KubeCon | CloudNativeCon Europe 2025

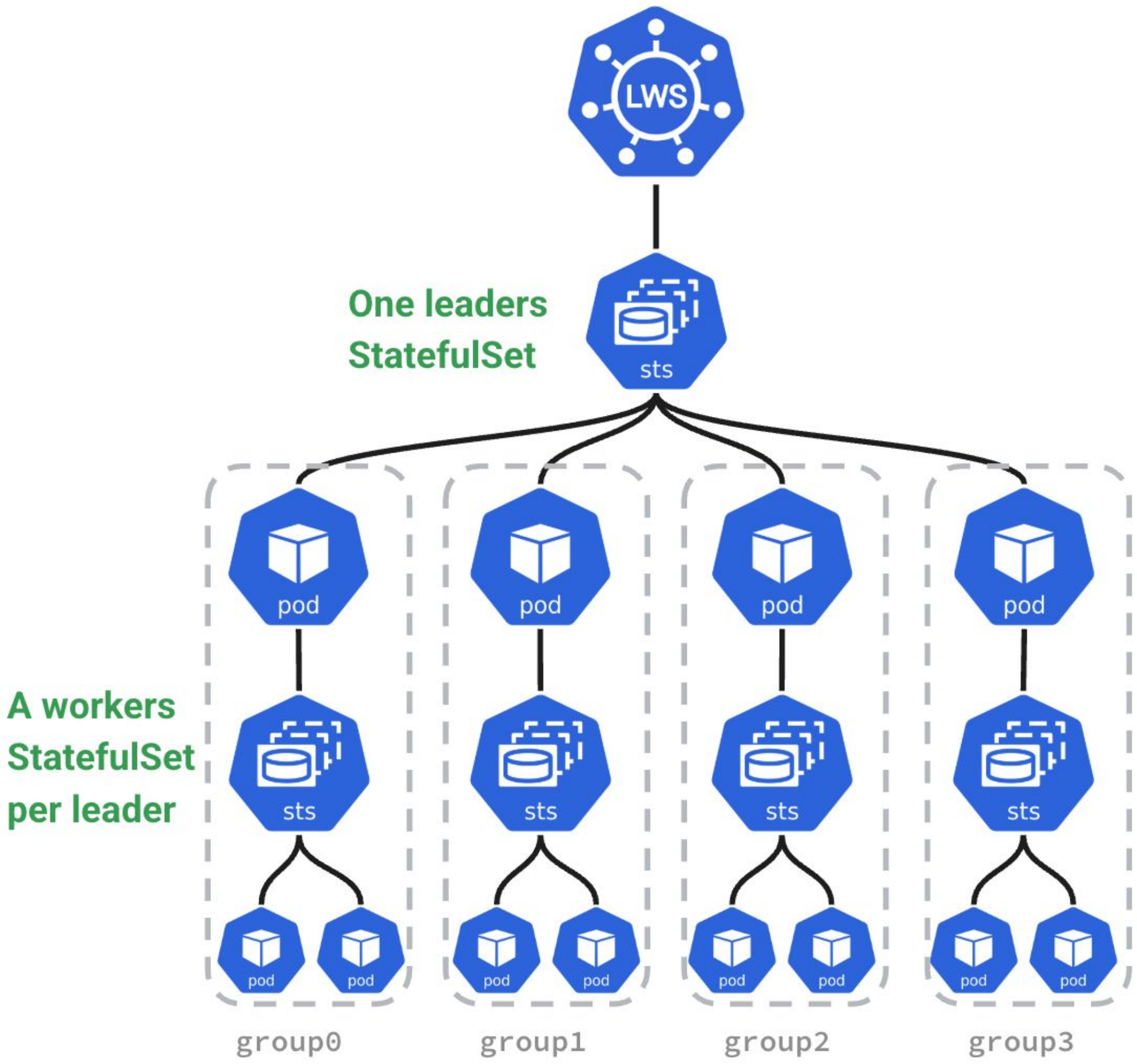
Sailing Multi-Host Inference with LWS

– Kante Yin

lws maintainer, farmer of InftyAI

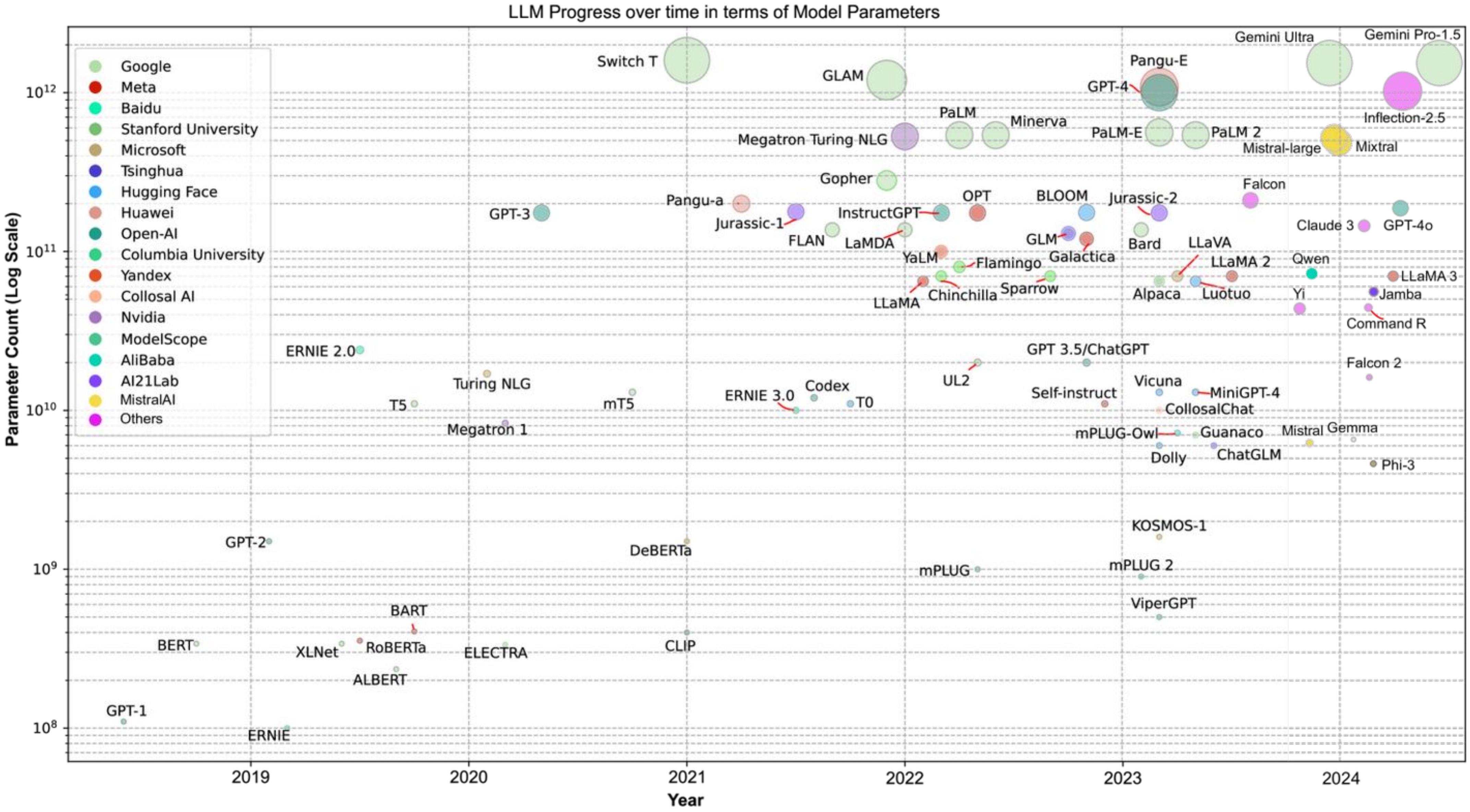
Concept

- Superpod as an unit
- Dual-template, one for leader and one for the workers
- A scale subresource
- Rollout and Rolling update
- Topology-aware placement
- All-or-nothing restart for failure handling



Why LWS

KubeCon CloudNativeCon Europe 2025



Adopters

- **Companies based on public documentations (Join us if you use LWS as well)**:
  - AWS
  - DaoCloud
  - Google Cloud
  - Nvidia
  - …
- **Project Integrations**:
  - **llmaz**: llmaz, serving as an easy to use and advanced inference platform, uses LeaderWorkerSet as the underlying workload to support both single-host and multi-host inference scenarios.
  - **sglang**: sglang, a fast serving framework for large language models and vision language models. It can be deployed with LWS on Kubernetes for distributed model serving, see documentation here
  - **vLLM**: vLLM is a fast and easy-to-use library for LLM inference, it can be deployed with LWS on Kubernetes for distributed model serving, see documentation here.

Updates

- **We just released the new v0.6.0 🎉, besides features**
  - we have a new website: https://lws.sigs.k8s.io/
  - we got 9 new contributors
- **Roadmap**
  - Disaggregated serving
  - Rolling update in place
  - Gang scheduling support
  - … we need your feedbacks
- **Join us if you like:**
  - Github: https://github.com/kubernetes-sigs/lws
  - Slack: under guidance of wg-serving