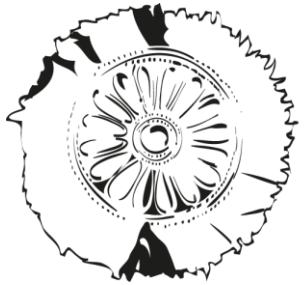
Exploring serverless service deployment in 5G for next generation media applications

Nikolaos Zioulis, nzioulis@iti.gr

(with Alexandros Doumanoglou & Petros Drakoulis, parts in cooperation with David Breitgand from IBM)



CERTH
CENTRE FOR RESEARCH & TECHNOLOGY HELLAS

Information Technologies Institute

VCL Visual Computing Lab
Information Technologies Institute

5G PPP 5G MEDIA

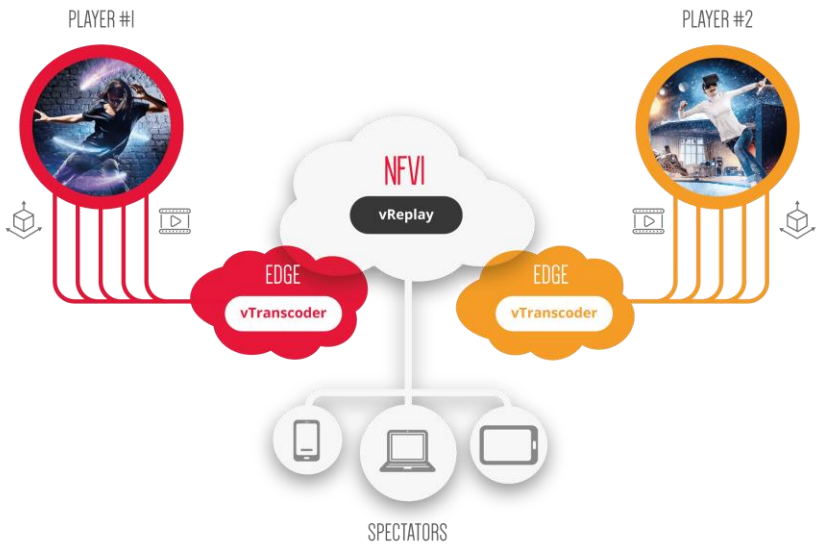
5G MEDIA H2020 Project

ENGINEERING
OTE GROUP OF COMPANIES
SingularLogic
IBM
rtve
UCL
Telefónica Investigación y Desarrollo
POLITÉCNICA
IRT NEXTWORKS ENGINEERING FORWARD
CERTH CENTRE FOR RESEARCH & TECHNOLOGY HELLAS
iti Information Technologies Institute
NETAS
InterInnov INTERNATIONAL INNOVATION
BitTubes
ΔΗΜΟΚΡΙΤΟΣ

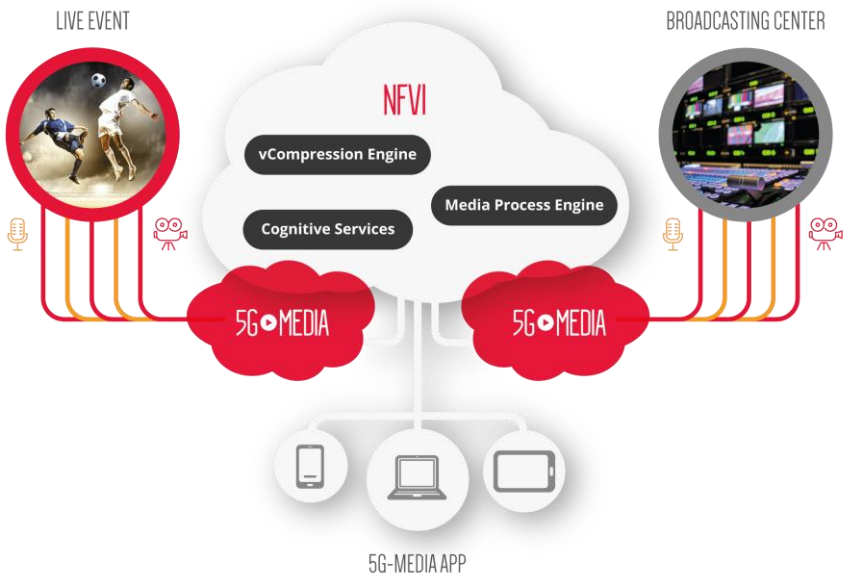
http://www.5gmedia.eu/

5G MEDIA H2020 Project

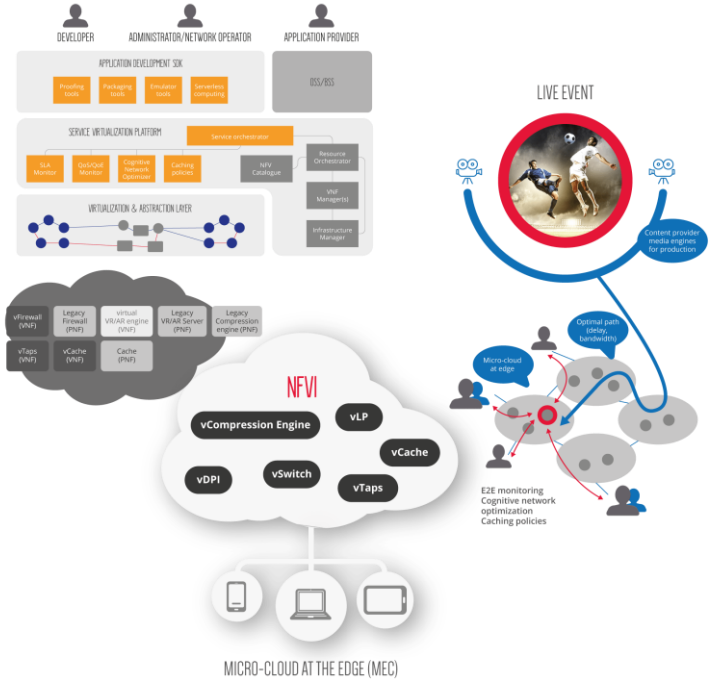
IMMERSIVE APPLICATIONS AND VIRTUAL REALITY



