Sustainable computing with RF2.0 at DESY

Dr Dwayne Spiteri, Konrad Kockler et al

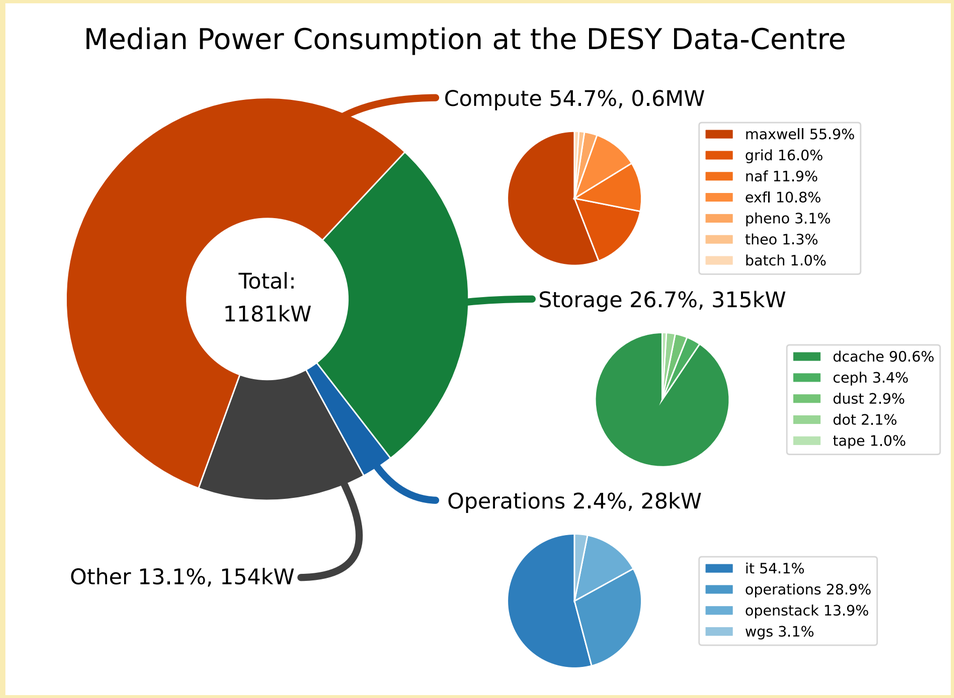
WLCG Environmental Sustainability Workshop - 11/12/2024

RF2.0 RESEARCH FACILITY
TOWARDS A MORE ENERGY-EFFICIENT AND SUSTAINABLE PATH

DESY

DESY Data Centre

- WLCG Tier2 with Over 3000 machines. ~1500 machines for compute (~50k HPC Cores and ~32k HTC cores) and ~1000 (~165PB) for storage
- The DESY Data centre is rated for ~1.6MW.
- Most of the power usage is driven by compute



DESY Data Centre

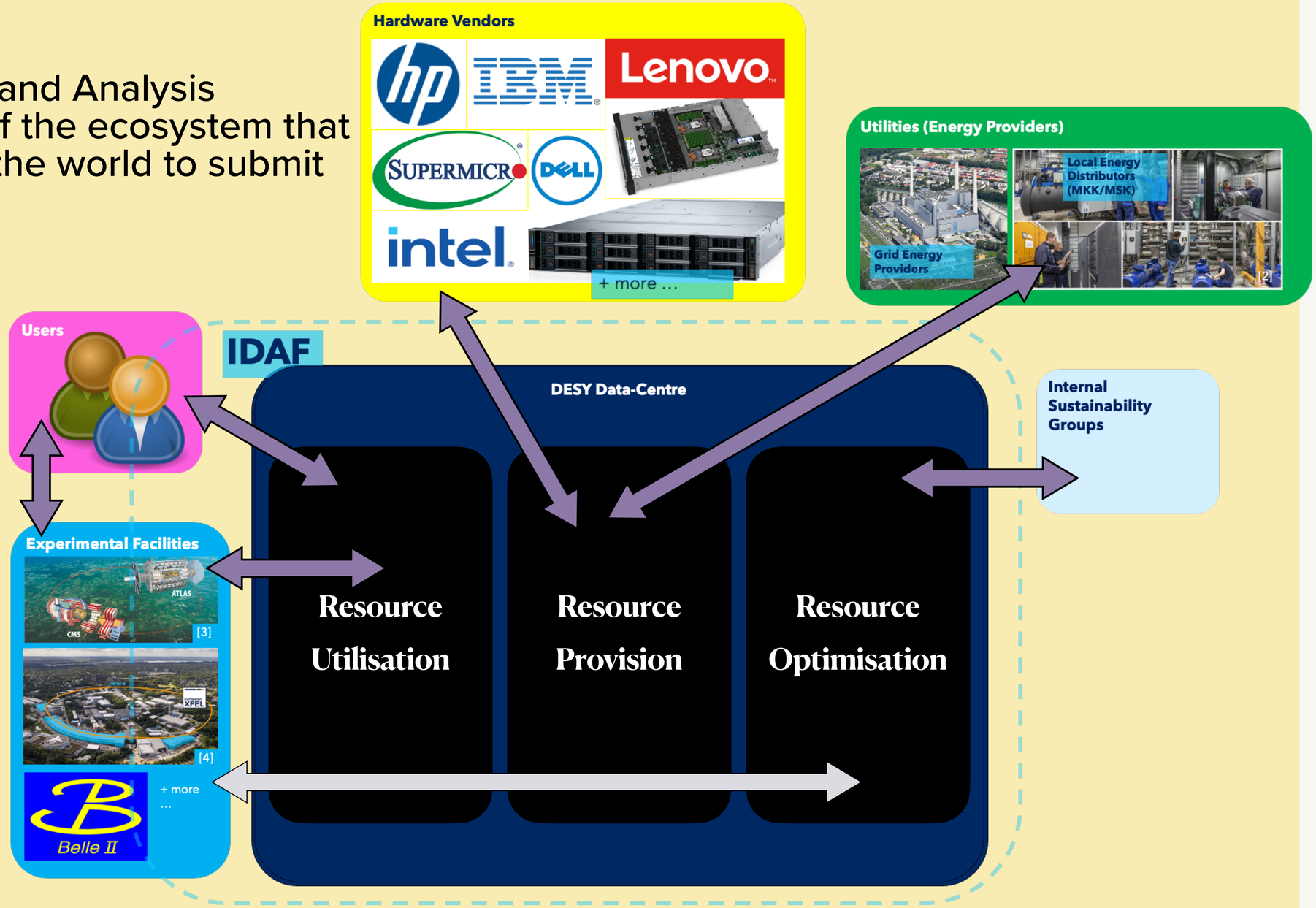
- Future data-centres will be even more power-hungry and are likely to form part of more energy-intensive ecosystems
- Data-centres, are large complex and have lots of moving parts

