SRC Technical Note
1997 - 024
October 1, 1997

Studying Balanced Allocations with Differential Equations

Michael Mitzenmacher

digital
Systems Research Center
130 Lytton Avenue
Palo Alto, California 94301
http://www.research.digital.com/SRC/

Studying Balanced Allocations with Differential Equations *

Michael Mitzenmacher †

Abstract

Using differential equations, we examine the GREEDY algorithm studied by Azar, Broder, Karlin, and Upfal for distributed load balancing [1]. This approach yields accurate estimates of the actual load distribution, provides insight into the exponential improvement GREEDY offers over simple random selection, and allows one to prove tight concentration theorems about the loads in a straightforward manner.

1 Introduction

Suppose that n balls are placed into n bins, with each ball being placed into a bin chosen independently and uniformly at random. Let the *load* of a bin be the number of balls in that bin after all balls have been thrown. It is well known that with high probability, the maximum load upon completion will be approximately $\frac{\log n}{\log \log n}$ [8]. (We use *with high probability* to mean with probability at least $1 - O(1/n)$, where n is the number of balls. Also, log will always mean the natural logarithm, unless otherwise noted.)

Azar, Broder, Karlin, and Upfal considered how much more evenly distributed the load would be if each ball had two (or more) choices [1]. Suppose that the balls are placed sequentially, and each ball is placed into the less full of *two* bins chosen independently and uniformly at random with replacement (breaking ties arbitrarily). In this case, they showed that the maximum load drops to $\frac{\log \log n}{\log 2} + O(1)$ with high probability. If each ball instead has d choices, then the maximum load will be $\frac{\log \log n}{\log d} + O(1)$ with high probability. Having two choices hence yields a qualitatively different type of behavior from the single choice case, leading to an exponential improvement in the maximum load; having more than two choices further improves the maximum load by only a constant factor. This result has important implications for distributed load balancing, hashing, and PRAM simulation [1].

Following Azar *et al.*, we refer to the algorithm in which each ball has d random choices as GREEDY(d). In this paper, we develop an alternative method of studying the performance of GREEDY(d) using differential equations. The differential equations describe the limiting performance of of GREEDY(d) as the number of balls and bins grows to infinity. As we will demonstrate, the description of the limiting performance proves highly

*A preliminary version of this work appeared in the *Proceedings of the 37th Annual Symposium on the Foundations of Computer Science*, October 1996.

†Much of this work was done at U.C. Berkeley, supported by a fellowship from the Office of Naval Research and by NSF grant CCR-9505448.

accurate, even when n is relatively small. Besides giving perhaps more insight into the actual performance of GREEDY, our methods provides a great deal of intuition behind the proof of the behavioral difference between one and two choices.

Our motivation in studying this problem is twofold. First, we wish to demonstrate and highlight this methodology, and encourage its use for studying random processes. While this methodology is by no means new, its uses have been surprisingly limited. The technical results justifying the relationship between families of Markov processes and differential equations date back at least to Kurtz [13, 14]. Karp and Sipser provided and early use of this technology to study an algorithm for finding maximum matchings in sparse random graphs [11]. Other past applications in the analysis of algorithms include [9, 12], and more recently many more have been found (see, for example, [2, 3, 15, 17, 21] to name a few).

Our second motivation is to demonstrate the power of using two choices. This idea dates back at least as far as the work of Eager, Lazowska, and Zahorjan [7], who examined the problem in a dynamic load balancing setting based on viewing processors as single server queues. In the static setting, this idea was also studied by Hajek [9], who used the same approach we undertake to determine the fraction of empty bins. The aforementioned exponential improvement in behavior was noted and proven first in a paper by Karp, Luby, and Meyer auf der Heide [10]. The work by Azar *et. al.* examined a simpler model that clarified the argument as well as provided many new results. Related work by the author [16, 17], as well as by others [20], examines the power of two choices in dynamic settings. Continued work in the area includes recent work by Stemann [19] and Czumaj and Stemann [6].

In the rest of the paper, we explain the derivation of the differential equations that describe the GREEDY strategy of [1] and compare the results from the differential equations with simulations. We also demonstrate how the equations give more insight into the behavior of GREEDY and how the equations relate to the work in [1].

2 The Differential Equations

In this section, we demonstrate how to establish a family of differential equations that can be used to model the behavior of the GREEDY strategy of [1]. Recall that we begin with m balls and n bins. Balls arrive sequentially, and upon arrival, each ball chooses d bins independently and uniformly at random (with replacement); the ball is then placed in the least loaded of these bins (with ties broken arbitrarily).

