Chapter 7

Distributed Storage

“Indeed, I believe that virtually every important aspect of programming arises somewhere in the context of [sorting and] searching!”

– Donald E. Knuth, The Art of Computer Programming

How do you store 1M movies, each with a size of about 1GB, on 1M nodes, each equipped with a 1TB disk? Simply store the movies on the nodes, arbitrarily, and memorize (with a global index) which movie is stored on which node. What if the set of movies or nodes changes over time, and you do not want to change your global index too often?

7.1 Consistent Hashing

Several variants of hashing will do the job, e.g. consistent hashing:

Algorithm 7.1 Consistent Hashing

1: Hash the unique file name of each movie $m$ with a known set of hash functions $h_i(m) \rightarrow [0, 1)$, for $i = 1, \ldots, k$
2: Hash the unique name (e.g., IP address and port number) of each node with the same set of hash functions $h_i$, for $i = 1, \ldots, k$
3: Store a copy of a movie $x$ on node $u$ if $h_i(x) \approx h_i(u)$, for any $i$. More formally, store movie $x$ on node $u$ if

$$|h_i(x) - h_i(u)| = \min_m \{|h_i(m) - h_i(u)|\},$$

for any $i$

Theorem 7.2 (Consistent Hashing). In expectation, Algorithm 7.1 stores each movie $kn/m$ times.

Proof. While it is possible that some movie does not hash closest to a node for any of its hash functions, this is highly unlikely: For each node ($n$) and each hash function ($k$), each movie has about the same probability ($1/m$) to be stored. By linearity of expectation, a movie is stored $kn/m$ times, in expectation. □
Remarks:

- Let us do a back-of-the-envelope calculation. We have $m = 1M$ movies, $n = 1M$ nodes, each node has storage for 1TB/1GB = 1K movies, i.e., we use $k = 1K$ hash functions. Theorem 7.2 shows that each movie is stored about 1K times. With a bit more math one can show that it is highly unlikely that a movie is stored much less often than its expected value.

- Instead of storing movies directly on nodes as in Algorithm 7.1, we can also store the movies on any nodes we like. The nodes of Algorithm 7.1 then simply store forward pointers to the actual movie locations.

- In this chapter we want to push unreliability to the extreme. What if the nodes are so unreliable that on average a node is only available for 1 hour? In other words, nodes exhibit a high churn, they constantly join and leave the distributed system.

- With such a high churn, hundreds or thousands of nodes will change every second. No single node can have an accurate picture of what other nodes are currently in the system. This is remarkably different to classic distributed systems, where a single unavailable node may already be a minor disaster: all the other nodes have to get a consistent view (Definition 4.4) of the system again. In high churn systems it is impossible to have a consistent view at any time.

- Instead, each node will just know about a small subset of 100 or less other nodes ("neighbors"). This way, nodes can withstand high churn situations.

- On the downside, nodes will not directly know which node is responsible for what movie. Instead, a node searching for a movie might have to ask a neighbor node, which in turn will recursively ask another neighbor node, until the correct node storing the movie (or a forward pointer to the movie) is found. The nodes of our distributed storage system form a virtual network, also called an overlay network.

7.2 Hypercubic Networks

In this section we present a few overlay topologies of general interest.

Definition 7.3 (Topology Properties). Our virtual network should have the following properties:

- The network should be (somewhat) homogeneous: no node should play a dominant role, no node should be a single point of failure.

- The nodes should have IDs, and the IDs should span the universe $[0, 1)$, such that we can store data with hashing, as in Algorithm 7.1.

- Every node should have a small degree, if possible polylogarithmic in $n$, the number of nodes. This will allow every node to maintain a persistent connection with each neighbor, which will help us to deal with churn.
• The network should have a small **diameter**, and routing should be easy. If a node does not have the information about a data item, then it should know which neighbor to ask. Within a few (polylogarithmic in \( n \)) hops, one should find the node that has the correct information.

![Figure 7.4: The structure of a fat tree.](image)

**Remarks:**

• Some basic network topologies used in practice are trees, rings, grids or tori. Many other suggested networks are simply combinations or derivatives of these.

• The advantage of trees is that the routing is very easy: for every source-destination pair there is only one path. However, since the root of a tree is a bottleneck, trees are not homogeneous. Instead, so-called **fat trees** should be used. Fat trees have the property that every edge connecting a node \( v \) to its parent \( u \) has a capacity that is proportional to the number of leaves of the subtree rooted at \( v \). See Figure 7.4 for a picture.

