EMC® ScaleIO™

v.1.3x

Technical Note
P/N 302-001-168
REV 05

April 30, 2015

This document describes best practices for maximizing performance in high-performance (more than 60,000 IOPS) ScaleIO v1.3x environments. Topics include:

- Revision history .......................................................... 2
- Tuning ScaleIO performance .............................................. 2
- Performance tuning during installation ............................... 3
- Performance tuning post-installation ................................. 3
- Additional network adjustments ........................................... 15
- Troubleshooting and getting help ........................................ 16

ScaleIO is a software-only solution that uses existing servers' local disks and LAN to create a virtual SAN that has all the benefits of external storage—but at a fraction of the cost and complexity. ScaleIO utilizes the existing local internal storage and turns it into internal shared block storage. For many workloads, ScaleIO storage is comparable to, or better than external shared block storage.
Revision history

The following table presents the revision history of this document:

<table>
<thead>
<tr>
<th>Revision</th>
<th>Date</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>June 09, 2014</td>
<td>First release of this document</td>
</tr>
<tr>
<td>02</td>
<td>September 15, 2014</td>
<td>Updates for ScaleIO v1.3x</td>
</tr>
<tr>
<td>03</td>
<td>December 1, 2014</td>
<td>Updates for ScaleIO v1.3x</td>
</tr>
<tr>
<td>04</td>
<td>January 26, 2105</td>
<td>Revised the Windows MTU procedure</td>
</tr>
<tr>
<td>05</td>
<td>April 30, 2015</td>
<td>Revised to expand scope to all high-performance environments</td>
</tr>
</tbody>
</table>

Tuning ScaleIO performance

You can improve ScaleIO system performance in terms of IOPS, latency, and bandwidth, by making environment-specific fine-tunings on the operating system, network, and ScaleIO components. This document describes these performance-related best practices.

**NOTICE**

To avoid the need to restart the system after making these configuration changes, it is strongly recommended to perform the changes before running any applications, or, minimally, before running applications with production data.

Performance tuning is very case-specific; to prevent undesirable effects, it is highly recommended to thoroughly test all changes. For further assistance, contact EMC Support.

We can separate performance tuning into during installation and post-installation activities.
Performance tuning during installation

This section describes steps to take during installation of ScaleIO to enhance performance.

The following table describes how to enhance initial performance optimization during installation. For more information, see the installation instructions in the documentation.

<table>
<thead>
<tr>
<th>Installation method</th>
<th>Do</th>
</tr>
</thead>
<tbody>
<tr>
<td>Installation Manager</td>
<td>In the CSV file, set Optimize IOPs = Yes</td>
</tr>
<tr>
<td>VMware deployment wizard</td>
<td>In the Add SDSs screen, select Optimize for Flash for each SDS.</td>
</tr>
<tr>
<td>Manual installation</td>
<td>Use the CONF=IOPS flag when installing the SDS</td>
</tr>
</tbody>
</table>

Performance tuning post-installation

This section describes steps to take after installation to enhance performance, in the following areas:

- “Tuning SDS nodes” on page 3
- “Tuning SDC nodes” on page 10
- “Additional network adjustments” on page 15

Tuning SDS nodes

You can achieve optimum performance by making the following adjustments on every SDS node. The first section is applicable to SDS nodes in all operating systems.

For all ScaleIO systems

Allocate more network memory buffers that the SDS can use for I/O. The more buffers added, the more I/Os the SDS can handle in parallel.

On every SDS node, configure the number of buffers by using the set_num_of_io_buffers command and setting its value for higher performance (default: 1, max: 10).

For higher performance, set the value to 3.
Performance tuning post-installation

**Note:** When using the replication splitter for RecoverPoint, set the value to 5.

**Syntax**

```bash
scli --set_num_of_io_buffers (--sds_id <ID> | --sds_name <NAME> | --sds_ip <IP>) --num_of_io_buffers <VALUE>
```

**Example**

```bash
scli --set_num_of_io_buffers --sds_ip 10.25.137.41 --num_of_io_buffers 5
```

**Linux**

Perform the following steps, for all NICs in the ScaleIO system.

