Samsung Research

VERGE: AI at the edge for 6G networks

June, 2023

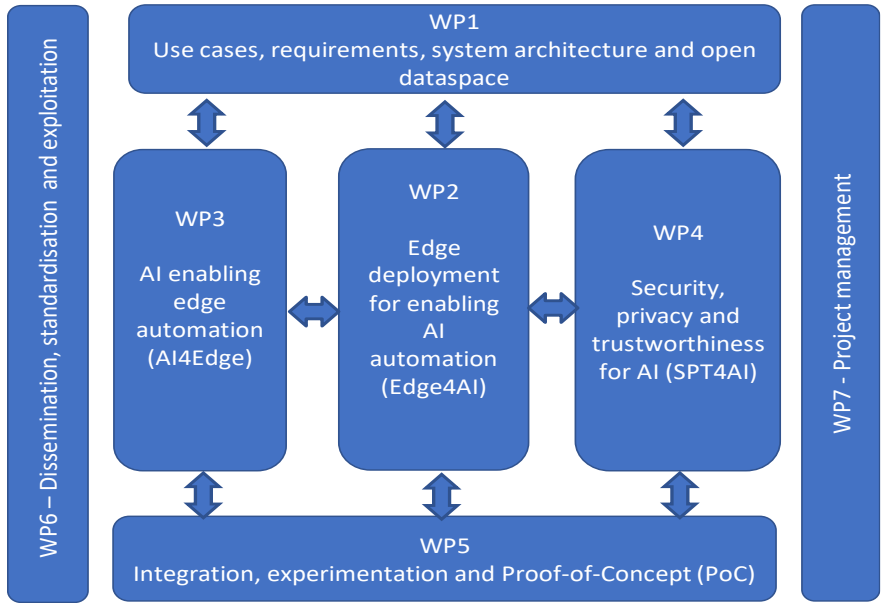
Dr. Joan Pujol-Roig
j.pujolroig@Samsung.com

Introduction

VERGE: AI-powered eVolution towards opEn and secuRe edGe architEctures

HORIZON-JU-SNS-2022

Participant organisation name
UNIVERSITAT POLITECNICA DE CATALUNYA
CENTRE TECNOLOGIC DE TELECOMUNICACIONS DE CATALUNYA
KING'S COLLEGE LONDON
OULUN YLIOPISTO
BARCELONA SUPERCOMPUTING CENTER-CENTRO N ACIONAL DE SUPERCOMPUTACION
SAMSUNG ELECTRONICS (UK) LIMITED
INTEL DEUTSCHLAND GMBH
ERICSSON ARASTIRMA GELISTIRME VE BILISIM HIZM ETLERI ANONIM SIRKETI
ERICSSON AB
TURKCELL TEKNOLOJI ARASTIRMA VE GELISTIRME A NONIM SIRKETI
NEARBY COMPUTING SL
HIRO MICRODATACENTERS B.V.
THALES ITALIA SPA
ARCELIK A.S.
HOLO-INDUSTRIE 4.0 SOFTWARE GMBH



© 2023. Samsung Research All rights reserved.

