HiSbase: Informationsfusion in P2P Netzwerken*

Tobias Scholl Bernhard Bauer Richard Kuntschke Daniel Weber
Angelika Reiser Alfons Kemper

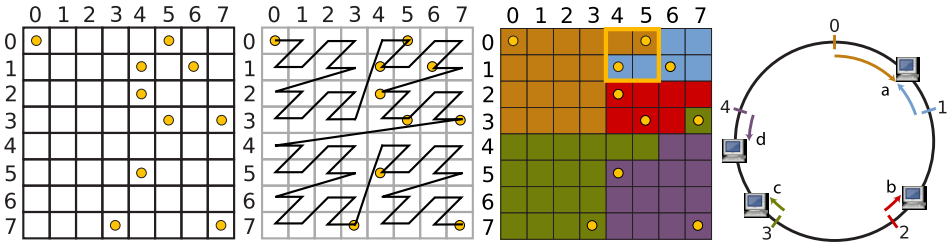
Technische Universität München
{scholl, bauerb, kuntschk, weberd, reiser, kemper}@in.tum.de

1 Einleitung und Motivation

E-Science Projekte vieler Fachrichtungen sehen sich mit den Herausforderungen einer stark wachsenden Datenflut konfrontiert. Die anwendungsspezifische, dynamische Fusion logisch verwandter Daten aus weltweit verteilten Quellen ist von höchstem wissenschaftlichen Interesse. Eine herkömmliche zentrale Datenhaltung stößt hierbei allerdings auf Grund der enormen Datenmengen und Transfervolumen an ihre Grenzen. HiSbase bietet einen Ansatz des Datenmanagements, der sowohl verfügbare Rechenleistung als auch freie Hauptspeicherkapazitäten innerhalb einer Forschungs-Community nutzt. Verteilte Hashtabellen (distributed hashtables, DHT) ermöglichen eine dezentrale und skalierbare Kommunikation und Datenhaltung. Für mehrdimensionale Daten verwenden wir eine Hashfunktion, welche die Daten gemäß einer *raumfüllenden Kurve* partitioniert. Dadurch bleibt die logische Nähe der Daten erhalten. Existierende Schieflagen (engl.: *skew*) der Daten gleichen wir durch ein *Verteilungs-Histogramm* aus. So verwaltet jeder Rechner einen ungefähr gleich großen Datencluster. Die Kombination dieser Technologien steigert die Leistungsfähigkeit der Anfragebearbeitung.

Wie in anderen Wissenschaftsbereichen, erwarten wir in der Astrophysik neben dem gewaltigen schon existierenden Datenvolumen exponentielle Wachstumsraten. Außerdem wird das Bedürfnis nach einer skalierbaren und effizienten Datenhaltung durch höhere Zugriffsraten verstärkt. In vielen Applikationen spielt die Fusion und Kombination von Beobachtungsdaten aus verschiedenen Datenquellen (die bspw. unterschiedliche Frequenzbereiche abdecken) eine Schlüsselrolle, um neue wissenschaftliche Erkenntnisse zu gewinnen. Das Erstellen von Wahrscheinlichkeitslandkarten für Galaxienhaufen oder die Klassifikation von spektralen Energieverteilungen sind derartige Anwendungen. Im Rahmen der deutschen e-Science und Grid Computing Initiative D-Grid wirken wir am Aufbau einer Plattform zur verteilten Informationsverarbeitung für die deutsche Astrophysik mit [KSH+04, KSKR05, KSH+06].

*Diese Arbeiten sind Teil des AstroGrid-D Projekts in der D-Grid Initiative und werden durch das Bundesministerium für Bildung und Forschung (BMBF) unter Vertrag 01AK804F und durch Microsoft Research Cambridge (MSRC) unter Vertrag 2005-041 gefördert.



(i) Datenraum (ii) Z-Kurve (iii) Regionen (iv) Peer-Zuordnung

Abbildung 1: Verarbeitungsschritte für zweidimensionale Beispieldaten.

2 HiSbase-Architektur

Die DHT-Struktur *Pastry* [RD01] verwaltet die HiSbase-Stationen und wickelt den Nachrichtenaustausch im Overlay-Netzwerk ab. Wie in Chord [SMK+01] verteilt Pastry die Daten gleichmäßig auf einen eindimensionalen, ringförmigen Schlüsselraum. Im Vergleich zu anderen DHT-Systemen ([SMK+01, RFH+01]) optimiert Pastry das Routing. In den ersten Phasen des Routings senden Peers bevorzugt Nachrichten zuerst an physische Nachbarn, was die Kommunikation über das Overlay-Netzwerk beschleunigt.

