

# ScaleIO Architecture Description

Cloud Execution Environment

SYSTEM ARCHITECTURE DESCRIPTION

**Copyright**

© Ericsson AB 2016-2018. All rights reserved. No part of this document may be reproduced in any form without the written permission of the copyright owner.

**Disclaimer**

The contents of this document are subject to revision without notice due to continued progress in methodology, design and manufacturing. Ericsson shall have no liability for any error or damage of any kind resulting from the use of this document.

**Trademark List**

All trademarks mentioned herein are the property of their respective owners. These are shown in the document Trademark Information.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Scope	1
<b>2</b>	<b>Architectural Overview</b>	<b>1</b>
2.1	MDM Cluster	3
2.2	SDS	4
2.3	SDC	5
2.4	ScaleIO GW	5
2.5	LIA	5
<b>3</b>	<b>ScaleIO System Logical Structure</b>	<b>6</b>
3.1	Protection Domain	6
3.2	Storage Pool	6
3.3	Fault Set	7
<b>4</b>	<b>ScaleIO Networking</b>	<b>8</b>
4.1	Frontend Storage Network	8
4.2	Backend Storage Network	9
4.3	Management Networks	10
4.4	Network Configuration Requirements	10
<b>5</b>	<b>Security</b>	<b>11</b>
<b>6</b>	<b>Fault Management</b>	<b>12</b>
<b>7</b>	<b>ScaleIO GUI</b>	<b>12</b>
<b>8</b>	<b>Logging</b>	<b>12</b>
<b>9</b>	<b>Maintenance</b>	<b>12</b>
<b>10</b>	<b>Upgrade Process</b>	<b>13</b>





# 1 Introduction

This document describes the architecture of EMC<sup>2</sup> ScaleIO 2.0 on Ericsson Cloud Execution Environment (CEE). ScaleIO is a software defined storage solution for distributed external block storage, designed and implemented with enterprise-grade resilience. In CEE, ScaleIO operates as a Cinder backend and is available in a Value Pack (VP).

In CEE, ScaleIO can be used in **managed** or **unmanaged** configuration

In **managed** configuration, a ScaleIO cluster is deployed during installation according to the set configuration options, on dedicated servers within the CEE region.

In **unmanaged** configuration, CEE uses a ScaleIO cluster which is already deployed outside of the CEE region, and is not managed by CEE.

As software defined storage, ScaleIO is hardware agnostic. The software works efficiently with various types of disks, including the following:

- Magnetic Hard Disk Drives (HDDs)
- Solid State Drives (SSDs)
- Flash PCI Express (PCIe) cards

## 1.1 Scope

This document provides a high level overview of the ScaleIO distributed storage solution used within CEE. The purpose of this document is to provide a general description of ScaleIO, both in managed and unmanaged configurations, when used in CEE.

# 2 Architectural Overview

In CEE, ScaleIO is deployed in a two-layer configuration, that is, a dedicated set of servers are used to create the distributed storage solution.

The ScaleIO solution consists of the following main software components:

- Meta Data Manager (MDM)
- ScaleIO Data Server (SDS)



- ScaleIO Data Client (SDC)
- ScaleIO Gateway (GW)
- Light Installation Agent (LIA)

These software components are described in the relevant section of the document [Dell EMC ScaleIO Version 2.x User Guide](#).

In CEE with managed ScaleIO, the main software components of ScaleIO are deployed in the following way:

- MDM cluster deployment consists of five members, hosted on five dedicated servers. As a consequence, a minimum of five dedicated servers are required for the use of ScaleIO in CEE.
- SDSs are deployed on additional disks dedicated for ScaleIO tenant data storage.
- SDC is a kernel module, installed on all hosts required to have access to the ScaleIO storage, that is, all compute hosts and vCIC hosts.
- CEE deployment supports one or more ScaleIO GWs.
- HAProxy is used to provide resilient access to MDM and ScaleIO GW management interfaces.

In CEE deployed with unmanaged ScaleIO solution, only the SDC and HAProxy components are handled by CEE. The remaining components are deployed on dedicated hardware outside of the CEE region. Therefore, CEE deployment requires the following data:

- ScaleIO GW IP
- ScaleIO GW username
- ScaleIO GW password

The prerequisites of CEE deployment with unmanaged ScaleIO are the following:

- ScaleIO cluster is deployed by the customer (SDSs, MDM cluster, ScaleIO GW).
- ScaleIO GW IP is accessible from the CEE deployment.
- Network connectivity is established between CEE and ScaleIO cluster.
- CEE storage frontend and backend networks are configured towards the ScaleIO cluster.

