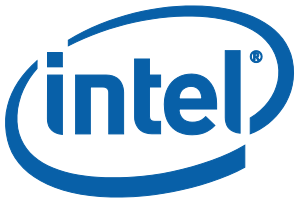

Maintaining the Ethernet Link to the BMC During Server Power Actions

Using the Advanced Manageability Feature of Intel Ethernet Controllers

Intel® LAN Access Division

Revision 1.0

October 2012

Legal

Revisions

Date	Revision	Description
October 2012	1.0	Initial Release

Contents

1 BMC Ethernet Connectivity

In most modern servers a Baseboard Management Controller (BMC) is responsible for monitoring things such as temperature, voltages and fans. The BMC can usually be configured to send out alerts (in the form of emails and/or SNMP alerts) when something in the system requires the attention of an administrator (such as the CPU starting to overheat).

The BMC usually has network connectivity via the sideband interface on some kind of LAN On Motherboard (LOM). The most common sideband interface is Network Controller Sideband Interface (NC-SI) defined by the DMTF and available on many Intel as well as other vendor's Ethernet Controllers. The NC-SI interface provides a mechanism by which the BMC can share Ethernet connectivity with the operating system.

The BMC can use this Ethernet connectivity to enable a remote management application to establish a connection. Over this connection a management application can do such things as view the temperature, voltage and fan sensors along with a great many other tasks. One popular usage for this connection is Serial over LAN (SoL) by which one can reboot a server and do tasks such as BIOS configuration. A more advanced feature is Keyboard, Mouse & Video (KVM) redirection, which enables the BMC to provide a remote desktop type of capability independent of the operating system.

1.1 The Physical Layer of the OSI Model

The Open System Interconnect (OSI) model has 7 layers to it. Layer 1 is the physical layer; this is usually what Intel refers to as a PHY on an Ethernet device. The PHY is the port that the Ethernet cable plugs into. On many PHY's you'll find LED's that indicate that there is power, as well as a link state. There are usually different colors for different link speeds such as 10 Mbps, 100 Mbps and 1 Gbps; if the link LED is off, it indicates no link.

1.2 Importance of Link When Server is Powered Off

As the BMC is used for remotely managing a server, which includes powering it on and off as well as rebooting it when the operating system hangs, it is important that the BMC have network connectivity even when the computer is operating only on standby power (basically plugged into the wall, but powered down).

In order to achieve this, the LOM must also be designed to work on standby power. This is a basic requirement for server manageability – the BMC and the LOM must be connected to standby power in order to manage a server that is powered down.

1.3 Times When the Link Goes Down

There is a time during normal operations of a server that the link can go down and then back up. When a server is in a powered off state and is powered on, the PE_RST_N (PERST#) signal is usually asserted, causing the Ethernet controller to reset and re-initialize. This action of resetting the Ethernet controller causes the link to momentarily be lost (usually less than a second).

If the software device driver (or BMC via NC-SI) renegotiates the link speed, the link also goes down momentarily.

1.4 BMC Connectivity During Link Down/Up Times

There are times when the network link momentarily goes down and then back up. If this is a short enough time period (a couple of seconds or so) this usually does not cause any major issues even if a management console is connected to the BMC. This is because many times the network traffic going to the BMC is UDP and the IPMI protocol has a retry mechanism defined.

In many cases, even if the traffic is TCP/IP based (say for example for a KVM session), it too will have a retry mechanism that would be tolerant of such a momentary loss of connectivity.

2 Sometimes a Quick Link Down/Up Causes Interruption in Connectivity for Minutes

There is a case where even if the link goes down and then back up within a second, a remote application is not able to communicate with the BMC over the network for up to several minutes. This occurs when the Spanning Tree Protocol (STP) is used within the switches between the management console and the BMC.

