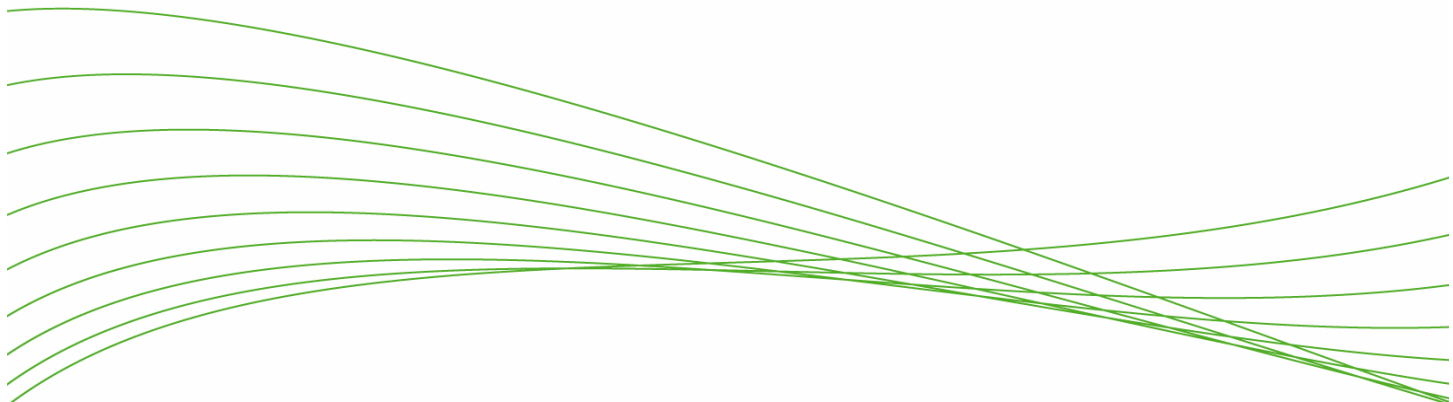


ProCurve Switch 8100fl Series Interconnect Fabric

Technical Brief



Introduction	2
Product Overview	2
Performance	3
Security	3
Scalability	3
High Availability	3
Hardware	3
Modules	4
Console Ports	5
Fan Trays	5
Chassis Power Supply	6
Software	6
Software Architecture	6
Initial Configuration	6
System Software Memories	6
Software Installation	6
Software Synchronization on Redundant Management Modules	7
Hardware Architecture	7
Data Plane	8
Packet Management	9
Packet Processors	9
Traffic Manager	10
Backplane Interface Between Traffic Manager and Fabric	10
Fabric	11
Data Plane Review: Packet Walkthrough	12
Control Plane	13
Management Plane	13
Summary	14
For more information	15

Introduction

The ProCurve Switch 8100fl Series represents a new class of product – Interconnect Fabric – for ProCurve Networking by HP, enabling enterprises to design and interconnect networks based on the ProCurve Adaptive EDGE Architecture. The 8100fl Series offers high-performance, high-availability, cost-effective connectivity for intelligent edge devices, while delivering a flexible, scalable, high port density Gigabit and 10 Gigabit Ethernet (10GbE) core networking solution. It complements ProCurve's traditional core product offerings and ProCurve's Intelligent Edge Switches, providing customers with investment protection and utmost choice and flexibility in designing their network.

The 8100fl Series, featuring multiport, modular switches that perform non-blocking, wirespeed, Layer 2 switching, Layer 3 routing and Layer 4 application switching, delivers exceptional functionality and cost-efficiency to the end user. It consists of two chassis configurations – the ProCurve Switch 8108fl and ProCurve Switch 8116fl – both capable of leveraging ProCurve Intelligent Edge Switch offerings designed for the ProCurve Adaptive EDGE Architecture.

This technical brief outlines the hardware, software and architecture of the ProCurve Switch 8100fl Series.



Figure 1. ProCurve Switch 8100fl Series Interconnect Fabric products.

Product Overview

The ProCurve Switch 8100fl Series consists of two models: the Switch 8108fl and the Switch 8116fl.

The ProCurve Switch 8108fl (J8727A) is an eight-slot chassis-based routing switch delivering 119 million pps, wirespeed non-blocking performance for up to eight 10GbE ports (16 10GbE ports using the two-port X2 10GbE module, oversubscribed 2:1) or 80 100/Gigabit Ethernet ports. The 8108fl is ideal for medium-to-large networks and provides high performance and highly available core switching and routing for ProCurve Adaptive EDGE Architecture applications as well as for collapsed backbones, data centers and server farms.

The ProCurve Switch 8116fl (J8728A) is a 16-slot chassis-based routing switch delivering 238 million pps, wirespeed non-blocking performance for up to 16 10GbE (32 10GbE ports using the two-port X2 10GbE module, oversubscribed 2:1) or 160 100/Gigabit Ethernet ports. The 8116fl is ideal for large networks and provides high performance and highly available core switching and routing for ProCurve Adaptive EDGE Architecture applications as well as for collapsed backbones, data centers and server farms.