2.1 Einspeisung der Daten

Alle Peers beziehen ihre Datenobjekte direkt von den Datenzentren, die in HiSbase integriert werden. Eine Community-spezifische Verteilungsfunktion bestimmt auf welchen Peers die Daten (unabhängig vom Ursprungs-Datenarchiv) abgelegt werden. Jeder Peer hält die Daten aus allen Archiven, die in die von ihm verwalteten Regionen fallen. Bei der Anbindung an die Datenhaltung abstrahiert HiSbase von konkreten Datenbanksystemen. So können sowohl traditionelle als auch hauptspeicherbasierte Datenbanksysteme eingesetzt und verglichen werden.

2.2 Mehrdimensionale Daten und Raumfüllende Kurven

In vielen Bereichen der Wissenschaft werden mehrdimensionale Daten verwendet, z.B. in der Klimaforschung, in der Medizin und besonders in der Astronomie. Objekte aus Beobachtungsdaten, die eine benachbarte Position im astronomischen Koordinatensystem haben, werden hier als logisch benachbart betrachtet. Gängig ist die Verwendung des sphärischen Koordinatensystems, das die Koordinaten in Rektaszension und Deklination bezogen zum Erdmittelpunkt angibt. Abbildung 1 zeigt das Vorgehen in HiSbase an einem kleinen vereinfachten Beispiel. In Abbildung 1(i) sind die Datenobjekte des zweidimensionalen Datenraums als gelbe Punkte dargestellt.

Um bei der Abbildung auf den eindimensionalen Datenring die Datenlokalität zu erhalten, verwendet HiSbase *raumfüllende Kurven* wie die Hilbertkurve [Hil91] oder die Z-Ordnung [OM84]. Raumfüllende Kurven wurden als Indexstruktur für mehrdimensionale Datenbanken bereits intensiv erforscht [OM84, Mar99]. Abbildung 1(ii) zeigt, wie die Bereiche des Datenraums mit Hilfe der Z-Kurve linearisiert werden.

2.3 Daten-Lastbalancierung durch Histogramme

Oft ist mit einer Schieflage der Datenverteilung zu rechnen, da manche Bereiche intensiver untersucht wurden oder dichter besiedelt sind (Abbildung 1(i)). Diese Schieflage kann bei der Wahl des Histogramms berücksichtigt und ausgeglichen werden. Dieses fungiert als Verteilungsfunktion und ist allen Stationen bekannt. Die Objekte werden bei Equi-Depth Histogrammen [PIHS96] gleichmäßig auf Regionen verteilt. Regionen mit einer höheren Punktdichte sind in diesem Fall kleiner. In Abbildung 1(iii) enthält jede Region (zusammenhängende Abschnitte der raumfüllenden Kurve) zwei Datenpunkte.

Das Histogramm wird in einer Trainingsphase bestimmt. Die Trainingsdaten können die Gesamtdaten oder eine zufällige Auswahl sein. Auch durch andere Methoden wie „Biased Sampling“ [KGKB03] können die repräsentativen Daten gewonnen werden.

2.4 Regionenzuordnung

Den Regionen (Histogrammabschnitten) werden Identifikatoren der DHT-Struktur zugeordnet, die gleichmäßig über den Schlüsselraum verteilt sind. Die Peers werden zufällig auf diesem angeordnet. Die Wahrscheinlichkeit, dass ein Peer einer Region zugeordnet wird, ist somit für alle Regionen gleich, auch wenn sie unterschiedlich große Datenbereiche abdecken. Abbildung 1(iv) illustriert eine Zuordnung der fünf Regionen (0-4) auf vier Stationen (a-d). Die Regionenanzahl wird im realistischen Einsatz viel größer sein, um Dynamik in der Teilnehmerzahl zu ermöglichen. Die Identifikatoren der DHT-Struktur sind nicht dargestellt. Die Stationen kommunizieren über die Region-Identifikatoren, um immer den verantwortlichen Peer zu erreichen.

2.5 Regionen-basierte Anfragen

Jede HiSbase-Station kann regionen-basierte Anfragen stellen und berechnet bei der Anfrageanalyse die relevanten Regionen. Falls diese Station selbst eine relevante Region verwaltet, wird sie Koordinator. Sonst wird unter den verantwortlichen Peers ein Koordinator für die Anfragebearbeitung bestimmt. Der Koordinator kontaktiert die anderen relevanten Peers und fügt deren Ergebnisse zusammen. Die Beispielanfrage aus Abbildung 1(iii) kann Station a vollständig beantworten, da die Station zufällig für beide Regionen zuständig ist, die mit der Anfrageregion überlappen. Wir wollen in zukünftigen Arbeiten auch die Anfrageintensität bei der Bestimmung der Histogramme berücksichtigen. Überlastsituationen, die durch viele Anfragen auf die gleiche Region entstehen, könnten auch durch mehrere redundante Pastry-Ringe ähnlich wie in HotRod [PNT06] abgeschwächt werden.

