Virtual machine resource allocation algorithm in cloud environment

Lei Zheng[1, 2*]

[1]*School of Information Engineering, Shandong Youth University of Political Science, Jinan 250103, Shandong, China*

[2]*Key Laboratory of Information Security and Intelligent Control in Universities of Shandong, Jinan 250103, Shandong, China*

Received 1 July 2014, www.cmnt.lv

Abstract

To resolve the problem that virtual machine deployment reservation scheme waste a lot of resources and single-objective deployment algorithm is not comprehensive, a virtual machine resource allocation algorithm based on virtual machines group multi-objective genetic algorithm is proposed. The algorithm is divided into group coding and resources coding. Resources coding integrated coding according to the history resource need of virtual machines to physical machine and integrate number of physical machine and resource need of physical machine occupied by virtual machine through improved crossover and mutation operations. The experimental results show that the algorithm is effective to reduce the number of physical machine and resource utilization of physical machine, saving energy as much as possible.

Keywords: Cloud computing, Resource allocation, Virtualization, Energy-saving, Genetic algorithms

1 Introduction

Cloud computing is a type of new computing model, providing all kinds of serviced through Internet. Users can gain access to the cloud service anytime, anywhere and on any device. Virtualization technology plays a critical role in management of cloud resource and dynamic configuration, for all kinds of underlying hardware resources can be encapsulated by virtualization technology and provides services to users with virtual machines as the basic resource unit [1]. However, since cloud computing platform includes highly dynamic and heterogeneous resources, virtual machines has to adapt to the dynamic cloud computing environment [2]. The purpose of virtual machines deployment strategy is realizing ideal result by changing layout and placement of overall virtual machines to optimize the objective in meeting constraint condition. The problem of virtual machines placement proved to be *NP* problem. Thus, it is a research hotspot in current cloud computing filed how to conduct virtual machines placement effectively.

Currently, single-objective resource allocation and deployment method are usually adopted in the field of virtualized server technology application. With the goal of maximizing utility, the literature [3] utilizes *NUM* model in computer network. One or more physical machines resources, like bandwidth of network link, distributes one or more job by virtualization technology, reaching higher level of allocation of computing resources of physical machines and optimizing it by algorithm. With the goal of energy-saving, the literature [4] dynamically deploys virtual machines application by energy-aware heuristic algorithm. The literature [5] put forward an improved preferential cooperation descending method to solve the problem of node bin packing. The method merely involves node integration during peak load situation, without consideration of constraint that the goods may be incompatible with the box. The literature [6] suggests an adaptable management frame for virtual machines placement, and studies on the solution of genetic algorithm to the overall placement of virtual machines, effectively reduce the number of physical machines and migration times, but not considering the integration of physical machines resources by virtual machines in the solving process. Nowadays, the optimal method of most virtual machines placement is transforming multi-objective optimization problem to several single-objective optimization problems to be solved in stages. It rarely happens that multi-objective is optimized at the same time. In most time, only partial optimal solution rather than global optimal solution is gained.

For the issue of server integration and resource allocation in cloud computing, a virtual machines resource allocation algorithm based on virtual machines multi-objective genetic algorithm is proposed. With the aim of reduce the number of physical machines and resource allocation, a best solution is searched by genetic algorithm to saving energy as much as possible.

2 Resource allocation algorithm in cloud computing

The definition of bin packing problem [7] is that a set S of M in size and a set P of N in size given, how all

* ***Corresponding author*** e-mail: 1292815809@qq.com

elements of *S* are packed in elements of *P* with least elements of *P* used. *BPP* problem, a difficult *NP* problem, cannot be done by a known optimal algorithm in polynomial time. The problem of virtual machines deployment is actually bin packing problem. In cloud computing, how to reasonably deploy virtual machines to relevant nodes shall be considered, realizing optimal usage of resources while meeting service objectives of different applications. The virtual machines placement may be regarded as vector bin packing problem. The goods being packed are the virtual machine under operation, and the resources of virtual machine are the changeable size of goods. The box is the physical node, and the capacity of the box is the usage threshold of node resource. The number of types of resources is the number of dimensions of vector bin packing problem. Assuming that the number of physical nodes is *M* and the number of virtual machines is *N*, the solution space from the virtual machines to the physical nodes is M^N . It is a *NP* problem similar with bin packing problem that requires an approximate optimal solution.

