UNIVERSITY OF TARTU

INSTITUTE OF COMPUTER SCIENCE

Mobile and Cloud Computing Laboratory

From Exploiting Cloud in Science to Enabling Internet of Things

Satish Srirama

satish.srirama@ut.ee

Mobile Cloud Lab

iktp

European Union
Regional Development Fund

Investing in your future

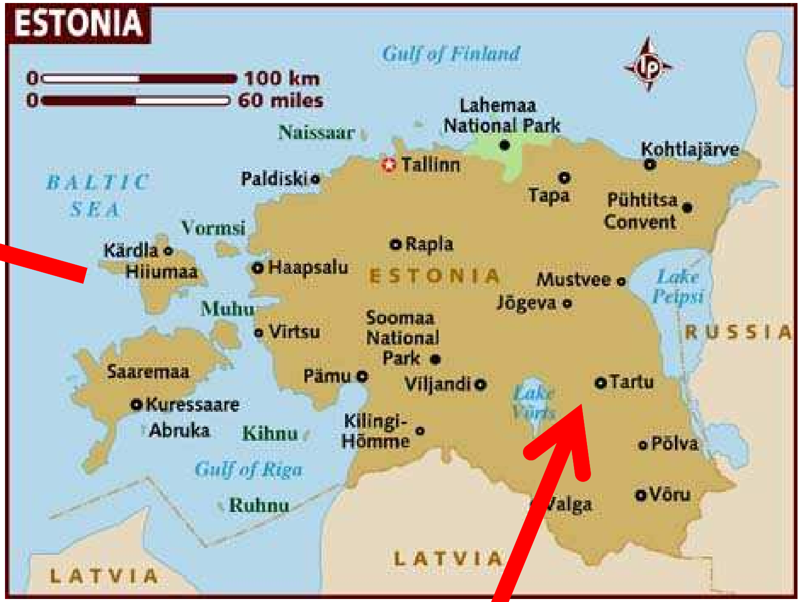
Who am I

- Head of Mobile & Cloud Lab, Institute of Computer Science, University of Tartu, Estonia

http://mc.cs.ut.ee

Mobile Cloud Lab

EUROPE
Greenland (DENMARK)
Jan Mayen (NORWAY)
Greenland Sea
Norwegian Sea
Barents Sea
ICELAND
Reykjavík
NORWAY
SWEDEN
FINLAND
Helsinki
Stockholm
Oslo
Tallinn
ESTONIA
LATVIA
Riga
LITHUANIA
Vilnius
BELARUS
Minsk
RUSSIA
Moscow
North Atlantic Ocean
North Sea
UNITED KINGDOM
London
IRELAND
Dublin
DENMARK
Copenhagen
POLAND
Warsaw
GERMANY
Berlin
NETH.
Amsterdam
BEL.
Brussels
FRANCE
Paris
CZECH REPUBLIC
Prague
SLOVAKIA
AUSTRIA
Vienna
HUNGARY
Budapest
UKRAINE
Kiev
MOLDOVA
ROMANIA
Bucharest
BULGARIA
Sofia
SERBIA AND MONTENEGRO
Belgrade
CROATIA
Zagreb
BOSNIA AND HERZEGOVINA
ALB.
GREECE
Athens
TURKEY
ITALY
Rome
SWITZ.
SPAIN
Madrid
PORTUGAL
Lisbon
Bay of Biscay
Mediterranean Sea
Black Sea
MOROCCO
ALGERIA
TUNISIA
Scale 1: 19,500,000
Lambert Conformal Conic Projection, standard parallels 40°N and 56°N



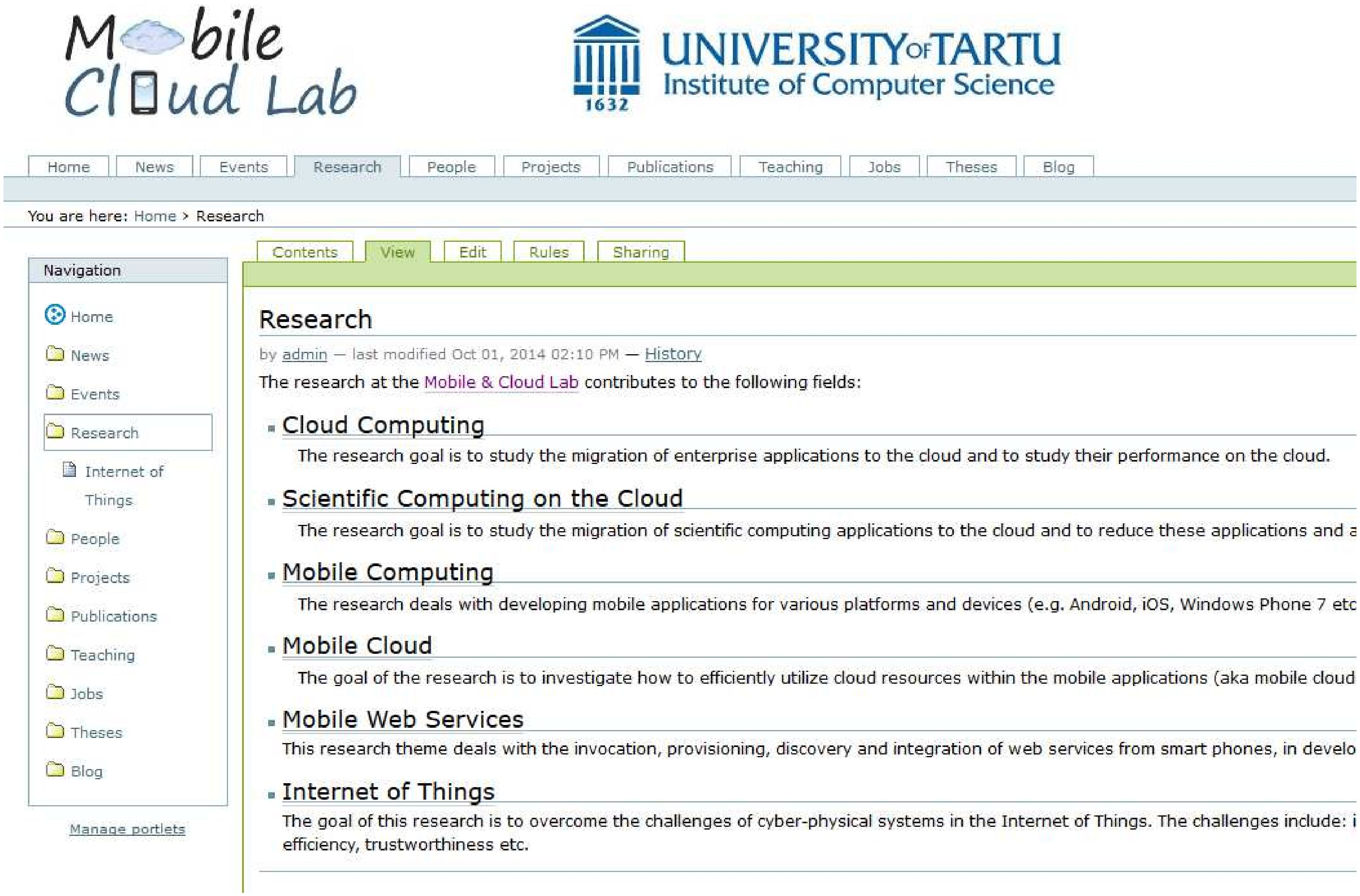
Estonia pop: 1,300,000

TARTU

Pop: 100,000

Academic excellence since 1632

Main Research Activities



Research

by admin — last modified Oct 01, 2014 02:10 PM — History

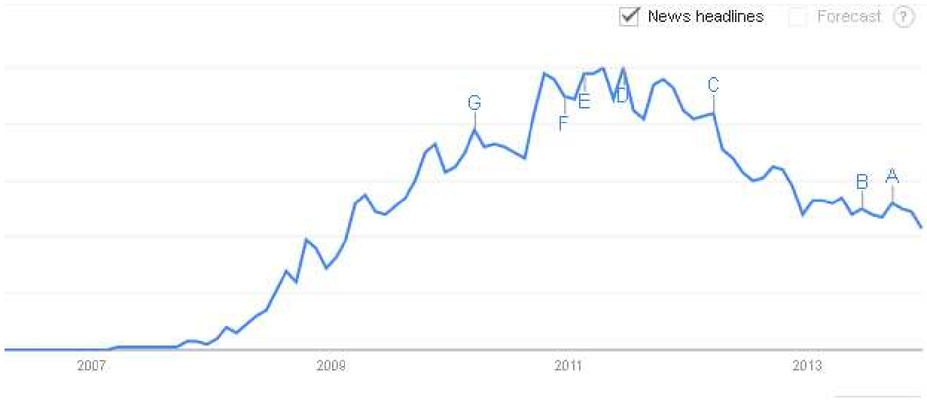
The research at the Mobile & Cloud Lab contributes to the following fields:

- **Cloud Computing**
  The research goal is to study the migration of enterprise applications to the cloud and to study their performance on the cloud.
- **Scientific Computing on the Cloud**
  The research goal is to study the migration of scientific computing applications to the cloud and to reduce these applications and a
- **Mobile Computing**
  The research deals with developing mobile applications for various platforms and devices (e.g. Android, iOS, Windows Phone 7 etc
- **Mobile Cloud**
  The goal of the research is to investigate how to efficiently utilize cloud resources within the mobile applications (aka mobile cloud
- **Mobile Web Services**
  This research theme deals with the invocation, provisioning, discovery and integration of web services from smart phones, in develo
- **Internet of Things**
  The goal of this research is to overcome the challenges of cyber-physical systems in the Internet of Things. The challenges include: i efficiency, trustworthiness etc.

