Modelling Scenarios for Carbon-Aware Geographic Load Shifting of Compute Workloads

Wim Vanderbauwhede
School of Computing Science
University of Glasgow
Glasgow, UK

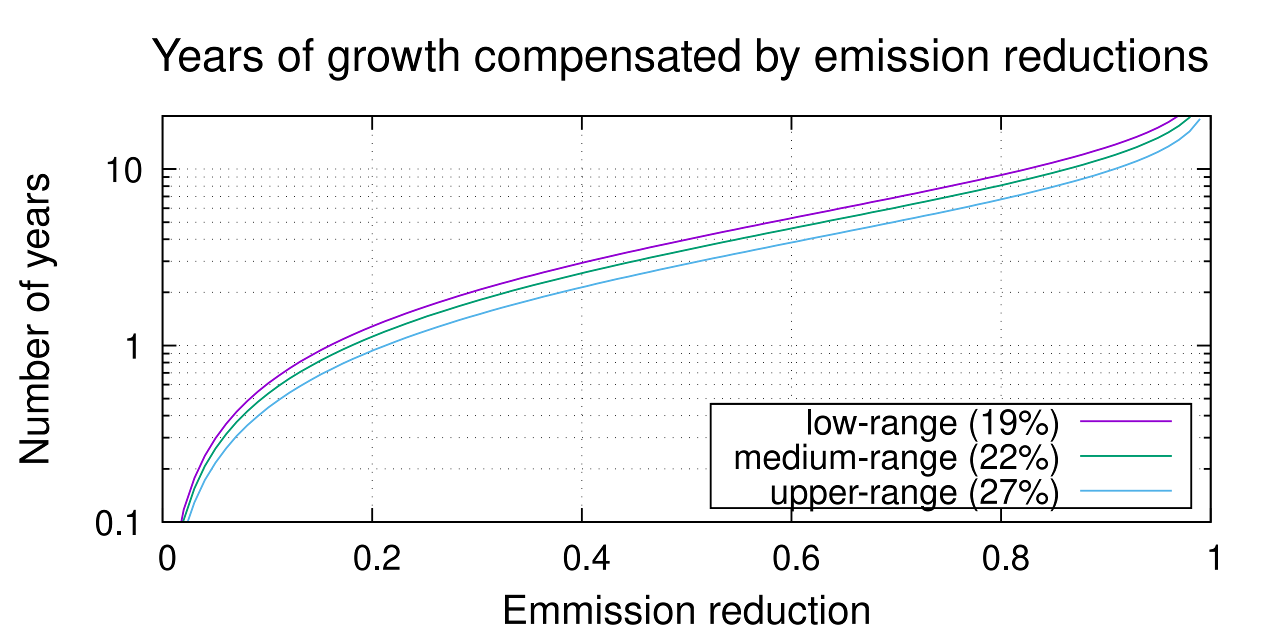
SC4RC, 7 May 2026, CERN

Context: Emissions from AI data centres

- Greenhouse gas emissions from ICT estimated at 4% & rising steeply
- Estimates from before popularity of generative AI – now main driver for growth in emissions from ICT
- International consulting firm McKinsey projects global growth in data centres until 2030 of between 19% and 27% annually
- Extrapolating their medium scenario of 22% growth to 2040 corresponds to a 20× increase in data centre energy consumption – far larger than the projected 4× by Belkhir

Context: Emissions from AI data centres

Any approach which reduces data centre emissions by a fixed amount compensates for a continued time of growth



If emissions reduced by 10% per year: reduction undone in less than a year (medium scenario, CAGR=22%)

$$\text{years}(r) = -\log(1-r)/\log(1+\text{CAGR})$$

Aim

- To model the reduction in overall data centre emissions from shifting work from a data centre in a high-emission region to one in a low-emission region
- Even without constraints on the grid ("perfect" grid = optimal amount of energy can always be delivered anywhere):

  What is the expected level of emission reductions from deploying geographic load shifting, assuming realistic data centres parameters?

