Improving energy efficiency on ARCHER2

Adrian Jackson, Alan Simpson, Andy Turner
EPCC, The University of Edinburgh
www.archer2.ac.uk
a.turner@epcc.ed.ac.uk

archer2

epcc

Outline

- ARCHER2 service
- Efficiency priorities
- ARCHER2 power draw
- Reducing power draw/energy use
  - CPU power BIOS setting
  - Default CPU frequency
- Lessons learned and summary

ARCHER2 Partners

epcc

epcc

THE UNIVERSITY of EDINBURGH

Hewlett Packard Enterprise

This work was partially funded by the UKRI Digital Research Infrastructure Net Zero Scoping Project (NE/W007134/1) https://net-zero-dri.ceda.ac.uk/

- **UK National Supercomputing Service** – based at EPCC at The University of Edinburgh
- Designed to **enable world-leading computation** for a wide range of research areas in the UK
- User base of **over 3000 researchers**
- **Huge range of software** for modelling and simulation each with different performance properties
- Allocations and usage measured via *residency* – how much resource you use for how long
- Service also includes training, support, outreach and software development on top of the hardware itself

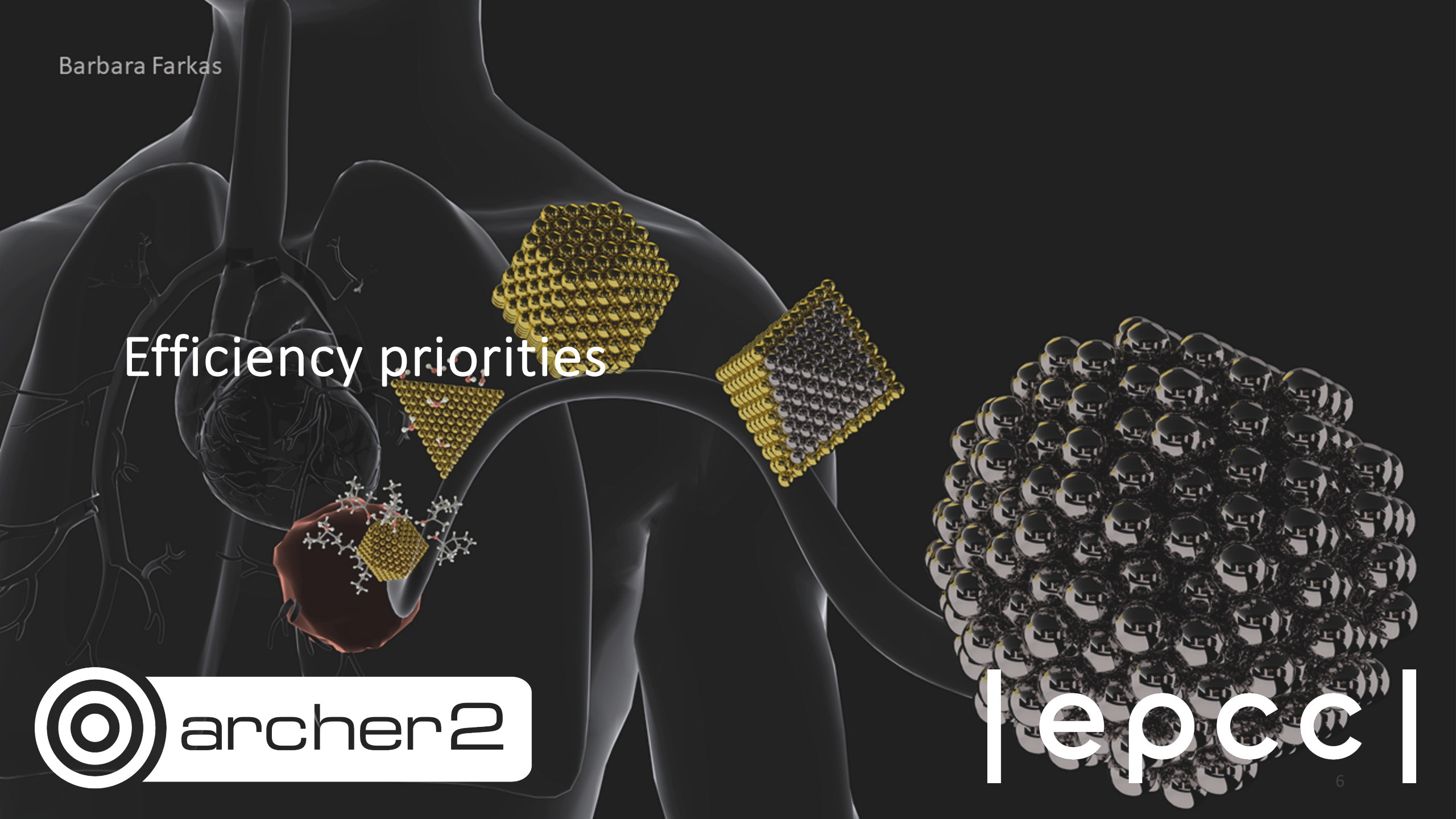
Application Type	Approx. % Use
Quantum Materials Modelling	40%
Earth Systems Modelling	20%
Computational Fluid Dynamics	15%
Biomolecular Modelling	15%
Classical Materials Modelling	5%
Plasma Physics	3%
Quantum Chemistry	2%

