效率与公平：最大加速比与博弈论上的公平预示下一代云计算新形态

中国科学院计算技术研究所副研究员　刘宇航

摘　要：云计算是近十年兴起的一种计算模式，具有与超级计算等传统计算模式不完全相同的价值导向。面对多样化的大量用户应用，如何将纷繁复杂的数据中心资源分配给各个用户应用，传统计算模式的价值导向是高通量、高吞吐率、高性能，述评的论文还考虑了公平，但不是“平均主义”意义上的公平，而是在考虑了系统整体性能基础上（在Amdahl定律逆定律的指导下将处理器分配给能换来最大加速比的应用），博弈论意义上的公平，对云计算数据中心的资源分配具有重要的理论意义和实用价值，从某种程度上预示了下一代云计算的一种新形态。

关键词：博弈论　市场均衡　公平　性能　云计算　资源分配

1 引 言

HPCA 2018会议的四篇最佳论文之一[3]，“数据中心时代的Amdahl定律：一个用于公平分配处理器的市场”由三位作者合作撰写。第一和第三作者来自美国杜克大学，第二作者来自VMware公司。

其中，第一和第三作者主题类似的研究成果曾获得ASPLOS 2014会议的最佳论文[2]。按照中国计算机学会的推荐列表，HPCA和ASPLOS都是计算机系统结构领域最高水平的会议，连续两次以同一主题获得最佳论文，在一定程度上反映了当前学术界的热点或倾向。

共享与独占是对立统一的两个方面。应用对性能有高要求，即期望能在短时间内被执行完毕。从理论上，应用被分配的处理器数量越多，潜在的可能的硬件并行度就越大。但是，同时运行的应用数量较多，这些应用在需求和重要性上有较大差异，系统的处理器总量有限，如何在这些应用之间分配处理器，关系到每个应用的性能（即执行时间的倒数）、系统的效率、公平性。

面对多样化的用户应用，如何将纷繁复杂的数据中心资源分配给各个用户应用，传统的价值导向是高通量、高吞吐率、高性能。在这篇论文中，作者还考虑了公平，但不是“平均主义”意义上（即资源完全平均划分或减速比相同）的公平，而是考虑了系统整体性能基础上（在Amdahl定律逆定律的指导下将处理器分配给能带来最大加速比的应用）的在博弈论意义上的公平。这对云计算数据中心的资源分配具有重要的理论意义和实用价值，从某种程度上预示了下一代云计算的一种新形态。

我们此次受邀撰写这篇论文评述，有以下研究基础和契机：

（1）这篇论文使用了效用函数，我们对超级计算三大定律Amdahl定律、Gustafson定律和Sun-Ni定律进行了深入研究，在SC 2015会议（全球超级计算领域顶级会议，属于CCF A类会议）上发表的长文[9]中提出的性能公式与述评论文中的效用函数有异曲同工之处；

（2）这篇论文使用了Amdahl定律的逆定律来估计用户应用的可并行化的比例，我们研读了Karp-Flatt方法的原文[10]（发表于1990年《美国计算机通讯》），述评的这篇论文使用了该方法估算应用的可并行化比例。

(3)我们精读并深入分析了冯·诺依曼所著的《计算机与人脑》,归纳整理为10个要点[4],其中一个就是Amdahl定律的雏形,可见这一定律具有历久弥新的基础性作用;

(4)我们正在研究数据中心资源的分配,发现提高资源利用率和保证服务质量之间存在着尖锐矛盾,为此提出了标签化体系结构[8],并提出了低熵云计算基础理论和DIP猜想[11];

(5)我们对2018年《美国计算机学会通讯》上一篇关于Amdahl定律与尾延迟之间的文章[1],进行了深入分析。

整体上来说,并行是节省时间的基本方式,Amdahl定律及其逆定律具有重要作用,这一点不仅对超级计算成立,也对云计算成立。

2 要点归纳

价值导向(同时兼顾性能与公平)、划分方式(基于授权的划分方式)、效用函数(最大加速比)、市场均衡(将数据中心中的资源分配的过程映射为市场中的商品交易过程)是述评论文的四个基本要点。

2.1 共享资源时的不同价值导向

任何不同的资源分配策略,均分别服务于各自的目标函数,更本质地说,是服务于各自的价值导向。性能、效率、资源利用率、公平性、用户体验或者它们的组合,这些都是资源分配时可以选择的优化目标。不同用户的任务之间可能具有不同的重要性和紧急性,而且不同的用户为系统的构造、运行、维护承担了大小不同的费用,因此需要考虑共享动机(Sharing Incentive,SI)和公平性的问题。

