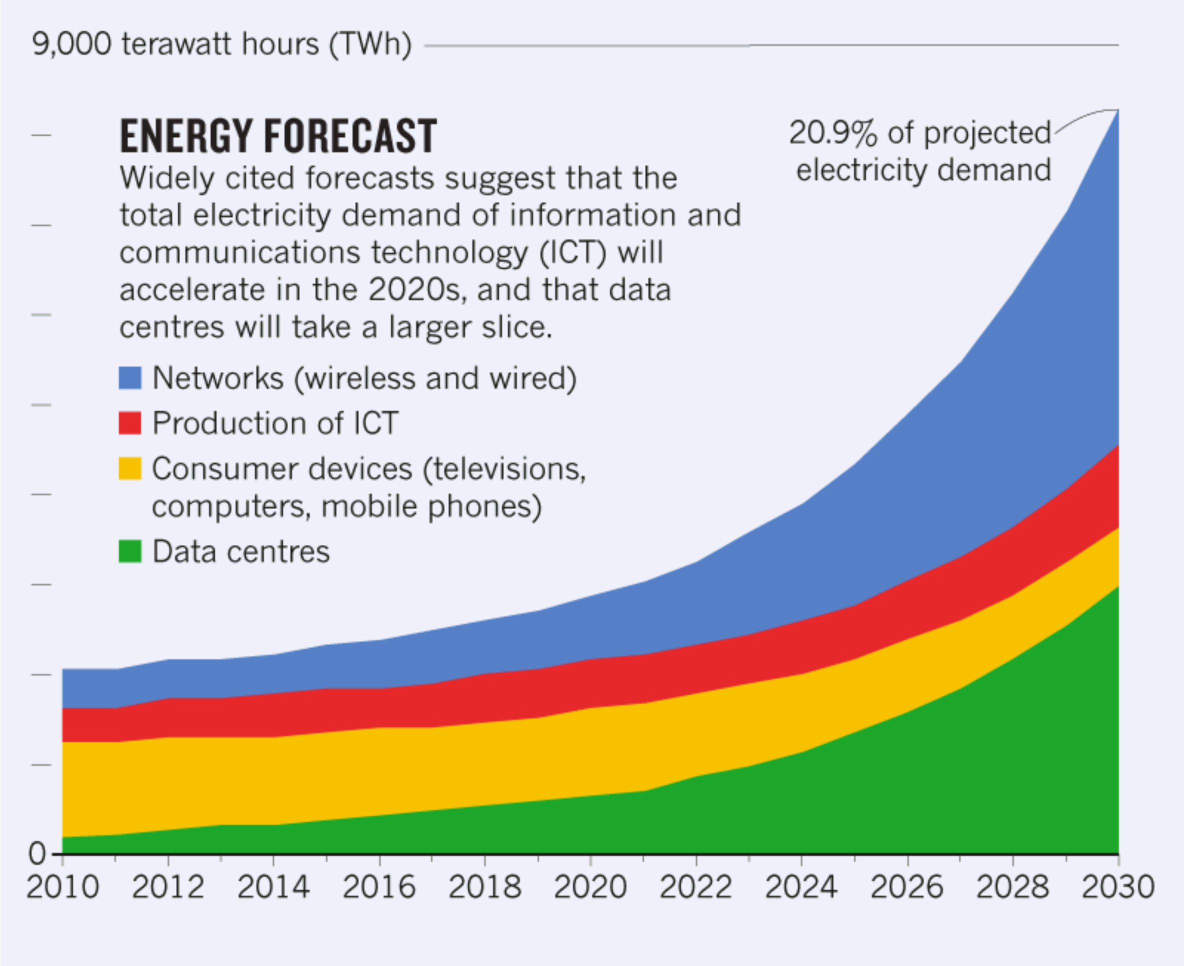
Reducing the Carbon Emissions of ML Computing
- Challenges and Opportunities -

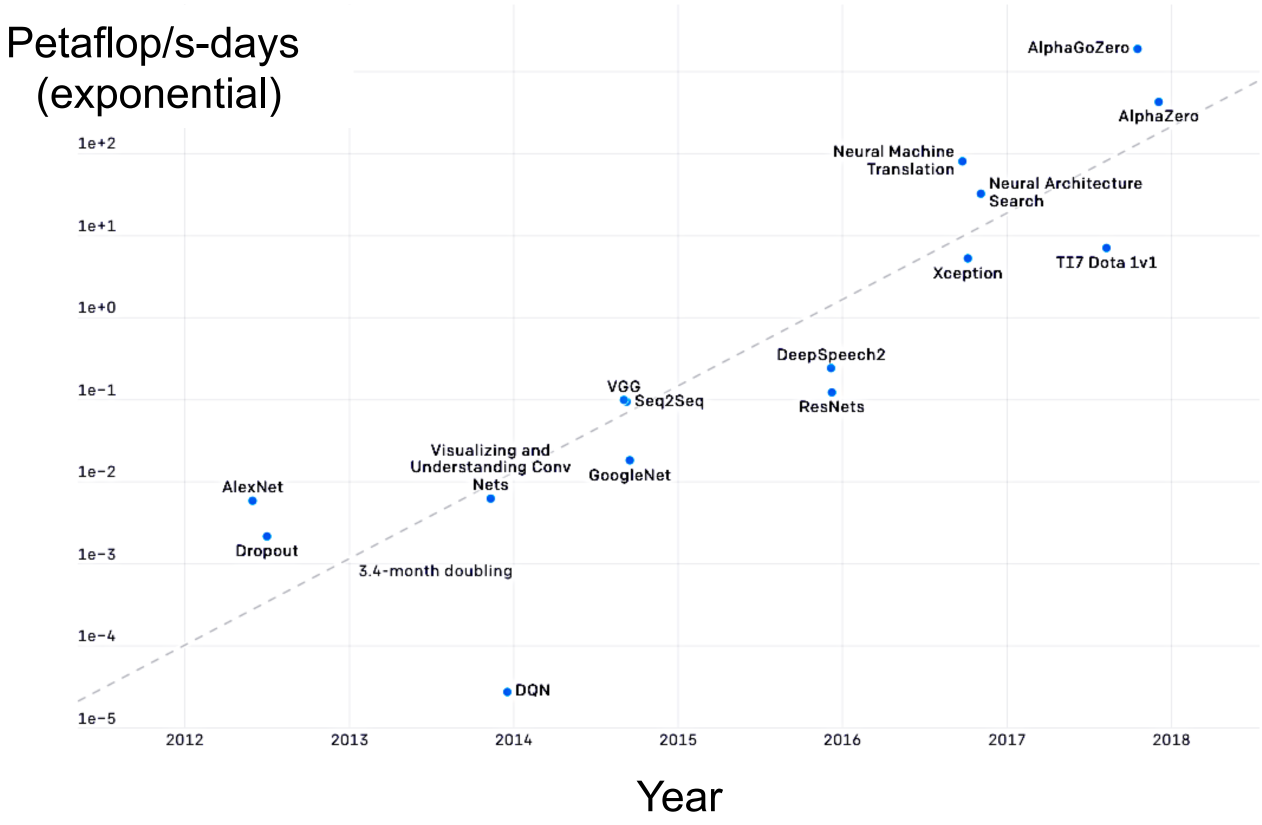
Vivienne Sze (@eems_mit)

Massachusetts Institute of Technology

Growing Demand for Computing



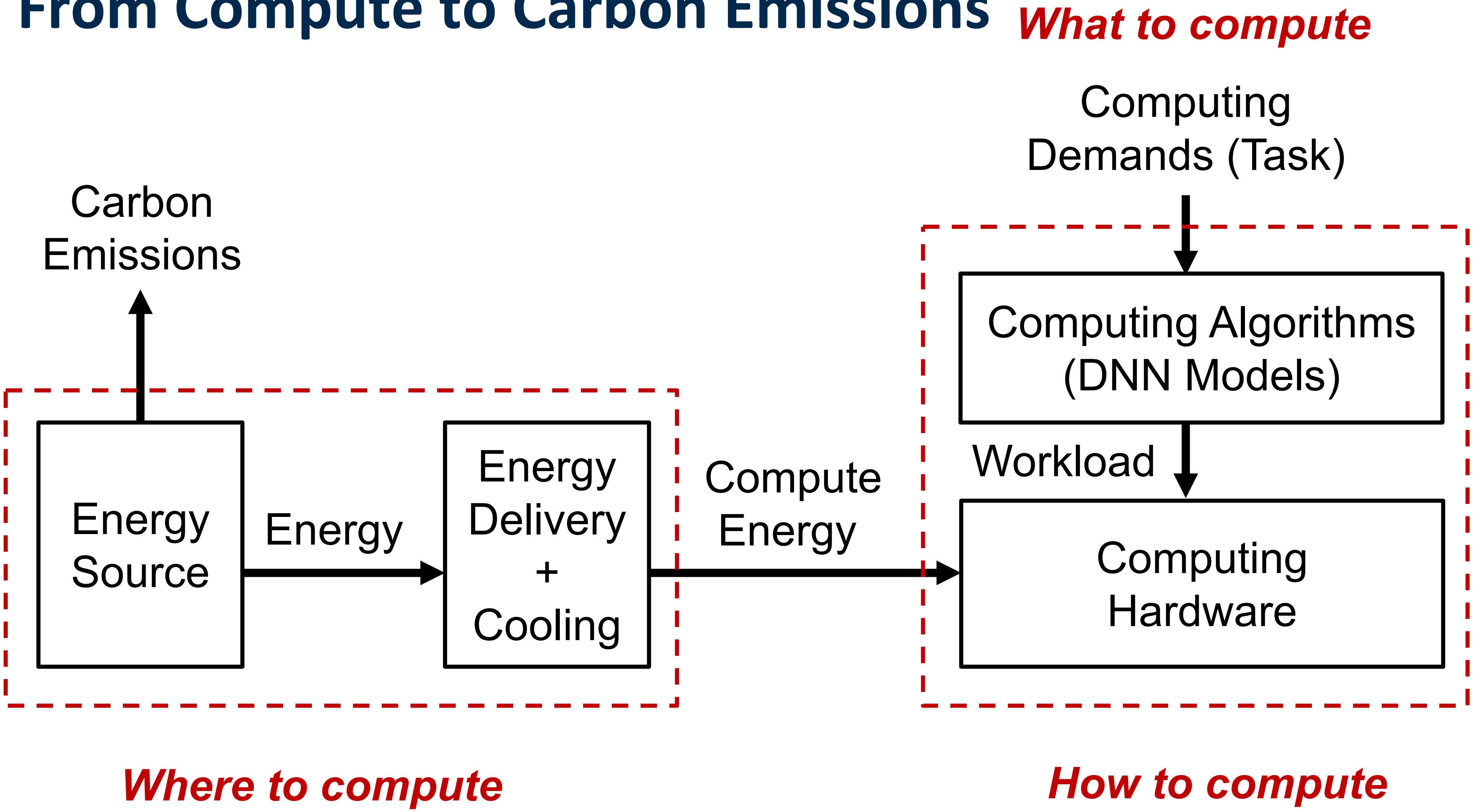
AlexNet to AlphaGo Zero: A 300,000x Increase in Compute



