RIBBON: Cost-Effective and QoS-Aware Deep Learning Model Inference using a Diverse Pool of Cloud Computing Instances

Baolin Li, Rohan Basu Roy, Tirthak Patel,

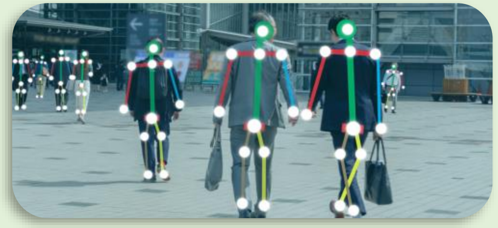
Vijay Gadepally, Karen Gettings, Devesh Tiwari

Northeastern University

MIT Lincoln Laboratory

Deep Learning Models Are Ubiquitous

Autonomous Driving



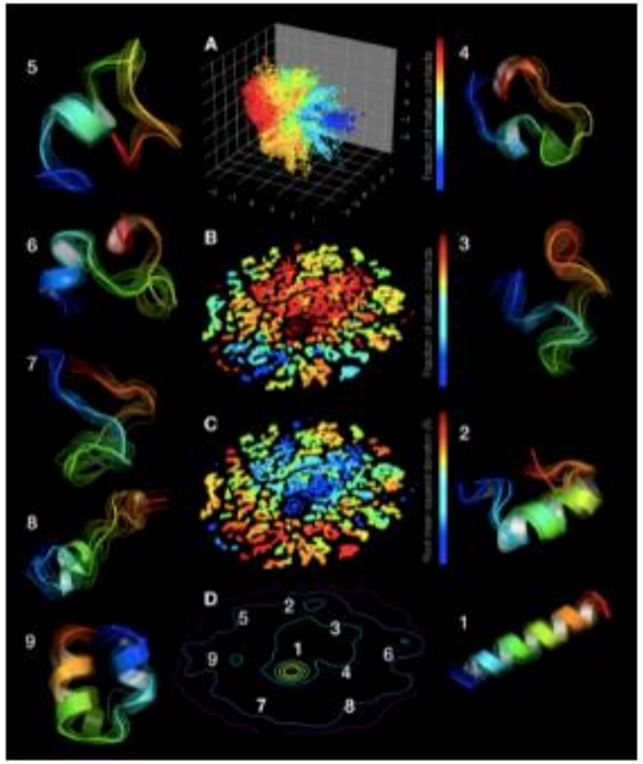
Recommendation Systems

amazon

NETFLIX

Drug Discovery

Deep Learning Models Are Ubiquitous



Using CANDLE deep learning to extract protein folding intermediate states. | National Cancer Institute

Model Inference Serving System Requirements

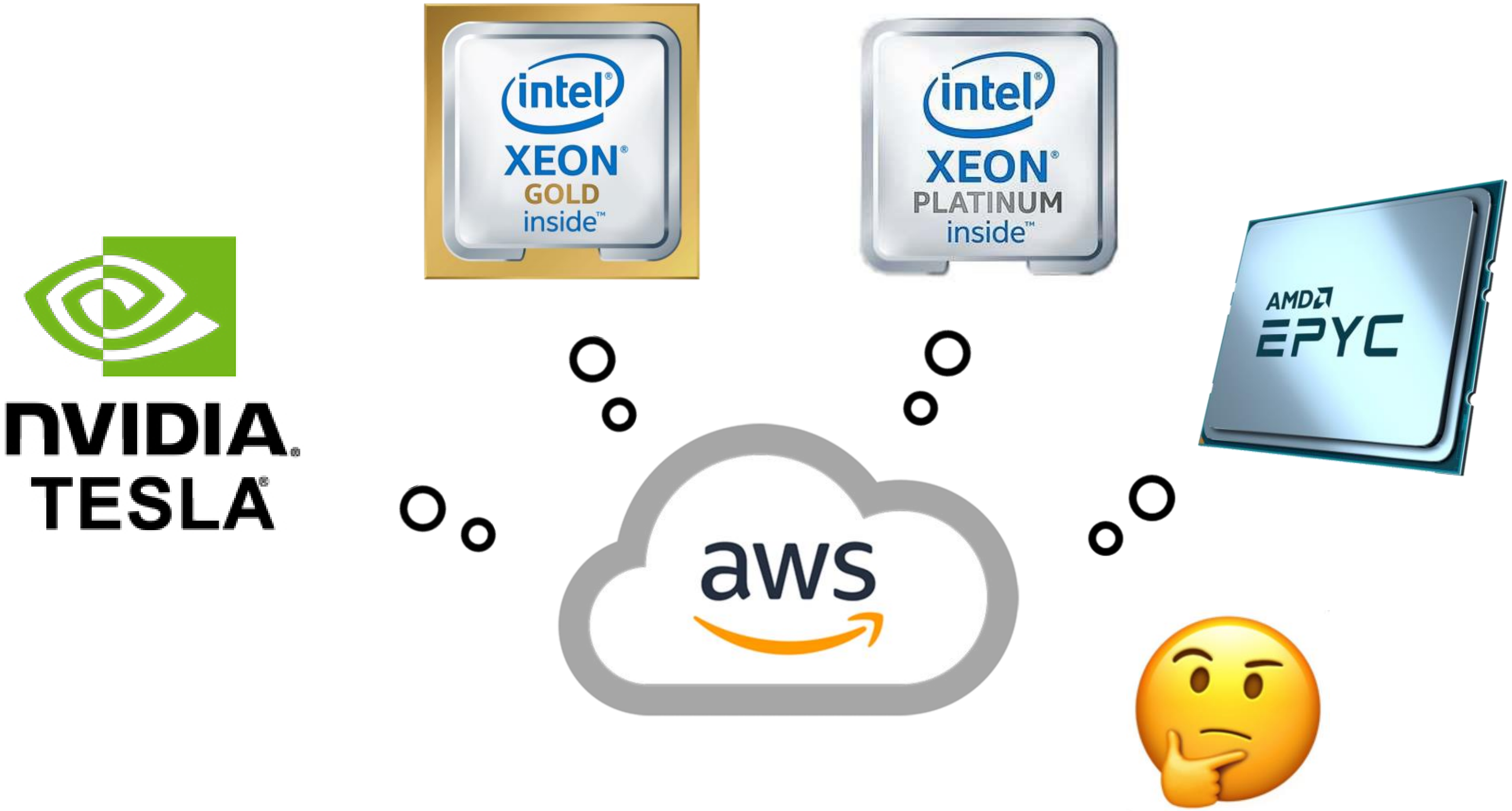
Meet Quality-of-Service (QoS)

