RESOURCE MANAGEMENT STRATEGIES FOR SDR CLOUDS

Vuk Marojevic, Ismael Gomez, Pere L. Gilabert, Gabriel Montoro, Antoni Gelonch
(Universitat Politecnica de Catalunya, Barcelona, Spain;
marojevic|ismael.gomez|plgilabert|montoro|antoni@tsc.upc.edu)

ABSTRACT

This paper analyzes different resource management strategies for SDR base stations implemented as SDR clouds. SDR clouds describe distributed antennas that connect to a data center for digital signal processing. The data center employs cloud computing technology, providing a virtualized computing resource pool. The service area of a single SDR cloud may be a medium sized metropolitan area with a high user density. Hence, the data center will execute thousands of SDR applications—waveforms associated to different radio cells—in parallel. Whenever a user initiates or terminates a wireless communications session, computing resources need to be allocated or deallocated in real time. We therefore propose a hierarchical resource management approach and present simulation results that demonstrate its feasibility.

1. INTRODUCTION

Today's base stations are equipped with a set of heterogeneous processing devices, including application-specific integrated circuits (ASICs), general-purpose processors (GPPs), digital signal processors (DSPs), and field-programmable gate arrays (FPGA). Each device typically executes the tasks that were specified at design time. The ongoing advances in radio engineering and digital signal processing however suggest the deployment of distributed processor arrays [1]. Automatic resource allocation algorithms may then allocate the required computing resources on demand. The approach of considering a data center as the computing core of a base station is then a natural evolution of wireless communications [2].

The digitalization of wireless communications and emerging SDR frameworks facilitates the reconfiguration of radio equipment [3]. This enables dynamic service and RAT adaptations as a function of environmental conditions and user preferences. Future SDR base stations will be correspondingly reconfigured, but on a much larger scale. An SDR base station will than need to concurrently run and manage many waveforms, serving a large number of users. (A waveform or SDR application defines the digital processing chain of a single user transmitter, receiver, or transceiver.) This paper analyzes different resource management strategies for SDR base stations implemented as SDR clouds. An SDR cloud describes distributed antennas that connect to a data center, which performs the digital signal processing tasks associated to all users within the coverage area of the SDR cloud. The data center will thus execute thousands of SDR applications in parallel.

SDR clouds provide a scalable solution for the evolution of wireless communications. Infrastructure resources may be shared among radio operators. Infrastructure or cloud operators may provide computing resources on demand to radio operators and, eventually, the end users. Users penetrate different radio cells and request different services at different times. Computing resources thus need to be dynamically allocated. Whenever a user or pair of users initiates a new wireless communications session, for instance, computing resources need to be allocated in the data center. Resource management is thus essential for SDR clouds.

SDR clouds are very large-scale systems with thousands of processors serving thousands of users. The tight timing constraints of wireless communications require developing efficient computing resource management approaches. Managing the entire data center as one large resource pool is computationally inefficient or impossible as we will demonstrate later. We suggest a two-level resource management approach: clustering and mapping. A computing cluster stands for a logical union of a determined number of processors. These clusters may be assigned to different radio operators, cells, or services. Mapping refers to the real-time assignment of computing resource (processors, interprocessor communication bandwidth, memory, and so forth) of a cluster or group of clusters to SDR applications. This paper analyzes different clustering approaches, that is, resource management strategies.

The rest of the paper discusses some resource management implications of SDR clouds. We then show that employing a two-level management approach considerably reduces the management complexity, enabling real-time computing resource allocation. We propose three resource management strategies and simulate four scenarios of

different communication patterns (user distributions, operator market shares, and so forth). The results indicate that the resource management strategy should dynamically adapt to the communication pattern for an effective SDR cloud resource usage.

