Using AI to improve AI Development and Deployment

Vijay Gadepally
Lincoln Laboratory Supercomputing Center

MIT LINCOLN LABORATORY
SUPERCOMPUTING CENTER

Acknowledgements


Siddharth Samsi


David Bestor


Nathan Frey


Mike Jones


Joseph McDonald


Matthew Weiss

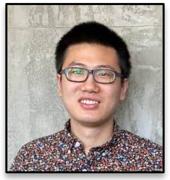

Daniel Edelman

U.S. AIR FORCE


Maj. Andrew Bowne


Capt. Lindsey McEvoy


Prof. Devesh Tewari


Baolin Li

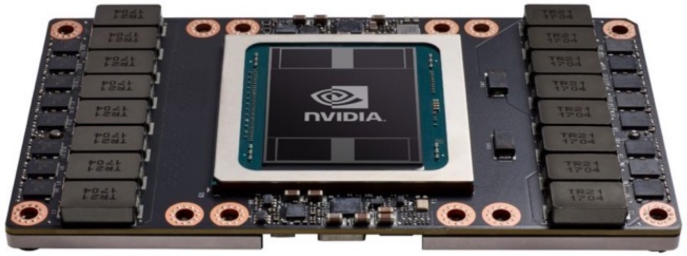
And many others…

- Charles Leiserson (CSAIL)
- Tim Kraska (CSAIL)
- Manya Ghobadi (CSAIL)
- Sam Madden (CSAIL)
- Mike Stonebraker (CSAIL)
- T.B. Schardl (CSAIL)
- Anson Cheng (USAF)
- Allan Vanterpool (USAF)
- Andrew Kirby (Alum)
- Emily Do (Alum)
- Matthew Hutchinson (Alum)
- William Arcand
- William Bergeron
- Chansup Byun
- Matthew Hubbell
- Michael Houle
- Hayden Jananthan
- Jeremy Kepner
- Anna Klein
- Peter Michaleas
- Lauren Milechin
- Julie Mullen
- Andrew Prout
- Albert Reuther
- Antonio Rosa
- Pat Ross
- Charles Yee
- Stephen Rejto
- Jeff Gottschalk
- Marc Zissman
- Dave Martinez
- Mark Veillette
- Bob Bond
- Jason Williams
- Brad Dillman
- Daniela Rus
- Col. Garry Floyd
- CK Prothmann

MIT Lincoln Laboratory Supercomputing Center

Low Carbon Emission

- **Significant increase in computing power for simulation, data analysis, and machine learning**
- **Critical computing power for simulation, data analysis, and machine learning**



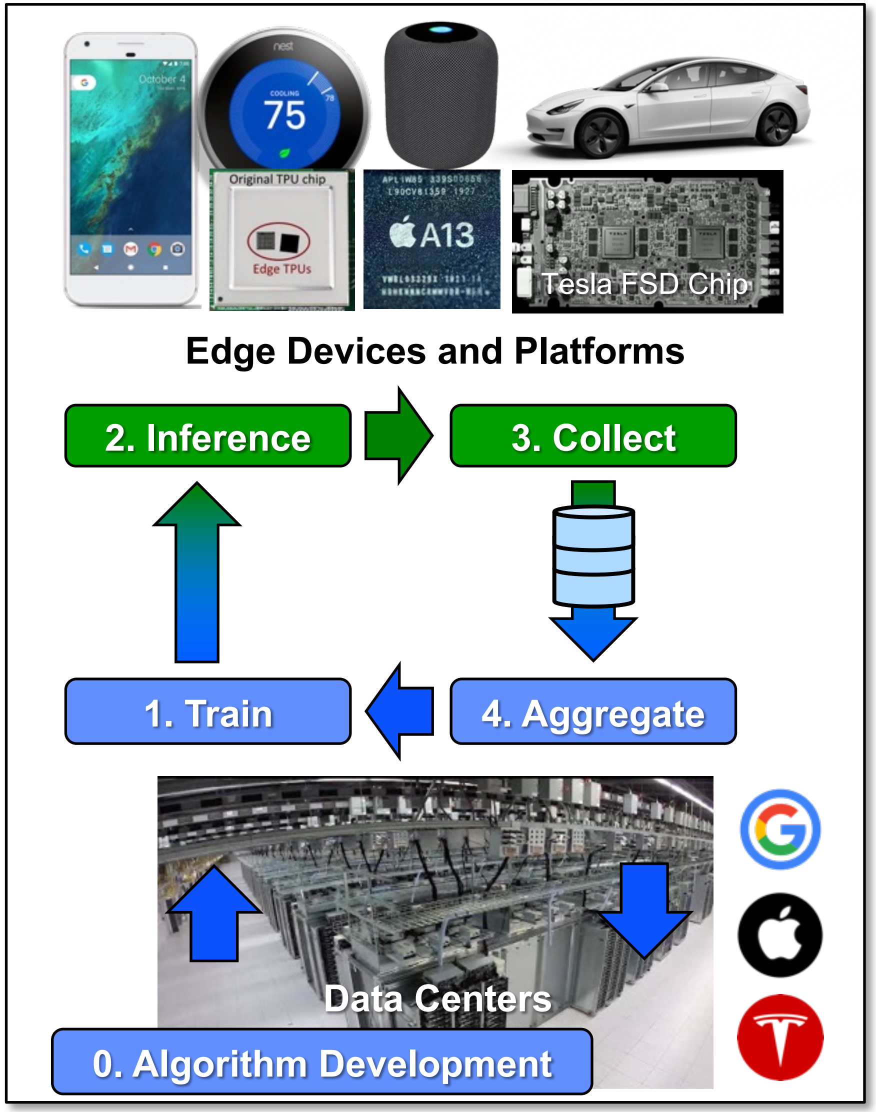
- **<u>Operates on renewable energy</u>**

	Capability
Processor	Intel Xeon & Nvidia Volta
Total Cores	737,000
Peak	7.4 Petaflops
Top500	5.2 Petaflops
Memory	172 Terabytes
Peak AI Flops	100+ Petaflops
Network Link	Intel OmniPath 25 GB/s

*Based on 2020 Top500.org
AI Flops = 4x4 matrix multiply half precision in, single precision out (mixed precision training)

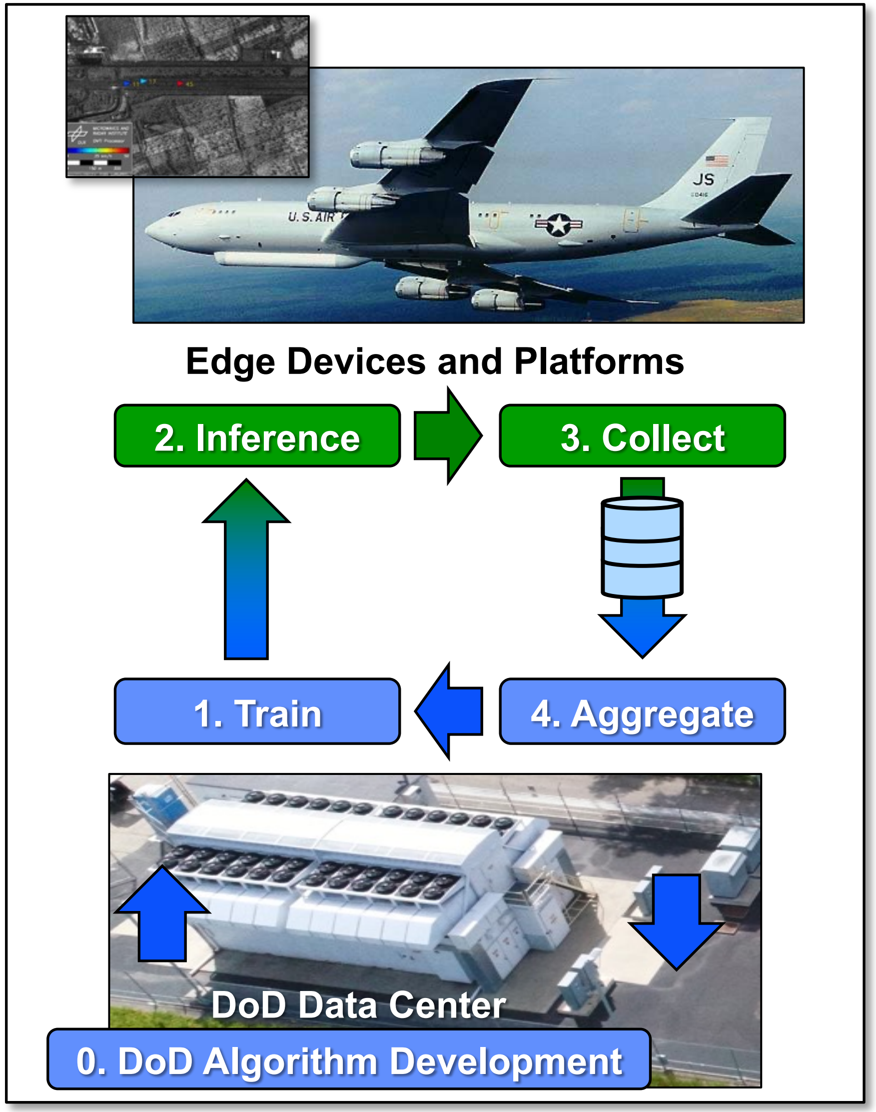
AI Development vs Deployment

Industry



Edge Computing (Deployment)

Data-Center Computing (Development)



