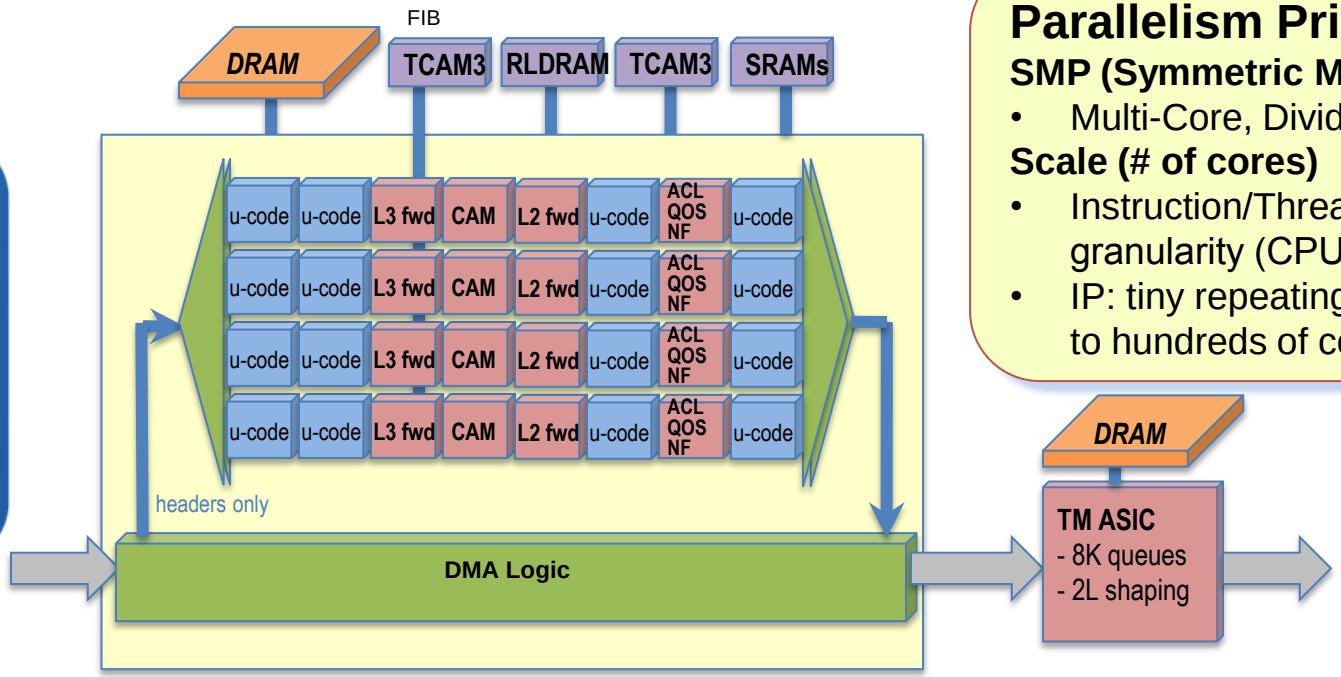
- multiplies instruction cycle budget
- allows faster execution (MHz)
- non-linear gain with pipeline depth
- typically not more than 8-10 stages



SMP Pipelining Programmable ASIC

2004: Engine5 (SIP) – Cisco 12000

✓16 Mpps, 10 Gbps
✓130/90nm, u-programmable
✓2 per LC (Rx, Tx)
✓240W/10G = 24 W/Gbps



Parallelism Principle #2:

SMP (Symmetric Multiprocessing)

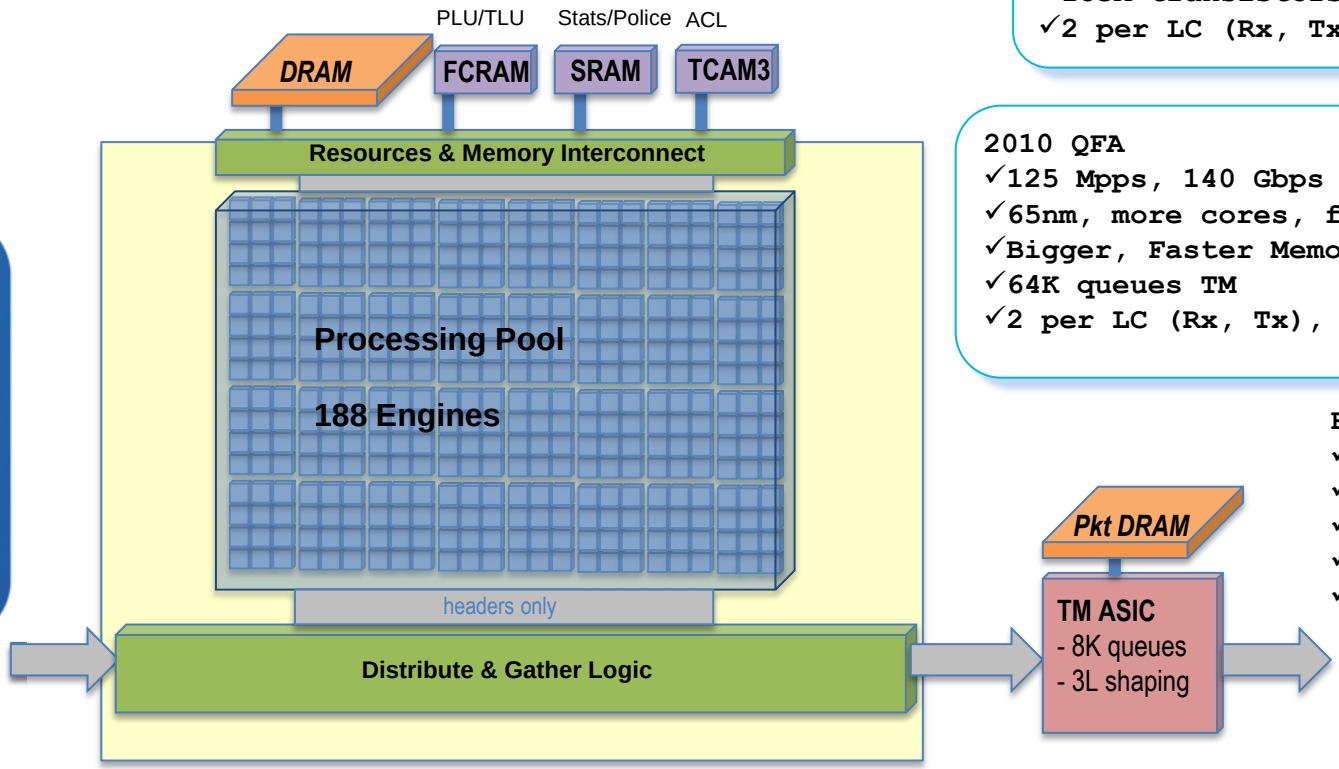
- Multi-Core, Divide & Conquer

Scale (# of cores)

- Instruction/Thread/Process/App granularity (CPU's: 2-8 core)
- IP: tiny repeating tasks (typically up to hundreds of cores)

SMP NPU

QFA (Quantum Flow Array) – CRS



2004 QFA (SPP)
- ✓80 Mpps, 40 Gbps
- ✓130nm, 188 cores
- ✓185M transistors
- ✓2 per LC (Rx, Tx),~9 W/Gbps

2010 QFA
- ✓125 Mpps, 140 Gbps
- ✓65nm, more cores, faster MHz
- ✓Bigger, Faster Memories (RLDRAM, TCAM4)
- ✓64K queues TM
- ✓2 per LC (Rx, Tx), ~4.2 W/Gbps

Future
- ✓40nm version
- ✓400G
- ✓More cores, more MHz
- ✓Integrated TM
- ✓Faster TCAM etc.

"If you were plowing a field, which would you rather use:
Two strong oxen or 1024 chickens?"

Seymour Cray

"What would Cinderella pick
to separate peas from ashes?"

Unknown IP Engineer

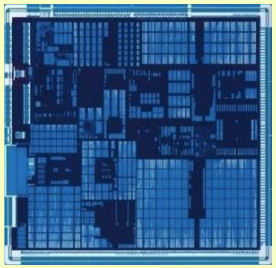
"Good multiprocessors are built
from good uniprocessors"

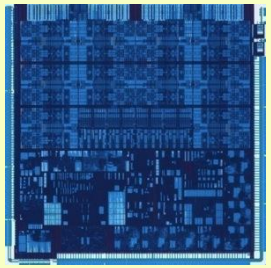
Steve Krueger

PPE (Packet Processing Elements)

Generic CPU's or COT?

- NP does not need many generic CPU features
  - floating point ops, BCD or DSP arithmetic
  - complex instructions that compiler does not use (vector, graphics, etc.)
  - privilege/protection/hypervisor
  - Large caches
- Custom improvements
  - H/W assists (TCAM, PLU, HMR…)
  - Faster memories
  - Low power
  - C language programmable! (portable, code reuse)





QFP:

- **>1.3B transistors**
- **>100 engineers**
- **>5 years of development**
- **>40 patents**

	Cisco QFP	Sun Ultrasparc T2	Intel Core 2 Mobile U7600
Total number processes (cores x threads)	160	64	2
Power per process	0.51W	1.01W	5W
Scalable traffic management	128k queues	None	None

Packaging Examples:

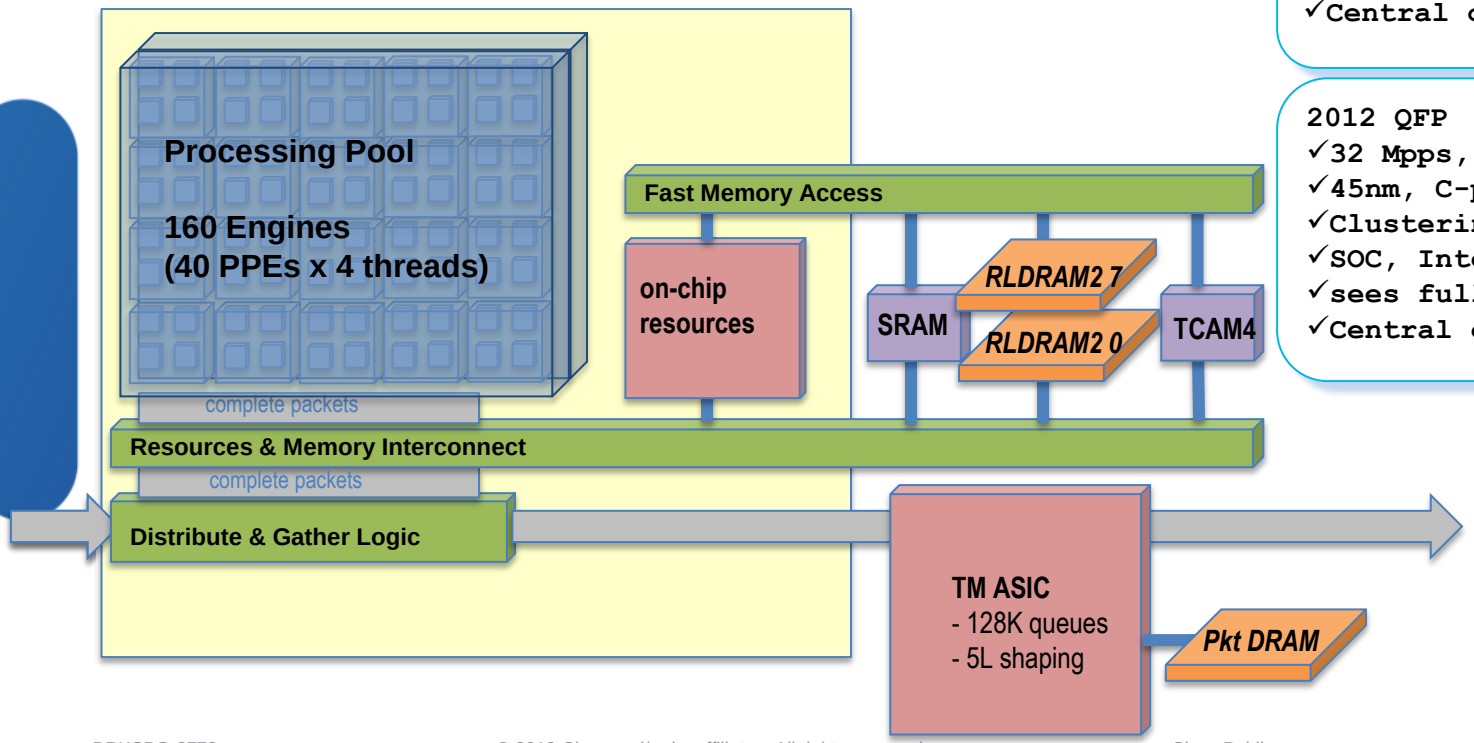
- **ESP5 = 20 PPEs @ 900MHz**
- **ESP10 = 40 PPEs @ 900MHz**
- **ESP20 = 40 PPEs @ 1200MHz**
- **etc.**

SMP NPU (full packets processing)

QFP (Quantum Flow Processor) – ASR1000 (ESP), ASR9000 (SIP)

