Driven by
Experience™

Load Balancing in Data Center

Author: Vivek V

Documentation Feedback

We encourage you to provide feedback so that we can improve our documentation.
Send your comments to design-center-comments@juniper.net. Include the document or topic name, URL or page number, and software version (if applicable).

Table of Contents

Introduction

This document provides a comprehensive deep dive into the various load-balancing mechanisms and their evolution to suit the needs of the Juniper Networks data center. The document elucidates the legacy equal cost multipath (ECMP)-based load balancing and its journey to more recent variants such as Dynamic Load Balancing (DLB) and the latest evolution to Selective Load Balancing (SLB) and Reactive Path Load Balancing (RLB) along with various configuration and validation examples.

Static Load Balancing

An ECMP group is formed when the routing table has multiple paths to the same destination with the same cost. Junos OS can be configured as such, where multiple next hops will exist in the forwarding table.

Figure 1: 3-Stage Clos Fabric with Two Paths from the Source Leaf to the Destination

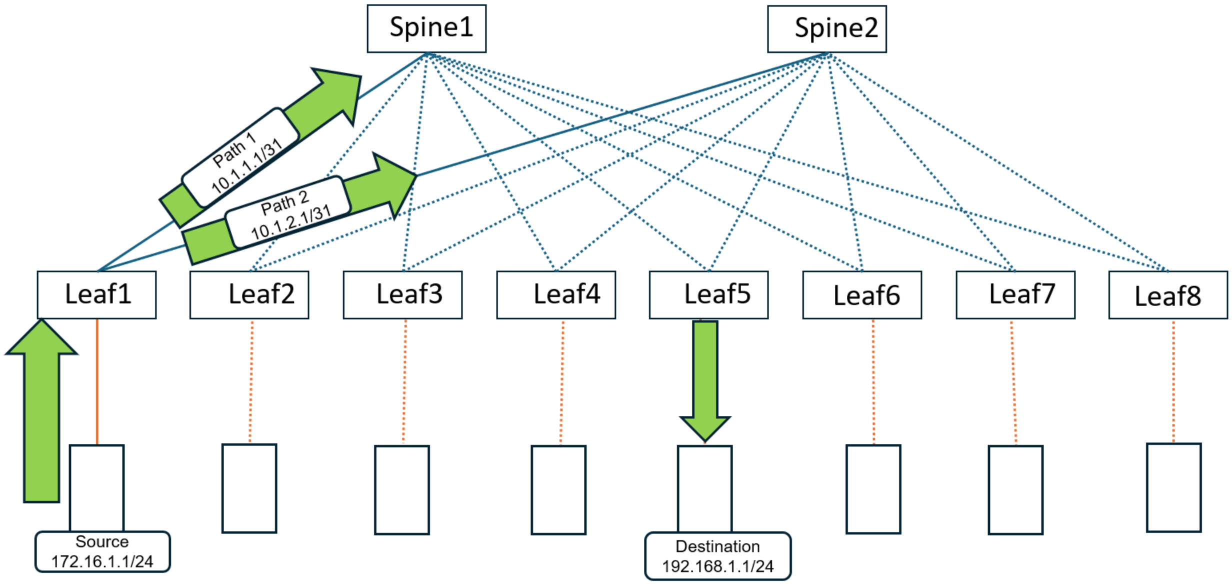


Figure 1 shows a typical 3-stage any-to-any Clos fabric wherein each leaf will have two or more equal cost paths to any given destination based on the number of spines and the number of links to the spines. In this example, the premise is that there are two equal-cost paths through Leaf1 for traffic from source 172.16.1.1 to destination 192.168.1.1 advertised through Spine1 and Spine2 respectively.

Output1: Multiple Equal Cost Paths to the Same Destination

```
leaf1> show route

inet.0: 1 destinations, 2 routes (1 active, 0 holddown, 0 hidden)
Restart Complete
+ = Active Route, - = Last Active, * = Both
192.168.1.0/24        *[BGP/170] 5d 15:25:18, localpref 100
                      AS path: 101 201 I, validation-state: unverified
                       to 10.1.1.0 via et-0/0/0.0
                    >  to 10.1.2.0 via et-0/0/2.0
                    [BGP/170] 5d 15:25:18, localpref 100
```

Once the multiple paths are installed in the forwarding table, the hashing algorithm must choose the appropriate path to be taken.

