The Entropy Economy: A New Paradigm for Carbon Reduction and Energy Efficiency for the Age of AI

Scott Evans
GE Research
Niskayuna, NY
evans@ge.com

Tapan Shah
GE Research
San Ramon, CA
tapan.shah@ge.com

Acahlesh Pandey
GE Digital
San Ramon, CA
achalesh.pandey@ge.com

2021 55th Asilomar Conference on Signals, Systems, and Computers | 978-1-6654-5828-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/IEEECONF53345.2021.9723336

***Abstract*—We introduce a new paradigm for minimization of carbon and maximization of energy efficiency: The Entropy Economy. Our approach addresses the predicted exponential rise in energy consumed by compute within the next decade and proposes Energy Aware Machine Learning (EAML) together with grid architectures and distributed High Performance Compute (HPC) infrastructure to jointly optimize learning, energy efficiency, and disposition of waste heat. We introduce a "Kolmogorov Learning Cycle", that enables characterization of the efficiency of the learning cycle, and assert that the precious resource to be conserved in the age of AI is entropy: maximizing the entropy reduction (in learning) while minimizing entropy flow loss (through thermodynamic inefficiency) to minimize carbon production, maximize energy efficient learning, and stabilize the grid. We present straw man case studies and initial EAML results showing how the Entropy Economy can reduce carbon reduction while leveraging trades between Machine Learning Model Quality, Energy Cost and throughput.**

***Index Terms*—Entropy, Energy Efficient Computing, Carbon Reduction**

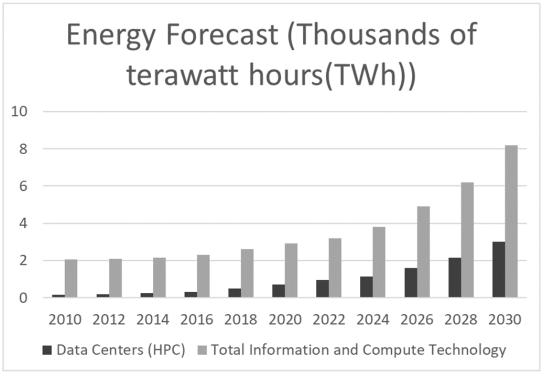

Fig. 1. Projected Energy Rise from Compute. By 2030, over 20% of energy consumed is projected to be consumed by compute, with 10% powering High Performance Compute (HPC). [2]

I. Introduction

The figure of merit for optimizing energy systems in the past decades has been Levelized Cost of Energy (LCOE) [1]. Optimization of LCOE has been effective in driving renewable energy and other carbon reducing energy generation methods, including battery storage and the hydrogen economy. But by 2030, over one-fifth of the energy consumed globally is estimated to be consumed by computation [2], see Figure 1.

To drive carbon production from AI to manageable levels, joint optimization of efficient energy production and distribution along with efficient and effective computational use of that energy – what we call Information Work – will be required.

In this paper, we introduce the concept of "Entropy Economy". This compliments and contrasts with our current "Energy Economy", that seeks to manage the precious resource of Kilowatt Hours. The key principle of the "Entropy Economy" is the joint optimization of computation and energy made possible by managing the precious resource of entropy flow that is shared by both energy and computational systems. Today energy systems and computational systems are optimized separately. The Energy Grid strives to provide low cost power everywhere in the world, irrespective of how efficiently that energy is used. In contrast, CERN, for example, has created a computational grid for sharing computational capacity for physicists throughout the world that optimizes use of computational capacity, irrespective of energy use. The Entropy Economy provides a means of jointly optimizing these resources, making possible exponential reduction in carbon compared to the non-jointly optimized case. After defining the concept of "Information Work" we introduce and quantify two means of optimization for the Entropy Economy: (1) moving information work to where clean, low cost energy exists through an optimized compute/energy grid architecture linking wind farms with collocated HPC capacity, and (2) creating computational methods that allow tradeoffs between energy consumed and learning through Energy Aware Machine Learning (EAML) techniques. Together these methods can provide a path to reducing carbon, optimizing efficiency, and stabilization of the Energy Grid in the age of AI.

This work was funded by GE Research.

