Hauptseminar „Inside Google“

Google Hardware Architektur

19. Dezember 2005

Lars Geiger

1 Einleitung

Ende September 2005 umfasste der Index von Google nach eigenen Angaben rund 8,1 Milliarden Dokumente und musste durchschnittlich fast 1000 Anfragen in jeder Sekunde bewältigen (siehe auch Tabelle 1). Jede dieser Anfragen liest dabei mehrere hundert Megabyte an Daten von Festplatten und verbraucht mehrere Milliarden Rechenzyklen. Die Suche im World Wide Web ist damit eine der Webanwendungen mit den höchsten Anforderungen an Rechenleistung und Speicherplatz.

Um diese Ansprüche möglichst wirtschaftlich erfüllen zu können, besitzt Google Computercluster, die mit den größten Supercomputer-Installationen der Welt vergleichbar sind, aber nur einen Bruchteil davon kosten. Im Folgenden soll gezeigt werden, woher die gegenwärtig große Beliebtheit von Clustersystemen stammt, wie Googles Cluster aufgebaut ist und welche Faktoren bei diesem Design eine Rolle gespielt haben.

2 Computercluster

Als Computercluster (oder kurz Cluster) bezeichnet man ein System aus vielen eigenständigen Rechnern (den „Knoten“), die über ein Kommunikationsnetz verbunden sind (im Gegensatz zu einer Kommunikation über gemeinsam genutzte Speicherbereiche), typischerweise zur Steigerung der Rechenleistung oder der Verfügbarkeit gegenüber einem einzelnen Rechner. Dabei ist darauf zu achten, dass keine einzelne Komponente eine solche Sonderstellung besitzt, dass ihr Ausfall das gesamte System außer Betrieb setzt (ein sogenannter „Single Point of Failure“); dies ist keine direkte Eigenschaft von Clustern, sondern sollte bei der Konzeption und Planung beachtet werden.

Datum	Index-Größe (Dokumente)	Anfragen/ Tag	Engine
Apr. 1994	110.000	1.500	WWW Worm
Nov. 1997	100.000.000	20.000.000	AltaVista
Dez. 2000	1.327.000.000	70.000.000	Google
Okt. 2005	8.168.684.336	81.646.533	Google

Tabelle 1: Wachstum des World Wide Web [Hennessy02, NN05]

Abbildung 1 zeigt beispielhaft einen Cluster, dessen Knoten über ein 1 Gbit schnelles Ethernet verbunden sind. Die einzelnen Knoten sind dabei mit einen oder zwei Prozesso-

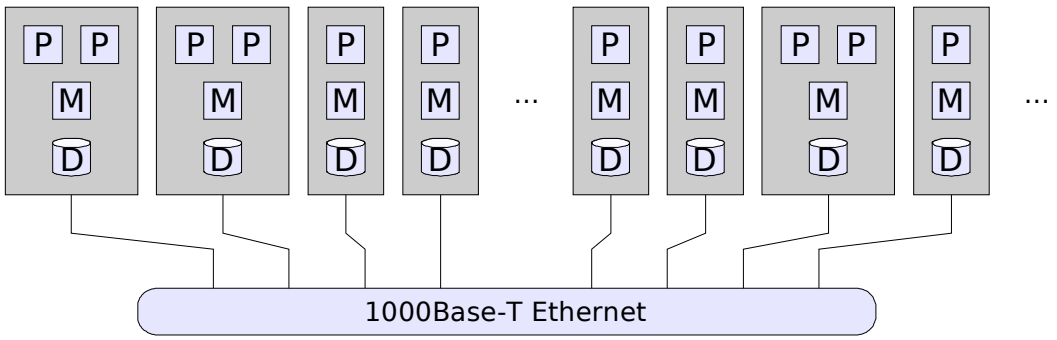

Abbildung 1: Cluster Organisation [Hennessy02]

ren (P), Arbeitsspeicher (M) sowie Festplatten als Massenspeicher (D) ausgestattet und laufen jeder mit einem eigenen Betriebssystem. Bei diesem Beispiel-Cluster ist das Verbindungsnetzwerk allerdings ein Single Point of Failure, ein Ausfall des Netzwerks könnte den Cluster außer Betrieb setzen.

