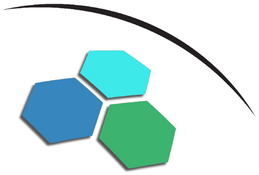
CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre



Walking the Walk:

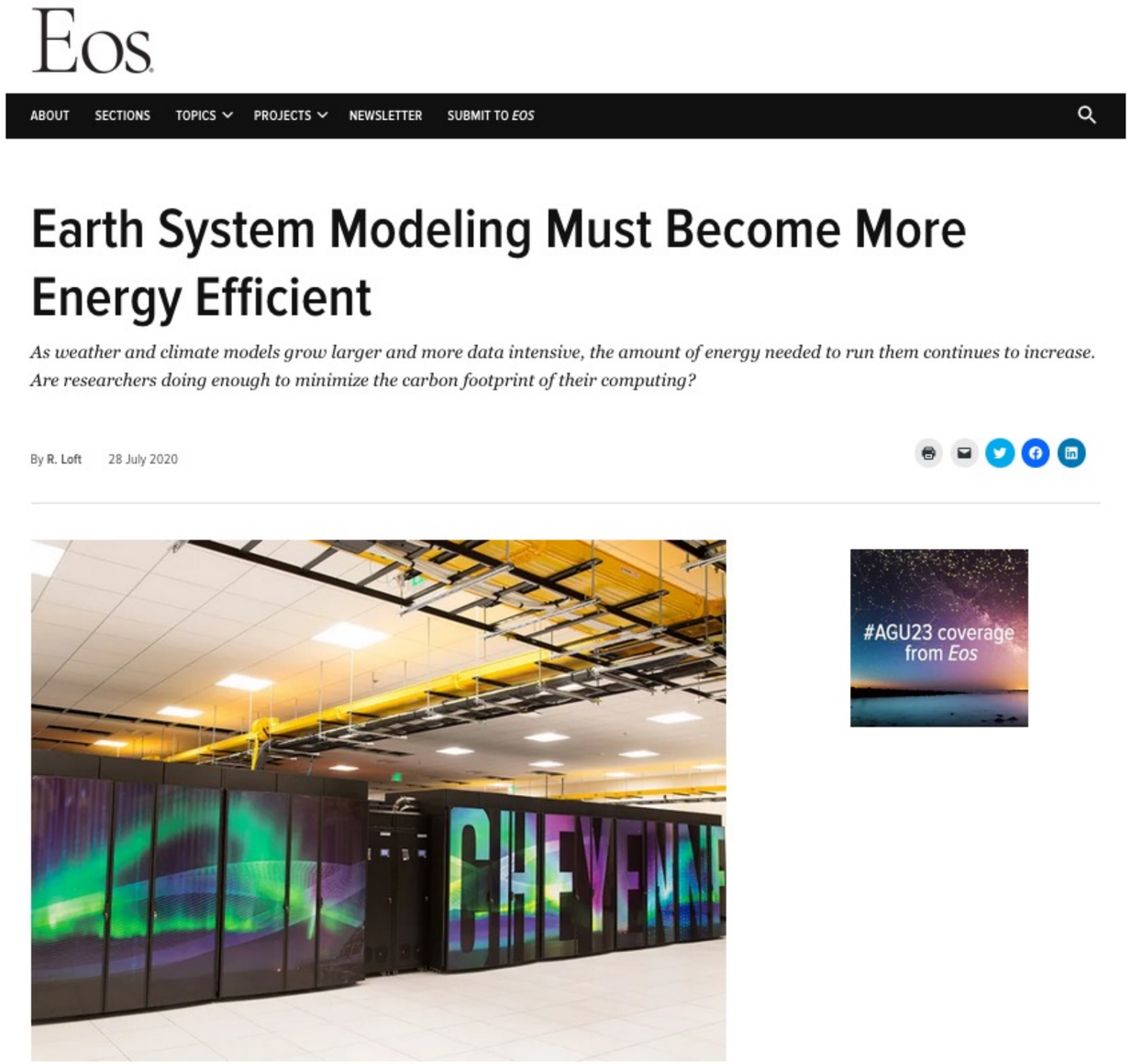
The ESM Community must be a Role Model in Energy Efficiency

PASC24, MS1F

June 3, 2024

Rich Loft (AreandDee LLC), **Will Sawyer (CSCS)**

Search for a motivational speaker



- B.Sc Chemistry, 1977, Harvey Mudd College
- Ph.D. Physics, 1988, University of Colorado, Boulder
- Thinking Machines Corporation. 1989-1994
- 1994 – 2021 National Center for Atmospheric Research (NCAR), Director of Technology Development in the Computational and Information Systems Laboratory
- 2001 Gordon Bell Prize Special Category working to develop HOMME
- 2023 Walter O. Roberts Scientific and Technical Advancement Award for the development of a GPU version of the MURaM solar physics model
- 2003-2012: Planning for NCAR-Wyoming Supercomputing Center

Guiding Principles

- You can’t fix an issue you don’t **own**.
- You can’t improve what you don’t **measure**.
- You can’t claim transparency if you don’t **report** what you measure.
- You can’t claim to care if you don’t **work to improve** your performance.
- **Mottainai:** A Japanese work roughly translated as a “waste not want not philosophy”.

