Hewlett Packard
Enterprise

Slingshot and HPC Storage – Choosing the right Lustre Network Driver (LND)

John Fragalla, Distinguished Technologist

Cray User Group 2023
May 11, 2023

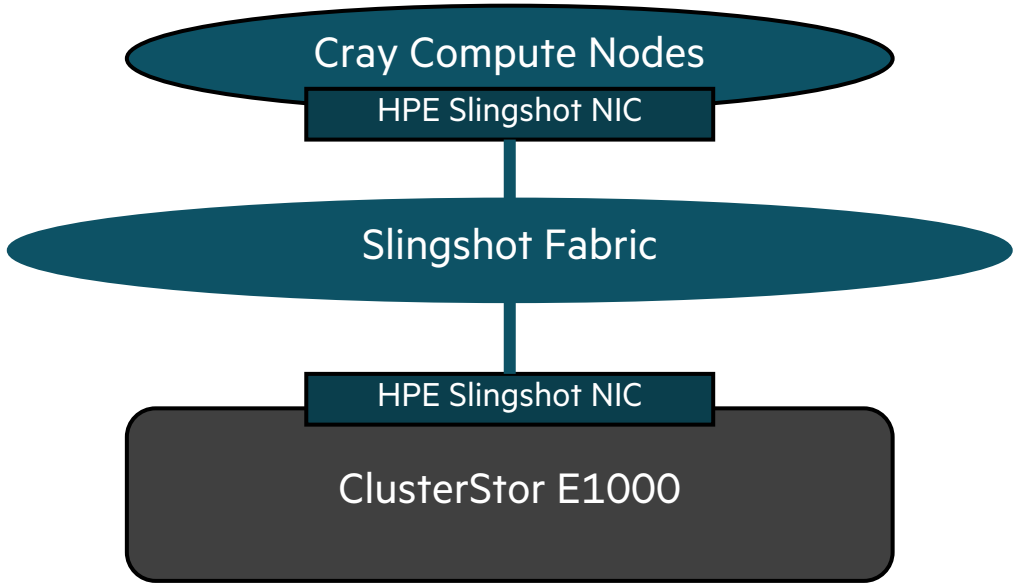
Agenda

- Recommended Lustre Network Driver (LND) for different HPE Slingshot Deployments
  - kkfilnd, ksocklnd, ko2iblnd via SW RoCE
- Performance comparing the three different LNDs on HPE Slingshot NIC deployments
  - IOR and MDTEST
- Lustre Network (LNet) Routing performance to a HPC Slingshot native ClusterStor System
  - Mount kkfilnd ClusterStor System from ko2iblnd Compute nodes through a set of LNet Routers
- ClusterStor Dual-LND on Ethernet
- Acknowledgements
- Backup Slides

Slingshot-11 Native Connectivity Option

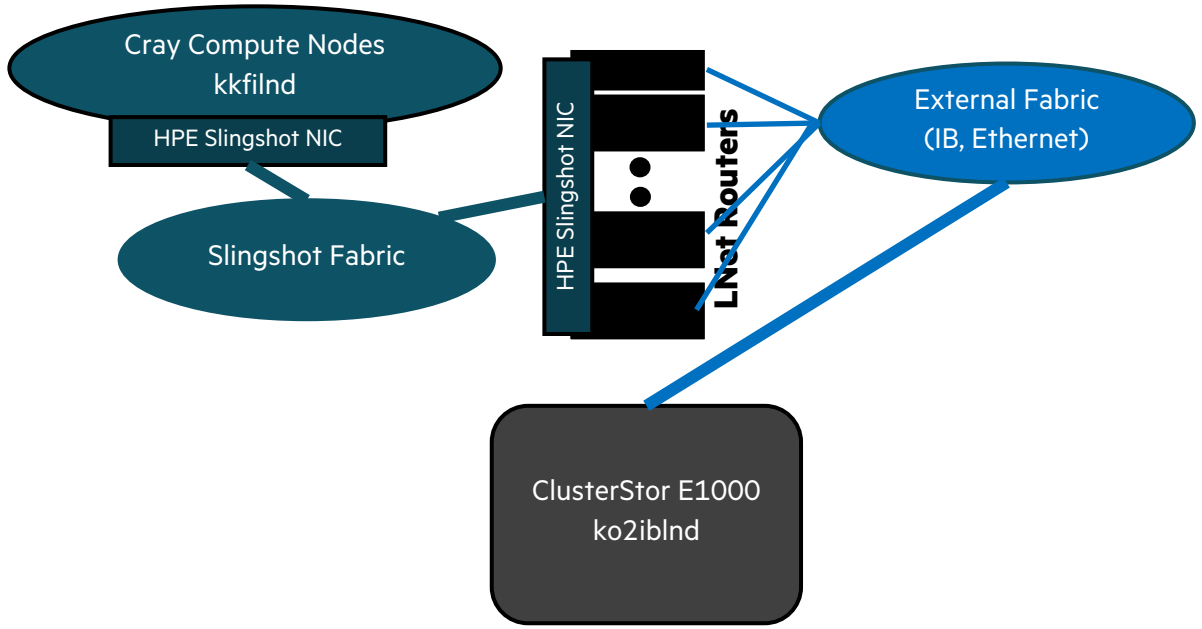
- E1000 supports native Slingshot-11 connectivity
- Recommend E1000 Lustre Network Driver (LND) to mount Lustre on Slingshot-11 attached clients is kkfilnd end-to-end.
  - kkfilnd provides the best I/O bandwidth on Slingshot-11 fabric, and is the primary LND recommended

- **NOTE**: Slingshot-11 does support ksocklnd, but not recommended to be used as an exclusive LND for I/O due to performance limitations.

