A Decade of Machine Learning Accelerators: Lessons Learned and Carbon Footprint

David Patterson, Google and UC Berkeley

Based on the following papers and talks:

The Design Process for Google's Training Chips: TPUv2 and TPUv3
IEEE Micro, March 2021
Thomas Norrie, Nishant Patil, Doe Hyun Yoon, George Kurian, Sheng Li, James Laudon, Cliff Young, Norman Jouppi, & David Patterson

Ten Lessons From Three Generations Shaped Google's TPUv4i
International Symposium on Computer Architecture, June 2021
Norman P. Jouppi, Doe Hyun Yoon, Matthew Ashcraft, Mark Gottscho, Thomas B. Jablin, George Kurian, James Laudon, Sheng Li, Peter Ma, Xiaoyu Ma, Thomas Norrie, Nishant Patil, Sushma Prasad, Cliff Young, Zongwei Zhou, & David Patterson

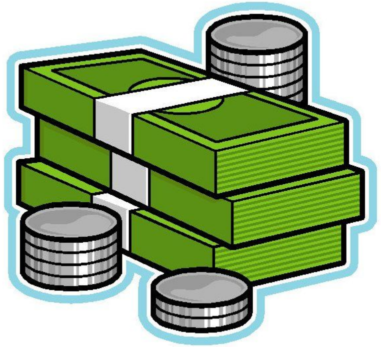
The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink
IEEE Computer, July 2022
David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, & Jeff Dean

Outline

- Introduction to the TPU Family (~5 minutes)
- Lessons learned and how they shaped TPUs (~15 minutes)
    - Preview: DNNs grow and evolve quickly, FLOPs are easy but memory is hard
- Dire projections of carbon emissions of ML (~5 minutes)
    - Preview: Some papers overestimate Google ML emissions by 1,000,000x
- “4Ms” of energy efficiency: Model, Machine, Mechanization, Map (~10 minutes)
    - Preview: Optimizing 4Ms can reduce energy consumption 10x, emissions 100x
    - Preview: ML is ~75% of Google’s FLOPs but < 15% of total energy
- Conclusion and Recommendations (~5 minutes)
- Acknowledgements
- Q&A
- (if time) Lessons from My Career

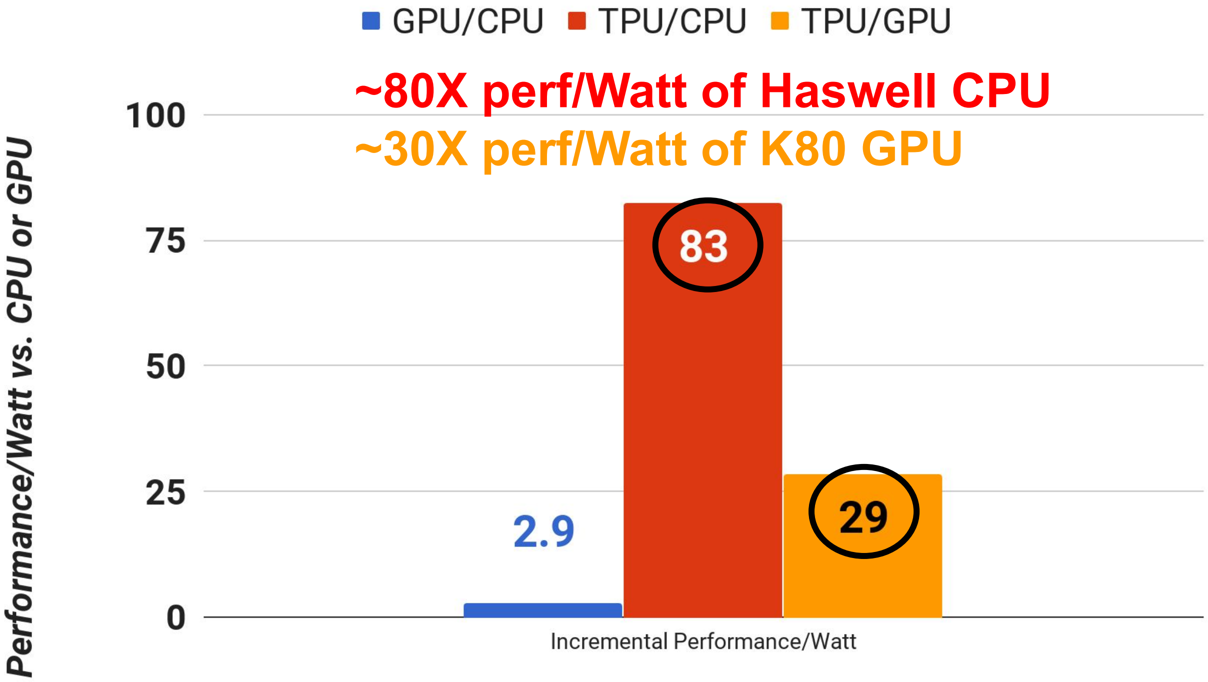
Introduction to the TPU Family

TPU Origin Story



- 2013: Prepare for success-disaster of new DNN apps
  - Scenario with 100M users speaking to phones 3 minutes per day: If only CPUs, double whole data center fleet!
- Goal: Custom *Domain Specific Architecture (DSA)* to reduce the Total Cost of Ownership (TCO) of DNN inference phase by 10X
  - Training "learns" parameters; Inference uses the trained model in production
  - Must run existing apps developed for CPUs and GPUs
- Very short development cycle
  - Started TPUv1 project 2014
  - Running in datacenter 15 months later: architecture invention, compiler invention, hardware design, build, test, and deploy

TPU v1 vs CPU & GPU: Performance/Watt



Jouppi, Norman P., Cliff Young, Nishant Patil, David Patterson, et al. In-datacenter performance analysis of a tensor processing unit, *ISCA*, 2017.

May 18, 2016 Google Announcement

"We've been running TPUs inside our data centers for more than a year, and have found them to deliver an ***order of magnitude better-optimized performance per watt for ML****."*

Google CEO Sundar Pichai

cloudplatform.googleblog.com/2016/05/Google-supercharges-machine-learning-tasks-with-custom-chip.html

See timecode 1:48:31 in the Google I/O keynote video (May 18, 2016): https://www.youtube.com/watch?v=862r3XS2YB0

Google

The Launching of “1000 Chips”

- Intel acquires DSA chip companies
  - Nervana: ($0.4B) August 2016
  - Movidius: ($0.4B) September 2016
  - MobilEye: ($15.3B) March 2017
  - Habana: ($2.0B) December 2019
- Alibaba, Amazon build inference chips
- >100 startups ($3B/yr) launch own bets
  - Coarse-Grained Reconfigurable Arch: SambaNova, ...
  - Analog computing: Mythic, …
  - Full silicon wafer computer: Cerebras, …
- Academia: TPUv1 paper ~5000 citations
- Most influential since RISC, Pentium Pro?

