تحسين الأداء القائم على تعليم الاله لحساب نواة SpMV لأنظمة إكساسكيل ذات الذاكرة الموزعة

سردار عثمان

مشرف د .اياد كاتب

المستخلص

تتطلب العديد من التطبيقات العلمية والهندسية والاقتصادية والاجتماعية حل أنظمة المعادلات الخطية المتفرقة ، ويتم استخدام التطبيقات بشكل متزايد للتطوير في الوقت المناسب ذكاء لتصميم والعمليات وإدارة ذكية للمدن والمجتمعات. المنتج المتناثر المصفوف المتفرق (SpMV) هو الأكثر شيوعاً واستهلاكاً للوقت من أجل الحل التكراري لـ أنظمة المعادلات الخطية المتفرقة. لذلك، الحوسبة عالية الأداء (HPC) تستخدم ميزات الحوسبة المتوازية للبرامج والأجهزة الأساسية لحل المشاكل الكبيرة بشكل أسرع .ويتم تطبيق HPC على SpMV والجبر الخطي ، ومشاكل أخرى لعدة عقود. البيانات الكبيرة والنهج القائمة على البيانات كانت تستخدم مؤخرا نسبيا في الحوسبة العلمية لمعالجة HPC التحديات ذات الصلة ، وهذا أدى إلى تقارب HPC والبيانات الكبيرة. علاوة على ذلك ، فإن الذكاء الاصطناعي (AI) يزداد تستخدم لتحسين البيانات الكبيرة HPC، الحوسبة العلمية ، وغيرها من المجالات المختلفة. وقد أدى هذا الاتجاه إلى تقارب كبير البيانات ، HPC وكذلك .AI وبالمناسبة تهدف هذه الرسالة إلى المساهمة في رفع متسوى هذا التقارب.

كخطوة نحو الحوسبة exascale وما بعدها ، وتطبيقها (تقارب المناطق الثلاثة) إلى مجال حسابات .SpMV هناك عدة عوامل تؤثر على أداء حسابات .SpMV هذه تشمل خصائص المصفوفة ، وتنسيقات التخزين ، والبرمجيات التطبيقات ، ومنصات الأجهزة. لذلك ، تحسين أداء تطبيق الأبنية متعددة النواة يمثل تحديا بسبب عدم التجانس وتنوع المباني. الآلات الحديثة لها مجموعة من الذاكرة المشتركة والموزعة ، والبنية المختلطة ، مع العديد من التسلسلات الهرمية التي تنطوي على اختفاء الاتصالات غير موحدة. النمط sparity من المصفوفة يؤثر على أداء SpMV الحسابات ، وخاصة في حالة الذاكرة الموزعة تطبيقات ، مما أدى إلى اختلال الحمل الذي يسبب كليهما حساب والتواصل النفقات العامة. العملية اليدوية للمحاكمة وتجربة الخطأ للعثور على أفضل عدد من العمليات ل حساب SpMV تستغرق وقتًا طويلاً. تقترح هذه الرسالة ZAKI و + ZAKI ، نهجين يعتمدان على البيانات والتعلم الآلي للتنبؤ بالشكل الأمثل عدد تكوينات العملية ورسم الخرائط الموازية الأمثل استراتيجيات لحساب SpMV لمصفوفة المتفرق التعسفي على آلة الذاكرة الموزعة. يتم استخدام ثلاث خوارزميات تعلم الآلة للتنبؤ - أشجار القرار ، والغابات العشوائية ، وتعزيز التدرج وقد تم تقييمها باستخدام ١٨٣٨ مصفوفات متفرقة في العالم الحقيقي المرتبطة ٤٥ مجالات التطبيق .تمثل البيانات المحلية تحديًا كبيرًا في تقديم الحوسبة exascale بسبب ارتفاع تكلفة الطاقة لحركة البيانات. هذه الأطروحة أيضا يساهم في مراجعة الأدبيات الأولى على موقع البيانات للمتقاربين HPC)، البيانات الكبيرة ، و (AI أنظمة .exascale بنية على أساس كما يتم توفير أنماط تصميم لتقارب HPC والبيانات الضخمة .وعلاوة على ذلك ، فإن التحقيق الأولي والاتجاهات المستقبلية هي المقدمة بشأن تطبيق الأداة المقترحة لدينا على الطاقة تحسين الكفاءة. وقد ساهمت هذه الأطروحة في نشر ما مجموعه

سبعةSCOPUSو أوأوراق ISI المفهرسة. على حد علمنا ، هذا أطروحة هي المحاولة الأولى ، حيث قمنا باستغلال هيكل المصفوفات للتنبؤ بالعدد الأمثل للعمليات وأفضلها استراتيجية رسم الخرائط لمصفوفة معينة في بيئة الذاكرة الموزعة عن طريق استخدام أساليب مختلفة في تعلم الآلة.

Machine learning based Performance Optimization of SpMV kernel for Distributed Memory Exascale Systems

By Sardar Usman

Supervised by: Dr. Iyad Katib

ABSTRACT

Numerous scientific, engineering, economic and social applications require the solution of sparse linear equation systems, and these applications are increasingly being used to develop timely intelligence for the design, operations, and management of smart cities and societies. Sparse matrix-vector product (SpMV) is the most important and time-consuming kernel for the iterative solution of sparse linear equation systems. High performance computing (HPC) typically exploits parallel computing features of the underlying software and hardware infrastructure to solve large problems faster. HPC has been applied to SpMV and linear algebra, and other problems for several decades. Big data and data-driven approaches have been used relatively recently in scientific computing to address HPC related challenges, and this has given rise to the convergence of HPC and big data. Moreover, artificial intelligence (AI) is increasingly being used to improve big data, HPC, scientific computing, and other problem domains. This trend has given rise to the convergence of big data, HPC and AI. This thesis aims to contribute to this convergence, as a step towards and beyond exascale computing, and applies it (the convergence of the three areas) to the area of SpMV computations.

Several factors affect the performance of SpMV computations. These include matrix characteristics, storage formats, software implementations, and hardware platforms. Performance optimization of an application on multicore architectures is challenging due to the heterogeneity and diversity of architectures. Modern machines have a range of shared and distributed memory, and hybrid architectures, with several hierarchies involving non-uniform communication latencies. The sparsity pattern of the matrix affects the performance of SpMV computations, particularly in case of distributed memory implementations, resulting in load imbalance that causes both computation and communication overheads. The manual process of trial and error experimentation to find the best number of processes to compute SpMV is time-consuming. This thesis proposes ZAKI and ZAKI+, two data-driven and machine-learning approaches to predict the optimal number of process configurations and optimal

parallelization mapping strategies for SpMV computations of an arbitrary sparse matrix on a distributed memory machine. Four machine learning algorithms are used for predictions -- Decision Trees, Random Forest, Xtreme and Gradient Boosting -- and have been evaluated using 1838 real-world sparse matrices associated with 45 application domains.

Data locality is a major challenge in delivering exascale computing due to the high energy cost of data movement. This thesis also contributes the first literature review on data locality for converged (HPC, big data, and AI) exascale systems. An architecture based on design patterns for convergence of HPC and big data is also provided. Moreover, a preliminary investigation and future directions are provided on the applicability of our proposed tool to energy efficiency optimization. This thesis has contributed a total of seven SCOPUS and/or ISI-indexed papers. To the best of our knowledge, this thesis is the first attempt, where we have exploited the structure of the matrices to predict the optimal number of processes and best mapping strategy for a given matrix in distributed memory environment by using different base and ensemble machine learning methods.