



CATE Storage and Backup Engineering  
CATE CitiSystems Engineered Systems  
Citigroup Architecture and Technology Engineering

---

# ScaleIO R1.3.1277

**CSBE-15007-PS**

**Project Scope**

**0.5**

©Citigroup Inc. 2015, All Rights Reserved.

This document contains material confidential to Citigroup. The contents are protected by specific employee and contractor agreements regarding confidentiality. Reproduction for distribution outside of Citigroup is prohibited.

## 1 Version Control

| Version | Date          | Author          | Reviewed                                                                          | Comments              |
|---------|---------------|-----------------|-----------------------------------------------------------------------------------|-----------------------|
| 0.1     | 17-April-2015 | Scott Torgerson | Chris Clayton,<br>Chris Howarth,<br>Bob Small,<br>Eliot Wilson.<br>Ludwig Derodel | Peer                  |
| 0.5     | 20-April-2015 | Scott Torgerson | Global Stakeholders                                                               | RFC                   |
| 1.0     |               | Scott Torgerson | Global Stakeholders                                                               | Release to Production |
|         |               |                 |                                                                                   |                       |
|         |               |                 |                                                                                   |                       |
|         |               |                 |                                                                                   |                       |
|         |               |                 |                                                                                   |                       |
|         |               |                 |                                                                                   |                       |
|         |               |                 |                                                                                   |                       |
|         |               |                 |                                                                                   |                       |
|         |               |                 |                                                                                   |                       |
|         |               |                 |                                                                                   |                       |

## 2 Legal Notice

Information contained herein is for Internal Use and may be used only for business purposes authorized by Citigroup.

CITI, CITIBANK, CITICORP, and CITIGROUP are trademarks and service marks of Citicorp and are used and registered throughout the world. Citibank, Citicorp, Citigroup and their subsidiaries also claim rights in certain other trademarks and service marks contained in these pages.

## 3 Contact

Please address any comments or questions regarding the information contained in this document via email to the [\\*CATE Global Engineering Support – ScaleIO](#) distribution list.

## 4 Latest Versions

It is recommended that the latest version of the document be used from the CSBE website at the following URL:  
<http://catecollaboration.citigroup.net/domains/platstor/gdse/products/default.aspx>.  
Newer versions may include certification on additional platforms, or comments on issues discovered after deployment.

## 5 Table Of Contents

|           |                                                 |           |
|-----------|-------------------------------------------------|-----------|
| <b>1</b>  | <b>VERSION CONTROL .....</b>                    | <b>2</b>  |
| <b>2</b>  | <b>LEGAL NOTICE.....</b>                        | <b>3</b>  |
| <b>3</b>  | <b>CONTACT.....</b>                             | <b>3</b>  |
| <b>4</b>  | <b>LATEST VERSIONS .....</b>                    | <b>3</b>  |
| <b>5</b>  | <b>TABLE OF CONTENTS .....</b>                  | <b>4</b>  |
| <b>6</b>  | <b>INTRODUCTION.....</b>                        | <b>5</b>  |
| <b>7</b>  | <b>PROJECT SCOPE.....</b>                       | <b>6</b>  |
| 7.1       | PLANNED COMPONENT AND FUNCTIONALITY MATRIX..... | 6         |
| <b>8</b>  | <b>CERTIFICATION SCOPE.....</b>                 | <b>7</b>  |
| 8.1       | FUNCTIONALITY IN SCOPE .....                    | 7         |
| 8.2       | FUNCTIONALITY NOT IN SCOPE .....                | 8         |
| <b>10</b> | <b>PROJECT CLASSIFICATION .....</b>             | <b>9</b>  |
| <b>11</b> | <b>PROJECT DELIVERABLES.....</b>                | <b>9</b>  |
| <b>12</b> | <b>APPENDIXES .....</b>                         | <b>10</b> |
| 12.1      | APPENDIX A: PLANNED TESTING .....               | 10        |
| 12.2      | APPENDIX B: ARCHITECTURE AND TOPOLOGY .....     | 12        |
| 12.2.1    | <i>ScaleIO Product Components.....</i>          | <i>12</i> |
| 12.2.2    | <i>Additional Architecture Features .....</i>   | <i>14</i> |
| 12.2.3    | <i>VMWare/ESXi implementation.....</i>          | <i>16</i> |
| 12.2.4    | <i>ScaleIO Diagram .....</i>                    | <i>16</i> |

