Journal of Computer Applications
计算机应用,2015,35(8):2153 -2157,2163
ISSN 1001-9081
CODEN JYIIDU
2015-08-10
http://www.joca.cn

文章编号:1001-9081(2015)08-2153-05
doi:10.11772/j.issn.1001-9081.2015.08.2153

计算集群中一种基于任务运行时间的组合预测方案

余　莹[1*],李肯立[2],徐雨明[2]
(1.衡阳师范学院 计算机科学系,湖南 衡阳 421002;　2.湖南大学 信息科学与工程学院,长沙 410082)
(*通信作者电子邮箱 yuying_kszx@126.com)

摘　要:针对现有单一预测策略不适用于所有异构任务的问题,提出一种基于本地任务与远程任务运行时间的组合预测方案(CPS)和预测精度保证(PAA)的概念。使用 GridSim 工具集来实现 CPS,将 PAA 作为定量评价由某一特定预测策略提供的预测运行时间精度的标准。仿真实验表明:与本地任务预测策略如 Last 和滑动窗口中值(SM)相比,CPS 的平均相对残差下降了 1.58%、1.62%;与远程任务预测策略如平均运行时间(RM)和加权移动平均值(ES)相比,CPS 的平均相对残差下降了 1.02%、2.9%。因此,PAA 能从综合策略所提供的结果中选择接近最优值的预测,CPS 增强了计算环境中本地任务和远程任务运行时间的 PAA。

关键词:计算集群;组合预测方案;预测精度保证;任务;运行时间
中图分类号:TP393.027;TP18　　**文献标志码**:A

Combined prediction scheme for runtime of tasks in computing cluster

YU Ying[1*], LI Kenli[2], XU Yuming[2]
(1. *Department of Computer Science, Hengyang Normal University, Hengyang Hunan* 421002, *China*;
2. *College of Computer Science and Electronic Engineering, Hunan University, Changsha Hunan* 410082, *China*)

Abstract: A Combined Prediction Scheme (CPS) and a concept of Prediction Accuracy Assurance (PAA) were put forward for the runtime of local and remote tasks, on the issue of inapplicability of the singleness policy to all the heterogeneous tasks. The toolkit of GridSim was used to implement the CPS, and PAA was a quantitative evaluation standard of the prediction runtime provided by a specific strategy. The simulation experiments showed that, compared with the local task prediction strategy such as Last and Sliding Median (SM), the average relative residual error of CPS respectively reduced by 1.58% and 1.62%; and compared with the remote task prediction strategy such as Running Mean (RM) and Exponential Smoothing (ES), the average relative residual error of CPS respectively reduced by 1.02% and 2.9%. The results indicate that PAA can select the near-optimal value from the results of comprehensive prediction strategy, and CPS enhances the PAA of the runtime of local and remote tasks in the computing environments.

Key words: computing cluster; Combined Prediction Scheme (CPS); Prediction Accuracy Assurance (PAA); task; runtime

0　引言

集中式和分布式大型高性能计算系统为越来越多的研究和生产应用提供服务。它们能够获得巨大的计算能力,这种计算能力可用于执行计算密集型的应用程序。集群计算环境协调属于不同组织和个人的分布式资源,允许聚集和共享异构资源,具有规模效益、成本效益和易于扩展的经济优势。然而,多站点异构资源结构和动态异构工作负载的挑战,限制了系统资源的有效实用性。

根据集群计算环境中的调度模型,用户需要提供所需计算资源的具体要求,包括任务提交后的实际运行时间。然而,不同规格任务通过不同平台执行,它的运行时间有很大差别。任务运行时间的高可变性使用户使用系统变得困难,因此,有一些预测策略已用来解决这个问题。预测准确,不仅是提前预留时协助有效调度和未来规划资源分配的需要,同时对提高资源用户和资源所有者的满意度也是非常有用的。从用户的角度来看,任务被合理分派给一个成本越低的资源,而所得到的服务也是越满意的。然而,从资源所有者来看,任务派遣的方式应该是所有资源满负荷工作以保持低成本,在没有突破预设的服务水平协议(Service Level Agreement, SLA)情况下获得最大收益[1]。毫无疑问,这些问题都需要准确预测运行时间。

然而,现有预测战略的特点决定了单一的策略无法适应各种异构任务。在集群计算环境中,有两种类型的任务:本地任务和远程任务。本地任务比远程任务具有更高的优先级,所以是一种抢占计算资源现象[2],因此,本文的研究工作是设计一个组合预测方案(Combined Prediction Scheme, CPS),通过综合现有预测策略优势分别来预测本地任务和远程任务的运行时间。同时还介绍了一种名为预测精度保证(Prediction Accuracy Assurance, PAA)的评价标准来评估通过

收稿日期:2015-02-10;**修回日期**:2015-04-05。　**基金项目**:国家自然科学基金资助项目(61370095,61370098,61070057,90715029);湖南省教育厅科学研究项目(13C074);衡阳市科技发展计划项目(2011KJ22);湖南省教育科学"十二五"规划课题(XJK014CGD006)。
作者简介:余莹(1982 -),女,湖南益阳人,讲师,硕士,主要研究方向:任务调度、并行计算;　李肯立(1971 -),男,湖南娄底人,教授,博士生导师,博士,主要研究方向:并行处理、网格计算、DNA 计算、实时与混成嵌入式系统;　徐雨明(1966 -),男,浙江衢州人,副教授,博士研究生,主要研究方向:任务调度。

