Computing for Climate Modelling

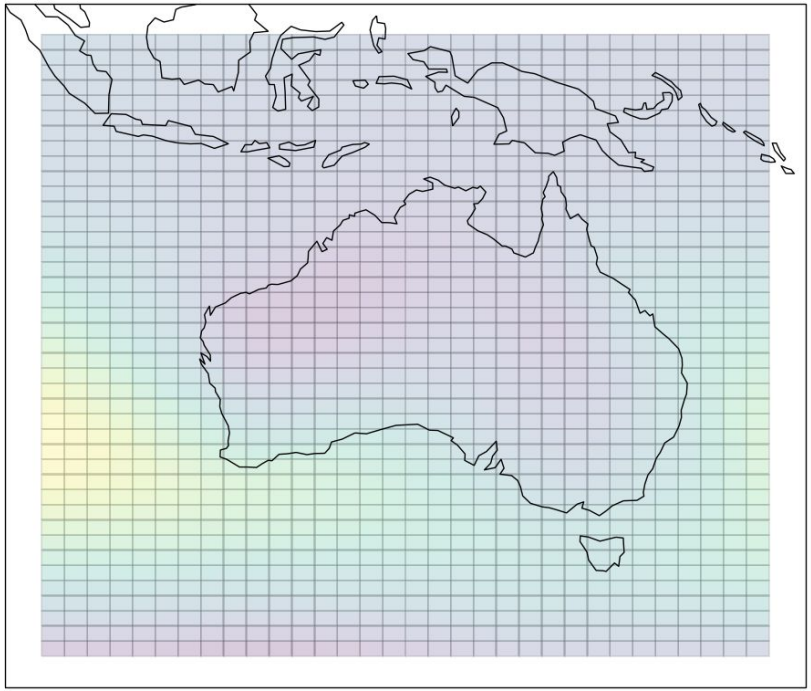
Scott Wales, CLEX CMS

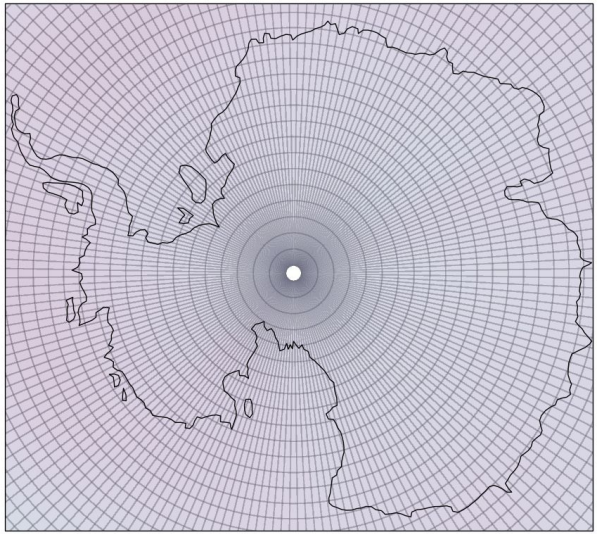
CLEX CMS | cws_help@nci.org.au

climate extremes
ARC centre of excellence

Australian Government
Australian Research Council

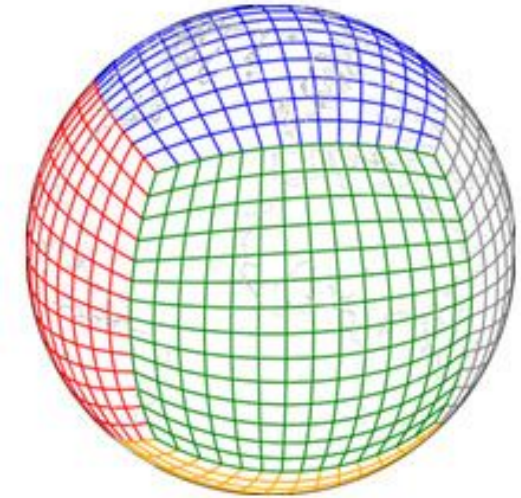
Discretisation





Alternate Grid Types

Cubed Sphere

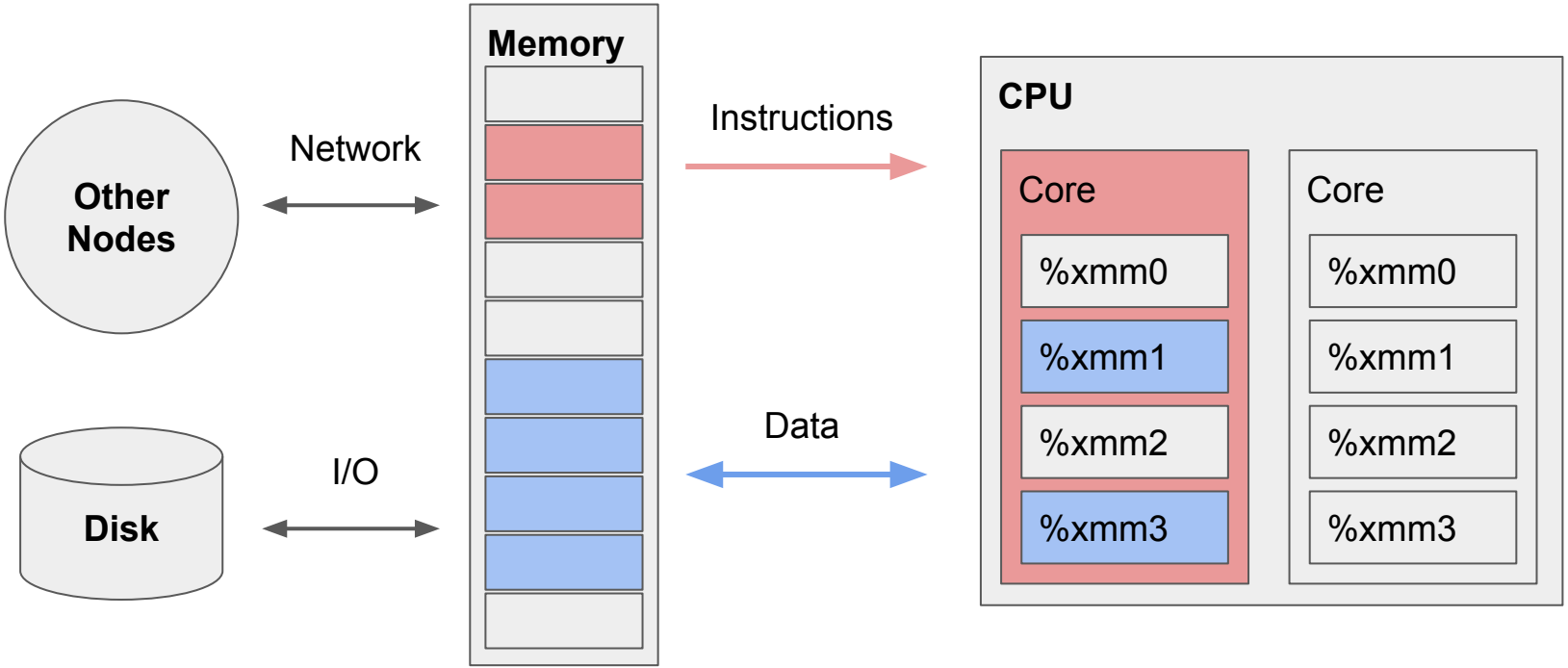


https://www.gfdl.noaa.gov/fv3/fv3-grids/

Voronoi Tessellation

https://mpas-dev.github.io/

Compute Node



Compiling

Fortran:

```
subroutine add(r, a, b)
    real, intent(in) :: a(4), b(4)
    real, intent(out) :: r(4)
    integer :: i

    do i=1,4