2. RESOURCE MANAGEMENT IMPLICATIONS

Wireless communications services have tight real-time constraints, typically specified as the minimum throughput and maximum latency. The data throughput will increase with the deployment of 4G systems, with tens of Mbps on the horizon. A higher data throughput means that more data needs to be processed per time interval. Many signal processing algorithms performing complex computing operations on a continuous data stream. Combined with high data rates, these algorithms consume a considerable amount of computing resources. The computing complexity of modern physical layer signal processing chains is in the order of 10-100 giga-operations per second (GOPS) per user [4] [5].

The end-to-end latency constraint of real-time services is specified for different services and radio standards. LTE specifies 10 ms as the end-to-end latency figure. In practice, however, 30-60 ms are tolerable for the voice service. The popularity of VoIP has, moreover, shown that users are willing to accept even longer delays. Other services (video, streaming) have more relaxed latency constraints. Note that services with tight latency constraints have relatively low throughput requirements.

The range or capacity of an SDR cloud data center will be limited by the latency constraint of wireless communications systems or services. The end-to-end latency here consists of the processing latency of the transmitter and receiver and the two-way transmission delay between the base station antenna and the data center. Since the physical layer (PHY) digital signal processing is a hard real-time SDR computing challenge, we suggest limiting the optical fiber transmission latency to less than 1% of the PHY processing latency. Assuming an end-to-end latency of 40 ms, 10 ms may be available for the PHY processing of the transmitter and another 10 ms for the PHY processing of the receiver. Then, the optical fiber transmission latency may be limited to 0.1 ms. Signals travel through optical fiber at a speed of approximately $2 \cdot 10^8$ m/s. Optical switches introduce negligible delays, which are in the order of 100 ns. A signal path of 20 km length through an optical fiber network would then have a latency of approximately 0.1 ms. A data center providing the computing resources for wireless communications services in a radius of 10 km would fit in this model. The 314 km^2 may cover an entire city with millions of inhabitants. (Barcelona, for instance, covers an area of 100 km^2 and has 1.6 million registered inhabitants.) In metropolitan areas of that size we may have a million wireless subscribers.

Wireless communications sessions are independently initiated and terminated. Users penetrate several geographical zones each day and require very different services. Statistical measurements have shown that an average user establishes 7 or 8 communication sessions per day of some 90 s duration in the mean [6]. At peak we may have some 20,000 wireless communications sessions in parallel (2 % of one million subscribers). If each session requires 10 GOPS for the PHY processing, 200,000 GOPS would be needed for the digital signal processing.

By the time when SDR clouds will be deployed, several million GOPS will be needed for serving the user demands. Only a very large array of multicore devices can provide this capacity. The processing capacity of a graphics processing unit (GPU), for instance, is in the order of one tera-floating point operations per second (TFLOPS) [7].

The SDR cloud resource manager hence needs to be able to dispatch a very large number of user sessions. The stringent timing constraint for assigning computing resources to waveforms, loaded during the session establishment period, moreover requires a very efficient allocation process. We, therefore, suggest a hierarchical resource management.

3. HIERARCHICAL RESOURCE MANAGEMENT

Computing resource allocation and management is a well-known problem in heterogeneous computing [8]. Real-time processing of continuous data flows needs to be ensured in the SDR context. The essentially directed data flow of the physical layer processing chains of radio transmitters and receivers facilitates a pipelined waveform processing. This greatly simplifies the scheduling process [4], which determines the execution order of software processes.

Mapping is the process of distributing the computing load among the available computing resources. The general mapping problem is known to be NP hard [9]. This means that sub-optimal algorithms or heuristics need to be applied in practice. The t_w-mapping is such a heuristic. It provides the necessary flexibility for SDR computing resource management [4]. It implicitly considers the waveform's timing constraints, employ *million operations per second* (MOPS) and *mega-bits per second* (Mbps) as the metrics for characterizing the processing and interprocessor bandwidth resources and the processing and data flow requirements.