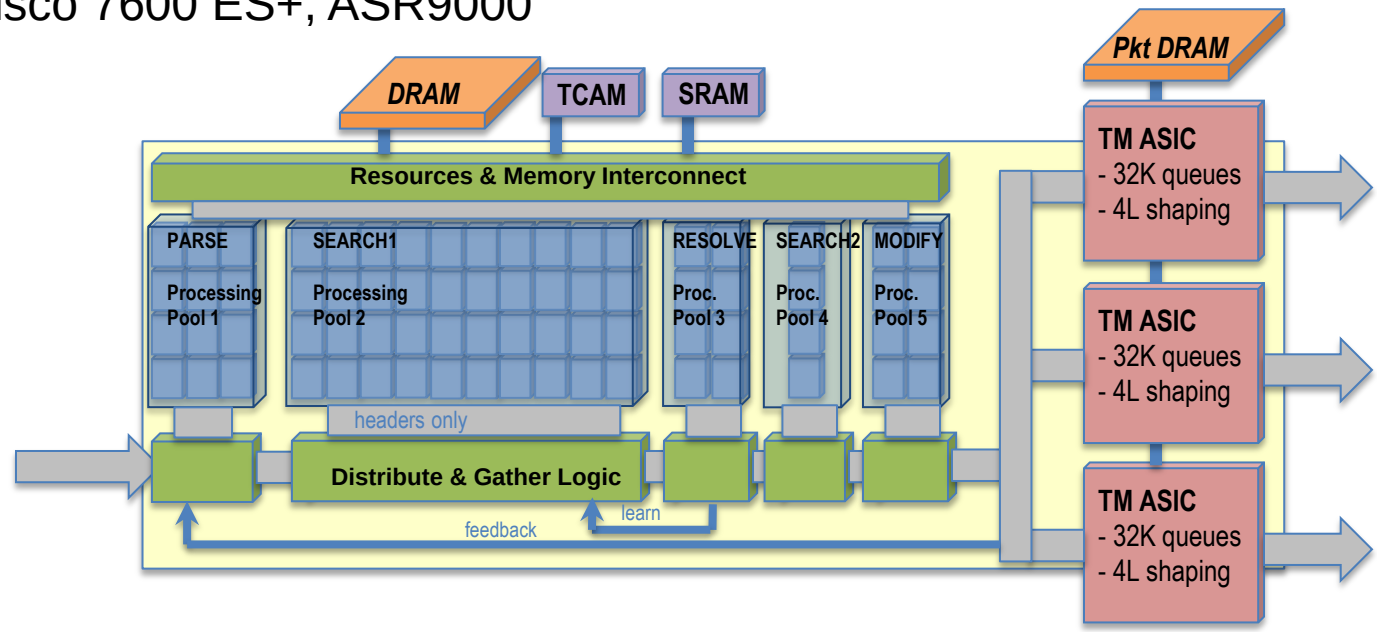
2008 QFP
- ✓16 Mpps, 20 Gbps
- ✓90nm, C-programmable
- ✓sees full packet bodies
- ✓Central or distributed

2012 QFP
- ✓32 Mpps, 60 Gbps
- ✓45nm, C-programmable
- ✓Clustering capabilities
- ✓SOC, Integrated TM
- ✓sees full packet bodies
- ✓Central engine (ASR1K)



Pipelining SMP NPU

Cisco 7600 ES+, ASR9000



2008 NP [Trident]:
- ✓28 Mpps, 30 Gbps
- ✓90nm, 70+ cores
- ✓3 on-board TM chips (2 Tx)
- ✓2,4 or 8 per LC
- ✓565W/120G = 4.7 W/Gbps
- ✓7600 ES+, ASR9K -L/-B/-E

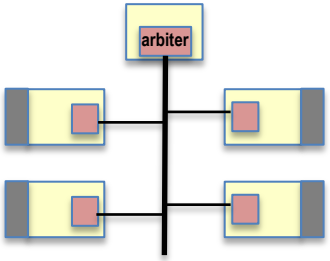
2011 NP [Typhoon]:
- ✓90 Mpps, 120 Gbps
- ✓55nm, lot more cores
- ✓Integrated TM and CPU
- ✓2,4 or 8 per LC
- ✓800W/240G = 3.3 W/Gbps
- ✓ASR9K -TR/-SE

Future
- ✓40nm version
- ✓200+G

Switching Fabrics

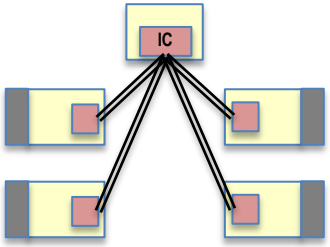
Interconnects Technology Primer

Capacity vs. Complexity



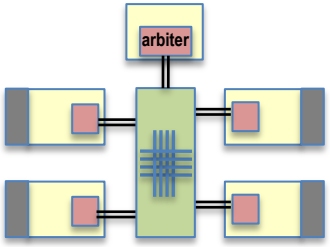
Bus

- half-duplex, shared media
- standard examples: PCI [800Mbps], PCIe [Nx 2.5Gbps], LDT/HT [25Gbs]
- simple and cheap



Serial Interconnect

- full-duplex, point-to-point media
- standard examples: SPI [10Gbps], Interlaken [100Gbps]
- Ethernet interfaces are very common (SGMII, XAUI,…)



Switching Fabric (cross-bar)

- full-duplex, any-to-any media
- proprietary systems [up to multiple Tbps]
- often uses double-counting (Rx+Tx) to express the capacity

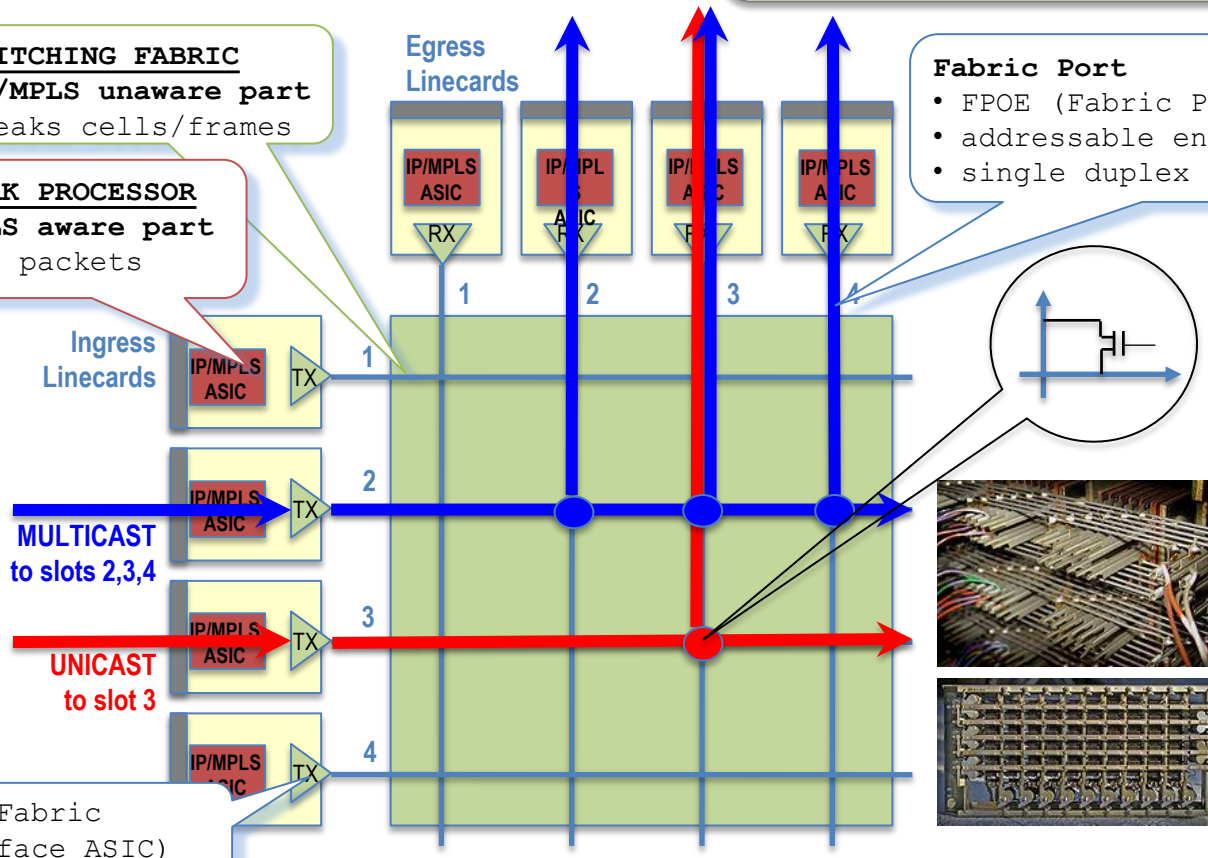
Switching Fabric

Q: What's the capacity? 4 fabric ports @ 10Gbps
A: (ENGINEERING) 4 * 10 = 40Gbps full-duplex
A: (MARKETING) 4 * 10 * 2 = 80Gbps

SWITCHING FABRIC
IP/MPLS unaware part
speaks cells/frames

NETWORK PROCESSOR
IP/MPLS aware part
speaks packets

Fabric Port
- FPOE (Fabric Point of Exit)
- addressable entity
- single duplex pipe

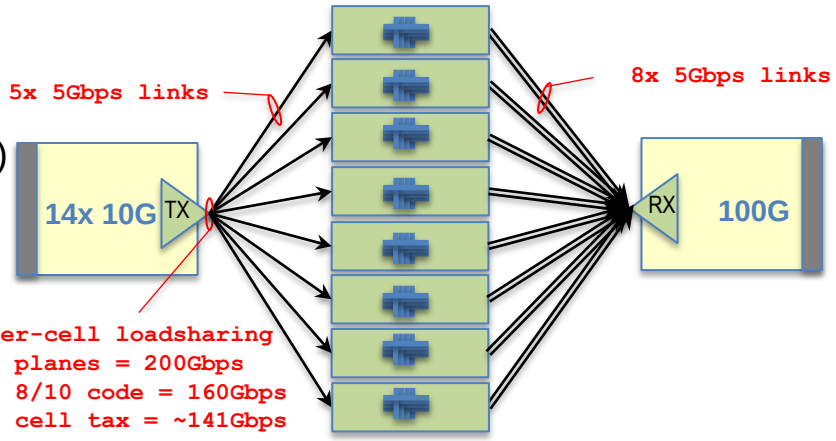


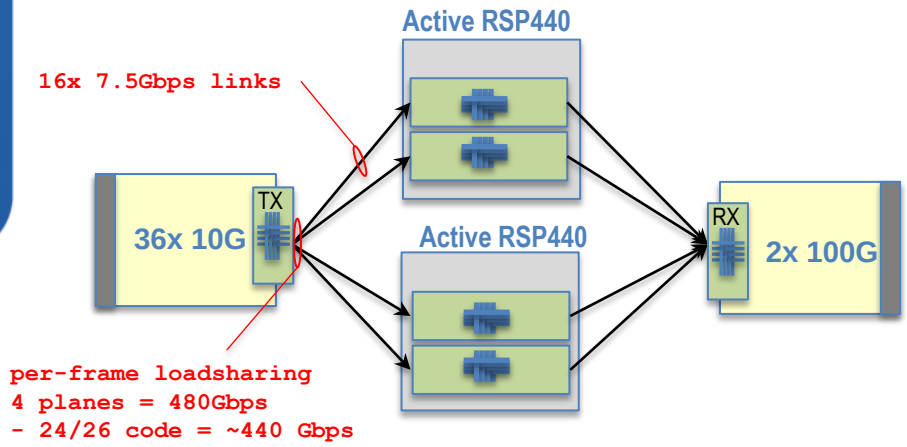
Type B (1960) Western Electric 100-point 6-wire crossbar switch

Fabric Port engineering – examples

CRS-3/16 – 4.48Tbps (7+1)

- 16 linecards, 2 RP
- linecard up to 140G (next 400G)
- backwards compatible



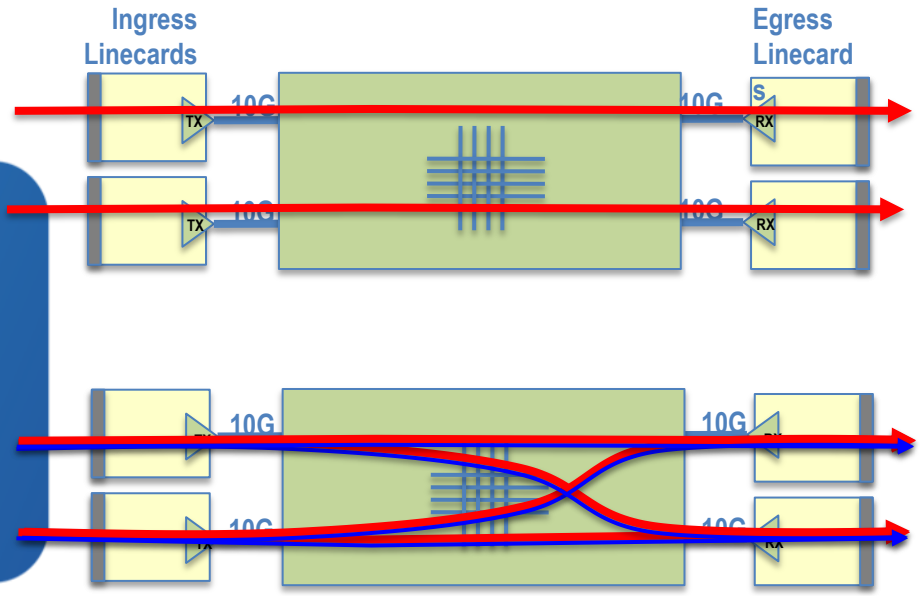


ASR9010 – 7.04Tbps (1+1)

- 8 linecards, 2 RSP
- linecard up to 360G (next 800G)
- backwards compatible

“Non-Blocking” voodoo

RFC1925: It is more complicated than you think.