Contributions

- A model to assess the extent of reduction in emissions that can result from geographic load shifting
- A number of scenarios, both for commercial AI data centres and HPC centres

Contributions

The model improves on the state of the art:

- Both embodied CO_2e emissions and operational CO_2e emissions calculated based on up-to-date, detailed & rigorous Life Cycle Analysis models
- Combines calculated emission values into a novel, simple but expressive analytical model that allows to calculate the emissions reduction arising from geographic load shifting
- Considers not only the maximum theoretically possible gains but focuses on realistic scenarios

Carbon-aware computing

- Running compute workloads when and where the electricity grid is powered by renewable energy
- Purpose is to reduce emissions from data centre based compute activities
- Requires shifting work in time (temporal) and/or space (geographic)
- Focus of this work is on geographic load shifting

Model for emissions reduction of geographic load shifting

Data centre emission contributions

- Embodied carbon incurred during server manufacturing process and infrastructure building
- Operational emissions based on:
  - Power consumption
  - Electricity generation carbon intensity (CI) of the regions

Model assumptions

The model is not grid-aware: assumes sufficient power in target region to power data centre at full capacity

- Not a fundamental restriction: available power only one factor determining available capacity at target site
- Reduced power availability means less compute capacity
  - Modeling power availability through load of data centre
  - Modeling reduced availability of renewables through CI

Model assumptions

The model conceptually simple:

- Only 2 site types (“high-emission” & “low-emission”)
- Compute embodied carbon emissions and operational emissions for each site assuming sites are identical in specification (twin data centres)
- Embodied emissions and operational emissions calculated using our separate LCA model

Model parameters

n_{nodes} Number of nodes in data centre

c_{emb} Embodied carbon of data centre ($kgCO_2e/y$, expressed per node)

c_{hi} Operational carbon emissions of high-emission site ($kgCO_2e/y$, per node, includes PUE)

c_{lo} Operational carbon emissions of the low-emission site ($kgCO_2e/y$, per node, includes PUE)

Model parameters

λ_{hi} Server load factor of high-emission site, $0 \leq \lambda_{hi} \leq 1$

λ_{lo} Server load factor of high-emission site, $0 \leq \lambda_{lo} \leq 1$

γ Idle power consumption as fraction of active power consumption, $0 \leq \gamma \leq 1$
(assumed to be the same for all sites)

Model parameters for geographic load shifting

- α Fraction of workload that can be moved, $0 \leq \alpha \leq 1$
- β Fraction of time workload can be moved, $0 \leq \beta \leq 1$.
- η Overhead factor for emissions incurred because of geographic load shifting, $0 \leq \eta \ll 1$, on a per-node basis

Total emissions from all nodes:

$$C_{\{emb,hi,lo\}} = c_{\{emb,hi,lo\}} \cdot n_{nodes} \quad (kgCO_2e/y)$$

Embodied emissions

LCA model to calculate embodied carbon C_{emb}:

- Our re-implementation of Boavizta model
- Updated estimates for parameters:
  - Chips: ACT methodology (Architectural Carbon modelling Tool): embodied emissions of manufacturing process & materials combined with die size, with updated parameters
  - Extended model to include non-integrated GPUs
- With embodied emissions of facility (building & furnishings, power & cooling) & networking equipment

Operational emissions

Per node, incl. network infrastructure overhead & cooling etc.

- Pnode: node power consumption
- E_{node}: year-averaged node energy consumption

$$E_{node} = P_{node} \times 24 \times 365/1000 \text{ (kWh)}$$

$$C_{\{hi,lo\}}(t) = E_{node} \cdot n_{nodes} \cdot (1 + \nu) \cdot PUE \cdot CI_{\{hi,lo\}}(t) \text{ (kgCO}_2\text{e)}$$

Operational emissions per site

$$C_{op,\{hi,lo\}}(t) = C_{\{hi,lo\}}(t) \cdot (\lambda_{\{hi,lo\}}(t) + \gamma \cdot (1 - \lambda_{\{hi,lo\}}(t)))$$

Total emissions per site

$$C_{\{hi,lo\}} = C_{op,\{hi,lo\}} + C_{emb}$$

Rationale for model construction

Model is linear: equations do not contain non-linear terms for any parameters

Per-site expressions:

- Embodied emissions are constant
- Total emissions of the form

$$C_{tot}(t) = A \cdot \lambda(t) \cdot C(t) + B \cdot C(t) + D$$

with $A = 1 - \gamma$, $B = \gamma$, $D = C_{emb}$

- Load $\lambda(t)$ and emissions $C(t)$ statistically independent

Rationale for model construction

- CI of electricity generation not influenced by load
- Unless data centre uses temporal shifting, load not influenced by CI either
  - Load and CI values are time series but total emissions at given time independent of values at other times
  - Treat $\lambda(t)$ and $C(t)$ as independent random variables
  - Use mean (expected value) = time-integrated values of load and emissions over one year

