OS-to-BMC Pass-through: A New Chapter in System Manageability

A Dell Technical White Paper

Brian Doty

Kalyani Khobragade

Rich Hernandez

Nick Harris

Narayanan Subramaniam

Pankaj Gupta

Contents

This document is for informational purposes only and may contain typographical errors and technical inaccuracies. The content is provided as is, without express or implied warranties of any kind.

February 2012| Rev 1.0

Abstract

This document describes the rationale and operation of the OS-BMC communication link available on the latest generation of Dell™ PowerEdge™ servers. The baseboard management controller (BMC) is a specialized service processor that monitors the physical state of a server. The BMC, running specialized firmware, is responsible for monitoring the health of server components, maintaining thermal characteristics, and aiding in remote access and recovery. Dell's BMC has evolved through several generations, expanding its monitoring capabilities and remote access and adding comprehensive remote management through the Lifecycle Controller. Dell's current generation of BMC is known as the Integrated Dell Remote Access Controller 7 (iDRAC7). The iDRAC's duties do not end at managing direct server management; the Dell OpenManage™ portfolio and other management consoles use iDRAC and Lifecycle Controller to simplify monitoring, deployment and updates.

The OS-to-BMC pass-through is another example of Dell's investment in iDRAC's management capabilities, and in this case, fabric convergence. Dell supports network I/O convergence in numerous ways, such as FCoE and iSCSI. In addition, OS-to-BMC is another way to converge iDRAC traffic over the Ethernet network controller by leveraging the LOM/NDC NC-SI communication channel. This provides many powerful capabilities for iDRAC and the management consoles that build on it.

The current paradigm

With the iDRAC7, the chipset is located on the server mainboard but is logically separate from the server platform and powered even when the server is off. Because of this, iDRAC7 is able to interrogate the network interface card (NIC) for configuration information, display it to an operator, and accept changes, all while in a pre-boot environment. These changes are subsequently stored in non-volatile memory until a boot is initiated and it is loaded into the NIC.

There are two connectivity paths to the iDRAC. The first path is a remote connection over the LAN, where security protocols such as SSL are in place. The LAN interface allows a user to access iDRAC remotely, even over WAN links, directly accessing the iDRAC IP address. The OS interface uses the IPMI protocol over the in-band KCS channel.

The second connectivity path is the KCS interface. The KCS OS-to-iDRAC link is based on an industry standard interface called Intelligent Platform Management Interface (IPMI[1]). The

Dell version of IPMI supports two interfaces, one connected via LAN (IPMI over LAN) and one connected to the host server operating system via the Keyboard Controller Style (KCS) interface.