        r(i) = a(i) + b(i)
    end do
end subroutine
```

Assembly:

```
add_:
# parameter 1: %rdi
# parameter 2: %rsi
# parameter 3: %rdx
        movups    (%rsi), %xmm1          #5.5
        movups    (%rdx), %xmm0          #5.5
        addps     %xmm0, %xmm1           #5.5
        movups    %xmm1, (%rdi)          #5.5
        ret                              #6.1
```

Memory Layout

Model Field

1,1	1,2	1,3	1,4
2,1	2,2	2,3	2,4
3,1	3,2	3,3	3,4
4,1	4,2	4,3	4,4

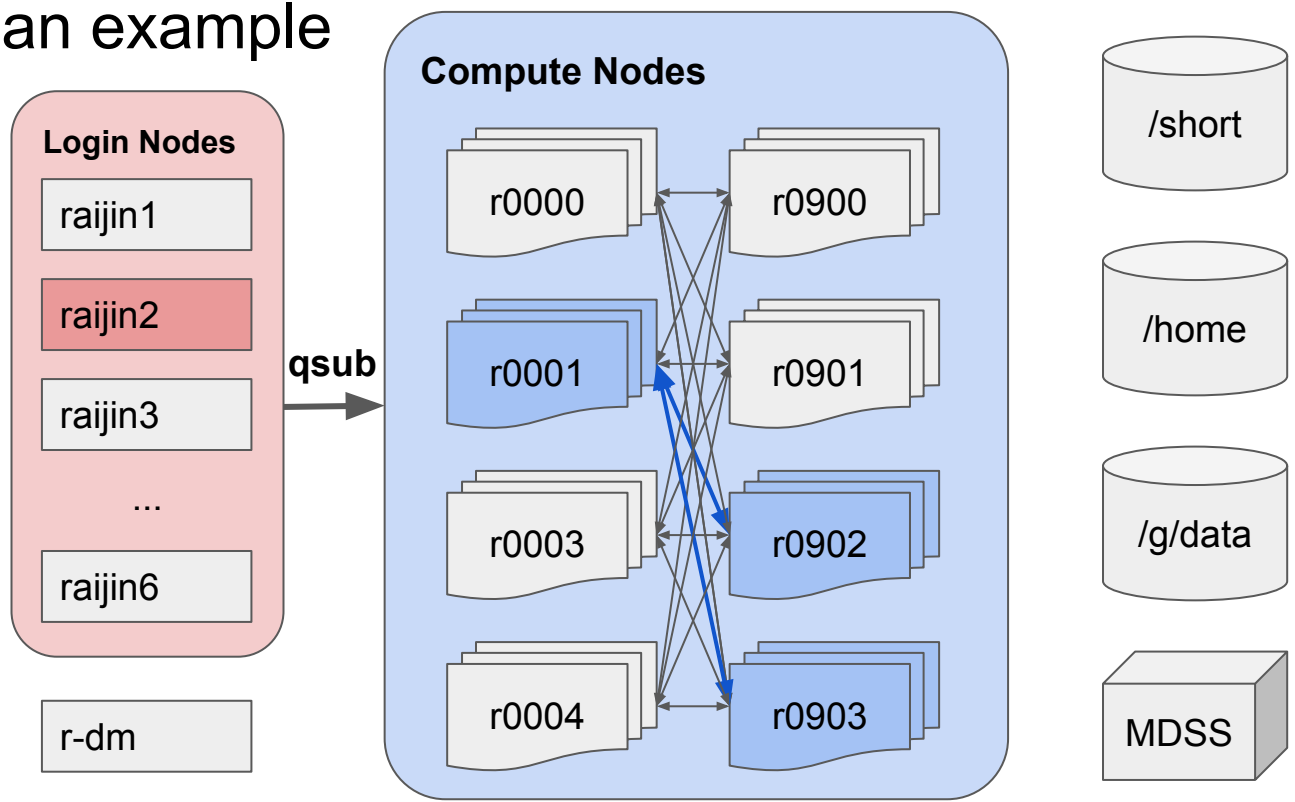
Computer Memory

Parallelisation

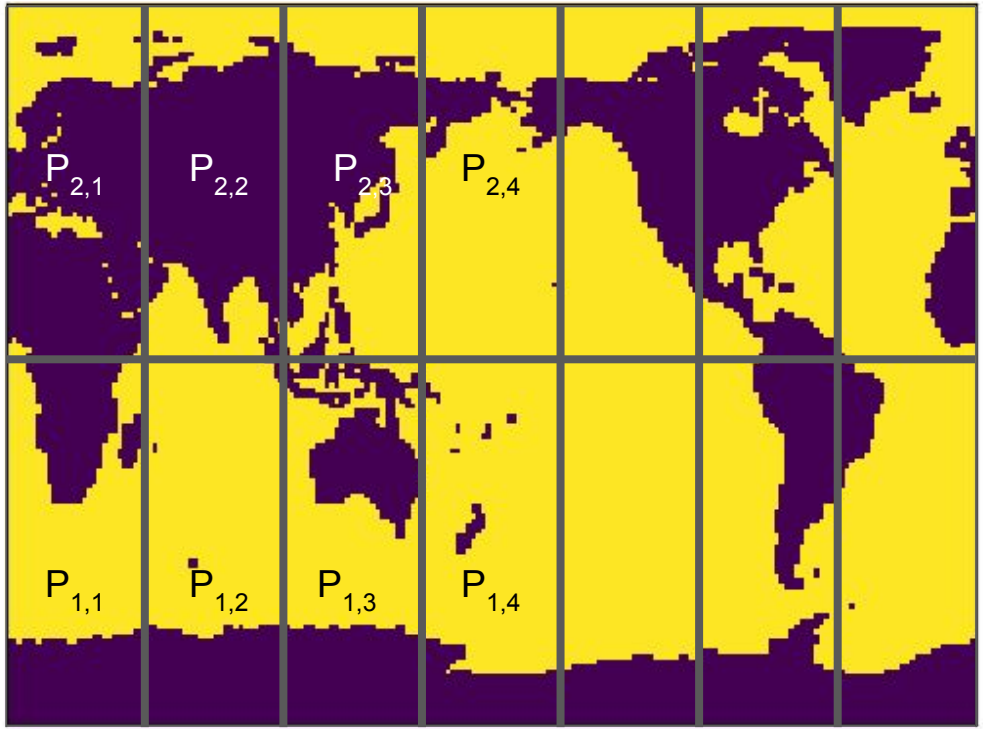
climate extremes
ARC centre of excellence
Australian Government
Australian Research Council