Rationale for model construction

For statistically independent random variables, regardless of distributions, expected values relate as:

$E(XY) = E(X) \cdot E(Y)$ and

$E(aX+bY+c) = A \cdot E(X) + B \cdot E(Y) + D$ (A, B and D constants)

Therefore, total emissions:

$$E(C_{tot}(t)) = A \cdot E(\lambda(t)) \cdot E(C(t)) + B \cdot E(C(t))$$

Or (notation) $C_{tot} = A \cdot \lambda \cdot C + B \cdot C + D$

Year-averaged data centre emissions: no need to calculate emissions at every time step, use averages instead of traces

Baseline emissions model

$$C_{base} = 2 \cdot C_{emb} + (\lambda_{hi} + (1 - \lambda_{hi}) \cdot \gamma) \cdot C_{hi} + (\lambda_{lo} + (1 - \lambda_{lo}) \cdot \gamma) \cdot C_{lo}$$

- Assume data centre has excess capacity that can be used for geographic load shifting, when no peak loads
- Excess capacity via load parameter λ and idle power consumption parameter γ

Geographic load shifting model

- Use some free capacity at low-emission site by moving load from high-emission site
- Parameters α and β represent shifted workload
  - α: constraints on workload
  - β: proportion of time load can be shifted
- When no shifting: emissions given by baseline model
- Yearly averages for α, β and γ without loss of generality

Geographic load shifting model

- At most $(1-\lambda_{lo})$ free capacity at low-emission site
- Shiftable load limited by freed-up capacity on high-emission side
  - When free capacity < load to shift: cap α at free capacity to not exceed free load
  - When free capacity > load of high-emission data centre: cap α to not shift more work than available

Geographic load shifting model

$$\alpha_{eff} = \begin{cases} \alpha & , \alpha \cdot \lambda_{hi} < 1 - \lambda_{lo} \\ (1 - \lambda_{lo}) / \lambda_{hi} & , \alpha \cdot \lambda_{hi} \geq 1 - \lambda_{lo} \end{cases}$$

$$\alpha_{eff} =$$

- Low side load: increases from λ_{lo} to $\lambda_{lo} + \alpha_{eff}.\lambda_{hi}$
- High side load: reduces from λ_{hi} to $\lambda_{hi}\,(1 - \alpha_{eff})$

Geographic load shifting model

Example: High side load: 0.9 / Low side load: 0.7

- Total load 1.6
- Move 20% of the work 50% of the time
  - Average load on high side: 0.9.(1.0 - 0.2) = 0.72
  - Average load on low side: 0.7+0.9×0.2 = 0.88
- Total remains 1.6

$$\lambda_{hi}(1-\alpha_{eff})+(\lambda_{lo}+\alpha_{eff}\cdot\lambda_{hi})=\lambda_{hi}+\lambda_{lo}$$

Geographic load shifting model

If several data centres in the high and low emission zones:

introduce parameters n_{hi} and n_{lo} (number of sites)

$$\alpha_{eff} = \begin{cases} \alpha & , \alpha \cdot n_{hi} \cdot \lambda_{hi} < n_{lo} \cdot (1 - \lambda_{lo}) \\ n_{lo} \cdot (1 - \lambda_{lo}) \,/\, n_{hi} \cdot \lambda_{hi} & , \alpha \cdot n_{hi} \cdot \lambda_{hi} \geq n_{lo} \cdot (1 - \lambda_{lo}) \end{cases}$$

$$\alpha_{eff} =$$

Final geographic load shifting model

$$
\begin{aligned}
C_{geo} = {} & (n_{hi} + n_{lo}) \cdot C_{emb} + \\
& n_{hi} \cdot (\lambda_{hi} \cdot (1 - \alpha_{eff}) + \\
& \quad (1 - \lambda_{hi} \cdot (1 - \alpha_{eff})) \cdot \gamma) \cdot C_{hi} + \\
& n_{lo} \cdot (\lambda_{lo} + \alpha_{eff} \cdot \lambda_{hi} \cdot n_{hi} / n_{lo} + \\
& \quad (1 - \lambda_{lo} - \alpha_{eff} \cdot \lambda_{hi} \cdot n_{hi} / n_{lo}) \cdot \gamma) \cdot C_{lo} + \\
& \eta \cdot \alpha_{eff} \cdot (n_{hi} C_{hi} + n_{lo} C_{lo})
\end{aligned}
$$