Defense

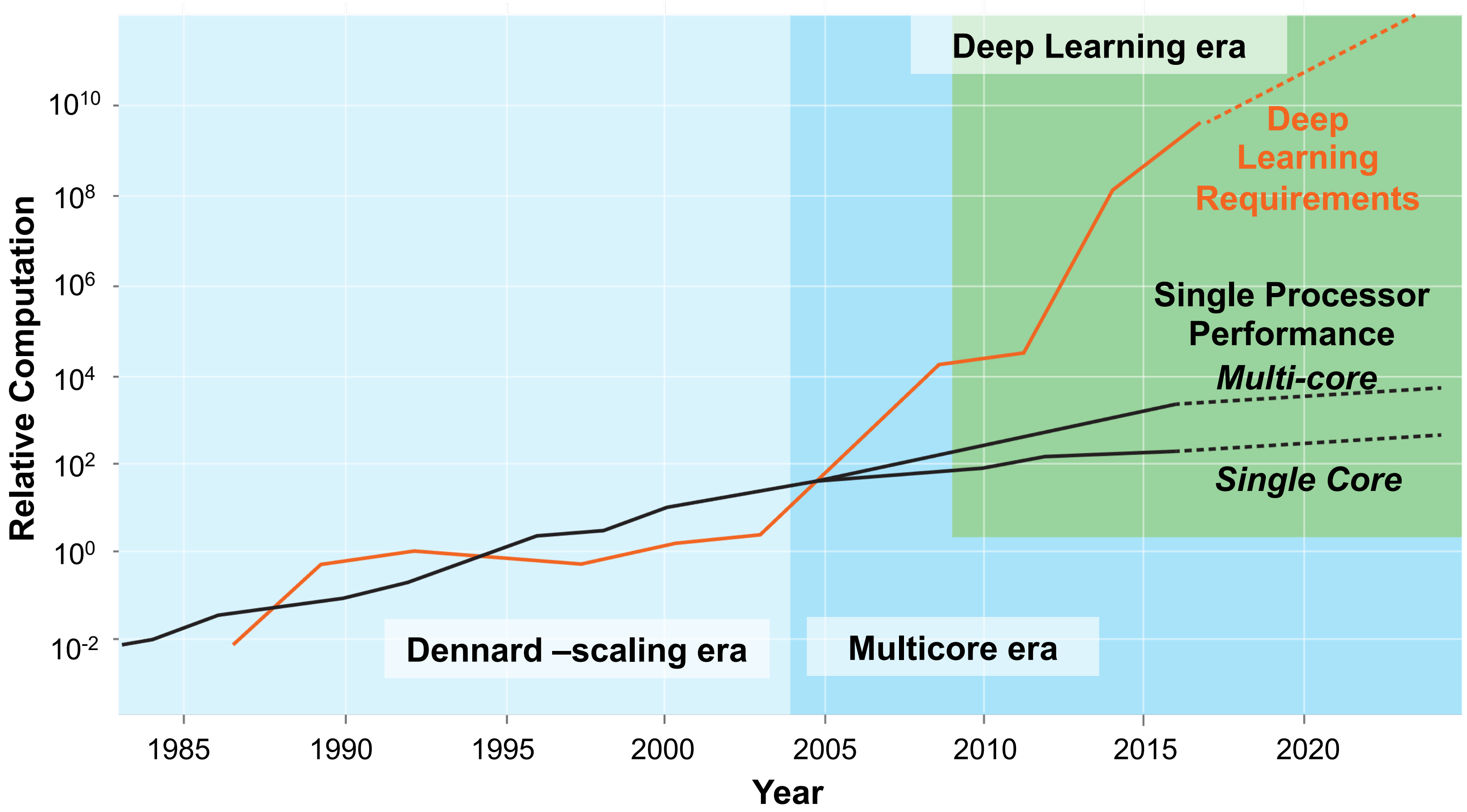
A Few Trends and Observations

Challenges	Trends and Observations
Computing Performance	Large number of many-cores; concurrency and locality; instruction level parallelism | Moore's law dead: Power and memory walls; clock rate limitations
Hardware Platforms	Domain Specific Accelerators: Heterogenous edge computing; legacy hardware solutions
Power and Energy	Unsustainable energy requirements: Power and energy walls; growing environmental impact
New Application Areas	Unknown requirements: new applications face "new" problems (e.g., seamless transition between development and deployment)
Research and Development	Education: Lack of trained computing engineers; Research: difficult to collect and develop solutions based on real data

Top-level trends:

- **Renewed resurgence of HPC solutions to power AI and research innovations**
- **Need for "seamless" transition between HPC and Deployment (Edge) environments**

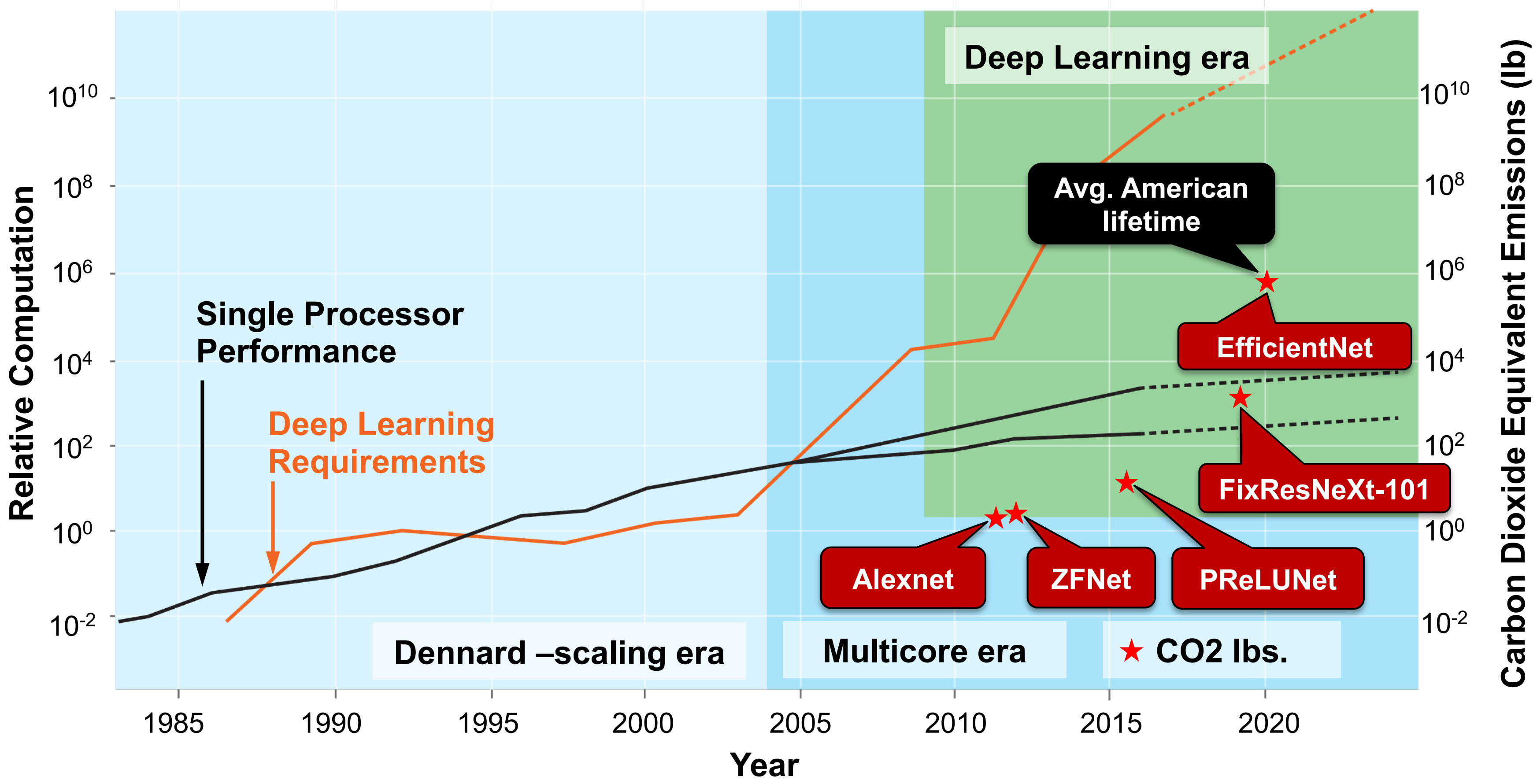
Trend 1: AI Development Computing Requirements Gap