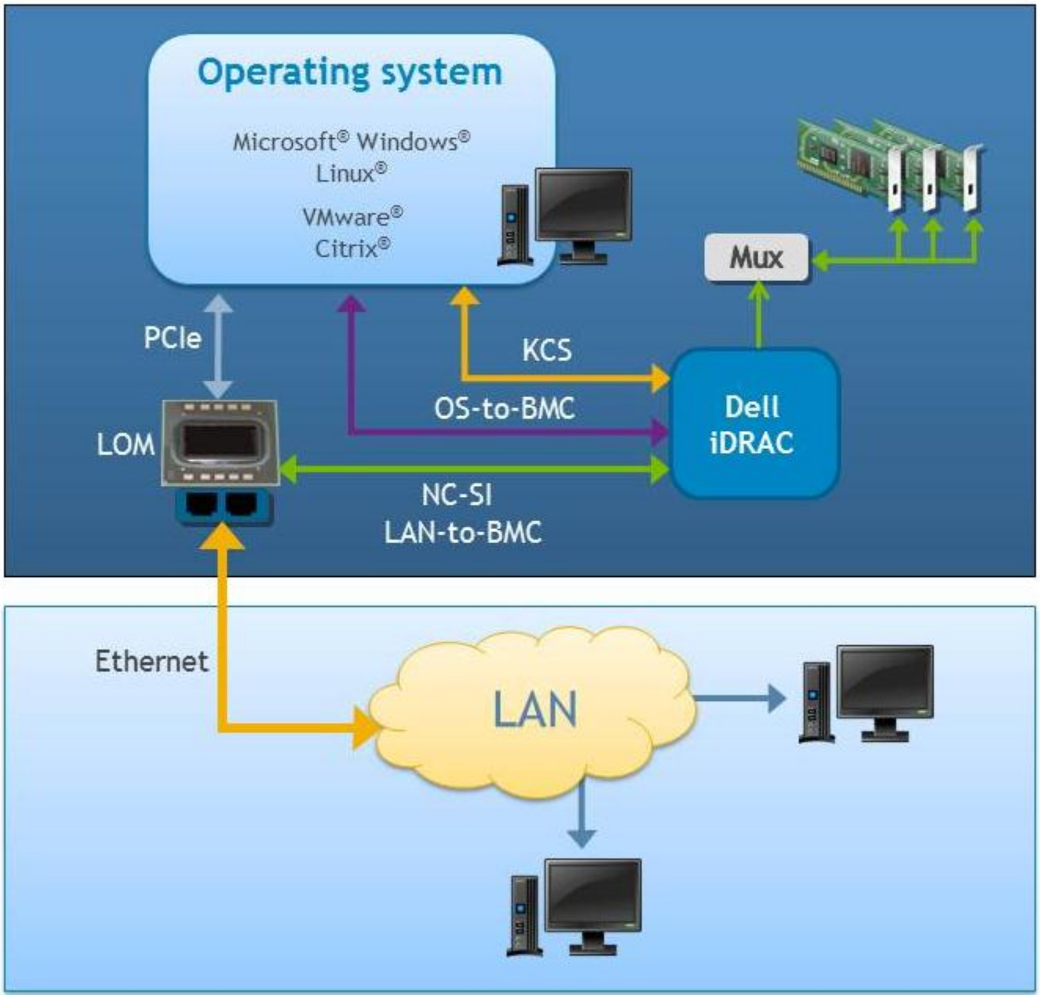
The KCS interface provides unauthenticated local access to iDRAC but requires third-party software agents running on the active OS system. The problem with this arrangement is two-fold: one being the presence of third party agents running on the OS and two, the KCS channel is not very robust. KCS is a relatively slow interface, using dual byte-aligned I/O ports, and is suitable for simple functions such as temperature monitoring, fan speed, or power status. The iDRAC uses this interface to gather information about the operating system and its current state. However, more advanced functions require greater bandwidth and programmatic flexibility than currently offered by KCS. For example KCS does not have the bandwidth to access the iDRAC Web GUI directly on the host.

[1] http://www.intel.com/design/servers/ipmi

The iDRAC7 solution

Dell saw opportunity in this problem paradigm and drove a solution to provide a robust link between the OS and iDRAC7. This was done by replacing the KCS channel with a secure and high-speed management network using the onboard network adapter. Since the eighth-generation of Dell PowerEdge servers, iDRAC has been able to use the embedded NIC for network access. This configuration is called shared mode. This mode allows you to minimize network cabling by providing a path for iDRAC to share bandwidth with the onboard embedded NIC. The OS-to-BMC pass-through builds on this interface. In this mode, the normal data path between OS and LOM/NDC is configured to redirect management frames to the iDRAC. This technology takes advantage of the high-bandwidth standard network stack in the OS and provides a much more robust and standardized link for management traffic. Also, the hardware design supports selecting any of the LOM/NDCs ports that have a path to iDRAC. For those systems with more than one LOM, the solution provides maximum flexibility by supporting any of the LOMs as an OS-to-iDRAC communications channel. Only one LOM can be configured as the OS-to-BMC channel at a time, unless the LOMs are teamed. In this case the pass-through capability is enabled on all ports. All teamed ports must support OS-to-iDRAC pass-through and shared for iDRAC connectivity to be enabled. The communication channel between the LOM and iDRAC is the same NC-SI channel used for network to iDRAC communication. This is illustrated in Figure 1.

Figure 1. Dell server communication diagram



Uses and benefits of the OS-to-BMC

OS-to-BMC pass-through operation requires an active link on the network side of the shared LAN controller. This allows the network OS stack to use the LAN controller interface for communications. The simplest advantage that this provides is to allow a user administrator seated at the server to bring up a Web browser session to interface with the iDRAC in that server. There are several other use cases described as follows:

- The OS-to-BMC pass-through operation allows in-band WSMAN access to the iDRAC CIMOM and other web services as defined by DMTF.
- It provides in-band OS access to iDRAC web browser for all system management functions. This allows customers to avoid using a remote management client for local tasks.
- The tools used for in-band and out-of-band operation are consistent, as there is no need to invest in special tools that depend on the slow KCS interface.
- It supports Dell's goals of an agent-free environment by providing access to standard operating system management protocols.
- It provides a fast and reliable communication path for the iDRAC to communicate status about server health and operation to the host operating system. It is not possible, for example, to use KCS to communicate complex power consumption statistics to the host OS.
- The standard LAN interface to the iDRAC includes security not present when using KCS. Using the OS-to-BMC technology, even a local user will be required to provide the appropriate authentication and security to communicate with the iDRAC.