Complete emissions model

Baseline + geographic load shifting

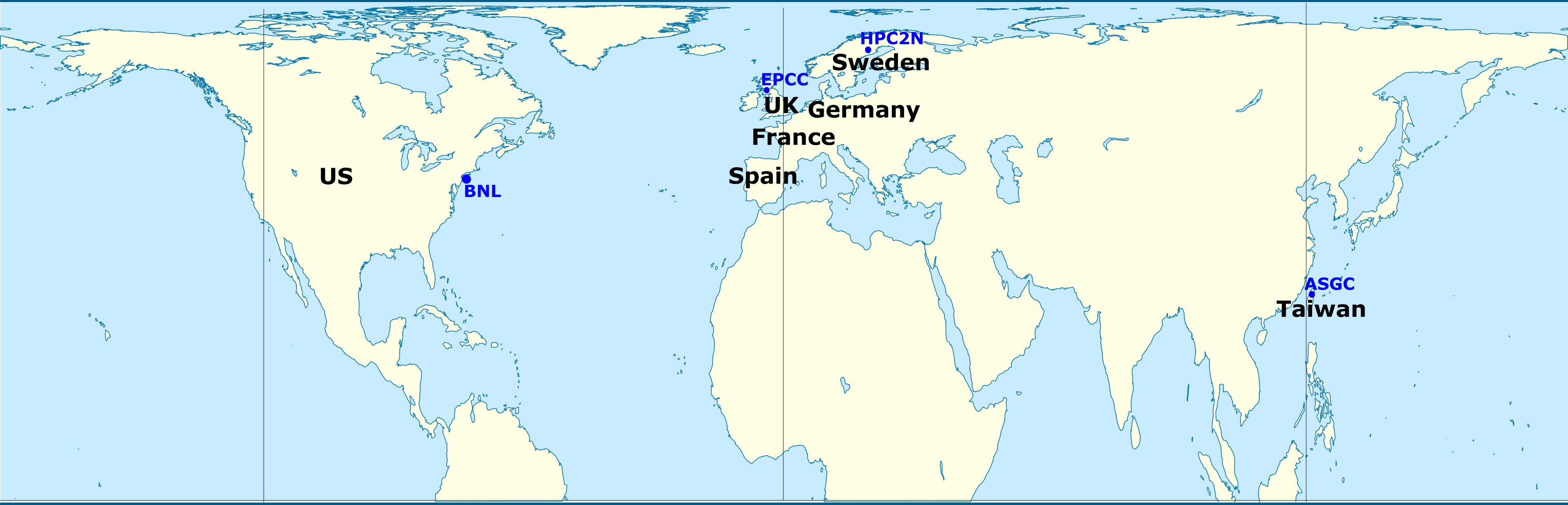
$$C = (1 - \beta) \cdot C_{base} + \beta \cdot C_{geo}$$

If $\beta = 0$ we have the baseline model

Scenarios for geographic load shifting

Scenarios for geographic load shifting

- **Commercial data centre scenarios:**
  - Optimise emissions through use of solar energy and wind energy
  - Locations: US, UK, Germany, France and Spain
- **HPC centres scenarios:**
  - Offloading work to centres in low-emission regions
  - Locations: BNL in Brookhaven, US; EPCC in Edinburgh, UK; HPC2N in Umeå , Sweden; ASGC in Taipei, Taiwan



Commercial data centre scenarios: US, UK, Germany, France and Spain

Supercomputer scenarios: BNL in Brookhaven, US; EPCC in Edinburgh, UK; HPC2N in Umeå, Sweden; ASGC in Taipei, Taiwan

