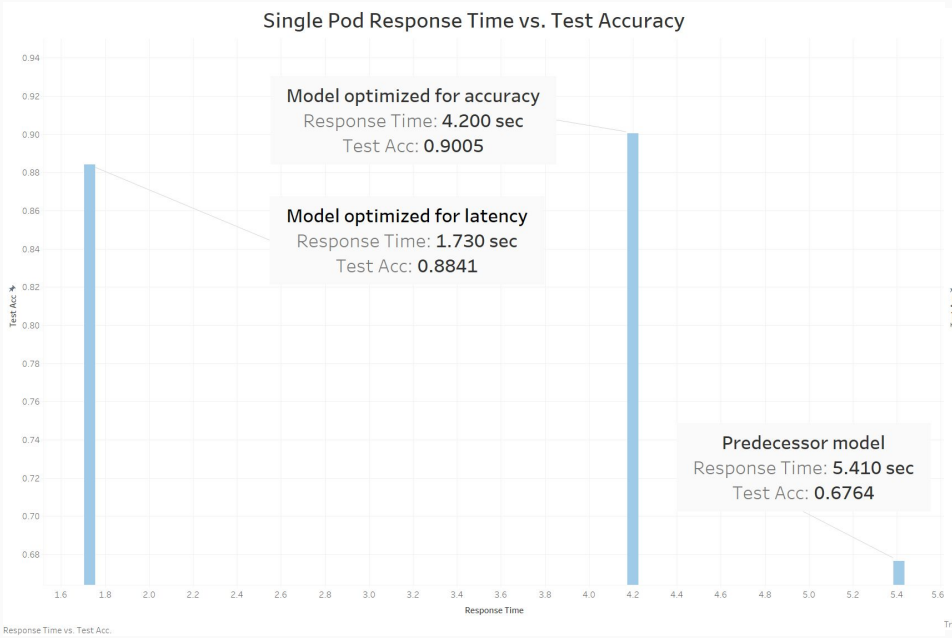
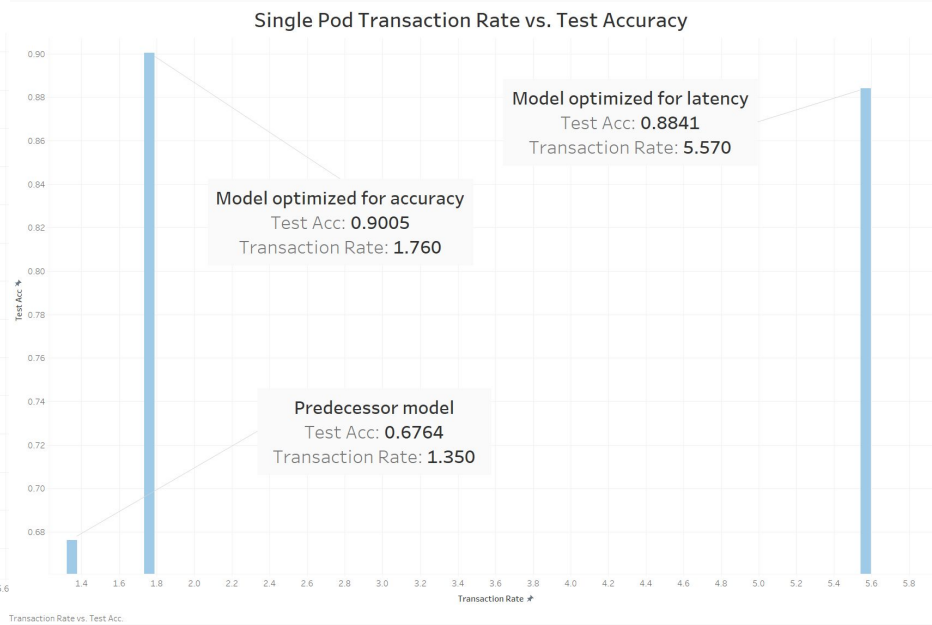
Model Choices

