Introduction to Gadi Supercomputer

NCI Training & Educational Events

Javed Shaikh | Staff Scientist | User Services

April 2024

Acknowledgement of Country

The National Computational Infrastructure acknowledges, celebrates and pays our respects to the **Ngunnawal** and **Ngambri** people of the Canberra region and to all First Nations Australians on whose traditional lands we meet and work, and whose cultures are among the oldest continuing cultures in human history.

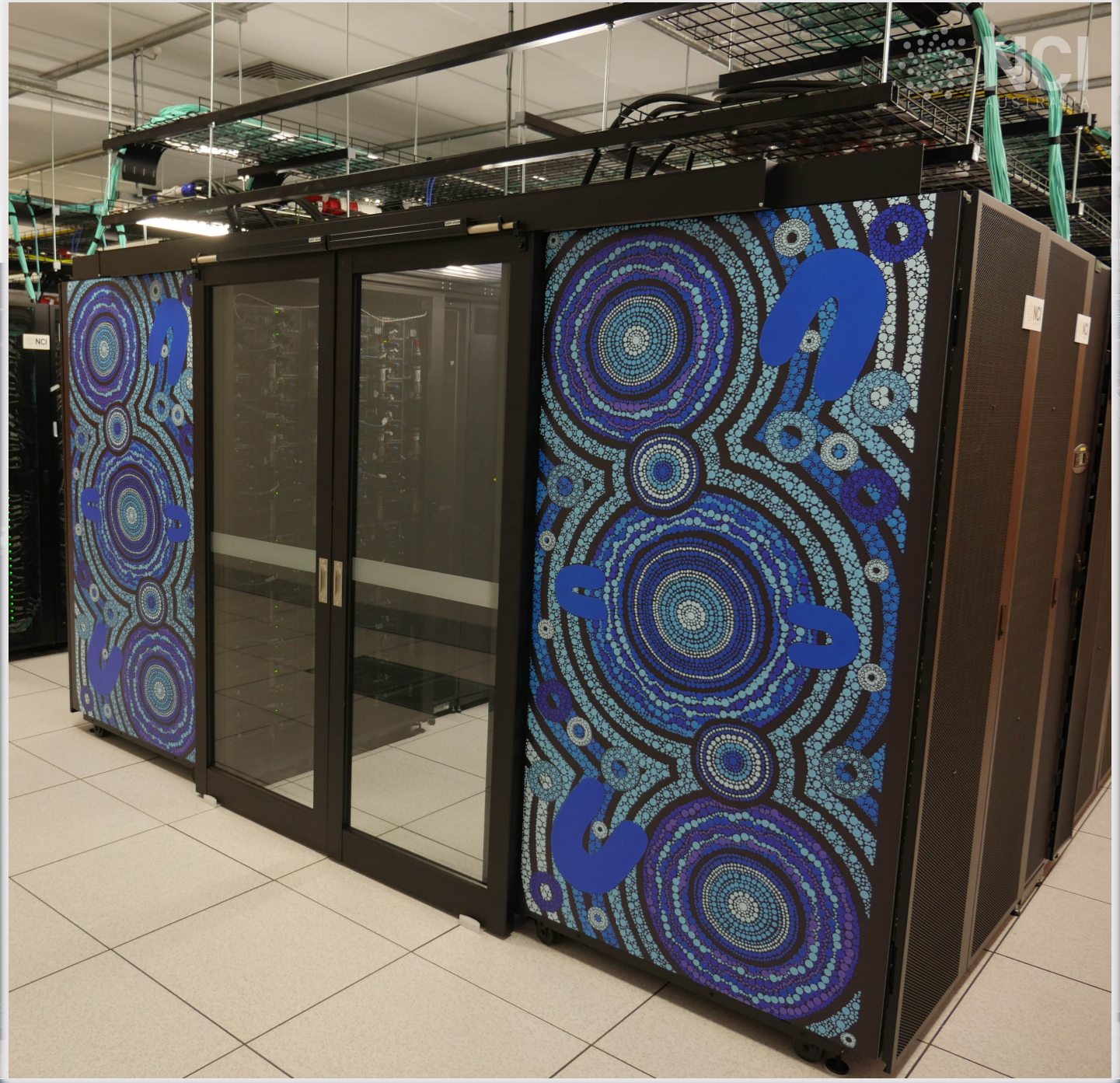
Agenda

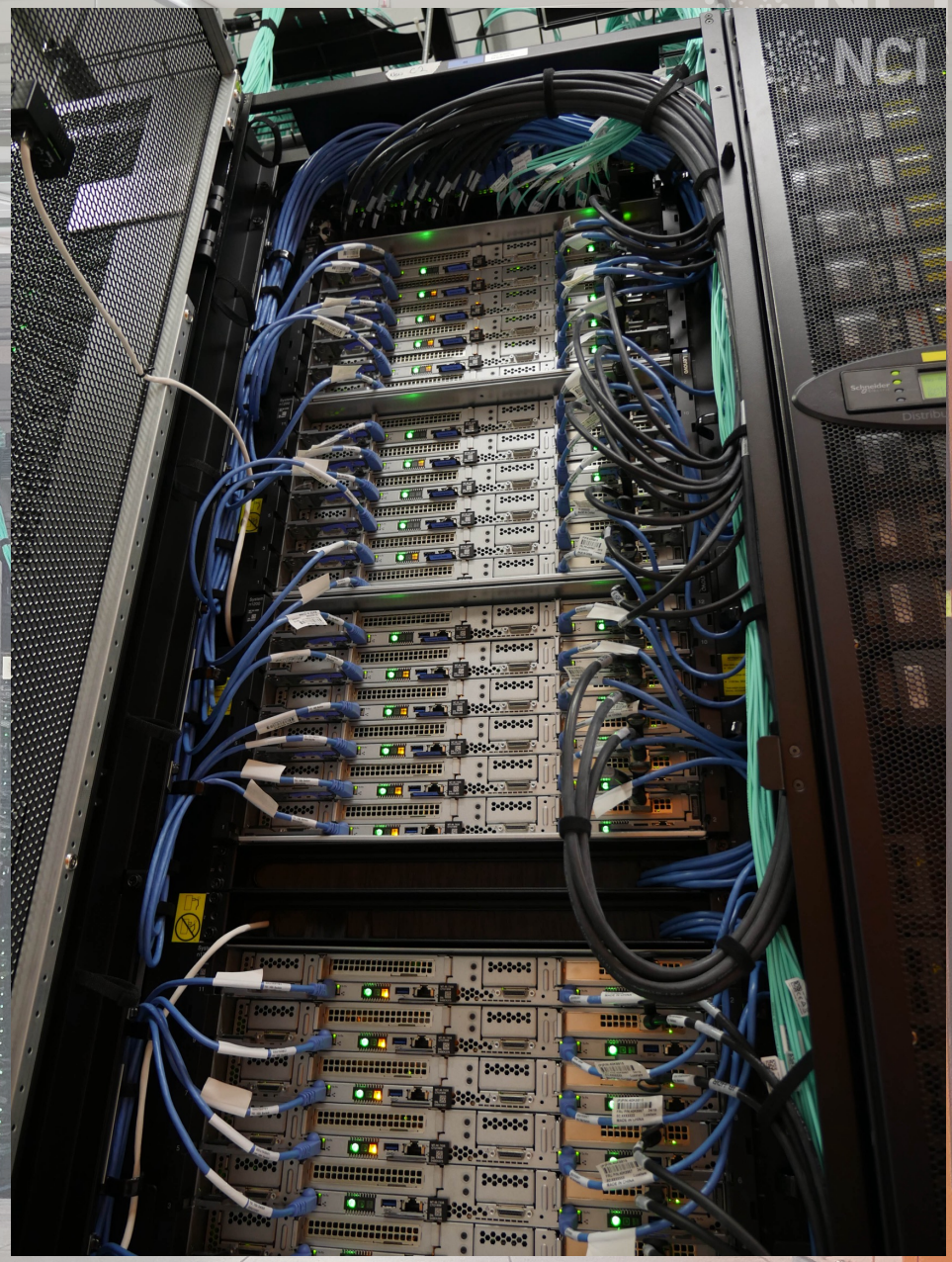
- Introduction
- Account
- Login
- Storage and Data Transfer
- Applications
- Jobs

