JISEA Joint Institute for Strategic Energy Analysis

Green Computing Catalyzer Kickoff: Green Artificial Intelligence (AI) Directions

Charles Edison Tripp

JISEA Annual Meeting

4/13/2022

NREL Transforming ENERGY · COLORADO SCHOOL OF MINES EARTH • ENERGY • ENVIRONMENT · COLORADO STATE UNIVERSITY · MIT Massachusetts Institute of Technology · Stanford University · University of Colorado Boulder

Background

- As demands for high volumes of data processing, data analysis, artificial intelligence grow, and computing systems permeate multitudinous aspects of commerce, industry, government, and life, computing has become a burgeoning consumer of energy and contributor of carbon emissions.

- While computational research at NREL has historically focused on applying computing resources to energy efficiency and renewable energy research questions, computing itself is now a serious energy research challenge.

- The Green Computing Catalyzer looks to establish green computing as a salient research domain at NREL, and crucially, to begin to address the looming computing energy crisis.

Background

- Demands for high volumes of data processing, data analysis, and high-performance artificial intelligence are growing rapidly. Computing systems are permeating many aspects of life.
- Computing has become a burgeoning consumer of energy and source of carbon emissions.

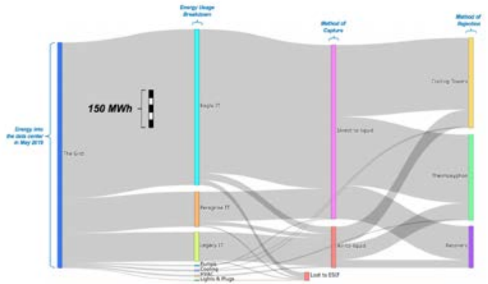
Model	Hardware	Power (W)	Hours	kWh·PUE	CO_2e	Cloud compute cost
$\text{Transformer}_{base}$	P100x8	1415.78	12	27	26	\$41–\$140
Transformer_{big}	P100x8	1515.43	84	201	192	\$289–\$981
ELMo	P100x3	517.66	336	275	262	\$433–\$1472
BERT_{base}	V100x64	12,041.51	79	1507	1438	\$3751–\$12,571
BERT_{base}	TPUv2x16	—	96	—	—	\$2074–\$6912
NAS	P100x8	1515.43	274,120	656,347	626,155	\$942,973–\$3,201,722
NAS	TPUv2x1	—	32,623	—	—	\$44,055–\$146,848
GPT-2	TPUv3x32	—	168	—	—	\$12,902–\$43,008

Table 3: Estimated cost of training a model in terms of CO_2 emissions (lbs) and cloud compute cost (USD).[7] Power and carbon footprint are omitted for TPUs due to lack of public information on power draw for this hardware.

Strubell, E., Ganesh, A., & McCallum, A. (2019, July). Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3645-3650).

A Brief History of Green Computing at NREL

As a datacenter practitioner with demanding HPC workloads NREL has successfully demonstrated a decade of energy efficiency and expertise.

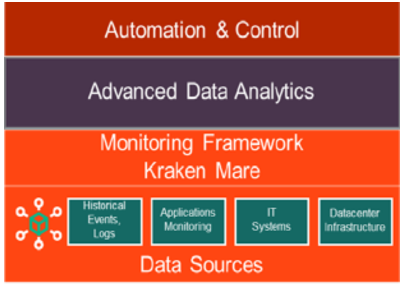


Energy flow (1.23MW average power) through the HPCDC infrastructure, May 2019

- World Class
- Energy Efficiency

- Application of Fuel Cells &
- Datacenter Computing

- Water Usage



- Datacenter Optimization
- (AIOps)

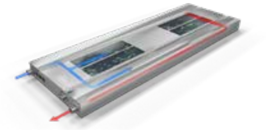
Transforming **ENERGY** through computational excellence

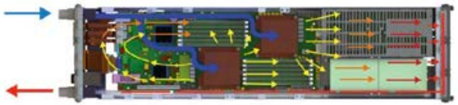
A Brief History of Green Computing at NREL

Hot Water Cooling

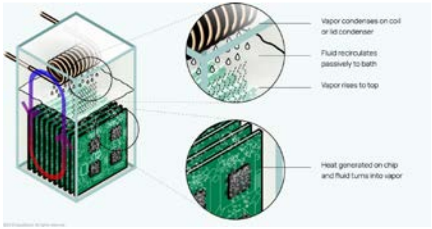
- Electronic components operate at 70-90C (160-194)F