传统的研究,默认所有用户必须参与共享资源。但是,事实上,如果用户从一种共享资源分配机制中获取的效用低于平均分配资源时的效用,那么这个用户就没有参与共享的动机,他宁可选择平均分配资源(也就是平均划分然后各自独占的方式)。

公平在直觉上与在博弈论上具有不同的定义,值得研究和探讨。有的研究认为,保持任务之间的性能损失(用减速比量化)一样,就是公平。但是,在博弈论中,一个公平的分配被定义为无嫉妒(envy-free, EF)而且帕累托高效(Pareto Efficient, PE)。所谓无嫉妒,就是如果让用户自己选择,每一个用户都只会选择自己的当前分配,而不是别人的分配。所谓帕累托高效,就是不能在不损害其他用户的效用的前提下,去改善一个用户的效用。通俗地说,如果已经达到了帕累托高效,再继续调整就相当于“拆东墙补西墙”。显然博弈论的定义相对传统的定义,具有更坚实的理论基础。

2.2 三种划分共享资源的方式

系统在多用户之间划分份额的方式包括预留、优先级、授权,具体分析如下:

(1)预留(Reservation):根据用户的预约请求进行资源分配。如果用户预约请求的资源超过了自己的实际需求,资源将不能被充分利用。

(2)优先级(Priorities):根据用户(任务)重要性的相对大小,进行资源分配,如果两个用户同时竞争同一资源,资源将分配给高优先级的用户。这种方式将会使用户之间的相互干扰暴露给用户(特别是优先级较低的用户),用户获得的性能具有不确定性。

(3)授权(Entitlements):授权这种方式,不是生硬地按照每个用户的预约请求去分配资源,而是每个用户拥有一个最低的分配额(所以具有优先级方式不具有的隔离性),没有被利用的资源被重新分配(所以相对预留方式能够提高资源利用率)。举一个实例,谷歌(Google)没有使用授权而使用优先级方式管理数据中心的资源,后果是不能保证用户之间的隔离。

整个处理器资源分配机制分为两个部分,一个是效用函数,一个是市场。这两个部分被协同设计,来快速发现市场均衡。

表1 三种资源分配方式的特点比较

	预留	优先级	授权
隔离性	✓	×	✓
资源利用率	×	✓	✓

2.3 Amdahl定律的逆定律与Amdahl效用函数

并行部分的比例比较难确定，程序员很少知道所编写算法或代码的并行部分的比例。但是，可以根据Amdahl定律的逆定律，通过测量加速比来估计并行部分的比例 F 。

Amdahl定律的逆定律 设 p 表示处理器数量，加速比 s_p 具有式(1).

$$s_p=\frac{T(p)}{T(1)}=\frac{T(1)f+\frac{T(1)(1-f)}{p}}{T(1)}=f+\frac{1-f}{p} \quad (1)$$

整理式(1)，可得式(2)，

$$f=\frac{1/s_p-1/p}{1-1/p} \quad (2)$$

由式(2)可得并行部分比例与加速比的关系式，如式(3)所示，

$$F=1-f=\frac{1-1/s_p}{1-1/p} \quad (3)$$

公式(3)被称为Amdahl定律的逆定律，用于估计应用在系统上的并行部分的比例。Amdahl效用函数定义为加速比的加权平均值，如式(4)，其中 w_{ij} 为用户 i 在服务器 j 上的任务量的比例。

$$u_i(x_i)=\frac{\sum_{j=1}^{m}w_{ij}F_{ij}(x_{ij})}{\sum_{j=1}^{m}w_{ij}} \quad (4)$$

2.4 市场模型与市场均衡

使用市场理论来分配处理器资源，具有多种性质。首先，这是一种有助于共享的方案，即每个用户总是可以获得自己的授权份额，有时还可以获得更多。其次，这是帕累托最优(Pareto-efficient)的，即不存在其他方案使得在不损害其他用户的收益的前提下，能使某一用户获益。第三，这是一种防范欺诈(Strategy-proof)的方案，用户众多且相互竞争，没有用户可以通过误报效用获得收益。

市场模型 用户 i 的分配向量 $x_i=(x_{i1},\cdots,x_{im})$，$x_{i1}$ 是在第1个服务器上分配的处理器核数，x_{im} 是在第 m 个服务器上分配的处理器核数。用户 i 的效用为 U_i，

$$\max\ u_i(x_i),\ s.t.\ \sum_{j=1}^{m}x_{ij}p_j\le b_i$$

市场均衡 在市场均衡中，所有的用户都得到最优的分配，处理器没有过剩或赤字。价格向量 $p*=(p_j^*)$，分配向量 $x*=(x_{ij}^*)$，在满足以下两个条件时构成一个均衡：(1)市场清空：每个服务器中所有的处理器核心都被分配。(2)最优分配：在满足每个用户预算的前提下，最大化效用。可见，“市场均衡”的定义之中包括的条件暗含了资源被充分利用(由“市场清空”规定)和方案经过了优化搜索已经最优(由“最优分配”规定)。

