Fog Computing and Networking:
A New Paradigm for 5G and IoT Applications

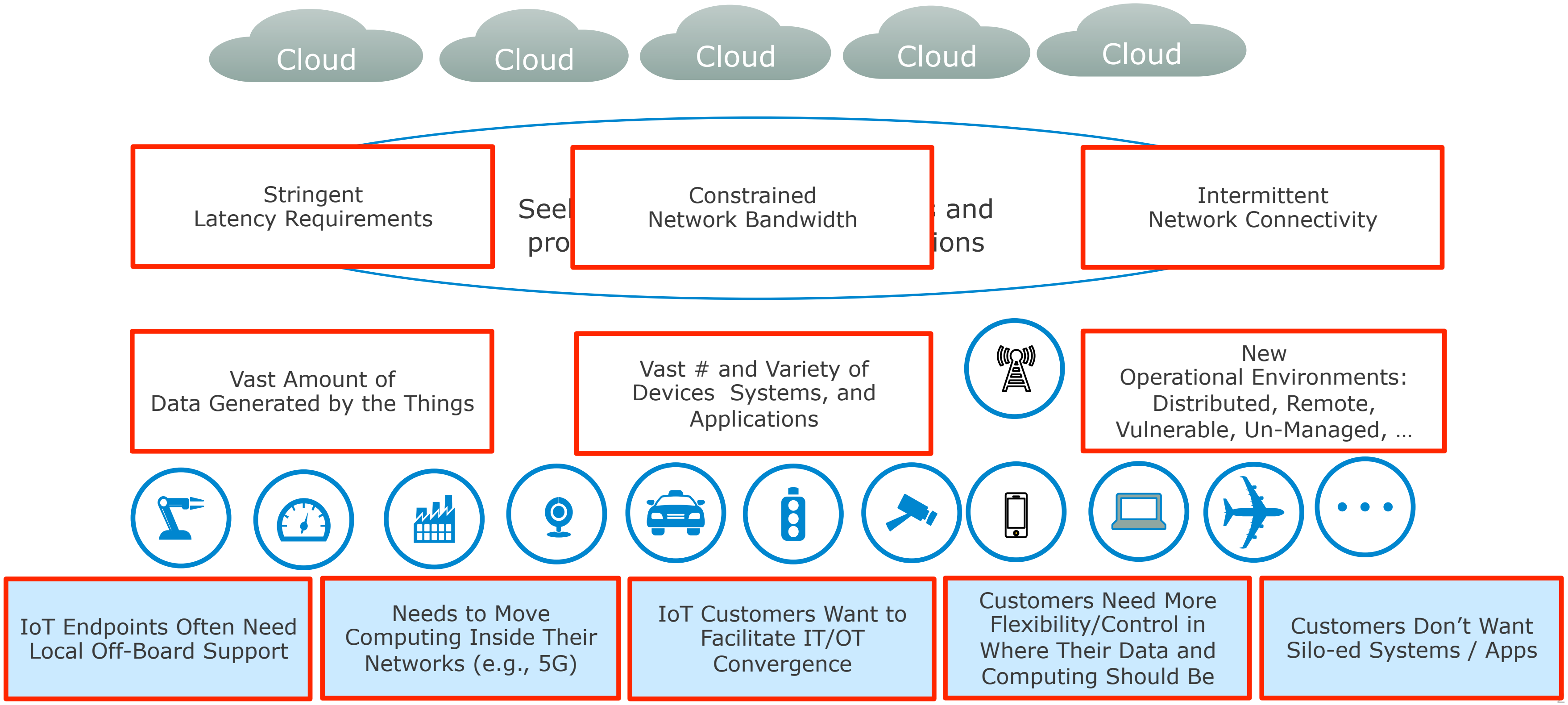
Tao Zhang, Cisco
Tony Quek, Singapore University of Technology and Design
Jianwei Huang, The Chinese University of Hong Kong
Ai-Chun Pang, National Taiwan University
Yang Yang, Shanghai Institute of Microsystem and Information Technology

IEEE ICC Tutorial, May 2017

The Era of Fog Computing & Networking

Tao Zhang
Cisco Corporate Strategic Innovation Group
Co-Founder and Board Director, OpenFog Consortium
tazhang2@cisco.com
May 21, 2017

Cloud-Only Computing Models Inadequate for IoT, 5G, Embedded AI, …



All Industries Are Developing Solutions, But…

Silo-ed
Systems

- For different networks: 5G, wired telecom, enterprises
- For different industry verticals: manufacturing, smart cities, ...
- For different applications inside same industry verticals
- For different types of edge devices: mobile edge, enterprise edge, users' edges, and more

Isolated
Systems and Applications

- Poor integration with the cloud
- Difficult to interoperate or collaborate with each other

Massively Confused
Market and Customers

- Edge Computing vs. Mobile Edge Computing vs. Multi-access Edge Computing vs. Mobile Edge Cloud vs. Cloud RAN vs. MiniCloud vs. Cloudlet vs. CORD vs. ...
- … and where does the Cloud fit in all these?

Out of the Chaos … Emerged an Important Trend:

Cross-industry need to move computing closer to users … or in other words … the need for

Fog Computing

What is Fog Computing and How is It Different?

Horizontal

- Support multiple network types and industry verticals
  (not silo-ed systems for different networks, industries, or application domains)

Works Over and Inside Wired or Wireless Networks

(no need for silo-ed platforms just for moving computing inside any specific network such as 5G)

E2E Architecture

- Distribute, use, manage, and secure resources & services
- Enable horizontal and vertical interoperability, orcherstration, and automation
  (not just placing servers, apps, or small clouds at edges)

Cloud-to-Thing Continuum

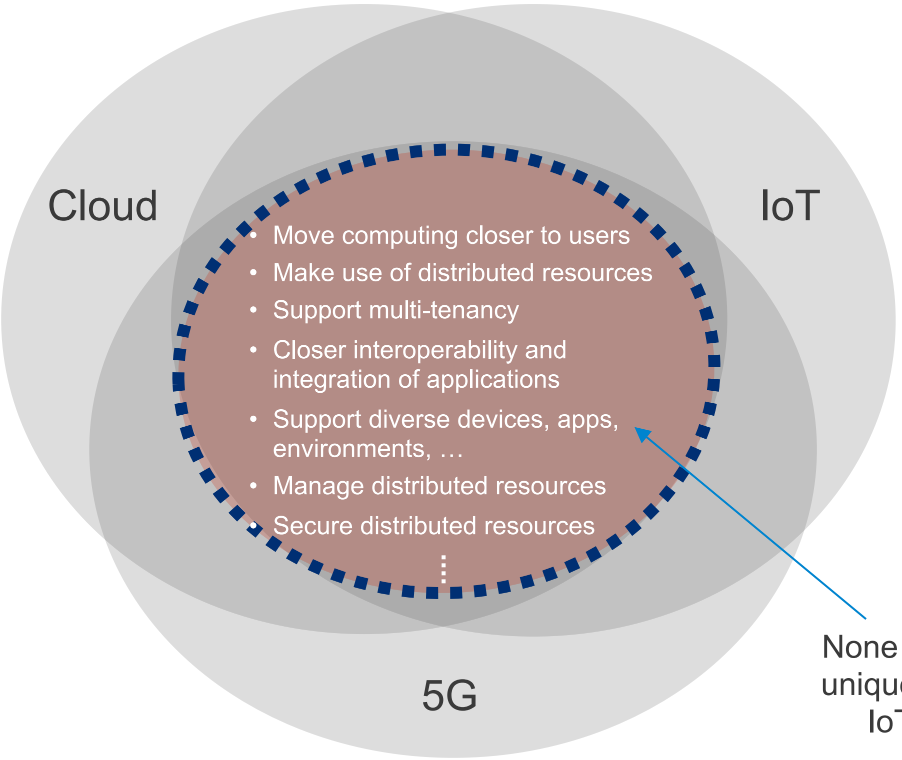
- Enable computing anywhere along the continuum
  (not just at any specific edge)
- Orchestrate resources in clouds, fogs, and things
  (not just isolated edge devices, systems, or apps)

Fog Computing Is Analogous to Previous Internet Revolutions: TCP/IP, WWW, …

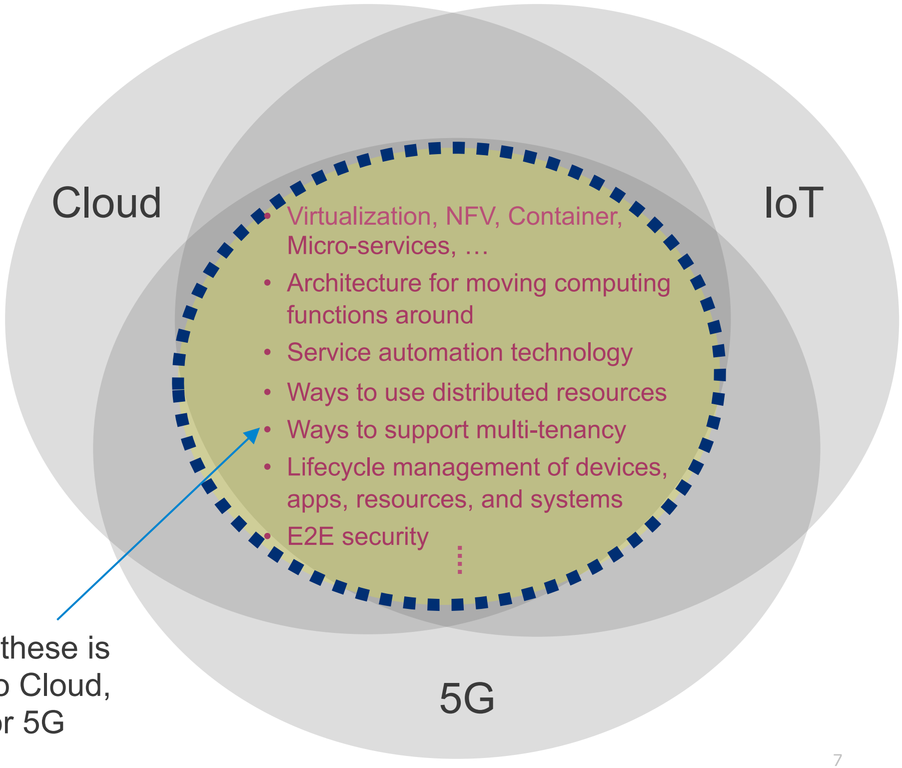
TCP/IP	WWW	Fog Computing and OpenFog Consortium
A horizontal framework for distributing data packets	A horizontal framework for accessing files anywhere	A horizontal framework for distributing computing functions and using, managing, & securing distributed resources and services
Wouldn’t it be better if we also had a TCP-for-2.5G? for 3G?, for 4G, … **NO**	Wouldn’t is be better if we also had a HTTP-for-2.5G? for 3G? for 4G? for wired telecom? … **NO**	Should we have a separate fog-like system for 5G? another for wired telecom? another for enterprises? another for smart city? another for manufacturing? … **NO**

IoT, Cloud, and 5G Are Converging

Common Functionalities



Shared Technologies



None of these is unique to Cloud, IoT, or 5G

Fog Disrupts Existing Industry Landscape

Causing convergence at the edge and laying foundation for IT/OT convergence

- ❑ Edge routers, switches, wireless APs, app servers, and storage have been converging into unified fog nodes lower costs, higher efficiency, closer app integration, and easier IT/OT convergence

Empowers and extends the cloud

- ❑ Connect things to the cloud
- ❑ Deliver cloud services to things

Brings new services

- ❑ Fog-based security services for IoT
- ❑ Fog as a Service

Allows players of all sizes to play

- ❑ Rise of local and regional fog system/service operators?

Transforms industries toward unified cloud-to-thing continuum of computing platform, services, and apps

- ❑ Reduce isolated or silo-ed edge computing systems and applications

Why Must We Care About Fog **Now**?

Service Orchestration
Putting back together applications and services

Microservices
Decompose applications and services

Container
Portablize applications and services

Network Function Virtualization (NFV)
Virtualize network elements

Hardware Virtualization
Abstract OSs away from HW

Software Defined Networking (SDN)
Separate control and data planes
Softwarize control functions

Technology Advancement

User and Business Needs

IT and OT Convergence

Customers don't want silo-ed systems or services

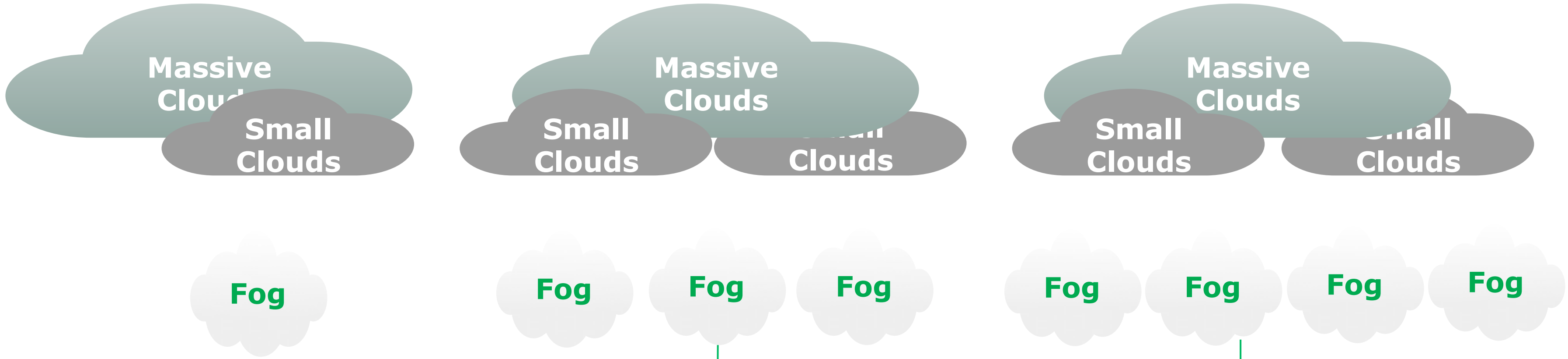
Diverse environments and requirements

Local off-board support

Time-sensitive local processing

Vast amount of data at the edge

Fog Is the Future …



Systems of Records

(Collecting/processing info or records)

- Web surfing
- HR, ERP, CRM, supply chain, marketing, and sales apps
- E-commerce
- Online content delivery

(larger scale and more real time)

Systems of Engagement

(People engaging with each other)

- Email
- Collaboration
- Social networking
- Virtual / Augmented Reality
- …

Systems of Control

(Real-time control between devices/people)

- Industrial control systems
- Mobile edge computing in 5G
- Connected or autonomous cars and trains
- Smart buildings and cities
- Smart grids, oil and gas systems
- Medical and consumer devices controlled by embedded AI
- Visual Security
- Drones – commercial and consumer
- Robotics – industrial and consumer
- …

Unique Architectural Requirements

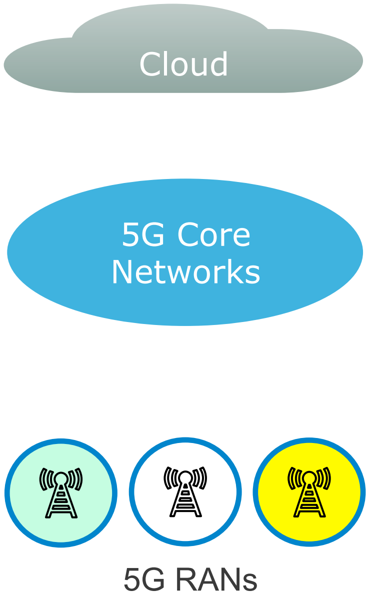
Distributed but **unified** computing platforms enabling seamless services along cloud-to-thing continuum	❑ Distribute resources / apps to many fogs: remote with diverse capabilities, operating environments, and user requirements ❑ Orchestrate resources in clouds, fogs, and things to enable seamless E2E services ❑ Support fog-based services and Fog-As-A-Service: interfaces, protocols, procedures, … ❑ Scalable and trustworthy monitoring ❑ Lifecycle management of distributed fog systems, resources, and apps, with higgh degree of automation
Hierarchical architecture	❑ Interactions and interfaces between hierarchical levels, between fogs on the same level, between fog and cloud, and between fog and things
Integration with Operational Technology (OT) systems	❑ Fogs integrated into the operations of end-user systems (e.g., machines, cars, trains, drones, …) will inherit requirements from these OT systems
Work **over** and **inside** wireless and wireline networks	❑ Moving computing into Radio Access Networks, wired telecom central offices, enterprise networks, …
Elastic architecture	❑ Support fogs that can vary widely in size, # of users, # of applications, capabilities, and user requirements
Security	❑ Protect distributed fog systems: remote, vulnerable environments, run by non-IT experts, … ❑ Handle new threats, unique operational constraints, resource constraints, … ❑ Enable fog-based security services (or fog-based security as a service)

5G Needs a New Computing Platform …

and also here

Today, we centralize most network control functions and mobile services here

Tomorrow, we want to move some control functions, services, and user apps in or close to here



or maybe an edge-to-core-to-cloud computing platform for 5G?

Perhaps, an edge-to-core computing platform for 5G?

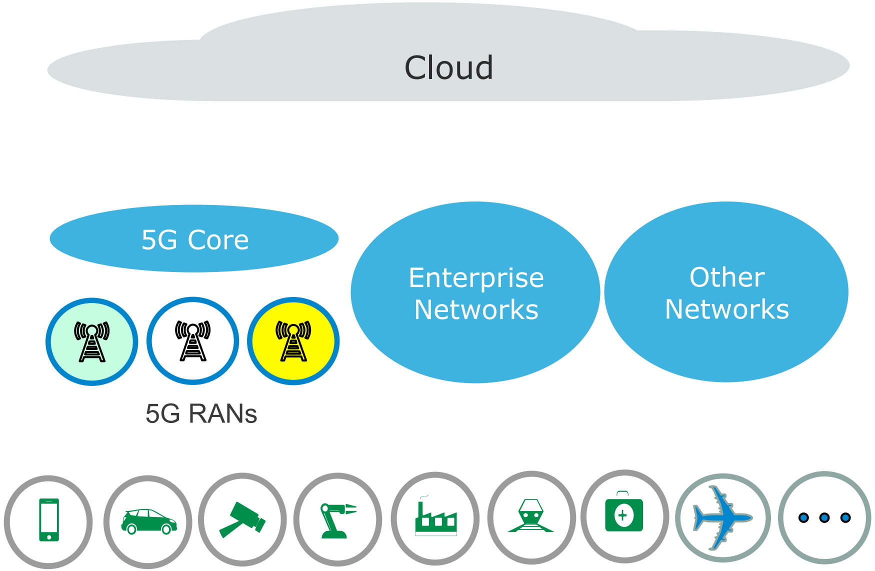
Should we develop an edge computing platform dedicated for 5G RANs?

What's really need?

- **Move** computing functions around
- **Make use** of distributed computing resources
  - ✓ Pool resources in different places
  - ✓ Automate services
  - ✓ Provide baseline services
  - ✓ Support multi-tenancy
- **Manage** distributed resources & apps
- **Provide** attractive app development environments
- **Secure** distributed resources & apps
- …

and, none of these is unique to 5G

Fog – the Computing Platform that Brings Cloud, IoT, and 5G Together



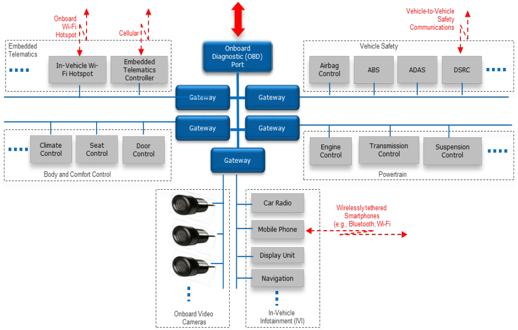
Horizontal Fog Platform
Integrated with Cloud and Things
for
Seamless Cloud-to-Thing Service Continuum

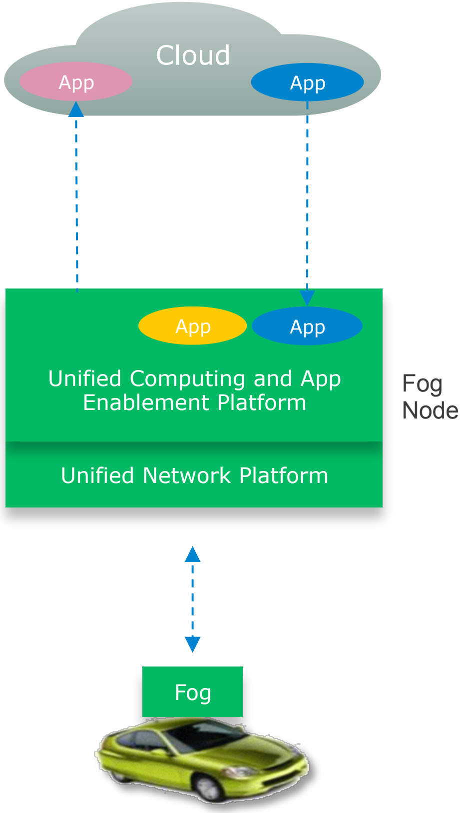
- Unified E2E computing platform: pooling, automation, management, security in the cloud, the fog, and even on the things
- Seamless services and apps: anywhere along the Cloud-Fog-Thing continuum
- Standards-based plug and play

Powered by a Core Set of Common Technologies

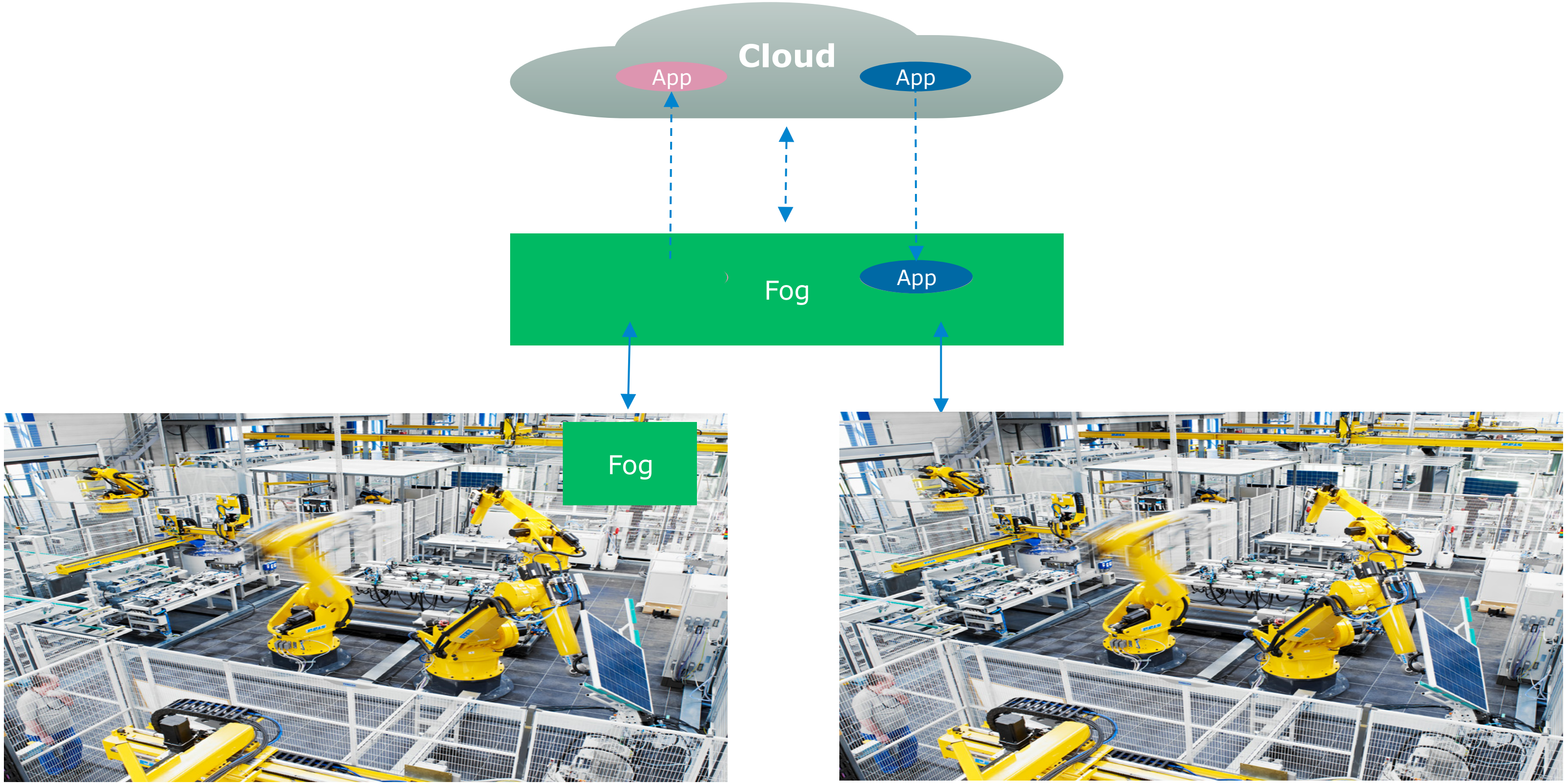
- Virtualization, NFV, Container, Microservices, …
- Resource/Function Distribution
- Service Automation
- Resource Pooling
- Lifecycle Management
- Security
- Standards-based APIs

Use Case: Connected Cars & Smart Cities





Use Case: NextGen Manufacturing



OpenFog Consortium

Build global consensus

Develop fog industries/markets

Develop reference architecture & technologies

Technology

Identify and share use cases and requirements

Develop an open reference architecture

Demonstrate technologies and business values with testbeds

Co-develop / Influence necessary new standards

Innovation

Foster industry-university and cross-industry partnerships to identify and tackle technical challenges

Provide a forum to share ideas and facilitate business development

Education

Evangelize fog concept and value, share best practices, showcase real-world applications

Educate through events, conferences, training courses, and publications

OPTIMIZATION AND RESOURCE MANAGEMENT FOR FOG NETWORKING

TONY Q.S. QUEK
ASSOCIATE PROFESSOR
ASSOCIATE HEAD OF ISTD PILLAR
DEPUTY DIRECTOR, SUTD-ZJU IDEA
GRADUATE CHAIR, ISTD PILLAR

ICC 2017

Singapore University of Technology and Design

Key Challenges

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN
Established in collaboration with MIT

Optimization and Resource Management in Fog Networking

- Local Architecture
- Global Architecture
- Distributed Optimization
- Centralized Optimization
- Fog RAN
- Cloud RAN
- Machine Learning
- Artificial Intelligence
- Network Economics
- Sharing Economy
- Big Data Analytics
- Network Virtualization
- IoT Services
- Tactile Internet
- Security
- Privacy

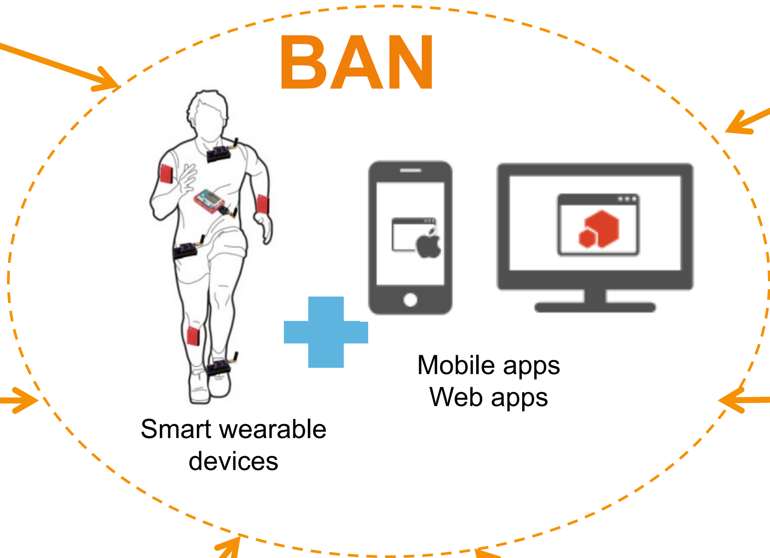
Adaptive MAC Scheduling in Fog Computing

IoT - Body Area Networks (BANs)

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN
Established in collaboration with MIT

Voice-guidance gymnastic training

Detecting presence of occupants & automated home appliances



Toddler fall detection

Pedestrian and cyclist safety

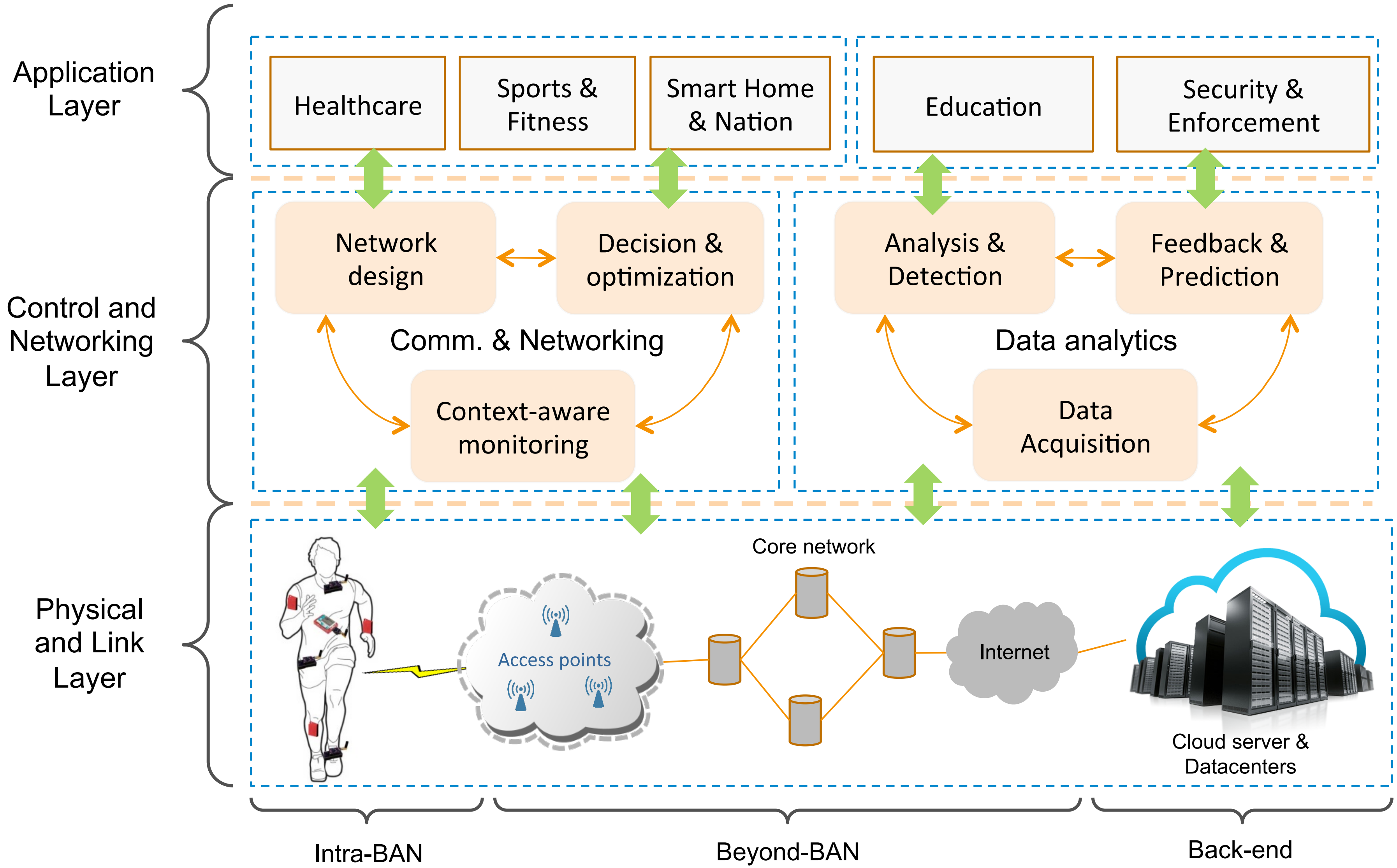
Elderly care

Patient rehabilitation

INFORMATION SYSTEMS TECHNOLOGY AND DESIGN

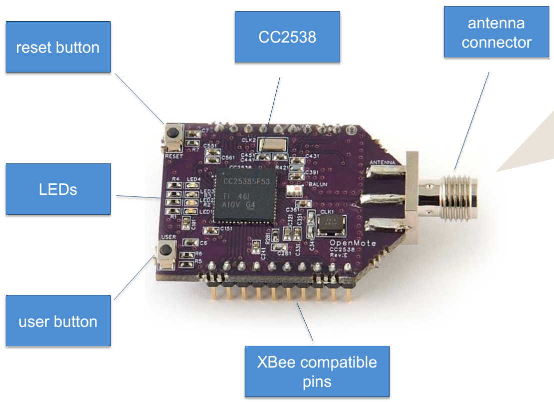
Sensing + Networking + Analytics + Apps

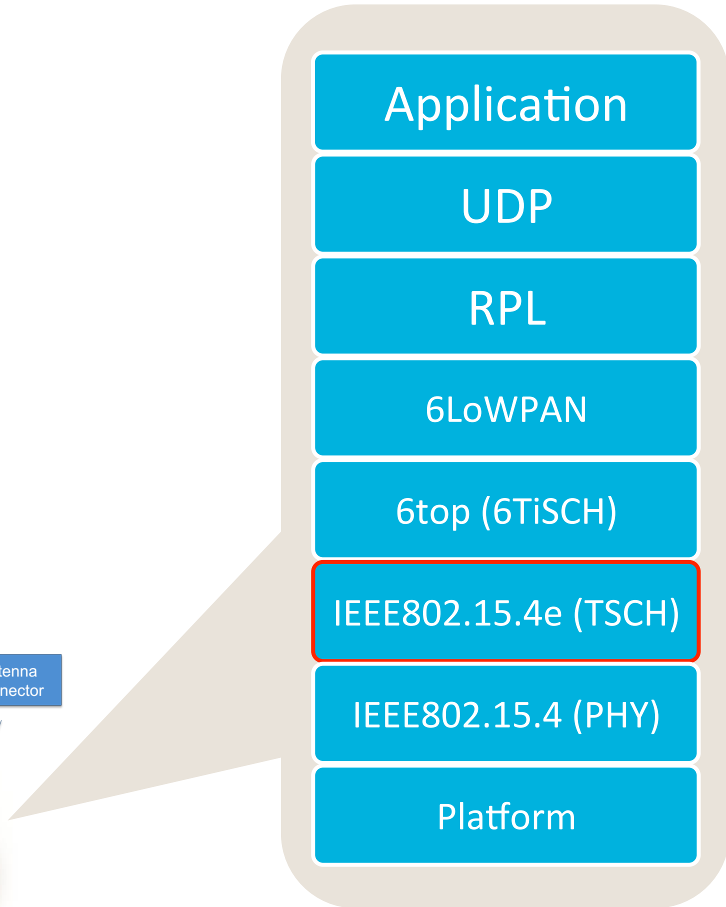
SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN
Established in collaboration with MIT



ZigBee Platform - IEEE 802.15.4

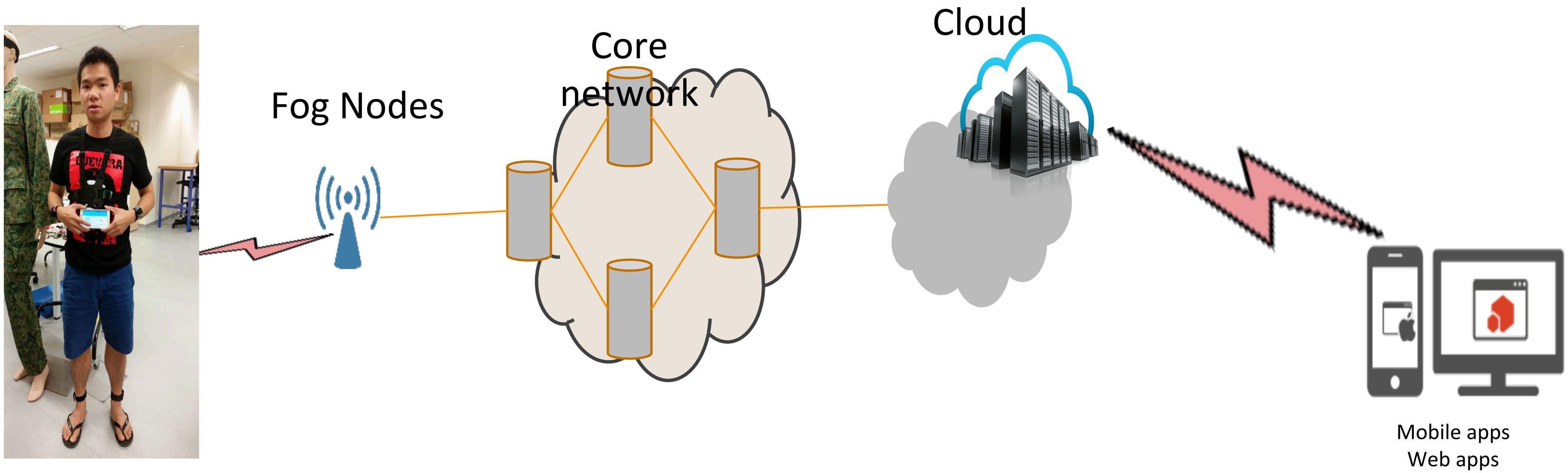
- **ZigBee/IEEE 802.15.4**, Bluetooth, BLE
  - Range, power consumption, mesh network
- CSMA vs TSCH (Time Slotted Channel Hopping)
  - Coexistence,
  - Flexible, scalable, interference and predictive latency.





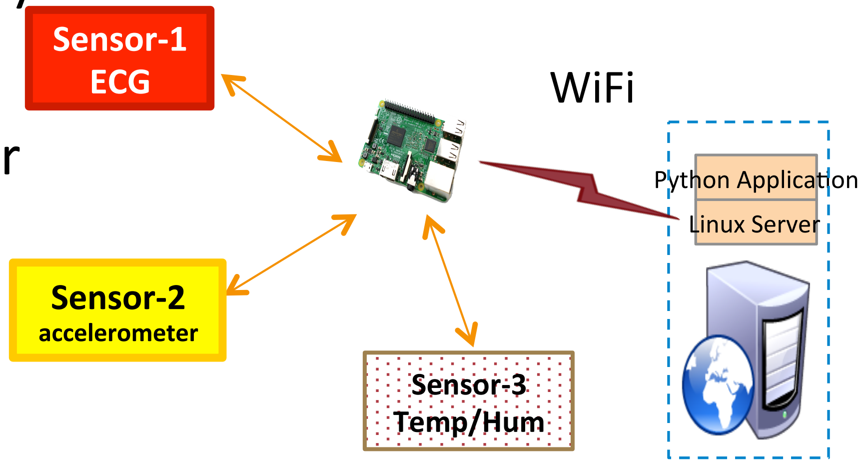
Adaptive Network Close to User

- Data analytics **nearby users**
  - Time critical, availability, privacy
- Energy efficiency: **Adaptive sampling**, **Bandwidth**.



