Imperial College London

The Hidden Environmental Cost of Machine Learning
Is it worth it?

Alexander Montgomerie, Diederik Vink, Aditya Rajagopal

intelligent Digital Systems Lab
Dept. of Electrical and Electronic Engineering

www.imperial.ac.uk/idsl

Examples of Machine Learning

Machine Learning is ubiquitous in everyday life

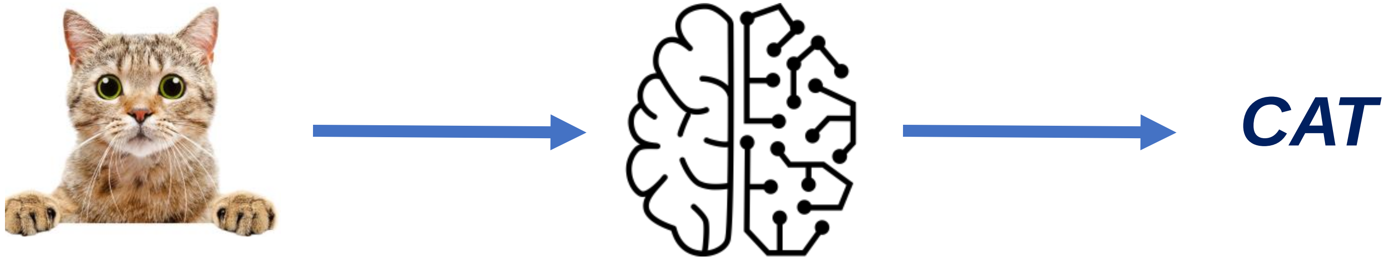
- **Home Assistants:**



- **Translation:**
- **Recommendation:**
- **Object Detection:**

Overview of Machine Learning

A blackbox model which can perform a given task



Training:

- Using a set of example real-world inputs to *learn* the parameters for the blackbox model

Inference:

- Running this blackbox model with the learned parameters on new real-world inputs

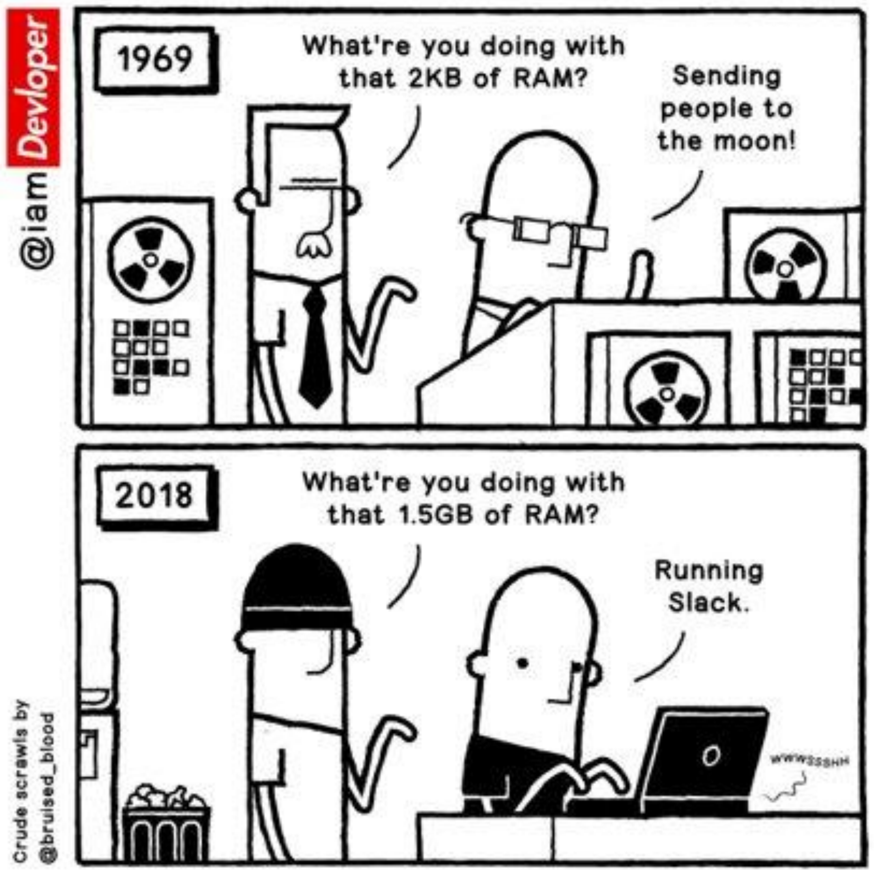
Imperial College London

Evolution of ML model sizes

What does it take to perform training and inference?