A. Prior Work

Opportunities to reduce Carbon by leveraging stranded Renewable power is explored in [3]. The concept of entropy applied to information and learning was introduced by Shannon in his seminal paper: [4]. Compression as learning is explored by Adriaans in [5], utilizing principles of Kolmogorov Complexity and Minimum Description Length [6], [7]. The thermodynamics of computing and related concepts have been recently developed by Wolpert et. al. [8], [9].

B. This Work

We bring together many of the concepts above into an overall structure we call the Entropy Economy with the thesis that that precious resource to be managed in the age of AI is entropy flow: - its reduction as learning takes place, over and against its loss as lost waste heat from thermodynamic and computational systems. We begin by defining the Entropy Economy and its relationship to learning and thermodynamics, and introduce a "Learning Structure Function" and "Kolmogorov Cycle" that build upon ideas of Kolmogorov Structure Function and Carnot Heat Cycles, respectively. We then describe grid architectures that can make use of the Entropy Economy to reduce carbon release, and discuss Energy Aware Machine Learning (EAML) as a driving force to enable optimization of the Entropy Economy through the ability to trade Machine Learning Model Energy Cost for Qualilty and/or Throughput. We conclude by discussing next steps.

II. The Entropy Economy

A. Equivalence between energy and information work

In 1961 the Physicist Rolf Landaur posited that the minimum possible amount of energy required to erase one bit of information was equal to:

$$E = ST = k_B T \log_e 2 \tag{1}$$

where S is entropy, k_B is the Boltzmann constant and T is temperature in degrees kelvins. At room temperature this comes to $0.0175eV$ and derives from the fact that $E = ST$ energy must be emitted into the environment if the added entropy $S = k_B \log_e 2$ flows to that environment [8]. The same thermodynamic entropy that drives thermodynamics is precisely the same entropy Shannon introduced when he launched Information Theory in 1948 [4], and enables the joint optimization of energy systems with computation systems. Just as entropy is central to efficiency in thermodynamic cycles such as the Carnot engine and the Rankine cycle, entropy flow is central to efficiency in computational systems, and is in fact a measure of information content and thus learning. The first law of thermodynamics states that energy is conserved. The second states that Entropy production is always positive. The Carnot cycle, shown in Figure 2 is a well known thermodynamic cycle that provides means of assessing the energy efficiency of heat engines. The inefficiency of the cycle can be determined by the area under the Temperature/Entropy Curve - the area under the lower curve is energy lost as waste heat.

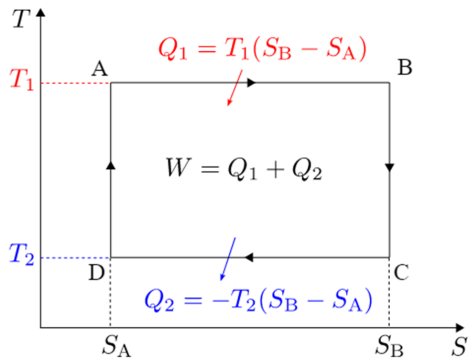

Fig. 2. The Carnot cycle enables the assessment of energy efficiency in the transformation of heat to work. Efficiency is high when a large temperature difference exists in the engine, as the medium transforms in entropy

In the same way we envision what we would like to call a "Komogorov Learning Cycle," shown abstractly in figure 5, where we can assess the thermodynamic efficiency of learning as measured by level of compression.

B. Compression as Learning

Entropy and Mutual Information are key drivers of numerous machine learning algorithms, with Shannon Entropy being the compression bound for a sequence. In [5], the author argues that Compression, as bounded by Kolmogorov Complexity and in the sense of a two part Minimum Description Length model of a given data set, represents optimal learning, where "optimal compression theoretically represents the optimal interpretation of the data" . Thus the extent to which a machine learning model can lead to compression of a data set represents the amount of "Learning Work" that has been achieved.

This provides an additional dimension to assess our joint optimization of compute and energy: The entropy released through learning of data.

The learning process can be viewed through the lens of Kolmogorov Complexity through the Kolmogorov Structure Function, shown in Figure 3. Kolmogorov Complexity, $K(X)$, is the size of the smallest program that can be written that will run on a universal computer and produce the string X as output. Algorithmic Information Theory and Minimum Description Length principles tell us that this quantity can be split into a two part code consisting of the size of the program to print out the typical set to which string X belongs, which we call $K(S)$, plus the number of bits needed to identify which of the entries in set S[1] corresponds to string X.

$$K(X) = K(S) + \lceil (\log_2 |S|) \rceil$$

We think of $K(S)$ as the "model cost" and $\log_2(|S|)$ as the "Data Cost." As shown in the figure, when zero bits are

[1]The term S in this context is size of the set and not the entropy.