Output2: RTAG7 Hashing Input Parameters for IPv4 Traffic

```
leaf1:pfe> show forwarding-options enhanced-hash-key
Current RTAG7 Settings
-------------------------
   Hash-Mode                 :layer2-payload
   Hash-Seed                 :232516805

inet  RTAG7 settings:
----------------------
inet packet fields
  protocol                  :yes
  Destination IPv4 Addr     :yes
  Source IPv4 Addr          :yes
  destination L4 Port       :yes
  Source L4 Port            :yes
  Vlan id                   :no
--snip--
```

In the output above, one can notice the term “RTAG7.” RTAG (0-7) are static hash generation mechanisms that are used by Broadcom ASICs which are applicable to ECMP as well as LAGs. RTAG7 in this case uses a 5-tuple (fields) input mechanism as shown in the output above and generates a unique hash for every micro-flow. Once the hash is determined, a modulus function is performed as described below.

In default (static) hash-based load balancing flows are assigned to members using mathematical mod (%) operation:

Member ID = Hash (key) mod (number of members in the group)

Any increase or decrease in the number of group members results in a complete remapping of flows to member IDs.

For example, hash key 10

10 mod 5 = 0: The member with id 0 is selected for flow.
10 mod 4 = 2: The member with id 2 is selected for the same flow when the number of members is decreased by 1.

This way, a static mapping is formed for every flow, and this remains the same as long as the number of paths (members) remains the same.

The RTAG7 algorithm provides the control that determines whether hashing is to be done purely based on the outermost Layer 2 (L2) header fields or the hashing includes fields after the outermost L2 header (Layer2-payload).

Output3: Hashing Modes Based on L2 Headers or Payloads

```
leaf1# set forwarding-options enhanced-hash-key hash-mode ?
Possible completions:
+ apply-groups Groups from which to inherit configuration data
+ apply-groups-except  Don't inherit configuration data from these groups
  layer2-header        Only layer2 header fields are considered for hashing
  layer2-payload       Only layer2 payload fields are considered for hashing
```

The drawback of this method of static link selection is that it doesn’t account for the load on the link or available underutilized paths in the network, which means that even if 1 of 5 available paths is at 99% load capacity and the others are at 1%; if the modulus mechanism chooses the first link for a new flow, it sends the traffic over that link and oversubscribe it.

To resolve such drawbacks and enable better utilization of the available paths, the next generation of load-balancing algorithms was introduced.

Dynamic Load Balancing

Dynamic Load Balancing (DLB) was conceptualized to resolve one of the main drawbacks of static hash based ECMP, which was to ensure that all paths are utilized equally in a network, avoid polarization(traffic aligning more to certain links), and introduce a mechanism that can change the path a flow takes based on dynamic network conditions.

This document focuses on the **flowlet** mode of operations, and other modes including per-packet load-balancing mode and assigned mode.

Output4: DLB Options

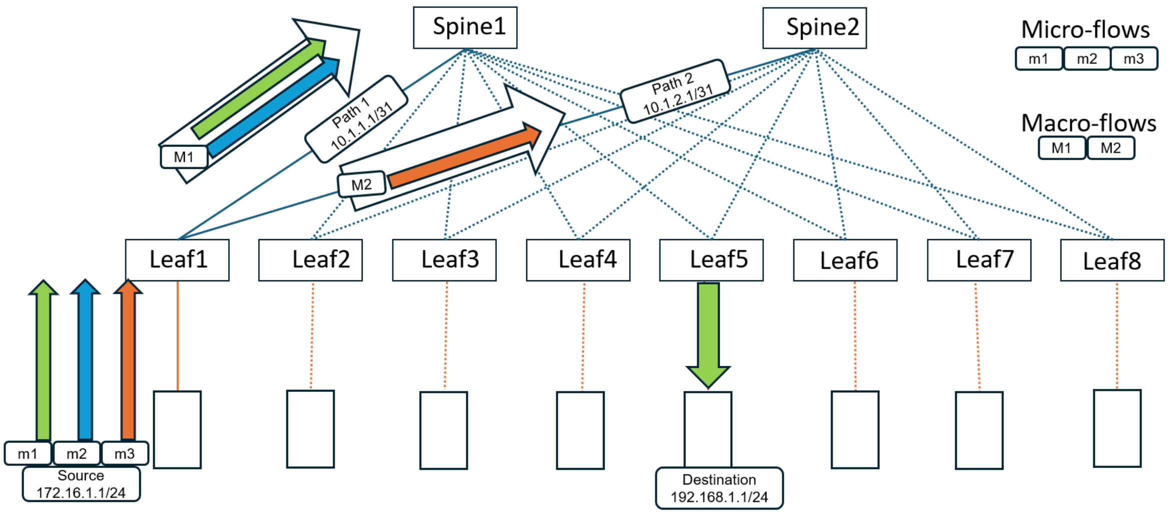
```
ECMP DLB Load Balancing Options:
------------------------------------------------
  Load Balancing Method                  : Flowlet
  Inactivity Interval                    : 16 (us)
  Flowset Table size                     : 256 (entries per ECMP)
  Reassignment Probability Threshold : 0
  Reassignment Quality Delta             : 0
  Egress Port Load Weight                : 50
  EgressBytes Min Threshold              : 10
  EgressBytes Max Threshold              : 50
```

