Carbon Measurement and Attribution for Processes and Hardware Devices in the Linux Kernel

Aditya Manglik
ETH Zürich, Switzerland

Contact: manglik.aditya@gmail.com

EcoCompute-2024
25 April, 2024

Brief Introduction

Graduate student at ETH Zürich, Switzerland

Brief Introduction

Graduate student at ETH Zürich, Switzerland

Research at the intersection of computer architecture, operating systems, and networks

Outline

Background

- Energy sources in computation systems:
  Direct: DC input / USB / Ethernet

Background

- Energy sources in computation systems:
  - Direct: DC input / USB / Ethernet
  - Battery

Background

- Energy sources in computation systems:
  - Direct: DC input / USB / Ethernet
  - Battery
  - Energy harvesting

Background

- Energy sources in computation systems:
  - Direct: DC input / USB / Ethernet
  - Battery
  - Energy harvesting
- We want to use the ~~maximum~~ minimum amount of energy to perform computation

Background

- Energy sources in computation systems:
  - Direct: DC input / USB / Ethernet
  - Battery
  - Energy harvesting
- We want to use the ~~maximum~~ minimum amount of energy to perform computation
- Energy (*battery*) capacity is a major design constraint for any computation platform, e.g., mobile phones or AR headsets

Outline

The lack of tools

Performance optimization is well-understood

The lack of tools

Performance optimization is well-understood

Measure latency using mature tools (e.g., perf) and consistent metrics (e.g., CPU clock cycles)

The lack of tools

Performance optimization is well-understood

Measure latency using mature tools (e.g., perf) and consistent metrics (e.g., CPU clock cycles)

Question: Tools to measure the application's energy?

Calculating Energy Consumption of a Process

$$\text{Energy Consumption} = \text{Power} \times \text{Latency}$$

Calculating Energy Consumption of a Process

Energy Consumption = Power $\times$ Latency

Power is reported by the CPU (e.g., RAPL for Intel) or datasheet

Calculating Energy Consumption of a Process

Energy Consumption = Power $\times$ Latency

Power is reported by the CPU (e.g., RAPL for Intel) or datasheet

Example: CPU reports $\approx$ 15 W

Calculating Energy Consumption of a Process

Energy Consumption = Power $\times$ Latency

Power is reported by the CPU (e.g., RAPL for Intel) or datasheet

Example: CPU reports $\approx$ 15 W

Latency can be measured using `time or perf`

Calculating Energy Consumption of a Process

Energy Consumption = Power $\times$ Latency

Power is reported by the CPU (e.g., RAPL for Intel) or datasheet

Example: CPU reports $\approx$ 15 W

Latency can be measured using `time or perf`

Example: Process A takes $\approx$ 5 ms

Calculating Energy Consumption of a Process

Energy Consumption = Power $\times$ Latency

Power is reported by the CPU (e.g., RAPL for Intel) or datasheet

Example: CPU reports $\approx$ 15 W

Latency can be measured using `time or perf`

Example: Process A takes $\approx$ 5 ms

Energy Consumption = 15 W $\times$ 5 ms = 75 mJ

Calculating Energy Consumption of a Process

Energy Consumption = Power × Latency

Power is reported by the CPU (e.g., RAPL for Intel) or datasheet

Example: CPU reports ≈ 15 W

Latency can be measured using `time or perf`

Example: Process A takes ≈ 5 ms

Energy Consumption = 15 W × 5 ms = 75 mJ

Problem: Does not reflect the ground truth!

Oversight in Calculation Model

The model assumes linear power draw

Oversight in Calculation Model

The model assumes linear power draw

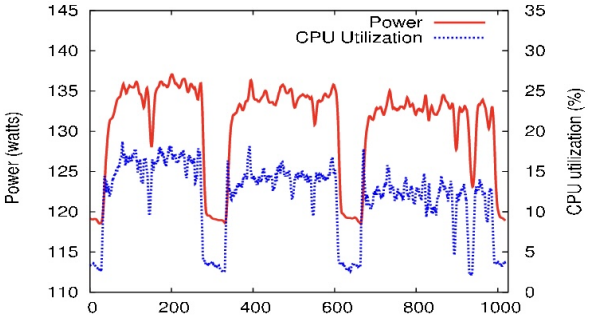


Figure: CPU power draw over time

Oversight in Calculation Model

The model assumes linear power draw

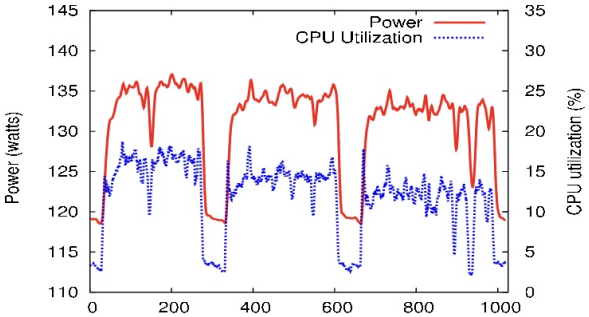


Figure: CPU power draw over time

Limitation 1: Power (on y-axis) is not constant over time (on x-axis) due to power-gating

