

Kthena Community Meeting Notes / Agenda

This meeting is for Volcano/Kthena design reviews and technical discussions and is open to Volcano/Kthena community team members. This document captures the agenda and notes from each meeting. If there is anything you want to discuss at this meeting, please add an item in the agenda.

When suggesting topics and agenda items for an upcoming meeting please make sure to include your name in the suggestion.

Schedule

Kthena Community Meeting runs weekly.

Meeting time:

- Community biweekly meeting for Asia: 16:00 - 17:00 (UTC+8) .Wednesday ([Convert to your timezone.](#))

Meeting link: [Kthena Community Meeting](#)

Kthena Community Links

- Github: <https://github.com/volcano-sh/kthena>
- Website: <https://kthena.volcano.sh/>
- Mailing List: <https://groups.google.com/g/volcano-sh>
- Slack: [Volcano Slack Channel](#) | [Join](#)
- Meeting Calendar: [Kthena Meeting calendar](#)

Sept 23, 2026 [Asia]:

1. Same-user FIFO in the fairness queue — bolubo
 - a. (<https://github.com/volcano-sh/kthena/pull/1791>)
 - b. User-Level FIFO
2. Behavior of the partition in ModelServing — lizhENCHENG9527
 - a. (<https://github.com/volcano-sh/kthena/issues/1797>)
 - b. Determines that the partition protects resources whose ordinal number is smaller than the partition.
3. KCD Hangzhou 2026 — lizhENCHENG9527
 - a. ([KCD Hangzhou 2026: Call for Speakers @ Sessionize.com](#))
 - b. CFP DDL: 30th Sept
4. Bootstrap accelerate with ModelExpress - hzxuzhonghu

- a. Use in combination in capacity expansion, failure recovery, and batch creation scenarios

Sept 16, 2026 [Asia]:

1. SessionSticky—FAUST-BENCHOU(<https://github.com/volcano-sh/kthena/pull/1144>)
2. Discuss In-PlaceRollingUpdate – kube-gopher([Proposal: Support In-Place Rolling Update by kube-gopher · Pull Request #1768 · volcano-sh/kthena](https://github.com/volcano-sh/kthena/pull/1768))
3. Share investigation of llm-d vs vllm (LFX Mentorship: Build a KVCache-aware scheduler E2E test suite) - Avinash Kumar (<https://github.com/volcano-sh/kthena/issues/1328>)
- 4.

Sept 9, 2026 [Asia]:

This week's Kthena community meeting is suspended.

Sept 2, 2026 [Asia]:

1. Router benchmark demo and final report — StLeoX
 - a. [A router-benchmark implement \(both tier1 and tier2\) by StLeoX · Pull Request #1285 · volcano-sh/kthena](https://github.com/volcano-sh/kthena/pull/1285)
2. Kthena Release v1.1 — LiZhenCheng9527
 - a. Highline Feature
 - i. support MaxSurge [done]
 - ii. support MaxSkew in RoleRollingUpdate
 - iii. Roll Back when rolling update failed
 - iv. Router support third party model
 - v. modelserving eviction protection
 - vi. Performance Improvements in the KV-Cache-Aware Plugin
 - vii. Announce the remove of the ModelBooster alert.
 - viii. Router support RL
 - b. Timeline: Sept 30
3. v1/response protocol
 - a. [kthena/docs/proposal/infer-gateway-session-sticky.md at main · volcano-sh/kthena](https://github.com/volcano-sh/kthena/blob/main/docs/proposal/infer-gateway-session-sticky.md)
 - b. Plan to support
- 4.

Aug 26, 2026 [Asia]:

1. Remove ModelBooster API: [\[Discuss\] Remove ModelBooster API · Issue #1679 · volcano-sh/kthena](#) - kube-gopher
2. proposal: coordinated proportional Role rolling updates : <https://github.com/volcano-sh/kthena/pull/1688> - haoeetee

Aug 19, 2026 [Asia]:

1. ModelServing Support rollback when rolling update is failed.
—acsoto
 - a. [docs: propose ModelServing revision history and rollback by acsoto · Pull Request #1607 · volcano-sh/kthena](#)
 - b. [feat: add ModelServing revision core by acsoto · Pull Request #1643 · volcano-sh/kthena](#)
2. Modify the Kubebuilder configuration for `spec.Replicas`, `role.Replicas`, and `workerReplicas`
3. Review on Local Tokenizer Proposal - ntxcybernet - <https://github.com/volcano-sh/kthena/pull/1265>

Aug 12, 2026 [Asia]:

1. Ratio Rounding Violates ratioConstraint
 - a. [Disaggregated autoscaler can patch a ratio outside ratioConstraint after integer rounding by miantalha45](#)

Aug 05, 2026 [Asia]:

1. Support PodGroup and subGroup affinity

- a. [docs: add ModelServing topology affinity proposal by VanderChen · Pull Request #1478 · volcano-sh/kthena](#)
2. Should RoleRollingUpdate support ServingGroup partition and ServingGroup maxUnavailable?
 - a. [Support maxUnavailable in all type Rolling Strategy by LiZhenCheng9527 · Pull Request #1429 · volcano-sh/kthena](#)
 - b. Extract the Rollout Configuration
 - c. open a issue to discuss two rollout strategy.
3. Modifying the pod name generation rule.
 - a. add a trigger about create headless service
 - b. add a good first issue

July 29, 2026 [Asia]:

- 1.[Feature Request] Coordinate Prefill/Decode rolling-update progress with maxSkew
<https://github.com/volcano-sh/kthena/issues/1406> FAUST-BENCHOU
2. [Feature Request] support role-scoped network topology policies. JIAHUAT
<https://github.com/volcano-sh/kthena/issues/1278>
<https://github.com/volcano-sh/kthena/pull/1301>
<https://github.com/volcano-sh/kthena/pull/1325>
apply a proposal in next conference.
3. Define InPlaceRollingUpdate for ModelServing: eligible image changes, update scope, readiness checks, and failure recovery. [Talha (github.com/miantalha45)]
4. Modifying the pod name generation rule.
to create a issue discuss about it.