2.1 DESCRIPTION OF ISSUES IN MULTI-OBJECTIVE VIRTUAL MACHINES DEPLOYMENT ALGORITHM

The resource asked by users to the cloud platform is equal to s virtual machine requiring specific resource, and the applications package of each user operates on their own virtual machines. It is an academic research hotspot how to save energy and utilize cloud computing resource as much as possible to deploy multi-objective virtual machines. Deploying multi-objective virtual machine problem is a multiple combination optimization problem, as well as multi-objective optimization problem. The available resource of each physical machine is multi-dimensional vector, with each dimension as one of all resources of physical machines, and the resource needed by each physical machine is also multi-dimensional vector. The objective is to allocate several virtual machines to several physical machines, maximizing each resource utilization rate of physical machines and minimizing the number of virtual machine immigration. The multi-objective deployment problem is described as follows:

Make N_{PM} as the physical machine set in cloud computing, N_{VM} as virtual machine set in cloud computing, *N* as the total number of virtual machines, N_R as the available allocated resources set in cloud computing and *K* as the total number of available allocated resources.

The objective: $max\sum_{m=1}^{M}\sum_{k=1}^{K}U_{m,k}$ and $min\sum_{m=1}^{M}P_m$.

$\forall n \in N_{VM}, \forall m \in N_{PM}$, among them, $U_{m,k}$ is usage rate of the *K* type of resource by physical machine *m*, P_m is the number of nodes of physical machines.

Constraint:

$$P_m \in \{0,1\} . \quad (1)$$

If $P_m = 1$, it means using new physical machine.

$$U_{m,k} < C_{m,k}, \quad (2)$$

$$\sum_n U_{n,k}^{\ m} < C_{m,k} . \quad (3)$$

Among them, $C_{m,k}$ means the threshold value of the *K* type of resource of *m* physical machine. $U_{n,k}^{\ m}$ is the usage rate of the *K* type of resource by the *n* virtual machine under *m* physical machine. The operation of each type of resource of each physical machine shall be less than the threshold value of each type of resource during the virtual machine deployment process. When there are several virtual machines under deployment in *m* physical machine, the total usage rate of resources of virtual machine under physical machine shall be less than the threshold value of each type of resource.

2.2 DESIGN AND REALIZATION OF MULTI-OBJECTIVE VIRTUAL MACHINES DEPLOYMENT ALGORITHM

For the huge cloud computing centre, combinational explosion may occur in combinatorial optimization. Genetic algorithm is one of the methods for solving combination problem now, since it can concurrently handle with all objectives and avoid priority ordering among objectives. Therefore, genetic algorithm is very suitable for solving multi-objective optimal issues [8]. A virtual machine resource allocation algorithm based on virtual machines multi-objective genetic algorithm is proposed, on the basis of multi-objective virtual machines deployment problem in cloud computing centre.

2.2.1 Coding

In the virtual machines deployment problem, there are three types of genetic coding methods: (1) the representation based on box; (2) the representation based on goods; (3) the representation based on group. Since the objective function of bin packaging problem relies on the goods group, the former two coding methods face single goods, with shortcoming of unclear grouping information. The shortcoming of the third coding method is relying on goods group, neglecting the difference of utilization of physical machine resources by each virtual machine in the crossover and mutation process. In the paper, combining with the coding method based on group

and goods, dynamic allocation genetic algorithm of cloud computing resources is proposed.

The coding based on goods in the paper mainly adopts the coding based on the resource need of virtual machine to the physical machine. The resource need of virtual machine to the physical machine taking CPU, disc and network as example, by *N* number of samplings of *i* virtual machine in a while *T*, calculates the number of operations of CPU, disc and network according to the sampling points. Then the number of operations is coded, and the number of operations of CPU, disc and I/O are showed as formulas (4), (5), and (6). In order to understand the change of demand of virtual machine for resources, the author of the paper adopts the energy efficiency models in the literature [9] for data sampling.

$$L_c^i(T)=\sum_{t=1}^{N}\left[C_f^i(t)\times C_u^i(t)\times C_c^i\times C_{fn}\right]. \quad (4)$$