The IDAF at DESY

- The Interdisciplinary Data and Analysis Facility (IDAF) forms part of the ecosystem that allows users from around the world to submit scientific work to DESY

- While the data-centre is at it's heart, sustainability efforts will be limited if the parts of this wider ecosystem don't talk to each other

- More sustainable future experiments → Research Facility 2.0

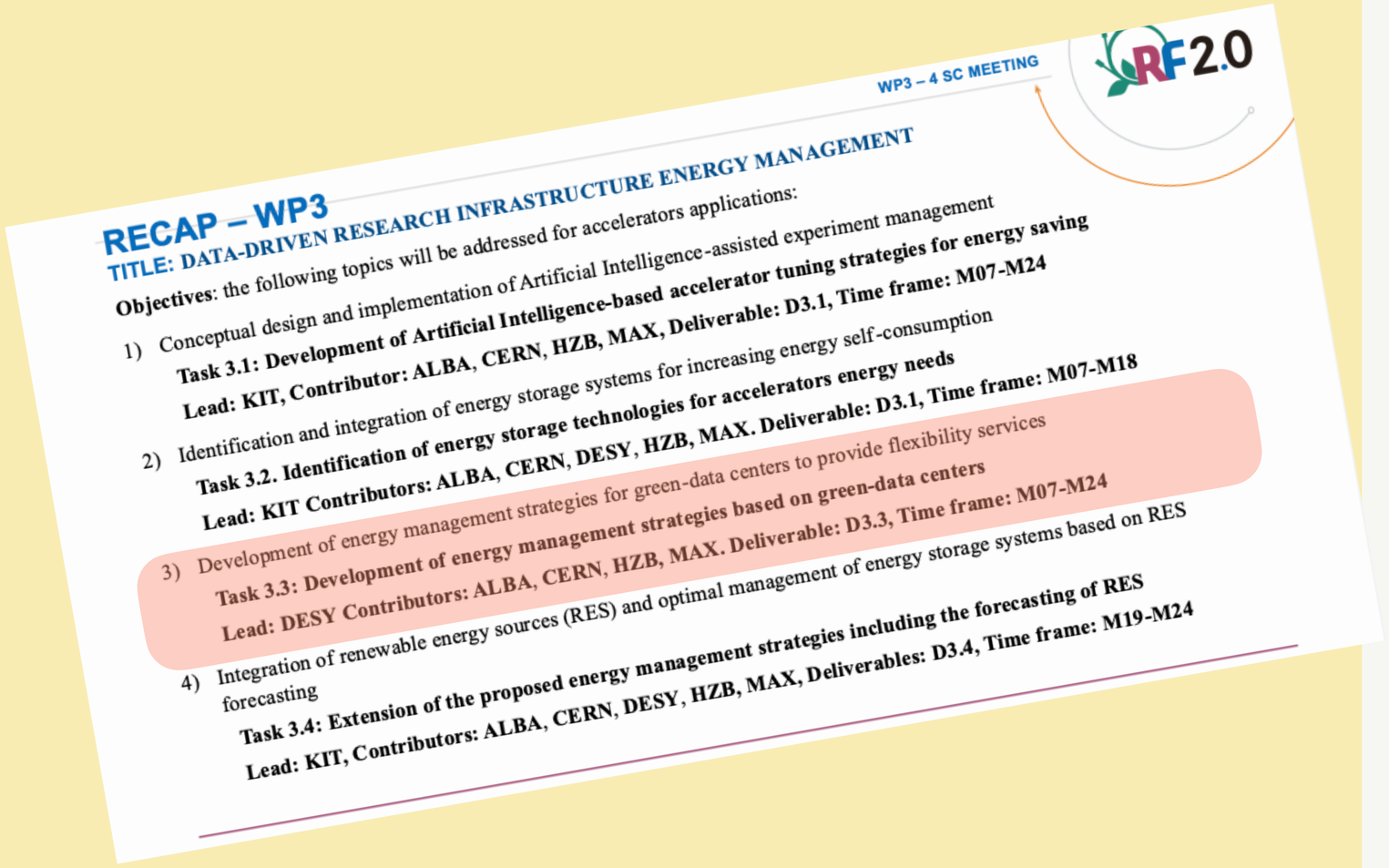


Research Facility 2.0

- An EU-funded project whose remit covers the design and use of technologies for use at future accelerators; **and the approaches we can take to manage energy at support infrastructures**

