Improvements of Carbon Emissions Efficiency from AI Workloads

Vincent Poncet
Google Cloud
April 2023

Based on “The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink”, July 2022 in *IEEE Computer* , and “Carbon Emissions and Large Neural Network Training” April 2021, arXiv preprint arXiv:2104.10350
by David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean

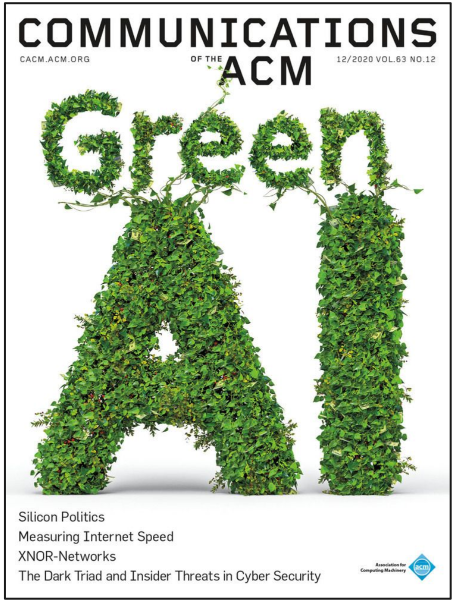
AI and Climate Change

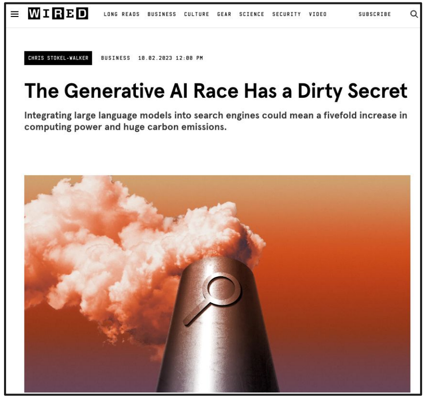
Lots of external interest on Energy Consumption and CO_2 emissions of ML recently:

- [Str19] Strubell, E., Ganesh, A. and McCallum, A., June 2019. Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:1906.02243
- [Lac19] Lacoste, A., Luccioni, A., Schmidt, V. and Dandres, T., Nov 2019 Quantifying the carbon emissions of machine learning
- [Tho20] Thompson, N.C., et al., 2020. The computational limits of deep learning. arXiv preprint arXiv:2007.05558.
- [Sch20] Schwartz, R., Dodge, J., Smith, N.A. and Etzioni, O., Dec 2020. Green AI. *Communications of the ACM*, 63(12), pp.54-63
- [Fre21] Freitag, C., et al, 2021. The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations. *Patterns*, 2(9).
- [Luc22] Luccioni, Viguier, Ligozat, 2022. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. arXiv:2211.02001

Media coverage is growing:

- The Generative AI Race Has a Dirty Secret. 2023/02/10. Wired
- The internet contributes 1.6 billion annual tons in greenhouse gas emissions. Google and Microsoft's AI search war will make it worse. 2023/02/13 Business Insider
- The mounting human and environmental costs of generative AI. 2023/04/12. Ars Technica





What we learned

- Can reduce energy up to 100X(!), reduce CO_2e up to 1000X(!!) via best practices: pick DNN, ML accelerator, datacenter, location carefully ("4Ms")
    - **Model matters**: Sparsely activated DNNs consume ~10X less energy than large dense DNNs
    - **Machine matters**: TPUs ~2-5X more energy efficient than standard processors
    - **Mechanization matters**: Cloud ~1.4-2X more energy efficient than an average datacenter
    - **Map location matters**: varies ~5-10X tCO_2e/KWh *within same country & organization*

Talk Outline

1. Case Study from Transformer (2017) to Evolved Transformer (2019) to Primer (2021) as vary processor and datacenter (the “4Ms”)

2. Case study energy consumption and CO_2 emissions for recent large NLP models: T5, Meena, GShard, Switch Transformer, GLaM, and GPT-3

3. Update prior CO_2 emissions estimates that are off by 100X–100,000X

4. Address FAQs: ML Energy growth, Access to Cloud

We studied Operational energy use, not Lifecycle

- Emissions can be classified as
  - *Operational*, energy cost of operating ML hardware including datacenter overheads (Scope 2), or
  - *Lifecycle* additionally includes embedded carbon emitted during manufacturing of all components, from chips to datacenter buildings (Scope 3).
- Like most prior work (papers cited above) we focus on operational emissions
- Estimating lifecycle emissions is a larger, more difficult, future study

