软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(2):429–442 [doi: 10.13328/j.cnki.jos.005026]

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

基于模型预测控制的数据中心节能调度算法*

赵小刚[1], 胡启平[1], 丁 玲[2], 沈志东[1]

[1](武汉大学 国际软件学院 软件工程系,湖北 武汉 430079)

[2](湖北科技学院 计算机科学与技术学院,湖北 咸宁 437100)

通信作者: 赵小刚, E-mail: zxgang302@whu.edu.cn

摘 要: 如今日益增长的数据中心能耗,特别是冷却系统能耗已日益受到重视,降低系统能耗能够减少数据中心碳排放.提出了一种基于模型预测控制(model prediction control,简称 MPC)的节能调度策略,该策略可以有效地减小数据中心冷却能耗.该方法采用动态电压频率调节技术来调整计算节点频率,从而减少节点间的热循环;所有节点的峰值温度可被保持在温度阈值下,在任务的执行中稳态误差较小.该方法可以通过动态频率调节来抑制由于负载类型变化造成的模型不确定性带来的内部扰动,分析结果表明,基于模型预测的温控算法系统开销较小,具有良好的可扩展性.基于该算法设计的控制器能够有效地降低输入温度,提高数据中心能耗效率.通过在实际数据中心内运行的模拟网上书店,该方法与安全最小热传递算法和传统反馈温控算法这两种经典方法相比,无论是在正常条件下还是在扰动存在的情况下都能取得较好的温度抑制效果,系统性能如吞吐率也达到最大.在相同的负载条件下,该方法能够获得最小的输入峰值温度和最小的冷却能耗.

关键词: 模型预测控制;反馈控制;热传递;阈值温度;内部扰动;能耗效率

中图法分类号: TP316

中文引用格式: 赵小刚,胡启平,丁玲,沈志东.基于模型预测控制的数据中心节能调度算法.软件学报,2017,28(2):429–442. http://www.jos.org.cn/1000-9825/5026.htm

英文引用格式: Zhao XG, Hu QP, Ding L, Shen ZD. Energy saving scheduling strategy based on model prediction control for data centers. Ruan Jian Xue Bao/Journal of Software, 2017,28(2):429–442 (in Chinese). http://www.jos.org.cn/1000-9825/5026.htm

Energy Saving Scheduling Strategy Based on Model Prediction Control for Data Centers

ZHAO Xiao-Gang[1], HU Qi-Ping[1], DING Ling[2], SHEN Zhi-Dong[1]

[1](Department of Software Engineering, Int'l School of Software, Wuhan University, Wuhan 430079, China)

[2](College of Computer Science and Technology, Hubei University of Science and Technology, Xianning, 437100, China)

Abstract: Today the ever-growing energy cost, especially cooling cost of data centers, draws much attention for carbon emission reduction. This paper presents an energy efficient scheduling strategy based on model prediction control (MPC) to reduce cooling cost in data centers. It uses dynamic voltage frequency scaling technology to adjust the frequencies of computing nodes of a cluster in a way to minimize heat recirculation effect among the nodes. The maximum inlet temperature of nodes can be kept under temperature limits with little stable error. The method can also deal with inner disturbance (system model variation) by dynamic frequencies regulation among the nodes. Analysis shows good scalability and small overhead, making the method applicable in huge data centers. A temperature-aware controller is designed to reduce inlet temperatures to improve energy efficiency of data centers. Using a simulated online bookstore run in a heterogeneous data center the proposed method is proved to have larger throughput in both normal and emergency cases compared with existing solutions such as safe least recirculation heat temperature controller and traditional feedback temperature controller. The

* 基金项目: 国家自然科学基金(61003185); 湖北省自然科学基金(201FFB04505)

Foundation item: National Natural Science Foundation of China (61003185); Natural Science Foundation of Hubei Province of China (201FFB04505)

收稿时间: 2015-01-14; 修改时间: 2015-09-10, 2015-12-22; 采用时间: 2016-01-05

MPC-based scheduling method also has less inlet temperature and cooling cost comparing with those two methods under same workload.
Key words: model prediction control; feedback control; heat recirculation; temperature threshold; inner disturbance; energy efficiency

随着信息社会的迅速发展,网上购物、在线搜索和云计算服务等使得数据中心日益成为计算密集和存储密集的中心环节.海量的计算和存储需求使得数据中心必须增大硬件规模来获得快速的响应.而硬件规模的增长带来的不仅仅是购买成本的迅速增加,庞大的硬件能耗更是给数据中心带来巨大的运行成本.Gartner 估计到2015 年,约占全世界数据中心 71%比例的大型数据中心能耗将超过 1 262 亿美金[1].而到 2016 年~2018 年,数据中心的能耗将达到 600 亿千瓦时~1 300 亿千瓦时[2].巨大的能耗成本使得数据中心的运行维护成本已经远远超过了数据中心的建设成本,即购买数据中心所必须的大量硬件设备和场地所花的费用[3].

