CMPSCI 677 Distributed Operating Systems **Spring 2025**

Lecture 14: March 26

Professor: Norman Bashir *Scribe: Richa Singh*

14.1 Overview

This lecture covers the following topics:

Part 1: Computing Demand Introduction

Part 2: Computing Demand Sustainability

Part 3: Using Computing To Improve Broader Sustainability

14.2 Lecture Notes

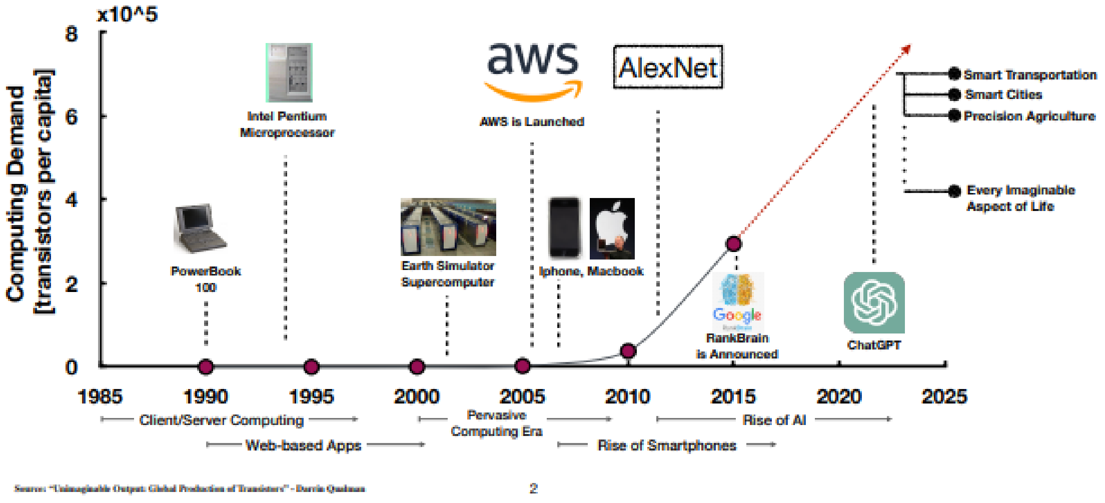
14.2.1 Part 1: Computing Demand Introduction

14.2.1.1 What Are Sustainable Systems?

Sustainable systems minimize environmental impact across their lifecycle. Key considerations:

- **Supply Chain:**
  - Evaluate material extraction, manufacturing, and transportation processes
  - Example: Complex hardware designs (e.g., GPUs) increase *embodied emissions*
- **Energy Efficiency:**
  - Optimize hardware/software to reduce power consumption
  - Koomey's Law: Energy efficiency doubles every 1.5–2.6 years
- **Energy Sources:**
  - Shift from fossil fuels (*brown energy*) to renewables (*green energy*)
- **End-of-Life Management:**
  - Recycling and responsible e-waste disposal
  - Example: Apple's lifecycle analysis shows 80% of iPhone emissions are embodied
- **Applications Beyond Compute:**
  - Buildings (appliance efficiency)
  - Transport (EVs charged with renewables)
  - Grid infrastructure

14.2.1.2 Computing Demands Increasing And Implications



Source: "Unimaginable Output: Global Production of Transistors" - Darrin Qualman

Figure 14.1: Computing demand vs time

As shown in figure 6.1, the key inventions of earlier years were not as frequent and didn't greatly increase demand. Currently, the computing demand is increasing exponentially and will continue to accelerate in growth.

The following key events have contributed to this increase in demand

1. Introduction of cloud services
2. Smart phone revolution
3. Rise of AI

While energy demands of computing is increasing, computing can be used to lower energy demands in other applications. For example, ride sharing could reduce the energy need to make cars since more people may not find the need to purchase them.

Successful apps inspire more people to use it. Jevons Paradox states that when you make an application more efficient, the energy to run the application reduces but the overall usage of the application increases. Thus, society's energy demand for a successful app increases.

Question: What is the effect of time efficiency on energy demand?
Answer: The figure doesn't account for increased energy effects as a result of the time efficiency.