## 6 Introduction

EMC ScaleIO is software that creates a SAN by aggregating servers' local direct-attached storage (DAS) to form globally shared block storage accessible over IP. The lightweight ScaleIO software components are installed on the application hosts and inter-communicate via a storage IP LAN to handle the application I/O requests sent to ScaleIO block volumes. This decentralized block I/O flow combined with a distributed, sliced volume layout results in parallelism which allows it to scale too many nodes (servers).

A ScaleIO volume is a block device that is exposed to one or more hosts (much like a traditional SAN LUN). ScaleIO breaks each volume into a large number of 1MB data chunks which are spread across the servers and disks within the ScaleIO cluster. Each volume also has a corresponding mirror on different nodes in the cluster. This layout goes a long way in eliminating hot spots across the cluster and allows for the scaling of the overall I/O performance of the system through the addition of nodes and or disks. Furthermore, this layout enables a single application that is accessing a single volume to use the full IOPs of all the cluster's disks.

Traditional storage systems typically combine system software with commodity hardware (which is comparable to application server hardware) to provide enterprise-grade resilience. ScaleIO provides similar enterprise-grade resilience by running the storage software directly on commodity servers. Designed for fault tolerance and high availability, ScaleIO handles all types of failures including failures of media, connectivity, nodes, software interruptions, and more.

The ScaleIO software features distributed auto-healing processes that can overcome media and node failures without requiring administrator involvement. Dynamic and elastic, ScaleIO enables administrators to add or remove nodes and capacity on the fly. The software immediately responds to changes by rebalancing the storage distribution to achieve a layout that optimally suits the new configuration. All performance and redistributions on ScaleIO can be monitored and observed on the ScaleIO dashboard which is a simple Java applet.

Because ScaleIO is hardware agnostic, the software works with various types of storage including magnetic disks, solid-state disks, and PCIe flash cards.

Based on the above statements, previous evaluation findings, and potential cost savings of Scale-Out commodity storage, it has been determined that EMC ScaleIO will be certified for production use in Citigroup.

## 7 Project Scope

The scope of this project is to certify the initial release of EMC ScaleIO for DEV, UAT, and POC use. As part of this certification, POC (Proof of Concept) environments have been created with Operational partners for what is considered appropriate use cases of ScaleIO. This document will outline all functionality and versions that will be in scope for this initial project release.

### 7.1 Planned Component and Functionality Matrix

| Component Matrix                                       |                                                                                        |
|--------------------------------------------------------|----------------------------------------------------------------------------------------|
| ScaleIO <sup>1</sup>                                   | 1.31.1277                                                                              |
| ScaleIO Dashboard                                      | 1.31.1277                                                                              |
| REST Gateway                                           | 1.31.1277                                                                              |
| ScaleIO Dashboard Supported O.S. versions <sup>2</sup> | Windows 2008 R2 (SOE Build)<br>Windows 2012 (SOE Build)<br>Windows 2012 R2 (SOE Build) |
| MDM                                                    | RHEL 6.X (SOE Build)                                                                   |
| Tie-Breaker                                            | RHEL 6.X (SOE Build)                                                                   |
| SDS                                                    | RHEL 6.X (SOE Build)                                                                   |
| SDC Supported O.S. versions <sup>3</sup>               | RHEL 6.X (SOE Build)<br>Windows 2008 R2/2012 (SOE Build)<br>ESXi 5.5 SP1 (SOE Build)   |

**Note:** This project only covers ScaleIO v1.3.1277 which is slated for DEV/UAT and POC uses only. ScaleIO v1.32 will be GA in May 2015 which will be the targeted release for ScaleIO production use, after the CATE certification completes.

<sup>1</sup> All servers that makeup the ScaleIO cluster (MDM, SDS, REST Gateway) will only be certified on Linux SOE built servers.