1) Case study: Transformer ⇒ Evolved Transformer ⇒ Primer

- **Model**: Compare Transformer on P100 GPUs in average US datacenter 2017 vs Evolved Transformer [So19] on TPUv2s in Google Iowa datacenter 2019 vs Primer [So21] on TPUv4s in Google Oklahoma datacenter 2021

- Algorithm/program improvement: save time, money, energy, and CO_2e
  - All models deliver same accuracy/error rate, some just faster
  - Evolved Transformer (2019) takes **1.3X** less time than Transformer ⇒ **1.3X** less CO_2e
  - Primer (2021) takes **4.2X** less time than Transformer ⇒ **4.2X** less CO_2e

- Four years later, **4.2X** better for same results from better ML model

1) Case Study: Processor P100 GPU ⇒ TPUv2 ⇒ TPUv4

- **Machine**: Net gain Evolved Transformer: TPUv2 (2019) performance/Watt is **5.7X** better than P100 GPU (2017) ⇒ another factor of **5.7X** less CO_2e from better processors
  - NVIDIA P100 GPU optimized for graphics, not for ML
- Net gain Primer: TPUv4 (2021) performance/Watt is **13.7X** better than P100 GPU (2017) ⇒ another factor of **13.7X** less CO_2e from better processors
- TPUs aim to improve performance/Total Cost of Ownership, almost perfectly linear correlated with performance/Watt* ⇒ saves time, money, energy, CO_2e
- 4 Years later, **13.7X** better for hardware optimized for ML vs standard processors

* 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., June 2021. Ten Lessons From Three Generations Shaped Google's TPUv4i, 48th International Symposium on Computer Architecture.

1) TPU Generations

92 teraops (int8)
Inference only

TPU v1
(2015)

46 teraops (bf16)
256 chips,
11.5 petaflops (bf16), 4TB HBM

Cloud TPU v2
(2017)

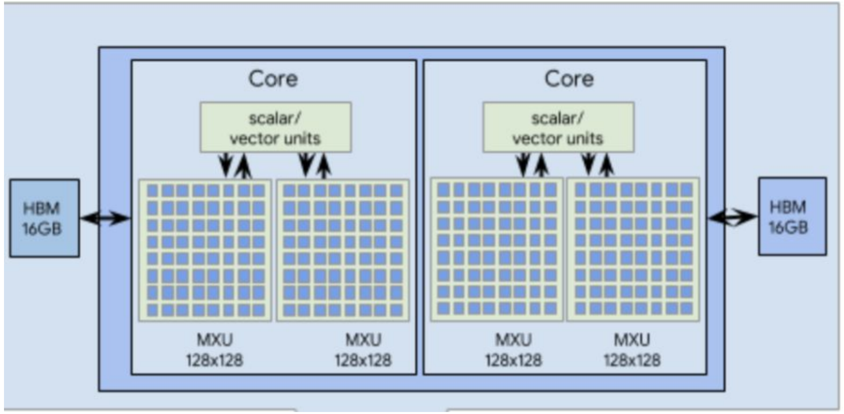
123 teraops (bf16)
1024 chips,
> 100 petaflops (bf16), 32TB HBM
2D Torus network
340TB/s all reduce

Cloud TPU v3
(2018)

Cloud TPU v4
(2021)

275 teraops (bf16)
4096 chips
> 1 exaflops (bf16), 128 TB HBM
3D Torus OCS
1.1PB/s all reduce

1) Specialized and more energy efficient architecture



A100 Tensor Core GPU has 7 GPCs, 7 or 8 TPCs/GPC, 2 SMs/TPC, up to 16 SMs/GPC, 108 SMs
64 FP32 CUDA Cores/SM, 6912 FP32 CUDA Cores/GPU
4 Tensor Cores/SM, 432 Tensor Cores per GPU.
40 MiB on chip memory
Integrated NVLink up to 8 GPUs
Ethernet or Infiniband for up to X Ks chips

