WHITE PAPER–OCTOBER 2019

Integration and Advanced Usage

Bitfusion Guide

vmware®

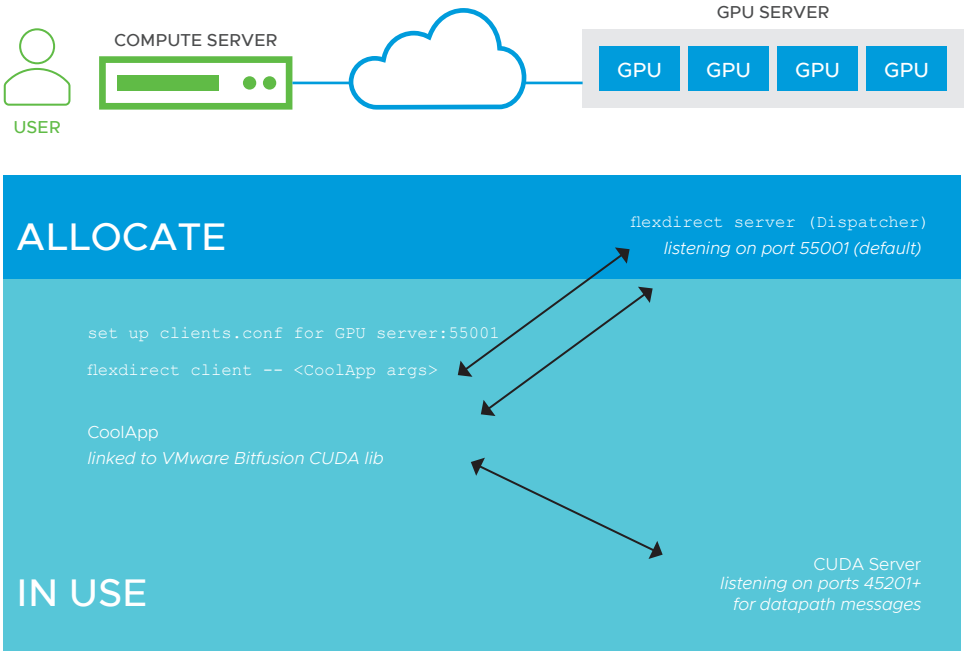
Table of contents

Introduction

For engineers integrating Bitfusion technology into their own resource scheduler or perhaps for advanced users needing more control over GPUs are resourcing, this guide shows how to start and invoke both server and client processes with low-level access. You will start a server daemon for a particular GPU configuration (e.g., partial memory) and write a client-side configuration file 'adaptor.conf' as shown in the examples below. We have done integrations for several job schedulers and resource managers, so *contact us* if you're looking for help.

Starting FlexDirect Daemons as Servers via CLI

The drawing below shows the four processes that are running on a client (or CPU) node and on a server (or GPU) node when you are interacting with the FlexDirect Server (Dispatcher). It should help you understand the concepts, commands and usage that this manual discusses. Only two processes are directly launched by the user. These are the ones shown in a fixed font as you would type them in a command shell. The drawing also shows the TCP ports used by the GPU server processes.



You must start FlexDirect as a server (which is called Dispatcher) on all the instances that have GPUs which you'd like to make available to your client nodes and applications.

Shell

```
flexdirect server [-p port]
```

You can also start a FlexDirect server (Dispatcher process) from the client machine with the `request_gpus` command. However, this requires that the GPU server is already running the resource scheduler. Advantages include:

- Prevents multiple users from trying to serve the same GPUs
- Creates adaptors.conf file for you
- Does not automatically deallocate the GPUs after a client application has finished so you can run several applications sequentially

However, this document covers manual launches of the FlexDirect server.

Run Client Applications with FlexDirect via flexdirect client

Once the FlexDirect Servers are running, run applications using `flexdirect client`. Pass the `-l` parameter as a list of the IP addresses of the nodes on which you have FlexDirect Server running. Use semicolons to separate the addresses. Replace `<application>` with the application you would like to run. Use a double dash `--` before the application if it requires its own arguments.