Eine solche Anordnung ist zum einen toleranter gegenüber Ausfällen einzelner Rechner: indem Prozesse oder Anfragen zur Laufzeit auf die Knoten verteilt werden, können ausgefallene Systeme einfach umgangen werden. Dadurch ist der Betrieb mit etwas weniger Rechenleistung trotz der (unvermeidlichen) Defekte von Hard- und Software gewährleistet.

Zum anderen lässt sich mit Clustern eine gesteigerte Rechenleistung erreichen, falls ein Algorithmus parallel ausgeführt werden kann: der Algorithmus wird dazu in kleinere – möglichst parallel ausführbare – Teile zerlegt und dann von den Knoten berechnet, die Knoten synchronisieren dabei die Ausführung via Nachrichten über das Verbindungsnetzwerk. Dabei ist zu beachten, dass für die Rechenleistung eines Clusters mit n Knoten folgendes gilt: $B(n) \leqslant n \cdot B(1)$, wobei $B(n)$ die Rechenleistung eines Clusters mit n Knoten, $B(1)$ entsprechend die eines einzelnen Knoten ist. Im Allgemeinen ist sogar $B(n) < n \cdot B(1)$, da ein gewisser Overhead für die Verwaltung des Clusters anfällt.

Im Allgemeinen wird man beim Entwurf eines Clusters einen Trade Off zwischen möglichst hoher Rechenleistung und möglichst guter Verfügbarkeit eingehen müssen.

Sollte die Rechenleistung eines Clusters eines Tages nicht mehr ausreichen, so kann er einfach durch das Hinzufügen zusätzlicher Knoten erweitert werden, eine angenehme Eigenschaft um kommenden Herausforderungen begegnen zu können.

Von der wirtschaftlichen Seite betrachtet sprechen die niedrigen Anschaffungskosten für Computercluster. Für die Knoten eines Clusters können einfache Standard-PCs verwendet werden, wie sie in vielen Büros auf der ganzen Welt zum Einsatz kommen. Durch speziell für Hochleistungsrechen-

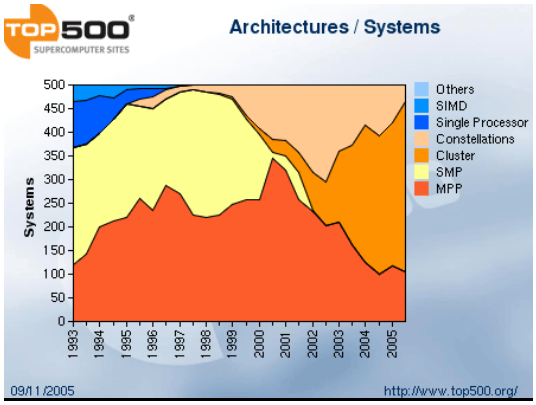

Abbildung 2: Zusammensetzung der TOP500 [TOP500]

anlagen entworfene Hardware ließe sich zwar prinzipiell eine höhere Leistung oder niedrigere Ausfallquoten erreichen. Dabei müssen aber die Entwicklungskosten auf die deutlich niedrigeren Stückzahlen umgelegt werden, wodurch die Geräte entsprechend teurer sind.

Den niedrigen Anschaffungskosten für Cluster stehen aber höhere Administrationskosten für Computercluster gegenüber: Ein Cluster mit n Knoten kostet in der Wartung rund das n-fache eines einzelnen Rechners, da im Prinzip jeder Knoten ein eigenständiges System ist. Bei einem – im Vergleich zu einem Cluster deutlich teureren – System mit n Prozessoren und vergleichbarer Rechenleistung (sofern ein System mit so vielen Prozessoren existiert) sind dagegen die Kosten für die Wartung kaum höher als bei einem einzelnen System mit nur einem Prozessor.

Cluster-Installationen sind sehr beliebt bei Systemen, bei denen es auf Leistung oder Ausfallsicherheit ankommt und machen zusammen mit den Constellations augenblicklich rund 80% der 500 schnellsten Supercomputer der Welt (TOP500) aus, wie man in Abbildung 2 sieht (als Constellation bezeichnet man dabei Cluster, bei denen die einzelnen Knoten mehr als einen Prozessor besitzen).