- Part of one of the work packages is energy management at Research infrastructures, and **DESY is the only institution on this project** looking to develop strategies for the energy management of “green” data-centres

- My contribution:
  - Create a digital twin of the DESY data-centre and use that to try and investigate energy/carbon saving strategies



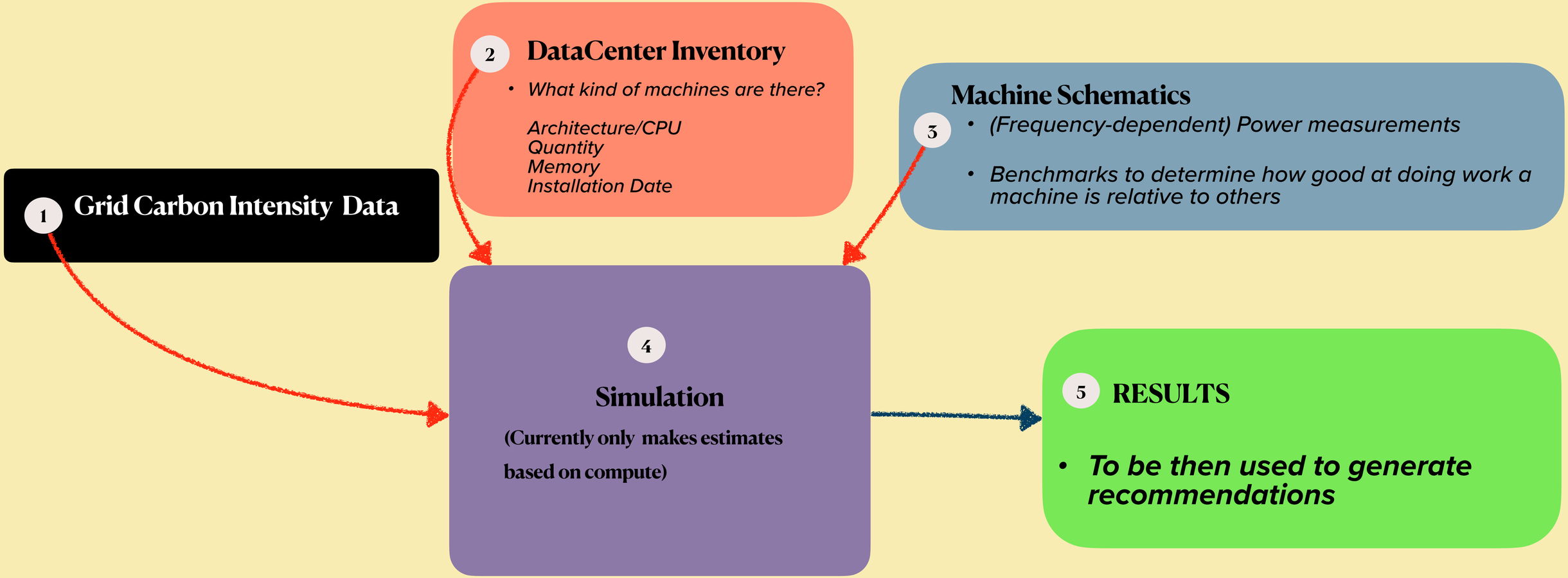
The Data-Centre Simulation Framework

- Initially created at the University of Glasgow - Currently being expanded using RF2.0 funding

