Dell PowerEdge M1000e Blade Enclosure and EqualLogic PS Series 10 GbE SAN Design Best Practices

A Dell EqualLogic Best Practices Technical White Paper

Storage Infrastructure and Solutions Engineering
Dell Product Group
October 2012

Table of Contents

Acknowledgements

This white paper was produced by the PG Storage Infrastructure and Solutions team of Dell Inc.

The team that created this white paper:

Clay Cooper, **Guy Westbrook**, and **Camille Daily**

Feedback

We encourage readers of this publication to provide feedback on the quality and usefulness of this information by sending an email to SISfeedback@Dell.com.

SISfeedback@Dell.com

1 Introduction

With the Dell™ EqualLogic™ PS Series storage arrays, Dell provides a storage solution that delivers the benefits of consolidated networked storage in a self-managing iSCSI storage area network (SAN) that is affordable and easy to use, regardless of scale. By eliminating complex tasks and enabling fast and flexible storage provisioning, these solutions dramatically reduce the costs of storage acquisition and ongoing operation.

To leverage the advanced features provided by an EqualLogic array, a robust, standards-compliant iSCSI storage area network (SAN) infrastructure must be created. When using blade servers in a Dell PowerEdge™ M1000e blade enclosure (also known as a blade chassis) as hosts, there are a number of network design options for storage administrators to consider when building the iSCSI SAN. For example, the PS Series array member network ports can be connected to the switches within the M1000e blade chassis or the blade server network ports can be connected to top of rack (TOR) switches residing outside of the blade chassis. After testing and evaluating a variety of different SAN design options, this technical white paper quantifies the ease of administration, the performance, the high availability, and the scalability of each design. From the results, recommended SAN designs and practices are presented.

1.1 Audience

This technical white paper is intended for storage administrators, SAN/NAS system designers, storage consultants, or anyone who is tasked with integrating a Dell M1000e blade chassis solution with EqualLogic PS Series storage for use in a production storage area network. It is assumed that all readers have experience in designing and/or administering a shared storage solution. Also, there are some assumptions made in terms of familiarity with all current and possibly future Ethernet standards as defined by the Institute of Electrical and Electronic Engineers (IEEE) as well as TCP/IP and iSCSI standards as defined by the Internet Engineering Task Force (IETF).

1.2 Terminology

This section defines terms that are commonly used in this paper and the context in which they are used.

TOR switch – A top of rack (TOR) switch, external to the M1000e blade chassis.

Blade IOM switch – A blade I/O module (IOM) switch, residing in an M1000e Fabric slot.

Stack – An administrative grouping of switches that enables the management and functioning of multiple switches as if they were one single switch. The switch stack connections also serve as high-bandwidth interconnects.

LAG – A link aggregation group (LAG) in which multiple switch ports are configured to act as a single high-bandwidth connection to another switch. Unlike a stack, each individual switch must still be administered separately and function as such.

Uplink – A link that connects the blade IOM switch tier to the TOR switch tier. An uplink can be a stack or a LAG. Its bandwidth must accommodate the expected throughput between host ports and storage ports on the SAN.

ISL – An inter-switch link that connects either the two blade IOM switches or the two TOR switches to each other. An ISL can be a stack or a LAG.

Blade IOM switch only – A category of SAN design in which the network ports of both the hosts and the storage are connected to the M1000e blade IOM switches, which are isolated and dedicated to the SAN. No external TOR switches are required. The ISL can be a stack or a LAG, and no uplink is required.

TOR switch only – A category of SAN design in which the network ports of both the hosts and the storage are connected to external TOR switches. For this architecture, 10 GbE pass-through IOM are used in place of blade IOM switches in the M1000e blade chassis. The ISL can be a stack or a LAG.

Blade IOM switch with TOR switch – A category of SAN design in which host network ports are internally connected to the M1000e blade IOM switches and storage network ports are connected to TOR switches. An ISL stack or LAG between each blade IOM switch and/or between each TOR switch is required. An uplink stack or LAG from the blade IOM switch tier to the TOR switch tier is also required.

Switch tier – A pair or more of like switches connected by an ISL which together create a redundant SAN Fabric. A switch tier might accommodate network connections from host ports, from storage ports, or from both. If all switches in a switch tier are reset simultaneously, for example if the switch tier is stacked and the firmware is updated, then the SAN is temporarily offline.

Single switch tier SAN design – A SAN design with only blade IOM switches or TOR switches but not both. Both host and storage ports are connected to the same type of switch and no uplink is required. Blade IOM switch only and TOR switch only designs are single switch tier SAN designs.

Multiple switch tier SAN design – A SAN design with both blade IOM switches and TOR switches. Host and storage ports are connected to different sets of switches and an uplink stack or LAG is required. Blade IOM switch with TOR switch designs are multiple switch tier SAN designs.

Host/port ratio – The ratio of the total number of host network interfaces connected to the SAN divided by the total number of active PS Series array member network interfaces connected to the SAN. A ratio of 1:1 is ideal for optimal SAN performance, but higher port ratios are acceptable in specific cases. The host to port ratio can negatively affect performance in a SAN when oversubscription occurs, that is when there are significantly more host ports or significantly more storage ports.

1.2.1 Terminology illustration

Figure 1 illustrates the basic SAN components involved when deploying an M1000e blade chassis with blade servers into an EqualLogic PS Series array SAN. When creating the SAN to connect blade server network ports to storage array member network ports, the SAN might consist of only blade IOM switches, only TOR switches, or both switch types together in two separate tiers. Note that the blade servers connect to the blade IOM switches internally with no cabling required. Blade servers can also be connected to TOR switches if the blade IOM switches are replaced with pass through IOM.

If only TOR switches or only blade IOM switches are used, this paper will refer to the SAN as a single switch tier design. When both switch types are used then the SAN will be referred to as a multiple switch tier design. In multiple switch tier SAN designs the multiple switch tiers will need to be connected by an uplink, which can be either a stack or a LAG. For both single and multiple switch tier SAN designs a matched switch pair will need to be interconnected by an inter-switch link or ISL. Like the uplink, the ISL

can be a stack or a LAG. The ISL is necessary to create a single layer 2 SAN Fabric over which all PS Series array member network ports can communicate with each other.

For a much more detailed description of each SAN design that was tested and evaluated see Section 4 titled, "Tested SAN designs".

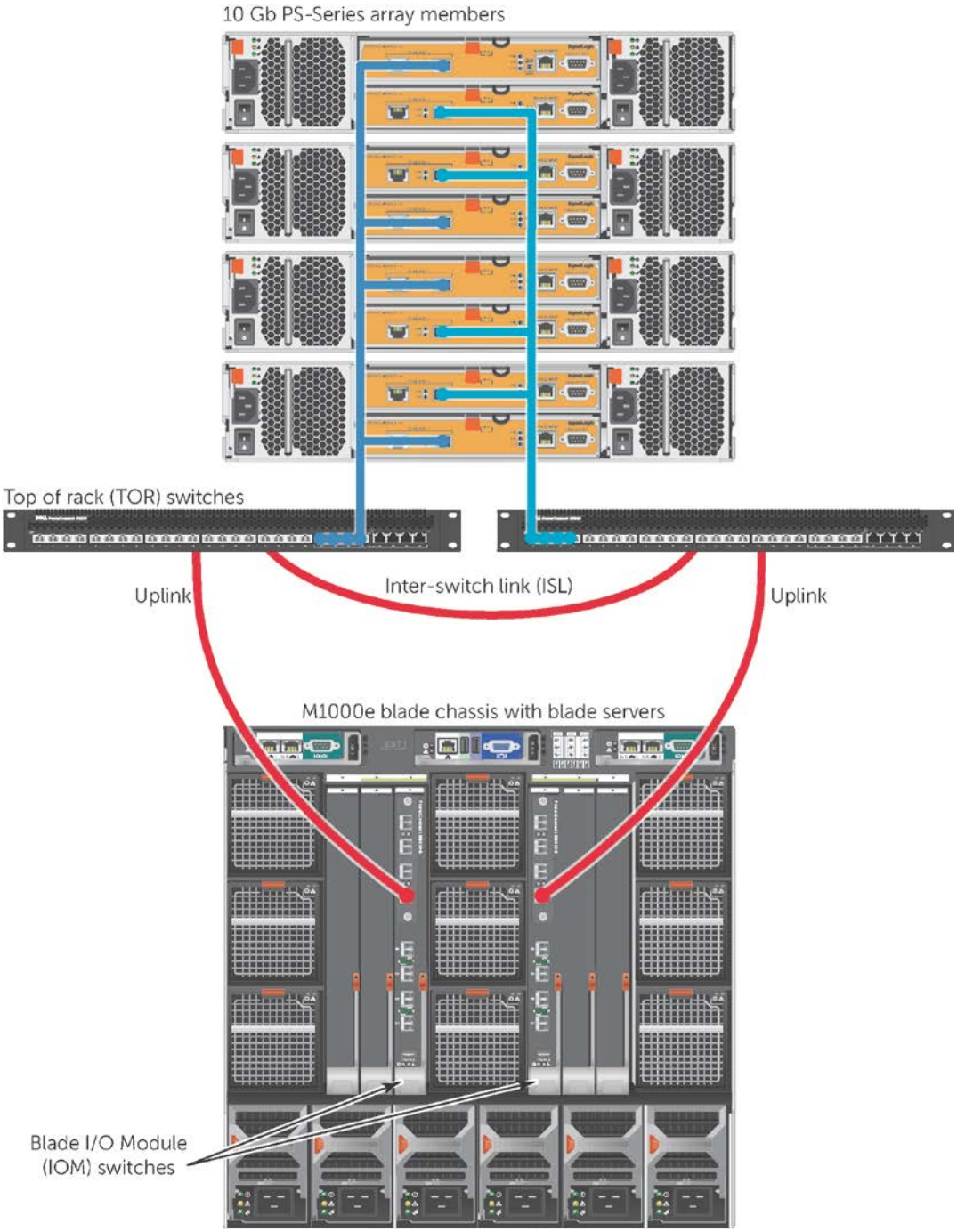

Figure 1 An example EqualLogic SAN consisting of PS Series array members, an M1000e blade chassis with blade servers, and TOR and Blade IOM switches.

2 Overview of M1000e blade chassis solution

The following section describes the M1000e blade chassis networking Fabrics consisting of IO modules, a midplane, and the individual blade server network adapters.

