

网格任务的脉冲响应模型与预测控制调度策略

陈迎迎¹, 李艳君², 吴铁军¹

(1. 浙江大学 控制科学与工程学系, 浙江 杭州 310027; 2. 浙江大学 城市学院, 浙江 杭州 310015)

摘要: 采用反馈控制策略, 处理网格环境中的任务调度问题. 利用任务并行度等内部结构信息, 在任务接纳速度与网格计算资源使用量之间, 建立了任务脉冲响应模型. 采用预测控制策略对任务接纳速度进行在线调节, 以消除网格动态不确定性因素对于任务执行的影响. 采用人工免疫算法进行优化求解, 最大化网格吞吐能力. 仿真结果验证了模型的正确性和本文算法的有效性.

关键词: 网格计算; 任务调度; 预测控制; 人工免疫算法

中图分类号: TP273 **文献标识码:** A

Impulse response model for grid task and predictive control strategy

CHEN Ying-ying¹, LI Yan-jun², WU Tie-jun¹

(1. Department of Control Science and Engineering, Zhejiang University, Hangzhou Zhejiang 310027, China;

2. City College, Zhejiang University, Hangzhou Zhejiang 310015, China)

Abstract: The feedback control policy is employed to solve the grid task scheduling problem. According to the parallelism of tasks, a task impulse response model is developed to represent the relation between the task acceptance speed and the grid resource usage. In order to reduce the adverse effect of dynamic uncertainties on tasks processing, a predictive control strategy is used to adjust the task acceptance speed online; and an artificial immune algorithm is used to maximize the throughput of the grid. Results of simulation study show the validity of the proposed model and the effectiveness of the algorithm.

Key words: grid computing; task scheduling; predictive control; artificial immune algorithm

1 引言(Introduction)

网格调度是实现高性能计算网格的基本要素^[1,2], 根据分布并行任务对资源的要求, 资源自身特性和状态等调度信息, 设计规则将任务分配到适当的计算资源上执行, 满足特定性能指标(如最小任务运行时间, 任务平均响应时间, 最大系统利用率或吞吐量). 与以往并行计算环境不同, 网格环境中存在多种不确定性动态因素, 如通信带宽和服务质量不稳定, 计算资源负载状态不确定性变化等, 造成调度任务所依据的资源状态信息无法预先准确得知, 增加了任务调度的难度. 为了解决这一问题, 不少学者对网格中的不确定性动态因素进行了研究并从概率上分析了计算资源本地负载状况的变化情况及其对任务运行时间的影响^[3,4]. 虽然这些研究揭示了一些不确定性因素的变化规律, 但基本上仅能提高平均性能, 且对于如网格结构变化, 通信带宽延迟等不确定性因素的研究还远远不够, 将所有不确定性因素综合考虑也存在困难^[5].

针对上述问题, 将控制理论应用于解决网格调度

问题也逐渐受到重视^[6,7]. 采用控制理论方法, 首先分析影响系统输出的关键因素, 并将其作为输入, 建立系统响应模型, 将其他非关键因素作为扰动, 采用反馈原理, 将系统性能输出的采样值与设定值进行比较, 根据比较结果通过控制算法调节输入, 使系统输出满足控制要求^[8,9].

本文采用反馈控制策略, 利用任务的并行度等内部结构信息, 参照控制系统中的脉冲响应, 建立任务接纳速度(单位时间内进入网格并开始运行的任务数)与网格计算资源使用量之间关系的脉冲响应模型, 将计算资源本地负载等不确定因素对任务运行的影响作为系统扰动, 采用预测控制结合人工免疫算法调节任务接纳速度, 最大化网格处理器资源使用量. 仿真结果表明, 使用本文提出的脉冲响应模型和调度方法, 可最大化资源使用量和网格任务吞吐量, 提高网格抵抗系统扰动的能力.