TPUv4 Cores with 2 MXUs per core

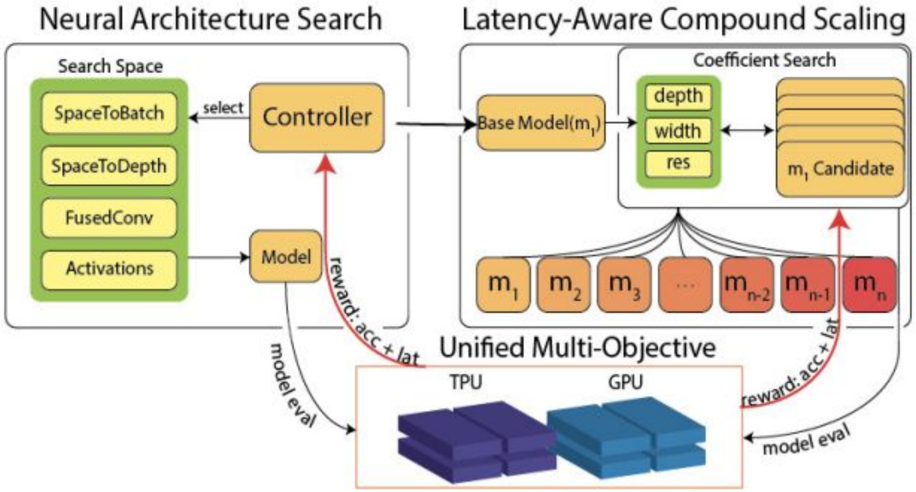
Embedding Lookup Unit, SparseCore
170 MiB on chip memory
Integrated Inter-Chip Interconnect to 4K chips through Optical 3D Torus Mesh

1) Case Study: Model/Hardware Efficiency

- More specialised hardware and platform-optimized models drives higher FLOPs efficiency

Model	# of Parameters (in billions)	Accelerator chips	Model FLOPS utilization
GPT-3	175B	V100	21.3%
Gopher	280B	4096 TPU v3	32.5%
Megatron-Turing NLG	530B	2240 A100	30.2%
PaLM	540B	6144 TPU v4	46.2%

- ML to optimize the efficiency of ML
  - Reinforcement Learning using Platform metrics
    - ALU energy/latency cost
    - Register/Cache/RAM/network energy/latency cost
  - Multi-Objectives
    - Training cost
    - Inference Latency



1) Case Study: PUE US Avg. Datacenter ⇒ Google Iowa

- **Mechanization**: *Power Usage Effectiveness (PUE)*: Energy overhead "wasted" in datacenter (doesn't get to computers)
  - Datacenter industry standard metric for energy efficiency
  - If 50% overhead, PUE is 1.50
  - Global average in 2017 was 1.60 (1.57 in 2021)
  - Google Iowa average in 2019 was 1.11
  - Google Oklahoma average in 2021 was 1.11
- PUE improvement is ~1.4X ⇒ 1.4X less CO_2e if use optimized datacenter
  - Cloud datacenters large, new warehouses optimized for energy efficiency vs on premise datacenters often 10X smaller, squeezed into space for other uses