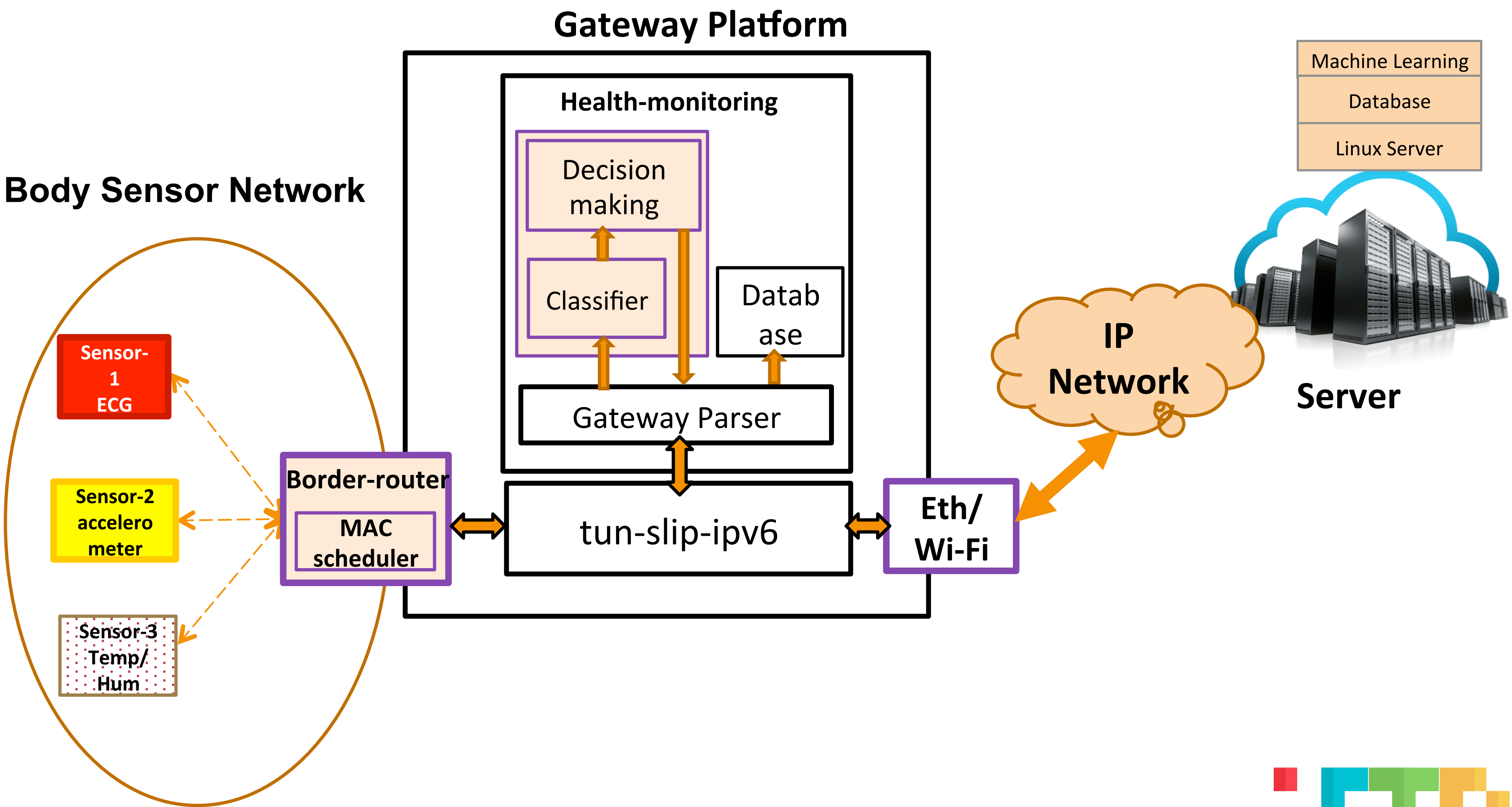
System Overview

- Collected, **analyzed** at fog node/gateway.
- Trigger alarm
- Increase sampling rate of specific sensor
  - ➔ High resolution data.
  - ➔ Higher bandwidth of network.
- Change the scheduling to provide more bandwidth.



Case	ECG sensor		Accelerometer		Core temperature	
	Sampling	**QoS**	**Sampling**	**QoS**	**Sampling**	**Qos**
Normal	64Hz	d<100ms, PDR > 90%	32 Hz	d<100ms, PDR > 90%	1 Hz	d < 1s, PDR > 90%
Dangerous	256 Hz	d<50ms, PDR >99%	The same	The same	The same	The same

System Software Architecture

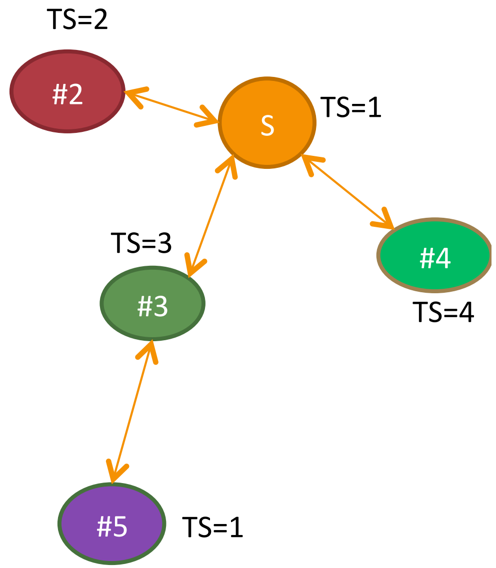


Time Slotted Channel Hopping (TSCH)

Time

Channel

Super-frame	N					N+1				
Timeslot	**TS1**	**TS2**	**TS3**	**TS4**	**TS5**	**TS1**	**TS2**	**TS3**	**TS4**	**TS5**
ASN	96	97	98	99	100	101	102	103	104	105
11	S→*									
12		2→S								
13			3→S							
14				4→S						
15										
16	5→3					S→*				
17							2→S			
18								3→S		
19									4→S	
…						5→3				
26										

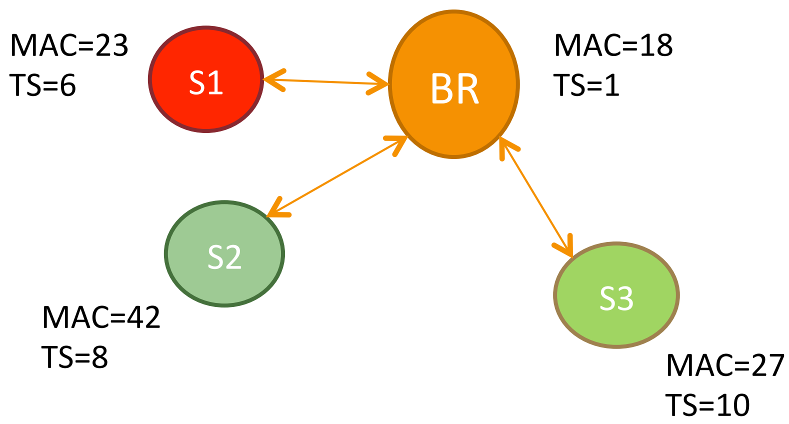


Adaptive MAC scheduler: Normal

	TS1	TS2	TS3	TS4	TS5	TS6	TS7	TS8	TS9	TS10	TS11	TS12	TS13	TS14	TS15	TS16	TS17	TS1	TS2	TS3 ...
Border -router	Tx(S)					Rx		Rx		Rx								Tx(S)		
Sensor 1	Rx					Tx(S)												Rx		
Sensor 2	Rx							Tx(S)										Rx		
Sensor 3	Rx									Tx(S)								Rx		

L = 17

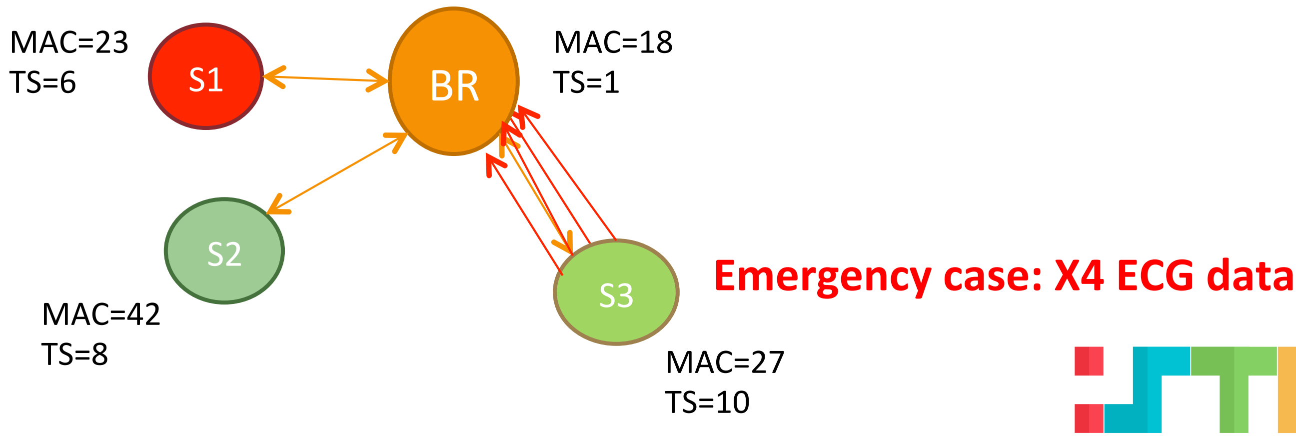
Slot = 10ms



Adaptive MAC scheduler: Urgent (1)

	TS1	TS2	TS3	TS4	TS5	TS6	TS7	TS8	TS9	TS10	TS11	TS12	TS13	TS14	TS15	TS16	TS17	TS1	TS2	TS3 ...
Border -router	Tx(S)					Rx		Rx		Rx	Rx	Rx	Rx					Tx(S)		
Sensor 1	Rx					Tx(S)												Rx		
Sensor 2	Rx							Tx(S)										Rx		
Sensor 3	Rx									Tx(S)	Tx	Tx	Tx					Rx		

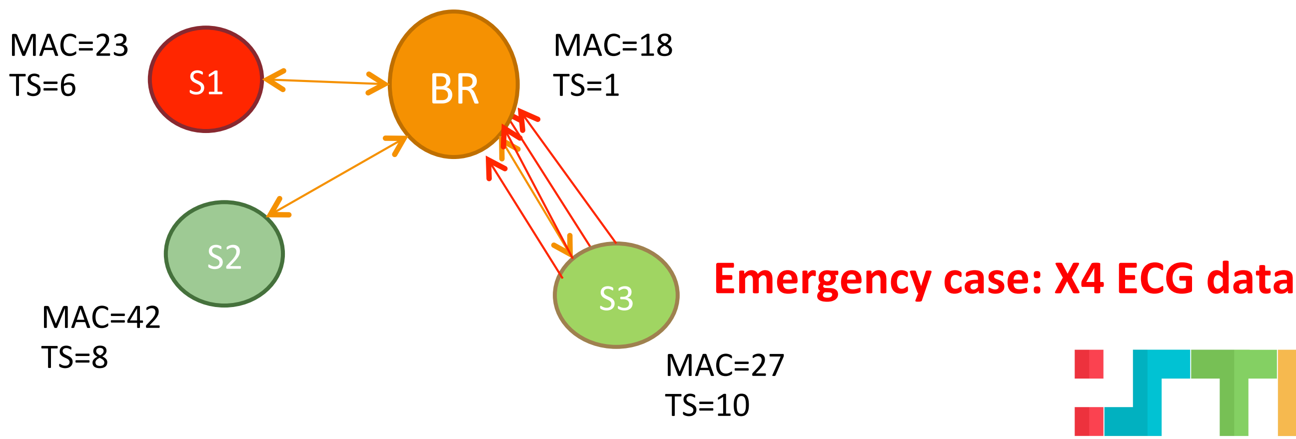
Consecutive timeslots allocation



Adaptive MAC scheduler: Urgent (2)

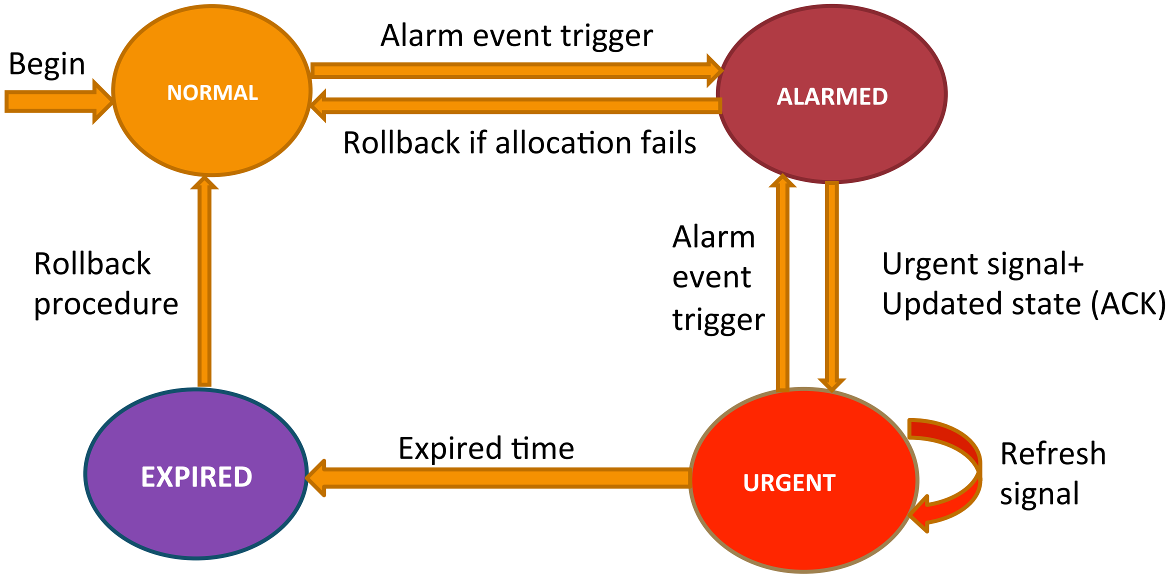
	TS1	TS2	TS3	TS4	TS5	TS6	TS7	TS8	TS9	TS10	TS11	TS12	TS13	TS14	TS15	TS16	TS17	TS1	TS2	TS3 ...
Border -router	Tx(S)	Rx			Rx	Rx		Rx		Rx				Rx				Tx(S)	Rx	
Sensor 1	Rx					Tx(S)												Rx		
Sensor 2	Rx							Tx(S)										Rx		
Sensor 3	Rx	Tx			Tx					Tx(S)				Tx				Rx	Tx	

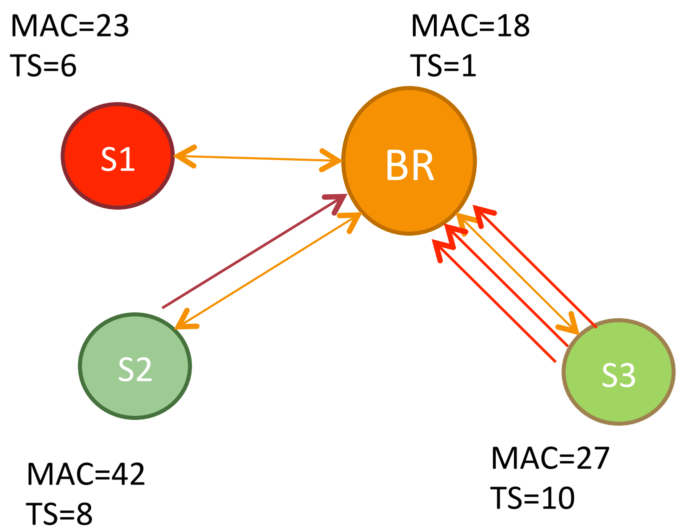
Equally spaced timeslots allocation



State Machine Representation

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN
Established in collaboration with MIT





NORMAL:
- BR: receive in the normal link + analyze
- Sensor: sample + send in the normal link

EXPIRED:
- BR: remove extra links
- Sensor: remove extra links, normal sampling rate.

ALARMED:
- BR: Allocation, send control message
- Sensor: get control signal, add timeslots, change state

URGENT:
- BR: receive in extra links & normal link
- Sensor: sampling at a higher rate

INFORMATION SYSTEMS TECHNOLOGY AND DESIGN

Jitter Minimization

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN
Established in collaboration with MIT

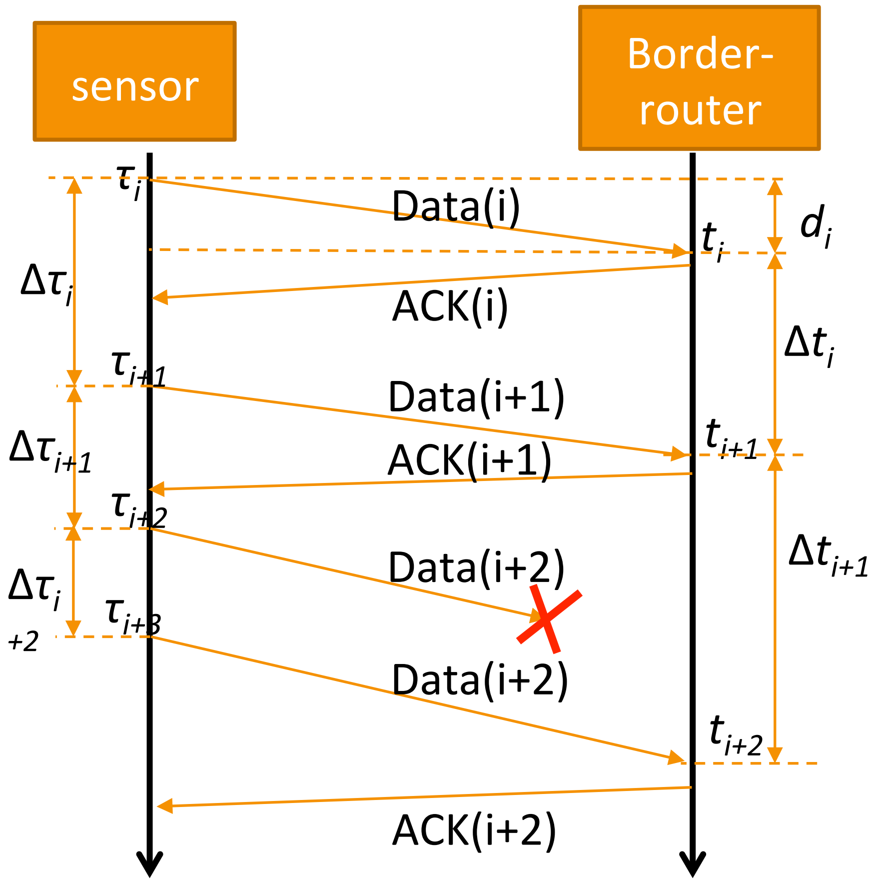
Minimize the variance of inter-arrival times:

$$\min_{\mathcal{F}} \mathrm{Var}[\Delta t_i],$$

Where:

$$\mathcal{F} = \{(\tau_1, \tau_2, \ldots, \tau_{K+1}) \mid 0 = \tau_1 < \tau_2 < \ldots < \tau_{K+1} = T\}$$

$$\Delta t_i = t_{i+1} - t_i,\ t_i = \tau_i + d_i \text{ and } d_i \sim \mathcal{N}(\mu, \sigma^2), \forall i.$$



Solution:

$$\tau_1 = 0,\ \tau_{K+1} = T,\ and\ \tau_{i+1} - \tau_i = \frac{T}{K},\ \forall i.$$

INFORMATION SYSTEMS TECHNOLOGY AND DESIGN

MITA (Minimal-Jitter Time-slot Allocation) Algorithm

ALGORITHM 1: MITA (Minimal-Jitter Time-slot Allocation) algorithm

Data: N_{ex}, t_{cur}, L
Result: $\mathcal{S}$

$\mathcal{S} \leftarrow \emptyset$;
$step \leftarrow \lfloor \frac{L}{N_{ex}}) \rfloor$;
$t_{tem} \leftarrow t_{cur}$;
$i \leftarrow 0$;
while $|\mathcal{S}| < N_{ex}$ & $i < L$ **do**
    $t_{tem} \leftarrow (t_{tem} + step) \mod L$;
    if t_{slot} *is available* **then**
        $S \leftarrow t_{tem}$;
    else
        $j \leftarrow 0$;
        while $j < step$ **do**
            $t_{up} = (t_{tem} + j) \mod L$;
            $t_{low} = (t_{tem} - j) \mod L$;
            if t_{up} *is available* **then**
                $\mathcal{S} \leftarrow t_{up}$;
                break;
            end
            if t_{low} *is available* **then**
                $\mathcal{S} \leftarrow t_{low}$;
                break;
            end
            $j \leftarrow j + 1$;
        end
    end
    $i \leftarrow i + 1$;
end

Data Analysis

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN
Established in collaboration with MIT

Data collection
- Accelerometers*
- ECG
- Temperature
- Store, and timestamp data

Data cleaning
- Labeling
- Sampling window =2s
- Data imputation

Training & Testing
- Features extraction
- Train: SVM, GNB, Decision Tree
- Testing

Real deployment
- Run at GW & Server
- Trigger emergency event
- Testing network's performance

Detect anomaly activities: fall detection, heart attack.

For now: running is an emergency event (a proof of concept)

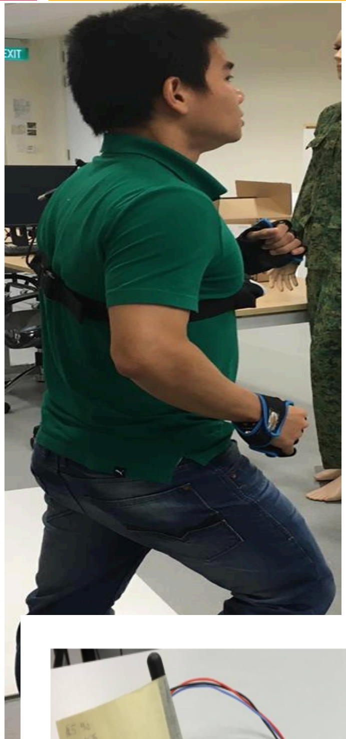
Table 1. Accuracy of machine learning models

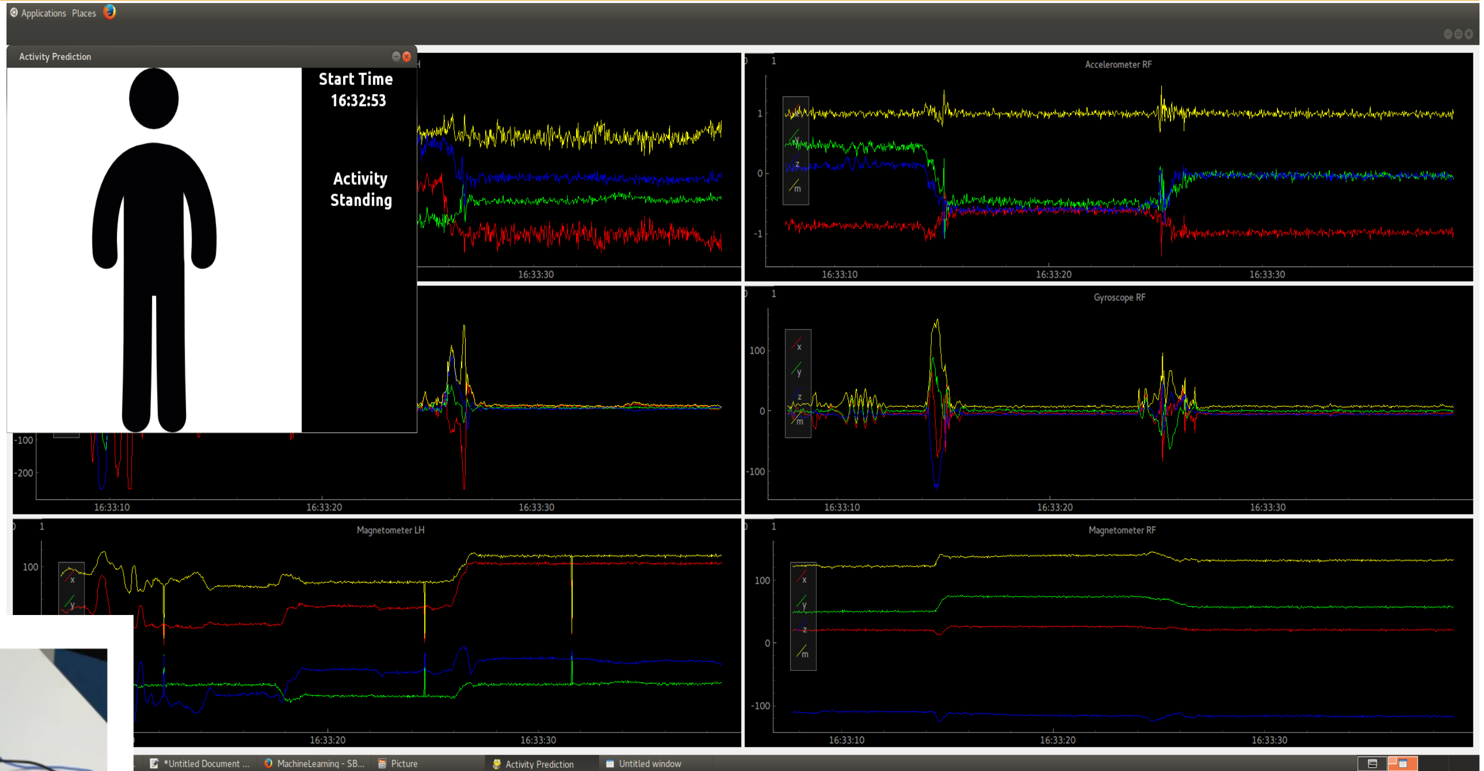
Model	Training accuracy (%)	Test accuracy (%)
Support vector machine	98.73	97.65
Decision tree	100	99.60
Gaussian Naive Bayes	98.40	98.11

INFORMATION SYSTEMS TECHNOLOGY AND DESIGN

Experiment & Results

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN
Established in collaboration with MIT





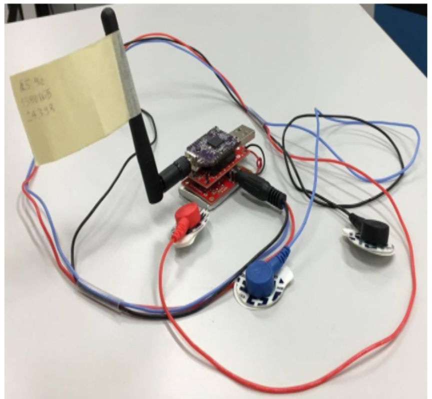
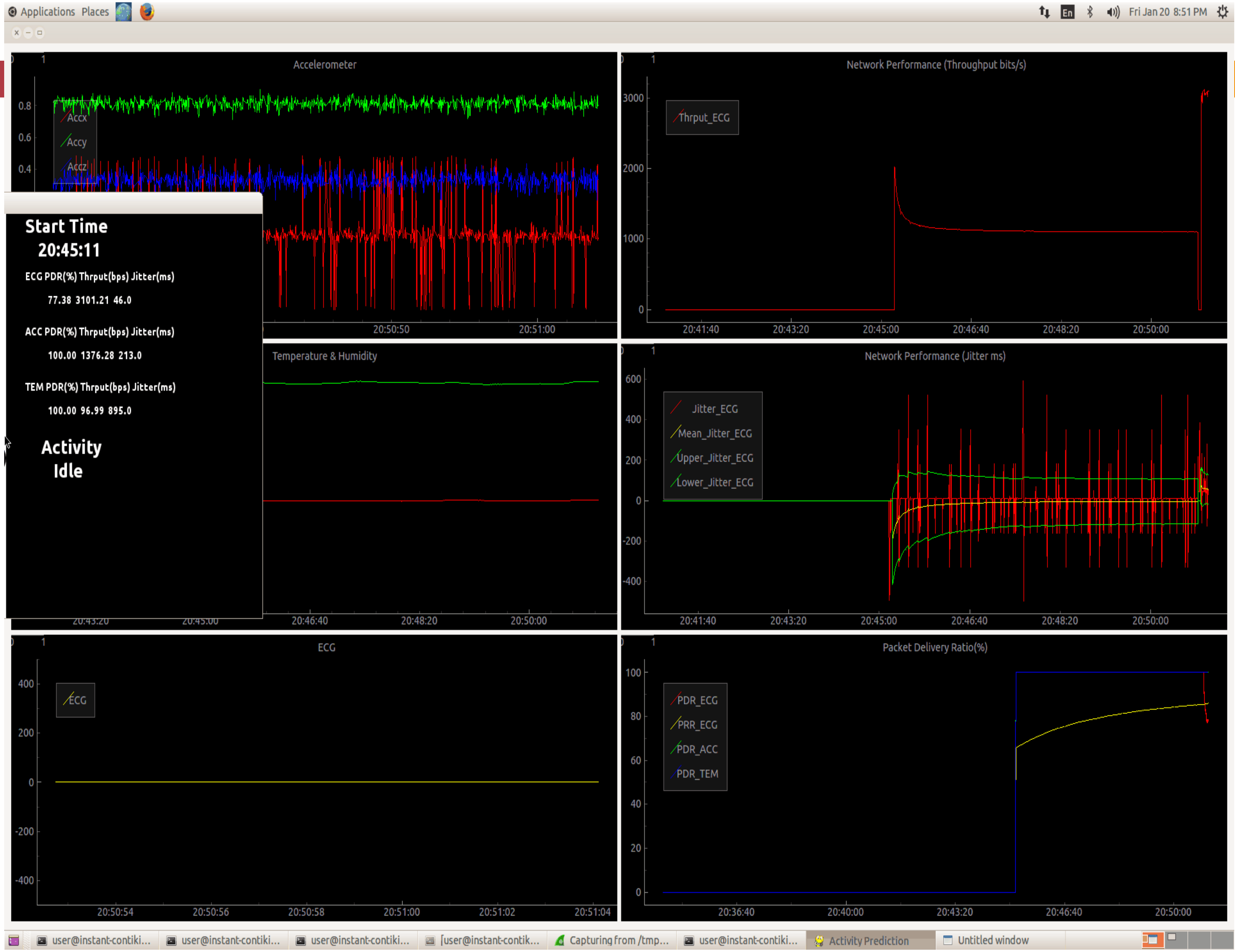


Fig. 8. ECG sensor with OpenMote-cc2538

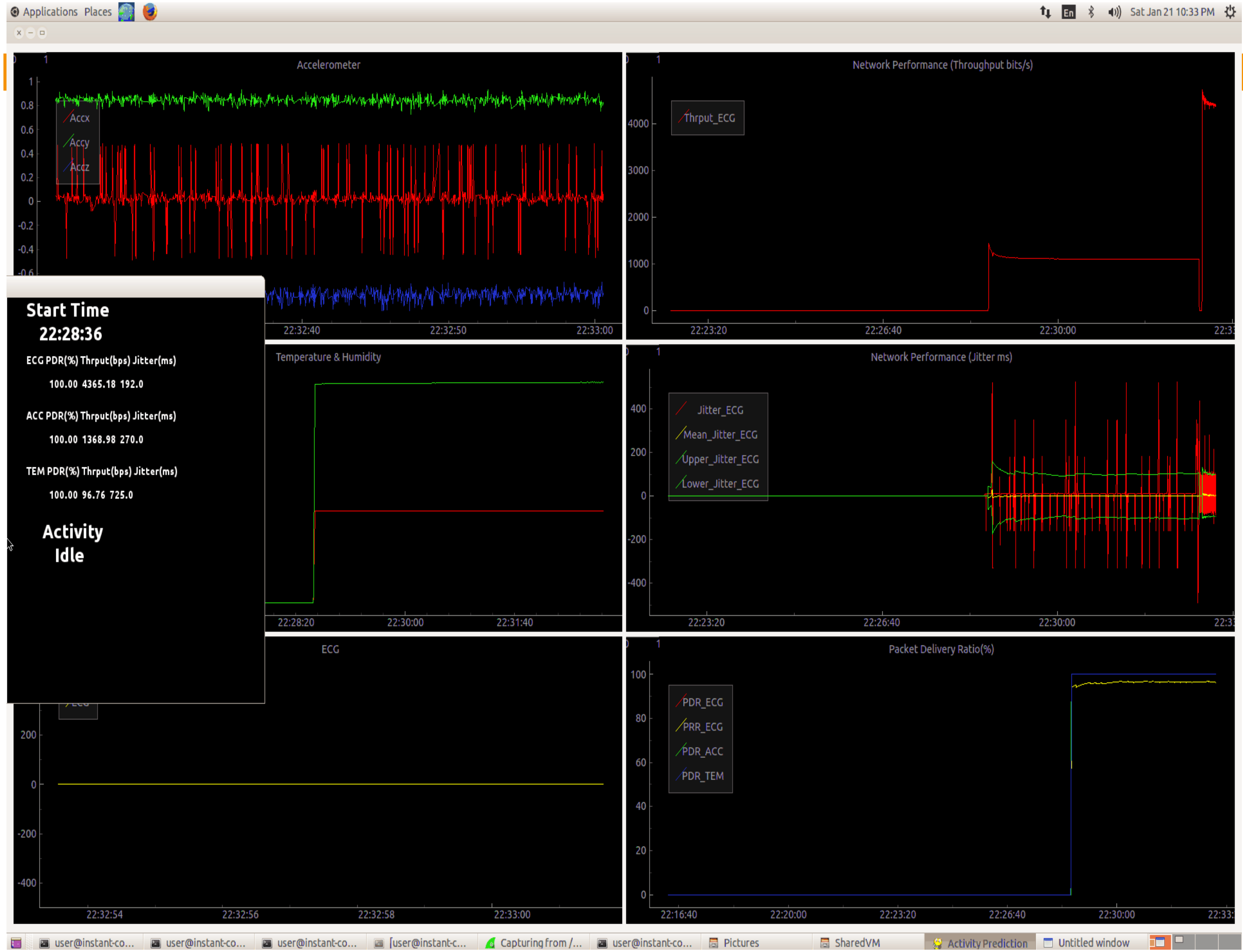
Table 2. The sampling frequency, and sending rate of sensors

Case	ECG		Accelerometer		Temperature/humidity	
	Sampling	Sending	Sampling	Sending	Sampling	Sending
Normal	64 Hz	2 Pkt/s	30 Hz	3 Pkt/s	2 Hz	1 Pkt/s
Urgent	256 Hz	8 Pkt/s	30 Hz	3 Pkt/s	2 Hz	1 Pkt/s

Without MITA Algorithm



With MITA Algorithm



Summary

- Design a fog-computing system for healthcare applications
  - An adaptive MAC scheduler
  - A MITA algorithm – minimize variance of jitter
  - Apply data analytics at the edge (fog node)
  - Deploy and test in the testbed
- Combine Adaptive MAC scheduler + Edged decision making

➔ High **reliability**, **availability**, **timely sensitivity** for healthcare in urgent situation

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Fog-Based Heterogeneous Sensor Data Fusion

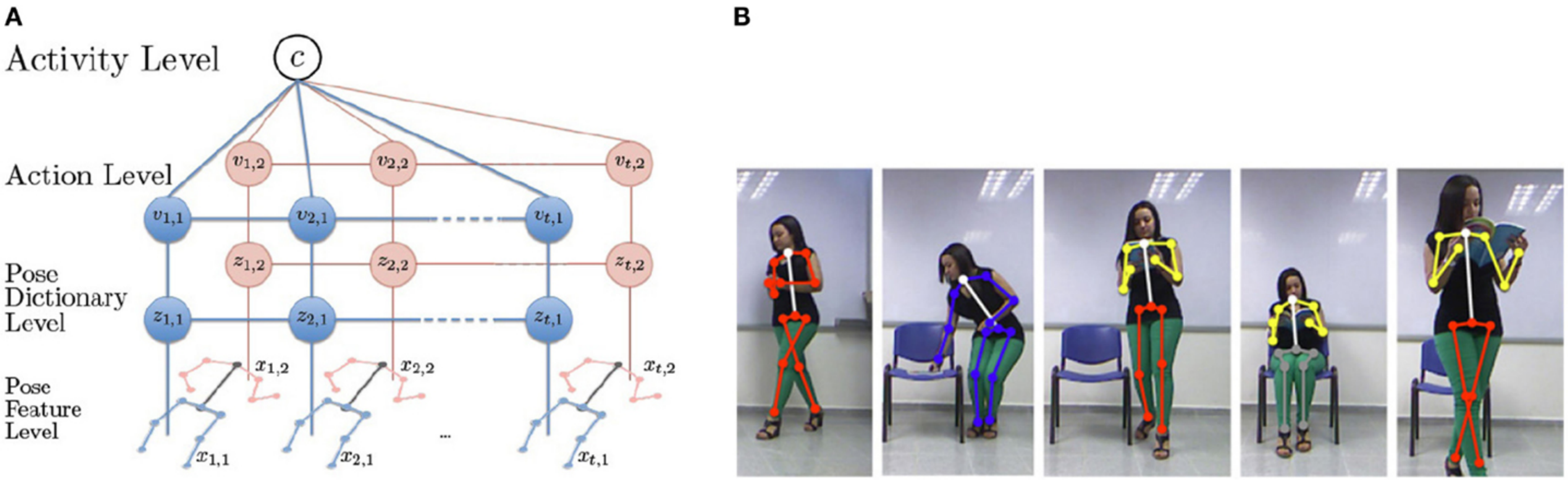
Selected Publications

- Z. Liu, W. Zhang, S. Lin, and T. Q. S. Quek, "Heterogeneous Sensor Data Fusion By Deep Multimodal Encoding," *IEEE J. Selected Topics Signal Processing*, Apr. 2017.