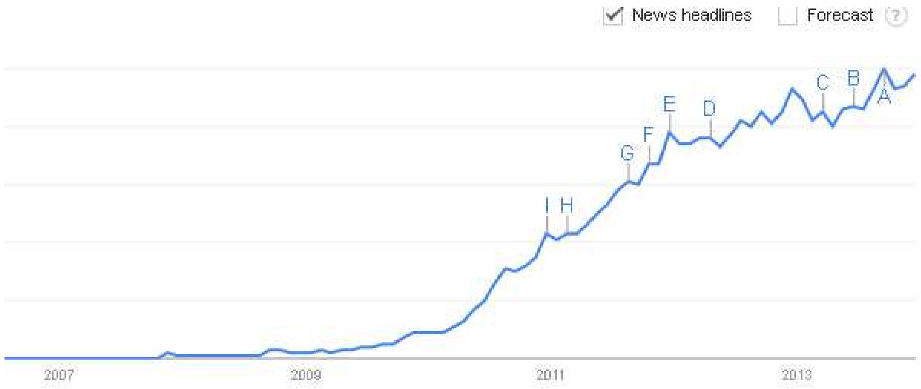
Outline

- Cloud computing
- Migrating enterprise applications to the cloud
- Scientific computing on the cloud
- Mobile Cloud
- Internet of Things

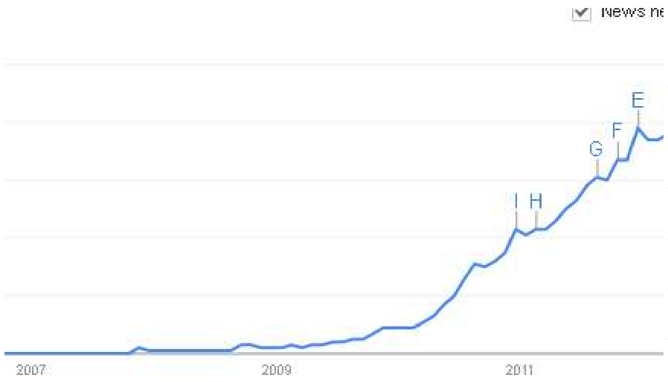
Some Recent Trends

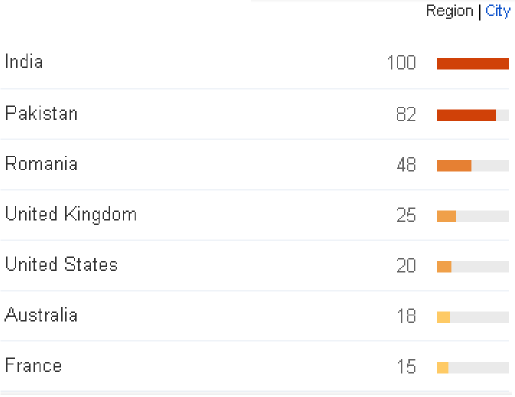


Cloud Computing



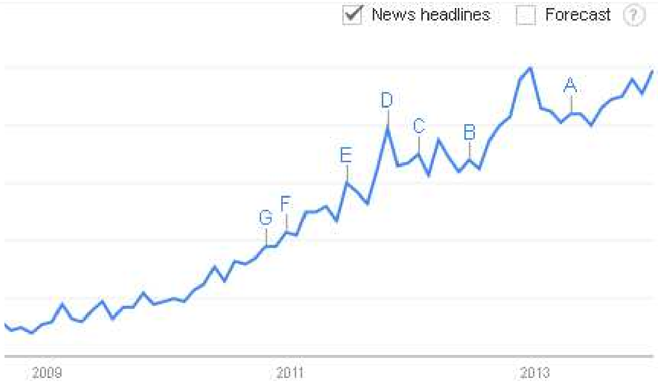
Smart Phone





Region | City		
India	100	🟥
Pakistan	82	🟥
Romania	48	🟧
United Kingdom	25	🟧
United States	20	🟧
Australia	18	🟨
France	15	🟨

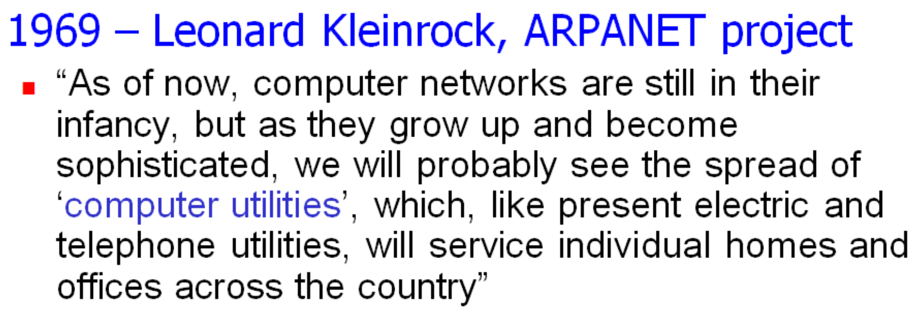
Android



Mobile Cloud

What is Cloud Computing?

- Computing as a utility
  - Utility services e.g. water, electricity, gas etc
  - Consumers pay based on their usage



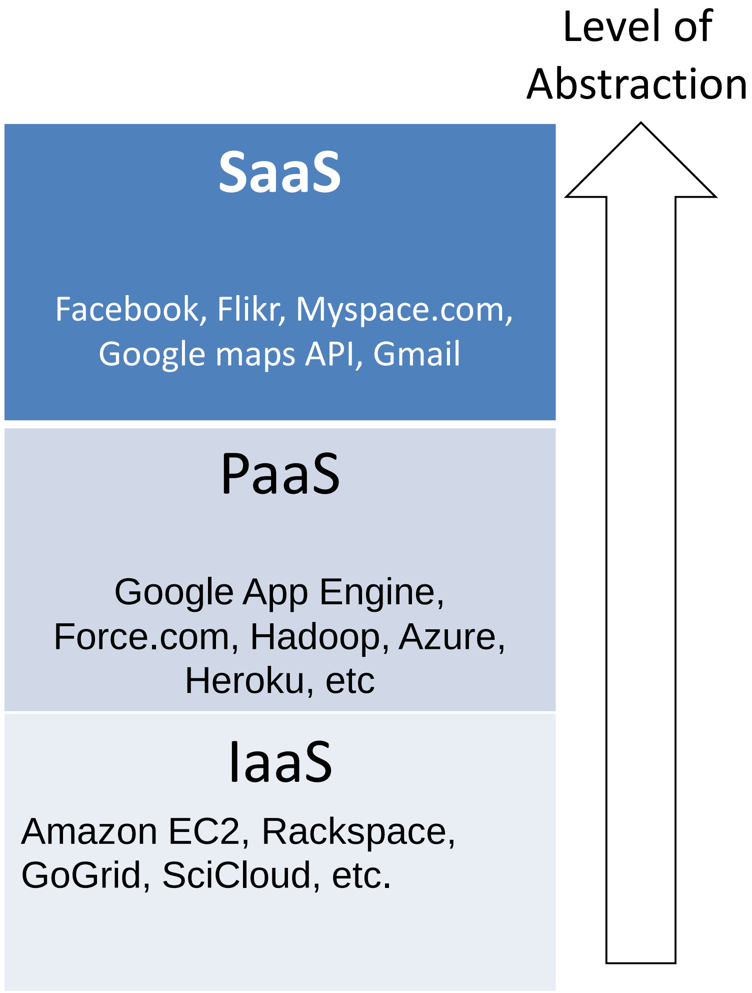
1969 – Leonard Kleinrock, ARPANET project

- "As of now, computer networks are still in their infancy, but as they grow up and become sophisticated, we will probably see the spread of 'computer utilities', which, like present electric and telephone utilities, will service individual homes and offices across the country"

- Cloud Computing characteristics
  - Illusion of infinite resources
  - No up-front cost
  - Fine-grained billing (e.g. hourly)

Cloud Computing - Services

- Software as a Service – SaaS
  - A way to access applications hosted on the web through your web browser
- Platform as a Service – PaaS
  - Provides a computing platform and a solution stack (e.g. LAMP) as a service
- Infrastructure as a Service – IaaS
  - Use of commodity computers, distributed across Internet, to perform parallel processing, distributed storage, indexing and mining of data
  - Virtualization



Level of Abstraction

SaaS

Facebook, Flikr, Myspace.com, Google maps API, Gmail

PaaS

Google App Engine, Force.com, Hadoop, Azure, Heroku, etc

IaaS

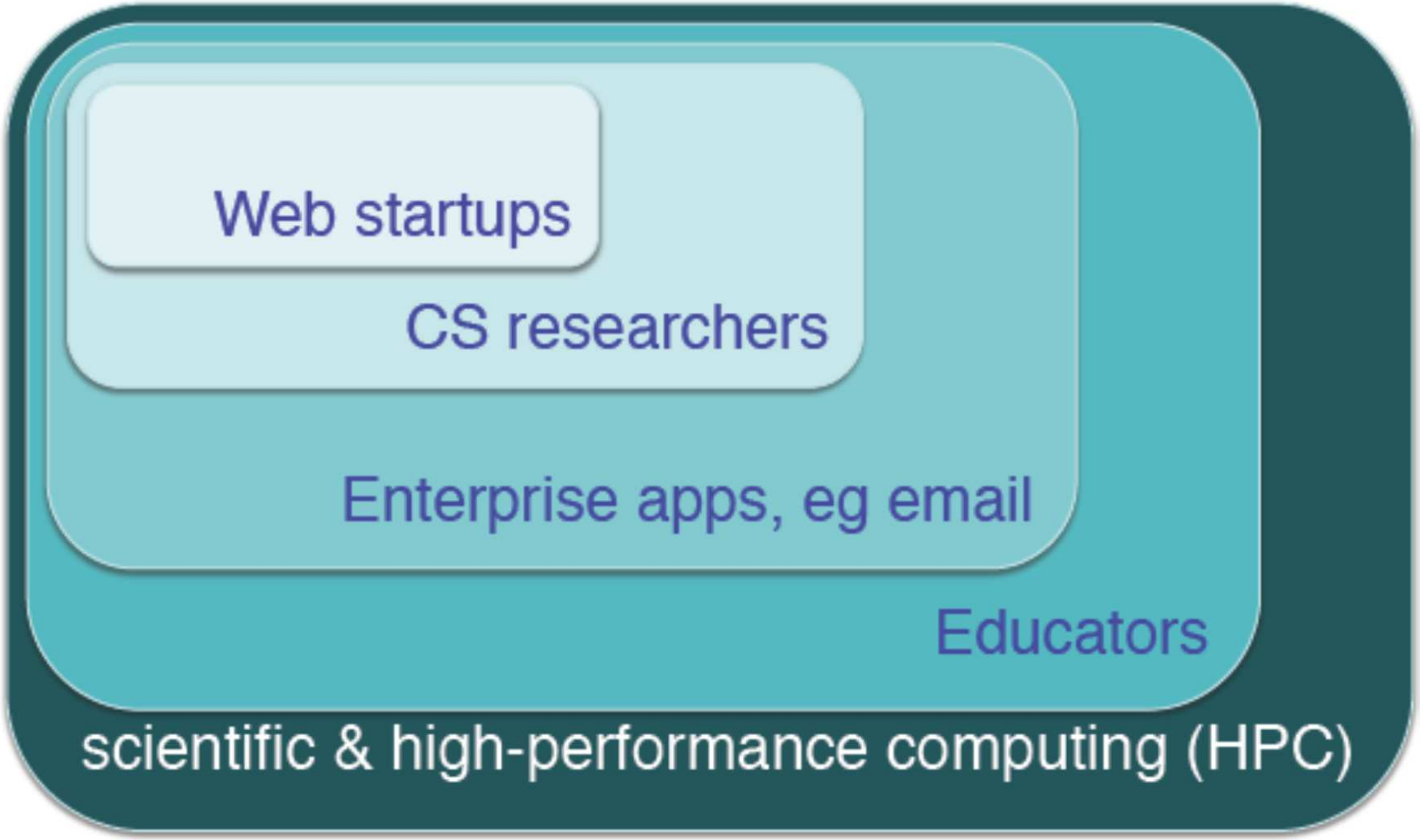
Amazon EC2, Rackspace, GoGrid, SciCloud, etc.

