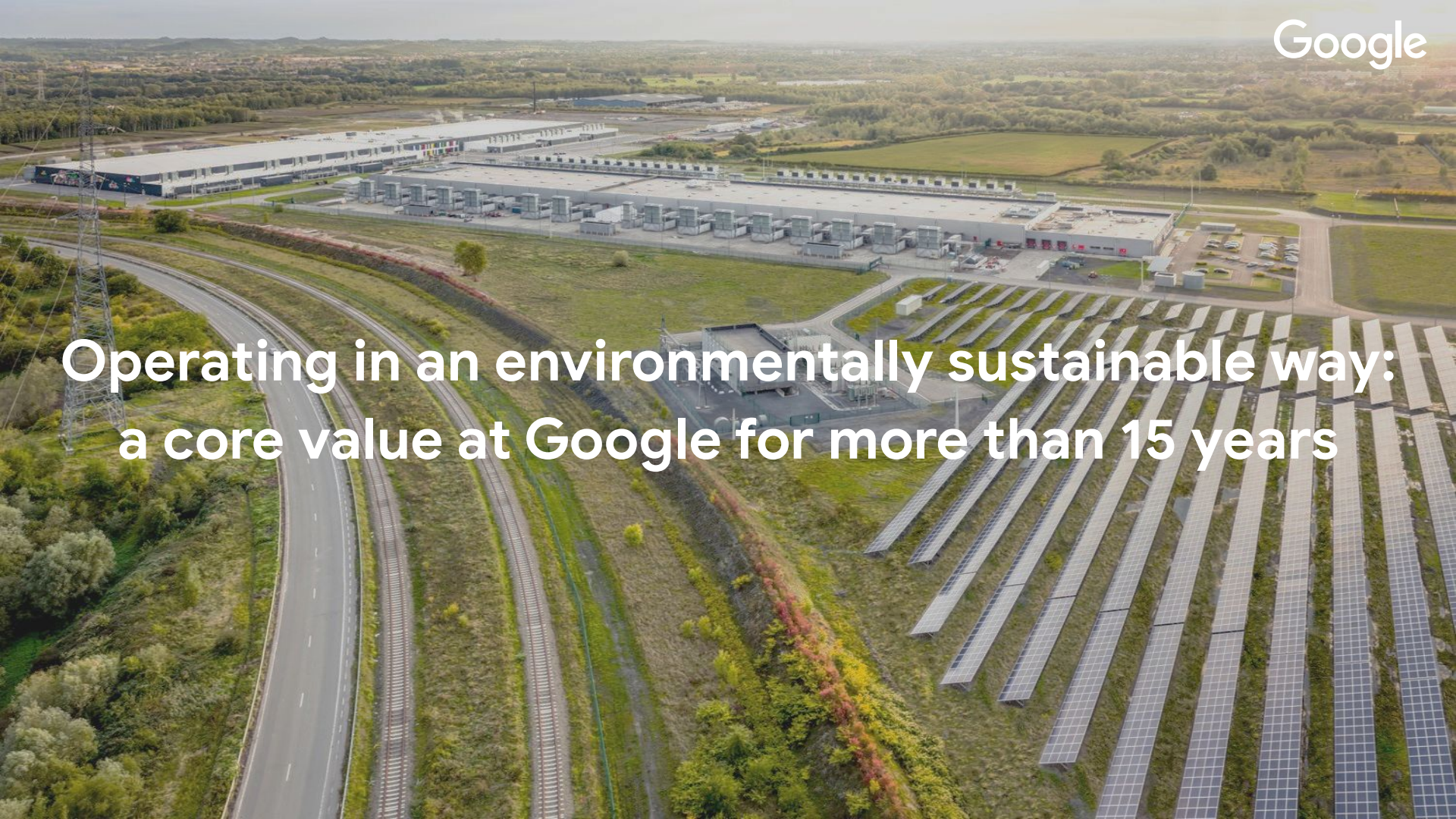

Google
Energy Usage of ML
Urs Hölzle
SVP, Technical Infrastructure
Presenting the work of **many** people at Google


Google
Operating in an environmentally sustainable way:
a core value at Google for more than 15 years

Milestones from our first two decades of climate action

Carbon Neutrality

(offsetting emissions)

2007

We became the first major company to be **carbon neutral**.

100% Renewable Energy

(reducing emissions)

2017

We reached 10 consecutive years of carbon neutrality and we became the first major company to match **100%** of our annual electricity use with **renewable energy**.

Google

Our third decade of climate action: Realizing a carbon-free future

Sep 14, 2020 · 4 min read

Sundar Pichai
CEO of Google and Alphabet

Eliminating our carbon legacy

As of today, we have eliminated Google's entire carbon legacy (covering all our operational emissions before we became carbon neutral in 2007) through the purchase of high-quality carbon offsets. This means that Google's lifetime net carbon footprint is now zero. We're pleased to be the first major company to get this done, today.

…excerpt…

Operating on carbon-free energy 24/7 by 2030

Since 2017 we've been matching all of our annual electricity consumption with 100 percent renewable energy. Now we're going even further: By 2030 Google is aiming to run our business on carbon-free energy everywhere, at all times.

This is our biggest sustainability moonshot yet, with enormous practical and technical complexity. We are the first major company that's set out to do this, and we aim to be the first to achieve it.

https://blog.google/outreach-initiatives/sustainability/our-third-decade-climate-action-realizing-carbon-free-future/

What's the difference between carbon-neutral, 100% renewable energy, and 24/7 carbon-free energy?

