Panel Discussion:
The Sustainability of Foundation Models
(Can AI be sustainable?)

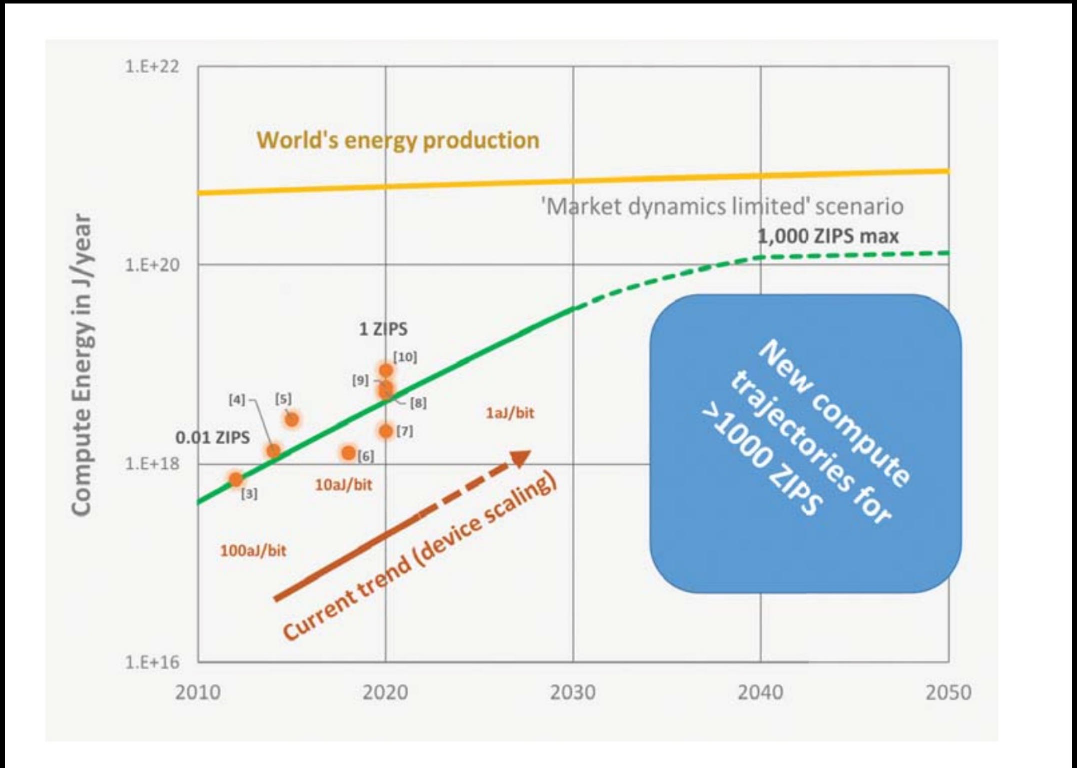
Martha Kim (Columbia University), Ramya Raghavendra (META),

Huamin Chen (RedHat), Andrew Chien (University of Chicago),

Sanjay Krishnan (University of Chicago)

Moderator: Eun Kyung Lee (IBM)

Ever rising energy demands for computing vs. global energy production is creating new risk, and new opportunities for **radically different computing paradigms to drastically improve energy efficiency**



31%

a years the energy consumption increase trend for hyperscalers in North America

>10%

of the world's power will be consumed by hyperscalers by 2030

Why this is important

Datacenter energy consumption and technology trends

Datacenter energy consumption will increase to **8% - 20%** by **2030**.

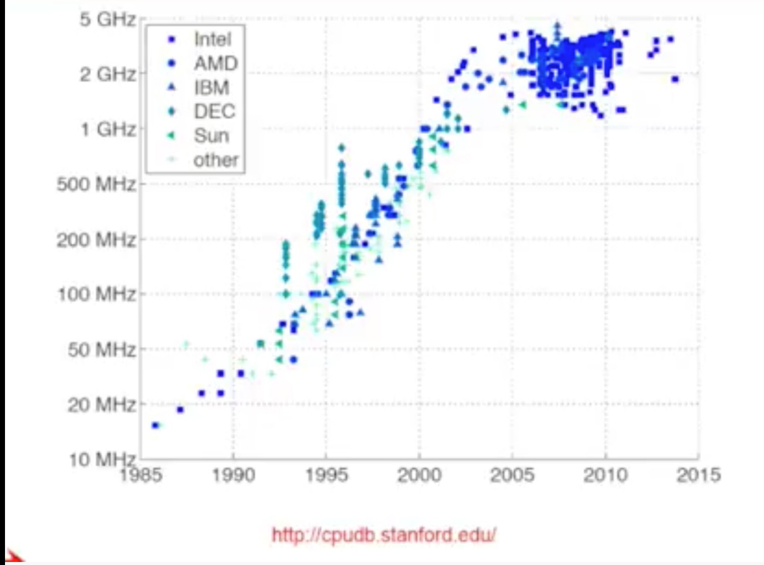
End of Dennard Scaling (Moore's Law)