2 调度问题描述(Problem statement)

在实际运算中, 有些并行任务的子任务数量在运

行过程中会发生变化. 因此, 其在运行过程中所需要的处理器数量也是变化的. 当有大量任务需处理而有限数量处理器资源不能同时满足所有任务对资源的要求时, 可通过调度任务的接纳速度, 将各任务安排在不同时段执行, 使得网络中的处理器资源得到充分利用, 提高网络任务吞吐量. 如图1(图中圆圈代表一台处理器运行一个单位时间)所示, 网络中共有4台处理器, 任务A的运行分为3个阶段, 第1, 第3阶段分别需一台处理器运行一个单位时间, 第2阶段需两台处理器分别运行一个单位时间. 当任务A可用的处理器数量足够多时, 可获得最短执行时间. 当有大量任务A需处理时, 从时刻0开始, 每个时刻点加入一个任务A并运行, 时刻2之后, 可以看到4台处理器全部被使用, 且每个任务都获得最小执行时间, 此时网络任务吞吐量最大.

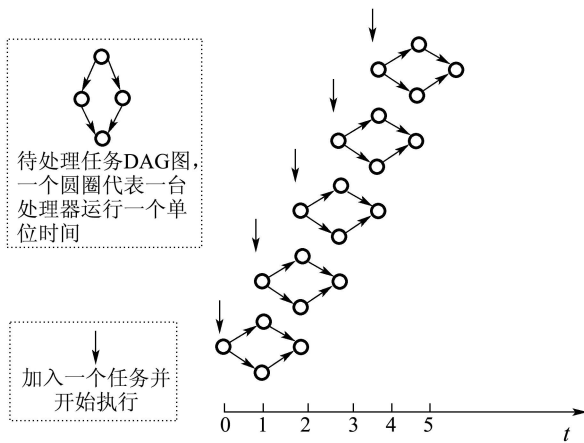


图1 并行任务执行过程

Fig. 1 Processing of parallel tasks

为调节任务接纳速度, 最大化处理器资源使用量, 本文建立了反映任务接纳速度与处理器资源使用量关系的任务脉冲响应模型, 并求得在网络无扰动, 各任务运行时间不变情况下的最优任务接纳速度序列. 由于网络中存在通信带宽和服务质量不稳定, 计算资源本地负载影响等扰动因素, 任务执行发生延迟, 在无扰动情况下求得的最优任务接纳速度会使网络系统发生任务拥塞. 这是因为在该速度下网络资源使用量最大, 网络处于饱和状态. 当任务未能按期完成离开系统时, 会与后期进入网络的任务争用资源, 使得一部分任务因得不到足够的资源而阻塞, 同时它们所占用的资源也无法释放, 最终可能导致网络中全部资源被阻塞, 网络进入死锁状态. 为避免这种情况的发生, 本文根据建立的任务脉冲响应模型, 实时采样任务对资源的使用信息, 调节预测控制器在线调节任务接纳速度, 采用人工免疫算法进行优化, 以最大化处理器资源使用量. 上述网络调度问题的一般数学描述如下:

$$\begin{aligned} \max_{u(t)} Q &= \int_0^{\infty} y(t) dt, \\ \text{s.t. } y(t) &= f(u(t), t), \\ y(t) &\leq N, \\ u(t) &\in \{0, \mathbb{Z}^+\}, \end{aligned} \quad (1)$$

其中: N 为网络中的资源总量, 即处理器总数, 各处理器同构, $y(t)$, $u(t)$ 分别为时刻 t 的资源使用量和任务接纳速度, $f(\cdot)$ 是关于 u, t 的函数. 目标值 Q 表示调度时间之内的资源使用总量, 当 Q 最大时网络任务吞吐量也最大. 约束条件说明网络中的处理器资源数量有限, 并且各时刻进入网络并运行的任务数为整数.