ScaleIO unmanaged installation includes the following:

- SDC kernel module installation in CEE hosts
- SDC configuration on each CEE host physical server



- OpenStack configured to use ScaleIO as volume backend

Figure 1 shows the architectural overview of CEE with the ScaleIO VP in a two-layer hardware configuration, running on dedicated servers with two pairs of 10 GE ports.

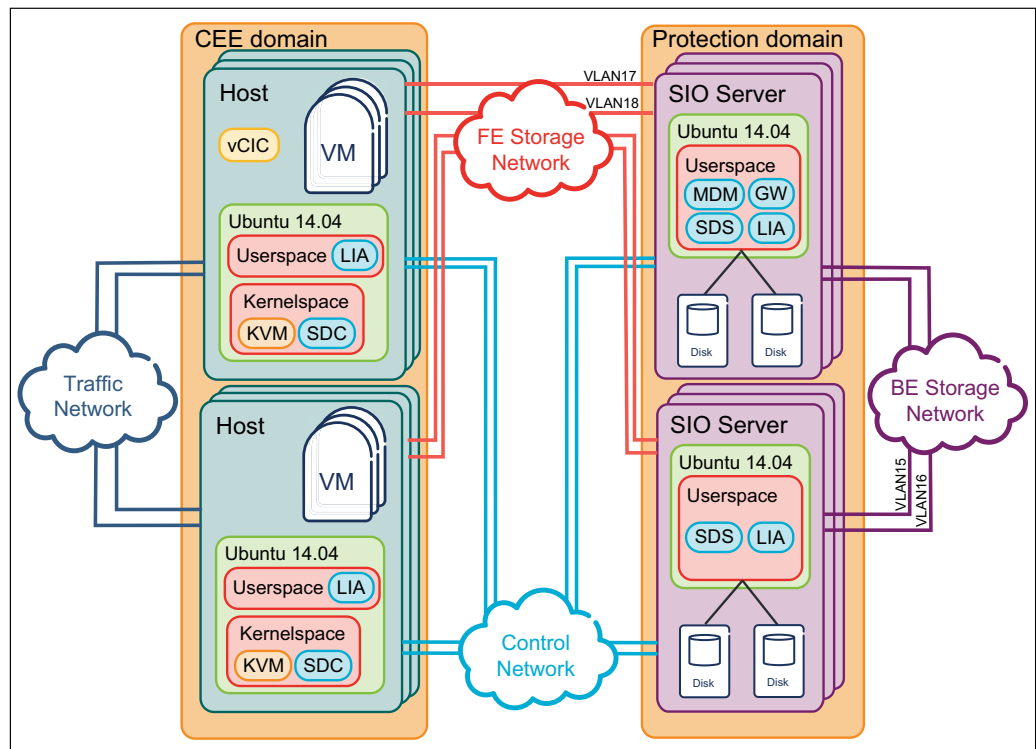


Figure 1 Architectural Overview of CEE with ScaleIO

## 2.1 MDM Cluster

The MDM is the monitoring and configuration agent of the ScaleIO system. The main purpose of the MDM is to manage ScaleIO, which includes rebuilds, migration, and all system-related functions. No I/O traffic goes through the MDM.

Any server with the MDM package installed can be used as MDM.

In CEE with managed ScaleIO, high availability is achieved through running five MDMs on separate dedicated ScaleIO servers, thus providing a system resilient to two simultaneous failures. Hosts can be assigned to the running MDM service at the time of CEE deployment.

In CEE with unmanaged ScaleIO, as this component is not deployed or managed by CEE, MDM cluster configuration is the responsibility of the user.

MDMs have one of the following roles:

- Master



- Slave
- Tie-Breaker (TB)
- Standby

Master and Slave MDMs are also referred to as manager MDMs.

The TB determines which manager MDM is the master.

A five-node MDM cluster consists of one master MDM, two slave MDMs and two TBs. If the number of MDMs defined at the time of deployment is more than five, the extra MDMs and TBs get standby MDM role. The maximum number of standby MDMs is eight.

