软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(Suppl.(2)):90−100

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

生物基因测序高性能计算日志的任务分析及建模*

曹志波[1], 董守斌[1], 王丙强[2], 左利云[1]

[1](华南理工大学 计算机科学与工程学院 广东省计算机网络重点实验室,广东 广州 510006)

[2](深圳华大基因研究院,广东 深圳 518083)

通讯作者: 董守斌, E-mail: sbdong@scut.edu.cn

摘 要: 生物基因测序是生物信息学分析中最常用的高性能计算任务.旨在通过分析生物基因测序日志找出生物基因测序日志中的任务特性,构建一种通用的适合分析生物基因测序的任务模型,并应用于面向基因测序的高性能计算系统的任务调度及性能优化.基于任务日志,主要分析了生物基因测序日志中任务到达时间的规律特性、任务运行时间和任务的并行尺寸等特性,通过这些任务特性利用指数分布、伽马分布、正态分布以及线性拟合构建了相应的局部任务模型,然后提出一种局部模型融合的方法,将各个局部模型合并为统一的任务模型.通过两种通用的模型评测方法对任务模型进行的评测结果显示,最终的任务模型与原有任务日志的 4 种任务属性趋于相同的分布,验证了所构建的任务模型具有很好的通用性.

关键词: 高性能计算;任务建模;模型评测;生物基因测序

中文引用格式: 曹志波,董守斌,王丙强,左利云.生物基因测序高性能计算日志的任务分析及建模.软件学报,2014,25(Suppl.(2)):90−100. http://www.jos.org.cn/1000-9825/14027.htm

英文引用格式: Cao ZB, Dong SB, Wang BQ, Zuo LY. Workload analysis and modeling of high performance computing trace of biological gene sequencing. Ruan Jian Xue Bao/Journal of Software, 2014,25(Suppl.(2)):90−100 (in Chinese). http://www.jos.org.cn/1000-9825/14027.htm

Workload Analysis and Modeling of High Performance Computing Trace of Biological Gene Sequencing

CAO Zhi-Bo[1], DONG Shou-Bin[1], WANG Bing-Qiang[2], ZUO Li-Yun[1]

[1](Key Laboratory of Communication and Computer Network, School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China)

[2](Shenzhen Huada Gene Research Institute, Shenzhen 518083, China)

Corresponding author: DONG Shou-Bin, E-mail: sbdong@scut.edu.cn

Abstract: Biological gene sequencing is one of the most common high-performance computing tasks in Bioinformatics analysis. This paper aims to find the main workload characteristics of biological gene sequence trace (BGST) and construct a general model to analyze the biological gene sequence (BGS), which can be used in high-performance computing scheduling and performance optimization with the BGS. The study mainly analyzes the job arrival, runtime and parallelism characteristics in BGST. Based on the analysis, it constructs several local models with exponential, Gamma, Gaussian and linear regression, then combines all the local models into a final model. The experimental results obtained by applying two general evaluation methods show that the new model has uniform distributed trend with BGST, which demonstrates the good versatility of the model.

Key words: high-performance computing; workload modeling; model estimate; biological gene sequence trace

高性能计算系统的性能与其上运行的任务负载有很大的关系[1].因此通过高性能系统上运行的任务负载提取任务负载的特征,进而构建更加通用的任务模型来优化高性能系统的性能也就成为一个重要的研究点.而

* 基金项目: 国家自然科学基金(61070092); 广州市科技计划(2012Y2-00043, 2013Y2-00041)
收稿时间: 2013-08-05; 定稿时间: 2014-03-13

生物基因测序技术作为生命科学研究的一项重要的技术,为人类基因图谱的绘制、各种疾病的诊断、生物病毒传播的预防等提供了很大的帮助,因此,通过分析生物基因测序高性能计算日志的任务负载特性,然后构建通用的生物基因测序的任务模型也就具有重要的研究价值.

虽然生物基因测序是最常用的生物信息学分析高性能计算任务,但目前针对该计算任务的任务模型还没有,而高性能计算系统的任务调度及性能优化极大地依赖于任务模型的建立.因此研究和构建生物基因测序的任务模型具有重要的应用意义.针对现有研究的重点,本文主要分析了生物基因测序日志中任务到达时间的规律特性、任务的运行时间、任务的并行尺寸,然后利用分析的结果构建了相应局部模型,随后提出了一种局部模型融合的方法,并将各个局部模型合并为最终的任务模型.通过两种通用的模型评测方法对任务模型进行了评测.

1 相关工作

在实际的高性能环境的性能评估中,主要有两种可选的方法[1]:(1) 利用由实际环境采集的任务日志直接进行仿真环境驱动的性能评估;(2) 通过对任务日志的分析构建的任务模型进行仿真环境驱动的性能评估.与任务日志相比,任务模型具有很多优势[2]:(1) 任务模型具有任务负载的各个任务特性,因此通过调节各个特性的参数,研究系统性能的变化情况.(2) 任务负载可能被失败的任务污染,而任务模型可以完全避免这种情况的发生.(3) 超强的伸缩特性,例如从128个节点上采集的任务日志构建的任务模型,通过调节参数生成256个节点上的任务负载.(4) 重复特性,任务模型相当于随机变量,而任务日志只是这个随机变量的一个采样点,任务模型可以重复地产生同样集群环境下的不同的任务负载.文献[3]将高性能环境下的任务负载分为两种类型:刚性任务负载和可塑性任务负载.刚性任务负载是指任务在运行过程中并行度固定不变.本文研究的生物基因测序日志属于刚性任务负载范畴,因此以下主要对刚性任务负载的相关工作进行探讨.