AI power consumption **doubles every 3 – 4 months.** Large AI training jobs have life cycle carbon footprint of 5 cars (red AI).

GPT-4 1-time training consumes 7.5 megawatt-hours (MWh) of energy

🏘🏘🏘🏘🏘🏘🏘🏘 . . . . .

700 household annual energy consumption



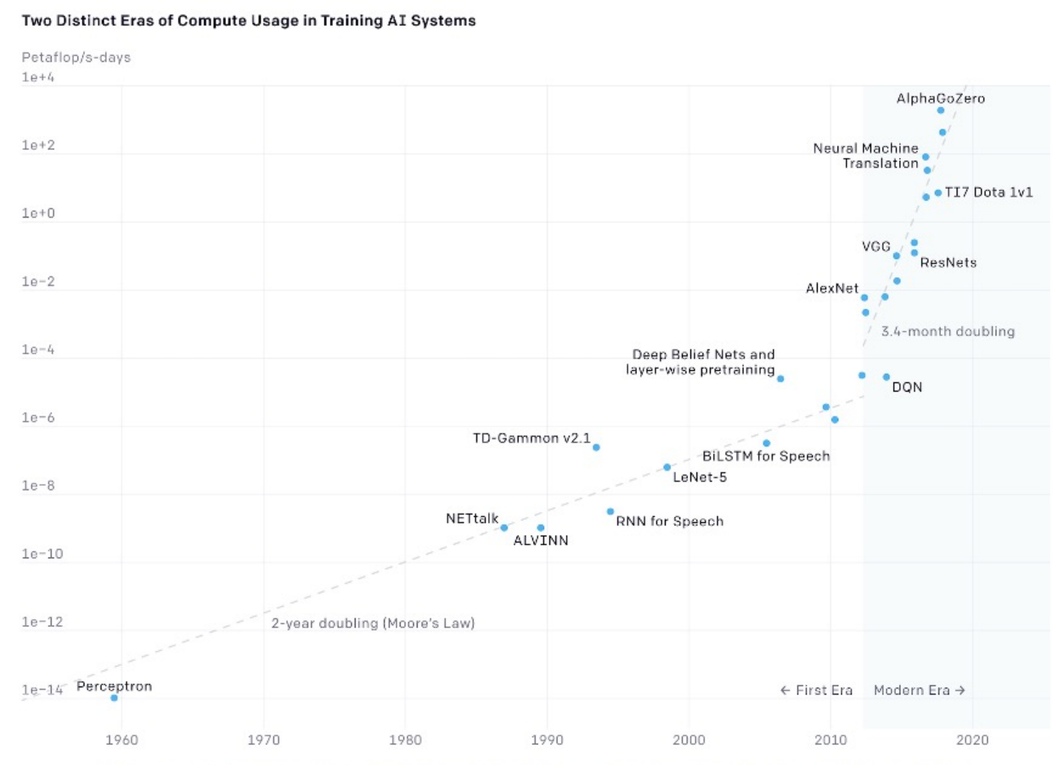


Figure 1 Compute usage doubles in 3-4 months in training AI models

Green AI, R. Schwartz, J. Dodge, N. A. Smith, O. Etzioni 2019

http://cpudb.Stanford.edu

Martha Kim

Columbia University

Can AI be sustainable? Yes!