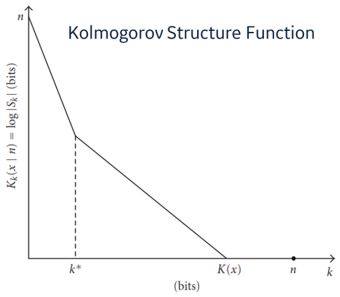

Fig. 3. Kolmogorov Structure Function: The x-axis represents the model cost and the y-axis represents the data cost.

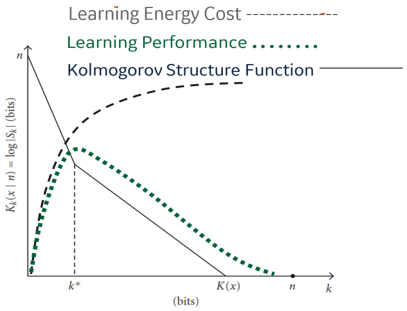

Fig. 4. Learning Structure Function: Overlaying energy cost of learning, learning performance on Kolmogorov Structure Function.

allocated to the model, the Kolmogorov Complexity estimate consists of the log of the size of the set of all possible strings of size n. As bits are allocated to the model, the size of the set decreases, reducing the "data cost" while increasing the "model cost." At the point k^*, shown in Figure 3, we have found the "Algorithmic Minimum Sufficient Statistic," [6], which represents the optimal model that captures all of the essence of the data without overfitting. In principle, the goal of Machine Learning is to find k^*: Models with less descriptive cost do not capture the full extent of learning available in the data, while models greater than k^* are subject to overfitting.

A key aspect of learning that is not included in the concept of Kolmogorov Complexity is the energy cost of learning the models. In Figure 4, we overlay the Kolmogorov Structure Function with "Learning Structure Functions" that denote the energy cost in learning the model as well as the Learning Performance for which a model is capable. Optimal Learning Performance can be achieved at the Algorithmic Minimum Sufficient Statistic, k^*. The energy cost increases as the search for the optimal model takes place. If the search for a model continues beyond $k*$, that energy is wasted, since model performance will be reduced due to over-fitting.

The idea of the "Learning Structure Function" enables us to characterize the Kolmogorov Learning Cycle shown in Figure 5. We define the "Learning Work" as the process of

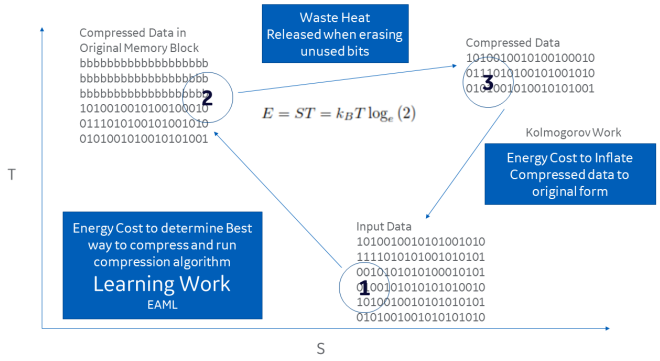

Fig. 5. Kolmogorov Learning Cycle. $1 \rightarrow 2$: Reduction in entropy of input data, $2 \rightarrow 3$: Bit erasure and increase in entropy, $3 \rightarrow 1$: Bit write (and temperature reduction).

compressing, or learning from an input data set to capture its essence. This learning work may consist of searching different types of Machine Learning Models to the best performance, and the effect of the learning will be to reduce the entropy of the input data set as the Kolmogorov Learning Cycle moves from state 1 to 2. The energy required to learn will increase the temperature of the thermodynamic compute system, and result in a compressed data set, along with a set of bits that are no longer needed and can be erased. When these bits are erased, energy is released into the system as the cycle moves from state 2 to 3, as previously discussed. Kolchinsky and Wolpert have recently characterized the progression from state 3 back to 1 - the amount of energy required to run a program of size $K(X)$ on a Turing Machine to produce x as the "Kolmogorov Work" [9]. This is the energy cost of exercising a machine learning model to make a prediction or reproduce the original data. This will involve the system writing out more bits of information, reducing entropy and drawing heat from the system, thereby reducing temperature.

