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).

Δ

- 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
Fotal 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_BW1 TB quota per userFor 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:

•

•





- For outside Shaheen, use scp command
 scp -r myusername@shaheen.hpc.kaust.edu.sa:/path/directory
- For large files, Globus via dtn6. Contact <u>help@hpc.kaust.edu.sa</u> 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

Filesystem quota	a limits fo	or user had	 rib							
Tier	Filesyste	em 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	Øk	0k	-	209596	0 10	00000	-	
lfs quota –uh s	SUSER /scr	atch								
Disk quotas for	r usr xxx	(uid xxx):								
Filesyster	n used	quota	limit	grace	files	quota	limit	grace		
/scratch	1 . 043	T Øk	11T	-	240782	2 0	1024000	-	-	
									5 or 2 0	
its quota -un s	SUSER /pro	јест							•	
Disk quotas for	r usr hadr	ib (uid 12	9285):						•	
• Filesyster	n used	quota	limit	grace	files	quota	limit	grace		
/project	t 7.778T	0k	0k	-	209596	0	1000000	• • • –		10

Storage Best Practices 3

Quota for /project, kpq project-id . Quota 80 TB per Pl. • kpg k10005 PI quota for : Principal Investigator Filesystem used limit files quota limit grace quota grace /project 16k 80T **0**k 0 0

- Don't be shy, ask for help. <u>help@hpc.kaust.edu.sa</u>
- 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 computation PI : Saber Feki	al scientists	
Project k01 expiry:	2030-12-31	
Allocations: 2024-02-08	10000000	
Total allocations: Core hours used:	10000000 1300532	
Remaining balance:	98699468	14

Queue

- FIFO policy with backfilling as long as
 - core hours available
 - Maximum of number of jobs/nodes not reached

• 5 QoS

0

- 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	C: Marken and a state of the state									
	0	heck the quota with k	oq, kuq on pi	roject and ov	verall						
	o I o I	lis quota -un \$0 Disk quotas for	USER /scra user usei	atch rname (ui	.d 12345	6):					
	0	Filesystem	used	quota	limit	grace	files qu	ota li	mit	grace	
	0	/scratch	402.7G	0 k	11T	- 1	123339	0 1024	000	- 16	

Applications available

- Over 500 version of libraries, tool and applications available
- Module avail ; module avail –S nameofsoftware