Shell

```
flexdirect client -l “172.31.51.20; 172.31.51.26” [--] <application>
```

You may specify a port number with the standard colon notation:

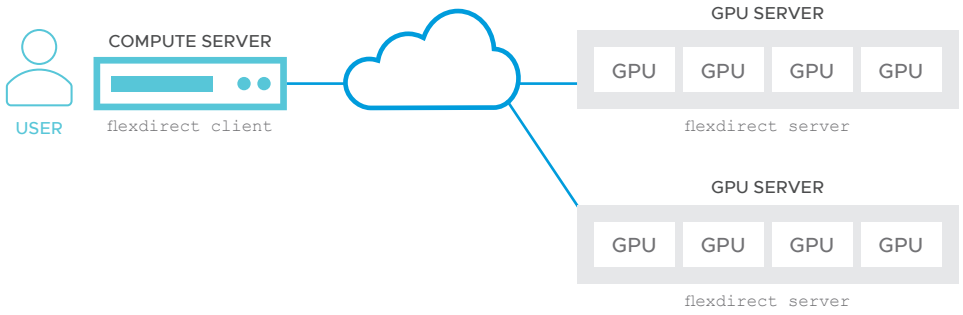
CPU Server Command Line

```
flexdirect client -l 172.31.51.20:55002 nvidia-smi
```

GPU Server Command Line

```
nvidia-smi
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 375.26                 Driver Version: 375.26                    |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla K80    Off  | 0000:01:00.0     Off |                  N/A |
| N/A%   53C    P8    29W /  149W |      0MiB /  11439MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID  Type  Process name                               Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```



Configuring IP Addresses as Part of Client Configuration

If you want to simplify the flexdirect client command, you can put your Bitfusion server IP addresses into the /etc/bitfusionio/adaptor.conf file. Override the default port by adding `:<port>`.

CPU Server Command Line

```
cat /etc/bitfusionio/adaptor.conf
172.31.51.20
172.31.51.26:57001
```

After writing adaptor.conf, simply run flexdirect client with a GPU application. For example, if you run `flexdirect client` with `nvidia-smi` it will list the GPUs configured.

CPU Server Command Line

```
flexdirect client nvidia-smi
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 375.26                 Driver Version: 375.26                    |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla K80           Off  | 0000:01:00.0     Off |                  N/A |
| N/A%   53C    P8    29W /  149W |     0MiB /  11479MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID  Type  Process name                               Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

Type `flexdirect help` or `flexdirect help [command]` for additional helpful commands and information.

Sample Output

```
$ flexdirect help

NAME:
    flexdirect - Run application with Bitfusion FlexDirect

USAGE:
   flexdirect <command> <options> "application"
   flexdirect <command> <options> -- [application]
   flexdirect help [command]

For more information, system requirements, and advanced usage please visit
https://www-review.vmware.com/solutions/business-critical-apps/hardwareaccelerators-
virtualization.html


COMMANDS:
        init, i           Initialize configuration.  Requires root priviledges.
        version, v        Display full FlexDirect version.
        localhealth, LH   Run health check on current node only.
        upgrade, U        Upgrade version.  Requires root priviledges.
        uninstall         Uninstall FlexDirect.  Requires root priviledges.
        dealloc           Deallocate license certificate.  Requires root priviledges.
        crashreport       Send crash report to Bitfusion.
        license           Check license status.
        list_gpus         List the available GPUs in a shared pool.
        help, h           Shows a list of commands or help for one command.

   Client Commands:
        client, c         Run application.
        health, H         Run health check on all specified servers and current node.
        request_gpus      Request GPUs from a shared pool.
        release_gpus      Release GPUs back into a shared pool. Options must match a
                          previous request_gpus command.
        run               Request GPUs from a shared pool, run a client command, then
                          release the GPUs.
        stats             Gather stats from all servers.
        smi               Display smi-like info for all servers.
        local             Run a CUDA application locally.
        net_perf          Gather network performance data from all SRS servers.

   Server Commands:
        server, s         Run server.
        resource_scheduler  Run FlexDirect resource scheduler (SRS) on GPU server


EXAMPLES:
   $ sudo flexdirect init -l <license_key>

   $ flexdirect resource_scheduler --srs_port 50001

   $ flexdirect run -n 4 -- <application>
```

Here are some `flexdirect` examples with explanatory comments.

