Measureing Carbon

NetZero and the IRISCAST Project

R. A. Owen[1,*] and *J.* Hays[1,**]

[1]School of Physical and Chemical Sciences, Queen Mary University of London,
327 Mile End Road, LONDON, E1 4NS, United Kingdom

Abstract.
Moving towards NetZero requires robust information to enable good decision making at all levels: covering hardware procurement, workload management and operations, as well as higher level aspects encompassing grant funding processes and policy framework development. The IRISCAST project is a proof-of-concept study funded as part of the UKRI DRI Net-Zero Scoping Project.
IRISCAST performed an audit of carbon costs across a multi-site heterogeneous infrastructure by collecting and analysing snapshots of actual usage across different facilities within the IRIS community (https://iris.ac.uk). Combining usage information with an analysis of the embodied carbon costs and careful mapping and consideration of the underlying assumptions resulted in an estimate of the overall carbon cost, an understanding of the key elements that contribute to the carbon cost, and the important metrics needed to measure it.
IRISCAST makes reccomendations to allow high level feedback of carbon costs to funding bodies to inform strategic decisions as well as low level feedback of carbon costs to users and user communities to drive changes in user code bases and behaviors to be more carbon efficient.
IRISCAST carbon modeling shows that estimates of carbon costs can vary by factors of $\sim$ 10, hence there is significant opportunity for the carbon footprint of Digital Research Infrastructures to be reduced.

1 Introduction and Objectives

On the 1st May 2019 the UK parliament declared an environment and climate emergency. Later that year the UK became the world's first major economy to adopt a legally binding target to reduce its greenhouse gas emissions to NetZero by 2050. On the back of that UK Research and Innovation (UKRI) – the UK government agency that allocates public funds to research – committed to becoming carbon NetZero by 2040. This includes its digital research infrastructure (DRI).

The URKI NetZero DRI Scoping Project [1] was formed and tasked with producing a roadmap to help achieve a NetZero DRI. The Scoping Project commissioned nine core consortium projects and seven community lead sandpit projects, IRISCAST among them, as can be seen in Figure 1 including IRISCAST as one the sandpit projects.

*e-mail: r.a.owen@qmul.ac.uk
**e-mail: j.hays@qmul.ac.uk

Moving towards Net-Zero for DRIs requires robust information to enable good decision making at all levels. This covers low level decision making about hardware procurement, workload management and operations, as well as higher level aspects encompassing grant funding processes and policy framework development.

IRISCAST was founded on the premise that making good decisions regarding DRI requires understanding, as much as possible, the full carbon costs associated with operating, maintaining, and using the infrastructure: going beyond accounting for electricity and cooling and including the full chain of costs embodied in the infrastructure.

The ambitious objectives of the IRISCAST project were to estimate carbon costs for scientific computing across a broad heterogeneous landscape, while identifying both the key drivers for carbon costs and the barriers to measuring carbon costs.

The project aimed to work coherently across different communities to collaborate on measuring and communicating the carbon costs, as only by communicating the costs can we drive change.

As a scoping project a learning by doing approach was adopted and workshops and meetings were planned to foster coherent work across different facilities with different remits, tooling, and capabilities.

Difficulties, issues and barriers were to be gathered and documented to help inform requirements for future work and decision making.

IRISCAST also sought to engage a range of academic communities and build a foundation for future action.