14.2.1.3 How Is Demand Served?

Computing demand can be satisfied as follows (see **Figure 6.3**)

1. Big data centers, which have thousands of servers in them. These are huge facilitates and consume a huge amount of energy (hundreds of megawatts).

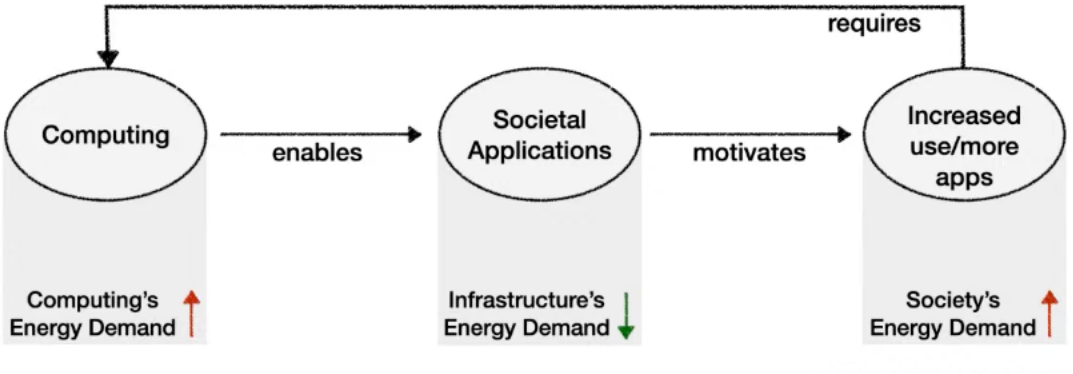


Figure 14.2: Implications of increasing computing demand

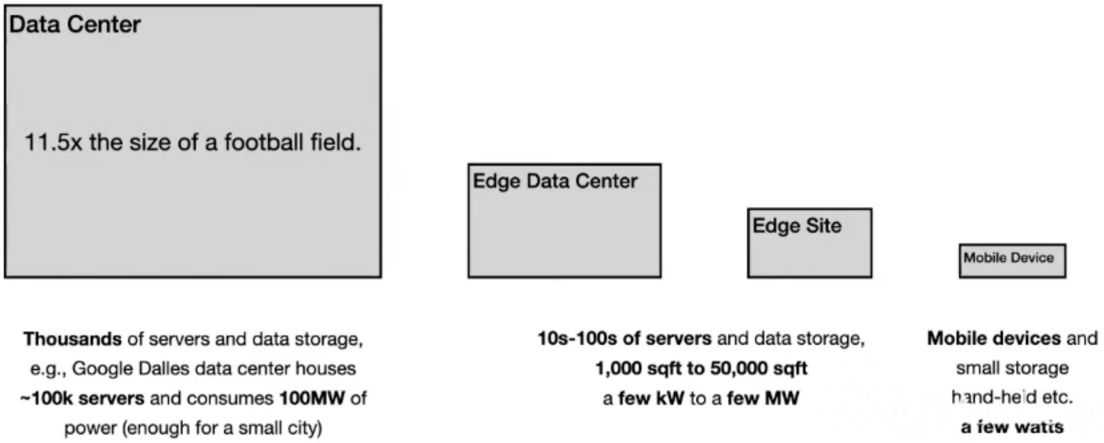

Data Center
11.5x the size of a football field.
Edge Data Center
Edge Site
Mobile Device
Thousands of servers and data storage, e.g., Google Dalles data center houses ~100k servers and consumes 100MW of power (enough for a small city)
10s-100s of servers and data storage, 1,000 sqft to 50,000 sqft a few kW to a few MW
Mobile devices and small storage hand-held etc. a few watts

Figure 14.3: Implications of increasing computing demand

2. Edge data centers. These are smaller and have hundreds of servers. They consumes a few kilowatts
3. Mobile devices, which consume few watts of power

14.2.2 Part 2: Computing Demand Sustainability

14.2.2.1 Contributions To Data Center Cost, Energy And Carbon Footprint

The following describe the cost, energy consumption and carbon footprint of data centers.

