Communication-aware Computing Resource Management in Super Base Station for Future 5G

Prof. Yiqing Zhou
Wireless Communication Technology Research Center
Institute of Computing Technology, CAS

A Joint Work with Guowei Zhai, Lin Tian and Jinglin Shi
2015.09
Outline

- Background
- System Model and Problem Formulation
- Load Diversity Aware Computing Resource Management
- Real-Time Guaranteed Processing for TDD Protocols
- Conclusions
Background—What for 5G?

- Target of 5G: By 2020, the capacity of mobile communication networks should be boosted by **thousands times** to support the huge data traffic.

- Possible Solutions
  - More spectrums
  - Enhanced technologies with high spectral efficiency
    - Massive MIMO; Full-duplex comm.; ..... 
  - Adopting heterogeneous cellular network architectures with dense wireless access nodes
    - Dense cellular network (DCN)
Background-Dense Cellular Network

- One dominant factor of gain for the wireless network capacity is the **density** of access points in the network
  - From the very beginning till now, mobile network capacity has been boosted by **1000 times** by increasing the density of BSs
  - personal indoor access (femtocell), street outdoor coverage (picocell), relay, remote radio head (RRH)

- **Dense Cellular Network:** A further extension/enhancement of **heterogeneous cellular networks** and an important feature of 5G
Background-Challenges in DCN

- In traditional distributed mobile networks, BSs are independent to each other. Each BS should be set up and maintained separately. As the number of BSs increases
  - Difficulty to find BS sites
  - Higher energy consumption
  - Higher maintaining cost
  - Low computing resource utilization
Centralized mobile network architectures are proposed, which keep only the antenna part at the original BS site and move all signal processing functions to a central unit:

- Much easier to find a site for an antenna than a BS
- Reduce energy consumption and maintaining cost
- Improve computing resource utilization

Various schemes: CRAN (China Mobile), Super Base Station (ICT), Wireless Network Cloud (IBM)
Background-5G Heterogeneous Network Architecture

Network side: Super BS converging communication and computing

Wireless Access: Dense Heterogeneous Network
Outline

● Background

● System Model and Problem Formulation

● Load Diversity Aware Computing Resource Management

● Real-Time Guaranteed Processing for TDD Protocols

● Conclusions
Super BS: Centralized Protocol Processing

CPU Processing Capability

<table>
<thead>
<tr>
<th>PPS1</th>
<th>PPS2</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU1</td>
<td>CPU1</td>
</tr>
<tr>
<td>CPU2</td>
<td>CPU2</td>
</tr>
<tr>
<td>CPU3</td>
<td>CPU3</td>
</tr>
<tr>
<td>CPU4</td>
<td>CPU4</td>
</tr>
</tbody>
</table>

PPS Processing Resource Management

PPS: Protocol Processing Server

VBS-PE: Virtual Base Station Protocol Entity
Problem 1: Under Utilization of Computing Resource

Mobile network load is dynamic: Tide-like Effect

Allocating computing resource according to the peak requirement results in serious under utilization of resources
Problem 2: Difficult to guarantee real-time processing

Multiple VBS-PEs are running on one CPU: Real-time processing cannot be guaranteed.

How to arrange the computing of communication tasks so that real-time processing can be guaranteed with high resource efficiency?
Outline

● Background
● System Model and Problem Formulation
● Load Diversity Aware Computing Resource Management
● Real-Time Guaranteed Processing for TDD Protocols
● Conclusions
How to exploit load diversity in computing resource management?

According to the load variation in mobile communication networks, VBS-PE should be mapped to PPS exploiting load discrepancy.

Main idea

For small area coverage:
Optimized computing
Resource allocation

For large area coverage:
Sub-optimal computing
Resource allocation
Framework of Protocol Processing Resource Management

Protocol Processing Server 1
- VBS-PE1
- VBS-PE2
- VBS-PE3
- VBS-PE4
- Processing Resource Management
  - Load Tracking
  - Processor Mapping
  - VBS-PE Migration

Protocol Processing Server 2
- VBS-PE5
- VBS-PE6
- VBS-PE7
- VBS-PE load statistics
- VBS-PE to PPS Mapping

Processing Resource Allocation Server
- VBS-PE1
- VBS-PE2
- VBS-PE3
- VBS-PE4
- VBS-PE5
- VBS-PE6
- VBS-PE7
Three types of computing resource consumption

Processor

VBS1 | VBS2 | VBS3 | VBS4 | VBS5

Protocol load processing ($W_{load}$)

Protocol task scheduling ($W_{sched}$)

Processing Resource Adjusting ($W_{adj}$)

VBS Migration

Processor

$W_{load}$

$W_{sched}$

$W_{adj}$
Problem Formulation

\[
\max \quad \text{Total Efficiency} \\
= \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{W_{load}[i, j]}{N} \sum_{i=1}^{M} W_{max}[i]
\]

\(S.T.\)
\(W_{load}[i, j] + W_{sched}[i] + W_{adj}[i, j] < W_{max}[i]\) 
\((\text{For any } i, j)\)

**i**: index of Protocol Processing Server (PPS)

**j**: index of time stage

**\(W_{max}\)**: the computing capacity of PPS

This optimization problem can be solved using dynamic programming (DP)
Load Diversity Based Allocation (LDA)

PPS maintains a high processing resources utilization in each time stage.
Applicable for small area coverage.
Distributed Load Diversity Based Allocation (D-LDA) for Large Area Coverage

LDA can be executed in each geographical region independently.