Ownership of our own climate impacts

Have we done everything to minimize the carbon footprint of our computing activities?

- **In conversations with colleagues, I have encountered the following counter-arguments:**
  - **Weather and climate modeling activities are only a small contributor to societal emissions.**
    - But the footprint is increasing and likely to increase further.
    - First, many actors outside the scientific community are now running these models for insights into future conditions.
    - Second, leaders in ESS community are pushing for km-scale ESMs [Shukla, Palmer]. Running high-resolution ES models approach the carbon footprint of training the largest AI models, which have recently come, along with activities like Crypto-currency mining, under heavy scrutiny by the environmental community.

Palmer, T. (2014), Climate forecasting: Build high-resolution global climate models, Nature, 515, 338–339, https://doi.org/1 0.1038/515338a.

Shukla, J., et al. (2010), Toward a new generation of world climate research and computing facilities, Bull. Am. Meteorol. Soc., 91(10), 1,407–1,412, https://doi.org/10.1175/2010BAMS2900.1.

Ownership, continued

Other reactions I have encountered in discussions with colleagues.

- **ESS research is too important to let it be slowed by these considerations.**
  - Scientists tend to see themselves as impassive observers. But scientists are part of the system they are studying.
  - Hypocrisy undermines credibility and our effectiveness in translating research to action.
  - In any event, this argument is out of step with emerging regulatory frameworks requiring emissions reporting.
- **Even raising the subject of emissions from ESS computing provides ammunition for political attacks on the science and scientists.**
  - Political (and personal!) attacks will come anyway.
  - Any whiff of hypocrisy is simply chum in the water.

Ownership

Have we done everything to minimize the carbon footprint of our computing activities?

- **No big deal. Just shift to Green Energy Sources and you're done.**
  - The environmental side effects of a future decarbonized energy portfolio are not well understood [Luderer, 2019], and
  - Switching to renewable energy sources often means buying credits, which can be traded, thereby obfuscating the actual energy source.

Luderer, G., et al. (2019), Environmental co-benefits and adverse side-effects of alternative power sector decarbonization strategies, *Nat. Commun.*, 10, 5229, https://doi.org/10.1038/s41467-019-13067-8.

Measure and Report

- Measure and report the Power Usage Effectiveness of the data-centers where the computations were done.
- Report the methodology used to compute energy consumption of the system.
- Report energy per simulated year for the experiments performed.
- Reporting leads to **inter-comparison** which can create a *virtuous cycle* of improvement through competition and innovation.

Working to Improve

Specific Factors that can improve carbon footprint of ESMs

- **Algorithm/Program Improvements (Avoiding Computations)**
  - **Numerical methods innovation:** e.g. [Bosler, 2020] SLT algorithms and more efficient communication patterns improved E3SM dynamical core throughput by 6x.
  - **Reduced precision:** 32 bit precision may be unnecessary for climate computations. [Paxton, 2022]
  - **AI-based weather models** (e.g. FourCastNet, Pangu Weather) are showing rapid advancement in skill for intermediate weather forecasting, may be orders of magnitude faster, and are, in some cases, better than traditional physics-based models. [Pathak, 2022] [Bi, 2023]
- **Processor Efficiency**
  - [Fuhrer, 2018] proposed energy to solution metric for meteorology for the COSMO model.
  - In 2020, my team at NCAR measured the power-efficiency of the NVIDIA V100 GPU as 3.8x that of a cluster composed of dual, 18-c Intel Xeon Broadwell nodes running the Model for Prediction Across Scales MPAS meteorological model.
- **Data Center Power Utilization Efficiency**
  - PUE can vary from datacenter to datacenter by up to 40% (Uptime Institute Survey).
- **Energy Mix**
  - The energy mix (and thus the carbon content) of the power source is highly dependent on the location of the datacenter and the utility provider.

Reference: “Computational Challenges in weather and climate modeling” presentation at Boise State University by Peter Bosler, Sandia National Laboratory (10/2020)

Watch here: https://www.youtube.com/watch?v=03iPJSRl5j8Paxton, A.E., Chantry, M., Kloewer, M., Saffin, L., Palmer, T., Climate Modeling in Low Precision: Effects of Both Deterministic and Stochastic Rounding, Journal of Climate, 2022, p.1215-1229. https://doi.org/10.1175/JCLI-D-21-0343.1

Pathak, J. et al. FourCastNet: a global data-driven high-resolution weather model using adaptive Fourier neural operators. Preprint at https://arxiv.org/abs/2202.11214 (2022).

Bi, K., Xie, L., Zhang, H. *et al.* Accurate medium-range global weather forecasting with 3D neural networks. *Nature* **619**, 533–538 (2023). https://doi.org/10.1038/s41586-023-06185-3.

