October 2023

DYNAMIC STACK PORT SERDES POWER UTILIZATION FOR A SUSTAINABLE DATA STACK

Amitabh Ranjan

Rakesh Maduvinakodi Rangaswamy

Girish Kumar Gupta

DYNAMIC STACK PORT SERDES POWER UTILIZATION FOR A SUSTAINABLE DATA STACK

AUTHORS:
Amitabh Ranjan
Rakesh Maduvinakodi Rangaswamy
Girish Kumar Gupta

ABSTRACT

Multiple network switches may be stacked, one atop another, and then interconnected through each switch's two stack ports (SPs). Within such a stack arrangement, techniques are presented herein that support dynamically reducing the power utilization of a switch's SP Serializer/Deserializer (SerDes) blocks without interrupting any data traffic. Under aspects of the presented techniques, a switch may transition between different power saving modes, which may include a normal mode (encompassing powering down the SerDes blocks of both of the switch's SPs), an optimized mode (encompassing reducing the speed of the SP SerDes blocks), and a smart mode (encompassing dynamically powering up and down one of the switch's SPs based on a budgeting of the network traffic). Under further aspects of the presented techniques, the selection of a power saving mode may be based on a switch's configuration (such as a standalone arrangement, part of a half-ring topology, or part of a full-ring topology) and a switch's input traffic bandwidth.

DETAILED DESCRIPTION

A modern network switching platform (which, for simplicity, may be referred to herein as a switch) may support a stacking architecture under which multiple switches may be arranged, one atop another, and then interconnected using stack cables. Such a stacking paradigm yields a centralized architecture comprising one primary switch, one standby switch (for purposes of redundancy), and one or more member switches. Such an arrangement may support a maximum of 16 switches in a single stack.

Under a stacking arrangement as described above, each of the two stack ports (SPs) of a switch (or stack member) may be connected to other switches using stack cables. Such connections may yield either a full-ring topology (as depicted in Figure 1, below) or a half-ring topology (as shown in Figure 2, below).