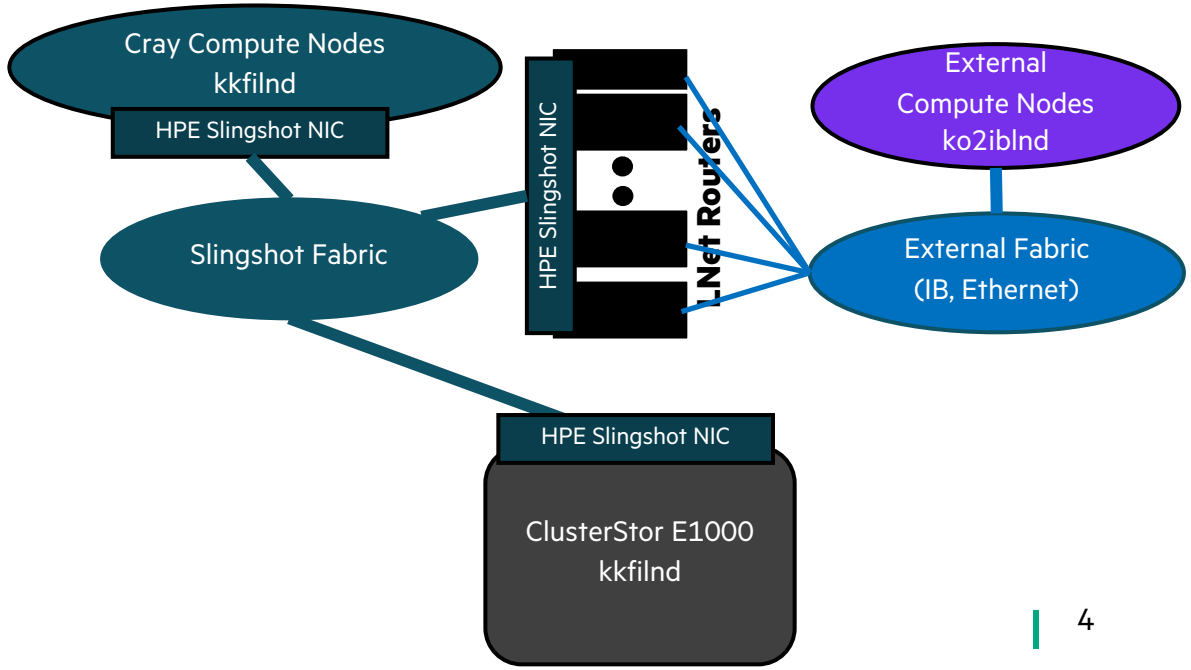


Lustre Network (LNet) Routing to and from Slingshot-11

- Compute configured with Slingshot-11 and the E1000 is **not** on the Slingshot fabric
  - **Use Case**: Share an already existing ClusterStor system to Slingshot-11 compute nodes through LNet Routers
  - Routing from IB/HW RoCE Based E1000 to Slingshot-11
    - ko2iblnd (storage) <-> kkfilnd (compute) Routing



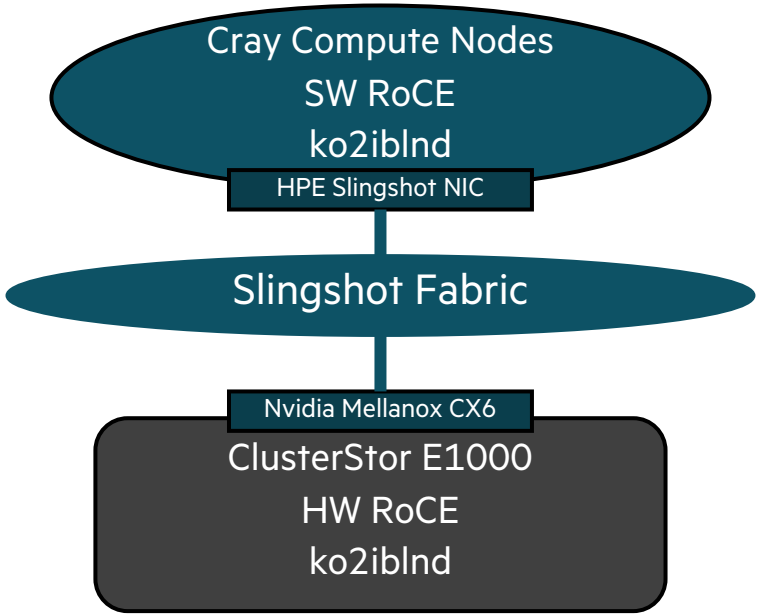
- E1000 configured with Slingshot-11 and access storage to **non-slingshot compute nodes**
  - **Use Case**: Share E1000 to external clients not on the Slingshot fabric through LNet Routers
  - Routing from E1000 with kkfilnd, and external clients with IB/HW RoCE Fabric
    - kkfilnd (storage) <-> ko2iblnd (compute) Routing



Software RoCE (ko2iblnd) on Compute Nodes w/ HPE Slingshot NIC

- SW RoCE (ko2iblnd) is a Compute node option with HPE Slingshot NIC starting in Cray OS (COS) 2.4+
- Allows direct access to ClusterStor E1000 I/O storage nodes using MLX CX6 HCAs running HW RoCE and LND is ko2iblnd.

- **Use cases** to enable SW RoCE on Compute Nodes w/ HPE Slingshot NIC
  - Migrating Slingshot-10 to Slingshot-11 on the Compute Nodes, while the E1000 stays with MLX CX6 HCAs
  - Requirement for routable RDMA LND through Slingshot Link Aggregation Group (LAG) and ClusterStor is being accessed by more than one Slingshot Fabric **without** LNet routers