INFORMATION SYSTEMS TECHNOLOGY AND DESIGN

Multimodal Sensor Fusion

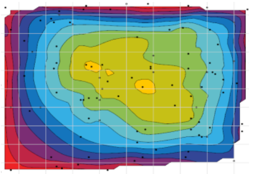
- **Multimodal** sensor data for many tasks
  - ✓ E.g., Human Activity Recognition
  - ✓ E.g., Health Condition, Monitor Environmental Condition, Control Intelligent System



Challenges of Machine Learning in Sensor Networks

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN
Established in collaboration with MIT

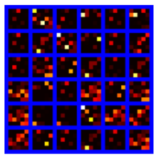
- Exploit spatial-temporal correlations in sensor data



After sparse coding

Missing Data, Sparse Sensor Deployment

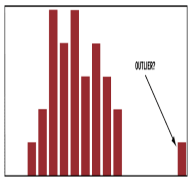
Smart Sensors (e.g. Plant Health, Crowd Sensing)



Compressed – 25 bytes

Data Compression for Transport and Storage

Fault Detection, Predictive Maintenance



Outlier Detection, Multimodal Analysis

Intruder Detection, Data Anonymization

INFORMATION SYSTEMS TECHNOLOGY AND DESIGN

Deep Fog-Based Sensor Fusion

- **Multimodal** sensor data for many tasks
  - ✓ E.g., Human Activity Recognition
  - ✓ E.g., Health Condition, Monitor Environmental Condition, Control Intelligent System

- Wireless sensor network data always **incomplete**
  - ✓ Missing data due to low battery, transmission loss or faulty sensor

- Goal: **Impute the missing values** in **multimodal** sensor dataset for inference in fog-based computing at the edge

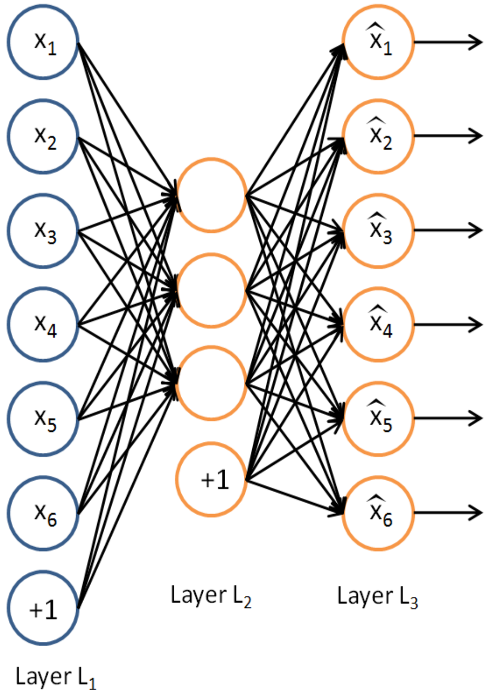
Sparse Autoencoder

Seminal paper on deep RBMs

Reducing the dimensionality of data with neural networks. Hinton-Salakhutdinov 2006

Sparse deep belief net model for visual area V2. Lee-Ekanadham-Ng 2007

Sparse RBMs



Given data vectors $y_1, y_2, \ldots, y_N$, let

$$x_i^{(0)} = y_i$$
$$x_i^{(1)} = f(\, A^{(0)} x_i^{(0)} + b^{(0)} \,)$$
$$x_i^{(2)} = f(\, A^{(1)} x_i^{(1)} + b^{(1)} \,)$$

Minimize over all weights $A^{(0)}, b^{(0)}, A^{(1)}, b^{(1)}$

Sparsity Penalty

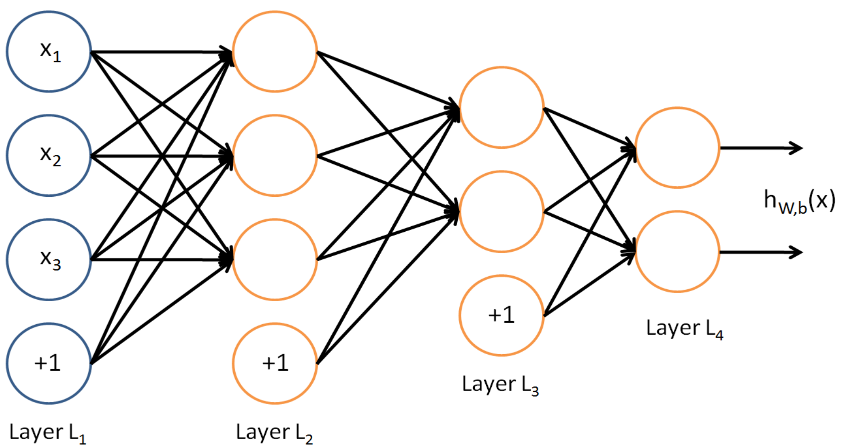
$$\Sigma_i(\| x_i^{(2)} - x_i^{(0)} \|_2^2 + \beta \| x_i^{(1)} \|_1)$$
$$+ \lambda \| A^{(0)} \|_2^2 + \lambda \| A^{(1)} \|_2^2$$

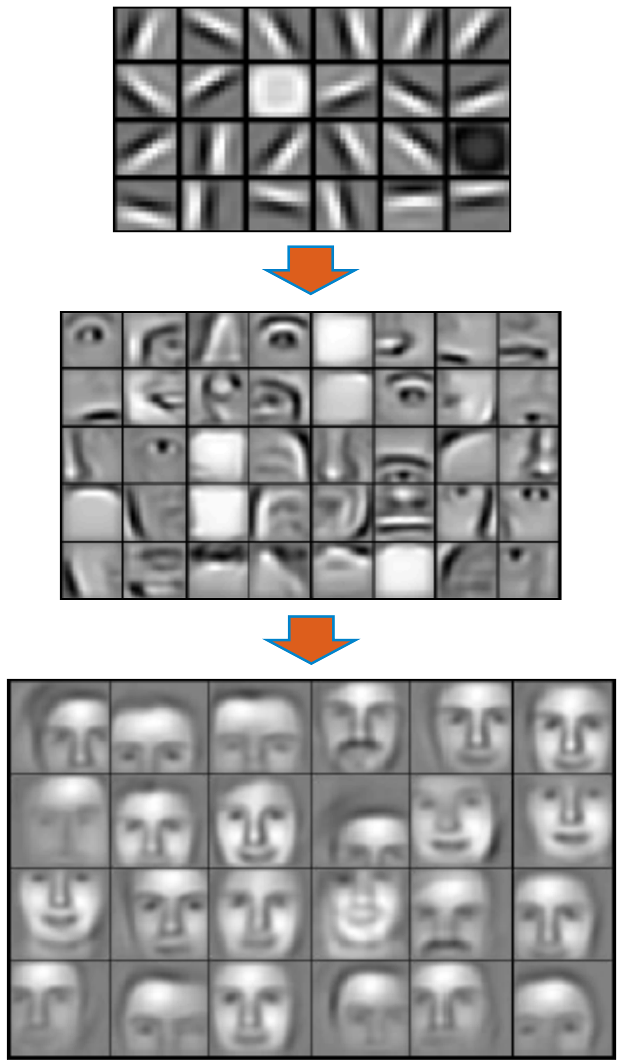
Weight Decay

- Backpropagation algorithm
- L-BFGS method for optimization
- Visualization of learned weights

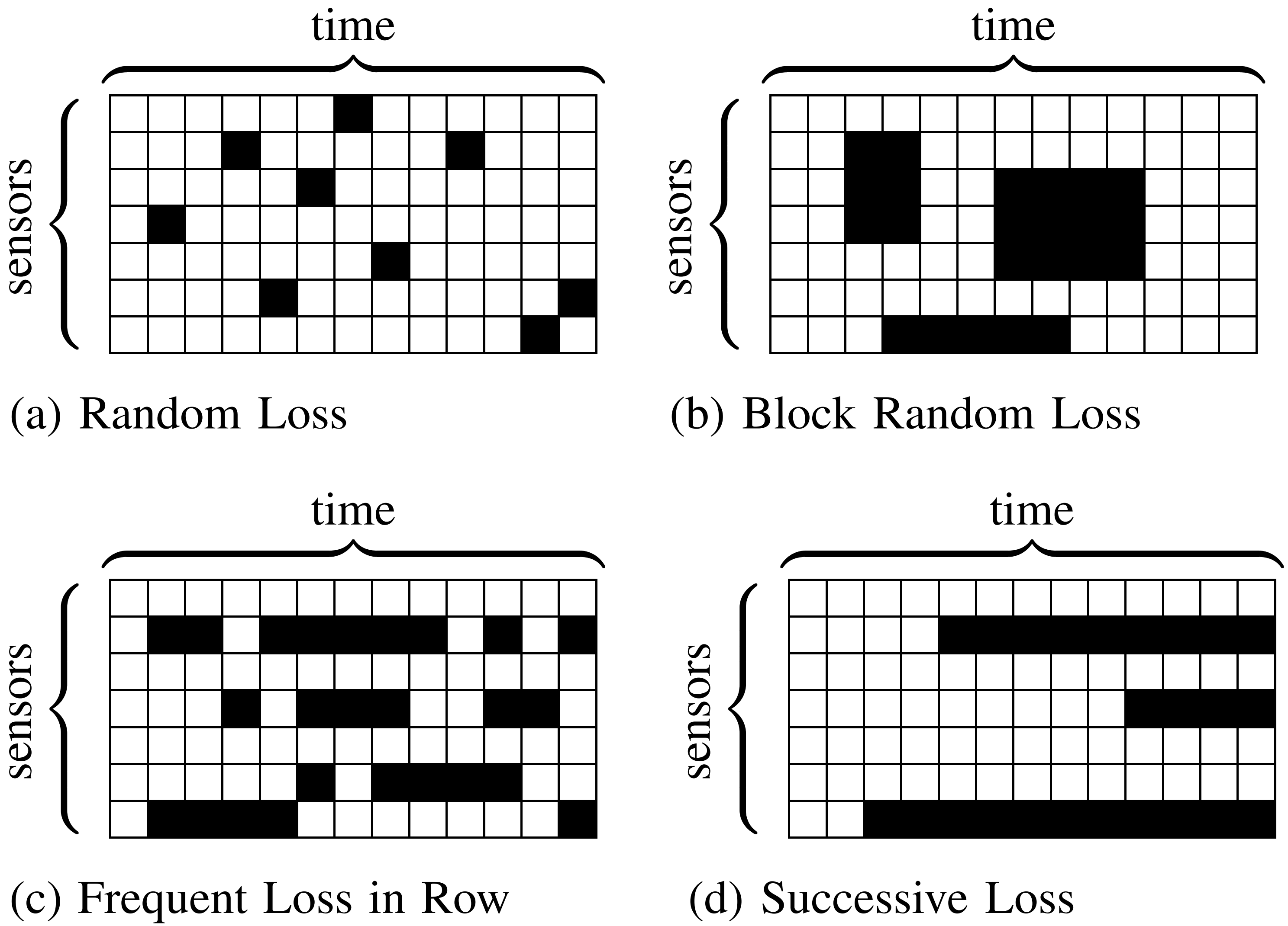
Deep Learning

- Discovers deep features in data
- Greedy layer-wise initialization
  - Learning each layer using contrastive divergence or sparse autoencoders
- Fine tuning of edge weights
  - Backpropagation



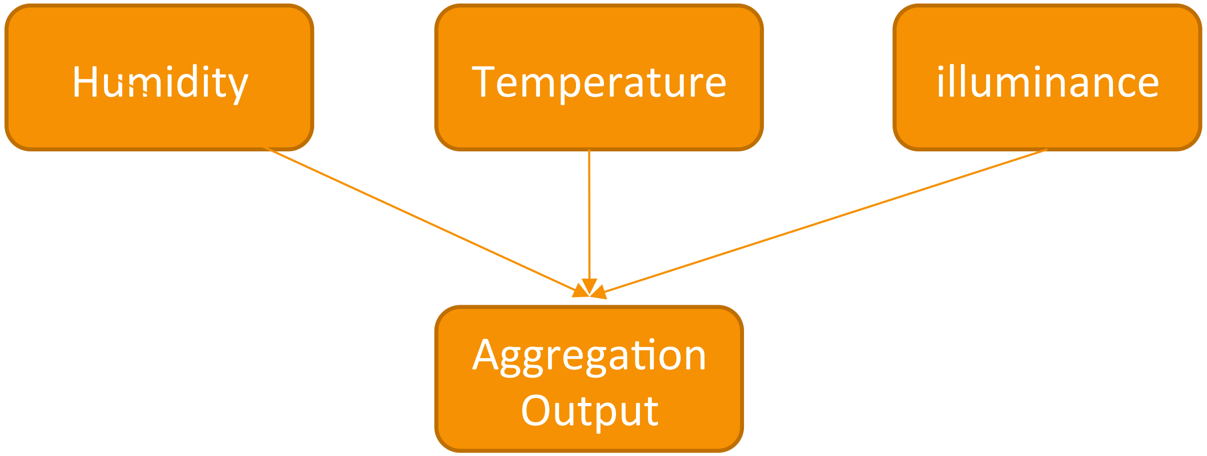


Types of Missing Data



Problem Setting

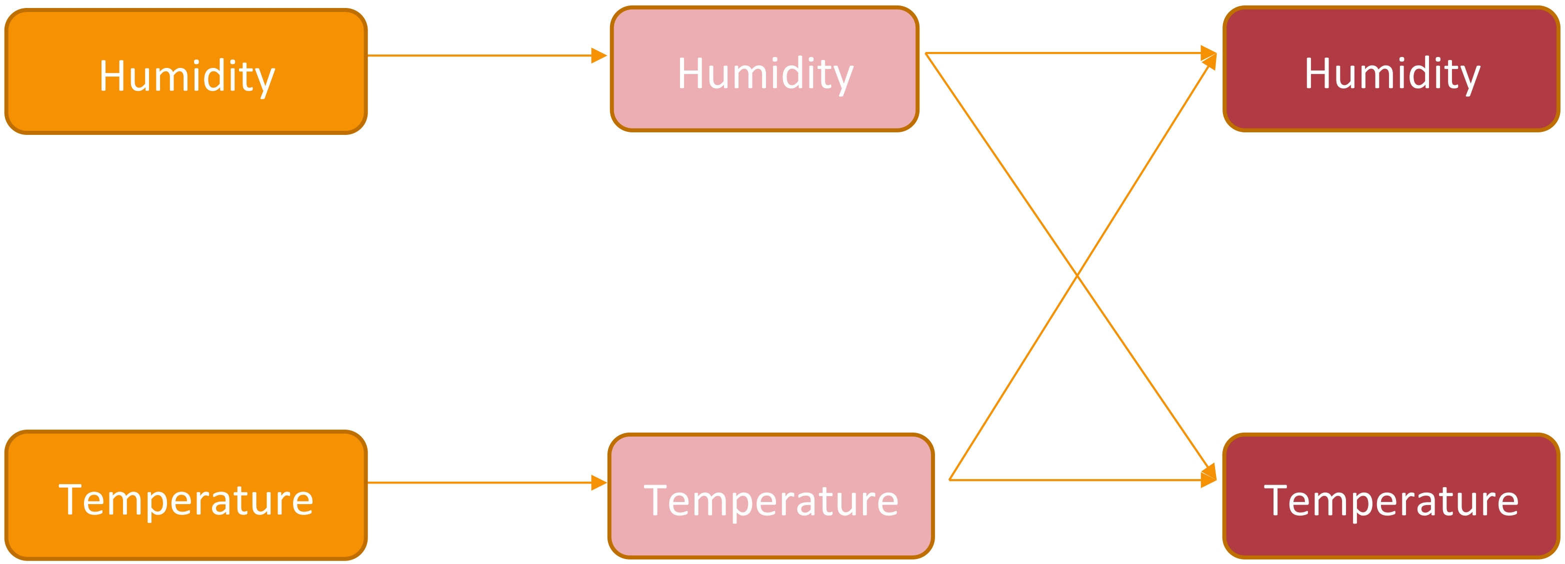
- **Multimodal** sensor data for many tasks
  - ✓ E.g., Human Activity Recognition
  - ✓ E.g., Health Condition, Monitor Environmental Condition, Control Intelligent System



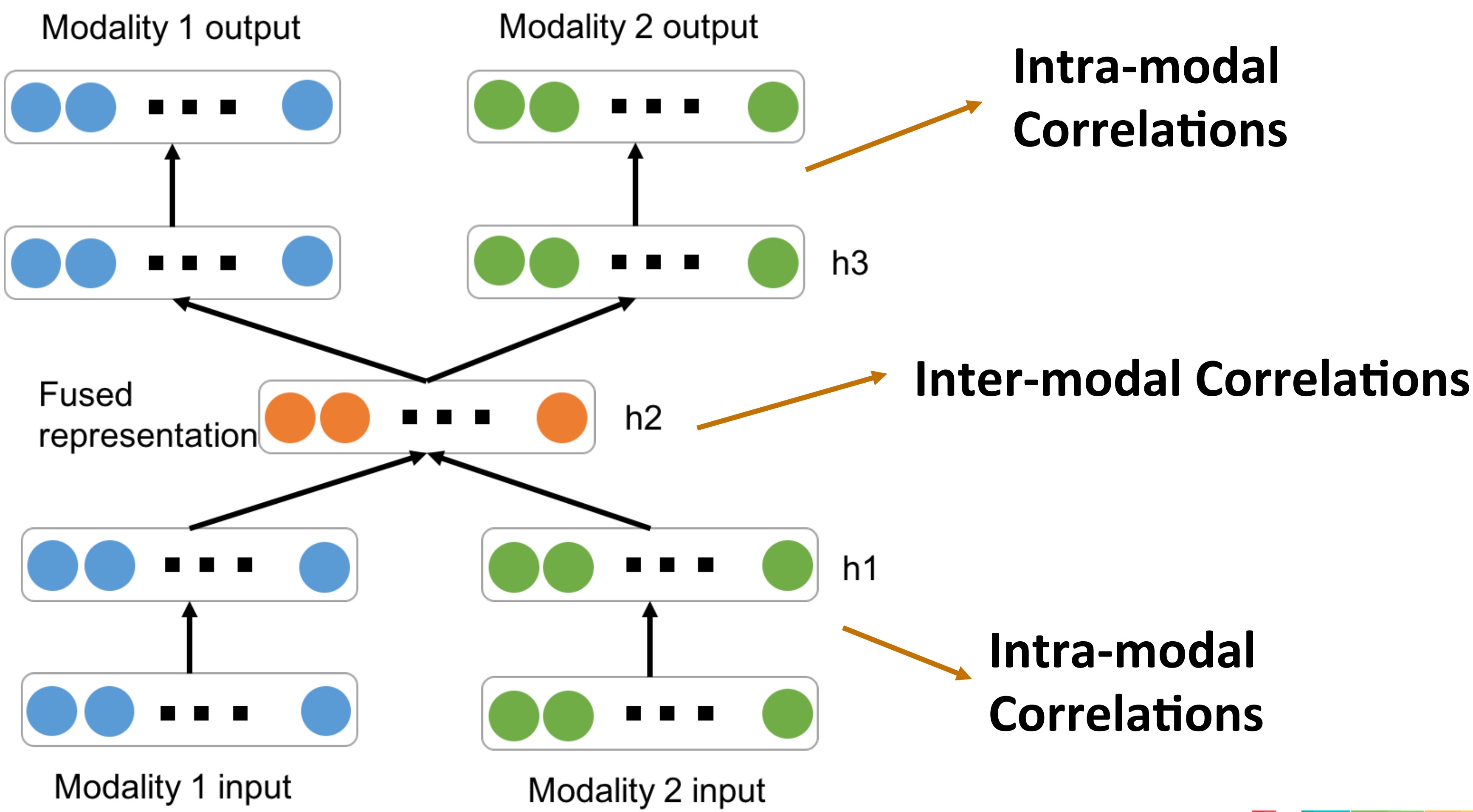
Agriculture Sensor Network

Proposed Method

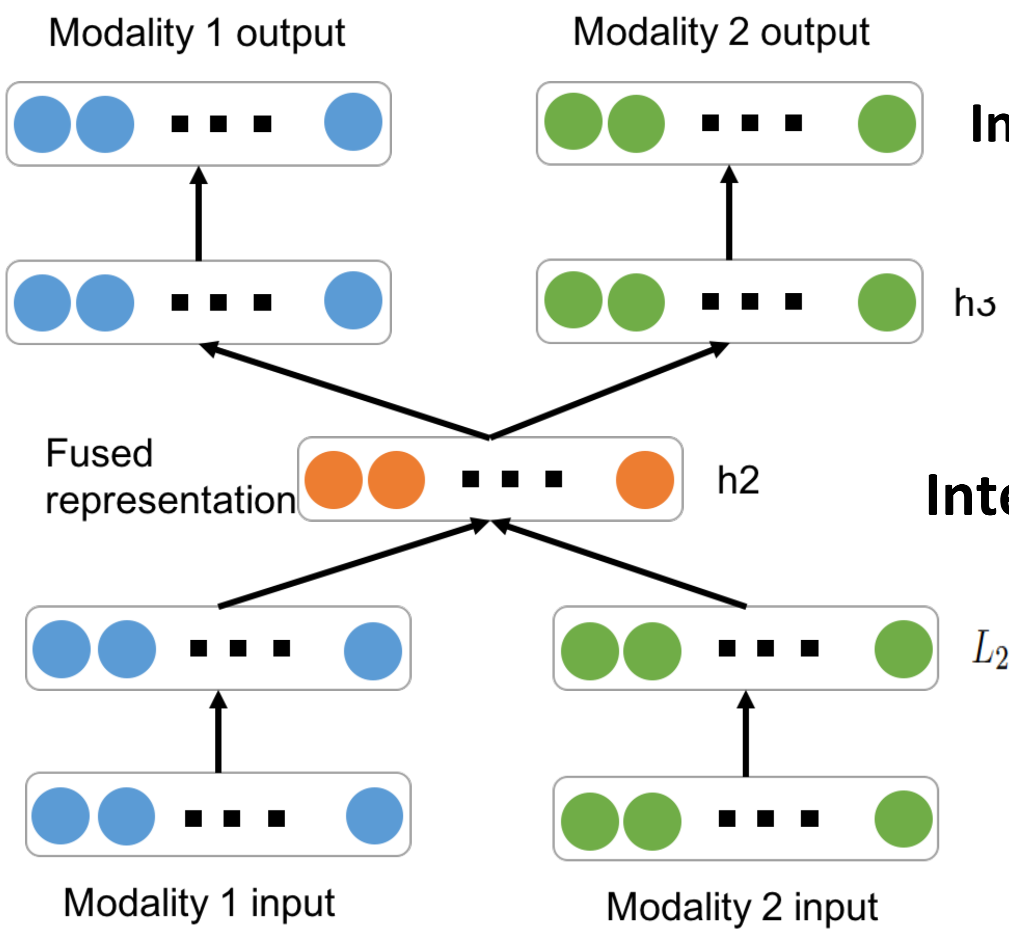
Our model consider both intra-modal and inter-modal correlations:



Deep Multimodal Encoder



Intra- & Inter-modal Correlations



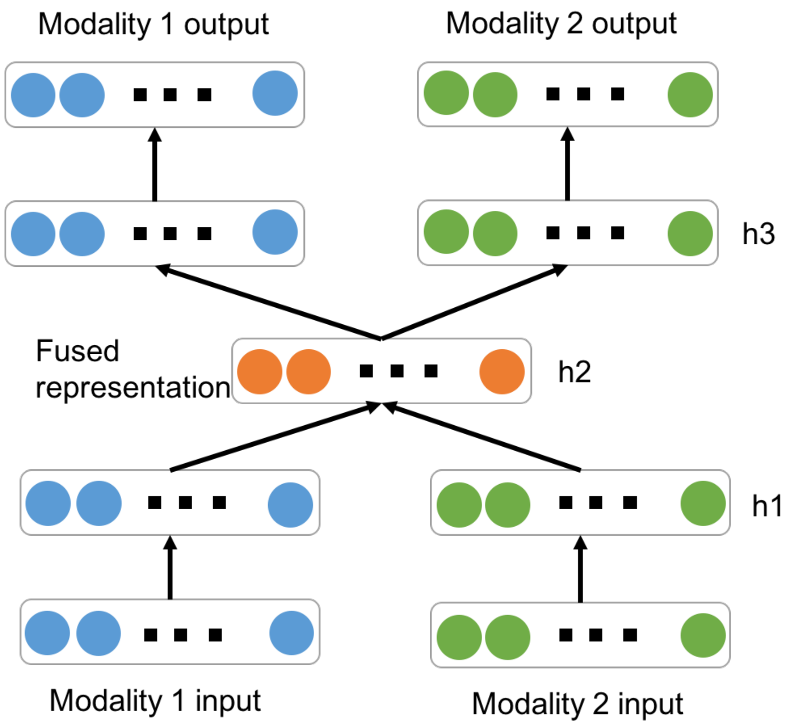
Intra-modal correlations objective:

$$\min L^{xy}(\boldsymbol{X},\boldsymbol{Y}) = L^{x}(\boldsymbol{X}) + L^{y}(\boldsymbol{Y})$$

Inter-modal correlations objective:

$$L_2(H) = \frac{1}{2N}\|\hat{\boldsymbol{H}} - \boldsymbol{H}\|_F^2 + \lambda^{xy}\|\boldsymbol{W}^{xy}\|_F^2 + \beta^{xy}\sum_{j=1}^{m_2} KL(\rho^{xy}\|\hat{\rho}_j)$$

Proposed Objective Function



$$\tilde{J}(\boldsymbol{W}, \boldsymbol{b}) = \frac{1}{2} \| \frac{1}{\mathbb{1}^N \otimes \boldsymbol{\theta}^{x\top}} \cdot (\hat{\boldsymbol{X}} - \boldsymbol{X}) \cdot \boldsymbol{S}^x \|_F^2$$

Only the observed values contribute to the final MSE objective function.

$$\min L^x(\boldsymbol{X}) = \frac{1}{2} \| \frac{1}{\mathbb{1}^N \otimes \boldsymbol{\theta}^{x\top}} \cdot (\hat{\boldsymbol{X}} - \boldsymbol{X}) \cdot \boldsymbol{S}^x \|_F^2 + \lambda^x \| \boldsymbol{W}^x \|_F^2 + \frac{\beta^x}{T^x} \sum_{m=1}^{m_x} \sum_{t=1}^{T^x} KL(\hat{\rho}_{m,t}^x \| \rho^x).$$

$$\hat{\boldsymbol{\rho}}^x = [(\boldsymbol{A}^x)^\top \boldsymbol{S}^x \cdot \frac{1}{\mathbb{1}^{m_x} \otimes \boldsymbol{\theta}^{x\top}}]^\top$$

Add sparse and weight penalty.

$$\min L^{xy}(\boldsymbol{X}, \boldsymbol{Y}) = L^x(\boldsymbol{X}) + L^y(\boldsymbol{Y})$$

Experiment

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN
Established in collaboration with MIT

Dataset:

Agriculture sensor data of Humidity, Temperature, Illuminance

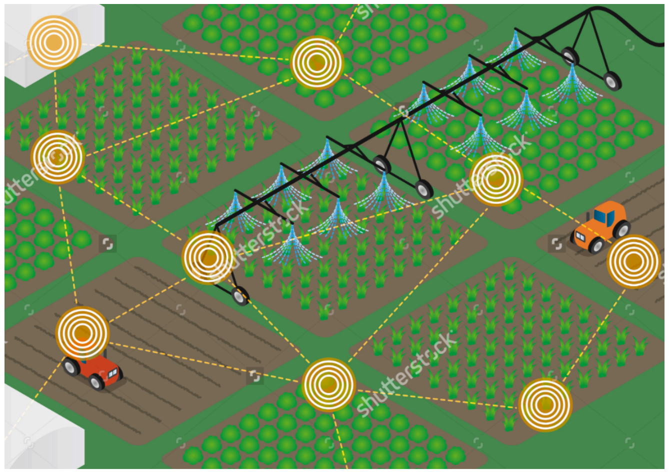
40 sensors for 4 months

3306 data samples each modality, each sample for of dimension $R\uparrow 144$ (for one day).

Training set: 2400; Validation set: 306; Test set: 600

Experiments:

1. Compression and Reconstruction Test
2. Missing value imputation test
3. Generalization Performance Test

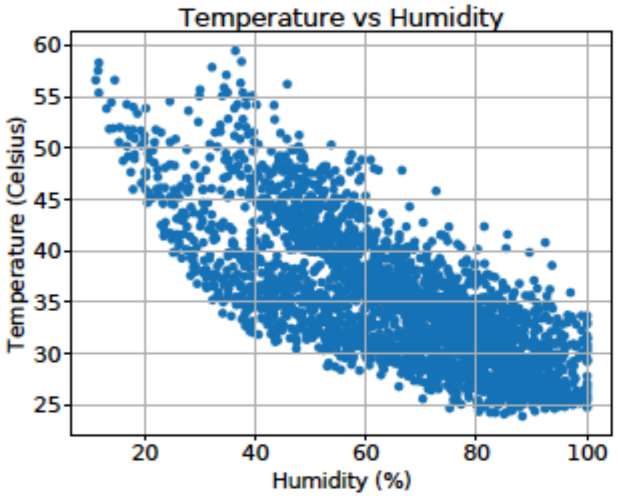


INFORMATION SYSTEMS TECHNOLOGY AND DESIGN

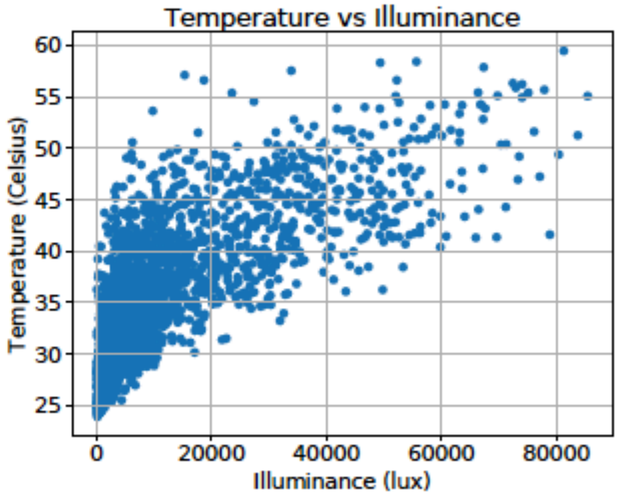
Data Statistics

TABLE I: Dataset Statistics

	Temp.	Hum.	Illum.
Min	21.16	9.58	0
Max	60.95	100.00	98295.30
Lower Quartile	25.90	72.63	0
Median	27.60	84.42	29.68
Upper Quartile	31.28	90.87	2411.29
Standard Deviation	5.03	16.16	6635.97



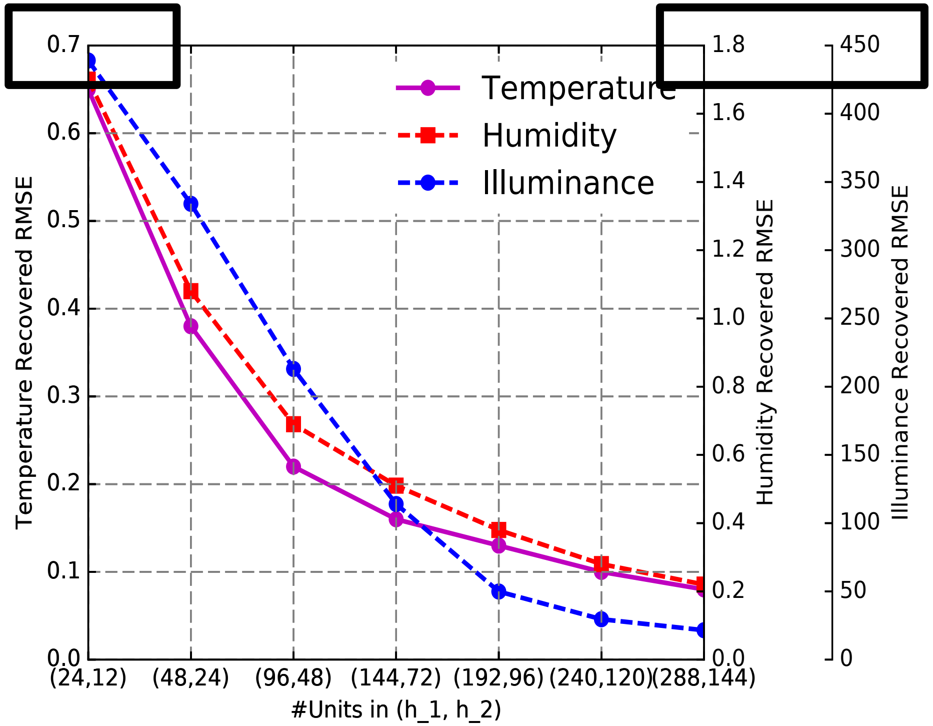
(a) Temp V.S. Hum



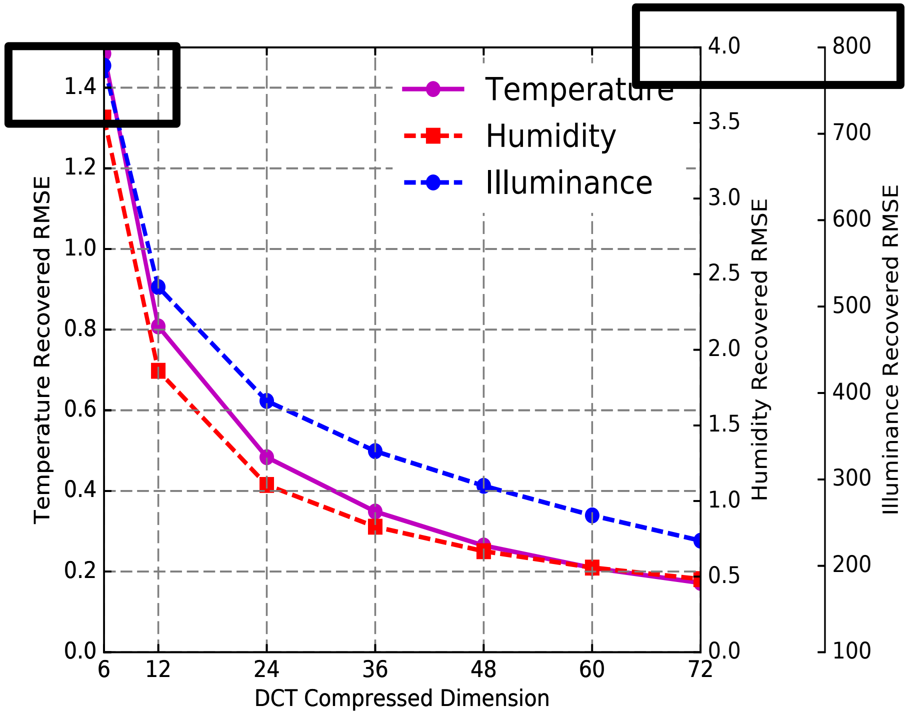
(b) Temp V.S. Illum

Data Compression Efficiency

Deep Multimodal Encoder



Discrete Cosine Transform



Missing Value Imputation

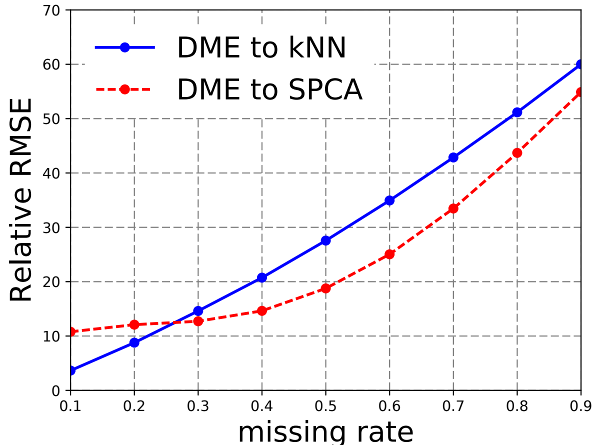
SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN
Established in collaboration with MIT

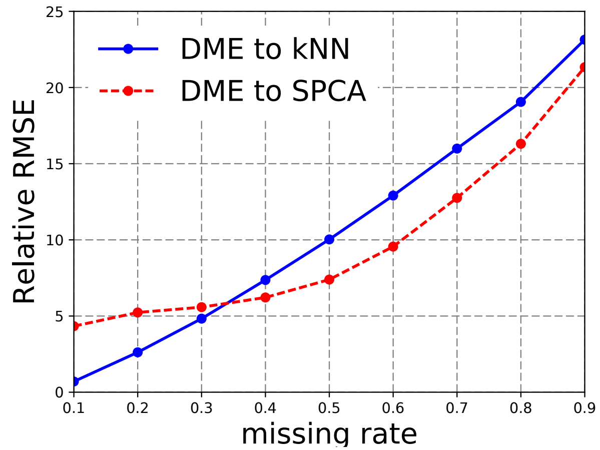
Humidity-Temperature:

	Humidity								Temperature							
Miss rate	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
KNN	8.31	14.56	20.98	27.90	35.41	43.56	52.36	61.8	2.45	4.66	7.12	9.81	12.71	15.83	19.13	22.70
S-PCA	15.45	17.88	19.09	21.79	26.62	33.66	42.97	54.40	6.09	7.28	7.88	8.67	10.08	12.49	15.90	19.96
UAE	5.09	5.98	6.79	7.73	8.13	9.00	10.07	11.06	2.03	2.32	2.48	2.71	2.95	3.10	3.27	3.69
CAE	**4.64**	5.98	6.40	7.36	8.05	8.87	9.75	11.07	**1.73**	2.07	2.32	2.52	2.78	**2.93**	3.28	3.81
DME	**4.64**	**5.79**	**6.37**	**7.15**	**7.85**	**8.62**	**9.50**	**10.69**	**1.73**	**2.04**	**2.29**	**2.45**	**2.68**	**2.93**	**3.14**	**3.65**

Analysis: Less RMSE & Robust performance

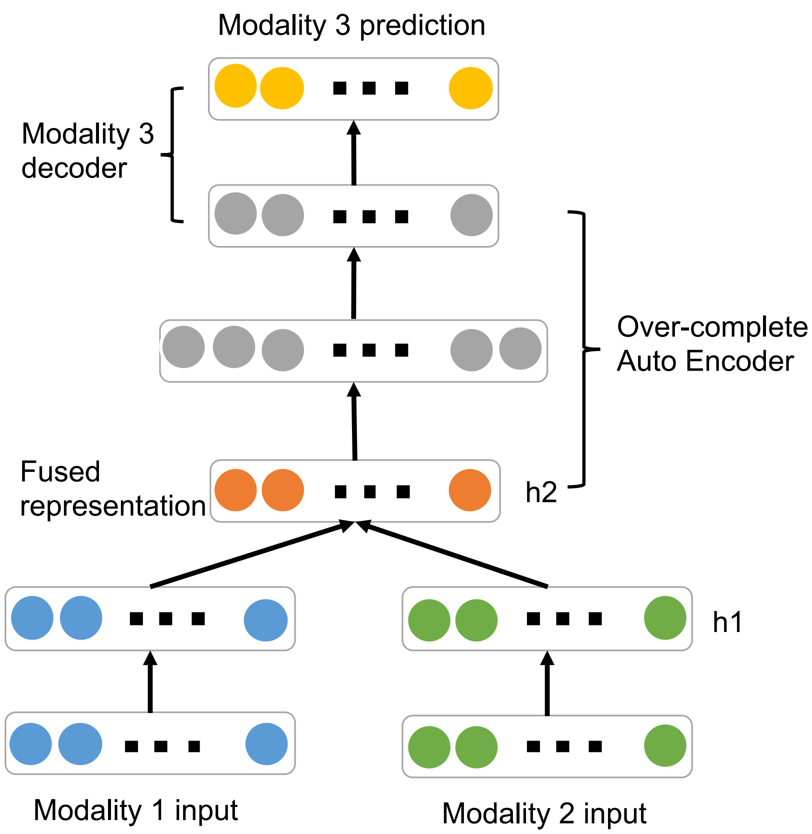
Humidity-Temperature Relative RMSE

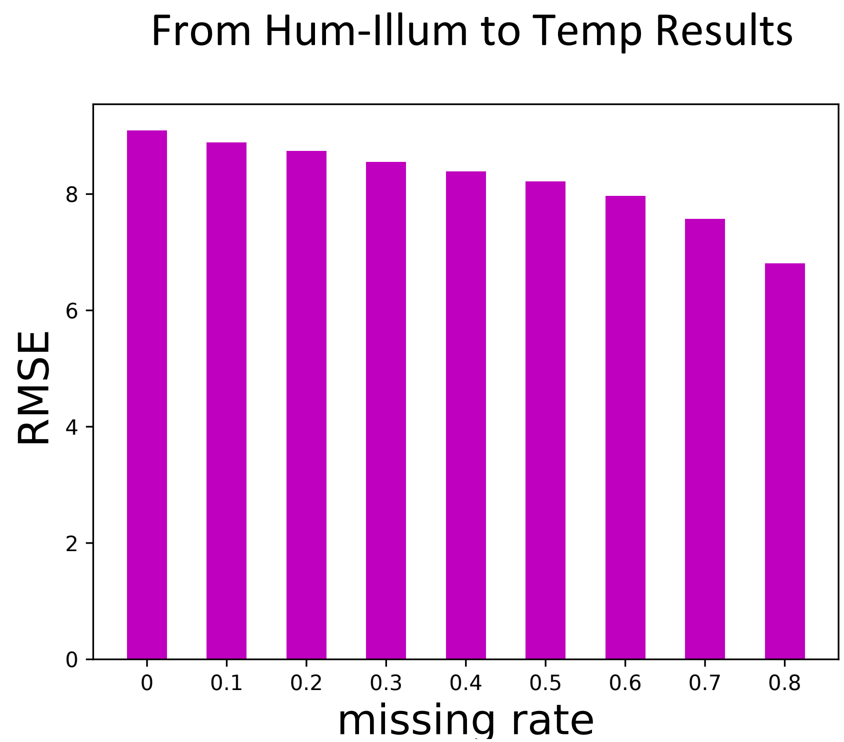




INFORMATION SYSTEMS TECHNOLOGY AND DESIGN

New Modality Prediction





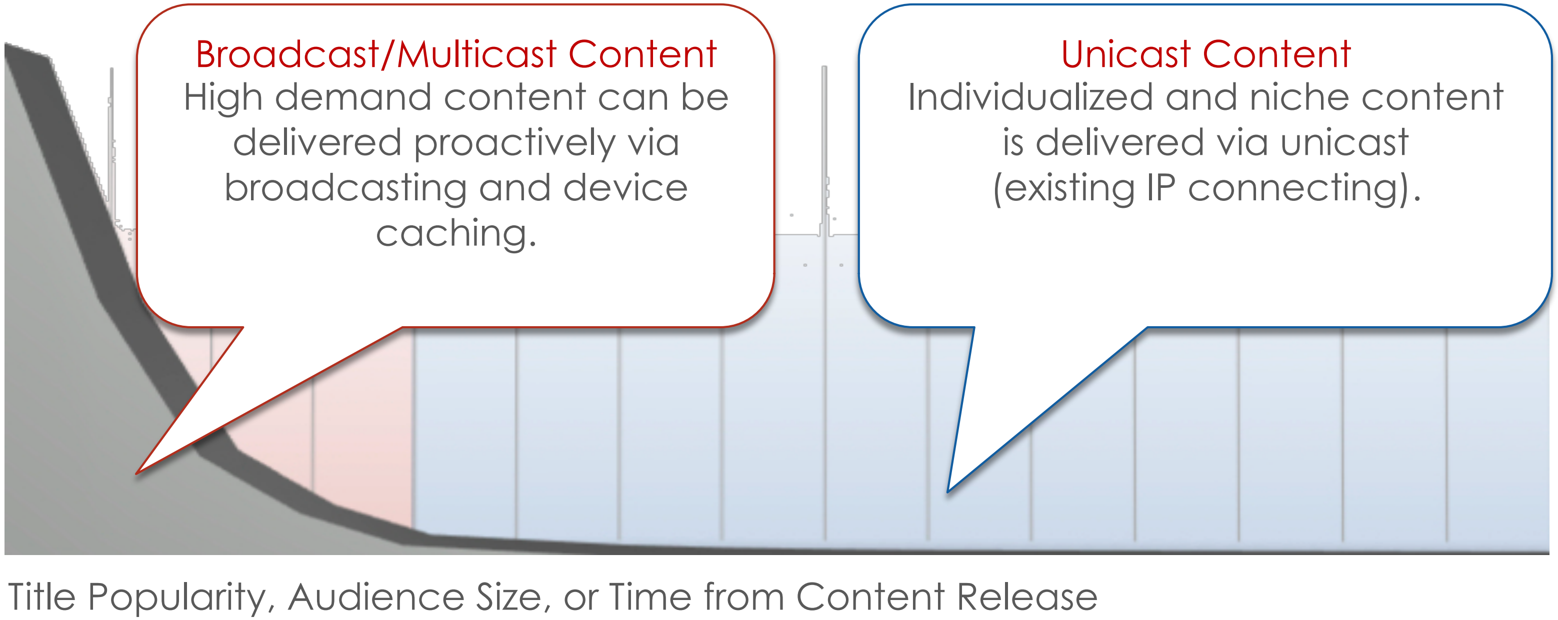
Construct a new modality with reasonable errors

Summary

- Applied neural networks to sensor data processing for fog computing applications
- Trained a multimodal framework together with a novel objective function
- Unsupervised learning for missing data imputations
- Demonstrated the performance with lots of experiments
- Application perspective: Decompose the hierarchical framework

Content-Aware Proactive Caching

Content-Centric Networks



- To exploit rich **storage** and **computing** resources
- Multicast/broadcast: transmit common content to multiple users on the same resource block
- Caching: bring contents closer to users by pre-fetching contents during off-peak time

Motivation

- ✓ Few works consider a complete procedure of video popularity prediction and cache replacement.
- ✓ Most of the caching works deal with videos after being published.
- ✓ Popularity distribution varies → the training set is required to be updated together with newly uploaded videos.
- ✓ Some information is not available for new videos.