ARCHER2 Technology

- HPE Cray EX Supercomputer
- 5,860 compute nodes
  - 750,080 CPU compute cores
- HPE Slingshot 10 interconnect
- Compute nodes:
  - Dual socket AMD EPYC™ 7742 Processors, 64c, 2.25 GHz
  - 256 GiB / 512 GiB memory per node
  - Two 100 Gbps HPE Slingshot 10 interfaces per node
- 4x ClusterStor L300 Lustre file systems, each 3.6 PB
- 1 PB ClusterStor E1000F solid state storage
- 4x NetApp FAS8200A file systems, 1 PB total

Barbara Farkas

Efficiency priorities



archer2

epcc

Different sites have different priorities

- Priorities and motivations vary between sites, and may include:
  - Reducing running costs
  - Reducing carbon emissions
  - Reducing energy use
  - Power demand control to improve integration between HPC centres and energy grids
  - Educating and enabling users to be energy-aware
  - Fair attribution of actual costs
- Different efficiency targets means different operational decisions
- Doing the “right” thing can be complicated

Carbon emissions vs energy

- Understanding carbon emissions is increasingly important for HPC in the context of reducing worldwide limits on such emissions
- A significant component of HPC emissions already comes from embodied emissions (from manufacture, delivery, decommissioning, etc.)
  - And fractional contribution will increase as more energy grids decarbonize
  - Can be hard to get firm numbers on embodied emissions
- When energy emissions are low, most emissions-efficient use is to run as fast as possible irrespective of energy cost
  - Get the most out of the embodied emissions before service is decommissioned
- However, this is a less energy-efficient approach to running an HPC service
- Tension between minimising total carbon emissions and minimising energy usage

Example: ARCHER2

- Estimates from UKRI DRI Net Zero project suggest around 1100 $kgCO_2e$ per compute node
- Using this figure and ignoring other components for simplicity
  - 5860 compute nodes
  - Total embodied emissions estimate = 6,446,000 $kgCO_2e$

Scenario	gCO_2/kWh	Energy Emissions: per annum[1] ($kgCO_2$)	Energy Emissions: 5 years ($kgCO_2$)	Embodied Emissions ($kgCO_2e$)	% Total emissions over 5 years
Green energy	~0	~0	~0	6,446,000	0%
South Scotland	48[2]	1,261,440	6,307,200	6,446,000	49%
UK	268[3]	7,043,040	35,215,200	6,446,000	85%
World	441[3]	11,589,480	57,947,400	6,446,000	90%

[1] Assuming 3 MW power draw
[2] Median value from 12 months: 1 Apr 2022 – 31 Mar 2023. https://electricityinfo.org/
[3] https://ourworldindata.org/grapher/carbon-intensity-electricity

Why consider energy efficiency?

- Increasing in importance in the Exascale era as both energy usage and costs rise
- Total Cost of Ownership of HPC centres used to be dominated by capital costs but energy costs may now make up a significant fraction
- Can maximise ”science per kWh”

- For the rest of this talk, we focus on reducing energy and power as these have practical impacts:
  - Reduces carbon emissions from systems that have already been procured
  - Reduces running costs and TCO
  - Increases control over power demand

