JISEA Joint Institute for Strategic Energy Analysis

Green Computing Opportunities & Strategy

Charles Tripp, Hilary Egan, Erik Bensen, Jordan Perr-Sauer, Nicholas Wimer, Ambarish Nag, Sagi Zisman, Jamil Garfur

4/26/2023

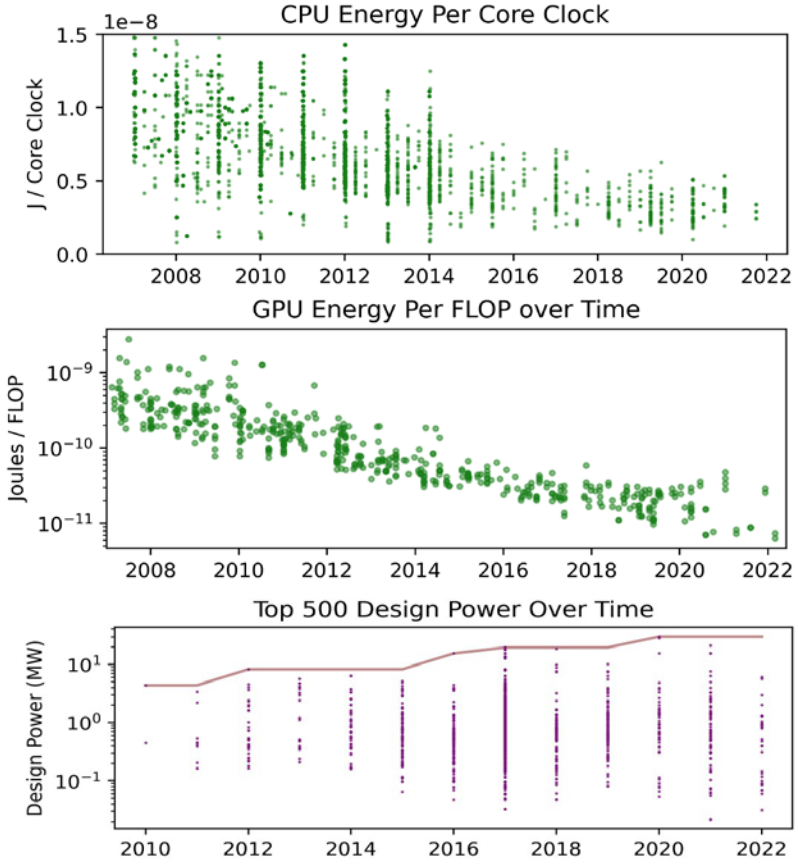
NREL Transforming ENERGY
Colorado School of Mines
Colorado State University
Howard University Office of Research
Massachusetts Institute of Technology
Stanford University
University of Colorado Boulder

Agenda

- A Brief History of Energy & Computing
- The Growing Energy Costs of Computing
- What Can We Do? Addressing the Challenge
- Green Computing @ NREL
  - Green AI @ NREL

- Collaborative Discussion

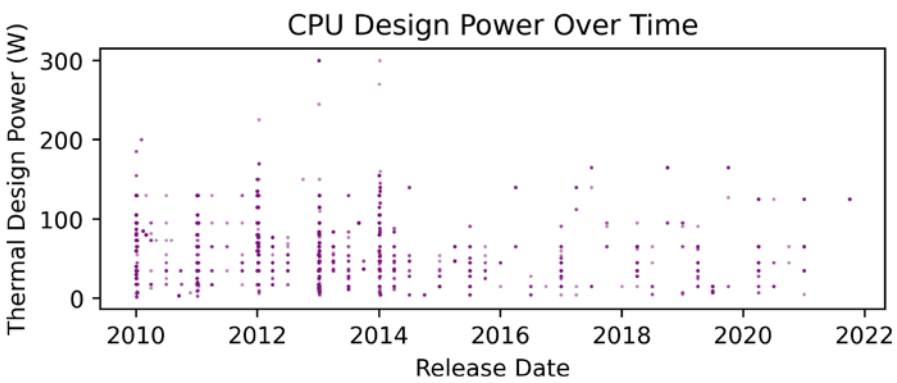
Scaling Through Hardware Advances is Intractable