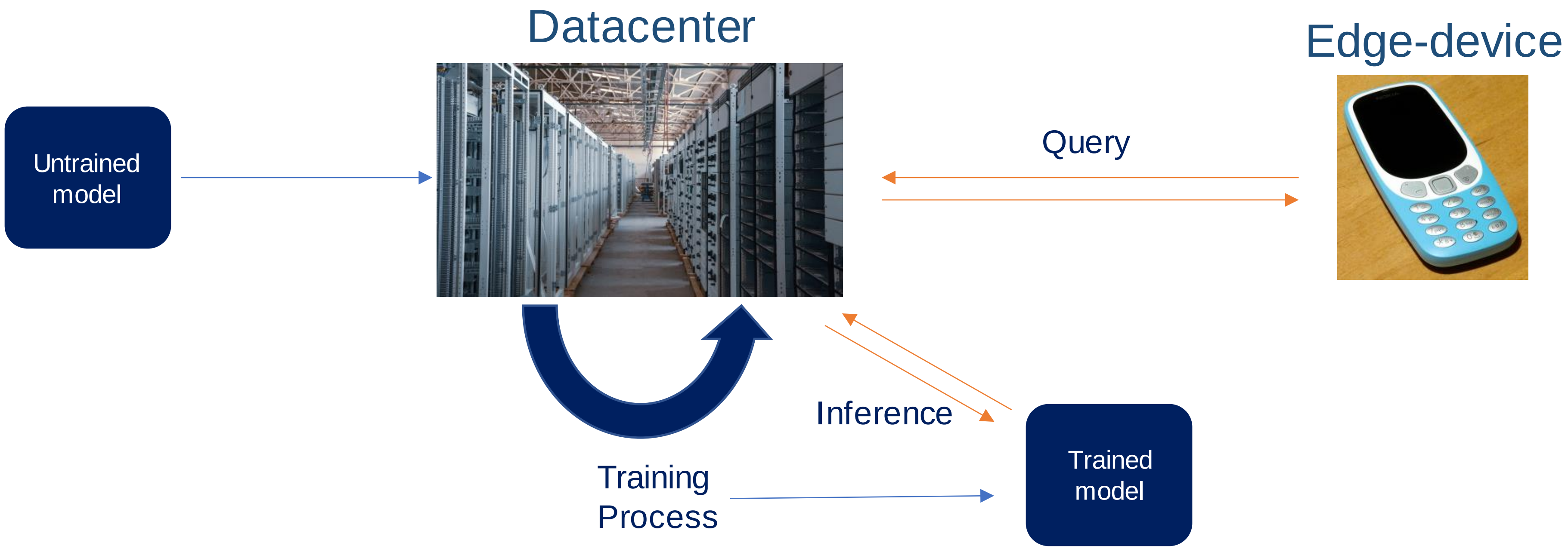
Architecture	Year	Application	# Parameters (million)
AlexNet	2012	CV	60
Seq2seq	2014	NLP	320
GoogLeNet	2014	CV	64
Transformer	2017	NLP	213
AmoebaNet	2019	CV	557
GPT-3	2020	NLP	175000

- CV = Computer Vision
- NLP = Natural Language Processing



ML Model Lifecycle

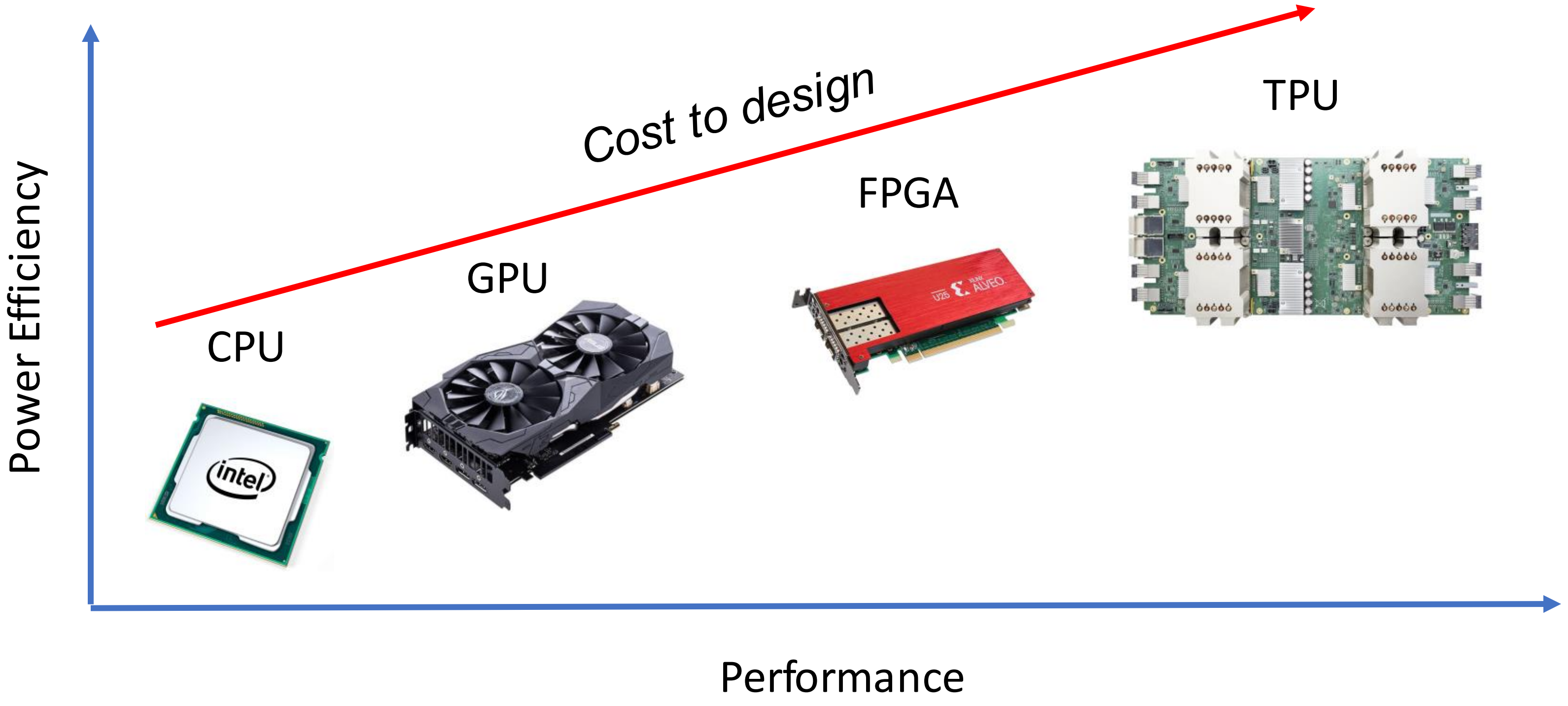
How is a machine learning model deployed?