Carbon-Neutral

offsets emissions

achieved by purchasing carbon offsets that reduce or prevent global emissions

100% Renewable

reduces emissions

achieved by purchasing enough renewable energy to match annual electricity use

24/7 Carbon-Free

eliminates emissions

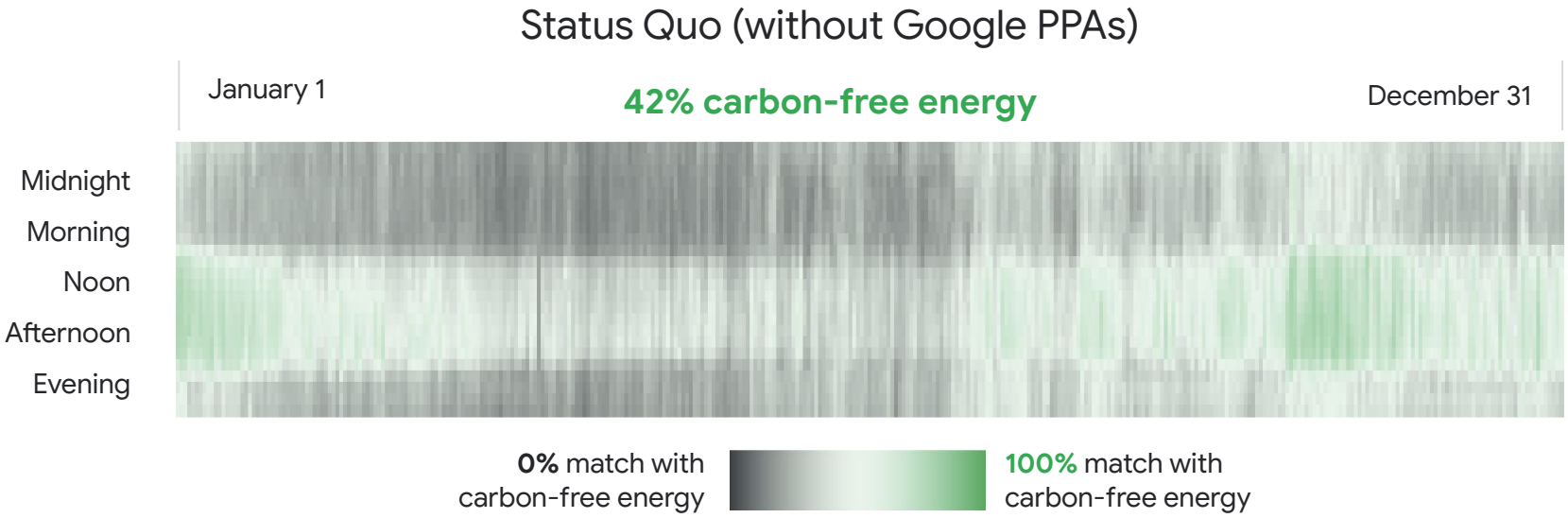
achieved by sourcing clean energy for every location and every hour of operations

2019: 61%
2020: 67%
2030 goal: 100%


Carbon-free energy
Google

Scenario: every hour of electricity use at Quilicura, Chile data center

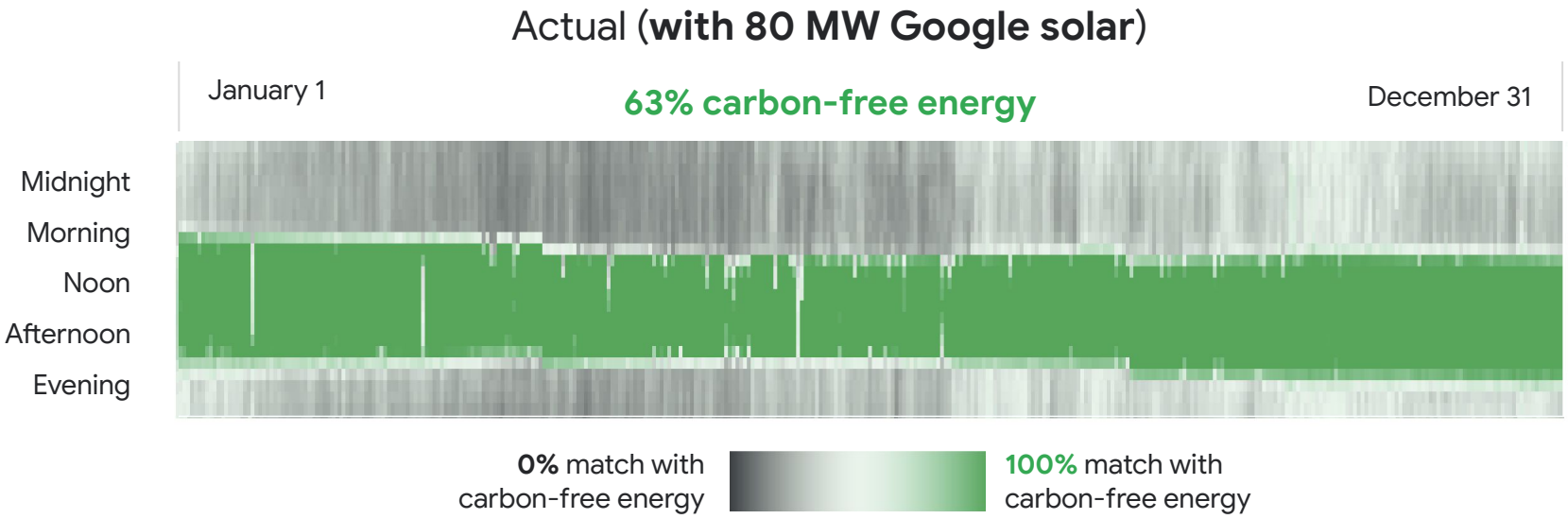
Without solar and wind Power Purchasing Agreements (PPAs), less than half our energy use in Chile would be matched with carbon-free sources on an hourly basis



Google

Actual: every hour of electricity use at Quilicura, Chile data center

Google's **first solar PPA** in Chile significantly increased our data center's carbon-free matching

