武漢大學 人工智能与软件工程暑期学校

2020-08

演讲人：何 强

Qiang HE received his first Ph. D. degree from Swinburne University of Technology (SUT), Australia, in 2009 and his second Ph. D. degree from Huazhong University of Science and Technology (HUST), China, in 2010. He is a currently an Associate Professor at Swinburne University of Technology. His major research interests include edge computing, software engineering, service-oriented computing and cloud computing. He has published 110+ papers at top venues, e.g., TPDS, TSE, TSC, TCC, TBD, JPDC, ICSE, WWW, ICDE, IJCAI, ICWS, ICSOC and CLOUD. He is recipient of the Best Student Paper Awards at SCC2018, ICWS2017, ICSOC2019, and the Best Paper Awards at ENASE2020 ICSOC2018.

Best Research Paper ICSOC2018, CORE A

Optimal Edge User Allocation in Edge Computing with Variable Sized Vector Bin Packing

Qiang HE

qhe@swin.edu.au

https://sites.google.com/site/heqiang

Outlines

- ***About Us***
- **Background**
- **Problem Identification**
- **Problem Statement**
- **Our Solution**
- **Experiments**

SWINBURNE UNIVERSITY OF TECHNOLOGY

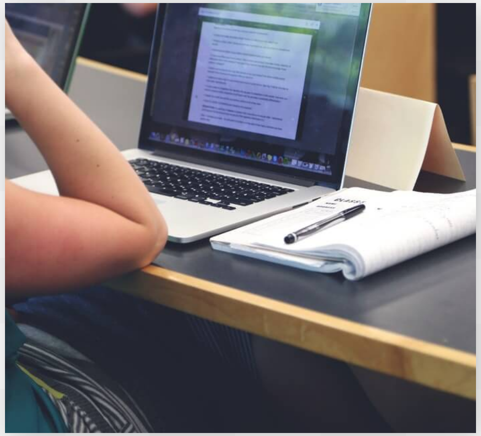
Swinburne University of Technology

45

top young universities in the world by QS World University Rankings 2021

62

2020 Times Higher Education Young University Rankings



327

QS World University Rankings 2021

129

Computer Science & Engineering ranked 129 over the world by Shanghai Ranking

Our Team

Introduction

Our team has 6 (+3) PhD students

Main Research Topics

- ✓ Edge Resources (Since 2018)
  - ✓ Edge User Allocation
  - ✓ Edge Data Caching
  - ✓ Edge Server Placement
  - ✓ Edge Demand Response
- ✓ Edge Security (Since 2019)
  - ✓ Edge Data Integrity
  - ✓ Edge Blockchain
- ✓ Edge AI (Since 2020)

Major Awards

- ✓ Best Research Paper ENASE2020 (CORE B)
- ✓ Best Research Paper ICSOC2018 (CCF B, CORE A)
- ✓ Best Student Paper ICWS2019 (CCF B , CORE A)
- ✓ Best Student Paper ICSOC2018 (CCF B , CORE A)
- ✓ Best Student Paper ICWS2017 (CCF B , CORE A)
- ✓ ...

Our Publications on Edge Computing (Since 2018)

Topic	Paper Title	Venue	Year	Rank	
Edge User Allocation	Optimal Edge User Allocation in Edge Computing with Variable Sized Vector Bin Packing	ICSOC	2018	CCF B	CORE A
	Edge User Allocation with Dynamic Quality of Service	ICSOC	2019	CCF B	CORE A
	A Game-theoretical Approach for User Allocation in Edge Computing Environment	TPDS	2020	CCF A	CORE A*
	Quality of Experience-Aware User Allocation in Edge Computing Systems: A Potential Game	ICDCS	2020	CCF B	CORE A
	Cost-Effective App User Allocation in an Edge Computing Environment	TCC	2020	JCR Q1	
	QoE-aware User Allocation in Edge Computing Systems with Dynamic QoS	FGCS	2020	JCR Q1	
	Interference-aware SaaS User Allocation Game for Edge Computing	TCC	2020	JCR Q1	
Edge Data Caching	Online Collaborative Data Caching in Edge Computing	TPDS	2020	CCF A	CORE A*
	Cost-Effective App Data Distribution in Edge Computing	TPDS	2020	CCF A	CORE A*
	Graph-based Optimal Data Caching in Edge Computing	ICSOC	2019	CCF B	CORE A
	Budgeted Data Caching based on k-Median in Mobile Edge Computing	ICWS	2020	CCF B	CORE A
	Graph-based Data Caching Optimization in Edge Computing	FGCS	2020	JCR Q1	

