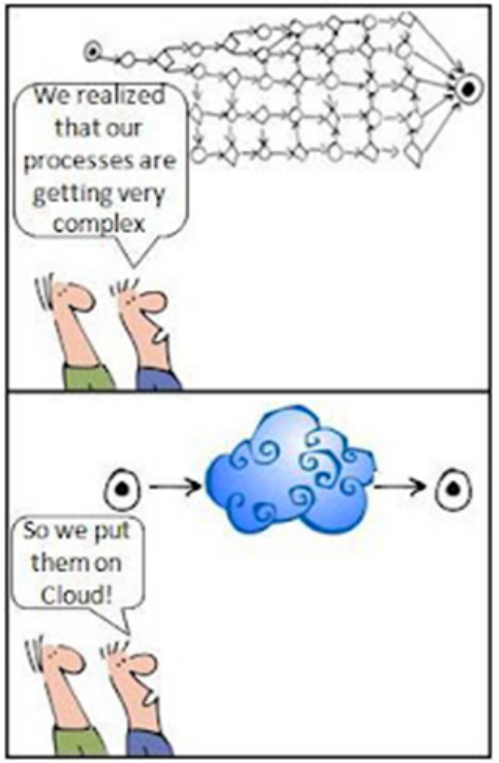
Mobile Cloud Computing and Probabilistic Cloud Performance Diagnosis and Prediction

Asst. Prof. Karan Mitra
Luleå University of Technology
Skellefteå, Sweden
karan.mitra@ltu.se
https://karanmitra.me

23rd December 2016



© http://tinyurl.com/nrqpww

Agenda

- Introduction
- M2C2: A Mobility Management System for Mobile Cloud Computing
- ALPINE: A Bayesian System for Cloud Performance Diagnosis and Prediction
- Summary

Introduction

- Cloud Computing

  *"Cloud computing is a model for enabling ubiquitous, convenient, on-demand **network access** to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."* [NIST,2011]

  - Characteristics: On-demand access, broad network access, resource pooling (multi-tenant model), rapid elasticity, measured service (metering, and transparency)

- Cloud-as-a-Utility
  - Like electricity and water
  - Illusion of infinite capacity
  - Massive economies of scale leading to low pay-as-you-go prices
  - No upfront commitment

Cloud Computing

Software-as-a-Service (SaaS)

- Email, CRM, audio/video processing, office suites, and numerous other applications

Platform-as-a-Service (PaaS)

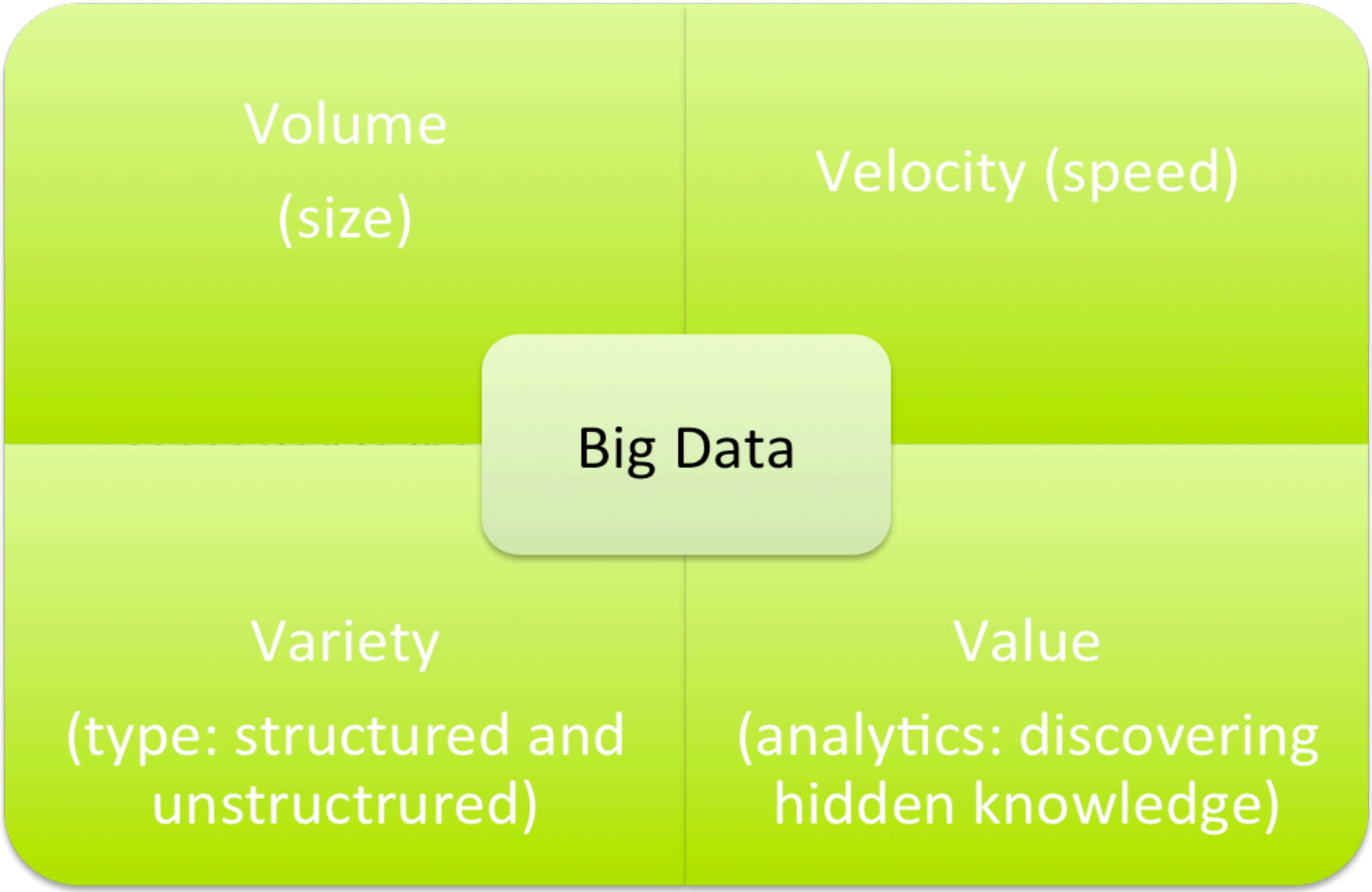
- Run Servers (e.g., Web, database and AAA)
- Programming languages (e.g., Java, PHP, Python and Ruby and Rails) and frameworks (e.g., CloudFront, Elastic MapReduce, and HDFS)
- Operating Systems (e.g., Ubuntu and Microsoft Windows Server 2008)

Infrastructure-as-a-Service (IaaS)

- Processing, Network and Storage

Monitoring-as-a-Service (MaaS), Network-as-a-Service (NaaS), BigData-as-a-Service (BDaaS), …,*aaS

