

FACILITY FOR LARGE-SCALE ADAPTIVE MEDIA EXPERIMENTATION

Service design and development patterns for interactive edge computing experiences

Michael Boniface

(mjb@it-innovation.soton.ac.uk)

IT Innovation Centre, University of Southampton

Urban Hacking in 5G

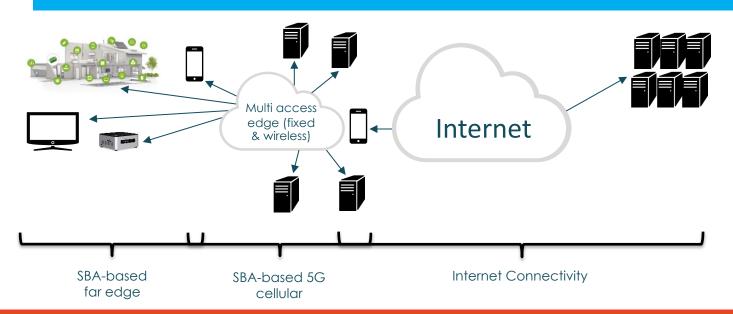
WWW.ICT-FLAME.EU 17/06/2019

Micro-Services From Far-Edge to Distant Cloud



Anything-as-a-Service (new interactive, immersive experiences, localized where possible)

Service-based architecture across all edge and the Internet



Well-proven Internet technology, such as web services, HTTP, IP, ... mixed with virtualization technology

Micro Services Architecture

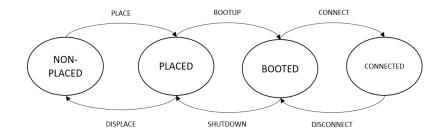


- Microservices architectures contain a collection of small selfcontained services typically implementing a single function
 - deployed independently, persist their own data/state, communicate through well-defined APIs, and may not share technology stacks
 - benefits include agility, resilience, scalability, etc.
 - challenges include complexity, testing, decentralised governance, etc.
- Microservices are old news but FLAME offers some distinct lifecycle management and control features that create useful patterns

Management and Control Features



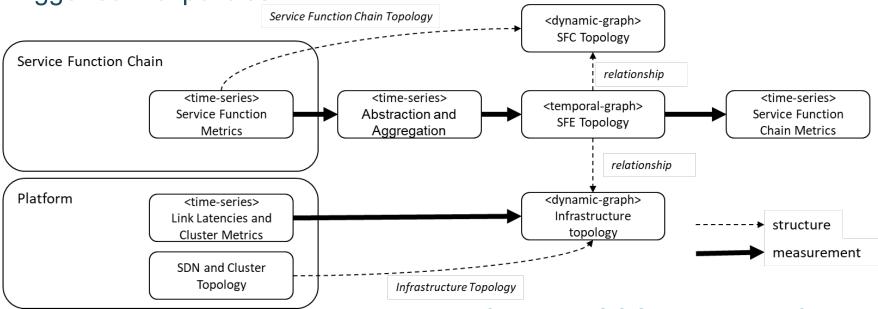
- FLAME uses policies to control the lifecycle state of service function endpoints
 - Placed Image deployed on cluster
 - Booted SFE booted on cluster
 - Connected SFE connected to network
- Defined in TOSCA resource specification



Monitoring and Alerting Features



FLAME uses time-series/ cross-layer graph monitoring and alerts to trigger control policies



Types of Scaling



- Scale Up (not FLAME)
 - moving to a larger instance or upgrading resources, typically traditional applications
 - E.g. increasing a server resource using OpenStack
- Scale Out (FLAME)
 - adding more instances to a service, system or application
 - E.g. internal load balancer such as Docker Swarm or Kubernetes
- Scale Geographically (FLAME)
 - Scaling a service to run in different geographical locations including mobile edge and other data centres
 - E.g. Triggers control states of SFEs (PLACED, BOOTED, CONNECTED) in specific compute nodes