The DLB mechanism as shown in Output4 achieves optimum distribution by reassigning flows based on whether the interframe gap (IFG) is greater than the inactivity timer for a micro-flow. The default inactivity interval is 16 microseconds as shown in the output above and can be customized to suit the network. The inactivity timer should be optimized to match the IFG (gap between two frames) so that the flow does not get reassigned due to incorrect timers.

To understand DLB better, one needs to first understand a few concepts that are listed below.

Micro-Flow and Macro-Flow

Figure 2: Diagram Representing Micro-Flows and Macro-Flows



In **Figure 2**, above, the source connected to Leaf1 is communicating with the destination that is connected to Leaf5. Based on the standard 5-tuple mechanism of RTAG7 (source IP, destination IP, source L4 port, destination L4 port, protocol), the individual flows are formed wherein all packets in the flow share the same headers and/or payloads. These flows are called **micro-flows** and are immutable.

Once the micro-flows are formed, they pick a flow index which is mapped to an interface. A collection of micro-flows that have picked the same flow index is called a **macro-flow**; each macro-flow is load-shared across the available paths by DLB.

NOTE: If there is high entropy between the flows and an optimal selection of a packet's hash input fields along with the hash algorithm, it is also possible that all micro-flows m1… m3 are treated individually as macro-flows M1…M3. This ensures the highest probability to give best possible traffic distribution.

The primary goal of DLB is to choose the best egress link for a given flowlet based on network quality. To do this, DLB employs two metrics:

- **Port Load Metric** – The amount of traffic in bytes transmitted per interval over each ECMP link.
- **Port Queue Metric** – The number of memory cells occupied while queuing at each ECMP.

The port quality of member links is categorized from 0-7. A port quality value of 0 indicates the lowest rating, for example a heavily loaded and/or degraded port; a port quality of 7 indicates the highest rating, such as an under-utilized and/or pristine port. The output below shows this for four ECMP paths, indicating that two of them are pristine.