综合策略得到的预测结果的精度,采用提供了最佳预测精度保证的预测结果。

1 相关工作

现在已尝试采用多种策略去建模和预测所提交任务的运行时间。文献[3]提出了以静态分析、分析基准和编译器方面为基础的一套方法和体系结构,其中程序是以段为单位进行分析的,这些段的执行时间组合在一起就是程序的总运行时间。文献[4]建议的预测策略采用 K-最近邻算法进行预测,统计模型和时间序列都是以历史数据为基础的,不需要任何内部设计和算法的知识,但不能缺少以前观测的历史数据。文献[5]指出,任务运行时间很可能依赖于某些特定的资源配置类型,并提出了一种预测运行时间的建模方法,以应用资源的使用行为为基础,不使用侵入性技术,如代码检查或测量。该模型是跨平台的,预测不需要在目标计算平台上第一时间被描述。神经网络方法也被用来预测运行时间,如文献[6]提出一种名为混合贝叶斯神经网络的方法,构建了一个贝叶斯网络来代表不同因素影响运行时间下的性能概率分布,神经网络则利用这些概率分布提供一个计算有效且精确的预测。文献[7]中描述了一种不同的策略,提出了一种使用"相似模板"的系统,任务的特点是考虑它们在不同网格基础设施水平的属性,然后通过确认被记录和提交任务有类似属性的最适合的模板来推导预测结果。上述方法对具体任务使用单一的预测策略。文献[8]提出了一种新颖的非线性时间序列预测模型,这种预测方法不仅适合小数据集的时间序列预测,而且对大数据集具有极高的计算效率。文献[9]提出了一种网格环境下任务的执行时间预测的新方法,该方法不需要参考历史数据,可以让用户在提交网格作业之前进行任务执行时间的精确预测。而文献[10]提出了一种预测运行时间的多策略协同模型,应用多种方法去建立预测,其基本思想类似于组合预测方案,但它没有将处理的在线任务细分到本地任务和远程任务。由此可见,以上介绍的现有的预测策略大多数是单一的或特定服务的。

本文中组合预测方案采用了多种方法来预测运行时间,将综合预测策略分成针对本地任务和远程任务的两个虚拟类,只有性能适用于预测具体任务运行时间的策略才会被执行,并引入了评价标准预测精度保证来确定最终所采取的预测策略。

2 组合预测方案

2.1 预测精度保证 PAA 的计算方法

假设计算环境由 N 个计算节点构成,所有的计算节点可以作为一个集合 $CN = \{CN_1, CN_2, \cdots, CN_N\}$,此外,使用 M/M/C 排队系统来表示负载压力和服务模型[11]。服务模型中的第 i 个计算节点可以表示为一个三元组$\langle \lambda_i, \mu_i, c_i \rangle$,其中:$\lambda_i$ 为单位时间平均到达节点 CN_i 的任务数,这是一个服从参数 λ_i 的泊松过程;μ_i 是单位时间能被服务完成的任务数,这是一个服从参数 μ_i 的指数分布。这些参数的值可以从历史数据得到,并动态更新。此外,c_i 是在节点 CN_i 上的 CPU 资源总数。节点 CN_i 的利用率[11] 可以表示为 $\rho_i = (\lambda_i \cdot c_i)^{-1} \cdot \mu_i$,$\rho_i < 1$。

本文认为,本地任务和远程任务的预测运行时间 t_p 均需满足以下两个条件才是合理的:1)t_p 时间段内,被分配的计算节点应该为任务提供必要的资源;2)t_p 应该与运行时间的期望值有最小误差。前者确保预测是有意义的,因为是按预测来进行调度和计划资源分配,若计算节点不能保证必要的资源,会导致任务的性能显著降低,使得预测没有意义;后者保证选择最精确的预测[12]。这两个约束条件是相互独立的。

由于 M/M/C 随机服务系统是用来描述负载压力和服务模型,预测精度保证由预测策略通过解释这两种约束得到下面的线性表达:

$$PAA = p\{\psi \leqslant \lfloor c_i \cdot \mu_i^{-1} \cdot t_p \rfloor - 1\} \cdot p\{|t_p - \gamma| \to 0\} \quad (1)$$

其中:$p\{\psi \leqslant \lfloor c_i \cdot \mu_i^{-1} \cdot t_p \rfloor - 1\}$ 表示在 t_p 时间段内,被分配的计算节点可以为任务提供必要资源的概率;$p\{|t_p - \gamma| \to 0\}$ 表示预测值与期望运行时间值的最小误差概率。

推导程序和引入符号说明如下:

假设变量 X 代表队列中正在等待执行的任务数。在队列 CN_i 中等待任务数的概率等于 $\bar{\omega}$,根据 M/M/C 随机服务模型描述,$\bar{\omega}$ 可以为式(2):

$$p\{X = \bar{\omega}\} = \begin{cases} \delta \cdot \dfrac{\rho_i^{\bar{\omega}+c_i} \cdot {c_i}^{c_i}}{c_i!}, & \bar{\omega} > 0 \\ \displaystyle\sum_{n=0}^{c_i} \delta \cdot \frac{(\rho_i \cdot c_i)^n}{n!}, & \bar{\omega} = 0 \end{cases} \quad (2)$$

其中:

$$\delta = \left[\sum_{n=1}^{c_i} \frac{(\rho_i \cdot c_i)^n}{n!} + \frac{(\rho_i \cdot c_i)^{c_i}}{c_i} \cdot \frac{1}{1-\rho_i} \right]^{-1}$$