Feitelson在文献[4]中,通过分析NASA iPSC/860集群日志[5]中的任务负载特征,发现该集群的任务到达时间间隔不但具有工作日周期性同时还具有节假日周期性.在文献[1]中,Lublin在分析了任务到达的工作日周期性后,采用伽马分布对任务的工作日周期性进行模拟.不同于Lublin的研究,本文在对生物基因测序日志的研究中除了考虑任务到达时间的工作日周期性以外,同时也考虑到任务到达时间的节假日周期性.

在文献[6,7]中,Hui 针对任务的到达速度和任务的运行时间进行了更为复杂的任务模型构建,其中针对任务到达速度的模型可以用于预测实际高性能环境下的任务到达的速度,进而优化资源的调度.在文献[8]中,Feitelson 指出在进行任务建模时需要找出任务负载中重要的任务属性,因为重要的任务属性对系统性能有至关重要的影响,然后 Feitelson 分析了任务负载到达的日周期特性对不同调度策略的影响,最终发现调度器在拥有日周期特性和没有日周期特性的任务负载的评估下性能的差距达到 50%左右,因此任务的日周期特性是任务建模中的一个关键的任务属性.同时,针对任务到达的时间规律存在长范围依赖和突发特性,文献[9]提出了一种 MWM(多面小波模型)来模拟任务到达规律中的长范围依赖特性,但是文献[9]中并没有很好地模拟突发特性,因此文献[10]针对突发特性提出了一种改进的MWM以更好地拟合任务到达时间规律上的突发特性.

2 基于高性能计算日志的任务特性分析

通过分析任务达到时间可以了解集群系统的负载变化情况;而通过分析任务的并行尺寸和任务的运行时间则可以更好地了解集群中运行任务的特性,进而优化任务的调度.目前许多高性能计算系统使用 SGE(Sun grid engine)[11]日志,SGE 日志中记录的任务属性可参考网站[11].针对 SGE 日志所能提供的信息,本文对任务日志的分析主要包括以下两个方面:(1) 任务到达时间的规律性;(2) 任务的并行尺寸同任务运行时间之间的关系.

2.1 任务特性分析方法

(1) 任务到达时间的规律性分析方法

文献[1,4]中指出,高性能计算日志在任务到达时间上存在工作日周期特性和节假日周期特性.因此本文针

对任务到达时间的规律性分析方法如下:首先将工作日的一天分成 24 个时间槽,时间槽 i ($\in[1,24]$)到达的任务数记为 x_i,第 i 个时间槽在第 j 天到达的任务数目记录为 x_{ij}. x_i 的期望值可表示为式(1),其中 m 表示工作日的总天数.同时也将节假日的一天分成 24 个时间槽,其中时间槽 i 到达的任务数记为 y_i,第 i 个时间槽在第 j 天到达的任务数目记录为 y_{ij}.于是, y_i 的期望值可表示为式(2),其中 n 表示节假日的总天数.通过找出式(1)和式(2)中的随机变量 $\overline{x}_i$ 和 $\overline{y}_i$ 在 24 个时间槽内的分布状况即可找出任务日志的工作日周期特性和节假日周期特性.

$$\overline{x}_i = \sum_{j=1}^{m} x_{ij} \tag{1}$$

$$\overline{y}_i = \sum_{j=1}^{n} y_{ij} \tag{2}$$

(2) 任务的并行尺寸同运行时间之间的规律性分析方法

文献[1,4,12,13]中研究指出,高性能计算日志中任务的并行尺寸同任务的运行时间之间存在相关性.为了方便对任务的并行尺寸和任务运行时间之间关系描述,本文用符号 P 表示任务的并行尺寸,而用符号 R 表示任务的运行时间.因此,本文针对这两种任务属性的规律性分析方法为:首先分析日志中 P 的概率分布,寻找合适的分布(指数分布、伽马分布等)进行拟合;然后分析不同取值范围的 P 内对应的所有任务的任务运行时间的概率分布,并寻找合适的分布进行拟合;最后分析任务并行尺寸与任务运行时间之间的关系,并采用合适的分布函数来拟合这两个属性之间的关系.

2.2 生物基因测序计算任务的日志分析

本文使用的任务日志是由深圳华大基因公司[14]提供的生物基因测序日志.该任务日志是在其高性能计算系统上产生的 SGE 任务日志.为了更方便地对该日志进行分析和模型构建,本文将其转化为标准化日志格式(SWF)[3].所生成的 SWF 格式的日志主要包括任务到达时间,任务的并行尺寸以及任务的运行时间主要的特性.

(1) 任务到达时间的规律性分析

依据第 2.1 节(1)中任务到达时间的规律性分析方法容易得出日志中的 $\overline{x}_i$ 和 $\overline{y}_i$ 的分布情况,如图 1 所示.由图 1 可以看出,任务到达时所在的时间槽与任务到达数目的期望值存在一种非线性关系.但是同一个时间槽每天到达的任务数目又不相同,因此需要找出每个时间槽内任务每天到达的任务数目的分布情况(即找出随机变量 x_i 和 y_i 的分布情况),进而构建合适的任务到达时间规律性的模型.

