# WELCOME!

## MODERN KUBERNETES: DEPLOYING CLOUD-NATIVE AI APP

**MMNOG WORKSHOP 2026**

- **PRESENTER:** KAUNG MYAT SOE
- **GOAL:** FROM ZERO TO A RUNNING AI CHAT APP IN 1.5 HOURS.
- **PLATFORM:** AGB CLOUD (AGBC.CLOUD)

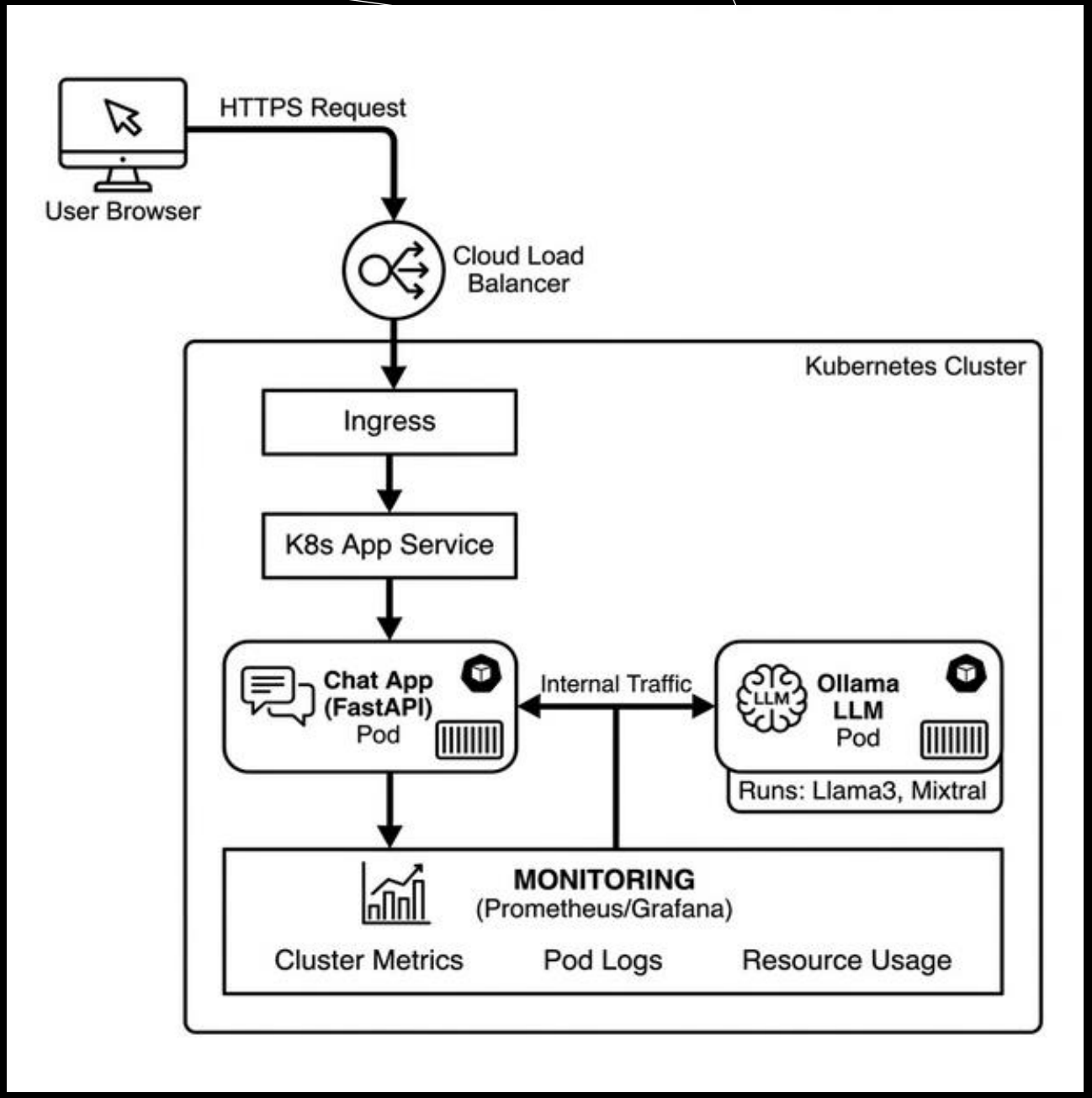
# AGENDA

- Why AI on Kubernetes?
- The Architecture
- LAB AI Model
- Workshop Roadmap
- Networking on AGB Cloud
- Monitoring with Prometheus

  Scaling with HPA

# WHY AI ON KUBERNETES?

- **Scalability:** Auto-scale models as demand grows.
- **Portability:** Run the same stack on any K8s cluster.
- **Resource Management:** Efficiently share CPUs/GPUs.
- **Self-Healing:** Kubernetes restarts models if they crash.