The complexity order of the t_w-mapping is $O(M \cdot N^{w+1})$, where M represents the number of processes (software blocks or modules in the processing chain) and N the number of processors. The algorithm complexity thus grows non-linearly with the number of processors; it is proportional to N^2 for $w = 1$. Table I shows some execution

Table I Execution times in seconds of the t_w-mapping process for different window sizes w and number of processors N as measured on an 2.67 GHz i7 Quadcore.

N	w		
	1	2	3
20	0.005	0.09	1.57
30	0.025	0.61	16.23
40	0.075	2.43	87.77
50	0.17	7.2	326.4
100	2.9	221.3	-
200	68.6	-	-
300	329.2	-	-

time measurements of a t_w-mapping implementation as a function of N and w for $M = 24$ processes.

The initiation of a wireless communications session requires the mapping of the corresponding waveform or waveforms to the available computing resources of the SDR cloud. The session establishment time is constrained to a fraction of a second. Hence, it is inappropriate to apply the t_w-mapping to large processor arrays.

We may, on the other hand, consider a greedy mapping algorithm, whose complexity grows linearly with the number of processors. Apart from the inferior performance, the execution time of the g_1-mapping [4] still becomes impractical for hundreds of processors. (We measured an execution time of 0.75 s for $N = 300$ processors.)

A central mapping engine is inefficient also from an SDR cloud implementation perspective. First of all, a single-user receiver or transmitter processing chain, probably containing 10-30 processing blocks, can be executed on a few processors (possibly consisting of several pico-processors). Moreover, many user sessions will be initiated practically simultaneously. These users are likely to be located in different parts of the SDR cloud service area, that is, in different radio cells. A distributed computing resource management approach should then be considered.

We suggest dividing the data center in clusters of a few processors. A high-level resource manager assigns users to clusters as a function of radio communications and cloud computing conditions. The low-level resource managers can then concurrently allocate and deallocate computing resources in real-time. That is, whenever a user initiates or terminates a wireless communications session, the corresponding low-level resource manager allocates or frees cluster computing resources. The maximum number of processors that are controlled by a low-level resource manager is then limited by the execution time of the mapping algorithm, the user arrival rate, and the timing constraint for the session establishment. Since user sessions are independently initiated, cluster resources should be assigned on a first-come first-serve basis.

4. RESOURCE MANAGEMENT STRATEGIES

The high-level resource manager specifies the resource management strategy. Here we suggest three simple strategies and analyze their performances in the Section 5.

4.1. Strategy 1: Operator Clusters

The data center resources are grouped in clusters and assigned to different radio operators. Radio operators may, for example, demand a certain number of clusters based on statistical measurements and the expected system load. The data center capacity should be scaled as a function of these measurements. Nevertheless, a proper management and accounting of actually used computing resources is necessary. Contracts between the cloud operator and the radio operators may, for instance, establish a minimum amount of guaranteed computing resources. The remaining resource pool should be fairly shared between the different parties. If conflicts or resource bottlenecks arise repeatedly, the data center should be upgraded.

Since only a few radio operators may initially exist, dividing the data center resources between radio operators may be insufficient for an efficient low-level resource management. This strategy may then be combined with another, such as strategy 2.

4.2. Strategy 2: Cell Clusters

The second strategy is based on assigning radio cells to computing clusters. Radio cells will be of different sizes and traffic characteristics. The assignment of cells to clusters should then be chosen as a function of the number of expected active users in a cell and their processing demands. Since the load varies over time, the assignment needs to be dynamically adjusted.

Strategy 2 seems to reflect the actual communications infrastructure. Nevertheless, the SDR cloud facilitates a dynamic refinement of clusters or cluster sizes assigned to the different radio cells. This strategy may simplify the data center architecture and access to the fiber optical communication network. Configurable switches enable joining clusters or creating different cluster sizes.

4.3. Strategy 3: Service Clusters

Strategy 3 assigns clusters to wireless communications services. Each service has more or less stringent timing and computing constraints. Voice services, have a low data throughput, but a very tight end-to-end latency requirement.