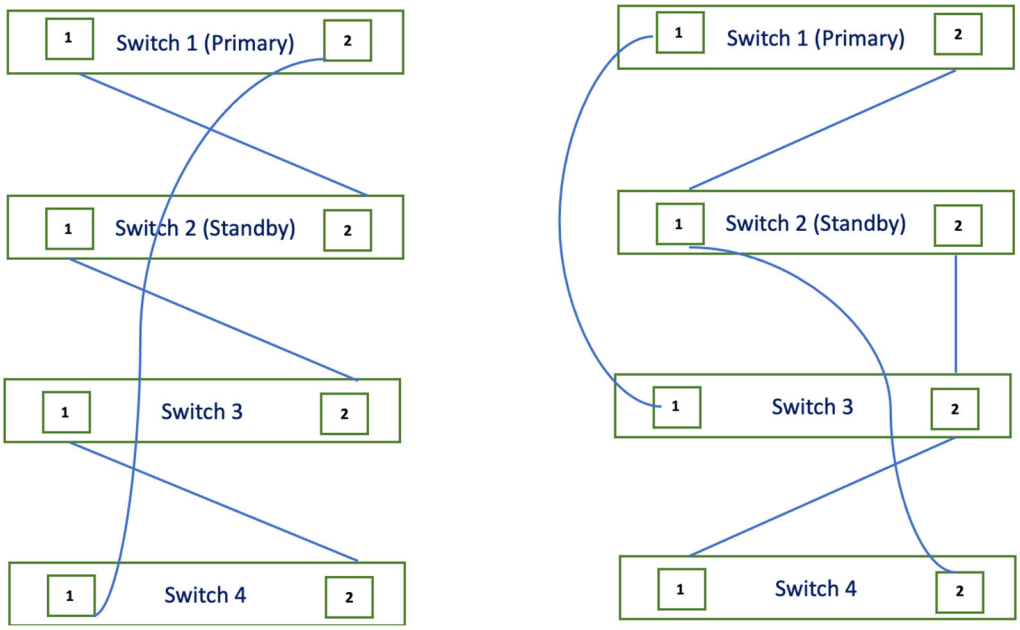

Figure 1: Four-Member Full-Ring Stack

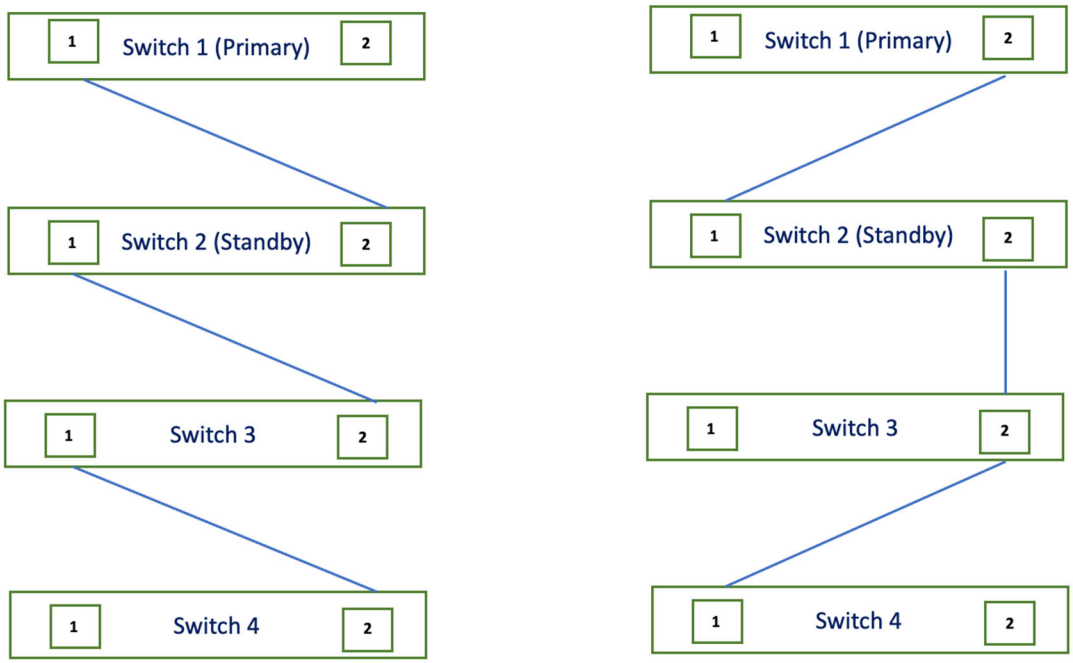

Figure 2: Four-Member Half-Ring Stack

In a stack configuration as depicted in Figures 1 and 2, both above, there are two types of traffic that may be observed transiting the SPs after the switches in such a stack formation enter a “Ready” state.

A first type comprises control traffic. Each member in a stack, including a standby switch, may periodically (e.g., every two seconds) send a keepalive message to the primary switch to monitor the health and topology of the stack. The control traffic may also encompass inter-process communication (IPC) messages that are exchanged between processes on two different switches.

A second type comprises data traffic. The incoming traffic on a switch’s network ports may be sent over the SPs if the traffic’s destination port is on a remote switch in the stack.

The total bandwidth of a switch’s SPs is equally distributed across that switch’s two SPs. A SP typically comprises one or more Serializer/Deserializer (SerDes) functional blocks, each of which is capable of supporting many gigabits (Gbs) of traffic. For example, within one representative switch that is on the market each SP consists of four 10Gb SerDes blocks combining to provide a total of 40Gb of bandwidth, as shown in Figure 3, below.

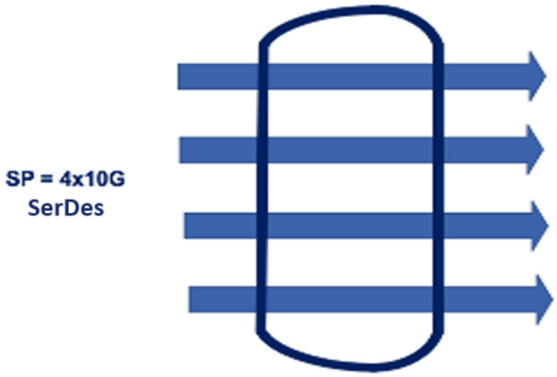

Figure 3: Switch SP Comprising Four 10Gb SerDes Blocks

Since, as described above, each stackable switch contains two SPs, the total SP bandwidth for such a switch will be 2 * 40Gb or 80Gb. It has been observed that much of there are fewer than 24 downlink ports and just a few uplink ports connected to a switch. In such cases, the incoming network traffic will be less than the bandwidth of one SP.