Cost:

1. Use of building and location of the data center
2. Replacing servers
3. Energy of data center

Energy:

1. Computing
2. Cooling equipment

Carbon Footprint:

1. **Embodied:** Carbon emissions from manufacturing computing hardware and data center building
2. **Operational:** Carbon emissions from energy used for computing and cooling

14.2.2.2 How To Serve Demand In A Sustainable Manner?

In the context of computing, **sustainability** is defined as operating the infrastructure in the least carbon intensive way. In other words, to be more sustainable is to emit less carbon when running your servers. Thus, the goal is to reduce embodied and operational carbon emissions. Carbon footprint can be computed as follows

$$\text{Carbon Footprint} = \frac{\text{Cycles per Unit Work} * \text{ Total Units of Work}}{\text{Computing's Energy Efficency} * \text{ Energy's Carbon Efficiency}}$$

Where **Cycles per unit work** is the amount of CPU cycles to perform a unit of work, **Total unit of work** is total work units computed, **Computing's energy efficiency** is how many cycles can be ran on a server per unit energy of work and **Energy's carbon efficiency** is how much unit energy is produced per 1 gram of carbon dioxide.

14.2.2.3 Drivers of Rising Emissions

AI & GPUs: Training large models (e.g., GPT-3) requires massive energy.
Data Center Growth: Global compute instances increased 6.4x (2010–2018) with only 5% energy rise (pre-AI).
Short Hardware Lifespans: Frequent upgrades increase embodied emissions.

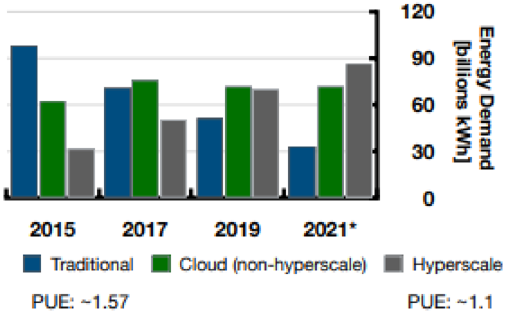
14.2.2.4 Motivations In Data Center Innovations

Power Usage Effectiveness (PUE) is a value multiplied by monthly server cost to get the total cost including cooling. In traditional data centers the PUE value is usually 2. State of the art data centers (hyperscale) can have a PUE value of 1.1. There has been a shift to hyperscale cloud providers over the years as shown in figure 6.2, which is caused by the cost of energy.

As shown in figure 6.3, predictions were made that energy demand will double every year, but some estimates show that is not the case, while others think that it has been worse. However, what is clear is that demand will eventually grow as it will no longer be offset with PUE savings.

14.2.2.5 Opportunities To Improve Carbon Foot Print

The following parameters to the carbon footprint equation are analyzed for their potential carbon emissions savings below



8 Source: Global data centre energy demand by data centre type, 2015-2021 - IEA

Figure 14.4: Energy demand with data center distinction

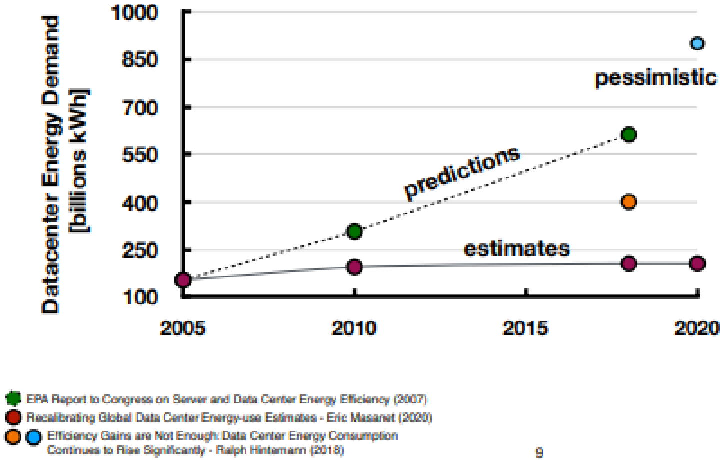


Figure 14.5: Data center energy demand vs time

Cycles per unit work: Algorithmic efficiency can help the CPU run work in less cycles, but there are limits. This parameter eventually has a bound.

