CATS Problem Statement, Use cases and Requirements

draft-ietf-cats-usecases-requirements-02

K. Yao, China Mobile
D. Trossen, Huawei
M. Boucadair, Orange
LM. Contreras, Telefonica
H.Shi,Y.Li,Huawei
S. Zhang, China Unicom
Qing An, Alibaba Group

Contents Outline

Table of Contents

- Merged one use case draft draft-an-cats-usecase-ai-01(expired)
- Merged a requirement from draft-yuan-cats-end-to-end-problem-requirement-01

Draft updates: Use cases

Computing-Aware AI Large Model Inference

- highly interactive and real-time translation applications are promising, which are typically based on Large Language Models(LLMs).
- training LLMs needs large scale computing clusters, around tens thousands of GPU cards, which are primarily implemented in DCs.
- training doesn't require CATS since the behaviors of compute groups in training are pre-defined.
- well-trained models are deployed for inference distributedly, which follows a Cloud-Edge-Device co-inference mode.
- selecting different edge inference instances in real-time is within CATS scope.

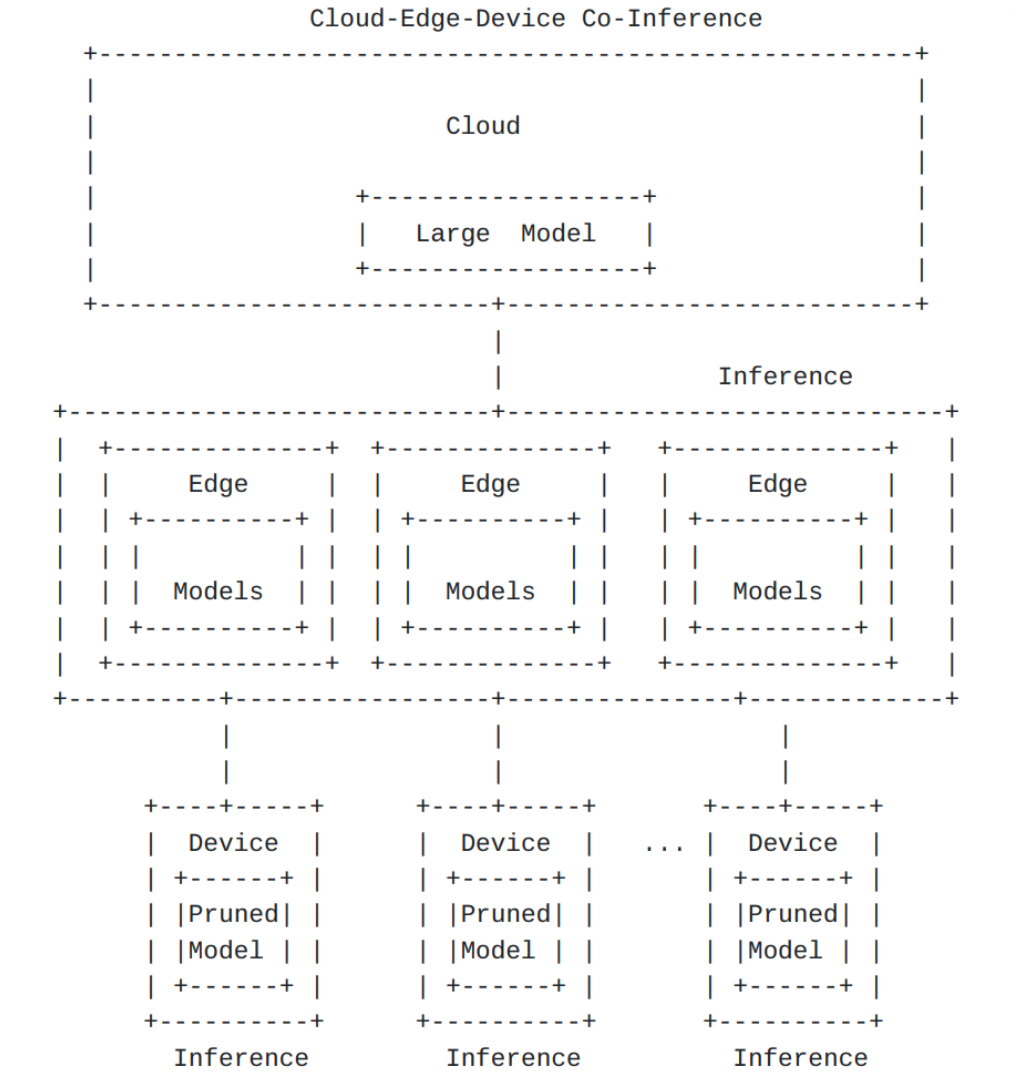


Figure 5: Illustration of Computing-aware AI large model inference

Draft Updates: Requirements

➢ Support Moderate Metric Distributing:

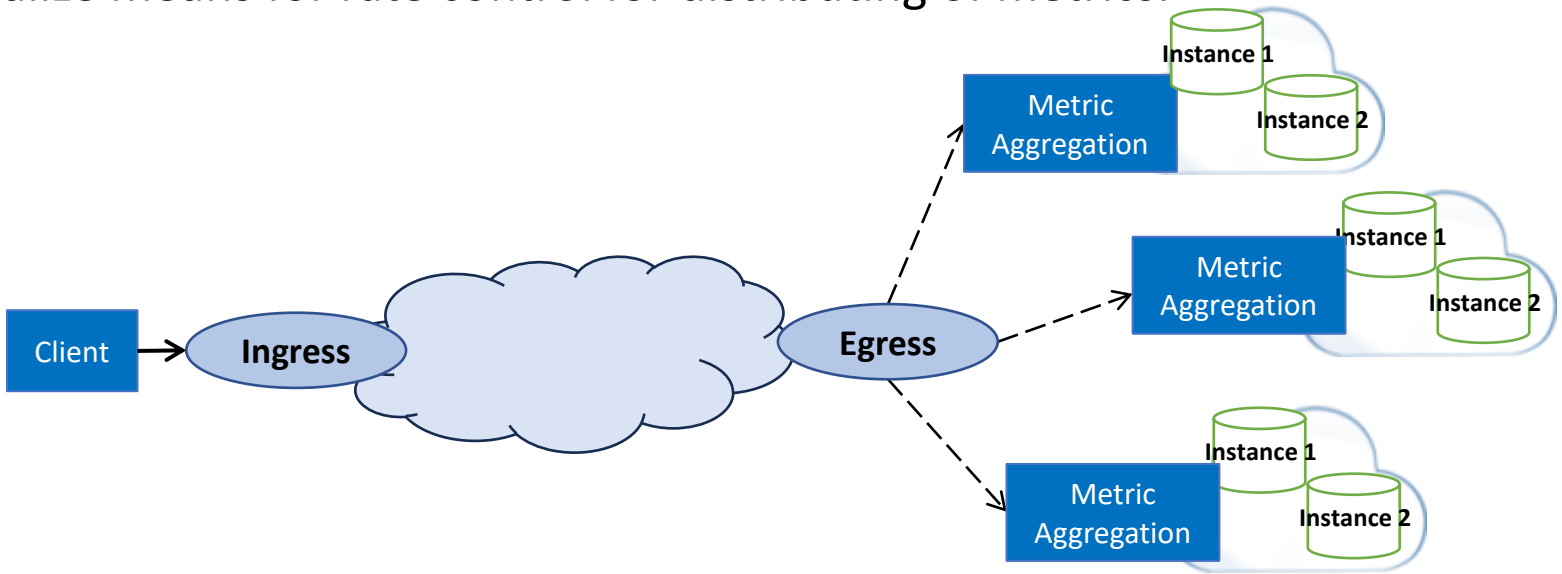
- R8: MUST provide mechanisms for metric collection.

  Collecting metrics from all of the services instances may incur much overhead for the decision maker, and thus hierarchical metric collection is needed. That is,

- (newly added)R9: SHOULD provide mechanisms to aggregate the metrics.

  CATS components do not need to be aware of how metrics are collected behind the aggregator.

- R10: MUST provide mechanisms to distribute the metrics.
- R11: MUST realize means for rate control for distributing of metrics.



Draft Updates: Requirements

➢ Support Alternative Definition and Use of Metrics:

- R12: a computing semantic model SHOULD be defined for the mapping selection.
- R13: In addition to common metrics that are agreed by all CATS components like processing delay, there SHOULD be some other ways for metrics definition, which is used for the selection of specific service instance.
- R14: MUST set up metric information that can be understood by CATS components.

  For metrics that CATS components do not understand or support, CATS will ignore them.

- ~~(Previous)R14: MUST include a default action for the interoperation of network nodes which may or may not support the specific metrics.~~

Therefore, a desirable system	Therefore, a desirable system
R13: MUST set up metric information that can be understood by CATS components.	R14: MUST set up metric information that can be understood by CATS components.
R14: MUST include a default action for the interoperation of network nodes which may or may not support the specific metrics.	For metrics that CATS components do not understand or support, CATS components will ignore them.

Summary of Discussion on mailing list:

- **Discussion on whether SFC should be a proper use case in CATS:**
  (https://mailarchive.ietf.org/arch/msg/cats/xtluUPFMjS3urPLDbYGbnWCGbC4/)

  - **About the Definition of "Service" in SFC and CATS:**
    - Service in CATS: Services that are addressed by client applications.
    - Service in SFC: Operator based services that could improve client applications.

  - **Differences:**
    - CATS is about sending traffic to a compute instance(running client applications), and then get a response.
    - SFC is about sending traffic through instances(running operator service functions) along a pre-defined path(SFP) towards a destination.
    - SFC may not need a response, unless it is a loop which is not an usual case.

  - **Similarities:**
    - Construction of SFC/SFP can be enhanced based on computing awareness, but this is not a use case in CATS WG. (out of scope)

Next Steps:

- Should we update the terms of “Computing Service” of the two WG documents to clarify that CATS is more about user traffic?
- Since metrics definition and metrics distribution are the primary work for the next step, are there any reflections on whether current requirements should be improved?
- Any other comments?

Thanks!