Helen of Troy
by Evelyn De Morgan

Google

TPU Generations

Year	Inference	Training & Inference	Peak Chip Performance*	TDP	Tech. Node	Chips/ Pod	Peak Pod Performance*
2015	TPU v1		92 TOPs/s	75 W	28 nm	- -	- -
2017		TPU v2	46 TFLOPs/s	280 W	16 nm	256	11 PetaFLOPs/s
2018		TPU v3	123 TFLOPs/s	450 W	16 nm	1024	125 PetaFLOPs/s
2020	TPU v4i (TPU v4 lite)		138 TFLOPs/s	175 W	7 nm	- -	- -
2021		TPU v4	275 TFLOPs/s	- -	7 nm	4096	≥1 ExaFLOPs/s
2023	TPU v5e (TPU v5 lite)		197 TFLOPs/s	- -	- -	256	

Jouppi et al., Ten Lessons From Three Generations Shaped Google's TPUv4i, ISCA, 2021

* Bfloat16 FLOPS

Ten Lessons and how they shaped TPUs

10 Lessons Learned Over ~10 Years

1. DNNs grow rapidly in memory and compute
2. DNN workloads evolve with DNN breakthroughs
3. Can optimize DNN as well as compiler and hardware
4. Inference SLO limit is P99 latency, not batch size
5. Production inference normally needs multi-tenancy

DNN Models

6. It’s the memory, stupid (not the FLOPs)
7. DSA Challenge: Optimize for domain while being flexible
8. Logic, Wires, SRAM, & DRAM improve unequally
9. Maintain compiler optimizations and ML compatibility
10. Design for performance per TCO vs perf per CapEx

Hardware/Architecture

Lesson 1: DNN Model Growth

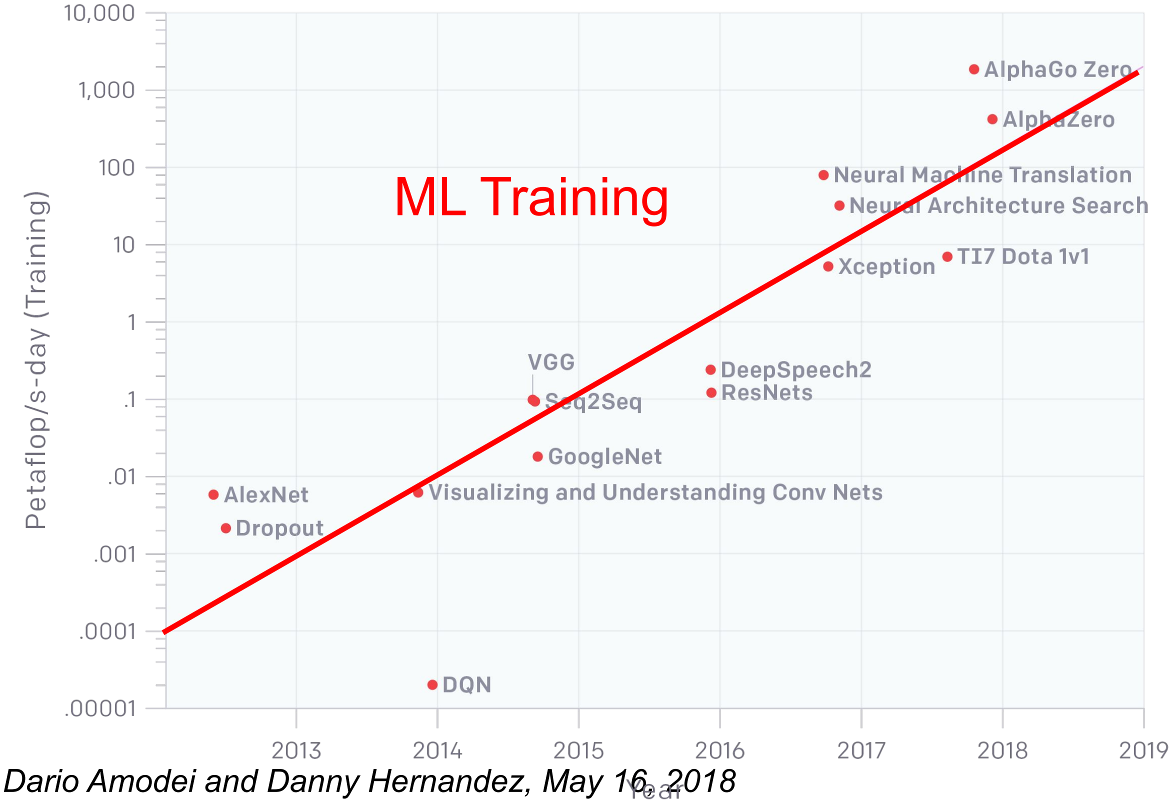
- For inference production DNNs, accelerators need headroom for growth in memory footprint and FLOPS over lifetime of deployment
  - ~1.5X per year in memory & FLOPs
- 1+ year design, 1+ year deployment, 3+ year service
  - 1.5^5 = ~8X!

Model	Annual Memory Increase	Annual FLOPS Increase
CNN1	0.97	1.46
CNN0	1.63	1.63
MLP0	2.16	2.16
MLP1	1.26	1.26

Lesson 1: DNN Model Growth

- New models getting even larger
- 2012-19, ML training compute SOTA 10X/year!
- GPT-3 “breakthrough” is simply 100X bigger: GPT-2 ⇒ GPT-3
  1.5B ⇒175B parameters

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



From “[AI and Compute](https://openai.com/blog/ai-and-compute/).” Dario Amodei and Danny Hernandez, May 16, 2018
https://openai.com/blog/ai-and-compute/

Lesson 2: DNN Workloads Evolve with DNN Breakthroughs

- Google DNN workloads 2016 vs 2020
- Past benchmarks still important (MLP, CNN)
- RNNs replaced LSTMs
- Added BERT models
  - Some apps switched from MLP to BERT
  - MLPerf 0.7 inference also added BERT
  - BERT published 2018!
- DSA needs to be general enough to handle new models

DNN Name	*2020*	*2016*
MLP0	25%	61%
MLP1		
CNN0	18%	5%
CNN1		
LSTM0	0%	29%
LSTM1		
RNN0	29%	0%
RNN1		
BERT0	28%	0%
BERT1		
TOTAL	100%	95%

Lesson 3: Can optimize DNN as well as compiler and hardware

- OK to change DNN as well as compiler and hardware to improve performance as long as maintain or improve DNN quality
  - Unlike CPUs where benchmark code is sacrosanct
- DNNs easier since 100s or 1000s of lines of TensorFlow code
  - Unlike CPUs where benchmarks can be 100,000s of lines of C++ code
- *Platform-aware AutoML** uses *Neural Architecture Search* (*NAS*) to Pareto-optimize ML model performance and quality on ML accelerators
  - Searches a space of more than O(2^{300}) candidates
- Discovered DNN is 1.6X performance at comparable quality for CNN1
- Using ML to improve ML performance!

