Cisco Expo 2009

Anatomy of Core Network Elements

from 1Gbps to 10Tbps

Josef Ungerman

CSE, CCIE#6167

© 2009 Cisco Systems, Inc. All rights reserved.

Agenda

1. Basic Terms
2. Router Architectures
3. Switch Architectures
4. Hybrid Architectures
5. Network Processors
6. Switch Fabrics

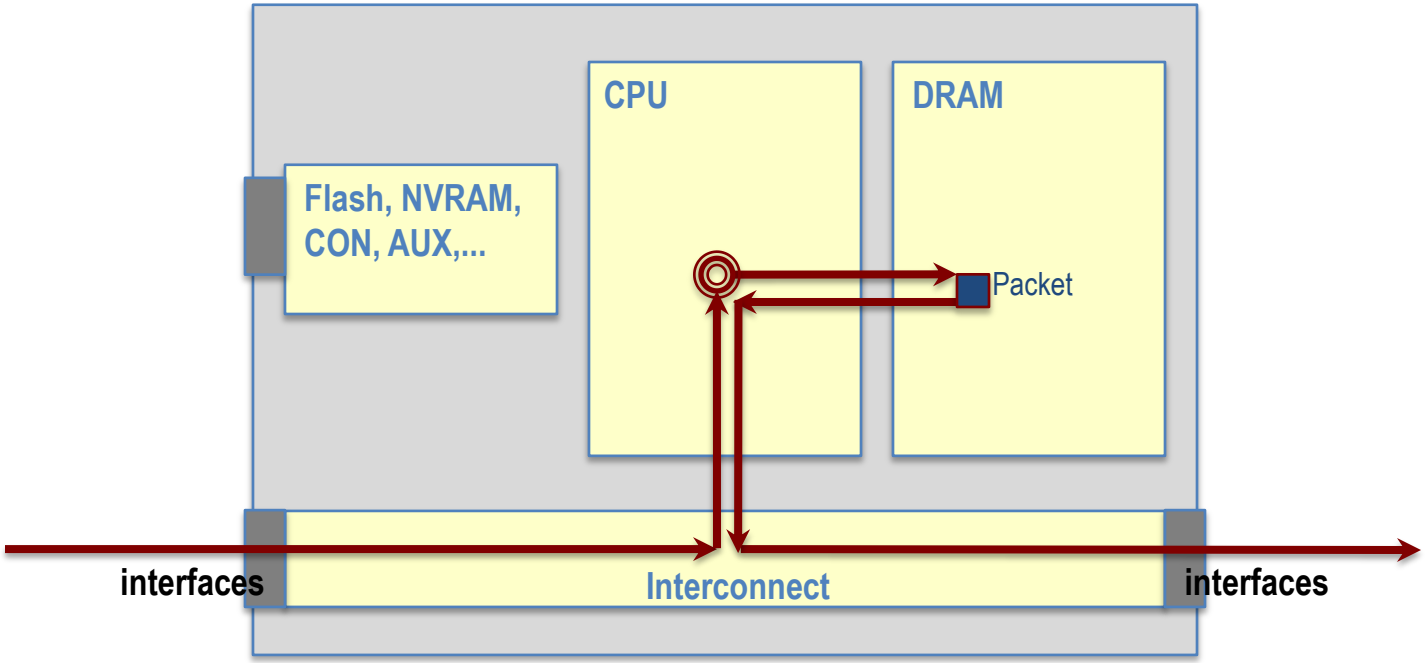
© 2009 Cisco Systems, Inc. All rights reserved.

Basic Terms

Chapter 1

© 2009 Cisco Systems, Inc. All rights reserved.

Cisco in 80's: Router Architecture

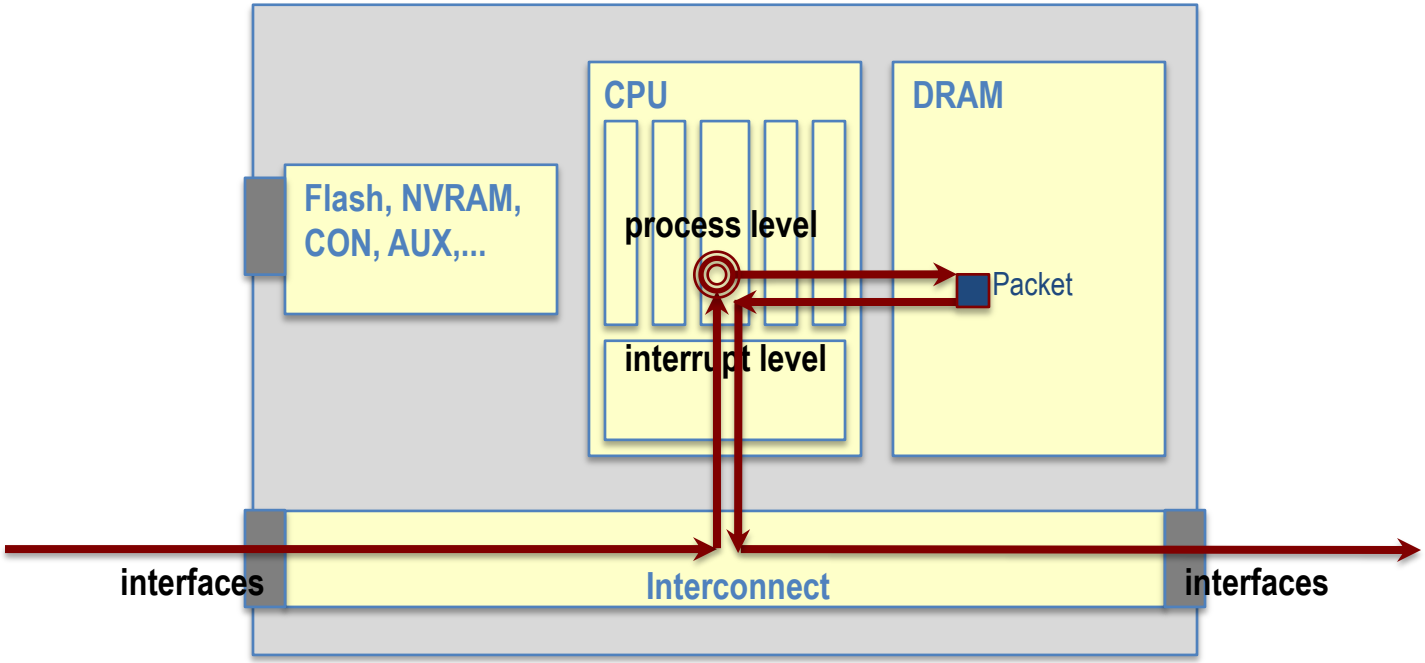


Store & Forward Switching – using packet buffers and QoS, handles WAN interfaces (very variable interface speeds)

© 2009 Cisco Systems, Inc. All rights reserved.

Real-Time Packet Processing

Process Switching

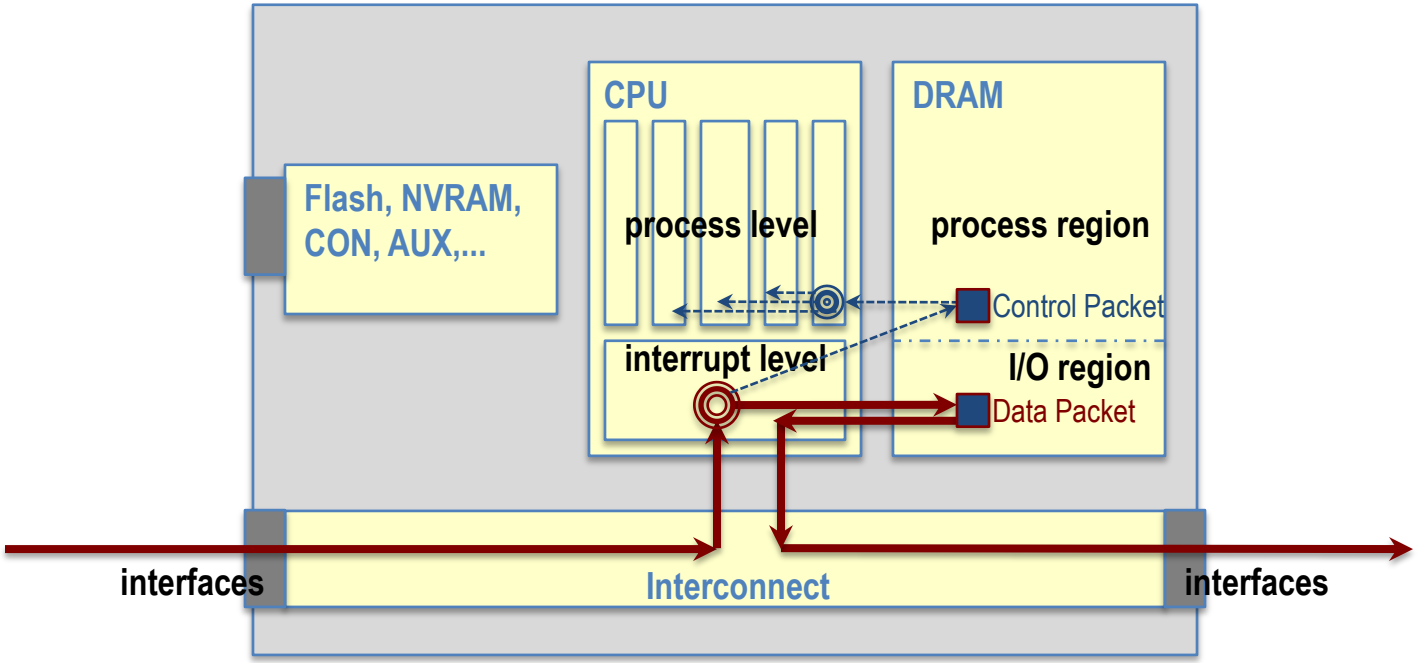


Process Switching – IOS Process handles the forwarding decision and other operations with the packet

© 2009 Cisco Systems, Inc. All rights reserved.

Real-Time Packet Processing

Data Plane vs. Control Plane



Data Plane – transit packets (aka. fast path)

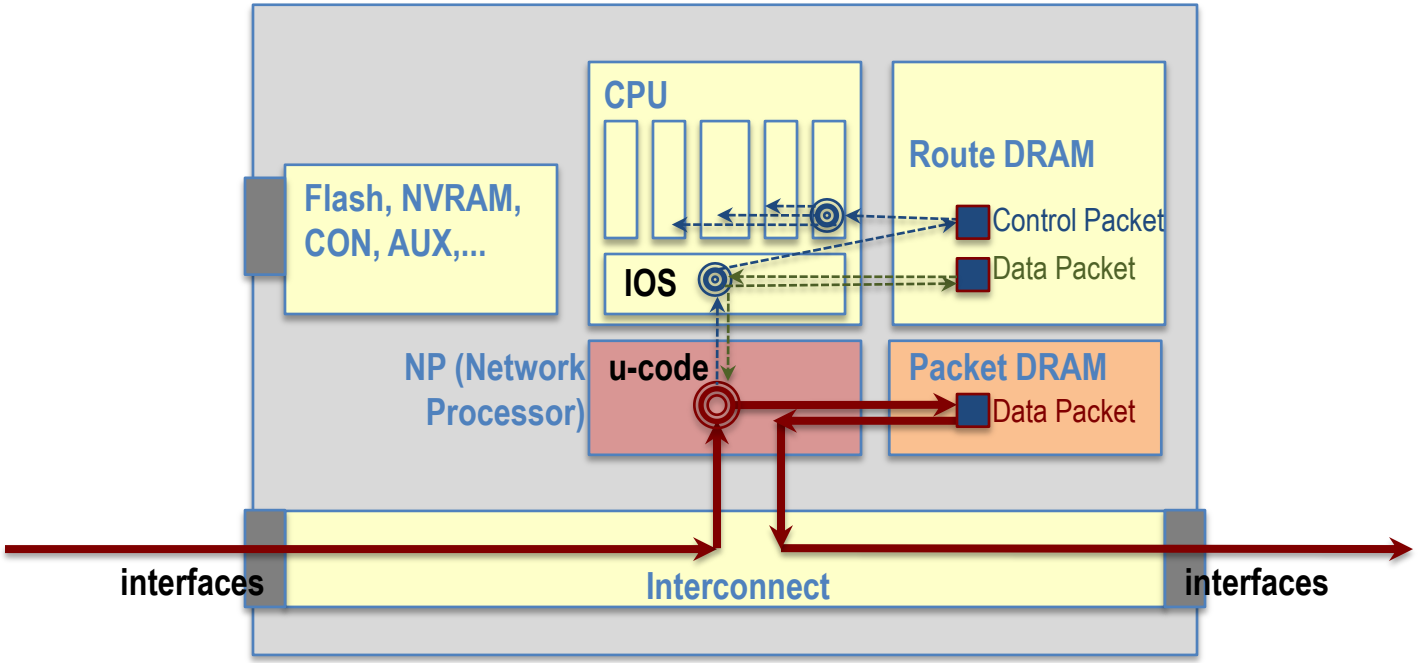
Control Plane – packets for the router (routing, management, exceptions)

- routing/control plane = routing and vital functions (OSPF, BGP, LDP, NTP, keepalives,...)
- management plane = access to the router (telnet, ssh, SNMP, ...)

© 2009 Cisco Systems, Inc. All rights reserved.

Real-Time Packet Processing

NP (Network Processor) – S/W vs. H/W router



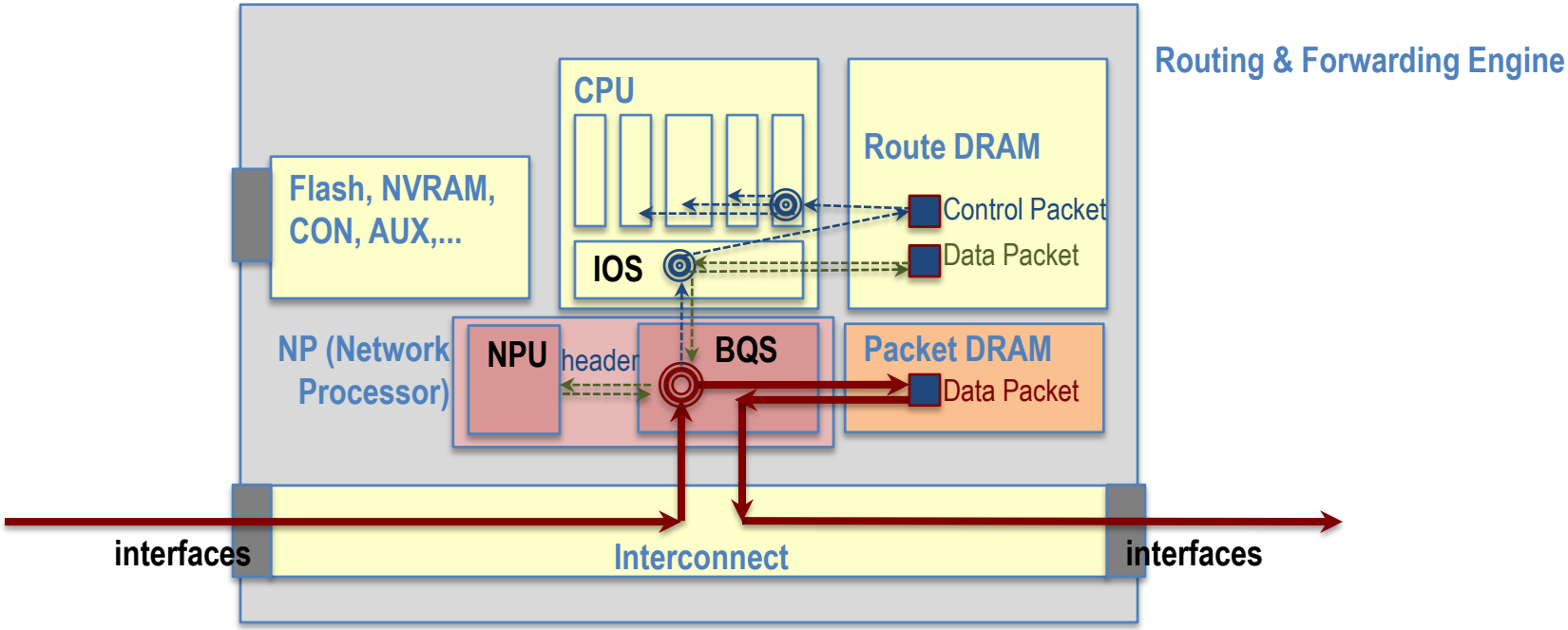
NP (Network Processor) – NP handles the data plane, not IOS (platform-dependent)
CPU – runs IOS – handles only the control plane (platform-independent)

Slow Path – IOS on the CPU can still forward some packets that NP cannot handle (eg. exceptions, non-IP protocols routing, unsupported features)

© 2009 Cisco Systems, Inc. All rights reserved.

Real-Time Packet Processing

NP (Network Processor) – S/W vs. H/W router



BQS (Buffering, Queuing, Scheduling) or TM (Traffic Manager) ASIC

– handles the memory access and QoS (packet body)

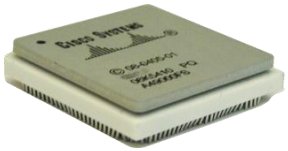
NPU (Network Processing Unit)

– handles only packet forwarding and operations (packet header)

© 2009 Cisco Systems, Inc. All rights reserved.

Summary – what is inside the router?

BBB – basic building blocks



- **Processor**
  - control-plane – OS processor
  - data-plane – network processor

- **Memory**
  - DRAM for OS memory and packet buffers
  - SRAM for caches
  - TCAM for fast lookups

- **Interconnects**
  - bus
  - serial link
  - switch fabric

We do not care about what is visible on the router

- chassis, fans, power supplies
- control ports – CON, AUX, BITS, Alarms, Disks
- data ports – LAN and WAN interfaces

© 2009 Cisco Systems, Inc. All rights reserved.

“It is always something

(corollary). Good, Fast, Cheap:

Pick any two (you can’t have all three).”

RFC 1925

“The Twelve Networking Truths”

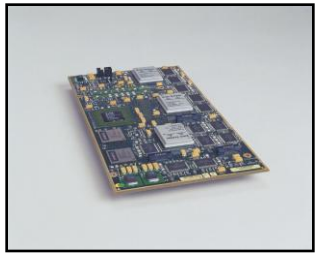
© 2009 Cisco Systems, Inc. All rights reserved.