Performance

Due to its packet forwarding hardware design built for 10 Gbps, the 8100fl Series offers line rate performance and the ability to support future technologies. In addition, since a portion of the operating system is run locally on the interface module, control plane processing is optimized and communication between the control plane and management module is minimized.

The 8100fl Series also provides the potential for rich Quality of Service (QoS) and bandwidth management features that enhance system performance. These include ingress rate limiting, QoS assignment, sophisticated queuing, and egress management, such as minimum bandwidth guarantees and maximum bandwidth shaping.

Security

All ProCurve products are built with industry-standard security protocols and offer utmost protection against malicious agents.

The 8100fl Series offers rate limiting and QoS features, which prevent denial of service (DoS) attacks to the control and management planes. In addition, the management plane provides secure interfaces, such as SSH and SCP (Secure Copy Protocol).

Scalability

The 8100fl Series offers exceptional scalability with its flexible, chassis-based system. The chassis itself is designed with performance upgrades in mind, with the ability to improve overall system performance in future generations without a chassis upgrade and maintain non-blocking, line rate performance. As such, the system will grow in accordance with future business and technical requirements.

In addition to hardware upgrades, the 8100fl Series can be scaled through future software releases.

High Availability

The 8100fl Series offers redundant management and fabric modules to provide high availability. Redundant paths between interface modules and fabric modules are supported, so the interface modules have seamless failover capabilities should the fabric modules receive errors. In addition, the management module supports image and configuration synchronization automatically. Redundant power supplies and redundancy in the fans promote maximum availability.

Hardware

The 8100fl Series is an Interconnect Fabric switch product, not a network edge switch product. As such, it is not a traditional plug-and-play device like ProCurve's Intelligent Edge Switches.

The Switch 8108fl is an eight-slot chassis that comes with one management module, one switch fabric module, and one power supply. A single power supply is sufficient to power up a fully loaded eight-slot chassis. An optional second power supply provides full power redundancy.

The Switch 8116fl is a 16-slot chassis with one management module, one switch fabric module and two power supplies. The two power supplies are sufficient to power up a fully loaded 16-slot chassis. An optional third power supply provides N+1 redundancy.

Both switch models share the same "fl" Series interface modules and redundant management module. The redundant fabric modules, however, are unique to each model, with an 8-slot version and a 16-slot version.

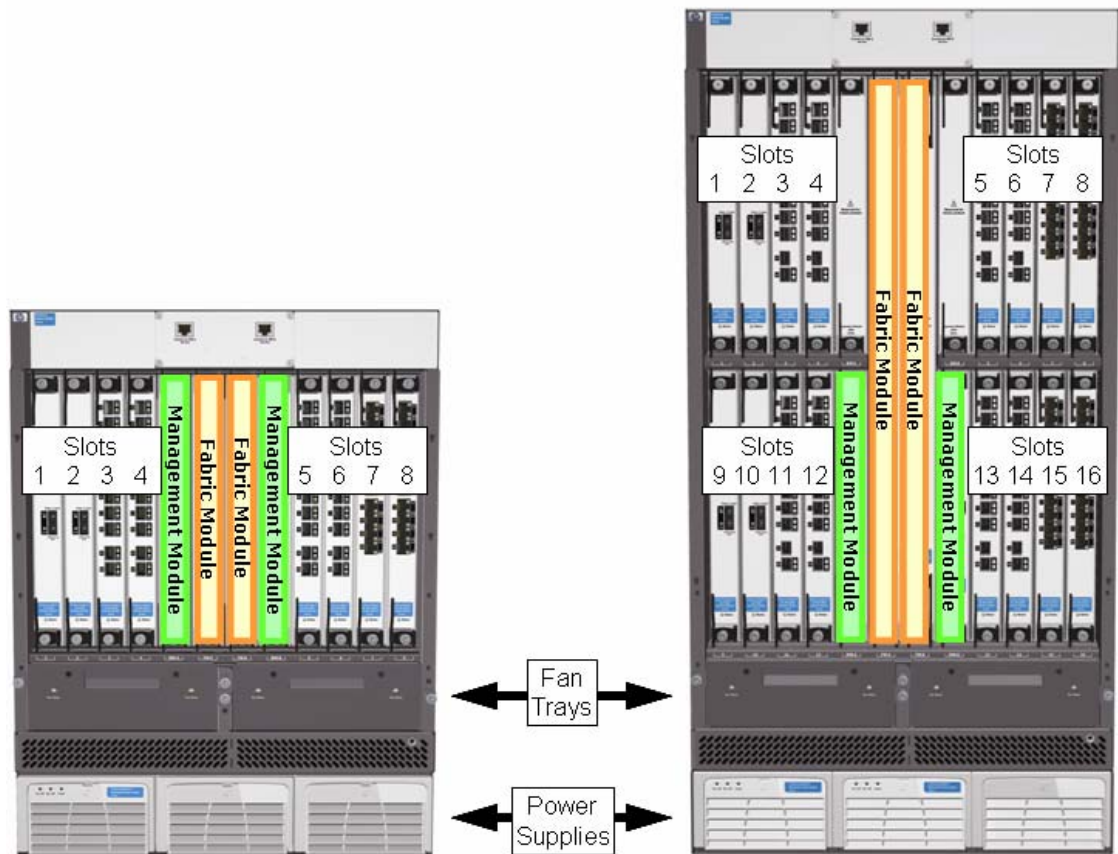


Figure 2. 8100fl Series chassis overview.

