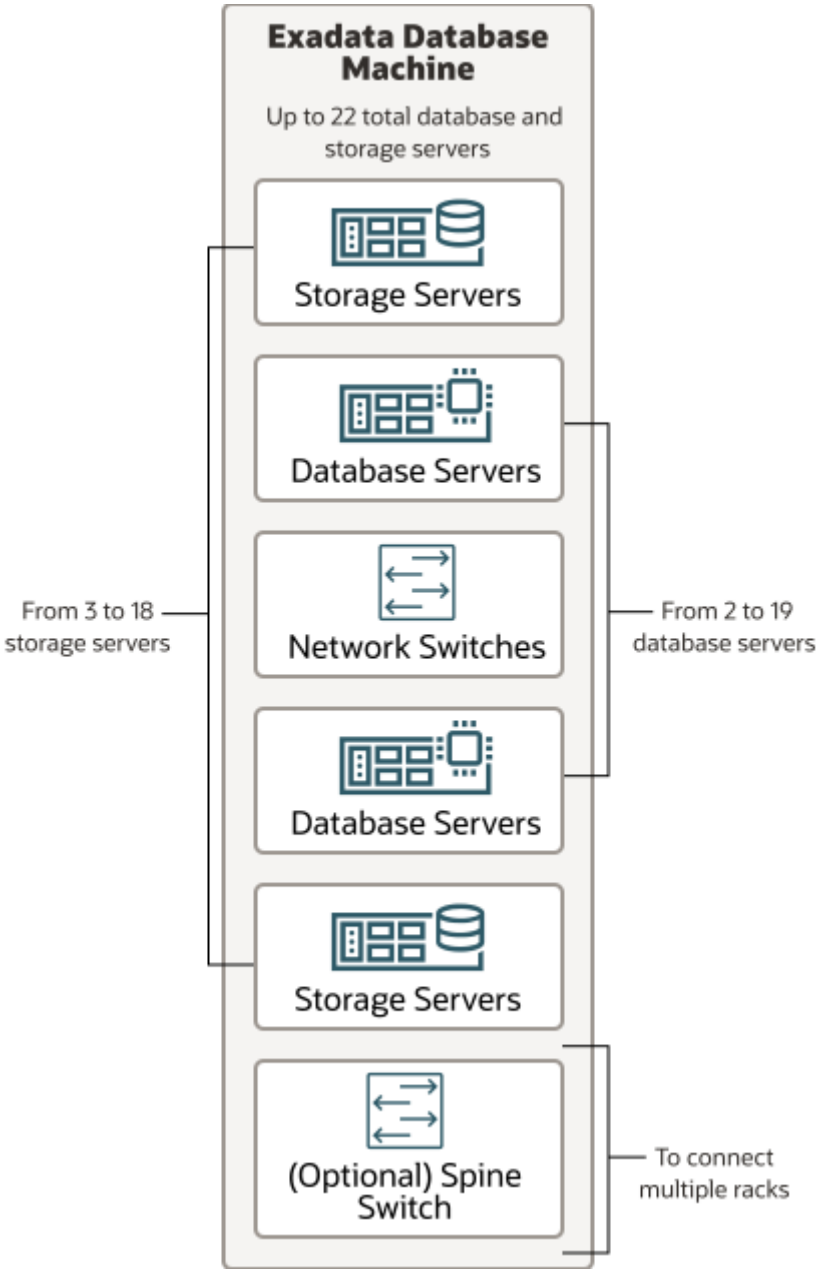
Oracle Exadata Database Machine Technical Architecture

ORACLE

Exadata Database Machine Rack Overview



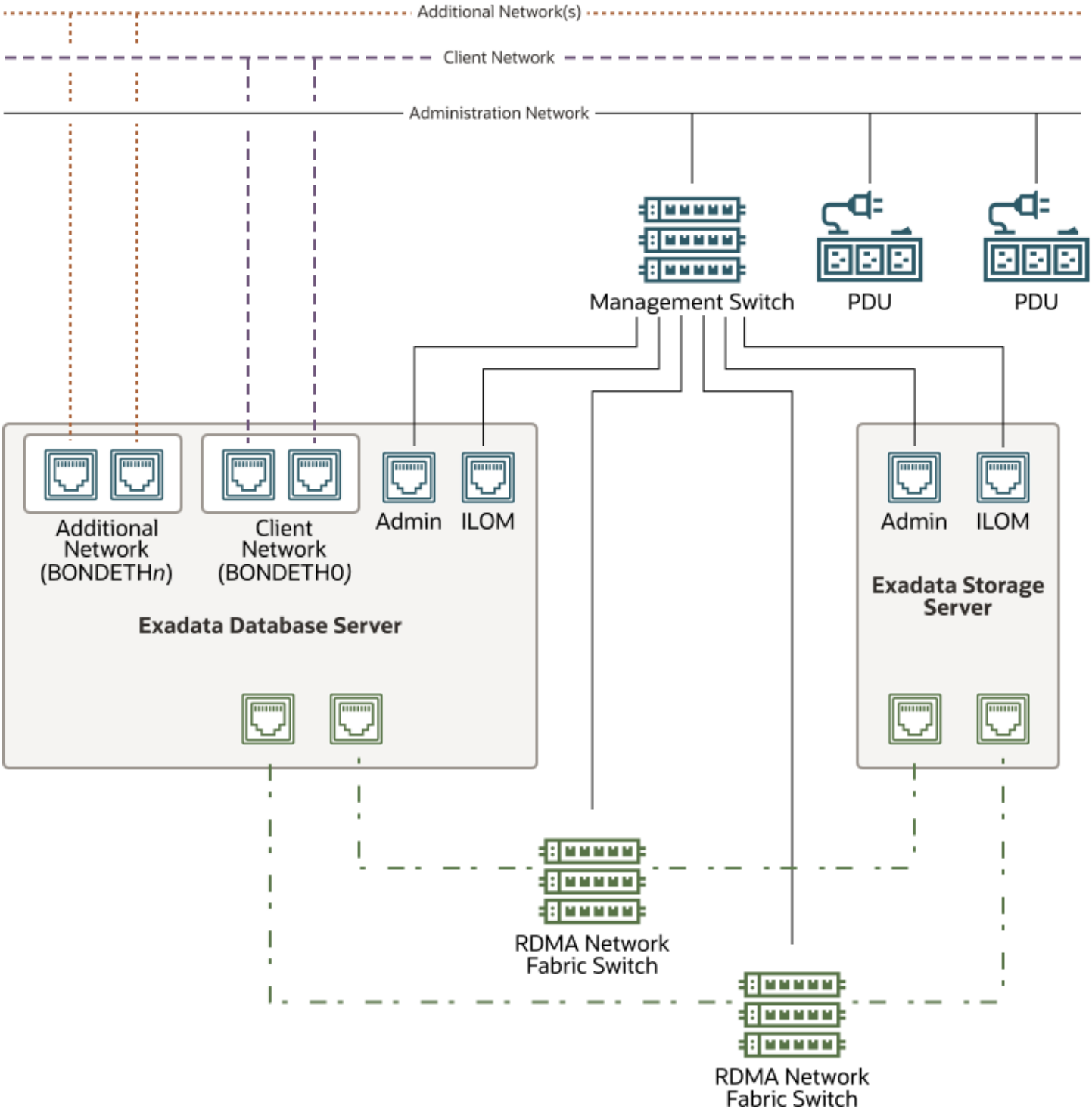
Oracle Exadata Database Machine features scale-out industry-standard database servers, scale-out intelligent storage servers, and high-speed internal RDMA Network Fabric that connects the database and storage servers.