<sup>2</sup> Due to the Java requirements, the GUI will only be certified for SOE built windows servers.

<sup>3</sup> Currently there is no ScaleIO support for Solaris and AIX.

## 8 Certification Scope

This document is only intended to list versions and functionality that will be in scope for this project. All installation and configuration instructions for this project will be recorded in the ScaleIO Solutions Guide document [CSBE-15007-SG](#), and all test results will be recorded in the ScaleIO Release Notes document [CSBE-15007-RN](#).

### 8.1 Functionality in Scope

Below is a list of items that will be in scope for this *initial* EMC ScaleIO certification. For a list of Planned tests for this certification see [Appendix "A"](#) of this document.

- Pre-requisites for ScaleIO deployments
- Network design and connectivity
- Installation / Configuration
- Hardware Configurations
- Provisioning
- Management Interface (Dashboard/CLI)
- ESX Platform Integration
- Tornado++ Hardware Environments
- Event Notification (Tivoli)
- Log Reconciliation (SEMs/ArcSight)
- Automated Configuration Checking (HPNA)
- Chargeback and Capacity Reporting (SRM Integration)
- REST Gateway<sup>4</sup>
- Security Process<sup>5</sup>
- POC (Proof of Concept) Environments
- Tiered Storage Pools
- ScaleIO Device throttling
- Over-Commitment<sup>6</sup>
- ScaleIO License Management

**Note:** Orchestration layer products are limited with this initial certification of ScaleIO v1.3.1277. This certification is aimed at DEV, UAT, and POC environments. The next ScaleIO certification (v1.32) will be targeted for production environments have more operational and cloud integration points.

<sup>4</sup> This will include integration testing with REST API with products like EMC SRM and HPNA.

<sup>5</sup> It is planned to take ScaleIO through the complete SAA process (ISRP, CIVA, Standard Build) as part of this certification and have an open RE at the end of the project.

<sup>6</sup> All over-commitment will be tested at the VMWare level with SDRS.

## 8.2 *Functionality not in Scope*

Functionality that will not be considered in scope for this project is listed below: However these items could be in scope in future ScaleIO certification project.

- Installation Manager and LIA
- O.S. level clustering software products in physical environments<sup>7</sup>
- Thin Provisioning
- Storage Level Remote Replication/COB<sup>8</sup>
- ScaleIO Snap Shots
- VMware plug-in
- Remote Vendor notification/support<sup>9</sup>
- ScaleIO level data Mobility<sup>10</sup>
- Integration with orchestration layers
- Integration with AD/LDAP for authentication/authorization<sup>11</sup>
- Data Integrity scrubbing checks<sup>12</sup>

<sup>7</sup> Currently in ScaleIO v1.3.1277 SCSI-3 reservations are only supported for Windows and ESX. Other host level clustering solutions such as VCS are not supported on ScaleIO by EMC. However, these type certifications are underway and are planned for future Citigroup ScaleIO releases.

<sup>8</sup> Currently there is no ScaleIO level replication functionality

<sup>9</sup> Currently vendor event notification is done via SMTP. Integration with EMC's ESRS is planned for future certifications.

<sup>10</sup> Currently there is no ability to move volumes between pools and or ScaleIO tiers.

<sup>11</sup> ScaleIO v2.0 (GA Dec 2015) will include AD/LDAP authentication

<sup>12</sup> ScaleIO v1.32 (GA May 2015) will introduce scrubbing software to perform a data integrity check by scanning the ScaleIO data.

## 10 Project Classification

This project is classified as a “TDL Create” with Work-ID **CAT140051467** under the Planview Project Management System. Progress of this project can be tracked by looking at the status page at the following URL: [EMC ScaleIO](#)

All comments pertaining to this project should be returned to CSBE by May 4<sup>th</sup> 2014 to the following: [\\*CATE Storage and Backup Engineering](#)

## 11 Project Deliverables

As a result of this project the following items will be completed:

### Documents

- **GDSE-15007-PS** – EMC ScaleIO R1.31.1277 Project Scope (*This Document*)
- **CSBE-15007-SG** – EMC ScaleIO R1.31.1277 Solutions Guide
- **CSBE-15007-RN** – EMC ScaleIO R1.31.1277 Release Notes
- **GDSE-15004-BP** – ScaleIO on T++ Hardware Best Practices

### POC Environments

The following POC environments will be built, configured, and maintained, while the application owners observe the applications running on ScaleIO.