Modules

Each 8100fl Series chassis comes with a management module and switch fabric module. An optional, second management module or second fabric module can be purchased for redundancy. All interface modules are ordered separately.

- Switch fl Redundant Management Module (J8731A)** – The 8100fl Series allows for two management modules in a single chassis. One is included with each chassis and an optional second management module can be installed for redundancy. Management modules can be inserted in either or both slot MM-A (management module A) and slot MM-B (management module B); there is no priority as to which slot actively manages the chassis. However, if two management modules are in a single chassis, only one is active at a time. The management modules have a built-in 10/100Base-T Ethernet port, which is not a part of the switching data plane and is used for management purposes only. This management port does not support Auto-MDIX and is configured as MDI, requiring the use of a crossover cable if connected directly to most PCs with 10/100 Ethernet ports. If connected to another switching device that does not provide Auto-MDIX detection, a straight-through cable will be required. The management modules communicate with other modules in the chassis through a 100 Megabit Full-Duplex Ethernet control plane. These modules have a PCMCIA card slot for future use (capabilities not yet available). The module has two LEDs on the bottom, one to indicate system status (a green/orange bicolor LED to indicate booting, running or fault conditions) and one to indicate whether it is the active or standby management module for the system (green LED).
- Switch fl 2-Port 10GbE X2 Module (J8737A)** – This two-port, X2 form factor 10 Gigabit module accepts X2 10Gbps transceivers, a 2:1 oversubscribed full duplex, and a 10 Gigabit module. Each port has Link and Activity LEDs. The module has a bicolor green/orange LED on the bottom to indicate system status (booting, running or fault conditions). The throughput capacity of the J8737A is 10Gbps, regardless of whether one or both ports are linked. This

provides connectivity to multiple 10Gbps links, and automatically balances ingress traffic between the two ports.

- **Switch fl 10-Port 100/1000-T Module (J8734A)** – This module has 10 ports of 100/1000Base-T, providing for up to line rate Gigabit connectivity. Each RJ-45 port has embedded LEDs: a bicolor LED to indicate link state and speed (green for Gigabit, orange for 100 bps), and another to indicate port activity. The module has a bicolor green/orange LED on the bottom to indicate system status (booting, running or fault conditions).
- **Switch fl 10-Port Mini-GBIC (SFP) Module (J8735A)** – This module delivers 10 ports of full-duplex, line rate Gigabit connectivity. It supports the B-versions of ProCurve Mini-GBIC accessories, including the new 1000Base-T Mini-GBIC device (J8177B). Each port has two LEDs, one to indicate link state and one to indicate port activity. The module has a bicolor green/orange LED on the bottom to indicate system status (booting, running or fault conditions).

In contrast to the modules that are common to both chassis, the fabric modules are unique to each system. There is one for the 8108fl chassis and another for the 8116fl chassis. Both modules have two LEDs, one to indicate system status (booting, running and fault conditions) and one to indicate whether it is the active or standby fabric module.

- **8108fl Redundant Switch Fabric Module (J8729A)** – This fabric module for the eight-slot chassis can be inserted into either of the two center (FM-x) slots and provides for near-hitless failover (0.2 second failover) for switched traffic when used with a second fabric module.
- **8116fl Redundant Switch Fabric Module (J8730A)** – Taller in size than the 8108fl Redundant Switch Fabric Module, this fabric module for the 16-slot chassis can be inserted into either of the two center (FM-x) slots and provides for near-hitless failover (0.2 second failover) for switched traffic when used with a second fabric module.

Console Ports

Each chassis contains two RJ-45 serial console ports, one for Management Module A and one for Management Module B. These ports are used to directly connect an RS-232 serial management console to the switch. The management module that is active (indicated by the “Active” LED or the CLI command “show redundancy”) dictates which console port should be utilized. The user can employ the console port only for RS-232 out-of-band communication; it cannot be used for a Telnet connection. For network connections (Telnet, SSH, FTP, TFTP, SCP) to the management module, use the 10/100Base-T port located on each of the management modules. An RJ45-to-DB9 adapter is included with every chassis.

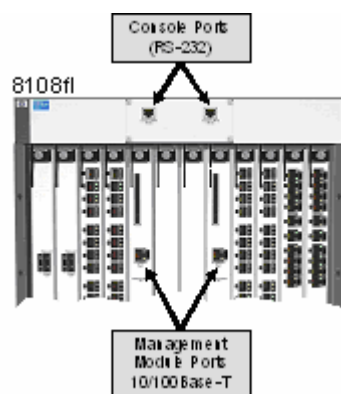


Figure 3. 8100fl Series console and management ports.

