

Energieeffiziente Datenzentren

Motivation

Yahoo Compute Coop

Temperaturabhängige Lastverteilung

Energieeffizientes Hadoop



- Ziel auf Datenzentrumsebene: Energieeffiziente Kühlung der Systeme
 - Stand der Kunst vor 2004
 - Empfehlung der ASHRAE [Americ. Society of Heating, Refrigerating and Air-Conditioning Engineers]
 - Optimale Temperatur für Datenzentren zwischen 16 und 18 °C
 - Umdenken seit 2004
 - Empirische Überprüfung der Empfehlung
 - Resultat: Temperaturen um 27 °C ausreichend
- Ziel auf Systemebene: Energieproportionale Systeme
 - Energieverbrauch ist abhängig von der geleisteten Arbeit
 - Ideal: Kein Stromverbrauch im Leerlauf
- Herausforderungen
 - Wie lässt sich ein Datenzentrum mit „natürlicher“ Klimaanlage bauen?
 - Wie lässt sich Anwendungswissen zum Stromsparen nutzen?
- Literatur
 -  Gregory Mone
Redesigning the data center
Communications of the ACM, 55(10):14–16, 2012.



Yahoo Compute Coop

- Grundkonzept für Klimatisierung
 - Normalfall: Kühlung mittels Umgebungsluft
 - An warmen Tagen: Zusätzliche Kühlung durch Verdunstungskälte
 - Kein Einsatz konventioneller Klimaanlage
 - Keine Entstehung von Kühlabwasser
- Beispielstandort: Lockport, New York, USA
 - Durchschnittstemperaturen im Juli (heißester Monat): 16–28 °C
 - Erwartungswerte
 - > 27 °C: 212 Stunden pro Jahr
 - > 32 °C: 34 Stunden pro Jahr
 - Ca. 20 km Distanz zu den Niagarafällen → Strom aus Wasserkraft
- Literatur
 -  A. D. Robison, Christina Page, Bob Lytle
Yahoo! Compute Coop (YCC): A next-generation passive cooling design for data centers
Technical Report DE-EE0002899, Yahoo Inc., 2011.



Kühlung des Datenzentrums

- Einsatz von Ventilatoren
 - Teil der Außenwände des Gebäudes
 - Steuerung des Luftstroms
- Kühlmodus des Datenzentrums abhängig von Außentemperatur
 - 21–29 °C: Nutzung unbehandelter Umgebungsluft
 - Umgebungsluft dringt durch Lüftungsschlitze in den Wänden ins Gebäude ein
 - Einsatz von Luftfiltern
 - Weiterleitung des Luftstroms durch die Server-Schränke
 - Abzug der warmen Luft durch Luftschlitze im „Dachboden“
 - 29–43 °C: Nutzung gekühlter Umgebungsluft
 - Luftstrom wie bei erster Variante
 - Zusätzlich: Kühlung der Luft mittels Verdunstungskälte
 - < 21 °C: Nutzung erwärmter Umgebungsluft
 - Luftstrom wie bei erster Variante
 - Zusätzlich: Rückführung eines Teils der Abluft zur Erwärmung der einströmenden kalten Umgebungsluft



Einsparungen beim Ressourcenverbrauch

■ Power Usage Effectiveness (PUE)

- Metrik für die Energieeffizienz von Datenzentren
- $PUE = \frac{\text{Gesamter Energieverbrauch}}{\text{Energieverbrauch der IT-Systeme}}$
- Idealwert: 1,0
- Üblicher Wert für industrielle Datenzentren: Wert zwischen 1,5 und 2,0
- Bisheriger Bestwert eines Yahoo-Datenzentrums
 - Standort: Wenatchee, Washington, USA
 - PUE-Wert: 1,25

■ Datenzentrum in Lockport [Robison et al.]

- PUE-Wert liegt zwischen 1,08 und 1,11
- Während mehr als 99 % der Betriebszeit reicht die natürliche Kühlung aus
- 99 % geringerer Wasserverbrauch als ein wassergekühltes Datenzentrum
- Nebeneffekt: Verfügbarkeitsgrad von 99,98 %

■ Ähnliches Konzept: Facebook (z. B. in Prineville, Oregon, USA)



Temperaturabhängige Lastverteilung

■ Problem

- Rechner produzieren je nach Auslastung unterschiedlich viel Wärme
 - Kühleffekt abhängig von der Distanz zur Klimaanlage
- Aufrechterhaltung einer einheitlichen Raumtemperatur nicht trivial

■ Temperaturabhängige Lastverteilung

- Detaillierte Temperaturmessung: Fläche + unterschiedliche Höhen
 - Platzierung von Prozessen abhängig von erwarteter Wärmeentwicklung
 - Ziele
 - Reduzierung der auftretenden Temperaturunterschiede
 - Minimierung der Höchsttemperatur
- Energieeinsparung durch Entlastung der Klimaanlage

■ Literatur



Ratnesh K. Sharma, Cullen E. Bash, Chandrakant D. Patel et al.
Balance of power: Dynamic thermal management for Internet data centers
IEEE Internet Computing, 9(1):42–49, 2005.