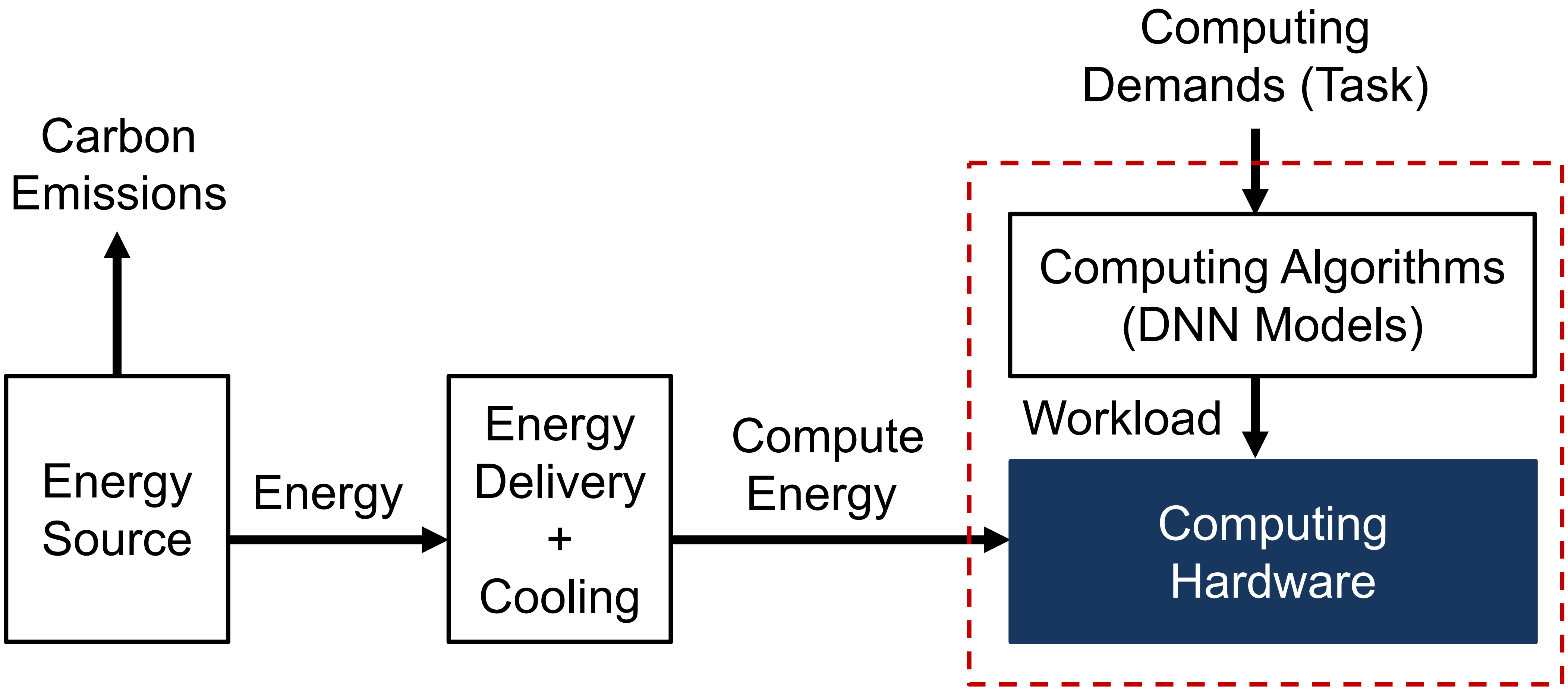
Source: Nature (https://www.nature.com/articles/d41586-018-06610-y)

Source: Open AI (https://openai.com/blog/ai-and-compute/)

From Compute to Carbon Emissions

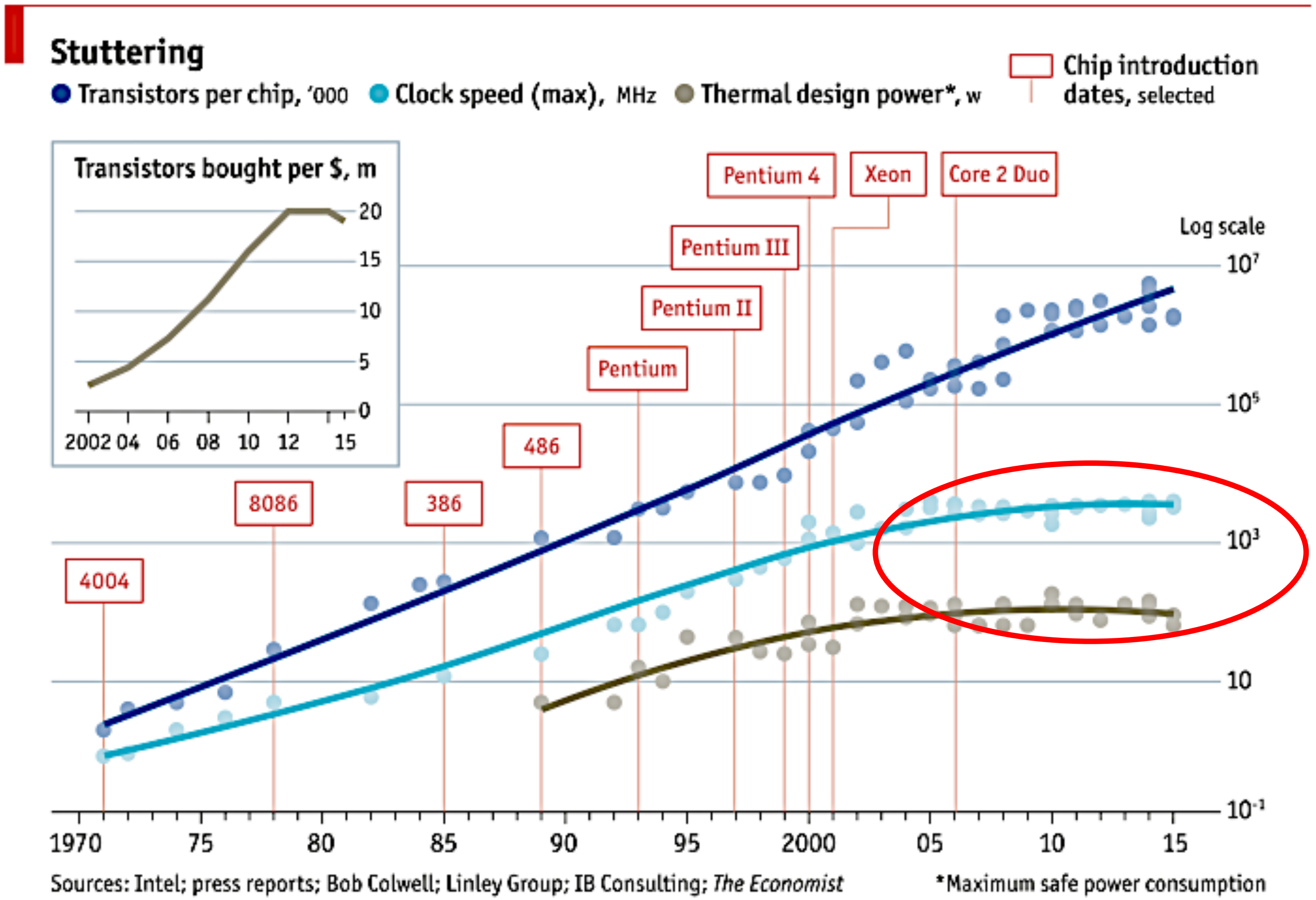


From Compute to Carbon Emissions



How to compute

Transistors Are Not Getting More Efficient



Slowdown

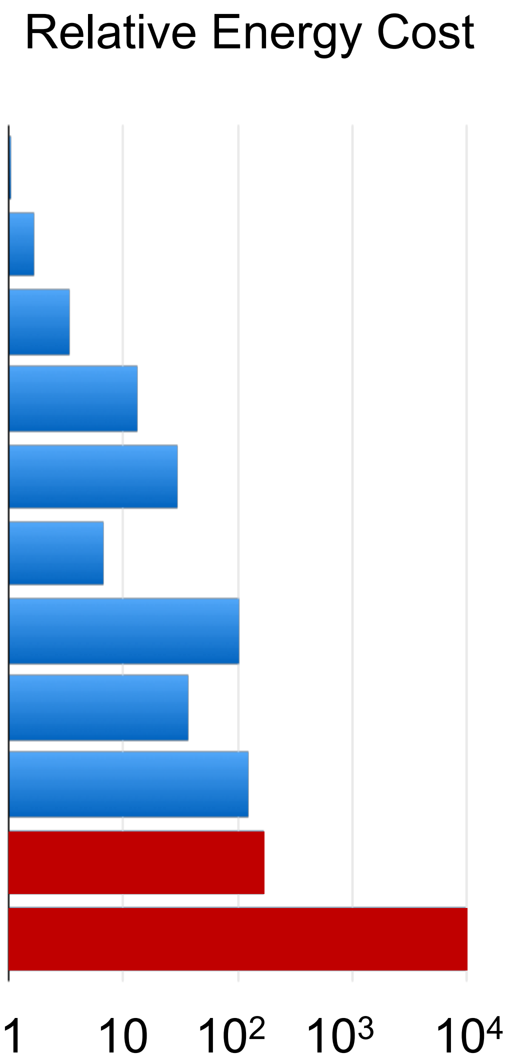
Slowdown of Moore's Law and Dennard Scaling

General purpose microprocessors are not getting faster or more efficient

Need **specialized / domain-specific hardware** for significant improvements in speed and energy efficiency

Energy Consumption Dominated by Data Movement

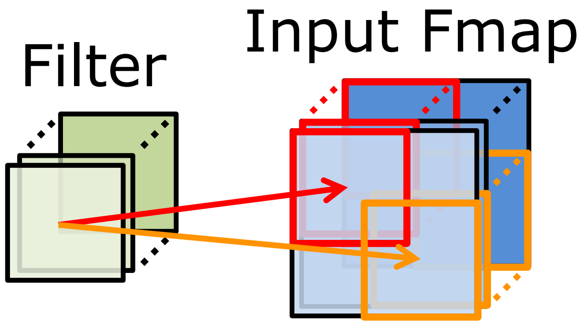
Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9
8b Multiply	0.2
32b Multiply	3.1
16b FP Multiply	1.1
32b FP Multiply	3.7
32b SRAM Read (8KB)	5
32b DRAM Read	640



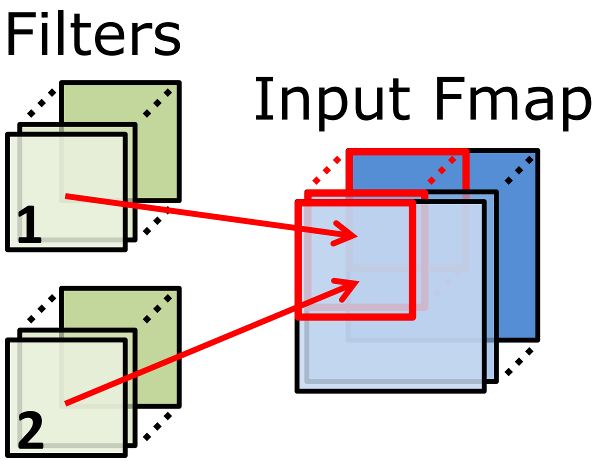
Memory access is **orders of magnitude** higher energy than compute

[**Horowitz**, *ISSCC* 2014]

