Shared with dev@kubernetes.io for commenting
Shared with wg-serving@kubernetes.io for editing

Our community meeting is weekly on Thursdays, alternating between 9:00 and 11:30 PDT (Zoom Link).
Meeting link: https://zoom.us/j/9955436256?pwd=Z2FQWU1jeDZkVC9RRTN4TlZyZTBHZz09
Recording link: https://www.youtube.com/playlist?list=PL69nYSiGNLP30qNanabU75ayPK7OPNAAS

Topics for next meeting

-

Apr 23, 2026 | WG-Serving: Benchmarking Tool Contributors Meeting

Attendees: Ashok Chandrasekar Jason Kramberger Brendan Slabe

Notes

- Open topics:
  - Otel tracing
    - Uploading to huggingface
    - Adding support to use trace via HF datasets in inference-perf
    - Proposal : Scheduling for agentic applications in llm-d
    - [Ashok] What is the source of traces?
      - https://www.exgentic.ai/
      - Sources from different agents
    - https://www.appliedcompute.com/research/inference-benchmark?utm_source=tldrai
- Issue triage
- Milestone review
- Open PRs
- Recently merged PRs

Action items

- [ ] 0.5.0 release with some bug fixes discussed

Apr 9, 2026 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees: Jason Kramberger, Brendan Slabe, Ashok Chandrasekar

Notes

- Open topics:
  - Propose adding v0.5.0, v0.4.0 was Feb 6
    - Needs fix for 0.5.0
      - https://github.com/kubernetes-sigs/inference-perf/pull/383
      - https://github.com/kubernetes-sigs/inference-perf/pull/410
      - https://github.com/kubernetes-sigs/inference-perf/pull/433
  - Contribution model, in coming months for 1.0.0
    - Proposal:
      - Before PR review and merge, issue must have needs-triage label removed and a priority/* label applied
      - Prototype and [WIP] PRs are still welcomed
    - [Sachin] If a new member joins how do they pickup a ticket
      - Needs-help label (good-first-issue), unassigned priority/* labels
      - Triage should outline expectations for next step, (should we add a label for needs-design?)
      - We need to update the Contribution guide to reflect these
    - [Brendan] Separate path for bugs vs feature requests?
      - Separate labels for bugs? Or do priority/* work well enough
  - Trace hosting / workload catalog
    - https://github.com/kubernetes-sigs/inference-perf/pull/432
    - https://huggingface.co/changelog/agent-trace-viewer
    - Otel traces will be uploaded to huggingface
      - Looking into other trace repositories
      - There are different repositories and formats, no clear winner yet
      - Inference-perf slack is likely a good place to collaborate and sync on trace hosting / generation
      - Some open questions on how otel support will be in HF wrt to Dataset cards
  - Otel updates
    - https://github.com/kubernetes-sigs/inference-perf/pull/434
    - https://github.com/kubernetes-sigs/inference-perf/pull/436
  - Distributed loadgen demo / PR walkthrough
    - https://github.com/kubernetes-sigs/inference-perf/issues/437
    - https://github.com/kubernetes-sigs/inference-perf/pull/438
- Milestone review
- Open PRs
- Recently merged PRs

Action items

- Jason Kramberger open milestone 0.5.0 with the linked issues/prs

Apr 9, 2026 |

WG-Serving: Benchmarking Tool Contributors Meeting

Attendees: Jason Kramberger Brendan Slabe Ashok Chandrasekar

Notes

- Open topics:
  - v1.0.0 delayed
  - Trace hosting
    - Where to host traces and how to display trace metadata like a model card
    - Some collaboration already with huggingface
    - https://github.com/kubernetes-sigs/inference-perf/pull/422#issuecomment-4215264733
    - https://github.com/kubernetes-sigs/inference-perf/pull/369
  -
- Milestone review - https://github.com/kubernetes-sigs/inference-perf/milestone/4
- Open PRs
- Recently merged PRs
  - https://github.com/kubernetes-sigs/inference-perf/pull/372
    - Presentation in llm-d sig-benchmarking?
    - Recorded demo to contribute

Action items

- [ ]

Apr 2, 2026 |

WG-Serving: Benchmarking Tool Contributors Meeting

Attendees: Jason Kramberger

Notes

- Open topics:
  - v1.0.0 indefinitely delayed
    - Want feature additions
      - Trace replays

                - Multi-modality datasets
                - Saturation detection
                - UI/UX improvements
                    - https://github.com/kubernetes-sigs/inference-perf/pull/377
                    - https://github.com/kubernetes-sigs/inference-perf/pull/384
            - Test and validation plan
                - Documentation audit
- Milestone review - https://github.com/kubernetes-sigs/inference-perf/milestone/4
- Open PRs
- Recently merged PRs

Action items

- [ ] Slack ask for contributors for multi modality support
- [ ] Remove blank issue and update issue template

Mar 26, 2026 11:30 AM PDT |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees: Ashok Chandrasekar Sachin MV

Notes

- Open topics:
    - Trace replay next steps
        - Otel trace replay for agentic benchmarks - https://github.com/kubernetes-sigs/inference-perf/pull/372
        - Shared prefix tree-of-thought - https://github.com/kubernetes-sigs/inference-perf/pull/369
        - Alibaba qwen trace replay (code generation, multi-turn user chat and API tool calling) - https://github.com/llm-d/llm-d-benchmark/blob/main/experimental/multi-turn/production-trace-replay-qwen.py
- Milestone review - https://github.com/kubernetes-sigs/inference-perf/milestone/4
- Open PRs
- Recently merged PRs

Action items

- [ ]

Mar 19, 2026 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees: Ashok Chandrasekar Brendan Slabe

Notes

- Milestone review - https://github.com/kubernetes-sigs/inference-perf/milestone/4
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
- Open PRs / recently merged PRs

Action items

- [ ]

Mar 12, 2026 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Milestone review - https://github.com/kubernetes-sigs/inference-perf/milestone/4
    - https://github.com/kubernetes-sigs/inference-perf/pull/362#issuecomment-4048328781
        - https://github.com/kubernetes/test-infra/pull/36123 merged, unblocked on k8s.io inference-perf image and helm chart (https://github.com/kubernetes-sigs/inference-perf/issues/218)
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
- Open PRs / recently merged PRs
- Open topics

Notes

-

Action items

- [ ]

Mar 12, 2026 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees: Jason Kramberger

Notes

- Milestone review - https://github.com/kubernetes-sigs/inference-perf/milestone/4
    - https://github.com/kubernetes-sigs/inference-perf/pull/362#issuecomment-4048328781
    - 0.4.1 patch release for recent bug fixes for random seeding
        - Evaluate cherry picking
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
- Open PRs / recently merged PRs
- Open topics
    -

Action items

- [ ] Open issues for other trace replay requests that we are aware of

---

Mar 5, 2026 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees: Jason Kramberger Brendan Slabe

