ARCHER2 TDS Experiences

Andy Turner, EPCC, The University of Edinburgh
a.turner@epcc.ed.ac.uk

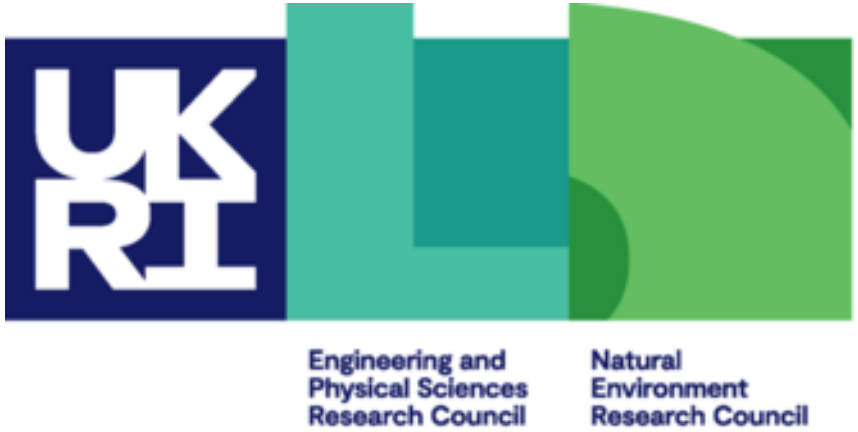
2 September 2020

www.archer2.ac.uk

Reusing this material

Partners

epcc



epcc

THE UNIVERSITY of EDINBURGH

CRAY®

a Hewlett Packard Enterprise company

Timescales and progress

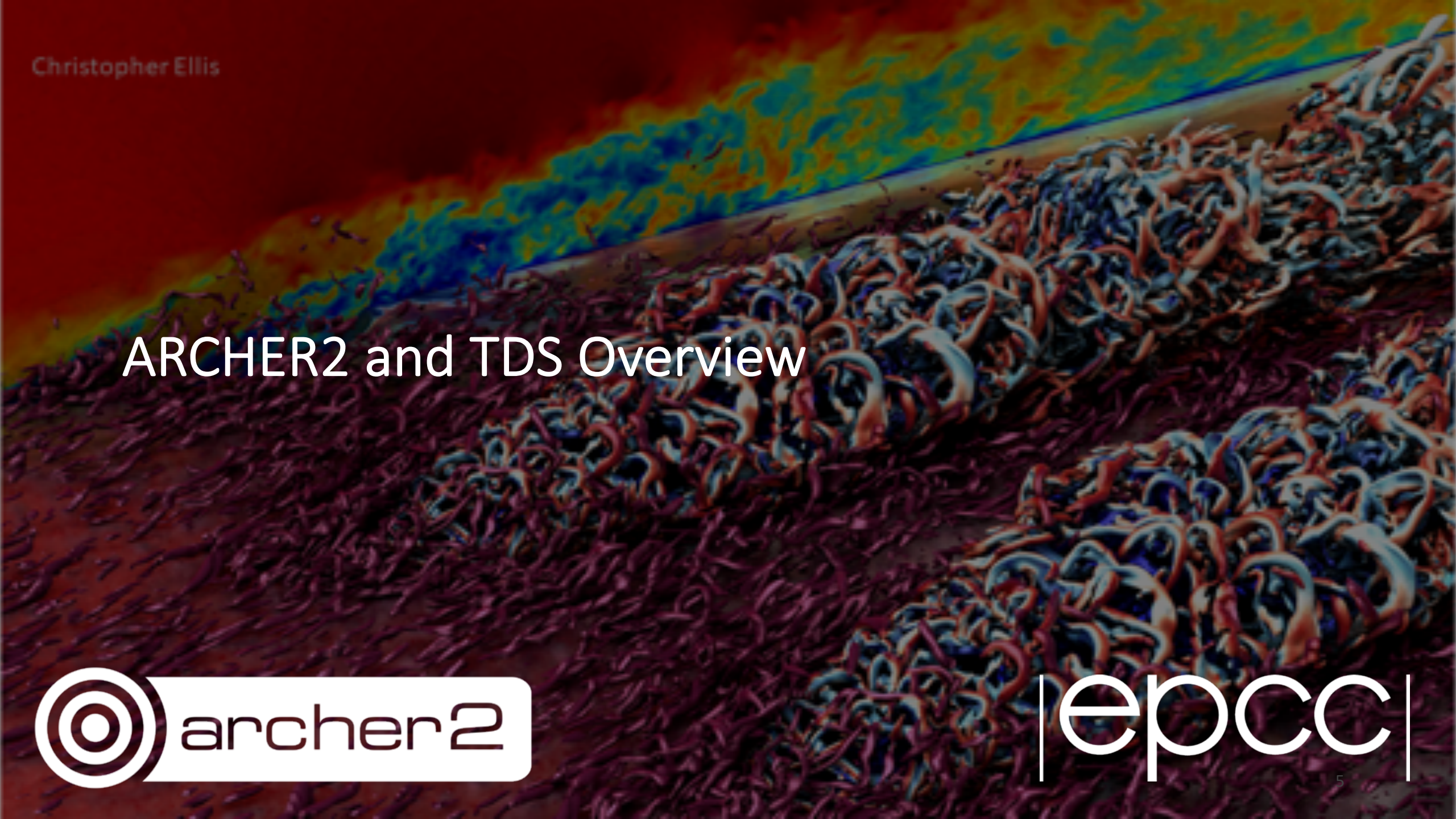
4 cabinet ARCHER2

Date	Item
25 June 2020	4 cabinet system passes factory acceptance test
13 July 2020	4 cabinet ARCHER2 system arrives at EPCC ACF
Mid-September 2020	4 cabinet ARCHER2 system handed to EPCC for commissioning
Early-October 2020	4 cabinet ARCHER2 early access starts
End-October 2020	4 cabinet ARCHER2 user service starts

Future Work

Date	Item
Q4 2020	ARCHER decommission and removal
Q4 2020	19 cabinet ARCHER2 arrives at EPCC ACF
Q1 2021	19 cabinet ARCHER2 user service starts
Q1 2021	Merge of 19 and 4 cabinet ARCHER2 systems to full 23 cabinet ARCHER2

Note: future timescales subject to change depending on progress



ARCHER2 and TDS Overview

ARCHER2 Full System Overview

- HPE Cray EX system
  - (formerly Shasta)
- Peak performance: ~28 PFlops
- 5,848 compute nodes
  - 748,544 AMD cores
- 14.5 PB Lustre (4 file systems)
- 1.1 PB solid state burst buffer
- 1 PB home (backed up) storage
- HPE Cray Slingshot interconnect

Compute Nodes: AMD EPYC 7742

- Dual socket AMD EPYC (Rome) 7742, 2.25GHz, 64 core
  - 128 cores per node (256 SMT)
- 5556 Standard Nodes: 256 GiB
- 292 High Memory Nodes: 512 GiB
- 2 Slingshot interfaces per node
  - 1 per socket