Non-blocking!

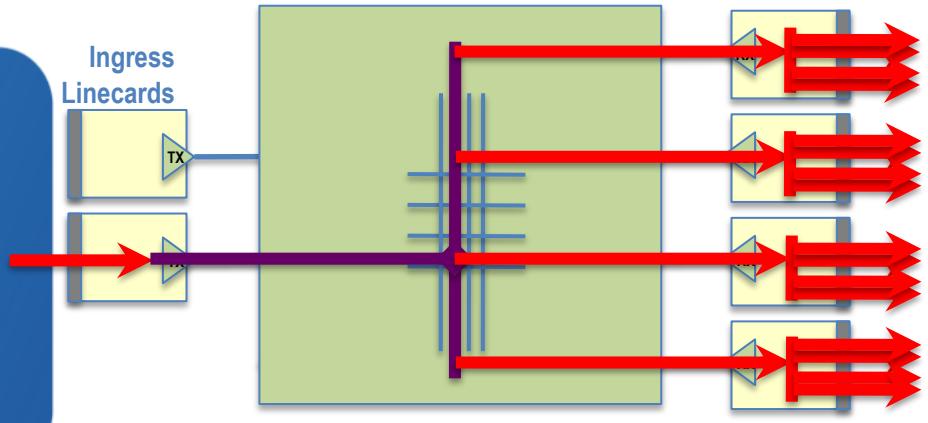
- zero packet loss
- port-to-port traffic profile
- certain packet size

Blocking (same fabric)..?

- packet loss, high jitter
- added meshed traffic profile
- added Multicast
- added Voice/Video

Example: 16x Multicast Replication

Egress Replication



Good:

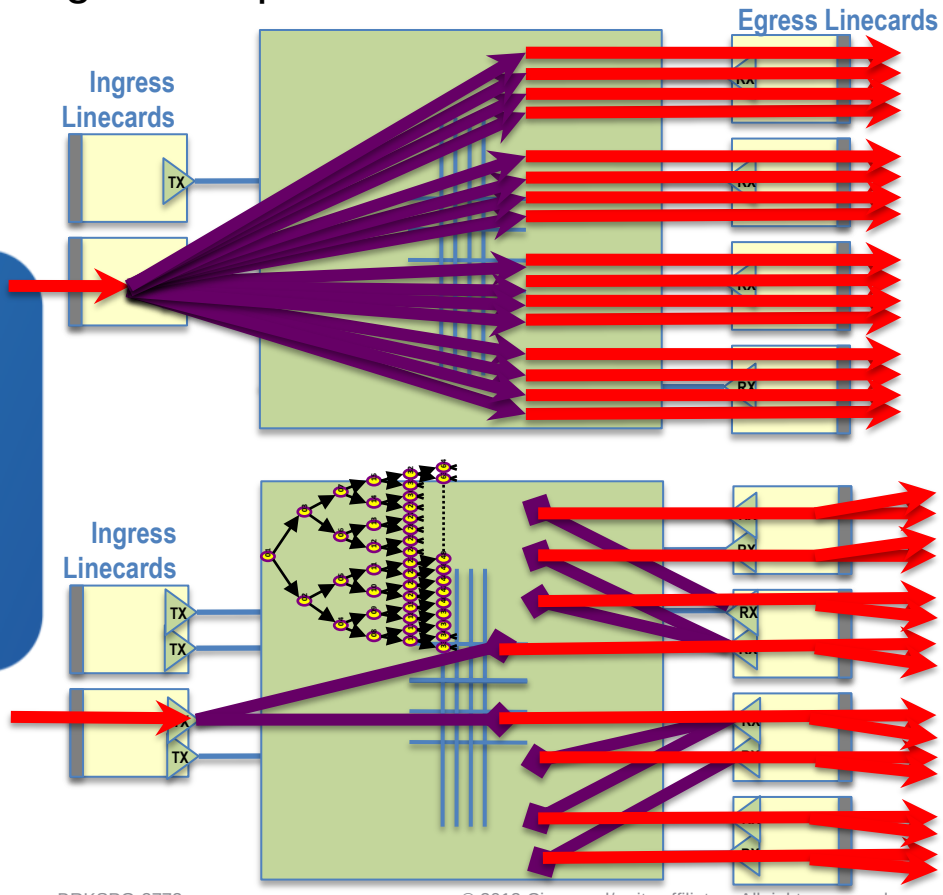
Egress Replication

- Cisco CRS, 12000
- Cisco ASR9K, 7600

10Gbps of multicast eats 10Gbps fabric bw!

What if the fabric can't replicate multicast?

Ingress Replication Flavors



Bad:

Ingress Replication

- central replication or encapsulation engines

*) of course, this is used in centralized routers

10Gbps of multicast
eats 160Gbps fabric bw!
(10G multicast impossible)

Good-enough/Not-bad-enough:

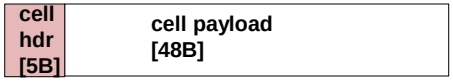
Binary Ingress Replication

- dumb switching fabric
- non-Cisco

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

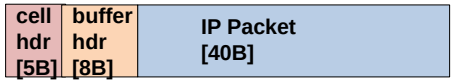
Cell dip explained

cell format example:



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

40B IP Packet:

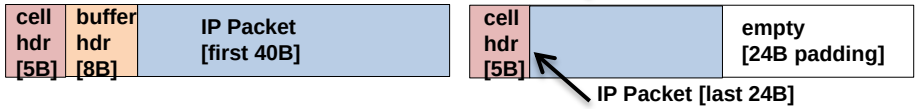


Good efficiency
1Mpps = 1Mcps
1Gb/s → 1.33Gb/s

41B IP Packet:

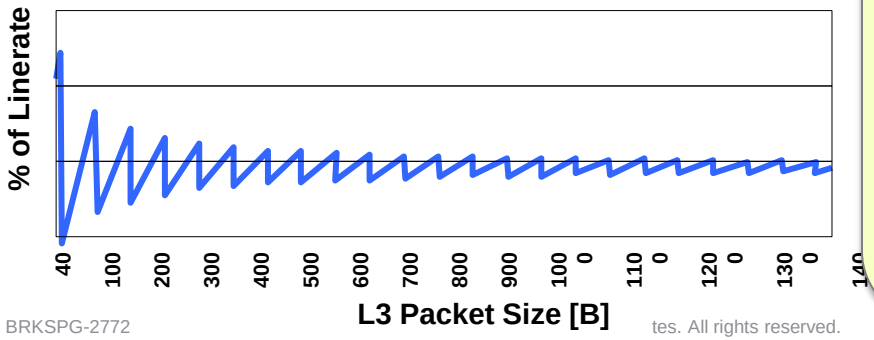
Poor efficiency
1Mpps = 2Mcps
1Gb/s → 2.6Gb/s

64B IP Packet:

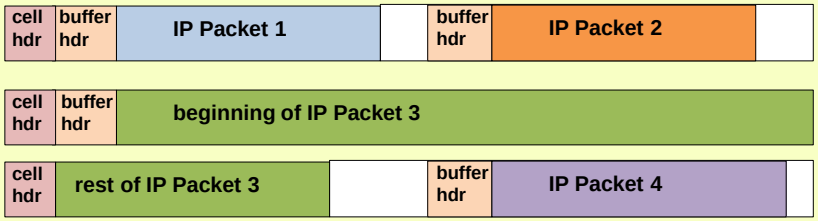


Fair efficiency
1Mpps = 2Mcps
1Gb/s → 1.7Gb/s

Cell Tax effect on traffic: saw-tooth curve

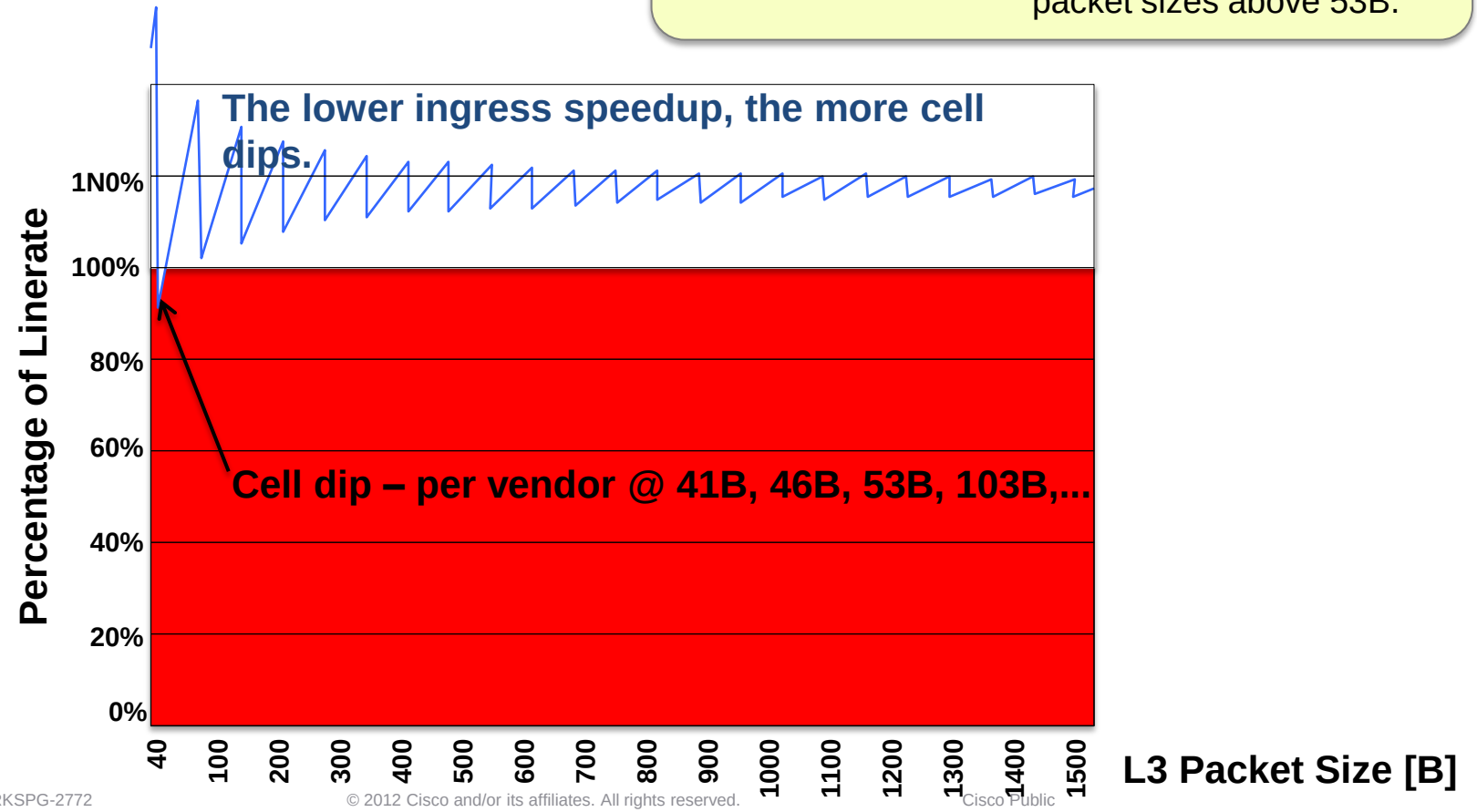


super-cell or super-frame (packet packing)