2.1 Multiple Fabrics

Each M1000e can support up to three separate networking Fabrics that interconnect ports on each blade server to a pair of blade IO modules within each chassis Fabric through a passive chassis midplane. Each Fabric is associated with specific interfaces on a given blade server as described in Table 2. Each blade server has a LAN on Motherboard (LOM) or a Network Daughter Card (NDC) that is mapped to the blade IOM located in the Fabric A slots in the M1000e chassis. In addition, each blade server has two mezzanine sockets for adding additional networking options such as 1 Gb or 10 Gb Ethernet, Infiniband, or Fibre Channel cards. These mezzanine cards are mapped to either the Fabric B or the Fabric C blade IOM.

Figure 2 illustrates the layout of the three Fabric blade IOM located in the back of the M1000e chassis.

Table 1 M1000e Fabric mapping

	LOM/NDC	Mezzanine B	Mezzanine C
Fabric	A	B	C

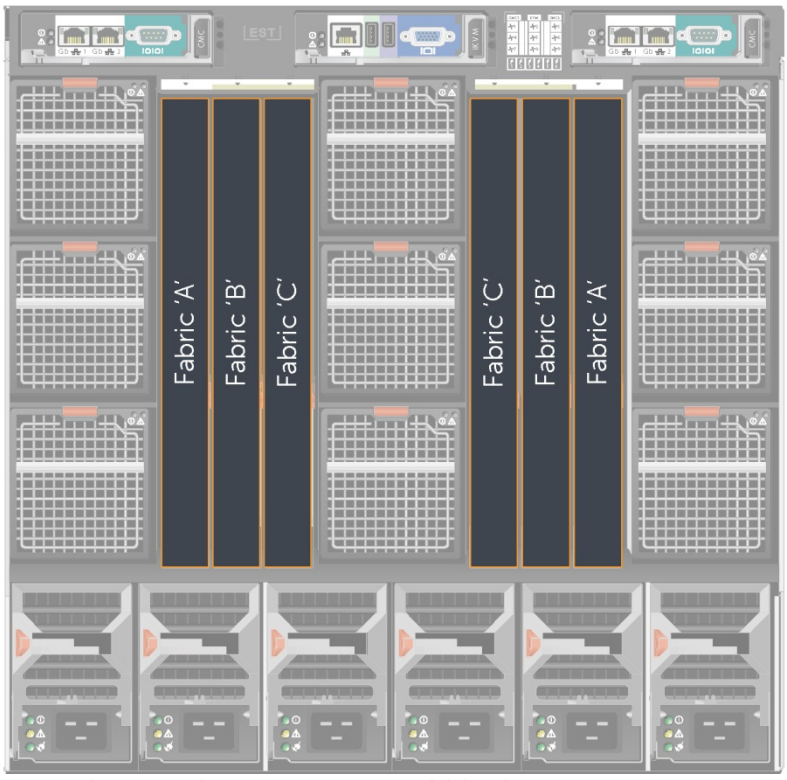

Figure 2 Blade IO Modules and M1000e Chassis

2.2 Blade IO modules

The following table lists the 10 GbE blade IO module options (available at the time of this publication) and the number of ports available for EqualLogic SAN solutions.

Table 2 1 GbE Blade IO Module options for EqualLogic

	10 GbE external ports	40 GbE uplink ports
PowerConnect M8428-k*	8	N/A
PowerConnect M8024-k	Up to 8	N/A
Force10 MXL*	Up to 8	Up to 6
10 GbE Pass-through	16	N/A

***Note:** Only the PowerConnect M8024-k and 10 GbE Pass-through IOM were used during testing, not the PowerConnect M8428-k or Force10 MXL.

3 Summary of SAN designs and recommendations

This section provides the high level conclusions reached after the course of comprehensive lab testing and analysis of various EqualLogic PS Series array SAN designs which incorporate M1000e blade server hosts on a 10 GbE network.

The recommendations assume two M8024-k switches or two pass-through IO modules per blade chassis, two SAN ports per host and, if applicable, 24-port PC8024F TOR switches.

For complete results and recommendations see Section 5 titled, "Detailed SAN design analysis and recommendations". For an illustration of each SAN design see Section 4 titled, "Tested SAN designs".

Note: Green cells indicate recommended SAN designs based on all factors considered during testing, while orange cells indicate designs that for various reasons, such as SAN availability or uplink bandwidth, might not be preferred.

Table 3 Summary of SAN designs and recommendations

	Switch tier topology	Ease of administration	Performance	High availability	Scalability
Blade IOM with ISL stack	Single	• A single switch stack to manage • No TOR switches required • Fewest cables • During ISL stack reload SAN is unavailable	• Equivalent performance during small scale SAN testing	• Blade IOM switch failure reduces host ports by 50%	• High host/storage port ratios even with a maximum five arrays and only eight blade servers per chassis
Blade IOM with ISL LAG	Single	• Two switches to manage • No TOR switches required • Fewest cables	• Equivalent performance during small scale SAN testing	• Blade IOM switch failure reduces host ports by 50%	• High host/storage port ratios even with a maximum five arrays and only eight blade servers per chassis
TOR with ISL stack	Single	• A single switch stack to manage • No blade IOM switches required • Most cables	• Equivalent performance during small scale SAN testing	• TOR switch failure reduces host ports by 50%	• Can support up to 12 arrays with four TOR switches • Additional TOR switches must be connected with ISL
TOR with ISL LAG	Single	• Two switches to manage • No blade IOM switches required • Most cables	• Equivalent performance during small scale SAN testing	• TOR switch failure reduces host ports by 50%	• Can support up to 12 arrays with four TOR switches • Additional TOR switches must be connected with ISL

	Switch tier topology	Ease of administration	Performance	High availability	Scalability
Blade IOM and TOR with 3-way LAG	Multiple	• Four switches to manage • Uplinks required between tiers	• Equivalent performance during small scale SAN testing	• Blade IOM or TOR switch failure reduces host ports by 50% • Blade IOM or TOR switch failure reduces uplink bandwidth by 50%	• Highly scalable, allowing 16 array members and a second blade chassis using four TOR switches • Additional TOR switches must be connected with ISL

3.1 Ease of administration

If reducing administrative overhead is the goal, a single switch tier design with an ISL stack is the simplest option. Because the storage is directly attached to the blade IOM switches, fewer cables are required than with the TOR switch only design, and the ISL stack allows the switches to be administered as a single switch.

If the availability of the SAN is critical, then an ISL LAG configuration may be preferred over stacking. If a switch tier ISL is stacked, then a switch stack reload (required for tasks such as switch firmware updates) will temporarily reset the entire switch tier making the SAN unavailable during that time. In this case, SAN downtime for firmware updates would have to be scheduled. A multiple switch tier design that avoids this is the three-way LAG design.

If TOR switches from a different vendor are used, then the simplest choice is to implement the TOR only design by cabling M1000e pass-through IOM directly to the TOR switches. If multiple switch tiers are desired, plan for an uplink LAG as the blade IOM switches will not be stack-compatible with the TOR switches from a different vendor.

3.2 Performance

The throughput values were gathered during the performance testing of each SAN design with four hosts and four arrays members at three common workloads. At this SAN scale, there were no significant performance differences among the SAN designs during any of the three tested workloads.

3.3 High availability

In both the TOR and blade IOM switch failure scenarios, all tested SAN designs suffer a temporary 50% reduction in the number of connected host ports and for the multiple switch tier three-way LAG SAN design, a 50% reduction in uplink bandwidth. In the case of the three-way SAN design this is problematic because the remaining 60 Gbps uplink bandwidth will not accommodate the throughput of more than six array members even though port count allows up to 12 array members. Therefore, when deploying a SAN consisting of more than six array members, the vulnerability of the multiple switch tier three-way LAG SAN design to uplink bandwidth insufficiency during switch failures should be kept in mind.

3.4 Scalability

The blade IOM switch only SAN design does not yield optimal host/storage port ratios even with the maximum number of arrays and only eight full-height blade servers per blade chassis.

With 12 array members and four TOR switches, the TOR switch only SAN design yields an acceptable host/storage port ratio, even with 16 blade servers per blade chassis.

The blade IOM and TOR switch SAN design is the most scalable of all, allowing 16 blade servers per chassis to achieve a 2:1 host/storage port ratio with 16 array members using four TOR switches.

Note: The scalability data presented in this paper is based primarily on available port count. Actual workload, host to array port ratios, and other factors may affect performance.

3.5 Conclusion

When using PowerConnect 8024-k IOM switches, the blade IOM switch only SAN design is not recommmended due to the high host/port ratios caused by the limited number of ports available for array member port connection.

The TOR switch only SAN design requires four switches to produce optimal host/port ratios. Though it reduces administrative overhead, an ISL stack should be avoided if SAN availability must be maintained during switch configuration changes or firmware updates.

The multiple tier SAN design with uplink and ISL LAG between the blade IOM and TOR switches allows up to 12 array members using two TOR switches and up to 16 array members and two M1000e blade chassis using four TOR switches. However, a loss of either a blade IOM or TOR switch temporarily reduces uplink bandwidth by an amount dependent on the number of switches within each tier.

4 Tested SAN designs

The following section describes each tested M1000e blade chassis SAN design in detail including diagrams and a table for comparison of important values such as bandwidth, maximum number of supported array members, and the host to storage port ratio. All information below assumes a single M1000e chassis and 16 half-height blade servers with two network ports each.

There are three categories of SAN designs for M1000e blade chassis integration:

1. **Blade IOM switch only** – Network ports of both the hosts and the storage are connected to the M1000e blade IOM switches. No TOR switches are required. The ISL can be a stack or a LAG, and no uplink is required.
2. **TOR switch only** – Network ports of both the hosts and the storage are connected to external TOR switches. 10 GbE pass-through IOM are used in place of blade IOM switches in the M1000e blade chassis. The ISL can be a stack or a LAG.
3. **Blade IOM switch with TOR switch** – Host network ports are connected to the M1000e blade IOM switches and the storage network ports are connected to TOR switches. An ISL stack or LAG between each blade IOM switch and/or between each TOR switch is required. An uplink stack or LAG from the blade IOM switch tier to the TOR switch tier is also required.