Calculation Model

- The calculation model focuses on the CPU

Calculation Model

- The calculation model focuses on the CPU
- **Limitation 2**: What about devices like memory (DRAM) and the network interface?

Calculation Model

- The calculation model focuses on the CPU
- **Limitation 2**: What about devices like memory (DRAM) and the network interface?
- Experimental data contrary to assumptions, corroborated by [1]

[1] Barroso, Luiz André, Urs Hölzle, and Parthasarathy Ranganathan. "The datacenter as a computer: Designing warehouse-scale machines." Synthesis Lectures on Computer Architecture 13.3 (2018): i-189.

Ground Truth

- Platform-specific interfaces: RAPL is available only on specific Intel processors

Ground Truth

- Platform-specific interfaces: RAPL is available only on specific Intel processors
- Conflicting values from datasheets

Ground Truth

- Platform-specific interfaces: RAPL is available only on specific Intel processors
- Conflicting values from datasheets
- **Limitation 3**: No uniform interfaces or data formats to report power reliably across different platforms and devices

Problem Summary

- We are *inaccurately* calculating only *a fraction* of a *specific* system's actual energy consumption!

Problem Summary

- We are *inaccurately* calculating only *a fraction* of a *specific* system's actual energy consumption!

- **Take away:** We cannot improve what we cannot measure.

Outline

Goal

Develop a framework to *accurately and reliably* measure the energy consumption of a process on Linux

Goal

Develop a framework to *accurately and reliably* measure the energy consumption of a process on Linux

Report the statistics to the

Goal

Develop a framework to *accurately and reliably* measure the energy consumption of a process on Linux

Report the statistics to the

- End-users: In an easy-to-understand and useful format

Goal

Develop a framework to *accurately and reliably* measure the energy consumption of a process on Linux

Report the statistics to the

- End-users: In an easy-to-understand and useful format
- Programmers: Via APIs that improve programmer actionability

Goal

- **Framework** = Models and Tools

Goal

- **Framework** = Models and Tools
- **Power models** = How we reason about and estimate a device's power draw over time

Goal

- **Framework** = Models and Tools
- **Power models** = How we reason about and estimate a device's power draw over time
- **Power models** are often not available or poorly understood for many devices, e.g., network interfaces

Goal

- **Framework** = Models and Tools
- **Power models** = How we reason about and estimate a device's power draw over time
- **Power models** are often not available or poorly understood for many devices, e.g., network interfaces
- **Tools** can be built to accurately calculate power based on the models, e.g., nvidia-smi for Nvidia GPUs

Goal

- **Framework** = Models and Tools
- **Power models** = How we reason about and estimate a device's power draw over time
- **Power models** are often not available or poorly understood for many devices, e.g., network interfaces
- **Tools** can be built to accurately calculate power based on the models, e.g., nvidia-smi for Nvidia GPUs
- **Summary:** We need accurate models and reliable tools to calculate energy consumption