Need for tools that bridge computing gap

Source: Neil Thompson, MIT CSAIL

Corollary Trend 1: Major Source of Carbon Emissions

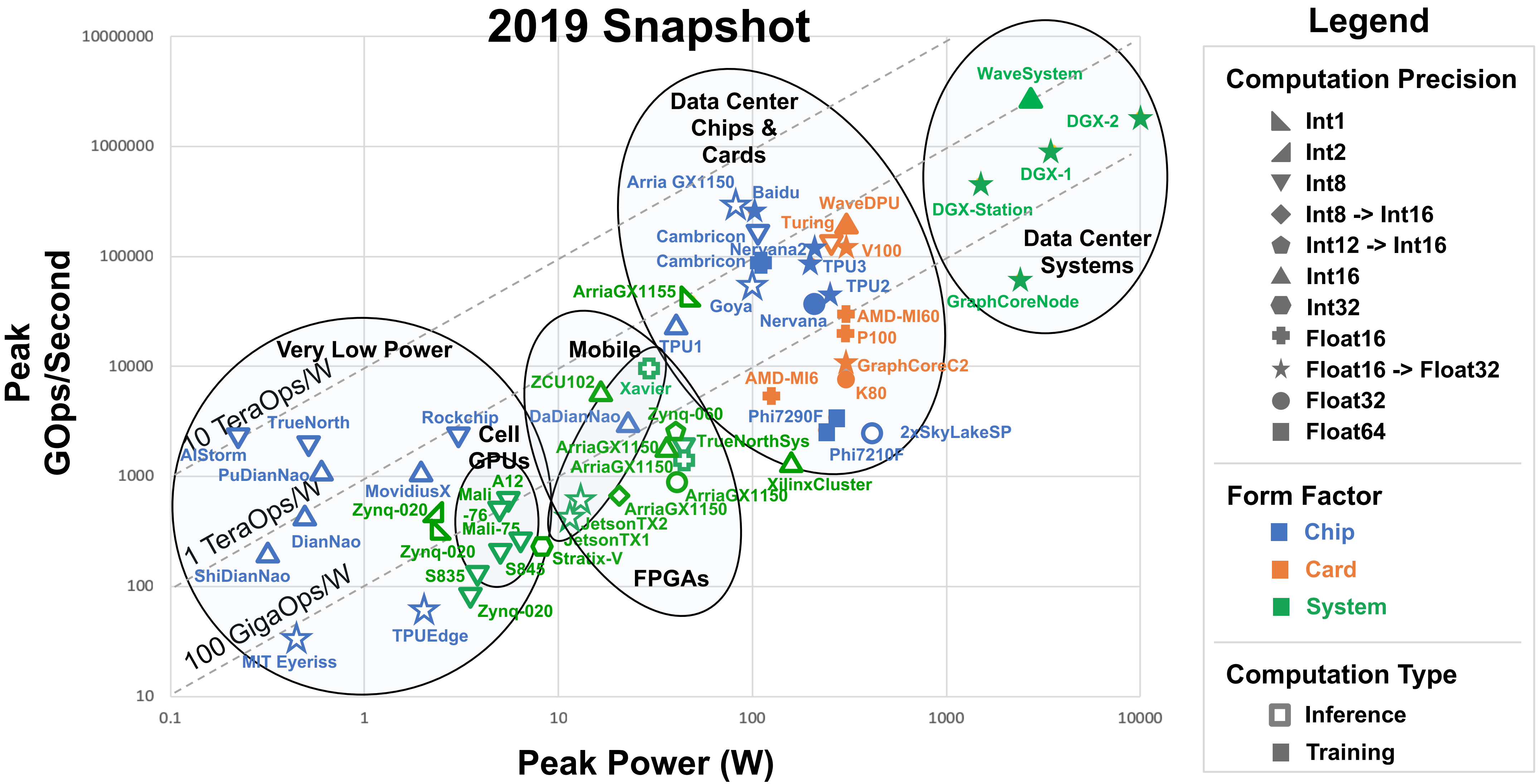


Deep learning energy requirements are growing unsustainably

[1] Thompson, Neil C., Kristjan Greenewald, Keeheon Lee, and Gabriel F. Manso. 2021. Deep Learning's Diminishing Returns: The Cost of Improvement is Becoming Unsustainable. IEEE Spectrum.

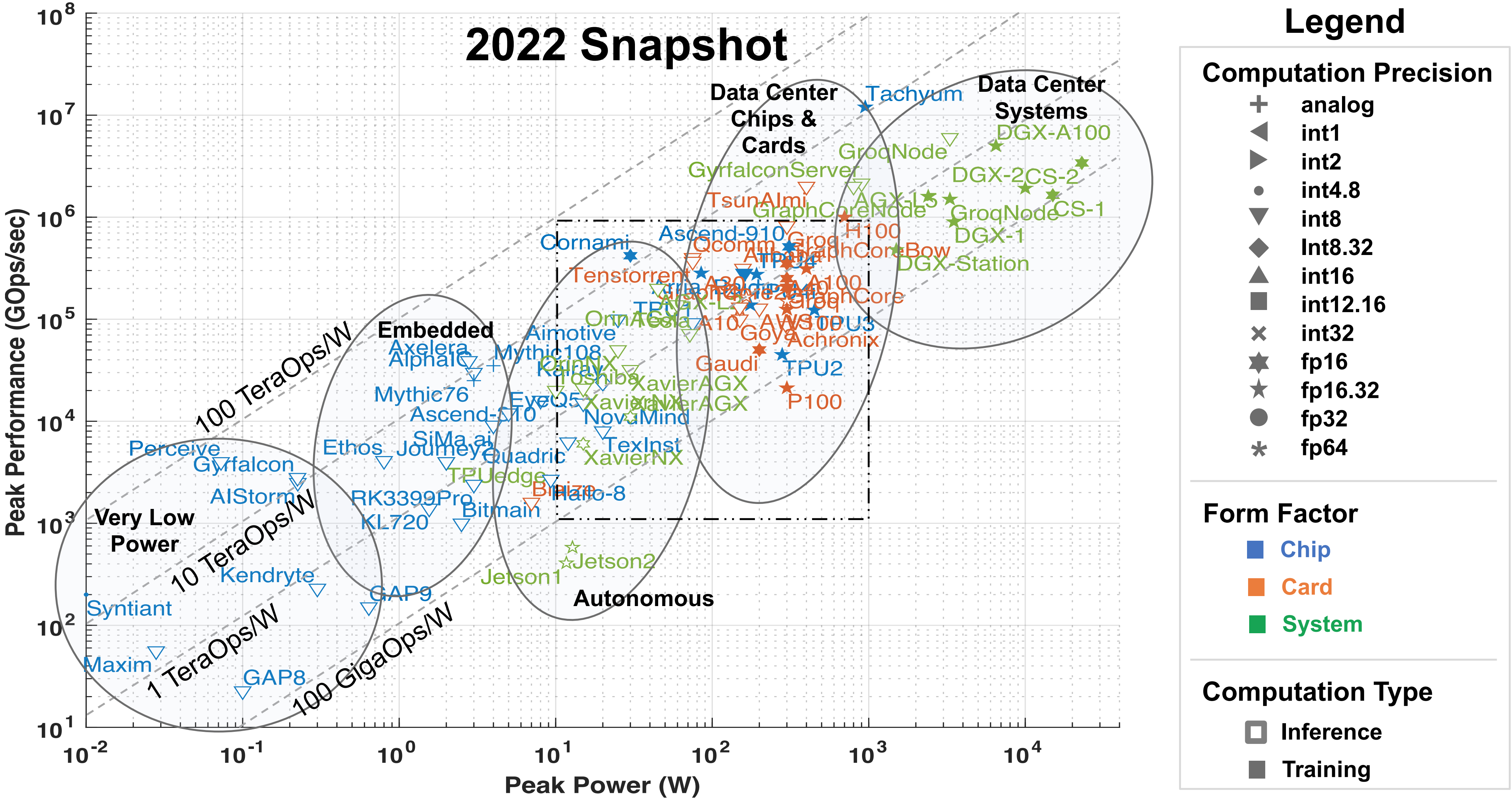
[2] The Energy and Carbon Footprint of Training End-to-End Speech Recognizers - Parcollet, T., & Ravanelli, M., 2021

Trend 2: Growing diversity of ML Accelerators



A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi and J. Kepner, "Survey and Benchmarking of Machine Learning Accelerators," *2019 IEEE High Performance Extreme Computing Conference (HPEC)*, 2019, pp. 1-10.

Trend 2: Growing diversity of ML Accelerators



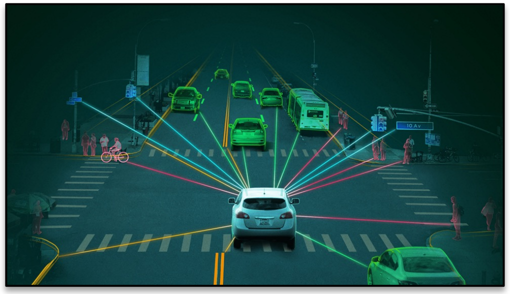
A. Reuther, P. Michaleas, M. Jones, V. Gadepally, S. Samsi and J. Kepner, "AI and ML Accelerator Survey and Trends," *2022 IEEE High Performance Extreme Computing Conference (HPEC)*, 2022, pp. 1-10.

Trend 3: Emerging Application Domains

Health Care

- Correlate data across millions of patients
- Evidence Based Medicine
- Data from different modalities
    - Image
    - Video
    - Signal
    - Text
    - ...

Transportation

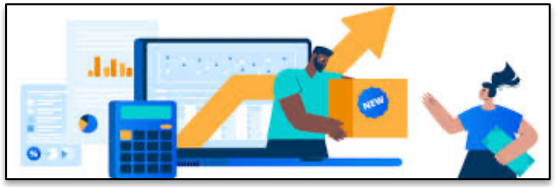


- Billions of vehicles
- Need to correlate high rate information from different vehicles
- More sensors -> More problems
- Can be used to improve quality of transportation systems

IoT/Smart XYZ

- Billions of small “edge” connected devices across homes, cities, countries, …
- Need to identify patterns of living and correlate for improved efficiency and safety

Retail



- Sell you things better, supply chain management, inventory management
- Dozens of existing enterprise systems connected to numerous management systems (credit card processing, FedEx, ...)

Application Example: Autonomous Vehicles

In-Vehicle data processing

Com = Communication
GPS = Global Positioning System
IMU = Inertial Measurement Unit

V2V: Vehicle to Vehicle
V2I: Vehicle to Infrastructure

Example Autonomous Vehicle Data Feeds and Speeds

Sensor Type	Frequency	Data rate	Data type
Lidar	10 Hz	8 MBps	Point cloud
Lower-res Lidar (x4)	55 Hz	1 MBps	Point cloud
Lower-res Camera (x4)	20 Hz	4 MBps	JPEG frames
High-res Camera	4 Hz	1 MBps	JPEG frames
CAN bus	900 Hz	50 KBps	Custom struct
IMU	50 Hz	30 KBps	Custom struct
Compass	100 Hz	10 KBps	Custom struct
GPS	6 Hz	$<$ 1 KBps	Custom struct

(40 min trip -> 30 GB Sensor Log)

Emerging applications have developing hardware requirements

“A framework for estimating driver decisions near intersections,” Gadepally, et. al., IEEE Transactions on Intelligent Transportation Systems, 2014

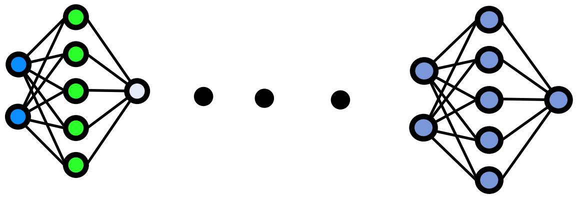
“Exploring big volume sensor data with Vroom,” Moll, et. al., VLDB 2017

Outline

- Motivation
- Reducing development computing demands
- Finding the right deployment environment
- Datacenter Challenge
- Summary and Air Force Perspective

