

Shaheen III: Storage

Bilel Hadri, KAUST Supercomputing Lab

February 4th, 2025

Shaheen Storage

"A supercomputer is a device for converting a CPU-bound problem into an I/O bound problem."
[Ken Batcher]

- HPC systems consist of three main components:
 - Compute nodes
 - High-speed interconnect
 - I/O infrastructure

- Most optimization work on HPC applications is carried out on
 - **Single node performance across many cores**
 - Network performance (communication)
 - **I/O only when it becomes a real problem (or when the sys-admin contact us...)**



Shaheen Storage

- I/O is commonly used by scientific applications to achieve goals like:
 - storing numerical output for later analysis
 - loading initial conditions or datasets for processing
 - checkpointing to files that save the state of an application in case of system failure
- On Supercomputer, you will need parallel file system

Shaheen Storage Why do we need parallel I/O?

- Imagine a 24 hours simulation on 16 cores.
 - 1% of run time is serial I/O.
 - You get the compute part of your code to scale to 6 nodes, ie 1152 cores.
 - Up to 72x speedup in compute: I/O is 39% of run time (22'16" in computation and 14'24" in I/O).
- Parallel I/O is needed to
 - Spend more time doing science
 - Not waste resources
 - Prevent affecting other users
 - Easier to use and get the performance without major change in the workflow.

Shaheen III Hardware Specifications: Storage

Characteristics Tiered storage	Shaheen III /scratch storage /scratch/username
Total Capacity (usable)	32 PB
Capacity tier (HDD)	25 PB
Capacity tier perf Read/Write	330/260 GB/s
BW tier capacity	6.8PB
BW Perf. tier Read/Write	3750/2500 GB/s
IOPS tier capacity	338 TB
IOPS tier IOPS (Read/Write)	10+M IOPS

Characteristics	Shaheen III /projects /project/kxxxxx
Total Capacity (usable)	57 PB*
Capacity perf Read/Write	~200/120 GB/s

Shaheen Storage

- **/scratch**
 - Read + Write on login and Compute nodes
 - Fast file system
- **/project**
 - Read only on Compute nodes
 - Read + Write on PPN and DTN
 - Keeping large file for the lifetime of the project
- **/home**
 - Only available on login nodes
- <https://docs.hpc.kaust.edu.sa/policy/shaheen3.html>



Tiered Storage /scratch

Shaheen 3 uses tiered storage for Scratch (80TB quota per PI)

- Capacity Tier: default
/scratch/username or cd \$SCRATCH
10 TB quota per user
- Bandwidth Tier
/scratch/username/bandwidth cd \$SCRATCH_BW
1 TB quota per user
For files which do require high bandwidth, large files...
- IOPs Tier
/scratch/username/iops or cd \$SCRATCH_IOPS
50 GB quota per user
For small files with high IOPs pattern requirements



Shaheen Data Transfer

- Data transfer between tiers: just **copy** the data with the **cp** command.
- Data transfer between scratch <-> project: SLURM job script:

```
user@login4:> cat dtn_copy_job.sh
#!/usr/bin/env bash
#SBATCH -N 1
#SBATCH -n 1
#SBATCH -t 20
#SBATCH -p dtn
#SBATCH --output=SLURM_LOGS/copy_out.txt
#SBATCH --error=SLURM_LOGS/copy_err.txt
#SBATCH --job-name=COPY_project_to_scratch
rsync <options> -v --progress /lustre2/project/k????? /scratch/user/
```

- For outside Shaheen, use scp command

scp -r [myusername@shaheen.hpc.kaust.edu.sa:/path/directory](#) .

- For large files, Globus via dtn6. Contact help@hpc.kaust.edu.sa for any guidance.

Storage Best Practices

- /scratch meant for temporary use during the lifetime of the job.
- /project for persistent storage during the lifetime of the project
- Users must delete their unused data from scratch and move necessary files to project filesystem.
- For overall quota on scratch:
`lfs quota -uh $USER /scratch`
- For quota on capacity tier on scratch:
`lfs quota -uh $USER --pool capacity /scratch`
- For quota on bandwidth tier on scratch:
`lfs quota -uh $USER --pool bandwidth /scratch`
- For quota on IOPS tier on scratch:
`lfs quota -uh $USER --pool iops /scratch`

Storage Best Practices 2

- My usage of storage quota: command `kuq`

`kuq`

Filesystem quota limits for user hadrib

Tier	Filesystem	used	quota	limit	grace	files	quota	limit	grace
capacity	/scratch	119.1G	0k	10T	-	240782	0	0	-
bandwidth	/scratch	924.1G	0k	1T	-	240782	0	0	-
iops	/scratch	60.88M	0k	50G	-	240782	0	0	-
project	/project	7.778T	0k	0k	-	209596	0	1000000	-

`lfs quota -uh $USER /scratch`

Disk quotas for usr xxx (uid xxx):