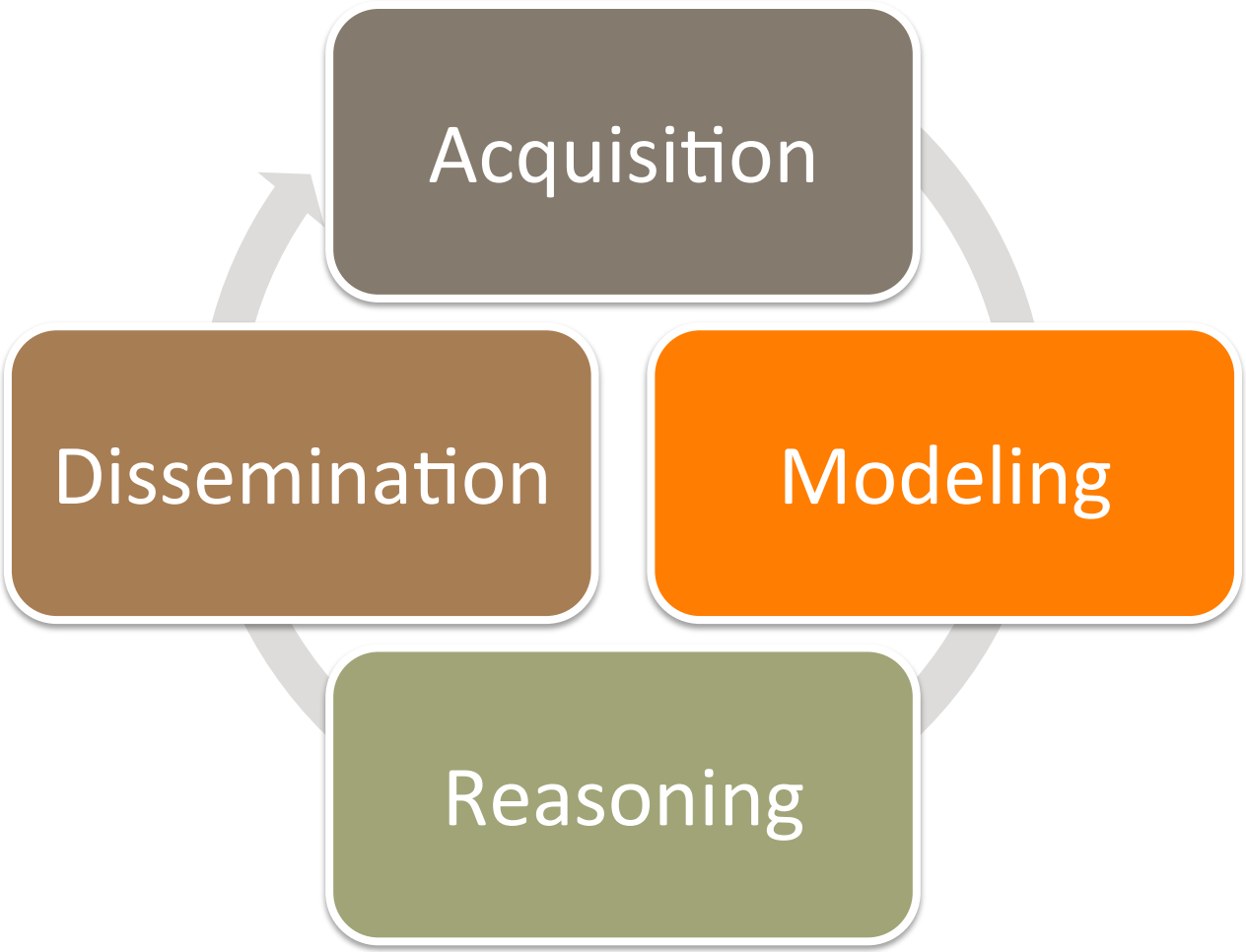
Internet-of-Things and Big Data



IoT and Context-Aware Computing

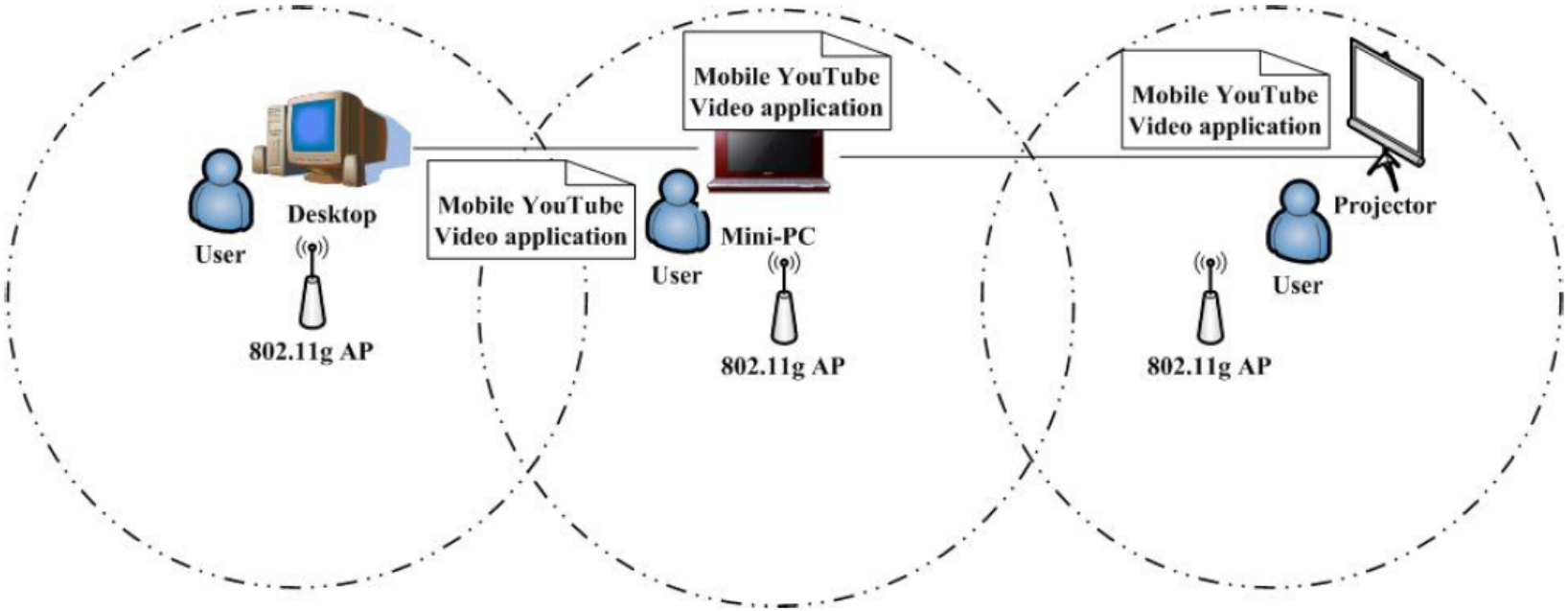
- *"The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it."* [Weiser, 1991]
  - Computer vanishes into the background
- *"Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves"* [Dey & Abowd, 1999]
- *"A system is context-aware if it uses context to provide relevant information and/or services to the user, where relevancy depends on the user's task."* [Dey & Abowd, 1999]
- Two decades of research led to numerous prototypes
  - Limited number of sensors (physical or virtual)
- IoT, Clouds and Big Data leading to resurgence of the research in context-aware systems