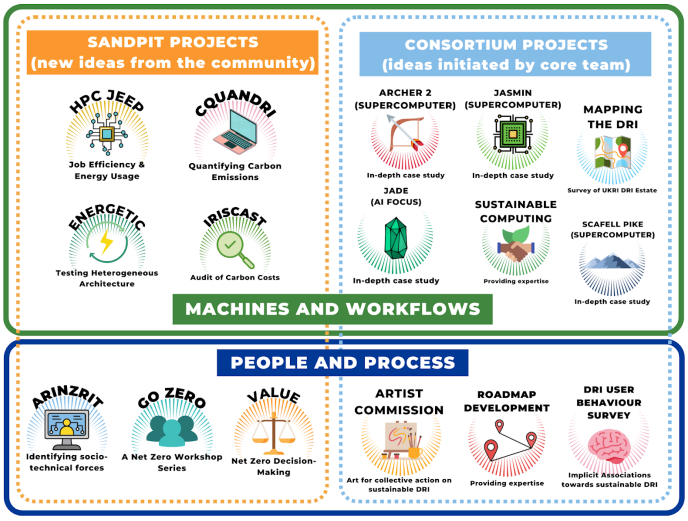

Figure 1. A representation of the constituent sub-projects of the UKRI NetZero DRI Scoping Project showing the core consortium projects and the community sandpit projects. The sub-projects can also be classified as pertaining to Machines and Workflows or People and Process. After [1]

2 IRISCAST Carbon Model

IRISCAST defined a carbon model [2] to calculate the total carbon cost (C_t^p) attributable to operating a DRI resource during a specific time period. This is in essence simply the sum of the active/operational (scope 1&2) carbon cost (C_a^p) for that period plus the embodied (scope 3) carbon cost (C_e^p) apportioned to that period.

$$C_t^p = C_a^p + C_e^p$$

The complexity comes when formulating what is included in C_a^p and C_e^p and how they are derived. For active carbon costs, C_a^p, the difficulty for this metric is deciding and defining

which resources are included in the DRI, apportioning the percentage of resources shared by the DRI and other infrastructure, and defining the scope of resources. The computer nodes involved in a specific DRI are relatively straightforward to identify as is the network equipment, although a little more care may be needed in identifying the network demarcation point. The wider campus network and internet were deemed out of scope. Facility carbon costs are also included in C_a^p where these facilities are identified as: Cooling systems for the DRI resources and buildings hosting the DRI resources; power distribution units and transformers supplying DRI resources and infrastructure; uninterruptible power supply (UPS) resources supporting DRI systems; Facility electricity usage, such as lighting, fire and security systems, as well as other ancillary systems within the data centre/building hosting the DRI resources. Clearly carbon costs of shared resources in this list must be apportioned to DRI and non-DRI uses as appropriate for that facility thus adding some uncertainty to the result.

Assuming all carbon costs are due to electricity usage it can be stated that

$$C_e^p = CM_e^p \left(E_{nodes}^p + E_{network}^p + E_{cooling}^p + E_{power}^p + E_{facility}^p \right)$$

Where E_{nodes}^p is the total energy used by all the nodes in the period, $E_{network}^p$ the total energy used by DRI network in the period, and $E_{cooling}^p$, E_{power}^p and $E_{facility}^p$ are the total apportioned energies for cooling, power distribution and other facility support in the period and CM_e^p is a factor to convert the energy used into carbon equivalent units derived from the electricity/energy supply mix for period p.

The embodied carbon costs require consideration of the carbon emitted in creating the resources in the first place and can be represented as:

$$C_e^p = \sum_{1}^{nodes} \sum_{1}^{p} C_{enode} + \sum_{1}^{network} \sum_{1}^{p} C_{enetwork} + \sum_{1}^{facility\ items} \sum_{1}^{p} C_{efacilities}$$

For embodied carbon, C_{enode} is the carbon emitted in creating, delivering, installing and disposing of a given node, likewise for network components and for the facility components discussed earlier. Embodied carbon is not dependent on the operational time period under consideration as the actual emissions happened at the time of creation of the resource. However, it is sensible to apportion those embodied carbon emissions to the period under consideration in relation to the expected lifetime of the resource. For instance a computer with an expected lifetime of 5 years would have 20% of its embodied carbon emissions apportioned to each year of its life. However a building with an expected 20 year lifetime would only have 5% of its embodied carbon cost apportioned to each year. For shared resources an additional apportionment to each shared use should be made.

