A Distributed Hash Table

by

Frank Dabek

S.B., Computer Science (2000); S.B. Literature (2000)
M.Eng., Computer Science (2001)
Massachusetts Institute of Technology

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology

September 2005

Author . . . .
Department of Electrical Engineering and Computer Science
November 4, 2005

Certified by . . . .
M. Frans Kaashoek
Professor
Thesis Supervisor

Certified by . . . .
Robert T. Morris
Associate Professor
Thesis Supervisor

Accepted by . . . .
Arthur C. Smith
Chairman, Department Committee on Graduate Students

A Distributed Hash Table

Frank Dabek

Submitted to the Department of Electrical Engineering and Computer Science
on November 4, 2005, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

DHash is a new system that harnesses the storage and network resources of computers distributed across the Internet by providing a wide-area storage service, DHash. DHash frees applications from re-implementing mechanisms common to any system that stores data on a collection of machines: it maintains a mapping of objects to servers, replicates data for durability, and balances load across participating servers. Applications access data stored in DHash through a familiar hash-table interface: `put` stores data in the system under a key; `get` retrieves the data.

DHash has proven useful to a number of application builders and has been used to build a content-distribution system [31], a Usenet replacement [115], and new Internet naming architectures [130, 129]. These applications demand low-latency, high-throughput access to durable data. Meeting this demand is challenging in the wide-area environment. The geographic distribution of nodes means that latencies between nodes are likely to be high: to provide a low-latency `get` operation the system must locate a nearby copy of the data without traversing high-latency links. Also, wide-area network links are likely to be less reliable and have lower capacities than local-area network links: to provide durability efficiently the system must minimize the number of copies of data items it sends over these limited capacity links in response to node failure.

This thesis describes the design and implementation of the DHash distributed hash table and presents algorithms and techniques that address these challenges. DHash provides low-latency operations by using a synthetic network coordinate system (*Vivaldi*) to find nearby copies of data without sending messages over high-latency links. A network transport (*STP*), designed for applications that contact a large number of nodes, lets DHash provide high throughput by striping a download across many servers without causing high packet loss or exhausting local resources. *Sostenuto*, a data maintenance algorithm, lets DHash maintain data durability while minimizing the number of copies of data that the system sends over limited-capacity links.

Thesis Supervisor: M. Frans Kaashoek
Title: Professor

Thesis Supervisor: Robert T. Morris
Title: Associate Professor

A Distributed Hash Table

Frank Dabek

Submitted to the Department of Electrical Engineering and Computer Science
on November 4, 2005, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

DHash is a new system that harnesses the storage and network resources of computers distributed across the Internet by providing a wide-area storage service, DHash. DHash frees applications from re-implementing mechanisms common to any system that stores data on a collection of machines: it maintains a mapping of objects to servers, replicates data for durability, and balances load across participating servers. Applications access data stored in DHash through a familiar hash-table interface: `put` stores data in the system under a key; `get` retrieves the data.

DHash has proven useful to a number of application builders and has been used to build a content-distribution system [31], a Usenet replacement [115], and new Internet naming architectures [130, 129]. These applications demand low-latency, high-throughput access to durable data. Meeting this demand is challenging in the wide-area environment. The geographic distribution of nodes means that latencies between nodes are likely to be high: to provide a low-latency `get` operation the system must locate a nearby copy of the data without traversing high-latency links. Also, wide-area network links are likely to be less reliable and have lower capacities than local-area network links: to provide durability efficiently the system must minimize the number of copies of data items it sends over these limited capacity links in response to node failure.

This thesis describes the design and implementation of the DHash distributed hash table and presents algorithms and techniques that address these challenges. DHash provides low-latency operations by using a synthetic network coordinate system (*Vivaldi*) to find nearby copies of data without sending messages over high-latency links. A network transport (*STP*), designed for applications that contact a large number of nodes, lets DHash provide high throughput by striping a download across many servers without causing high packet loss or exhausting local resources. *Sostenuto*, a data maintenance algorithm, lets DHash maintain data durability while minimizing the number of copies of data that the system sends over limited-capacity links.

Thesis Supervisor: M. Frans Kaashoek
Title: Professor

Thesis Supervisor: Robert T. Morris
Title: Associate Professor

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.

WHAT WILL YOU BUILD? | **sales@docketalarm.com** | **1-866-77-FASTCASE**

fastcase®
Smarter legal research.