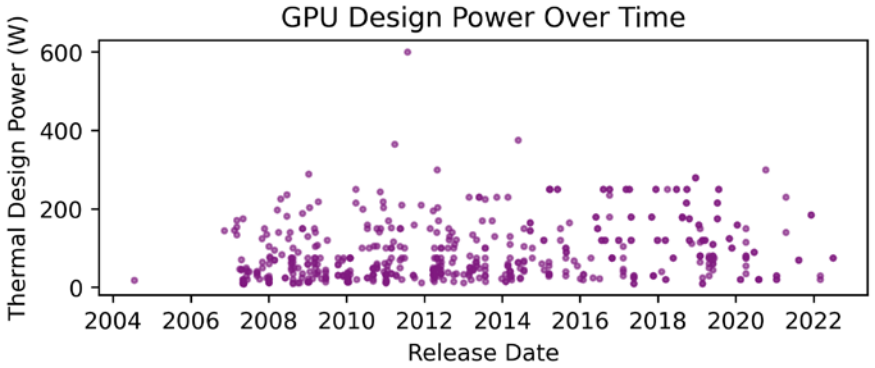


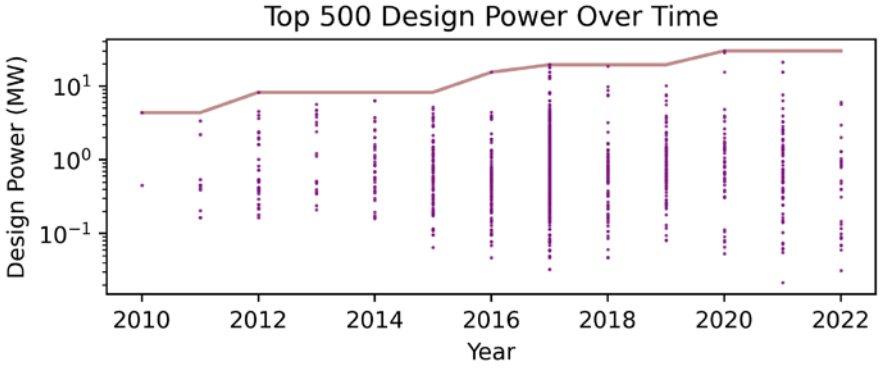
- Gains in hardware efficiency have reduced the energy per operation cost, but not as rapidly as computing demand has grown.
- Worse, efficiency improvements are beginning to level-off.
- Example: the fastest computing systems are consuming an exponentially increasing amount of power despite gains in efficiency

Figure data from: ark.intel.com, github.com/toUpperCase78/intel-processors,
Danowitz, A., Kelley, K., Mao, J., Stevenson, J. P., & Horowitz, M. (2012). CPU DB: recording microprocessor history. *Communications of the ACM*, *55*(4), 55-63.,
Top 500, top500.com

Computing Systems are Gradually Using More Energy







- Per-device power is increasing
- Number of devices is increasing
- Total energy consumption is increasing

- Growth in mobile-centered products may confound these plots somewhat, making it appear that design power is not increasing overall.

The Rising Energetic Cost of Computing

- Total computing energy usage has been growing exponentially for more than a decade.
- If this trend continues, computing is forecasted to consume >8 PWh, or **~20% of generated electricity by 2030**
- **It is time to invest in efficient algorithms, software, and operating strategies.**