My Footprints

2007-2010
Dual PhD program

2010-2013
Running business

2013-2014
Post-doc fellowship

2014-2018
Lecturer

2018-2019
Senior Lecturer

2020-present
Associate Professor

My Publications and Awards

- **Journal Papers (54)**
  - **28 x ACM/IEEE Transactions**
  - **2 x TPDS, 5 x TSE, 8 x TSC, 5 x TCC, 5 x TBD**
- **Conference Papers (58)**
  - **CCF A**: IJCAI, ICDE, WWW, ICSE
  - **CCF B**: 8 x ICSOC, 12 x ICWS, 8 x SCC, ICDCS, ICDM, AAMAS

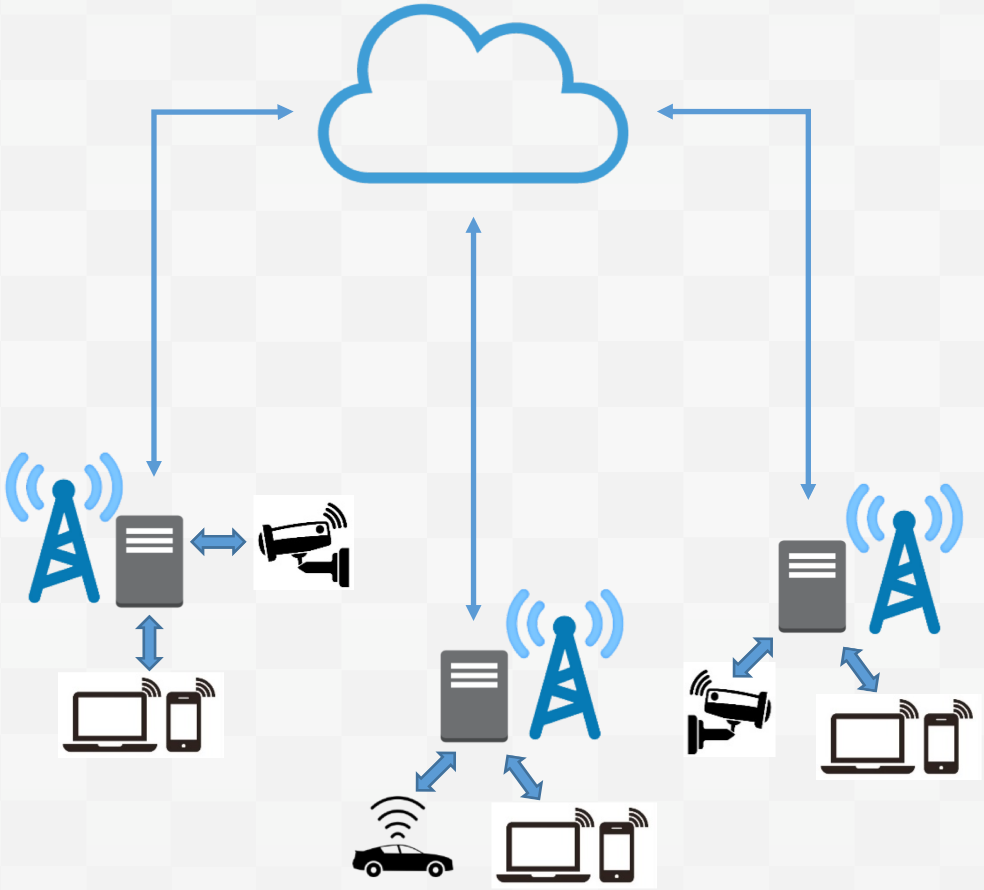
Major Awards

- **Best Paper Award**, ENASE2020
- **Best Student Paper Award**, ICWS2019 (CCF B)
- **Best Paper Award**, ICSOC2018 (CCF B).
- **Best Student Paper Award**, SCC2018
- **Best Student Paper Award**, ICWS2017 (CCF B)
- **FSET MCR Award**, 2020
- **FSET MCR Award**, 2018

Outlines

- About Us
- Background
- Problem Identification
- Problem Statement
- Our Solution
- Experiments

What is edge computing and why?



