Powering the Service Responsiveness of Deep Neural Networks with Queuing Models

Evgenia Smirni

Feng Yan, Yuxiong He, Olatunji Ruwase,

Paper appeared at Supercomputing 2016

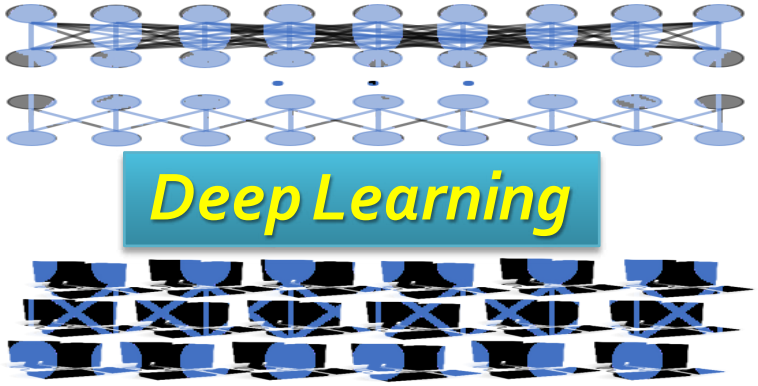
WHY DEEP LEARNING?

350 million *photos* uploaded daily to Facebook

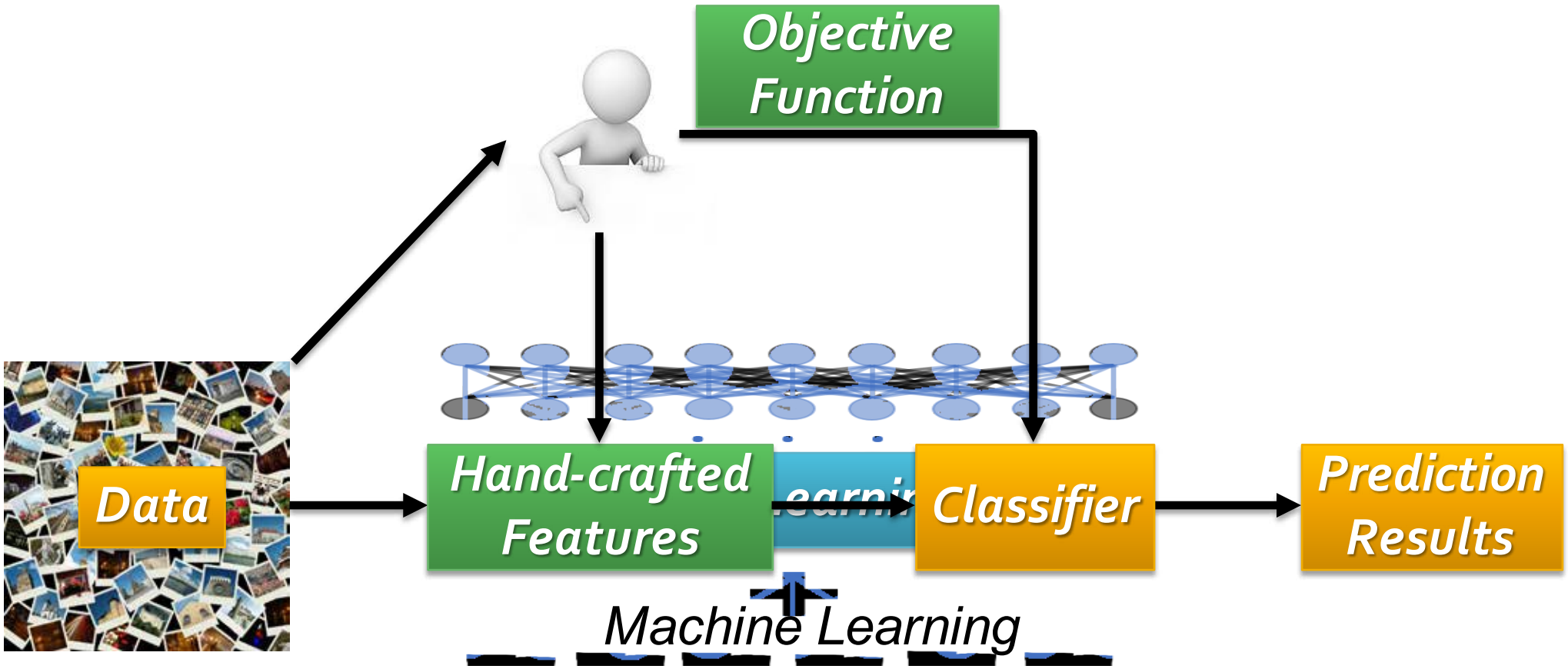
300 hours of new *video* uploaded every minute to YouTube