- Mongo as a Service (Maas)
- Oracle RAC as a Service (Raas)

### Software Packages

- Windows ScaleIO TPM Packages  
SDC – CITI\_WIN\_EMCScaleIOSDC\_131277\_00  
GUI – CITI\_WIN\_EMCScaleIOGUI\_131277\_00
- Linux Repository Containing the following packages  
MDM – EMC-ScaleIO-mdm.x86\_64  
SDS – EMC-ScaleIO-sds.x86\_64  
Tie-Breaker – EMC-ScaleIO-tb.x86\_64  
REST Gateway – EMC-ScaleIO-gateway.noarch  
SDC – EMC-ScaleIO-sds.x86\_64  
Carrier-mon\_1.0 – Citi package to setup ScaleIO network interfaces  
Tonado\_routes\_1.0 – Citi package to configure routing tables T++ nodes.  
Java 1.7 – jre.x86\_64
- Vmware Repository Containing the following packages  
SDC vib – sdc-1.31.216.1-esx5.zip

## 12 Appendixes

### 12.1 Appendix A: Planned Testing

The following high-level areas of testing are planned as part of this EMC ScaleIO certification. A detailed test plan will be created as part of the certification process and included in the [CSBE-15007-RN](#) Release Notes document.

| Test | Storage Type                 | Host Platform                                         | Test                                                                 |
|------|------------------------------|-------------------------------------------------------|----------------------------------------------------------------------|
| 1    | Local Storage on x86 Servers | RHEL 6.X SOE Build                                    | Cluster Installation/Configuration                                   |
| 2    | Local Storage on x86 Servers | RHEL 6.X SOE Build                                    | Cluster components rolling upgrade                                   |
| 3    | Local Storage on x86 Servers | RHEL 6.X SOE Build                                    | Cluster various failure testing (MDM, SDS, SDC)                      |
| 4    | Local Storage on x86 Servers | RHEL 6.X SOE Build                                    | Cluster Self-Healing (Automatic node storage additions and removals) |
| 5    | Local Storage on x86 Servers | RHEL 6.X SOE Build                                    | Cluster Fault set functional testing                                 |
| 6    | Local Storage on x86 Servers | RHEL 6.X SOE Build                                    | Cluster network testing                                              |
| 7    | Local Storage on x86 Servers | RHEL 6.X SOE Build                                    | SDS Pool/Device throttling                                           |
| 8    | Local Storage on x86 Servers | RHEL 6.X SOE Build                                    | SDS RAM Read Cache testing                                           |
| 9    | Local Storage on x86 Servers | RHEL 6.X SOE Build                                    | SDS disk replacement testing                                         |
| 10   | Various SDC Client servers   | RHEL 6.X SOE Build/<br>Windows 2008 R2/2012 SOE Build | SDC network and various other failure testing                        |
| 11   | Local Storage on x86 Servers | RHEL 6.X SOE Build                                    | CLI functionality                                                    |

|    |                                |                                |                                                                        |
|----|--------------------------------|--------------------------------|------------------------------------------------------------------------|
| 12 | Local Storage on x86 Servers   | Windows 2008 R2/2012 SOE Build | ScaleIO Dashboard                                                      |
| 13 | Local Storage on x86 Servers   | ESXi 5.5 SP1/SP2               | Virtualized environment Testing. Integration with VMWare               |
| 14 | Local Storage on x86 Servers   | ESXi 5.5 SP1/SP2               | Thick provisioning                                                     |
| 15 | Local Storage on x86 Servers   | RHEL 6.X SOE Build             | Standalone/Physical environment Testing.                               |
| 16 | VMWare Cluster                 | ESXi 5.5 SP1                   | Compatibility of ESXi SDC with other VMWare functionality used in Citi |
| 17 | Local Storage on x86 Servers   | RHEL 6.X SOE Build             | REST Gateway/API Integration                                           |
| 18 | Local Storage on x86 Servers   | RHEL 6.X SOE Build             | Event Notification (Tivoli)                                            |
| 19 | Local Storage on x86 Servers   | RHEL 6.X SOE Build             | Log Reconciliation (SEMs)                                              |
| 20 | Local Storage on x86 Servers   | RHEL 6.X SOE Build             | Capacity/Chargeback reporting (SRM)                                    |
| 21 | Local Storage on x86 Servers   | RHEL 6.X SOE Build             | Automated SBCC (HPNA)                                                  |
| 22 | Local Storage on x86 Servers   | RHEL 6.X SOE Build             | Various Security Testing <sup>13</sup>                                 |
| 23 | Tornado++ Hardware environment | RHEL 6.X SOE Build             | Various Testing on Tornado++ hardware environment.                     |

