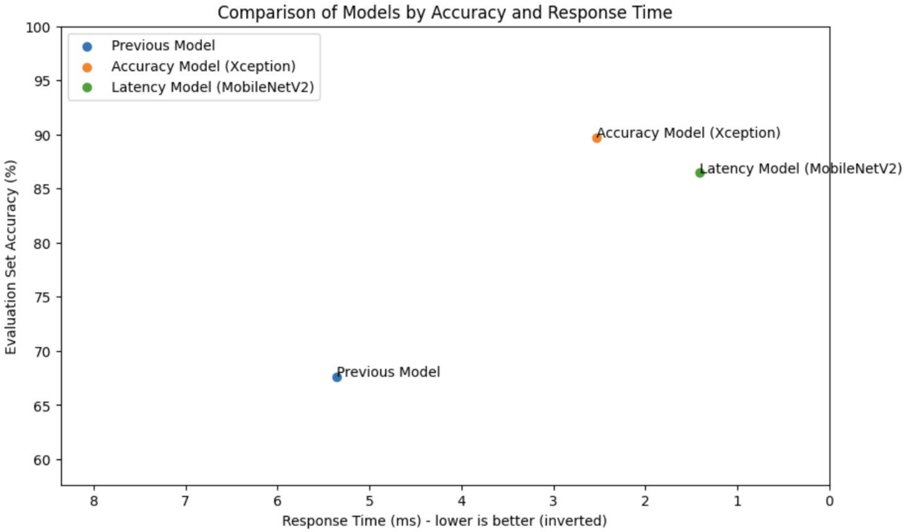
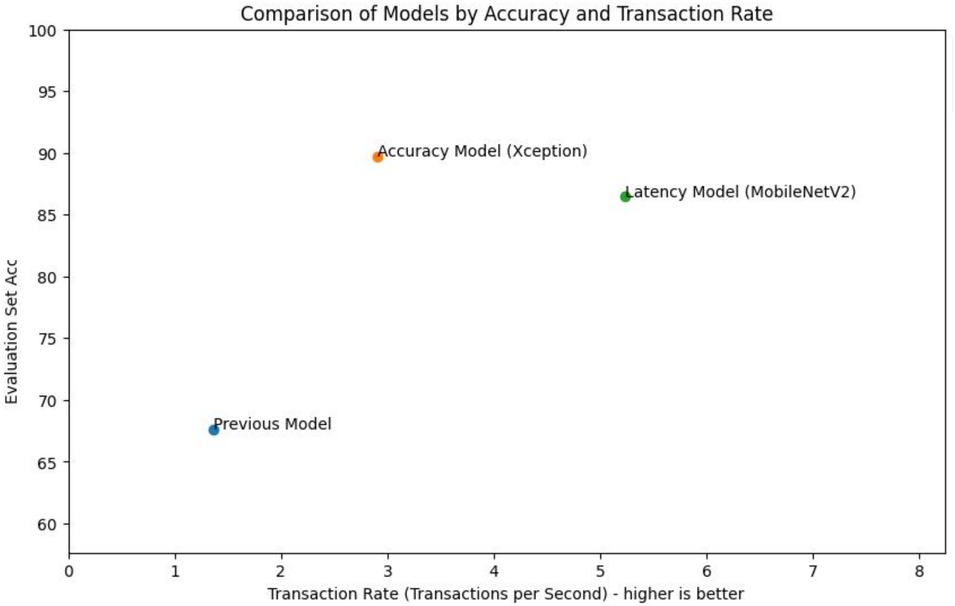
Table of model training choices

	Optimizing Accuracy (Xception)	**Optimizing Latency (MobileNetV2)**
Data Transformation/Augmentation	Random rotation, zoom, width shift, height shift, horizontal flip	Random rotation, zoom, width shift, height shift, horizontal flip
Base Model (include name, size, top-1 accuracy, CPU inference time))	Name: Xception Size: 88 MB Top-1 Accuracy: 79.0% CPU Inference Time: 109.4 ms	Name: MobileNetV2 Size: 14 MB Top-1 Accuracy: 71.3% CPU Inference Time: 25.9 ms
Number of epochs, Optimizer, and learning rate used to train classification head	20 epochs, ADAM @ 0.01 (early stopping stopped epochs at 4)	5 epochs, ADAM @ 0.01
number of layers un-frozen	5	5
Number of epochs, Optimizer, and learning rate used to further fine-tune the model	30 epochs, Adam @ 0.0001	10 epochs, ADAM @ 0.0001
final accuracy on evaluation set (test set)	89.75%	86.47%

