

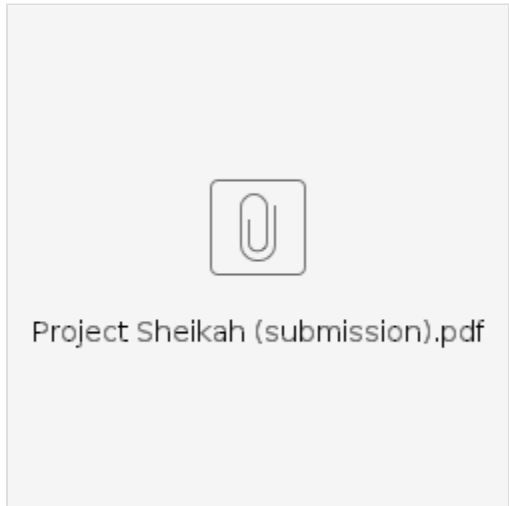
Edge AI Virtual Agents

Project Technical Lead:

Project Committers detail:

Committer	Committer Company	Committer Contact Info	Committer Bio	Committer Picture	Self Nominate for PTL (Y/N)
Qi Tang	Individual developer	qitang2023@gmail.com	Ph.D., Hardware Engineer at Google, Ex-Meta, Connectivity Research, ETSI MEC Delegate	blocked URL	Y
Wei Wu		willwuwork@gmail.com	LLM researcher at Amazon		
Yi Han	Individual developer	yihan7206@gmail.com	Developer, Finance Manager at Achieve, M.Sc University of Southern California; B.S Peking University		
Sharu Jiang	Individual developer	jsrshark110@gmail.com	Software Engineer at Google, Tech LeadTech lead at Google		

Presentation



a video presentation: <https://www.youtube.com/watch?v=r2GfqvbA0hk&t=360s>

Use Case Details

Attributes	Description	Informational
Type	New Blueprint for AI Services at the Edge	
Blueprint Family - Proposed Name	AI Edge Blueprint Family	
Use Case	Running Real-Time Generative AI Models at the Edge	
Blueprint proposed Name	Serverless	
Initial POD Cost (capex)	Leverage Unicycle POD - less than \$150k	
Scale & Type	Up to 7 servers x86/ARM server or deep edge class GPU Accelerators Memory Cache	
Applications	Edge Servers Hosting Real-Time Generative AI Services, e.g. Speech-to-Speech Translation, LLM-based Virtual Assistant Services, Text-to-Image/Video Content Generation Services, etc.	
Power Restrictions	Less than 10Kw	

Infrastructure orchestration	Kubeless Docker 1.13.1 or above and K8s 1.10.2 or above- Container Orchestration OS - Ubuntu 16.x Under Cloud Orchestration - Airship v1.0	
SDN	OVS	
Workload Type	Containers	
Additional Details	Video Streaming: gRPC, Pytorch, Tensorflow Non-Sql Database: MongoDB	