4.1 Blade IOM switch only

This SAN design category includes configurations in which the EqualLogic PS Series array member ports are directly connected to the blade IOM switch ports within the blade chassis. In these scenarios, dual PowerConnect M8024-k switches in the M1000e chassis were used. Two SAN designs of this type were tested:

- M8024-k switches connected with an ISL stack, and
- M8024-k switches connected with an ISL LAG

4.1.1 M8024-k switch with ISL stack

This SAN design provides 30 Gbps of ISL bandwidth between the two M8024-k switches using three 10 Gb SFP+ ports on each switch for stack connections. Since there is only a single tier of switches, there is no uplink to TOR switches. The five remaining external ports on each M8024-k (10 ports total) can accommodate the connection of five 10 GbE PS series array members, each of which require two ports for the active and passive controllers combined. The host/storage port ratio with the maximum number of array members is 6.4:1.

The following diagram illustrates how four PS6110XV array members directly connect to the two M8024-k switches in Fabric C of the M1000e blade chassis and how the two M8024-k switches are stacked using the 10 Gb SFP+ ports. This network design requires the use of a 10 GbE expansion modules in each of the M8024-k switches. Each array member controller connects to both M8024-k switches for SAN redundancy. Note that the port on the passive controller is connected to a different switch than the port on the active controller, ensuring that the port-based failover of the PS6110 array member will connect to a different switch upon port, cable or switch failure. Management and host LAN networks are shown for reference.

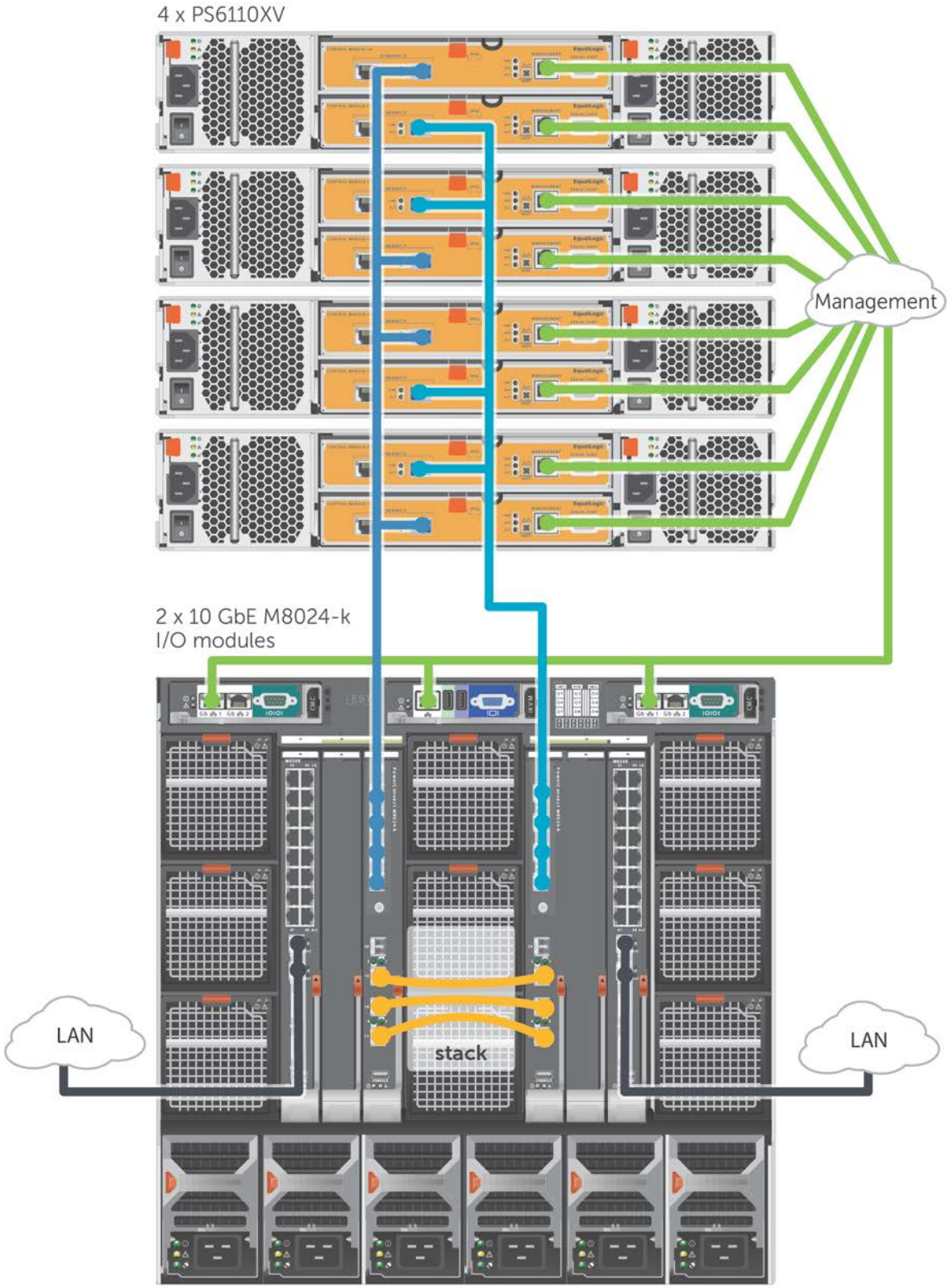

Figure 3 Blade IOM switch only with ISL stack

4.1.2 M8024-k switch with ISL LAG

This SAN design provides 30 Gbps of ISL bandwidth between the two M8024-k switches using three integrated 10 GbE SFP+ ports on each switch to create a LAG. Since there is only a single tier of switches, there is no uplink to the TOR switches. The remaining five external ports on each M8024-k (10 ports total) can accommodate the connection of five 10 GbE PS series array members, each of which require two ports for the active and passive controllers combined. The host/storage port ratio with the maximum number of array members is 6.4:1.

The following diagram illustrates how four PS6110XV array members directly connect to the two M8024-k switches in Fabric C of the M1000e blade chassis and how the two M8024-k switches are connected by a LAG using three 10 GbE SFP+ ports on each switch. This network design requires the use of 10 GbE expansion modules in each of the M8024-k switches. Note how each array member controller connects to both M8024-k switches for SAN redundancy. Also note that the port on the passive controller is connected to a different switch than the port on the active controller, ensuring that the port-based failover of the PS6110 array member will connect to a different switch upon port, cable or switch failure. Management and host LAN networks are shown for reference.

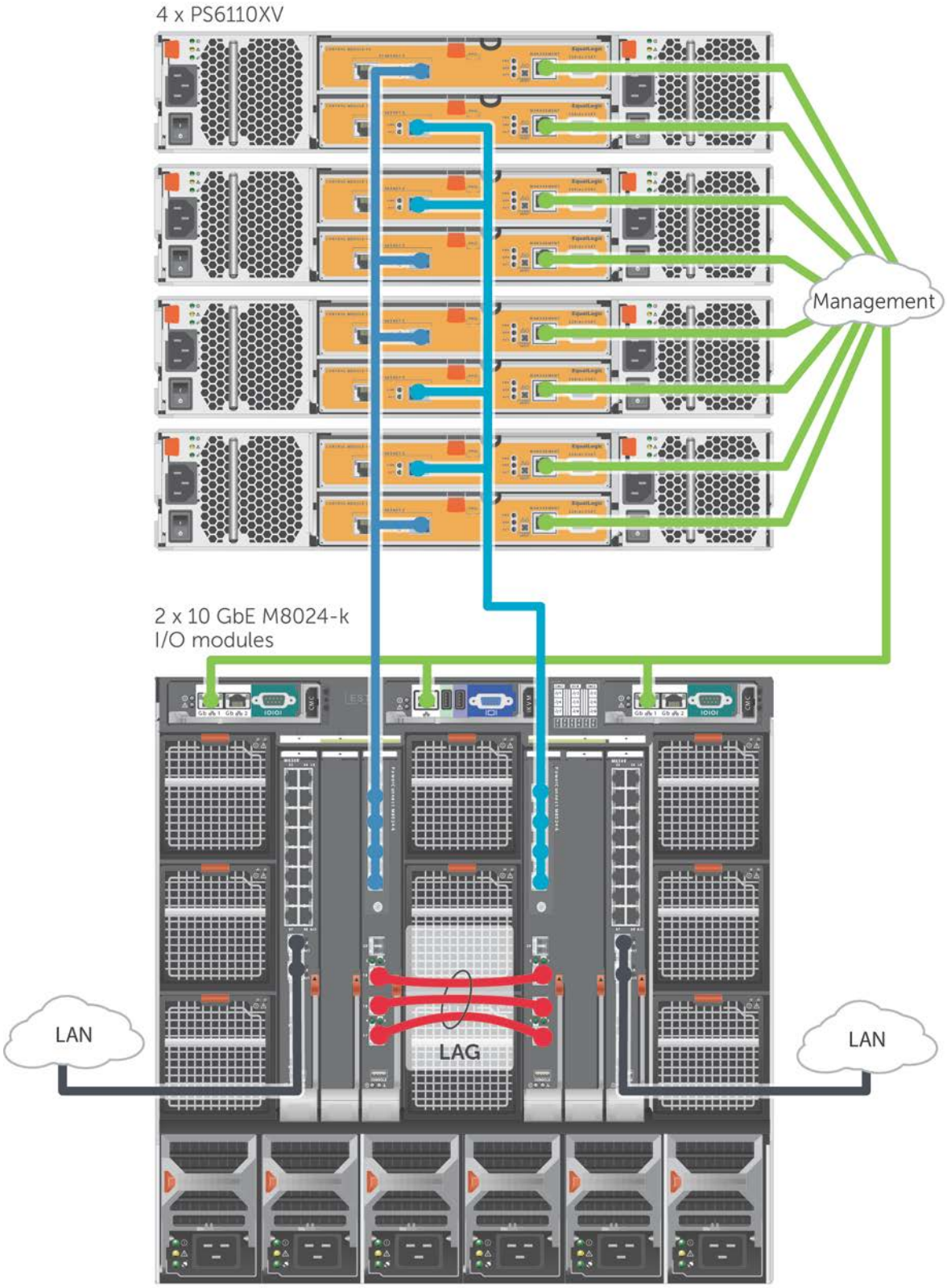

Figure 4 Blade IOM switch only with ISL LAG

4.2 TOR switch only

These SAN designs include configurations where the blade server host ports are directly connected to TOR switches using 10 GbE pass-through IOM in the M1000e blade chassis. The storage ports are also connected to the TOR switches, in this case a pair of PowerConnect switches. Two SAN designs of this type were tested:

- PC8024F switches connected with an ISL stack, and
- PC8024F switches connected with an ISL LAG

4.2.1 PC8024F switch with ISL stack

This SAN design provides 30 Gbps of ISL bandwidth between the two PC8024F switches using three 10 GbE SFP+ ports on each switch to create a stack. Since there is only a single tier of switches, there is no uplink from the blade IOM pass-through module. 16 ports on each TOR switch are required for the connections of the 16 hosts with two network ports each. The remaining five ports on each PC8024-k (10 ports total) can accommodate the connection of five 10 GbE PS series array members, each of which require two ports for the active and passive controllers combined. The host/storage port ratio with the maximum number of array members is 6.4:1.