Notes

- Milestone review - https://github.com/kubernetes-sigs/inference-perf/milestone/4
    - [Jason] Working on test coverage report and final new timeline for v1.0.0
    - [Jason] Helm chart & k8s image repository has some traction, hopefully unblocked soon
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
- Open PRs / recently merged PRs
    - https://github.com/kubernetes-sigs/inference-perf/pull/357

    - https://github.com/kubernetes-sigs/inference-perf/pull/358
- Open topics
    -

Action items

Feb 26, 2026 | WG-Serving: Benchmarking Tool Contributors Meeting

Attendees: Ashok Chandrasekar Jason Kramberger Sachin MV

Notes
- Milestone review - https://github.com/kubernetes-sigs/inference-perf/milestone/4
    - [Ashok] Add a “no-config” option to run from the CLI
    - [Jason] Re: Hardware metrics from prometheus https://github.com/llm-d/llm-d-benchmark/issues/597
    - [Jason] Should we move the milestone date?
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
- Open PRs / recently merged PRs
    -
- Open topics
    - Linting tasks don’t block PRs
        - Prow configuration for auto merge

Action items
- [ ] UX tract items for v1.0.0 (helm, image, no-config) Jason Kramberger
- [ ] Final goal for testing for v1.0.0 Jason Kramberger

Feb 19, 2026 | WG-Serving: Benchmarking Tool Contributors Meeting

Attendees: Sachin MV

Notes
- Milestone review - https://github.com/kubernetes-sigs/inference-perf/milestone/4
    - K8s image and helm chart

- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
- Open PRs / recently merged PRs
- Open topics

Action items

- [ ] Check guest access to meeting note doc

Feb 12, 2026 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees: ~~Jason Kramberger~~

Notes

- Milestone review - https://github.com/kubernetes-sigs/inference-perf/milestone/4
  - K8s image and helm chart
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
- Open PRs / recently merged PRs
- Open topics
  - https://github.com/kubernetes-sigs/inference-perf/pull/320
    - Request_queue.put needs workid specified instead of preferred_worker_id, issue to be added
  - New meeting members
    - Chang Min Bark, meta production eng, interested in ML infra and k8s
  -

Action items

- [ ] Check guest access to meeting note doc

Feb 5, 2026 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees: Jason Kramberger Sachin MV Brendan Slabe

Notes

- Milestone review - https://github.com/kubernetes-sigs/inference-perf/milestone/4
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
- Open PRs / recently merged PRs
- Open topics
    - v0.4.0
    - v1.0.0 Roadmap https://github.com/kubernetes-sigs/inference-perf/issues/321

Action items
- Jason to create v0.4.0 roadmap
- Sachin cut a release on 02/06

Jan 29, 2026 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees: Jason Kramberger Sachin MV

Notes
- Milestone review - https://github.com/kubernetes-sigs/inference-perf/milestone/4
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
- Open PRs / recently merged PRs
- Open topics
    - V1.0.0 Roadmap https://github.com/kubernetes-sigs/inference-perf/issues/321
    - Meeting calendar invite link

Action items
- ☐

Jan 22, 2026 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees: Jason Kramberger Brendan Slabe Sachin MV

Notes

- Milestone review - https://github.com/kubernetes-sigs/inference-perf/milestone/4
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
- Open PRs / recently merged PRs
- Open topics
    - V1.0.0 Roadmap https://github.com/kubernetes-sigs/inference-perf/issues/321

Action items

- [ ]

Jan 15, 2026 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees: Jason Kramberger Brendan Slabe Sachin MV

Notes

- Milestone review
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
- Open PRs / recently merged PRs
- Open topics
    - https://github.com/kubernetes-sigs/inference-perf/issues/262
    - Look through assigned issues that may be stale and assign appropriately.

Action items

- [ ]

Dec 18, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees: ~~Ashok Chandrasekar~~

- Jason Kramberger

- Brendan Slabe
- Sachin MV

Notes

- Milestone review 1.0 - https://github.com/kubernetes-sigs/inference-perf/issues/243
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
- Open PRs / recently merged PRs
- Open topics
    - WG-Serving disband, Serving-catalog merge

Dec 11, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees: ~~Ashok Chandrasekar~~

- Jason Kramberger
- Brendan Slabe
- Sachin MV

Notes

- Milestone review 1.0 - https://github.com/kubernetes-sigs/inference-perf/issues/243
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
- Open PRs / recently merged PRs
- Open topics
    - WG-Serving disband, Serving-catalog merge

Action items

- [ ]

Nov 20, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Brendan Slabe
- Sachin MV
- Jason Kramberger

Notes

- Milestone review 1.0 - https://github.com/kubernetes-sigs/inference-perf/issues/243
    -
- New issue triage -
- Open PRs / recently merged PRs
- Open topics
    - Llm-d-benchmark ask for high concurrency / multi client benchmark
    - Helm chart https://github.com/kubernetes/k8s.io/pull/8776 https://github.com/kubernetes/k8s.io/pull/8777

Action items

Oct 30, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Brendan Slabe
- Sachin MV

Notes

- Milestone review 1.0 - https://github.com/kubernetes-sigs/inference-perf/issues/243
    - Trace / replay - waiting for Aishwarya
    - Multi-turn chat - Xia is looking into this starting with shared prefix dataset
    - CI/CD
    - Concurrency support
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
    - https://github.com/kubernetes-sigs/inference-perf/issues/258
        - Publish initial helm chart repository
- Open PRs / recently merged PRs
- Open topics

Action items

- [ ]

Oct 23, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Jason Kramberger
- Brendan Slabe

Notes

- Milestone review 1.0 - https://github.com/kubernetes-sigs/inference-perf/issues/243
    - Trace / replay - waiting for Aishwarya
    - Multi-turn chat - Xia is looking into this starting with shared prefix dataset
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
    - https://github.com/kubernetes-sigs/inference-perf/issues/258
        - Publish initial helm chart repository
- Open PRs / recently merged PRs
- Open topics
    - Kubecon slides and results for review hopefully next week

Action items

- [ ]

Oct 16, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Jason Kramberger
- Brendan Slabe

Notes

- Milestone review 1.0 - https://github.com/kubernetes-sigs/inference-perf/issues/243
    - Trace / replay - waiting for Aishwarya
    - Multi-turn chat - Xia is looking into this starting with shared prefix dataset

- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
- Open PRs / recently merged PRs
- Open topics
    - What kind of load we will use to compare inference optimizations?
        - Generic hardware with model optimizations - quantization, speculative decoding, etc.
        - Saturation detection and full curve is the way to go

Action items

- [ ]

Oct 9, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

Notes

- Cancelled

Action items

- [ ]

Oct 2, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Jason Kramberger
-

Notes

- Roadmap - https://github.com/kubernetes-sigs/inference-perf/issues/243

- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
- Open PRs / recently merged PRs
- Open topics