3 任务脉冲响应模型(Task impulse response model)

网络中的任务具有一定的生命周期, 并且根据上文所述, 在运行过程中对计算资源数量要求不同, 这与控制系统中被控对象的脉冲响应形式类似. 根据上述特点, 以任务在各阶段所需的资源数量为模型参数, 本文建立反映任务接纳速度与资源使用量关系的任务脉冲响应模型. 任务 J 定义为 $((t_1, p_1), (t_2, p_2), \dots, (t_i, p_i), \dots, (t_m, p_m))$, $i = 1, 2, \dots, m$, 其中 (t_i, p_i) 表示 J 在第 i 阶段需要 p_i 台处理器资源运行 t_i 时间, $t_m = 1, p_m = 0$ 表示任务执行完毕, 退出网络系统, 任务各阶段是依次执行的. 定义 $p(t, T)$ 为从时刻 T 开始运行的任务在时刻 t 需要的处理器资源数量, 即:

$$p(t, T) = (p_i | l = \sup\{i : \sum_{j=1}^i t_j + T \leq t\}), i \leq m. \quad (2)$$

由此可得出

$$y(t) = \int_0^t u(T) p(t, T) dT. \quad (3)$$

由于 $u(t)$, $p(t, T)$ 均不是解析式, 本文对其进行离散化处理, 设采样周期为 τ , 有 $\left\lceil \frac{t_i}{\tau} \right\rceil = n_i$, 任务 J 的离散化描述为 $((n_1, p_1), (n_2, p_2), \dots, (n_i, p_i), \dots, (n_m, p_m))$, 有:

$$p(k, K) = (p_i | l = \sup\{i : \sum_{j=1}^i n_j + K \leq k\}), i \leq m, \quad (4)$$

$$y(k) = \sum_{K=0}^k u(K) p(k, K). \quad (5)$$

其中 $k = 0, 1, 2, \dots, L$ 为采样时刻(L 为整个调度过程经历的采样时间长度). 当 $K \leq k - n$, $n = \sum_{i=1}^m n_i$ 时 $p(k, K) = 0$, 由式(5)得到任务脉冲响应模型:

$$y(k) = \begin{cases} \sum_{K=k-n+1}^k u(K)p(k, K), & k \geq n, \\ \sum_{K=0}^k u(K)p(k, K), & k < n, \end{cases} \quad (6)$$

$$k = 0, 1, 2, \dots, L.$$

基于上述模型, 式(1)定义的网格调度问题可转化为以下形式:

$$\begin{aligned} \max_{u(K)} Q &= \sum_{k=0}^{\infty} y(k), \\ \text{s.t. } y(k) &= \begin{cases} \sum_{K=k-n+1}^k u(K)p(k, K), & k \geq n, \\ \sum_{K=0}^k u(K)p(k, K), & k < n. \end{cases} \\ y(k) &\leq N, \\ u(K) &\in \{0, \mathbb{Z}^+\}, k = 0, 1, \dots, L. \end{aligned} \quad (7)$$

根据式(7), n 个采样时间过去后, 处理器资源使用量 $y(k)$ 就由包括第 k 采样时间在内的前 n 个任务接纳速度 $[u(k-n+1) \ u(k-n+2) \ \dots \ u(k-1) \ u(k)]$ 决定. 在任务各阶段执行时间不变的情况下, 当 $u(k)$ 序列呈周期性变化时, $y(k)$ 也将呈周期性变化. 因此, 可寻找一条长度为 n 的最优任务接纳速度序列 $\mathbf{u}^* = [u(0) \ u(1) \ \dots \ u(n-1)]$, 将该序列中的值依次作为网格的任务接纳速度, 使得1个周期内的处理器资源使用量最大. 为方便表述, 仅对第2个周期内的处理器资源使用量进行讨论. 该周期第1个采样时刻的处理器资源使用量 $y(n)$ 由上1周期第2个采样时刻的接纳速度 $u(1)$ 至最后1个采样时刻的 $u(n-1)$ 以及当前采样时刻的任务接纳速度 $u(n)$ 决定. 依次类推, 各采样时刻的处理器资源使用量均由包括当前采样时刻在内的前 n 个接纳速度决定, 即:

$$\begin{cases} y(k) = \sum_{K=k-n+1}^{k-1} u(K)p(k, K) + u(k)p(k, k), \\ k = n, n+1, \dots, 2n-1. \end{cases} \quad (8)$$