Vision

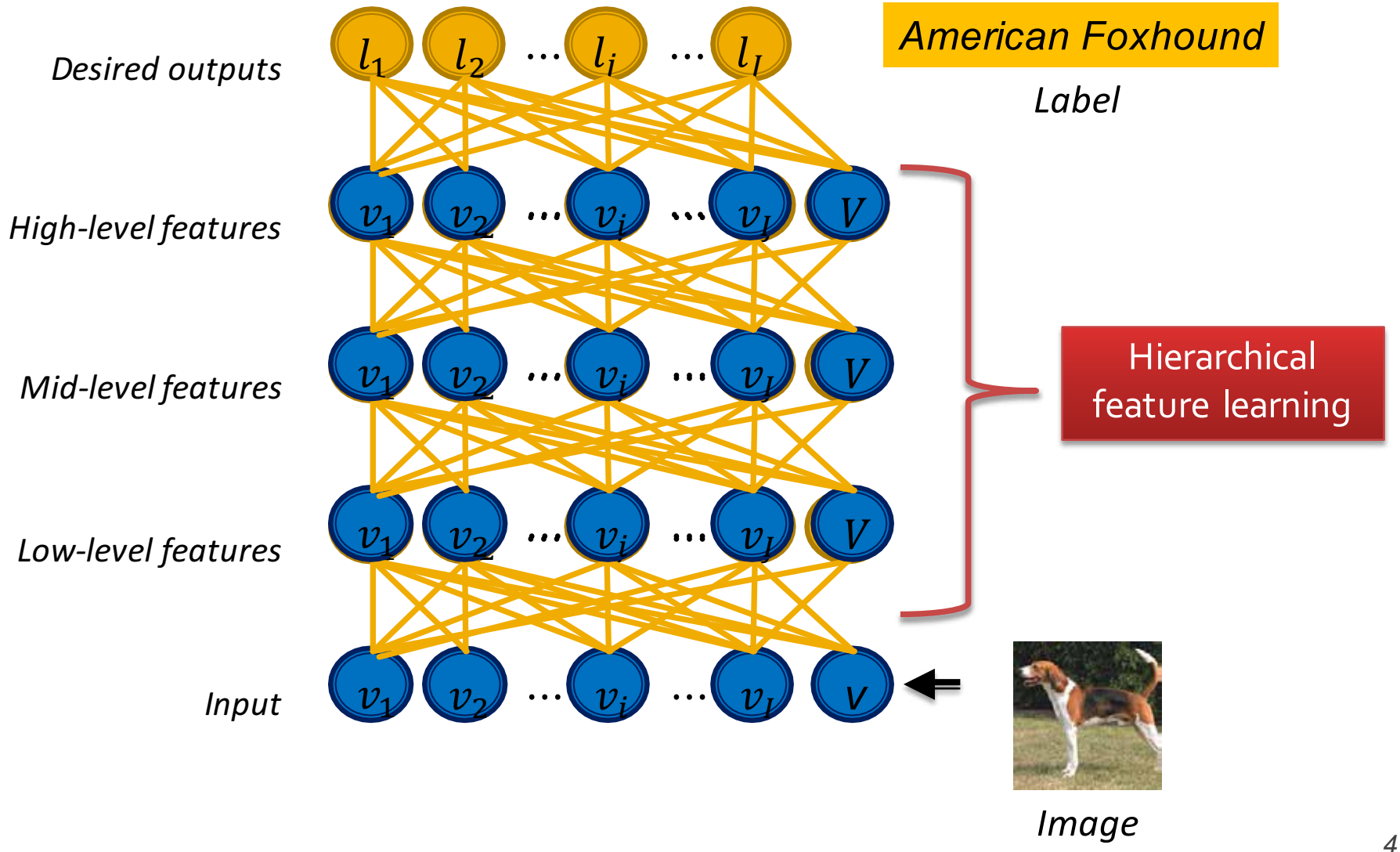
Speech



WHAT IS DEEP LEARNING?



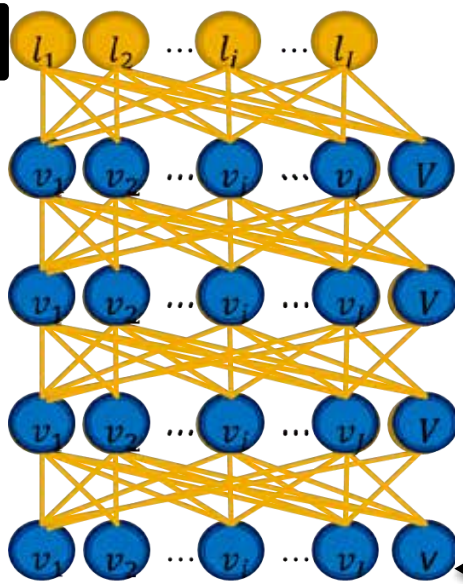
DEEP NEURAL NETWORK (DNN)



DEEP LEARNING SERVICE

Serving Cluster

Multi-core Processor



Feed-forward Evaluation

Service: Classify Dog

Arrival: random arrival of requests

(Load: requests / sec)

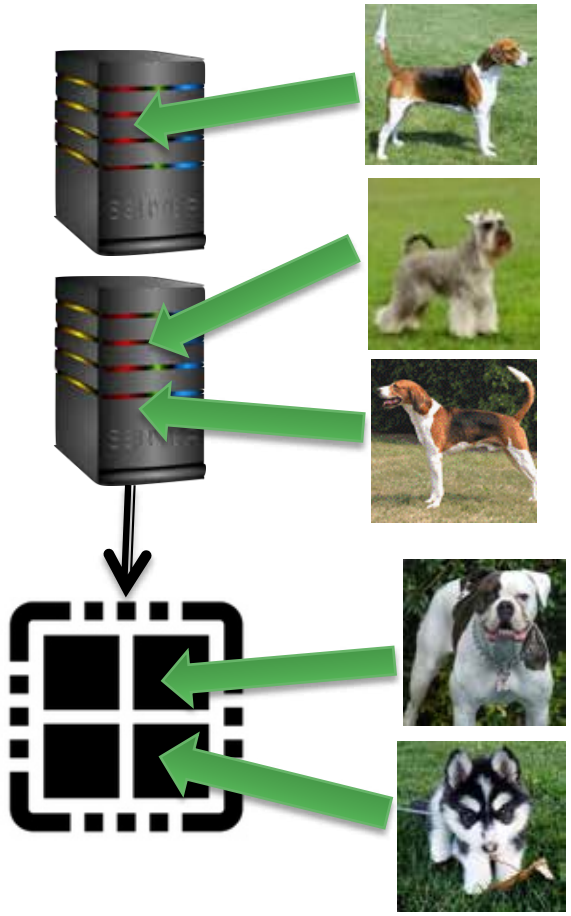
Performance Issues

Sceniaro1: Given cluster, minimize Latency?

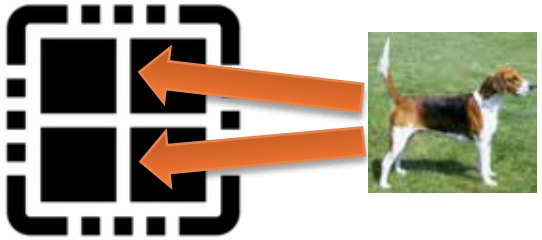
Sceniaro2: Given SLO, minimize Resources?