You can select an eighth rack with 2 database servers and 3 storage servers or an elastic configuration with up to 22 total database and storage servers, including 2-19 database servers and 3-18 storage servers.

Exadata Database Machine also includes network switches to connect the database servers to the storage servers, and you can add an optional spine switch to connect multiple racks.

Note: All specifications are for Exadata X9M-2 racks. For details on all models, see Exadata Database Machine Hardware Components by Model.

Networking



Oracle Exadata Database Machine includes equipment to connect the system to your network. The network connections allow clients to connect to the database servers and also enable remote system administration.

The network has the following components:

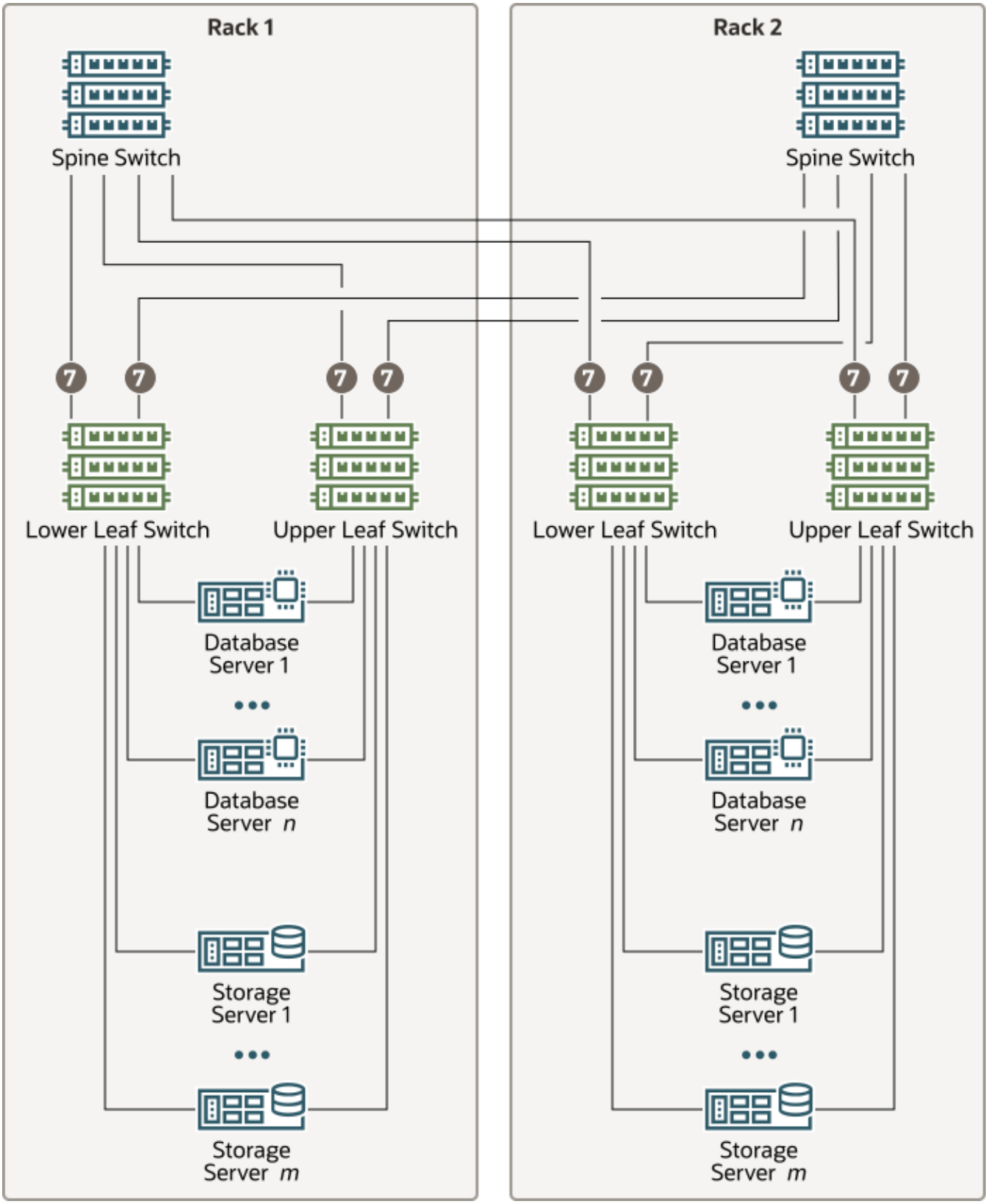
- One Management switch
- One database server, which represents all the database servers in the rack
- One Exadata Storage Server, which represents all the storage servers in the rack
- Two RDMA Network Fabric switches
- Two power distribution units (PDUs)

Exadata Database Machine provides the following networks and interfaces:

- The administration network connects to the PDUs and the Management switch. Through the Management switch, the administration network connects to dedicated administration and ILOM ports on every database server and storage server, and also to each of the RDMA Network Fabric switches.
- The client network is physically connected to every database server. The diagram shows a pair of bonded physical connections. In a bonded network configuration, the client network bonded interface name is BONDETH0.
- The private network, also known as the RDMA Network Fabric, interconnects all of the database servers and storage servers using a pair of RDMA Network Fabric switches. On each server, one port is connected to each switch, which maximizes throughput and availability.
- Database servers can optionally connect to additional networks using the available open ports that are not used by the administration network and the client network. The diagram shows a pair of bonded physical connections for an additional network. In a bonded network configuration, the first additional network bonded interface name is BONDETH1, the second additional network bonded interface name is BONDETH2, and so on.

For details on the networking requirements, see Understanding the Network Requirements for Exadata Database Machine.

