SAN Guidelines for Maximizing Pure Performance

The SAN is a very common component in many customer issues. The aim of this article is to provide a brief overview of our suggested guidelines to achieve the best possible performance; or, alternatively, remove your SAN from the list of variables to review when troubleshooting.

Summary of ideal standards:

- Use all of FlashArray's Fibre Channel ports (If your environment allows it, given host limitations)
- Single initiator -> multi-target zoning.
- Avoid ISLs if possible. If not possible, watch for frame discards on ISLs
- Verify all paths are clean; address any CRCs or similar errors
- Use consistent ports speeds fabric wide, i.e. do not connect 2Gb to 8Gb

**Fibre Channel**

- Do not route iSCSI
- VLAN tagging is only supported in Purity 4.6.0+
- Use a MTU of 9000 across the entire path
- **Use all of the FlashArray's interfaces** (critical for iSCSI performance)
- Verify all paths are clean; address any CRCs or similar errors
- Create at least 8 sessions per host (or, again, use all interfaces on Pure)

**iSCSI**

- Use all external ports for bladed servers; these are heavily oversubscribed
- Check our [Solutions KB](#) for host-specific Best Practice Guides

**Host-Specific**
When configuring your SAN, it's important to remember that the more hops you have, the more latency you will see. For best performance, the ideal topology is a “Flat Fabric” where the FlashArray is only one hop away from any applications being hosted on it. For iSCSI, we recommend that you do not add routing to your SAN.

### Topological Bottlenecks

#### ISLS

To illustrate, we'll use an actual support case without names to protect the innocent. In this example, we were seeing terrible performance from the test host. Latency varied wildly, with several hundred milliseconds of peak latency. Bandwidth never surpassed 100MB/sec.

**Assumed topology by Pure Support and customer**

**Actual topology**

Make sure you know the topology. Please consult with your switch vendor's documentation on how to confirm your topology.

Some examples of helpful switch tools:

- **Brocade:** `topologyshow`, `fabricshow`, and `islshow`
- **Cisco:** `show interface brief` (look for `e_ports`) or `show topology`
Blade Servers

Many of our customers use a CPU chassis such as a Cisco UCS or a HP c7000. These systems commonly have a number of bladed servers that connect to an embedded switch over a copper bus. All but UCS use a type of "dumb" switch (no zoning) which connects to a core fabric switch (this is true for FC and iSCSI). UCS connects to an additional switch/bridge, a "Fabric Interconnect" and then to a core switch.

Each one of these steps increases oversubscription.

For example, a bladed chassis might have 16 discrete servers. Each of these servers connects to an internal HBA which connects to the embedded switch. This switch takes these 16 servers and performs a form of NAT, forwarding all of their traffic to a lesser number of ports; commonly 4, and as many as 16 ports. These will log into a core switch passing frames over to storage. The oversubscription rate can get quite high if you use a hypervisor for your discrete servers. Add Virtual Machines to each blade, let’s say 4 VMs per blade, and what do we end up with?

4VMs X 16 blades = 64 initiators

64 initiators share sixteen 8Gb ports. Sixteen 8Gb ports are funneled into an embedded switch with eight 8Gb external ports. We now have 8 entrance points for 64 hosts to communicate with storage, backup, virtual devices, etc. On an 8Gb switch, this is eight hosts for each 8Gb port. For daily operations, this is usually fine, but if you have several high demand systems on this chassis; a database, development systems, this configuration can behave like a bottleneck. This is the driving force behind 16Gb Fibre Channel and the coming 32Gb standard.

In one support case, each chassis only had two iSCSI connections to the core switch, providing, in the real world use, substantially less than 20Gb of bandwidth for all 64 hosts.

This configuration is particularly devastating for iSCSI. From VMware’s Best Practices [pg 14] (emphasis is mine):

For iSCSI and NFS, make sure that your network topology does not contain Ethernet bottlenecks, where multiple links are routed through fewer links, potentially resulting in oversubscription and dropped network packets. Any time a number of links transmitting near capacity are switched to a smaller number of links, such oversubscription is a possibility.

Recovering from these dropped network packets results in large performance degradation. In addition to time spent determining that data was dropped, the retransmission uses network bandwidth that could otherwise be used for new transactions.

VMware adds this additional tip:

Be aware that with software-initiated iSCSI and NFS the network protocol processing takes place on the host system, and thus these might require more CPU resources than other storage options.

The cumulative impact of additional CPU overhead is another factor when laying out your iSCSI network. In other words, err on the side of too much bandwidth instead of too little.
Physical Paths

**Applies to:** Fibre Channel, iSCSI

Assuming you have plenty of network ports, please do avail yourself of all of Pure's ports. You will need to make sure that you balance between maximizing connections to the Pure Storage FlashArray, and any host limitation you may have on the number of connections.

