"Expand: High Performance Storage System for HPC and Big Data Environments" (TED2021-131798B-I00)

High Performance Storage Sytems for HPC and Big Data (Expand)

D 1.1
Report on New prototype of Expand with fault tolerance support for Big Data and HPC applications

Universidad Carlos III de Madrid

June, 2025

CONTENTS

1. EXPAND FOR HPC APPLICATIONS

The first version of Expand[1] was a parallel file system for heterogeneous clusters installed in user space, which facilitates its deployment in any POSIX-based environment [1], [2]. Expand works as a client-server architecture, building a distributed virtual partition across all the servers. When an application uses Expand as a file system, data is divided into blocks, whose size is defined at the virtual partition level, that are distributed in parallel among all the servers composing the distributed partition.

This first version of the Expand file system relied on standard servers and protocols, such as Network File System (NFS), File Transfer Protocol (FTP), or GridFTP, which facilitated its integration in heterogeneous systems, as well as the reuse and aggregation of existing resources and parallel access to the data. Furthermore, it did not require server changes (e.g., NFS) because all data management was done on the client. To achieve this goal, Expand provided different connectors, allowing the creation of virtual partitions with diverse storage servers that implement different technologies, such as NFS, FTP, etc. For example, an NFS server and an FTP server could form a virtual partition, being transparent to the application that it was using Expand.

Over the years, Expand has been evolving, and now this parallel and distributed file system allows the usage of its own servers that run in user space and exploit the local storage of the compute nodes where they run. Multiple communication protocols are supported in Expand nowadays: it has a specific server for distributed environments using sockets as a communication mechanism, and for HPC applications, Expand provides MPI connectors.

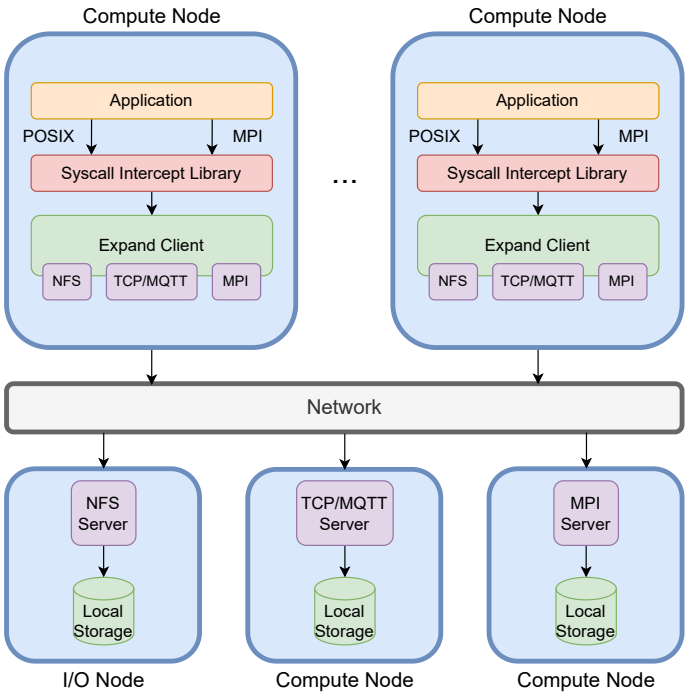


Figure 1.0.1: General Expand components diagram.

[1] `https://xpn-arcos.github.io/`

This chapter will explain the Expand file system's design and implementation for HPC applications. To this aim, we have completely redesigned the initial Expand file system, using MPI for communications between clients and servers.

Figure 1.0.1 shows the Expand components diagram with all elements. As may be seen, all of them could be integrated into a virtual partition without any problem by using on the client side the connectors included in Expand for each kind of server.

In the following sections, we explain the design and operation of Expand for HPC applications. We also show how they can be used as an ad-hoc file system in HPC environments to alleviate potential bottlenecks in the I/O systems created by data-intensive applications running on supercomputers.

1.1. The design of Expand for HPC applications

Expand file system for HPC applications architecture is based on an ad-hoc file system, where data servers that run on different compute nodes communicate with each other using the MPI standard (see Figure 1.1.1). We decided to use the MPI standard for communications in this version of Expand because, since its appearance in 1994, it has been heavily used in HPC. As a result, it is available in all supercomputers and is highly optimized for the different network technologies, following the MPI authors' goals: performance, scalability, and portability. In addition, MPI has already been used in distributed storage systems [3], and current MPI implementations have been upgraded to support modern high-performance communication technologies such as libfabric, and UCX, among others [4]–[6]. This allows MPI to remain a good choice for distributed storage systems.

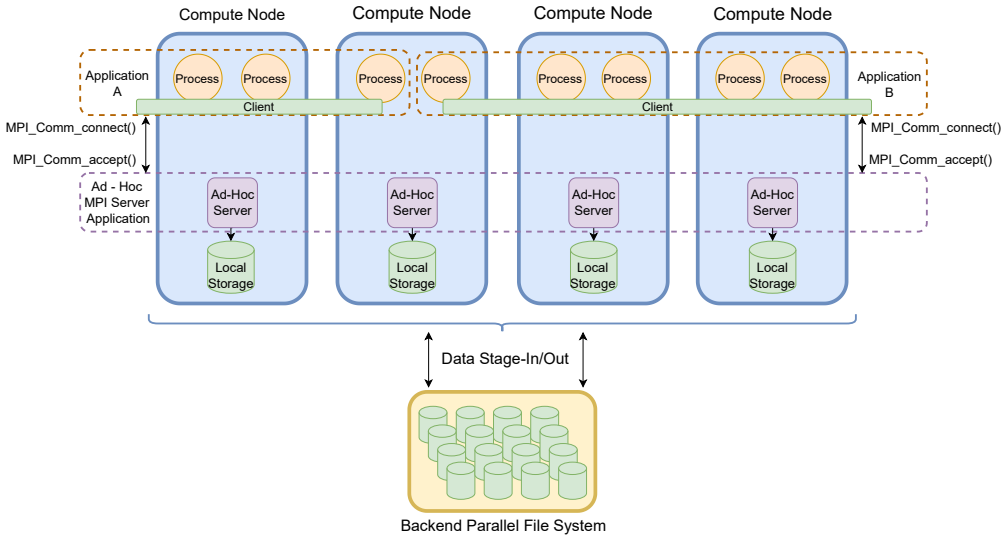


Figure 1.1.1: Architecture of the Expand File System.