假设变量 Y 表示在 t_p 时间内到达队列 CN_i 的任务数。根据排队论,在一个时间单元内到达的任务数要服从参数 λ_i 的泊松分布,因此在 t_p 时间内到达的任务数的概率等于 k,k 可以描述为式(3):

$$p\{Y = k\} = \frac{{\lambda_i}^{-t_p \cdot k}}{(t_p \cdot k)!} \cdot e^{-1/\lambda_i} \quad (3)$$

t_p 时间内位于队列 CN_i 中的任务总数包括已在队列中等待的任务数和将到达的任务数。因为变量 X 和 Y 是相互独立的,可以得出一个结论,用式(4)来描述:

$$p\{X = \bar{\omega}, Y = k\} = \begin{cases} \delta \cdot \dfrac{\rho_i^{\bar{\omega}+c_i} \cdot {c_i}^{c_i}}{c_i!} \cdot \dfrac{{\lambda_i}^{-t_p \cdot k}}{(t_p \cdot k)!} \cdot e^{-1/\lambda_i}, & \bar{\omega} > 0 \\ \displaystyle\sum_{n=0}^{c_i} \delta \cdot \frac{(\rho_i \cdot c_i)^n}{n!} \cdot \frac{{\lambda_i}^{-t_p \cdot k}}{(t_p \cdot k)!} \cdot e^{-1/\lambda_i}, & \bar{\omega} = 0 \end{cases} \quad (4)$$

根据 M/M/C 随机服务系统的特点还可以得到如下信息:$c_i \cdot \mu_i^{-1}$ 代表在一个单位时间可以完成的平均任务数,在时间 t_p 内可以完成的平均任务数可以表示为 $\mu_i^{-1} c_i \cdot t_p$。考虑到本地任务具有抢占优先权,远程任务运行过程中会因为本地任务的到达和执行发生中断,被分配的计算节点应该在时间 t_p 内为任务提供必要的资源,确保预测结果有意义。所以在时间 t_p 内节点 CN_i 的任务总数表示为 ψ,ψ 值应小于等于 $\lfloor c_i \cdot \mu_i^{-1} \cdot t_p \rfloor - 1$,因此得出一个结论,在时间 t_p 内被分配的计算节点可以为任务提供必要资源的概率,等于在时间 t_p 内总任务数位于节点 CN_i 上的概率,总任务数应小于或等于 $\lfloor c_i \cdot \mu_i^{-1} \cdot t_p \rfloor - 1$,见式(5):

$$p\{\psi \leqslant \lfloor c_i \cdot \mu_i^{-1} \cdot t_p \rfloor - 1\} = \sum_{(\bar{\omega}+k)=0}^{\lfloor c_i \cdot \mu_i^{-1} \cdot t_p - 1 \rfloor} p\{X = \bar{\omega}, Y = k\} = \sum_{(\bar{\omega}+k)=0}^{\lfloor c_i \cdot \mu_i^{-1} \cdot t_p - 1 \rfloor} (p\{\bar{\omega} = 0, k\} + p\{\bar{\omega} > 0, k\}) =$$

$$\sum_{(\bar{\omega}+k)=0}^{\lfloor c_i\cdot\mu_i^{-1}\cdot t_p-1\rfloor}\sum_{n=0}^{c_i}\delta\cdot\frac{(\rho_i\cdot c_i)^n}{n!}\cdot\frac{\lambda_i^{-t_p\cdot k}}{(t_p\cdot k)!}\cdot e^{-1/\lambda_i}+\sum_{(\bar{\omega}+k)=1}^{\lfloor c_i\cdot\mu_i^{-1}\cdot t_p-1\rfloor}\delta\cdot\frac{\rho_i^{\bar{\omega}+c_i}\cdot c_i^{c_i}}{c_i!}\cdot\frac{\lambda_i^{-t_p\cdot k}}{(t_p\cdot k)!}\cdot e^{-1/\lambda_i}\tag{5}$$

任务 $T_i(1\leqslant i\leqslant N)$ 的运行时间服从参数 $1/\gamma$ 的指数分布。期望运行时间为 γ,方差是 γ^2,可以从统计的历史数据获得,并动态更新。在 $|t_p-\gamma|$ 非常小的情况下,预测运行时间 t_p 被采用,因此假设 $H_0:t=t_p$,即预测运行时间 t_p 和实际运行时间 t 没有差别。接下来需要统计测试来得知该假设是被接受还是拒绝。根据中心极限定理[12],可以得到一个式(6),其中 n 表示统计数据的大小。从测试看来,预测值和期望值之间的差别可通过变化数据来测量。

$$U=\frac{t_p-\gamma}{\gamma/\sqrt{n}}\sim N(0,1)\tag{6}$$

如果统计结果 U 足够小,则可以接受假设 H_0;否则拒绝。应该确定一个概率 $\alpha(0<\alpha<1)$,表示实现的可信度,如式(7),问题被简化为获得适当的 α 值。一般来说,需要预先设置一个 α 值,然后通过查找标准正态分布表[12]获得 $u_{1-\alpha/2}$ 的值。通过比较标准正态分布表中 U 的绝对值来得到适当的 α 值。如果查找到的是对应适当可信度 α 的下界值,且大于 U 的绝对值,就可以接受假设且 H_0 的接受概率 α 在一个适当水平。α 的值与 U 的绝对值成反比,可以得出一个结论,预测运行时间与期望运行时间的最小偏差等于可信度,可表示为 $p\{t_p-\gamma\to 0\}=\alpha$。

$$\left|\frac{t_p-\gamma}{\gamma/\sqrt{n}}\right|<\lfloor u_{1-\alpha/2}\rfloor\tag{7}$$