**Why?** Storage devices are often oversubscribed in today's SAN. By adding more physical paths, you help maintain oversubscription; you provide more pathways, more resiliency, more performance, mitigation of physical problems, and last but not least; you take better advantage of our CPU allocation.

**How to Check?**

**GUI**

Open our UI and click on SYSTEM -> Connections and you should see the below:

![Connections UI](image)

Do note that at the bottom of the connections we list our own ports as “Target Ports” and show our connection speed. This is a nice way to easily verify if you connected to some rogue port fixed at a lower speed.

**CLI**

Here’s the command syntax and output:

```
pureuser@myarray> pureport list --initiator
Initiator WWN  Initiator Portal     Initiator IQN                                           Target    Target WWN  Target Portal
-              172.28.109.37:52143  iqn.1998-01.com.vmware:vs-ucs3-b200-m2-17-4ccc62f8 CT0.ETH4  -
172.28.109.120:3260  iqn.2010-06.com.purestorage:flasharray.137a6af57d9<wbr/>4535c
-              172.28.109.37:52147  iqn.1998-01.com.vmware:vs-ucs3-b200-m2-17-4ccc62f8 CT0.ETH5  -
172.28.109.121:3260  iqn.2010-06.com.purestorage:flasharray.137a6af57d9<wbr/>4535c
-              172.28.109.37:52151  iqn.1998-01.com.vmware:vs-ucs3-b200-m2-17-4ccc62f8 CT0.ETH6  -
172.28.109.122:3260  iqn.2010-06.com.purestorage:flasharray.137a6af57d9<wbr/>4535c
```

©2018 Copyright Pure Storage. All rights reserved.
In this example, we are using an FA-450 that has 8 physical ports. Now, you don't necessarily have to zone/connect all 8 paths to each host like here; you can simply connect 4 ports (2 per controller) to each host. Alternate for other hosts. For example, if you have 20 hosts, 10 can use any 4 ports and the other 10 hosts can use the other ports.

Due to the inherent overhead in iSCSI, we recommend using all possible interfaces per host. You will also want to use at least 8 sessions per host. This happens by default when you attach a host to all 8 Pure ports. If you do not do this, or if your model of FlashArray has less than 8 ports, configure for multiple sessions per host.

Here's the main point: you can see the same Initiator IQN (and it would be Initiator WWN for Fibre Channel) across 8 ports on Pure, the "Target IQNs."

If you see a host, but it is not mapped to any target port, this means that someone configured a host, manually entered an IQN or WWN, but that host has not logged into Pure. In other words, that host is offline for some reason.

### Clean Paths

A surprisingly large number of performance cases have been resolved by replacing cables. Touching the ends of fibre optic cables, or letting them dangle in a rack (we all do it) leads to contamination. The SFP+ connections for iSCSI are *not* immune to this and are just as unforgiving.
You can clean the cable tips. But most seem to just replace the entire cable. Physical layer errors are insidious, they can destroy performance during peak loads, are often overlooked, and yet is the easiest fix for any performance problem.

The best way is to log into your switch and check for physical layer errors. For 10Gb switch vendors the reporting here varies; some report very little. But almost all 10Gb switch vendors report CRCs. Any port with physical layer errors (like CRCs) should have the cable cleaned or replaced. If that doesn't work, test/replace your switch SFPs. Avoid patch panels if at all possible, if not possible be sure to bypass it for testing.

For Fibre Channel, Cisco has limited diagnostics for physical layer errors. Brocade reports every error in two ways; per port (portstatsshow like Cisco's "show interface details") and in a full Excel like table named "porterrshow."

Porterrshow is a powerful troubleshooting tool only available through Brocade's CLI:

```
<table>
<thead>
<tr>
<th>frames</th>
<th>enc</th>
<th>crc</th>
<th>crc</th>
<th>too</th>
<th>too</th>
<th>bad</th>
<th>enc</th>
<th>disc</th>
<th>link</th>
<th>loss</th>
<th>loss</th>
<th>frjt</th>
<th>fbsy</th>
<th>c3timeout</th>
<th>pcs</th>
</tr>
</thead>
<tbody>
<tr>
<td>tx</td>
<td>rx</td>
<td>in</td>
<td>err</td>
<td>g_eof</td>
<td>shrt</td>
<td>long</td>
<td>eof</td>
<td>out</td>
<td>c3</td>
<td>fail</td>
<td>sync</td>
<td>sig</td>
<td>tx</td>
<td>rx</td>
<td>err</td>
</tr>
<tr>
<td>0:</td>
<td>51.4m</td>
<td>3.8m</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1:</td>
<td>173.0m</td>
<td>89.5m</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
```

Between the columns "frames rx" and "enc out" are your physical layer errors. In the above example, the paths are pristine.