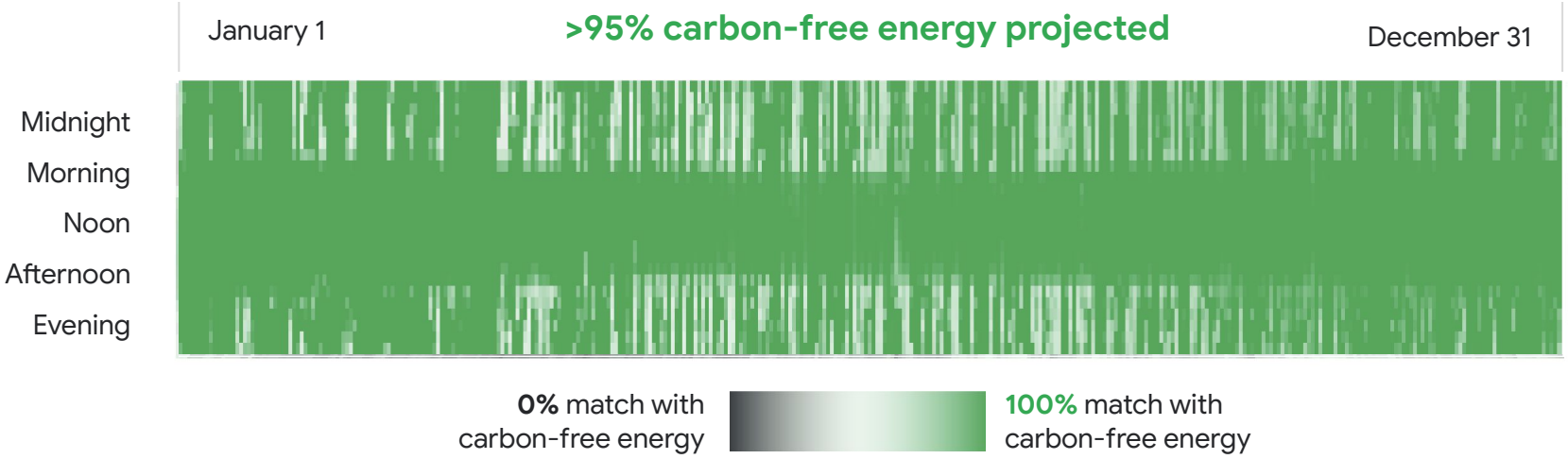


Google

Projected: every hour of electricity use at Quilicura, Chile data center

A **new solar + new wind PPA** will fill in the gaps, enabling us to match almost 100% of our electricity use with carbon-free resources on an hourly basis

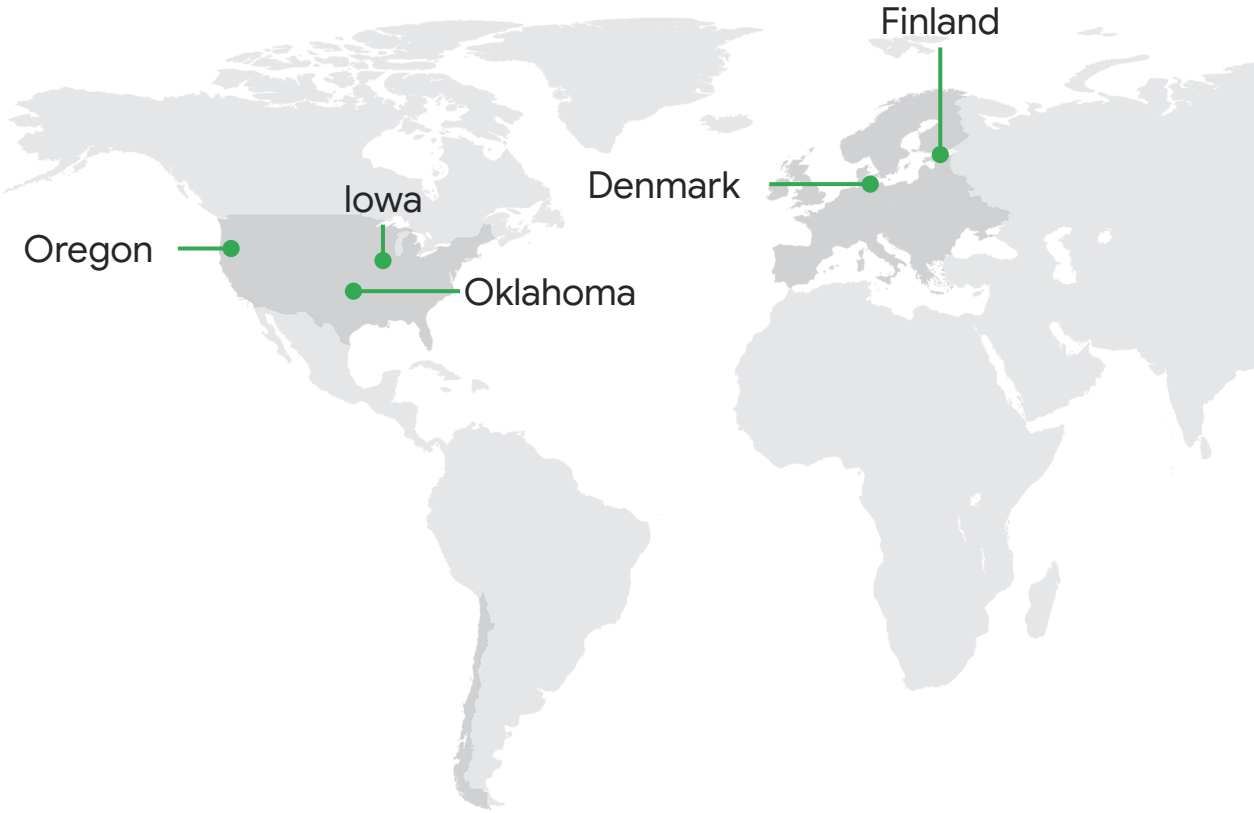
Projected for 2022 (with 80 MW Google solar + **new 35 MW solar** + **new 90 MW wind**)