Text

```
Initialize flexdirect license before the first run of server on a system
$ sudo flexdirect init -l <license_key>

Run a flexdirect server with default port 55001
$ flexdirect server

Run a flexdirect server with a different port
$ flexdirect server -p 55010

Run an application with a server running local with default port 55001
$ flexdirect client -l “localhost” <application>

Run an application with multiple servers, local or remote
$ flexdirect client -l “192.168.0.2:55010; 192.168.0.6:51234” <application>

Run an application with servers specified in one of the default config files (~/.
bitfusionio/adaptor.conf and /etc/bitfusionio/adaptor.conf in priority order)
$ flexdirect client <application>

Run an application with servers specified in a config file
$ flexdirect client -f <path_to_config_file> <application>

Run a server with a resource scheduler on a custom port
$ flexdirect resource_scheduler --srs_port 50001 --port 55010

Run an application with 4 shared GPUs
$ flexdirect run -n 4 <application>

Run an application with 2 shared GPUs, using half the available memory, and a custom
servers.conf
$ flexdirect run -n 2 -p 0.5 -s servers.conf <application>

Run an application with 4 shared GPUs with InfiniBand
$ flexdirect run -n 4 <application>

Run an application locally, restricted to only half the physical GPU memory
$ flexdirect local -p 0.5 <application>

Request 8 remote GPUs
$ flexdirect request_gpus -s servers.conf -f adaptor_8gpu.conf -n 8

Run an application with the generated config file
$ flexdirect client -f adaptor_8gpu.conf <application>

Release the 8 remote GPUs after the application has finished
$ flexdirect release_gpus -f adaptor_8gpu.conf

Get help on a specific command (the client command in this example)
$ flexdirect help client
```

Advanced: Cluster Communications

If you are unable to open up the default 45201-46225 port range for in-cluster communication, you can override this range by exporting these environment variables on your GPU servers before running the FlexDirect Server (also called Dispatcher):

GPU Server Command Line

```
$ export BF_SERVER_PORT_MIN=<port number>
$ export BF_SERVER_PORT_MAX=<port number>
```

Advanced: Flexible Dynamic GPU Configurations

The examples below assume that you have a four-GPU server at IP address 123.45.67.890. We will use this one GPU node for three different client applications with slightly different resource configurations, all sharing the same GPU node.

NOTE

Note how as we progress through the examples, we use different ports so that each server process is utilizing unique ports for communication.

BF_VISIBLE_DEVICES refers to the ID number of each GPU device, which starts at 0. If you have a 4 GPU instance, the IDs would be 0, 1, 2, and 3 respectively. You can see the devices and their specific IDs by running `nvidia-smi`.

Exposing One GPU out of the Four GPUs for Apps

Start the FlexDirect Server (also called Dispatcher) on the first GPU device (out of the four we are assuming for these examples) with the following command:

GPU Server Command Line

```
BF_VISIBLE_DEVICES=0 flexdirect server -p 55001
Dispatcher listening...
Listening on 0.0.0.0 : 55001
```

Now run the FlexDirect Client on your CPU node. In this example, we'll do it with application "nvidia-smi", but you could replace this with the application you would like to run using FlexDirect virtualization.