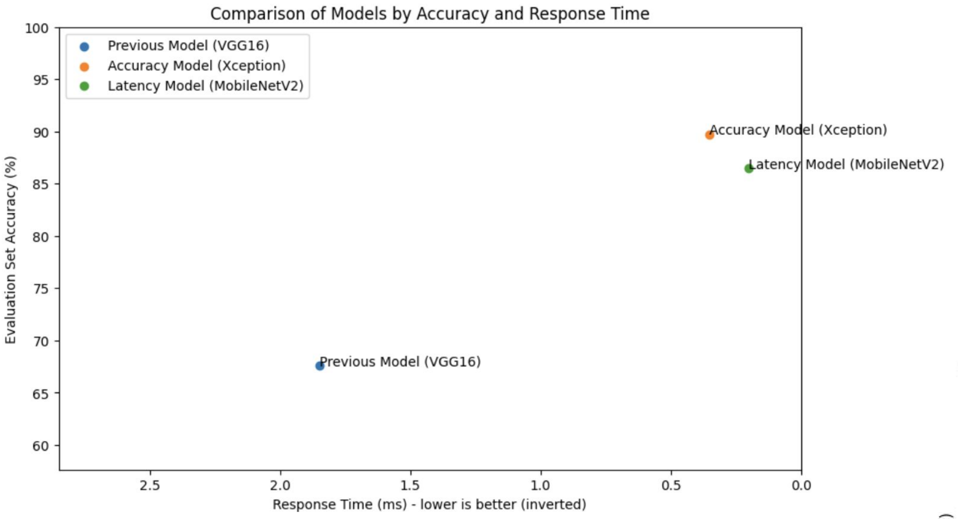
Performance of models when deployed as a single pod



Legend
Previous Model
Accuracy Model (Xception)
Latency Model (MobileNetV2)



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



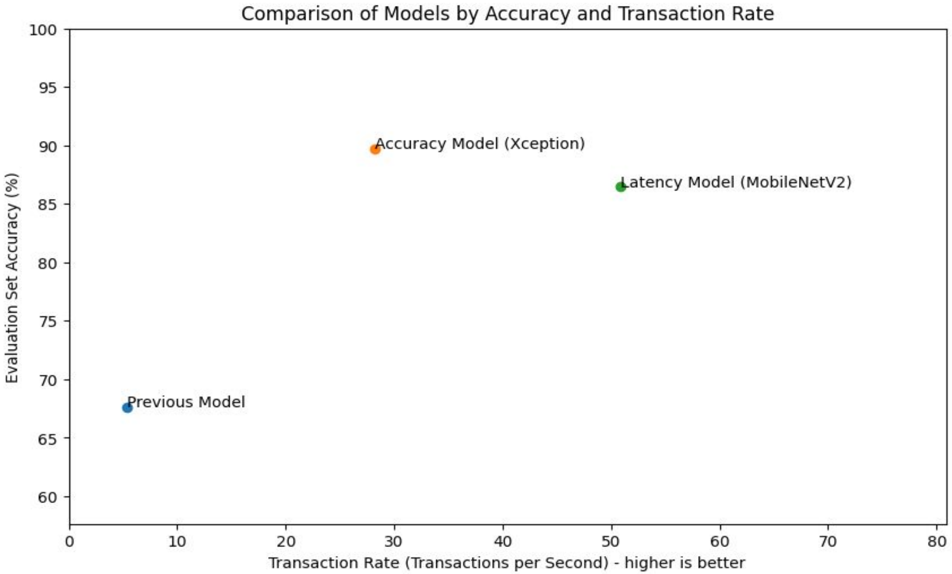
Comparison of Models by Accuracy and Response Time

Legend

Previous Model

Accuracy Model (Xception)

Latency Model (MobileNetV2)