Dr Alfonso Bueno Orovio

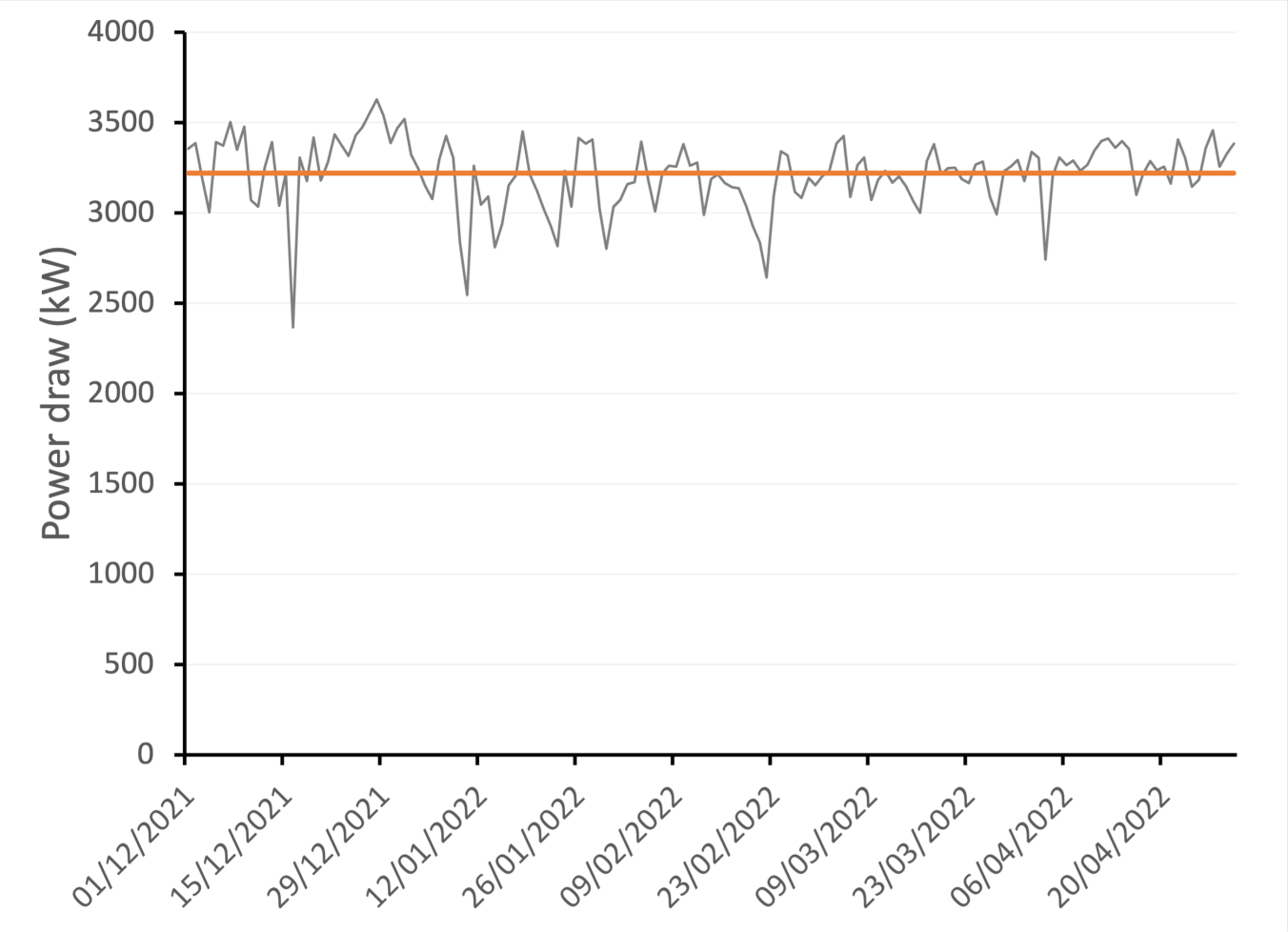
ARCHER2 power draw

archer2

epcc

Historical power draw measurements

- Power draw before any changes made
- Utilisation on ARCHER2 is consistent – just over 90%
- Mean power draw from compute node cabinets: 3220 kW
- Measurements taken from the chassis management infrastructure in Mountain cabinets