Connecting Multiple Racks



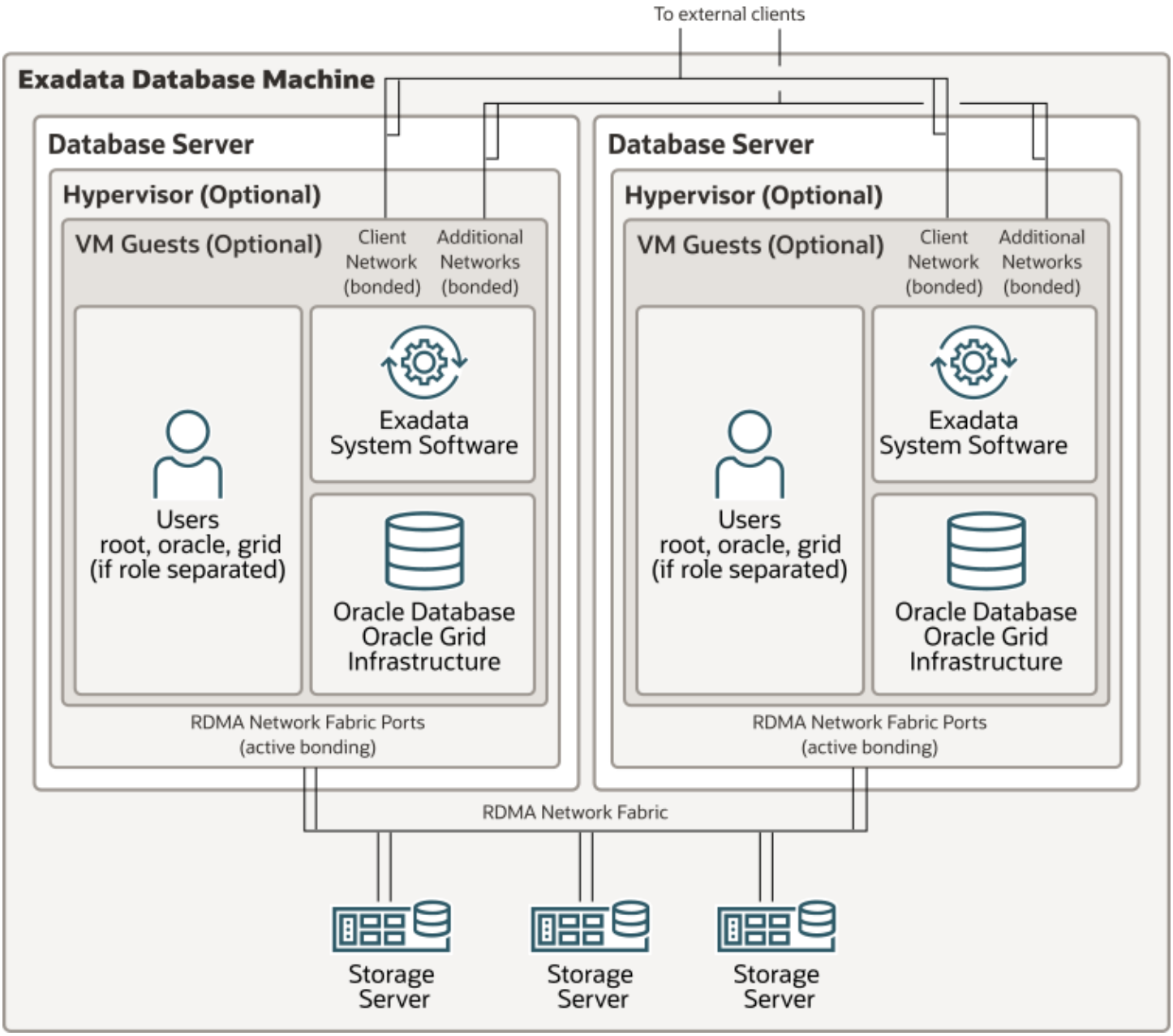
You can connect up to 12 RoCE-based Exadata racks together before external RDMA Network Fabric switches are required.

The diagram shows the RDMA Network Fabric architecture for two interconnected X9M racks. Each rack shows two database servers (1 and *n*) and two storage servers (1 and *m*), which represent all the database and storage servers in the rack.

Each rack has one spine switch and two leaf switches. Each spine switch has seven connections to each leaf switch. Leaf switch-to-leaf switch interconnection is not required. All database and storage servers connect to both leaf switches, the same as in a single rack.

For details on connecting more than two Exadata X9M racks, see Multi-Rack Cabling Tables for Oracle Exadata Rack X9M.

Database Servers



When deploying Oracle Exadata Database Machine, you can optionally implement virtual machines (VMs) on each database server. (Exadata Database Machine models with RDMA Network Fabric use Oracle Linux KVM. Earlier models that use Infiniband Network Fabric use Oracle VM.)

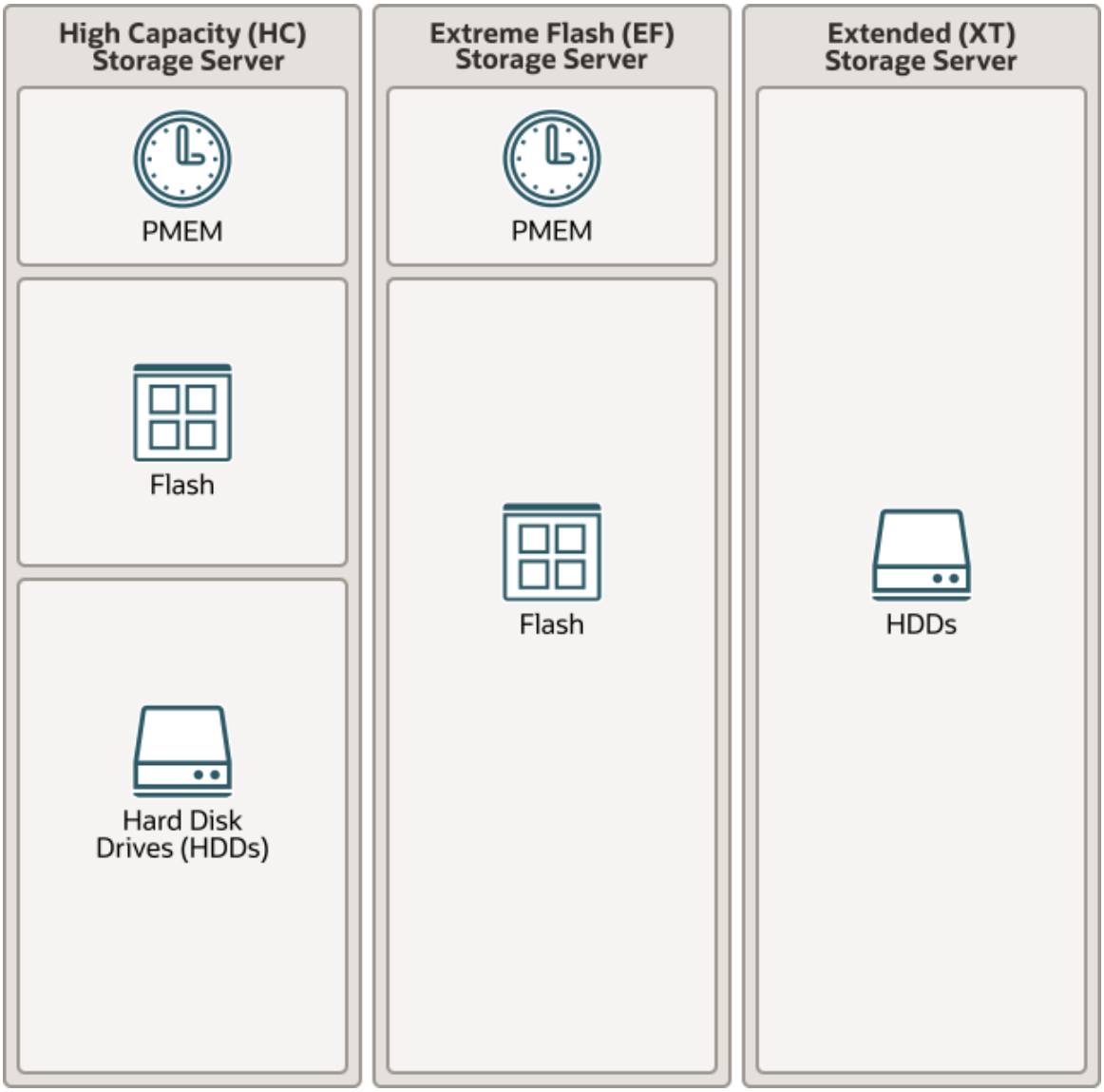
Each hypervisor can support multiple VM guests per database server. The number of VMs depends on the database server model and RDMA network technology. For details on VM guest specifications, see Managing Oracle Linux KVM Guests or Managing Oracle VM User Domains.

