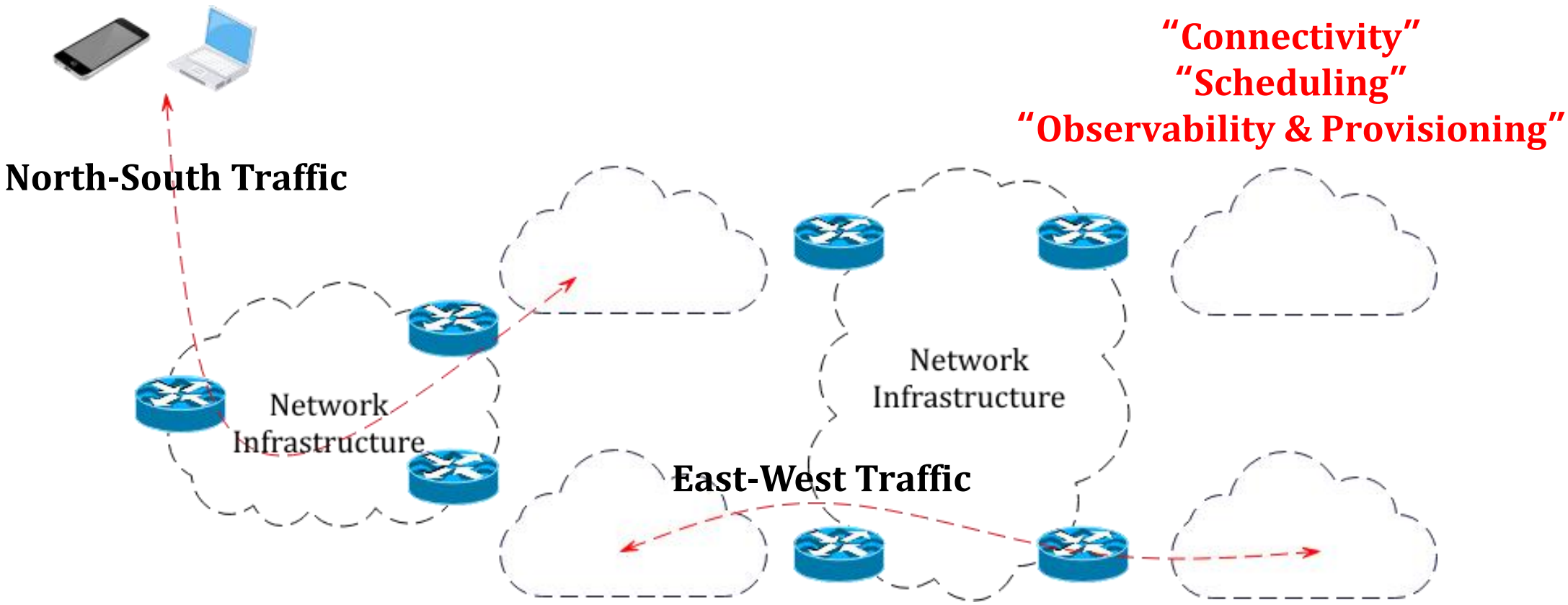
Use Cases-Standalone Service ID in Routing Network

draft-huang-rtgwg-us-standalone-sid-00

Ge Chen, Jie Liang,	China Telecom
Feng Yang,	China Mobile
Yan Zhang,	China Unicom
Daniel Huang, Dongyu Yuan,	
Huakai Fu, Cheng Huang, Yong Guo,	ZTE Corporation

