EXPLORING THE DELL POWEREDGE M1000e NETWORK FABRIC ARCHITECTURE

By John Loffink

Modular components can be key to keeping pace with fast-moving IT advancements. The architecture of the new Dell™ PowerEdge™ M1000e modular blade server enclosure and its 10th-generation Dell server technology is designed to support both current and future network technologies, helping protect enterprise modular server investments.

Related Categories:

Blade servers

Dell PowerEdge blade servers

Dell PowerEdge servers

Visit DELL.COM/PowerSolutions for the complete category index.

Designed from the ground up to support current and future generations of server, storage, networking, and management technologies, the PowerEdge M1000e modular blade server enclosure includes the headroom for multiple generations of connectivity. As part of its modular architecture, key PowerEdge M1000e components, such as the LAN on Motherboards (LOMs) and mezzanine cards in the server blades and the I/O modules (IOMs) and midplane in the enclosure, are designed to support three high-speed fabrics per server blade: one standard Gigabit Ethernet (GbE) fabric and two additional customizable fabrics providing sufficient bandwidth to support interconnects such as 10 Gigabit Ethernet (10GbE), Fibre Channel, and InfiniBand. The enclosure uses redundant hot-pluggable components throughout to help maximize system availability.

The PowerEdge M1000e enclosure supports up to 16 half-height server modules, each occupying a slot accessible in the front of the enclosure. This article discusses the PowerEdge M1000e fabric architecture when using these half-height blades. Blades in larger form factors can support a proportionally higher level of fabric connectivity.[1]

MODULAR SERVER I/O

To understand the PowerEdge M1000e architecture, it is necessary to first define four key terms: *fabric*, *lane*, *link*, and *port*.

A *fabric* is defined as a method of encoding, transporting, and synchronizing data between multiple devices. Examples of fabrics are GbE, Fibre Channel, or InfiniBand. Fabrics are carried inside the PowerEdge M1000e system between server modules and IOMs through the midplane. They are also carried to the outside world through the physical copper or optical interfaces on the IOMs.

A *lane* is defined as a single fabric data transport path between I/O end devices. In modern high-speed serial interfaces, each lane comprises one transmit and one receive differential pair. In reality, a single lane is four wires in a cable or traces of copper on a printed circuit board: a transmit positive signal, a transmit negative signal, a receive positive signal, and a receive negative signal. Differential pair signaling provides improved noise margin for these high-speed lanes. Various terminologies are used by fabric standards when referring to lanes. PCI Express (PCIe) calls this a lane, InfiniBand calls it a physical lane, and Fibre Channel and Ethernet call it a link.

[1] For more information on the PowerEdge M1000e, see "The Next-Generation Dell PowerEdge M1000e Modular Blade Enclosure," by Chad Fenner, in *Dell Power Solutions*, February 2008, DELL.COM/Downloads/Global/Power/ps1q08-20080206-Fenner.pdf.

> **“Key PowerEdge M1000e components are designed to support three high-speed fabrics per server blade: one standard Gigabit Ethernet fabric and two additional customizable fabrics.”**

A *link* is defined here as a collection of multiple fabric lanes used to form a single communication transport path between I/O end devices. Examples are four-lane (x4), eight-lane (x8), and sixteen-lane (x16) PCIe, or four-lane 10GBase-KX4. PCIe, InfiniBand, and Ethernet call this a link. The differentiation has been made here between *lane* and *link* to help prevent confusion over Ethernet’s use of the term *link* for both single- and multiple-lane fabric transports. Some fabrics such as Fibre Channel do not define links, as they simply run multiple lanes as individual transports for increased bandwidth. A link as defined here provides synchronization across the multiple lanes, so they effectively act together as a single transport.

A *port* is defined as the physical I/O end interface of a device to a link. A port can have single or multiple lanes of fabric I/O connected to it.

SERVER BLADE I/O ARCHITECTURE

There are three supported high-speed fabrics per PowerEdge M1000e half-height server module, with two flexible fabrics using optional plug-in mezzanine cards on the server. The server blades used in the PowerEdge M1000e enclosure are designed to support multiple network topologies and provide sufficient bandwidth to accommodate future upgrades. I/O fabric integration encompasses LANs, storage area networks (SANs), and Interprocess Communication (IPC) networks. The blade ports connect through the enclosure midplane to the associated IOMs in the back of the enclosure, which then connect to the LAN, SAN, or IPC network.

As shown in Figure 1, the first embedded high-speed fabric, referred to as fabric A, comprises dual GbE LOMs and their associated enclosure IOMs. The LOMs are based on the Broadcom BCM5708 NetXtreme II Ethernet controller and support TCP/IP Offload Engine (TOE) and Internet SCSI (iSCSI) boot. Although the blades currently have a dual GbE LOM configuration, the enclosure midplane design allows future support for up to four GbE LOMs in each half-height blade. The blade LOMs and Ethernet fabric B and C mezzanine cards also support Wake-on-LAN (WOL). SAN boot is supported by iSCSI- and Fibre Channel–enabled cards.