We first ask how many bins remain empty after the protocol GREEDY(d) terminates. This question has a natural interpretation in the task-processor model: how many of our processors are not utilized? The question can also be seen as a matching problem on random bipartite graphs: given a bipartite graph with n vertices on each side such that each vertex on the left has d edges to vertices chosen independently and uniformly at random on the right, what is the expected size of the greedy matching obtained by sequentially matching vertices on the left to a random unmatched neighbor? Our attack, again, is to consider this system as $n \to \infty$. This question has been previously solved by Hajek using entirely similar techniques [9]. We shall begin by briefly repeating his argument with some additional insights. Once we show how to answer the question of the number of empty bins, we shall extend it to the more general load balancing problem.

2.1 The empty bins problem

We set up the problem of the number of empty bins by developing a Markov chain with a simple state that describes the balls ans bins process. We first establish a concept of time. Let $Y(T)$ be the number of non-empty bins after T balls have been thrown. Then $\{Y(i)\}$, $i = 0 \ldots m$, is clearly a Markov chain. Moreover

$$\mathsf{E}[Y(T+1) - Y(T)] = 1 - \left(\frac{Y(T)}{n}\right)^d,$$

since the probability that a ball finds all non-empty bins among its d choices is $(Y(T)/n)^d$.

The notation becomes somewhat more convenient if we scale by a factor of n. We let t be the time at which exactly nt balls have been thrown, and we let $y(t)$ be the fraction of non-empty bins. The the claim is the random process described above is well approximated by the the trajectory of the differential equation

$$\frac{dy}{dt} = 1 - y^d, \tag{1}$$

where this equation has been obtained by using for the limiting value dy the expected change in y. That the random process given by the Markov chain closely follows the trajectory given by the differential equation follows easily from known techniques, such as for example Kurtz's Theorem, or the similar work on random graphs by Wormald. (As previously mentioned, the balls and bins process has a natural interpretation in terms of random bipartite graphs.) This connection with the differential equation yields the following theorem:

Theorem 1 *Suppose cn balls are thrown into n bins according to the* GREEDY(d) *protocol for some constant c. Let Y_{cn} be the number of non-empty bins when the process terminates. Then* $\lim_{n\to\infty} \mathsf{E}[\frac{Y_{cn}}{n}] = y_c$*, where* $y_c < 1$ *satisfies*

$$c = \sum_{i=0}^{\infty} \frac{y_c^{id+1}}{(id+1)}.$$

Proof: The preconditions for Kurtz's theorem (see [18, Theorem 5.3] or [14, Chapter 8]) are easily checked for the one-dimensional system described by (1), so by Kurtz's theorem we have that this differential equation is the correct limiting process.[1] Instead of solving (1) for y in terms of t, we solve for t in terms of y: $\frac{dt}{dy} = \frac{1}{1-y^d} = \sum_{i=0}^{\infty} y^{id}$. We integrate up to some time t, yielding

$$t = \sum_{i=0}^{\infty} \frac{y(t)^{id+1}}{(id+1)}. \tag{2}$$

From equation (2), given d we can solve for $y(t)$ for any value of t using for example binary search. One can also attempt to find an equation for y in terms of d and t; standard integral tables give such equations when $d = 2, 3$ and 4, for example. When $t = c$, all of the balls have been thrown, and the process terminates.

[1] It appears that there might be a problem here since we consider events occurring at discrete time steps, instead of according to random times from a Poisson process. One can always adopt the convention that each discrete time step corresponds to an amount of time given by an exponentially distributed random variable. In the limiting case, this distinction disappears.

Plugging $t = c$ into equation (2) yields the theorem, with $y_c = y(c)$. ■

We may actually use Kurtz's Theorem to obtain a concentration result (again, see [18, Theorem 5.3] or [14, Chapter 8]).