System Model

- A macro-cell based station (BS) connecting to a server through a limited backhaul link.
- BS serves a set of K users.
- Server originally contains a set of (old) videos with available statistical information.
- Time is divided into frames with the same duration.
- Server updates new videos at the beginning of every frame (with ratio R = new/old).

Average Backhaul Load

- s_i : size of video i.
- $p_{i,t}$: popularity of video i at time t.
- $\alpha_{i,t}$ portion of video i cached at time t.
- The average load on the backhaul link:

$$\sum_{i=1}^{V_t} \left(1 - \alpha_{i,t}\right) s_i p_{i,t}$$

Total number of videos in the server at time t

Problem Formulation

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN
Established in collaboration with MIT

- Minimize the average backhaul load

$$\min_{\alpha_{i,t}} \sum_{i=1}^{V_t} \left(1 - \alpha_{i,t}\right) s_i p_{i,t}$$

subject to cache utility as follows:

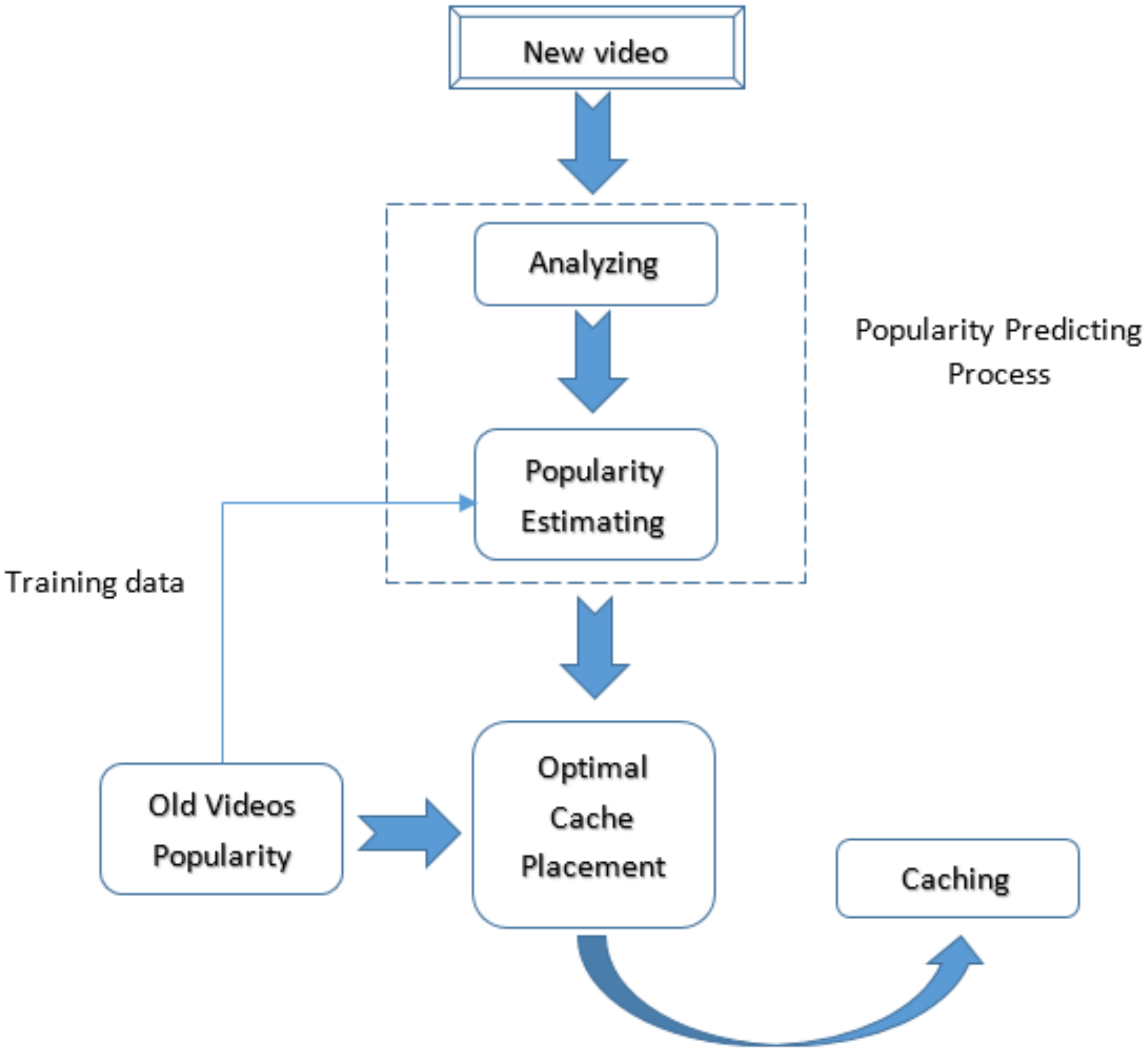
$$\sum_{i=1}^{V_t} \alpha_{i,t} s_i = M$$

Cache capacity

$$0 \le \alpha_{i,t} \le 1$$

INFORMATION SYSTEMS TECHNOLOGY AND DESIGN

Joint Popularity Prediction & Cache placement



Popularity Prediction

- Collaborative filtering (user & item-based)
  - Focus on relationship between users & items.
  - Similarity between item 1 & 2 is determined by the rating given from users who rated both.
- Content-based filtering
  - Focus on properties of items.
  - Similarity between items is determined by the similarity between their properties.

Features Extraction

Raw features extraction step is done using 3D-CNN.

- CNN has a breakthrough in an image domain.
- Advantages:
  - Can extract sophisticated features.
  - Features are condensed into a vector which is machine-readable form.
  - Not require human assist.

Features Mapping to Probability

Feature vector conversion step.

- Raw feature vectors are mapped to G-dimensional space using SVM.
- Probability is obtained by solving

Prob. that an observation belongs to class i

$$p_i^c = \sum_{j:j\neq i} \left(\frac{p_i^c + p_j^c}{G-1} \right) \mu_{ij}, \forall i$$

$$\text{s.t.} \sum_{i=1}^{G} p_i^c = 1, p_i^c \geq 0, \forall i$$

Prob. of belonging to i given i and j only.

Popularity Prediction Model

- By ignoring the noise, the long-term popularity has a strong linear correlation to the early one.

$$p_i(t) = m(t_i^r, t)\, p_i(t_i^r) + \varepsilon_i(t)$$

Parameter presenting the linear relationship

Noise RV with 0 mean

Popularity of video i at time t.

Reference time (after publishing)

Popularity Score

- Each user is assign a vector of preference

$$\zeta^i = (\zeta_1^i \ \zeta_2^i \ \ldots \ \zeta_G^i)$$

- Each element is a popularity score (PS) that user gives to a specific VC.
- Sum of elements is the same fore every user.
- PS of a VC is the total PS from every user.
- PS of a video is the weighted sum of PS of all the VCs, where weights are elements of it feature vector.

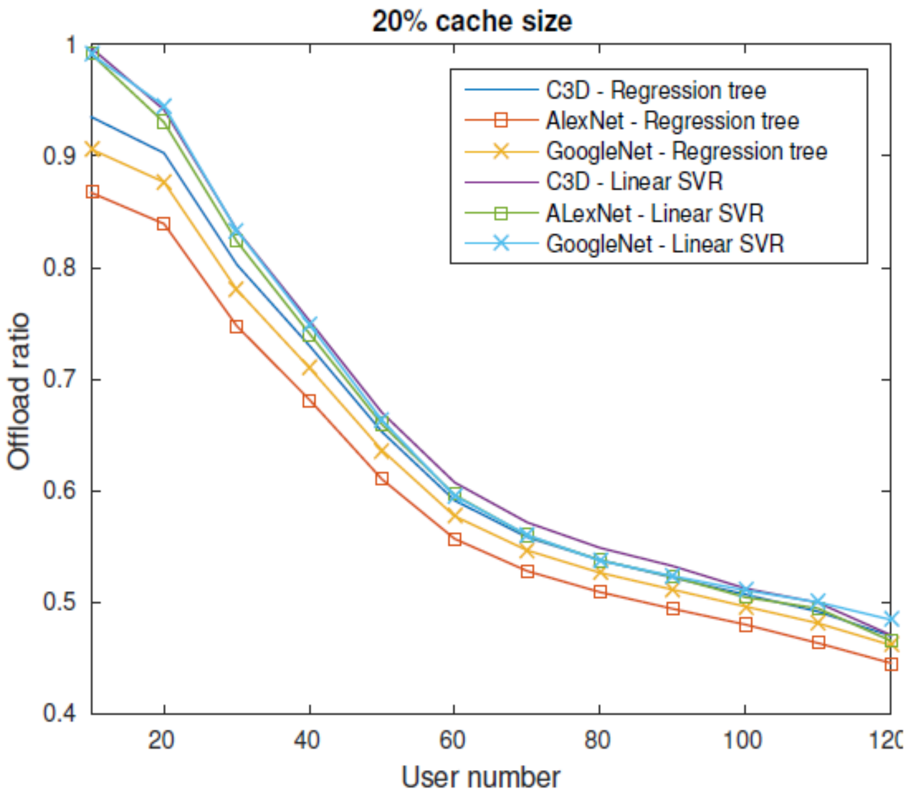
Experiments

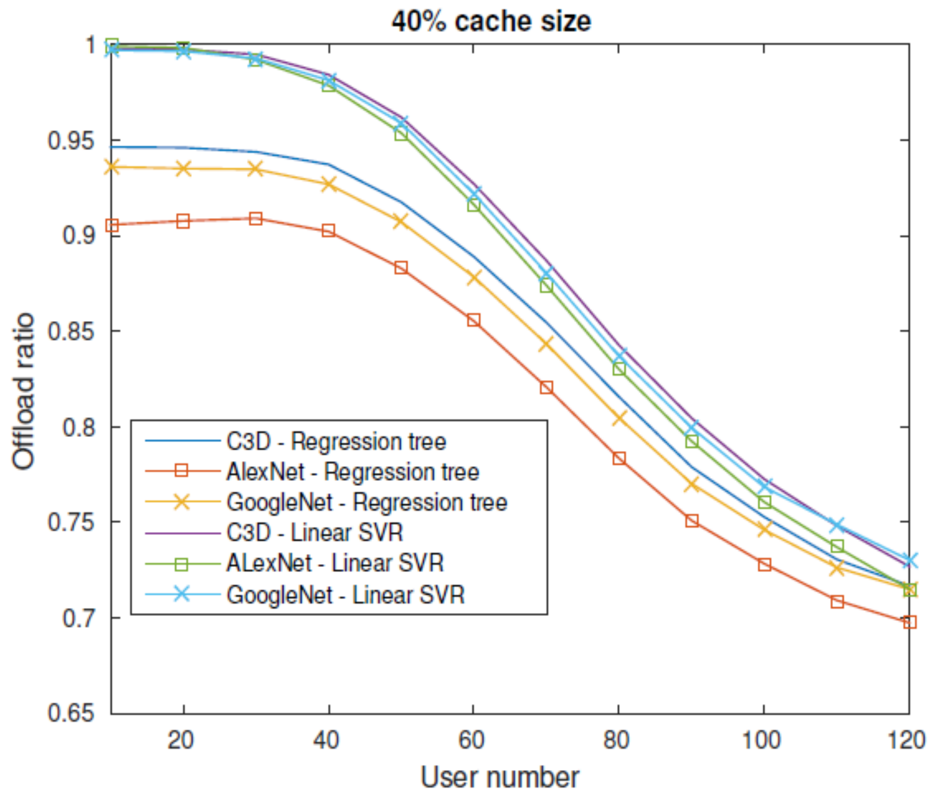
- Dataset = YUPENN + UCF101
- Totally, G = 115 VCs and 13740 videos
- Video size = 1 for […]ry video
- Ea[…] VC[…]i.e.[…]e element in his pr[…]antl[…]er than the others.

420 videos of 14 types

13320 videos of 101 types

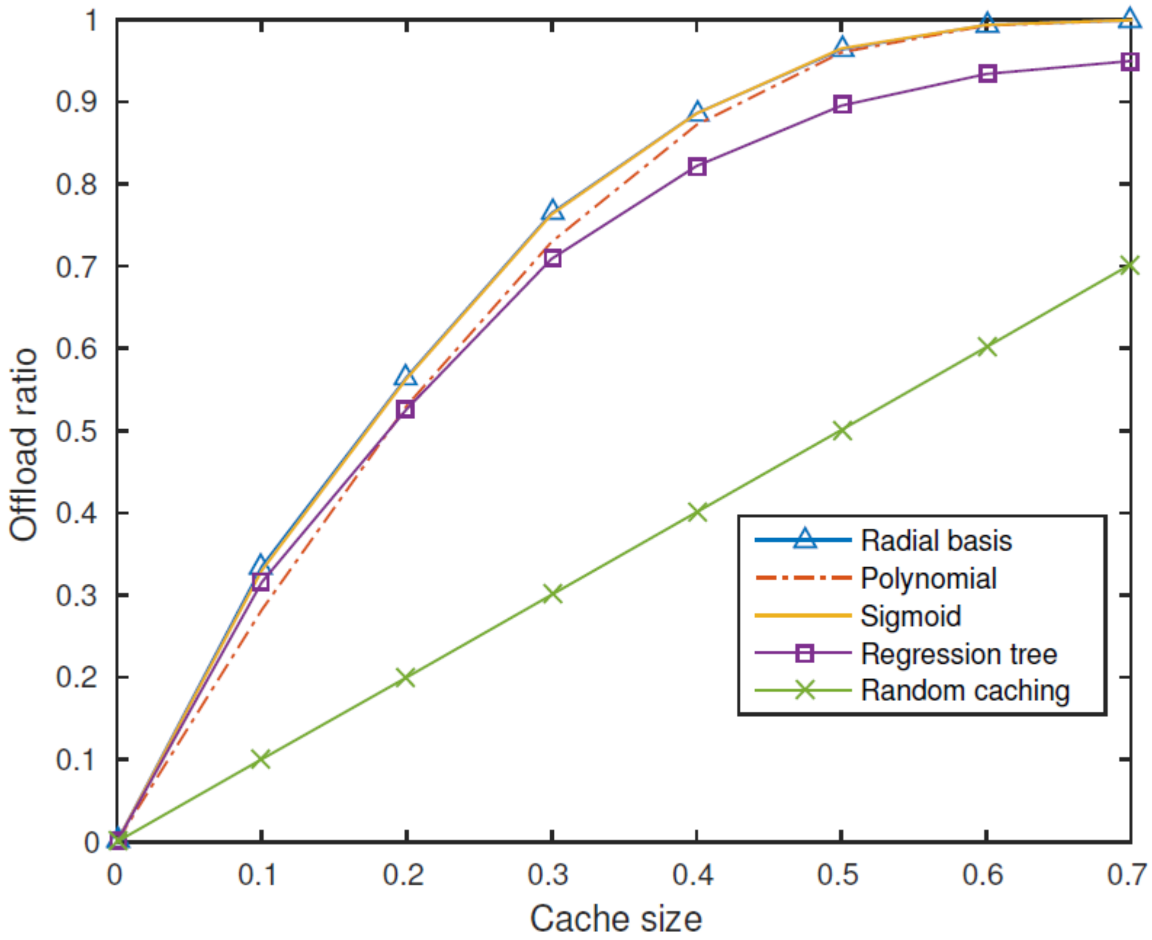
Proactive Caching - Offload Ratio





Offload ratio with respect to the load variation (adjusted by the user number) under small and large cache capacity conditions.

Popularity Prediction



Prediction accuracy of different regression models in comparison to the baseline of random caching (no prediction is required)

Summary

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

- Build a popularity predicting – caching framework to make cache placement decision before videos are published based on extracted features.
- Investigate the performance of different CNN types and prediction models in caching.
- Study the joint effect from the user's preference intensity and the number of users as well as the influence from the cache capacity.

INFORMATION SYSTEMS TECHNOLOGY AND DESIGN

Network Economics of Fog Computing

Jianwei Huang

Network Communications and Economics Lab (NCEL)

Department of Information Engineering

The Chinese University of Hong Kong (CUHK)

Fog Computing

- Rely on close collaboration of end-user clients or near-edge devices
- Carry out a substantial amount of storage, communication, control, configuration, measurement, and management
- Many exciting technology challenges and opportunities

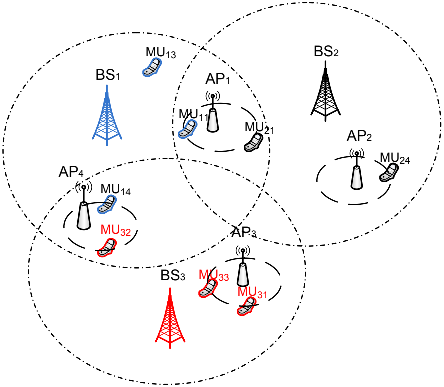
How to Make Fog Computing Successful?

- Solving technology challenges alone is not enough
- Addressing economic and incentive issues are critical for the success
  - Why should users and devices collaborate?
  - How to compensate for the costs due to collaboration?
  - How to share the benefits of collaboration?
  - How to make the algorithms distributed, fair, and robust?

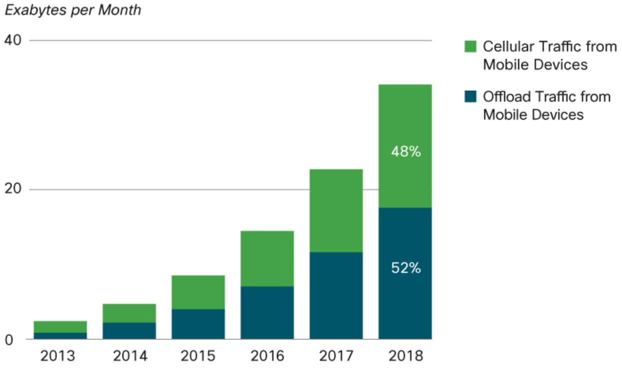
Three Case Studies

- Mobile data offloading
  - Coordination among cellular base stations and Wi-Fi APs
- Crowdsourced mobile video streaming
  - Distributed resource sharing for QoS-sensitive applications
- Mobile crowdsensing
  - Diversity-driven social-aware computing

Mobile Data Offloading



Trend of Data Offloading



Mobile Traffic Offloading Prediction (source: Cisco VNI Mobile 2014)

- Mobile offloading will increase from 45% in 2013 to 52% in 2018

Fast On-Demand Data Offloading

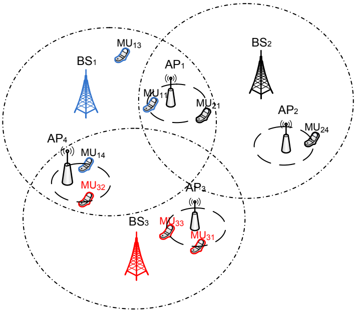
- Mobile Operators (MOs) do not own WiFi Access Points (APs).
- Private APs are already there, owned by business and personal owners.
- MOs will lease APs for on-demand offloading.
- Real-time decisions to catch up the demand fluctuations

Key Problems

- From the MO's Perspective: How much traffic should each MO offload to each AP, and how much to pay?
- From the AP owner's Perspective: How much traffic should each AP offload for each MO, and how much to charge?

System Model

- Each MO is represented by one Base Station (BS)
- $\mathcal{M} \triangleq \{1, ..., M\}$: the set of BSs
- $\mathcal{I} \triangleq \{1, ..., I\}$: the set of APs



Example: $\mathcal{M} = \{1, 2, 3\}$ and $\mathcal{I} = \{1, 2, 3, 4\}$.

For Each BS $m \in \mathcal{M}$

- x_{mi}: offloading request to AP i
- $\boldsymbol{x}_m \triangleq (x_{mi}, \forall i \in \mathcal{I})$: offload request vector to all APs
- $J_m(\boldsymbol{x}_m)$: the utility function of BS m
  - Positive, increasing, and concave

For Each AP $i \in \mathcal{I}$

- y_{im}: offload admission for BS m
- $\boldsymbol{y}_i \triangleq (y_{im}, \forall m \in \mathcal{M})$: offload admission vector for all BSs;
- $V_i(\boldsymbol{y}_i)$: the cost function of AP i
  - Positive, increasing, and convex.
- C_i: capacity constraint

A Benchmark Problem

Social Welfare Maximization (Efficiency)

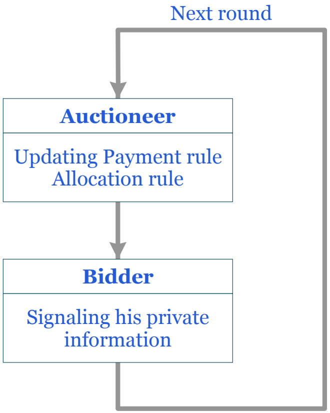
maximize	$\sum_{m \in \mathcal{M}} J_m(\boldsymbol{x}_m) - \sum_{i \in \mathcal{I}} V_i(\boldsymbol{y}_i)$	......*Social Welfare*
subject to	(i) $\sum_{m \in \mathcal{M}} y_{im} \leq C_i, \ \forall i \in \mathcal{I},$	......*Capacity constraint*
	(ii) $x_{mi} = y_{im}, \ \forall m \in \mathcal{M}, i \in \mathcal{I},$	......*Feasibility*
variables	$\boldsymbol{x}_m, \boldsymbol{y}_i, \forall m, \forall i.$	

Local Interests with Private Information

- BSs and APs want to optimize their own payoffs (not social welfare)
- Both utility functions and cost functions are private information
- Solution: incentive mechanism design

Iterative Double Auction (IDA)

- Conducts multiple rounds of double-auction
- Bidders: BSs and APs



Design Principles of IDA

Step 1: Rules

Socially Optimal Rules

$\Uparrow$

Step 2: Bidding

Individual Optimal Bids $\Longleftrightarrow$ Socially Optimal Bids

$\Uparrow$

Step 3: Allocation

Auctioneer Optimal KKT $\Longleftrightarrow$ Socially Optimal KKT

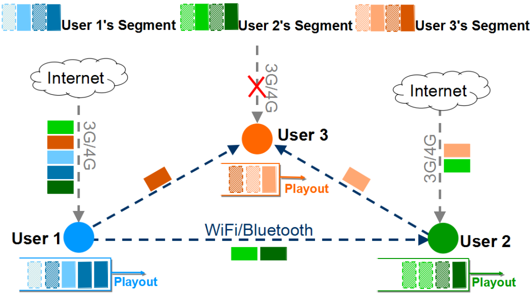
IDA - Properties

Properties of IDA

Under properly designed auction rules, the IDA is

- Efficient
  - Achieves the social welfare maximization;
- Weakly Budget Balanced
  - Does not require money input from the auctioneer;
- Incentive Compatible
  - Incentivizes all bidders to reveal information truthfully;
- Individually Rational
  - Offers all bidders non-negative payoffs.

Crowdsourced Mobile Video Streaming



Single-User Video Streaming

My downloading speed is 0.5Mbps, want to watch video.

A

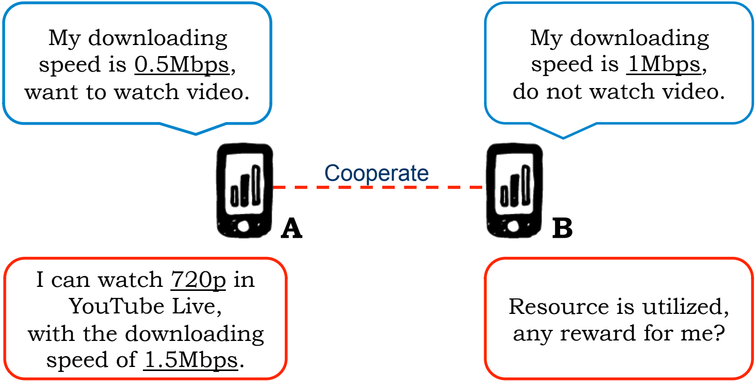
I can watch 240p in YouTube Live, with the downloading speed of 0.5Mbps.

My downloading speed is 1Mbps, do not watch video.

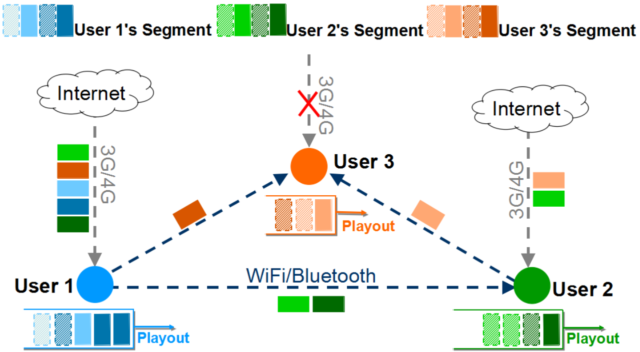
B

My resource is idle.

Multi-User Cooperative Video Streaming

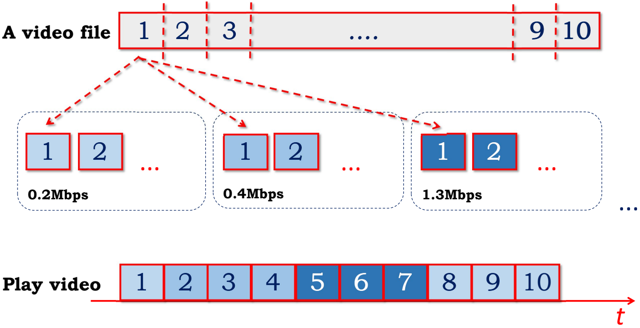


Crowdsourced Mobile Video Streaming



- Crowdsource network resources from multiple near-by mobile users from potentially different service providers.
- Each mobile user watches a different video.

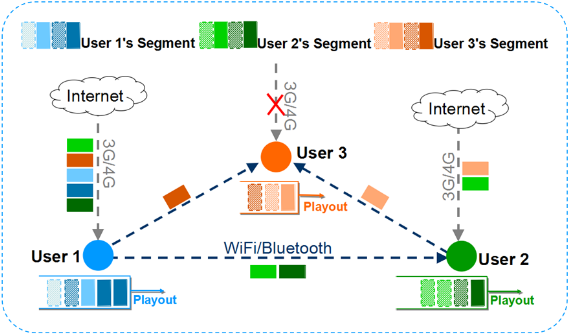
Adaptive BitRate Streaming



- To achieve flexible Quality of Experience in wireless video streaming
- Single user case: choose the bitrate of each video segment based on real-time network conditions and user QoE preferences.

Multi-User Collaborative Video Streaming

- Three decisions when downloading a video segment



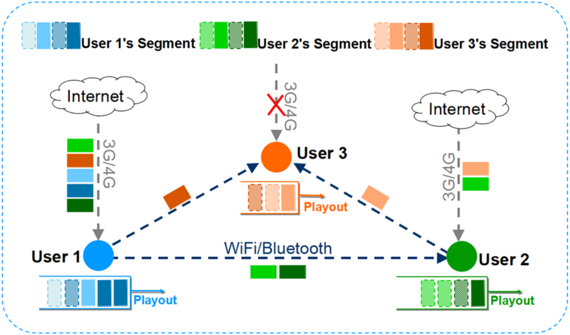
Receiver Selection:
Whose segment ?

Bitrate Adaptation
What bitrate ?

Cost Compensation
How much to pay ?

Multi-User Collaborative Video Streaming

- Three decisions when downloading a video segment



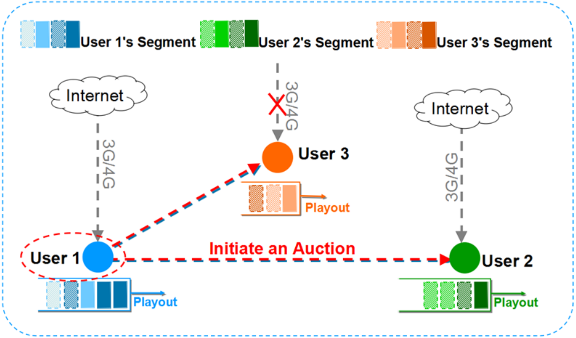
Receiver Selection: Whose segment ?

Bitrate Adaptation What bitrate ?

Cost Compensation How much to pay ?

- Need decentralized and asynchronous algorithm without complete network information

Auction-Based Incentive Mechanism

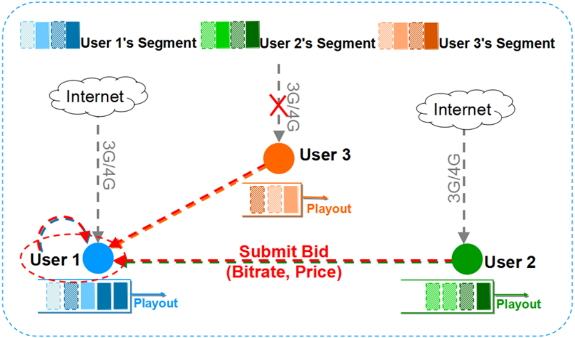


User 1
(Ready to download)
Initiate an auction

User 1, 2, 3
Submit bid with (bitrate, price)
-Bitrate Adaptation

User 1
Winner & Payment
-Receiver Selection
-Cost Compensation

Auction-Based Incentive Mechanism

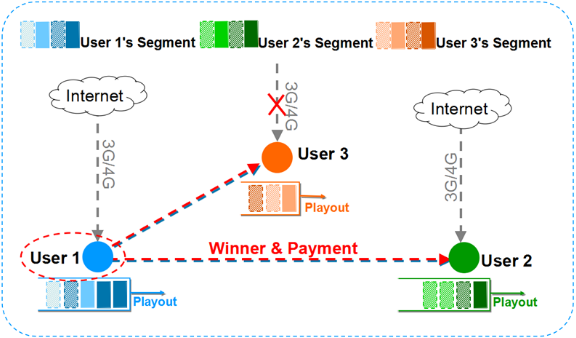


User 1
(Ready to download)
Initiate an auction

User 1, 2, 3
Submit bid with (bitrate, price)
-Bitrate Adaptation

User 1
Winner & Payment
-Receiver Selection
-Cost Compensation

Auction-Based Incentive Mechanism



User 1
(Ready to download)
Initiate an auction

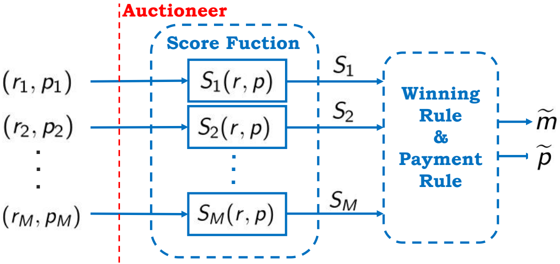
User 1, 2, 3
Submit bid with (bitrate, price)
-Bitrate Adaptation

User 1
Winner & Payment
-Receiver Selection
-Cost Compensation

Challenge: Multi-Dimensional Bids

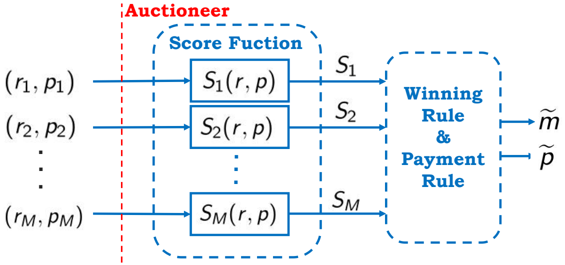
- Each bid is multi-dimensional: (bitrate, price)
  - (0.2Mbps, 20¢) vs. (0.4Mbps, 35¢) vs. (1.3Mbps, 70¢)
- How to rank vectors to decide the winner and the payment?
- Solution: Second Score Auction

Score Function



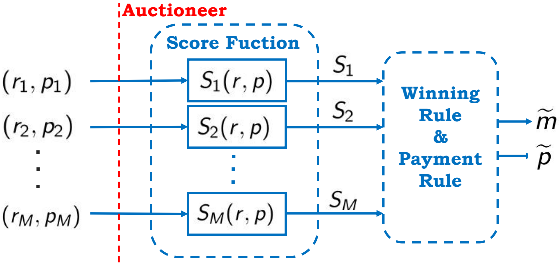
- Score function: transforms a multi-dimensional bid to a scalar
  - Determined by the auctioneer (mechanism design)
  - Each user m can have a unique score function $S_m(r, p)$

Score Function



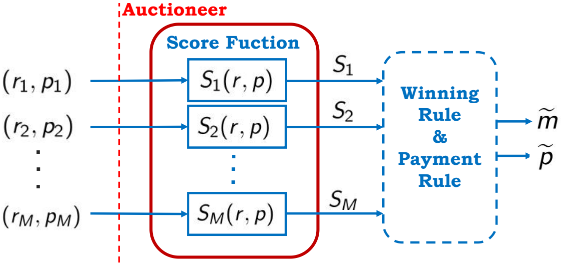
- Score function: transforms a multi-dimensional bid to a scalar
  - Determined by the auctioneer (mechanism design)
  - Each user m can have a unique score function $S_m(r, p)$
- Winner: bidder with the highest score
- Payment: determined by the second highest score

Score Function



- Score function: transforms a multi-dimensional bid to a scalar
  - Determined by the auctioneer (mechanism design)
  - Each user m can have a unique score function $S_m(r, p)$
- Winner: bidder with the highest score
- Payment: determined by the second highest score
- How to choose the score function?

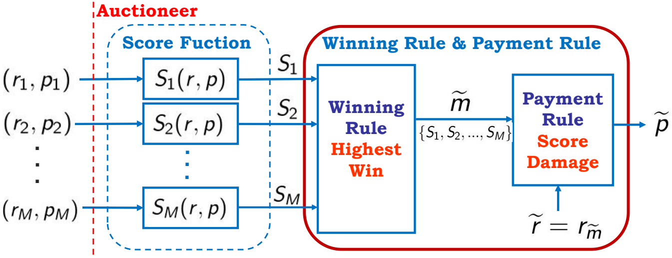
Additive Score Function



$$S_m(r, p) = p - C_n(r)$$

- Difference between the bidder m's price and the downloader n's cost
- All bidders have the same score function (related to downloader n)

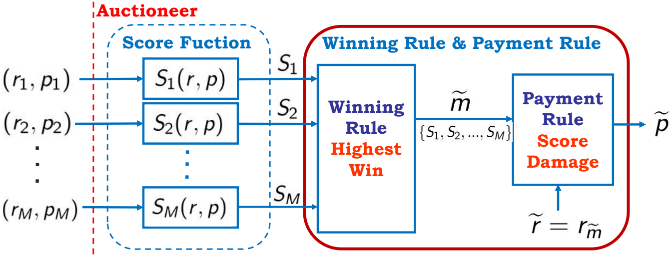
Winner Selection and Payment Determination



- **Winner** = the bidder with the highest score

$$m^* = \arg\max_{m \in \mathcal{N}_n} (p_m - C_n(r_m))$$

Winner Selection and Payment Determination

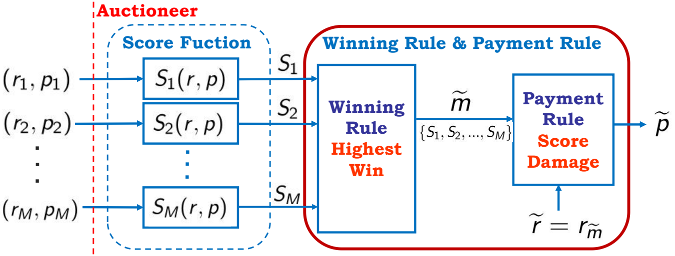


- **Winner** = the bidder with the highest score

$$m^* = \arg\max_{m \in \mathcal{N}_n} (p_m - C_n(r_m))$$

- Winner's **bitrate** = the winner's bid bitrate r_{m^*}

Winner Selection and Payment Determination



- **Winner** = the bidder with the highest score

$$m^* = \arg\max_{m \in \mathcal{N}_n} (p_m - C_n(r_m))$$

- Winner's **bitrate** = the winner's bid bitrate r_{m^*}
- Winner's **payment** = the score damage to other users

Property of the Auction

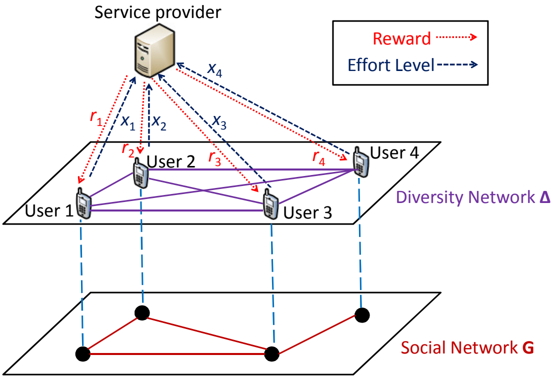
Theorem (Efficient Auction)

Under the following score function

$$S_m(r, p) = p - C_n(r),$$

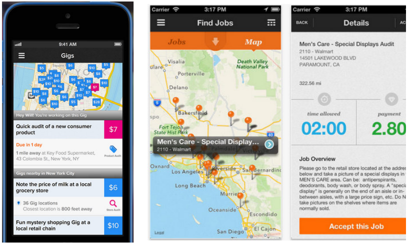
the auction is efficient as it maximizes the social welfare.