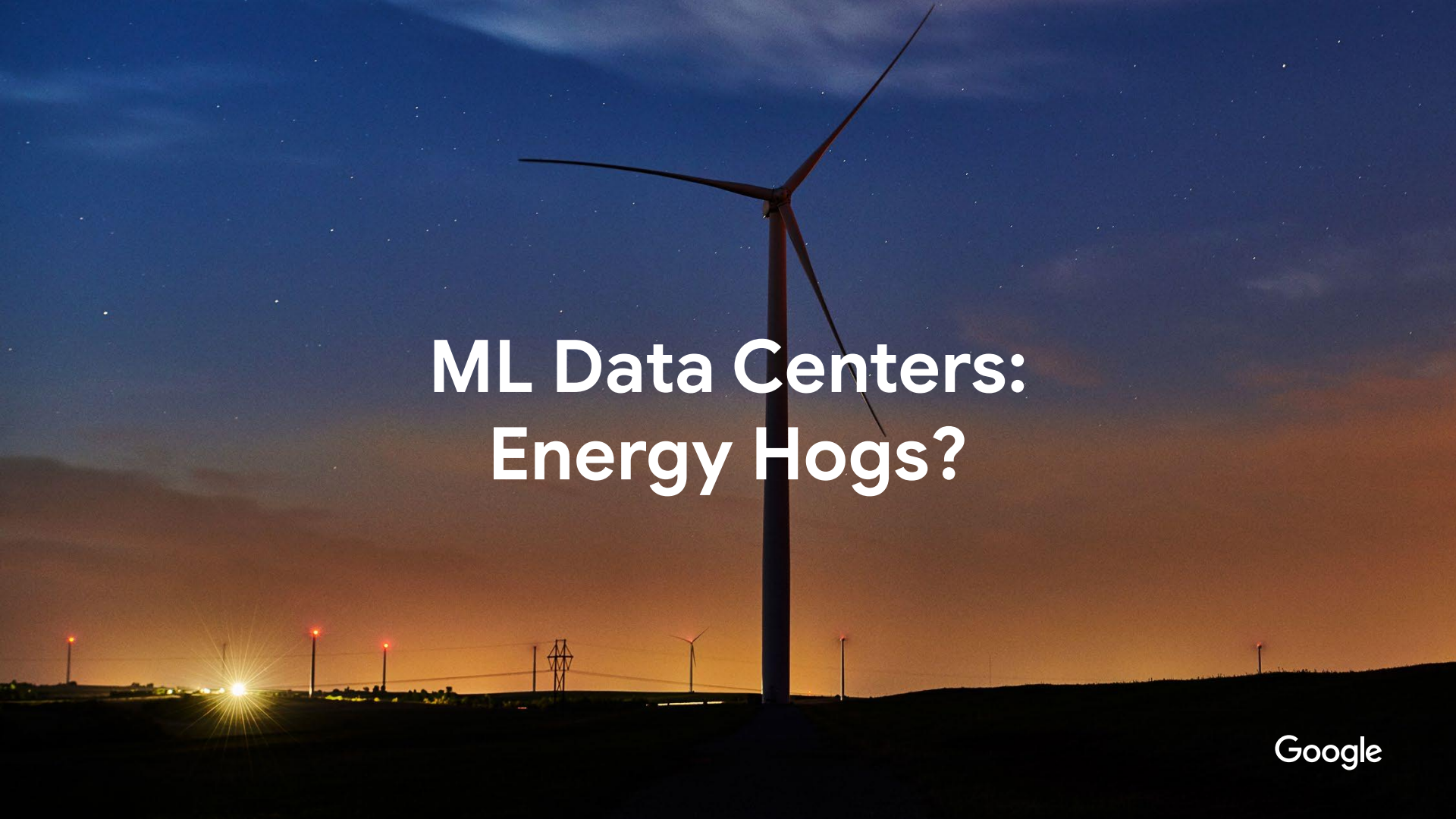
Google

5

data centers now operate near or at 90% carbon-free energy, as of end of 2021



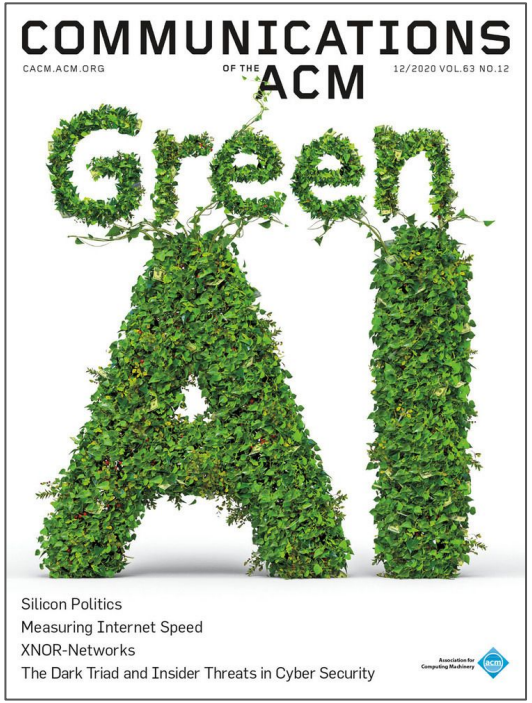
Google


ML Data Centers:
Energy Hogs?
Google


1.3 km
Google

ML Emissions

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, ACL 2019, 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).

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

* The ML task is image classification using the Imagenet benchmark to reduce the error rate to 5% from 11.5% when article was written.

Google

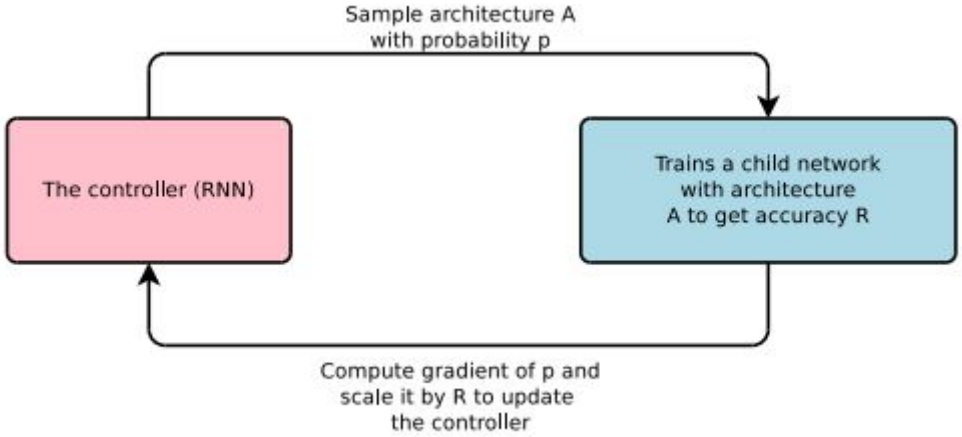
These predictions are based on inaccurate estimates and interpretations

Two key misunderstandings:

Neural Architecture Search (NAS)

Assuming Static Technology

Neural Architecture Search



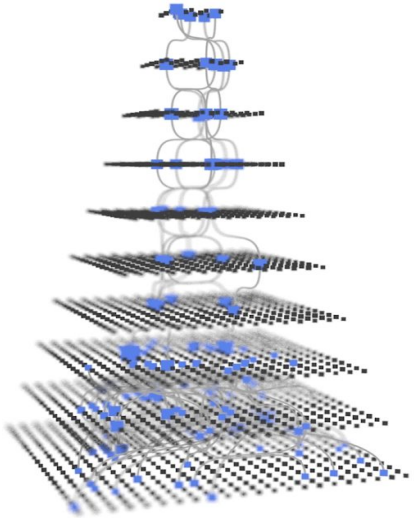
Idea: model-generating model trained via reinforcement learning

(1) Generate ten models
(2) Train them for a few hours
(3) Use loss of the generated models as reinforcement learning signal