The following diagram illustrates how four PS6110XV array members directly connect to the two TOR PC8024-k switches and how the two switches are connected by an ISL stack using three 10 GbE SFP+ ports on each switch. It also shows the connection of four server blades each with two host ports to the PC8024-k switches using the 10 GbE pass-through IOM in Fabric C of the M1000e chassis. Note how each array member controller connects to both PC8024-k switches for SAN redundancy. Also note that the port on the passive controller is connected to a different switch than the port on the active controller, ensuring that the port-based failover of the PS6110 array member will connect to a different switch upon port, cable or switch failure. Management and host LAN networks are shown for reference.

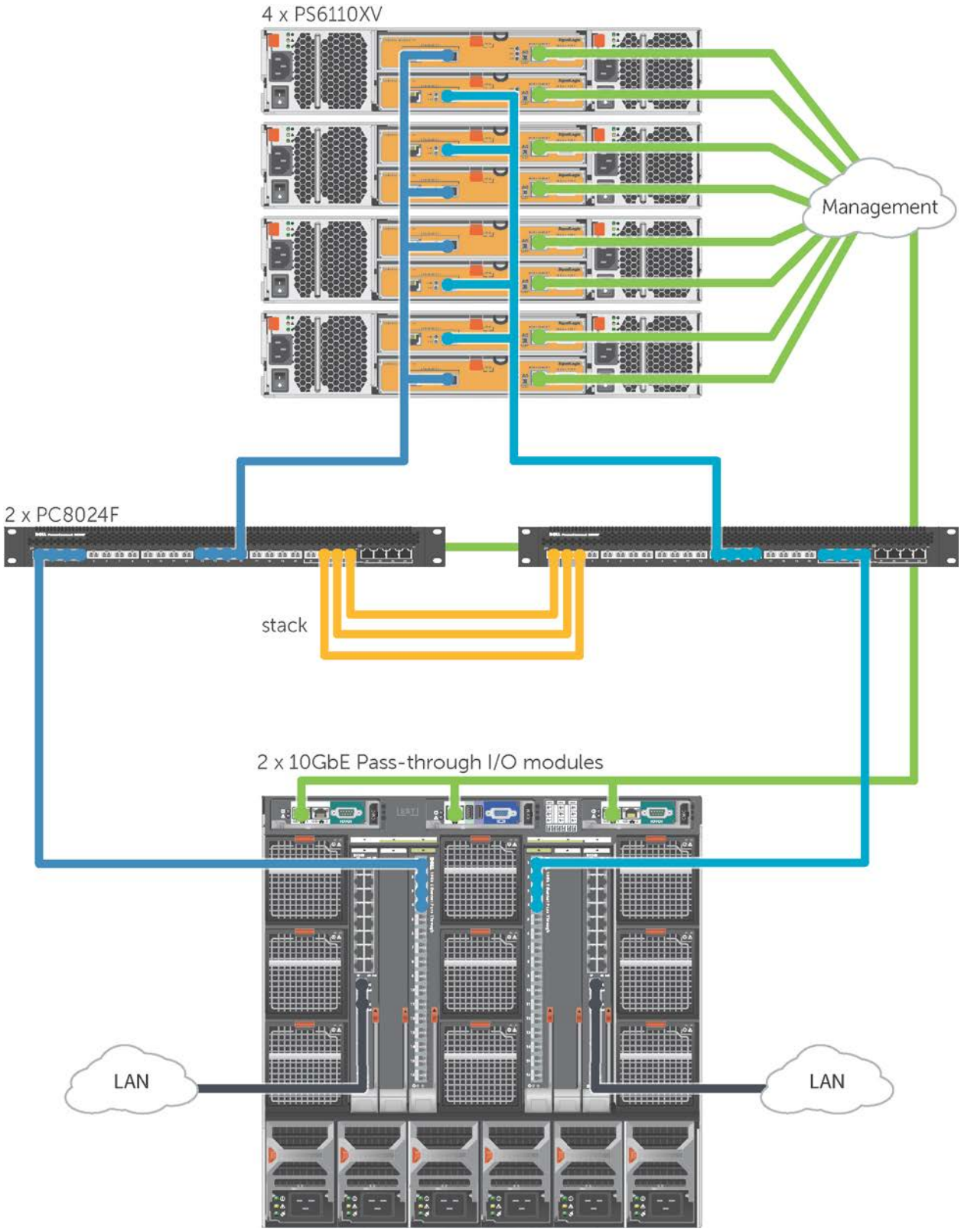

Figure 5 TOR switch only with ISL stack

4.2.2 PC8024F switch with ISL LAG

This SAN design provides 30 Gbps of ISL bandwidth between the two PC8024F switches using three 10 GbE SFP+ ports on each switch to create a LAG. Since there is only a single tier of switches, there is no uplink from the blade IOM pass-through modules. 16 ports on each switch are required for the connections of the 16 hosts with two network ports each. The remaining five ports on each PC8024-k (10 ports total) can accommodate the connection of five 10 GbE PS series array members, each of which require two ports for the active and passive controllers combined. The host/storage port ratio with the maximum number of array members is 6.4:1.

The following diagram illustrates how four PS6110XV array members directly connect to the two TOR PC8024-k switches and how the two switches are connected by an ISL LAG using three 10 GbE SFP+ ports on each switch. It also shows the connection of four server blades each with two host ports to the PC8024-k switches using the 10 GbE pass-through IOM in Fabric C of the M1000e chassis. Note how each array member controller connects to both PC8024-k switches for SAN redundancy. Also note that the port on the passive controller is connected to a different switch than the port on the active controller, ensuring that the port-based failover of the PS6110 array member will connect to a different switch upon port, cable, or switch failure. Management and host LAN networks are shown for reference.

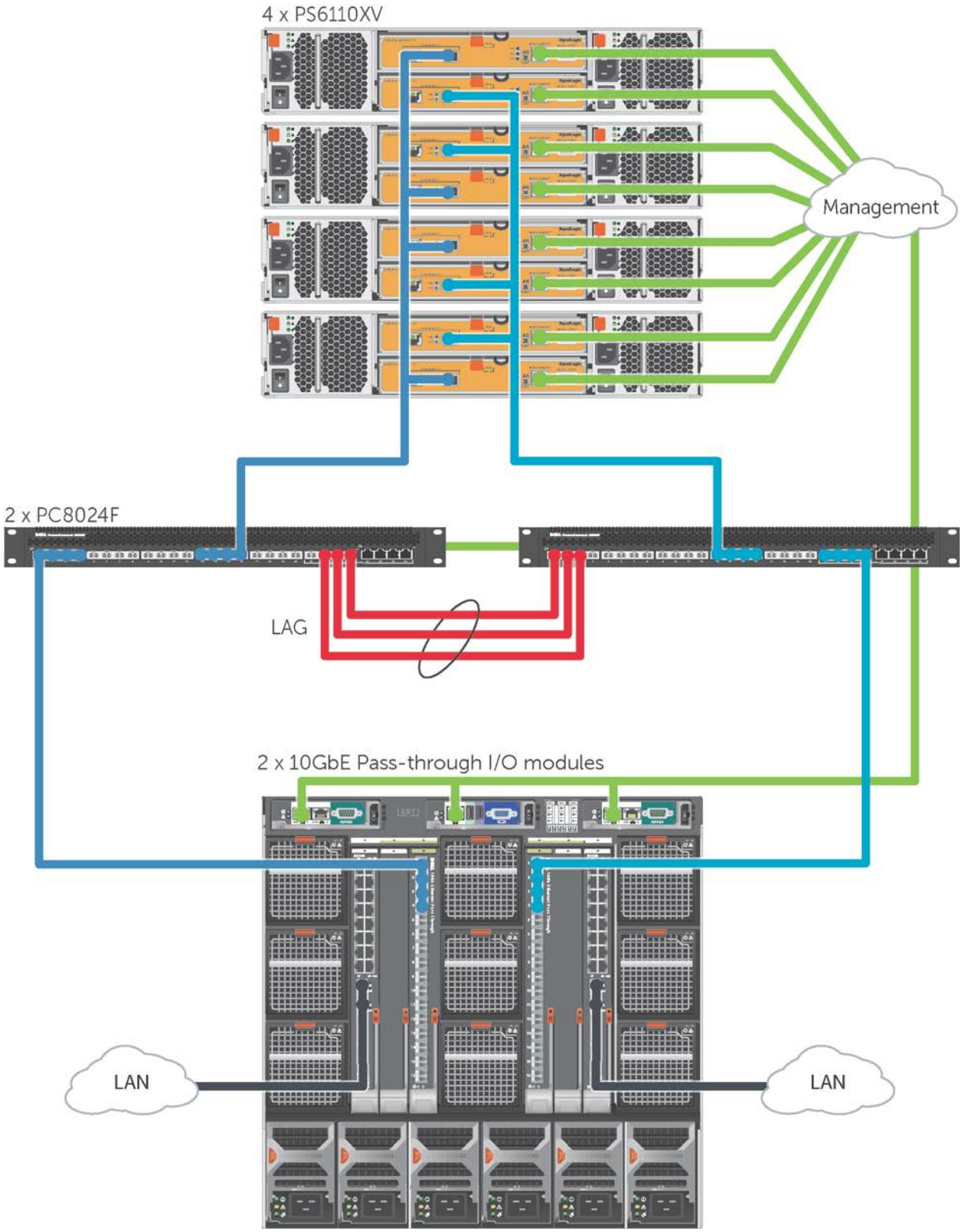

Figure 6 TOR switch only with ISL stack

4.3 Blade IOM switch with TOR switch

These SAN designs include configurations in which the EqualLogic PS Series array member ports are connected to a tier of TOR switches while the server blade host ports are connected to a separate tier of blade IOM switches in the M1000e blade chassis.