REMOTE AND SMART MEDIA PRODUCTION INCORPORATING USER GENERATED CONTENT



DYNAMIC AND FLEXIBLE UHD CONTENT DISTRIBUTION OVER 5G CDNS



5G MEDIA H2020 Project

CERTH
CENTRE FOR RESEARCH & TECHNOLOGY HELLAS

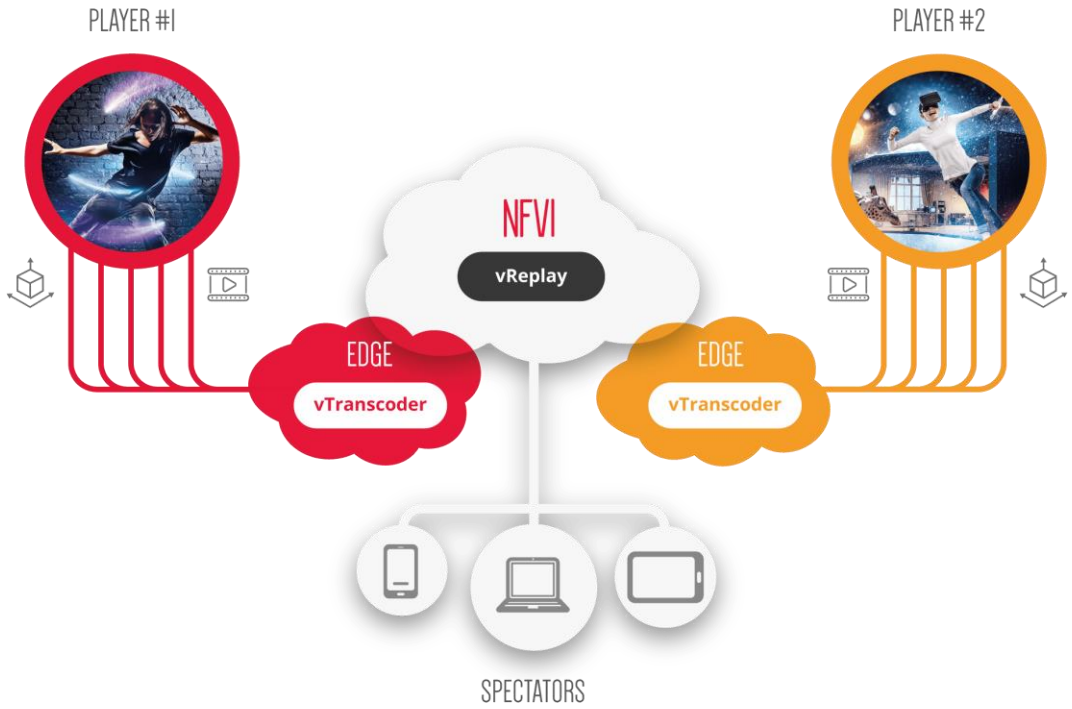
iti Information Technologies Institute

IBM

OTE GROUP OF COMPANIES

ΔΗΜΟΚΡΙΤΟΣ

IMMERSIVE APPLICATIONS AND VIRTUAL REALITY



Overview

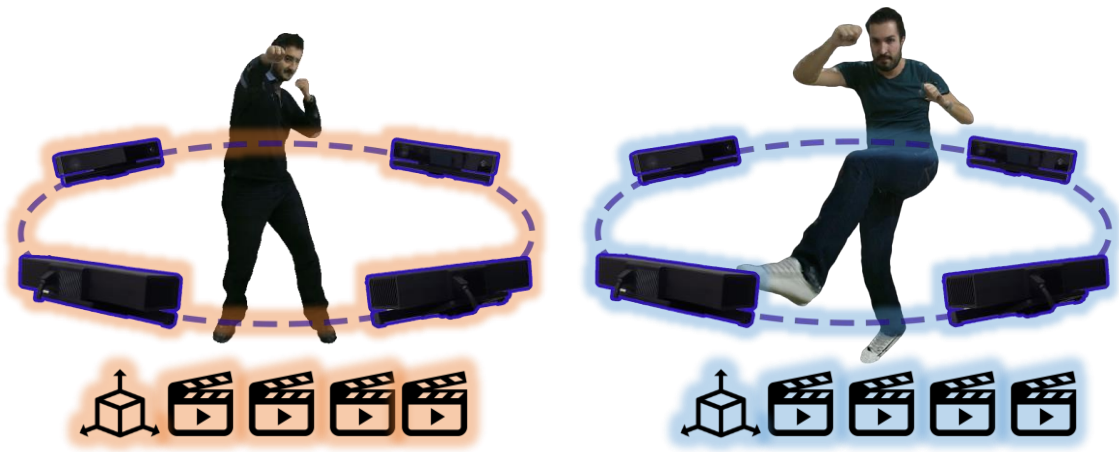
- New medium
- 5G
- Serverless
- Why Serverless?
- 5G-Media

Holograms

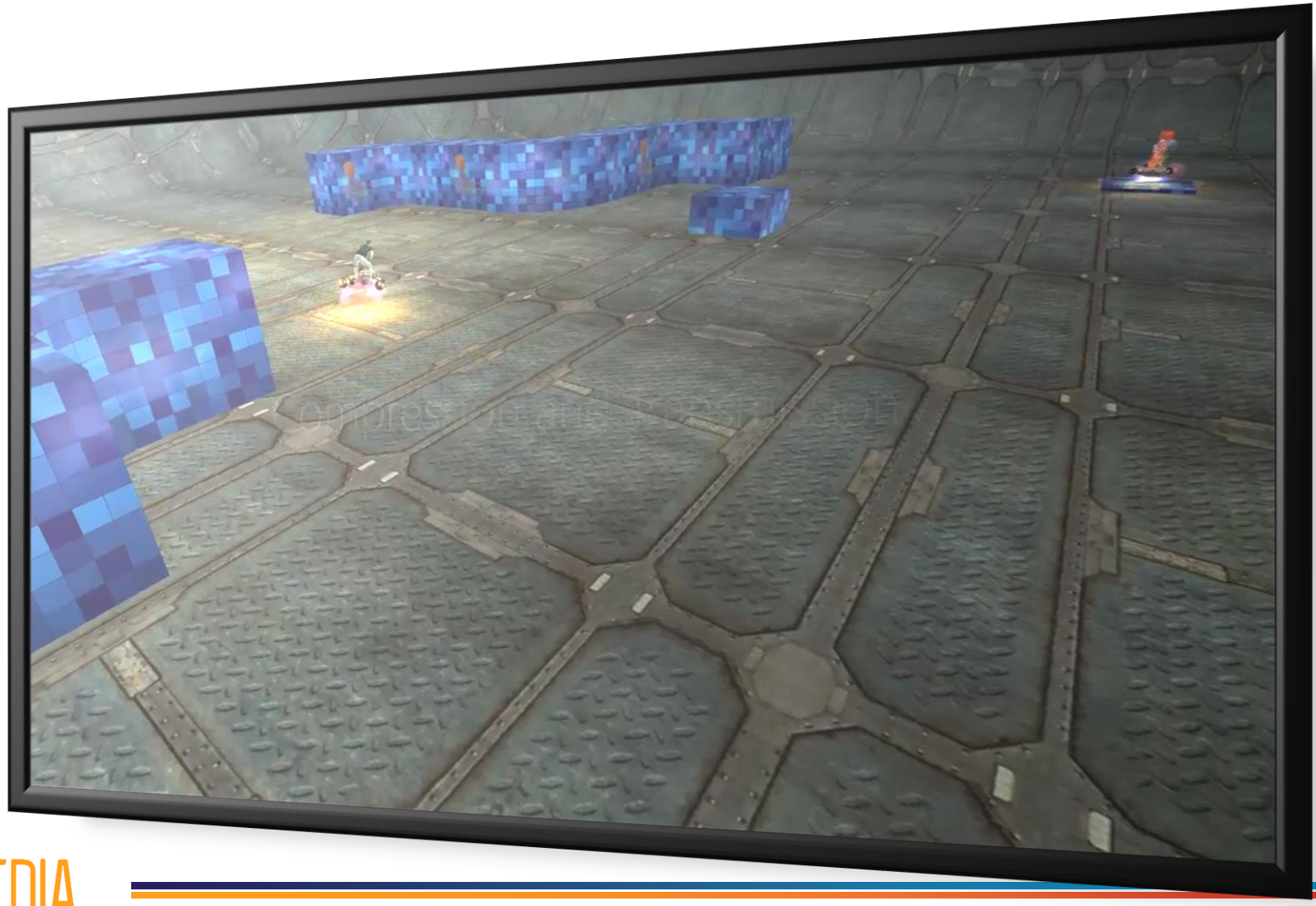
- Immersive Communications
- Approaching Physical Co-presence
- 3D Multimedia