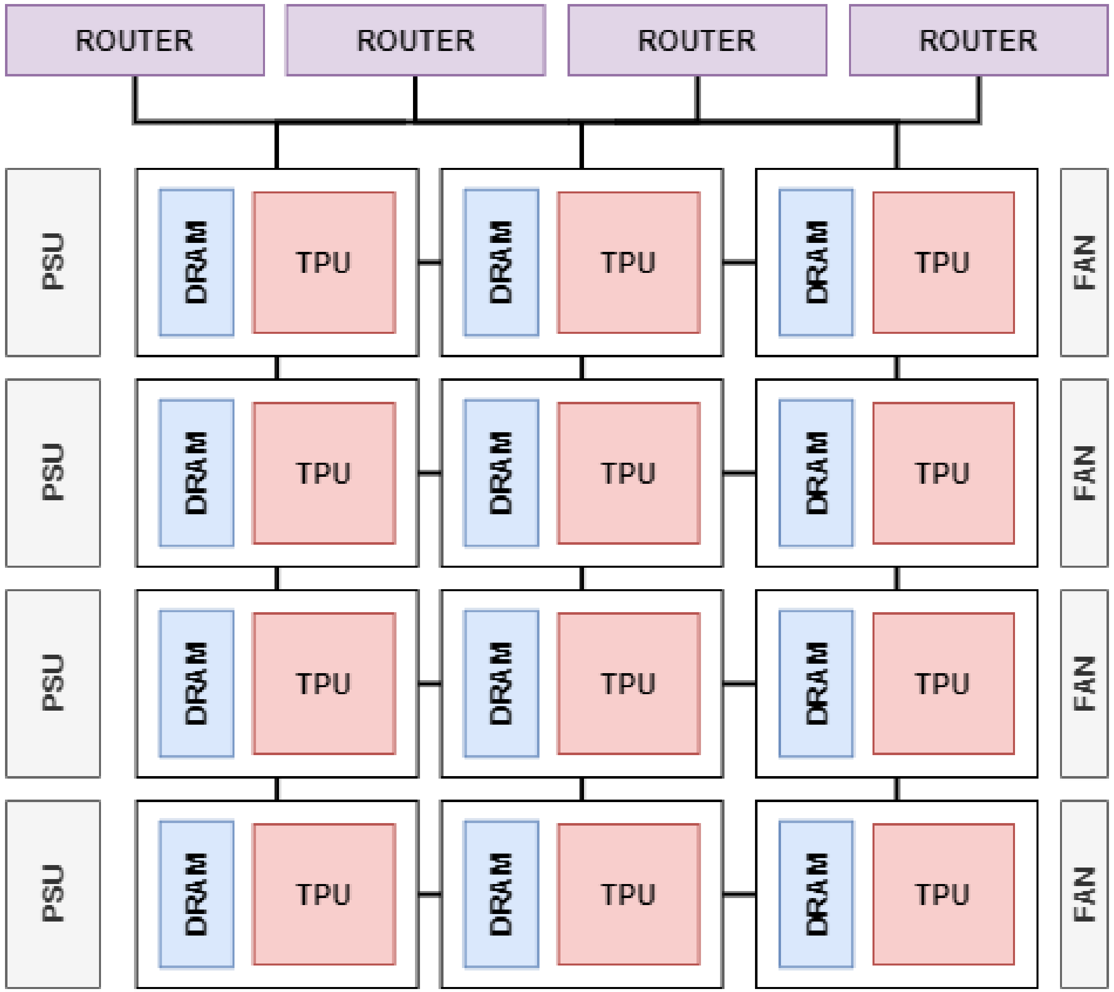
Hardware for Machine Learning

What are the desirable properties for hardware to run machine learning workloads?

- Handle matrix multiplication efficiently
- Ability to store millions of parameters



Hardware for Machine Learning - Datacenter



Facility Components:
- PSU
- Heating Control

Communication:
- Routers
- Switches

Computing nodes:
- ML Accelerator
- Memory

Quantifying the Carbon Footprint – Power Usage Effectiveness

How can we summarise the power consumed by a datacenter?

$$Power\ Usage\ Effectiveness\ (PUE) = \frac{IT\ Energy\ Consumption + Datacenter\ Overhead}{IT\ Energy\ Consumption}$$

- *IT Energy Consumption* = Power consumed by the machine learning hardware
- *Datacenter Overhead* = Power consumed by other parts of the datacenter (Heating and Cooling, Power supply inefficienies, etc …)

Describes efficiency of a datacenter

Quantifying the Carbon Footprint – Greenhouse Gasses

How can we relate energy consumption to green house gas emissions?

CO_2e : Carbon Dioxide Equivalent

- Measure of total greenhouse gas emissions from a process
- For datacenters, can describe the tonne of CO2 per MWh
- The US average value for CO2e is 0.71 tonnes of CO2 per MWh

Quantifying the Carbon Footprint – Total Footprint

What is the total carbon footprint for a Machine Learning Application?

$$(Training_{energy} + queries * Inference_{energy}) * PUE * CO_2e$$

- One-time cost of training
- Continued cost of inference throughout lifetime
- The datacenter charateristics are important (PUE and CO2e)

Data Center Environmental Impact

Supporting infrastructure has significant impact on data center construction

Imperial College London

- Supporting infrastructure originally made up 2/3 of floor space
- Centers host for users:
  - Computational servers
  - Data storage units
  - Network servers
  - Supporting infrastructure