Mobile Crowdsensing

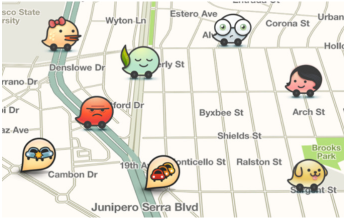


Mobile CrowdSensing (MCS)

- Smartphones with a rich set of embedded sensors:
  - Camera, microphone, GPS, accelerometer, ...
- A large number of individuals using their smartphones to collectively extract and share information:
  - Mobile market research, traffic monitoring, ...



©Gigwalk & Field Agent



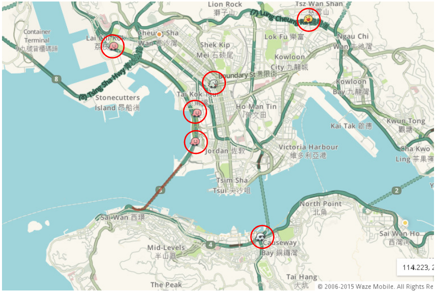
©Waze

Social Effect

- Social effect: an extra incentive to participate in the collaboration.
  - Wave: Allows users to connect accounts with Facebook or Twitter.
  - Gigwalk: Encourage users to share experiences with friends via Facebook or Twitter.

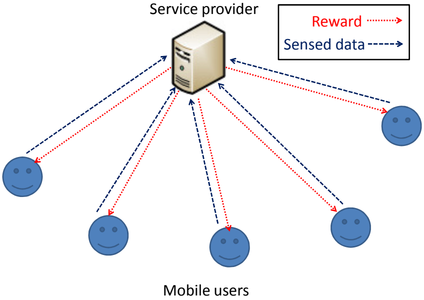
User Diversity

- User diversity improves the sensing quality for service provider:
  - Traffic monitoring: Users at less-crowded area provide more "valuable" information.
  - Mobile market research: Participants with diverse backgrounds may together provide a more comprehensive view.



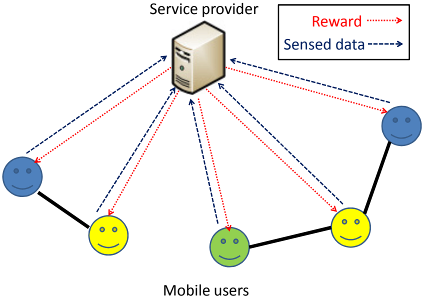
Diversity-Driven Social-Aware MCS

- Previous common assumptions:
  1. Users' decisions are independent.
  2. Value of information (VoI) depends on users' sensing efforts only.

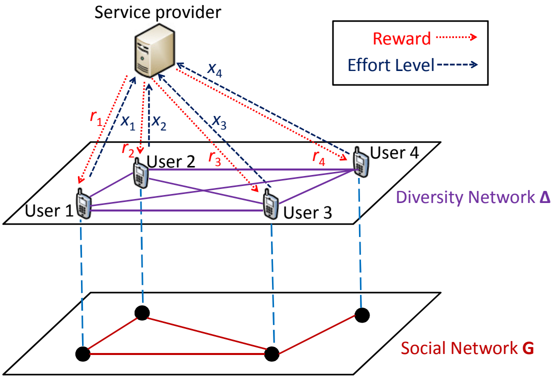


Diversity-Driven Social-Aware MCS

- New formulations:
  1. Users' decisions depends on social relationships.
  2. Value of information (VoI) depends on users' sensing efforts & diversity.

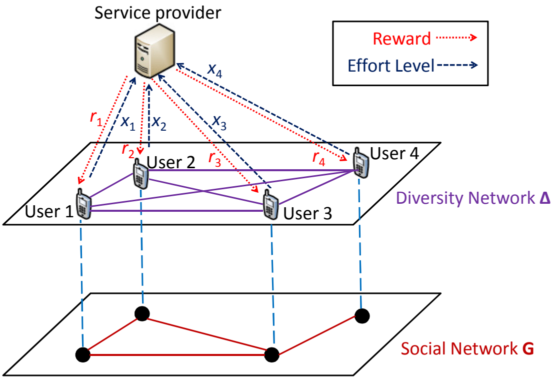


Two-Stage Model



- Stage I: How should the service provider determine the rewards $\boldsymbol{r} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_I)$ to all users?
- Stage II: How should users determine their effort levels $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_I)$?

Social Effect



- Social network $\boldsymbol{G} = [g_{ij}]_{I \times I}$ to model users' peer effects:
  - $g_{ij} \geq 0$: Social influence of user j on user i.
  - $g_{ii} = 0, \forall\, i \in \mathcal{I}$.

Stage II: Mobile User's Participation Game

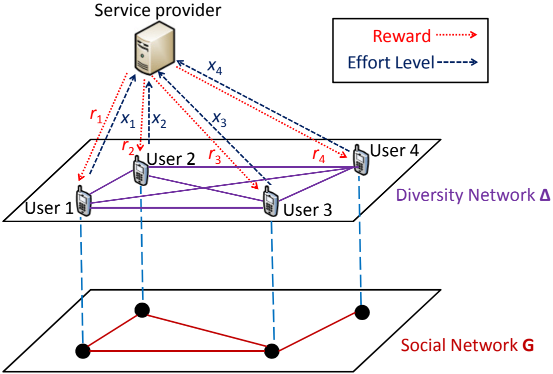
Mobile User Participation Game

Given reward $\boldsymbol{r}$, mobile users' participation game on social network $\boldsymbol{G}$:

- Players: Mobile users.
- Strategies: Effort levels $\boldsymbol{x} = (x_1, \ldots, x_I)$.
- Payoff: user i's payoff

$$U_i(x_i, \boldsymbol{x}_{-i}, r_i) \triangleq \underbrace{r_i x_i}_{\text{Reward}} - \underbrace{c_i x_i^2}_{\text{Convex cost}} + \underbrace{\sum_{j \in \mathcal{I}} g_{ij} x_i x_j}_{\text{Social effect}}.$$

Measuring VoI from Diversity

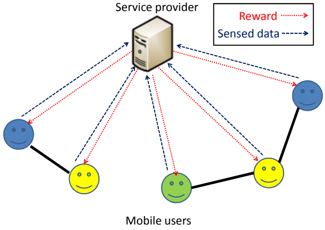


- Diversity network $\mathbf{\Delta} = [\delta_{ij}]_{I \times I}$:
  - $\delta_{ij} \geq 0, i \neq j$: Diversity or dissimilarity between users i and j.
  - $\delta_{ii} = 0, \forall\, i \in \mathcal{I}$.
  - E.g., $\delta_{ij} = \text{dist}(i, j)$ in location-dependent MCS.

Measuring VoI from Diversity

- VoI $\phi(\boldsymbol{x}) \triangleq \underbrace{\sum_{i\in\mathcal{I}} q_i x_i}_{\text{Indiv. quality}} + \underbrace{\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{I}} \delta_{ij} x_i x_j}_{\text{User Diversity}}.$
  - q_i: Sensing capability of user i.

- VoI depends on both users' sensing efforts & diversity.



Stage I: Service Provider's Profit Maximization

- Service provider chooses the reward to maximize its profit:

$$\underset{\boldsymbol{r}}{\text{maximize}} \quad \underbrace{\phi\big(\boldsymbol{x}(\boldsymbol{r})\big)}_{\text{VoI}} - \underbrace{\sum_{i \in \mathcal{I}} r_i x_i(\boldsymbol{r})}_{\text{Payment}}.$$

Optimal Rewards

Theorem

Service provider's optimal rewards in Stage I is

$$\boldsymbol{r}^* = \frac{\mathbf{s}q}{2} + \left(\mathbf{s}\Delta + \frac{\boldsymbol{G}^T - \boldsymbol{G}}{2}\right)(4\boldsymbol{C} - \boldsymbol{G} - \boldsymbol{G}^T - 2\boldsymbol{\Delta})^{-1}\boldsymbol{q}.$$

- What is the physical meaning of $\boldsymbol{r}^*$ in (1)?

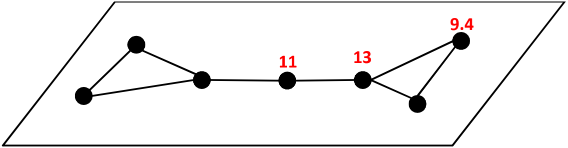
Katz Centrality

- Katz centrality: A node is important if it is linked from other important nodes or if it is highly linked.

$$\psi(\boldsymbol{A}, \alpha, \boldsymbol{w}) \triangleq (\boldsymbol{I} - \alpha \boldsymbol{A})^{-1} \boldsymbol{w}.$$

  - $\boldsymbol{A} = [A_{ij}]_{I \times I}$: Graph.
  - $\alpha \geq 0$: Scalar.
  - $\boldsymbol{w}$: Weight vector.

- Example: $\alpha = 1/3$ and $\boldsymbol{w} = \mathbf{1}$ (i.e., unweighted).



Reward as Weighted Katz Centrality

Theorem

The optimal reward to a user is the weighted sum of Katz centrality of other users in the superimposed graph $\boldsymbol{G} + \boldsymbol{\Delta}$.

Conclusion

- Technologies make fog computing feasible.
- Economics make fog computing successful.

National Taiwan University

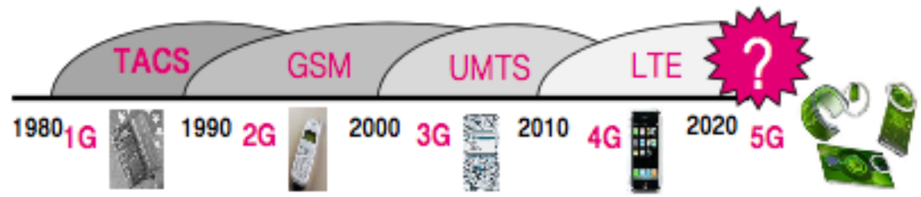


Enabling Low-Latency Application in Fog-Radio Access Network

Ai-Chun Pang, Professor

Grad. Inst. of Networking & Multimedia
Dept. of Comp. Sci. & Info. Engr.
National Taiwan University, Taiwan

Trends for Future Wireless Comm.



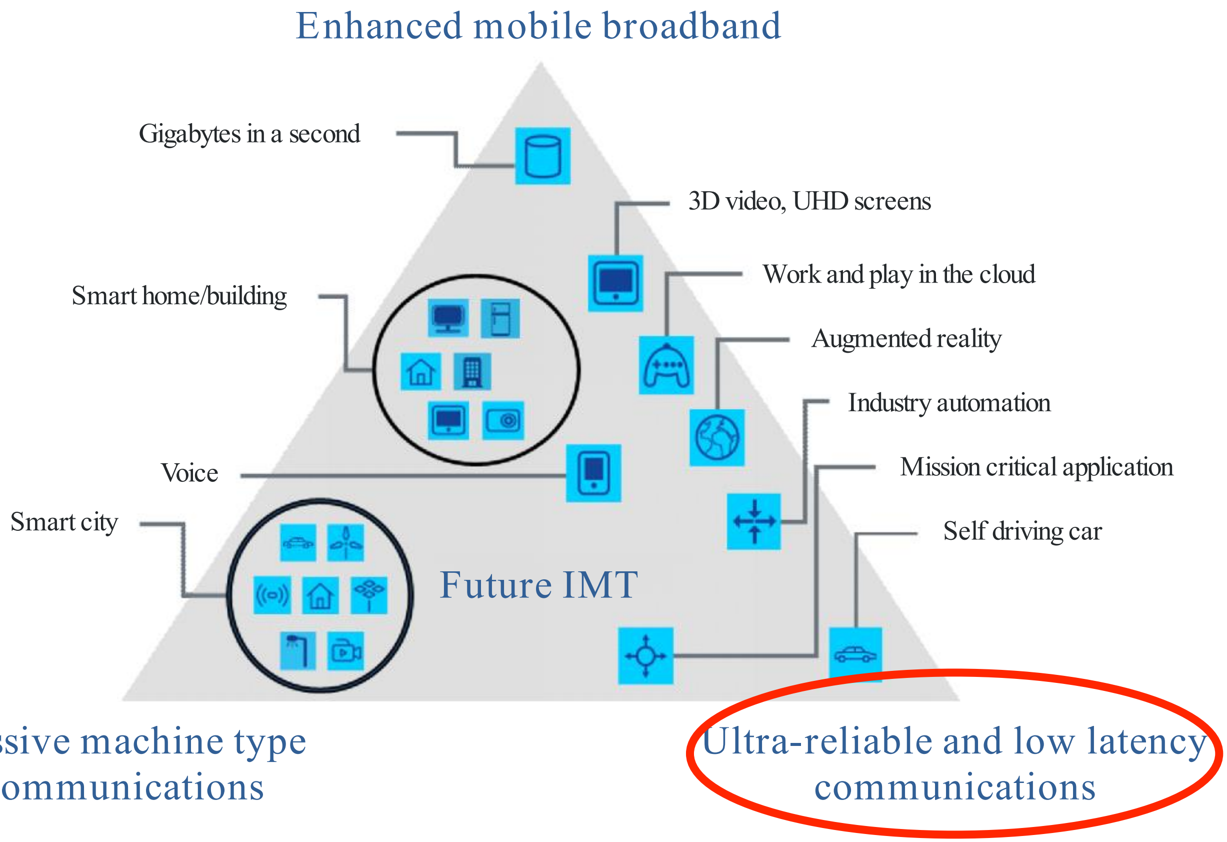
- Data traffic avalanche
- Massive growth of connected devices
- Diversification of services and equipment
- Vertical markets

臺灣大學 National Taiwan University



Vision for 5G

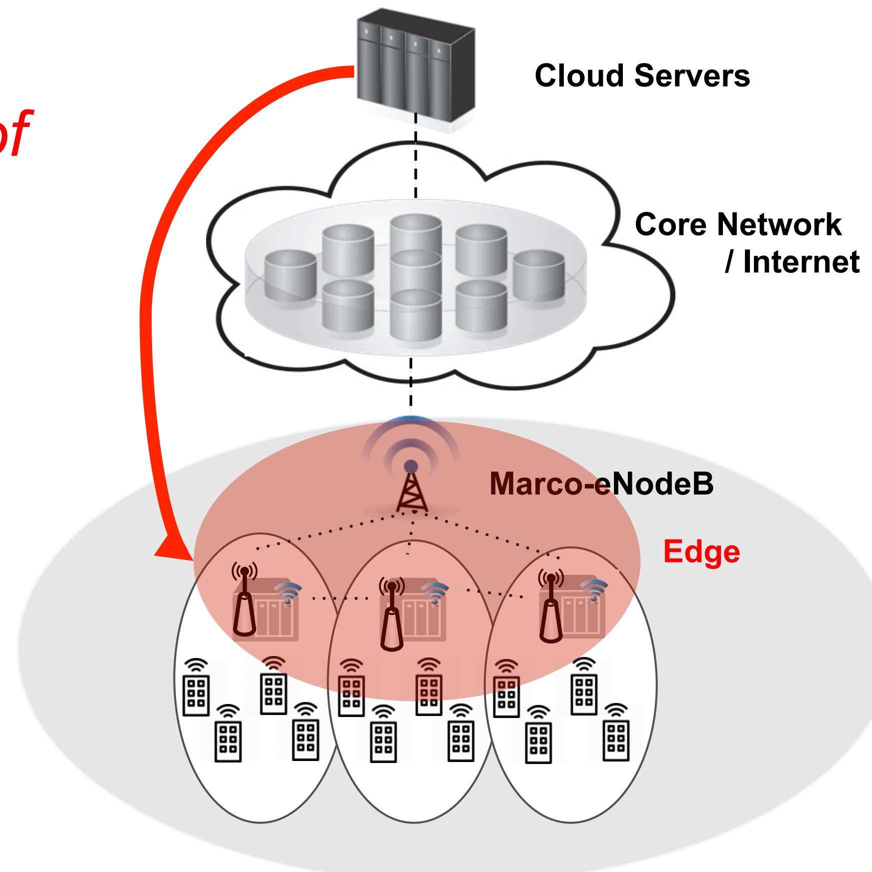
New use scenarios will emerge calling for requirement enhancement:
Mobile Broadband, Massive Connectivity, Low Latency.



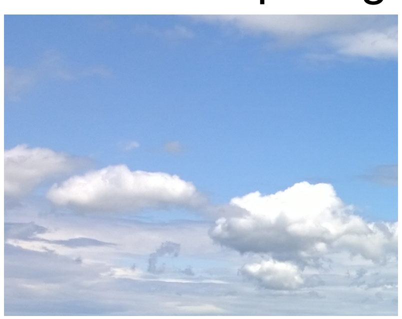
Ultra Low Latency Realization in 5G

In order to realize the latency of several ms, new technology will be required.

*"Push **everything** to the edge of network for low latency"*



Cloud Computing



- Centralized pooling
- Efficient resource utilization

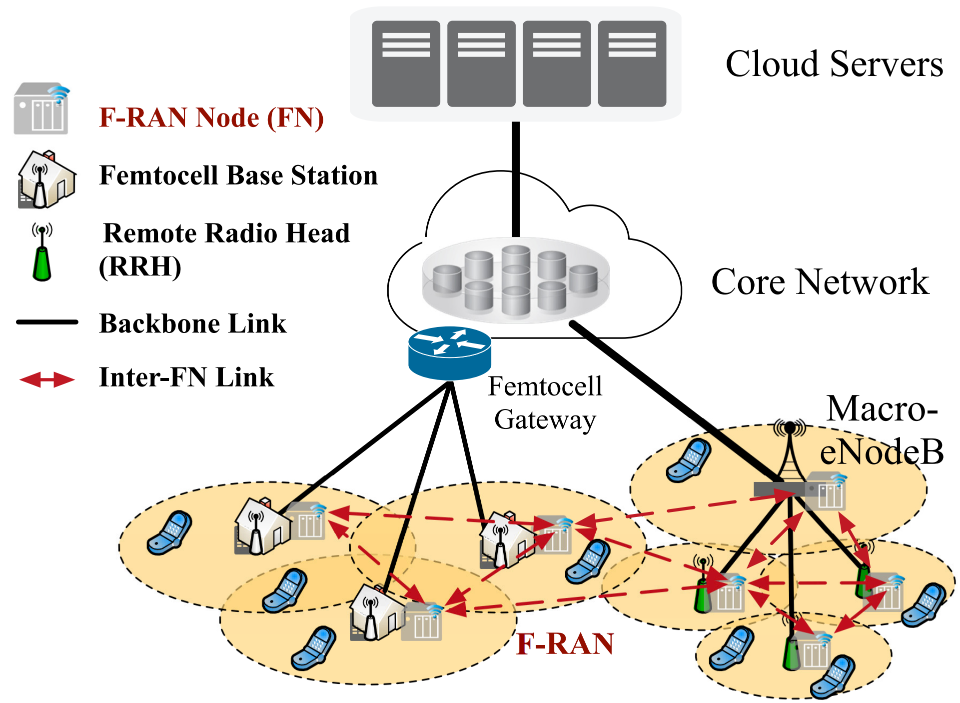
Fog Computing

- Close to the edge
- Low latency



Fog-Radio Access Network (F-RAN)

- Equipment **in the RAN**
  - responsible for both **communication** (protocol/signaling) and **application services** (data processing and storage)
  - can communicate with each other directly



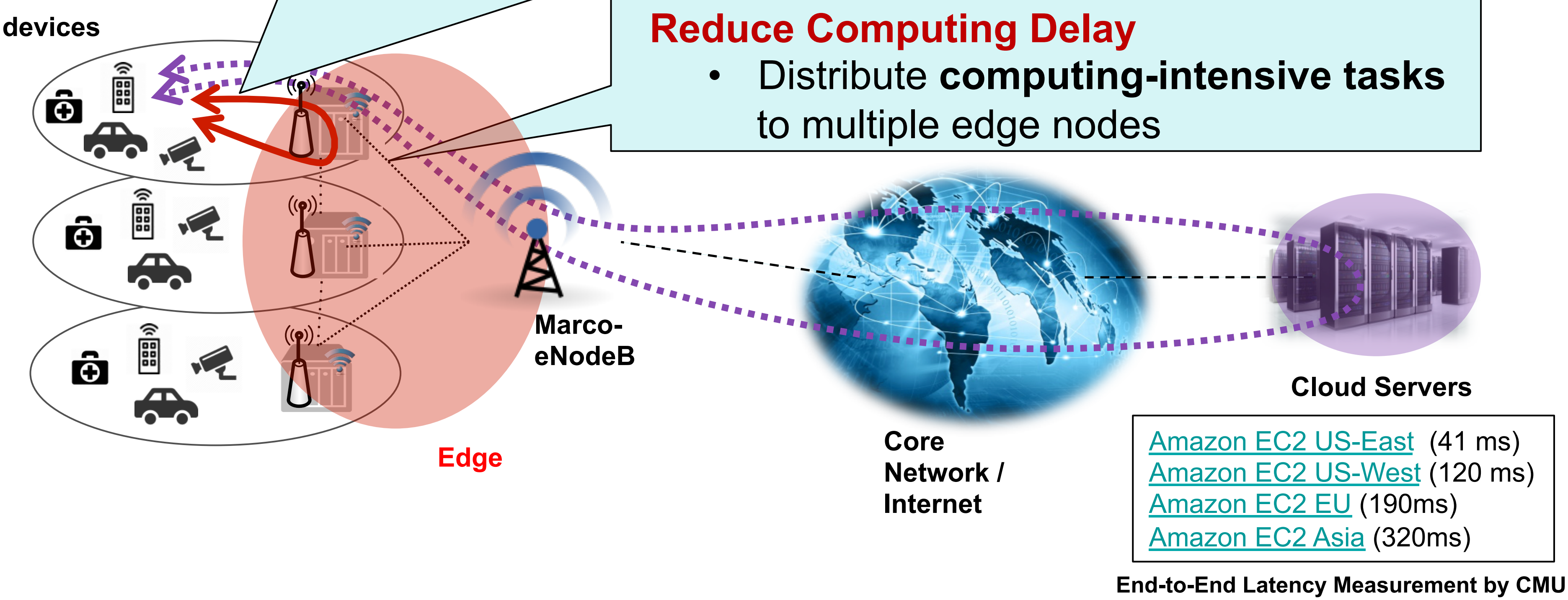
- "Enabling Low-Latency Applications in Fog-Radio Access Network," *IEEE Network*, January/February, 2017.
- "5G Radio Access Network Design with Fog Paradigm: Confluence of Communications and Computing," accepted and to appear in *IEEE Communications Magazine*.

National Taiwan University

How F-RAN Works

Reduce Communication Delay
- **WAN latency** is hard to improve
- Some applications require **bulk processing data** for computing-intensive tasks (e.g., real-time video analytics)

Reduce Computing Delay
- Distribute **computing-intensive tasks** to multiple edge nodes





Research on F-RAN

1. **Resource Management**
   - Joint design on computing and communication resource allocation
   - "Latency-Driven Cooperative Task Computing in Fog-Radio Access Networks," *IEEE ICDCS 2017*
2. **Service Provisioning**
   - Container-based virtualization for provisioning wearable applications in WiFi access points
   - "A Virtual Local-hub Solution with Function Module Sharing for Wearable Devices," *IEEE MSWiM 2016*
3. **Fog-based Platform**

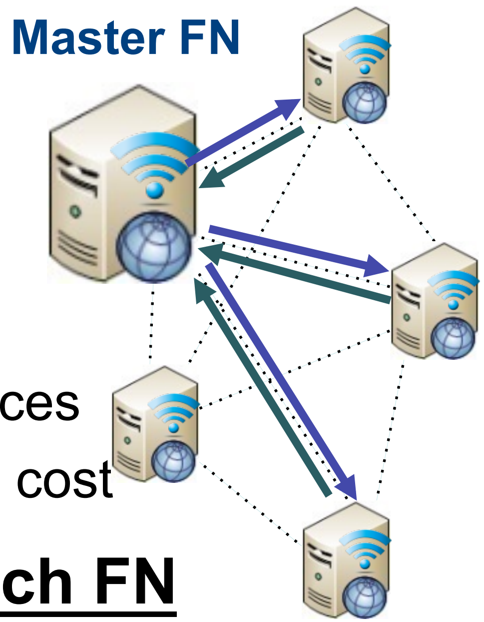


R1: Challenges for Computing in F-RAN

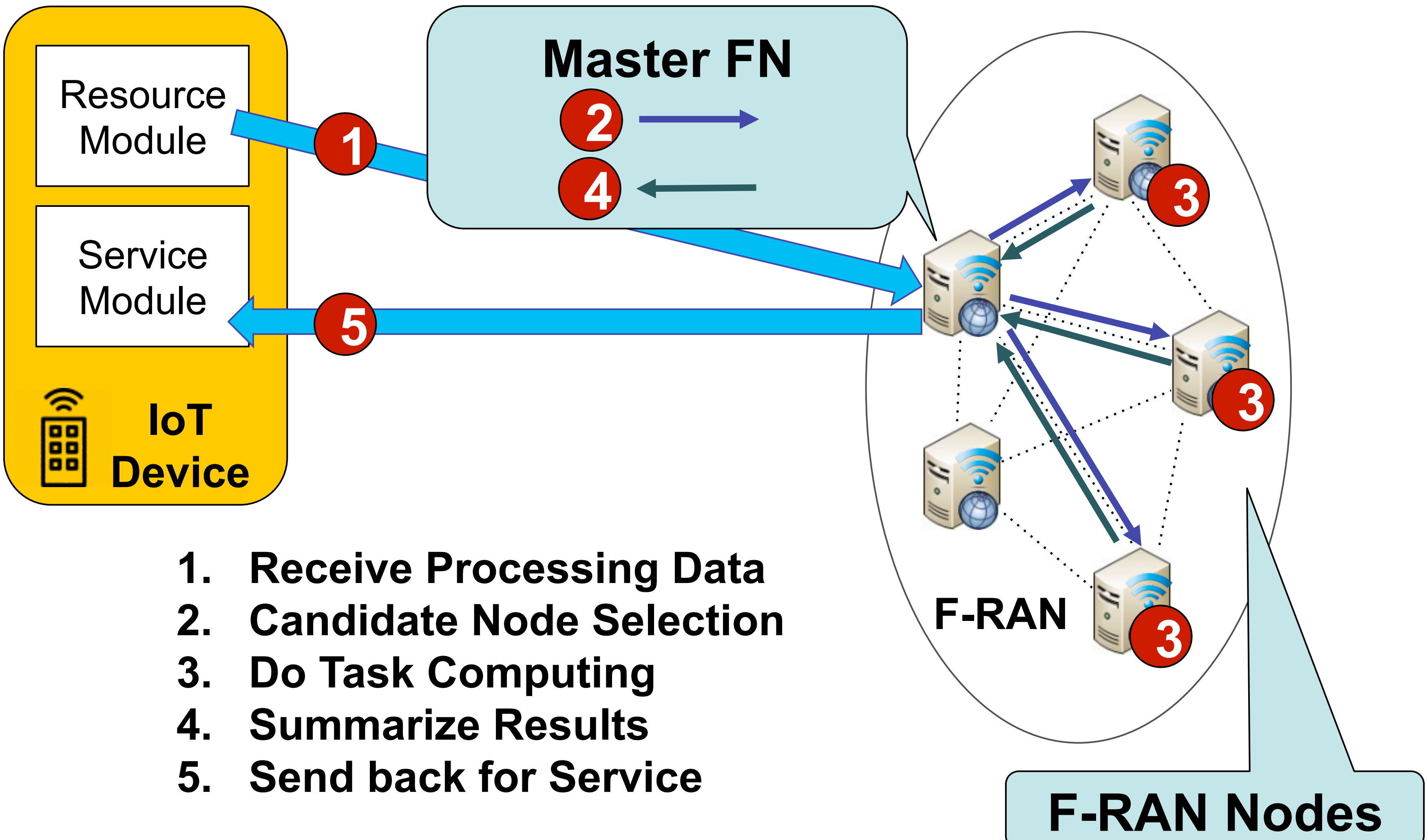
- **The computing capability of an F-RAN node (FN) is very limited.**
  - Single FN is not capable for computing-intensive tasks.
  - Propose to do the application-layer **computing collaboratively involving multiple FNs**.
- **How to decide how many and which FNs to be involved**
  - A new type of **cost (communication/computing)-performance tradeoff** where the **temporal equivalency** of the two physically different resources needs to be built.

A New Type of Comm. and Comp. Tradeoff

- Need to tackle the issues considering the **tradeoff between communication and computing** in **temporal domain**
  - More FNs ► Higher computing power for all system (lower comp. delay) but lower communication resources for each FN (higher comm. delay)
- **Decide which FNs to be selected**
  - Attributes of master FN ► communication resources
  - Distances between master FN and FNs ► comm. cost
- **Decide amount of computing tasks for each FN**
  - Attributes of FNs ► computing resources
  - Loading of FNs ► computing cost



Cooperative Computing in F-RAN



1. Receive Processing Data
2. Candidate Node Selection
3. Do Task Computing
4. Summarize Results
5. Send back for Service

Problem Formulation

- Objective:
  - **To minimize total service latency** (**communication** + **computing**) :

$$Objective = \min\max_{\forall n \in N} \left(\min\max_{\forall f \in F} I_f^n \times \left(D_f^n / \delta_f^n \times \gamma_f + C_f^n / \theta_f^n \times \rho_f \right) \right)$$

Minimize all user's latency
which is decided by the
last user with longest latency

Each user's service latency
which is decided by the last
cooperative FN

Each user's communication latency

Each user's computing latency

- Subject to:
  - Communication and Computing Resource Feasibility:

Heterogeneous Resources

$$\sum_{\forall n \in N} \sum_{\forall f \in F} I_f^n \times \delta_f^n \le \delta \text{ and } \sum_{\forall n \in N} I_f^n \times \theta_f^n \le \theta_f, \forall f$$

  - Processing Data and Computing Tasks Assurance:

$$\sum_{\forall f \in F} I_f^n \times D_f^n \ge D^n, \forall n \text{ and } \sum_{\forall f \in F} I_f^n \times C_f^n \ge C^n$$



Cooperative Task Computing Operation (1/2)

- **<u>Special case for one user</u>**:
  - Design a *dynamic programming* approach (CTC-DP)
  - Proof of *optimal* solution for minimum service latency
  - Based on recursive formula $g(r, c, f)$ to build a DP table

$$g(r,c,f) = \begin{cases} 0, & \text{if } c = 0 \\ \infty, & \text{else if } r = 0 \text{ or } f = 0 \\ \min\limits_{\hat{r}\in[1,r],\hat{c}\in[1,c]} \left(\max(g(r-\hat{r}, c-\hat{c}, f-1), t^{f}_{\hat{r},\hat{c}}), g(r,c,f-1)\right), & \text{otherwise} \end{cases}$$

  - Two procedures:
    - ✓**FILL-TABLE()**: fills the DP table by $g(r, c, f)$
    - ✓**BACK-TRACE()**: selects the feasible set of FNs with cooperative tasks assignment



Cooperative Task Computing Operation (2/2)

- **<u>General case for multiple users</u>**:
  - Design a heuristic algorithm (CTC-All)
    - ✓ Propose *one-for-all* concept to consider other's *side-effect*
  - Avoid *resource starvation* and *utilization degradation*
  - Two stages:
    - ✓ **Heterogeneous resource allocation**
      - ◻ Decide comm. resources based on processing data weight
      - ◻ Dynamic comp. resource allocation under distributed architecture
    - ✓ **Cooperative task computing**
      - ◻ Leverage CTC-DP with *one-for-all* concept for solving each user's cooperative task computing

Simulation Setup

- Communication considers **path loss**, **shadowing**, and **multipath fading**
- Computing ability are estimated by **ARtoolKit** [1] **Valgrind** [2]
- Frame Width: QCIF 176×144 pixels [2][3] (Encode with H.264)
- Bits/pixel: 8 bits (Gray scale)
- Max RB number: 100 (Based on **LTE specification -** *3GPP TS 36.211*)
- Data rate per RB: 9.6, 14.4, 19.2, 21.6 Kbps
- Max FN number: 20
- Platform: Intel i7 Core 2.5GHz, Dual core, 8G RAM
- Computing Power: 700 - 1700 Million Instructions/sec

[1] ARtoolKit, Available: http://artoolkit.sourceforge.net
[2] Valgrind, Available: http://valgrind.org/
[3] Video sequences, Available: http://trace.eas.asu.edu/yuv/
[4] J. Ha, K. Cho, F.A. Rojas, H.S. Yang, "Real-time scalable recognition and tracking based on the server-client model for mobile Augmented Reality", in IEEE ISVRI, Mar. 2011.

Exemplary Ultra-Low Latency Result

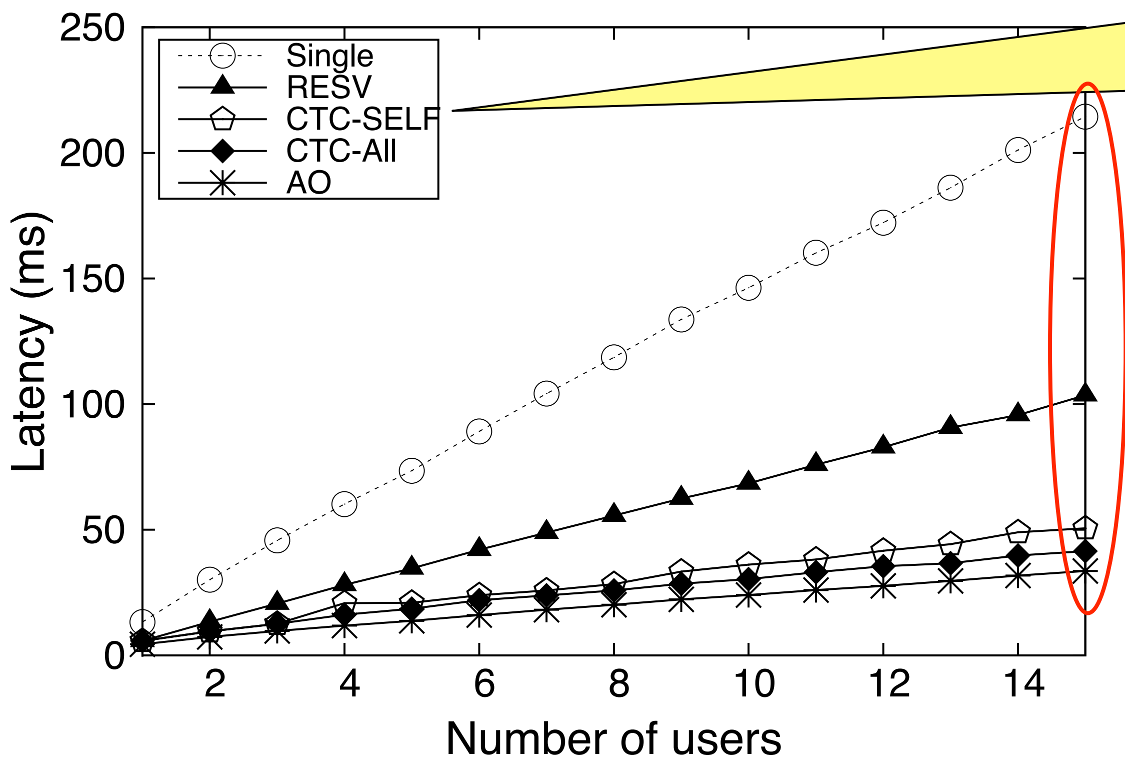


Fig. 1 Impacts of the number of users on total service latency.

Single: Lower bound
RESV: Reservation idea
CTC-SELF: Selfish goal
CTC-All: Our scheme
AO: Upper bound

CTC-All achieves **173ms (4.2x) less latency** than Single, **62ms (1.5x) less latency** than RESV and **9ms (24%) less latency** than CTC-SELF

National Taiwan University

Other Matrices

Single: Lower bound **RESV**: Reservation idea
CTC-SELF: Selfish goal **CTC-All**: Our scheme
AO: Upper bound

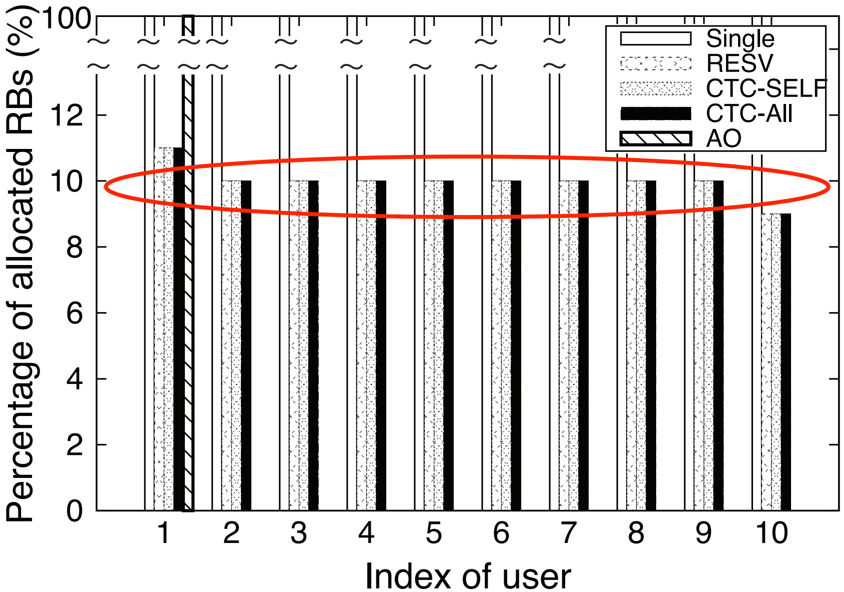


Fig. 3 Percentage of allocated RBs of each user.

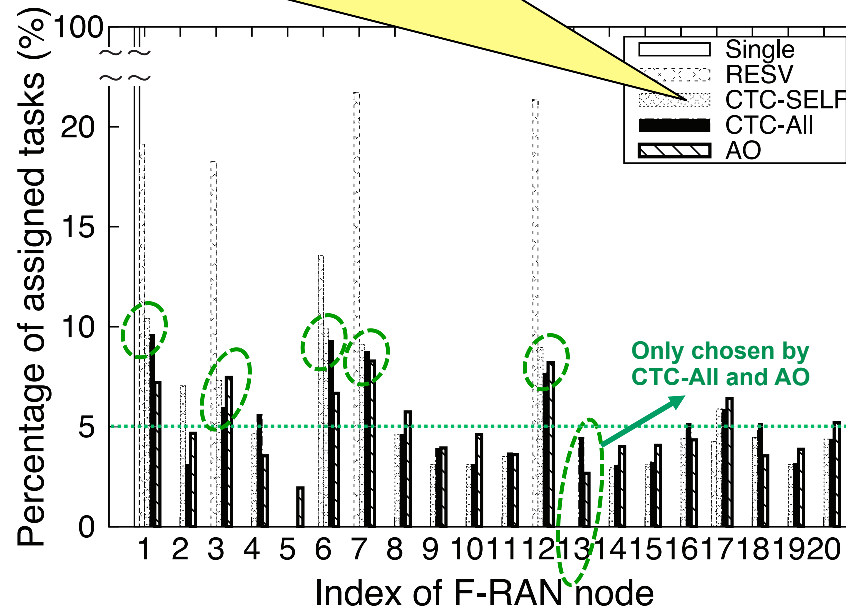


Fig. 4 Percentage of assigned tasks of each FN.

