JUNIPER NETWORKS | Engineering Simplicity

DAY ONE GREEN: OPTIMIZING NETWORKS WITH EFFICIENT SYSTEM DESIGN

Juniper has made concerted efforts to develop power-efficient features at both the hardware and software level that you can implement today.

By Kapil Jain, Eswaran Srinivasan, and Unmesh Agarwala

Day One Green

Optimizing Networks with Efficient System Design

Developing power-efficient networking equipment is a key design metric for Juniper Networks, driven by its customers to minimize their operational expenses and energy consumption. This in turn reduces their carbon footprint and can help them reach their carbon targets during the forthcoming decade. Juniper has been purposefully addressing greenhouse gas (GHG) emissions by making a concerted effort to develop power-efficient networks at both the hardware and software levels.

A typical Juniper modular router platform has the following components:

- Common hardware components including power supplies and fan trays.
- One or two Routing Engines (REs)
- One or more Flexible PIC Concentrators (FPCs)
- One or more Physical Interface Cards (PICs) or Modular Interface Cards (MICs) per FPC
- One or more Switch Interface Boards (SIBs)

At a high level, the REs provide management connectivity to a router. With dual REs, it also provides hot redundancy for a RE failure scenario. The main purpose of the REs is to have a common routing control plane in a router.

Each FPC contains one or more Packet Forwarding Engines (PFEs) for the processing and forwarding the packets.

The PICs/MICs hosted by a FPC can be fixed as part of the FPC and or can be pluggable. A PIC/MIC supports multiple optics cages with different port speeds for providing WAN connectivity. Certain PICs/MICs have direct connectivity between the optics cages and PFE complex while certain PICs/MICs have gearboxes/retimers between the optics cages and PFE complex.

Each of the PFE complexes in a FPC is connected to all the SIBs in a router to provide a full mesh connectivity without head-of-line blocking.

Figure 1 depicts the common hardware components of a modular router platform.

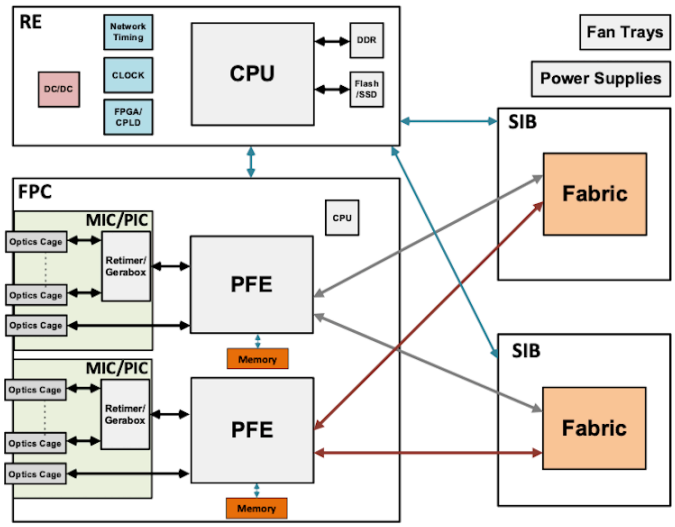

Figure 1 *A Routing and Switching Platform Architecture*

In a networking system, the major power consuming components are: PFEs and their external memories, retimers, switch fabric, CPU subsystem, and fans. Figure 2 details the power distribution of major components as a percentage of total system power dissipation and is based on the measured power dissipation on a Juniper PTX10001-36MR device. Please note that the exact power distribution will vary among different systems but the trend should remain the same.

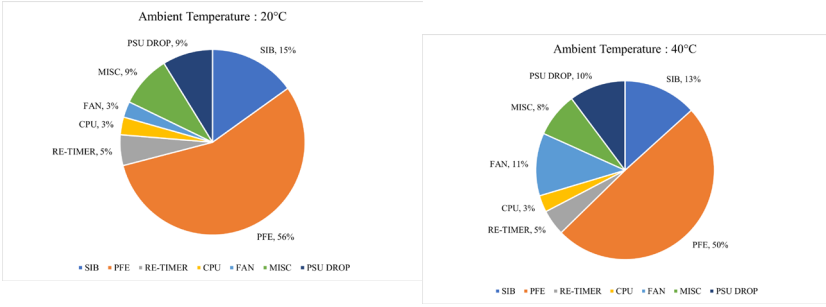

Figure 2 *Power Distribution of Major Components*

To develop a power-efficient system, the focus is to reduce power consumption of these components using the following strategies.

Efficient Thermal Design

Efficient thermal design in our products contributes to reduced system power consumption. (See Day One Green: Optimized Thermal Design). Efficiently applying thermal management solutions can reduce a lot of power consumption in the system. For example, system fans used for cooling the system are the second-highest power consuming components and lowering the fan speed can provide direct power savings. Carefully designed thermal policy makes sure that the fans can run at the lowest speed possible without violating component specifications.