<sup>13</sup> It is planned that both the ISRP Questionnaire and CIVA Scan will be completed on v1.3.1227

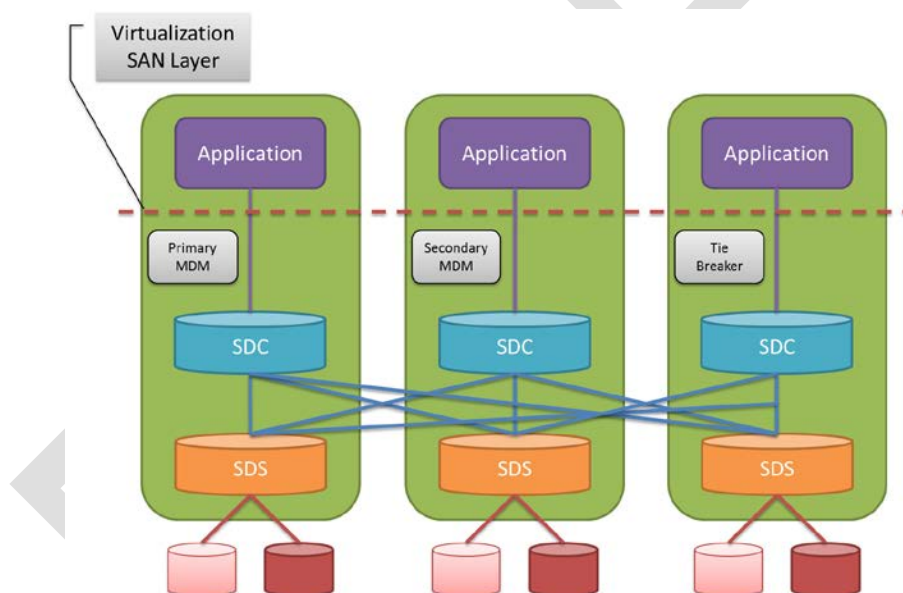
## 12.2 Appendix B: Architecture and Topology

ScaleIO is scale-out software-only solution running on commodity hardware that leverages servers with local disks communicating with a proprietary protocol over TCP to create a shared pool of **block storage**. It is hardware agnostic, and will work with HDD, SSD as well as flash PCIe cards. ScaleIO does not provide file or object storage.

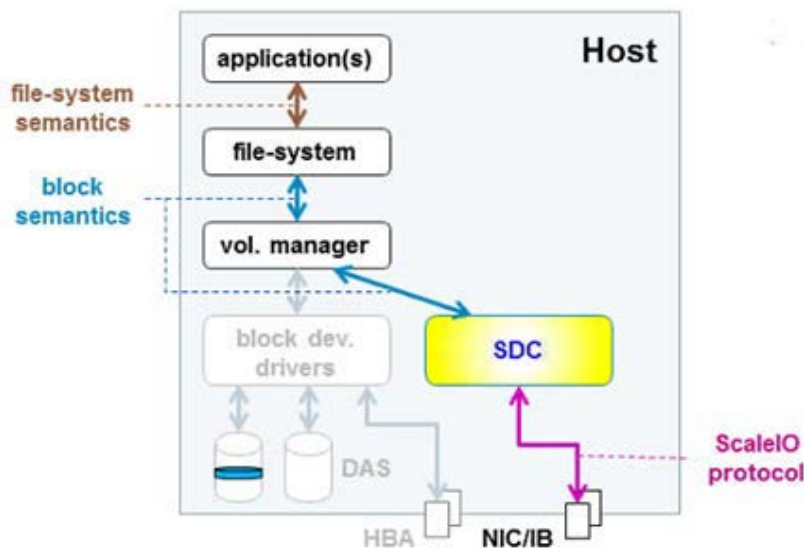
ScaleIO has been designed with auto-monitoring & auto-healing processes that deal with loss of disks and nodes without requiring administrator intervention. Nodes can also be added or removed on the fly, and the software will immediately respond to the changes rebalancing the storage across the new configuration.