Packet Processing Technology Primer

Performance vs. Flexibility

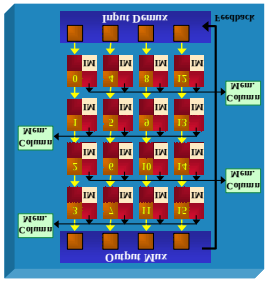
CPU (Central Processing Unit)

- multi-purpose processors (CISC, RISC)
- high s/w flexibility [weeks]
- low performance stability [cca 1Mpps today]
- **usage example:** access routers (ISR's)



ASIC (Application Specific Integrated Circuit)

- mono-purpose hard-wired functionality
- low engineering flexibility [2 years]
- high performance stability [over 200 Mpps today]
- **usage example:** switches (Catalysts), core routers



NP (Network Processor) = "something in between"

- performance + programmability
- moderate s/w flexibility [months]
- moderate and stable performance [4Mpps – 40 Mpps+]
- can be expensive, power-hungry, can have low code memory
- **usage**: fast feature-rich edge and aggregation

© 2009 Cisco Systems, Inc. All rights reserved.

Memory Technology Primer

Capacity vs. Access Speed

Two basic memory technologies are in use today:

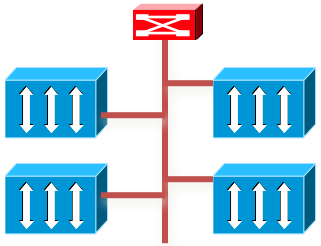
- **Static RAM** (SRAM, SSRAM)
- **Dynamic RAM** (DRAM, EDO DRAM, SDRAM, DDR)

SRAM	DRAM
High Power	**Low Power**
High Speed [10-20ns]	**Low Speed** [40-60ns]
Low Density [eg. 16M per chip]	**High Density** [eg. 1G per chip]

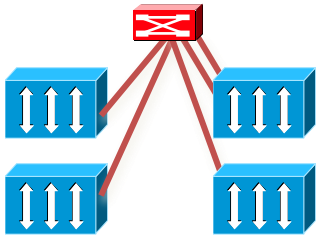
© 2009 Cisco Systems, Inc. All rights reserved.

Interconnects Technology Primer

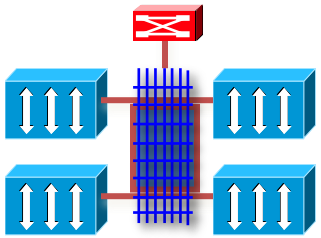
Capacity vs. Complexity



- **Bus**
  - half-duplex, shared medium
  - for example PCI [800Mbps to 25Gbps+ today]
  - simple and cheap



- **Serial Lane (Point-to-Point Link Set)**
  - dedicated, unidirectional or full-duplex line
  - for example SPI4.2 [11.2Gbps+ today]



- **Switching Fabric (cross-bar, exchange)**
  - non-blocking, full-duplex, any-to-any
  - for example GSR, CRS [40Gbps to 9.6Tbps+ today]

© 2009 Cisco Systems, Inc. All rights reserved.

Example: Lookup Problem

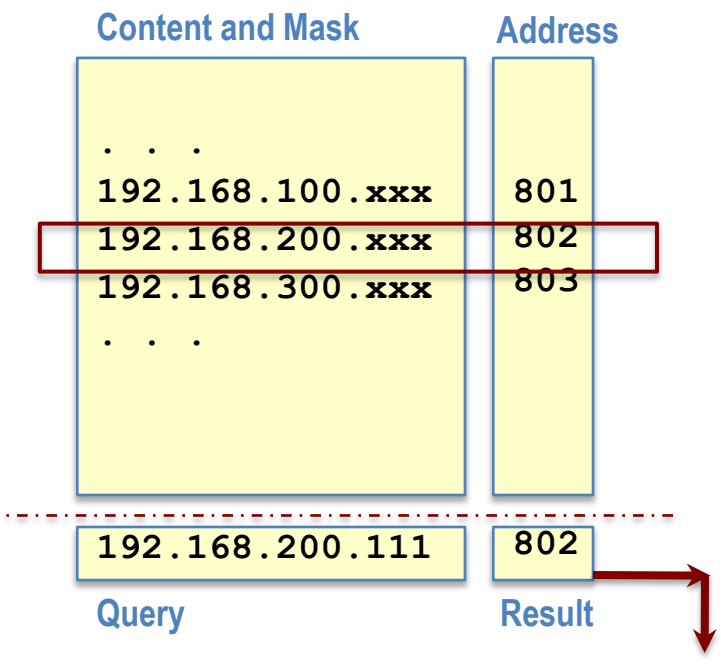
memory vs. processing

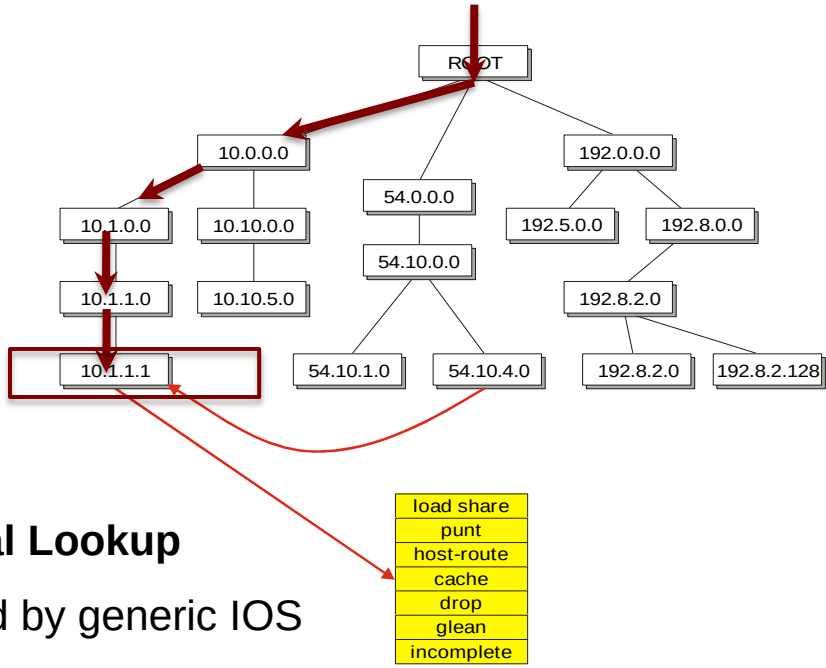
TCAM (Ternary Content Addressable Memory)

SRAM with a comparator at each cell

1 step – very fast, but very expensive

parallel, order independent lookups
(ACL, QoS, Netflow, even FIB)





Tree or Serial Lookup

8-8-8-8 used by generic IOS

16-8-8 used by the C12000, 11-8-5-8 used by C10K

memory vs. speed tradeoff!

- could be 8-1-1-1-1-1-1-1-1-1-1-1... (low SRAM)
- could be used also for ACL, uRPF, accounting

© 2009 Cisco Systems, Inc. All rights reserved.

Router Anatomy

Chapter 2

© 2009 Cisco Systems, Inc. All rights reserved.

Fundamental Building Blocks

Module, card

I/O module (hardware module with I/O interfaces

Switch Fabric (any-to-any full-duplex switching element)

Mux/demux, fabric interface (typically including a tiny buffer)

F

Forwarding ASIC (a complex of hardware elements and SRAM's handling data plane)

Q

Queuing ASIC (BQS – Buffering/Queuing/Scheduling, TM – Traffic Manager, etc.)

NP

NP, Network Processor (programmable hardware element handling data plane)

buff.

Packet buffering, packet memory, QoS point

IOS

Control Plane element : CPU + DRAM + Flash + NVRAM and control interfaces. CPU (Central Processing Unit) is a general purpose micro-processor running the OS (Operating System)

Simplex serial link set

Duplex serial link set

Active and backup backplane connection (serial duplex link set)

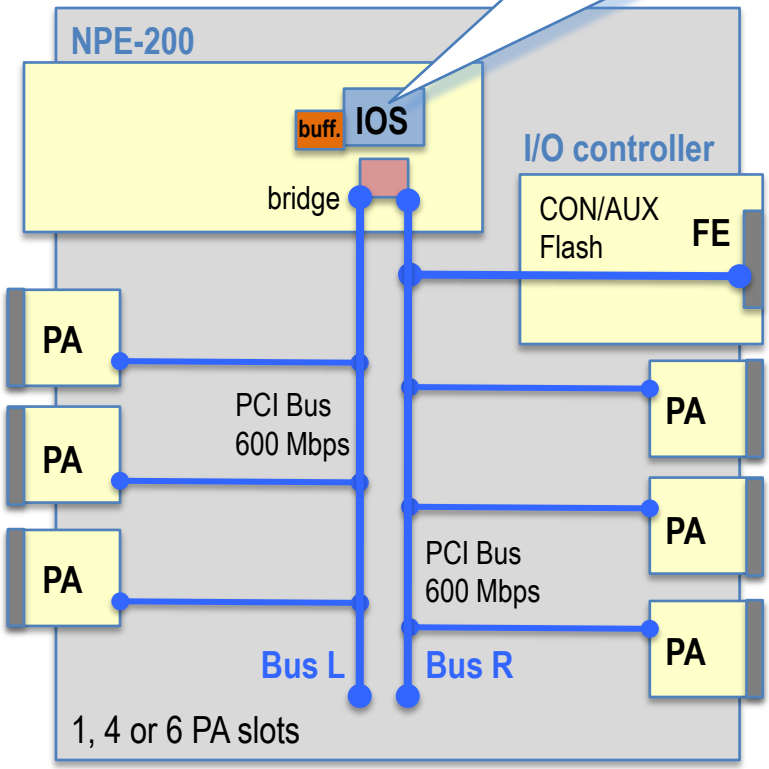
Bus

© 2009 Cisco Systems, Inc. All rights reserved.

Cisco 7200 (1990's)

architecture

Software Router
- data plane = IOS interrupt level
- control plane = IOS processes

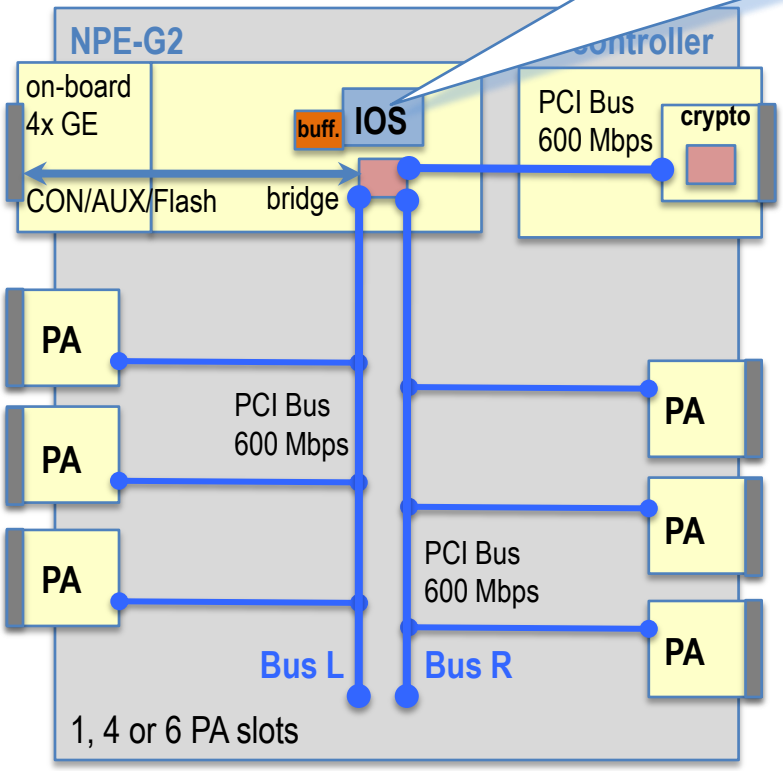


© 2009 Cisco Systems, Inc. All rights reserved.

Cisco 7200 – NPE-G1/G2 upgrade

architecture

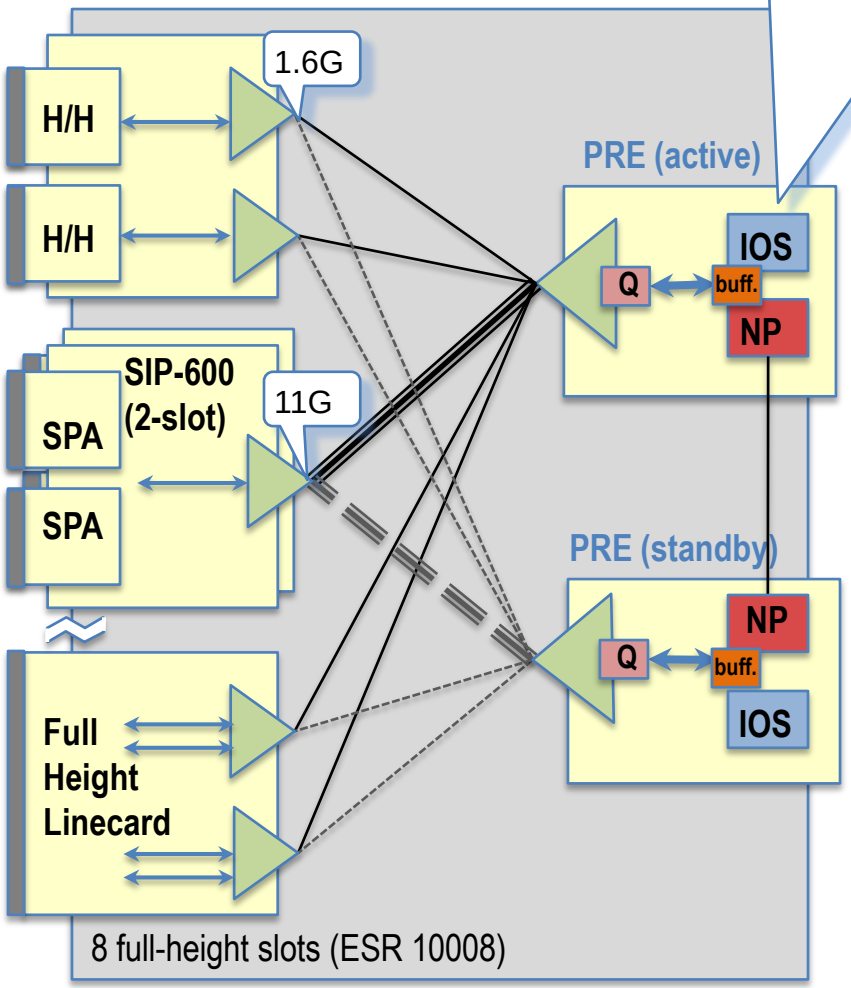
Software Router
- no change, just faster CPU/memory



© 2009 Cisco Systems, Inc. All rights reserved.

Cisco ESR10000
architecture

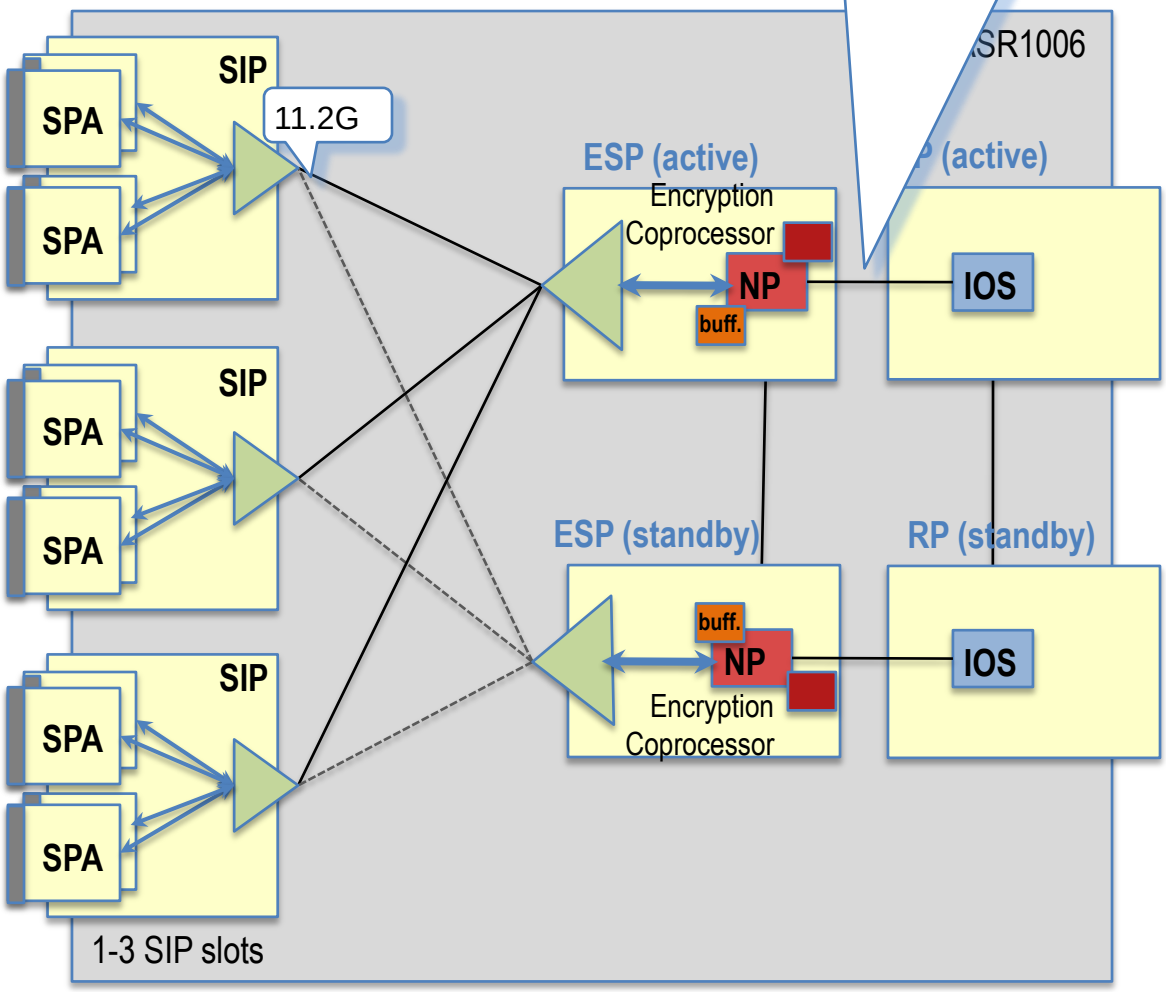
Hardware Router
- data plane = PXF chip (u-code)
- control plane = CPU with IOS
- DMA chip for packet memory



© 2009 Cisco Systems, Inc. All rights reserved.