Next Generation Immersive Media

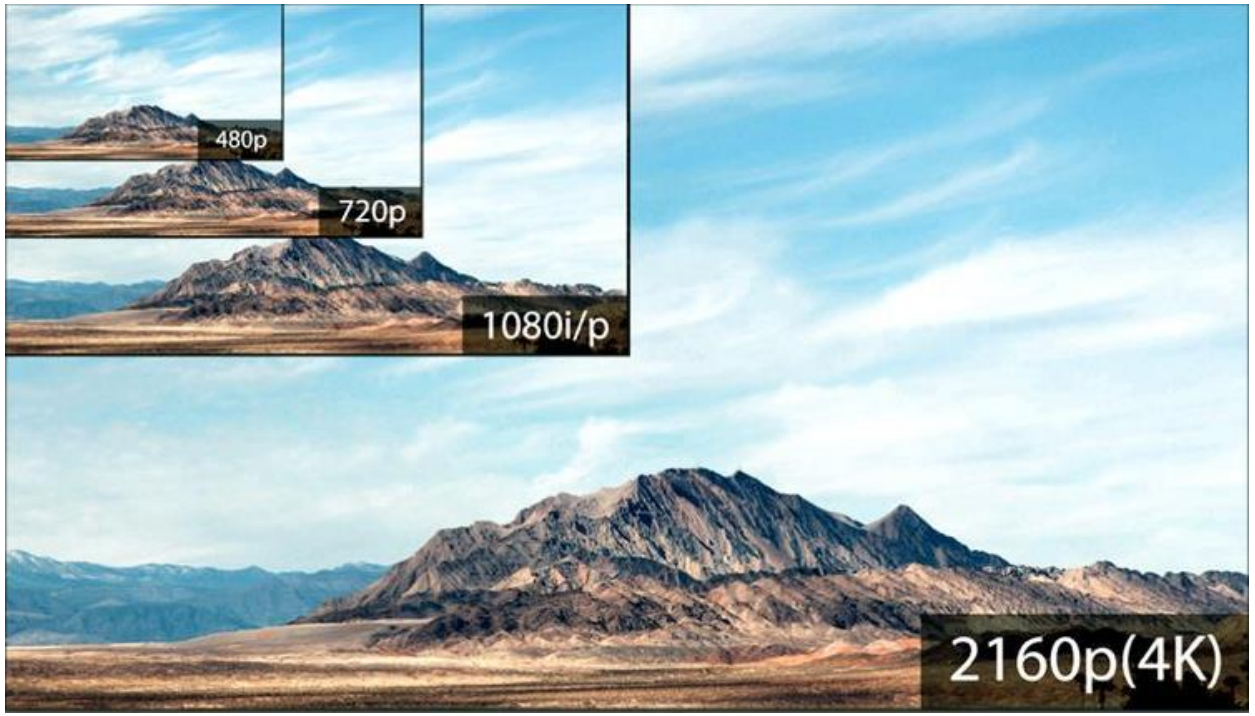
- Real-time 3D Capturing
- Live 3D Multimedia Stream
- Geometry & Texture Data
- Proof-of-concept 2 player game



Next Generation Immersive Media



New Medium ⇒ New Challenges



3.0 Megabits per second — Standard definition (SD) bandwidth

5.0 Megabits per second — High definition (HD) bandwidth

25 Megabits per second — Ultra HD (4K UHD) bandwidth

~20-40 Mbps per stream

New Medium ⇒ New Challenges

Requirements	Solutions
High Bandwidth Media	**Real-time** Adaptive Streaming
Interaction Latency	**Minimal** Transcoding Processing
Heterogeneous Consumption	**Multiple** Encodings

Next Generation Networks

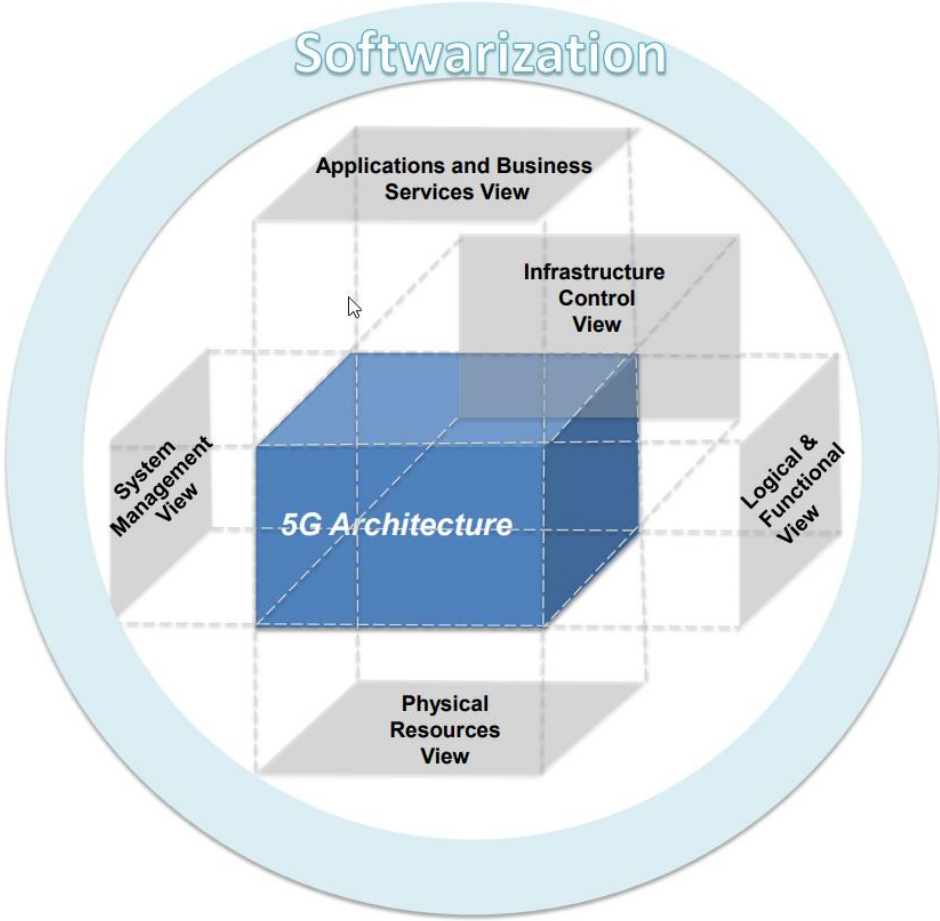
- New types of immersive experiences
- VR/AR 5G **eMBB** Applications
- 5G Technology
  - 5G New Radio
  - 5G NextGen

5G Next Generation Architecture

- *Software defined networking (SDN)*
- ***Network functions virtualisation (NFV)***
- *Network slicing*
- ***No magic bandwidth increase !***