3 Der Google Cluster

Die folgenden Aspekte spielten eine wichtige Rolle bei der Entscheidung für einen Cluster und seinem endgültigen Design:

Eine Suchmaschine wie Google wird von Menschen auf der ganzen Welt benutzt und muss deshalb rund um die Uhr erreichbar sein. Außerdem entstehen durch den Betrieb der Suchmaschine laufende Kosten, die durch die Einnahmen gedeckt werden müssen. Fällt der Dienst durch einen Defekt aus, so entfallen die Einnahmen in dieser Zeit und es entstehen deutliche Verluste. Ausfälle von Hard- oder Software können zwar nicht verhindert werden, ein Cluster kann ausgefallene Knoten jedoch bis zur Beseitigung des Problems auslassen und Anfragen entsprechend an andere Knoten stellen. Dazu ist es jedoch nötig, dass Kopien eines Datenbestands auf mehreren Rechnern zur Verfügung stehen. Diese Technik wird auch als Replikation bezeichnet. Gleichzeitig entsteht dadurch auch ein höherer Aufwand, wenn die Daten verändert werden müssen (wenn Seiten geändert wurden oder neue hinzukamen), weil alle Kopien der Daten aktualisiert werden müssen.

Außerdem werden Suchmaschinen üblicherweise interaktiv von Menschen benutzt, deshalb muss darauf geachtet werden, die Geduld des Benutzers nicht zu überschreiten. Das erklärte Ziel von Google ist es, alle Anfragen in weniger als 0,5s zu beantworten. Um einen Datenbestand von über 8.000.000.000 Dokumenten effizient handhaben zu können und damit solch kurze Suchzeiten möglich zu machen, zerlegt Google den Suchindex in kleinere Teile und verteilt diese auf einzelne Knoten, wobei wie oben bereits angedeutet mehrere Knoten über identische Kopien einzelner Teile verfügen. Durch dieses – als Partitionierung bekannte – Aufteilen der Daten wird es möglich Suchanfragen hochgradig parallel abzuarbeiten, unter voller Ausnutzung der kombinierten Leistung der Knoten eines Clusters.

Im nächsten Kapitel wird gezeigt, wie eine Suchanfrage bei Google bearbeitet wird und welche logischen Aufgaben daran beteiligt sind.

3.1 Suchablauf

Wenn ein Benutzer eine Anfrage (http://www.google.com/search?q=stuttgart) an Google stellt, dann muss zuerst der Browser via DNS (Domain Name Service) den Namen www.google.com in eine IP Adresse auflösen. Bei der DNS Auflösung findet eine erste Lastverteilung statt, indem die Adresse eines geografisch möglichst nahe beim Benutzer gelegenen Data Centers zurückgegeben wird. Danach schickt er die Anfrage an den Google Web Server.

Im Folgenden übergibt der Google Web Server (GWS) die einzelnen Suchbegriffe parallel an die Index Server, die zu jedem Begriff eine Liste mit ID-Nummern von passenden Dokumenten suchen und zurückgeben (document IDs, kurz: docids). Gleichzeitig werden die Suchwörter von einem Spell checker überprüft, um mögliche Schreibfehler als „Meinten Sie …“ auf der Antwortseite alternativ anbieten zu können, und ein Ad Server sucht zu der Anfrage passende Anzeigen heraus und liefert sie an den GWS zurück.

Als nächstes fordert der Web Server die zu den docids gehörenden Dokumente von den Dokumenten Servern an. Nachdem er diese bekommen hat, werden die Ergebnisse bewertet, mit der entsprechenden Reihenfolge, kurzen Ausschnitten aus den jeweiligen Seiten, der Werbung und evtl. alternativen Schreibvorschlägen wird dann die Antwortseite der Google Suche erzeugt und an den Browser des Benutzers zurückgegeben. Abbildung 3 zeigt diesen Ablauf.

Wie man sieht, ist die Suche in viele kleine, parallel arbeitende Abläufe unterteilt. Außerdem kann man erkennen, welche Rollen ein einzelner Knoten im Google Cluster innehaben kann: Web Server, Index Server, Dokumenten Server, Ad Server oder Spell Checker. Bei Google ist die Funktion einem Knoten fest zugewiesen und ändert sich nicht – anders als bei Clustern üblich, bei denen Prozesse im Allgemeinen dynamisch zur Laufzeit auf den Knoten verteilt werden. In der Abbildung nicht zu sehen sind die Crawler, die das WWW kontinuierlich nach neuen oder geänderten

