TOMORROW starts here.

CISCO

Cisco live!

Cisco Nexus 3548 Switch Architecture

BRKARC-2013

V1.6

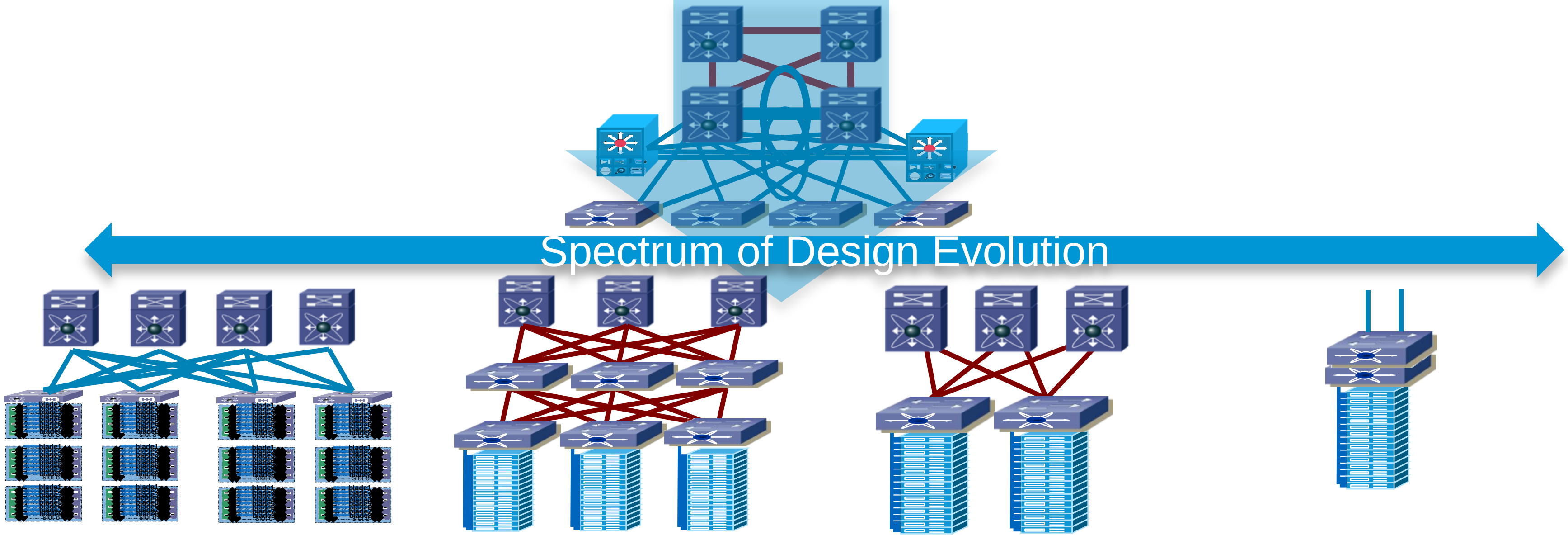
Lucien Avramov

Technical Marketing Engineer

Data Center Group – Business Unit [Nexus]

lucien@cisco.com

Four Data Center Architecture Trends



Virtualized Data Center

- SP and Enterprise
- Hypervisor Virtualization
- Shared infrastructure Heterogenous
- 1G Edge moving to 10G
- Nexus 1000v, 2000, 5500, 7000 & UCS

Warehouse Scale

- Layer 3 Edge (iBGP, ISIS)
- 1000’s of racks
- Homogeneous Environment
- No Hypervisor virtualization
- 1G edge moving to 10G
- Nexus 2000, 3000, 5500, 7000 & UCS

HPC/GRID

- Layer 3 & Layer 2
- No Virtualization
- iWARP & RCoE
- Nexus 2000, 3000, 5500, 7000 & UCS
- 10G moving to 40G

High Frequency Trading

- Layer 3 & Multicast
- No Virtualization
- Limited Physical Scale
- Nexus 3000 & UCS
- 10G edge moving to 40G

Cisco Nexus 3548 Switch Architecture

BRKARC-2013

V1.6

Lucien Avramov

Technical Marketing Engineer

Data Center Group – Business Unit [Nexus]

lucien@cisco.com

Session Goal – BRKARC-2013

Provide a framework and context around understanding 3548

- Understand how to measure performance at nanosecond

- Understand the Nexus 3548 architecture and benefits

- Look into design examples

Agenda – Nexus 3548 – BRKARC-2013

- Benchmarking
- Architecture
- Designs

Agenda – Nexus 3548 – BRKARC-2013

- Benchmarking
  - Line Rate
  - Throughput
  - Latency
  - Jitter
- Architecture
- Designs

Benchmarking

DILBERT, DO YOU HAVE THE BENCHMARK RESULTS?
scottadams@aol.com
www.dilbert.com
DO YOU WANT THE TEN-MINUTE EXPLANATION OF WHY THE DATA ARE USELESS, OR A SIMPLE "HERE YOU GO"?
© 2004 Scott Adams, Inc./Dist. by UFS, Inc.
I'M IN SALES.
HERE YOU GO.

Benchmarking: What is line rate

- "Line Rate" CAN be measured in terms of "Frame Rate":

Frame Rate FR = Transmit-Clock-Frequency / (Frame-Length*8 + Minimum_Gap + Preamble + Start-Frame Delimiter)

Example for 1 GB Ethernet speed with 64-byte frames:

- FR= 1,000,000,000 /(64*8 + 96 + 56 + 8)
- FR= 1,000,000,000 / 672
- FR= 1,488,095.2 Frames per Second.

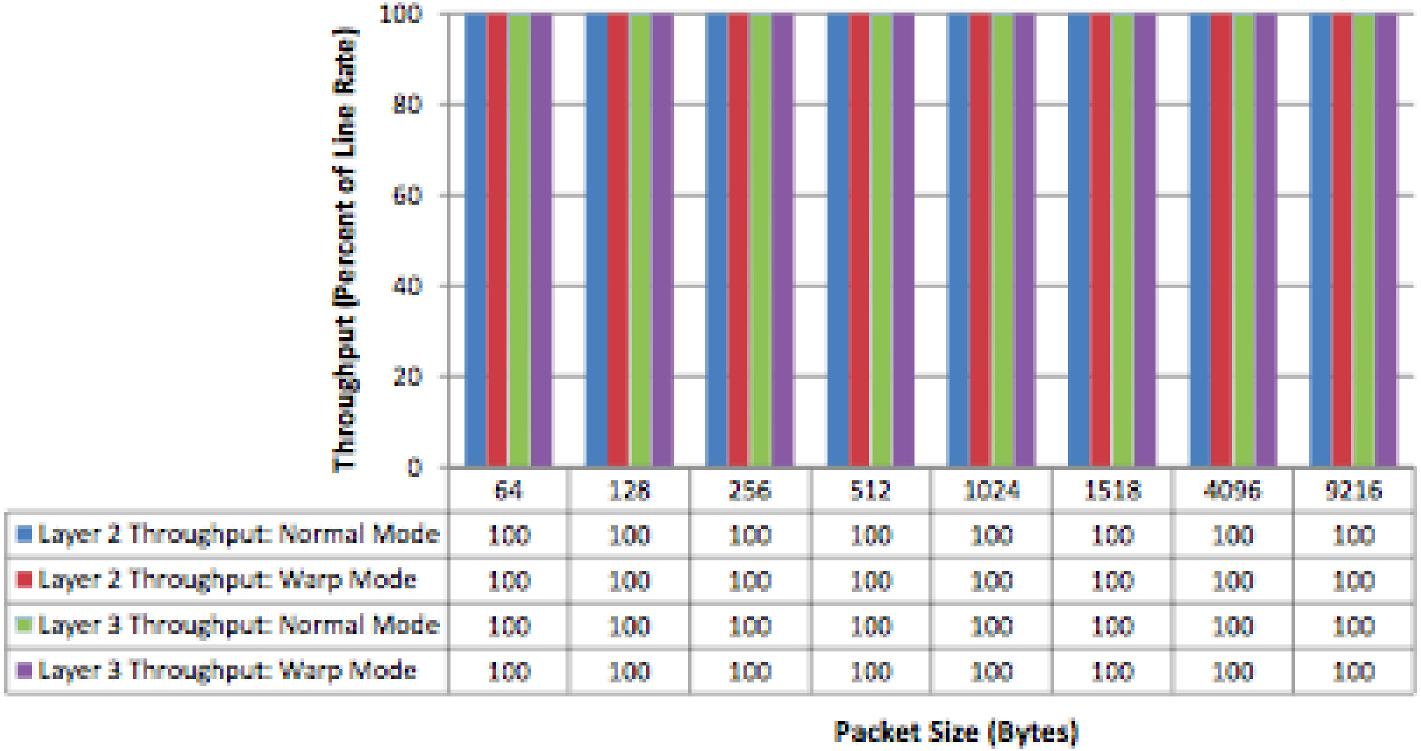
Benchmarking: How to measure linerate

- Traffic generator sending traffic is required

- In a production network, it is very unlikely to see precise line rate over a very brief period.
- There is no observable difference between dropping packets at 99% of line rate and 100% of line rate

Line rate CAN measured at 100% of line rate with a -100PPM adjustment.
Line rate SHOULD be measured at 99,98% with 0 PPM adjustment.

Benchmarking Nexus 3548
10GE Throughput – L2 and L3 - RFC 2544



	64	128	256	512	1024	1518	4096	9216
🟦 Layer 2 Throughput: Normal Mode	100	100	100	100	100	100	100	100
🟥 Layer 2 Throughput: Warp Mode	100	100	100	100	100	100	100	100
🟩 Layer 3 Throughput: Normal Mode	100	100	100	100	100	100	100	100
🟪 Layer 3 Throughput: Warp Mode	100	100	100	100	100	100	100	100

Spirent third party performance report for Nexus 3548

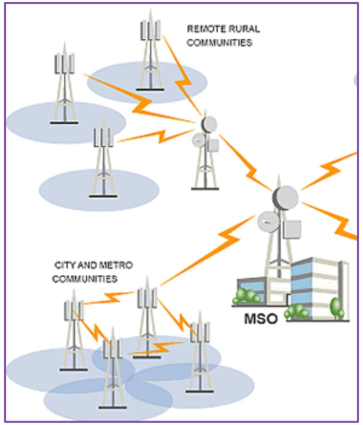
What is Latency?

- Definition of latency: delay introduced in the communication between the time sender initiates it and the receiver receives and processes the information.

- Example: Voice Over IP, Radar, Satellite Communication, Real time application

- Different requirements / different user experience
  - Example of market data (user experience vs. machine trading..)
  - Telecommunication

Financial trading



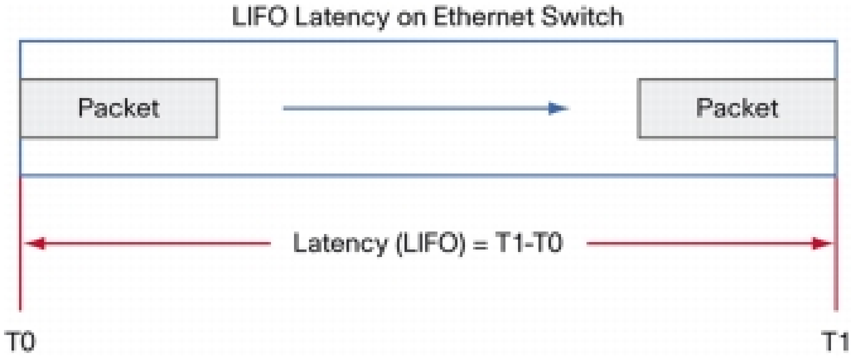
Telecommunications

Ultra Low Latency : sub 1 usec

Benchmarking: Latency

- From RFC 1242:
- For store and forward devices:

The time interval starting when the last bit of the input frame reaches the input port and ending when the first bit of the output frame is seen on the output port.

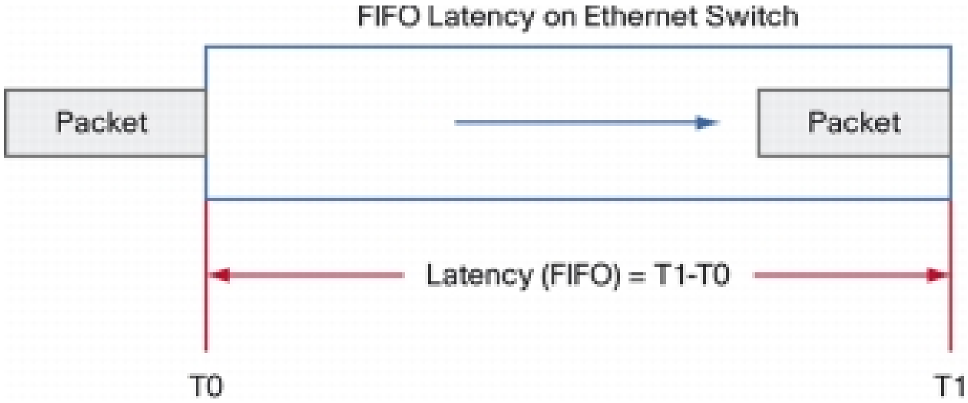


LIFO: Last In First Out

Benchmarking: Latency

- From RFC 1242:
- For bit forwarding devices *(cut-through devices)*:

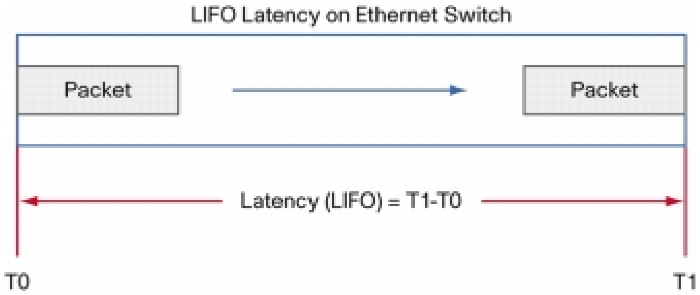
The time interval starting when the end of the first bit of the input frame reaches the input port and ending when the start of the first bit of the output frame is seen on the output port

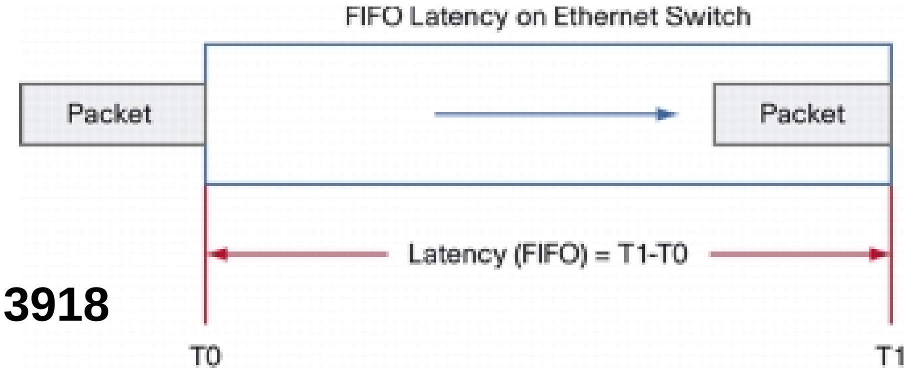


FIFO: First In First Out

Benchmarking: Latency

- Measurement method: LIFO or FIFO?
- LIFO = FIFO - (Packet size in bits/Link speed)
- Cable length: identical cable type and length
- Identical amount of ports to test
- Identical testing equipment:
  - Chassis
  - Testing cards
  - Software Revision
- Typical Latency tests: **RFC 2544, 2889, 3918**



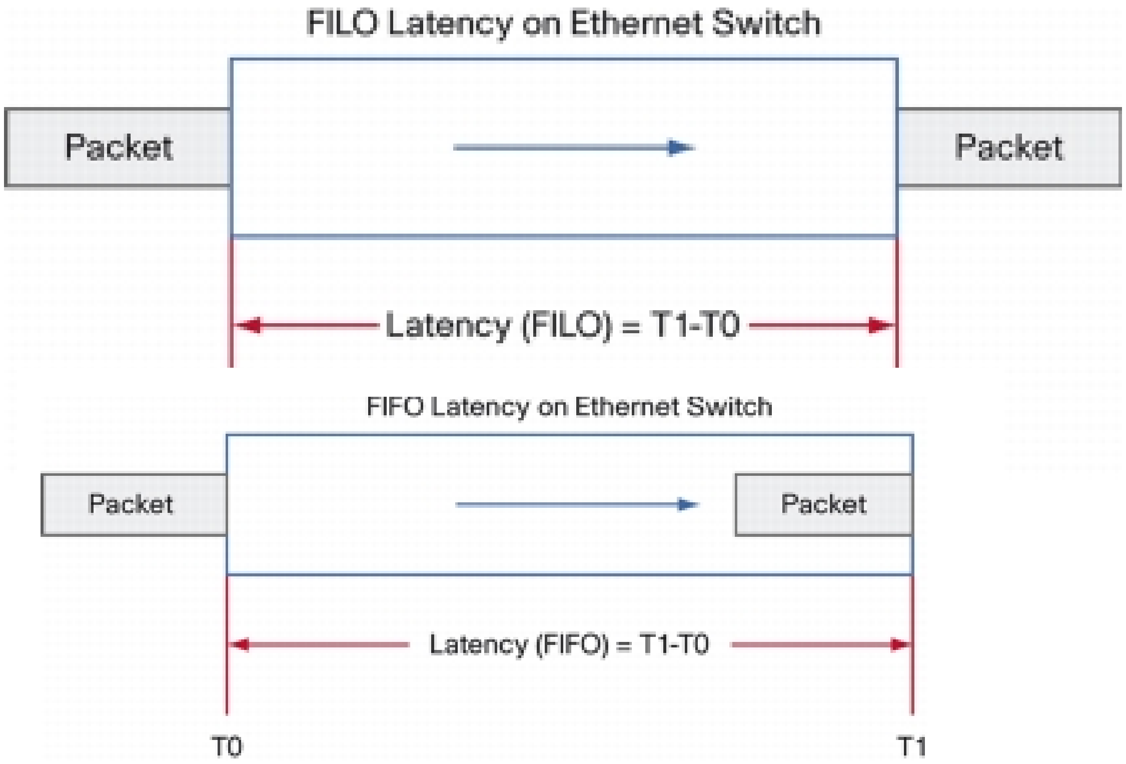


Benchmarking: Latency Comparison

HOW???

Benchmarking: Latency Comparison

HOW???



Benchmarking: Latency Comparison

HOW???

The measuring methods to use for benchmarking purposes are as follow:

1) FILO **MUST** be used as a measuring method, as this will include the latency of the packet; and today the application commonly need to read the whole packet to process the information and take an action.

2) FIFO **MAY** be used for certain applications able to proceed data as the first bits arrive (FPGA for example)

3) LIFO MUST not be used, because it subtracts the latency of the packet; unlike all the other methods.

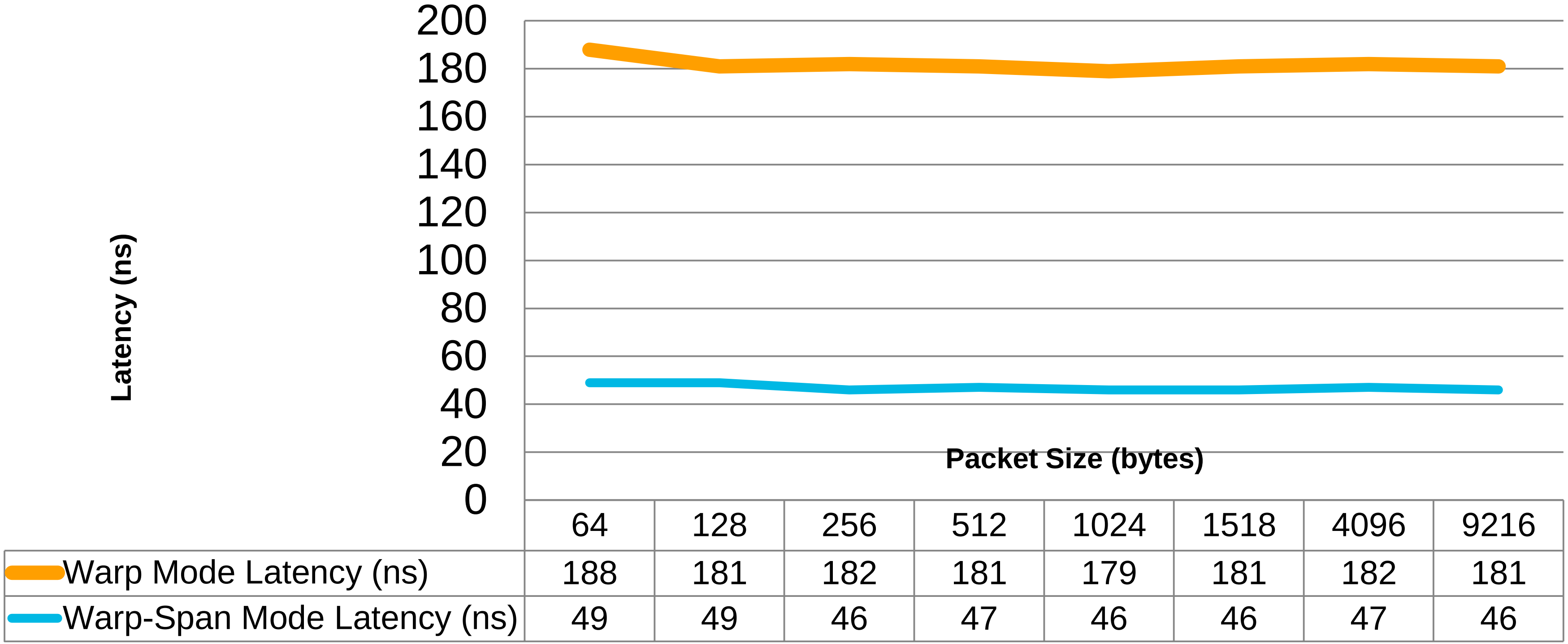
Benchmarking: Jitter

- IP Packet Delay Variation is commonly known as Jitter

- The jitter MUST be measured when sending packets of the same size.
- Jitter MUST be measured as packet to packet delay variation and delta between min and max packet delay variation of all packets sent.
- A histogram MAY be provided as a population of packets measured per latency or latency buckets.