* Li, S., Tan, M., Pang, R., Li, A., Cheng, L., Le, Q.V. and Jouppi, N.P., 2021. Searching for Fast Model Families on Datacenter Accelerators. Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition

Lesson 4: Inference SLO Limit is Latency, Not Batch Size

- Some accelerators claim batch size must be 1 to keep latency low. In reality:

Type	DNN	Latency constraint	Batch size
Production	MLP0	7 ms	512
	MLP1	20 ms	128
	CNN0	1 ms	16
	CNN1	32 ms	16
	RNN0	60 ms	8
	RNN1	10 ms	32
	BERT0	5 ms	128
	BERT1	10 ms	64
MLPerf	Resnet50	15 ms	16
	SSD	100 ms	4
	GNMT	250 ms	16

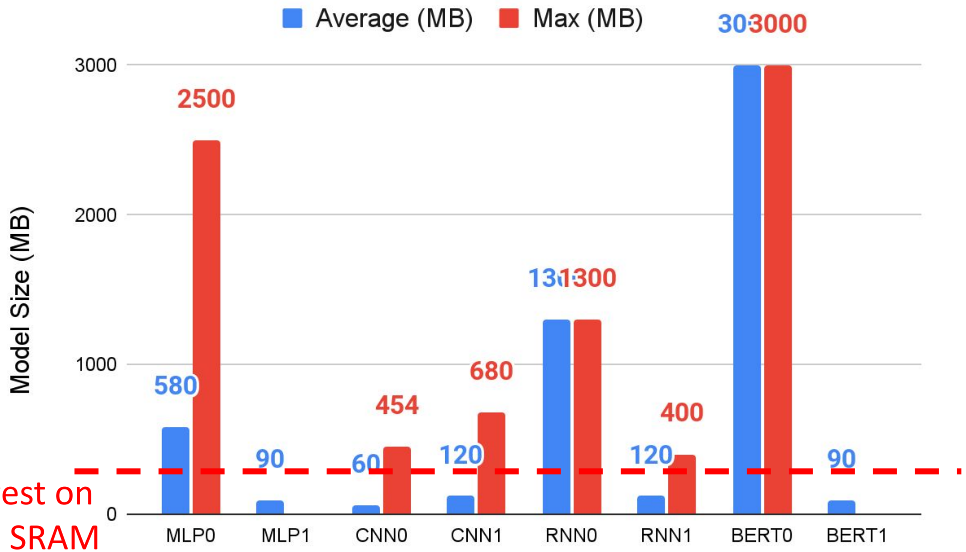
- Google’s production workloads have ~9X larger batch size despite ~7X stricter latency limit than MLPerf

Lesson 5: Production Inference Needs Multi-tenancy

- Many inferencing applications need to support multiple models
  - Want near zero switching time between models (e.g., <100 µs)
- Examples:
  - Translate - many different language pairs and models
  - Development - Main model plus experimental models
  - Multiple batch sizes to balance throughput and latency

Lesson 5: DNN Tenancy and Size (Feb 2020)

	Multi-tenancy?	*Avg # Programs (StdDev), Range*
MLP0	Yes	27 (±17), 1-93
MLP1	Yes	5 (±0.3), 1-5
CNN0	No	1
CNN1	Yes	6 (10), 1-34
RNN0	Yes	13 (±3), 1-29
RNN1	No	1
BERT0	Yes	9 (±2), 1-14
BERT1	Yes	5 (±0.3), 1-5

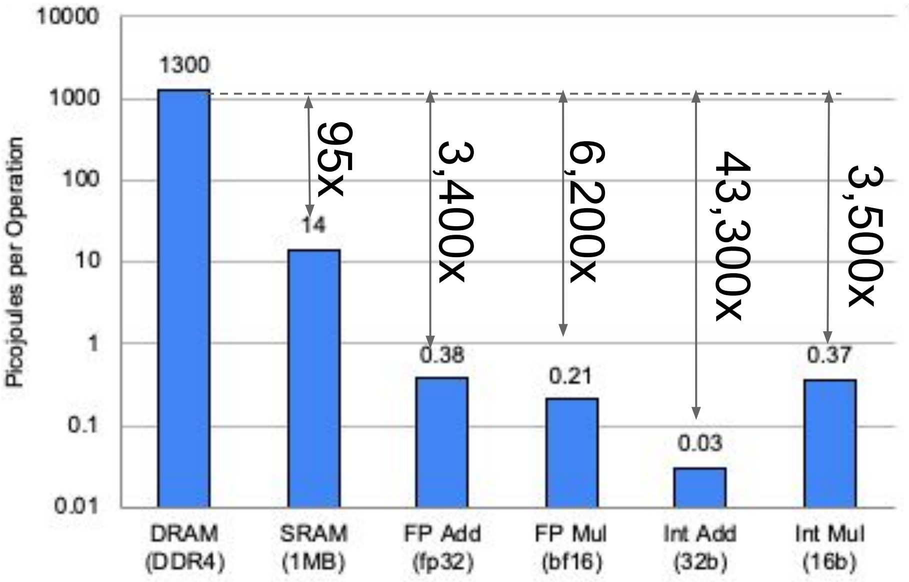


Largest on chip SRAM (~300 MB)

- 10s of ms context switching if reloading parameters from CPU host
- Need to fast DRAM to swap multiple models

Lesson 6: It's the memory, stupid! (not the FLOPs)

- Energy limits modern chips, not number of transistors
- External memory access energy ~100X on chip memory access ~ 10,000X arithmetic operation
- Easy to scale up FLOPs/sec by adding many ALUs to balance energy of memory accesses
    - Also why DNN model developers should focus on reducing memory accesses versus reducing FLOPs

