Sustainable Computing

& Computing for Sustainability

Noman Bashir - MIT
nbashir@mit.edu
02/26/2024
|---|---|---|---|

| MIT Cluster Carbon Footprint [In preparation] | Embodied Accounting [HotCarbon'23] |

- **Supply Chain**
  - Extract, Process, Manufacture, Transport

- **Datacenter/Edge/Device**
  - software algorithms, scheduling
  - energy idle, dynamic
  - carbon grid, solar, battery

- **End of Life (EOL)**

- **Apps**

- **Electric Grid**
  - Buildings
  - Transportation

- **Sustainability of Computing**

- **Computing for Sustainability**
Computing’s Demand is Growing Exponentially

- Society continues to find useful applications

Source: “Unimaginable Output: Global Production of Transistors” - Darrin Qualman
Implications of Increasing Computing Demand

Computing's Energy Demand

enables

Societal Applications

motivates

Increased use/more apps

requires

Computing

Society's Energy Demand

Infrastructure's Energy Demand
How is Computing Demand Served?

**Data Center**

11.5x the size of a football field.

- **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)

**Edge Data Center**

- **10s-100s of servers** and data storage,
- **1,000 sqft to 50,000 sqft**
- **a few kW to a few MW**

**Edge Site**

**Mobile Device**

- **Mobile devices** and small storage
- **hand-held etc.**
- **a few watts**

*note: figures are not drawn to scale.*
What Contributes to Data Center’s Cost, Energy, Carbon Footprint?

Cost
- **Servers**: Cost a lot and are replaced every 3-5 years.
- **Building**: Capital investment, depends on location.
- **Energy**: Major cost of datacenter, depends on location.

Energy
- **Computing**: Become more energy efficient over time.
- **Cooling**: Wasted energy, significantly reduced over years.

Carbon
- **Embodied**: Carbon emissions from manufacturing/building.
- **Operational**: Emissions from energy use for compute and cooling.
How to Serve Computing’s Demand in a Sustainable Manner?

**Sustainable** —> least carbon intensive way.

**Carbon**
- **Embodied**: Carbon emissions from manufacturing/building.
- **Operational**: Emissions from energy use for compute and cooling.
  - From the energy used to run the servers.
  - From the energy used to cool the servers.

Reduce Embodied Emissions and Reduce Operational Emissions
Carbon Footprint = \frac{\text{Cycles per Unit Work} \times \text{Total Units of Work}}{\text{Computing's Energy Efficiency} \times \text{Energy's Carbon Efficiency}}

Carbon Footprint = \frac{10 \text{ cycles per inference request} \times 100 \text{ inference requests}}{5 \text{ cycles per kWh} \times 1 \text{ kWh per gCO2eq}}

Carbon Footprint = 200 \text{ gCO2eq}
History: Driving Factors Behind Innovations in Data Centers

**Cost of Energy Has Been Driving Innovation**

- Assume 100,000 servers
- **Monthly cost of 1 server**
  - 500W server
  - Cost = (Watts X Hours / 1000) * cost per kWh
  - Always-on server monthly cost = $50
- **Monthly cost of 100k servers = $5M**
- What about the cost of cooling?
  - Use Power Usage Effectiveness (PUE)
  - PUE = 2 → double the cost
  - PUE = 1.2 → 10% extra on $5M ($6M)

**Shift from Traditional Data Centers to Cloud**

- Energy Demand [billions kWh]
  - 2015: Traditional (PUE: ~1.57), Cloud (non-hyperscale), Hyperscale
  - 2017: Traditional (PUE: ~1.1), Cloud (non-hyperscale), Hyperscale
  - 2019: Traditional (PUE: ~1.1), Cloud (non-hyperscale), Hyperscale
  - 2021*: Traditional (PUE: ~1.1), Cloud (non-hyperscale), Hyperscale

Source: Global data centre energy demand by data centre type, 2015-2021 - IEA
Energy Efficiency Gains Moving Forward

- Most optimistic estimates suggest 6% increase from 2010-2018

![Chart showing data center energy demand with estimates and predictions for 2005 to 2020, with labels for pessimistic, predictions, and estimates.]