Among them, $C_f^i(t)$ is the CPU frequency of *i* virtual machine at *t* time; $C_u^i(t)$ is the usage rate of CPU of *i* virtual machine at *t* time; C_c^i is the number of CPU cores of *i* virtual machine; C_{fn} is the number of calculation of floating point in each period.

$$L_d^i(T)=\sum_{t=1}^{N}\left[D_r^i(t)+D_w^i(t)\right]. \quad (5)$$

Among them, $D_r^i(t)$ is the data amount read from disc of *i* virtual machine per second at *t* time; and $D_w^i(t)$ *i*s the data amount written to disc of *i* virtual machine per second at *t* time.

$$L_n^i(T)=\frac{1}{2}\sum_{t=1}^{N}\left[N_r^i(t)+N_w^i(t)\right]. \quad (6)$$

Among them, $N_r^i(t)$ is the data amount received by network card of *i* virtual machine per second at *t* time; and $N_w^i(t)$ is the data amount sent by network card of *i* virtual machine per second at *t* time.

Then calculate the probability of the calculation amount of three types of resources needed by each virtual machine in the calculation amount of physical machine recourses and conduct normalization processing. The formula of probability of the calculation amount is as the following formula (7).

$$P_i=\frac{\mu_i}{\sum_{i=1}^{Z}\mu_i}. \quad (7)$$

Among them, μ_i is the calculation amount of CPU, disc and I/O of *i* virtual machine; and *Z* is the number of virtual machines in physical machines of *i* virtual machine. P_i may be the CPU calculation probability of CPU, disc and I/O of *i* virtual machine.

Then normalization processing is conducted according to the ratio of probability *H* of in the entire set of all types of resources in current physical machine. The value is in the range of [0, 10]. and the probability distribution is shown in figure 1.

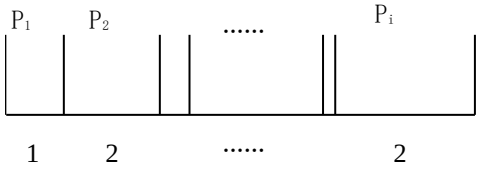

FIGURE 1 Probability distribution graph

Make c_i , m_i , n_i represent the coding of the resource need by CPU, disc and network, and make the proportionality distribution of all types of resources of *i* virtual machine after normalization processing as the resource coding of *i* virtual machine.

Constraint: $c_i\in[0-9], c_i\in N^*$, $m_i\in[0-9], m_i\in N^*$, $n_i\in[0-9], n_i\in N^*$.

Make T_c as the threshold value of CPU resource of physical machine, T_m as the threshold value of disc resource of physical machine and $T_{I/O}$ as threshold value of CPU resource of physical machine. In order to better illustrate the algorithm rose in the paper, the values of T_c, T_m and $T_{I/O}$ in the genetic operation process is 8. The threshold less than 10 is for reserving part of resource space for immigration of virtual machines.

The coding method is shown as figure 2. It mainly takes group as chromosome that has uneven length for the inconsistent number of genes in chromosome. There are three types of deployment of 9 virtual machines: the length of EBA and FCQ are the same but the types of deployment are different; both the length of chromosome of EBA and FCQ as well as the types of deployment are different.

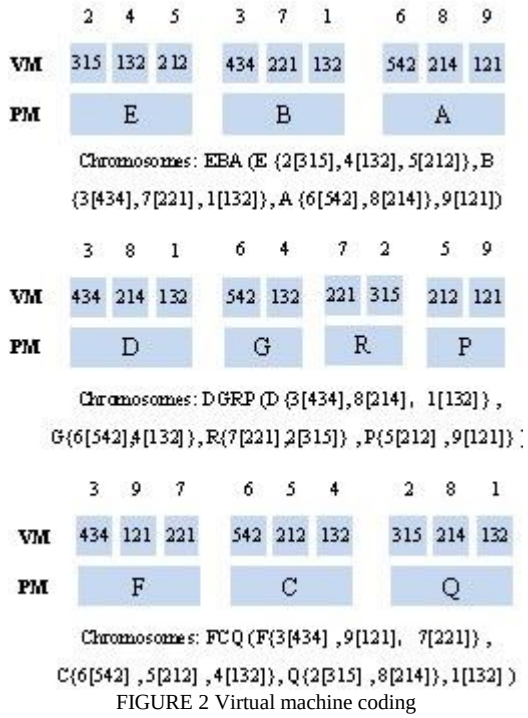

FIGURE 2 Virtual machine coding

2.2.2 Evaluation of fitness

Genetic algorithm evaluates the pros and cons of individuals according to fitness. Fitness function means the corresponding relevance between the whole subjects and their fitness. Evaluation of fitness conducts evaluation of each individual and prepare for the next genetic operation. Since the objective function adopts physical machines as less as possible to place virtual machines as more as possible, the fitness function is shown as the following formula (8), according to the objective function and constraint condition:

$$Fitness_j = \sum_{i=1}^{P_j} \sum_{k=1}^{K} U_{i,k}{}^{j} / P_j . \quad (8)$$

Among them, j means the number of father node, P_j as the total number of physical machines used by father node, and $U_{i,k}$ as the utilization rate of k type of resource of i virtual machine.

2.2.3 Crossover

The main function of the crossover process in genetic algorithm is letting the next generation inherit the excellent genes from the parents and have chance to produce more excellent generations. There are two parts of the crossover process: one is crossover process based on group coding aiming to minimize the number of physical machines, and the other one is crossover process based on resource coding aiming to maximizing the resource of physical machines.

Crossover process based on group coding is shown in figure 3 with steps as follows:

1. Randomly select two father nodes, cross part and cross dot.
2. Insert a selected virtual machine to father node to form new deployed physical machine setoff virtual machines.
3. Delete the repetitive virtual machines in the new physical machine set.

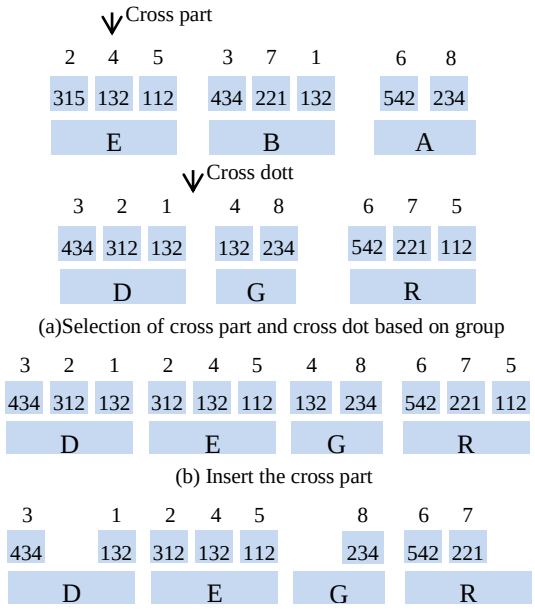

FIGURE 3 Coding cross process based on group

Based on the resource coding crossover process and group coding crossover process, the genetic operation of resource coding can fasten the convergence speed of the group coding genetic process, and integrate the resource of physical machine occupied by the virtual machines. In order to reserve the group crossover result, the resource coding crossover process will reserve the inserted virtual machine group of group coding, not conducting resource coding crossover to it.

Crossover process based on resource coding is shown in figure 4 with steps as follows:

1. Select the remaining virtual machine of the first father node and the second father node (excluding the cross part) of the group coding crossover result as the father nodes. Select the cross part and cross dots. Now the cross part is the virtual machine but not the virtual machine group.
2. Insert the selected virtual machine to virtual machine group.
3. Combine independent virtual machines and delete repetitive physical machine as well as physical machine without any virtual machine.
4. Combine the results of group coding crossover process and resource coding crossover process to get the crossover process result.