Performance to meet the p99 tail latency

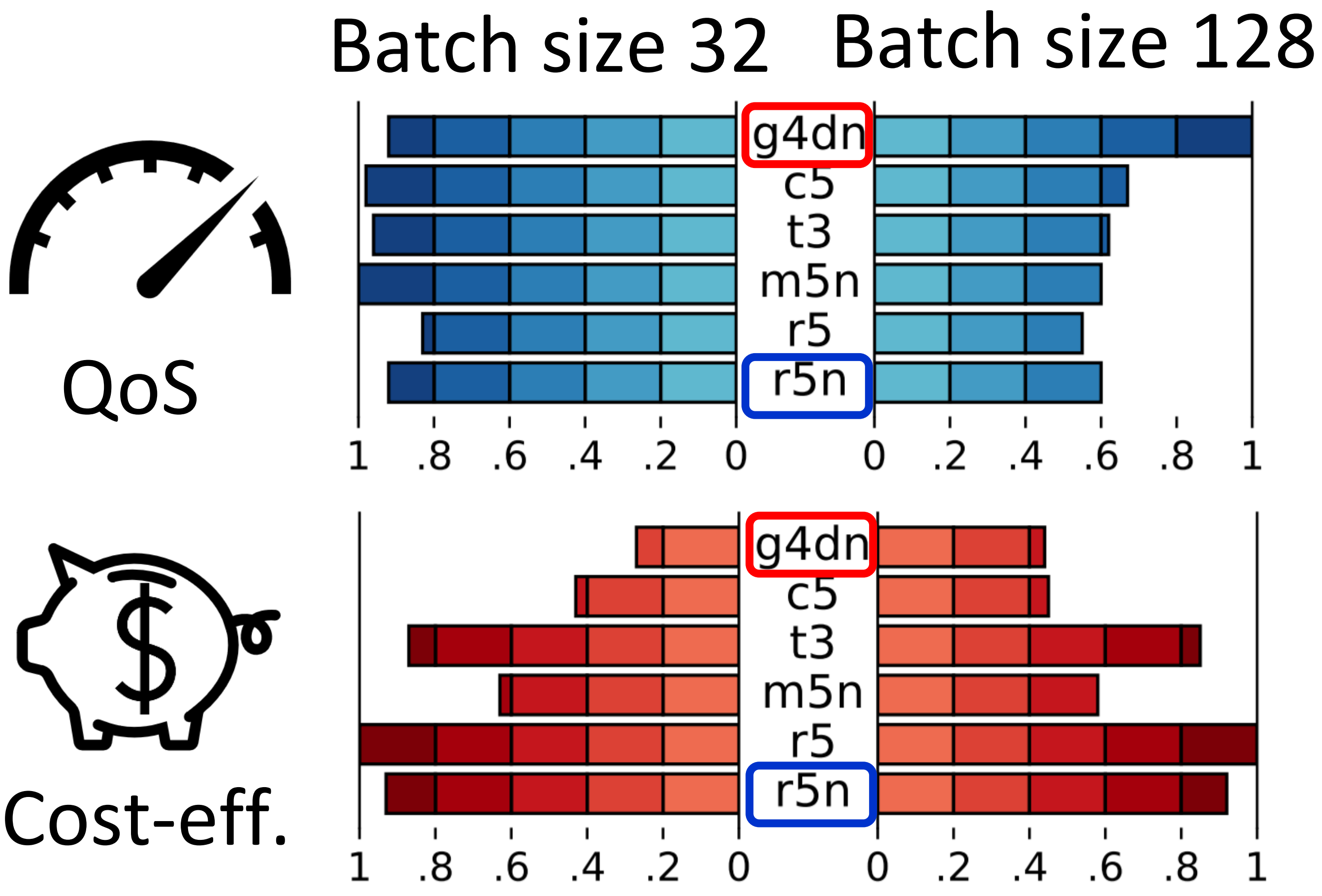
Find cost-effective solution

Minimize TCO, hardware renting fee

Cloud computing resources: too many choices with different trade-offs!

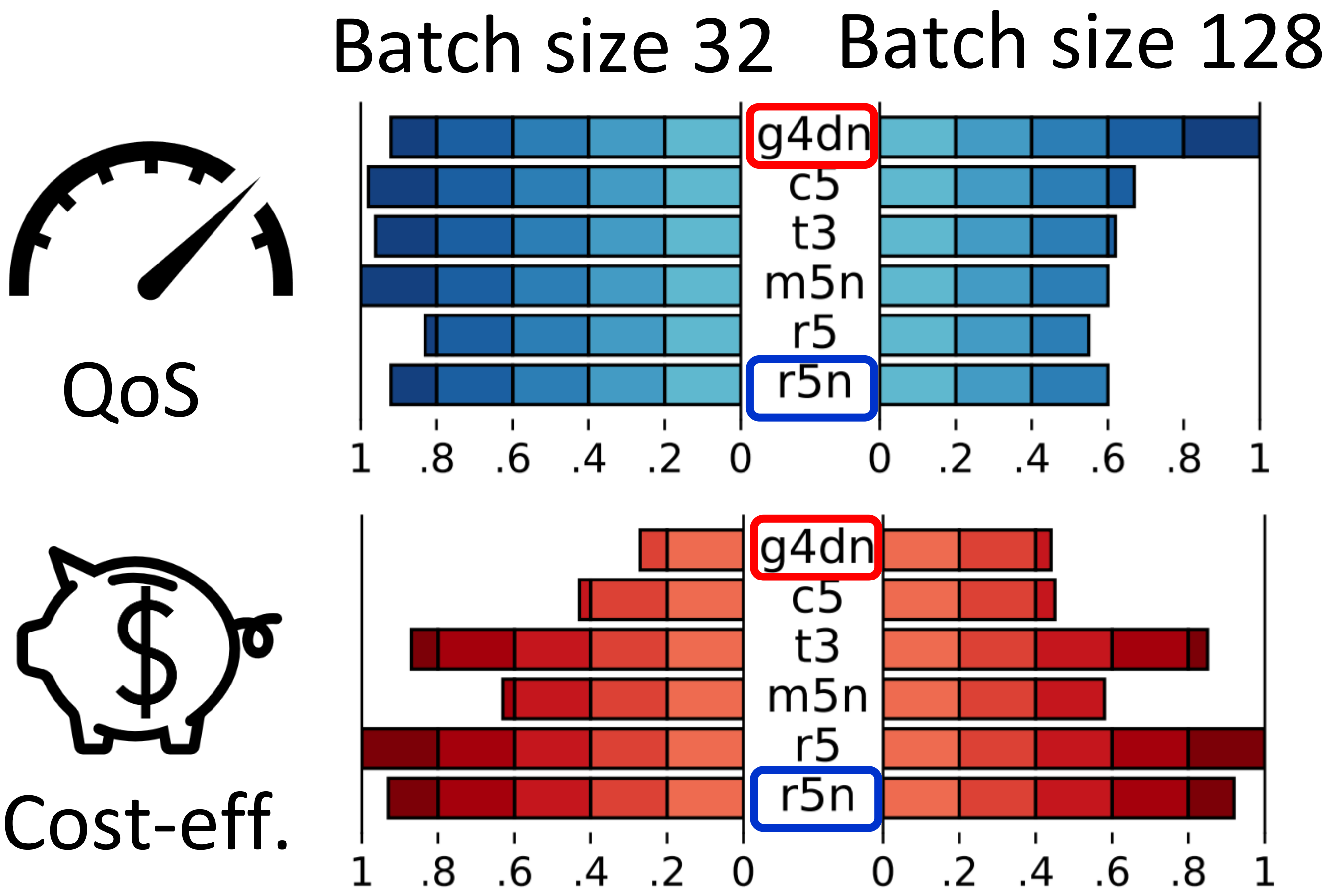


QoS and cost-effectiveness are at odds!