Supercomputers architecture:

Raijin as an example

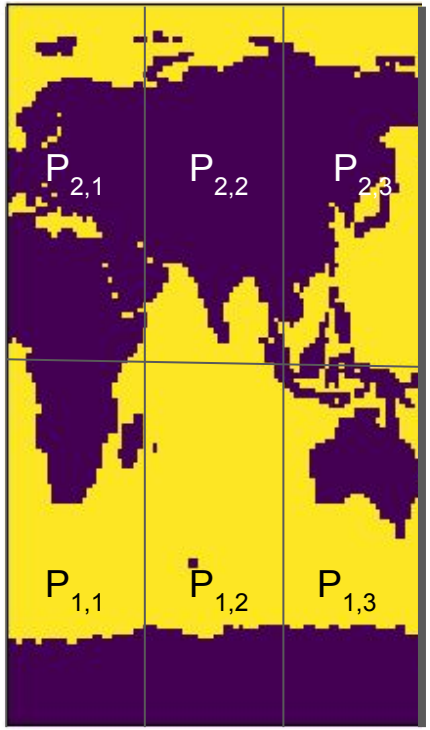


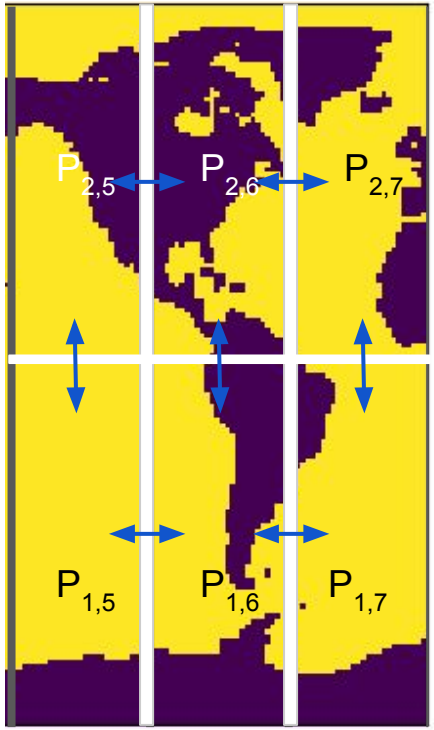
Decomposition



Parallelisation

Shared Memory





Distributed Memory

Shared Memory (OpenMP)

```
!$OMP PARALLEL DO SHARED(u, dudt)
do j=0, nj
  do i=0, ni
    du2dx2 = (u[i+1,j] - 2*u[i,j]
              + u[i-1,j])/deltax**s
    du2dy2 = (u[i,j+1] - 2*u[i,j]
              + u[i,j-1])/deltay**2

    dudt[i,j] = alpha*(du2dx2 +
                       du2dy2)
  end do
end do
!$OMP END PARALLEL DO
```

Each processor has access to the same memory

Specific sections are marked as parallel - e.g. loops

Variables can be either shared so all threads see the same data, or private so each thread has a different variable

Distributed Memory (MPI)

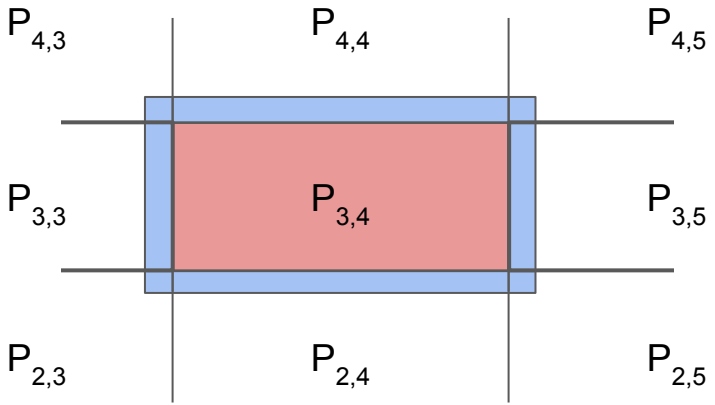
```
call halo_exchange(u, MPI_COMM_WORLD)

do j=local_j0, local_nj
  do i=local_i0, local_ni
    du2dx2[i,j] = (u[i+1,j] - 2*u[i,j]
                + u[i-1,j])/deltax**s
    du2dy2[i,j] = (u[i,j+1] - 2*u[i,j]
                + u[i,j-1])/deltay**2

    dudt[i,j] = alpha * (du2dx2[i,j]
                + du2dy2[i,j])
  end do
end do
```

