XCloud: Extensible Performance Management for Cloud Data Services

Olga Papaemmanouil
Brandeis University
olga@cs.brandeis.edu

ABSTRACT

Cloud computing is drastically transforming our perception of how data-centric applications are deployed by reducing data processing services to commodities that could be acquired and paid for on-demand. Despite the increased adoption of cloud data services, challenges related to application management still exist. These challenges include tasks such as provisioning cloud resources to meet application-specific Quality-of-Service (QoS) goals, assigning incoming query workloads to the reserved resources to optimize resource utilization, monitoring performance to ensure acceptable QoS levels. These tasks strongly depend on the application-specific workload characteristics and performance criteria, therefore their implementation burden is left on the application developers.

Existing efforts on cloud databases (e.g., [1, 2]) offer solutions to some of these challenges but these mechanisms are designed for specific performance metrics (typically query latency). However, the diversity of data management applications and workloads (e.g., scientific, financial, etc.) unavoidably implies the need for customizable systems that support equally diverse performance objectives (e.g., throughput, response latency, etc.). Today, the lack of such customizable performance management tools leaves developers to rely on ad-hoc mechanisms for addressing their individual application management needs. But such solutions eventually drastically hinder the application's implementation and maintainability. For instance, if the optimization metrics and performance criteria need to be modified during the lifetime of the application, its designers are required to implement, test and deploy new resource and workload management algorithms often from scratch.

In this abstract, we argue that the performance management of data processing applications should itself be offered to developers as a cloud-based automated service. Specifically, we envision resource, workload and QoS management services that hide the complexities and the volatility of the cloud environment from the developers as well as provide them with the necessary abstractions to declaratively express and realize performance requirements of their applications. We believe that these services will simplify the development of cloud data management applications and enable the seamless creation of new ones.

To realize our vision we propose XCloud, an extensible cloud-based performance management service platform. The key feature of XCloud is performance extensibility, i.e., the ability to express customized performance models and constraints and use them to automatically extend the functionality of core application tasks, such as resource and workload management, towards meeting the applications's performance objectives. XCloud supports performance extensibility for cloud databases by (a) enabling performance criteria and performance Service-Level-Agreements (SLAs) to be defined at the application level and (b) offering a suite of application management services, including workload, resource and QoS management, that can seamlessly customize their functionality towards satisfying application-specific performance objectives.

We envision XCloud as a suite of services operating on an Infrastructure-as-a-Service (IaaS) cloud (e.g., Amazon, GoGrid) and complementing a Database-as-a-Service (DBaaS) utility. XCloud's clients are the application designers that deploy an application on a cloud-based database service. XCloud acts as a "middleware" between the application and the database service by providing automatic capacity planning, workload allocation, as well as QoS specification and management services.

XCloud supports performance extensibility through a declarative language, XCLang (XCloud Language), for specifying performance metrics. The language makes available through an API: (a) a set of built-in metrics (e.g., query response time, throughput), (b) performance prediction models and (c) properties of the underlying cloud infrastructure (e.g., I/O rate, CPU load, etc), database engine (e.g., buffer pool size), queries (e.g., query plan cost) and operators (e.g., operator selectivity). Its unique feature is that is allows applications to formulate and register their own performance metrics and prediction models either through black-box style functions or through declarative specifications that combine the available API functions. XCLang performance metric specifications are then translated to executable code that customizes XCloud's performance monitoring and QoS management modules (e.g., query admission and scheduling, resource provisioning).

Our proposed system involves a novel combination of extensible performance APIs and customizable application management solutions that will significantly improve the quality of existing cloud databases as well as ease the development efforts of application designers. In this talk we will motivate the need for such an extensible performance management service and overview the design of XCloud and the key challenges towards its implementation.

1. REFERENCES

[1] J. Duggan, U. Cetintemel, O. Papaemmanouil, and E. Upfal. Performance prediction for concurrent database workloads. In *SIGMOD*, 2011.

[2] Xiong et al. Intelligent management of virtualized resources for database systems in cloud environment. In *ICDE*, 2011.

7th Biennial Conference on Innovative Data Systems Research (CIDR '15) January 4-7, 2015, Asilomar, California, USA