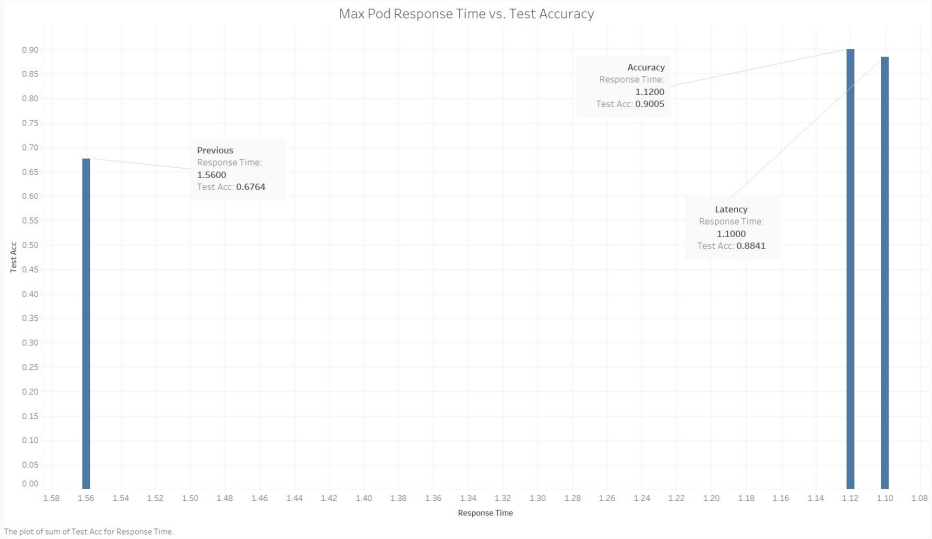
	Optimizing Accuracy	Optimizing Latency
Transforms used in data augmentation	rescale: Scales pixel values to [0,1]. rotation_range: Randomly rotates images up to ±10 degrees. zoom_range: Randomly zooms images up to 10%. width_shift_range: Shifts image width by up to 10%. height_shift_range: Shifts image height by up to 10%. shear_range: Shears images by 0.1 degrees. horizontal_flip: Randomly flips images horizontally. vertical_flip: Randomly flips images vertically. fill_mode: Fills new pixels using the nearest method. brightness_range: Adjusts brightness by 80% to 120%. channel_shift_range: Shifts color channels by up to ±20.	rescale: Scales pixel values to [0,1]. rotation_range: Randomly rotates images up to ±10 degrees. zoom_range: Randomly zooms images up to 10%. width_shift_range: Shifts image width by up to 10%. height_shift_range: Shifts image height by up to 10%. shear_range: Shears images by 0.1 degrees. horizontal_flip: Randomly flips images horizontally. vertical_flip: Randomly flips images vertically. fill_mode: Fills new pixels using the nearest method. brightness_range: Adjusts brightness by 80% to 120%. channel_shift_range: Shifts color channels by up to ±20.
Base model (Name, Size, Top-1 ACC, CPU Inference Time)	ResNet101V2, 171MB, 77.2%, 72.7ms	MobileNetV2, 14MB, 71.3%, 25.9ms
# of epochs. optimizer, and learning rate for training classification head	39 epochs, Adam, 0.01	30 epochs, Adam, 0.001
# of layers un-frozen	8	5
# of epochs, optimizer, and learning rate for fine-tuning	50 epochs, Adam, 0.0005	100 epochs, Adam, 0.0005
Final accuracy	0.9005	0.8841