PARALLEL CONFIGURATION

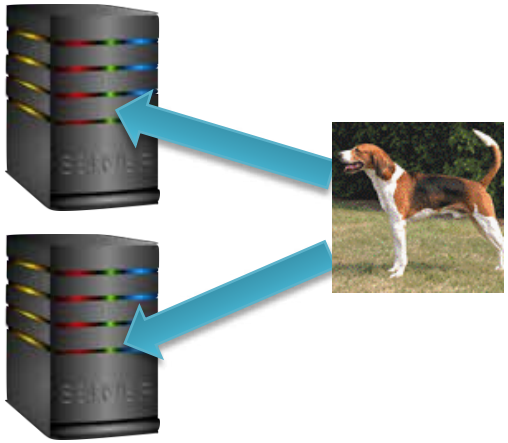
Service Parallelism



Intra-node Parallelism



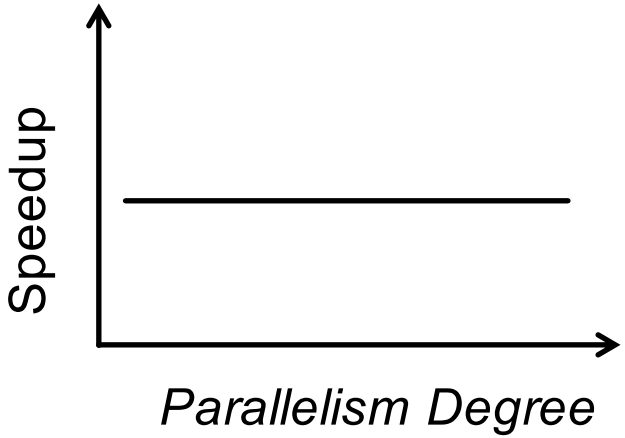
Inter-node Parallelism



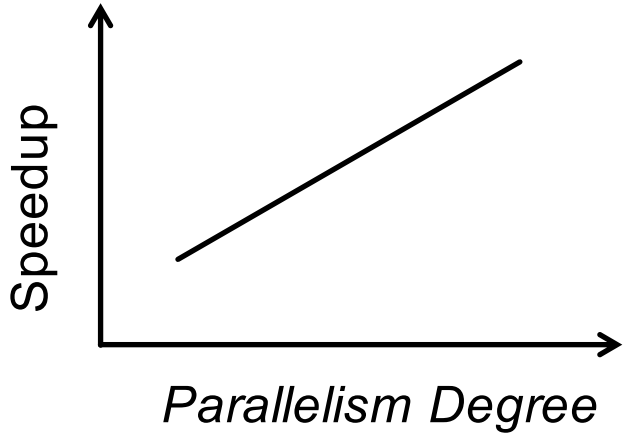
IDEAL SPEEDUP

Hypothesized Speedup Metric: Service Rate

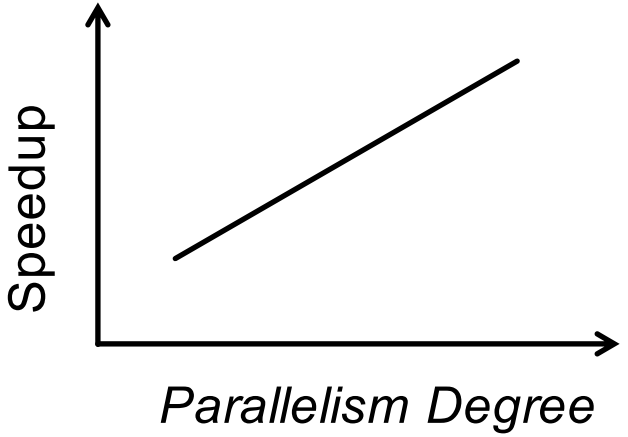
Service Parallelism (Same Node)



Intra-node Parallelism



Inter-node Parallelism



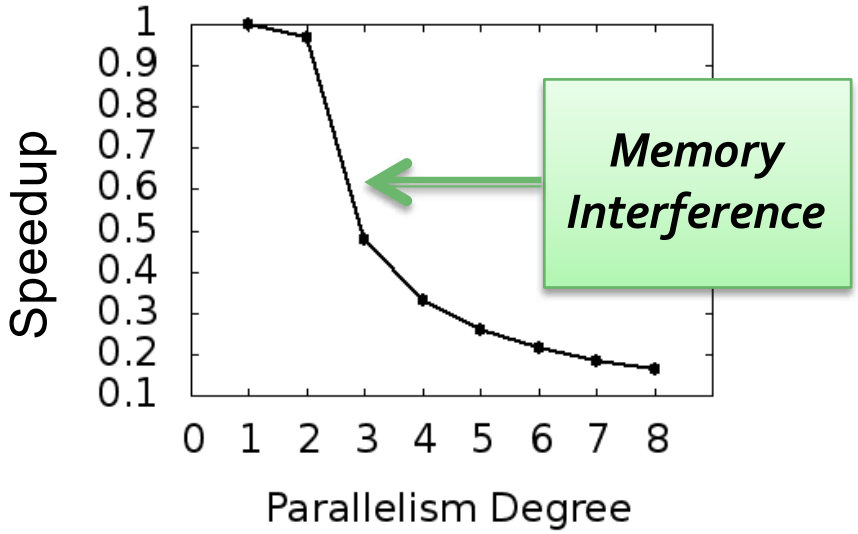
REALITY

App: ImageNet 22K

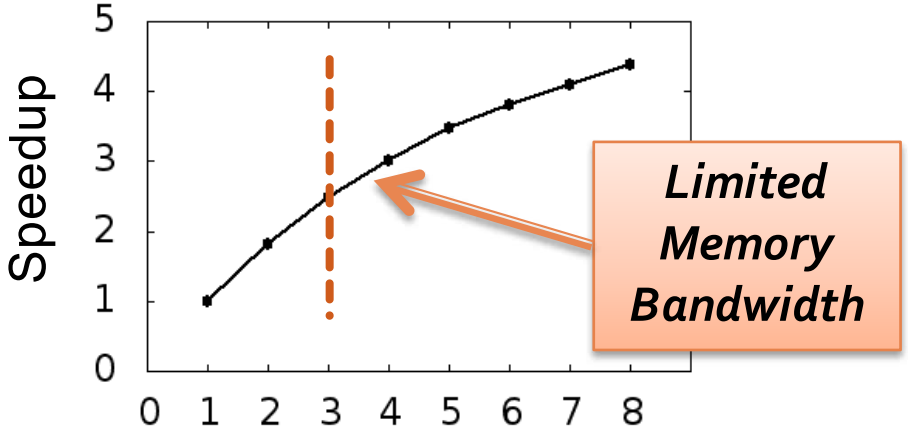
Machine: 8-core, 2.1GHz Processor
64G Memory

Metric: Service Rate

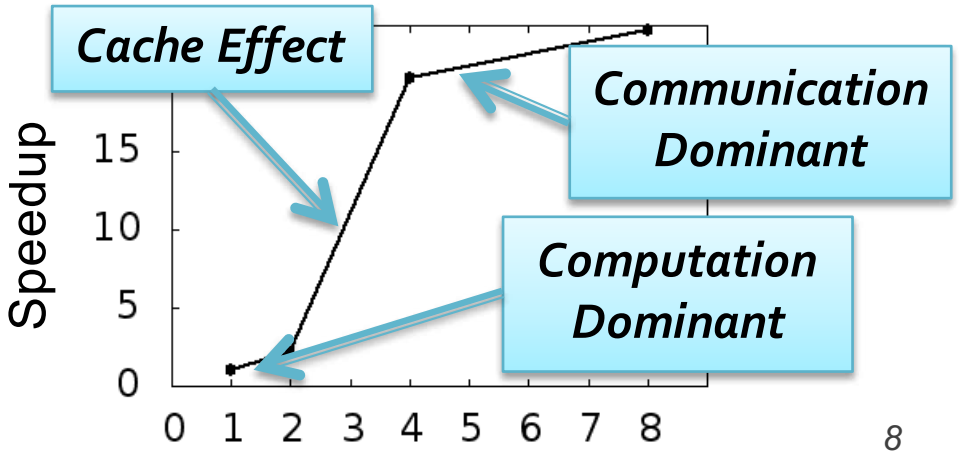
Service Parallelism (Same Node)