During configuration you install the Exadata system software, Oracle Database, and Oracle Grid Infrastructure on each VM guest. For details on the database server software, see About Database Server Software.

The default user accounts include oracle, root, and grid (if you select role separation). For the full list of default users, see Default User Accounts for Oracle Exadata.

External clients connect to the database servers through the client and additional networks with bonded network interfaces. RDMA Network Fabric interconnects all of the database servers and storage servers using a pair of RDMA Network Fabric switches. On each server, one port is connected to each switch. The RDMA Network Fabric ports use active bonding.

Storage Server Types



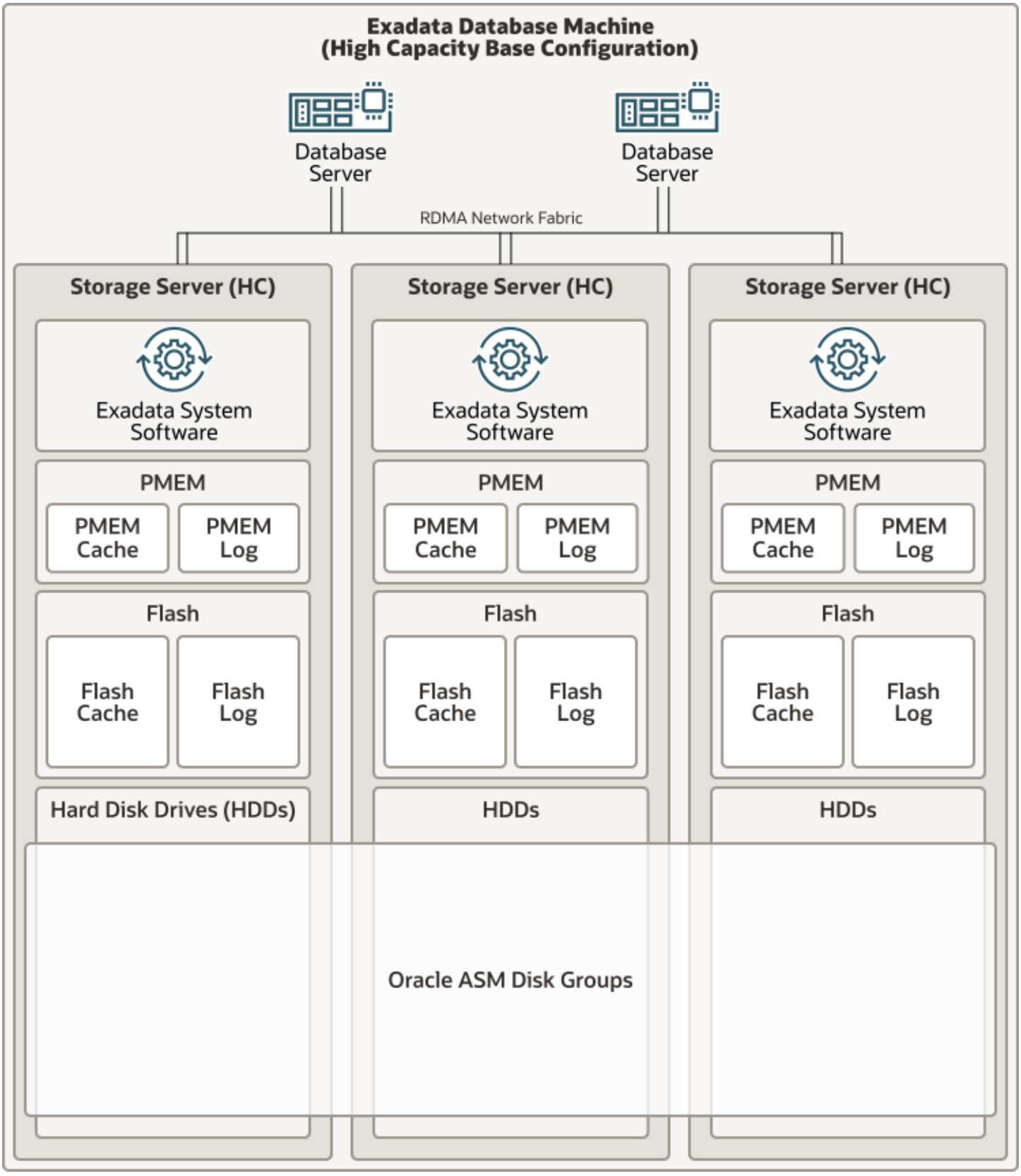
When configuring Oracle Exadata Database Machine, you can choose High Capacity (HC) or Extreme Flash (EF) Storage Servers. You can also add Extended (XT) Storage Servers to store rarely accessed data that must be kept online.

The storage server types have the following hardware components:

- HC Storage Servers include persistent memory (PMEM), flash, and hard disk drives (HDDs).
- EF Storage Servers have an all-flash configuration with PMEM.
- XT Storage Servers have HDDs only.

Note: All specifications are for Exadata X9M-2 Storage Servers. For details on hardware components for all models, see Oracle Exadata Storage Server Hardware Components.

High Capacity (HC) Storage Servers



Each Oracle Exadata Storage Server High Capacity X9M-2 (HC) server includes the following hardware components:

- Four 6.4 TB NVMe flash devices (PCIe)
- Twelve 18 TB hard disk drives (HDDs)
- Twelve 128 GB persistent memory (PMEM) modules

Note: All specifications are for Exadata X9M-2 Storage Servers (not including eighth rack configurations). For details on hardware components for all configurations and models, see Oracle Exadata Storage Server Hardware Components.

RDMA Network Fabric interconnects all of the database and storage servers using a pair of RDMA Network Fabric switches.