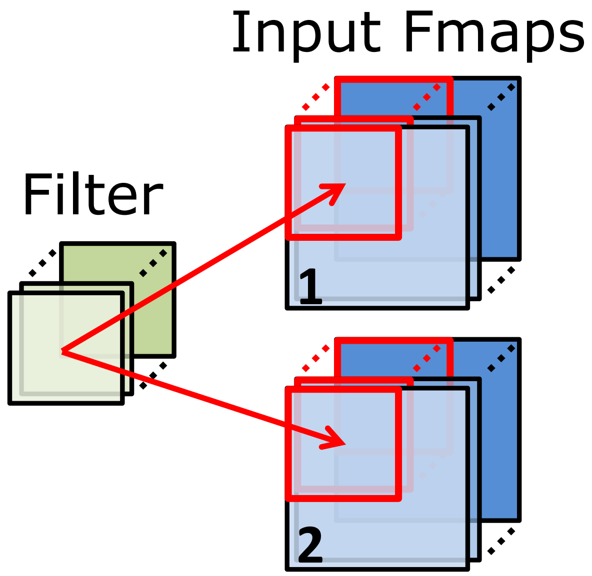
Exploit Data Reuse Opportunities in DNNs



Convolutional Reuse
(Activations, Weights)
CONV layers only
(sliding window)

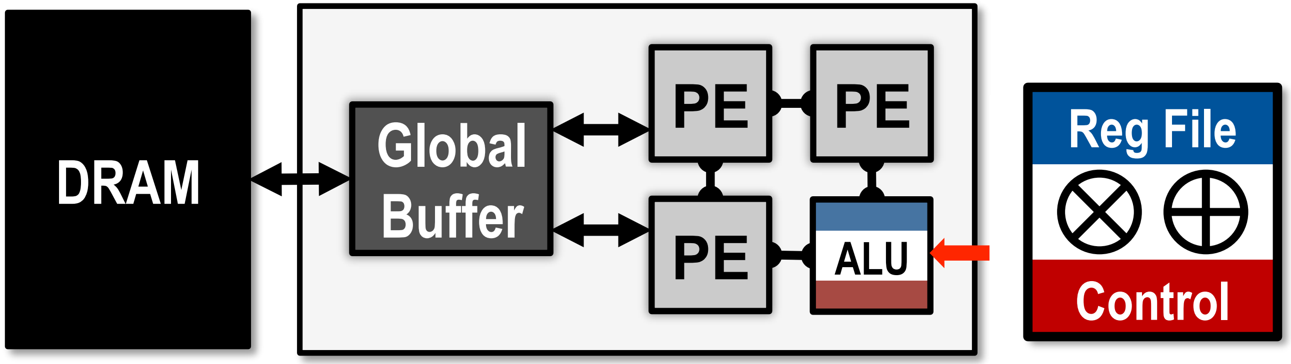


Fmap Reuse
(Activations)
CONV and FC layers

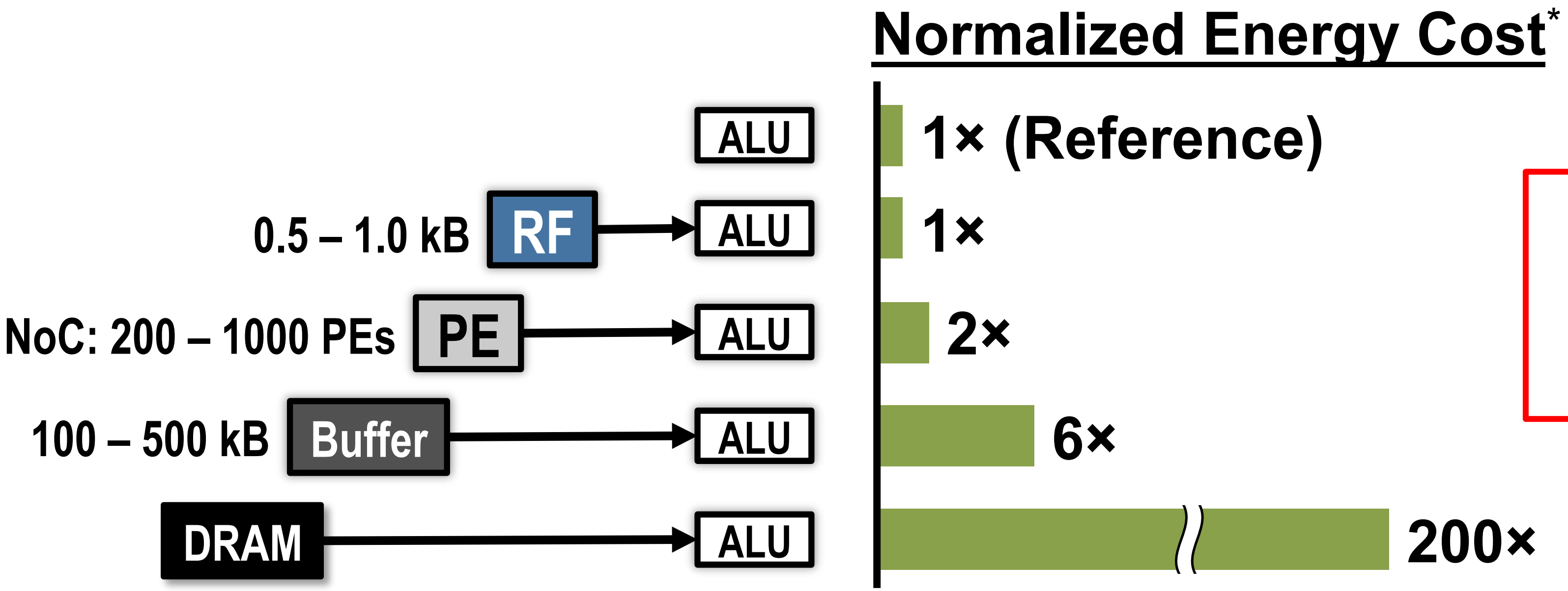


Filter Reuse
(Weights)
CONV and FC layers
(batch size > 1)

Exploit Data Reuse at Low-Cost Memories



Specialized hardware with small (< 1kB) low cost memory near compute

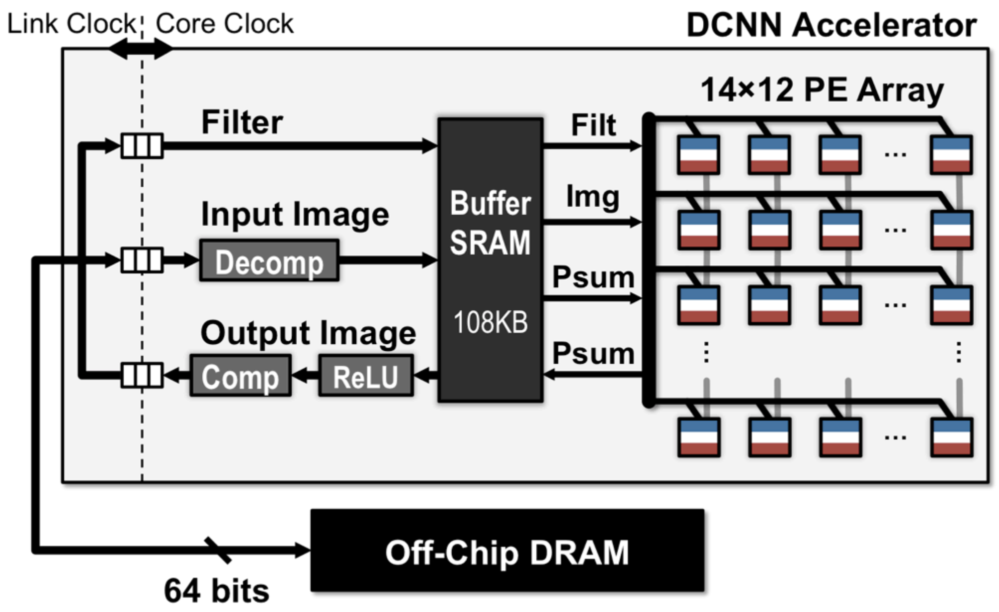


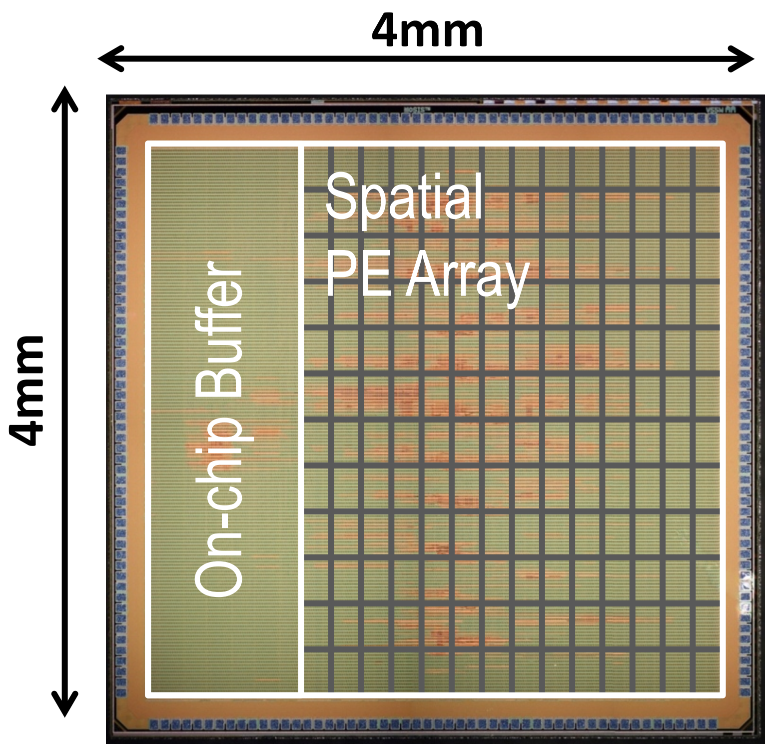
Farther and **larger** memories consume more power

* measured from a commercial 65nm process

Energy-Efficient Dataflow

Eyeriss





Eyeriss Project Website: http://eyeriss.mit.edu

[**Chen**, *ISSCC* 2016],[**Chen**, *ISCA* 2016] **Micro Top Picks**

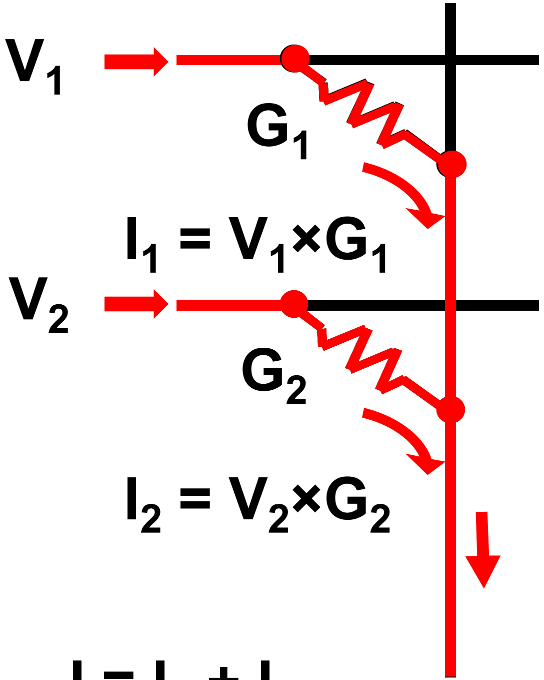
Exploits data reuse for **100x** reduction in memory accesses from global buffer and **1400x** reduction in memory accesses from off-chip DRAM

