NTT

NTT Information Sharing Platform Laboratories
NTT 情報流通プラットフォーム研究所

User Demand Prediction from Application Usage Pattern in Virtual Smartphone

Joon Heo *(presenter), Kenji Terada, Masashi Toyama,*
Shunsuke Kurumatani and Eric Y. Chen

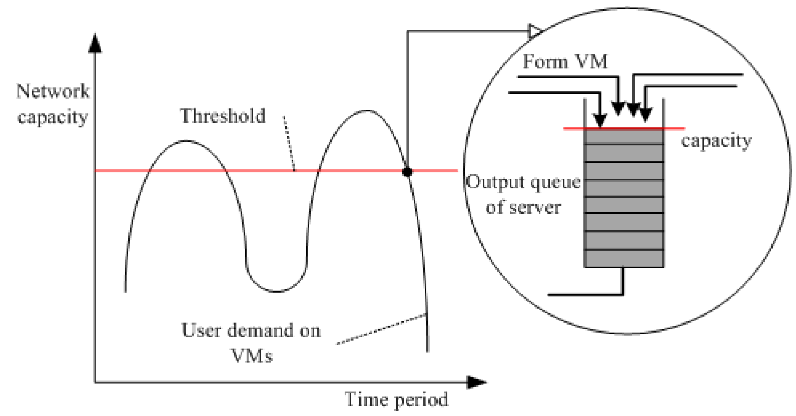
NTT Information Sharing Platform Laboratories, NTT Corporation