Benchmarking Nexus 3548
10GE Latency and Jitter – L2 and L3 - RFC 2544

measured in FIFO



	64	128	256	512	1024	1518	4096	9216
Warp Mode Latency (ns)	188	181	182	181	179	181	182	181
Warp-Span Mode Latency (ns)	49	49	46	47	46	46	47	46

Spirent third party performance report for Nexus 3548

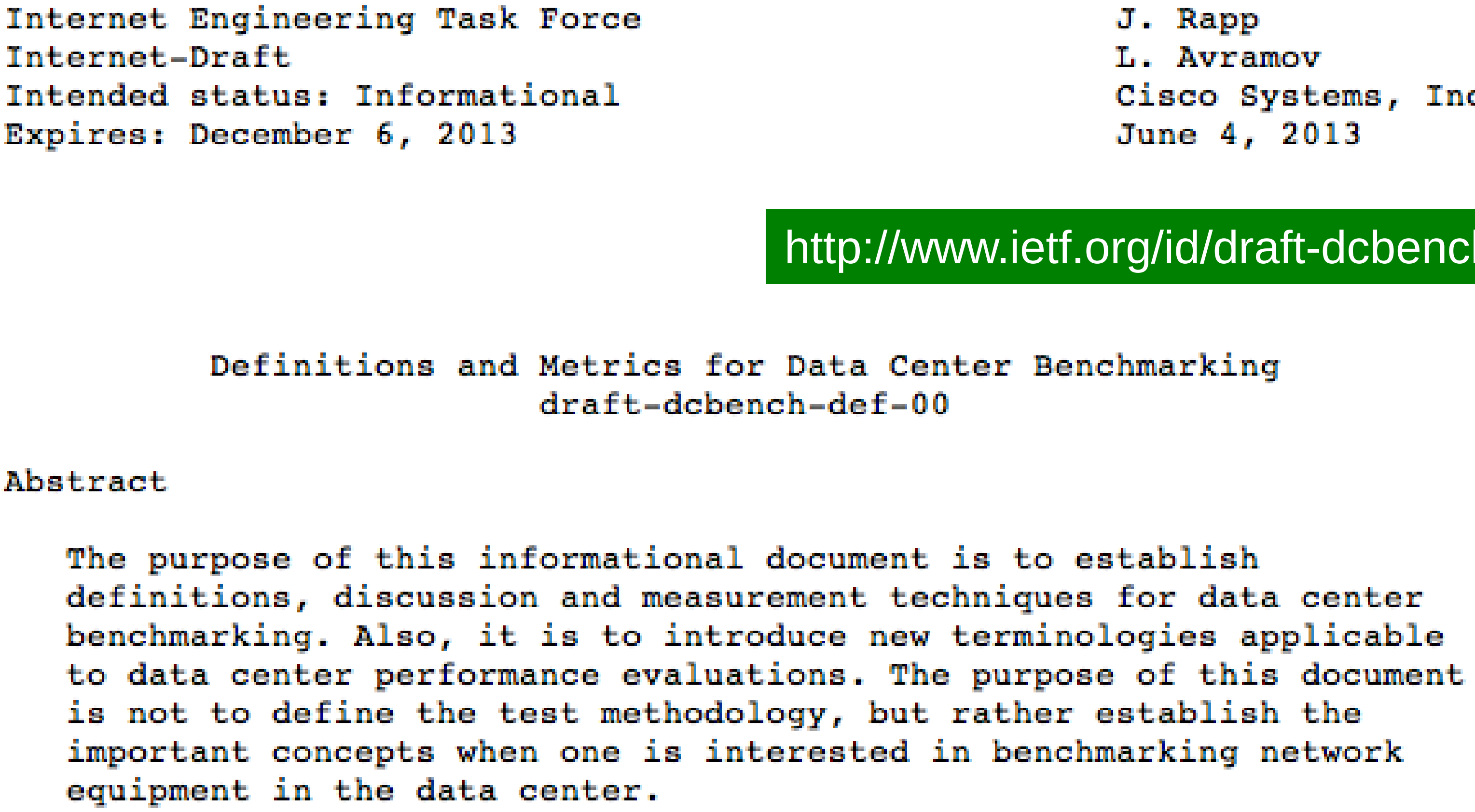
The MAX Latency value for 3548 and 3064 switches

	Control Protocols Running	MAX Latency (nsec)	AVG Latency (nsec)	MIN latency (nsec)
Nexus 3548 Layer 2 test	Default Enabled: LLDP/CDP/STP	~500	~250	~220
	Disabled: LLDP/CDP Enabled: STP	~340	~250	~220
	Disabled: LLDP/CDP/STP	~280	~250	~220
Nexus 3064 Layer 2 Test (64 B)	Default Enabled: LLDP/CDP/STP	~1120	~840	~800
	Disabled: LLDP/CDP Enabled: STP	~920	~840	~800
	Disabled: LLDP/CDP/STP	~860	~840	~800

Max latency increase due to control packets is more visible on ULL switches

RFC Draft !

Onto Data Center Benchmarking RFC



```
Internet Engineering Task Force                                J. Rapp
Internet-Draft                                              L. Avramov
Intended status: Informational                      Cisco Systems, Inc
Expires: December 6, 2013                                 June 4, 2013
```

http://www.ietf.org/id/draft-dcbench-def-00.txt

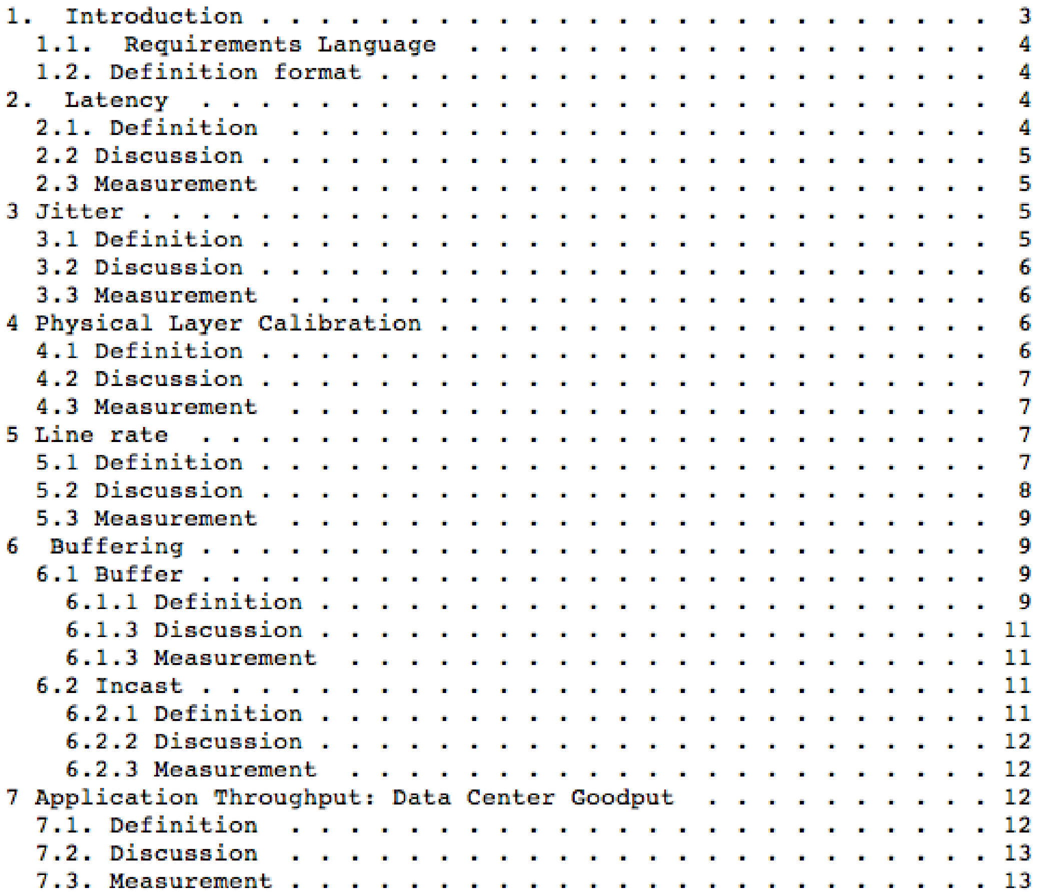
Definitions and Metrics for Data Center Benchmarking
draft-dcbench-def-00

Abstract

The purpose of this informational document is to establish definitions, discussion and measurement techniques for data center benchmarking. Also, it is to introduce new terminologies applicable to data center performance evaluations. The purpose of this document is not to define the test methodology, but rather establish the important concepts when one is interested in benchmarking network equipment in the data center.

RFC Draft !

Onto Data Center Benchmarking RFC

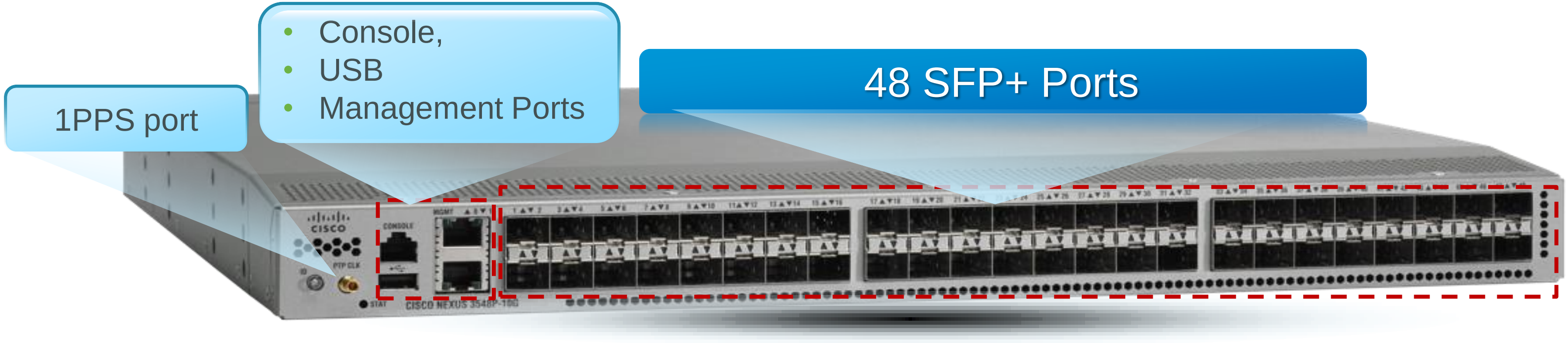


Provides definitions, explanations and guidance on key data center benchmarking topics

Agenda – Nexus 3548 – BRKARC-2013

- Benchmarking
- Architecture
  - **Product Overview**
  - Architecture
  - Features
  - Analytics
  - Scripting
- Designs

Nexus 3548 - Rear View

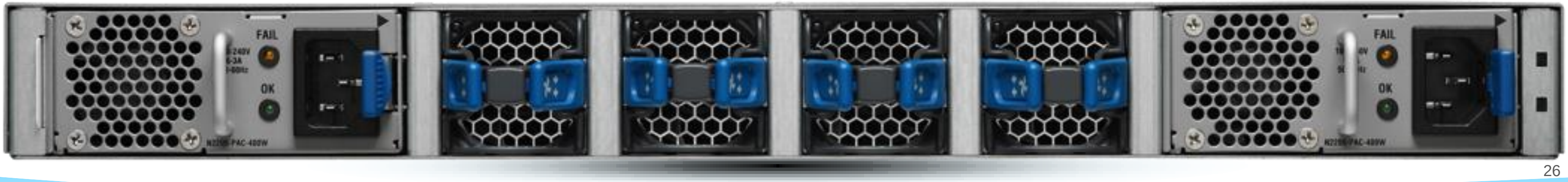


Rear View

Nexus 3548 – Front View

4 Individual Fan Trays

2 Redundant Power Supplies

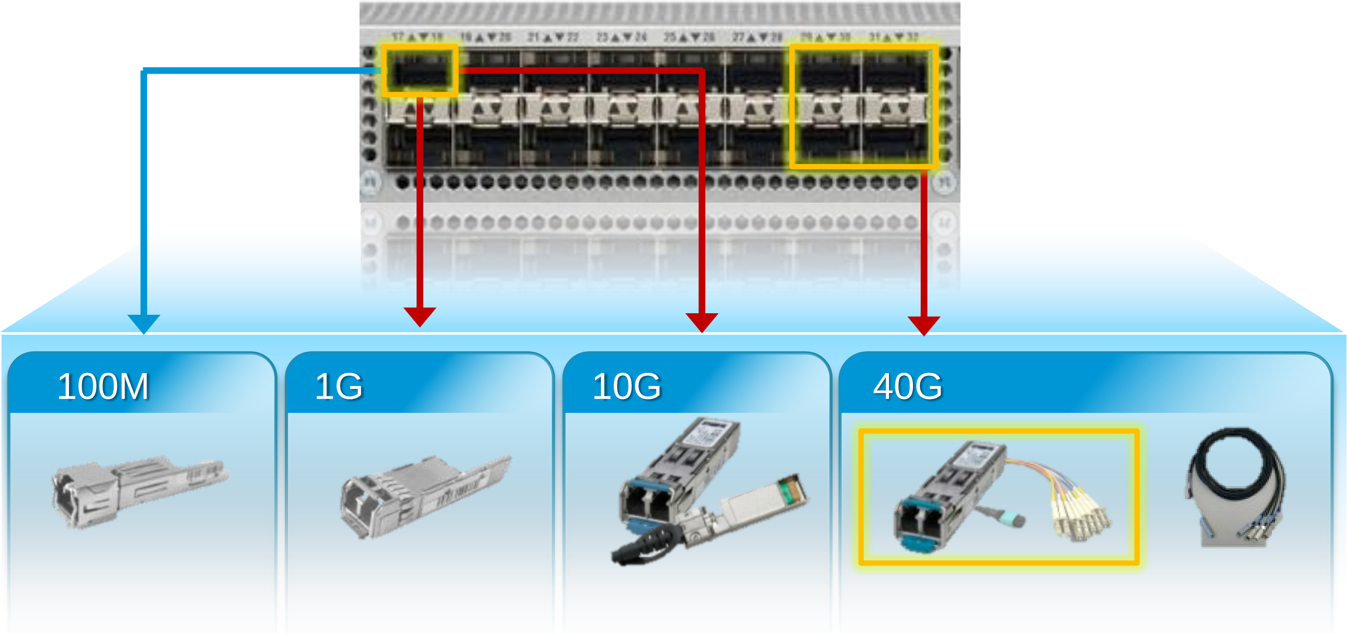


Front View

Size DxHxW	17x1.72x17.3 in (43 x 4.3 x 44 cm)
Weight	19 lbs (8.6 kg)

A/C and D/C Power with Forward and Reverse Airflow

Nexus 3548 – Physical Media Type

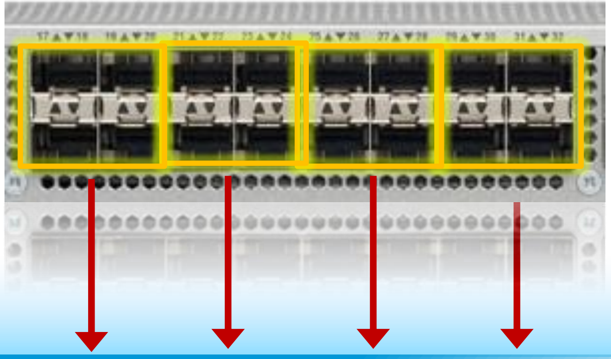


All Speeds from 100m to 40GE native are supported on same form-factor switch

Nexus 3548 – native 40GE cable options

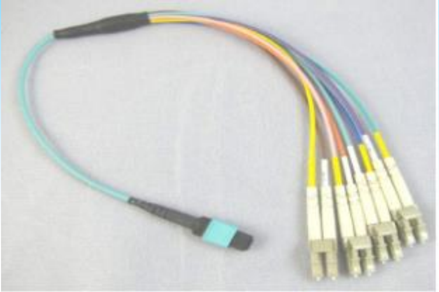
For Your Reference

```
3548(config)# interface eth 1/1
3548(config-if)# speed 40000
```



40Gi Breakout Fiber Cable Options

QSFP+ to SFP Copper Breakout Cable



Each 4 consecutive ports form a single native 40GE interface. No reload required.

Introduction to the Nexus 3548

Nexus 3548 Specifications

- 48x SFP+ – 100M / 1G / 10G / 40G
- Line rate L2/L3, Unicast & Multicast
- 18MB Packet Buffer
- 24K IPv4 Route, 8K MC, 64K Host
- 4K Flexible ACL / QoS
- Data Center TCP (DCTCP/ECN)

Algorithm Boost Features

- Ultra Low Latency – ~250 nsec
- WARP Mode - ~190nsec
- WARP SPAN - ~50nsec
- NAT @ Ultra Low Latency
- Active Buffer Monitoring
- Intelligent Traffic Mirroring
- IEEE-1588 PTP w/Pulse Per Second

Algo Boost Engine

Nexus 3548- Performance and Scalability

- Performance and Scalability

Hardware Feature	Value
Latency (64-9216Bytes)	~250 nanoseconds
Switching Capacity	960Gbps (720MPPS)
IPv4/v6 Routing Table	24K
IPv4/v6 Host Table	64K/16K
IP Multicast Routes	8K
IGMP Snooping Groups	8K
MAC address table	64K
ACL TCAM	4K
Supported L3 Interfaces	8K
VLANs	4,096
ECMP	32-Way
Etherchannel/Maximum members	48/32
VRF	4K
SPAN/ERSPAN Sessions	8 bidir
NAT Table	2K*

Please check Verified scalability Guide for more details.

Nexus 3548 – Software Features

Unicast Routing	Static, RIPv2, EIGRP, OSPF, BGP, HSRP, VRRP, 24K Routes, 64K Adjacencies
Multicast Routing	PIM-SM, SSM, IGMP v2/3, MSDP, 8K IGMP Groups, 8K Multicast routes, PIM Bi-dir
Layer2	RPVST+ (.1w), MST(.1s),STP extensions, LLDP, Storm Control, LACP, PVLAN*
Security & Segmentation	VRF-Lite, PACL, VACL, I/E Routed ACLs, Unicast RPF (uRPF), Static NAT
QoS	Modular QoS CLI, SP/Deficit Weighted Round Robin (DWRR), Classification, Marking, ECN, CoPP, PFC, Flow Control
System Management	AAA, SPAN, CallHome, SNMP, PTP, ERSPAN

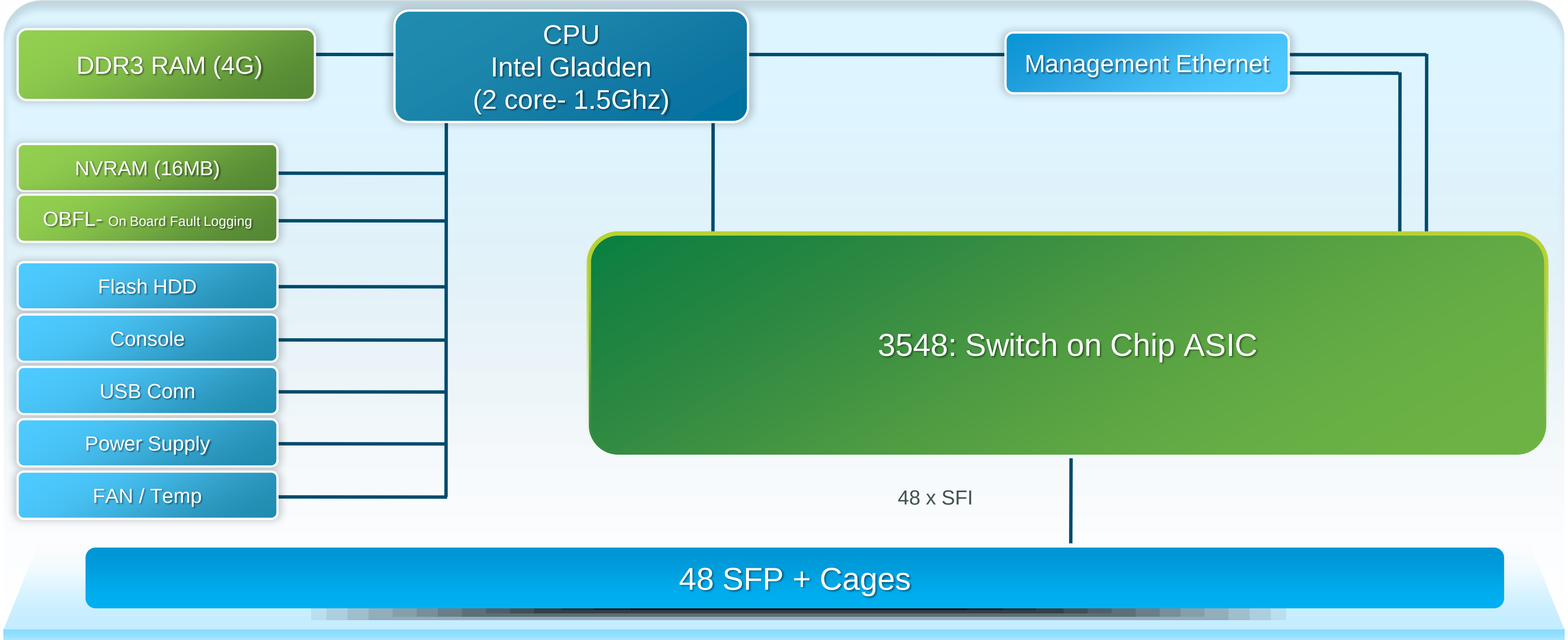
*Post-FCS

Agenda – Nexus 3548 – BRKARC-2013

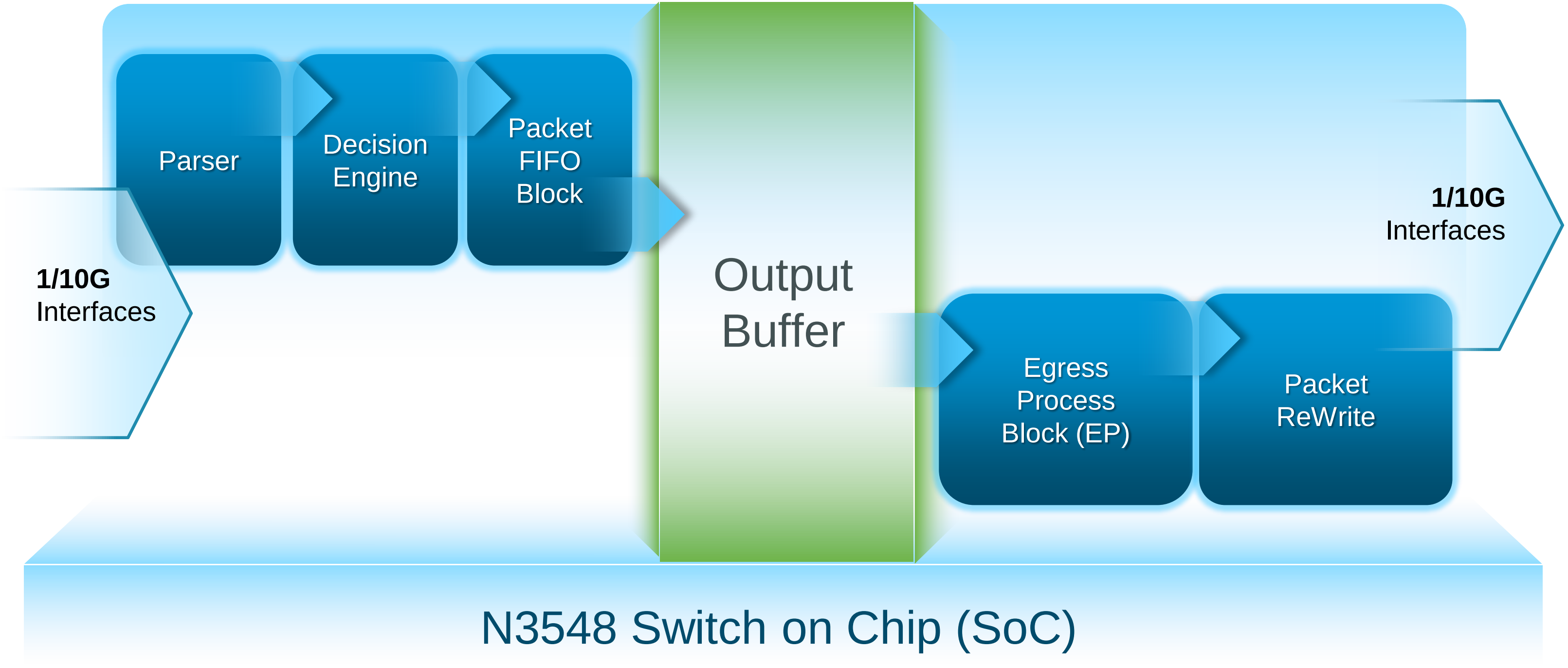
- Benchmarking
- Architecture
  - Product Overview
  - **Architecture**
  - Features
  - Analytics
  - Scripting
- Designs

Cisco live!

