AI400X2 Turbo 24x13.9TiB nvme 8xHDR200 - 32 clients

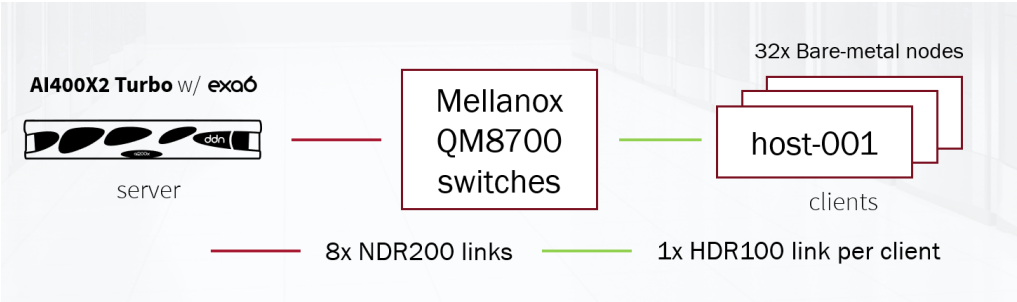
DDN

August 10, 2024

Contents

1 Hardware

The hardware is composed of 32 clients and one server connected to the same switch.



1.1 Client

The clients are **bare-metal** nodes (compute) with 1 CPU socket and 92GiB of RAM based on Intel S2600BPB motherboard. All clients are configured the same (hw & sw). The clients communicate with the server through IB using one dedicated link.

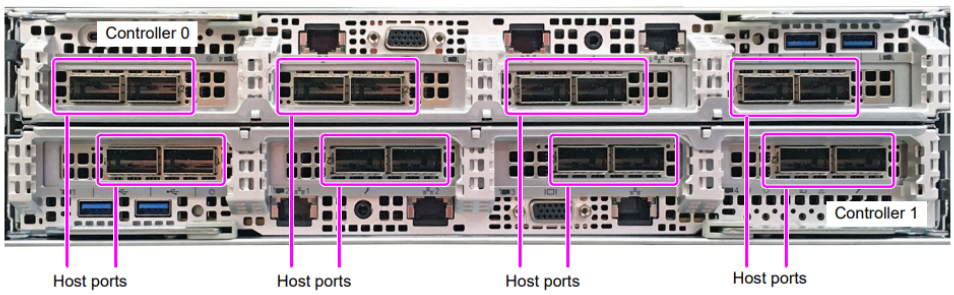
- 32x HDR100 links total (one per client)

1.2 Server

The server is **a commercial DDN AI400X2 Turbo** appliance (storage) providing a single EXAScaler filesystem to the clients (one filesystem total). The configuration is the default one with 24x Phison NVME drives with 14TiB per drive. The server communicate with the clients through IB.

- 8x dual ports NDR200 adapters for a total of 8x NDR200 links (1x link per HCA)

Figure 27. Host Ports on AI400X2T



1.3 Switch

Mellanox QM8700 switches. The server and clients are connected to it.

2 Software

2.1 Client

The clients OS is Rocky Linux 8.8 (Green Obsidian). The Mellanox OFED version is `MLNX_OFED_LINUX-23.07-0.5.0.0`

2.2 Server

The AI400X2 Turbo runs EXAScaler SFA Rocky 6.3.0-3 on top of the SFAOS stack (SFAOS version 12.4.0). The EXAScaler driver used on the server is `2.14.0_ddn145_8_g2dc656d`. This is the GA version of the product with no modification applied. The OFED version is OFED-internal-23.10-2.1.3.

2.3 Switch

NA

3 Settings

3.1 Client

The standard EXAScaler 6.3.0-3 driver (GA) is installed on the client using the default installation procedure. The driver is used to access the remote EXAScaler filesystem.

3.1.1 EXAScaler driver configuration

The EXAScaler configuration file has been generated automatically following the default installation procedure. This configuration file is necessary to mount the filesystem. Content of /etc/modprobe.d/lustre.conf is:

```
# This file has been generated by exa-client-deploy
#
# Do not edit unless exa-client-deploy service is stopped & disabled
# e.g: 'systemctl status exa-client-deploy'
#

options lnet networks="o2ib(ib0)"
options lnet lnet_transaction_timeout=100
options lnet lnet_retry_count=2
options ko2iblnd peer_credits=32
options ko2iblnd peer_credits_hiw=16
options ko2iblnd concurrent_sends=64
options ksocklnd conns_per_peer=0
```

As optimizations and customizations to the storage system are allowed in CLOSED division (section 6. of the MLPerfV.1.0 submission guidelines), the following tunings were applied on each client:

```
# UNet3D clients tunings
  lctl set_param osc.*.max_pages_per_rpc=1M \
       osc.*.max_rpcs_in_flight=32 \
       osc.*.checksums=0
# Cosmoflow & Resnet50 client tunings
  lctl set_param osc.*.max_pages_per_rpc=1M \
       osc.*.max_rpcs_in_flight=8 \
       osc.*.checksums=0
```

3.2 Server

No tuning were applied on the DDN AI400X2 Turbo after the initial installation. The inital installation was automated using the latest version of our manufacturing script. The filesystem was mounted on the client after the end of this script.

3.3 Switch

No tuning applied

4 Misc

4.1 Drop cache

The cache of both the client and server were dropped between iterations using the following function. Since 'clush -S' was used, both commands did work.

```
clear_caches() {
    # Clear cache on client
    clush -abS "sync; echo 3 > /proc/sys/vm/drop_caches"  || return 1
    # Clear cache on EXAScaler VMs
    ssh root@${exafsip}  "clush -abS 'sync; echo 3 > /proc/sys/vm/drop_caches'" \
        || return 1
}
```

4.2 Script

The run script used the following bash options to ensure that all commands worked.

```
#!/bin/bash

set -u
set -e
set -x
set -o pipefail
```