As Expand is based on a client-server architecture, first, the Expand servers must be run as an MPI application on the set of compute nodes allocated to the application. If several application processes are run in one node, only one server is created in this node. Next, the application using the ad-hoc file system is deployed across all nodes with the Expand client. The way of deploying both elements using MPI communications is shown in Figure 1.1.2.

```
// MPI Ad-Hoc Server pseudocode

/* Phase 1: init + port + publish */
MPI_Init(argc, argv);

MPI_Open_port(MPI_INFO_NULL, port_name);
MPI_Publish_name(srv_name, port_name);

/* Phase 2: accept + work session + disconnect */
MPI_Comm_accept(port_name, MPI_INFO_NULL, 0,
      MPI_COMM_SELF, cli_comm);
...
MPI_Comm_disconnect(cli_comm);

/* Phase 3: finalize */
MPI_Unpublish_name(srv_name, MPI_INFO_NULL,port_name);
MPI_Close_port(port_name);
MPI_Finalize();
```

```
// MPI Ad-Hoc Client pseudocode

/* Phase 1: init + dns */
MPI_Init(argc, argv);


MPI_Lookup_name(srv_name, MPI_INFO_NULL, port_name);

/* Phase 2: connect + work session + disconnect*/
MPI_Comm_connect(port_name, MPI_INFO_NULL, 0,
     MPI_COMM_WORLD, srv_comm);
...
MPI_Comm_disconnect(srv_comm);

/* Phase 3: finalize */


MPI_Finalize();
```

Figure 1.1.2: Pseudocode of the MPI interconnection between ad-hoc servers and clients.

The Expand servers, in the first phase, initialize MPI using `MPI_Init` and each one takes care of creating a port and publishing it to the name service used (e.g., XPN DNS, hydra_nameserver, etc.). Once initialized, in a second phase, the servers wait for the connections from the clients, using `MPI_Comm_accept` on `MPI_COMM_SELF`. Once the connection is established, clients and servers use point-to-point calls to send and receive messages to perform the different operations on the file system until the client indicates to the server the end of the work session. In the last phase, each server unpublishes its port in the name service used, closes the port, and terminates MPI. This connection mechanism allows the servers to be able to serve different applications simultaneously, as they only need to know the connection end-points.

On the other hand, the ad-hoc client associated with the application that uses Expand in the first phase initializes MPI and looks up the port associated with a specific server in the name service. In the second phase, it connects to the server using `MPI_Comm_connect` on the `MPI_COMM_WORLD` on the previously obtained port. A working session is established between clients and servers until the application terminates. When this happens, the server is notified, and the disconnection is performed. Finally, in the third phase, the ad-hoc clients are finalized.

In Expand file system for HPC applications, local storage devices such as HDD, SSD, or Shared Memory on various servers deployed ad-hoc for an application are used to create distributed partitions. The goal is to store data near the application, thus reducing the need for remote access to the backend parallel file system. To store data locally, we rely on services provided by the local operating system, such as POSIX.

The application's processes and the Expand servers might be deployed on the same compute nodes because Expand does not need dedicated storage servers. Although applications and servers run on the same compute nodes, their interference is minimal. This is because applications mainly perform computations, while ad-hoc servers mainly perform I/O. Furthermore, in operating systems, while a process is blocked waiting for an I/O operation, another process runs on the CPU. Hence, the I/O operations of ad-hoc servers and the computation of applications overlap. In addition, today's multicore architectures help to alleviate this problem for two reasons: there is a large number of cores available for both the ad-hoc servers and the application, and it is possible to pin more cores to the application, if required, with minimal interference to the ad-hoc server.

This is the recommended deployment because it minimizes the number of accesses to remote data. Nevertheless, the processes of the application and the Expand servers can be deployed on different compute nodes if is needed. In Figure 1.1.1, you can see both alternatives.

Listing 1.1 shows an example template of a deployment using SLURM (Simple Linux Utility for Resource Management) [7]. First, the machinefile is created with the allocated nodes using `scontrol` command. Second, the ad-hoc servers are deployed, and, third, the application using Expand is deployed.

Listing 1.1: Example of SLURM job template.

```
#!/bin/bash

#SBATCH --job-name=ior_xpn_test
#SBATCH --nodes=32

#Get allocated nodes by SLURM
scontrol show hostnames ${SLURM_JOB_NODELIST} > machinefile

# Expand Ad-Hoc servers deployment
srun -n 32 -w machinefile ./xpn_mpi_server &

# IOR benchmark with Expand Ad-Hoc deployment
export LD_PRELOAD=./xpn_bypass.so

srun -n 256 -w machinefile ./ior -w -r -o /tmp/expand/xpn/ior.txt -t 1024k -b 1024k -s 4096 -i 10
```

Figure 1.1.3 describes the internal design of Expand as an ad-hoc parallel file system. This figure shows the different software layers (syscall intercept library, Expand client, etc.) involved from when an application performs an I/O operation until it is completed.

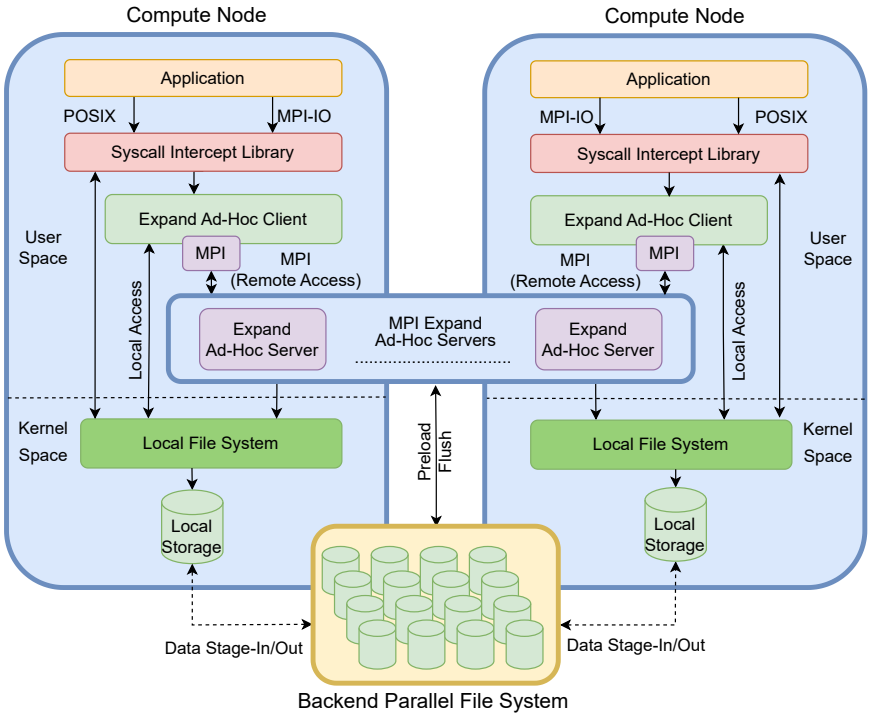


Figure 1.1.3: Expand architecture details.