Advantages

- Computation offloading
- Energy saving
- Low latency

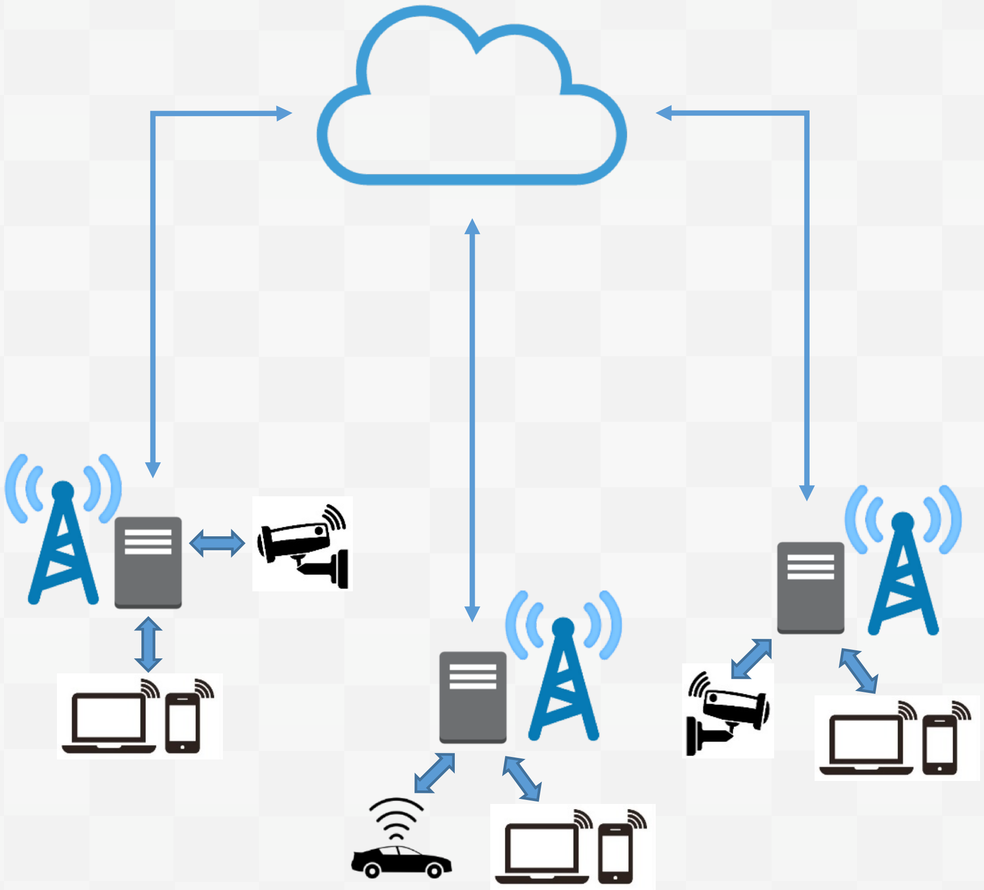
Applications

- Driverless cars
- Mobile gaming
- Augmented reality
- Remote healthcare
- Smart manufacturing
- …

Outlines

- About Us
- Background
- **Problem Identification**
- Problem Statement
- Our Solution
- Experiments

Current Research Problem and Perspectives



Computation Offloading

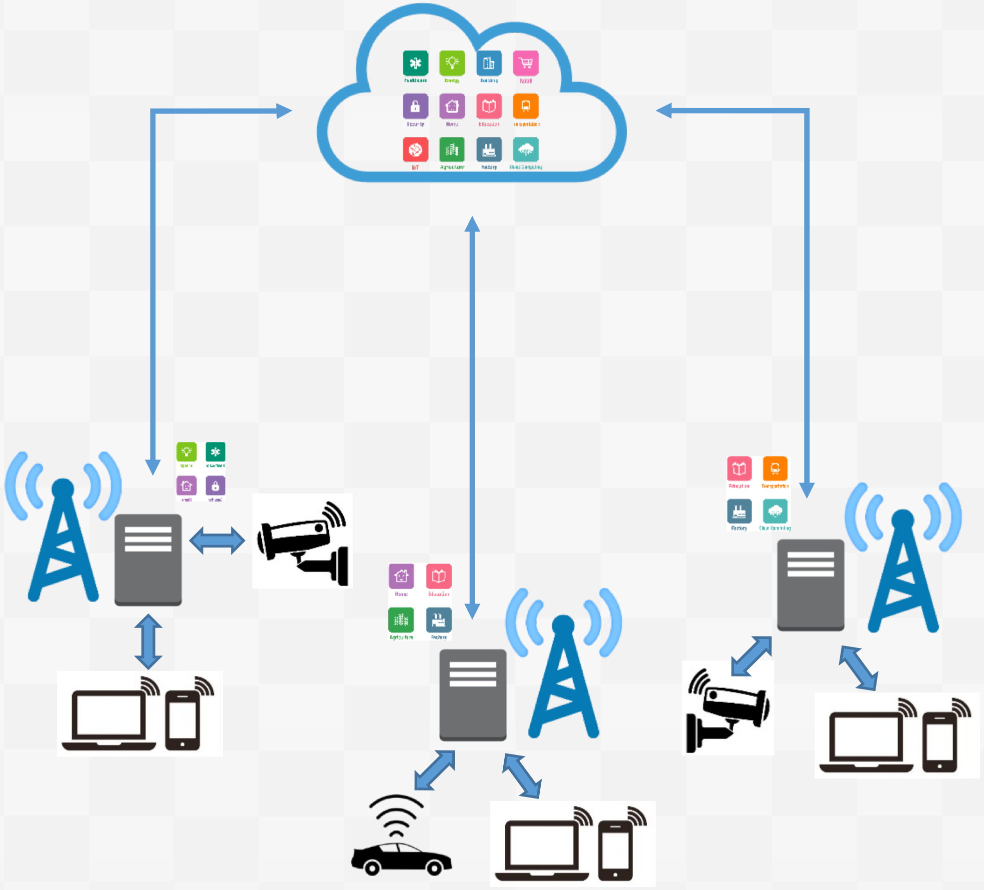
Mobile/IoT users and devices

- Energy efficiency
- Latency

Edge infrastructure provider

- Network throughput
- Workload balance

What about app vendors?



