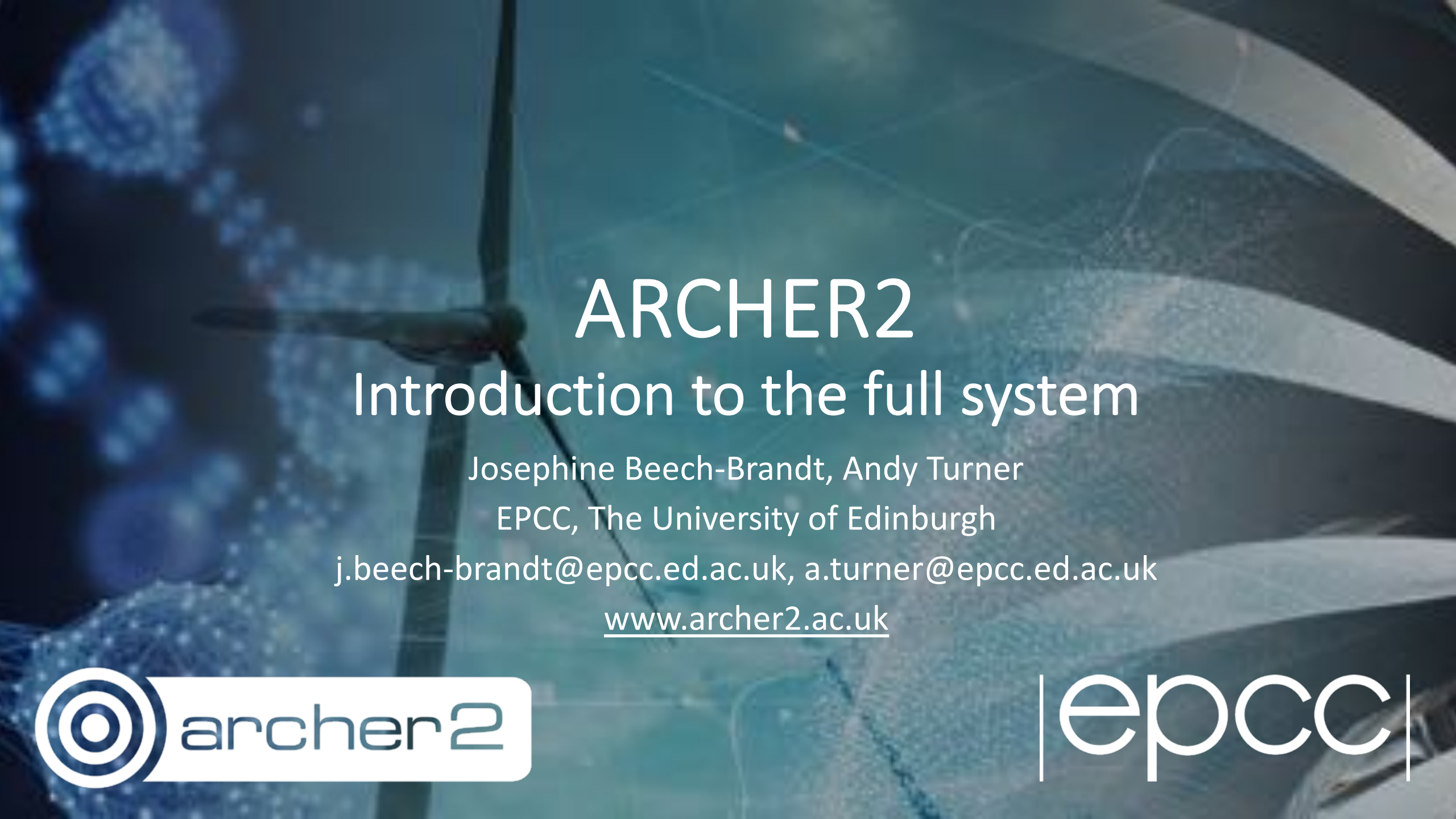


ARCHER2

Introduction to the full system

Josephine Beech-Brandt, Andy Turner
EPCC, The University of Edinburgh
j.beech-brandt@epcc.ed.ac.uk, a.turner@epcc.ed.ac.uk