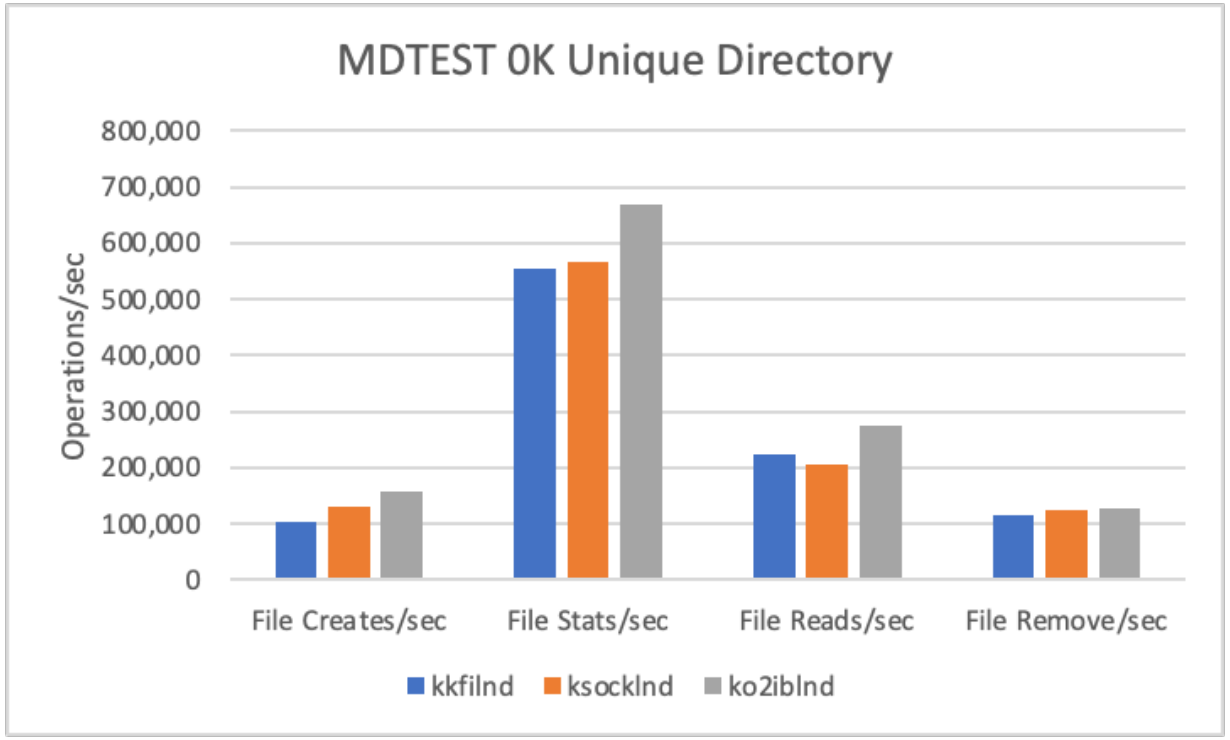


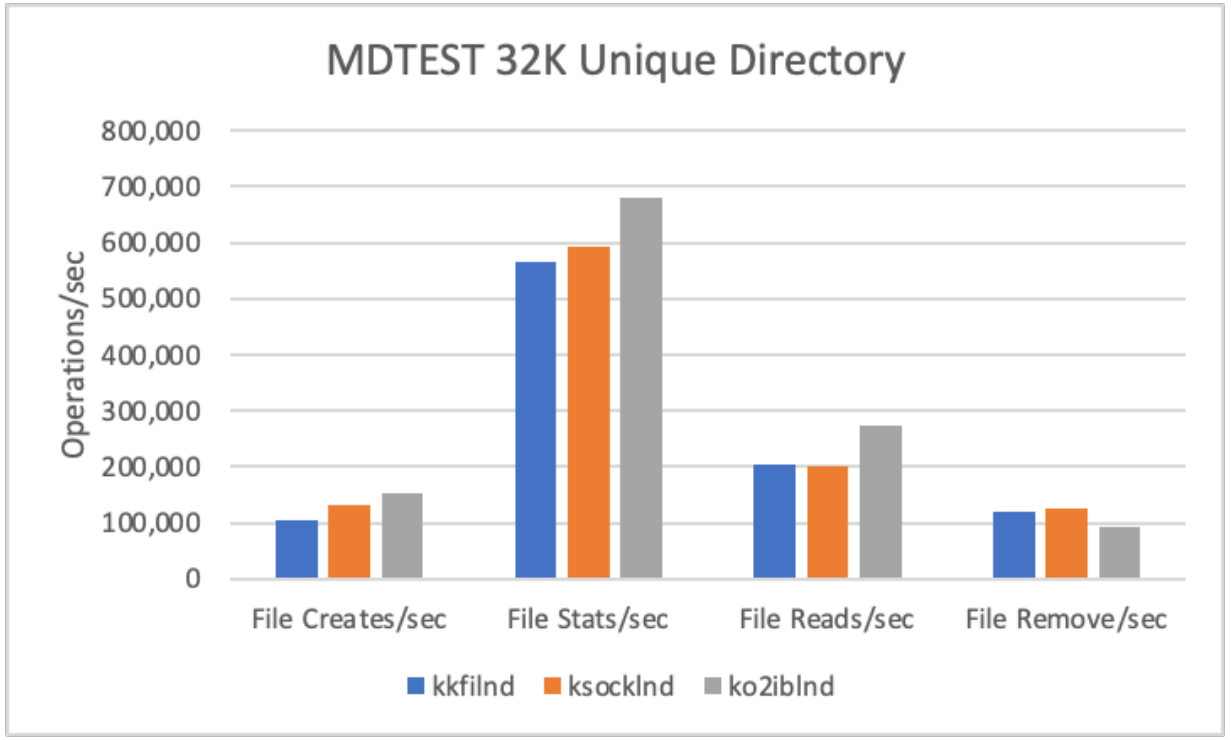
Engineering Benchmark Performance Results

- Performance results are comparing three Lustre LNDs
  - kkfilnd
    - HPE Slingshot NIC on Compute nodes
    - HPE Slingshot NIC on ClusterStor E1000 nodes
  - ksocklnd
    - HPE Slingshot NIC on Compute nodes
    - HPE Slingshot NIC on ClusterStor E1000 nodes
  - ko2iblnd
    - HPE Slingshot NIC on Compute Nodes using SW RoCE
    - Nvidia Mellanox CX6 NIC on ClusterStor E1000 using HW RoCE

Slingshot-11 Engineering I/O Performance MDTEST Results

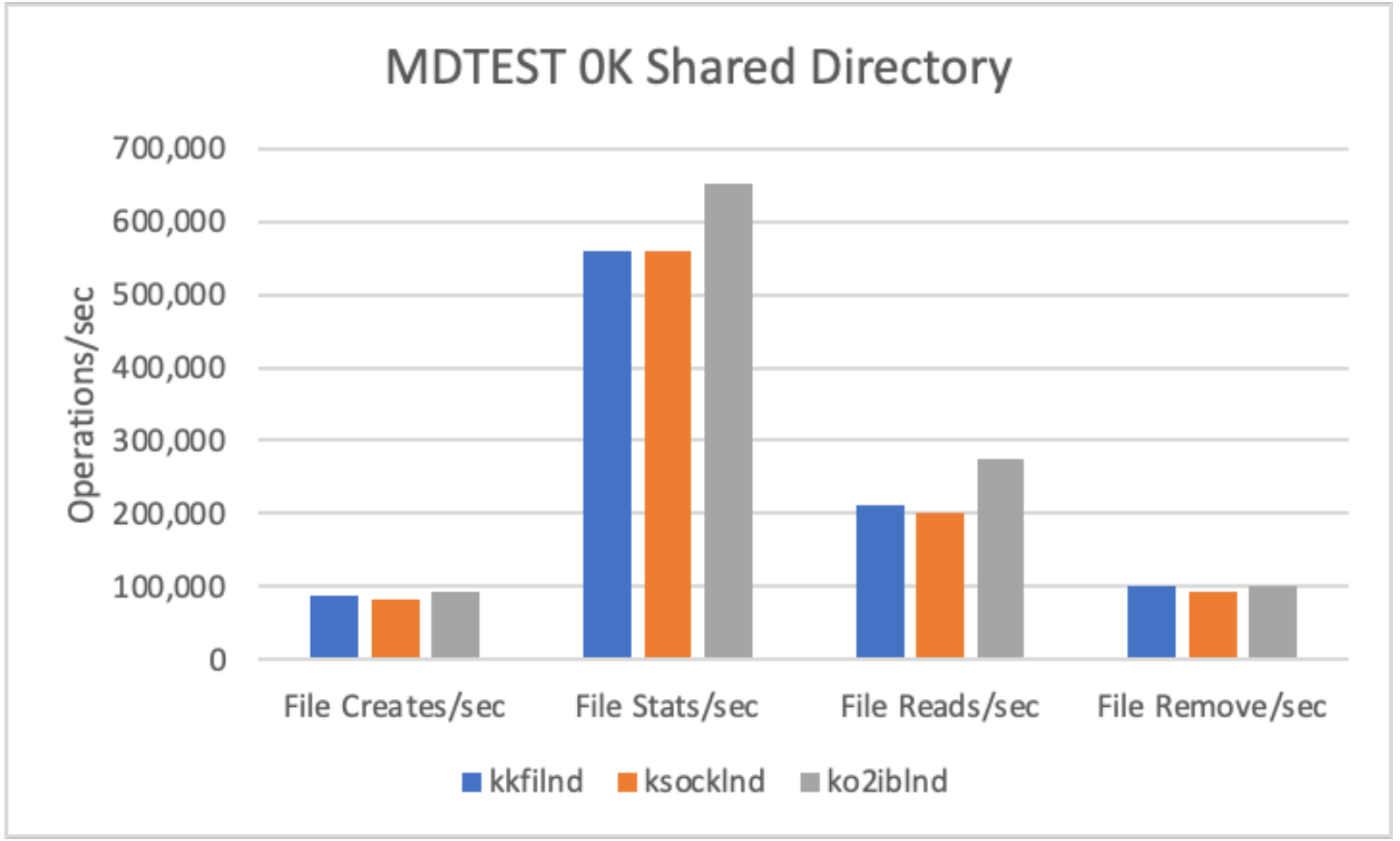
kkfilnd vs ksocklnd vs ko2iblnd on HPE Slingshot NIC (Slingshot 2.0.2 RC3), LDISKFS