Intra-node Parallelism



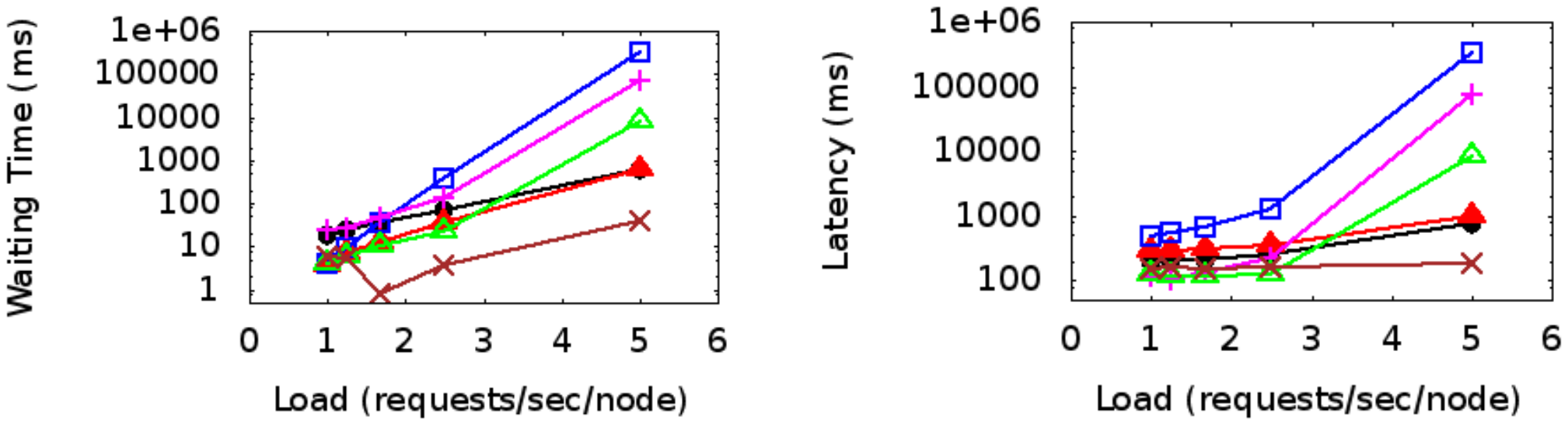
Inter-node Parallelism



EVEN MORE COMPLEX

App: ImageNet 22K
Machine: 8-core, 2.1GHz CPU
64G Memory
Measure: Waiting Time (ms)
Latency (ms)

Config.	Service	Inter-node	Intra-node
Config1	1	1	8
Config2	2	1	4
Config3	4	1	2
Config4	1	4	8
Config5	2	4	4
Config6	4	4	2



Queuing effects add complexity

SIMPLE SOLUTION

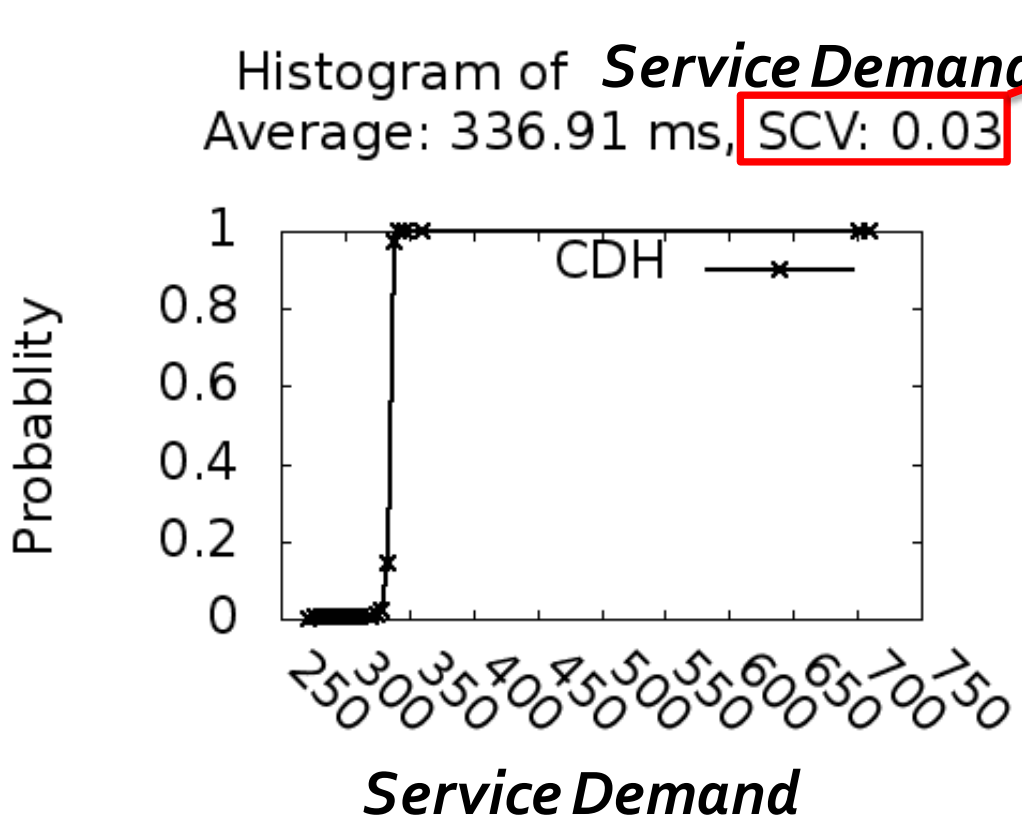
Simple Solution1: Extensive Profiling

 # Configs * # Load Level * Ti… = Profiling Cost

80 * 10 …nn = 19 days

Failed due to Cost

OBSERVATION



Almost no variance

SCV(Squared Coefficient of Variation): $SCV = \frac{Variance}{mean^2}$

- Erlang: <1
- Exponential: =1
- Hyper-exponential: >1

Deterministic Service Demand

OUR APPROACH

Lightweight Profiling

Profile Service Demand under different Parallelism (i.e., no random arrival)

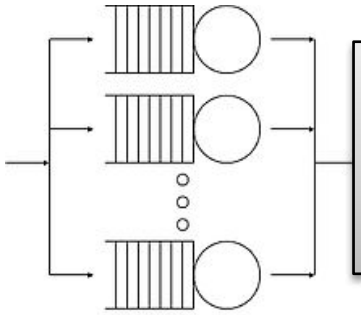
Configs * Time for Each Exp = Profiling Cost

80 * 0.07 min = 5.5 min (compared to 19 days)

+

Queuing–based Prediction Model: Captures Queuing Effects

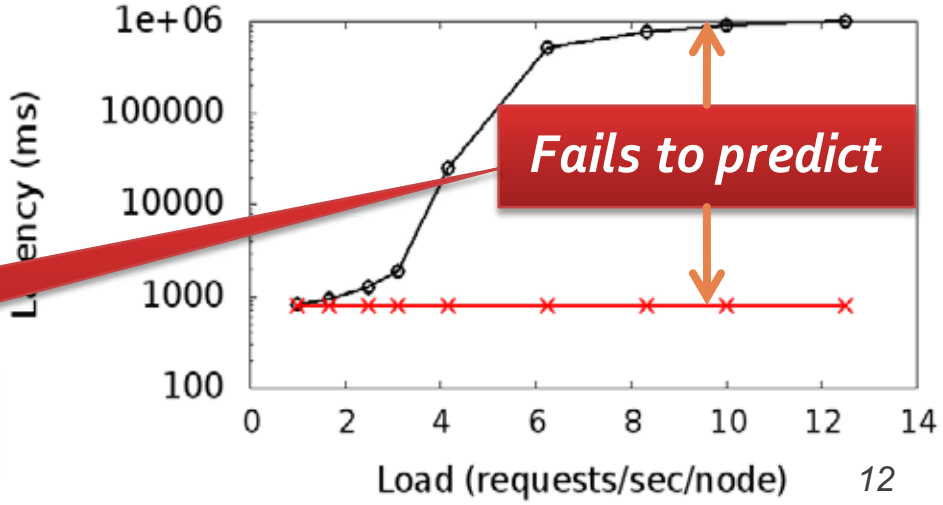
Use M/D/c queue to estimate the latency



M: random arrivals

D: deterministic service

c: multiple abstracted servers



Problem: load-dependent service

OUR APPROACH

Lightweight Profiling

Profile Service Demand under different Parallelism (i.e., no random arrival)