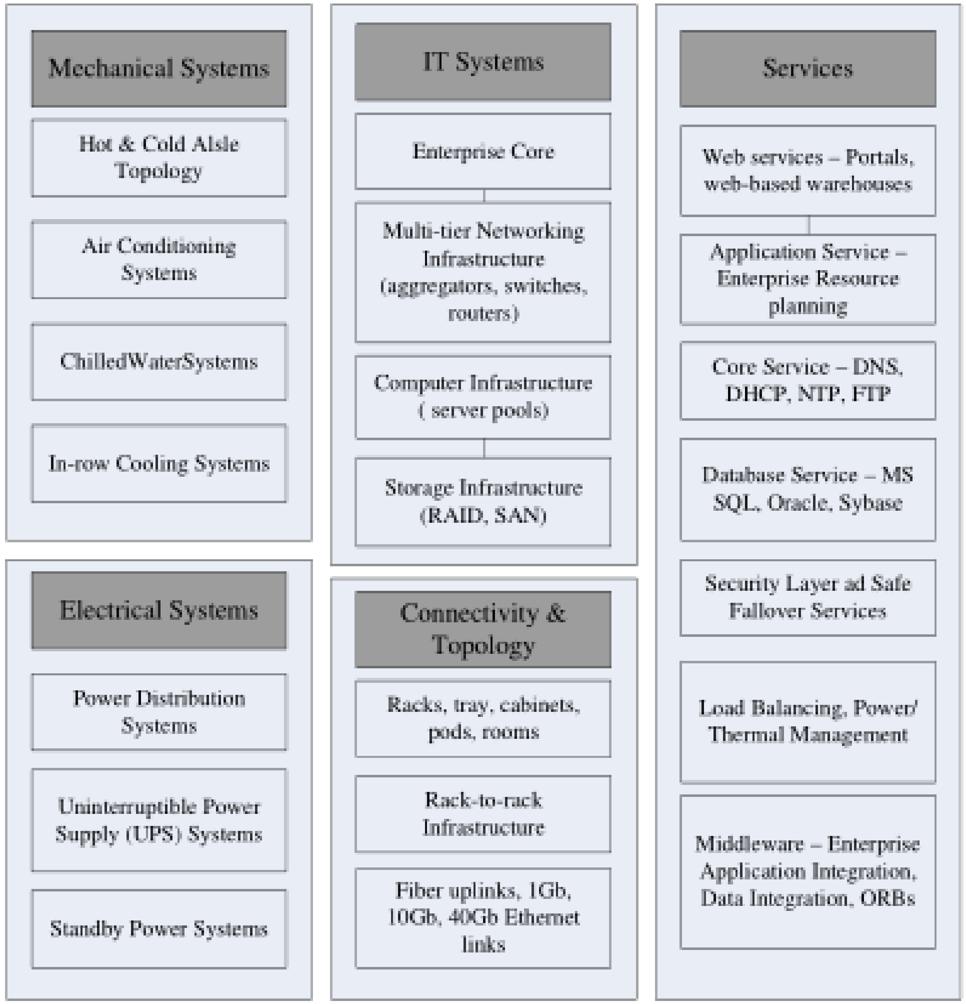


Figure: Data Center Construction Breakdown [17]

3 features differentiate hyperscale data centers

Hardware to support simplified power distribution [24]

- 48V instead of 12V motherboards to reduce "stepping down" voltage

Improved Virtualization [25]

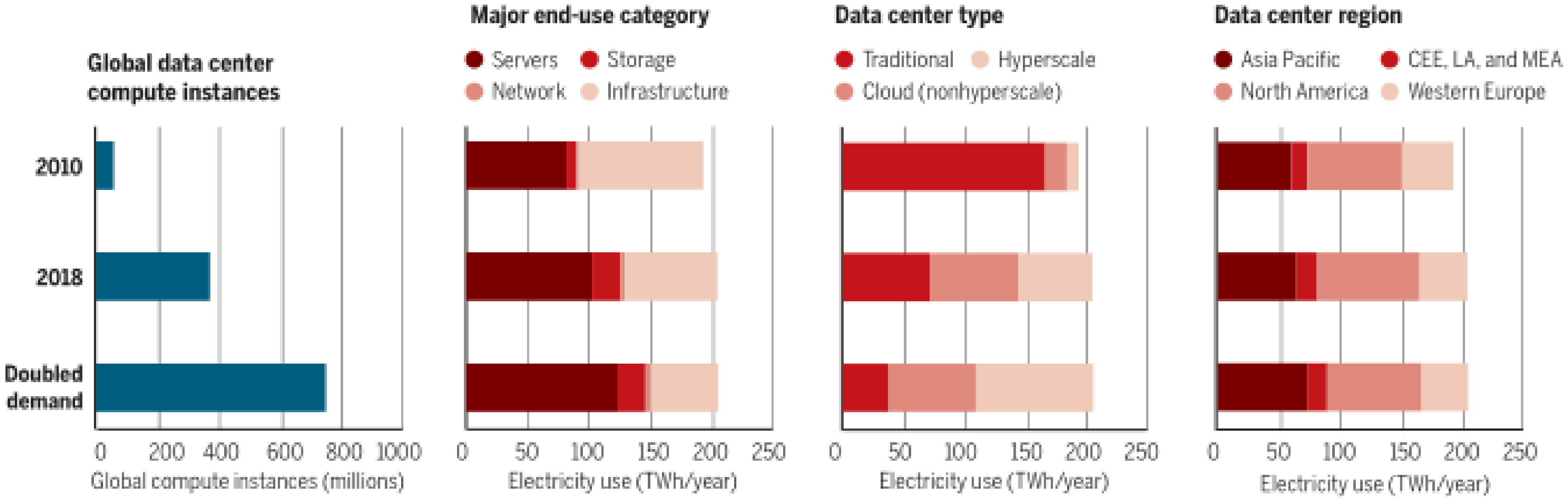
- Includes predictive scheduling
- Workload reshuffling
- Hardware reallocation

Advanced cooling systems [24]

- Kyoto Cooling (indirect air)
- Membrane-based evaporative cooling (Facebook)
- Water to the chip (Google)
- Rear-door chilling units (LinkedIn).

Imperial College London

Data Center utilization is growing faster than power consumption



Data Center Energy Breakdown in 2014 [16]

