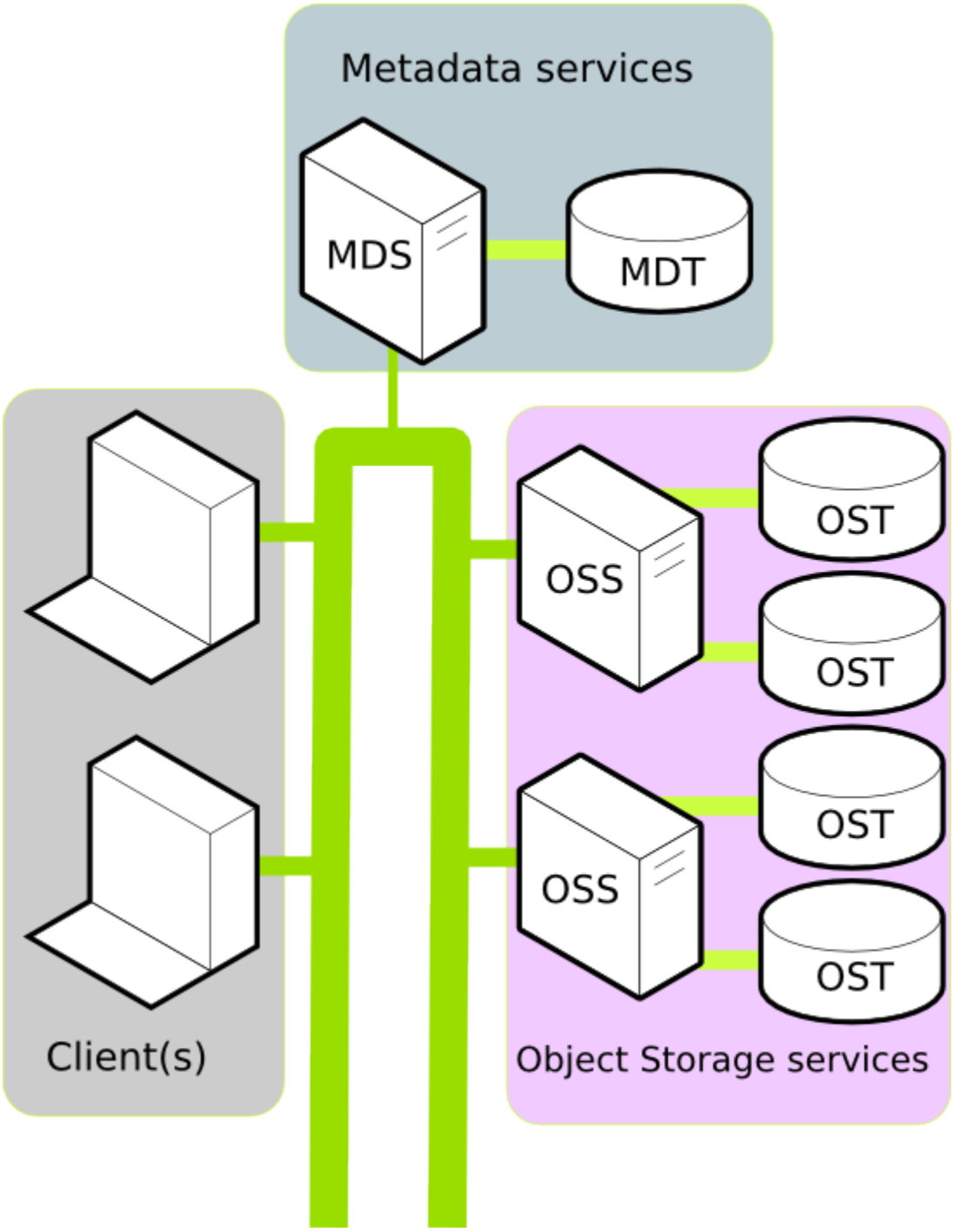
Components of a Lustre filesystem

A Lustre file system consists of a number of machines connected together and configured to service filesystem requests. The primary purpose of a file system is to allow a user to read, write, lock persistent data. Lustre file system provides this functionality and is designed to scale performance and size as controlled, routine fashion.

In a production environment, A Lustre filesystem typically consists of a number of physical machines. Each machine performs a well defined role and together these roles go to make up the filesystem. To understand why Lustre scales well, and to consider if Lustre is suitable in your use case, it is important to understand the machine roles within Lustre. A typical Luster installation consists of:

- One or more Clients.
- A Metadata service.
- One or more Object Storage services.



Client

A Client in the Lustre filesystem is a machine that requires data. This could be a computation, visualization, or desktop node. Once mounted, a Client experiences the Lustre filesystem as if the filesystem were a local or NFS mount.

Further reading:

- Lustre Manual

Metadata services

In the Lustre filesystem metadata requests are serviced by two components: the Metadata Server (MDS, a server node) an Metadata Target (MDT, the HDD/SSD that stores the metadata). Together, the MDS and MDT service requests like: Where is the data for file XYZ located? Do I have permission to write to file ABC?

All a Client needs to mount a Lustre filesystem is the location of the MDS. Currently, each Lustre filesystem has only one active MDS. The MDS persists the filesystem metadata in the MDT.

Further reading:

- Lustre Manual

Object Storage services

Data in the Lustre filesystem is stored and retrieved by two components: the Object Storage Server (OSS, a server node) and the Object Storage Target (OST, the HDD/SSD that stores the data). Together, the OSS and OST provide the data to the Client.

A Lustre filesystem can have one or more OSS nodes. An OSS typically has between two and eight OSTs attached. To increase the storage capacity of the Lustre filesystem, additional OSTs can be attached. To increase the bandwidth of the Lustre filesystem, additional OSS can be attached.

Further reading:

- Lustre Manual

Networking

Lustre uses LNET to configure and communicate over the network between clients and servers. LNET is designed to provide maximum performance over a variety of different network types, including InfiniBand, Ethernet, and networks inside HPC clusters such as Cray XP/XE systems.

Further reading:

- Lustre Manual

Tools

A Lustre filesystem uses modified versions of e2fsprogs and tar. Managing a large Lustre filesystem is a task that is simplified by community and vendor supported tools. Details of these tools are available at the **COMMUNITY PAGE OF THE WIKI**

Further reading

- Lustre design considerations white paper
- Notes from the Lustre Centre of Excellence at Oak Ridge National Laboratory
- Understanding Lustre Filesystem Internals white paper from National Centre for Computational Sciences.
- HDFS Scalability: the limits to growth