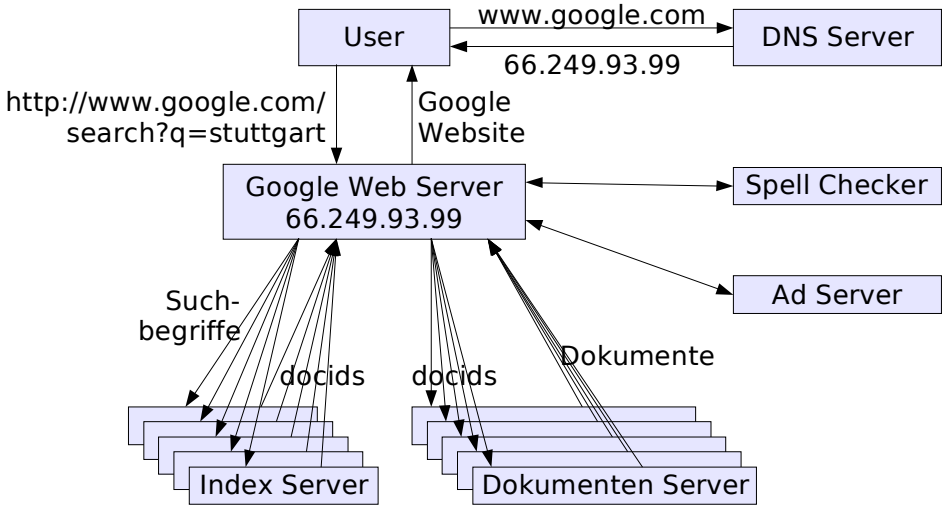

Abbildung 3: Google Suchablauf [Barroso03]

Seiten durchsuchen und den Index von Google entsprechend aktualisieren, sowie Systeme, die ein Load Balancing durchführen, also Anfragen nach Auslastung und Verfügbarkeit auf die vielen Google Web Server verteilen.

3.2 Knoten

Aus wirtschaftlichen Gründen setzt Google bei den Knoten seines Clusters auf Rechner, deren Ausstattung an normale Desktop PCs erinnert. Durch die hohe Parallelisierung und Redundanz des Google Clusters ist die Rechenleistung eines einzelnen Knotens nicht so wichtig, dass sie entsprechende Aufpreise für Server Hardware rechtfertigen würde. Im März 2000 beispielsweise kostete ein Pentium III 800MHz rund $800, ein Celeron 533MHz dagegen nur rund $200. Der Leistungsunterschied zwischen beiden Prozessoren rechtfertigte aber nicht den vierfachen Preis.

Im Jahr 2003 besaßen die Knoten des Google Clusters folgende Ausstattung:

- 1-CPU Celeron mit 533 MHz bis 2-CPU Pentium III mit 1,4 GHz
- DRAM (teilweise mit ECC zur leichteren Erkennung von Fehlern in den Speicherbausteinen)
- 80 GB IDE-Festplatte (Dokumenten-Server auch mehr)
- 100 Mbit Ethernet Anschluss zum Verbindungsnetzwerk
- Ein modifiziertes Redhat GNU/Linux als Betriebssystem

Durch diese relativ einheitliche Ausstattung ist es Google möglich, die bereits angesprochenen höheren Verwaltungskosten von Clustern zu senken. Außerdem wird bei den Knoten darauf geachtet, dass die Teile mit hohen Ausfallquoten (hauptsächlich Arbeitsspeicher, Netzteile und Festplatten) leicht zugänglich sind, um sie so ohne größeren Aufwand ersetzen zu können.

Die durchschnittliche Lebenszeit eines Rechners im Google Cluster liegt zwischen zwei und drei Jahren. Danach ist die Rechenleistung von Neugeräten so weit fortgeschritten, dass ein Austausch sinnvoll wird.