A higher processing throughput can be achieved employing parallel processing units, where distributed computing resources jointly process a waveform or set of waveforms. Latency control, on the other hand, requires

more sophisticated resource management (mapping and scheduling) solutions. The fiber optical communication network and the data movement within the data center also introduce delays and, therefore, need to be considered. Latency-constrained services may then be assigned to those computing resources that have lower data flow delays. This strategy, moreover, facilitates applying service-dependent resource optimization goals. Rather than minimizing the processing delay, other optimization objectives may be defined for the vast amount of data-intensive services with fewer timing restrictions.

This strategy may be combined with the first, for instance. Clusters may then be assigned to the different operators as a function of the services they offer.

5. SIMULATIONS

We simulate different scenarios for analyzing the performance of the proposed strategies. We therefore consider a reduced service area of, for example, 1.6 x 1.6 km^2 with distributed antennas providing full service coverage. This area is divided into sixty-four radio cells of 200 x 200 m^2 each. We neglect inter-cell interference and radio resource management issues. The distributed antennas connect to the data center via an optical fiber network.

The data center features 256 quad-cores, or 1024 processors. Each core has a computing capacity of 12 GOPS. Since we do not analyze the implications of the data center communication architecture here, we may assume that each quad-cores connects to a central logical switch of 40 Gbps ports. This switch enables full connectivity between the processors and also serves as the access point to the optical fiber network.

5.1. Scenarios and Strategies

Two 3G service providers offer voice and data services at different data rates: 64 kbps (voice), 128 kbps, 384 kbps, and 1024 kbps. The four SDR applications correspond to the chip- and bit-rate digital signal processing chain of the UMTS receiver model [4]. The corresponding transmitters are not considered here. We, rather, assume that they are allocated to additional computing resources.

We define four scenarios. Scenario I considers a homogeneous market share, that is, 50 % of the users are associated with each operator. The voice versus data service demand is also balanced, that is, 50 % voice and 50 % data. The exact probabilities are 0.5, 0.2, 0.2, and 0.1 for 64, 128, 384, and 1024 kbps. Users are uniformly distributed across the entire services area.

Scenario II is similar to I except for the market share, which is 75 % for operator 1 and 25 % for operator 2. Scenario III is characterized by 25 % voice and 75 % data traffic. More precisely, the user demand 64, 128, 384 and 1024 kbps with equal probability. A Gaussian distribution of user locations finally defines scenario IV. Table II summarizes the main characteristics of these scenarios.

Table II Simulation scenarios.

	OP 1	OP 2	64 kbps *voice*	128 kbps *data*	384 kbps *data*	1024 kbps *data*	User distr.
I	50 %	50 %	50 %	20 %	20 %	10 %	Uni-form
II	75 %	25 %	50 %	20 %	20 %	10 %	Uni-form
III	50 %	50 %	25 %	25 %	25 %	25 %	Uni-form
IV	50 %	50 %	50 %	20 %	20 %	10 %	Gaus-sian

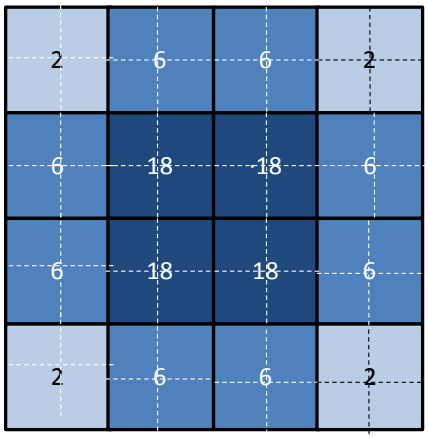


Fig. 1 Service area divided into 4x4 zones, each consisting of 2x2 radio cells. The amount of computing resources assigned to each cell is a function of the strategy. The given numbers correspond to the number of clusters (quad-cores) per operator due to strategy 2.b (Section 5.2.4).