Overall **>10x energy reduction** compared to a mobile GPU

[Joint work with Joel Emer]

In-Memory Computing

Activation is input voltage (V_i)
Weight is resistor conductance (G_i)



Psum is output current

$I = I_1 + I_2$
$= V_1 \times G_1 + V_2 \times G_2$

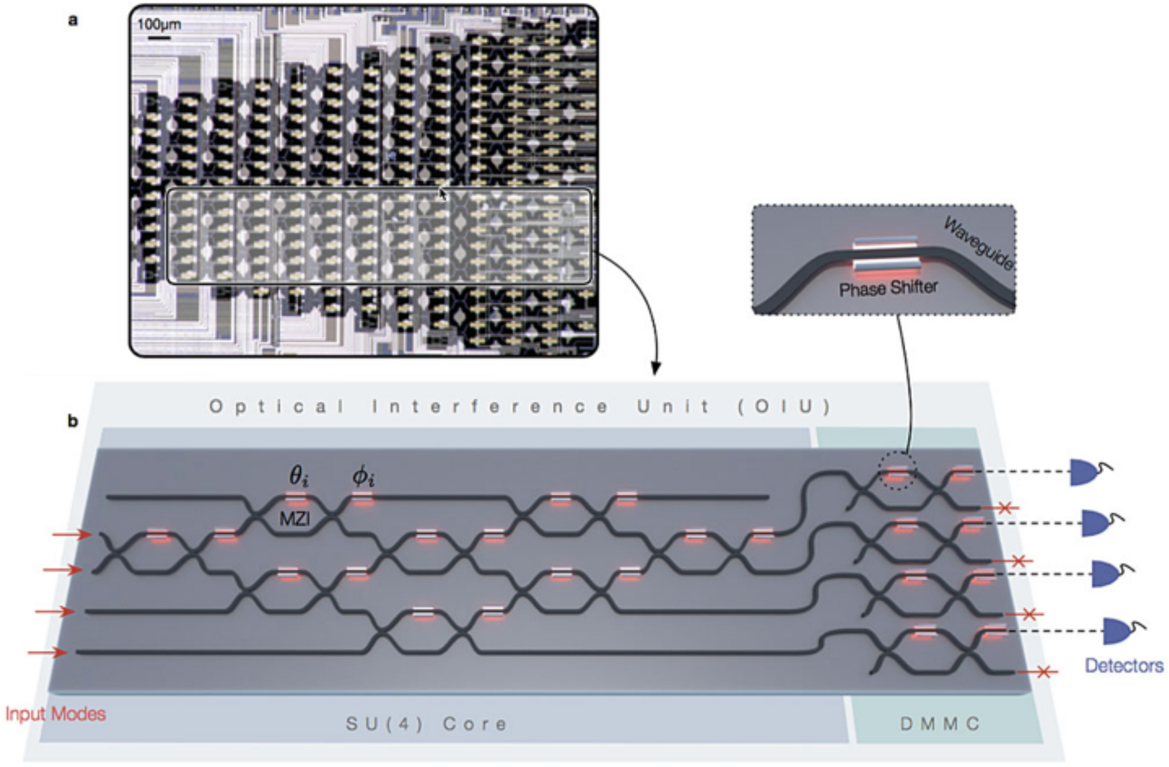
Image Source: [**Shafiee**, *ISCA* 2016]

- Reduce data movement by **moving compute into memory**
- Compute with memory storage elements
- **Analog Compute**
  - Activations, weights and/or partial sums are encoded with analog voltage, current, or resistance
  - Increased sensitivity to circuit non-idealities
  - A/D and D/A circuits to interface with digital domain
- Leverage **emerging memory device technology**

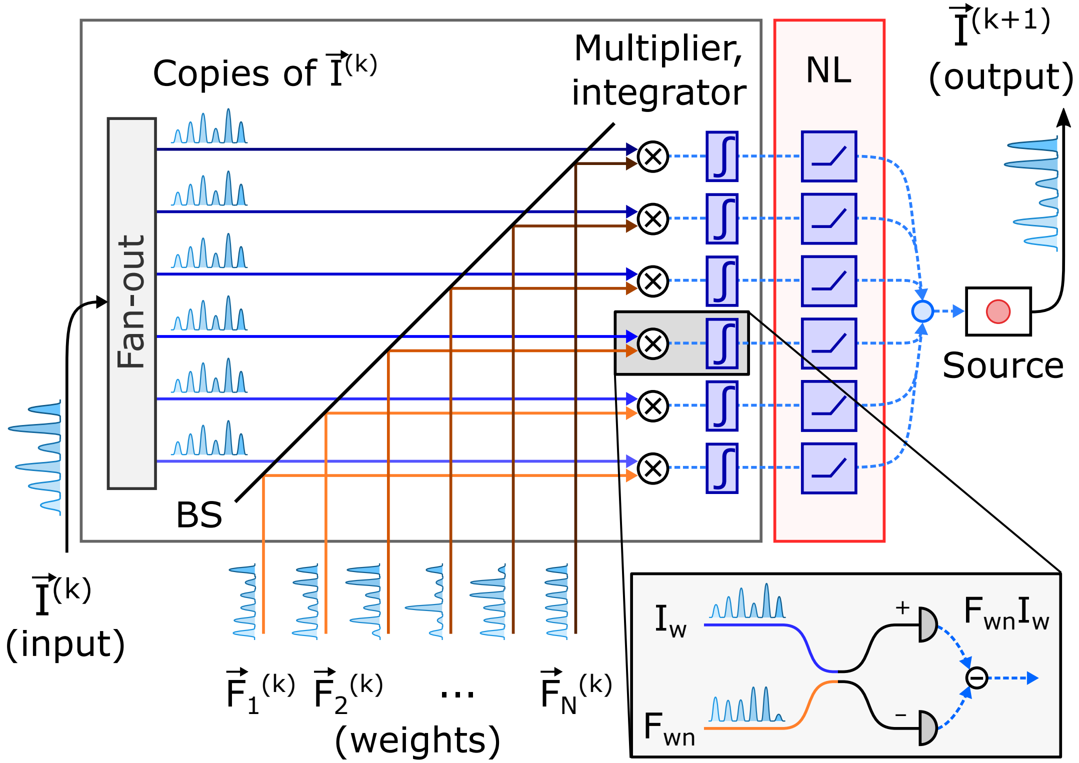
eNVM:[**Yu**, *PIEEE* 2018], SRAM:[**Verma**, *SSCS* 2019]

Computing With Light

- Cost of moving a photon can be **independent** of distance
- Multiplication can be performed **passively**

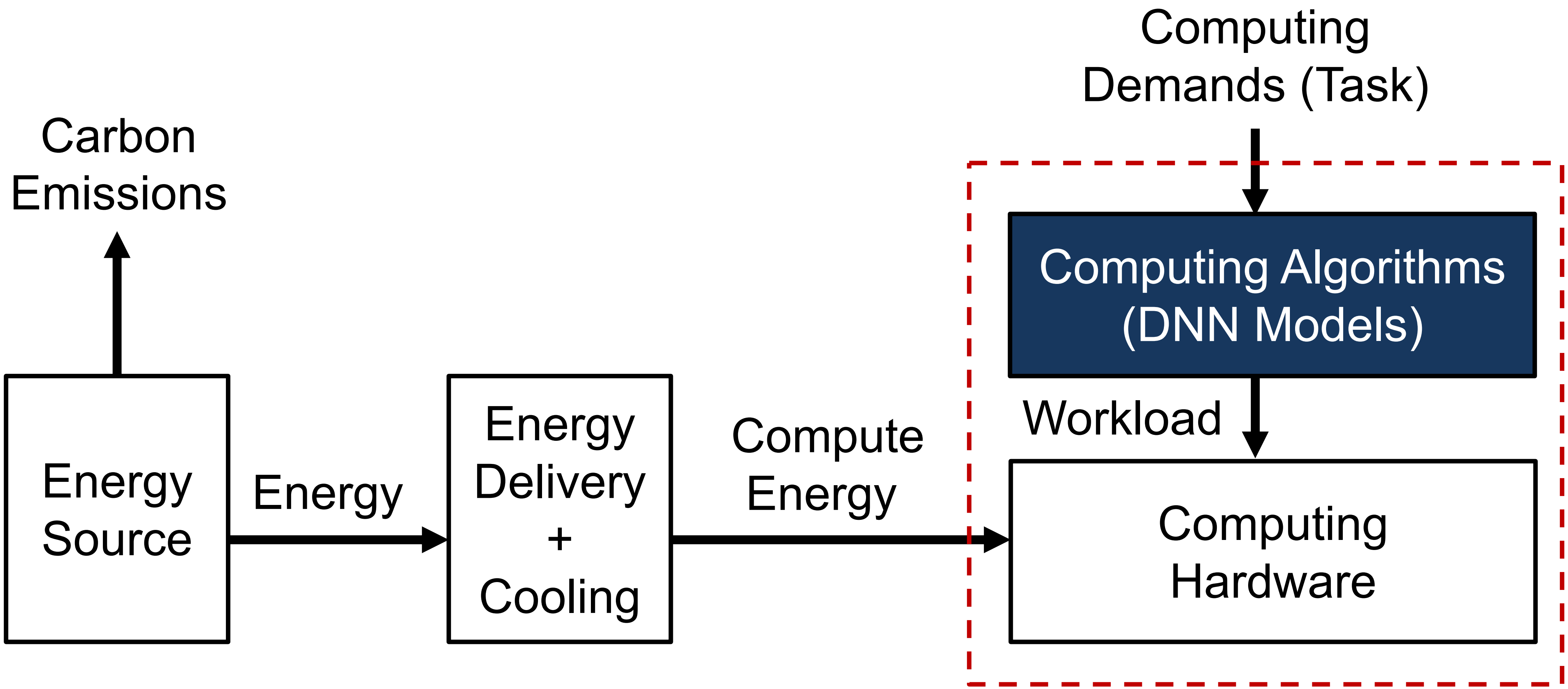


[**Shen**, *Nature Photonics* 2017]



[**Bernstein**, *CLEO* 2020]