1) Case Study: Energy US Avg. ⇒ Iowa ⇒ Oklahoma

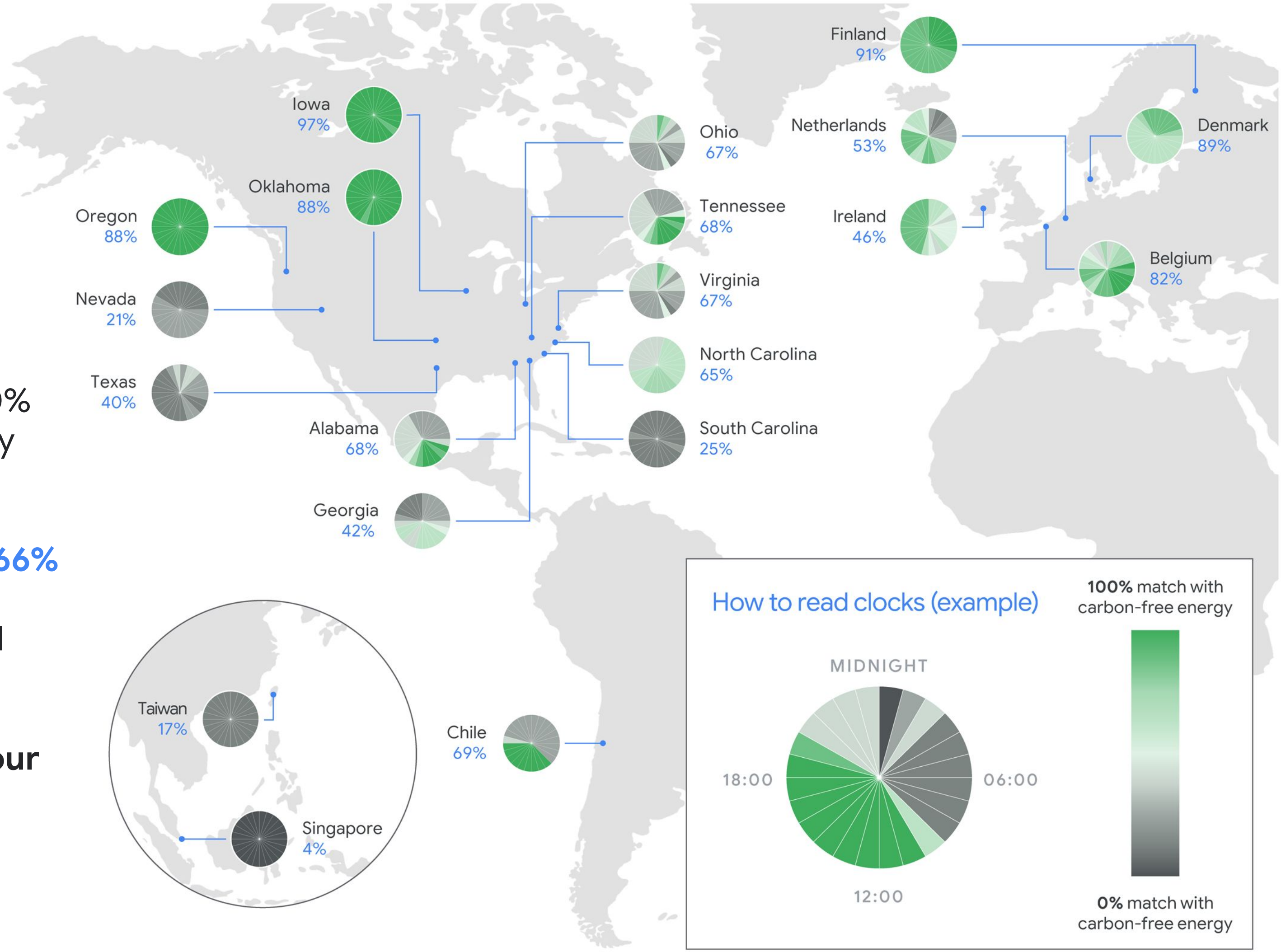
- **Map location**: Average US energy mix in 2017 ⇒ 0.488 kg CO_2e / KWh
- Google Iowa energy mix in Q4 2019 ⇒ 0.080 kg CO_2e / KWh (**5.4X** less)
- Google Oklahoma energy mix in Q2 2021 ⇒ 0.054 kg CO_2e / KWh (**9.0X** less)

- Google contracts with renewable energy projects since 2010, attained 100% renewable energy matched at the annual global basis since 2017
- Google announced all datacenters will use 100% carbon free energy (CFE) by 2030
  - 61% CFE in 2019, 66% CFE in 2021
  - “24x7”: Google purchases on same local grid in same hour ⇒ lowers net CO_2e/KWh value

Since 2017, Google is 100% Renewable Energy globally annually matched.