• Fat trees belong to a family of networks that require edges of non-uniform capacity to be efficient. Networks with edges of uniform capacity are easier to build. This is usually the case for grids and tori. Unless explicitly mentioned, we will treat all edges in the following to be of capacity 1.

**Definition 7.5 (Torus, Mesh).** Let \( m, d \in \mathbb{N} \). The \((m, d)\)-**mesh** \( M(m, d) \) is a graph with node set \( V = [m]^d \) and edge set

\[
E = \left\{ \{(a_1, \ldots, a_d), (b_1, \ldots, b_d)\} \mid a_i, b_i \in [m], \sum_{i=1}^{d} |a_i - b_i| = 1 \right\},
\]

where \([m]\) means the set \(\{0, \ldots, m - 1\}\). The \((m, d)\)-**torus** \( T(m, d) \) is a graph that consists of an \((m, d)\)-mesh and additionally wrap-around edges from nodes \((a_1, \ldots, a_{i-1}, m - 1, a_{i+1}, \ldots, a_d)\) to nodes \((a_1, \ldots, a_{i-1}, 0, a_{i+1}, \ldots, a_d)\) for all
\( i \in \{1, \ldots, d\} \) and all \( a_j \in [m] \) with \( j \neq i \). In other words, we take the expression \( a_i - b_i \) in the sum modulo \( m \) prior to computing the absolute value. \( M(m, 1) \) is also called a path, \( T(m, 1) \) a cycle, and \( M(2, d) = T(2, d) \) a \( d \)-dimensional hypercube. Figure 7.6 presents a linear array, a torus, and a hypercube.

The structure of \( M(m, 1), T(4, 2), \) and \( M(2, 3) \).

Remarks:

- Routing on a mesh, torus, or hypercube is trivial. On a \( d \)-dimensional hypercube, to get from a source bitstring \( s \) to a target bitstring \( t \) one only needs to fix each “wrong” bit, one at a time; in other words, if the source and the target differ by \( k \) bits, there are \( k! \) routes with \( k \) hops.

- If you put a dot in front of the \( d \)-bit ID of each node, the nodes exactly span the \( d \)-bit IDs \([0, 1)\).

- The Chord architecture is a close relative of the hypercube, basically a less rigid hypercube. The hypercube connects every node with an ID in \([0, 1)\) with every node in exactly distance \( 2^{-i}, i = 1, 2, \ldots, d \) in \([0, 1)\). Chord instead connect nodes with approximately distance \( 2^{-i} \).

- The hypercube has many derivatives, the so-called hypercubic networks. Among these are the butterfly, cube-connected-cycles, shuffle-exchange, and de Bruijn graph. We start with the butterfly, which is basically a “rolled out” hypercube.

**Definition 7.7 (Butterfly).** Let \( d \in \mathbb{N} \). The \( d \)-dimensional butterfly \( BF(d) \) is a graph with node set \( V = [d + 1] \times [2]^d \) and an edge set \( E = E_1 \cup E_2 \) with

\[
E_1 = \{ (i, \alpha), (i + 1, \alpha) \mid i \in [d], \alpha \in [2^d] \}
\]

and

\[
E_2 = \{ (i, \alpha), (i + 1, \beta) \mid i \in [d], \alpha, \beta \in [2^d], |\alpha - \beta| = 2^i \}.
\]

A node set \( \{(i, \alpha) \mid \alpha \in [2^d]\} \) is said to form level \( i \) of the butterfly. The \( d \)-dimensional wrap-around butterfly \( W-BF(d) \) is defined by taking the \( BF(d) \) and having \((d, \alpha) = (0, \alpha) \) for all \( \alpha \in [2^d] \).
Remarks:

- Figure 7.8 shows the 3-dimensional butterfly $BF(3)$. The $BF(d)$ has $(d+1)2^d$ nodes, $2d \cdot 2^d$ edges and degree 4. It is not difficult to check that combining the node sets $\{(i, \alpha) \mid i \in [d]\}$ for all $\alpha \in [2]^d$ into a single node results in the hypercube.

- Butterflies have the advantage of a constant node degree over hypercubes, whereas hypercubes feature more fault-tolerant routing.

- You may have seen butterfly-like structures before, e.g. sorting networks, communication switches, data center networks, fast fourier transform (FFT). The Benes network (telecommunication) is nothing but two back-to-back butterflies. The Clos network (data centers) is a close relative to Butterflies too. Actually, merging the $2^i$ nodes on level $i$ that share the first $d-i$ bits into a single node, the Butterfly becomes a fat tree. Every year there are new applications for which hypercubic networks are the perfect solution!