From Compute to Carbon Emissions

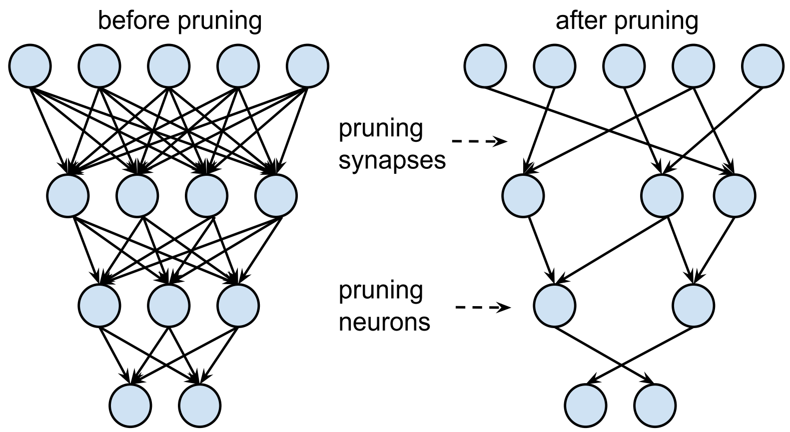


How to compute

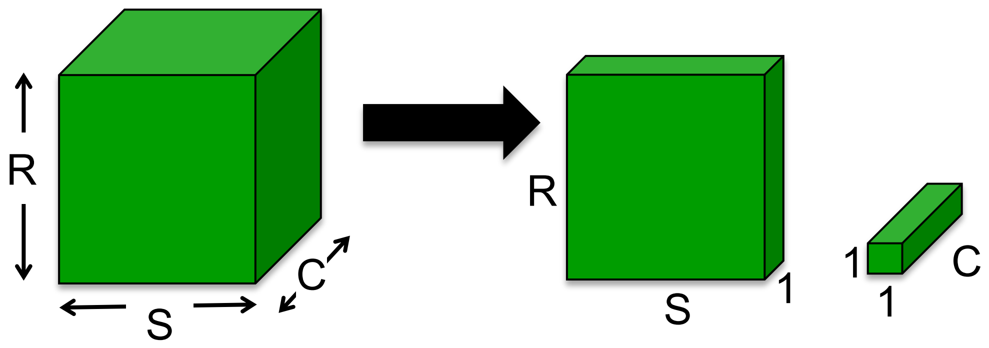
Design of Efficient DNN Algorithms

Popular efficient DNN algorithm approaches

Network Pruning



Efficient Network Architectures



Examples: SqueezeNet, MobileNet

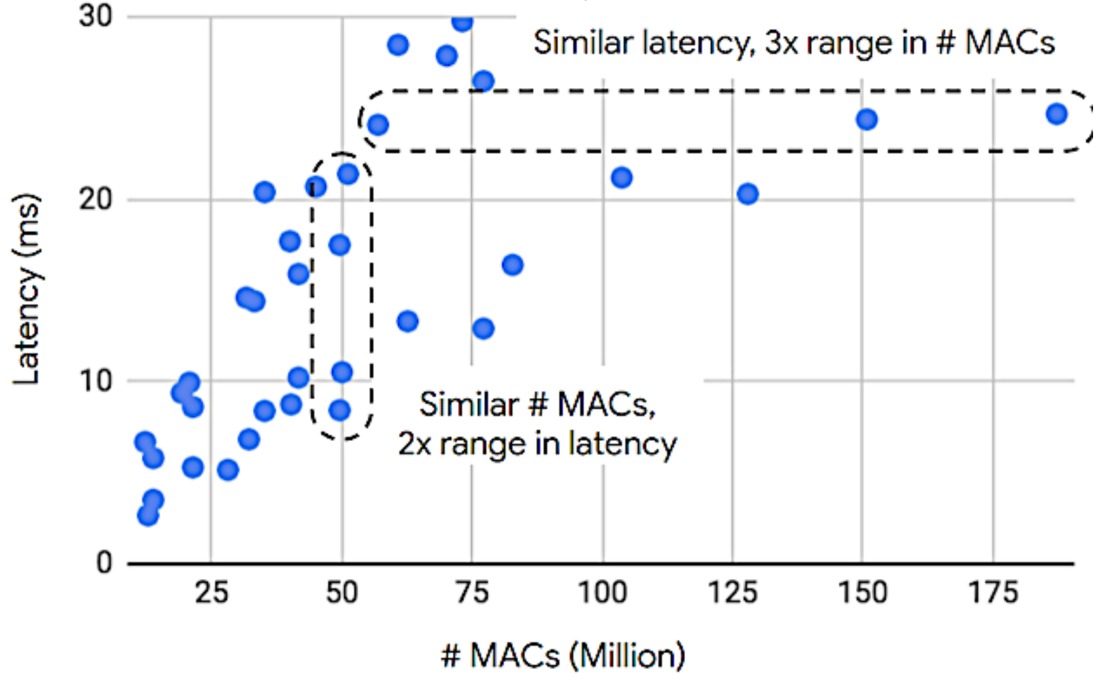
... also reduced precision

- Focus on reducing **number of MACs and weights**
- **Does it translate to energy savings and reduced latency?**

[**Chen*, Yang***, *SysML* 2018]

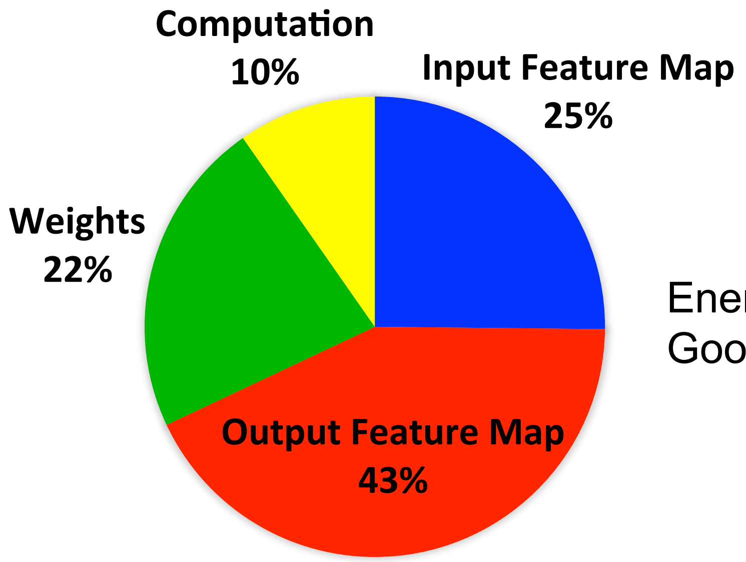
Number of MACs and Weights are Not Good Proxies

of operations (MACs) does not approximate latency well



Source: Google (https://ai.googleblog.com/2018/04/introducing-cvpr-2018-on-device-visual.html)

# of weights ***alone*** is not a good metric for energy (**All data types** should be considered)



Energy breakdown of GoogLeNet

https://energyestimation.mit.edu/

[**Yang**, *CVPR* 2017]

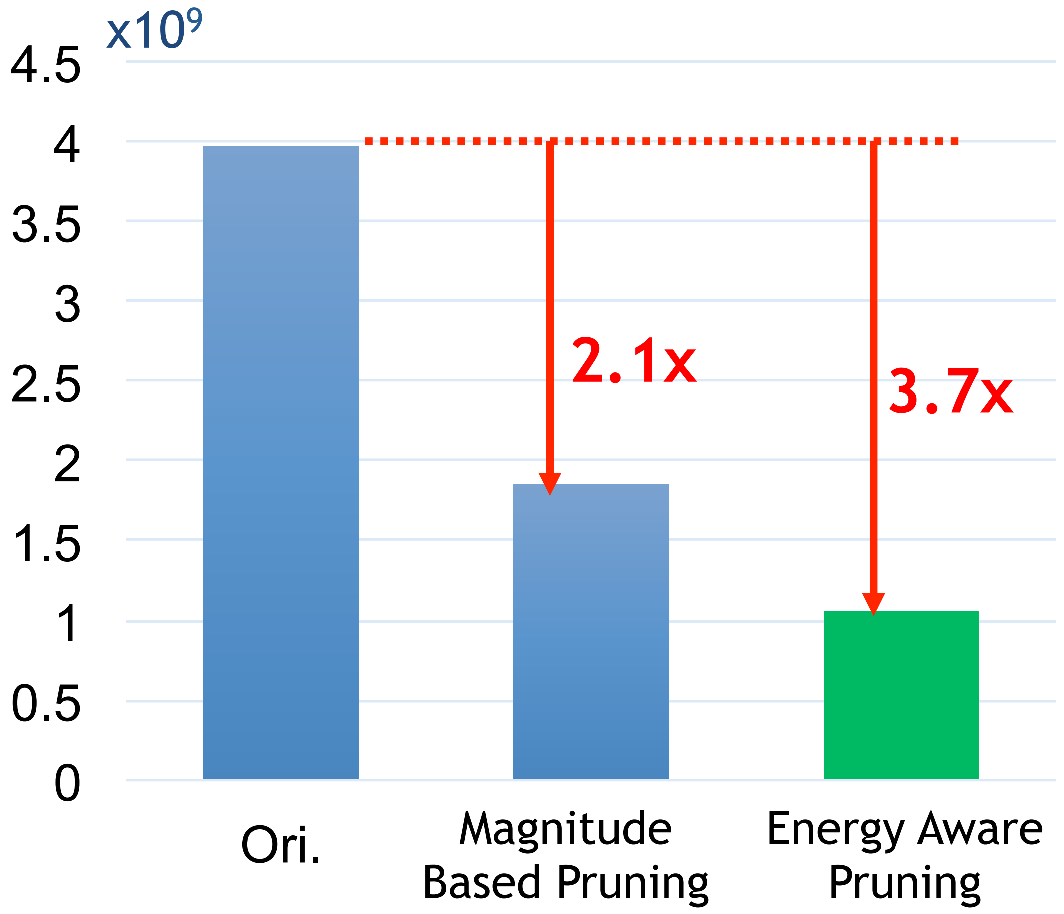
Energy-Aware Pruning

Directly target energy
and incorporate it into the optimization of DNNs to provide greater energy savings

- Sort layers based on energy and prune layers that consume the most energy first
- **Energy-aware pruning** reduces AlexNet energy by **3.7x** w/ similar accuracy
- Outperforms magnitude-based pruning by **1.7x**

[**Yang**, *CVPR* 2017]

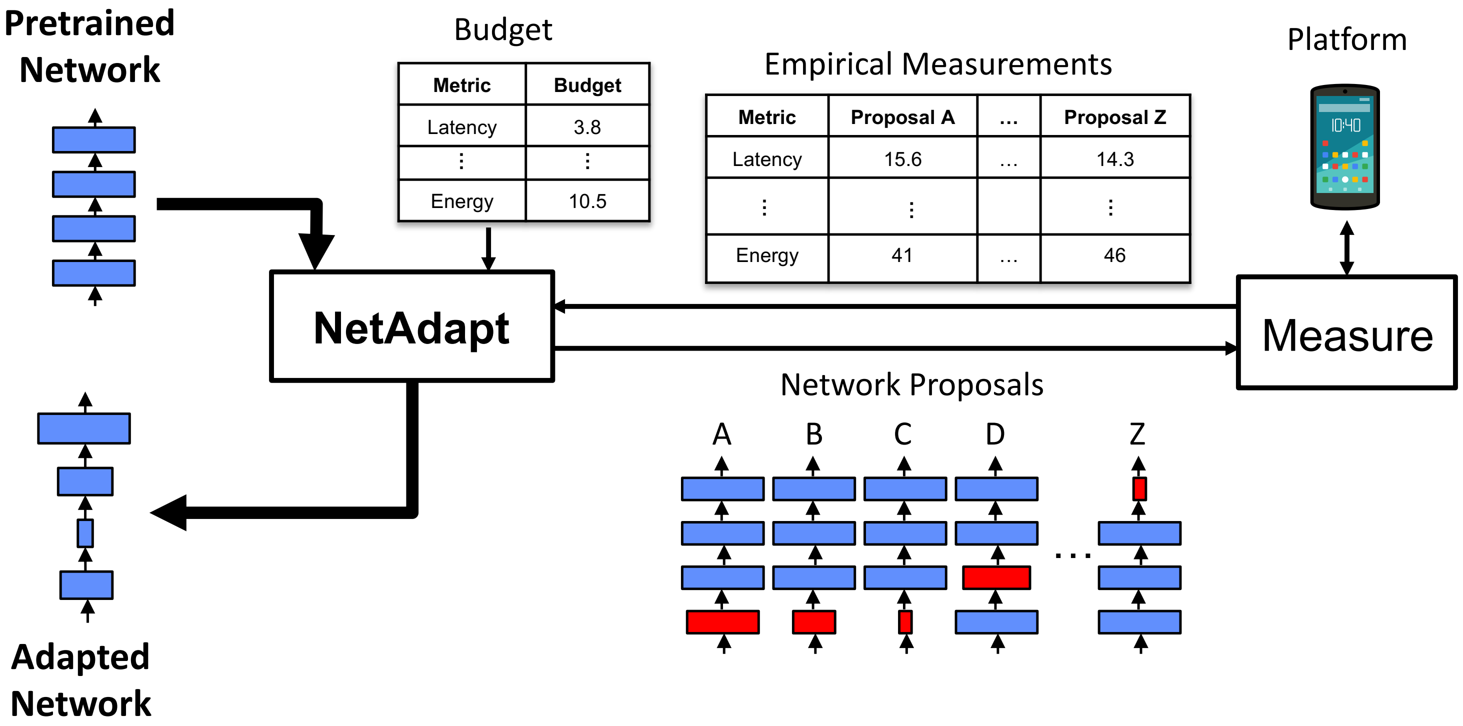
Normalized Energy (AlexNet)