In 2021, Google reached **66% carbon-free energy** globally on an **hourly grid basis.**

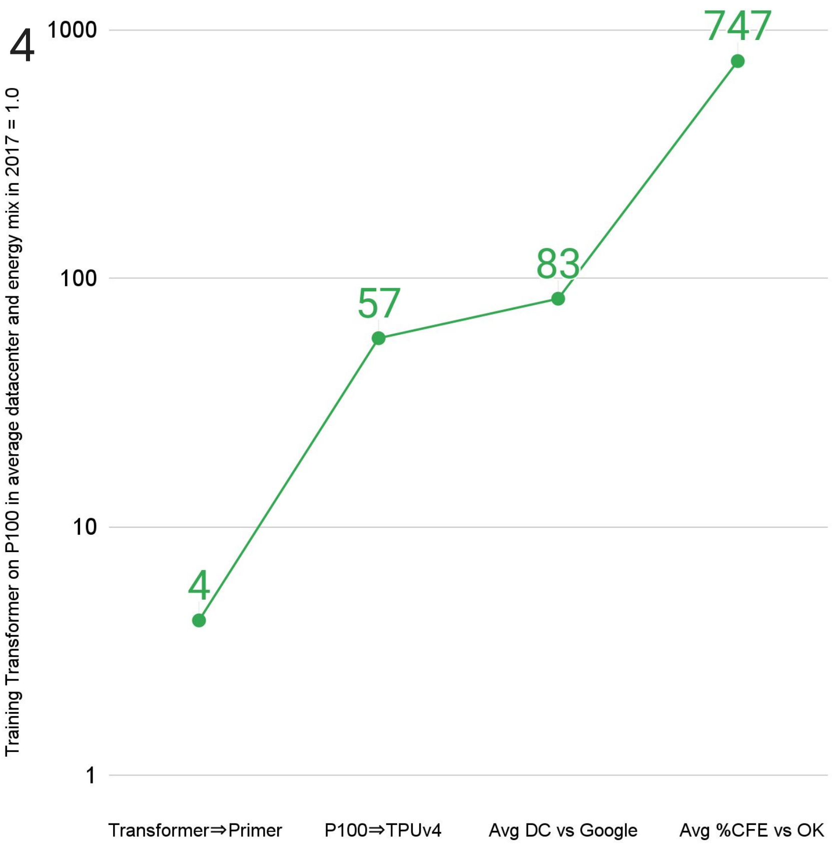
In the same year, **five of our data centers** operated **at or near 90%.**



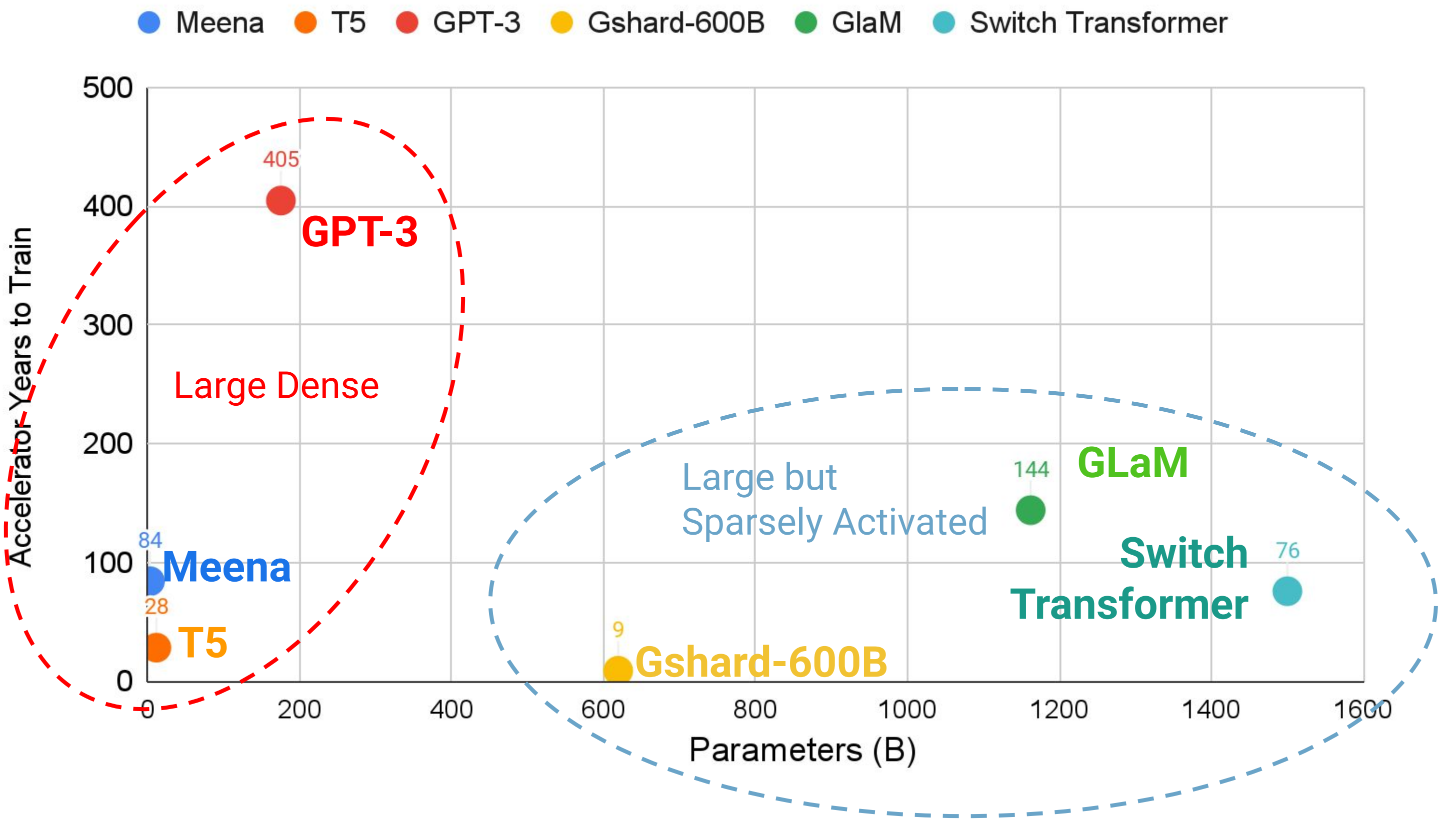
Cumulative Benefits: Reduce energy 100X, CO_2e 1000X!