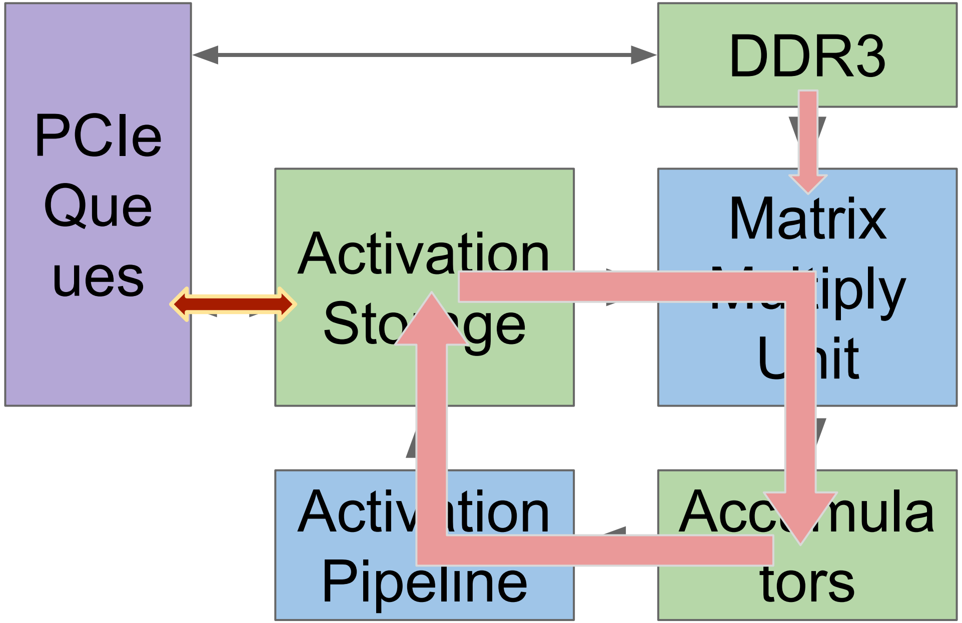


Jouppi, N., Yoon, D-H, Jablin, T., Kurian, G., Laudon, J., Li, S., Ma, P., Ma, X., Patil, N., Prasad, S., Young, C., Zhou, Z., and Patterson, D., 2021. Ten Lessons From Three Generations Shaped Google's TPUv4i, In Proc. 48th International Symposium on Computer Architecture.

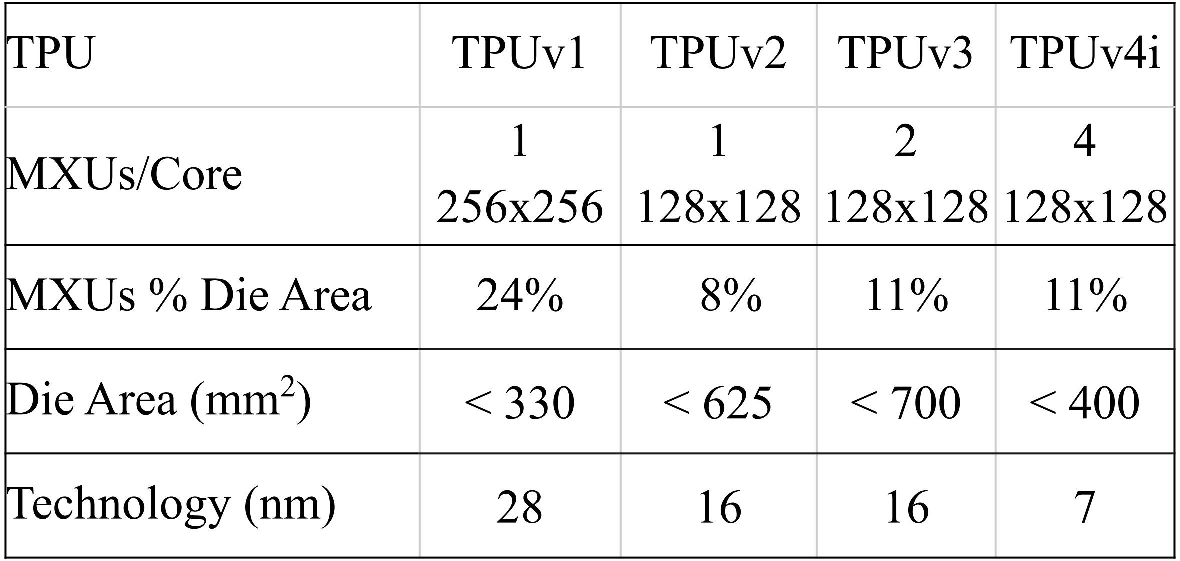
Paper showing that FLOPS Misleading: Dehghani, M., Arnab, A., Beyer, L., Vaswani, A. and Tay, Y., 2022. The efficiency misnomer. ICLR.

TPUv1: High-level Chip Architecture

- The Matrix Unit: 65,536 (256x256) 8-bit multiply-accumulate units
  - >25X as many MACs vs GPU
  - >400X as many MACs vs CPU
- 700 MHz clock rate
- Peak: 92T operations/second
  - 65,536 * 2 * 700M
- 4 MiB of on-chip Accumulator memory
- 24 MiB of on-chip Activation Storage
  - 3.5X on-chip memory vs GPU
- Two 2133MHz DDR3 DRAM channels for weights (8 GiB)



Easier to Scale FLOPs/sec as Logic improves quickest



TPU	TPUv1	TPUv2	TPUv3	TPUv4i
MXUs/Core	1 256x256	1 128x128	2 128x128	4 128x128
MXUs % Die Area	24%	8%	11%	11%
Die Area (mm^2)	< 330	< 625	< 700	< 400
Technology (nm)	28	16	16	7

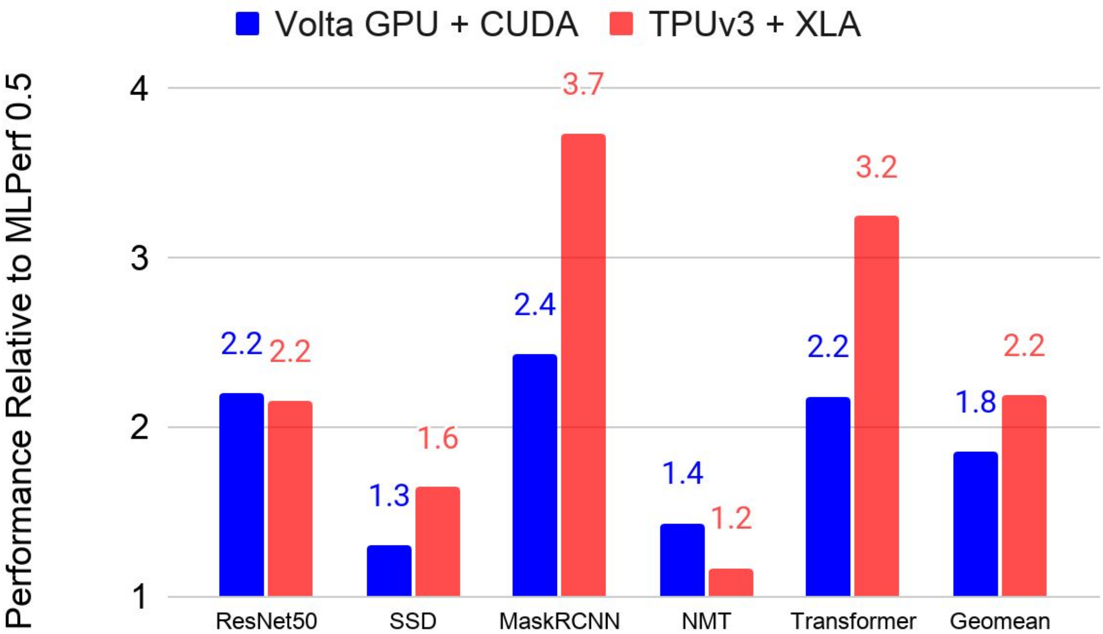
Jouppi et al., Ten Lessons From Three Generations Shaped Google's TPUv4i, ISCA, 2021