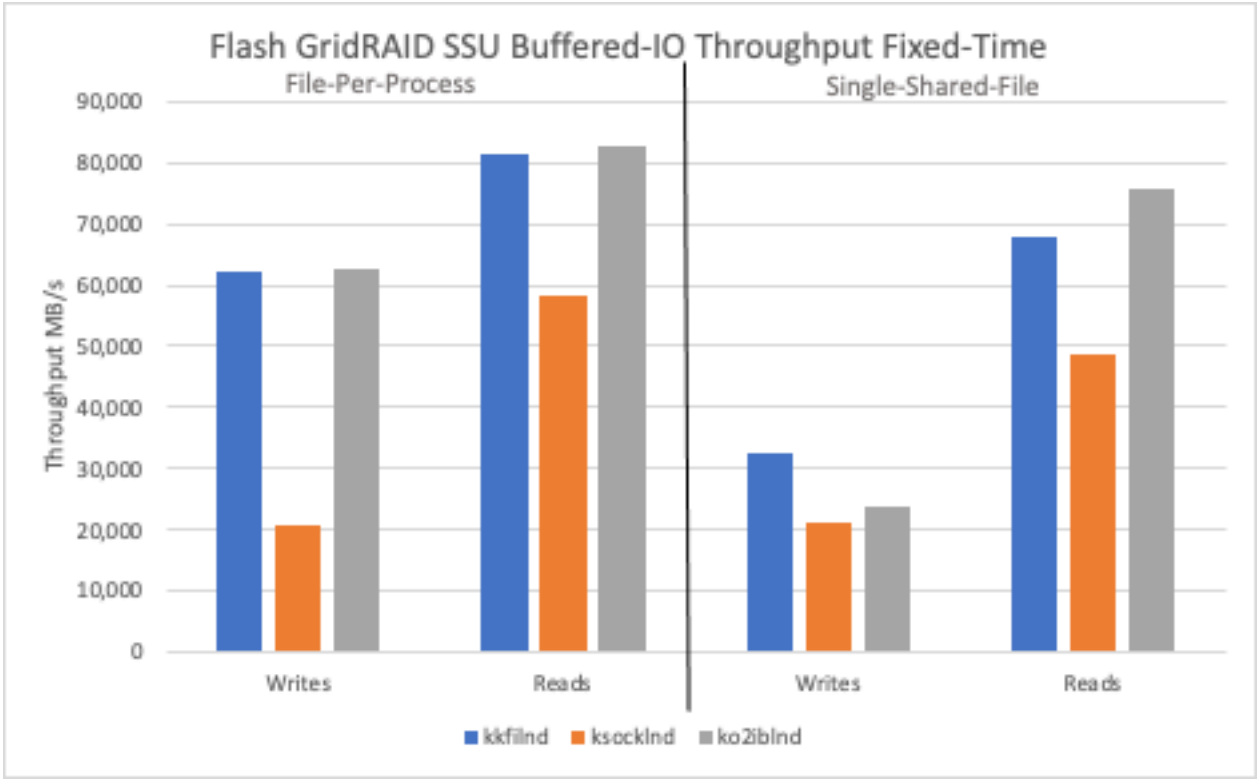
Slingshot-11 Engineering I/O Performance MDTEST Results

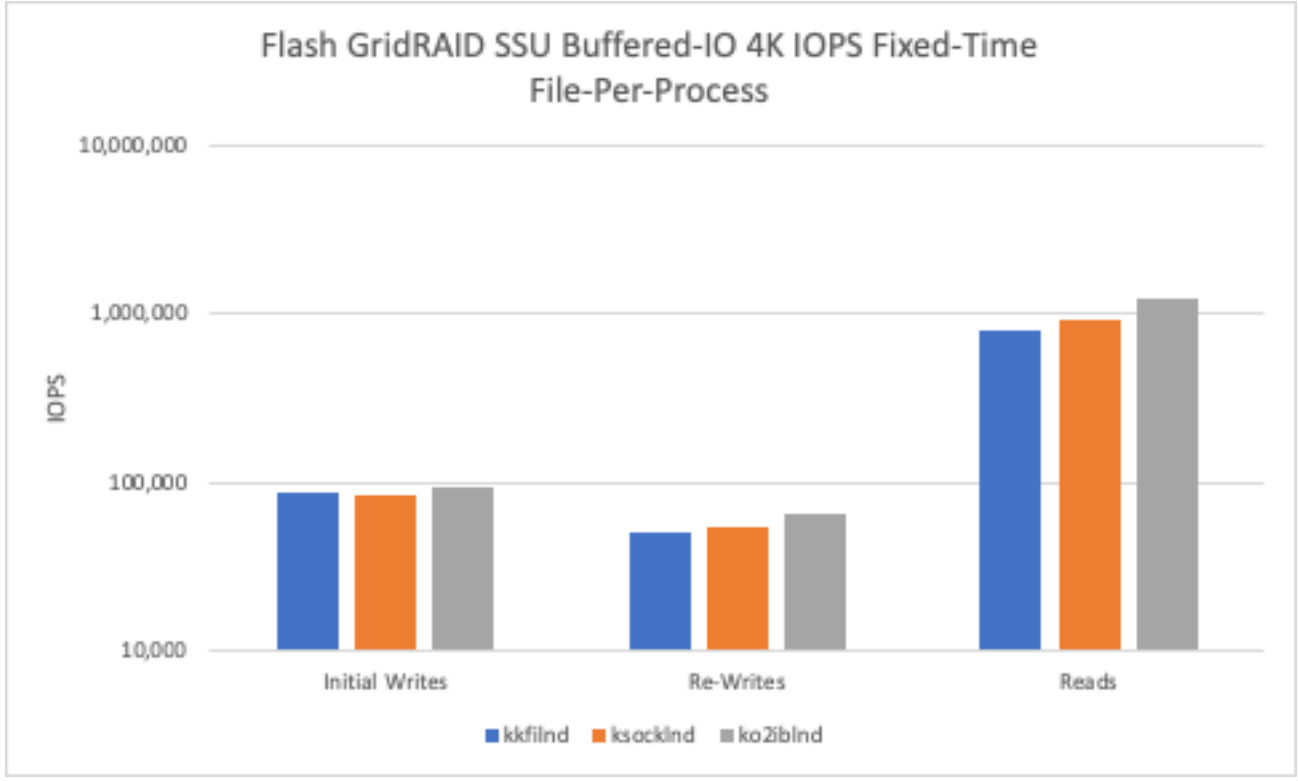
kkfilnd vs ksocklnd vs ko2iblnd on HPE Slingshot NIC (Slingshot 2.0.2 RC3), LDISKFS



