iDRAC7 Networking and Improved Remote Access Performance

This paper describes how TCP throughput impacts the performance of virtual media, and other remote access features, in various iDRAC7 networking configurations.

Andy Butcher and Tim Lambert

Enterprise Solutions Group

March 2012| Rev 2.0

Contents

Executive summary

There are variables that can affect the performance of iDRAC7 remote management features. This paper will explore some of those variables and help Dell customers achieve optimal performance from their iDRAC7. Understanding networking protocols such as TCP and the iDRAC7's position in the network aids customers in this process.

As an example of a systems management application, Virtual Media is a licensed feature for Rack and Tower systems, and it is available by default on blade servers. This feature allows an administrator to mount a "virtual drive" to a remotely-located server from a CD, DVD, USB drive, or an ISO image file on his local machine. Together with the iDRAC7 virtual console for remote desktop, the administrator can copy files, install drivers, applications, or operating systems to a server in a different location.

Introduction

In the Dell™ PowerEdge™ 12th generation servers, the iDRAC7 processor is significantly faster than previous generations. The CPU clock speed is more than double the 11th generation speed (576MHz vs. 220 MHz), and the SoC (System on a Chip) also features a dual-execution pipeline giving further gains in processing capacity; as a result, throughput of a networking intensive application such as Virtual Media running on iDRAC7 is much greater than before. In some cases, networking devices and protocols have a more noticeable impact on the performance than the core CPU speed on iDRAC7.

Virtual Media is the example used in this paper as a "networking-intensive application," because its data rate is important to the user, and it could be affected by the network and iDRAC7 configuration. To implement virtual media, the iDRAC7 emulates a USB mass storage device interfacing with the server's main processor. Over the management network, the server's iDRAC7 issues commands to the administrator's client machine that are translated from the SCSI commands from the main processor and used to read the storage device; in return, the client machine returns data to the iDRAC7. TCP protocol is used to provide a mechanism for any lost packets to be retransmitted.

There are several interfaces for Virtual Media that could represent a choke point: the USB link to the main processor, the iDRAC7 network interface, the server network interface to the external switch, any switch interface along the way, and the client NIC interface to its accompanying switch.

iDRAC7 Networking

Overview

The Dell PowerEdge 12th generation server products offer several enhancements to the iDRAC7 server management networking options; these enhancements provide improved connectivity, flexibility, performance and resilience options, as well as reduced cabling and switch port management effort. Figure 1 illustrates three of the possibilities for the iDRAC7 network interface; they are configured with F10 during server boot, or with the iDRAC7 Web interface, or with the RACADM command line utility:

- The dedicated management port, which is enabled with the iDRAC7 Enterprise license, has improved to support gigabit Ethernet.
- A shared network port provides the capability to directly manage the server using the iDRAC7 through ports on the LAN-on-Motherboard (LOM).
- A dedicated management port for the iDRAC7 is used for blade servers. It operates at 100Mbps for compatibility with the previous generation chassis.