Action items

- [ ]

Sep 25, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Brendan Slabe
- Jason Kramberger
- Sachin MV

Notes

- v0.2 release is out!
- v1.0 Planning (Nov 3 is the deadline, Nov 11 is Kubecon, Blog on the week of Kubecon)
    - P0 - Trace replay - replay both inference-perf trace and other trace datasets - llm-d
    - P0 - Multi-turn chat support
        - Idea is for datagen and loadgen to make sure the same worker is able to send all the requests in the conversation as a single user
    - P1 - Support different input / output distribution for different stages - helps with autoscaling (related to trace replay too since this might need new abstraction)
    - P1 - Multi-model or LoRA support and traffic splitting - inference gateway
        - Model is already part of completion API, header would define the lora adapter
    - P1 - Multi-modal support - can we benchmark vision language models
    - P2- SLO support and conformance of specific latency SLOs
        - More on the analysis side and we can show additional ones like goodput
    - P2 - GPU utilization / other hardware metrics from Prometheus
    - P0 - Testing - ability to automate scale testing with a mock server and test different datasets, loadgen, etc.
- Open items
    - Trace replay open PR - https://github.com/kubernetes-sigs/inference-perf/pull/198/files

        - In general, the current PR is fine
        - We need to run it through the mock server and look at scale

Action items

- [ ]

Sep 18, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Jason Kramberger
- Brendan Slabe

Notes

- v0.2 tracking
    - https://github.com/kubernetes-sigs/inference-perf/milestones
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
- Open PRs / recently merged PRs
- Open topics
    - V1 timeline / scope
        - 0.2 to be released this week
        - V1 milestone pre-kubecon would be ideal
        - Scoping - can we do a meeting to finalize that?
            - Let's add a list and go from there
    - Inference-perf users - OSS communities
        - Inference gateway is using inference-perf, but haven't fully moved to it yet
        - Llm-d - https://github.com/llm-d/llm-d-benchmark
        - Can we update the readme to highlight who our users are?

Action items

- [ ]

Sep 11, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Jason Kramberger
- Brendan Slabe

Notes

- v0.2 tracking
    - https://github.com/kubernetes-sigs/inference-perf/milestones
        - Request rate sweeps
            - Generation type is the only non-default
            - Can we default to linear?
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
- Open PRs / recently merged PRs
    - Added generated report and dataset in the PR description: https://github.com/kubernetes-sigs/inference-perf/pull/198
- Open topics

Action items

- [ ]

Sep 4, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Brendan Slabe
- Jason Kramberger
- Aishwarya Raimule
- Sachin MV

Notes

- v0.2 tracking
    - Tracing
        - Python echo server can be used to load test

        - https://github.com/kubernetes-sigs/inference-perf/issues/40
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
- Open PRs / recently merged PRs
    - https://github.com/kubernetes-sigs/inference-perf/pull/211
        - Dataset abstraction to handle different distributions makes sense, but we need to make sure it is efficient
        - Pre-processing adds overhead which might not work for a lot of smaller runs
        - https://github.com/kubernetes-sigs/inference-perf/pull/204
- Open topics
    - V1.0 before Kubecon
        - V1 feature request - can we provide visual comparison of metrics maybe across reports - Sachin to create an issue

Action items

- [ ]

Aug 28, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Jason Kramberger
- Brendan Slabe
- Aishwarya Raimule

Notes

- v0.2 tracking
    - Trace support - https://github.com/kubernetes-sigs/inference-perf/pull/198
        - Timeout issue with trace replay even on a smaller model
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
- Open PRs / recently merged PRs
- Open topics
    - Azure tracegen results -

Action items

- [ ]

Aug 21, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Jason Kramberger
- Brendan Slabe
- Aishwarya
- Sachin

Notes

- v0.2 tracking
    - Scheduling delay improvement - https://github.com/kubernetes-sigs/inference-perf/pull/157/files
        - Trace replay - would the new multi process implementation work with trace replay as is?
            - Multi-turn chat might be the difficult part to implement
            - Trace data is usually 60 MB+
            - If the entire dataset is going to be in memory anyway, this doesn't make much of a difference
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
- Open PRs / recently merged PRs
- Open topics
    - https://github.com/kubernetes-sigs/inference-perf/pull/193

Action items

- [ ]

Aug 14, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Aishwarya
- Brendan Slabe
- Jason Kramberger
-

Notes

- v0.2 planning
    - Tracing
        - Timestamps, input and output tokens
        - Do we want to allow timestamps only and generate data ourselves?
        - Data and loadgen config
            - New loadgen to follow timestamps from distribution
            - New datagen to follow data from the trace
    - Request rate sweeps
    - Adding a runtime - TGI
        - Ollama was validated and it works because OpenAI API format is supported by most model servers
        - Sachin looked into Triton https://github.com/kubernetes-sigs/inference-perf/issues/18
        - Aishwarya looking into adding TGI
        - Metrics standardization - [External] Standardizing Large Model Server Metrics in Kubernetes
            - For now, it would be good to go with metrics definitions as reported by model servers
    - Multi-turn chat / rounds per user
    - Multi-model - let's check with inference gateway team
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
- Open PRs / recently merged PRs
- Open topics

Action items

- [ ]

Aug 7, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Aishwarya

Notes
There were no items to discuss - canceled.

Jul 31, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:
- Ashok Chandrasekar
- Brendan Slabe
- Jason Kramberger
- Chen Wang
- Aishwarya

Notes
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
- Recently merged PRs
- Open PRs
    - https://github.com/kubernetes-sigs/inference-perf/pull/153 Accounts created and will proceed with testing. To be merged by this week.
    - https://github.com/kubernetes-sigs/inference-perf/pull/42 to be updated next week.
- Open topics
    - KubeCon accepted. How to plan?
        - Congrats to speakers!
        - Tutorial:
            - 1 hr 30 min with 5 speakers
            - What content do we want to cover?
            - Jason Kramberger would like to demo llm-d with inference-perf harness
        - Inference-perf talk - Sachin, Brendan
            - Would be good to share once the presentation is ready

Action items
- [ ]

Jul 24, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Brendan Slabe
- Aishwarya

Notes

- V0.1.1 readiness - release pipeline and pending issues
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
- Open PRs
    - https://github.com/kubernetes-sigs/inference-perf/pull/152
        - Let's go with this for offline and eventually make tokenizer optional
- Open topics
    -

Action items

- [ ]

Jul 17, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Brendan Slabe
- Chen Wang
- Vivek Karunai Kiri Ragavan
- Sachin MV

Notes

- Pypi release https://github.com/kubernetes-sigs/inference-perf/issues/117
- [chenw]Fmperf renaming libraries => deployer/orchestrator?
    - deployer

Action items

-

Jul 10, 2025 | WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Jason Kramberger
- Nick Masluk
- SachinMV
- Aishwarya
- Vivek Karunai Kiri Ragavan

Notes

- https://github.com/kubernetes-sigs/inference-perf/issues/124
    - Added heterogeneous resources to the API
    - Combined specific metrics across machines
    - We do not want to replicate the whole k8s manifest in the API to keep it more readable and simple
        - Can have a metadata field to add more adhoc things
    - We need it to be more flexible
    - Quickstart should be simple to do
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
- Inference Perf: Project Update (recording)
- Can we cut 0.1.1 release?
    - https://github.com/kubernetes-sigs/inference-perf/pull/137

Action items

- [ ]

Jul 3, 2025 | WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Aishwarya Raimule
- Chen Wang

Notes:

- There are no items to discuss, so the meeting is canceled.