Output5: DLB Port Quality Metric

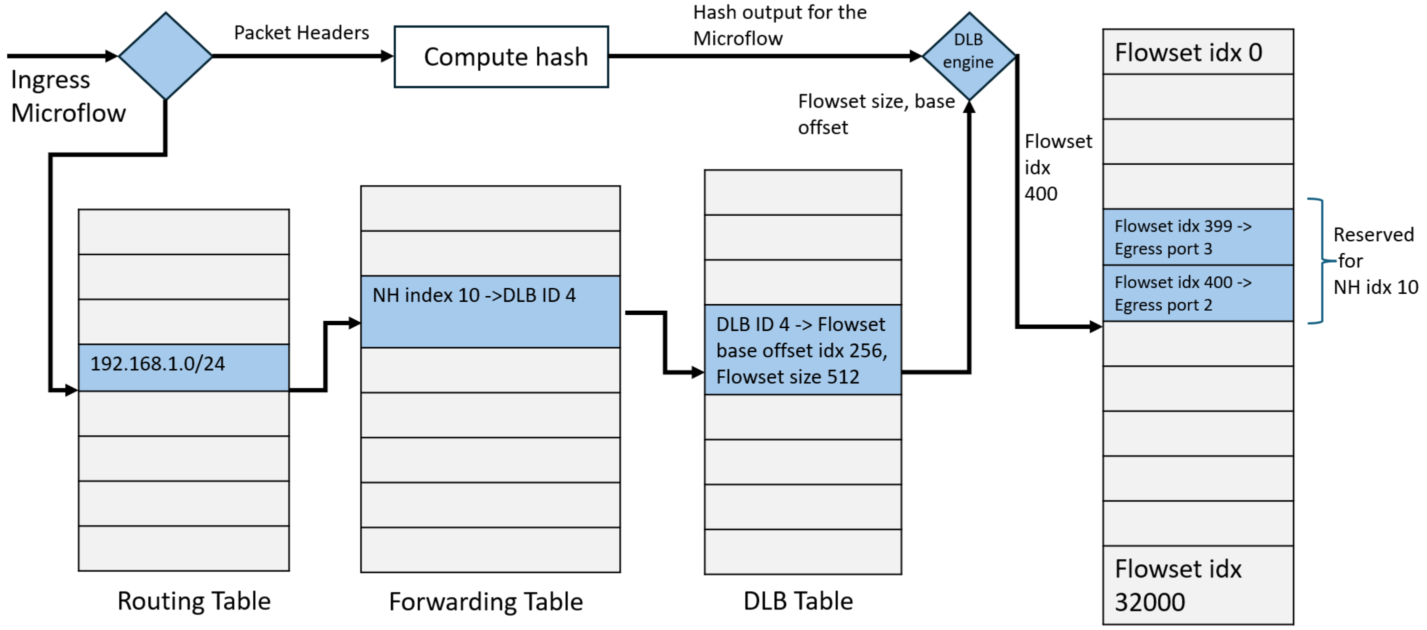
```
Leaf1:pfe> show evo-pfemand nh dlb-enabled index 51033
 DLB ECMP NhId : 51033

 ============ ==============  ======= ============
 Hw-Ecmp-NhId Hw-Unicast-NhId Hw-Port Port-Quality
 ============ ==============  ======= ============
   203972      100001           5       4
   203972      100002           9       7
   203972      100003           17       4
   203972      100004           25       7
```

The QFX forwarding ASIC is programmed to sample the port load on the egress ECMP members and update the quality. The default sampling rate is **62500 samples** per second.

Figure 3 shows the path selection mechanism in DLB. Micro-flow will go through two processes in parallel which is the routing lookup and the hash generation. Once the routing lookup is done and the forwarding table has mapped the next hop index to a DLB ID, this information is passed onto the DLB table which assigns a base offset ID and forwards this along with the flowset size to the DLB engine which further assigns a flowset ID, once this is done, it maps the flowset ID to an egress port in an ECMP member and sends the flow out. One or more macro-flows can be sent out of the same interface based on the flowset idx(index) to egress port mapping.

Figure 3: DLB Micro-flow – Macro-flow Diagram



The flowset table size allocated to the DLB ECMP groups can be increased to better load balance the flows on the ECMP member links when the number of micro-flows is less. By increasing the flowset size, it is possible to accommodate a higher number of macro-flows thereby ensuring better distribution of the flows.

Output6: DLB Flowset Table Size Customization

```
set forwarding-options enhanced-hash-key ecmp-dlb flowlet flowset-table-size ?
Possible completions:
  1024                 Flowset size assigned to DLB group is 1024
  16384                Flowset size assigned to DLB group is 16384
  2048                 Flowset size assigned to DLB group is 2048
  256                  Flowset size assigned to DLB group is 256 (default)
  32768                Flowset size assigned to DLB group is 32768
  4096                 Flowset size assigned to DLB group is 4096
  512                  Flowset size assigned to DLB group is 512
  8192                 Flowset size assigned to DLB group is 8192
```

Egress Quantization Thresholds

Let’s consider a scenario where there are two ECMP member links. Link1 is at 90% capacity and link2 is at 95% capacity. By default, the port quality metric algorithm will assign the same number 0 (lowest quality) to both links. If a new flow comes in that needs 7% utilization for example, since the quality is the same on both links, it will randomly assign a member link to the flow. In our example, link1 can handle this flow without loss but link2 cannot and can lead to a loss. To mitigate such a scenario, we can manually assign quality bands to provide more granularity to the DLB engine and optimize traffic distribution.