All the Juniper router platforms are designed with efficient cooling systems to support this. As an example, please refer to the following product documentation for more details about the cooling system for PTX10008 and MX10008 router platforms:

- https://www.juniper.net/documentation/us/en/hardware/ptx10008/topics/topic-map/ptx10008-cooling-system.html
- https://www.juniper.net/documentation/us/en/hardware/mx10008/topics/topic-map/mx10008-cooling-system.html

Operational Temperature

Juniper routers are built to operate under different temperature conditions. In general, the power consumption is directly proportional to the ambient temperature. A CLI configuration is available to specify the ambient temperature of a chassis which can help reduce the overall power consumption of the HW FRUs and the provisioned power.

Please refer to the following product documentation for more details about the CLI configuration to specify the ambient temperature of a chassis.

- https://www.juniper.net/documentation/us/en/software/junos/chassis/topics/ref/statement/chassis-ambient-temperature.html

Continuous Monitoring

The temperature of various components are continuously monitored in the router. This includes the PFE ASICs and its external memories. On certain Juniper router platforms, the PFE capacity is dynamically reduced by software when the temperature of the PFE ASIC and/or the external memories goes over a threshold. This mechanism helps to have a better thermal solution with reduced power consumption.

Refer to the following product documentation for more about supporting this functionality for MPC10E-10C-MRATE MPCs on MX240/480/960 router platforms:

- https://www.juniper.net/documentation/us/en/hardware/mx-module-reference/topics/concept/mpc10e-10c-mrate.html

Reset Unused WAN Ports

In many cases, some of the WAN ports are not used in the FPC. It is observed that even though WAN ports are not in use the connected PFE continues to consume power. From day one hardware is designed in such a way that each PFE can be kept in reset if not used. Keeping unused PFEs in reset has a significant impact on overall system power consumption. For example, in the PTX10001-36MR, if an unused PFE is kept in reset, then direct power savings of 150W per PFE is achieved. A CLI command is provided in software for users to keep PFEs in reset when the ports are not in use.

Please refer to the following product documentation for more details about the CLI configuration to power ON/OFF a PFE:

- https://www.juniper.net/documentation/us/en/software/junos/chassis/topics/topic-map/chassis-guide-tm-managing-power.html

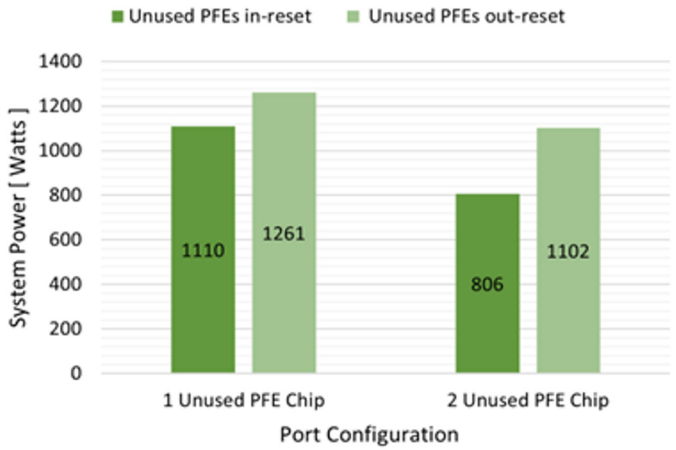

Figure 3 *Comparison of power consumption between cases where unused PFEs are in-reset and out-of-reset.*

NOTE Similarly, provision is made in hardware to keep individual Gearbox in reset if not used. This provides power savings of ~18 W per retimer in a few router platforms.

PFE Configuration Flexibility

The external memories connected to the PFEs offer various functionalities including packet buffers and data memory for lookup in subsystem. The external memory consumption is significant. Juniper ASICs are built with configuration flexibility to use these external memories as needed and the decision can be runtime. The Junos Operating

System (SW) is designed to take advantage of this by default and no CLI configuration is required for this. With it, the external memories are used only when needed which significantly reduces the overall power consumption of a PFE complex.

FPCs Support for Pluggable MICs

On certain Juniper router platforms, the FPCs support pluggable MICs. Certain FPC types offer a lot of forwarding features without relying on the traffic on the WAN ports connected to the MICs. Under these conditions, the MICs are not required to be plugged into the FPCs. When no MICs are plugged into the FPC, Junos won't include the power required to operate the MICs and so the overall power consumption of the FPCs is reduced without compromising the forwarding features. This in turn reduces the overall provisioned power required for the routers. This feature is referred to as MIC-aware power management. See Juniper documentation here:

- https://www.juniper.net/documentation/us/en/software/junos/chassis/topics/topic-map/chassis-guide-tm-managing-power.html

When SerDes Links Are Not Initialized

In the case of the FPCs with multiple PFEs, Junos is implemented to initialize links in such a way that if a PFE is not present the corresponding fabric device SerDes links are not initialized.

Similarly, if the fabric chip is not present the corresponding PFEs fabric SerDes side links are not initialized. Using this approach of initializing SerDes links based on the presence of FPC/SIB results in a net power saving of 11W per FPC and 33W per SIB.

