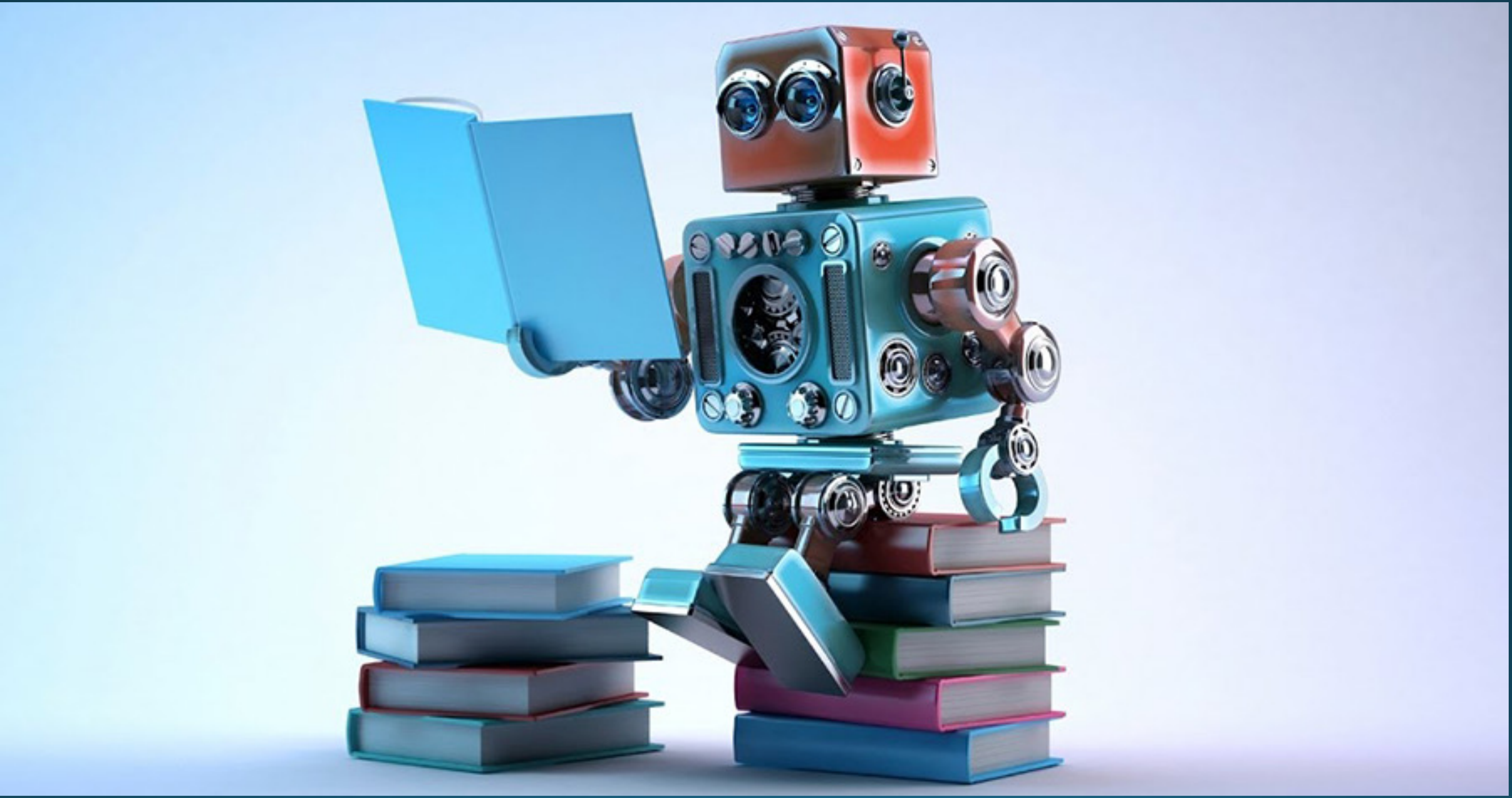
Environmental Implications, Challenges and Opportunities