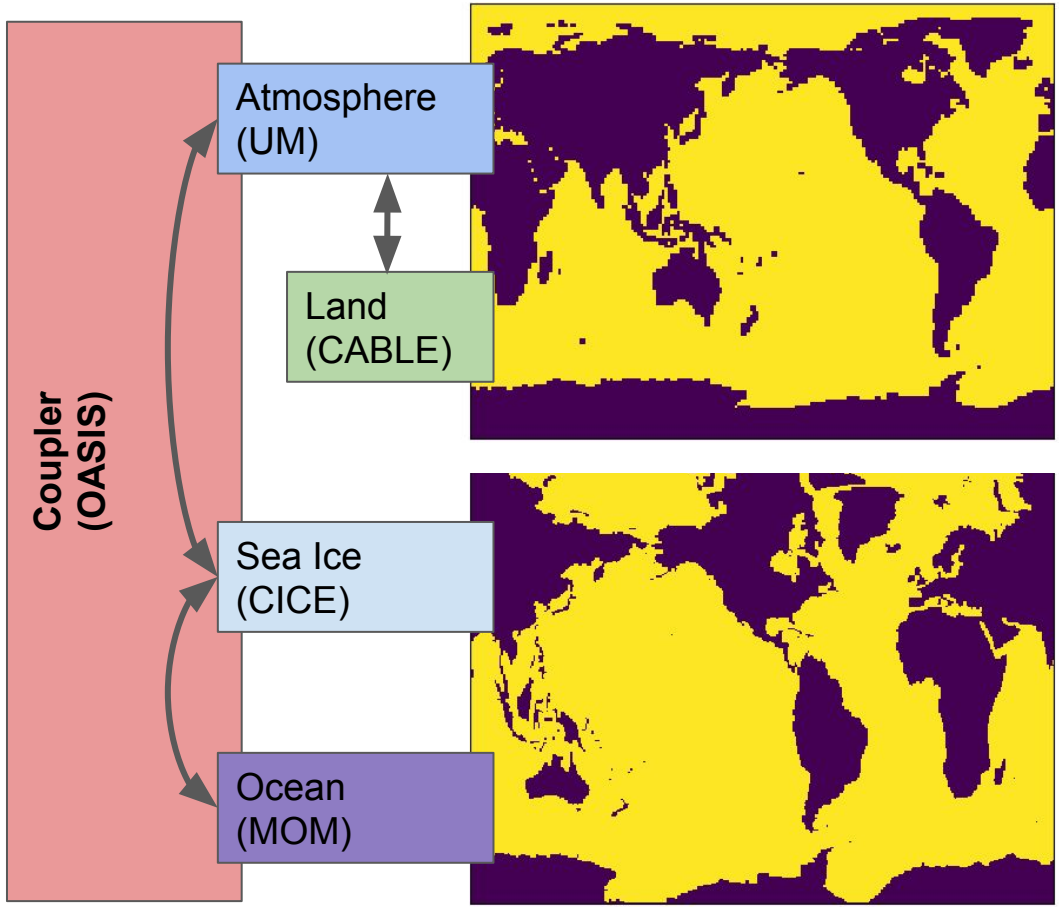
Each processor's data is kept separate

Neighbouring data is kept as a "halo" around each field



climate extremes
ARC centre of excellence
Australian Government
Australian Research Council

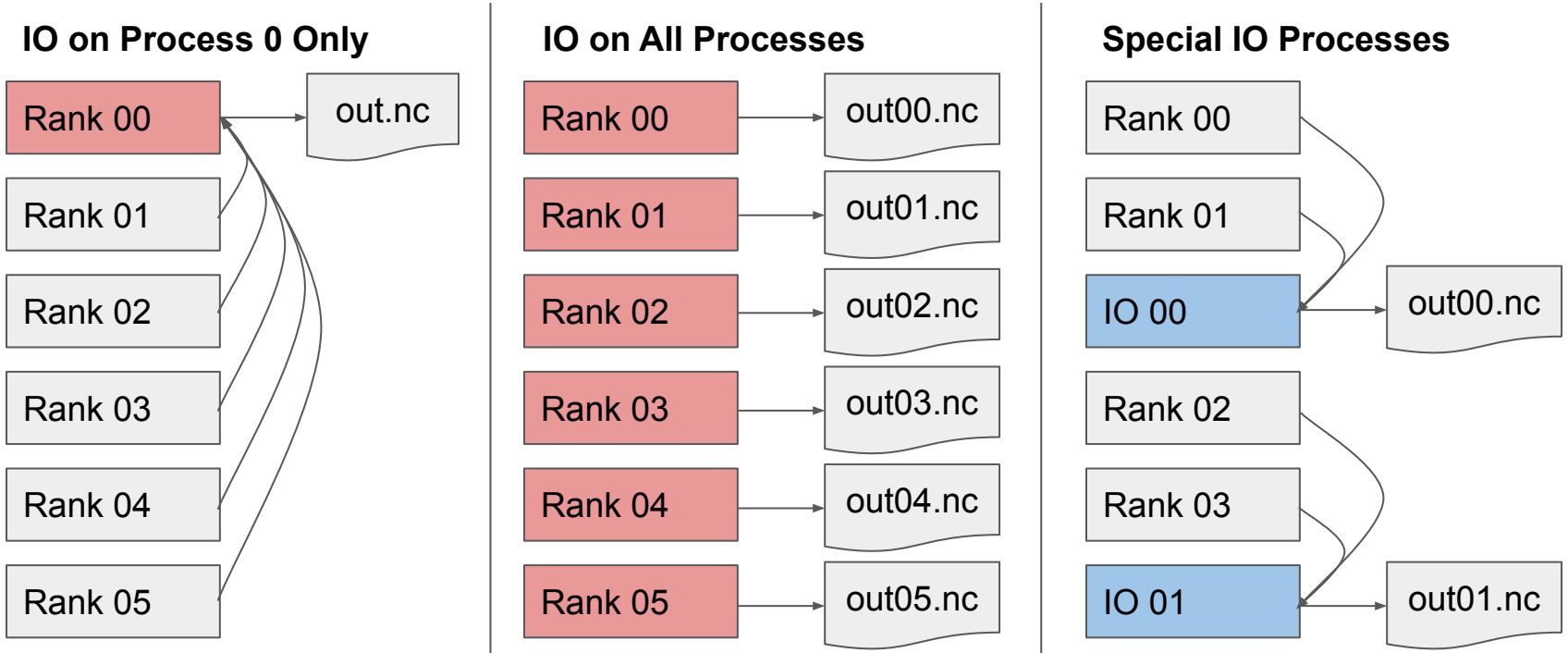
Coupling


Coupler
(OASIS)
Atmosphere
(UM)
Land
(CABLE)
Sea Ice
(CICE)
Ocean
(MOM)
Atmosphere
Grid
Ocean
Grid