Cell dip

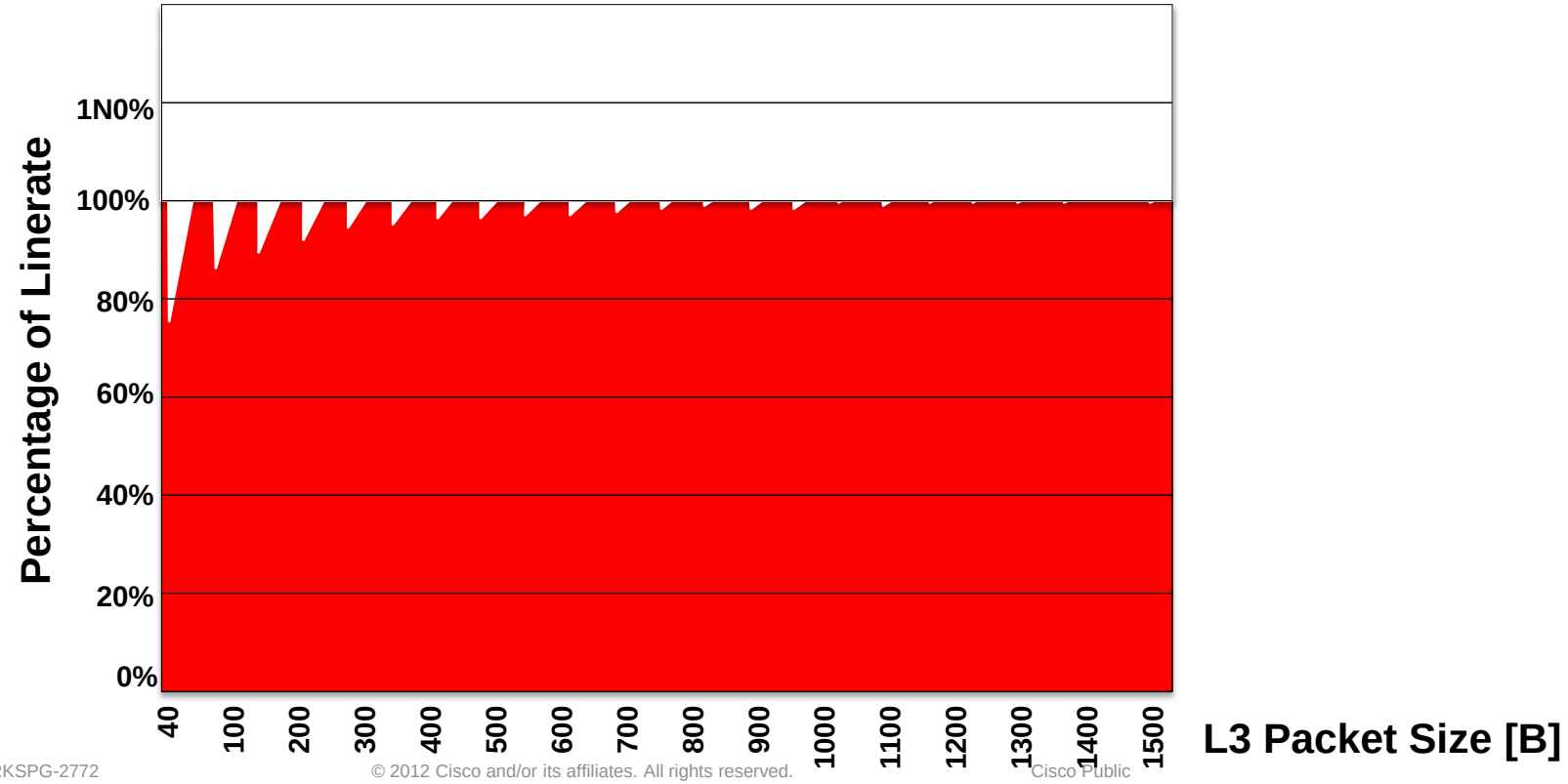
Q: Is this SF non-blocking?
A: (MARKETING) Yes Yes Yes !
A: (ENGINEERING) Non-blocking for unicast packet sizes above 53B.



Cell dip gets bad – too low speedup (non-Cisco)

MARKETING: this is non-blocking fabric*

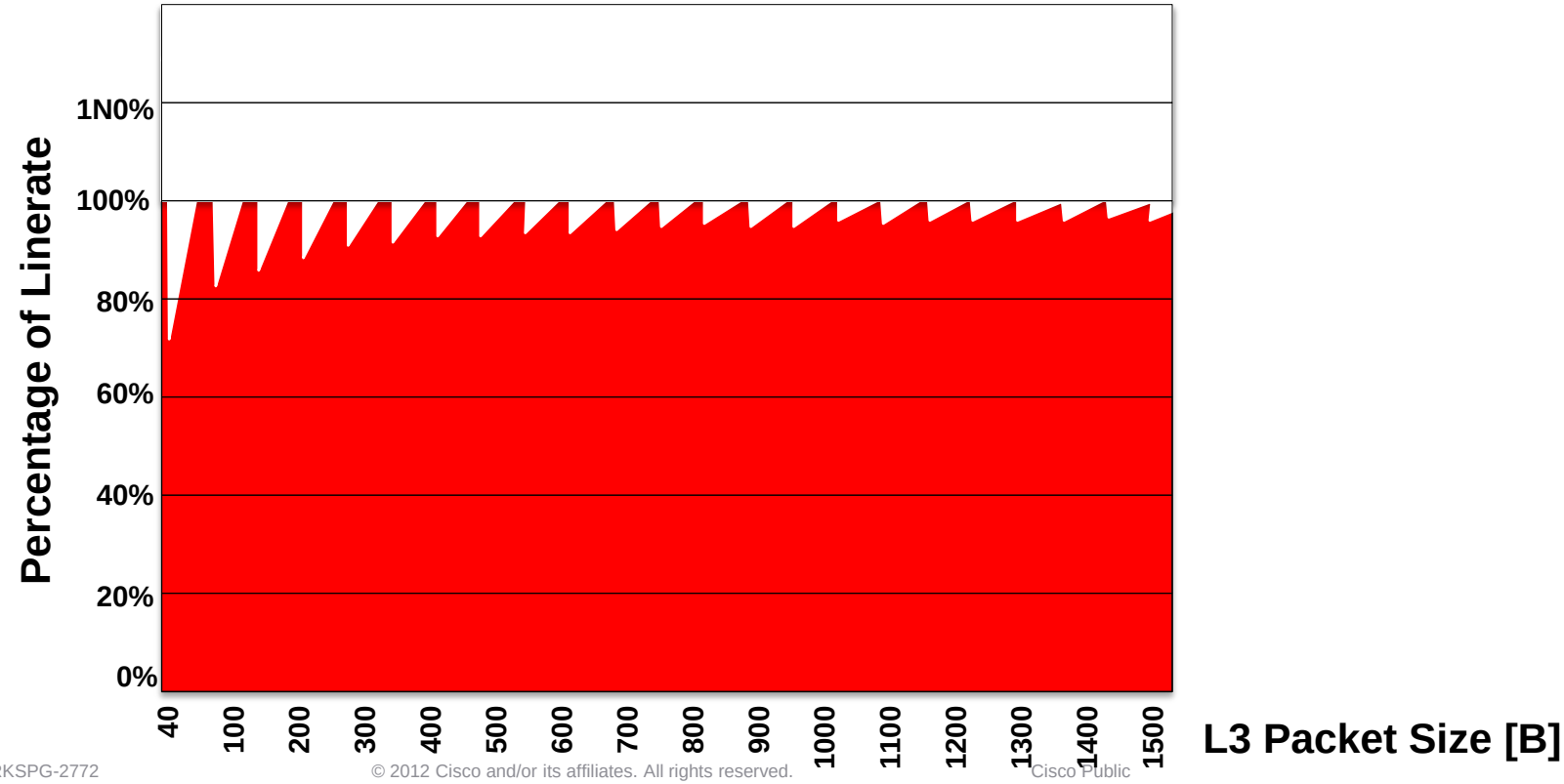
*) because we can find at least one packet size that does not block



Cell dip gets worse – multicast added (non-Cisco)

MARKETING: this is non-blocking fabric*

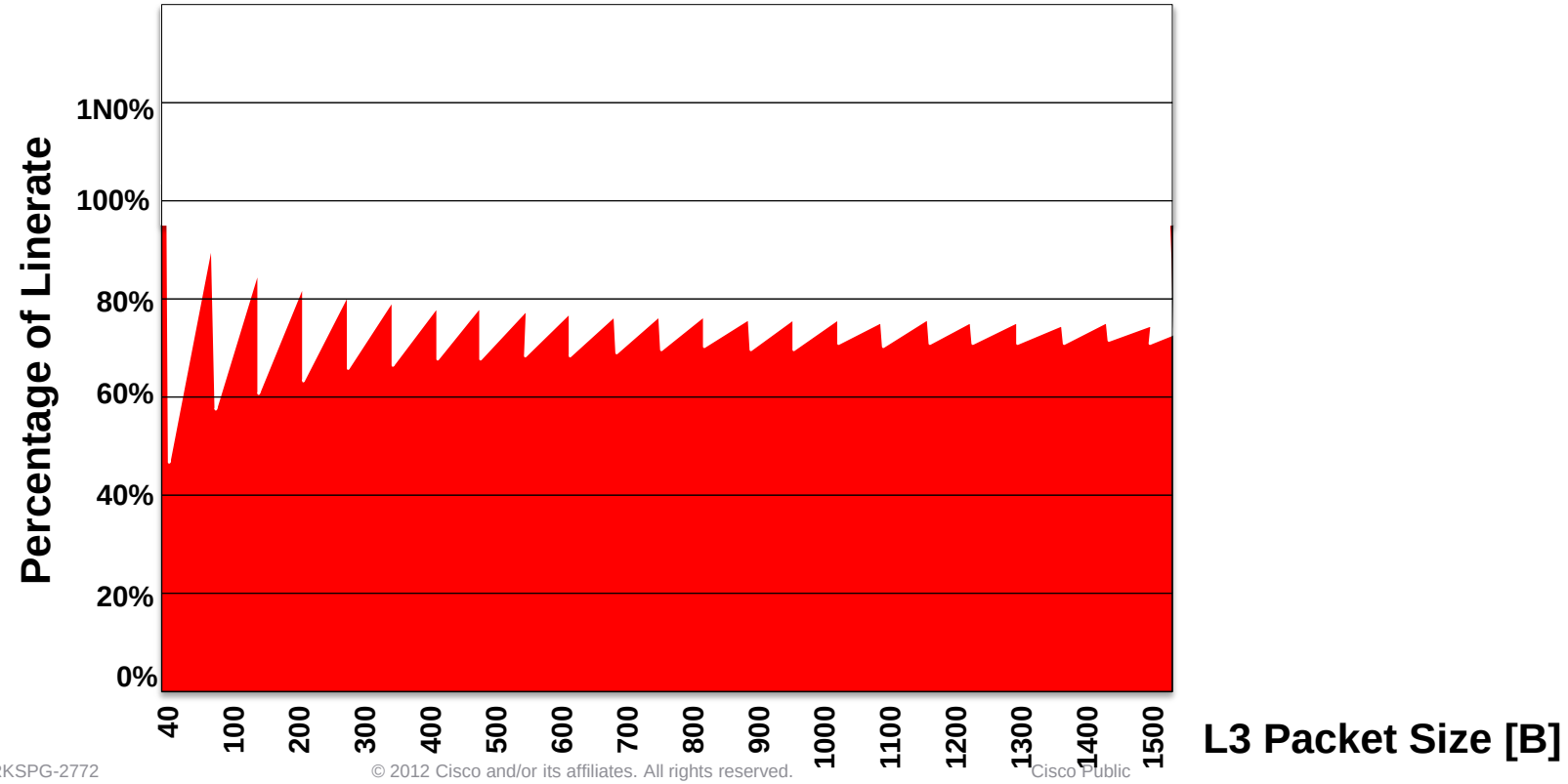
*) because we can still find at least one packet size that does not block, and your network does not have that much multicast anyway



Router is blocking – 1 fabric cards fails (non-Cisco)

MARKETING: this is non-blocking fabric*

*) because nobody said the non-blocking capacity is “protected”



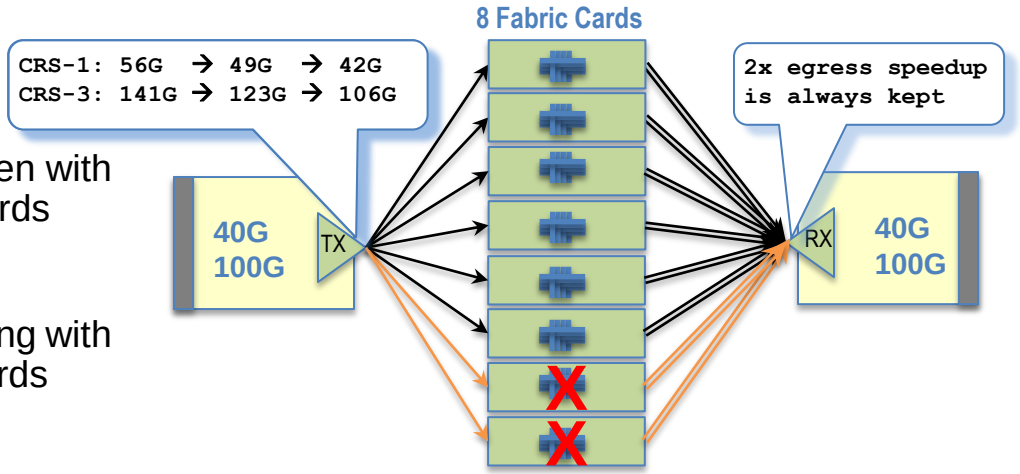
What is “Protected Non-Blocking”

