Architecture Overview – Networking – ML model deployment

1. Kubeflow instructs Kserve to deploy a model
2. Kserve queries the model from Minio storage
3. Kserve updates/"serves" the model in all Kserver inference Pods

Service name	Service type	Note
Vehicle API	NodePort	externalTrafficPolicy: Local Externally Accessible through Anycast/NodeIP:nodePort
Kserve Inference	ClusterIP	internalTrafficPolicy: Local
Kepler	ClusterIP	
Prometheus	ClusterIP	
Kserve	ClusterIP	
Kubeflow	ClusterIP	GUI accessible through Ingress
Minio Storage	ClusterIP	
Grafana	ClusterIP	GUI accessible through Ingress