- Far too early in technology lifecycle to declare defeat
- Ample opportunity to improve (even with sub-optimal carbon models)

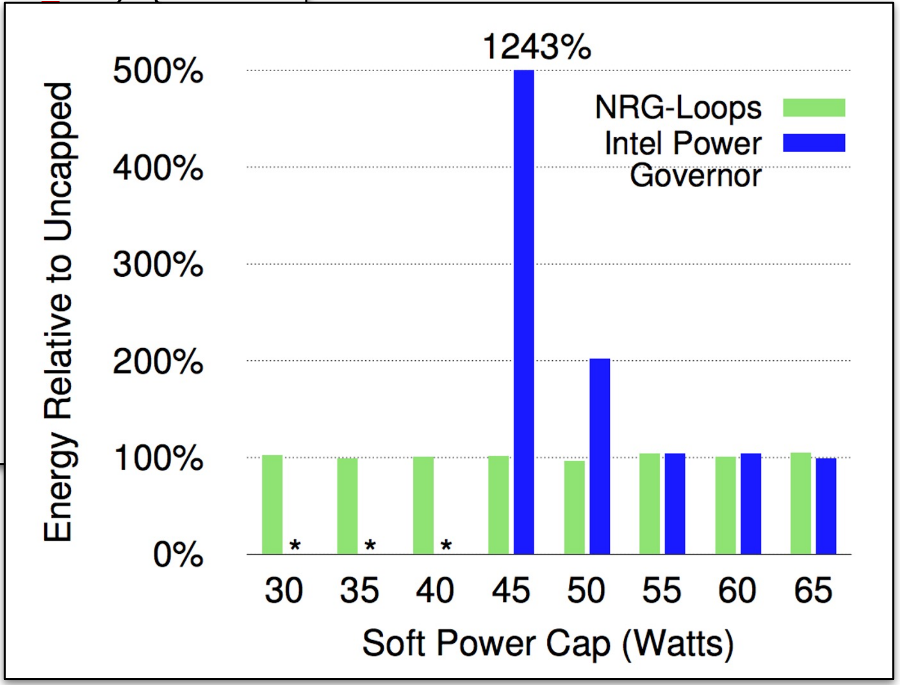
$$\frac{\text{Carbon}}{\text{Task}} = \underset{\substack{\Big\uparrow \\ \text{Application efficiency} \\ \text{(HW + SW)}}}{\frac{\text{J used}}{\text{Task}}} \;\text{x}\; \underset{\substack{\Big\uparrow \\ \text{Datacenter PUE}}}{\frac{\text{J supplied}}{\text{J used}}} \;\text{x}\; \underset{\substack{\Big\uparrow \\ \text{Carbon intensity} \\ \text{of power source}}}{\frac{\text{Carbon}}{\text{J supplied}}}$$

- Can probably optimize what we’re doing today
- Closed loop between application and system is very powerful

Power Capping from Inside Application

Substring search, with adaptive thread count

```
NRG_ADAPT_for (int i=0; i<STRINGS_TO_CHECK; ++i && NRG_AVG_P <= SOFT_CAP) {
    if (num_threads < MAX) num_threads += 2;
    // num threads search concurrently for substring
} NRG_ALTERNATE {
    num_threads -= 2;
    if (num_threads < MIN) num_threads = MIN;
    // num threads perform search
}
```