About NCI

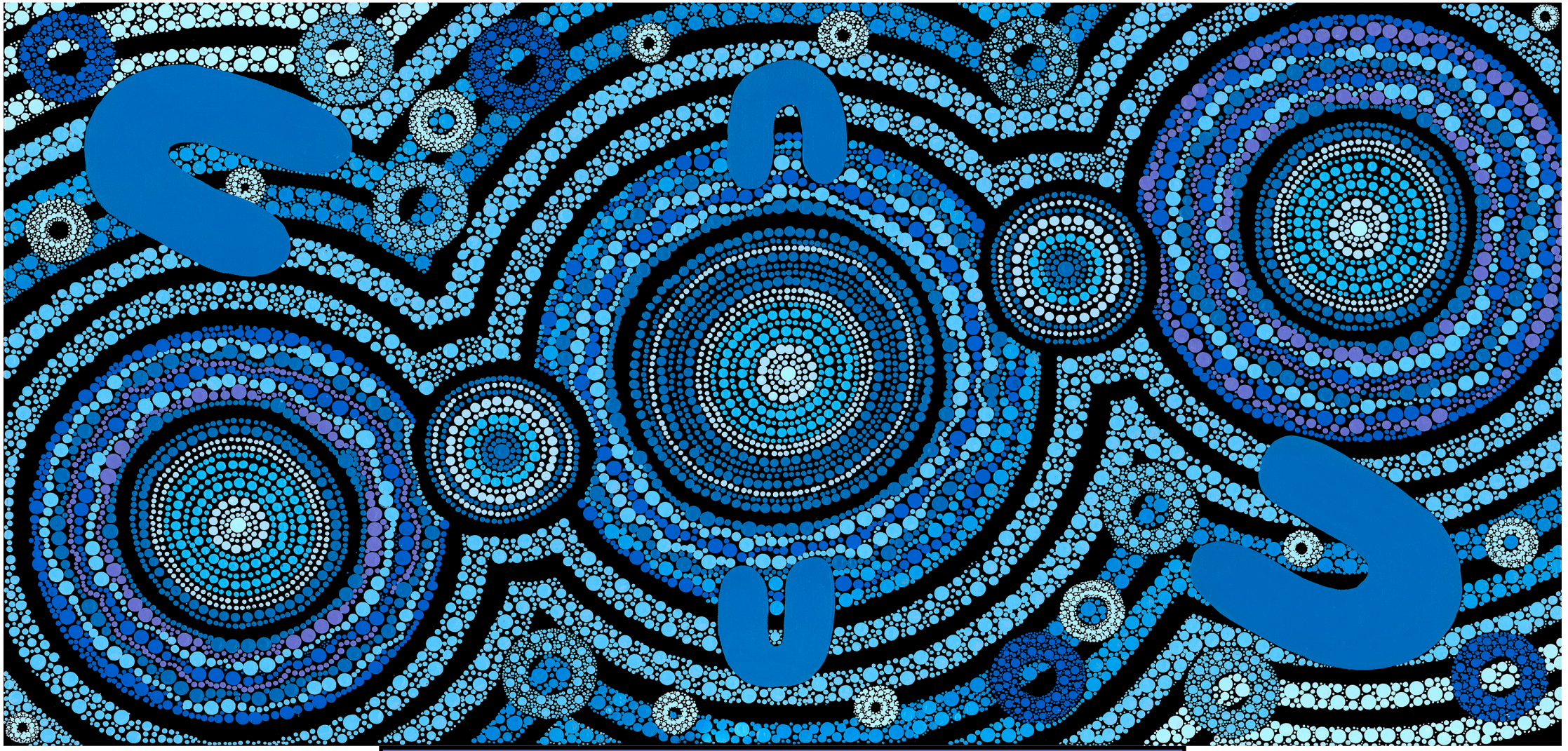
- NCI is the premier facility providing:
  - High-performance computing – **GADI**
  - Cloud computing – **NIRIN**
  - Data storage and services – **Global Filesystems**

- NCI is part of The Australian National University and located in Canberra



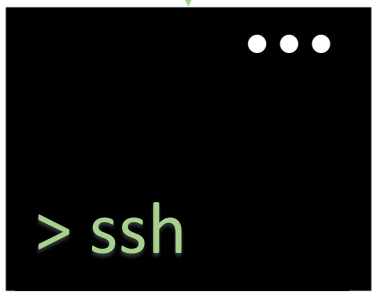


Gadi Artwork



Artist: Lynnice Letty Church – Tribes: Ngunnawal, Wiradjuri & Kamilaroi (ACT and NSW)
Gadi - "to search for" in Ngunnawal language - January 2020 for NCI Gadi Supercomputer