In addition to fabric A, the blades support additional fabrics B and C by installation of two optional dual-port I/O mezzanine cards in each half-height blade. These cards can currently support a wide array of Ethernet (including iSCSI), Fibre Channel, and InfiniBand technologies and use a common form factor to provide flexibility in fabric configuration. Fabric B and C I/O mezzanine cards have identical mechanical, electrical, and management specifications, providing a high level of flexibility and modularity.

The optional mezzanine cards connect to the blade chipsets through eight-lane (x8) PCIe interfaces, providing up to 16 Gbps of bandwidth per mezzanine card with Gen1 PCIe. Both the PCIe fabrics and external fabrics are routed through high-speed, 10 Gbps–capable air dielectric connector pins through the planar and

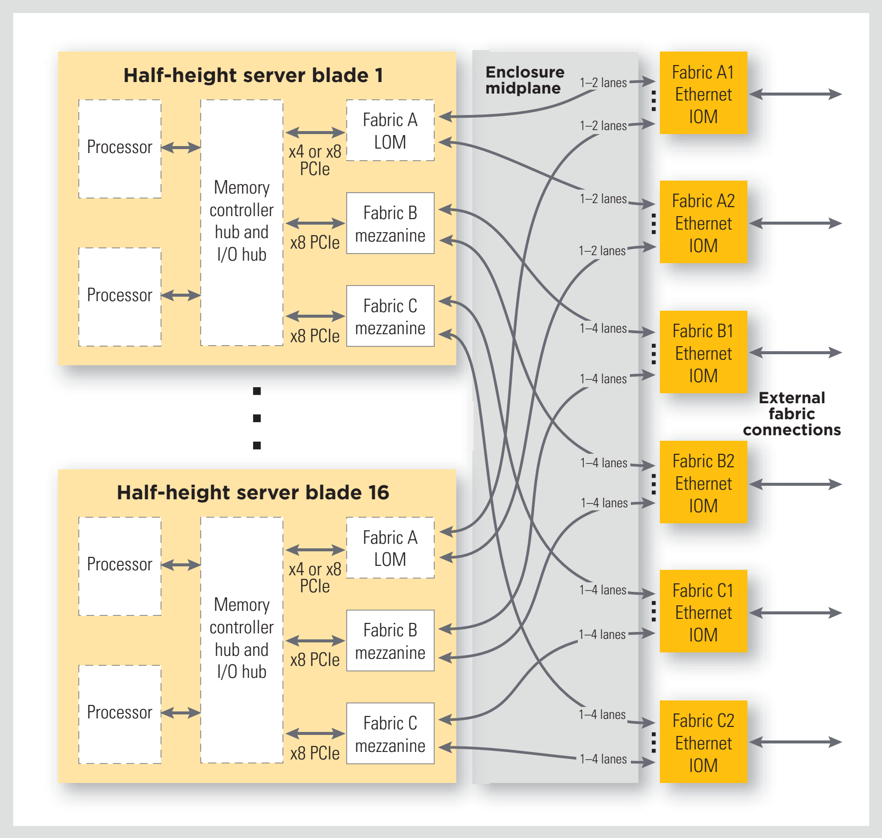


Figure 1. *Dell PowerEdge M1000e high-speed fabric architecture*

midplane. To enhance signal integrity, the signals isolate transmit and receive signals to help minimize crosstalk. Differential pairs are isolated with ground pins, and signal connector columns are staggered—helping minimize signal coupling.

Because of the modular architecture and flexibility to configure mezzanines and IOMs in the system, administrators could inadvertently hot plug a blade with the wrong mezzanine card into the system. To help prevent the accidental activation of mis-configured fabrics on a blade, the PowerEdge M1000e systems management hardware and software include fabric consistency checking. For example, if administrators have configured Fibre Channel IOMs in the fabric C slots, then all blades must have either a Fibre Channel mezzanine card or no mezzanine card in that fabric slot. If, in this case, they hot plugged a blade into the enclosure that had a GbE mezzanine card in its fabric C slot, the system would automatically detect this misconfiguration and alert administrators of the error so they could appropriately reconfigure the blade.

ENCLOSURE I/O ARCHITECTURE

A key capability introduced with the PowerEdge M1000e is comprehensive 10/100/1,000 Mbps Ethernet support when using Ethernet pass-through modules. In the past, pass-through connections were limited to 1,000 Mbps or GbE speeds. The PowerEdge M1000e enables organizations to connect to legacy 10/100 Mbps infrastructures using Ethernet pass-through or switch technology. This feature uses in-band signaling on a 1000Base-KX transport and does not require administrator interaction to function.