Direct Immersion

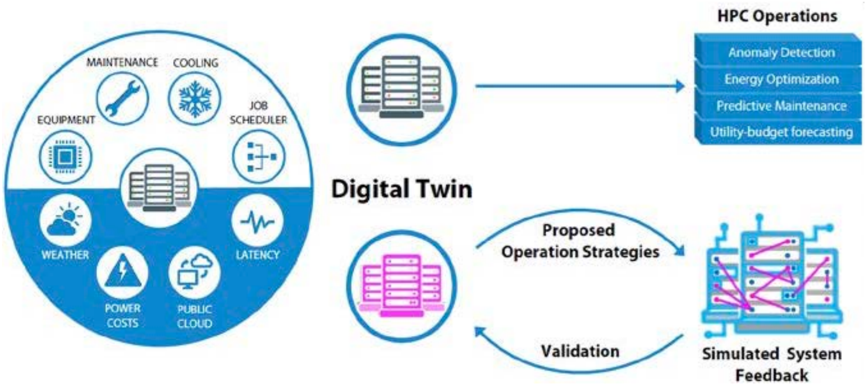




Two Phase



(image from Liquid Stack)



Digital Twins

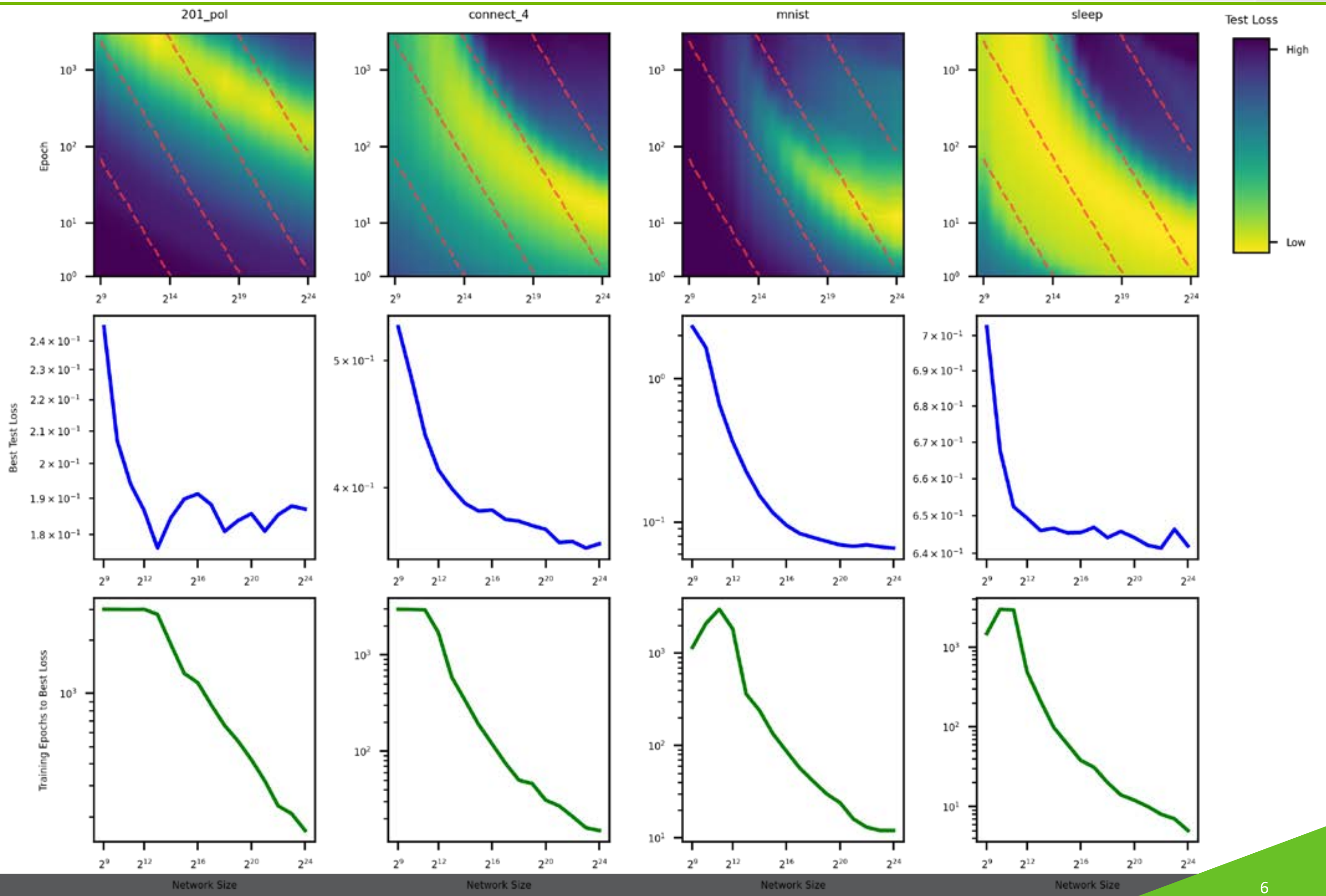
- Complex systems analysis
- Realistically model large complex dynamic environments prior to building or investing

Transforming **ENERGY** through computational excellence

Slide Credit: Aaron Andersen

Empirical Deep Learning Efficiency Study

Larger models can perform better and achieve peak performance in fewer training epochs.