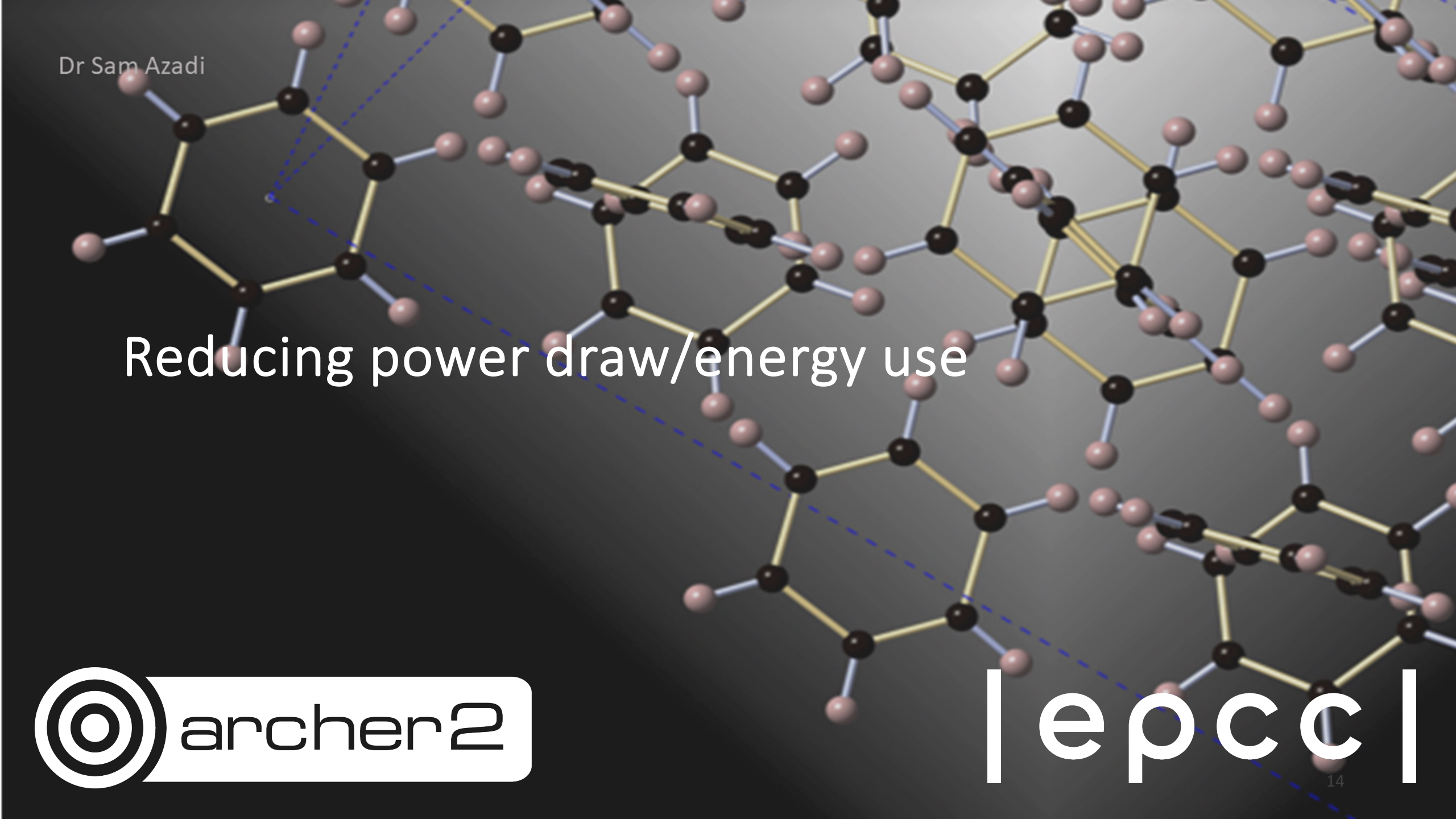
Power draw by component

Estimated loaded power draws for ARCHER2 components:

- Some values measured by experiments and others provided by HPE engineers

Component	Notes	Idle (each)	Loaded (each)	Approx. %
Compute nodes	5860 nodes	1350 kW (0.23 kW)	3000 kW (0.51 kW)	80%
Slingshot interconnect	768 switches	100-200 kW (0.10-0.25 kW)	540 kW (0.70 kW)	10%
Other Cabinet Overheads	23 cabinets	100-200 kW (4.3-8.7 kW)	210 kW (9.1 kW)	6%
Coolant Distribution Units	6 CDUs	96 kW (16 kW)	96 kW (16 kW)	3%
File systems	5 file systems	40 kW (8 kW)	40 kW (8 kW)	1%
Service nodes	Negligible	-	-	
Total		1800 kW	3900 kW	

- Energy use dominated by compute cabinets; storage power not important
- Idle power draw of compute nodes is high – likely dominated by memory and NIC