Interfaces to Gadi



aarnet

jupyter

R

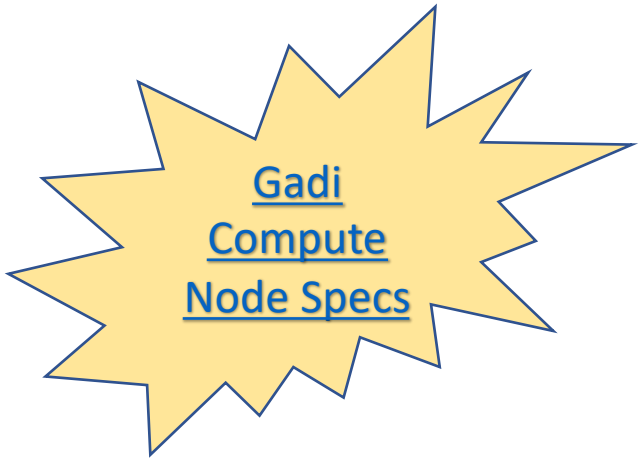
are
australian research environment

NCI AUSTRALIA

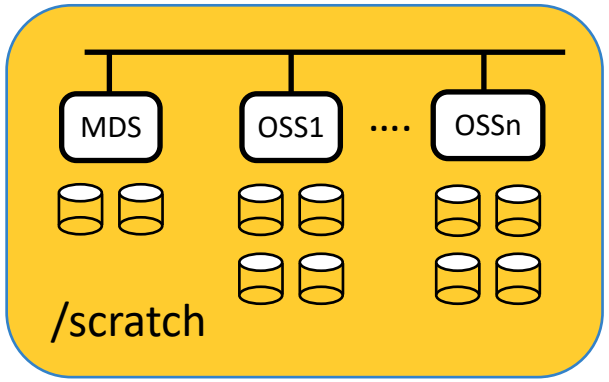
OPEN OnDemand

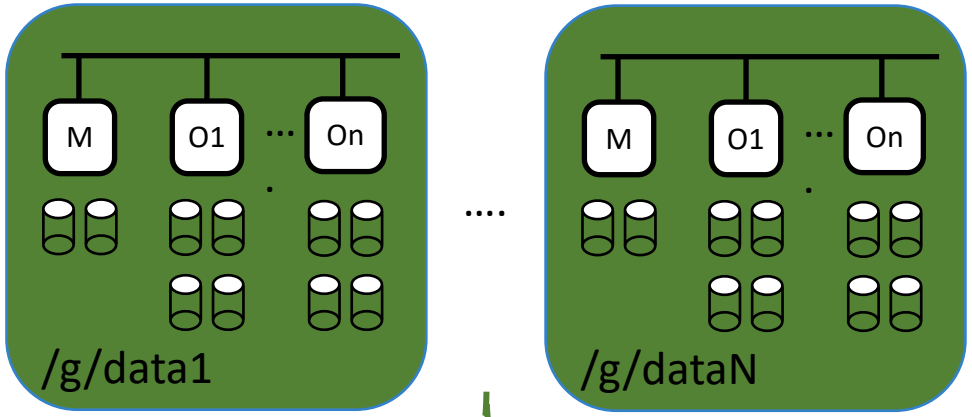
Gadi Specifications

- Australia's fastest CPU-based research supercomputer with
  - 185,880 compute cores (Intel Cascade Lake, Skylake, Broadwell)
  - 640 NVIDIA V100 GPUs in 160 nodes, 2 NVIDIA DGX A100 nodes
  - 22 PiB of high speed scratch storage with max IO speeds of 490 GiB/s
  - 200 Gb/s Infiniband HDR network
  - Operating System: Rocky Linux
  - **15.14** (peak) / **9.26** (sustained) Pflop system ranked 24th fastest in the world on debut in 2020 (currently #83) – https://www.top500.org/lists/top500/2023/11/
  - +74,880 compute cores (Sapphire Rapids) in 720 nodes = ~260760 cores



Gadi Ecosystem





data services

/home

/apps

/scratch

/g/data

massdata

Infiniband 200 Gb/s

login nodes
01 .. 10

data mover nodes
01 .. 06

compute nodes
4800

aarnet

External systems

- Global data filesystems (gdata)
  - A collection of Lustre parallel filesystem blocks to store large data files for longer period
  - 80 PiB storage space now and counting
  - Space managed by stakeholders
  - Similar to scratch filesystem in terms of access and usage
- Massdata
  - 70 Petabytes of archival project data in state-of-the-art magnetic tape libraries
  - Multiple copies over multiple locations for disaster management
  - Access on Gadi through special utility ***mdss***