由于 $u(k)$, $u(K)$ 分别为 \mathbf{u}^* 中的 $u(k-n)$ 和 $u(K-n)$, 因此第2个周期内处理器资源使用量最大化问题可表述为:

$$\begin{aligned} \max_{\mathbf{u}^*} Q &= \sum_{k=n}^{2n-1} y(k), \\ \text{s.t. } y(k) &= \sum_{K=k-n+1}^{k-1} u(K-n)p(k, K) + u(k-n)p(k, k), \\ y(k) &\leq N, \\ u(k) &\in \{0, \mathbb{Z}^+\}, k = n, n+1, \dots, 2n-1. \end{aligned} \quad (9)$$

当网格中无扰动发生时, 式(9)优化问题中求得的 \mathbf{u}^* 可使网格处理器资源使用量最大. 但是当网格中存在扰动时, 将 \mathbf{u}^* 作为任务接纳速度则可能发生如上文所述的死锁现象. 为避免该情况的发生, 需根据实时采集的任务执行情况, 采用预测控制结合人工免疫算法对 u 进行在线调整, 使网格系统正常运行, 最大化任务吞吐量.

4 预测控制器设计(Predictive controller design)

模型预测控制是目前在工业过程控制中应用比较广泛的先进控制方法, 它的核心思想是多步预测、滚动优化和系统校正. 在各采样点, 控制器预测系统未来一段采样时间长度 M (预测窗口)内的性能, 选择一条输入轨迹, 最优化关于这 M 采样时间内的性能.

采用预测控制策略对本文问题进行求解. 首先在采样时刻 k 根据任务脉冲响应模型对资源使用量 y 进行 M 步预测, 得模型输出 \mathbf{y}_m :

$$\mathbf{y}_m(k+1) = H_1 \mathbf{u}_1(k) + H_2 \mathbf{u}(k+1), \quad (10)$$

其中 $\mathbf{u}_1(k)$ 为 k 时刻之前的 $(n-1)$ 个任务接纳速度, $\mathbf{u}(k+1)$ 待优化变量表示 k 时刻及之后的 $M-1$ 个任务接纳速度. 式(10)中向量分别为:

$$\begin{cases} \mathbf{u}_1(k) = [u(k-n+1) \ u(k-n+2) \ \dots \ u(k-1)]', \\ \mathbf{u}(k+1) = [u(k) \ u(k+1) \ \dots \ u(k+M-1)]', \\ \mathbf{y}_m(k+1) = [y(k) \ y(k+1) \ \dots \ y(k+M-1)]. \end{cases} \quad (11)$$

H_1, H_2 为模型参数矩阵,

$$\begin{cases} H_1 = \begin{bmatrix} h_n & h_{n-1} & \dots & h_2 \\ & h_n & \dots & h_3 \\ & & \ddots & \vdots \\ 0 & & & h_n \dots h_{M+1} \end{bmatrix}_{M \times (n-1)}, \\ H_2 = \begin{bmatrix} h_1 & & & \\ h_2 & h_1 & & 0 \\ \vdots & \vdots & \ddots & \\ h_M & h_{M-1} & \dots & h_1 \end{bmatrix}_{M \times M}. \end{cases} \quad (12)$$

其中 $h_i = p(i, 0)$, $i = 1, 2, \dots, n$. 任务的运行时间可能会因为系统扰动而延长, 由于只能检测到任务是否延迟进行下一阶段, 所以系统采样值为当前时刻未能按期开始下一阶段运行的任务数 $\Delta \mathbf{u}_1(k) = [\Delta u(k-n) \ \Delta u(k-n+1) \ \dots \ \Delta u(k-n+i) \ \dots \ \Delta u(k-2)]$, 根据 $\Delta \mathbf{u}_1(k)$ 对 $\mathbf{u}_1(k)$ 进行修正得 $\mathbf{u}_1(k)_p$:

$$\mathbf{u}_1(k)_p = \mathbf{u}_1(k) + (E - F)\Delta \mathbf{u}_1(k), \quad (13)$$