GreenHDFS

Übersicht

■ Log-Daten-Analyse eines Yahoo-Hadoop-Cluster im Produktiveinsatz

- 2600 Server, 34 Millionen Dateien, 6 Petabytes Daten
- Ergebnisse für eines der Hauptverzeichnisse
 - 99 % der Dateien werden innerhalb von 2 Tagen nach dem Anlegen gelesen
 - 80 % der Dateien werden max. 8 Tage nach dem Anlegen letztmalig gelesen
 - 80 % der Dateien werden später als 20 Tage nach dem letzten Lesen gelöscht
- Folgerungen
 - „Heiße Phase“: Relativ häufige Zugriffe kurz nach dem Anlegen der Daten
 - „Kalte Phase“: Anschließende, vergleichsweise lange Phase ohne Zugriffe

■ GreenHDFS: Energieeinsparung durch Anpassung an Nutzungsprofil

■ Literatur



Rini T. Kaushik, Milind Bhandarkar, and Klara Nahrstedt
Evaluation and analysis of GreenHDFS: A self-adaptive, energy-conserving variant of the Hadoop distributed file system
Proceedings of the 2nd International Conference on Cloud Computing Technology and Science (CLOUDCOM '10), S. 274–287, 2010.



GreenHDFS

Grundkonzept

■ Heiße Zone (Hot Zone)

- Verwaltung von Daten, die sich gerade in ihrer „heißen Phase“ befinden
- Mehrheit (z. B. 75 % [Kaushik et al.]) der Rechner im Cluster
- Rechner mit hoher Leistungsfähigkeit
- Durchgängiger Betrieb der Rechner

■ Kalte Zone (Cold Zone)

- Verwaltung von Daten, die sich gerade in ihrer „kalten Phase“ befinden
- Restliche Rechner im Cluster
- Rechner mit geringerer Leistungsfähigkeit aber vielen Festplatten
- Betrieb eines Rechners jeweils nur nach Bedarf (z. B. per *Wake-on-LAN*)

■ Üblicher Lebenszyklus einer Datei

- Erzeugung in der heißen Zone
- Bei einem bestimmten Alter der Datei: Migration in die kalte Zone
- Löschung der Datei in der kalten Zone



- Anpassung der Replikationslogik des verteilten Dateisystems [z. B. HDFS]
 - Definition eines *Covering Subset*: Untergruppe von Rechnern des Cluster
 - Modifikation des Auswahlmechanismus für Replikate
 - Mindestens ein Replikat jedes Datenblocks muss Teil des Covering Subset sein
 - Selektion der anderen Replikate wie bisher
- Vorteile
 - Covering Subset im Normalfall für Verfügbarkeit ausreichend
 - Sonstige Rechner nur beim Speichern der Ergebnisse erforderlich
- Nachteile
 - Reduzierung des Grads an Parallelität
 - Erhöhte Latenzen sowohl im Normal- als auch im Fehlerfall
- Literatur
 - Jacob Leverich and Christos Kozyrakis
On the energy (in)efficiency of Hadoop clusters
Operating Systems Review, 44(1):61–65, 2010.



- Phasenweise Bearbeitung von MapReduce-Jobs
 - Gebündelte Ausführung einzelner Jobs auf allen Rechnern
 - Anschließend: Versetzen des kompletten Cluster in den Energiesparmodus
 - Reaktivierung des Cluster zu Beginn der nächsten Phase
- Vorteile
 - Keine Einschränkung der für einen Job verfügbaren Ressourcen
 - Breite Lastverteilung der Dateisystemanfragen möglich
 - Keine Modifikationen am Dateisystem erforderlich
- Nachteile
 - Erhöhte Latenzen für Jobs, die während einer Energiesparphase eintreffen
 - Einzelner Job kann Wechsel in den Energiesparmodus aufhalten
- Literatur
 - Willis Lang and Jignesh M. Patel
Energy management for MapReduce clusters
Proceedings of the VLDB Endowment, 3(1-2):129–139, 2010.



BEEMR

Übersicht

- Analyse eines Facebook-Hadoop-Cluster im Produktiveinsatz
 - 3000 Server, 45 Tage, mehr als 1 Million MapReduce-Jobs
 - Tägliche Lastspitzen um Mittag und Mitternacht
 - Identifikation zweier Job-Klassen
 - *Interaktive*, zeitsensitive Jobs: Ad-hoc-Anfragen von Entwicklern
 - * Eingabedaten decken nur einen kleinen Teil der Gesamtdaten ab
 - * Viele dieser Jobs arbeiten auf denselben bzw. ähnlichen Eingabedaten
 - Nicht-zeitsensitive *Batch-Jobs*
- *Berkeley Energy Efficient MapReduce (BEEMR)*
 - Energieeinsparung durch Ausnutzung der Eigenschaften interaktiver Jobs
 - Gebündelte Ausführung von Batch-Jobs
- Literatur
 - Yanpei Chen, Sara Alspaugh, Dhruba Borthakur, and Randy Katz
Energy efficiency for large-scale MapReduce workloads with significant interactive analysis
Proceedings of the 7th European Conference on Computer Systems (EuroSys '12), S. 43–56, 2012.



BEEMR

Grundkonzept

- Interaktive Zone
 - Kleiner Teil der Rechner im Cluster
 - Durchgängiger Betrieb der Rechner
 - Vorrangige Bearbeitung interaktiver Jobs
- Batch-Zone
 - Restlicher Teil des Cluster
 - Normalzustand: Rechner im Energiesparmodus
 - Periodische Aktivierung aller Rechner zur Abarbeitung von Batch-Jobs
 - Rückkehr in den Energiesparmodus, sobald alle für die jeweilige Phase geplanten Batch-Jobs beendet wurden
- Start eines neuen Jobs
 - Klassifizierung mittels in der Studie ermittelter Schwellenwerte
 - Für interaktive Jobs
 - Überprüfung, ob Eingabedaten bereits in der interaktiven Zone verfügbar
 - Bei Bedarf: Holen der Eingabedaten während der nächsten Batch-Phase