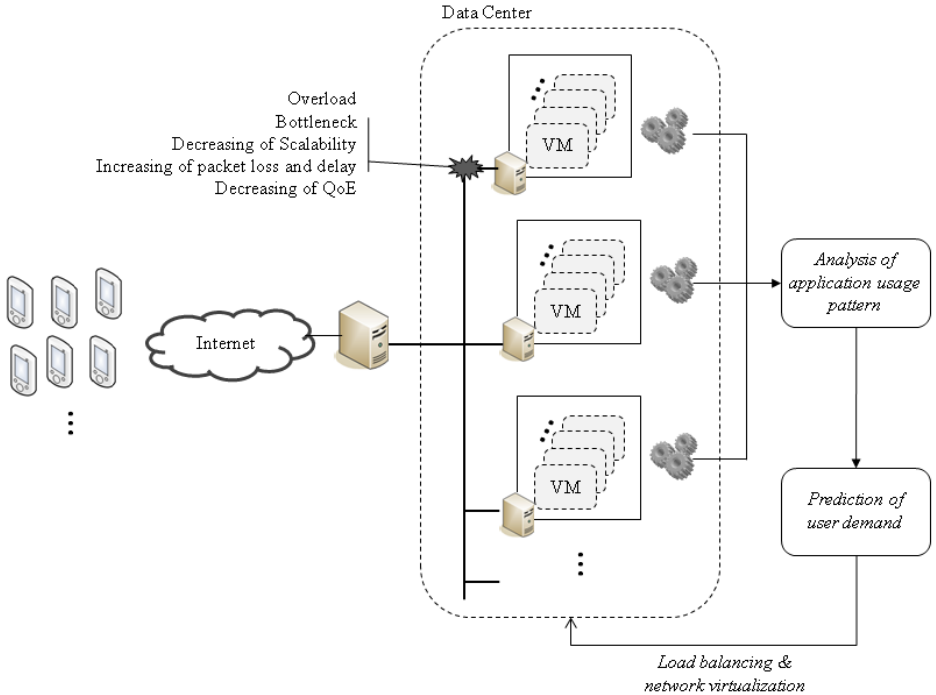
IEEE CloudCom 2010
December 02, 2010

Contents

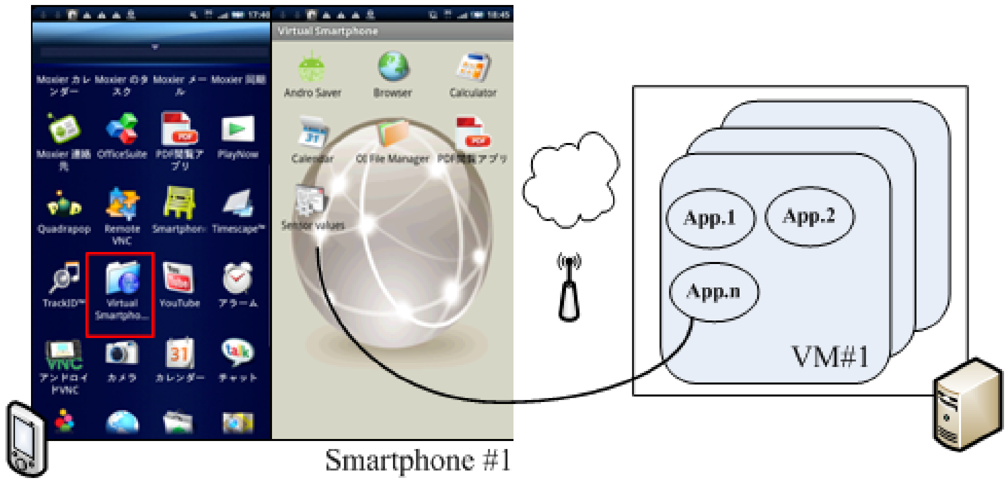
Motivation

- Cloud computing : virtualization of computing resources and storage
- Applications continue to become more data-intensive
- Overload problems
- Network virtualization and load balancing technique can be used
- Accurate prediction of user demand is need firstly "can guide resource allocation of server and enhance network performance with avoiding congestion and bottleneck"
- Find application usage pattern and predict user demand of each VM and server.
- Be reflected to load balancing and network virtualization scheme





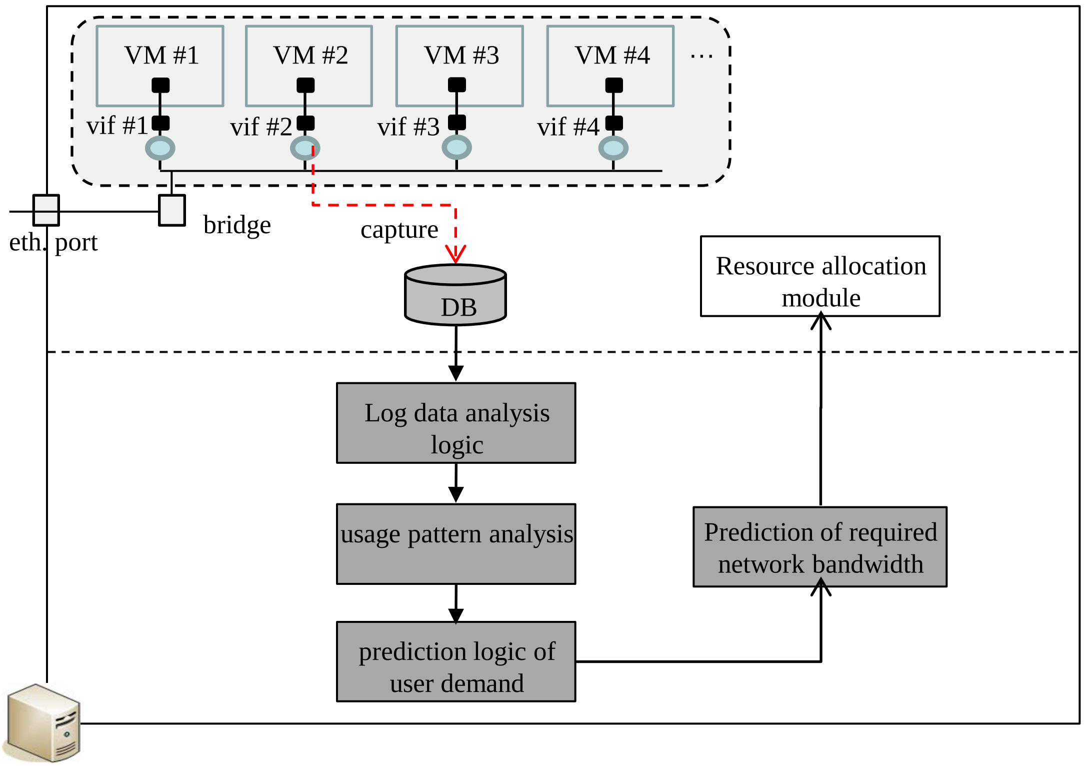
Virtual Smartphone System



• A variety of smartphones equipped with increasingly faster CPUs and larger memory; however, hardware capacity is still very limited.
• Virtual smartphone system, which provides a cloud computing environment
• To enable smartphone users to more easily tap into the power of the cloud and to free themselves from the limits of the physical smartphone

• Users control their virtual smartphone images through a dedicated client application
• Most users to access their virtual smartphone images through an unstable network such as 3G
• Be in the same state when the user is connected again after user is disconnected in an unexpected manner.