Increase in forecasted energy consumption driven by data centers and networks

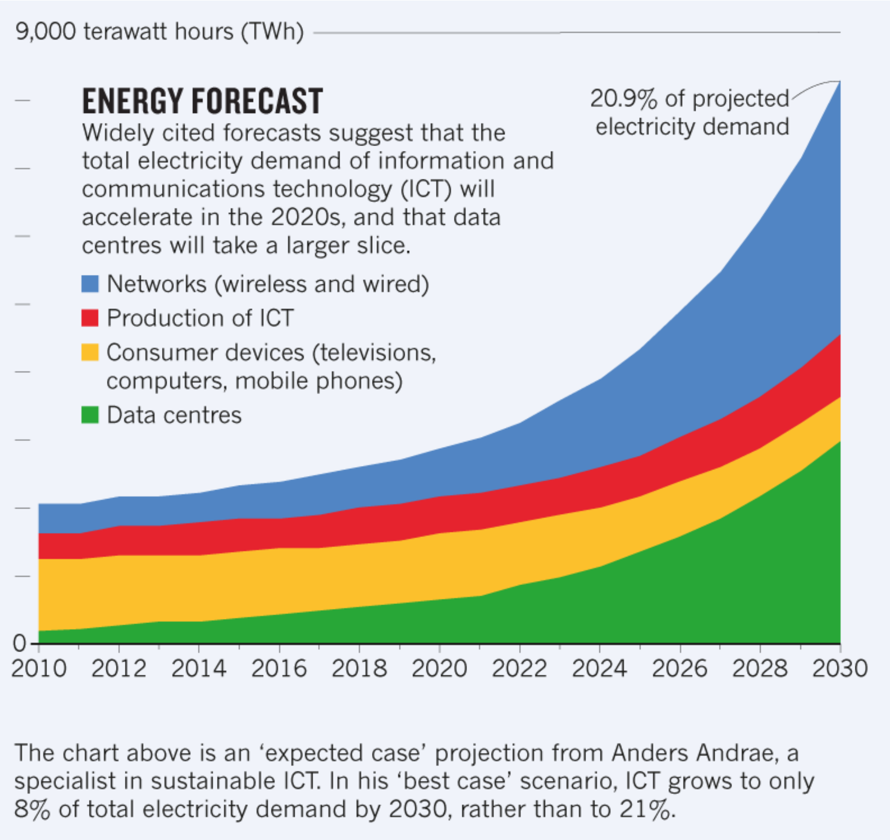


Figure: Data Center energy forecasts [19]

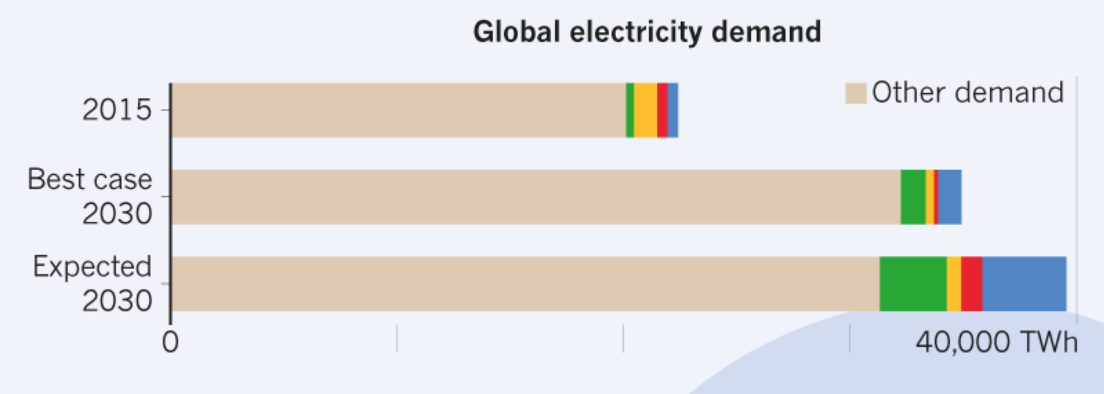
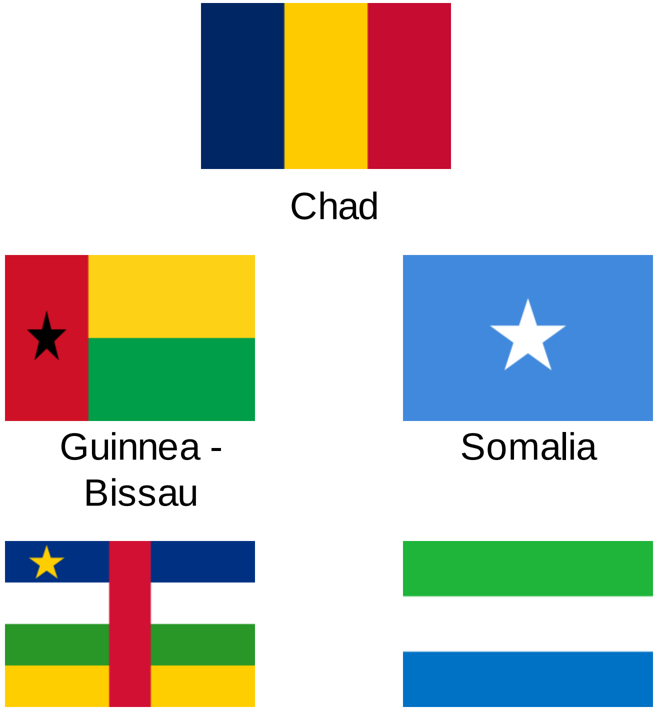


Figure: ICT Proportion of Global Electricity Demand [19]

Perspective | Relative energy consumption of data centers and the ICT industry



Chad

Guinnea - Bissau

Somalia

Central African Republic

Sierra Leone

Equivalent energy consumption of global music downloads [21]

1% Proportion of worldwide energy consumption attributable to ICT industry [23]

200 TWH Annual energy consumption of data centers and ICT industry [19]

20% Predicted share of global electricity share attributed to ICT power consumption (2030) [21]

Large relative environmental impact of data centers and ICT industry

Global emissions from ICT are equivalent to those of entire aviation industry

- Accounts for 2.3% of global greenhouse gas emissions [23]
- 25% directly from Data Centers [15]

Data centers:
3.15x10^7 tons of CO_2e emission [26]

Bitcoin:
5.8x10^7 tons of CO_2e emission [1]

>

2040 prediction: 14% of world emissions will be produced by storing digital data: [20]

- Same proportion as US today
- Data Centers are fastest growing emitter in ICT

Primary causes of data center inefficiencies

- Cloud computing has 20-40% CPU utilization [22]
  - Idle compute causes large inefficiency
- Warm climate based data centers [19]
- Power surging [19]

Primary approaches to reducing carbon footprint

- Systems used which include the type of hardware, cooling systems, layout etc.
- Location
- Carbon Offsetting
- Google specific – Georgia does not have a supply of Carbon Free Energy
  - Relocate the server to Oklahoma where Google can average 95.6% net CFE (Location)
  - Purchase the equivalent MWh of CFE in Montana (Carbon Offsets)