The enclosure midplane can support up to four GbE links per blade for fabric A, providing up to 4 Gbps of bandwidth per half-height blade. Fabrics B and C are routed as two sets of four lanes from mezzanine cards on the blades to the IOMs in the back of the enclosure. Supported bandwidth ranges from 1 to 10 Gbps per lane depending on the fabric, or up to 80 Gbps per mezzanine card.

Because each mezzanine card connects to the blade chipset through an eight-lane PCIe link, the system has no throttle points constricting I/O bandwidth. Dell anticipates that when multi-lane 10GBase-KR technology is available, blades will have moved to PCIe 2.0 or better, providing full end-to-end I/O bandwidth from the server blades to the enclosure IOMs.

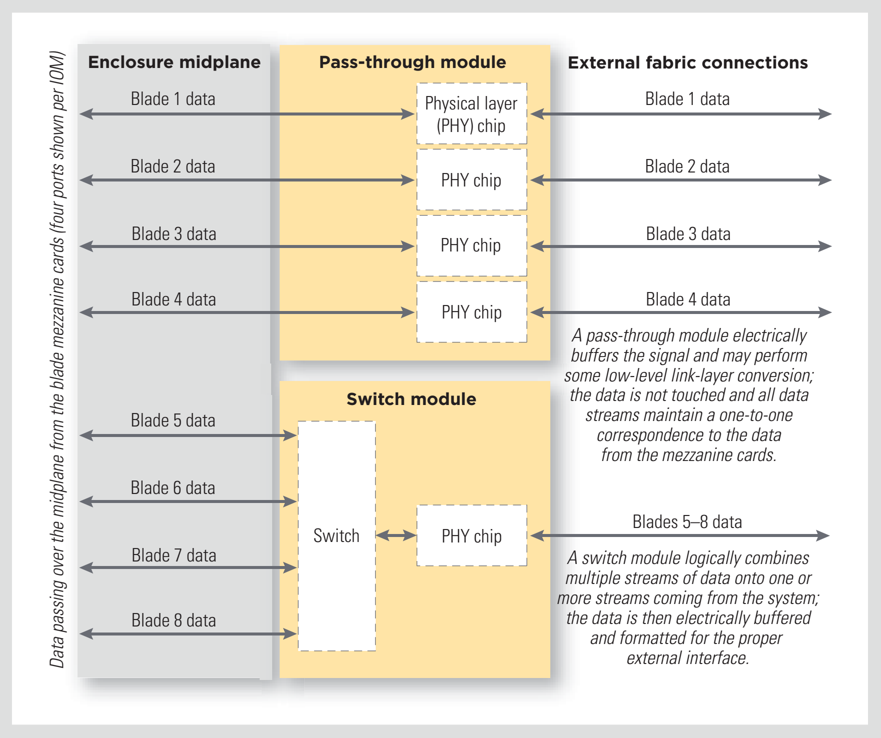
Midplane

The midplane—a large printed circuit board providing power distribution, fabric connectivity, and systems management infrastructure—is the focal point for all connectivity within the PowerEdge M1000e enclosure, and is designed to provide scalable bandwidth for both current and future generations of servers and infrastructure. As is required for fault-tolerant systems, the PowerEdge M1000e midplane is passive, with no hidden stacking midplanes or interposers with active components. The I/O fabrics and systems management infrastructure are designed to be fully redundant for each hot-pluggable component.

I/O fabrics are routed through 10 Gbps–capable high-speed connectors and dielectric material. The I/O channels have been simulated to 10GBase-KR channel models. Following industry standards, the fabrics internally support a bit error rate of 10^{-12} or better. Dell has made a high level of investment to help ensure scalable bandwidth for current and future generations of servers and infrastructure.

I/O modules

The back of the PowerEdge M1000e enclosure contains systems management, cooling, power, and I/O components. The IOMs are used as pairs, with two fully redundant modules for each blade fabric, and can be pass-through or switch modules. Pass-through modules provide direct one-to-one connectivity from each LOM or mezzanine card port on each blade to the external network. Switches provide an efficient way to consolidate links from the LOM or mezzanine cards



A pass-through module electrically buffers the signal and may perform some low-level link-layer conversion; the data is not touched and all data streams maintain a one-to-one correspondence to the data from the mezzanine cards.

A switch module logically combines multiple streams of data onto one or more streams coming from the system; the data is then electrically buffered and formatted for the proper external interface.