**Note:** Prior to activating MTU settings on the logical level, you must set Jumbo frames = MTU 9000\9126 on the physical switch ports that are connected to the server. Failure to do so may lead to network disconnects and packet drops.

Refer to your relevant vendor guidelines on how to configure Jumbo Frame support.

1. Perform one of the following:
   - For non-reboot persistent configurations, type the following command:
     ```bash
     Ifconf ethx MTU 9000
     ```
   - For persistent configurations, add the following line to the file `/etc/sysconfig/network-scripts/ifcfg-eth[n]`:
     ```bash
     MTU=9000
     ```
     Type the command: `Linux:ifconfig <NIC_NAME> MTU 9000`
     To make it reboot persistent, type the command: `echo 'MTU=9000' >> /etc/sysconfig/network-scripts/ifcfg-<NIC_NAME>`
     To apply the changes, type the command: `service network restart`
     To test the command, type the command: `ping -M do -s 8972 <DESTINATION_IP_ADDRESS>`
Output similar to the following should be displayed:

```
ping -M do -s 8972 10.10.221.23
PING 10.10.221.23 (10.10.221.23) 56(40) bytes of data.
64 bytes from 10.10.221.23: icmp_seq=1 ttl=64 time=0.636 ms
64 bytes from 10.10.221.23: icmp_seq=2 ttl=64 time=0.183 ms
64 bytes from 10.10.221.23: icmp_seq=3 ttl=64 time=0.235 ms
64 bytes from 10.10.221.23: icmp_seq=4 ttl=64 time=0.128 ms
64 bytes from 10.10.221.23: icmp_seq=5 ttl=64 time=0.236 ms
64 bytes from 10.10.221.23: icmp_seq=6 ttl=64 time=0.236 ms
64 bytes from 10.10.221.23: icmp_seq=7 ttl=64 time=0.236 ms
64 bytes from 10.10.221.23: icmp_seq=8 ttl=64 time=0.151 ms
64 bytes from 10.10.221.23: icmp_seq=9 ttl=64 time=0.236 ms
64 bytes from 10.10.221.23: icmp_seq=10 ttl=64 time=0.119 ms
64 bytes from 10.10.221.23: icmp_seq=11 ttl=64 time=0.236 ms
64 bytes from 10.10.221.23: icmp_seq=12 ttl=64 time=0.154 ms
64 bytes from 10.10.221.23: icmp_seq=13 ttl=64 time=0.236 ms
64 bytes from 10.10.221.23: icmp_seq=14 ttl=64 time=0.236 ms
64 bytes from 10.10.221.23: icmp_seq=15 ttl=64 time=0.236 ms
--- 10.10.221.23 ping statistics ---
15 packets transmitted, 15 received, 0% packet loss, time 1424ms
rtt min/avg/max/mdev = 0.119/0.227/0.636/0.118 ms
```

2. Perform one of the following:
   - For persistent configurations, change the `txqueuelen` parameter to 10,000, by adding the following line to the file `rc.local`:
     ```
     ifconfig <NIC_NAME> txqueuelen 10000
     ```
   - For non-reboot persistent configurations, type the following command:
     ```
     /sbin/ifconfig <NIC_NAME> txqueuelen 10000
     ```

3. To modify the I/O scheduler of the devices, type the following on each server, for each SDS device:
   ```
   echo noop > /sys/block/<device_name>/queue/scheduler
   ```
   **Example**
   ```
   echo noop > /sys/block/sds/queue/scheduler
   ```
   **Hint:** To make these changes effective immediately, run `sysctl -p` after making the changes.

4. It is recommended to change the kernel tunables by copying the content of `/opt/emc/scaleio/sds(cfg/emc.conf into /etc/sysctl.conf (while leaving /opt/emc/scaleio/sdscfg/scaleio.conf as is).