As seen in this figure, an Expand server is deployed on each compute node, which is responsible for accessing the local storage devices. Furthermore, to mitigate bottlenecks, this server has an I/O request queue that allows buffering of the request peaks, as well as the local file system overload.

In addition, it can also be seen that Expand uses the local file system on the compute nodes to access the local storage devices. This lets Expand adapt to different local storage configurations (HDD, SSD, NVMe, or SHM) in a transparent way.

Finally, it should be noted that applications access Expand services using the system call interception library since this is responsible for calling the Expand services, as will be detailed in Section 1.7.

1.2. Definition of the Expand virtual partition

To define the Expand partitions, it is necessary to create a configuration file in XML format, as shown in the example in Listing 1.2. Each partition is defined in this configuration file by specifying its name, the Expand block size (partitioning size), and the URI of the nodes that comprise it, along with other parameters. The URI of the node includes the protocol (e.g., mpi_server, nfs3, etc.), the hostname (or IP address), and the mount point in the compute node's local storage.

The configuration file displayed in Listing 1.2 outlines the definition of a partition named `p1`. This partition is comprised of `two` ad-hoc servers that use the `mpi_server` protocol, with a block size set at `64 KiB`. Additionally, the file paths for each server's respective subfiles are specified as `path1` and `path2`.

Before running an application that uses Expand, the user must create the XML file defining Expand virtual partition. Expand provides a simple script (`mk_conf.sh`) that can generate the content of this file given all the partition parameters. The Expand client library relies on a configuration file to establish network connections between all Expand clients and servers.

Listing 1.2: Example of configuration file used to define an Expand virtual partition

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<xpn_conf>
 <partition name="p1" bsize="64k">
  <data_node id="<node ID>"
             url="mpi_server://<server>/<path1>"/>
  <data_node id="<node ID>"
             url="mpi_server://<server>/<path2>"/>
 </partition>
</xpn_conf>
```

1.3. Data distribution

The Expand file system uses the MPI standard to combine multiple servers to create a distributed partition. Each server provides one or more directories merged to create the partition on the compute nodes.

When the Expand is used, the files in a virtual partition are divided into blocks, which will then be distributed among all the Expand servers that are associated with that partition. It's important to note that each partition can have a different block size. Each server will hold a portion of the original file, resulting in multiple subfiles that make up the complete file stored in Expand file system. Those subfiles are transparent to the users.

To distribute the blocks equally among all the deployed ad-hoc servers, a cyclic or "round-robin" pattern is followed. This process is described in Figure 1.3.1, which shows the way the blocks are distributed among the subfiles in the parallel partition.

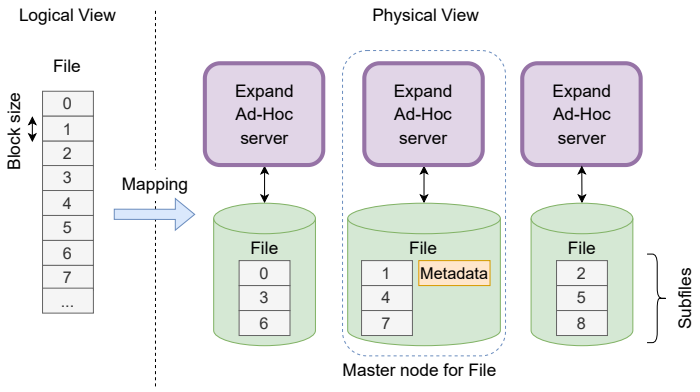


Figure 1.3.1: Data distribution in Expand.

1.4. Metadata and directory management

Expand does not use any kind of metadata manager to simplify the file naming process and reduce potential bottlenecks in the metadata server. The Expand clients are in charge of metadata management. For this purpose, Expand uses two levels of metadata management: metadata associated with the file, and metadata associated with Expand subfile management.

In the metadata associated with the file, we find the corresponding POSIX metadata, such as size, date, permissions, owner, etc. Those metadata are managed by the local file system of each node and are therefore distributed among all nodes. In this way, the metadata management is distributed among all nodes, avoiding creating bottlenecks in one node. On the one hand, to retrieve most of the attributes of a file, it is only necessary to consult the metadata from one subfile. However, some file attributes require the information of all subfiles, such as the file size. To calculate the size of a file stored in Expand, since it is divided into subfiles, you have to query the size of each of them and add these sizes together.

On the other hand, in Expand servers, the metadata required for managing subfiles is saved in a header at the start of each subfile (see Figure 1.3.1). The metadata saved in this header includes, among others; the number of servers; the base node, which the identifier of the ad-hoc server storing the first block of the file; the file distribution pattern used (at the moment, only "round-robin" distribution is available); and the block size, which is the same as the virtual partition block size where this file belongs. Even though all subfiles can accommodate the header, currently, only the master node, which may not be the base node, stores the metadata.

$$(\sum_{i=1}^{i=strlen(nameFile)} nameFile[i]) \bmod numServers \tag{1.4.1}$$

To determine the master node of a file among all servers of the partition, Expand uses a hash function based on the file name. The file name is converted to a node number by applying the following hash function (see

Equation 1.4.1) to identify the master node. Therefore, as the assignment of the master node depends on the file name, when a user modifies this name, the master node must also be changed. So the algorithm used in Expand to rename a file is the one shown in Listing 1.3.

Listing 1.3: Algorithm to rename a file.