App vendor (service provider, content provider)

- Important stakeholder
- Major users of edge infrastructure

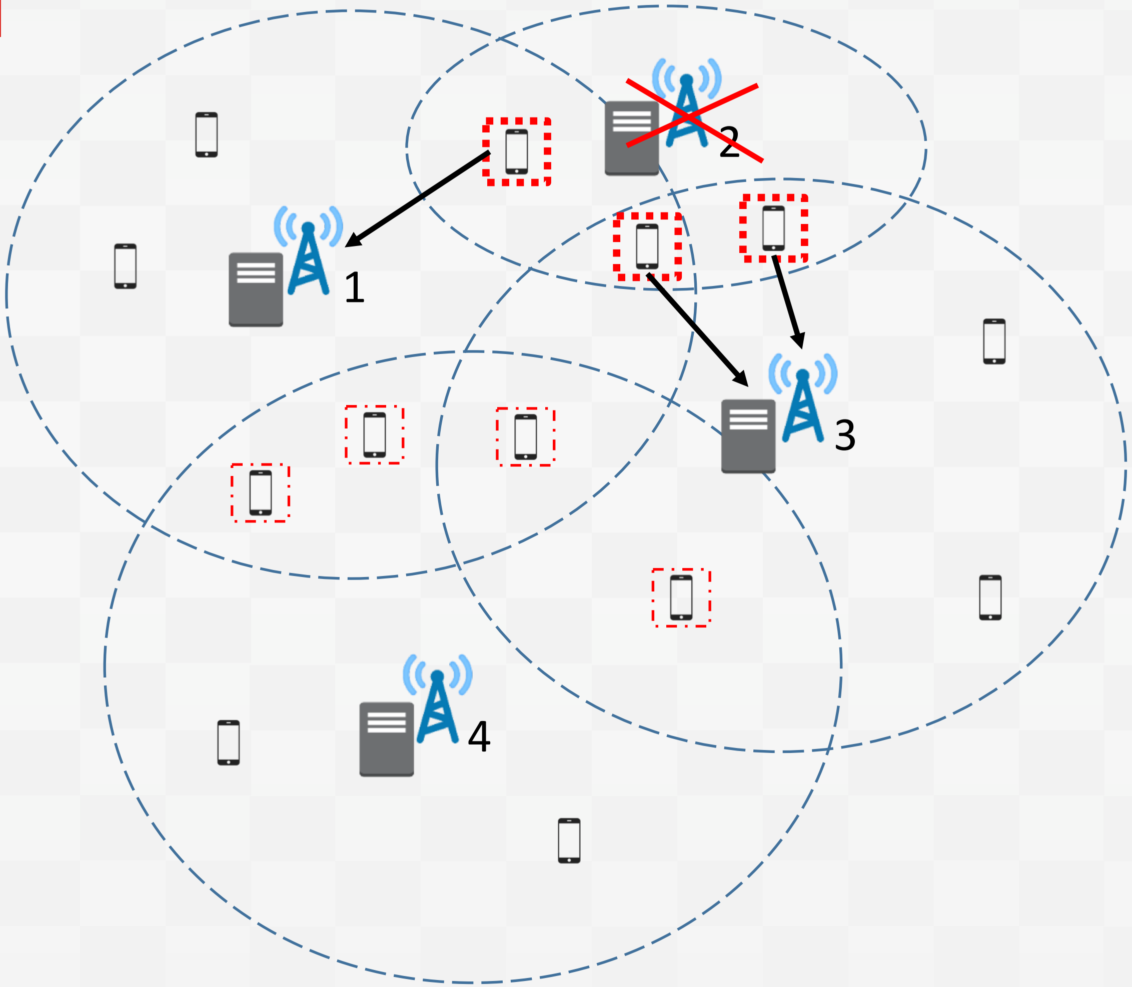
Major concerns

- Benefit by serving app users
- Cost by hiring computing resources on edge servers

Outlines

- About Us
- Background
- Problem Identification
- Problem Statement
- Our Solution
- Experiments