Reducing Development Environment Computing Demands

Model Development



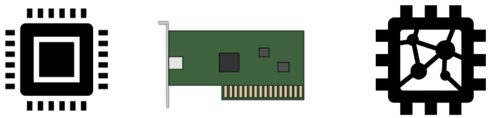
Challenge

- Model design, testing, and development
- AI training & inference

Proposed Approach

- AI-enabled Model Discovery[1]
- Knowledge Informed Models

Hardware Usage Strategies



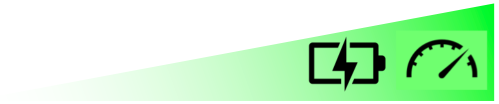
Challenge

- Hardware variety
- Matching workload needs to hardware capabilities

Proposed Approach

- Hardware-based interventions
- ML-based hardware selection[2]

Performance & Energy Tuning



Challenge

- Hardware power modulation

Proposed Approach

- Power limiting[3]
- Clock frequency scaling[3]
- Auto-tuning complex applications[4]

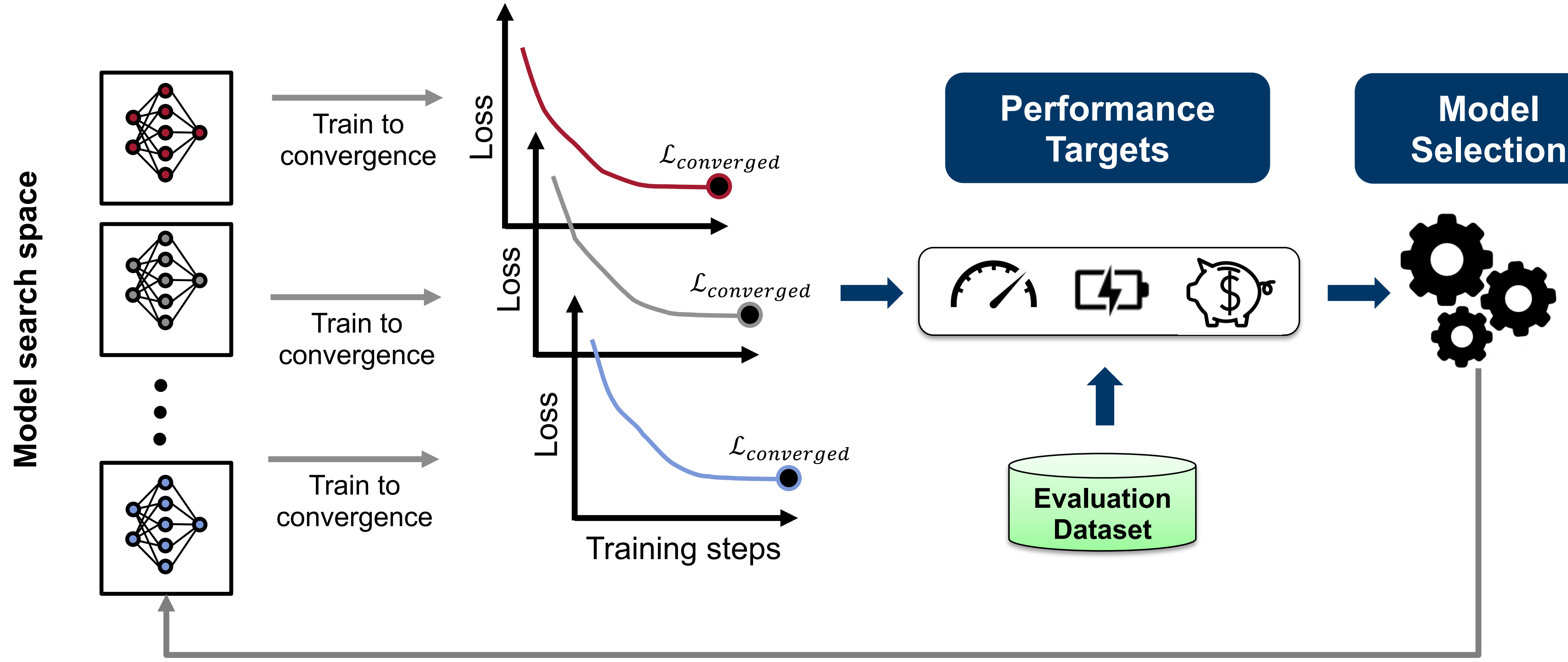
[1]Neural Scaling of Deep Chemical Models – Frey, et. al, *Nature Machine Intelligence (submitted)*

[2] DASH: Scheduling Deep Learning Workloads on Multi-Generational GPU-Accelerated Clusters – Li, et. al., IEEE HPEC 2022

[3] Great Power, Great Responsibility: Recommendations for Reducing Energy for Training Language Models – McDonald, et. al., NAACL 2022

[4] Bliss: auto-tuning complex applications using a pool of diverse lightweight learning models – Roy, et. al., *PLDI 2021*

AI-enabled Model Discovery: Neural Architecture Search and Hyperparameter Optimization



Architecture searches and parameter optimization has significant compute requirements

[1] Energy-aware neural architecture selection and hyperparameter optimization – Frey, et. al,, *IEEE IPDPS ADOPT 2022*
[2] Neural Scaling of Deep Chemical Models – Frey, et. al, *Nature Machine Intelligence (submitted)*

Modeling performance: training speed estimation (TSE)

How do we speed up *time to performance* for new models and datasets?

Area under training loss curve

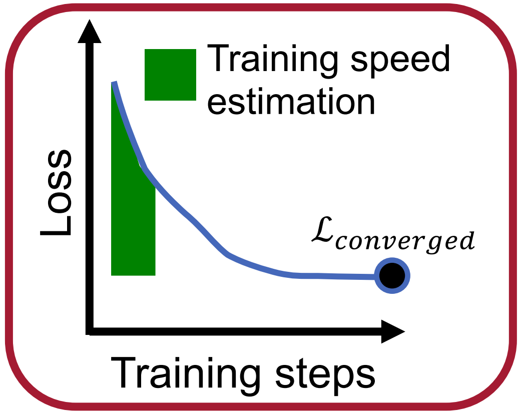
$$\mathrm{TSE} = \sum_{t=1}^{T} \left[\frac{1}{B} \sum_{i=1}^{B} \ell \left(f_{\theta_{t,i}}(\mathbf{X}_i), \mathbf{y}_i \right) \right]$$

Features and Associated Labels

Neural Network

Number of Completed Epochs

Loss Function

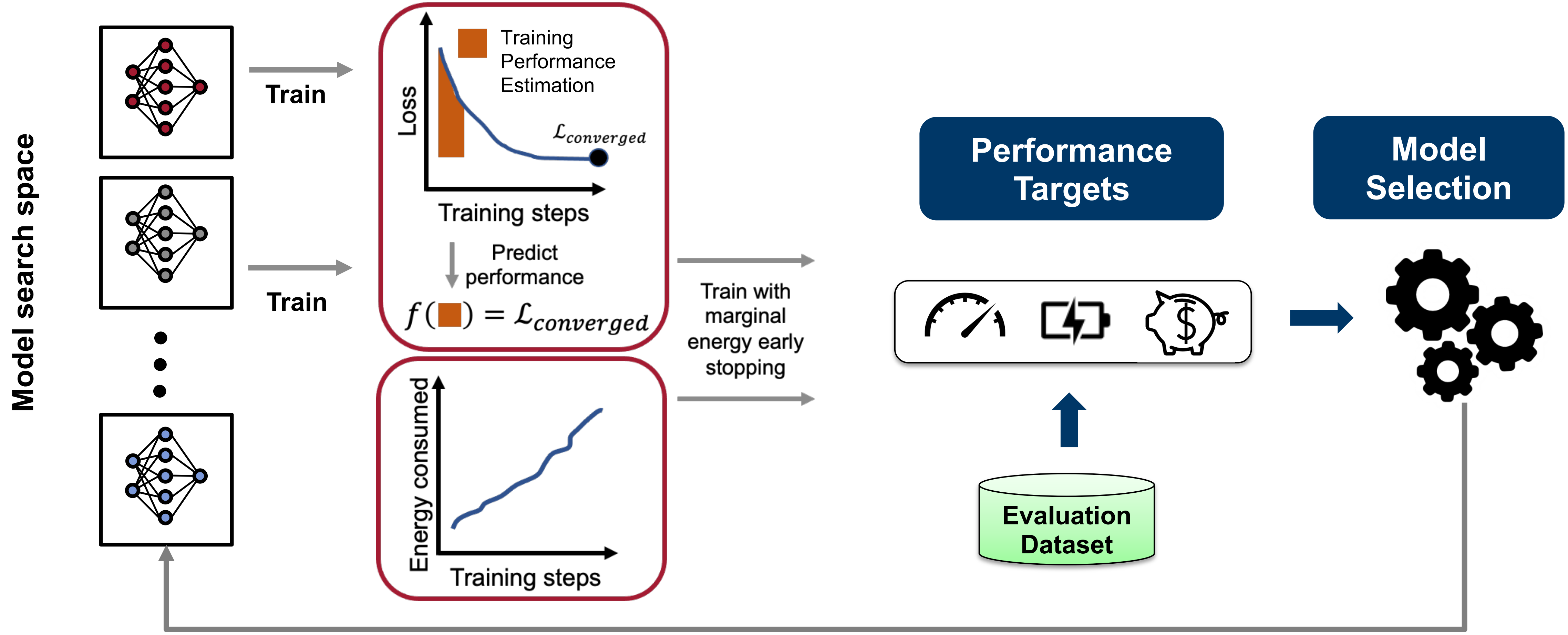


- **TSE is a simple, efficient, computationally cheap method for neural architecture search and hyper-parameter optimization**

Ru, Robin, et al. "Speedy Performance Estimation for Neural Architecture Search." *Advances in Neural Information Processing Systems* 34 (2021).