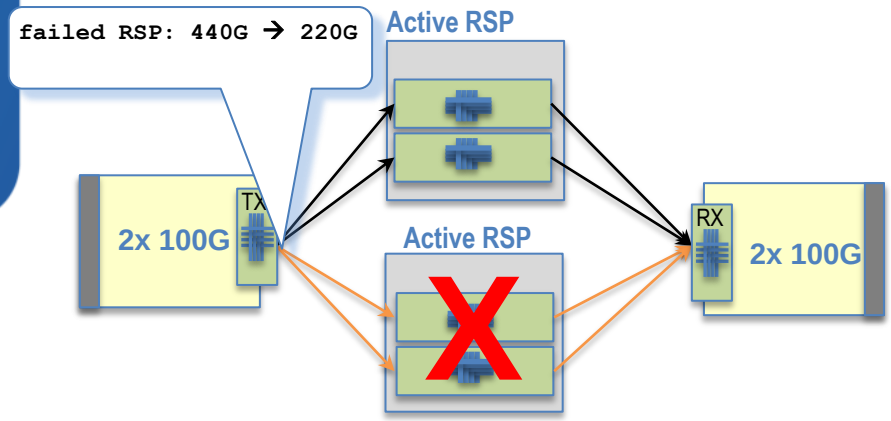
CRS-1

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

CRS-3

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



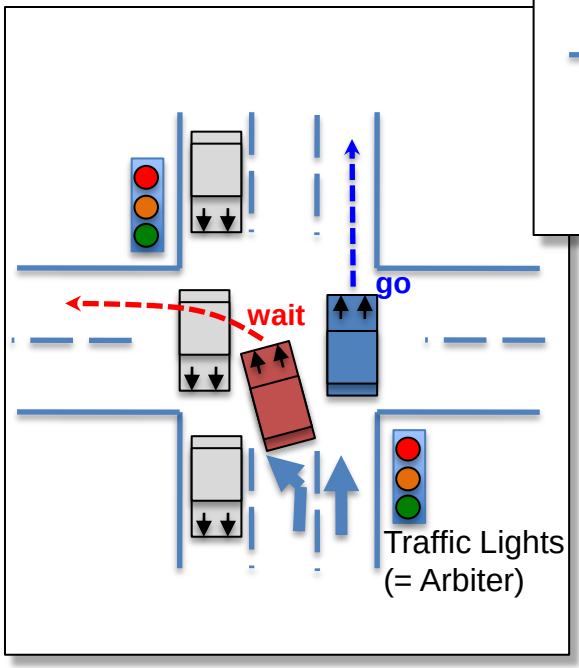


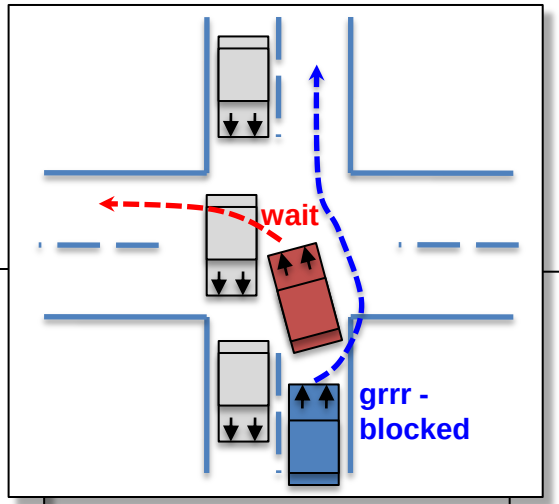
ASR9000

200G non-blocking even with a
failed RSP

HoLB (Head of Line Blocking) problem

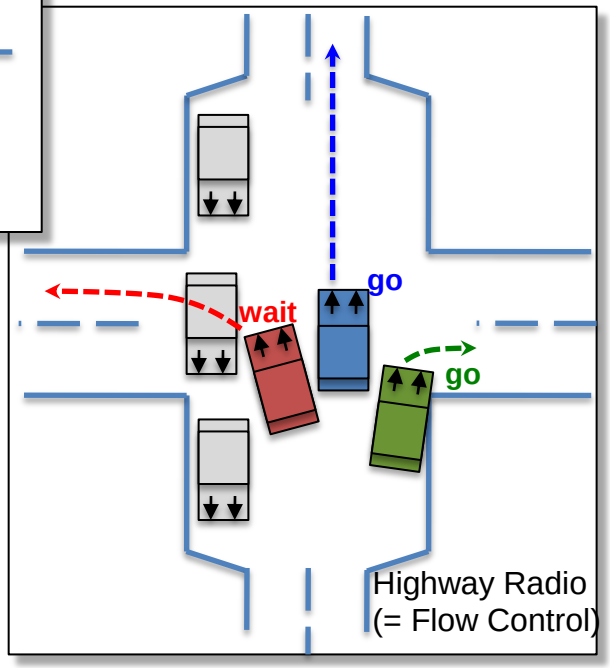
Solution 1:
Traffic Lanes per direction
(= Virtual Output Queues)





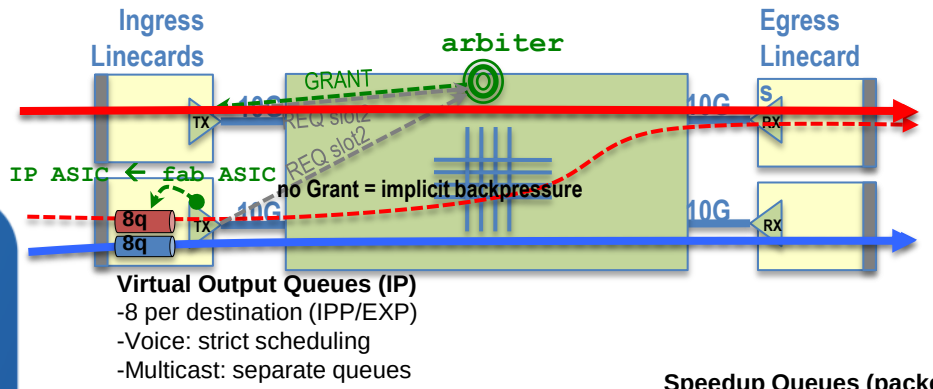
Red Light, or "Traffic Jam Ahead" message
(= Backpressure)

Solution 2:
Enough Room ☺
(= Speedup)



Good HoLB Solutions

Fabric Scheduling + Backpressure + QoS

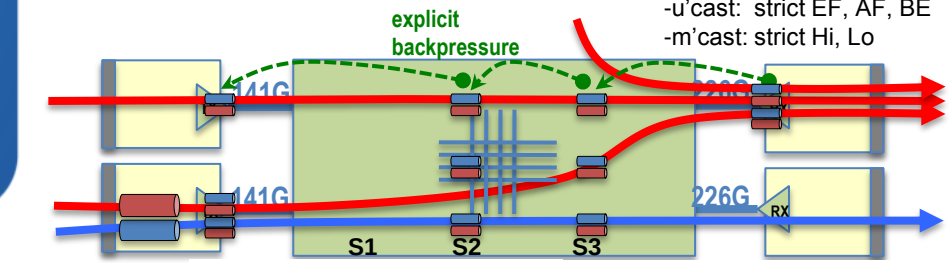


Virtual Output Queues (IP)
-8 per destination (IPP/EXP)
-Voice: strict scheduling
-Multicast: separate queues

Arbitrated: Implicit backpressure

+ Virtual Output Queues (VOQ)

- Cisco 12000 (also ASR9000)
- per-destination slot queues VOQ (Virtual Output Queues)
  - **GSR:** 16 slots + 2 RP's * 8 CoS + 8 Multicast CoS =152 queues per LC



Speedup Queues (packets)
-u'cast: strict EF, AF, BE
-m'cast: strict Hi, Lo

Input Q's (IP)
-configurable
-shaped

Destination Queues (packets)
-u'cast: strict Hi, Lo
-m'cast: strict Hi, Lo

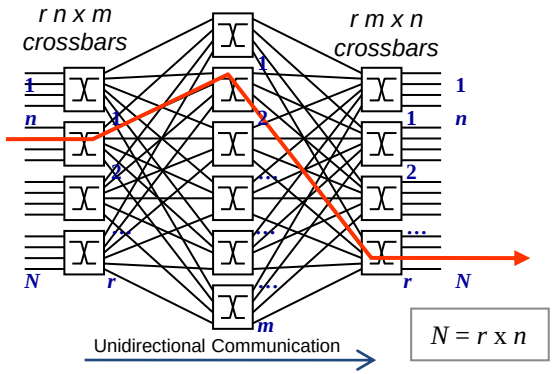
Fabric Queues (cells)
-u'cast: strict Hi, Lo
-m'cast: strict Hi, Lo

Output Buffered: Explicit backpressure

+ Speedup & Fabric Queues

- Cisco CRS (1296 slots!)
- 6144 destination queues
- 512 speedup queues
- 4 queues at each point (Hi/Lo UC/MC) + vital bit

Multi-stage Switching Fabrics

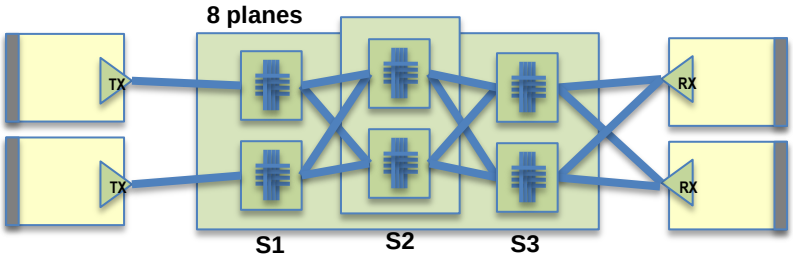


Multi-stage Switching Fabric

• constructing large switching fabric out of smaller SF elements

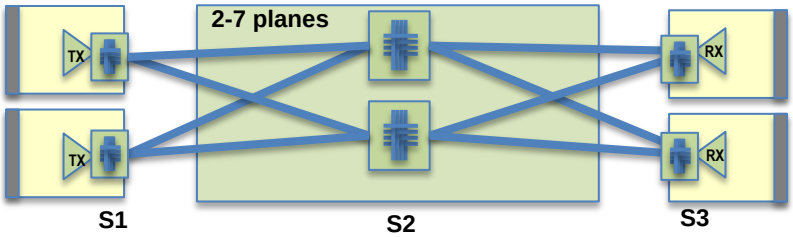
50's: Ch. Clos – general theory of multi-stage telephony switch

60's: V. Beneš – special case of *rearrangeably non-blocking* Clos (n = m = 2)



CRS-1 – Benes

- Multi-chassis capabilities (2+0, N+2,…)
- Massive scalability: up to 1296 slots !!!
- Output-buffered, speedup, backpressure

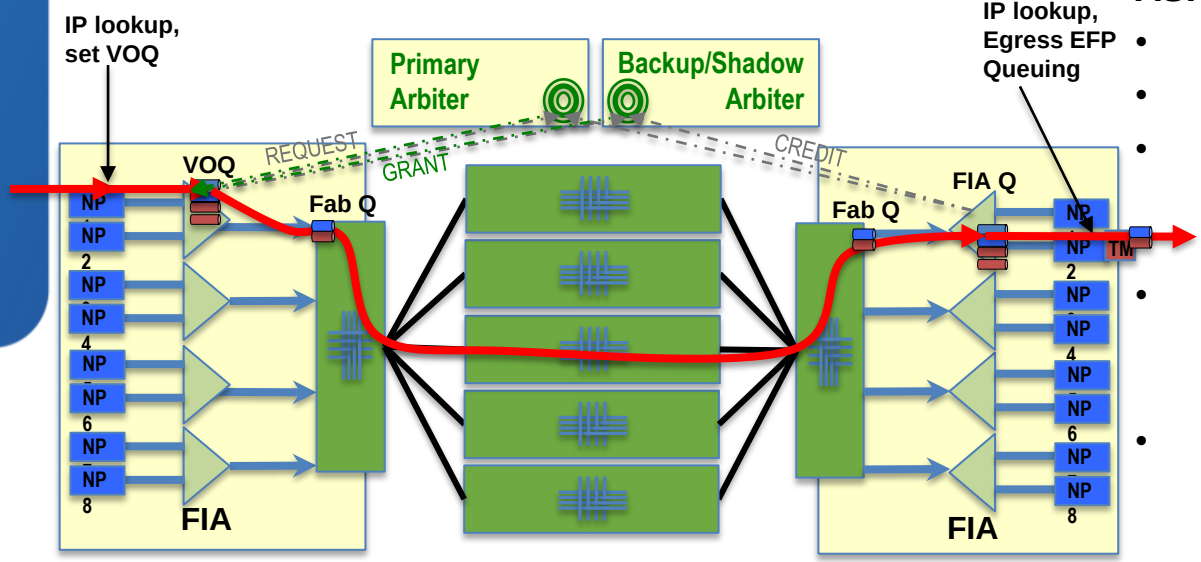


ASR9000 – Clos

- Single-chassis so far
- Scales to 22 slots today
- Arbitrated VOQ's

Virtual Output Queuing

- VOQ on ingress modules represents fabric capacity on egress modules
- VOQ is “virtual” because it represents egress capacity but resides on ingress modules, however it is still physical buffer where packets are stored
- VOQ is not equivalent to ingress or egress fabric channel buffers/queues
- VOQ is not equivalent to ingress or egress NP/TM queues