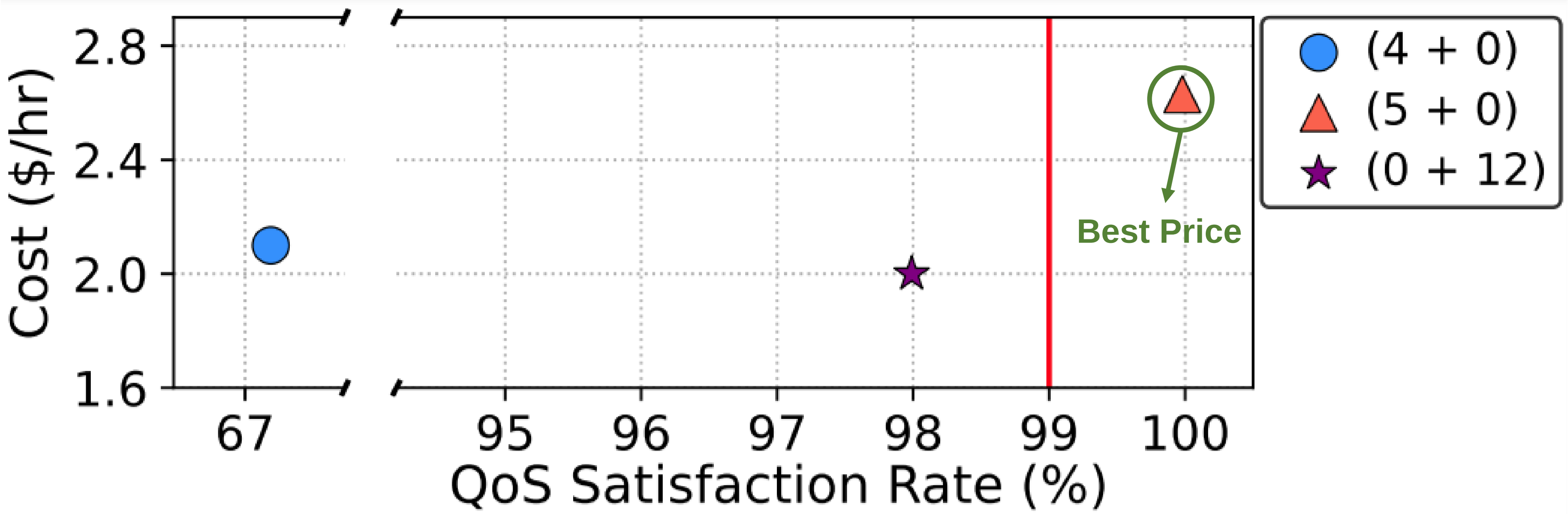
Incoming query stream may have queries of different batch sizes

QoS and cost-effectiveness are at odds!



Traditional inference serving system puts together a bunch of homogenous instance types that *must* satisfy the QoS

Best homogenous pool is the one with the least cost that satisfies the QoS constraints



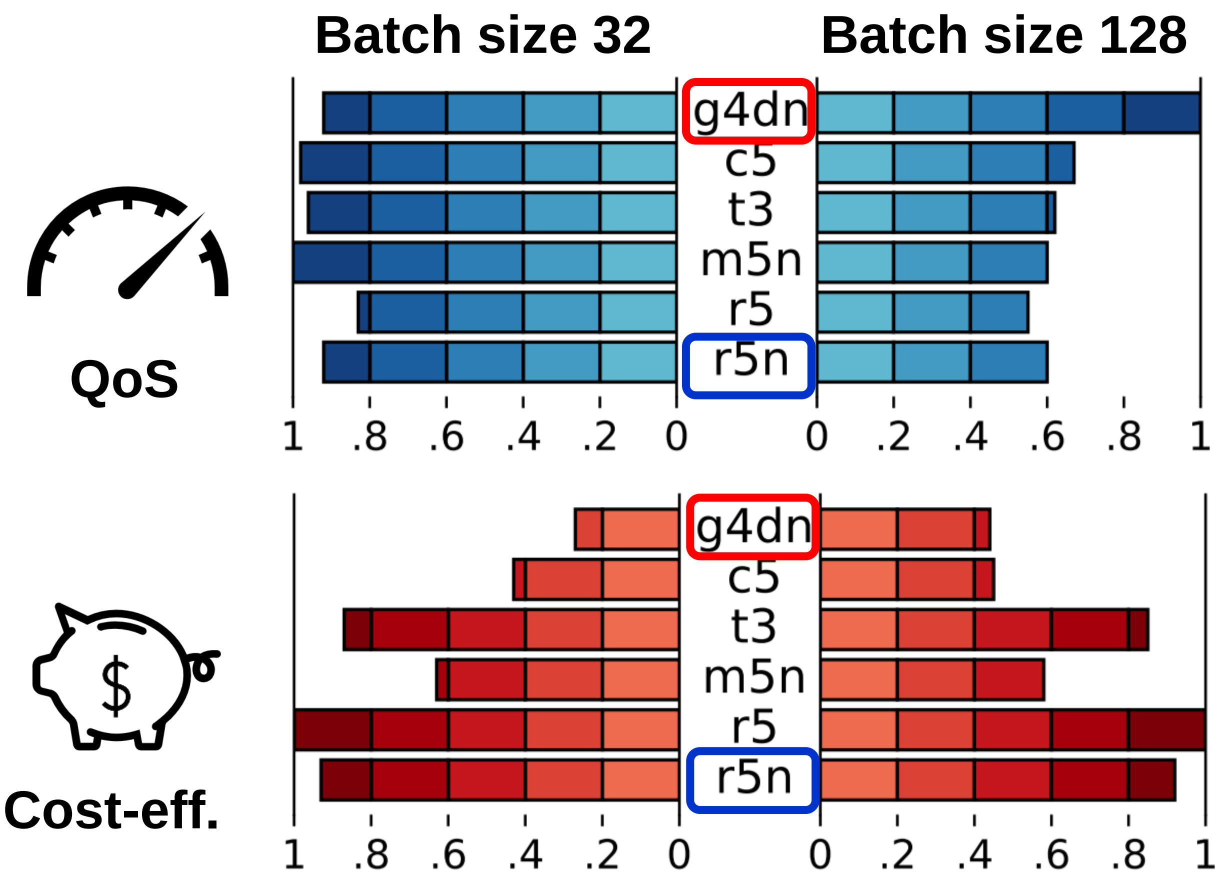
Instance pool is represented as (X + Y)