Neural Scaling of Deep Chemical Models – Frey, et. al, *Nature Machine Intelligence (submitted)*

Training Performance Estimator (TPE) for Efficient Neural Architecture Search and Hyperparameter Optimization

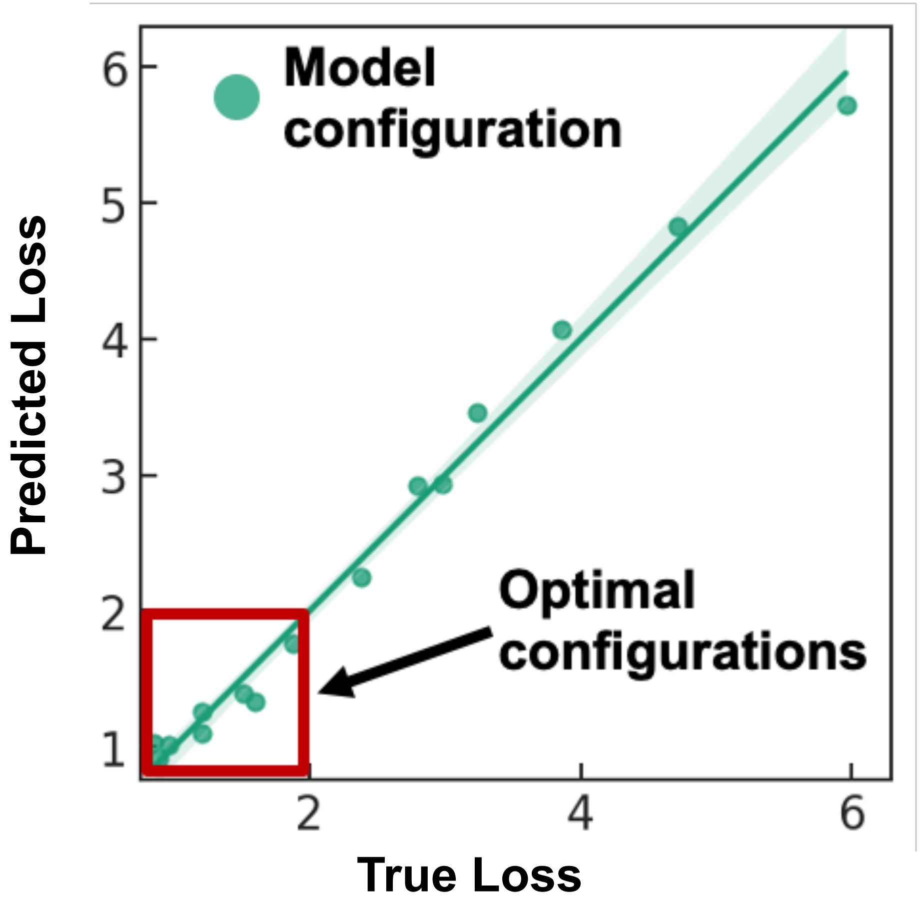


Training performance estimation (TPE) combines training speed estimation and energy consumption tracking to minimize energy expenditure

[1] Energy-aware neural architecture selection and hyperparameter optimization – Frey, et. al,, *IEEE IPDPS ADOPT 2022*
[2] Neural Scaling of Deep Chemical Models – Frey, et. al, *Nature Machine Intelligence (submitted)*

Neural Architecture Optimization for GNNs

Predicted Model Performance for SchNet[3]



***80%* total computing savings with early identification of optimal training configurations**

[1] Energy-aware neural architecture selection and hyperparameter optimization – Frey, et. al,, *IEEE IPDPS ADOPT 2022*
[2] Neural Scaling of Deep Chemical Models – Frey, et. al, *Nature Machine Intelligence (submitted)*
[3] Schnet: A continuous-filter convolutional neural network for modeling quantum interactions, Schutt, et. al, *NeurIPS 2017*

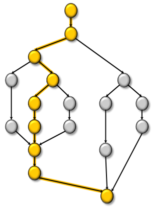
Outline

- Motivation
- Reducing development computing demands
- Finding the right deployment environment
- Datacenter Challenge
- Summary and Air Force Perspective

A Few AI Deployment Challenges…

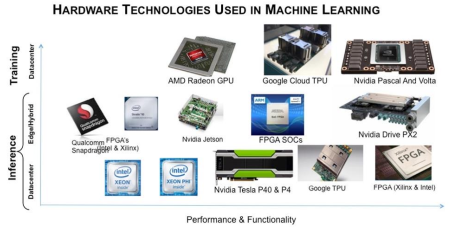
Challenge

Compilers/Middleware



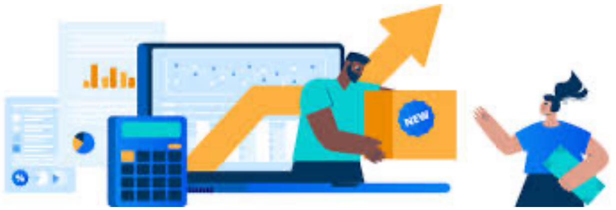
- **Inefficient AI middleware**
- **Particularly with newer hardware platforms**

Hardware Capabilities



- **Huge spectrum of capabilities**
- **Changing Mission Needs**

Application Demands



- **Dynamic Requirements**
- **Transitioning between "datacenter" and "edge"**

Proposed Approach

- **TapirXLA[1] Compiler for Tensorflow**
- **AI-enabled auto-tuning and workflow scheduling[2]**
- **RIBBON[3]: Leveraging heterogenous computing for dynamic**

[1] TapirXLA: Embedding fork-join parallelism into the XLA compiler in Tensorflow using tapir. – Schardl, Samsi, *IEEE HPEC 2019*.

[2] Mashup: making serverless computing useful for HPC workflows via hybrid execution – Roy, et al., *PPoPP 2022*

[3] RIBBON: cost-effective and qos-aware deep learning model inference using a diverse pool of cloud computing instances – Li, et al., *SC 2021*

Serving Inference Queries Under Evolving Requirements

DoD Mission

Task
- Inference
- Recommendation System
- Time series analysis

Constraints
- Minimize latency
- Maximize throughput
- Minimum QoS
- Minimize Energy

Hardware
- CPU
- GPU
- TPU
- …
- Custom hardware

Optimize

AI Recommendation

Datacenter/Secure cloud

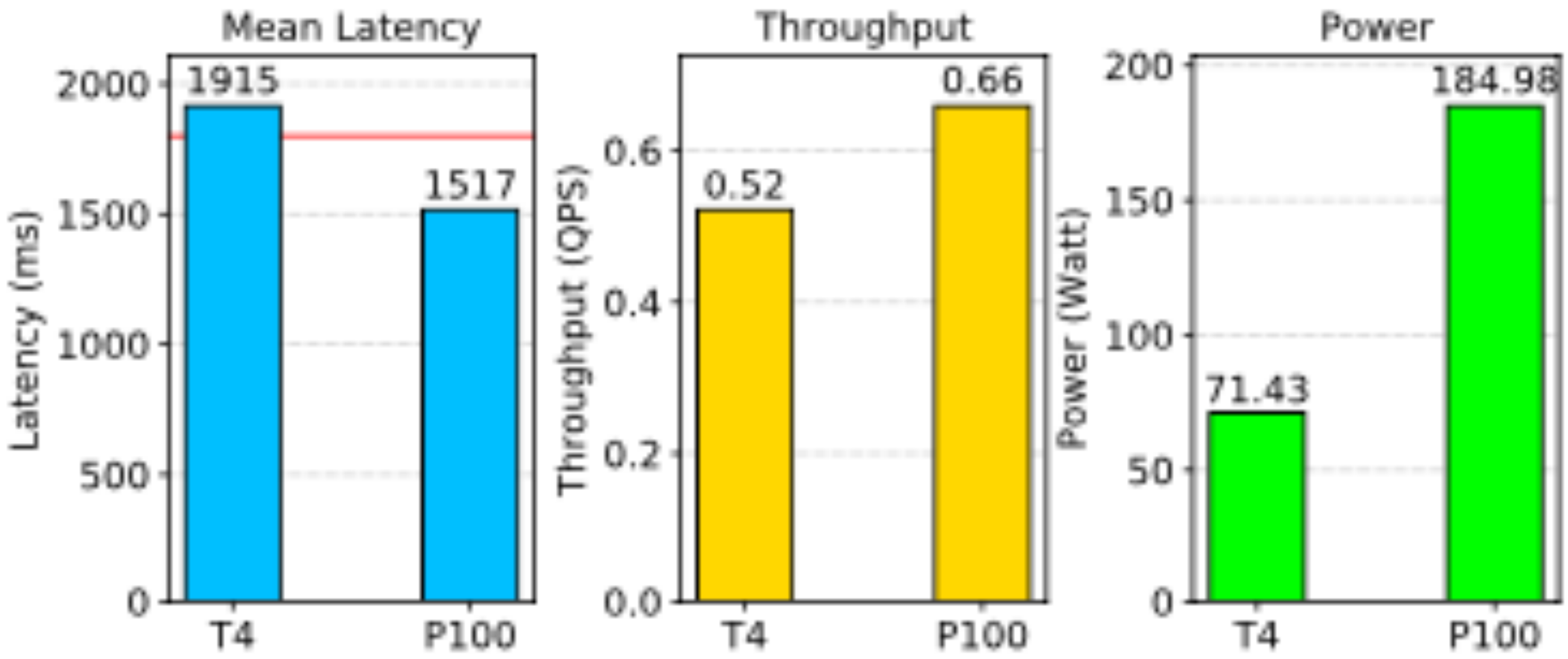
Command Post Systems

Edge Systems

Dynamic mission and hardware constraints need automated hardware selection

Observation: Different hardware platforms provide capabilities at different costs

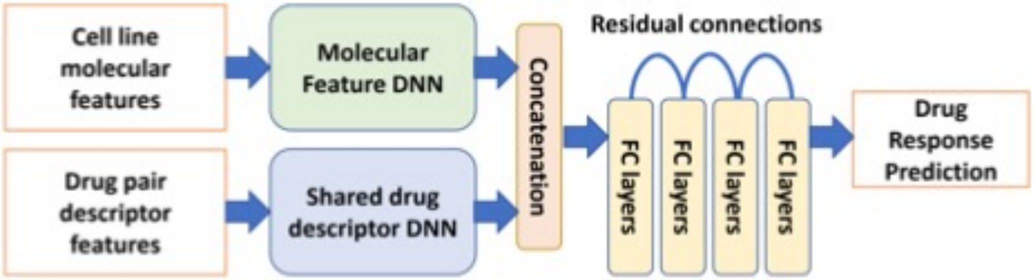
Application: Weather Forecasting



Idea: Mix and Match Hardware that satisfies high-end goals while minimizing other functions