Nexus 3548– Data Plane and SoC Architecture



Nexus 3548 Packet Flow



Nexus 3548 Forwarding Paths

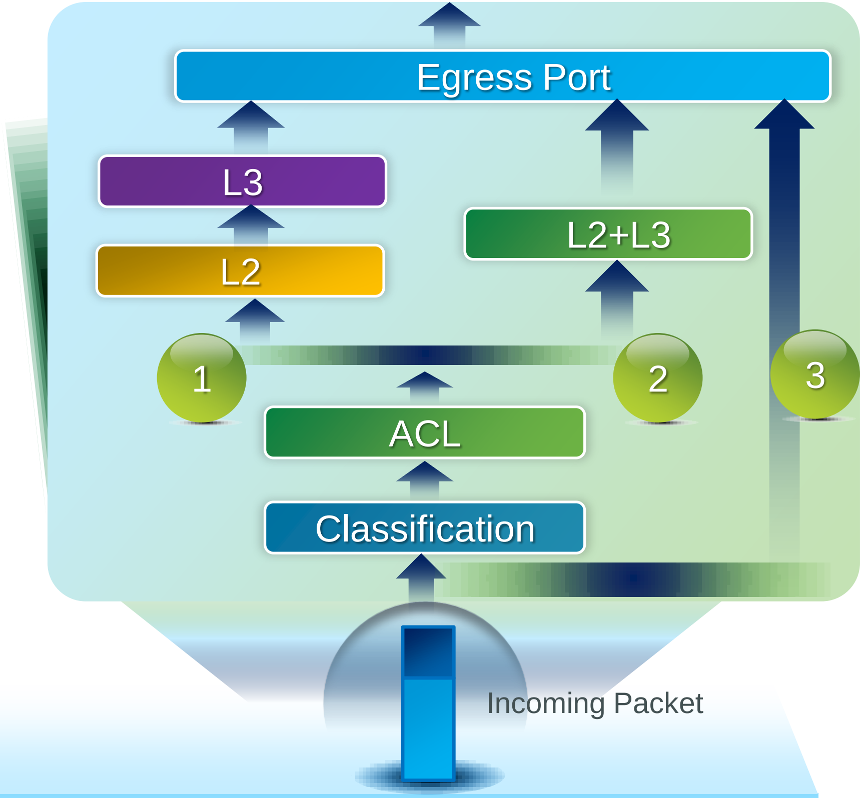
Logical Diagram

Forwarding Paths:

1 Normal

2 WARP

3 WARP SPAN

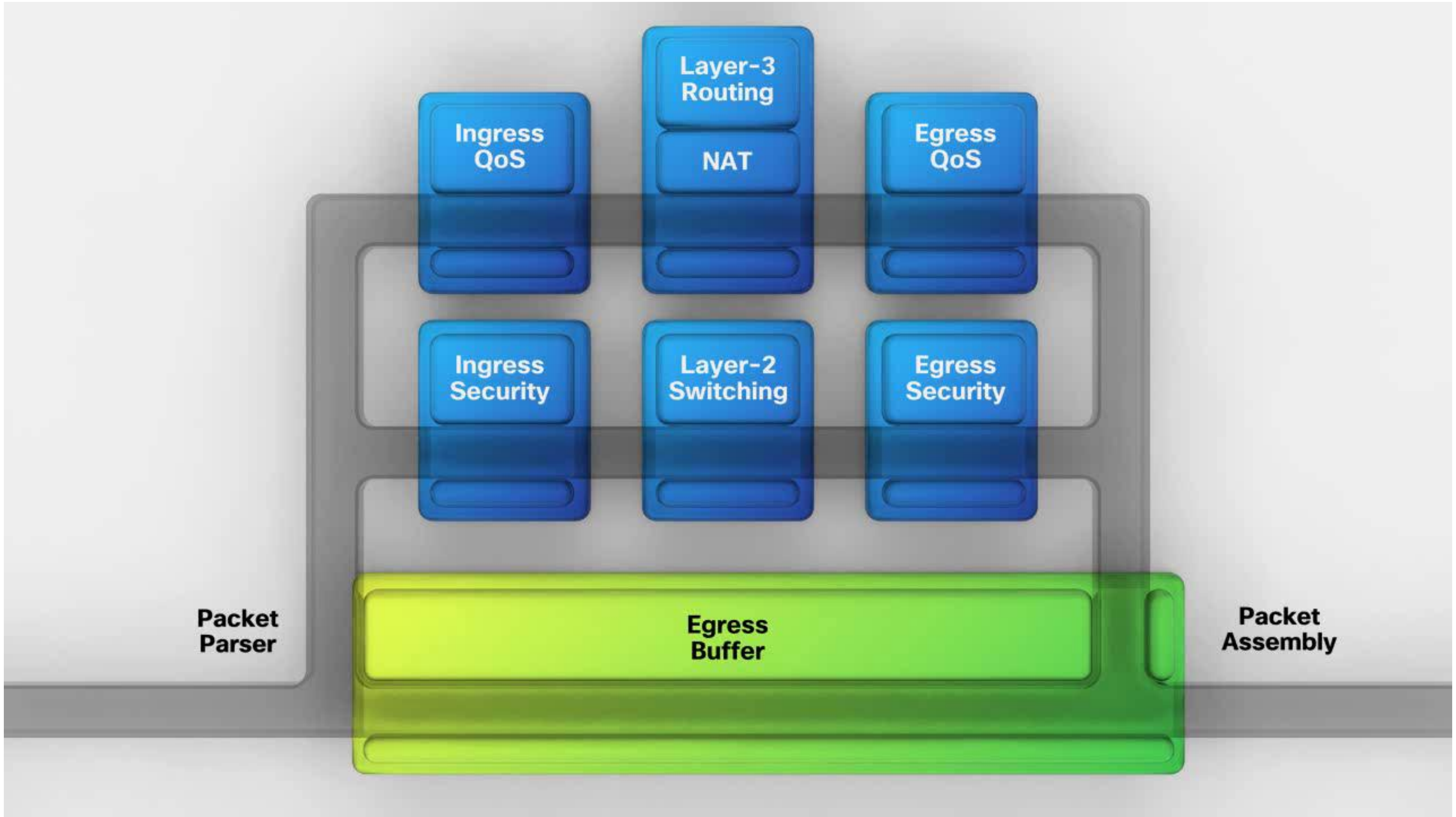


Nexus 3548 Forwarding Paths

Comparison Normal/Warp Mode

Feature	Normal Mode	Warp Mode
Latency (FIFO)	~250 nsec	~190 nsec
NAT	Yes	Yes
Ingress RACL/VACL	Yes	Yes
Multicast Routes	8K	8K
Unicast Route	24K	4K
Host Route and MAC Table	64K each	8K Each
L3 ECMP	Yes	No
Egress ACL/PACL	Yes	No

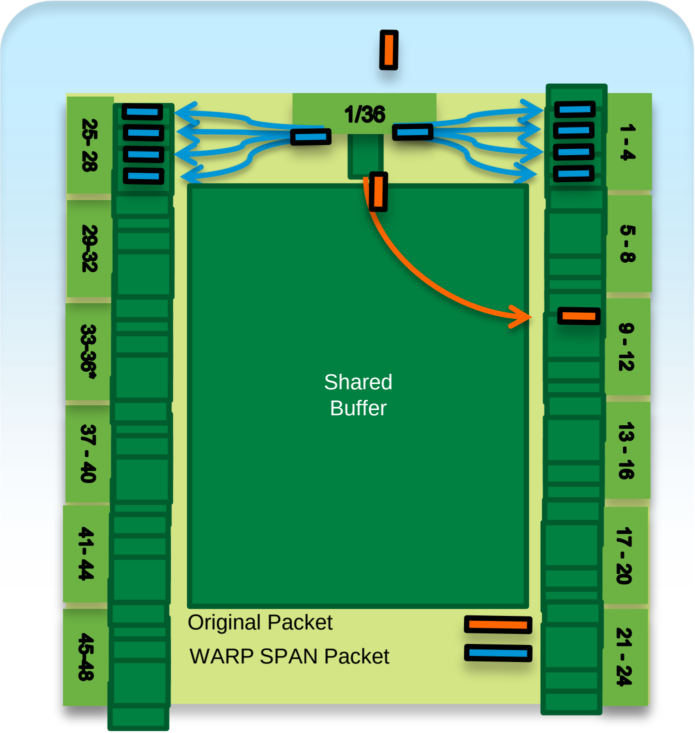
Design Consideration #5 – Feature Set – Impact on Maximum Latency



Nexus 3548 Differentiators

WARP SPAN

- WARP SPAN enables mirroring of all the ingress traffic on a dedicated port to user configurable group of ports
- WARP SPAN can be enabled both in normal and WARP mode
- The Latency of the WARP SPAN'd packets would be ~50 nanosec
- WARP SPAN source has to be port Ethernet 1/36
- WARP SPAN destination would be group of 4 ports as shown.
- WARP span source and destination ports has to be 10Gig, no mix of 1Gi and 10Gi

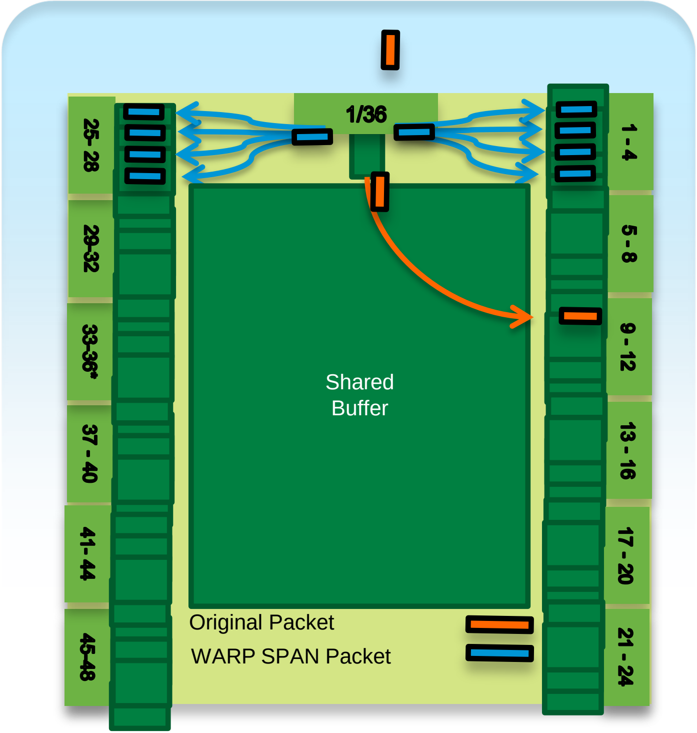


Nexus 3548 Differentiators

WARP SPAN

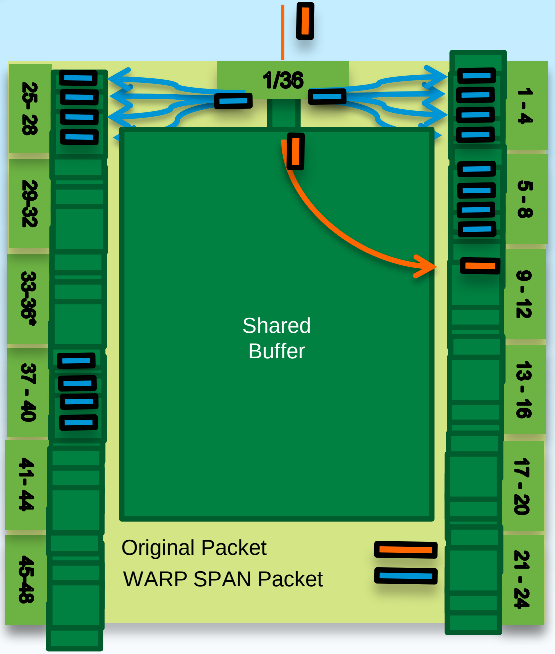
- The traffic received on the WARP SPAN source will be forwarded normally along with the WARP SPAN.

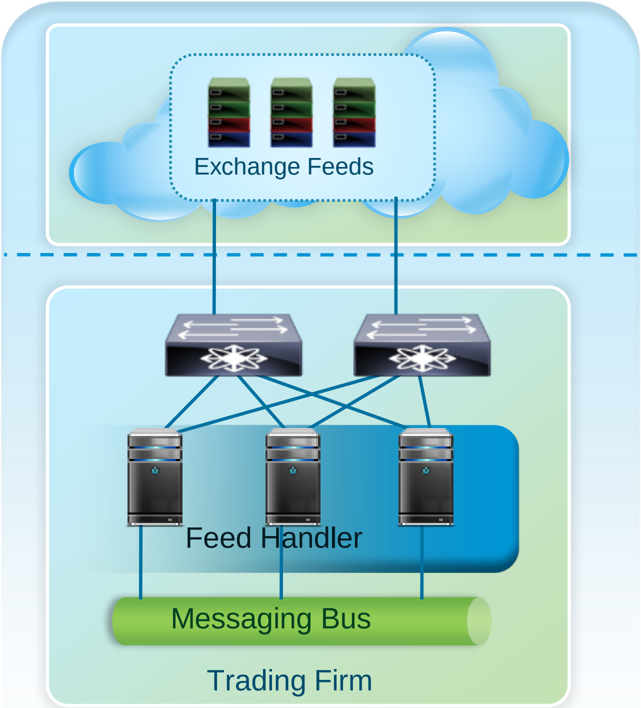
- WARP destination ports are dedicated destination port. These ports cannot be used to receive traffic. However, Other ports in the switch can be used as normal L2/L3 ports.



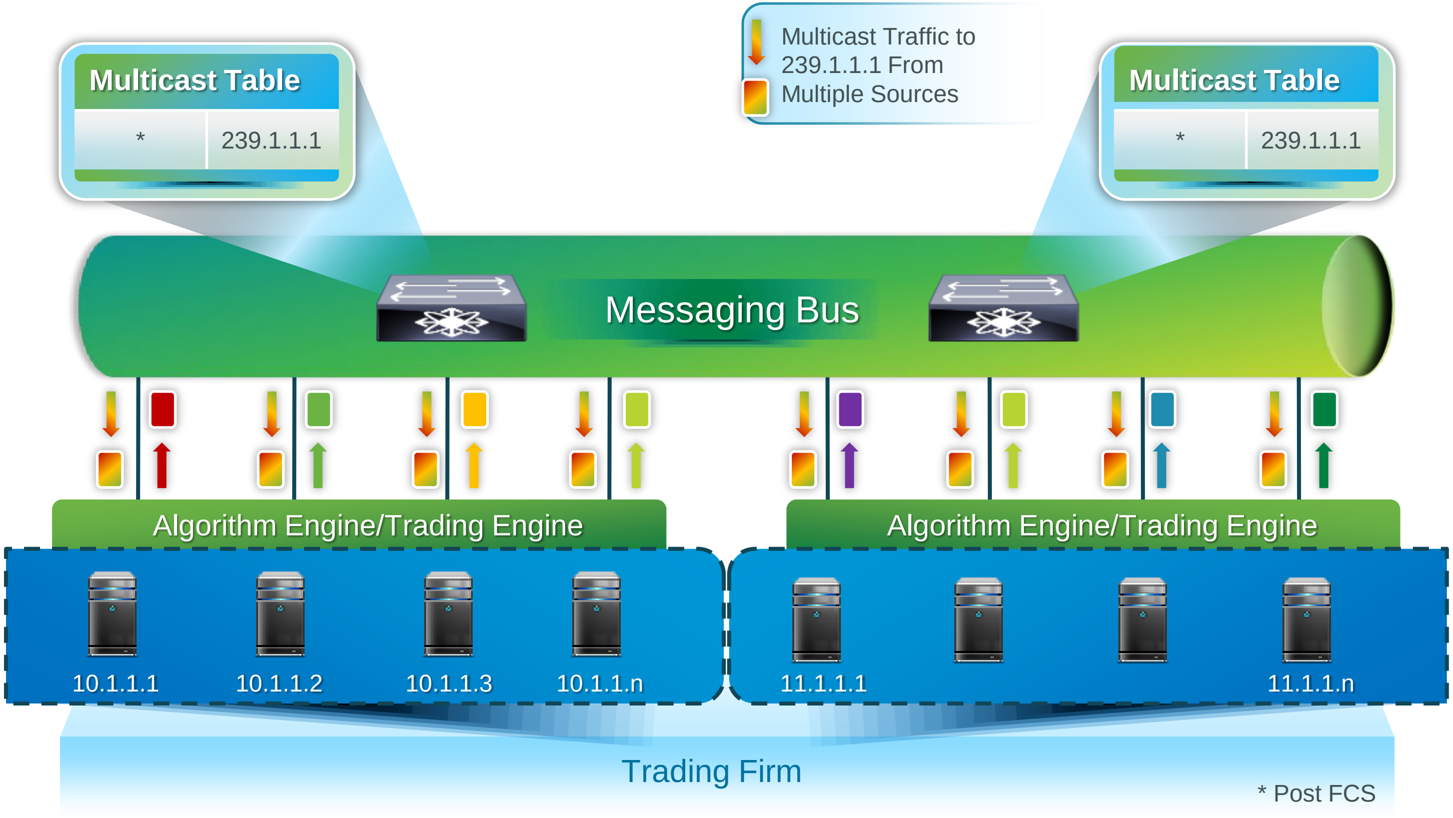
Nexus 3548 Differentiators

WARP SPAN – Use Case – Serving Feed Handler





Nexus 3548 Multicast: adding Pim Bi-Dir



* Post FCS

Nexus 3548 Multicast enhancements

- PIM Bi-Dir support
- Separate Multicast CoPP policer for RPF failure, Source Registration & Other Multicast traffic
- In addition, Bloom filter to limit RPF failure traffic to CPU
- Per multicast group traffic counter
- Per input port Multicast RPF failure counter*

QoS

Nexus 3548 QOS

Classification and Marking

Classification

- Classification is done using ACL TCAM (256 ACE)
- Supports classification based on:
  - Layer2 CoS
  - Layer3 Prec/DSCP
  - IPv4 ACL
  - IPV6 ACL/MAC ACL/VLAN ACL*

Marking

- Layer2 COS
- Layer3 Prec/DSCP

Nexus 3548 QOS

Congestion Avoidance and CoPP

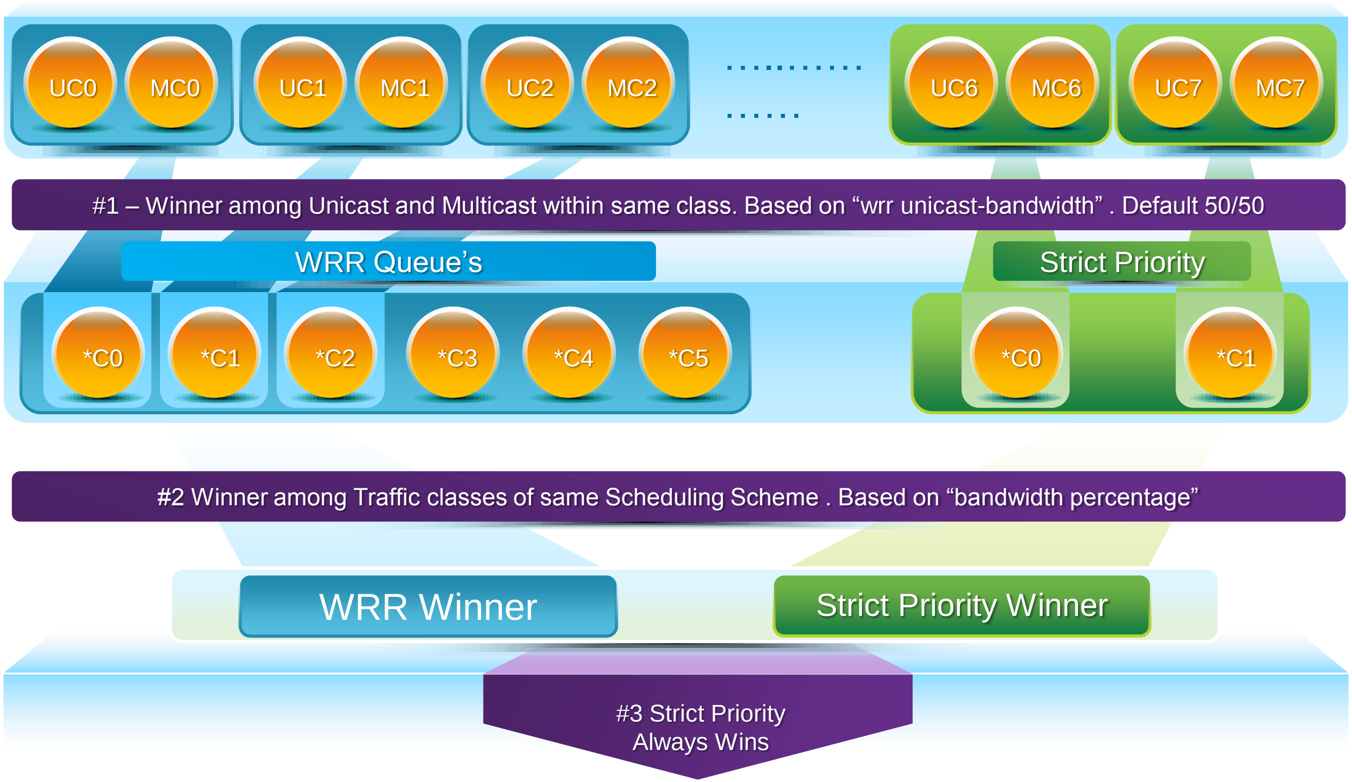
Congestion Avoidance

- By Default WRR queue does tail drop if congestion is experienced
- ECN (DCTCP) marking is done if queue reaches the SW configured threshold