Lesson 9: Maintain compiler optimizations and ML compatibility

- XLA (*accelerated Linear Algebra*) compiler does whole-program analysis and optimization
  - Divided into HLO ops (machine independent) and LLO ops (machine dependent)
  - HLO optimizations apply to all TPU/GPU/CPU systems, changes at LLO level OK
- XLA exploits huge parallelism represented in a TensorFlow input dataflow graph
  1. Multicore Parallelism: Up to 4096 chips
  2. Data Level Parallelism: 2D vector and matrix functional units
  3. Instruction Level Parallelism: VLIW instruction set (format 322–400 bits)
- 2D vector registers, compute units ⇒ good data layout in units & memory
- No caches ⇒ XLA manages all memory transfers
- DSA software stacks less mature than CPU SW stacks; how fast improve?

Lesson 9: Maintain compiler optimizations and ML compatibility

- Compilers take time to mature and produce good quality code
  - Learning curve for new architecture and new DSA apps
  - Speedup MLPerf 0.7 (7/2020) vs. MLPerf 0.5 (11/2018)



Google

Dire Projections of Carbon Emissions for ML Training

Malthusian Predictions about ML Training

- Environmental cost to improve ML task (2024)?* *“The answers are grim: Training such a model would cost* ***US $100 billion*** *and would* ***produce as much carbon emissions as New York City does in a month****. And if we estimate the computational burden of a 1 percent error rate, the results are considerably worse.”*

  Thompson, N.C., et al., October 2021. Deep Learning's Diminishing Returns: The Cost of Improvement is Becoming Unsustainable, *IEEE Spectrum*

- *“In fact, by 2026, the training cost of the largest AI model predicted by the compute demand trend line would* ***cost more than the total U.S. GDP****.” [$20T]*

  Lohn, J. and Musser, M., January 2022. AI and Compute—How Much Longer Can Computing Power Drive Artificial Intelligence Progress? Center for Security and Emerging Technology

January 2022

AI and Compute

How Much Longer Can Computing Power Drive Artificial Intelligence Progress?

CSET Issue Brief

CSET
CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

AUTHORS
Andrew J. Lohn
Micah Musser

What's Next for Deep Learning › Another AI winter or eternal sunshine? P.26

Inside DeepMind's Robot Lab › An AI powerhouse takes on “catastrophic forgetting” P.34

The 7 Biggest Weaknesses of Neural Nets › Surprise! One of them is math P.42

FOR THE TECHNOLOGY INSIDER
OCTOBER 2021

IEEE Spectrum

Why Is AI So Dumb?
A SPECIAL REPORT

IEEE

Google

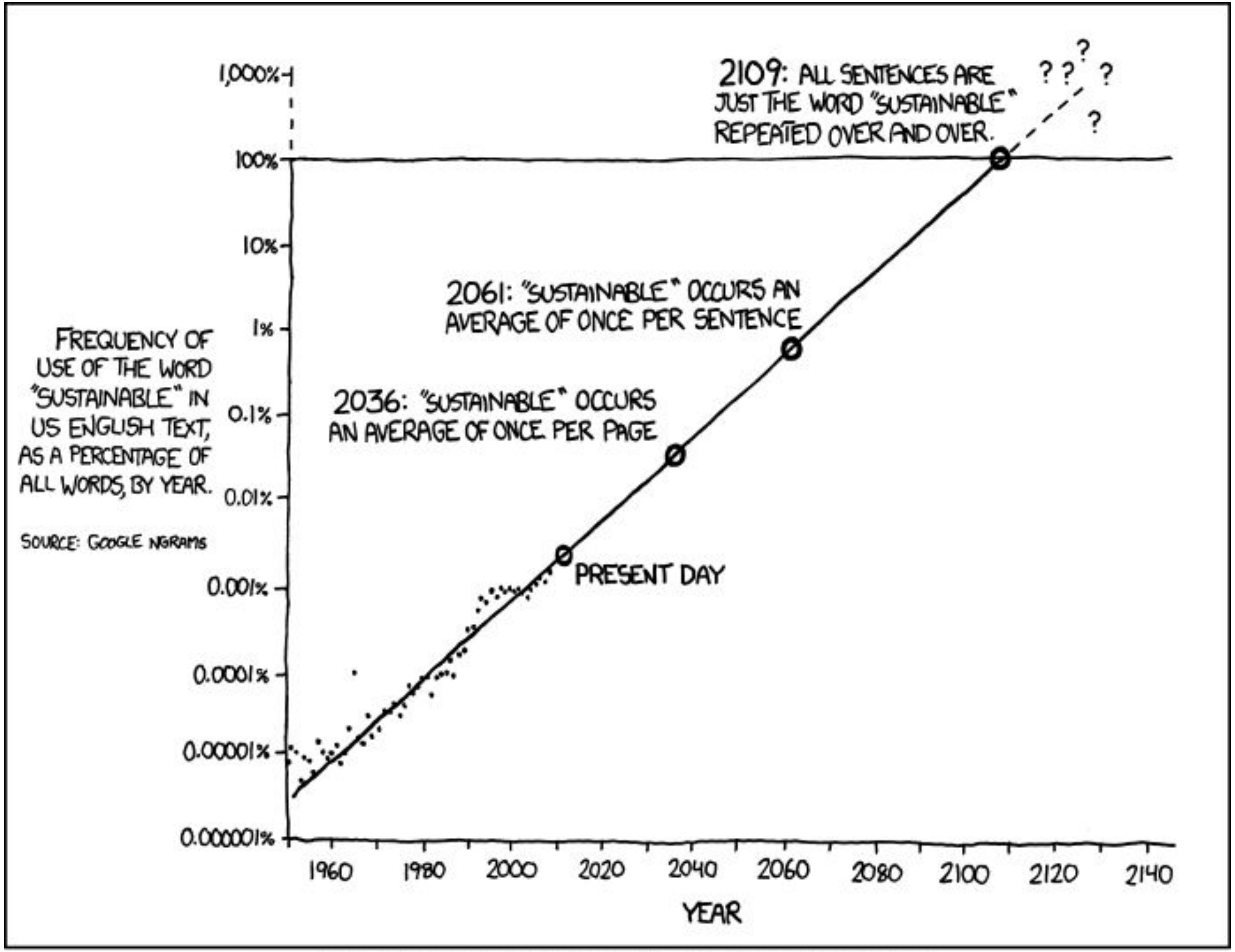
* The ML task is object recognition using the Imagenet benchmark
to reduce the error rate for an ML task* to a 5% from 11.5% today.