E 为 $n-1$ 阶单位矩阵, F 为 $n-1$ 阶移位矩阵,

$$F = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}_{(n-1) \times (n-1)}, \quad (14)$$

用 $\mathbf{u}_1(k)_p$ 替换式(10)的 $\mathbf{u}_1(k)$ 得预测输出 \mathbf{y}_p :

$$\mathbf{y}_p(k+1) = H_1 \mathbf{u}_1(k)_p + H_2 \mathbf{u}(k+1), \quad (15)$$

根据调度要求, 将滚动优化目标定义为:

$$\begin{aligned} \max_{\mathbf{u}(k+j)} Q &= \sum_{j=0}^{M-1} y_p(k+j), \\ \text{s.t.} \quad &\begin{cases} \mathbf{y}_p(k+1) = H_1 \mathbf{u}_1(k)_p + H_2 \mathbf{u}(k+1), \\ y_p(k+j) \leq N, \\ u(k+j) \in (0, \mathbb{Z}^+), j = 1, 2, \dots, M. \end{cases} \end{aligned} \quad (16)$$

式(16)为带约束的非线性规划整数问题, 本文采用人工免疫算法进行数值求解. 将求得的任务接纳速度序列 $\mathbf{u}(k+1)$ 的第一个值 $u(k)$ 作为下一采样时刻间隔的任务接纳速度, 略去序列中的其他取值. 在各个采样时刻重复上述滚动优化过程.

5 基于人工免疫算法的优化求解(Optimization based on IA)

由于本文的调度过程是一个滚动优化过程, 候选解的某些特征会重复出现, 而人工免疫算法^[10]具有记忆功能, 较之其他数值优化方法, 更适合此处的优化问题. 首先对任务接纳速度 $\mathbf{u}(k+1)$ 采用如下方式进行编码:

$$\mathbf{u}(k+1) = [e_k \ e_{k+1} \ \cdots \ e_{k+l} \ \cdots \ e_{k+M-1}], \quad (17)$$

其中等位基因 $e_{k+l} \in [0, 1, 2, \dots, E_m]$, $l = 0, 1, \dots, M-1$, 且

$$E_m = \left\lfloor \frac{N}{np^*} \right\rfloor, \quad p^* = \min\{p_e, e = 1, 2, \dots, m\},$$

抗体 i , j 的相似度 $C_{u(k)_i^j}$, 适应度值 $J_{u(k+1)_i^j}$ 和繁殖概率 $P_{u(k+1)_i^j}$ 分别定义为:

$$\begin{cases} C_{u(k+1)_i^j} = \frac{1}{N_1 + N_2} \sum_{i=1}^{N_1+N_2} \xi_{ij}, \\ \xi_{ij} = \begin{cases} 1, & \|\mathbf{u}(k+1)_i^j - \mathbf{u}(k+1)_i^i\| \leq \varepsilon, \\ 0, & \text{其他}. \end{cases} \\ J_{u(k+1)_i^j} = \|H_1 \mathbf{u}_1(k)_p + H_2 \mathbf{u}(k+1)_i^j\|^2, \\ P_{u(k+1)_i^j} = \frac{J_{u(k+1)_i^j} / C_{u(k+1)_i^j}}{\sum_{i=1}^{N_1+N_2} J_{u(k+1)_i^j} / C_{u(k+1)_i^j}}. \end{cases} \quad (18)$$

交叉算子为等位替换算子^[11], 而变异算子则是将随机选取的某位基因用 $[0, 1, 2, \dots, E_m]$ 中任一项代替.

6 仿真与结果分析(Simulation)

本节对算法进行评估. 在每个采样周期对以概率 p_d 产生干扰任务, 使其延迟进入下一阶段. 设网格中计算资源总量为100, 任务 J 为 $((2, 1), (4, 3), (1, 2), (5, 1), (1, 0))$, 完成一个任务需要12个采样周期.