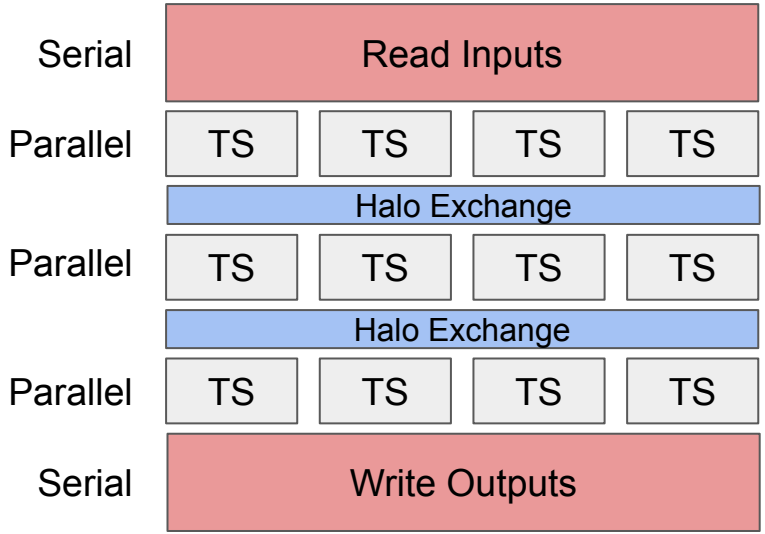
Optimisation

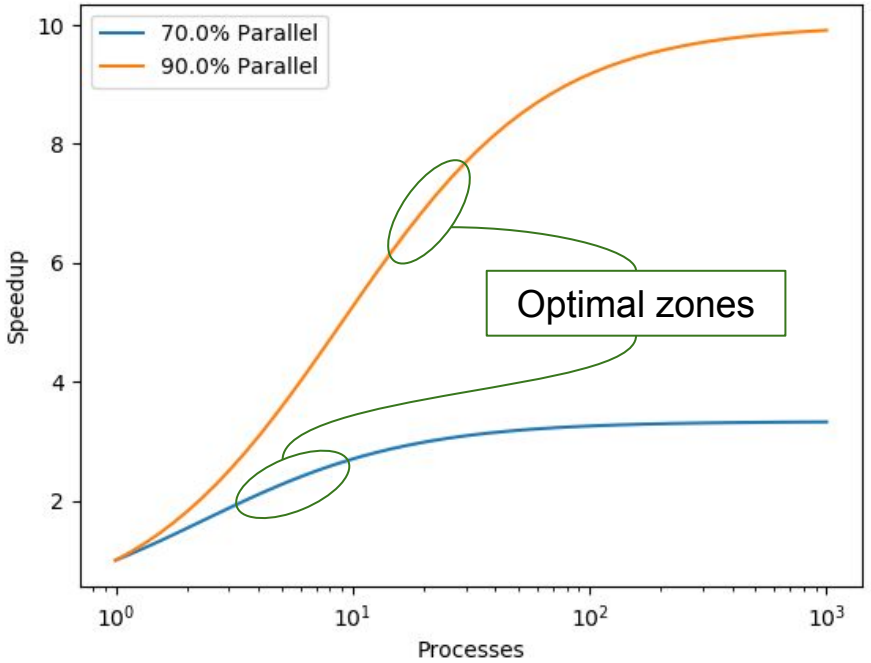
climate extremes
ARC centre of excellence
Australian Government
Australian Research Council

IO Strategies



Scaling

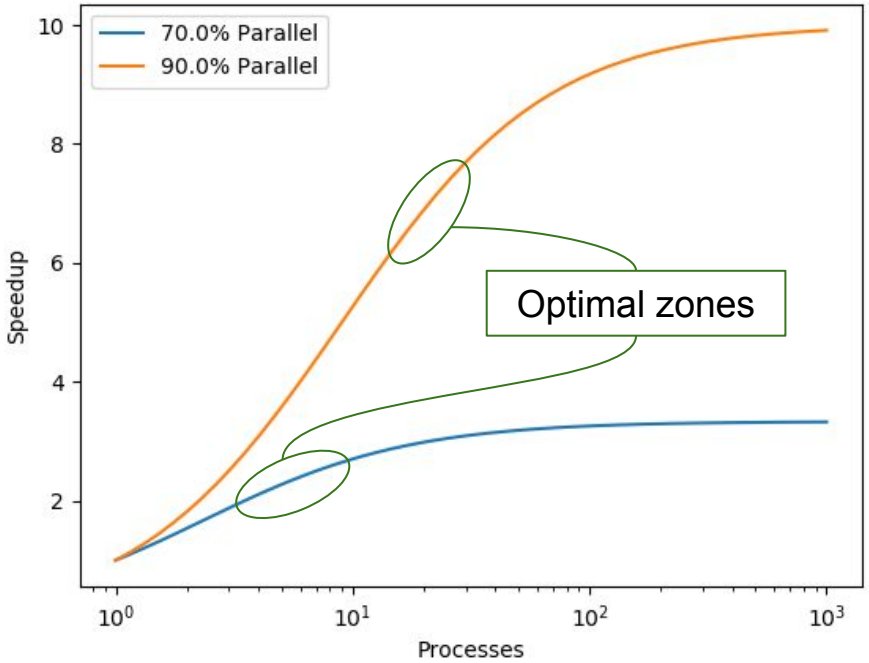




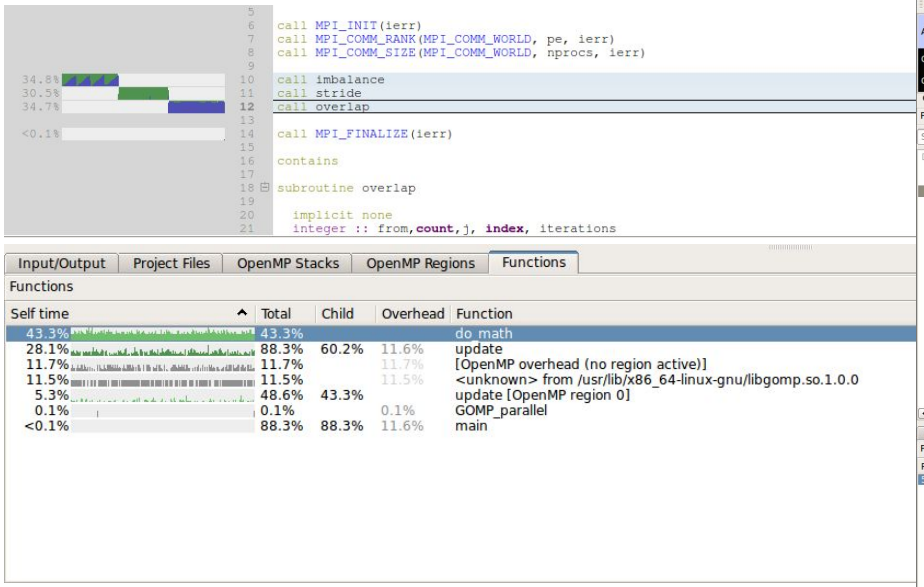
Optimising Efficiency

Experiment with short model runs

Optimise **(SU cost x walltime)** for different numbers of CPUs to find the best efficiency for your experiment