Can meet a broader range of power caps at significantly less energy

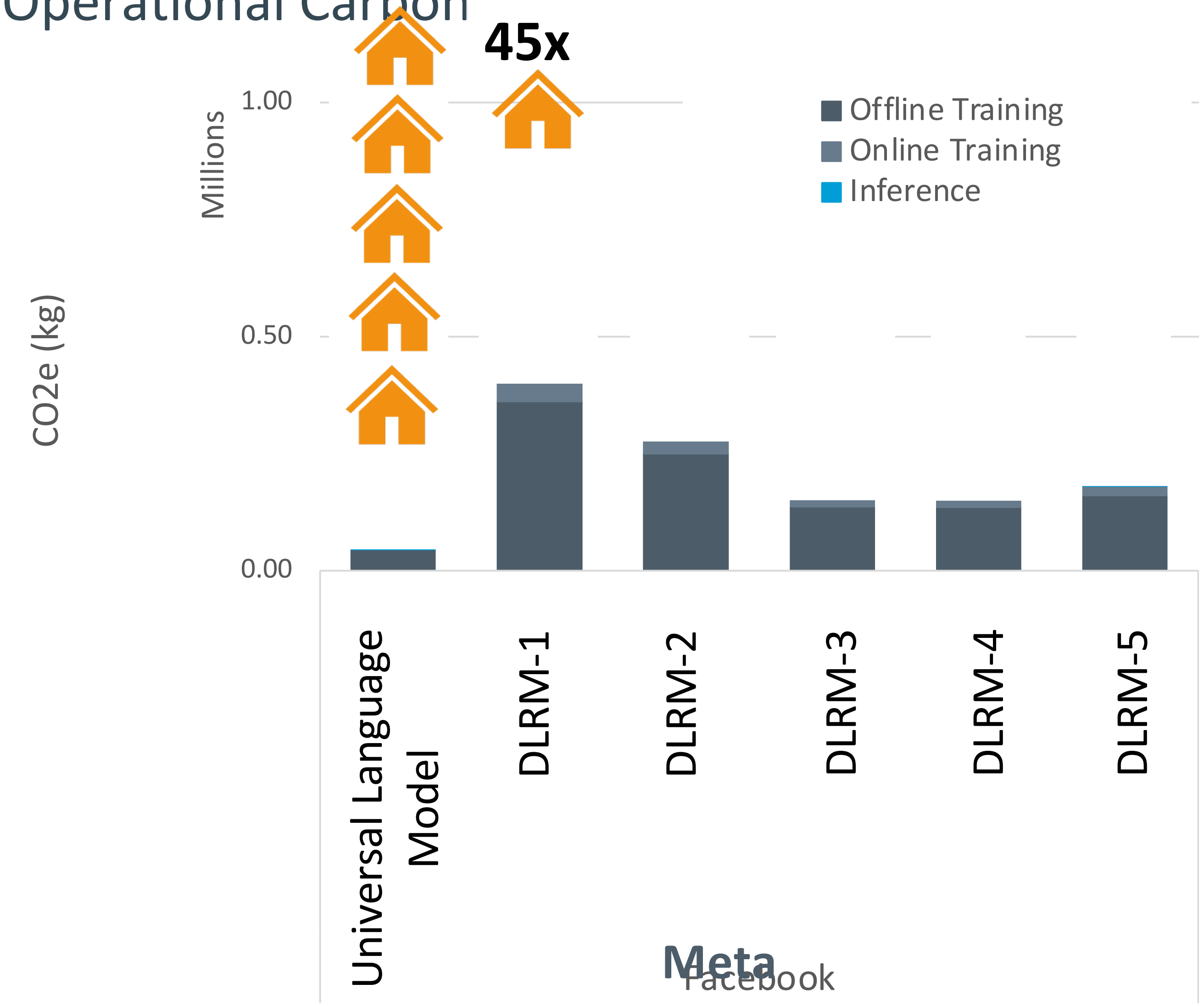
[“NRG-loops: adjusting power from within applications.” CGO ‘16]