Control Plane Policing (CoPP)

- 64 CoPP Policer
- Separate Multicast CoPP policer for
  - RPF failure,
  - Source Registration
  - Other Multicast traffic.
- Bloom filter to limit RPF failure traffic to CPU

Nexus 3548 – Multilevel Scheduler



#1 – Winner among Unicast and Multicast within same class. Based on "wrr unicast-bandwidth" . Default 50/50

#2 Winner among Traffic classes of same Scheduling Scheme . Based on "bandwidth percentage"

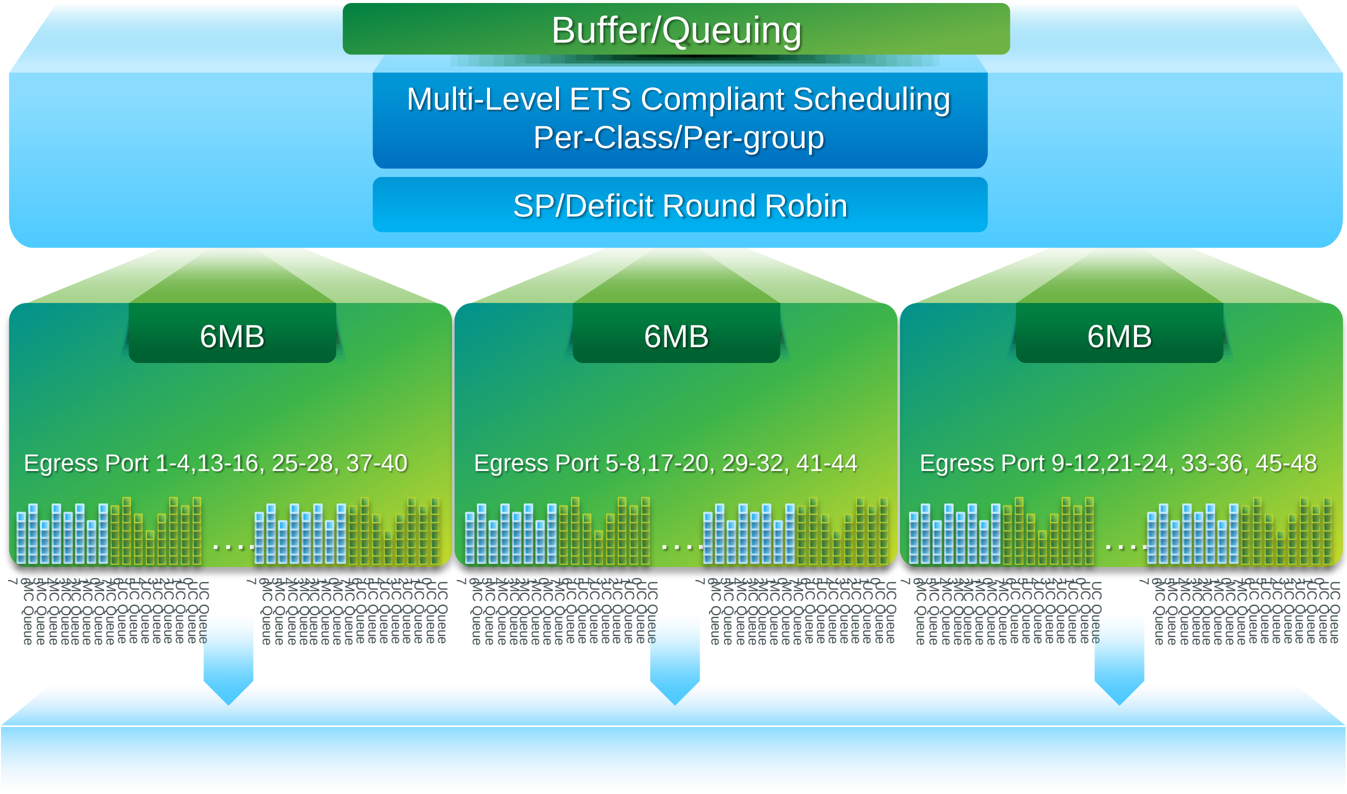
WRR Winner

Strict Priority Winner

#3 Strict Priority
Always Wins

Nexus 3548 QOS

Buffer/Queuing



Buffer/Queuing

Multi-Level ETS Compliant Scheduling
Per-Class/Per-group

SP/Deficit Round Robin

6MB

6MB

6MB

Egress Port 1-4,13-16, 25-28, 37-40

Egress Port 5-8,17-20, 29-32, 41-44

Egress Port 9-12,21-24, 33-36, 45-48

Nexus 3548 QOS

CLI Knob for Buffer tuning

Per Class shared buffer threshold:

```
policy-map type network-qos <policymap-name>
  class type network-qos <class>
    queue-threshold <percentage>
```

Per Port/class buffer threshold:

```
policy-map type network-qos <policymap-name>
  class type network-qos <class>
    port-queue-threshold <percentage>
```

SPAN buffer threshold:

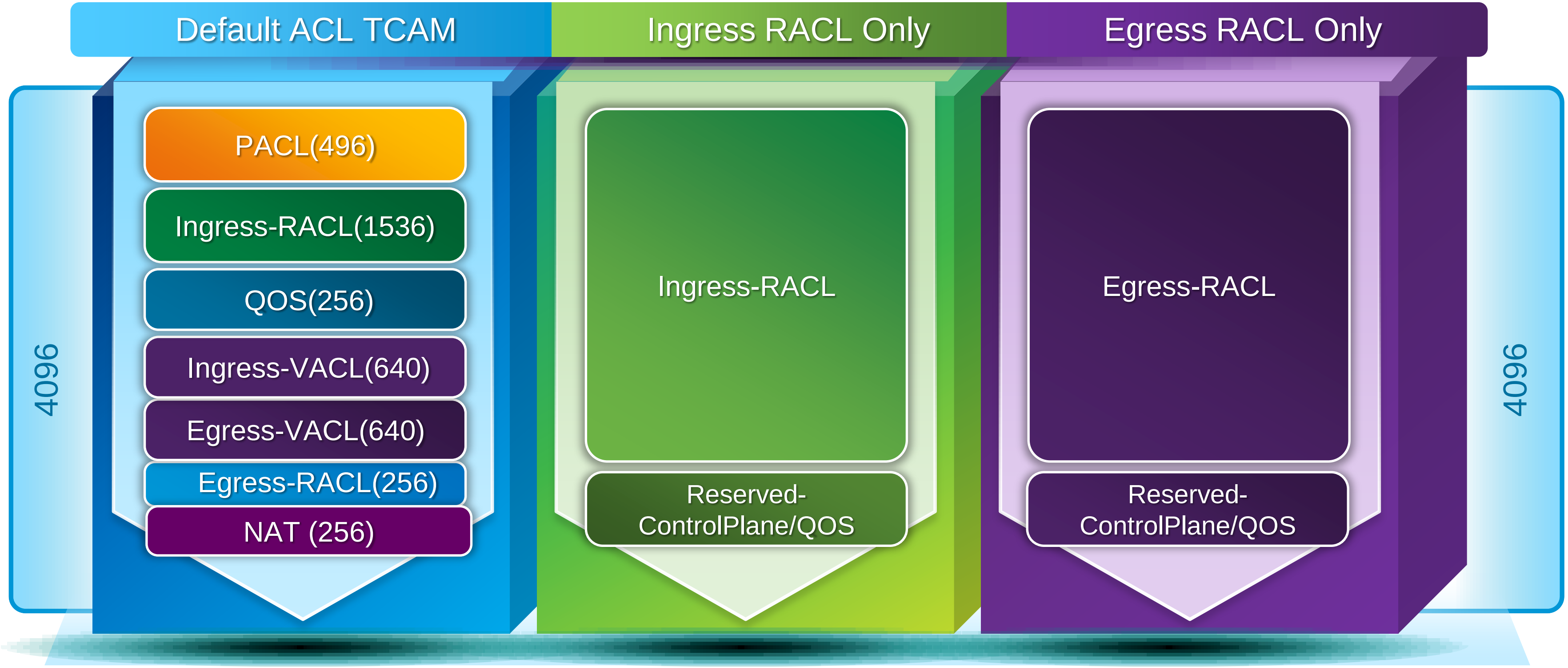
```
hardware profile buffer span threshold <percentage>
```

Agenda – Nexus 3548 – BRKARC-2013

- Benchmarking
- Architecture
  - Product Overview
  - Architecture
  - **Features**
  - Analytics
  - Scripting
- Designs

Nexus 3548 Features

Flexible ACL TCAM



Default ACL TCAM

4096

PACL(496)

Ingress-RACL(1536)

QOS(256)

Ingress-VACL(640)

Egress-VACL(640)

Egress-RACL(256)

NAT (256)

Ingress RACL Only

Ingress-RACL

Reserved-ControlPlane/QOS

Egress RACL Only

Egress-RACL

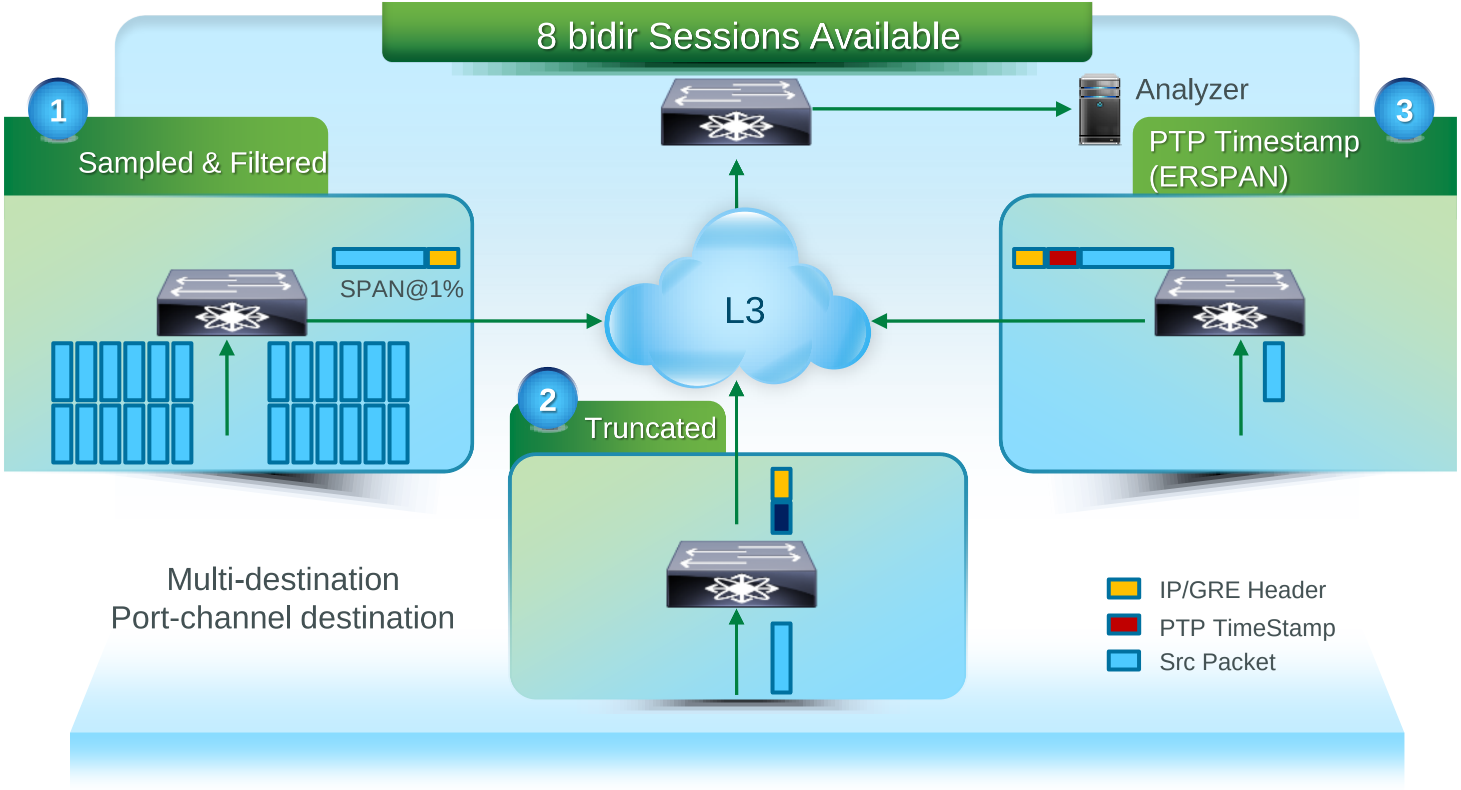
Reserved-ControlPlane/QOS

4096

hardware profile tcam region {**arpacl** | **e-racl**} |**ifacl** | **ipsg** | **nat** | **qos**} |**qoslbl** | **racl**} | **vacl** }*tcam_size*\

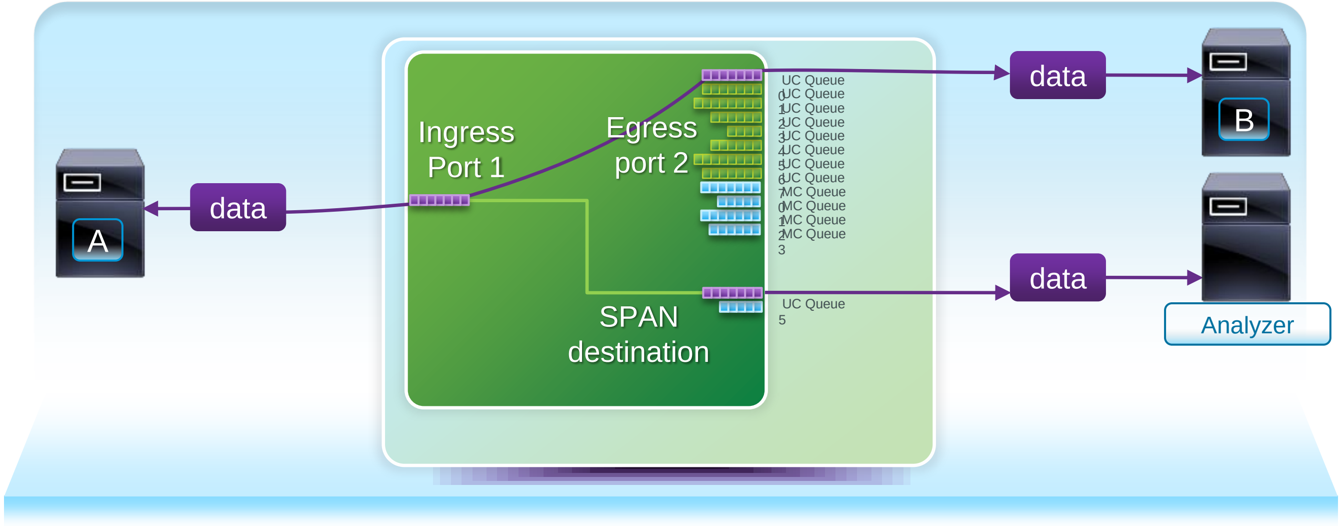
Nexus 3548 Features

Advanced SPAN and ERSPAN



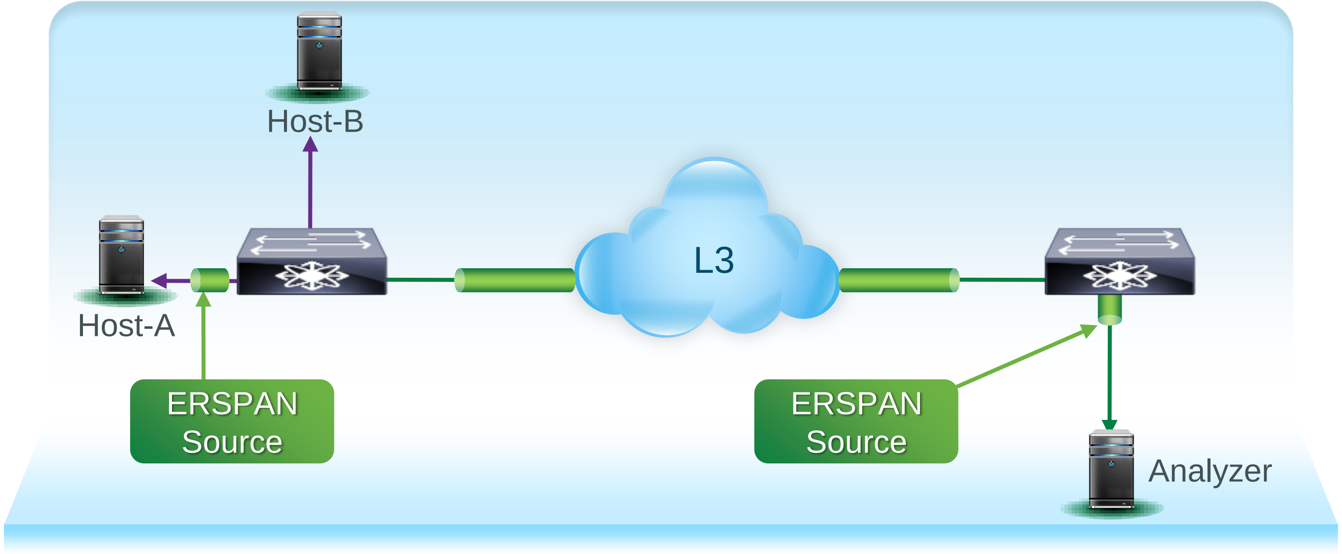
Nexus 3548 SPAN Flow

- Traffic to be replicated is marked in the ingress flow
- The replication occurs in the switch ASIC, there is no latency impact on the product traffic
- Dedicated queue (queue 5) for SPAN traffic
- SPAN buffer utilization is limited by threshold, max 200 pages (38KB), and this threshold is configurable



ERSPAN

- Encapsulated Remote Switch Port Analyzer (ERSPAN) allows the analyzer to be placed on one location and multiple switches can send mirrored traffic to this analyzer
- Allows to analyze traffic from any port on the network on any remote switch without physically moving the analyzer tool



NAT

NAT

Different Flavors

Static NAT:

This provides option to configure static mapping between local and global addresses and UDP/TCP ports (in case of static PAT)

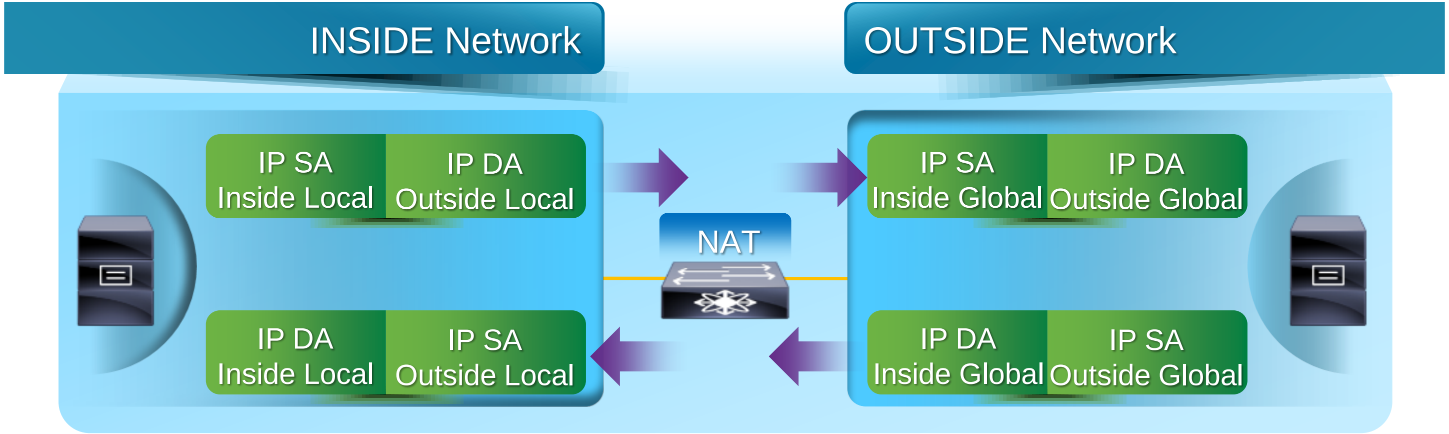
Dynamic Address Translation:

The user can establish dynamic mapping between the local and global addresses, by describing the local addresses to be translated and the pool of addresses from which to allocate global addresses, and associating the two.

Port Address Translation (PAT):

This provides option to map multiple IPv4 address to fewer number of IPv4 address using different TCP/UDP port numbers

Nexus 3548 - NAT Terminology



Inside local address—The actual IP address assigned to the host.

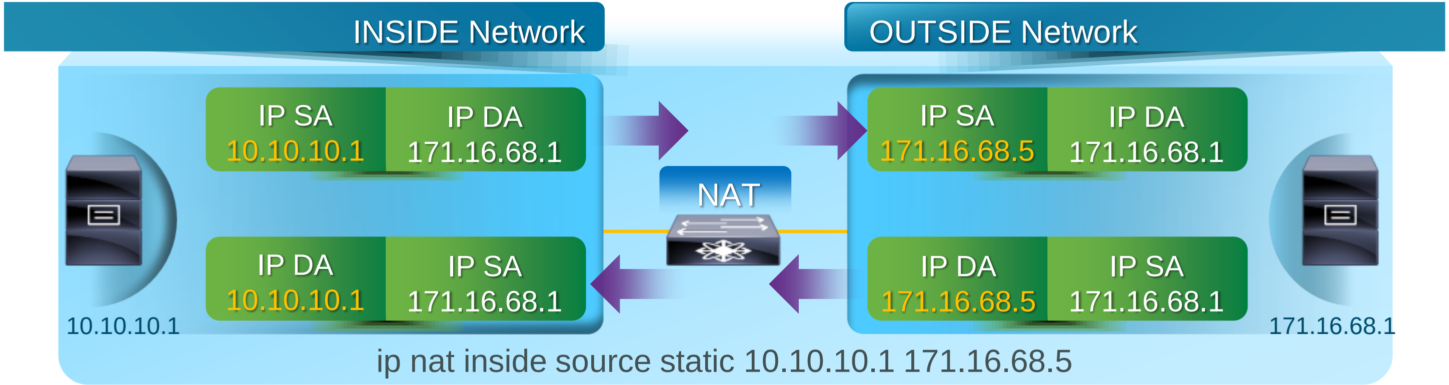
Outside local address—The IP address of an outside host as it appears to the inside network.