Slingshot-11 Engineering I/O Performance Flash IOR Results

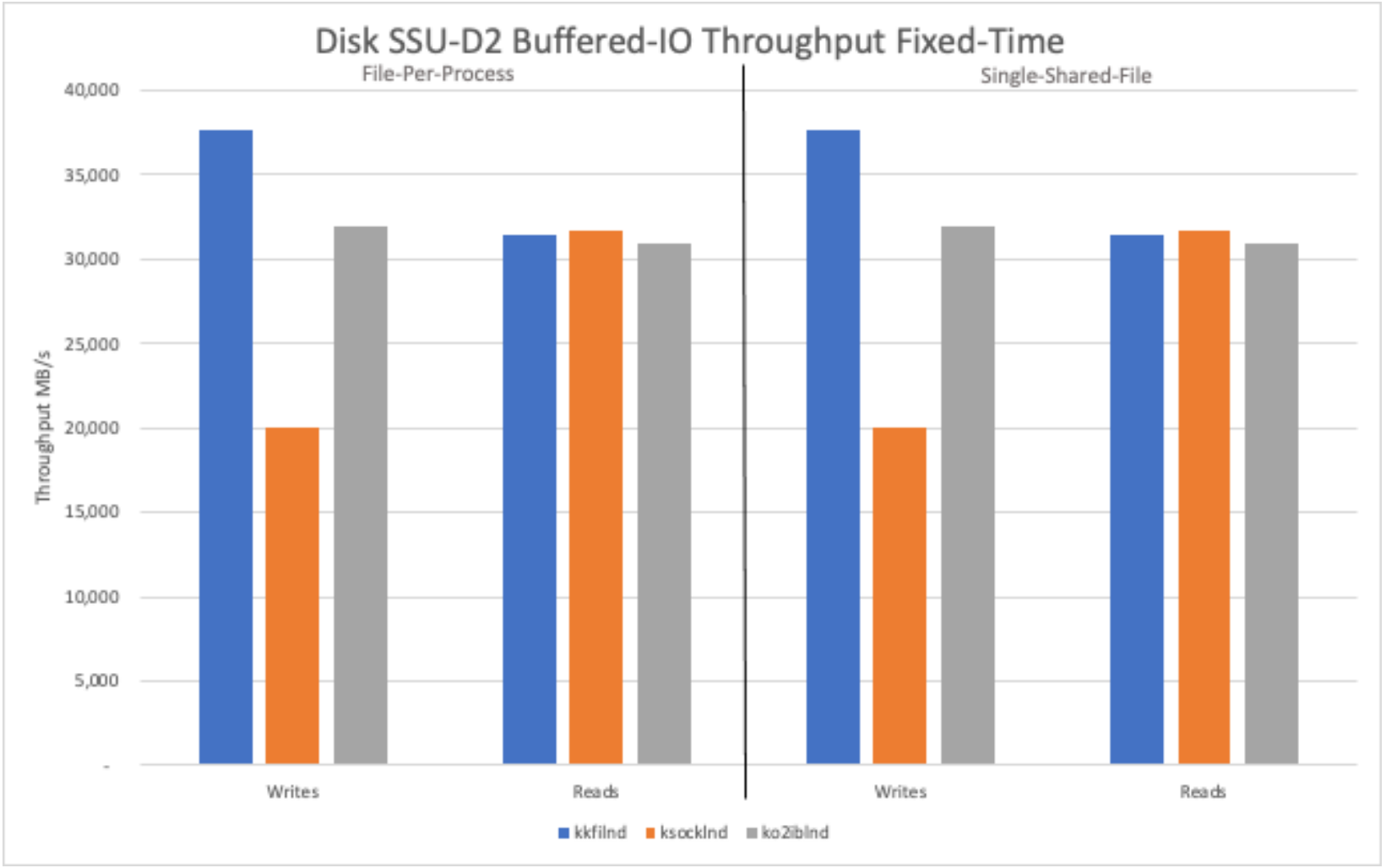
kkfilnd vs ksocklnd vs ko2iblnd on HPE Slingshot NIC (Slingshot 2.0.2 RC3), LDISKFS





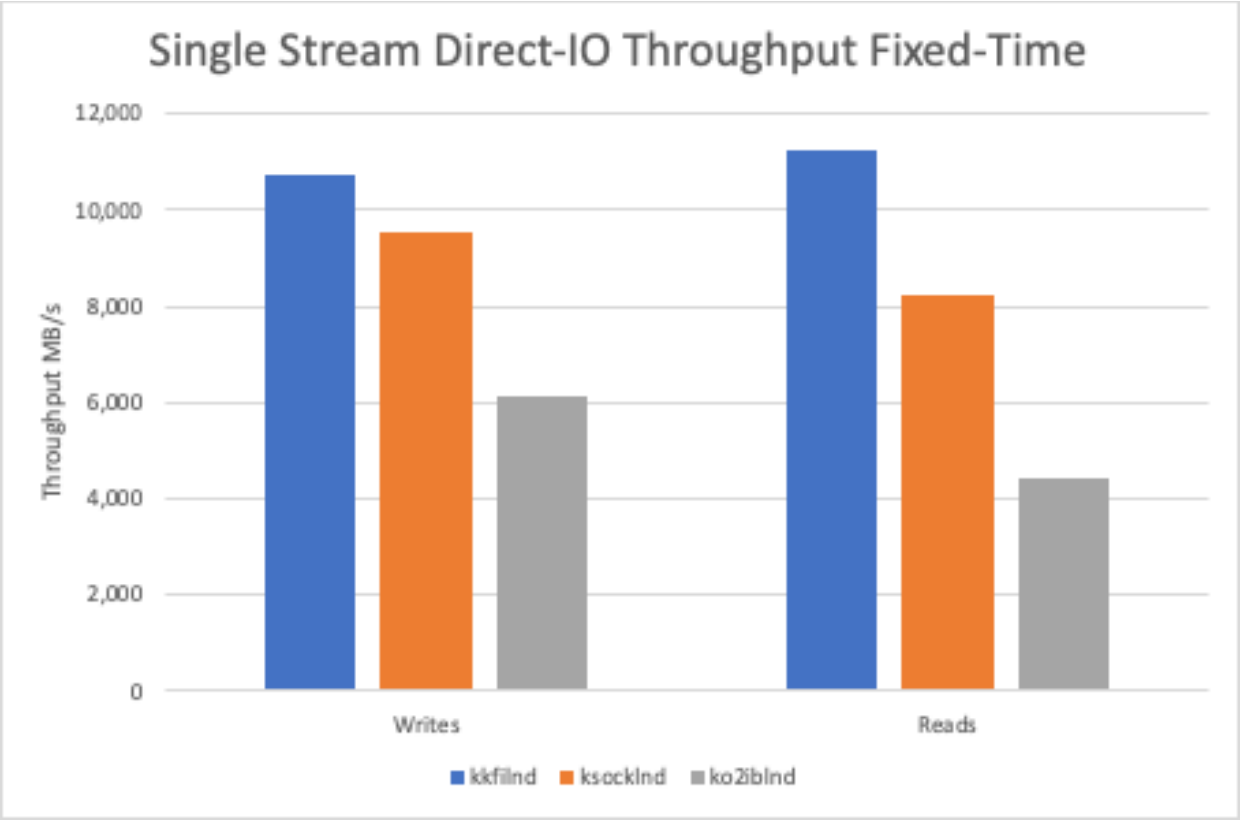
Slingshot-11 Engineering I/O Performance Disk IOR Results