Dr Sam Azadi

Reducing power draw/energy use

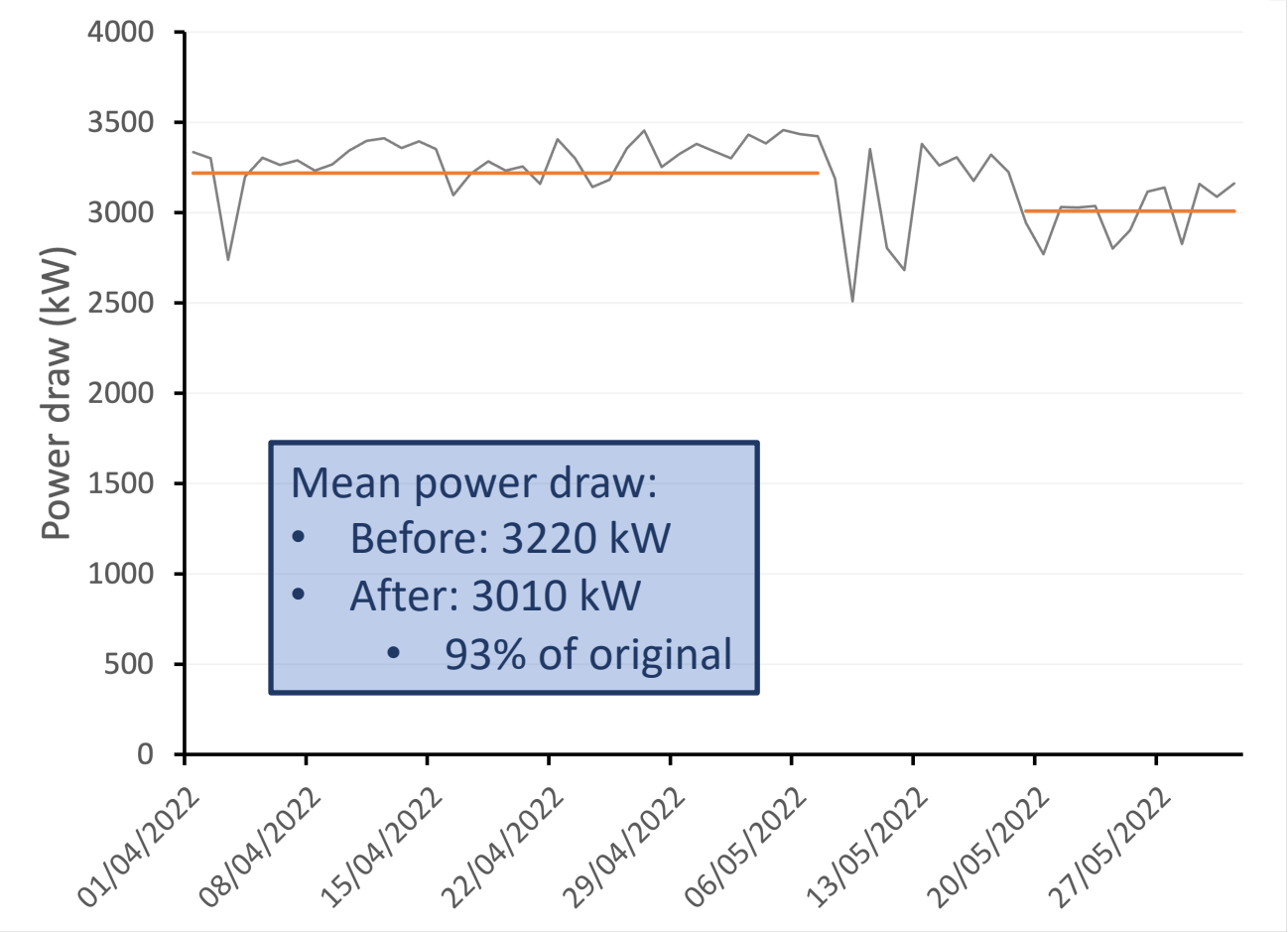
archer2

epcc

Power/Performance Determinism

- In May 2022 the ARCHER2 compute nodes had a CPU BIOS setting changed from *Power Determinism* mode to *Performance Determinism* mode

- *Performance Determinism* keeps node performance more in-sync
  - Performance of multi-node parallel applications is determined by slowest node
  - Any extra power draw for performance above the slowest node is wasted power



https://www.amd.com/system/files/2017-06/Power-Performance-Determinism.pdf

Impact on application performance

Application benchmark	Number of nodes	Performance ratio PerfMode:PowerMode	Energy[1] ratio PerfMode:PowerMode
CASTEP Al Slab	16	0.99	0.94
OpenSBLI TGV 1024^3	32	1.00	0.90
VASP TiO_2	32	0.99	0.93

[1]Energy measured from on-node energy use counters – only reflects node energy use

- Performance impact is generally low – expected to be lower where more nodes are used
- Energy savings measured using cabinet power in line with energy savings measured on compute nodes
  - Suggests that overheads on top of compute node power do not affect conclusions