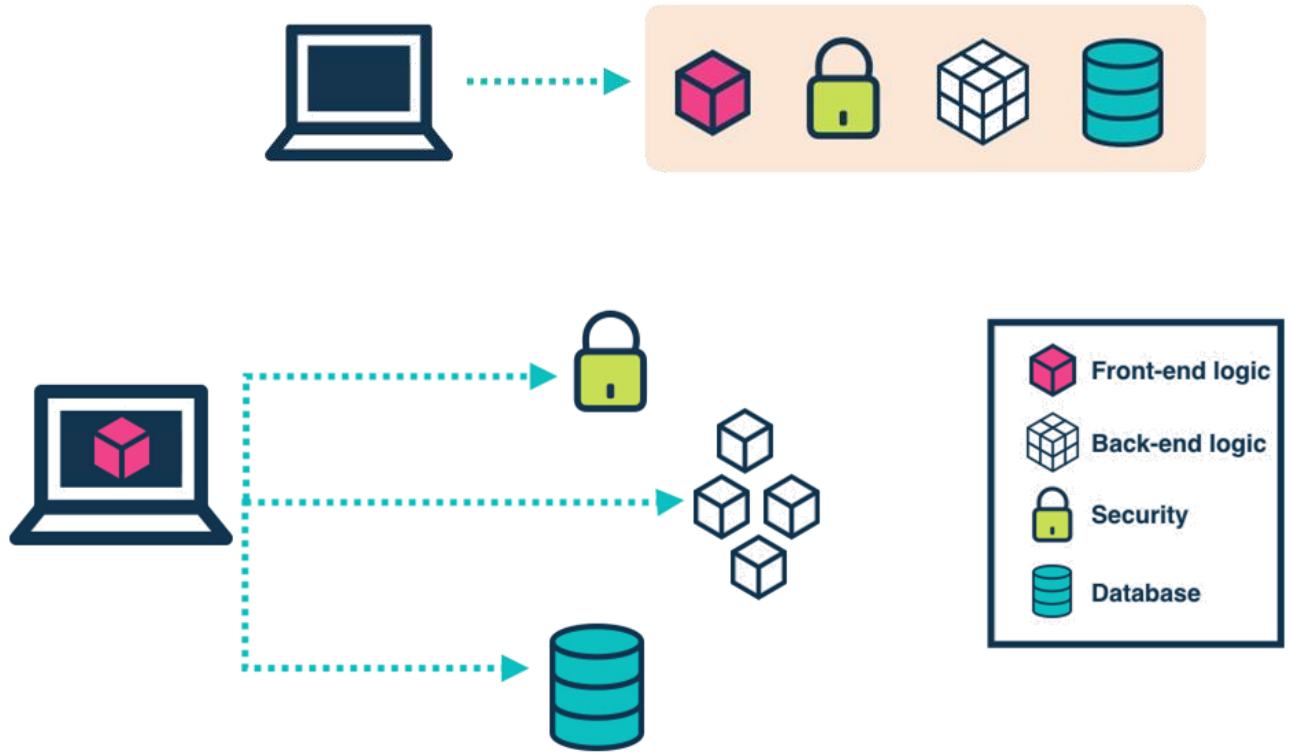
5G Next Generation Architecture

- ***Network functions virtualisation (NFV)***
  - *Network virtualization*
  - *Network programmability*
  - *Network management & orchestration*
  - *Edge computing*
- ***Service/Network Boundary Blurring***
  - *Application virtual functions*



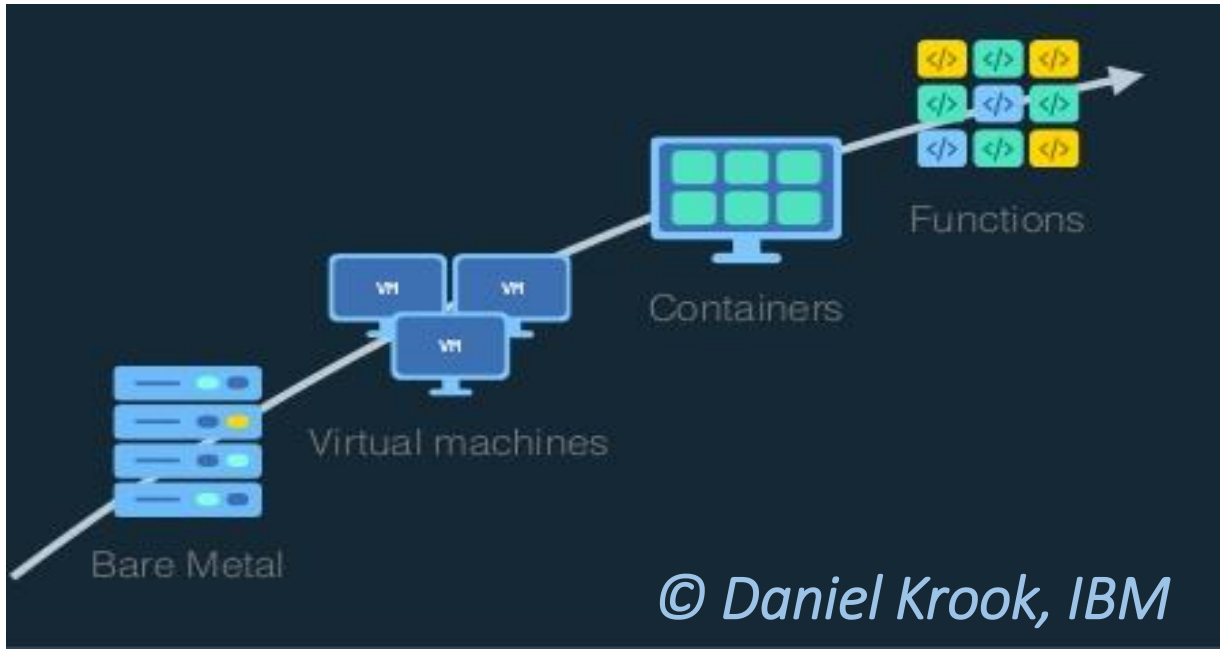
Serverless

- No Servers?
- Micro-services



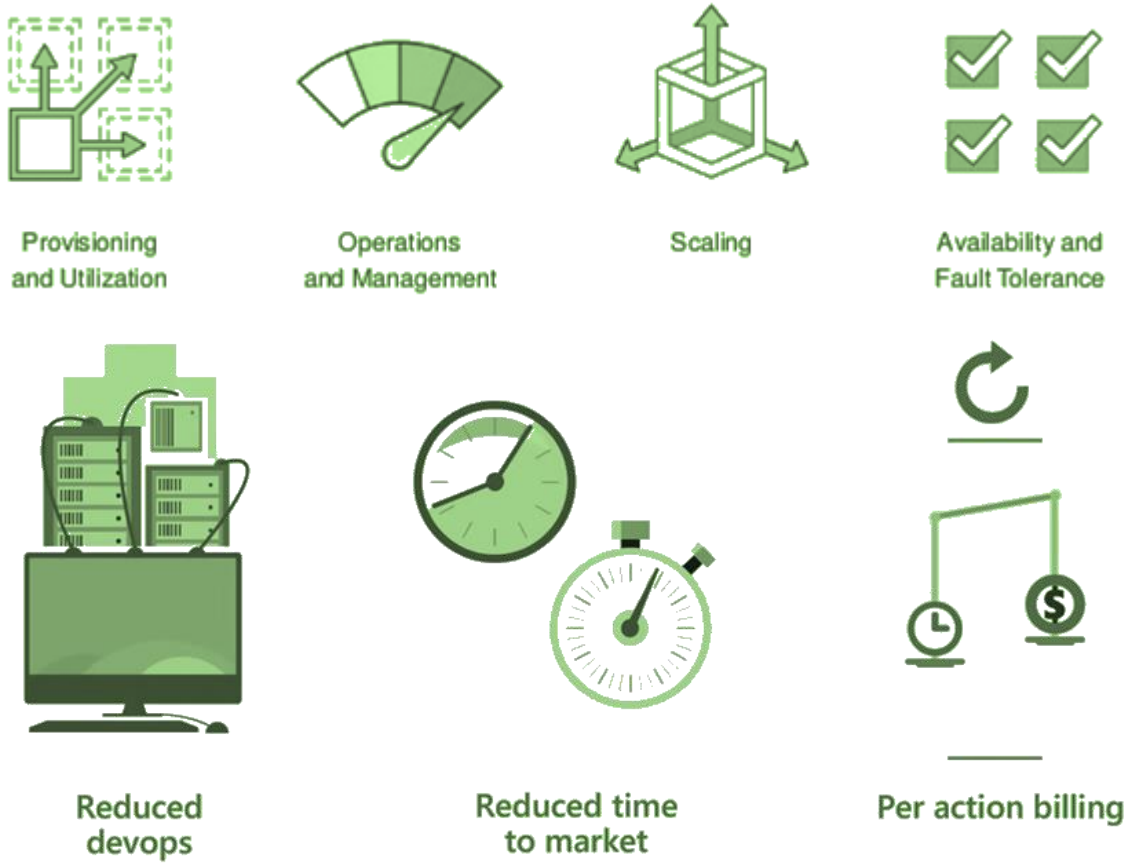
Function-as-a-Service (FaaS)