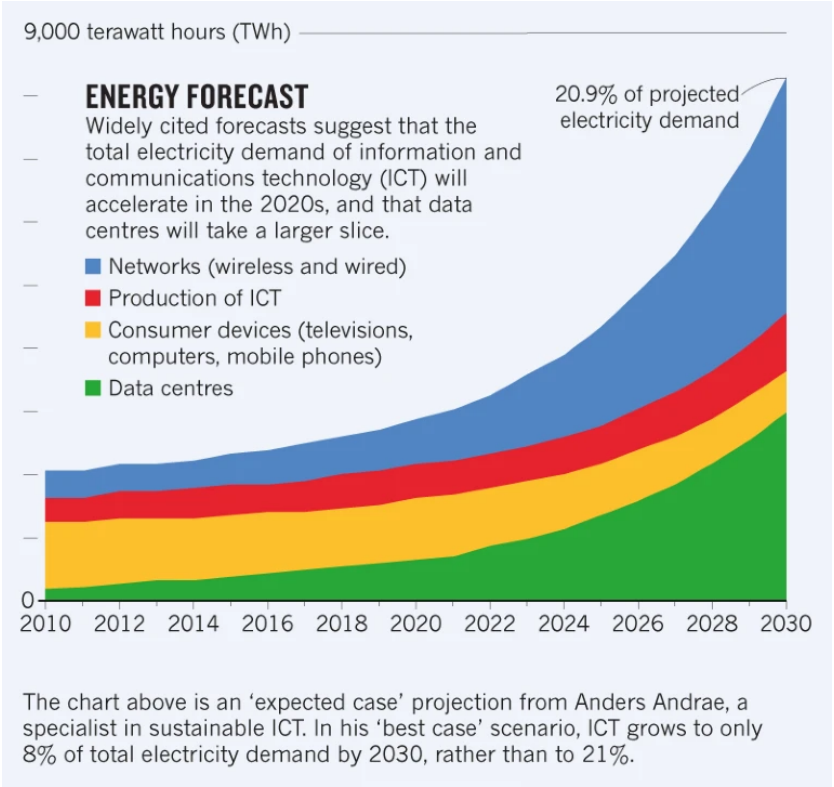


Figure from Jones, N. (2018). How to stop data centres from gobbling up the world's electricity. *Nature*, *561*(7722), 163-167. https://www.nature.com/articles/d41586-018-06610-y

The Rising Energetic Cost of Computing

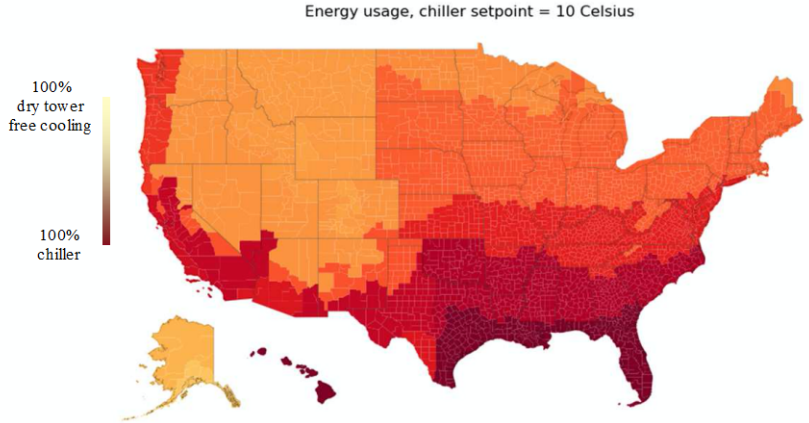
- National Lab Scientists: Computing becomes constrained by energy use.
  - Allocated a computing energy budget?
  - Can you afford the energy to do the computation?
- DOE: Failing to foresee and address a rapidly escalating consumer of energy.
  - May take a decade or more to catch up
- Industry: Commercial activity and innovation limited by computing energy demands.
  - Energy for computing becomes the de-facto currency
- National Security: Internal and external stresses on computing energy demands breed instability through conflicts between energy consumers, producers, and stakeholders.
- Humanity: Long-increasing benefits from computing stagnate, stagnating human progress.
  - Computing energy production unequally impacts disadvantaged groups while primarily benefitting advantaged groups.
- Nature: Increasing carbon emissions and energy production demands reduce quality and diversity of life on the planet.

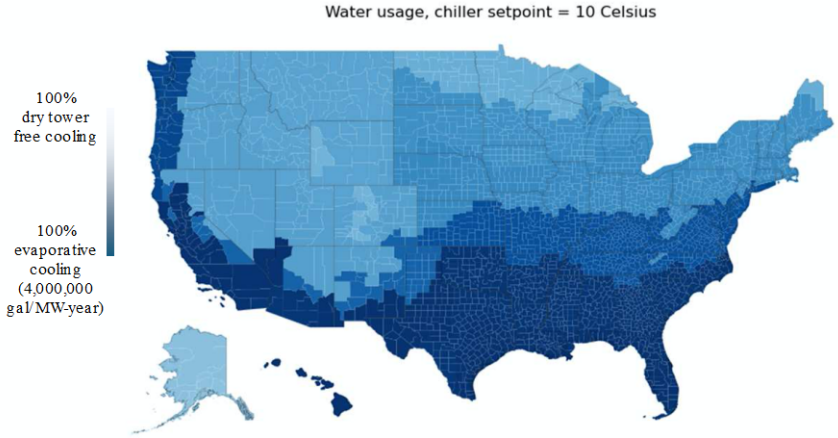
Addressing the Challenge

- Systems Management Efficiency
  - Dynamic, grid-integrated load-shifting and throttling strategies
  - Advanced cooling systems
    - Improved efficiency, reliability, and component longevity
  - Virtualization and multi-tenant systems
- Architectural efficiency
  - Optimizing system layouts for the most intensive workloads
    - network, system, operating system, and chip level
- Total environmental and energy footprint reduction
  - Accounting for the impacts of manufacturing, operation, and decommissioning may indicate surprising strategies to reduce environmental impacts of computing
- Algorithmic Efficiency
  - In breakthrough cases, gains can be exponential in nature