From analysis of this model it can be seen that the key inputs to the model are: the electricity usage of a DRI; the Grid Carbon intensities for the relevant period; and an inventory of the DRI equipment including embodied carbon costs and expected lifetimes.

3 IRISCAST Snapshot

Six UKRI DRI resources contributed to the IRISCAST audit. To fit within the timescale and resources of the project each DRI resource chose the scope of the equipment they surveyed for IRISCAST and the inventories below reflect that. IRISCAST conducted a 24 hour snapshot simultaneously across sites. Due to failures at some sites in the first snapshot period a second snapshot was made at a sub-set of sites and the data from the first successful snapshot at each site was further analysed.

The QMUL GridPP T2 cluster deemed four racks of equipment in scope for audit as per the inventory in table 1. QMUL chose to collect cumulative energy readings from the APC Power Distribution Units (PDUs) via SNMP thus measuring power into the racks. Cumulative energy consumption for each compute node was gathered by using FreeIPMI to query a node's Baseboard Management Controller (BMC). A method to query the switch energy consumption was not found. At the sub-node level turbostat was used to gather energy consumption of CPU and RAM using the CPU's Running Average Power Limit (RAPL) facilities[3]. Job level information was obtained from the Slurm scheduler logs. Slurm was not configured to collect job energy information. QMUL found that there was no capability to measure energy usage at the facility level; neither the air conditioning electricity feed nor the server room electricity feed had any metering available.

The Imperial GridPP T2 cluster was unable to make measurements at the facility level as the cluster is housed in a shared datacenter outside of the cluster administrators' control. PDU measurements were unavailable for the snapshot periods. Imperial chose to return data on 241 nodes broadly comprising seven models of hardware, having been procured in batches over a number of years. The inventory details can be seen in table 1. Instantaneous power usage from each node's BMC was logged via IPMItool run out of a python script which also logged the operating system load average and CPU utilisation.

Cambridge University's Research Computing Services selected 60 IRIS funded compute nodes to be in scope for audit. The inventory is shown in Table 1. Power usage data was collected from each node's BMC using Prometheus Redfish while Prometheus Node Exporter collected data regarding CPU and RAM usage and Idle time. Facility level energy usage including cooling was not available.

The DiRAC clusters COSMA7 and COSMA8 hosted at Durham University participated in the IRISCAST audit. Measurements of power or current to each rack were retrieved from PDU's via ssh. The IPMI protocol was used to collect node level data from BMCs. The DiRAC inventory from Durham can be seen in Table 1. The COSMA clusters are managed by a Slurm job scheduler which was configured to collect energy usage per-job which was was extracted from the Slurm accounting database along with other job data. Facility level energy usage including cooling was not available.

The Scientific Computing Application Resource for Facilities (SCARF) cluster run by STFC at the Rutherford Appleton Laboratory participated in the IRISCAST audit. The SCARF equipment shown Table 1 is housed in 27 racks. Power to the racks was measured by querying PDU's over SNMP. The Mellanox switches were able to report energy usage over SNMP but the other switches could not. IPMItool was used to make instantaneous power readings from BMCs on the 571 SCARF nodes of 11 different Supermicro or Dell models purchased between 2015 and 2021. This heterogeneity highlighted that IPMI is not the same on all models and different models/BMC's provide different energy or power readings. The instantaneous power reading from the IPMItool was used for the IRISCAST analysis as it was the most consistent across all node types. Job level information was queried from the Slurm scheduler accounting database but this did not include job energy usage as Slurm was not configured to collect this data. Facility level energy usage including cooling was not available.

The STFC Cloud hosted at RAL is a private OpenStack cloud service for STFC and IRIS users to create and run virtual machines (VMs) co-located with the SCARF cluster described immediately above. Facility level energy usage including cooling was not available. However the STFC Cloud equipment in Table 1 housed in 30 racks was subject to IRISCAST audit.