AI at the Edge

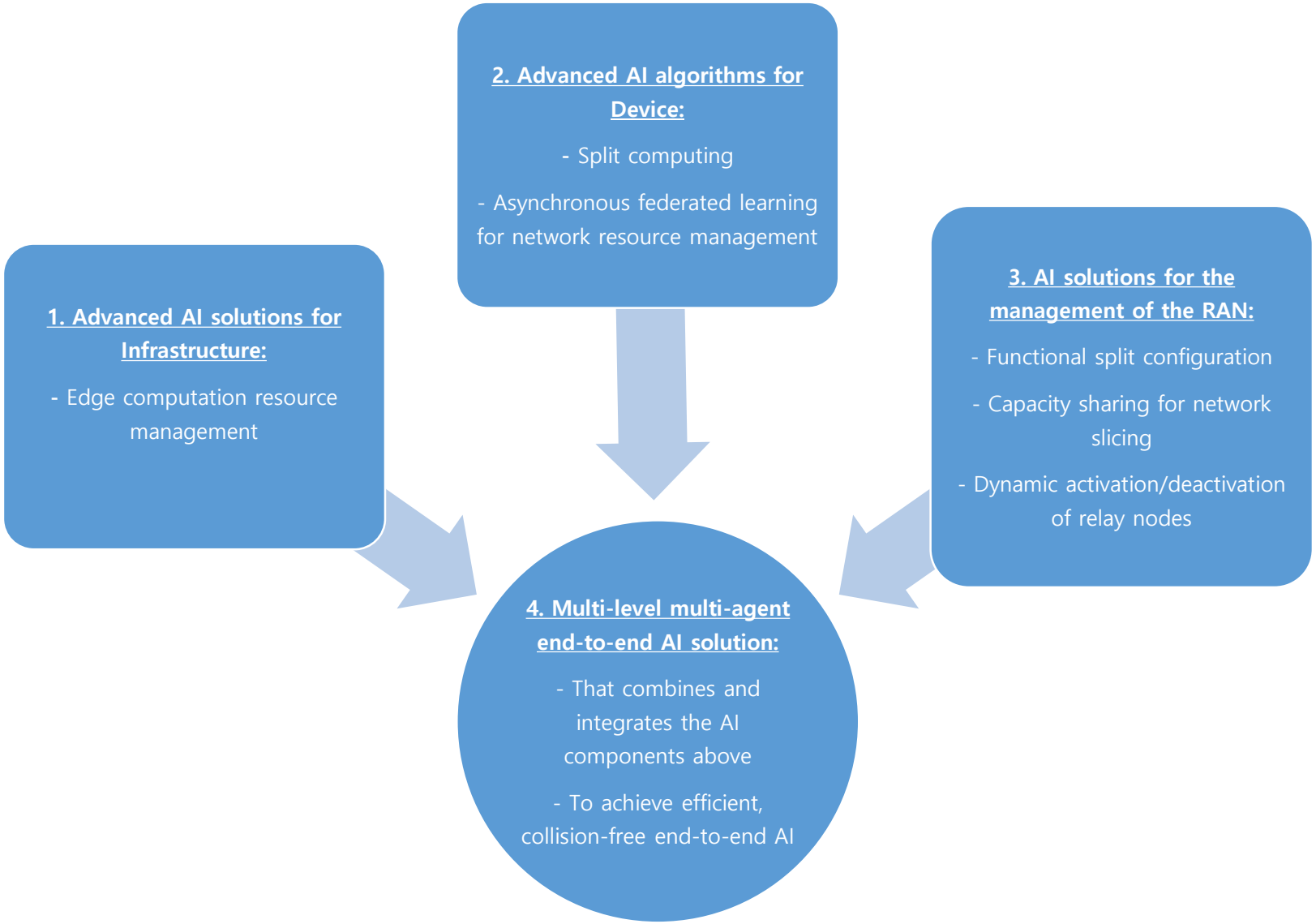
- AI at the Edge is a technology paradigm that brings computation, data storage, and AI capabilities directly to the edge or UEs where data is generated, away from centralized cloud data centres.

- Benefits:
  - Reduced Latency: Edge AI drastically reduces the time taken to transport data.
  - Enhanced Privacy: By processing data locally on the device itself, Edge AI significantly enhances data privacy
  - Cost Savings: The volume of data transferred to the cloud is drastically reduced, reducing transmission and storage costs.

- Real-world examples:
  - Self-Driving Car: Decisions need to be made in real-time for safety.
  - XR/VR: Real-time data processing for immersive experiences.
  - Healthcare: Wearable devices can employ edge AI to monitor health. By processing data locally, these devices can provide r eal-time feedback and alerts, ensuring quicker response times.

© 2023. Samsung Research All rights reserved.

VERGE Objectives



© 2023. Samsung Research All rights reserved.

