基于非对称 balls-into-bins 的高效平衡负载分配模型*

皇甫先鹏,罗雪山
(国防科技大学 信息系统与管理学院,湖南 长沙 410073)

摘 要:在大规模数据中心和P2P覆盖网络等复杂网络负载平衡分配中,前人提出了多种多样的负载分配方法,但许多方法为了达到更好的平衡负载指标,追求越来越复杂的算法,使得时间复杂度和算法复杂度很难控制在合理的范围之内。本文在研究了经典 balls-into-bins、Azar balls-into-bins 和 balls into non-uniform bins 等模型的基础上,提出了一种新颖高效的非对称 balls-into-bins 平衡负载分配模型,该模型具有异构的 balls、异构的 bins,以及不同的 bin 选择概率,能以很高的概率将最大负载均衡地控制在合理的范围内,通信负载很小,且具有很好的可扩展性,通过拓展,该模型在负载平衡的诸多领域都将有广阔的应用空间。

关键词:非对称 balls-into-bins;平衡负载分配;复杂系统
中图分类号:TP393 **文献标志码**:A **文章编号**:1001-2486(2013)03-0067-05

An efficient and balanced load allocation model based on non-uniform balls-into-bins

HUANGFU Xianpeng, *LUO Xueshan*
(College of Information System and Management, National University of Defense Technology, Changsha 410073, China)

Abstract: In balanced load allocation problem in complex systems like large-scale data center and P2P overlay network, the various load allocation methods is proposed. In order to achieve better balanced load index, many methods, however, are in pursuit of more and more complicated algorithms, which makes the time and algorithm complexity hard to control. Based on the study of the original balls-into-bins model, Azar balls-into-bins model and balls into non-uniform bins model, the paper brings forward an efficient and balanced non-uniform balls-into-bins load allocation model, which is provided with heterogeneous balls, heterogeneous bins and different bin selection probabilities. The model can achieve rational largest load with high probabilities, at the cost of little time and algorithm complexity. The model is extensible and can be applied in many domains.

Key words: non-uniform balls-into-bins; balanced load allocation; complex system

在大规模数据中心[1-3]和P2P覆盖网络[4-6]等复杂网络中,寻找简单、高效、平衡的负载分配方法始终是近年来研究的热点领域。但一种误区逐渐地显现,为了取得良好的实验结果,负载分配算法越来越复杂,计算周期也越来越长,虽然最大负载及负载偏差幅度等指标得到了提升,但离应用却渐行渐远,时间复杂度和算法复杂度已经超出了复杂系统所能接受的范围。本文另辟蹊径,提出了一种新颖高效的非对称 balls-into-bins 平衡负载分配模型,该模型将计算和服务等用户请求映射为 balls,将系统资源和计算能力等映射为 bins,那么负载分配问题就可以看作是如何将 balls 更加均匀地分配到各个 bins 中去。该模型具有异构的 balls、异构的 bins,以及不同的 bin 选择概率,能以很高的概率将最大负载均衡控制在合理的范围内,其时间复杂度和算法复杂度均很小,具有很好的可扩展性,可以适用于不同规模大小的复杂系统负载分配。图1为复杂系统中的平衡负载分配问题。

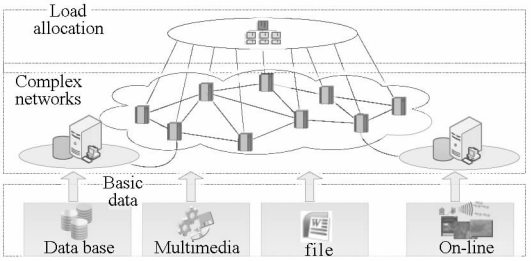

图1 复杂系统中的平衡负载分配问题

Fig. 1 The balanced load allocation of complex systems

* **收稿日期**:2012-09-20
基金项目:国家自然科学基金资助项目(61070216, 71071160, 61170284);国家部委资助项目;高等学校博士学科点专项科研基金项目(20114307110011);湖南省研究生创新资助项目(CX2010B022);国防科大研究生创新资助项目(B100501)
作者简介:皇甫先鹏(1982—),男,河南商丘人,博士研究生,E-mail:xphuangfu@gmail.com;
罗雪山(通信作者),男,教授,博士,博士生导师,E-mail:xsluo@nudt.edu.cn