   **Note:** You need to determine whether these parameters are beneficial for your environment.
Windows

Perform the following steps, for all NICs in the ScaleIO system.

**Note:** Prior to activating MTU settings on the logical level, you must set Jumbo frames = MTU 9000\19126 on the physical switch ports that are connected to the server. Failure to do so may lead to network disconnects and packet drops.

Refer to your relevant vendor guidelines on how to configure Jumbo Frame support.

1. Change the Maximum Transmission Unit (MTU) setting to 9,000, or the highest value that is supported by the switch and the connected nodes. First, determine the appropriate NIC name by typing the command:

   `Netsh interface ipv4 show interface`

   Output similar to the following should appear:

<table>
<thead>
<tr>
<th>Idx</th>
<th>Me</th>
<th>MTU</th>
<th>State</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>4294967295</td>
<td>connected</td>
<td>Loopback Pseudo-Interface 1</td>
</tr>
<tr>
<td>17</td>
<td>5</td>
<td>1500</td>
<td>connected</td>
<td>10G_Data</td>
</tr>
<tr>
<td>18</td>
<td>5</td>
<td>1500</td>
<td>connected</td>
<td>10G_Hgnt</td>
</tr>
</tbody>
</table>

   In this example, index 17 is the appropriate network.

2. Type the command:

   `netsh interface ipv4 set subinterface <network_ID>
   mtu=9000 store=persistent`

   where `network_ID` is the ID from the output in the previous step, in this case 17
3. In the Advanced tab of the Adapter Properties dialog for your vendor and driver, change the value of Jumbo Packet to 9000, as illustrated in the following figure:

![Advanced tab of Adapter Properties dialog](image)

4. Click OK. The network connection may disconnect briefly during this phase.

5. Test that the configuration is working, by typing the command:

   `ping -f -l 8972 <Destination_IP_Address>`

   Output similar to the following should be displayed:
Performance tuning post-installation

**Note:** Ensure that the switch supports 10 GB ethernet.

**ESX**

Perform the following steps, for all NICs in the ScaleIO system.

**Note:** Prior to activating MTU settings on the logical level, you must set Jumbo frames = MTU 9000\9126 on the physical switch ports that are connected to the server. Failure to do so may lead to network disconnects and packet drops.

Refer to your relevant vendor guidelines on how to configure Jumbo Frame support.

1. Change the Maximum Transmission Unit (MTU) setting to 9,000 on the vSwitches and on the SVM (be sure to make the change in /etc/sysconfig/network/ifcfg-ethX):
   a. Type the command:
   
   ```
   esxcfg-vswitch -m 9000 <vSwitch>
   ```
   b. Create VMKernel with Jumbo Frames support by typing the following commands:
   
   ```
   1. esxcfg-vswitch -d
   2. esxcfg-vswitch -A vmkernel# vSwitch#
   3. esxcfg-vmknic -a -i <ip address> -n <netmask> -m 9000 <portgroup name>
   ```
   
   **Note:** If you change JumboFrames on an existing vSwitch, the packet size for VMKernel does not change, and therefore, the existing VMKernel must be deleted and a new one must be created.

2. Double the CPU and memory assigned to the SVM. In general, 4 CPUs and 4GB of memory is sufficient, but this may vary in your environment.
3. From the **Resources** tab of the **Virtual Machine Properties** window, select **Reserve all guest memory (All locked)**.

**Note:** In addition to the above settings, every SVM is also a Linux machine, and should *also* be tuned according to the SDS and SDC settings described for Linux.
Performance tuning post-installation

Tuning SDC nodes

You can achieve optimum performance by making the following tunings on every SDC node, in both physical and virtual environments.

Linux

1. Edit the file `/etc/init.d/scini` on each SDC node, by adding the following parameters in the `/sbin/insmod $DRV_BIN` line. Adjust the parameter values to the needs of your workload.