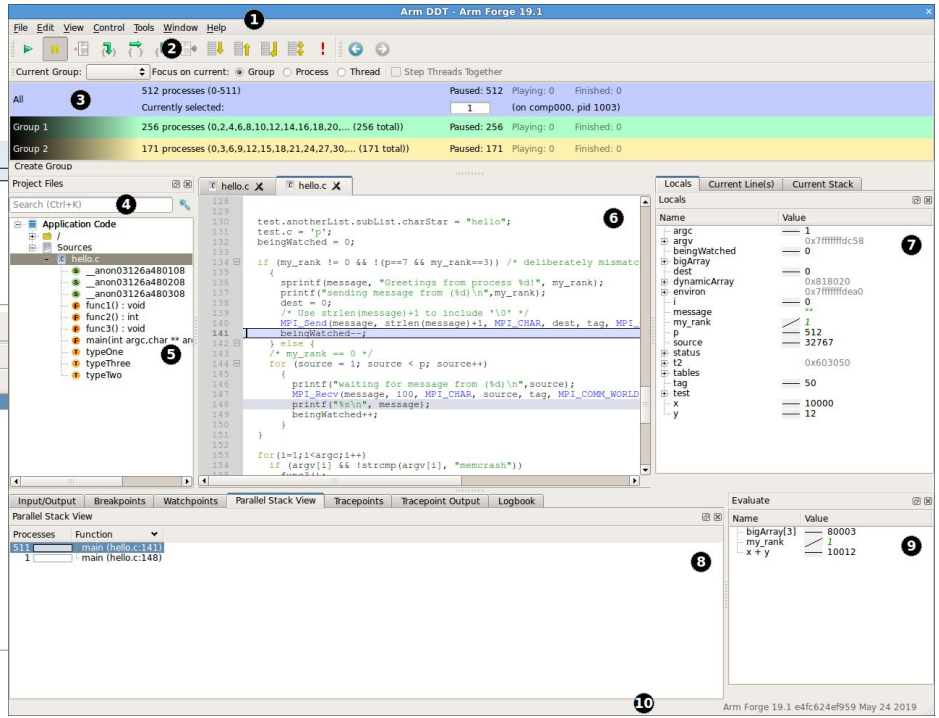


Profilers…



https://www.arm.com

and Debuggers



Model Output

climate extremes
ARC centre of excellence

Australian Government
Australian Research Council

Data Publishing

FAIR Principles

Findable - DOI, metadata

Accessible - Standard formats

Interoperable - Common vocabularies

Reusable - Provenance, licence

NCI Data Catalogue

https://geonetwork.nci.org.au

Research Data Australia Catalogue

https://researchdata.ands.org.au/

Australian Government
Australian Research Council

NetCDF

```
netcdf
tas_Amon_ACCESS1-0_historical_r1i1p1_185001-200512 {
 dimensions:
    time = UNLIMITED ; // (1872 currently)
    lat = 145 ;
    lon = 192 ;
variables:
   double time(time) ;
          time:units = "days since 0001-01-01" ;
          time:calendar = "proleptic_gregorian" ;
          time:standard_name = "time" ;
    float tas(time, lat, lon) ;
          tas:standard_name = "air_temperature" ;
          tas:long_name = "Near-Surface Air Temperature" ;
          tas:units = "K" ;
          tas:cell_methods = "time: mean" ;
          tas:cell_measures = "area: areacella" ;
// global attributes:
          :source = "ACCESS1-0 2011." ;
```

Climate and Forecasting Conventions
cfconventions.org

ncdump -h - Show metadata

cdo, nco - Stats, climate metrics, regridding

libnetcdf, libnetcdff - C and Fortran libraries

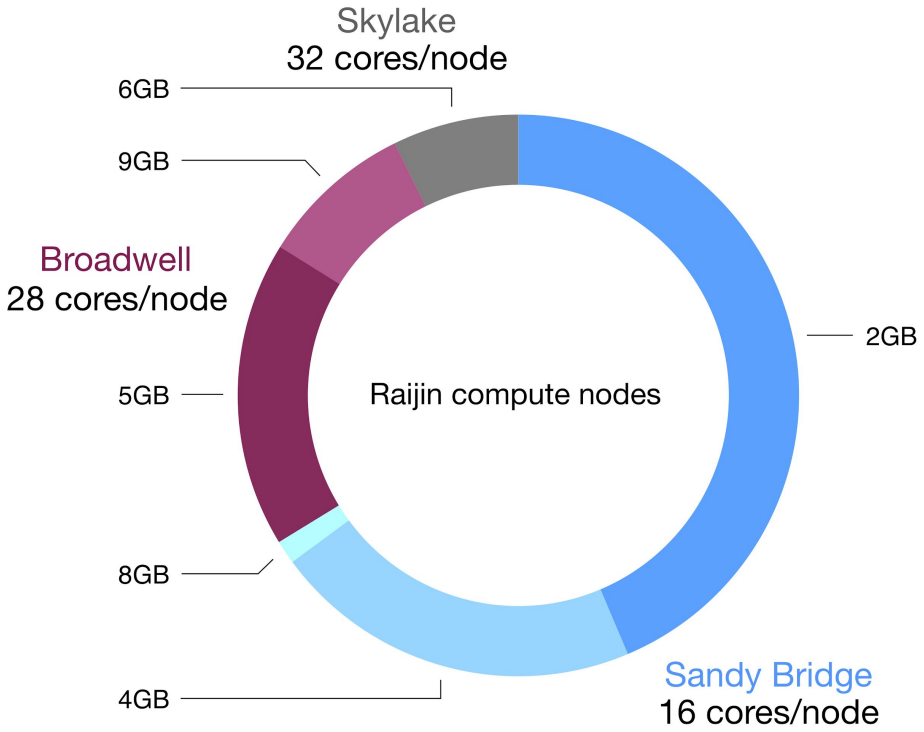
xarray, iris - metadata-aware Python libraries

Files can be read over the internet with THREDDS/OpenDAP

climate extremes
ARC centre of excellence
Australian Government
Australian Research Council

NCI Raijin

Raijin Nodes



Special Nodes:

- **gpu**: 32 nodes with GPUs
- **hugemem**: 10 nodes with 1TB memory
- **knl**: 32 nodes with Knights Landing CPUs

Tenjin (Openstack Cloud)

- 100 16-core nodes
- Access to /g/data filesystems

climate extremes
ARC centre of excellence
Australian Government
Australian Research Council

Submitting jobs: PBSPro Scheduler

```
#!/bin/bash
#PBS -q normal
#PBS -l cpus=16
#PBS -l mem=32gb
#PBS -l walltime=01:00:00
#PBS -l wd

module load openmpi

mpirun ./wrf.exe
```