Jun 26, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Nick Masluk
- Jason Kramberger
- Sachin MV
- Chen Wang

Notes

- Nick - At IBM Research working on model optimization
- V0.1.0 followup and bugs from testing
    - Load generator is doing well in terms of scale - needs the right machine to generate the specified load
    - We were misreporting the latency because we were getting the end time before the request completed
    - https://github.com/kubernetes-sigs/inference-perf/pull/125 adds QPS observability
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
    - https://github.com/kubernetes-sigs/inference-perf/issues/126 - can we improve config documentation
        - Can we do this automatically and make some config not a requirement?
        - Can we build an interactive CLI which takes information from people and builds the config to run?
    - Benchmarking input / output API - https://github.com/kubernetes-sigs/inference-perf/issues/124
        - Unified format
        - Extends
- [chenw]Release bot image built issue, will send out PR. https://github.com/kubernetes-sigs/inference-perf/issues/119

Action items

- [ ]

Jun 12, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Brendan Slabe
- Sachin MV

Notes

- v0.1.0 followups
    - Release tag - requires permission to be added. Chen to add that.
    -
- New issue triage - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
    - [chenw]
    - Do we launch via helm or library/tool: https://github.com/kubernetes-sigs/inference-perf/issues/113
    - Do we want the helm chart to be deployed by the library?
- Unifying input and output data format
    - Input format config and yaml looks good
    - Output format - json works
        - Can we make this available via the library?
        - Time to second token (for disaggregation)
            - To second decoding token, to second last decoding token latencies.
    - How do we add support other file storage providers?
        - https://github.com/kubernetes-sigs/inference-perf/tree/main/inference_perf/client/filestorage
        - Will need to add other file storageclass (chenw will create an issue).
- Deploy inference-perf to llm-d-benchmark
    - https://github.com/llm-d/llm-d-benchmark/pull/42
    - [chenw] will prioritize the fmperf library merging to inference-perf: https://github.com/kubernetes-sigs/inference-perf/pull/42

Action items

- [ ]

Jun 5, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Brendan Slabe
- Jason Kramberger
- Aishwarya Raimule
- Sachin MV

Notes

- Milestone v0.1.0 review
    - Ran into a blocking issue for release - https://github.com/kubernetes-sigs/inference-perf/pull/112
- Publishing python package in pypi
    - It would be good to publish the v0.1.0 package so it can be easily installed via pip
    - Create an issue for this
- New issue triage
    - https://github.com/kubernetes-sigs/inference-perf/issues?q=is%3Aissue%20state%3Aopen%20label%3Aneeds-triage
        - Create an issue for guidance around chat template with vLLM and possibly other servers
        - Moving to PDM - https://github.com/kubernetes-sigs/inference-perf/issues/111
            - Convenient to manage dependencies in a more fine grained way
            - Can keep makefiles for now and look into moving to PDM
- Helm chart for deploying
    - Can help with helm installing directly on to a cluster
- Pypi package for inference-perf
    - Chen to look into this
- Demo + update in the wg-serving meeting next time

Action items

- [ ]

May 29, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Jason Kramberger
- Brendan Slabe
- Sachin MV
- Aishwarya
- Chen Wang

Notes

- Milestone v0.1.0 review - https://github.com/kubernetes-sigs/inference-perf/milestones
- https://github.com/kubernetes-sigs/inference-perf/pull/99
    - Up for early review, several todos to address
    - What is the QPS we can reach? 1k to 2k QPS
    - How can we ensure QPS is met?
        - Scheduling is done in one place
        - Actual QPS can be reported
- Image available on quay.io/inference-perf/inference-perf:latest
    - https://github.com/kubernetes-sigs/inference-perf/pull/97
- Testing sub tasks
    - https://github.com/kubernetes-sigs/inference-perf/pull/96
    - Can we
- Streaming request support in progress
    - Config updates: https://github.com/kubernetes-sigs/inference-perf/pull/81
- Multi-stage report gen for Prometheus metrics
    - https://github.com/kubernetes-sigs/inference-perf/pull/95

Action items

- [ ]

May 22, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Chen Wang
- Ashok Chandrasekar
- Jason Kramberger
- Aishwarya
- Sachin MV

Notes

- Milestone v0.1.0 review
    - Tentative date: May 30
    - Prefix caching - we can add a simple random generator with prefix match for now
    - Extend in the future to support other loadgens
- Scaling the loadgen Jason Kramberger
    - https://github.com/kubernetes-sigs/inference-perf/issues/86
    - Fmperf's deployer can wrap the deployment of inference-perf config
    - Two main changes:
        - Loadgen needs to spawn multiple process
        - Client refactoring since client publishes to reportgen
            - Move metrics collection from client into loadgen and client should only process the request and return the response
- Fmperf library PR (library renaming and output data code clean up) Chen Wang
    - GuideLLM, LMBench, etc. are merged to fmperf
    - Will be adding them in the next PR
    - Current setup separates loadgen from the inference stack
    - Because loadgens have their own specs, and inference stacks have their own specs, they can vary
- KubeCon Benchmarking Talk submission Chen Wang
    - Tutorial for GenAI benchmarking
        - https://docs.google.com/document/d/1tDscEc6HWGVMVfFQ_IXICr4SIofZ3M7Z3LUgR1qJ-ho/edit?usp=sharing
    - Inference-perf breakout session
        -
- https://github.com/kubernetes-sigs/inference-perf/pull/84
- Container image
    - https://github.com/kubernetes/k8s.io/tree/main/registry.k8s.io#managing-kubernetes-container-registries
    - Backup choice will be quay
    - Chen to take this up

Action items

- [ ]