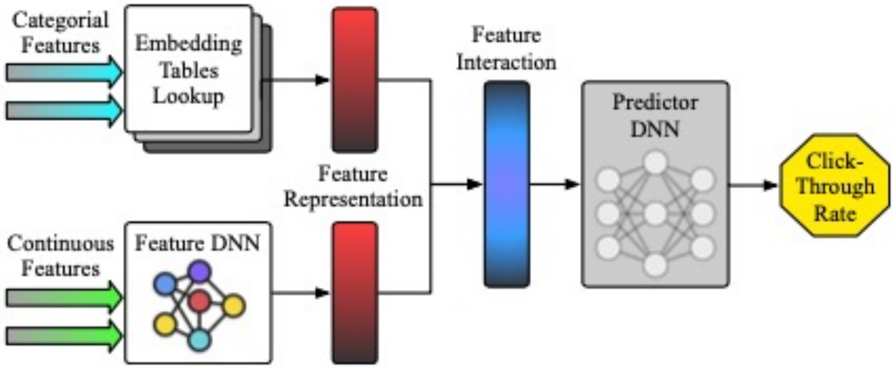
Example (Streaming) Inference Serving Tasks

Cancer Tumor Prediction (CANDLE)



- Large-scale fully-connected DNN model in Cancer Distributed Learning Environment (CANDLE) project
- Predicts tumor cell line response to drug pairs

Deep Learning Recommender Models (MT-WND, DIEN)



- Multi-Task Wide and Deep – model used for YouTube video recommendations
- Deep Interest Evolution Network – model used in e-commerce recommendations (Alibaba)

Inference Serving System Requirements

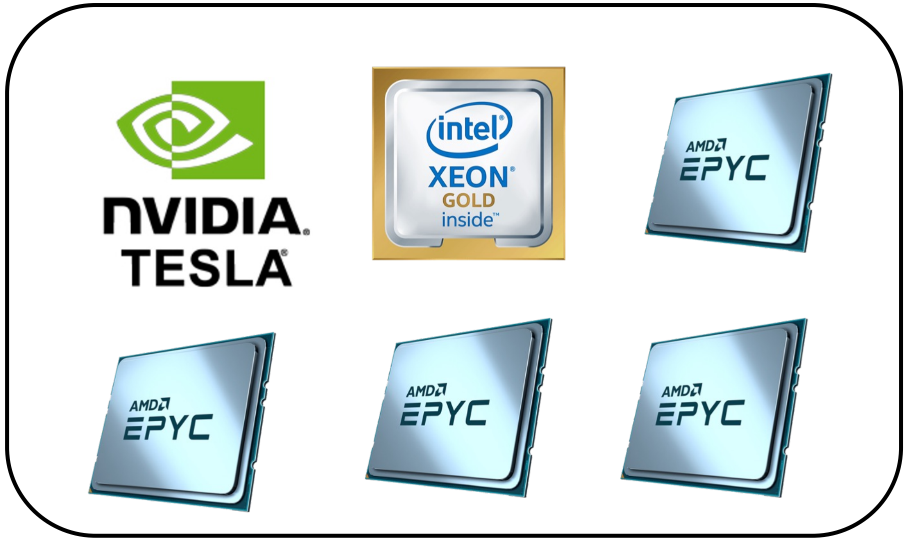
Meet Quality-of-Service (QoS)

Performance to meet the p99 tail latency

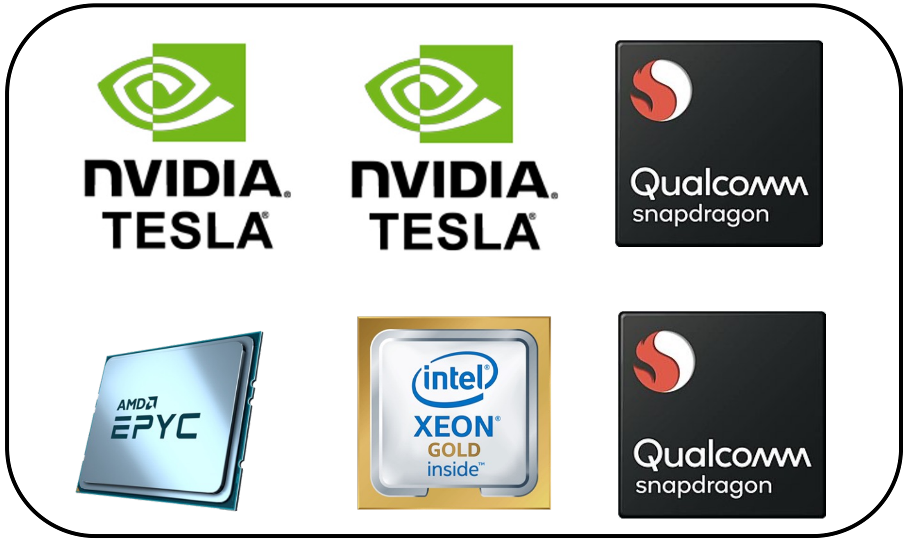
Find cost-effective solution

Minimize TCO, hardware renting fee

Problem Statement



Vs.

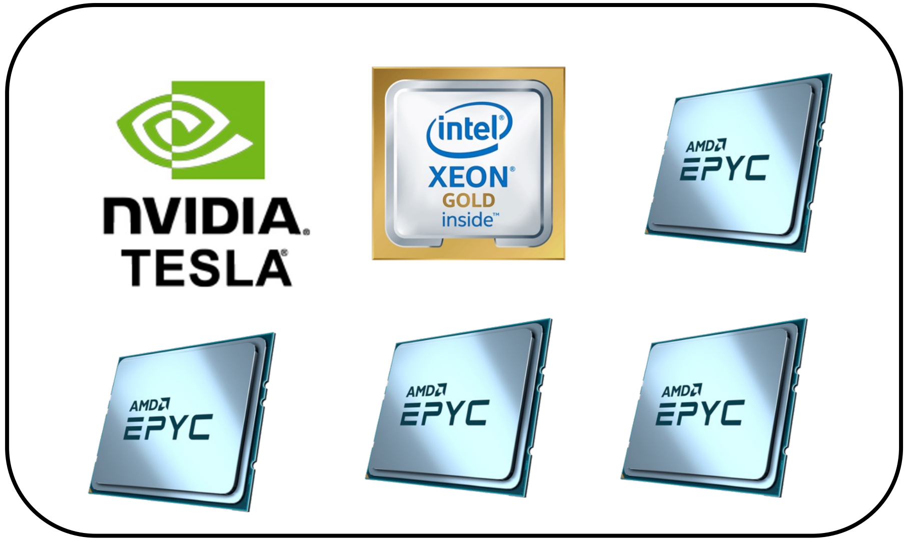


Find the least expensive* optimal diverse configuration pool while meeting the inference query QoS target

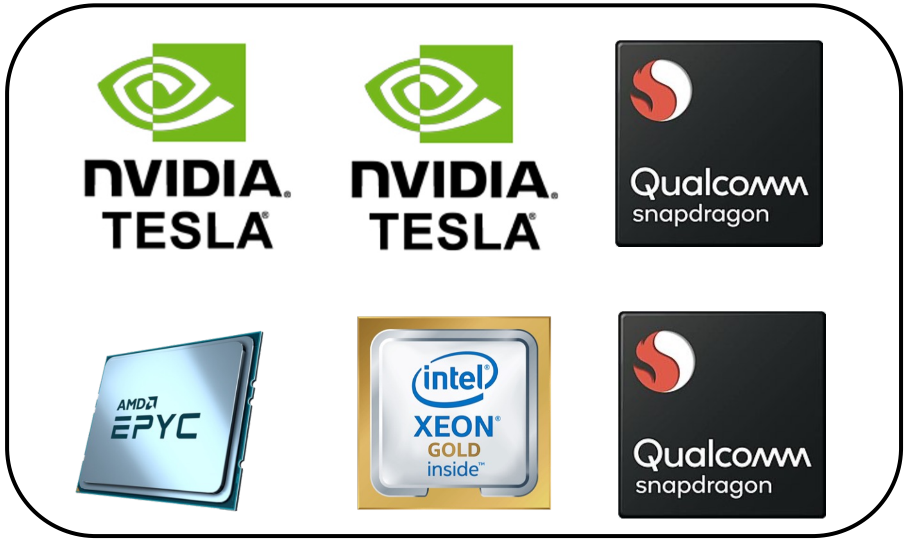
***Cost could be $$$, Energy, …**

RIBBON: cost-effective and qos-aware deep learning model inference using a diverse pool of cloud computing instances – Li, et al., *SC 2021*

Problem Statement



Vs.



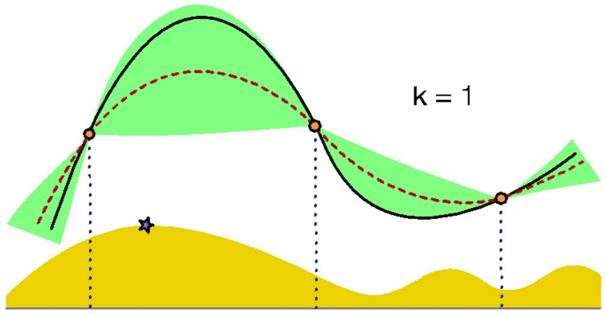
Given a certain heterogeneous instance types (e.g., X, Y, Z), how to determine the optimal number of each instance type in the heterogeneous pool (i.e., c1*X + c2*Y +c3*Z)?