With the multiple switch tier designs it is a best practice to connect all array member ports to the TOR switches and not the blade IOM switches in the M1000e chassis. This allows the M1000e chassis to scale independently of the array members. The switches within each switch tier are connected to each other by an ISL stack or LAG. It is also a best practice to have the ISL span the TOR switches connecting the array members to better facilitate inter-array member communication. The switch tiers themselves are connected by an uplink stack or LAG.

In this case the TOR switches in the storage tier are PowerConnect 8024F and the blade IOM switches in the host tier are PowerConnect M8024-k residing in the M1000e chassis. Note that because the PC8024F switch and the M8024-k switch are not stack-compatible, SAN designs that require the uplinks to be stacked were not an option. Also, the uplink LAG with ISL stack SAN design was excluded because the available external ports on the M8024-k were not of a sufficient number to create an ISL stack of equivalent bandwidth to the TOR switch ISL stack. Therefore, only one multiple switch tier SAN design was tested.

4.3.1 M8024-k/PC8024F switches with three-way LAG

This SAN design uses the 10 GbE SFP+ ports of the PC8024F and the M8024-k to setup two separate uplink LAGs and one ISL LAG between the TOR PC8024F switches. It provides 120 Gbps of uplink bandwidth between the storage tier of PC8024F switches and the host tier of M8024F switches, while providing 30 Gbps of ISL bandwidth. Although not shown in the diagram, an additional 20 Gbps of ISL LAG could be created using the remaining external ports on the M8024-k switches. While not providing sufficient ISL bandwidth to accommodate the possible 60 Gbps of ISL traffic generated by a full twelve arrays, a second ISL LAG would prevent half of the host ports from being disconnected from the SAN in the event of a TOR switch failure. Under normal operation with both ISL LAGs active, the second ISL LAG would necessitate that Spanning Tree Protocol be enabled and would need to be assigned a higher path cost than the other 3 ISLs in this design. This ensures it would only be active if one of the other, primary ISLs were to fail.

With 12 ports on each PC8024F (24 ports total) remaining after the uplinks and ISL connections the three-way LAG design can accommodate the connection of 12 10 GbE PS 6110 array members, each of which require two ports for the active and passive controllers combined. The host/storage port ratio with the maximum number of array members is 2.67:1.

The following diagram illustrates how four PS6110XV array members connect to the two TOR PC8024F switches and how all four switches are interconnected with two uplink LAGs and one ISL LAG between the TOR switches. This network design requires the use of a 10 GbE expansion modules in each of the M8024-k switches. Each array member controller connects to both PC8024F switches for SAN redundancy. Note that the port on the passive controller is connected to a different switch than the port on the active controller, ensuring that the port-based failover of the PS6110 array member will connect to a different switch upon port, cable, or switch failure. Management and host LAN networks are shown for reference.

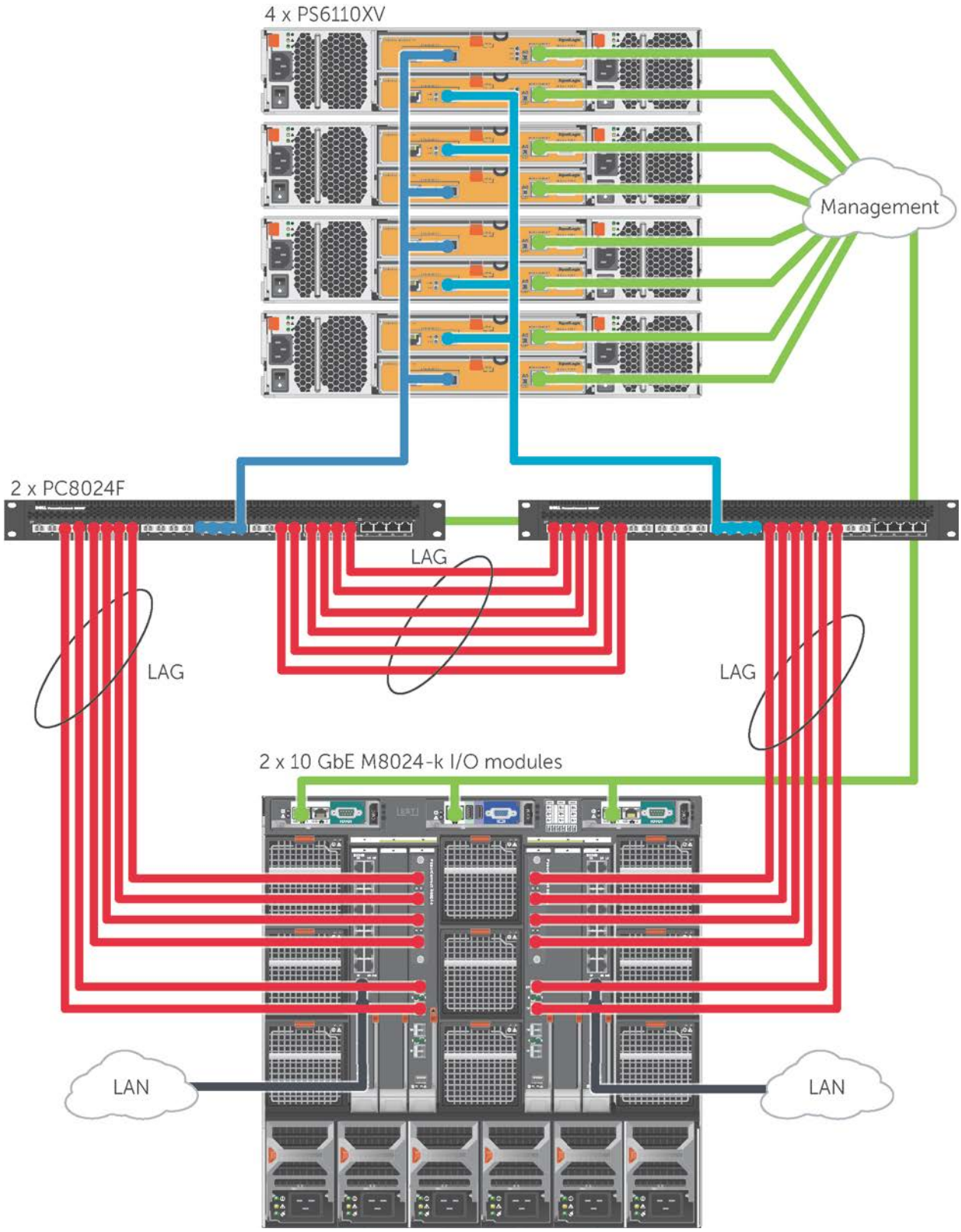

Figure 7 Blade IOM switch with TOR switch and a three-way LAG

4.4 Summary table of tested SAN designs

The following table assumes one fully populated M1000e blade chassis with 16 half-height blade servers each using two network ports (32 host ports total) and the maximum number of PS Series array members accommodated by the available ports of the array member switches -- either dual TOR PC8024F switches or dual M8024-k switches in a single M1000e blade chassis I/O Fabric.

In single switch tier designs, increasing the number of total host ports per chassis decreases the number of ports available for array member port connection. Total host ports can be increased either by increasing the number of host ports per server blade or increasing the number of blade servers per chassis.

Note: Green cells indicate the recommended SAN design within each design category based on all factors considered during testing, while orange cells indicate designs that might not be preferred.

Table 4 A comparison of all tested SAN designs

	Host switch type	Array member switch type	Total uplink bandwidth	Total ISL bandwidth	Maximum number of hosts	Maximum number of arrays members	Port ratio with maximum hosts/array members
Blade IOM with ISL stack	Blade	Blade	N/A	30 Gbps	16	5	6.4:1
Blade IOM with ISL LAG	Blade	Blade	N/A	30 Gbps	16	5	6.4:1
TOR with ISL stack	TOR	TOR	N/A	30 Gbps	16	5	6.4:1
TOR with ISL LAG	TOR	TOR	N/A	30 Gbps	16	5	6.4:1
Blade IOM and TOR with 3-way LAG	Blade	TOR	120 Gbps	60 Gbps	16	12	2.7:1

5 Detailed SAN design analysis and recommendations

The following section examines each M1000e blade chassis and EqualLogic PS Series SAN design from the perspectives of administration, performance, high availability, and scalability. In addition, SAN bandwidth, host to storage port ratios, and SAN performance and high availability test results are provided as a basis for SAN design recommendations.

5.1 Administration

In this section, SAN designs are evaluated by the ease of hardware acquisition and installation as well as initial setup and ongoing administration. Administrative tasks such as physical installation, switch configuration, and switch firmware updates play a role in determining the merits of a particular SAN design. The following paragraphs provide a list of common administration considerations and how each is affected by SAN design choice.

5.1.1 Stack vs. LAG

One characteristic that all SAN designs share is the requirement for connections between switches. Even designs with a single tier of switches, like the blade IOM only designs, will still have an ISL between switches. For multiple switch tier designs, the uplink between switch tiers needs sufficient bandwidth to prevent constraining the throughput of SAN traffic. 10 GbE SFP+ ports or proprietary stacking ports are the best solution for an ISL or an uplink and should be used whenever possible. The PC8024F switch has 24 10 GbE SFP+ ports, while the M8024-k blade IOM switch has up to eight 10 GbE SFP+ ports when using the 10 Gb SFP+ expansion module.

From an administrative perspective, a switch stack may be preferred because it allows the administration of multiple switches as if they were one physical unit. First, on the PowerConnect switches, the initial stack is defined by configuring the correct cabling and completing a few simple steps. Then, all other tasks such as enabling flow control or updating firmware must be done only once for the entire stack.

One important thing to note with this configuration is that a switch stack reset will bring down all switch units simultaneously and if switches within a tier are stacked together, the SAN becomes unavailable. The resulting SAN downtime must be scheduled.

Another important note is that the PC8024F and the M8024-k switches are not stack compatible so any stacking will have to be between like switches in a single switch tier. The uplink connections between the tier of M8024-k and PC8024F switches cannot be stacked; only LAG can be created.

The alternative inter-switch connection to the administrative switch stack is a link aggregation group (LAG). Multiple switch ports are configured to act as a single connection to increase throughput and provide redundancy, but each individual switch must still be administered separately. Creating a LAG between two PowerConnect switches is very straightforward and administrative complexity is not a concern.