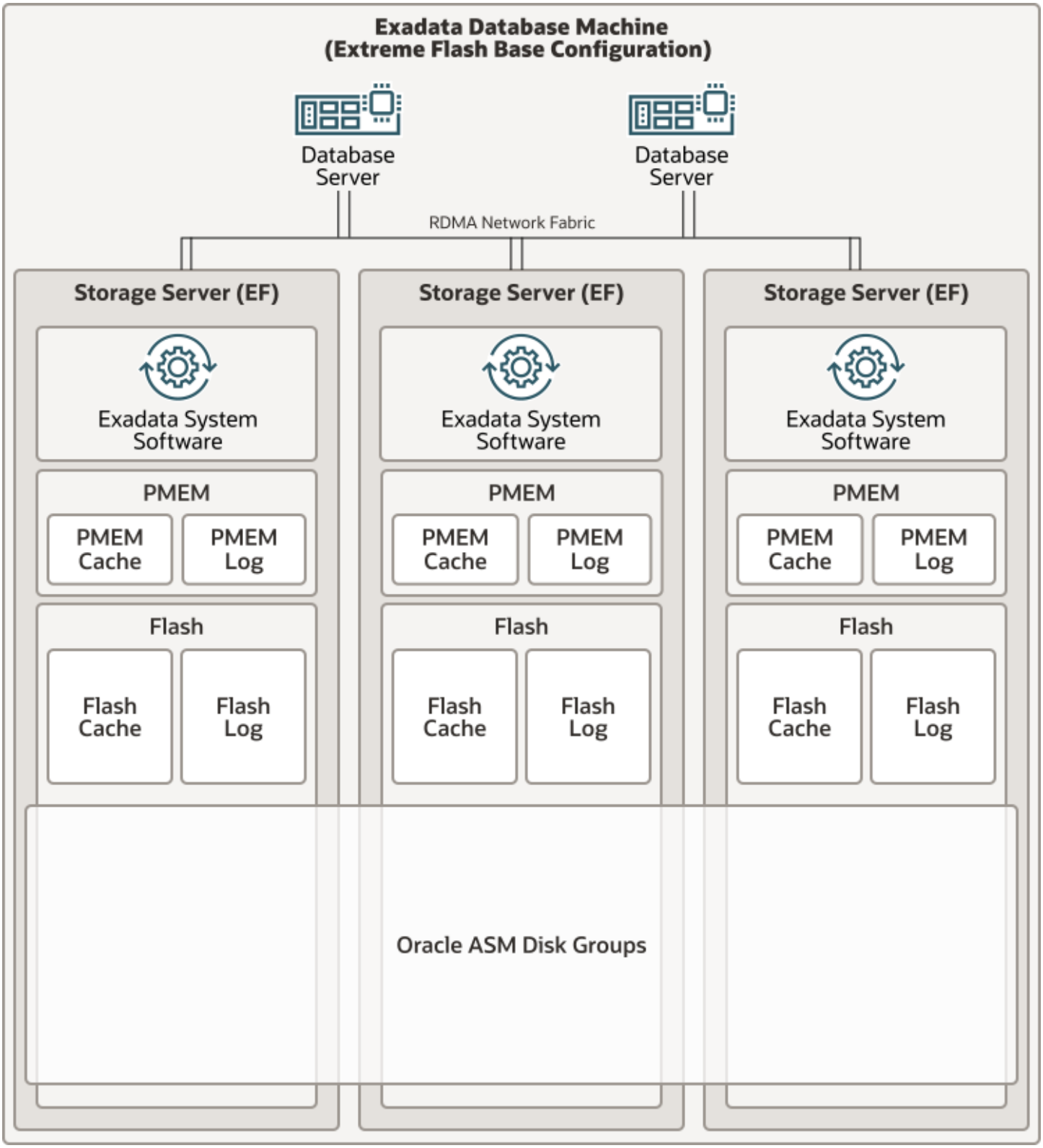
Each storage server runs Oracle Exadata System Software to process data at the storage level and pass only what is needed to the database servers. For details on the software components, see Introducing Oracle Exadata System Software.

On HC Storage Servers, you typically configure the flash as a flash cache (Exadata Smart Flash Cache), which automatically caches frequently used data in high-performance flash memory. The Exadata Smart Flash Log also uses a small portion of flash memory as temporary storage to reduce latency for redo log writes. For details on Smart Flash, see Smart Flash Technology.

Each storage server also includes a PMEM cache, also called the Persistent Memory Data Accelerator, in front of the flash cache to provide direct access to persistent memory through RDMA. Additionally, PMEM contains recently written log records (not the entire redo log) in the Persistent Memory Commit Accelerator. For details on PMEM, see Persistent Memory Accelerator and RDMA.

You configure Oracle Automatic Storage Management (ASM) disk groups to store and manage your data across the HDDs on multiple HC Storage Servers to improve performance and provide redundancy to protect against disk failures. For details on ASM, see About Oracle Automatic Storage Management.

Extreme Flash (EF) Storage Servers



Each Oracle Exadata Storage Server Extreme Flash X9M-2 (EF) server includes the following hardware components:

- Eight 6.4 TB NVMe flash devices (PCIe)
- Twelve 128 GB persistent memory (PMEM) modules

Note: All specifications are for Exadata X9M-2 Storage Servers (not including eighth rack configurations). For details on hardware components for all configurations and models, see Oracle Exadata Storage Server Hardware Components.

RDMA Network Fabric interconnects all of the database and storage servers using a pair of RDMA Network Fabric switches.