1 balls-into-bins 模型

本节从模型定义、参数选择和最大负载三个方面讨论经典 balls-into-bins 模型、Azar balls-into-bins 模型和 balls into non-uniform bins 模型的性质。

1.1 经典 balls-into-bins 模型

假如顺序地将 n 个 balls 投入到 n 个 bins，策略是从 n 个 bin 中随机独立均匀地(independently and uniformly at random，i. u. a. r.)选择任意一个 bin。该经典 balls-into-bins 问题在相关文献中进行了广泛的分析，最经典的结论就是：当抛球结束的时候，以很高的概率(也就是以概率 $1-o(1)$)，n 个 bin 中最大负载为 $(1+o(1))\log n/\log\log n$ 个 ball。

1.2 Azar balls-into-bins 模型

假如不是随机地从 n 个 bin 中任意选择一个，而是将球抛入负载最小的那个 bin，最大的负载将迅速减少为 $\log\log n/\log d+O(1)$。在随机分配负载过程中，这个很小的抛球选择策略上的变化，比经典抛球理论模型成指数型地降低了最大负载。

更进一步，当 m 个 ball 顺序地抛入 n 个 bin 中时，从中 i. u. a. r. 选择 $d\geqslant 2$ 个 bin，将球抛入负载最小的 bin 中，当所有球均被抛完的时候：

当 $n\to\infty$ 且 $m\geqslant n$ 时，bin 中的最大负载将以很高的概率 $1-o(1)$ 为 $(1+o(1))\ln\ln n/\ln d+\Theta(m/n)$。

当 $n\to\infty$ 且 $m=n$ 时，bin 中的最大负载将以很高的概率 $1-o(1)$ 为 $\ln\ln n/\ln d+\Theta(1)$。

其他任何将球随机抛入 d 个 bin 中的策略都会导致负载球数的增加。

同时，该抛球理论也证明，当 $m\geqslant n$ 时，负载的偏差独立于抛球的数量 m[7]。

1.3 异构 bins 和不同 bin 选择概率的 balls into non-uniform bins 模型

Byers 等在文献[8]研究了以下问题：由于受像 Chord 这样的 P2P 网络属性的驱动，bin 的选择不再是 i. u. r. 。每个 bin 用 Chord 环上的一个点表示，球随机地选择 $d\geqslant 2$ 个环上的点，每个点和最靠近它的 bin 相关联，那么每个球就将和 d 个可选 bin 的集合中存储其他项目最少的 bin 相关联。假如最大弧的长度能达到大于平均弧长度的 $\log n$，最大负载将同样能以很高的概率达到 $\ln\ln n/\ln d+\Theta(1)$。同时也证明了，即使 bin 选择的概率比平均概率大 $O(\log n)$ 时，最大负载也只有很小的偏差变动。

Berenbrink 等在文献[8]中假定系统包含异构的 bin，其中每个 bin i 具有容量 c_i，并用 bin 中的 ball 数目除以 bin 的容量来衡量每个 bin 的负载。分析表明，Bin 的最大负载并不依赖于 bin 的总体容量 $C=\sum_{i=1}^{n}c_i$，而是依赖于 bin 的数量，当抛球数量 $m=\sum_{i=1}^{n}c_i$ 时，那么 Bin 的最大负载是 $\log\log n+O(1)$。表 1 为 balls-into-bins 模型及负载总结。

表 1 Balls-into-bins 模型及负载

Tab. 1 The balls-into-bins models and load

名称	Balls(m)	Bins(n)	可选择 bin 数(d)	Bin 选择概率(Pr)	负载(L)
经典 balls-into-bins	同构 $m=n$	同构	/	/	$(1+o(1))\log n/\log\log n$
Azar balls-into-bins	同构 $m\geqslant n$	同构	$d\geqslant 2$	$1/n$	$(1+o(1))\ln\ln n/\ln d+\Theta(m/n)$
	同构 $m=n$	同构	$d\geqslant 2$	$1/n$	$\ln\ln n/\ln d+\Theta(1)$
Balls into non-uniform bins	同构 $m=\sum_{i=1}^{n}c_i$	异构，容量 c_i	$d\geqslant 2$	c_i/C	$2\cdot\log\log n+O(1)$