Power to the racks was queried over SNMP from PDU's. Energy usage was queried over SNMP from the Mellanox switches but not from the other switches. The hypervisor nodes, storage nodes and auxiliary nodes were monitored by IPMItool yielding instantaneous power

DRI		Specification		
Resource	Model	CPU	RAM	Quantity
QMUL	Dell R640	-	-	118
QMUL	Melanox SN2410	-	-	4
QMUL	APC APDU9953	-	-	12
Imperial	Dell R410	-	-	68
Imperial	Dell R430	-	-	60
Imperial	Dell R440	-	-	15
Imperial	Dell R6526	-	-	30
Imperial	HPE SL2x170z G6	-	-	24
Imperial	SYS-6028TP-HTR	-	-	12
Imperial	X9DRT	-	-	24
Imperial	Generic Servers	-	-	8
Cambridge	Dell C6320	Intel E5-2690 v4	256 GB	60
Durham	Dell C6420	Intel Gold 5120	512 GB	452
Durham	Dell C6525	AMD EPYC 7H12	1024 GB	360
STFC SCARF	-	AMD EPYC 7502	256 GB	246
STFC SCARF	-	Intel Gold 6126	192 GB	164
STFC SCARF	-	Intel E5-2650v4	128 GB	201
STFC SCARF	-	Intel E5-2650v3	128 GB	88
STFC SCARF	Network Switches	-	-	-
STFC Cloud	Dell C6420	Intel Xeon 4108	96 GB	96
STFC Cloud	Dell C6525	AMD EPYC 7452	512 GB	138
STFC Cloud	Supermicro	AMD EPYC 7452	512 GB	238
STFC Cloud	Supermicro	Intel 6130	384 GB	74
STFC Cloud	Dell	Various	Various	10
STFC Cloud	GPU Nodes	Various	Various	94
STFC Cloud	FPGA Node	Intel 6148	192 GB	1
STFC Cloud	Control Plane	Various	Various	12
STFC Cloud	Storage Nodes	Various	Various	105
STFC Cloud	Network Switches	-	-	-

Table 1. Inventory detailing the equipment subject to carbon audit at the six DRI resources included in IRISCAST

usage data. IRISCAST could not find a tool to monitor the payload energy usage, which for a cloud system is the Virtual Machine (VM) energy usage. PowerTop seemed like promising tool but can only really be used on laptops. The Prometheus tool should be investigated in future.

The Opensearch dashboard system at RAL was already configured to collect STFC Cloud data via a continuous data pipeline. IRISCAST used this same Opensearch dashboard to collate the data from the other DRI resources by importing from site supplied CSV, XML and json data files as needed. Additional data analysis was conducted in Python using pandas dataframes.

4 IRISCAST Results

IRISCAST asked each DRI Resource to measure energy usage at the Facility, Enclosure, Node and Payload levels. However data was only readily available at the Enclosure and Node levels measuring using PDU's and BMC's respectively. Table 2 shows the total energy usage measured over the 24 hour snapshot period at the different DRI resources. It can be seen that where PDU and BMC data is available the BMC data read is $\sim$ 20% lower than

DRI	Energy measured in kWh by				Number
Resource	Facility	PDU	BMC	Turbostat	of Nodes
QMUL	1299	1299	1279	1214	118
Imperial	944	-	944	-	117
Cambridge	261	-	261	-	59
Durham	8154	8154	6267	-	876
STFC SCARF	4271	4271	3292	-	571
STFC Cloud	3831	-	3831	-	721
Total	18760				

Table 2. Table showing the results of IRISCAST energy measurements. The Facility column shows the best estimate of total Facility energy usage based on PDU data where available and BMC data otherwise.