Figure 2 shows the five-node MDM cluster with standby MDMs.

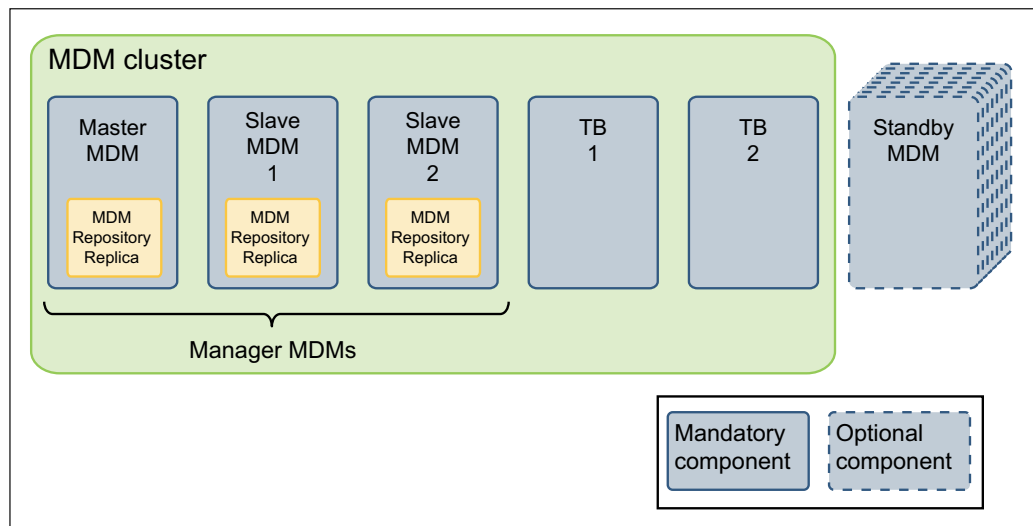


Figure 2 MDM Cluster

In CEE with unmanaged ScaleIO, changes to the MDM cluster must be propagated in the CEE region manually. For more information, refer to the relevant section of the document [Runtime Configuration Guide](#).

## 2.2 SDS

The purpose of the SDS is to perform backend I/O operations requested by an SDC.

In CEE with managed ScaleIO, SDSs run on dedicated ScaleIO hosts. The SDS contributes its own local storage to the ScaleIO storage pool. A minimum of three SDS servers with a minimum of 100 GB storage capacity each are required for each ScaleIO storage pool. Each SDS host can be installed, removed, and upgraded independently. SDS hosts can be assigned to a ScaleIO Storage Pool during CEE deployment or later through region expansion. For more information, refer to the document [Region Expansion](#).





In CEE with unmanaged ScaleIO, as this component is not deployed or managed by CEE, SDS configuration is the responsibility of the user. However, some changes to the SDSs must be propagated in the CEE region manually. For more information, refer to the relevant section of the document [Runtime Configuration Guide](#).

## 2.3 SDC

The SDC is a lightweight block device driver that exposes the shared ScaleIO volumes to applications. The SDC runs on the same server as the application, and communicates with other nodes over the TCP/IP-based protocol.

SDCs can access only specific volumes on a ScaleIO system that are mapped to that particular SDC. SDC presents the volumes to the OS as standard block devices which appear as `/dev/scini<x>`, where `x` is a letter, starting from `a`, for example: `/dev/scinia`, `/dev/scinib`. The maximum number of partitions for a `scini` disk is 15.

In CEE, both with managed and unmanaged ScaleIO, SDCs are deployed automatically on each vCIC and compute host at the time of CEE deployment. SDCs use two separate networks to connect to the ScaleIO system and handle all the multipath-related functions.

**Note:** SDC mapping is similar to LUN mapping, as it only allows volume access to clients that are explicitly mapped to the volume. The amount of bandwidth and storage available for each SDC and for each volume can be configured through CLI and REST API.

## 2.4 ScaleIO GW

The ScaleIO GW provides a REST API to expose the following:

- ScaleIO system monitoring
- ScaleIO system provisioning
- SNMP trap sender functionality

The REST server is installed as part of the ScaleIO GW.

## 2.5 LIA

The LIA component is required for upgrade and maintenance operation, and does not take part in the normal operation of the ScaleIO system.

LIA is installed during CEE deployment, on each dedicated ScaleIO host.