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. **Map** (geographic location, energy source). Avg %Carbon Free Energy (2017) to Google OK %CFE is 9x (2021)

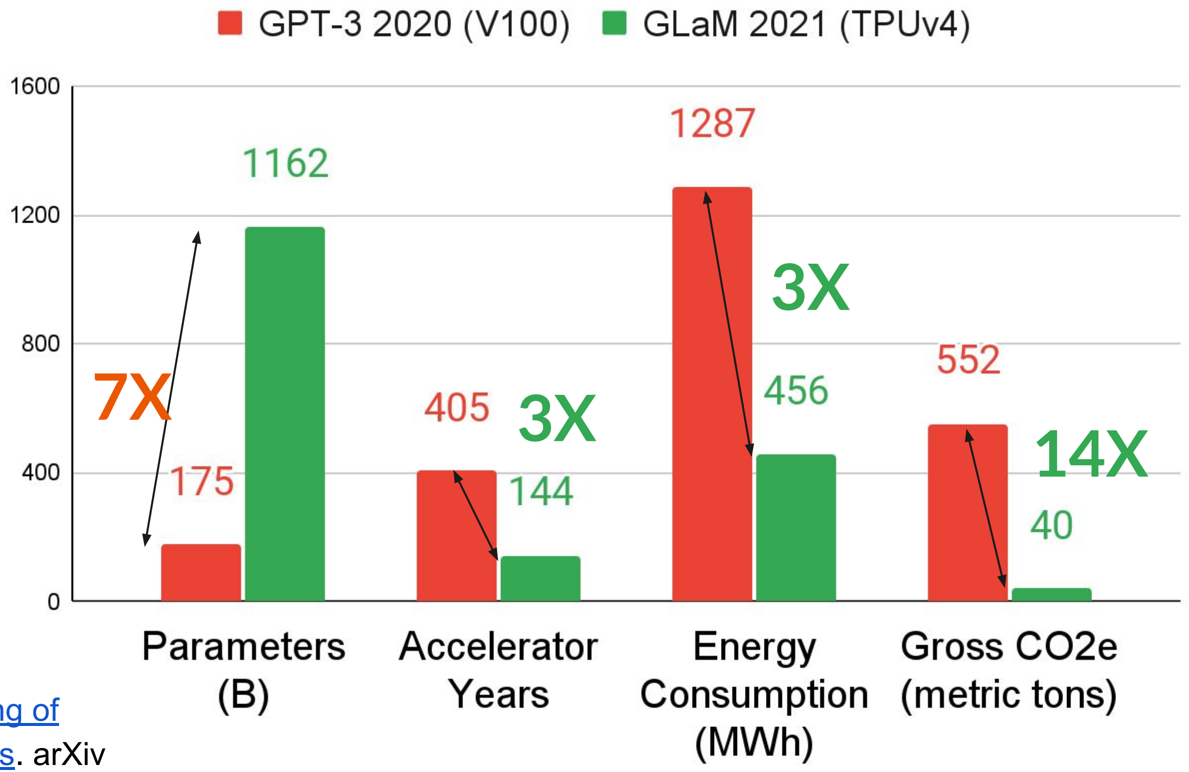


2) Large Recent NLP Models: Increase in Parameters vs Total Compute Time



GLaM (TPUv4, Google Oklahoma datacenter, 2021) vs GPT-3 (V100 GPU, Microsoft datacenter, 2020)

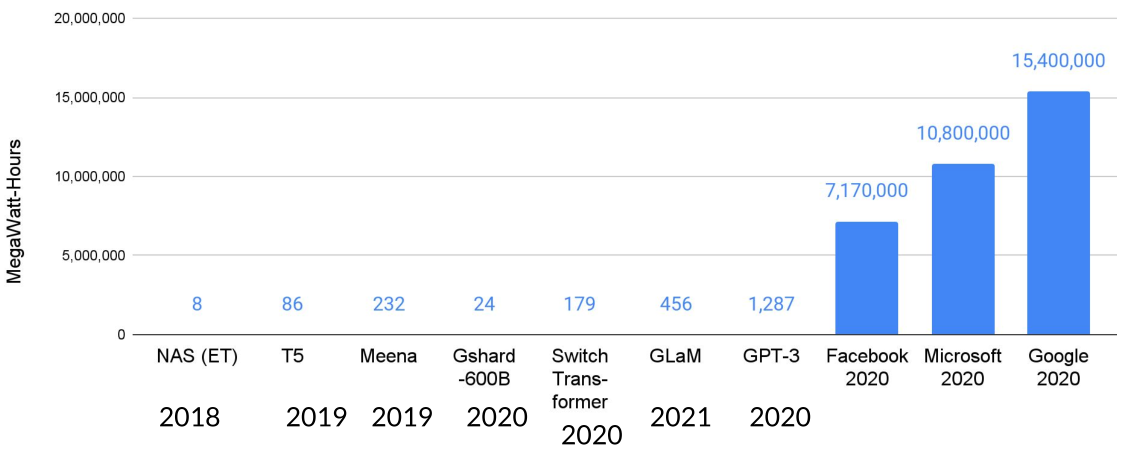
- 18 months after GPT-3
- GLaM has *better accuracy* for same tasks as GPT-3
- 7X more parameters
- Mixture of experts: 8% parameters/token
- 3X less time, energy
- 14X less CO_2e



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

Discussion: Is Training a large % of Cloud footprint?

- Google total energy consumed 2020 = 15.4 TeraWatt-hours
- Microsoft total energy 2020 = 10.8 TW-h
- Facebook 2020 = 7.1 TWh
- Energy for NAS + Meena + T5 + Gshard-600B + Switch Transformer + GLaM + GPT-3 is round off error



Li, S., Tan, M., Pang, R., Li, A., Cheng, L., Le, Q. and Jouppi, N.P., 2021. Searching for Fast Model Families on Datacenter Accelerators. [arXiv preprint arXiv:2102.05610](https://arxiv.org/abs/2102.05610).

Google's custom-designed Cloud TPU v4 Pods

Breathtaking Scale & Speed

- **Industry-leading interconnect:** 6 Tbps per host allowing to significantly speed up training
- **Embeddings acceleration:** perfect for embedding heavy models such DLRM

Price-Performance & Efficiency

- **Flop-per-dollar gains:** 2.2x more peak FLOPs and ~1.4x more peak FLOPs per dollar vs Cloud TPU v3
- **Exceptionally high utilization of these FLOPs** at scale up through thousands of Cloud TPU v4 chips

Sustainability / Zero carbon footprint

- **~90% direct clean energy supply** in our world's largest ML cluster in Oklahoma with up to 9 exaflops of peak compute
- **Lower carbon impact:** remaining operational carbon emissions are fully offset

ML Engagements team

- **A team of top ML experts:** responsible for ensuring customer success on any framework and developing end-to-end innovative and cutting edge solutions on Cloud TPUs

TensorFlow

What's so special about Google's custom-designed Cloud TPU v4 Pods?