perftools perftool	s-lite perftools-lite-ev	ents perftools-lite-gpu pe	rftools-lite-hbm perftools-lite-	loops perftools-preload	GIE11163				
PrgEnv-acc/8.4.9(default) PrgEnv-acc/8.5.0 PrgEnv-acc/8.5.0 PrgEnv-acc/8.5.0 PrgEnv-acv/8.5.0 PrgEnv-acv/8.5.0 PrgEnv-acv/8.5.0 PrgEnv-mk5.5.0 PrgEnv-mk5.6.4.9(default) PrgEnv-mk5.6.4.9(default) PrgEnv-mk5.6.4.9(default) PrgEnv-mk5.6.4.9(default) PrgEnv-mk5.6.4.9(default) PrgEnv-mk5.6.4.9(default) PrgEnv-mk5.6.4.9(default) dr/3.15.2 cce/14.6.40 cce/12.3.40(default) cce/23.30(default) cpe/23.85	cp=-cuda/23.05 cp=-cuda/23.07 cp=-cuda/23.07 cp=-cuda/23.07 cr=y=d/4.3.1 cr=y=-cuda/25.07 cr=y=-cuda/25.07 cr=y=-cuda/25.07 cr=y=-cuda/25.07 cr=y=-cuda/25.07 cr=y=-cuda/25.07 cr=y=-cuda/25.07 cr=y=-dr=17.07 cr	efault) efault) default) (default) (default) 2.0 3.1 3.1 4.3 4.6(default) 1.3 7.(default) 1.3 7.(default) 1.3 7.0 5.3 7.0 5.3 7.0 5.3 7.5 7.5 7.5 7.5 7.5 7.5 7.5 7.5	<pre>cray-hdf6-parallel/1.12.2.3 cray-hdf6-parallel/1.12.2.7(default cray-hdf6-parallel/1.12.2.7) cray-hdf6-parallel/1.12.2.7(default 1.12.7).475.90.475.475.475.475.475.475.475.475.475.475</pre>	<pre></pre>	cra cra cra cra cra cra cra cra cra cra	<pre>wy-python/3.18.18(default) wy-python/3.11.5 y-tstr/.12.12 wy-tstr/.12.12 wy-tstr/.12.12 wy-tstr/.12.12 wy-tstr/.12.12 wy-tstr/.12.12 wye/2.7.23 wy/2.7.23 wy/2.7.2</pre>	gcc/18.3.0 gcc/13.2.0 gcc/13.2.0 gcc/13.2.0 gcc/12.3.0 gdxhpc/.4.15.1 gdxhpc/.4.15.1 gdxhpc/.4.15.1 gdxhpc/.4.15.1 gdxhpc/.4.15.1 gdxhpc/.4.15.1 gdxhpc/.4.15.1 gdxhpc/.1.1 papi/7.0.1.1 gdxhpc/.1.1 papi/7.0.1.2 papi/7.0.2 papi/7.0.1.2 papi/7.0 papi/7.0 papi/7.0 papi/7		
chapel/1.30.0 cray-lustre-client-ofed/2.15.0	.7 rc2 cray 25 ga33b7d9-2.5 5.2	4 ga33b7d9745.shasta(default	cudatoolkit/23.3_11.8) cudatoolkit/23.3_12.0(default)	/opt/cray/modulefiles		xpmem/2.6.2-2.5_2.27gd067c3f.shasta(d	efault)		
cudatoolkit/23.3_11.0			libfabric/1.15.2.0(default)	/ont/modulafiles					
amduprof/4.2.850	aocc/4.2.0(default)	intel/19.0.5.281	intel/2023.1.0(default)	intel-classic/2023.1.0(default) int	el-oneapi/2023.1.0(d	iefault)			
craype-accel-amd-gfx90a craype craype-accel-amd-gfx908 craype craype-accel-amd-gfx940 craype	-accel-host craype-arm-gi -accel-nvidia70 craype-hugepi -accel-nvidia80 craype-hugepi	ace craype-hugepages2M ges1G craype-hugepages4M ges2G craype-hugepages8M	craype-hugepages16M crayp craype-hugepages32M crayp craype-hugepages64M crayp	 /opt/cray/pe/craype-targets/default/m ee-hugepages128M craype-network-none ee-hugepages256M craype-network-ofi te-hugepages512M craype-network-ucx 	craype-x86-geno craype-x86-mila craype-x86-mila	pa craype-x86-rome craype-: in craype-x86-spr in-x craype-x86-spr-hbm	x86-trento		
abinit/9.4.2 abinit/9.10.3 abinit/9.10.3 abinit/9.10.4 alimode/1.3.0 alimode/1.3.0 alimode/1.4.2 alimote/1.4.2 amber/14_mpiplic amber/14_mpiplic amber/14_mpiplic amber/18_mpi amber/18_mpi amber/18_mpi amber/18_mpi amber/18_mpi amber/18_mpi amber/18_mpi amber/18_mpi amber/18_mpi amber/14_mpiplic a	berkelevgw/2.1 berkelevgw/3.1e bio/bwaf/s.1e bio/bwaf/s.7.7 bio/dts/2.2.0 bio/gatk/4.1.6 bio/gatk/4.1.6 bio/gatk/4.1.6 bio/gatk/1.78 bio/link/2.8e bio/link/2.8e bio/amtois/1.8 bio/amtois/1.8 bio/trap2/24.1.3 bost/1.85 bittrap/12.5 bolttrap/24.1.5 bost1/1.85 chembhl/23.0.1 chembhl/23.0.1 chembhl/23.0.1	converge/3.8.22_udf converge/3.1.4 converge/3.1.5 converge/3.1.5 converge/3.1.5 converge/3.1.6 converge/3.1.6 converge/3.1.8 converge/3.1.8 converge/3.1.8 converge/3.1.8 converge/3.1.8 converge/3.1.8 converge/3.1.8 converge/3.1.8 converge/3.2.1 converge/3.2.1 converge/3.2.1 converge/3.2.1 converge/3.2.1 converge/3.2.2 converge/3.2.2 converge/3.2.4 converge/3.4.4 converge/3.2.4 converge/3.4.4 c	egsnrc/2028 egsnrc/2028 els/7.3.2 els/6.2.2 els/6.2.35.601 espress/6.4.1 espress/6.4.1 espress/6.8.else espress/6.8.else espress/6.8.else espress/7.2 else espress/7.2.1 espress/7.3.1 e	<pre></pre>	<pre>molpro/2012.1p16 moltemplate/2.02.2 molt/2.