- Next we define the cube-connected-cycles network. It only has a degree of 3 and it results from the hypercube by replacing the corners by cycles.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{bf3.png}
\caption{The structure of BF(3).}
\end{figure}

**Definition 7.9** (Cube-Connected-Cycles). Let $d \in \mathbb{N}$. The cube-connected-cycles network $CCC(d)$ is a graph with node set $V = \{(a, p) \mid a \in [2]^d, p \in [d]\}$ and edge set

\[ E = \{ \{(a, p), (a, (p + 1) \mod d)\} \mid a \in [2]^d, p \in [d]\} \]

\[ \cup \{\{(a, p), (b, p)\} \mid a, b \in [2]^d, p \in [d], a = b \text{ except for } a_p \} . \]
Remarks:

- Two possible representations of a CCC can be found in Figure 7.10.
- The shuffle-exchange is yet another way of transforming the hypercubic interconnection structure into a constant degree network.

**Definition 7.11 (Shuffle-Exchange).** Let $d \in \mathbb{N}$. The $d$-dimensional shuffle-exchange $SE(d)$ is defined as an undirected graph with node set $V = [2]^d$ and an edge set $E = E_1 \cup E_2$ with

$$E_1 = \{ \{(a_1, \ldots, a_d), (\bar{a}_1, \ldots, \bar{a}_d)\} \mid (a_1, \ldots, a_d) \in [2]^d, \ \bar{a}_d = 1 - a_d \}$$

and

$$E_2 = \{ \{(a_1, \ldots, a_d), (a_d, a_1, \ldots, a_{d-1})\} \mid (a_1, \ldots, a_d) \in [2]^d \}.$$

Figure 7.12 shows the 3- and 4-dimensional shuffle-exchange graph.

**Definition 7.13 (DeBruijn).** The $b$-ary DeBruijn graph of dimension $d$ $DB(b,d)$ is an undirected graph $G = (V,E)$ with node set $V = \{v \in [b]^d \}$ and edge set $E$ that contains all edges $\{v,w\}$ with the property that $w \in \{(x,v_1,\ldots,v_{d-1}) : x \in [b]\}$, where $v = (v_1,\ldots,v_d)$. 

Figure 7.12: The structure of SE(3) and SE(4).
Remarks:

- Two examples of a DeBruijn graph can be found in Figure 7.14.
- There are some data structures which also qualify as hypercubic networks. An example of a hypercubic network is the skip list, the balanced binary search tree for the lazy programmer:

Definition 7.15 (Skip List). The skip list is an ordinary ordered linked list of objects, augmented with additional forward links. The ordinary linked list is the level 0 of the skip list. In addition, every object is promoted to level 1 with probability 1/2. As for level 0, all level 1 objects are connected by a linked list. In general, every object on level \( i \) is promoted to the next level with probability 1/2. A special start-object points to the smallest/first object on each level.

Remarks:

- Search, insert, and delete can be implemented in \( O(\log n) \) expected time in a skip list, simply by jumping from higher levels to lower ones when overshooting the searched position. Also, the amortized memory cost of each object is constant, as on average an object only has two forward links.
- The randomization can easily be discarded, by deterministically promoting a constant fraction of objects of level \( i \) to level \( i + 1 \), for all \( i \). When inserting or deleting, object \( o \) simply checks whether its left and right level \( i \) neighbors are being promoted to level \( i + 1 \). If none of them is, promote object \( o \) itself. Essentially we establish a maximal independent set (MIS) on each level, hence at least every third and at most every second object is promoted.
- There are obvious variants of the skip list, e.g., the skip graph. Instead of promoting only half of the nodes to the next level, we always promote all the nodes, similarly to a balanced binary tree: All nodes are part of the root level of the binary tree. Half the nodes are promoted left, and half the nodes are promoted right, on each level. Hence on level \( i \) we have have \( 2^i \) lists (or, if we connect the last element again with the first: rings) of about \( n/2^i \) objects. The skip graph features all the properties of Definition 7.3.
- More generally, how are degree and diameter of Definition 7.3 related? The following theorem gives a general lower bound.
Theorem 7.16. Every graph of maximum degree \( d > 2 \) and size \( n \) must have a diameter of at least \( \lceil (\log n)/(\log(d-1)) \rceil - 2 \).