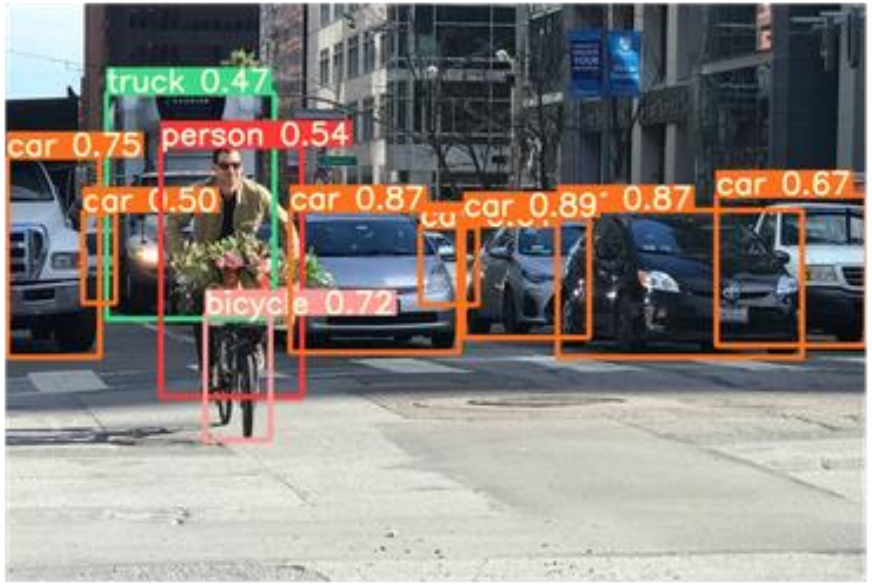
Advanced AI solutions for Devices

© 2023. Samsung Research All rights reserved.

Advanced AI algorithms for Device: Split computing

Overview:

- Soaring use of complex deep neural networks in mobile applications.
- XR/VR, mobile hologram, digital replicas… require extensive computation capabilities for a real-time immersive user experience.
- Expected increase of model complexity for 6G solutions.
- Inference on mobile has been elusive due to the excessive computational demands of modern DNNs.
- Popular alternative → offloading processing to powerful cloud-based solutions:



Example: Yolo v5 Image segmentation

© 2023. Samsung Research All rights reserved.

Advanced AI algorithms for Device: Split computing

Computation offloading:

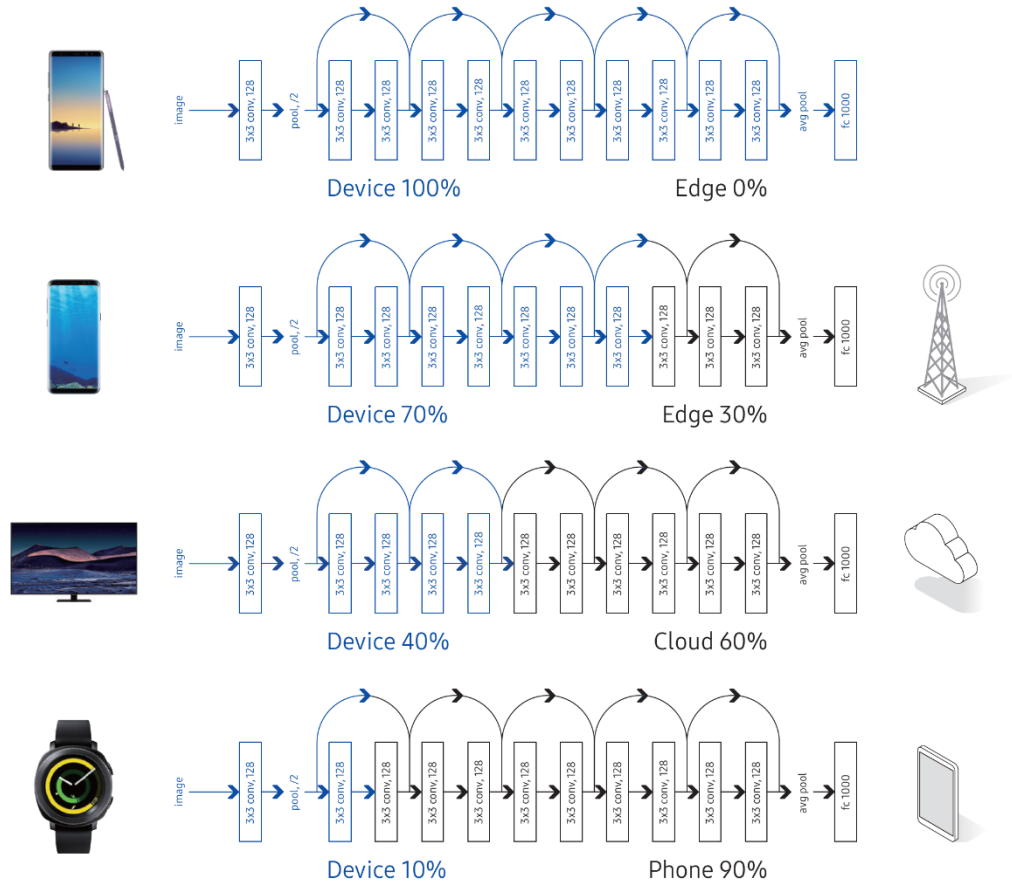
- Issue→ suffer from the dynamic connectivity conditions and the uncertain availability of the cloud which impacts:

  1. Latency critical applications

  2. High-mobility applications

Proposed solution: AI-based split computing decision block-> optimal balance between performance and latency.

Goal: General Framework for dynamic progressive inference.



Source: Samsung Research 6G vision paper

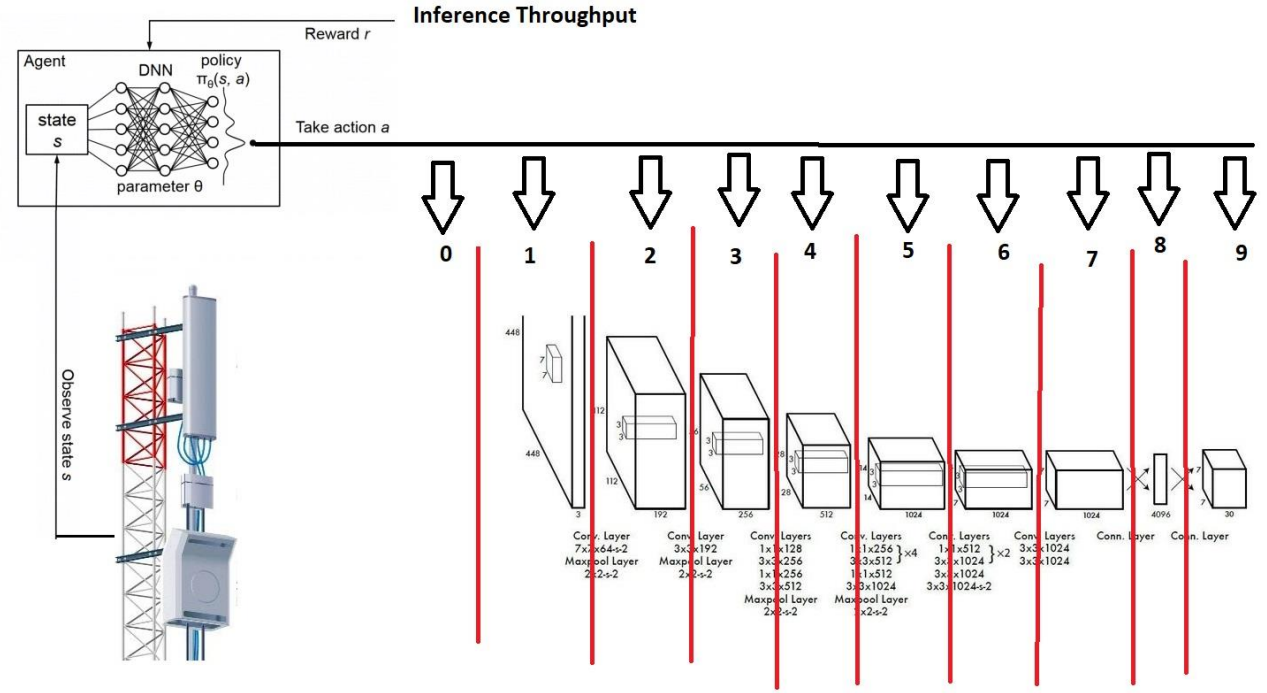
© 2023. Samsung Research All rights reserved.