```
xpn_rename(old-name, new-name) {
   old-master = hash(old-name);
   new-master = hash(new-name);
   move-subfile-metadata(old-master, new-master);
   storage-rename(old-name, new-name);
}
```

As depicted in Figure 1.4.1, directories are replicated on all ad-hoc servers within a partition. Thus, a directory-specific metadata manager is not necessary.

Expand uses the local file system of the compute nodes to store the data; this local file system is in charge of maintaining the consistency of the data at the node level. On the other hand, if multiple applications use Expand simultaneously, then these applications are responsible for maintaining the desired consistency.

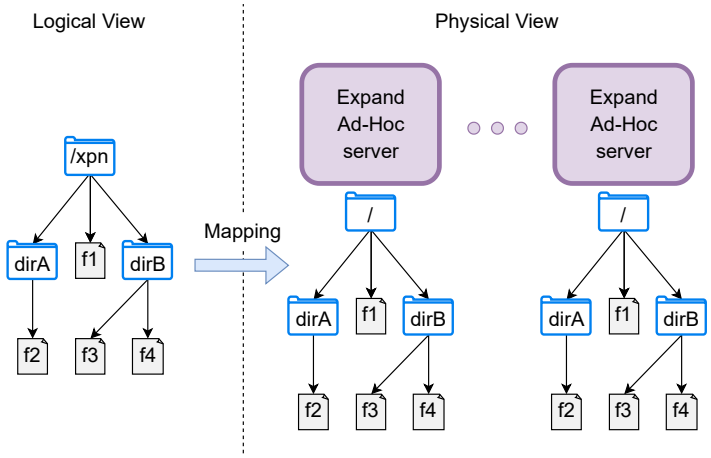


Figure 1.4.1: Metadata and directory mapping in Expand.

1.5. Parallel access

When a working session with a file is started (for example, when opening a file), a virtual file handler is created in Expand. This virtual file handler is used for all operations performed on this file during that session. Expand can access all the associated subfiles containing the associated data through this virtual file handler.

When a subfile is accessed, the Expand library uses the virtual file handler and breaks down the application operation into separate parallel sub-requests that are sent to the servers involved. This way, if k ad-hoc servers are involved in a request, Expand can perform k requests simultaneously to the servers using threads to speed up the process. This applies to both data and metadata operations carried out in Expand.

1.6. Data locality

As explained in Section 1.1, Expand servers can be executed on the compute nodes where the parallel application is being run. When this occurs, the application data is distributed among all compute nodes. In some cases, this allows access to the data requested by an application process using the local file system of the compute node by executing this process in the same node where the data is stored (see arrow tagged as "Local Access" in Figure 1.1.3). This helps to optimize data access from the Expand client and avoids delays and network-related issues that might occur when using remote data access.

It is important to keep in mind that data locality is crucial when multiple parallel applications (workflows) need to process the same dataset on the same nodes.

1.7. System call interception library

Expand offers a fully compliant POSIX and MPI-IO interface. With this aim, Expand Ad-Hoc provides a POSIX system call interception library that allows the use of Expand without modifying the source code of existing or legacy applications. This library intercepts the POSIX system calls made while a program is running. If a file is stored in Expand, which is detected because the file pathnames include the Expand mount point, the corresponding Expand API function is called. However, if the system call is made on the compute node's local file system, the corresponding `libc.so` system call will be executed (see Figure 1.1.3).

To load the intercept library before everything else, including the Linux `libc.so`, without needing super-user permissions, we have chosen to use `LD_PRELOAD`. Although there are other technologies to perform the same operation, such as FUSE [8], FUSE uses a block size of 128 KiB, which limits the block size of the ad-hoc File System to this value. Furthermore, FUSE is in the Linux kernel and requires super-user permissions, making it difficult to use, especially in HPC environments.

Also, Expand provides a native API with functions that are very similar to the POSIX API calls, both in their names (e.g. `xpn_open, xpn_read`, etc.) and in the arguments they receive (e.g. `xpn_read(fd, buf, nb)`). To use the Expand API, the source code must be altered, and the application must be compiled with the Expand dynamic library.

1.8. I/O Data staging

If an application needs initial data when using an ad-hoc file system, data staging is needed to preload (stage-in) data from the backend file system (e.g., GPFS) to the ad-hoc file system servers. The transfer of data between the backend file system and Expand file system is done through an MPI program that runs on the Expand servers. This program creates a replicated directory tree and the subfile in parallel on each server for every file stored in the backend.

Likewise, when the application that uses the ad-hoc file system finalizes its execution, persistent data must be stored in the backend file system because the ad-hoc servers will no longer be available. To complete this task, a flush operation (stage-out) is provided to move data from the Expand servers to the backend file system.

In Expand, an MPI program, similar to the one mentioned earlier, is used for this data transfer. The program runs on the ad-hoc servers and reads in parallel subfiles stored in the servers, generating the file that will be saved in the backend file system.

Furthermore, the flush operation can also be used during the application execution to store the generated data persistently in the back-end file system. Therefore, this operation helps mitigate data loss if a software or hardware issue occurs during the application execution. Moreover, as this operation is performed on demand, if the data generated by the application is temporary, then by skipping this operation, we avoid the potential overhead caused by its execution.

The Figure 1.1.3 shows both preload and flush operations.

2. EXPAND FOR BIG DATA APPLICATIONS

The solution proposed for Expand in Big Data applications involves designing a connector that enables using Apache Spark by Expand. The main goal is to exploit the benefits provided by both platforms. Spark applications run on the backend file system by default, but there are connectors for other frameworks, such as HDFS or Amazon S3. Their use is specified by the use of URIs: `file://` for the backend file system; `hdfs://` for HDFS; and `s3://` for Amazon S3. The use of Expand is expected to be similar to these platforms.

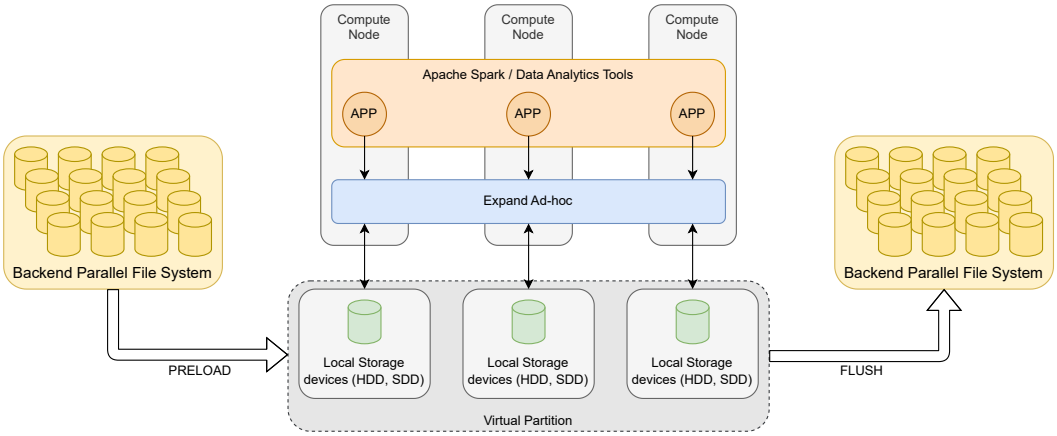


Figure 2.0.1: Workflow for Spark using Expand.