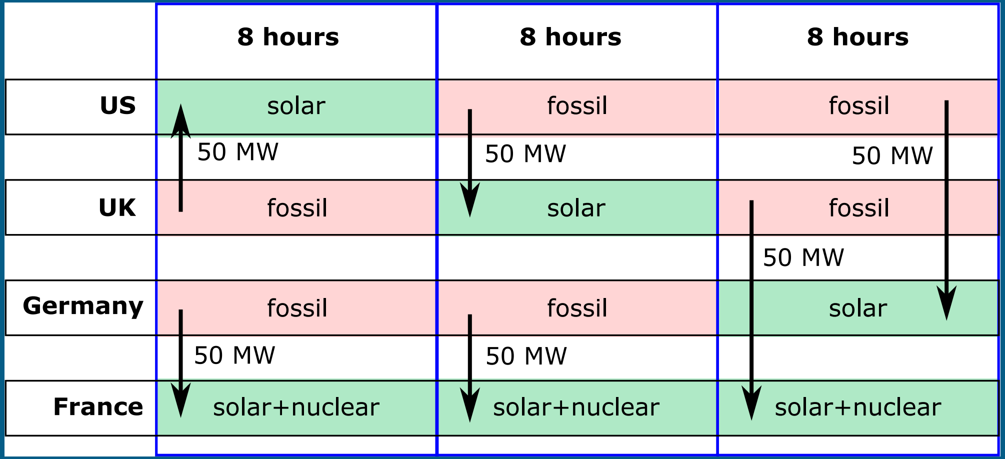
Commercial data centre scenarios

- Based on assumptions GLS paper Lindberg et al (2022):
  - Four identical data centres in different locations with different CI
  - Maximum load is 300 MW
  - Nominal load is 250 MW
- Additional assumptions:
  - GPU nodes similar to Nvidia DGX-A100
  - Data centre PUE is 1.16

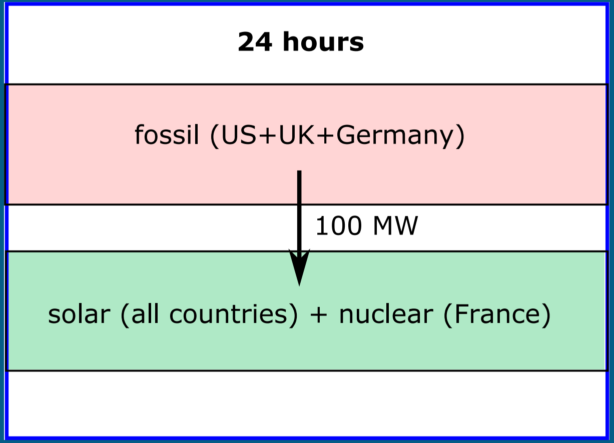
Commercial data centre scenarios

	Scenario 1 (Solar)	Scenario 2 (Wind)
λ_{hi}	0.83 (250/300)	0.83 (250/300)
λ_{lo}	0.83 (250/300)	0.83 (250/300)
γ	0.30	0.30
α	0.20 (50/250)	0.20 (50/250)
β	0.52	0.63

Solar scenario: ideal



equivalent to :



Ideal assumptions:

Two sites are low-CI in any 8-hour period

100 MW can be shifted

Solar scenario: more realistic

	8 hours	**8 hours**	**8 hours**
US	solar	fossil	fossil
	50 MW	50 MW	
UK	fossil	solar	fossil
Germany	fossil	solar	fossil
	50 MW		50 MW
France	solar+nuclear	solar+nuclear	solar+nuclear

More realistic assumptions:

Two, three or one low-CI sites in an 8-hour period)

Only (100+50+50)/3 MW can be shifted

Wind scenario: ideal

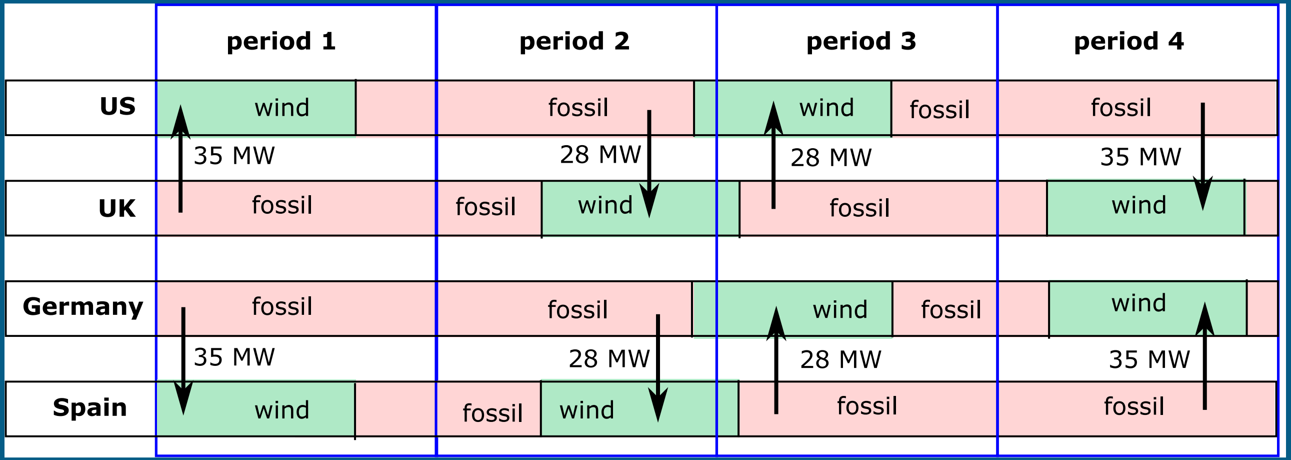
	period 1	period 2	period 3	period 4
US	wind	fossil	wind	fossil
	↑ 50 MW	↓ 50 MW	↑ 50 MW	↓ 50 MW
UK	fossil	wind	fossil	wind
Germany	fossil	fossil	wind	wind
	↓ 50 MW	↓ 50 MW	↑ 50 MW	↑ 50 MW
Spain	wind	wind	fossil	fossil