User Demand Prediction Method (1/6)

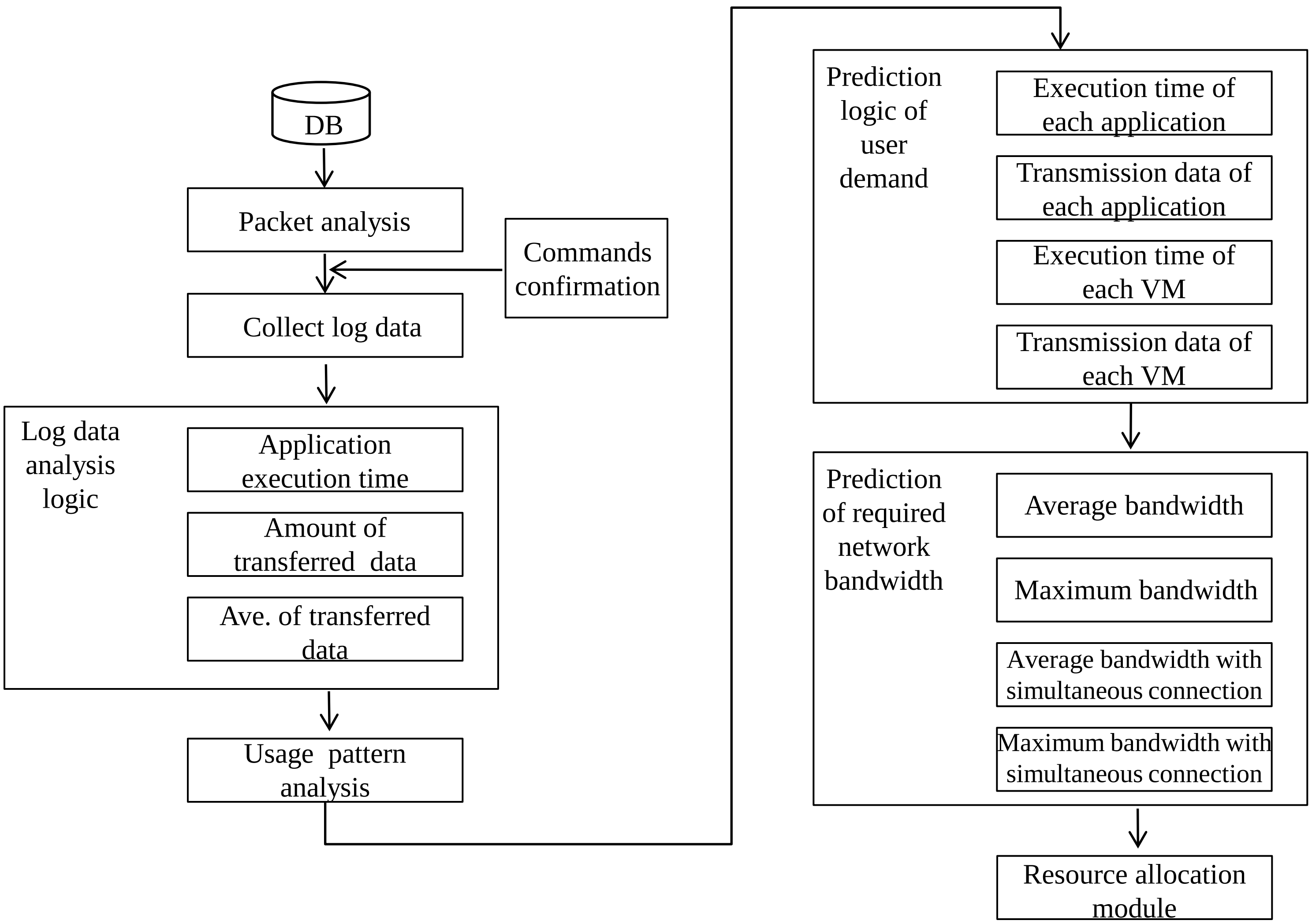


- The application execution by a specific user follows the patterns such as airtime and preferred control

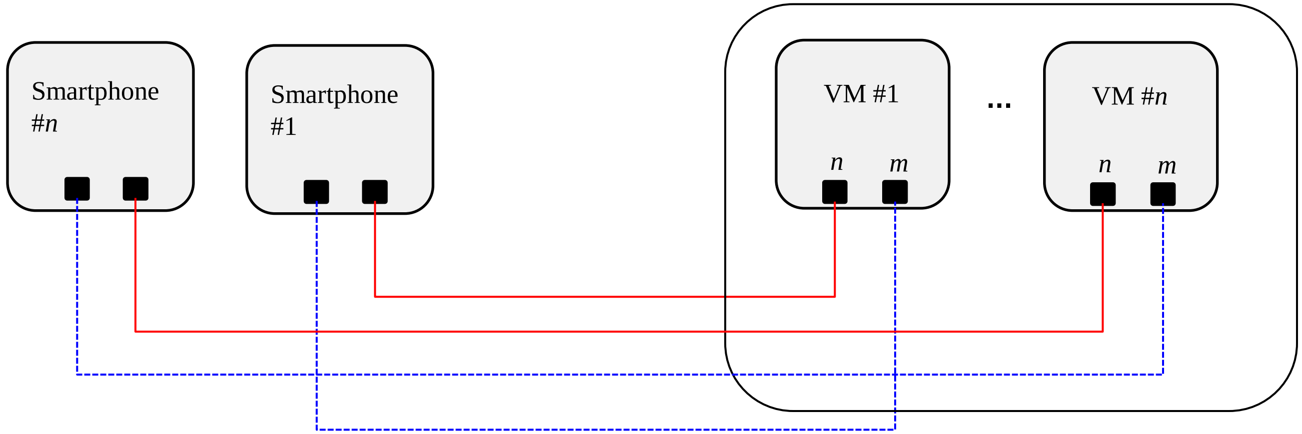
- Each VM is allocated to a specific smartphone user → connect to this VM to execute applications and can receive changed screens or application data

