Adaptive Resource Allocation in High-Performance Distributed Multimedia Computing

Yang, R.

2011

Summary

Adaptive Resource Allocation in High-Performance Distributed Multimedia Computing

In recent years, the increasing importance of multimedia data, in particular in the form of still pictures and video, has boosted demands for automatic extraction, comparison, and processing of features from such data - and has led to the new research domain of Multimedia Content Analysis (MMCA). In the very near future, computerized access to the content of multimedia data will be a problem of phenomenal proportions, as digital video may produce data at extremely high rates, and multimedia archives steadily run into petabytes of storage. At the same time, applications in MMCA must often run under strict time constraints. For example, to avoid delays in queues of people waiting, a biometric authentication system must identify a persons identity within several seconds. Large autonomous applications, such as the automatic detection of suspect behavior in video data obtained from surveillance cameras, may even need to work under real-time restrictions. As individual computers cannot satisfy the high computational demands, distributed supercomputing on large collections of compute clusters (Grids) is rapidly becoming indispensable.

In our research, we restrict ourselves only on the image processing of the multimedia elements. In this kind of services-based execution scenario, a client program (typically a local desktop computer) connects to one or more remote multimedia servers, each running on a (different) compute cluster. At application run-time, the client application sends video/audio frames (e.g., captured by a camera) to any number of available servers, each performing the analysis in a data parallel manner. A highly complicating factor is the strong variability in the availability of hardware and software resources. Therefore, a fundamental problem of parallel computing in a Grid environment is to achieve high execution efficiency in the presence of the dynamically changing hardware and software availability.

In MMCA applications there are several sources of dynamic behavior. First, the sources of computer power are often shared by numerous applications that may make the available capacity scarce and varying over time. Although technological improvements will increase the bandwidth of single wide-area links beyond the Gbits/second range, the demands of emerging MMCA problems will increase - rather than reduce - competition for resources. Second, in MMCA applications the amount of data that needs to be processed often changes wildly over time. For instance, in the application for the comparison of objects and individuals in video streams obtained from multiple surveillance cameras, the job-arrival process is fairly constant and predictable, whereas in the iris-scan applicatie the arrival process has a random nature, and the job-arrival epochs may be hard to predict at small time scales. Because of their dynamic behavior, MMCA applications must be made variability-tolerant by means of controlled adaptive resource utilization. This raises the need for new stochastic runtime performance-control methods that properly react to changing circumstances.

Apart from satisfying the time constraint of MMCAs, it is also essential to minimize the utilization costs because the sources of computer power are often shared by numerous applications that may make the available capacity scarce and varying over time. It gives rise to a class of stochastic models that aims to minimize the average utilization costs by proper on-the-fly actions, while at the same time meeting a time constraint. In this thesis, we focus on the development, analysis and optimization of this kind of stochastic control schemes.

After an introduction in Chapter 1, we brief discuss the basic concepts of prediction methods and Markov Decision Processes (MDPs) in Chapter 2 that are important and needed for the other chapters, and we provide a literature overview of these methods in the context of Grid computing.

In Chapter 3 we discuss the so-called "resource utilization" (RU) problem and the "just-in-time" (JIT) communication problem in MMCA applications in which the amount of data that needs to be processed is enormous. The RU problem focuses on determining the optimal number of compute nodes used by each multimedia server, properly balancing the complex tradeoff between computation and communication. The JIT problem aims to tune the transmission of newly generated data sent to remote servers, so as to obtain the highest service utilization, while minimizing the need for buffering. For the RU problem, we develop a simple and easy-to-implement method to determine the optimal number of nodes to be employed, which is based on the

classical binary-search method for non-linear optimization and is independent of the specifics of the system. The JIT problem is addressed by a smart adaptive control method that properly reacts to the continuously changing circumstances in large-scale Grid environment. Extensive experimental validation shows that our optimization approaches are highly effective.

In Chapter 4 we study optimal resource allocation in time-reservation systems. In such systems, jobs arrive at a service facility and receive service in two steps; in the first step information is gathered from the customer, which is then sent to a pool of computing resources, and in the second step the information is processed, after which the customer leaves the system. Here, two decisions should be made: (1) *when* to reserve computing power from the pool of resources, such that the job does not have to wait for the start of the second service step and that the processing capacity is not wasted due to the job still being serviced at the first step, and (2) *how many* processors to allocate for the second processing step such that reservation and holding costs are minimized. We decompose the problem into two parts. First, we show that the near-optimal reservation moment is given by the difference of the mean service time in the first step and the mean reservation time. Then, we apply dynamic programming to show that the near-optimal resource-allocation policy follows a step function with as extreme policy the bang-bang control for given structures of the cost function and the service rate function.

In Chapter 5, we extend the second part of the model in Chapter 4 to multi-queue systems. In such systems, each service facility poses a constraint on the maximum expected sojourn time of a job. The decision should be made to dynamically allocate the servers over the different facilities such that the sojourn-time constraints are met at minimal costs. We model this problem as a Markov decision problem and derive the structural properties of the relative value function via standard induction-based arguments. These properties, which are hard to derive for multi-dimensional systems (together with the properties described in Chapter 6), give a full characterization of the optimal policy.

In Chapter 6 we study a special case of the model described in Chapter 5, in which there is only a single queue. We show via dynamic programming that (1) the optimal allocation policy is work-conserving, and (2) the optimal number of servers follows a step function with as extreme policy the bang-bang control policy. Moreover, (3) we provide conditions under which the bang-bang control policy is optimal. The derivation of these results is based

on a combination of direct arguments and induction, which are not generalizable to multiple queues. Therefore, these characterizations of the optimal policy are not directly applicable in the multi-queue setting. However, it provides a good foundation of exploring the additional optimal allocation properties in multi-queue systems.

In Chapter 7, we study the optimal allocation policy for multi-queue systems with time-varying arrivals. We consider the problem under two different cases. In the first case, the time-varying parameters are known beforehand, and we show how the optimal policy can be obtained numerically. On the contrary, the second case considers the optimal allocation problem *without* full knowledge of the job arrival rates. For this case, we use both a prediction method and a stochastic approximation method to track the time-varying parameters to obtain near-optimal policies. Numerical results show that our techniques are highly effective.

Finally, in Chapter 8 we validate our resource-allocation policies in an experimental setting. The results show that our methods are extremely effective in practical scenarios.