# Systematic Change for Sustainable Supercomputing

**Michael Rudgyard**

michael.rudgyard@alces-flight.com

alcesflight

© 2024 Alces Flight Limited

# A leading provider of managed services for HPC/AI

- One of the largest HPC-focussed commercial teams in the UK (Bicester HQ)

- We currently provide end-to-end managed services for:
  - Systems at 20+ academic and research instutions; commercial HPC/AI systems including three Formula One teams
  - ~8,000 end-users across more than 400,000 x86 compute cores + gpus
  - On-prem and co-located systems from 1 to 100s racks (in UK, Europe and US)

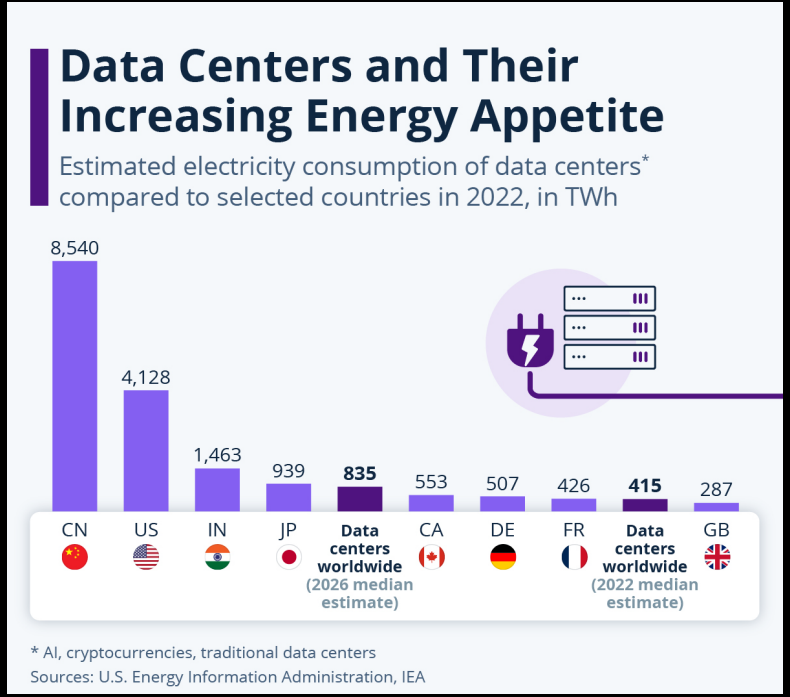
- Presently advising a number of customers on cost and sustainability


© 2024 Alces Flight Limited

# The Carbon Impact of (High Performance) Computing

- Datacentres currently use 1-2% of the world's energy
- "AI is poised to drive a 160% increase in data center power demand by 2030" (source: Goldman Sachs)*
  - A ChatGPT query requires 10 times as much electricity as a Google search....

* Other estimates are significantly higher !



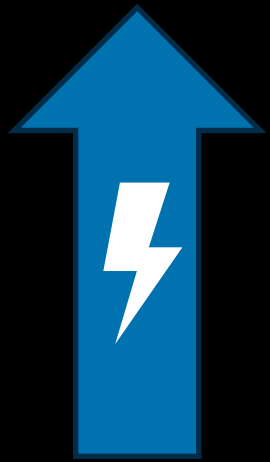


© 2024 Alces Flight Limited

# The Relentless Rise of Electricity Prices ?

- Although geo-political/economic factors can cause significant short-term variation in wholesale electricity markets, the real-term price to consumers and business has continued to increase


Source: gov.uk



© 2024 Alces Flight Limited

# Reducing our Carbon Footprint (& Cost ?)

- What and how do we try to optimize ?
- Can the choices we make result in 'less-bad' outcomes?
  - Is entering into a  Power Purchasing Agreement for renewable energy enough?
  - Or can we make choices that:
    - encourage investment in future infrastructure for renewables
    - help rather than hinder the energy transition
    - help to minimize overall power demand


© 2024 Alces Flight Limited

# IT and Data Centre Efficiency


alcesflight

© 2024 Alces Flight Limited

# Maximizing IT efficiency: ‘science per £ and gCO$_2$e’

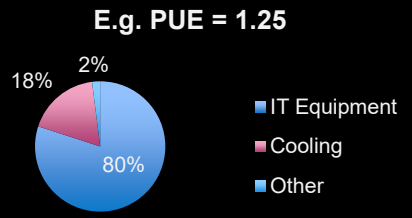
- Maximize throughput at the platform level:
  - Maximize system reliability
  - Minimize idle cycles
  - Manage user adoption & minimize user errors
  - Ensure that applications can scale to the resources demanded
- Optimize throughput at the application level
  - Software optimization, scalability and I/O
  - Workflow optimization
  - Target applications for most suitable platforms
- Implies an increased focus on system management expertise and research software support


© 2024 Alces Flight Limited

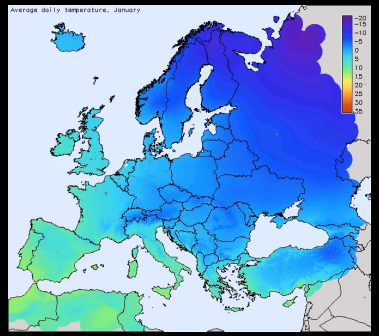
# Data-centre efficiency: Power Usage Effectiveness