Neural Architecture Search with Reinforcement Learning, Zoph & Le, ICLR 2016
arxiv.org/abs/1611.01578

Neural Architecture Search to find a model architecture

Controller: proposes ML model architectures

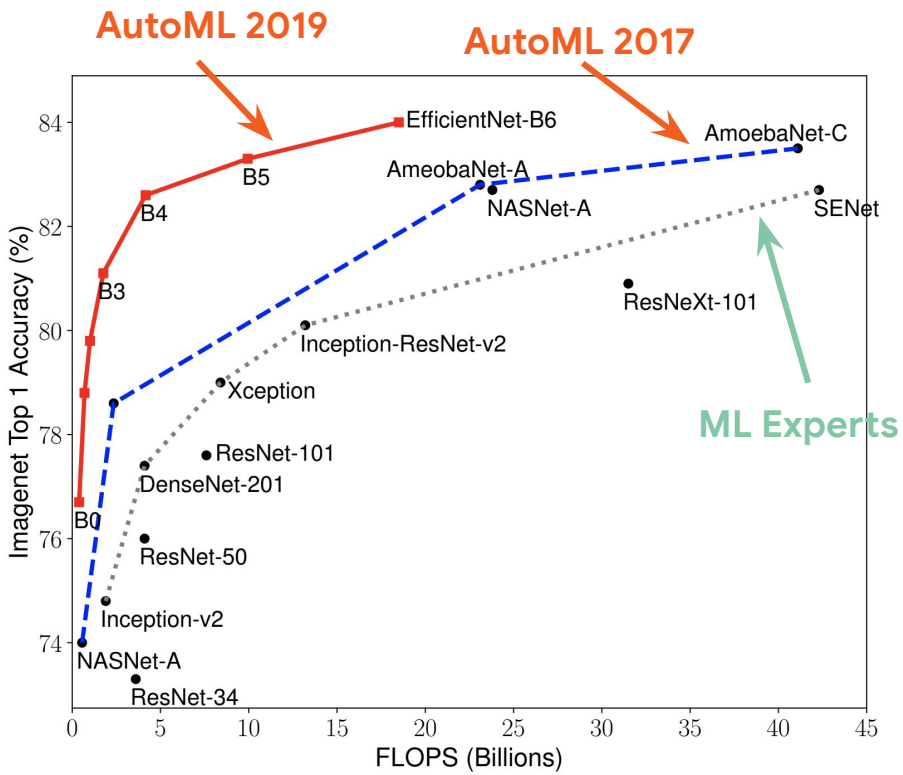


20K times

Iterate to find the most accurate model

Train & evaluate models

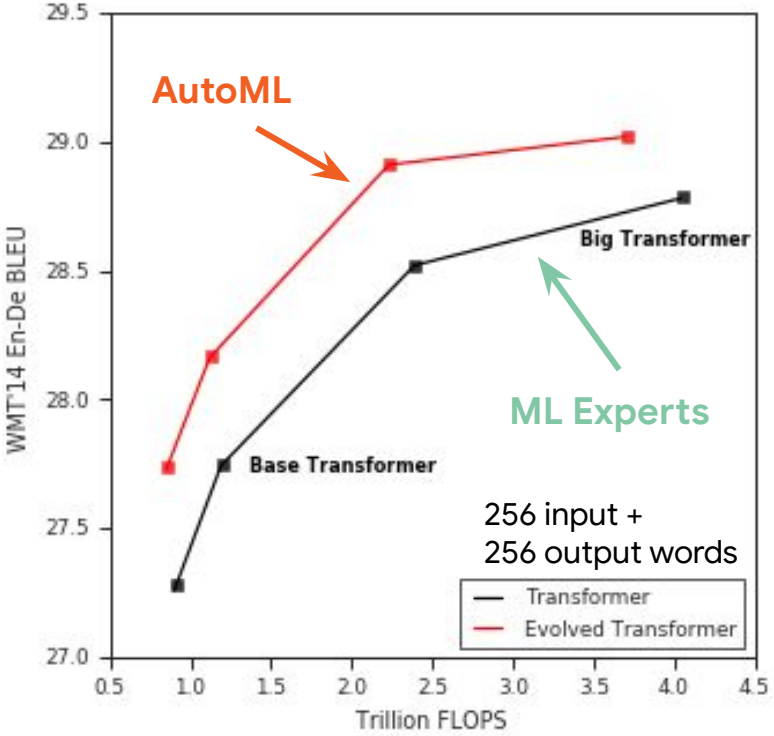
Image Recognition



Tan *et al.* EfficientNet: Rethinking Model Scaling for Deep Convolutional Neural Networks, ICML 2019, arxiv.org/abs/1905.11946

Google

Language Translation



So *et al*. The Evolved Transformer, 2019, arxiv.org/abs/1901.11117

Google

Neural Architecture Search And Efficiency/CO_2e Emissions Concerns

Misconception #1: Neural architecture search is done on every problem, rather than a one-time cost per problem-domain/search space

Discovered model architectures often open-sourced. e.g.:

github.com/tensorflow/tpu/blob/master/models/official/efficientnet/efficientnet_model.py

github.com/tensorflow/tensor2tensor/blob/master/tensor2tensor/models/evolved_transformer.py

github.com/google-research/google-research/tree/master/primer

Reused thousands of times for different problems

More efficient models lead to overall energy savings and lower CO_2e emissions

Neural Architecture Search And Efficiency/CO_2e Emissions Concerns

Misconception #2: Neural Architecture Search is done on full-sized problems, when in fact search is done using much smaller proxy tasks

Proxy tasks make the search itself much more efficient

MIT Technology Review

Training a single AI model can emit as much carbon as five cars in their lifetimes

Deep learning has a terrible carbon footprint.