Output7: Egress Quantization Threshold Customization

```
root@leaf1# set forwarding-options enhanced-hash-key ecmp-dlb egress-quantization min ?
Possible completions:
  <min>                Min in percentage (1..100)

root@leaf1# set forwarding-options enhanced-hash-key ecmp-dlb egress-quantization max ?
Possible completions:
  <max>                Max in percentage (1..100)
```

NOTE: The default values for different DLB modes are given below:

In per packet mode-

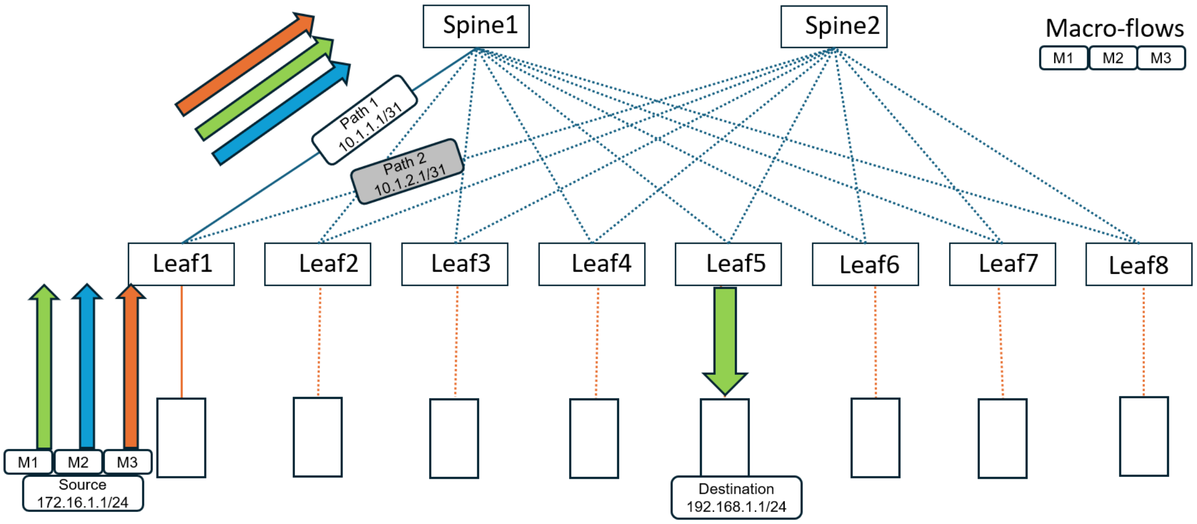
Egress Port Load Weight	: 50
EgressBytes Min Threshold	: 80
EgressBytes Max Threshold	: 99

In flowlet mode -

Egress Port Load Weight	: 50
EgressBytes Min Threshold	: 20
EgressBytes Max Threshold	: 50

Reactive Path Load Balancing

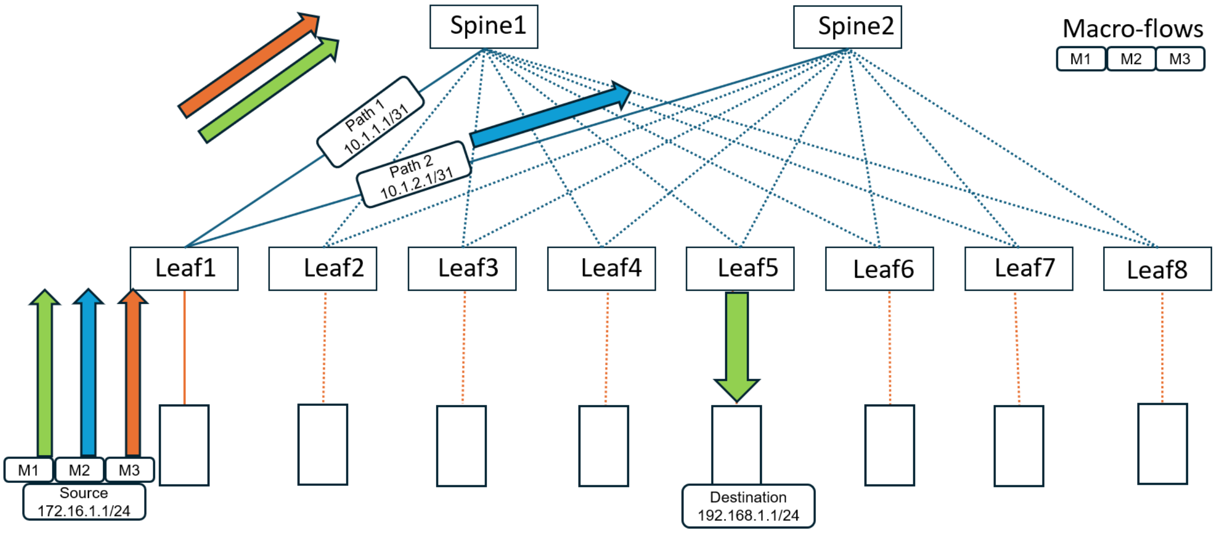
Figure 4: Macro-Flow Polarization at the Beginning



In DLB flowlet mode, macro-flows can get initially assigned to a specific ECMP member link, such as Path 1 in **Figure 4**.