Figure 2. *Pass-through and switch modules as part of the Dell PowerEdge M1000e architecture*

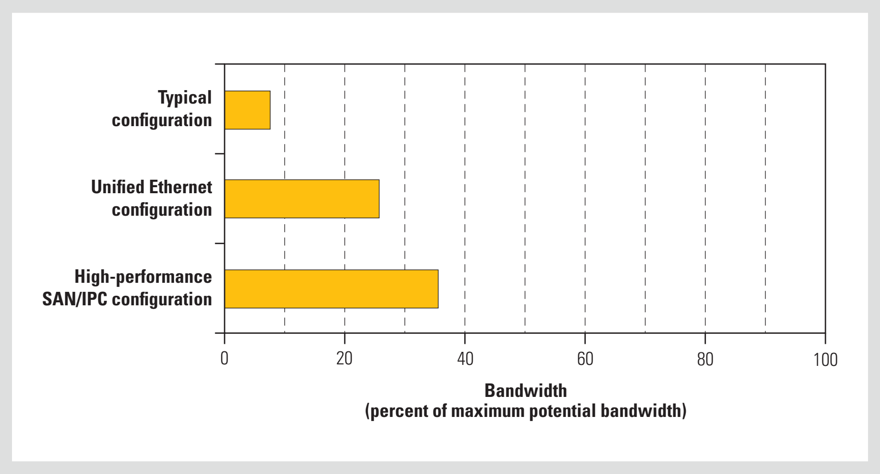


Figure 3. *Bandwidth requirements for three possible Dell PowerEdge M1000e configurations compared with total potential system bandwidth*

on the blades to network uplinks. IOMs are designed to be fully compatible across slots. Figure 2 illustrates these different components as part of the PowerEdge M1000e architecture.

The Dell PowerConnect™ M6220 Ethernet switch and Cisco Catalyst blade switch (3032, 3130S-G, and 3130S-X) for the PowerEdge M1000e enclosure demonstrate the advanced modularity of the IOMs. These Ethernet switch modules introduce sub-module I/O expansion to help maximize flexibility and extension. They provide a single hardware design that allows scaling from a cost-effective GbE configuration to one that utilizes switch stacking to interconnect multiple switches within or between enclosures or to one that adds support for 10GbE uplinks to the core network with both copper and optical interfaces.

Scalable system bandwidth

Assuming a full population of GbE and 10GbE lanes in all three fabrics, the PowerEdge M1000e architecture can support a total potential bandwidth of 5.44 Tbps. As shown in Figure 3, a typical configuration of four lanes of GbE and two lanes of Fibre Channel per server blade uses less than 10 percent of this total potential bandwidth. A unified Ethernet configuration—one that aggregates all network, storage, and IPC-over-Ethernet links by using two lanes of GbE and four lanes of 10GbE per blade—uses only about 25 percent of the total potential bandwidth. Even a high-performance SAN/IPC configuration using two lanes of GbE, two lanes of 8 Gbps Fibre Channel, and two lanes of QDR InfiniBand still uses only about 35 percent of the total potential bandwidth.

The PowerEdge M1000e is also designed to provide comprehensive support for near-, medium-, and long-term I/O infrastructure requirements in a flexible, cost-effective way. An example of this flexibility is the routing of dual two-lane paths for fabric A and dual four-lane paths for fabrics B and C. In the near term, 10GBase-KX4 routing supports 10GbE connectivity. 10GBase-KX4 routing uses all 4 lanes, each running at a 2.5 Gbps data rate to achieve total link data bandwidth of 10 Gbps. Today and in the near-term future, 10GBase-KX4 is the most cost-effective, ubiquitous solution for 10GbE fabrics over midplanes. A fully configured PowerEdge M1000e system supports two dual GbE fabrics and two dual 10GbE fabrics with 10GBase-KX4 routing, supporting the leading edge of unified network topologies. Such a configuration could, for instance, provide redundant 10GbE links per blade for traditional network traffic, and another set of redundant 10GbE links for iSCSI or Fibre Channel Over Ethernet networked storage, and still maintain dual two-lane GbE links for systems management or other low-bandwidth requirements.

MODULAR BLADE SERVER ARCHITECTURE

The architecture of the Dell PowerEdge M1000e modular blade server enclosure and its 10th-generation Dell server technology has been designed specifically with modularity in mind, providing customizable multi-lane fabrics that can support both current and future network technologies and help protect enterprise blade server investments. As part of the Dell focus on simplifying IT, there is no confusing support matrix to follow for PowerEdge M1000e mezzanines or IOMs and no multiplication of mezzanine and IOM form factors and design standards, helping avoid potential compatibility and configuration difficulties. The PowerEdge M1000e supports one mezzanine design standard and one IOM design standard to enable true modularity at the system and subsystem level, helping simplify extension and enhancement now and in the future.

John Loffink is an engineer/strategist in the Dell Server Advanced Engineering Group. He has over 20 years' background in servers, enterprise storage, hardware design, and high-availability computing. John holds a B.S. in Electrical Engineering from the Florida Institute of Technology.

MORE ONLINE
DELL.COM/PowerSolutions

QUICK LINK

Dell PowerEdge M1000e:
DELL.COM/Servers