2 非对称 balls-into-bins 模型及负载

上节讨论了几种不断演化的 balls-into-bins 模型，从经典模型中任意选择一个相同的 bin 到以不同的概率选择 $d\geqslant 2$ 个不同构的 bin，最大 bin 负载以及负载的偏差幅度都显著降低了，而且模型的适用场景及范围也更加广泛。本节将延续 balls-into-bins 模型的发展脉络，考虑在不同构的 balls、不同构的 bins，以及不相同的 bin 选择概率的情况下的非对称的 balls-into-bins 模型。

2.1 非对称 balls-into-bins 模型的符号和定义

假设有 m 个不相同的 balls，顺序地抛入 n 个

不相同的 bins 中,其中每个 bin 具有正整数的容量属性“Capacities”,分别用 $c_1,\cdots,c_n$ 表示 n 个 bin 的容量,并用 $C=\sum_{i=1}^{n}c_i$ 表示 n 个 bin 的容量之和;同时每个 ball 具有正整数的体积属性“Volume”(此处体积的含义直观上应理解为 ball 在 bin 中占据的立方体的体积,而不是狭隘的圆球的体积),分别用 $v_1,\cdots,v_m$ 表示 m 个 ball 的体积,并用 $V=\sum_{j=1}^{m}v_j$ 表示 m 个 ball 的体积之和。

考虑单个 ball 的体积相较于单个 bin 的容量是较小的情形,这样对于模型建立以及模型的应用也更有意义。首先考虑可选择 bin 数 $d=2$,并且 ball 的体积之和 $V=\sum_{j=1}^{m}v_j$ 和 bin 的容量之和 $C=\sum_{i=1}^{n}c_i$ 相等的情况,也就是 $V=C$ 或 $\sum_{j=1}^{m}v_j=\sum_{i=1}^{n}c_i$。在此基础上,能较容易地扩展至 $d>2$,并且 $(V=\sum_{j=1}^{m}v_j)>(C=\sum_{i=1}^{n}c_i)$ 和 $(V=\sum_{j=1}^{m}v_j)<(C=\sum_{i=1}^{n}c_i)$ 的情形。

假如 m_j 个 ball(体积 $V=\sum_{j=1}^{m}v_j$)被分配到第 n_i 个 bin(容量 $c_i\geqslant 1$)中,那么第 n_i 个 bin 的负载可以用分配到该 bin 的 ball 的体积之和除以该 bin 的容量来表示,也就是 $l_i=\lceil(V=\sum_{j=1}^{m}v_j)/c_i\rceil$。

那么,在 balls 和 bins 的体积和容量均为异构的情况下,从 d 个可选的 bin 中选择其中一个 bin 的选择概率和经典 balls-into-bins 的概率 $1/n$ 也必将不同,此时第 i 个 bin 的选择概率由其容量 c_i 占据整个容量 C 的比值表示,也就是 $Pr_i=c_i/C=c_i/\sum_{i=1}^{n}c_i$ 来表示,该选择概率正比于每个 bin 的容量。故对于 d 个可选的 bin,也可以用 bin 的容量 c_i 占据 d 个可选的 bin 的容量之和的比值来表示,同样正比于 d 中每个 bin 的容量,也就是 $Pr_i=c_i/d$ 个可选 bin 的容量。两种计算方式对于 bin 的选择概率是一致的。图 2 为非对称 balls-into-bins 模型。

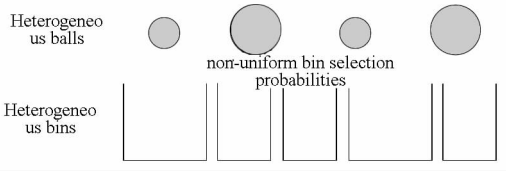

图 2　非对称的 balls-into-bins 模型

Fig. 2　Non-uniform balls-into-bins model

定义 1　(Majorization 优于,$\succ$)给定两个向量 $P=(p_1,\cdots,p_n)$ 和 $Q=(q_1,\cdots,q_m)$,其中 $\bar{p}_i$ 和 $\bar{q}_i$ 分别是标准化向量 $\bar{P}$ 和 $\bar{Q}$ 的第 i 项,对于所有 $k=1,\cdots,\min\{m,n\}$,当且仅当 $\sum_{i=1}^{k}\bar{p}_i\geqslant\sum_{i=1}^{k}\bar{q}_i$ 时,我们说 P 优于 Q,记为 $P\succ Q$。