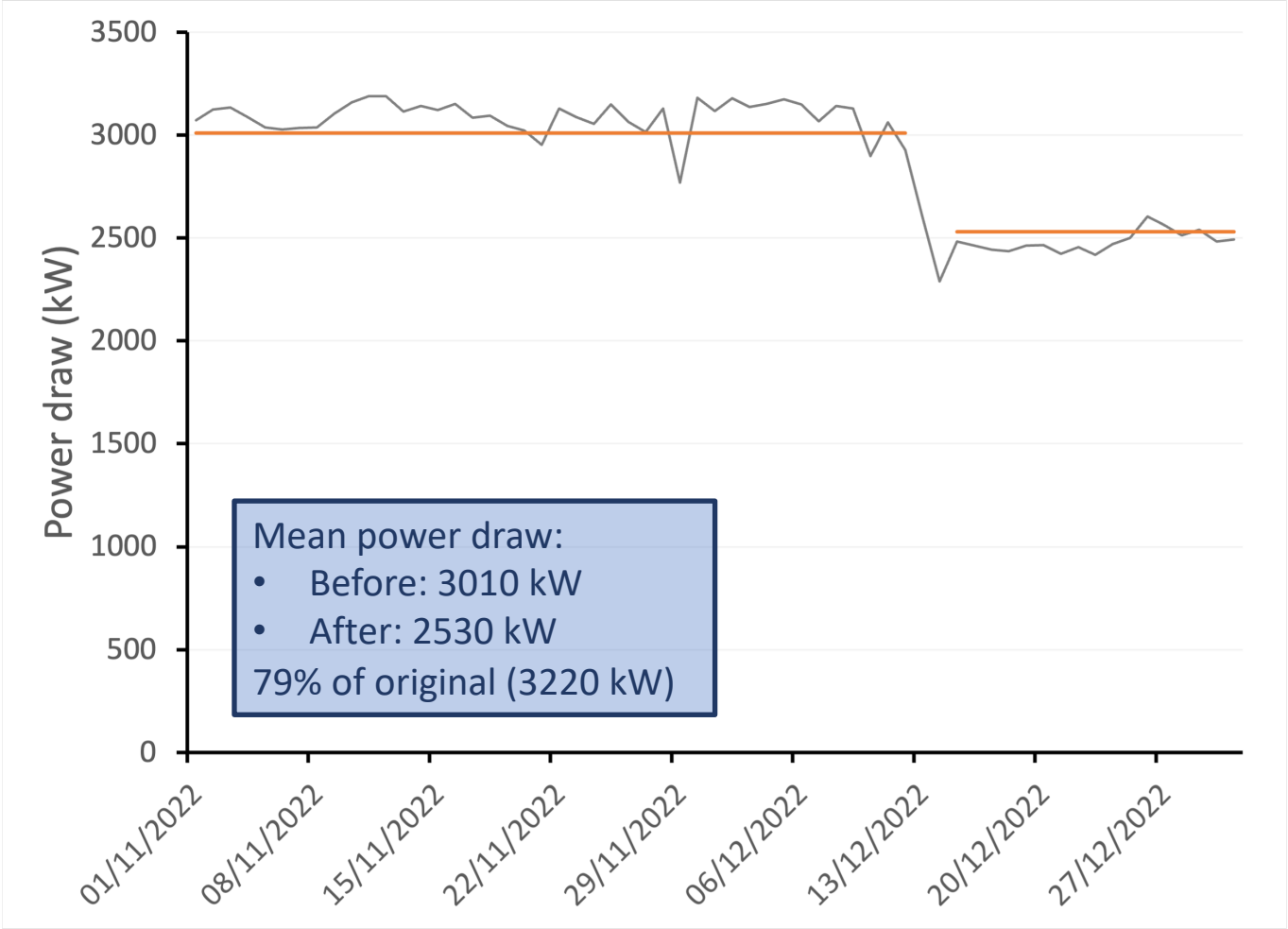
CPU Frequency – impact on power draw

Changed on 12 Dec 2022

Default CPU frequency:

- Before: 2.25 GHz (can turbo boost)
  - Typically boosts to ~2.8 GHz when all cores running intensively
- After: 2.00 GHz (no turbo boost)
- Some applications kept at original 2.25 GHz setting (with turbo boost)

Freed up significant power on the local electricity grid during period of potential electricity shortages