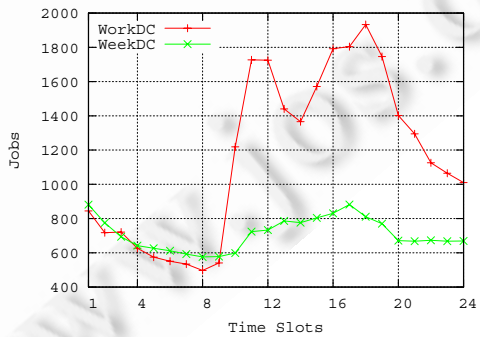


Fig.1 Work daily cycle (WorkDC) and weekend daily cycle (WeekDC) in the trace

图 1 任务日志的工作日周期性(WorkDC)和节假日周期性(WeekDC)

通过记录工作日和节假日中 24 个时间槽每天到达的任务数目,容易计算出每个时间槽内不同任务到达数目的概率密度,如图 2 和图 3 所示.由图 2 可以看出工作日中每个时间槽内每天任务到达数目的分布规律:时间槽 1~9 内趋向于指数分布,而时间槽 10~24 内则趋向于伽马分布.由图 3 可以看出节假日中每个时间槽内任务到达数目的分布规律;时间槽 1~24 内均趋向于指数分布.

通过对任务到达时间的规律性分析,工作日和节假日中的时间槽内到达任务数目的期望值成一种非线性

关系;而在工作日中,时间槽 1~9 内每天到达的任务数目趋向于指数分布,时间槽 10~24 内每天到达的任务数目则趋向于伽马分布,在节假日中,每个时间槽内每天到达的任务均趋向于指数分布.

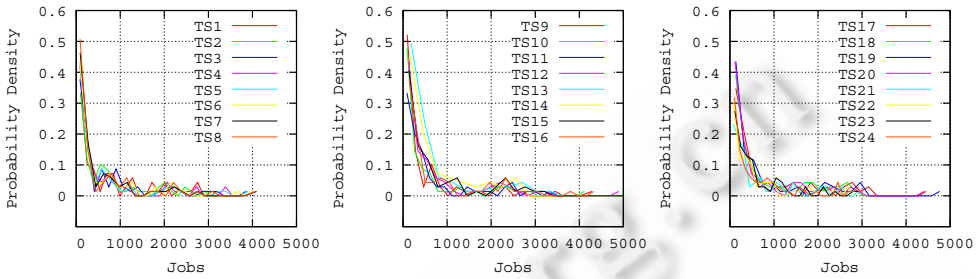


Fig.2 Daily arrived jobs' probability density of the 24 time slots (TS) in workday

图 2 工作日中 24 个时间槽内每天到达任务数目的概率密度分布

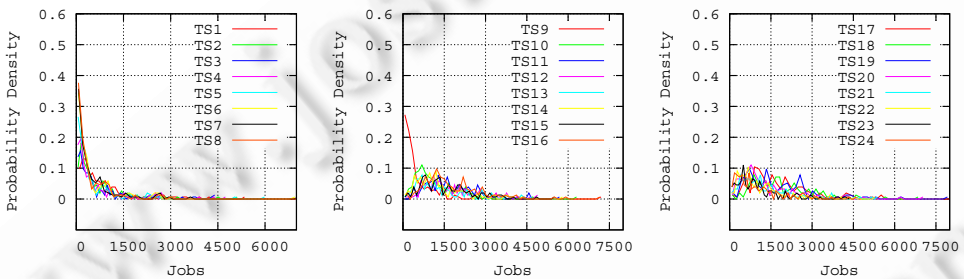


Fig.3 Daily arrived jobs' probability density of the 24 time slots (TS) in Weekend

图 3 节假日中 24 个时间槽内每天到达任务数目的概率密度分布

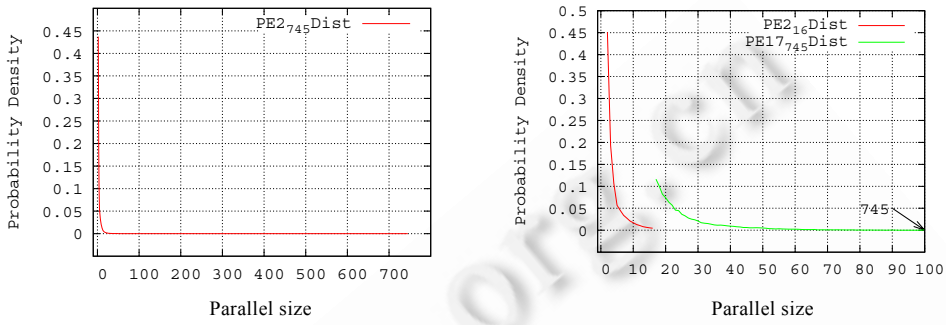
(2) 任务的并行尺寸与运行时间之间的关系

在 BGI 标准化(SWF)的日志中,$P=1$ 的任务占 79.6%,而 $P\geqslant 2$ 的任务占 20.4%,因此受 $P=1$ 时任务数目的影响,从整体上拟合 P 的分布情况将会变得异常困难.所以为了更好地构建 P 与 R 之间的任务模型,本文主要分析 $P\geqslant 2$ 时 P 的分布情况,以及 P 与 R 之间的关系.

1) 当 $P\geqslant 2$ 时 P 的分布情况.易得图 4.图中 PE 后面的数字和数字下标表示 P 的取值范围,在图 4(a)中,可以看出 P 的分布趋向于长尾分布[15],随着 P 的增大,P 的概率密度急剧下降,最大取值超过了 700.关于长尾分布的数据生成可以通过 matlab 的 gprnd 函数产生服从长尾分布的数据,但是拟合的效果较差.因此,本文通过将 $P\geqslant 2$ 时的分布进行分段分析,然后寻找出每一段的分布情况.考虑到 $P\geqslant 2$ 时 P 的概率密度分布情况,将其分成两部分:$2\leqslant P\leqslant 16$ 时和 $17\leqslant P\leqslant 745$ 时 P 的分布情况.由图 4(b)可知,两个区间内 P 的概率密度分布均服从指数或者伽马分布,由于指数分布是伽马分布的特例,本文选用伽马分布进行拟合.