根据两个向量之间的比较关系,可以容易地扩展到系统以及过程的优势比较。

定义 2　(负载优于,为了避免符号过多导致混乱,同样用 $\succ$ 表示)A 和 B 分别为将 m 个 ball 投入 n 个 bin 的抛球过程,其中总容量 $C_A=C_B$ 和总体积 $V_A=V_B$。考虑 $d=2$,向量 $\tau=(\tau_1,\cdots,\tau_{2m})$ 表示对于 m 个球的所有 $2m$ 个随机地 bin 选择,其中 $\tau_j\in[n]$。对于所有 $j=1,\cdots,m,\tau_{2j-1}$ 和 τ_{2j} 表示第 j 个 ball 的两个 bin 选择。$L^A(\tau)$ 和 $L^B(\tau)$ 分别表示用随机选择 τ 的方式分配 A 和 B 的负载向量,那么

(1) 对于所有的随机选择 $\tau\in[n]^{2m}$,A 和 B 间存在单调映射函数 $f:[n]^{2m}\to[n]^{2m}$,表示 A 优于 B,记为 $A\succ B$。

(2) 对于所有的随机选择 $\tau\in[n]^{2m}$,A 和 B 的最大负载间存在单调映射函数 $f:[n]^{2m}\to[n]^{2m}$,表示 A 的最大负载优于 B 的最大负载,记为 $L^A(\tau)\succ L^B(f(\tau))$。

集合 $[n]$ 表示 $\{1,\cdots,n\}$,我们说事件 A 以很高的概率发生,意为对于 $a>0$,事件 A 发生的概率 $Pr[A]\geqslant 1-n^{-a}$。

2.2　非对称 balls-into-bins 模型的性质证明

非对称 balls-into-bins 模型由于假定的 ball 不同、bin 不同以及 bin 的选择概率亦不同,因此使用贪婪算法从 d 个可选 bin 中选择负载最小的 bin。由于非对称 balls-into-bins 模型放宽了约束条件,其应用对象和范围也有了很大的扩展,其性质也区别于表 1 中的模型。下面讨论非对称 balls-into-bins 模型的最大负载的性质。

定理 1　对于任意常量 $l>0$。选择适当的常量 $\varepsilon=\varepsilon(l)$ 和 $d>0$。那么以概率 $1-1/n^l$,每个 bin 的负载最多为 $2(1+\varepsilon)+1$。

证明　这是 Chernoff 边界的一个应用,考虑单个 ball 的体积相较于单个 bin 的容量是较小的情形,假如 $c_i\geqslant d\log n$ 时,$m=C=\sum c_i$。一个 ball 抛入一个 bin b_i 的概率最多为 $2c_i/C=2c_i/m$。在 m 个 ball 抛出后,抛入 bin b_i 的 ball 的期望个数小于 $(2c_i/m)\times m=2c_i$。设 m_i 为选择 bin i 作为被抛入对象的 ball 的数量,那么根据 chernoff 边界定理,

$$Pr(m_i \geqslant (1+\varepsilon)\cdot 2c_i) \leqslant e^{-\varepsilon^2\cdot 2c_i/3}$$
$$\leqslant e^{-\varepsilon^2\cdot 2\cdot d\log n/3}$$
$$= n^{-\varepsilon^2\cdot 2\cdot d/3} = n^{-l}$$

那么,选择合适的 $d = d(\varepsilon)$,以至少 $1 - 1/n^l$ 的概率,bin 最多能被抛入 $(1+\varepsilon)\cdot 2c_i$ 个 ball,这也就是 bin 中抛入球的数量的上界。那么,选择合适的 ε 和 d,此时的负载以高概率 $1 - 1/n^l$,最多为

$$l_i = \lceil m_i/c_i \rceil = \lceil (1+\varepsilon)\cdot 2c_i/c_i \rceil$$
$$= \lceil 2\cdot(1+\varepsilon) \rceil$$
$$\leqslant 2\cdot(1+\varepsilon) + 1$$