RIBBON: cost-effective and qos-aware deep learning model inference using a diverse pool of cloud computing instances – Li, et al., *SC 2021*

RIBBON Builds Inference Serving System Using Diverse Computing Instances

QoS targets

**Objective:
most cost-effective serving
systems while meeting QoS targets**



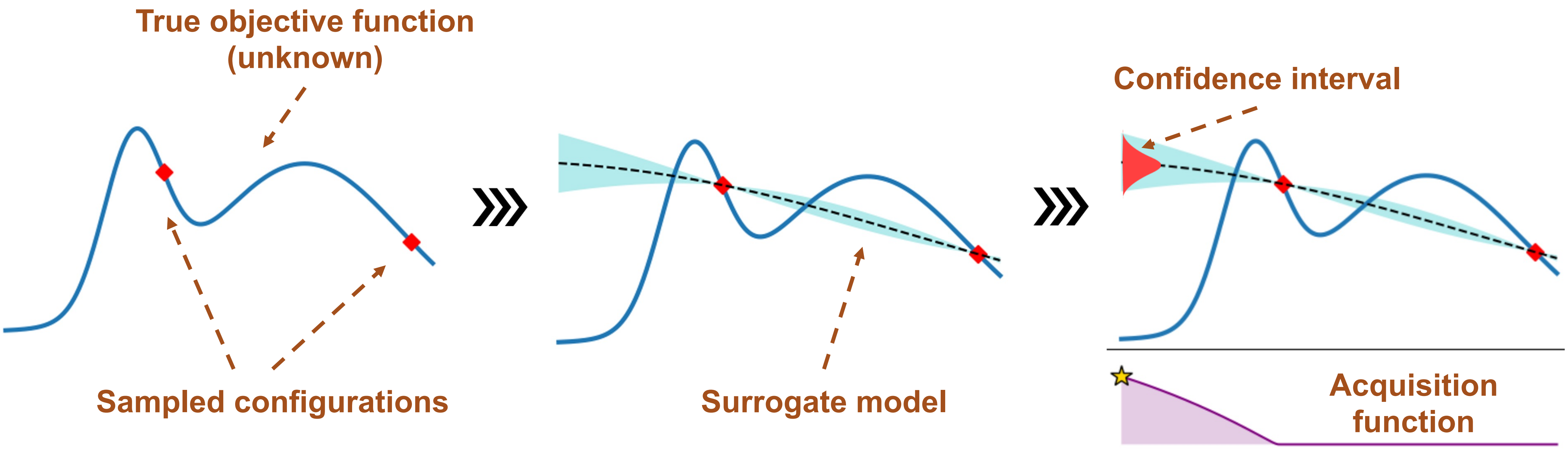
**RIBBON's Bayesian
Optimization Engine**

Minimal cost

RIBBON: cost-effective and qos-aware deep learning model inference using a diverse pool of cloud computing instances – Li, et al., *SC 2021*

RIBBON Bayesian Optimization Engine

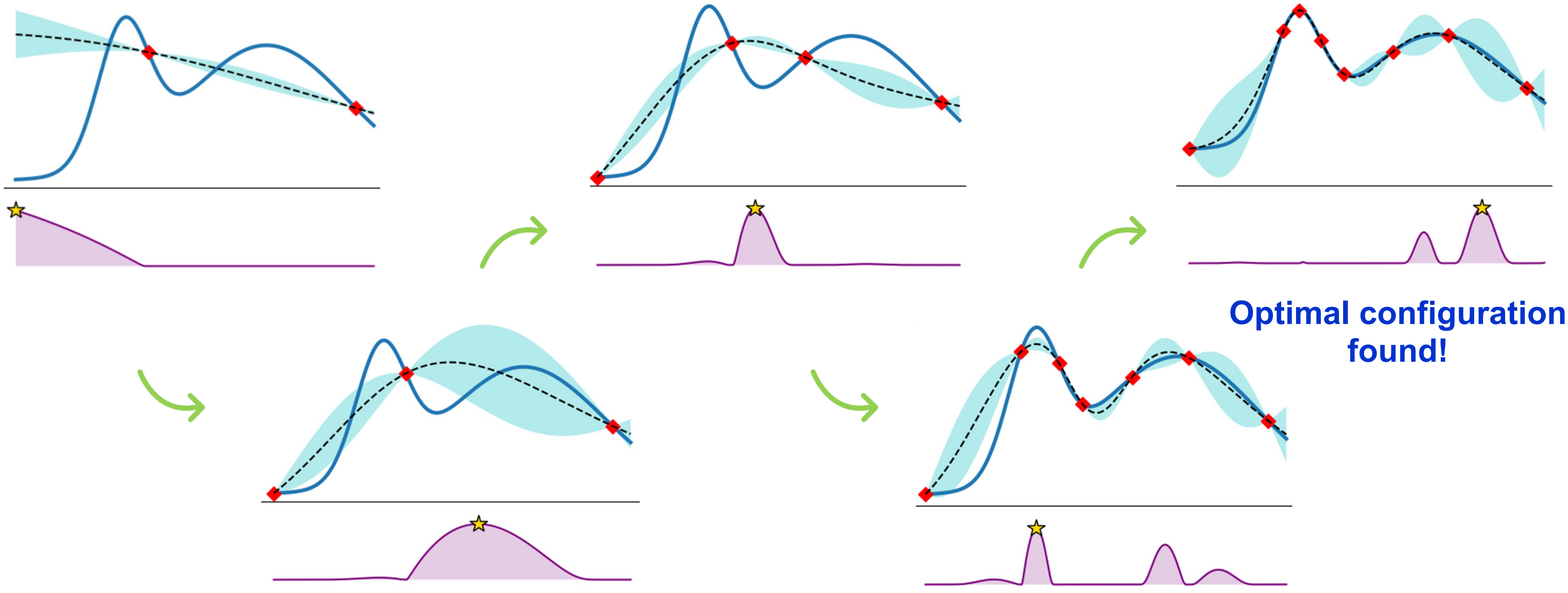
Bayesian Optimization: performs strategic global sampling to optimize unknown objective with limited total samples.



RIBBON: cost-effective and qos-aware deep learning model inference using a diverse pool of cloud computing instances – Li, et al., *SC 2021*

RIBBON Bayesian Optimization Engine

With more sampled configurations, surrogate model becomes closer to true objective function



Optimal configuration found!

RIBBON: cost-effective and qos-aware deep learning model inference using a diverse pool of cloud computing instances – Li, et al., *SC 2021*

Design Considerations

Surrogate model	Covariance function	Acquisition function
Gaussian process	**Matern 5/2 kernel**	**Expected improvement**
Tree Parzen estimator	Dot product	Upper confidence bound
Polynomial estimator	Rational Quadratic	Probability of improvement

Guides optimizer towards QoS satisfaction smoothly

$$f(x) = \begin{cases} \frac{1}{2} \cdot \frac{R_{sat}(x)}{T_{qos}} & \text{if violates QoS,} \\ \frac{1}{2} + \frac{1}{2} \cdot \left(1 - \frac{\sum_{i=1}^{n} p_i \cdot x_i}{\sum_{i=1}^{n} p_i \cdot m_i}\right) & \text{otherwise.} \end{cases}$$

Objective function

- **Minimize cost**
- **While meeting QoS targets**

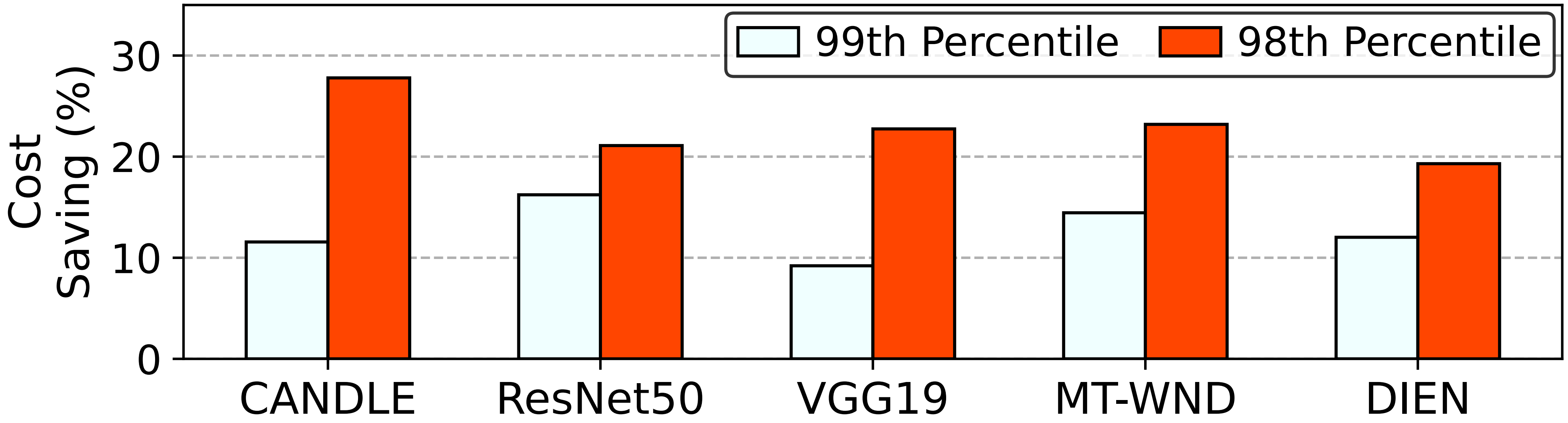
Ensures QoS satisfaction

Normalized total serving cost

RIBBON: cost-effective and qos-aware deep learning model inference using a diverse pool of cloud computing instances – Li, et al., *SC 2021*

Significant cost savings across inference tasks while meeting various QoS targets

Cost savings of RIBBON suggested hardware pool vs. best homogenous configuration



CANDLE	CANcer Distributed Learning Environment drug response model
ResNet50	CNN model with residual operations, applied in image classification
VGG19	Popular computer vision model
MT-WND	Multi-Task Wide-and-Deep, deep learning model for YouTube video recommendation
DIEN	Deep Interest Evolution Network, used for e-commerce recommendation

