Problem Statement for ROSA

Dirk Trossen (Huawei), Luis Contreras (Telefonica), J. Finkhaeuser (Interpeer), P. Mendes (Airbus)

25.07.2023
Service Routing as Done Today

Explicitly resolving service name onto network locator

• Done through explicit indirection infrastructure
• Client issues resolution request first
• Client then utilizes resolved network locator for the actual application-level request

Example Technologies

• DNS
  • Global Server Load Balancing (GSLB) for CDN-based load balancing
• ALTO (application layer traffic optimizations)
• HTTP redirect
• LISP (to some extent)
Issues observed across our use cases

Centered around distribution (of service instances), dynamicity of choice and availability, as well as efficiency to utilize chosen service instance; more specifically

1. **Anycast**: more than one service instance available
2. **Dynamic decisions**: may want to change choice of service instance frequently, down to single requests even
3. **Ephemeral service instances**: may deploy (or ramp down) service instance frequently, e.g., using microservice architectures
4. **Latency**: explicit resolution costs time, possibly exceeding delay budget of application or making frequent changes expensive
5. **Service-specific selection**: choice of anycast may be based on service-specific policies, in addition to network criteria
6. **Distributed deployment**: service instances may exist in different/many network locations, not even in PoPs
7. **Namespace mapping**: not all services use URLs
8. **Service chaining**: a (composite) service experience may consists of many chained services, accumulating issues on, e.g., latency
WHAT IF...

...a similar end-to-end procedure of data communication between a client and a 'best' choice of service instances (out of set of possibly many) existed that significantly reduced the aforementioned latency, while it allowed for updating the assignments at rates that are more aligned with the possibility to establish new service instances in distributed locations?
Problems to address include

1. How can we **make decisions** on anycast-based service instance assignments **at high rate**, even down to every service request?
   - How to possibly **remove the need for an explicit out-of-band discovery** step, which incurs additional latencies before any data transfer can commence?

2. How could we improve on the **update speed** for the assignments between service name and 'best' IP locator for the service instance to be used...to **align the rate** of the possible anycast assignment update with that of the possible availability of the service instance resource?

3. How could we allow for **incorporating service-specific policies** into the anycast selection?

4. How can we support **any application identifier space** (within the governance defined for that identifier space) beyond domain names?

5. How could the **chaining** of more than one service be realized without explicit discovery latency incurred?

6. How can we possibly preserve such client-driven operation, and thus **avoid transaction state in the network**?
Sidemeeting on 24.07.2023

Material (agenda and slides) at https://github.com/dirk-trossen-huawei/IETF117_ROSA

1. Presented (i) use cases (web browsing and content delivery), (ii) problem statement, (iii) gaps in existing technologies, (iv) possible blueprint

2. Discussed on
   1. validity of problem
   2. possible ways forward

3. Agreed to [will be added verbally]
IETF117 – RTG WG
https://datatracker.ietf.org/doc/draft-mendes-rtgwrg-rosa-use-cases/

Please join the discussion on rosa@ietf.org

Contributors wanted!

THANKS!
25.07.2023