However, switches do not currently implement any SP power saving techniques. Under normal circumstances, both of a switch’s SPs are always enabled – when the switch

is in a standalone mode (i.e., an arrangement comprising just one switch where the switch's SPs are not connected) and when the switch is in a stack configuration where the volume of incoming traffic to the switch is less than one-half of the total switch stack bandwidth.

Techniques are presented herein that address such an inefficiency by dynamically reducing the power utilization of SP SerDes blocks without interrupting the traffic that may be flowing over the data stack. The techniques may be employed in different switch configurations including a standalone arrangement, a half-ring topology, or a full-ring topology. The presented techniques support three different power-saving modes, each of which will be described and illustrated in the narrative that follows.

Under a first, or normal, power saving mode, if a switch is being used in a standalone configuration, then the SerDes blocks of both of the switch's SPs may be powered down. Alternatively, if a switch is part of a half-ring topology, then the SerDes blocks of the switch's SP which is not being used may be powered down. Through such actions, switch power is not consumed for the unconnected or idle SPs.

Under a second, or optimized, power saving mode, when a switch comes up in a stack configuration the SP SerDes block speed is predetermined and does not change with each reboot operation. Even if the sum total of all of the incoming traffic across all of the connected network ports is less than current stack interface bandwidth, the SP SerDes block speed remains the same. In an optimized power saving mode, a user is allowed to boot a switch stack with a reduced SerDes block speed without impacting the topology.

For example, consider a switch having 48 1 Gb downlink ports, four 10Gb uplink ports, and two SPs. The maximum possible incoming traffic is 88Gb (i.e., 48Gb + 4 * 10Gb). The default stack SerDes block speed may be 40Gb for each SP, thus providing 80Gb of bandwidth across the switch's two SPs. If the bandwidth of the incoming traffic, based on the number of connected ports, is less than one-half of the total stack bandwidth, this second power saving mode allows a user to boot the switch with a reduced SP speed (at, for example, a 20Gb SerDes block speed for each SP).

Figure 4, below, presents elements of a process flow that is possible according to the second power saving mode of the presented techniques and which is reflective of the above discussion.