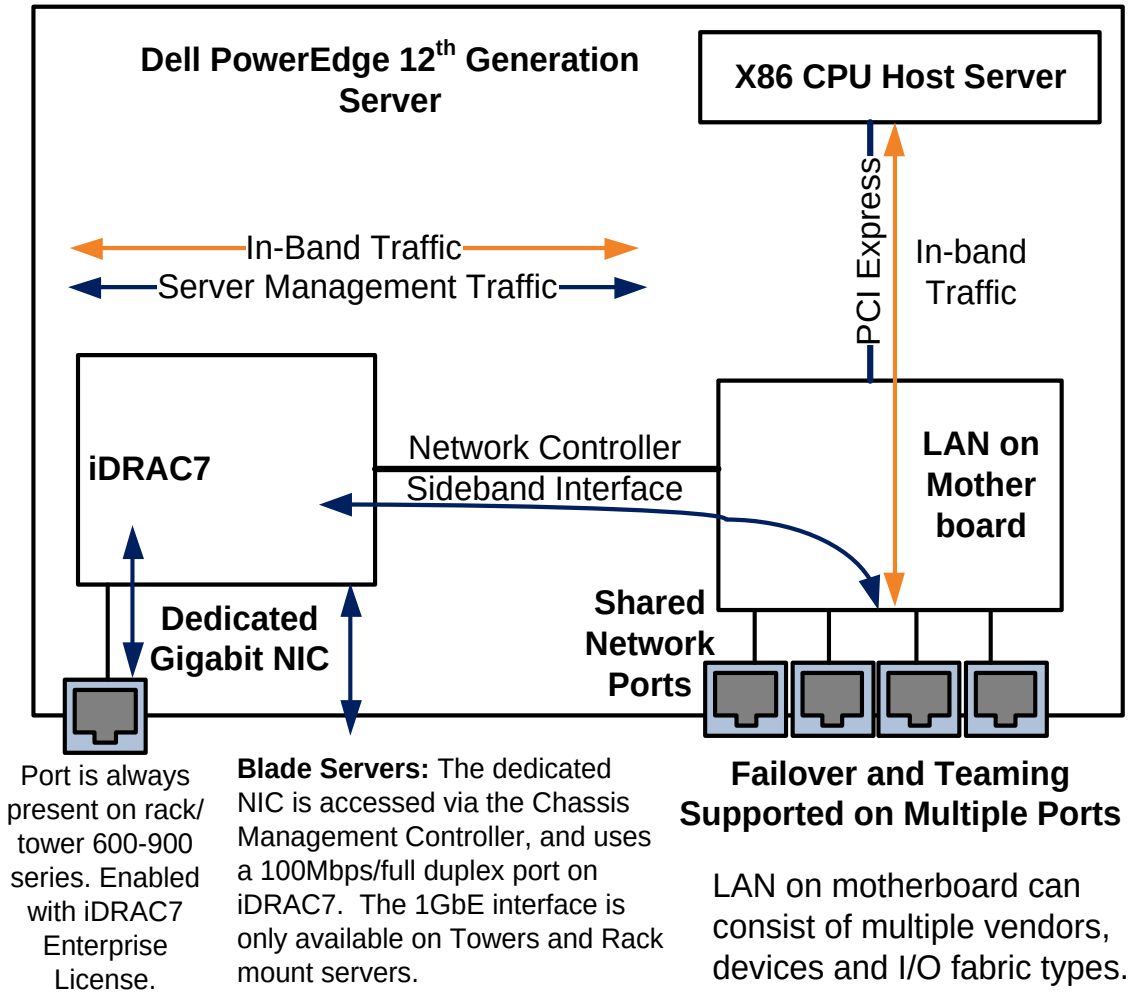

Figure 1: Server Management Network Options

Dedicated Network Interface

Dell PowerEdge servers offer a dedicated management Ethernet port for communicating with iDRAC7 to provide various server management functions. The dedicated NIC is useful to customers who commonly provide physical network isolation between their management network and other networks such as those utilizing the main system LOMs (LAN-on-motherboard).

The 12th generation server 600 models and above, such as the R620 and R720, always include the iDRAC7 dedicated management port hardware. In previous generations, such as on the R610 and R710, the dedicated NIC port was an optional iDRAC6 Enterprise module. On the R620 and R720 servers, the iDRAC7 Enterprise software license, available at the time of sale or later, enables the dedicated management port functionality, along with additional features such as Virtual KVM and Virtual Media.

Additionally, the dedicated NIC increases from 10Mbps or 100Mbps to gigabit capable. Gigabit capability helps administrators minimize or avoid any special configuration of gigabit switches. The dedicated NIC provides remote access while the server is powered on or off and also supports IPv4, IPv6, auto-negotiation, forced speed and duplex, static IP and DHCP.

The network settings can be configured locally via the F2 System Setup Menu→ iDRAC7 Settings→ Network Menu or using RACADM or the iDRAC7 Web interface.

Shared Network Ports

12th generation servers, such as the R620 and R720, continue to support shared network port functionality. Shared network ports provide a network path through the LOMs to communicate with the iDRAC7; this communication is available for remote management when the server is on or off and all enabled server management features are supported.

Benefits of Shared Network Ports

- Since management traffic can be combined with host server traffic on the same physical cable(s), fewer physical network cables are required which helps with server cable management. Use of Virtual LANs is supported to logically partition management traffic from host server traffic.
- Due to the use of VLANs, less physical switch ports are required to support and manage.
- Management link redundancy reduces the chance for loss of remote server management communication. Automatic link failover can be configured where any LOM port can be assigned to be the primary shared network port. The failover network can be designated to any specific LOM port or to all LOM ports in a round robin fashion for systems supporting more than 2 LOM ports, such as the R620 and R720. These options provide flexibility in failover control, especially in configurations where LOMs can be of different network types, such as 2 gigabit ports and 2 10 gigabit ports.
- More fabric flexibility. Since flexible LOMs provide options that include 1 gigabit, 10 gigabit, copper and SFP interfaces, the management network interface can utilize any of these or future interfaces or fabrics.

