Disk Farm Installation and Administration Guide

Version 1.6

Krzysztof Genser, Tanya Levshina, Igor Mandrichenko, Miroslav Siket

Fermi National Accelerator Laboratory

Table of Contents

Terminology 3
What is Disk Farm ? 3
Disk Farm Design 3
Installing Disk Farm 4
Requirements 4
FUE Installation 4
Disk Farm Configuration 6
Set vfssrv 6
Set cell 6
Set storage 7
Set cell_class 8
Set quota 8
Running Disk Farm 8
Maintenance and Troubleshooting 9
Checking Disk Farm Status 9
Shutting Disk Farm Down 10
Backup Procedures 10
Moving Data between PSAs 10
Disk Farm Reconfiguration 11

Terminology

In this document,

Disk Farm or dfarm – name of the product.

Disk farm – particular instance of Disk Farm, collection of nodes, this instance consists of.

Disk farm cell – individual disk farm node.

What is Disk Farm ?

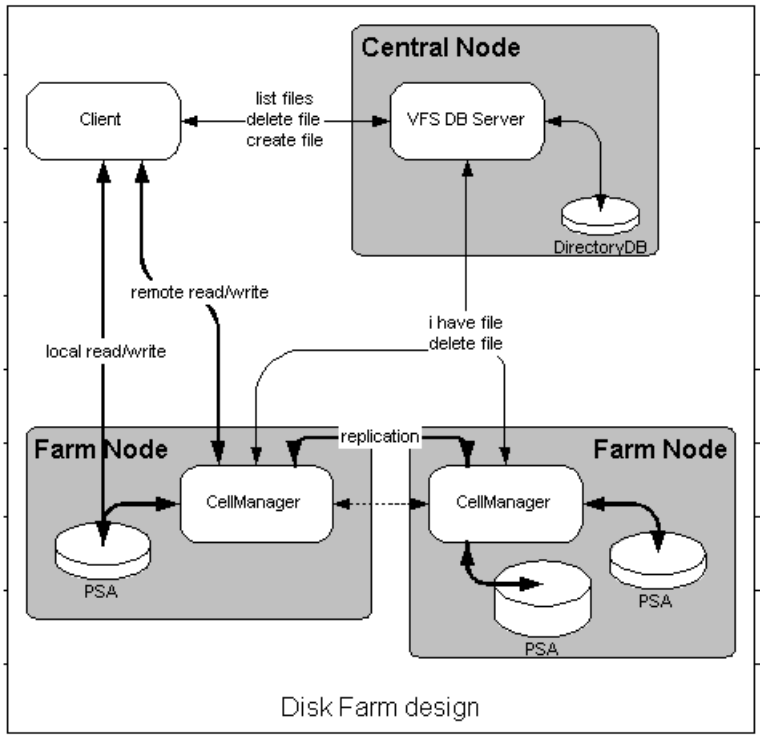
Disk Farm is a distributed data storage system. It is designed to organize the space available on individual nodes of a computing farm into storage system for temporary data. The purpose of Disk Farm is to compliment, not to replace other data storage systems which should be used for permanent data storage. Disk Farm itself does not have any data backup/recovery functionality. Data stored in disk farm can be permanently lost or become temporarily unavailable due to a failure of a disk or one of farm computers. Therefore, disk farm should be used only as temporary storage of reproducible data, or data with limited expected lifetime.

Disk Farm users can use data replication feature to increase availability of data stored in disk farm. If required, additional data recovery functions can be implemented as a set of external administrative tools.

Disk Farm product includes an example of VFS database backup procedure as described below.

Disk Farm Design

Disk Farm system consists of Virtual File Name Space (VFS) Database Server (VFS DB Server, vfssrv), Cell Managers (cellmgr) and user interface executables. VFS DB Server runs under non-privileged account on one (preferably, most reliable) computer of the disk farm. One instance of Cell Manager runs as root on each node of the disk farm.



VFS DB Server is responsible for maintaining virtual file name space database. The database is stored on the central node of the disk farm, under the directory specified in the Disk Farm configuration.

Cell Manager stores and maintains user data on one or more local disks organized into Physical Storage Areas (PSAs). Also, Cell Manager maintains physical file location database. This database maps virtual file path into physical location of the file in one of PSAs.

Installing Disk Farm

Disk Farm (dfarm) distribution complies with Fermilab UNIX Environment (FUE) standards and is packaged as UPD/UPS product. This document describes installation procedure in FUE environment.