Proof. Suppose we have a graph \( G = (V,E) \) of maximum degree \( d \) and size \( n \). Start from any node \( v \in V \). In a first step at most \( d \) other nodes can be reached. In two steps at most \( d \cdot (d-1) \) additional nodes can be reached. Thus, in general, in at most \( k \) steps at most
\[
1 + \sum_{i=0}^{k-1} d \cdot (d-1)^i = 1 + d \cdot \frac{(d-1)^k - 1}{(d-1)-1} \leq \frac{d \cdot (d-1)^k}{d-2}
\]

nodes (including \( v \)) can be reached. This has to be at least \( n \) to ensure that \( v \) can reach all other nodes in \( V \) within \( k \) steps. Hence,
\[
(d-1)^k \geq \frac{(d-2) \cdot n}{d} \quad \Leftrightarrow \quad k \geq \log_{d-1}((d-2) \cdot n/d).
\]
Since \( \log_{d-1}((d-2)/d) > -2 \) for all \( d > 2 \), this is true only if \( k \geq \lceil (\log n)/(\log(d-1)) \rceil - 2 \).

Remarks:
- In other words, constant-degree hypercubic networks feature an asymptotically optimal diameter.
- Other hypercubic graphs manage to have a different tradeoff between node degree and diameter. The pancake graph, for instance, minimizes the maximum of these with \( d = k = \Theta(\log n/\log \log n) \). The ID of a node \( u \) in the pancake graph of dimension \( d \) is an arbitrary permutation of the numbers \( 1, 2, \ldots, d \). Two nodes \( u, v \) are connected by an edge if one can get the ID of node \( v \) by taking the ID of node \( u \), and reversing (flipping) the first \( i \) numbers of \( u \)'s ID. For example, in dimension \( d = 4 \), nodes \( u = 2314 \) and \( v = 1324 \) are neighbors.
- There are a few other interesting graph classes which are not hypercubic networks, but nevertheless seem to relate to the properties of Definition 7.3. Small-world graphs (a popular representations for social networks) also have small diameter, however, in contrast to hypercubic networks, they are not homogeneous and feature nodes with large degrees.
- Expander graphs (an expander graph is a sparse graph which has good connectivity properties, that is, from every not too large subset of nodes you are connected to an even larger set of nodes) are homogeneous, have a low degree and small diameter. However, expanders are often not routable.

7.3 DHT & Churn

Definition 7.17 (Distributed Hash Table (DHT)). A distributed hash table (DHT) is a distributed data structure that implements a distributed storage. A DHT should support at least (i) a search (for a key) and (ii) an insert (key, object) operation, possibly also (iii) a delete (key) operation.
7.3. DHT & CHURN

Remarks:

- A DHT has many applications beyond storing movies, e.g., the Internet domain name system (DNS) is essentially a DHT.
- A DHT can be implemented as a hypercubic overlay network with nodes having identifiers such that they span the ID space $[0, 1)$.
- A hypercube can directly be used for a DHT. Just use a globally known set of hash functions $h_i$, mapping movies to bit strings with $d$ bits.
- Other hypercubic structures may be a bit more intricate when using it as a DHT: The butterfly network, for instance, may directly use the $d+1$ layers for replication, i.e., all the $d+1$ nodes are responsible for the same ID.
- Other hypercubic networks, e.g. the pancake graph, might need a bit of twisting to find appropriate IDs.
- We assume that a joining node knows a node which already belongs to the system. This is known as the bootstrap problem. Typical solutions are: If a node has been connected with the DHT previously, just try some of these previous nodes. Or the node may ask some authority for a list of IP addresses (and ports) of nodes that are regularly part of the DHT.
- Many DHTs in the literature are analyzed against an adversary that can crash a fraction of random nodes. After crashing a few nodes the system is given sufficient time to recover again. However, this seems unrealistic. The scheme sketched in this section significantly differs from this in two major aspects.
- First, we assume that joins and leaves occur in a worst-case manner. We think of an adversary that can remove and add a bounded number of nodes; the adversary can choose which nodes to crash and how nodes join.
- Second, the adversary does not have to wait until the system is recovered before it crashes the next batch of nodes. Instead, the adversary can constantly crash nodes, while the system is trying to stay alive. Indeed, the system is never fully repaired but always fully functional. In particular, the system is resilient against an adversary that continuously attacks the “weakest part” of the system. The adversary could for example insert a crawler into the DHT, learn the topology of the system, and then repeatedly crash selected nodes, in an attempt to partition the DHT. The system counters such an adversary by continuously moving the remaining or newly joining nodes towards the areas under attack.
- Clearly, we cannot allow the adversary to have unbounded capabilities. In particular, in any constant time interval, the adversary can at most add and/or remove $O(\log n)$ nodes, $n$ being the total number of nodes currently in the system. This model covers an adversary
which repeatedly takes down nodes by a distributed denial of service attack, however only a logarithmic number of nodes at each point in time. The algorithm relies on messages being delivered timely, in at most constant time between any pair of operational nodes, i.e., the synchronous model. Using the trivial synchronizer this is not a problem. We only need bounded message delays in order to have a notion of time which is needed for the adversarial model. The duration of a round is then proportional to the propagation delay of the slowest message.