Advanced AI algorithms for Device: Split computing

Goal:

- Implement Progressive inference for Deep neural networks.
- The AI solution splits any DNN model according to UE, network and MEC server conditions.

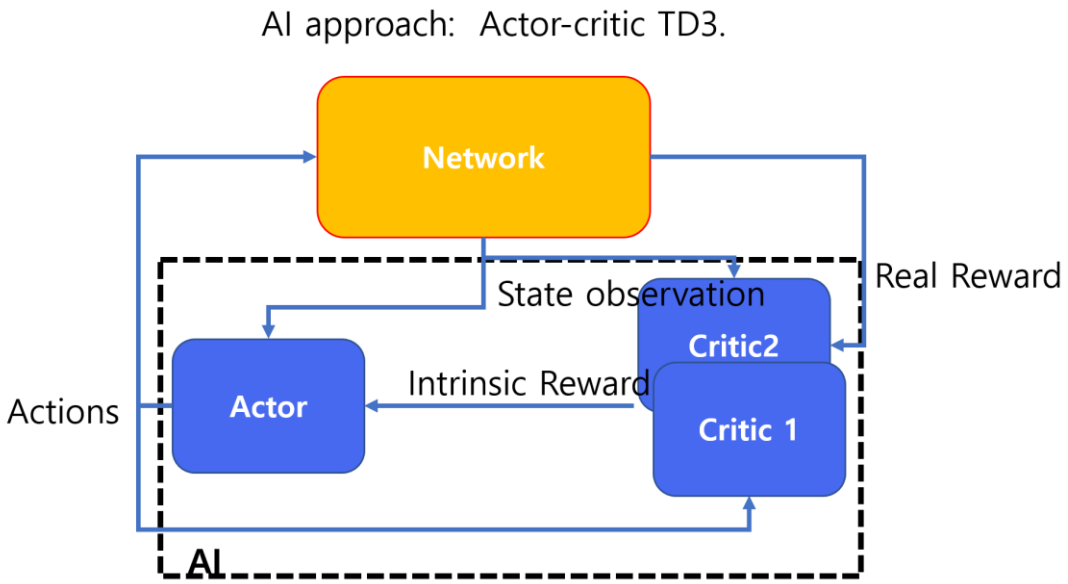
- To do so it:
  1. Identifies all candidates splitting points in the model.
  2. Estimates performance and latency on the client and server.
  3. obtains these initial timings along with the target SLAs and run-time conditions and decides on the optimal splitting point.



© 2023. Samsung Research All rights reserved.

Advanced AI algorithms for Device: Split computing

Developed solution: Dynamic profiling of DNNs and inference splitting using Deep Reinforcement Learning.



© 2023. Samsung Research All rights reserved.

Advanced AI solutions for Infrastructure

© 2023. Samsung Research All rights reserved.

Advanced AI algorithms for Infrastructure: Autoscaling

Samsung Research

Background & problem statement:

- 5G/B5G Networks are evolving to a containerized micro-service based infrastructure that allow dynamic scaling of the network
- Container orchestration relying on existing threshold-based autoscaling solutions, which does not always adapt to the requirements posed by mobile network traffic.
  - Especially under the dynamic changing contexts in the mobile networks
  - Existing metrics do not seem to reflect the load of the network
  - Existing auto-scaling techniques (i.e. Horizontal and Vertical) are not efficient in all scenarios.

AI-assisted VNF Autoscaling solution:

- The development of an artificial intelligence (AI) based autoscaling solution for VNF
- The customization & optimization of scaling metrics through AI algorithms, according to changing network conditions & requirements
- Development of novel Machine Learning (ML) driven techniques for optimizing autoscaling decisions.