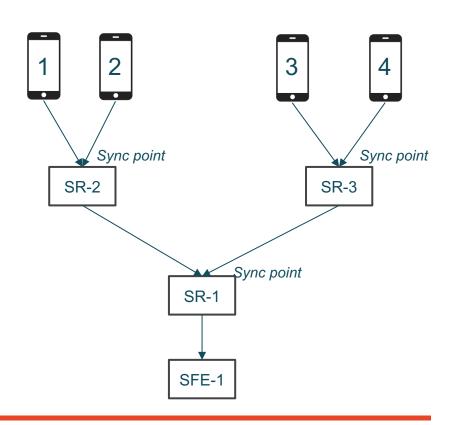
WWW.ICT-FLAME.EU 6



Sync'd Playout



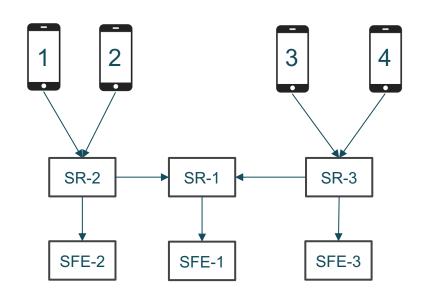
- High bandwidth video streaming (4K VR) to many clients from one playout point
- Provider value
 - Cost reduction (reduced AWS network usage) as video is front loaded to the edge
 - Cost reduction of server usage (http request suppression when multicast occurs)
- Experience value
 - Reduction in startup time for playout (avoids download)



Nearest Playout



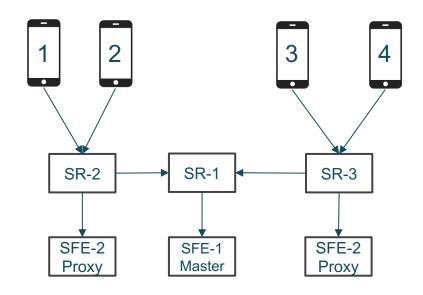
- Serve request from the closest service function
- Provider value
 - Cost reduction in network usage
 - Cost reduction since no replication required
- Experience value
 - Reduction in startup time for playout
 - Reaction time to latency changes
 - Content can be uploaded/placed anywhere and access anywhere
- Assumption the closest gives the best performance



Proxy Cache Playout



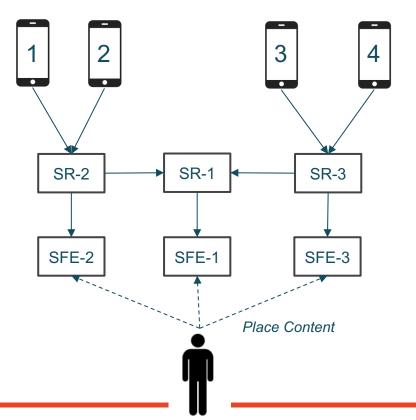
- Serve request from the closest service function proxy cache
- Provider value
 - Cost reduction in network usage
 - No preloading required
- Experience value
 - Reduction in startup time for playout
 - Reaction time to latency changes



Content Placement



- Place content within specific service function endpoint at a location within the network
- Provider value
 - Cost reduction in network usage
 - Preload for expected demand
- Experience value
 - Reduction in startup time for playout
 - Reaction time to latency changes

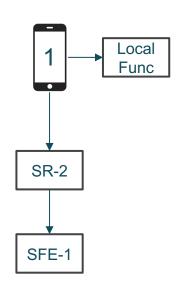


Application Function Offloading



12

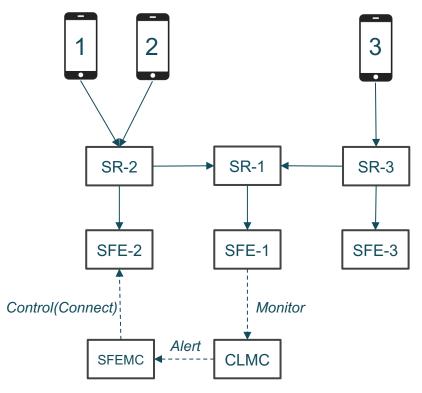
- Offload local terminal-centric functions to network elements
- Provider value
 - Terminal improvements (e.g., battery)
- Experience value
 - Utilize better device capabilities (e.g., better displays)



Scale Geographically



- Place content within specific service function endpoint at a location within the network
- Provider value
 - Cost reduction in network usage
 - Preload for expected demand
- Experience value
 - Reduction in startup time for playout
 - Reaction time to latency changes







This project received funding from the European Union's Horizonhas 2020 research and innovation programme under grant agreement No 731677