Sustainable AI

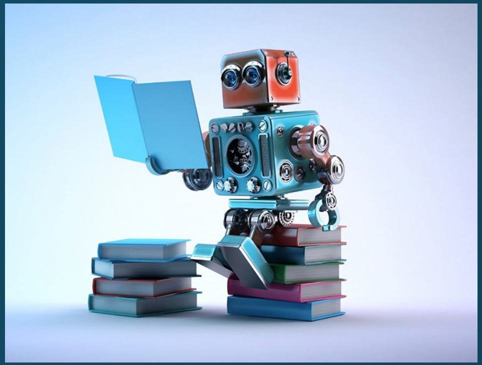
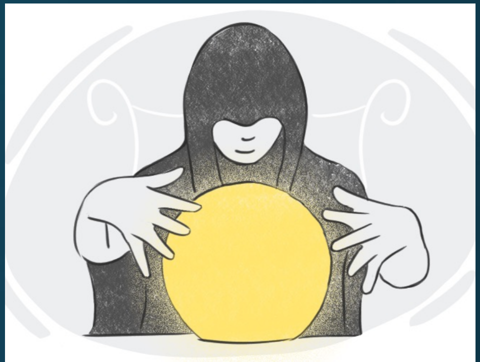
Elevator Pitch

The cost of AI



Training, a lot!



It’s a system of carbon emission.

Solution?

- ✓ Technical
- ✓ Design

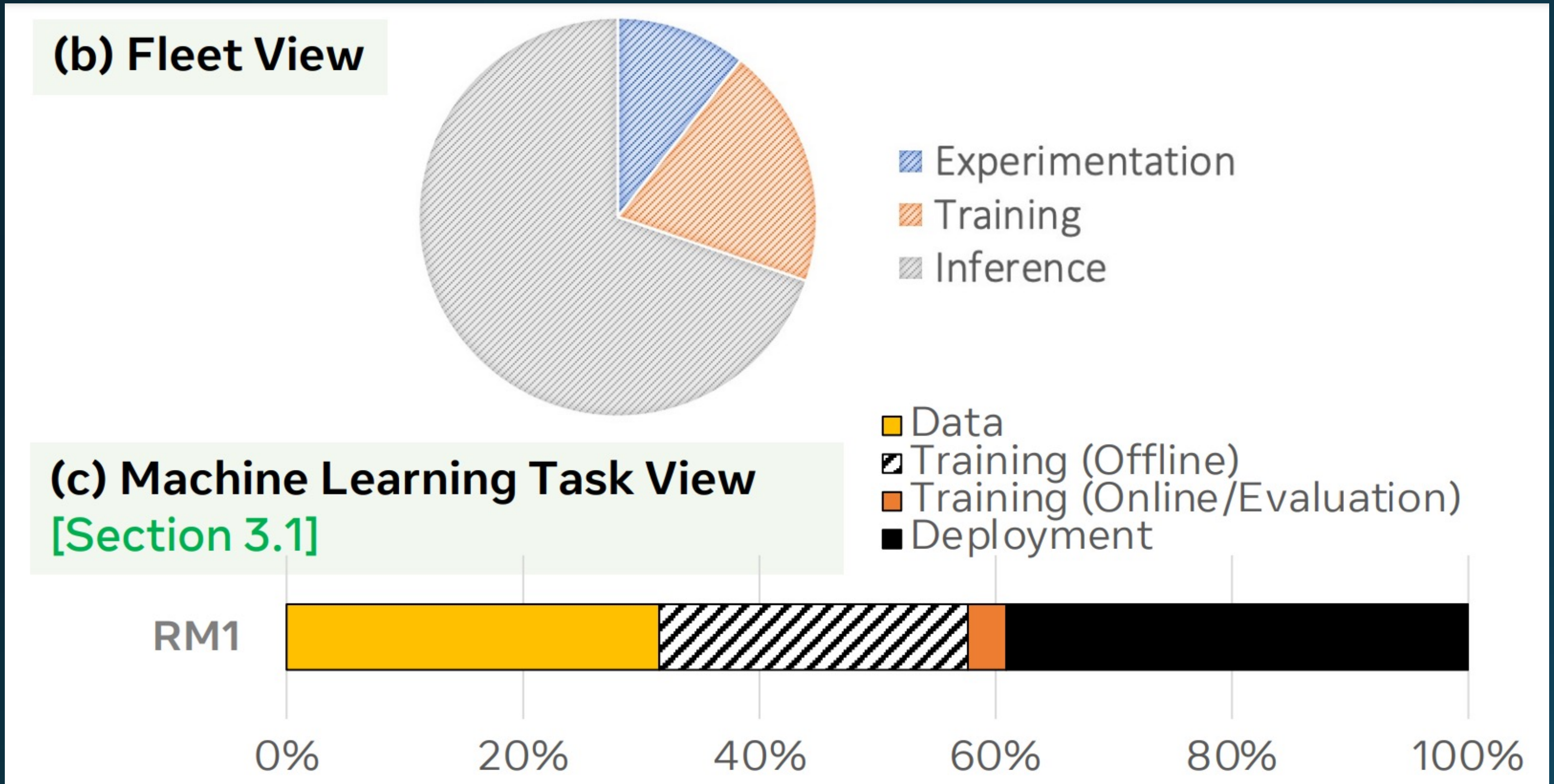
Write-up

Questions

Q1. What is the problem?