Requirements

Disk Farm requires that the following products are installed on the system:

- Python 1.5 or 2.1
- FCSLIB library package available from Fermitools

Currently, Disk Farm is supported on Linux, IRIX, OSF1 and SunOS. However, the product is highly portable and most likely will run on other flavors of UNIX.

All the nodes of a disk farm must have the same IP broadcast number, which means that they have to be on the same IP subnet.

FUE Installation

1. Download and install the product distribution. The simplest way to install Disk Farm is to use UPD:

```
upd install –R dfarm
```

Disk Farm has to be installed for each platform of OS the cluster computers run under. See UPD documentation for instructions on product installation for multiple platforms. "upd install –R dfarm" command will install Python and FCSLIB products unless they have not been installed already.

Also, Disk Farm and FCSLIB distribution packages for each platform can be downloaded from fnkits product repository

2. After dfarm product is installed, it has to be declared with UPS for each flavor of the cluster, for example:

```
ups declare –f IRIX -c -m dfarm.table -r dfarm/v1_6/IRIX \
             dfarm v1_6
```

It is important to include the specification of UPS action table file with "-m dfarm.table".

3. Create or designate an account to run Disk Farm VFS DB server. VFS DB server does not have to run as root. This account will be referred to as <dfarm-user>.

4. Create a directory, preferably in NFS-shared area, for Disk Farm configuration files, databases and utility scripts. This directory should be shared by all nodes of the disk farm and owned by <dfarm-user>. This directory will be referred to as <DFARM_ROOT> or $DFARM_ROOT.

Typically, <DFARM_ROOT> is a directory under NFS-shared home area of <dfarm-user>. At FNAL, it usually is ~farms/dfarm_root.

5. Tailor the product for each UPS flavor using "ups tailor" command:

```
ups tailor –f IRIX –O <DFARM_ROOT> dfarm v1_6
```

Tailoring option (-O) is the path to <DFARM_ROOT>. This will create scripts $DFARM_DIR/bin/setup_dfarm.sh and $DFARM_DIR/bin/setup_dfarm.csh.

6. Under <DFARM_ROOT>, create subdirectories:

- db – for disk farm VFS database
- cfg – for configuration file(s)
- bin – for maintenance scripts

Optionally, create subdirectories:

- backup – for VFS database backups
- log – for VFS database server log files

Subdirectories db, backup, log must be owned by the VFS DB server account. Other subdirectories must be readable by VFS DB server.

7. Copy template of dfarm configuration file from $DFARM_DIR/cfg/dfarm.cfg to <DFARM_ROOT>/cfg directory. Edit the template according to your configuration. Configuration file format is described below.

8. Copy maintenance scripts $DFARM_DIR/bin/start*.sh, …/kill*.sh, …/restart*.sh into <DFARM_ROOT>/bin directory:

```
cd <DFARM_ROOT>/bin
cp $DFARM_DIR/bin/start*.sh .
cp $DFARM_DIR/bin/kill*.sh .
cp $DFARM_DIR/bin/restart*.sh .
chmod +x *.sh
```

Unless your UPS installation bootstrap scripts setups.sh and setups.csh are located in "standard" location /fnal/ups/etc, edit the maintenance scripts in <DFARM_ROOT>/bin following instructions found in those files.

Disk Farm Configuration

Disk Farm configuration is defined in the configuration file located in <DFARM_ROOT>/cfg/dfarm.cfg. The file consists of sections or "sets". Each set defines a group of parameters. Some parameters are optional, others are required.

Set vfssrv

This is a set of parameters for VFS Server. The parameters are:

- host – (required) IP address of the computer running VFS Server
- cellif_port – (required) TCP server port number for the port used for communication between VFS Server and Cell Managers. This should be some number greater than 1024.
- api_port – (required) TCP server port number for the port used for communication between disk farm clients and VFS Server. This should be some number greater than 1024. It must be different from cellif_port.
- db_root – (required) Path to the VFS database directory (usually it is <DFARM_ROOT>/db)
- log – (optional) Path and name of VFS Server log file. If left unspecified, no log file will be produced. Log file will be closed and reopened every day. Log files from previous days will be kept with suffixes ".1", ".2", etc. appended to the file name.