2) P 与 R 之间的关系.本文分析 P 的每一个不同取值与 R 之间的关系.标准化的日志中显示,当 $P>13$ 时,P 对应的任务数目急剧下降,因此很难找出每一个 P 的取值同 R 之间的关系,所以本文设定了一个任务数目的阈值(=10000),当一个 P 对应的任务数目小于这个阈值时,就加上下一个 P 对应的任务数,直到任务数超过这个阈值为止.易得图 5.图 5 表示不同 P 的取值范围内运行时间的变化情况,其中横坐标轴采用的是对数坐标.由图 5 中关于不同 P 下 R 的概率密度分布可以看出,P 和 R 之间是长尾分布的关系.同时,从图中和原有数据中可将 P 和 R 之间的关系进行分类:当 $P=1$ 时,P 和 R 之间趋向于一种长尾分布;当 $2\leqslant P\leqslant 16$ 时,P 和 R 之间趋向于另一种长尾分布;当 $17\leqslant P\leqslant 745$ 时,P 和 R 之间趋向于另一种长尾分布.同样,对 3 种长尾分布的"尾巴"进行单独拟合,本文设定分割 3 种长尾分布的两个阈值分别为 R_{low_th}=100s 和 R_{mid_th}=10000s,然后对大于和小于这两个

阈值的运行时间的分布情况进行分析可得图 6.由图 6 可以看出,在 P 的 3 个取值范围内,在 R_{low_th} 以及 R_{mid_th} 的限制下,任务的运行时间的概率密度趋向于指数或伽马分布.由于指数分布是伽马分布的特例,因此本文采用伽马分布进行拟合.



(a) (b)

Fig.4 Distribution of the job's parallel size (≥2)

图 4 任务的并行尺寸大于等于 2 时的分布情况

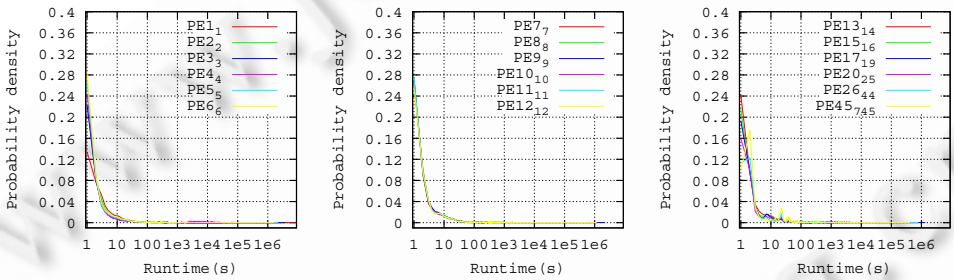
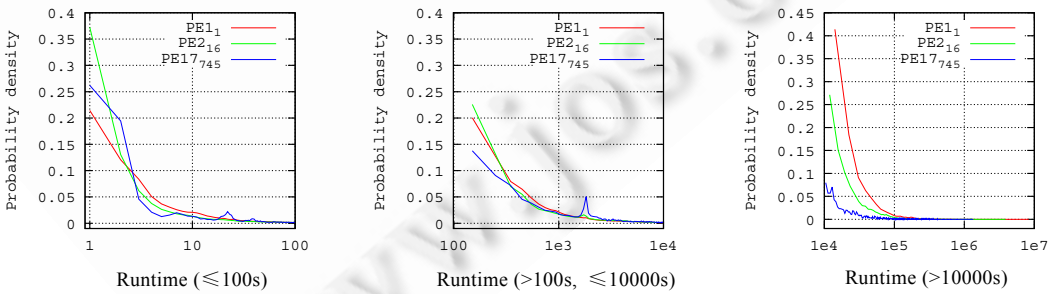


Fig.5 Relationship between the job's parallel size and runtime

图 5 任务的并行尺寸与任务运行时间之间的关系



(a) (b) (c)

Fig.6 Distributions of runtime greater and less than the threshold of the three parallel types

图 6 大于和小于阈值的 3 种并行尺寸任务的运行时间分布情况

2.3 生物基因测序计算任务的任务特性

通过以上对日志数据的分析,我们发现生物基因测序计算任务呈现如下特性:

(1) 任务到达时间上的特性.在工作日周期特性上:从 0:00 到 8:00 之间,每个小时内任务到达数目逐渐下降,且在 8:00 时任务到达数目到达的任务数目最少.从 8:00 到 24:00,每个小时到达的任务数目先增高后降低,且出

现两次波峰,分别为 12:00 和 17:30 左右.节假日周期特性:与工作日周期特性类似,但是从 8:00 到 24:00 只出现了一次任务到达数目的波峰,在 17:00 左右.

(2) 任务的并行尺寸 P 与任务运行时间 R 之间的关系.整体来看, P 与 R 之间存在长尾分布的关系特性,但是由于长尾分布利用相应的长尾函数进行拟合,会存在很大的偏差,因此本文将任务的并行尺寸分为 3 个区间,我们发现,在这 3 个取值区间内, P 与 R 之间存在伽马分布的关系.

因此,深入了解生物基因测序计算任务的任务特性有助于为该计算任务设计合适的调度算法,从而优化系统性能.