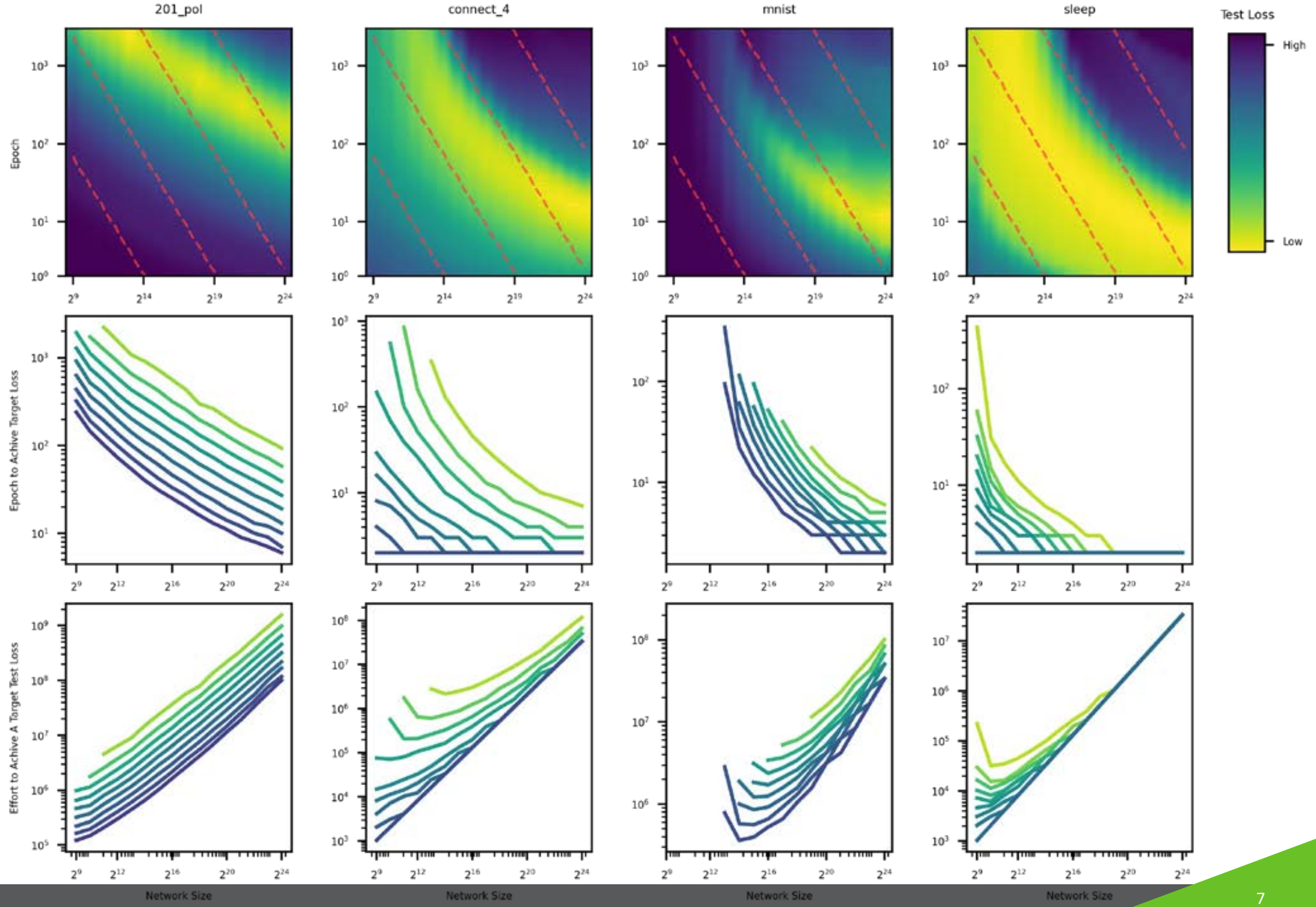


Empirical Deep Learning Efficiency Study

Larger models can perform better and achieve peak performance in fewer training epochs.