5.1.2 Hardware requirements

The SAN design will determine the type and quantity of hardware and cabling required. Implementing a multiple tier switch SAN design will obviously require at least twice the number of switches as other more simple designs.

The blade IOM switch only SAN design requires the fewest cables, with only the array member ports and a single ISL stack or LAG at the M1000e chassis to cable. The blade IOM switch with TOR switch SAN designs require the addition of two uplink stacks or LAGs, and finally the TOR switch only designs (with pass-through IOM), while needing only one ISL stack/LAG, require a cable for each of the host ports; up to 32 cables for an M1000e chassis with 16 half-height blade servers with two host ports per server.

5.1.3 Using alternate switch vendors

While the choice of switches for use within an M1000e blade chassis is limited to the blade IOM product offering, TOR switches can be of any type or vendor as long as uplinks are not stacked. So for example if a SAN consisting of EqualLogic PS Series array members and an M1000e blade chassis were being deployed in a datacenter with an existing layer of non-PowerConnect switches, there are blade IOM switch with TOR switch designs and TOR switch only designs which could accommodate such a scenario. For more information on EqualLogic SAN components see the EqualLogic Compatibility Matrix at http://en.community.dell.com/techcenter/storage/w/wiki/2661.equallogic-compatibility-matrix.aspx.

Note: Multiple switch tier SAN designs that do require uplink stacks would only be possible with stack-compatible Dell TOR switches.

Also note that while setting up a LAG between two PowerConnect is very straightforward, configuring a LAG between a PowerConnect and a switch of a different vendor might be more complex. While there is an industry-standard protocol for link aggregation, LACP, some switch vendors have their own unique implementations and additional diligence might be required to ensure a properly functioning LAG. Thus even multiple tier SAN designs with an uplink LAGs require a bit more administrative planning when TOR switches of alternate switch vendors are used.

5.1.4 Recommendations

In summary, when reducing administrative overhead is the goal, a single switch tier design with an ISL stack is the simplest option. Because the storage is directly attached to the blade IOM switches, fewer cables are required than with the TOR switch only design, and the ISL stack allows the switches to be administered as a single switch.

If the availability of the SAN is critical, an ISL LAG configuration may be preferred over stacking. If a switch tier ISL is stacked, then a switch stack reload (required for tasks such as switch firmware updates) will temporarily reset the entire switch tier making the SAN unavailable during that time. In this case, SAN downtime for firmware updates would have to be scheduled. A multiple switch tier design that avoids this is the three-way LAG design.

If TOR switches from a different vendor are used, the simplest choice is to implement the TOR only design by cabling M1000e pass-through IOM directly to the TOR switches. If multiple switch tiers are desired, plan for an uplink LAG as the blade IOM switches will not be stack-compatible with the TOR switches from a different vendor.

5.2 Performance

The second criterion by which SAN designs will be evaluated is their performance relative to each other. This section reports the performance results of each SAN design under two common IO workloads.

5.2.1 Test environment

In order to determine the relative performance of each SAN design we used the performance tool vdbench to capture throughput values at three distinct I/O workloads. Vdbench is "a disk and tape I/O workload generator for verifying data integrity and measuring performance of direct attached and network connected storage. "(http://sourceforge.net/projects/vdbench/)

Each performance test was conducted with the hardware and software listed below.

> Note: All EqualLogic SAN best practices, such as enabling flow control and Jumbo frames, were implemented.
> See Appendix A for more detail about the hardware and software infrastructure.
> See Appendix B for a list of vdbench parameters.

Hosts:

- Four PowerEdge M610 blade servers each with:
  - Windows Server 2008 R2 SP1
  - Dell EqualLogic Host Integration Toolkit v4.0.0
  - Two 10 GbE ports on the SAN

Storage:

- Two EqualLogic PS6110XV array members each with:
  - Firmware: 5.2.4 R255063 H1
  - One active 1 GbE ports on the SAN

- Four iSCSI volumes dedicated to each host

> Note: There were a total of eight host ports and four storage ports for a 2:1 ratio.

The following three vdbench workloads were defined:

- 8KB transfer size, random I/O, 67% read
- 256KB transfer size, sequential I/O, 100% read
- 256KB transfer size, sequential I/O, 100% write

Each vdbench workload was run for one hour and the I/O rate was not capped (the vdbench "iorate" parameter was set to "max"). The throughput values used in the relative performance graphs are the sums of the values reported by each of the four hosts.

5.2.2 Bandwidth

All SAN designs provide different amounts of ISL bandwidth between the two switches within each switch tier. While single switch tier designs have host and storage ports connected to the same switches, multiple switch tier SAN design require an uplink stack or LAG between switch tiers. Each multiple switch tier design provides a different amount of uplink bandwidth between the host and storage switch tiers.

Uplink bandwidth should be at least equal to the aggregate bandwidth of all active PS Series array member ports. For example, four array members with one active 10 Gb port each would require 40 Gbps of uplink bandwidth. Choosing a SAN design that maximizes uplink bandwidth is of particular importance for larger scale SANs.

ISL bandwidth is also important. Since it is a best practice to create a redundant SAN Fabric with at least two switches in each switch tier, SAN traffic will often have to cross the ISL. Assuming a worst case scenario of 100% of all SAN traffic crossing the ISL in both directions (half going one way and half going the other) the ISL bandwidth requirements are 50% of the uplink bandwidth. The ISL bandwidth of each SAN design should be considered accordingly.

The following table shows the uplink and ISL bandwidth of each tested SAN design. Each of the single switch tier designs provide adequate ISL bandwidth for the maximum number of array members that their port counts accommodate. The multiple switch tier three-way LAG design provides adequate ISL and uplink bandwidth for up to 12 array members.

Note: Green cells indicate the recommended SAN design within each design category based on all factors considered during testing, while orange cells indicate designs that might not be preferred.

Table 5 A comparison of the bandwidth provided by all SAN designs

	Switch tier topology	Total uplink bandwidth	Total ISL bandwidth	Maximum number of host ports	Maximum number of array member active ports	Port ratio with maximum hosts/array members
Blade IOM only with ISL stack	Single	N/A	30 Gbps	32	5	6.4:1
Blade IOM only with ISL LAG	Single	N/A	30 Gbps	32	5	6.4:1
TOR only with ISL stack	Single	N/A	30 Gbps	32	5	6.4:1
TOR only with ISL LAG	Single	N/A	30 Gbps	32	5	6.4:1
Blade IOM and TOR with three-way LAG	Multiple	120 Gbps	60 Gbps	32	12	2.67:1

5.2.3 Results

The following three figures show the relative aggregate vdbench throughput of all four hosts within each SAN design at three different I/O workloads. Each throughput value is presented as a percentage of a baseline value. In each chart, the PC8024F with ISL LAG design was chosen as the baseline value. All throughput values were achieved during a single one hour test run.

8 KB random I/O, 67% read workload

The following figure shows the aggregate vdbench throughput of all four hosts within each SAN design at an 8 KB random I/O, 67% read workload. All SAN designs yielded throughput results within 1% of the baseline value.

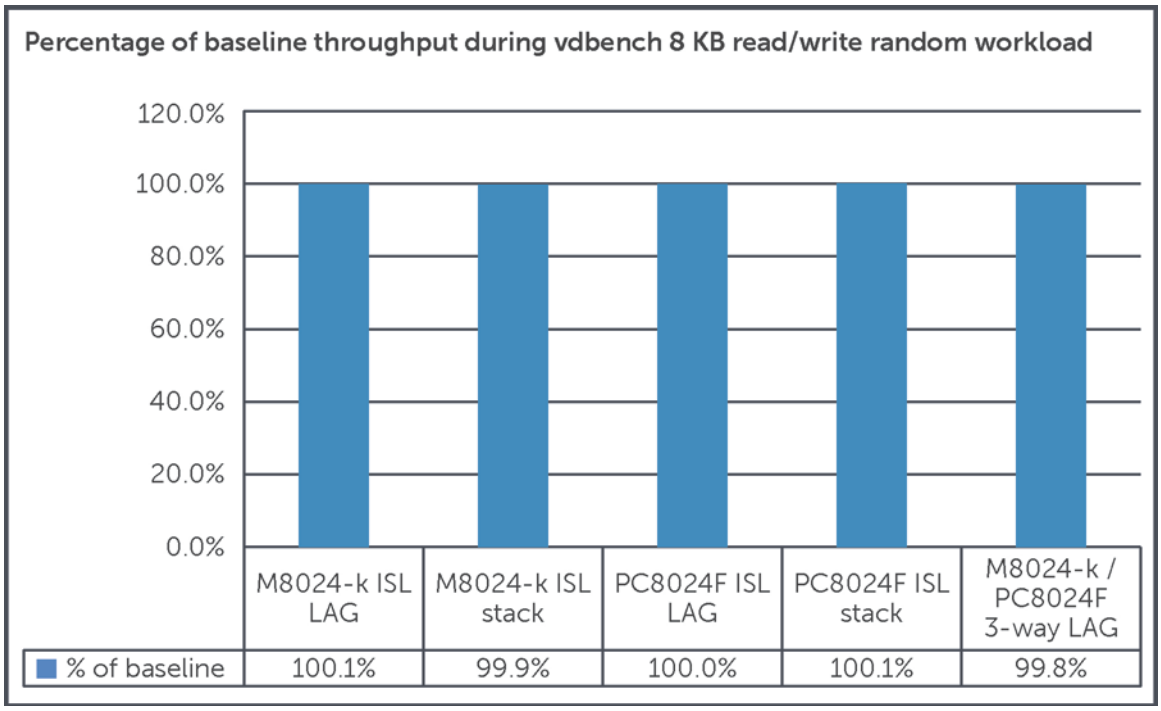

Figure 8 Aggregate vdbench throughput as a percentage of the baseline value in each SAN design during an 8 KB random I/O, 67% read workload

256 KB sequential I/O, read workload

The following figure shows the aggregate vdbench throughput of all four hosts within each SAN design at a 256 KB sequential I/O, read workload. All SAN designs yielded throughput results within 4% of the baseline value.

