University of Zurich[UZH]

BACHELOR THESIS – Communication Systems Group, Prof. Dr. Burkhard Stiller

Design and Prototypical Implementation of ...

Author
City, Country
Student ID: 00-711-999

Supervisor: ...
Date of Submission: January 1, 2006

University of Zurich
Department of Informatics (IFI)
Binzmühlestrasse 14, CH-8050 Zürich, Switzerland

ifi

Bachelor Thesis
Communication Systems Group (CSG)
Department of Informatics (IFI)
University of Zurich
Binzmühlestrasse 14, CH-8050 Zürich, Switzerland
URL: http://www.csg.uzh.ch/

Abstract

Das ist die Kurzfassung...

Acknowledgments

Optional

Contents

Chapter 1

Introduction

1.1 Motivation

Big data blabla

1.2 Description of Work

1.3 Thesis Outline

Chapter 2

Related Work and State of the Art

Although the widest usage of the MapReduce programming model to date may be the one of Hadoop [1] with its centralised master-slave architecture, recently, there have also been a number of attempts to transport it to a more decentralised setting in an endeavour to support currently popular Cloud platforms, pervasive grids, and/or mobile environments.

[4] recognised the problem of centralised master-slave architectures to not cope well with dynamic Cloud infrastructures, where nodes may join or leave the network at high rates. Thus, they introduce a peer-to-peer model to manage node churn but also master failures and job recovery in a decentralised way. Each node may become a master or a slave at any given time dynamically, preserving a certain master/slave ratio. Slaves are assigned tasks to perform by the masters, which handle management, recovery, and coordination. To reduce job loss in case of master failures, each master may act as a backup master for a certain job, only executing it if the master primarily responsible for that job fails. Overall, the structure of different entities resembles very much the one of Hadoop, simply ported to a P2P setting. They were able to show that such an implementation provides better fault tolerance levels compared to a centralised implementations of MapReduce and only limited impact on network overhead.

In a very similar direction goes the idea of CloudFIT [8], which the authors advertise as a "Platform-as-a-Service (PaaS) middleware that allows the creation of private clouds over pervasive environments". The idea is to use private computers to set up a private "Cloud", and that MapReduce jobs are then distributed to this environment and executed in a P2P fashion. CloudFIT is, thus, built to support various P2P overlays and the authors show the performance of CloudFIT with both PAST and TomP2P against Hadoop. The results demonstrated CloudFIT to be able to achieve similar execution speeds as Hadoop while omitting the need of a dedicated cluster of computers. Data in CloudFIT is directly stored within the DHT of the corresponding overlay, such that it is replicated by a factor of k. Although such an implementation may not guarantee n-resiliency (meaning, replicating the data on each node), it is a reasonable assumption that fault tolerance may be ensured more than enough (provided at least a number of nodes stay alive) while drastically improving storage performance. Furthermore, not all the data is broadcasted but only the keys of each task, further reducing the network usage the authors showed in another

study [9] to be one of the main problems in distributed environments for the performance of MapReduce tasks.

Chapter 3

Methods and Prototype

In this section, the used methods are laid out and the implemented prototype is described.

3.1 Methods used

3.1.1 MapReduce

At the very core of this thesis lies the idea of MapReduce, a programming model made popular by Google and presented in [3]. Having large data sets to analyse that may not be handled by only few computers, like the millions of Websites gathered by Google each day that have to be indexed fast and reliably, MapReduce provides a way of automatic parallelisation and distribution of large-scale computations. Users only need to define two types of functions: a map and a reduce function. Each computation expressed by these functions takes a set of input key/value pairs and produces a set of output key/value pairs. A map function takes an input pair and produces a set of *intermediate* key/value pairs. Once the intermediate values for each intermediate key I are grouped together, they are passed to the reduce function. The reduce function then takes an intermediate key I and the corresponding set of intermediate values (usually supplied as an iterator to the reduce function) and merges them according to user-specified code into a possibly smaller set of values. The resulting key and value are then written to an output file, allowing to handle lists of values that are too large to fit in memory.

A very simple yet often used MapReduce operation is to count words in a collection of documents, the so called *WordCount*. Below pseudocode demonstrates how one can count all the words in documents using map and reduce functions.

```
function MAP(String key, String value)
    // key: document name, value: document content
    for word w in values do
        EmitIntermediate(w, "1");
    end for
end function
```

```
function REDUCE(String key, Iterator values)
    // key: a word, values: a list of "1"'s
    int result = 0;
    for v in values do
        result += ParseInt(v);
    end for
    Emit(AsString(result));
end function
```

The map function splits the received text (*value*) into all corresponding words and for each word, emits its occurrence count (simply a 1). The reduce function then takes, for each word (*key* of the reduce function) individually, these occurrences (*values*) and sums them up, eventually emitting for each word the sum of occurrences (*result*). Although above pseudocode suggests inputs and outputs to be strings, the user specifies the associated types. This means that input keys and values may be of a different domain than the output keys and values (note that the *intermediate* keys and values are from the same domain as the output keys and values). Conceptually, this is expressed as follows:

map	(k1, v1)	→ list(k2, v2)
reduce	(k2, list(v2))	→ list(v2)

The traditional way to view the workflow in a MapReduce program is indicated by Figure 3.1. A master schedules execution and assigns tasks to workers, which then perform the actual map or reduce on the assigned data. Input data is partitioned into M splits with a size of around 16-64 MB per piece to be processed in parallel by different machines, and the intermediate key space is partitioned into R pieces, where R usually is the number of available workers for executing reduce but may also be specified by the user. To evenly distribute these pieces, a partitioning function like *hash(key) mod R* may be used. In such a setting, same intermediate keys may not be grouped together, which is done before the reduce action is performed. Additional combination in between may be applied to reduce the amount of data to be distributed among the workers.

3.2 Peer-to-Peer Overlay Networks

3.2.1 Definitions and Flavours

Peer-to-peer (P2P) systems and applications are distributed systems without any centralized control or hierarchical organisation, where the software running at each node is equivalent in functionality [10]. Instead of a central server (as present in Client/Server architectures [7]), many hosts (usually referred to as peers) that posses desired contents also handle user requests for them [2]. According to [7], a P2P network is given if the participants share a part of their own hardware resources (like e.g. processing power or storage capacity), which provide the service and content offered by the network and which are directly accessible by other peers. Peers are, thus, both resource (service and content) providers and requesters. There exist different flavours of P2P overlay networks but the

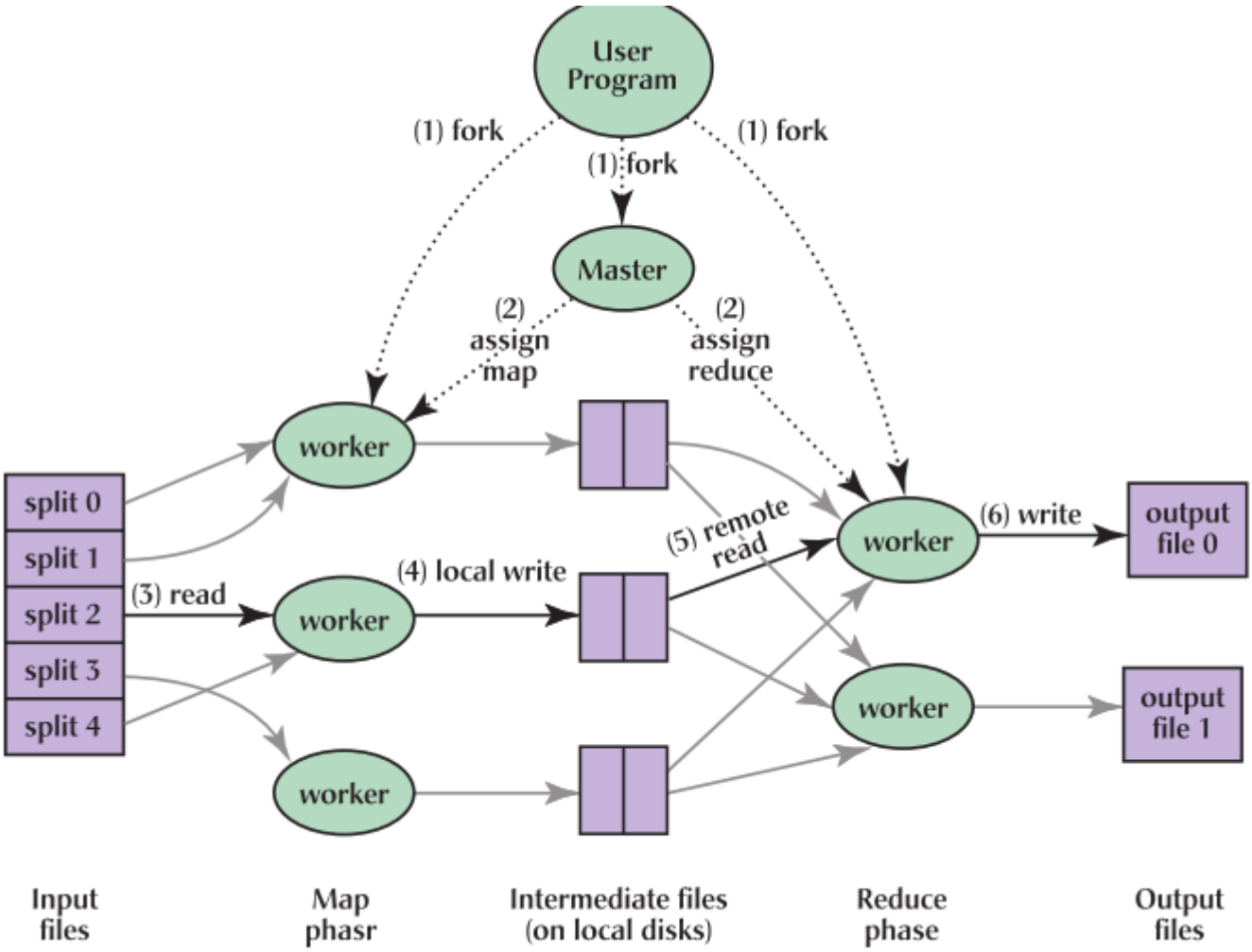

Figure 3.1: Execution overview in a typical MapReduce program

overall themes may be summarised as "pure" P2P and "hybrid" P2P systems. In a *pure P2P system*, any entity may be removed without the network suffering any service loss. A *hybrid P2P system*, on the other hand, requires a central entity to provide parts of the offered networks. An example of such a hybrid system was Napster[2].

3.2.2 Discovering Content and Distributed Hash Tables

The definitions given above demonstrate the need for efficiently discovering content and/or services without using a central server. Central servers may provide O(*1*) lookup of a file's location in the P2P network but also suffers from the problem of being a single point of failure (SPOF), which made Napster very efficient for lookup but also vulnerable for lawsuits[2] and certainly not scalable at that time[6]. However, although being rather resilient to node crashes, without efficient lookup capabilities, P2P networks are *unstructured* and as there is no constraint on where files are placed in the overlay network, queries need to be flooded across the overlay to find the location of the desired content (requiring O(n) lookups). Such overlay networks like e.g. Gnutella were, therefore, not scaling. To overcome this, distributed hash tables (DHTs) were introduced, of which the Content-Addressable Network (CAN, [6]) may be one of the first such design outlines intended to be used in P2P networks. Keys are mapped onto values in a DHT, operations include (among others) typical hash table or hash map operations like put(*key*, *value*), get(*key*): *value*, etc. The idea is that the problem of P2P systems is not the file transfer process which is inherently scalable but finding the peer from whom to retrieve the file[6]. Thus the indexing scheme needs to be scalable to make the whole P2P system scalable. Importantly, it requires no centralized control and there is no rigid hierarchical naming structure like e.g. in IP or DNS routing. DHTs make overlay networks *structured* and provide hash-table-like semantics on internet-like scales [6]. Efficient routing protocols (e.g. Chord[10], Kademlia[5]) reduce lookup complexity to O(log n) while lowering node state (the routing entries to other nodes) to O(log n) as well. Although there are some limitations to DHTs (e.g. a high churn rate requires O(log n) repair operations, keyword searches DHTs by default may only handle exact-match lookups, and many queries may not require an exact recall as they are mostly for well-replicated files and thus, do not justify the overhead of a DHT[2]), they provide a good balance between node state and communication overhead. As TomP2P uses Kademlia as its DHT, its functionality will be briefly outlined next. For more detailed information, please consult the accompanied references.

Kademlia

Kademlia is a P2P DHT with XOR-based metric topology [5]. It combines provable consistency and performance, latency-minimising routing, and a symmetric, unidirectional topology. It minimises number of configuration messages nodes must send to learn about each other, and configuration information spreads automatically as a side-effect of key lookups. Furthermore, nodes have enough knowledge and flexibility to route queries through low-latency paths. Kademlia contacts only O(log n) nodes while searching the system containing n nodes. Keys are 160-bit opaque quantities of which each participating

computer is assigned one node ID in the 160-bit key space. Key-value pairs are stored on nodes with IDs "close" to the key. Closeness is defined as the bitwise exclusive or (XOR) distance of two nodes' IDs (d(x, y) = x $\oplus$ y, where x and y are two 160-bit identifiers). A node's ID is a large random number intended to achieve uniqueness. Thus, distance is not used in a geographical sense as "neighbour" nodes may be spread around the world and are only logically considered close in the overlay network due to their small XOR distance. Kademlia treats nodes as leaves in a binary tree, with each node's position determined by the shortest unique prefix of its ID. In a fully-populated binary tree of 160-bit IDs, the magnitude of the distance between two IDs is the hight of the smallest subtree containing them both. If the binary tree is not fully populated, the closest leaf to an ID is the leaf whose ID shares the longest common prefix. For every node, the binary tree is divided into a series of successively lower subtrees that do not contain the node. Every node knows at least one node in each of its subtrees if the subtree contains a node. By successfully querying the known node, contacts are found in lower subtrees until the lookup finally converges to the target node.Thus, any node can locate any other node by its ID. A Kademlia node stores contact information about each other to route query messages for nodes of distance 2^i and 2^{i+1} (called k-buckets). k-buckets are sorted by time last seen (least-recently (head) to most-recently (tail) Appropriate k-buckets are updated whenever a Kademlia node receives any message (request or reply) from another node.

3.3 Prototype

Chapter 4

Evaluation

Chapter 5

Summary and Conclusions

Bibliography

[1] Autoren: Titel, Verlag, `http://...`, Datum.

Abbreviations

AAA	Authentication, Authorization, and Accounting

Glossary

Authentication

Authorization Authorization is the decision whether an entity is allowed to perform a particular action or not, e.g. whether a user is allowed to attach to a network or not.

Accounting

List of Figures

List of Tables

Appendix A

Installation Guidelines

Appendix B

Contents of the CD

Bibliography

[1] Welcome to Apache Hadoop!

[2] Yatin Chawathe, Sylvia Ratnasamy, Lee Breslau, Nick Lanham, and Scott Shenker. Making gnutella-like P2P systems scalable. *Proceedings of the 2003 conference on Applications technologies architectures and protocols for computer communications SIGCOMM 03*, 25:407, 2003.

[3] Jeffrey Dean and Sanjay Ghemawat. MapReduce. *Communications of the ACM*, 51(1):107, jan 2008.

[4] Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. P2P-MapReduce: Parallel data processing in dynamic Cloud environments. *Journal of Computer and System Sciences*, 78(5):1382–1402, sep 2012.

[5] Petar Maymounkov and David Mazières. Kademlia: A peer-to-peer information system based on the xor metric. In *Revised Papers from the First International Workshop on Peer-to-Peer Systems*, IPTPS '01, pages 53–65, London, UK, UK, 2002. Springer-Verlag.

[6] Sylvia Ratnasamy, Paul Francis, Mark Handley, Richard Karp, and Scott Shenker. A scalable content-addressable network. *SIGCOMM Comput. Commun. Rev.*, 31(4):161–172, August 2001.

[7] R. Schollmeier. [16] a definition of peer-to-peer networking for the classification of peer-to-peer architectures and applications. In *Proceedings of the First International Conference on Peer-to-Peer Computing*, P2P '01, pages 101–, Washington, DC, USA, 2001. IEEE Computer Society.

[8] Luiz Angelo Steffenel and Manuele Kirch Pinheiro. CloudFIT , a PaaS platform for IoT applications over Pervasive Networks. 2015.

[9] Luiz Angelo Steffenel and Manuele Kirch Pinheiro. Leveraging Data Intensive Applications on a Pervasive Computing Platform: The Case of MapReduce. *Procedia Computer Science*, 52:1034–1039, 2015.

[10] Ion Stoica, Robert Morris, David Karger, M. Frans Kaashoek, and Hari Balakrishnan. Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications. *Conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM '01)*, pages 149–160, 2001.