Inside global address—A legitimate IP address assigned by the service provider that represents one or more inside local IP addresses to the outside world.

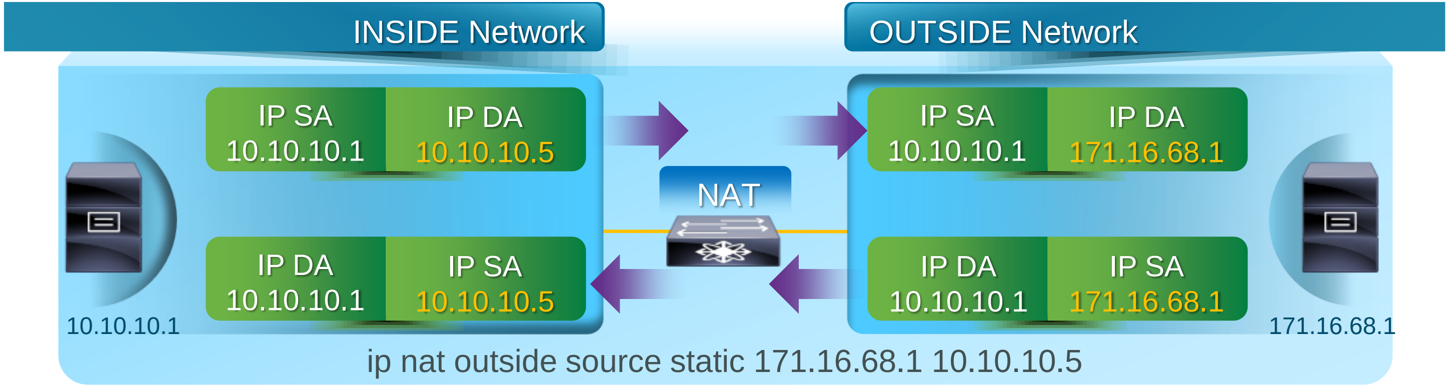
Outside global address—The actual IP address assigned to a host on the outside network.

NAT Example

Inside/Outside Source Translation



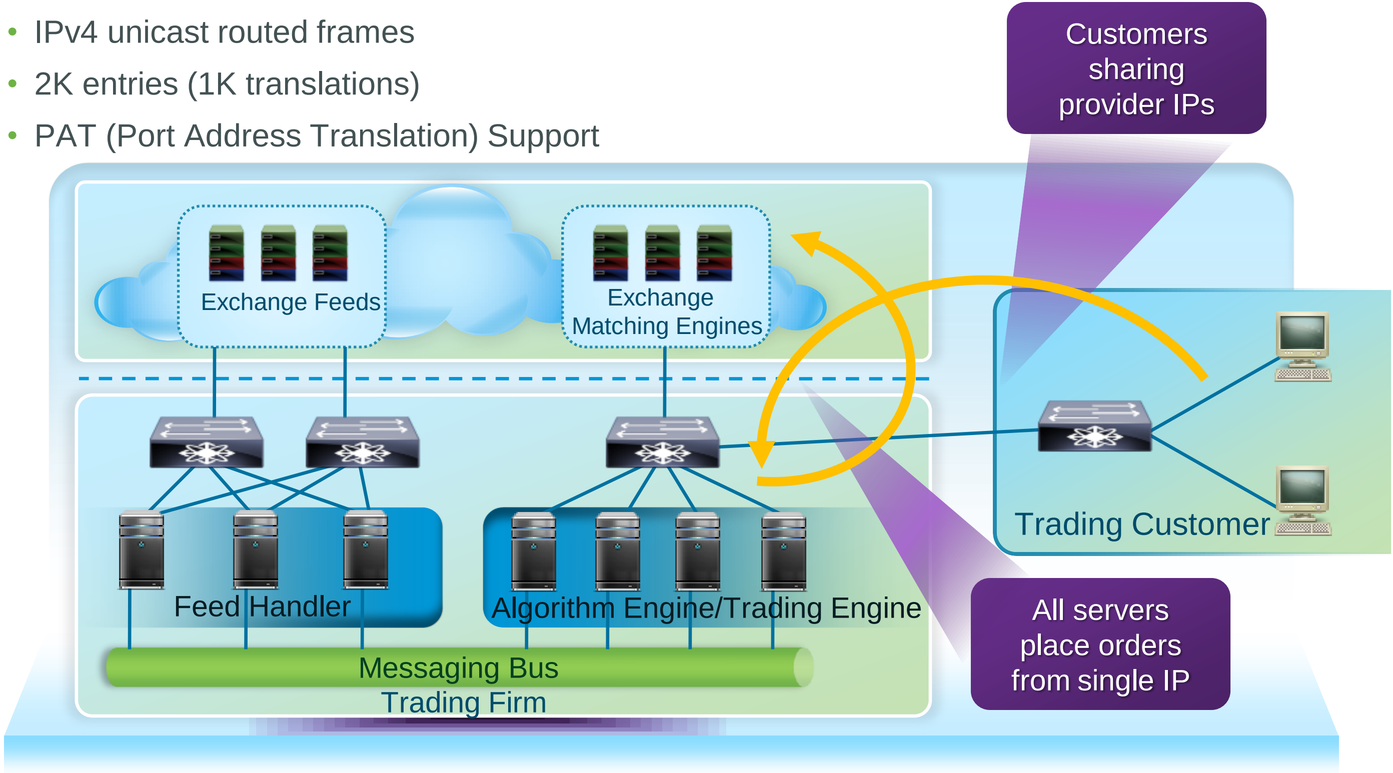
```
ip nat inside source static 10.10.10.1 171.16.68.5
```



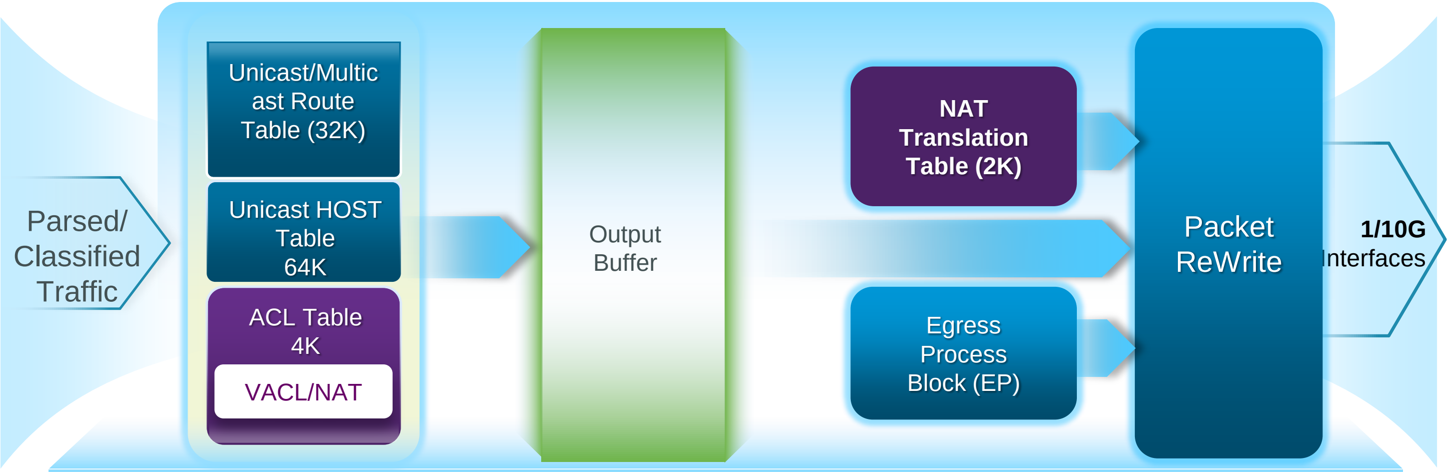
```
ip nat outside source static 171.16.68.1 10.10.10.5
```

Nexus 3548 NAT

- IPv4 unicast routed frames
- 2K entries (1K translations)
- PAT (Port Address Translation) Support


Customers sharing provider IPs
Exchange Feeds
Exchange Matching Engines
Trading Customer
Feed Handler
Algorithm Engine/Trading Engine
All servers place orders from single IP
Messaging Bus
Trading Firm

Nexus 3548 NAT Implementation


Parsed/ Classified Traffic
Unicast/Multicast Route Table (32K)
Unicast HOST Table 64K
ACL Table 4K
VACL/NAT
Output Buffer
NAT Translation Table (2K)
Egress Process Block (EP)
Packet ReWrite
1/10G Interfaces

N3548 NAT/PAT Classification and Translation

- NAT uses VACL space for classifying and identifying the traffic for NAT translation based on ingress interface
- NAT translation table would provide actual translation info for packet ReWrite block for packet modification before sending the packet out of NAT interface
- For Static NAT, ACL and Translation Table are updated as soon as the NAT static config is added
- For dynamic NAT*, first packet is punted to CPU after ACL classifies it to be NAT flow and then software updates the translation table based on the flow info

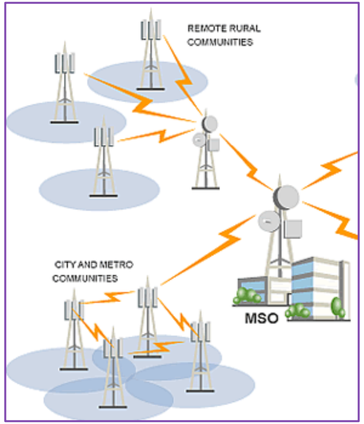
* Post FCS

IEEE 1588 PTP

IEEE 1588 – PTP - Application Precision

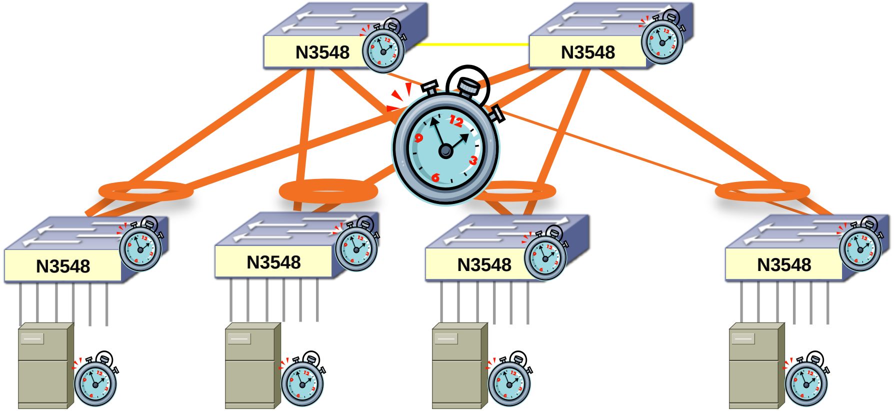
- Precision Time Protocol: IEEE 1588v2
- Nanosecond Precision

```
Meinberg Lantime M600 GPS
GPS: NORMAL OPERATION
NTP: Offset PPS: -10ns PTP:ok
Tue, 13.06.2006 UTC 09:06:52
```



Telecommunications

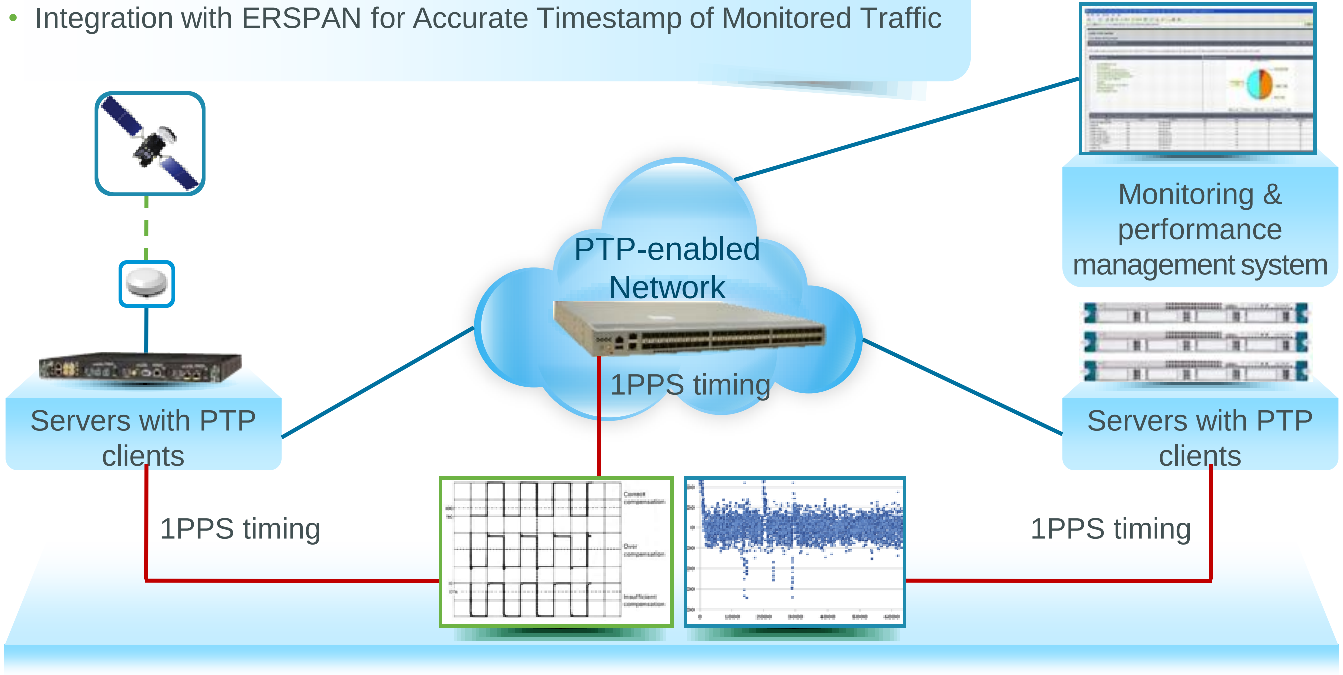
Financial trading



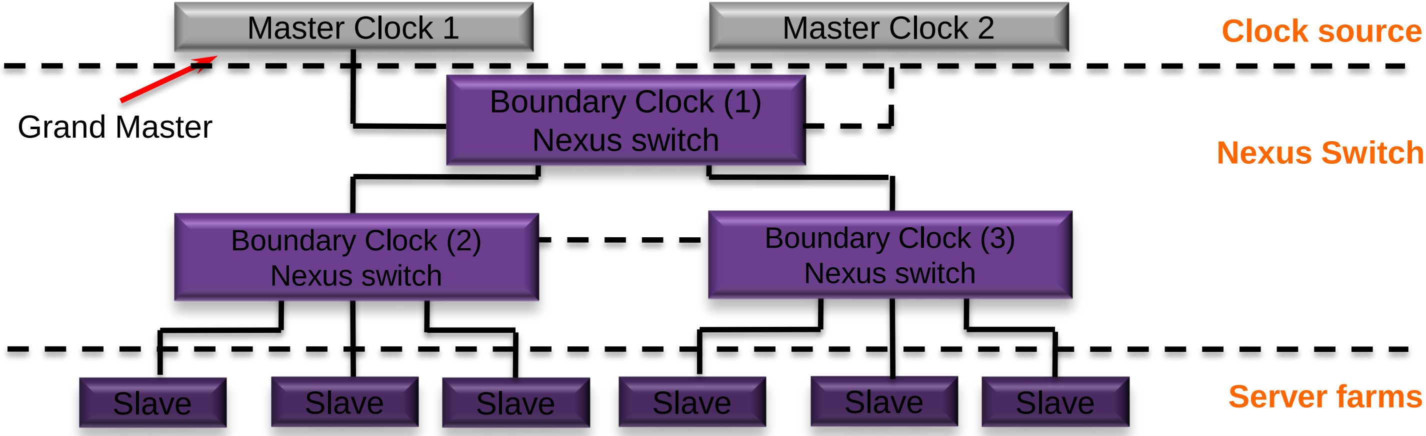
IEEE 1588 Implementation in N3548

Applications @ Switch

- Verify accuracy with 1PPS output
- Integration with ERSPAN for Accurate Timestamp of Monitored Traffic



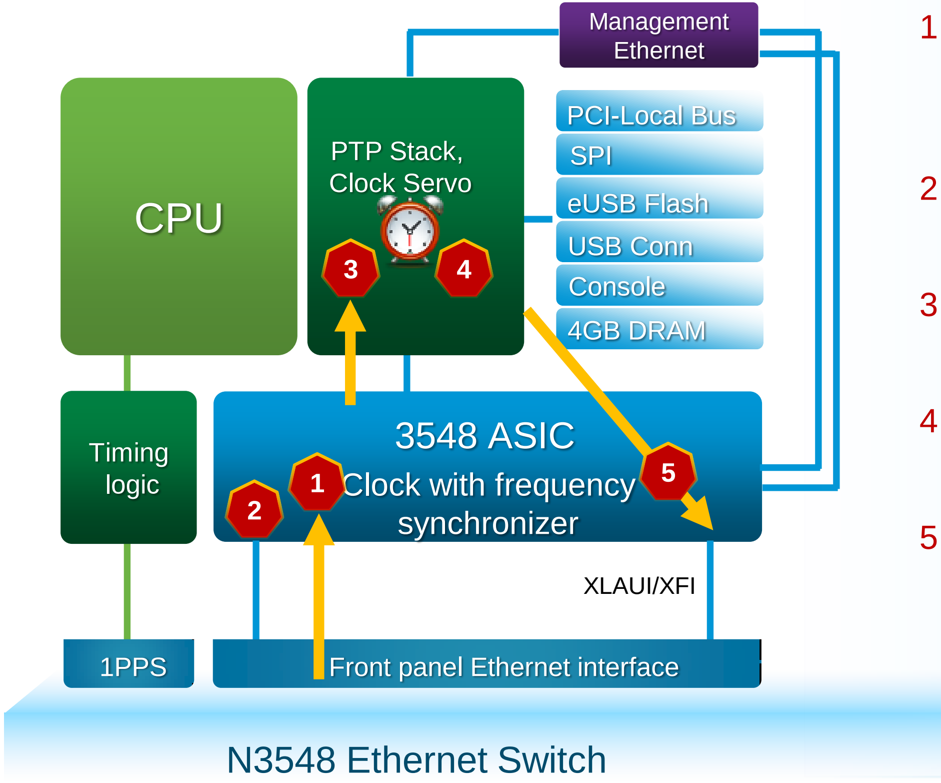
IEEE 1588 Implementation in N3548



1 Elect the grand master, form a master-slave hierarchy. Grand master is selected based on Best Master Clock selection Algorithm (BMCA). (Master clock 1 is selected as Grand Master in the diagram)

2 Each slave clock synchronizes itself to the master clock

IEEE 1588 Implementation in N3548

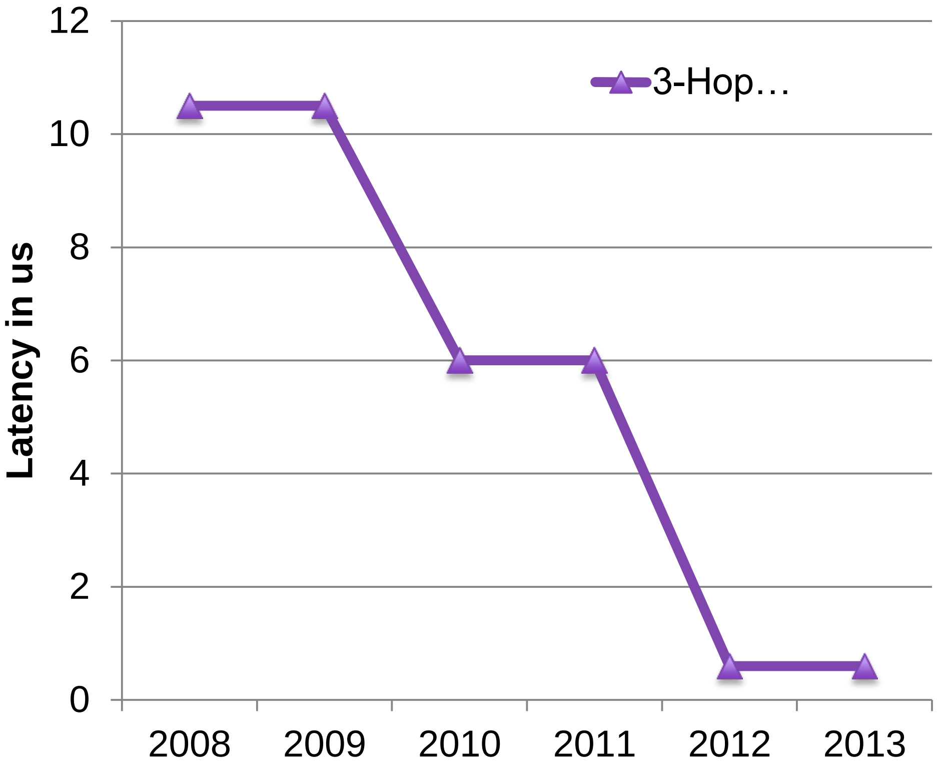


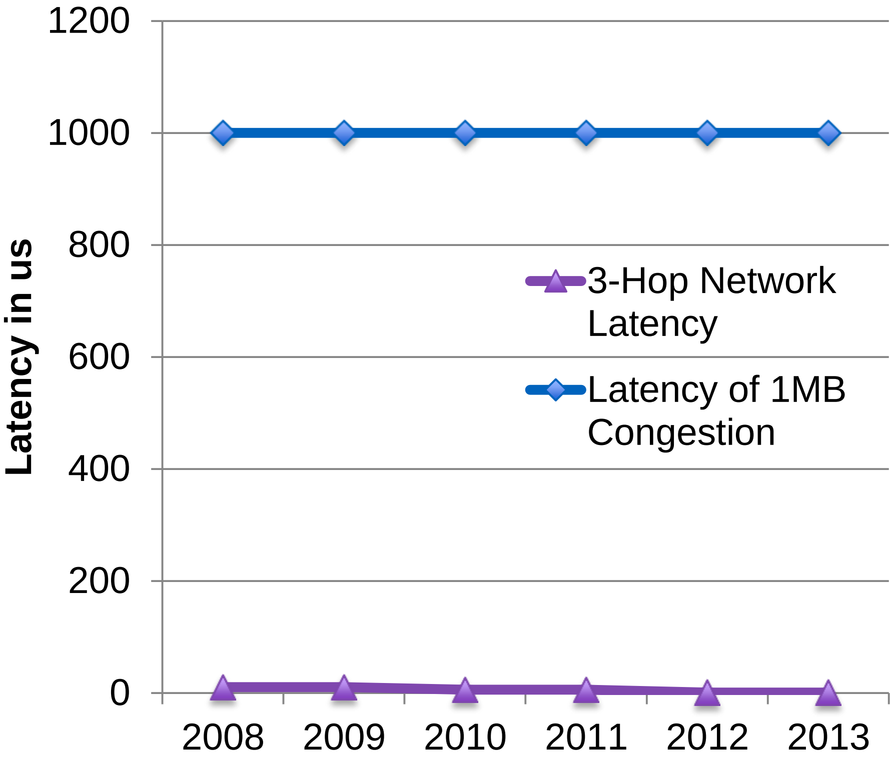
1. 1588 packet is timestamped at ingress of ASIC to record the arrive time (t_2)
2. Timestamp points to the first bit of the packet (following SFD)
3. Packet is copied to CPU with timestamp and destination port
4. The packet goes through PTP stack and other process
5. The packet is sent out at egress port. (The corresponding timestamp for the TX packet is available from the FIFO TX time stamp) ASIC records the packet’s departure timestamp and delivers it to the PTP stack.