GPU Server Command Line

```
flexdirect client -l 123.45.67.89:55001 nvidia-smi
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 375.26                 Driver Version: 375.26                    |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla K80           Off  | 0000:01:00.0     Off |                  N/A |
| N/A%   53C   P8   29W /  149W |     0MiB /  11479MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
+-----------------------------------------------------------------------------+
| Processes:                                                       GPU Memory |
|  GPU       PID  Type  Process name                               Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

Exposing Two GPUs out of the Four GPUs for Apps

Start the FlexDirect Server (also called Dispatcher) on the four-GPU node with the following command:

GPU Server Command Line

```
BF_VISIBLE_DEVICES=0,1 flexdirect server -p 55002
```

Run the FlexDirect Client. In this example, we'll do it with application "nvidia-smi:, but you could replace this with the application you would like to run using FlexDirect virtualization.

GPU Server Command Line

```
flexdirect client -l 123.45.67.89:55002 nvidia-smi
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 375.26                 Driver Version: 375.26                    |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla K80           Off  | 0000:01:00.0     Off |                  N/A |
| N/A%   53C   P8   29W /  149W |     0MiB /  11479MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
+-----------------------------------------------------------------------------+
|                                                                             |
|   0  Tesla K80           Off  | 0000:01:00.0     Off |                  N/A |
| N/A%   53C   P8   29W /  149W |     0MiB /  11479MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
+-----------------------------------------------------------------------------+
|Processes:                                                       GPU Memory  |
|  GPU       PID  Type  Process name                               Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
```

Partial GPU Configurations

NVIDIA GPU SETTING TO ALLOW SHARING

When you partition a GPU, presumably you want to be able to use both partitions simultaneously. NVIDIA GPUs have a compute mode that should be set to "Default" (not "Exclusive") so that multiple applications can share access. Use the `nvidia-smi -a` command to see the current compute mode setting. And set the mode to "Default" with the command `sudo nvidia-smi -c 0`.

Server-side commands shown, see above on how to invoke the client.

1/2-GPU available on port 55001

This is done by setting environmental variable `BF_GPU_DEVICE_MEMORY_LIMIT` to half of the GPUs memory.

```
BF_VISIBLE_DEVICES=0 BF_GPU_DEVICE_MEMORY_LIMIT=6291456000 flexdirect server -p 55001
Dispatcher listening...
Listening on 0.0.0.0 : 55001
```

Two GPUs Available on Port 55001

GPU Server Command Line

```
BF_VISIBLE_DEVICES=0,1 flexdirect server -p 55001
Dispatcher listening...
Listening on 0.0.0.0 : 55001
```

2 1/2-GPUs Available on Port 55001

For two half-sized GPUs:

GPU Server Command Line

```
BF_VISIBLE_DEVICES=0,1 flexdirect server -p 55001
Dispatcher listening...
Listening on 0.0.0.0 : 55001
```

16 1/2 GPUs Assigned to Two Different Clients (Across two four-GPU nodes). Each client sees eight partial GPUs.

Use two different port numbers, one for each client. Comments are interlaced with commands:

GPU Server Command Lines

```
#server 1:
$ BF_GPU_DEVICE_MEMORY_LIMIT=6291456000 flexdirect server -p 55001 &
$ BF_GPU_DEVICE_MEMORY_LIMIT=6291456000 flexdirect server -p 55002

#server 2:
$ BF_GPU_DEVICE_MEMORY_LIMIT=6291456000 flexdirect server -p 55001 &
$ BF_GPU_DEVICE_MEMORY_LIMIT=6291456000 flexdirect server -p 55002
```

Sample Output

```
#server 1:
Dispatcher listening...
Listening on 0.0.0.0 : 55001
Dispatcher listening...
Listening on 0.0.0.0 : 55002


#server 2:
Dispatcher listening...
Listening on 0.0.0.0 : 55001
Dispatcher listening...
Listening on 0.0.0.0 : 55001
```

Configuration Files

```
#client 1 adaptor.conf:
<server 1 ip> 55001
<server 2 ip> 55001

#client 2 adaptor.conf:
<server 1 ip> 55002
<server 2 ip> 55002
```

16 1/2 GPUs (across two four-GPU nodes) available. Two different clients each allocate one partial GPU.

GPU Server Command Lines