### 12.2.1 ScaleIO Product Components

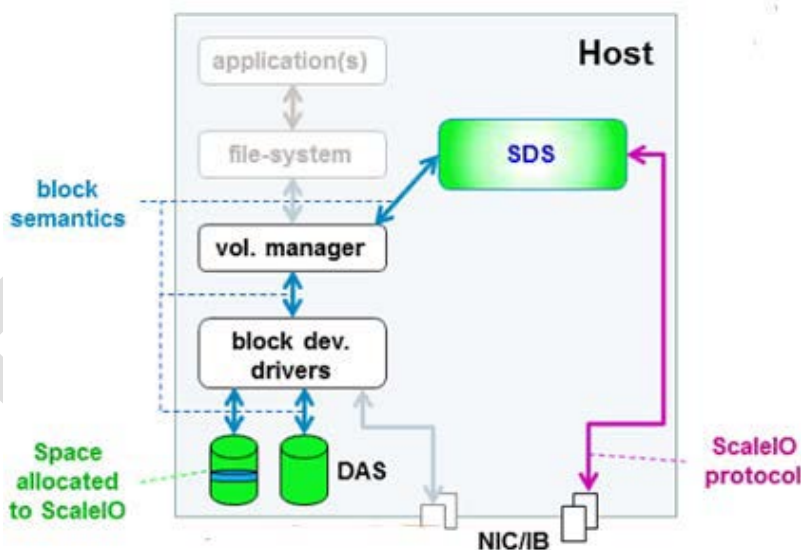
ScaleIO has three main components:



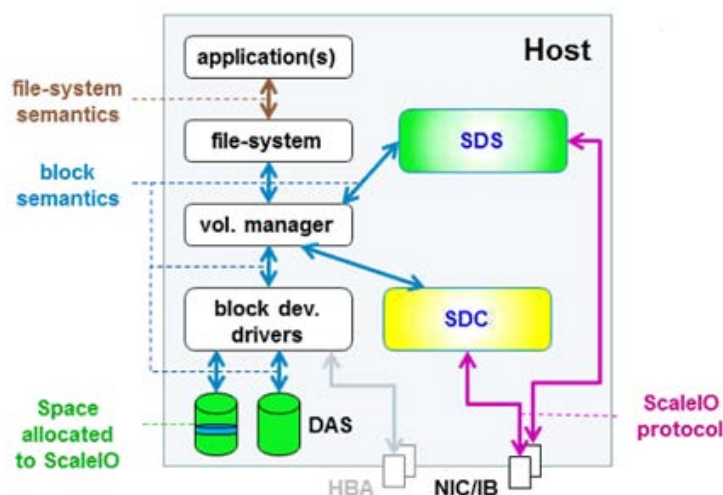
**SDC – ScaleIO Data Client** – Device driver resident on each host whose applications require access to ScaleIO virtual SAN block devices. An SDC is somewhat similar to an iSCSI initiator. When an SDC gets a read request from an application, it uses lookup information to determine which SDS has the data, and then interacts directly with that SDS over IP networks.



**SDS – ScaleIO Data Server** – Software component resident on each host that contributes local storage to the central ScaleIO virtual SAN. An SDS is somewhat similar to an iSCSI target.