- Filter and capture packets at the front of the virtual interface of each VM
- Analyze the usage pattern and calculate the average amount of transferred data

User Demand Prediction Method (2/6)



User Demand Prediction Method (3/6)

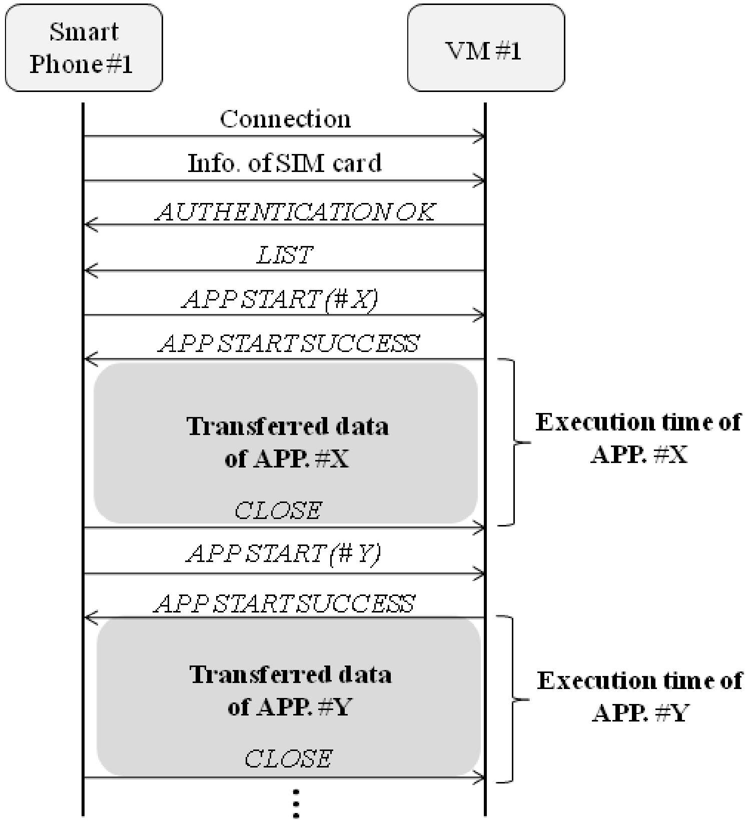


- A general link, port *n*, is used for maintenance of link connections, authentication and disconnection.
- A transmission link, port *m*, is used for transmission of executed application data or changed VM screens.

Sever port	Role
n	Maintenance of link connection
	Authentication
	Commands exchange
	Disconnection
m	Transmission of application data
	Transmission of executed result

User Demand Prediction Method (4/6)

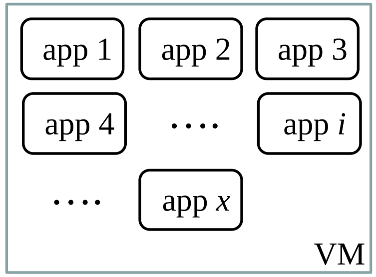
Type of command	Transmission	Description
AUTHENTICATION OK	*VM→S.Phone*	Notice of authentication by SIM card information
LIST	*VM→S.Phone*	Transmission of current application list
APP START	*S.Phone→VM*	Notice of chosen Application
APP START SUCCESS	*VM→S.Phone*	Notice of application Execution
CLOSE	*S.Phone→VM*	Notice of application Closing



- New application layer commands to connect and to control the VM; all of these commands are exchanged through port *n*.

- Overall process and exchange of defined commands from initial connection.
- *execution time* of the specific application that was selected by the user from APP START SUCCESS to CLOSE and the *transferred data* during this time.