Cisco ASR1000
architecture

Split data and control plane
- RP = control-plane only
- ESP = data-plane (QFP chip)
  20Gbps, 16Mpps, C-programmable



© 2009 Cisco Systems, Inc. All rights reserved.

“It is more complicated than you think.”

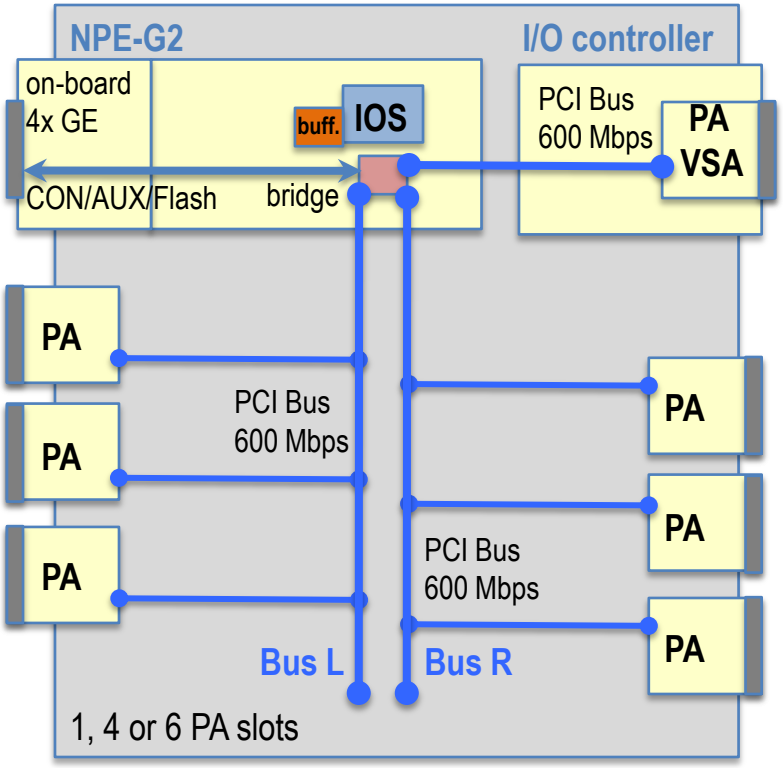
RFC 1925

“The Twelve Networking Truths”

© 2009 Cisco Systems, Inc. All rights reserved.

Cisco 7200: centralized single processor architecture

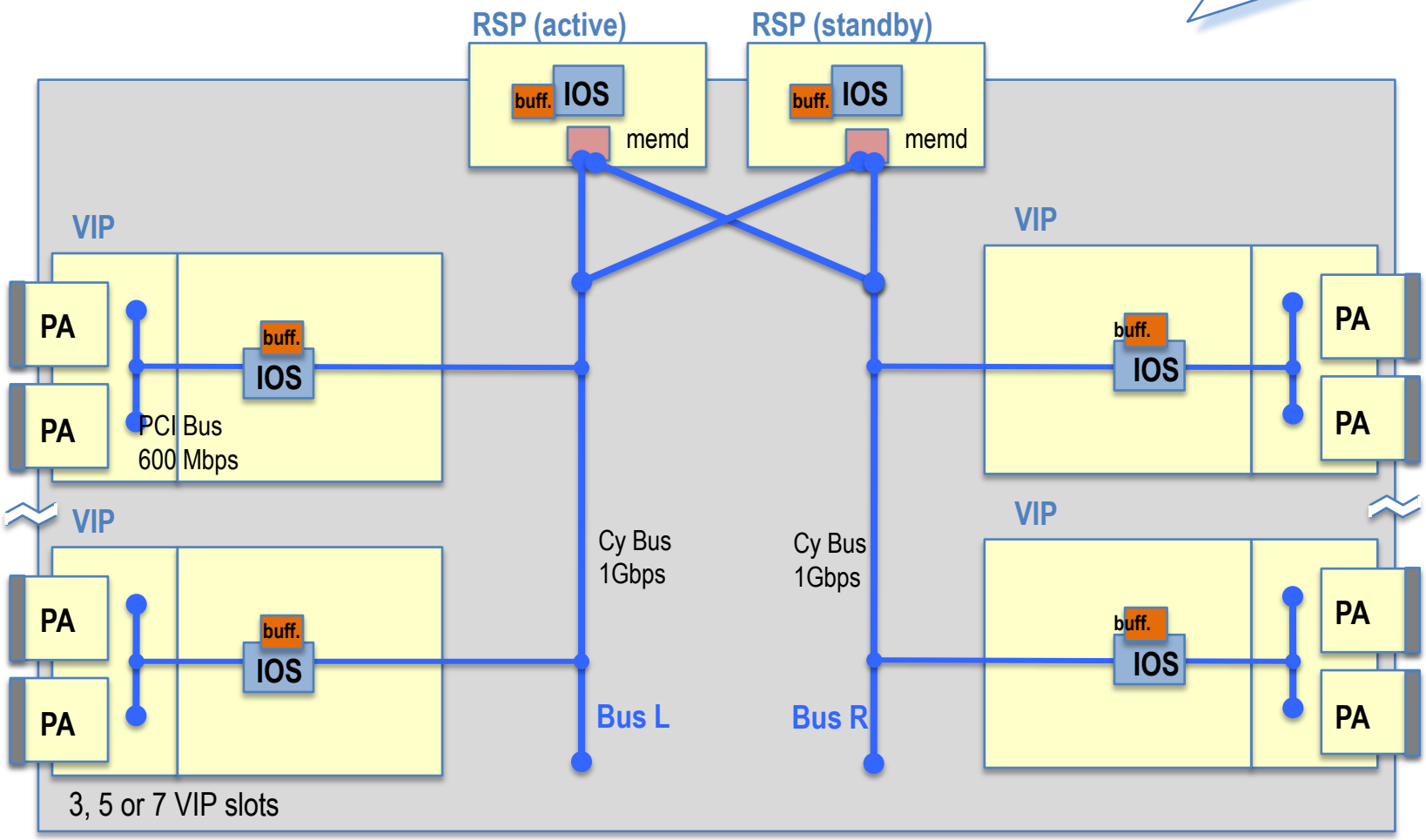
Single-Processor
• one CPU for everything



© 2009 Cisco Systems, Inc. All rights reserved.

Cisco 7500: distributed multi-processor architecture

Multi-Processor
• distributed, parallel CPU's

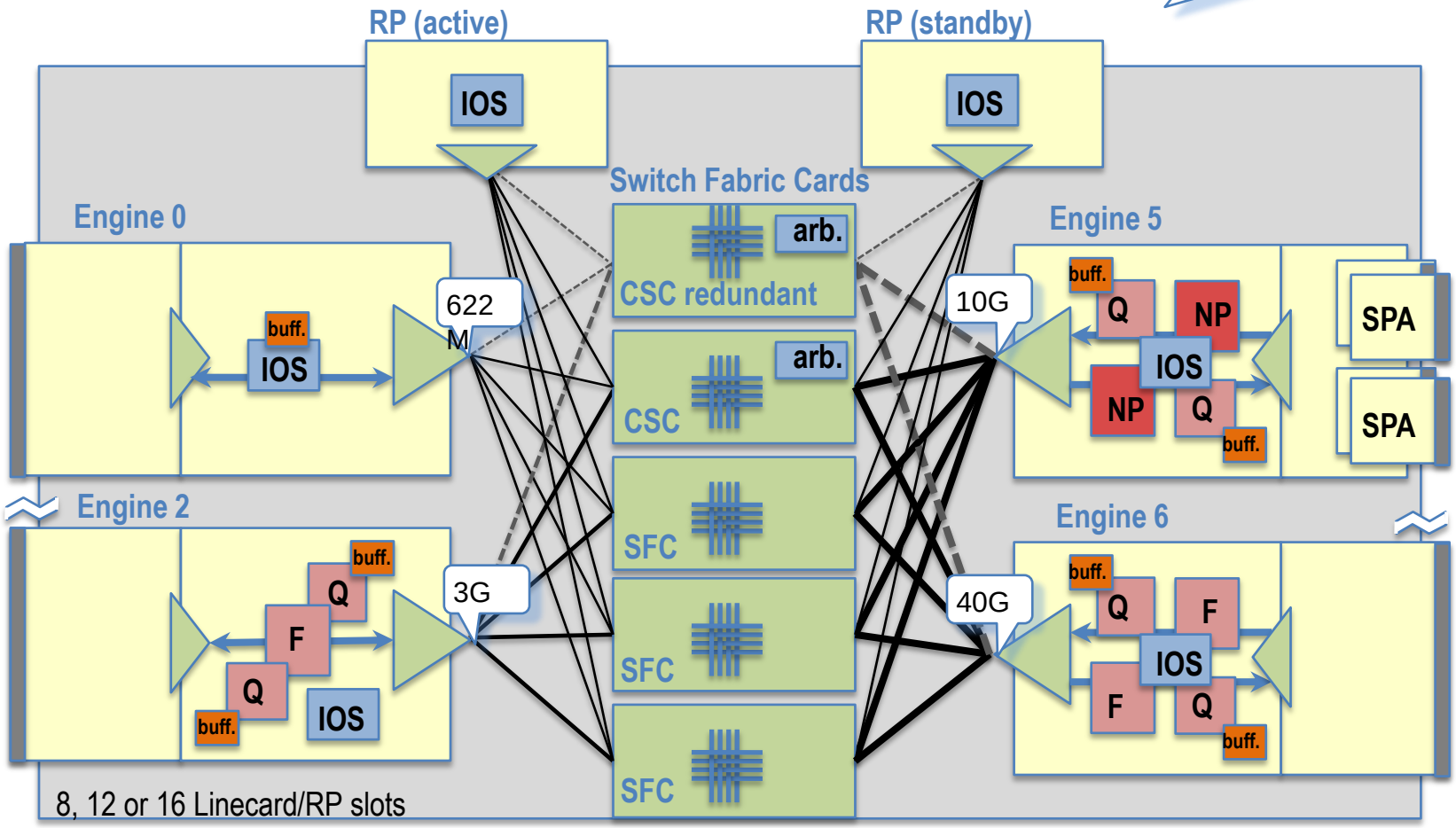


© 2009 Cisco Systems, Inc. All rights reserved.

Cisco 12000 – switch fabric

architectural evolution

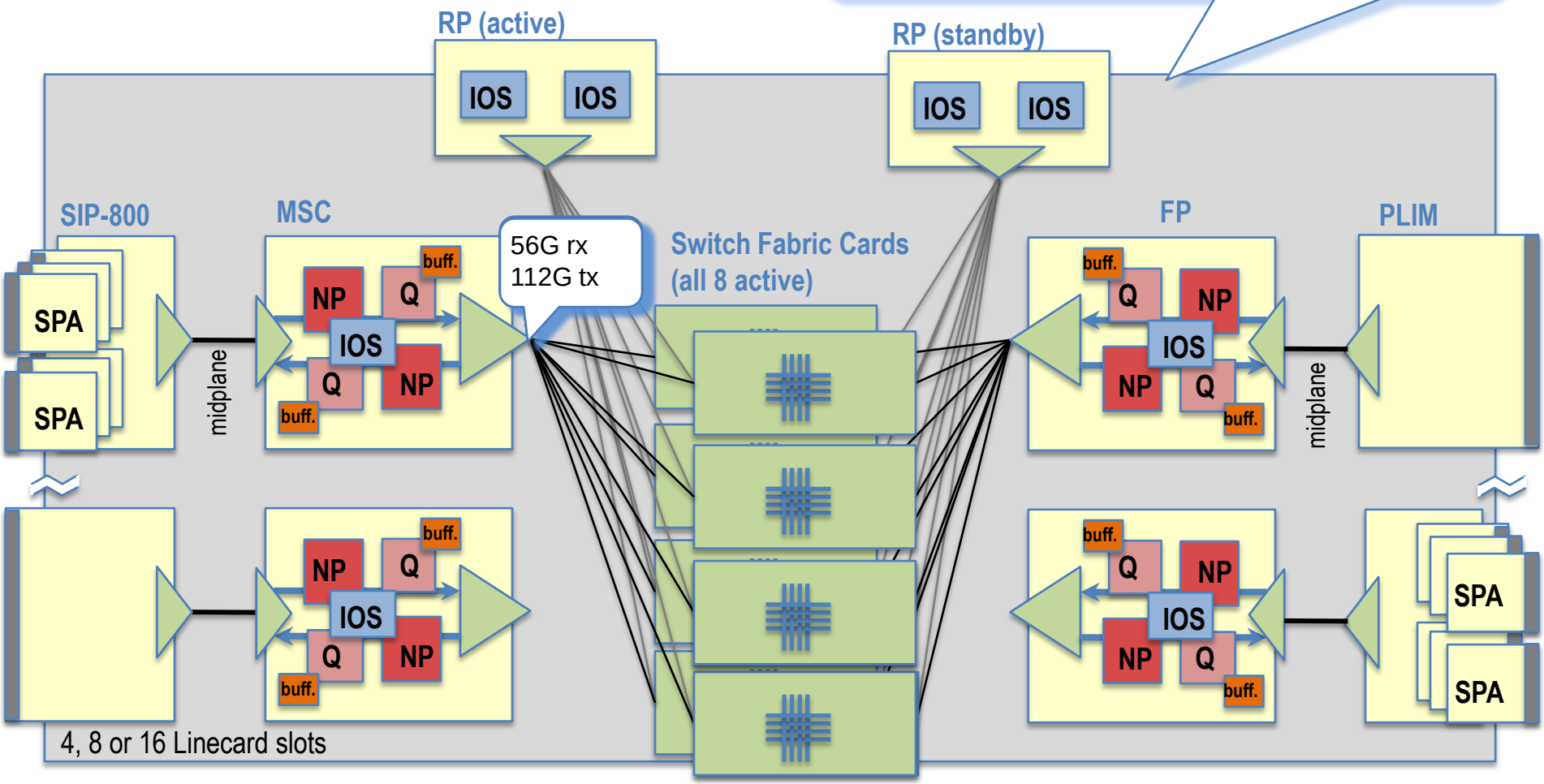
Distributed Forwarding Architecture
• up to 600Gbps today



© 2009 Cisco Systems, Inc. All rights reserved.

Cisco CRS-1
architecture

Distributed Forwarding Architecture
- up to 1.28Tbps today
- 8 multi-chassis: 10.24Tbps today

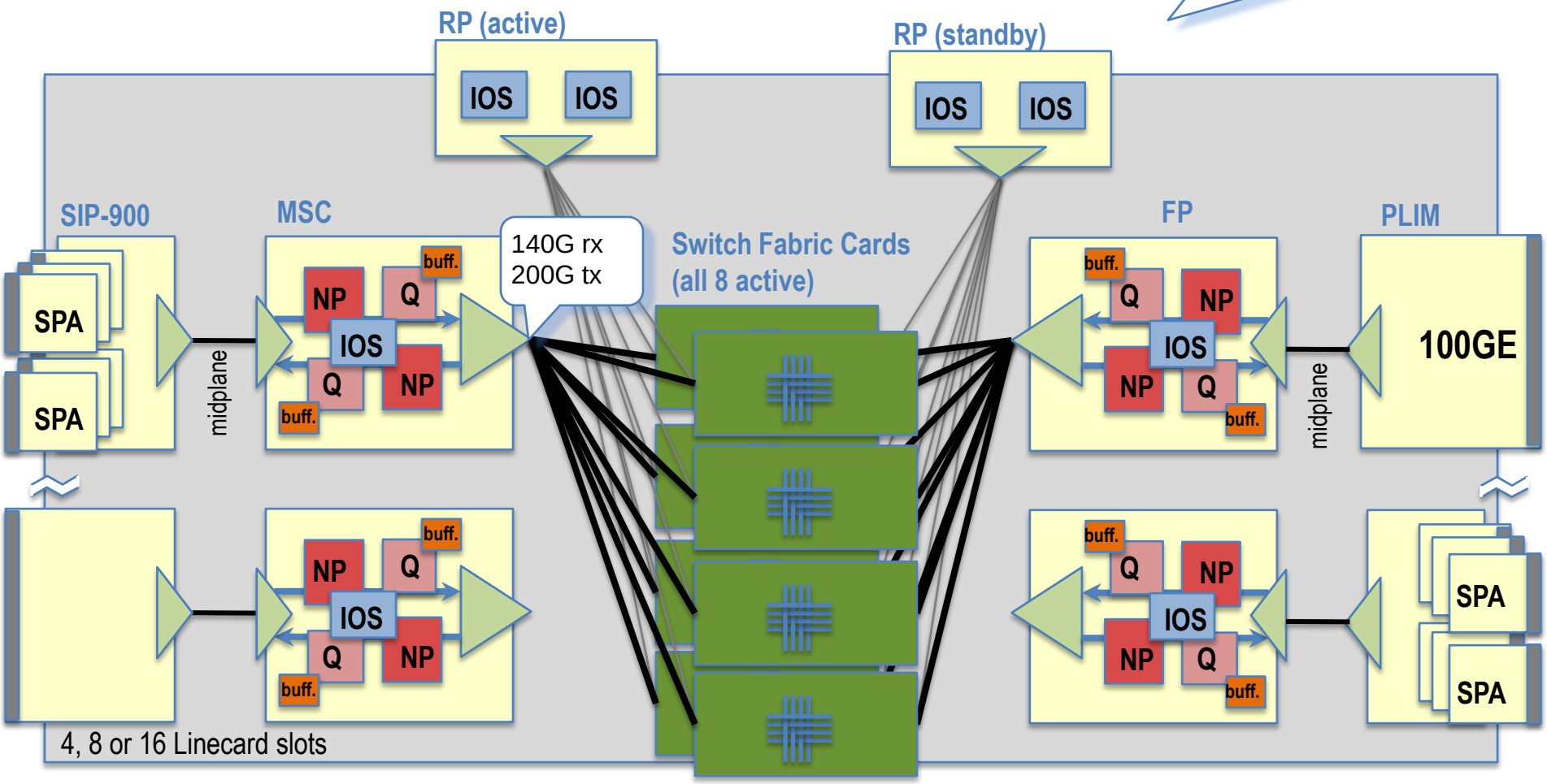


© 2009 Cisco Systems, Inc. All rights reserved.

Cisco CRS-1

evolution to 100GE

Distributed Forwarding Architecture
• up to 3.84Tbps single node in 2010



© 2009 Cisco Systems, Inc. All rights reserved.

Switch Anatomy

Chapter 3

© 2009 Cisco Systems, Inc. All rights reserved.

Switch vs. Router

- **Definition 1** (old): switch is L2, router is L3

  UNTIL PEOPLE INVENTED L3 SWITCH...