Performance of models when deployed as a single pod





Performance of models when deployed as a "max-size" deployment



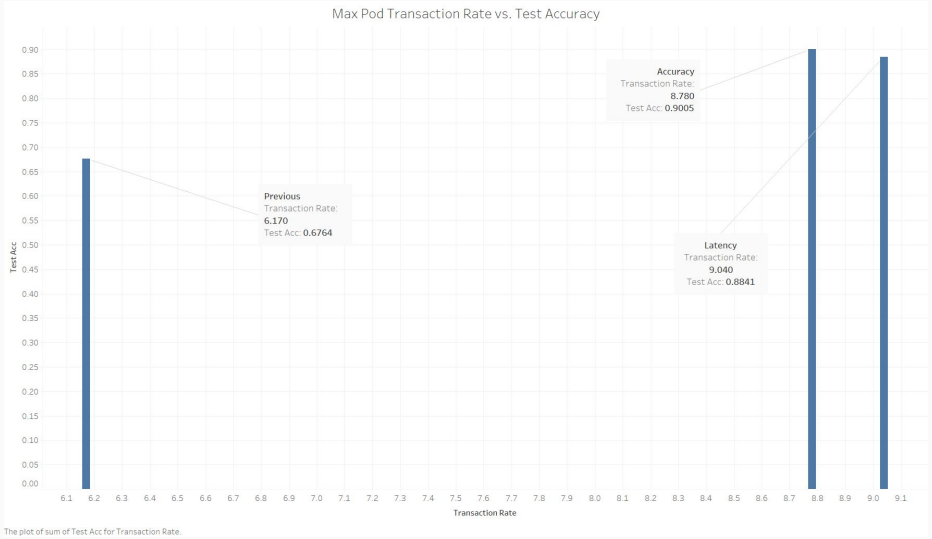


Table of configurations for max-size deployment

	Previous	Accuracy	Latency
# of Replicas (Actually)	4	4	2
CPU Resource Requests	2405m (60%)	2405m (60%)	1405m (35%)
Mem Resource Requests	4627034Ki (57%)	4627034Ki (57%)	2529882Ki (31%)
CPU Resource Limits	4400m (110%)	4400m (110%)	2900m (72%)
Mem Resource Limits	6544139520 (79%)	6544139520 (79%)	3859784960 (46%)

Table of horizontal scaling configurations

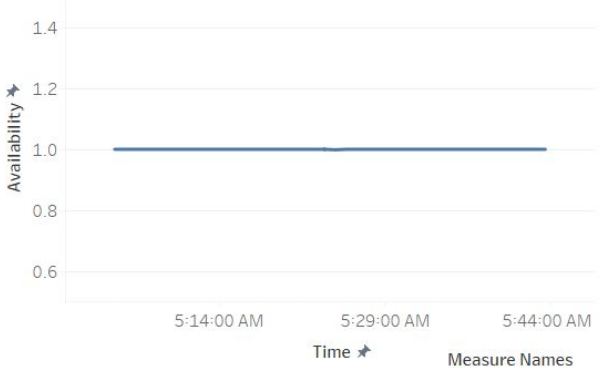
	ACCURACY	LATENCY
MinReplicas	2	2
MaxReplicas	10	8
Target CPU Utilization	60%	50%
CPU Resource Requests	1405m (35%)	405m (10%)
Mem Resource Requests	1481306Ki (18%)	443115520 (5%)
CPU Resource Limits	3400m (85%)	1400m (35%)
Mem Resource Limits	5470397696 (66%)	1175430400 (14%)