User Demand Prediction Method (5/6)



Collect log data

$$\overrightarrow{d_p} = (d_p^1, d_p^2, d_p^3) \ \ldots(1)$$

- $\overrightarrow{d_p}$ is the vector value of d *th* log data in period p
- d_p^k is the time domain data of d *th* log data in period p
  - ✓ d_p^1: executed application
  - ✓ d_p^2 : transmitted traffic volume
  - ✓ d_p^3 : execution time

if we have x applications and y log data in period p

Total transferred data volume

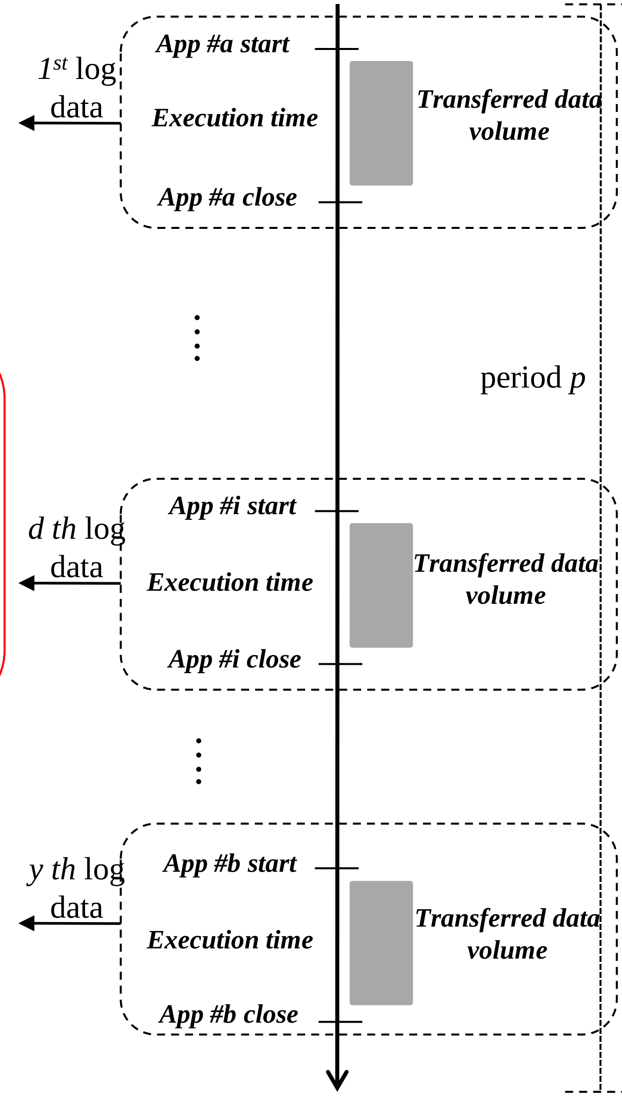
$$\overrightarrow{v_p} = (v_p^1, \ldots, v_p^n) \text{, where } v_p^i = \sum_{d=1}^{y} d_p^2, (d_p^1 = i) \ \ldots(2)$$

Total execution time

$$\overrightarrow{T_p} = (T_p^1, \ldots, T_p^n) \text{, where } T_p^i = \sum_{d=1}^{y} d_p^3, (d_p^1 = i) \ \ldots(3)$$

Average transmission data volume

$$V_p^i = \frac{v_p^i}{T_p^i} \ \ldots(4)$$



User Demand Prediction Method (6/6)

We can predict total execution time and transmission data volume of n+1 period

for $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1$ (use Holt's linear method).

The execution time follows the equations:

$$(level)\ L_n^t = \alpha T_n^i + (1-\alpha)(L_{n-1}^t + b_{n-1}^t) \quad \ldots(5)$$

$$(trend)\ b_n^t = \beta(L_n^t - L_{n-1}^t) + (1-\beta)b_{n-1}^t \quad \ldots(6)$$

$$T_{n+1}^i = L_n^t + b_n^t \quad \ldots(7)$$

And the transmission data volume follows the equations:

$$(level)\ L_n^e = \alpha V_n^i + (1-\alpha)(L_{n-1}^e + b_{n-1}^e) \quad \ldots(8)$$

$$(trend)\ b_n^e = \beta(L_n^e - L_{n-1}^e) + (1-\beta)b_{n-1}^e \quad \ldots(9)$$

$$V_{n+1}^i = L_n^e + b_n^e \quad \ldots(10)$$

We can forecast the transmission load of each application (s_{n+1}^i) and VM (Q).

$$s_{n+1}^i = T_{n+1}^i * V_{n+1}^i \quad \ldots(11)$$

$$Q = \sum_{i=1}^{x} s_{n+1}^i \quad \ldots(12)$$

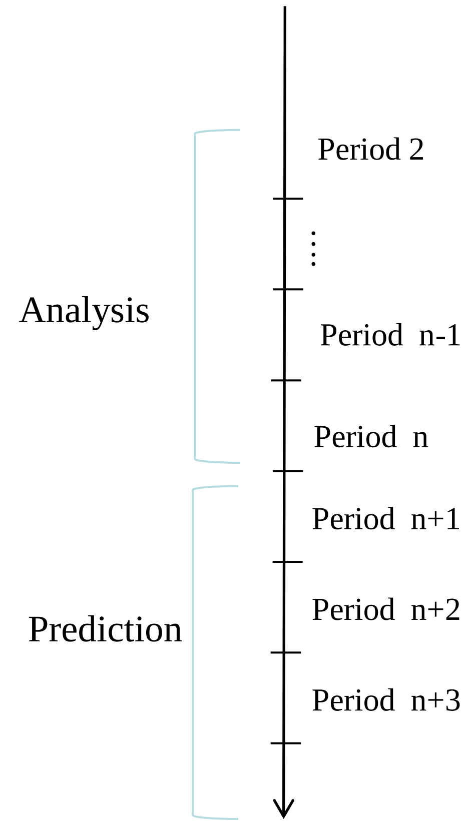
We can forecast the total execution time of each VM (F)