# THE ARCHITECTURE

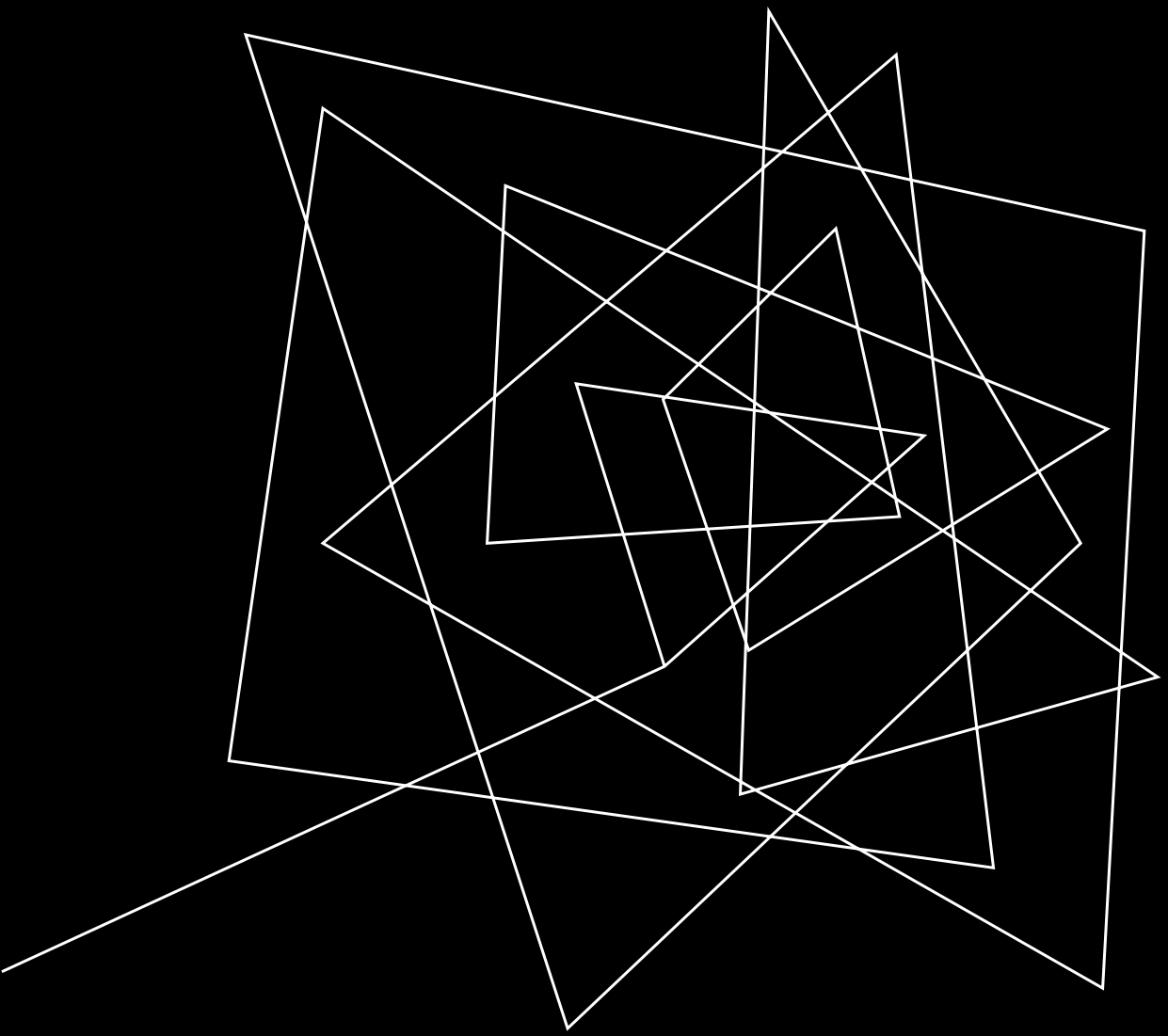
# LAB AI MODEL

**MODEL:** GEMMA3:1B

**SIZE:** ~815MB

**SPEED:** OPTIMIZED FOR CPU-ONLY INFERENCE.

**CAPABILITY:** GENERAL-PURPOSE CHAT, SUMMARIZATION, AND CODING ASSISTANT.



# WORKSHOP ROADMAP

- **LAB 00:** TOOL CHECK (KUBECTL, DOCKER)
- **LAB 01:** CONNECT TO **AGB CLOUD**
- **LAB 02:** RUN **OLLAMA** & DOWNLOAD LLM
- **LAB 03:** DEPLOY THE CHAT **UI**
- **LAB 04: AUTO-SCALE** UNDER LOAD
- **LAB 05: MONITOR** PERFORMANCE (PRE-INSTALLED!)

# NETWORKING ON AGB CLOUD

- **Public IP:** Access your cluster via Cluster Public IP.
- **NodePorts:** We use fixed ports to route traffic:
  - **30706:** Chat Application (mapped to port 8000).
  - **31856:** Grafana Dashboard (mapped to port 3000).
- **Port Forwarding:** Use the AGB Cloud Panel to link your Public IP to these internal ports.

# MONITORING WITH PROMETHEUS

- **Auto-Deployed:** Our setup.sh installs the full stack for you.
- **Prometheus:** Scrapes metrics from our pods every 30s.
- **Grafana:** Visualizes CPU, Memory, and Network traffic.
- **Why?** Essential for debugging performance bottlenecks and seeing scaling in action.

# SCALING WITH HPA

- **The HPA (Horizontal Pod Autoscaler)**:
  - Watches CPU usage.
  - If CPU > 60%, it spins up more replicas (Max: 8).
  - Once load drops, it scales back down (Min: 2).
- **In Action:** We'll use hey to stress-test our chat app!

# READY? LET'S GO!


**Ready? Let's go!**

- **Repo:** https://github.com/kaungmyatsoe/mmnogworkshop.git
- **Facilitators:** We are here to help!
- **First Step:** Open labs/lab-00-prerequisites.md

# THANK YOU

Kaung Myat Soe

+959420150304

kaungmyatsoe@agbcommunication.com

www.agbcommunication.com