Map Reduce on a Chord Distributed Hash Table

Andrew Rosen Brendan Benshoof Matt Erwin Anu Bourgeois
Department of Computer Science, Georgia State University
34 Peachtree St NW
Atlanta, Georgia 30303
rosen@cs.gsu.edu

***Abstract*—Large problems in Computer Science are often able to be split into many smaller, functionally identical tasks, the results of which can be reduced into the desired solution. MapReduce is a framework for performing distributed computations of these types of problems, but most of these frameworks are strongly dependent on a heirarchical structure. Chord is a distributed hash table that provides $\log_2 n$ connectivity among processors. Our experiments show that a Chord ring has number of desirable advantages for implementing MapReduce, such as a lack of a single point of failure, fault-tolerance, an even distribution of work, scalability, and minimal overhead.**

I. Introduction

Ever since Google published a paper on the subject [5], MapReduce has rapidly become an integral part in the world of data processing. Using MapReduce, a user can take a large problem, split it into small, equivalent parts and send those parts off to be worked on by other processors. These results of these computations are then sent back to the user and combined until one large answer results. Numerous programming tasks - word counts, reverse indexing, sorting, Monte-Carlo approximations - can be efficiently distributed using MapReduce [5].

The most popular platform for MapReduce is Hadoop [2]. Hadoop is an open-source Java implementation deveolped by Apache and Yahoo! [12]. Haddop has two components, the Hadoop Distributed File System (HDFS) and the Hapdoop MapReduce Framework [8]. Under HFDS, nodes are arranged in a heirarchical tree, with a master node, called the NameNode, at the top. The NameNode is responsible for keeping track of whcih DataNodes posess which files, as well as coordinating the work done under a MapReduce job **double check the latter**.

However, what if we desire a less heirarchical structure among our nodes? A single node in charge is a single point of failure. We need a system that can scale, is fault tolerant, has a minimal amount of latency, and distributes files evenly.

Chord [16] is a distributed hash table (DHT) that possesses these qualities. Chord guarantees a worst-case $\log n$ lookup time for a particular node or file in the network. It is highly fault-tolerant to node failures and churn. It scales extremely well and there is little maintenance required by the network as a whole to handle individual nodes. Files in the network are distributed evenly among its members.

However, rather that viewing Chord solely as a means for sharing files, we see it as a means for distributing work. We have developed a system, called CHRONUS, that builds off the concepts that make Chord a powerful means of evenly distributing files and applies it toward distributing work among member nodes[1]. We built a framework on CHRONUS for performing Map and Reduce tasks on Chord ring. CHRONUS leverages the underlying protocol to distribute map and reduce tasks to nodes evenly, provide greater data redundancy, and guarantee a greater amount of fault tolerance. The network has no single point of failure.

At the same time we avoid the architectural and file system constraints of systems like Hadoop. Nodes in CHRONUS can be setup in a cluster for high performance or they can be deployed the Internet, for volunteer computing tasks. Our experiments demonstrate that our framework is highly scalable, solving problems significantly faster when distributed. The larger the problem is, the greater the speedup gained incorperating by incorperating more nodes into the problem.

Section X covers the specifics of the Chord Protocol and the design choices we made for implementation. We describe MapReduce and Hadoop in more detail in Section QA4162. Related Work is discussed in Section 42. Details of CHRONUS's implementation and code is described in Section Q, while our experiments and their results are covered in Section Z. Lastly, Section L discusses fruitful avenues of future research.

II. Chord

The Chord protocol [16] takes in some key and returns the identity (ID) of the node responsible for that key. These keys are generated by hashing a value of the node, such as the IP address and port, or by hashing the filename of a file. The hashing process creates a m-bit hash identifier.

The nodes are then arranged in a ring from the lowest hash-value to highest. Chord then takes the hashed files and places each in the node that has the same hashed identifier as it. If no such node exists, the node with the first identifier that follows this value. This node responsible for the key κ is called the $successor$ of κ, or $successor(\kappa)$. Since the overlay is a circle, this assignment is computed in module 2^m space. For example, if there were some portion of the network with nodes 20, 25, and 27, node 25 could be responsible for the files with the keys (21,22,23,24,25). If node 25 were to decide to leave the network, it would inform node 27, who would then be

[1] repetitive

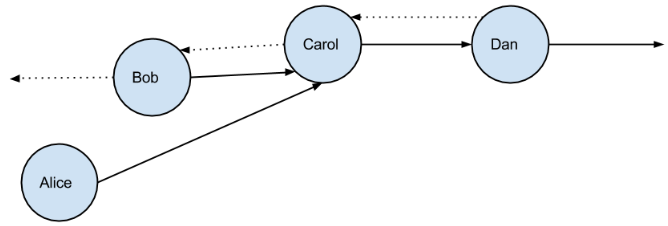

Fig. 1: Alice has incorrectly determined that Carol is her appropriate successor. When Alice stabilizes, Carol will let her know about Bob.

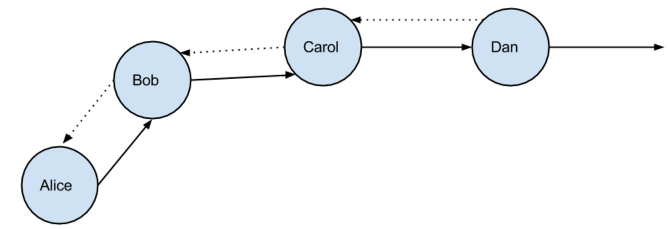

Fig. 2: After completing stabilize, Alice makes Bob her successor and notifies him. Bob then made Alice as his predecessor.

responsible for all the keys node 25 was covering. An example Chord network is drawn in in Figure INSERT FIG HERE.

With this scheme, we can reliably find the node responsible for some key by asking the next node in the circle for the information, who would then pass the request through the circle until the successor was found. We can then proceed to directly connect with the successor to retrieve the file. This naive approach is largely inefficient, and is a simplification of the lookup process, but it is the basis of how Chord theoretically works.

To speed up the lookup time, each node builds and maintains a *finger table*. The *finger table* contains the locations of up to m other nodes in the ring. The ith entry of node n's *finger table* corresponds to the node that is the $successor(n + 2^{i-1})$ $mod\ 2^m$. Because hash values won't be perfectly distributed, it is perfectly acceptable to have duplicate entries in the *finger table*.

When a node n is told to find some key, n looks to see if the key is between n and $successor(n)$ and return $successor(n)$'s information to the requester. If not, it looks for the entry in the finger table for the closest preceding node n' it knows and asks n' to find the successor. This allows each step in the to skip up to half the nodes in the network, giving a $\log_2(n)$ lookup time. Because nodes can constantly join and leave the network, each entry in the table is periodically checked and updated.

To join the network, node n first asks n' to find $successor(n)$ for it. Node n uses the information to set his successor, but the other nodes in the ring will not acknowledge n's presence yet. Node n relies on the stabilize routine to fully integrate into the ring.

The stabilize routine helps the network integrate new nodes and route around nodes who have left the networls. Each node periodically checks to see who their successor's predecessor is. In the case of a static network, this would be the checking node. However, if the checking node gets back a different node, it looks at that returned node's hash value and changes their successor if needed. Regardless of whether the checking node changes its successor, that node then notifies the (possibly) new successor, essentially telling the successor "based on the information I have, I'm your predecessor. Check to see if you need to update your predecessor information," to which the successor obliges. A more concrete example:

Suppose Alice, Bob, Carol, and Dan are members of the ring and everyone happens to be ordered alphabetically (Figure 1). Alice is quite sure that Carol is her successor. Alice asks Carol who her predecessor is and Carol says Bob is. Since Bob is closer than Carol, Alice changes her successor to Bob and notifies him.

When Bob sees that notification, he can see Alice is closer than whoever his previous predecessor is and sets Alice to be his predecessor. During the next stabilization cycle, Alice will see that she is still Bob's predecessor and notify him that she's still there (Figure 2).

One of the major design choices for Chord implementation is not figuring which node is responsible for a given key, but figuring out who decides which node is responsible for a given key. In our implementation, a node n is responsible for the keys ($predecessor(n)$, n]. In other words, when n gets a message, it considers itself the intended destination for the message if the message's destination hash is between $predecessor(n)$ and n. A node that does not have a predecessor assigns itself as its own predecessor and considers itself responsible for all messages it receives.

When a node n changes his successor, n asks if the successor is holding any data n should be responsible for. The successor looks at all the data n is better suited to hold onto, packages it up, and sends it along to n. The successor does not have to delete this data. If fact, keeping this data as a backup is beneficial to the network as a whole, as n could decide to leave the network at any point.

Due to the potentially volatile nature of a peer-to-peer network, Chord has to be able to handle (or at the very least, tolerate) an arbitrary amount of churn. We already detailed how Chord gradually guides nodes into their correct locations after they join the network. The same is true for when a node leaves the network; the stabilize procedure will guide nodes to their correct successors and predecessors. However, we can exert more control over how to handle nodes leaving the network

A node can leave the ring in one of two ways. A node can either suddenly drop out of existence, or a node can tell the network he is about to leave, letting his successor and predecessor immediately perform the needed changes.

When a node politely quits, he informs both his successor and predecessor and gives them all the information they need to fill the gap that would be left over. He also sends all of the data he is responsible for to his successor, who would become responsible for that data when that node left. Fingers that pointed to that node would be corrected during the finger maintenence period. This allows for the network to adjust to the change with a minimum of fuss.

Unfortunately, it is impossible that every time a node leave the network it will do so politely. If a node suddenly quits, the data it had stored goes with it. To prevent data from becoming irretrievable, a node can periodically send backups to its successor. So as not to overwhelm the ring with a cascade of backups of backups, the node only passes along what it considers itself responsible for, which changes as nodes enter and leave the network. If the backup leaves, he send his stuff to his successor, since the backup's successor would be the one responsible for the info now.

A. Extensions of Chord

The Cooperative File System (CFS) is an anonymous, distributed file sharing system built on top of Chord [4]. In CFS, rather than storing an entire file at a single node, the file is split up into multiple chunks around 10 kilbytes in size. These chunks are each assigned a hash and stored in nodes corresponding to their hash in the same way that whole files are. The node that would normally store the whole file instead stores a *key block*, which holds the hash address of the chunks of the file.

The chunking allows for numerous advantages. First, it promotes load balancing. Each piece of the overall file would (ideally) be stored in a different node, each with a different backup or backups. This would prevent any single node from becoming overwhelmed from fulfilling multiple requests for a large file. It would also prevent retrieval from being bottlenecked by a node with a relatively low bandwidth. Finally, when Chord uses some sort of caching scheme like that described in CFS [4], caching chunks as opposed to the entire file resulted in about 1000 times less storage overhead.

Chunking also opens up the options for implementing additional redundancy such as erasure codes[13]. With erasure codes, redundant chunks are created but any combination of a particular number of chunks is sufficient to recreate the file. For example, a file that would normally be split into 10 chunks might be split into 15 encoded chunks. The retreival of any 10 of those 15 chunks is enough to recreate the file. Implementing erasure codes would presumably make the network more fault tolerant, but that is an exercise left for future work.

III. MapReduce and Hadoop

1) What is it: MapReduce became a major topic of interest in 2004 when Google published a paper detailing their implementation, called MapReduce [5].

At its core, MapReduce [5] is a system for division of labor, providing a layer of speration between the programmer and the nastier parts of parallel processing. The programmer sends a large task to a master node, who then divides that task among slave nodes (which may further divide the task). This task has two distinct parts: Map and Reduce. Map performs some operation on a set of data and then produces a result for each map operation. This intermediate data can then be reduced, combining these sets of intermediate data into a set, which is further combined with other sets. This process continues until one set of data remains.

The classic example given for MapReduce is counting the occurrence of each word in a collection of documents. The master node splits up the documents into multiple chunks and sends them off to workers. Each worker then goes through each chunk and creates a small word frequency list. These lists are then used by other workers, who combine them into larger and larger lists, until the master node is left with a word frequency list of all the words in the documents.

One very popular implementation of MapReduce is Apache's Hadoop. [2].

2) Requirements for a viable MapReduce platform: scalability and fault-tolerence.

Chord is highly scalable and fault tolerent. Chord guarantees a latency of $\log(n)$, making it extremely scalable. Chord's ability to automatically assign nodes responibilty as the ring adds and loses nodes makes it extremely fault tolerant.

3) Why are we an interesting and viable alternative to Hadoop: Here is a list.

- Much easier to program on ours? Citation of complaints.
- Hadoop has a specific architecture. Ours is really easy to setup. Step 1: add node to ring. Step 2: there is no step 2.
- Hadoop is bound to it's filesystem. For our system to use a database, make the query during stage(). Or if part of the job requires making queries to the database, just write those calls in.
- If I'm reading this correctly, Hadoop keeps track of where the data is stored and asks the node where the data is stored to do the map tasks. In CHRONUS we don't make any assumption about the data, including where it's coming from. This means we can lose time feeding in our data
- Hadoop has master node bottle neck issue [CITATION]. Ours can have that happen but it's much less of an issue because we can collate on the way back. This would take extra $waittime \log(n)$ time for the reduce step.
- Hadoop is more latency tolerant. There are ways to work around that. BRENDAN EXPOUND HERE.
- I seem to get the implication from [8] that Hadoop must wait for all maps to finish before reducing. CHRONUS doesn't.
- CHRONUS doesn't rely in a scheduler; we just send messages.
- There is no single point of failure for CHRONUS. Our master node can go down and we don't care. Hadoop's name node is a single point of failure [15].
- DataNodes register with the NameNode, so that the DataNode's identity and responsibilities persist even if the node is restarted under different address or port.

- no need for heartbeats. HFDS has nodes send a heartbeat to the NameNode every 3 seconds [15].

IV. Related Work

THIS SECTION NEEDS OOMPH due to narrowing of focus.

P2P-MapReduce [11], is similar to our work, but looks only at MapReduce; MapReduce is only one of the services CHRONUS provides. It consumed more network resources than the traditional centralized implementation, but was much more tolerant to churn and lost less time when nodes?jobs failed[2]. P2P-MapReduce was not implemented on a large scale; the test results for larger networks were derived from simulations.

Closest to our work is Lee et al.'s work *Parallel Processing Framework on a P2P System Using Map and Reduce Primitives* [9]. Their work, like ours, draws attention to the fact that a P2P network can be much more than a way to distribute files and demonstrates how to accomplish different tasks using map and reduce functions.

4) Underlying protocol: Chord vs Symphony: [9][3] is implemented on top of BruNet[3], which itself is an implantation of the DHT protocol Symphony [10]. Symphony and Chord share a great deal of symetry. Both protocol create an overlay in the shape of a ring, both use a hash to assign files to a node that corresponds to that hash, and both use a finger table[4] to create shortcuts across the ring.

The difference is that Symphony seeks to exploit the small world phenomena [7], where the fingers are chosen at random along a probability distribution function. The further away a node is, the less likely it will be chosen as a finger [5]. Like Chord, messages travel along the paths that bring them closest to their target destination, which is the node responsible the destination's hash value. In a network with N nodes, each with k fingers, a message will take on average $O(\frac{1}{k}\log^2(N))$ hops to reach its destination. In comparison, the average lookup time in Chord is $\frac{1}{2}\log(N)$ [16].

To speed up routing in Symphony, the fingers are bidirectional, rather than unidirectional (does this mean when k is 4 there's effectively 8 fingers? I think it does according to the # tcp connection). Symphony also has nodes maintain a 1-lookahead list for each finger [6].

In the simulation of a 2^15 node network in [10], this allowed the Symphony to perform at nearly identical speeds with Chord while utilizing only 4 bidirectional fingers and a 1-lookahead. With 27 bidriectional finger and 1-lookahead, Symphony's lookups took half the time that Chord did.

However, for large numbers of fingers, keeping a lookahead list becomes expensive. Unless their metrics for Symphony are actually log base 10, then it's unbelievable amazing looking. I have my doubts there.

5) Maintenence: Symphony is extremely effective in a (relatively) smaller network because of the simplicity of the maintenance.

It's okay to have $< k$ fingers or different k for a finger.

6) Fault Tolerence: At a glance Symphony and Chord seem to handle churn the same way, backing up nodes for the data.

Nodes entering the a Symphony ring choose their address uniformly at random, do they hash it? If not, then nodes will basically be unable to resume their place in the network.

7) Implementation of map reduce:

8) Important thing is the paradigm shift:

9) Does Brunet address security?: Doesn't look like it, so we should especially talk about our encrypted chat later and make mention of it heare

10) We're looking at a giant computer, a distributed operating system.: Our work handles issues of fault tolerance and reassigning lost jobs. We also don't explicitly form a hierarchical structure.

V. CHRONUS

CHRONUS bagan its life as a much more generalized response to the question "Is there something I can do with peer-to-peer networks *other* that distrubute files?."

A. Implementation

Paragraph here about the bidding nature of the amazon market. age of the inherent ring structure.

1) Code Details: We implemented the Chord based on the psuedocode in the Stoica's paper [16], using Python instead of C++. We also sent messages instead of performing remote procedure calls.

2) Latency: The latency depends heavily on deployment. Obviously if the network's node are distriuted all over the world, you won't be able to expect the types of gains you'd want in a cluster, but the type of work that you would distribute over the Internet as a whole is different that the work you would want to do in a local cluster. And you'd have different expectations of each

3) Targets for CHRONUS: Who are we marketing to.

B. Distributed MapReduce

1) Operation: In our implementation of a distributed map reduce, each node takes on responsibilities of both a worker and master, much in the same way that a node in a p2p file-sharing service will act as both a client and a server. To start a job, the user contacts a node a specified hash address. The node he contacts to be the stager may be his own computer, and this is preferable when the job involves dividing up a large pile of data.

The job of this stager is to take the work and divide it into *data atoms*, which are the smallest individual units that work can be done on. This might a line of text in a document, the

[2]Fix this, actual quote was "In summary, the experimental results show that even if the P2P-MapReduce system consumes in most cases more network resources than a centralized implementation of MapReduce, it is far more efficient in job management since it minimizes the lost computing time due to jobs failures."

[3]I need to find a better way to reference this; it doesn't have a catchy name.

[4]For simplicity, we use the Chord terminology in discussing Symphony concepts. The *long-distance links* are analogous to Chord's fingers and the *short-distance links* correspond to the predecessor and successors of a node.

[5]Check this with brendan.

[6]these speedups could be implemented on chord.

result of a summation for a particular intermediate value, or a subset of items to be sorted. The specifics of how to divide the work are defined by the user in a *stage* function. These *data atoms* are then given a random hash and sent to that hash address, guaranteeing an even distribution of the data atoms throughout the network. The *data atoms* also contain the map function and reduce function, as defined by the user.

Nodes which receive data atoms apply the map function to the data to create intermediate data atoms, which are sent to other BLANK. DESCRIBE HOW TWO INTERMEDIATE ATOMS GET TO THE SAME PLACE. Nodes that receive at least two intermediate values merge them into one data atom using the reduce function. The atoms are continually merged until only one remains at the hash address of the stager.

Once the reductions are finished, the user gets his results from the node at the stager's address. This may not be be the stager himself, as once the stager has sent all the data atoms, his job is done. The stager does not need to collect the results work, since the work is sent to the stager's hash address, rather than the stager itself. Thus, the stager could quit the network after staging, and both the user and the network would be unaffected by the change.

2) What happens if the stager goes down while staging:

3) Recap the issues of normal MapReduce and why ours is better: We present our implementation of MapReduce not as a direct competitor to MapReduce, but as proof of a more versatile system able to support many complex operations.

The big advantage of our system is the ease of development.

The developer does not need to worry about distributing work evenly, nor does he have to worry about any node in the network going down. The underlying Chord ring handles that automatically. If a node joins the ring as the MapReduce process is running, that node can be assigned work automatically. The stager does not need to keep track of the status of the network. In a node goes down while performing an operation, his successor takes over for him.

All a developer needs to do to write three functions: the staging function, map, and reduce. These define how to split up the work into manageable portions, the work to be performed on each portion to get some results, and how to combine these results into a single result.

4) Calculating Pi:

5) Word Count:

6) My computer can do way more work: In which case you boot up more instances of a node, declaring you can do that much more of the network's work.

C. In which we address possible criticism

1) Mirco versus small: We had huge issues with micro instances because of their sporadic cpu power. Less of an issue with Small instances.

2) If we hash a filename to get the identifier, then you can't have two same file names.:

3) Determining when you're done is expensive: As a whole yes, but merging two sorted lists is $O(n)$ time.

4) Disjoint rings: We can assign every node a *ring id*. When a node creates a ring, he randomly chooses a number for the ring id. Nodes joining that ring use that ring's id. If a node finds another node on differing ring, he compares the their ring id's and leaves if

5) The security we didn't do: We can have multiple non interacting secure file systems.

6) File system via hash prefixing: First k bits of hash used to describe files

VI. Experiments

A set of experiments were run on large groups of Amazon EC2 Micro Instances[1].

We ran word count on James Joyce's *Ulysses* [6].

VII. Results

VIII. Conclusion and Future Work

Our stuff [14] is awesome. CHRONUS can be adapted to handle mutable data, unlike other MapReduce schemes. Tackle security and distrubuted authentication.

References

[1] Amazon.com. Aws documentation: Micro instances. http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/concepts_micro_instances.html

[2] Dhruba Borthakur. The hadoop distributed file system: Architecture and design, 2007.

[3] P. Oscar Boykin, Jesse S. A. Bridgewater, Joseph S. Kong, Kamen M. Lozev, Behnam Attaran Rezaei, and Vwani P. Roychowdhury. A symphony conducted by brunet. *CoRR*, abs/0709.4048, 2007.

[4] Frank Dabek, M Frans Kaashoek, David Karger, Robert Morris, and Ion Stoica. Wide-area cooperative storage with cfs. *ACM SIGOPS Operating Systems Review*, 35(5):202–215, 2001.

[5] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[6] James Joyce. *Ulysses*. Editions Artisan Devereaux, 1924.

[7] Jon M Kleinberg. Navigation in a small world. *Nature*, 406(6798):845–845, 2000.

[8] Kyong-Ha Lee, Yoon-Joon Lee, Hyunsik Choi, Yon Dohn Chung, and Bongki Moon. Parallel data processing with mapreduce: a survey. *ACM SIGMOD Record*, 40(4):11–20, 2012.

[9] Kyungyong Lee, Tae Woong Choi, A. Ganguly, D.I. Wolinsky, P.O. Boykin, and R. Figueiredo. Parallel processing framework on a p2p system using map and reduce primitives. In *Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW), 2011 IEEE International Symposium on*, pages 1602–1609, 2011.

[10] Gurmeet Singh Manku, Mayank Bawa, Prabhakar Raghavan, et al. Symphony: Distributed hashing in a small world. In *USENIX Symposium on Internet Technologies and Systems*, page 10, 2003.

[11] 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, 2012.

[12] Andrew Pavlo, Erik Paulson, Alexander Rasin, Daniel J Abadi, David J DeWitt, Samuel Madden, and Michael Stonebraker. A comparison of approaches to large-scale data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 165–178. ACM, 2009.

[13] Luigi Rizzo. Effective erasure codes for reliable computer communication protocols. *ACM SIGCOMM Computer Communication Review*, 27(2):24–36, 1997.

[14] Andrew Rosen, Brendan Benshoof, and Matt Erwin. Chronus. https://github.com/BrendanBenshoof/CHRONUS.

[15] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. The hadoop distributed file system. In *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on*, pages 1–10. IEEE, 2010.

[16] Ion Stoica, Robert Morris, David Karger, M. Frans Kaashoek, and Hari Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications. *SIGCOMM Comput. Commun. Rev.*, 31:149–160, August 2001.