- **Definition 2:**
  - switch is optimized for LAN – FE, GE, 10GE
  - router is optimized for WAN – many speeds 64Kbps – 40Gbps

  UNTIL ETHERNET CAME TO WAN/CORE TOO...

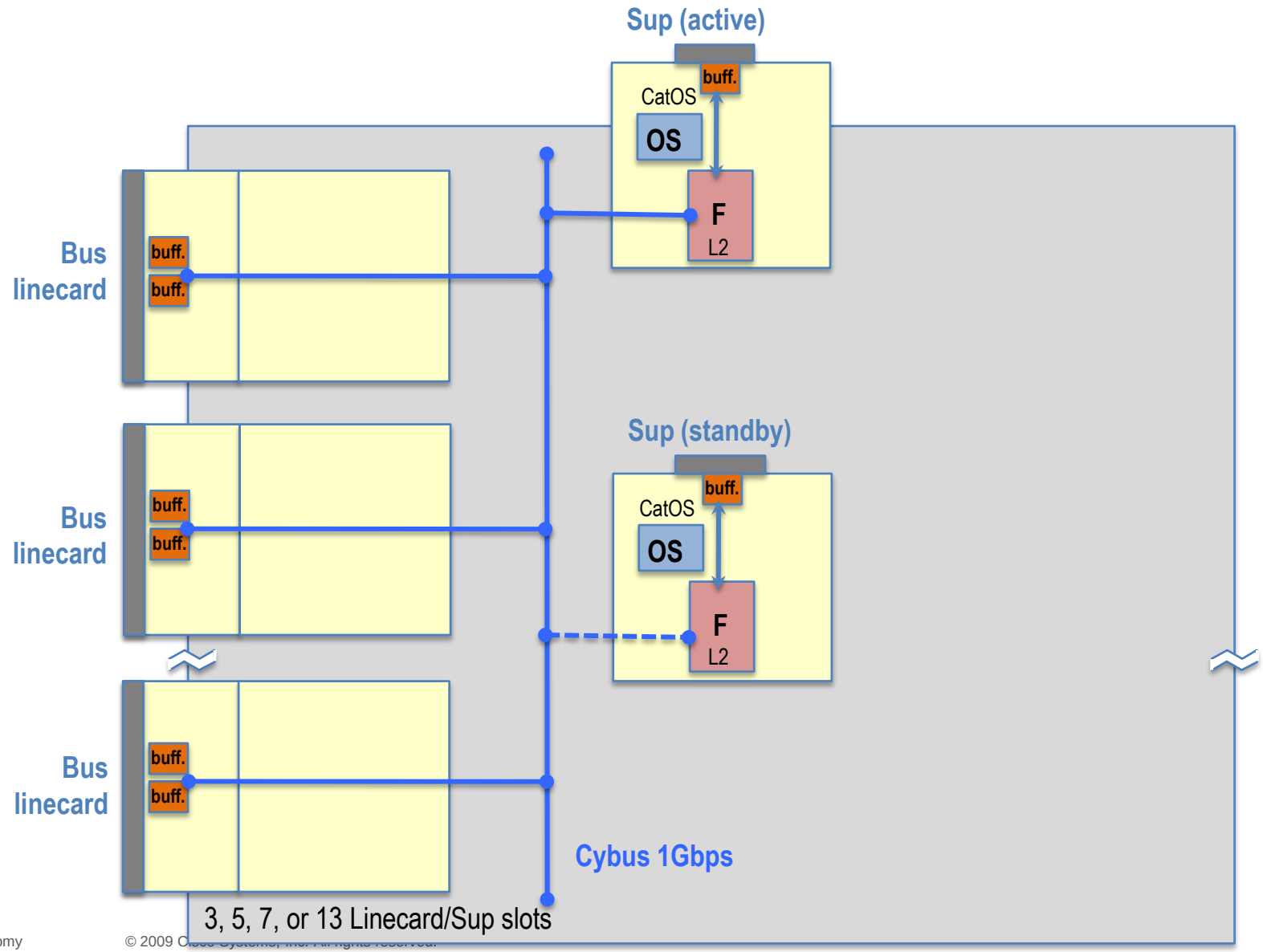
- **Definition 3:**
  - switch is optimized for speed (fixed functionality)
  - router is optimized for functionality (programmable, modular)

	SWITCH	ROUTER
Programmable Forwarding Engine	NO	YES
Packet Buffers	shallow, max. ~16MB	deep, hundreds of MB
Queues and Shapers	4-8 per port	thousands per port
Hardware Modularity	1-level	3-level+

© 2009 Cisco Systems, Inc. All rights reserved.

Catalyst 5000 (90's)

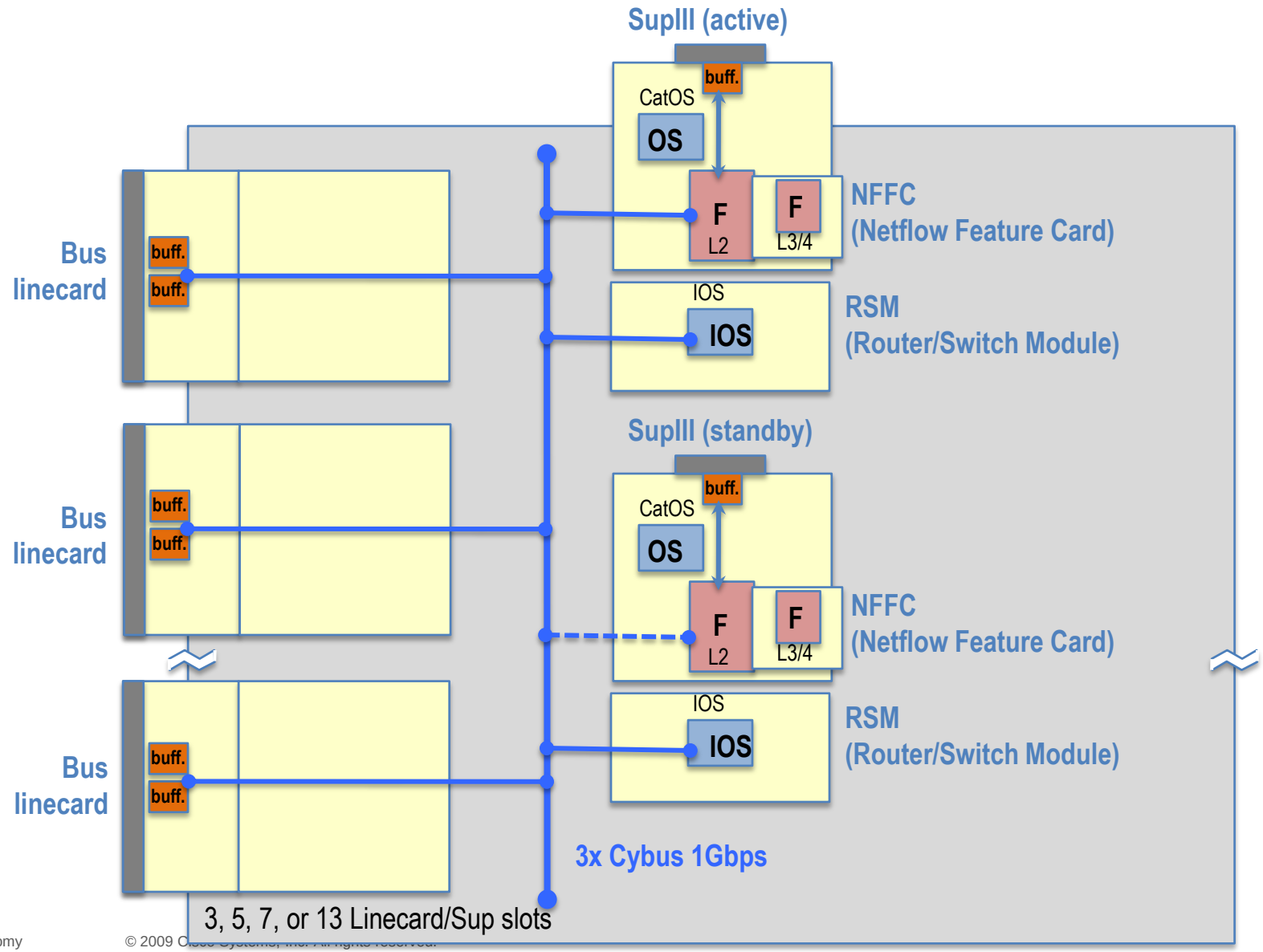
centralized switching architecture



© 2009 Cisco Systems, Inc. All rights reserved.

Catalyst 5500

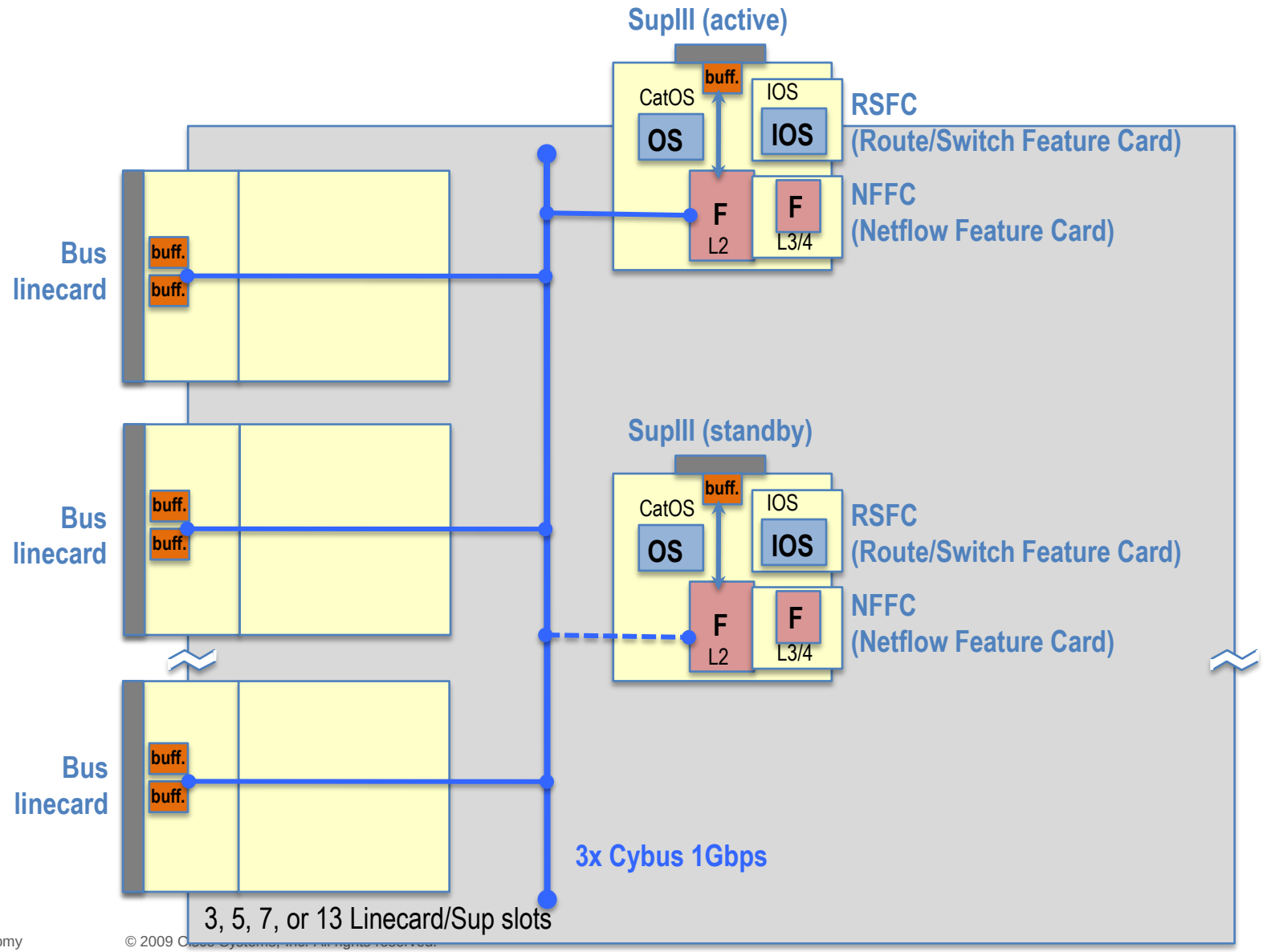
Netflow Switching: NFFC + RSM



© 2009 Cisco Systems, Inc. All rights reserved.

Catalyst 5500

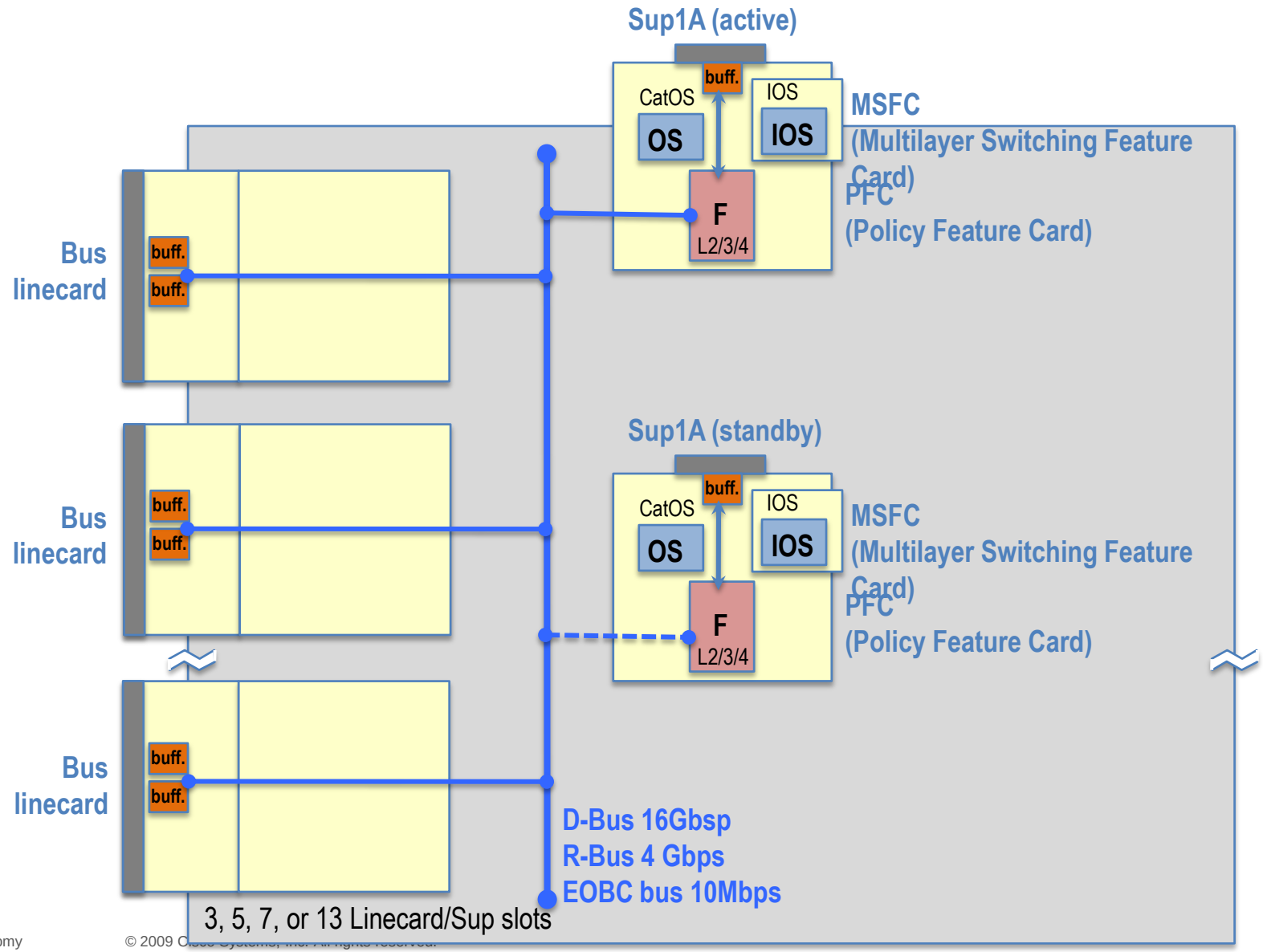
Netflow Switching: NFFC + RSFC



© 2009 Cisco Systems, Inc. All rights reserved.

Catalyst 6000

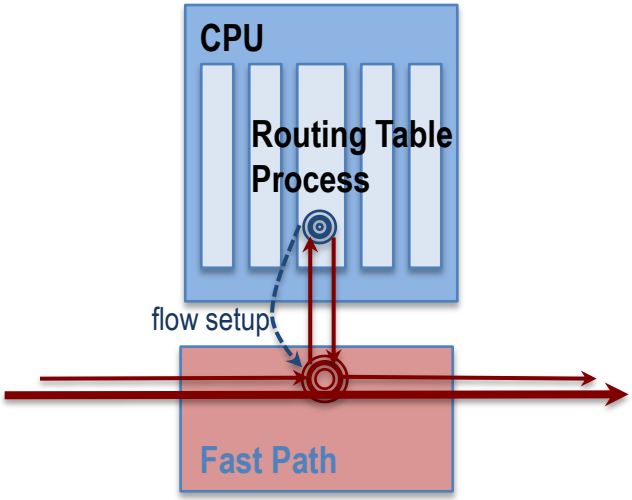
MLS (Multilayer Switching): PFC + MSFC

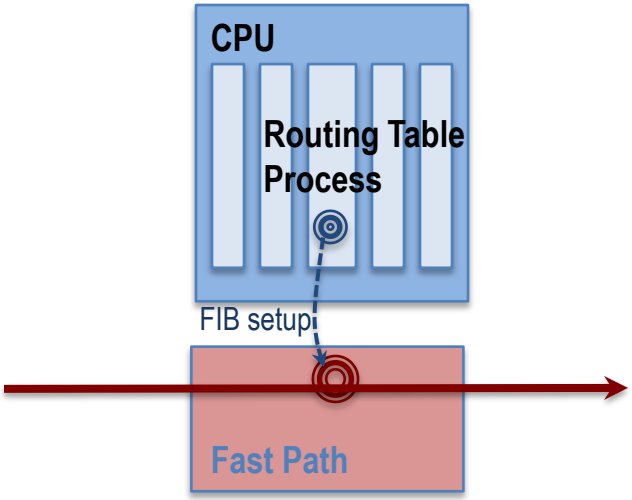


© 2009 Cisco Systems, Inc. All rights reserved.