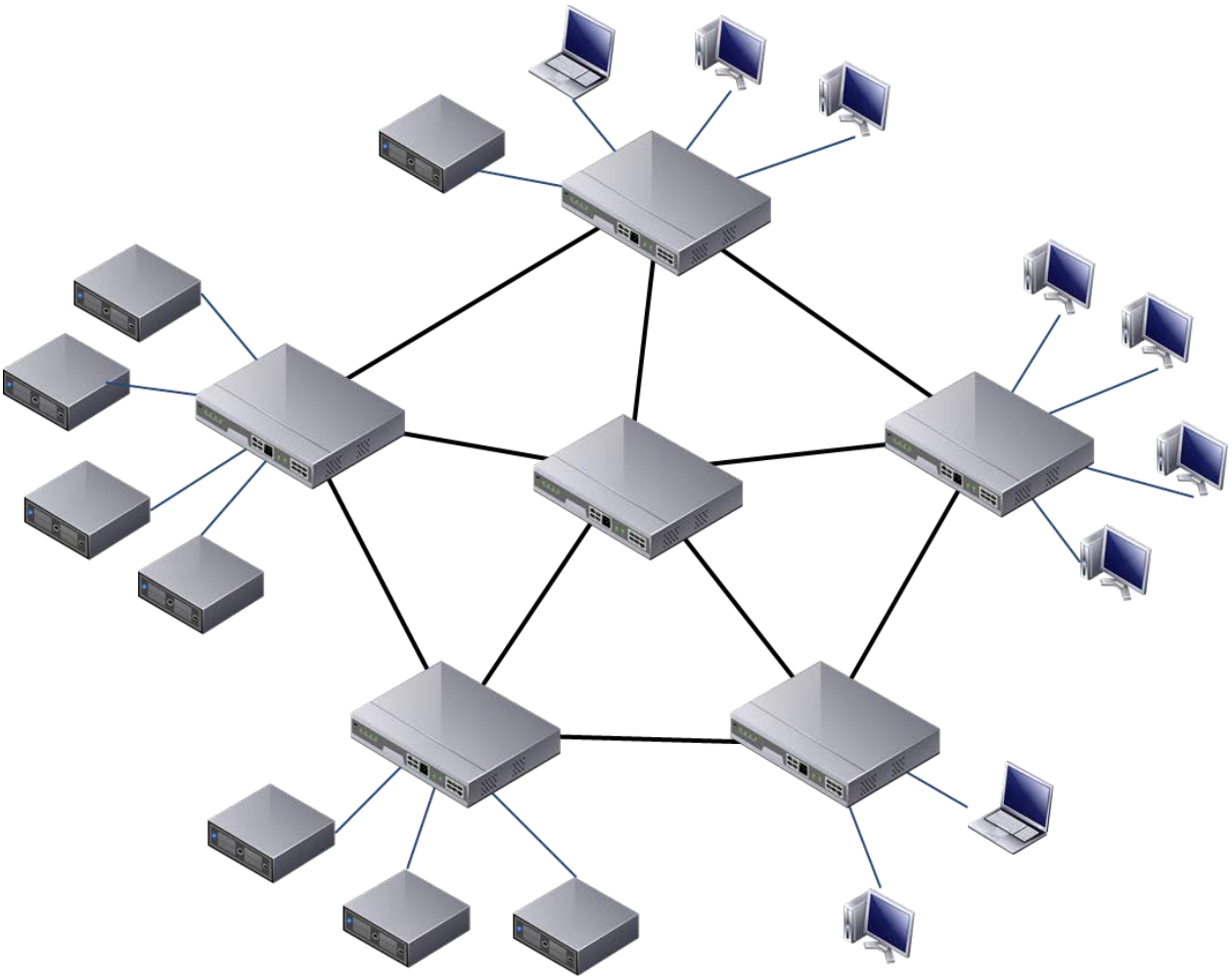

Figure 1 Interconnected Switches

With spanning tree, many switches are interconnected that among other things provides redundancy. Spanning tree makes it so that a packet does not get repeated – meaning only one switch passes the directed packet along.

When a switch detects a link down/link up condition on a port, it reactivates the STP, as it doesn't know which topology reconnected after the link was re-established and whether new loops where created. It might take minutes until the link is moved by the STP to active state again. During this time the BMC is not able to communicate with the remote management console.

The end result is that if you are wanting to reboot your server to use SoL in order to go into BIOS and configure something – you won't be able to do so because by the time the STP re-establishes connectivity to the BMC, the BIOS setup option has likely long since passed and the server running the operating system.

3 The Solution – the Critical Session Feature

Having recognized this issue many years ago, Intel Ethernet controllers have a feature Intel calls a critical session by which the BMC can tell the Intel Ethernet controller to NOT reset the device and to keep that PHY connection steady just as it is.

Using this feature, the BMC can configure the critical session via the sideband interface and prevent the loss of link and the potential loss of connectivity.

3.1 Related Technology – Low Power Link Up

Wake on LAN (WoL) is another technology that requires the Ethernet device and PHY to be powered when the server is running on standby power. It enables you to power down your server until needed and then have it power on after your Ethernet device receives a WoL packet.

One usually powers down a server that is not needed and has it enabled for WoL in order to save on power usage – no need to have the server fully powered until it is needed. So if your Ethernet controller is a 1 Gbps device such as the Intel® Ethernet Controller I350, many would consider it a waste of power for the Ethernet device to maintain a 1 Gbps link to the switch. Note that a 10 Mbps or 100 Mbps connection works just as well for WoL, yet requires a lot less power.

Intel Ethernet controllers have the Low Power Link Up (LPLU) feature that enables the link to be renegotiated down to a lower speed when the server is running on standby power. So what this means is, if you power off your server, the Intel Ethernet controller (if so configured) attempts to negotiate the link speed down to the lowest possible link speed in order to save power.