We apply the three strategies of Section 4. Each strategy has two implementations: (a) uniform clusters assignment and (b) cluster assignment adapted to the expected demand. Figure 1 shows the service area divided into 16 zones. Each zone is further divided into 4 radio cells. The numbers inside each zone indicate the number of allocated clusters in case of strategy 2.b. More details about the strategies are provided later.

5.2. Results and Analysis

Five thousand users enter the system one after another. Each user tries initiating a wireless communications session. The operator, service, and location characteristics of each user are random, assuming the probabilities of Table II. Users are assigned to computing clusters as a function of the resource

management strategy. A user is accepted if and only if the t_1-mapping allocates enough computing resources for the real-time execution of the corresponding SDR application. No user exits the system.

5.2.1. Scenario I

Here we employ strategies 1.a, 2.a. and 3.a. Strategy 1.a assigns four clusters to each radio cell, two per operator. There are 64 cells in the system, so that each operator is assigned a total of 128 clusters.

Strategies 2.a and 2.b divide the service area into 16 zones with 4 cells in each zone (Fig. 1). Sixteen clusters are allocated to each zone, eight per operator. In case of strategy 2.a, the computing resources of eight clusters are shared among the users of one operator and zone.

Strategy 3.a allocates clusters as a function of service. It dedicates 50 % of the clusters to voice (64 kbps) and 50 % to data services (128, 384, 1024 kbps). Four clusters are shared among the users in one zone obtaining the same service type (voice or data) from one operator.

Figure 2 shows the allocation results of the three strategies. These strategies perform very similar. The reason for this is that they are all adapted to the given, homogenous traffic demand. Strategy 2.a balances the possible non-uniformities in the user distribution of four neighboring cells and, therefore, reaches the maximum processor utilization earlier. Strategy 3.a performs slightly worse because the data services consume more computing resources than the voice service.

5.2.2. Scenario II

In the second scenario, we employ strategy 1.a and strategy 1.b. Strategy 1.a is the same as above. Strategy 1.b assigns three clusters to radio operator 1 and one cluster to operator 2 for each cell. This leads to a total of 192 and 64 clusters allocated to operators 1 and 2, respectively.

Figure 3 shows the results. Strategy 1.b better adapts to the actual resource demand. It is therefore capable of serving more users at reasonable system loads. Since users do not exit, that is, remain active over the entire simulation time span, both curves saturate at the same data center resource occupation level of approximately 87 % (Fig. 3.b).

5.2.3. Scenario III

In Scenario III we compare strategies 3.a and 3.b. Strategy 3.a is the same as in Section 5.2.1. For each radio cell, strategy 3.b allocates one cluster for processing voice services and 1 cluster for each of the three data services. Users are thus assigned to clusters as a function of their service request, rather than their operator affiliation. Again, there are 4 clusters per cell, making a total of 256 clusters.

The results of Fig. 4 show that strategy 3.b outperforms 3.a at medium system load (1000-3500 active users). Strategy 3.b better adapts to the given service distribution. However, higher data rate services require more computing resources than lower data rate services [4]. Clusters are assigned according to service probability and not the actual resource demand. The clusters resources providing high data rate services (1024 kbps) begin to saturate earlier for strategy 3.b than for strategy 3.a, where the data services share two clusters per cell. This sharing explains the slightly higher maximum resource utilization of strategy 3.a (Fig. 4, 4000-5000 users).

5.2.4. Scenario IV

In this scenario user are not uniformly distributed over the service area, but rather follow a Gaussian distribution. The mean of this distribution is the center of the service area with as standard deviation of 250 m. We define a new strategy, strategy 2.b, which allocates resources as a function of the expected communication traffic in each zone. The allocation of clusters to zones is indicated in Fig. 1: 4x18 clusters are dedicated to the four central zones, 4x2 clusters to the four edge zones and 8x6 clusters to the remaining lateral zones. This makes a total of 128 clusters dedicated radio operator 1. Another 4x18 + 4x2 + 8x6 clusters are allocated the same way to operator 2. The *x* clusters in any zone are shared among all users associated with one operator. That is, 18 clusters are shared in a central zone, 6 in a lateral zone, and 2 in an edge zone.