- Power-usage effectiveness is a measure of datacentre (in-)efficiency:

$$\text{PUE} = \frac{\text{Total power used by the Data Centre}}{\text{Power used by the IT Equipment alone}}$$




- A 'good' datacentre will have a PUE < 1.2 ; a 'bad' data-centre may have a PUE > 1.5
  - Lots of on-prem HPC datacentres are in the 'bad' category
  - Rear-door and immersive cooling technologies can drive PUEs closer to 1 (especially the latter)
  - Geographical location can also have significant impact on PUE




© 2024 Alces Flight Limited

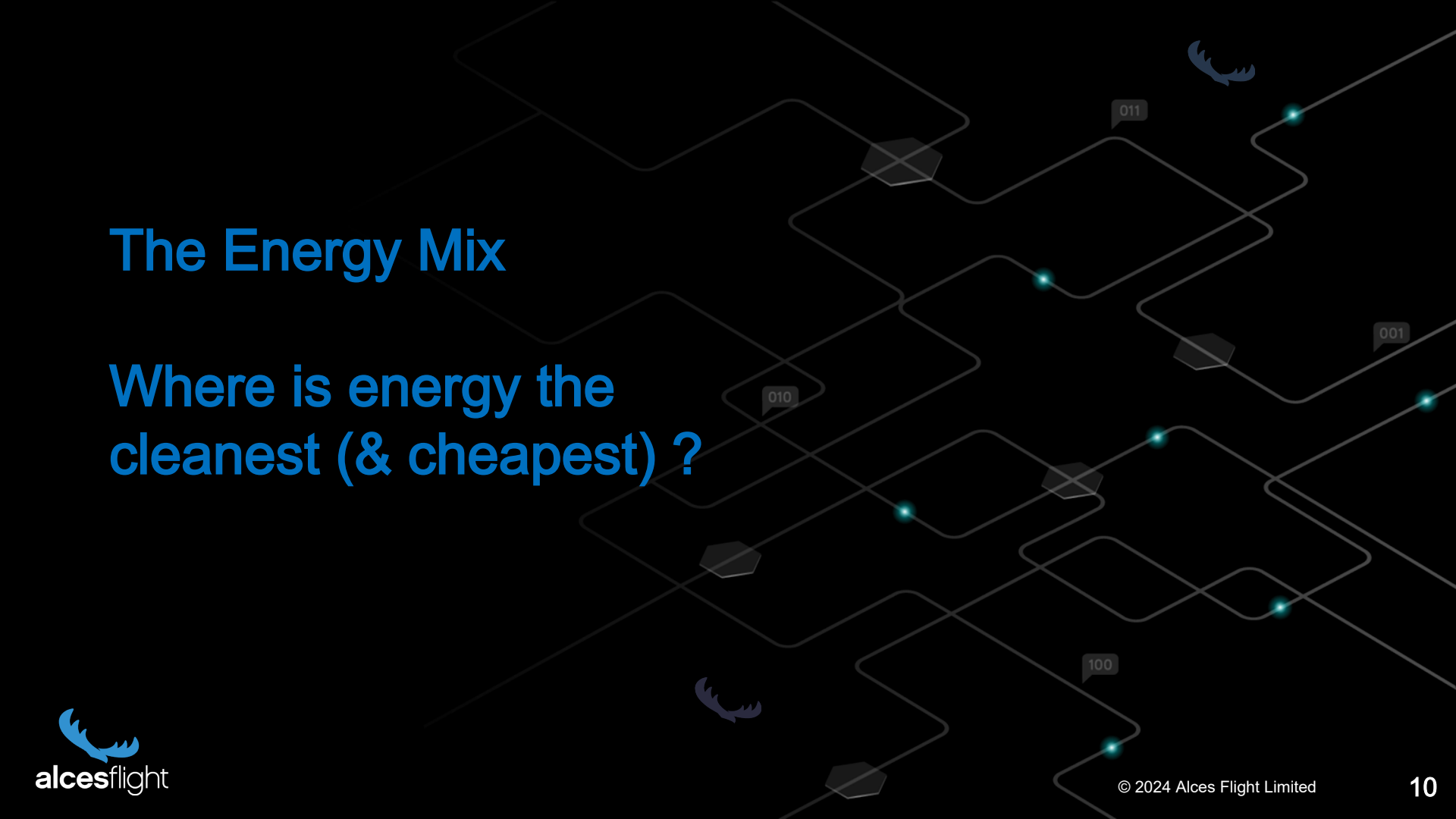
# On-site (or near -site) generation and heat re -use

- The reduction in capital costs for renewables is starting to drive the adoption of on-site solar and/or wind generation (albeit insufficient to power large datacentres)
- Forward-thinking datacentres are also being designed for heat re-use, which can offset $CO_2e$ and reduces contention for electricity
  - Waste heat can be used for district heating systems, swimming pools, greenhouses or industrial processes
  - Sadly, most existing datacentres have not been designed nor sited with this in mind…
- A new generation of smaller 'edge' datacentres, with on-site solar and heat re-use will help to offset both emissions and cost