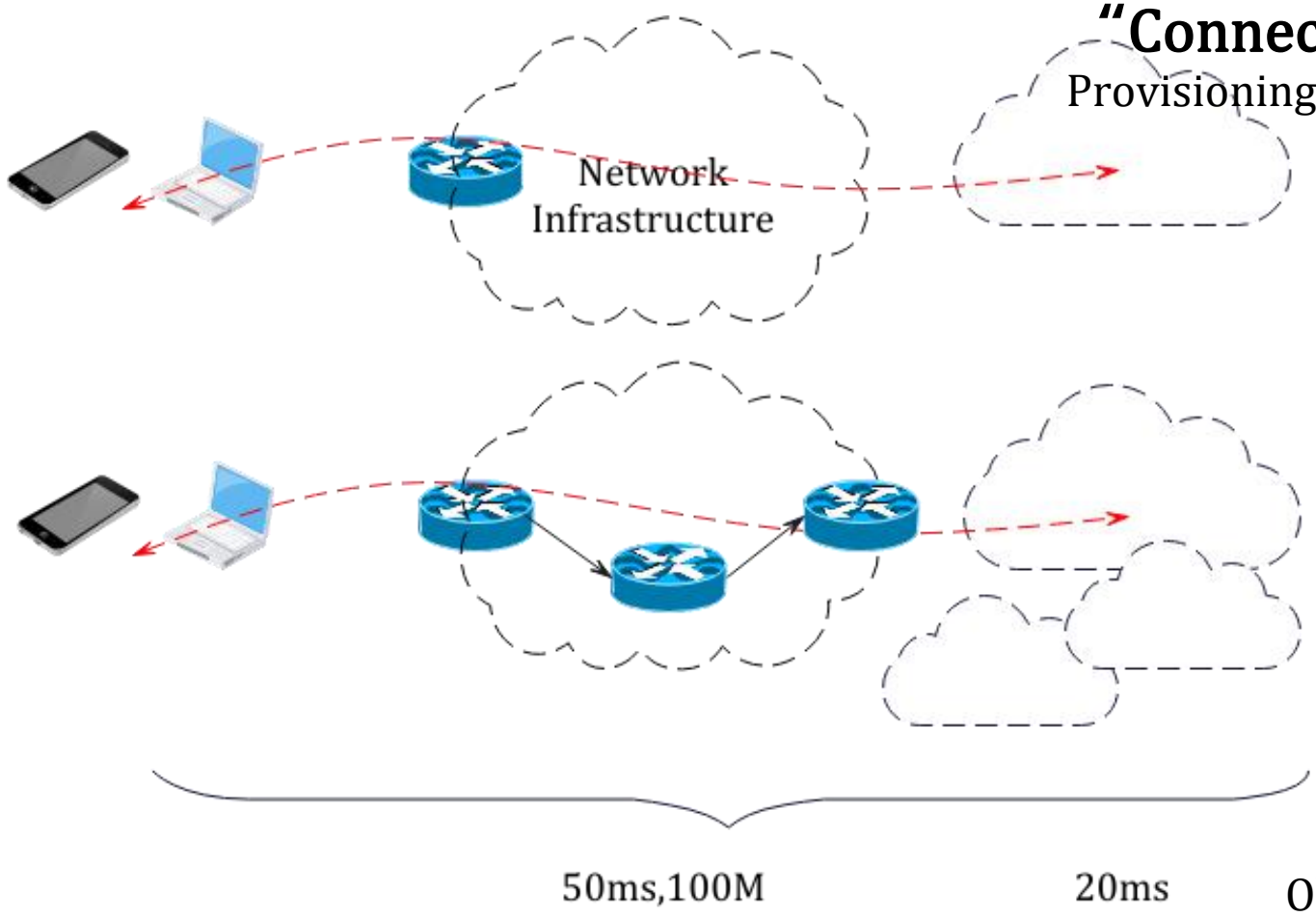
IETF-119, Mar 2024

Apps and service diversity, coordination among micro-services



Identify and guarantee SLAs of multi-tasking flows, scheduling and orchestrating forwarding path for transient life cycle of computing tasks.

Gaps of south-north traffic scenario



“Connectivity” : Network Resources Provisioning and Differentiated Network Services

“Scheduling” : Traffic Steering among Multiple Service Instances (service-based interconnection and routing would be more aligned with the essential demands and expectations of clients and services)

“Observability & Provisioning” : Observability of End-to-End Services (To locate performance bottlenecks, prevent failures, and improve resource utilization efficiency)