Outline

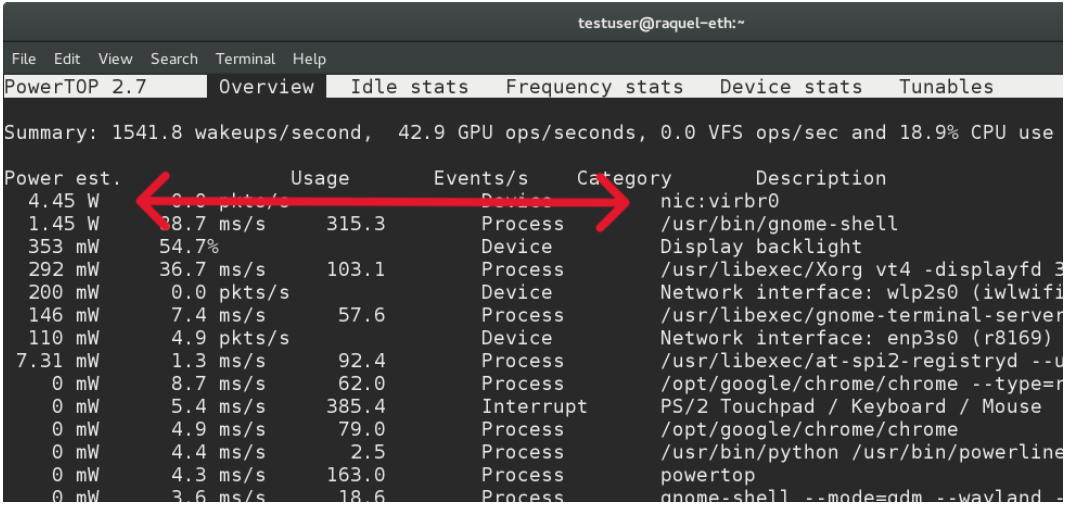
Hardware Solution

- Probe the wires or input supply

Hardware Solution

- Probe the wires or input supply
- Reliable but does not scale!

PowerTOP


testuser@raquel-eth:~
File Edit View Search Terminal Help
PowerTOP 2.7 Overview Idle stats Frequency stats Device stats Tunables
Summary: 1541.8 wakeups/second, 42.9 GPU ops/seconds, 0.0 VFS ops/sec and 18.9% CPU use
Power est. Usage Events/s Category Description
4.45 W 0.0 pkts/s Device nic:virbr0
1.45 W 38.7 ms/s 315.3 Process /usr/bin/gnome-shell
353 mW 54.7% Device Display backlight
292 mW 36.7 ms/s 103.1 Process /usr/libexec/Xorg vt4 -displayfd 3
200 mW 0.0 pkts/s Device Network interface: wlp2s0 (iwlwifi
146 mW 7.4 ms/s 57.6 Process /usr/libexec/gnome-terminal-server
110 mW 4.9 pkts/s Device Network interface: enp3s0 (r8169)
7.31 mW 1.3 ms/s 92.4 Process /usr/libexec/at-spi2-registryd --u
0 mW 8.7 ms/s 62.0 Process /opt/google/chrome/chrome --type=r
0 mW 5.4 ms/s 385.4 Interrupt PS/2 Touchpad / Keyboard / Mouse
0 mW 4.9 ms/s 79.0 Process /opt/google/chrome/chrome
0 mW 4.4 ms/s 2.5 Process /usr/bin/python /usr/bin/powerline
0 mW 4.3 ms/s 163.0 Process powertop
0 mW 3.6 ms/s 18.6 Process gnome-shell --mode=gdm --wayland

PowerTOP

It is possible to use Powertop to view the "power estimate" of a process/device/interrupt/timer.

PowerTOP

It is possible to use Powertop to view the "power estimate" of a process/device/interrupt/timer.

Challenges:

1. Power estimate is a discrete-time event. Energy consumption is a continuous process with a higher correlation to battery drain.

PowerTOP

It is possible to use Powertop to view the "power estimate" of a process/device/interrupt/timer.

Challenges:

1. Power estimate is a discrete-time event. Energy consumption is a continuous process with a higher correlation to battery drain.
2. Vendor-specific implementation

PowerTOP

It is possible to use Powertop to view the "power estimate" of a process/device/interrupt/timer.

Challenges:

1. Power estimate is a discrete-time event. Energy consumption is a continuous process with a higher correlation to battery drain.
2. Vendor-specific implementation
3. Actionability of this data for programmers

PowerTOP

It is possible to use Powertop to view the "power estimate" of a process/device/interrupt/timer.

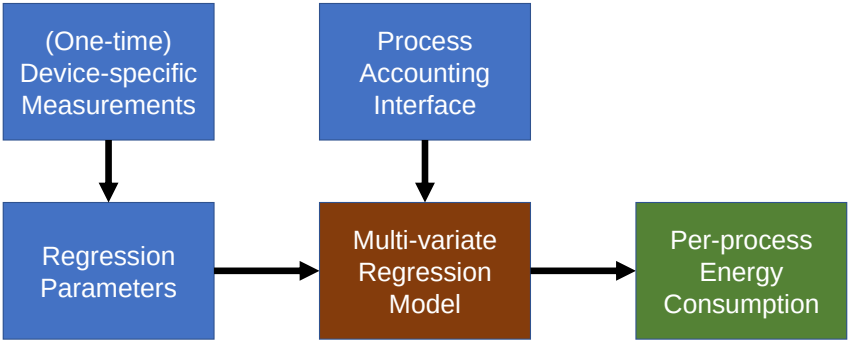
Challenges:

1. Power estimate is a discrete-time event. Energy consumption is a continuous process with a higher correlation to battery drain.
2. Vendor-specific implementation
3. Actionability of this data for programmers

   Process X consumes 1.45 Watts. What should the programmer do to optimize it?