As Expand is designed to be deployed for specific applications, the workflow shown in Figure 2.0.1 must be followed. First, the data needed by the application shall be distributed among the compute nodes. This operation is called preload. When the preload operation is completed, Spark or any data analytics application can be deployed using the distributed data in the node's local file system. Finally, the data produced by the application that is wanted to be preserved shall be saved in the backend file system. This operation is called flush. Both the preload and flush operations are available in the Expand utilities.

To use Expand in Big Data environments, a two-layer connector has been designed as shown in Figure **??**. The top-level layer consists of tailoring Expand to the Hadoop environment by extending the `File System` class from the Hadoop project [9]. This will allow the creation of Expand clients through the Spark tasks. However, since the Expand client functions are written in C, it is necessary to create another layer to translate them to Java functions. To achieve this purpose, the Java Native Interface (JNI) is used to convert the client requests to the servers and the server's response to the clients so that they can communicate. The following sections explain in more detail the design of these layers.

The connector is offered in a `jar` package for ease of use. For this, the Maven [10] tool is used, which makes it easier to compile and package the connector. In addition, it allows the inclusion of the project dependencies.

2.1. Expand JNI layer

The implementation of the JNI layer is divided into four stages. First, it is necessary to create the Java interface with the sign of the functions required for implementing the Expand clients. Secondly, this interface must be compiled, and the header file of the necessary functions must be generated in C. Afterwards, it is required to implement the functions in C. In most cases, the goal is to call the Expand functions; however, some of them will have other requirements that are necessary for proper system performance. Finally, compiling these last sources and generating a dynamic library for the connector will be necessary.

The JNI layer must translate the Expand client functions, the POSIX standard `Stat` structure, and the POSIX standard flags used by the Expand client functions. For each of the last two, an object is created to be translated into a structure that the servers can handle. To use the functions by the upper layer, it is necessary to specify the signature of the functions that the Hadoop layer will use. For ease of use, the same name is used as the Expand client functions preceded by the `jni_` prefix, as shown in Table 2.1.1.

Table 2.1.1: JNI Expand client functions.

Function	Parameters
jni_xpn_chdir	String path
jni_xpn_chmod	String path, short mode
jni_xpn_close	int fd
jni_xpn_creat	String path, long mode
jni_xpn_destroy	
jni_xpn_fstat	int fd
jni_xpn_get_block_locality	String path, long offset, String[] url_v
jni_xpn_getcwd	String path, long size
jni_xpn_init	
jni_xpn_lseek	int fd, long offset, long whence
jni_xpn_mkdir	String path, short mode
jni_xpn_open	String path, long flags
jni_xpn_read	int fd, ByteBuffer buf, long size
jni_xpn_rename	String src, String dst
jni_xpn_rmdir	String path
jni_xpn_stat	String path
jni_xpn_unlink	String path
jni_xpn_write	int fd, ByteBuffer buf, long count

After creating the Java objects and the needed interface, compiling the sources to obtain the header file is necessary. This can be achieved with the `javac` command specifying the `-h` option.

Using the header file, it is possible to create a C file to call the Expand client functions. These functions require capturing the parameters specified in the signatures and translating them to C-intelligible values. Then, the Expand client functions are invoked, and the result returned by the server is converted so that it can be passed to the clients.

Finally, it is possible to compile these last sources as any C program by indicating the Expand dependencies. That way, a dynamic library can be obtained, the Expand servers can handle the incoming Java requests, and their responses can also be translated.

2.2. Expand Hadoop layer

The main goal of this layer is to convert Expand to a Hadoop Compatible File System (HCFS) so that the Spark tasks can handle the clients. To achieve this goal, extending the `File System` class available in the Hadoop project [9] is necessary. This class contains the abstract methods shown in Table 2.2.1 that must be overwritten. The functions shown in this table allow Spark to perform I/O and metadata operations such as list file status (`getFileStatus`), list directory status (`listStatus`), create directories (`mkdirs`), read files (`open`), write files (`append` and `create`), rename files (`rename`), or delete files and directories (`delete`).

Table 2.2.1: Overwritten Hadoop File System abstract methods.

Function	Parameters
`append`	`Path f, int bufferSize, Progressable progress`
`create`	`Path f, FsPermission permission, boolean overwrite, int bufferSize, short replication, long blockSize, Progressable progress`
`delete`	`Path f, boolean recursive`
`getFileStatus`	`Path f`
`getUri`	
`getWorkingDirectory`	
`listStatus`	`Path f`
`mkdirs`	`Path path, FsPermission permission`
`open`	`Path f, int bufferSize`
`rename`	`Path src, Path dst`
`setWorkingDirectory`	`Path new_dir`

Besides, the functions shown in Table 2.2.2 must be overwritten to ensure the correct operation and improve the system performance. In these functions, Expand clients can be raised (`initialize`), paths can be classified into a file or a directory (`isDirectory`), task locality is implemented (`getFileBlockLocations`), and files and directories permissions can be modified (`setPermission`).

Table 2.2.2: Overwritten Hadoop File System methods.

Function	Parameters
`initialize`	`URI uri, Configuration conf`
`isDirectory`	`Path f`
`getFileBlockLocations`	`FileStatus file, long start, long len`
`setPermission`	`Path path, FsPermission perm`

3. FAULT TOLERANT SUPPORT AND MALLEABILITY

3.1. Fault tolerance

The fault tolerance design in Expand Ad-Hoc is based on block replication. This means that the data is replicated as many times as the replication factor that the user selected. The replication factor is the number of copies of a block.