Fan Trays

8100fl Series chassis contain two fan tray assemblies to provide airflow across the switch modules. The fan tray assembly is hot swappable to allow for easy replacement of fans. Each fan tray assembly contains three fans to ensure no single fan failure is catastrophic. However, a failure of the entire fan tray renders half the system unusable due to temperature warnings and power shutdowns.

Each interface, management and fabric module has a temperature sensor, and temperature warning thresholds are user-configurable. When the temperature reaches the user-defined warning level, an error message and SNMP trap are generated. When the temperature reaches a critical level (preset at the factory), the electrical power to that module is turned off to prevent damage to the components. Remaining modules continue to operate as long as they remain below the critical temperature level.

Chassis Power Supply

The Switch 8108fl comes standard with one power supply and the Switch 8116fl comes standard with two power supplies, sufficient for a fully loaded configuration. However, up to three power supplies can be installed in a single chassis for redundancy. Each power supply operates between 100 and 240 VAC, and the system components (management module, interface module, etc.) regulate their own power needs. Multiple power supplies are bus connected, so it does not matter which one is powered on. However, a separate power cord must be utilized for each power supply (located on the back of the chassis). The chassis utilize a C19 connector for power cord connections (notable for rack cabinets with integrated power cables).

Software

Software Architecture

The 8100fl Series features a distributed software system, designed to ensure high availability and performance. Every card, including the interface modules and redundant fabric modules, runs an operating system.

The distribution of software to multiple CPUs optimizes tasks and CPU load to relevant cards. The management module CPU provides overall system management and does not forward data packets. The interface module CPU provides exception packet processing and local hardware control.

Initial Configuration

As noted earlier, the Switch 8100fl Series is unlike Intelligent Edge Switches as its place in network designs is slated for the core of the network. It requires some initial configuration to be connected properly to other switches. By default, the ports are disabled in a "shutdown" state. Once they are configured, individual ports must be enabled by issuing a "no shutdown" command.

System Software Memories

8100fl Series management modules have a 512 MB Compact Flash memory device (device name "flash:"). This device is used for intermediate storage of files during software upgrade procedures as well as storage for system crash dumps. These crash dumps can be offloaded and sent to ProCurve Support for further analysis.

Each management module reserves an area on this Compact Flash device for two banks of system software – Bank-A and Bank-B – each of which can hold a different version of the system software. These different versions of software can be run using the same configuration file.

Software Installation

When software is loaded onto the system, an image is transferred from an outside source onto the 8100fl Series' Compact Flash memory. This can occur using a number of protocols, such as TFTP, FTP or SCP (Secure Copy). Once the image is copied to the flash device, it can be installed in either Bank-A or Bank-B. The system is then configured to boot from one or the other bank (this setting is stored in nonvolatile RAM for the next reboot cycle and initially factory-set for Bank-A). When the system boots the operating system is extracted from the designated bank. Once the management module is up and running, each interface module receives its software over the management plane.

8100fl Series software installation and management procedures are different from most ProCurve devices. For example, on a ProCurve Switch 5300xl series, the command for TFTP-to-Flash copying is:

```
copy tftp flash 15.15.15.10 full.bin.CY.01.02.0050 primary
```

The 8100fl Series offers a “flash:” storage device on the system, complete with “directory,” “mkdir,” “rmdir” and “cd” commands applicable to a Unix storage device. The software installation task on an 8100fl Series is a two-step process: Copy the file from a server onto the 8100fl “flash:” device, specifying the source device as a URL, and then install the image into the management module(s) memory:

```
copy tftp://15.15.15.10/full.bin.CY.01.02.0050 flash:
image install flash:full.bin.CY.01.02.0050
```

The software for all modules is contained in the single system software image. Installation of the software images for each fabric or interface module is performed from this single image installed and running on the active management module during the boot process.

To utilize other file transfer protocols such as FTP or SCP, the username must be specified in the URL and the system will prompt the user for password access once the connection is established to the file server. For example:

```
copy ftp://{UName}:{passwd}@15.15.15.10/{FILENAME} flash:
```

```
copy scp://{remoteUser}@15.15.15.10/{FILENAME} flash:
```

The system will then query the user for the {remoteUser} password.

Copying files or configurations from an 8100fl Series to a TFTP or FTP server is a similar process, where the user specifies the target destination as a URL. For example:

```
copy flash:My_Config tftp://15.15.15.10/Config_Oct05.txt
```

```
copy startup-config ftp://{UName}:{passwd}@15.15.15.10/{FILENAME}
```

Software Synchronization on Redundant Management Modules

The “image install” command has options to install software on individual management modules as well as specific banks, although users cannot install software onto the running bank.

System software installation, by default, is targeted to the bank (A or B) that is not currently the running bank. For example, if the management module was booted from Bank-A, installation of new software will be to Bank-B.