3.2 How to Use the Critical Session Feature

This feature requires a two-step process. It must be first enabled in the NVM/EEPROM of the Intel Ethernet controller (which is the default), and then the BMC must send a command to the Intel Ethernet controller over the sideband interface.

This feature is supported over both the NC-SI interface using an Intel OEM NC-SI command (Set Intel Management Control Command) and over the SMBus interface. Refer to product specific documentation (http://www.intel.com/products/ethernet/resource.htm) for details.

3.3 Potential Side Effects of the Critical Session Feature

Once this feature is enabled, the link state does not change (unless of course the link partner, such as the switch changes), the software device driver cannot change the link speed and if the link partner attempts to renegotiate the link speed, the only option available is that of the current speed.

3.3.1 Excess Power Usage Link Speed Stays High

Let's take an example where the server is fully powered and the operating system is up and running and has negotiated the link of the Intel Ethernet port to 1 Gbps. For some reason the BMC gets reset (perhaps a firmware update occurred). The BMC might, as part of its configuration setup the Intel critical session feature. All is good right?

Well, maybe not, depends on your desired usage. If the server is then powered down, the link speed remains at 1 Gbps – drawing more power. The Low Power Link Up (LPLU) feature is overridden when the critical session feature is active – it does not renegotiate to a lower link speed to save power.

This means that your server might now be drawing more power than you want in a powered down state.

3.3.2 Reduced Link Speed in the Operating System

Let's imagine that the Intel Ethernet controller you are using (say the upcoming Intel® Ethernet Controller I210) has the Low Power Link Up (LPLU) enabled. The server at first has no power to it and then it is plugged in. The BMC goes and gets itself all initialized, while at the same time the Intel® Ethernet Controller I210, using the Low Power Link Up (LPLU) feature, negotiates a nice power saving link speed of 10 Mbps. The BMC now configures the critical session feature.

Sometime later the server is powered on (by pressing the power button, via the BMC or even by WoL) and booted to the operating system. Since the critical session feature is configured, when the software device driver in the operating system goes to initialize the port on the Intel® Ethernet Controller I350, it recognizes that the critical session feature is active and leaves the link as it is, at 10 Mbps.

So now you have your 1 Gbps Ethernet device only running at 10 Mbps, at 1% of potential speed.

3.4 Suggested Use of the Critical Session Feature

While the critical session feature available in the Intel Ethernet controllers prevents the link from dropping during power ups, there are some potential side effects if not used with caution and wisdom.

In most cases, this feature should be only activated when the BMC has an active session with a remote management application. When there is no active session, the feature should

be deactivated.

The following flow charts provide the basic algorithms.