Shared network port functionality is available with or without iDRAC7 Enterprise license. If iDRAC7 Enterprise license is enabled, then the shared network ports are still available for use. The shared network ports support auto-negotiation, forced speed and duplex, static IP and DHCP. Shared network ports support IPv4, IPv6, auto-negotiation, forced speed and duplex, static IP and DHCP. The same iDRAC7 MAC address is used if the server is configured to use dedicated or shared network ports. In either configuration, the iDRAC7 MAC address is available by default on the front, external server pullout tag, or internally on a motherboard label, or using the front LCD (if equipped) and the F2 System Setup Menu -> iDRAC7 Settings -> Network Menu.

The shared network port NIC can be configured locally using the Boot F2 Setup → iDRAC7 Settings → Network Menu, or using RACADM or the Web GUI. In addition, the LOM or integrated network card can be disabled from the host server while still usable for only server management purposes; this is configured using the Boot F2 Setup → Integrated Devices menu. If the LOMs are set to disabled and

the dedicated NIC is used by the iDRAC7, then the LOMs are put into their lowest possible power state.

Limitations of Shared Network Ports

- When the server has AC input power applied but is powered off, LOMs typically negotiate down the external link speed for reduced power consumption while still supporting Wake-on-LAN and/or shared network ports. These re-negotiations can create a momentary external link loss which may or may not interrupt any active remote management connectivity. If a connected switch is configured in a way to not allow for this dynamic speed negotiation, then remote management connectivity may not be available when the server is powered down.
- Shared network port use does not support teaming or failover of LOMs with other NICs in the system, such as those in general PCI Express slots; this is because inbound traffic on teamed NICs can arrive on any port, including the NICs to which the iDRAC7 does not have network controller side-band connectivity. Failover for iDRAC7 is supported with the LOM device ports.
- Dedicated and shared network ports cannot be used simultaneously, and dynamic failover between the dedicated NIC port and LOM ports is not supported. A user can manually modify the active port setting locally using the F2 System Setup menu, or remotely using RACADM or the Web interface.
- Remote server management sessions such as virtual KVM, virtual media, telnet, and SSH accessed using shared network ports are robust during normal server operations; however, momentary loss of remote connectivity is possible during particular server events that affect the external link. Non-session based server management traffic, such as IPMI commands, could be affected during momentary link losses. Examples of the server events that can cause link loss include server power on and off, warm reboot, link failover, UEFI network driver load, and operating system network card device driver load or unload.
    - In a network configuration with a single network switch, the number of times the session is lost due to system events is small; this number increases with more complex network topologies, or when using 10 gigabit Ethernet ports that may exhibit longer link negotiation times. On network switches, when spanning tree protocol is enabled and portfast disabled, link down times will be significantly increased on shared network ports.
    - Most of these specific server events result in a momentary interruption in service that may be noticeable to the remote user, such as a pause in the virtual KVM video. In rare cases where a session loss occurs, re-launching that feature may be required. If session loss is observed and precludes intended server management operations, then it is recommended to use the dedicated NIC or perform functions locally on the server. To help mitigate the effect of a pause in the virtual KVM video, features are available such as Video Boot Capture under the troubleshooting tab of the iDRAC7 Web interface.
    - Further resilience to reduce or eliminate the effects of shared network port link interruptions following specific server events may be achieved by not

explicitly interacting with the console mouse or keyboard during the interruption. If possible on the client's OS, modifying the client system registry to increase the number of TCP data retransmissions can help.

Influence of Networking Environment on Throughput

The iDRAC7 SoC and firmware are not solely responsible for determining the achievable throughput; instead, it is the endpoint in a larger network, with a number of devices in between it and the management station. Figure 2 illustrates that traffic can be incoming to the machine at 1Gbps rates, while being routed to the iDRAC7 at the slower 100Mbps. This section describes this and other phenomenon that will impact performance, which apply to these two cases, but are not a factor when using the dedicated management port on rack and tower servers.