`qsub` - Submit a job

`qstat -U $USER` - See your jobs

`nqstat -P $PROJECT` - Project jobs

qdel $JOBID - Stop a job

qcat $JOBID - See job output

MPI programs get the number of processes from the queue system

Queues

normal - 1 SU / cpu / wall hour
normalbw - 1.25 SU / cpu / wall hour
normalsl - 1.5 SU / cpu / wall hour

express - 3 SU / cpu / wall hour
expressbw - 3.75 SU / cpu / wall hour

copyq - 1 SU / cpu / wall hour
Max of 1 cpu per job
Access to internet for file downloads/ uploads

hugemem - 1.25 SU / cpu / wall hour
gpu - 3 SU / cpu / wall hour
knl - 0.25 SU / cpu / wall hour

Storage

/home
- Backed up
- 2 GB / user

/short/$PROJECT
- Shared per-project quota
- Short-term storage
- /short/public for sharing files

/g/data/$PROJECT
- Shared per-project quota
- Medium-term storage
- Cloud access

MDSS (Tape storage)
- Shared per-project quota
- Slow access
- Long-term archival
- Redundant (2 sites)
- **mdss** command to access

$PBS_JOBFS (Node storage)
- Storage directly on the node
- Disappears after a job completes

climate extremes
ARC centre of excellence
Australian Government
Australian Research Council

Central Dataset Storage

Reanalyses

- **ERA5, ERA-Interim**
- **JRA55**
- **OSTIA-SST**

Intercomparisons

- **CMIP5**
- **CMIP6**

Rainfall

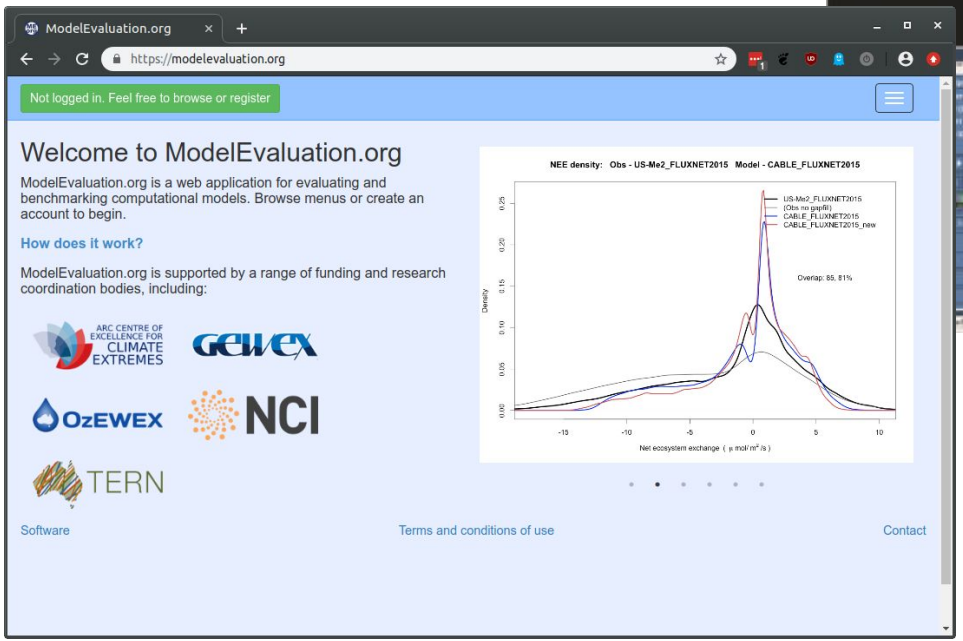
- **TRMM**
- **AWAP/AGCD**
- **Open Radar**

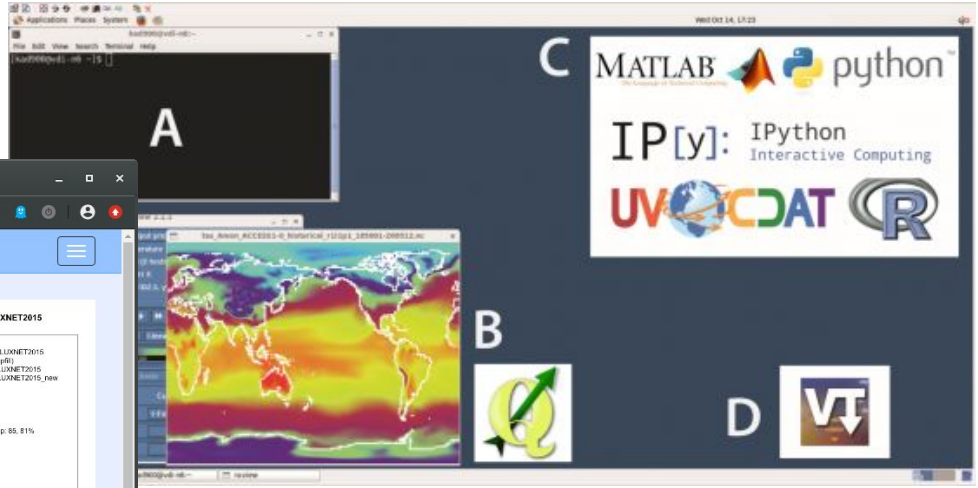
Centrally downloaded / published datasets are outside of the CLEX storage quota

You can request useful datasets by emailing us at **cws_help@nci.org.au**

http://climate-cms.wikis.unsw.edu.au/Category:Dataset

Cloud Computing





CLEX Resources at NCI

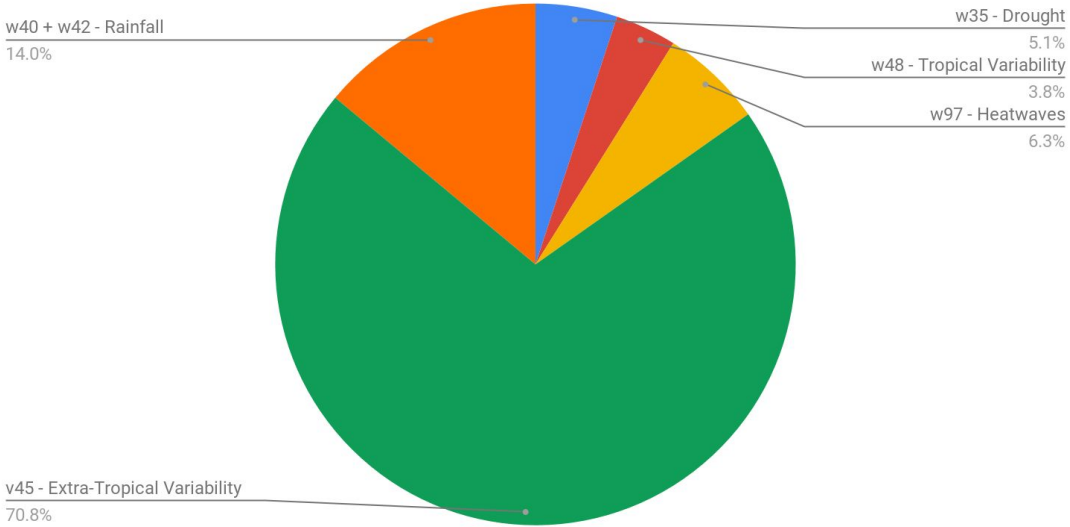
climate extremes
ARC centre of excellence