Context-Aware Computing



Mobile Cloud Computing

- Users are going mobile!
  - 6.1 billion smartphones worldwide [Techcrunch
  - Minimize battery consumption
  - Mobility



Mobile Cloud Computing Challenges

- End user mobile devices
  - Limited compute, storage and battery capacity
  - Network: intermittent connectivity, throughput, delay & jitter
  - Variability: both mobile networks and clouds
  - Mobility Management

Smart healthcare

Emergency management

Quality of Experience

- *"Quality of experience (QoE) is a metric that depends on the underlying QoS along with a person's preferences towards a particular object or service where his/her preferences are defined by his/her personal attributes related to expectations, experiences, behaviour, cognitive abilities, object's attributes and the environment surrounding that person".* [Mitra, Zaslavsky, Åhlund, 2015]
  - QoE = f(QoS, Context)
- Humans recognize:
  - Faces in 370 ms (best case) & 620 ms (worst case)!
  - Short speech phrases: 300ms to 450 ms
  - Detect human voice : 4ms
- VR application's perceptual stability requires 16ms

Maximize QoE!

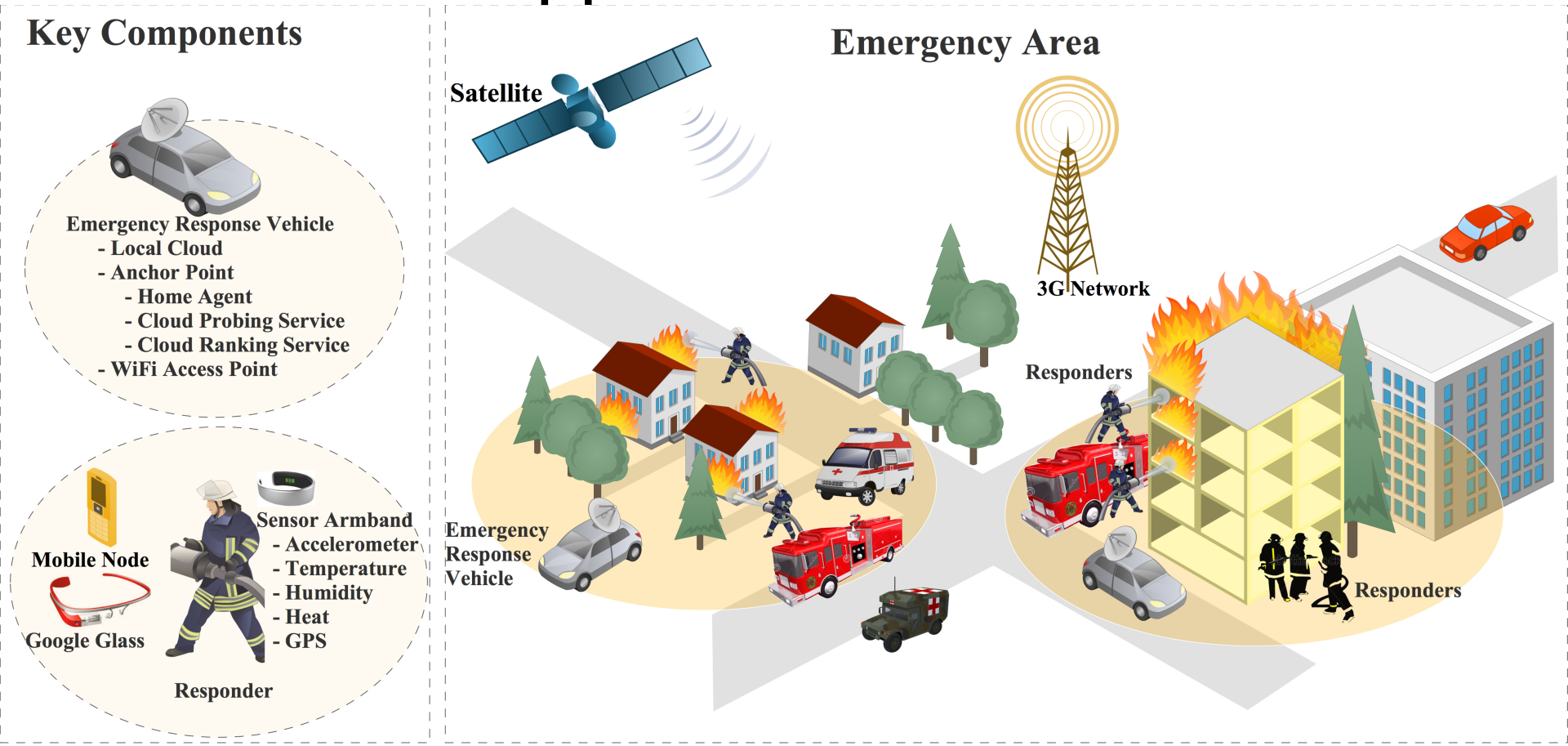
Karan Mitra, Arkady Zaslavsky and Christer Åhlund, *"Context-Aware QoE Modelling, Measurement and Prediction in Mobile Computing Systems"*, IEEE Transactions on Mobile Computing, IEEE, 2015

M2C2: A Mobility Management System for Mobile Cloud Computing

- Aim: *To select the best cloud and the best network while users roam in heterogeneous access networks*
- Proposed and developed M2C2
  - Multihoming: being able to connect to several access networks together (e.g., WiFi and LTE)
  - Cloud and network probing mechanisms
  - Cloud and network selection mechanisms

- Karan Mitra, Saguna Saguna, Christer Åhlund and Daniel Granlund, *"M2C2: A Mobility Management System for Mobile Cloud Computing"*, in Proceedings of the 2015 IEEE Wireless Communications and Networking Conference (IEEE WCNC 2015), 2015.