May 15, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Brendan Slabe

- Jason Kramberger
- Aishwarya Raimule

Notes

- Open PRs
  - Report formatting - https://github.com/kubernetes-sigs/inference-perf/issues/56
    - [Aishwarya] Why do we need the report to be distinct with the different metrics collectors?
    - We can go with benchmarking metrics, model server metrics and accelerator metrics
  - https://github.com/kubernetes-sigs/inference-perf/pull/74/files
- Recently merged PRs
  - Config refactor for model servers - https://github.com/kubernetes-sigs/inference-perf/pull/72
  - Fill in defaults for config - https://github.com/kubernetes-sigs/inference-perf/pull/71/files
  - Input distribution and synthetic dataset - https://github.com/kubernetes-sigs/inference-perf/pull/66
  - Prometheus metrics for vllm - https://github.com/kubernetes-sigs/inference-perf/pull/64
    - Created a follow up issue for handling multi-stage runs
    - Aishwarya to send up a follow up PR to handle this
- Backlog grooming
- Open topics
  - New model server support - TGI
  -

Action items

- [ ]

May 8, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Brendan Slabe

Notes

- Open PRs
  - Config defaulting - https://github.com/kubernetes-sigs/inference-perf/pull/71

    - Report format - https://github.com/kubernetes-sigs/inference-perf/pull/67
    - Input output distribution + synthetic dataset support
- Recently merged PRs
    - Prometheus client for new model server metrics
    - GCS support for report storage
- Backlog
    - Add TPOT, req latency, input and output token throughput
    - Follow it up with streaming request support and TTFT, ITL measurements
-

Action items

- [ ]

May 1, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Jason Kramberger
- Sachin MV
- Brendan Slabe
- Aishwarya Raimule

Notes

- Open PRs
    - https://github.com/kubernetes-sigs/inference-perf/pull/62 - lint checks
    - https://github.com/kubernetes-sigs/inference-perf/pull/61 - storage client and GCS support to start with
        - Sachin to review
    - https://github.com/kubernetes-sigs/inference-perf/pull/46 - Prometheus client
        - Most of the comments are addressed
        - Aishwarya to merge from a new account
        - Future changes to go in a separate PR
- Recent changes
- Issues in backlog
    - Brendan to make lint checks blocking for PR merges
- Can we add a mini-kube set up with inference-perf, model server and prometheus?
    - Sachin to create an issue
- Input / output distribution

    - Can we allow configuring input and output length along with aggregate requirements like mean, standard deviation, etc.?

Action items

- [ ]

Apr 24, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Brendan Slabe
- Sachin MV

Notes

- https://github.com/kubernetes-sigs/inference-perf/pull/49
    - Prometheus metrics one might be delayed by a few more days
    - Do we wait for stage 1 requests to complete?
        - We have a configurable wait time in between
    - Burst request support
        - Can we use stage request rate, time and cool down period in between to support burst in traffic
- Open PRs
    - Prometheus metrics PR
        - Metrics client is collecting the metrics and storing it in memory temporarily. Can we have a more durable way to do this? But not a blocker for this one
    - https://github.com/kubernetes-sigs/inference-perf/issues/51 - add support for Kubernetes deployment
        - Brendan to take this up - add a job to deploy the tool along with a config map for configuration, provide a way to get the report from the job
    - CSV report
        - Let's create a separate issue for full CSV dump of request level stats
    - Accurate token counting - https://github.com/kubernetes-sigs/inference-perf/issues/45
        - Sachin to follow up with Vivek on next steps
- Open issues and what to prioritize next
    - Add additional metrics
    - Deploy on Kubernetes
    - Have a well formatted report

Action items

- [ ]

Apr 17, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Meeting canceled

Apr 10, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Brendan Slabe
- Aishwarya Raimule

Notes

- Open PRs
    - Prometheus client for metrics
        - [Aishwarya] Do we expect all model servers to follow this?
            - Yes, good to go with this for now since most of the model servers support it
    - Multi-stage run PR
        - [Ashok] What is the main motivation?
            - [Sachin] Motivation is to run multiple stages - you can specify different QPS, time window, etc.
            - Gateway use case was around supporting different models, lora adapters
- [Ravi Sharma] Looking for MFU with benchmarks
    - Are there open PRs for this?
    - Working at RedHat and customers are looking to compare cost for inference. Can we use MFU and HFU for inference?
- Kubecon learnings
    - Inference-perf slide was included in wg-serving talk

Action items

- [ ]

Apr 3, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Meeting canceled due to Kubecon EU

Mar 27, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Brendan Slabe
- Chen Wang (IBM Research)
- Aishwarya
- Vivek Karunai Kiri Ragavan

Next meeting topics:

- Add release bot PR: https://github.com/kubernetes-sigs/inference-perf/pull/41
  - Should allow us to automatically release tagged versions
- Add fmperf library framework PR: https://github.com/kubernetes-sigs/inference-perf/pull/42
  - Currently pulls model, cluster and workload spec from fmperf
  - Need to review and merge
- Docker registry exploration:
  - quay.io vs registry.k8s.io
  - Some SIGs already have a repo at registry.k8s.io
  - SIG scalability doesn't have a K8s registry repo yet
  - Can we start with quay.io?
    - Follow up with Yuan and Kevin to confirm
- PR for containerizing: https://github.com/kubernetes-sigs/inference-perf/pull/38
- Tokenization PR: https://github.com/kubernetes-sigs/inference-perf/pull/43

Mar 20, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Vivek Karunai Kiri Ragavan
- Brendan Slabe
- Chen Wang (IBM Research)

Notes

- https://github.com/kubernetes-sigs/inference-perf/issues/35
    - Gateway API inference extension has a proto config that they are using to deploy model server and benchmark to simplify
        - https://github.com/kubernetes-sigs/gateway-api-inference-extension
    - Would it make sense to use protobug for our config?
    - Follow up with the gateway extension team and gather requirements
- Record demo with inference-perf e2e
    - Vivek to help out with configuring sharegpt data generator
- Can we set up CI/CD to push to docker hub?
    - Chen to take this up
    - Can we check with Yuan on if there is a wg-serving account that we can use for Dockerhub?
- KubeCon roadmap of fmperf merging to In
    - Chen planning to add a slide on inference-perf and how fmperf is planning to migrate functionality over
    - Planning to contribute some PRs before Kubecon
- Cutting a release
    - Can we label some of the open issues on which ones are needed for the initial release?
    - Can we integrate with a release bot which can automatically cut a release when tagged?
        - Chen can help set up a default one
- Let's keep the meeting next week and cancel the week of Kubecon

Action items

- [ ]

Mar 13, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Sachin MV
- Aishwarya
- Vivek Karunai Kiri Ragavan

Notes

- Batch WG benchmarking effort - AI Batch Datacenter Benchmarking Suite Design
- Parameterization PR
- GH actions on PR
    - How can we enforce this on PR pre-submit?
    - We will create issues for these
- Client side tokenization - Vivek is working on this
- ShareGPT data set is now supported
- Mock metrics client is added
    - Let's extend to include prometheus metrics and others
    - Aishwarya is planning to take this up
- Let's plan for a proof of concept demo in WG Serving

Action items

- [ ]