引理1 设 A 为总体积为 V 的异构 ball 且总容量为 C 的异构 bin 的抛球系统,B 为单位容量为 1 的总容量仍为 C 的同构抛球系统,那么 $A \prec B$。

证明 将 $V = \sum_{j=1}^{m} v_j$ 均分为单位体积的 ball,根据文献[8]的引理1得证。

定理2 对于总体积为 V 的异构 ball 且总容量为 C 的异构 bin 的非对称 balls-into-bins 模型,假如将 m 个 ball 分配到总容量为 m 的 $n \leqslant m$ 个 bin 中,那么以概率 $1 - 1/n^l$,最大负载为 $2\cdot \text{loglog}n + O(1)$。

证明 考虑一个 ball i 选择两个 bin b_1 和 b_2。投入一个容量为 c 的 bin 的概率为 c/m。那么根据引理1,两个选择 bin 负载以概率 $1 - 1/(m_s)^l \geqslant 1 - 1/n^l$ 最大为 $\text{loglog}m_s + O(1) = \text{loglog}n + O(1)$。因此,bin 的最大负载至多为 $\text{loglog}n + O(1) + 2\cdot(1+\varepsilon) + 1 = \text{loglog}n + O(1)$。对于总体积为 V 的异构 ball 且总容量为 C 的异构 bin 的非对称 balls-into-bins 模型,应用定理和引理1,得到最大负载为 $\text{loglog}(m) + O(1) \leqslant \text{loglog}(n^2) + O(1) = 2\cdot\text{loglog}(n) + O(1)$。

3 基于非对称 balls-into-bins 模型的负载分配和实验分析

考虑复杂网络中动态资源分配的场景,一个用户或一个进程必须从一定数目的完成相似功能使命的在线资源中选择其中负载最小的资源,直观的做法是用户将查看所有资源的实时负载,而后相互比较,选择负载最小的资源。这个过程由于用户向每个资源发出实时负载查询请求,同时每个资源均要反馈用户的负载查询请求,如果资源的数目较大,那么计算最小负载也需要消耗一定的计算资源,所以整个过程会大幅增加通信和计算的负载。还有一种方法就是将任务随机地分配给其中的一个资源,而不管其实时负载的大小,此种方法的通信负载最小,但是如果每个用户都随机地选择资源,其资源负载的不平衡性将会非常大,负载偏差幅度也变得不可控制,甚至会达到对数级数的程度。因此,在负载系统的动态资源分配中,将资源分配与非对称 balls-into-bins 模型进行映射,资源即为 bins,负载请求即为 balls。

定理3 复杂系统中应用非对称 balls-into-bins 模型动态资源分配的最大负载的数学期望以很高概率 $(1 - n^{-a})$ 是应用 Azar balls-into-bins 模型负载分配算法的 $2\cdot\text{loglog}n/\ln\ln n$。

考虑 $m = n$ 时,Azar balls-into-bins 模型负载以高概率 $(1 - n^{-a})$ 为 $\ln\ln n/\ln 2$,用随机变量 X 来表示该数学期望,也就是 $X = \ln\ln n/\ln 2$。

非对称 balls-into-bins 模型负载则以高概率 $(1 - n^{-a})$ 为 $2\cdot\text{loglog}n$,用随机变量 Y 来表示该数学期望,也就是 $Y = 2\cdot\text{loglog}n$。

定义随机变量 $Z = Y/X$,那么随机变量 Z 当 $m = n$ 时出现 $(2\cdot\text{loglog}n)/(\ln\ln n/\ln 2)$ 的概率是

$$P(Z) = (2\cdot\text{loglog}n/\ln\ln n) \mid ((X = \ln\ln n/\ln 2) \cap (Y = 2\cdot\text{loglog}n))$$
$$= P(X = \ln\ln n/\ln 2)\cdot P(Y = 2\cdot\text{loglog}n)$$
$$= (1 - n^{-a})\cdot(1 - n^{-a})$$
$$= 1 - n^{-a}$$