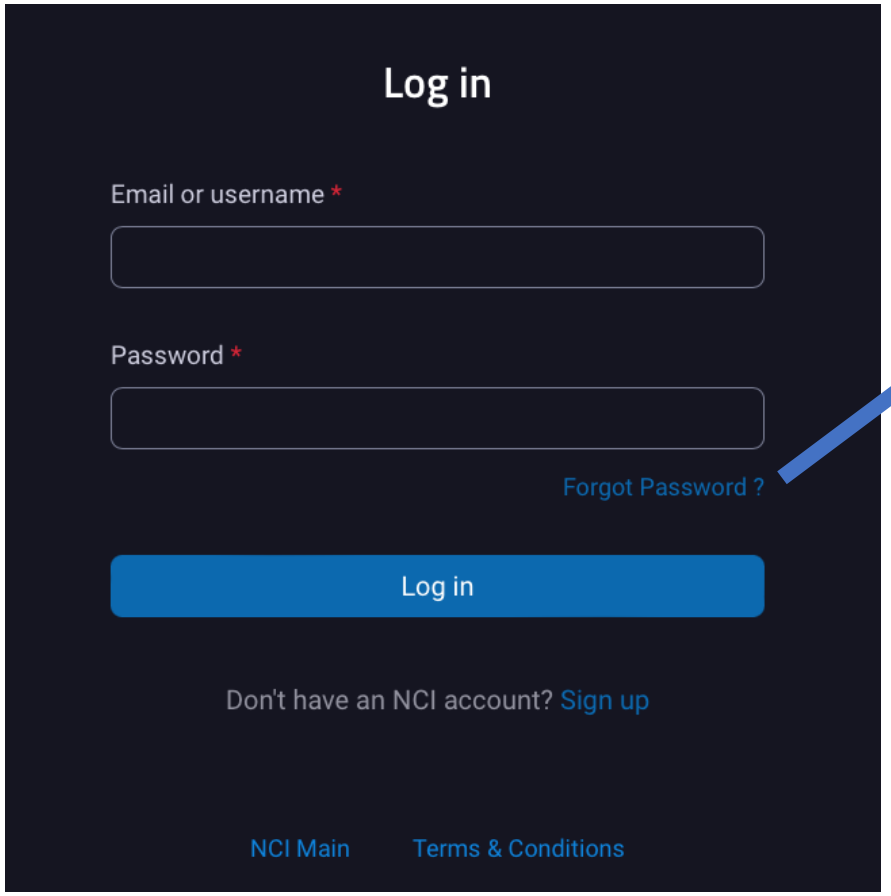

myNCI

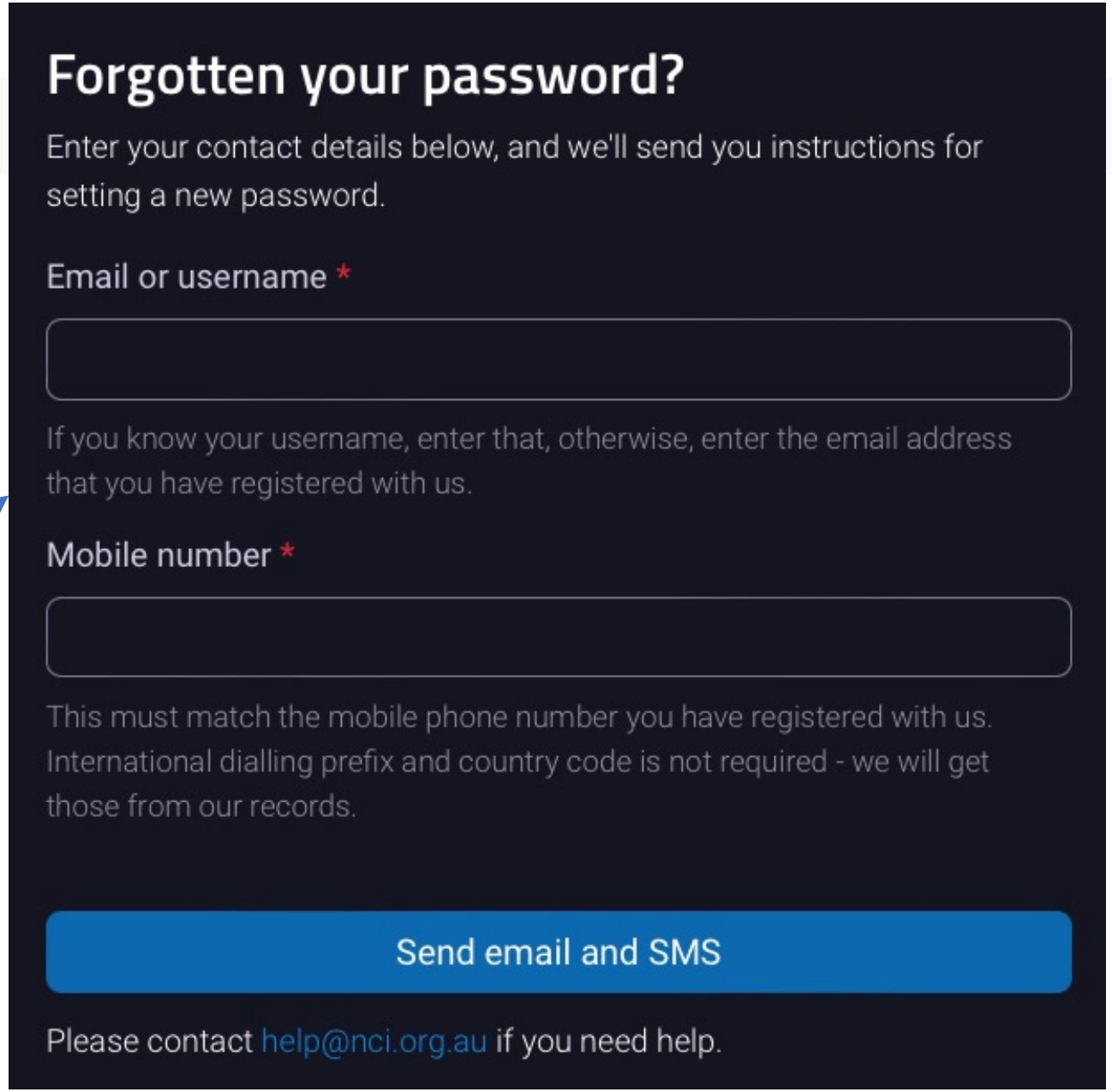
NCI Account

- Account is for a **lifetime**
- Always keep contact information up-to-date
- Recertify once a year. This includes changing your password and accepting **Conditions of Use** agreement. A reminder email sent to registered email address one month prior to “Recertification due date”
- If not recertified in time, account will go into suspended mode for 120 days. Beyond that it will be deactivated
- A deactivated account can always be revived by writing to NCI Helpdesk (help@nci.org.au)

NCI Account

my.nci.org.au





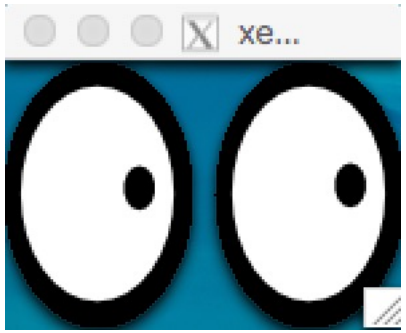

Login

ssh to gadi

ssh jjj777@gadi.nci.org.au

Mac	ssh XQuartz
Windows	Putty MobaXterm
Linux	ssh startx

```
me@local:~ $ ssh -Y jjj777@gadi.nci.org.au
jjj777@gadi.nci.org.au's password:
[jjj777@gadi-login-05 ~]$ xeyes
[jjj777@gadi-login-05 ~]$ exit
me@local:~ $
```



Login Environment

- Round-robin login
- Message of the day (motd)
- Account status information

- Environment check : whoami, hostname, default shell, gadi project, home dir

- Linux commands quick reference : pwd, ls, cd, mkdir, cp, mv, cat, less, vim, man, etc.

- Setting default linux shell and gadi project in ~/.config/gadi-login.conf
  - ~/.bashrc for SHELL=/bin/bash or ~/.cshrc for SHELL=/bin/csh etc.
  - Caution: Incorrect editing may lock you out !

Login nodes

- Access restrictions

- On a login node you can:
  - Edit files, build programs, install software in your home/project space, etc.
  - Download/upload small amount of data
  - Run/test/debug programs:
    - Not exceeding 30-minute CPU cumulative time limit
    - Not exceeding 4GiB memory
  - Submit and monitor PBS jobs

- …

Storage and Data Transfer

Storage areas

Filesystems	Path	Critical Info
home	/home/institutionCode/username /home/777/jjj777	• Personal space • Backed up
scratch	/scratch/project/username /scratch/c25/jjj777	• Project space providing fastest large scale IO speeds • Temporary storage for input/output files to/from HPC applications • Files not accessed for 100 days will be automatically removed • Not backed up. Data once deleted can *never* be recovered
global data	/g/data/project/username /g/data/c25/jjj777	• Project space for long term data storage • Can also be storage space for input/output files to/from HPC applications • Data can be made visible via other interfaces • Not backed up. Data once deleted can *never* be recovered
mass data	mdss –P c25 ls -l	• Tape based backup system for archiving large data files of a project • Need to use *mdss* utility to access dirs in massdata store
applications	/apps/software/version /apps/python3/3.10.4	• Centrally installed software applications and their module files • Readonly access