Gateways and proxies in the way of east-west traffic

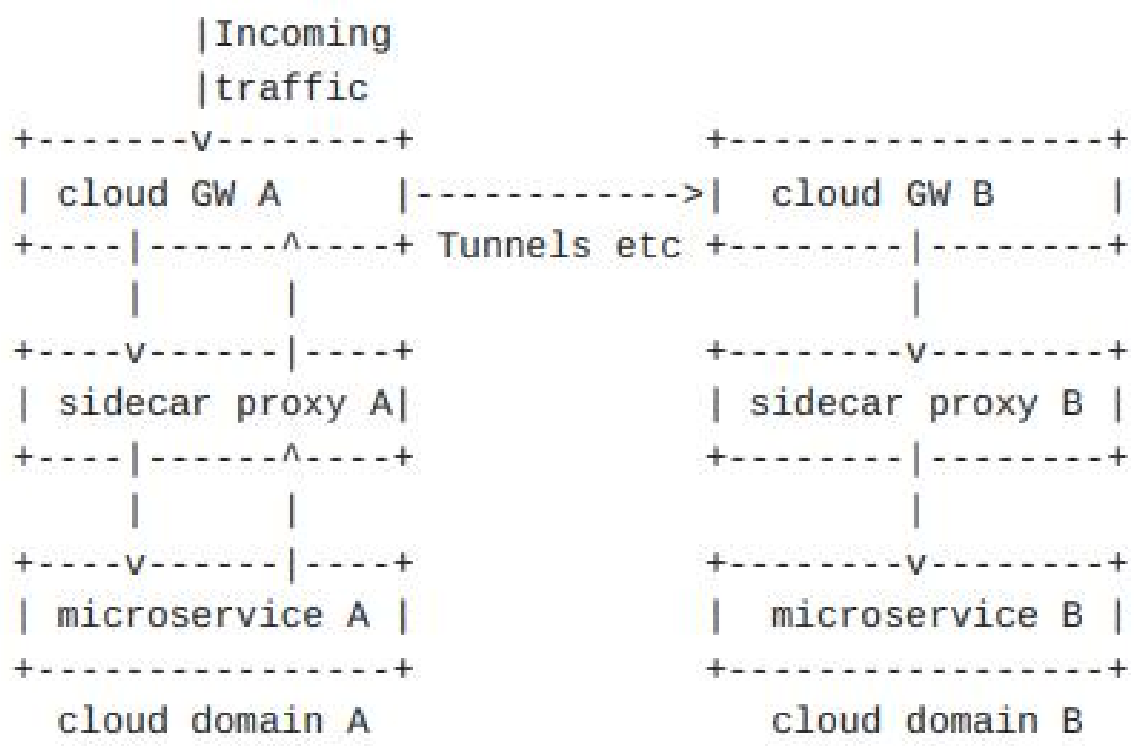


Figure 1: Inter-service communication within multi-domains

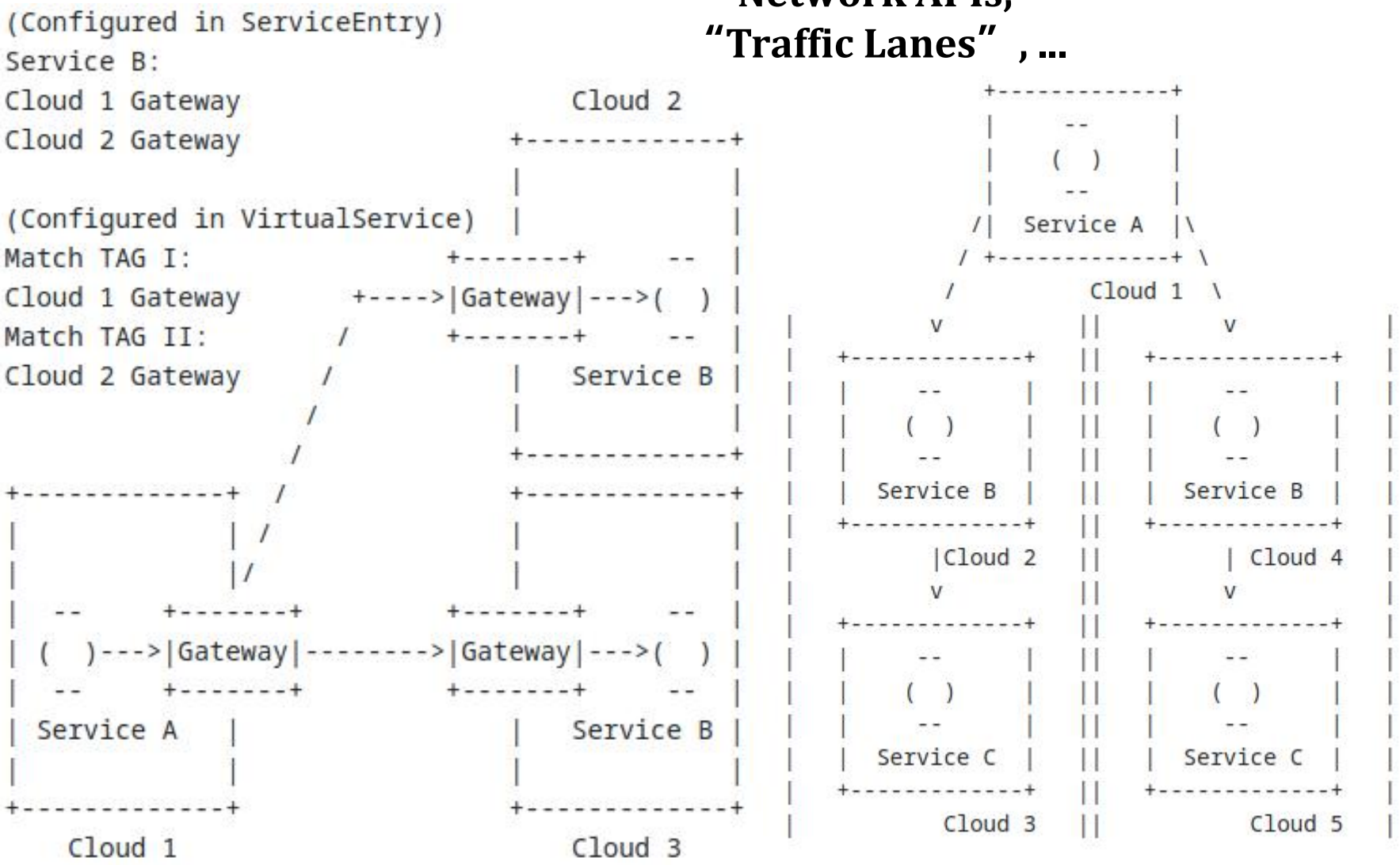
- a proxy intercept traffic and conduct processing
- application semantics are considered
- consists of three TCP segments