Strategy 2.a is the same as in Section 5.3.1. It assigns 16 clusters to each zone, 8 per operator. The computing resources of 8 clusters are thus shared among a user group.

Figure 5 shows the results. We observe that strategy 2.b, which is better matched to the non-uniform spatial distribution of users, accepts more users. The cluster assignment does not closely match the user distribution, because of the limited granularity of 16 zones. If there were a sufficiently large number of very small zones, the distribution of clusters could be almost perfectly adapted to the user distribution in theory. In practice, however, the exact user distribution will not be known in advance and the data center communication network will neither allow defining any clustering combination.

6. CONCLUSIONS AND FUTURE WORK

This paper has discusses the resource management implications of SDR clouds and proposed as two-level resource management approach. The high-level resource manager pre-assigns computing clusters to SDR applications as a function of the user location (radio cell), operators, or services, for instance. It defines the resource management strategy. The distributed low-level resource managers allocate and deallocate computing resources on demand and in real-time, whenever a user initiates or terminates a wireless communications session. This approach seems

suitable for SDR clouds as it facilitates efficiently managing the large-scale computing system.

We have introduced three resource management strategies and analyzed their performance in different simulation scenarios. The results highlighted the importance of adapting the strategy to the given wireless communications traffic distribution. Adaptive resource management strategies can maximize the resource usage in each moment for serving as many users as possible with the available computing resources. Although we have tried to isolate the analysis of the different strategies here, a combination of strategies may be more flexible and better adapt to the heterogeneous user demands in practical systems.

This work omitted the analysis of different data center communication architectures and bandwidths. Both, architecture and bandwidths will have implications on the definition and performance of resource management strategies. Data locality is an important issue for large-scale computing system and we will deal with it in future work.

ACKNOWLEDGMENT

This work was supported by Spanish Government (MICINN) and FEDER under project TEC2008-06684-C03-03.

REFERENCES

[1] S. Zhou, M. Zhao, X. Xu, J. Wang, Y. Yao, "Distributed wireless communication system: a new architecture for future public wireless access," *IEEE Commun. Mag.*, vol. 41, iss. 3, pp. 108-113, March 2003.

[2] Y. Lin, L. Shao, Z. Zhu, Q. Wang, R. K. Sabhikhi, "Wireless network cloud: architecture and system requirements," IBM J. Res. & Dev., vol. 54, no. 1, paper 4, Jan./Feb. 2010.

[3] I. Gomez, V. Marojevic, A. Gelonch, "ALOE: an open-source SDR execution environment with cognitive computing resource management capabilities, *IEEE Commun. Mag.*, vol. 49, iss. 9, pp. 76-83, Sept. 2011.

[4] V. Marojevic, "Computing resource management in software-defined and cognitive radios," Ph.D. Dissertation, Universitat Politècica de Catalunya (UPC), Barcelona, July 2009. Available at http://flexnets.upc.edu/trac/wiki/Publications

[5] M. Woh, Sangwon Seo, S. Mahlke, T. Mudge, C. Chakrabarti, K. Flautner, "AnySP: anytime anywhere anyway signal processing," *IEEE Micro*, vol. 30, iss. 1, pp. 81-91, Jan./Feb. 2010.

[6] Junqiang Guo, Fasheng Liu, Zhiqiang Zhu. "Estimate the call duration distribution parameters in GSM system based on K-L divergence method," *Proc. IEEE Int. Conf. Wireless Communications, Networking and Mobile Computing (WiCom 2007)*, Shanghai, 21-25 Sept. 2007, pp. 2988-2991.

[7] J. Kim, S. Hyeon, S. Choi, "Implementation of an SDR system using graphics processing unit," *IEEE Commun. Mag.*, March 2010.