Cloud Computing - Themes

- Massively scalable
- On-demand & dynamic
- Only use what you need - Elastic
  - No upfront commitments, use on short term basis
- Accessible via Internet, location independent
- Transparent
  - Complexity concealed from users, virtualized, abstracted
- Service oriented
  - Easy to use SLAs

SLA – Service Level Agreement

Cloud Computing Progress


Web startups
CS researchers
Enterprise apps, eg email
Educators
scientific & high-performance computing (HPC)

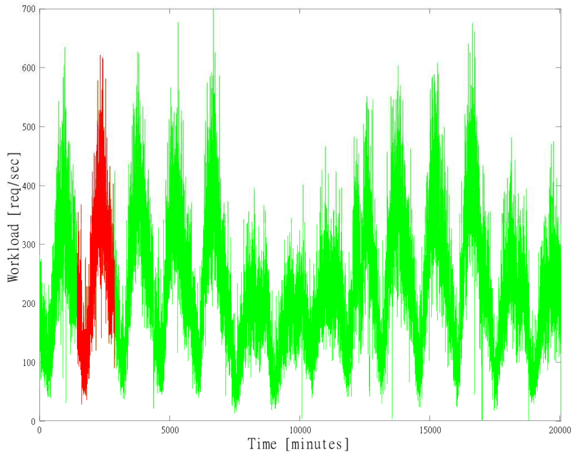
[Armando Fox, 2010]

Research Challenges

MIGRATING ENTERPRISE APPLICATIONS TO THE CLOUD

Enterprise applications on the cloud

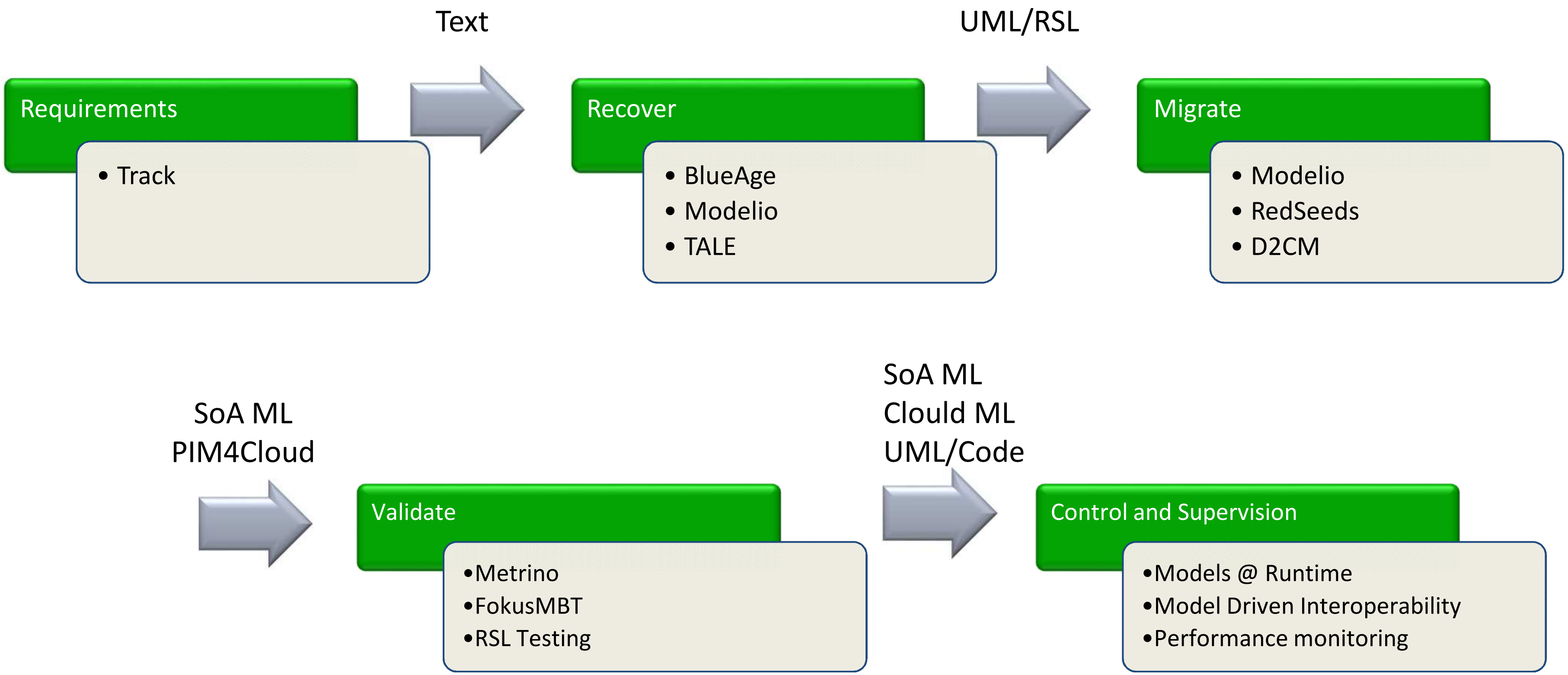
- Enterprise applications are mostly based on SOA and componentized models
- Fault tolerance, high availability & scalability
  - Essential prerequisites for any information system
- Cloud with its promise of virtually unlimited resources can offer the above prerequisites
  - Availability zones
  - Elasticity and horizontal scaling
  - Utility computing



REMICS

http://www.remics.eu/

- Reuse and migration of legacy applications to the cloud



Text

UML/RSL

Requirements

- Track

Recover

- BlueAge
- Modelio
- TALE

Migrate

- Modelio
- RedSeeds
- D2CM

SoA ML
PIM4Cloud

SoA ML
Clould ML
UML/Code

Validate

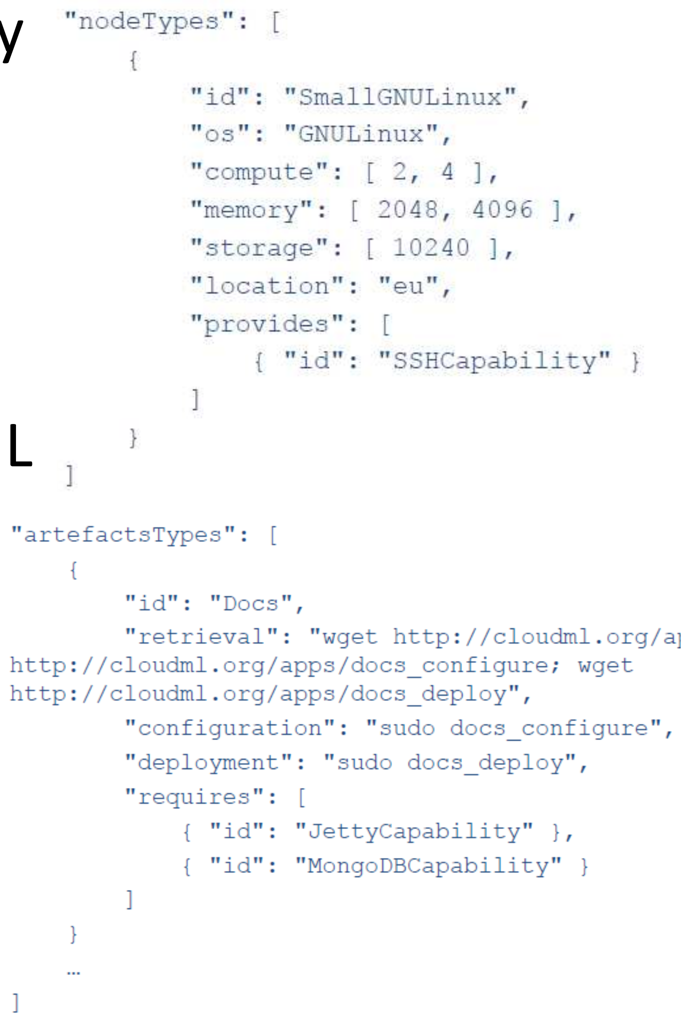
- Metrino
- FokusMBT
- RSL Testing

Control and Supervision

- Models @ Runtime
- Model Driven Interoperability
- Performance monitoring

CloudML

- Developed to tame cloud heterogeneity
- Domain-specific language (DSL) for modelling the provisioning and deployment at design-time
  - Nodes, artefacts and bindings can be defined
- Different means to manipulate CloudML models
  - Programmatically via Java API
  - Declaratively, via serialized model (JSON)
- Models@Runtime
  - Dynamic deployment of CloudML based models



```
"nodeTypes": [
    {
        "id": "SmallGNULinux",
        "os": "GNULinux",
        "compute": [ 2, 4 ],
        "memory": [ 2048, 4096 ],
        "storage": [ 10240 ],
        "location": "eu",
        "provides": [
            { "id": "SSHCapability" }
        ]
    }
]
"artefactsTypes": [
    {
        "id": "Docs",
        "retrieval": "wget http://cloudml.org/a
http://cloudml.org/apps/docs_configure; wget
http://cloudml.org/apps/docs_deploy",
        "configuration": "sudo docs_configure",
        "deployment": "sudo docs_deploy",
        "requires": [
            { "id": "JettyCapability" },
            { "id": "MongoDBCapability" }
        ]
    }
    …
]
```

[Ferry et al, Cloud 2013]

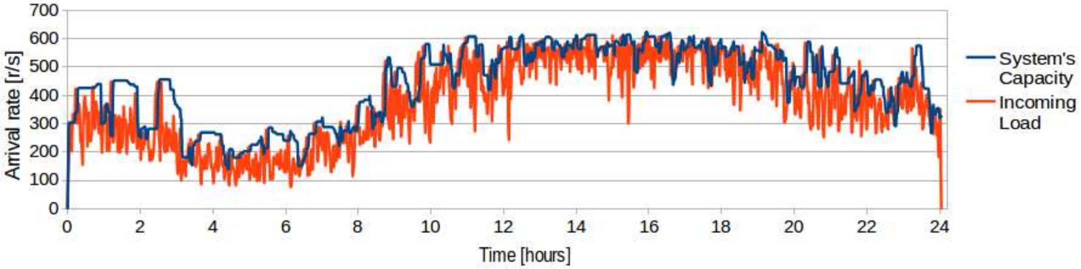
Auto-Scaling enterprise applications on the cloud