Total Units of work: This is unbounded, there is no end to the amount of work a data center does. In fact, there is a monetary incentive to increase total work.

Computing's energy efficiency: Laundar's principle states that theoretical efficiency limit of CMOS will be hit by 2050, but could be practically sooner. Kommey's law states that energy efficiency doubles every 1.6-2.6 years. Jevan paradox states that gains in efficiency does not reduce demand. This parameter also has a bound.

Energy's carbon efficiency: A lot of gains can be made by transitioning to low carbon energy. Most of the work is done to optimize this parameter.

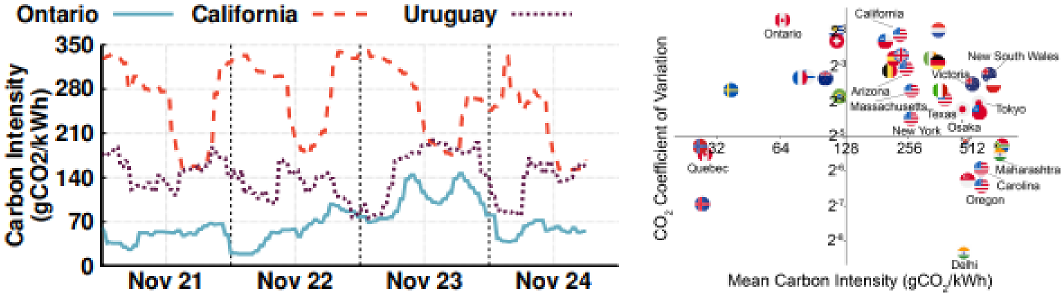


Figure 14.6: Carbon intensity in different regions

14.2.2.6 Grids Carbon Intensity And Clean Energy

Carbon intensity is how much carbon is emitted for a unit of energy created. While some countries have been improving in this metric, it is not possible to control how much carbon intensity the world is emitting.

Since carbon intensity tends to be variable, as shown in figure 6.4, there are opportunities of doing computation when carbon intensity is low. The variation in intensity is due to availability of solar energy. Computing also gives the ability to shift demands into regions with low intensity variations.

14.2.2.7 Computing Advantages

There are many options for running a processing job

1. Run immediately
2. Run somewhere else
3. Run later
4. Run slower/faster
5. Run intermittently

These options are useful when it comes to carbon intensity. For example, you can run a job faster when intensity is low or move a job to an area with less carbon intensity.

Why would companies want to improve sustainability? One reason is that customers care about it and the company wants to look green in their eyes. Additionally, companies selling other sustainable products need to know about their computing carbon foot print.

To read more about ”How can we leverage carbon intensity variations and computing’s flexibility?”, have a look at the papers on ”Enabling Sustainable Clouds: The Case for Virtualizing the Energy System” by Noman Bashir published at SoCC’21, ASPLOS’23.

14.2.2.8 Accounting For And Reducing Embodied Carbon

It is important to measure the emissions of a computing data center. There is disagreement about whether embodied or operational is more important to optimize. For example 80% of emissions of iphone is embodied, but the opposite could be true for servers. One view is that embodied emissions is for the producer of the good to optimize. In other words, everyone should focus on lowering their operational emission, which in turn lowers both types of emission. Another view is weighting the two equally.

14.2.2.9 Implications for Sustainable Computing

The following steps should be taken as an action plan for a carbon free grid. First, terms such as “Carbon-free”, “carbon-neutral”, “zero-carbon” and “100% renewable” should be clarified to prevent false impressions. Next, enhancing carbon visibility is important to know how much carbon you are emitting. Then, the focus should be shifted on carbon and not energy consumption. Finally, using computing flexibility to balance the grid.