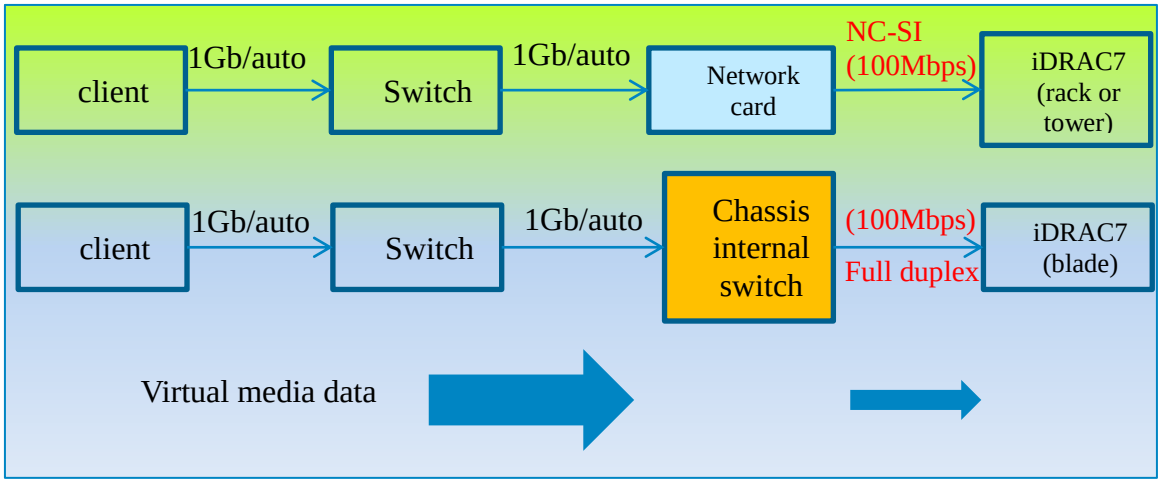

Figure 2: Speed difference on blades and in rack/tower server shared mode

Insight into TCP protocol

TCP (transmission control protocol) provides different mechanisms for acknowledgement and retries of network packets. Fundamentally, using Virtual Media as an example, packets sent from the management station ("client" in Figure 2) are acknowledged by the receiver (iDRAC7) with a network packet that reflects the next sequence number expected. TCP specifies the retransmission of lost packets by one of two mechanisms: fast retransmissions or retransmissions after a timeout. The former occurs when the sender detects "duplicate ACKs" in which the next sequence number expected is the same as a previous acknowledgement. This can only happen if the receiver observes a packet arriving out of order; the acknowledgements will continue to contain the sequence number expected in the packet that was missed or dropped. In this case, the sender resends lost data without delay, and throughput is not severely impacted. However, in case the sender is waiting on an acknowledgement that it never receives, it will wait for a timeout period (250 ms might be typical) and then try to resend the lost data; either of these cases is possible if there are dropped packets.

Turning to the switch behavior in the network, packets are received on the inbound port and routed to the correct outbound port based on the address in the packet header. In Figure 2 the switch (or network card) sending packets to iDRAC7 will sooner or later fill its internal buffers, because the data on the 1GbE network is coming in faster than it can be sent out; the sender does not necessarily wait for every acknowledgement before sending the next packet. The network device buffer management effectiveness varies from one device to another.

Flow control using PAUSE packets specified by IEEE 802.3x is one obvious option for avoiding dropped packets. However, this is not universally enabled in network devices. Also, pause packets will be sent based on a watermark defined for the port's receive buffer, and if used, will halt all traffic coming into that port regardless of the destination; if one end device is being throttled, it will affect the performance of all end devices serviced by that switch. Also, in the case of the server LOM shared mode, the cards are designed to give priority to the server networking traffic and will not halt these streams to avoid dropped packets for the management controller (iDRAC7).

As a result the management network is relying on TCP to control the flow of data to iDRAC7. There are two built in mechanisms help with this; first, the receiver defines a window size that signifies to the sender the amount of data it is prepared to receive. This size is communicated in the acknowledgment packets, and the sender will not attempt to send more data than the endpoint receiver can handle. Second, the transmitter tracks a transmission window size that is determined by computing the minimum of the observed receiver window and its own varying size based on its congestion avoidance algorithm; the latter is affected by whether or not it is observing missing acknowledgements. However, neither of these mechanisms alone will protect against possible dropped packets in the network switches.

Effect of dropped packets

Dropped packets alone do not necessarily make performance slow; however, there is an impact when the client is doing retransmissions due to packet timeouts. As illustrated in Figure 3 , retransmission timeouts can cause brief interruptions in data streams, and therefore slow down overall throughput. The timeouts occurred in this example when the network switch in the blade chassis dropped packets because it could not keep up with the incoming data rate. The throughput kept up, but it was not ideal because of the breaks in the transmission while the sender was waiting for an acknowledgement that never came for a packet that was dropped.