3 基于高性能计算日志的任务建模

本文利用上一节分析的结果构建相应的任务模型,主要包括任务到达时间的模型(简记为 DCModel)以及任务的并行尺寸同任务运行时间的模型(简记为 PRModel),最后将这两个任务模型合并成最终的模型.合并后的任务模型可以生成类似实际环境下生物基因测序日志的任务负载,即可生成不同规模集群的任务负载,进而可以通过仿真进行不同规模集群的性能评估并优化资源的配置.DCModel 主要用于产生任务到达的时间间隔;PRModel 则主要用于产生任务的并行尺寸和任务的运行时间长度.另外,本文主要使用了 3 种分布函数的数据生成方法:指数分布、伽马分布以及正态分布.具体的数据生成方式参见文献[15].本文正态分布的数据生成方法采用的是 Box-Muller[16]方法,同时采用了线性回归分析.

3.1 任务到达时间的模型

本文使用的主要符号变量有: x_i , y_i , $\overline{x}_i$ 和 $\overline{y}_i$ 这 4 个变量的具体含义参考上一节对任务到达时间规律性分析; t_i 表示第 i 个时间槽任务到达的时间间隔; α_i 和 β_i 表示工作日中服从伽马分布的第 i 个时间槽的两个参数; $WorkDC$ 表示工作日周期性, $WeekDC$ 表示节假日周期性.上一节关于任务到达时间的规律性的分析结果可以通过式(3)和式(4)进行描述.通过式(3)和式(4)可以产生服从指数分布和伽马分布的每个时间槽内到达的工作日和节假日的任务数量.由文献[1]和概率论的知识可知,集群中任务到达的时间间隔服从指数分布,且本文中单个时间槽的时间长度是 3 600s,因此可得式(5).通过式(5)可以产生工作日和节假日的任务到达时间间隔,即可完成 DCModel 的模型构建.

$$WorkDC \Rightarrow \begin{cases} x_i \sim E(\overline{x}_i), & 1 \leqslant i \leqslant 9 \\ x_i \sim \Gamma(\alpha_i, \beta_i), & 10 \leqslant i \leqslant 24 \end{cases} \tag{3}$$

$$WeekDC \Rightarrow y_i \sim E(\overline{y}_i), \quad 1 \leqslant i \leqslant 24 \tag{4}$$

$$t_i \sim \begin{cases} E(3600/x_i), & 1 \leqslant i \leqslant 24 \\ E(3600/y_i), & 1 \leqslant i \leqslant 24 \end{cases} \tag{5}$$

$$a_1 i^5 + a_2 i^4 + a_3 i^3 + a_4 i^2 + a_5 i + a_6 = \overline{x}_i \tag{6}$$

$$b_1 i^5 + b_2 i^4 + b_3 i^3 + b_4 i^2 + b_5 i + b_6 = \overline{y}_i \tag{7}$$

下面计算标准化任务日志中的 $\overline{x}_i$, $\overline{y}_i$, α_i 和 β_i 的取值.首先,由上一节关于任务到达时间规律性的分析可知,任务到达时间所在的时间槽与 $\overline{x}_i$ 及 $\overline{y}_i$ 存在非线性关系,因此可以通过两个多项式拟合这种非线性关系,本文选用两个五次多项式来拟合,如式(6)和式(7)所示,拟合的结果见表 1.可以通过 matlab 的 gamfit()函数计算出工作日中时间槽 10~24 的伽马分布的参数值,见表 2 和表 3.图 7 是 DCModel 生成时间间隔的伪代码.伪代码中的 sample 函数用于产生相应分布函数的样本点.

Table 1 Regressed results of the workday and weekend cycles

表 1 关于工作日和节假日任务到达数目周期性的拟合参数值

k	1	2	3	4	5	6
a_k	0.00 7	−0.391	6.2	−18.156	−115.618	1 007.48
b_k	0.00 3	−0.159	2.667	−10.408	−62.059	935.408

Table 2 Gamma distribution parameters of the time slots between 10 and 27

表 2 工作日中时间槽 10~17 之间任务到达数的伽马分布的参数值

i	10	11	12	13	14	15	16	17
α_i	1.353	1.135	1.485	1.121	1.147	1.197	1.467	1.377
β_i	4.165	3.613	2.439	4.862	4.868	3.767	2.435	2.698

Table 3 Gamma distribution parameters of the time slots between 18 and 24

表 3 工作日中时间槽 18~24 之间任务到达数的伽马分布的参数值

i	18	19	20	21	22	23	24	18
α_i	1.296	1.306	1.265	1.161	1.192	1.218	1.093	1.296
β_i	2.678	3.127	4.326	5.256	5.909	6.251	8.142	2.678

Conditions: $\overline{x}_i$, $\overline{y}_i$, α_i and β_i

Input: St —The start time, $Size$ —The number of jobs to generate

Output: Tl —The time length, Tll —The list of Tl

```
WHILE Size →0 THEN
| IF St in Workday THEN
| |     IF St in time slots (1,9) THEN
| |     |     x_i = E(x̄_i).sample(); Tl = E(3600/x_i).sample();
| |     ELSE IF St in time slots (10,24) THEN
| |     |     x_i = Γ(α_i,β_i).sample(); Tl = E(3600/x_i).sample();
| |     END
| ELSE IF St in Weekend THEN
| |     y_i = E(ȳ_i).sample(); Tl = E(3600/y_i).sample();
| END
| Tll.add(Tl);
END
```

