

Thesis Title	The Runtime Estimation Framework for Black Box Scientific Applications
Thesis Credits	12
Candidate	Miss Sarunya Pumma
Thesis Advisor	Assoc. Prof. Dr. Tiranee Achalakul
Program	Master of Engineering
Field of Study	Computer Engineering
Department	Computer Engineering
Faculty	Engineering
Academic Year	2013

Abstract

The growth of data and computation in the past decade has brought about the needs for cloud infrastructure. Cloud leverages the Internet as a tool through which remote computers can share resources on-demand. The cloud infrastructure can be utilized as a high performance computing (HPC) platform which contains flexible and excessive computing resources. In order to efficiently run the HPC applications in the cloud, a great deal of technical knowledge is required. One of the challenges is how to estimate the runtimes of applications accurately because an inaccuracy in runtime estimation can lower the overall performance of a computer system. Moreover, runtime is an important attribute for tasks scheduling. For instance, most well known scheduling algorithms, such as, Backfilling and Heterogeneous Earliest Finish Time (HEFT), use runtime to determine the schedule of the tasks. In this thesis, we have proposed a runtime estimation method for unknown-profile applications in the cloud computing environment. Unlike other approaches, we also provide a procedure to collect the profiles of applications, which are the metrics that represent the execution behavior of an application. This allows our approach to predict the runtime of the HPC applications even if the metadata is not provided. In order to predict a runtime of a workload, only two steps are required. In the first step, the application will be classified into a class based on the similarity of the execution characteristics. In our work, we have adopted the Berkley's Dwarfs taxonomy to define the classes. The classification result will be used to choose a runtime prediction equation for the workload. In the next step, the runtime will be predicted by using the equation that is selected in the previous step. The runtime prediction equations are constructed by using the Artificial Bee Colony (ABC) and the linear regression techniques. In order to verify the practicality of our framework, we predicted the runtimes of the HPC applications on three types of virtual machines, General purpose, Compute Optimized, and Memory Optimized instances, provided by Amazon EC2. Our method can yield low prediction error percentages in most cases. Moreover, it can provide more accurate runtime prediction results in comparison to the user-estimation method.

Keywords : Amazon EC2 / Berkley's Dwarfs / Cloud Computing / Job Scheduling / Runtime Estimation

หัวข้อวิทยานิพนธ์	การคำนวณเวลาในการประมวลผลแอปพลิเคชันเชิงวิทยาศาสตร์
หน่วยกิต	12
ผู้เขียน	นางสาวศรัณยา ภูมิมา
อาจารย์ที่ปรึกษา	รศ. ดร. ชีรณี อจลากุล
หลักสูตร	วิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
ภาควิชา	วิศวกรรมคอมพิวเตอร์
คณะ	วิศวกรรมศาสตร์
ปีการศึกษา	2556