An Application Scenario

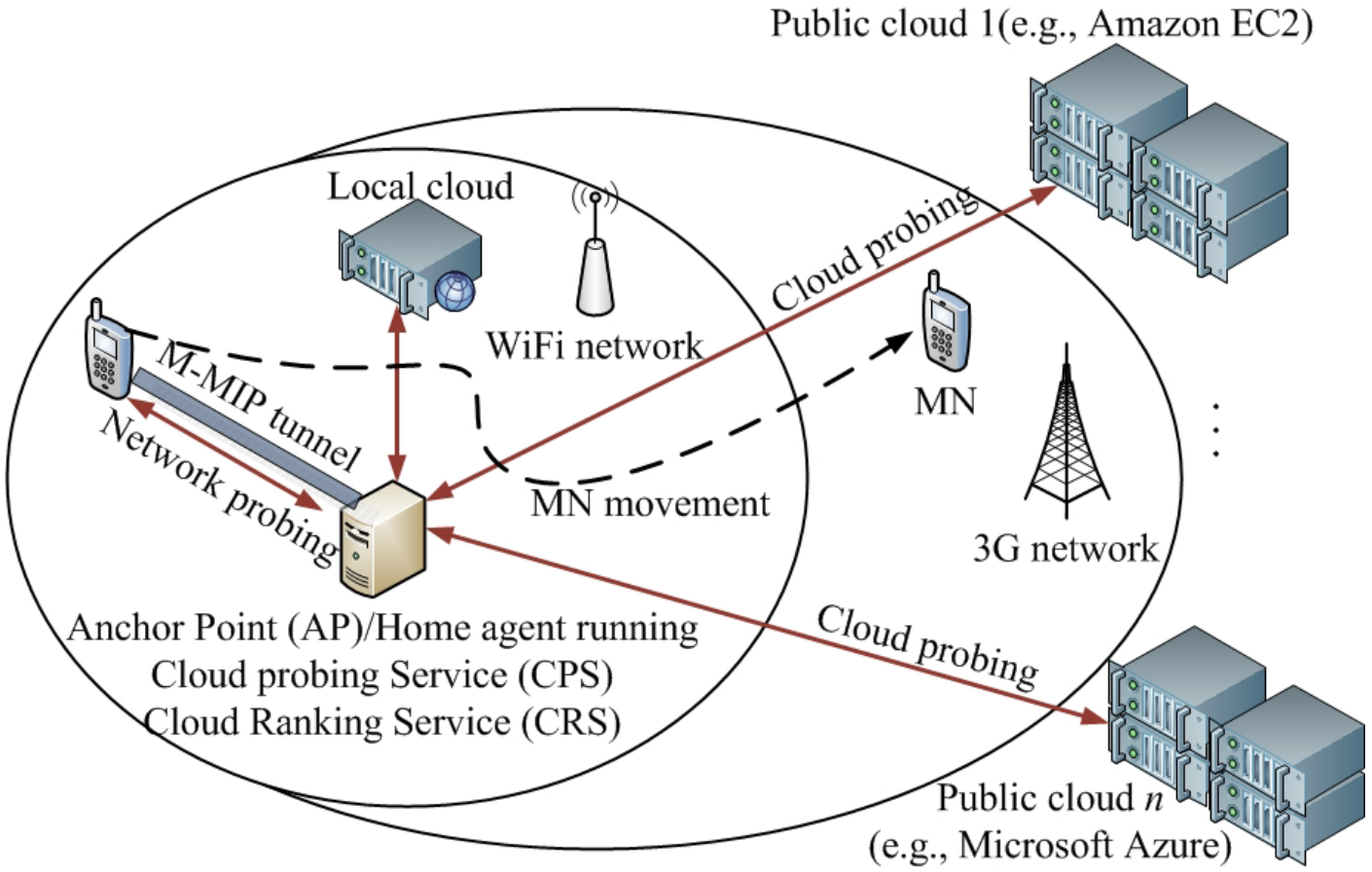


- Karan Mitra, Saguna Saguna and Christer Åhlund, *"A Mobile Cloud Computing System for Emergency Management,"* Cloud Computing, IEEE, vol. 1, no. 4, pp. 30–38, 2014.

M2C2: Mobility Management in Mobile Cloud Computing

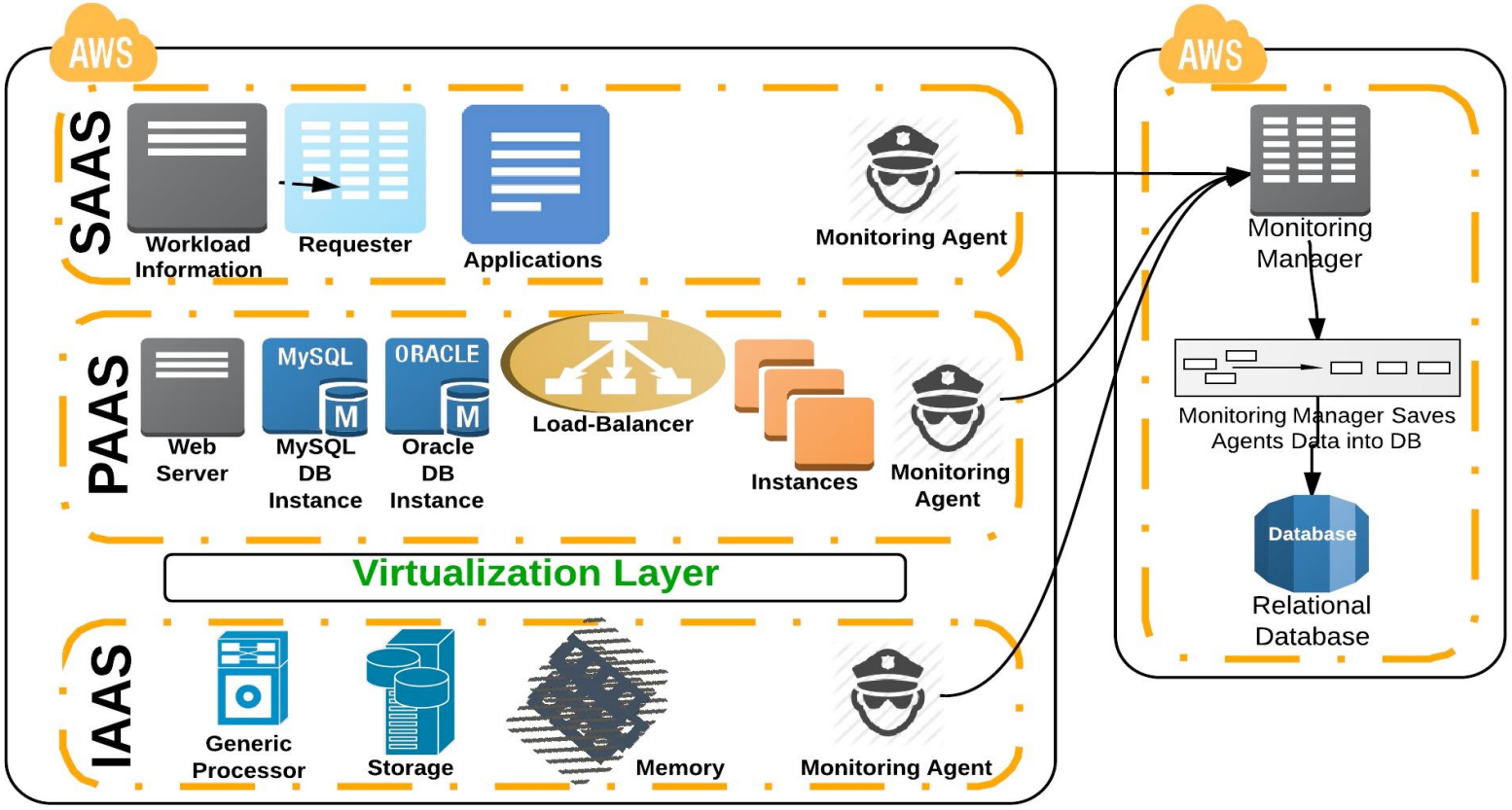
- Comprise several components:
  - Anchor Point
    - Cloud and network awareness
      - Cloud Probing Service
      - Cloud Ranking Service
    - Cloud probing and ranking: RESTful Web services
  - Home Agent
    - Network path probing via M-MIP tunnel
  - Mobile Node
    - Network selection using Relative Network Load metric

M2C2: Mobility Management in Mobile Cloud Computing



M2C2: system architecture

Cloud Monitoring as-a-Service



- Khalid Alhamazani, Rajiv Ranjan, Karan Mitra, Prem Prakash Jayaraman, Huang Zhiqian, Lizhe Wang and Fethi Rabhi, *"CLAMS: Cross-Layer Multi-Cloud Application Monitoring-as-a-Service Framework,"* in Proceedings of the 11th IEEE International Conference on Services Computing (IEEE SCC 2014). IEEE, 2014.