LDA cannot get an apparent good result if most VBSs in a geographical region have similar load variation.
Hybrid Load Diversity Based Allocation (H-LDA) for Large Area Coverage

LDA can be executed in each VBS group

The proportion of VBSs with different load variation in each group is similar
Simulation Configurations

VBSs are running in PPSs with 4 processors, which have a same resource capacity. CFS (Completely Fair Scheduler) are used as scheduler.

The values of the three types of processing resource consumption are obtained by running the LTE protocol stack software developed by our center.

The variation of load can be represented by the variation of active user number.
Simulation Result

- Computing efficiency as a function of VBS number
- The user density of each VBS is fixed
- Two area types (Business and residential area)

![Graph showing efficiency as a function of VBS number. The graph indicates an increase of +58%.]
Outline

- Background
- System Model and Problem Formulation
- Load Diversity Aware Computing Resource Management
- Real-Time Guaranteed Processing for TDD Protocols
- Conclusions
Real-time Guaranteed TDD Protocol Processing

Analyze the main procedures in TDD protocol processing

Decide the processing method of different procedures

Main idea

Focusing on the DL/UL processing which is of high and dynamic computing load, design real-time guaranteed computing resource management scheme
Analysis of Main Procedures in Protocol Processing

Various Procedures in VBS-PE

Low load/ almost static
Allocate fixed and small portion of computing resource

High load/ Dynamic
Different and variable computing resources are allocated to VBS-PE

Focus of protocol processing

CPU1
- VBS-PE 1
- VBS-PE 2
- VBS-PE 3

CPU2
- 4
- 5
- 6
- 7

CPU3
- 8
- 9

Fixed load
Reserved

接入控制
连接管理
信令处理

用户及数据调度
上下行数据处理
无线资源管理中用户或业务相关管理

24
DL/UL Protocol Processing-1

Physical layer subframe configuration

<table>
<thead>
<tr>
<th>D</th>
<th>S</th>
<th>U</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>S</th>
</tr>
</thead>
</table>

Protocol processing subframe cycle

Case 1:

- Downlink load: Uplink load = 3 : 2
- Npre = 2

Case 2:

- Downlink load: Uplink load = 3 : 2
- Npre = 3

Real-time DL processing cannot be ensured!
Using fixed DL pre-processing, either real time cannot be guaranteed or computing resource is under utilization!

Case 2

Downlink load: Uplink load = 3:7
Npre = 3

Computing resource is under utilization!
Modeling Real-time Protocol Processing Requirement-1

- **System parameters**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_f$</td>
<td>The total number of subframes, UL subframes and DL subframes in a protocol process frame</td>
</tr>
<tr>
<td>$N_{uf}$</td>
<td>Length of a subframe (1ms in TD-LTE)</td>
</tr>
<tr>
<td>$N_{df}$</td>
<td>The ratio of UL load and DL load</td>
</tr>
<tr>
<td>$N_{pre}$</td>
<td>The number of subframes that DL processing should be moved ahead</td>
</tr>
<tr>
<td>$T_{idp}$</td>
<td>Without UL processing, the time needed to process the DL data in a protocol processing frame</td>
</tr>
</tbody>
</table>
Modeling Real-time Protocol Processing

Requirement-2

\[ T_1 = (N_{pre} - N_{Uf}) \times Ts \]
\[ T_2 = (T_{idp} \times N_{Df} - T_1) \times 2 \]
\[ T_1 + T_2 > (N_{pre} + N_{Df} - 1) \times Ts \]

Case 1: Computing resource is fully utilized, however, DL real-time processing cannot be guaranteed.

\[ T_{idp} < Ts \]

Case 2: DL real time processing is guaranteed, however, computing resource is under utilized.

DL processing must be moved \( N_{pre} \) sub-frames earlier.

\[ N_{pre} = \lceil N_f \times K/(1 + K) \rceil + N_{Uf} \]

Guarantee real-time processing and improve the utilization of computing resources.
Dynamic Npre Based Real-time Protocol Processing

- Calculate Npre according to K
- Process DL subframes consecutively
- Avoid the impact from UL processing

Decide the ratio of UL Load and DL load K

Physical layer subframe configuration

Protocol processing subframe cycle

Case 1: Continuous processing
- Downlink load: Uplink load = 3:7
- Npre = 2

Case 2: Continuously processing
- Downlink load: Uplink load = 1:1
- Npre = 3
Simulation Configurations

Experimental Setting

- Protocol processor: Intel Xeon CPU, 2.9Ghz, 8-core, memories 16G
- OS: Linux2.6.32
- CPU Task Scheduling: CFS (Completely Fair Scheduler)
- Communication Protocol: TDD-LTE protocol developed by ICT
Simulation Results-1

Improve the efficiency of computing resource

- Based on dynamic Npre, the proposed scheme provides higher computing resource efficiency than those with fixed Npre.
The proposed scheme can ensure the real time processing of DL data
Conclusions

- The Super BS based architecture is promising for future 5G to reduce energy consumption, decrease construct and maintaining cost and enable various advanced techniques.

- To develop a highly efficient Super BS, communication and computing should be well converged. Initial investigations have been carried out in this direction.
  - Load diversity aware
  - Real time processing guaranteed

- Many more problems to be solved: virtualization of base band processing, impact of fronthaul, etc.
Thank you!