NewScientist

Creating an AI can be five times worse for the planet than a car

The **one-time** Evolved Transformer NAS search on TPU v2 hardware in a Google datacenter in Georgia generated 3.2t of CO_2e, not 284t of CO_2e

→ ~88X less CO2e than estimated

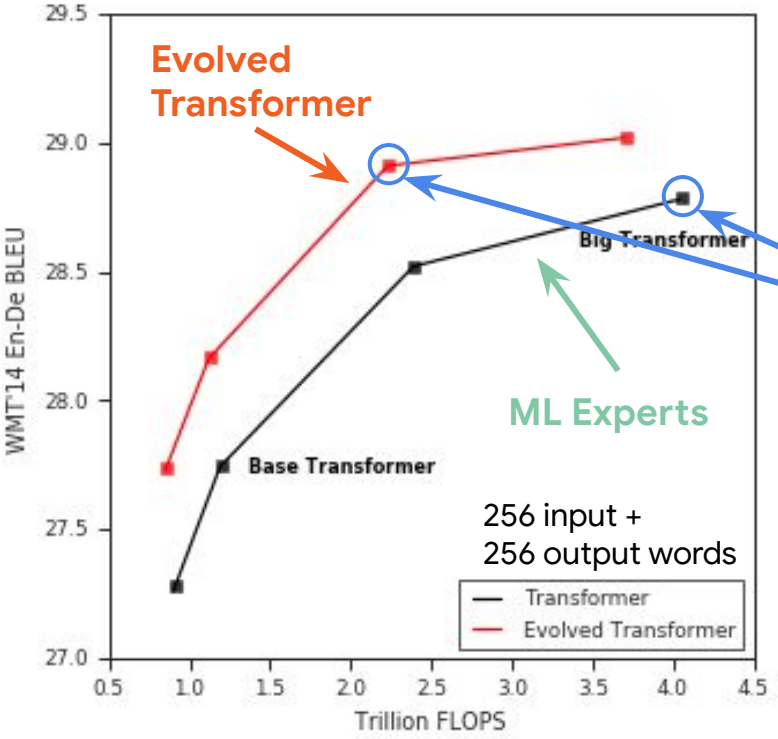
(1) Modeled P100 vs TPU v2, and US averages vs Google DC: **actual NAS was 5X lower**

(2) Assumed use of full model vs small proxy task for search (as described by So *et al.*): **actual NAS was 19X less compute/emissions**

"Five car lifetimes" → 0.00004 car lifetimes (**120,000x less**)

Better Models, Across Multiple Modalities/Domains

Language Translation



Many fewer FLOPs (and less energy) to reach same or higher accuracy

e.g. In Google Iowa datacenter:

On P100 GPUs: **185 KWh vs. 221 KWh (-16%)**
On TPU v2: **30 KWh vs 40 KWh (-25%)**

So *et al*. The Evolved Transformer, 2019, arxiv.org/abs/1901.11117

Google

Even Better: Rapid Efficiency Improvements Every Year

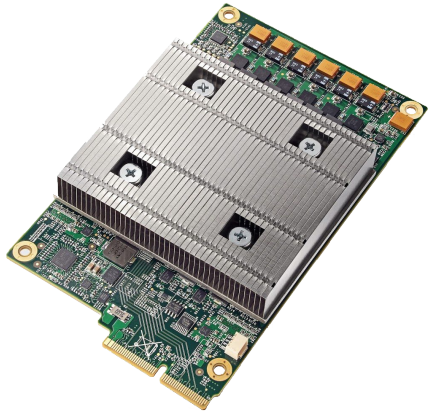
Research: Transformer (2017) ⟶ Evolved Transformer (2019) ⟶ Primer (2021): 4.2x faster at same accuracy level

Hardware: Energy per performance has been improving rapidly

- Architectures fully optimized for ML (not general-purpose GPUs)
- Optimized communication and memory access (**FLOPs often aren't primary energy consumption driver**!)
- Better mapping of models to available hardware
- Specialized pods for large-model training

TPUv1: Google's first Tensor Processing Unit (TPU)

Google-designed chip for neural net **inference**

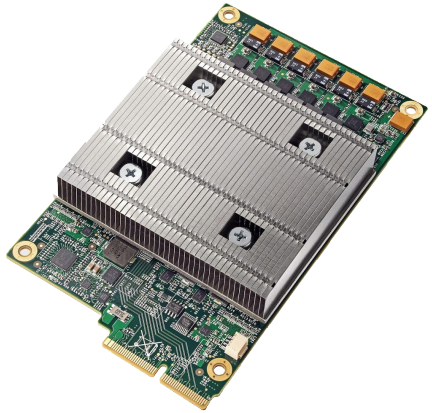


In production use since 2015: used on search queries, for neural machine translation, for speech, for image recognition, for AlphaGo match, ...

In-Datacenter Performance Analysis of a Tensor Processing Unit, Jouppi, Young, Patil, Patterson *et al.*, ISCA 2017, arxiv.org/abs/1704.04760

TPUv1: Google's first Tensor Processing Unit (TPU)

Google-designed chip for neural net **inference**

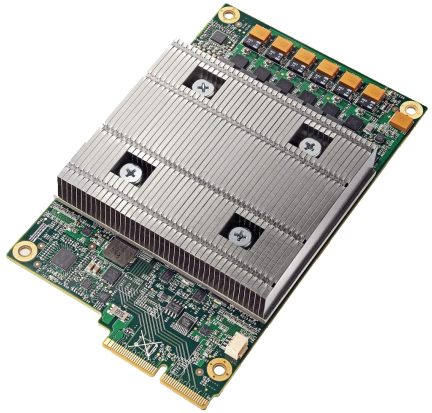


In production use since 2015: used on search queries, for neural machine translation, for speech, for image recognition, for AlphaGo match, ...

In-Datacenter Performance Analysis of a Tensor Processing Unit, Jouppi, Young, Patil, Patterson *et al.*, ISCA 2017, arxiv.org/abs/1704.04760

TPUv1: Google's first Tensor Processing Unit (TPU)

Google-designed chip for neural net **inference**



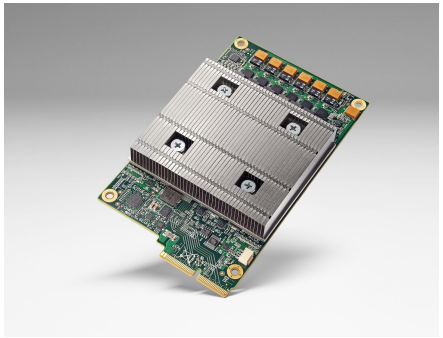
In production use since 2015: used on search queries, for neural machine translation, for speech, for image recognition, for AlphaGo match, ...