- Auto-Scaling
  - Scaling policy **->** When to Scale
  - Resource provisioning policy **->** How to scale
- Threshold-based scaling policies are very popular due to their simplicity
  - Observe metrics such as CPU usage, disk I/O, network traffic etc.
  - E.g. Amazon AutoScale, RightScale etc.
  - However, configuring them optimally is not easy

Optimal Resource Provisioning for Auto-Scaling Enterprise Applications

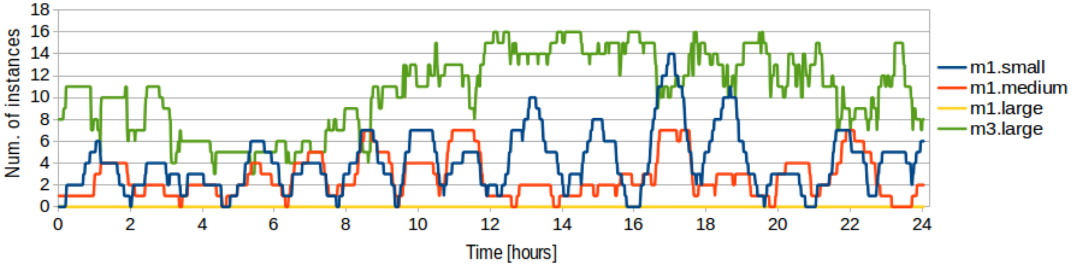
- Cloud providers offer various instance types with different processing power and price
  - Can it be exploited in deciding the resource provisioning policy?
  - Makes the policy to be aware of current deployment configuration
- Another challenge: Cloud providers charge the resource usage for fixed time periods
  - E.g. Hourly prices of Amazon cloud
- Developed an LP based optimization model which considers both the issues [Srirama and Ostovar, CloudCom 2014]

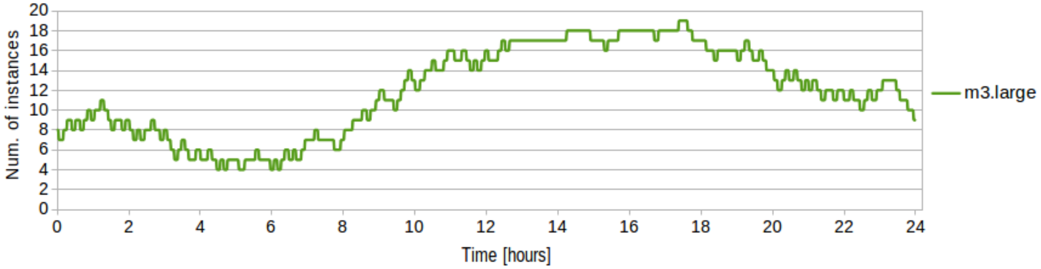
Scaling enterprise application with the optimization model



Incoming load and scaling curves of Optimization model

Instance type usage curves of Optimization model





Scaling with Amazon AutoScale

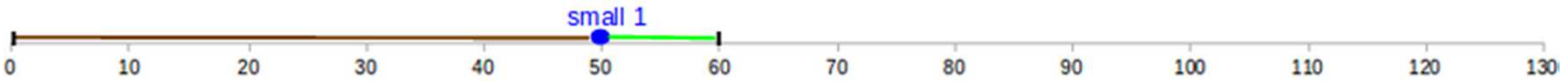
[Srirama and Ostovar, CloudCom 2014]

Optimization Model

Intuition behind instance lifetime consideration

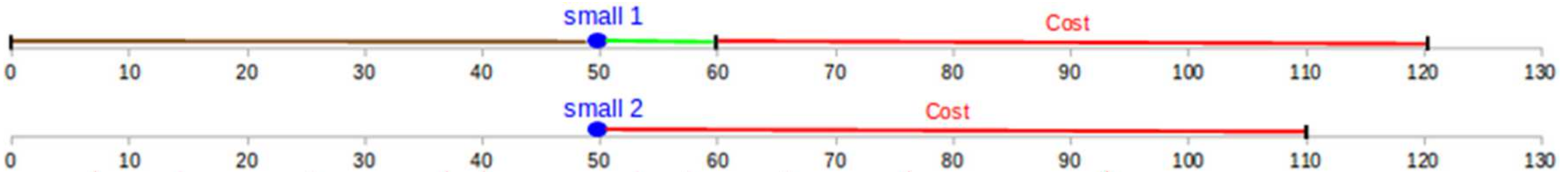
- Consider 2 instance types
  - Small instance(PW = 6r/s, Price = $0.25/h),
  - Medium instance(PW = 12r/s, Price = $0.4/h)

Load is 6r/s



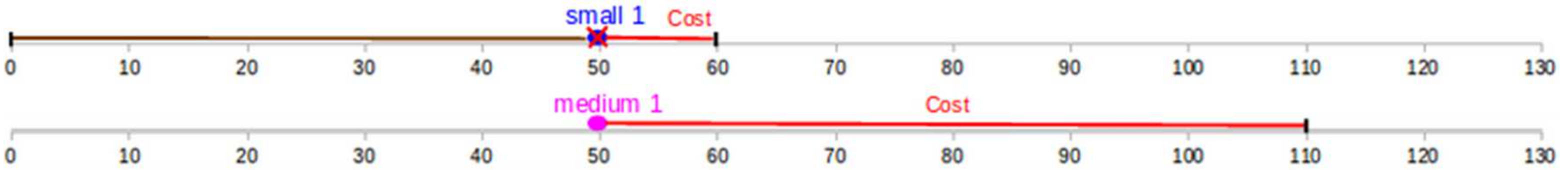
Load increases to 12r/s => ?

Solution 1



Cost = (cost of two small instances) - (10-min profit of a small instance) = 0.5 - 0.04 = $0.46

Solution 2



Cost = (cost of a medium instance) + (10-min cost of a small instance) = 0.4 + 0.04 = $0.44

- Saved cost with solution 2 : 0.46 – 0.44 = 0.02$
- So can we find this automatically?

Optimization Model

Some key definitions

- Region:
  - A task with its own independent characteristics
  - Each region can have its own capacity of instances
- Instance Type:
  - Each region can include multiple instance types
  - It is associated with processing power, price per period, capacity constraint, and configuration time
- Time bags:
  - Time interval where an instance is at a particular time
- Killing Cost:
  - Money lost when an instance is killed before it fills its paid period
- Retaining Cost:
  - The cost of the lived duration of the paid period

[Srirama and Ostovar, CloudCom 2014]

Optimization Model

- ***Cost Function:***

Cost of new instances

Configuration cost of new instances

$$Min\ \left(\boxed{\sum_{i=1}^{n}\sum_{j=1}^{m} N_{r_i,t_j} * C_{r_i,t_j}} + \boxed{N_{r_i,t_j} * (CT_{r_i,t_j} * CTB_{r_i,t_j})} + \right.$$

$$\boxed{\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{q} S_{r_i,t_j,tb_k} * KC_{r_i,t_j,tb_k}} +$$

→ Cost of killed instances

$$\left.\boxed{\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{q} (X_{r_i,t_j,tb_k} - S_{r_i,t_j,tb_k}) * RC_{r_i,t_j,tb_k}}\right)$$

→ Cost of retained instances

- ***Constraints:***

$$\sum_{j=1}^{m}(N_{r,t_j} + (\sum_{k=1}^{q} X_{r,t_j,tb_k} - S_{r,t_j,tb_k})) * P_{r,t_j} \geq W_r$$

→ Workload constraint

$$\sum_{j=1}^{m}(N_{r,t_j} + (\sum_{k=1}^{q} X_{r,t_j,tb_k} - S_{r,t_j,tb_k})) \leq CC_r$$

→ Cloud capacity constraint

$$N_{t_r} + (\sum_{k=1}^{q} X_{t_r,tb_k} - S_{t_r,tb_k}) \leq CCT_{t_r}$$

→ Instance type capacity constraint

$$S_{tb_{r,t}} \leq X_{tb_{r,t}}$$

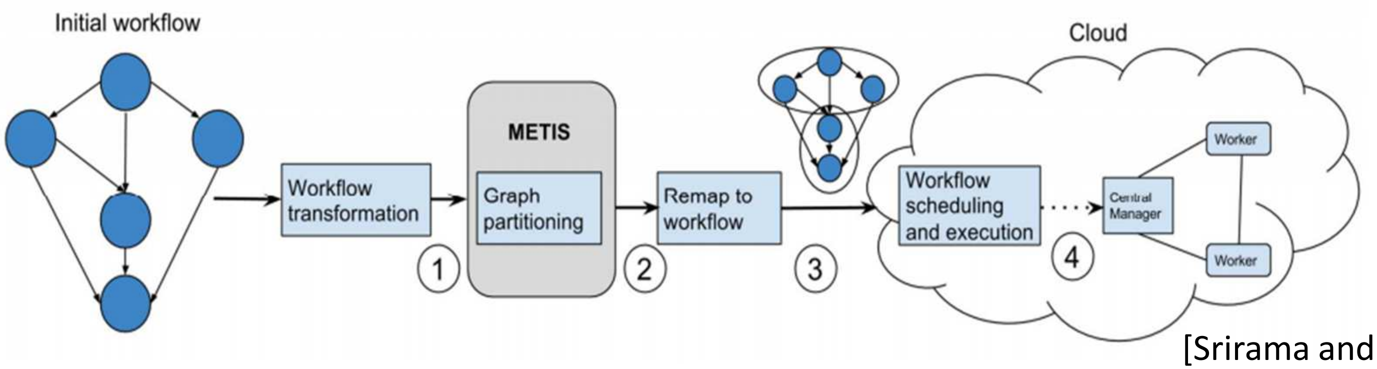
→ Shutdown constraint

$$N_{r,t} \geq 0$$
$$S_{r,t} \geq 0$$

[Srirama and Ostovar, CloudCom 2014]

Current Interests

- Remodeling enterprise applications for the cloud migration
  - Cloud has huge troubles with communication/transmission latencies [Srirama et al, SPJ 2011]
  - Intuition: Reduce inter-node communication and to increase the intra-node communication
- Auto-scale them based on optimization model and CloudML



[Srirama and Viil, HPCC 2014]

Research Challenges

SCIENTIFIC COMPUTING ON THE CLOUD

Scientific Computing on the Cloud

- Public clouds provide very convenient access to computing resources
  - On-demand and in real-time
  - As long as you can afford them
- High performance computing (HPC) on cloud
  - Virtualization and communication latencies are major hindrances [Srirama et al, SPJ 2011; Batrashev et al, HPCS 2011]
    - Things have improved significantly over the years
  - Research at scale
    - Cost-to-value of experiments