Ideal assumptions:

Wind load factor 50% for all sites, no correlation

100 MW can be shifted

Wind scenario: more realistic



More realistic assumptions:

Wind load factor 35% and correlation 10%

Only 63 MW can be shifted

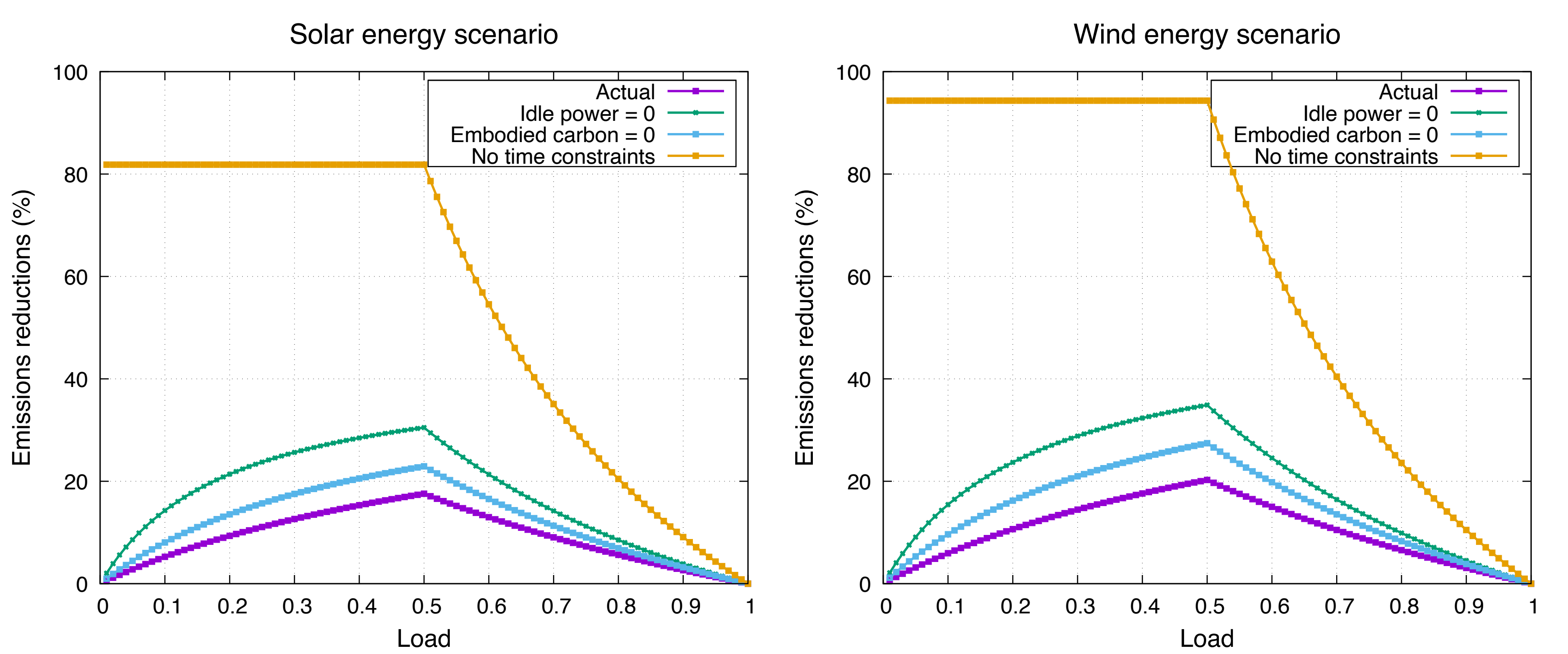
Commercial data centre conclusion

Reduction in emissions around 5% (solar: 4.6% / wind: 5.3%)