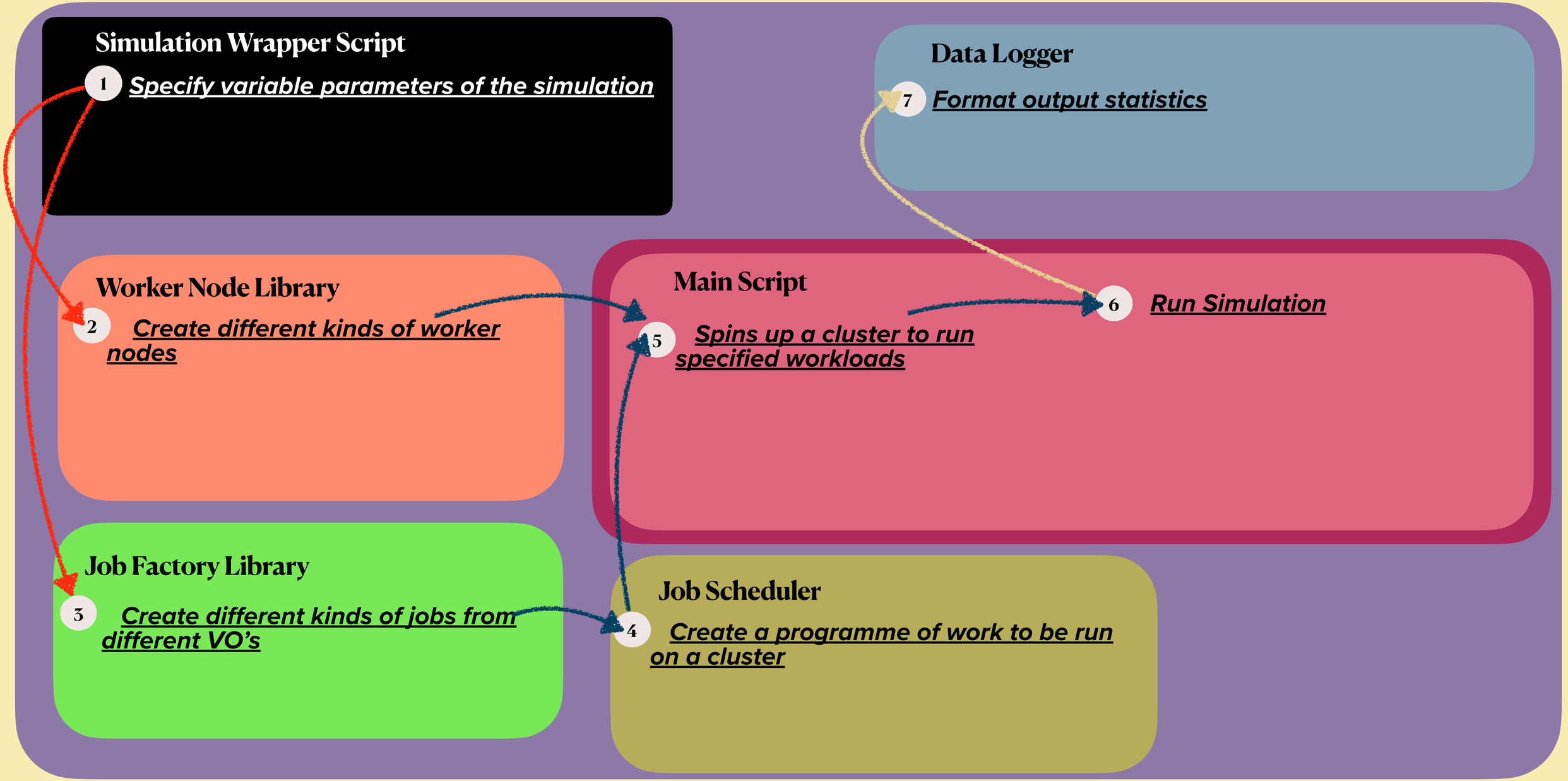
ScotGrid Scottish Grid Service

University of Glasgow

- Mainly aimed at simulating data-centre compute and outputting carbon usage data



Simple Simulation Schematic



Simple Simulation Schematic

Simulation Wrapper Script

1 ***Specify variable parameters of the simulation***

- The number and type of nodes your cluster is made from (ampere, dell, grace)
- The amount of starting jobs and how many jobs are submitted per hour
- Maximum length of the simulation

Data Logger

7 ***Format output statistics***

Worker Node Library

2 ***Create different kinds of worker nodes***

Main Script

5 ***Spins up a cluster to run specified workloads***

6 ***Run Simulation***

Job Factory Library

3 ***Create different kinds of jobs from different VO's***

Job Scheduler

4 ***Create a programme of work to be run on a cluster***

Simple Simulation Schematic

Simulation Wrapper Script

1 *Specify variable parameters of the simulation*

- The number and type of nodes your cluster is made from (ampere, dell, grace)
- The amount of starting jobs and how many jobs are submitted per hour
- Maximum length of the simulation

Data Logger

7 *Format output statistics*

Worker Node Library

2 *Create different kinds of worker nodes*

- Different types of worker node Attributes like hostnames, cores, memory, max power consumed, frequency
- Formulas for scaling power consumption
- Methods for automatically clocking up and down nodes
- Updates with whether the job is finished per timestep

Main Script

5 *Spins up a cluster to run specified workloads*

6 *Run Simulation*

Job Factory Library

3 *Create different kinds of jobs from different VO's*

Job Scheduler

4 *Create a programme of work to be run on a cluster*

Simple Simulation Schematic

Simulation Wrapper Script

1 ***Specify variable parameters of the simulation***

- The number and type of nodes your cluster is made from (ampere, dell, grace)
- The amount of starting jobs and how many jobs are submitted per hour
- Maximum length of the simulation

Data Logger

7 ***Format output statistics***

Worker Node Library

2 ***Create different kinds of worker nodes***

- Different types of worker node Attributes like hostnames, cores, memory, max power consumed, frequency
- Formulas for scaling power consumption
- Methods for automatically clocking up and down nodes
- Updates with whether the job is finished per timestep

Main Script

5 ***Spins up a cluster to run specified workloads***

6 ***Run Simulation***

Job Factory Library

3 ***Create different kinds of jobs from different VO's***

- Assume jobs run for samples amount of time drawn from previously measured distributions (for testing all jobs are set to be 5hrs long)
- Require amounts of memory and cores to be used

Job Scheduler

4 ***Create a programme of work to be run on a cluster***

Simple Simulation Schematic

Simulation Wrapper Script

1 ***Specify variable parameters of the simulation***

- The number and type of nodes your cluster is made from (ampere, dell, grace)
- The amount of starting jobs and how many jobs are submitted per hour
- Maximum length of the simulation

Data Logger

7 ***Format output statistics***

Worker Node Library

2 ***Create different kinds of worker nodes***

- Different types of worker node Attributes like hostnames, cores, memory, max power consumed, frequency
- Formulas for scaling power consumption
- Methods for automatically clocking up and down nodes
- Updates with whether the job is finished per timestep

Main Script

5 ***Spins up a cluster to run specified workloads***

6 ***Run Simulation***

Job Factory Library

3 ***Create different kinds of jobs from different VO's***

- Assume jobs run for samples amount of time drawn from previously measured distributions (for testing all jobs are set to be 5hrs long)
- Require amounts of memory and cores to be used

Job Scheduler

4 ***Create a programme of work to be run on a cluster***

- Initialises jobs from ones requested from types of ones available
- Updates with jobs to be submitted to the cluster per time-step

Simple Simulation Schematic

Simulation Wrapper Script

1 ***Specify variable parameters of the simulation***

- The number and type of nodes your cluster is made from (ampere, dell, grace)
- The amount of starting jobs and how many jobs are submitted per hour
- Maximum length of the simulation