The Carbon Footprint of Training

Carbon Footprint of Training - Components

- Network Architecture Search (NAS)
  - Generate model architecture by framing search space of possible network architectures as a learning problem and optimizing for target metric
- Prototyping
  - Hyperparameter optimization
- Final Training Run
  - Single training run resulting in final model

Carbon Footprint of Training – Final Training Run

- Most commonly measured and reported

Net CO_2e			TPUv3		V100 GPU
	Meena	T5	Gshard-600B	Switch Transformer	GPT-3
Metric Tons	96	47	59	4	552
SF-NY Roundtrips	0.53	0.26	0.33	0.022	3.07

Carbon Footprint of Training – NAS

- Commonly presumed to be energy inefficient
  - Using NAS to develop the Evolved Transformer [5] resulted in a model with 37% fewer parameters and 25% less energy than a vanilla Transformer [1]
  - These transformers were used to train the Meena DNN [6] and the energy savings obtained from using a NAS developed model was approx. 15x larger than the cost of NAS [1]

Carbon Footprint of Training – Prototyping

- Currently very difficult to measure or estimate

- Power consumption varies with
  - Device used
  - Time of day
  - Location of server
  - Architecture of network
  - Size of dataset

Carbon Footprint of Training – Measuring

- Current approaches to estimate include:
  - Peak performance per Watt
    - Peak is higher than measured by on average 1.6x for TPUs and 3.5x for GPUs
  - Modelling
    - ML Emissions [7] and Green Algorithms [8] differ from the measured energy consumption by on average 0.92x and 1.48x respectively
- New approaches [9],[10] aim to facilitate easy measurement instead of estimation so that accurate reporting of carbon footprint to training and deploying a neural network can be performed

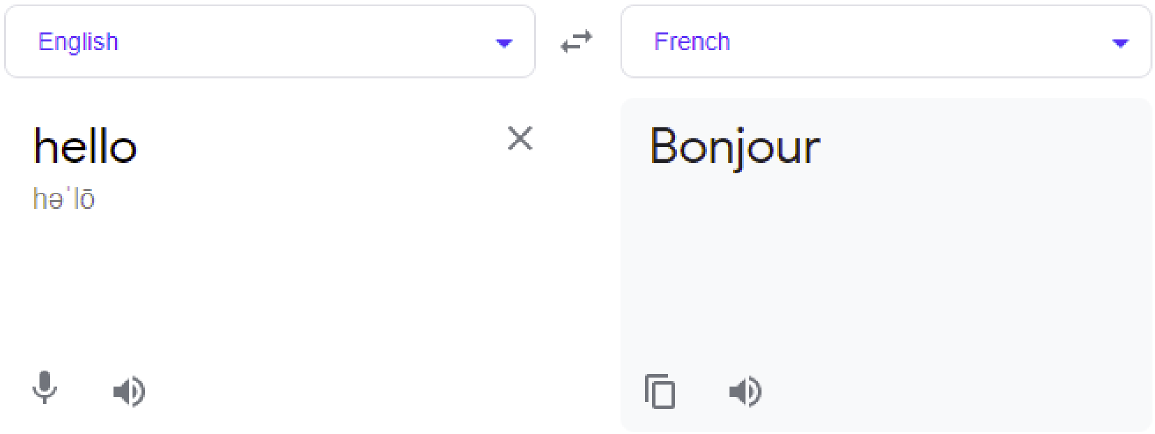
Environmental Impact of Inference

Environmental Impact of Inference

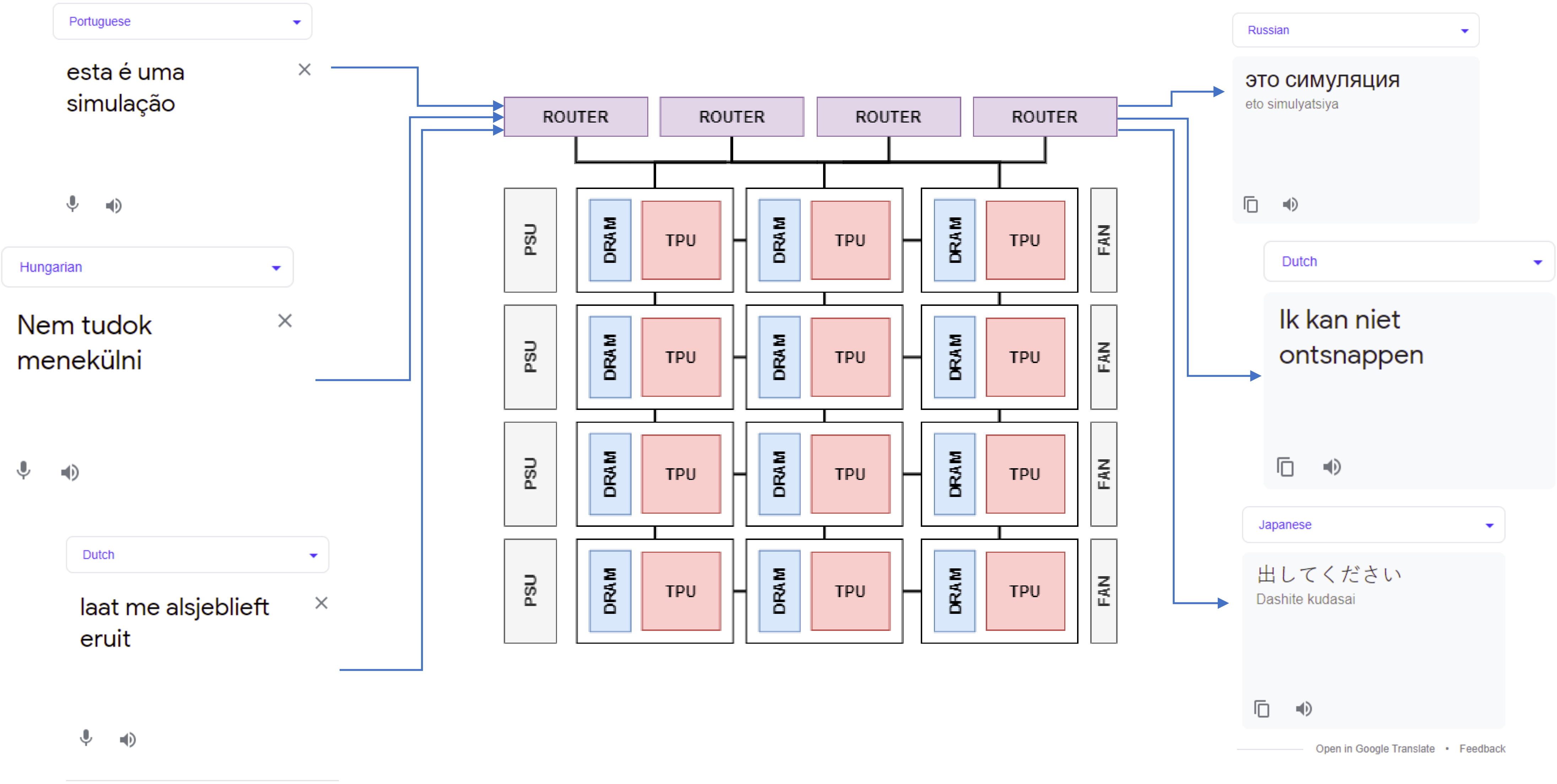
- What is inference?
  - When the ML model is being run to perform the given task
- Why is the impact of inference important?
  - Amazon estimates that 90% of the ML cloud compute is used for inference [1]
  - ML will become even more prevalent in the future
- How significant is the impact of inference on the environment?