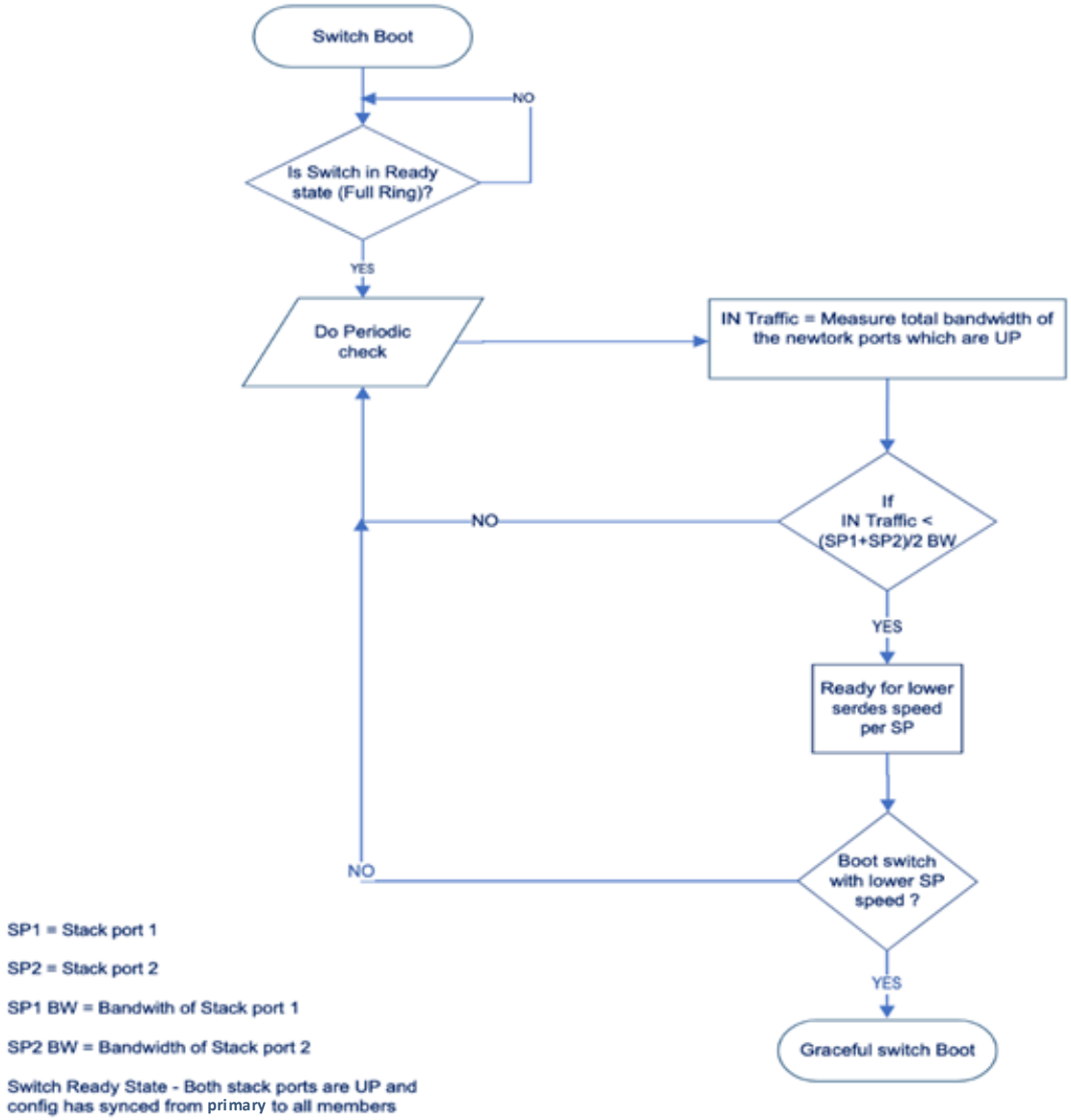

Figure 4: Exemplary Optimized Power Saving Mode Process Flow

In Figure 4, above, SP1 identifies a switch's first SP, SP2 identifies a switch's second SP, BW identifies the bandwidth of those SPs (where SP1 BW is the bandwidth of SP1 and SP2 BW is the bandwidth of SP2), and a switch being in a "Ready" state indicates that the switch's two SPs are up and configuration has synchronized from a primary switch to all of the stack members.

The advantages of this power saving mode include the switch consuming less power when a stack SerDes block is operating at a lower speed with, importantly, the maintenance of SP redundancy as both SPs are active, albeit each is working at a lower

speed. It should be noted that under this power saving mode, a switch stack would require a reload operation in order for a switch to boot at a reduced stack SerDes block speed.

Under a third, or smart, power saving mode, in the case of a full-ring stack topology one of the SPs on each switch may be dynamically powered up and down based on a budgeting of the incoming traffic and the traffic that is destined for a remote switch. According to this power saving mode, the stack may boot as normal with all of the SPs enabled. Some amount of time (e.g., five minutes) after the stack enters a "Ready" state the algorithm of this power saving mode may engage. In connection with that algorithm, Figure 5, below, presents elements of a process flow that is possible according to the third power saving mode of the presented techniques and which is reflective of the above discussion.