- EPA Report to Congress on Server and Data Center Energy Efficiency (2007)
- Recalibrating Global Data Center Energy-use Estimates - Eric Masanet (2020)
- Efficiency Gains are Not Enough: Data Center Energy Consumption Continues to Rise Significantly - Ralph Hintemann (2018)
Algorithmic Efficiency can be further improved, but has limits. Industry has strong incentive to improve the algorithmic efficiency. Recent focus on ML training and Crypto-mining.

Carbon Footprint = \( \text{Cycles per Unit Work} \times \text{Total Units of Work} \)

Computing’s Energy Efficiency \( \times \) Energy’s Carbon Efficiency

[Koomey’s Law: Energy efficiency doubles every 1.5-2.6 years] transition to cloud, dedicated hardware;

[Laundar’s Principle: Theoretical limit to be reached in 2050, practical sooner] [bounded]

Zero-carbon energy means carbon efficiency can be infinite.

Industry has helped subsidize zero-carbon energy.

Datacenter capacity increased by 6X from 2010-2018.

Crypto-mining and ML demand is outpacing Moore’s law.

Industry has strong incentive to maintain and accelerate growth.

Industry has helped subsidize zero-carbon energy.

Crypto-mining and ML demand is outpacing Moore’s law.

Industry has strong incentive to maintain and accelerate growth.
Grid’s Carbon Intensity Has Been Decreasing

- Energy’s carbon efficiency in the US has improved by 45.6% over 2001-2017

Carbon intensity may never truly reach 0gCO2eq per kWh. It may actually increase in parts of the world.

Source: Ember Global Electricity Review (2022)
Source: BP Statistical Review of World Energy
Source: Ember European Electricity Review (2022)
Carbon Intensity of Electricity Varies Across Space & Time

**Spatial Variations**: Move to the greenest data center possible

**Temporal Variations**: Move to a time slot with the lowest carbon emissions

Run when and where low-carbon energy is available.
Clean Energy is Variable and Unreliable

- Carbon intensity variation: **less than 50g** to **more than 800g** across time and geographical regions.

More regions in the world would look like Ontario in near future.

Source: electricityMap
Driven by efforts to improve user experience & scale

Driven by efforts to reduce costs, improve user experience, and scale.
How can we leverage carbon intensity variations and computing’s flexibility?
Enabling Sustainable Clouds: The Case for Virtualizing the Energy System


Collaborators:
* University of Massachusetts Amherst
^ Worcester Polytechnic Institute (WPI)
^^ California Institute of Technology (Caltech)
Ecovisor: A Virtual Energy System for Carbon-Efficient Applications
Ecovisor: A Virtual Energy System for Carbon-Efficient Applications
Ecovisor: Design and API

<table>
<thead>
<tr>
<th>Function Name</th>
<th>Type</th>
<th>Input</th>
<th>Return Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>set_container_powercap()</td>
<td>Setter</td>
<td>ContainerID, kW</td>
<td>N/A</td>
<td>Set a container's power cap</td>
</tr>
<tr>
<td>set_battery_charge_rate()</td>
<td>Setter</td>
<td>kW</td>
<td>N/A</td>
<td>Set battery charge rate</td>
</tr>
<tr>
<td>set_battery_max_discharge()</td>
<td>Setter</td>
<td>kW</td>
<td>N/A</td>
<td>Set max battery discharge rate</td>
</tr>
<tr>
<td>get_solar_power()</td>
<td>Getter</td>
<td>N/A</td>
<td>kW</td>
<td>Get virtual solar power output</td>
</tr>
<tr>
<td>get_grid_power()</td>
<td>Getter</td>
<td>N/A</td>
<td>kW</td>
<td>Get virtual grid power usage</td>
</tr>
<tr>
<td>get_grid_carbon()</td>
<td>Getter</td>
<td>N/A</td>
<td>g-CO2/kW</td>
<td>Get current grid carbon intensity</td>
</tr>
<tr>
<td>get_battery_discharge_rate()</td>
<td>Getter</td>
<td>N/A</td>
<td>kW</td>
<td>Get current rate of battery discharge</td>
</tr>
<tr>
<td>get_battery_charge_level()</td>
<td>Getter</td>
<td>N/A</td>
<td>kW</td>
<td>Get energy stored in virtual battery</td>
</tr>
<tr>
<td>get_container_powercap()</td>
<td>Getter</td>
<td>ContainerID</td>
<td>kW</td>
<td>Get a container's power cap</td>
</tr>
<tr>
<td>tick()</td>
<td>Notification</td>
<td>N/A</td>
<td>N/A</td>
<td>Invoked by ecovisor every Δt</td>
</tr>
</tbody>
</table>