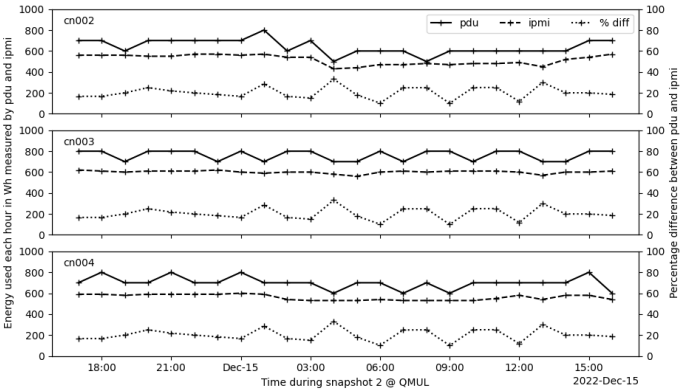

Figure 2. Graph showing energy measurements made at the PDU port outlet and corresponding node BMC/IPMI energy measurements at QMUL in the second snapshot period. The percentage difference in the energy measurements is also shown and in all cases there is a discrepancy of $\sim 20\%$

the PDU measurements, with the exception of QMUL where the difference is only $\sim 1.5\%$. This $\sim 20\%$ discrepancy can perhaps be attributed to losses in node power supplies, although the exact origin is unclear. Further measurements at QMUL, which can be seen in Figure 2 using different APC AP8459WW PDU's, which could report per-port energy usage, did show a $\sim$ 20% difference to BMC measurements made by IPMI. Clearly there is more to investigate here but in all cases it appears that the BMCs somewhat underestimate energy usage compared to PDUs.

The administrators at each DRI resource were able to choose their own methods and tools to measure electricity usage. Table 3 shows the devices, protocols and tools used at each site. Universally BMCs were used to monitor individual nodes electricity usage. Predominantly these were read using the IPMI protocol often with ipmitool. Most DRI resources also were able to measure electricity usage at the enclosure level using PDUs, often from the vendor APC, these were predominantly read using SNMP using a variety of tools.

Electricity usage was measured variously as instantaneous current or power draw, which required integrating to give energy usage, or read out directly as cumulative energy usage. It is concluded that measuring directly as cumulative energy usage is preferred as this is less error prone due to inaccuracies in integrating power or current measurements and more robust against missing measurements. However it is acknowledged that not all equipment has this facility. For APC PDU's cumulative energy can often be read by SNMP, in the rather strange unit of hectowatthours, by querying `PowerNet-MIB::rPDU2DeviceStatusEnergy.1` for global measurements or `PowerNet-MIB::rPDU2OutletMeteredStatusEnergy.n` for a per port measurement from port n. Some BMCs also can yield cumulative energy usage for

DRI	Enclosure Level			Node Level		
Resource	Device	Protocol	Tool	Device	Protocol	Tool
QMUL	PDU	SNMP	Net-SNMP	BMC	IPMI	FreeIPMI
Imperial	-	-	-	BMC	IPMI	ipmitool
Cambridge	-	-	-	BMC	Redfish	Prometheus
Durham	PDU	SSH	ssh	BMC	IPMI	unknown
STFC SCARF	PDU	SNMP	LibreNMS	BMC	IPMI	ipmitool
STFC Cloud	PDU	SNMP	LibreNMS	BMC	IPMI	ipmitool

Table 3. Table showing devices, protocols and tools used to measure electricity usage at each DRI resource in the IRISCAST audit.

	Scenario		
Factor	Low	Medium	High
Carbon Intensity (gCO_2/kWh)	50	175	300
Power Usage Effectiveness (PUE)	1.1	1.3	1.6
Server Embodied Carbon ($KgCO_2$)	400	-	1100
Server Lifespan (years)	3	5	7

Table 4. Model scenarios using low, medium and high estimates of values that IRISCAST could not measure.