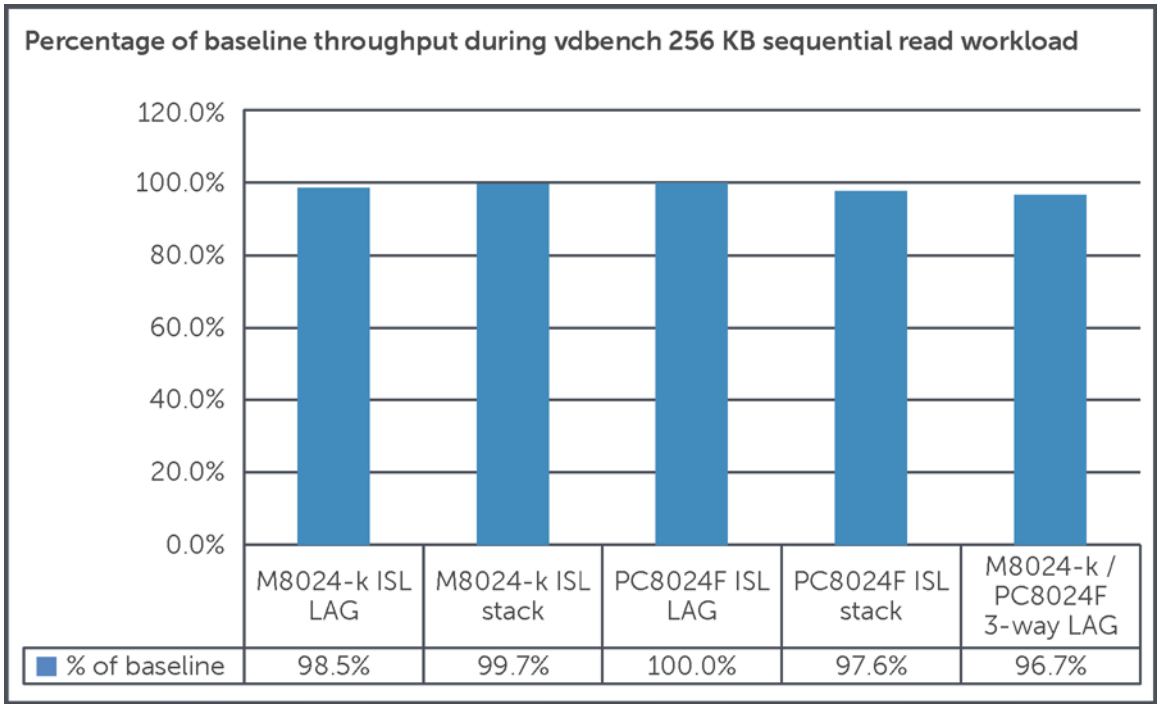

Figure 9 Aggregate Vdbench throughput as a percentage of the baseline value in each SAN design during a 256 KB sequential I/O, read workload

256 KB sequential I/O, write workload

The following figure shows the aggregate vdbench throughput of all four hosts within each SAN design at a 256 KB sequential I/O, write workload. All SAN designs yielded throughput results within 1% of the baseline value.

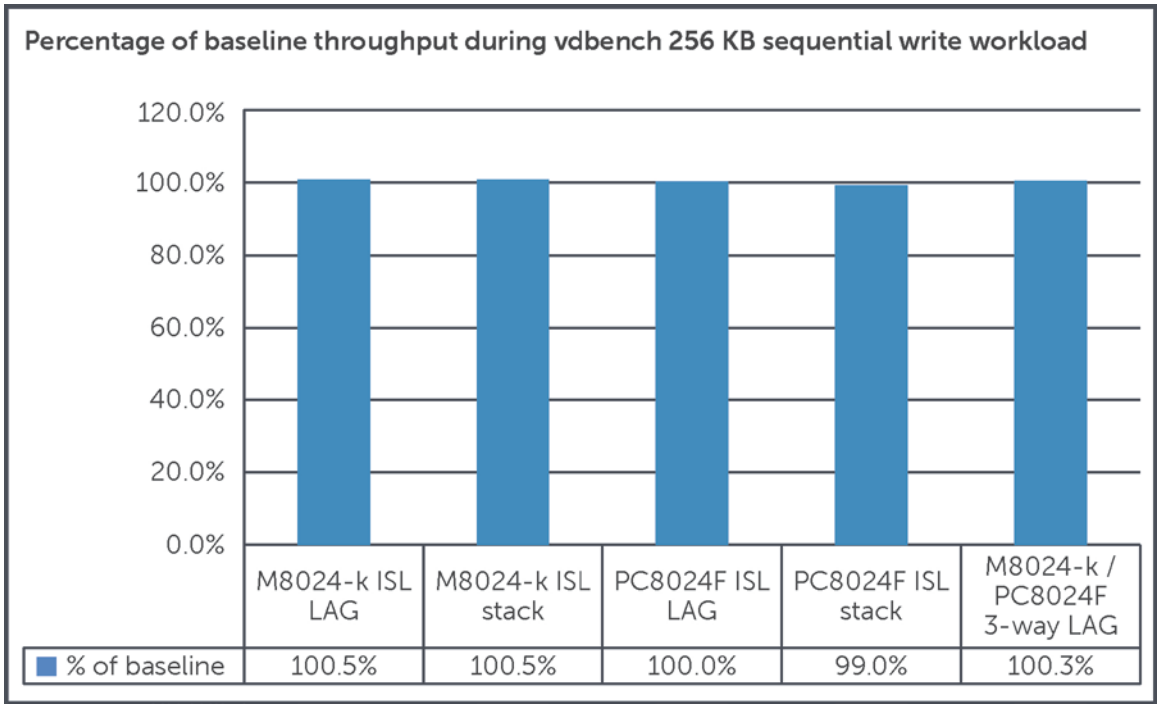

Figure 10 Aggregate Vdbench throughput as a percentage of the baseline value in each SAN design during a 256 KB sequential I/O, write workload

5.2.4 Recommendations

The throughput values were gathered during the performance testing of each SAN design with four hosts and four arrays members at three common workloads. Among all SAN designs, there were no significant performance differences during any of the three tested workloads.

5.3 High availability

The third criterion by which SAN designs will be evaluated is how each design tolerates a switch failure. This section quantifies how the loss of different switches within the SAN Fabric affects the available bandwidth and the total number of connected host ports. The results below assume a single M1000e chassis and 16 half-height blade servers with two SAN ports each for a total of 32 host ports.

Note that storage port disconnection is not addressed in the tables because the PS6110XV controller port failover ensures that no single switch failure will cause the disconnection of any array member ports. Previous generations of PS Series arrays did not have individual port failover and a single port, cable or switch failure could reduce the number of connected array member ports.

To test SAN design high availability, an ungraceful switch power down was executed while the SAN was under load. The test environment was the same as the environment that was used during performance testing, and the workload was 256 KB sequential I/O write using vdbench.

In all cases, vdbench I/O continued without error and no iSCSI volume disruptions were observed. In cases where host ports were disconnected, iSCSI connections were appropriately migrated to the remaining host ports. In these cases, since the loss of 50% of the host ports merely reduced the host/storage port ratio from 2:1 to 1:1, there was no significant reduction in vdbench throughput.

5.3.1 TOR switch failure

The following table shows how each SAN design is affected by the loss of a TOR switch. Note that this failure is not applicable to the blade IOM switch only designs in which both host and storage ports are connected to blade IOM switches.

In all applicable SAN designs, a TOR switch failure reduces the number of connected host ports by 50% and in the multiple switch tier SAN design the uplink bandwidth is also reduced by 50%. ISL bandwidth is eliminated in all applicable SAN designs as all array member ports migrate to the remaining TOR switch with which the remaining host ports have either a direct or an uplinked connection.

As discussed in Section 4.3.1, an additional 20 Gbps ISL LAG could be created using the remaining external ports on the M8024-k switches. While not providing sufficient ISL bandwidth to accommodate the possible 60 Gbps of ISL traffic generated by a full twelve arrays, a second ISL LAG would prevent half of the host ports from being disconnected from the SAN in the event of a TOR switch failure. Under normal operation with both ISL LAGs active, the second ISL LAG would necessitate that Spanning Tree Protocol be enabled and would need to be assigned a higher path cost.

Note: Green cells indicate the recommended SAN design within each design category based on all factors considered during testing, while orange cells indicate designs that might not be preferred.

Table 6 A comparison of the way each SAN designs tolerates a TOR switch failure

	Reduction in connected host ports	Reduction in uplink bandwidth	Reduction in ISL bandwidth
Blade IOM only with ISL stack	N/A	N/A	N/A
Blade IOM only with ISL LAG	N/A	N/A	N/A
TOR only with ISL stack	32-->16	N/A	30 Gbps-->N/A**
TOR only with ISL LAG	32-->16	N/A	30 Gbps-->N/A**
Blade IOM and TOR with three-way LAG	32-->16	120 Gbps-->60 Gbps	60 Gbps-->N/A**

**ISL bandwidth is no longer relevant because the switch failure eliminates the ISL. Note that this happens in conjunction with the loss of the 50% of the host ports connected to the remaining TOR switch.

5.3.2 Blade IOM switch failure

The following table shows how each SAN design is affected by the loss of a blade IOM switch. Note that this failure is not applicable to TOR switch only designs in which both host and storage ports are connected to the TOR switches.

In all applicable SAN designs, a TOR switch failure reduces the number of connected host ports by 50% and in the multiple switch tier SAN design the uplink bandwidth is also reduced by 50%. ISL bandwidth is eliminated in all single switch tier SAN designs as all array member ports migrate to the remaining TOR switch with which the remaining host ports have either a direct or an uplinked connection. However, the multiple switch tier three-way LAG design retains all ISL bandwidth.

Green cells indicate the recommended SAN design within each design category based on all factors considered during testing, while orange cells indicate designs that might not be preferred.

Table 7 A comparison of the way each SAN designs tolerates a blade IOM switch failure

	Reduction in connected host ports	Reduction in uplink bandwidth	Reduction in ISL bandwidth
Blade IOM only with ISL stack	32-->16	N/A	30 Gbps -->N/A**
Blade IOM only with ISL LAG	32-->16	N/A	30 Gbps -->N/A**
TOR only with ISL stack	N/A	N/A	N/A
TOR only with ISL LAG	N/A	N/A	N/A
Blade IOM and TOR with three-way LAG	32-->16	120 Gbps-->60 Gbps	60 Gbps-->60 Gbps

**ISL bandwidth is no longer relevant because the switch failure eliminates the ISL. Note that this happens in conjunction with the loss of the 50% of the host ports connected to the failing blade IOM switch.