Real-Time Packet Processing

Fast-Path Forwarding (aka. Switching)





Flow Switching (aka. Fast Switching)

- the 1st packet of a flow goes to slow-path
- the flow paths sets up the fast cache entry
- next packets of the flow take the fast-path

FIB Switching (aka. CEF)

- RIB (Routing Information Base)
- FIB (Forwarding Information Base)
- RIB is mirrored to fast-path FIB first, before any packets are switched
- traffic doesn't collide with control plane

© 2009 Cisco Systems, Inc. All rights reserved.

“Every old idea will be proposed again with a different name and
a different presentation, regardless of whether it works.”

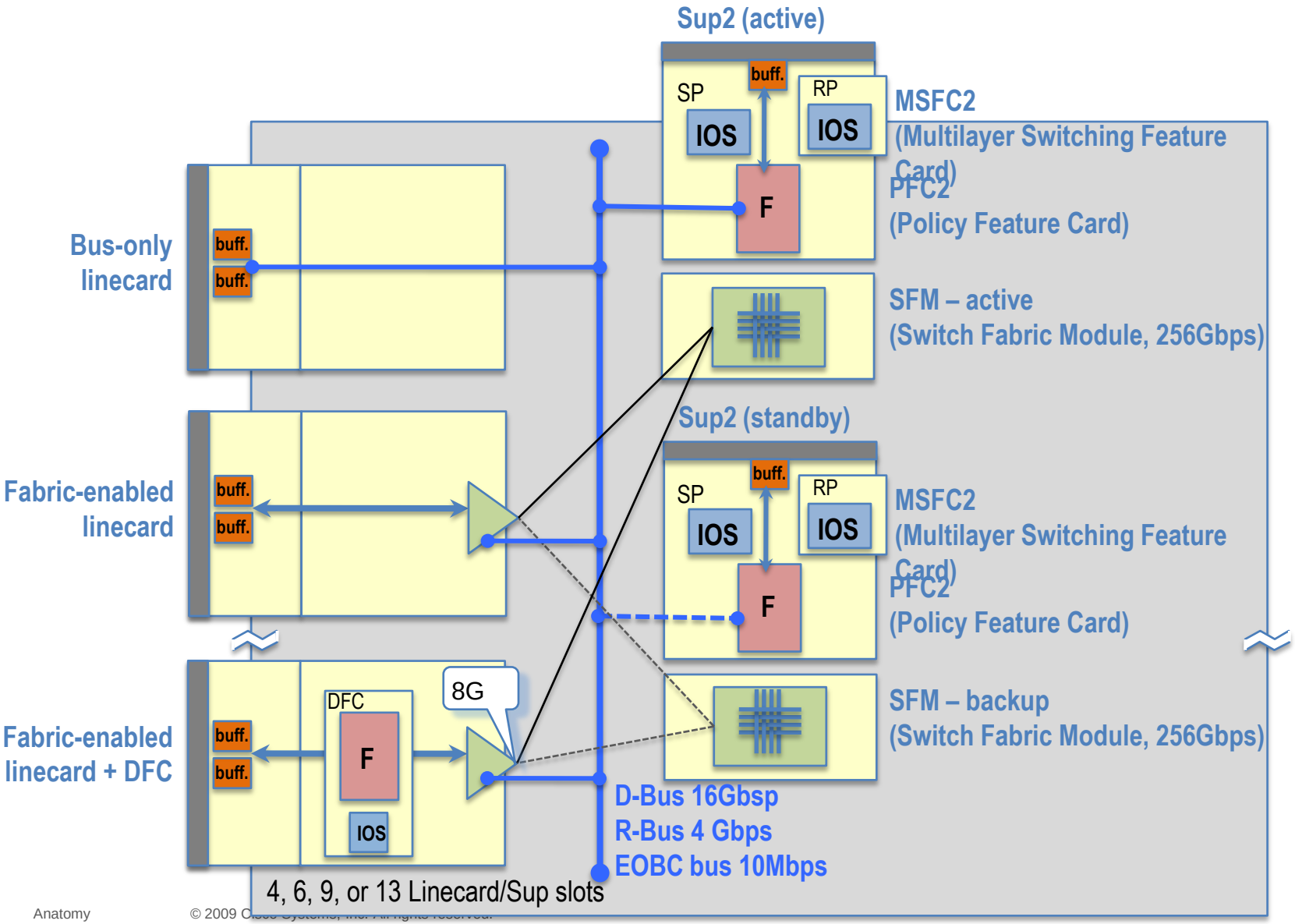
RFC 1925

“The Twelve Networking Truths”

© 2009 Cisco Systems, Inc. All rights reserved.

Catalyst 6500

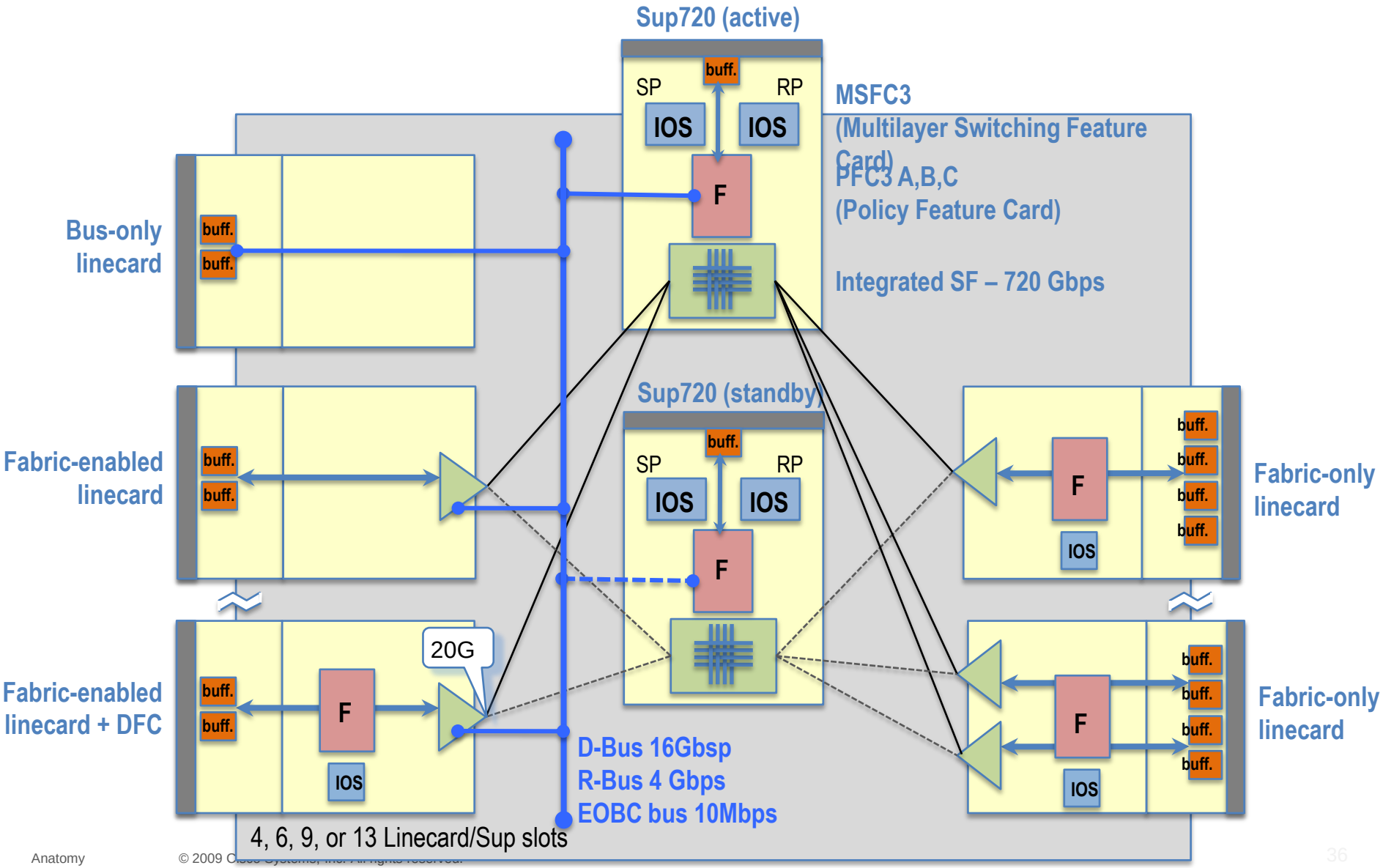
Switch fabric



© 2009 Cisco Systems, Inc. All rights reserved.

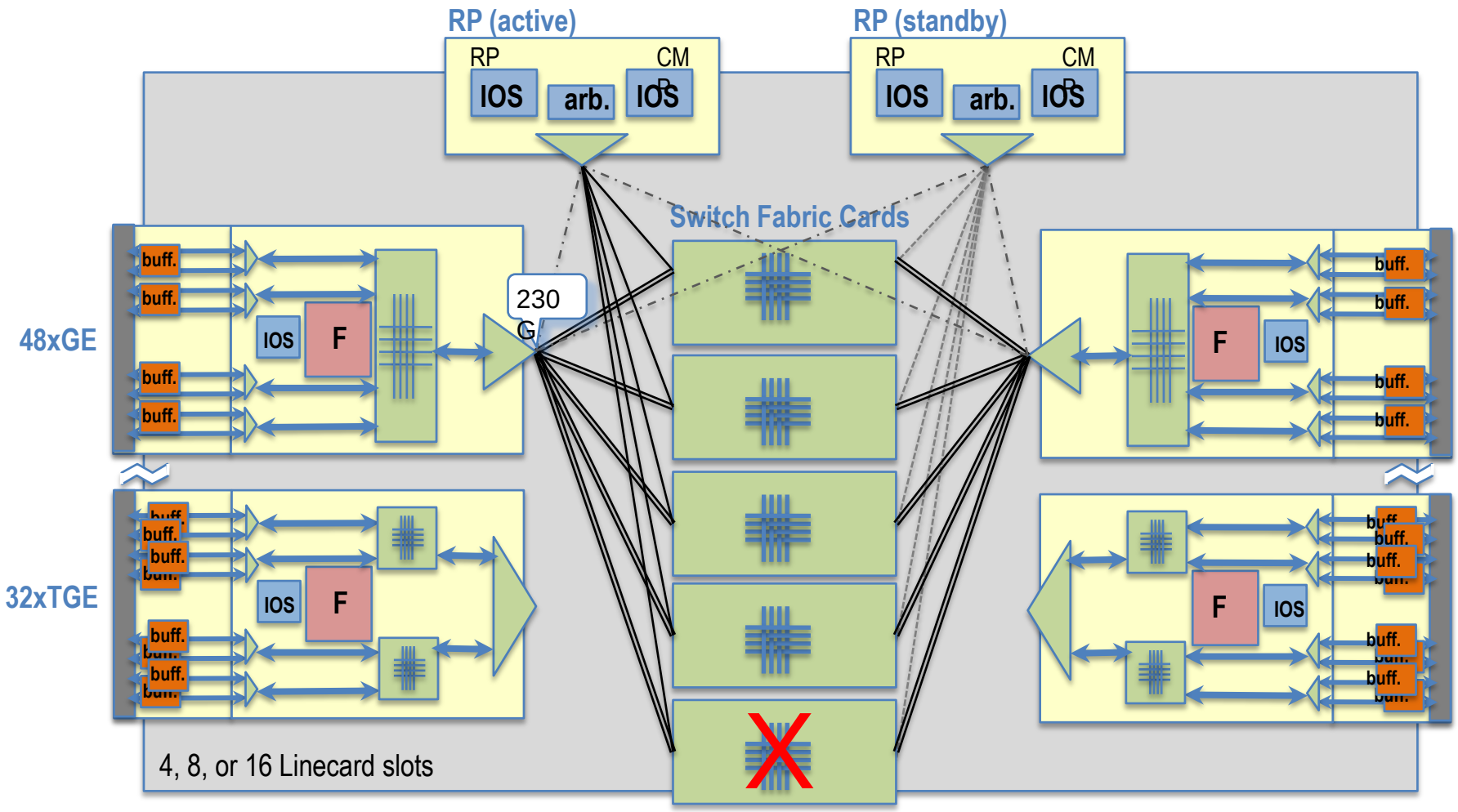
Catalyst 6500

Sup720



© 2009 Cisco Systems, Inc. All rights reserved.

Nexus 7000
architecture



© 2009 Cisco Systems, Inc. All rights reserved.

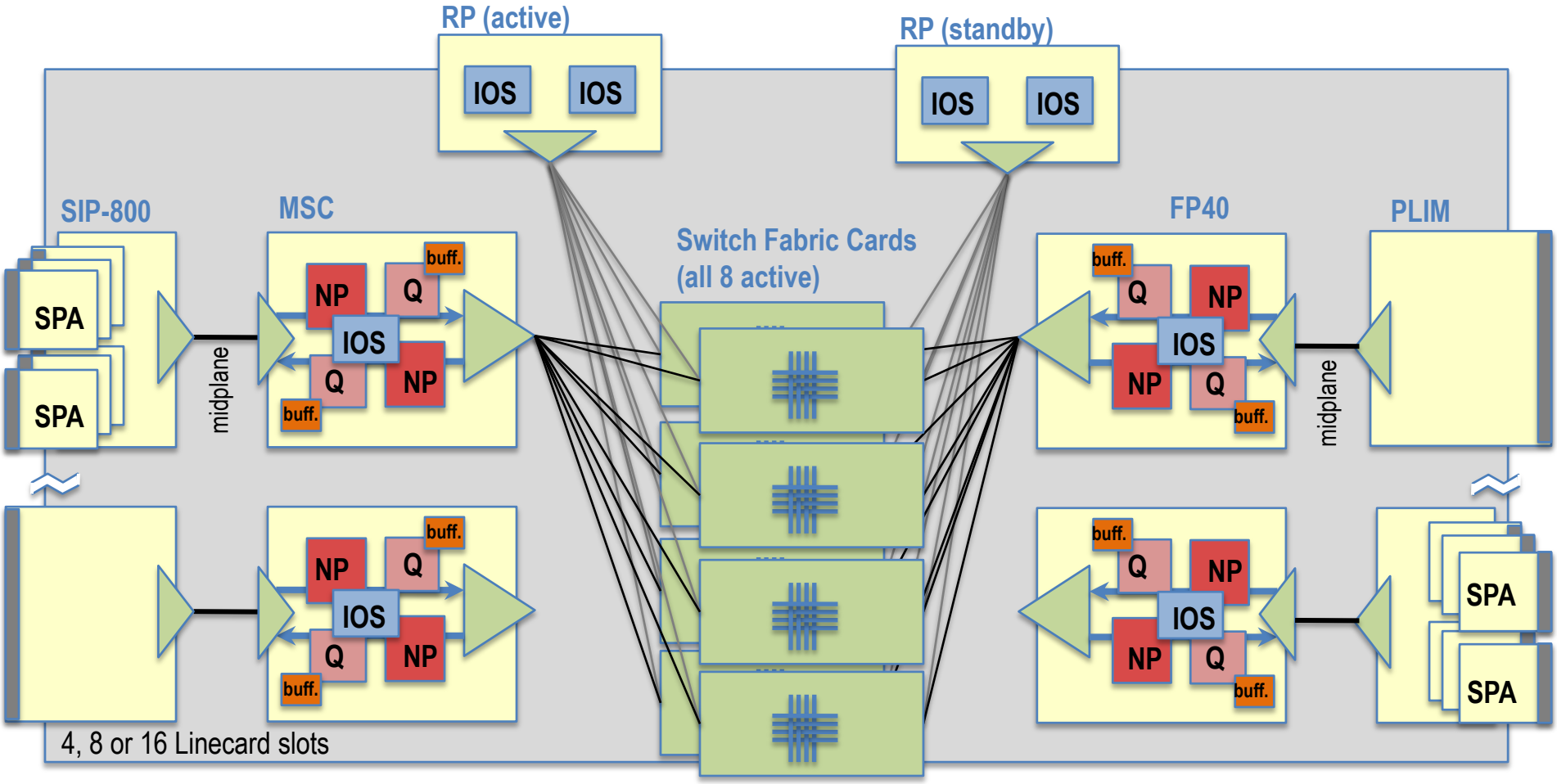
Switch/Router Anatomy

Chapter 4

© 2009 Cisco Systems, Inc. All rights reserved.

Cisco CRS-1

extremely modular architecture

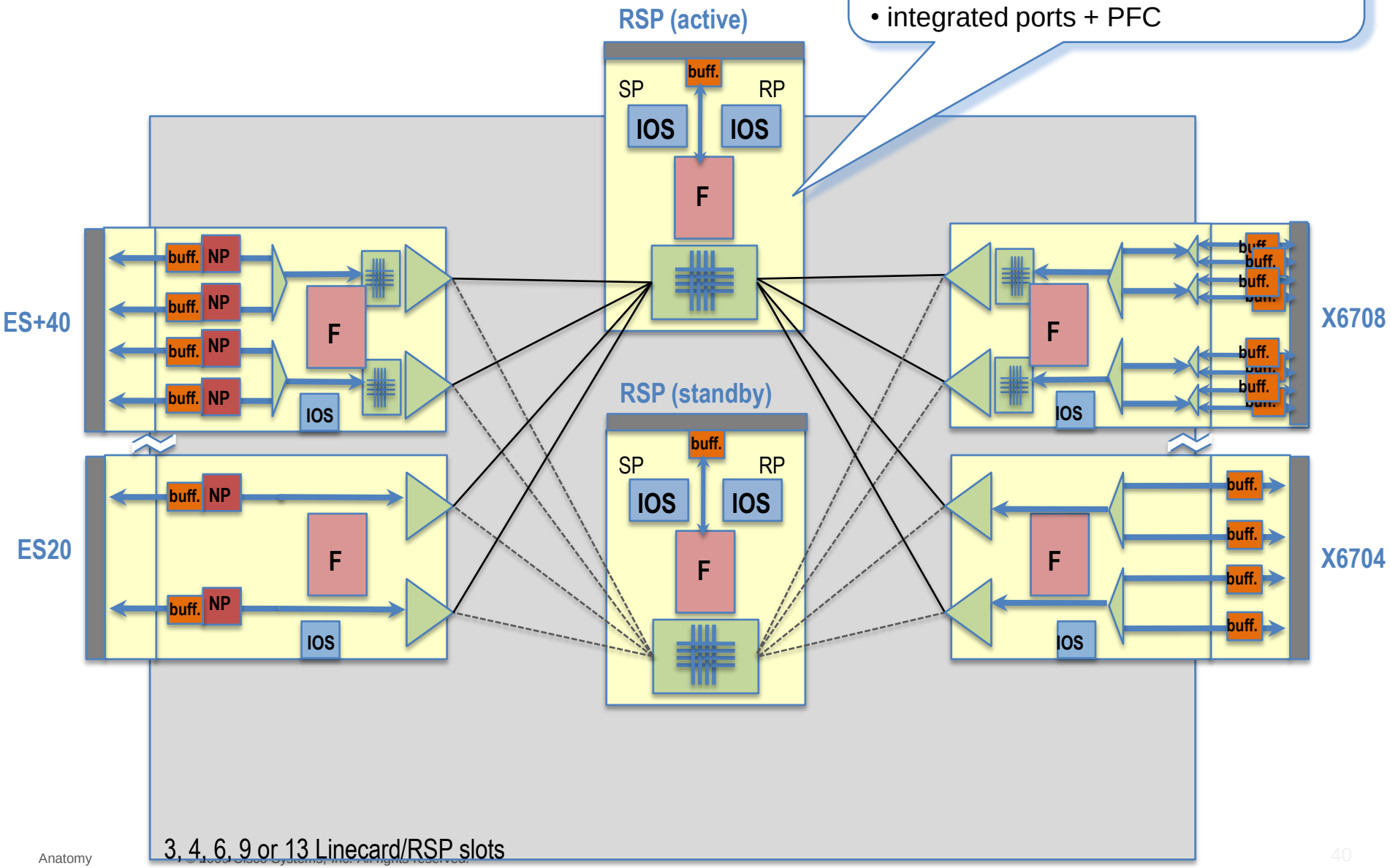


© 2009 Cisco Systems, Inc. All rights reserved.