Breathtaking Scale & Speed

- **Industry-leading interconnect:** 6 Tbps per host allowing to significantly speed up training
- **Embeddings acceleration:** perfect for embedding heavy models such DLRM

Price-Performance & Efficiency

- **Flop-per-dollar gains:** 2.2x more peak FLOPs and ~1.4x more peak FLOPs per dollar vs Cloud TPU v3
- **Exceptionally high utilization of these FLOPs at scale** up through thousands of Cloud TPU v4 chips

Sustainability / Zero carbon footprint

- **~90% direct clean energy supply** in our world's largest ML cluster in Oklahoma with up to 9 exaflops of peak compute
- **Low carbon impact:** remaining operational carbon emissions are fully offset

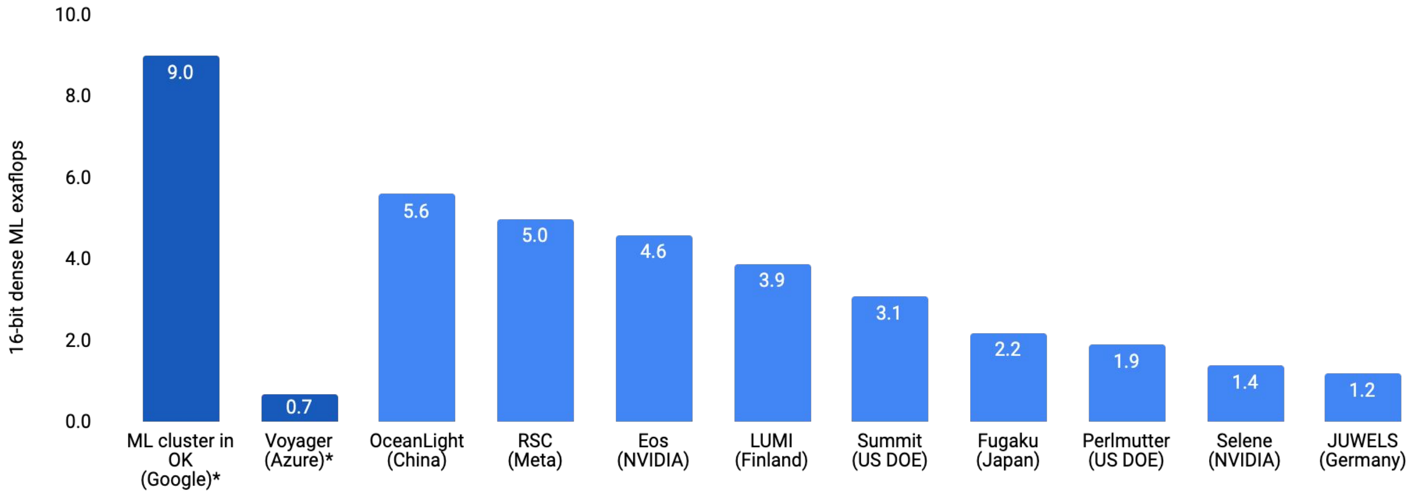
ML Engagements team

- **A team of top ML experts:** responsible for ensuring customer success on any framework and developing end-to-end innovative and cutting edge solutions on Cloud TPUs

TensorFlow

World's largest ML hubs / supercomputers

(16-bit dense / BF 16 ML exaflops)



* Clusters available in public cloud are highlighted in dark blue (compiled by jekbradbury@ based on publicly available information)

Google Cloud

Summary

- Can reduce energy up to 100X(!), reduce CO_2e up to 1000X(!!) via best practices: pick DNN, ML accelerator, datacenter, location carefully ("4Ms")
    - **Model matters**: Sparsely activated DNNs consume ~10X less energy than large dense DNNs
    - **Machine matters**: TPUs ~2-5X more energy efficient than standard processors
    - **Mechanization matters**: Cloud ~1.4-2X more energy efficient than an average datacenter
    - **Map location matters**: varies ~5-10X tCO_2e/KWh *within same country & organization*

- Cloud provides access to the highest specialized supercomputers on demand with the highest efficiency and lowest carbon footprint