Configs * Time for Each Exp = Profiling Cost

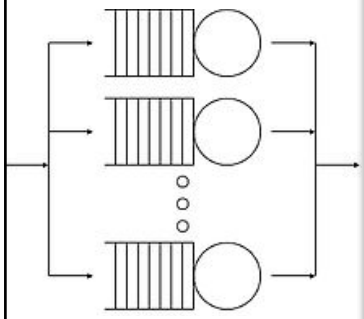
80 * 0.07 min = 5.5 min

+

Queuing–based Prediction Model: Captures Queuing Effects

Leverage: Cosmetatos' Approximation (uses M/M/c to approximate M/D/c)

Extension: use M/M-interf./c queue to approximate M/D-interf./c queue



M: random arrival

M-interf.: interference-aware

D: deterministic service

D-interf.: interference-aware d… service

c: multiple abstracted servers

Solve M/D-interf./c queue

Extend Cosmetatos' Approximation

OUR APPROACH

Queuing–based Prediction Model: Captures Queuing Effects

Leverage: Cosmetatos' Approximation (use M/M/c to approximate M/D/c)

Extension: use M/M-interf./c queue to approximate M/D-interf./c queue

$$W^{M/D_{interf}/c}(\lambda) \approx \sum_{i=1}^{c} \frac{p_{i-1}}{\mu_i} + \frac{\prod_{i=1}^{c} \rho_i}{\mu_c \cdot c! \cdot (1-\rho)}$$

Average Service Time

$$+ \frac{1}{2}(1 + f(s) \cdot g(\rho)) \cdot \frac{p_0 \cdot \prod_{i=1}^{c} \rho_i}{\lambda \cdot c!} \cdot \frac{\rho}{(1-\rho)^2}$$

Average Waiting Time

$$\left(p_n = \begin{cases} \frac{\prod_{i=1}^{n} \rho_i}{n!} \cdot p_0 & (0 \le n \le c-1) \\ \frac{\rho^{n-c} \cdot \prod_{i=1}^{n} \rho_i}{c!} \cdot p_0 & (n \ge c), \end{cases} \quad f(s) = \frac{(c-1) \cdot (\sqrt{4+5c}-2)}{16c}, \quad g(\rho) = \frac{1-\rho}{\rho} \right)$$

EVALUATION

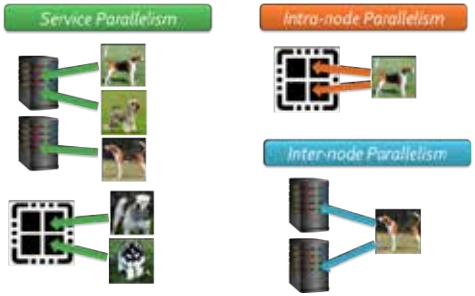
Experiment Setup

Image Recognition Task: ImageNet-22K

- 256x256 RGB images in 22,000 categories
- ~2Bn. Parameters model
- Random Arrivals

Distributed DNN Serving System

- Based on Adam [OSDI'14]
- Support: Service, Intra-node, Inter-node parallelisms



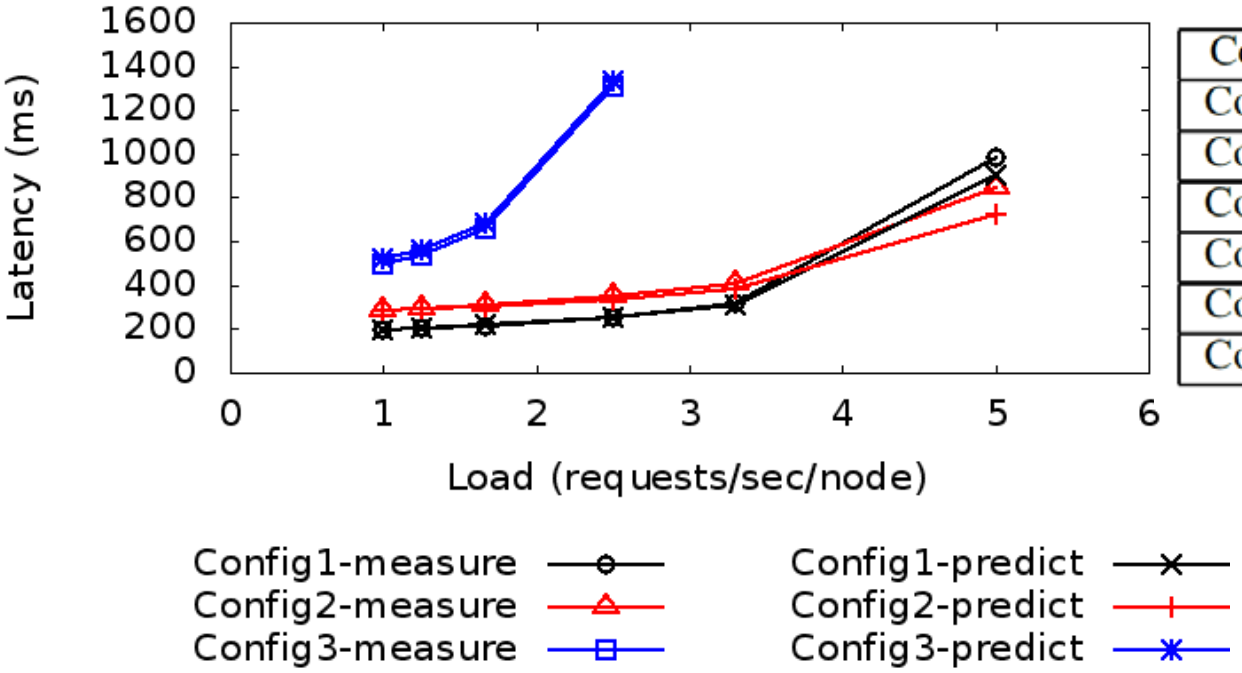
Hardware

- 20 nodes, 10 Gbps Ethernet cluster
- Intel Xeon E5-2450: 2.1GHz, 16 core, 64GB RAM

EVALUATION

Prediction Accuracy

ImageNet-22K: Latency VS Load



Config.	Service	Inter-node	Intra-node
Config1	1	1	8
Config2	2	1	4
Config3	4	1	2
Config4	1	4	8
Config5	2	4	4
Config6	4	4	2

