Contribution ID: **237** Type: **not specified**

A systematic approach to building cloud special purpose systems

Thursday 6 July 2023 18:15 (15 minutes)

Scheduling tasks and allocating resources in a cloud (distributed) system is significantly different from resource management of a single computer. Cloud (global) schedulers view the system as a large pool of resources to which they have full access. At the same time, the most urgent task remains to maintain a balance, which consists in managing each individual computing task in such a way that the restrictions associated with it are met, while the total load of the system would meet the requirements of its owner (complete load, return on the provision of resources, etc.). However, as a rule, the execution of a single task is controlled independently of other tasks and the state of the entire system as a whole. In this case, priorities, scheduling options for individual tasks and their parameters are not taken into account. The absence of global coordinating mechanisms leads both to an increase in the execution time of individual tasks and to the underutilization of resources. On the other hand, end users do not have information about the timing of obtaining a solution or service capabilities, which leads to a loss of quality of service. Allowing the user to control system performance expectations for each job improves the quality of service for a particular job, but may have a negative impact on the performance of other jobs.

Classical schedulers are built either on minimizing response time (real time) or on maximizing total resource utilization (time sharing). And since the purpose of global schedulers is to improve the state of the system as a whole, the requirements of individual consumers are practically not taken into account and user tasks can be performed for hours. To minimize service time, it is necessary to adjust the strategy of classical schedulers so that, on the one hand, take into account the interests of users, task priorities and their execution time, and on the other hand, information about the state and distribution of system resources for general optimization of its operation.

The purpose of this work is to present a systematic approach to the construction of high-performance specialized computing systems (SCS) that perform resource-intensive tasks related to a special class, the execution methods of which can be defined as random enumeration with an unknown outcome [1,2]. Here, obtaining a solution is based on enumeration algorithms and comes down to searching for a fragment with predetermined properties in a large array of initial data. Such an array, as a rule, consists of separate, indivisible, identical in size, meaningfully significant fragments. Each task is considered solved as soon as a unique element can be identified in some piece of data. Tasks are of different types. This includes searching for a graphic object on map fragments for its recognition, and searching the Internet for some text on a given fragment, and encryption and decryption tasks.

Based on previous works [3-6], the following conclusions can be drawn. The use of classical schedulers in SCS does not allow to fully realize all their capabilities and reduces productivity and efficiency.

To eliminate this, methods for managing such systems based on intelligent agents (IAs) have been developed. Such IAs, having no information about the initial state of the system, according to the obtained statistical data on its functioning, can significantly increase the productivity and improve the efficiency of the SCS.

The management of the passage of tasks in the SCS is carried out by IAs based on the parameters assigned by them for each task, without having an analytical description of the entire system. Based on this, this type of control can be attributed to artificial intelligence systems.

The report discusses a systematic approach to the construction of various SCS control schemes based on IAs. Approaches to measuring the maximum performance will be shown and an analysis of the quality of work of such systems will be carried out.

Литература

Малашенко Ю.Е., Назарова И.А. Модель управления разнородными вычислительными заданиями на основе гарантированных оценок времени выполнения. // Изв. РАН. ТиСУ. 2012. No 4. С. 29-38.
Купалов-Ярополк И.К., МалашенкоЮ.Е., НазароваИ.А. и др. Модели и программы для системы управления ресурсоемкими вычислениями. М.: ВЦ РАН, 2013. http://www.ccas.ru/depart/malashen/papper/ronzhin2012preprint.pdf.
Голосов П.Е., Гостев И.М. О некоторых имитационных моделях планировщиков операционных систем // Телекоммуникации. 2021 No 6, ст. 10-21.
Голосов П.Е., Гостев И.М. Об имитационном моделировании функционирования операционной системы с вытесняющим планированием // Телекоммуникации. 2021. No 8. С. 2–22.
Голосов П.Е., Гостев И.М. Имитационное моделирование серверов с прерываниями в больших многопроцессорных системах // Известия вузов. Приборостроение. 2021. Т. 64. No 11. С. 879–886. https://doi.org/10.17586/0021-3454-2021-64-11-879-886.
Golosov P.E., Gostev I.M. About one cloud computing simulation model // Systems of Signals Generating and Processing in the Field of on Board Communications, Conference Proceedings. 2021. P. 9416100. https://doi.org/10.1109/IEEECONF51389.
Golosov P.E., Gostev I.M. Cloud computing simulation model with a sporadic mechanism of parallel problem solving control. Scientific and Technical Journal of Information Technologies, Mechanics and Optics, 2022, vol. 22, no. 2, pp. (in Russian).

Summary

Author: Prof. GOSTEV, Ivan (IITP RAS)

Co-author: Dr GOLOSOV, Pavel (Russian Presidential Academy of National Economy and Public Administration)

Presenter: Dr GOLOSOV, Pavel (Russian Presidential Academy of National Economy and Public Administration)

Session Classification: Distributed Computing Systems

Track Classification: Distributed Computing Systems