- Khalid Alhamazani, Rajiv Ranjan, Prem Jayaraman, Karan Mitra, Chang Liu, Fethi Rabhi, and Lizhe Wang,*"Cross-Layer Multi-Cloud Real-Time Application QoS Monitoring and Benchmarking As-a-Service Framework"*, IEEE Transactions on Cloud Computing, 2015.

M2C2: Mobility Management in Mobile Cloud Computing

- Cloud Service Selection via Cloud Ranking Service
  - Simple Additive Weighting (SAW)

$$\Re_k = w_{ma}(QoS_{nk}) + (1 - w_{ma})(Cost_{nk}),$$

- Network Selection
  - Relative Network Load metric

$$RNL = Z_n + cJ_n \quad (1)$$

$$Z_n = \frac{1}{h} RTT_n + \frac{h-1}{h} Z_{n-1} \quad (2)$$

$$RTT_n = R_n - S_n \quad (3)$$

$$D_n = RTT_n - RTT_{n-1} \quad (4)$$

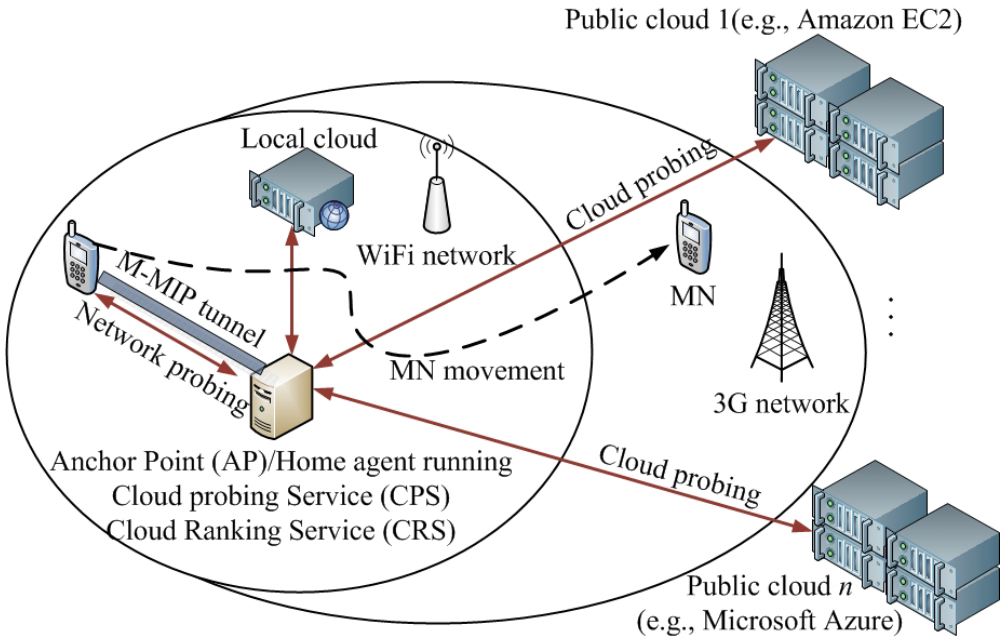
$$J_n = \frac{1}{h} |D_n| + \frac{h-1}{h} J_{n-1} \quad (5)$$

M2C2: Mobility Management in Mobile Cloud Computing

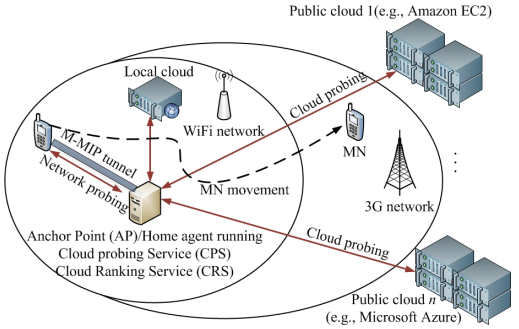
ALGORITHM 1: An algorithm for cloud and network selection for mobile cloud computing

Input: Avaliable networks, I; available clouds, K
Output: Perform activity recogntion on selected cloud, $k \in K$ using selected network, $i \in I$

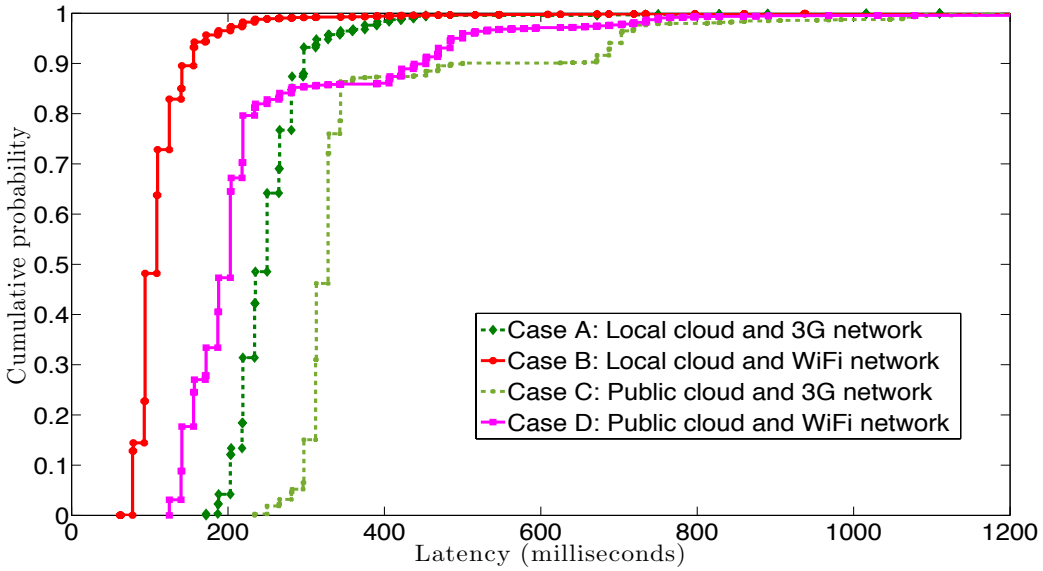
1 **Initialization:** $RNL_{i=\{1...n\}} \leftarrow 0$
2 Using RSSI, discover I;
3 **foreach** $(i \in I)$ **do**
4 **if** $RSSI_i \geq Threshold(RSSI_i)$ **then**
5 Connect to i
6 **end**
7 **end**
8 **foreach** $(i \in I)$ **do**
9 Establish tunnel with the Home Agent (HA)
10 **end**
11 **foreach** $(i \in I)$ **do**
12 Compute RNL_i
13 **end**
14 Connect to i with $\min(RNL_i)$
15 Retrieve the best cloud k from the Anchor Point (AP)
16 Perform activity recognition on k while connected to i