Data Logger

7 ***Format output statistics***

Worker Node Library

2 ***Create different kinds of worker nodes***

- Different types of worker node Attributes like hostnames, cores, memory, max power consumed, frequency
- Formulas for scaling power consumption
- Methods for automatically clocking up and down nodes
- Updates with whether the job is finished per timestep

Main Script

6 ***Run Simulation***

5 ***Spins up a cluster to run specified workloads***

- Defines things like amount of memory, cores available to outside sources from input worker nodes
- Define how you run the cluster in the event you want to try and run it differently - clock down nodes at certain times of day for example

Job Factory Library

3 ***Create different kinds of jobs from different VO's***

- Assume jobs run for samples amount of time drawn from previously measured distributions (for testing all jobs are set to be 5hrs long)
- Require amounts of memory and cores to be used

Job Scheduler

4 ***Create a programme of work to be run on a cluster***

- Initialises jobs from ones requested from types of ones available
- Updates with jobs to be submitted to the cluster per time-step

Simple Simulation Schematic

Simulation Wrapper Script

1 ***Specify variable parameters of the simulation***

- The number and type of nodes your cluster is made from (ampere, dell, grace)
- The amount of starting jobs and how many jobs are submitted per hour
- Maximum length of the simulation

Data Logger

7 ***Format output statistics***

Worker Node Library

2 ***Create different kinds of worker nodes***

- Different types of worker node Attributes like hostnames, cores, memory, max power consumed, frequency
- Formulas for scaling power consumption
- Methods for automatically clocking up and down nodes
- Updates with whether the job is finished per timestep

Main Script

5 ***Spins up a cluster to run specified workloads***

- Defines things like amount of memory, cores available to outside sources from input worker nodes
- Define how you run the cluster in the event you want to try and run it differently - clock down nodes at certain times of day for example

6 ***Run Simulation***

- Calculates the total power used and CO2e emitted per timestep (10 minutes)
- Takes Jobs from the scheduler if able
- Passes data from the worker nodes to the DataLogger
- Ends when you run out of work, or out of time

Job Factory Library

3 ***Create different kinds of jobs from different VO's***

- Assume jobs run for samples amount of time drawn from previously measured distributions (for testing all jobs are set to be 5hrs long)
- Require amounts of memory and cores to be used

Job Scheduler

4 ***Create a programme of work to be run on a cluster***

- Initialises jobs from ones requested from types of ones available
- Updates with jobs to be submitted to the cluster per time-step

Simple Simulation Schematic

Simulation Wrapper Script

1 ***Specify variable parameters of the simulation***

- The number and type of nodes your cluster is made from (ampere, dell, grace)
- The amount of starting jobs and how many jobs are submitted per hour
- Maximum length of the simulation

Data Logger

7 ***Format output statistics***

- Total (and average): CPU used, time elapsed, jobs started/ completed, (peaktime) power used and estimated C02e emissions.

Worker Node Library

2 ***Create different kinds of worker nodes***

- Different types of worker node Attributes like hostnames, cores, memory, max power consumed, frequency
- Formulas for scaling power consumption
- Methods for automatically clocking up and down nodes
- Updates with whether the job is finished per timestep

Main Script

5 ***Spins up a cluster to run specified workloads***

- Defines things like amount of memory, cores available to outside sources from input worker nodes
- Define how you run the cluster in the event you want to try and run it differently - clock down nodes at certain times of day for example

6 ***Run Simulation***

- Calculates the total power used and CO2e emitted per timestep (10 minutes)
- Takes Jobs from the scheduler if able
- Passes data from the worker nodes to the DataLogger
- Ends when you run out of work, or out of time

Job Factory Library

3 ***Create different kinds of jobs from different VO's***

- Assume jobs run for samples amount of time drawn from previously measured distributions (for testing all jobs are set to be 5hrs long)
- Require amounts of memory and cores to be used

Job Scheduler

4 ***Create a programme of work to be run on a cluster***

- Initialises jobs from ones requested from types of ones available
- Updates with jobs to be submitted to the cluster per time-step

Data Logger

7 ***Formats output statistics***

- Total (and average): CPU used, time elapsed, jobs started/ completed, (peaktime) power used and estimated CO2e emissions.

```
========
Summmary
========

Total Simulated-time Duration       :  5.4 days
Total Real-time Duration            : 10.2 minutes

Jobs Started                        :  50000
Jobs Finished                       :  50000

Total CPU duration                  : 2000000.0 hours
Average CPU duration                : 5.00 hours

Total energy consumed by compute    : 1428.75 kWh
Peaktime (5-9pm) energy consumption: 256.65 kWh
Average energy consumption per job : 28.57 Wh

Estimated CO2e emmissions           : 112.188 kg
Estimated Peaktime CO2e emmissions : 21.009 kg
Average CO2e emmissions per job     : 2.244 g
Peaktime CO2e emmissions percentage: 18.726 %
```

Output

- Data here is an example from when I ran on the Glasgow Data-centre
- Each time the simulation is called, a file gets produced with the following information

Simulated and Real-time duration of the simulation

Job information

Total and Average CPU duration

Estimated energy used in total, during peak times and job-average

Estimated CO_2 (e)quivalent emissions for said work

Use Case 1 - Can you save carbon by shifting work?

- Insert jobs to run for 7 days of simulated time. Do you save carbon by clocking down nodes when the carbon intensity of the grid is forecast to be high?