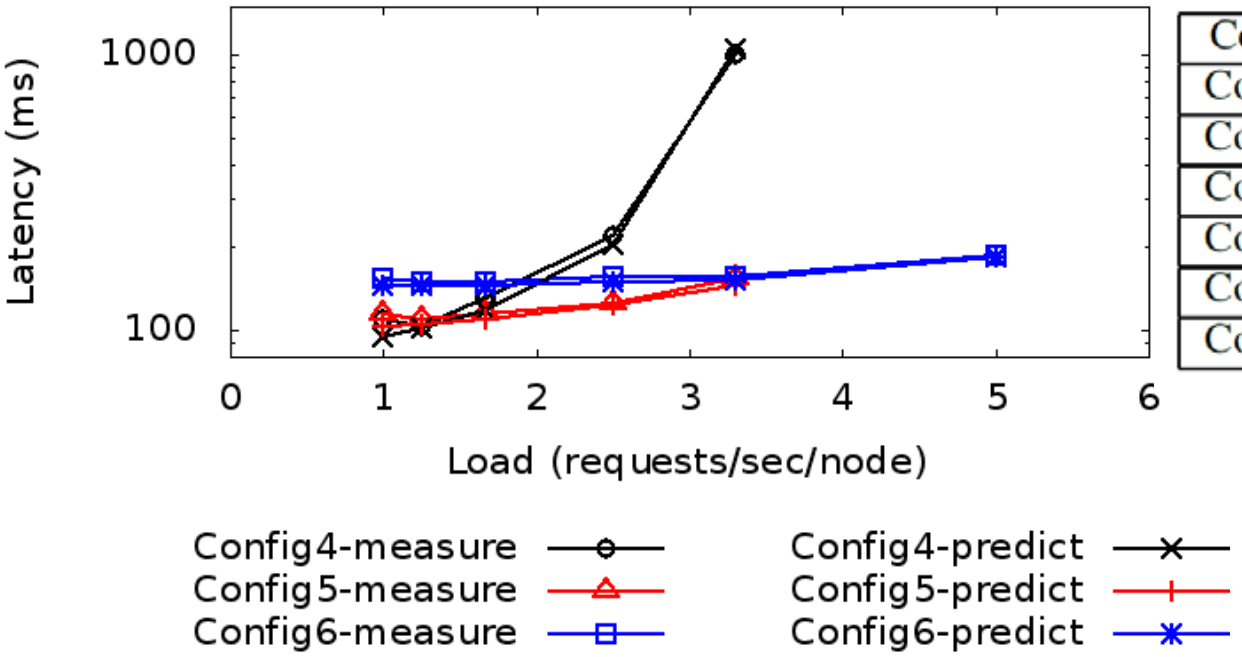
Captures Trend Change

√ Good Accuracy

EVALUATION

Prediction Accuracy

ImageNet-1K: Latency VS Load



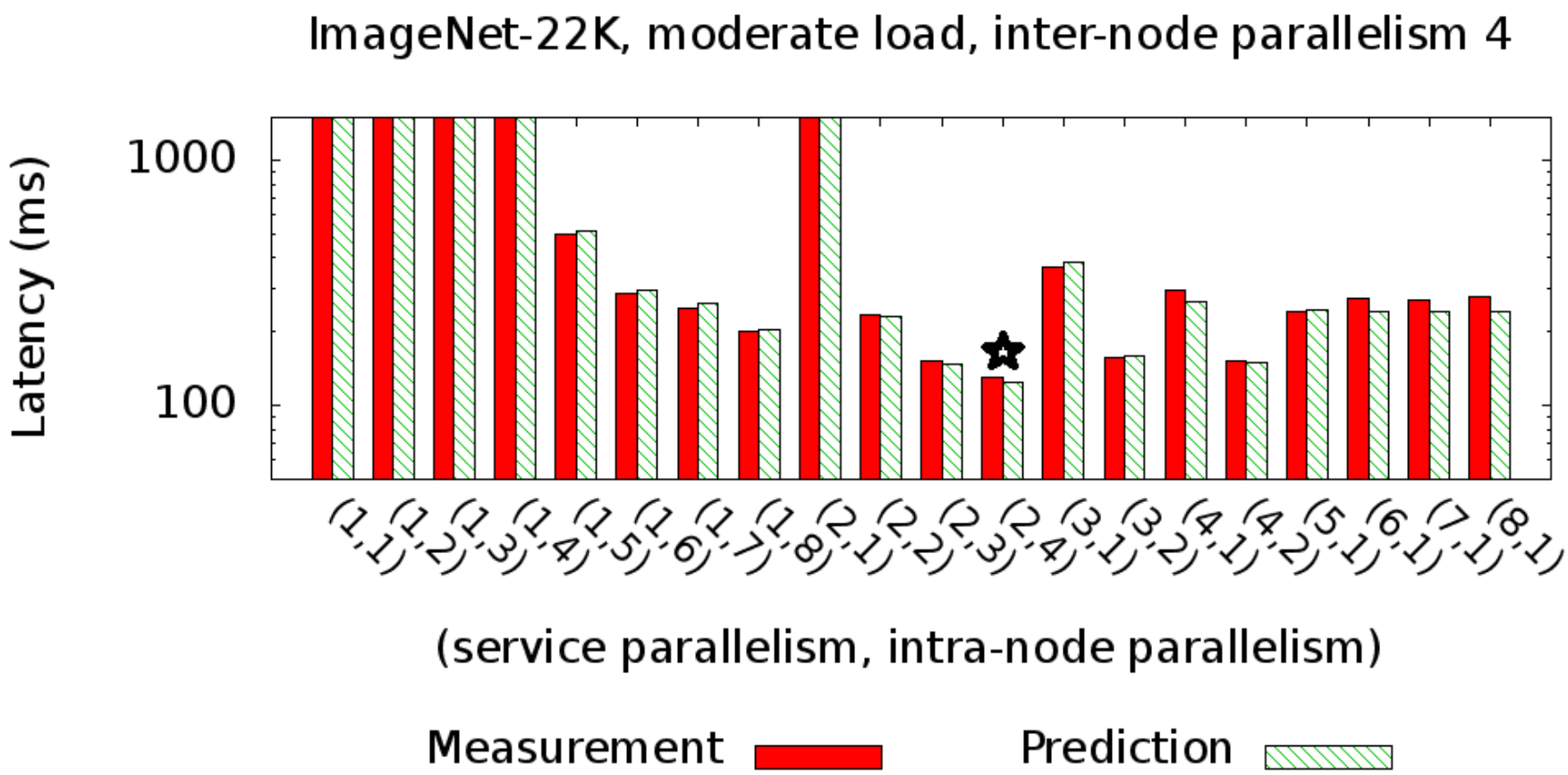
Config.	Service	Inter-node	Intra-node
Config1	1	1	8
Config2	2	1	4
Config3	4	1	2
Config4	1	4	8
Config5	2	4	4
Config6	4	4	2