X is high performance cost-ineffective instance type

Y is low performance but cost-effective instance type

RIBBON: Opportunity

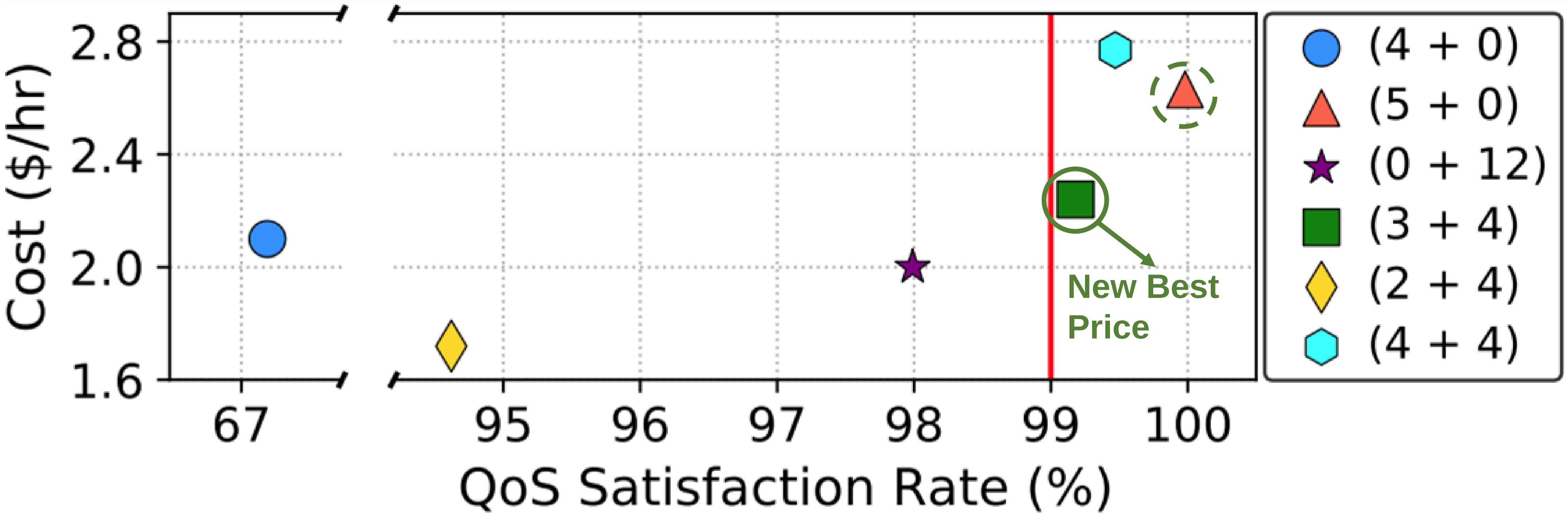
Batch size 128:
r5n ≪ g4dn
But at size 32:
r5n ≈ g4dn



Mix-up high performance cost-ineffective instances with low performance cost-effective instances

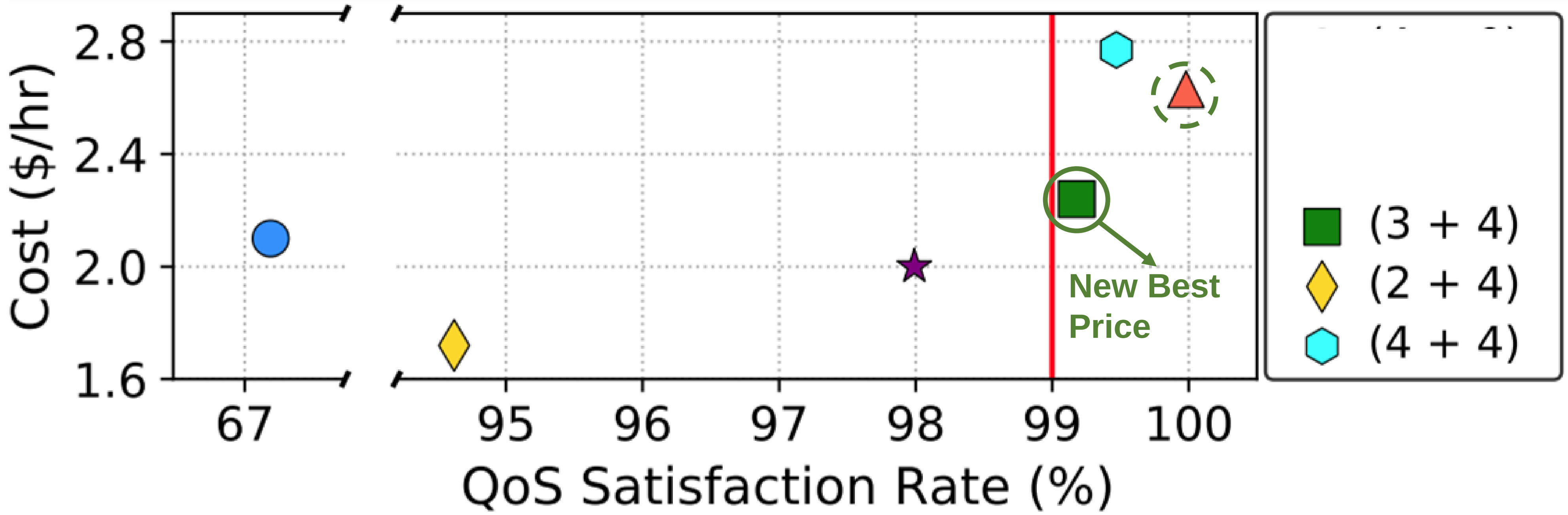
Key Insight

"Some" heterogeneous/diverse pools can be more cost-effective than the best homogenous pool

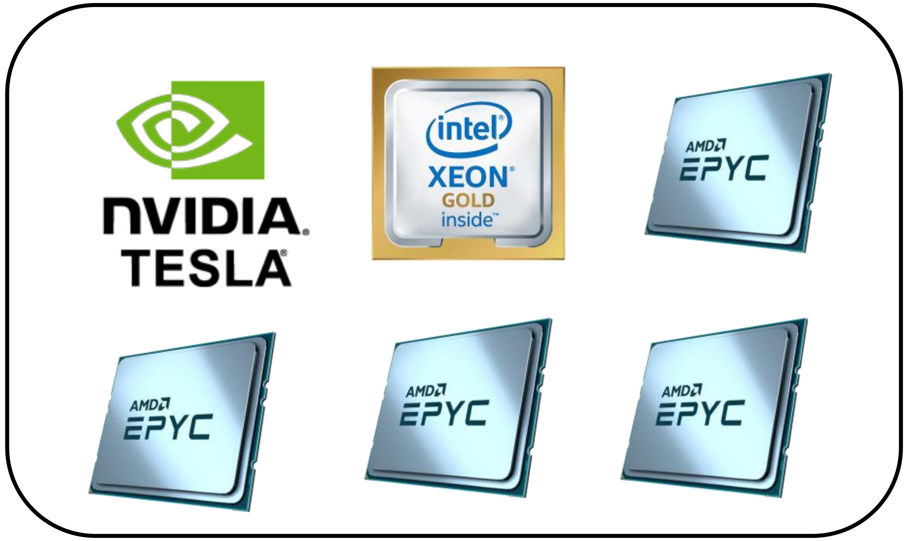


Key Insight

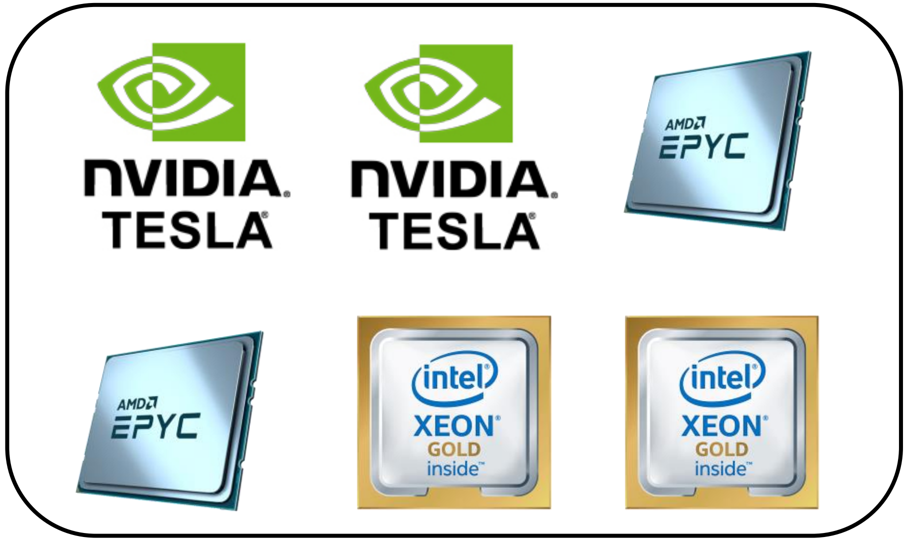
But finding such promising heterogeneous configurations is challenging, let alone optimal