Additional processing delay at the gateways for each hop service-to-service communication
---> Large end-to-end delay

Interfaces are needed to establish appropriate mapping between different technologies in two adjacent domains
---> High management complexity

Cross registries and isolations from service interconnections

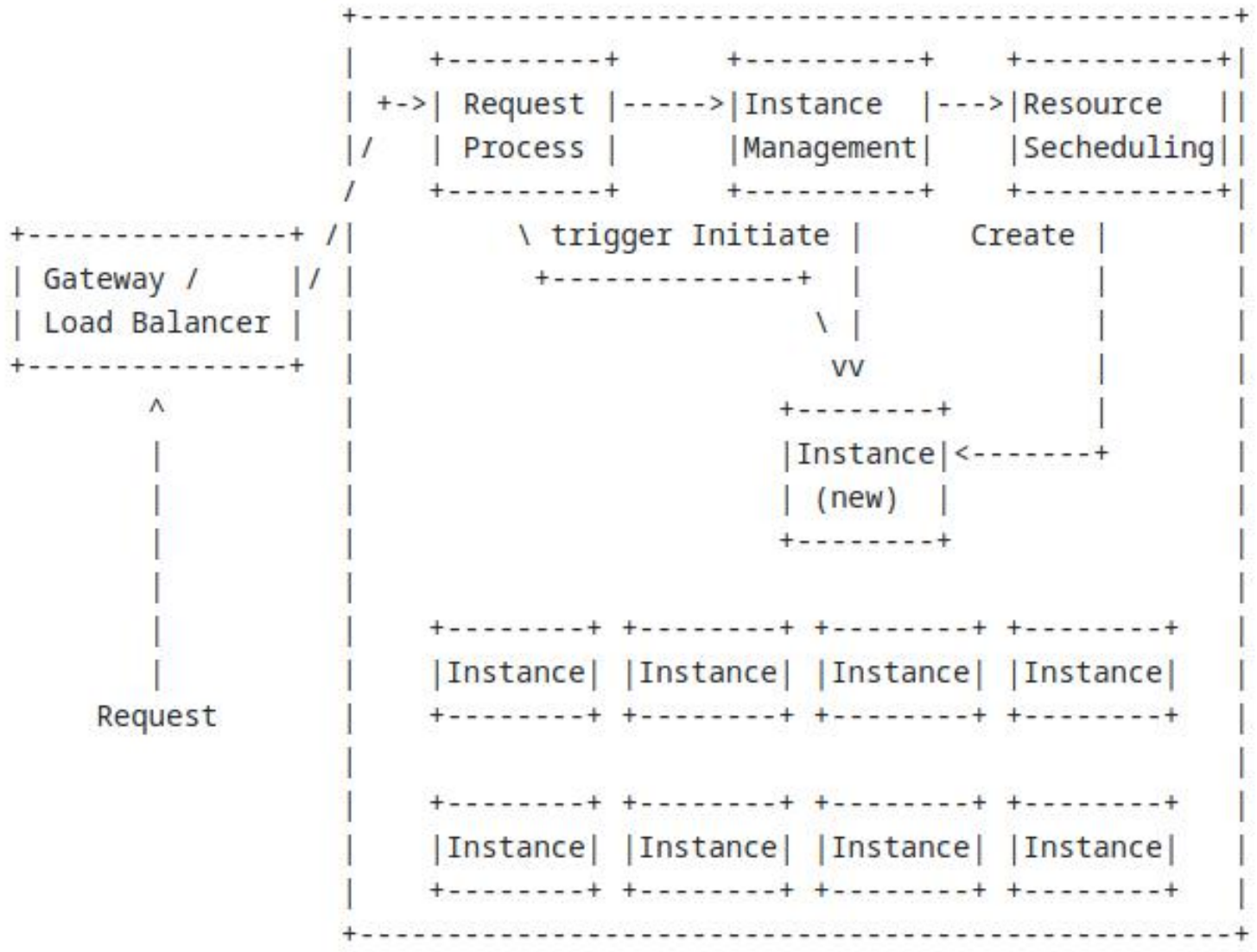
Network APIs,
“Traffic Lanes” , ...