```
#Server 1:
BF_VISIBLE_DEVICES=0 BF_GPU_DEVICE_MEMORY_LIMIT=6291456000 flexdirect server -p 55001 &
BF_VISIBLE_DEVICES=1 BF_GPU_DEVICE_MEMORY_LIMIT=6291456000 flexdirect server -p 55002 &
BF_VISIBLE_DEVICES=2 BF_GPU_DEVICE_MEMORY_LIMIT=6291456000 flexdirect server -p 55003 &
BF_VISIBLE_DEVICES=3 BF_GPU_DEVICE_MEMORY_LIMIT=6291456000 flexdirect server -p 55004 &
BF_VISIBLE_DEVICES=4 BF_GPU_DEVICE_MEMORY_LIMIT=6291456000 flexdirect server -p 55005 &
BF_VISIBLE_DEVICES=5 BF_GPU_DEVICE_MEMORY_LIMIT=6291456000 flexdirect server -p 55006 &
BF_VISIBLE_DEVICES=6 BF_GPU_DEVICE_MEMORY_LIMIT=6291456000 flexdirect server -p 55007 &
BF_VISIBLE_DEVICES=7 BF_GPU_DEVICE_MEMORY_LIMIT=6291456000 flexdirect server -p 55008 &

#Server 2:
BF_VISIBLE_DEVICES=0 BF_GPU_DEVICE_MEMORY_LIMIT=6291456000 flexdirect server -p 55001 &
BF_VISIBLE_DEVICES=1 BF_GPU_DEVICE_MEMORY_LIMIT=6291456000 flexdirect server -p 55002 &
BF_VISIBLE_DEVICES=2 BF_GPU_DEVICE_MEMORY_LIMIT=6291456000 flexdirect server -p 55003 &
BF_VISIBLE_DEVICES=3 BF_GPU_DEVICE_MEMORY_LIMIT=6291456000 flexdirect server -p 55004 &
BF_VISIBLE_DEVICES=4 BF_GPU_DEVICE_MEMORY_LIMIT=6291456000 flexdirect server -p 55005 &
BF_VISIBLE_DEVICES=5 BF_GPU_DEVICE_MEMORY_LIMIT=6291456000 flexdirect server -p 55006 &
BF_VISIBLE_DEVICES=6 BF_GPU_DEVICE_MEMORY_LIMIT=6291456000 flexdirect server -p 55007 &
BF_VISIBLE_DEVICES=7 BF_GPU_DEVICE_MEMORY_LIMIT=6291456000 flexdirect server -p 55008 &
```

Sample Output

```
#Server 1:
Dispatcher listening...
Listening on 0.0.0.0 : 55001
Dispatcher listening...
Listening on 0.0.0.0 : 55002
Dispatcher listening...
Listening on 0.0.0.0 : 55003
Dispatcher listening...
Listening on 0.0.0.0 : 55004
Dispatcher listening...
Listening on 0.0.0.0 : 55005
Dispatcher listening...
Listening on 0.0.0.0 : 55006
Dispatcher listening...
Listening on 0.0.0.0 : 55007
Dispatcher listening...
Listening on 0.0.0.0 : 55008

#Server 2:
Dispatcher listening...
Listening on 0.0.0.0 : 55001
Dispatcher listening...
Listening on 0.0.0.0 : 55002
Dispatcher listening...
Listening on 0.0.0.0 : 55003
Dispatcher listening...
Listening on 0.0.0.0 : 55004
Dispatcher listening...
Listening on 0.0.0.0 : 55005
Dispatcher listening...
Listening on 0.0.0.0 : 55006
Dispatcher listening...
Listening on 0.0.0.0 : 55007
Dispatcher listening...
Listening on 0.0.0.0 : 55008
```

Sample Output

```
#Client 1 adaptor.conf (first partial GPU of server 1):
<server 1 ip> 55001

#Client 2 adaptor.conf (second partial GPU of server 1):
<server 1 ip> 55002
```