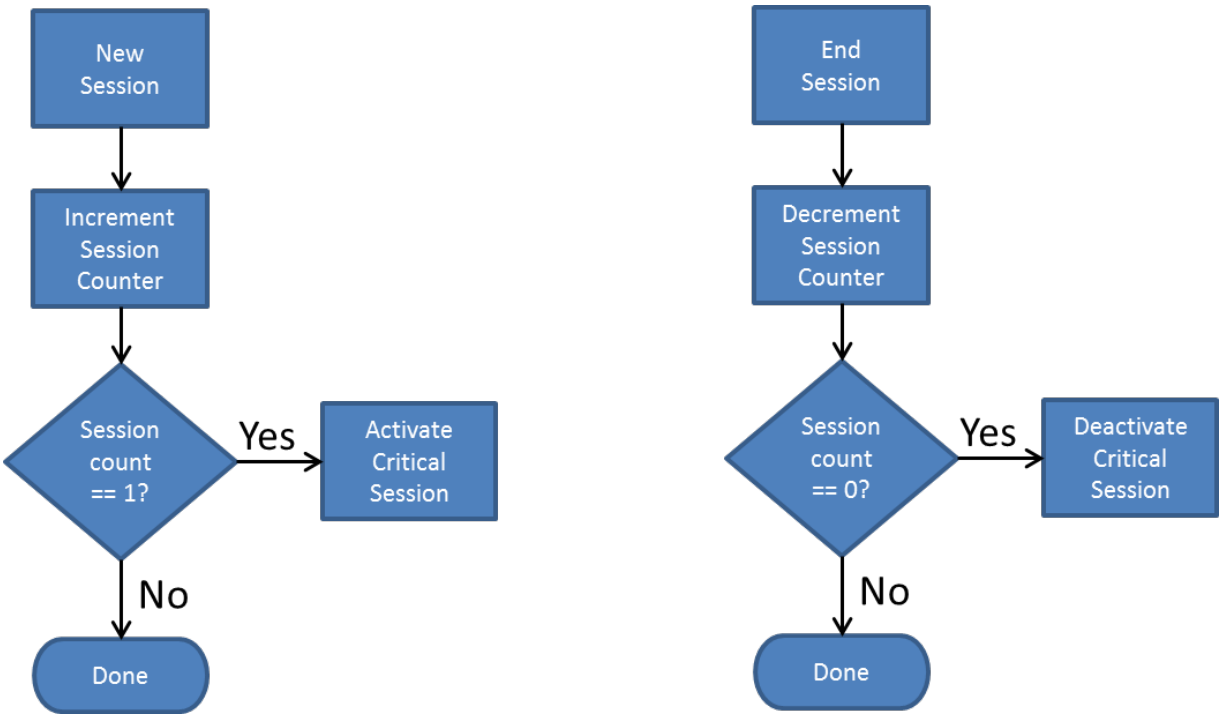

Figure 2 Critical Session Usage Algorithms

Only the BMC knows when there is an active connection with a remote management console, the Ethernet controller has no specific knowledge of when a session begins or ends. For this reason, the BMC should specifically enable and disable this feature as needed.

3.5 Lifetime of the Critical Session Feature

Intel recognized that some BMC's might simply configure the critical session feature and then forget to turn it off. As such, on many of the Intel Ethernet controllers that support this manageability feature, it is disabled after a reset has occurred.

Technically speaking, when the Main Power OK (MAIN_POWER_OK) pin is negated, the critical session feature is turned off and the BMC must re-enable (re-arm) it. In this way, the BMC must explicitly re-arm the feature and the potential side effects does not occur.

The MAIN_POWER_OK pin does not exist on the cost-optimized Intel Ethernet controllers such as the Intel® 82574 GbE Controller and the Intel® Ethernet Controller I210. For these devices, the BMC must explicitly enable and disable the feature.

4 Not All Customer Designs are the Same

Server manufactures design different servers for different environments. Some servers using Intel 1 Gbps devices require that the link always be 1 Gbps and never negotiate down to a lower link speed to save power.

Sometimes the BMC designers want to enable the critical session feature and have it stay enabled until a complete loss of power.

The following are a series of recipe' that should cover most of the desired usage models.

4.1 Enable Critical Session Only Once

In this situation, the BMC wants to enable the feature and never have to re-arm it. On the Intel® 82574 GbE Controller and the Intel® Ethernet Controller I210, this is the default behavior. On other devices this feature is un-armed after the negation of the MAIN_POWER_OK pin.

In order to make it so that the BMC can arm the critical session feature and the forget about it, the board designers must make sure that the MAIN_POWER_OK pin does not get asserted. The recommended method to do this is to tie the MAIN_POWER_OK pin to the AUX POWER pin.

The only activity the MAIN_POWER_OK pin does is to signal to the Intel Ethernet controller that it should reset the critical session feature, as such by tying it to AUX POWER, the feature is never reset unless AUX POWER is lost.

4.2 Always Link at 1 Gbps

There are a couple of ways to achieve this. First that if the BMC is using NC-SI to connect to the Intel Ethernet controller, it can always use the NC-SI commands to set the link speed as it sees fit.

The other way is to, in the NVM/EEPROM of the Intel Ethernet controller, configure it so that Low Power Link Up (LPLU) links only at 1 Gbps. Talk to your Intel account representative on how to obtain an NVM/EEPROM image configured in such a way.

4.3 Session Began When Server was Powered Down and at Low Link Speed

If Low Power Link Up (LPLU) is configured to save power and has negotiated the link to a low link speed and then the BMC session is established and the BMC, as suggested, enables the critical session feature and at a later time, while the BMC session is still active, the server is powered on and boots in the operating system, the link remains at the low link speed as described in Section 3.3.2.

Since the software device driver only tries to autonomously negotiate the link speed when the device is initialized, there are only a couple options available. After the BMC session has ended, it could either reboot the server again, and when the software device driver in the operating system is loaded it negotiates the link to the desired speed.

Another option is to have the BMC (if it is using NC-SI) renegotiate the link on behalf of the operating system using the NC-SI Set Link command. The drawback of this solution is that if for some reason the operating system were using the Ethernet port, it would lose connectivity briefly if not using STP on the switches or up to minutes if it was.

5 Summary

The critical session feature is available in many Intel Ethernet server controllers. This feature can keep the BMC from losing connectivity, in particular when STP is being used within the switches upstream from the server.

To make the most out of the critical session feature, the BMC should be an active participant in the usage of this feature, enabling it when there is an active remote management connection to the BMC and disabling it when finished with the session.

6 Additional Resources

Intel Ethernet Controller documentation

http://www.intel.com/products/ethernet/resource.htm

Manageability – Intel® Sideband Technology

http://download.intel.com/design/network/applnots/321786.pdf

NC-SI specification

http://dmtf.org/