Results Analysis



- Prototype implementation and experimentation
  - Activity recognition application
- Experiment 1: local clouds vs. public clouds
  - Computation should be offloaded to local clouds using WiFi



Results Analysis

- Experiment 2: Cloud and Network Selection

```
INFO: The total number of Clouds listed in our account are: 2
INFO: The start time is :Tue Sep 09 15:11:51 CEST 2014
INFO: The end time is :Tue Sep 09 15:16:51 CEST 2014
INFO: Cloud with IP: [54.77.183.180]CPU utilization statistics: 0.67% (avg), 1.69% (max), 0% (min)
INFO: The start time is :Tue Sep 09 15:11:52 CEST 2014
INFO: The end time is :Tue Sep 09 15:16:52 CEST 2014
INFO: Cloud with IP: [54.77.218.113]CPU utilization statistics: 100% (avg), 100% (max), 0% (min)
```

1. Cloud probing by CPS

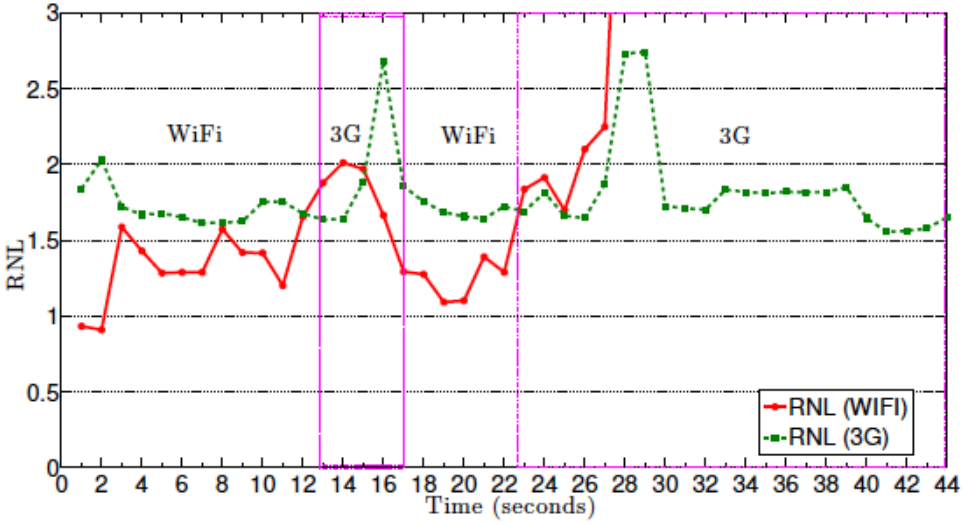
```
INFO: Retrieving the best URL...
INFO: The Best cloud URL: [54.77.183.180]
INFO: [54.77.183.180]
```

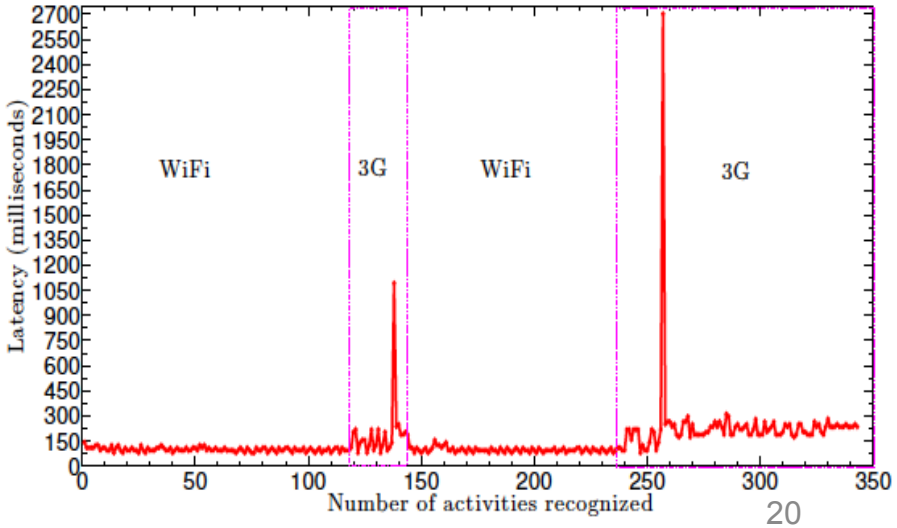
2. Cloud selection by CRS

```
Mobile Cloud Application is running...
INFO: Getting sensor readings:
Sep 09, 2014 4:23:11 PM com.ltu.mobilecloud.mobilecloudapplications.MobileCloudApplications
performActivityRecognition
INFO:
http://54.77.183.180:8080/ActivityRecogntionAppService/webresources/getactivity?type=xml&x=4
43&y=701&z=659
Sep 09, 2014 4:23:12 PM com.ltu.mobilecloud.mobilecloudapplications.MobileCloudApplications
performActivityRecognition
INFO: The recognized activity is: <InferredActivity>sitting</InferredActivity>
Sep 09, 2014 4:23:12 PM com.ltu.mobilecloud.mobilecloudapplications.MobileCloudApplications
performActivityRecognition
INFO: The end-to-end latency for performing AR on cloud:  54.77.183.180:8080 is: 176.0 milliseconds
```

Results Analysis

- Experiment 3: Impact of mobility
  - Mobile node roaming in WiFi and 3G networks
  - Seamless handoffs with no packet loss
  - Activity recognition continued successfully
    - Variation in latency based on access network





Conclusion and Future Work

- Proposed, developed and validated M2C2
  - A novel system for mobility management in mobile cloud computing
    - Multihoming
    - Cloud and network probing
    - Cloud and network selection

Future Work:

- Power consumption on mobile devices
- Extend the metrics for power-aware computation and storage placement
- Real-world case studies for smart regions

Cloud Performance Diagnosis and Prediction

- Cloud performance in terms of QoS is stochastic
- Affected by a large no. of parameters
  - Virtual machine types
  - Regions
  - Application workloads
  - Wide-area network delay
  - Throughput
  - Time-of-Day
  - Day-of-Week

Cloud Performance Diagnosis and Prediction

- Cloud benchmarking assists in holistic awareness of applications' performances w.r.t the underlying Cloud resources
  - Determining the baseline performance
    - Understanding application performance before its deployment
  - Comparing continual comparison of applications QoS performance

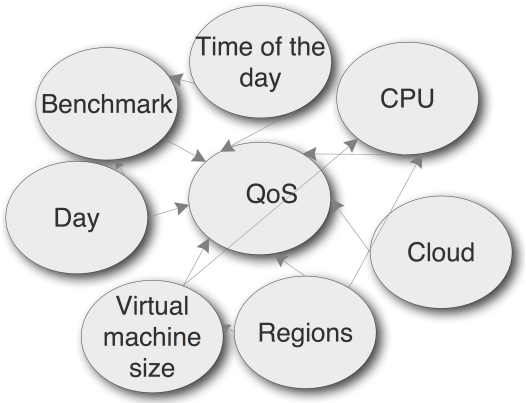
GOGRID amazon web services™ Microsoft Azure