Ramya Raghavendra

Meta

AI’s Carbon Footprint

Operational Carbon



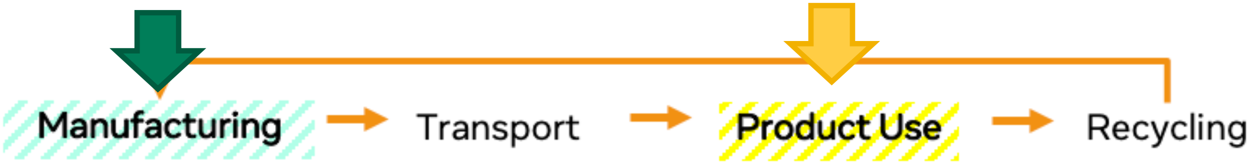
Universal Language Model Training

≈5 Home’s Annual

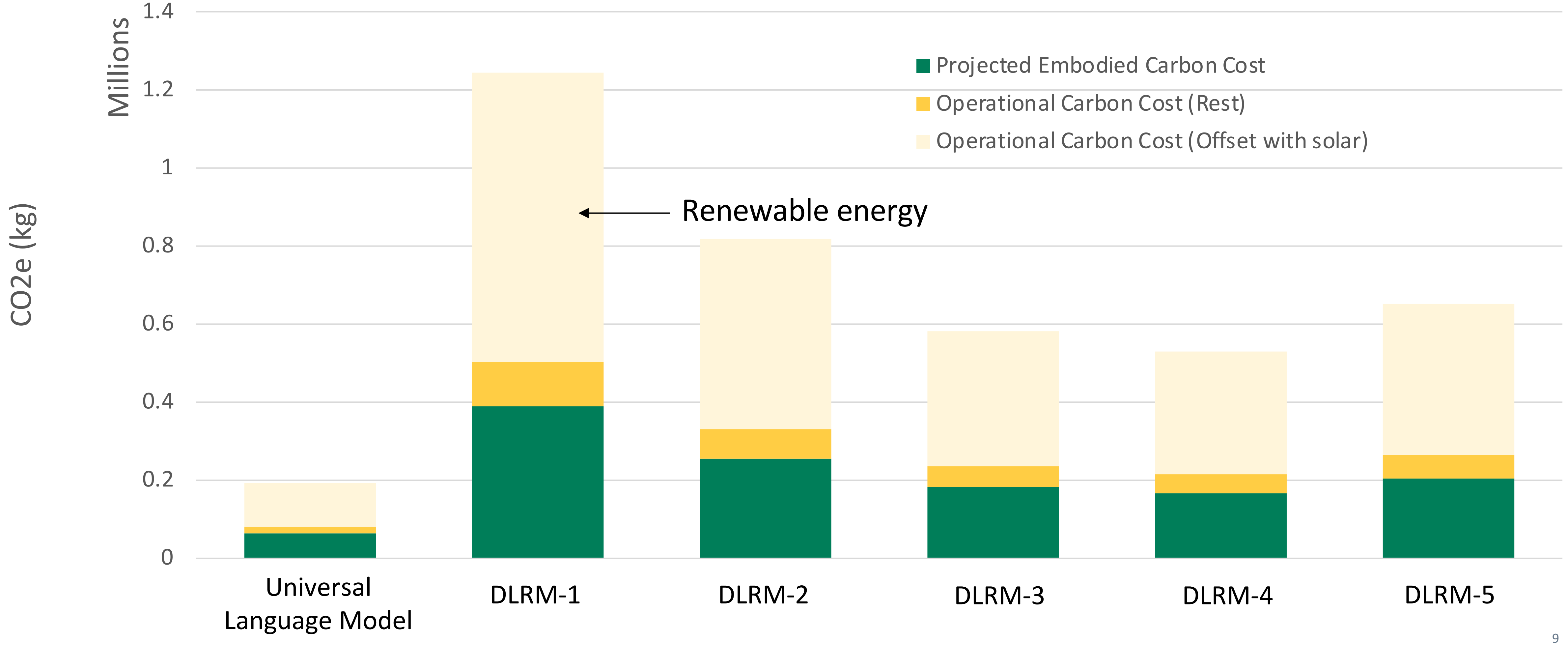