   - `netConSchedThrd=8`  `netSockSndBufSize=4194304`
   - `netSockRcvBufSize=4194304`  `mapTgtSockets=4`

   After the editing, the line should look similar to this:

   ```bash
   /sbin/insmod $DRV_BIN netConSchedThrd=8
   netSockSndBufSize=4194304  netSockRcvBufSize=4194304  mapTgtSockets=4
   ```

2. Restart the service by typing the following command:

   ```bash
   service scini restart
   ```

   **NOTICE**

   If this command does not succeed in restarting the service, it could be because an application or file system is using the storage. Unmount the potential user and try again. If you are still unsuccessful, you will need to restart the machine, as directed at the end of this procedure.
3. It is recommended to change the kernel tunables by copying the content of  
/opt/emc/scaleio/sdc/cfg/emc.conf into /etc/sysctl.conf (while  
leaving /opt/emc/scaleio/sdc/cfg/scaleio.conf as is).

**Note:** You need to determine whether these parameters are beneficial for your  
environment.

**Note:** Prior to activating MTU settings on the logical level, you must set  
Jumbo frames = MTU 9000\9126 on the physical switch ports that are connected  
to the server. Failure to do so may lead to network disconnects and packet drops.

Refer to your relevant vendor guidelines on how to configure Jumbo Frame  
support.

4. For all NICs in the ScaleIO system, perform the following (if the SDC resides on the  
same host as an SDS where you have already configured these parameters,  
proceed to the next step):

   a. Perform one of the following:

      – For non-reboot persistent configurations, type the following command:
        ```bash
        ifconf ethx MTU 9000
        ```

      – For persistent configurations, add the following line to the file  
        /etc/sysconfig/network-scripts/ifcfg-eth[n]:

        ```
        MTU=9000
        ```

        Type the command:  
        ```bash
        Linux: ifconfig <NIC_NAME> MTU 9000
        ```

        To make it reboot persistent, type the command:  
        ```bash
        echo 'MTU=9000' >> /etc/sysconfig/network-scripts/ifcfg-<NIC_NAME>
        ```

        To apply the changes, type the command:  
        ```bash
        service network restart
        ```

        To test the command, type the command:  
        ```bash
        ping -M do -s 8972 <DESTINATION_IP_ADDRESS>
        ```
b. Perform one of the following:

- For persistent configurations, change the `txqueuelen` parameter to 10,000, by adding the following line to the file `rc.local`:

  ```
  ifconfig <NIC_NAME> txqueuelen 10000
  ```

- For non-reboot persistent configurations, type the following command:

  ```
  /sbin/ifconfig <NIC_NAME> txqueuelen 10000
  ```

  **Note:** The previous changes will take effect on the next restart. To make the changes effective immediately, also run the commands from the Linux shell.

5. Restart the SDC node.

   Restarting is only necessary if the service did not restart in step 2 on page 10.

**VMware**

From ScaleIO v1.31, the SDC can be installed either directly on ESX (preferred method), or on an SVM (ScaleIO Virtual Machine). If you install the SDC directly on ESX, after the SDC is installed, type the following esxcli command:

```
esxcli system module parameters set -m scini -p "netConSchedThrd=4 mapTgtSockets=4 netSockRcvBufSize=4194304 netSockSndBufSize=4194304"
```

Furthermore, if you issue this command, ESX will delete other existing parameters. Therefore, the SDC GUID and MDM IP address should be provided as part of the same command.

For example:

```
esxcli system module parameters set -m scini -p "netConSchedThrd=4 mapTgtSockets=4 netSockRcvBufSize=4194304 netSockSndBufSize=4194304
IoctlIniGuidStr=12345678-90AB-CDEF-1234-567890ABCDEF
IoctlMdmIPStr=192.168.144.128"
```

To increase per device queue length (which can be lowered by default by ESX to 32), type the following esxcli command:

```
esxcli storage core device set -d <DEVICE_ID> -O <QUEUE_LENGTH>
```

where `<QUEUE_LENGTH>` can be number in the range 32-256 (default=32).
For example:

```
esxcli storage core device set -d
eui.16bb852c56d3b93e3888003b00000000 -O 256
```

**Windows**

Perform the following steps, for all NICs in the ScaleIO system.