Similar to how the Carnot Engine Cycle makes possible refinement of the efficiency of heat engines to increase, the Kolmogorov Learning Cycle illustrates key areas where the Entropy Economy can be optimized to maximize learning while minimizing wasted and needless entropy flow loss. Careful attention to how much energy is expended on "Learning Work" can vastly reduce the energy used in the cycle. And simply using fewer bits when possible will result in fewer bits erased, and more efficient use of the entropy economy. This motivates Energy Aware Machine Learning as a prime opportunity for optimization.

C. Optimizing Information Work Systems

Data centers and high performance computing (HPC) centers can be considered information work machines: It inputs data and energy, transforms the data and releases heat to the environment. A lot of research is conducted in efficient transformation of this data thereby minimizing the entropy flow. For example in [8], the authors prove generalized Landaur's

bound:

$$S(p_0) - S(p_1) \leq \text{Heat (Entropy) Flow from the system}$$

where $S(p_0)$ is the (Shannon) entropy of the input data and $S(p_1)$ is the (Shannon) entropy of the transformed data. Assuming equality, minimizing the heat flow can be is equivalent to minimize the difference in the entropy of the input and output distributions. The authors propose a circuit optimization problem to minimize the heat loss.

D. Joint Optimization of Power Source and Information Work Systems

Traditionally, the power source (steam engine, wind turbine etc.) and the information work systems are optimized individually, what we can call the energy economy and information economy. However, we propose a joint optimization of the power source and the information work systems. The trade off between energy use and prediction has been explored by Still et. all [9], we present a vision for the application to the electric grid in a holistic way.

Consider a wind HPC co-located at a wind farm. Under high wind conditions and low grid demand, the wind farm typically curtails the power produced by each turbine for grid safety. At any given time instant

$$P_{total} = P_{prod} + P_{other} + P_{curt}$$

where P_{prod} is the power fed to the grid, P_{other} is the power produced but not transferred to the grid and P_curt is the curtailed power i.e power that could have been produced/used but is allowed to go *waste*. Several methods have been proposed to minimize P_{curt} by either battery storage [10] or generating hydrogen [11], [12]. Recently, there have been research on using the curtailed power to power a high performance computing center Figure 6, either partially or completely [3], [13]. We build upon this work to propose a joint optimization of power source and information work systems, To illustrate the need for joint optimization, we explain a simple example.

Suppose a Biderectional Encoder Representations from Transformers (BERT) [14], a machine learning based language model is to be trained on a HPC using a large training data set with certain minimum accuracy [2]. The HPC can be powered, both by recovering curtailed power (which is free) as well as the grid power which needs to be paid for. The optimization than boils down to modifying the training algorithm such that to maximize the use of recovered power and at the same time achieving the desired minimum accuracy. A simple modification is to dynamically change the numeric precision of representation depending on the available recoverable power. For example, we use 32 bit precision when large amount of power can be recovered and a lower bit precision when smaller amount of power can be recovered. The key thing to note is that reducing precision can *drastically* reduce the power requirement [16], [17]. This is a constrained optimization problem where the desired accuracy defines the constraints. In this research, we mathematically formulate this problem, discuss the potential solutions which can reduce the carbon emission associated training of a neural network.

[2]Training asingle BERT adds 1500 lbs of CO2, 79 hours of compute time and 1500 KWH of energy [15]

