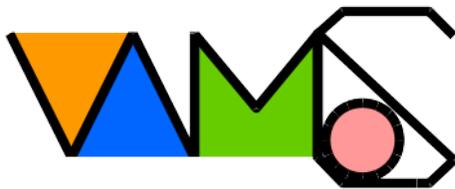


Thesis or HiWi: Parallelized ML Scheduling

Machine Learning hat im letzten Jahrzehnt eine Revolution erlebt und ist eine der wichtigsten Entwicklungen in der Technologie. Es erfordert jedoch enorme Mengen an Rechenleistung und daher muss die Last auf parallele Cluster mit vielen Arbeitern verteilt werden. Die für ML verwendeten Algorithmen weisen aber Planungsbeschränkungen auf, die den Datenaustausch zwischen den parallelen Aufgaben erfordern. Dies unterscheidet sich von den analytischen Modellen für Parallelsysteme, die von völlig unabhängigen Parallelaufgaben ausgehen. Wir entwickeln analytische Modelle für parallele Systeme mit Synchronisation Beschränkungen und benötigen reale Experimente, um diese Forschung zu validieren und zu leiten.



Beispielaufgaben:

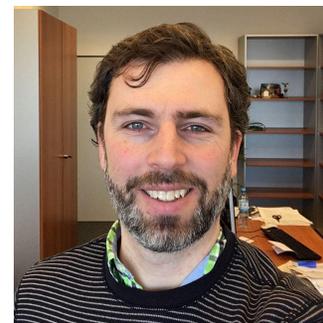
- Führen Sie Beispiel-Workloads aus, die die MLLib von Spark auf unserem Cluster verwenden.
- Analysieren Sie Spark Logs, um die Leistung verschiedener Workloads in MLLib zu charakterisieren.
- Untersuchen Sie verschiedene Möglichkeiten zur Parallelisierung von ML-Workloads.
- Wählen Sie eine parallelisierte ML-Plattform (außer Spark) und stellen Sie sie in unserem Cluster bereit.
- Führen Sie darauf Training- und Test-Workloads aus. Analysieren Sie die Logs und die Dokumentation, um Synchronisierungspunkte zu identifizieren.

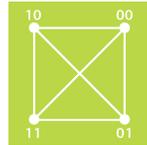
Nützliche Fähigkeiten:

- Programmiererfahrung: idealerweise Scala, Python oder Java.
- Vertrautheit mit einer Parallelverarbeitung Plattform: Hadoop, Spark, MPI, OpenMP.
- Vertrautheit mit Konzepten des maschinellen Lernens: Input- und Output Space, Training vs. Test Daten.
- Grundlegende Statistiken.

Kontaktieren Sie bitte:

Brenton Walker
Institut für Kommunikationstechnik
Appelstr 9a; Raum 1431, 14. OG
brenton.walker@ikt.uni-hannover.de
Tel: +49 511 762 2827

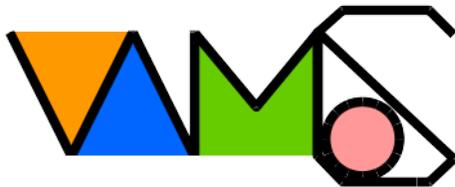




Thesis or HiWi: Parallelized ML Scheduling

Machine Learning has undergone a revolution in the past decade, and is one of the most important developments in technology. However it requires massive amounts of computational power, and therefore the load must be distributed across parallel clusters of many workers. The types of algorithms used for ML often have scheduling constraints that require exchanges of data between the parallel tasks. This is different from the analytical models for parallel systems, which assume completely independent parallel tasks.

We are developing analytical models for parallel systems with such synchronization constraints, and we need real experiments to validate and guide this research.



Example tasks:

- Run example workloads that use Spark's MLLib on our cluster.
- Analyze Spark logs to characterize the performance of different workloads in MLLib.
- Investigate different ways ML workloads are parallelized.
- Choose a parallelized ML platform (other than Spark) and deploy it on our cluster.
- Run training and testing workloads on it. Analyze the logs and documentation to identify synchronization points.

Useful Skills:

- Programming experience: ideally Scala, Python, or Java.
- Familiarity with a parallel processing platform: Hadoop, Spark, Flink, MPI, OpenMP.
- Familiarity with machine learning concepts: input and output space, training vs testing.
- Basic statistics.

Please contact:

Brenton Walker
Institut für Kommunikationstechnik
Raum 1431, 14. OG
Appelstr 9a
brenton.walker@ikt.uni-hannover.de
Tel: +49 511 762 2827