In Fig.3, ***dynamic computing resource allocation*** is the key to perform effective cooperative task computing

In Fig.4, CTC-All with *one-for-all* achieves ***load-balancing***

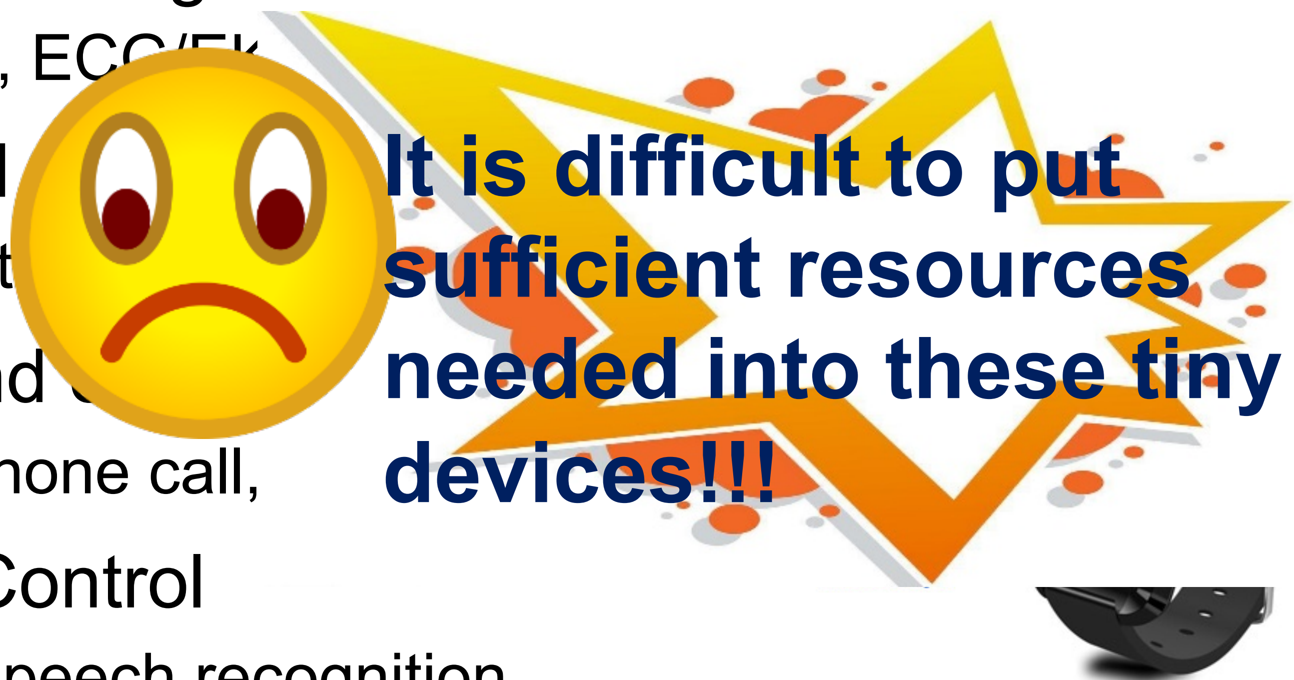
R2: Fog-based Wearable Applications

- Clothing or accessories worn on human body incorporating computer and advanced electronic technologies
  - Sensors
  - Processing and storage capacities
  - Wireless connectivity (BLE、Wi-Fi)
  - Display

- Characteristics
  - Light weight: easy to wear
  - Low power consumption

Applications (1/2)

- Health monitoring
  - Heart-rate, ECG/EK
- Fitness and
  - Step count
- Sociality and
  - Texting, phone call,
- Intelligent Control
  - Gesture, speech recognition…
- Entertainment and Others
  - Streaming, gaming, navigation…



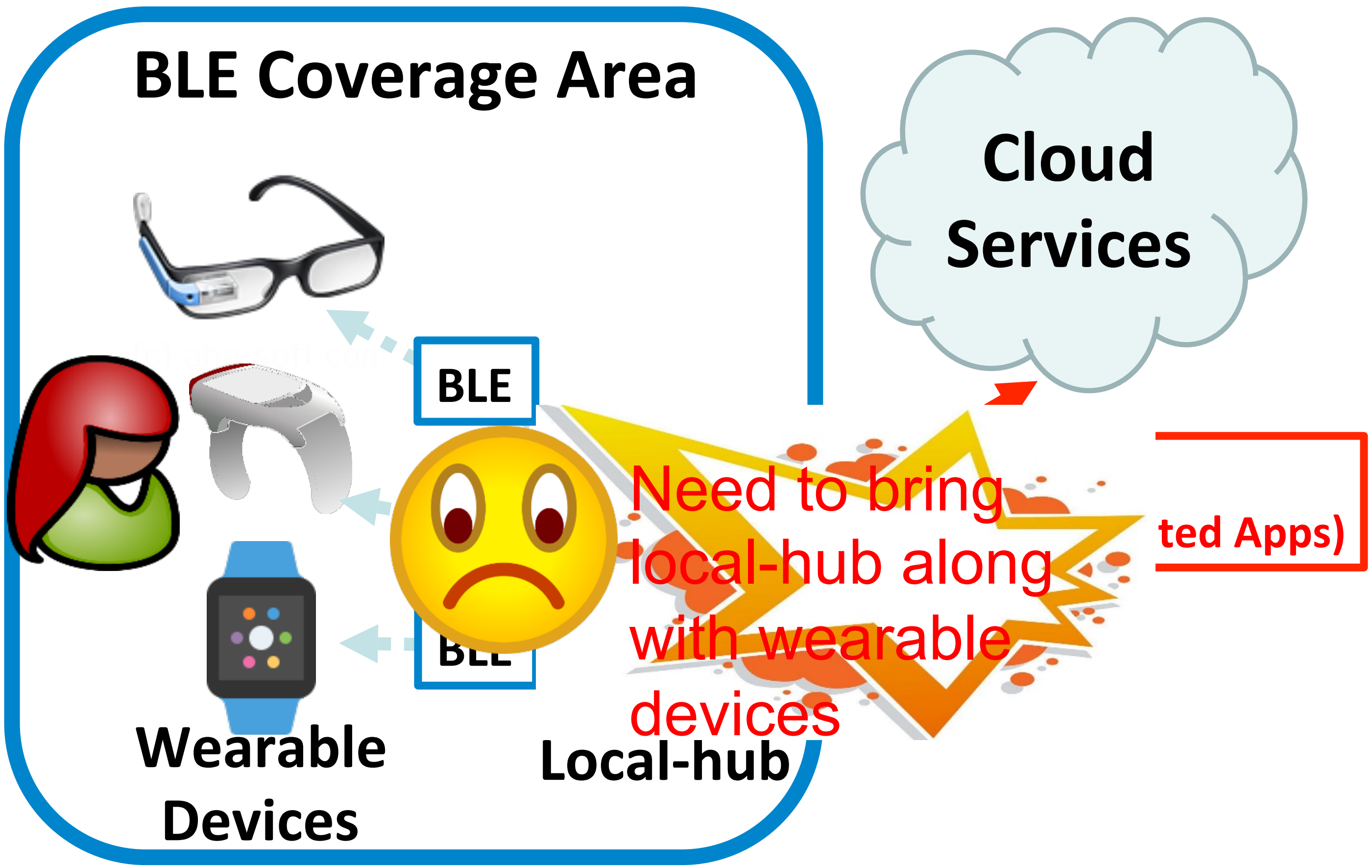
Applications (2/2)

- Requirements of wearable applications
  - Powerful processing capacity
  - Sufficient storage capacity
  - Internet connectivity
- The requirements are in conflict with the characteristics of wearable device
- Existing solution: local-hub
  - Adopt powerful devices to replenish capacities

Local-hub

- Usually a smart-phone or tablet, installed with applications related to wearable devices
- Wearable devices are connected with a local-hub via low power wireless technologies, e.g., BLE

Physical Local-hub Scenario

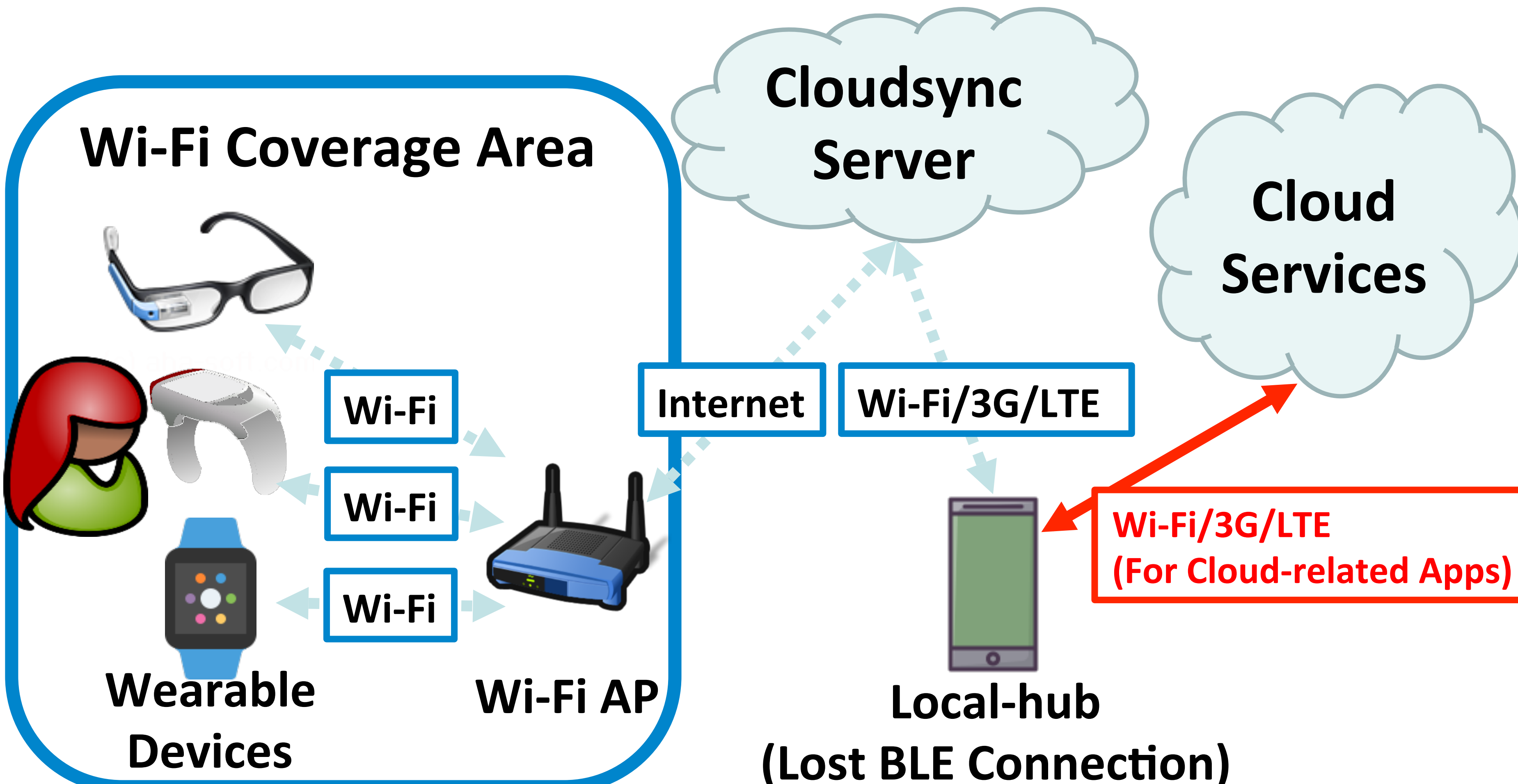




Inconvenience of Physical Local-hub

- Wearable devices are useless if local-hub is not nearby, for example,
  - Working out in a gym
  - Swimming in a pool
- Local-hub functionalities drawdown the battery of smart phone
- Current solutions
  - Google: Android Wear **Cloudsync**
  - Apple: **Compatible Wi-Fi** for Apple Watch

Android Wear Cloudsync Scenario





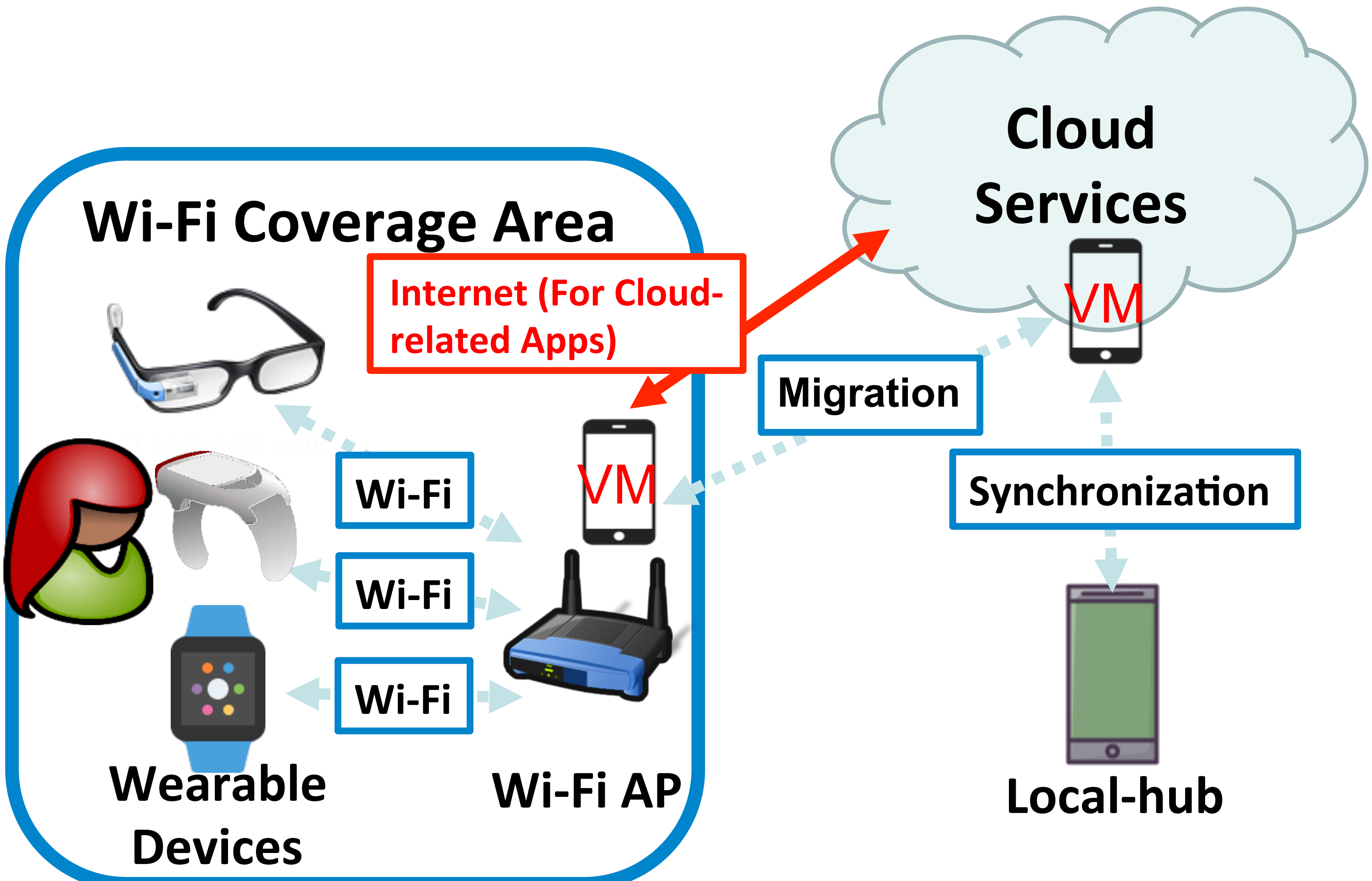
Limitation of Current Wi-Fi Solution

- Long response time
  - Raw data traveling time
    - Among wearable device, cloud, and local-hub over the Internet
    - Pre-processing of the raw data should be done on local-hub
  - Indirect data exchange
    - Cloudsync server intermediates data exchanged between wearable device and local-hub
- Shortcoming
  - Poor user experience (waiting time)
  - More power consumption (screen-on time)

Concept of Virtual Local-hub (VLH)

- Virtual Local-hub (VLH)
  - Wearable devices can utilize network edge nodes nearby to serve as their local-hub instead of smartphones
- Basic ideas make VLH to be practicable
  - Fog computing
  - Virtualization technology
- Intuitive idea of VLH
  - Virtualize all applications of local-hub in a smartphone as a virtual machine (VM)
  - Migrate the whole VM to edge nodes (e.g., Wi-Fi AP) nearby the user

Intuitive Idea of VLH



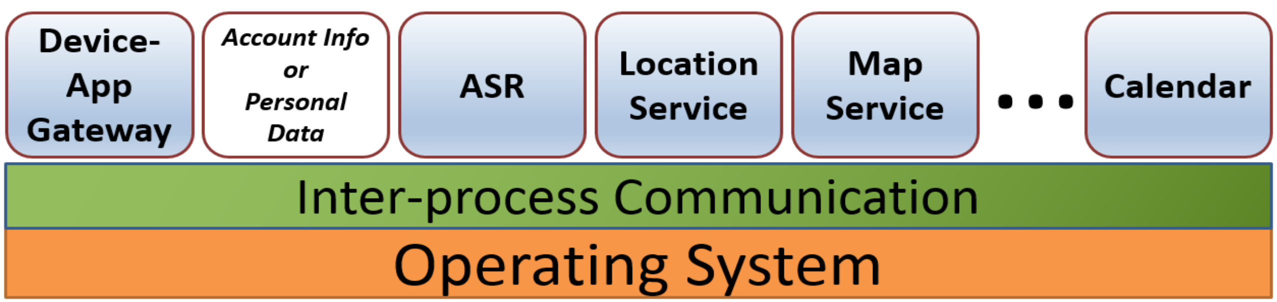


Issues of VM Migration

- Long migration time of whole VM
  - Size of a VM is quite large (about hundreds of MBs)
- Capacity limitation of a Wi-Fi AP
  - Processing/storage resources are restricted on an AP
  - A Wi-Fi AP may only accommodate few VMs
- Not a cost-effective solution

VLH System Design

- Idea 1: Fog Computing realized by a group of Wi-Fi APs
  - Wi-Fi APs can connected with each other on a LAN
- Idea 2: Container-based Virtualization
  - Modular programming environment for mobile APP
    - Developers can adopt existing function modules to build the applications for wearable devices
  - To virtualize function modules as containers



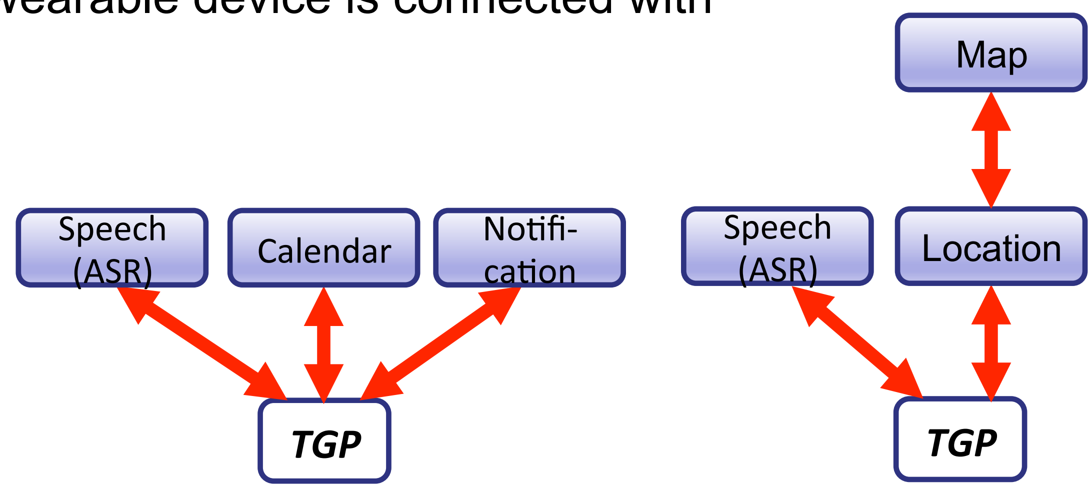
System Model (1/2)

- Wearable Devices
  - Equipped with both BLE and Wi-Fi radios
- Edge Network (VLH Network)
  - Deployed by the operators with the technology of container-based virtualization and fog computing
  - APs are capable of executing applications as the local-hub
- Function Module (FM) Sharing
  - Pre-install the images of function modules on Wi-Fi APs
  - Function modules can be shared by service requests
    - Every function-module instance has an upper bound for sharing.



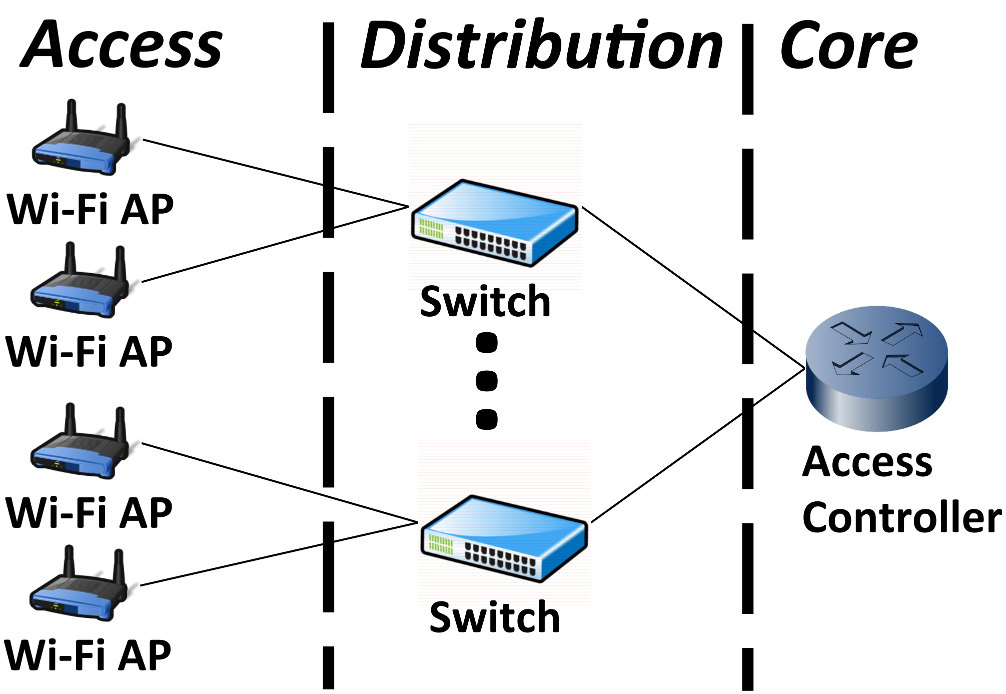
System Model (2/2)

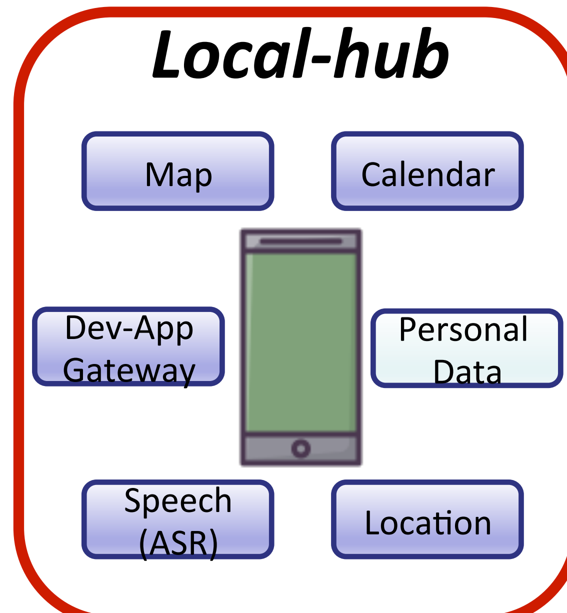
- Service Request
  - A service/application is performed by a serial execution of function modules
    - Call Graph: A serial call of function modules
    - TGP: Traffic Gateway Point
      - AP that wearable device is connected with



How It Works (1/3)

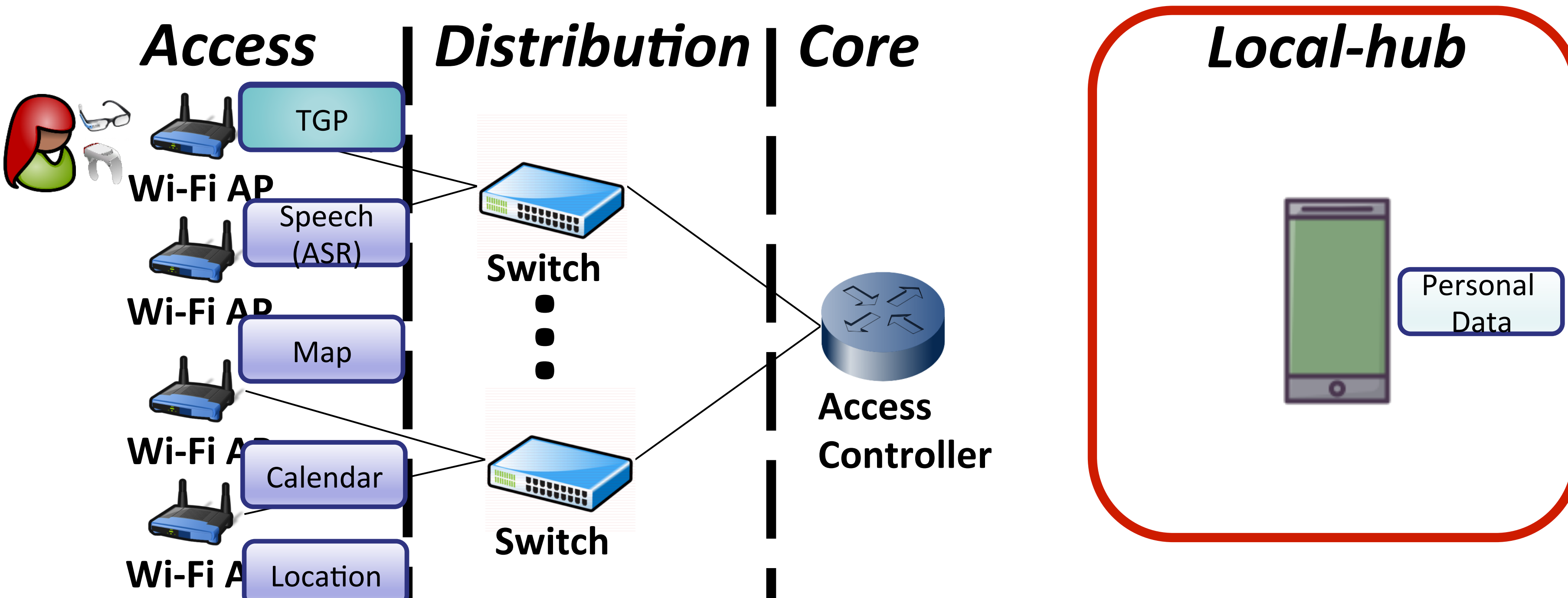
- Shift functionalities of local-hub from smart phone to edge network
  - Pre-install shareable function module on APs





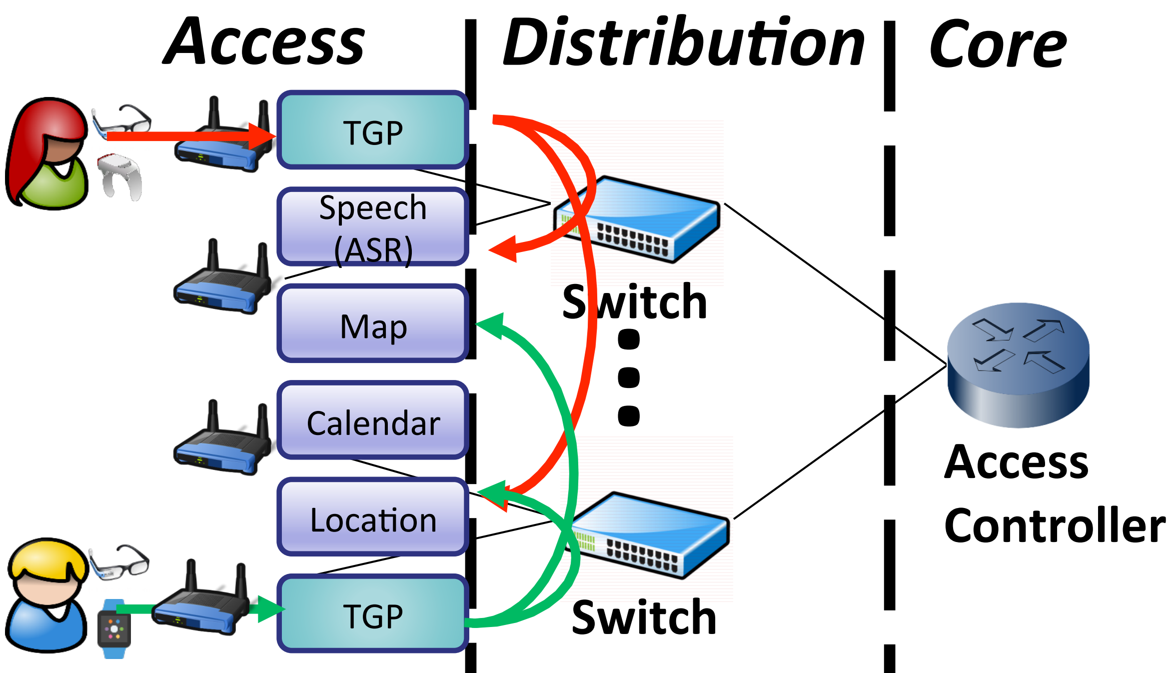
How It Works (2/3)

- Role of the AP wearable device connects to
  - Download personal data from cloud or smart phone
  - Perform TGP functionality for wearable device



How It Works (3/3)

- Fulfill user's requests
  - Allocate sufficient FM instances on right locations
  - Call graph mapping: FM instances sharing decision





Objective

- To mitigate the side-effect of function module sharing
  - To Minimize the total bandwidth consumption of edge network
- Challenges
  - How many FM instances should be executed on VLH network?
    - **Resources usage** decision
  - How to allocate these FM instances?
    - **Migration** decisions
    - **Allocation** decisions
  - How to share these FM instances?
    - **Call graph mapping** decisions

Proposed Algorithm

- **<u>Nearest Serving Node (NSN) Algorithm</u>** (Greedy-based)
  - Key Idea: A FM instance should serve those requests as near as possible

Resources Usage Decision → **FM Instances Allocating Procedure**

Choose **the least FM instances** for allocation based on sharing limit

For each FM instance

- Try **every node** on edge network
  - Migration bandwidth consumption
  - Bandwidth consumption of serving these SRs
- Choose the **least bandwidth consumption** one



Performance Evaluation

- Simulation Setup
  - Number of Wi-Fi APs: 100
  - Available bandwidth capacity: 1 Gbps
  - Available computing capacity: 1000
  - Number of function module types: 20
  - Bandwidth requirement of types: 1-150 Kbps
  - Computing requirement of types: 5-100
  - Package size of function module types: 1-15 MB
  - Number of call graph types: 20
  - Number of service requests: 500



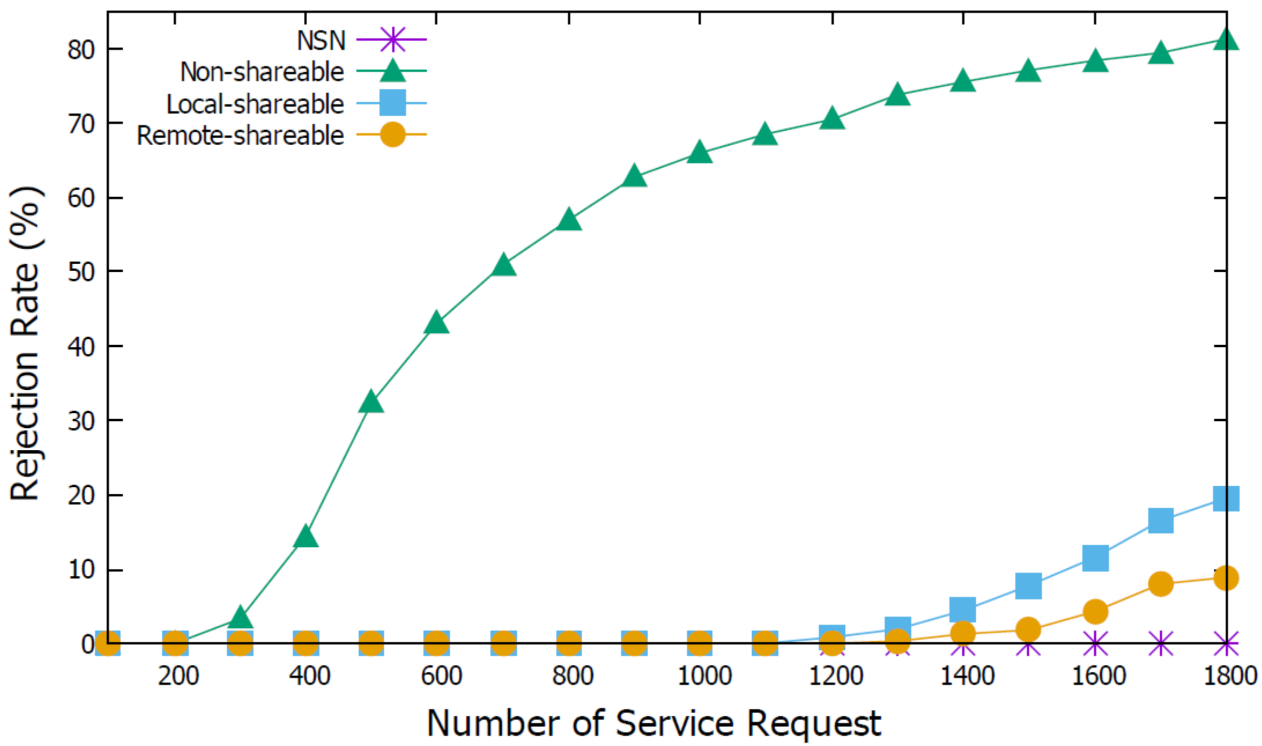
Performance Evaluation

- We conduct two kinds of comparison
  - Comparison of Different Sharing Strategies
    - To assess the impact of different function module sharing strategies on the rejection rate
      - Non-shareable
      - Local-shareable
      - Remote-shareable
  - Comparison of Different Allocation Strategies
    - To investigate the performance of total bandwidth consumption
      - First In First Out (FIFO)
      - Random

National Taiwan University

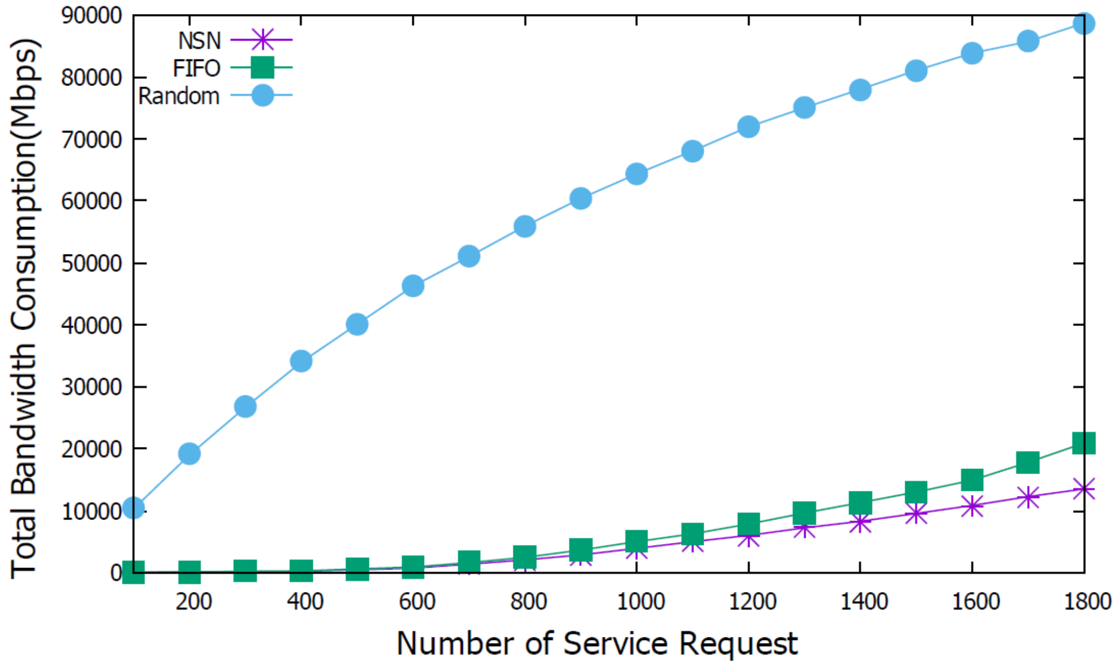
Comparison of Sharing Strategies

- Non-shareable suffers from high rejection rate
  - Up to 80% service requests cannot be accommodated
- Remote FM sharing can reduce rejection rate significantly



Comparison of Allocation Strategies

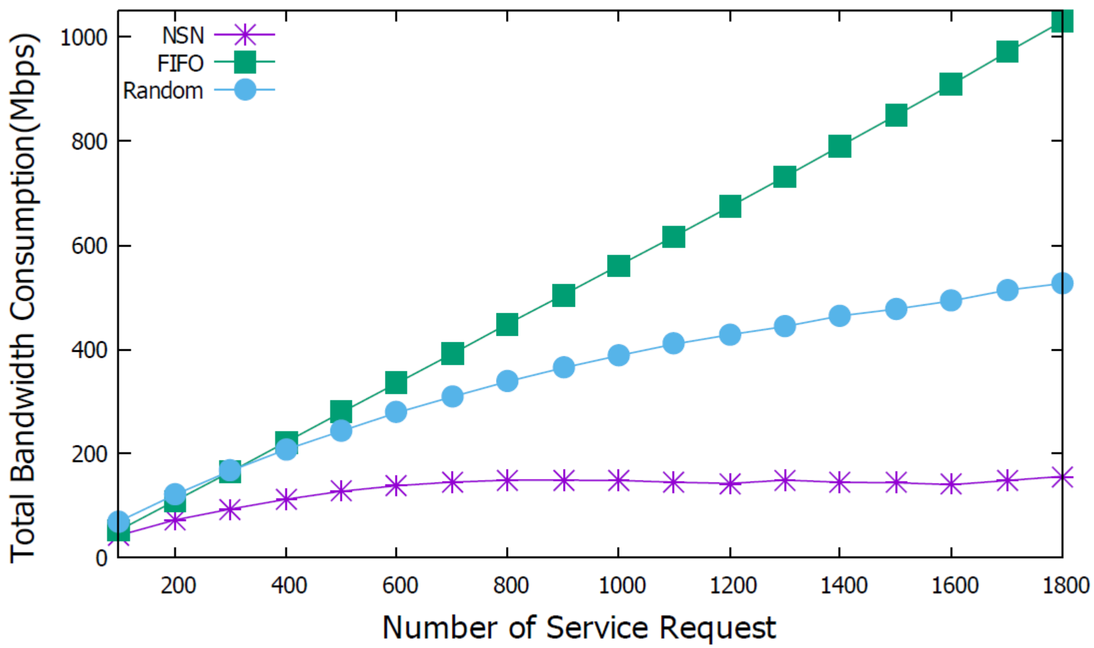
- Impact of the number of service requests
(migration occurs due to limited storage size)



(a) Total bandwidth consumption

Comparison of Allocation Strategies

- Impact of the number of service requests (No migration)





R3: OmniEyes: Fog-based Video Management Platform

- The generation of video data has started a paradigm shift from the content provider to individuals and now the "things"

NETFLIX → facebook →

12B hours of Netflix in Q4 2015 and 2015 revenue sales: 6.78B USD

Snapchat: 7B video views per day
Periscope: 100m live broadcasts in 10 months

Video → **Video & Sharing** → **Video & Sharing + Mobility & Analytics**

We want to become the "Mobile Video" King of the physical world

To Change the way people explore the physical world
with our **omnipresent videos**

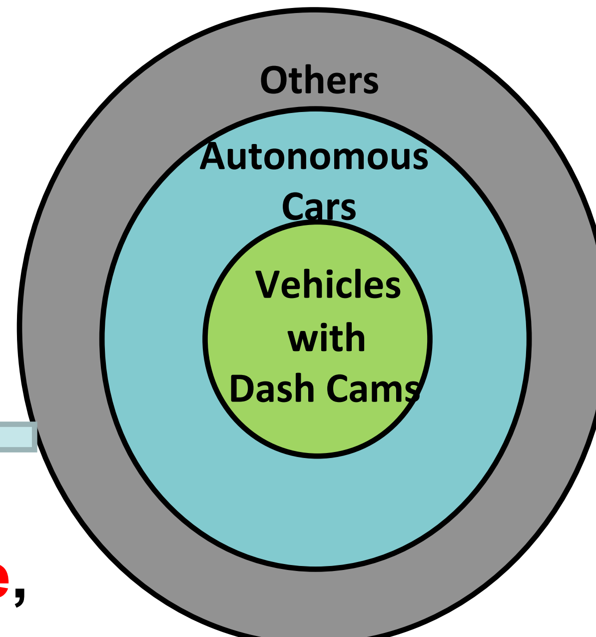
New ways of Searching, Driving, and Tracking
New ways of Mobile Advertisement and Auto Insurance

National Taiwan University

Our OmniEyes Platform

1. Street View Live | **2. Harperscope** | **3. Time Machine**

Fog Based Video Platform



A platform to share, to fuse, to analyze, and to render location-dependent, video data + information



Conclusion

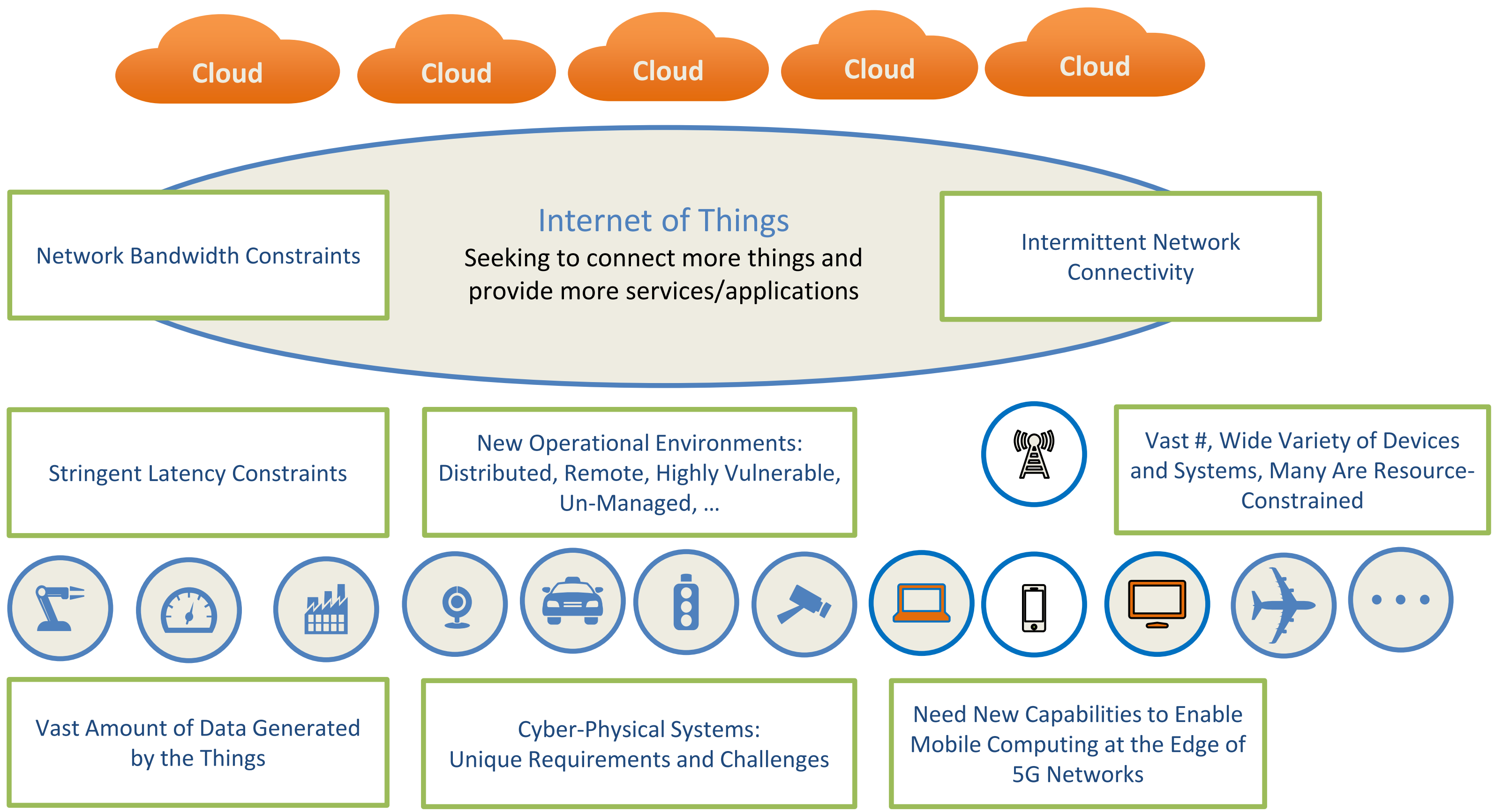
- **Low latency** is required by many existing and new usage scenarios for future communications.
- Fog computing is the key to realize low-latency communications.
  - It also makes ISP/carrier turn from dump-pipe into smart-pipe.
  - Orchestration of fog and cloud
- There will be huge research and business opportunities following this direction.