- A super-linear growth in AI
- Data, models, infrastructure

- **Limited knowledge for the environmental impact holistically**



Energy cost break-down

Q2. Insights?

- Several phases for ML dev;
- Different energy cost:
  - Inference is the worst.

Q3. What is the solution?

1. Optimize the energy efficiency and reducing carbon footprint:
    - Model, platform, infrastructure, hardware.

2. Efficient designs with a sustainability mindset:
    - Data, experimentation, system utilization, telemetry.

Q4. What is the takeaway message?

- Big environmental cost for ML development;
- Solutions need to be holistic.

Q5. test of time award?

- Don't think so.
  - But a good synthesis of a few existing ideas;
- Qualitatively, not enough theoretical insights;
- Quantitatively, need in-depth analysis for problem size and the effectiveness of the solution.

Q6. Accept or Reject?

- Accept at a conference.
- However,
  - close to social science research
  - not enough depth for a top journal publication for social science.

Q & A

Thank you

Zeus: Understanding and Optimizing GPU Energy Consumption of DNN Training

Shanqing Lin

March 21st, 2022

PARALLEL SYSTEMS ARCHITECTURE LAB

EPFL

Elevator Pitch

- Power consumption of the DNN training should be considered
- Zeus
  - Trade-off power and training time (saving up to 70% energy consumption)
  - Automatic and transparent
  - Adaptive to various types of DNN

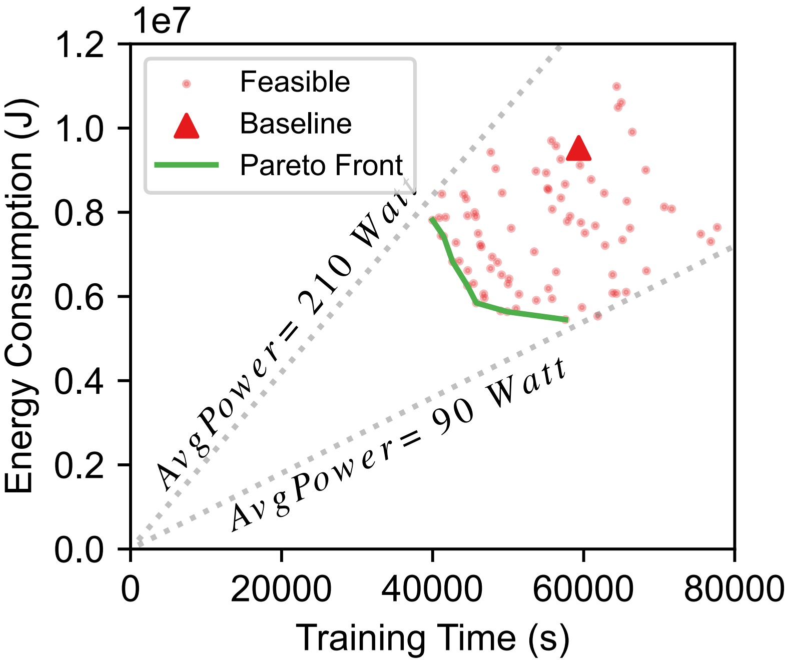
Problem

PARALLEL SYSTEMS ARCHITECTURE LAB

- DNN training is energy-hungry
- Training parameters are selected unaware of the energy efficiency
  - Maximize training throughput
  - Or, follow the setting suggested by the original paper

Insight: Opportunity

Default parameters are far away from pareto frontier.



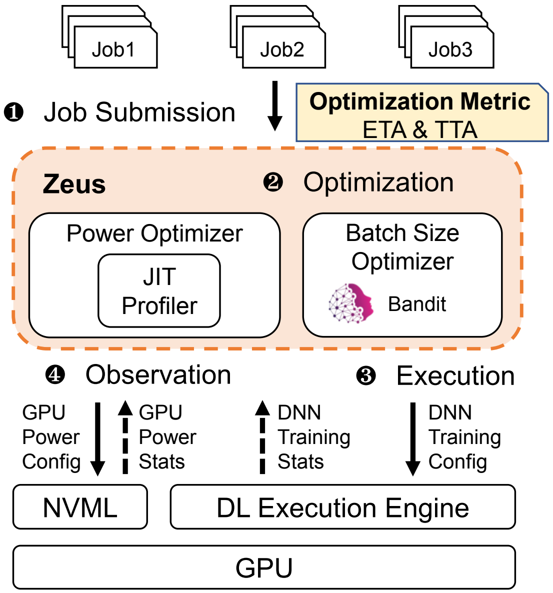
Implicit opportunity: training does not have tight latency requirement as inference.