CPU Frequency – impact on performance

Application benchmark	Research areas	Performance ratio 2.0 GHz:2.25 GHz	Node Energy ratio 2.0 GHz:2.25 GHz
VASP CdTe	Materials science, Mineral physics	0.95	0.88
GROMACS 1400k atoms	Biomolecular simulation	0.83	0.92
CP2K H2O 2048	Materials science	0.91	0.93
LAMMPS Ethanol	Materials science, Engineering, Biomolecular modelling	0.74	0.92
CASTEP Al Slab	Materials science	0.93	0.88
ONETEP hBN-BP-hBN	Materials science	0.92	0.82
Nektar++ TGV 128 DoF	Engineering	0.80	0.80

- All applications are more energy efficient at 2.0 GHz
- Looking at cost-efficiency would suggest:
  - Frequency set to 2.25 GHz: GROMACS and LAMMPS, Nektar++ [due to increased residency costs]
  - Frequency set to 2.0 GHz: VASP, CASTEP, ONETEP, CP2K
- Default frequency: 2.0 GHz with strong advice to users to test impact on their software

CPU Frequency – impact on performance

- What is the impact on energy use beyond just node energy use?
- Reserved a full cabinet (256 nodes) and filled with copies of benchmarks
- Initially focussed on applications which would be running at 2.0 GHz

Experiment	Cabinet energy use (kWh)[1]	Node energy use (kWh)[2]	Overheads (kWh)	% Overheads	Cabinet ratio to 2.25 GHz	Node ratio to 2.25 GHz
8-node VASP, 256 nodes, 2.25 GHz	43.9	35.3	8.6	19.6%		
8-node VASP, 256 nodes, 2.00 GHz	38.5	30.4	8.1	21.0%	0.88	0.86

Experiment	Cabinet energy use (kWh)[1]	Node energy use (kWh)[2]	Overheads (kWh)	% Overheads	Cabinet ratio to 2.25 GHz	Node ratio to 2.25 GHz
4-node ONETEP, 256 nodes, 2.25 GHz	128.2	108.3	19.8	15.5%		
4-node ONETEP, 256 nodes, 2.00 GHz	107.8	88.5	19.3	17.9%	0.84	0.82

[1] Calculated from instantaneous cabinet power draw measurements during benchmark runtime
[2] Sum of energies from all calculations in set that filled 256 nodes

- Energy savings measured at the node level clearly propagate to full cabinet energy use
  - Cabinet energy use includes interconnect switches and power overheads