[8] A. Khokhar, V. K. Prasanna, M. Shaaban, C. L. Wang, "Heterogeneous computing: challenges and opportunities," *IEEE Computer*, vol. 26, iss. 6, pp. 18-27, June 1993.

[9] S. H. Bokhari, "On the mapping problem," *IEEE Trans. Comput.*, vol. C-30, no. 3, pp. 207-214, March 1981.

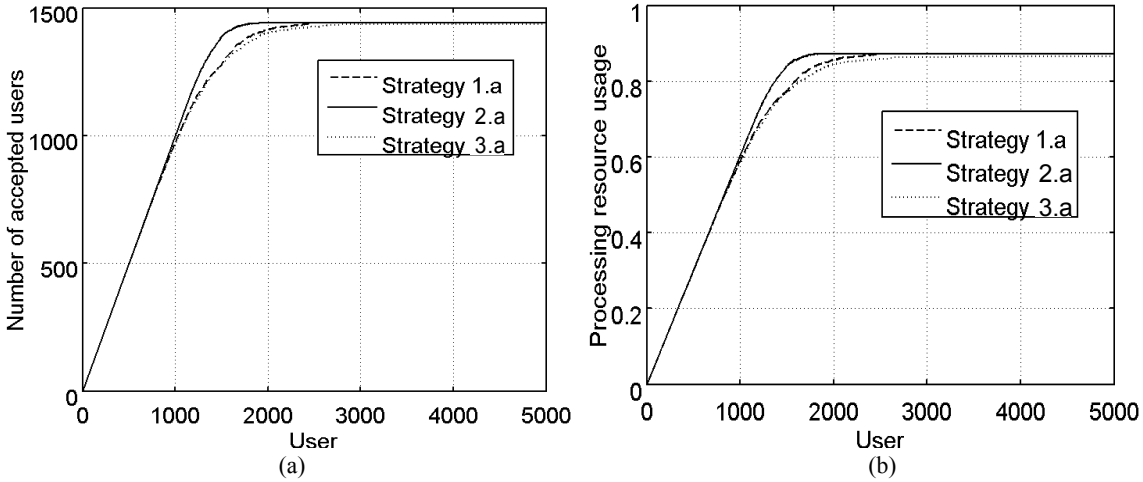
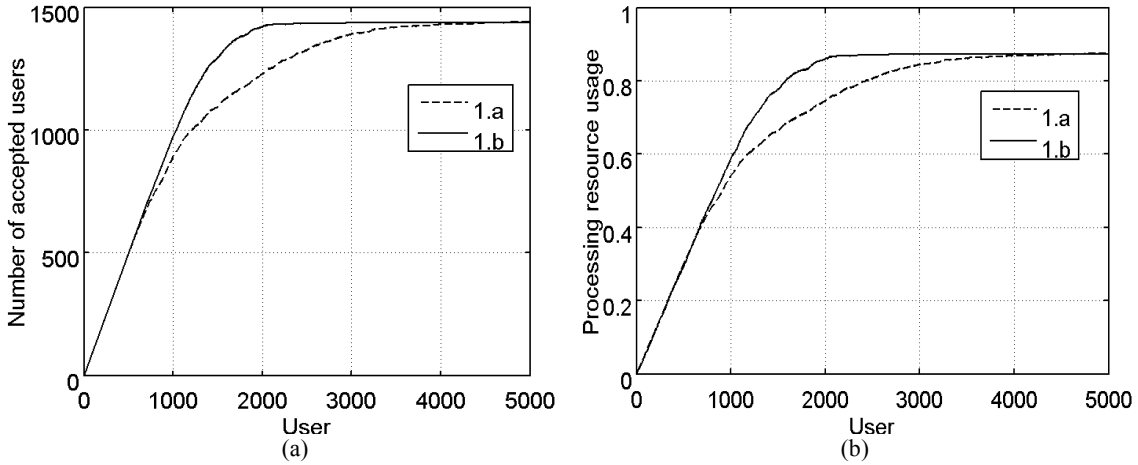
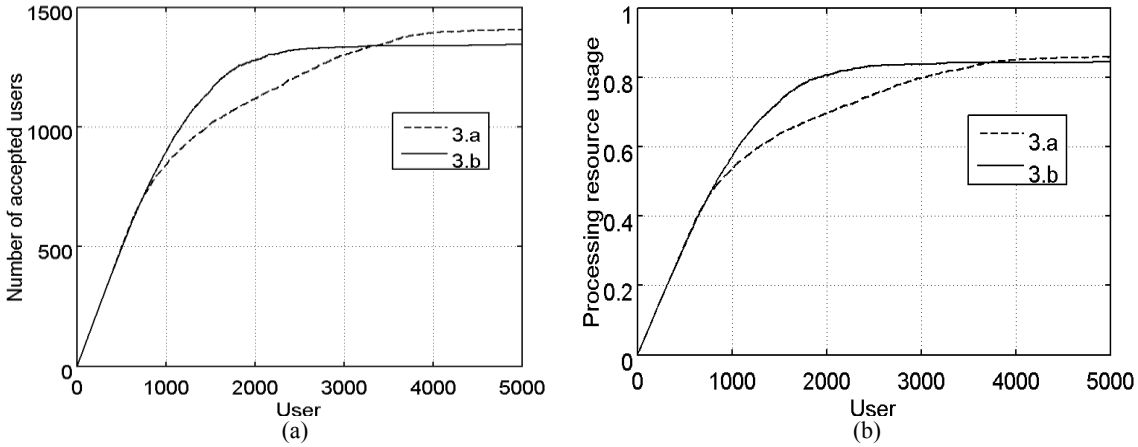