Insight: Challenge

- Training energy is hard to model accurately
  - Highly sensitive with workloads, dataset, and hardware
  - Stochastic training process
- Opportunity: The same network is trained repeatedly
  - Reason: New data is added to update model weights
  - Implication: Profiles are sufficient

High-level Solutions: Feedback

- Profile GPU power online
- Predict the optimal GPU power limit and batch size



Technical Solutions

PARALLEL SYSTEMS ARCHITECTURE LAB

- Separate the optimization goal to Epochs(b), Throughput(b, l), and Power(b, l)
  - Throughput and power can be profiled real time
  - Stable to batch size, workload, and hardware
  - Optimal GPU power limit can be predicted using batch size, without chaos
- Model the Epochs(b)
  - Multi-armed bandit (Reinforcement Learning) with gaussion distribution belief
  - Thompson sampling: accelerate convergence
- Early stopping to avoid struggling

Takeaway Message

- ML research with awareness of energy consumption
- Stochastic effects in ML can be modeled by RL, with enough profiles
- Feedback can be applied to solve optimization problem

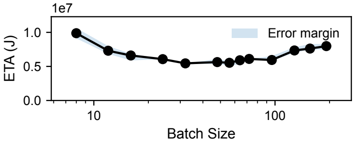
Test-of-time Award?

- Yes
- Awareness of energy
- General solutions for other optimization
- Modeling stochastic training process with RL

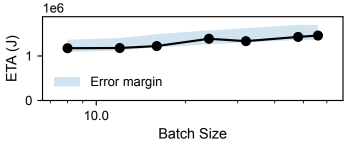
Non-acceptance Reason

PARALLEL SYSTEMS ARCHITECTURE LAB

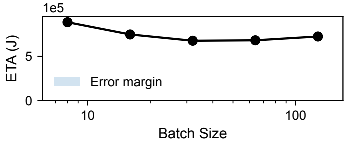
No explanation to the opportunity!



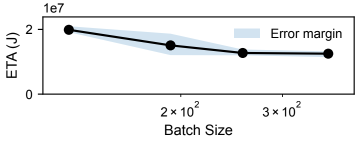
(a) DeepSpeech2



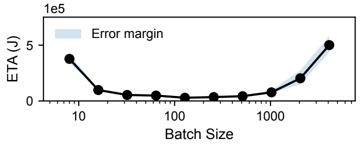
(b) BERT (QA)



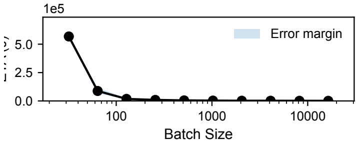
(c) BERT (SA)



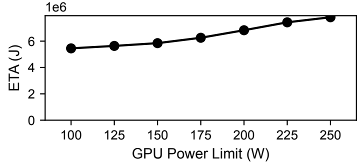
(d) ResNet-50



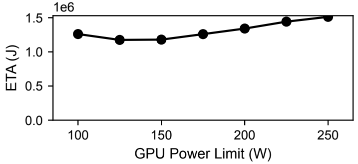
(e) ShuffeNet V2



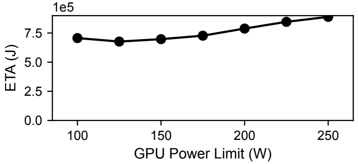
(f) NeuMF



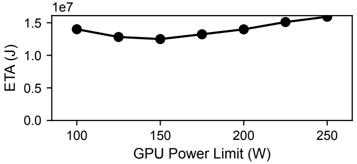
(a) DeepSpeech2



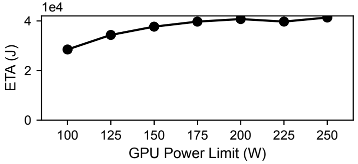
(b) BERT (QA)



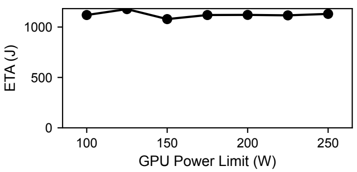
(c) BERT (SA)



(d) ResNet-50



(e) ShuffeNet V2



(f) NeuMF

Thank You!

For more information, please visit us at

parsa.epfl.ch