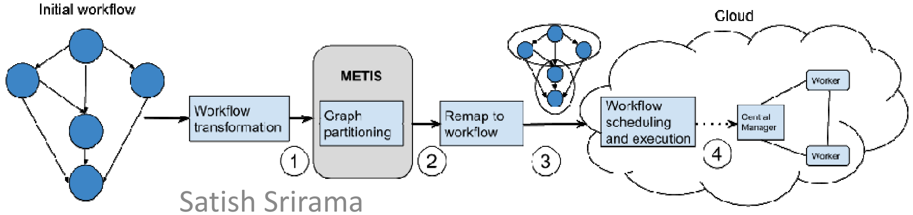
Desktop to Cloud Migration (D2CM) tool

- Seamless migration of desktop virtual machine images to the Cloud – For domain scientists
  - Transform local VM images into cloud compatible VM images
    - Extract the file system, package kernels, Install additional software etc.
  - Move it to the target cloud
- Create deployment template to describe the configuration
  - Define roles
    - Instance type, Number of instances
  - Define actions for each role
    - Uploads, Initialization commands, Run commands, Deployment ending conditions, Downloads
- Describe experiment once, run anywhere
  - Reuse the deployment template to generate new experiments and change the parameters as needed

[Srirama et al, HPCS 2013]

Migrating Scientific Workflows to the Cloud

- Workflow can be represented as weighted directed acyclic graph (DAG)
- Partitioning the workflow into groups with graph partitioning techniques
  - Such that the sum of the weights of the edges connecting to vertices in different groups is minimized
  - Utilized Metis' multilevel k-way partitioning
- Scheduling the workflows with tools like Pegasus
  - Considered peer-to-peer file manager (Mule) for Pegasus



Economics of Cloud Providers

- Cloud Computing providers bring a shift from high reliability/availability servers to commodity servers
  - At least one failure per day in large datacenter
- Why?
  - Significant economic incentives
    - much lower per-server cost
- Caveat: User software has to adapt to failures
  - Very hard problem!
- Solution: Replicate data and computation
  - MapReduce & Distributed File System

Typical Large-Data problem

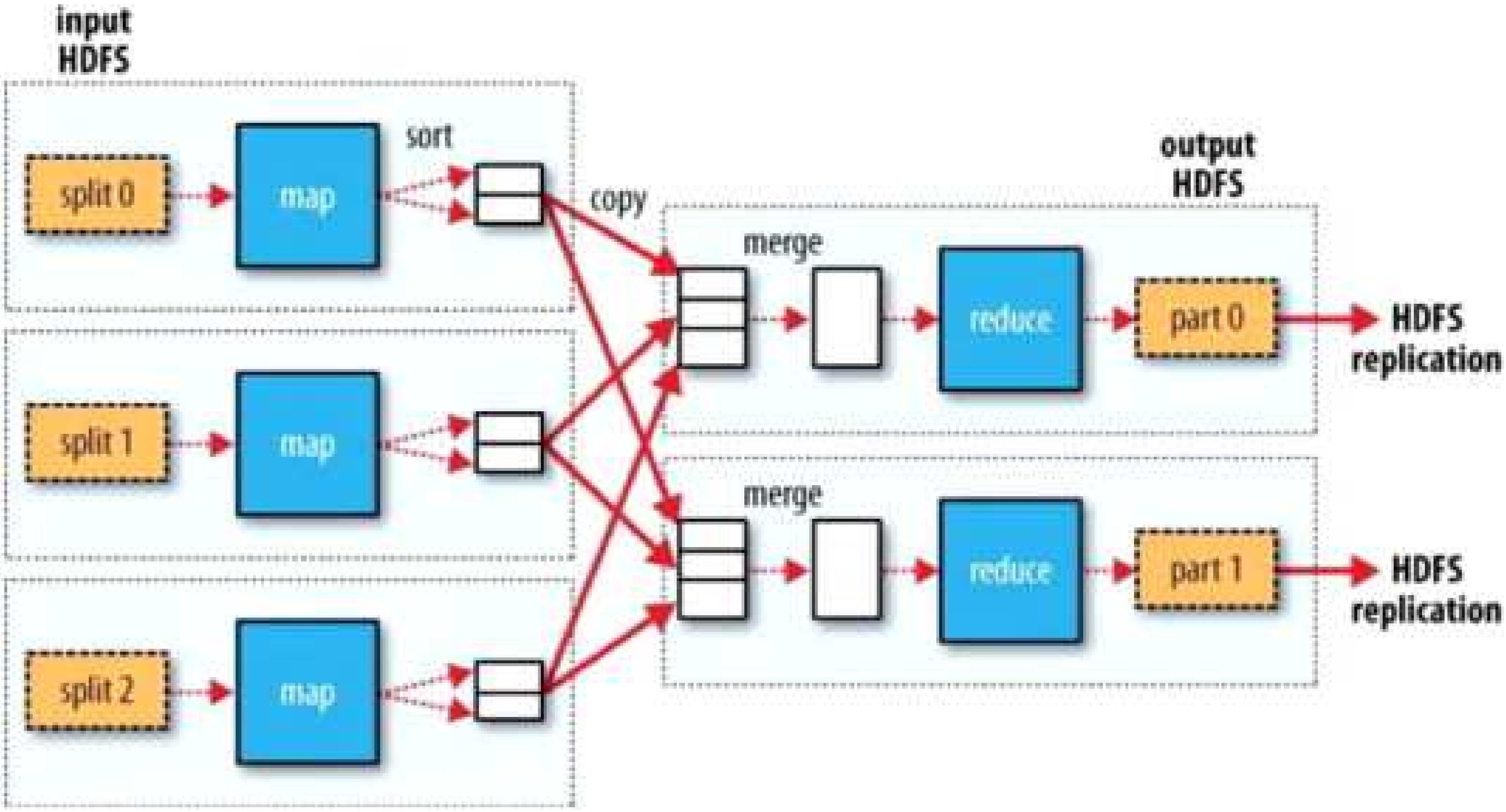
Map

- Iterate over a large number of records
- Extract something of interest from each
- Shuffle and sort intermediate results
- Aggregate intermediate results
- Generate final output

Reduce

Some material adapted from slides by Jimmy Lin, Christophe Bisciglia, Aaron Kimball, & Sierra Michels-Slettvet, Google Distributed Computing Seminar, 2007 (licensed under Creation Commons Attribution 3.0 License)

MapReduce model



https://tedwon.atlassian.net/wiki/display/SE/Apache+Hadoop

Apache Hadoop MapReduce

- Most prominent Open Source solution
- The user only has to write Map and Reduce functions
- Framework handles everything else for the user
  - Scheduling, data distribution, synchronization, errors and faults
- Parallelism is achieved by executing Map and Reduce tasks concurrently

Adapting Computing Problems to Cloud

- Reducing the algorithms to cloud computing frameworks like MapReduce [Srirama et al, FGCS 2012]
- Designed a classification on how the algorithms can be adapted to MR
  - Algorithm → single MapReduce job
    - Monte Carlo, RSA breaking
  - Algorithm → *n* MapReduce jobs
    - CLARA (Clustering), Matrix Multiplication
  - Each iteration in algorithm → single MapReduce job
    - PAM (Clustering)
  - Each iteration in algorithm → *n* MapReduce jobs
    - Conjugate Gradient
- Applicable especially for Hadoop MapReduce

Issues with Hadoop MapReduce

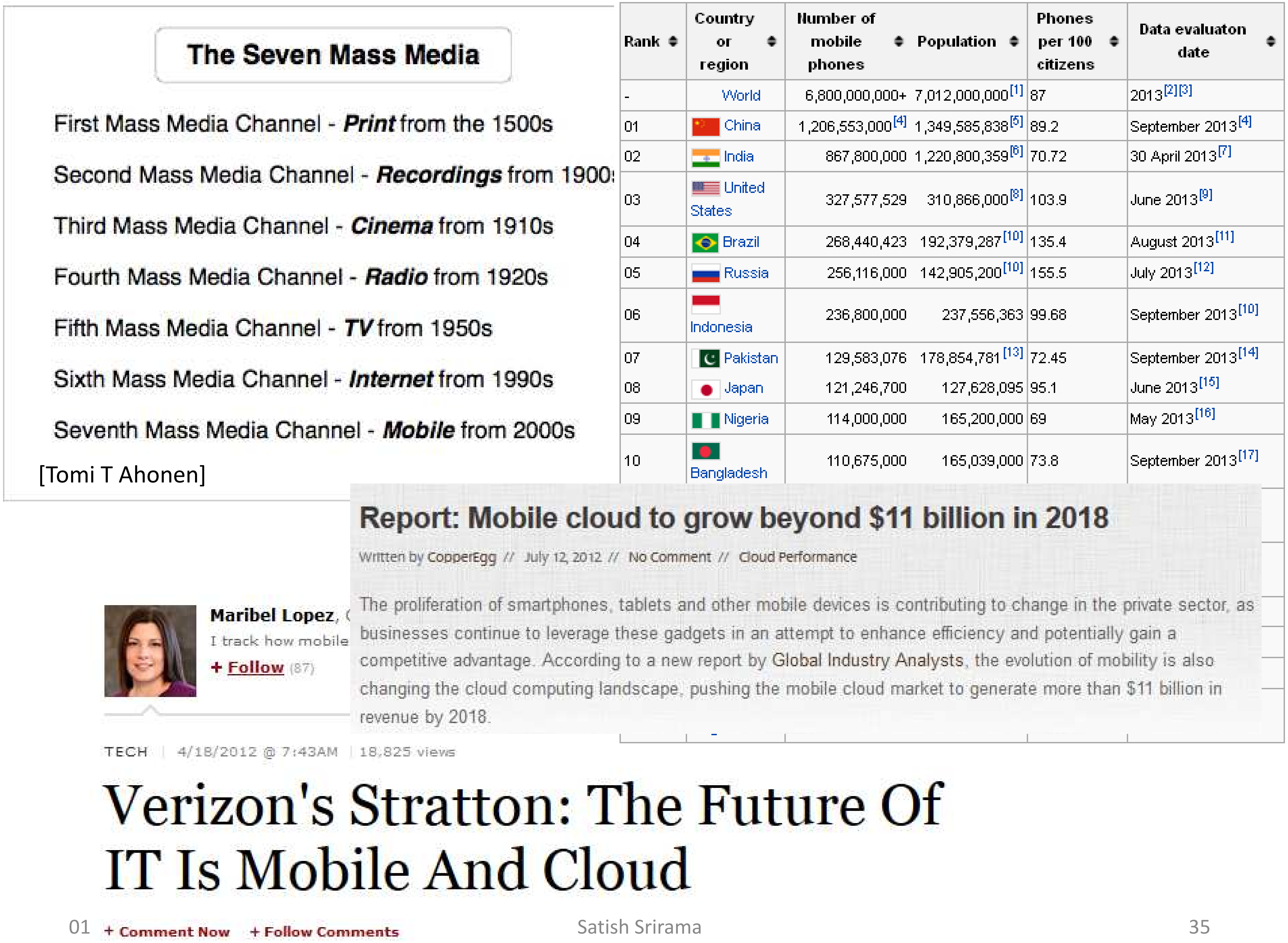
- It is designed and suitable for:
  - Data processing tasks
  - Embarrassingly parallel tasks
- Has serious issues with iterative algorithms
  - Long „*start up*“ and „*clean up*“ times **~17** seconds
  - No way to keep important data in memory between MapReduce job executions
  - At each iteration, all data is read again from HDFS and written back there at the end
  - Results in a significant overhead in every iteration