Problem Statement

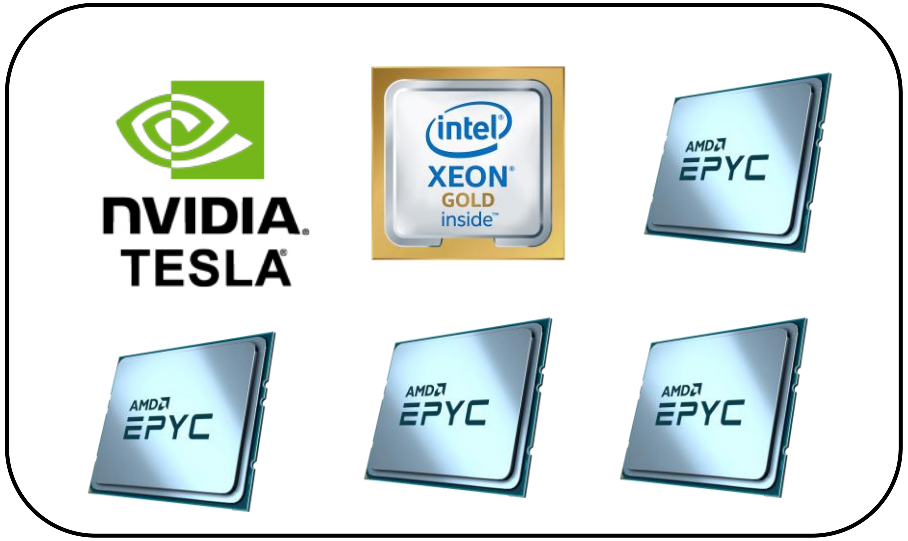


Vs.

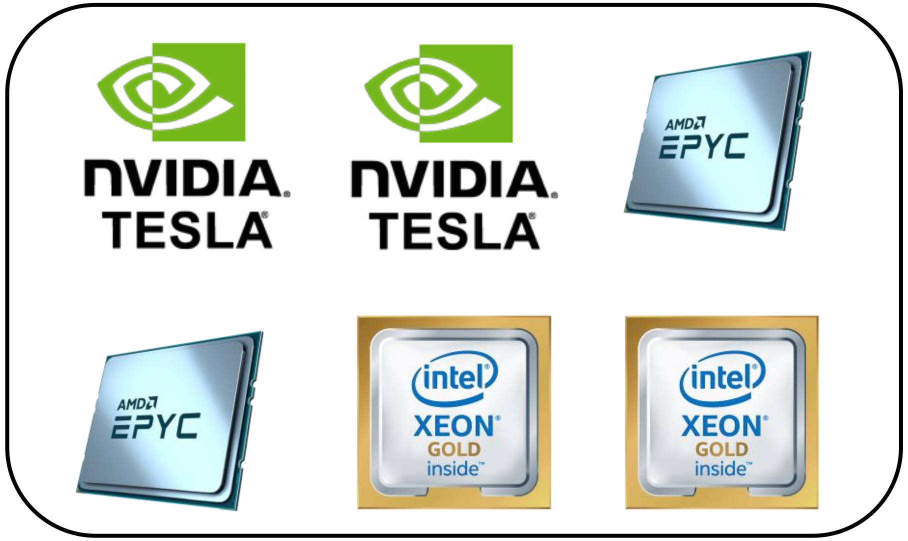


Find the optimal diverse configuration pool which is least expensive while meeting the inference query QoS target

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 diverse pool (i.e., c1*X + c2*Y +c3*Z)?

Large configuration space of heterogeneous configurations

Complex interaction between configured diverse pool and QoS

Why Building an Inference Serving System with Optimal Heterogeneous Pool So Challenging?

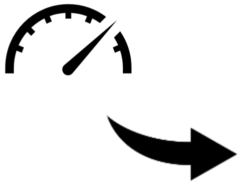
Evaluating each heterogeneous configuration is expensive

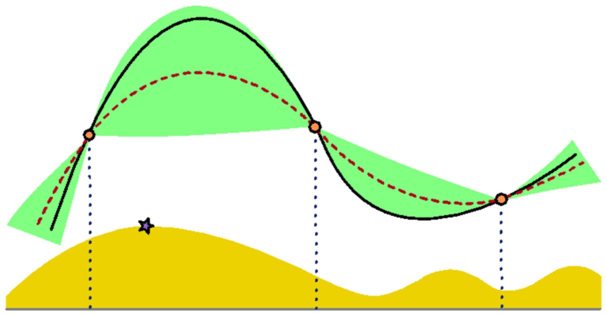
QoS and cost rankings of configurations are different

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

Request Inferencing Based On Bayesian OptimizatioN

Bayesian Optimization: strategy for global optimization of black-box, expensive-to-evaluate functions

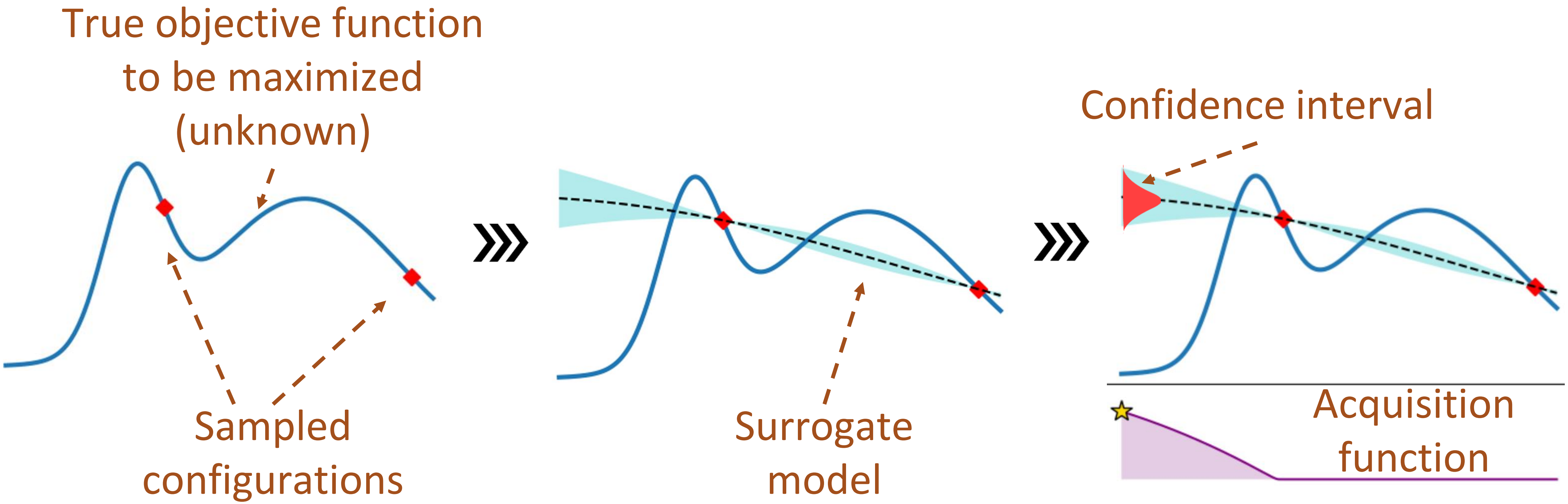
BO maintains a balance between exploration (unsampled configurations) and exploitation (sampled good configurations).