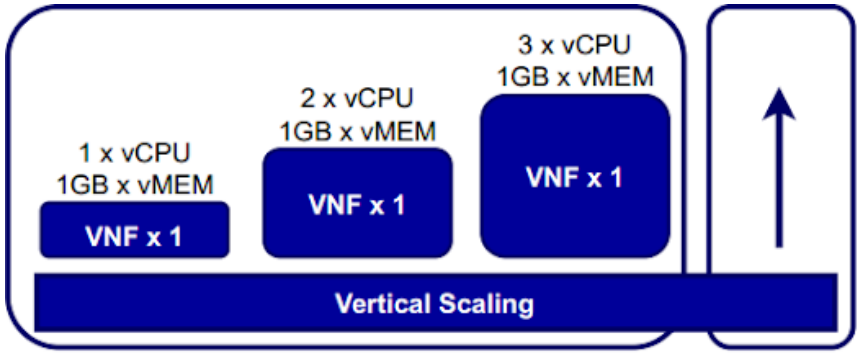
© 2023. Samsung Research All rights reserved.

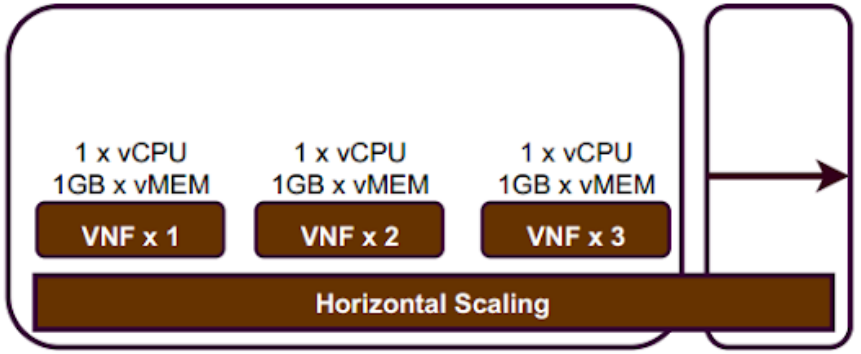
Advanced AI algorithms for Infrastructure: Autoscaling

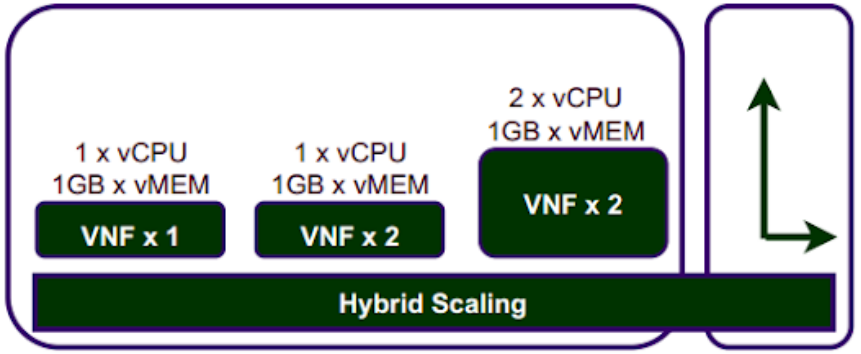
Vertical Scaling
- Data consistency
- Customized scaling
- Lower cost
- Low availability
- No load balancing

Horizontal Scaling
- High availability
- Load balancing
- Resiliency
- Costly
- Data inconsistency
- Resource underutilization

Hybrid Scaling
- High availability
- Load balancing
- Customized scaling
- Affordable
- Data consistency
- Resource efficiency