Agenda – Nexus 3548 – BRKARC-2013

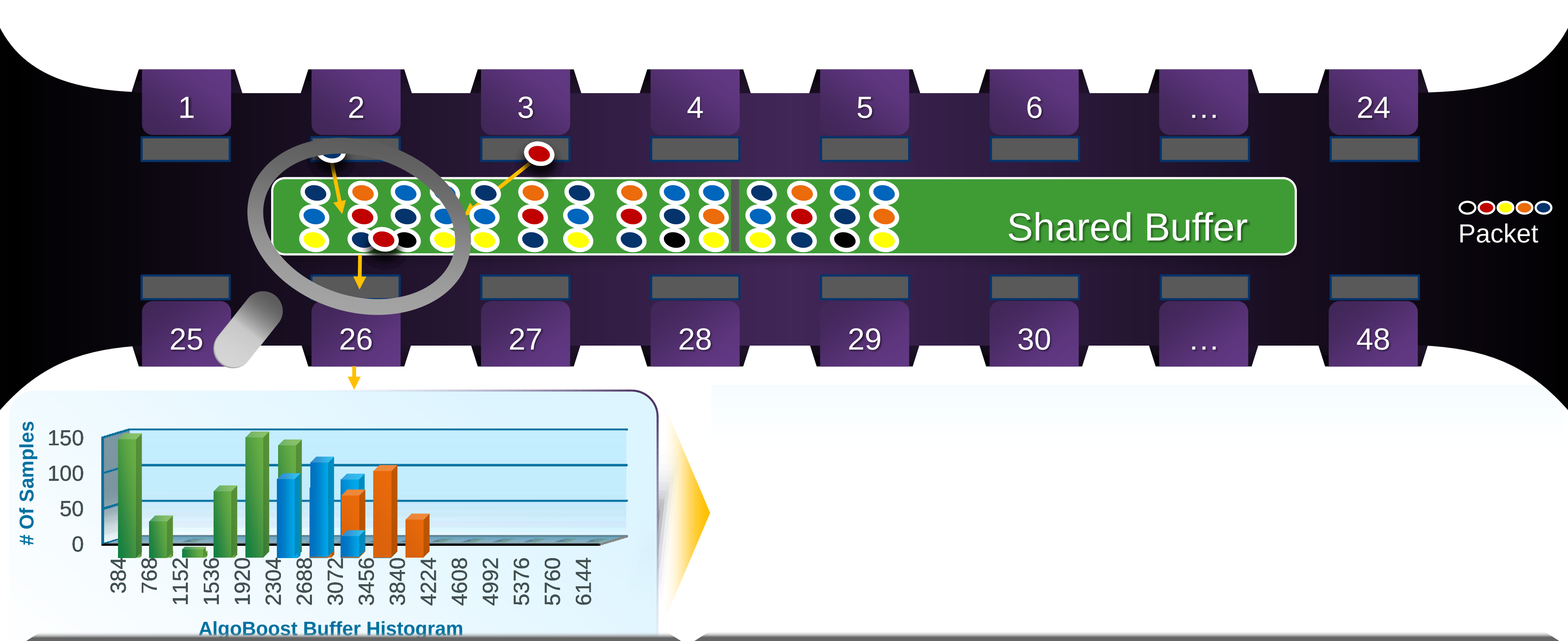
- Benchmarking
- Architecture
  - Product Overview
  - Architecture
  - Features
  - **Analytics**
  - Scripting
- Designs

Why are analytics important?





Nexus 3548 – Active Buffer Monitoring [ABM]



Hardware Polling

Software Polling

N3548 – ABM Output

```
Nexus3548# sh hardware profile buffer monitor
brief
Brief CLI issued at: 09/10/2012 22:15:34
                  Maximum buffer utilization detected
                 1sec     5sec    60sec     5min      1hr
                -----    -----    -----    -----    -----
Buffer Block 1      0K       0K       0K       0K

Total Shared Buffer Avaliable = 20528
Class Threshold Limit = 13872
----------------------------------------------------------
Ethernet1/9     0K       0K       0K       0K       0K
<snip>
Ethernet1/4  2304K    3072K    3072K             3072K
```

```
Nexus3548#show hardware profile buffer monitor  interface ethernet 1/4 detail
Detail CLI issued at: 09/10/2012 22:15:42
Legend -
384KB  - between   1 and 384KB of shared buffer consumed by port
768KB  - between 385 and 768KB of shared buffer consumed by port
307us  - estimated max time to drain the buffer at 10Gbps
Active Buffer Monitoring for port Ethernet1/4 is: Active
KBytes              384  768 1152 1536 1920 2304 2688 3072 3456 3840 4224 4608
4992 5376 5760 6144
us @ 10Gbps         307  614  921 1228 1535 1842 2149 2456 2763 3070 3377 3684
3991 4298 4605 4912
                    ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ---- ----
09/10/2012 22:15:38    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
09/10/2012 22:15:37   34    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
09/10/2012 22:15:36  139  111    0    0    0    0    0    0    0    0    0    0    0    0    0    0
09/10/2012 22:15:35    0   67  179    4    0    0    0    0    0    0    0    0    0    0    0    0
09/10/2012 22:15:34    0    0    0  174   76    0    0    0    0    0    0    0    0    0    0    0
09/10/2012 22:15:33    0    0    0    0  102  148    0    0    0    0    0    0    0    0    0    0
09/10/2012 22:15:32    0    0    0    0    0   30  178   43    0    0    0    0    0    0    0    0
09/10/2012 22:15:31    0    0    1    0    0    1    0  208    0    0    0    0    0    0    0    0
09/10/2012 22:15:30    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0    0
```

- HW samples default buffer occupancy every 4msec, and update the bin counters. [down to 10ns]
- SW polls buffer Histogram counters every second

N3548 - Active Buffer Monitoring

Configuration

Enable/Disable Buffer Monitoring:

```
[no] hardware profile buffer monitor  [unicast|multicast]
```

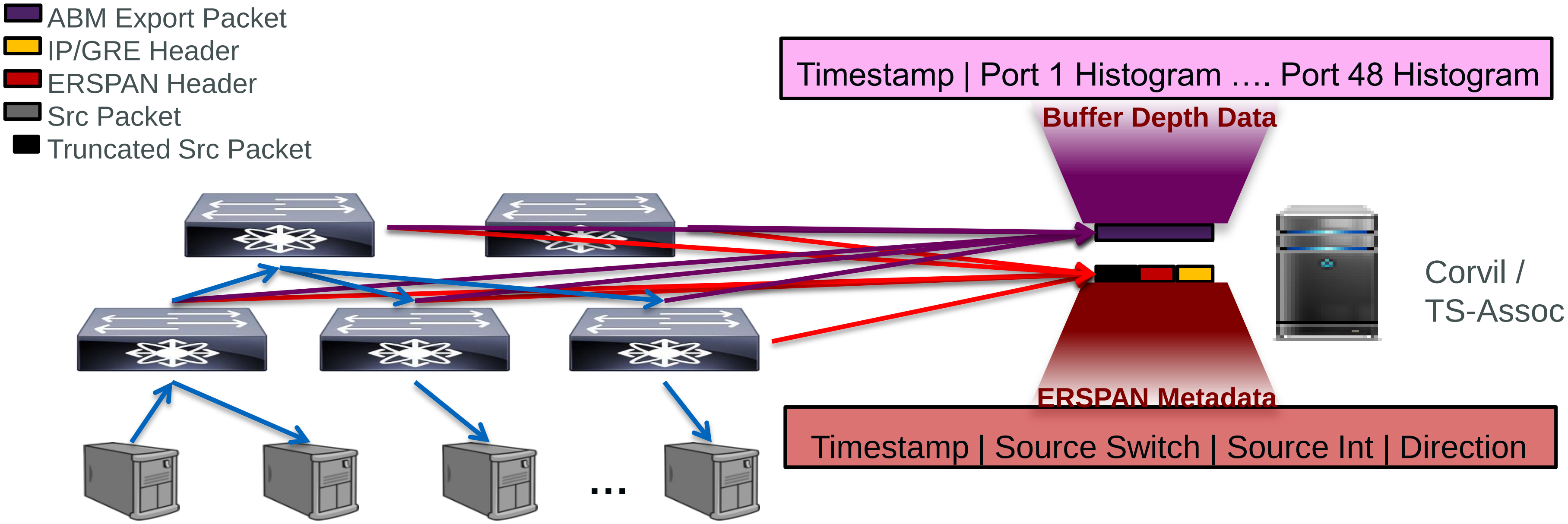
Show/Clear Commands:

```
show hardware profile buffer monitor [summary | brief | detail |
Interface | output-block | multicast]

clear hardware profile buffer monitor
```

Setting Buffer Usage Threshold for Notification & Hardware sampling Interval:

```
hardware profile buffer  monitor  sampling <sampling  value>
   // Default 4millisec

hardware profile buffer  monitor threshold  <384-6144KB>
    //  Have  “logging level mtc-usd 5” set to get the syslog message
```

Combine PTP + ERSPAN for Live Latency monitoring



- 🟪 ABM Export Packet
- 🟨 IP/GRE Header
- 🟥 ERSPAN Header
- ▨ Src Packet
- ■ Truncated Src Packet

Timestamp | Port 1 Histogram …. Port 48 Histogram

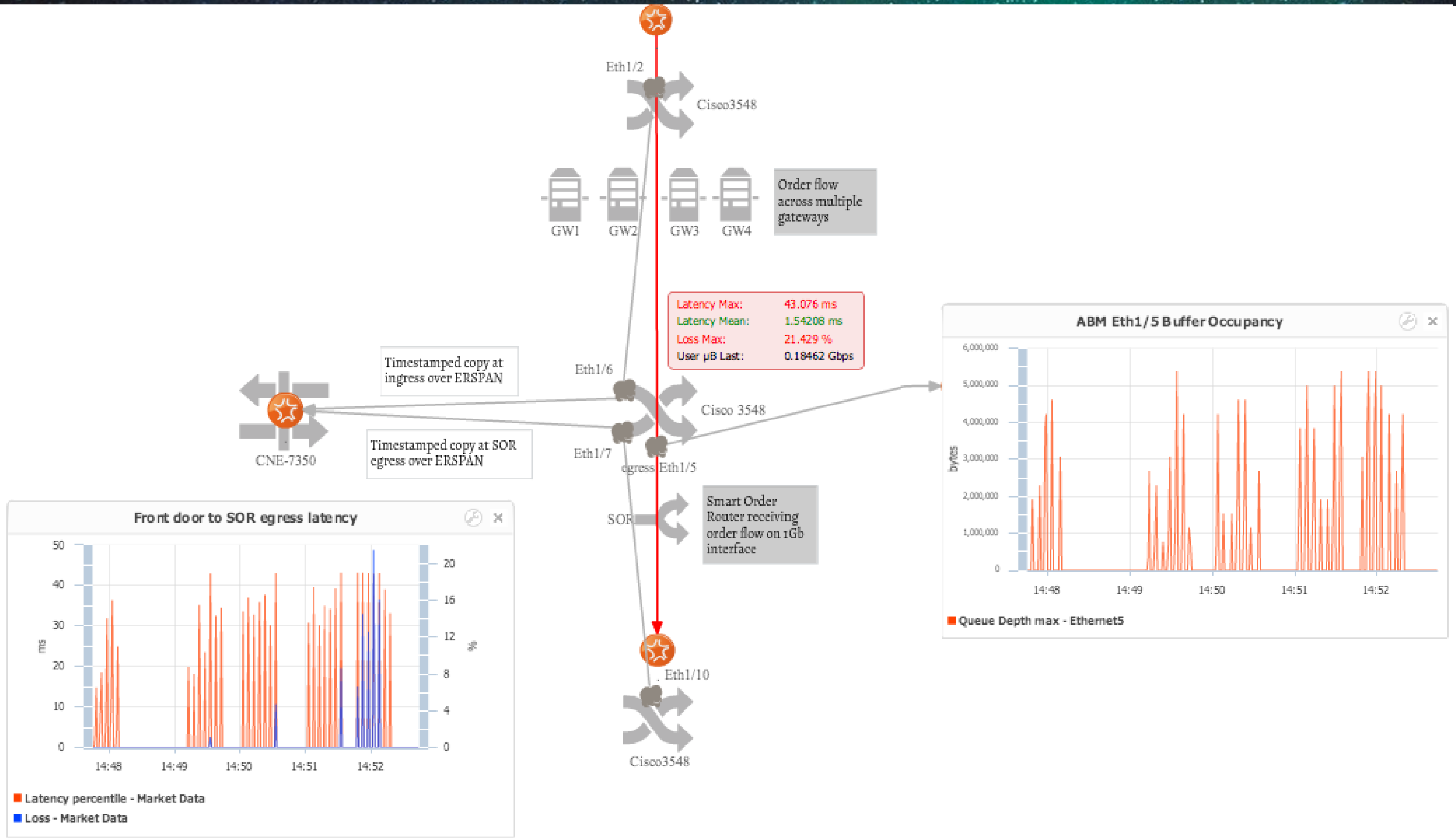
Buffer Depth Data

ERSPAN Metadata

Timestamp | Source Switch | Source Int | Direction

Messaging Bus

Nexus 3548 Analytics – Example with Corvil



Order flow across multiple gateways

Latency Max:	43.076 ms
Latency Mean:	1.54208 ms
Loss Max:	21.429 %
User µB Last:	0.18462 Gbps

Timestamped copy at ingress over ERSPAN

Timestamped copy at SOR egress over ERSPAN

Smart Order Router receiving order flow on 1Gb interface

Agenda – Nexus 3548 – BRKARC-2013

- Benchmarking
- Architecture
  - Product Overview
  - Architecture
  - Features
  - Analytics
  - **Scripting**
- Designs

Which port is connected?

```
n3548-001# show interface brief

--------------------------------------------------------------------------------
Ethernet      VLAN   Type Mode   Status  Reason                   Speed     Port
Interface                                                                   Ch #
--------------------------------------------------------------------------------
Eth1/1        1      eth  access up      none                        10G(D) --
Eth1/2        1      eth  access up      none                        10G(D) --
Eth1/3        1      eth  access up      none                        10G(D) --
Eth1/4        1      eth  access up      none                        10G(D) --
Eth1/5        1      eth  access up      none                        10G(D) --
.
.
Eth1/33       1      eth  access up      none                        10G(D) --
Eth1/34       1      eth  access up      none                        10G(D) --
Eth1/35       1      eth  access down    SFP not inserted            10G(D) --
Eth1/36       1      eth  access down    SFP not inserted            10G(D) --
Eth1/37       1      eth  access down    Administratively down       10G(D) -
.
```

What is connected there? Classic Network View

```
n3548-001# show mac address-table dynamic
Legend:
        * - primary entry, G - G                    uted MAC, O - Overlay
MAC
        age - seconds since firs                    ntry using vPC Peer-
Link
   VLAN     MAC Address       Typ                   e NTFY    Ports
---------+-----------------+----                    --+----+---------------
---
* 1          e8b7.484d.a208    dynamic   60570        F    F   Eth1/31
* 1          e8b7.484d.a20a    dynamic   60560        F    F   Eth1/31
* 1          e8b7.484d.a73e    dynamic   60560        F    F   Eth1/34
* 1          e8b7.484d.a740    dynamic   60560        F    F   Eth1/34
* 1          e8b7.484d.ad15    dynamic   60560        F    F   Eth1/28
* 1          e8b7.484d.ad17    dynamic   60560        F    F   Eth1/28
* 1          e8b7.484d.b3e9    dynamic   60570        F    F   Eth1/25
* 1          e8b7.484d.b3eb    dynamic   60560        F    F   Eth1/25
.
.
```

MAC Addresses of the connected devices … and the port they are on…

But, what is really connected and what is running?

```
n3548-001# portServerMap
=====================================
Port    Server FQDN
-------------------------------------
Eth1/1 c200-m2-10g2-001.cluster10g.com
…
Eth1/38 c200-m2-10g2-011.cluster10g.com
```

```
n3548-001# trackerList
=======================================
Port     Server                Server Port
---------------------------------------
Eth1/2   c200-m2-10g2-002         50544
Eth1/3   c200-m2-10g2-003         41909
Eth1/4   c200-m2-10g2-004         36480
Eth1/5   c200-m2-10g2-005         38179
Eth1/6   c200-m2-10g2-006         51375
Eth1/7   c200-m2-10g2-031         41915
Eth1/8   c200-m2-10g2-008         50983
Eth1/9   c200-m2-10g2-009         37056
Eth1/11 c200-m2-10g2-011         35882
Eth1/12 c200-m2-10g2-012         44551
.
```

We build scripts and share them on github! Github/datacenter

Which node is using the buffer?

```
n3548-001# bufferServerMap
=================================================================
Port     Server                         1sec     5sec     60sec    5min     1hr
-----------------------------------------------------------------
Eth1/1   c200-m2-10g2-001               0KB      0KB      0KB      0KB      0KB
Eth1/2   c200-m2-10g2-002               384KB    384KB    1536KB   2304KB   2304KB
Eth1/3   c200-m2-10g2-003               384KB    384KB    1152KB   1536KB   1536KB
Eth1/4   c200-m2-10g2-004               384KB    384KB    2304KB   2304KB   2304KB
Eth1/5   c200-m2-10g2-005               384KB    384KB    768KB    1536KB   1536KB
Eth1/6   c200-m2-10g2-006               384KB    2304KB   2304KB   2304KB   2304KB
Eth1/7   c200-m2-10g2-031
Eth1/8   c200-m2-10g2-008
Eth1/9   c200-m2-10g2-009
Eth1/11  c200-m2-10g2-011
.
.
```

Eth1/1(c200-m2-10g2-001) has 0 buffer usage because it’s the name node

See instantaneously the buffer use per node name

What’s running on this cluster and buffer correlation!

```
n3548-001# jobsBuffer
Hadoop Job Info ...
===================================================================
1 jobs currently running
JobId           RunTime(secs)   User    Priority
job_201306131423_0009   120     hadoop  NORMAL
===================================================================
Buffer Info - Per Port
Port    Server                  1sec    5sec    60sec   5min    1hr
-------------------------------------------------------------------
Eth1/1  c200-m2-10g2-001        0KB     0KB     0KB     0KB     0KB
Eth1/2  c200-m2-10g2-002        384KB   384KB   768KB   768KB   768KB
Eth1/3  c200-m2-10g2-003        384KB   384KB   1152KB  1152KB  1152KB
Eth1/4  c200-m2-10g2-004        384KB   1536KB  1536KB  1536KB  1536KB
Eth1/5  c200-m2-10g2-005        384KB   768KB   1152KB  1152KB  1152KB
```

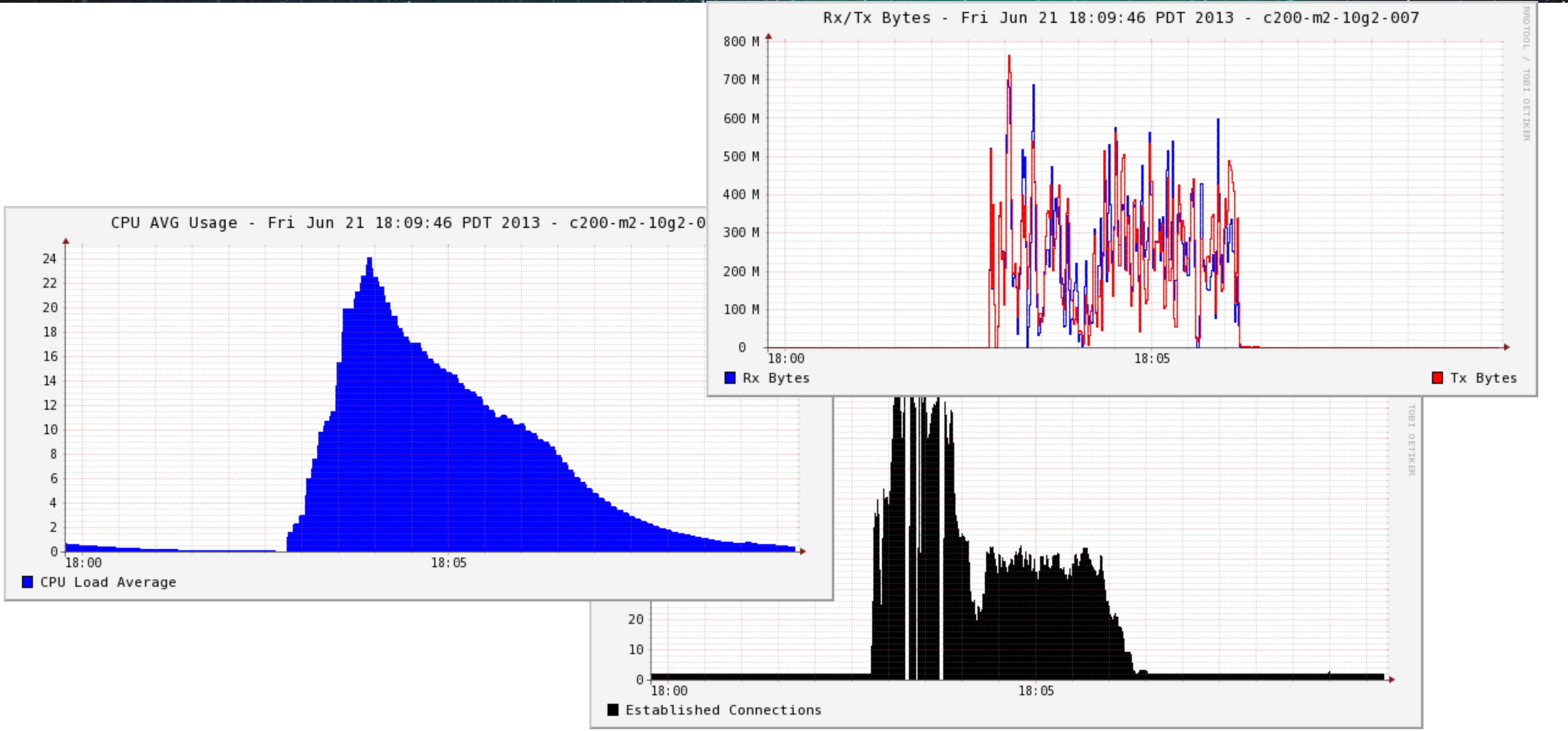
What jobs were running during peak buffer usage … and for how long were they running