Key components of the RIBBON BO engine

- Surrogate model
- Acquisition function
- Sampled configuration

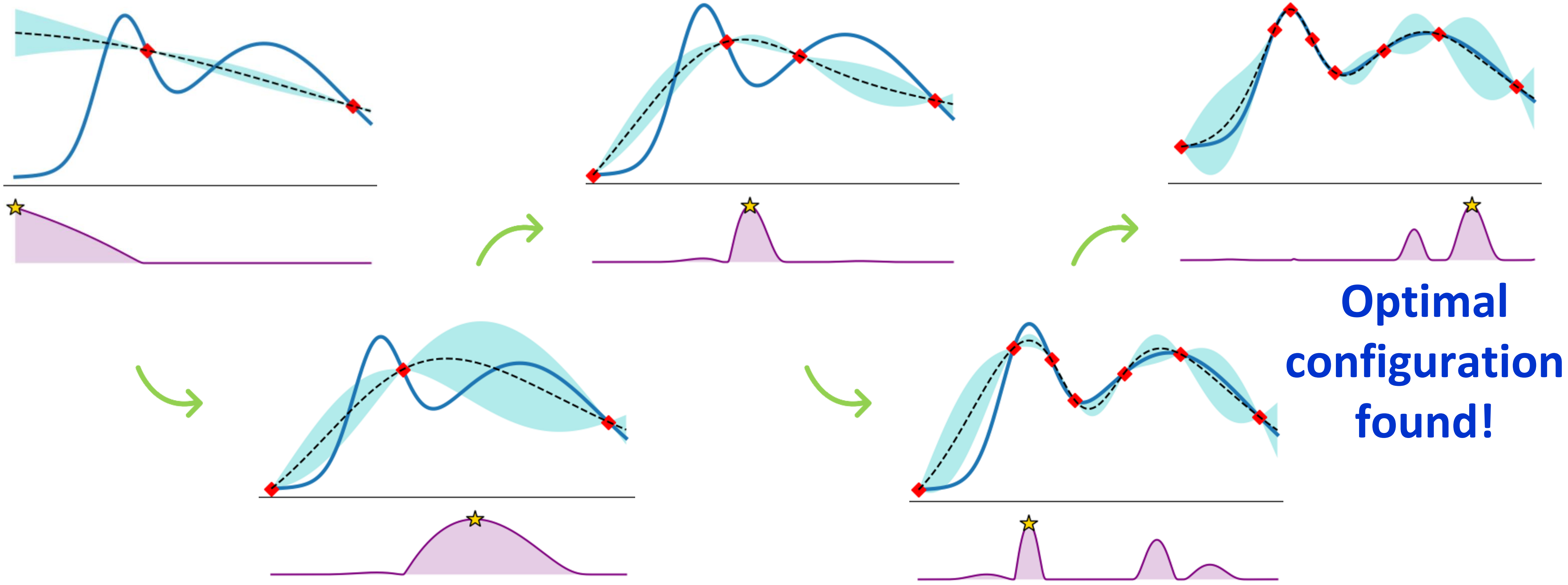
RIBBON Bayesian Optimization Engine

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



RIBBON Bayesian Optimization Engine

As more configurations get sampled, the surrogate model becomes closer to the true objective function



RIBBON: 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

Objective function

- Minimize cost
- While meeting QoS targets

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

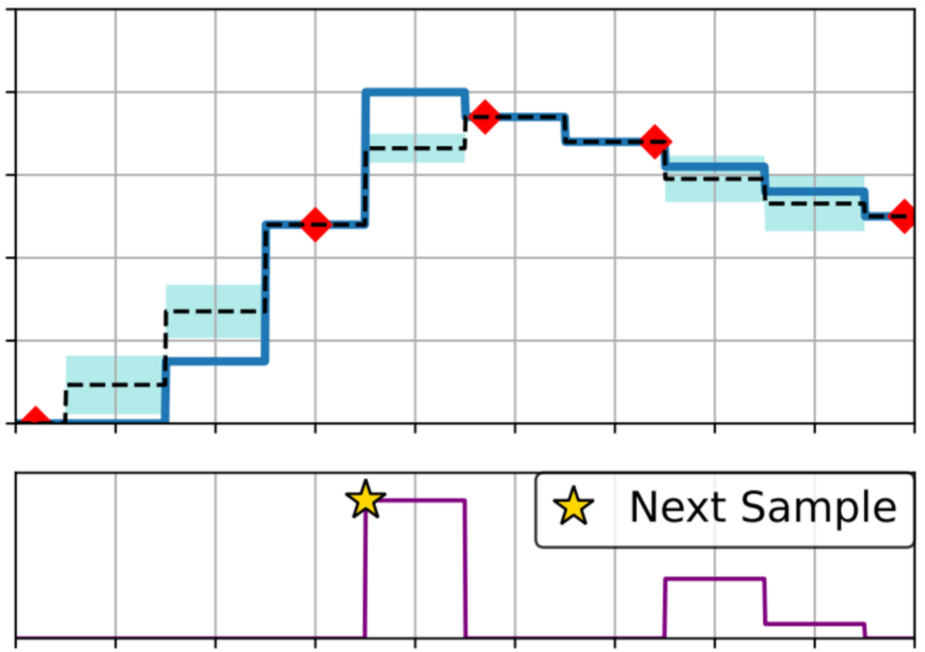
Ensures QoS satisfaction

Normalized total serving cost

RIBBON: Design Considerations

Kernel design for discrete inputs

Applies rounding to GP kernel



Active pruning to reduce search space

If (3X,4Y,5Z) violates QoS, so will (2X,3Y,4Z)

Promptly respond to load change

Maintains a record of sample history

RIBBON: Experimental Methodology

Evaluated Models

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

Inference Query Characteristics

QoS: 99-th percentile tail latency

Inference arrival rate and other characteristics modeled after Industry-grade trace