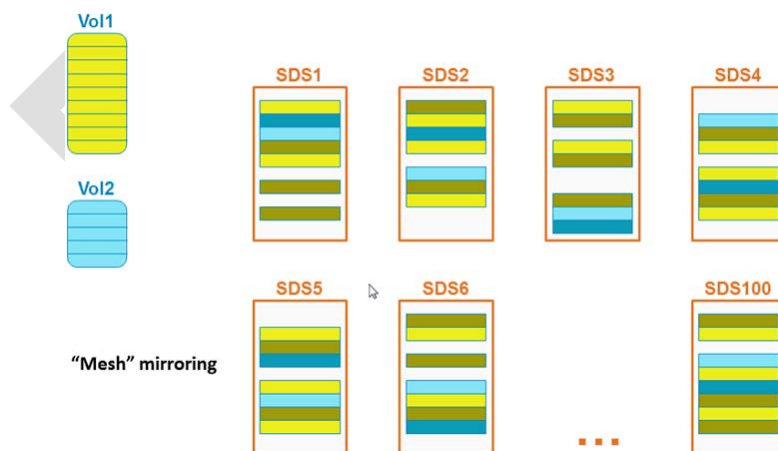
An SDC and an SDS can also run in the same host in a “converged” configuration. In this model, application servers would also be acting as providers of a shared pool of storage.



**MDM (Metadata manager)** – This software component is the authority for re-mapping and data chunk ownership, and is also responsible for data migration decisions (such as rebuilds and rebalancing). User data never passes via the MDM (passes directly from SDSs to SDC) and hence is not a bottleneck. A tie-breaker instance is used to avoid split-brain.

### 12.2.2 Additional Architecture Features

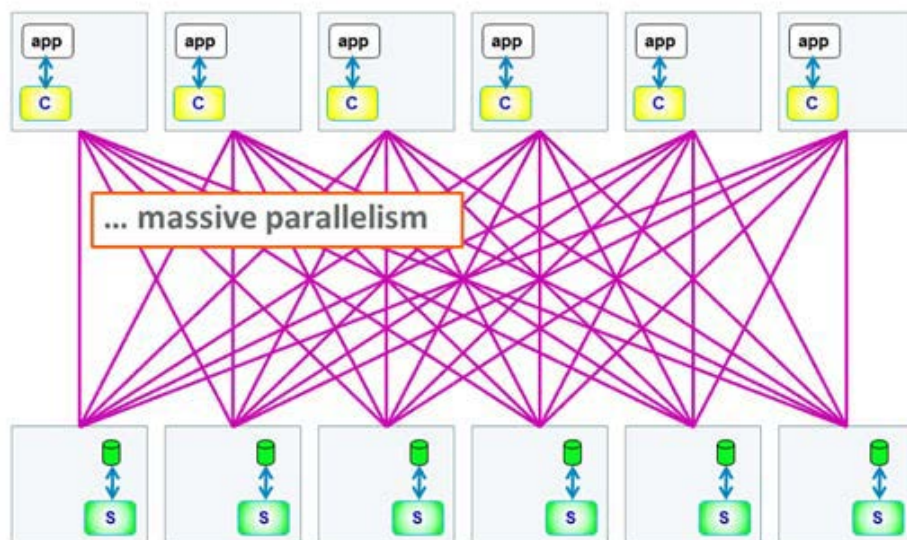
Each volume presented to hosts is broken into 1MB data chunks (of which there are 2 copies on separate SDS nodes disk i.e. a mesh mirror very similar to XIV) that are scattered across the cluster in a balanced manner, providing parallel access by an application to all nodes and their disks.



In the event of a loss, ScaleIO immediately starts to create another mirror for the data chunks that were lost in the failure.

Each SDC knows where to retrieve its data from and hence data flow is decentralized, which the vendor claims leads to linearly scalable performance and a

solution can scale to hundreds or even thousands of nodes. The more nodes the configuration has, the greater the numbers of parallel IO channels are available and the higher the aggregated IO bandwidth and IOPS.