Cloud Performance Diagnosis and Prediction

- Cloud performance benchmarking and diagnosis has attracted significant interest from industry and academia
- Still remains a highly challenging problem
  - *N no. cloud providers* x *M no. of parameters*
    - Each combination of parameters may lead to variation in performance
  - Lack of data sets
  - Limited experimental results under different use cases, constraints and experimental setups
  - Comparison with prior-work missing
- Hard to extract meaningful knowledge!
  - Root-cause diagnosis

ALPINE: A Bayesian System for Cloud Performance Diagnosis and Prediction

- We proposed and developed ALPINE for efficient cloud performance diagnosis and prediction
- Utilizes Bayesian Networks (BNs) to model uncertain and complex relationships between several factors
- Use the Expectation Maximization algorithm for learning BN model parameters
  - Handles missing information



ALPINE: A Bayesian System for Cloud Performance Diagnosis and Prediction

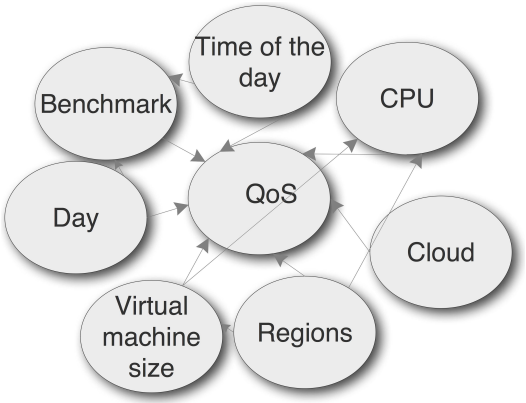
- Benchmark data is collected using tools such as CloudWorkBench, AWS CloudWatch and HTTPerf
- Data is pre-processed and stored in a database
- A BN is created by the expert or is learned using structural learning methods
- Modelled BN is used for Cloud performance diagnosis and prediction
- If BN(s) are deemed to be suitable, they are used by the stakeholders

ALPINE: A Bayesian System for Cloud Performance Diagnosis and Prediction

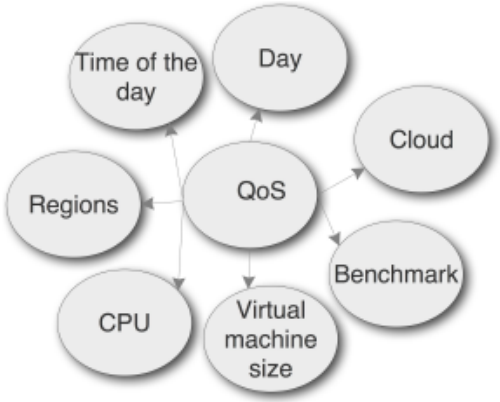
Definition 1. *A Bayesian network (BN) is a directed acyclic graph (DAG) where, random variables form the nodes of a network. The directed links between nodes form the causal relationships. The direction of a link from X to Y means that X is the parent of Y. Any entry in the Bayesian network can be calculated using the joint probability distribution (JPD) denoted as:*

$$P(x_1, ..., x_m) = \prod_{i=1}^{m} P(x_i | Parents(X_i)) \blacksquare \quad (1)$$

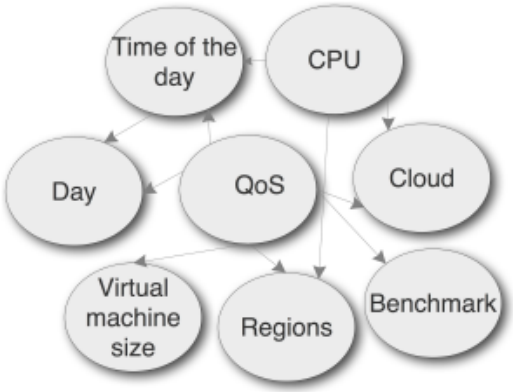
where, parents(X_i), denotes the specific values of Parents(X_i). Each entry in the joint distribution is represented by the product of the elements of the conditional probability tables (CPTs) in a BN



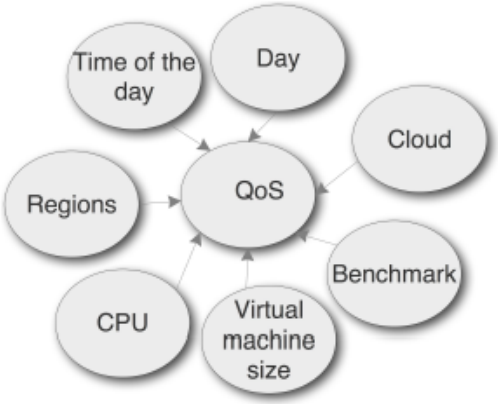
ALPINE: Multiple Types of Bayesian Networks



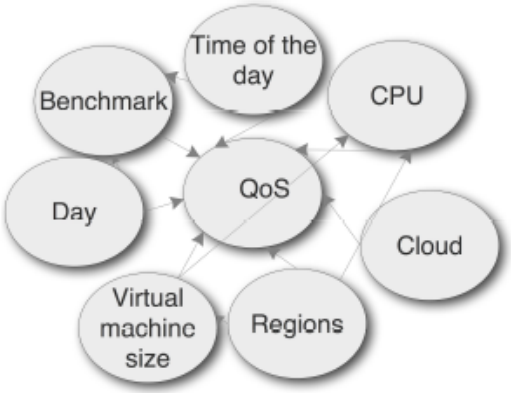
(a) Naive Bayes' Network (NBN).



(b) Tree-Augmented Naive Bayes' Network (TAN).



(c) Noisy-Or Network (NOR).



(d) Complex Bayesian Network (CBN).

ALPINE: Results Validation

- Cloud benchmark dataset from (Leitner & Cito, 2014)
  - 30,140 unique records for Amazon EC2 and Google Compute Engine
  - Collected over one month
  - Five benchmarks
    - **CPU:** total time taken to check 20,000 natural numbers for primeness (in seconds, lower the better)
    - **MEM:** measure read/write memory speed in MB/s (higher the better)
    - **Compile:** total cloning and compilation time in seconds (lower the better)
    - **I/O:** measure the disk read/write speed for a 5GB file for 3 seconds (Mb/sec, higher the better)
    - **OLTP:** measure the average no. of MySQL queries/sec for 10,000 rows in 3 minutes (queries/sec, higher the better)

ALPINE: Results Validation

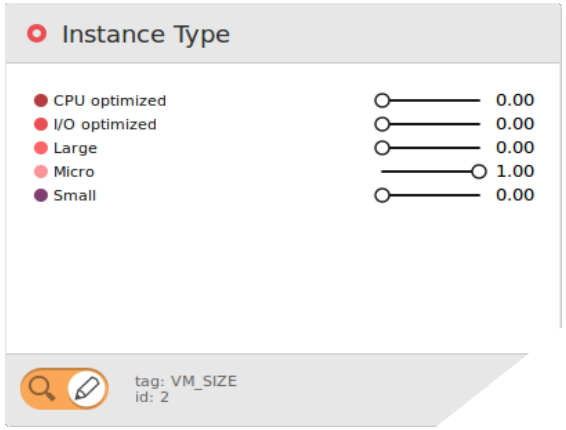
Table 1: Statistics related to all QoS values present in the dataset (Θ).