3.3 Racks

Normale PC Gehäuse eignen sich offensichtlich nicht für die wirtschaftliche Unterbringung einer großen Anzahl von Servern auf kleinstem Raum. Zu diesem Zweck werden die Rechner in sogenannte Racks oder Serverschränke montiert. Diese bis zu 2m hohen Schränke nehmen Geräte einer standardisierten Größe auf: die Geräte sind 19“ breit und 1 oder mehrere Höheneinheiten hoch (1 Höheneinheit = 1U = 1,75“). Die Google Racks nehmen 44 Höheneinheiten auf und enthalten auf Vorder- und Rückseite jeweils 40 PCs mit je 1U Höhe und

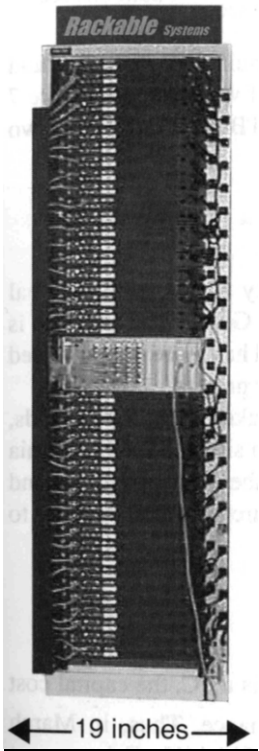

Abbildung 4: Google Rack [Hennesy02]

auf mittlerer Höhe einen 100 Mbit Ethernet Switch mit 4U. Dieser Switch verbindet die Server auf einer Seite des Racks miteinander und stellt über einen 1 Gbit schnellen Uplink die Verbindung zum Foundry Switch und damit zu anderen Racks her (auf die Vernetzung der Racks wird detailiert in Kapitel 3.4 eingegangen). Abbildung 4 zeigt ein Google Rack mit 20 Servern oben und unten. Links erkennt man die von jedem Rechner ausgehenden Netzwerkkabel, die zum Switch in der Mitte führen. Das von dem Switch ausgehende einzelne Kabel rechts ist einer der Uplinks, der dieses Rack unter den Bodenplatten mit mit dem Foundry Switch verbindet. Auf der rechten Seite sieht man außerdem die Stromkabel der einzelnen Rechner, ausgehend von einer zentralen Stromversorgung. Die Rückseite des Racks ist identisch bestückt und enthält einen weiteren Switch und 40 weitere Rechner.

Zwischen den Rechnern der Vorder- und Rückseite bleibt ein ca. 8cm großer Zwischenraum, über den mittels Lüftern die warme Abluft nach oben aus dem Rack abtransportiert wird.

3.4 Data Center

Im Jahr 2000 war der Google Cluster auf insgesamt drei Data Center verteilt: zwei davon befanden sich in Silicon Valley, eines in Virginia. Damit sollte sichergestellt werden, dass der Betrieb auch im Falle von Naturkatastrophen wie Erdbeben oder Überschwemmungen zumindest von einem Data Center fortgesetzt werden kann.

Abbildung 5 zeigt den Floorplan eines solchen Data Centers: die einzelnen Racks sind über eine Gbit Ethernet Leitung mit zwei redundanten Foundry Switches mit jeweils 128 Anschlüssen verbunden (zur Übersichtlichkeit sind in der Abbildung nur die Verbindungen des jeweils obersten und untersten Racks in einer Reihe eingezeichnet). Ein solches Data Center ist über eine OC48 Leitung mit 2488 Mbit/s Übertragungsrate ans Internet angebunden. Zur Sicherheit existiert noch eine OC12 Standleitung mit 622 Mbit/s, die zu einem anderen Google Center verbindet. Damit kann bei einem Ausfall der Internet-Verbindung ein Data Center immer noch über die OC12 Standleitung und die Internet-Verbindung eines anderen Data Centers Suchanfragen beantworten. Das Risiko, dass gleichzeitig die

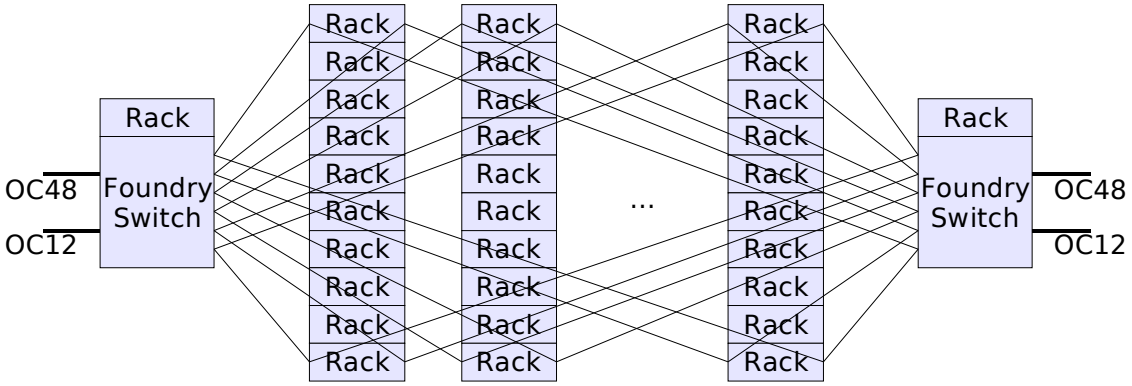

Abbildung 5: Google Cluster von oben [Hennesy02]

Verbindungen zweier Data Center ausfallen, ist gering, da die Leitungen von unterschiedlichen Providern bereitgestellt werden.