Storage areas

Filesystems	Data Ownership	Allocations	iNode Limit
home	User	• Fixed default 10GiB	
scratch	Project	• 1TiB default • Managed by NCI • More space allocated if reasonable justification is provided	• Limited
global data	Project	• Managed by sponsoring scheme/institution • For more space discuss with project CI / scheme manager	• Limited
mass data	Project	• Managed by sponsoring scheme/institution • For more space discuss with project CI / scheme manager	• Limited
applications	NCI		

Storage utilities

Util	Information
quota	• Provides home quota and usage *quota -s*
lquota	• Provides quota and usage for all connected project spaces on scratch and/or gdata filesystem *lquota*
nci_account	• As above + gives the sponsoring scheme name • Also total compute allocations, and compute time usage by each user *nci_account -P c25 -v -p 2024.q1*
nci-files-report	• Gives the data foorprint for a project data on scratch and/or gdata *nci-files-report -p c25 -f scratch*
nci-file-expiry	• Scratch data expiry management tool *nci-file-expiry list-quarantined*

Data transfer

```
me@local:~ $ scp -p newsample.mph jjj777@gadi-dm.nci.org.au:/scratch/c25/jjj777/
jjj777@gadi-dm.nci.org.au's password:

newsample.mph                                 100%  218MB    2.2MB/s   01:38
```

upload

```
me@local:~ $ scp -p jjj777@gadi-dm.nci.org.au:/g/data/c25/jjj777/README.pdf
/Users/me/Downloads/
jjj777@gadi-dm.nci.org.au's password:

README.pdf                                    100%  299KB    2.8MB/s   00:00
```

download

Data transfer utilities

- Secure copy (scp), secure file transfer protocol (sftp)
- rsync, aspera, aws client
- Filezilla, WinSCP
- …

Applications

Applications

- Central software repository with 200+ applications in /apps directory
- All built from source code and optimised for Gadi
- A given application is available via its module
- For an application not in central repository you can download and install in home/project dir
- NCI recommends **Intel compilers** and **OpenMPI** to compile and run applications

Applications: Modules

- module {avail, show, load, list, unload, purge}

- module load
  - modifies search/exec path
  - loads dependencies
  - handles conflicts
  - configures environment to define how the application runs

- Do:
  - Always start working in a clean environment
  - Always load specific version of application

Applications: License module and software group

- Restricted modules available to specific group of users
- Software groups control access to license modules
  - Example: matlab, ansys
- License modules tell the application where to checkout license
- Software groups control access to applications
  - Example: vasp

- To join a software group on my.nci.org.au:
  - search for the software group
  - read project overview
  - ensure eligibility criteria is being met
  - submit the membership request
  - wait for approval email
  - … takes roughly 30 minutes after the approval email for membership to be synchronised throughout the system


Jobs

Data transfer example

```
#!/bin/bash

#PBS -P c25
#PBS -q copyq
#PBS -l ncpus=1
#PBS -l mem=4GB
#PBS -l walltime=00:30:00
#PBS -l storage=gdata/c25
#PBS -l wd

export SOURCEDIR=/g/data/c25/jjj777/archive
export DSTDIR=/scratch/c25/jjj777/test

time cp -avr $SOURCEDIR $DSTDIR > /scratch/c25/jjj777/cp.log
```

PBS commands

- Submit standard or interactive jobs with *qsub*
- Check job status with *qstat*
- *qcat* is useful to see job error and output files during the jobrun
- *qdel* deletes jobs specified by their ids

Compute resource

- In order to run a job, a project needs to have compute allocation i.e. service units (SU)
- 1 SU gets you 30mins of 1 cpu time in a *normal* queue
- PBS will calculate and reserve the total number of SUs required to run your job:

  Charging rate in SU **X** Number of Cpus (or MemUnits) **X** Walltime
- Once compute allocations are exhausted, a job will be held in the queue until project gets more SU
- Compute allocations are usually made on quarterly basis, but can be increased/decreased/transferred to another project (under same stakeholder) anytime of the quarter:
  - Discuss with project chief investigator (CI) and/or allocation scheme manager of your institution
- If it is expected, allocations will not be used with-in a quarter, they can be rolled-over to next quarter in first two weeks of current quarter
- A project can have minimum 1000 SU i.e. 1KSU