C_j 表示服务器 j 上的处理器数量。$\sum_j C_j p_j^*=B$，B 是所有用户预算的总和。b_i 为用户 i 的预算。用户 i 在服务器 j 上被授权的处理器数量为 $x_{ij}^{ent}=(b_i/B)C_j$，从这个式子可以看出，用户 i 被授权的处理器数量 x_{ij}^{ent} 正比于自己的预算 b_i。

用户在不同服务器上划分预算的标准：正比于从不同服务器上获取的效用。市场在不同用户上划分资源的标准：正比于不同用户的投标。根据1954年Arrow和Debreu发表的文献“Existence of an equilibrium for a competitive economy”(“在竞争经济中存在着均衡”)中的结论，因为Amdahl效用函数是连续的且是凹的，所以费希尔(Fisher)市场均衡一定存在。

系统的进展是多用户进展的加权平均，其中权值为用户预算与全部预算的比例。

3 五种资源分配算法的比较

该文提出的算法与其他四种算法进行了比较。

(1)Amdahl投标算法(Amdahl Bidding, AB)：这是该文提出的算法，用户根据效用和服务器价格对处理器进行投标，在此过程中一些处理器核心会从并行度较低的作业移动分配到并行度较高的作业上。

(2)贪心算法(Greedy, G)：这是一个以性能为中心的机制，猜测在不同数量的处理器核心上的加速比，贪心地将每一个处理器核心分配到产

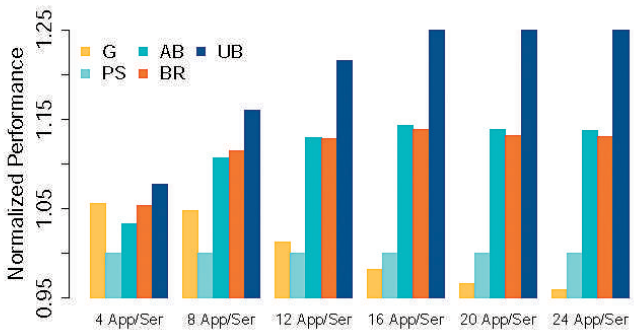

图1　横轴表示负载密度(每个服务器上的应用数量),从左到右负载密度依次增大,资源稀缺程度越来越高,纵轴表示归一化的系统性能,图中第3个矩形对应AB。

生最大加速比或进展的负载上。

(3)上界算法(Upper-Bound, UB): 与贪心算法一样,这也是一个以性能为中心的机制,但是它的目标是最大化系统进展,根据系统进展的定义,这个机制在追求性能时偏向具有较大预算和授权的用户。

(4)按比例共享算法(Proportional Sharing, PS) :对每一个服务器,正比于用户的授权,按比例进行处理器的分配。如果一个用户在一个服务器上没有进行计算,他的份额将在该服务器上的其他用户之间按授权比例进行分配。这个算法在各个服务器内部严格遵循了授权,但在多服务器整体上违反了授权。

(5)最佳响应算法(Best Response, BR):与AB算法一样,这个算法是一个市场机制,在公平性和性能之间实现平衡。用户对资源进行投标,随后市场宣布新价格,然后用户按照内部点方法(interior point method)优化投标,这个过程迭代地进行下去,直至算法收敛。由于Amdahl效用函数是凹函数,内部点方法可以在多项式时间内发现最优投标。

Amdahl投标算法(AB)与最佳响应算法(BR)有重要区别,体现在更新投标的开销和适合的系统规模方面。首先,在更新投标的开销方面,AB的开销较小,原因是AB在新价格时,通过一个清晰明确的方程来更新投标,相比之下,BR通过求解一个优化问题来更新投标,在大规模系统中,BR的做法会导致极高的开销。其次,AB更适合大规模高竞争系统,AB在用户之间自由竞争、接受但不影响价格(price-taking,即假设投标不显著影响价格)时发现费希尔(Fisher)市场均衡,相比之下,BR在用户自己意识到他们的投标可以改变价格(这一意识会影响自己投标)时发现纳什(Nash)市场均衡。在小规模系统中,单个用户的投标更可能改变价格。

从图1可以看出:

(1)AB的性能比PS的要高,原因是PS只关注授权,忽略了性能,有些处理器核心分配给某个应用时,可能只能获得少量的增益,如果分配给其他应用,可能会获得较大的增益。