In-Datacenter Performance Analysis of a Tensor Processing Unit, Jouppi, Young, Patil, Patterson *et al.*, ISCA 2017, arxiv.org/abs/1704.04760

TPU Chip Family

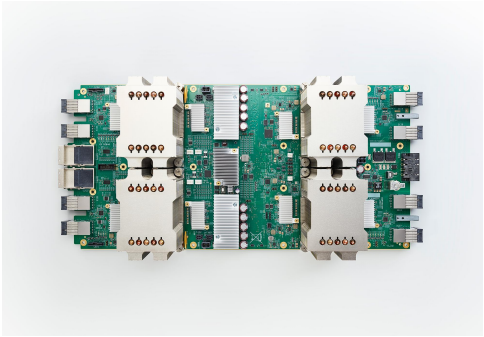
g.co/cloudtpu

TPU v1 (2015)
92 teraops
(inference only)



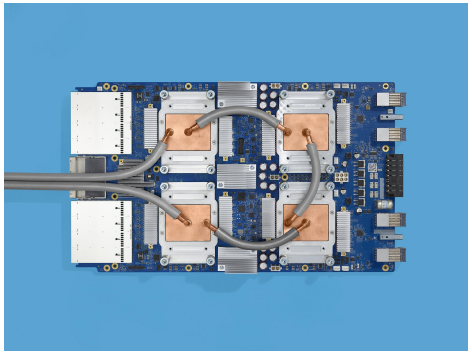
In-Datacenter Performance Analysis of a Tensor Processing Unit, Jouppi *et al.*, ISCA 2017

TPU v2 (2017)
45 teraflops
/ chip

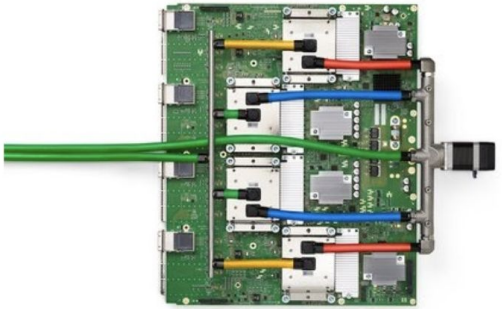


A Domain-Specific Supercomputer for Training Deep Neural Networks, Jouppi *et al.*, CACM 2020

TPU v3 (2018)
105 teraflops
/ chip



TPU v4 (2020)
275 teraflops
/ chip



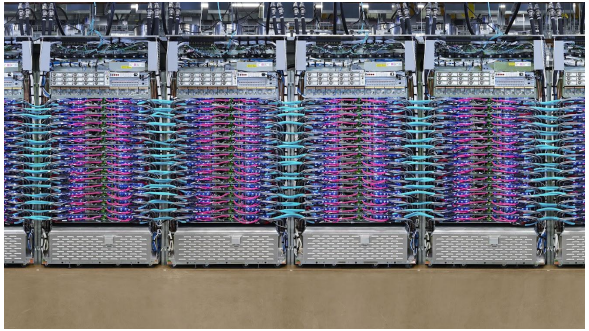
Ten lessons from Three Generations Shaped Google's TPUv4i, Jouppi *et al.*, ISCA 2021.

Google

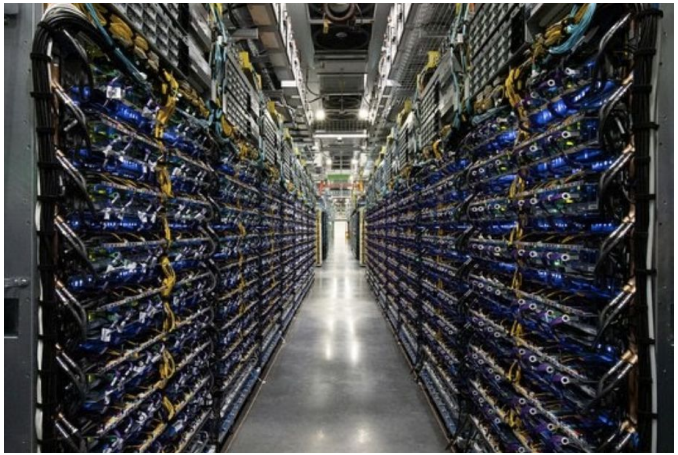
TPU Pods

TPU v2 Pod (2017)
11.5 petaflops, 256 chips,
2-D toroidal mesh network

g.co/cloudtpu



TPU v3 Pod (2018)
105 petaflops, 1024 chips,
liquid cooled



TPU v4 Pod (2020)
1.1 exaflops, 4096 chips,
liquid cooled

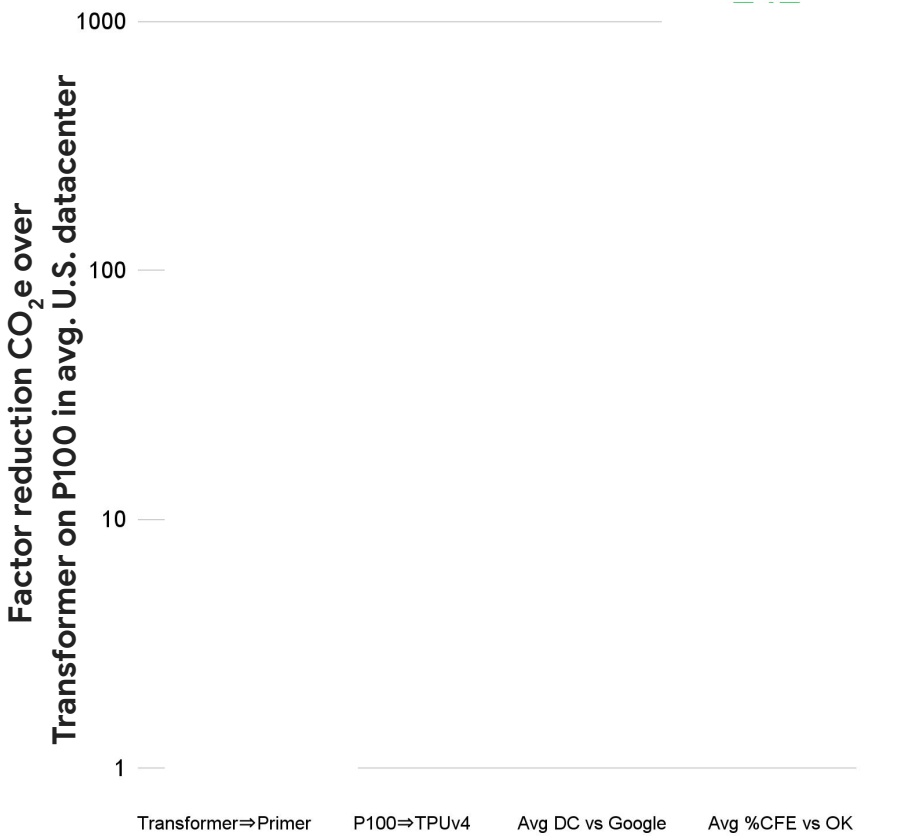
Google

Result: Reduce energy ~100X, CO_2e by 747x!