Refer to the following product documentation for more details about the CLI show commands used to display the fabric plane status:

- https://www.juniper.net/documentation/us/en/software/junos/system-mgmt-monitoring/chassis/topics/ref/command/show-chassis-fabric-summary.html

When SerDes Links Are Initialized

WAN ports are connected to PFEs using PAM4 SerDes lanes. To save power, SerDes initialization is done only when the optics are inserted into the port. This results in direct power savings as unused ports' SerDes links are not kept in power off. Using this approach ~5 Watts power per port is saved.

Please refer to the following product documentation for more details about the CLI commands that can be used to sanitize the health of the SerDes used for a WAN port and to display the number of SerDes lanes used by a WAN port:

- https://www.juniper.net/documentation/us/en/software/junos/interfaces-ethernet/topics/task/collecting-prbs-statistics.html

Power-off Unused WAN SerDes Lanes Based on the Port Speed

Building on the idea of initializing only those WAN SerDes lanes where optics are presence, you can also power-off unused WAN SerDes lanes based on the port speed. For example, if a port is configured at 100G speed, then only 4 lanes are active versus a port configured for 400G speed where all 8 lanes are used. This approach results in power savings of 0.6 W per SerDes lane and for each 100G port it would be 2.4W per port. Moving away from static SerDes initialization to dynamic initialization methods, based on optics presence and speed configuration can result in significant power saving.

Refer to the following product documentation for more details about the CLI commands that can be used to sanitize the health of the SerDes used for a WAN port and to display the number of SerDes lanes used by a WAN port:

- https://www.juniper.net/documentation/us/en/software/junos/interfaces-ethernet/topics/task/collecting-prbs-statistics.html

Clock Gating MACsec Blocks

Juniper's MACsec feature is supported on all ports of a PFE complex up to 400G port speed. MACsec blocks are initialized as part of the ASIC initialization process. MACsec blocks are clock-gated during init to stop this power drain. In case you want to use MACsec there is a CLI command to enable and disable clock gating of MACsec. Using this approach of clock gating MACsec blocks when not in use can result in net power savings of 20W per PFE.

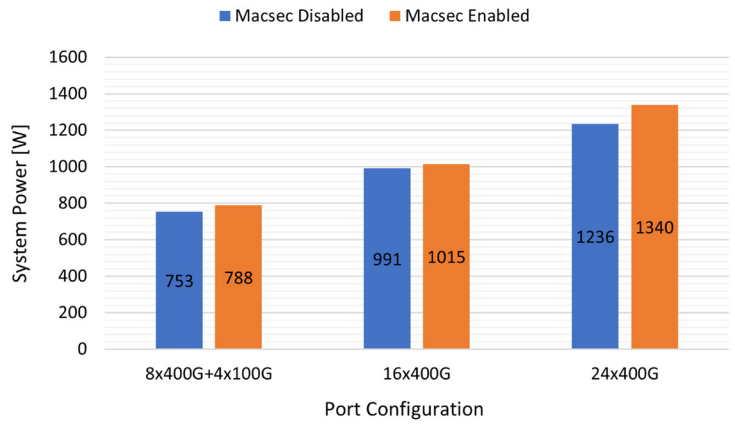

Figure 4 *Comparison of Power Consumption Between MACsec Enabled vs. Disabled*

Refer to the following product documentation for more details about configuring and managing MACsec:

- https://www.juniper.net/documentation/en_US/day-one-books/DO_MACsec_UR.pdf

PFE Chip and Switch Fabric Chip

PFE and Switch Fabric are some of the most power-consuming components in the system. So to reduce overall power consumption of the system, PFE chip and Switch Fabric chip must be addressed. In general, the total power consumed by the ASIC is the sum of static power and dynamic power. Static power is directly proportional to the voltage^2 while dynamic power is directly proportional to the voltage^2 and frequency.

To reduce the ASIC chip power consumption, core clock frequency reduction and core voltage reduction is done without compromising on chip performance.

For example, on PTX10001-36MR, core clock frequency of PFE was reduced to 800 MHz and core voltage is reduced by 40 mV based on PFE load and temperature, which resulted in net power savings of 20 W per PFE chip. Similarly, for switch fabric 10W per chip was net power savings. This is done by default and no CLI configuration is required.

As an another example, MPC8E for MX2008/2010/2020 platforms can be configured in 960G or 1.6T per-slot bandwidth mode. It is worth noting that 960G per-slot bandwidth is the default mode and 1.6T per-slot bandwidth mode can be enabled using a CLI configuration command.

In 960G per-slot bandwidth mode, the PFE ASIC is configured with the core clock frequency of 768 MHz for datapath and 562 MHz for lookup subsystem. In 1.6T per-slot bandwidth mode, the PFE ASIC is configured with the core clock frequency of 862 MHz for datapath and 937 MHz for lookup subsystem.

Refer to the following for more details about configuring MPC8E in 1.6T per-slot bandwidth mode:

- https://www.juniper.net/documentation/us/en/software/junos/chassis/topics/ref/statement/bandwidth-edit-chassis-fpc.html

Summary

You can see there are definitely ways to reduce your power consumption with Juniper devices. You can do these today and weave them into your operational best practices. Always test these techniques in the lab before moving into production environments.