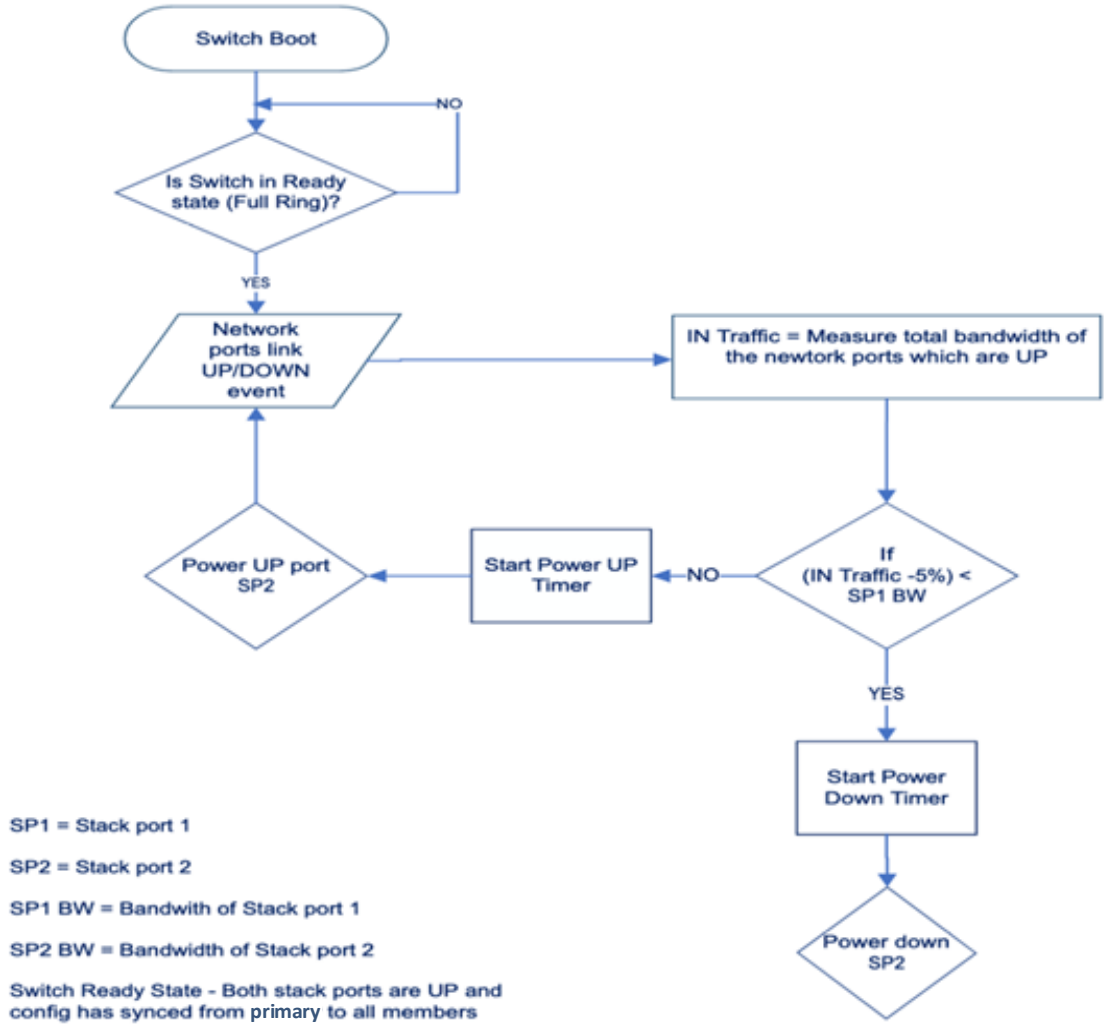

Figure 5: Exemplary Smart Power Saving Mode Process Flow

In Figure 5, above, SP1 identifies a switch's first SP, SP2 identifies a switch's second SP, SP1 BW is the bandwidth of SP1, and a switch being in a "Ready" state indicates that the switch's two SPs are up, and configuration has synchronized from a primary switch to all of the stack members.

As shown in Figure 5, above, the software of this mode's algorithm may calculate the bandwidth of all of the network ports that are up based on their maximum line rate. If the maximum incoming traffic is less than the bandwidth of a SP, then all of the SerDes blocks of one SP may be powered down. This event may take place every time that a network port comes up or goes down. Additionally, such a powering down and up of a SP may be preceded by a deactivate and activate timer, both of which may be configured by a user.

The advantages of this power saving mode include keeping both of the SPs enabled (i.e., powered on) only as needed, with one SP of each stack member powered down at other times. In such a state, the stack functions in a half-ring topology but saves a significant amount of power over time. By budgeting the incoming and outgoing stack traffic, the second SP may be dynamically enabled without any noticeable delay.