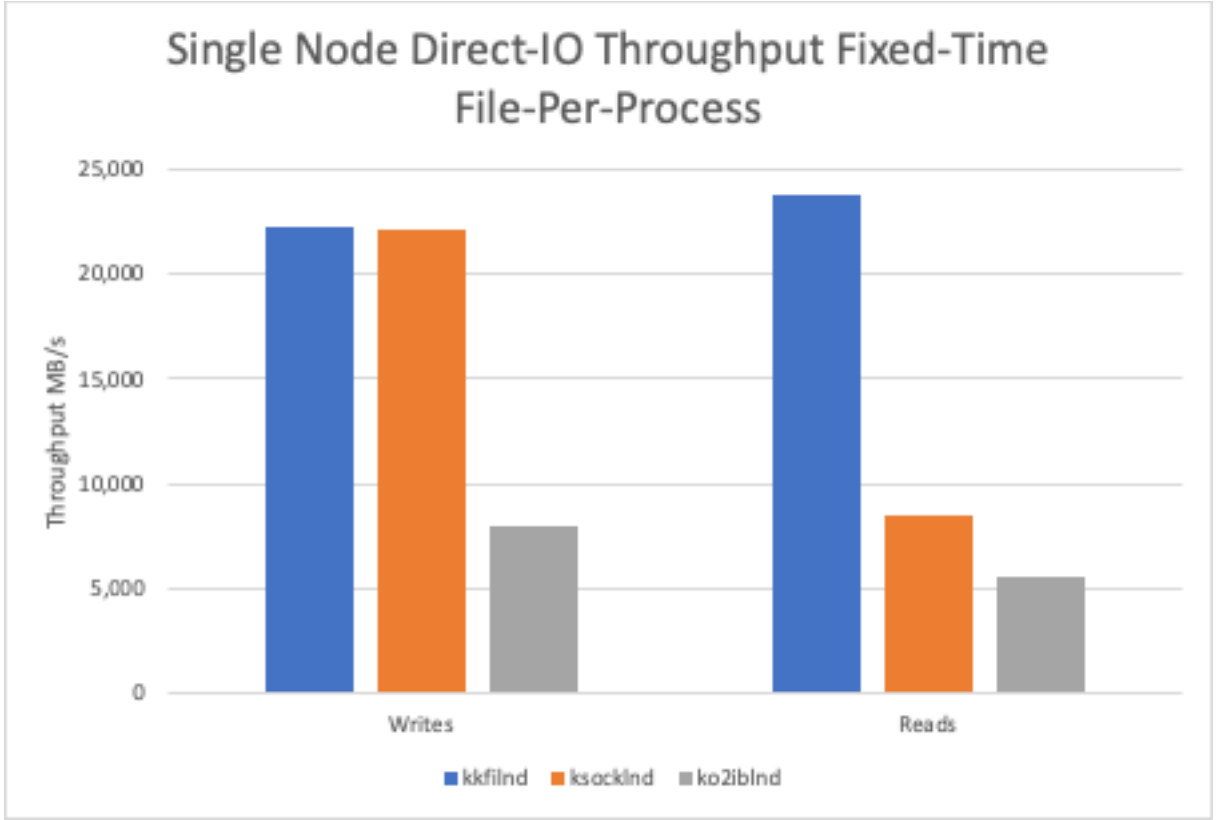
kkfilnd vs ksocklnd vs ko2iblnd on HPE Slingshot NIC (Slingshot 2.0.2 RC3), LDISKFS



Slingshot-11 Engineering I/O Performance Single Stream/Node IOR Results

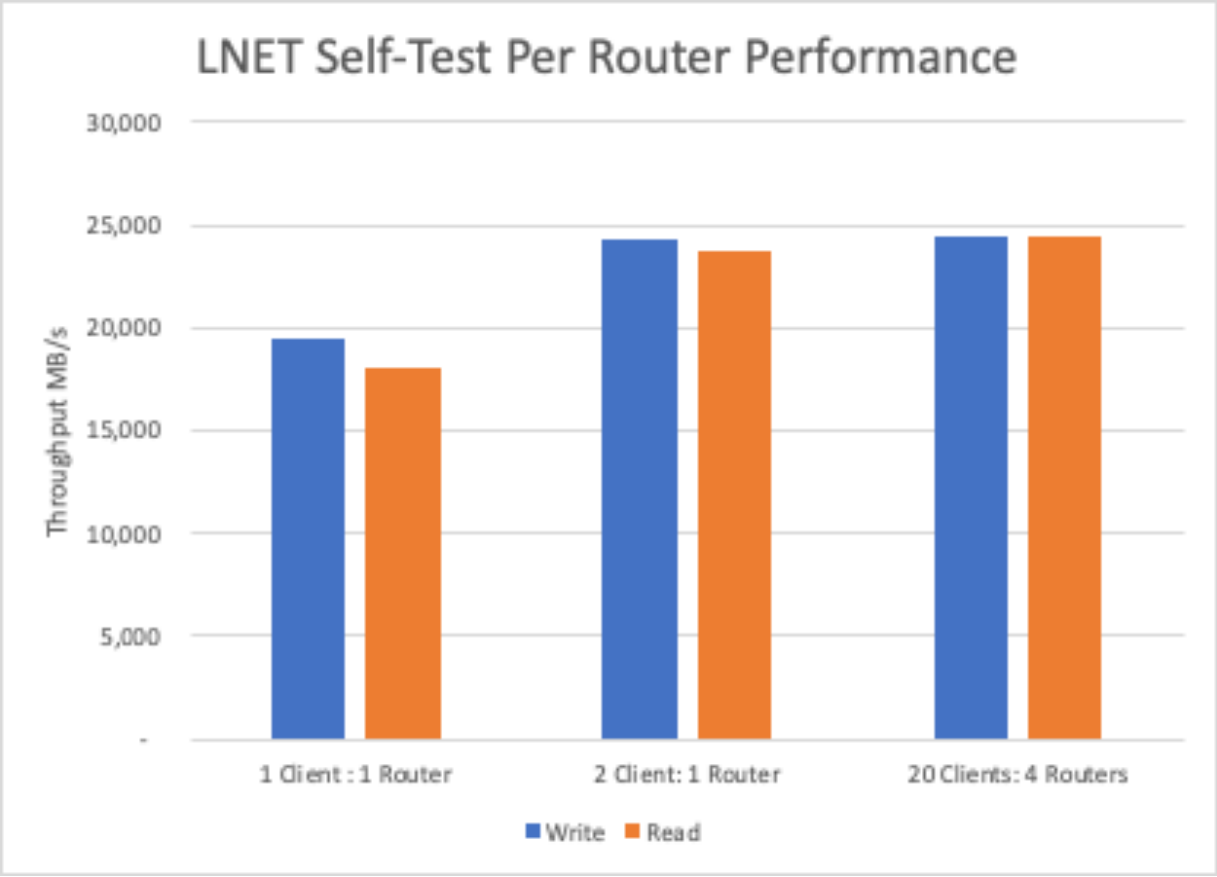
kkfilnd vs ksocklnd vs ko2iblnd on HPE Slingshot NIC (Slingshot 2.0.2 RC3), LDISKFS