This approach allows a failure in up to $N - 1$ ad-hoc servers, where N is the replication factor. This is possible because, with any N replication factor, the blocks and the metadata are stored in N ad-hoc servers, as shown in Figure 3.1.1.

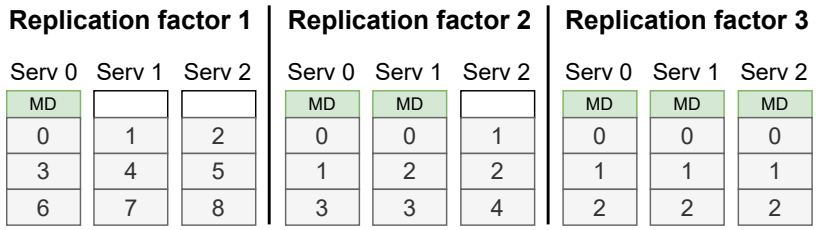


Figure 3.1.1: Data block distribution with replication factor 1, 2, and 3 in Expand Ad-Hoc.

Given a replication factor R, where $R \in \mathbb{N}$, the block mapping function is formally expressed as follows when Expand Ad-Hoc uses replication:

$$f_{mappingRepl} : Off \rightarrow (Serv \times Off)^R$$
$$f_{mappingRepl}(off) = ((serv_i, off_j) \ldots (serv_k, off_l))$$

For $R = 2$, in other words, a replication factor of 2, the block mapping function is:

$$f_{mappingRepl}(off) = ((serv_0, off_0), (serv_1, off_1))$$

Where off represents a file offset, off_0 is the subfile offset in $serv_0$, and off_1 is the subfile offset in $serv_1$ where the off is replicated.

In a round-robin distribution pattern, all the blocks are distributed among the ad-hoc, as illustrated in Figure 3.1.1 with 1, 2, and 3 replication factors. As with the previous scenario, the function that performs this mapping with replication is of complexity $O(1)$, which can be seen in Algorithm 1. This algorithm returns only the i element of the tuple $((serv_i, off_j) \ldots (serv_k, off_l)$.

There are two possible approaches to perform the detection of the servers with error. One uses any MPI implementation by default, which can detect when an ad-hoc client attempts to connect to one ad-hoc server and the server has problems. It is marked as erroneous, and the client will not use that server. Another approach uses the OpenMPI 5.0 implementation that offers the module ULFM (User-Level Fault Mitigation) [11]. With this

Algorithm 1 : Algorithm for the mapping function for round-robin and R replication factor, that only calculate the i element of the tuple $((serv_i, off_j) \dots (serv_k, off_l))$

function A_MAPREPL(off, i)
  $block = off/b\_size$
  $b\_repl = block \cdot R + i$
  $b\_line = b\_repl/nserv$
  $serv = (b\_repl + first\_node) \mod nserv$
  $serv\_off = b\_line \cdot b\_size + (off \mod b\_size)$
  return $(serv, serv\_off)$
end function

module, the ad-hoc client can detect in the middle of the application that if one communication with a server fails, it marks that server as erroneous to stop using it.

Thanks to the replication factor implemented with the fault tolerance, we have developed an optimization in the read operations that takes advantage of the increase in data locality. This is achieved by having the ad-hoc client check whether the data requested by the user is replicated in one ad-hoc server deployed in the same compute node. If that is the case, the ad-hoc client can directly access the data in the local storage, avoiding the communication overhead with the ad-hoc server.

3.2. Malleability

Three distinct malleability algorithms were developed, each employing a unique methodology and showing different advantages and disadvantages.

The primary distinction between these algorithms lies in the method utilized to move the data from the existing partition to the newly configured one. All the algorithms stop and start the servers to facilitate the brand-new start of the new partition.

It is important to note that, as previously explained, all malleability algorithms take into account the data replication utilized in the fault tolerance model. In the event that a server is down, data will be obtained from replications on other servers.

In a formal context, the malleability operation can be defined as the movement of a partition, stored in a set of servers designated as S, to another set of servers, designated as R. The number of servers in each set allows for classifying this operation as either expand or shrink: an expand operation occurs when $|S| < |R|$, and a shrink operation occurs when $|R| < |S|$.

The relationship between S and R can be conceptualized in the following manner:

$$S \subset R$$

$$R \subset S$$

$$S \cap R = \emptyset$$

To perform this operation, we formally define the following malleability function:

$$f_{malleablity} : (S, Off)^P \rightarrow (R, Off)^Q$$
$$f_{malleablity}(Serv_i, Off_i)^P = (Serv_j, Off_j)^Q$$

Applying this function to each file enables the transfer of data block-by-block from a partition S with the replication factor P to a partition S with replication factor Q.

Given a malleability mapping, we define the cost function of this malleability function as follows:

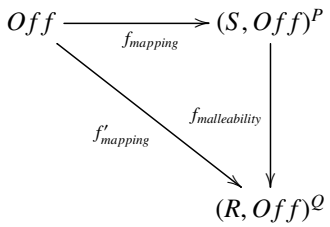
$$Cost : (f_m, S, R, File) \mapsto \mathbb{N}$$

This function determines the number of operations required to move the file designated as $File$ from partition S to partition R where the malleability function is f_m.

In order to accurately execute the cost function, it is necessary to obtain the blocks in a file, taking into account the replication factor. The function is defined as follows:

$$f_{blocks}(File) \rightarrow \mathbb{N}$$

Given a malleability $f_{malleability}$ and a block mapping function $f_{mapping}$, both functions are related as follows:



The new mapping function in the new partition can be defined as:

$$f'_{mapping} = (f_{malleability} \circ f_{mapping})$$

The following sections present three different algorithms, ordered according to their efficiency calculated with this cost function.

3.2.1. Malleability: backend rebuild

The initial algorithm employs the supercomputer's backend file system to facilitate the data transfer process. This is achieved by first transferring all data from the current Expand Ad-Hoc partition to the backend file system via an MPI application that runs in parallel on all the nodes that contain servers. Upon completion of this transfer, the data is then transferred in the opposite direction to the new configuration of the partition. A more illustrative representation can be observed in Figure 3.2.1.

The cost of this malleability function can be defined as follows:

$$Cost(f_{backendRebuild}, S, R, File) = f_{blocks}(File) \times 2$$

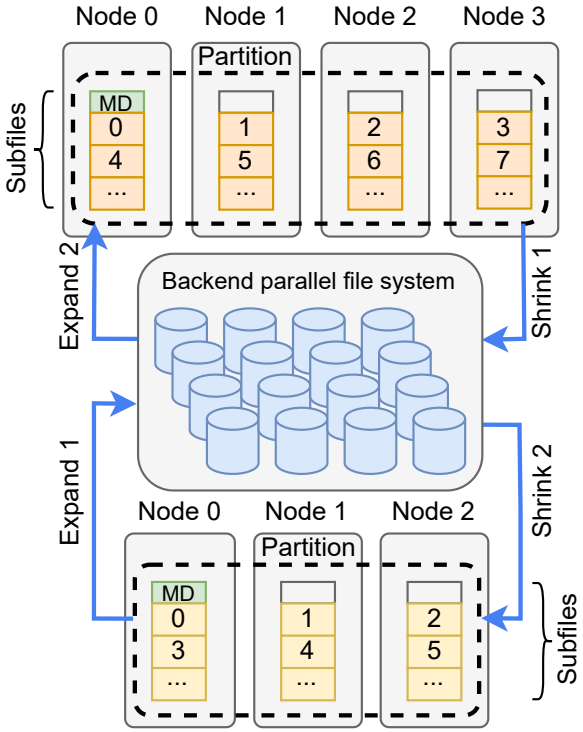


Figure 3.2.1: Malleability: backend rebuild.