结合式(5)和(7),可以推断出预测精度保证的值等于被分配的计算节点 CN_i 在时间 t_p 内可为任务提供必要资源的概率的乘积,且预测运行时间与期望运行时间有一个最小偏差,如式(8):

$$PAA=p\{\psi\leqslant\lfloor c_i\cdot\mu_i^{-1}\cdot t_p\rfloor-1\}\cdot p\{t_p-\gamma\to 0\}=\left(\sum_{(\bar{\omega}+k)=0}^{\lfloor c_i\cdot\mu_i^{-1}\cdot t_p-1\rfloor}\sum_{n=0}^{c_i}\delta\cdot\frac{(\rho_i\cdot c_i)^n}{n!}\cdot\frac{\lambda_i^{-t_p\cdot k}}{(t_p\cdot k)!}\cdot e^{-1/\lambda_i}+\sum_{(\bar{\omega}+k)=1}^{\lfloor c_i\cdot\mu_i^{-1}\cdot t_p-1\rfloor}\delta\cdot\frac{\rho_i^{\bar{\omega}+c_i}\cdot c_i^{c_i}}{c_i!}\cdot\frac{\lambda_i^{-t_p\cdot k}}{(t_p\cdot k)!}\cdot e^{-1/\lambda_i}\right)\cdot\alpha\tag{8}$$

2.2　CPS 模型的体系结构

CPS 模型的体系结构如图 1 所示。它起到一个联合装置的作用,服务可以作为一个门户提供给调度程序,或者也可以设计为一个中间件。主扩展模块的描述说明如下:

1)预测模块。由于模型的可扩展性,各种运行时间预测策略可以集成和实现在预测模块中。当调度决策或资源分配计划需要任务运行时间的有关信息时,每个可用的策略可以根据需求提供一个预测结果。提交任务的信息(如提交任务和计算资源之间的映射关系)对预测模块是必不可少的,它由调度程序提供。另外,计算开销和效率也很重要,因此根据任务的种类(本地任务和远程任务)将综合预测策略分成两个虚拟类别。提交任务时只启用对应类别的策略,关闭其他策略以减少计算开销。

2)决策管理器。使用计算预测精度保证的方法来计算每个预测结果的 PAA 值,采用提供最优 PAA 预测结果的策略作为模型的输出。

3)统计数据库。使用以前的观测和统计建模来对这些预测策略提供历史数据,新执行结果会添加到数据库更新统计数据。统计数据库包含以前获得的历史数据的详细说明,反映了对每种类型任务的计算环境的性能。使用这种技术产生的输出是特定于每一个计算环境的体系结构,如果执行跨平台运行时间的预测,则对每种类型的平台都需要重复。CPS 使用这种方法构建包含每类任务历史运行时间的统计数据库。然而,如果运行时间预测需要在一个具有不同体系结构的计算平台上进行,需要收集历史数据到该体系结构中,并且用一些适当的数据来刷新数据库。也就是说,该预测模型是通过应用程序的需求驱动的,简单描述计算平台来收集历史数据。这也意味着该预测模型的可移植性受限于体系结构,其历史数据在统计数据库中是可用的。

4)控制开关。为实验分析和比较,在模型中设置一组控制开关在操作过程打开或关闭预测策略。当所有的控制开关都关闭时,表示调度程序不需要任何运行时间的信息。

由于是可扩展的设计架构,该模型能综合其中现有的预测策略,利用其他研究人员提出的某些特定于任务的策略的优势。因为本文的重点是设计一个组合预测模型,当前模型针对本地任务和远程任务分成两个虚拟类别,每个类别集成了两个典型简单的预测策略,应用于任务运行时间的时间序列。

本地任务预测策略包括:①Last,它提供的预测运行时间为最后一个观测任务运行时间;②滑动窗口中值(Sliding Median, SM),它提供的预测运行时间为观测任务运行时间的滑动窗口的中值,本文取窗口大小为 5[13]。

远程任务预测策略包括:①平均运行时间(Running Mean, RM),它提供的预测运行时间为所有观测任务的平均运行时间;②加权移动平均值(Exponential Smoothing, ES),它提供的预测运行时间为观测任务运行时间的加权移动平均值。加权移动平均法对最新数据较敏感,而计算环境展示了许多典型时段变化的工作负载,到达数量在很短的时间内变化非常明显,经常有突发任务到达,所以引入一个参数 $\omega(0<\omega\leqslant 1)$ 用来控制敏感性的平滑度。第 i 个观测结果的给定权值为 $\omega_i=m_i/l$,其中:l 代表层数;$m=1,2,\cdots,l$。假定 m_i 代表第 i 层的观测结果。详细的信息可见文献[14]。

集成的两类预测策略对所提交任务用它们各自的规则提供预测结果。CPS 将采用提供最优 *PAA* 的预测结果,并传送给调度程序。

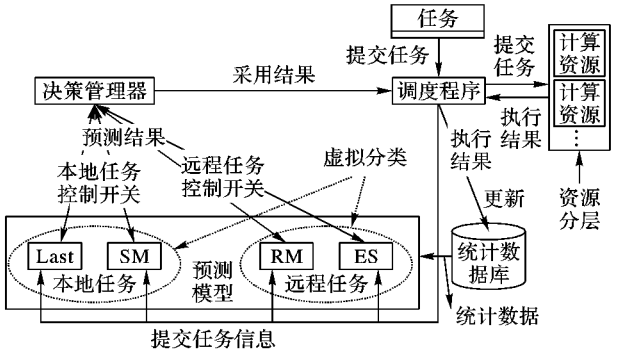

图 1　CPS 的体系结构

2.3　组合预测算法

在组合预测算法中把复制策略[15]考虑进来,将同一任务的几个副本在不同的计算资源上执行,而由于资源的异构性,

这些分配的资源具有不同的 $p\{\psi \leqslant \lfloor c_i \cdot \mu_i^{-1} \cdot t_p \rfloor - 1\}$ 值。一旦任务的其中一个副本完成执行，所有其他副本都被取消。一次任务调度只是资源请求到计算资源的一个映射，可以表示成一个向量 $\boldsymbol{V} = T_i \times CN \rightarrow \{0,1\}$。元素 $V_{i,j} = 1$ 表示第 j 个计算节点分配给了第 i 个任务。如果计算环境中有多重任务，全部现有任务的映射可以表示为一个由 $\boldsymbol{V}$ 组成的矩阵 $\boldsymbol{M}$。为了比较，对同一任务使用相同的调度策略却使用不同的运行时间预测策略。在一个任务存在多个副本的条件下，主要使用循环分配法(Round Robin Policy, RR_P) 的典型调度策略思想来分派每个副本。