		Total carbon footprint estimate ($kgCO_2$) (Percentage active carbon)		
Server embodied carbon	**Server lifespan**	**PUE Low Carbon Intensity Low**	**PUE Medium Carbon Intensity Medium**	**PUE High Carbon Intensity High**
Low	**3**	1950 (55%)	5293 (83%)	10186 (91%)
	5	1600 (67%)	4943 (89%)	9836 (95%)
	7	**1449** (74%)	4792 (92%)	9685 **(96%)**
High	**3**	3483 **(31%)**	6826 (65%)	**11719** (79%)
	5	2519 (42%)	5862 (75%)	10755 (86%)
	7	2106 (51%)	5449 (81%)	10342 (90%)

Table 5. Total carbon footprint of the IRISCAST resources over the snapshot period for a range of scenarios. The scenarios are summarised in Table 4

example some Dell BMC's can be queried over IPMI with the freeipmi command `ipmi-oem dell get-power-consumption-data`.

Having made measurements and collected inventories the carbon model discussed in section 2 could be run. It proved difficult to obtain precise values for embodied carbon costs. As a result the model was used to evaluate carbon costs under the range of estimates detailed in Table 4. The carbon model results are shown in in Table 5. It can be seen that under the scenario with the least favorable estimates the modeled carbon footprint of the IRISCAST 24 hour snapshot is put at 11719 $KgCO_2$, while under the most favorable scenario the footprint is 1449 $KgCO_2$. This is an order of magnitude different and shows that there is a great potential to reduce the carbon footprint of Digital Research Infrastructures.

5 IRISCAST Recommendations

Having built a community through two workshops and having made measurements and models and analysed what was learned IRISCAST essentially made seven recommendations in two groups[2]. To enable high-level feedback to inform decisions at the strategic and policy levels on the timescale of months and years IRISCAST proposed the following.

1. Future DRI procurement to include a score based on embedded carbon costs and equipment energy usage.
2. New computer hardware to include energy measurement capability such as IPMI (or per port PDUs) and require the supplier to provide best estimates of embedded carbon costs.
3. Measure energy used by the cooling infrastructure and the computing infrastructure.
4. Facilities to keep an inventory of equipment including embedded carbon cost and idle power draw.
5. Monthly (or other periodic) reporting of carbon usage by facilities based on 3 and 4 above. Incorporate these reports into the standard grant reporting regime.

To enable low-level feedback to end users and user communities to inform tactical and operational decisions on the time scale of hours to weeks IRISCAST proposed the following.

6. Collect per job (or VM) energy usage by using tools like Slurm (correctly configured). Combine this with embedded carbon from inventory and electricity carbon intensity to feedback job carbon cost to the end user to drive improvements in user code and workflow.
7. Identify user communities and the authors of community codebases so that useful feedback can be given to them to drive the development of more carbon efficient code and workflows.

6 Acknowlegements

The IRISCAST project was: lead by Jon Hays (QMUL), Nic Walton (Cambridge),Adrian Jackson (Edinburgh) and Alison Packer (STFC); and staffed by Alex Owen (QMUL), Alex Ogden (Cambridge), Anish Mudaraddi (STFC); and funded by the UKRI NetZero DRI Scoping Project [1]. IRISCAST was supported by volunteer effort at the DRI resources that took part in the IRISCAST audit, namely: Dan Traynor (QMUL), Derek Ross (STFC), Alexander Dibbo (STFC), Jon Roddom (STFC), Martin Summers (STFC), Jacob Ward (STFC), Dan Whitehouse (Imperial) and Alastair Basden (Durham).

References

[1] M. Juckes, M. Bane, J. Bulpett, K. Cartmell, M. MacFarlane, M. MacRae, A. Owen, C. Pascoe, P. Townsend, Tech. rep., UKRI (2023), this work is funded by the UKRI Digital Research Programme on grant NERC (NE/W007134/1). The project website is net-zero-dri.ceda.ac.uk/., https://doi.org/10.5281/zenodo.8199984

[2] J. Hays, N. Walton, A. Jackson, A. Mudaraddi, A. Packer, R. Owen, Tech. rep., UKRI (2023), https://doi.org/10.5281/zenodo.7692451

[3] Intel 64 and IA-32 Architectures Software Developer's Manual Volume 3B: System Programming Guide, Part 2, (2023), https://www.intel.com/content/www/us/en/developer/articles/technical/intel-sdm.html