Cisco 7600

extremely compact architecture

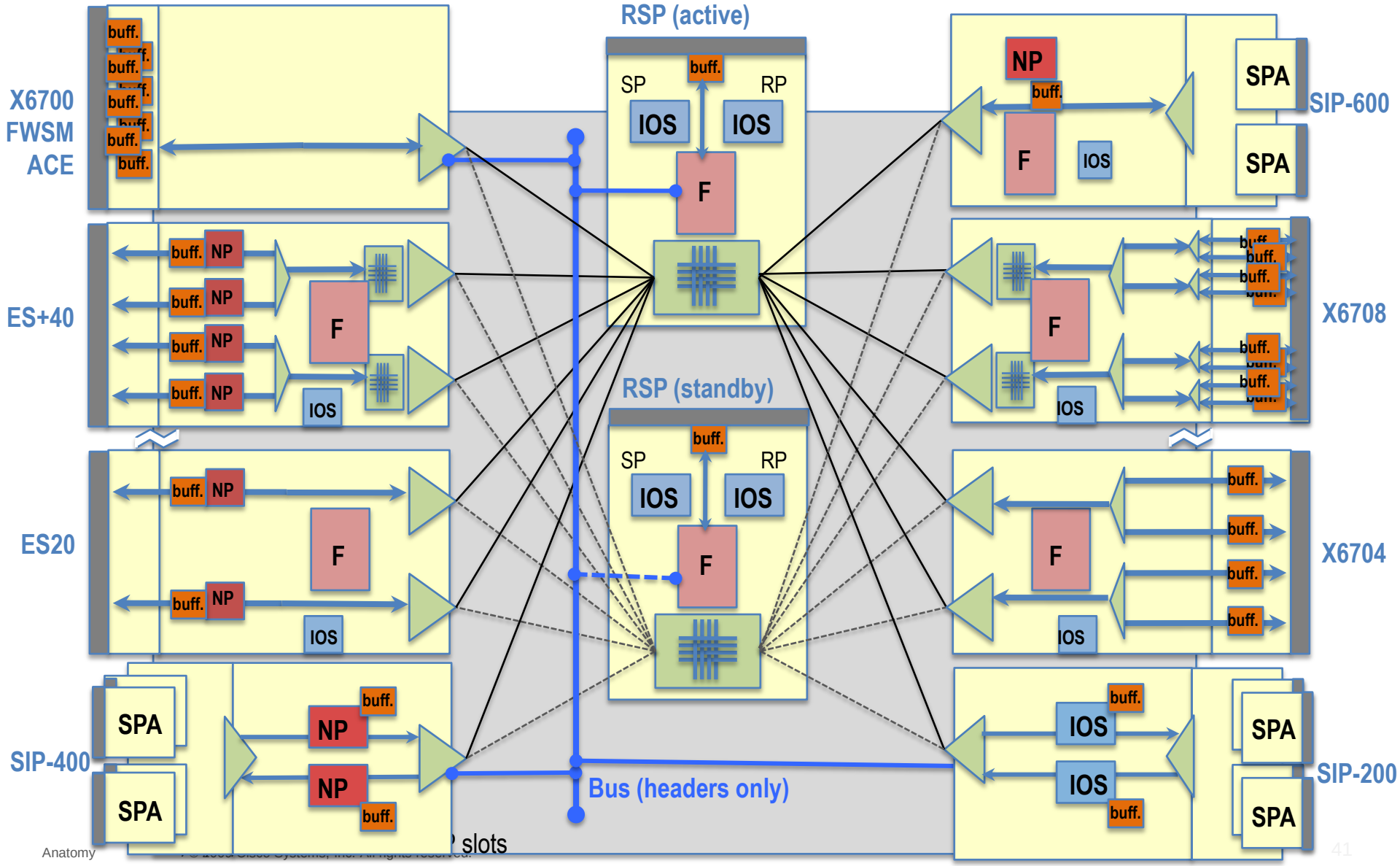
RSP (Route/Switch Processor)
- CPU + Switch-fabric
- active/standby SF
- integrated ports + PFC



© 2008 Cisco Systems, Inc. All rights reserved.

Cisco 7600

switch/router architecture



Cisco Systems, Inc. All rights reserved.

“It is always possible to agglutinate multiple separate problems
into a single complex interdependent solution.

In most cases this is a bad idea.”

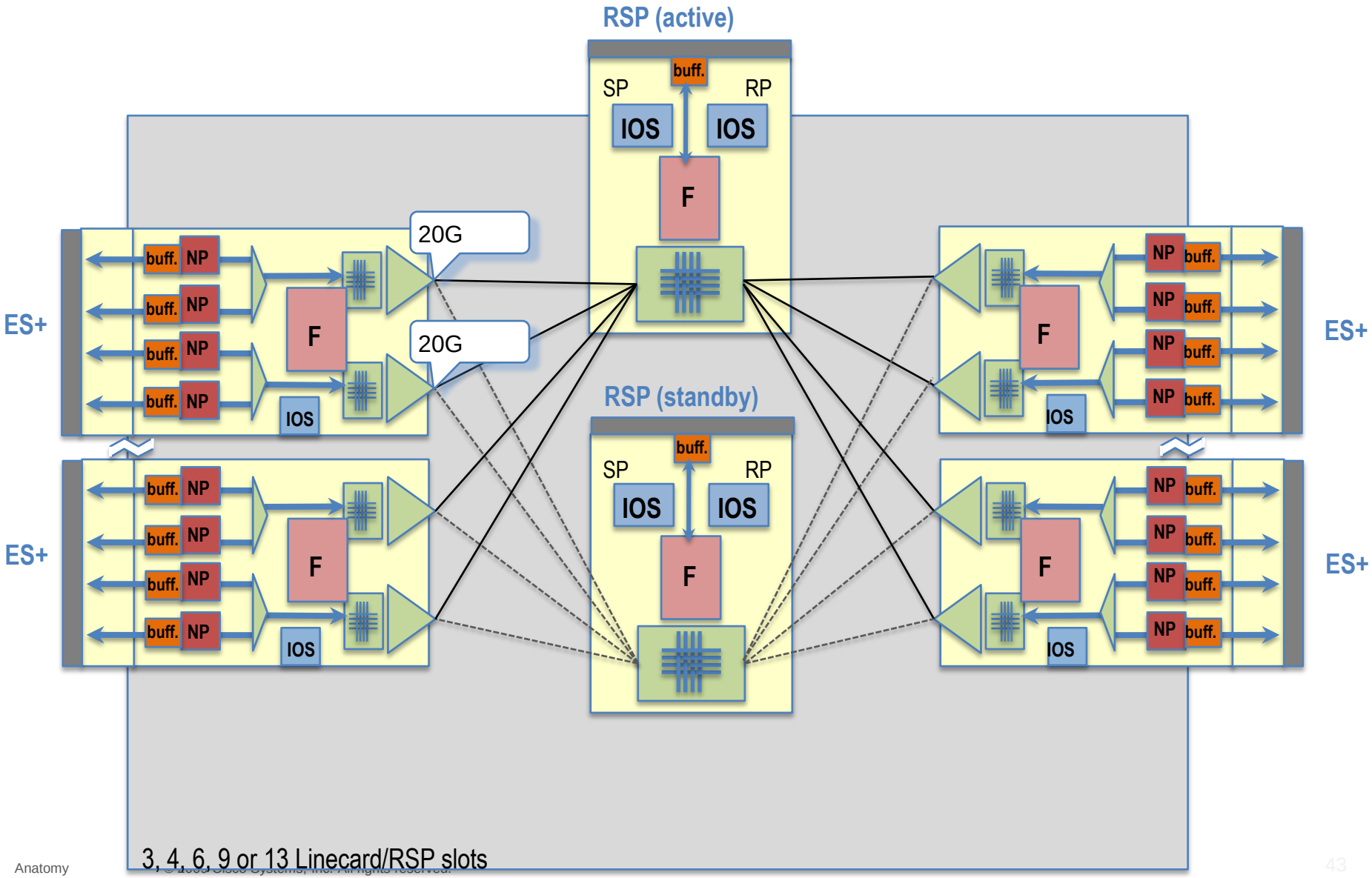
RFC 1925

“The Twelve Networking Truths”

© 2009 Cisco Systems, Inc. All rights reserved.

Cisco 7600 with ES+

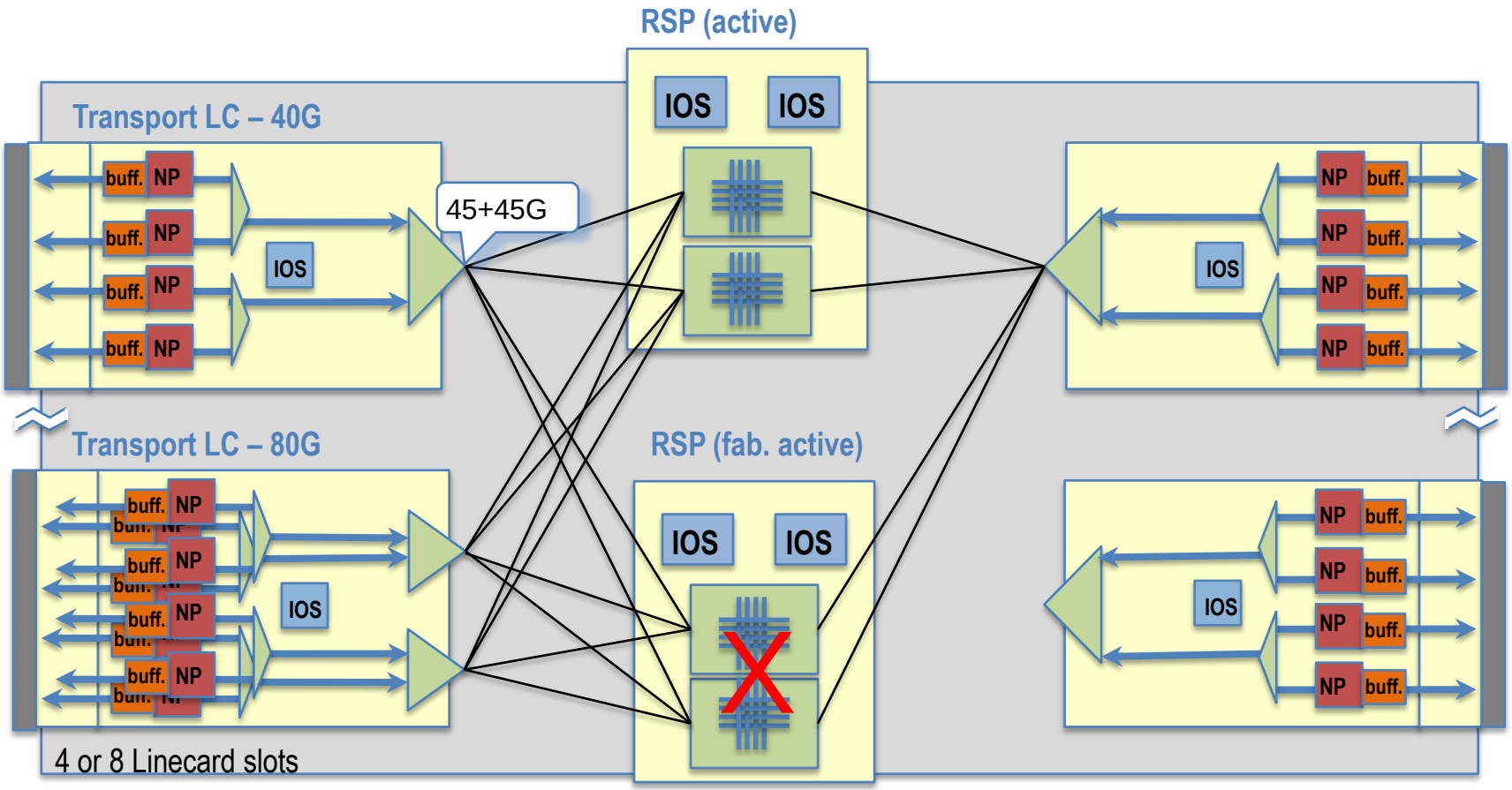
new generation – nice and clean



© 2008 Cisco Systems, Inc. All rights reserved.

Cisco ASR9000

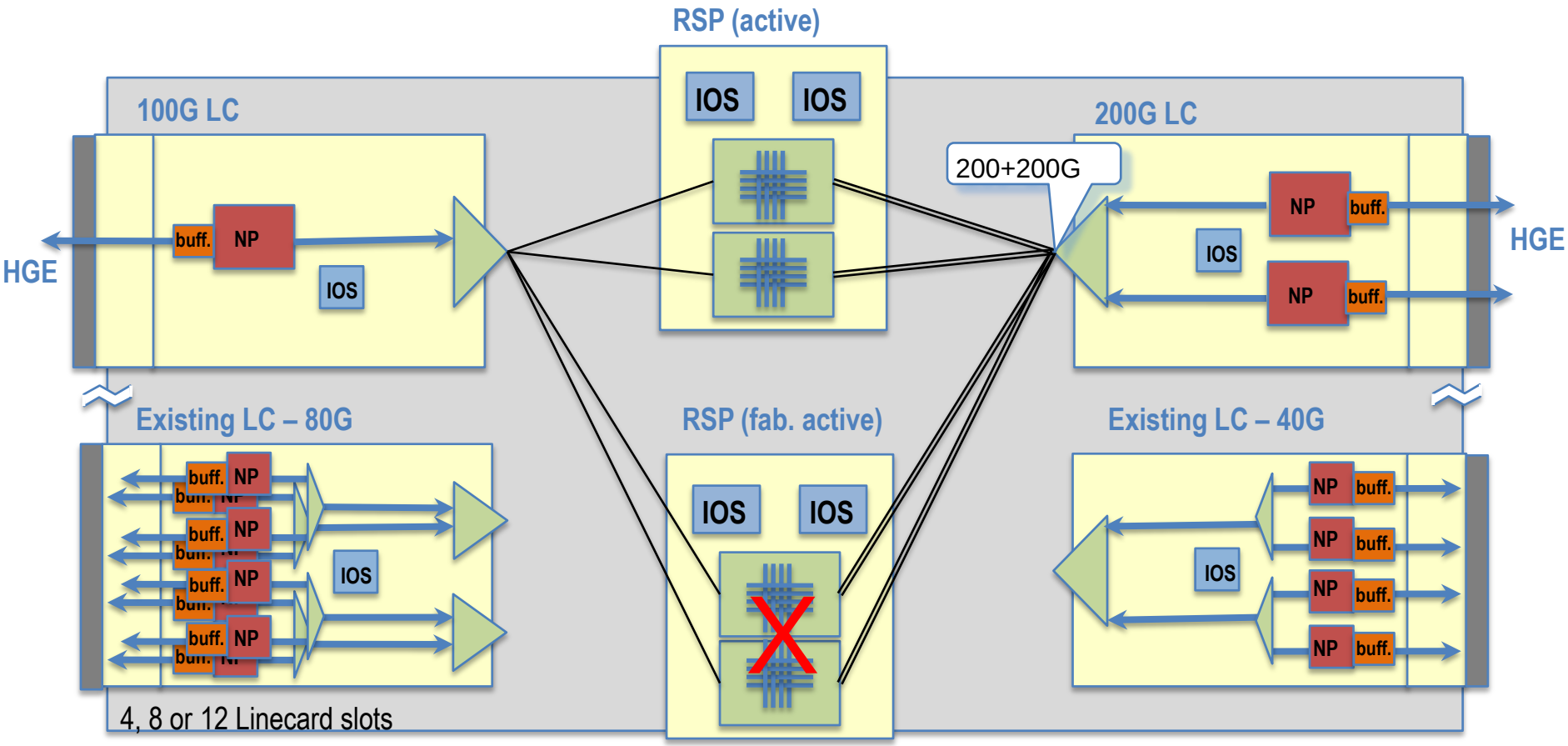
router/switch architecture



© 2009 Cisco Systems, Inc. All rights reserved.

Cisco ASR9000

double 100GE evolution



© 2009 Cisco Systems, Inc. All rights reserved.

Network Processors

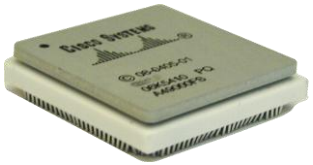
Chapter 5

© 2009 Cisco Systems, Inc. All rights reserved.

Forwarding ASIC's and Network Processors

Price vs. Performance vs. Flexibility

Pipelining L2/L3 Switch ASIC
- example – Catalyst 4500
- 250Mpps, 320Gbps



Parallel L2/L3 Switch ASIC
- example – Catalyst 6500
- 48Mpps, 80+Gbps

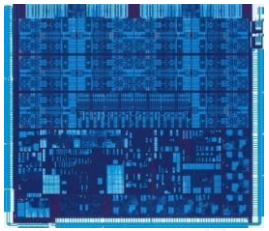
Pipelining ASIC with u-program. stages
- example – Cisco 12000
- 16Mpps, 10Gbps
- complex features – cRTP, ATM, MVPN...