**Algorithm 7.18 DHT**

1: Given: a globally known set of hash functions $h_i$, and a hypercube (or any other hypercubic network)
2: Each hypercube virtual node (“hypernode”) consists of $\Theta(\log n)$ nodes.
3: Nodes have connections to all other nodes of their hypernode and to nodes of their neighboring hypernodes.
4: Because of churn, some of the nodes have to change to another hypernode such that up to constant factors, all hypernodes own the same number of nodes at all times.
5: If the total number of nodes $n$ grows or shrinks above or below a certain threshold, the dimension of the hypercube is increased or decreased by one, respectively.

**Remarks:**

- Having a logarithmic number of hypercube neighbors, each with a logarithmic number of nodes, means that each node has $\Theta(\log^n n)$ neighbors. However, with some additional bells and whistles one can achieve $\Theta(\log n)$ neighbor nodes.

- The balancing of nodes among the hypernodes can be seen as a dynamic token distribution problem on the hypercube. Each hypernode has a certain number of tokens, the goal is to distribute the tokens along the edges of the graph such that all hypernodes end up with the same or almost the same number of tokens. While tokens are moved around, an adversary constantly inserts and deletes tokens. See also Figure 7.19.

- In summary, the storage system builds on two basic components: (i) an algorithm which performs the described dynamic token distribution and (ii) an information aggregation algorithm which is used to estimate the number of nodes in the system and to adapt the dimension of the hypercube accordingly:

**Theorem 7.20 (DHT with Churn).** We have a fully scalable, efficient distributed storage system which tolerates $O(\log n)$ worst-case joins and/or crashes per constant time interval. As in other storage systems, nodes have $O(\log n)$ overlay neighbors, and the usual operations (e.g., search, insert) take time $O(\log n)$. 
7.3. DHT & CHURN

Figure 7.19: A simulated 2-dimensional hypercube with four hypernodes, each consisting of several nodes. Also, all the nodes are either in the core or in the periphery of a node. All nodes within the same hypernode are completely connected to each other, and additionally, all nodes of a hypernode are connected to the core nodes of the neighboring nodes. Only the core nodes store data items, while the peripheral nodes move between the nodes to balance biased adversarial churn.

Remarks:

- Indeed, handling churn is only a minimal requirement to make a distributed storage system work. Advanced studies proposed more elaborate architectures which can also handle other security issues, e.g., privacy or Byzantine attacks.

Chapter Notes

The ideas behind distributed storage were laid during the peer-to-peer (P2P) file sharing hype around the year 2000, so a lot of the seminal research in this area is labeled P2P. The paper of Plaxton, Rajaraman, and Richa [PRR97] laid out a blueprint for many so-called structured P2P architecture proposals, such as Chord [SMK+01], CAN [RFH+01], Pastry [RD01], Viceroy [MNR02], Kademlia [MM02], Koorde [KK03], SkipGraph [AS03], SkipNet [HJS+03], or Tapestry [ZHS+04]. Also the paper of Plaxton et. al. was standing on the shoulders of giants. Some of its eminent precursors are: linear and consistent hashing [KLL+97], locating shared objects [AP90, AP91], compact routing [SK85, PU88], and even earlier: hypercubic networks, e.g. [AJ75, WIT81, GS81, BA84].

Furthermore, the techniques in use for prefix-based overlay structures are related to a proposal called LAND, a locality-aware distributed hash table proposed by Abraham et al. [AMD04].

More recently, a lot of P2P research focussed on security aspects, describing for instance attacks [LMSW06, SENB07, Lar07], and provable countermeasures [KSW05, AS09, BSS09]. Another topic currently garnering interest is using P2P to help distribute live streams of video content on a large scale [LMSW07]. There are several recommendable introductory books on P2P computing, e.g.
CHAPTER 7. DISTRIBUTED STORAGE

[SW05, SG05, MS07, KW08, BYL08].

Some of the figures in this chapter have been provided by Christian Scheideler.

Bibliography


[LMSW06] Thomas Locher, Patrick Moor, Stefan Schmid, and Roger Wattenhofer. Free Riding in BitTorrent is Cheap. In 5th Workshop on Hot Topics in Networks (HotNets), Irvine, California, USA, November 2006.