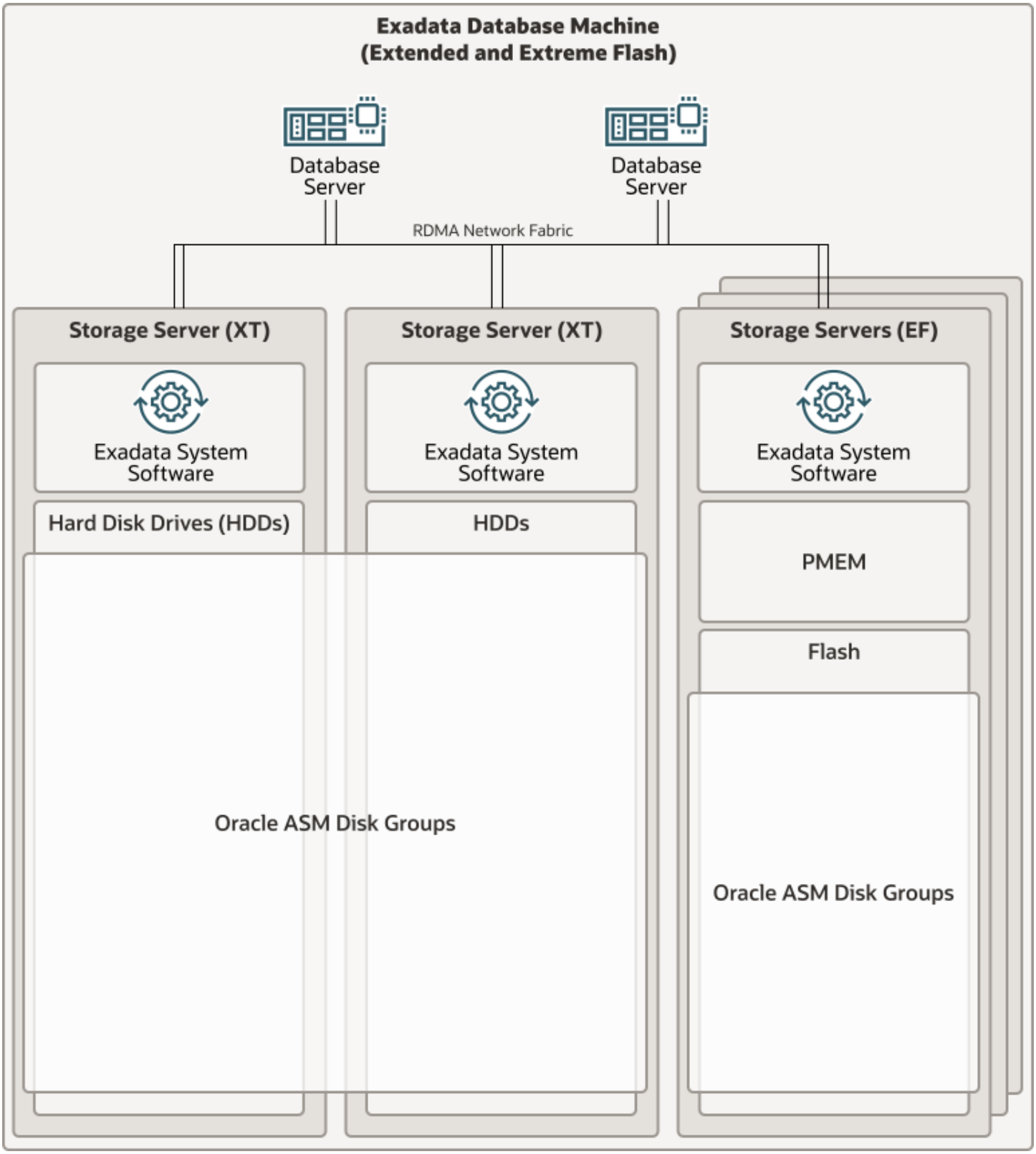
Each storage server runs Oracle Exadata System Software to process data at the storage level and pass only what is needed to the database servers. For details on the software components, see Introducing Oracle Exadata System Software.

On EF Storage Servers, all of the data resides in flash so you don't need the Exadata Smart Flash Cache for normal caching. However, you still use the Exadata Smart Flash Cache to host the columnar cache, which caches data in columnar format and optimizes various analytical queries. The Exadata Smart Flash Log also uses a small portion of flash memory as temporary storage to reduce latency for redo log writes. For details on Smart Flash, see Smart Flash Technology.

Each storage server also includes a PMEM cache, also called the Persistent Memory Data Accelerator, in front of the flash cache to provide direct access to persistent memory through RDMA. Additionally, PMEM contains recently written log records (not the entire redo log) in the Persistent Memory Commit Accelerator. For details on PMEM, see Persistent Memory Accelerator and RDMA.

You configure Oracle Automatic Storage Management (ASM) disk groups to store and manage your data across the flash devices on multiple EF Storage Servers to improve performance and provide redundancy to protect against disk failures. For details on ASM, see About Oracle Automatic Storage Management.

Extended (XT) Storage Servers



Each Oracle Exadata Storage Server Extended X9M-2 (XT) server includes twelve 18 TB hard disk drives (HDDs). Unlike Extreme Flash (EF) and High Capacity (HC) storage servers, XT servers don't contain flash or persistent memory (PMEM).

Note: All specifications are for Exadata X9M-2 Storage Servers (not including eighth rack configurations). For details on hardware components for all configurations and models, see Oracle Exadata Storage Server Hardware Components.

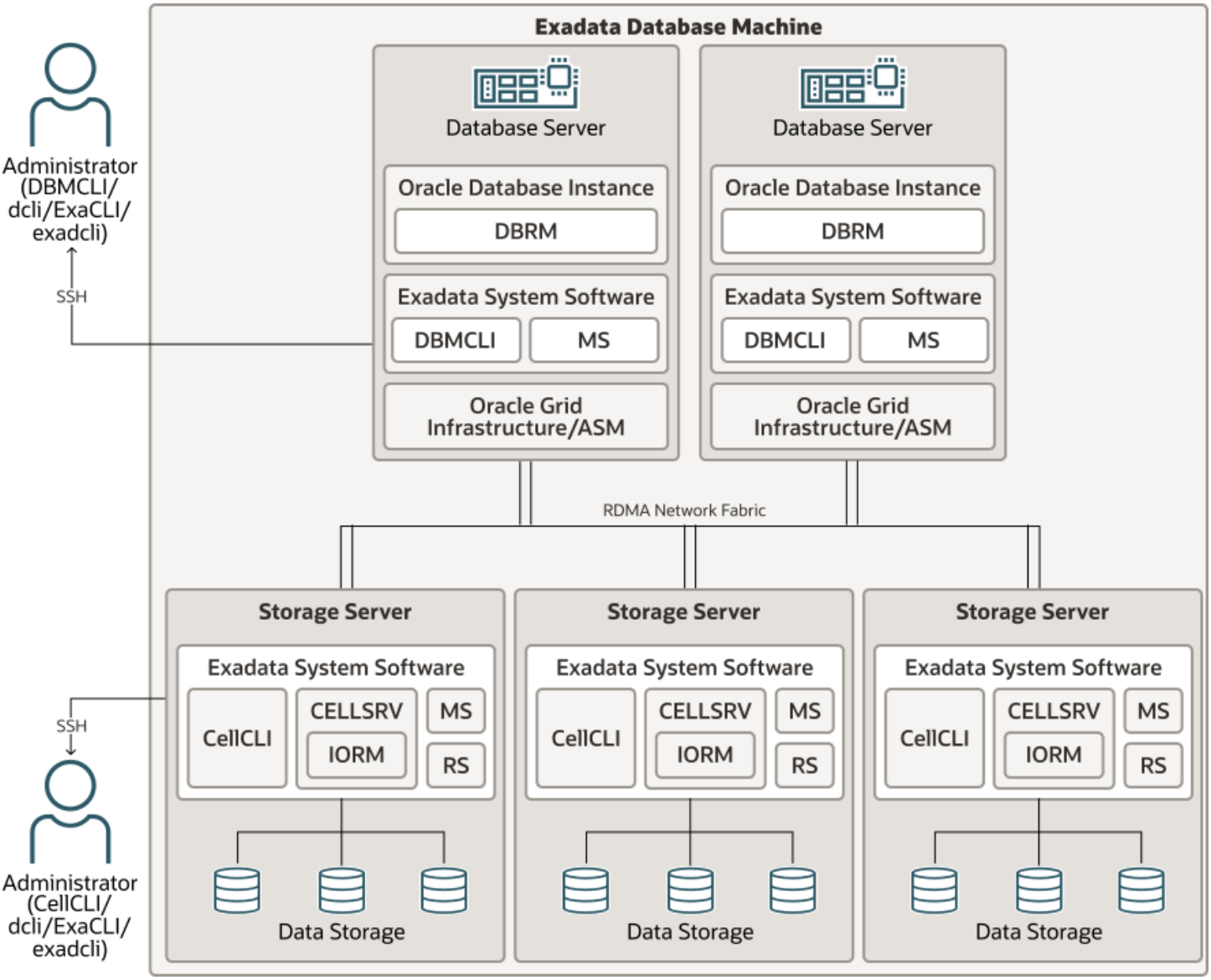
You can add two or more XT Storage Servers to existing EC or HC Storage Servers to store rarely accessed data that must be kept online. (The diagram shows two XT Storage Servers in addition to three EF Storage Servers.)