Caveats:

1. For Cisco and Brocade, these numbers are only as good as the switch uptime *or* since stats were last cleared.
2. An older version of Cisco NX-OS didn't report any user command evidence of stats being cleared, but now does provide this under "show interface details" in newer versions.

We do collect physical layer errors on our ports and these are pulled from system statistics in hex format. Not very user-friendly to be sure and we're working to roll this data out so that you can review port errors in our UI. In the meantime, be assured that the array's physical layer errors are one of the first items we review for performance escalations and we're happy to provide these stats anytime you need it (for no additional charge, we'll also convert the data from hex to decimal).

**Port Connection Speeds**

Applies to: Fibre Channel, iSCSI

Often customers aren't aware of the rogue switch port that was hard fixed to 4Gb; or that the company's mission-critical database server, which cannot suffer any downtime, is still using 4Gb HBAs with outdated drivers. We have caught the occasional 2Gb host as well.

The item to bear in mind is that an SSD based array, with 8Gb or 16Gb HBAs, using Fibre Channel, can achieve extraordinary bandwidth with sub-millisecond latency. When you zone this to a 4Gb host there's a good chance that the host will cause back pressure, unable to keep up, forcing fame discards further down the path.
The best way to avoid performance problems with slower hosts or switches would be to:

1. Add more physical paths if possible and avail yourself multipathing, as per your hosts best practices.
2. In the event of having to use hardware 2 full port speeds from Pure (2Gb in a 8Gb SAN or 4Gb in a 16Gb SAN), you may need to fix the port speeds that Pure connects to one speed down.

Here's an example:

Let's say you just bought an FA-450 with 16Gb HBAs to put into your brand new 16Gb SAN (awesome!). The hosts you are using to migrate data over to Pure are 4Gb. The switch is now in a position to manage 16Gb speeds for Pure, but 4Gb speeds for your host. This can lead to enormous backpressure causing frame discards to Pure, and potentially to all connected devices! (Each frame discard can trigger up to 12 seconds of paused IO between devices for error recovery between ra_tov and ed_tov fabric values).

Therefore, you may need to down-clock port speeds at the switch where Pure connects to 8Gb. This should sufficiently stop discards and allow you to, at least, push IO at the host speed.

To check port speeds for Pure:

**CLI**

```
pureuser@problemchild> purehw list
Name       Status  Identify  Slot  Index  Speed       Temperature  Details
CT0        ok      off       -     0      -           -
CT0.ETH0   ok      -         -     0      1.00 Gb/s   -
CT0.ETH1   ok      -         -     1      1.00 Gb/s   -
...          
CT0.ETH8   ok      -         5     8      10.00 Gb/s  -
CT0.ETH9   ok      -         5     9      10.00 Gb/s  -
...          
CT0.FC0    ok      -         6     0      8.00 Gb/s   -
CT0.FC1    ok      -         6     1      0.00 b/s    -
CT0.FC2    ok      -         7     2      0.00 b/s    -
CT0.FC3    ok      -         7     3      8.00 Gb/s   -
```

**GUI**

Click on SYSTEM, Host Connections, and then look at the bottom of the main page. These are the speeds we've established with the switch, as well as our WWNs and IQNs which you can copy and paste if need be.
The best way to check port speeds for your switch is with the following (iSCSI left out due to the abundance of vendors):

**Brocade - switchshow**

Switchshow is a terrific command, displaying a list of all connected devices; the port speeds, port status, physical location, Fibre Channel address, etc.

```
switchshow:
... switchBeacon: OFF

Index Slot Port Address Media Speed State     Proto
==================================================<wbr/>=
  0 1 0 6a0000 id  N4  Online  FC  F-Port  50:06:0e:80:10:1a:dd:e4
  1 1 1 6a0100 id  N4  No_Light FC
  2 1 2 6a0200 id  N4  Online  FC  F-Port  52:4a:93:7e:27:89:c5:00
... 137 1 25 6a8900 id  N1  Online  FC  F-Port  50:06:0b:00:00:07:f2:f0
  138 1 26 6a8a00 id  N4  Online  FC  F-Port  21:00:00:24:ff:0d:0f:ab
  139 1 27 6a8b00 id  N2  Online  FC  F-Port  50:06:0b:00:00:39:6a:8e
  140 1 28 6a8c00 id  N2  Online  FC  F-Port  50:06:0b:00:00:39:6a:8c
  141 1 29 6a8d00 id  N4  Online  FC  F-Port  21:00:00:24:ff:02:5b:6f
  142 1 30 6a8e00 id  N4  Online  FC  F-Port  21:00:00:e0:8b:85:96:1e
  143 1 31 6a8f00 id  N4  Online  FC  F-Port  21:00:00:24:ff:0d:0e:03
  16 2 0 6a1000 id  N4  Online  FC  F-Port  50:06:0e:80:10:1a:dd:e5
```

Much of the output has been snipped, but take a look at the Speed column. The "N" before the number means that the port is set to auto-negotiate, and the following number is the speed that the device settled on.