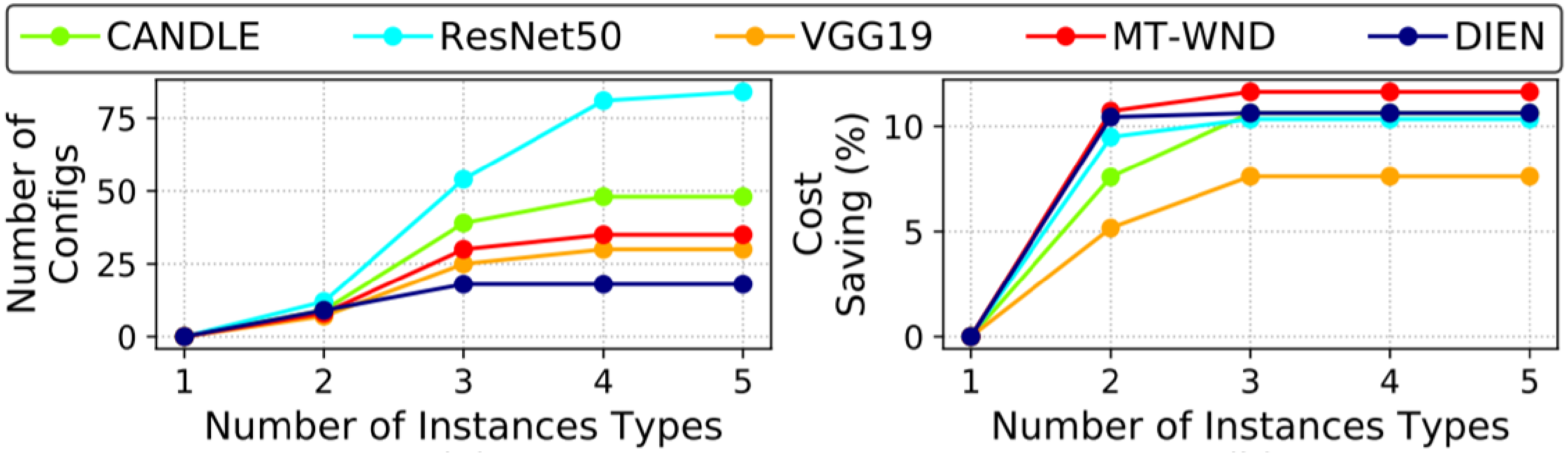
RIBBON: Experimental Methodology

Building diverse pool with different AWS instances

- DNN: c5a, m5, t3
- Recommendation: g4dn, c5, r5n

Three different instance types used

- Avoid search space explosion
- Diminishing returns of potential cost savings



RIBBON: Figure of Merit and Competing Schemes

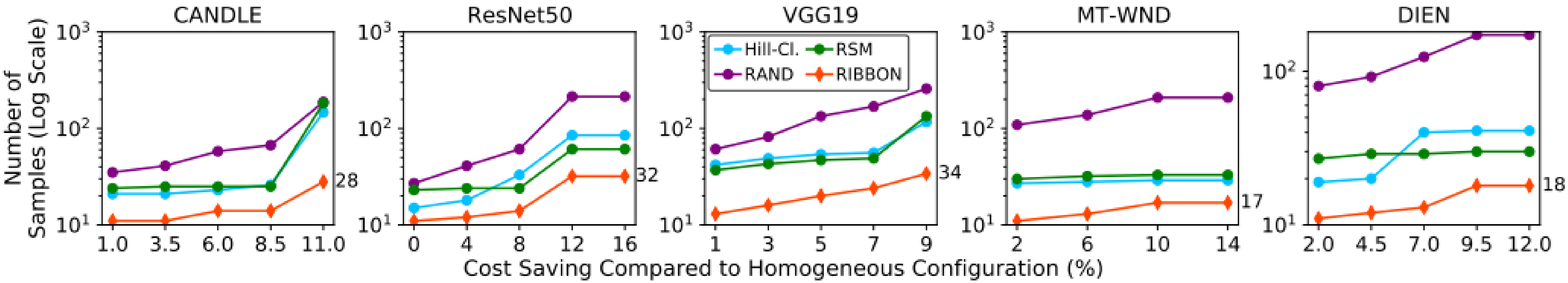
Metrics

- ✓ Cost of the inference serving system
- ✓ Time to find the (near) optimal configuration

Competing Schemes

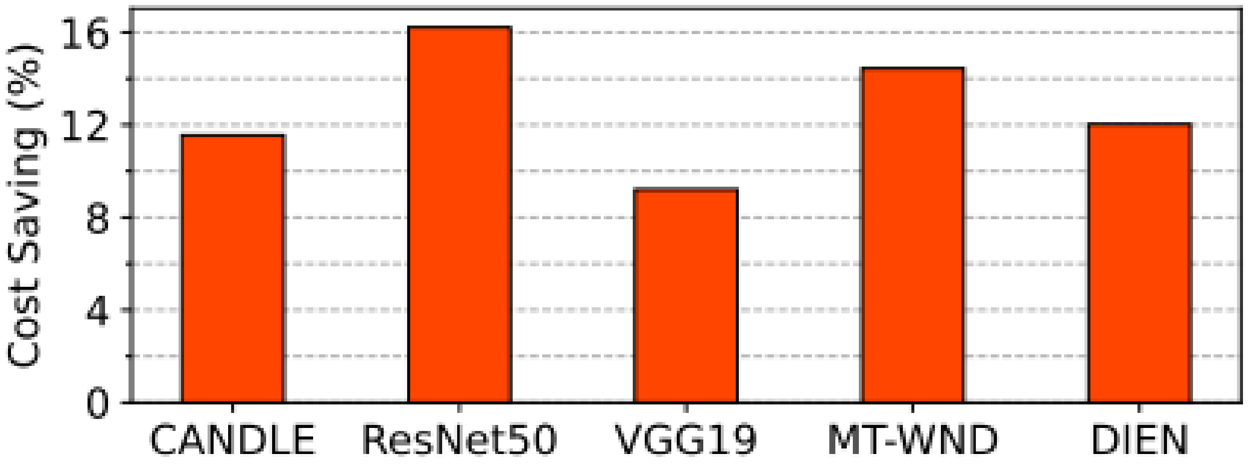
- ✓ Best homogenous pool (baseline)
- ✓ RAND: randomly explore
- ✓ Hill-Climb: go to a better neighbor
- ✓ RSM: response surface methodology

RIBBON's Bayesian Optimization based method determines the most cost-effective diverse configuration the quickest.



RIBBON’s diverse pool approach yields significant cost savings across all models over homogenous pool while meeting QoS targets.

CANDLE	CANcer Distributed Learning Environment drug response model
ResNet50	CNN model with residual operations, applied in image classification
VGG19	Another famous 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