Example:

```
%set vfssrv
host = fnsfo.fnal.gov
cellif_port = 4568
api_port = 4569
db_root = /home/farms/dfarm_root/db
log = /tmp/vfssrv.log
```

Set cell

This set defines common configuration parameters for disk farm nodes (cells):

- listen_port – (required) UDP server port number used for communication between dfarm clients and Cell Managers. This should be some number greater than 1024.
- broadcast – (required) Broadcast address for the disk farm subnet
- farm_name – (optional) Logical name of the disk farm instance. In cases when it is necessary to run more than one instance of disk farm within the

same subnet (with the same IP broadcast address), this parameter will be used to associate individual nodes with particular instance. If this parameter is omitted, then all nodes with given IP broadcast address will be considered in the same disk farm instance.

- domain – (optional) Common suffix of IP node names in the disk farm. Usually this is IP domain name. This parameter is used to translate IP host names into logical names of the disk farm nodes. For example, if IP node name is "fnpc123.fnal.gov", and domain name is defined to be "fnal.gov", disk farm logical node name will be "fnpc123". Another example: if node name is "srv1.phys.nwu.edu", and domain is "nwu.edu", logical node name will be "srv1.phys". If this parameter is omitted, then full IP node names will be used as logical names.
- log – (optional) Path and name of Cell Manager log file. If left unspecified, no log file will be produced. Log file will be closed and reopened every day. Log files from previous days will be kept with suffixes ".1", ".2", etc. appended to the file name.
- max_get, max_put, max_rep – (optional) maximum allowed number of concurrent transactions of corresponding types (get, put and replication) for each individual disk farm node (cell). Their defaults are 3, 1 and 2 respectively.
- max_txn – (optional) maximum allowed total number of all types of transactions for each disk farm node. The default is 3.

Example:

```
%set cell
listen_port = 4567
broadcast = 131.225.167.255
farm_name = fixed-target
domain = fnal.gov
log = /tmp/cellmgr.log
```

Set storage

Disk Farm may consist of computers with different disk capacities or other characteristics. In this case, similar computers are grouped into *Storage Classes*. Storage Classes are assigned logical names. Disk Farm configuration file must have one "storage" set for each Storage Class. Set "storage" consists of definitions of *Physical Storage Areas (PSA)* allocated to the Disk Farm on each node of the class. For each PSA, set "storage" has a line in the following format:

```
<PSA name> = <PSA top directory> <allocated size in MB>
```

For example, the following two sets define storage classes "single" and "double". Nodes of class "single" have 1 30GB PSA, while nodes of class "double" have 2 8GB PSAs.

```
%set storage big_computer
st2 = /local/stage2/dfarm 30000

%set storage small_computer
st1 = /local/stage1/dfarm 8000
st2 = /local/stage2/dfarm 8000
```

Disk Farm PSA is organized into 2 directory sub-trees under the directory specified in the configuration file:

- Subdirectory "info" is used to store metadata.
- Actual data is stored under subdirectory "data".

These 2 subdirectories are created and initialized by Cell Manager when it starts.

Set cell_class

This set defines the association between disk farm node names and storage classes they belong to. This set has one line per node in the format:

```
<logical node name> = <storage class name>
```

For example:

```
%set cell_class
fnpc1 = big_computer
fnpc2 = big_computer
fnpc101 = small_computer
fnpc3 = big_computer
```

Set quota

If necessary, Disk Farm administrator can set disk farm space utilization quotas for individual users. If present, set "quota" has one line per user in the format:

```
<username> = <quota in Megabytes>
```

Here, "username" is the UNIX username of the user the quota is to be applied to. Special "username" "*" can be used to implicitly specify quota for all users except explicitly mentioned. For example:

```
%set quota
alice = 100000     # 100 GB for Alice
bob = 70000        # 70 GB for Bob
* = 50000          # 50 GB for all others
```

This set may be omitted entirely. In this case, every user has unlimited quota.

Running Disk Farm

In order to run Disk Farm, the administrator has to start the following processes:

- VFS Server runs on one of disk farm nodes, referred to as "central node". This process maintains Virtual File Space (VFS) database and user space utilization/quota information. It does not have to run as root. It has to have write permissions for <DFARM_ROOT>/db directory.

In order to start VFS Server, use script <DFARM_ROOT>/bin/start_vfssrv.sh

- One instance of Cell Manager must run as root on all disk farm nodes defined in the configuration as members of Storage Classes. Cell Managers are responsible for storing and maintaining data on their nodes.