Australian Government
Australian Research Council

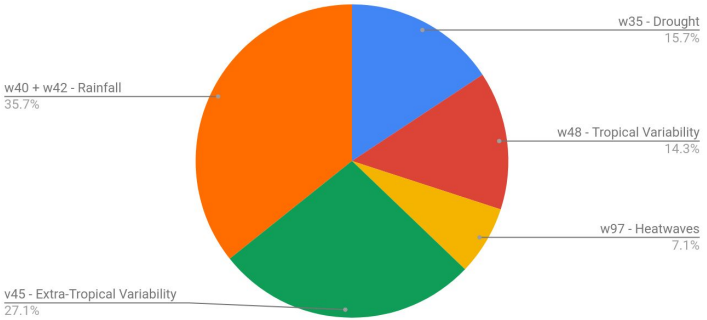
CLEX Resources

Shared between projects: v45, w35, w40, w42, w48, w97

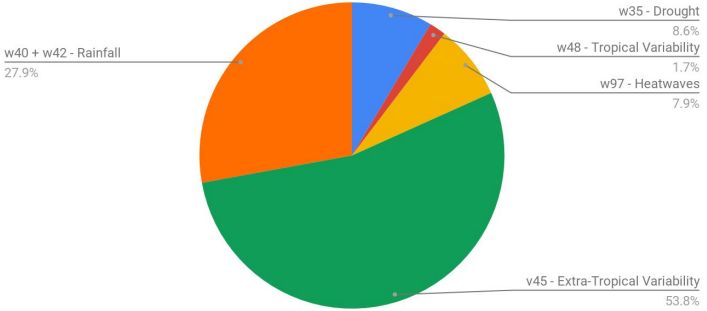
SU - Approx 3.1 MSU / quarter



/g/data Disk - 0.42 PB



MDSS Tape - 0.75 PB



CLEX CMS | cws_help@nci.org.au

climate extremes
ARC centre of excellence

Australian Government
Australian Research Council

NCI Projects

w35 - Drought

w40 - Rainfall (Lane)

w42 - Rainfall (Sherwood)

v45 - Extra-Tropical Variability

w48 - Tropical Variability

w97 - Heatwaves

Join projects at

https://my.nci.org.au

ua8 - Publication Storage
hh5 - LIEF Storage
access - ACCESS model input data

+ projects for specific datasets

http://climate-cms.wikis.unsw.edu.au/Category:Dataset

Managing Allocations

nci_account - Current allocation and usage

short_files_report - Project disk usage by user
gdata1_files_report, …

lquota - Storage limits

nf_limits - Maximum queue resources

CMS Conda environment tools

```
module use /g/data3/hh5/public/modules
module load conda/analysis3
```

ncimonitor - Graphs of per-project compute and data usage

nccompress - Compress NetCDF files under a directory

More Info

CLEX CMS

wiki	- climate-cms.wikis.unsw.edu.au
training	- weekly over VC
email	- cws_help@nci.org.au
chat	- arccss.slack.com

NCI

web	- nci.org.au
docs	- opus.nci.org.au
email	- help@nci.org.au






climate extremes
ARC centre of excellence

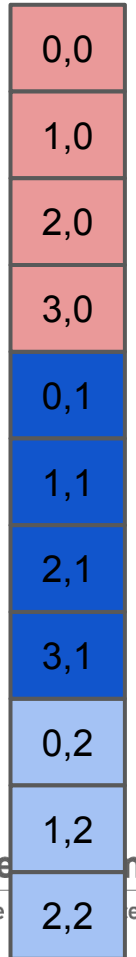
Australian Government
Australian Research Council

Memory Layout

Model Field

0,0	0,1	0,2	0,3
1,0	1,1	1,2	1,3
2,0	2,1	2,2	2,3
3,0	3,1	3,2	3,3

Computer Memory



Raijin Nodes

Sandy Bridge (**normal**/**express**):
- 3503 nodes
- 16 cpus/node
- 32, 64, 128 GB memory/node (16:8:1)

Broadwell (**normalbw**/**expressbw**):
- 814 nodes
- 28 cpus/node
- 128, 256 GB memory/node (2:1)

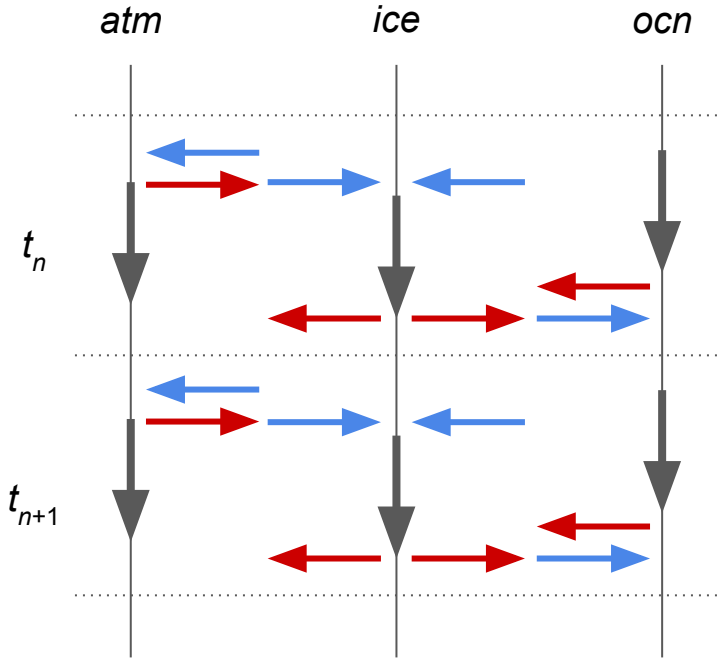
Skylake (**normalsl**):
- 192 nodes
- 32 cpus/node
- 192 GB memory/node

Special Nodes:
- **gpu**: 32 nodes with GPUs
- **hugemem**: 10 nodes with 1TB memory
- **knl**: 32 nodes with Knights Landing CPUs

Tenjin (Openstack Cloud)
- 100 16-core nodes
- Access to /g/data filesystems

climate extremes
ARC centre of excellence
Australian Government
Australian Research Council

Coupling



Coupling (test other slide format)