- No Servers?
- Micro-services
- Stateless compute – FaaS
- Developer Friendly Abstractions



New Business Models

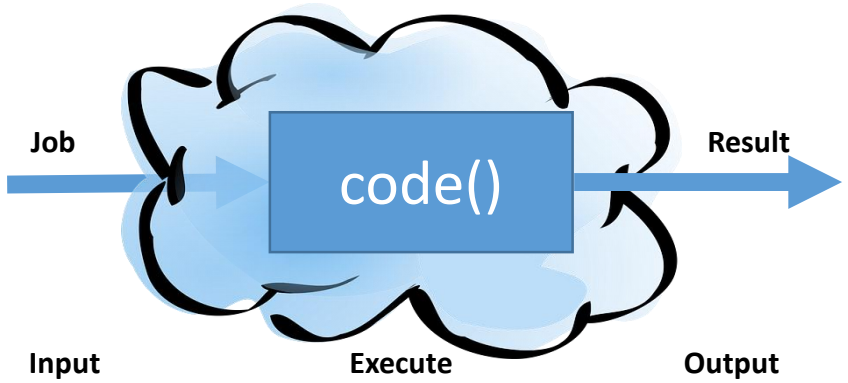
- Pay-per-use
- Zero Administration
- Deployment Elasticity
- Auto-scaling

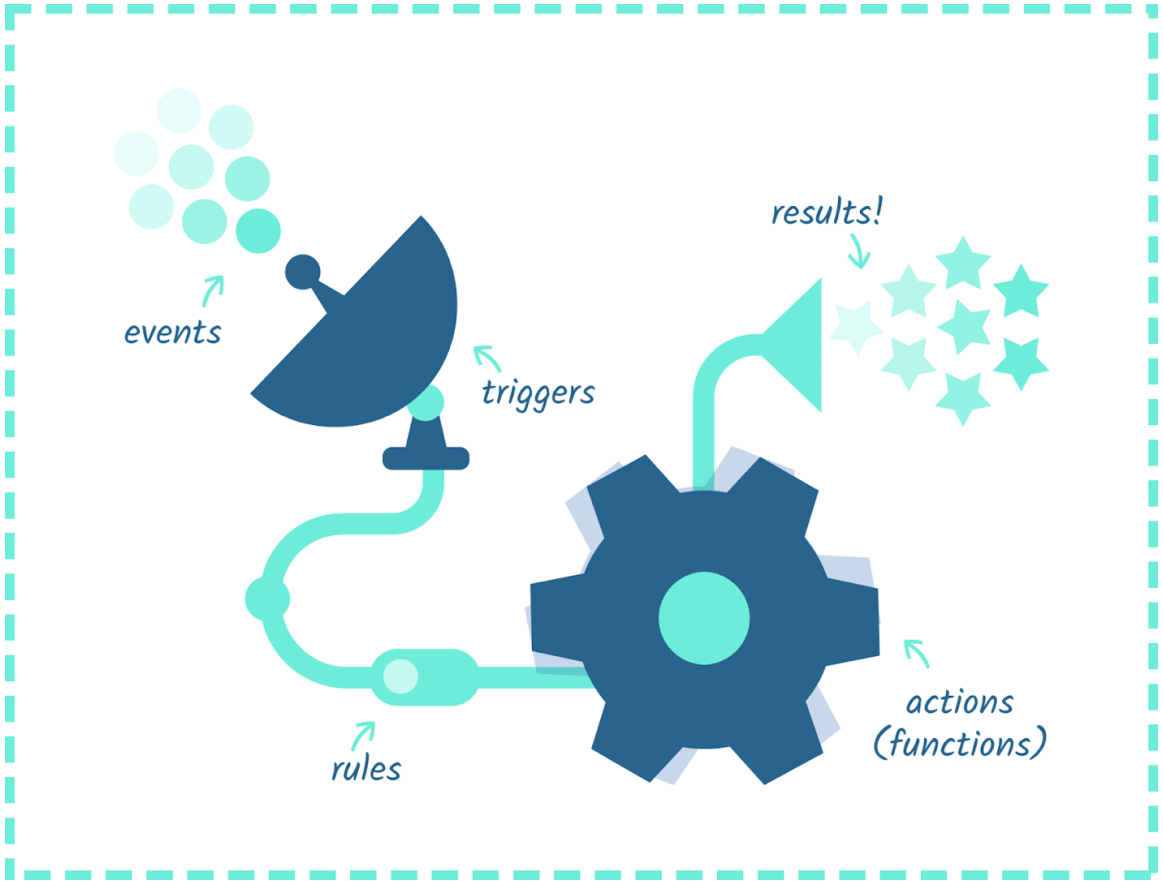


Apache OpenWhisk

APACHE OpenWhisk™

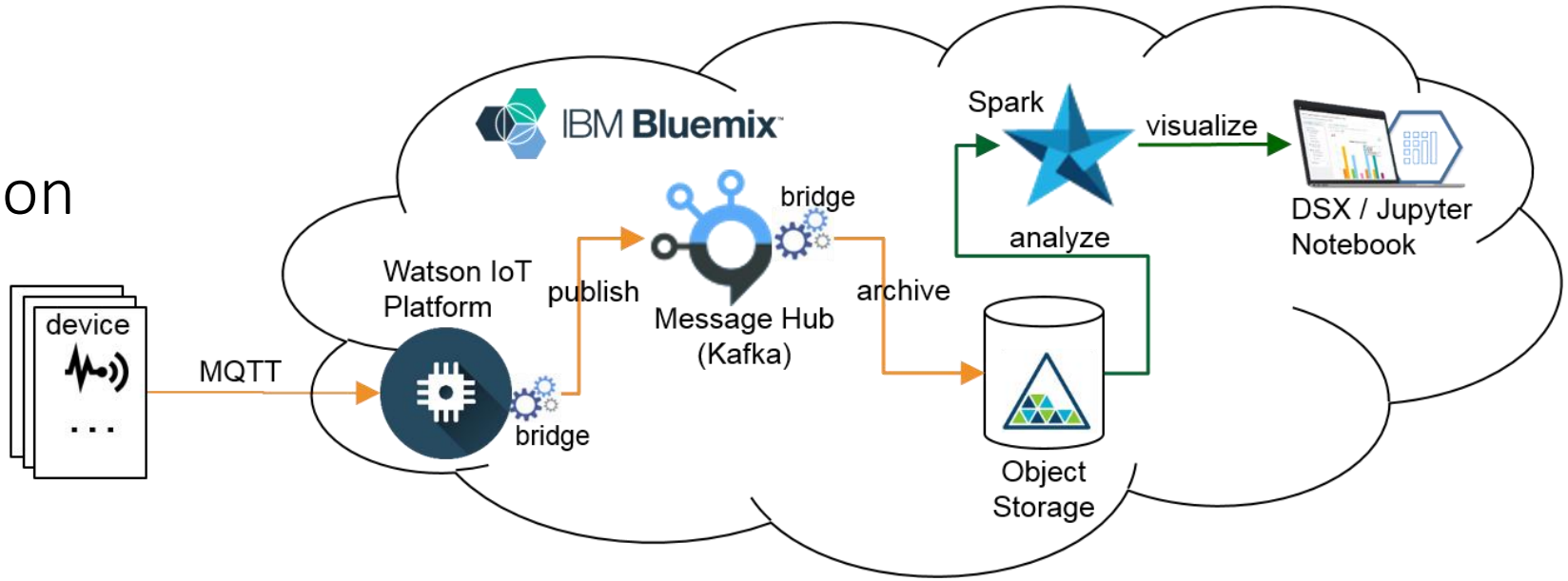
- A cloud-native programming model
- Geared for event-driven use cases
- Built-in autoscaling
- No administration/provisioning
- *"Think only about your code"*
- Billing @ $100ms$ resolution