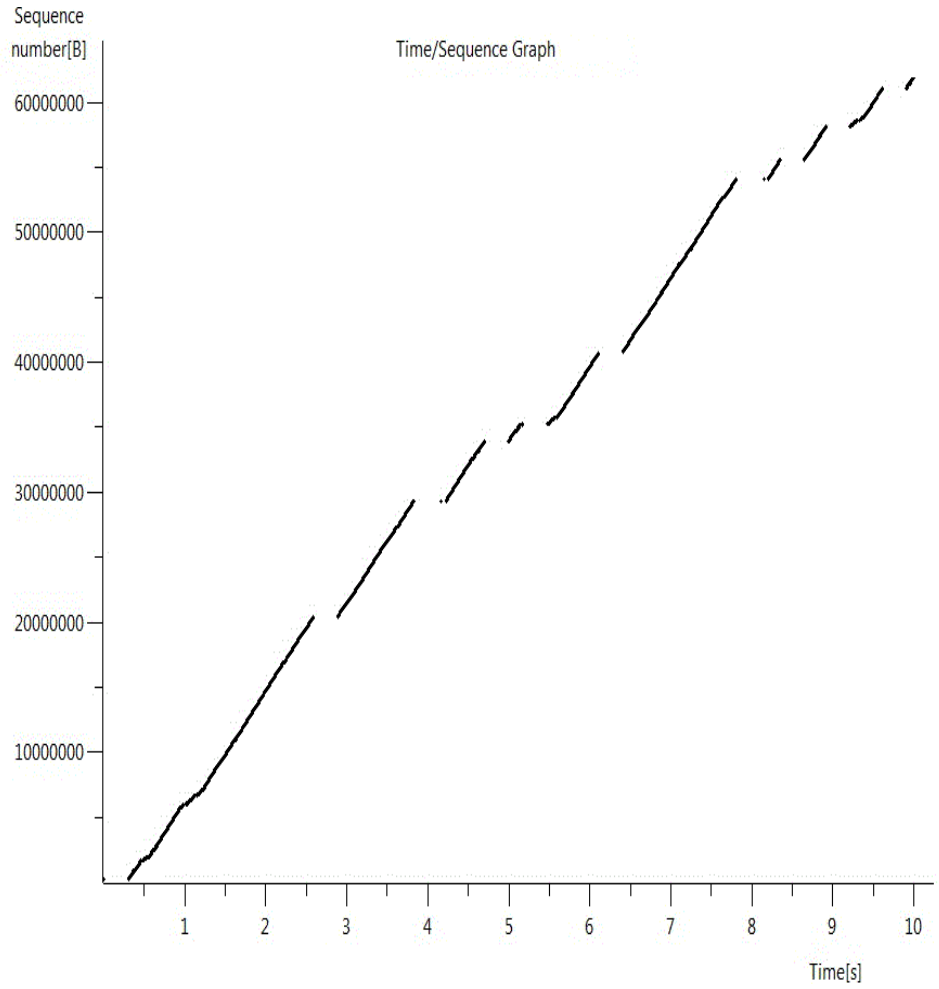

Figure 3: Illustration of retransmission timeouts during a TCP stream

Suggestions

There are several things that can be done to maximize network throughput to iDRAC7 on rack and tower servers. First, using the dedicated 1Gbps port that comes with the iDRAC7 Enterprise license will provide an uninterrupted 1Gbps Ethernet pipe for management traffic. Second, as is documented in previous and current iDRAC7 user manuals, either disable spanning tree protocol (STP) on the managed switch connected to the server, or enable the portfast setting to avoid unnecessarily long interruptions in network traffic for ports that are connected to endpoints rather than other switches. In the past, it has been possible to modify windows registry settings to affect TCP stack parameters, such as retransmission timeouts; these are not included in these suggestions because they are not available on all operating system versions. When not possible to use a dedicated management network with the 1GbE port, there are some additional recommendations offered here for use if applicable in the IT environment.

Avoid Duplex mismatch

The need for half duplex during the advent of computer networks has created a present-day phenomenon that can have extreme effects on network performance. If a half-duplex port is talking to a link partner whose setting is for full duplex, erroneous collision detection can occur and invoke legacy recovery mechanisms used when CSMA/CD was necessary on local area networks. Duplex mismatch can occur if network ports are intentionally configured for a certain duplex setting, but it can also occur if auto-negotiating ports fall back to half duplex to talk to a full-duplex partner; be aware of this phenomenon, and ideally configure all network ports to 1Gbps/auto-negotiate.