© 2024 Alces Flight Limited

# The Energy Mix

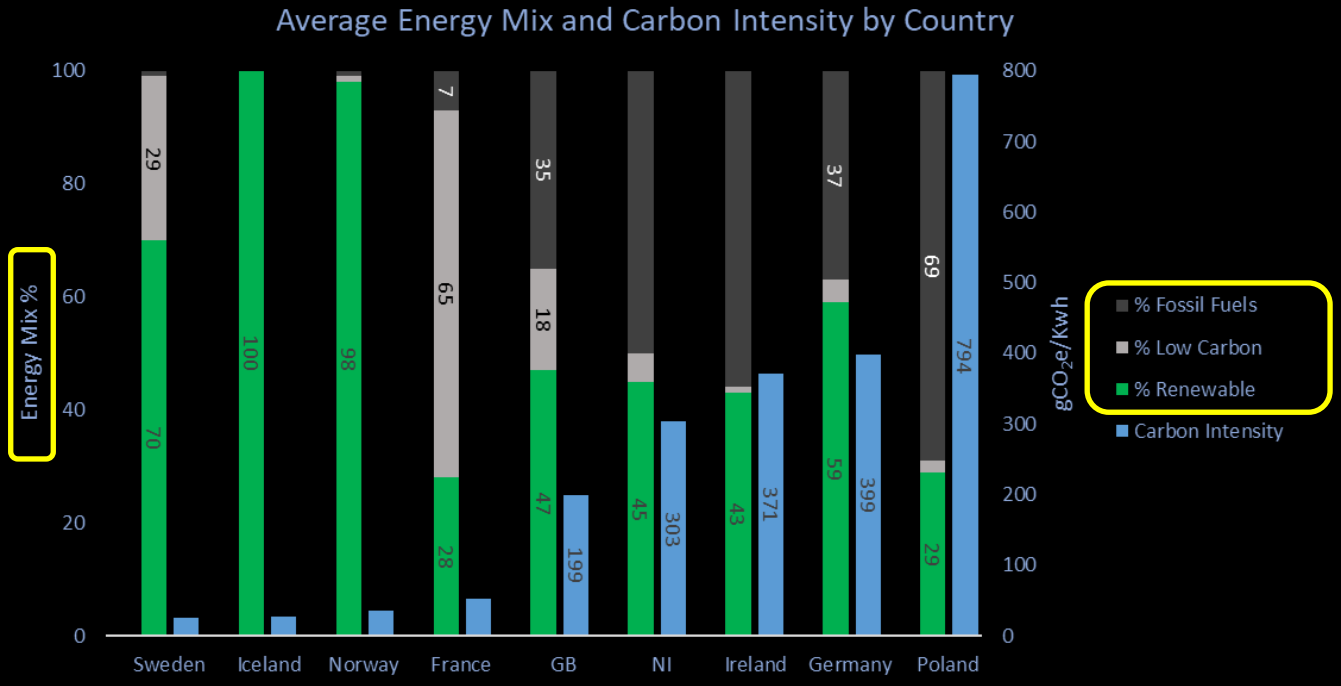
## Where is energy the cleanest (& cheapest) ?