(2)正如其名称所示,UB的性能最好,是性能上界的值,AB的性能达到了UB的90%以上,说明AB在兼顾公平性的同时,在性能方面已经挖掘了绝大部分的潜力。

(3)G的性能随着负载密度的增加而减少,G将处理器核心分配给加速比可能最大的用户,这种分配很可能是错误的,因为系统进展的定义中以授权作为权值,但G完全没有考虑授权。当负载密度增大,更多的用户共享同一服务器,处理器资源变得越来越稀缺,这时每个处理器核心都比较重要。

(4)AB与BR的性能相当,但AB对显式方程进行求解,计算开销与用户数量、负载数量、服务器数量无关,BR利用内部点法或登山法进行求解,计算开销与用户数量、负载数量、服务器数量有关,所以AB在计算开销方面具有明显优势。

图2中,预算正比于等级,例如等级4用户的预算是等级2用户的预算的2倍,等级5用户的预算是等级1用户的预算的5倍。

从图2可以看出:

(1)UB偏向等级较高的用户,G没有偏向等级较高的用户。

(2)AB在不同等级的用户中获得的性能是类似的(BR也是如此),从博弈论意义上实现了公平。

(3)AB和BR在性能上超过PS。将预算分解为对处理器核的投标,等价于用运行并行度较低任务的服务器上的处理器核心,来交换并行度较

高任务的服务器上的处理器核心，这样的交换使得AB和BR在性能上超过PS。

4 启发意义

4.1 显式方程具有机器学习不具有的优势

封闭形式的方程（Close-form equation）是显式的计算表达式，是传统科学理论的基本形式，比如牛顿三大运动定律、麦克斯韦电磁方程，它们具有可解释、可证明、计算开销小、确定性强的特点。相比之下，机器学习那样的黑盒子方法不可解释、不可证明、计算开销大、不确定性强。

效用函数是一个封闭形式的方程，具有关键作用，能够反映软件、硬件之间的交互，即从硬件的资源到软件的效用的转化关系，从这个意义上来说，效用函数是软件定义的效用。

4.2 时间具有基本量意义上的重要性

时间（Time）是贯穿计算基础理论与计算机设计的最重要的基本量之一。性能、用户体验、服务质量、功耗、温度、可靠性等都与时间密切相关，具体来说，应用的执行时间的倒数即为应用的性能；用户的尾延迟是否低于用户能容忍的阈值，将决定用户体验；满足的体验要求的用户数量越多，服务质量越高；功耗管控电路与应用的执行（忙碌与空闲）方式有密切关系，温度也是如此；可靠性一般用平均故障时间（MTBF, Mean Time Between Failure）来度量，显然与时间密切相关。

一个任务能否计算、能否快速地计算，分别是可计算性理论和高性能计算两大领域的核心问题。其中，一个任务能否计算，不关心在多久时间范围内完成计算，这是功能的考虑；一个任务能否快速地计算，则关心在多久时间范围内完成计算，这是性能的考虑。从概念上讲，性能是具有较高时间要求的功能，所以可以认为是功能的特例。

时间具有一维单向性，即历史长河中的时间是一去不复返的，所以时间的节省要依靠复用。具体来说，在任务量一定时，时间的节省有两种方式，一是时间复用，在同一时间段内，多个任务同时运行，本质是挖掘开发利用并发性（Concurrency）；一种是空间复用，高速缓存中的数据被多次复用，即通过一次长路径长延迟的数据移动，带来多次短路径短延迟的数据访问，本质是开发局部性（Locality）。从概念上讲，在第二种方式中，可以认为后续发生的多次短路径短延迟的数据访问，已经虚拟地与第一次长路径长延迟的数据移动同时开始了，所以第二种方式可以认为是第一种方式的特例。

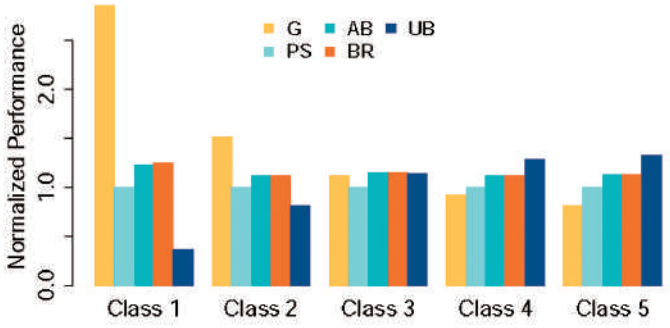

图2　每一等级用户的平均性能（等级越高，预算越多）

4.3 性能模型可以帮助改进linpack和戈登·贝尔奖应用

述评的论文使用了1990年Karp和Flatt发表的工作[10]，由于国内对这一工作的了解还相对较少，我们在这里介绍一下。由Amdahl定律，应用的串行部分对系统效率具有制约作用，这是众所周知的，但是Karp和Flatt给出了更清晰的表达。整理式（1），可得式（5），

