Cloud Computing

Nicola Santillo

23/01/2024

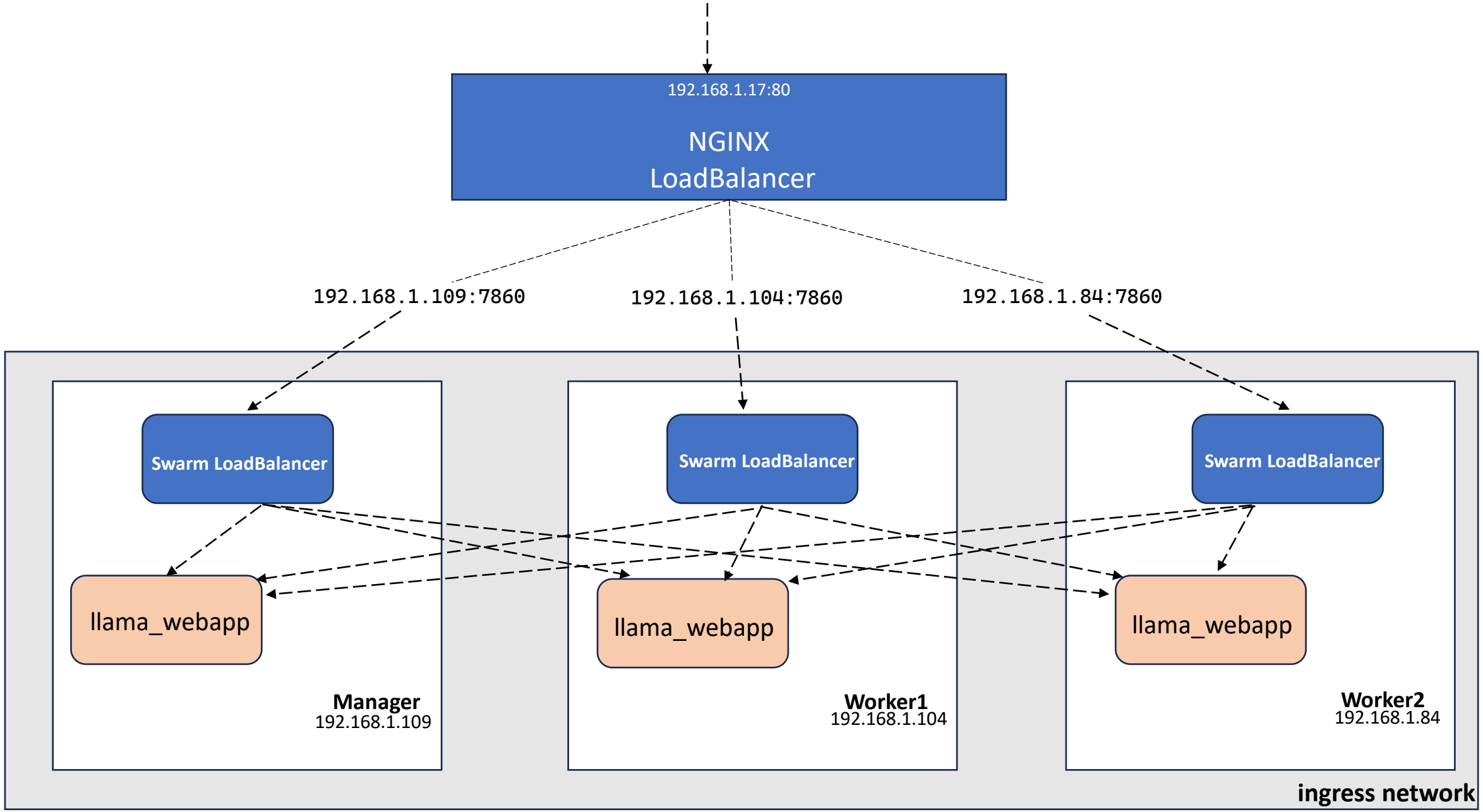
Llama-Webapp

Webapp scritta in python che gestisce la formattazione dei messaggi di input e output, utilizzando le librerie:

- llama_cpp, che utilizza modelli LLaMA via CPU
- gradio, usata come interfaccia per chatbot

Il modello opensource utilizzato è llama-2–7b-chat.Q2_K.gguf che è la versione più compressa dei modelli di chat da 7B parametri.

Docker engine in swarm mode using NGINX loadbalancer



Kubernetes deployment using Kind