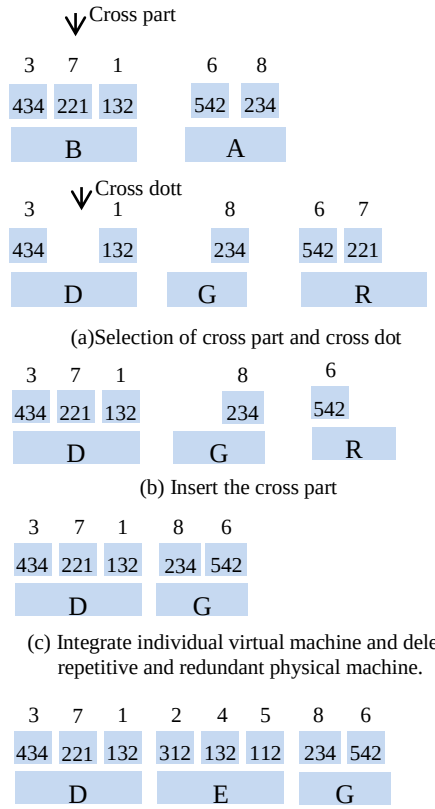

FIGURE 4 Cross process based on resource coding

3 Experimental analysis

In order to verify the proposed algorithm, the author conducts simulation experiment in CloudSim [10]. With the purpose of verifying the effectiveness and deployment scheme, we select the following two classical virtual machine deployment algorithm （Multi-object virtual machines resource allocation Algorithm, MOA） to compare with the multi-objective virtual machine resources distribution algorithm.

Best Fit Algorithm (BFA) means to select the physical machine that meets the resource need of virtual machine with least remaining resource during the virtual machine deployment process, making the physical machine least remaining resource. First Fit Algorithm (FFA) means to search physical machines in order during the virtual machine deployment process, letting virtual machine directly deployed in the physical machine that meets the resource need of virtual machine.

Experiment 1 Calculation of number of physical machines

Deploy 100 virtual machines in 50 physical machines using three types of algorithm independently, with the same nature of physical machines and virtual machine tasks excepting the deployment method and resource threshold value. Among them, the crossover proportionality and mutation proportionality of multi-objective virtual machine deployment algorithm is 0.7 and 0.5 respectively, and the genetic algebra is set as 10. There are load parameter and change of virtual machine resource need during the experiment, taking 10 minute as a time unit to record the change of number of physical machines in 10 time units by three types of algorithm. The experiment result is shown as figure 5:

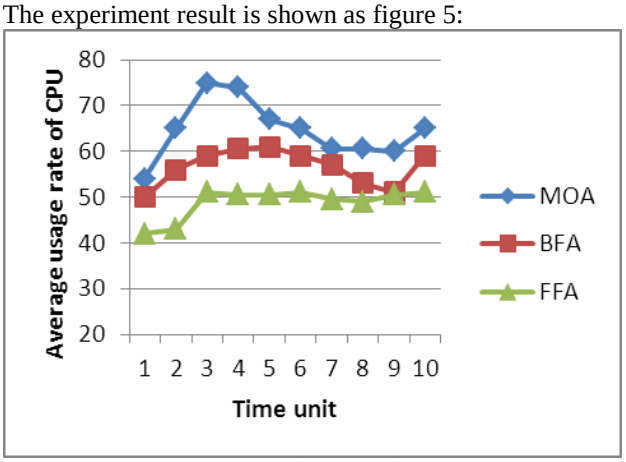

FIGURE 5 Comparisons of number of physical machines

The experiment shows that as time goes on, with the dynamic change of virtual machine's need of resource, the number of physical machines in MOA algorithm is less than that of BFA and FFA. It is because that in a dynamic process, MOA algorithm searches the least number of physical machines in generic operation that meets the constraint condition. It shows that MOA algorithm can effectively reduce the number of physical machines.

Experiment 2 Calculation of resource utilization rate

Calculate the average resource utilization rate of physical machines by three types of algorithm, taking 10 minute as a time unit to record the change of usage rate of CPU and inner storage of physical machines in 10 time units, and calculate the average resource utilization rate. The experiment results of average usage rate of CPU and inner storage by three types of algorithm are shown in figure 6 and 7.