Outline

System Design



Device-Specific Measurements

Goal: Determine regression parameters

Device-Specific Measurements

Goal: Determine regression parameters
Algorithm:

Device-Specific Measurements

Goal: Determine regression parameters
Algorithm:

1. Minimize system load by turning off all devices

Device-Specific Measurements

Goal: Determine regression parameters

Algorithm:

1. Minimize system load by turning off all devices
2. Reliably measure battery drain over multiple intervals

Device-Specific Measurements

Goal: Determine regression parameters

Algorithm:

1. Minimize system load by turning off all devices
2. Reliably measure battery drain over multiple intervals
3. Turn on a single target device

Device-Specific Measurements

Goal: Determine regression parameters

Algorithm:

1. Minimize system load by turning off all devices
2. Reliably measure battery drain over multiple intervals
3. Turn on a single target device
4. Sweep target device parameters from low to high while measuring battery drain

Device-Specific Measurements

Goal: Determine regression parameters

Algorithm:

1. Minimize system load by turning off all devices
2. Reliably measure battery drain over multiple intervals
3. Turn on a single target device
4. Sweep target device parameters from low to high while measuring battery drain
5. Turn off target device or set parameter to low

Device-Specific Measurements

Goal: Determine regression parameters

Algorithm:

1. Minimize system load by turning off all devices
2. Reliably measure battery drain over multiple intervals
3. Turn on a single target device
4. Sweep target device parameters from low to high while measuring battery drain
5. Turn off target device or set parameter to low
6. Repeat step 3-5 for all target devices

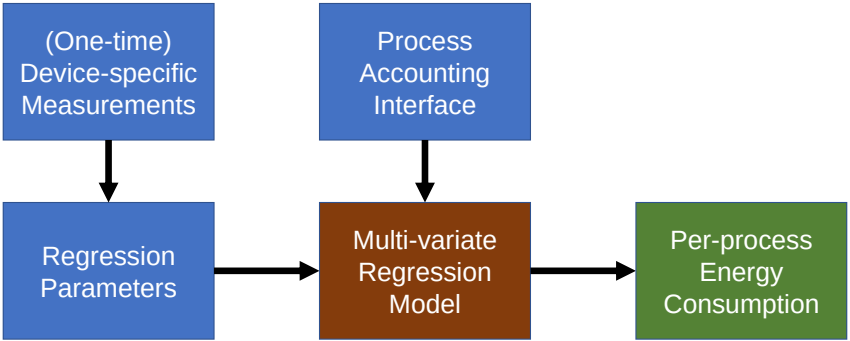
Device-Specific Measurements

Goal: Determine regression parameters

Algorithm:

1. Minimize system load by turning off all devices
2. Reliably measure battery drain over multiple intervals
3. Turn on a single target device
4. Sweep target device parameters from low to high while measuring battery drain
5. Turn off target device or set parameter to low
6. Repeat step 3-5 for all target devices
7. Solve for regression parameters (A)

System Design



Kernel Process Accounting Infrastructure

Goal: Determine regression inputs

Kernel Process Accounting Infrastructure

Goal: Determine regression inputs

Algorithm:

Kernel Process Accounting Infrastructure

Goal: Determine regression inputs

Algorithm:

1. Determine PID and group processes

Kernel Process Accounting Infrastructure

Goal: Determine regression inputs

Algorithm:

1. Determine PID and group processes
2. Poll the process accounting infrastructure for the PID

Kernel Process Accounting Infrastructure

Goal: Determine regression inputs

Algorithm:

1. Determine PID and group processes
2. Poll the process accounting infrastructure for the PID
3. Calculate CPU time allocation, memory set, open file handles (disk), screen wakeups, and network sockets.