## 3 ScaleIO System Logical Structure

The physical layer and the virtualization layer in a ScaleIO system are linked by the following elements:

- Protection domains, see Section 3.1 on page 6
- Storage pools, see Section 3.2 on page 6
- Fault sets, see Section 3.3 on page 7

For more information, refer to Dell EMC ScaleIO Version 2.x User Guide.

These elements must be taken into consideration when configuring a ScaleIO system.

### 3.1 Protection Domain

A protection domain is a logical entity that includes a group of SDSs that mutually provide backup for each other. Each SDS belongs exclusively to one protection domain, therefore each protection domain is a unique set of SDSs. The maximum recommended number of SDSs for each protection domain is 100.

When CEE is deployed with managed ScaleIO, at least one protection domain must be defined.

### 3.2 Storage Pool

A storage pool is a set of physical storage devices in a protection domain. Storage pools enable forming different storage tiers within the same ScaleIO system. Each storage device belongs exclusively to one storage pool.

When CEE is deployed with managed ScaleIO, at least one storage pool must be defined.

When a volume is created on the virtualization layer, it is distributed over all storage devices of the storage pool. Each block is stored in two copies on different SDSs, providing data availability in case a single storage device fails. Data can be available even after multiple storage device fail, if each failure occurs in a separate storage pool.

To provide consistent performance, it is recommended that each storage device in a storage pool has similar storage properties, for example size or read and write speed. Mixing storages with different properties or mixing different types of storage media (for example, HDDs and SSDs) within the same storage pool is supported but not recommended, as performance is limited to the least-performing member of the storage pool, due to the distribution of data.



Large disk size differences (for example, if one disk of a storage pool is the same size as the other disks combined) can cause suboptimal performance.

Figure 3 shows an example configuration of two storage pools in a protection domain with five SDSs, each including an HDD and an SSD storage device.

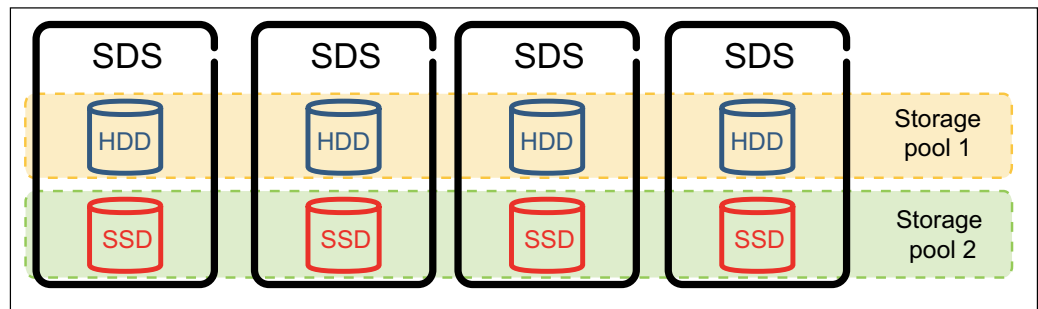


Figure 3 Example Storage Pool Configuration

### Zero Padding

Zero padding ensures that every read on a previously non-written area returns zeros. Zero padding also ensures that reading does not return previously deleted data. Some applications that depend on that reading return consistent data or zeros. If zero padding is not enabled, a read from a previously unwritten area returns unknown content that can change on subsequent reads.

Each storage pool can be configured with zero padding enabled or disabled. Zero padding can be selected for each storage pool independently at the time of storage pool creation, for example, at CEE with managed ScaleIO deployment. Enabling zero padding causes performance overhead on the first write to every area of the volume.

**Note:** Zero padding policy must be defined for each storage pool before adding the first disk to the storage pool. After the addition of the first storage device to the storage pool, zero padding policy can no longer be changed.

## 3.3 Fault Set

A fault set is a logical entity containing a group of SDSs within a protection domain which have a higher chance of failing simultaneously, for example, SDSs powered by the same subrack. The purpose of fault sets is to make sure that data on any SDS in a fault set is mirrored in other fault sets.

The configuration of fault sets is optional, but if configured, a minimum of three fault sets must be configured. In this case, data is mirrored twice, each to a separate fault set.

When defining fault sets, the term “fault unit” is used to refer to the following:

- A fault set



- An SDS not associated with any fault set, which can be considered a single-SDS fault set

Fault sets can be configured at CEE with managed ScaleIO deployment or added later during region expansion.