$$F = \sum_{i=1}^{x} T_{n+1}^i \quad \ldots(13)$$

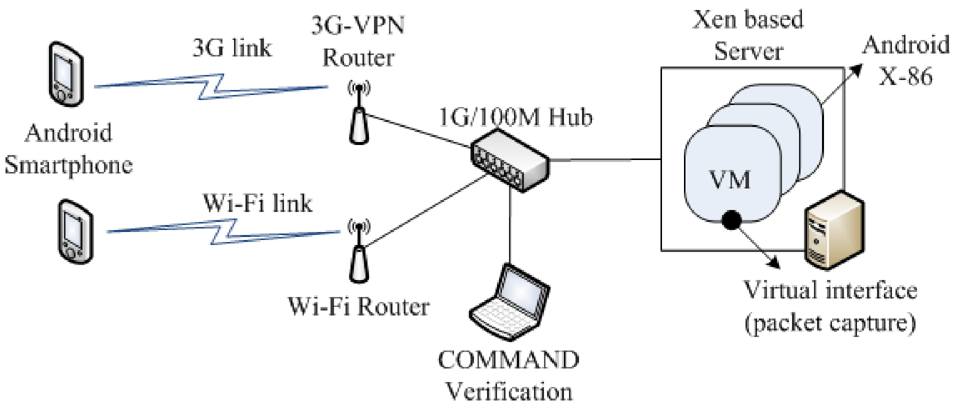
Suppose that a specific server has k VMs.

The predicted total volume of transmission data of server during n+1 period:

$$V_Q^{n+1} = Q_1 + Q_2 + \ldots + Q_k \quad \ldots(14)$$

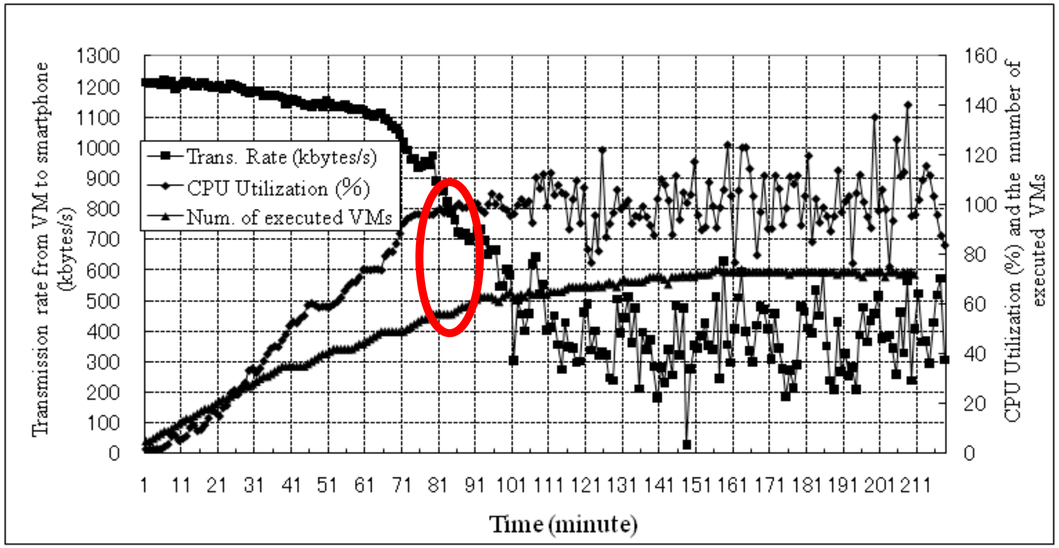


Experiment Results (1/5)



- Android by Google
- smartphone (XPERIA X10 by Sony Ericsson)
- 100 Android-x86 images as the virtual machine
- on Xen hypervisor