Theorem 2 *In the notation of Theorem 1, $|\frac{Y_{cn}}{n} - y_c|$ is $O\left(\sqrt{\frac{\log n}{n}}\right)$ with high probability, where the constant depends on c.*

One can also obtain entirely similar bounds using Y_{cn} more straightforward martingale arguments. In the following, we assume familiarity with basic martingale theory; see, for example, [4, Chapter 7] for more information. We use the following form of the martingale tail inequality due to Azuma [5]:

Lemma 3 **[Azuma]** *Let $X_0, X_1, \ldots X_m$ be a martingale sequence such that for each k,*

$$|X_k - X_{k-1}| \leq 1.$$

Then for any $\alpha > 0$,

$$\mathbf{Pr}(|X_m - X_0| > \alpha\sqrt{m}) < 2\mathrm{e}^{-\alpha^2/2}.$$

Theorem 4 *In the notation of Theorem 1, $\mathsf{Pr}(|Y_{cn} - \mathsf{E}[Y_{cn}]| > \alpha\sqrt{cn}) < 2\mathrm{e}^{-\alpha^2/2}$ for any $\alpha > 0$.*

Proof: For $0 \leq j \leq cn$, let $\mathcal{F}_j$ be the σ-field of events corresponding to the possible states after j balls have been placed, and $Z_j = \mathsf{E}[Y_{cn}|\mathcal{F}_j]$ be the associated conditional expectation of Y_{cn}. Then the random variables $\{Z_j\}_{j=0}^{cn}$ form a Doob martingale, and it is clear that $|Z_j - Z_{j-1}| \leq 1$. The theorem now follows from Lemma 3. ■

Theorem 4 implies that Y_{cn} is within $O(\sqrt{n \log n})$ of its expected value with high probability; however, the martingale approach does not immediately lead us to the value to which Y_{cn}/n converges. This is a standard limitation of the martingale approach: in contrast, the differential equations approach allows us to find the mean as well as prove concentration around the mean.

2.2 Bins with fixed load

We can extend the previous analysis to find the fraction of bins with load k for any constant k as $n \to \infty$. We first establish the appropriate Markov chain. Let $s_i(t)$ be the fraction of bins with load at least i at time t, where again at time t exactly nt balls have been thrown. Then the corresponding differential equations regarding the growth of the s_i (for $i \geq 1$) are easily determined:

$$\begin{cases} \dfrac{ds_i}{dt} & = (s_{i-1}^d - s_i^d) \text{ for } i \geq 1\,; \\ s_0 & = 1. \end{cases} \tag{3}$$

	$d = 2$ Prediction	1 million Simulation	$d = 3$ Prediction	1 million Simulation
s_1	0.7616	0.7616	0.8231	0.8230
s_2	0.2295	0.2295	0.1765	0.1765
s_3	0.0089	0.0089	0.00051	0.00051
s_4	0.000006	0.000007	$< 10^{-11}$	0
s_5	$< 10^{-11}$	0	$< 10^{-11}$	0

Table 1: Predicted behavior for GREEDY(d) and average results from 100 simulations with 1 million balls.

The differential equations have the following simple interpretation: for there to be an increase in the number of bins with at least i balls, the d choices of a ball about to be placed must all be bins with load at least $i - 1$, but not all bins with load at least i.

In contrast to Section 2.1, where we could derive a formula for the fraction of empty bins, we are not aware of how to determine explicit formulae for $s_i(t)$ in general. These systems of differential equations can be solved numerically using standard methods, however; for up to any fixed k and t, we can accurately determine $s_k(t)$. By applying Kurtz's theorem (as in Theorem 2) or martingale arguments (as in Theorem 4) one can show that these results will be accurate with high probability.

We also demonstrate that our technique accurately predicts the behavior of the GREEDY(d) algorithm by comparing with simulation results. The first and third columns of Table 1 shows the values of $s_i(1)$ for $d = 2$ and $d = 3$ as caclulated from the differential equations. We use these values as predictions for the process where we throw n balls into n bins. From the predictions with $d = 2$, one would not expect to see bins with load five until billions of balls have been thrown. Similarly, choosing $d = 3$ one expects a maximum load of three until billions of balls have been thrown. These results match our simulations of several hundred runs with up to thirty-two million balls, the largest simulation we have attempted. We also present the averages from one hundred simulations of one million balls for $d = 2$ and $d = 3$, which demonstrate the accuracy of the technique in predicting the behavior of the system. Further simulations reveal that, in general, the solution given by the limiting system of differential equations becomes more accurate as n grows, and the deviation from this solution is small, as one would expect. This accuracy is a marked advantage of this approach; previous techniques have not provided ways of concretely predicting actual performance.

2.3 Relationship to $O(\log\log n)$ bounds

We can also use the differential equations to provide an alternative derivation of a key idea from the proof of the upper bounds on the maximum load of GREEDY(d). The approach of looking at the underlying differential equations provides insight into how the s_k decrease, which is essential to determining the $O(\log\log n)$ bounds.

We begin by focusing on the case where the number of balls n equals the number of bins n, and consider the limiting description given by the differential equations as $n \to \infty$.