Comparison of Models by Accuracy and Transaction Rate

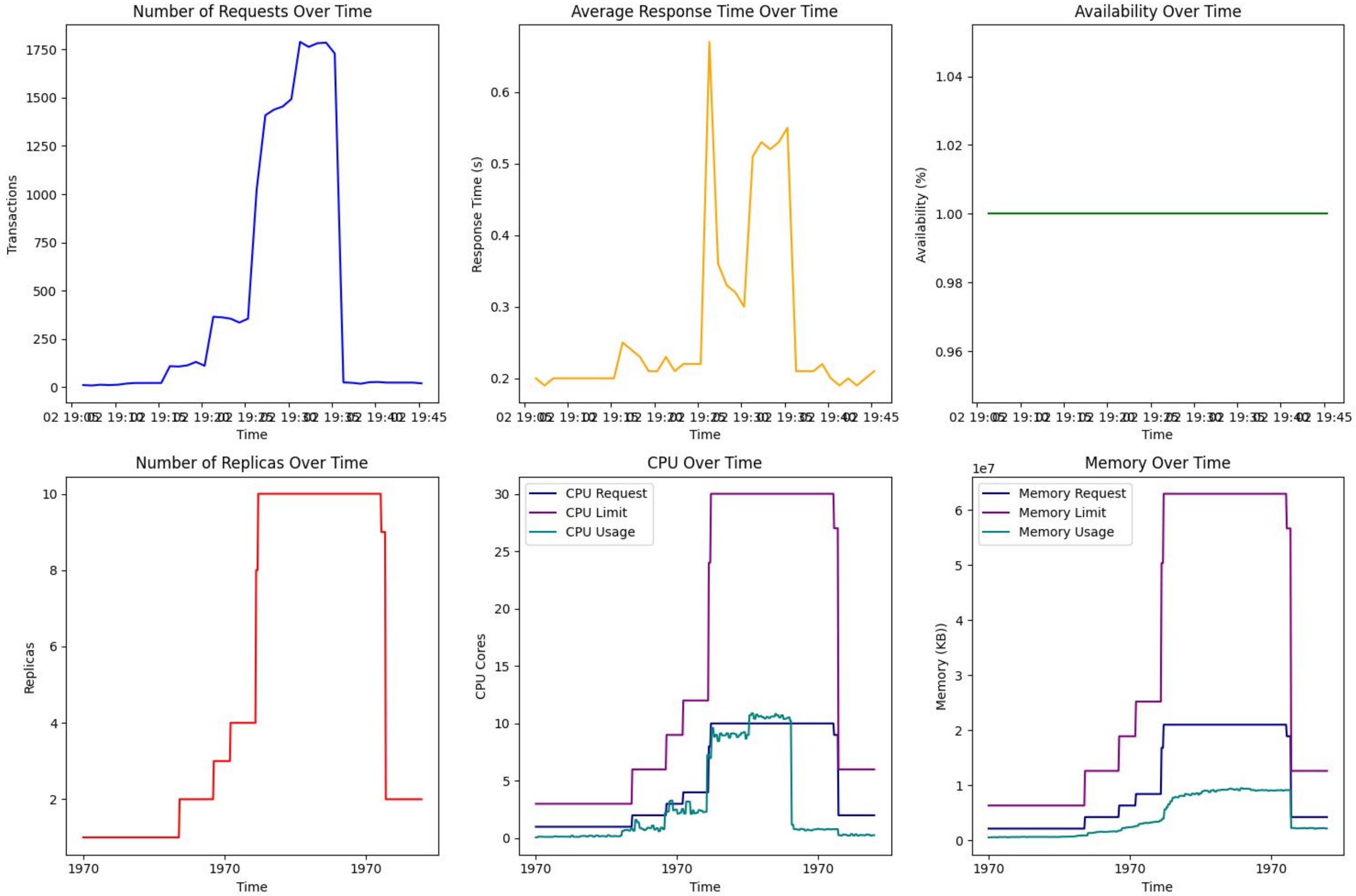
Table showing replicas and resource request configurations for "max-size" deployment

Criteria	Previous	Accuracy (Xception)	Latency (MobileNet)
Number of replicas	20	12	9
CPU resource requests	0.1	0.2	1
Memory resource requests	0.5Gi	1Gi	15 Mi
CPU resource limits	2	2	2
Memory resource limits	4Gi	4Gi	4Gi