</pre>	packmol/20.14.3 paraview5.11.2	<pre>smins/20220122 sod/0.47 sod/0.47 sod/0.47 sum/7.3.8 sum/7.3.8 sum/7.3.8 sum/7.3.8 sum/7.3.8 sum/7.3.8 sum/7.3.8 sum/7.3.8 sum/7.3 temporf/202322 temporf/202322 temporf/202322 temporf/202322 temporf/202323 temporf/202323 temporf/202323 temporf/2014 temporf/202323 turboms/7.1 sum/7.1 sum/7.5 sum/7.</pre>	<pre>vaspkit/1.5.1 vasputi/6.1 vitual/16/15/7 visit/3.3.3 vmd/1.0.4 wannierbol/2.1.0 wannierbol/2.1.0 wannierbol/2.1.0 wannierbol/2.5.1 wannierbol/2.5.1 wannierbol/2.1.elpa wierk/2.1.elpa wierk/2.1.elpa wierk/2.1.elpa wierk/2.1.elpa wierk/2.5.cruy warf/4.5.2.cruy wff/4.5.2.intol xcrydden/1.5.00</pre>	
ansys vzsk-riulds ansys vzsk-riulds ans/1/4.2,0(default) ass/3.10.6 ass/3.22.1 ask/2019.03spl atompaw/4.2.0.3 autodockvina/1.2.3 bader/1.46 bader/1.46 bader/1.46 bader/1.66 bazel/6.16 bazel/6.16	c1r201/1.2.18 cnak(3.28.3 4 c0lumbus/7.2 columbus/7.2 comsol/6.2_tmp converge/3.8.13 converge/3.8.13 udf converge/3.8.13 udf converge/3.8.25 converge/3.8.25 converge/3.8.26 converge/3.8.26	crrt1c/1.1stable crrt1l/1.0.3 deepspeed/0.14.0 dftpDjus/1.2 dftqd/2.5.0 djpDj/4.09 djpDj/5.1.0 djpDj/5.1.0 djpDj/5.1.0 dipDj/6.2 eddp/0.2 eddf/0.2 eddf/0.2 eddf/0.2	<pre>rmpeg/s.l.1(default) rhiaims/230216.3 rhiaims/231212.1 rlks/23.212.1 rlks/23.212.1 rlus/23.6.4 fourphonon/2.11 gamess/38spt22222.2 gamess/38spt22222.2 gamess/38spt222232 gamess/38spt222322 gamess/38spt222322 gamess/38spt22232 gamess/38spt2232 gamess/38spt2232 gamess/38spt2232 gamess/38spt2232 gamess/38spt2232 gamess/38spt2232 gamess/38spt2232 gamess/38spt2232 gamess/38spt223 gamess/38spt23 gamess/38spt223 gamess/38spt223 gamess/38spt223 gamess/38spt23 gamess/38s</pre>	1Dox7.9.1.7 1Dox7.9.2.7 1Dox7.9.2.7 1Dox167.9.0 materstudio/2023 milo/1.0.3 mclio/1.0.3 mclio/1.0.3 mclio/1.2.1 mclio/1.2.2 mclio/1.2.2 mclio/1.2.2 mclio/1.2.2 mclio/1.2.2 mclio/1.2.2 mclio/1.3.3.2 mclio/1.3.3.3.3.3 mclio/1.3.3.3.3.3.3.3.3.3.3.3.3.3.3.3.3.3.3.3	octopus/13.8 commf/2.0alpha3 commf/2.1alpha8 openbabel/3.1.1 openrolcas/23.10 openmolcas/23.10 openm/3.9.9 orca/5.0.4 ovita/2.9.0 ovita/3.0.1 p4vasp/0.3.30 packem/20200322 packem/18.169	<pre>qcnem/6.1 qchem/6.2 quantumite/1.1 qvasp/2.23 rasps/2.2.3 seiss/1/1.1.4_gnu shengbte/1.5.6 sisti/.1.5 sis</pre>	<pre>vssp/s.4pl2_ltbsef vssp/s.4pl2_ltbsef vssp/5.4pl2_ltbsef vssp/5.4pl2_occmat vssp/5.4pl2_occmat vssp/5.4pl2_vsspol vssp/5.4pl2_vsspol vssp/5.4pl2_vsspol vssp/5.4pl2_vsspol vssp/6.4pl2_vstr78 vssp/6.4.2_stsp0 vssp/6.4.2_stsp0 vssp/6.4.2_vstr198 vssp/6.4.2_vstr198</pre>	<pre>xtb/6.4. xtb/6.4. xtb/6.7.0 xtb/6.7.0 xtb/6.7.1 yambo/5.0.4 yambo/5.0.4 yambo/5.2.1 yambo/5.2.1 signer zZpack/2.2.0 zendmn/4.1</pre>	17

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





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
- 10:00am
- **10:15**am
- 10:30am

10:50 a

orkflow

- Job Scheduling Coffee Break
- Storage overview & Best practices

Applications software example: VASP workflow

Applications software example: CFD applications Applications software example: Bio informatics

11:20-11.30am Q&A and Open Discussion