www.archer2.ac.uk

archer2

epcc

Reusing this material

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

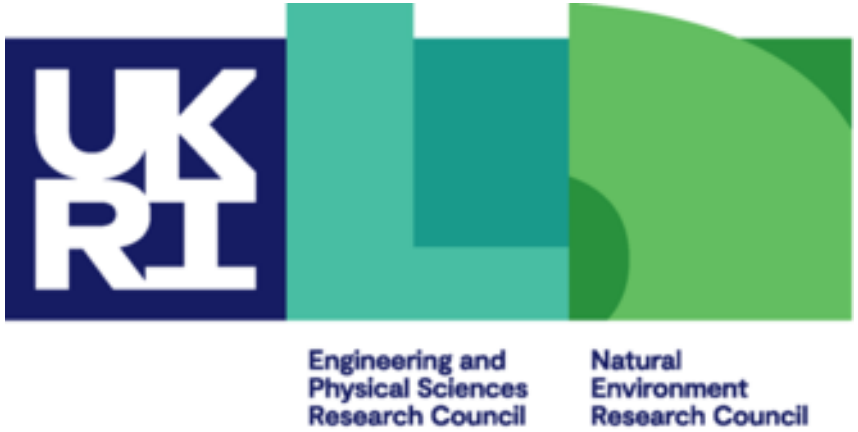
https://creativecommons.org/licenses/by-nc-sa/4.0/

This means you are free to copy and redistribute the material and adapt and build on the material under the following terms: You must give appropriate credit, provide a link to the license and indicate if changes were made. If you adapt or build on the material you must distribute your work under the same license as the original.

Note that this presentation contains images owned by others. Please seek their permission before reusing these images.

Partners

epcc



epcc

THE UNIVERSITY *of* EDINBURGH

Hewlett Packard Enterprise

Overview

- Onboarding process for users and projects
  - Projects and Allocations
  - Accounts and access
  - File systems and data migration
- Differences from 4-cabinet ARCHER2 system
  - Hardware and software
  - Slurm scheduler layout

20	Piz Daint - Cray XC50, interconnect , NVIDIA Swiss National Supers Switzerland					
21	Trinity - Cray XC40, Xe Xeon Phi 7250 68C 1.4 DOE/NNSA/LANL/SNL United States					
22	ARCHER2 - Cray XE, A Slingshot-10, HPE EPSRC/University of E United Kingdom					
23	SuperMUC-NG - Think 24C 3.1GHz, Intel Omn Leibniz Rechenzentru Germany					
24	Hawk - Apollo 9000, A Mellanox HDR Infiniba HLRS - Höchstleistung Germany					
25	Ghawar-1 - HPE Cray Slingshot-10, HPE Saudi Aramco Saudi Arabia					

11 entries found.

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
22	ARCHER2 - Cray XE, AMD EPYC 7742 64C 2.25GHz, Slingshot-10, HPE EPSRC/University of Edinburgh United Kingdom	716,800	19,539.0	25,804.0	
67	Cray XC40, Xeon E5-2695v4 18C 2.1GHz, Aries interconnect , HPE United Kingdom Meteorological Office United Kingdom	241,920	7,038.9	8,128.5	
95	DiRAC, Tursa - BullSequana XH2000, AMD EPYC 7302 32C 3GHz, NVIDIA A100 40GB, Mellanox HDR Infiniband, Atos University of Edinburgh United Kingdom	55,552	5,228.0	8,580.0	244
117	Cray XC40, Xeon E5-2695v4 18C 2.1GHz, Aries interconnect , HPE ECMWF United Kingdom	126,468	3,944.7	4,249.3	1,897
118	Cray XC40, Xeon E5-2695v4 18C 2.1GHz, Aries interconnect , HPE ECMWF United Kingdom	126,468	3,944.7	4,249.3	1,897

ARCHER2 in at number 22 in the Nov 2021 Top500 list

Largest current system in the UK

https://top500.org/

Transition to the Full System

Projects and Allocations

- All projects with an active allocation on 1st October on 4-cabinet system were transferred to the full system
  - Project codes are the same on both systems
  - The ARCHER2 Compute Unit (CU) = 1 ARCHER2 Node hour is used on both systems
  - There will be a period of at least 30 days where users have access to both systems. During this period:
    - Usage on full system will be uncharged
    - Usage on the 4-cabinet system will be charged
  - The uncharged period is scheduled to end on January 4th but this will be confirmed by UKRI

Accounts and access

- Login addresses:
  - ARCHER2 4-cabinet system: login-4c.archer2.ac.uk
  - ARCHER2 full system: login.archer2.ac.uk
- User Accounts:
  - Same username, password and ssh key as ARCHER2-4-cabinet

- IMPORTANT! Ordering of credentials is switched on full system
  - 4-cabinet system:
    1. machine account password
    2. passphrase for SSH key pair
  - Full system:
    1. passphrase for SSH key pair
    2. machine account password