Case Study – Google Translate

- Employs the **GNMT** machine learning model
- Used by roughly **500 million people** worldwide [11]
- Handles around **100 billion words a day**
- **TPU accelerators** are used in Google Datacenters for processing ML workloads



Case Study – Google Translate



Imperial College London

Inference Power Model

How can we estimate the power consumption of Google Translate?

$$P_{total} = N_{TPU} \cdot (\frac{T_{requests}}{T_{TPU} \cdot N_{TPU}} \cdot P_{TPU}^{busy} + (1 - \frac{T_{requests}}{T_{TPU} \cdot N_{TPU}}) \cdot P_{TPU}^{idle}) \cdot PUE$$

- P_{total} = Total power consumed for inference
- N_{TPU} = Total number of accelerators used
- $T_{requests}$ = queries per second for the application
- T_{TPU} = queries a second a single accelerator can handle
- P_{busy} = The power consumed by an accelerator when processing a query
- P_{idle} = The power consumed by an accelerator when idle
- PUE = Power Usage Effectiveness

Case Study – Google Translate

- Google handles roughly **1.2 million translation queries a second** [11]
- A single **TPU** die can handle **175 queries a second** [3]
- A 4-die **TPU** card draws **384W when busy** and **290W when idle** [12]
- Google has an average **PUE of 1.1** across all their datacenters [13]
- **How many accelerators?**

Parameter	Value
$T_{requests}$ (queries/s)	1,200,000
T_{TPU} (queries/s)	175
P_{busy} (W)	96
P_{idle} (W)	72.5
PUE	1.1

Case Study – Google Translate

Let's look at an **ideal** system and a more typical **redundant** system:

Number of TPUs	Efficiency (%)	Total Power (KW)	Energy per Year (MWh)	CO2e (t)
7000	98	660	5780	4100
100000	7	8100	71000	50000

To put this into perspective, let's compare the following:

	CO2e (t)	Efficient	Redundant
Round Trip NY-SF 1 Passenger	0.9	x4555	x56000
Average Human per Year	5	x820	x10000
Lifetime of Average Car	57	x72	x880

Inference Power Model - Limitations

What are the limitations of this model?

- **Communications** overhead is ignored
- Assumption that all TPUs have the **same PUE** and **CO2e**
- More **fine-grain** understanding of the **relationship between workload** and **power consumption**

Environmental Impact of Inference – what can be done

What can be done to reduce the environmental impact of inference?

- Transparency from machine learning companies
- More efficient hardware for running the models
- More efficient models to run
- Reducing the usage of Machine Learning hardware all together

Link to Our Work

Link to our own work – Low Power Accelerator Design for Inference

- Focus on reducing power consumption in accelerators
- Designed power modelling tools for CNN Accelerators
- Researching ways to address memory power consumption (roughly 30% of the system) for CNN Accelerators
- Designing a highly customisable CNN Accelerator architecture for FPGAs

Link to our own work – Edge device training

- [2] mention that a significant energy cost of training is the retraining of models after deployment in order to improve their performance

- Exploration of low-cost training of CNNs on edge devices
  - Development of tools to easily program and deploy a variety of architectures on low powered FPGA devices
  - Estimating network activations to finetune just the FC layer
    - Results suggest that applying the proposed methodology on a CPU can achieve a 10x speedup compared to retraining the entire network on a GPU with little difference in achieved accuracy

Conclusion – Is it worth it?

Datacenter Environmental Impact:

- General trend of more efficient datacenters
- Choice in datacenter and location important

PROMISING

Carbon Footprint of Training:

- More concrete methods of measuring impact needed
- Where do the true costs lie?

NEEDS WORK

Environmental Impact of Inference:

- The most significant contributor
- Not addressed in Industry

NEEDS WORK

Index of Terms

PUE	Power Usage Effectiveness. Ratio of total facility power to power delivered.
Performance per Watt	Number of operations done per second, per watt.
CO2e	Carbon Dioxide equivalent. This is the measure for the impact power usage has on the environment

Discussion Points

- *What can academics do to ensure they are not having a negative impact on the planet?*
- *Ways to hold companies more accountable for the cost of training?*
- *Ways to hold companies more accountable for the cost of inference?*
- *Is the impact on the environment justified?*
- *What is it that we are in the dark about in terms of environmental impact of datacenters?*
- *Moving to mobile/edge devices, is this the trend that might reduce power consumption of ICT energy effect?*
- *What are alternatives for ML computing?*