1. The service endpoints are not aware of the resources and circumstances of network capabilities, and scheduling strategies tend to be static.

2. Collaborative logic across microservices and endpoints is not perceived and maintained by infrastructure.

Limited resources and burdening configs in edge site scheduling



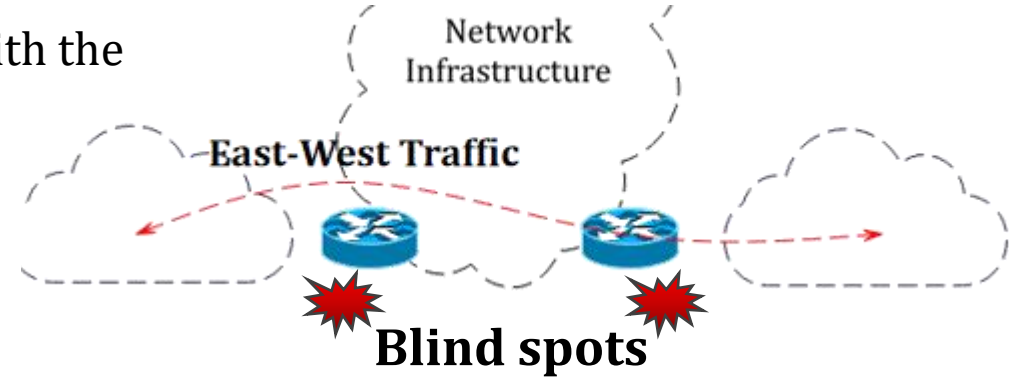
When a service request reaches the GW or load balancer of an edge cloud, a request process module queries the instance management module to determine whether there are available idle instances.
....

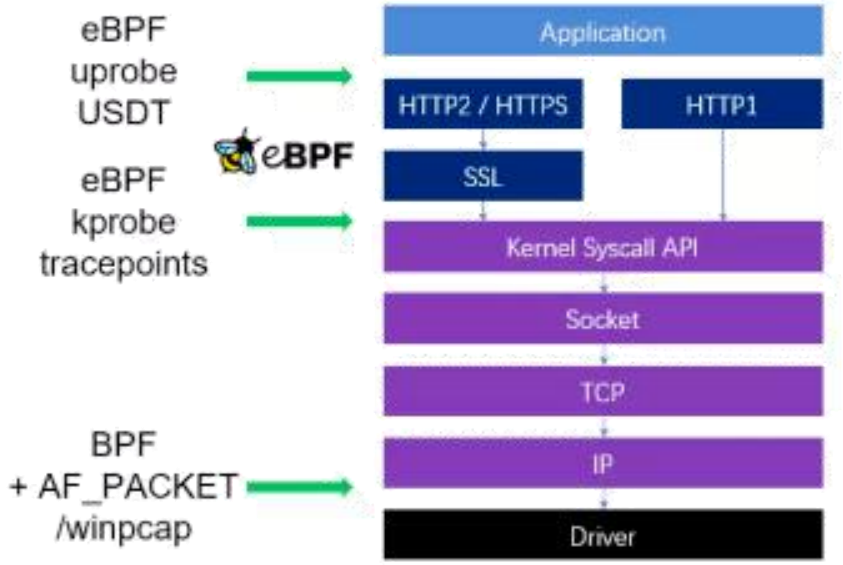
3. Difficulties and burdens in managing instances with dynamic scaling conditions in control plane.