However, if a better-quality member becomes available and reactive path load-balancing is enabled, one or more of the macro-flows can get reassigned to newly available link (Path 2) as shown in **Figure 5**.

Figure 5 - Macro-Flow Reassignment by Reactive Path Load Balancing



Note that when this reassignment occurs, there is a chance that the destination will receive out-of-order packets as packet reordering can occur during the transition period.

On the QFX5240 platform, it is possible to configure this flowlet reassignment based on the quality-delta. Note that an incorrect value of quality delta might result in the flows being re-assigned unnecessarily leading to traffic issues.

```
set forwarding-options enhanced-hash-key ecmp-dlb flowlet reassignment quality-delta <value>
```

Selective DLB

DLB is enabled globally and although there are mechanisms to define flowset size and macro-flow assignment, there was a need for a way in which one can manually define which micro-flow can be filtered for SLB versus DLB. The Selective DLB mode was developed so that one or more micro-flows can be filtered and selectively sent across for DLB.

Method:

- DLB is enabled globally.
- A firewall filter mechanism is used to match traffic type as shown in Output8.
- Per-flow / per packet DLB can be enabled for those filtered packets/flows.
- All other flows will use SLB.

Output8: Selective DLB Configuration Example

```
leaf1# set forwarding-options enhanced-hash-key ecmp-dlb ?
Possible completions:
+ apply-groups         Groups from which to inherit configuration data
+ apply-groups-except  Don't inherit configuration data from these groups
  assigned-flow        Flow-based fixed link assignment
> egress-quantization  Configure egress attributes quantization
> ether-type           Ether type
> flowlet              Inactivity-based flowlet link assignment (default)
  per-packet           Per-packet optimal spraying

--snip—

leaf1# set firewall family inet filter f1 term t1 from ?
Possible completions:
```

```
+ apply-groups         Groups from which to inherit configuration data
+ apply-groups-except  Don't inherit configuration data from these groups
> destination-address  Match IP destination address
+ destination-port     Match TCP/UDP destination port
+ destination-port-range-optimize  Optimize the destination port range
> destination-prefix-list  Match IP destination prefixes in named list
+ dscp                 Match Differentiated Services (DiffServ) code point
  first-fragment       Match if packet is the first fragment
> flexible-match-mask  Match flexible mask
> flexible-match-range  Match flexible range
+ icmp-code            Match ICMP message code
+ icmp-type            Match ICMP message type
> interface            Match interface name
+ ip-options           Match IP options
  is-fragment          Match if packet is a fragment
+ packet-length        Match packet length
+ precedence           Match IP precedence value
+ protocol             Match IP protocol type
+ rdma-opcode          Match on InfiniBand Base Transport header opcode
+ rdma-opcode-except   Do not match on InfiniBand Base Transport header opcode
> source-address       Match IP source address
+ source-port          Match TCP/UDP source port
+ source-port-range-optimize  Optimize the source port range
> source-prefix-list   Match IP source prefixes in named list
  tcp-established      Match packet of an established TCP connection
  tcp-flags            Match TCP flags (in symbolic or hex formats)
  tcp-initial          Match initial packet of a TCP connection
+ ttl                  Match IP ttl type

leaf1# set firewall family inet filter f1 term t1 then accept
leaf1# set firewall family inet filter f1 term default then dynamic-load-balance disable
--snip—
```

Summary

This document details the different load-balancing mechanisms available with the QFX Series switches and their evolution. It elucidates the methods by which each of them functions along with configuration and validation examples. The DLB algorithms can be customized in various ways; however, one should exercise caution to make sure that the timers and other customizations are suited for individual network types and scenarios.

Authors Acknowledgement

Avinash Patil, Suraj Kumar, and Sanoop Rajan

Corporate and Sales Headquarters
Juniper Networks, Inc.
1133 Innovation Way
Sunnyvale, CA 94089 USA
Phone: 888.JUNIPER (888.586.4737)
or +1.408.745.2000
Fax: +1.408.745.2100
www.juniper.net

APAC and EMEA Headquarters
Juniper Networks International B.V.
Boeing Avenue 240
1119 PZ Schiphol-Rijk
Amsterdam, The Netherlands
Phone: +31.207.125.700
Fax: +31.207.125.701