Visualize the deployment of the “accuracy” model over time

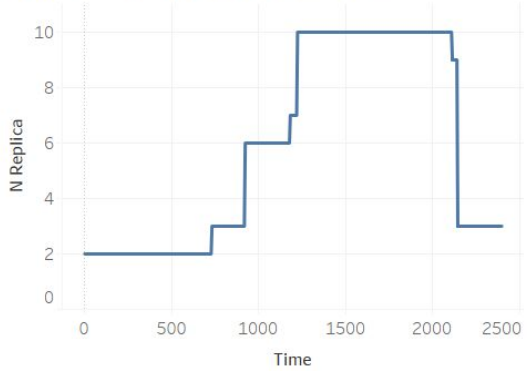
Number of requests over time

Average response time over time

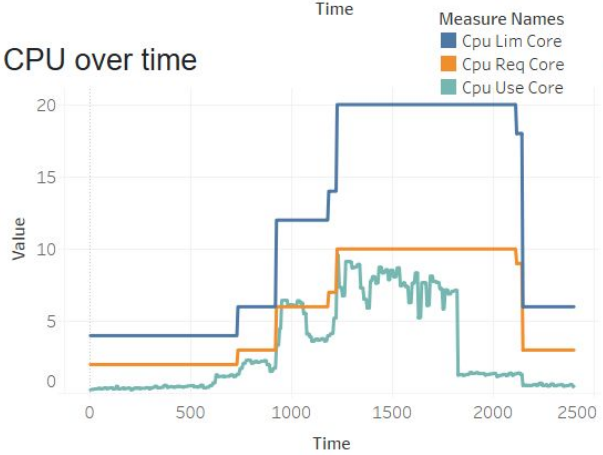
Availability over time



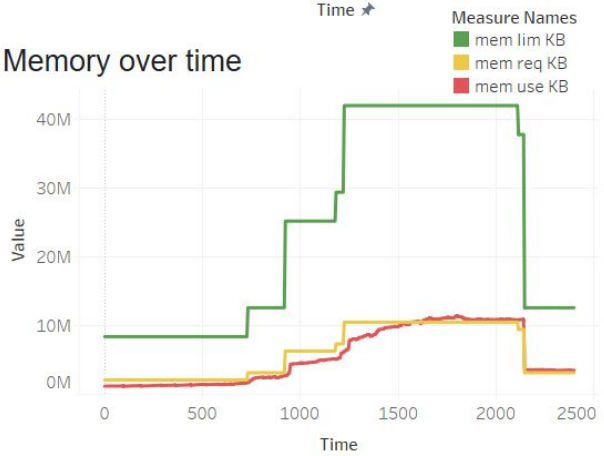
Number of replicas over time



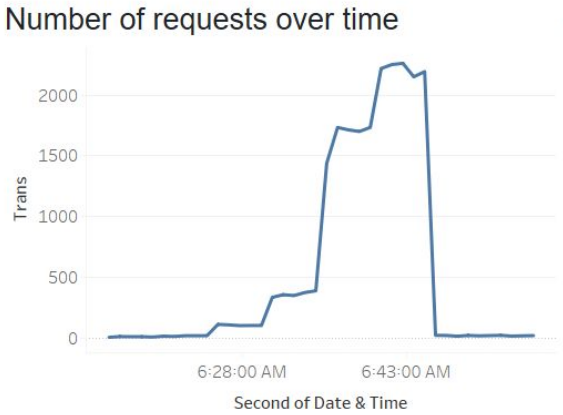
CPU over time

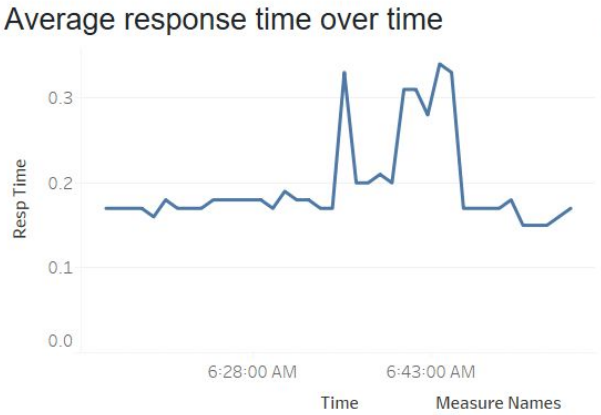


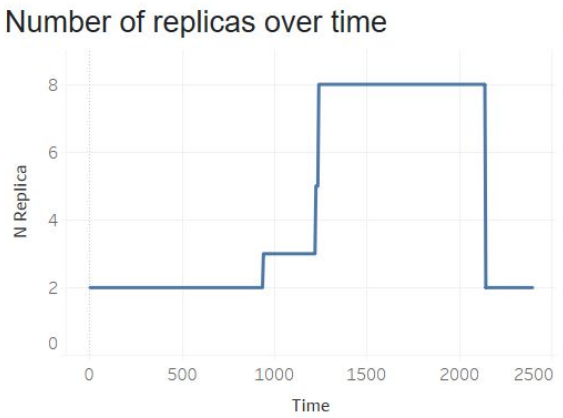
Memory over time

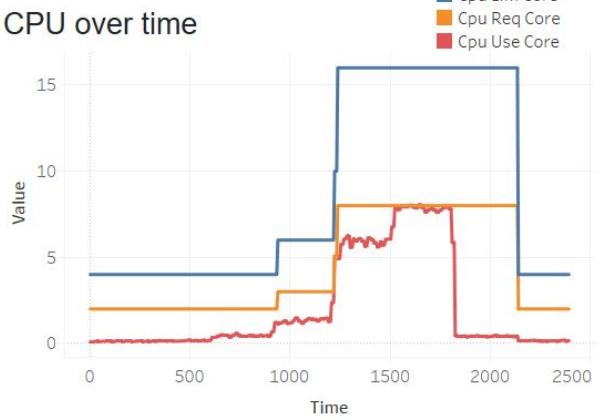


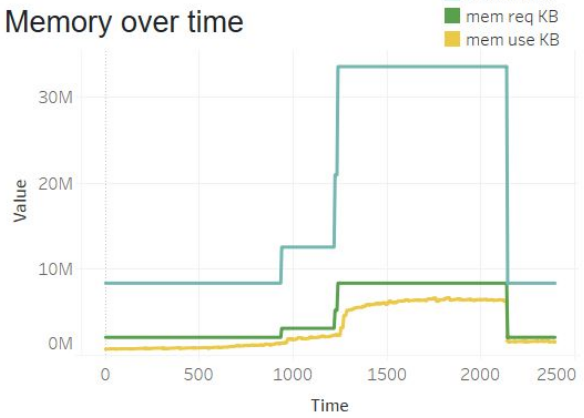
Visualize the deployment of the “ latency” model over time











Contributions

- Experimented different base models
- Presented 2 models with more Accuracy and less Latency
- Experimented different configurations of resource usage for both 2 model
- Deployed models to a K8s cluster on a 3-node server for load balancing