Alternative Approaches

- Restructuring algorithms into non-iterative versions
  - CLARA instead of PAM [Jakovits & Srirama, Nordicloud 2013]
- Alternative MapReduce implementations that are designed to handle iterative algorithms [Jakovits and Srirama, HPCS 2014]
  - E.g. Twister, HaLoop, Spark
- Alternative distributed computing models
  - Bulk Synchronous Parallel model [Valiant, 1990] [Jakovits et al, HPCS 2013]
  - Built a fault-tolerant BSP framework (NEWT) [Kromonov et al, HPCS 2014]

Research Challenges

MOBILE CLOUD



The Seven Mass Media

First Mass Media Channel - ***Print*** from the 1500s

Second Mass Media Channel - ***Recordings*** from 1900s

Third Mass Media Channel - ***Cinema*** from 1910s

Fourth Mass Media Channel - ***Radio*** from 1920s

Fifth Mass Media Channel - ***TV*** from 1950s

Sixth Mass Media Channel - ***Internet*** from 1990s

Seventh Mass Media Channel - ***Mobile*** from 2000s

[Tomi T Ahonen]

Rank	Country or region	Number of mobile phones	Population	Phones per 100 citizens	Data evaluaton date
-	World	6,800,000,000+	7,012,000,000[1]	87	2013[2][3]
01	China	1,206,553,000[4]	1,349,585,838[5]	89.2	September 2013[4]
02	India	867,800,000	1,220,800,359[6]	70.72	30 April 2013[7]
03	United States	327,577,529	310,866,000[8]	103.9	June 2013[9]
04	Brazil	268,440,423	192,379,287[10]	135.4	August 2013[11]
05	Russia	256,116,000	142,905,200[10]	155.5	July 2013[12]
06	Indonesia	236,800,000	237,556,363	99.68	September 2013[10]
07	Pakistan	129,583,076	178,854,781[13]	72.45	September 2013[14]
08	Japan	121,246,700	127,628,095	95.1	June 2013[15]
09	Nigeria	114,000,000	165,200,000	69	May 2013[16]
10	Bangladesh	110,675,000	165,039,000	73.8	September 2013[17]

Report: Mobile cloud to grow beyond $11 billion in 2018

Written by CopperEgg // July 12, 2012 // No Comment // Cloud Performance

The proliferation of smartphones, tablets and other mobile devices is contributing to change in the private sector, as businesses continue to leverage these gadgets in an attempt to enhance efficiency and potentially gain a competitive advantage. According to a new report by Global Industry Analysts, the evolution of mobility is also changing the cloud computing landscape, pushing the mobile cloud market to generate more than $11 billion in revenue by 2018.

Maribel Lopez, (

I track how mobile

+ Follow (87)

TECH | 4/18/2012 @ 7:43AM | 18,825 views

Verizon's Stratton: The Future Of IT Is Mobile And Cloud

+ Comment Now **+ Follow Comments**

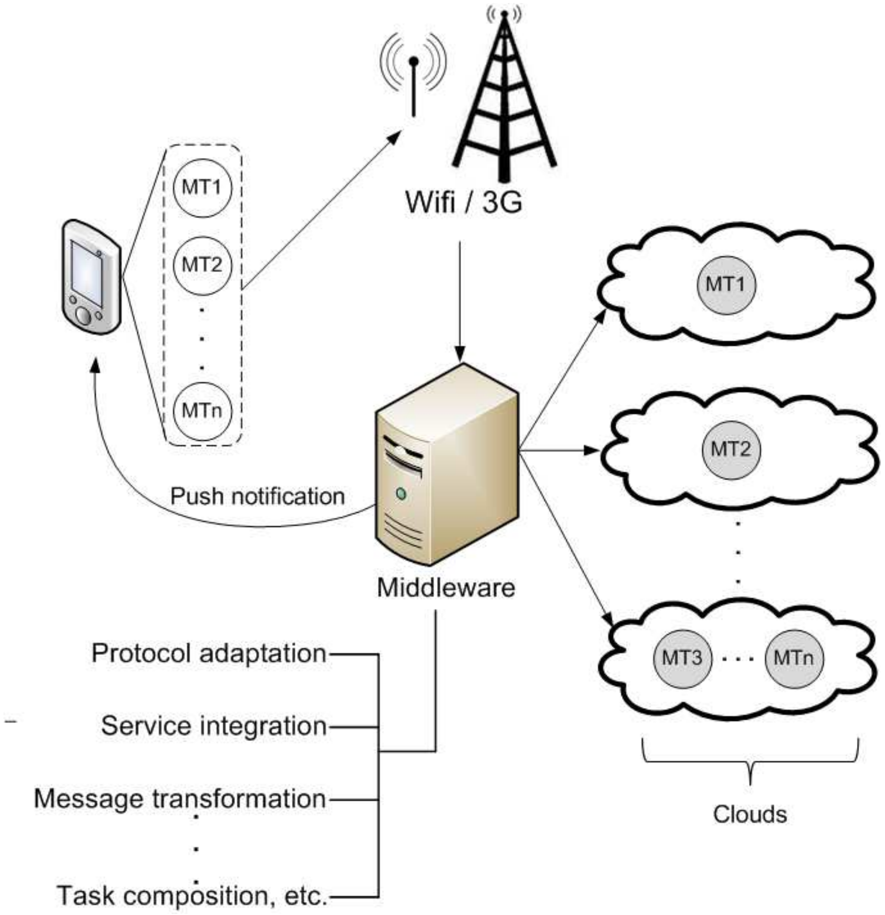
Mobile Applications

- One can do interesting things on mobiles directly
  - Today’s mobiles are far more capable
  - Location-based services (LBSs), mobile social networking, mobile commerce, context-aware services etc.
- It is also possible to make the mobile a service provider
  - Mobile web service provisioning [Srirama et al, ICIW 2006; Srirama and Paniagua, MS 2013]
  - Challenges in security, scalability, discovery and middleware are studied [Srirama, PhD 2008]
  - Mobile Social Network in Proximity [Chang et al, ICSOC 2012; PMC 2014]

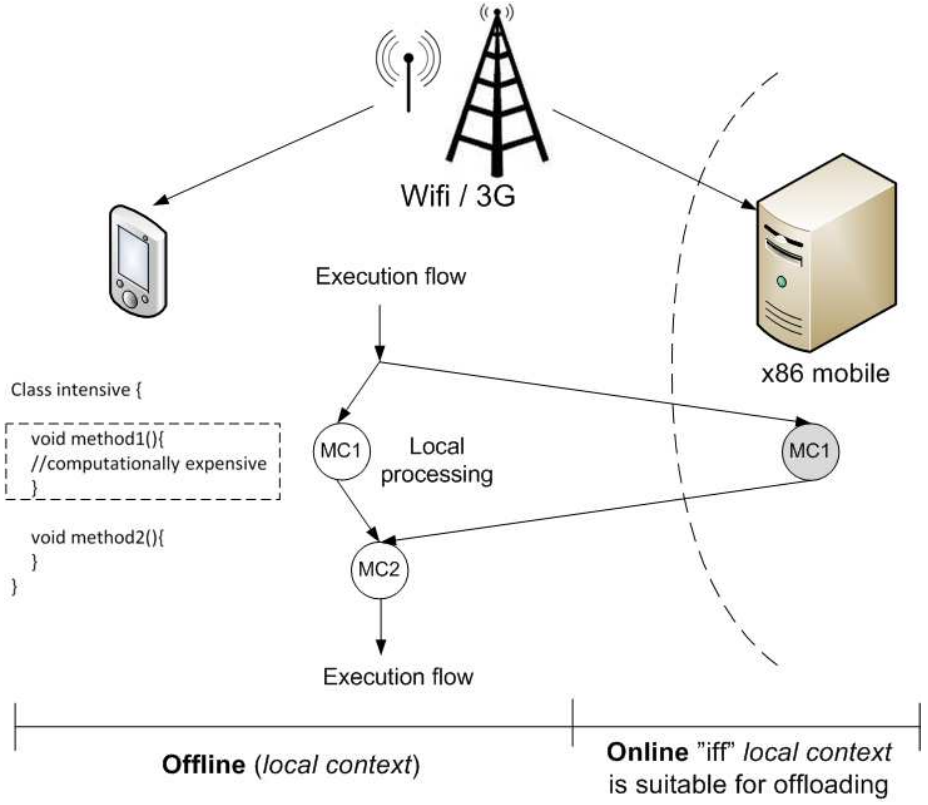
Mobile Cloud Applications

- Bring the cloud infrastructure to the proximity of the mobile user
- Mobile has significant advantage by going cloud-aware
  - Increased data storage capacity
  - Availability of unlimited processing power
  - PC-like functionality for mobile applications
  - Extended battery life (energy efficiency)