Edge User Allocation Problem



Objectives

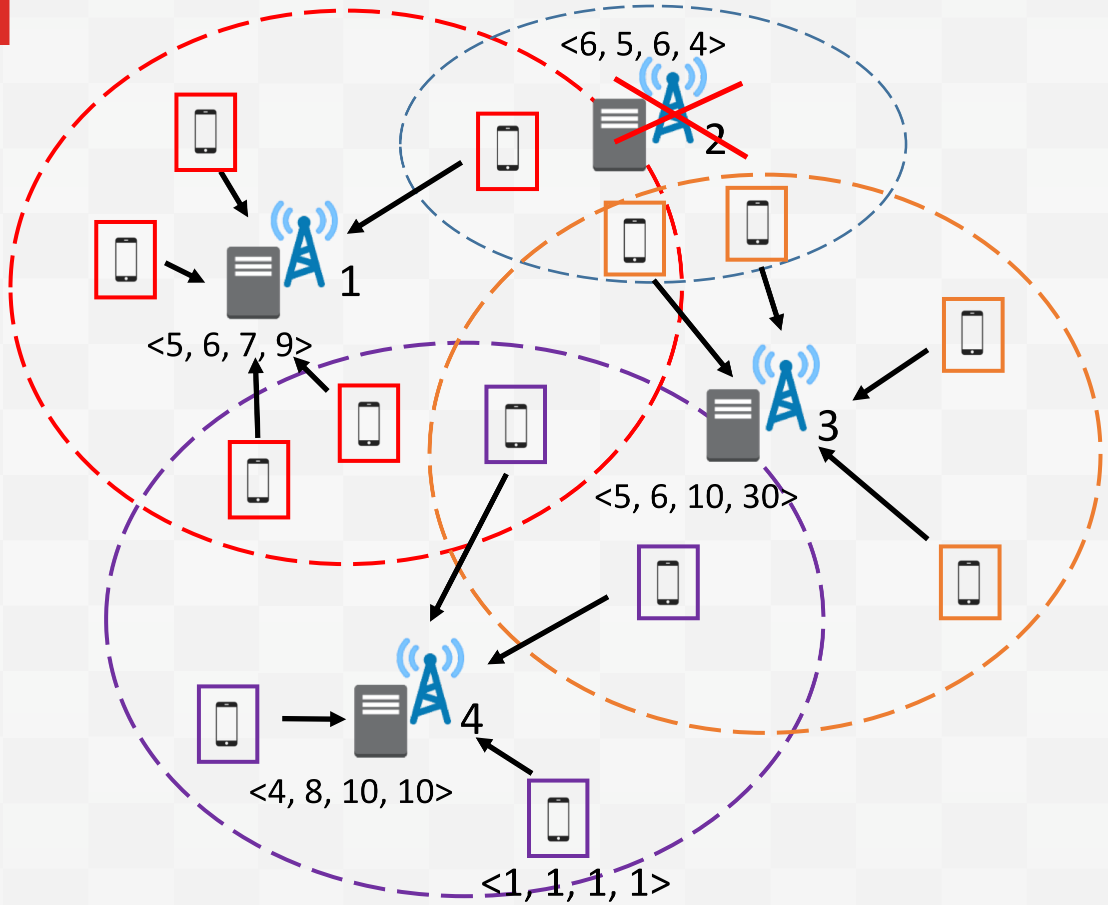
- Maximize number of app users allocated
- Minimize number of edge servers used

Proximity constraint

- An edge server can only serve users within its coverage

SWINBURNE UNIVERSITY OF TECHNOLOGY

Edge User Allocation Problem



Capacity vector: <CPU, RAM, VRAM, Bandwidth>

Objectives

- Maximize number of app users allocated
- Minimize number of edge servers used

Proximity constraint

- An edge server can only serve users within its coverage

Capacity constraint

- Demands of users allocated to an edge server must not exceed its remaining capacities