Lemma 5 *For the family of differential equations given by (3),* $s_i(1) \leq [s_{i-1}(1)]^d$.

Proof: We wish to know the values of $s_i(1)$. Because the s_i are all non-decreasing over time and non-negative, from (3)

$$\frac{ds_i}{dt} = s_{i-1}^d - s_i^d \leq [s_{i-1}(1)^d]$$

for all $t \leq 1$ and hence by integrating

$$s_i(1) \leq [s_{i-1}(1)]^d. \tag{4}$$

■

One can calculate $s_1(1)$ directly from Theorem 1, and it follows from a simple induction on (4) that

$$s_i(1) \leq [s_1(1)]^{d^{i-1}}.$$

In other words, the $s_i(1)$, which represent the limiting fraction of bins with load at least i after all balls have been thrown, decrease *doubly exponentially*, as the i is in the second level of the exponent. Using Kurtz's theorem one obtains a high probability result for the case of a finite number of balls n.

Theorem 6 *Let w_i^n be the fraction of bins with load at least i when n balls are thrown into n bins using* GREEDY(d). *Then for any $\epsilon > 0$ and fixed i, $w_i^n \leq (1+\epsilon)[s_1(1)]^{d^{i-1}}$ with probability at least $1 - \mathrm{e}^{c(\log i)n\epsilon^2}$ for some suitable constant c.*

Proof: The proof is a direct application of Kurtz's Theorem, with the error bounds as given in [18, Theorem 5.3]. ■

This doubly exponential decrease in the $s_i(1)$ (or, equivalently, of the w_i^n) is a key step of the proof of Azar *et al* in [1], where it is proven via an inductive use of Chernoff bounds. Theorem 6 shows that this induction can be replaced by applying Kurtz's theorem to the differential equations, at least up to any fixed constant value of i. This approach has some advantages. Most importantly, Lemma 5 and the corresponding inductive bound for $s_i(1)$ seem quite natural and make transparent how the s_i decrease. Additionally, using this approach can improve the additive $O(1)$ term in Theorem 4 of [1].

Intuitively, this doubly exponential decrease suggests that if we look at bins with load at least $i^* = \lceil \frac{\log \log n}{\log d} + \gamma \rceil$, where $\gamma = (\log 2 - \log \log(1/s_1(1)))/\log d$, then then

$$s_{i^*+1} \leq [s_1(1)]^{d^{\log \log n/\log d + \gamma}} \leq \frac{1}{n^2},$$

and hence with high probability there will be no bins with load at least $i^* + 1$.

Note, however, that as stated in Theorem 6 the high probability bounds only hold up to any fixed i, and not for values of i up to $\Omega(\log \log n)$. This weakness in the result stems from directly applying the differential equations approach and Kurtz's theorem, which requires the underlying state space to be finite dimensional, and hence loads up to only some fixed constant can be considered. By reverting back to the explicit martingale argument that underlies Kurtz's Theorem, we can circumvent this restriction somewhat (up until the point where s_i is $1/n^{1/2+\delta}$ for any $\delta > 0$), but at some point when s_i is sufficiently small it seems we have to explicitly handle this case directly, as is done in Theorem 4 of [1]. We omit details of these more extensive arguments, since their form would be almost entirely a restatement of Theorem 4 of [1], replacing their use of Chernoff bounds with an equivalent martingale argument. A general framework that would allow us to apply the differential equations in this instance in a more straightforward manner would clearly be appealing, since the differential equations do make clear the behavior of the s_i.

In the case of m balls and n bins, a similar argument to Lemma 5 and Theorem 6 holds when $m = cn$ for some constant c. As in Section 2.1, when $m = cn$, the infinite process runs until time c; if m is not a linear function of n, the time until the process terminates is dependent on n, and Kurtz's theorem cannot be applied. For Lemma 5, the appropriate result becomes

$$s_i(c) \leq c[s_{i-1}(c)]^d.$$

By noting that $s_{c^2+c}(c) \leq 1/(c+1)$ (the fraction of bins with load at least x cannot be more than c/x), one may again inductively show as before that the tails of the loads are doubly exponentially decreasing. Improvements can be made in the constants by using the differential equations to find better starting points than $s_{c^2+c}(c) \leq 1/(c+1)$ for the induction.