Pipelining u-program. NP
- example – Cisco 10000, 7600, 7300
- 9Mpps, 18Gbps
- complex features – PPPoE, ISG...

Pipelining SMP (symmetric multiprocessing) u-program. NP
- example – Cisco 7600, ASR9000
- 20Mpps, 20Gbps
- very complex features – VPLS, PPPoE, ISG, Vidmon...

Massive SMP u-program. NP
- example – Cisco CRS-1
- 80Mpps, 40Gbps
- 185M transistors
- core/edge/eth. features

Massive SMP C-program. NP
- example – Cisco ASR1000
- 16Mpps, 20Gbps
- 1.3B transistors
- fully programmable in C
- >40 patents
- 100 engineers, 5 years
- sees full packet bodies! (firewall, IPSec, DPI, ISG,...)

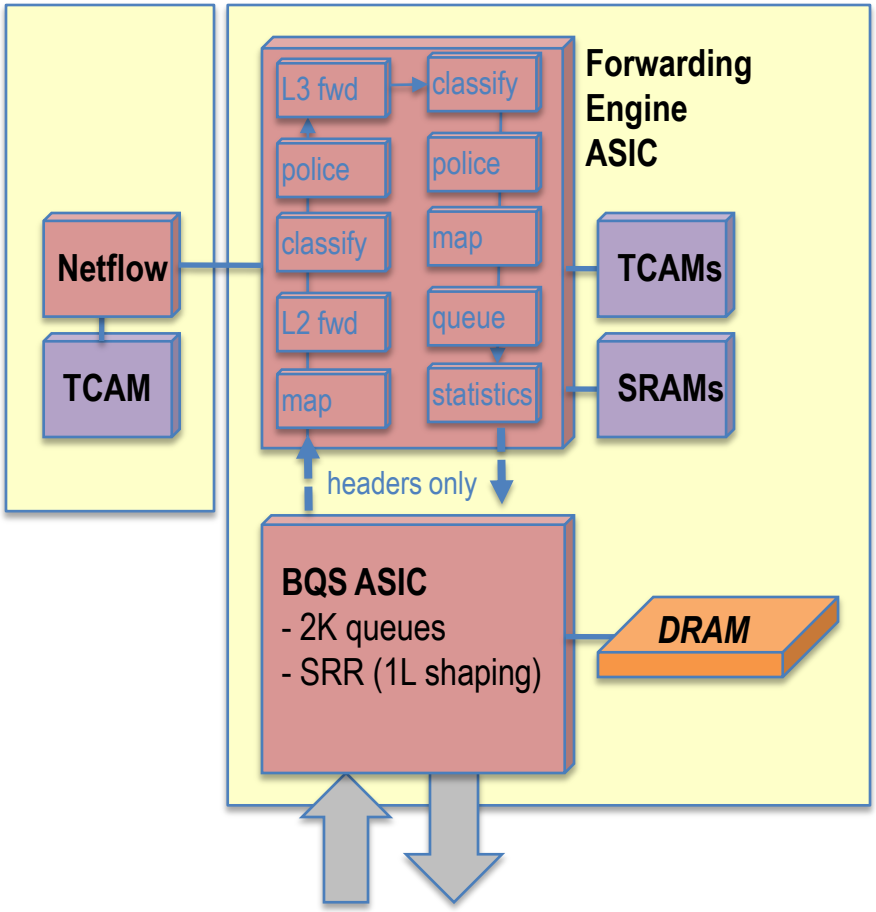


© 2009 Cisco Systems, Inc. All rights reserved.

Non-programmable L2/L3 Switching ASIC

Catalyst 45/4900

Forwarding [Mpps]: 250
Throughput [Gbps]: 320
Programmability: None



Linecards (8x 3Gbps fdx per LC)

© 2009 Cisco Systems, Inc. All rights reserved.

“One size never fits all.”

RFC 1925

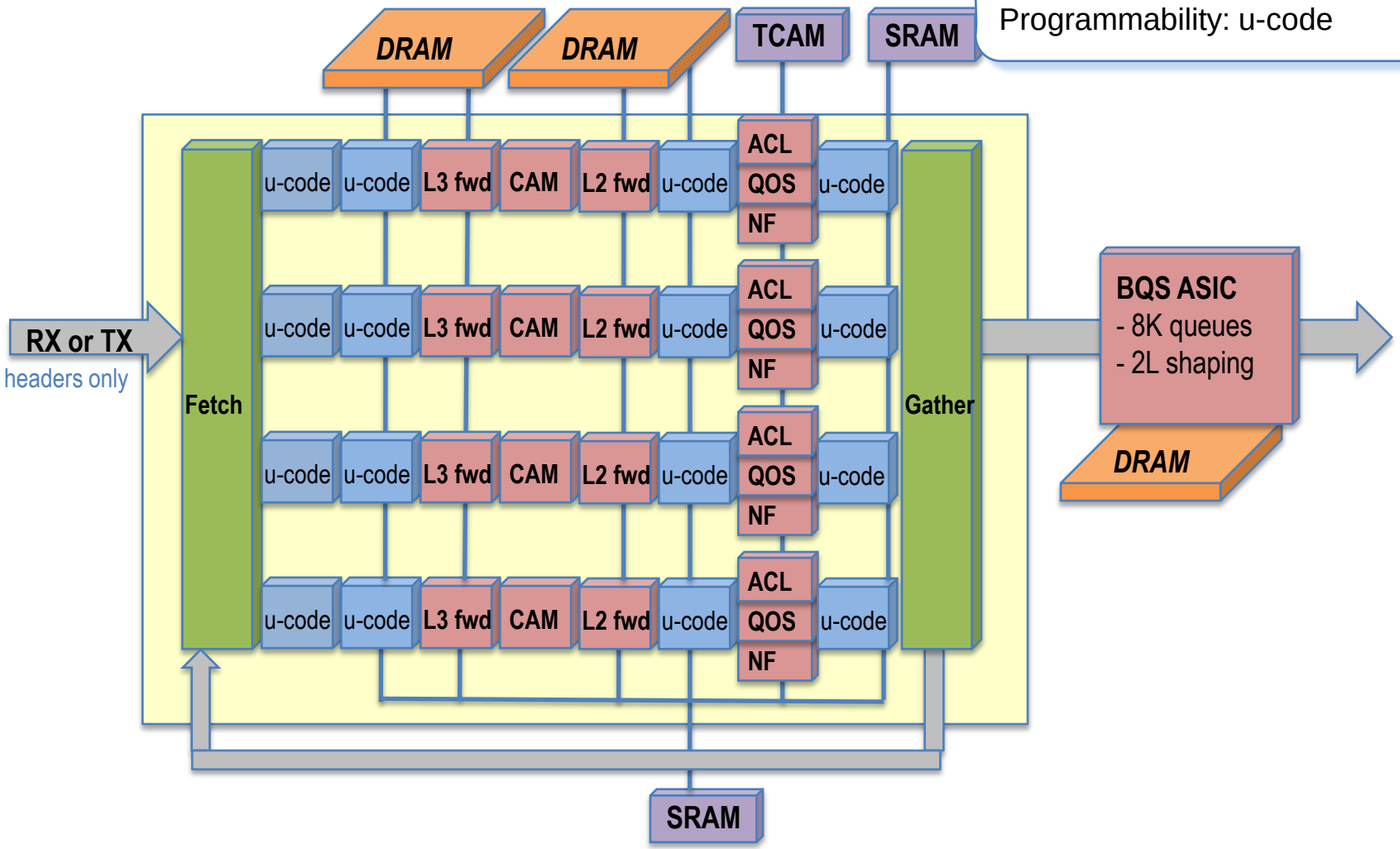
“The Twelve Networking Truths”

© 2009 Cisco Systems, Inc. All rights reserved.

Pipelining Parallel Programmable ASIC

GSR – Engine5 (SIP-x01)

Forwarding [Mpps]: 16
Throughput [Gbps]: 10
Programmability: u-code



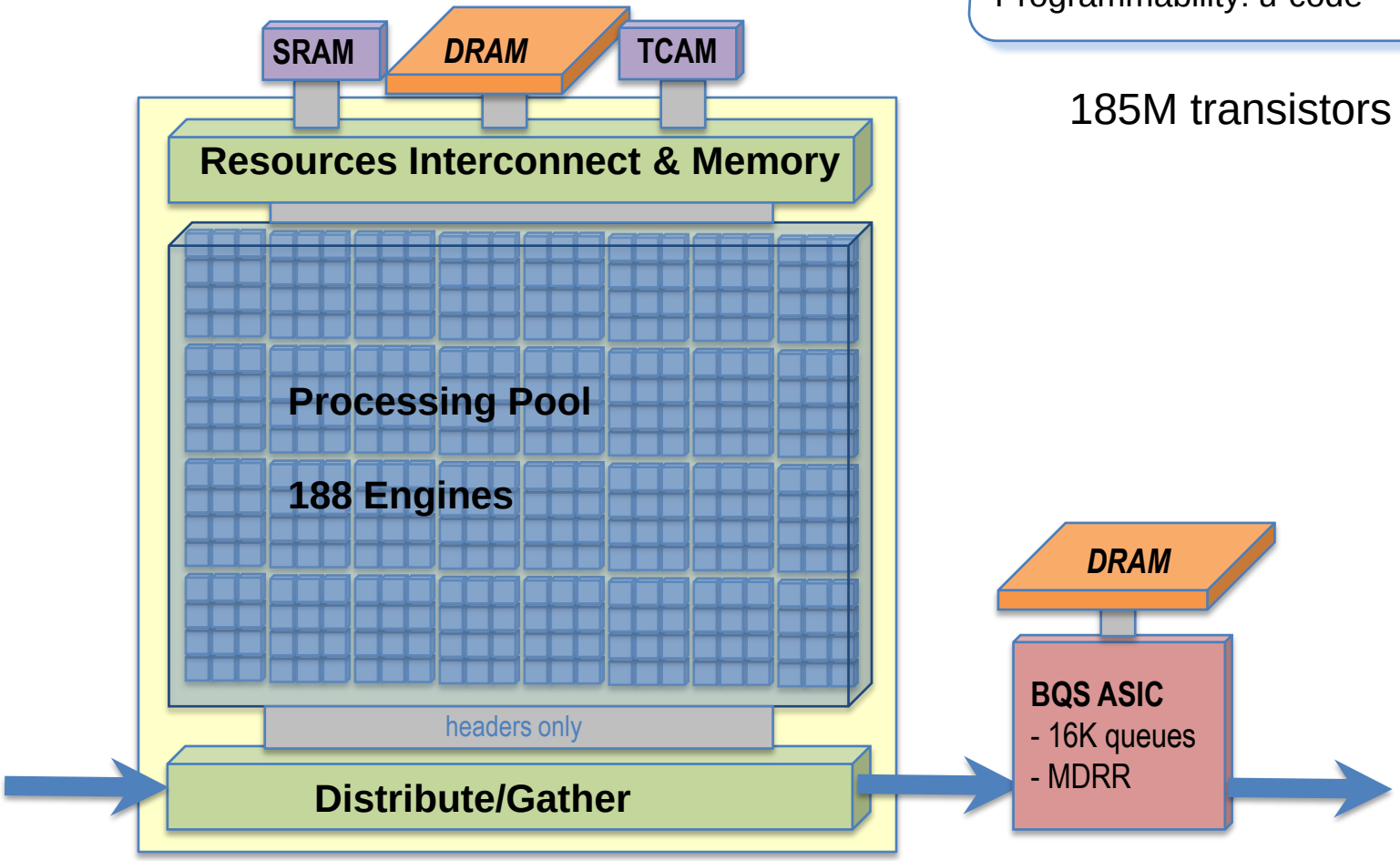
© 2009 Cisco Systems, Inc. All rights reserved.

u-Programmable SMP NPU

CRS – SPP (Silicon Packet Processor)

Forwarding [Mpps]: 80
Throughput [Gbps]: 40
Programmability: u-code

185M transistors



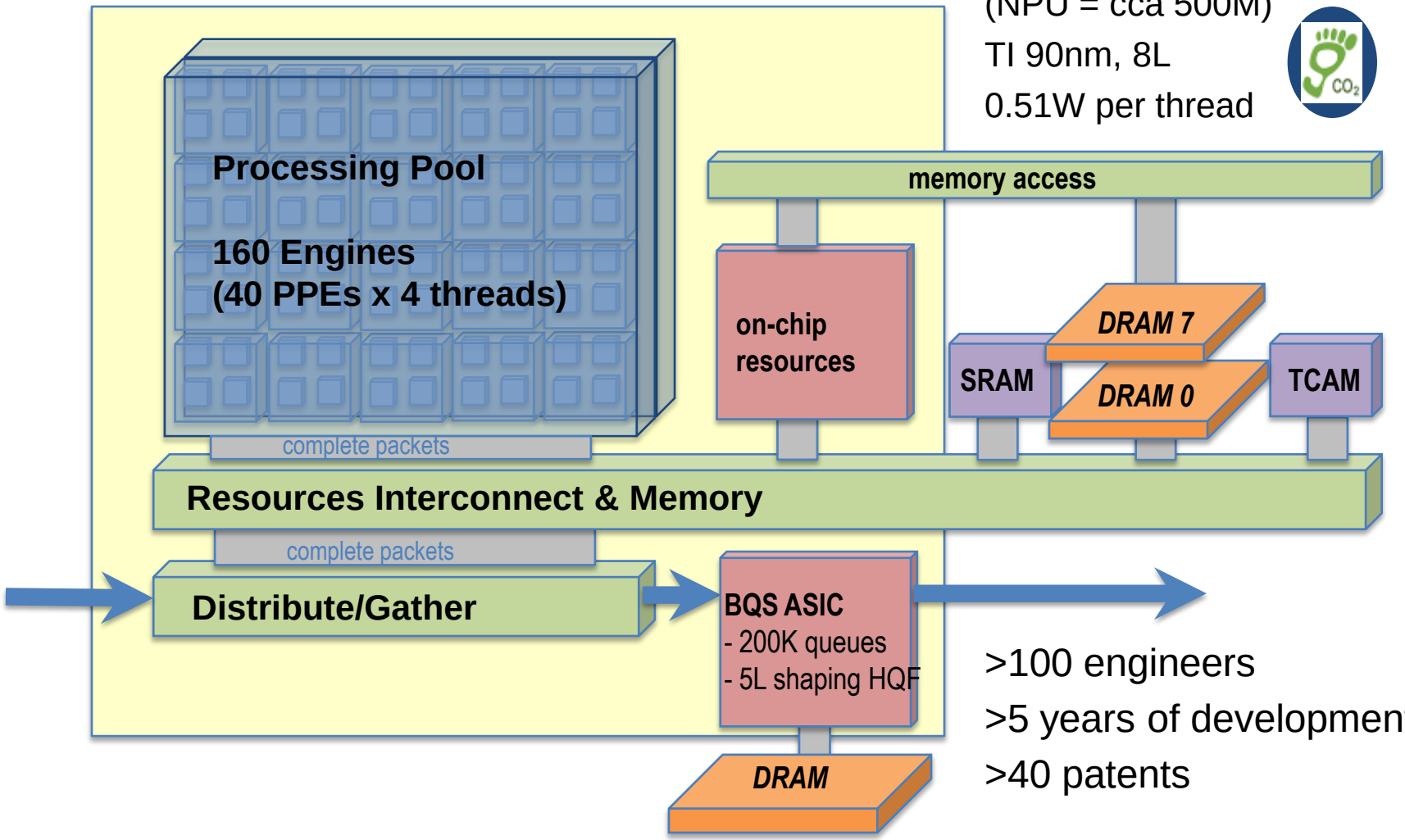
© 2009 Cisco Systems, Inc. All rights reserved.

C-Programmable SMP NPU with complete packet processing

ASR – QFP (Quantum Flow Processor)

Forwarding [Mpps]: 16
Throughput [Gbps]: 20
Programmability: C code

1.3B transistors
(NPU = cca 500M)
TI 90nm, 8L
0.51W per thread



>100 engineers
>5 years of development
>40 patents

© 2009 Cisco Systems, Inc. All rights reserved.

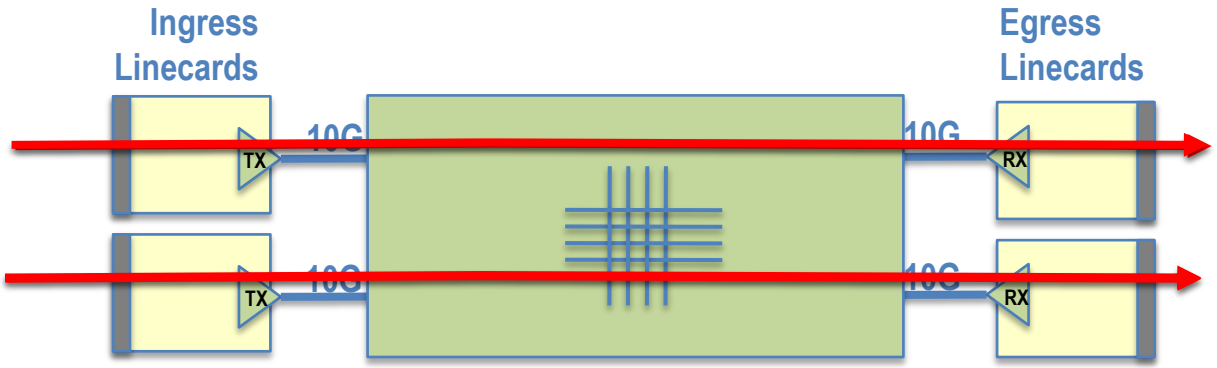
Switch Fabrics
Chapter 6

© 2009 Cisco Systems, Inc. All rights reserved.

“Non-Blocking” woo-doo

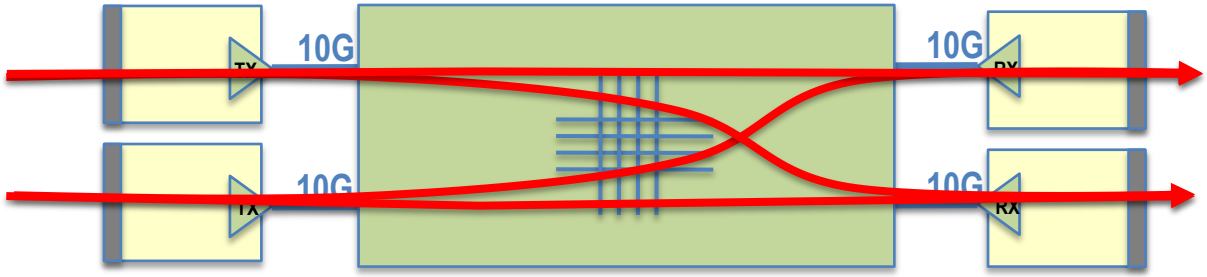
RFC1925: it is more complicated than you think!