No Changes

```
========
Summary
========

Total Simulated-time Duration      : 168.0 hours
Total Real-time Duration           : 156.0 minutes

Jobs Started                       : 466536
Jobs Finished                      : 450576

Total CPU duration                 : 2285107.9 hours
Average CPU duration               : 4.90 hours

Total energy consumed by compute   : 10339.39 kWh
Peak time (5-9pm) energy consumption: 1649.79 kWh
Average energy consumption per job : 22.55 Wh

Estimated CO2e emissions           : 688.678 kg
Estimated Peak time CO2e emissions : 118.386 kg
Average CO2e emissions per job     : 1.502 g
Peak time CO2e emissions percentage: 17.190 %
```

17% reduction in jobs

25% peak time energy reduction

20% overall CO_2 reduction

Forecasted Clock-down

Each job uses 3% less CO_2

```
========
Summary
========

Total Simulated-time Duration      : 168.0 hours
Total Real-time Duration           : 174.3 minutes

Jobs Started                       : 392392
Jobs Finished                      : 376432

Total CPU duration                 : 2313757.3 hours
Average CPU duration               : 5.90 hours

Total energy consumed by compute   : 8613.15 kWh
Peak time (5-9pm) energy consumption: 1243.06 kWh
Average energy consumption per job : 22.41 Wh

Estimated CO2e emissions           : 560.153 kg
Estimated Peak time CO2e emissions : 88.647 kg
Average CO2e emissions per job     : 1.457 g
Peak time CO2e emissions percentage: 15.825 %
```

Use Case 2 - What do different procurements look like?

- An example type of recommendation - Running fixed work of 50,000 jobs, what new machines will lower your impact? (Same number of new cores each)

No Changes (With Old Kit)

```
Total Simulated-time Duration      : 27.8 hours
Total Real-time Duration           :  1.0 minutes

Jobs Started                       : 50000
Jobs Finished                      : 50000

Total CPU duration                 : 250451.5 hours
Average CPU duration               : 5.01 hours

Total energy consumed by compute   : 1362.10 kWh
Peaktime (5-9pm) energy consumption: 292.48 kWh
Average energy consumption per job : 27.24 Wh

Estimated CO2e emmissions          : 94.188 kg
Estimated Peaktime CO2e emmissions : 19.462 kg
Average CO2e emmissions per job    : 1.884 g
Peaktime CO2e emmissions percentage: 20.663 %
```

30% CO_2 reduction

Replacing older nodes w/ x86 - AMD Siena

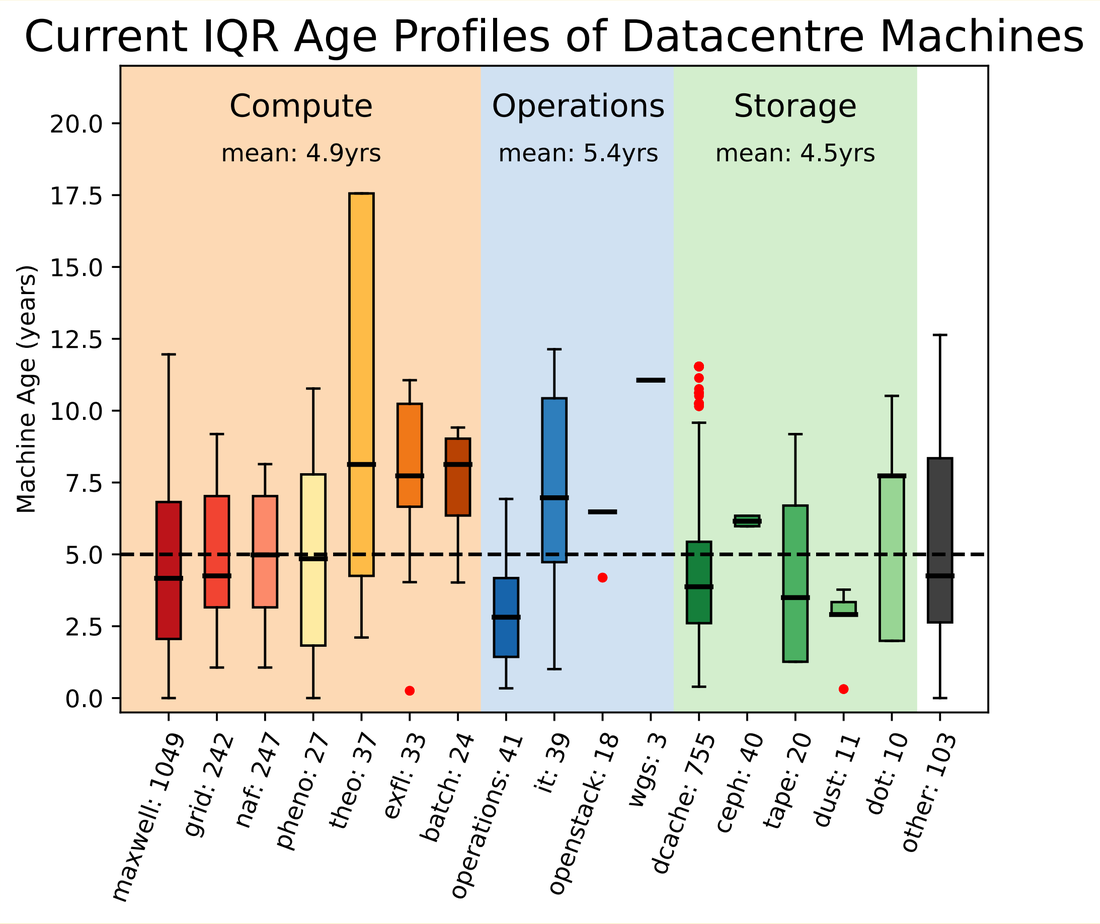
```
Total Simulated-time Duration      : 20.0 hours
Total Real-time Duration           :  0.6 minutes

Jobs Started                       : 50000
Jobs Finished                      : 50000

Total CPU duration                 : 259273.7 hours
Average CPU duration               : 5.19 hours

Total energy consumed by compute   : 969.80 kWh
Peaktime (5-9pm) energy consumption: 211.61 kWh
Average energy consumption per job : 19.40 Wh

Estimated CO2e emmissions          : 66.048 kg
Estimated Peaktime CO2e emmissions : 13.810 kg
Average CO2e emmissions per job    : 1.321 g
Peaktime CO2e emmissions percentage: 20.909 %
```

