موازنة العبء لحوسبة البيانات الكبيرة مع الوعي بمحلية البيانات

ياسر عرفات

بحث مقدم لنيل درجة الماجستير في العلوم (علوم الحاسبات)

أشراف

أ.د / راشد محمود
د / عياد البشرى

قسم علوم الحاسبات
كلية الحاسبات وتقنية المعلومات
جامعة الملك عبدالعزيز
جدة- المملكة العربية السعودية
رجب ١٤٣٨هـ - أبريل ٢٠١٧ م

موازنة العبء لحوسبة البيانات الكبيرة مع الوعي بمحلية البيانات

ياسر عرفات

المستخلص

الهدف من رسالتي هو تصميم تقنيات جديدة التي تحسن من أداء الحلول الخطية المتكررة لمتن.-موازنة التحميل (Loadbalancing) للمهام الحسابية والبيانات تلعبان دورا حاسما في نظم البيانات الضخمة (Big Data systems). بشكل عام، موازنة التحميل تساعد على تحسين الحوسبة (computation) أو على أداء النظام. على سبيل المثل الإنتاجية (throughput) وزمن الحل (أو زمن الاستجابة) واستخدام الموارد. من خلال موازنة التحميل يمكن تحقيق التوزيع الأمثل للبيانات وأعباء العمل إلى الموارد المتاحة. وتهدف تقنيات مركزية البيانات (Data locality) لجدولة المهام مع العقد (nodes) وذلك عن طريق البيانات التي لها صلة ببعض في كلا الطرفين. أهداف أداء (موازنة التحميل ومركزية البيانات) عادة ما تكون على خلاف. ولذلك، هناك حاجة لإيجاد استراتيجيات مثلى لتحقيق أقصى قدر من الأداء.

الهدف من هذه الرسالة هو تطوير محليات البيانات مع الاخذ بالحسبان بما يتعلق بعمليات موازنة الاحمال للبيانات الكبيرة وتطبيقها على المشاكل العملية ذات الأهمية العالية. تم عرض مسح ادبي تفصيلي للمؤلفات لتحديد التحديات الرئيسية في أبحاث البيانات الكبيرة. تشمل هذه التحديات، من بين أمور أخرى، موازنة الاحمال ومحلية البيانات. أيضاً، قمنا بتطوير تقنية موازنة الاحمال والتي اخذت بعين الاعتبار مواقع تلك البيانات وتم تطبيقها على مشكلة النقل البري القائم على الرسم البياني. لقد قمنا بتصميم نماذج شبكة الطرق الأمريكية بأكملها والتي تحتوي على حوالي ٢٤ مليون قمة (vertices) و ٥٨ مليون قوس (arcs). الهدف من هذا البحث هو تحديد نقاط مختلفة ذات الأهمية (POI) في المناطق بما في ذلك أماكن المعيشة ومراكز الرعاية الصحية. هذه النقاط استخدمت في وقت لاحق للعثور على أقصر الطرق لأغراض التخطيط والعمليات. قمنا بعرض الخوارزميات التي قمنا بتطويرها في هذه الأطروحة. قد تم تنفيذ الخوارزميات باستخدام منصة سبارك للبيانات الكبيرة على الحاسوب خارق الاداء عزيز، وهو احد الحواسب المصنفة في قائمة TOP500 على مستوى العالم وفقا لتصنيف يونيو ونوفمبر ٢٠١٥. تم جمع النتائج المتعلقة بأقصر الطرق وتم عرض شبكات الطرق كرسوم بيانية. تم تحليل أداء موازنة الاحمال وتقنيات الوعي بمكان البيانات باستخدام الحاسب خارق الاداء عزيز وذلك من خلال استخدام اعداد مختلفة من الاجهزة (nodes) وتم الحصول على سرعة جيدة. تم استخلاص استنتاجات وتوجيهات لما يمكن عمله في المستقبل.

Load Balancing for Big Data Computing with Data Locality Awareness

By

Yasir Arfat

A thesis submitted for the requirements of the degree of Master of Science in Computer Science

Supervised by

Prof. Rashid Mehmood

Dr. Aiiad Albeshri

FACULTY OF COMUTING & INFORMATION TECHNOLOGY

KING ABDULAZIZ UNIVERSITY

JEDDAH-SAUDI ARABIA
Rajab 1438H – April 2017G

Load Balancing For Big Data Computing with Data Locality Awareness

Yasir Arfat

ABSTRACT

Load balancing of computational tasks and data plays a critical role in big data systems. Load balancing attempts to optimize the overall computation or system performance, such as throughput, solution (or response) time, and resource usage. This is achieved by an optimum allocation of data and workloads to the available resources. Data locality techniques aim to map tasks to the nodes where the relevant data resides. The two performance objectives (i.e. load balancing and data locality) are usually at odds. There is a need to find optimal strategies to maximize the performance.

The aim of this thesis is to develop data locality aware load balancing techniques for big data computing and apply these to practical problems of high significance. A detailed review of the literature is presented to identify major challenges in big data research. These challenges include, among others, load balancing and data locality. A load balancing technique that is also aware of data locality has been developed and is applied to a graph-based road transportation problem. We have modelled the entire US road network data that contains approximately 24 million vertices and 58 million arcs; the specific aim of this research is to identify various points of interests (PoIs) in the regions including living places and healthcare centres. These PoIs are subsequently used to find the shortest paths among them for planning and operations purposes. The relevant algorithms that we have developed are presented in the thesis. The algorithms have been implemented using the Spark big data platform tools on the Aziz supercomputer, a Top500 supercomputer in the world according to the June and November 2015 rankings. The results are collected in terms of the shortest paths and the road networks are visualised as graphs. The performance of the load balancing and data locality awareness techniques is analysed against a varying number of nodes on the Aziz supercomputer and a good speedup has been reported. Conclusions are drawn with directions for future work.