Mar 6, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Brendan Slabe
- Vivek Karunai Kiri Ragavan
- Sachin Varghese

Notes

- [Victor] Benchmarking landscape - what has been discussed
- [Sachin] Parameterization of tool https://github.com/kubernetes-sigs/inference-perf/issues/29
- Metrics Client PR https://github.com/kubernetes-sigs/inference-perf/pull/32

- [Vivek] Adding a ShareGPT data generator - https://github.com/kubernetes-sigs/inference-perf/pull/33
- Github workflows are passing again
    - Brendan to check in the PR to add tests
- [Ganesh] How to contribute?
    - https://github.com/kubernetes-sigs/inference-perf/blob/main/CONTRIBUTING.md
    - Mainly signing CLA and sending PR should be enough
- Open issues - PTAL and leave a comment to take them up

Action items

- [ ]

Feb 27, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Vivek Karunai Kiri Ragavan

Notes

- vLLM Client initial impl https://github.com/kubernetes-sigs/inference-perf/pull/27
- Things to follow up
    - Tests and how we can make them blocking pre-submit tests
    - Check the Github workflow failures on linting
- Vivek is going to pick up the dataset integration
    - Start with ShareGPT or OpenOrca
- Let's aim for initial end to end working version of the tool and first release for early April to see if we can do it before Kubecon EU
- [Aishwarya] Interested in adding model server metrics support
- [Yue] DCGM metrics collection is done in other places - can add a reference to it
    - fmperf includes a way to pull dcgm metric for GPU power for energy analyze
        - https://github.com/fmperf-project/fmperf/blob/b2e216c4b63066150d4b1575557713b0295f3f0d/fmperf/loadgen/run.py#L228
    - Do you need to have a way to allow direct DCGM metric collection using dcgmi cmd?

Action items

- [ ]

Feb 20, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Sachin MV

Notes

- Orchestrator follow up - https://github.com/kubernetes-sigs/inference-perf/issues/22
    - [Sachin] Agree on serving catalog or other options being more customization friendly
    - Next steps
        - [Yue] Chen and Yue to come up with a short write up based on the feedback on the issue
    - [Sachin] How we want to package / run this + model servers meanwhile
        - We can package the benchmarking tool as a container and point to existing deployments
- Inference gateway ask - https://github.com/kubernetes-sigs/inference-perf/issues/23
- HF Dataset - https://github.com/kubernetes-sigs/inference-perf/issues/25
    - Options to start with
        - ShareGPT
        - OpenOrca
    - Leave a comment before starting
    - Can be generic and can support multiple datasets based on the use case
    - [Ganesh] We are using Sonnet and synthetic generation based on that
    - [Ganesh] Can we use sharegpt to generate specific ISL and OSL?
        - ShareGPT or other datasets are meant to mimic a specific varied use case
        - Having input to the load generator which will allow us to configure the different dataset / input output distributions would be ideal
        - Multi-turn chat use case is different and is something genai-perf is adding soon
- vLLM Model Client - Next steps
    - https://github.com/kubernetes-sigs/inference-perf/issues/14
- [Ganesh] Would like to update the genai-perf section in the proposal
- [Ganesh] API format for model server clients - we are going with Open AI API format and there should be customizability at the model server client level
- [David] kServe uses llm-load-test today - RedHat with the recent Neuralmagic acquisition is using another tool - https://github.com/neuralmagic/guidellm/tree/main
    - Keeping tabs on inference-perf
- [Ganesh] Trace based routing - replay customer payload. You can create a load shaper and send it to the load generator

Action items

- [ ]

Feb 13, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Ashok Chandrasekar
- Sachin MV
- Chen Wang
- Yue Zhu
- Cong Liu
- Brendan Slabe

Notes

- [Sachin] Basic Loadgen Impl https://github.com/kubernetes-sigs/inference-perf/pull/21
- [Cong] Benchmark the inference gateway extension - https://github.com/kubernetes-sigs/gateway-api-inference-extension
    - Can we extend this tool to use for gateway benchmarking?
    - Creating Github issues would be the way to go to bring up new requests.
    - Sharing learnings from inference gateway extension benchmarking.
    - [Yue] Automation framework - can we use fmperf or merge it?
- [chenw] how does the orchestration of servers and jobs fit into the existing structure of code?
    - Current structure works well with loadgen, dataset, etc.
    - How to include k8s orchestration?
        - Fmperf today uses python library for k8s orchestration where it can deploy by pointing to a specific cluster
        - https://github.com/fmperf-project/fmperf/tree/main/fmperf
    - Chen to create Github issue to add python orchestration
    - [Sachin] How do we colocate benchmarking tool alongside model server in the cluster?
    - [Chen] fmperf assumes / runs load tester and model server together in a single cluster
    - [Yue] Should we split this out and make sure they are separated out to make it easier to consume for different cases?
-

Action items

- [ ]

Feb 6, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

- Chen Wang
- David Yastremsky
- Gerry Seidman
- Vivek Karunai Kiri Ragavan
- Sachin MV
- Brendan Slabe
- Jason Kramberger
- JooHo Lee
- Tony Song
-

Notes

- [Ashok] Issues created in the repo for new features
- [Sachin] Load generator implementation
    - Start with the python loadgen
    - Add k6 or other integrations at a later time
    - If we can get a first draft of loadgen ready soon.
- [chenw] how to merge with fmperf
    - Are we using Poetry?
    - We are using pdm and it is similar
    - Code formatting tool: ruff, lint
- [chenw] Is going to KubeCon EU to talk about existing benchmarking tool with NVIDIA.
    - Can help get the word out about inference-perf
- [david] from NVIDIA and implementer of genai-perf
    - Looking forward to contributing
- [Sachin] How do we connect payloads?
    - Synthetic vs other datasets
    - Tokenizers and model specific things
    - Welcome Tony!
- [Tony] Looking to contribute
- [Ashok] Request body and how we want to define it
    - [Sachin] OpenAI API spec might be good to start with

Action items

- ☐ [chenw] start merging code from fmperf to Inference Perf
- ☐

Jan 30, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

Notes
- [Ashok] Follow up on the k6 proposal - https://github.com/kubernetes-sigs/inference-perf/issues/2
- [Brendan] PRs in review for model server client and metrics client
- [Vivek] Can we create issues to start implementing specific pieces?
    - Ashok to create issues for loadgen and model server clients

Action items
- ☐

Jan 23, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

Notes
- [Sachin] Loadgen proposal to use k6 - https://github.com/kubernetes-sigs/inference-perf/issues/2
    - Can we have python bindings for the k6 library?
    - Tokenization and other extensions make it easier if it were in python
    - Will investigate further and follow up next week
    - Vivek to follow up on this as well
- [Brendan] PR in review - Add presubmit lint and format checks - https://github.com/kubernetes-sigs/inference-perf/pull/3
    - Start with Github actions for now to do linting and formatting
- [Ashok] Next steps
    - Initial vLLM integration - start with a basic OpenAI API request and measure just the throughput and latency.
    - We should be able to port the existing integration from available benchmarking tools.
    - Extend from there to make it more feature rich and robust.
- [Brendan] is looking into adding the vLLM integration too.
- [Ashok] Can we do a Kubernetes blog and make the tool widely available - maybe a release before Apr 1?
    - We can include it in the wg-serving maintainer track talk
    - Other avenues?