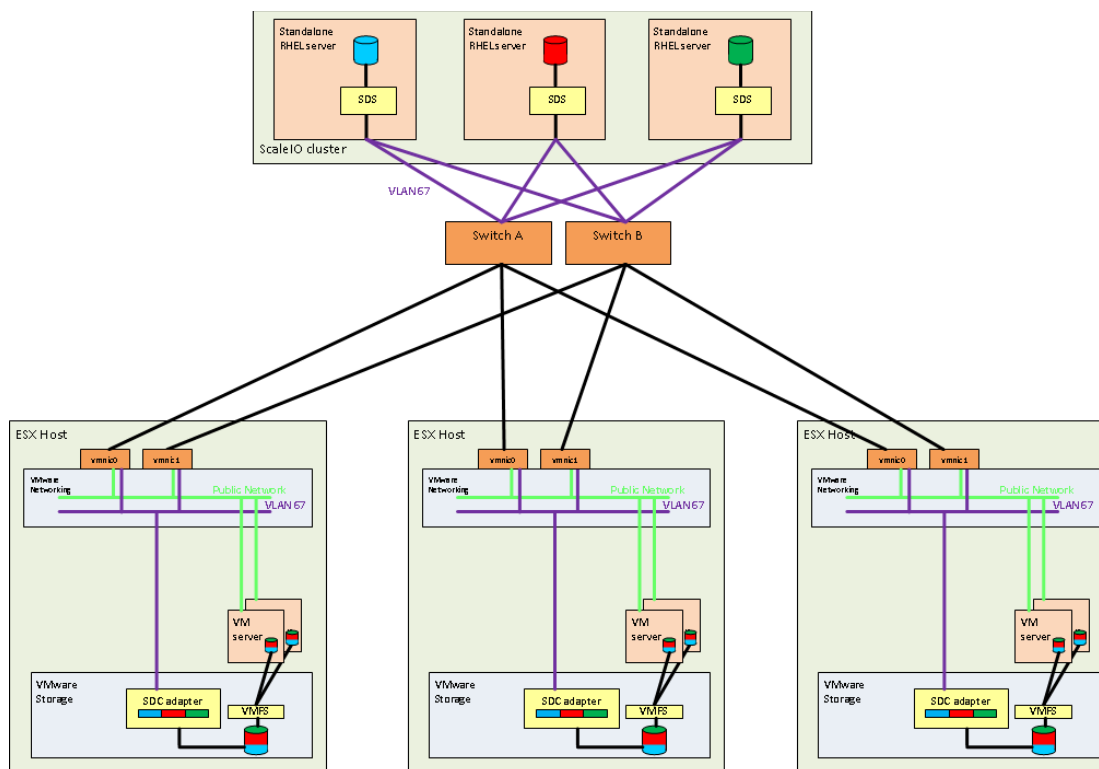


ScaleIO has the facility for administrators to set maximum IOPS or bandwidth values per client or datastore.

ScaleIO also supports redirect-on-write (ROW) snapshots of volumes, and these snapshots can be consistent across multiple volumes and servers. This is currently no support for remote replication within ScaleIO.

### 12.2.3 VMWare/ESXi implementation

ScaleIO is able to provision LUN devices directly to the ESXi Hypervisor. The diagram below provides an architectural overview of this.



The top diagram shows a 3 node ScaleIO cluster. This cluster aggregates the storage of the 3 servers to create LUN devices. Each ESXi host has the ScaleIO SDC client driver installed on it. The ScaleIO cluster maps LUN devices to all the ESXi hosts. As shown above each LUN is composed of chunks of data from all 3 physical ScaleIO hosts. The LUNs are then formatted as VMFS devices and a Datastore is created. At this point administration of the storage is performed by VMware.

Note that the same Datastore is visible on all ESXi hosts and none of this storage is physically local. Consequently storage VMotion operations are quick.

### 12.2.4 ScaleIO Diagram

The following diagram shows a typical ScaleIO deployment. All servers that make up the ScaleIO Cluster (MDM, Tie-Breaker, and all SDS's) would be built out on Citigroup SOE Built Linux servers. On the cluster is built and configured, all disks on the SDS nodes would be grouped together to form a storage pool. After the pool is created supported O.S. types (virtualized or physical) on the LAN could then load the ScaleIO SDC and have LUNs provisioned from the cluster.