Fig.7 The psuodo-code of DCModel

图 7 DCModel 生成时间间隔的伪代码

3.2 任务并行尺寸同运行时间的模型

本文使用的符号变量有: P, R, R_{low_th} 和 R_{mid_th} 具体含义参考上一节对任务并行尺寸和任务运行时间关系的分析; *pdf* 表示概率密度函数.上一节关于 P 的分布规律可以用式(8)来描述, P 和 R 之间的关系可以通过式(9)~式(11)来描述.可以通过日志求得 3 个公式中伽马分布的参数值,见表 4.同时,需要计算出 P 在 3 个区间内出现的概率值用来选定 P 服从的分布函数,以及对应 P 的每个区间内 $R \leqslant R_{low\_th}$ 的概率值, $R_{low\_th} < R \leqslant R_{mid\_th}$ 和 $R > R_{mid\_th}$ 的概率值用来选定 R 的分布函数.利用标准化的日志容易计算出以上的概率值,见表 5.

$$P \sim \begin{cases} U(1,1), & P=1 \\ \Gamma(\alpha_{11},\beta_{11}), & 2 \leqslant P \leqslant 16 \\ \Gamma(\alpha_{12},\beta_{12}), & 17 \leqslant P \leqslant 745 \end{cases} \tag{8}$$

$$R(\leqslant R_{low\_th}) \sim \begin{cases} \Gamma(\alpha_{21},\beta_{21}), & P=1 \\ \Gamma(\alpha_{22},\beta_{22}), & 2 \leqslant P \leqslant 16 \\ \Gamma(\alpha_{23},\beta_{23}), & 17 \leqslant P \leqslant 745 \end{cases} \tag{9}$$

$$R(> R_{low\_th} \,\&\, \leqslant R_{mid\_th}) \sim \begin{cases} \Gamma(\alpha_{31},\beta_{31}), & P=1 \\ \Gamma(\alpha_{32},\beta_{32}), & 2 \leqslant P \leqslant 16 \\ \Gamma(\alpha_{33},\beta_{33}), & 17 \leqslant P \leqslant 745 \end{cases} \tag{10}$$

$$R(>R_{mid\_th})\sim\begin{cases}\Gamma(\alpha_{41},\beta_{41}), & P=1\\ \Gamma(\alpha_{42},\beta_{42}), & 2\leqslant P\leqslant 16\\ \Gamma(\alpha_{43},\beta_{43}), & 17\leqslant P\leqslant 745\end{cases} \tag{11}$$

Table 4 The Gamma distribution parameters of the equations from Eq.(8)~Eq.(11)

表 4 式(8)~式(11)中伽马分布的参数值

i	(α_{1i},β_{1i})	(α_{2i},β_{2i})	(α_{3i},β_{3i})	(α_{4i},β_{4i})
1	(2.8574, 1.3648)	(0.6466, 24.5366)	(0.7349, 2109.2)	(0.9640, 52143)
2	(6.2272, 4.3883)	(0.5338, 25.5052)	(0.7037, 2333.5)	(1.1162, 38474)
3	(NA, NA)	(0.5900, 27.2626)	(0.8369, 2375)	(1.1102, 40013)

Table 5 Distributed probability of the job's parallel size and runtime

表 5 任务并行尺寸以及运行时间分布的概率值

$pdf(P,R)$ & $pdf(P)$		$P=1$	$2\leqslant P\leqslant 16$	$17\leqslant P\leqslant 745$
$pdf(P,R)$	$R\leqslant R_{low\_th}$	0.625	0.638	0.576
	$R_{low\_th}<R\leqslant R_{mid\_th}$	0.334	0.322	0.378
	$R>R_{mid\_th}$	0.041	0.04	0.046
$pdf(P)$	NA	0.796	0.198	0.006

另外,为了选取 P 和 R 的概率分布函数,需要 4 个伪随机函数发生器来进行选择.将这 4 个伪随机函数发生器分别标记为 Rnd_{11}, Rnd_{21}, Rnd_{22} 及 Rnd_{23}.其中, Rnd_{11} 用于选择 P 所在的区间,利用所在区间的分布函数产生 P 的并行尺寸; Rnd_{21} 用于选择 $P=1$ 内的 R 的分布函数; Rnd_{22} 用于选择 $2\leqslant P\leqslant 16$ 内的 R 的分布函数; Rnd_{23} 用于选择 $17\leqslant P\leqslant 745$ 内的 R 的分布函数.

3.3 合并的任务模型

DCModel 主要负责产生生物基因测序高性能计算日志中任务到达时间的间隔,因此可以用来控制集群系统的负载大小.PRModel 主要负责产生到达任务的任务运行时间和并行尺寸等特性.本文将 DCModel 产生的任务到达时间间隔作为一个触发点来触发 PRModel 产生任务的运行时间和并行尺寸,然后产生一个完整的任务记录,最后完成两个任务模型的融合.DCModel 和 PRModel 融合的具体流程是:首先通过 DCModel 产生一个任务的时间间隔 T ;最后通过 PRModel 产生任务的并行尺寸 P 和任务的运行时间 R ,然后格式化为 SWF 格式[3]的负载.图 8 是整个流程的伪代码,其中, *size* 代表产生的任务负载的数目, *load* 表示需要产生的任务负载使用的集群的规模(默认为 1,1 表示现有集群的规模,当 *load* 为 2 时代表生成日志的集群规模为现有的 2 倍).

```
Conditions: DCModel, PRModel
Input:  size, load
Output:  swf -The workload list
DCModel.setLoad( load );
WHILE  size →0 THEN
| T =DCModel.generateTimeLength();
|( P,R )=PRModel.getPR();
| swf .add(SWFFormat( T,P,R ));
END
```

Fig.8 The pseudo-code of the final model

图 8 合并的任务模型产生任务负载的伪代码

4 任务模型的评测

本文使用两种评测方法:Kolmogorov-Smirnov 评测和 Anderson-Darling 评测,分别简记为 KSTest 和 ADTest.