Facility supply temperature key driver for energy use

50 ºF





Common facility temperatures 10-32°C. Example to illustrate effect of facility Temperature on Energy use

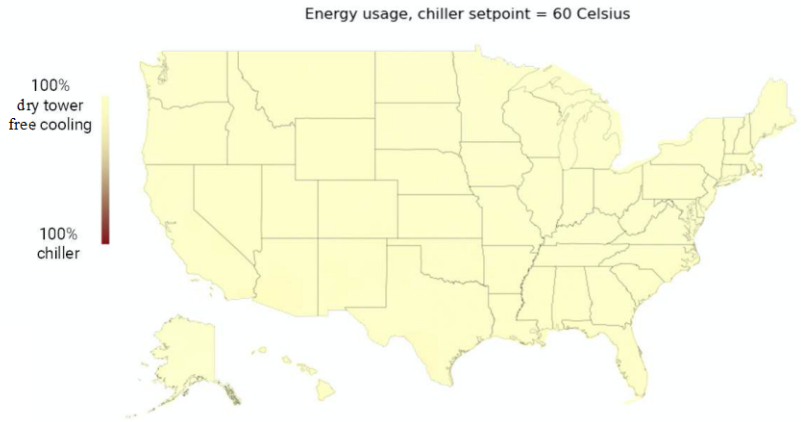
- The price paid for the standard supply temperature low is Excessive Energy Use
  - Chiller has to run most of the year – 0.75 quads for cooling
  - Water is consumed in most locations – *approx. total of >500 billion gallons of water use attributable to US data centers (~57% sourced from potable water)*

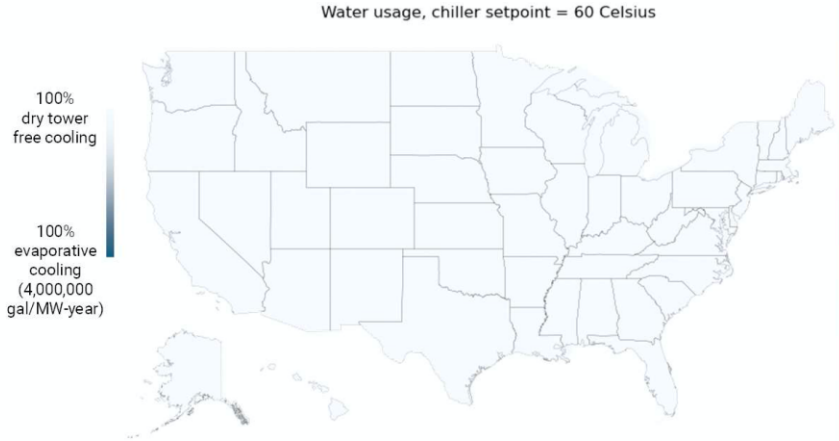
https://www.nature.com/articles/s41545-021-00101-w

Credit: Peter de Bock (from DOE ARPA-E *Cooling Compute Systems Efficiently, Anytime, Anywhere Workshop, December 2021*)

Efficient heat rejection can Change the Landscape

140 °F





If technology is developed to reject heat from future servers 10 x more efficient in secondary loop (chip to facility supply); facility temperatures can be evaluated, and cooling energy is saved

Bonus features:

+ Location/ climate independence
+ Minimal/ No need for water usage
+ Reduced footprint
+ Heat rejection >60°C facilitates future WHR

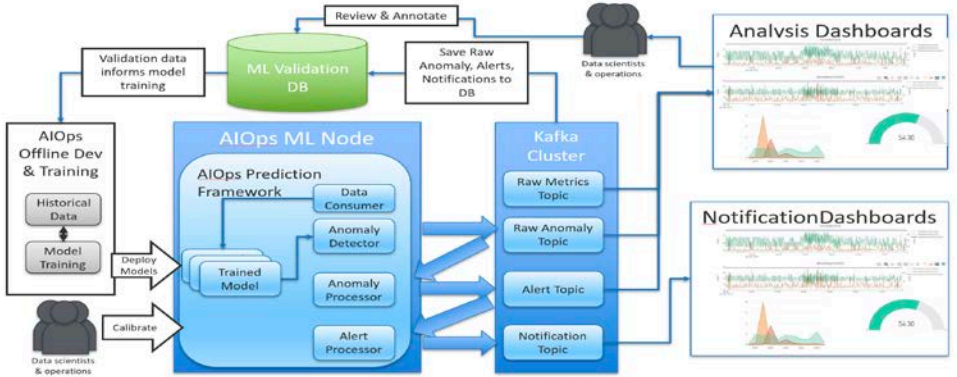
Credit: Peter de Bock (from DOE ARPA-E *Cooling Compute Systems Efficiently, Anytime, Anywhere Workshop, December 2021*)