บทคัดย่อ

อัตราการเพิ่มขึ้นของข้อมูลและการประมวลผลในช่วงสิบปีที่ผ่านมาซึ่งความต้องการในการใช้บริการการประมวลผลแบบกลุ่มเมฆหรือคลาวด์ ทั้งนี้ระบบคลาวด์ใช้ระบบอินเทอร์เน็ตเป็นเครื่องมือในการรวมทรัพยากรคอมพิวเตอร์ไว้ด้วยกันเพื่อตอบสนองความต้องการของผู้ใช้ โดยระบบคลาวด์สามารถถูกนำมาใช้เป็นแพลตฟอร์มสำหรับการประมวลผลแอปพลิเคชันที่ต้องการทรัพยากรในการประมวลผลสูงได้ เนื่องจากผู้ใช้งานสามารถร้องขอทรัพยากรในการประมวลผลหรือจัดเก็บข้อมูลได้ตามความต้องการ อย่างไรก็ตามการประมวลผลบนคลาวด์ต้องอาศัยความรู้เชิงเทคนิคอย่างมาก ซึ่งหนึ่งในปัจจัยที่ควรจะต้องพิจารณาคือวิธีการที่มีประสิทธิภาพในการคำนวณเวลาที่ใช้ในการประมวลผลแอปพลิเคชัน เนื่องจากความคลาดเคลื่อนในการคาดการณ์เวลาที่ใช้ในการประมวลผลอาจส่งผลกระทบต่อประสิทธิภาพโดยรวมของระบบคอมพิวเตอร์ได้นอกจากนี้ตัวจัดลำดับงาน (Scheduler) ที่เป็นที่ยอมรับ เช่น Backfilling และ Heterogeneous Earliest Finish Time (HEFT) ใช้เวลาในการประมวลผลเป็นตัวแปรที่สำคัญในการจัดลำดับงานภายในระบบ ดังนั้นงานวิจัยนี้จึงนำเสนอวิธีการในการคำนวณเวลาในการประมวลผลสำหรับแอปพลิเคชันเชิงวิทยาศาสตร์บนระบบคลาวด์ โดยแอปพลิเคชันดังกล่าวมักจะต้องทรัพยากรในการประมวลและจัดเก็บข้อมูลจำนวนมาก ทั้งนี้วิธีการที่นำเสนอครอบคลุมไปถึงวิธีการเก็บโปรไฟล์ (Profile) ของแอปพลิเคชันซึ่งถูกจัดเก็บในรูปแบบของเมตริก (Metrics) ซึ่งใช้ในการอธิบายลักษณะเฉพาะในการประมวลผลของแอปพลิเคชันนั้นๆ ทั้งนี้การเก็บโปรไฟล์ของแอปพลิเคชันจะทำให้สามารถคาดการณ์เวลาในการประมวลผลได้โดยไม่ต้องอาศัย เมตาดาต้าของแอปพลิเคชัน (Metadata) ซึ่งการคำนวณเวลาในการประมวลผลสำหรับงานวิจัยนี้ประกอบด้วย 2 ขั้นตอนหลัก โดยในขั้นแรกแอปพลิเคชันจะถูกแบ่งออกเป็นคลาส (Class) โดยพิจารณาจากลักษณะเฉพาะในการประมวลผล ทั้งนี้คณะนักวิจัยได้อ้างอิงการแบ่งคลาสตามงานวิจัยของมหาวิทยาลัยแคลิฟอร์เนีย เบิร์กลีย์ ซึ่งแบ่งแอปพลิเคชันเชิงวิทยาศาสตร์ออกเป็นหมวดหมู่ซึ่งถูกเรียกว่า คนแคระ (Dwarf) ตามลักษณะการประมวลผลของแต่ละแอปพลิเคชัน ทั้งนี้ผลการแบ่งคลาสจะถูกนำไปใช้ในการเลือกสมการในการคำนวณเวลาในการประมวลผลของแอปพลิเคชันในขั้นตอนต่อไป สมการดังกล่าวถูกสร้างขึ้นโดยใช้ อัลกอริทึมการจำลองฝูงผึ้ง (Artificial Bee Colony หรือ ABC) และการวิเคราะห์การถดถอยเชิงเส้น (Linear Regression) สำหรับการประเมินประสิทธิภาพของวิธีการที่นำเสนอ นั้น คณะนักวิจัยได้นำวิธีดังกล่าวไปใช้

ในการคำนวณเวลาในการประมวลผลของแอปพลิเคชันต่างๆ บนเครื่องคอมพิวเตอร์เสมือน (Virtual Machine) 3 รูปแบบซึ่งให้บริการโดยอเมซอนอีซีทู (Amazon EC2) คือ เครื่องแบบทั่วไป (General Purpose) เครื่องสำหรับการคำนวณที่ใช้หน่วยประมวลผลสูง (Compute Optimized) และเครื่องสำหรับการคำนวณซึ่งใช้หน่วยความจำสูง (Memory Optimized) ทั้งนี้วิธีการที่นำเสนอสามารถคาดการณ์เวลาที่ใช้ในการประมวลผลด้วยความคลาดเคลื่อนต่ำ รวมทั้งยังมีความแม่นยำมากกว่าวิธีที่ให้ผู้ใช้งานเป็นผู้ประมาณเวลาในการประมวลผลอีกด้วย

คำสำคัญ : การคำนวณเวลาในการประมวลผล / การจัดลำดับงาน / คนแกระของเบิร์กกลีย์ / ระบบคลาวด์
 / อเมซอนอีซีทู