Mobile Cloud Binding Models



Task Delegation



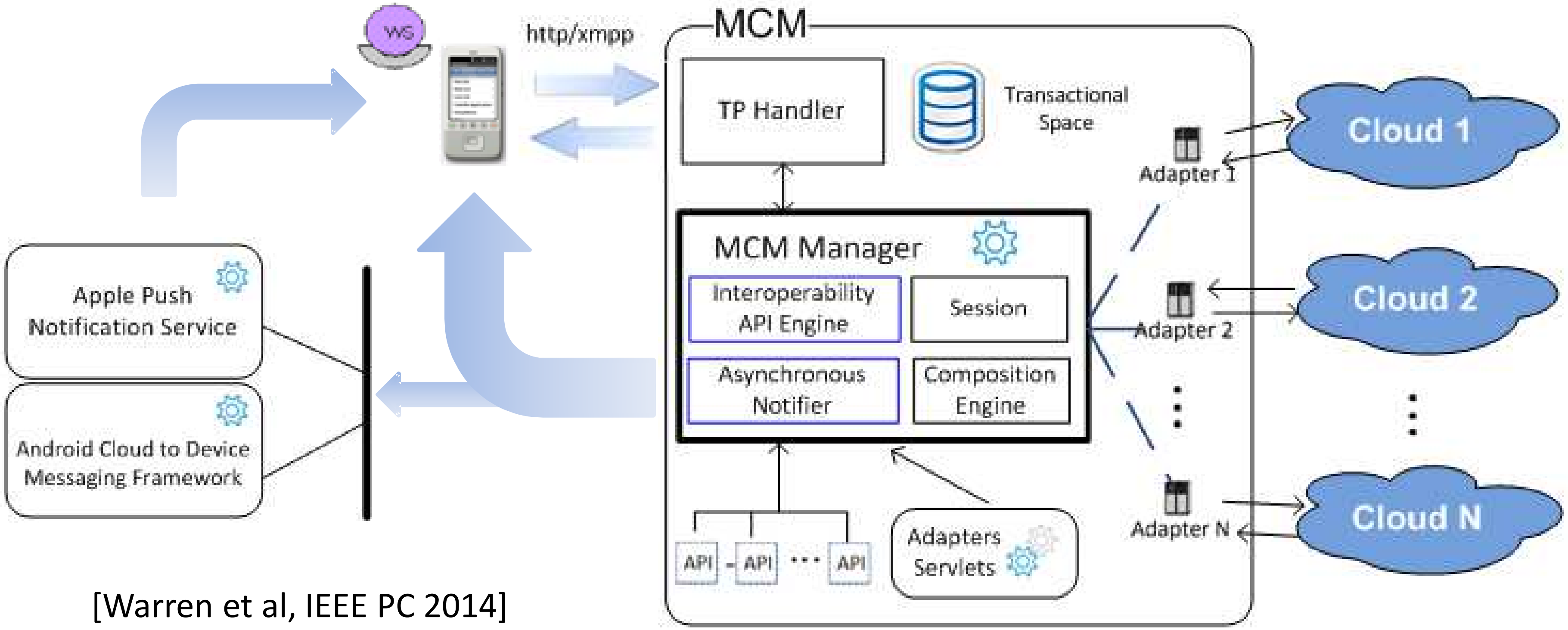
Wifi / 3G
Execution flow
x86 mobile
MC1
Local processing
MC1
MC2
Execution flow
Offline (*local context*)
Online "iff" *local context* is suitable for offloading

```
Class intensive {

  void method1(){
  //computationally expensive
  }

  void method2(){
  }
}
```

Code Offloading

[Flores & Srirama, JSS 2013]

Mobile Cloud Middleware



[Srirama and Paniagua, MS 2013]

[Warren et al, IEEE PC 2014]

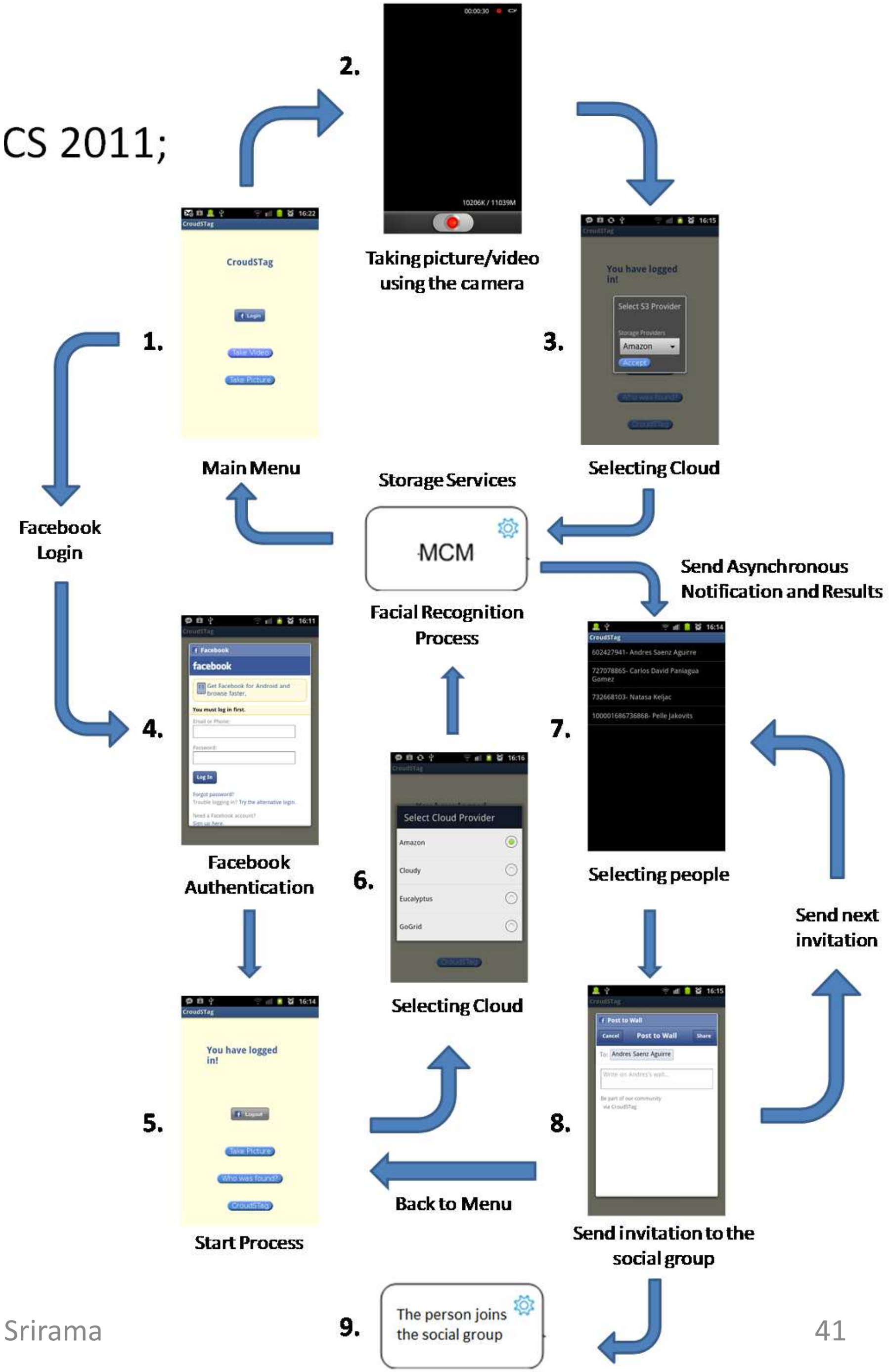
[Flores et al, MoMM 2011; Flores and Srirama, JSS 2013]

CroudSTag – Scenario

- CroudSTag takes the pictures/videos from the cloud and tries to recognize people
  - Pictures/Videos are actually taken by the phone
  - Processes the videos
  - Recognizes people using facial recognition technologies
- Reports the user a list of people recognized in the pictures
- The user decides whether to add them or not to the social group
- The people selected by the user receive a message in facebook inviting them to join the social group

CroudSTag [Srirama et al, PCS 2011; SOCA 2012]

- Cloud services used
  - Media storage on Amazon S3
  - Processing videos on Elastic MapReduce
  - face.com to recognize people on facebook
  - Starting social group on facebook

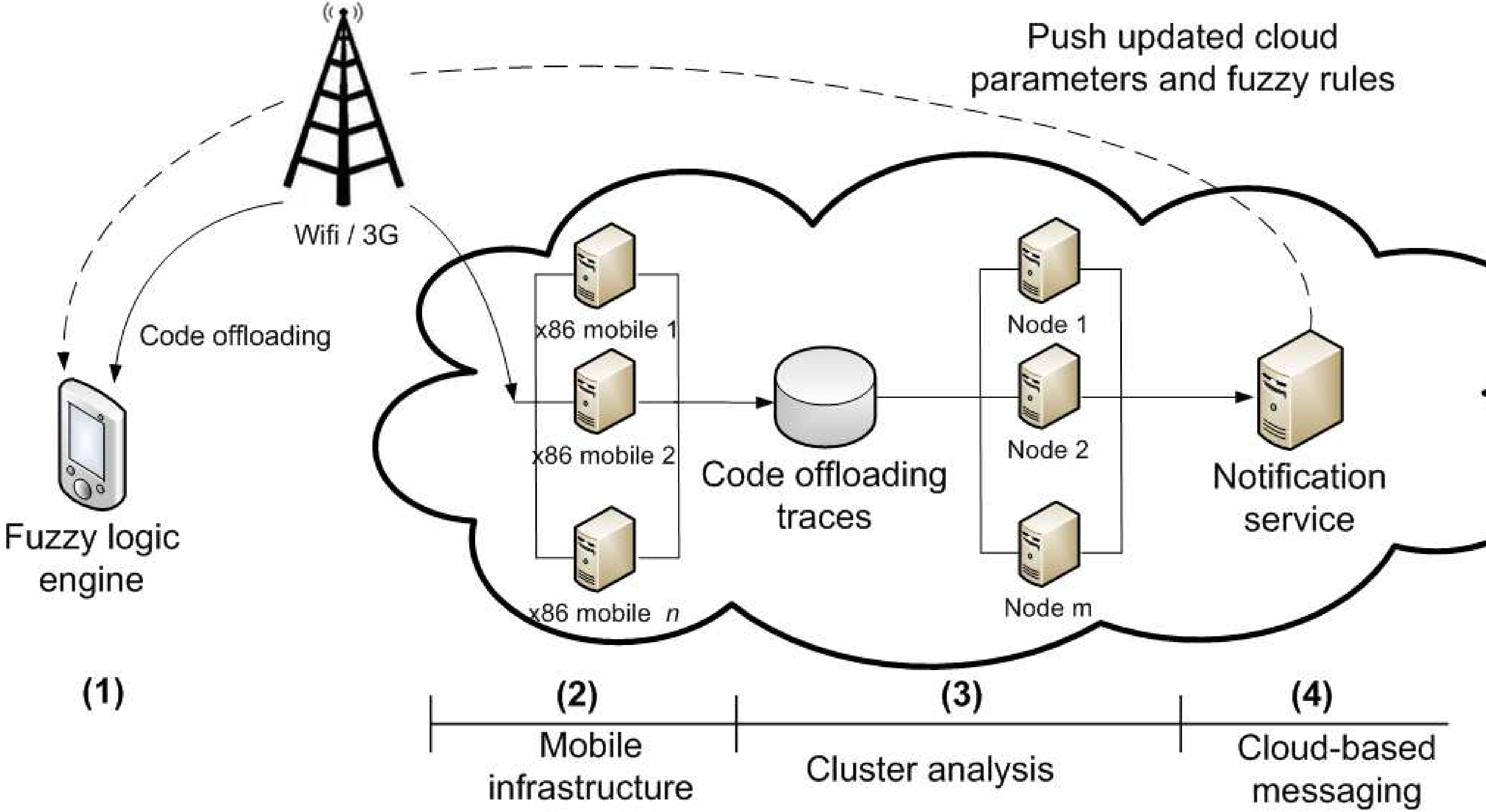


Code Offloading

- Studied extensively by community [MAUI, Cloudlets etc.]
- Is *Mobile Cloud* taking full advantage of *Cloud Computing*?
  - Parallelization and elasticity are not exploited
- Offloading from a different perspective
  - *"Offloading is a global learning process rather than just a local decision process"* [Flores and Srirama, MCS 2013]
- How it can learn?
  - Analysis of code offloading traces which are generated by the massive amount of devices that connect to cloud