Google Demonstration: Millionfold better in 2 years???

- Same authors of those papers said training Evolved Transformer by Google in 2019 took:
  2M GPU hours*,
  5X lifetime emissions of a car**,***
- Will show 2 years later trained an equivalent model with a millionfold less emissions!

* Thompson, N.C., et al., 2020. The computational limits of deep learning. arXiv:2007.05558.
** Freitag, C., et al., 2021. The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations. *Patterns 2(9)*.
*** Dodge, J., et al., 2022, Measuring the carbon intensity of AI in cloud instances. ACM Conference on Fairness, Accountability, and Transparency.

https://xkcd.com/1007/



THE WORD "SUSTAINABLE" IS UNSUSTAINABLE.

Google

We studied Operational energy use, not Embodied

- Responsible AI is a broad topic; this focus is carbon emissions from ML training (matching much of the attention in ML community and public)
- Emissions can be classified as
  - *Operational*: energy cost of operating ML hardware including datacenter overheads , or
  - *Embodied:* operational + embedded carbon emitted during manufacturing of all components, from chips to datacenter buildings
- Like prior work we focus on operational emissions
  - Estimating embodied emissions is a larger, more difficult, future study
- Emissions measured as tCO_2e = 1000 kg of *CO_2 equivalent emissions*
  - Includes greenhouse gases like methane

How to document energy use and CO_2e

KWh = Hours to train × Number of Processors × Average Power per Processor × PUE

- Google, Facebook publish quarterly PUE for all regions (e.g., Iowa, Oklahoma)
  - *Power Usage Effectiveness*: energy overhead “wasted” in datacenter (doesn’t get to computers); if overhead is 50%, PUE = 1.5
- ML experts already know Hours to Train and Number of Processors
- Average Power per Processor:
  - Measure power while running like we did
  - Or reuse our Google average power numbers
    - TPUv2: 228 Watts ± 5% (Transformer, Evolved Transformer, NAS)
    - P100 GPU: 284 Watts ± 10% (Transformer, Evolved Transformer, NAS)
    - TPU v3: 283 Watts ± 10% (T5, Meena, Gshard, Switch Transformer)
    - V100 GPU: 325 Watts ± 2% (GPT-3, Transformer Big)

tCO_2e = KWh × tCO_2e per KWh

- Ask datacenter operator for tCO_2e per KWh
  - Google publishes %carbon free energy per datacenter

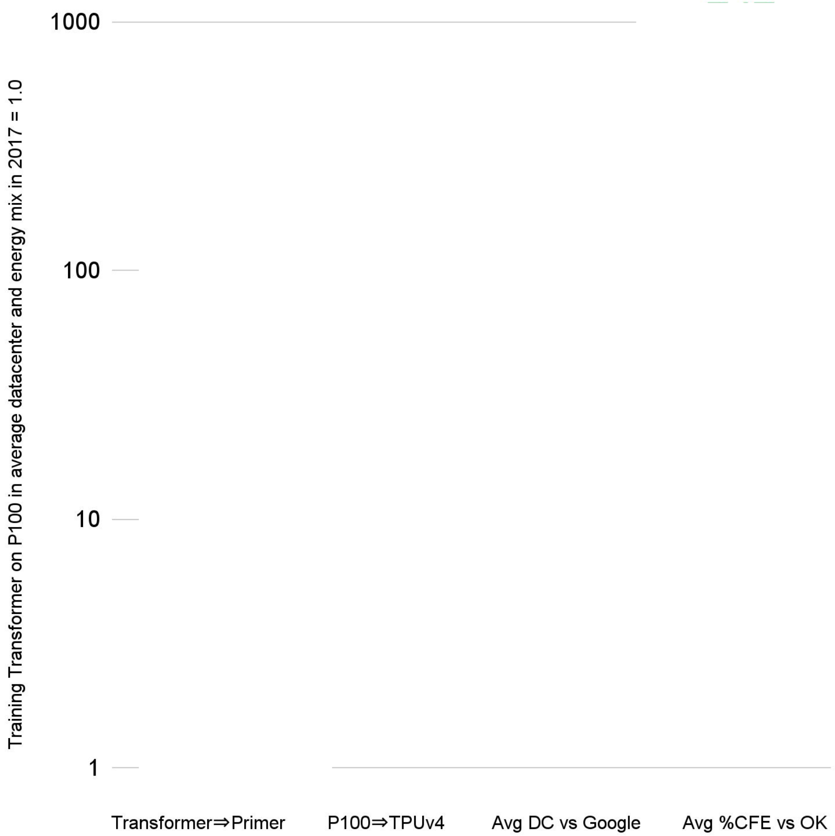
4Ms of Energy Efficiency for ML

Good News #1: Reduce energy 100X, CO_2e 1000X!

CO_2 equivalent emissions (CO_2e) include greenhouse gases

Energy efficiency in ML can be improved by 4 (multiplicative) best practices "4Ms of ML Energy Efficiency"