组合预测算法中的符号解释如下：t_p^k 表示预测策略 S_k 的预测运行时间；$p\{\psi \leqslant \lfloor c_i \cdot \mu_i^{-1} \cdot t_p^k \rfloor - 1\}$ 表示被分配的第 j 个计算节点在时间段 t_p^k 内可以为任务提供必要资源的概率；$p_{\max}\{\psi \leqslant \lfloor c_i \cdot \mu_i^{-1} \cdot t_p \rfloor - 1\}$ 表示使用复制策略后，被分配的计算节点在时间段 t_p^k 内可以为任务提供必要资源的最大概率；$PAA_{\max}$ 表示预测精度保证的最大值。

组合预测算法 CPS 描述如下：

输入　资源分配映射矩阵 $\boldsymbol{M}$(矩阵 $\boldsymbol{M}$ 由 $\boldsymbol{V}$ 组成；元素 $V_{i,j} = 1$ 表示第 j 个计算节点分配给了第 i 个任务)；服务模型中的第 i 个计算节点相关信息 $\langle \lambda_i, \mu_i, c_i \rangle$；统计数据来源数据库。

输出　具有最优 PAA 的预测运行时间，本地任务 $t_{\text{optimal-l}}$，远程任务 $t_{\text{optimal-r}}$。

```
开始
  对每个提交的本地任务 T_{i-l}
    PAA_{max-l} = 0, p_{max-l}{ψ ≤ ⌊c_i · μ_i^{-1} · t_{p-l}⌋ - 1} = 0;
  对每个本地任务预测策略 S_{k-l}(Last,SM)
    得到预测运行时间 t_{p-l}^k;
    对每个计算节点(j = 1 to N)
      如果 V_{i,j} = 1
        计算 p_{j-l}{ψ ≤ ⌊c_i · μ_i^{-1} · t_{p-l}⌋ - 1};
        如果 p_{max-l}{ψ ≤ ⌊c_i · μ_i^{-1} · t_{p-l}⌋ - 1} < p_{j-l}{ψ ≤
          ⌊c_i · μ_i^{-1} · t_{p-l}⌋ - 1}
          执行 p_{max-l}{ψ ≤ ⌊c_i · μ_i^{-1} · t_{p-l}⌋ - 1} = p_{j-l}{ψ ≤
            ⌊c_i · μ_i^{-1} · t_{p-l}⌋ - 1};
    计算 U_l;
    确定概率 p{|t_{p-l} - γ| → 0};
    计算 PAA_{k-l} = p_{max-l}{ψ ≤ ⌊c_i · μ_i^{-1} · t_{p-l}⌋ - 1} ·
      p{|t_{p-l} - γ| → 0};
    如果 PAA_{max-l} < PAA_{k-l}
      执行 PAA_{max-l} = PAA_{k-l};
      执行 t_{optimal-l} = t_{p-l}^k
  对每个远程任务预测策略 T_{i-r}
    PAA_{max-r} = 0, p_{max-r}{ψ ≤ ⌊c_i · μ_i^{-1} · t_{p-r}⌋ - 1} = 0;
  对每个远程任务预测策略 S_{k-r}(RM,ES)
    得到预测运行时间 t_{p-r}^k;
    其他步骤同上;
    如果 PAA_{max-r} < PAA_{k-r}
      执行 PAA_{max-r} = PAA_{k-r};
      执行 t_{optimal-l} = t_{p-r}^k
结束
```

3 实验和分析

3.1 实验内容

使用 GridSim 工具集[16]来实现组合预测方案，但由于是缺乏充足的工作负载实际描述和详细的计算站点性能参数，很难呈现在实际的操作环境中。通过选择分析可用的痕迹来作比较，以满足客观需求。获得的每个观测任务的痕迹包括以下字段：提交时间、实际运行时间、在用户模式下的运行时间、在内核模式下的运行时间、需要的计算节点数量、提交节点的标识符和最后运行节点的标识符等。

用来测试的 3 个任务都是 CPU 密集型应用程序，并在实验中还运行一些其他程序作比较。这些用作比较的应用程序在测试阶段有着不同的提交时间和提交数量，以模拟突发性的特点。对每个任务，使用复制策略和轮循调度策略的关键思想来发送它和它的副本 (R) 到多个计算资源，以提高执行可靠性；同时，也可以观察在不同副本数量条件下组合预测方案的性能。本文只研究该模型在 $R = i(i = 1,2)$ 条件下的执行情况，i 表示副本的数量。如果一个任务有多个副本，选择第一个完成副本的要素作为当前记录，如实际的运行时间和执行环境的信息。