Possible DNS Warning:

login.archer2.ac.uk now being used for the full system. You may see the following error when trying to log in for the first time

```
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@       WARNING: POSSIBLE DNS SPOOFING DETECTED!          @
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
The ECDSA host key for login.archer2.ac.uk has changed,
and the key for the corresponding IP address 193.62.216.43
has a different value. This could either mean that
DNS SPOOFING is happening or the IP address for the host
and its host key have changed at the same time.
Offending key for IP in /Users/username/.ssh/known_hosts:11
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@    WARNING: REMOTE HOST IDENTIFICATION HAS CHANGED!     @
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
IT IS POSSIBLE THAT SOMEONE IS DOING SOMETHING NASTY!
Someone could be eavesdropping on you right now (man-in-the-middle attack)!
It is also possible that a host key has just been changed.
The fingerprint for the ECDSA key sent by the remote host is
SHA256:UGS+LA8I46LqnD58WiYNlaUFY3uD1WFr+V85CB07fUg.
Please contact your system administrator.
```

To resolve, remove the entry from `.ssh/known_hosts` file

If you require assistance, please contact: support@archer2.ac.uk

File systems

- Home file systems
  - Available on both ARCHER2 4-cabinet and ARCHER2 full systems
  - No action required and home file systems will be readable and writeable on both services
- Work file systems
  - Different work file system for 4-cabinet and the full systems
  - Work file system on the 4-cabinet system will be available during transition period
  - New work file systems on full system will typically be 2x quota of work on 4-cabinet system
  - IMPORTANT! Users are responsible for transferring any required data from the 4-cabinet work file system to the full system work file systems
  - You have at at least 30 days to transfer the data before the 4-cabinet work file system is removed
- RDFaaS fiile systems
  - RDFaaS file systems (/epsrc and /general) will be available on both systems

Data Migration : work file systems

- Data transfer from ARCHER2 4-cabinet system to ARCHER2 full system
  - Requires accounts on both systems
  - Only requires one form of authentication – password will suffice (no key required)

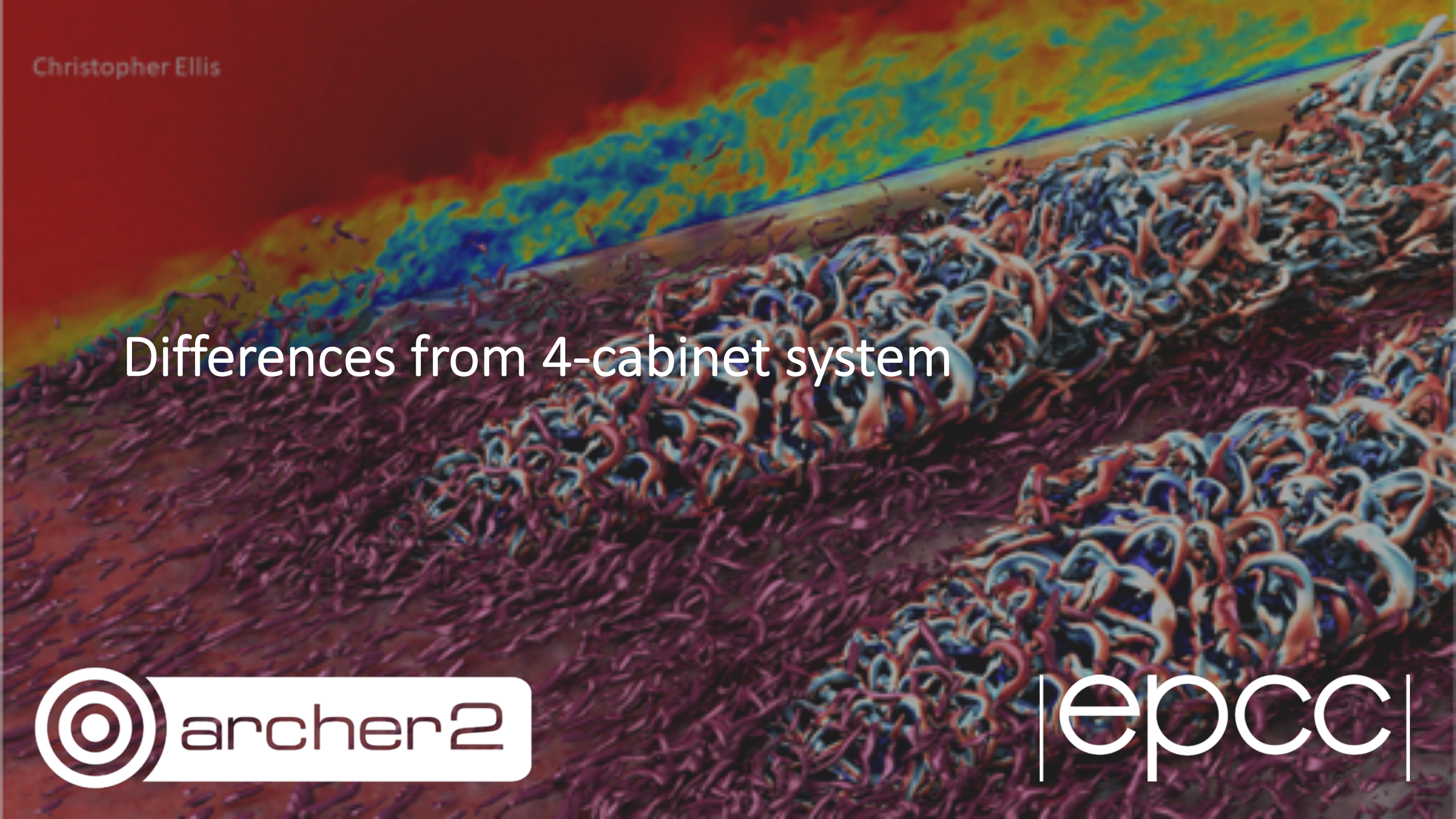
  - Examples of transferring data using `rync` or `scp` can be found at:

  https://docs.archer2.ac.uk/archer2-migration/data-migration/

Help

- Full instructions and information available at:
- https://docs.archer2.ac.uk/archer2-migration/

- If you need further assistance or have any questions:
- support@archer2.ac.uk



Differences from 4-cabinet system

ARCHER2 full system

- HPE Cray EX Supercomputer
- Hosted at EPCC, The University of Edinburgh
- 5,860 compute nodes (750,080 CPU compute cores)
- HPE Cray Slingshot interconnect
- Compute nodes:
  - Dual socket AMD EPYC 7742 (Rome), 64c, 2.25 GHz
  - 256 GIB / 512 GiB memory per node
  - Two 100 Gbps Slingshot interfaces per node
- 3x ClusterStor L300 Lustre file systems, each 3.6 PB
- 1 PB ClusterStor E1000F solid state storage
  - Not available at start of service
  - Will be available to users via Slurm BB/DW directives
- 4x NetApp FAS8200A file systems, 1 PB total

Comparison of systems

	ARCHER2 4-cabinet system	ARCHER2 full system
Compute nodes (Cores)	1,024 (131,072)	5,860 (750,080)
High memory nodes	✗	✓ (at least 292 nodes with 512 GiB)
Login nodes	2	4
Parallel file systems	1 (3.6 PB total)	4 (14.4 PB total, 3 at start of service)
Solid state file system	✗	✓ (not at start of service)
Job scheduler	Slurm	Slurm
Data analysis nodes (serial)	✗	✓
Module system	TCL Environment Modules v4	Lmod
Default programming environment	20.10	21.04
Compilers (default versions) [others]	Cray (10.0.4) [10.0.3, 11.0.3] Gnu (10.1.0) [9.3.0, 10.2.0] AMD (2.1.0.3) [2.2.0.1]	Cray (11.0.4) [12.0.3] Gnu (10.2.0) [9.3.0, 10.3.0, 11.2.0] AMD (2.2.0.1) [2.2.0, 3.0.0]
Singularity containers	✓	✓
RDF file system mount	✓	✓

Key differences: general system

- 5860 compute nodes on the full system
- 584 high memory nodes (512 GiB memory)
- Data analysis (serial) nodes available
- Modules provided by Lmod
- Some older versions of software/libraries are not available
- Scheduler layout expanded
- No need for `--reservation=shortqos` when using the short QoS
- No need for `module load epcc-job-env` in job scripts

Key differences: programming environment

- Default version is 21.04 (20.10 on 4-cab) with 21.09 available
  - Means that default compiler and library versions are newer
- Some software library modules are hidden by default due to dependencies
  - E.g. NetCDF modules, need appropriate HDF5 module loaded first
- Two MPI transport layers, OpenFabrics (OFI) and UCX are available and supported on the system
  - The default is OFI
  - Switching to UCX can have a positive impact on performance (particularly if performance depends on MPI collectives)
  - You do not need to recompile to switch to UCX, just use:

```
module load craype-network-ucx
module swap cray-mpich cray-mpich-ucx
```