Compute resource: Charging policy

Queue	SU / cpu / hour	SU / MemUnit / hour
copyq	2	2 (MemUnit=4GiB)
normal	2	2 (4GiB)
express	6	6 (4GiB)
hugemem	3	3 (32GiB)
megamem	5	5 (64GiB)
gpuvolta	3	3 (8GiB)
dgxa100	4.5	4.5 (16GiB)
normalsr	2	2 (5GiB)
expresssr	6	6 (5GiB)

You are charged on max of (ncpus, memUnits)

A job running in normal queue on 48cpus and mem <= 190GiB, with walltime of 4 hours will consume:

2SU x 48cpu x 4hours = 384SUs

A job running in normal queue on 1cpu and 12GiB mem, with walltime of 4 hours will consume:

2SU x 3mem x 4hours = 24SUs

Compute resource : Accounting with nci_account

- Provides compute allocation and usage to-date for a project for a given quarter
- Shows total SU usage by users of the project
- Displays SU reserved by PBS for user jobs in real time
- Also prints:
  - Total storage allocation and usage for scratch and/or gdata project space
  - Massdata usage
- Lists the sponsoring stakeholder/scheme name(s) for compute and storage allocations

Jobs: Putting it all together

- Compute resource: Service Units
- Storage resource:
  - Home directory (default)
  - Project space on scratch (default)
  - Project space on gdata (optional)
- Application(s)
- Time estimation

```
#!/bin/bash

#PBS -P c25
#PBS -q normal
#PBS -l ncpus=4
#PBS -l mem=8GB
#PBS -l storage=gdata/c25
#PBS -l walltime=00:10:00
#PBS -l wd

module load openmpi/4.1.3

cd ~/code/hpl-2.3/bin
mpirun -np 4 ./xhpl > /g/data/c25/jjj777/xhpl.out
```

Job monitoring

- nqstat_anu <job id>

```
                                  %CPU  WallTime  Time Lim     RSS    mem memlim cpus
12345678 R abc123  x11 myTest       33  10:53:56  20:00:00 58.7GB 58.7GB  200GB    96
19145286 R abc123  x11 atmos_ma    96  01:32:41  03:30:00  369GB  369GB 2625GB   768
19149497 R abc123  x11 coupled.    84  00:34:25  04:30:00  320GB  320GB 1440GB   720
19149708 R abc123  x11 netcdf_c    71  00:36:30  02:00:00 12.0GB 12.0GB 12.0GB     1
19150248 R abc123  x11 atmos_ma    86  00:22:27  03:30:00  345GB  345GB 2625GB   768
```

- qps <job id>
  - prints the snapshot of the current processes in the job
  - launches a *ps* query on each node running the job
  - accepts most flags *ps* would take
- qps_gpu <job id>
- qcat <job id>
  - print the job's standard streams
- Realtime using *top*
  - Login to the compute node and run *top* utility

- Compile program with *-g* (gcc) or *-g -traceback* (Intel compilers).

```
module load padb
padb -X pbs_job_id
```

- pstack
- attach gdb and get a stacktrace

Jobs submission options

- Interactive: *qsub -I -lstorage=gdata/c25+scratch/x11,wd job.sh*

- Other PBS directives:

```
#PBS -M <abc123>@<gmail.com>          #Sends you email at the start
#PBS -l software=matlab_nci           #Wait until matlab license is available
#PBS -e /scratch/c25/abc123/error.log #Redirect error to file
#PBS -l storage=gdata/c25+scratch/z00 #Project areas to be made visible
#PBS -a 202303241300                  #Wait until 1pm to start
```

- PBS Directives Explained

Why my job...

- has waited so long ?

```
qstat -u $USER -Esw
```

  - Insufficient amount of resource: ncpus
  - Project doesn't have sufficient allocation to run job
  - One of the project areas is already over disk quota
  - Waiting for software licenses
  - Job would not finish before dedicated time

- failed ?

```
Check job error/output files
```

  - File/directory not found [check -lstorage directive in jobscript]
  - Exceeding jobfs / memory / walltime limit [check job summary in output file]
  - Disk quota exceeded [quota, lquota, nci-files-report]

Helpdesk

Help us help you ☺

- Gadi User Guide
- help@nci.org.au
- When writing to helpdesk, always include following information:
  - Username, project code
- For job related queries:
  - Include job id or absolute paths to jobscript, error and output files
  - Avoid attachments; Screenshots are ok
- For additional allocations:
  - Compute – discuss with project chief investigator (CI) / scheme manager
  - Storage – gdata/massdata – discuss with CI / scheme manager
    scratch – discuss with NCI

NCI AUSTRALIA

Thank you !

NCI Training and Educational Events