DNS based service discovery for CATS

draft-liu-cats-dns-service-discovery-00

Xiang Liu, Rongwei Yang, Yu Zhang, Di Ma

IETF 123 @ Madrid

Introduction

- A key requirement of CATS framework is providing a discovery and resolving method for the mapping of a service identifier to a specific address
- DNS-SD (RFC 6763) provides a standard mechanism for service discovery and resolution
- We propose to extend DNS-SD to fulfill the CATS requirement and integrate with CATS framework

Service (Instance) Name

```
Service Instance Name =
<Instance>.<Service>.<Domain>
```

- Instance: a user-friendly name for the instance
- Service: indicates the name of a specific type of service
- Domain: indicates the domain name where the service registered

edge-inference-1._cats-inference._tcp.example.com

Service Parameters and TXT records

use TXT records to advertise parameters for each service instance:

- "cpu": CPU capacity in normalized units (integer)
- "mem": Memory capacity in MB (integer)
- "lat": Expected processing latency in milliseconds (float)
- "load": Current load level (0-100) as a percentage (integer)
- "gpu": GPU availability and type (string)
- "accel": Other accelerator availability and type (string)
- "vers": Service version (string)
- "caps": Capabilities as a comma-separated list (string)
- "prio": Priority tier (integer, lower values indicate higher priority)
- "cost": Relative cost metric (integer)
- "avail": Availability status (0=offline, 1=online, 2=degraded)

cpu=8 mem=16384 lat=15.5 load=35 gpu=nvidia-t4 vers=1.2 caps=inference,training prio=1 cost=2 avail=1

SRV records for service location

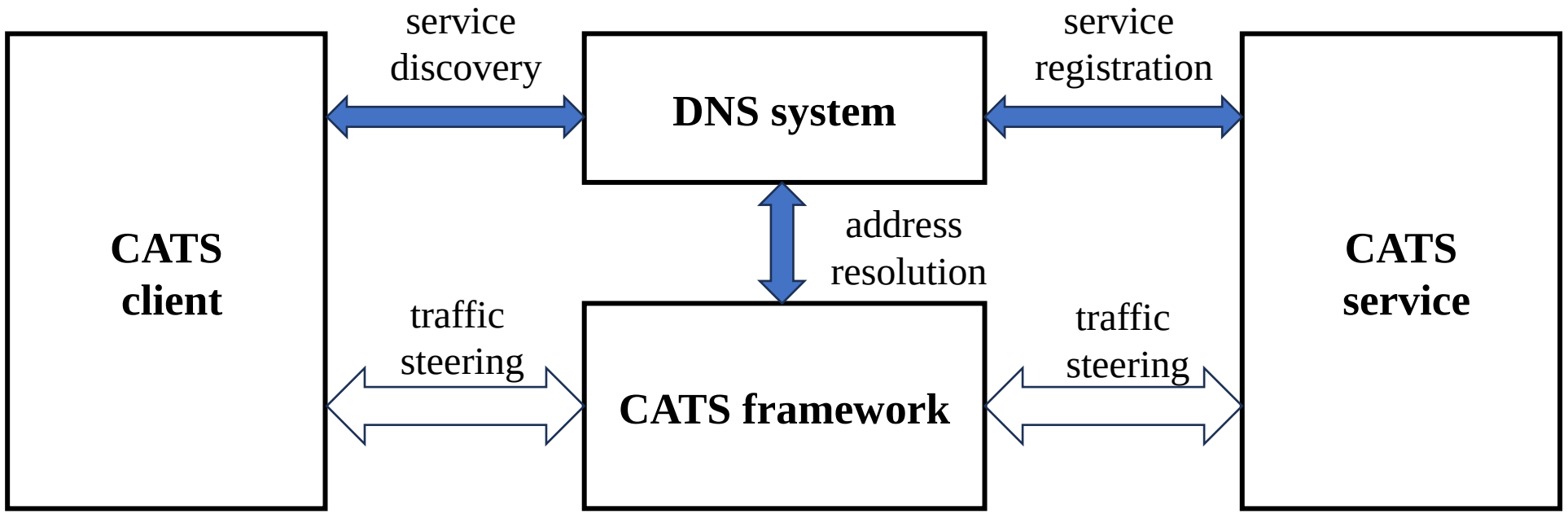
- for each service instance, an SRV record must be published according to RFC 2782, in the following format:

<instance-name>._cats._tcp.<domain> IN SRV <priority> <weight> <port> <target>

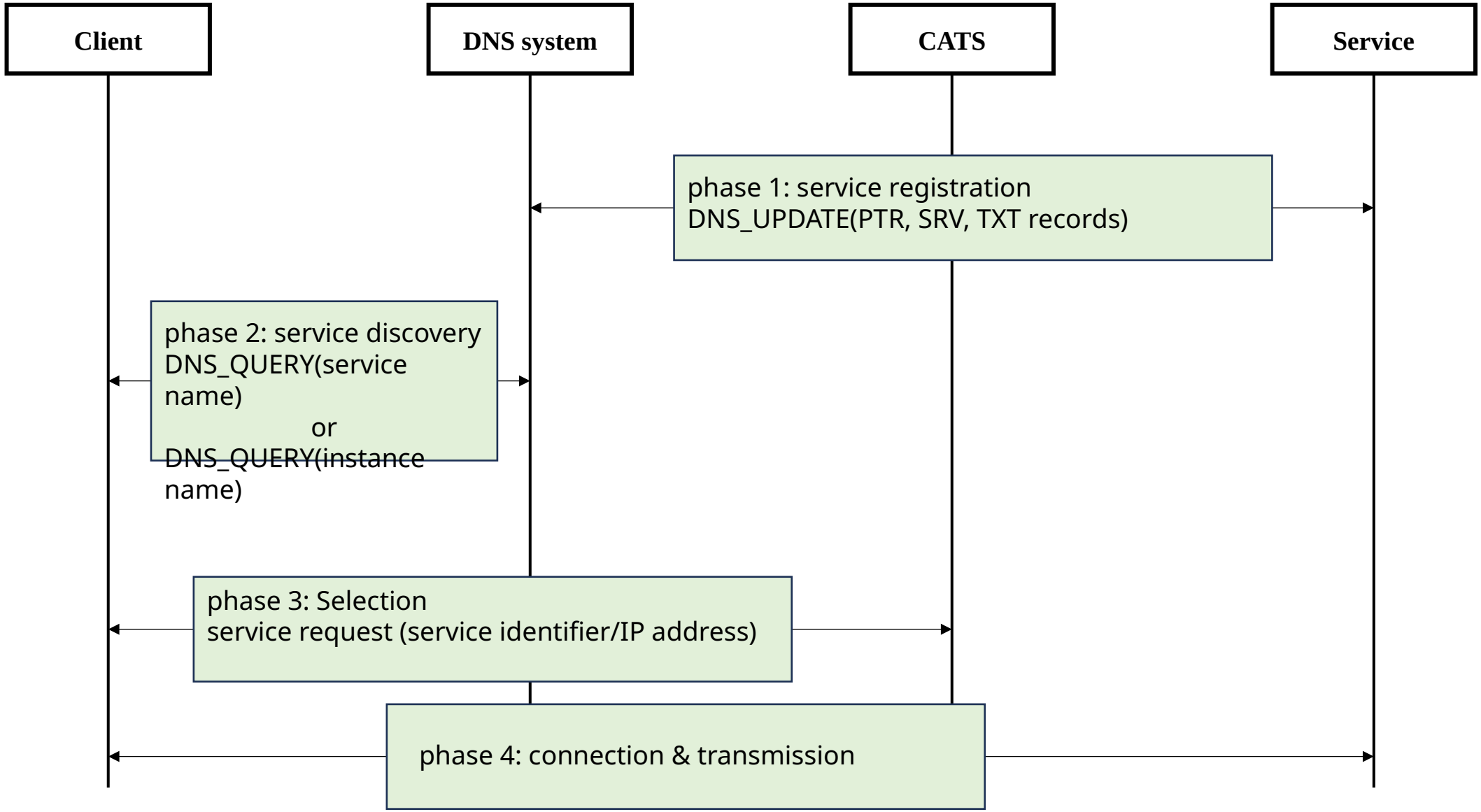
- <priority>: the priority of the target host
- <weight>: used for load balancing
- <port>: port number
- <target>: the hostname of the machine providing the service

For example: edge-inference-1._cats._tcp.example.com. SRV 0 5 8080 compute1.example.com

Integration with CATS framework



Workflow


Client
DNS system
CATS
Service
phase 1: service registration
DNS_UPDATE(PTR, SRV, TXT records)
phase 2: service discovery
DNS_QUERY(service name)
or
DNS_QUERY(instance name)
phase 3: Selection
service request (service identifier/IP address)
phase 4: connection & transmission

Thanks for your attention

We are looking forward to comments and discussions

liux15@pcl.ac.cn