For more information on fault sets, refer to the [Dell EMC ScaleIO Version 2.x User Guide](#).

## 4 ScaleIO Networking

This section describes the network configuration requirements of ScaleIO.

**Note:** The storage network architecture of the system must consider ScaleIO traffic patterns, especially in case of large ScaleIO deployments. The communication among ScaleIO software components (MDMs, SDSs, and SDCs) is predictable and is to be considered at storage network design.

This section describes the following network types used by ScaleIO:

- Frontend storage network
- Backend storage network
- Management network

Isolation of frontend and backend traffic is not required but strongly recommended in case of two-layer deployments, that is, in systems where storage servers and compute hosts act independently.

In CEE with unmanaged switches, the networks used by ScaleIO must be configured manually on the switches.

### 4.1 Frontend Storage Network

The ScaleIO frontend networks are used for the following purposes:

- The majority of frontend storage traffic consists of SDC to SDS communication, including all read and write traffic arriving at, or originating from a client. In case of multitenancy, SDC to SDS traffic can be isolated using VLANs and network firewalls.

SDC to SDS communication has a high throughput requirement.



- The MDM to SDC channel is used by the Master MDM for asynchronous communication with SDCs when the data layout changes. Data layout is changed in the following cases:
  - SDSs are added
  - SDSs are removed
  - SDSs enter maintenance mode
  - SDSs go offline

MDM to SDC traffic requires a reliable low latency network.

In CEE, the networks `sio_fe_san_pda` and `sio_fe_san_pdb` are used by default as ScaleIO frontend, with `mos_names`, `scaleio-frontend-left`, and `scaleio-frontend-right`, respectively. The ScaleIO frontend networks must be present on all nodes, including vCICs, compute hosts, and ScaleIO nodes. The ScaleIO frontend networks must be configured on the storage switches.

If CEE is deployed with unmanaged ScaleIO, the frontend networks must be configured on the ScaleIO nodes as described in the host networking template for the corresponding `mos_names`. For more information, refer to the section on ScaleIO network configuration of the document [Configuration File Guide](#).

It is possible to configure ScaleIO frontend networking using `iscsi_san_pda` and `iscsi_san_pdb`, or even a single network if redundancy is not required; however, these configurations are not recommended and are not described in CEE documentation.

## 4.2 Backend Storage Network

The ScaleIO backend networks are used for the following purposes:

- The majority of backend storage traffic consists of SDS to SDS communication, including writes mirrored between SDSs, rebalance traffic, and rebuild traffic.

SDS to SDS traffic has a high throughput requirement.

- MDM to MDM communication is used to coordinate operations within the cluster and establish a quorum. MDMs issue directives to ScaleIO for rebalance, rebuild and redirect traffic. Members of the MDM quorum communicate to maintain a shared understanding of data layout. MDM to MDM communication does not carry, or interfere with, I/O traffic.

MDM to MDM traffic requires a reliable low latency network, with a lower throughput than is required for SDS or SDC communication.

- MDM to SDS communication is used by the Master MDM to issue rebalance and rebuild directives.

MDM to SDS traffic requires a reliable low latency network.



In CEE with managed ScaleIO, the networks `sio_be_san_pda` and `sio_be_san_pdb` are used by default as ScaleIO backend with `mos_names`, `scaleio-backend-left`, and `scaleio-backend-right`, respectively. The ScaleIO backend networks must be configured on the dedicated ScaleIO nodes. The ScaleIO backend networks must be configured on the traffic switches.

It is possible to configure ScaleIO backend networking using `iscsi_san_pda` and `iscsi_san_pdb`. It is possible to configure the frontend networks as backend as well; in this case, the bandwidth is shared between the frontend and backend traffic, and both use the storage switching domain. However, these configurations are not recommended and are not described in CEE documentation.

If CEE is deployed with unmanaged ScaleIO, the ScaleIO backend network is not configured in the CEE region.