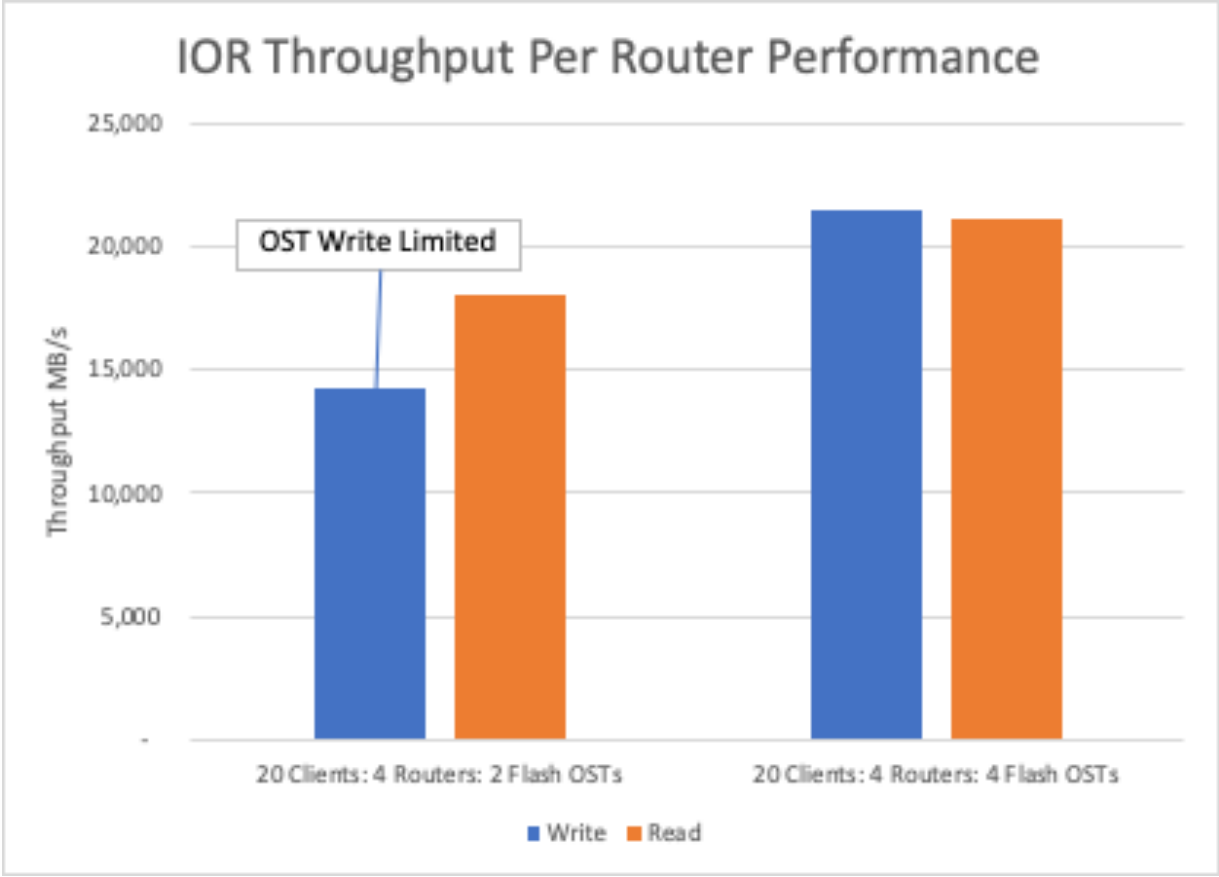




Lustre Network (LNet) Routing Performance

ko2iblnd Compute nodes <-> LNet Routers <-> kkfilnd Clusterstor Storage (single-injection)





ClusterStor Dual-LND Option for Ethernet

- Starting in ClusterStor Release 6.3 and later, Dual-LND feature is available to enable. Feature allows ksocklnd to be enabled at the same time as a RDMA NID, such as kkfilnd.

- For example, on a single OSS node with two CXI Links, by default the NIDs will be

```
74@kfi
73@kfi
```

- After the secondary NID is enabled, a single OSS node will have the following:

```
74@kfi
73@kfi
10.230.60.2@tcp
10.230.60.3@tcp
```

- This feature is useful if there exists two different set of clients, one running RDMA traffic and the other using ksocklnd (TCP/IP) and want to mount a single ClusterStor system without the use of LNet Routers.

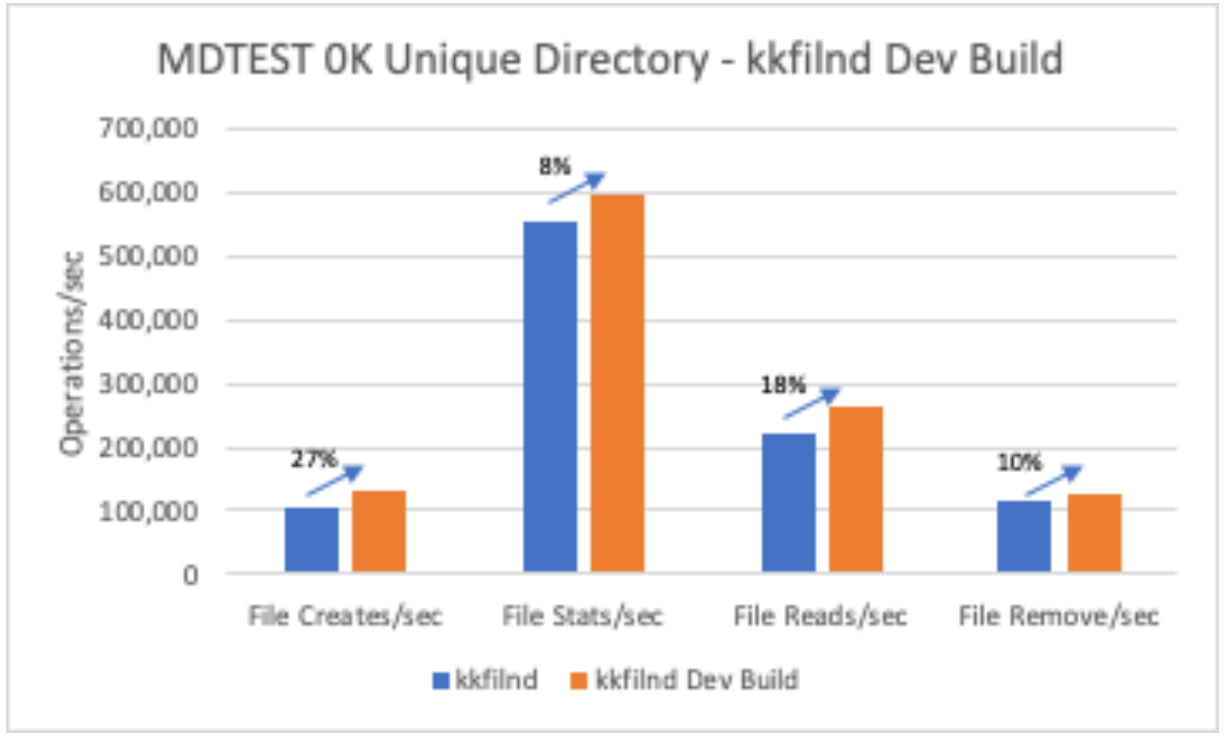
- User Defined Selection Policy (UDSP) is recommended to prioritize the RDMA NID in a Dual-LND configuration, for example