Since a SP is a bundle of two or more separate SerDes block lanes, the power savings may be enhanced by disabling the unused SerDes block lanes of a SP if the outgoing traffic over the SP is small enough to fit in one, or possibly more, SerDes blocks but not necessarily require all of the blocks.

Under this power saving mode, the switch stack topology changes from a full-ring to a half-ring. In the event that a SP goes down, the software will, according to this mode, dynamically bring up the powered-down SPs. Further, the switches in a stack may split and then merge to again form a stack. However, those activities may delay the realization of redundancy.

The benefits that may arise from use of the presented techniques may be demonstrated through an illustrative example. Under that example, the technique's three different power saving modes (as described and illustrated above) were applied in a controlled setting to a representative switch under different configurations. As will be discussed below, a power savings of up to 5.5 watts (W) per switch was realized. Such a savings, through application of the presented techniques, would represent an operating

expense reduction, provide a market differentiator for a switch vendor, and aid a switch vendor in achieving its sustainability goals.

To establish an initial baseline, the switch's power consumption with no power saving mode enabled was found to be 22,000 milliwatts (mW) or, equivalently, 22W.

With the switch operating in a standalone mode, the first (e.g., normal) power saving mode was applied. As described above, under this mode all of the switch's SP SerDes blocks were powered down. The switch's power consumption was then found to be 16,500mW, for a power savings (through the first mode) of 22,000mW – 16,500mW = 5,500mW or, equivalently, 5.5W.

Next, the second (i.e., optimized) power saving mode was applied to the switch. Under this mode, the stack SerDes block speed was reduced to 20Gb per SP by powering down two out of the four SerDes blocks on each SP or reducing the speed of each SerDes lane to half. The switch's power consumption was then found to be 18,000mW, for a power savings (through the second mode) of 22,000mW – 18,000mW = 4,000mW or, equivalently, 4W.

Finally, with the switch part of a three-member stack the third (i.e., smart) power saving mode was applied. Under this mode, one of the switch's SP was powered down since there were just two network ports connected. The switch's power consumption was then found to be 19,250mW, for a power savings (through the third mode) of 22,000mW – 19,250mW = 2,750mW or, equivalently, 2.75W.

As demonstrated by the above illustrative example, since a SerDes block operates at a high speed, powering it down when it is not in use (through, for example, the first or third power saving modes) significantly reduces a switch's power consumption. Further, even when a SP SerDes block is working at a lower than maximum speed (through, for example, the second power saving mode) there is also a significant power saving.

While the above discussion focused on SPs and SerDes blocks, it is important to note that the concepts behind the presented techniques (including, for example, the dynamic allocation of a component – such as a SerDes block – based on bandwidth) may also be applied to the network ports of a switch.

It is also important to note that redundancy and failure recovery are an integral part of the presented techniques. The chief element of the second (i.e., optimized) power saving

mode is the maintenance of a full-ring stack topology, even at a lower stack speed, while consuming less switch power. Thus, redundancy is maintained. In the case of the third (i.e., smart) power saving mode, the SerDes blocks of one SP are powered down based on input traffic bandwidth. As a result of such an action, the stack will change to a half-ring topology and transition to a power saving mode. If a failover of the active stack link is subsequently detected, the other SPs may be quickly brought up. While the stack may split for a few seconds, all of the members of the stack will be restored in the stack as earlier. Such a small glitch may be considered to be a cost for maintaining a sustainable data stack.

In summary, techniques have been presented herein that support dynamically reducing the power utilization of a switch's SP SerDes blocks without interrupting any data traffic. Under aspects of the presented techniques, a switch may transition between different power saving modes, which may include a normal mode (encompassing powering down the SerDes blocks of both of the switch's SPs), an optimized mode (encompassing reducing the speed of the SP SerDes blocks), and a smart mode (encompassing dynamically powering up and down one of the switch's SPs based on a budgeting of the network traffic). Under further aspects of the presented techniques, the selection of a power saving mode may be based on a switch's configuration (such as a standalone arrangement, part of a half-ring topology, or part of a full-ring topology) and a switch's input traffic bandwidth.