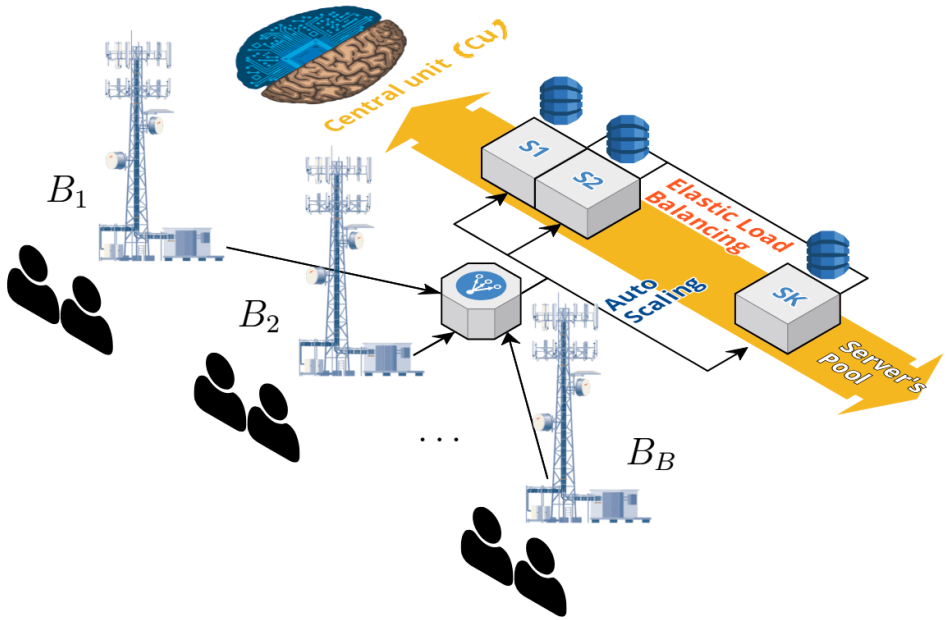


© 2023. Samsung Research All rights reserved.

Advanced AI algorithms for Infrastructure: Autoscaling

Goal:

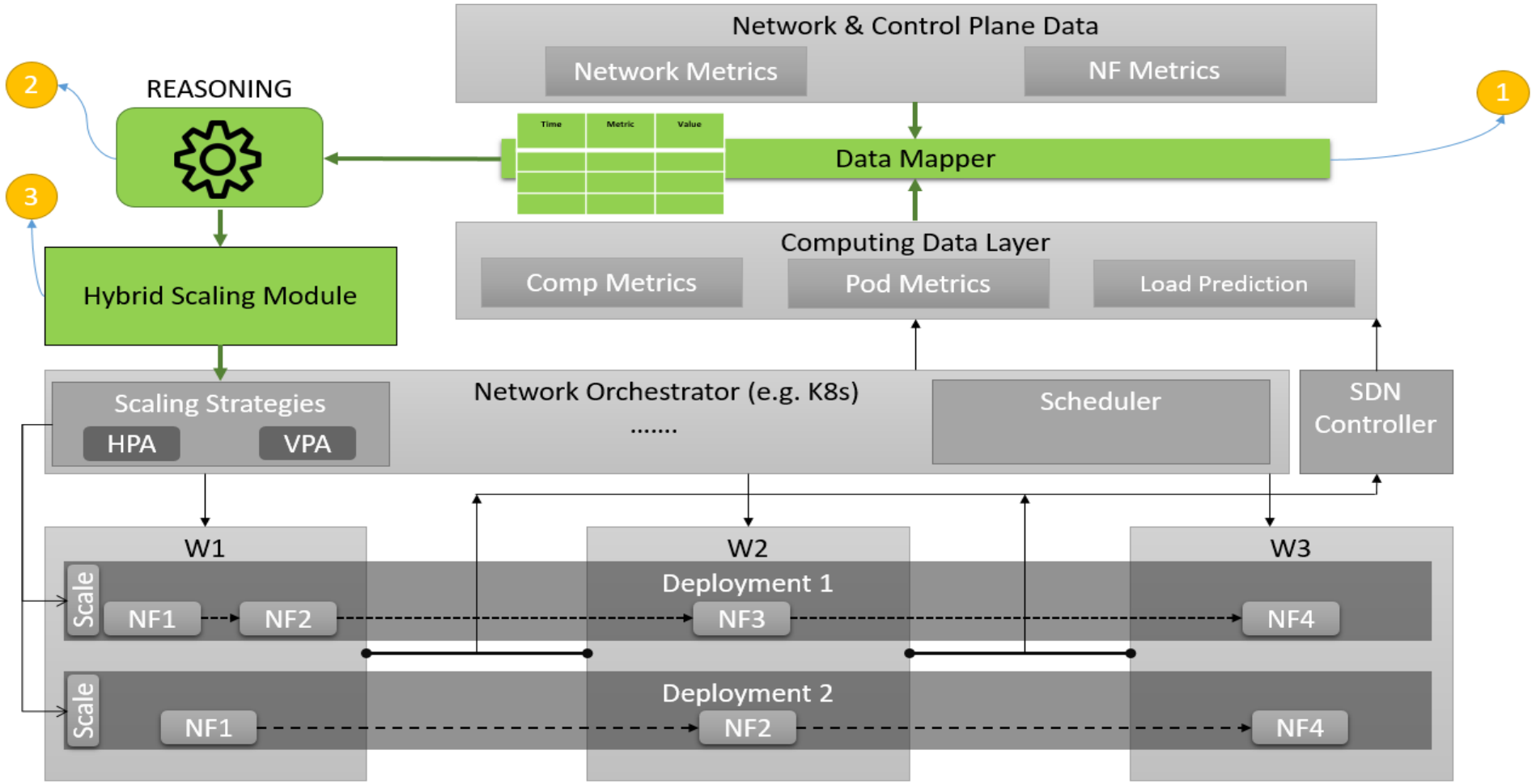
- Leverage AI to learn to scale VNFs.
- Based on the current system state, e.g., traf fic arrivals, service rates, service level agree ment (SLA), etc. The AI solution learns to scale a VNF:
  - Vertically: add more physical resources.
  - Horizontally: instantiate new containers.
- The AI aims to minimize a cost function which takes the shape of multi-objective optimization of different KPIs, e.g., latency, QoS, etc.