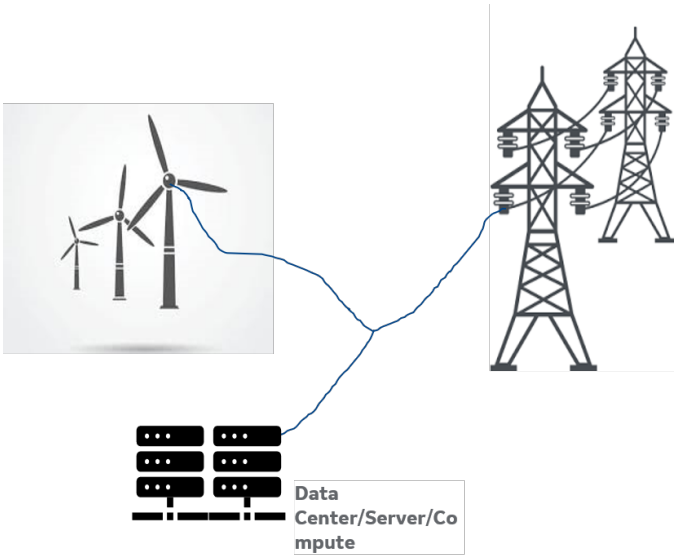

Fig. 6. A simple schematic of HPC powered directly by a wind farm as well as the grid.

While the previous example shows how energy at a single location can be harnessed to create a single machine learning model at sufficient quality while minimizing energy (and thus also carbon), the vision is for the entire grid to jointly optimize energy and minimization of carbon by employing energy aware machine learning in a distributed manner. Consider if HPC's were distributed throughout the grid that could deliver "Smart Load" services by making the tradeoffs shown in Figure 7.

Through Energy Aware Machine Learning algorithms, the energy load consumed at a given site can be traded off against the throughput (number of machine learning jobs being executed) and quality of prediction output. In this way, energy load can be balanced throughout the grid while maximizing effective creation of AI and minimizing production of carbon. Rather than considering independent thermodynamic systems, we consider a single system comprising of the power source and the information work generator. Using the Landaur's bound and the entropy computation of a wind turbine [18], the joint entropy flow of this thermodynamic system EF is bounded as follows

$$EF \geq S(p_1) - S(p_0) - P_{other} + \text{constant}$$

Thus the joint optimization of information work and power source is equivalent to minimizing the total entropy flow in the combined system.

III. Energy Aware Machine Learning

Energy Aware Machine learning (EAML) provides the capability trade energy cost of a Machine Learning Algorithm

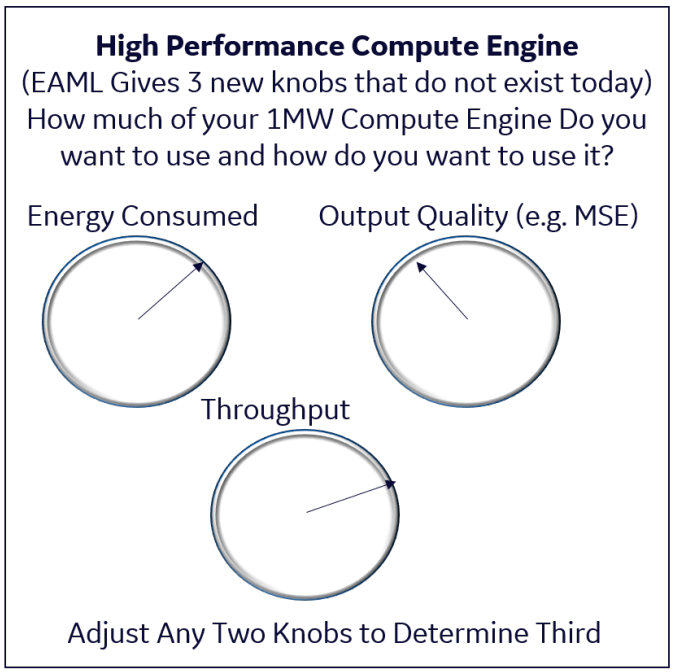

Fig. 7. Energy Aware Machine Learning Algorithms Enable HPCs to trade off Throughput, Energy, and Quality

for quality and/or throughput. This is achieved by a number of means, including: 1) reducing number of Quantization bits of the data and/or coefficients, 2) reducing the number of trees, bootstraps and/or Bayesian optimizations in model creation, 3) dimensional reduction.

Results from a simple regression problem are shown in Figure 8. Here numerical precision of the data as well as the gradients (while using the gradient descent algorithm) is varied and the learning error is determined for inputs of various entropy. Input data with higher entropy requires more precise learning in order to perform well, but lower entropy input data has almost identical performance using 4 bit precision as with using 14 bit precision. As discussed earlier, using fewer bits at the same performance implies fewer bits that will eventually need to be erased - this improving efficiency and reducing carbon cost.