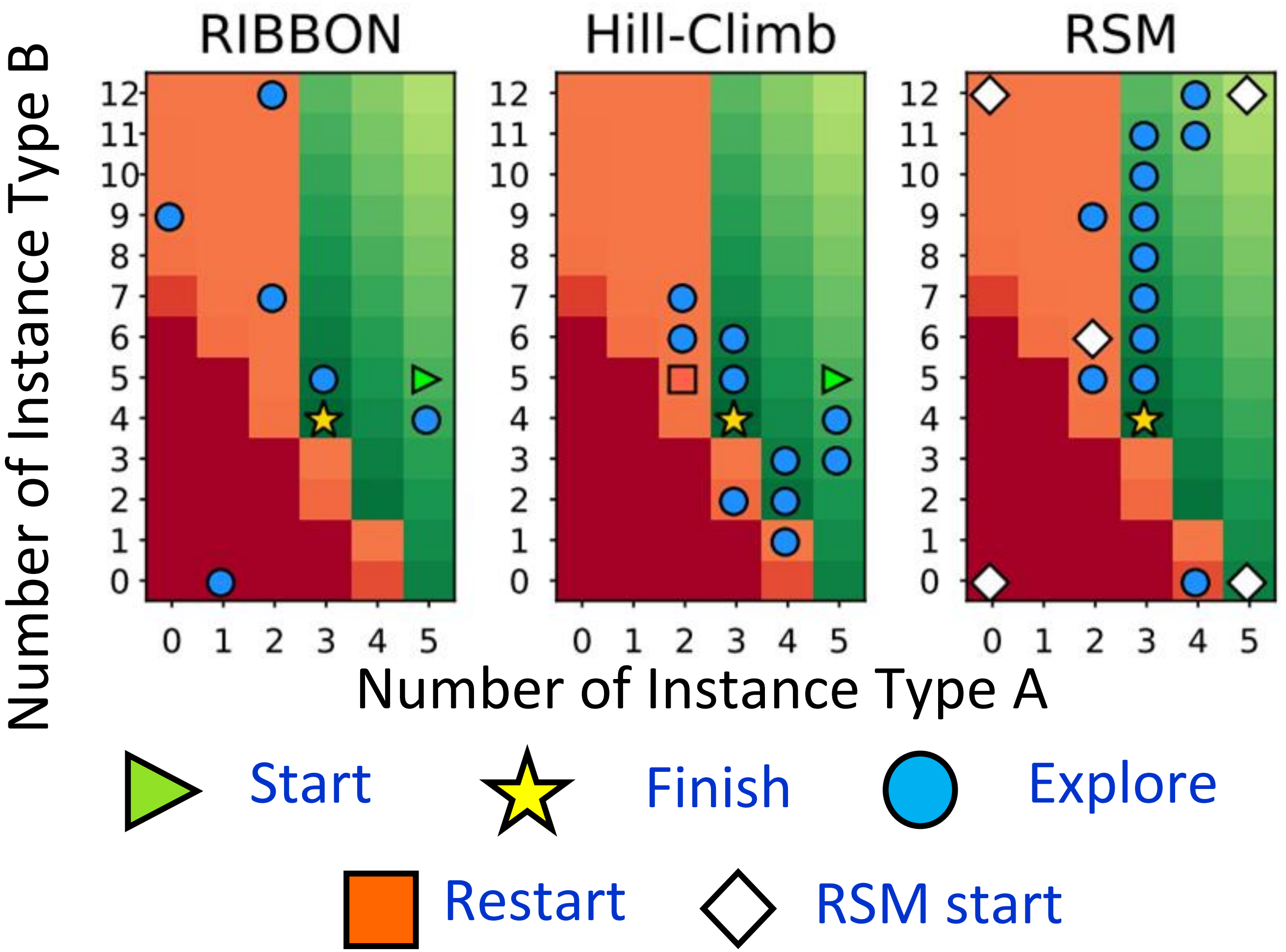
Why does RIBBON outperform other competing techniques?

QoS-violating configurations

More QoS-violations ⋙ Less QoS-violations

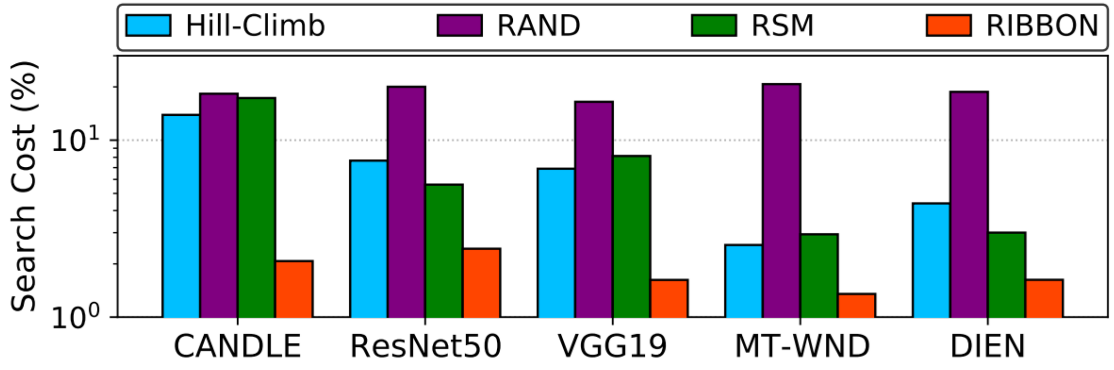
QoS-honoring configurations

Expensive ⋙ Cheaper

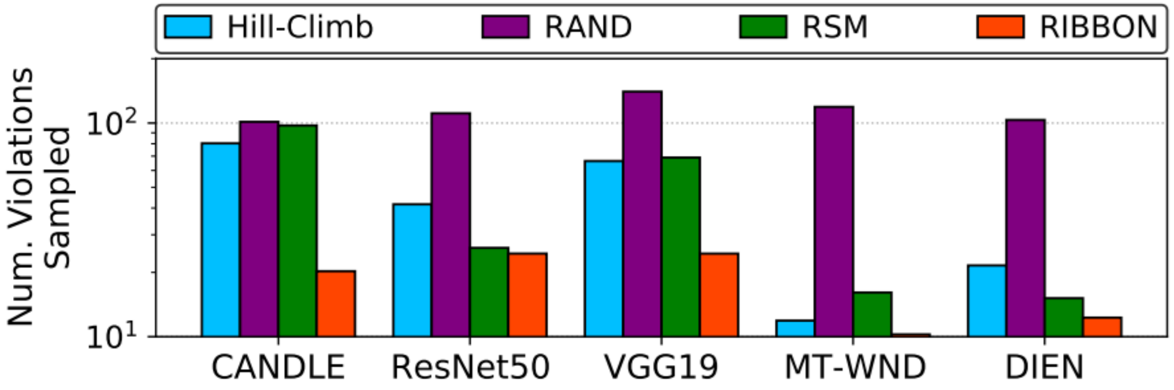


RIBBON performs most cost-effective and QoS-friendly configuration exploration process.

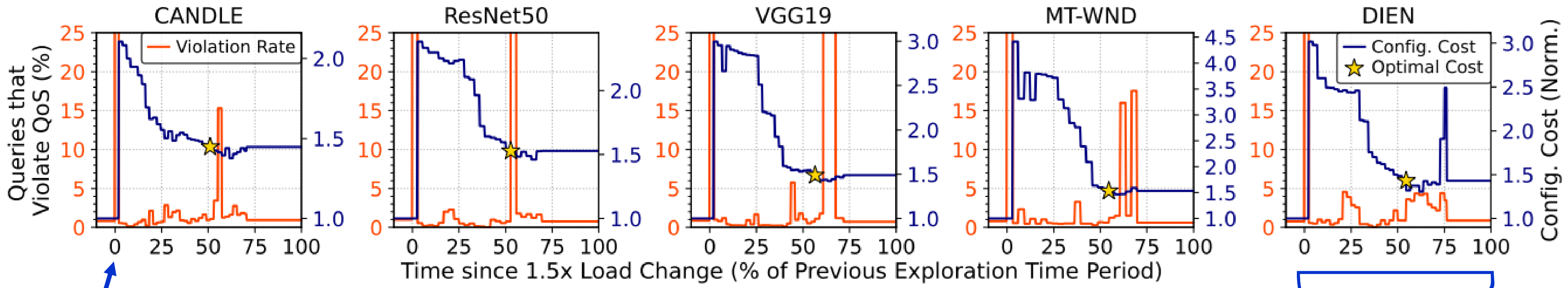
Low cost incurred during optimal configuration exploration



Low QoS violation during optimal configuration exploration



RIBBON adapts to load changes quickly.



Load Change

New optimal after load change

Previous convergence time

Load change: inference queries arrive 1.5x more frequently

RIBBON finds a new configuration with lowest cost while meeting QoS quickly and with low QoS violation rate

Tie New Ribbons!

Baolin Li <li.baol@northeastern.edu>

Research was sponsored by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.