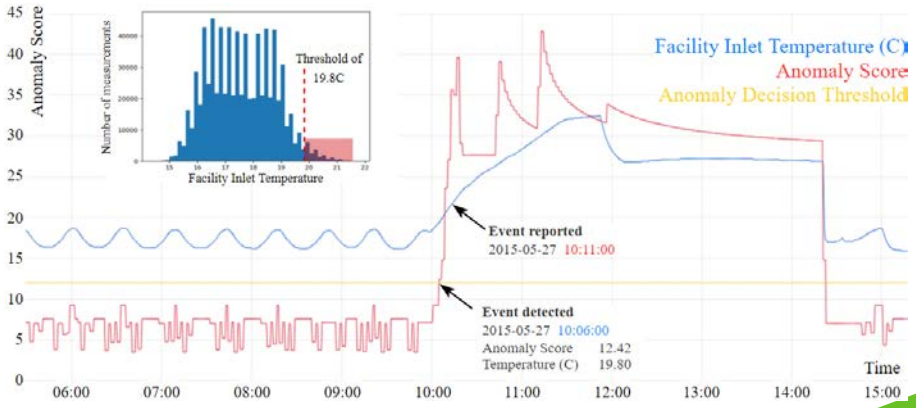
Algorithms for Sustainable Operations: AI for Data Center Operations (AIOPs)

- Enabling efficient HPC operation through
  - Anomaly detection
  - Predictive maintenance
  - Operational optimization (PUE)
  - Root cause analysis

- Real-time streaming data
  - **Eagle:** 1M metrics/min
  - **ESIF data center:** 4k metrics/min

- AI models and visualizations trained with historical data and deployed on real-time data

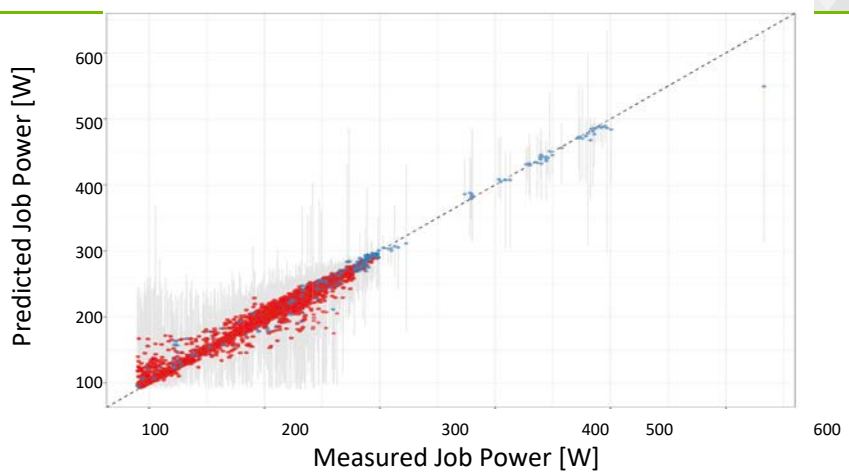


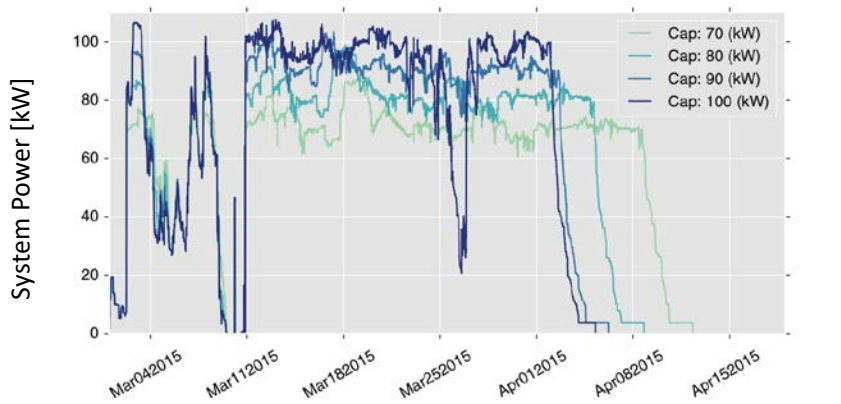


Work led by Hilary Egan

Dynamic job scheduling for load shaping

- Predicting job power usage a priori can enable sophisticated scheduling to minimize grid impacts
- Can respond to grid variability, dynamic electricity pricing/renewable availability, etc.
- Future work could address where and on what resources a job is run, not just when





