

Computing Information Description in CATS

draft-du-cats-computing-modeling-description-01

Zongpeng Du, Yuexia Fu, Cheng Li, Daniel Huang, Zihua Fu

duzongpeng@chinamobile.com, fuyuexia@chinamobile.com,
c.l@huawei.com, huang.guangping@zte.com.cn, fuzhijia@h3c.com

IETF117

Outline

- Interim discussion
- Modifications of version 01
- Next step

[Cats] Thoughts about CATS metrics

Adrian Farrel <adrian@olddog.co.uk> Thu, 04 May 2023

- Hi WG,
- **A small group had a one hour call last week to discuss metrics for CATS.**
- This email is to let you know a summary of what we concluded and out proposed next steps. It brings the discussion onto the public mailing list (where it belongs). In keeping with all IETF work, these off-list discussions do not constrain the working group in any way, but hopefully they will stimulate progress.
- Please continue the discussion. You're all encouraged to review and comment on [draft-du-cats-computing-modeling-description](#).
- Best Adrian

Summary of the discussion

- It is important the metric scheme used is **flexible and extensible to support future requirements** for metrics or metric-combination schemes we haven't thought of yet.
- For simplicity of specification and implementation, **the initial metrics** specification should cover only those metrics we think we need to solve immediate problems.
- On the call, the only requirement that we identified that would be used to select a server/instance in the immediate use cases is "**delay**": that includes network propagation time and processing time. But, as above, it must be possible to add new metrics in the future, and email discussions have suggested that we might want a "composite metric", bandwidth, or server capacity.
- The precise meaning of delay for compute and **other possible metrics** needs to be discussed and standardised **if it is to be useful**, because the ingress edge and all implementations of the same service need to have a common understanding.

Modifications of version 01

- We modify mainly the section 5 “Computing Resource Modeling”, with the **intent to start with simple metrics.**

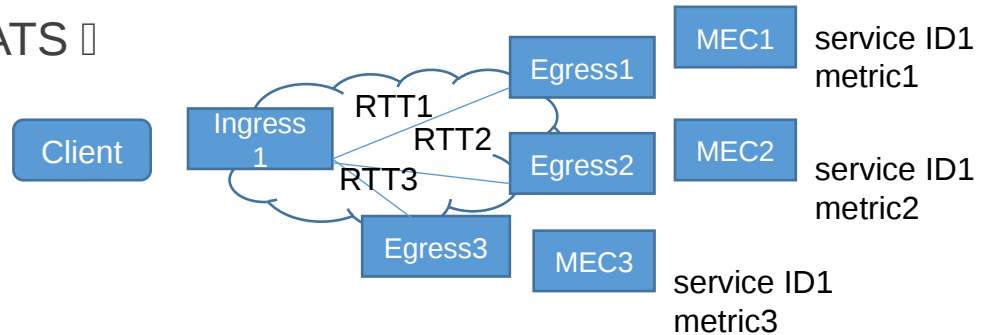
5. Computing Resource Modeling	6
5.1. Consideration of Using in CATS	7
6. Network Resource Modeling	8
6.1. Consideration of Using in CATS	8
7. Application Demands Modeling	9
7.1. Consideration of Using in CATS	9
8. Security Considerations	9
9. IANA Considerations	9
10. Acknowledgements	9
11. Contributors	9
12. Informative References	10
Appendix A. Related Works on Computing Capacity Modeling	11
Appendix B. Architecture of Computing Modeling	12
B.1. Computing Capacity	13
B.1.1. Types of Chips	13
B.1.2. Type of Computing	14
B.1.3. Relation of Computing Types and Chips	15
B.2. Communication, Cache and Storage Capacity	15
B.3. Comprehensive Computing Capability Evaluation	16
Authors' Addresses	16

5. Computing Resource Modeling	6
5.1. Requirements of Using in CATS	7
5.2. Consideration of Using in CATS	8
6. Network Resource Modeling	9
6.1. Consideration of Using in CATS	9
7. Application Demands Modeling	10
7.1. Consideration of Using in CATS	10
8. Security Considerations	10
9. IANA Considerations	10
10. Acknowledgements	10
11. Contributors	10
12. Informative References	11
Appendix A. Related Works on Computing Capacity Modeling	12
Authors' Addresses	13

5.1. Requirements of Metric Using in CATS

- We believe that the advertise / propagate / use of metric in CATS are all related, and would influence each other
- So we recall the scenario of the CATS

- ❑ The same service can be provided in multiple places in the CATS.
- ❑ They have the same service ID, but different metrics.