TDS System

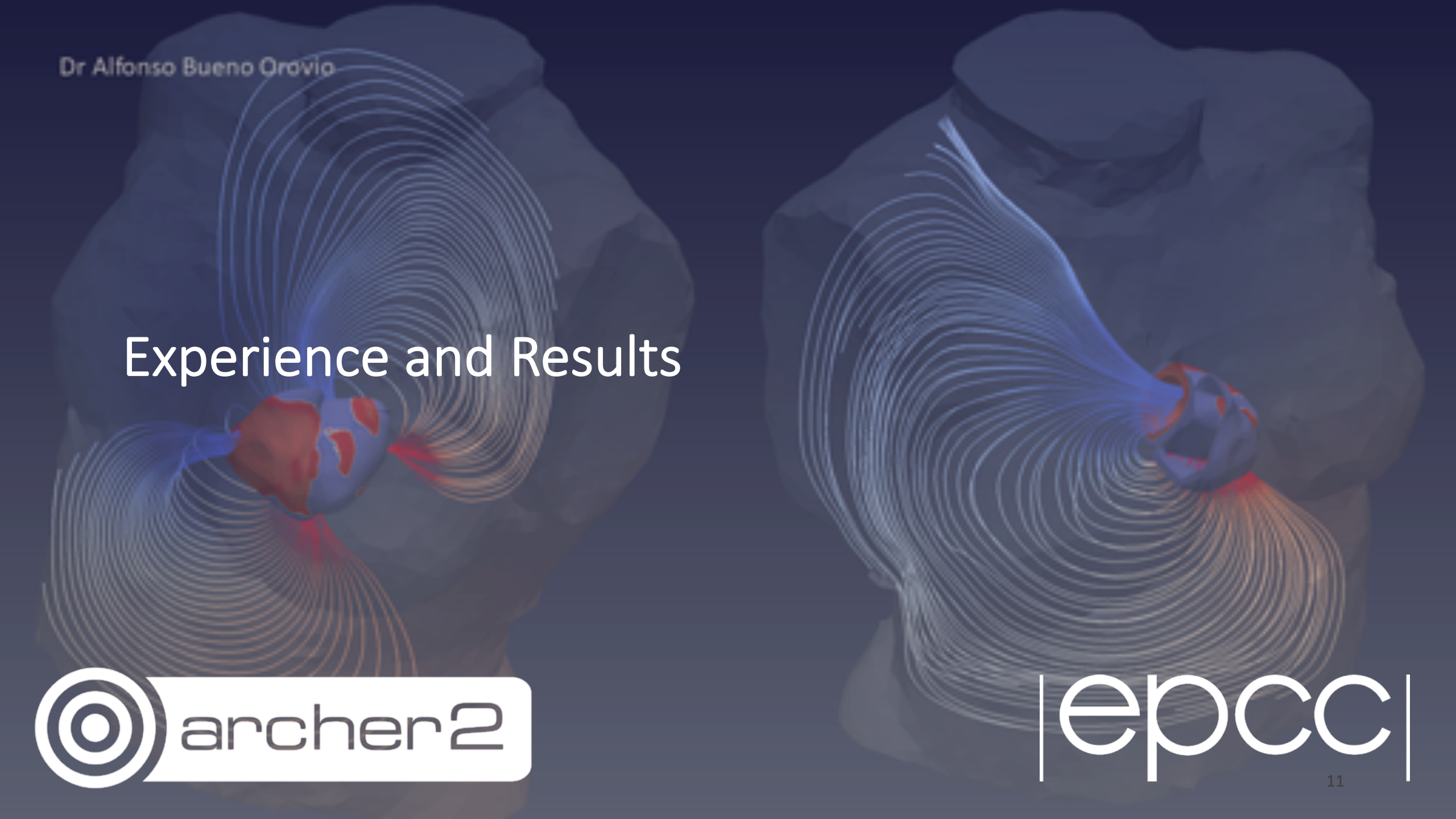
- Hosted in HPE Cray factory in USA
- Single cabinet HPE Cray system running system software v1.2 (pre-release version)
  - Only 8 nodes (4 of which have high memory)
  - Pre- and post- processing node being used as login node
  - Home file system not connected, using local storage instead
- Long term
  - Will be hosted at EPCC ACF in Edinburgh
  - Test and Development System (TDS) will be used to test new software, tools and libraries without disruption to the main system
- Short term
  - First hands-on experience of the HPE Cray EX ARCHER2 hardware and software prior to HPE Cray EX 4 cabinet, 19 cabinet and 23 cabinet systems

Initial Priorities for TDS Access

- HPC Systems team develop experience managing the system
- Integration of key systems (user authentication, SAFE)
- Functional test of software e.g. compilers, tools, batch system, process/thread placement
- Developing experience with the new ARCHER2 hardware
- Improve and enhance user documentation
- Port key application codes to the system
- Internal runs of relevant training courses
- Performance evaluation
  - …not possible to evaluate – more later

Initial Priorities for TDS Access

- HPC Systems team develop experience managing the system
- Integration of key systems (user authentication, SAFE)
- Functional test of software e.g. compilers, tools, batch system, process/thread placement
- Developing experience with the new ARCHER2 hardware
- Improve and enhance user documentation
- Port key application codes to the system
- Internal runs of relevant training courses
- Performance evaluation
  - …not possible to evaluate – more later