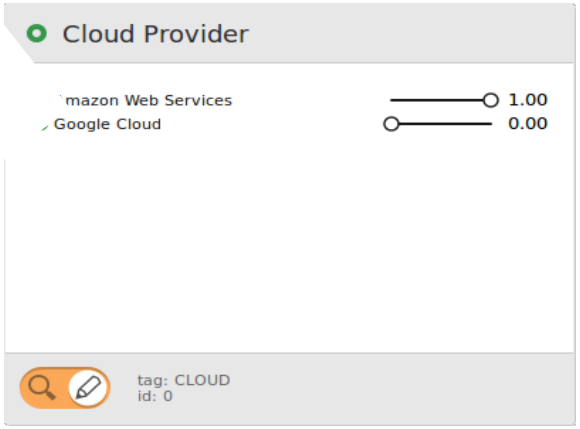
QoS Para.	Min.	Max.	Mean	Std. Dev.	Count
CPU	8.41	132.08	46.89	38.90	6894
Compile	0	2654.5	230.07	171.50	7319
Memory	611.65	6316.1	4114.5	1692.7	4581
I/O	1	1009.6	17.96	51.11	7377
OLTP	15.38	1130.25	310.05	281.74	3969
Combined	0	6316.1	737.19	1584.2	30140

ALPINE Results Validation: Diagnosis

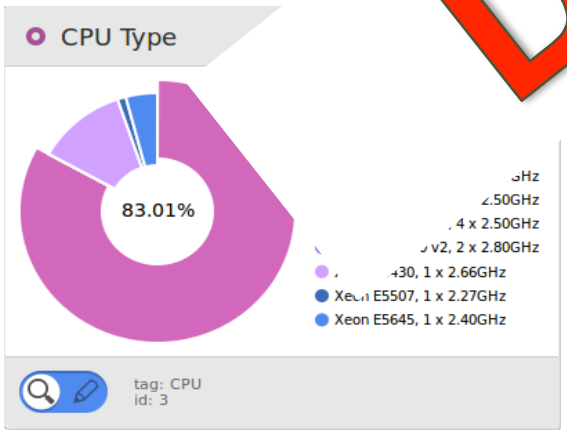
- CPU performance diagnosis
  - *"For a certain QoS value, what is the most likely instance type, CPU type, and the region"*
  - Using a single query we are trying to infer three factors:
    - Instance type
    - CPU type
    - Region
- What is the impact of time-of-the-day, and day-of-the-week

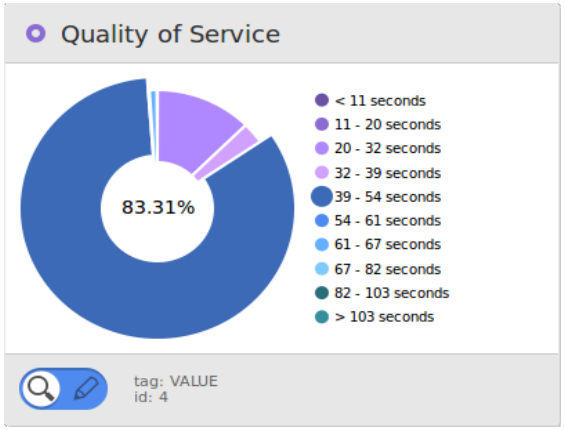
ALPINE Results Validation: Diagnosis





Demo!





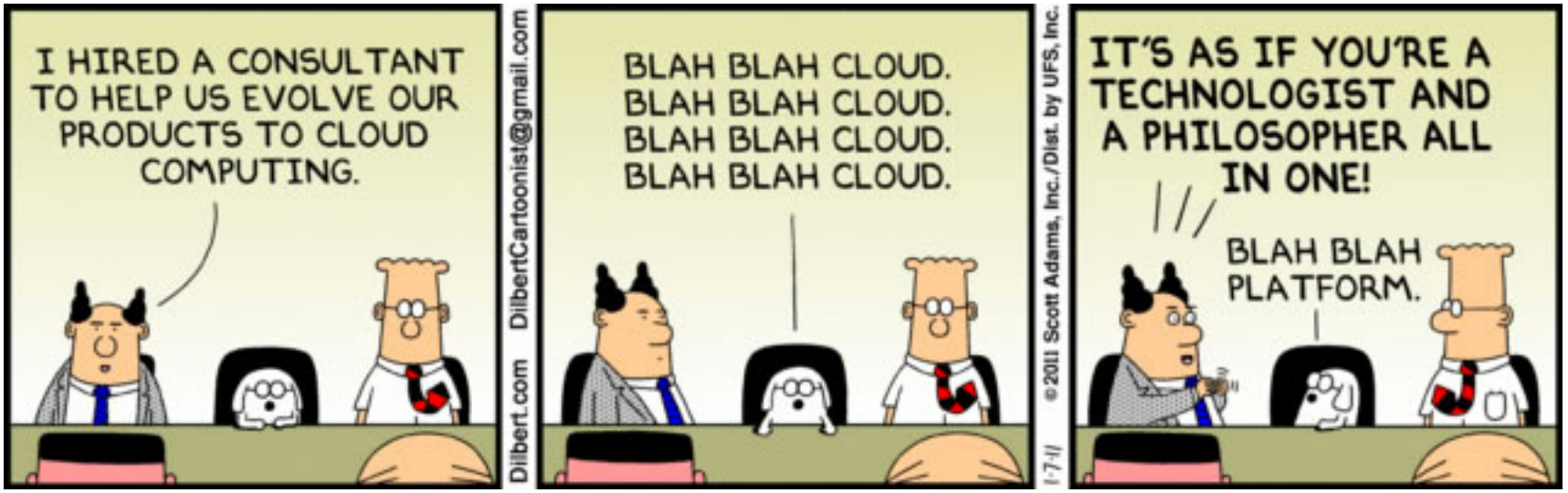
ALPINE Results Validation: Prediction

- BN can be modelled in several ways
  - We tested 4 different types of BNs as mentioned previously
  - 10-fold cross validation
  - 91.93% accuracy

BN Type	CPU	Compile	Memory	OLTP	I/O
NBN	97.12	95.93	89.54	97.40	76.21
TAN	99.24	96.08	92.20	97.40	76.17
NOR	99.24	95.65	91.42	97.40	76.08
CBN	99.24	96.09	92.70	97.40	76.04

Conclusions and Future Work

- Proposed, developed and validated ALPINE
  - Cloud diagnosis and prediction
- Data collection for multiple Cloud providers
- Finalize the prototype
- Extend our work on cloud orchestration to dynamically provision Cloud resources



Thank you for your attention!

Questions?