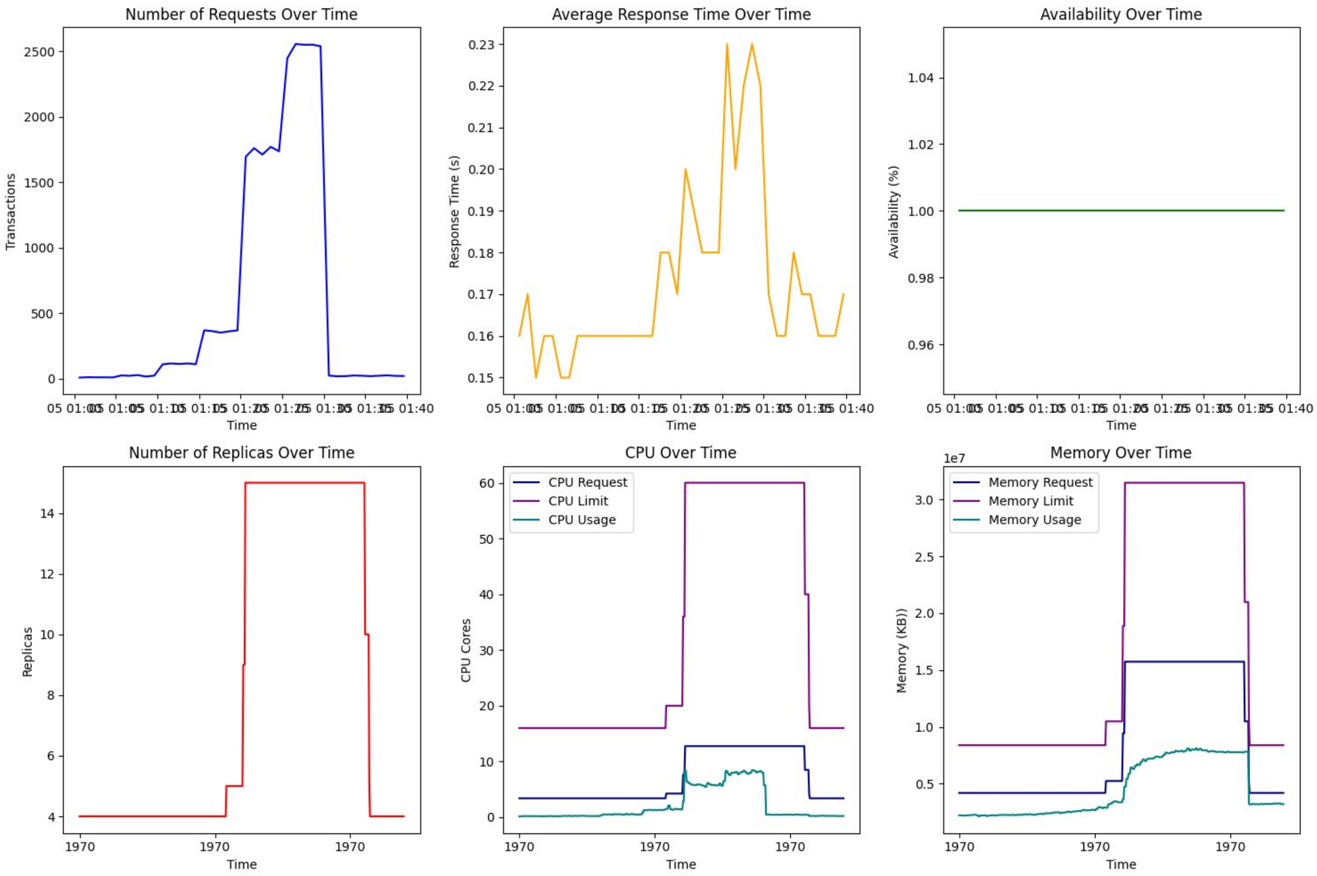
Table showing horizontal scaling configurations

	Accuracy (Xception)	Latency (MobileNetV2)
minReplicas	3	4
maxReplicas	10	15
targetCPUUtilizationPercentage	75%	40%
CPU resource requests	3.2 cores	0.85
Memory resource requests	6Gi	1Gi
CPU resource limits	4 cores	4 cores
Memory resource limits	8Gi	2Gi

Visualization of deployment for "accuracy" model over time



Visualization of deployment for "latency" model over time



Summarize your contributions

All in all,

- The **previous model** has low accuracy (67.64%) and high inference time (1.28 seconds) . When deployed, It has high response time, low availability, low transaction rate. **There is much scope for improvement.**
- We implement **Xception** Model to focus on accuracy. Which has good accuracy & good enough model size. We implement horizontal scaling and set configurations such that the system is **highly available and scales during high requests**. We see that it's accuracy is **89.75% and inference time is (1.19 seconds)**.
- We implement **MobileNetV2** Model to focus on latency. Which has better accuracy than previous model & very less inference time per step. We implement horizontal scaling and **allocate more resources such it reduces inference time** and scales during high requests. We see that it's accuracy is **86.74% and inference time is (0.05 seconds).**