Angenommen, eine Antwort auf eine Suchanfrage ist üblicherweise 4000 Bytes groß, so ist der Bandbreitenbedarf von Google:

$$\text{Bandbreite} = \frac{81.646.533 \frac{\text{Anfragen}}{\text{Tag}} \cdot 4.000 \frac{\text{Bytes}}{\text{Anfrage}} \cdot 8 \frac{\text{bits}}{\text{Byte}}}{24 \cdot 60 \cdot 60 \frac{\text{s}}{\text{Tag}}} \approx \frac{2.612.689 \cdot 10^6 \text{bits}}{86.400 \text{s}}$$

$$\text{Bandbreite} \approx 30 \cdot 10^6 \frac{\text{bits}}{\text{s}} = 30 \frac{\text{Mbit}}{\text{s}}$$

Der Bandbreitenbedarf für das Crawling des Webs liegt durch die größere Anzahl abzurufender Dokumente zwar höher, aber immer noch innerhalb der verfügbaren Bandbreite der OC48 Leitung von 2.488 Mbit/s

Die beiden Racks bei den Foundry Switches enthalten einige PCs, die Load Balancing- oder Monitoring-Aufgaben ausführen, sowie Unterbrechungsfreie Stromversorgungen (USV). Zusätzlich zu den USVs besitzen die Data Center noch eine unabhängige Notstromversorgung über Diesel-betriebene Generatoren.

Mit dem gezeigten Aufbau erhält man folgende maximale Größe: Die beiden Foundry Switches besitzen je 128 Anschlüsse, ein Rack benötigt davon 2. Ein einzelnes Data Center in der angegebenen Konfiguration kann also 64 Racks aufnehmen. Bei 80 Servern in jedem Rack enthalten die Google Data Center also höchstens 5.120 Server. Im Jahr 2003 besaß der Google Cluster ca. 15.000 Knoten, verteilt auf die drei Data Center. Damit wäre praktisch die größte Ausbaustufe erreicht gewesen, eine weitere Vergrößerung würde entweder ein weiteres Data Center oder eine Veränderung des bestehenden Aufbaus erfordern.

4 Leistungsaufnahme

Die Leistungsaufnahme moderner Prozessoren steigt trotz neuer Technologien und Fertigungsprozesse stetig an. So blieb in den letzten drei Hardwaregenerationen des Google Clusters die Rechenleistung je Watt Leistungsaufnahme näherungsweise konstant bzw. mit der steigenden Rechenleistung stieg auch die Leistungsaufnahme des Clusters.

Ein aktueller x86-basierter Server in der unteren Preis- und Leistungsklasse kostet einen Preis von unter $2000 und verbraucht durchschnittlich 200W Leistung, in Spitzenzeiten auch 300W und mehr. Durch Verluste in den Netzteilen oder zusätzlichem Verbrauch für Kühlung steigt dieser Wert auf ungefähr das Doppelte an.

Angenommen, ein solcher Server ist insgesamt 3 Jahre in Betrieb und eine Kilowattstunde Strom kostet $0,09, so machen die Energiekosten bereits fast die Hälfte der Anschaffungskosten des Servers aus:

$$\text{Energiekosten gesamt} = 0{,}4\,\text{kW} \cdot 3 \cdot 365 \cdot 24\,\text{h} \cdot 0{,}09 \frac{\$}{\text{kWh}} = \$946{,}08$$

Deshalb setzt Google nach Möglichkeit energiesparende Technologien ein, allerdings muss darauf geachtet werden, dass die dadurch höheren Anschaffungskosten die eingesparten Energiekosten nicht übersteigen.