Fuhrer, O., et al. (2018), Near-global climate simulation at 1.km resolution: Establishing a performance baseline on 4888 GPUs with COSMO.5.0, Geosci. Model Dev., 11, 1,665–1,681, https://doi.org/10.5194/gmd-11-1665-2018.

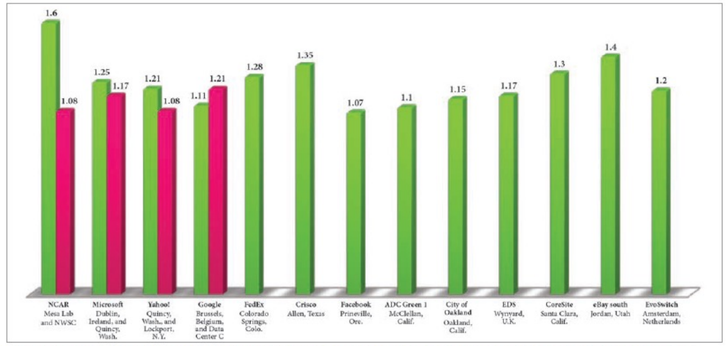


Fig. 1. A comparison of power usage efficiency (PUE) at a variety of energy-efficient data centers. All of these data centers were built within the past few years, with the exception of NCAR's Mesa Laboratory, which was built in 1967. The average PUE in 2019 among more than 600 data centers surveyed by the Uptime Institute was 1.67.

Working to Improve

Recommendations

- Improving the efficiency models is an interdisciplinary issue.
- There is a need for vendor power measurement tools.
- Use citable research supporting energy and CO2 emissions estimation methodologies.
- Governments should increase research into energy-saving numerical techniques, machine learning models, novel architectures, and reduced precision computations for ESMs.

How the AI community (e.g. Meta) addressing this issue?

What walking the walk looks like…

1. Acknowledge the issue!
2. Cite your methodology for computing power and emissions.
3. Compare to other models (OPT and BLOOM)
4. Provide estimates for various model configurations (7B, 13B, 65B… roughly equivalent to model resolution in ESM modeling)

6 Carbon footprint

The training of our models have consumed a massive quantity of energy, responsible for the emission of carbon dioxide. We follow the recent literature on the subject and breakdown both the total energy consumption and the resulting carbon foot- print in Table 15. We follow a formula for Wu et al. (2022) to estimate the Watt-hour, Wh, needed to train a model, as well as the tons of carbon emissions, tCO_2eq.

	GPU Type	GPU Power consumption	GPU-hours	Total power consumption	Carbon emitted (tCO_2eq)
OPT-175B	A100-80GB	400W	809,472	356 MWh	137
BLOOM-175B	A100-80GB	400W	1,082,880	475 MWh	183
LLaMA-7B	A100-80GB	400W	82,432	36 MWh	14
LLaMA-13B	A100-80GB	400W	135,168	59 MWh	23
LLaMA-33B	A100-80GB	400W	530,432	233 MWh	90
LLaMA-65B	A100-80GB	400W	1,022,362	449 MWh	173

Table 15: **Carbon footprint of training different models in the same data center.** We follow Wu et al. (2022) to compute carbon emission of training OPT, BLOOM and our models in the same data center. For the power consumption of a A100-80GB, we take the thermal design power for NVLink systems, that is 400W. We take a PUE of 1.1 and a carbon intensity factor set at the national US average of 0.385 kg CO_2e per KWh.

Text and Table extracted from: Touvron, H., et al. “LLaMA: Open and Efficient Foundation Language Models”. arXiv:2302.13971v1 [cs.CL])

https://doi.org/10.48550/arXiv.2302.13971, February 2023.

How does ESM energy consumption compare to LLMs?

A point of comparison

- A version of CESM running on CPUs at 15 km/58 level AMIP simulation has a throughput .15 SYPD (simulated years per day) on 84 nodes (10,752 CPU cores) of the Derecho supercomputer at NCAR.
- The power required to run a CPU node of Derecho is ~1 kW (2.6 MW total).
- Take the PUE to be 1.1. Derecho operates in a facility capable of this efficiency… the NCAR-Wyoming Supercomputing Center.
- 7.25 MW-hr to simulate 1 year of climate at 15 km/58 levels.
- Scale this in time to (1 century) or in resolution (3 km) and we’re in the ball-park of the energy costs for training an LLM (see table on previous slide).

Comments from co-chair

Rich’s message: hypocrisy diminishes our credibility

- It is not sufficient that we just do our best to increase energy efficiency
- But we need to consider, quantify and report the full footprint of our computing activities
- Society and human-kind looks to us to find the path forward. We are scientists on the one hand, but also role models

Suggestions for professional carbon-awareness

- Maintain a personal carbon diary, also for work, following the suggestion of environmental economist Dieter Helm
- Commuting and office space is a non-negligible proportion of our overall carbon footprint. Desk sharing and public transportation make a difference.
- Air travel is a *significant* proportion of our overall footprint
  - Optimize travel for multiple purposes
  - Attend conferences which can be reached by train travel, attend virtually the others, many of which allow full remote participation
  - Petition the other good conferences to allow remote participation

Thank you very much!