本文中其他参数量化如下：每个计算节点的 CPU 资源总数设置为 $c_i = 2$，假设到达的任务数量服从参数 $\lambda = 1$ 的泊松分布[2]，完成一个服务需求的平均服务时间服从参数 $\mu_i = 1$ 的指数分布。通过对比只使用一种预测策略的组合预测方案的预测结果来评估最终采用结果的准确性，提出一个度量标准——相对残差(e)，如式(9) 所示，预测结果和实际值之间的偏差可以从该度量标准得出。

$$e = (t_p - t)/t \tag{9}$$

3.2 分析结果

对每个观测任务设置 5 个检测点，它们的间隔是周期性的，通过故意设置任务的行为分化模拟突发性的特点，以探讨对分析结果的影响。工作负载突发性的特点给本研究带来了难度和挑战。

对于这两组观测的本地任务和远程任务，图 2 代表了所提模型 CPS、Last 和 SM 的实验结果，图 3 代表了所提模型 CPS、RM 和 ES 的实验结果。

从图 2、3 中可以看到，所采用的组合预测方案的预测值与使用单一预测策略所得结果中的最优值几乎是一致的。但也有一些错误，如 $R = 1$ 条件下远程任务 1 在检测点 3 的观测结果(见图 3(a))。从实验结果来看，预测策略 RM 提供了最优结果，最接近真正的运行时间。然而 CPS 通过比较综合预测策略提供的每个 PAA 值，选择了 ES 提供的预测结果作为最佳输出。这种偏差是由概率的计算偏差引起的。在 PAA 定义的推导过程中，较大的预测结果可能有较大的 $p\{\psi \leqslant \lfloor c_i \cdot \mu_i^{-1} \cdot t_p \rfloor - 1\}$ 值，而 $p\{|t_p - \gamma| \rightarrow 0\}$ 值与 U 的绝对值成反比，U 的绝对值与预测值成正比。也就是说，较大的预测结果有较小的 $p\{|t_p - \gamma| \rightarrow 0\}$ 值。如果两个预测值非常接近，较大的预测结果可能有较大的 $p\{\psi \leqslant \lfloor c_i \cdot \mu_i^{-1} \cdot t_p \rfloor - 1\}$ 值，而具有较小的 $p\{|t_p - \gamma| \rightarrow 0\}$ 值；较小的预测结果可能有较小的 $p\{\psi \leqslant \lfloor c_i \cdot \mu_i^{-1} \cdot t_p \rfloor - 1\}$ 值，而具有较大的 $p\{|t_p - \gamma| \rightarrow 0\}$ 值。根据 PAA 的定义，最后采用的预测结果将提供更大的 PAA 值。在这些偏差情况下，所采用的预测值与最优结果有着微小的偏差，然而这些偏差并没有导致性能的损失。与本地任务预测策略 Last 和 SM 相比，CPS 的平均相对残差下降了 1.58%、1.62%，与远程任务预测策略 RM 和 ES 相比，CPS

的平均相对残差下降了 1.02%、2.9%,因此,本文提出的评价标准是有效的,与采用单一的预测策略相比,组合预测模型提供了一个增强任务预测运行时间精度保证的方案。

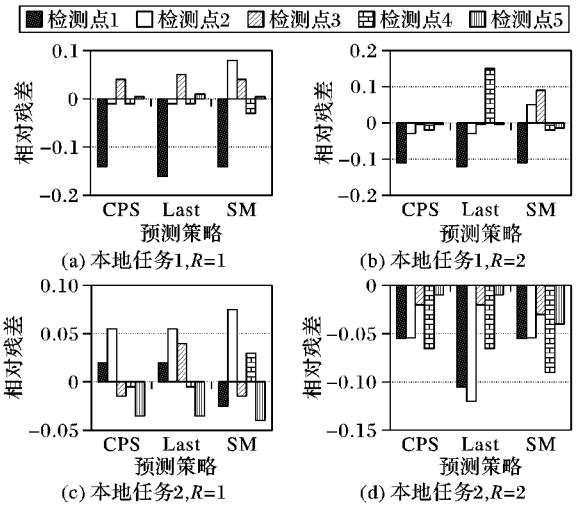

图 2　本地任务运行时间预测结果

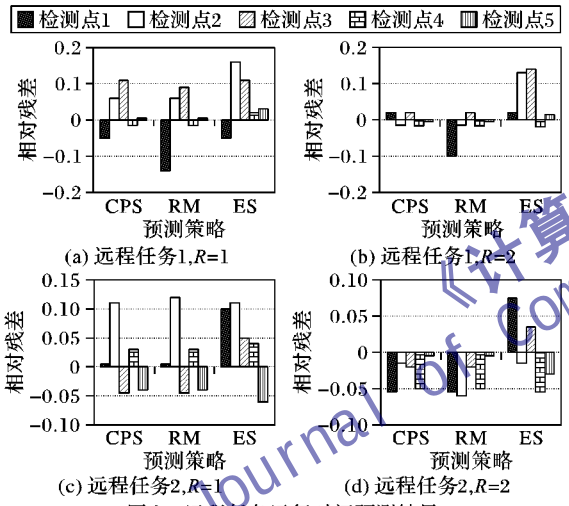

图 3　远程任务运行时间预测结果

通过分析实验结果,每个观测任务有着不同的副本数量,在每个检测点使用不同的预测方案。在有着不同副本数量的条件下,本地任务预测策略 Last、SM 在检测点 1 的性能相对较差,远程任务预测策略 RM 在检测点 1 的性能相对较差,这是因为工作负载的突发性到达。在检测点 2,每个预测方案与实际值有相对较大的偏差,因为在观测点有意想不到的工作负载规模的减少。在检测点 3,SM 的性能相对较差,关键原因是由于不稳定的现象。如果在每个滑动窗口的中值突发,这种现象会更加严重,将会有更大的偏差。当统计数据的规模较小时,这种现象相对更明显,当规模增大时,它趋向平滑,这可以在检测点 5 的观察结果中看到。不过也不是绝对的,这种现象是否严重仍然主要取决于执行环境。