## 4.3 Management Networks

Management networks are used for installation, management, and reporting. This includes the following communication:

- Traffic to the ScaleIO GW, including REST GW and SNMP trap sender communication
- Traffic to and from the LIA
- Reporting and management traffic to the MDMs, for example, `syslog` for reporting and LDAP for administrator authentication

## 4.4 Network Configuration Requirements

ScaleIO has the following networking requirements:

- **The GW server must have connectivity to all the nodes that are installed. In CEE, separate networks are used for management and data, and the server hosting the GW must have connectivity with both networks.**
- The minimum network configuration is the following:
  - 2x1 GE redundant management network
  - 2x10 GE redundant storage network, where frontend and backend ScaleIO traffic shares the same physical network
- If there are four 10 GE ports available on each dedicated blade, physically separate ScaleIO frontend and backend networks are recommended.
- Network connectivity is established for all components.
- Bandwidth and latency among all nodes is acceptable, according to the requirements of the intended application.



- The required bandwidth is supported by switches among the network nodes.
- Maximum Transmission Unit (MTU) settings are consistent across all servers.
- The following ports must be open on the local firewall and unused by other applications on the hosts of the following nodes:

ScaleIO component	Port
MDM	6611, 9011
Single SDS	7072
Multiple SDSs	7073-7076
ScaleIO GW (including REST GW, installation manager, and SNMP trap sender)	80, 4443
LIA	9099

- The following SDBG ports must be open on the local firewall and unused by other application on the hosts of the following nodes:

ScaleIO component	Port
MDM	25620
Single SDS	25640
Multiple SDSs	25641-25644

- The UDP port 162 (SNMP traps) must be open on the local firewall of all servers.

## 5 Security

ScaleIO cluster access is based on user defined credentials. LDAP authentication is not supported by EMC. For more information, refer to the document [Dell EMC ScaleIO Version 2.x Security Configuration Guide](#).



## 6 Fault Management

When using managed ScaleIO, ScaleIO alarms, alerts, and events are processed by CEE. For more information, refer to the document [Distributed Storage Alarm](#).

## 7 ScaleIO GUI

In CEE, the use of ScaleIO GUI is supported and optional. For more information on the GUI, refer to the document [Dell EMC ScaleIO Version 2.x User Guide](#).

## 8 Logging

When using managed ScaleIO, ScaleIO logs are stored locally on the ScaleIO dedicated hosts and are also transferred to vCICs central log storage area through syslog. For more information, refer to the section on remote logging in the document [Audit and Security Logging](#).

## 9 Maintenance

When using managed ScaleIO, region expansion and server replacement of dedicated ScaleIO hosts are performed by executing the relevant CEE procedures. For more information, refer to the documents [Server Replacement](#) and [Region Expansion](#).





## 10 Upgrade Process

In CEE deployed with managed ScaleIO, the upgrade process is handled by CEE.

In CEE deployed with unmanaged ScaleIO, only LIA and SDC are part of the CEE upgrade process. Ensuring ScaleIO cluster compatibility is the responsibility of the user.

The upgrade of ScaleIO components must be performed in a strict order. Figure 4 shows a high level overview of the upgrade process.

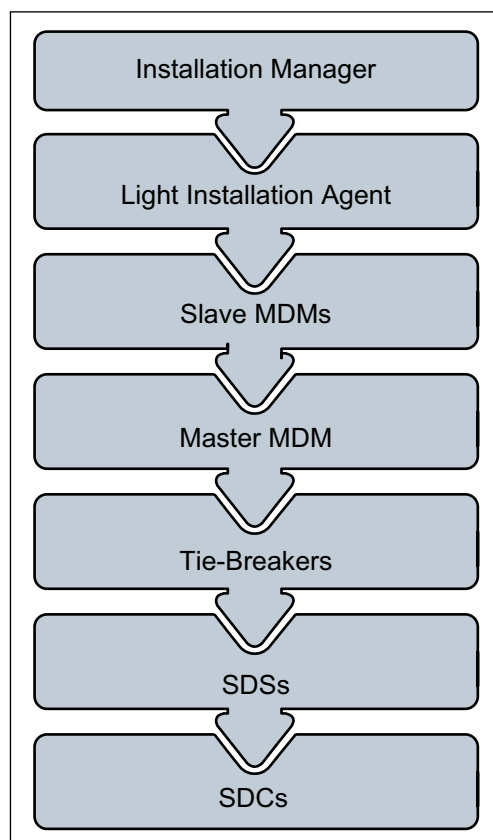


Figure 4 ScaleIO Upgrade Process

For more information, refer to the document [Dell EMC ScaleIO Version 2.x User Guide](#).