Pruned models available at
http://eyeriss.mit.edu/energy.html

NetAdapt: Platform-Aware DNN Adaptation

- **Automatically adapt DNN** to a mobile platform to reach a target latency or energy budget
- Use **empirical measurements** to guide optimization (avoid modeling of tool chain or platform architecture)
- **>1.7x speed up** on MobileNet w/ similar accuracy
- **Few hyperparameters** to reduce tuning effort



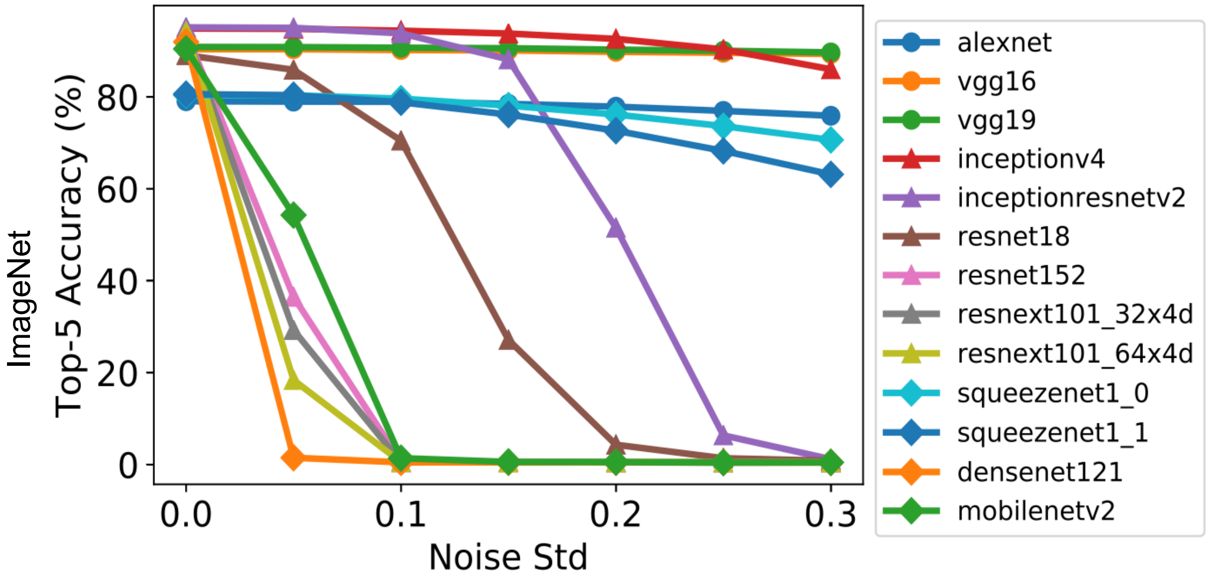
[**Yang**, *ECCV* 2018]

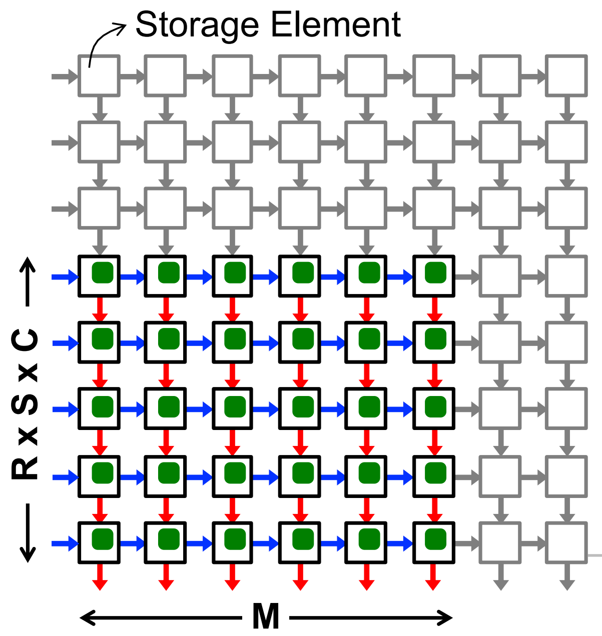
Code available at
http://netadapt.mit.edu

[In collaboration with Google's Mobile Vision Team]

Designing DNNs for In-Memory Computing (IMC)

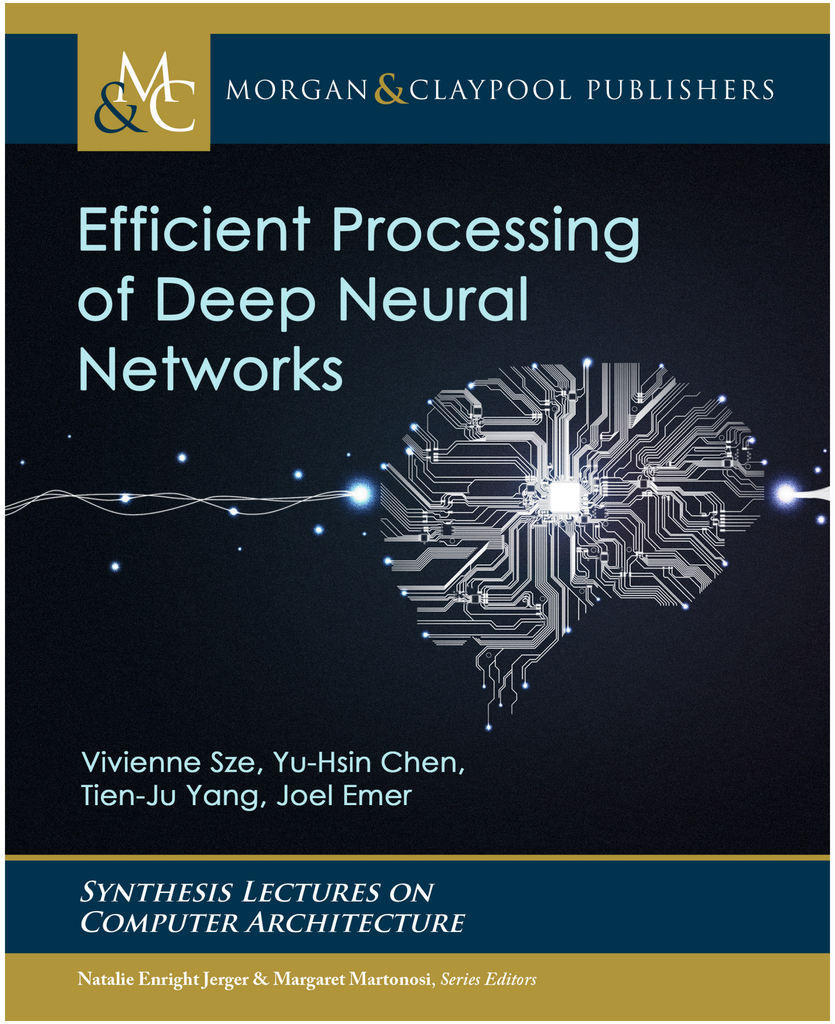
- Designing DNNs for IMC may differ from DNNs for digital processors
- Highest accuracy DNN on digital processor may be different on IMC
  - Accuracy drops based on robustness to non-idealities
- Reducing number of weights is less desirable
  - Since IMC is weight stationary, may be better to reduce number of activations
  - IMC tend to have larger arrays → fewer weights may lead to low utilization on IMC
- For IMC, may be preferable to do shallower and larger filters
  - Differs from current trend of deeper and smaller filters





[**Yang**, *IEDM* 2019]

Book on "How to Compute" Efficiently



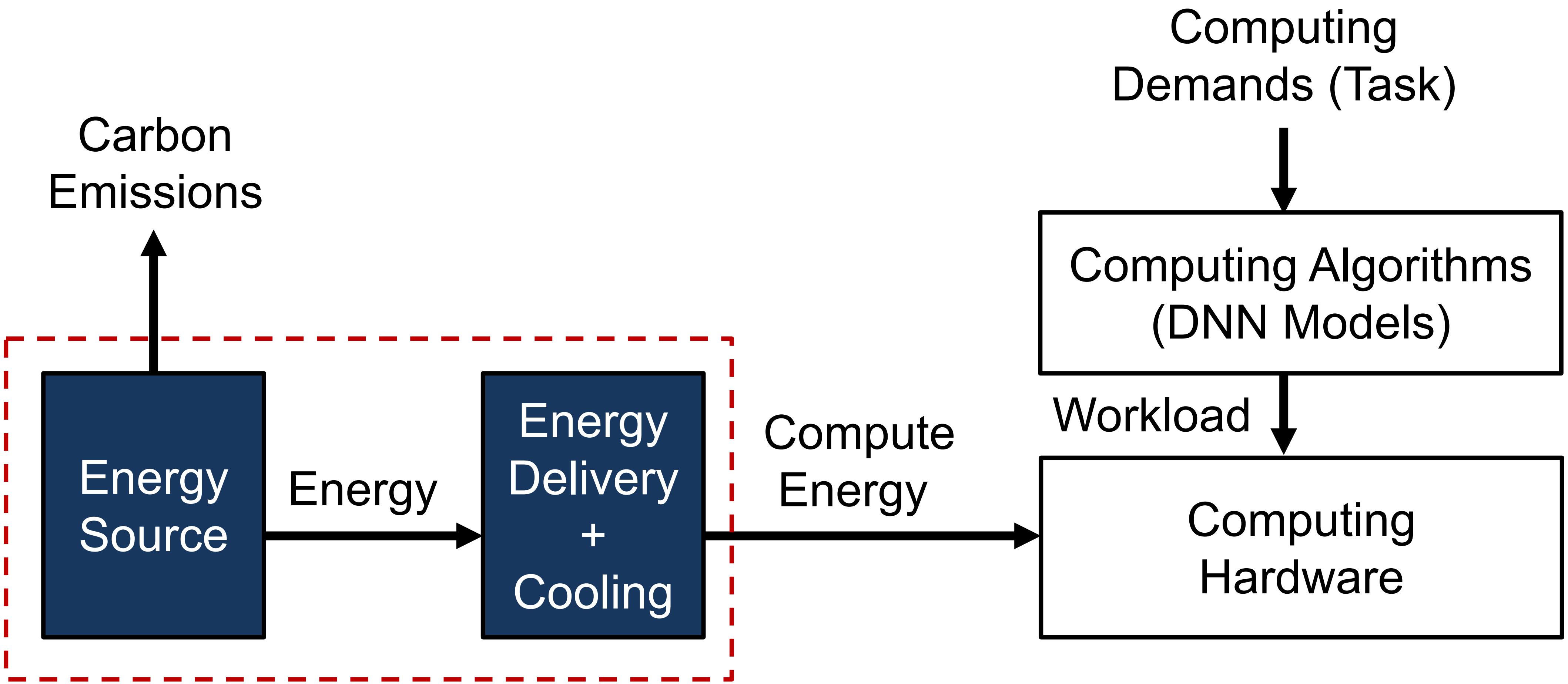
Part I Understanding Deep Neural Networks
Introduction
Overview of Deep Neural Networks