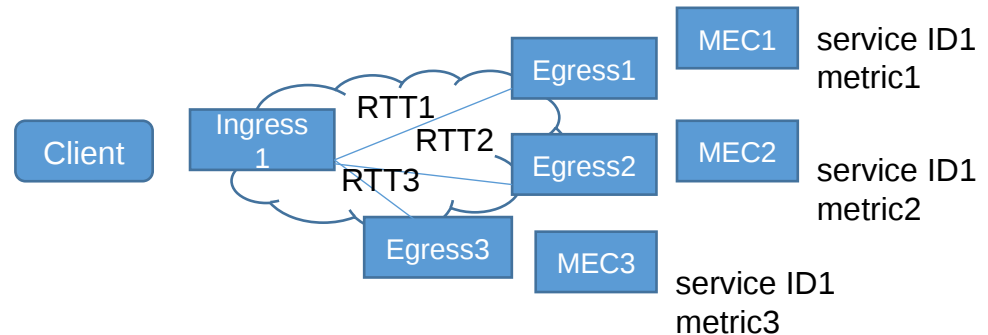
- Step1: the service points need to collect some specific computing information. In this step, only necessary computing information should be considered
- Step2: the service points send the computing information into the network by some means, and update it periodic or on demand
- Step3: the decision point receives the computing information, and makes a decision for the specific service related to the service ID. Hence, the route for the service ID on the Ingress is established or updated
- Step4: the traffic for the service ID reaching the Ingress node would be identified and steered according to the policy in the step3

Some requirements

- What to send, how to send, and the optimization objective of the policy are all related to the design of the computing resource modeling in CATS, meanwhile
- They would influence each other.

- ❑ The same service can be provided in multiple places in the CATS.
- ❑ They have the same service ID, but different metrics.
- Requirements:

- The optimization objective of the policy in the decision point may be various. For example, it may be the lowest latency of the sum of the network delay and the computing delay, or it may be an overall better load balance result, in which we would prefer the service points that could support more clients.
- The update frequency of the computing metrics may be various. Some of the metrics may be more dynamic, and some are relatively static.
- The notification ways of the computing metrics may be various. According to its update frequency, we may choose different ways to update the metric.
- Metric merging process should be supported when multiple service instances are behind the same Egress.



Some design principles can be considered

- The target in CATS mainly concerns about the service point selection and traffic steering in Layer3, in which we do not need all computing information of the service points.
 - Hence, we can **start with simple cases** in the work of the computing resource modeling in CATS.
- Some design principles can be considered:
 - 1 □ The computing metrics in CATS should be few and simple, so as to avoid exposing too much information of the service points.
 - 2 □ The computing metrics in CATS should be evolveable for the future extensions.
 - 3 □ The computing metrics in CATS should be vendor-independent, and OS-independent.

5.2. Consideration of Using in CATS

- We can start with simple cases
- Case1 □ the optimization objective of traffic steering in this scenario is the minimal total delay for the client
 - In this case, the decision point can collect the network delay and the computing delay, and make a decision about the optimal service point accordingly
 - The computing delay can be generated by the server, which has the meaning of “the estimate of the duration of my processing of request”
- Case2 □ Another metric that can be considered is the server capability.
 - For example, one server can support 100 simultaneous sessions and another can support 10,000 simultaneous sessions.
 - The value can be generated by the server when deploying the service instance.
- For some other optimization objectives, we can also consider other metrics, even metrics about energy consumption.

Next step

- Call for comments, and refine the draft
- Modify the draft according to other related documents, for example, the framework draft

Thanks