Blind spots in APM and eBPF

Enhanced Berkeley Packet Filter (eBPF) technologies

- the byte streams collected by the eBPF technologies generally do not contain service semantics, it is difficult to conduct aggregation
- difficult to correlate a failure in the underlay network with the overlay tunnel and take action in a timely manner.



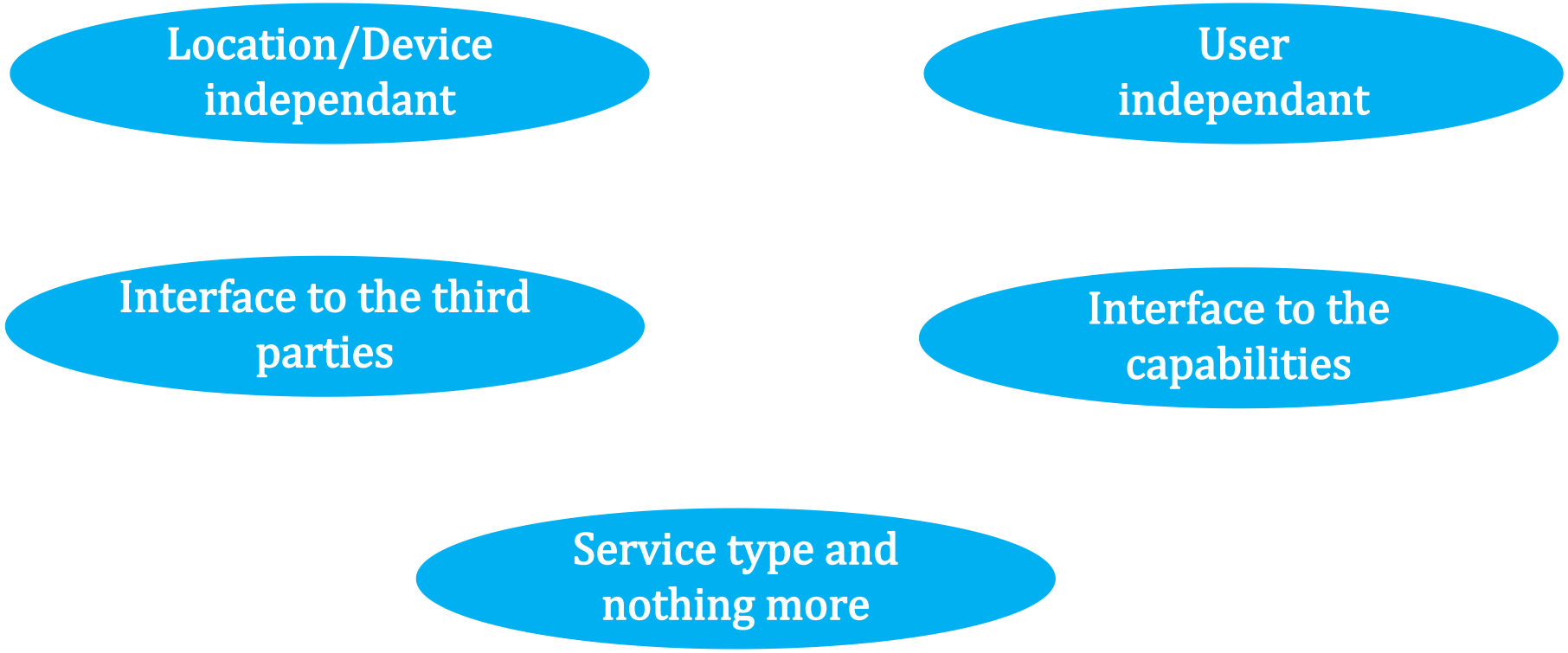


source: https://deepflow.io/blog/039-qcon-2022-in-depth-practice-of-cloud-native-observability-based-on-ebpf/

Application Performance Management (APM)

- On service-to-service communication paths, instrumention is either not allowed or leads to maintenance difficulties such as conflicts of codes
- Fail to obtain measurements regarding the infrastructure such as system calls and network transmissions

Requirements of standalone service ID in routing network



Terminology: we do not believe “service ID” is the best and it remains open for better suggestions.

Next Steps

- Refine the drafts upon comments and suggestions.
- More comments, suggestions and contributions would be welcome.