Recommendation Model Training

≈45 Home’s Annual

AI’s (Operational & Embodied) Carbon Footprint



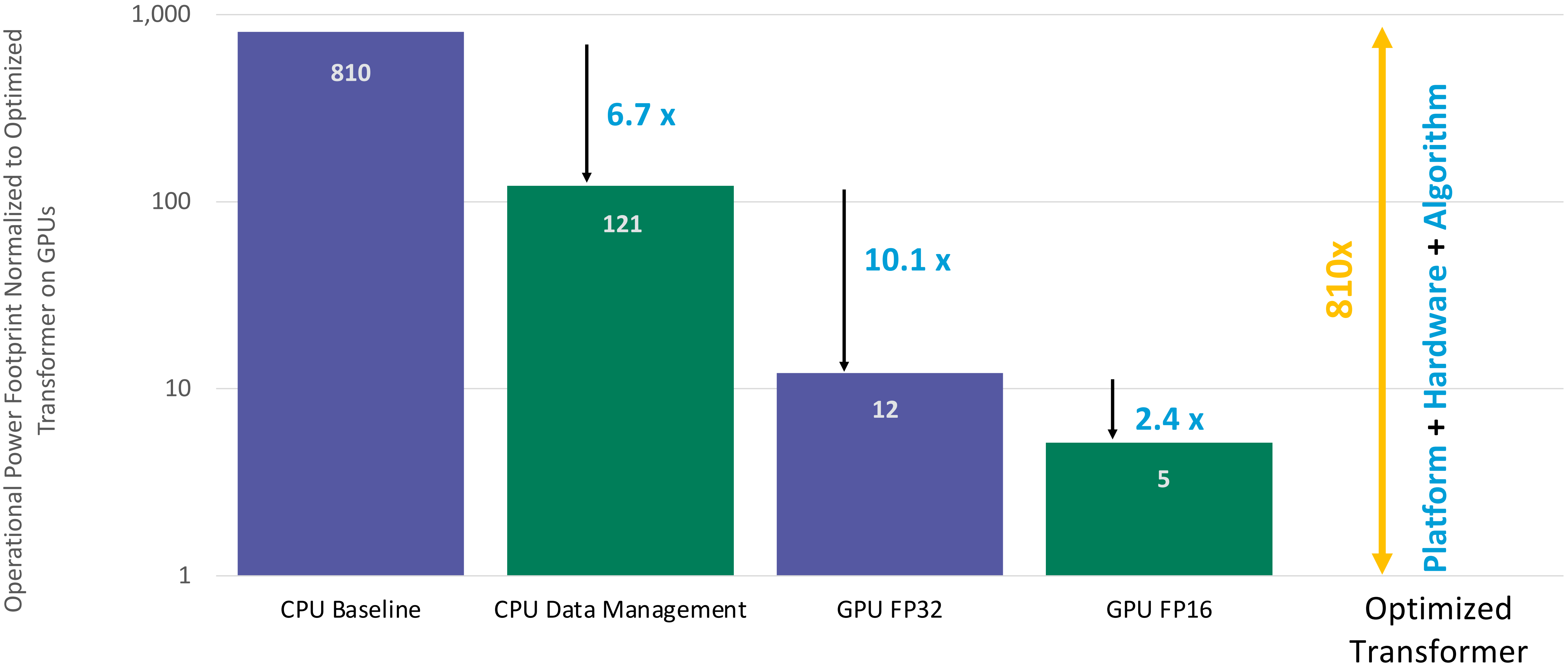
Manufacturing → Transport → Product Use → Recycling