alcesflight

© 2024 Alces Flight Limited

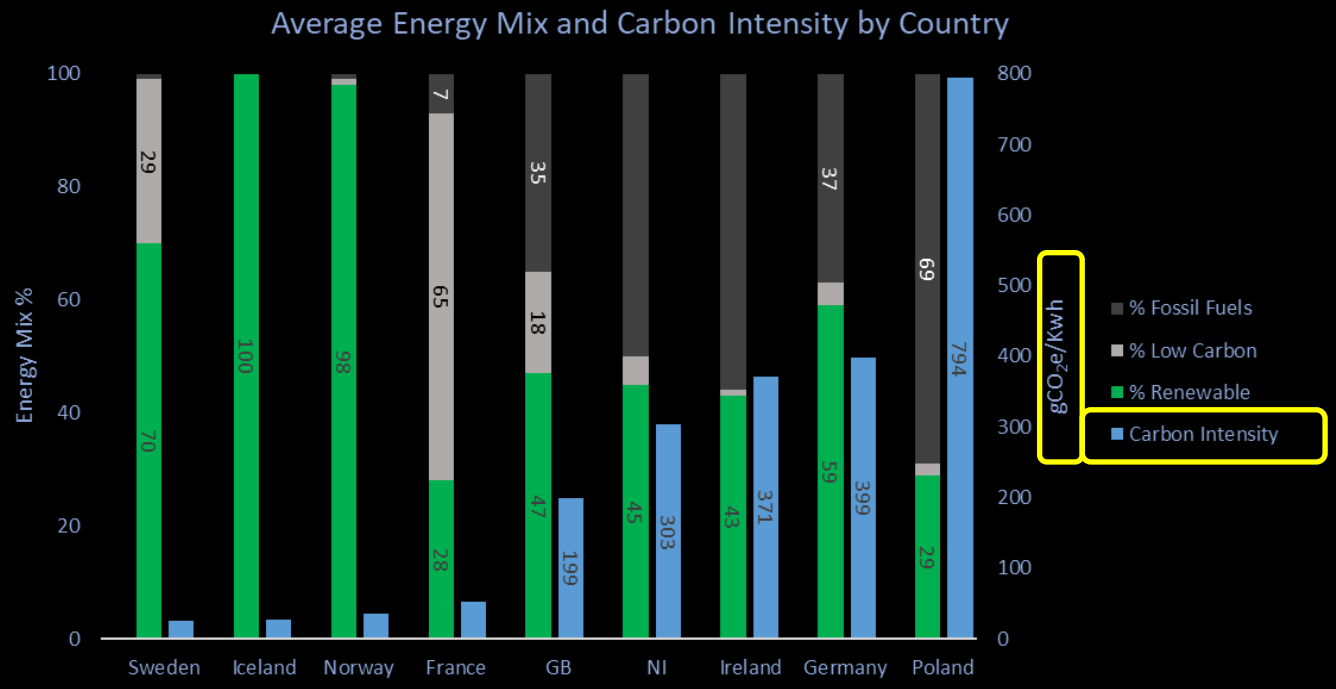
# Electricity generation and renewables




2023 Data. Source: app.electricitymaps.com


© 2024 Alces Flight Limited

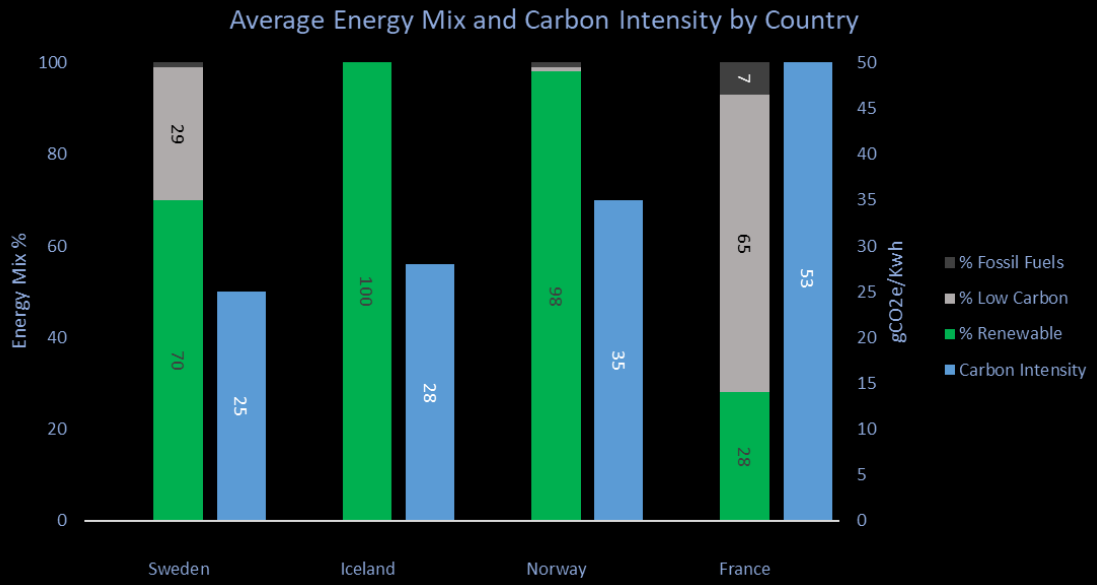
# Electricity generation and renewables

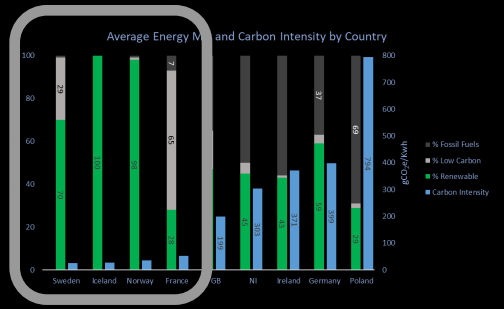



2023 Data. Source: app.electricitymaps.com

© 2024 Alces Flight Limited

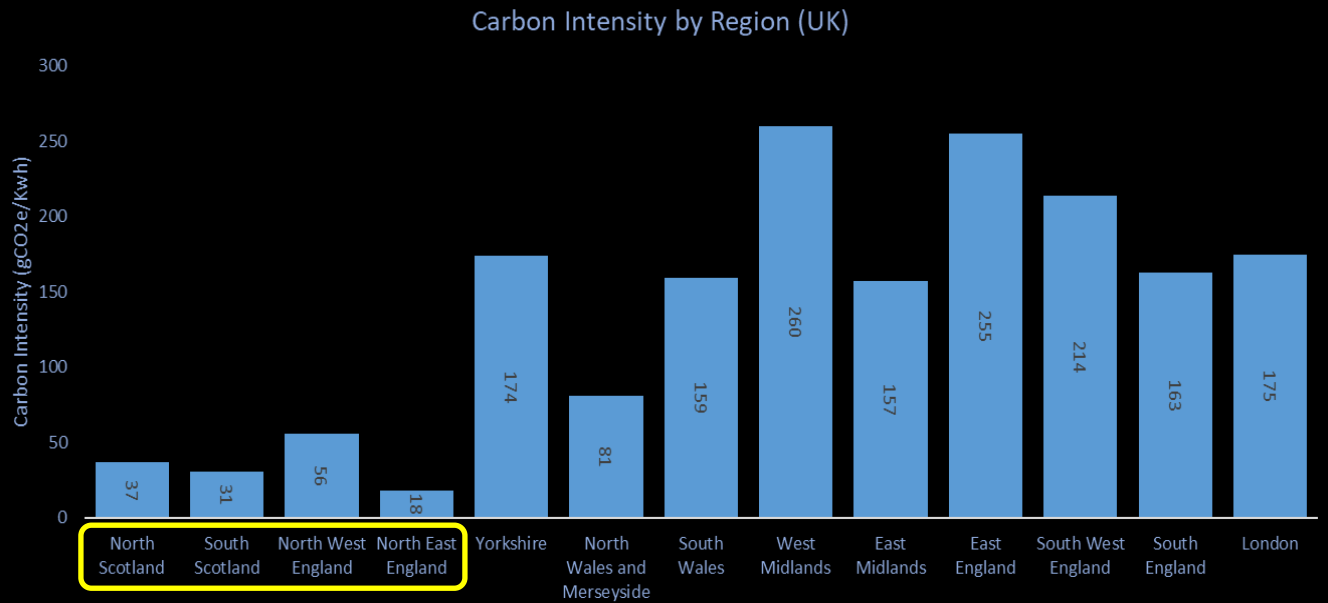
# Electricity generation and renewables




© 2024 Alces Flight Limited

# Carbon intensity by region (UK)




2023 Data. Source: nationalgrideso.com

© 2024 Alces Flight Limited

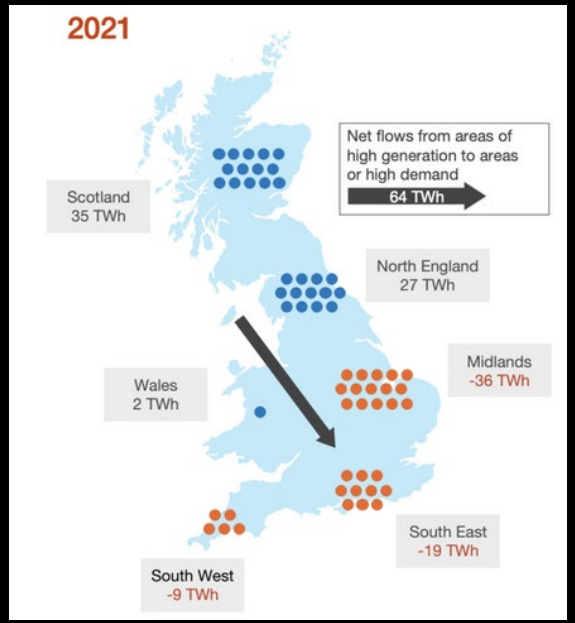
# Location, location, location

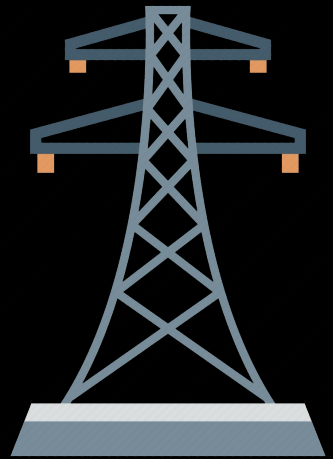
# Why does it matter for HPC in the UK ?

© 2024 Alces Flight Limited