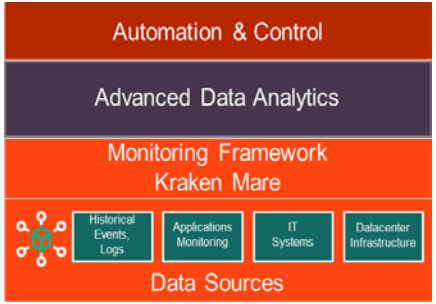
Work led by Caleb Phillips, Hilary Egan

HPC Green Computing Efforts @ NREL

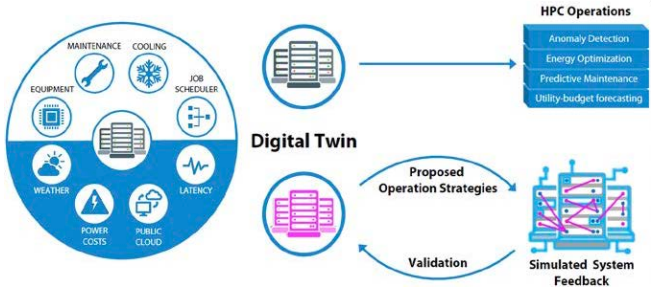
HPE

HPE Track

NREL

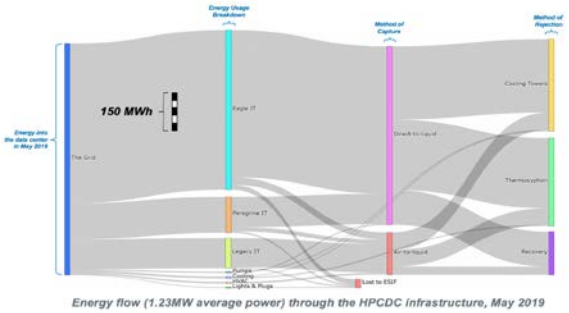


AIOps

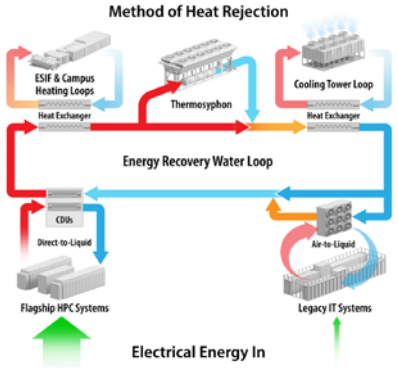


Digital Twin -Data Centers

Hydrogen Fuel Cell



Data Center Efficiency

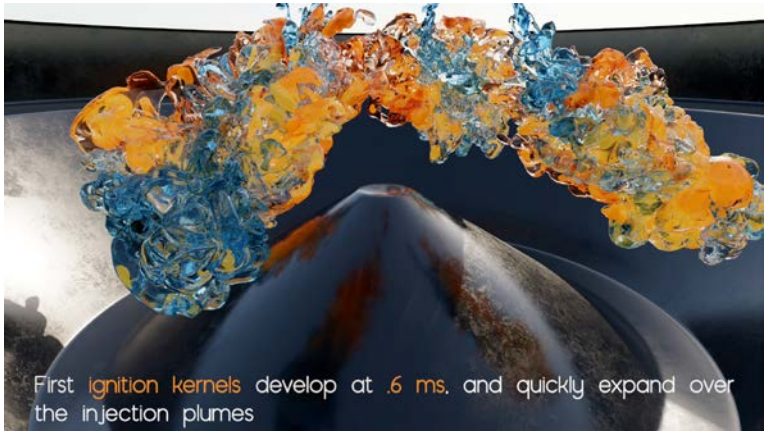


Advanced Cooling

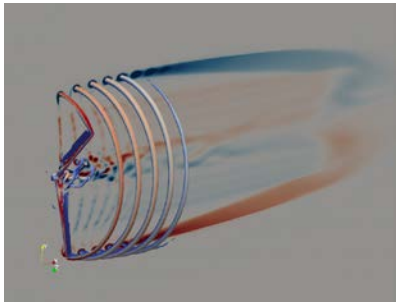
Water Usage

Computational Fluid Dynamics Research @ NREL

- Computational Fluid Dynamics (CFD)
  - Uses highly parallelized computer software to solve the underlying equations of fluid motion
  - CFD is used in a wide range of engineering applications from wind energy, combustion science, bio-reactors, aerodynamics, etc.
  - These simulations routinely require thousands of computational nodes to evolve the millions to billions of computational grid cells needed to resolve relevant physical phenomena
- Exascale Computing
  - New CFD codes have recently been developed at NREL to leverage exascale HPC systems, namely ExaWind and the Pele Suite of exascale codes
  - These codes have been parallelized on up to 56,000 GPUs
  - They are highly optimized for computational efficiency on both CPUs and GPUs
  - Energy efficiency analysis on these codes will provide a new metric with which to evaluate the performance on this class of software
  - Future development can then be aimed at developing numerical algorithms that balance both computational and energy efficiency metrics



