Joint Exposure of Network and Compute Information for Infrastructure-Aware Service Deployment

<draft-rcr-opsawg-operational-compute-metrics>

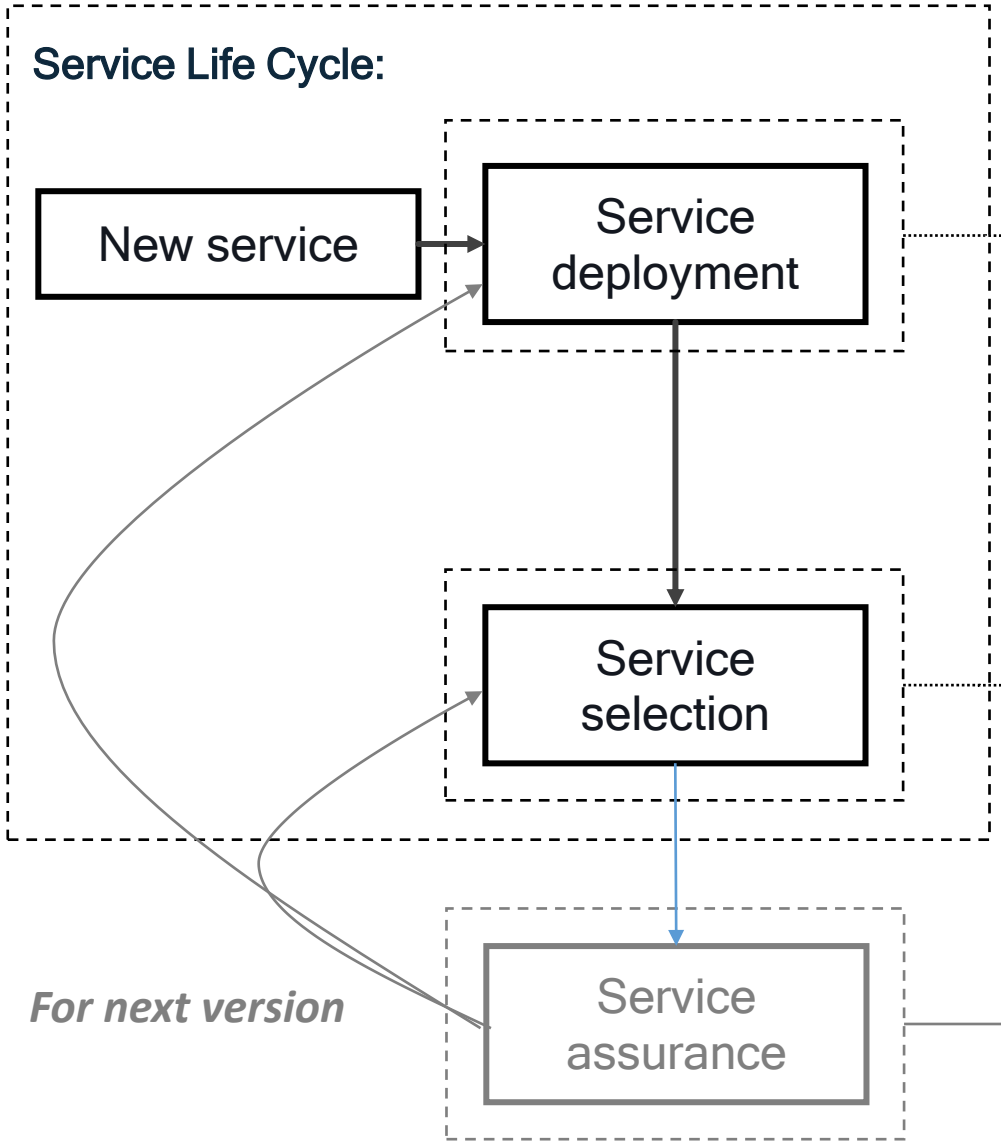
Sabine Randriamasy (*Nokia Bell Labs*), Luis Contreras (*Telefonica*),
Jordi Ros Giralt (*Qualcomm Europe, Inc.*), Roland Schott (*Deutsche Telekom*)

IETF 120, Vancouver, July 2024

Motivation (recap)

- Standardization of network information is quite mature but is in progress for compute information.
- There is a need to define a set of compute metrics to support various use cases being served in the IETF.
- Some ad hoc work exists in the IETF:
  - CATS (e.g., draft-du-cats-computing-modeling-description)
  - ALTO (e.g., draft-contreras-alto-service-edge)
  - OPSAWF (e.g., RFC 7666 MIB)
- Metrics are also defined in other bodies such as the Linux Foundation, DMTF, ETSI NFV, etc:
  - Raw compute infrastructure metrics (e.g., processing, memory, storage)
  - Compute virtualization resources and service quality metrics (e.g., VNF resources in VMs)
  - Service metrics including compute-related information (e.g., service delay, availability)

Problem space



Action to take	Information needed	Who needs it
• Service placement	• Compute • Communication	• Service provider

Action to take	Information needed	Who needs it
• Service selection • Service Path selection	• Compute • Communication	• Network provider • Application (client or proxy)

Action to take	Information needed	Who needs it
• Service assurance (corrective actions)	• Compute • Communication	• Network provider • Application (client or proxy)

Relation to compute metrics in CATS (new slide)

- [draft-du-cats-computing-modeling-description-03] proposes 2 service-level metrics
  - Combined network delay and computing delay (same unit)
  - Server capacity in terms of e.g. sessions
    - Can be used for load balancing restricted to servers with acceptable combined delay
- [draft-rcr-opsawg-operational-compute-metric] focuses on compute metrics
  - Collected and defined at different levels of granularity
    - Depending on the service lifecycle action and information
  - Proposes 2 abstracted generic metrics reflecting server performance and policy-based cost
    - To accommodate need for simplicity and/or restricted information access
  - Addresses service deployment use-case that requires fine grain resources-level info

History and updates from IETF 119

- Draft presented in IETF 118 (-01) and IETF 119 (-03)
- Updates (now in -06 version)
  - Added new section for “Study of the Kubernetes Metrics API and Exposure Mechanism”
    - Understanding the Kubernetes Metrics
    - Example of How to Map the Kubernetes Metrics API with the IETF CATS Metrics Distribution
    - Available Metrics from the Kubernetes Metrics API
  - References to exposure solutions moved to a specific section
  - Editorial fixing

New section 7 on Kubernetes Metrics API

- To address deployment use-case and explore aggregation levels
- To see what can be used in CATS service selection
- 3 sections
  - Section 7.1 lists Kubernetes metric collection Architectures
  - Section 7.2 explores possible mapping with the CATS framework
  - Section 7.3 lists available Metrics from the Kubernetes Metrics API
    - Low-level (container, POD, Node)
- Low-level node resources description for selection are not needed for CATS service selection
  - They are for service placement and assurance

History and updates from IETF 119

- New version to be submitted during IETF 120 week
- Updates (in -07 version under preparation)
  - Consideration of Service Assurance phase for the service lifecycle (section 1)
  - Refinements on metric dimensions to be considered (section 6.5)
  - Editorial fixing
    - Re-wording and typo in section 7.1
    - Revise Table 6

Next steps

- Collect feedback from CATS WG
  - Feedback has been requested to the chairs and to the mailing list
  - Feedback expected today during IETF 120 CATS meeting

- Prepare new version for IETF 121
  - Potentially, to add further details on metrics from some existing solutions