  In order to start Cell Manager, use script <DFARM_ROOT>/bin/start_cellmgr.sh

Maintenance and Troubleshooting

Checking Disk Farm Status

The quickest way to see if Disk Farm is running is to issue 2 commands:

```
$ dfarm ping
     fnpc72.fnal.gov     1ms    0w    0r
     fnpc80.fnal.gov     1ms    1w    0r
     fnpc69.fnal.gov     1ms    0w    1r
     fnpc62.fnal.gov     1ms    0w    2r
     …
--- 90 nodes ---------------------------
               total            5w    3r
                 min     1ms
             avegare     1ms
                 max     1ms
$ dfarm ls
fnsfo> dfarm ls
drwr-   - alice                      -               /alice/
drwr-   - bob                        -               /bob/
```

First command ("dfarm ping") shows basic status information about each node of the farm. It shows IP node name, its response time and how many "read" and "write" operations are in progress at the time. If a cell manager on a node is not running, the node will not be shown in the list. However, another (unlikely) reason for a node to be missing is that it is simply too busy to answer in time.

Second command lists top of VFS name space tree. If VFS Server was not running, this command would fail. Therefore, if you see some reasonable output, it indicates that VFS Server is up and running.

Another way of checking whether Disk Farm components are running is to use UNIX "ps" command:

```
$ ps -ef | grep vfssrv
   farms       1928          1  … /bin/sh -f /…/vfssrv.sh
   farms       2059       1928  … python /…/vfssrv.py

$ ps axuw | grep cellmgr
root       969  0.0  0.1  1632 … sh -f /…/cellmgr.sh
root      1190  0.0  0.3  4304 … python /…/cellmgr.py
```

More detailed status information for an individual disk farm node can be obtained with "dfarm stat" command:

```
fnsfo> dfarm stat fnpc51
            Area       Size       Used   Reserved       Free
---------------- ---------- ---------- ---------- ----------
             st2      30000        170          0      34026
Txn type Status VFS Path
-------- ------ --------
      RD      * /ivm/qq2
```

First part of the output lists all PSAs on the node and disk space utilization statistics for them. For each PSA, it shows its logical size, how much disk space is in use currently, how much is reserved for current incoming transactions, and how much space is physically available on the physical volume the PSA is located on. All these numbers are in Megabytes.

Second part shows all current I/O transactions for the node. First column shows transaction type (RD – read, WR – write and RP – outgoing replication). Second column is status (* - in progress, I – initialized). Third column is VFS path of the file being transferred.

Shutting Disk Farm Down

In order to shut down disk farm components, the administrator should use scripts kill_cellmgr.sh and kill_vfssrv.sh located in <DFARM_ROOT>/bin directory.

Backup Procedures

Critical piece of information which may be considered for periodic backups is VFS Database normally located under <DFARM_ROOT>/db. This directory may be periodically recursively archived and restored when recovery is necessary. Disk Farm distribution includes a template of backup creation and maintenance script <DFARM_ROOT>/bin/db-backup.cron.sh. Some modification of this script may be necessary to reflect specifics of the installation. It is recommended to run this script as a cron job. Frequency of backup should be determined based on the requirements of the specific installation. Every time the backup script runs, it creates a database snapshot and stores it in compressed tar file under the directory specified in the beginning of the script.

It is recommended to use "tar xfvN <tarfile>" command to restore contents of VFS database.

Moving Data between PSAs

Under certain circumstances, it may be necessary to move data from one PSA to another, for example, if a disk farm node is about to be shot down. In order to move data from one PSA to another:

1. Log in as root

2. Make sure the destination PSA has enough room for the data. Use "dfarm stat" command.

2. Shut down both source and destination Cell Managers

3. Create tar file with original data:

```
cd /stage1/dfarm         # cd to the top of source PSA
tar cf /tmp/data.tar .
```

4. Move the tar file to the destination computer

5. Unwind the tar file into the destination PSA

```
cd /stage2/dfarm         # cd to the top of destination PSA
tar xf /tmp/data.tar
```

6. If necessary, delete source data on the source computer

```
cd /stage1/dfarm         # cd to the top of source PSA
rm –rf /stage1/data/*
rm –rf /stage1/info/*
```

7. Start destination Cell Manager and, if necessary, source Cell Manager

Disk Farm Reconfiguration

Disk Farm configuration file modifications take effect only after corresponding daemon process is restarted. If you are not sure which component should be restarted, restart all of them.