When a system has two management modules installed, the default version of the “image install” command attempts to install the software on the same bank on both management modules. If there is a mismatch of the two management modules (e.g., Management Module A is running on Bank-A while Management Module B is running on Bank-B), an error message is displayed. This allows the user to boot the other management module to the same bank as the active management module.

Assuming Management Module B is the standby module (not active) and Management Module A is running on Bank-A, the command to reboot Management Module B to Bank-A is:

```
boot system management b bank-a
```

Since Management Module B is the standby module, no interruption to the 8100fl Switch service will occur by rebooting this standby module. Once Management Module B is up and running on Bank-A, the “image install” command can be issued again to properly synchronize both management modules with the same software; in this case, into Bank-B.

Hardware Architecture

8100fl Series chassis consist of one or two management modules, one or two fabric modules, and one to 16 interface modules. When a second management module is present, the

management modules are redundant components that provide failover of system management. One is active and the other is standby. Configurations and images are synchronized and each module is actively monitoring the state of the other.

Similarly, when a second fabric module is present, hardware and software components are monitoring the state of each. The standby fabric module is always maintaining the state of the data path, even though it is not actively forwarding traffic. This allows for rapid failover since the secondary fabric module does not need to re-establish state if the primary fabric module fails.

In terms of system logic, the 8100fl Series is represented by three major subsystems: the data plane, the control plane and the management plane. The data plane contains the elements that forward network traffic. The control plane dynamically configures and monitors the data plane and implements network protocols. The management plane provides user and network management interfaces and statically configures and monitors the entire system.

Figure 4 provides a system-wide view of all three planes. The data plane is confined to elements on the interface modules and fabric modules. The control plane is distributed throughout all modules in order to provide optimum processing. The management plane is located mostly on the management modules though some monitoring components (for statistic counters) are located on other modules. Note that the management plane has completely out-of-band interfaces – separate from the control and data planes - for access to configuration and system control.

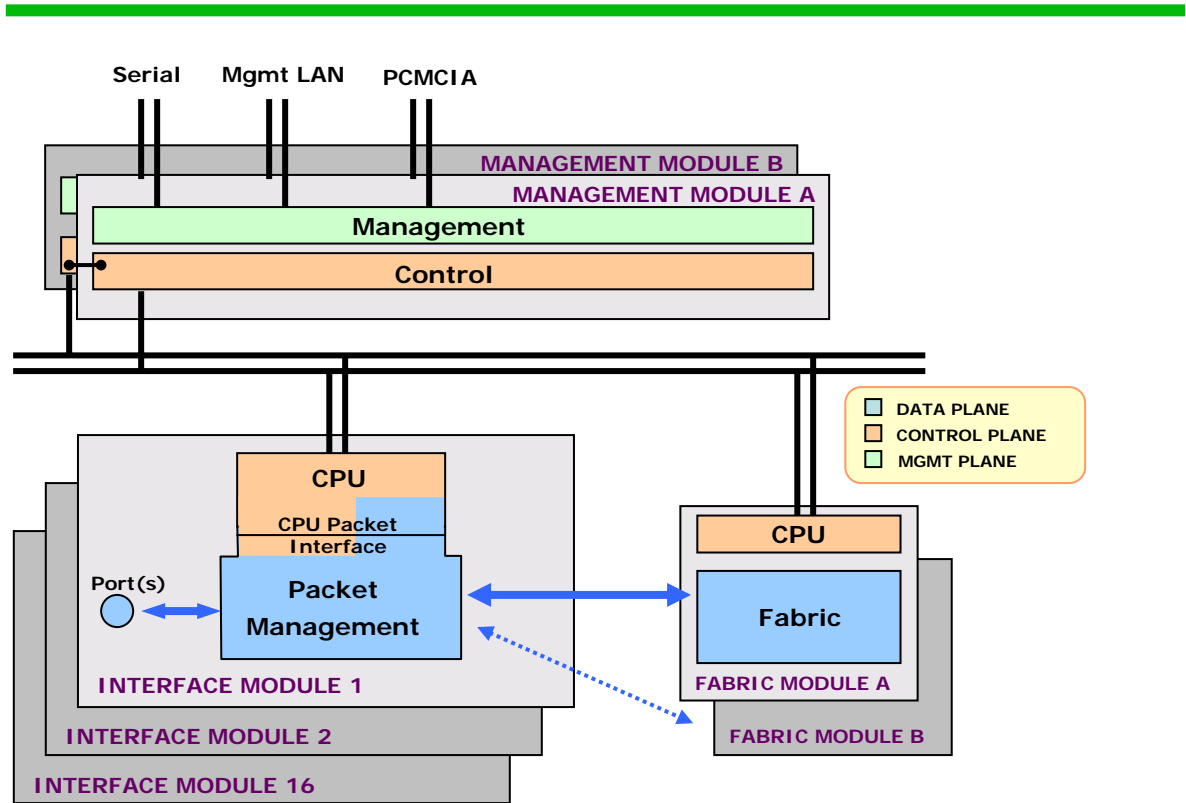


Figure 4. 8100fl Series architecture.