当网格无扰动发生时, 所有的任务都没有发生延迟, 易求得此时的最优处理速度序列 $\mathbf{u}^* = [5 \ 5 \ 5 \ 4 \ 5 \ 5 \ 5 \ 4 \ 5 \ 5 \ 5 \ 4]$, 资源使用量随时间变化情况如图2中曲线所示, 此时平均每一采样时刻完成4.75个任务, 在12个采样周期过后, 网络进入稳定状态, 资源使用量 y 始终保持在资源总量100附近.

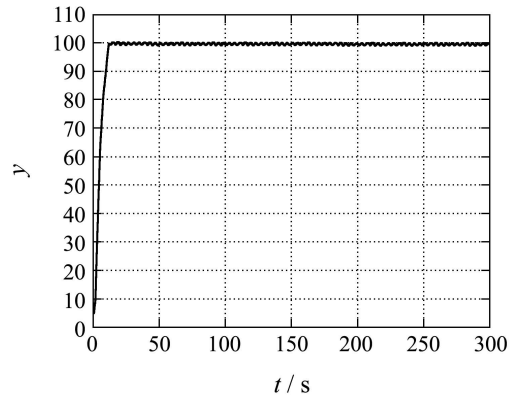
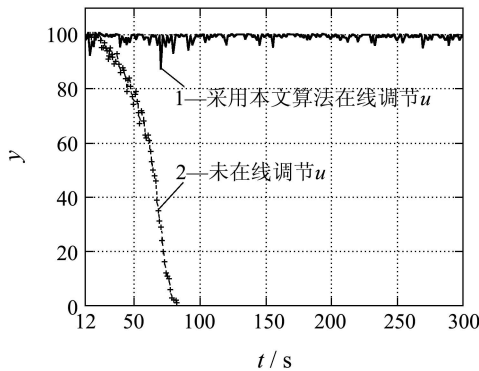
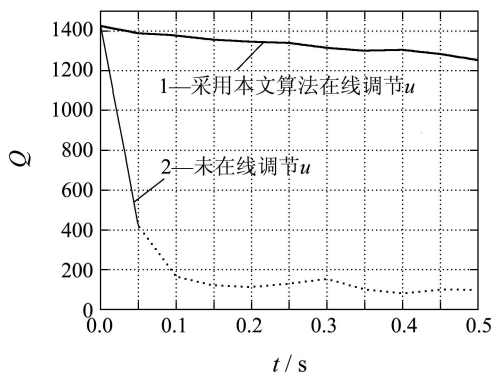


图2 无扰动时处理器使用量

Fig. 2 Processors usage without disturbance

当网格系统有扰动时, 采用预测控制在线调整 u 后, 资源使用量能够保持在资源总量100附近, 没有死锁的情况发生, 如图3中曲线1所示. 由该曲线还可得知, 处理器使用量这一性能与无扰动时(如图2)相比有所下降. 这说明扰动的存在影响了系统的性能. 在一些采样时刻, 资源使用量明显低于100, 这是由于当扰动发生时, 本文采用的预测控制算法在对预测时段内的性能进行评估后, 计算出此时应降低任务接纳速度, 减少处理器使用量, 才能避免在后面的时段发生死锁. 作为比较, 本文也对不使用预测算法在线调节 u 的情况进行了仿真, 如图3中曲线2所示, 经过一段时间后, 处理器使用量降为0, 系统进入了死锁状态.

随着延迟概率 p_d 增大, 网格系统在仿真时间内的计算资源使用量 Q 逐渐减小. 但是, 采用本文算法调节 u 后, Q 值虽然有所下降, 但下降幅度不大且与 p_d 的增长几乎呈线性关系, 如图4中曲线1所示; 而当仍依 \mathbf{u}^* 确定 u 时, 网络吞吐能力急剧下降, 且 p_d 越大死锁发生的时间越早, 如图4中曲线2所示.