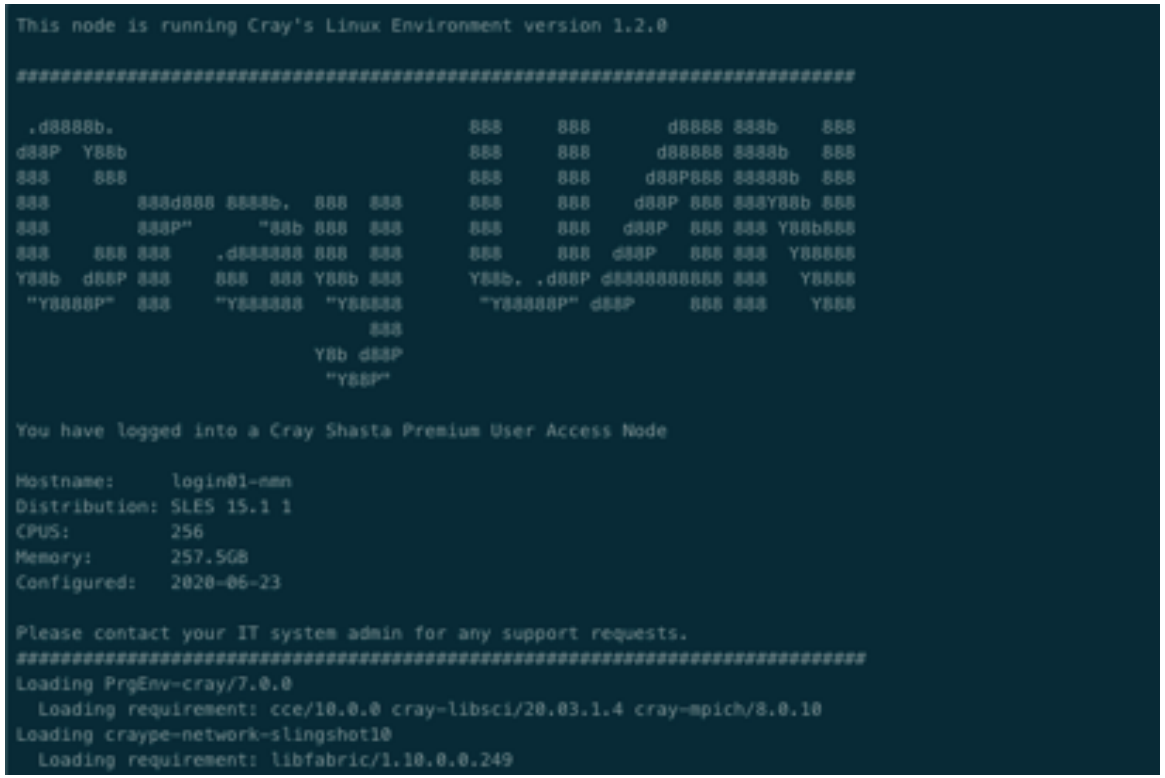
Experience and Results

archer2

epcc

Functional tests

- The vast majority of the functionality just works
- Compilers tested
  - Cray Compiler Environment (CCE) and GCC
  - AOCC not yet available
- Parallel tools tested
  - Cray MPT: MPI, OpenSHMEM, CAF
  - OpenMP
- Performance and debugging tools tested
  - gdb4hpc, valgrind4hpc
  - CrayPAT
- Singularity
  - Requires configuration to local user database in Edinburgh



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 --cpu-bind=cores 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

Scheduler: MPI+OpenMP jobs

```
#!/bin/bash
#SBATCH --job-name=Example_MPI_Job
#SBATCH --time=0:20:0
#SBATCH --nodes=4
#SBATCH --tasks-per-node=8
#SBATCH --cpus-per-task=16
#SBATCH --account=t01

export OMP_NUM_THREADS=16
export OMP_PLACES=cores


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

Select 4 nodes and 8 (MPI) processes per node, stride of 16 cores between processes to make space for OpenMP threads

Set OpenMP threads to 16 and specify placement to get correct pinning to cores

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 the processes and threads to the correct cores

More complex placements are possible – e.g. MPI processes cyclic across NUMA regions with block distribution of threads within NUMA regions.

Applications porting

- Generally, porting has been straightforward
- Applications ported successfully:
  - CASTEP, Code_Saturne, CP2K, GROMACS, LAMMPS, MITgcm, NAMD, Quantum Espresso, VASP
- Still working on:
  - PyChemShell, Elk, FEniCS, Met Office UM, NEMO, NWChem, ONETEP, OpenFOAM
- Also started work on evaluating Cray Python distribution
  - Document usage and functionality available
  - Understand how to build on top of the distribution

Documentation

- Used the TDS access to put draft ARCHER2 documentation in place:

https://docs.archer2.ac.uk

- Will be refined once the 4 cabinet ARCHER2 system is in place with production version of system software
- Public contributions to the documentation are welcome! Issue a PR against the documentation source on Github:

https://github.com/ARCHER2-HPC/archer2-docs

Performance

Not been able to evaluate performance properly on the TDS:

- Only one Slingshot interface active per node
- Issues with power capping of processors

Will revisit on the 4-cabinet ARCHER2 system with the production system software…

Mr. Alejandro de la Calle

Next Steps

epcc

Next Steps

- Commission 4-cabinet ARCHER2 once handed over by HPE Cray
- Repeat functional tests on 4-cabinet ARCHER2 once available
- Install and configure research software applications for user service
- Update application-specific documentation
- Evaluate performance of applications on 4-cabinet ARCHER2
- …release the system to researchers!