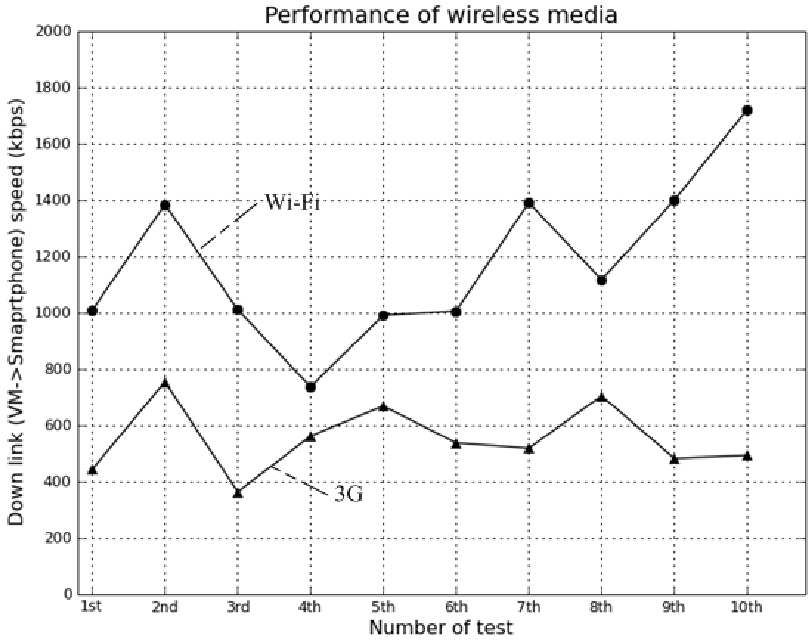
- 3G and Wi-Fi
- libpcap library
- Wireshark tool



- Performance impacts, CPU utilization of host server and transmission speed of VM, by increasing the number of VM at a same time.

Experiment Results (2/5)

- Network applications on smartphones heavily depend on wireless media factors, such as bandwidth, latency, and bit rate.
- 3G and Wi-Fi performances in the experiment environment.
- The transferred data volume is about 1,080 KB



Performance of wireless media in experiment environment

RTT Average of Data Transmission Link

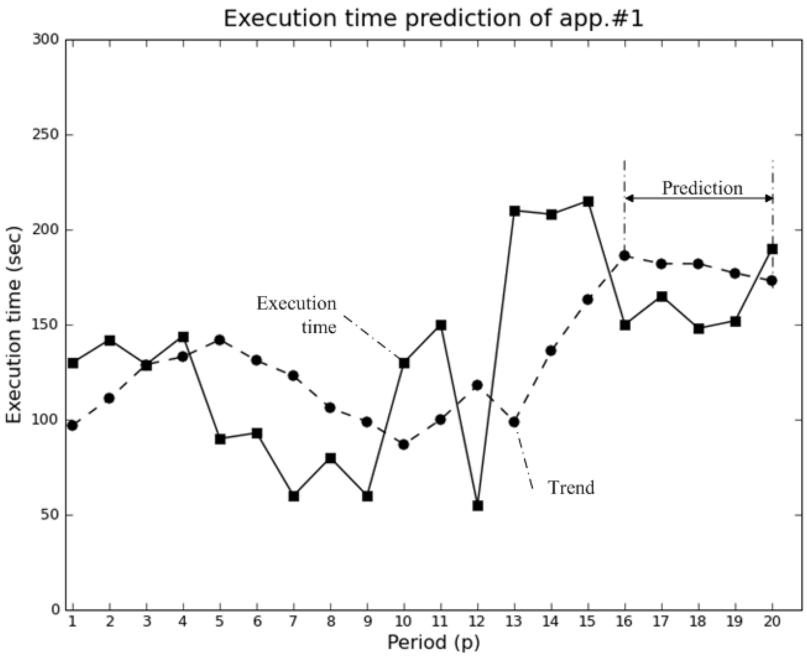
Num. of test	Wi-Fi	3G
1	23 ms	352 ms
2	19 ms	242 ms
3	29 ms	225 ms
4	20 ms	391 ms
5	18 ms	197 ms
6	21 ms	248 ms
7	25 ms	261 ms
8	19 ms	291 ms
9	15 ms	278 ms
10	19 ms	213 ms

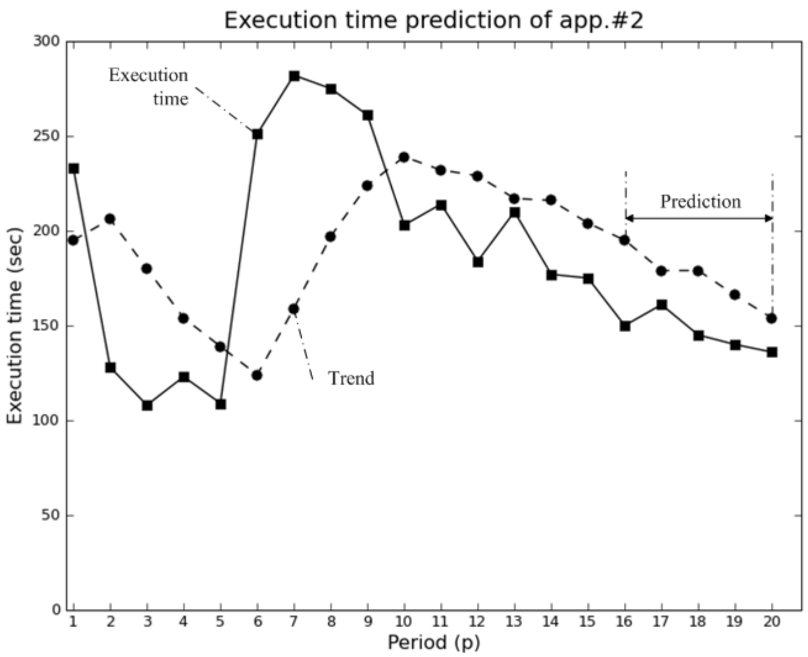
Experiment Results (3/5)