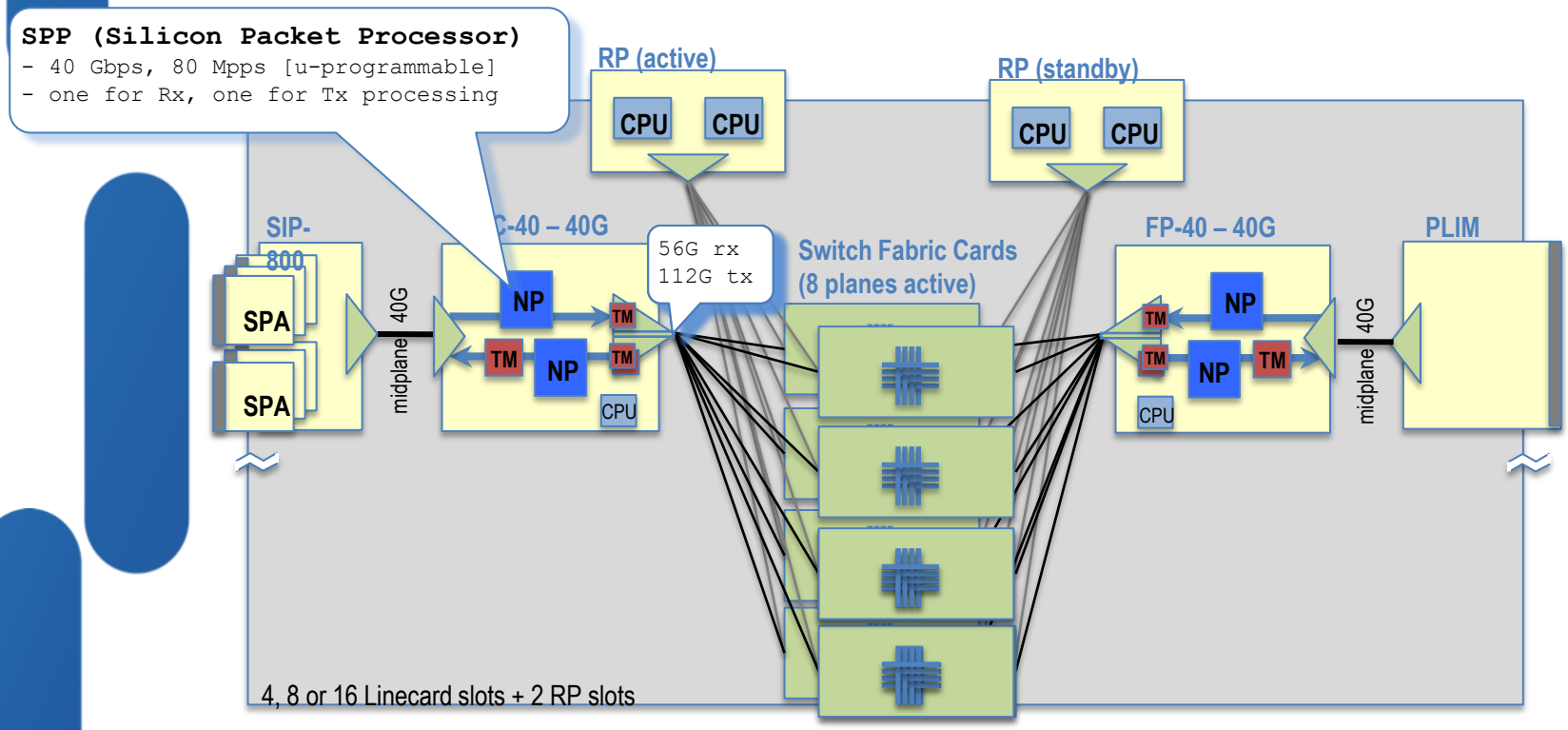


ASR9000

- Multi-stage Fabric
- Granular Central VOQ arbiter
- VOQ set per destination
  - Destination is the NP, not just slot
  - 4 per 10G, 8 per 40G, 16 per 100G
- 4 VOQ’s per set
  - 4 VOQ’s per destination, strict priority
  - Up to 4K VOQ’s per ingress FIA
- Example (ASR9922):
  - 20 LC’s * 8 10G NP’s * 4 VOQ’s
  - = up to 640 VOQ’s per ingress FIA

Router Anatomy

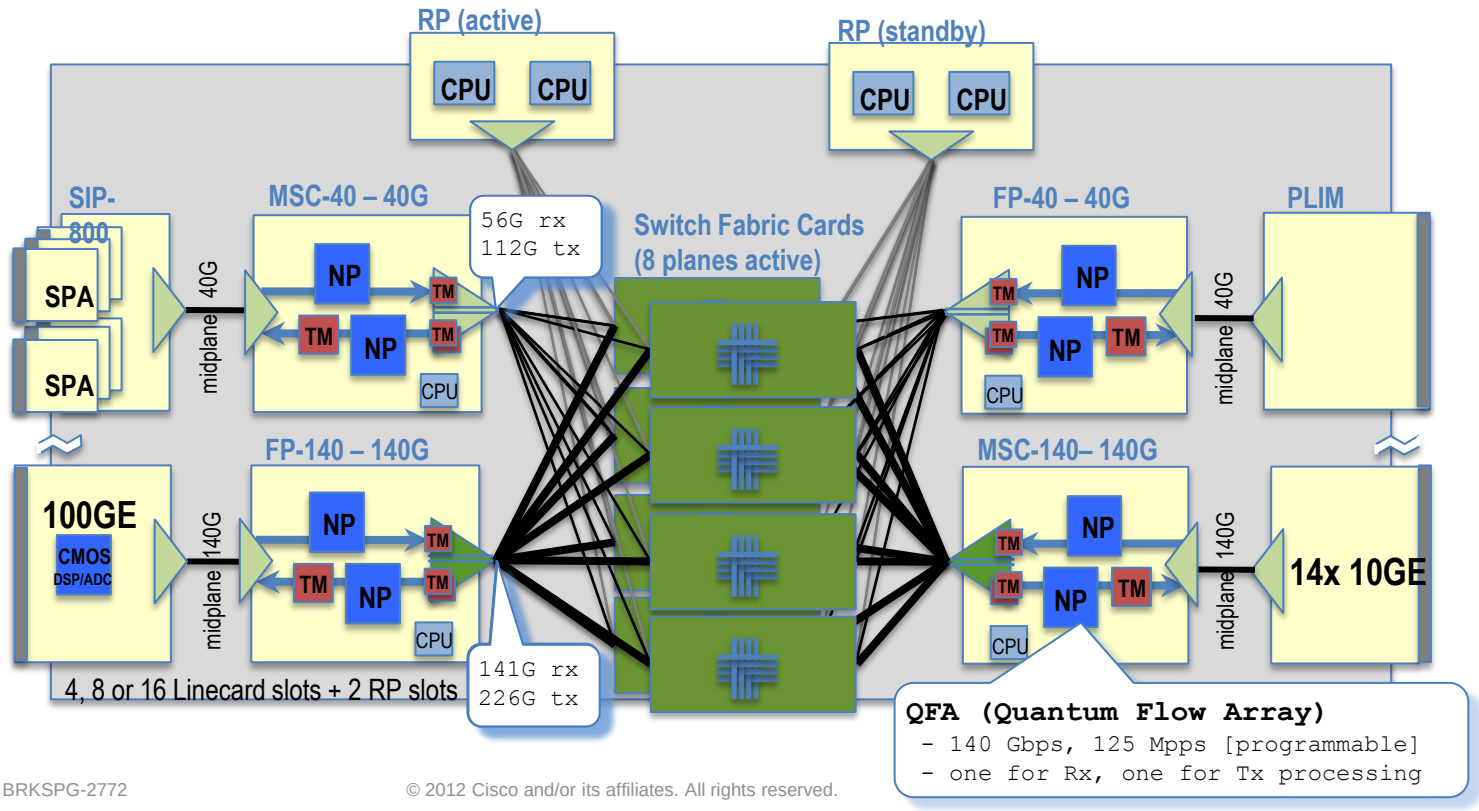
2004: Cisco CRS – 40G+ per slot



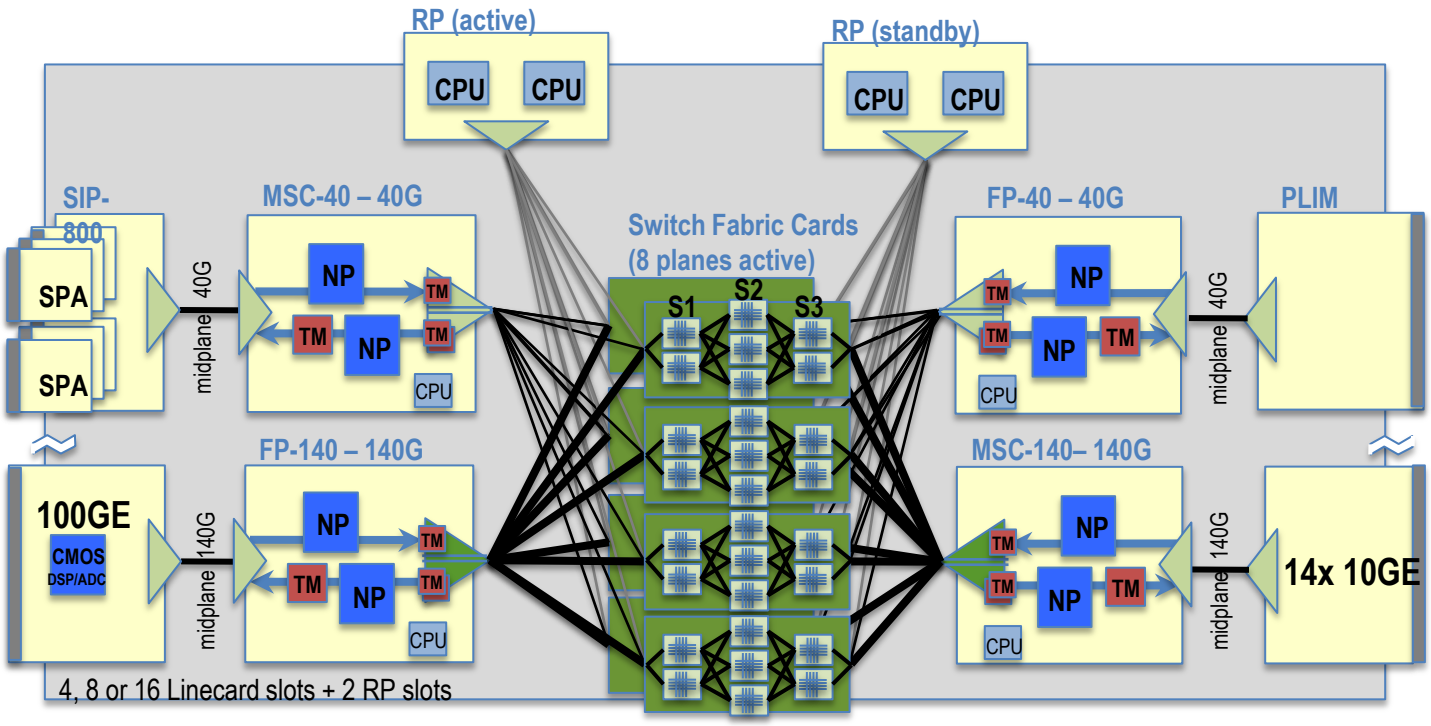
2010: Cisco CRS – 100G+ per slot

Next: 400G+ per slot (4x 100GE)

