In-Network Computing for App-Centric Micro-Services

draft-sarathchandra-coin-appcentres-03

D. Trossen, C. Sarathchandra, M. Boniface

COIN RG IETF 109 @ 18.11.2020

Premise of the Draft

*"*The application-centric deployment of 'Internet' services has increased over the past ten years with many million applications providing user-centric services, executed on increasingly more powerful smartphones that are supported by Internet-based cloud services in distributed data centres, the latter mainly provided by large scale players such as Google, Amazon and alike. *This draft outlines a vision of evolving those data centres towards executing app-centric micro-services; we dub this evolved data centre as an* ***AppCentre****."*

- Draft outlines use cases and research challenges for this vision

General Structure

Still included but planned to be moved to new use cases draft (headlines already included there)

Still unclear if this will become ultimately part of joint use case (and requirements) draft but will likely update in next revision with clearer linkage to Section 5

Added more text in a number of sub-sections

5.3 Compute Inter-Connection at Layer2

- References to using L2 switching for interconnecting distributed compute resources
  - 5GLAN efforts in 3GPP
  - Edge computing in 5G

- Service routing critical in such distributed (L2) environment, similar to intra-DC service scheduling problem

5.5 Constraint-based Forwarding Decisions

- Evolved from previous ‘service pinning’ placeholder in V2 of draft
- Extends Section 5.4 discussion on service routing by including **constraints** into forwarding decision between one or more service instance candidates
- Load/latency may not be the only constraints
  - App/service-specific ones may be needed
- Matching operations in intermediary routers over such constraints may be coordinated across several routers to achieve **service scheduling** capability (across distributed compute resources)
- Section 5.4 already provides references to ongoing work, such as CFN-dyncast, ICNRG work etc. that include constraints in forwarding decision

5.6 Collective Communication

- Pattern exhibited in number of micro-service scenarios (outlined in Section 2 – use cases) is not just 1:1 but 1:N, N:1, N:M
- Patterns may be short lived with possibly as short-lived as single requests
- Solutions required for supporting such spontaneous formation of multipoint relations
- References to be added to ongoing work in this space (e.g., BIER)

Future Plans

- Move use cases into updated use case draft
  - > Use clear references in various remaining sub-sections
- Link requirements more clearly to Section 5 sub-sections
- Fill in missing sub-sections for Section 5
- More clearly link to other COIN drafts in relevant areas, e.g., computing frameworks, programmable forwarding nodes
- **Adopt as RG draft?**