Action items

- [ ]

Jan 16, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:

Notes

- [Ashok] Inference-perf repo created - https://github.com/kubernetes-sigs/inference-perf
- [Ashok] PR on directory structure out - https://github.com/kubernetes-sigs/inference-perf/pull/1
- [Sachin] Interested in contributing to loadgen
- [Yue] Do we want to support different load generators / datasets including synthetic ones?
-
- [Vivek] Interested in working on loadgen / reportgen
- [Yue] Energy metrics - collection and reporting
- [Sachin] PR process - any requirements today?
    - CLA approval and approval from code owners will be needed
- [Sachin] Metrics collection - how do we want to add it?
    - Let's add the initial performance metrics and extend to include accelerator and model server metrics
- [Sachin] API spec - how do we want to interact with model servers? Do we need protobuf or do we want to standardize on an API spec?
    - Existing tools have example implementations with OpenAI API spec being the most widely supported

Action items

- [ ]

Jan 9, 2025 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees: Ashok Chandrasekar Yuan Tang

Notes

- Proposal approved - https://github.com/kubernetes-sigs/wg-serving/pull/26
    - Got verbal approval from SIG scalability lead
    - Who can create the repo once we get approval?

    - We need to create another another issue under kubernetes-community repo to create the benchmarking repo
    - Follow up with the NVIDIA folks as well
    - Identify a set of initial contributors
- [Vivek] Interested in contributing - currently a student at Purdue
    - Where to start?
    - We will have Github issues and pick new contributor items
    - This is a new project, so it will take some time to get to a more mature state
- [Sachin] Main structure of the project
    - Different components and where to start? What do we pull in from existing tools?
    - How do we design the loadgen component - especially if we want a distributed loadgen that we need to scale out?
    - K6 is what I currently use that allows scaling in k8s to more pods.
        - https://github.com/grafana/k6-operator
        - https://grafana.com/docs/k6/latest/testing-guides/running-distributed-tests/
- [Ashok] Are you planning to use the benchmarking tool in Purdue?
    - [Vivek] We are currently using vLLM benchmarking tool
    - Once available, we can use this
-

Action items

- [ ] Send out the email to SIG scalability group

Dec 5, 2024 |
WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:
Recording:

Notes

- Naming - most votes for **inference-perf (7 votes)**
    - **[chenw]+1**
- Proposal here for approval - https://github.com/kubernetes-sigs/wg-serving/pull/26
- SIG scalability sponsorship update
- [Sachin] What does sponsorship mean? New to the process. From Capital One and looking to see how I can work more closely.
- [Yue] Metrics are mostly traditional - throughput, latency, etc. Do we want to support RAG related workflows?
    - What metrics matter for RAG?
    - Can be a good extension
- [Samuel] Proposal overall looks good - captures things we have discussed before
- [Sachin] Naming looks good

Action items

- [ ]

Nov 21, 2024 |

WG-Serving: Benchmarking Tool Contributors Meeting

Attendees:
Recording: https://www.youtube.com/watch?v=jvM6bdqizsA
Notes

- Benchmarking tool ownership update - SIG scalability
- Tool naming - finalize on a name so we can create the subproject repo
    - Inference-perf
    - Inference-perf-tester
    - Inference-performance-evaluator
    - Serving-perf-tester
    - Genai-benchmark
    - GenAI-serving-benchmark
    - ~~GenAIPerf~~ (taken by NVIDIA)
    - AIServingPerf
    - NNServeBench
    - Fmperf
    - Fmperf-bench (perf refers performance)
    - Should we focus on just genai, what about outside of serving use cases? training?
        - [Yuan] Might be better to avoid having GenAI in the name so it can be expanded to other cases if needed
        - A tool that can measure serving performance is our focus - so other models can be benchmarked as well outside of GenAI
        - [Samuel] Embedding models were tested as well
        - [Chen] Our focus has been on LLMs mainly
            - Embeddings is probably the common term to describe vectorizing and tokenizing parts
            - Main focus is
                - Serving
                - Benchmarking
                - ~~Kubernetes~~
                    - ~~[Yuan] this will be under k8s-sigs so we can remove the redundant “k8s” in the name~~
        - [Yue] Do we want to cover accelerator performance or e2e performance?
    - Create a poll on the Slack channel

- [Yue] Energy measurement on accelerator
    - Can we include energy measurement?
        - Per token energy efficiency
        - Raw data - allow user to compute derived metrics
        - We report it for observability, doesn't have to be in the leaderboard or to measure how well a model / server performs
        - We can make this an optional metric to report instead of must have. Makes it easier to add other accelerators like TPU
- Kubecon EU submission
    - [Chen] submit one with NVIDIA speakers on step to step hands on tutorial.
        - Fmperf and GenAIPerf talk has been submitted
        - Can we get NVIDIA folks to contribute to this as well?
            - They have perf-analyzer / genai-perf which they use with Tensorrt-llm
            - Neelay Shah from NVIDIA who is the contributor on genai-perf
    - [Yuan and Eduardo] WG Serving session
        - Speakers not finalized yet
        - We can contribute to the slides and include more about the benchmarking tool
    -

Action items

- [ ]

Oct 31, 2024 | WG-Serving: Benchmarking Tool Discussion

Attendees:
Recording:
Notes:

- [Ashok / Chen / Yuan] GenAI benchmarking tool proposal - [PUBLIC] GenAI Benchmarking Tool - WG Serving Subproject Proposal
    - [Uma] Can the load tester be in Go or other languages
        - [chenw] Orchestrator should be able to take in different load tester
        - [Ashok] There will be a default load tester, probably in Python
        - [samuel] default language for python to be fewer code to change, [Ashok] seconds that.
        - [Ashok] we need to define things to collect. So we can add to the proposal more metrics if needed.
    - [chenw] output format and metric list: please contribute to the document
        - [Felix] Client-side metrics (TO BE DETERMINED)
        - [Felix] Server-side metrics, server-side traces
        - [Ashok] with or without streaming, we need to be able to extend it