Data Plane

The 8100fl Series data plane provides all of the processing and forwarding of packets throughout the system, which are primarily implemented in hardware via application specific integrated circuits (ASICs). Only rarely, and for specific Internet Protocol (IP) functionality such as fragmentation, will packets be intercepted and forwarded via software. In such cases, all forwarding is handled locally on the ingress interface module, not by a management module.

The data plane components are located on the 8100fl Series interface modules and fabric modules. Each interface module has redundant data plane connection to both fabric modules. Since the redundant fabric modules are in either an active or standby state, only one of the connections is active. However, a standby fabric module is always maintaining the dynamic state of the active module for rapid failover.

Figure 5 details the components of the 8100fl Series data plane. On the interface module, the packet management block consists of packet processors – one for ingress and one for egress – and a traffic manager. On the fabric module, the fabric consists of a crossbar switch and a bandwidth manager. The data path is completely full duplex. Packets simultaneously transit the components of the data plane in both directions with no performance degradation. The ports on an interface module include components not shown, such as physical layer (PHY) optics and media access control (MAC).

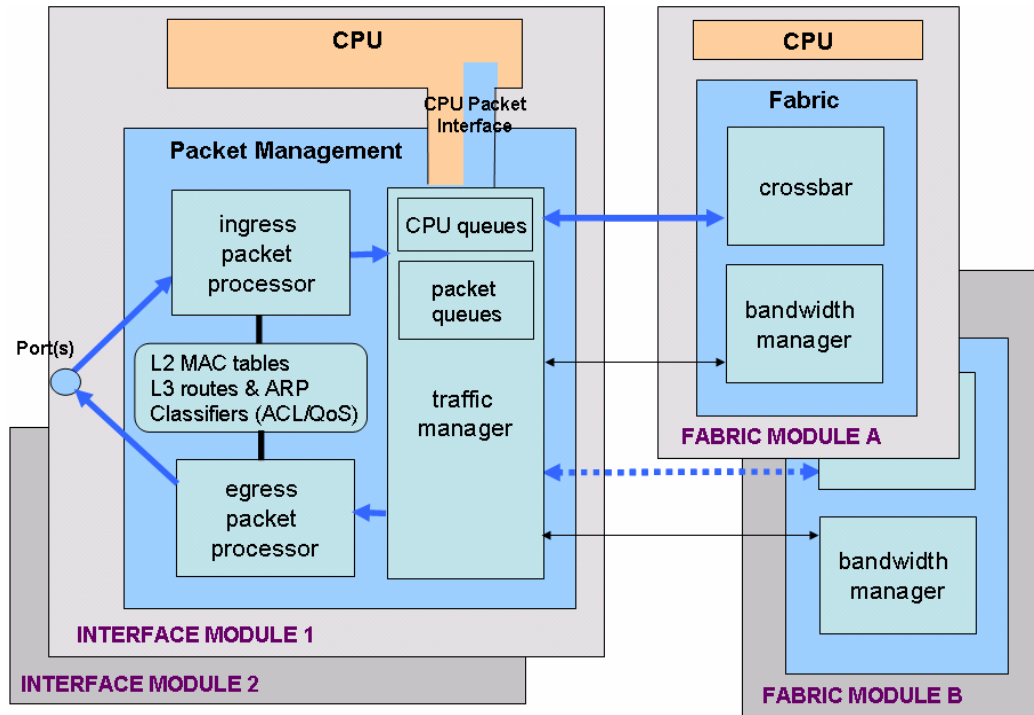


Figure 5. Components of the 8100fl Series data plane.

Packet Management

The packet management block is responsible for receiving and sending packets from ports and either managing their access to the fabric, or sending them to the local CPU for software processing. This block determines the destination of all forwarded packets by performing line rate routing and bridging. In the event of congestion on the egress port, packets are queued in very large, high-speed packet buffers. These queued packets can be treated with highly granular QoS features.

Packet Processors

The ingress packet processor ensures the security of inbound packets by processing their headers at line rate, enforcing and assigning their VLAN and preparing them for bridging and routing processing.

All Layer 2 bridging and Layer 3 routing forwarding decisions are made in hardware by the ingress packet processor. Every interface module has this information (distributed by the management module); therefore, local bridging and routing functions are distributed throughout the system. Since forwarding decisions are made locally on the interface modules and not by the central management module, system performance is optimized.

In the event that new Layer 2 MAC addresses must be learned for bridge updates, the hardware automatically refers the MAC addresses to the CPU for processing.

Packets are matched to rich access control entries by the ingress packet processor. The first release allows for 1000 entries with resultant actions to accept or deny. The design allows for a variety of potential actions on matched ACLs such as rate limit, assign QoS, policy route and mirror.

The egress packet processor updates various Layer 2 and Layer 3 packet headers such as VLAN tags, MAC addresses, 802.1p and IP-TOS class of service, IP TTL, etc. It also performs final replication of packets, if needed, for multicast.