KSTest 就是通过计算日志中的样本点的累加函数值与任务模型中相应样本点的累加函数值之间的最大距离来评测模型的优劣.具体的评测过程参见文献[17].KSTest 评测的是模型产生的样本点的累加概率密度与原日志文件中样本点的累加概率密度之间的最大差值,但没有给出整体样本拟合的评测效果.ADTest 则是在 KSTest 基础上改进的一种评测方法,这种评测方法可以对模型同原日志文件之间进行整体的评测,具体的评测过程参见文献[18].为了更好地评测本文构建的任务模型所产生日志文件的可靠性,本文使用 KSTest 和 ADTest 对最终任务模型在生成的任务到达时间间隔 T, 并行尺寸 P 和任务运行时间 R 上的分布趋势进行了评测.

4.1 KSTest评测

原有日志文件的文件规模,即日志记录的任务数目大约为 550 万条.为了使结果更具说服力,本节将原有日志拆分为两份任务日志,第 1 份任务日志含有 300 万条任务记录(标记为 LOG_1),第 2 份任务日志含有 250 万条任务记录(标记为 LOG_2).然后,利用任务模型分别产生相同数量的任务日志对 LOG_1 和 LOG_2 中的 P, R 和 T 这 3 个任务属性进行 KSTest 评测.其中 KSTest 评测结果的取值范围为[0,1],结果越小说明任务模型与原有日志之间的趋势拟合越好.为了更好地评测任务模型的可靠性,本文分别对 LOG_1 和 LOG_2 进行了 5 次评测,评测结果见表 6,对 LOG_1 和 LOG_2 的评测显示,任务模型产生的 P, R 和 T 这 3 个任务属性中,评测结果均小于 0.1,平均结果显示 P, R 和 T 的评测结果均保持在 0.08 以下,而 P 的评测结果则保持在 0.05 以下,因此任务模型产生的任务日志在 P, R 和 T 这 3 个属性上与原日志趋于相同的分布.

4.2 ADTest评测

对 KSTest 中针对 LOG_1 和 LOG_2 生成的 10 个日志文件进行 ADTest 评测,评测结果见表 6.在表 6 中,ADTest 对 P,R 和 T 的评测结果在 0.8~42 之间不等,这是因为 ADTest 评测是在 KSTest 评测基础上对整体样本的一种评测,是 KSTest 评测结果的标准差,评测结果同时取决于被评测的样本的数量.因此,ADTest 针对 4 种属性评测的具体含义是: P 的平均评测结果 0.64 表示任务模型生成的任务的并行尺寸平均偏差为 0.64; R 的平均评测结果 35 表示生成的任务运行时间的平均偏差为 35s; T 的平均评测结果 1.85 表示生成的任务时间间隔的平均偏差为 1.85s.ADTest 评测结果说明任务模型的 P 和 T 的分布趋势与原日志文件的趋势基本一致,而 R 的偏差略大.这是因为, R 的分布呈现长尾分布,同时,样本数目(即不同的运行时间数目)达到 8 万多个,样本点最大达 10E+06s,因此对 R 的分布趋势准确拟合有很大的难度.

Table 6 Results of regression trends between the workload model and the original trace by KSTest and ADTest

表 6 KSTest 和 ADTest 对任务模型同原日志文件拟合趋势的评测结果

Experiment	KSTest			ADTest		
	P	*R*	*T*	*P*	*R*	*T*
LOG_1(1)	0.037 16	0.080 19	0.055 33	0.452 10	41.495 16	1.764 09
LOG_1(2)	0.037 72	0.079 92	0.053 03	0.448 45	41.380 02	1.988 41
LOG_1(3)	0.037 34	0.080 15	0.058 88	0.446 56	41.425 22	1.882 14
LOG_1(4)	0.037 32	0.080 20	0.052 09	0.458 74	41.513 23	1.712 83
LOG_1(5)	0.037 50	0.079 40	0.051 28	0.449 31	41.152 71	1.689 44
LOG_2(1)	0.057 39	0.061 40	0.064 95	0.840 91	28.828 80	1.846 22
LOG_2(2)	0.057 45	0.060 88	0.068 17	0.852 02	28.686 44	1.854 07
LOG_2(3)	0.057 35	0.061 13	0.062 57	0.849 82	28.577 76	1.965 61
LOG_2(4)	0.057 16	0.060 21	0.061 91	0.840 04	28.577 84	1.971 12
LOG_2(5)	0.057 69	0.061 01	0.066 96	0.844 71	28.805 68	1.881 69
Average	0.047 41	0.070 45	0.059 52	0.648 27	35.044 29	1.855 56

5 总 结

本文首先通过对生物基因测序日志的任务到达时间规律以及任务的并行尺寸和运行时间进行了分析.分析出日志文件中的任务在到达时间上具有工作日周期特性和节假日周期特性,同时在不同的时间槽内服从指数分布和伽马分布;任务的并行尺寸具有长尾分布特征,而任务的并行尺寸同任务的运行时间也存在长尾分布.

然后,对任务到达时间的特性、任务的并行尺寸与运行时间之间的关系利用数学关系式进行概括,并根据概括后的数学公式进行任务模型构建.最后对构建的任务模型进行了评测.评测主要包括 KSTest 和 ADTest 评测.KSTest 评测的平均结果显示任务模型生成的任务并行尺寸 P,任务运行时间 R 以及任务的时间间隔 T 同原日志中相应属性的趋于相同的分布,其中 P,R 和 T 的评测结果均保持在 0.08 以下,而 P 的评测结果则保持在 0.05 以下.ADTest 的评测结果显示, P,R,T 的平均评测结果分别为 0.65,35s 和 1.85s.