(*Natural Fit for IoT*)

- Programmability
- Data transformation
- Event-based
  - Trigger
  - Rule



Server-based vs Serverless Streaming

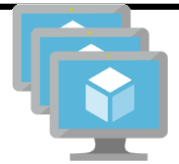
APACHE OpenWhisk™

- Calculate initial size

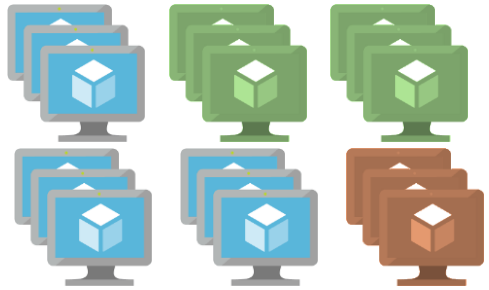
- Built-in Load Balancing

Server-based vs Serverless Streaming

- Calculate initial size



- Traffic ↑
- Traffic ↓

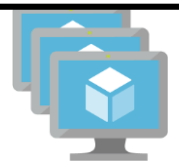


APACHE OpenWhisk™ docker

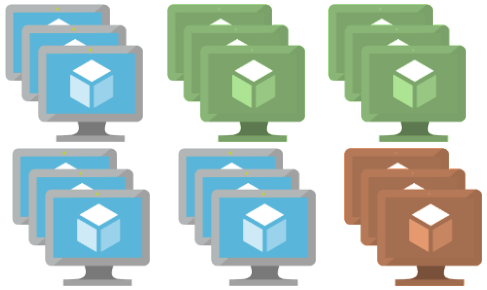
- Built-in Load Balancing
- Elastic Deployment
  - ✓On a per session basis
- Fast Deployment

Server-based vs Serverless Streaming

- Calculate initial size



- Traffic ↑
- Traffic ↓



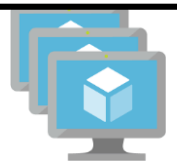
- Latency ↓ ⇔ Costs ↑

APACHE OpenWhisk™ docker

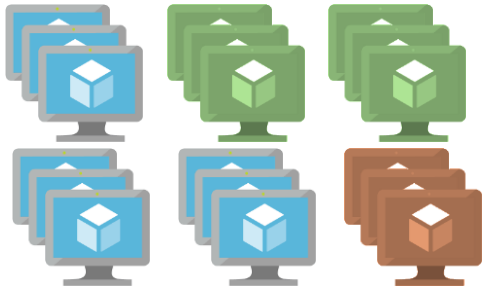
- Built-in Load Balancing
- Elastic Deployment
  - ✓On a per session basis
- Fast Deployment
- Run on-demand
  - ✓Better Resource Utilization
  - ✓Costs ↓

Server-based vs Serverless Streaming

- Calculate initial size



- Traffic ↑



- Traffic ↓
- Latency ↓ ⇔ Costs ↑

Capacity Engineering
Constant Sizing Problem

APACHE OpenWhisk™ docker

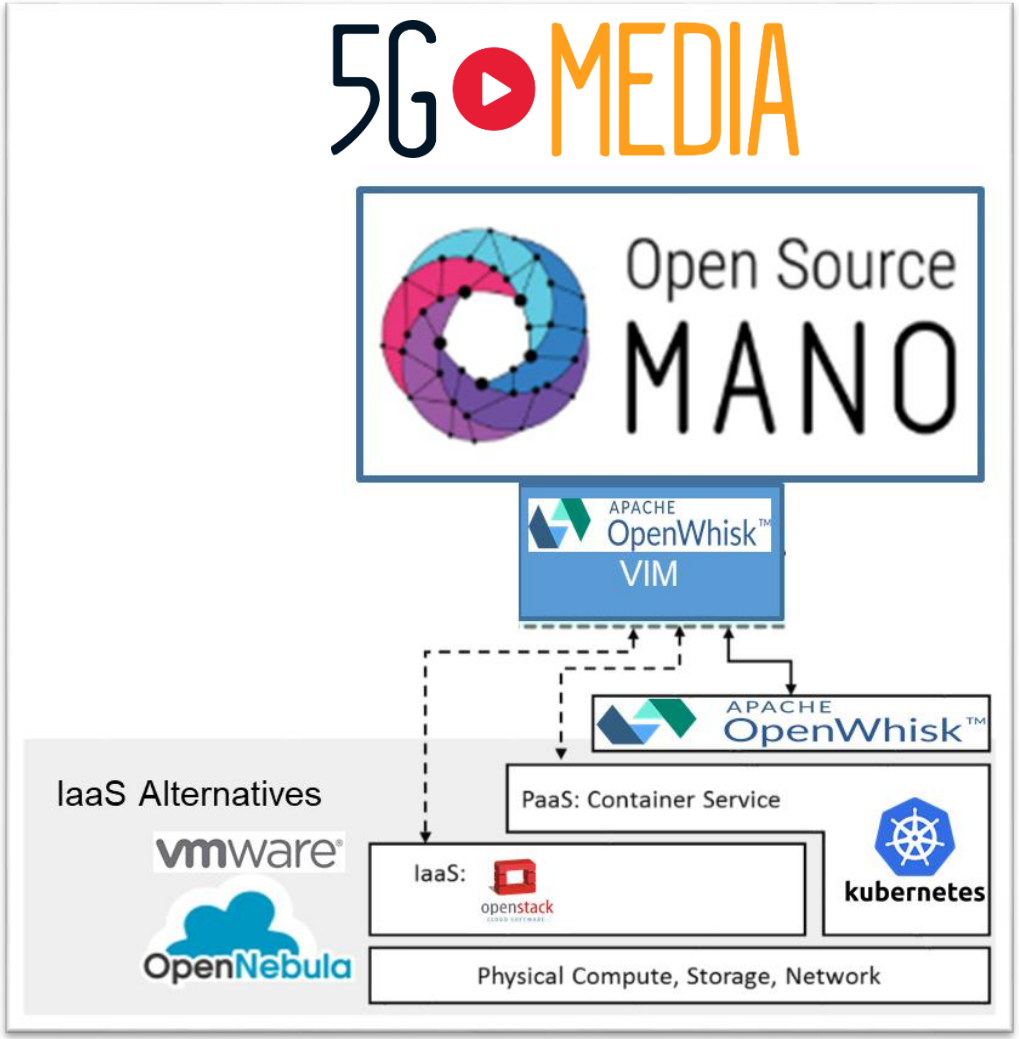
- Built-in Load Balancing
- Elastic Deployment
  - ✓On a per session basis
- Fast Deployment
- Run on-demand
  - ✓Better Resource Utilization
  - ✓Costs ↓