Imperial College London

References

1. David A. Patterson and Joseph Gonzalez and Quoc V. Le and Chen Liang and Lluis-Miquel Munguia and Daniel Rothchild and David R. So and Maud Texier and Jeff Dean (2021). Carbon Emissions and Large Neural Network Training. CoRR, abs/2104.10350.
2. Dhar, P. The carbon impact of artificial intelligence. Nat Mach Intell 2, 423–425 (2020)
3. Vijay Janapa Reddi and Christine Cheng and David Kanter and Peter Mattson and Guenther Schmuelling and Carole-Jean Wu and Brian Anderson and Maximilien Breughe and Mark Charlebois and William Chou and Ramesh Chukka and Cody Coleman and Sam Davis and Pan Deng and Greg Diamos and Jared Duke and Dave Fick and J. Scott Gardner and Itay Hubara and Sachin Idgunji and Thomas B. Jablin and Jeff Jiao and Tom St. John and Pankaj Kanwar and David Lee and Jeffery Liao and Anton Lokhmotov and Francisco Massa and Peng Meng and Paulius Micikevicius and Colin Osborne and Gennady Pekhimenko and Arun Tejusve Raghunath Rajan and Dilip Sequeira and Ashish Sirasao and Fei Sun and Hanlin Tang and Michael Thomson and Frank Wei and Ephrem Wu and Lingjie Xu and Koichi Yamada and Bing Yu and George Yuan and Aaron Zhong and Peizhao Zhang and Yuchen Zhou. "MLPerf Inference Benchmark". CoRR 2019; abs/1911.02549
4. Masanet, E., Shehabi, A., Lei, N., Smith, S. and Koomey, J., 2020. Recalibrating global datacenter energy-use estimates. Science , 367(6481), pp.984-986. https://datacenters.lbl.gov/sites/default/files/Masanet_et_al_Science_2020.full_.pdf .
5. So, D., Le, Q. and Liang, C., 2019, May. The Evolved Transformer. In International Conference on Machine Learning 2019 (pp. 5877-5886). PMLR. arXiv preprint arXiv:1901.11117
6. Adiwardana, D. , Luong, M., R. So, D., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., and Le. Q. Towards a Human-like Open-Domain Chatbot . arXiv preprint arXiv:2001.09977 .
7. Lacoste, A., Luccioni, A., Schmidt, V. and Dandres, T., 2019. Quantifying the carbon emissions of machine learning. arXiv preprint arXiv:1910.09700 .
8. Lannelongue, L., Grealey, J. and Inouye, M., 2020. Green algorithms: Quantifying the carbon footprint of computation. arXiv: 2007.07610 .
9. Peter Henderson and Jieru Hu and Joshua Romoff and Emma Brunskill and Dan Jurafsky and Joelle Pineau. "Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning". CoRR 2020; abs/2002.05651.
10. Lasse F. Wolff Anthony and Benjamin Kanding and Raghavendra Selvan. "Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models". CoRR 2020; abs/2007.03051.

11. https://blog.google/products/translate/ten-years-of-google-translate/

12. In-Datacenter Performance Analysis of a Tensor Processing Unit

13. https://www.google.co.uk/about/datacenters/efficiency/

14. https://www.epa.gov/energy/greenhouse-gases-equivalencies-calculator-calculations-and-references

15. Whitehead, Beth & Andrews, Deborah & Shah, Amip & Maidment, Graeme. (2014). Assessing the environmental impact of data centres part 1: Background, energy use and metrics. Building and Environment. 82. 151–159. 10.1016/j.buildenv.2014.08.021.

16. Masanet, Eric & Shehabi, Arman & Lei, Nuoa & Smith, Sarah & Koomey, Jonathan. (2020). Recalibrating global data center energy-use estimates. Science. 367. 984-986. 10.1126/science.aba3758.

17. Aishwarya T, Anusha K S, Gagana S, Megha V, 2021, Survey on Energy Consumption in Cloud Computing, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) NREST – 2021 (Volume 09 – Issue 04),

18. Castellazzi, Luca & Maria, AVGERINOU & Bertoldi, Paolo. (2017). Trends in data centre energy consumption under the European Code of Conduct for data centre energy efficiency. 10.2760/358256.

19. Jones, Nicola. (2018). How to stop data centres from gobbling up the world's electricity. Nature. 561. 163-166. 10.1038/d41586-018-06610-y. (nature)

20. C. Trueman, "What impact are data centres having on climate change?," Computerworld, 09-Aug-2019. [Online]. Available: https://www.computerworld.com/article/3431148/why-data-centres-are-the-new-frontier-in-the-fight-against-climate-change.html. [Accessed: 01-Sep-2021]. (computer world)

21. J. Robey, "How to take responsibility for your DATA CENTRE'S environmental impact," *Capgemini Worldwide*, 17-Mar-2020. [Online]. Available: https://www.capgemini.com/2020/01/the-more-sustainable-data-center/. [Accessed: 01-Sep-2021]. (capgemini)

22. A. Ezra, "Council post: Renewable energy alone can't address Data Centers' adverse environmental impact," *Forbes*, 03-May-2021. [Online]. Available: https://www.forbes.com/sites/forbestechcouncil/2021/05/03/renewable-energy-alone-cant-address-data-centers-adverse-environmental-impact/?sh=1ef33c4a5ddc. [Accessed: 01-Sep-2021]. (forbes)

23. DATA4, "How do we reduce the environmental footprint of data centers?," *DATA4*, 05-Jun-2020. [Online]. Available: https://www.data4group.com/en/csr/how-do-we-reduce-the-environmental-footprint-of-data-centers/. [Accessed: 01-Sep-2021]. (data4group)

24. DP Facilities, "Differentiating HDC from traditional enterprise data Center – Part 3," *DP Facilities*, 2020. [Online]. Available: https://www.dpfacilities.com/blog/differentiating-hdc-from-traditional-enterprise-data-centers/. [Accessed: 02-Sep-2021].

25. R. Miller, "How hyperscale customers & data centers are different," *Data Center Frontier*, 11-Oct-2019. [Online]. Available: https://datacenterfrontier.com/hyperscale-customers-data-centers-different/. [Accessed: 02-Sep-2021].

26. Md Abu Bakar Siddik, Arman Shehabi, & Landon Marston (2021). The environmental footprint of data centers in the United States. *Environmental Research Letters, 16(6), 064017.*