Control Power Supply and Demand
Asynchronous Notifications
Get Energy System Information
Ecovisor: Prototype Implementation

- **Software**: REST API
- **Hardware**: 60 Rock64 nodes

1. Reducing carbon (ML training, MPI)
2. Budgeting carbon (webserver)
3. Leveraging batteries (web server, Spark)
4. Leveraging solar (MPI, straggler)
Ecovisor: Optimizing Carbon/Performance Trade-off

• **Evaluation objectives:** Demonstrate carbon savings, show applications should do optimizations.
• **Baseline:** (WaitAWhile - Middleware ’21), **Proposed:** Application-specific (Wait&Scale) policy

---

**PyTorch ML Training**

- Optimal Scale = 2X
- Two follow-up papers, CarbonScaler (system) and RORO (theory), on leveraging Elasticity will appear at SIGMETRICS’24.

---

**BLAST**

- Optimal Scale = 3X
- Embarrassingly parallel job.
Computing for Sustainability

Computing enables Societal Applications, which motivates Increased use/more apps. Societal’s Energy Demand requires Computing’s Energy Demand.
Computing Use Cases

Improving Buildings and Transportation Sectors

- Building as an example of a distributed system
  - **Sense** monitor energy, temperature, occupancy etc.
  - **Analyze** data using computational tools.
  - **Control** lights, HVAC, doors to reduce energy usage.

- Transportation as an example of a distributed system
  - Sense?
  - Analyze?
  - Control?

- Agriculture as an example of computing use case
  - Sense?
  - Analyze?
  - Control?
Building Monitoring

- Power metering at different levels
  - Outlet-level monitoring
  - Meter-level monitoring

Wemo smart plug  eGauge meter with interface  smart meter
Analyzing the data

• Energy monitors / sensors provide real-time usage data
  • Building monitoring systems (BMS) data from office / commercial buildings

• Modeling, Analytics and Predictions
  • Use statistical techniques, machine learning and modeling to gain deep insights
  • Which homes have inefficient furnaces, heaters, dryers?
  • Are you wasting energy in your home?
  • Is an office building’s AC schedule aligned with occupancy patterns?
  • When will the aggregate load or transmission load peak?
Reduce Energy Use —> Learning Thermostat

sensed data

HOME  AWAY  HOME

typical day

MONDAY  TUESDAY  SUNDAY

occupancy

schedule
Use Low Carbon Energy —> Use Solar Power

• Significant growth in renewable energy adoption
  • Roof top wind turbines, solar PV, solar thermal (water heating)

• Highly intermittent
  • Impacted by cloud cover, temperature, environmental variables
Forecasting Solar Energy

- Predictive analytics to model and forecast solar energy generation
  - Use machine learning and NWS weather forecasts to predict solar generation

- Better forecasts of near-term generation; “Sunny load” scheduling
Use Case - EV Charging

• Solar panels installed in parking lots, rest areas, paid garages
  • Possible use case in offices and car rental services