14.2.2.10 Embodied and Operational Strategies

- **Reducing Embodied Emissions:**
  - **Design Simplification:**
    - ACT [ISCA’22]: Hardware design framework minimizing manufacturing complexity (15% emission reduction in FPGA prototypes).
    - Apple’s modular iPhone design cut embodied emissions by 15% over 3 generations.
  - **Lifetime Extension:**
    - Microsoft’s ”Junkyard Computing” reuses decommissioned SSDs/RAM in clusters (ASPLOS’23).
    - Extending server lifespan from 3 to 5 years reduces embodied emissions by 40%.
- **Reducing Operational Emissions:**
  - **Energy Efficiency:**
    - Koomey’s Law: Energy efficiency doubles every 1.5–2.6 years (e.g., Google TPU v4: 2x efficiency over v3).
    - Google’s PUE dropped from 1.5 (2008) to 1.2 (2023) via liquid cooling and AI-driven airflow optimization.
  - **Carbon-Aware Scheduling:**
    - **Temporal Shifting:** ”Let’s Wait Awhile” policy defers batch jobs (e.g., ML training) to low-carbon periods (30–55% savings with 18-hour delays).
    - **Spatial Shifting:** CDN-Shifter [SoCC’24] redirects traffic to low-carbon grids (80% savings in Europe with 60ms latency trade-offs).
    - **CarbonScaler:** Dynamically scales GPU resources based on carbon intensity (51–55% emissions cut without deadline extensions).
- **Trade-offs in Practice:**

Policy	**Carbon (g)**	**Energy (MWh)**	**Latency (ms)**
Carbon-Aware	4.17	39.7	139
Energy-Aware	11.2	6.1	28.6
Latency-Aware	258	23	5.5

14.2.3 Part 3: Using Computing To Improve Broader Sustainability

14.2.3.1 Computing Use Cases

There are ways to use computing in order to improve the energy consumption of other applications. For example, a building could have sensors to monitor temperature. That data could be analyzed to learn when people are in the building. This can be extended to automatically control lights and HVAC for the building. This is an example of a **Sense, Analyze, Control model**.

Another example is monitoring pests/disease in agriculture or sensing other cars to avoid congestion in transportation.

14.2.3.2 Power Monitoring And Analysis Of Data

At the home, there is outlet-level monitoring and meter-level monitoring. Data collected from these monitors can be used with machine learning or statistics to figure out patterns and inefficiencies.
Example: occupancy sensor feeds data to a smart thermostat to design schedule for heating a home.

14.2.3.3 Low Carbon Energy

There has been huge growth in renewable energy. The problem is that the energy is intermittent. One solution to this problem is to use past and weather data to forecast solar energy. A use case for this is electric vehicle charging. By analyzing when solar is high, you can delay their charging until then.

14.2.3.4 Computing for Global Sustainability

- **Smart Grids:**
  - Google's "24/7 Carbon-Free Energy" matches compute demand with solar/wind availability (90% carbon-free in Iowa data centers).
  - AI predicts renewable output to balance grid supply-demand (DeepMind's UK grid experiments: 10% efficiency gain).
- **Precision Agriculture:**
  - IBM's Watson Decision Platform cuts water/fertilizer use by 30% via soil moisture sensors and ML.
  - John Deere's autonomous tractors reduce diesel consumption by 20%.
- **Environmental Monitoring:**
  - Climate TRACE uses satellite imagery + AI to track global deforestation and methane leaks (30% faster detection).
  - Ocean Cleanup's ML models predict plastic accumulation zones with 95% accuracy.
- **Carbon Capture & Storage (CCS):**
  - DeepMind's AlphaFold optimizes enzyme designs for CO_2 sequestration (2x efficiency over manual methods).

14.2.3.5 Summary

1. Sustainable Computing
   (a) Demand is growing
   (b) Need to make server demand sustainable
   (c) Use computing unique advantages to optimize carbon footprint
   (d) Reduce operational carbon emission as well as embodied
2. Computing for Sustainability
   (a) Leverage computing so other sectors can reduce energy consumption + enhance use of carbon energy

Question: Does companies claimed PUE correct? How can companies claim to be zero carbon if they are running in places such as india?
Answer: PUE values are accurate. Companies do accounting magic by buying carbon offsets and investing in renewable and gaining carbon credits.
Question: What is your main research focus?
Answer: Lecturer has worked on all angles discussed in this presentation. Leveraging computing to make grid more efficient. Recently, has worked on sustainable computing itself.