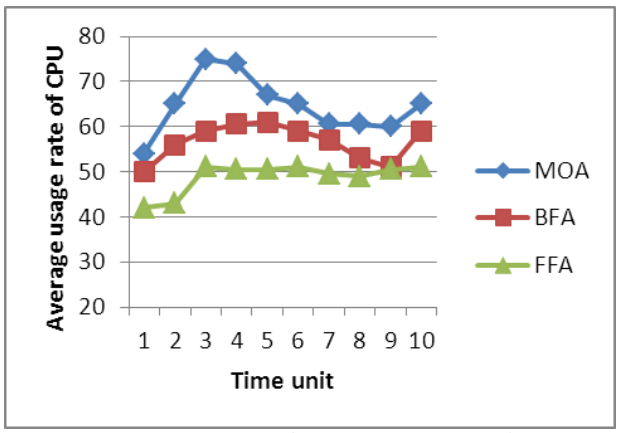

FIGURE 6 Comparisons of the average usage rate of CPU

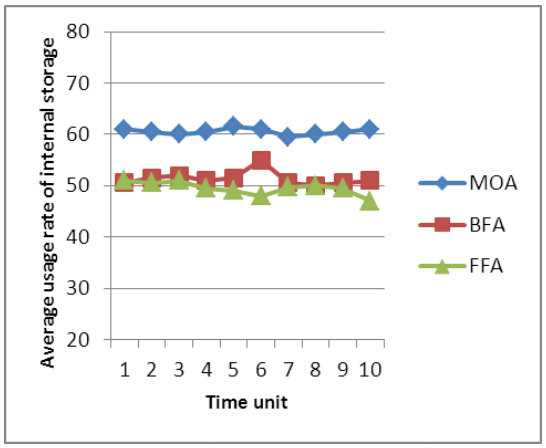

FIGURE 7 Comparisons of the average usage rate of internal storage

It shows from the compared experimental results that the average usage rate of CPU and internal storage of physical machines deployed by MOA algorithm is evidently higher than that of BFA and FFA algorithm. It is because MOA try to improve the resource usage rate as much as possible by using genetic algorithm to adjust virtual machine group during the deployment process, while BFA algorithm try to deploy physical machines as less as possible but not considering improving the resource usage rate, and there are randomness in FFA algorithm that does not consider the resource usage rate. It shows that MOA algorithm can improve the resource usage rate and save energy to some extent.

4 Conclusion

In the paper, after analysing the research situation of virtual machine deployment scheme in cloud computing, the author put forward improved genetic algorithm, conduct group coding and resource need coding of virtual machine, and improve the crossover and mutation operation to resolve the problem of energy waste in cloud computing. Under experimental condition, it shows that the algorithm can not only reduce the number of physical machines but also improve the resource utilization rate. In further research, the issue whether the performances among virtual machines are related will be introduced.

References

[1] Jinhua Hu, Jianhua Gu, Guofei Sun, et al. 2010 A scheduling strategy on load balancing of virtual machine resources in cloud computing environment *Third international symposium on parallel architectures, algorithms and programming* **89**(96) 18-20
[2] Cherkasova L, Gupta D, Vahdat A 2007 When virtual is harder than real: Resource allocation challenges in virtual machine based it environments *Technical Report HPL-2007-25*
[3] Truong H L, Dustdar S 2010 Composable cost estimation and monitoring for computational applications in cloud computing environments *Procedia computer science* **1**(1) 2175-84
[4] Bo Li, Jianxin Li, Jinpeng Huai, et al. 2009 EnaCloud: an energy-saving application live placement approach for cloud computing environments *IEEE international conference on cloud computing* **17**(24) 21-5
[5] Ajiro Y, Tanaka A 2007 Improving packing algorithms for server consolidation *Proceedings of the 33rd International Computer Measurement Group Conference* San Diego 399-406
[6] Qi G, Ji Q, Pan J Z, Du J 2011 Extending description logics with uncertainty reasoning in possibilistic logic *International journal of intelligent systems* **26** 353-81
[7] Aktas H, Cagman N 2007 Soft sets and soft groups *Information sciences* **177** 2726-35
[8] Sun Y L, Perrott R, Harmer T, Cunningham C, Wright P 2010 An SLA focused financial services infrastructure *Proceedings of the 1st International Conference on Cloud Computing Virtualization (CCV 2010)*, Singapore, 2010
[9] Rudolph S 2011 Foundations of description logics In: Polleres, A., D'Amato, C., Arenas, M., Handschuh, S., Kroner,P., Ossowski, S. and PatelSchneider, P.F., Eds. *Reasoning Web. Semantic Technologies for the Web of Data, Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg 76-136
[10] Calheiros R N, Ranjan R, De Rose C A F, et al. 2009 Cloud-Sim: A novel framework for modeling and simulation of cloud computing infrastructures and services *Parkville VIC*: The University of Melbourne Australia, Grid Computing and Distributed Systems Laboratory.

Authors

Lei Zheng, born on August 3, 1980, China

Current position, grades: researcher at Shandong Youth University of Political Science, China.
University studies: master's degree in Computer Software and Theory from Shandong Normal University, China in 2006.
Scientific interests: cloud computing and distributed computing.