3 Demonstrationsübersicht

Wir demonstrieren den HiSbase-Prototypen an einer Installation mit 16 Standard-PCs (jeweils 1,6 GHz Prozessoren und 512 MB RAM). Die Implementierung setzt auf der Java-Bibliothek *FreePastry* auf und verwendet als Datenbanksystem momentan IBM DB2 V8.1. Die Verarbeitung regionen-basierter SQL-Anfragen und deren Koordination über mehrere Stationen wird an einer Beispielanwendung aus der Astrophysik gezeigt. Wir vergleichen verschiedene Histogramm-Strategien hinsichtlich der Qualität des durch sie erreichten Clusterings und ihrer Effizienz.

4 Zusammenfassung und Ausblick

HiSbase ist ein P2P-Datenbanksystem, das es ermöglicht, Daten aus global verteilten Archiven effizient zu fusionieren. Dies haben wir an einem Beispiel aus der Astrophysik illustriert. Da regionen-basierte Anfragen dort eine zentrale Rolle spielen, werden raumfüllende Kurven für die Datenverteilung auf den eindimensionalen Datenraum des DHT-Systems verwendet. Equi-Depth Histogramme gleichen Schieflagen in den Daten aus und erzielen eine Balancierung der Datenlast über alle HiSbase-Stationen. Als nächste Schritte werden wir Leistungsanalysen einer weiträumig verteilten HiSbase-Instanz durchführen, um die Skalierbarkeit von HiSbase zu evaluieren. Wir vergleichen unterschiedliche Histogramme und bewerten Durchsatz und Antwortzeit unseres System bei vielen parallelen Anfragen. Bei einer ausreichend großen Anzahl an Peers erwarten wir durch den Einsatz von „reinen" hauptspeicherbasierten Datenbanksystemen einen noch höheren Effizienzgewinn, da alle relevanten Daten im Hauptspeicher gehalten werden können.

Literatur

[Hil91] D. Hilbert. Über die stetige Abbildung einer Linie auf ein Flachenstück. *Math. Ann.*, 38:459–460, 1891.

[KGKB03] G. Kollios, D. Gunopulos, N. Koudas und S. Berchtold. Efficient Biased Sampling for Approximate Clustering and Outlier Detection in Large Data Sets. *IEEE Transactions on Knowledge and Data Engineering*, 15(5), 2003.

[KSH+04] R. Kuntschke, B. Stegmaier, F. Häuslschmid, A. Reiser, A. Kemper, H.-M. Adorf, H. Enke, G. Lemson und W. Voges. Datenstrom-Management für e-Science mit StreamGlobe. *Datenbank-Spektrum*, 4(11):14–22, November 2004.

[KSH+06] R. Kuntschke, T. Scholl, S. Huber, A. Kemper, A. Reiser, H.-M. Adorf, G. Lemson und W. Voges. Grid-based Data Stream Processing in e-Science. In *Proc. of the IEEE Intl. Conf. on e-Science and Grid Computing*, Amsterdam, The Netherlands, Dezember 2006. Accepted for publication.

[KSKR05] R. Kuntschke, B. Stegmaier, A. Kemper und A. Reiser. StreamGlobe: Processing and Sharing Data Streams in Grid-Based P2P Infrastructures. In *Proc. of the Intl. Conf. on Very Large Data Bases*, Seiten 1259–1262, Trondheim, Norway, August 2005.

[Mar99] V. Markl. *MISTRAL: Processing Relational Queries using a Multidimensional Access Technique*. Dissertation, Technische Universität München, 1999.

[OM84] J. Orenstein und T. Merrett. A Class of Data Structures for Associative Searching. In *Proc. ACM SIGACT-SIGMOD Symp. on Principles of Database Sys.*, Seiten 181–190, Waterloo, Ontario, Canada, April 1984.

[PIHS96] V. Poosala, Y. E. Ioannidis, P. J. Haas und E. J. Shekita. Improved Histograms for Selectivity Estimation of Range Predicates. *SIGMOD*, 25(2):294–305, Juni 1996.

[PNT06] T. Pitoura, N. Ntarmos und P. Triantafillou. Replication, Load Balancing, and Efficient Range Query Processing in DHT Data Networks. In *EDBT*, 2006.

[RD01] A. I. T. Rowstron und P. Druschel. Pastry: Scalable, Decentralized Object Location and Routing for Large-Scale Peer-to-Peer Systems. In R. Guerraoui, Hrsg., *Middleware*, Jgg. 2218 of *Lecture Notes in Computer Science*, Seiten 329–350. Springer, 2001.

[RFH+01] S. Ratnasamy, P. Francis, M. Handley, R.M. Karp und S. Shenker. A Scalable Content-Addressable Network. In *SIGCOMM*, Seiten 161–172, 2001.

[SMK+01] I. Stoica, R. Morris, D. R. Karger, M. F. Kaashoek und H. Balakrishnan. Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications. In *SIGCOMM*, Seiten 149–160, 2001.