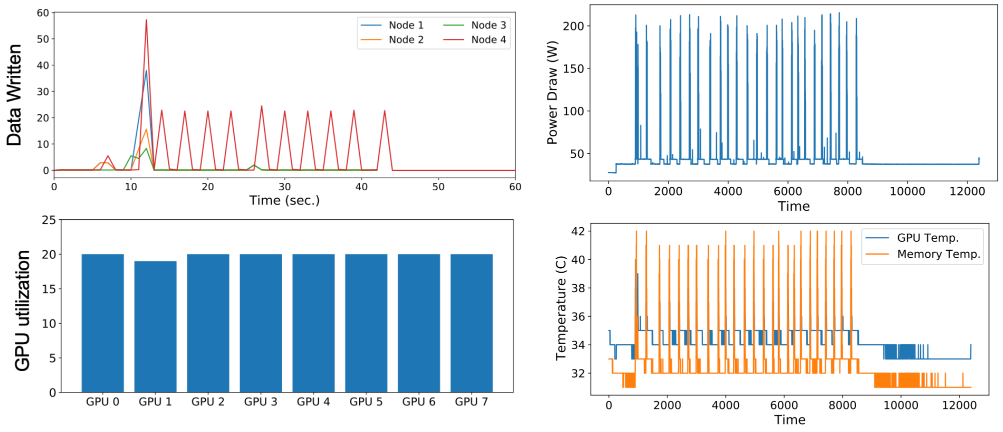
RIBBON: cost-effective and qos-aware deep learning model inference using a diverse pool of cloud computing instances – Li, et al., *SC 2021*

Outline

- Motivation
- Reducing development computing demands
- Finding the right deployment environment
- Datacenter Challenge
- Summary and Air Force Perspective

Datacenter Challenge

- **Challenge to enable datacenters that can:**
  - **Predict and identify system failures**
  - **Optimize system scheduling for improved resource consumption**
  - **Suggest optimization pathways for users**
- **Open-source data to improve operational capabilities on a variety of AI workloads**
- **Contents:**
  - **Scheduler Logs**
  - **CPU/GPU timeseries**
  - **BMS/Environmental Data**
  - **Labelled workloads**



The Fast AI Datacenter Challenge aims to foster innovation in AI approaches to the analysis of large scale datacenter monitoring logs

https://dcc.mit.edu/
https://news.mit.edu/2022/taking-magnifying-glass-data-center-operations-0824

Current Status

- **Over 2+ TB of time series data collected, parsed, anonymized and ready for distribution**
  - **Resource utilization from ~500K jobs**
  - **Includes ~100K GPU workloads**
  - **Labelled dataset of 3,425 known deep learning workloads from Vision, NLP and GNN**
    - **Mixture of Tensorflow and pytorch implementations**
- **Data dissemination**
  - **Available on Amazon AWS Open Data Registry :**

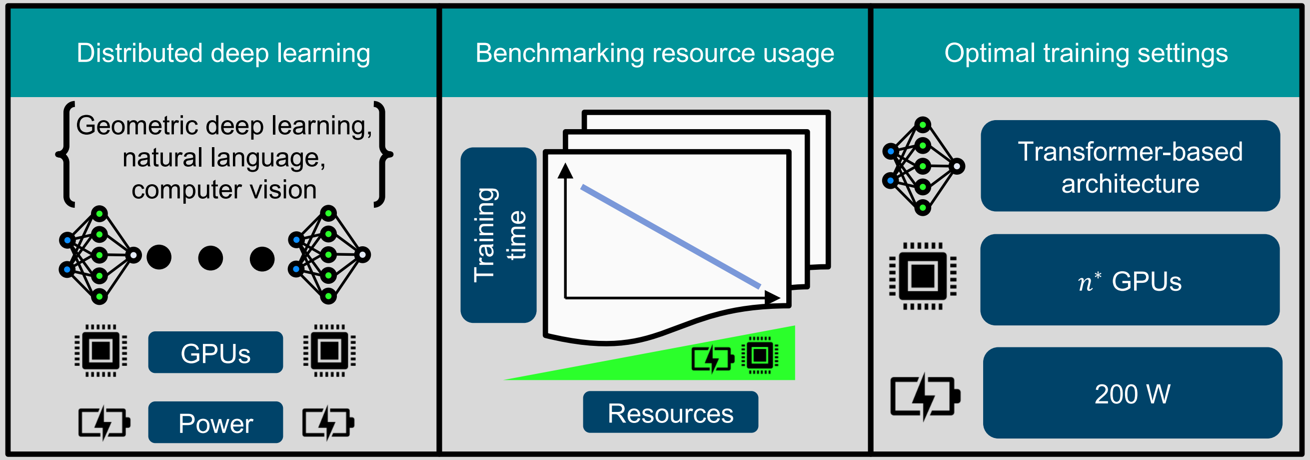
    `s3://mit-supercloud-dataset/datacenter-challenge`
  - **Scripts and data loaders**

    `https://github.com/MIT-AI-Accelerator`
- **Relevant Publications:**
  - **The MIT Supercloud Dataset, IEEE *HPEC'21***
  - **AI-Enabling Workloads on Large-Scale GPU-Accelerated System: Characterization, Opportunities, and Implications, *HPCA'22***

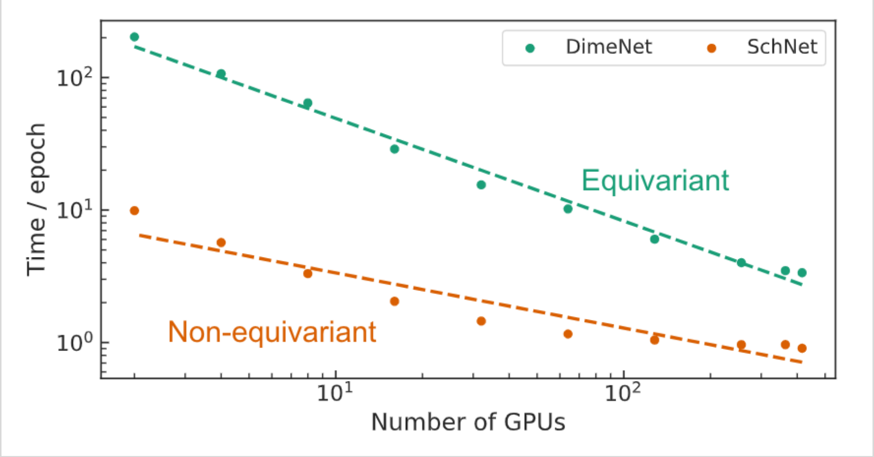
NLP – Natural Language Processing
GNN – Graph Neural Networks

Example Research:
Efficient, Scalable AI training on HPC Systems

- Performed over 3,400 deep learning workload experiments on LLSC systems
- Trained 6 state-of-the-art neural networks across vision, natural language processing, chemistry, and materials science domains on up to 424 GPUs



Benchmarking experiments on more than 400 GPUs with controlled hardware settings reveal optimal settings for large-scale deep learning workflows.



Training time versus number of GPUs is well-described by empirical power laws.

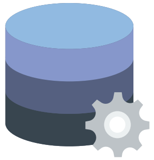
Findings will guide high-performance computing providers in optimizing resource usage

Outline

- Motivation
- Reducing development computing demands
- Finding the right deployment environment
- Datacenter Challenge
- Summary and Air Force Perspective

DAF-MIT AI Accelerator (AIA)
Bringing World-Class Research to USAF Missions

Air Force Missions


Cyber


ISR and Autonomy


Air Defense

USAF-MIT AI Accelerator

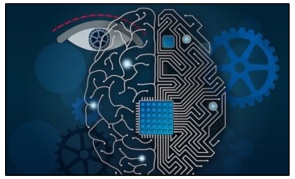
AI Researchers

MIT Campus

MIT Lincoln Laboratory

US Air Force

Mission-ready Projects


Foundational AI Research


AI Challenge Problems


AI Education and Training

Air Force Operations

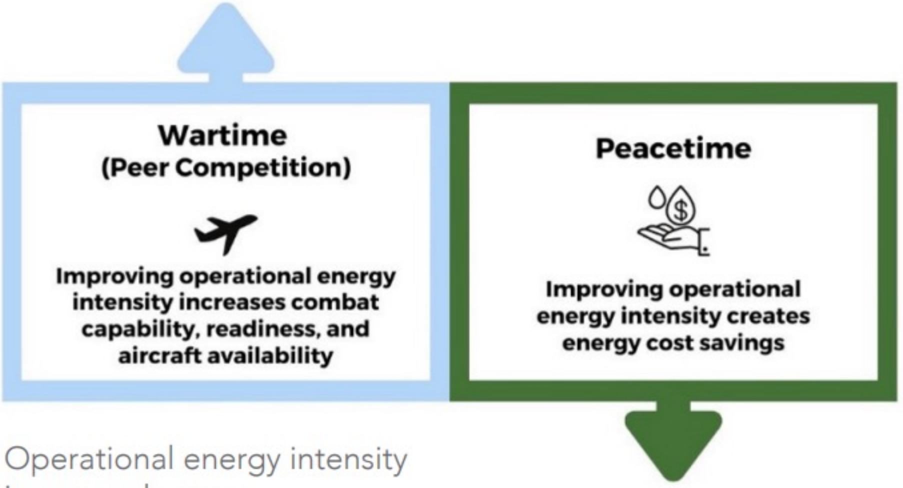

Personnel Recovery


Data Center Operations


Pilot Safety

Impact on Department of the Air Force

- The DAF's mission is to deter conflict and if necessary, defeat adversaries across the air and space domains
  - Constraints: time, cost, law/policy, operational environment, weather/climate, energy
  - Enablers: AI, capable allies & partners, access to cutting edge research
- The research conducted by MIT:
  - Advances the goals of the DAF's Climate Action Plan
    - Optimize energy use & make climate-informed decisions
  - Optimizes performance on DAF's existing hardware, saving costly tech refresh cycles
  - Can increase throughput or extend battery life on edge devices
  - Is supported by embedded Airmen
    - Collaborative R&D with continuous end-user feedback



Operational energy intensity in war and peace.

Summary

- Need for tools that bridge gap between development and deployment environments
- Challenges:
  - Increasing computing requirements
  - Energy / cooling limits
  - Hardware diversity
  - Evolving missions/workloads
- Opportunity to leverage AI to mitigate challenges
- LLSC looking for talented postdocs/staff. If interested, email me!

vijayg@ll.mit.edu