Fig. 2. Accepted users (a) and processor resource occupation (b) for scenario I.



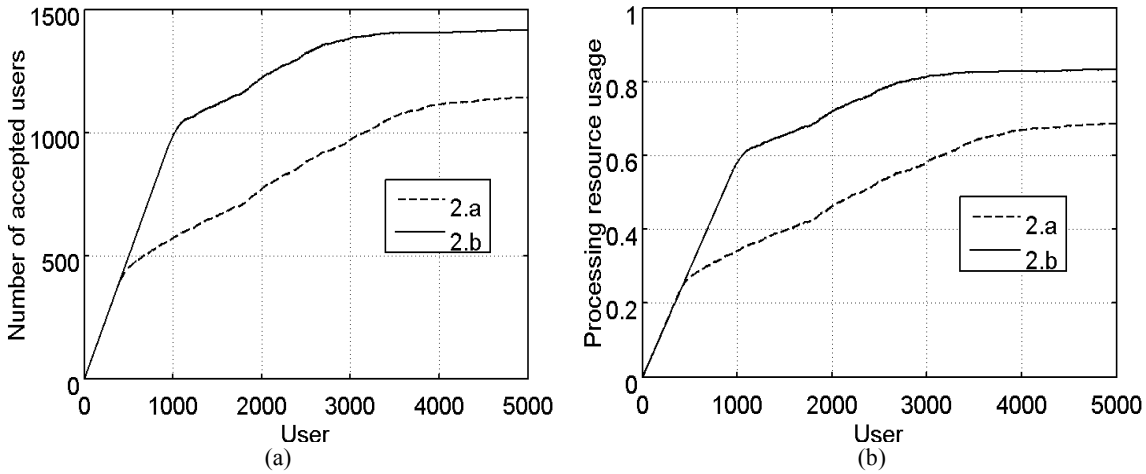
(a) (b)

Fig. 3. Accepted users and processor resource occupation for scenario II.



(a) (b)

Fig. 4. Accepted users and processor resource occupation for scenario III.



(a) (b)

Fig. 5. Accepted users and processor resource occupation for scenario IV.