Development Flexibility

FaaS plugged into 5G

- On-demand Instantiation
- Co-existence with VMs
- ETSI MANO compatible
- PaaS and IaaS neutral

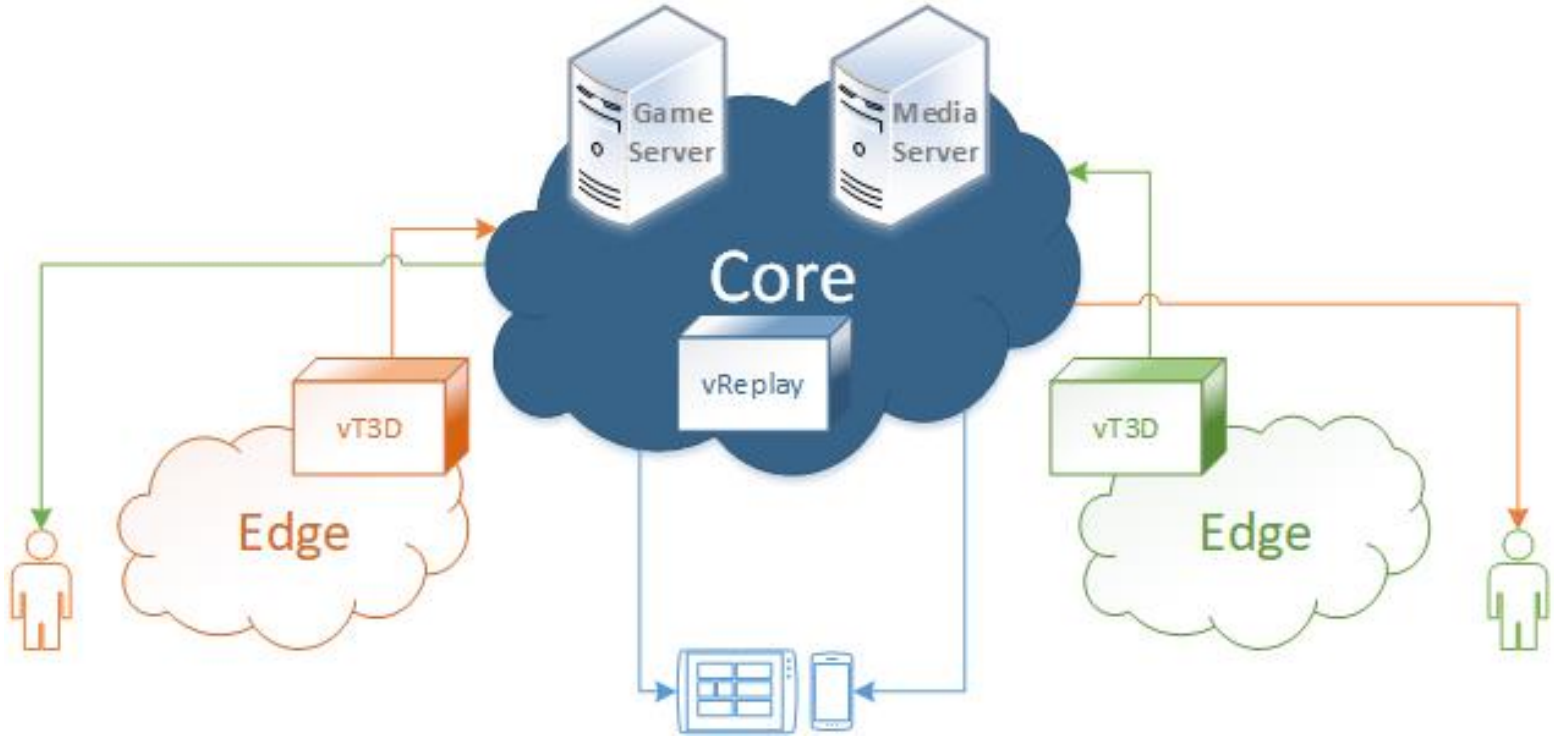
- **Basic flow**
  1) OSM initiates
  2) FaaS VIM invokes VNF as an OpenWhisk action
  3) OpenWhisk offloads to K8s
  4) K8s provides networking and placement



Serverless Media Streaming

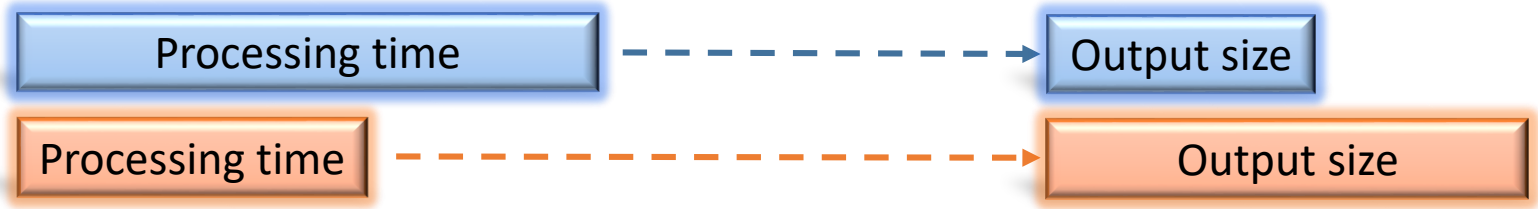
- **2** Playing Users
- $N >> 2$ Spectating Users

- **Micro**-transcoders
  - Per user
  - Per session
  - On the edge
  - No communication



Serverless Media Streaming

- Real-time Transcoding



- GPGPU
  - Nvidia k8s plugin
  - Minimize Processing Latency
  - Produce Multiple Qualities Simultaneously

- Optimization
  - **Location** (edge, regional or central clouds)
  - **Computational capabilities** (e.g. availability of GPUs)
  - **Cost** (edge nodes incur higher costs due to **low resources availability of** & **higher demand**)