尽管组合预测模型可以从提供的综合预测策略中挑选出好的预测结果,但即使是最好的结果,所采用的预测值与实际值之间仍然存在差异。导致性能不好的原因有很多,主要原因有两个:1)实验中突发提交情况的发生;2)综合预测策略的效率依赖于固定的时间序列。这些策略在时间序列平稳的情况下表现出良好的可预测性[14]。也就是说,它应该有一个恒定的长期均值和方差,而实际情况往往缺乏稳定性。

4　结语

在集群计算环境中,现有预测策略的特点决定了单一预测策略并不适合所有类型的本地任务和远程任务,因此,本文提出了一个更有效的选择策略——组合预测方案(CPS)。CPS 使用现有的运行时间预测策略,分别针对本地任务和远程任务生成组合协同预测方案,具有良好的可扩展性。CPS 也可以被移植到其他的计算基础设施,唯一需要做的是用从新特定平台上收集的历史数据来更新统计数据库。

为了评估预测的精度,提出了预测精度保证(PAA)作为评价标准,系统采用提供最优预测精度保证的预测结果方案。实验结果表明,采用的组合预测方案的预测值与使用单一的预测策略方案所提供的最优预测结果几乎是一致的,尽管有些偏差是由概率的计算偏差所导致,但并没造成性能损失,因此评价标准是有效的。对于所采用的预测值与实际值之间仍存在显著差异,本文也总结了两个主要原因。下一步将其他的一些运行时间预测策略纳入该模型,来提高预测效率并减少计算开销。

参考文献:

[1] LEFF A, RAYFIELD J T, DIAS D M. Service-level agreements and commercial grids [J]. IEEE Internet Computing, 2003, 7(4): 44 - 50.

[2] GONG L, SUN X, WASTON E. Performance modeling and prediction of non-dedicated network computing [J]. IEEE Transactions on Computers, 2002, 51(9): 1041 - 1055.

[3] KIRAN M, HASHIM A H A, KUAN L M, *et al*. Execution time prediction of imperative paradigm tasks for grid scheduling optimization [J]. International Journal of Computer Science and Network Security, 2009, 9(2): 155 - 163.

[4] PHINJAROENPHAN P, BEVINAKOPPA S, ZEEPHONGSEKUL P. A method for estimating the execution time of a parallel task on a grid node [C]// EGC 2005: Proceedings of the 2005 European Grid Conference on Advances in Grid Computing, LNCS 3470. Berlin: Springer, 2005: 226 - 236.

[5] SADJADI S M, SHIMIZU S, FIGUEROA J, *et al*. A modeling approach for estimating execution time of long-running scientific applications [C]// IPDPS 2008: Proceedings of the 22nd IEEE International Symposium on Parallel and Distributed Processing. Piscataway: IEEE, 2008: 1 - 8.

[6] DUAN R, NADEEM F, WANG J, *et al*. A hybrid intelligent method for performance modeling and prediction of workflow activities in grids [C]// Proceedings of the 9th IEEE/ACM International Symposium on Cluster Computing and the Grid. Piscataway: IEEE, 2009: 339 - 347.

[7] NADEEM F, FAHRINGER T. Using templates to predict execution time of scientific workflow applications in the grid [C]// Proceedings of the 9th IEEE/ACM International Symposium on Cluster Computing and the Grid. Piscataway: IEEE, 2009: 316 - 323.

[8] LI A, QIN Z. Moving windows quadratic autoregressive model for predicting nonlinear time series [J]. Chinese Journal of Computers, 2004, 27(7): 1004 - 1008. (李爱国, 覃征. 滑动窗口二次自回归模型预测非线性时间序列[J]. 计算机学报, 2004, 27(7): 1004 - 1008.)

(下转第 2163 页)

博客应用的有效性。将来的工作重点将尝试在实际网络环境中部署 BlogCloud,并在大量的实际应用中对其进一步完善。

参考文献:

[1] HE F. Design and implementation of blog system based on SSH mode [D]. Changsha: Hunan University, 2012: 9 - 18. (何发胜. 基于 SSH 模式架构的博客系统的设计与实现[D]. 长沙: 湖南大学, 2012: 9 - 18.)

[2] LIU Q. Design and implementation of enterprise blog system based on SSH framework [D]. Jinan: Shandong University, 2013: 5 - 7. (柳青. 基于 SSH 框架的企业内博客系统的设计与实现[D]. 济南: 山东大学, 2013: 5 - 7.)

[3] WIKIPEDIA. Cloud storage[EB/OL]. (2014-12-06) [2015-03-18]. http://en.wikipedia.org/wiki/Cloud_storage.

[4] GHEMAWAT S, GOBIOFF H, LEUNG S-T. The Google file system [C]// SIGOPS 2003: Proceedings of the 2003 Operating Systems Review. New York: ACM Special Interest Group on Operating Systems, 2003: 29 - 43.

[5] WU Z. The analysis of the core technology of cloud computing [M]. Beijing: People's Posts and Telecommunications Press, 2011: 7 - 8. (吴朱华. 云计算核心技术剖析[M]. 北京: 人民邮电出版社, 2011: 7 - 8.)

[6] BORTHAKUR D. The Hadoop distributed file system: architecture and design [R/OL]. [S. l.]: The Apache Software Foundation, 2007 [2015-03-22]. http://svn.apache.org/repos/asf/hadoop/common/tags/release-0.13.1/docs/hdfs_design.pdf.