Renewable energy

Carbon Optimization via HW-SW Co-Design

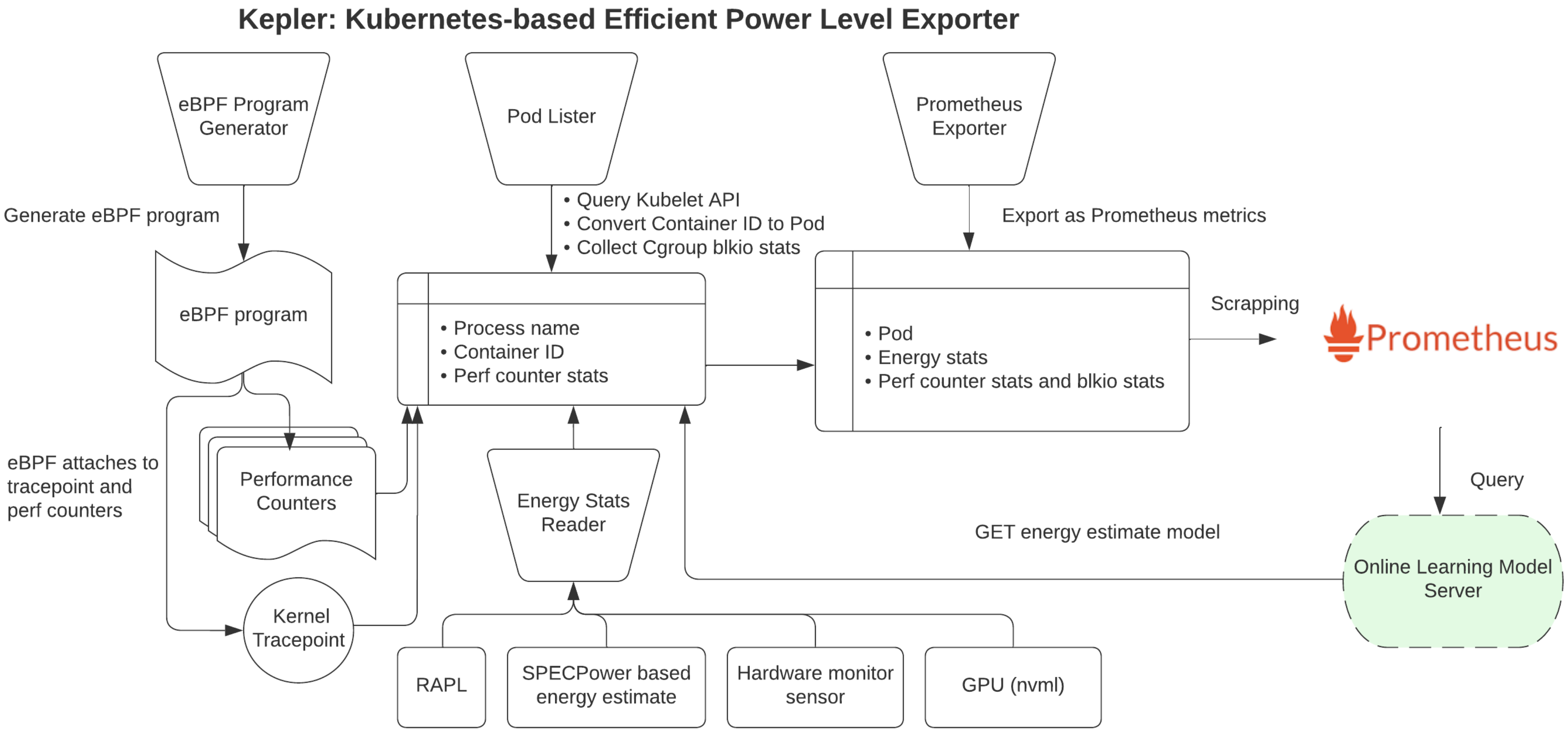
Universal Language Translation



Huamin Chen

RedHat

Kepler



Kepler: Kubernetes-based Efficient Power Level Exporter
eBPF Program Generator
Pod Lister
Prometheus Exporter
Generate eBPF program
• Query Kubelet API
• Convert Container ID to Pod
• Collect Cgroup blkio stats
Export as Prometheus metrics
eBPF program
• Process name
• Container ID
• Perf counter stats
• Pod
• Energy stats
• Perf counter stats and blkio stats
Scrapping
Prometheus
eBPF attaches to tracepoint and perf counters
Performance Counters
Energy Stats Reader
Query
GET energy estimate model
Online Learning Model Server
Kernel Tracepoint
RAPL
SPECPower based energy estimate
Hardware monitor sensor
GPU (nvml)

Andrew Chien

University of Chicago

What problem? Foundation models are a key LEVERAGE in reducing the Carbon Impact of Generative AI

Andrew A. Chien[1,2],

[1]University of Chicago [2]Argonne National Laboratory

All authors contributed equally

ACM HotCarbon '23,
Boston, USA

Training of Foundation Models is not the problem; Inference is the major sustainability problem

- Per our ChatGPT study (earlier today), for a successful foundation model (GPT-3), even one application is 25x the cost of one training
  - Inference already dominates
- 100x increase in use is coming, Slack, Msft Office, etc.
  - Moderate additional training
- Inference will really dominate for these applications
  - 25 x 100 => 2500x training ???

Business Balance and “Value engineering”

- Why did we go the moon in the 60’s and 70’s, and never go back? (until maybe 2025)
  - Investment was unsustainable, not supported by financial returns
  - Training cost higher than inference is financially unsustainable
- It makes no business sense to spend more to build a product, than can be earned back by its sales/use.
  - Foundational models that capture large volume use will be sustained, others will fail, and training in them will decline
  - Inference revenue must be greater than training cost, or the business is unsustainable
- Inference cost will dominate increasingly in the future, as the AI market matures.

Apollo 11, 1969

Artemis, 2025?

Could there be a case where Inference doesn’t dominate?

- For this to happen, there would have to be **“really high value inferences”**
  - So not that many inferences could have enough value to justify the cost of training

- Hmm…
  - Such applications could exist
  - Generative AI is not that application
    - Lots of wrong answers
    - Lost of low-value answers
    - ChatGPT does inferences for cheap, microcents

Summary

- Inference cost dominates; Inference carbon is the key problem
- Foundation models are not the problem, as their use reduces model Embodied carbon
  - Reducing and sharing training per application
- As unsustainable investment fades, Inference cost will dominate to an increasing degree

- => We should focus on and work on inference cost for foundation (and all) models

Sanjay Krishnan

University of Chicago

CHIDATA

My Research Group

Algorithmic and systems foundations for large-scale sensing.