RFC1925: the fast, the GOOD, and the cheap



Port-to-Port traffic

•zero packet loss



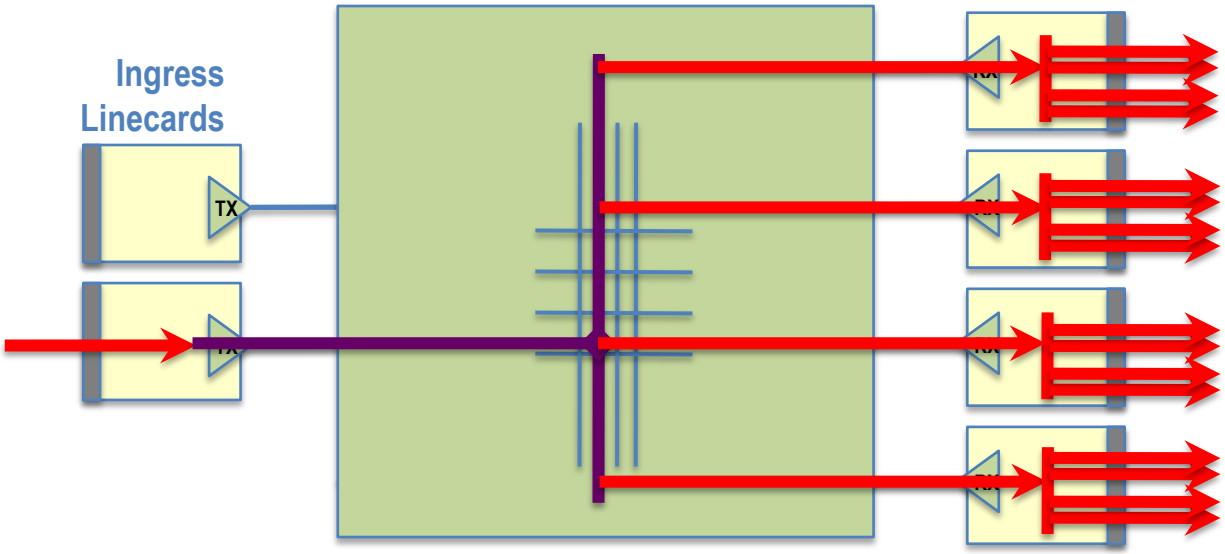
Any-to-Any traffic

- Voice/Video/Data
- Unicast/Multicast
- non-zero packet loss!

© 2009 Cisco Systems, Inc. All rights reserved.

Fabric Multicast Replication

Egress vs. Ingress Replication

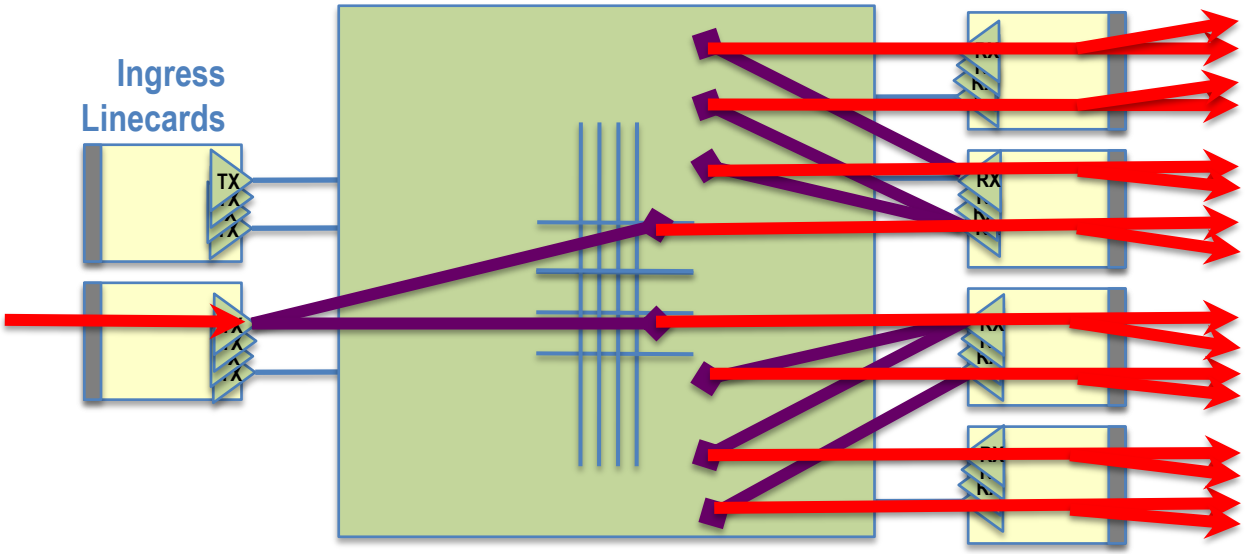


Good:

Egress Replication

- Cisco CRS-1, 12000
- Cisco ASR9K, 7600

10Gbps of multicast eats 10Gbps fabric bw!



Good-enough: Binary Ingress Replication

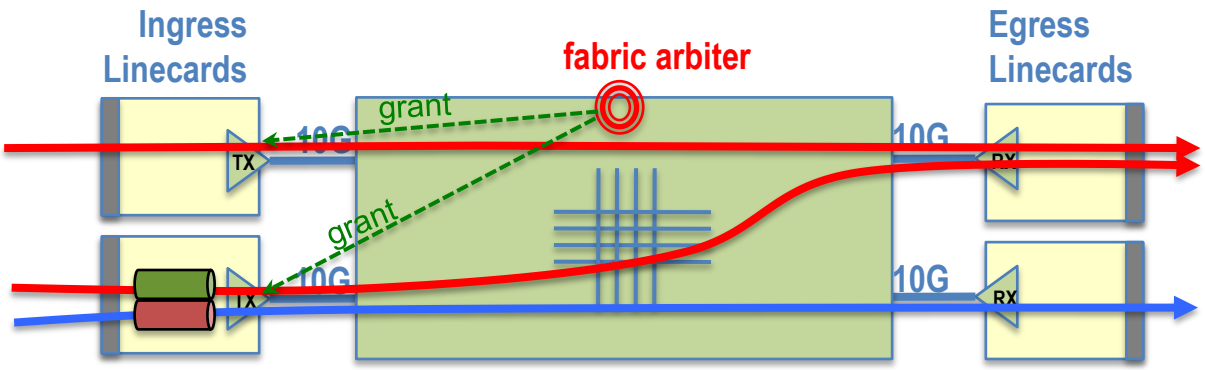
- dumb switch fabric
- non-Cisco

10Gbps of multicast eats 80Gbps fabric bw!!

© 2009 Cisco Systems, Inc. All rights reserved.

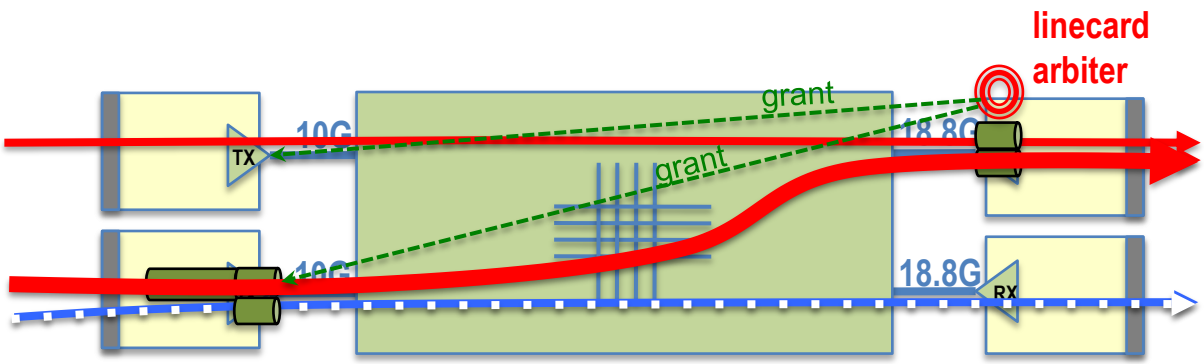
Fabric QoS

Head-of-Line Blocking and Arbitration



Good Fabric QoS example

- GSR, ASR9000, Nexus
- per-destination queues
- strict priority for Voice/TV
- thousands of queues



Good-enough Fabric QoS example

- non-Cisco
- traffic unaware arbiter (overspeed is needed)
- only 2 queues, no priority
- drops voice/video if loaded
- loss at >66% load

© 2009 Cisco Systems, Inc. All rights reserved.

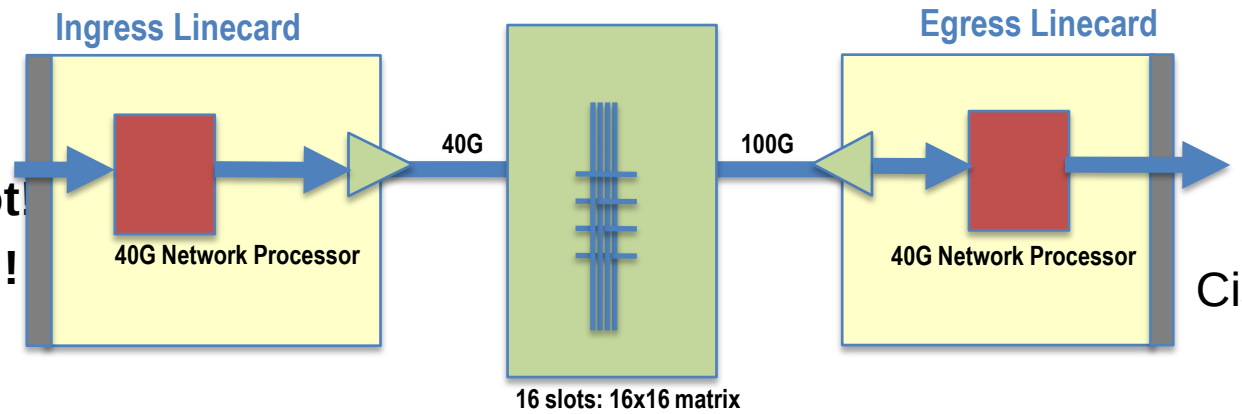
Marketing Math

40 ≠ 40

Good:

40 Gbps per slot!

non-blocking!!!

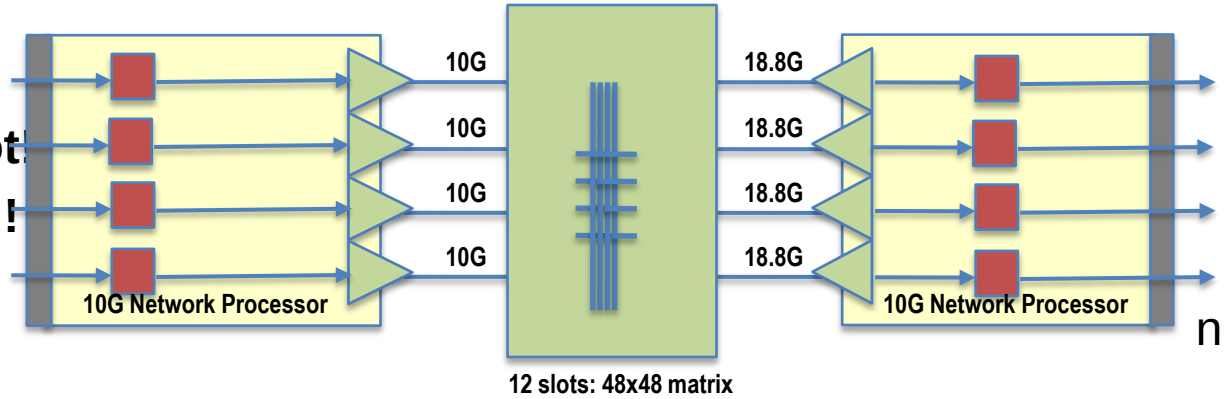


Cisco CRS-1

Good-enough:

40 Gbps per slot!

non-blocking!!!



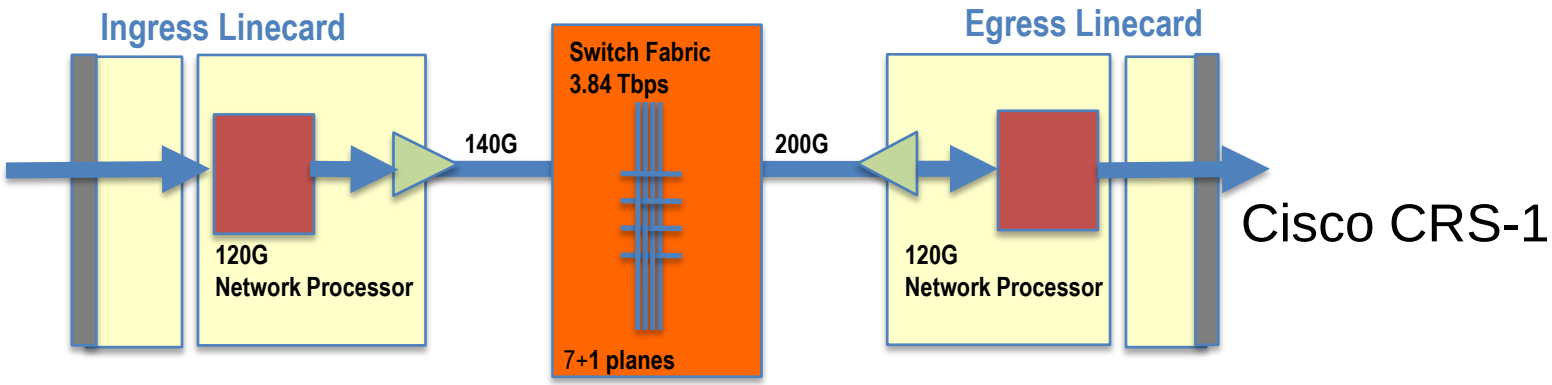
non-Cisco

© 2009 Cisco Systems, Inc. All rights reserved.

Marketing Math

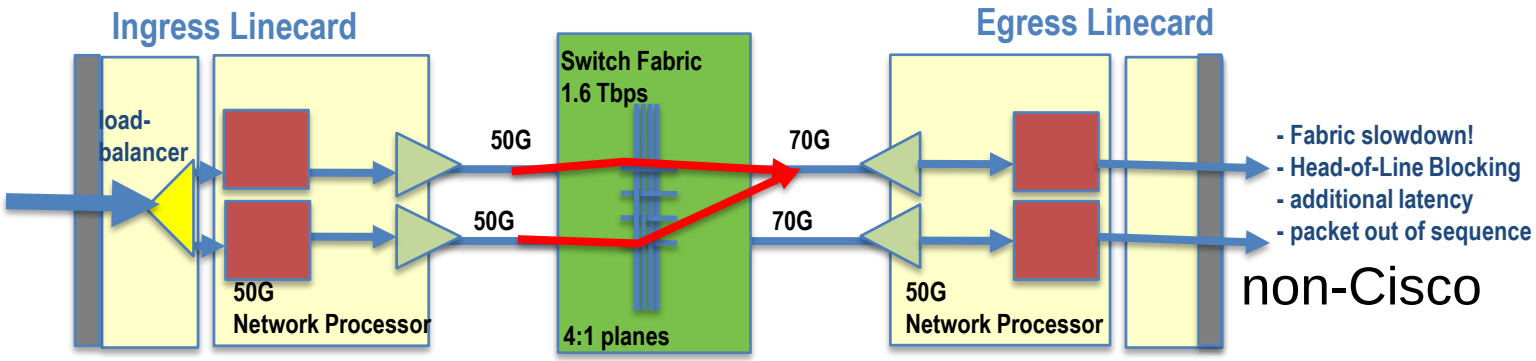
100 ≠ 100

100GE



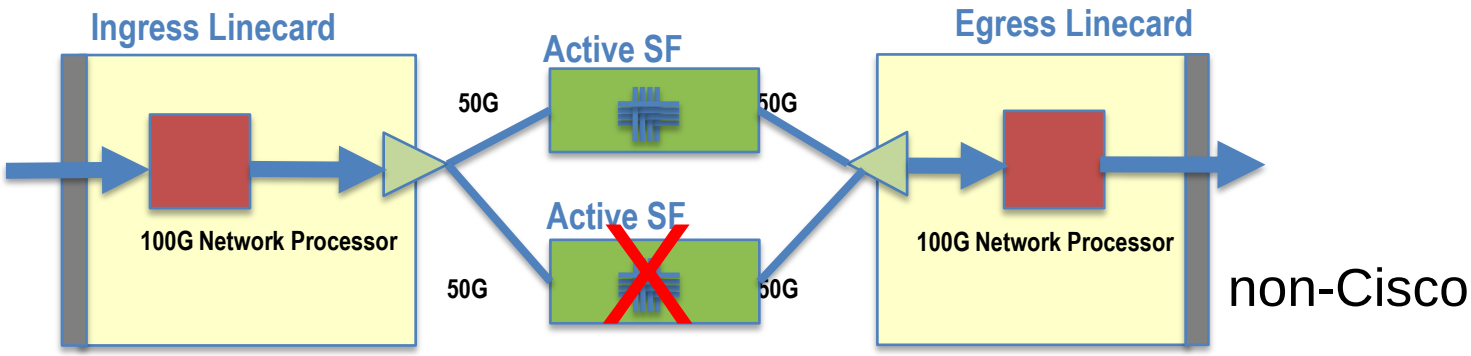
100GE?!

No 100G processor available, internal load-balancing across two old 50G processors



100GE??!

No independent Switching Fabric available, if one CMP is removed, only 50Gbps throughput (no redundancy)



© 2009 Cisco Systems, Inc. All rights reserved.

Ferrari = red car with horse

© 2009 Cisco Systems, Inc. All rights reserved.

Conclusion

The art of engineering optimization

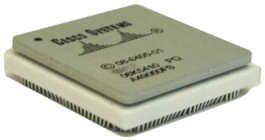
• The Fast, the Good, and the Cheap

Router and Switch Architectures

•Centralized, Distributed, Hybrid

Network Processors

• Programmable? Body visibility? Tunneling?



Switch Fabrics

• Is 40G really 40G? Multicast Replication? Redundancy? QoS? 100GE?

© 2009 Cisco Systems, Inc. All rights reserved.

© 2009 Cisco Systems, Inc. All rights reserved.