# Transmission losses

- In 2022, 22Twh of electricity generation, or 8% of the total electricity generated, was lost in transmission and distribution (source: gov.uk)

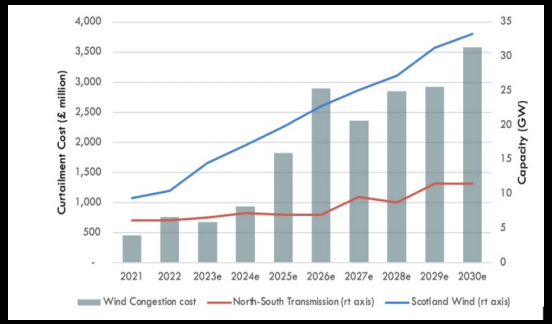


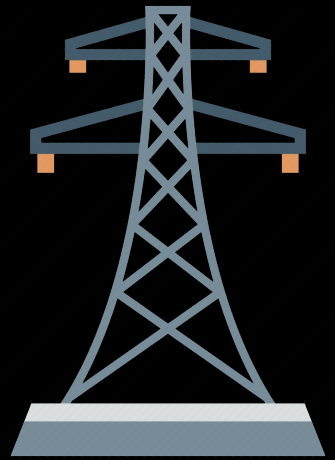
© 2024 Alces Flight Limited

# Energy Curtailment

- 3.2TWh of wind energy was curtailed in 2023, adding 1.4 $MtCO_2e$ and costing £570M on the wholesale market (Source: UK Wind Curtailment Monitor)
- "The practice of powering up gas power plants in England and Wales and switching off wind farms in Scotland cost bill-payers £920M in 2023. Approximately £670M was due to limited bandwidth of the UK's transmission network on the Scottish / English border" (source: Field Energy)