What’s running on this cluster + Buffer usage per server

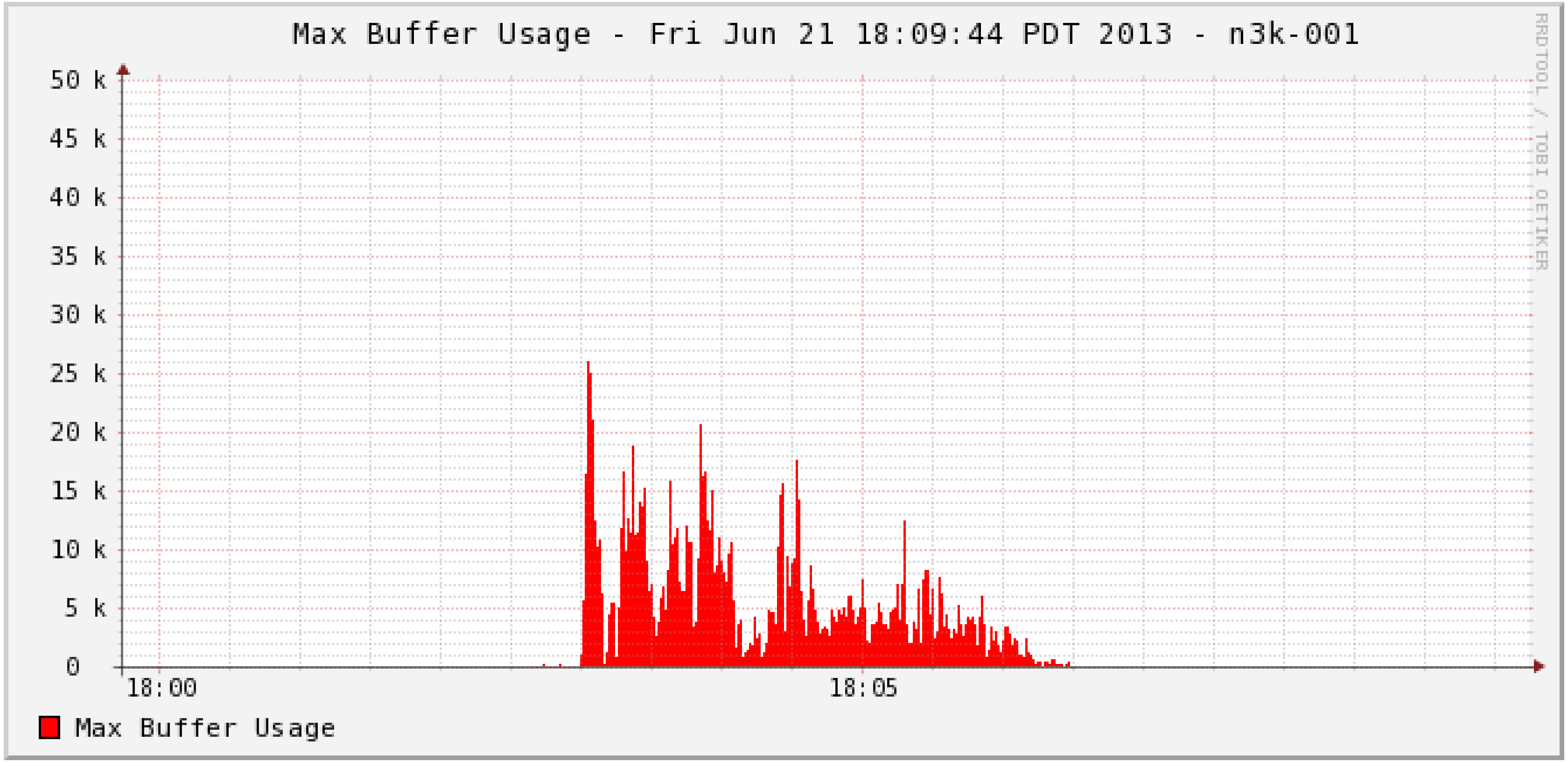
```
n3548-001(config)# jobsBuffer
Hadoop Job Info ...
===================================================================
0 jobs currently running
JobId           RunTime(secs)   User    Priority
===================================================================
Buffer Info - Per Port
Port    Server                  1sec    5sec    60sec   5min    1hr
-------------------------------------------------------------------
Eth1/1  c200-m2-10g2-001        0KB     0KB     0KB     0KB     0KB
Eth1/2  c200-m2-10g2-002        0KB     0KB     0KB     1920KB  1920KB
Eth1/3  c200-m2-10g2-003        0KB     0KB     0KB     2304KB  2304KB
Eth1/4  c200-m2-10g2-004        0KB     0KB     0KB     2688KB  2688KB
Eth1/5  c200-m2-10g2-005        0KB     0KB     0KB     2304KB  2304KB
Eth1/6  c200-m2-10g2-006        0KB     0KB     0KB     2304KB  2304KB
Eth1/7  c200-m2-10g2-031        0KB     0KB     0KB     1920KB  2688KB
.
```

Historical data is also captured

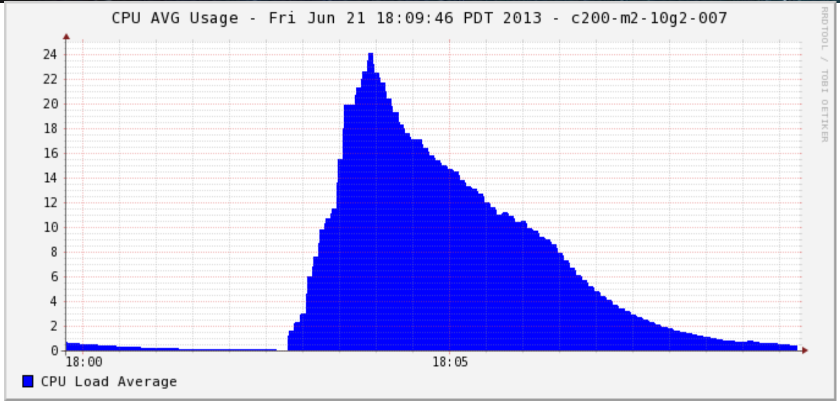
Server Resource Monitoring – CPU, Connections, etc.

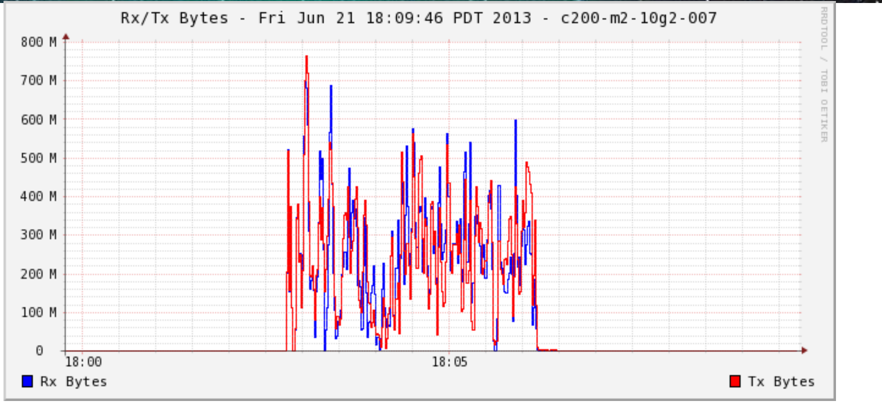


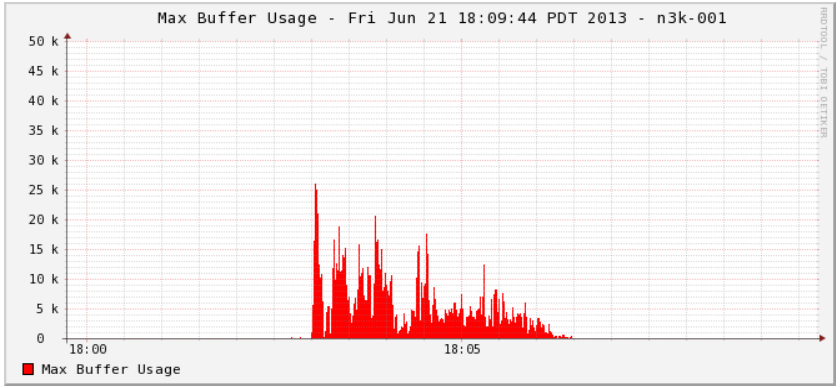
Network Resource Monitoring – Buffer Counters etc.

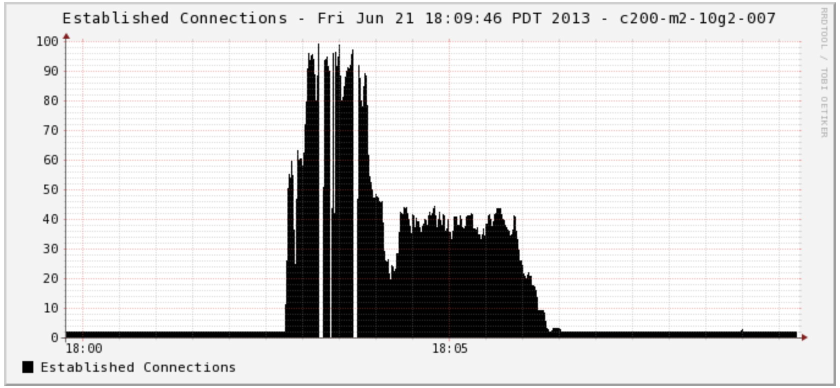


All in one view correlate app, server and network!









Where to find more information about scripts?

Software download on product page
cisco.com/go/nexus3548

GitHub Bootcamp If you are still new to things, we've provided a few walkthroughs to get you started.

1 **Set Up Git**
A quick guide to help you get started with Git.

2 **Create A Repository**
Create the place where your commits will be stored.

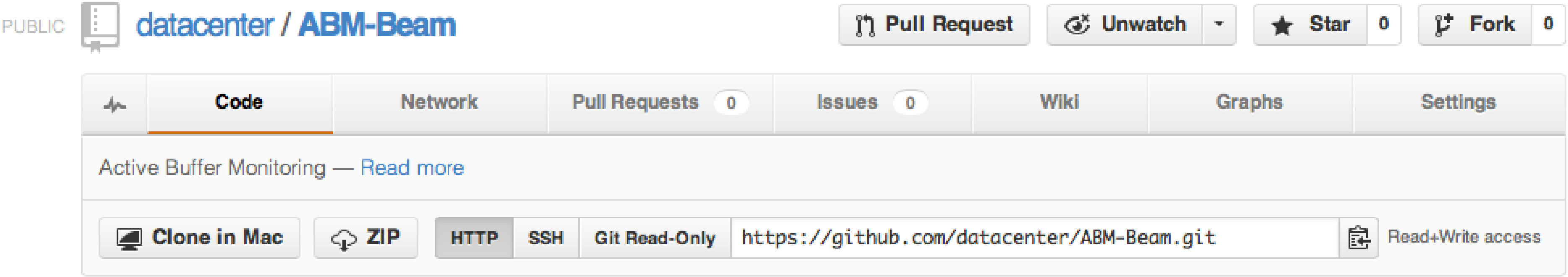
3 **Fork a Repository**
Copy a repo to create a new, unique project from its contents.

4 **Be social**
Follow a friend.
Watch a project.

Github.com/datacenter

Example: the ABM script

Github.com/datacenter



PUBLIC datacenter / ABM-Beam

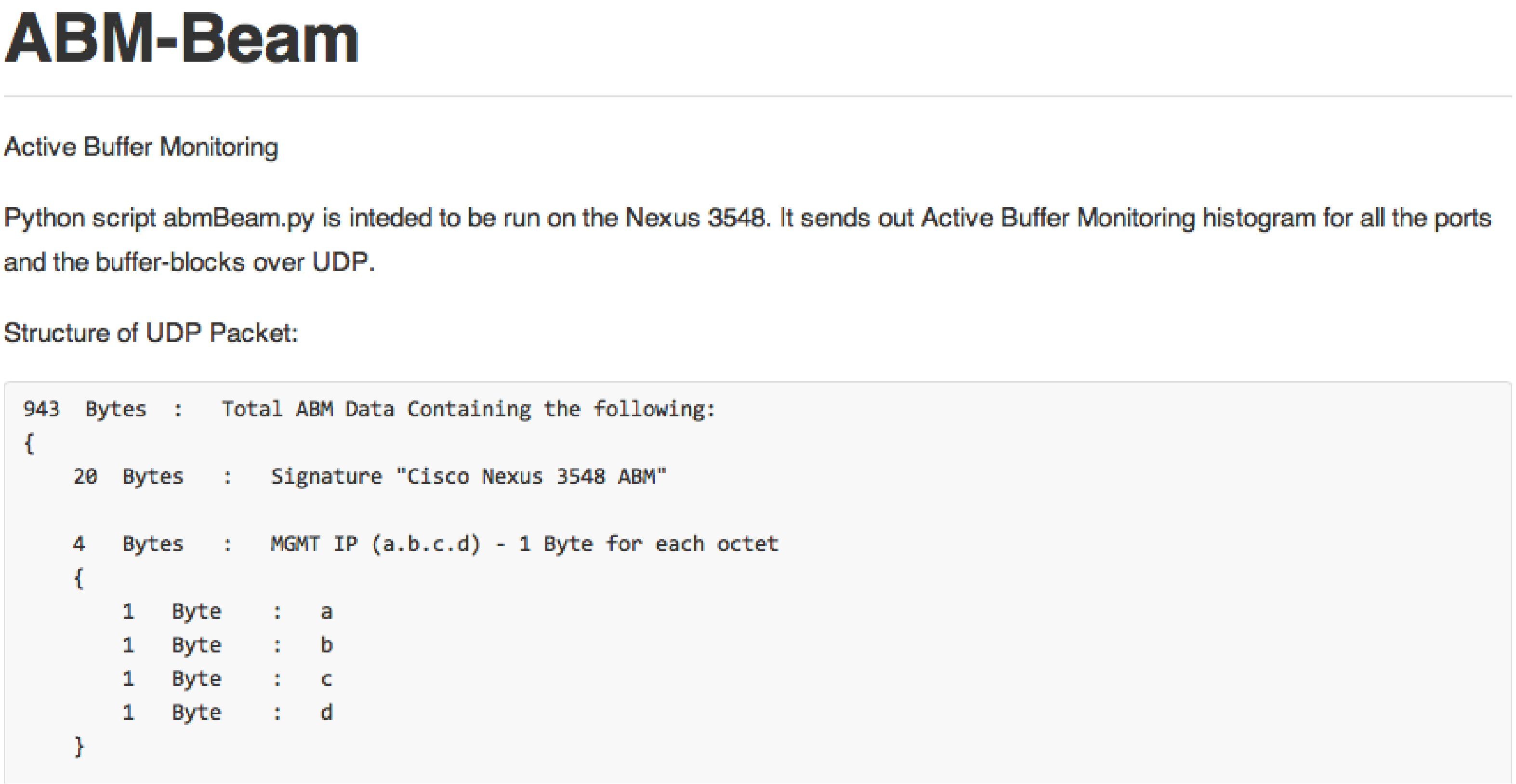
Active Buffer Monitoring — Read more

https://github.com/datacenter/ABM-Beam.git

Read+Write access

A script for the 3548 to stream active buffer monitoring

Example: the ABM script



ABM-Beam

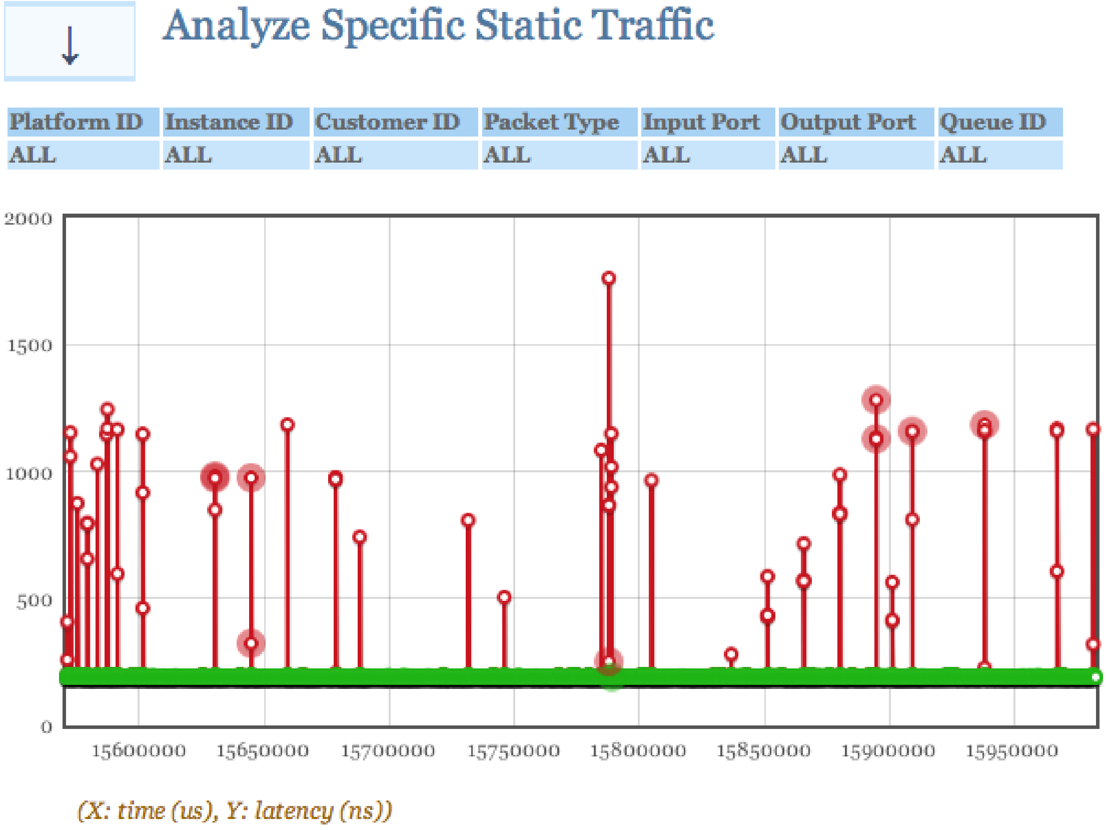
Active Buffer Monitoring

Python script abmBeam.py is inteded to be run on the Nexus 3548. It sends out Active Buffer Monitoring histogram for all the ports and the buffer-blocks over UDP.

Structure of UDP Packet:

```
943  Bytes  :   Total ABM Data Containing the following:
{
    20  Bytes   :   Signature "Cisco Nexus 3548 ABM"

    4   Bytes   :   MGMT IP (a.b.c.d) - 1 Byte for each octet
    {
        1   Byte    :   a
        1   Byte    :   b
        1   Byte    :   c
        1   Byte    :   d
    }
```

Nexus 3548 Analytics – Build your tools!



↓

Analyze Specific Static Traffic

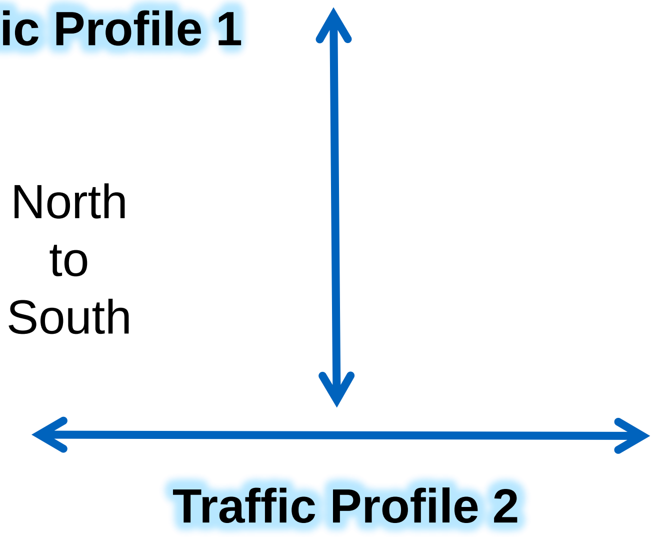
Platform ID	Instance ID	Customer ID	Packet Type	Input Port	Output Port	Queue ID
ALL	ALL	ALL	ALL	ALL	ALL	ALL

(X: time (us), Y: latency (ns))

Agenda – Nexus 3548 – BRKARC-2013

- Benchmarking
- Architecture
- Designs
  - HFT
  - HPC

Case Study – Trade Flow Example

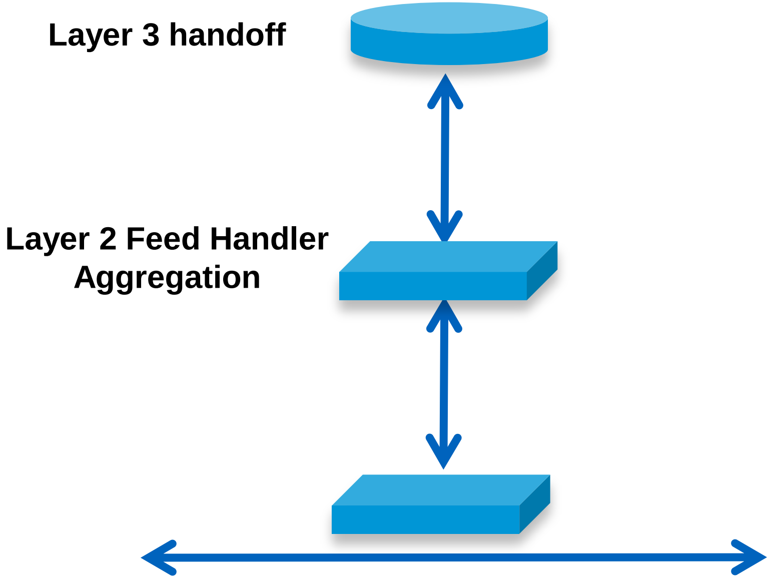


Traffic Profile 1

North
to
South

Traffic Profile 2

West to East



Layer 3 handoff

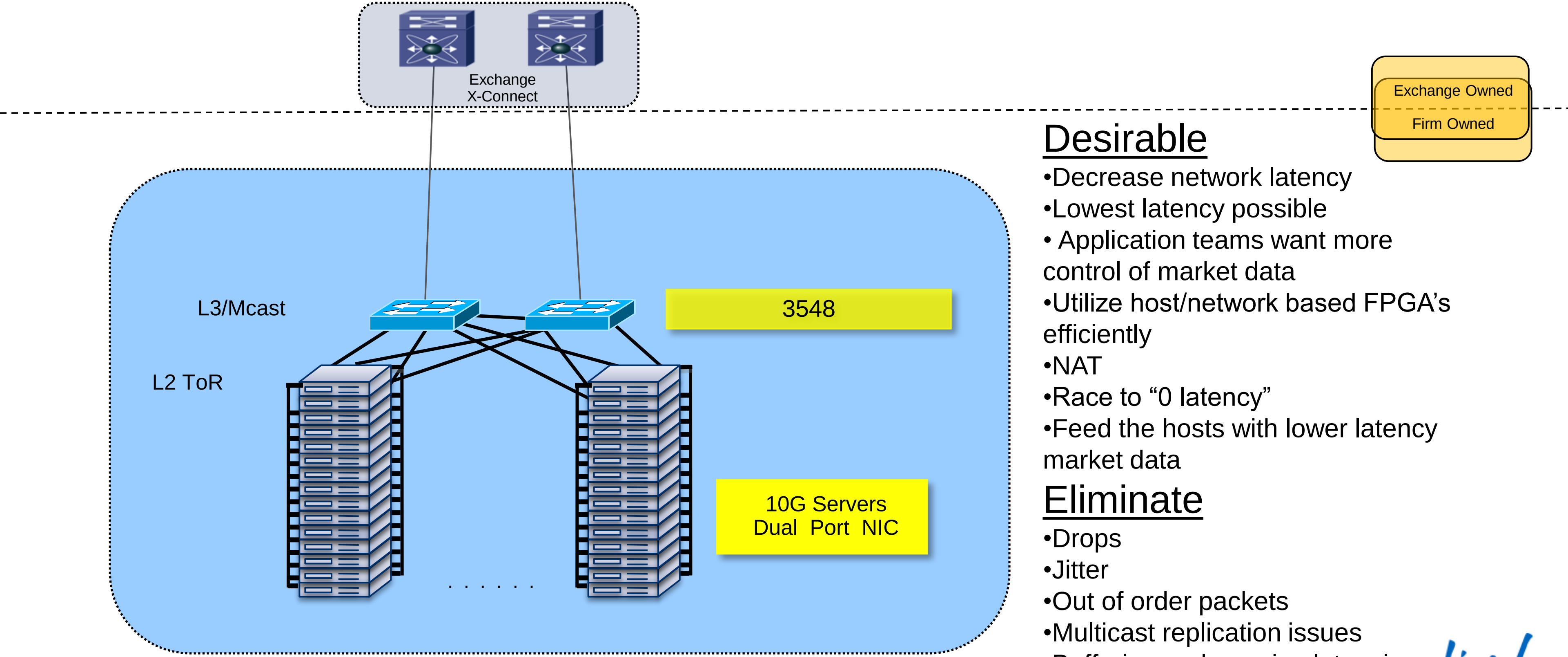
Layer 2 Feed Handler
Aggregation

Layer 2 message bus

What is more important?