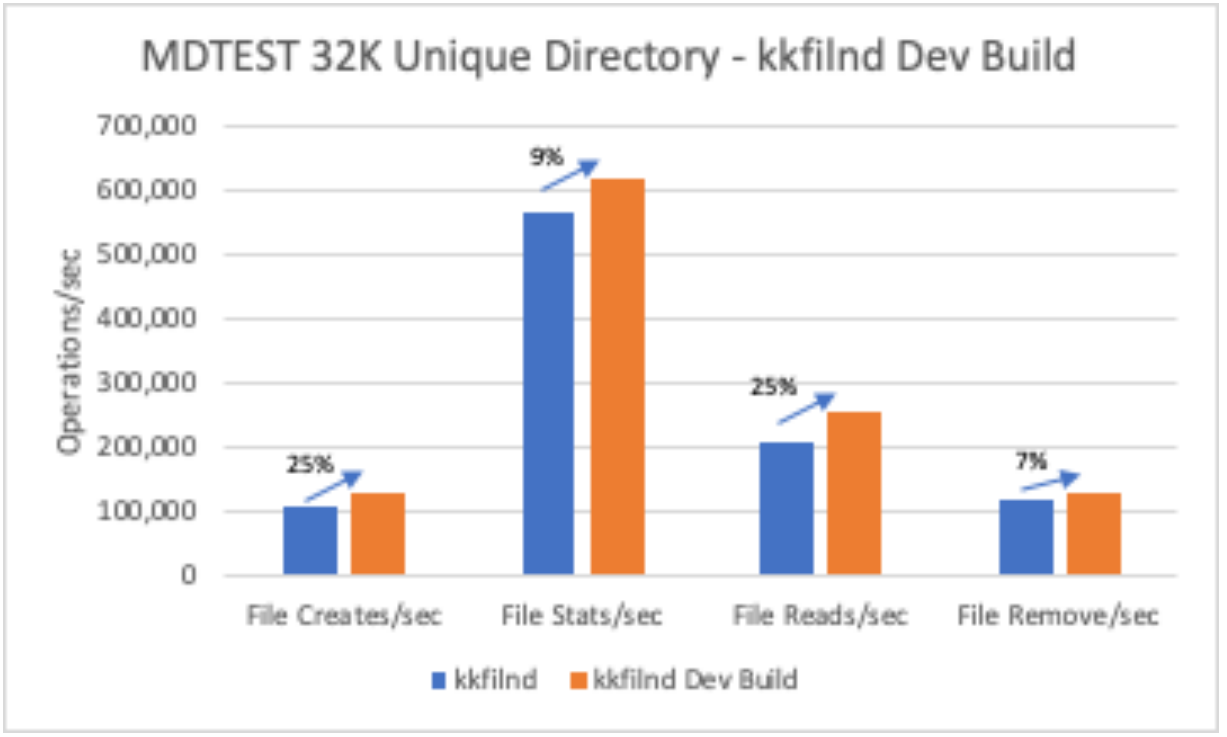
```
lnetctl udsp add –src kfi –priority 0
```

Late Breaking Results – MDTEST using kkfilnd Development Build

Development investigation to improve Metadata performance

- Development build reducing kkfilnd thread count and priority.





Summary

- HPE team continues to make performance improvements with kkfilnd, focusing on MDT and Data IOPS
- ksocklnd is a supplemental LND to expand connectivity options with ClusterStor when dual-LND feature is enabled
- ko2iblnd with Soft RoCE on HPE Slingshot NIC is an option to be used when routable RDMA is required and the use of LNet Routers are not possible.
- kkfilnd LNet single-injection router performance can saturate a 200Gbps Link

Acknowledgements

- HPE Virtual Performance Team
- Fort Collins Data Center Team
- Lustre Development Team
- Slingshot/Cassini Development Team
- Chris Horn, Ian Ziemba, Ron Gredvig, Bill Loewe, Michael Moore, Wolfgang Szoecs, Petros Koutoupis, Cory Spitz, Jenny Hack, Prudvi Danda, Christer Lundin, Bob Pearson, Kent Devenport

Thank You

john.fragalla@hpe.com

Backup Slides

Define Lustre Network Terms

- NID – Lustre Network Identification
- LND - Lustre Network Driver that defines the specific NID configured on Lustre nodes
- kkfilnd – klibfabric RDMA Lustre driver used for Slingshot-11 end-to-end (HPE Slingshot NIC)
- ksocklnd – TCP/IP Lustre driver
- ko2iblnd – RDMA Lustre driver used for SW RoCE on HPE Slingshot NIC (also used for IB or HW RoCE)

Benchmark Setup

- kkfilnd/ksocklnd Benchmark Setup
  - ClusterStor E1000 with Software release 6.4-010.75 (1 MDU, 1 Hybrid SSU, 1 Flash SSU), HPE Slingshot NICs
  - 24 Compute Nodes, RHEL based 8.4, Lustre 2.15 client (2.15.0.7_rc2_cray_3_g412d1c5), HPE Slingshot NIC
  - SHS 2.0.2 RC3 and Slingshot Switch 2.0.2 RC2
  - IOR/MDTEST 3.3.0
  - Compute
    - `options kkfilnd peer_credits=16 traffic_class=bulk_data immediate_rx_buf_count=8`
    - `options lnet lnet_retry_count=0 lnet_transaction_timeout=126`
    - `options cxi-core ioi_enable=0`
  - Storage:
    - `options kkfilnd peer_credits=16 traffic_class=best_effort immediate_rx_buf_count=32`
    - `options lnet lnet_retry_count=0 lnet_transaction_timeout=126`
    - `options cxi-core ioi_enable=0`

- SW RoCE/ko2iblnd Benchmark Setup
  - ClusterStor E1000 with Software release 6.2-010.65 (1 MDU, 1 HDD D2 SSU, 1 Flash SSU), Nvidia Mellanox CX6 NICs
  - 32 Compute Nodes, COS 2.4-109
  - SHS 2.0.2 RC3 and Slingshot Switch 2.0.2 RC2
  - IOR/MDTEST 3.3.0
  - ko2iblnd tuning
    - `Client: options ko2iblnd conns_per_peer=2 peer_credits=42 concurrent_sends=84 ntx=2048 credits=1024`
    - `Server: options ko2iblnd conns_per_peer=2 ntx=2048 peer_credits=42 concurrent_sends=84 credits=1024 map_on_demand=1`

ClusterStor Dual-LND Option for Ethernet - details

- Starting in ClusterStor Release 6.3 and later, Dual-LND feature is available to enable. Feature allows ksocklnd to be enabled at the same time as a RDMA NID, such as kkfilnd.
- For example, on a single OSS node with two CXI Links, by default the NIDs will be

```
74@kfi
73@kfi
```

- One can enable ksocklnd running the following command from the primary ClusterStor SMU node

```
cscli lustre lnet set_nid -s -i 0
cscli lustre_network apply -y
```

  - Where `-i 0` is the NID for ksocklnd, which can be changed to any numerical value
- After secondary NID is enabled, the single OSS node will have the following:

```
74@kfi
73@kfi
10.230.60.2@tcp
10.230.60.3@tcp
```

- This feature is useful if there exists two different set of clients, one running RDMA traffic and the other using ksocklnd (TCP/IP) and want to mount a single ClusterStor system without the use of LNet Routers.
- If a Compute node has both RDMA and ksocklnd NID enabled, it is recommended to setup User Defined Selection Policy (UDSP) on the Lustre clients, to prioritize the RDMA NID for performance reasons. On each compute node, run the following

```
lnetctl udsp add -src kfi -priority 0
```