"EMCO: Evidence-based mobile code offloading"

Evidence-based Mobile Code Offloading



[Flores and Srirama, MCS 2013]

Process-intensive Tasks on Cloud

- Media processing
  - CroudSTag demonstrates image and video processing
- Sensor data analysis
  - Human activity recognition [Srirama et al, NGMAST 2011]
  - Context aware gaming
  - MapReduce based sensor data analysis [Paniagua et al, MobiWIS 2012]

Research Challenges

INTERNET OF THINGS

Internet of Things (IoT)

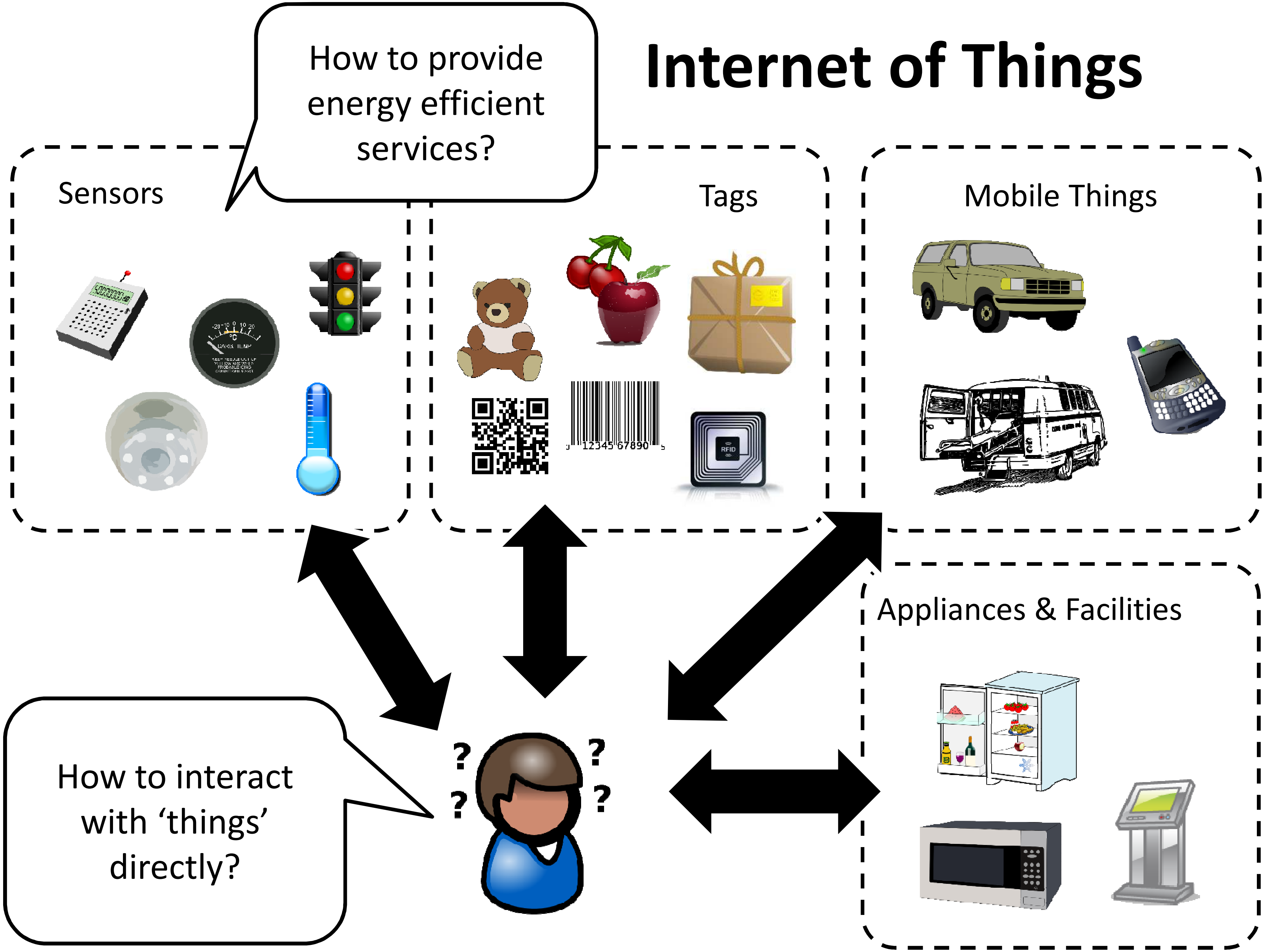
“*The Internet of Things allows people and things to be connected **Anytime**, **Anyplace**, with **Anything** and **Anyone**, ideally using **Any path/network** and **Any service***”—(Guillemin and Friess, 2009)

US National Intelligence Council has predicted that “*by 2025 Internet nodes may reside in everyday things–food packages, furniture, paper documents, and more*”

UK + Germany governments supported £73 million for IoT (2014).

China government is planning to invest $800 million by 2015.

Internet of Things



How to provide energy efficient services?

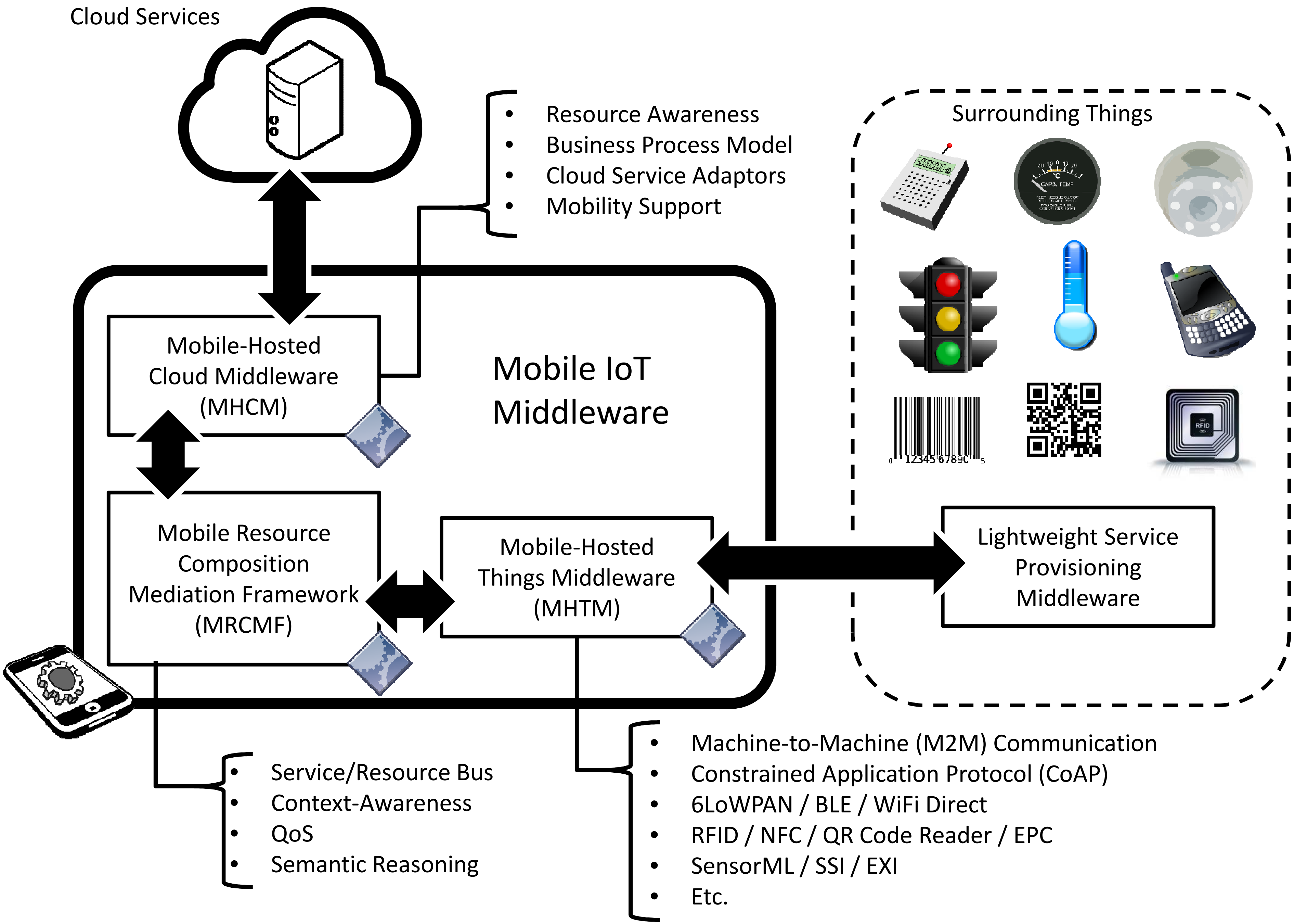
Sensors

Tags

Mobile Things

Appliances & Facilities

How to interact with ‘things’ directly?


Cloud Services
Resource Awareness
Business Process Model
Cloud Service Adaptors
Mobility Support
Surrounding Things
Mobile-Hosted Cloud Middleware (MHCM)
Mobile IoT Middleware
Mobile Resource Composition Mediation Framework (MRCMF)
Mobile-Hosted Things Middleware (MHTM)
Lightweight Service Provisioning Middleware
Service/Resource Bus
Context-Awareness
QoS
Semantic Reasoning
Machine-to-Machine (M2M) Communication
Constrained Application Protocol (CoAP)
6LoWPAN / BLE / WiFi Direct
RFID / NFC / QR Code Reader / EPC
SensorML / SSI / EXI
Etc.
RFID

Research Results

- Participated in a number of EU-funded projects
- Partner in the Estonian Center of Excellence in Computer Science
- Partner in Software Technology and Applications Competence Centre (STACC)
  - An R&D center that conducts industry-driven research projects in the fields of software engineering and data mining
- Output resulted in several SMEs
  - Plumbr, ZeroTurnaround etc.

Garage48, Startups, SME-s, .. #estonianmafia

e-ESTONIA

iktp

European Union
Regional Development Fund

Investing in your future

Eesti Teadusagentuur
Estonian Research Council

srirama@ut.ee

THANK YOU FOR YOUR ATTENTION