Physical World

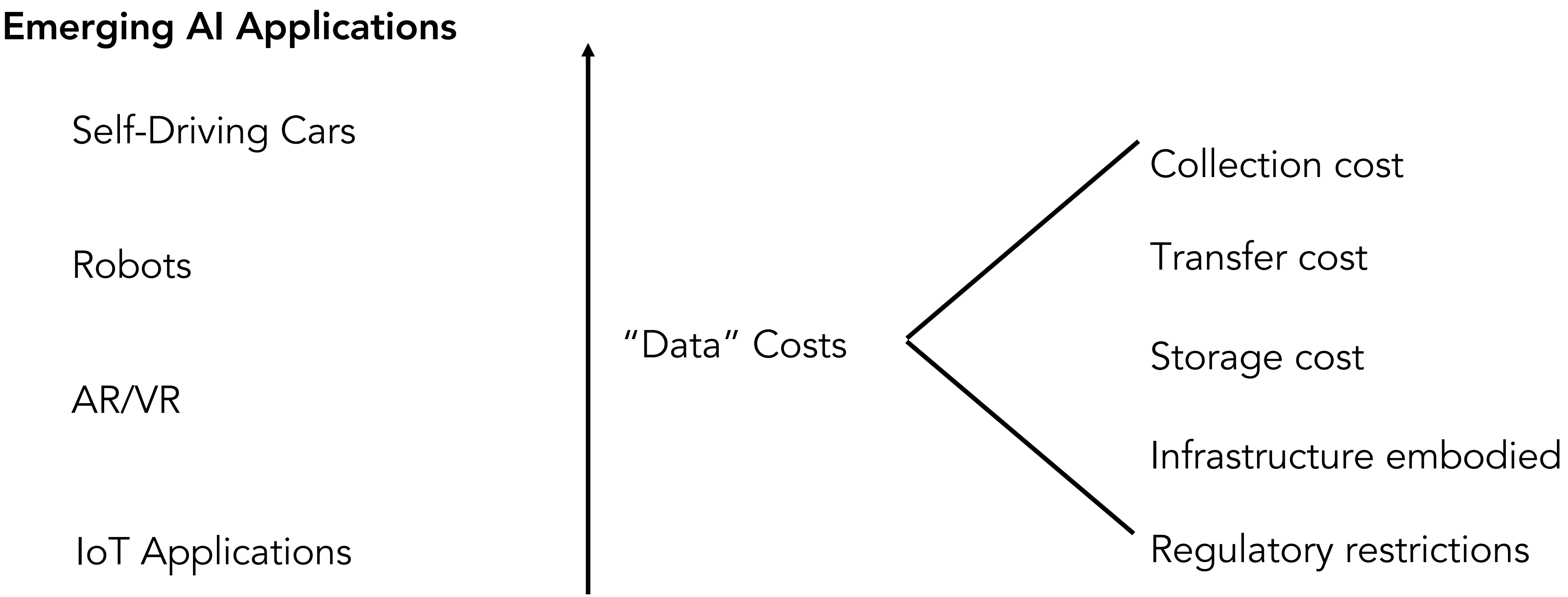
Digital World

Database/Machine Learning Group

“Simple” Research Question

What is the cost of data collection/transfer/storage in emerging AI applications?

Why is it important?



Emerging AI Applications

Self-Driving Cars

Robots

AR/VR

IoT Applications

“Data” Costs

Collection cost

Transfer cost

Storage cost

Infrastructure embodied

Regulatory restrictions

Carbon footprint of the data lifecycle will become a dominant factor.

Is the Problem Real? How serious?

(Increasing carbon footprint of using AI models)

Embodied Carbon Footprint

Operational Carbon Footprint

Would Standardization be Helpful?

Carbon Quantification

Accuracy/validation

Training Carbon Footprint

Inferencing Carbon Footprint

Other Carbon Footprint

(Data processing, fine-tuning)

Feasible HW and SW Solutions?
Research Directions?

Community Efforts?

Any Other Discussions?