RDMA Network Fabric interconnects all of the database and storage servers using a pair of RDMA Network Fabric switches.

Each XT Storage Server runs the same Oracle Exadata System Software as HC and EF Storage Servers. XT Storage Servers do not require licenses for Exadata System Software and include Hybrid Columnar Compression. If you enable SQL Offload features on XT Storage Servers, Exadata System Software licenses are required. For details on the software components, see Introducing Oracle Exadata System Software.

You configure Oracle Automatic Storage Management (ASM) disk groups to store and manage your data across the HDDs on multiple XT Storage Servers to improve performance and provide redundancy to protect against disk failures. You set up different disk groups for XT, HC, and EF Storage Servers. For details on ASM, see About Oracle Automatic Storage Management.

Exadata System Software



Oracle Exadata System Software provides database-aware storage services, such as the ability to offload SQL and other database processing from the database server. The database and storage servers both contain components of the Exadata System Software.

Each database server includes the following software components:

- Oracle Database instance, including the Oracle Database Resource Manager (DBRM) for managing resource allocation
- Exadata System Software, including the DBMCLI command-line interface for managing the Exadata System Software on the database servers

- Management Server (MS), which works in cooperation with and processes most of the commands from the DBMCLI
- Oracle Grid Infrastructure, including Oracle Automatic Storage Management (ASM), which is the cluster volume manager and file system used to manage the data stored on the storage servers

Each storage server includes data storage hardware (disks or flash) and Exadata System Software to manage the data. The software includes the following components:

- Cell Control Command-Line Interface (CellCLI) for managing the Exadata System Software on the storage servers
- Cell Server (CELLSRV), which provides the majority of the storage server services, including the advanced SQL offload capabilities and the I/O Resource Management (IORM) functionality to meter out I/O bandwidth to the various databases and consumer groups issuing I/O calls
- MS, which works in cooperation with and processes most of the commands from the CellCLI
- Restart Server (RS), which monitors the heartbeat with the MS and the CELLSRV processes and restarts the servers if they fail to respond within the allowable heartbeat period

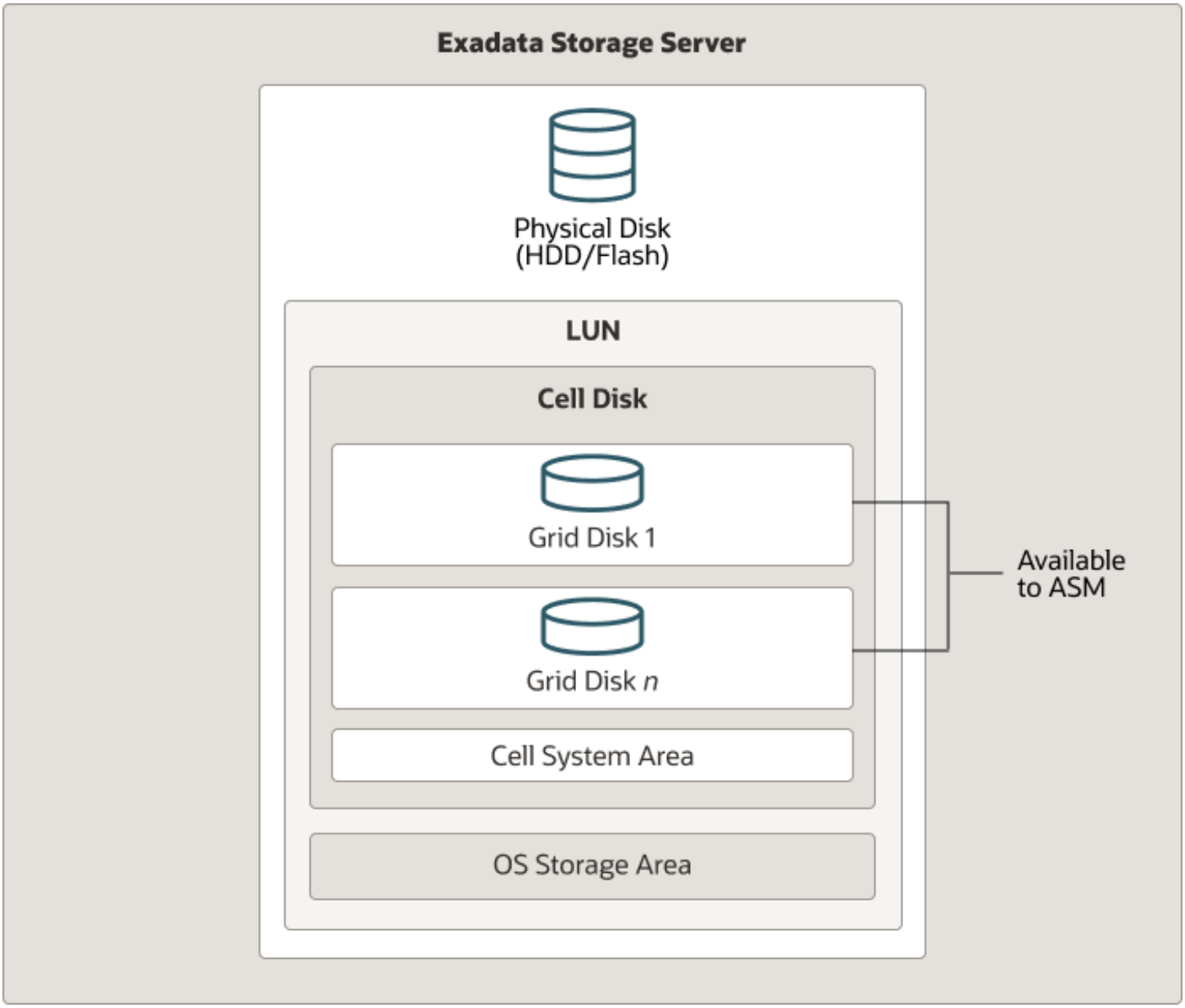
RDMA Network Fabric interconnects all of the database and storage servers using a pair of RDMA Network Fabric switches.