        - [samuel] We need to determine a default set of metrics, which are not server dependent, and then we expand it.
- [Ashok] Tool naming
    - [chenw] perf, bench in the name, llmperf, genaiperf has been taken
    - [ashok] genAI, llm in the name
    - [samuel] inference in the name or not?
        - inference-perf-tester, serving-perf-tester?
    - Candidates:
- [Samuel] Repo Creation?
    - [Kellen] We cut a PR in this repo first to quickly get consensus from wg-serving group: https://github.com/kubernetes-sigs/wg-serving/tree/main/proposals/
- [Samuel] KubeCon Europe 2025 proposal?
    - CFP closes 24 November
    - Maintainers call from wg-serving group?
    - Who wants to attend as presenters?
        - Samuel Monson,

Oct 24, 2024 | WG-Serving: Benchmarking Tool Discussion

Attendees: Chen Wang, Ashok
Recording:
Notes

- [ashok]Updates since last meeting
- [Samuel] Added llm-load-test for comparison
- [felix] Introduce load-tester supporting RBS from IBM
    - Supports RPM today - async and batched mode
- [Samuel] Output metrics collected by llm-load-test are from client-side
    - TTFT latency per request? (for streaming scenario, from)
    - ITL(TPOT) per token, or per request?
    - Average output token latency (can be measured from the client side), including TTFT in the ITL.
- [Abdulla] Rely on model server for ITL (only at the decoding phase) and TTFT per request.
    - [kellen] seconds on the server side metrics
- [ashok] we need TTFT from client side as well, or even the request latency.
    - [Abdullah] for non streaming use case, TTFT may not be important
    - [ashok] client-side TTFT is important for streaming use case.
- [chenw] raw data as an optional option.
- [chenw & ashok] writing to persistent volume makes sense.
- [felix] options to dump all requests to persistent volume,
- [samuel] json is a good default, would also be nice down the line to support other tools as output:
    - e.g. AIM, prometheus, etc.

- [ashok]Sub-project proposal for the new benchmarking tool under k8s
    - Start new project proposal on that, ashok will start the doc.
- [chenw] do we want to separate benchmarking libraries from load testers?
    - [ashok] good to keep benchmarking libraries itself usable for different load testers
    - [samuel] orchestrator to be very optional from the whole stack. Load-tester is the real value added here, it would be easier to have a standard here.
    - [felix] orchestrator should be able to support new load-tester image easily.
    - [ashok] we would still want default loadgen to be available in orchestrator to be used.
- Contributors interested in contributing to the tool
    - Thomas Parnell, Ghazi Syed, Chen Wang, Samuel Monson, Yue Zhu, Felix George, Umamaheshwari Devi, Ashok Chandrasekar
- Next steps
- [chenw]To discuss
    - Next step TODOs?
        1. Standardize the output of various load-testers to be used by Orchestrator.
            1. How?
                a. What metrics to collect?
                b. How we collect those data?
                    i. Persistent volume?
                    ii. Prometheus?
        2. Support different load-testers and have a reference load-tester for output.
        3. Non-k8s container support
-

Action items:

- [ ]

Oct 17, 2024 | WG-Serving: Benchmarking Tool Discussion

Notes

- Fmperf vs latency profile generator feature comparison link

Action items:

- [ ]

Oct 3, 2024 | WG-Serving: Benchmarking Tool Discussion

Attendees: Yuan Tang, Ashok Chandrasekar, Chen Wang, Evan Jones, Vishnu Challa, Dan Sun
Recording: https://youtu.be/uOUQCgtyMsc?si=lNTnL6ZGnRxwXVcg

Notes

- Doc: [PUBLIC] Benchmarking LLM Workloads for Performance Evaluation and Aut…
- Go over the updates from discussions so far.

    - Benchmark as code vs benchmark as data
- Red Hat and IBM to present their OSS benchmarking tool.
    - Github: fmperf
        - Features:
            - Python native
            - Launch inference servers on k8s from developer machine and measure performance based on load generation
                - Heterogeneous inference workloads
                - Metrics: TTFT, ITL, throughput
            - Can swap model server
        - Goals:
            - Distributed load testing
            - Plan to make loadgen extensible
        - Slides (TBA)
- GKE to share what they use today.
    - Github: Benchmarking on GKE
    - Used locust initially
    - Prototype status
    - Metrics:
        - TPOT, throughput
    - Can swap model server
        - TensorRT support:
            - Supported, but cumbersome
            - generate requests, basic benchmarking; deployment of TensorRT
    - Can connect to prometheus to generate basic metric outputs
    - Goals:
        - Distributed load testing to support higher concurrency load gen
        - Streaming support
- Openshift HTTP Benchmarking of RAG applications
- Next steps on picking an approach, proposing and sponsoring the project.
    - Using fmperf as starting point
        - What features does it have?
        - What do we need to add?
    - How do we want to integrate it into WG-serving?
        - Willing to donate to Kubernetes-SIGs
    - Preferences on starting from scratch or building on top of fmperf
        - [Ashok] Cannibalize best parts of fmperf and combine with what's working well in our tool
        - [Chen] As long as the user interface of fmperf doesn't change too much, or better yet improves, then there's no problem

Action items

- ☑ ~~Ashok - take closer look at fmperf to identify gaps from GKE perspective~~

Fmperf vs latency profile generator feature comparison

Features	fmperf	Latency profile generator	nvidia genai-perf	vllm benchmark	llm-load-test
Benchmark as code (can be used as a library)	Yes	Yes (requires change)	Yes (requires changes)	Yes (requires changes)	Yes (requires changes)
Multiple model server support	Yes (vLLM, TGIS)	Yes (vLLM, Triton, TGI, Jetstream)	Partial (Triton, NIM)	Partial (vLLM, TGI)	Yes (vLLM, TGIS, HF-TGI, Caikit)
Extensible to support new model server	Model server agnostic	Model sever agnostic	Hardest to extend	Second hardest to extend	Model server agnostic
Benchmarking time control	Yes	Yes	No	No	Yes
RPS support (along with various request rates)	No (can be added)	Yes	Yes	Yes	No (planned)
Multi-user support (concurrency)	Yes	No in	Yes	No	Yes
Containerization	Yes	Yes	No	No	Yes
Run inside a K8s cluster	Partial (Requires steps to create and mount prompt file alongside model weights via host machine)	Yes	No	No	Yes
Tokenization / token level metrics	Yes (only supports server side now, would require model server level changes)	Yes (client side)	Yes	Yes	Yes (client side)
Dataset support	Yes (quac)	Yes (sharegpt)	Yes (openorca, cnn_dailymail)	Yes (sharegpt, sonnet, etc.)	Yes (openorca, code_alpaca,

					pluggable)
Streaming requests	Yes	No	Yes	Yes	Yes
Pulling metrics from GPU	Yes	No (Can be added)	Yes	No	No
Pulling metrics from model server	No (Can be added)	Yes	No	No	No
Multi-model support for LoRA benchmarking	No (Can be added)	Yes	Yes	No	No
Deploy inference server / cluster (Ochestrator)	Yes (via code)	Yes (via terraform)	No	No	No
Support running outside of Kubernetes	No		?	?	Yes?