Factors limiting gains:

- CI of current renewable technology not zero
- CI of predominantly fossil fuel generation lower than worst case
- Important contribution of embodied carbon in low-emission case
- Idle power consumption not zero
- In realistic scenarios not possible to move sufficient load to use up all excess capacity in low-emission region all the time

Commercial data centre scenarios



Emission reduction as function of load

Supercomputer centre scenarios

Centre	Actual CI (gCO_2e/kWh)
ASGC,Taiwan	642
HPC2N, Sweden	36
BNL, US	369
EPCC, UK	211

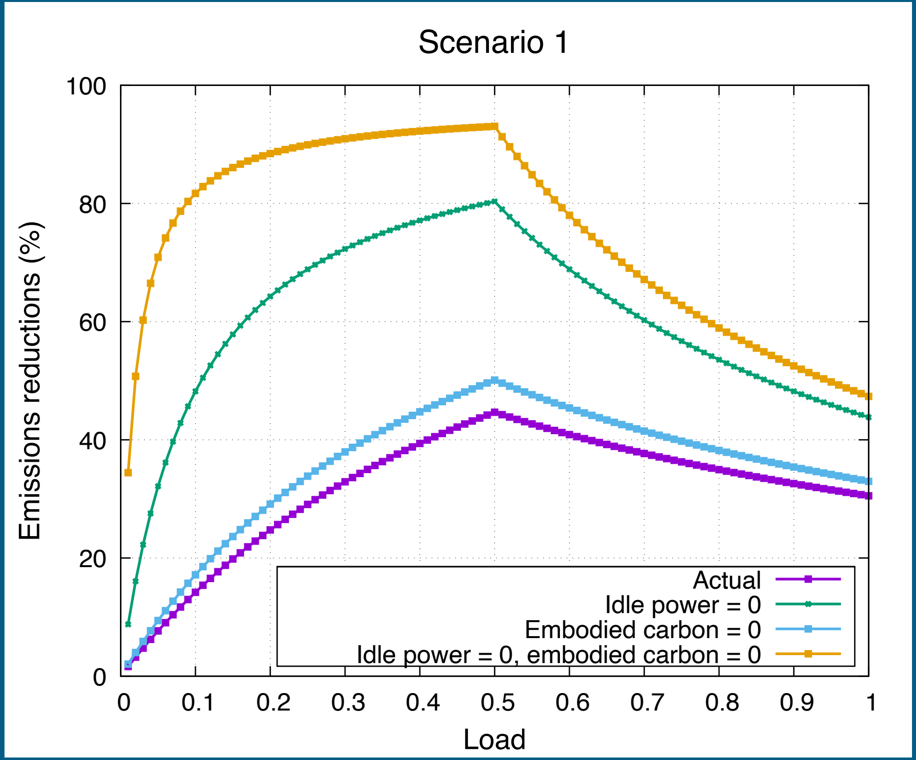
Data: "Carbon intensity of electricity generation - Ember and Energy Institute," 2025, energy Institute - Statistical Review of World Energy (2024) - major processing by Our World in Data

Supercomputer centre scenarios

	C_{hi}	C_{lo}	λ_{hi}	λ_{lo}	γ	α	β
Scenario 1 (ASGC - HPC2N): excess capacity is 50%; max work moved	10,831	390	1.00	0.50	0.30	1.00	1.00
Scenario 2: as scenario 1 but excess capacity is 20%	10,831	390	0.80	0.80	0.30	1.00	1.00
Scenario 3: as scenario 2 but move only 12.5% of work	10,831	390	0.80	0.80	0.30	0.25	0.50
Scenario 4: as scenario 2 but move between BNL (US) and EPCC (UK)	3,879	1,304	0.80	0.80	0.30	1.00	1.00
Scenario 5: Scenario 3 with CI of scenario 4	3,879	1,304	0.80	0.80	0.30	0.25	0.50