However,

- The cost of increasing model size outpaces efficiency gains from fewer training epochs.

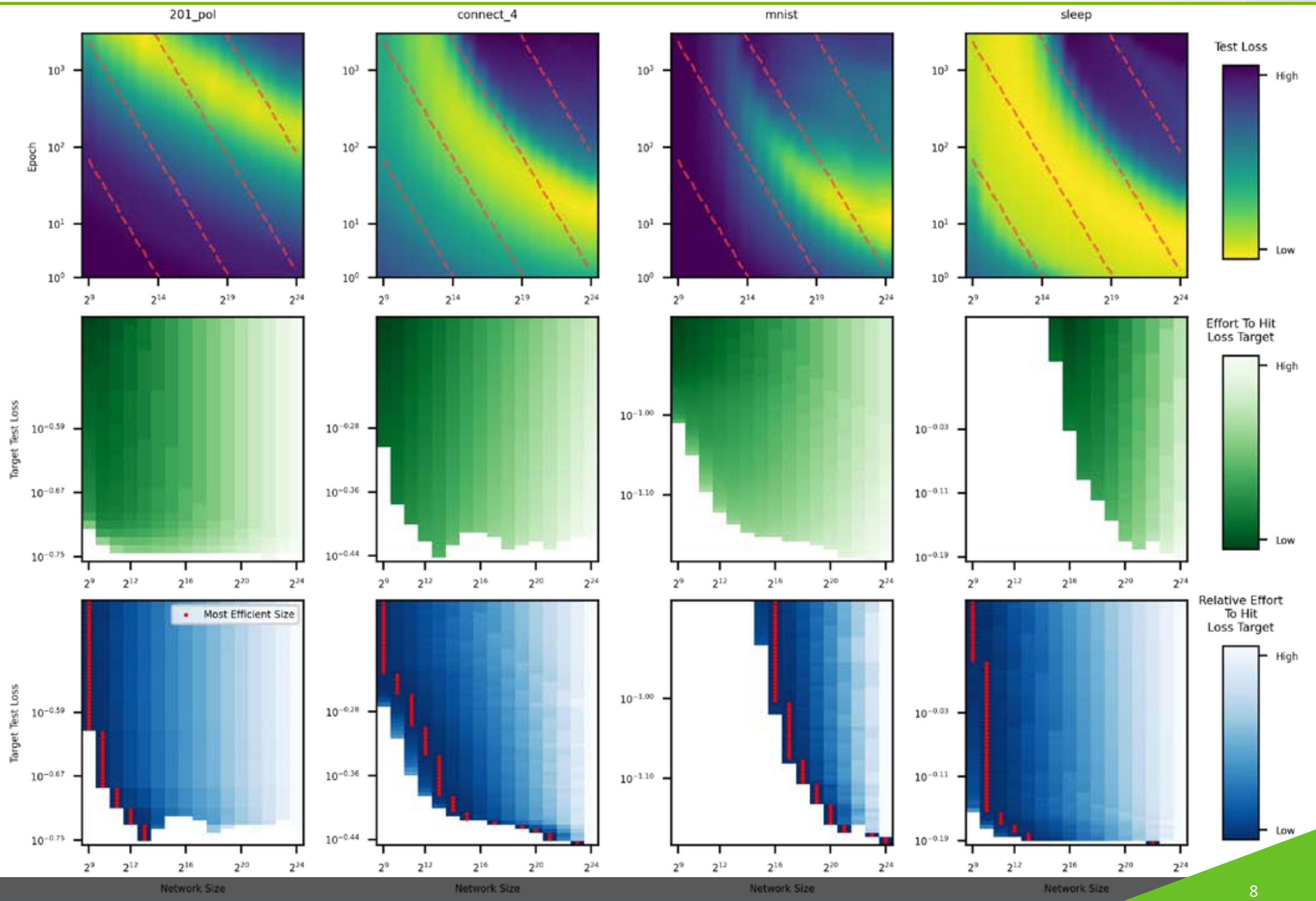


Empirical Deep Learning Efficiency Study

Larger models can perform better and achieve peak performance in fewer training epochs.

However,

- The cost of increasing model size outpaces efficiency gains from fewer training epochs.
- The most efficient network to achieve a target loss level is only slightly larger than the smallest network that can achieve it.
- Starting small and training increasingly larger networks until the network's performance is acceptable is likely to be considerably more efficient than starting with a heavily oversized network.