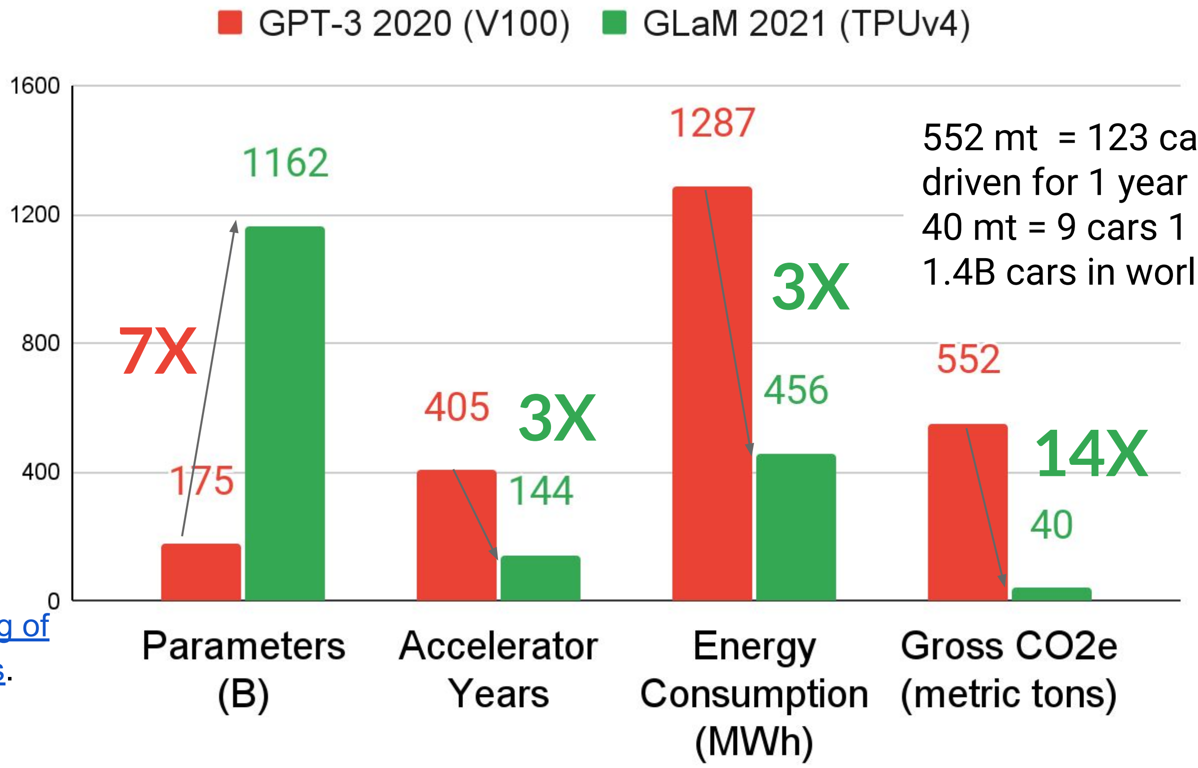
1. **Model. Transformer (2017) to Primer (2021) is 4x**
2. **Machine. P100 (2017) to TPUv4 (2021) is 14x**
3. **Mechanization (datacenter efficiency). PUE from global average to Google average is 1.4x**
4. **Maps (geographic location, energy source). Avg %Carbon Free Energy (2017) to Google OK %CFE is 9x (2021)**



Thanks to Cliff Young for 4M mnemonic!

4Ms for NLP: GLaM (TPUv4, Google Oklahoma datacenter, 2021) vs GPT-3 (V100 GPU, Microsoft datacenter, 2020)

- 18 months after GPT-3
- GLaM has *better accuracy* for same tasks as GPT-3
- **7X** more parameters
- Mixture of experts: **8% parameters**/token
- **3X** less time, energy
- **14X** less CO$_2$e



552 mt = 123 cars driven for 1 year
40 mt = 9 cars 1 year
1.4B cars in world

Du, N., et al 2021. GLaM: [Efficient Scaling of Language Models with Mixture-of-Experts](). arXiv preprint arXiv:2112.06905.

What about ML on the Edge?

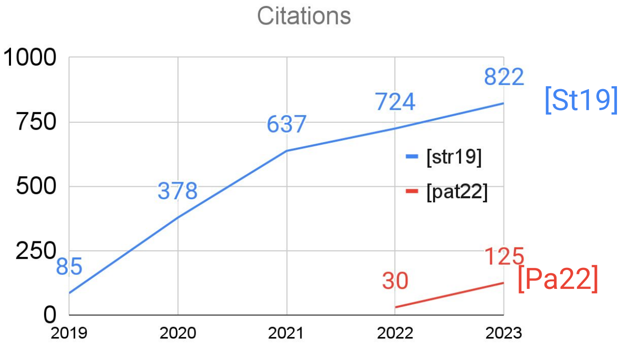
- New paper Feb 2024 *Communication of the ACM*
- “Energy and Emissions of Machine Learning on Smartphones versus the Cloud: A Google Case Study” by David Patterson, Jeffrey M. Gilbert, Marco Gruteser, Efren Robles, Krishna Sekar, Yong Wei, and Tenghui Zhu,
- Preview
  - ML is <3% of smartphone energy consumption
  - Smartphone charger inefficiency is a much larger energy consumption issue than ML
    - Chargers were responsible for 80% of energy use (wireless + multiple chargers)
  - While training ML models on smartphones has inherent advantages for privacy, it can have 100× the carbon footprint of training in the cloud

Climate change is one of our most important challenges

- **But must get numbers right to ensure work on biggest challenges**

Good News #3: Dire Predictions too high by >100,000x

- [So19] NAS for Evolved Transformer
  - Neural Architecture Search done once per problem domain+architectural search space
  - Didn’t include energy or emissions
- Concerns rightly raised about CO$_2$e of ML
- [St19] estimated emissions of this NAS
  - Cited ~2700 times
  - Used P100 vs TPUv2, US averages vs Google DC: 5X too high for NAS (4Ms)
  - + Used full model vs small proxy for search: 19X (88X too high for NAS)
- Some papers citing [Str19] confused NAS with Training
  - NAS emissions ~1300x training emissions of DNN model found in search
- 5 × 19 × 1300 = 120,000x (different 4Ms × flawed NAS × searching vs training)
- 2.4 kg vs 248,019 kg; 5 car lifetimes to < 0.00005 car lifetimes



[Str19] Strubell, E., Ganesh, A. and McCallum, A., June 2019. Energy and policy considerations for deep learning in NLP. *Annual Meeting of the Association for Computational Linguistics*.

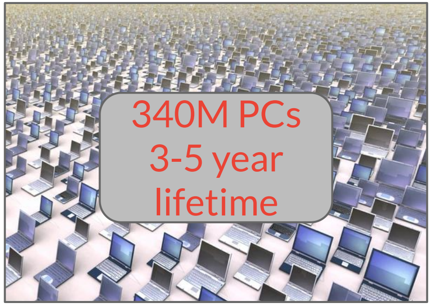
[So19] So, D., Le, Q. and Liang, C., 2019. The Evolved Transformer. In International Conference on Machine Learning (ICML).

[Pa22] Patterson, D., et al., 2022. The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink, *IEEE Computer*.