Filesystem	used	quota	limit	grace	files	quota	limit	grace
/scratch	1.043T	0k	11T	-	240782	0	1024000	-

`lfs quota -uh $USER /project`

Disk quotas for usr hadrib (uid 129285):

Filesystem	used	quota	limit	grace	files	quota	limit	grace
/project	7.778T	0k	0k	-	209596	0	1000000	-

Storage Best Practices 3

- Quota for /project, kpg project-id . Quota 80 TB per PI.

```
kpg k10005
```

```
-----  
PI quota for : Principal Investigator  
-----
```

Filesystem	used	quota	limit	grace	files	quota	limit	grace
/project	16k	0k	80T	-	4	0	0	-

- Don't be shy, ask for help. help@hpc.kaust.edu.sa
- Limit the number of files per directory
- Clean up after running a job. (remove temporary, slurm output...)
- Important files shall be copied to your personal workstation.

Shaheen III: Best Practices summary

Bilel Hadri, KAUST Supercomputing Lab

February 4th 2025

Who can use Shaheen III?

- Every Shaheen user must be an official member of at least one project, and every project must originate from an approved organization.
- PI needs to be a faculty/manager to endorse the project.
- Access to Shaheen is available to a select group of academic and industry partners.
- Following the terms and conditions
- Submitting all necessary documents

Charges of core hours

- Unit of measure is core hours for the CPU nodes
 - Each node has 192 cores.
 - For exclusive usage, full node (ie 192 cores) will be charged
- Check the usage of the project: `sb k1xxxx`

```
Title: KSL computational scientists
PI : Saber Feki
-----
Project k01      expiry: 2030-12-31
-----
Allocations:
2024-02-08      100000000
-----
Total allocations:      100000000
Core hours used:        1300532
Remaining balance:      98699468
-----
```


Queue

- FIFO policy with backfilling as long as
 - core hours available
 - Maximum of number of jobs/nodes not reached
- 5 QoS
 - workq (entire system up to 4608 nodes)
 - Limit per user 2048 nodes max, 500 job running, 1000 max job queued
 - shared (16 nodes in total)
 - debug (4 nodes in total)
 - 1 node per user
 - 72hours: up 128 nodes partition shared by all users
 - 16 nodes maximum per user, 3 job running, maximum 32 job queued
 - ppn nodes: post processing nodes: 15 nodes

Storage quotas for each tiers,

- /scratch
 - Total aggregated PI 80 TB default
 - Per user 1M files, 10 TB in /scratch/username and 1 TB in bandwidth and 50 GB in IOPS
- /project
 - Total aggregated PI 80 TB default
 - Per user 1M files

- Check:

- Check the quota with `kpq, kuq` on project and overall

- `lfs quota -uh $USER /scratch`

- `Disk quotas for user username (uid 123456):`

- | Filesystem | used | quota | limit | grace | files | quota | limit | grace |
|------------|--------|-------|-------|-------|--------|-------|---------|-------|
| /scratch | 402.7G | 0k | 11T | - | 123339 | 0 | 1024000 | -16 |

Acknowledging

Terms and Conditions regarding Research Publications

Whenever the results of research conducted on the HPC systems at KAUST are published, or the research involved personnel from KAUST Supercomputing Laboratory (KSL), Principal Investigators (PIs) are required to acknowledge the usage of the HPC systems at KAUST and/or the involvement of KSL personnel in their research in their publications. For example, the following statement could be used:

“For computer time, this research used Shaheen III managed by the Supercomputing Core Laboratory at King Abdullah University of Science & Technology (KAUST) in Thuwal, Saudi Arabia.

Best Practices / Getting help

- When submitting a ticket to help@hpc.kaust.edu.sa requesting help, you will likely get faster resolution by supporting a few best practices:
 - Where possible, provide helpful details that can help speed the process. For example: Project ID, relevant directories, job scripts, jobIDs, modules at compile/runtime, login name, etc.
- One issue per ticket. Do not add unrelated questions on existing tickets
- Do not open multiple tickets on the same unresolved topic.
- Let us know if your issue is fixed or you solved it (and let us know what worked!)
- **MOST IMPORTANT:** do not hesitate to contact us help@hpc.kaust.edu.sa;

Best Practices

- Check the documentations: <https://docs.hpc.kaust.edu.sa/policy/shaheen3.html>
- Read the announcements/newsletter sent by emails and available in the website:
 - hpc.kaust.edu.sa/

Thanks !

Agenda

- 8:30am Welcome
- 8:35am Shaheen III Hardware Overview
- 8:55am How to apply on Shaheen III
- 9:05am Getting Started on Shaheen III
- 9:15am Software Environment
- 9:35am Job Scheduling
- 10:00am Coffee Break
- 10:15am Storage overview & Best practices
- **10:30am Applications software example: VASP workflow**
- 10:50 am Applications software example: CFD applications
- 11:10 am Applications software example: Bio informatics workflow
- 11:20-11.30am Q&A and Open Discussion