Four (multiplicative) best practices

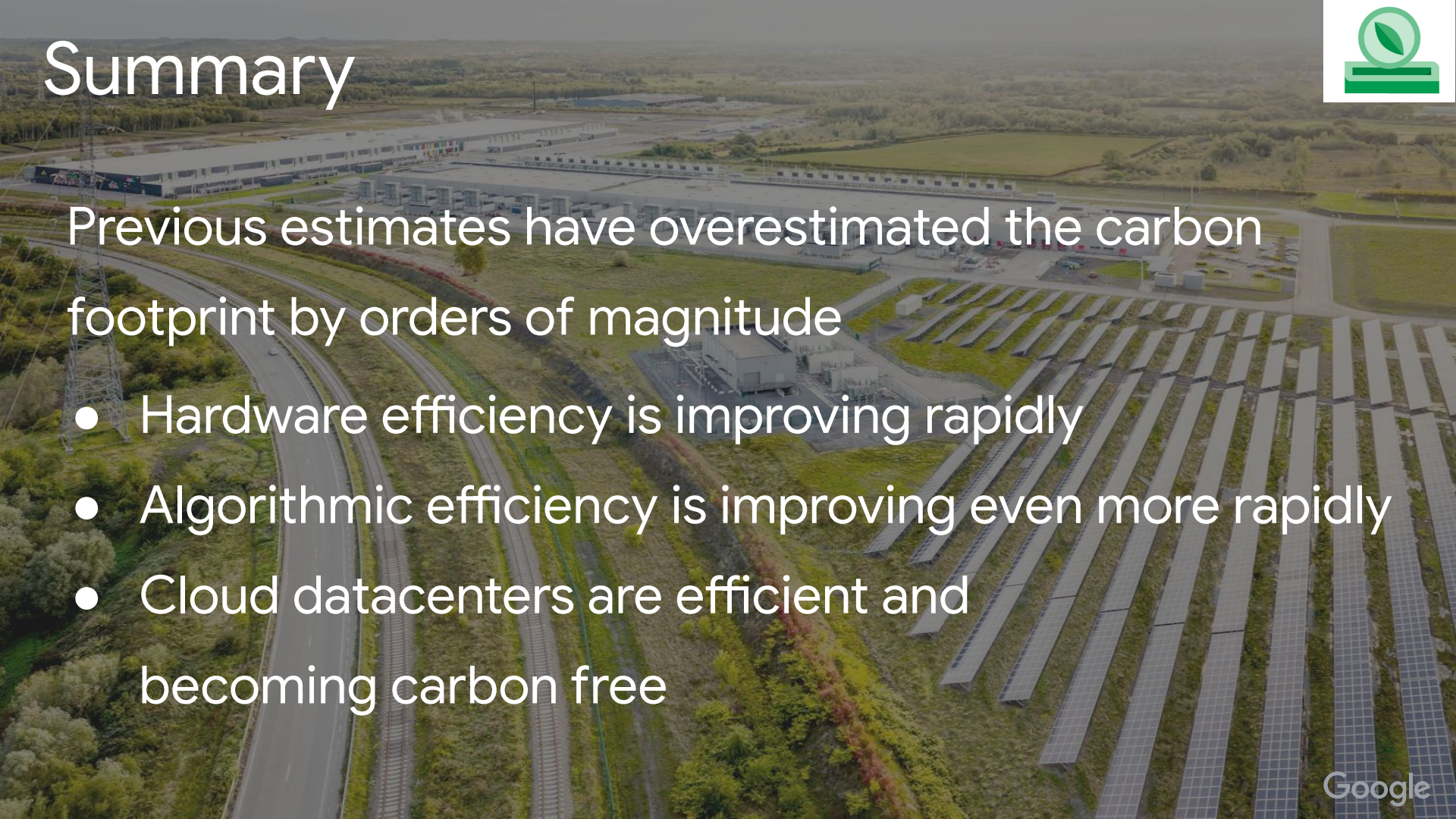
"4Ms of ML Energy Efficiency"

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



Google

Summary



Previous estimates have overestimated the carbon footprint by orders of magnitude

- Hardware efficiency is improving rapidly
- Algorithmic efficiency is improving even more rapidly
- Cloud datacenters are efficient and becoming carbon free

Bending The Curve Requires All Of Us

- ***Cloud providers:*** publish efficiency, %CFE, and CO2e/MWh per location to enable informed location choice
- ***ML practitioners:*** train using the most effective processors
- ***ML researchers:*** continue to develop more efficient ML models and approaches; publish energy consumption and carbon footprint

If we all ML follow best practices, we will create a **virtuous circle** that will bend the curve to flatten and eventually shrink CO_2e

Details: IEEE Computer Article

The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink

David Patterson[1,2], Joseph Gonzalez[2], Urs Hölzle[1], Quoc Le[1], Chen Liang[1], Lluis-Miquel Munguia[1], Daniel Rothchild[2], David So[1], Maud Texier[1], and Jeff Dean[1]

Abstract: *Machine Learning (ML) workloads have rapidly grown in importance, but raised concerns about their carbon footprint. Four best practices can reduce ML training energy by up to 100x and CO_2 emissions up to 1000x. By following best practices, overall ML energy use (across research, development, and production) held steady at <15% of Google's total energy use for the past three years. If the whole ML field were to adopt best practices, total carbon emissions from training would reduce. Hence, we recommend that ML papers include emissions explicitly to foster competition on more than just model quality. Estimates of emissions in papers that omitted them have been off 100x–100,000x, so publishing emissions has the added benefit of ensuring accurate accounting. Given the importance of climate change, we must get the numbers right to make certain that we work on its biggest challenges.*

1. Introduction

Over the past few years, a growing number of papers have highlighted the carbon emissions of machine learning (ML) workloads. While this work has been instrumental in rightfully elevating the discussion around carbon emissions in ML, some studies significantly overestimated actual emissions, which in turn led to worrisome extrapolations [1,2]:

Thank you!

Google