Ein weiteres Problem, das sich durch den höheren Energieverbrauch ergibt, ist die Leistungsdichte eines Google Racks. Schon bei normalen Serverracks liegt die Leistungsaufnahme bzw. die Wärmeabgabe eines Racks teilweise an den Grenzen dessen, was normale Data Center anbieten. Die bei Google eingesetzten Racks sind aber besonders dicht gepackt, um eine möglichst große Anzahl von Servern auf kleinstem Raum zu ermöglichen.

Ein Server, wie er bei Google beispielsweise im Jahr 2003 im Einsatz war, hatte einen Energieverbrauch von ca. 90W (zwei Pentium III mit 1,4GHz: 55W, Arbeitsspeicher und Mainboard: 25W, Festplatte: 10W). Bei einem Netzteil mit 75% Effizienz, ein üblicher Wert bei normalen PCs, bedeutete dies rund 120W Leistungsaufnahme für einen Server oder rund 10.000W für ein Rack mit 80 Servern. Bei einer Standfläche von 2m² liegt die Leistungsdichte eines solchen Racks damit bei $5000\frac{W}{m^2}$. Übliche Data Center bieten rund $2000\frac{W}{m^2}$, es entstehen für Google entsprechend zusätzliche Kosten für die Energieversorgung und die notwendige Kühlung.

Aus diesem Grund setzt Google große Hoffnungen auf Chip Multi-Processing Designs, wie z.B. die aktuell schon erhältlichen CPUs mit zwei Kernen auf einem Chip. So übertrifft zum Beispiel AMDs Opteron 275 mit zwei Kernen das entsprechende Modell mit nur einem CPU Kern, den Opteron 248, um ungefähr 80% bei einigen Benchmarks bei einer nur um 7% höheren Leistungsaufnahme. Damit ist die Energieeffizienz (Rechenleistung pro Watt) der CPU mit zwei Kernen deutlich besser als die der mit nur einem Kern.

5 Zusammenfassung

Cluster bieten eine hohe Leistung bei relativ geringen Kosten und konnten sich dadurch eine feste Stellung sichern, wenn es um Rechenleistung oder hohe Verfügbarkeit eines Systems bzw. Dienstes geht.

Google betreibt einen großen Materialaufwand für die Suche im World Wide Web. Der Google Cluster braucht den Vergleich mit anderen Hochleistungsrechnern nicht zu scheuen, es wird davon ausgegangen, dass es sich dabei um einen der größten Cluster in der Hand eines privaten Unternehmens handelt. Um einen so großen Computercluster wirtschaftlich entwerfen und betreiben zu können, ist eine genaue Analyse und eine sorgfältige Planung notwendig.

In den letzten Jahren wurde allerdings die steigende Leistungsaufnahme und die damit verbundenen höheren Anforderungen an Kühlsysteme zu einem immer größeren Problem, vor allem bei Clustern. Diesem Trend versucht man durch stromsparende Komponenten sowie der Entwicklung von Prozessoren mit mehreren Kernen auf einem Chip entgegen zu treten. Es bleibt abzuwarten, inwiefern es wirtschaftlich vertretbar sein wird, AMDs und Intels Dual-Core- und Multi-Core-

Prozessoren oder sogar Designs wie SPARCs Niagara in den Knoten des Google Clusters zu verwenden und wie sich solche Neuerungen auf den Energieverbrauch und die Rechenleistung auswirken werden.

6 Literatur

[Hennessy02] J. Hennessy, D. Patterson: „Computer Architecture“. 3rd edition, Morgan Kaufmann Publishers. 2002

[Barroso03] L. A. Barroso, J. Dean, U. Hölzle: „Web Search for a Planet: The Google Cluster Architecture“. IEEE Computer Society. 2003

[Cheney05] Matthew Cheney, Mike Perry: „A Comparison of the Size of the Yahoo! and Google Indices“. 2005 http://vburton.ncsa.uiuc.edu/indexsize.html

[NN05] Nielsen//NetRatings Press Release. 13.12.2005. http://www.netratings.com/pr/pr_051213.pdf

[TOP500] TOP500 Supercomputer Sites. http://www.top500.org

[Barroso05] L. A. Barroso: „The Price of Performance“. ACM Queue vol. 3, no. 7. 2005 http://acmqueue.com/modules.php?name=Content&pa=showpage&pid=330

[Ghemawat03] S. Ghemawat, H. Gobioff, S. Leung: „The Google File System“. SOSP'03. 2003