© 2023. Samsung Research All rights reserved.

Advanced AI algorithms for Infrastructure: Autoscaling

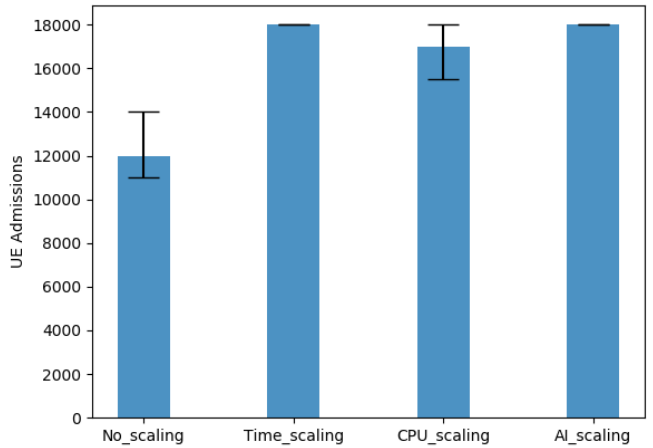
Samsung Research



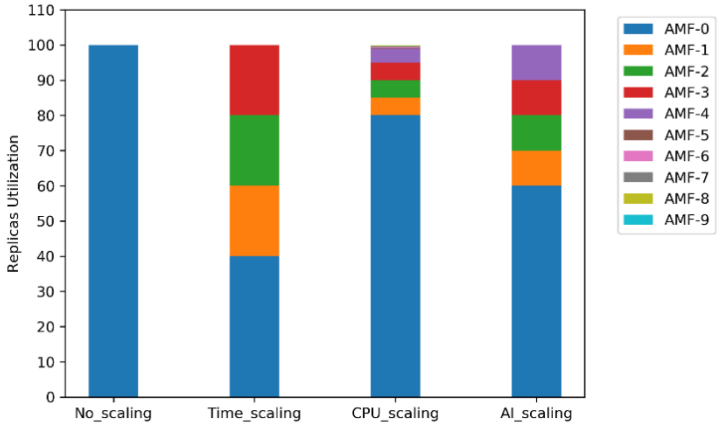
© 2023. Samsung Research All rights reserved.

Advanced AI algorithms for Infrastructure: Autoscaling

- Results show that scaling is needed for handling such a high number of users
- In the cases of Time based scaling (every 6h) and AI based scaling, no UE is left out of the system
  - New replicas are spawned and the original AMF does not reach its limits
  - CPU based scaling spawns new replicas constantly, even if not needed
    - Creating stability issues for the system -> UEs are rejected
- Using the scaling results to higher utilization of the replicated AMF services rather than the original one
- This can be projected to better energy efficiency for the AI case
  - The AMF replica is only spawned when needed, and destroyed afterwards



# of admitted UEs



# of replicas

© 2023. Samsung Research All rights reserved.

Samsung Research

Thank you