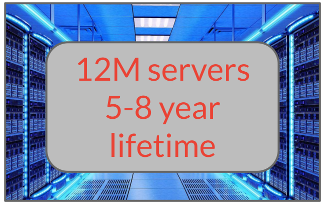
Get numbers right to ensure working on the actual biggest information technology challenge

- **Within IT, more likely climate challenge is embodied cost of manufacturing computing equipment of all types/sizes vs operational cost of ML training**

1700M cell phones in 2021
2-3 year lifetime



340M PCs
3-5 year
lifetime



12M servers
5-8 year
lifetime

Conclusion and Recommendations

Conclusion

- 10 Lessons learned from previous TPU generations drove next design
    1. DNNs grow rapidly in memory and compute
    2. DNN workloads evolve with DNN breakthroughs
    3. Can optimize DNN as well as compiler and hardware
    4. Inference SLO limit is P99 latency, not batch size
    5. Production inference normally needs multi-tenancy
    6. It’s the memory, stupid (not the FLOPs)
    7. DSA Challenge: Optimize for domain while being flexible
    8. Logic, Wires, SRAM, & DRAM improve unequally
    9. Maintain compiler optimizations and ML compatibility
    10. Design for performance per TCO vs perf per CapEx
- Four generations of TPU significantly improve Perf/TCO and Emissions of ML
    - 2019−2021: ML 70%−80% of FLOPS but only 10%−15% of Google energy use

Recommendations for ML Research and ML Practice

- **Model**: ML researchers keep developing more efficient ML models: 2x–4x
  - Research Challenge: Reduce cost of training and inference of giant models like GPT-3, GlaM
    - Focus on memory accesses vs FLOPS
  - Practice: Also publish energy consumption and carbon footprint of model to
    - Foster competition beyond ML quality e.g., speed, emissions
    - Ensure accurate accounting of their work (external estimates were off 100x–100,000x)
- **Machine**: Build faster, more efficient ML HW (e.g., A100 GPU, TPU v4): 2x–4x
  - Research Challenge: Leverage Sparsity with Systolic Arrays
  - Research Challenge: How to do embodied costs, not just operational costs
- **Mechanization**: Data center operators publish datacenter efficiency (PUE): 1.4x
  - Practice: Also publish % carbon free of energy supply per location
- **Map**: ML practitioners use greenest data centers per region, often in Cloud: 5x-10x
  - Practice: Increase carbon free energy per location (2 in Europe, 3 in US ~90% carbon free energy)
- Co-optimize 4Ms to realize the amazing potential of ML to positively impact many fields in a sustainable manner

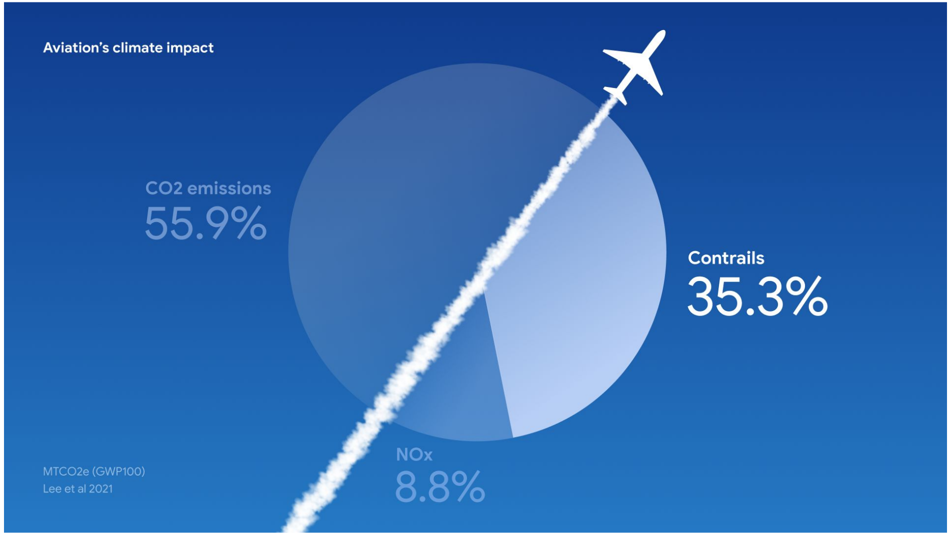
Thanks to the TPU Team, Including:

Gaurav Agrawal, Catherine Ahlschlager, Ahmet Akyildiz, Ashby Armistead, Sandeep Bhatia, Rich Bonderson, Oliver Bowen, Roger Carpenter, Andrew Casper, Clifford Chao, Dehao Chen, Chiachen Chou, William Chwee, Xiangyu Dong, Houle Gan, Rakesh Gautam, Peter Gavin, Arnd Geis, Ben Gelb, Russ Gibbons, Sandeep Giri, Vinayak Gokhale, Pareesa Golnari, Rajendra Gottipati, Nils Graef, Jesse Guss, Benjamin Gwin, David Haskell, Blake Hechtman, Matthew Hedlund, Jian Ho, Doug Hogberg, Jerry Huang, Michael Hsu, Adam Hutchin, Mike Hutton, Berkin Ilbeyi, Srikrishna Iyer, Arpith Jacob, Indira Jayaram, Chetan Kale, Pankaj Kanwar, Srinidhi Kestur, Teju Khubchandani, Woon-Seong Kwon, Namhoon Kim, Andy Koch, Alan Kulawik, Poorna Kumar, Alice Kuo, Steve Lacy, Joshua Lang, Chester Li, Avinash Lingamneni, Derek Lockhart, Stephen Longfield, Fong Lou, Tao Liu, Kyle Lucke, Adriana Maggiore, David Majnemer, Seth Merriman, Rolf Mueller, David Munday, Mandar Munishwar, Hithesh Murthy, Lifeng Nai, Spoorthy Nanjaiah, Andrew Noonan, Alexander Nguyen, Vinh Nguyen, Tayo Oguntebi, Virag Parekh, Jose Baiocchi Paredes, Sang-Keun Park, Tejas Parikh, Omkar Pathak, Ram Babu Penugonda, Andy Phelps, Vaishali Raghuraman, Guru Rajamani, Andrew Ranck, Paul Rodman, Bjarke Roune, Ohad Russo, Amit Sabne, Amir Salek, Kirk Sanders, Julian Schrittwieser, Chris Severn, Boone Severson, Hamid Shojaei, Jaideep Singh, Tej Soni, Jaswanth Sreeram, Dan Steinberg, Jim Stichnot, Qian Sun, Mercedes Tan, Hua Tang, Horia Toma, Alex Thomson, Ani Udipi, Dimitris Vardoulakis, Sandeep Venishetti, Jack Webber, Monica Wong-Chan, Hsin-Jung Yang, Mingyao Yang, Xiaoming Yu, Lu Yuan, Sara Zebian, Feini Zhang, Ce Zheng, and many others.

Backup Slides

Reducing Contrails with AI

- Contrails ~35% Aviation CO2e
  - Increases warming since trap heat
  - Also reflect sunlight (during the day)
- Only occur when planes fly through humid regions
- Google AI predicts path for planes to reduce contrails
- American Airlines flew 70 test flights over 6 months
  - 0.3% more fuel for whole fleet
  - 54% reduction in contrails
  - Cost per tCO2e saved $5-$25
  - Cost to extract tCO2e from air ~$1000



- 2020 Data centers/data transmission ~330 Mt
- Aviation in 2019 ~900 Mt CO2e
  - ~2% of global energy-related CO2e
- AI might save 54% x 35% x 900 = ~170 Mt CO2e