Closeness to market data access? Or performance between servers?

Financial Colocation – 12-20 Servers per exchange



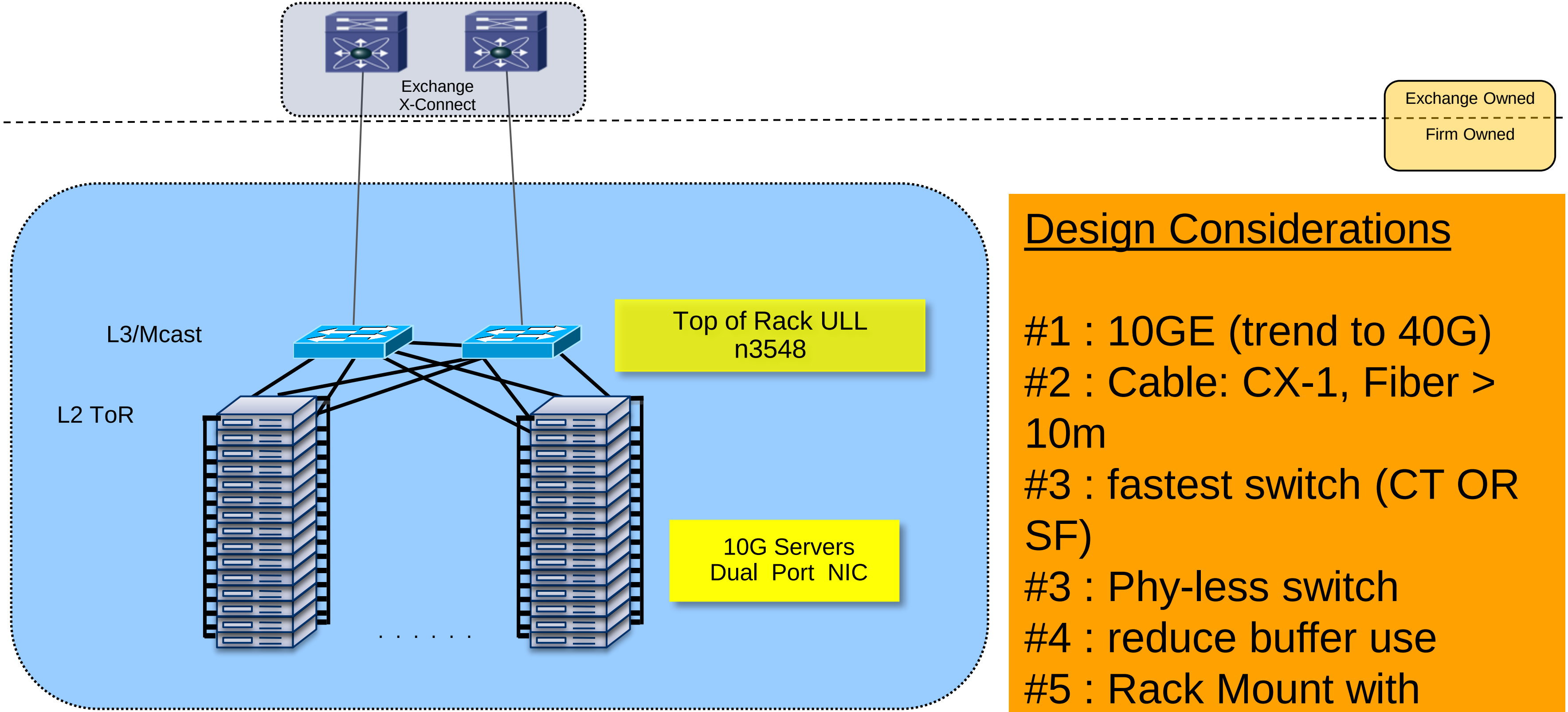
Desirable

- •Decrease network latency
- •Lowest latency possible
- • Application teams want more control of market data
- •Utilize host/network based FPGA’s efficiently
- •NAT
- •Race to “0 latency”
- •Feed the hosts with lower latency market data

Eliminate

- •Drops
- •Jitter
- •Out of order packets
- •Multicast replication issues
- •Buffering and queuing latencies

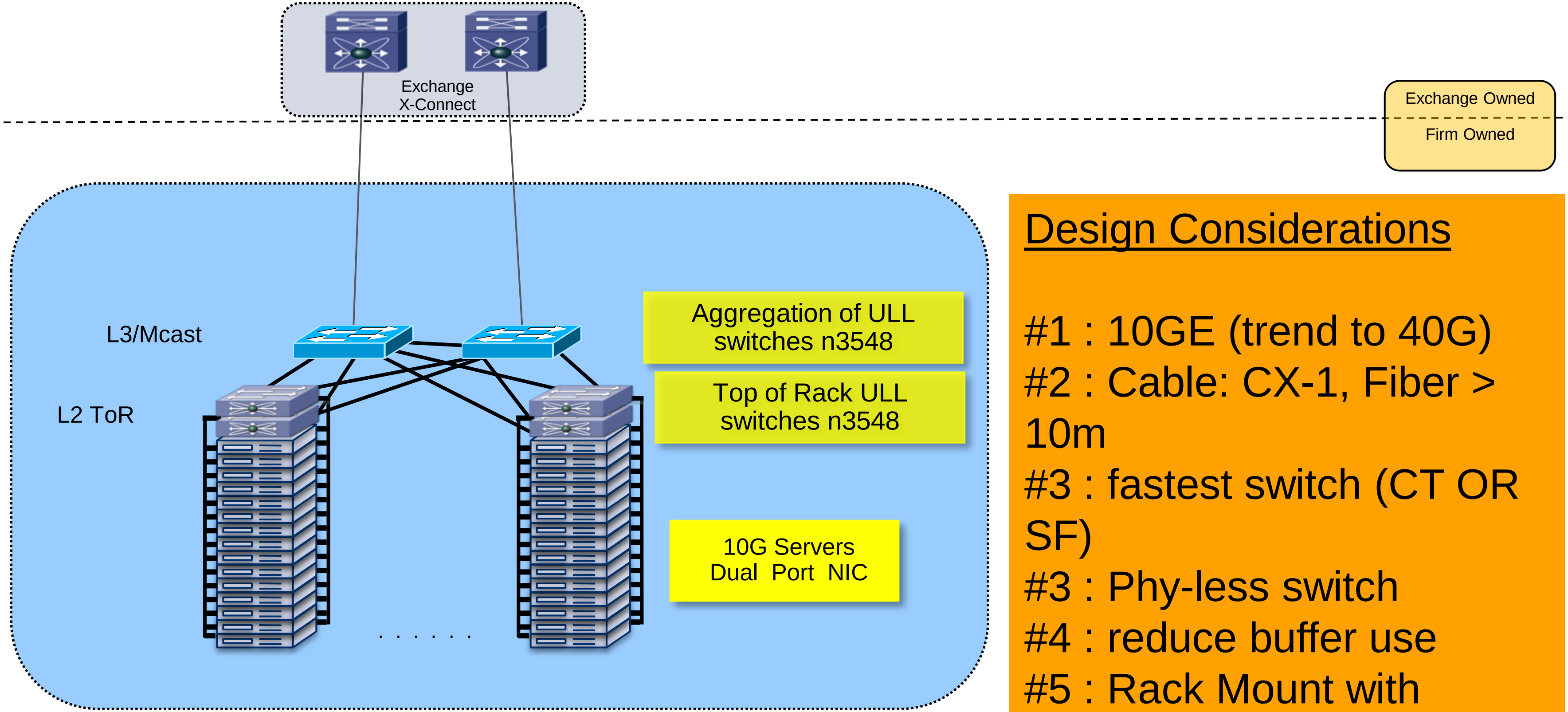
Financial Colocation – 12-20 Servers per exchange



Design Considerations

#1 : 10GE (trend to 40G)
#2 : Cable: CX-1, Fiber > 10m
#3 : fastest switch (CT OR SF)
#3 : Phy-less switch
#4 : reduce buffer use
#5 : Rack Mount with accelerated NIC

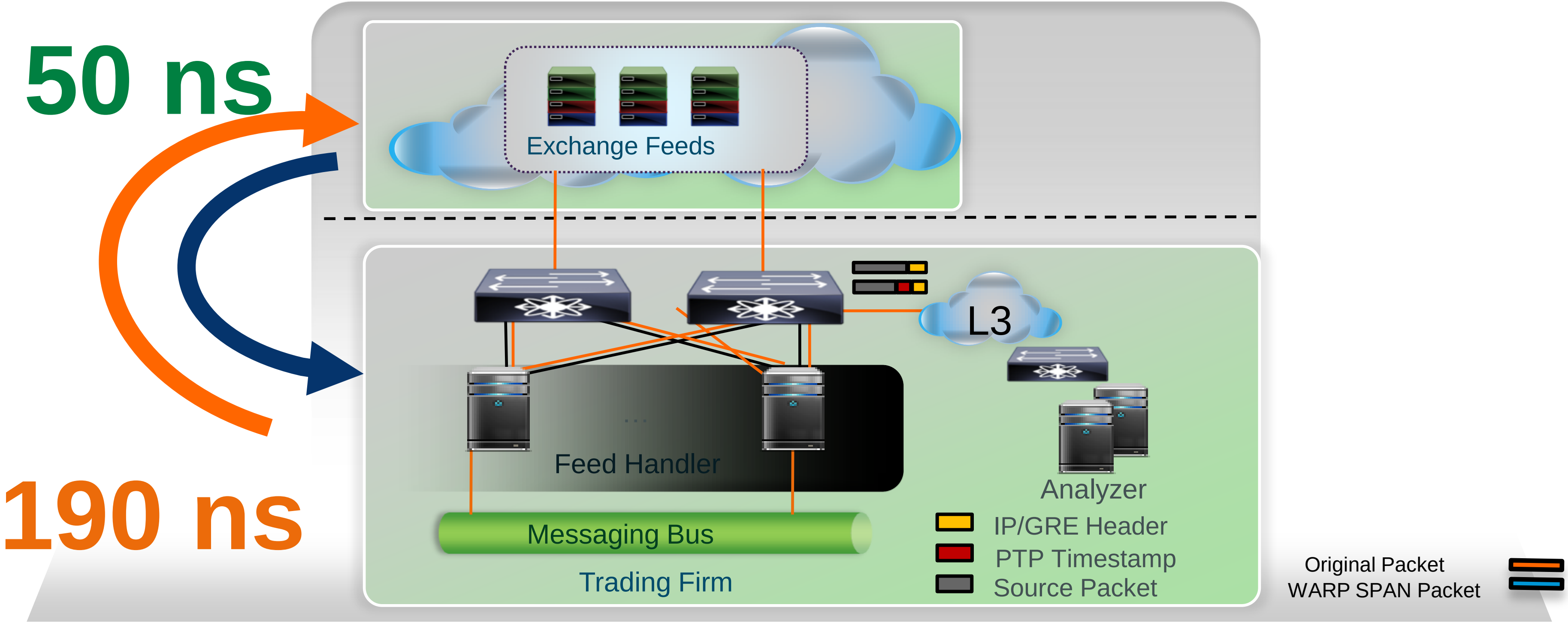
Financial Colocation – 20-48 Servers per exchange



Design Considerations

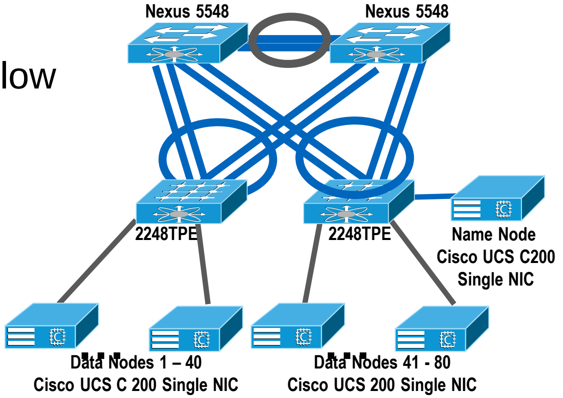
#1 : 10GE (trend to 40G)
#2 : Cable: CX-1, Fiber > 10m
#3 : fastest switch (CT OR SF)
#3 : Phy-less switch
#4 : reduce buffer use
#5 : Rack Mount with accelerated NIC

Financial Colocation – Innovate Design at 50ns!



High Performance Compute-HPC

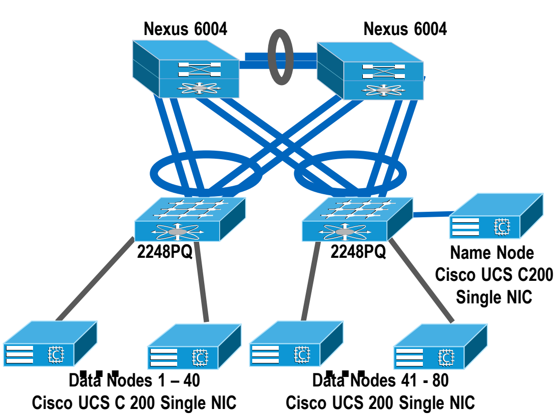
- Hadoop Network Topology – Unified Fabric / TOR DC
- Integration with Enterprise architecture – essential pathway for data flow
  - Architecture
  - Consistency
  - Management
  - Risk-assurance
  - Enterprise grade features
- Consistent Operational Model
  - NxOS, CLI, Fault Behavior and Management
- Though higher BW east-west compared to traditional transactional networks
- Over the time it will have multi-user, multi-workload behavior
  - Need enterprise centric features
  - Security, SLA, QoS etc.
- Big Data is just another application



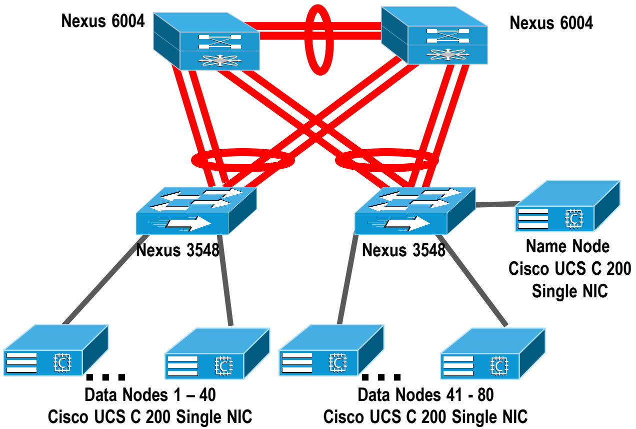
Traditional DC Design Nexus 55xx/2248

High Performance Compute-HPC

- Higher Density and Faster with 6004 and 3548!



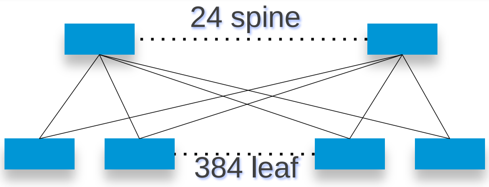
Ultra Low Latency DC Design Nexus 6004/2248



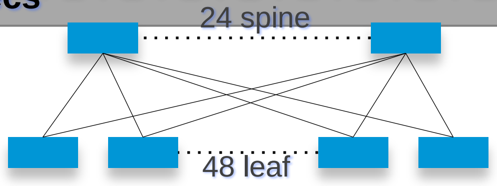
High Scale, lower latencyNexus 6K/3K TOR based Topology

Evolution of High Performance at High Density

10G Fabric



10G ports: 9,216   Target Latency: ~1.4 usecs



10G port s: 1,152   Target Latency: ~600 nsecs

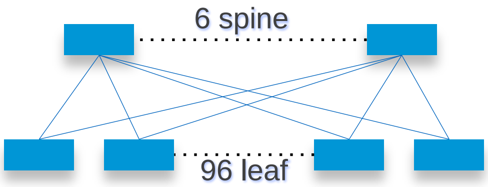
N6004

N3548

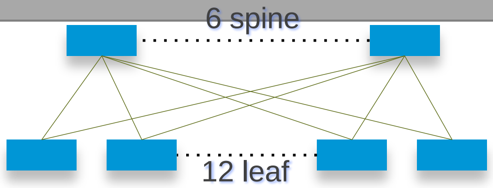
3548

3548

40G Fabric



40G port count: 576   Target Latency: ~1.5 usecs



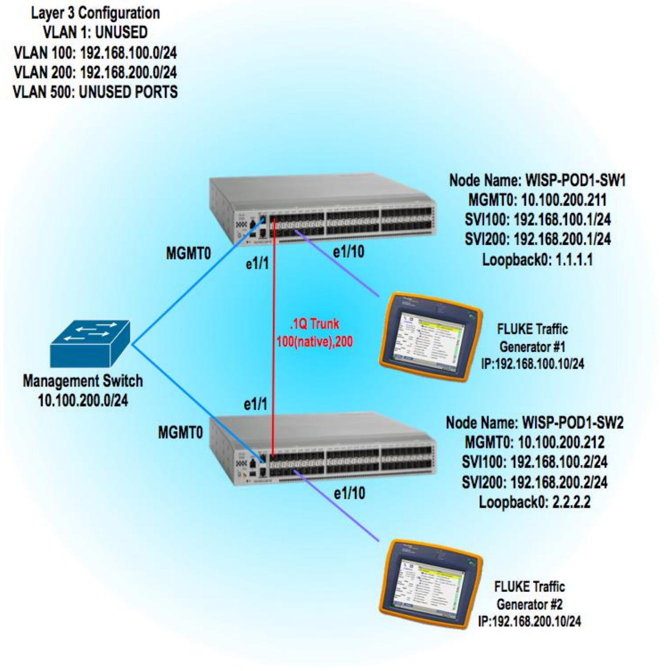
40G port count: 72   Target Latency: ~750 nsecs

Summary

- Measure Latency with the appropriate tools and compare switches with FILO
- Predictable latency at scale **and** lowest latency is **now** possible
- Understanding the simple architecture of the Nexus 3548 is **key** in designing solutions for application driven datacenter
- Nexus 3548 provides an application view and correlation to the network real time with open source scripts, allowing to design solutions for application driven datacenters, high performance data centers, and compact top of rack with L2/L3 and services

Practice!

- WISP LAB at BOOTH 158 for Nexus 3548!



Layer 3 Configuration
VLAN 1: UNUSED
VLAN 100: 192.168.100.0/24
VLAN 200: 192.168.200.0/24
VLAN 500: UNUSED PORTS

Node Name: WISP-POD1-SW1
MGMT0: 10.100.200.211
SVI100: 192.168.100.1/24
SVI200: 192.168.200.1/24
Loopback0: 1.1.1.1

MGMT0
e1/1
e1/10

.1Q Trunk
100(native),200

FLUKE Traffic
Generator #1
IP:192.168.100.10/24

Management Switch
10.100.200.0/24

e1/1
MGMT0

Node Name: WISP-POD1-SW2
MGMT0: 10.100.200.212
SVI100: 192.168.100.2/24
SVI200: 192.168.200.2/24
Loopback0: 2.2.2.2

e1/10

FLUKE Traffic
Generator #2
IP:192.168.200.10/24

- Get familiar with NX-OS
- Practice the ABM feature
- Practice WARP Mode

Complete Your Online Session Evaluation

- Give us your feedback and you could win fabulous prizes. Winners announced daily.
- Receive 20 Passport points for each session evaluation you complete.
- Complete your session evaluation online now (open a browser through our wireless network to access our portal) or visit one of the Internet stations throughout the Convention Center.

Don’t forget to activate your Cisco Live Virtual account for access to all session material, communities, and on-demand and live activities throughout the year. Activate your account at the Cisco booth in the World of Solutions or visit www.ciscolive.com.

cisco TM