The functionality of the ASIC-based packet processors are firmware upgradeable. This allows for future proofing and investment protection since new features can be added over time. For example, the ingress packet processors are capable of defining robust access control entries that control the assignment of QoS classes or the means by which packets are rate limited with highly granular controls.

The packet processors perform all functions at line rate.

Traffic Manager

The Traffic Manager contains large, high-speed packet buffers in the event that the destination (egress) port is congested. However, an egress will only become congested if high levels of traffic from multiple ingress ports are sent to a common egress port and the sum of the traffic is greater than the media speed of that port. Since packets are buffered before the fabric on ingress, the architecture uses advanced technology, called virtual output queues, which allocates queues for every egress port in order to prevent head-of-line blocking on the input port. The virtual output queue architecture allows for rich bandwidth management and QoS features.

The architecture can accommodate up to eight queues per egress port. The first release will provide five queues and will for Differentiated Services (DiffServ) per-hop behaviors such as three assured forwarding classes, an expedited forwarding service class and default class. Each queue has weighted random early detection (WRED) with three levels of drop precedence per queue.

The size of a queue is allocated dynamically by hardware as buffer space is required. Each interface module provides 128 MB of memory for this purpose. This allows queues with more congestion to get a greater amount of space when needed than other queues resulting in an optimum use of resource. In summary, the total buffer space per egress queue, system-wide, will vary between 64 KB to 25 MB, with 2.5 MB as an approximate average.

The selection of the queue is determined by the ingress packet processor based on packet header processing.

Packets destined for the CPU, such as protocol packets for the control plane, are treated by dedicated QoS queues and rate limiting features. This provides a secure interface to the CPU and prevents DoS attacks on the control plane, affecting regular switch data plane traffic.

In addition, the traffic manager has separate multicast queues to avoid unicast traffic blocking and provides efficient replication of multicast.

Backplane Interface Between Traffic Manager and Fabric

Interface modules are connected physically via the chassis backplane to the fabric modules. In order to provide redundancy, each interface module is connected via separate channels to each fabric module. The channel to a fabric module consists of two types of interfaces: a packet data interface and a control interface.

The packet data interface is a 20 Gbps full-duplex (bidirectional) interface. It includes fabric speedup, which allows the fabric to be completely non-blocking, and line rate forwarding, for a fully loaded system. This results in a raw fabric capacity of 640 Gbps in the 8116fl chassis and 320 Gbps in the 8108fl chassis. The effective throughput is 10 Gbps full duplex per interface module.

For investment protection, the chassis backplane is built with additional interfaces to every interface and fabric module to accommodate future capacity expansion. This will allow for line rate forwarding of higher port count interface modules in the future.

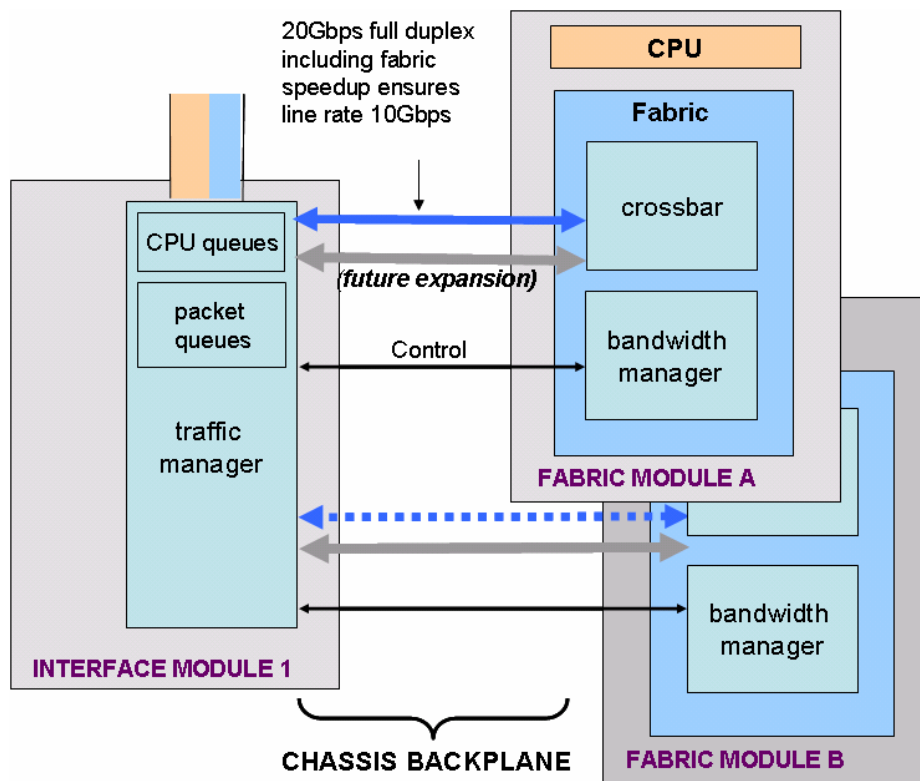
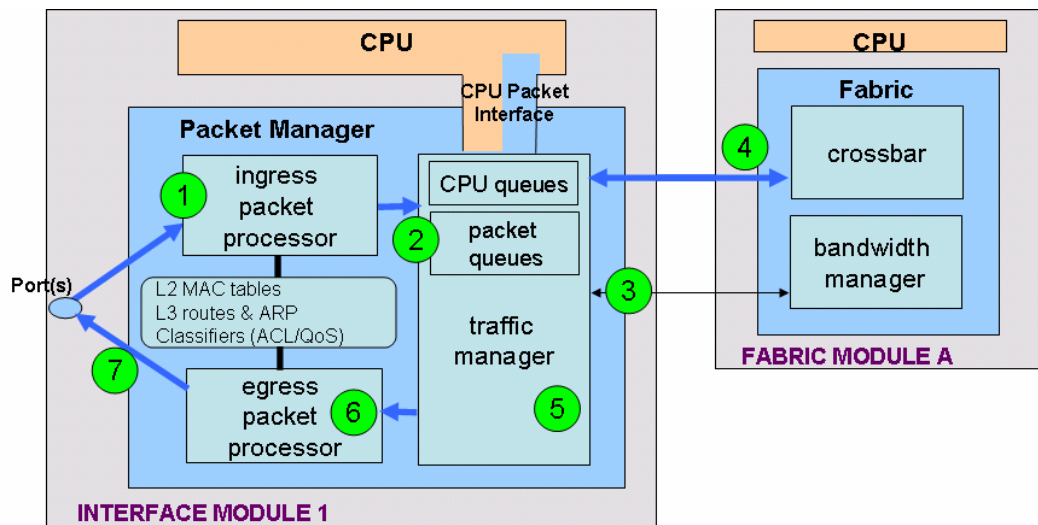