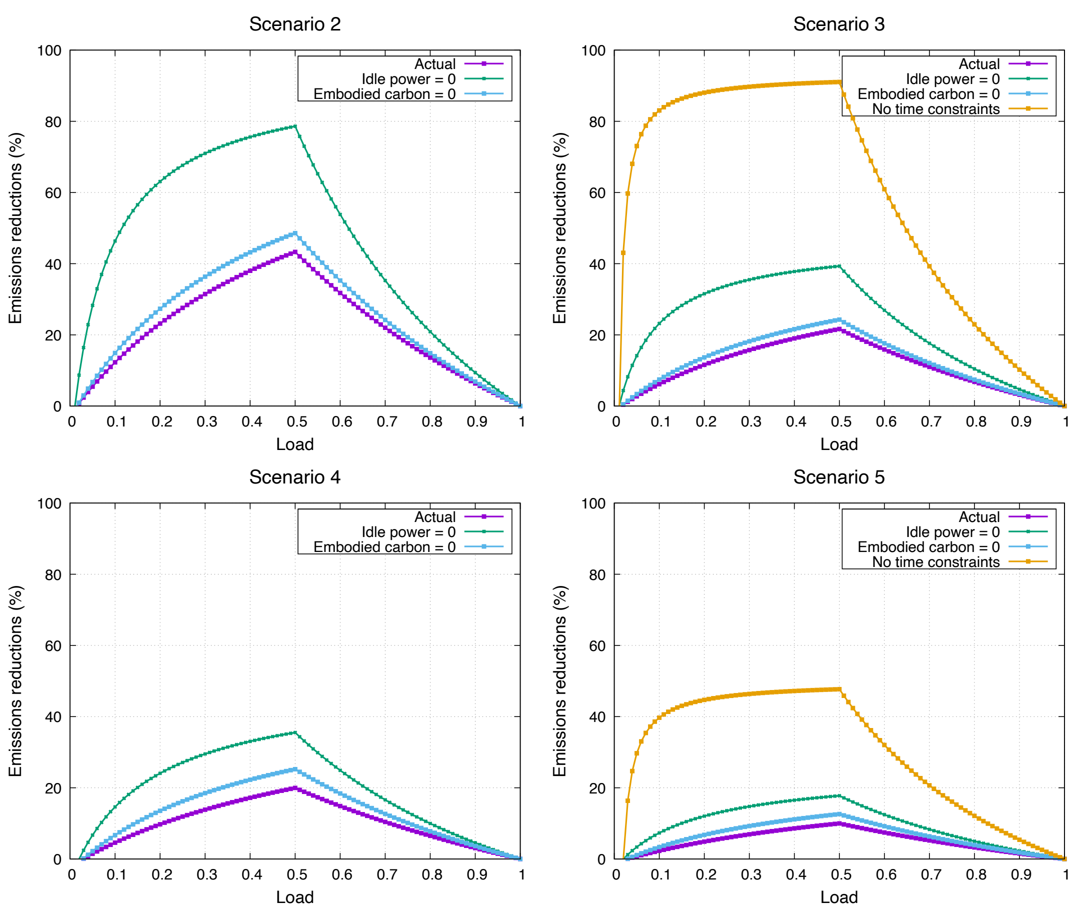
Emissions in kgCO_2e/y

Supercomputer centre scenarios



Emission reduction as function of load – effect of load, idle power and embodied emissions

Supercomputer centre scenarios



Emission reduction as function of load – effect of load, idle power, embodied emissions and time constraints

Supercomputer centre conclusion

Even if data centres don’t provision excess capacity & thus don’t incur excess embodied carbon, **geographic load shifting only results in large reductions if <u>all</u> of the following apply:**

- Free capacity at the low-emission HPC centre is very high
- High-emission site has very high emissions
- Low-emission site has very low emissions
- Overhead for moving workloads is very small

Supercomputer centre conclusion

- **Reduction in emissions around 5%** – unless complete flexibility in moving workloads
- **But many scientific computing workloads are limited:** large data sets and large volumes of data produced
- Much higher reductions with frequency downscaling for I/O-limited workloads or heat re-use – should be priority
- Reducing embodied carbon through server life extension increasingly important as grid decarbonises

Conclusions

- **Model for emissions reduction from geographic load shifting**
  - Using realistic scenarios for commercial AI data centres & HPC centres
  - Using linear analytical model with key server and data centre parameters
- **Achievable reductions in emissions typically less than 5%**
  - For both commercial data centres and HPC centres
  - Even if assuming perfect grid & ignoring performance penalties

Conclusions

- **Much larger reductions in emissions needed to counter current growth in global data centre capacity**

- **Only viable strategy is demand reduction**

Thank you!

For more details and references

https://github.com/Adora-Foundation/carbon-aware-geographic-load-shifting-model