Part II Design of Hardware for Processing DNNs
Key Metrics and Design Objectives
Kernel Computation
Designing DNN Accelerators
Operation Mapping on Specialized Hardware

Part III Co-Design of DNN Hardware and Algorithms
Reducing Precision
Exploiting Sparsity
Designing Efficient DNN Models
Advanced Technologies

https://tinyurl.com/EfficientDNNBook

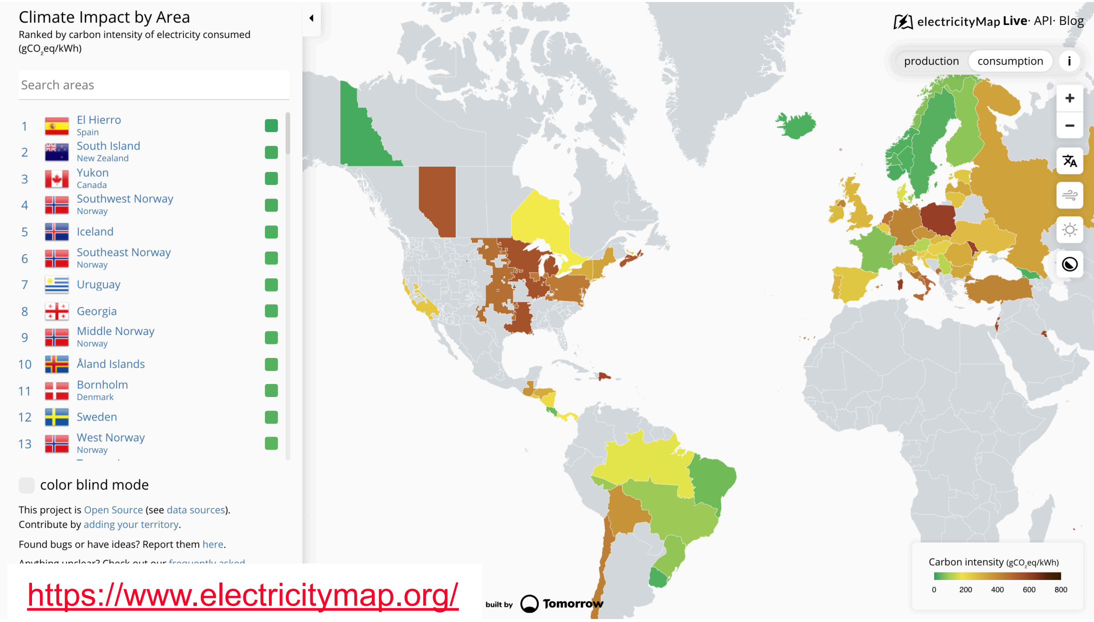
From Compute to Carbon Emissions



Where to compute

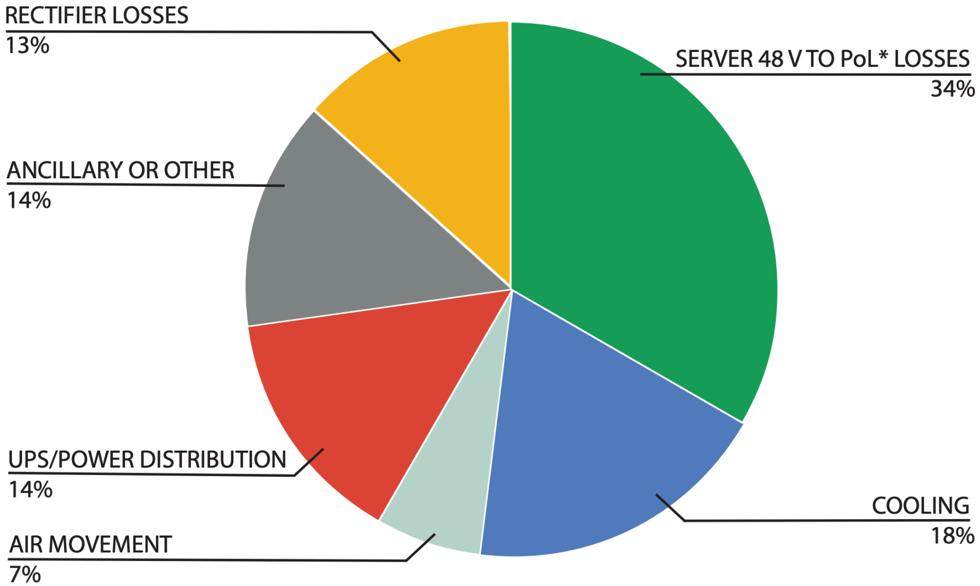
Where to Compute?

- **Energy Source** *(Carbon Emissions → Energy)*
  - Carbon Intensity (gCO_{2eq} / kWh) of Energy Source
    - Varies by region
  - Percentage of Renewable Energy
    - Varies with time of day



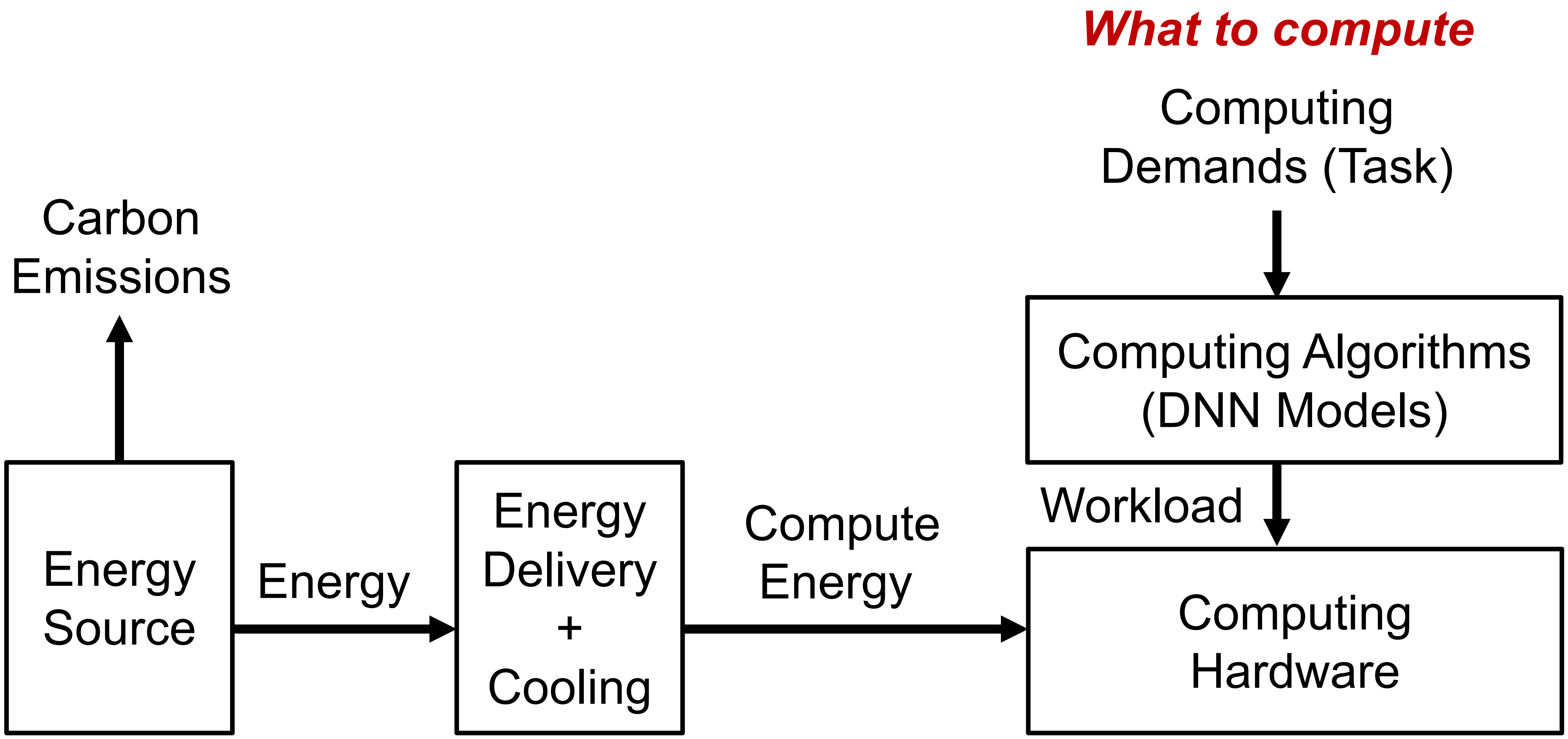
https://www.electricitymap.org/

- **Energy Delivery** *(Energy → Compute Energy)*
  - Power Conversion & Cooling Cost
  - Example: Data Centers
    - Power Usage Effectiveness (PUE) = Energy/Compute Energy
    - Typically in the range of 2.0 to 1.1 (1.0 is optimal)
  - ***Use ML to improve efficiency***



[**Barroso**, *Synthesis Lecture in Computer Architecture*, 2018]

From Compute to Carbon Emissions



What to Compute?