Software modules: Lmod

- Lmod is used on the full system
  - TCL Environment Modules v4 is used on the 4-cabinet system
  - Many commands are the same: `module load`, `module swap`, `module avail`, `module remove`
  - Some new commands: `module spider`, `module restore`
  - Commands no longer available: `module restore PrgEnv-*`
- Now switch to different compiler environments using module load
  - e.g. `module load PrgEnv-gnu` (**not** `restore` as on 4-cabinet system)
- Some modules not visible to module avail until dependent modules loaded
  - e.g. cray-netcdf, cray-netcdf-hdf5parallel
  - Use `module spider` to query all modules, including those that are not currently visible
- https://docs.archer2.ac.uk/user-guide/sw-environment/

MPI: OFI and UCX transport layers

- Openfabrics (OFI) and UCX transport layers available to use with MPI
  - (These are both available on 4-cab but use of UCX was not generally recommended.)
  - Default is OFI, UCX may give better performance/scaling in some cases
- You can switch to UCX without needing to recompile:
  - E.g. for switching to UCX for an application compiled with GCC:

```
module load PrgEnv-gnu
module load craype-network-ucx
module swap cray-mpich cray-mpich-ucx
```

- It is always useful to add the ldd executable command to your job submission script to check that the executable is really using UCX versions if you want to switch
- We will provide specific advice on applications as we gain experience with tests
- https://docs.archer2.ac.uk/user-guide/dev-environment/#message-passing-interface-mpi

Slurm layout

- More functionality and flexibility than the 4-cabinet system
- New additions:
  - highmem: Access to high memory nodes
  - taskfarm: Ability to run larger number of smaller jobs
  - serial: Access to data analysis nodes
- Changes:
  - short: supports jobs up to 32 nodes (8 node limit on 4-cabinet)
  - standard: supports jobs up to 2048 nodes (256 node limit on 4-cabinet)
  - largescale: supports jobs up to 5860 nodes (940 node limit on 4-cabinet)
- Ability to run from 1 node (128 cores) to 5860 nodes (750,080 cores)
- Reservations can allow for jobs that exceed standard limits
- https://docs.archer2.ac.uk/user-guide/scheduler/

Scheduler: MPI jobs

```
#!/bin/bash
#SBATCH --job-name=Example_MPI_Job
#SBATCH --time=0:20:0
#SBATCH --nodes=4
#SBATCH --tasks-per-node=128
#SBATCH --cpus-per-task=1
#SBATCH --account=t01

export OMP_NUM_THREADS=1

srun –hint=nomultithread –distribution=block:block my_executable.x
```

Select 4 nodes and 128 (MPI) processes per node

Set OpenMP threads to 1 to prevent libraries from using threading by default

srun uses the distribution from the job options to launch the correct number of MPI processes and place them on the correct nodes and pin to the correct cores

No longer need to load the “epcc-job-env” module in job scripts

Data analysis nodes

- 2 nodes:
  - Each with 128 cores (same processors as compute nodes)
  - Each with 512 GB memory
- Shared access: multiple users using nodes at the same time
- Request cores and/or memory
  - Maximum of 32 cores and/or 128 GB memory per user
  - Allocated 2 GB memory per core if you do not specify memory
  - Allocated 1 core if you do not specify number of cores
- All file systems available (home, work, RDFaaS)
- External network access for data transfer
- Jobs are uncharged (but you need a valid budget to run)

Scheduler: Data analysis nodes

```
#!/bin/bash
#SBATCH --job-name=Example_Serial_Job
#SBATCH --time=0:20:0
#SBATCH --ntasks=1
#SBATCH --mem=4G
#SBATCH --account=t01

my_executable.x
```

Select 1 process and 4 GB memory

Or, for an interactive shell on the compute nodes with graphical output via X:

```
srun --time=00:20:00 --partition=serial --qos=serial \
        --account=t01 --ntasks=1 --mem=4G --x11 --pty /bin/bash
```

(This assumes you have logged into ARCHER2 with X enabled (-X or –Y option) and that you have a local X client installed in your local system.)

Future plans

- Addition of solid state storage
- Continuing optimisation of Slurm configuration based on production use statistics
- Evaluation of comparative performance of different compilers
- Evaluation of comparative performance of OFI vs UCX
- Evaluation of Singularity containers
  - HPE Programming Environment containers
  - Generic MPICH containers

Summary

- User access for full system planned for w/b 22 Nov
  - Jobs will be uncharged for at least 30 days following full system access
- 4-cabinet system will be available for at least 30 days following full system access
- Current 4-cab username, password and SSH keys used to access system
- Home file systems shared across both systems
- Separate work file systems on both systems
  - Users responsible for copying any data need across to full system
- RDFaaS file systems shared across both systems

https://docs.archer2.ac.uk/archer2-migration/