In this example, Pure is on port 2 (we always start with a WWN of 52) and it is set to N4. So this is likely a 4Gb switch (as we are only 8 or 16Gb). Glance downward and notice the various port speeds. This customer hopefully does not intend to use the 1Gb device, as it is two generations behind 4Gb. Without using all eight FC ports on Pure, we would expect this customer to be bandwidth limited.

**Cisco - show int brief**

```
`show interface brief`

--------------------------------------------------<wbr/>-----------------------------
Interface  Vsan   Admin  Admin   Status          SFP    Oper  Oper   Port
Mode   Trunk                          Mode  Speed  Channel
Mode                                 (Gbps)
--------------------------------------------------<wbr/>-----------------------------
fcl/29  11 auto on up    swl  F    8   --
fcl/30  11 auto on up    swl  F    8   --
fcl/31  11 auto on up    swl  F    8  131
fcl/32  11 auto on up    swl  F    8  131

```

©2018 Copyright Pure Storage. All rights reserved.
Cisco reports the port speeds, but you'll have to make a note separately as to what connects to what interface (use "show flogi database" to know which WWN is connected to which Interface). In this example, all devices are connected at 8Gb.

**Zoning**

**Applies to**: Fibre Channel

Zone any single initiator to as many Pure ports as you like (for a dual fabric environment, use 4 ports through each fabric to each host port WWN).

Back in the day FC switch vendors recommended 1 host port to 1 storage port per zone. This was when a RSCN was sent to all devices and when a large switch was 32 ports. We don't recommend this, unless you have the desire and time to manage 4 to 8 zones per device.

Indeed, Brocade and Cisco no longer suggest 1 to 1 zoning:

**Brocade**: (takes you to a pdf of best practices, below quote is taken from page 14)

Zoning Recommendations
• Use single initiator single target or single initiator and multiple target zone sets. In a large fabric, zoning by single HBA requires the creation of possibly hundreds of zones; however, each zone contains only a few members. Zone changes affect the smallest possible number of devices, minimizing the impact of an incorrect zone change. This zoning philosophy is the preferred method and avoids RSCN performance concerns with multiple initiators in the same zone.

**Cisco**: (From MDS Configuration Guide)

The following guidelines must be considered when creating zone members:

• Configuring the same initiator to multiple targets is accepted.
• Configuring multiple initiators to multiple targets is not recommended.

**Jumbo Frames**

**Applies to**: iSCSI

A jumbo frame is an Ethernet frame that's larger than 1,518 bytes. The default MTU (Maximum Transmission Unit) for most devices is set to 1500. The FlashArray can support MTU up to 9000. Configuring the MTU to 9000 on the FlashArray, switch(es) and hosts will enable your environment for Jumbo Frames. In order to take advantage of the performance gains of using Jumbo Frames, you **must** enable the setting on the full path (Initiator -> Switch -> Target).

**Changing MTU on the FlashArray**

Configure Jumbo Frames using CLI or GUI by setting the MTU to 9000

**GUI**
CLI

```
pureuser@mv-sup-fa420> purenetwork setattr ct0.eth2 --mtu 9000
```

<table>
<thead>
<tr>
<th>Name</th>
<th>Status</th>
<th>Address</th>
<th>Mask</th>
<th>Gateway</th>
<th>MTU</th>
<th>MAC</th>
<th>Speed</th>
<th>Services</th>
<th>Slaves</th>
</tr>
</thead>
<tbody>
<tr>
<td>ct0.eth2</td>
<td>enabled</td>
<td></td>
<td></td>
<td>9000</td>
<td>74:86:7a:d4:e5:1a</td>
<td>1.00 Gb/s</td>
<td>iscsi</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

**Changing MTU on the Switch and Host**

Please refer to your vendor documentation on how to change the MTU. Here are some links to get you started:

- Brocade: Changing the MTU
- Cisco: Configuring the MTU
- Dell: How to configure MTU for Jumbo Frames on Dell Networking Force10 Switches
- AIX 7.1: Setting MTU Sizes
- RHEL: How do I set the MTU for my network?
- VMware: iSCSI and Jumbo Frames configuration on VMware ESXi/ESX (1007654)
- Oracle: Setting the MTU Property