$$\frac{p}{s_p}=f\times(p-1)+1 \tag{5}$$

设 e 为系统效率，则得式（6），

$$\frac{1}{e}=f\times(p-1)+1 \tag{6}$$

假设 f 与 p 无关，两边对 p 进行求导，得

$$\frac{\mathrm{d}}{\mathrm{d}p}\frac{1}{e}=f \tag{7}$$

从式（7）可以清晰地看到，随着处理器数量的增加，效率的倒数以 f 的速度上升，所以效率在下降。可以推广式（7）到Gustafson和Sun-Ni定律的情形。

从文献[10]，可以看到Karp-Flatt公式可以帮助改善超级计算的基准应用Linpack和获得“戈登·贝尔”奖的应用的性能，可见其不仅具有重要理

论价值，也具有强大现实影响力。

4.4 市场理论和博弈论可以用于数据中心的资源分配

将市场经济理论应用于云计算数据中心的资源分配与管理中，是一个研究热点，在最近几年的顶级会议上有多篇论文与此相关，且有几篇为最佳论文。

市场理论、博弈论等具有社会科学背景的理论，应该而且可以应用于计算机科学，这再次说明了社会科学与自然科学之间，应该而且可以交叉融合。

追求的目标是性能还是公平性，还是兼有两者，是价值观的反映。什么是公平，值得研究。平均划分，不一定是公平；性能损失一样，也不一定是公平，HPCA 2018的这篇最佳论文给出了一个实用的思路。

5 结束语

将传统经济学中的市场均衡理论，用于计算机科学中云计算数据中心资源分配，是学科融合意义上的一大创新。Amdahl定律是并行计算的三大基本规律之一，有悠久的历史。将由Amdahl定律推导出的Karp-Flatt公式和市场均衡理论，一起应用到云计算数据中心的资源分配，是比较新颖且成功的尝试，新旧结合，推陈出新，是历史思维、计算思维、数据思维、结构思维[7]的综合体现。我国的学术界，可以学习借鉴包括这篇文章在内的国际一流成果，做出具有持久历史影响力的基础研究成果，或者做出具有千万级用户的具有强大现实影响力的自主芯片[5]，并使两者相互促进、相互牵引、相互制导。

●参考文献

[1] Christina D, Christos K. Amdahl's law for tail latency[J]. Communications of the ACM, 2018,61(8):65-72.

[2] Zahedi S M, Lee B C. Resource Elasticity Fairness with Sharing Incentives for Multiprocessors[C]// International Conference on Architectural Support for Programming Languages & Operating Systems. ACM,2014.

[3] Zahedi S M, Llull Q, Lee B C. Amdahl's Law in the Datacenter Era: A Market for Fair Processor Allocation[C]// 2018 IEEE International Symposium on High Performance Computer Architecture (HPCA). IEEE,2018.

[4] 刘宇航. 冯·诺伊曼《计算机与人脑》要点归纳与启发. 中国计算机学会通讯(CCCF) ,中国计算机学会,2018,4,15 (4): 40-45.

[5] 刘宇航. 后中兴事件时代的科研模式转型. 中国计算机学会通讯(CCCF) ,中国计算机学会,2018,7,15(7):1-4.

[6] Luis C, Mark D H, Tomas F W, 鄢贵海, 王颖, 刘宇航译. 计算机体系结构2030：未来15年的研究愿景. 中国计算机学会通讯(CCCF), 中国计算机学会,2017,7,15 (7): 46-51.

[7] 刘宇航. 未来十年计算机科学研究需要的四种思维. 中国计算机学会通讯(CCCF), 中国计算机学会,2016,6,15 (6): 32-37.

[8] Bao Y G, Wang S. Labeled von Neumann Architecture for Software-Defined Cloud[J]. Journal of Computer Science and Technology, 2017, 32(2):219-223.

[9] Liu Y H, Sun X H. C^2-bound: a capacity and concurrency driven analytical model for many-core design[J]. 2015.

[10] Karp A H, Flatt H P. Measuring parallel processor performance[J]. Cacm, 1990, 33(5): 539-543.

[11] Zhi W X, Chun D L. Low-entropy Cloud Computing Systems[J]. SCIENTIA SINICA Informationis, 2017.47(9):1149-1163.