**Note:** Prior to activating MTU settings on the logical level, you must set Jumbo frames = MTU 9000\9126 on the physical switch ports that are connected to the server. Failure to do so may lead to network disconnects and packet drops.

Refer to your relevant vendor guidelines on how to configure Jumbo Frame support.

1. Change the Maximum Transmission Unit (MTU) setting to 9,000, or the highest value that is supported by the switch and the connected nodes (if the SDC resides on the same host as an SDS where you have already configured these parameters, proceed to step 6 on page 15). First, determine the appropriate NIC name by typing the command:

```
Netsh interface ipv4 show interface
```

Output similar to the following should appear:

<table>
<thead>
<tr>
<th>Idx</th>
<th>Met</th>
<th>MTU</th>
<th>State</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>4294967295</td>
<td>connected</td>
<td>Loopback Pseudo-Interface 1</td>
</tr>
<tr>
<td>17</td>
<td>5</td>
<td>1500</td>
<td>connected</td>
<td>10G_Data</td>
</tr>
<tr>
<td>18</td>
<td>5</td>
<td>1500</td>
<td>connected</td>
<td>10G_Hgmt</td>
</tr>
</tbody>
</table>

In this example, index 17 is the appropriate network.

2. Type the command:

```
netsh interface ipv4 set subinterface <network_ID>
mtu=9000 store=persistent
```

where *network_ID* is the ID from the output in the previous step, in this case 17
3. In the **Advanced** tab of the **Adapter Properties** dialog for your vendor and driver, change the value of **Jumbo Packet** to **9000**, as illustrated in the following figure:

![Image of Advanced tab in Adapter Properties dialog](image)

4. Click **OK**. The network connection may disconnect briefly during this phase.

5. Test that the configuration is working, by typing the command:

   ```
   ping -f -l 8972 <Destination_IP_Address>
   ```

   Output similar to the following should be displayed:

   ![Ping output example](image)
**Note:** Ensure that the switch supports 10 GB ethernet.

6. Edit the SDC registry

   HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\services\scini\Parameters as follows:
   
   - "mapTgtSockets"=dword:00000004
   - "netSockRcvBufSize"=dword:04194304
   - "netConSchedThrd"=dword:00000008

7. Restart the host machine.

   For a description of the parameters, see “Linux” on page 10.

8. In the **Controllers** section of the **Windows Device Manager**, disable and then enable the ScaleIO HBA device.

   **NOTICE**

   If the device does not restart, it could be because an application or file system is using the storage. Unmount the potential user and try again. If this does not work, restart the machine.

---

### Additional network adjustments

The following network adjustments can also be performed, to increase network performance.

**Windows**

Disable delayed ack by modifying the following `REG_DWORD` to 1 on all ScaleIO network interfaces:

```
HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\Tcpip\Parameters\Interfaces\Interface GUID\TcpAckFrequency
```
Troubleshooting and getting help

EMC support, product, and licensing information can be obtained as follows:

**Product information** — For documentation, release notes, software updates, or information about EMC products, go to EMC Online Support at:

https://support.emc.com

**Technical support** — Go to EMC Online Support and click Service Center. You will see several options for contacting EMC Technical Support. Note that to open a service request, you must have a valid support agreement. Contact your EMC sales representative for details about obtaining a valid support agreement or with questions about your account.

Copyright © 2015 EMC Corporation. All rights reserved. Published in the USA.

Published April 30, 2015

EMC believes the information in this publication is accurate as of its publication date. The information is subject to change without notice.

The information in this publication is provided as is. EMC Corporation makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose. Use, copying, and distribution of any EMC software described in this publication requires an applicable software license.

EMC², EMC, and the EMC logo are registered trademarks or trademarks of EMC Corporation in the United States and other countries. All other trademarks used herein are the property of their respective owners.

For the most up-to-date regulatory document for your product line, go to EMC Online Support (https://support.emc.com).