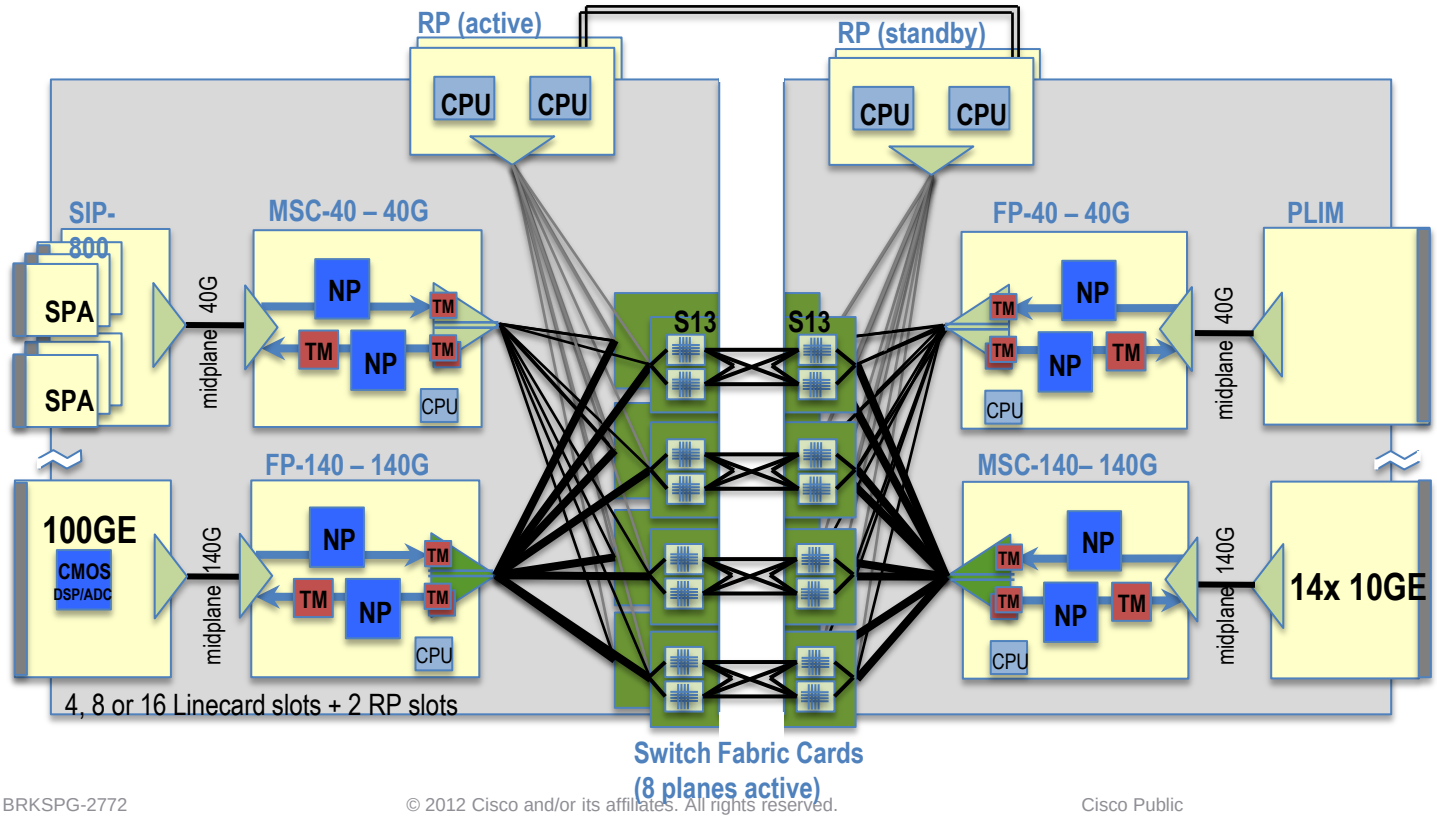
same backward-compatible architecture, same upgrade process



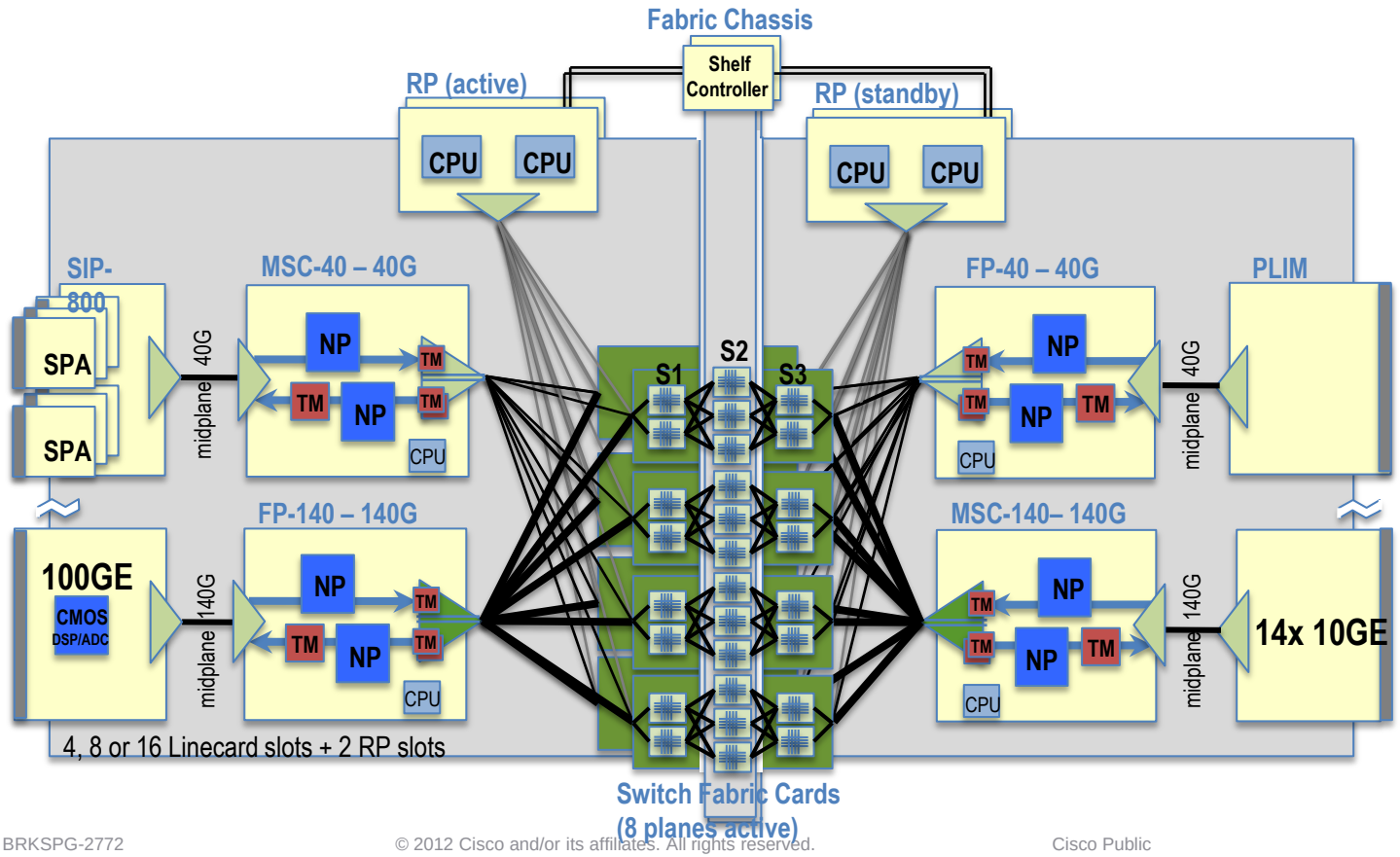
CRS Multi-Chassis



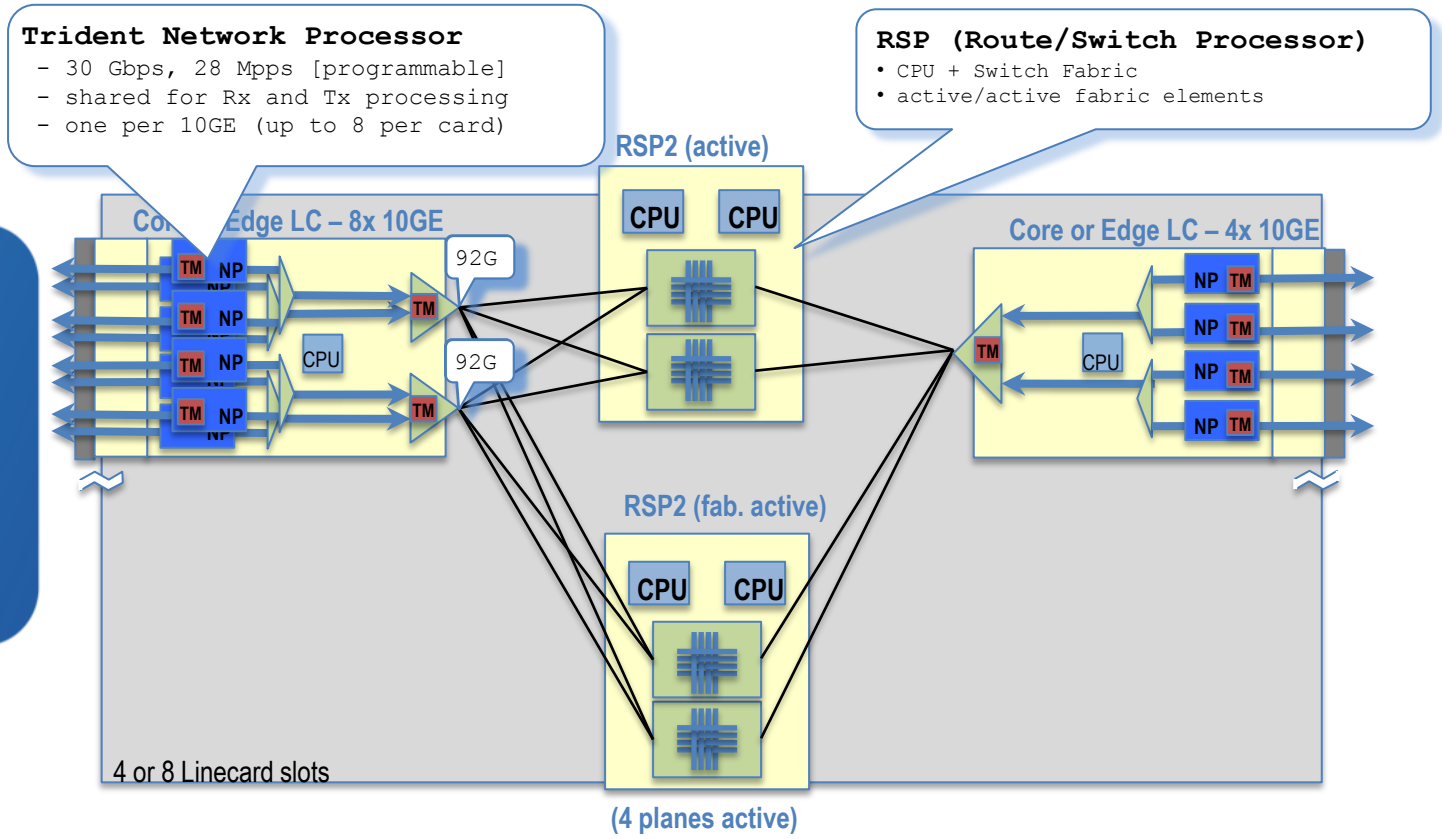
CRS Multi-Chassis (Back-to-Back, 2+0)



CRS Multi-Chassis (N+1, N+2, N+4)