Carbon Tracker Initiative 2023

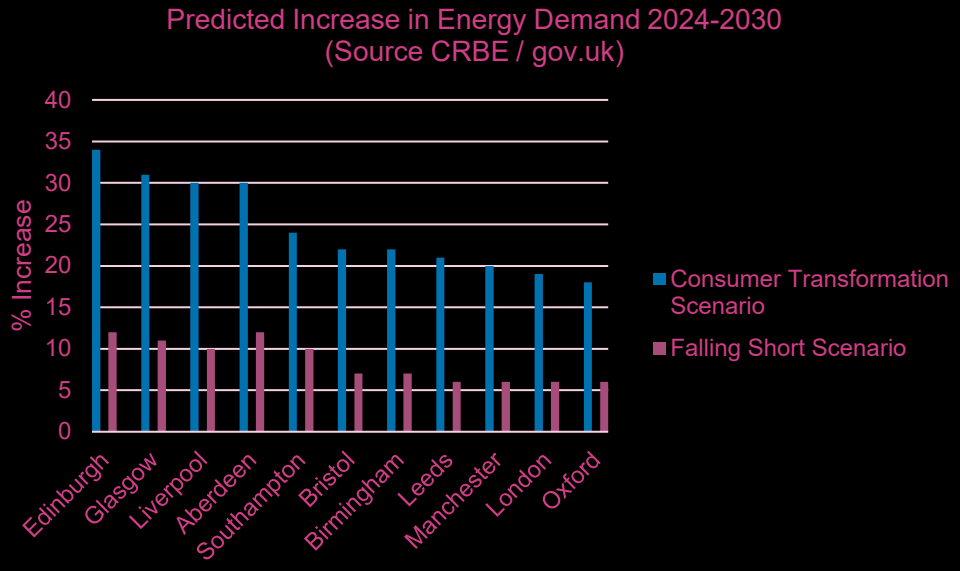


- PS. National Grid has recently announced a £30B investment from 2025-2030, including upgrading key bandwidth limitations.


© 2024 Alces Flight Limited

# Electricity Demand in Built -up Areas

- Meeting net-zero targets and associated regulation will lead to increased electricity demand (and cost?) within UK cities
- Exacerbated by the increased demand for (and cost of) land




*Excluding updated predictions on the effect of the growth in AI/ML

© 2024 Alces Flight Limited

# Embedded Carbon


alcesflight

© 2024 Alces Flight Limited

# Embedded Carbon – Data Centres

- Understanding the Embedded Carbon footprint of a newly-built, existing or renovated datacentre is very difficult!

- Some new datacentres are designed with Scope 3 in mind. E.g.:
  - They may use existing buildings where feasible
  - They may be built by local tradesmen using locally-sourced wood instead of concrete…


© 2024 Alces Flight Limited

# Embedded Carbon – HPC hardware

- Companies (e.g. Dell and HPE) increasingly publish data on embedded carbon equivalents for their products, detailing:
  - Manufacturing: generally, well over 90%
  - Shipping to customer: often negligible
  - EOL: often a small (1-3%) positive or negative %
  - Significant variation in estimates !
- Note:
  - "SSDs dominate in terms of impact in manufacturing phase. They contribute between 48% and 62%, depending on the product" – Dell Factsheet
- Understanding the Carbon footprint of an HPC system with very specific hardware configurations can be difficult, but things are improving….


© 2024 Alces Flight Limited

# Sustainability & the cost of HPC


alcesflight

© 2024 Alces Flight Limited

Very Approximate !

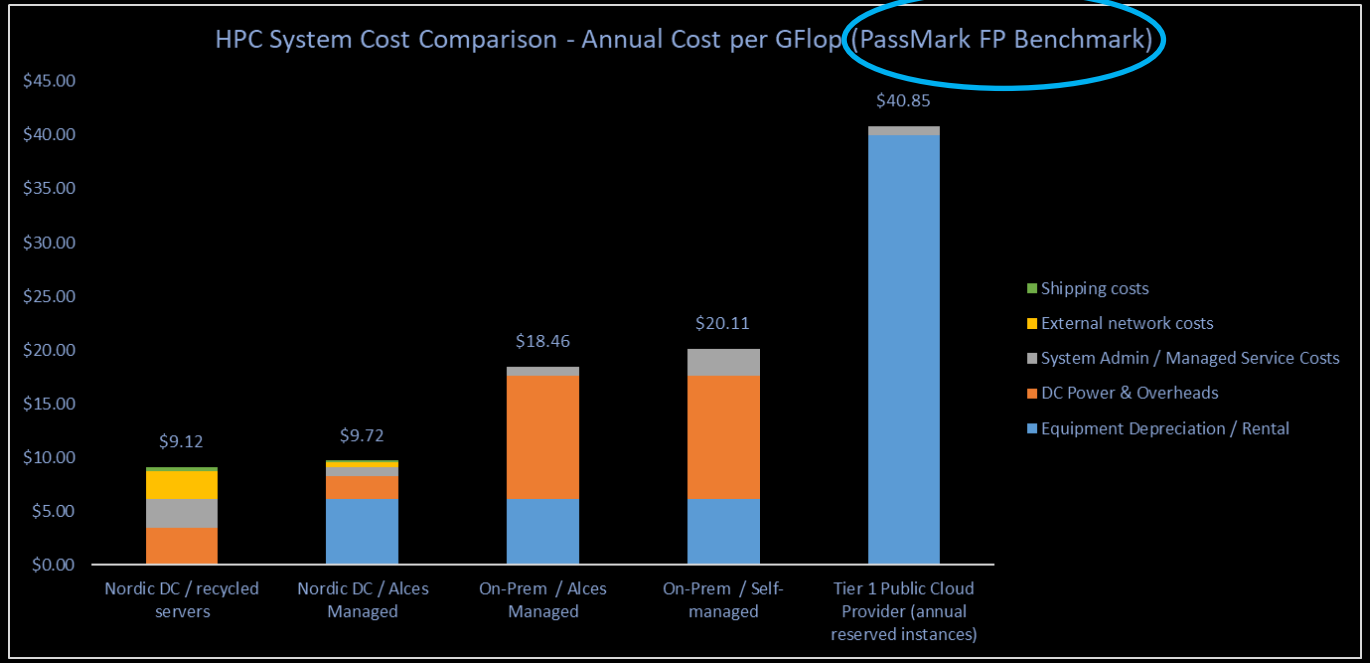
# What if we compared the costs of…..