32% CO_2 reduction

Replacing older nodes w/ ARM - AltraMax M128-30

```
Total Simulated-time Duration      : 18.0 hours
Total Real-time Duration           :  0.5 minutes

Jobs Started                       : 50000
Jobs Finished                      : 50000

Total CPU duration                 : 252801.8 hours
Average CPU duration               : 5.06 hours

Total energy consumed by compute   : 939.53 kWh
Peaktime (5-9pm) energy consumption: 217.55 kWh
Average energy consumption per job : 18.79 Wh

Estimated CO2e emmissions          : 63.599 kg
Estimated Peaktime CO2e emmissions : 14.197 kg
Average CO2e emmissions per job    : 1.272 g
Peaktime CO2e emmissions percentage: 22.323 %
```

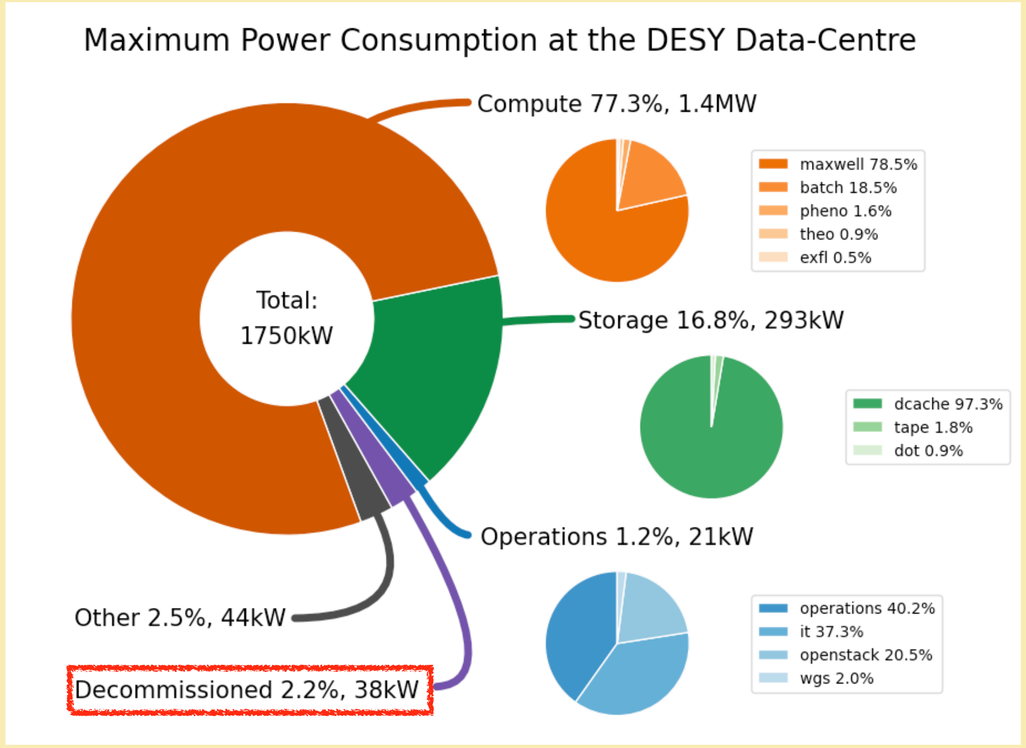
Other Ecosystem Improvements at DESY



Life cycle analysis of machines

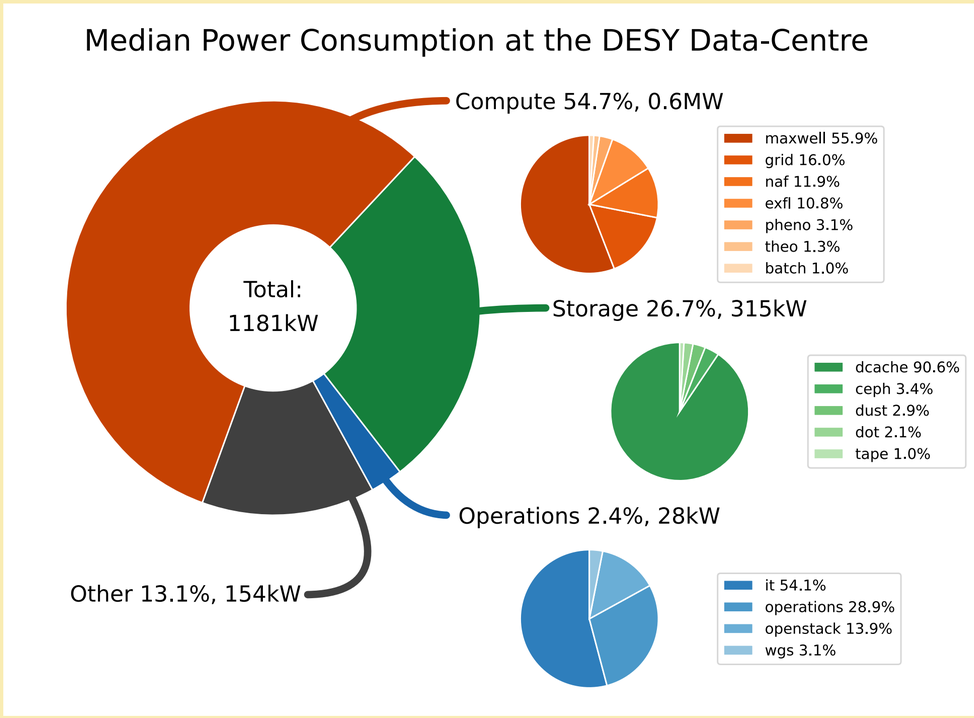
- Is it carbon Cost effective to recycle machines to other types of datacenter service?
- When and how can we decommission machines?