Administrators manage the database and storage servers through Secure Shell (SSH) or local access over the admin network (not shown in the diagram). Administrators can use the following command-line interfaces:

- DBMCLI for managing the database servers
- CellCLI for managing the storage servers
- dcli for automating operating system commands on a set of database or storage servers
- ExaCLI for managing database and storage servers remotely
- exadcli for centrally managing an Oracle Exadata system by automating ExaCLI commands

Note: This slide lists the most relevant Exadata System Software components. For the full list of components, see Oracle Exadata System Software Components.

Data Storage



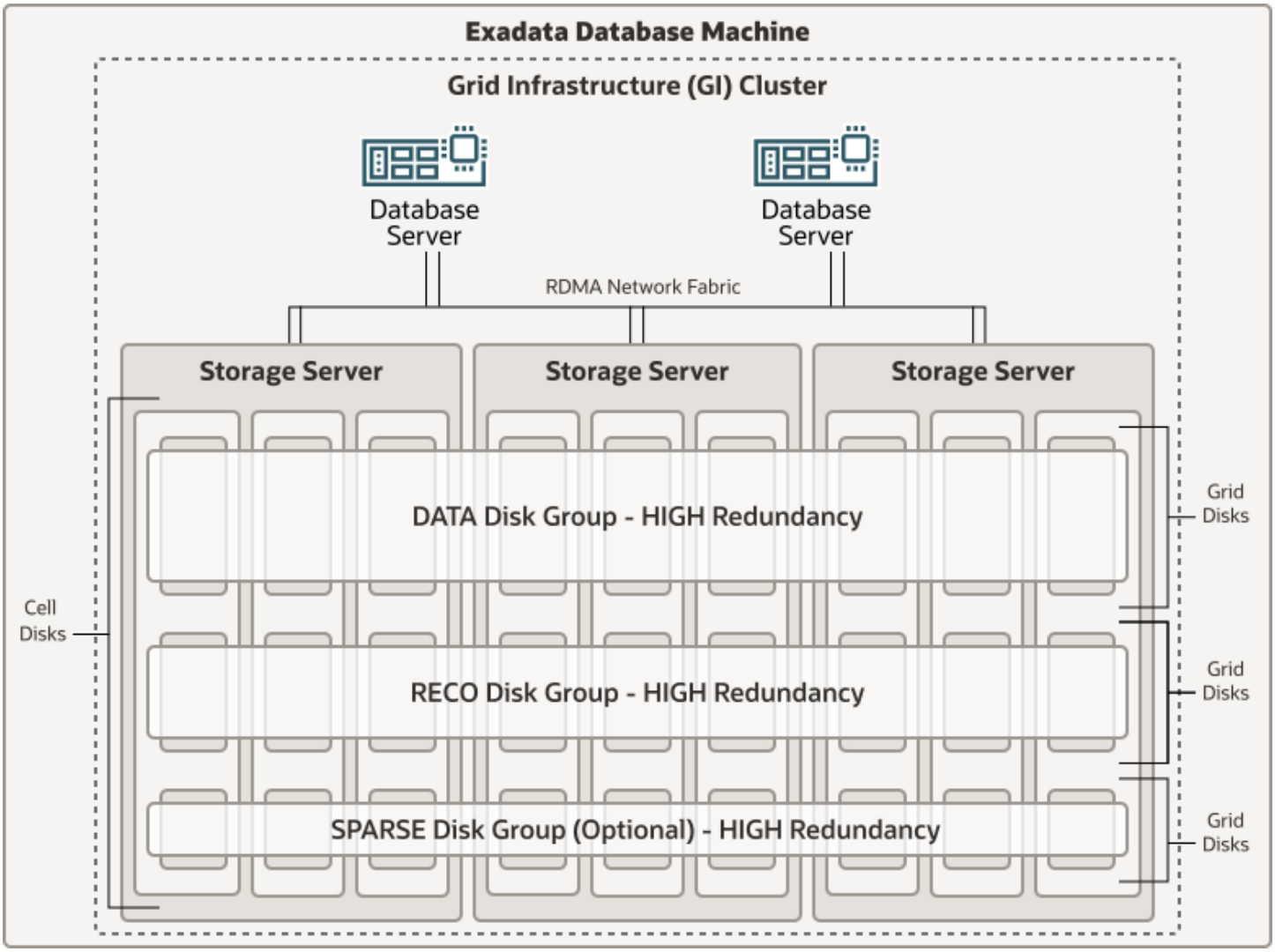
Oracle Exadata Storage Servers include physical disks, which can be hard disk drives (HDDs) or flash devices. Each physical disk has a logical address, called a logical unit number (LUN), which makes it available to the operation system (OS) and contains an OS storage area.

The cell disk is a higher level of abstraction that represents the data storage area on each LUN. You can divide a cell disk into multiple grid disks, which are directly available to Oracle Automatic Storage Management (Oracle ASM). For example, the diagram shows a cell disk divided into two grid disks. The cell disk also contains a segment called the cell system area, which is used by the Oracle Exadata System Software.

This level of virtualization enables multiple Oracle ASM clusters and multiple databases to share the same physical disk.

For details on the storage entities and relationships, see About Oracle Exadata System Software.

Oracle ASM Grid Disks and Disk Groups



Oracle Exadata Database Machine uses Oracle Automatic Storage Management (Oracle ASM) as the cluster volume manager and file system to manage data storage.

When configuring your Exadata rack, you define one or more Oracle Grid Infrastructure (GI) clusters, and you assign database and storage servers to the cluster. For example, the diagram shows one GI cluster with two database servers and three storage servers.

RDMA Network Fabric interconnects all of the database and storage servers using a pair of RDMA Network Fabric switches.

Each storage server includes physical disks, which can be hard disk drives (HDD) or flash devices. One physical disk corresponds to one cell disk. You divide the cell disks into multiple grid disks. (For simplicity, the diagram shows only three cell disks per storage server and three differently sized grid disks per cell disks.)

Your assign grid disks to ASM disk groups, which span cell disks across storage servers to improve performance and provide redundancy to protect against disk failures.

You typically configure the following Oracle ASM disk groups:

- DATA is the primary data disk group.
- RECO is the primary recovery disk group, which contains the Oracle Database Fast Recovery Area (FRA).
- SPARSE is an optional sparse disk group that supports Exadata snapshots.
- XTND is the default name for the disk group for Extended (XT) Storage Servers (not shown in the diagram).

You also configure the redundancy level for each disk group:

- HIGH redundancy (recommended) requires at least three storage servers and maintains three copies of every data block. (The diagram shows HIGH redundancy.)
- NORMAL redundancy requires at least two storage servers and maintains two copies of every data block.

For details on ASM disk groups, see About Oracle Automatic Storage Management.