Kernel Process Accounting Infrastructure

Goal: Determine regression inputs

Algorithm:

1. Determine PID and group processes
2. Poll the process accounting infrastructure for the PID
3. Calculate CPU time allocation, memory set, open file handles (disk), screen wakeups, and network sockets.
4. Calculate the fraction for each process over total time

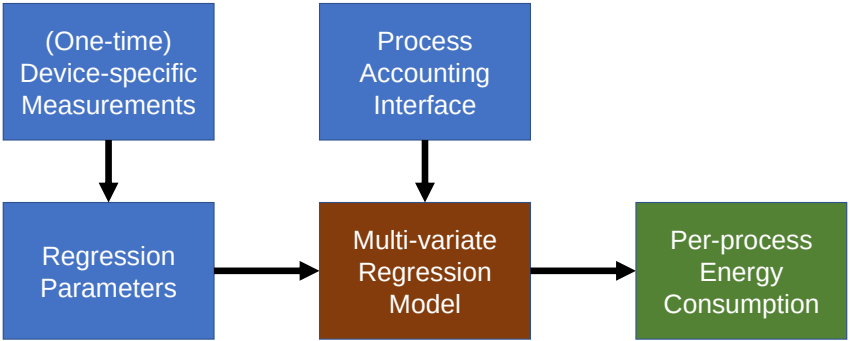
Kernel Process Accounting Infrastructure

Goal: Determine regression inputs

Algorithm:

1. Determine PID and group processes
2. Poll the process accounting infrastructure for the PID
3. Calculate CPU time allocation, memory set, open file handles (disk), screen wakeups, and network sockets.
4. Calculate the fraction for each process over total time
5. Input the fraction (X) in the regression model

System Design



Challenge: System Design

- **Estimated value**: All models are wrong, but some are useful

Challenge: System Design

- **Estimated value**: All models are wrong, but some are useful
- **Accuracy and Bias trade-off**: Accurate models generate larger systemic load that biases observations

Challenge: Data Collection

- There are millions of devices and billions of ICs inside these devices.

Challenge: Data Collection

- There are millions of devices and billions of ICs inside these devices.
- The power estimates can range across 2-3 orders of magnitude.

Challenge: Data Collection

- There are millions of devices and billions of ICs inside these devices.
- The power estimates can range across 2-3 orders of magnitude.
- How can we develop **accurate & reliable** power models across this diversity of devices?

Challenge: Validation of Ground Truth

- There is often significant difference between estimated values (from the model) and actual values (ground truth)

Challenge: Validation of Ground Truth

- There is often significant difference between estimated values (from the model) and actual values (ground truth)
- How to **identify divergence** from ground truth without hardware measurements or datasheets for validation?

Challenge: Privacy

- To develop **accurate & reliable** power models, we need data from different devices and users

Challenge: Privacy

- To develop **accurate & reliable** power models, we need data from different devices and users
- **Privacy**: Should users share this data to a "centralized" server?

Carbon Emissions for a Process

$$\text{Carbon Footprint} = \text{Energy Consumption} \times \text{Energy Composition}$$

Carbon Emissions for a Process

$$\text{Carbon Footprint} = \text{Energy Consumption} \times \text{Energy Composition}$$

$$\text{Energy Consumption} = \text{Power} \times \text{Latency}$$

Carbon Emissions for a Process

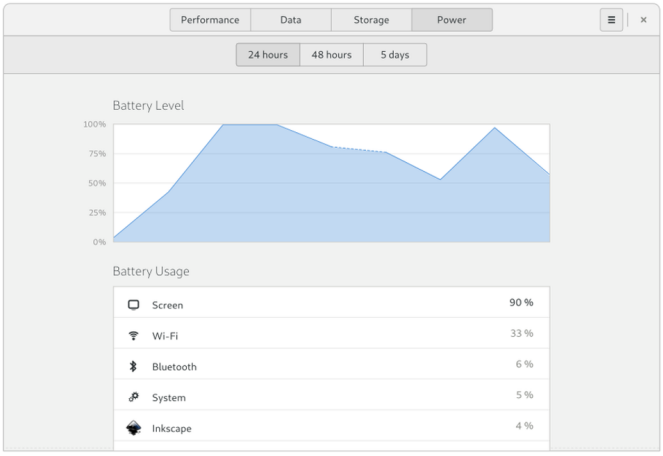
Carbon Footprint = Energy Consumption $\times$ Energy Composition

Energy Consumption = Power $\times$ Latency

Energy Composition depends on multiple factors, including geography, time of use, sourcing, and grid load

Outline

End-users



UI Credits: Allan Day, GNOME

Programmers

Command-line API for programmers: Indicate processes with high energy consumption

Example use-case: Energy-efficient code optimization suggestions in the coding platform

Outline

Key Takeaways

We cannot improve what we cannot measure.

Key Takeaways

We cannot improve what we cannot measure.

Non-CPU system components may dominate the overall energy consumption.

Thank you!

Feedback? manglik.aditya@gmail.com

Follow-up?