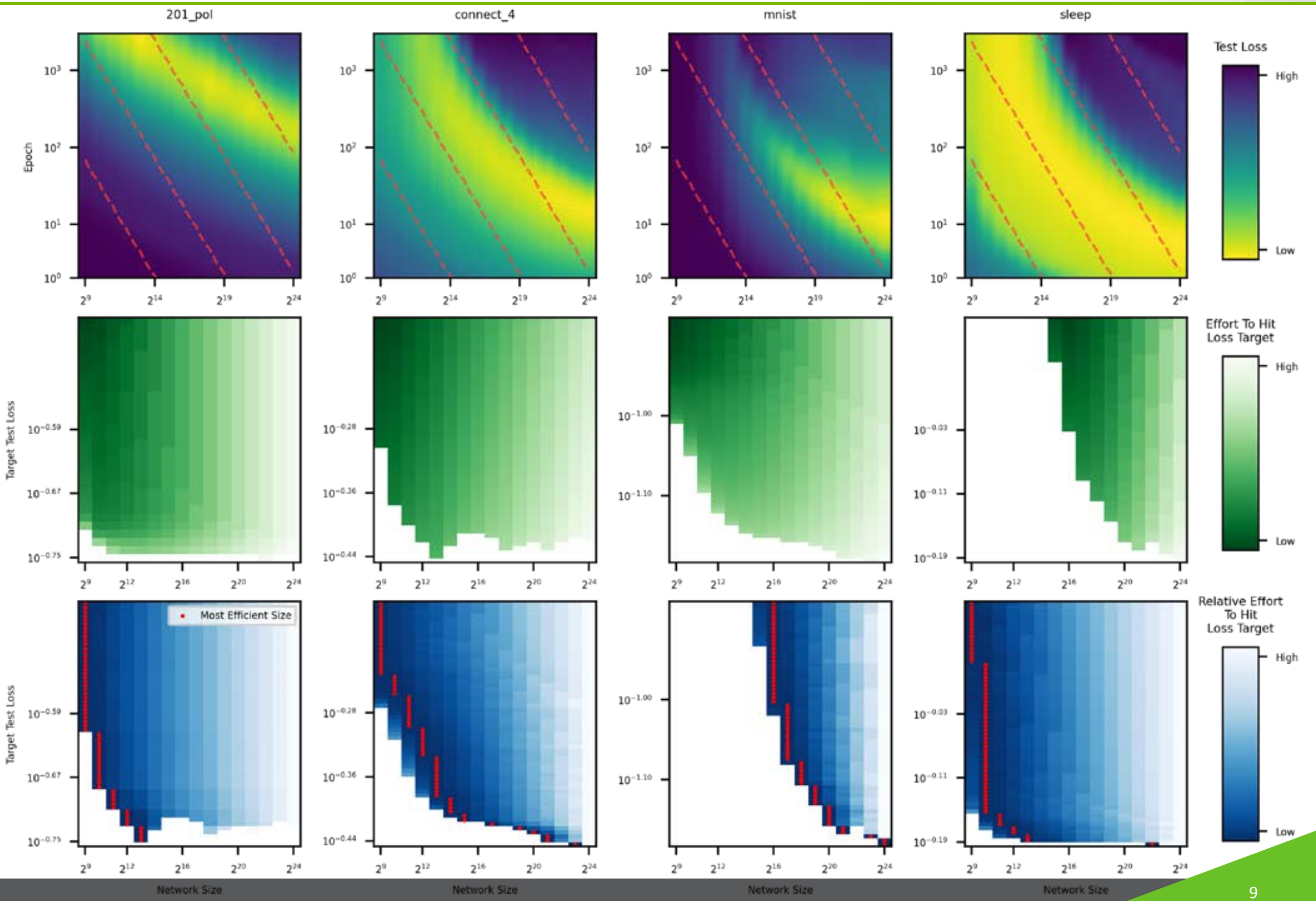


AI Efficiency Research Directions

AI is a primary driver of the looming computing energy crisis.

However, with more research:

- Large networks contain redundant and non-functional components which could be elided.
- We could interleave topological and parameter optimization to find efficient networks efficiently.
- More efficient model types beyond backpropagation based neural networks could yield massive efficiency gains over deep learning approaches.



Thank you!

NREL/PR-6A50-82594

www.jisea.org

This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the Joint Institute for Strategic Energy Analysis. The views expressed herein do not necessarily represent the views of the DOE, the U.S. Government, or sponsors.

JISEA Joint Institute for Strategic Energy Analysis

NREL Transforming ENERGY
COLORADO SCHOOL OF MINES EARTH • ENERGY • ENVIRONMENT
COLORADO STATE UNIVERSITY
MIT Massachusetts Institute of Technology
Stanford University
University of Colorado Boulder