Future work will produce EAML algorithms capable of adapting to a give energy profile as shown in Figure 9. Algorithms that can adapt in this way can then be used to prescribe energy profiles, which, together with scheduling algorithms can be used to stabilize the grid while maximizing energy efficient learning and reducing carbon.

IV. Impact of the Entropy Economy

Joint optimization of compute, energy, and waste heat made possible by the entropy economy will not only reduce carbon production by efficiently creating ”good enough” machine learning and AI models with maximal green energy, but also provide a means of stabilizing the grid. The Energy Economy Infrastructure creates the ultimate ”Smart Load” capability that will allow the grid to move or shed significant load

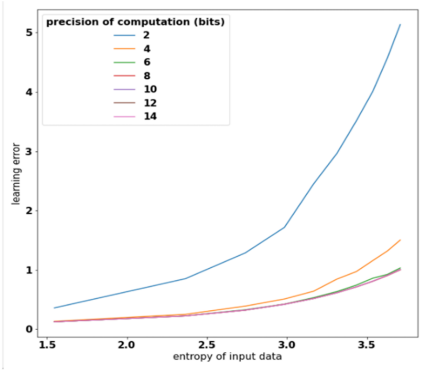

Fig. 8. Machine Learning model quality as a function of numerical precision of computation.

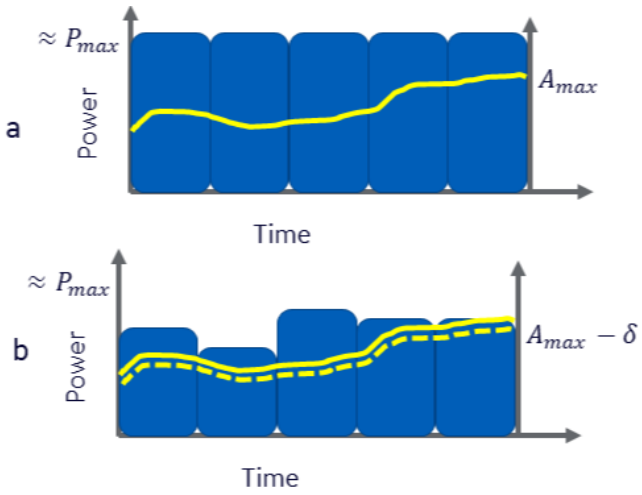

Fig. 9. Future scope of work: In a typical scenario (a), the power is not a bottleneck while learning. In (b), power available while learning is varying and the ML algorithm needs to adapt to it.

without negative consequences. Moreover, since load can be moved easily to where green energy exists, Renewable Energy Projects need not be limited by the amount of load the current grid can take, but can be built to capture more of the Wind, Solar or Hydro entitlement from the resources at hand.

A Scheduling architecture for the Entropy Economy is show in Figure 10.

The disruptive transformation made possible by the Entropy Economy is shown in Figure 11.

V. Conclusion

In this paper, we demonstrate how joint optimization of the power source and the information work can significantly reduce the carbon emissions as well as the financial cost. We coin the term ”entropy economy” and initiate a discussion on responsible and efficient information work. We identify Energy Aware Machine Learning algorithms (EAML) as key technologies to develop the entropy economy by creating means of trading energy for model quality and throughput (number of models able to be created at HPCs. Next steps

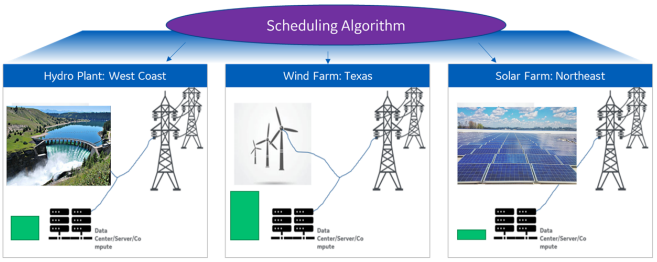

Fig. 10. Distributed Optimization of the Entropy Economy: Scheduling algorithms will move information work jobs to assets, depending on HPC and Green Energy Capacity