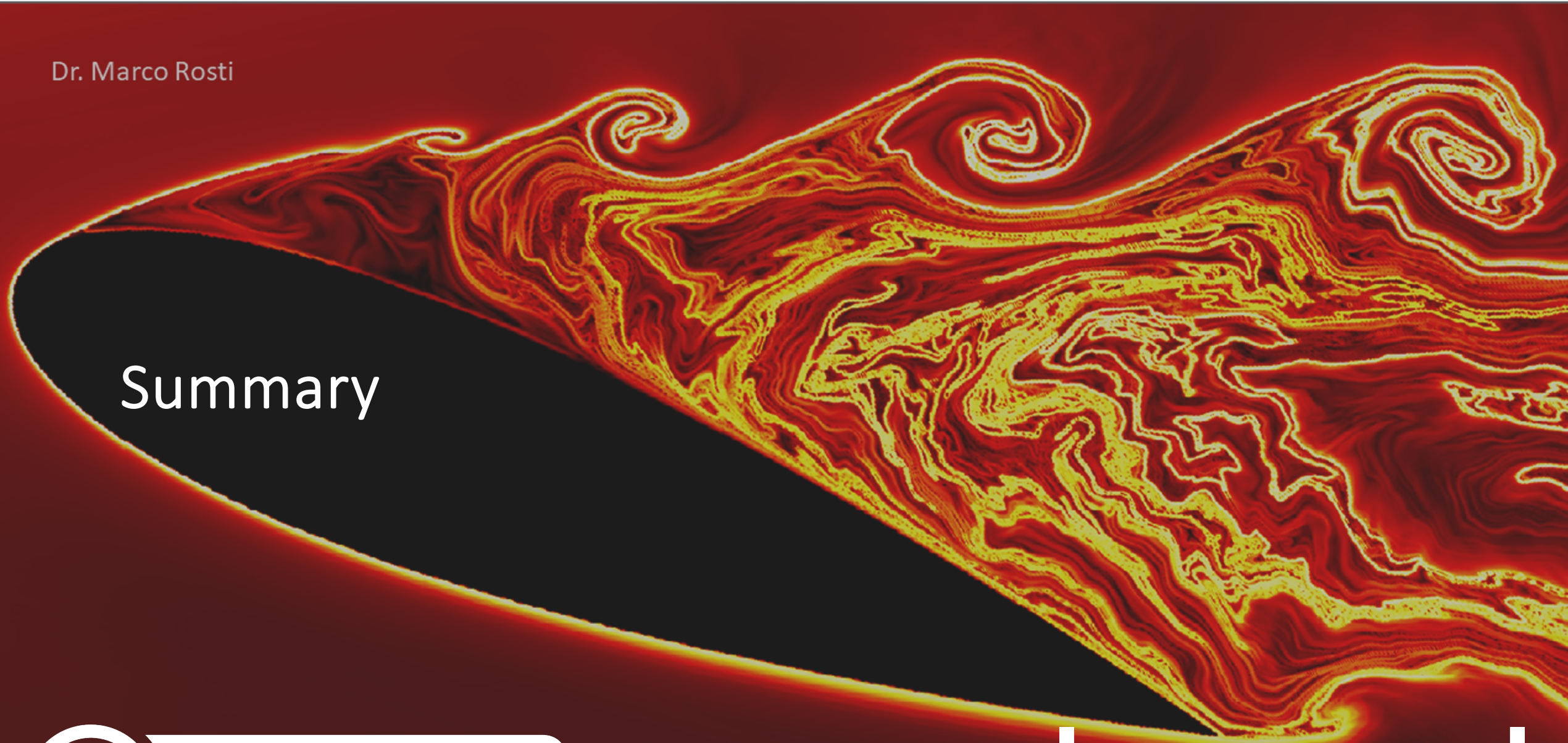
Understanding power draw

- Used single cabinet reservations to try and understand power draw better
  - 256 nodes, 32 switches
- Run enough copies copies of benchmarks to fill 256 nodes – all at 2.25 GHz with turbo-boost enabled
- Compare cabinet power draw to compute node power draw (from on-node counters)

Experiment	Median node power draw	Total node power draw	Cabinet power draw	Non-node power draw	% Overhead compared to node power draw
Idle	230 W	58.9 kW	75.6 kW	16.7 kW	28%
1-node HPL	513 W	131.3 kW	150.8 kW	19.5 kW	15%
8-node VASP	497 W	127.2 kW	149.2 kW	22.0 kW	17%
16-node OSU Alltoall	489 W	125.1 kW	156.7 kW	31.6 kW	20%

- Non-node power draw overheads increase as communication intensity increases
- Implies that interconnect switch power draw is the significant other component in compute cabinets in terms of power draw

Dr. Marco Rosti



Summary

Lessons learned

- High utilisation levels are critical for efficiency due to high idle power draw
  - The sector should investigate ways to reduce idle power draw of components
- Instrumentation of energy use needs to improve
  - Compute nodes are generally well covered but other key components (e.g., switches) are not
  - Makes it challenging to fully understand energy use or to introduce energy-based charging
- High quality information from vendors on embodied carbon associated with hardware is critical for good operational decision making
  - The current level of information is generally poor
- Need to know what your priorities are in order to make appropriate choices
  - Carbon emissions, energy, power, cost,…

Summary

- Changes which are quick to implement can have a large effect on energy use
- Gives flexibility to respond to particular requirements
  - Being asked to reduce demands on grid during specific periods
  - Reducing power when cooling infrastructure is under pressure
- Changing the CPU BIOS setting saves energy for large jobs and has negligible impact on performance
- Reducing the default processor frequency is worth considering
  - All application benchmarks showed lower energy use at 2.0 GHz
- On ARCHER2, we reduced energy usage by around 700 kW (21%)
  - With only modest impact on performance
  - Reducing demand on the power grid over winter
  - Making significant savings on running costs
- Future work planned to improve carbon auditing/modelling of large scale HPC systems

Recommendations

Improve the quality of data on embodied emissions from hardware vendors

- Embodied emissions are likely the largest component of emissions from HPC
- As the energy grid continues to decarbonise, this will become more pronounced
- Current quality of information is poor

Instrumentation of energy use needs to improve

- Compute nodes are generally well covered but other key components (e.g., switches) are not
- Makes it challenging to fully understand energy use or to introduce energy-based charging

HPC services should analyse their emissions based on operational data

- Emissions balance depends on location, lifetime and hardware

HPC services should plan how to maximise impact as a function of emissions

- How this is done depends on emissions balance and may lead to reduced energy efficiency

High utilisation levels are critical for efficiency due to high idle power draw

- The sector should investigate ways to reduce idle power draw of components

HPC sector should work to better understand emissions associated with in silico experiments vs physical experiments to enable correct policy and strategic decisions by government/funders