This algorithm relies on the supercomputer's backend file system for both data writing and reading, so it depends on that system's bandwidth. As a result, the transfer speed from one Expand Ad-Hoc partition to another is dependent on the speed of the backend file system. It is evident that this approach is not the most efficient one. Therefore, we developed the next algorithm, which is explained in the following section.

3.2.2. Malleability: direct rebuild

We developed a new MPI application with this next algorithm to eliminate dependency on the backend file system. This MPI application directly transfers the data blocks from the old Expand Ad-Hoc partition to the new one. To make this possible, we consider the two sets of nodes:

$$S = \{S_1, S_2, ..., S_k\}$$
$$R = \{R_1, R_2, ..., R_r\}$$

The variable S refers to the set of servers in the old partition, and R corresponds to the set of servers in the new partition.

The MPI application starts a number of processes equal to $|S|+|R|$. For the sake of convenience, the processes

in the set S are designated as *readers* having $|S|$ as the number of *readers*, while the set of processes in R are designated as *writers* having $|R|$ as the number of *writers*. The *readers* are responsible for reading each block, calculating its new position and server, and sending it to the corresponding *writer* process. Consequently, the *writers* are awaiting the reception of the blocks and their respective positions. Upon the receipt of a block, the writer only writes it.

The cost of this malleability function can be defined as follows:

$$Cost(f_{directRebuild}, S, R, P, File) = f_{blocks}(File)$$

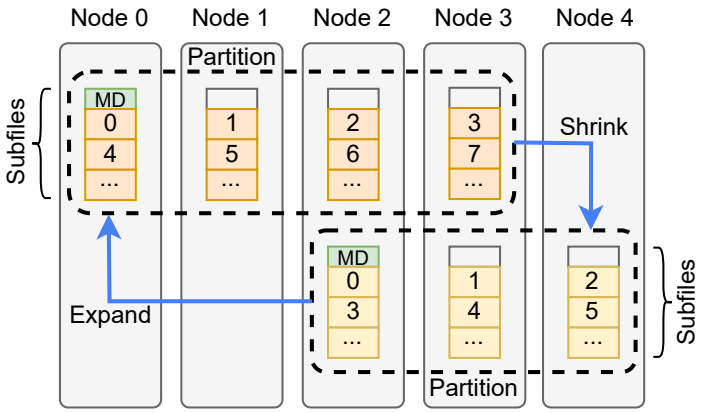


Figure 3.2.2: Malleability: direct rebuild.

One example of this malleability operation can be observed in Figure 3.2.2, which illustrates the expanding and shrinking of servers within the Expand Ad-Hoc malleability. In this example, we can see the expanding of three servers to four servers, as well as the shrinking of four servers to three servers. For the expanding, the set of readers is $S = \{Node_2, Node_3, Node_4\}$ and the set of writers is $R = \{Node_0, Node_1, Node_2, Node_3\}$, so the number of process to run is $|S| + |R| = 7$, being readers $|S| = 3$, and writers $|R| = 4$. Likewise, for the shrinking, the set of readers is $S = \{Node_0, Node_1, Node_2, Node_3\}$ and the set of writers is $R = \{Node_2, Node_3, Node_4\}$, so the number of processes to run is $|S| + |R| = 7$, being readers $|S| = 4$, and writers $|R| = 3$.

3.2.3. Malleability: metadata rebuild

We realized that the previous algorithm was not the most efficient design, leading to the creation of this newer and more efficient algorithm. The objective of this algorithm is to move only the minimal and necessary data from one partition to another.

For example, when there is an expand operation in which only additions to the new partition are made, the data can be accessed through the old servers and does not need to be moved. In other cases, when there is a shrink operation, only the data that cannot be accessed because the servers are removed must be moved.

This function's cost depends on whether the operation is of expand or shrink type. The cost for an expand

operation is defined as follows:

$$Cost(f_{metadataRebuild}, S, R, P, File) = 0$$

The cost for a shrink operation can be defined as:

$$Cost(f_{metadataRebuild}, S, R, P, File) = \frac{f_{blocks}(File)}{|S|} * (|S \setminus R|)$$

In order to achieve this malleability, we take advantage of the metadata and save the sections of the file in which the malleability was performed. A new mapping function with a $O(n)$ complexity was developed for the new file structure, where n is the number of reconstructions. This allows the data to be located within the files, regardless of the number and order of malleability operations. Despite the complexity being $O(n)$, the typical number of server reconstructions is low, resulting in a time complexity close to that of the previous mapping function with a complexity of $O(1)$.

The formality of this mapping function is identical to that of the mapping with replication, as it also takes into account data replication.

$$f_{mappingMalleability} : Off \rightarrow (Serv \times Off)^R$$
$$f_{mappingMalleability}(off) = ((serv_i, off_j) \ldots (serv_k, off_l))$$

For $R = 2$, in other words, a replication factor of 2, the block mapping function is:

$$f_{mappingMalleability}(off) = ((serv_0, off_0), (serv_1, off_1))$$

The mapping algorithm consists of a loop that iterates over each malleability section defined in the metadata, calculating the position and server of the blocks. For performance reasons, the algorithm initially verifies the absence of malleability and redirects to the original function, thereby maintaining the complexity in constant time ($O(1)$). If malleability is present, the algorithm iterates through each segment, calculating whether the block is located within the actual segment and, therefore, calculating its position and server. A segment is the section of the file that has been written to a partition. For example, if 3 blocks are written in one partition and malleability is performed to another partition, in which 2 new blocks are written, we would have two segments in the file, one of 3 blocks and one of 2 blocks. It can be a shrinking or expanding segment when the shrink or expand malleability operation has been performed respectively. In the case of the block situated within a shrinking segment, a further recalculation is performed to determine the new position and server. A more visual representation of the mapping algorithm can be found in Algorithm 2.

So, the MPI application developed to perform this algorithm is responsible for calculating the data that needs to be inserted into the metadata. If it is a shrink operation, it is also responsible for moving only the data from the server that will be removed to the other ones. The data movement is performed similarly to the previous malleability algorithm, in which the servers that will be removed are the *readers* and the others are the *writers*.

For illustrative purposes, three examples of the new mapping can be found in Figures 3.2.3, 3.2.4, and 3.2.5.

Algorithm 2 : Algorithm for the malleability mapping function with metadata for round-robin and R replication factor, returning only calculate the i element of the tuple $((serv_i, off_j) \dots (serv_k, off_l))$.