FROM	TO
Spinning Reserve of (Dumb) Generation Capacity	Spinning Reserve of Information Work Jobs, Movable Across Gid
Machine Learning Algorithms the Do Not Consider Energy Costs	EAML algorithms that act to Optimize and Stabilize Grid while maximizing Algorithm Quality of Service for Energy Cost
Stand Alone HPC's	Integrated Network of HPC's leveraging Stranded Power
Efficiency of Computation not considered	tCO2e / Cost Benefit of Computation
High I2R losses and Congested Grid	Grid Optimized and Stabilized through HPC network
Renewable Projects Limited by Grid Capacity	Renewable Projects optimized for Wind/Solar/Hydro Resource Available
Carbon Credits Giving the Right to Pollute	Joint Optimization of Energy and AI Reducing total Carbon to Generate Compute Work

Fig. 11. The Entropy Economy Paradigm will move from the current state on the left to the disruptive state on the right in the table above.

are to develop EAML algorithms as well as portfolio optimization scheduling algorithms and deploy them in prototype architectures to reduce carbon, maximize energy efficiency in learning, and stabilize the grid.

References

[1] "Levelized Costs of New Generation Resources in the Annual Energy Outlook 2021," p. 25, 2021.

[2] N. Jones, "How to stop data centres from gobbling up the world's electricity." *Nature*, vol. 561, no. 7722, pp. 163–167, 2018.

[3] F. Yang and A. A. Chien, "Large-scale and extreme-scale computing with stranded green power: Opportunities and costs," *IEEE Transactions on Parallel and Distributed Systems*, vol. 29, no. 5, pp. 1103–1116, 2018.

[4] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[5] A. P., "Learning as data compression," *Lecture Notes in Computer Science*, vol. 4497, 2007.

[6] M. Li and P. M. Vitnyi, *An Introduction to Kolmogorov Complexity and Its Applications*, 3rd ed. Springer Publishing Company, Incorporated, 2008.

[7] A. Barron, R. Jorma, and B. Yu., "The minimum description length principle in coding and modeling," *IEEE Transactions on Information Theory*, pp. 2743–2760, 1998.

[8] D. H. Wolpert, "Stochastic thermodynamics of computation," *Journal of Physics A: Mathematical and Theoretical*, vol. 52, no. 19, p. 193001, May 2019, arXiv: 1905.05669. [Online]. Available: http://arxiv.org/abs/1905.05669

[9] A. Kolchinsky and D. H. Wolpert, "Thermodynamic costs of turing machines," *Physics Review Research*, vol. 2, no. 3, pp. 1–22, 2020.

[10] C. Root, H. Presume, D. Proudfoot, L. Willis, and R. Masiello, "Using battery energy storage to reduce renewable resource curtailment," in *2017 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, 2017, pp. 1–5.

[11] M. A. Pellow, C. J. M. Emmott, C. J. Barnhart, and S. M. Benson, "Hydrogen or batteries for grid storage? a net energy analysis," *Energy Environ. Sci.*, vol. 8, pp. 1938–1952, 2015. [Online]. Available: http://dx.doi.org/10.1039/C4EE04041D

[12] G. Zhang and X. Wan, "A wind-hydrogen energy storage system model for massive wind energy curtailment," *International journal of hydrogen energy*, vol. 39, no. 3, pp. 1243–1252, 2014.

[13] J. Zheng, A. A. Chien, and S. Suh, "Mitigating curtailment and carbon emissions through load migration between data centers," *Joule*, vol. 4, no. 10, pp. 2208–2222, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2542435120303470

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv:1810.04805 [cs]*, May 2019, arXiv: 1810.04805. [Online]. Available: http://arxiv.org/abs/1810.04805

[15] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," *arXiv:1906.02243 [cs]*, Jun. 2019, arXiv: 1906.02243. [Online]. Available: http://arxiv.org/abs/1906.02243

[16] H. Zhang, J. Li, K. Kara, D. Alistarh, J. Liu, and C. Zhang, "ZipML: Training Linear Models with End-to-End Low Precision, and a Little Bit of Deep Learning," in *International Conference on Machine Learning*. PMLR, Jul. 2017, pp. 4035–4043, iSSN: 2640-3498. [Online]. Available: http://proceedings.mlr.press/v70/zhang17e.html

[17] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Quantized neural networks: Training neural networks with low precision weights and activations," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6869–6898, 2017.

[18] E. Asgari and M. Ehyaei, "Exergy analysis and optimisation of a wind turbine using genetic and searching algorithms," *International Journal of Exergy*, vol. 16, no. 3, pp. 293–314, 2015.