Captures Trend Change

√ Good Accuracy

EVALUATION

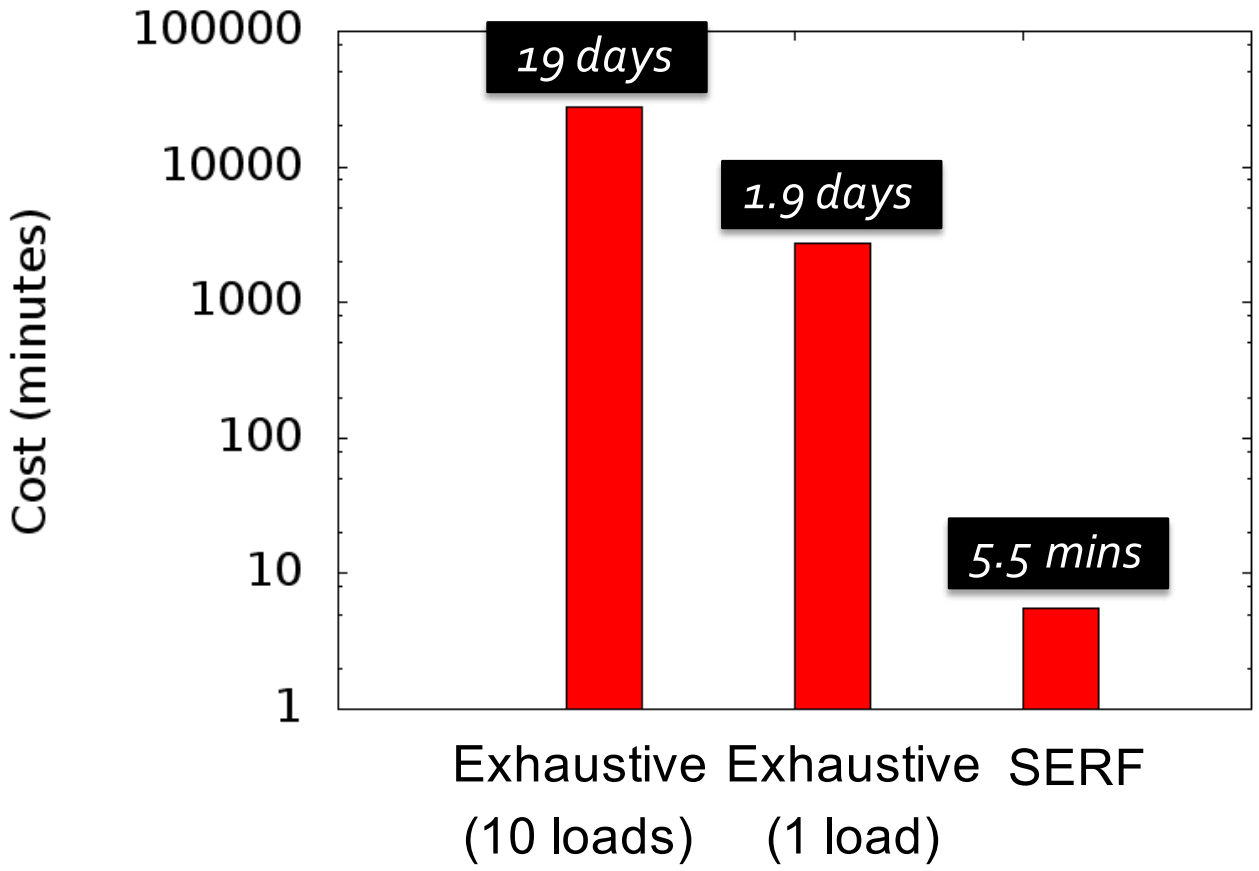
Prediction Accuracy



Identify Optimal Config.
√ Good Accuracy

EVALUATION

Cost



Fast Deployment For Online Use

Low Profiling Cost
√ Lightweight

SUMMARY

Scheduling framework for Deep Neural Network Serving

- ✓ *Automatic*

Take Away:

Balance measurements with modeling cost and complexity.

Make the model simple enough but not too simple…

Accurate performance model:

- *Performance prediction with <5% error*

Efficient Scheduler:

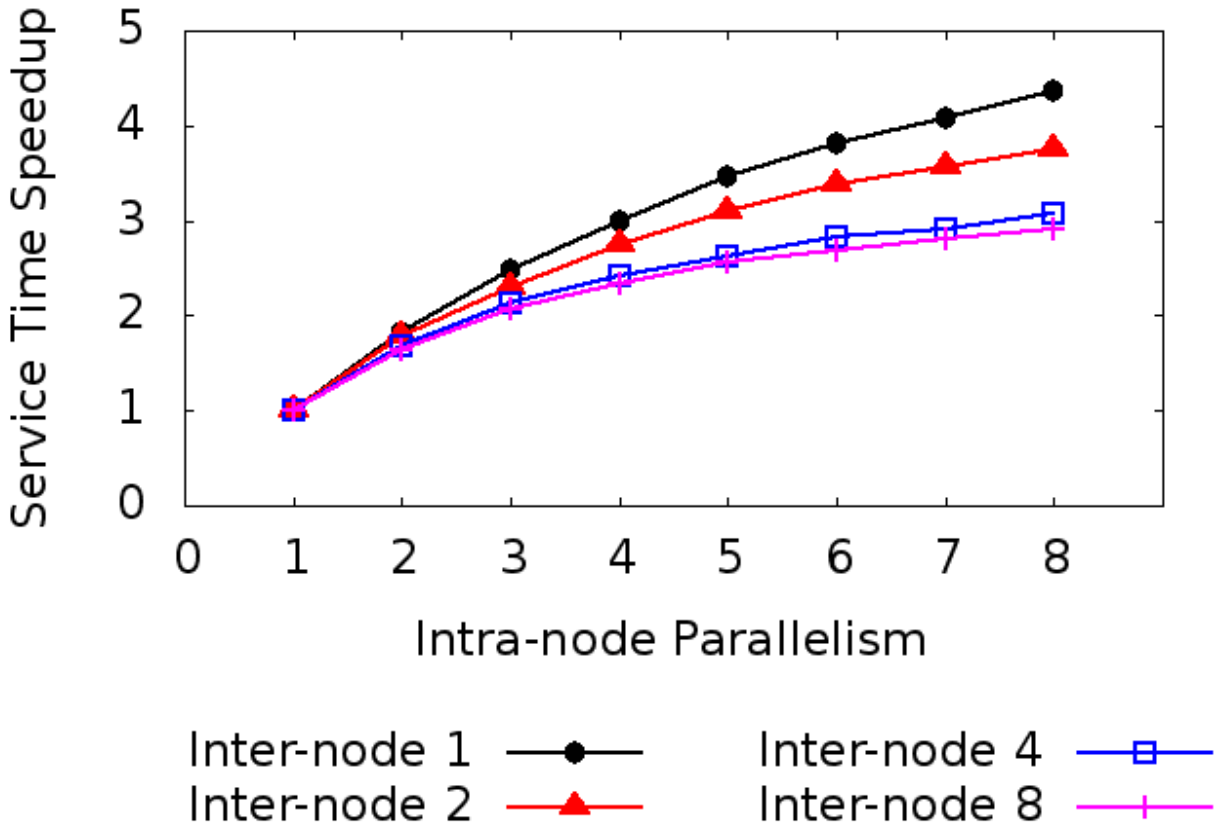
- *Adapts to dynamic load*
- *Supports various scheduling requirements*

THANK YOU!

Questions?