Other Ecosystem Improvements at DESY



<u>Downscaling/Slow Decommissioning</u>

- We don't operate close to maximum capacity



- Turn off the oldest machines, but don't get rid of them. Turn them on when overall usage is high

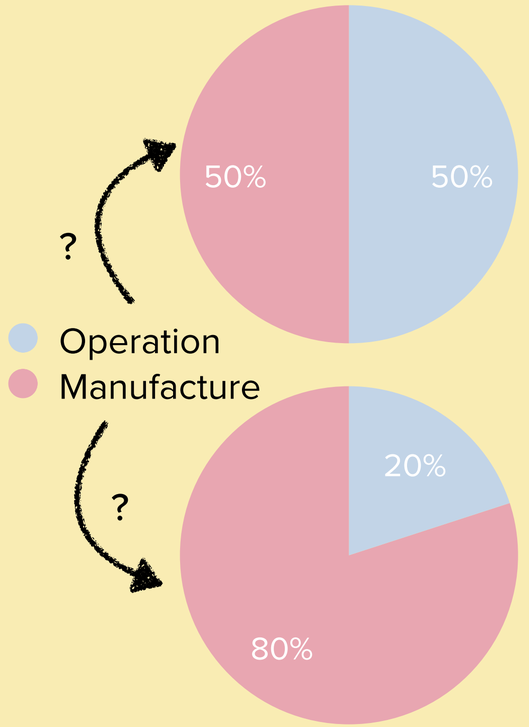
Other Ecosystem Improvements at DESY

Liasing with Local Energy Providers

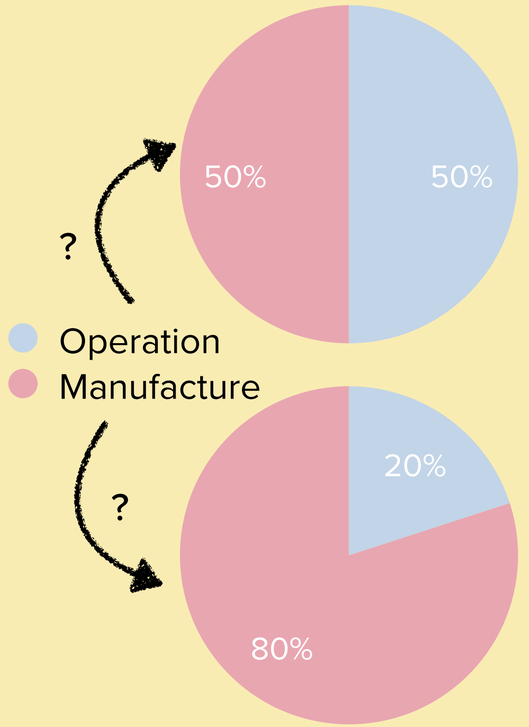
- Sustainability is more than $C0_2$. Water usage and other resources (like land) are important. Water will become increasingly more important in the future
- Liaising with local energy/water distribution MKK group at DESY
  - Dynamic Energy Loads (follow a signal - save costs for turning off green energy £1B wasted in UK last year)
  - Minimise Water Wastage

Embedded Carbon

- A significant component of carbon in a servers lifetime is in the embedded carbon
- Need to start pressuring hardware vendors to give us or produce some carbon lifecycle analyses - Procurement?
- The improvements listed are only on the carbon opportunity cost of RUNNING work. Assume an total operational carbon cost of Y and an embedded carbon cost of X



Embedded Carbon



- Attribute it all to purchase and treat operational carbon as independent?
  Total Carbon 2025 = Y(2025) -> Run in a way that reduces carbon
- Assume a set lifetime of operation (5 years) and split the cost for each year - X/5?
  Total Carbon 2025 = X/5 + Y(2025) -> Optimisations for Y(2025) less impactful
- Split the embedded carbon cost over every job you run?
  Total Carbon 2025 = X(2025) + Y(2025) -> Reduction in Jobs wastes embedded carbon
- Or completely separate the two and have a carbon budget for the data-centre itself to account for embedded carbon

Conclusions and Future Work

- A simulation framework has been created to try and test different kinds of operation of various data-centres. The simulation framework is currently private, but the plan is to make it freely available soon
- DESY will use it alongside its current strategies to help evaluate their effectiveness, and use it to define new ones
- Currently generates outputs based off compute information - but this is the largest component to every use of most data-centres. Storage will be looked at in the future
- The improvements listed are only on the carbon opportunity cost of **RUNNING** work. Improvements will be tempered by how we treat embedded carbon in the future
- It’s clear that there’s not much we can do to have a large impact without talking to external partners
- Hopefully the bigger badder AI data centres of the future can use some of these techniques to make them better

RF2.0

KIT Karlsruhe Institute of Technology

ALBA

CERN

DESY.

HZB Helmholtz Zentrum Berlin

MAX IV

COMMTIA

Zaphiro

Cryoelectra

ELYTT ENERGY

For further information and to follow our project progress visit www.rf20.eu

and our Social Media accounts: RF2.0 Project @rf20_project

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101131850 and from the Swiss State Secretariat for Education, Research and Innovation (SERI).

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI