

## Lecture 25: 4.21.05

Lecturer: Christos

Scribe: Amir Kamil

**Disclaimer:** *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

## 25.1 Load Balancing/Balls and Bins

Consider the problem of balancing a load of  $m$  requests across  $n$  servers. We would like to do this in a decentralized way to minimize the cost of the balancing. The simplest load balancing algorithm is to randomly choose one of the servers for each request, which is equivalent to randomly throwing  $m$  balls into  $n$  bins. Using this algorithm, what is the maximum load on any server?

### 25.1.1 The Birthday Paradox

The *birthday paradox* states that in a set of 20 people, it is very likely that two of them share a birthday. More generally, when throwing  $m$  balls into  $n$  bins, if  $m < \sqrt{n}$ , the maximum number of balls in any bin is likely to be one, but if  $m > \sqrt{n}$ , it is likely to be more than one.

There is at most one ball per bin when the second ball doesn't collide with the first, the third collides with neither the first nor the second, and so on. Thus we have

$$\begin{aligned} \Pr[\leq 1 \text{ ball/bin}] &= 1 \cdot \left(1 - \frac{1}{n}\right) \cdot \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{m-1}{n}\right) \\ &\leq e^{-\frac{1}{n}} \cdot e^{-\frac{2}{n}} \cdots e^{-\frac{m-1}{n}} \\ &= e^{-\frac{m(m-1)}{2n}}. \end{aligned}$$

This expression is very small if  $m > \sqrt{n}$ , but close to 1 if  $m < \sqrt{n}$ . Thus a collision is likely in the former case but unlikely in the latter.

### 25.1.2 Hashing

In hash tables, the number of buckets is generally about the same as the number of elements, so  $m \approx n$ . What is the maximum number of elements in any bucket in this case? It turns out that the exact bound,  $\frac{\log n}{\log \log n}$  is difficult to prove, so we only prove that the maximum is likely to be bounded by  $O(\ln n)$ .

First, we compute the probability that bin 1 has at least  $k$  balls. The probability that bin 1 has exactly  $k$  balls is

$$\Pr[k \text{ balls in bin 1}] = \binom{m}{k} \cdot \left(\frac{1}{n}\right)^k \cdot \left(1 - \frac{1}{n}\right)^{m-k}.$$

The first term above corresponds to choosing  $k$  out of the  $m$  balls to put in bin 1, the second to the probability of them all going to bin 1, and the third the probability that none of the rest of the balls go in bin 1. The probability that bin 1 has at least  $k$  balls is then

$$\begin{aligned} \Pr[\geq k \text{ balls in bin 1}] &= \sum_{i=k}^m \binom{m}{i} \cdot \left(\frac{1}{n}\right)^i \cdot \left(1 - \frac{1}{n}\right)^{m-i} \\ &\approx \sum_{i=k}^m \binom{m}{i} \cdot \left(\frac{1}{n}\right)^i && \text{(disregarding the last term)} \\ &\leq \sum_{i=k}^m \left(\frac{me}{in}\right)^i && \text{(Stirling's approximation)} \\ &\leq \sum_{i=k}^m \left(\frac{me}{kn}\right)^i. \end{aligned}$$

Let  $k = 2e \frac{m}{n} + 2 \ln n$ . Then

$$\begin{aligned} \Pr[\geq k \text{ balls in bin 1}] &\leq \sum_{i=k}^m \left(\frac{me}{kn}\right)^i \\ &\leq \sum_{i=k}^m \left(\frac{1}{2}\right)^i \\ &\leq 2 \left(\frac{1}{2}\right)^k \\ &\leq \frac{1}{n^2}. \end{aligned}$$

Now in order to compute the probability that any bin has at least  $k$  balls, we use the *union bound*. The union bound relies on the *inclusion/exclusion principle*:

**Theorem 25.1 [Inclusion/Exclusion Principle]** *Given  $n$  events  $E_1, \dots, E_n$ , the probability that at least one occurs is*

$$\Pr[\bigcup_{i=1}^n E_i] = \sum_{i=1}^n \Pr[E_i] - \sum_{\{i,j\} | i \neq j} \Pr[E_i \cap E_j] + \sum_{\{i,j,k\} | i \neq j, i \neq k, j \neq k} \Pr[E_i \cap E_j \cap E_k] - \dots \pm \Pr[\bigcap_{i=1}^n E_i].$$

In the above expression, each term is necessarily smaller than (or equal to) the preceding one, so ignoring all but the first term, we obtain the union bound:

**Theorem 25.2 [Union bound]** *Given  $n$  events  $E_1, \dots, E_n$ , the probability that at least one occurs is*

$$\Pr[\bigcup_{i=1}^n E_i] \leq \sum_{i=1}^n \Pr[E_i].$$

Using the union bound, the probability that any bin has at least  $k$  balls is

$$\begin{aligned} \Pr[\geq k \text{ balls in any bin}] &\leq \sum_{i=1}^n \Pr[\geq k \text{ balls in bin } i] \\ &\leq \frac{1}{n}. \end{aligned}$$

Thus, the probability that any bin has more than  $k = 2e\frac{m}{n} + 2\ln n$  balls is very low, so the maximum in any bin is  $k$ , which is in  $O(\ln n)$  when  $m \approx n$ .

### 25.1.3 Higher Loads

Suppose that we now have  $m > n \ln n$  balls. We show that with high probability, the maximum number of balls in any bin is  $\frac{m}{n} + \sqrt{8\frac{m}{n} \ln n}$ .

In order to prove this, we require the *Chernoff bound*:

**Theorem 25.3 [Chernoff bound]** *Given  $n$  independent random variables  $x_1, \dots, x_n \in \{0, 1\}$ , let  $\mu = \sum_{i=1}^n \Pr[x_i = 1]$ . Then for any  $\delta \leq 2e - 1$ ,*

$$\Pr \left[ \sum_{i=1}^n x_i \geq (1 + \delta)\mu \right] \leq e^{-\frac{1}{4}\delta^2\mu}$$

Let  $x_i$  denote whether or not ball  $i$  falls in bin 1. Then  $\Pr[x_i = 1] = \frac{1}{n}$ . Let  $\mu = \frac{m}{n}$ ,  $\delta = \sqrt{8\frac{m}{n} \ln n}$ , so that  $k = (1 + \delta)\mu$ . Then

$$\begin{aligned} \Pr[\geq k \text{ balls in bin 1}] &= \Pr \left[ \sum_{i=1}^n x_i \geq k \right] \\ &= \Pr \left[ \sum_{i=1}^n x_i \geq (1 + \delta)\mu \right] \\ &\leq e^{-\frac{1}{4}(8\frac{m}{n} \ln n)\frac{m}{n}} && \text{(Chernoff bound)} \\ &= e^{-2 \ln n} \\ &= \frac{1}{n^2}. \end{aligned}$$

Applying the union bound as before, we get

$$\Pr[\geq k \text{ balls in any bin}] \leq \frac{1}{n},$$

so the maximum number of balls in any bin is likely to be  $k = \frac{m}{n} + \sqrt{8\frac{m}{n} \ln n}$ .

### 25.1.4 The Power of Two Choices

Suppose that instead of choosing a single bin for each ball, we randomly choose two bins and place the ball in the bin with lower load. When  $m \approx n$ , the maximum number of balls in any bin becomes  $O(\log \log n)$ , exponentially better than with only a single bin. We prove that this is the case for  $m \approx \frac{n}{8}$ .

Picture the balls and bins as a graph, where each node corresponds to a bin and each edge corresponds to a ball, where the endpoints of the edge are the ball's two bin choices. Run the following algorithm on the graph:

```
for (int i = 1; ; i++) {
    find all nodes of degree < 13;
    delete them;
}
```

We show the following facts:

- (1) Suppose bin  $b$  is deleted at stage  $i$ . Then its load cannot exceed  $13i$ .
- (2) With high probability, all bins will be deleted within  $\log \log n$  stages.

We prove the first fact by induction. The base case trivially holds, since  $b$  has at most 12 edges and therefore load at most 12. For the inductive step, note that  $b$  has at most 12 nodes connected to it at stage  $i$ .  $b$  could also have many edges to deleted nodes, which by the inductive hypothesis, all must have load at most  $13(i-1)$ . Now consider what happens when the balls are added to bins. If at some point  $b$ 's load exceeds  $13(i-1)$ , the remaining balls for which  $b$  is a choice all choose the other bin, since its load is less than  $13(i-1)$ . The exceptions to this are the  $\leq 12$  balls whose other choice is not yet deleted. But even if they all go to  $b$ ,  $b$ 's load is still at most  $(13(i-1) + 1) + 12 \leq 13i$ , so the first fact holds.

In order to prove the second fact, we require the following lemma:

**Lemma 25.4** *With high probability,  $\forall k \geq 10 \log n$ , no subset of  $k$  nodes will have  $\geq k-1$  edges between them. Also with high probability,  $\forall l \leq 10 \log n$ , no subset of  $l$  nodes will have  $\geq 3l$  edges between them.*

**Proof:** Recall that there are a total of  $\frac{n}{8}$  edges. Consider an arbitrary subset  $S$  of  $j$  nodes. The probability that an arbitrary edge  $e$  is within this subset is

$$\Pr[e \in S] = \left(\frac{j}{n}\right)^2,$$

since the probability that each endpoint is in the subset is  $\frac{j}{n}$ . The probability that a set of edges  $e_1, \dots, e_i$  are all in the subset is then

$$\Pr[e_1, \dots, e_i \in S] = \left(\frac{j}{n}\right)^{2i},$$

and the probability that the subset has any  $i$  edges is

$$\Pr[S \text{ has } i \text{ edges}] = \binom{\frac{n}{8}}{i} \left(\frac{j}{n}\right)^{2i}.$$

The probability that any subset of  $j$  nodes has  $i$  edges is then

$$\Pr[\exists S . |S| = j \wedge S \text{ has } i \text{ edges}] \leq \binom{n}{j} \binom{\frac{n}{8}}{i} \left(\frac{j}{n}\right)^{2i}.$$

Setting  $j = k$  and  $i = k-1$ , we see that this probability is very small. Similarly for  $j = l$  and  $i = 3l$ . ■

Since a connected component of  $k$  nodes requires at least  $k-1$  edges, the first part of lemma 25.4 implies that the connected components are of size at most  $\log n$ . The second part implies that the average degree of a node is at most 6 in each connected component, so less than half the nodes in a connected component have degree more than 12. This means that at least half are deleted in the first stage of the algorithm above. The remaining nodes in each connected component again have average degree at most 6, so at least half the remaining nodes are deleted in the second stage. Continuing this process, we see that each connected component  $C$  is deleted in  $\log |C|$  stages. Since the connected components are of size at most  $\log n$ , at most  $\log \log n$  stages are needed to delete all connected components.

Since at most  $\log \log n$  stages are required to delete all bins, fact (1) implies that each bin has load at most  $13 \log \log n$ .