Fog Computing for Open 5G Development

Dr. Yang Yang

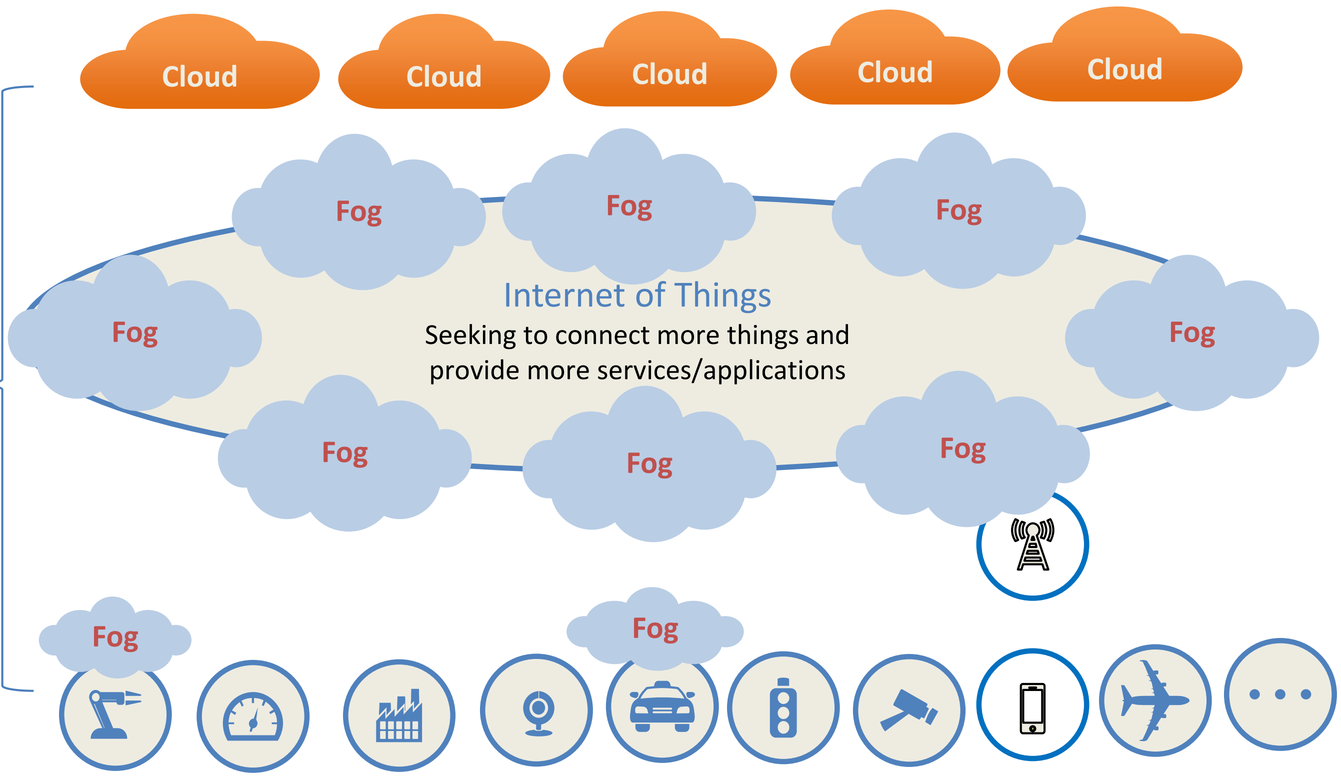
Shanghai Research Center for Wireless Communications
Key Lab of Wireless Sensor Network and Communication
SIMIT, Chinese Academy of Sciences

Current Computing Paradigm Inadequate

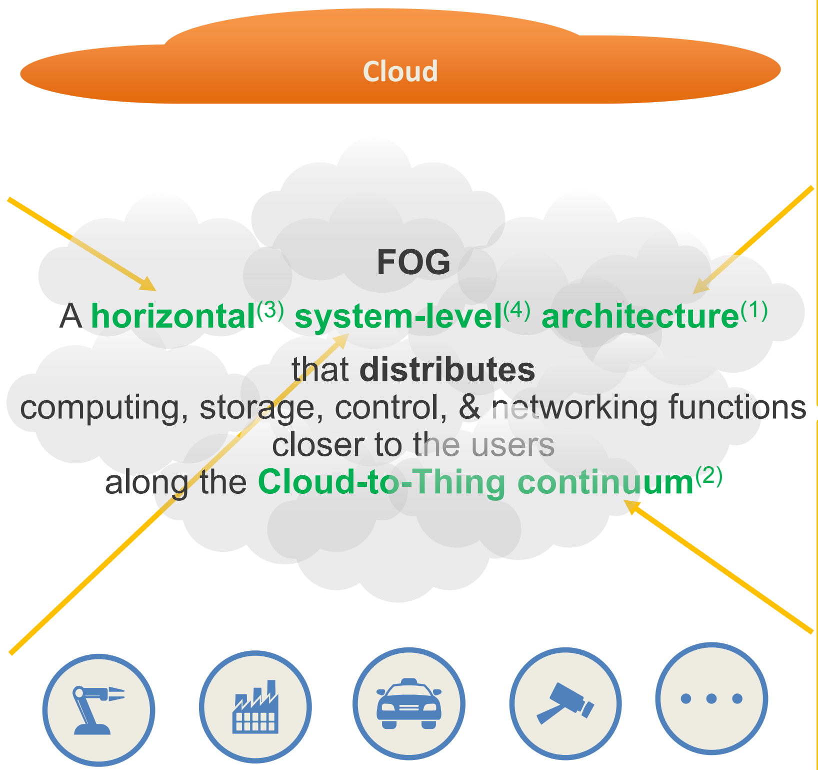


Fog Computing is the Future

Fog *Enables Computing anywhere along the Cloud-to-Thing Continuum*



What is Fog Computing?



1. Architecture

with its enabling **tools** for distributing, orchestrating, managing, securing resources and services

(not just placing servers, apps, or small clouds at the edge)

2. Cloud-to-Thing Continuum

Distributes resources and services to anywhere along the continuum

(not just at the edge)

Converged Cloud/Fog services

(not just isolated edge computing devices / apps)

3. Horizontal

Supports multiple industries

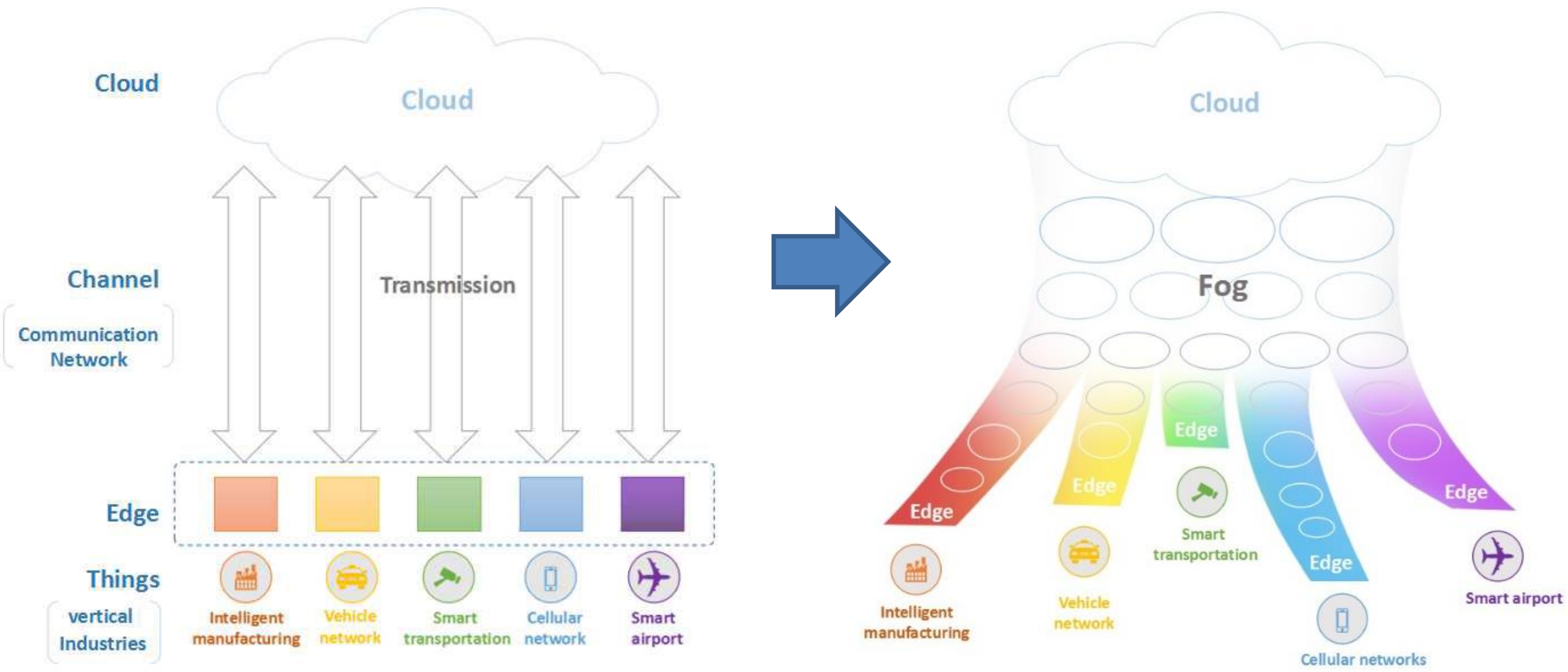
(not limited to any specific industry, network type, or application domain)

4. System-Level

from Things to the Edge, and over the Core to the Cloud, spanning multiple protocol layers

(works over and inside wireless and wireline networks along the Cloud-to-Thing Continuum)

Cloud, Fog, Edge and Things



Fog Is Needed Everywhere

Real-Time Adaptive Traffic Control,
Connected/Autonomous Car Apps (safety, Internet access, …)

Positive Train Control,
Real-Time Monitoring,
Internet Access, …

Industrial Control Applications,
Local Data Analytics, …

Local Control and Data Analytics with Intermittent Internet Connectivity

▪▪▪ 5G,
Oil & Gas,
Smart Cities and Homes,
Internet Services,
Robotics,
Smart Grid,
Visual Security,
Drones,
Virtual/Augmented Reality,
Embedded AI,
…

TCP/IP

A standard and universal framework
to
distribute packets

Fog

A standard and universal framework
to
distribute resources and services

plus
Manage, orchestrate, and secure
the distributed resources and services

Why Must We Care About Fog Now?

Fog Fills Critical Technology Gaps and Enable New Services

Address Challenges in Emerging Systems/Apps (IoT, 5G, Imbedded AI, …)	• Stringent latency/delay requirements • Resource constraints (endpoints, network bandwidth, …) • Intermittent network connectivity • Large # and many types of "Things" • Distributed, remote operations by non-IT experts
Empower the Cloud	• Fog as proxy of Things to connect more Things to Cloud • Fog as proxy of Cloud to deliver services to Things
Enable New Services	• Fog-based services • Fog-enabled 5G • Converged Cloud-Fog platforms and services • User controlled Fog services • Fog-enabled dynamic networking at the edge

Fog Will Disrupt Existing Business Models

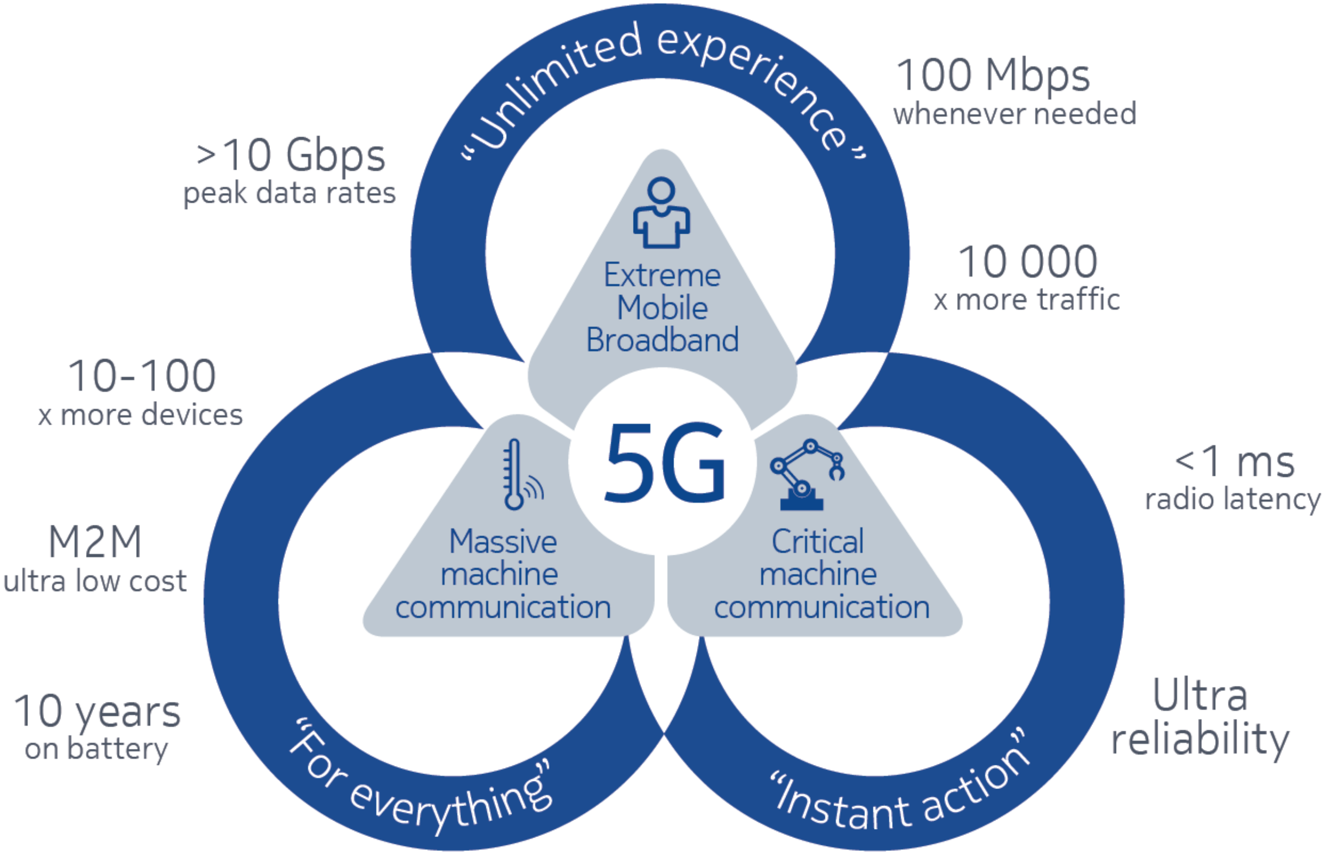
Reshaping Industry Landscape	• Routers, switches, application servers, and storage servers converge into unified fog nodes
Disruptive New Service Models	• Players of all sizes, not just massive cloud operators, build/operate fogs and offer fog services → "WiFi Model" and the rise of local/regional fog eco-systems and operators?
Integrated/Converged Cloud–Fog Services	• For a business to function as a cohesive whole, cloud and fog will converge into one common infrastructure for integrated and unified cloud and fog services: development, deployment, monitoring, management, security, …
Rapid Development and Deployment of Fog Systems and Applications	• Rapid deployment of localized applications → shifting from "build the cloud and see what services we can put on it" to "find what customers want and quickly put together a fog for them"

5G: a Game Changer

上海科技大学
ShanghaiTech University
地址：上海市浦东新区华夏中路393号 邮编：201210
Add: 393 Middle Huaxia Road, Pudong, Shanghai 201210, China

5G Technical Requirements

- Can **one** 5G network satisfy **all** diversified requirements?
- How to make 5G networks super flexible and adaptive?



Source: Nokia

FCC, July 14, 2016

- U.S. leadership in 5G is a national priority.
- There are others around the world who are saying, “No, we want to figure out what the standards are and then figure out how to do the spectrum.” We think that’s backwards.

Tom Wheeler, FCC Chairman

Licensed			Unlicensed
27.5GHz-28.35GHz	37GHz-38.6GHz	38.6GHz-40GHz	64GHz-71GHz

Source: FCC

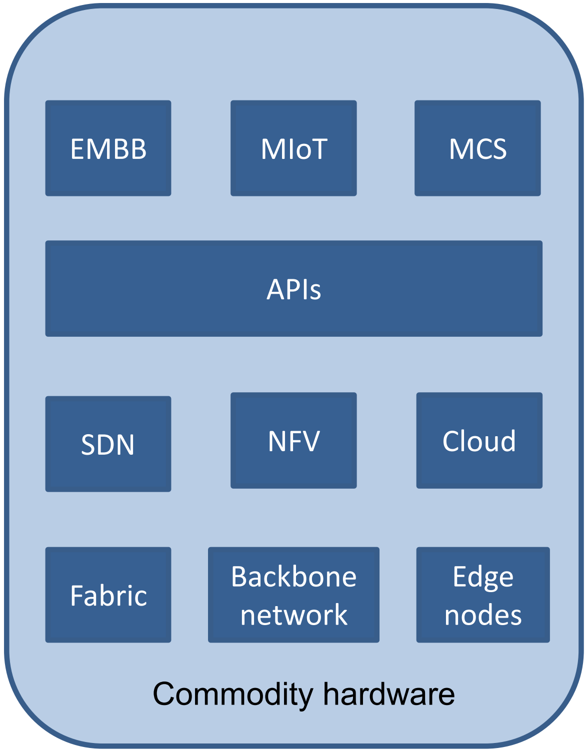
White House, July 15, 2016

- Advanced Wireless Research Initiative, USD 400 million, led by the NSF.
- **Deployment of four city-scale testing platforms for advanced wireless research.**
- (To) allow academics, entrepreneurs, and the wireless industry to test and develop advanced wireless technology ideas, some of which may translate into key future innovations for 5G and beyond.

Strong support from public and private sectors	
NSF	DARPA
NIST	NTIA
AT&T	Carlson Wireless
HTC	CommScope
Intel	InterDigital
NI	Juniper Networks
Nokia	Keysight
Oracle	Qualcomm
Viavi	Samsung
Sprint	Shared Spectrum
Verizon	T-Mobile
ATIS	CTIA
TIA	Source: White House

Google: target at 5G networks

- Google is partnering with leading mobile network operators globally, including Bharti Airtel and SK Telecom, to build a platform for operators to run their network services
- Google will bring their expertise in SDN, NFV and Cloud to the carrier ecosystem, thus accelerate the transition to 5G and enable new features such as the application of machine learning
- The platform will provide plenty of APIs which will enable new operational models and help operators bring new features
- The platform is based on commodity hardware instead of dedicated hardware provided by telecom manufacturers

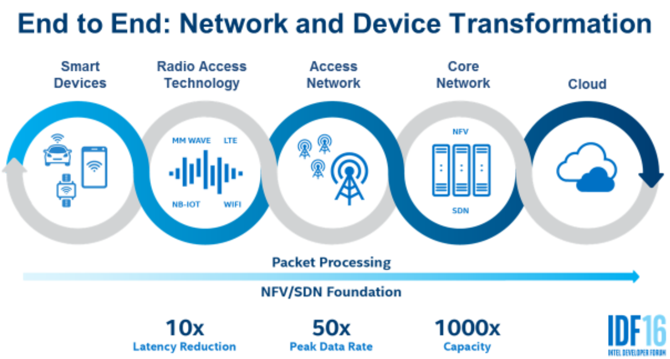


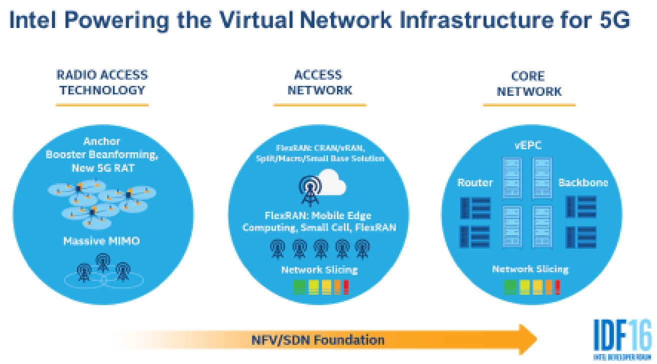
Google Edge Nodes

Map of metros where at least one Edge node (GGC) is present.

Intel's 5G Strategy

- Provide a full suite of products for covering almost every part of the new networks that will all seamlessly interact
- 5G networks will have to be designed to be more flexible, relying on software that can be reprogramming to handle different tasks running on more generic hardware, instead of being built on more customized hardware dedicated to specific tasks
- Links between different parts of the 5G network all made by Intel will be able to interact more efficiently and quickly, while Intel software gives users a smooth experience





Source: Intel

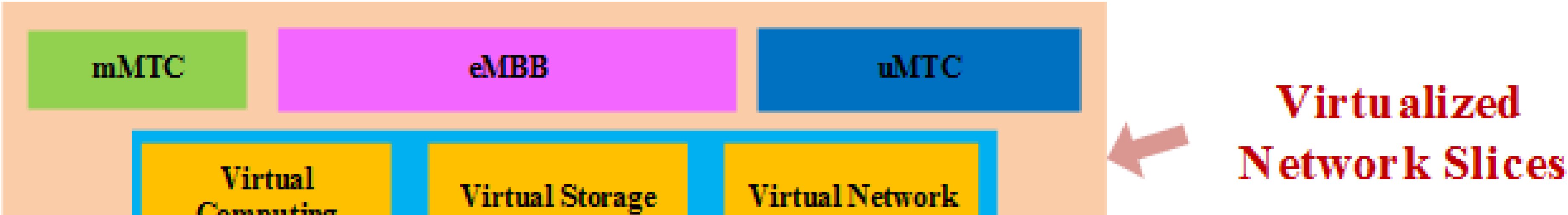
TIP, February 22, 2016

- The Telecom Infra Project (TIP) is an engineering-focused initiative driven by operators, infrastructure providers, system integrators and other technology companies that aim to reimagine the traditional approach to building and deploying telecom network infrastructure.
- Focus areas: access, backhaul, and core and management.
- Open and collaboration!

Members (growing)		
AMN	ACACIA	IP access
ADVA	Amarisoft	Juniper
ASOCS	Aricent	LEMKO
AW2S	Athonet	Lumentum
Axiata	BaiCells	MTN
Bandwidth	BlueStream	Nexius
Broadcom	Coriant	Nokia
EE	T-Mobile	Quortus
Equinix	Facebook	Radisys
Globe	Harman	Horizon
HCL	SK Telecom	iDirect
SS7	Starsolutions	Sysmocom
Intel	Indosat	Telefonica

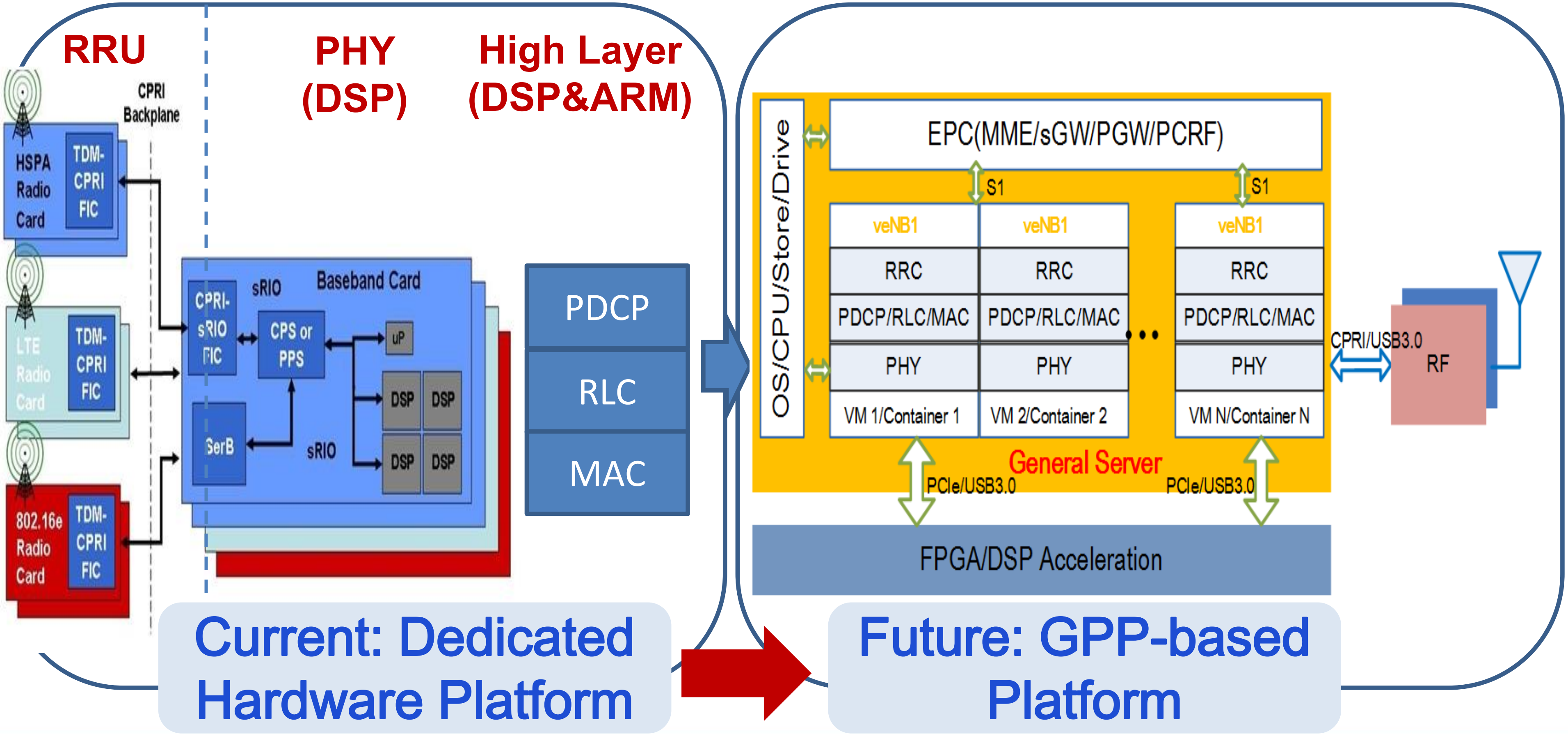
5G Vision: GPP-based Platform

- Software defined mobile network and resource/network function virtualization could meet different diversified 5G use cases and business models, i.e. eMBB, mMTC and uMTC.



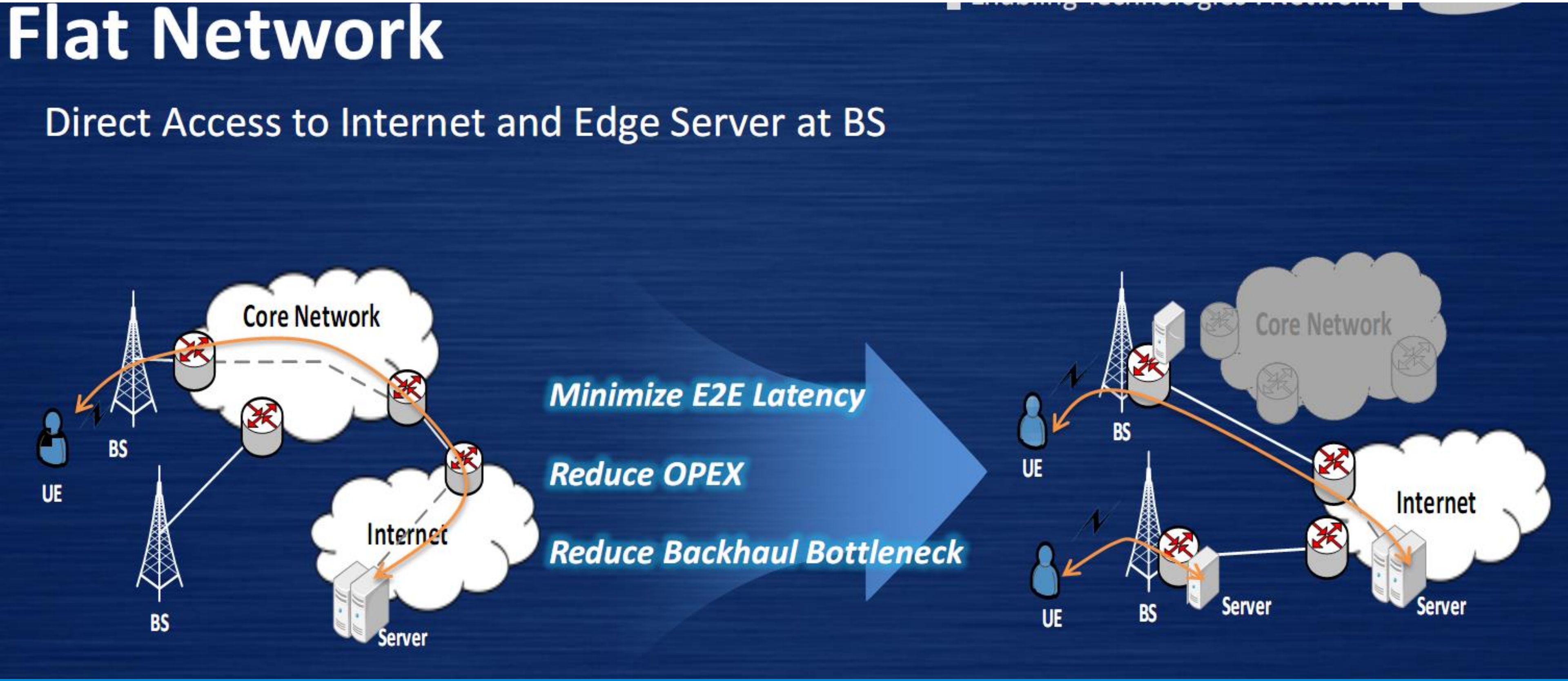
Motivation: Flexible and Adaptive

- To decouple software and hardware designs
- To realize flexible deployment of network functions



5G Vision: a flat network

- Dedicated core network: Again? No!
- EPC and Internet convergence: Yes!