Still from exascale simulation using PeleC. Simulation was run on thousands of GPUs evolving trillions of variables to compute the combustion phenomena in a clean burning diesel engine.



Simulation showing the flow structures from a 5 MW wind turbine using Nalu-Wind, an exascale CFD software developed out of the ExaWind project at NREL.

Red AI: Deep Learning's Energy Footprint

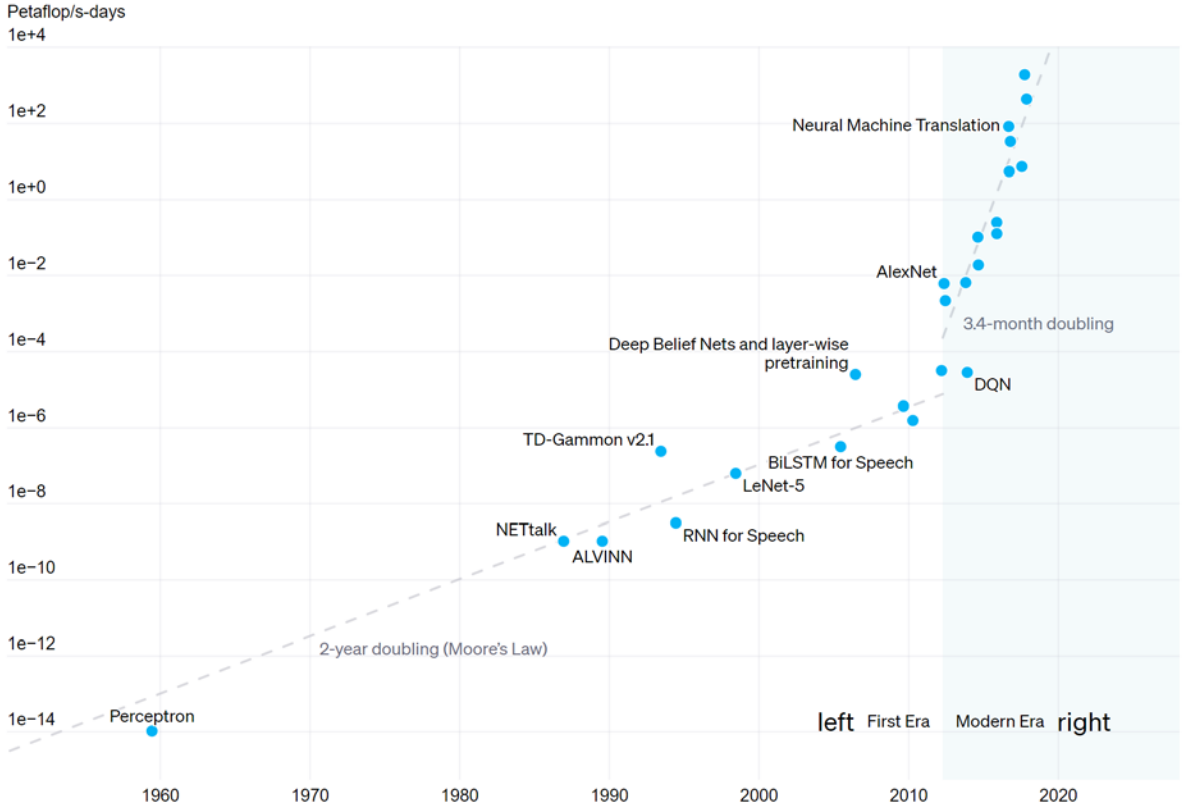


Figure From Amodei, D. and Hernandez, D. AI and compute, 2018. *Open AI Research Blog*. https://openai.com/research/ai-and-compute

Deep Learning's Energy Footprint

- Computing is an increasing consumer of energy resources worldwide
- AI is a popular consumer of computing resources
- Within AI, Deep Learning is both highly prevalent and computationally expensive
- Training and applying popular neural network models are consuming vast quantities of energy and incurring large carbon footprints.

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.