- A new (mid-size) HPC system hosted in the UK
  - Written off over 5 years
  - In a data centre with PUE of 1.4
  - Electricity @ 35p per kWh
  - With data-centre overheads (staff costs, equipment depreciation & maintenance, etc) of £100,000 p.a
- The same HPC system hosted in a Nordic data centre
  - Including power and hosting costs, network charges, shipping, managed services, etc
- A previous-generation HPC system given a new lease of life in a Nordic data centre


© 2024 Alces Flight Limited

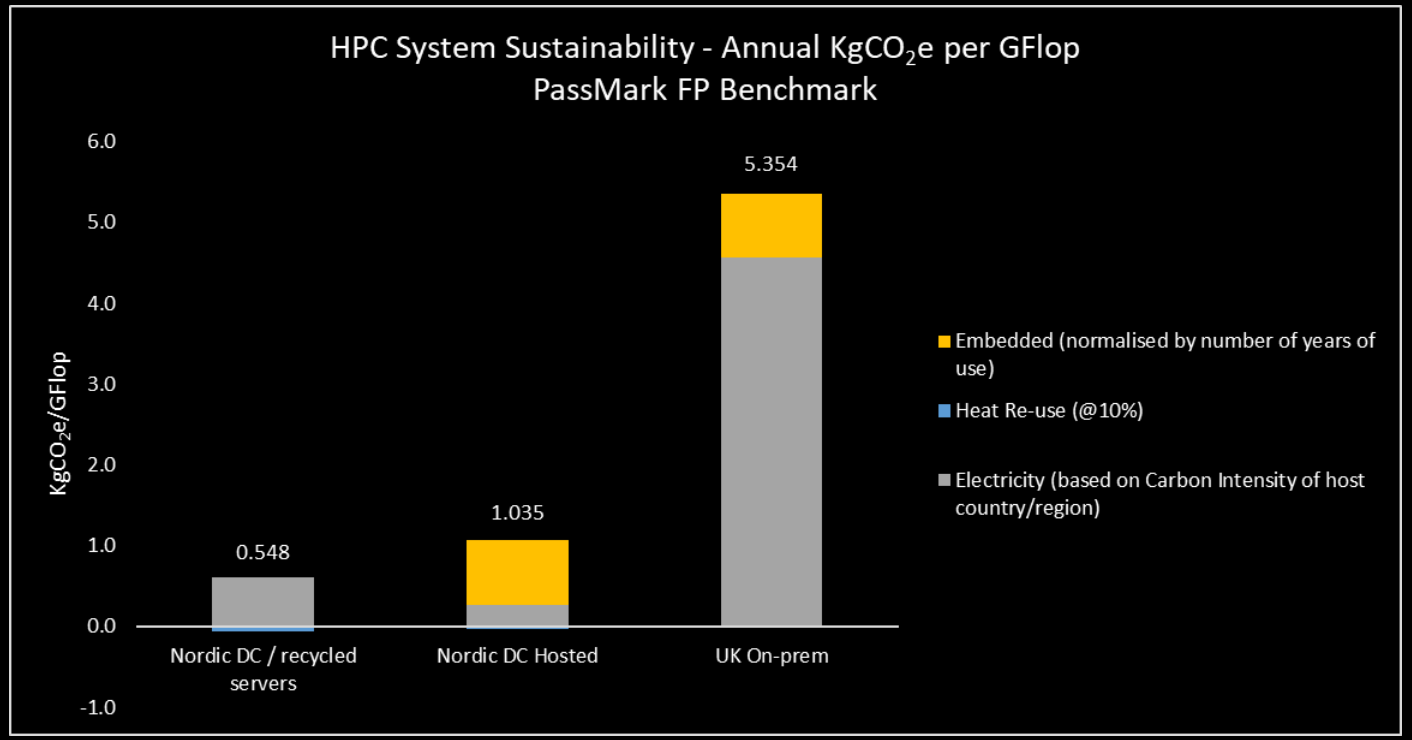
# Annual Costs per GFlop

Synthetic Floating Point Benchmark




© 2024 Alces Flight Limited

# Annual Carbon Footprint per Gflop




© 2024 Alces Flight Limited

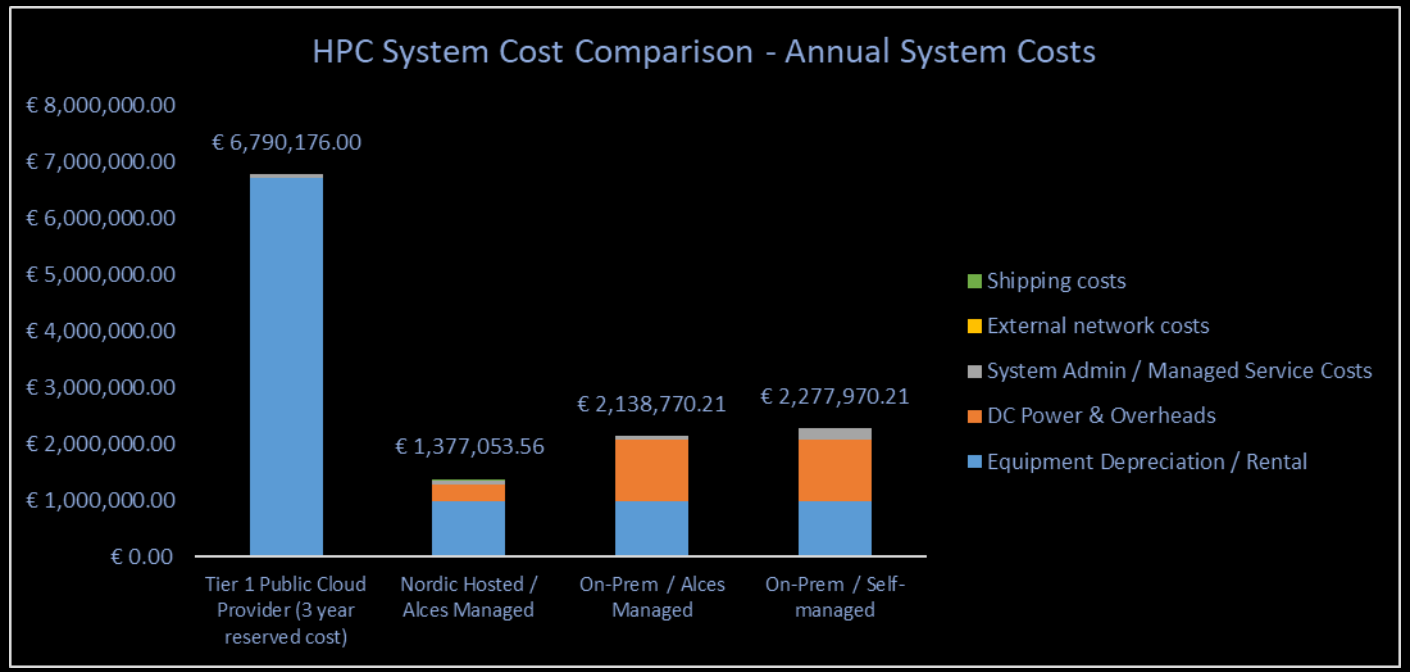
Very Approximate !

# What if we also compared the costs of…..

- A new larger HPC/AI system hosted in the UK
  - Mixture of H100 and A100 GPUs
  - Compute oriented (little storage)
- The same HPC/AI system hosted in a Nordic data centre
- A similar spec'ed system on a Tier 1 Cloud provider


