IGI PUBLISHING
701 E. Chocolate Avenue, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.igi-global.com

ITJ 5156

This paper appears in the publication, Journal of Database Management, Volume 20, Issue 3
edited by **Keng Siau © 2009, IGI Global**

Towards Autonomic Workload Management in DBMSs

Baoning Niu, Taiyuan University of Technology, China and Queen's University, Canada

Patrick Martin, Queen's University, Canada

Wendy Powley, Queen's University, Canada

ABSTRACT

Workload management is the discipline of effectively managing, controlling, and monitoring work flow across computing systems. It is an increasingly important requirement of database management systems (DBMSs) in view of the trends towards server consolidation and more diverse workloads. Workload management is necessary so the DBMS can be business-objective oriented, can provide efficient differentiated service at fine granularity, and can maintain high utilization of resources with low management costs. The authors see that workload management is shifting from offline planning to online adaptation. In this article, the authors discuss the objectives of workload management in autonomic DBMSs and provide a framework for examining how current workload management mechanisms match up with these objectives. They then use the framework to study several mechanisms from both DBMS products and research efforts. They also propose directions for future work in the area of workload management for autonomic DBMSs. [Article copies are available for purchase from InfoSci-on-Demand.com]

Keywords: *Autonomic Computing; Workload Characterization; Workload Management*

INTRODUCTION

Workload management involves the monitoring and control of work entering a system. Its goal is to ensure that sufficient resources are allocated to a workload so that a business can meet its management objectives (IBM Corporation, 2003a). Workload management is becoming increasingly important to businesses for several reasons. First, their workloads are becoming more diverse and complex. Second, the emerging trend of server consolidation has led to an environment with increased competition for shared resources between applications from potentially disjointed organizations, which results in a workload with diverse and dynamic resource demands and often competing performance objectives. Third, Web-based applications, especially Web services (Erickson & Siau, 2008), introduce a need for flexible and guaranteed application service levels because they tend to involve unpredictable workloads, with a high rate of overall growth in workload size (D. H. Brown Associate, Inc., 2004). Allocating database management system (DBMS) resources to competing workloads to meet performance objectives is a challenge.

We believe that workload management stands to benefit greatly from the new paradigm of autonomic computing. The goal of autonomic computing is to simplify system complexity by governing all computations within a given system (Mainsah, 2002). Autonomic computing systems have four fundamental characteristics (Ganek & Corbi, 2003): self-configuring, self-healing, self-optimizing, and self-protecting. Self-configuring means systems can automatically adapt to dynamically changing environments. New features, applications, and servers can be dynamically added to the enterprise infrastructure with no disruption of services. Self-healing is the ability of systems to discover, diagnose, and react to disruptions. Such a system needs to be able to predict problems and take necessary actions to prevent the failures from impacting the services. For a system to be self-optimizing, it should monitor itself and tune resources automatically to maximize resource utilization to meet users' performance requirements. Self-protecting means systems are able to anticipate, detect, identify, and protect themselves from attacks from anywhere.

The goal of this article is to specifically examine the progress made towards providing autonomic workload management in DBMSs and, based on this examination, to identify directions for future research. Previous surveys of the autonomic computing area, such as those by Salehie and Tahvildari (2005) and Elnaffar, Powley, Benoit, and Martin (2003), tend to have a more general focus and use criteria focusing on the existence of facilities to support the general features of self-configuration, self-healing, self-optimization, and self-protection. We propose a framework, Autonomic Workload Management Framework (AWMF), for the examination of autonomic workload management and then analyze current products and research efforts with respect to it. The framework specifies a model of the key processes and functions in autonomic workload management as well as identifying criteria for evaluation.

The remainder of the article is structured as follows. The following section discusses the objectives of workload management. We then outline our framework, AWMF, for autonomic workload management and examine workload management mechanisms taken from current DBMS products and research efforts and compare them using AWMF. A number of observations are derived from the comparison of the workload management mechanisms. We then validate AWMF using experiments with Query Scheduler, a prototype implementation of AWMF and finally summarize our survey and point out possible directions for future work.

MANAGEMENT OBJECTIVES

Workload management has evolved through three phases, from its infancy as a means for capacity planning (Lo & Douglas, 2007), to resource-oriented workload management, with the primary goal being resource utilization (Castro, Tezulas, Yu, Berg, Kim, & Gfroerer, 2001; D. H. Brown Associate, Inc., 2004), and finally to today's performance-oriented workload management, with a focus on business objectives (IBM Corporation, 2003a). The objectives of workload management have remained the same, namely cost sharing and meeting Service Level Objectives (SLOs) (IBM Corporation, 2003a; Menascé, 2004), but its style has changed from offline analysis to online adaptation.

Cost sharing consolidates multiple applications onto a single server to improve resource utilization and cut down costs. An SLO objectively defines the service level delivered to users in terms of system level metrics or business process level metrics. System level metrics include system availability, transaction response times and distribution, transaction volumes, and resource provisioning turnaround time. Business process level metrics include, for example, the time-frame for response and problem resolution (IBM Corporation, 2003a). Workload management is only interested in the

Related Content

Integrity Constraints in an Active Database Environment
Juan M. Aleand Mauricio Minuto Espil (2002). *Database Integrity: Challenges and Solutions (pp. 113-143).*
www.irma-international.org/chapter/integrity-constraints-active-database-environment/7880

An Efficient Method of Tooth Segmentation Under Massive Medical Data
Tian Ma, Yizhou Yang, Yun Li, Zhanli Liand Yuancheng Li (2022). *Journal of Database Management (pp. 1-22).*
www.irma-international.org/article/an-efficient-method-of-tooth-segmentation-under-massive-medical-data/309414

Semi-Supervised Event Extraction Incorporated With Topic Event Frame
Gongqing Wu, Zhuochun Miao, Shengjie Hu, Yinghuan Wang, Zan Zhangand Xianyu Bao (2023). *Journal of Database Management (pp. 1-26).*
www.irma-international.org/article/semi-supervised-event-extraction-incorporated-with-topic-event-frame/318453

Blockchain Application in Retirement Planning Investment: Improving Transparency and Viability
Anup Sharmaand Nitin Gupta (2022). *Applications, Challenges, and Opportunities of Blockchain Technology in Banking and Insurance (pp. 246-257).*
www.irma-international.org/chapter/blockchain-application-in-retirement-planning-investment/306466

Multi-Label Classification: An Overview
Grigorios Tsoumakasand Ioannis Katakis (2009). *Database Technologies: Concepts, Methodologies, Tools, and Applications (pp. 309-319).*
www.irma-international.org/chapter/multi-label-classification/7918