对非对称 balls-into-bins 模型及 Azar balls-into-bins 模型进行如下仿真验证。假设 bin 的数目从5到50,考虑在 $m = n$ 的情况下,balls 和 bins 的容量和体积服从随机分布,比较 Azar 和非对称 balls-into-bins 的负载情况。从图3看出,对于应用 Azar balls-into-bins 模型的算法,其最大负载落于1.8~2.7,并且随着负载请求的增加,其负载变化较快,而非对称 balls-into-bins 模型负载落于0.7~1.5,其负载变化率也更为缓和。所以,非对称 balls-into-bins 模型在复杂系统中能达到高效平衡的负载分配。

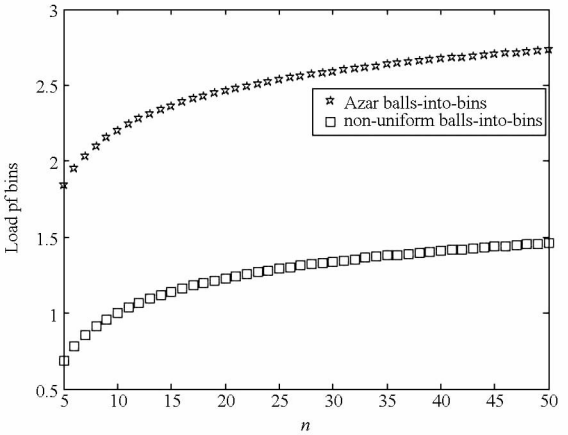

图3 Azar 和非对称 balls into bins 模型负载比较

Fig. 3 The load comparison of Azar and non-uniform balls into bins model

4　总结

本文首先从模型参数和主要结论等方面介绍了经典 balls-into-bins 模型、Azar balls-into-bins 模型及 balls into non-uniform bins 模型,沿着这一思路,继续拓展模型参数和应用范围,研究了非对称 balls-into-bins 模型,该模型放宽了模型参数的约束,ball 和 bin 均可以是异构的,bin 的选择概率也依照 bin 自身容量的不同而不相同,其适用范围更加广泛。通过研究,非对称 balls-into-bins 模型具有优异的性质,在最大负载及负载偏差幅度方面表现突出,其时间复杂度和算法复杂度均很小。该模型在负载平衡的诸多应用领域都将有广阔的应用空间。

参考文献(References)

[1] Fares M A, Loukissas A, Vahdat A, A scalable, commodity data center network architecture [C]//Proceedings of SIGCOMM, Seattle, Washington, USA, 2008:63 - 74.

[2] Guo C, Wu H, Tan K, et al. Dcell: A scalable and fault-tolerant network structure for data centers[C]//Proceedings of SIGCOMM, Seattle, Washington, USA,2008:75 - 86.

[3] Guo G,Lu G,Li D,et al. Bcube: A high performance, server-centric network architecture for modular data centers[C]// Proceedings of SIGCOMM, Barcelona, Spain, 2009:63 - 73.

[4] Wang C,Xiao L, Liu Y,et al. Dicas: an efficient distributed caching mechanism for p2p systems[J]. IEEE Transactions on Parallel and Distributed Systems, 2006,17(10):1097 - 1109.

[5] Liu Y. A two-hop solution to solving topology mismatch[J]. IEEE Transactions on Parallel and Distributed Systems, 2008, 19(11):1591 - 1600.

[6] He Y, Liu Y. Vovo: Vcr-oriented video-on-demand in large-scale peer-to-peer networks[J]. IEEE Transactions on Parallel and Distributed Systems, 2009,20(4):528 - 539.

[7] Azar Y, Broder A Z, Karlin A R, et al. Balanced allocations [C]//Proceedings of the 26th ACM Symposium on Theory of Computing,1994:593 - 602.

[8] Berenbrink P, Brinkmann A, Friedetzky T, et al. Balls into non-uniform bins [C]//Proceedings of the 24th IEEE International Parallel and Distributed Processing Symposium (IPDPS), 2010:1 - 10.