Screenshot of distinguished link and commands exchange

- port 5700 as the *n*
- port 5901 as the *m*

Experiment Results (4/5)

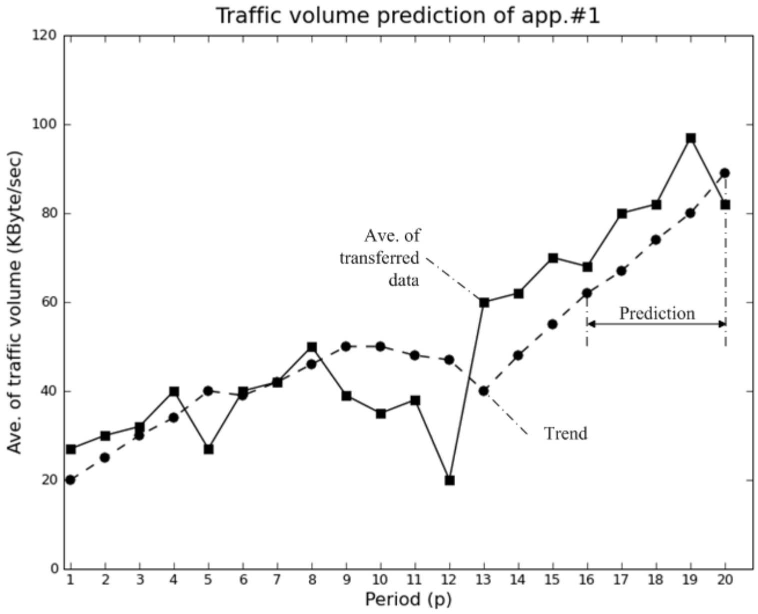


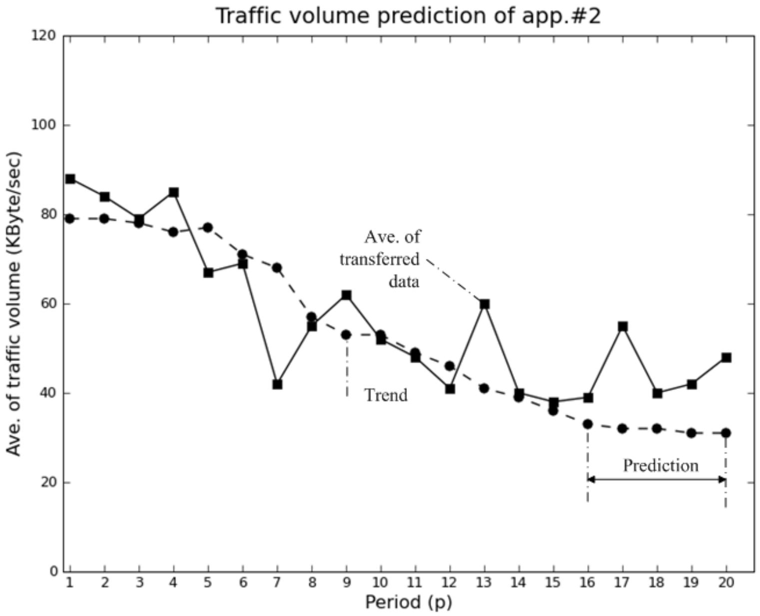


- To show the prediction values and to compare with actual values, we chose a VM which has continuing connection from smartphone in trial server of test bed.

- analyzed the application usage pattern of user who connected to this VM and executed some applications
- time period *p* for this analysis was 30 minutes.

Experiment Results (5/5)





- Since we defined a short time period, our experiment results do not show periodic changes such as airtime and preferred execution.

- Can find periodic characteristic of each application when we used long-term log data for practical users

Conclusion and Future works

- ✓ Should maintain the optimum status of the server to overcome the performance limitations
- ✓ The prediction of user demand and workload are significant factors

- ✓ User demand prediction method that uses analysis results of application usage patterns

- ✓ Can calculate the execution time and transferred data volume of each application, each VM and server

- ✓ As the future works…
    - To develop a more enhanced prediction method and parameters
    - To find user profiling
    - To develop a load balancing and network virtualization method

Q & A