100Mbps configuration

If duplex mismatch can be avoided, and the illustrations in Figure 2 apply to your environment and management workstations are used exclusively for that activity, consider configuring the NIC card and the accompanying managed switch port to 100Mbps speed/ full duplex. By doing this, the traffic from the potential Virtual Media data sender is matched to the speed that it is sent to its destination. Flow control becomes less of an issue, and dropped packets become unlikely, because the devices routing to the iDRAC7 are sending as fast as they are receiving.

Disabling large send offload

If this setting is available in your client machine (management station) network interface card settings, it can be disabled to decrease the chances of dropped packets. Large send offload is a feature of network cards that allows a software application to send data of sizes larger than the MTU (maximum transmission unit - 1500 bytes) to the network stack. For example, the application can send 64Kbytes, and without intervention from the CPU, the NIC will break this up into MTU packets for transmission on the wire. While having the intended advantage of CPU offload, when used in environments as illustrated in Figure 2, this can create fast bursts of data filling the switch or network device buffers en route to iDRAC7. By disabling this setting, there is artificial throttling on the client machine that brings the data rate down to a level that can be handled without packet loss by the 100Mbps pipe at the end of the network to the iDRAC7 endpoint. Note that this does not work in every case, and is highly dependent on the client machine and its CPU load.

Minimize bandwidth consumed by Virtual Console

The typical usage of Virtual Media includes concurrent usage of the iDRAC7 Virtual Console vKVM (virtual keyboard, video, and mouse). There are two ways to minimize the network bandwidth consumed by the Virtual Console session, and in turn leave more bandwidth available for the Virtual Media TCP stream. First, virtual media can be launched without the virtual console. For example, in the iDRAC7 Web interface if the Virtual Console is disabled and the user elects to launch the console, he will be prompted to invoke only the Virtual Media application. There is also a command line utility VMCLI, documented in the iDRAC7 User Manual, that allows you to use Virtual Media without the remote desktop. Second, in the Virtual Console session, there is a pull-down menu titled “Performance” that allows the user to select a tradeoff between video quality determined by color depth and frame rate vs. speed; selecting a preference for “maximum speed” will reduce the

necessary bandwidth for the desktop display, leaving more available for the Virtual Media TCP stream.

12G Speeds and Results

Dell tested a wide variety of client machines (management stations), and operating systems, with the PowerEdge product line. Table 1 represents a small sample of the results; it is not intended to be exhaustive, but rather demonstrate the achievable speeds. The baseline CD speed for reference is 150Kbytes per second or 1.2 Mbps, so 10X would represent 12Mbps. Note that although the iDRAC7 CPU processing efficiency is not the subject of this paper, decryption can have a significant impact on the throughput. Therefore, the user of Virtual Media should make an informed decision whether encryption is necessary in their application, or if significantly increased throughput is more advantageous.

Managed Server	Management Client	Encryption	Dedicated/Shared Mgmt. network	Speed (X CD)
R720 Rack	Windows7 64-bit/ ActiveX plug-in	Disabled	Dedicated	52.6X
R720 Rack	Windows7 64-bit/ ActiveX plug-in	Enabled	Dedicated	20.5X
R620 Rack	Windows7 64-bit/ Java plug-in	Enabled	Shared	17.1X
M620 blade	Windows7 64 bit/ Java plug-in (Note1)	Disabled	N/A	39.24

Table 1: Sample of results from Virtual Media testing

Note1: This number demonstrates a gain achieved by configuring the client to 100Mbps speed.

Conclusion

Network performance can vary drastically, and it has been the subject of many years of intense research and development. By understanding some fundamental characteristics of your network environment, performance can be optimized. Dell PowerEdge 12th Generation servers offer compelling opportunities for performance improvements in the IT professional's management network. The iDRAC7 in particular has been upgraded to improve the systems management experience for those charged with that task.

About the authors

Andy Butcher is a Principal Software Engineer and Tim Lambert is a Senior Systems Engineer who work together on the management subsystem solution for PowerEdge servers. The authors would like to particularly acknowledge and credit Doug Roberts, software engineer, for the tireless collection of data and insight provided for this paper. A great deal of credit is deserved also to Florin Dragan, software engineer, and our colleagues in our Bangalore Design Center, Shine K A, Anshul Simlote, and Tharakarama Raju for their attention to performance improvements in iDRAC7. Thanks also to Chip Webb and Rich Hernandez for their technical expertise, and to Phil Webster for creating experimental firmware images.