- Compute demands depend on number of requests, amount of data, and required quality of result (e.g., accuracy)
- Reduce number of requests
  - Make hyper-parameter tuning easier (e.g., reduce the number of hyper-parameters to tune)
  - Reproducibility is critical for reducing **unnecessary** requests due to replication difficulties → ***also good for advancing research in ML***
    - On-going efforts in ML (Reproducibility Challenges) and Systems (Artifact Evaluation Badges)
- Reduce amount of needed data
  - Exploit data reuse since data movement is expensive
  - Explore data-efficient ML techniques & ML models that incorporate prior knowledge
- Evaluate carbon emissions versus quality of result tradeoffs
  - Cost-benefit evaluation (e.g., Is the accuracy improvement worth the carbon emissions?)
  - Deeper consideration of quality of result for a given task

[**Henderson**, *arXiv preprint arXiv:2002.05651*, 2020]

Recommended Best Practices

- Make energy-efficient settings the default setting – or easy to set
  - Software and framework support for reduced precision and specialized hardware
- Measure and report energy consumption and carbon emissions
  - Software and hardware support for measuring energy consumption
  - System support for reporting carbon intensity of energy source
  - Frameworks for standardized reporting [**Henderson**, *arXiv preprint arXiv:2002.05651,* 2020]
    - https://github.com/Breakend/experiment-impact-tracker
- Run experiments in locations / at times with low carbon intensity
- Ensure reproducibility to avoid unnecessary experiments
- **We can do much of this today (or in the very near future)!**

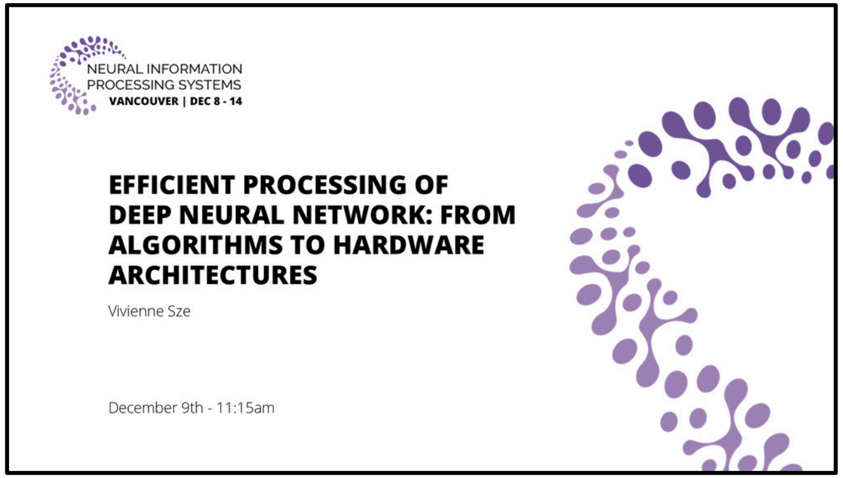
Key Takeaways

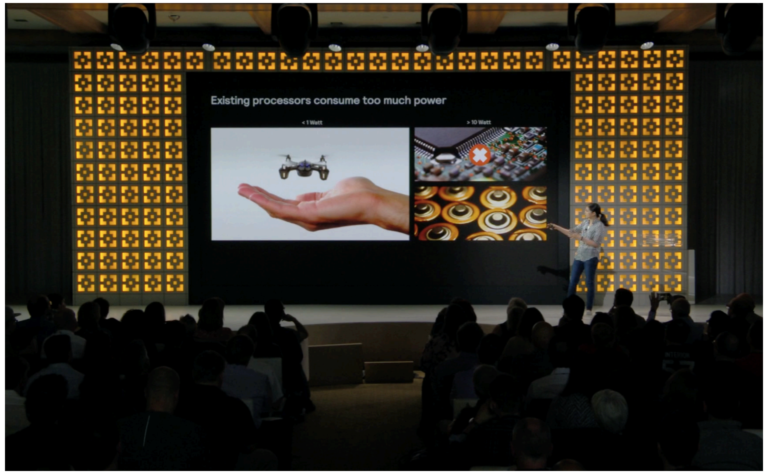
- **Jointly consider energy efficiency and accuracy in ML research**
  - Consider the accuracy-energy tradeoffs and data-efficient ML techniques
  - Design ML algorithms that directly target energy consumption
  - Design specialized ML hardware to reduce data movement
- **Incorporate energy efficiency considerations into best practices**
  - Lower the barrier to using existing energy-efficient computing options and reporting/measuring energy consumption and carbon emissions
  - Compute at locations with lowest carbon intensity and highest power efficiency
  - Reduce unnecessary computing demands by ensuring reproducibility

Additional Resources

Talks and Tutorial Available Online

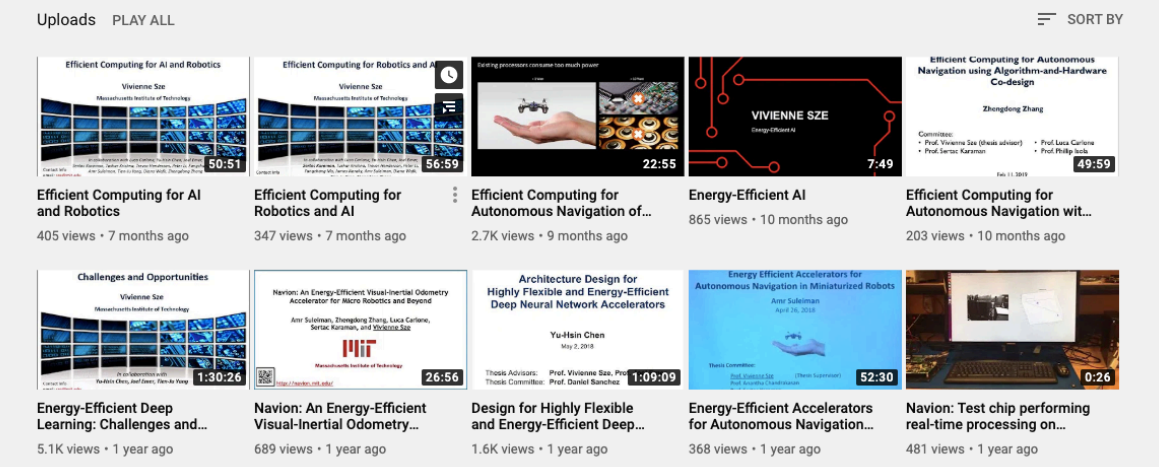
https://www.rle.mit.edu/eems/publications/tutorials/



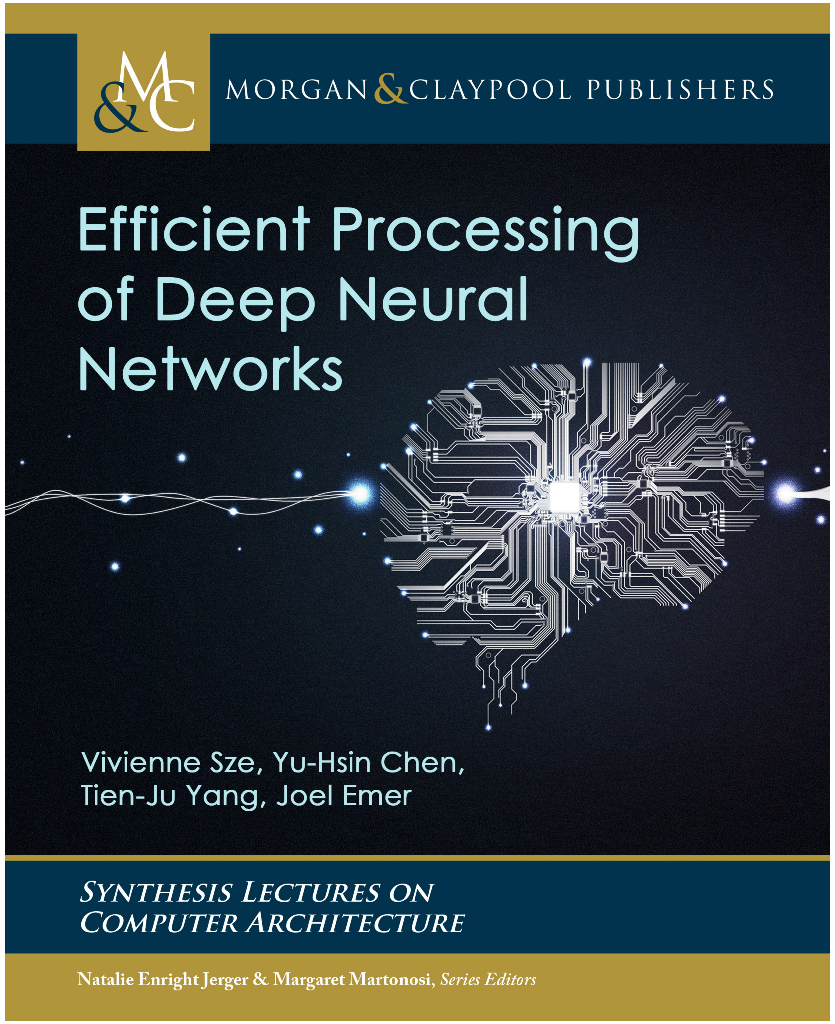


YouTube Channel

EEMS Group – PI: Vivienne Sze



Book on Efficient Processing of DNNs



Part I Understanding Deep Neural Networks
Introduction
Overview of Deep Neural Networks

Part II Design of Hardware for Processing DNNs
Key Metrics and Design Objectives
Kernel Computation
Designing DNN Accelerators
Operation Mapping on Specialized Hardware

Part III Co-Design of DNN Hardware and Algorithms
Reducing Precision
Exploiting Sparsity
Designing Efficient DNN Models
Advanced Technologies

https://tinyurl.com/EfficientDNNBook

References

- **Computing Hardware**

  - Y.-H. Chen, T. Krishna, J. Emer, V. Sze, “Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks,” IEEE Journal of Solid-State Circuits (JSSC), ISSCC Special Issue, Vol. 52, No. 1, pp. 127-138, January 2017. http://eyeriss.mit.edu
  - Y.-H. Chen, J. Emer, V. Sze, “Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks,” International Symposium on Computer Architecture (ISCA), pp. 367-379, June 2016.
  - Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and Marin Soljačić, “Deep learning with coherent nanophotonic circuits," Nature Photonics, 2017
  - L. Bernstein, A. Sludds, R. Hamerly, V. Sze, J. Emer, and D. Englund, Digital optical neural networks for large-scale machine learning, Conference on Lasers and Electro-Optics (CLEO), 2020

References

- **Computing Algorithm**

  - Y.-H. Chen*, T.-J. Yang*, J. Emer, V. Sze, “Understanding the Limitations of Existing Energy-Efficient Design Approaches for Deep Neural Networks,” SysML Conference, February 2018.
  - V. Sze, Y.-H. Chen, T.-J. Yang, J. Emer, “Efficient Processing of Deep Neural Networks: A Tutorial and Survey,” Proceedings of the IEEE, vol. 105, no. 12, pp. 2295-2329, December 2017. http://eyeriss.mit.edu/tutorial.html
  - T.-J. Yang, Y.-H. Chen, V. Sze, “Designing Energy-Efficient Convolutional Neural Networks using Energy-Aware Pruning,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
  - T.-J. Yang, A. Howard, B. Chen, X. Zhang, A. Go, V. Sze, H. Adam, “NetAdapt: Platform-Aware Neural Network Adaptation for Mobile Applications,” European Conference on Computer Vision (ECCV), 2018. http://netadapt.mit.edu/
  - T.-J. Yang, V. Sze, “Design Considerations for Efficient Deep Neural Networks on Processing-in-Memory Accelerators,” IEEE International Electron Devices Meeting (IEDM), Invited Paper, December 2019.
  - Sze, V., Chen, Y. H., Yang, T. J., & Emer, J. S. (2020). Efficient Processing of Deep Neural Networks. *Synthesis Lectures on Computer Architecture*, *15*(2), 1-341. https://tinyurl.com/EfficientDNNBook

References

- **Where & What to Compute**

  - Barroso, L. A., Hölzle, U., & Ranganathan, P. (2018). The datacenter as a computer: Designing warehouse-scale machines. Synthesis Lectures on Computer Architecture, 13(3), i-189.
  - Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D. and Pineau, J., 2020. Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning. arXiv preprint arXiv:2002.05651.