• Assumptions
  • Arrival/departure times for EVs
  • Accurate solar predictions

• Intelligent charging
  • When to charge?
  • Which EV to charge?
  • How much to charge?
Climate and Sustainability Implications of Generative AI

Noman Bashir\textsuperscript{1}, Priya L. Donti\textsuperscript{2,3}, James Cuff\textsuperscript{4}, Sydney Sroka\textsuperscript{1}, Marija Ilic\textsuperscript{2,3}, Vivienne Sze\textsuperscript{4,5,6,7}, Christina Delimitrou\textsuperscript{7}, Elsa A. Olivetti\textsuperscript{1,8}

\textsuperscript{1} MIT Climate & Sustainability Consortium (MCSC), \textsuperscript{2} MIT Electrical Engineering and Computer Science (EECS), \textsuperscript{3} MIT Laboratory for Information & Decision Systems (LIDS), \textsuperscript{4} MIT Office of Research Computing & Data (ORCD), \textsuperscript{5} MIT Research Lab of Electronics (RLE), \textsuperscript{6} MIT Microsystems Technology Laboratories (MTL), \textsuperscript{7} MIT Computer Science & Artificial Laboratory (CSAIL), \textsuperscript{8} MIT Materials Science & Engineering (DMSE)
Unfettered Growth and Its Key Drivers

- ChatGPT 1 million users in 5 days
- Only 15% of the users are from US
- Interest in Gen-AI perceived benefits
Unfettered Growth and Its Key Drivers

- ChatGPT: 1 million users in 5 days
- Only 5% of the users are from the US
- Interest in Gen-AI
- Consolidation of AI capabilities
- Lack of regulatory oversight
- Efficiency improvements

ChatGPT reached 1 million users in 5 days. Only 5% of the users are from the US. Interest in Gen-AI is rising. Consolidation of AI capabilities, lack of regulatory oversight, and efficiency improvements are key drivers of growth.
Need for Comparative Benefit-cost Evaluation Capability

• Scope
  • E.g., a search query.

• Boundaries
  • Geographical: A given region or a data center.
  • Temporal: A given window of time.
  • Conceptual: A search query.

• Baselines and scenarios
  • A standard Google search as a baseline.
  • Various GPT models as scenarios.

• Metrics and data
  • Energy usage, GHG emissions, water usage, and raw material.
**Illustrative Example: Generative AI-based Search**

1. **Baseline**: User needs $M$ queries.
2. **Gen-AI**: User needs $N$ queries.
3. **Baseline & Gen-AI**: Both incur costs during data processing and R&D phase.
4. **Gen-AI**: Model training is an additional cost.
5. **Baseline & Gen-AI**: Both incur per-query costs, which may differ.
6. **Baseline & Gen-AI**: Both incur costs during supply chain and end-of-life phases.
7. **Baseline & Gen-AI**: User’s actions have system-level socioeconomic impacts.

The **computing-related** costs include raw material usage, energy consumption, waste generation, and water use.

The **immediate application** impacts include the reduced time spent on search and quality of response.

The **system-level** impacts include broader socioeconomic impacts computing as well as user using the Gen-AI for search.
Stakeholder Engagement for Responsible Development in Gen-AI

- Policymakers and legal experts
- AI practitioners and engineers
- Energy and supply chain experts
- Economists
- Social scientists
- Civil society
- End users
Leveraging Benefit-cost Evaluation Framework

• Monitoring the evolution of Gen-AI as a sector
• Identifying opportunities to improve benefit-cost ratio
• Facilitating eco-economic decoupling and constrained growth
Summary

• Sustainable Computing
  • Demand for computing is growing
  • Need to serve the demand sustainably
  • Energy efficiency gains reducing
  • Computing has unique advantages
  • Try to optimize computing’s carbon efficiency
  • Reduce operational as well as embodied carbon

• Computing for Sustainability
  • Leverage computing to reduce energy consumption
  • Leverage computing to enhance use of low carbon energy