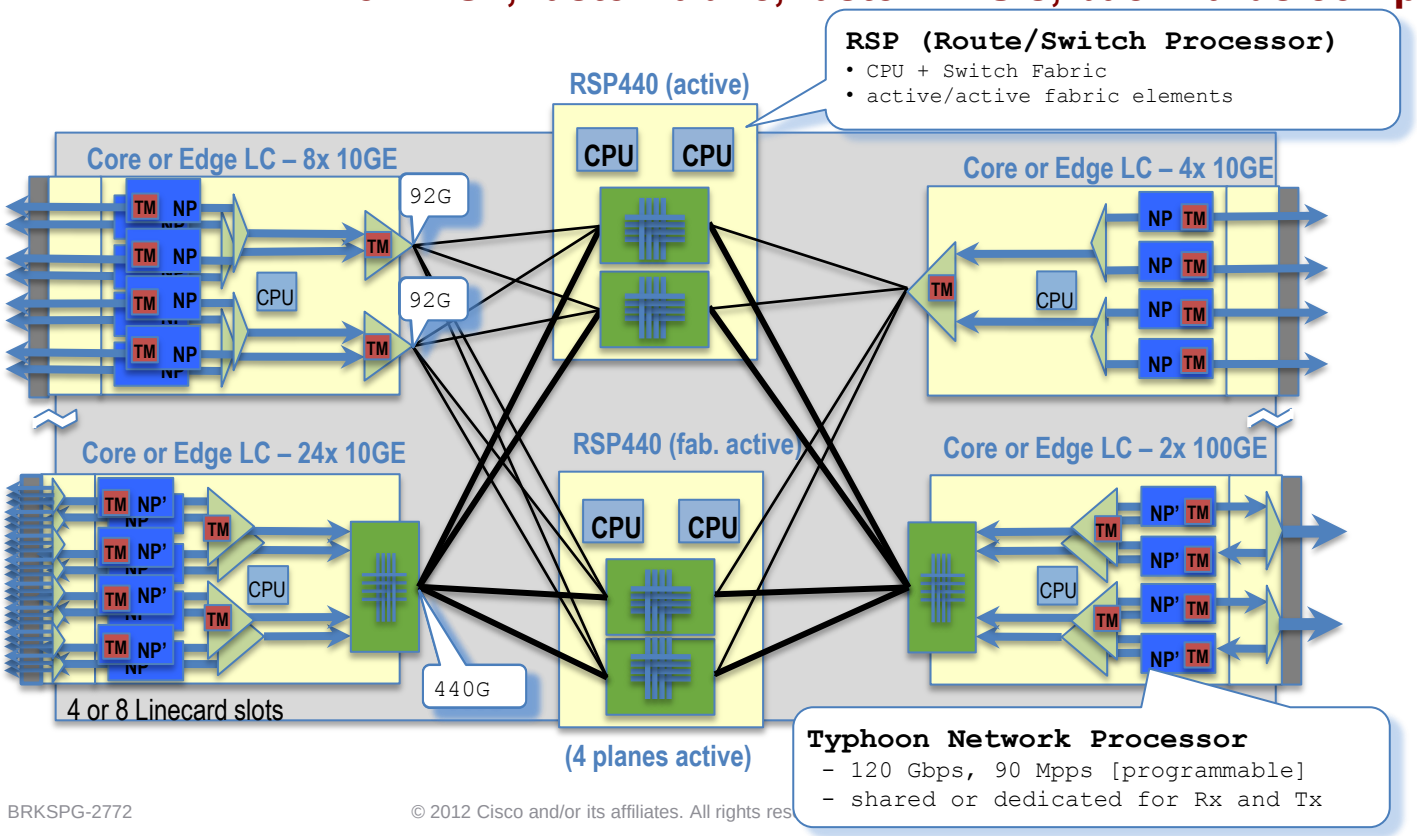
2009: Cisco ASR9000 – 80G+ per slot



2011: Cisco ASR9000 – 200G+ per slot

Next: 800G+ per slot

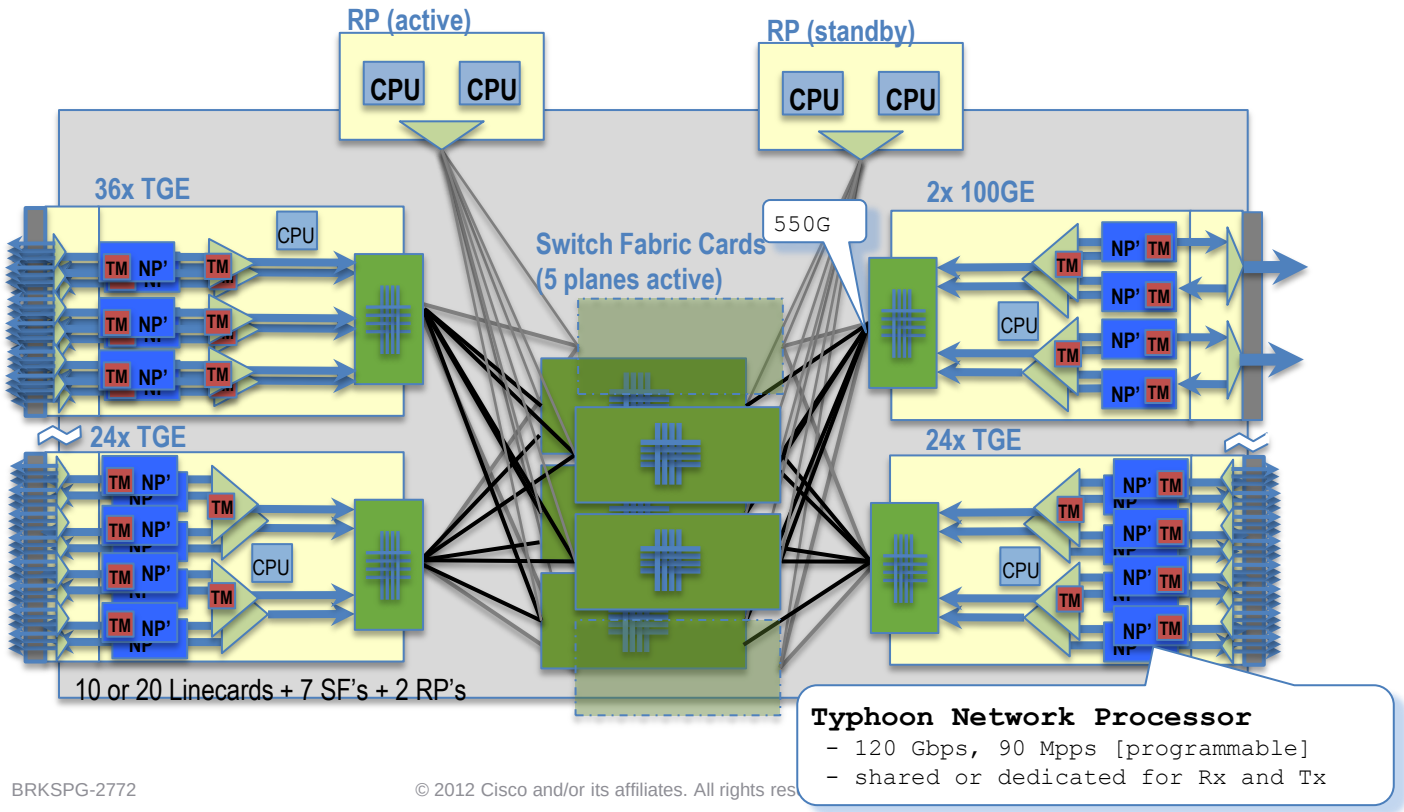
new RSP, faster fabric, faster NPU's, backwards compatible



2012: Cisco ASR9922 – 500+G per slot

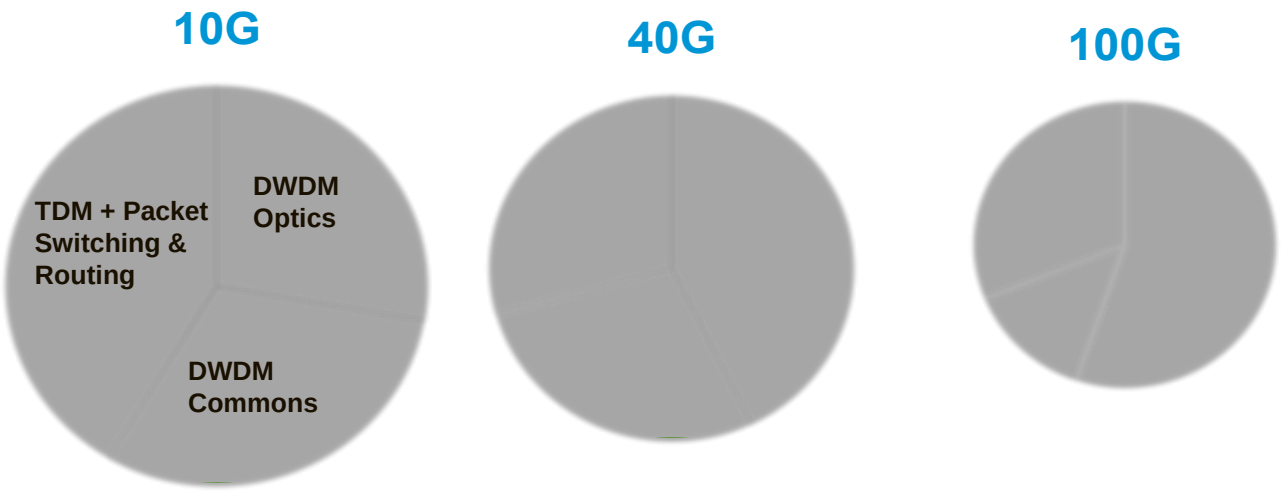
Next: 1.5T+ per slot

7 fabric cards, faster traces, faster NPU's, backwards compatible



Entering the 100GE world

Router port cost break-down



Core Routing Example

- 130nm (2004) → 65nm (2009): 3.5x more capacity, 60% less Watt/Gbps, ~8x less $/Gbps
- 40nm (2013): up to 1Tbps per slot, adequate Watt/Gbps reduction…

Silicon keeps following Moore's Law
Optics is fundamentally an analog problem

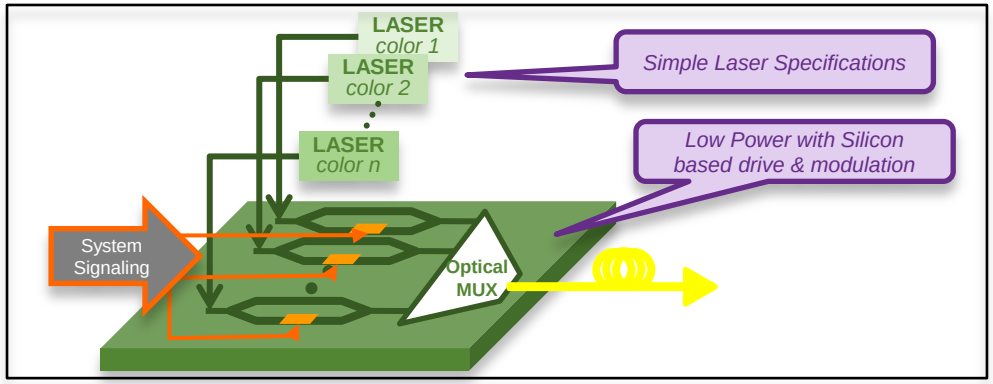
! Cisco puts 13% revenue (almost 6B$ annually) to R&D cca 20,000 engineers

Terabit per slot…

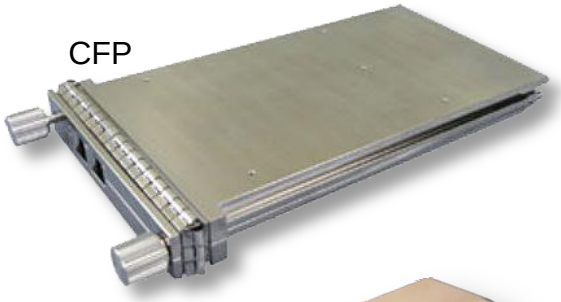
CMOS Photonics

What is CMOS Photonics?

- Silicon is semi-transparent for SM wavelengths
- Use case: Externally modulated lasers



CFP



CPAK

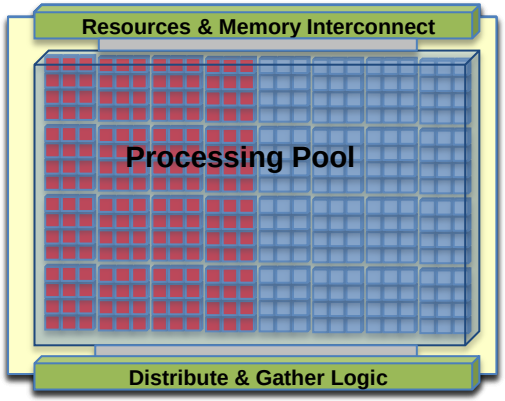
10x 100GBase-LR ports per slot

- 70% size and power reduction!
- <7.5W per port (compare with existing CFP at 24W)
- 10x 10GE breakout cable (100x 10GE LR ports per slot)

Terabit per-slot…

How to make it practically useful?

NPU Model (40nm)

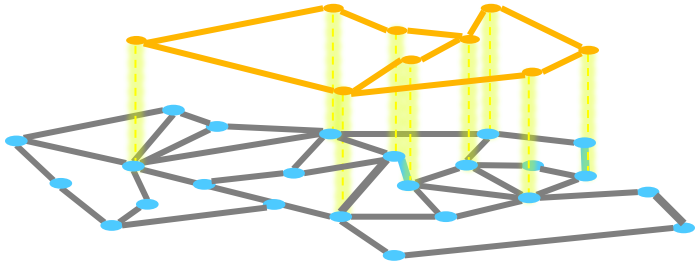


Silicon Magic @ 40nm

- Power zones inside the NPU – low power mode
- Duplicate processing elements – in-service u-code upgrade

Optical Magic @ 100G

- Single-carrier DP-QPSK modulators for 100GE (>3000km)
- CMOS Photonics
- ROADM



Data Plane Magic – Multi-Chassis Architectures

- nV Satellite
- DWDM and OTN shelves

Control Plane Magic – SDN (Software Defined Networks)

- nLight: IP+Optical Integration
- MLR – Multi-Layer Restoration
- Optimal Path & Orchestration (DWDM, OTN, IP, MPLS)

Evolution: Keeping up with Moore's Law

higher density = less W/Gbps, less $/Gbps

Switching Fabrics

- Faster, smaller, less power hungry
- Elastic – multi-stage, extensible
- Integrated – VOQ systems, arbiter systems, multi-functional fabric elements

Packet Processors

- 45nm, 40nm, 28nm, 20nm process
- Integrated functions – TM, TCAM, CPU, RLDRAM, OTN,…
- ASIC slices – Firmware ISSU, Low-power mode, Partial Power-off

Router Anatomy

- Control plane enhancements – SDN, IP+Optical
- DWDM density – OTN/DWDM satellites
- 100GE density – CMOS optics, CPAK/CFP4
- 10GE density (TGE breakout cables) and GE density (GE satellites)

Thank You.