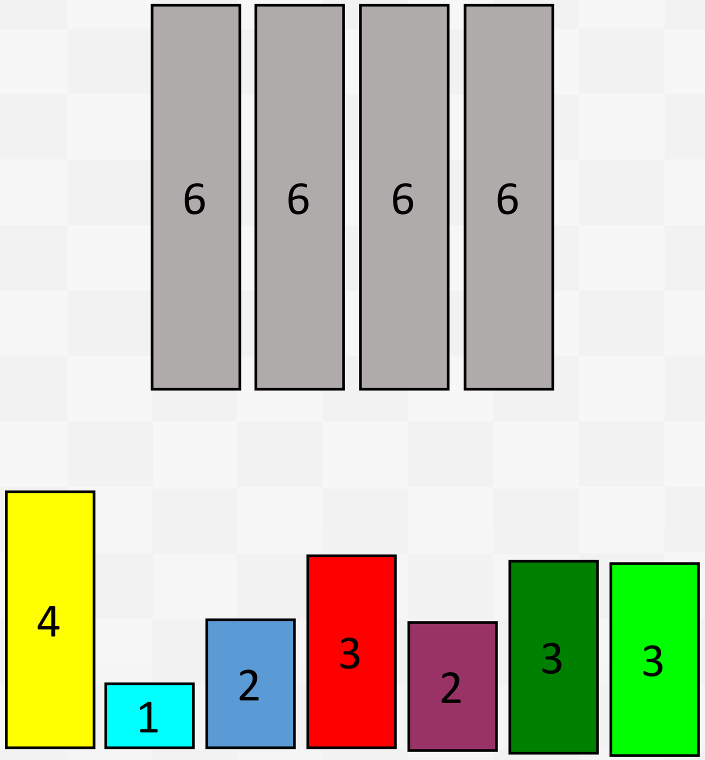
SWINBURNE UNIVERSITY OF TECHNOLOGY

Outlines

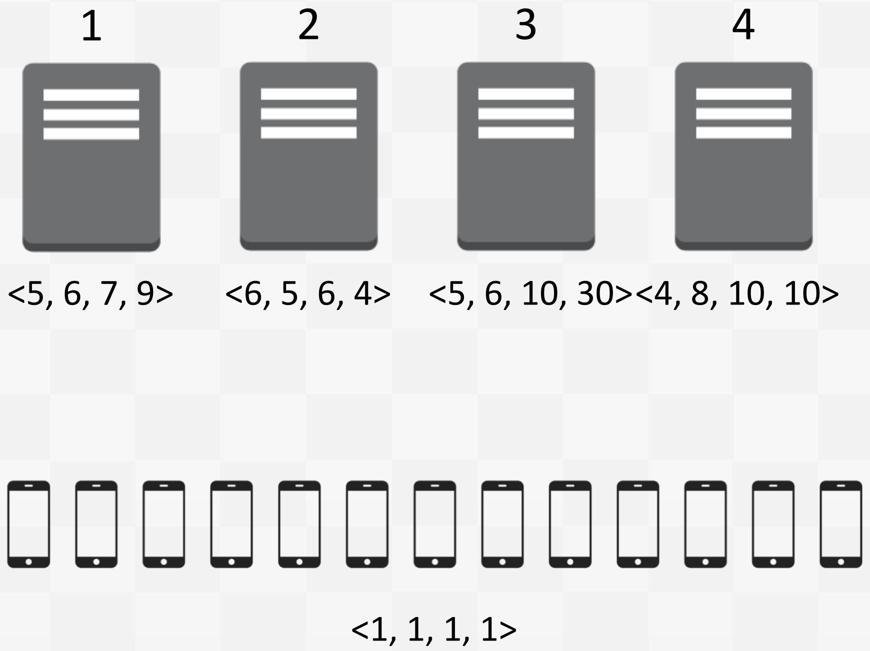
- About Us
- Background
- Problem Identification
- Problem Statement
- Our Solution
- Experiments

Problem Modelling

Bin Packing Problem (*NP*-hard)



Variable Sized Vector Bin Packing Problem (*NP*-hard)



Capacity vector: <CPU, RAM, VRAM, Bandwidth>

Integer Programming Optimization

- **Optimization objectives:**
  - Maximize the number of allocated users
  - Minimize the number of used servers
- **Constraints:**
  - Proximity constraint
  - Capacity constraint
- **Solver:** IBM CPLEX Optimizer

Notation	Description
$S = \{s_1, s_2, ..., s_i\}$	finite set of edge server s_i, where $i = 1, 2, ..., m$
$C_i = \langle C_i^1, C_i^2, ..., C_i^d \rangle$	$d-$dimensional vector with each dimension C_i^k being a resource type, such as CPU utilization or disk I/O, representing the remaining capacity of an edge server s_i, $k \in \{1, 2, ..., d\}$
$U = \{u_1, u_2, ..., u_j\}$	finite set of user u_j, where $j = 1, 2, ..., n$
$w_j = \langle w_j^1, w_j^2, ..., w_j^d \rangle$	$d-$dimensional vector representing the size of the workload incurred by user u_j. Each vector component w_j^k is a resource type, $k \in \{1, 2, ..., d\}$
$U(s_i)$	set of users allocated to server s_i. $U(s_i) \subset U$
d_{ij}	geographical distance between server s_i and user u_j
$cov(s_i)$	coverage radius of server s_i

$$maximize \sum_{j=1}^{n}\sum_{i=1}^{m} x_{ij} \quad (1)$$

$$minimize\ E = \sum_{i=1}^{n} y_i \quad (2)$$

subject to:

$$\sum_{j=1}^{n} w_j^k x_{ij} \leq C_i^k y_i, \forall i \in \{1, ..., n\}; \forall k \in \{1, ..., d\} \quad (3)$$

$$d_{ij} \leq cov(s_i), \forall i \in \{1, ..., n\}; \forall j \in \{1, ..., n\} \quad (4)$$

$$\sum_{i=1}^{m} x_{ij} \leq 1, \forall j \in \{1, ..., n\} \quad (5)$$

$$y_i \in \{0, 1\}, \forall i \in \{1, ..., n\} \quad (6)$$

$$x_{ij} \in \{0, 1\}, \forall i \in \{1, ..., n\}; \forall j \in \{1, ..., n\} \quad (7)$$

where:

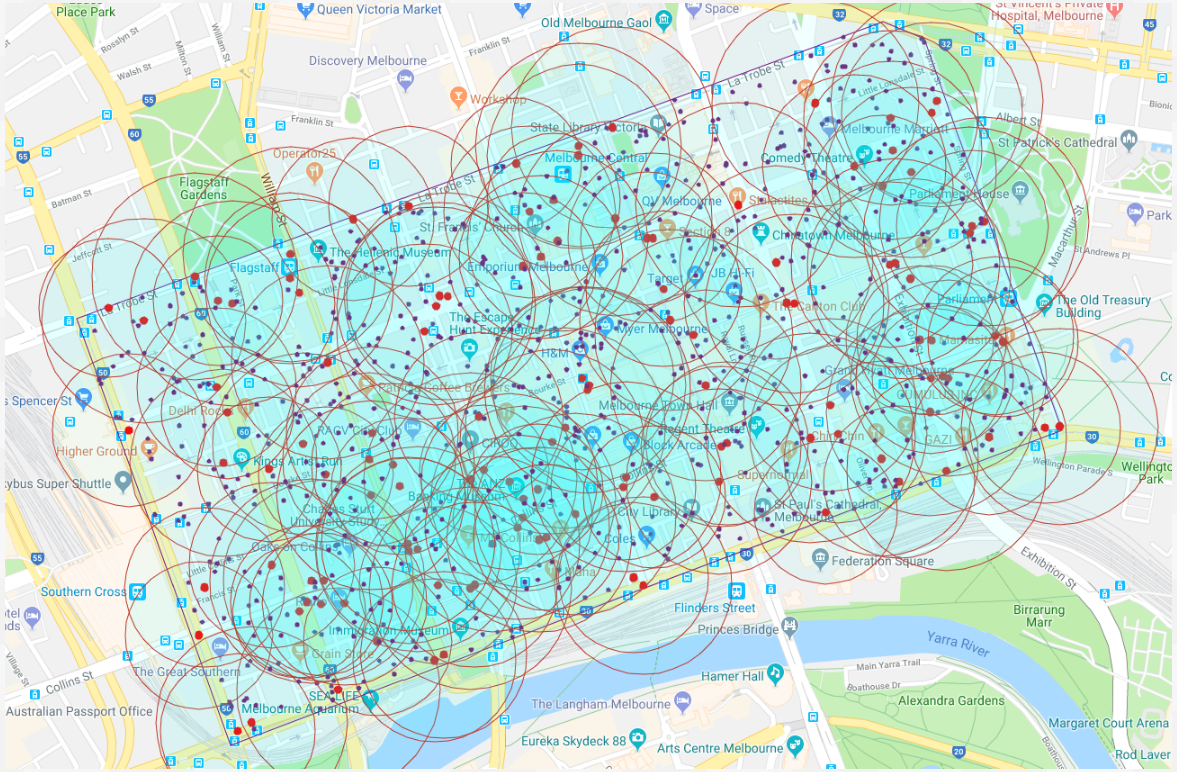
$y_i = 1$ if server s_i is hired.
$x_{ij} = 1$ if user u_j is allocated to server s_i.
$cov(s_i)$ is provided by edge computing providers.

Outlines

- About Us
- Background
- Problem Identification
- Problem Statement
- Our Solution
- Experiments

Experiment Data

- **Area:** Simulated Melbourne CBD area
- **Edge servers:**
  - 125 Telstra base stations in the CBD area.
  - Coverage: 450-750m.
- **End-users:**
  - 550 users in the CBD area.



Datasets: https://github.com/swinedge/eua-dataset or https://sites.google.com/site/heqiang/eua-repository, containing ***95,562*** base stations in Australia and ~***131,000*** users.

SWINBURNE UNIVERSITY OF TECHNOLOGY

Experiment Settings

- **Comparing approaches**
  - **Random**: Randomly allocates end-users
  - **Greedy**: Always allocates the most end-users to an edge server.

- **Parameters settings**
  - **Number of end-users**: randomly select 4, 8, 16, ..., 512 users.
  - **Number of edge servers**: 10%, 20%, ..., 100% of the number of the servers are available.
  - **Remaining server capacity**: 100%, 150%, ..., 300% of the combined user workload are available.