Figure 6. Backplane Interface between Traffic Manager and Fabric.

Fabric

A bandwidth manager coordinates with all interface module traffic managers for global switch scheduling. Each traffic manager provides continuous status on all queues to bandwidth managers on both the active and standby fabric modules. The bandwidth manager determines which traffic manager is granted access to the switch crossbar via sophisticated bandwidth allocation algorithms. The algorithms allow for features in guaranteed minimum bandwidth provisioning in 1 Mbps increments and also, weighted fair queuing or strict priority scheduling. Future capabilities — such as egress traffic shaping (maximum bandwidth) — can be accommodated. The crossbar provides the non-blocking fabric to interconnect all traffic managers for all interface modules.

For efficiency and minimal use of crossbar bandwidth, the fabric also performs first-stage packet replication for multicast packets sent to different interface modules (one packet into fabric and one packet out to each interface module). The second stage of multicast replication is conducted by the receiving traffic manager for the set of egress ports and the final stage of multicast replication is conducted by the egress packet processor to a single egress port. This multistage replication of multicast provides the most efficient use of system-wide resources and bandwidth.



1. Ingress packet processing performed
2. Traffic Manager queues packet
3. Bandwidth Manager coordinates scheduling
4. Crossbar fabric interconnects Interface Modules
5. Traffic Manager on egress card receives packet
6. Egress packet processing performed
7. Packet transmitted to physical interface

Figure 7. Data Plane Packet Walkthrough.

Data Plane Review: Packet Walkthrough

Figure 7 shows the numbered steps as data packets traverse the switch.

1. The packet is received by an interface module port via physical layer interfaces (PHY), such as optics, and media access control (MAC), and then sent to the ingress packet processor.

- Ingress header processing (e.g., VLAN assignment)
- Forwarding lookup (e.g., Ethernet bridging, IP routing)
- In-line classification (e.g., ACLs, filters, policy for QoS and future rate limiting)
- Policing (future rate limiting with three-color marker and 1 Mbps granularity)
- Assign Class of Service (based on 802.1p, DiffServ, etc.)
- Link aggregation and ECMP (Equal Cost Multi-Path) trunk selection

2. The traffic manager receives QoS prioritized packet from the input packet processor and enqueues the packet.

- Packets assigned to virtual output queues, up to 5 CoS queues per egress port
- Separate multicast queues avoid unicast traffic blocking
- Dynamic memory assignment to individual queues from large memory pools
- Congestion management (WRED) performed on individual queues based on packet drop-precedence ("color") with three profiles per queue

3. Bandwidth manager coordinates with all interface module traffic managers for global switch scheduling.

- Each traffic manager provides continuous status on all queues to active and standby bandwidth managers on each fabric module
- Bandwidth manager determines which traffic manager is granted access to crossbar fabric via sophisticated bandwidth allocation algorithm

For more information

To learn more about ProCurve networking solutions, contact your local ProCurve sales representative or visit the company's web site at www.procurve.com.

For a list of ProCurve Elite Partners that can provide ProCurve solutions, go to www.procurve.com and click on "Resellers."

To find out more about
ProCurve Networking
products and solutions,
visit our web site at

www.procurve.com



© 2007 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice. The only warranties for HP products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HP shall not be liable for technical or editorial errors or omissions contained herein.

4AA1-1914ENW, 04/2007