[7] Hbase Development Team. Hbase: bigtable-like structured storage for Hadoop HDFS [EB/OL]. (2011-12-20) [2014-11-15]. http://wiki.apache.org/hadoop/Hbase.

[8] GUAN L. P2P technology secret: the principle and typical system development of P2P network technology [M]. Beijing: Tsinghua University Press, 2011: 41 - 59. (管磊. P2P 技术揭秘——P2P 网络技术原理与典型系统开发[M]. 北京: 清华大学出版社, 2011: 41 - 59.)

[9] XU F, YANG G, JU D. Design of distributed storage system on peer-to-peer structure [J]. Journal of Software, 2004, 15(2): 268 - 277. (徐非, 杨广文, 鞠大鹏. 基于 Peer-to-Peer 的分布式存储系统的设计 [J]. 软件学报, 2004, 15(2): 268 - 277.)

[10] TIAN R, LU X, HOU M, *et al.* P2P-based distributed storage system [J]. Computer Science, 2007, 34(6): 47 - 48. (田荣华, 卢显良, 侯孟书, 等. P2P 分布式存储系统 [J]. 计算机科学, 2007, 34(6): 47 - 48.)

[11] YANG D, XU L, ZHANG J. Blue whale distributed file system metadata service [J]. Computer Engineering, 2008, 34(7): 4 - 6. (杨德志, 许鲁, 张建刚. 蓝鲸分布式文件系统元数据服务 [J]. 计算机工程, 2008, 34(7): 4 - 6.)

[12] ZHANG J, ZHANG J, ZHANG J, *et al.* Isolation technology of metadata from data in blue whale file system [J]. Computer Engineering, 2010, 36(2): 28 - 30. (张敬亮, 张军伟, 张建刚, 等. 蓝鲸文件系统中元数据与数据隔离技术 [J]. 计算机工程, 2010, 36(2): 28 - 30.)

[13] GROSSMAN R, GU Y. Data mining using high performance data clouds: experimental studies using sector and sphere [C]// KDD '08: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2008: 920 - 927.

[14] GU Y, GROSSMAN R. Sector and sphere: the design and implementation of a high-performance data cloud [J]. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2009, 367(1897): 2429 - 2445.

[15] BOWERS K D, JUELS A, OPREA A. HAIL: a high-availability and integrity layer for cloud storage [C]// CCS '09: Proceedings of the 16th ACM Conference on Computer and Communications Security. New York: ACM, 2009: 187 - 198.

[16] MA W, WU H, LIU P. MassCloud cloud storage system architecture and reliability mechanism [J]. Journal of Hohai University: Natural Sciences, 2011, 39(3): 348 - 352. (马玮骏, 吴海佳, 刘鹏. MassCloud 云存储系统构架及可靠性机制 [J]. 河海大学学报: 自然科学版, 2011, 39(3): 348 - 352.)

[17] SUN J, YU H, ZHENG W. Index of meta-data set of the similar files for inline de-duplication in distributed storage systems [J]. Journal of Computer Research and Development, 2013, 50(1): 197 - 205. (孙竞, 余宏亮, 郑纬民. 支持分布式存储删冗的相似文件元数据集合索引 [J]. 计算机研究与发展, 2013, 50(1): 197 - 205.)

[18] YANG W. Sina micro-blog technical architecture analysis [EB/OL]. (2010-11-16) [2015-01-10]. http://tech.sina.com.cn/i/2010-11-16/14434871585.shtml. (杨卫华. 新浪微博技术架构分析[EB/OL]. (2010-11-16) [2015-01-10]. http://tech.sina.com.cn/i/2010-11-16/14434871585.shtml.)

[19] AIMEI. Analysis of the 2011 China cloud storage market development [EB/OL]. (2012-03-05) [2014-12-13]. http://www.iimedia.cn/26270.html. (艾媒网. 2011 中国云存储市场发展状况研究分析[EB/OL]. (2012-03-05) [2014-12-13]. http://www.iimedia.cn/26270.html.)

(上接第 2157 页)

[9] JIANG Y. Research of task execution time prediction technology in grid computing environments [J]. Computer Engineering and Design, 2011, 32(10): 3428 - 3430. (蒋炎华. 网格环境下任务的执行时间预测技术研究[J]. 计算机工程与设计, 2011, 32(10): 3428 - 3430.)

[10] TAO M, DONG S, ZHANG L. A multi-strategy collaborative prediction model for the runtime of online tasks in computing cluster/grid [J]. Cluster Computing, 2011, 14(2): 199 - 210.

[11] GROSS D, SHORTLE J F, THOMPSON J M, *et al.* Fundamentals of queuing theory [M]. New York: Wiley, 1998: 68 - 82.

[12] DEKKING F M, KRAAIKAMP C, LOPUHAA H P, *et al.* A modern introduction to probability and statistics: understanding why and how [M]. Berlin: Springer, 2005: 103 - 114.

[13] WOLSKI R. Experiences with predicting resource performance online in computational grid settings [J]. SIGMETRICS Performance Evaluation Review, 2003, 30(4): 41 - 49.

[14] BROCKWELL P J, DAVIS R A. Introduction to time series and forecasting [M]. Berlin: Springer, 2002: 317 - 330.

[15] CIRNE W, BRASILEIRO F, PARANHOS D, *et al.* On the efficacy, efficiency and emergent behavior of task replication in large distributed systems [J]. Parallel Computing, 2007, 33(3): 213 - 234.

[16] BUYYA R, MURSHED M. GridSim: a toolkit for the modeling and simulation of distributed resource management and scheduling for grid computing [J]. Concurrency and Computation: Practice and Experience, 2002, 14(13/14/15): 1175 - 1220.