- **Performance metrics**
  - **Percentage of users allocated**
  - **Percentage of used edge servers**
  - **Execution time (CPU time)**

Table 2: Experiment Settings

Factor	Number of users	Number of servers	Remaining server capacity
Set #1	4, 8, ..., 512	100%	300%
Set #2	512	10%, 20%, ..., 100%	300%
Set #3	512	100%	100%, 150%, ..., 300%

Experiment Results

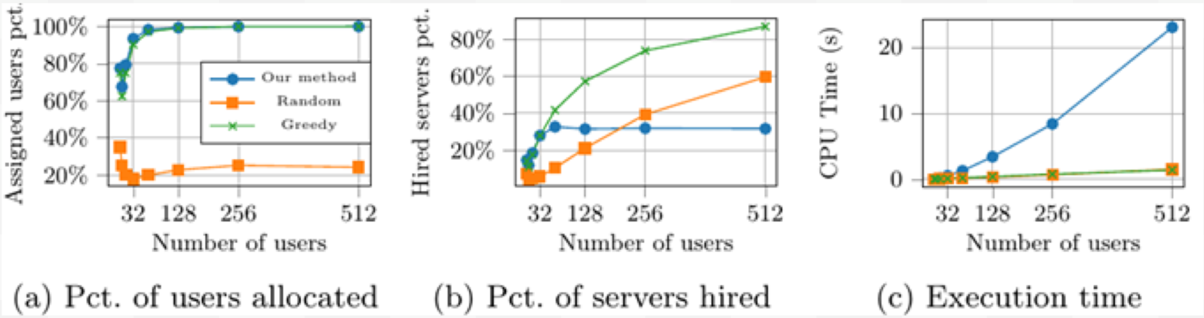


Fig. 1 Results of experiment set #1

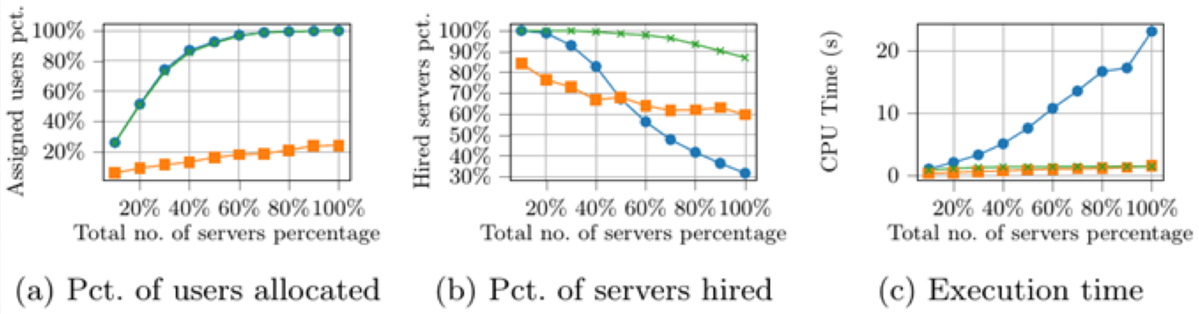


Fig. 2 Results of experiment set #2

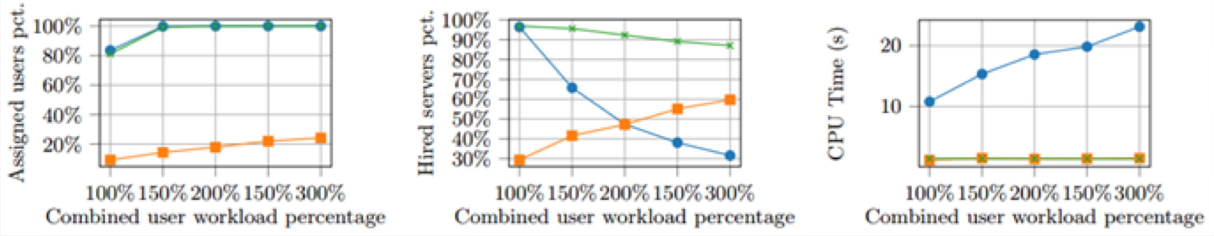


Fig. 3 Results of experiment set #3

SWINBURNE UNIVERSITY OF TECHNOLOGY

Our Major Contributions

- ✓ Identify the new edge user allocation (EUA) problem
- ✓ Model EUA as Variable Sized Vector Bin Packing Problem
- ✓ Propose a solution based on Integer Programming

Q & A

More Work on Edge User Allocation

- Edge User Allocation with **Dynamic Quality of Service** (ICSOC2019, CCF B)
- A **Game-theoretical** Approach for User Allocation in Edge Computing Environment (TPDS2019, CCF A, JCR Q1)
- **Quality of Experience-Aware** User Allocation in Edge Computing Systems: A Potential Game (ICDCS2020, CCF B)
- **Cost-Effective** App User Allocation in an Edge Computing Environment (TCC2020, JCR Q1)
- **QoE-aware** User Allocation in Edge Computing Systems with **Dynamic QoS** (FGCS2020, JCR Q1)
- **Interference-aware** SaaS User Allocation Game for Edge Computing (TCC2020, JCR Q1)

THANK YOU

Swinburne University of Technology

qhe@swin.edu.au

http://sites.google.com/site/heqiang