© 2024 Alces Flight Limited

# Annual System Costs – GPU system




© 2024 Alces Flight Limited

# Conclusions…

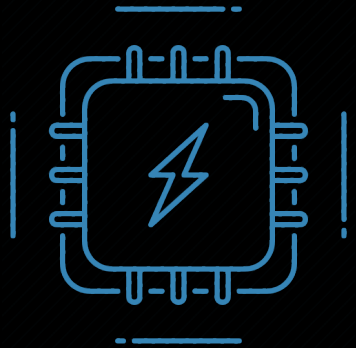

alcesflight

© 2024 Alces Flight Limited

# Our Conclusions

- Cost and sustainability are (increasingly) strongly correlated
- Users should carefully consider where compute takes place
  - It is easier (and cheaper) to move data to the power than power to the data (for applications for which bandwidth, latency or data sovereignty / security are not major issues)
  - Near-prem solutions, with heat re-use and some solar, are increasingly viable and may be technically and 'politically' preferable in some cases
  - A mix of near-prem, co-located and on-demand public cloud resource will make sense for larger HPC/AI users

    (NB. There is increasing choice of public cloud providers for HPC/AI !)
- Understanding and reporting on efficiency, cost and sustainability is not only increasingly important, but is a requirement for larger organisations




© 2024 Alces Flight Limited