与任务日志相比,任务模型具有高伸缩性、灵活的参数配置以及可重复特性.因此,本文基于任务日志构建的任务模型超越了任务日志的局限性,具有很好的通用性,可用于生物基因测序环境中针对任务到达时间间隔、任务运行时间以及任务的并行尺寸特性的调度策略和性能评估方面的研究,从而使针对生物基因测序的资源调度策略和性能评估的研究更为便利.

References:

[1] Lublin U, Feitelson DG. The workload on parallel supercomputers: Modeling the characteristics of rigid jobs. Journal of Parallel and Distributed Computing, 2003,63(11):1104−1122. [doi: http://dx.doi.org/10.1016/S0743-7315(03)00108-4]

[2] Feitelson DG. Workload modeling for performance evaluation. In: Calzarossa MC, Tucci S, eds. Performance Evaluation of Complex Systems: Techniques and Tools. Berlin, Heidelberg: Springer-Verlag, 2002. 113−141. [doi: 10.1007/3-540-45798-4_6]

[3] Steve JC, Walfredo C, Feitelson DG, Jones JP, Leutenegger ST, Schwiegelshohn U, Smith W, Talby D. Benchmarks and standards for the evaluation of parallel job schedulers. In: Feitelson DG, Rudolph L, eds. Job Scheduling Strategies for Parallel Processing. Berlin, Heidelberg: Springer-Verlag, 1999. 67−90. [doi: 10.1007/3-540-47954-6_4]

[4] Feitelson DG, Nitzberg B. Job characteristics of a production parallel scientific workload on the NASA Ames iPSC/860. In: Feitelson DG, Rudolph L, eds. Job Scheduling Strategies for Parallel Processing. Berlin, Heidelberg: Springer-Verlag, 1995. 337−360. [doi: 10.1007/3-540-60153-8_38]

[5] Parallel workloads archive. http://www.cs.huji.ac.il/labs/parallel/workload/

[6] Li H, Buyya R. Model-Based simulation and performance evaluation of grid scheduling strategies. Future Generation Computer Systems, 2009,25(4):460−465. [doi: http://dx.doi.org/10.1016/j.future.2008.09.012]

[7] Li H. Realistic workload modeling and its performance impacts in large-scale escience grids. IEEE Trans. on Parallel and Distributed Systems, 2010,21(4):480−493. [doi: 10.1109/TPDS.2009.99]

[8] Feitelson DG, Shmueli E. A case for conservative workload modeling: Parallel job scheduling with daily cycles of activity. In: IEEE Int'l Symp. on Modeling, Analysis and Simulation of Computer and Telecommunication Systems. London: IEEE, 2009. 1−8. [doi: 10.1109/MASCOT.2009.5366139]

[9] Li H. Long range dependent job arrival process and its implications in grid environments. In: Primet PVB, Welzl M, eds. Proc. of the 1st Int'l Conf. on Networks for grid applications. Belgium: Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, 2007. 26−33.

[10] Minh TN, Wolters L. Modeling job arrival process with long range dependence and burstiness characteristics. In: Proc. of the 9th IEEE/ACM Int'l Symp. on Cluster Computing and the Grid. Shanghai: IEEE, 2009. 323−330. [doi: 10.1109/CCGRID.2009.35]

[11] SGE Manual Pages. Sun Grid Engine accounting file format. http://arc.liv.ac.uk/SGE/htmlman/manuals.html

[12] Feitelson DG. Packing schemes for gang scheduling. In: Feitelson DG, Rudolph L, eds. Job Scheduling Strategies for Parallel Processing. Berlin, Heidelberg: Springer-Verlag, 1996. 89−110. [doi: 10.1007/BFb0022289]

[13] Jann J, Pattnaik P, Franke H, Wang F, Skovira, Riordan J. Modeling of workload in MPPs. In: Feitelson DG, Rudolph L, eds. Job Scheduling Strategies for Parallel Processing. Berlin, Heidelberg: Springer-Verlag, 1997. 94−116. [doi: 10.1007/3-540-63574-2_18]

[14] BGI-Shenzhen. http://www.genomics.cn/index

[15] Jain R. The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling. New York: John Wiley & Sons, 1991. 234−243.

[16] Box GEP, Muller ME. A Note on the Generation of Random Normal Deviates. The Annals of Mathematical Statistics, 1958,29(2): 610−611. [doi: 10.1214/aoms/1177706645]

[17] Stephens MA. EDF statistics for goodness of fit and some comparisons. Journal of the American statistical Association, 1974, 69(347):730−737. [doi: 10.1080/01621459.1974.10480196]

[18] Law AW, Kelton WD. Simulation Modeling and Analysis. 4th ed., New York: McGraw Hill, 2000. 197−199.

曹志波(1985－),男,河北邯郸人,博士,主要研究领域为云计算,网格计算,任务建模.

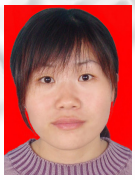
E-mail: caozhibo@126.com

董守斌(1967－),女,博士,教授,博士生导师,主要研究领域为高性能计算,海量信息处理.

E-mail: sbdong@scut.edu.cn

王丙强(1977－),男,博士,研究员,主要研究领域为高性能计算.

E-mail: wangbingqiang@genomics.cn



左利云(1980－),女,副教授,主要研究领域为云计算,资源评估.

E-mail: yuerly666@126.com