图3 延迟概率 $p_d = 0.5$ 时的处理器使用量Fig. 3 Processors usage with the $p_d = 0.5$ 图4 延迟概率 p_d 增加时的任务吞吐量Fig. 4 Total processors usage with the increasing of p_d

7 结束语(Conclusions)

本文研究了网格环境中的任务调度问题, 根据任务运行过程中对处理器资源数量要求的变化, 建立了任务脉冲响应模型, 使用预测控制方法, 根据系统的性能输出, 调节任务处理速度, 抑制了网格资源状态变化对系统性能的影响, 保证了系统性能. 虽然本文工作主要处理大量相同任务在同构处理器组成的网格环境中的调度问题, 但所提出的方法同样适用于异构环境下多种任务的调度. 只需建立多输入多输出的任务脉冲响应模型描述各任务的接纳速度对各种资源使用量的影响. 由于各任务的运行时间长度不同, 在确定预测窗口时有一定困难, 所以还须进一步的研究.

参考文献(References):

- [1] XHAFAA F, ABRAHAMB A. Computational models and heuristic methods for Grid scheduling problems[J]. *Future Generation Computer Systems*, 2010, 26(4): 608 – 621.
- [2] YU J, BUYYA R, RAMAMOHANARAO K. Workflow scheduling algorithms for grid computing[M] //Xhafa F, Abraham A. *Meta-heuristics for Scheduling in Distributed Computing Environments*. Berlin: Springer, 2008, 146: 173 – 214.
- [3] LI K. Optimal load distribution in nondedicated heterogeneous cluster and grid computing environments[J]. *Journal of Systems Architecture*, 2008, 54(1/2): 111 – 123.
- [4] WU M, SUN X. Grid harvest service: a performance system of grid computing[J]. *Journal of Parallel and Distributed Computing*, 2006, 66(10): 1322 – 1337.
- [5] DONG F, AKL S G. Scheduling algorithms for grid computing: state of the art and open problems, Technical Report 2006-504[ER/OL]. Kingston, Ontario: School of Computing, Queen's University, 2006.
- [6] ZHANG Y, KOELBEL C, COOPER K. Hybrid re-scheduling mechanisms for workflow applications on multi-cluster grid[C] //The 9th IEEE/ACM International Symposium on Cluster Computing and the Grid. Piscataway, NJ: IEEE, 2009: 116 – 123.
- [7] NOU R, KOUNEV S, JULIÀ F, et al. Autonomic QoS control in enterprise Grid environments using online simulation[J]. *Journal of Systems and Software*, 2009, 82(3): 486 – 502.
- [8] LU C, WANG X, KOUTSOUKOS X. Feedback utilization control in distributed real-time systems with end-to-end tasks[J]. *IEEE Transactions on Parallel and Distributed Systems*, 2005, 16(6): 550 – 561.
- [9] LIU X, ZHU X, SINGHAL S, et al. Adaptive entitlement control of resource containers on shared servers[C] //The 9th IFIP/IEEE International Symposium on Integrated Network Management. Piscataway, NJ: IEEE, 2005: 163 – 176.
- [10] HIGHTOWER R. Computational aspect of antibody gene families[D]. Albuquerque, New Mexico: University of New Mexico, 1996.
- [11] 蔡自兴, 龚涛. 免疫算法研究的进展[J]. *控制与决策*, 2004, 19(8): 841 – 846. (CAI Zixing, GONG Tao. Advance in research on immune algorithms[J]. *Control and Decision*, 2004, 19(8): 841 – 846.)

作者简介:

陈迎迎 (1979—), 女, 博士研究生, 研究方向为网格资源调度、智能优化算法、复杂系统控制与优化, E-mail: yychen@iipc.zju.edu.cn;

李艳君 (1973—), 女, 教授, 硕士生导师, 研究方向为智能控制与优化、计算智能、复杂系统建模, E-mail: yjlee@iipc.zju.edu.cn;

吴铁军 (1950—), 男, 教授, 博士生导师, 研究方向为复杂系统控制与优化、智能机器人系统、导航制导与控制, E-mail: tjwu@zju.edu.cn.