A deeper look at the Dell solution

As mentioned earlier, iDRAC supports a shared mode in the LOM that serves as the basis for the OS-to-BMC technology. When the LOM is configured in shared mode, it provides L2 filters for the following traffic flows based on destination MAC Address and optionally VLAN ID:

- OS-to-iDRAC
- iDRAC-to-OS
- Network-to-iDRAC
- iDRAC-to-network

All forwarding decisions are based on L2 filtering (destination MAC and VLAN). NC-SI automatically filters ARP, DHCP client, DHCP server, and NetBIOS broadcast packets. It also has specific filters for neighbor discovery and router discovery multicasts. For the outbound host traffic, NC-SI filter settings are used for broadcast and multicast traffic. The same set of multicast and broadcast filters can be used for both in-bound network and out-bound host traffic to determine whether the packet gets forwarded to the BMC or not.

For each of the packet-filtering rules, the NIC management firmware independently configures whether the traffic is forwarded to the iDRAC or not. The execution of the following NC-SI commands by the NIC management firmware results in forwarding the packet to iDRAC for the packet-filtering rules:

- Enable VLAN and set VLAN filter
- Set MAC address
- Enable broadcast filter
- Enable global multicast filter

OS-initiated traffic

Network I/O traffic from the host OS handled by an interface enabled for OS-to-BMC pass-through is treated as follows:

- The shared LOM will filter traffic coming from the host.
- If the destination MAC address matches the iDRAC MAC address, it will forward the packet to the NC-SI side band interface for iDRAC processing. Otherwise, it would send it out on the wire to the external network.
- VLAN ID check can also be performed if provisioned
- Broadcast and multicast packets should be sent on the side-band NC-SI channel and on the wire.

iDRAC-originated traffic

Network I/O traffic from the iDRAC handled by an interface enabled for OS-to-BMC pass-through is treated as follows:

- The shared LOM will filter traffic coming from the iDRAC.
- If the destination MAC address matches the OS side host MAC address known to the management firmware, it will forward the packet to the host OS. Otherwise, it would send it out on the wire. Broadcast packets are sent to the host OS and on the wire.
- If the shared LOM is configured in the promiscuous mode, then all outbound BMC traffic shall be forwarded to the host.

Network configuration

The host and iDRAC can be on the same subnet, and therefore routing is not required. This supports a dedicated (static or DHCP) iDRAC IP address for the OS-BMC channel interface unique per system and user configured to be on the same subnet as the host IP address.

Alternatively, the host and iDRAC can be on separate subnets where routing will be required thorough the configuration of a default gateway. In this case, the host OS and iDRAC can communicate even if they are on separate VLANs or subnets because packets would go all the way to the router and then back. Since, iDRAC and the host are on separate VLANs, they would be on separate IP subnets, and the OS routing table would guide the IP layer to use default gateway to route OS-to-iDRAC packets.

As with standard L3 routing, the destination MAC address used in these packets originating from host would be router or default gateway MAC and not the iDRAC MAC. Similarly, packets originating from iDRAC would have the router MAC and not the host MAC address.

These packets would go all the way to a router and get routed to the iDRAC OS-BMC VLAN/subnet. In both cases, the network interface link must be up since the OS-iDRAC channel is down when the Ethernet link is down.

How to use OS-BMC

You can enable OS-BMC PT through the remote access controller admin (RACADM) CLI, WSMAN, or IPMI command, and it can be enabled at the iDRAC, which will automatically enable the LOM as well. The OS-BMC is not yet available with iDRAC's dedicated NIC mode. Once enabled, you can simply launch a browser in the host OS (or guest OS) and connect to the iDRAC GUI.

Conclusion

The latest generation of Dell PowerEdge server and adapter technology is based on the foundation of making server operation and management easier and more effective. The transition from 1 Gb to 10 Gb and beyond is symbiotic with the Dell's current focus on consolidation and convergence. The OS-to-BMC pass-through feature in Dell LOMs and NDCs is an example of these themes in action, which provides you with higher performance and enhanced usability.