MORE COMPLEX

Relation between Inter-node and Intra-node Parallelism

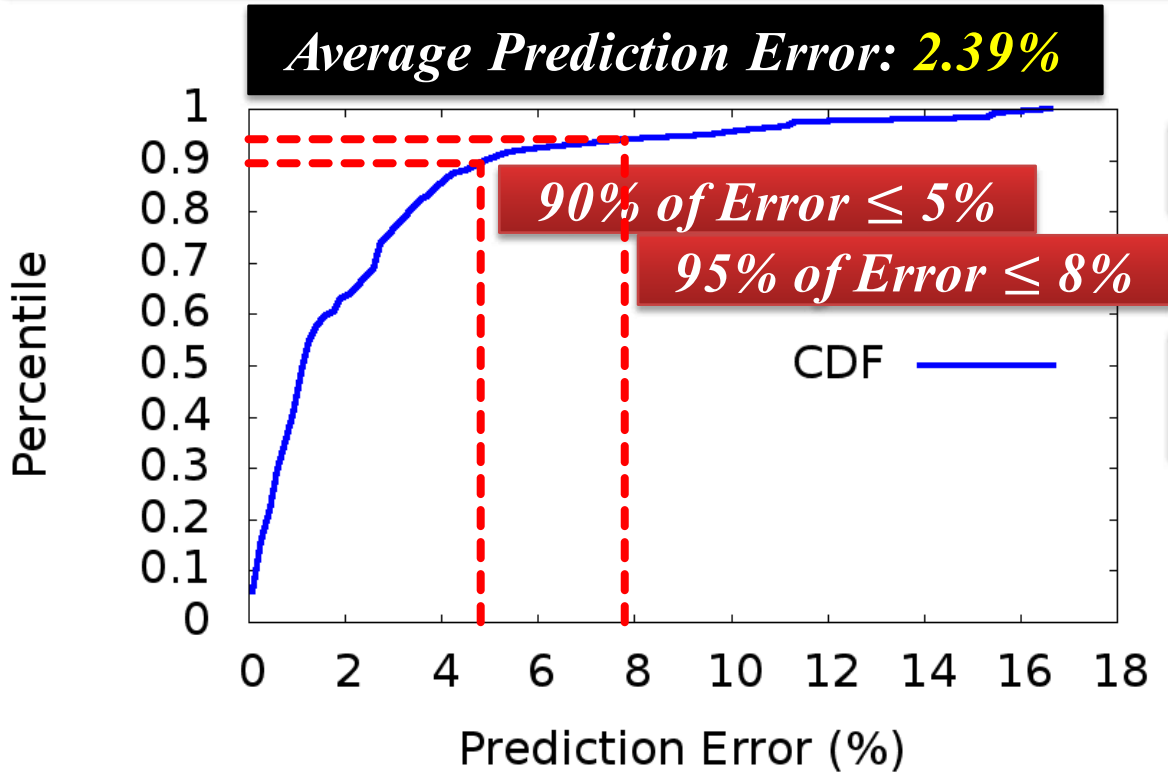


One parallelism can affect another parallelism!

EVALUATION

Prediction Accuracy

ImageNet-22K: Prediction Error Distribution

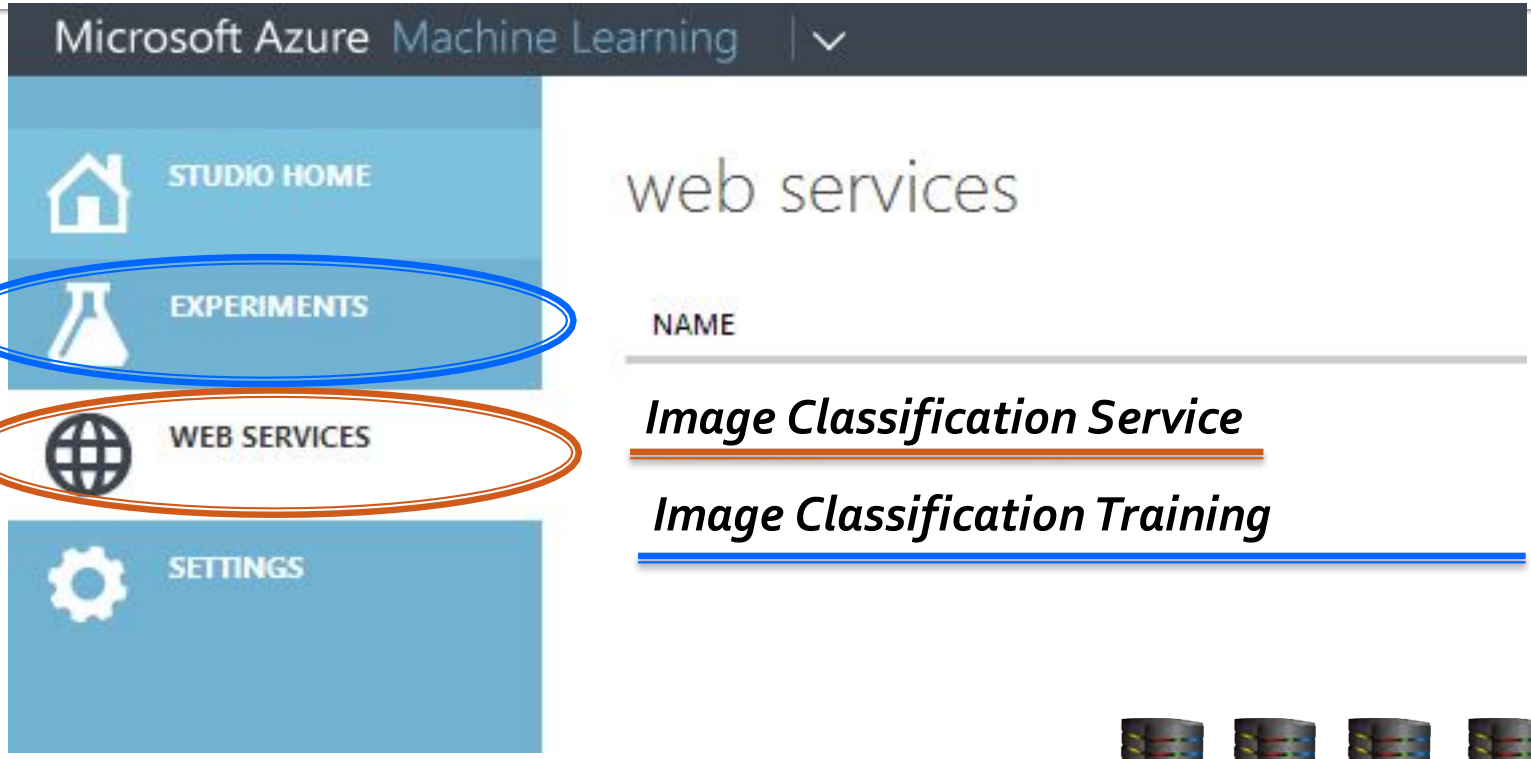


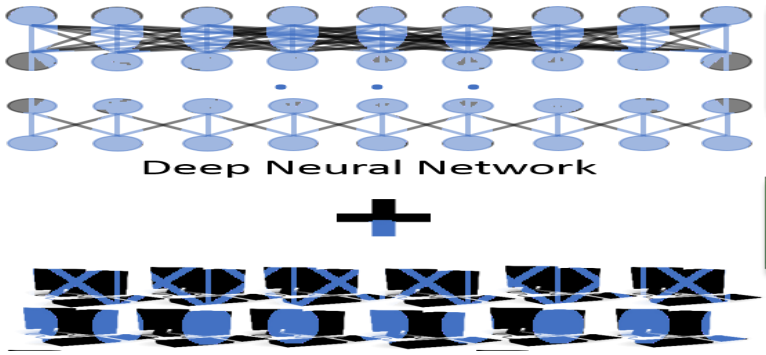
800 Prediction cases

CDF: Cumulative Distribution Function

Low Prediction Error
√ Good Accuracy

DEEP LEARNING SERVICE





Trained DNN

1. Fast Deployment

Deploy

2. Good Performance

Serving Cluster