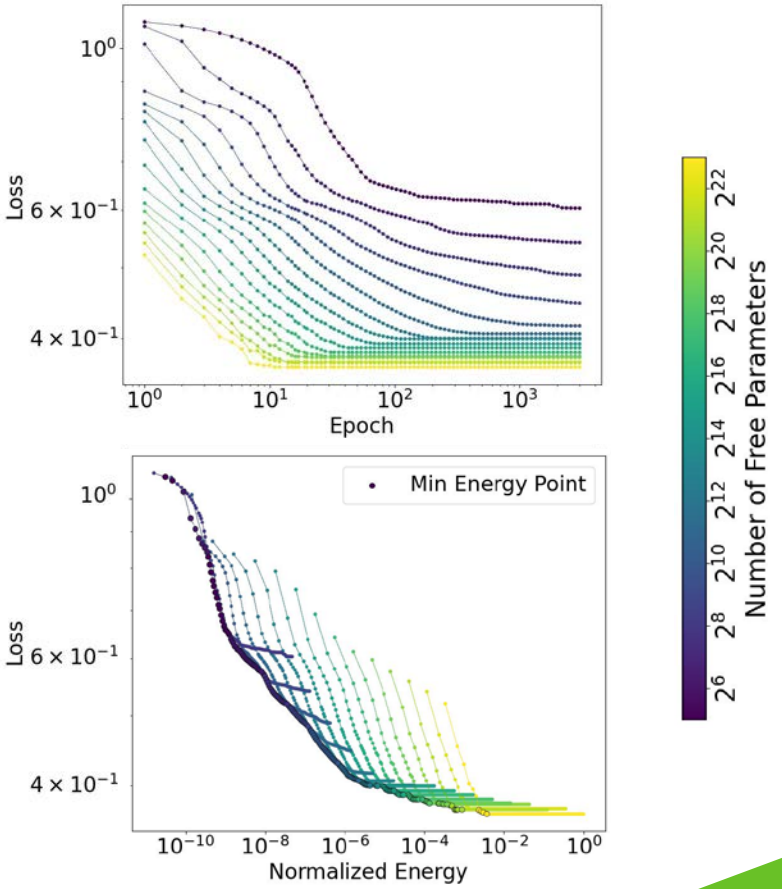
Table from [3] Strubell, E., Ganesh, A., & McCallum, A. (2019, July). Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3645-3650).

Green AI: Right-sizing neural networks

- Smaller neural networks are more efficient: empirical evidence shows that 90-99% of many models can be discarded without losing performance

- We are developing methods for following the lowest-energy training pathway by dynamically growing neural networks during training, following the lowest energy point along the loss-energy pareto frontier.
  - 10x energy efficiency gains are possible

- Further order-of-magnitude gains are likely possible by strategically pruning networks to optimize for energy efficiency.



Work led by Charles Tripp

Addressing the Challenge at NREL

- By investing in green computing research now, we are establishing a program that will be well-prepared for the increasing calls to reduce computing's energy demands.
  - *Current research portfolio addresses sustainability in computing **from algorithms to facilities***

- Establishing Partnerships
  - **Industry interest in green computing is strong and universal**
  - Funding agency interest is moderate
  - Engaging Industry & Academic Partners to form a Green Computing Consortium

  - **Collaborations across NREL**: Computing, buildings, mechanical engineering, policy
  - **Industry partnerships**: Hardware, software & cloud providers, applications (blockchain)
  - **Community engagement**: HPC Data Center Community, OCP (Open Compute Project) Heat Reuse group, Energy Efficiency HPC Working Group, Data Center Dynamics, other pre-competitive research consortia

What should we do today to maximize our impact?

- What are your ideas for establishing productive Green Computing partnerships & collaborations?
  - Industry, Academic, Research Groups, Labs, etc.
  - Which potential partners we should engage?
    - Do you have / know of any connections to initiate the process?
  - Possible networking opportunities

- How can we form a strong coalition of green computing partners?
  - How can we propagate strong industry interest to funding agencies?
  - How to fund our efforts in the interim?

What should we do today to maximize our impact?

- Which topics should we focus our research efforts on to maximize our success?
  - Measurement capability
  - Optimizing HPC/Datacenter Operations
  - Optimizing Datacenter Cooling Efficiency
  - Algorithmic Optimization
    - CFD
    - Green AI
  - Are there other areas we should consider breaking into?

Thank you!

NREL/PR-6A50-87115

www.jisea.org

This work was authored 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, and the National Renewable Energy Laboratory. 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 · Colorado State University · Howard University Office of Research · Massachusetts Institute of Technology · Stanford University · University of Colorado Boulder