5.3.3 Recommendations

In both the TOR and blade IOM switch failure scenarios, all tested SAN designs suffer a 50% reduction in the number of connected host ports and for the multiple switch tier three-way LAG SAN design, a 50% reduction in uplink bandwidth. In the case of the three-way SAN design this temporary reduction in uplink bandwidth caused by a switch failure would constrain throughput if more than six array members were in use.

5.4 Scalability

The final criterion by which SAN designs will be evaluated is scalability. Note that the scalability data presented in this section is based primarily on available port count. Actual workload, host to array port ratios, and other factors may affect performance. The three sections below will illustrate the scalability of each SAN design category when M1000e chassis and/or switches are added.

The following tables show the maximum number of array members supported by each SAN design along with the host/storage port ratios at 16 blade servers and at eight blade servers per chassis. The scalability data assumes two M8024-k switches or two pass-through IO modules per blade chassis, two SAN ports per host and, if applicable, 24-port PC8024F TOR switches.

Note that with M8024-k and PC8024F switches, a switch stack and a switch LAG use the exact same port type and bandwidth, therefore the data below does not differentiate between stack and LAG SAN designs.

Green cells indicate SAN design which generates optimal to acceptable host/storage port ratios. Orange cells represent less than acceptable host/storage port ratios.

5.4.1 Blade IOM switch only

The following table shows the scalability data for Blade IOM switch only designs where both host and storage are connected to the M1000e Blade IOM switches. It is only through additional chassis that switches can be added to this SAN design. Because of the limited number of external ports on an M8024-k switch and the increasing number of ports needed to establish the ISL connections between switches, the number of array members able to be accommodated by the remaining switch ports actually decreases with additional blade chassis, creating high host/storage port ratios. Even with the maximum five array members and only eight full-height blade servers per blade chassis this SAN design category yields a host/storage port ratio of over 3:1.

Table 8 A scalability matrix of the Blade IOM switch only SAN design

	Chassis #	Blade IOM #	TOR #	Array #	Host/storage port ratio (16 blade servers)	Host/storage port ratio (8 blade servers)
Blade IOM only	1	2	0	5	6.4:1	3.2:1
Blade IOM only	2	4	0	4	16:1	8:1
Blade IOM only	3	6	0	4	24:1	12:1

5.4.2 TOR switch only

The following two tables show the scalability data for TOR switch only designs where both host and storage are connected to PC8024F switches. Because changing the number of total host ports actually changes number of switch ports available for array members, the 16 blade server and eight blade server scalability data is presented in separate tables. Unlike the Blade IOM switch only SAN design, the number of switches can be increased without adding M1000e chassis, leading to much more optimal host/storage port ratios, particularly with eight full-height blade servers per chassis. If using 16 blade servers per blade chassis, having less than four TOR switches or having more than one blade chassis is not recommended.

An ISL connection with 50% of the uplink bandwidth should connect all TOR switches in a chain. Keep in mind that increasing the number of TOR switches will cause increased hop-counts and latency. Each storage pool should consist of array members that are all connected to the same pair of adjacent switches so that inter-array member traffic does not span more than one ISL.

Table 9 A scalability matrix of the TOR switch only SAN design with 16 blade servers per chassis

	Chassis #	Blade IOM #	TOR #	Array #	Host/storage port ratio (16 blade servers)
TOR only	1	0	2	5	6.4:1
TOR only	1	0	3	10	3.2:1
TOR only	1	0	4	12	2.7:1
TOR only	2	0	4	6	10.7:1

Table 10 A scalability matrix of the TOR switch only SAN design with 8 blade servers per chassis

	Chassis #	Blade IOM #	TOR #	Array #	Host/storage port ratio (8 blade servers)
TOR only	1	0	2	10	1.6:1
TOR only	1	0	3	14	1.1:1
TOR only	1	0	4	16	1:1
TOR only	2	0	4	12	2.7:1

5.4.3 Blade IOM and TOR switch

The following table shows the scalability data for Blade IOM and TOR switch design where hosts are connected to the M1000e Blade IOM switches and storage is connected to the PC8024F switches. Even though the TOR switches have to accommodate both ISL and uplink port connections, this multiple switch tier SAN design can still allow for a higher number of array members even with 16 blade servers per chassis.

An ISL connection with 50% of the uplink bandwidth should connect all TOR switches in a chain. Keep in mind that increasing the number of TOR switches will cause increased hop-counts and latency. Each storage pool should consist of array members that are all connected to the same pair of adjacent switches so that inter-array member traffic does not span more than one ISL.

Table 11 A scalability matrix of the Blade IOM and TOR switch only SAN design

	Chassis #	Blade IOM #	TOR #	Array #	Host/storage port ratio (16 blade servers)	Host/storage port ratio (8 blade servers)
Blade IOM and TOR with three-way LAG	1	2	2	12	2.7:1	1.3:1
Blade IOM and TOR with three-way LAG	1	2	3	14	2.3:1	1.1:1
Blade IOM and TOR with three-way LAG	1	2	4	16	2:1	1:1
Blade IOM and TOR with three-way LAG	2	4	4	16	4:1	2:1

5.4.4 Recommendations

The following recommendations assume two M8024-k switches or two pass-through IO modules per blade chassis, two SAN ports per host and, if applicable, 24-port PC8024F TOR switches.

The blade IOM switch only SAN design does not yield acceptable host/storage port ratios even with the maximum number of arrays and only eight full-height blade servers per blade chassis.

With 12 array members and four TOR switches, the TOR switch only SAN design yields an acceptable host/storage port ratio, even with 16 blade servers per blade chassis.

The blade IOM and TOR switch SAN design is the most scalable of all, allowing 16 blade servers per chassis to achieve a 2:1 host/storage port ratio with 16 array members using four TOR switches.

For additional information on multi-blade chassis SAN designs, including the number of supported array members and the maximum recommended stack size for different blade chassis IOM switches, see the Blade Server Chassis Integration section of the Dell EqualLogic Configuration Guide (ECG).

The ECG is available at: http://en.community.dell.com/techcenter/storage/w/wiki/2639.equallogic-configuration-guide.aspx

Appendix A Solution infrastructure detail

The following table is a detailed inventory of the hardware and software configuration in the test environment.

Table 12 A detailed inventory of the hardware and software configuration in the test environment

Solution configuration - Hardware components:		Description
Blade Enclosure	Dell PowerEdge M1000e chassis: CMC firmware: 4.00	Storage host enclosure
10 GbE Blade Servers	(4) Dell PowerEdge M610 server: Windows Server 2008 R2 SP1 BIOS version: 6.1.0 iDRAC firmware: 3.35 (2) Intel® Xeon® X5650 24GB RAM Dual Brocade BR1741M-k 10 GbE CNA Driver v3.0.1.0 Dell EqualLogic Host Integration Toolkit v4.0.0	Storage hosts for configs: Blade IOM switch only with ISL stack Blade IOM switch only with ISL LAG TOR switch only with ISL stack TOR switch only with ISL LAG Blade IOM switch and TOR switch with 3-way LAG
10 GbE Blade IO modules	(2) Dell PowerConnect M8024-k Firmware v4.2.2.3 10 Gb expansion module (2) Dell 10 Gb Ethernet Pass-through module	IO modules for configs: Blade IOM switch only with ISL stack Blade IOM switch only with ISL LAG TOR switch only with ISL stack TOR switch only with ISL LAG Blade IOM switch and TOR switch with 3-way LAG
10 GbE TOR switches	(2) Dell PowerConnect 8024F Firmware v4.2.2.3	Switches for configs: TOR switch only with ISL stack TOR switch only with ISL LAG Blade IOM switch and TOR switch with 3-way LAG
10 GbE Storage	(4) Dell EqualLogic PS6110XV: (24) 146GB 15K SAS disks – vHN63 (2) 10 GbE controllers Firmware: 5.2.4 R255063 H1	Storage arrays for configs: Blade IOM switch only with ISL stack Blade IOM switch only with ISL LAG TOR switch only with ISL stack TOR switch only with ISL LAG Blade IOM switch and TOR switch with 3-way LAG

Appendix B Vdbench parameters

Vdbench workloads were executed using the following parameters in the parameter file.

Common parameters:

```
hd=default

hd=one,system=localhost
```

iSCSI volumes (random IO):

```
sd=sd3,host=*,lun=\\.\PhysicalDrive3,size=256000m,threads=5

sd=sd4,host=*,lun=\\.\PhysicalDrive4,size=256000m,threads=5

sd=sd5,host=*,lun=\\.\PhysicalDrive5,size=256000m,threads=5

sd=sd6,host=*,lun=\\.\PhysicalDrive6,size=256000m,threads=5
```

iSCSI volumes (sequential IO):

```
sd=sd3,host=*,lun=\\.\PhysicalDrive3,size=1m,threads=5

sd=sd4,host=*,lun=\\.\PhysicalDrive4,size=1m,threads=5

sd=sd5,host=*,lun=\\.\PhysicalDrive5,size=1m,threads=5

sd=sd6,host=*,lun=\\.\PhysicalDrive6,size=1m,threads=5
```

8KB 67% read, random I/O workload:

```
wd=wd1,sd=(sd3-sd6),xfersize=8k,rdpct=67
```

256KB read, sequential I/O workload:

```
wd=wd1,sd=(sd3-sd6),xfersize=262144,rdpct=100,seekpct=sequential
```

256KB write, sequential I/O workload:

```
wd=wd1,sd=(sd3-sd6),xfersize=262144,rdpct=0,seekpct=sequential
```

Runtime options:

```
rd=rd1,wd=wd1,iorate=max,elapsed=3600,interval=30
```

Additional resources

Support.dell.com is focused on meeting your needs with proven services and support.

DellTechCenter.com is an IT Community where you can connect with Dell Customers and Dell employees for the purpose of sharing knowledge, best practices, and information about Dell products and your installations.

Referenced or recommended Dell publications:

- Dell EqualLogic Configuration Guide: http://en.community.dell.com/techcenter/storage/w/wiki/2639.equallogic-configuration-guide.aspx

For EqualLogic best practices white papers, reference architectures, and sizing guidelines for enterprise applications and SANs, refer to Storage Infrastructure and Solutions Team Publications at:

- http://dell.to/sM4hJT