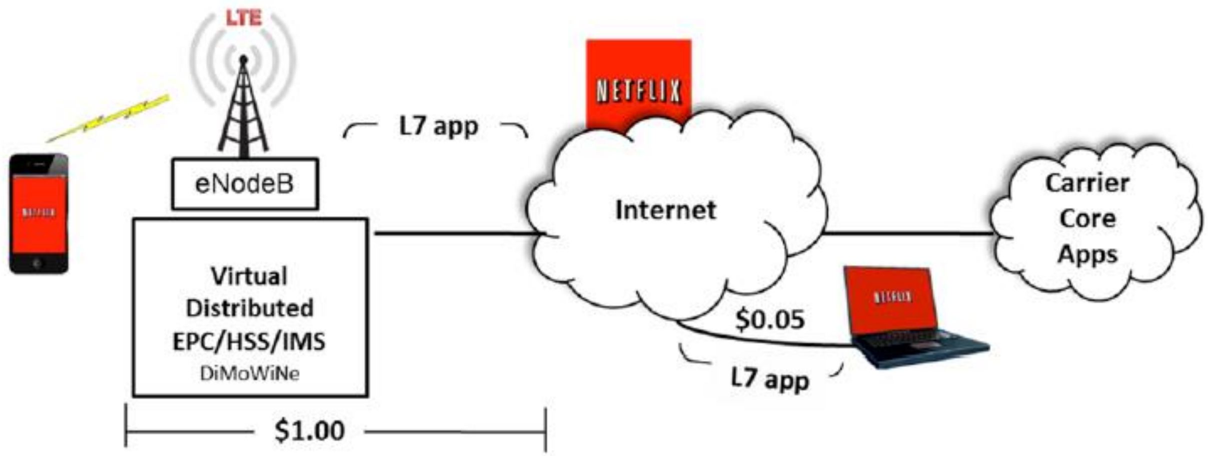


Motivation: Efficiency and Cost

- Good News for Mobile Internet Users

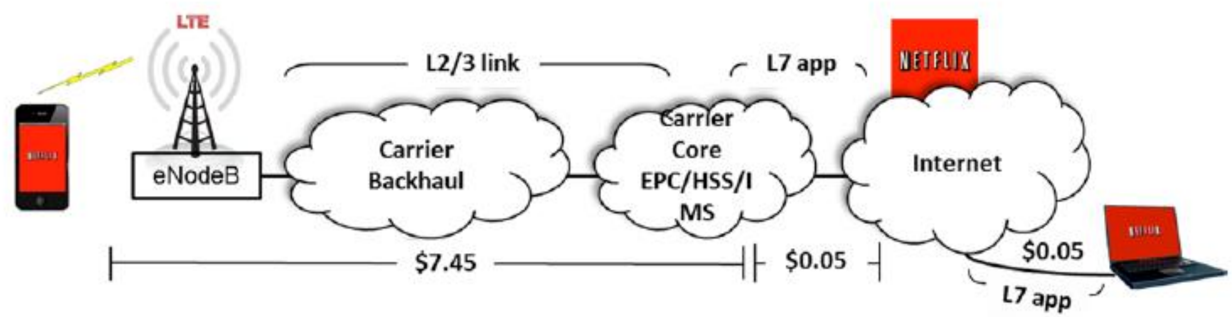
Future

Virtual, Distributed, SDN



Current

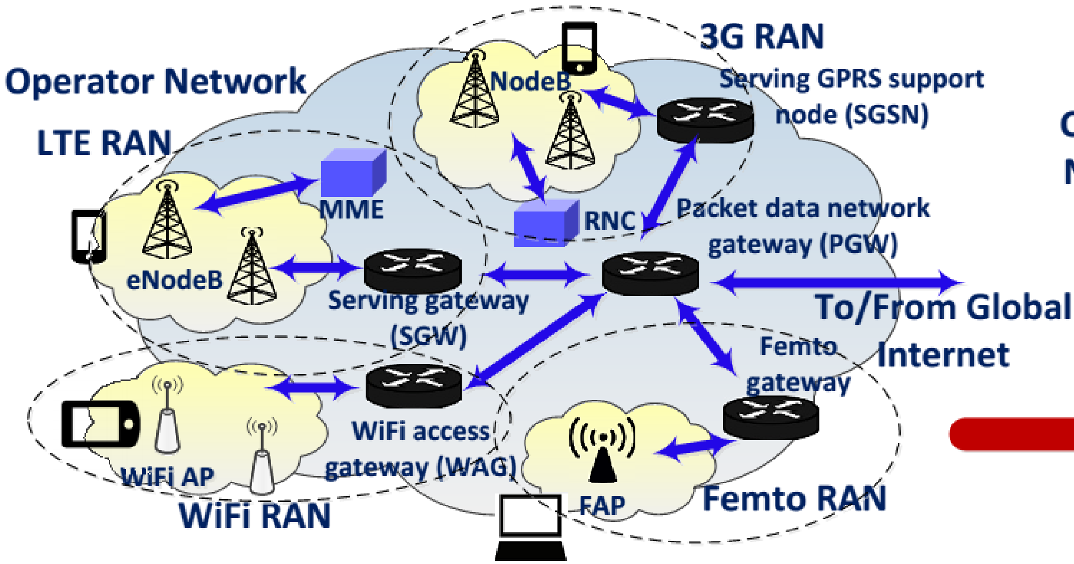
Traditional LTE Centralized EPC

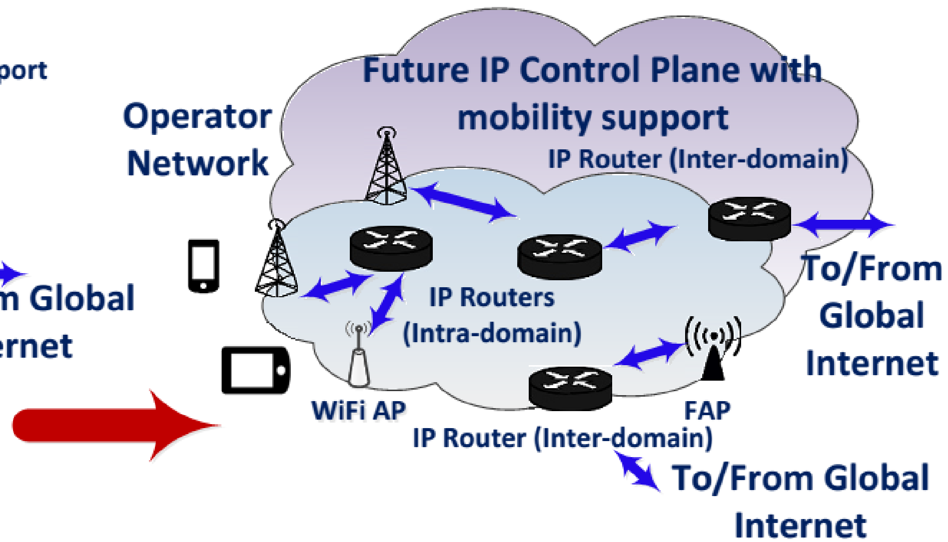


EPC and Internet Convergence

Current:
Separated, Isolated, Closed, Dedicated, Layered Management

Future:
Integrated, Collaborative, Shared, Flat, User Centric

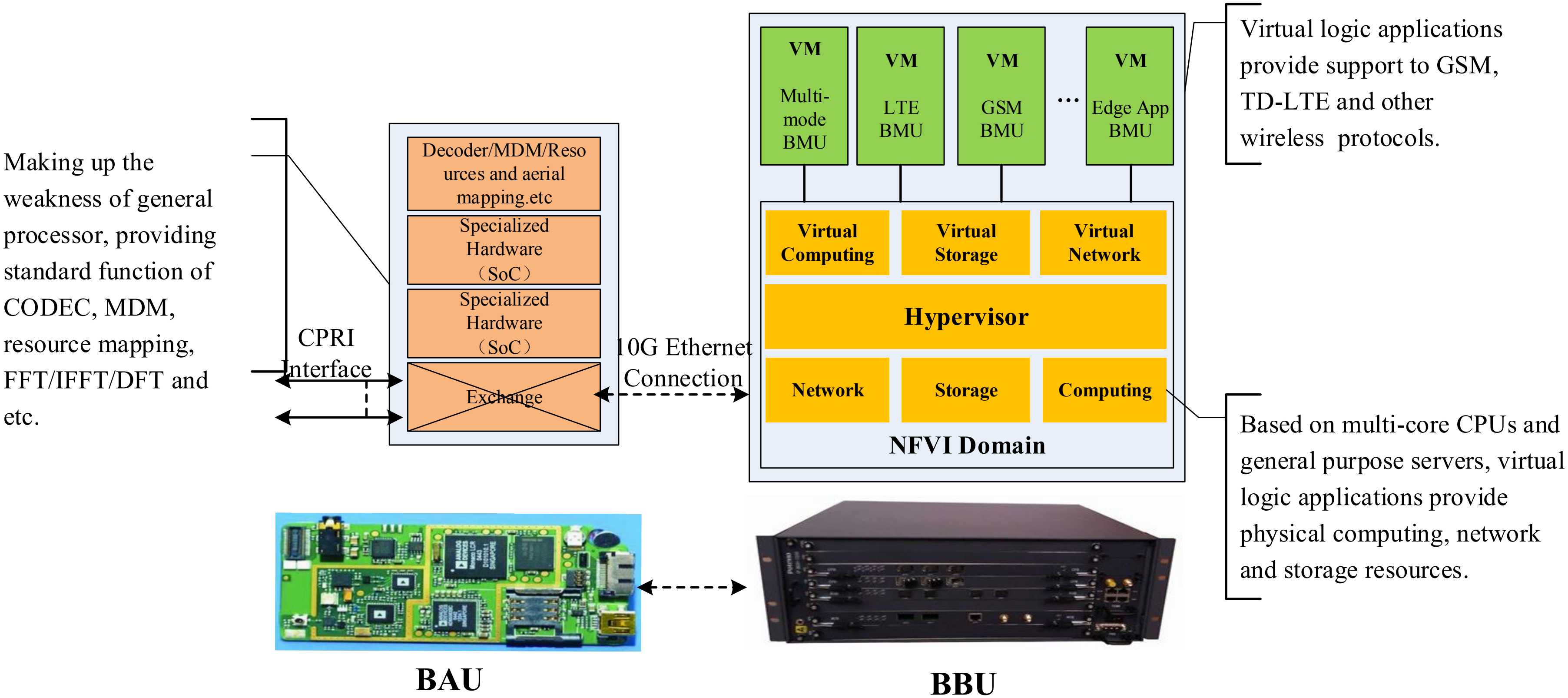




Outline

- 5G Vision: a user-centric flat network
- Approach: software defined mobile network
- Challenges:
  - Real-time processing
  - Industry applications
  - Your participation

Software Defined RAN



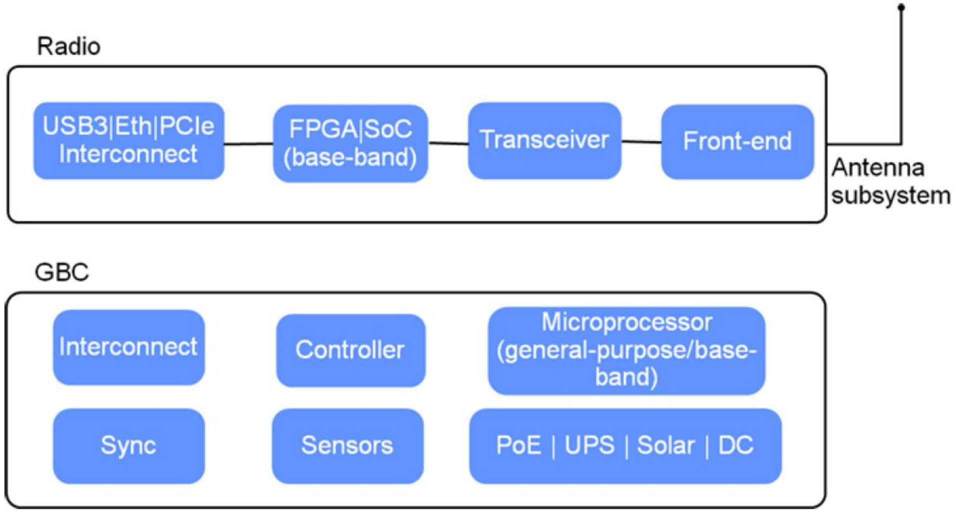
Dedicated accelerations with FPGA and DSP

Virtualization of baseband resources

Source: Alcatel-Lucent Shanghai Bell

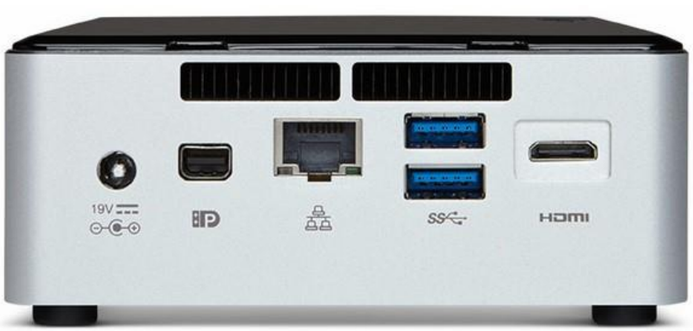
Facebook OpenCellular: an Open Source Wireless Access Platform

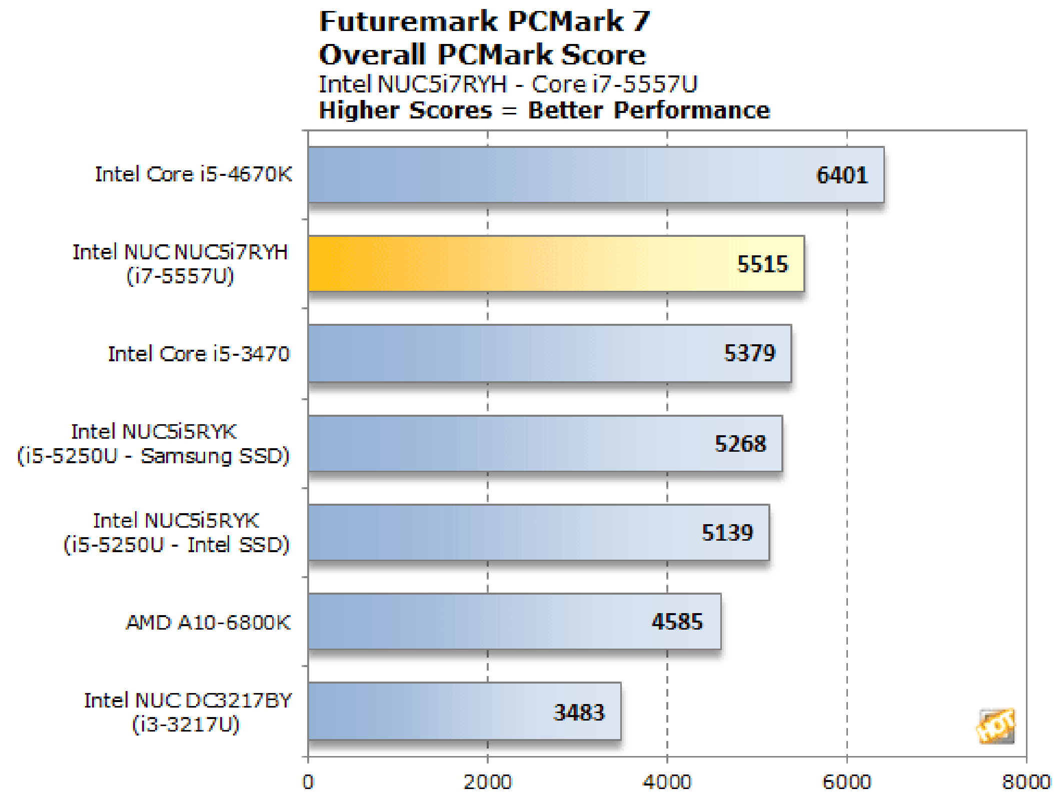
- **Radio:** Radio with integrated front-end, which is based on SDR/SoC and supports network-in-a-box or access point.
- **GBC:** General Baseband Computing
- **Function:** SMS messages, voice calls, basic data connectivity using 2G implementation.



Source: Facebook

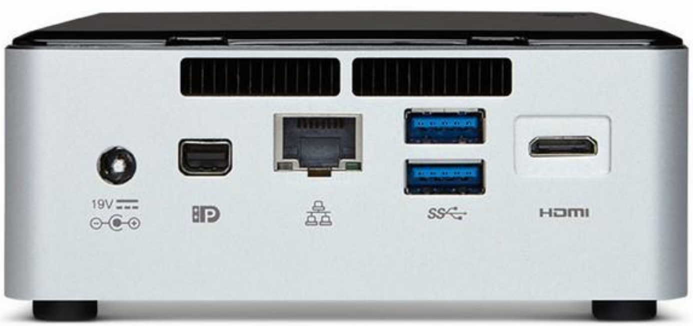
it is a just mini PC





Source: Internet

You think it is a just mini PC

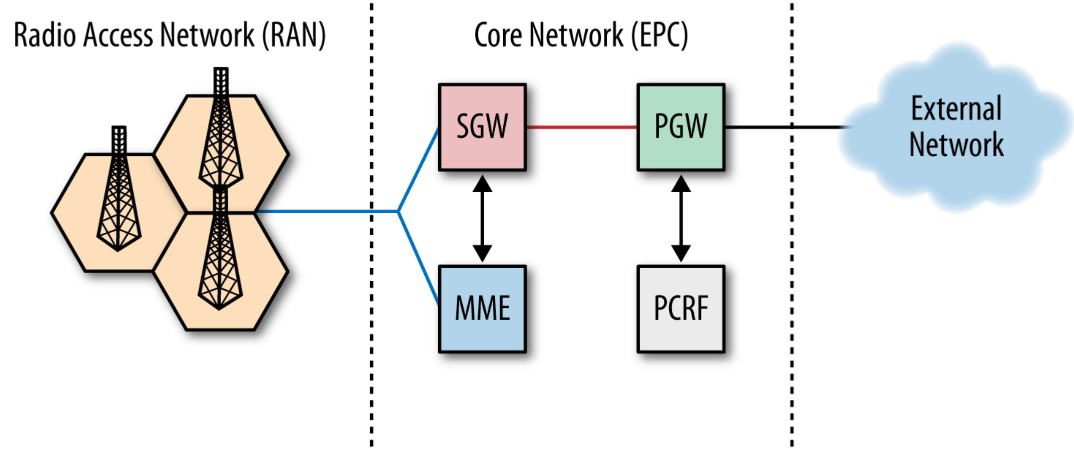


=

eNodeB

+

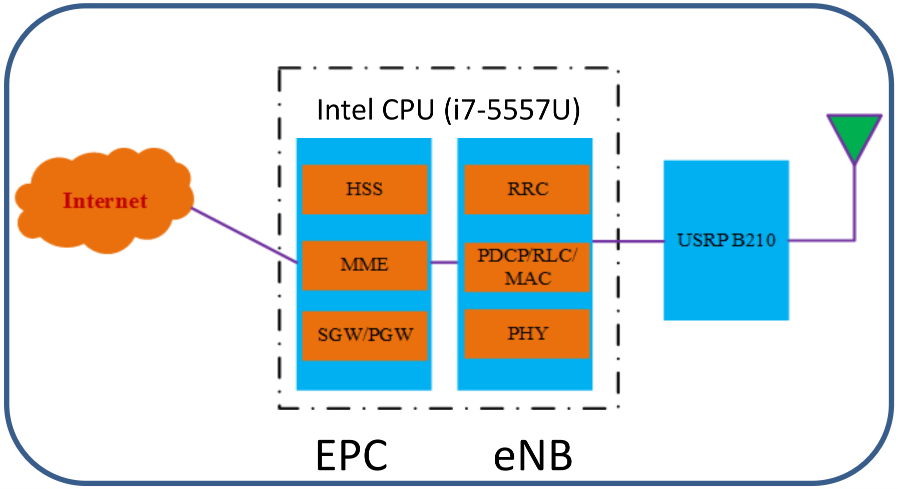
EPC



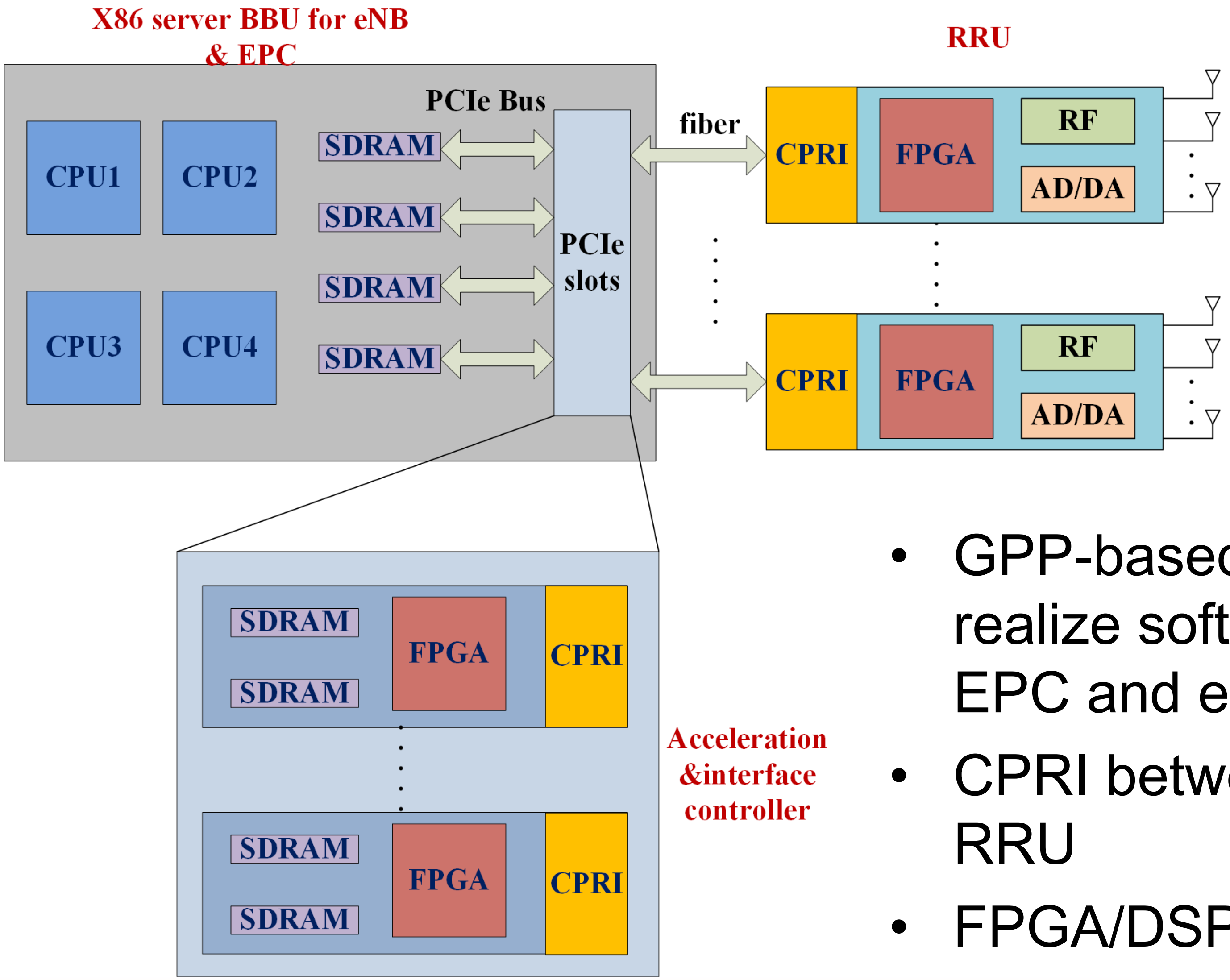
Source: Internet

Software Defined Mobile Network

- Based on OAI open-source LTE platform
- Real-time software defined LTE network (including RAN and EPC) on a multi-core GPP-based platform
- FDD and TDD modes
- Support multiple commercial LTE mobile terminals for each eNB
- Support video streaming and web browsing traffic



Open 5G Platform Architecture



- GPP-based platform to realize software defined EPC and eNB BBU
- CPRI between BBU and RRU
- FPGA/DSP Acceleration

Outline

- 5G Vision: a user-centric flat network
- Approach: software defined mobile network
- Challenges:
  - Real-time processing
  - Industry applications
  - Your participation

System-level simulation is SLOW

- Computation complexity increases exponentially as more antennas are adopted; wireless channel characteristics and interference calculation are time-consuming and resource-hungry.

One TTI (1ms)	FFT	Matrix Inversion	Matrix Multiplication
Computation Complexity	798474240	817152	13074432

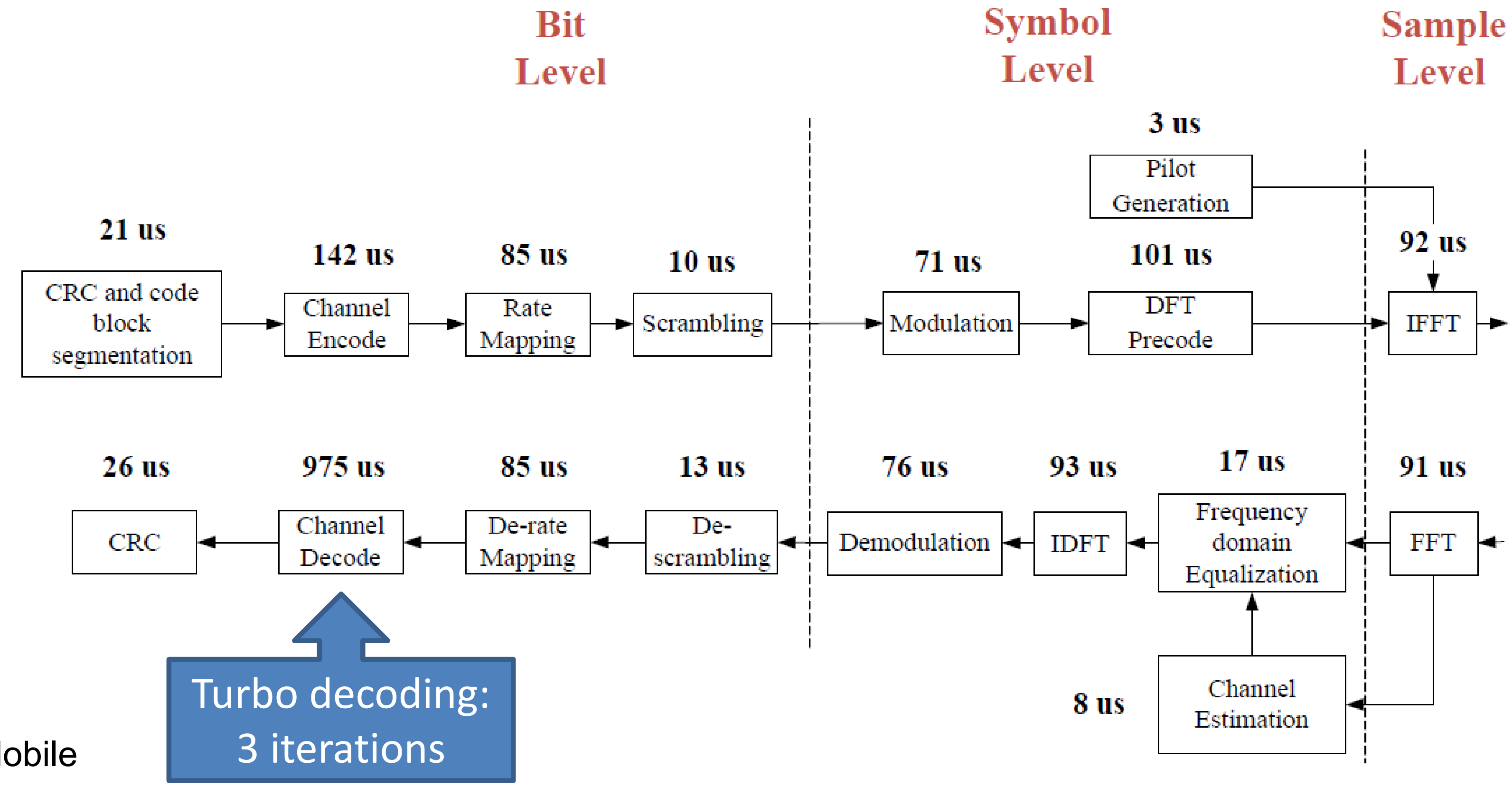
Simulation of one TTI (1ms) in real systems

19*3*15 (Cell*Sector*User)	Total Time (sec)	Interference Calculation + Pre-coding (sec)	Message Processing (sec)
Serial Simulation	8228.76	8037.46	156.56
Parallel Simulation	64.66	12.35	52.31
Hardware in the loop	≈127	0.012	≈1000

Hardware Platform: Intel Xeon Ivy Bridge E5, 256GB, USRP-RIO 2593, PXImc

Delay of baseband signal processing

- TD-LTE uplink and downlink on a GPP-based platform;
- Multi-core parallel computing achieves real-time requirements.



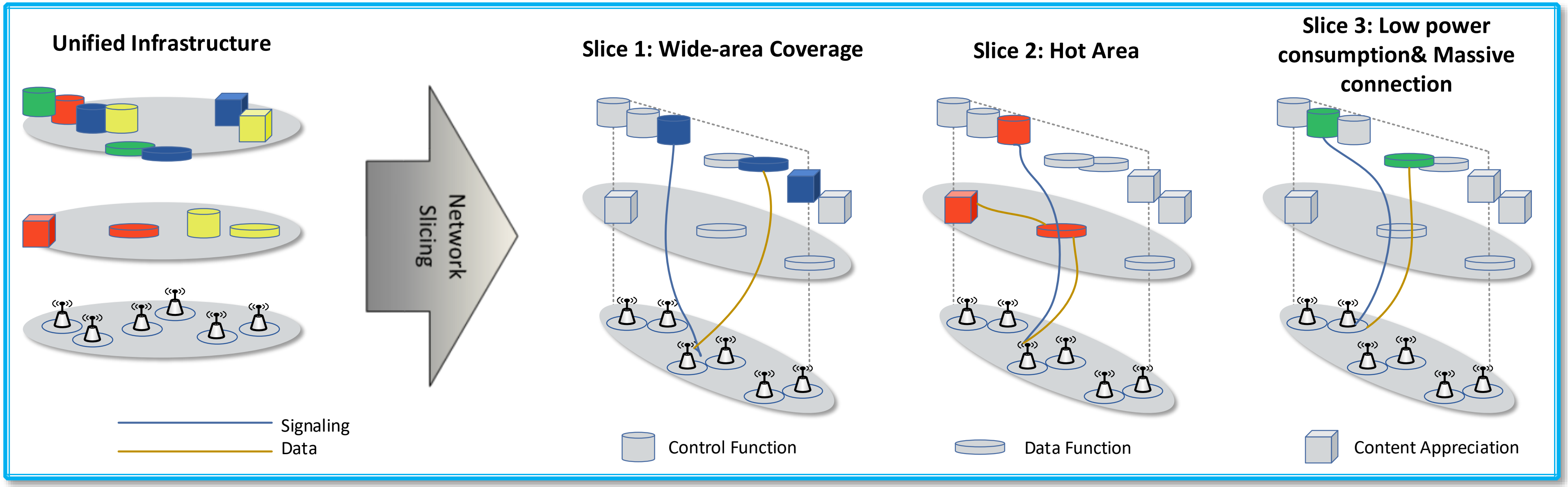
Source: China Mobile

Delay of baseband signal processing

- Our GPP-based platform: IBM System x3400 M3 with 2.13GHz CPU, quad-core Intel Xeon E5606, 4G RAM, 256G HDD, Linux Debian 7 OS with the version 64 bits Ubuntu 14.04 DeskTop.
- Turbo decoding is the bottleneck for real-time processing.

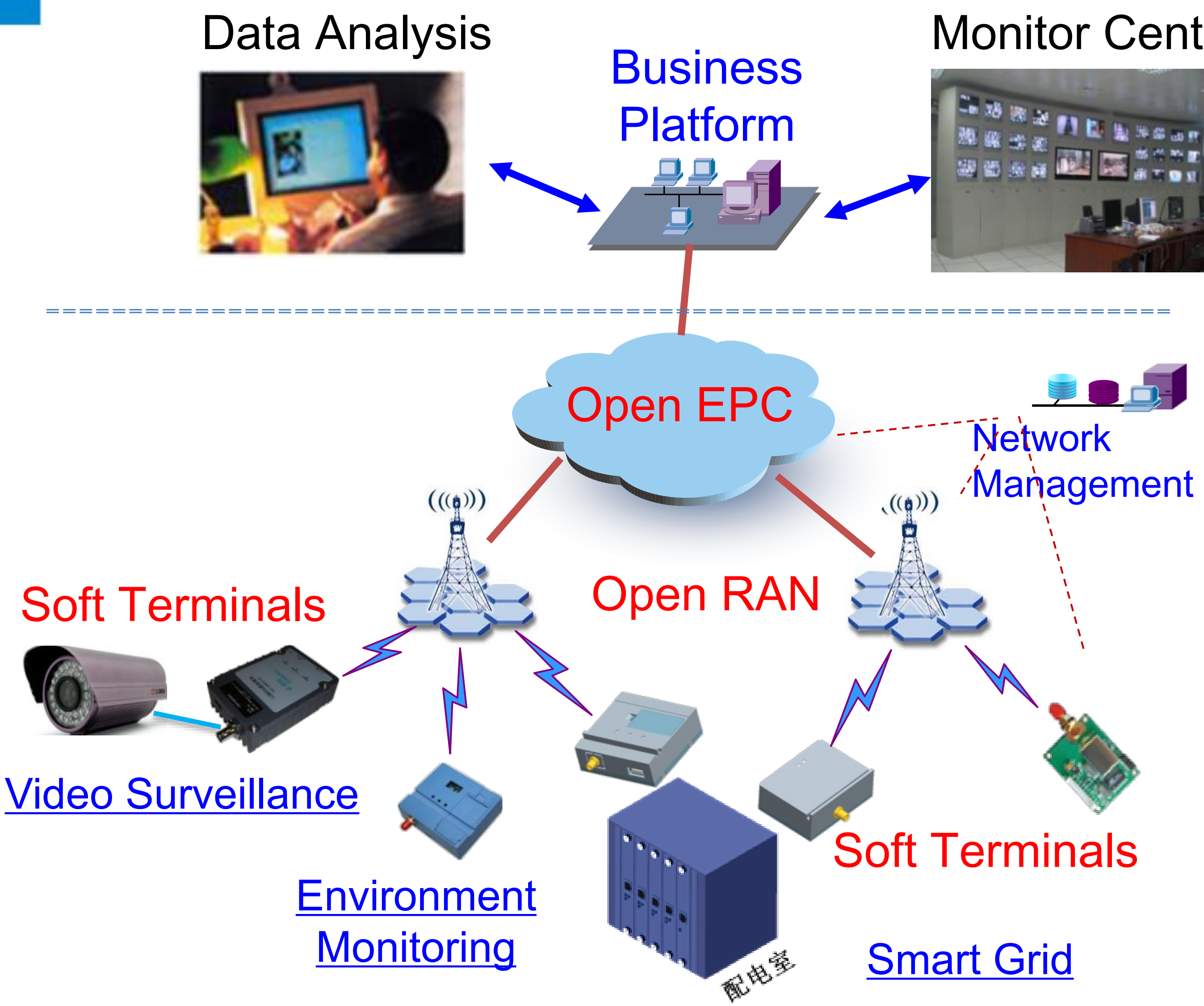
Rate (Mbps) / Processing Time(μs) / Function	2.152	8.76	13.536	17.56
De-scrambling	7.96	21.93	33.38	43.26
De-modulation	7.89	13.72	15.94	17.84
De-interleaving	6.27	30.19	48.68	72.11
Turbo decoding	113.44	465.01	734.86	1047.61

Network Slicing for Various Use Cases



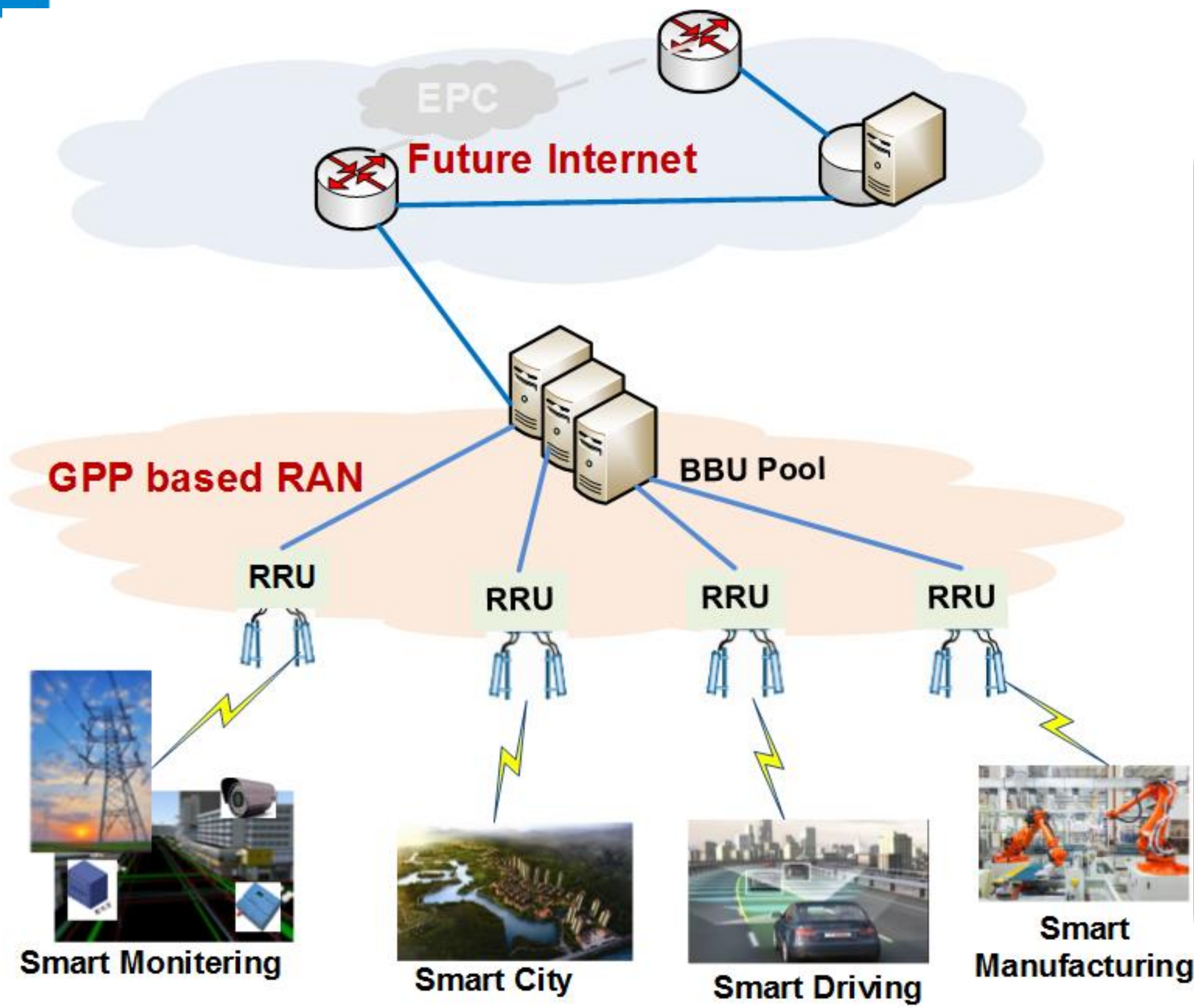
- **Open Source Software:** to build a collaborative community and ecosystem for innovations in EPC, eNB and terminals.
- **GPP-based Hardware:** to replace dedicated hardware (e.g. ASIC), thus enabling flexible and adaptive service creations and deployments for various use cases and business models.

Customized Industry Applications



- GPP-based 5G network supports **fast prototyping** of special industry requirements on soft terminals, open RAN and open EPC.
- **Quick deployment** of dedicated network slices for customized industry applications.

GPP-based 5G Network for IoT Applications



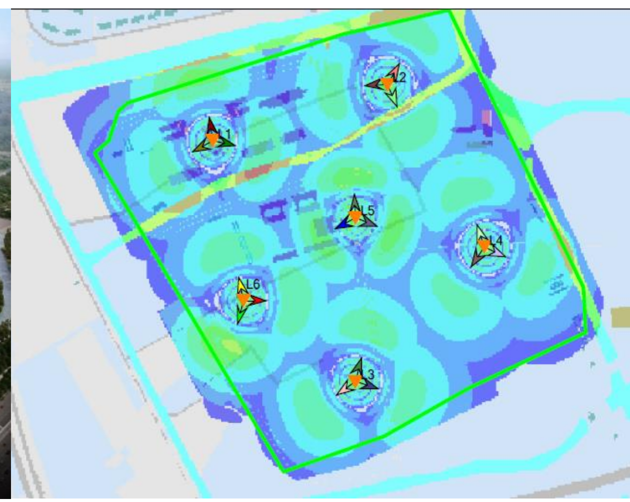
- GPP-based 5G network supports a variety of IoT applications
- Massive and low rate connections
- Low power consumption and depth coverage
- Low latency and high reliability

Heterogeneous Wireless Testbed

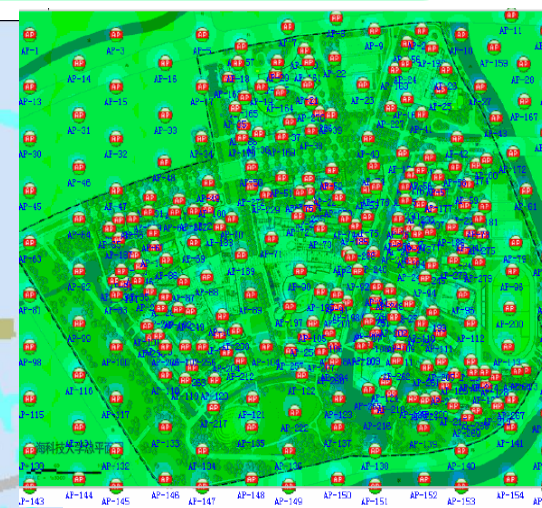
➢ **LTE + 5G hierarchical network architecture**
- 6 macro-cell base stations
- 10~20 micro-cell base stations
- 100+ small base stations
- Trial of GPP-based BSs

➢ **802.11ac high speed WLAN**
- 100~200 outdoor APs
- 1000~10000 indoor APs
- UDN, multi-carriers
- Trial of GPP-based APs

ShanghaiTech University

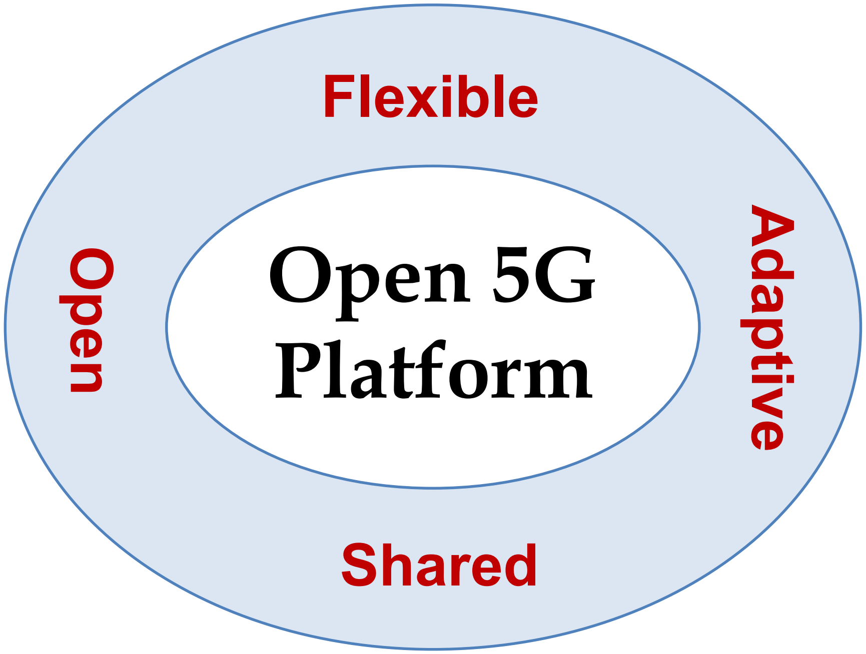


LTE+5G macro-cell BSs



802.11ac outdoor APs

Conclusion: More Innovation and Impact



- **Professor:** evaluation of creative ideas
- **Student:** learning by doing
- **Industry:** fast prototyping and trials of new products
- **Application:** cross-domain customized services