```
function A_MAPPINGMALLEABILITY(off, i)
    using F_mR = F_mappingRepl
    if not malleability then
        return F_mappingRepl(off, i)
    end if
    for segment in segments do
        data_nserv = mdata.data_nserv[i]
        offset = mdata.offset[i]
        calculate_segment_serv_off(off, i)
        if block in segment then
            serv_off, serv = F_mR(off, i)
            serv_off = add_segment_serv_off()
        end if
        if block in segment_shrink then
            serv_off, serv = F_mR(serv_off, i)
            serv_off = add_segment_serv_off()
        end if
    end for
    return (serv ,serv_off)
end function
```

As shown in Figure 3.2.3, the expanding process can be observed. Figure 3.2.4 presents the process of shrinking. Figure 3.2.5 demonstrates a combination of expansion and contraction. In these figures, the metadata fields "data_nserv" refer to the number of servers present in each segment, while "offset" specifies the block at which the segment begins.

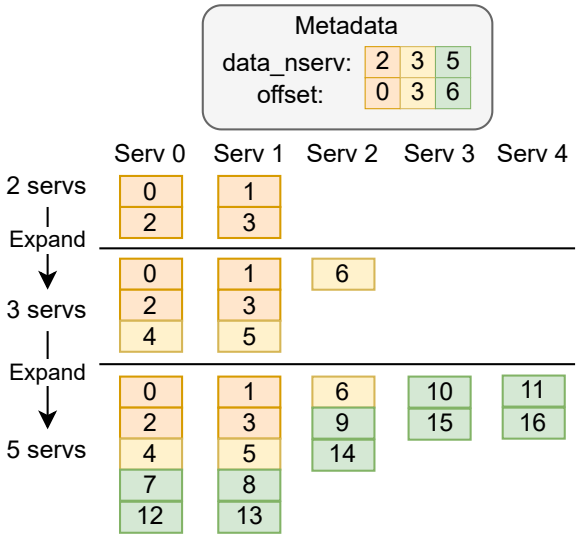


Figure 3.2.3: Malleability: metadata rebuild using expand. From 2 to 3 to 4 servers.

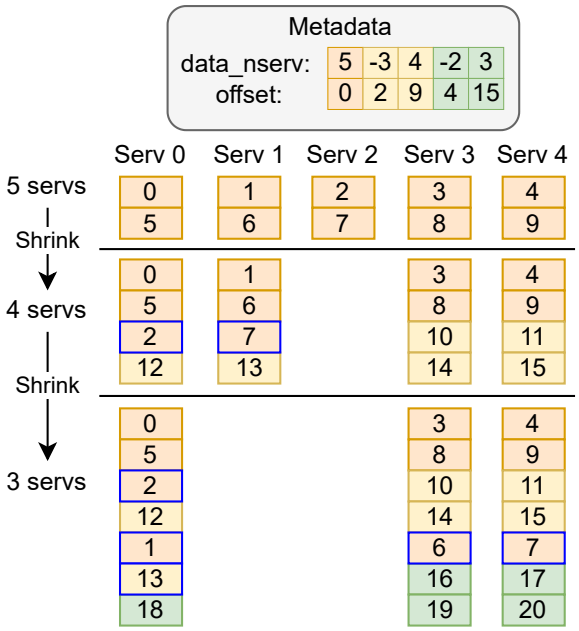


Figure 3.2.4: Malleability: metadata rebuild using shrink. From 5 to 4 to 3 servers.

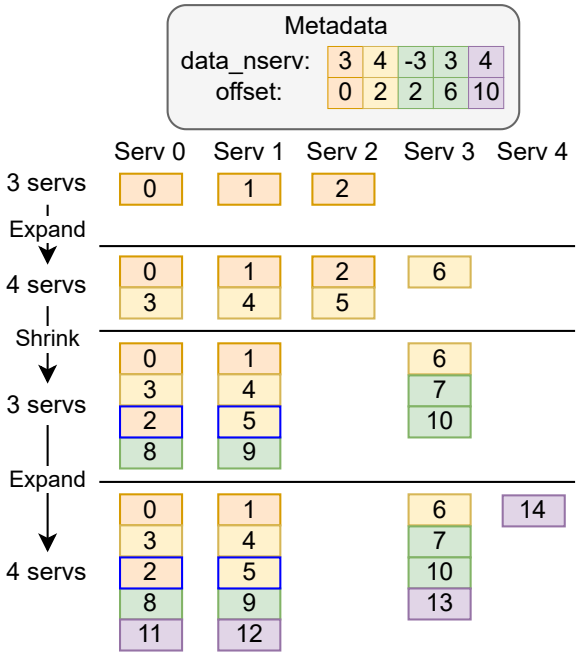


Figure 3.2.5: Malleability: metadata rebuild using both expand and shrink. From 3 to 4 to 3 to 4 servers.

BIBLIOGRAPHY

[1] F. Garcia-Carballeira, A. Calderon, J. Carretero, J. Fernandez, and J. M. Perez, "The design of the Expand parallel file system," *The International Journal of High Performance Computing Applications*, vol. 17, no. 1, pp. 21–37, 2003.

[2] F. García-Carballeira, J. Carretero, A. Calderón, J. D. García, and L. M. Sanchez, "A global and parallel file system for grids," *Future Generation Computer Systems*, vol. 23, no. 1, pp. 116–122, 2007. DOI: https://doi.org/10.1016/j.future.2006.06.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X06001282.

[3] J. A. Zounmevo, D. Kimpe, R. Ross, and A. Afsahi, "Using MPI in high-performance computing services," in *Proceedings of the 20th European MPI Users' Group Meeting*, 2013, pp. 43–48.

[4] L. A. N. L. Howard Pritchard, *Open MPI and recent trends in network APIs*, Accessed Nov. 9, 2023. [Online], 2016. [Online]. Available: https://www.openfabrics.org/images/eventpresos/2016presentations/110mpiandapi.pdf.

[5] A. MPICH, *MPICH readme*, Accessed Nov. 9, 2023. [Online], 2023. [Online]. Available: https://www.mpich.org/static/downloads/4.1.2/mpich-4.1.2-README.txt.

[6] M. A. Open, *Open MPI readme*, Accessed Nov. 9, 2023. [Online], 2023. [Online]. Available: https://docs.open-mpi.org/en/main/tuning-apps/networking.

[7] A. B. Yoo, M. A. Jette, and M. Grondona, "Slurm: Simple linux utility for resource management," in *Job Scheduling Strategies for Parallel Processing*, D. Feitelson, L. Rudolph, and U. Schwiegelshohn, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 44–60.

[8] B. K. R. Vangoor *et al.*, "Performance and resource utilization of FUSE user-space file systems," *ACM Transactions on Storage (TOS)*, vol. 15, no. 2, pp. 1–49, 2019.

[9] A. Hadoop, *Github repository*, 2024. [Online]. Available: https://github.com/apache/hadoop.

[10] Apache Software Foundation, *Welcome to Apache Maven*, 2024. [Online]. Available: https://maven.apache.org/.

[11] W. Bland, A. Bouteiller, T. Herault, G. Bosilca, and J. Dongarra, "Post-failure recovery of mpi communication capability: Design and rationale," *The International Journal of High Performance Computing Applications*, vol. 27, no. 3, pp. 244–254, 2013. DOI: 10.1177/1094342013488238.