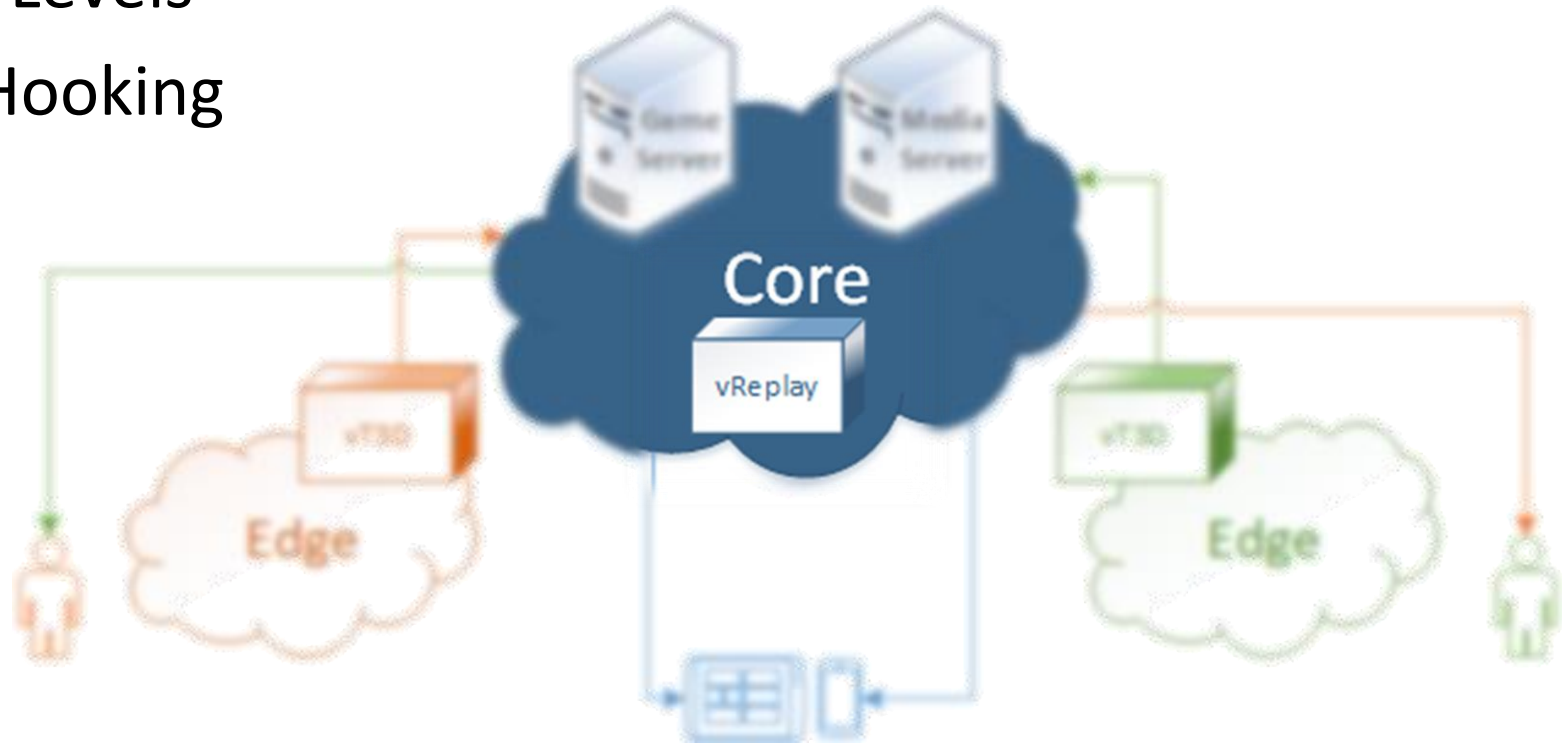
OW vs k8s

- Deployment time tests
- ~1% overhead
- 30 experiments
- Well behaved tests

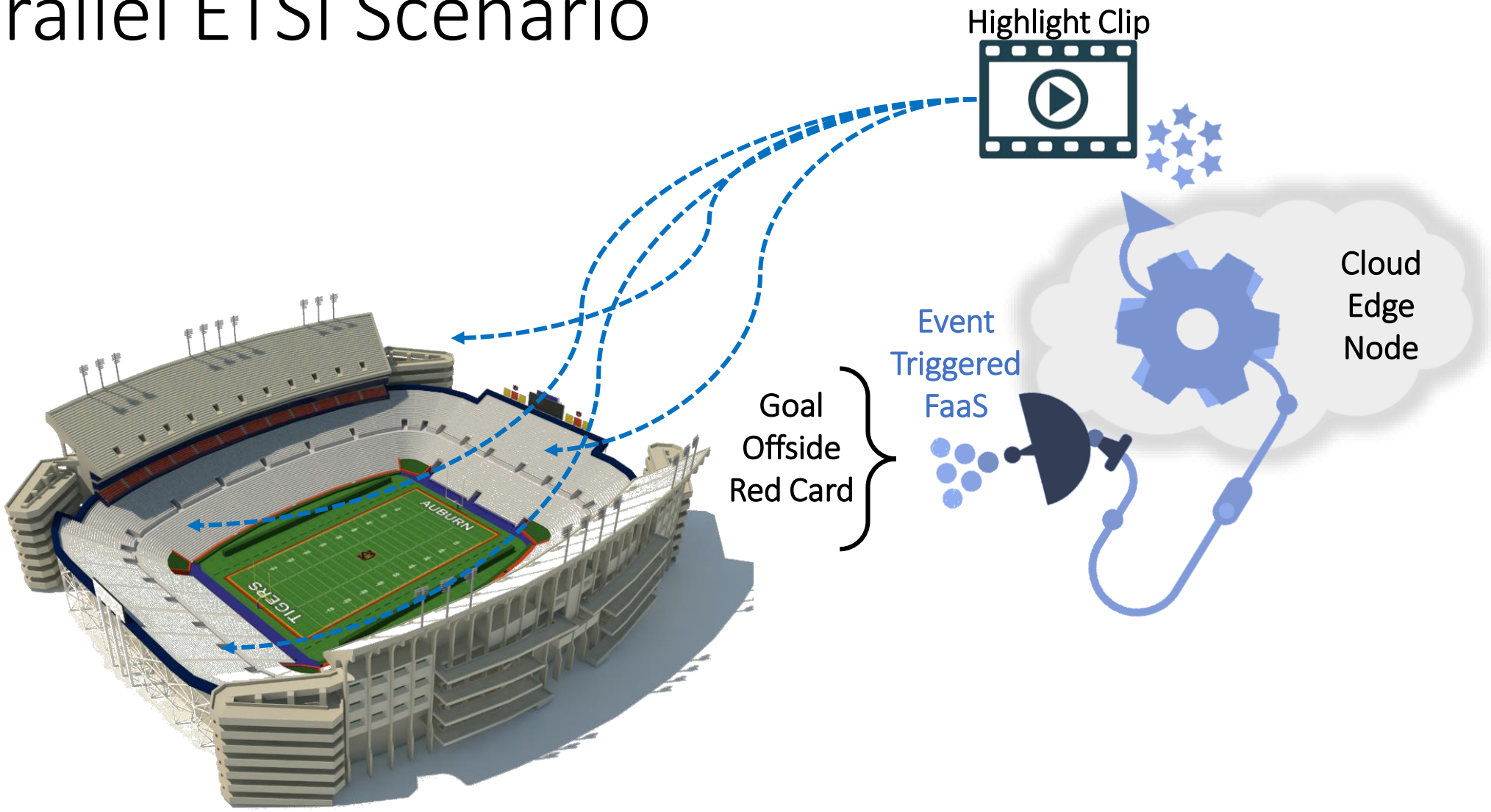
Statistics	k8s	OW
Mean	7.07s	7.8s
Confidence Interval for mean (at 0.05)	±0.2045s	±0.2778s
Minimum	6.492s	6.351s
Maximum	8.161s	8.868s
Standard Deviation	0.495s	0.673s
Kurtosis	0.044	-0.558
Skewness	1.07	-0.445

Event-based FaaS - vReplay

- Application Level Triggering
- Unknown Occurrence Levels
- Third-party Function Hooking



Parallel ETSI Scenario




danke
ありがとう
teşekkur ederim
спасибо
thank you
gracias
dziękuje
go raibh maith agat
obrigado
sukriya
arigato
merci
ευχαριστώ