3 Conclusion

There are significant advantages to using differential equations to study randomized load balancing problems. The insight one gains about the problem and the numerical accuracy one obtains are quite convincing. Moreover, when the corresponding state spaces are finite dimensional, applying Kurtz's Theorem can yield simple proofs of the limiting behavior. A general framework for dealing with spaces that are not necessarily finite dimensional would greatly simplify using this approach in developing bounds such as the $O(\log\log n)$ bounds of [1]. We expect this approach will find a great deal of further use in the analysis of load balancing schemes, as well as other algorithmic areas.

References

[1] Y. Azar, A. Broder, A. Karlin, and E. Upfal, "Balanced Allocations", *Proceedings of the 26th ACM Symposium on the Theory of Computing*, 1994, pp. 593–602.

[2] D. Achlioptas and M. Molloy, "The Analysis of a List-Coloring Algorithm on a Random Graph", *Proceedings of the 38th IEEE Symposium on Foundations of Computer Science*, 1997.

[3] M. Alanyali and B. Hajek, "Analysis of simple algorithms for dynamic load balancing", preprint. To appear in Mathematics of Operations Research.

[4] N. Alon, J. Spencer, and Paul Erdös, **The Probabilistic Method**, 1992, John Wiley and Sons.

[5] K. Azuma, “Weighted sums of certain dependent random variables”, *Tôhoku Mathematical Journal*, Vol. 19, 1967, pp. 357-367.

[6] A. Czumaj and V. Stemann, “Randomized Allocation Processes”, *Proceedings of the 38th IEEE Symposium on Foundations of Computer Science*, 1997.

[7] D. L. Eager, E. D. Lazowska, and J. Zahorjan, “Adaptive load sharing in homogeneous distributed systems”, *IEEE Transactions on Software Engineering*, Vol. 12, 1986, pp. 662–675.

[8] G. Gonnet, “Expected Length of the Longest Probe Sequence in Hash Code Searching”, *Journal of the ACM*, Vol. 28, 1981, pp. 289–304.

[9] B. Hajek, “Asymptotic Analysis of an Assignment Problem Arising in a Distributed Communications Protocol”, *Proceedings of the 27th Conference on Decision and Control*, 1988, pp. 1455–1459.

[10] R. M. Karp, M. Luby, and F. Meyer auf der Heide, “Efficient PRAM Simulation on a Distributed Memory Machine”, *Proceedings of the 24th ACM Symposium on the Theory of Computing*, 1992, pp. 318–326.

[11] R. M. Karp and M. Sipser, “Maximum Matchings in Sparse Random Graphs”, *Proceedings of the 22nd IEEE Symposium on Foundations of Computer Science*, 1981, pp. 364-375.

[12] R. M. Karp, U. V. Vazirani and V. V. Vazirani, “An Optimal Algorithm for On-line Bipartite Matching”, *Proceedings of the 22nd ACM Symposium on the Theory of Computing*, 1990, pp. 352–358.

[13] T. G. Kurtz, “Solutions of Ordinary Differential Equations as Limits of Pure Jump Markov Processes”, *Journal of Applied Probability* Vol. 7, 1970, pp. 49-58.

[14] T. G. Kurtz, **Approximation of Population Processes**, SIAM, 1981.

[15] M. Luby, M. Mitzenmacher, M. A. Shokrollahi, D. Spielman, and V. Stemann. “Practical Loss-Resilient Codes”, *Proceedings of the 29th ACM Symposium on the Theory of Computing*, 1997, pp. 150–159.

[16] M. Mitzenmacher, “Load Balancing and Density Dependent Jump Markov Processes”, *Proc. of the* 37^{th} *IEEE Symp. on Foundations of Computer Science*, 1996, pp. 213–222.

[17] M. Mitzenmacher, “The Power of Two Choices in Randomized Load Balancing”, Ph.D. thesis, University of California, Berkeley, September 1996.

[18] A. Shwartz, A. Weiss, **Large Deviations for Performance Analysis**, 1995, Chapman & Hall.

[19] V. Stemann, “Parallel Balanced Allocations”, *Proceedings of the Eighth Annual ACM Symposium on Parallel Algorithms and Architectures*, 1996, pp. 261–269.

[20] N.D. Vvedenskaya, R.L. Dobrushin, and F.I. Karpelevich. “Queueing System with Selection of the Shortest of Two Queues: an Asymptotic Approach”, *Problems of Information Transmission*, Vol 32, 1996, pp. 15–27.

[21] N. C. Wormald, “Differential Equations for Random Processes and Random Graphs”, *Annals of Appl. Prob.*, Vol 5, 1995, pp. 1217–1235.