减小数据中心能耗是学术界和工业界目前的共识[4−7].数据中心能耗的一部分为硬件运行时需要的能耗,称为计算能耗.很多硬件设计公司和数据中心建设公司都致力于减小计算能耗,并已经从不同层次开发出了硬件节能技术.Intel 等芯片公司主要从芯片级进行节能,如低电压集成电路、动态电压频率调整技术(dynamic voltage and frequency scaling,简称 DVFS)[8]等.HP,IBM 等服务器公司则主要从机架级进行节能,如刀片服务器集合层能耗管理[9]、硬件能耗监控和能耗优化的一体化方案[10,11].很多研究侧重于从数据中心负载的特性来减少能耗:如按照系统功率需求动态提供能耗的虚拟能耗技术,该技术基于可存储电能的电池设备[12];尽量减少 CPU 空闲时间的后填充任务发射技术[13];当 CPU,硬盘或网络空闲时将其转入睡眠态节能[14,15].这些技术都取得了较好的效果,但都局限于部件级的节能,缺乏对温度引起的冷却能耗的关注.

数据中心能耗的另一部分为冷却能耗.密集硬件运行时的高能耗会使数据中心产生大量热量,而热量的增加会使硬件的温度增加,工作稳定性降低.为了维持数据中心的持续健康运行,需要较低的环境温度,而由此产生的热循环现象是目前数据中心能耗效率低下的主因.研究表明,比较理想的制冷能耗和硬件能耗为 1:1.2011 年对 500 个大型数据中心进行调查发现[16],绝大多数的数据中心制冷能耗和硬件能耗比为 1.8:1.由此可见,数据中心的制冷成本非常高,而制冷能耗过大对环境也有较高的要求.一个传统的 15kW 的水冷数据中心需要一天 36 000 加仑的水来制冷[17].

冷却能耗的计算涉及到比较复杂的热力学过程,机器表面材质的散热系数、刀片服务器本身的间隔、数据中心内空气的流速等都会影响冷却能耗大小.服务器厂商和数据中心建设厂商主要从这个方面进行节能研究[18].C-Oracle 通过分析数据中心节点目前温度分布状态,估计其采用不同的作业调度算法后节点未来的温度变化,最终来选择合适的调度方法,是一种离线作业调度策略[18];Thermocast 通过时间序列方法估计未来 5 分钟内数据中心节点温度的变化情况,但缺乏实际的温度调节方案[19];Thermostat 则从仿真角度出发,以流体热力学理论建模数据中心,准确度很高,但没有给出具体的仿真参数及过程[20].这些调度方法都缺乏在线作业调度机制,难以对温度进行实时控制,特别是对紧急情况引起的温度异常的处理.

数据中心作业调度算法对数据中心冷却能耗影响较大.不同的调度算法会使数据中心节点的峰值温度不同,从而使得冷却设备的供应温度有差异.数据中心的冷却能耗与冷却设备的供应温度成反比[21,22].目前大多数调度算法都是通过降低节点的峰值温度来减小对空调的制冷能耗要求.Co-Con 采用控制算法来协调集群内主机能耗和主机上虚拟机的响应时间,但没有考虑温度的变化[23];算法 MinHR[24]根据每个节点的热传递系数(heat recirculation fraction,简称 HRF)来分配任务,对降低冷却能耗有一定的效果;算法 Xint[22]采用基因算法来离线进行计算作业调度,能够求得较优的冷却能耗,缺点是计算时间过长,不能用于在线作业调度.这些算法都能提高数据中心能耗效率,但都没有从温度角度考虑作业的性能要求.

数据中心的供应温度高低直接影响数据中心冷却能耗的大小.因此保持数据中心各节点温度的稳定性对于设定合理的供应温度,降低数据中心能耗有关键作用.本文拟研究温度敏感的节点能耗调度算法对数据中心温度的影响,着重于冷却能耗的降低.该算法采用先进的模型预测控制算法在节点间分配功耗,然后根据任务温度时域模型估算各节点温度;采用温度反馈控制算法实时控制各节点频率,以达到准确跟踪参考温度的目的;如果在任务执行过程中发生温度异常情况造成热点产生,可以利用调度算法来动态地调整节点频率,在达到降温

目的的同时,吞吐率受影响较小,从而最小化系统冷却能耗.

本文第 1 节将简述节点功耗模型,数据中心热交换模型和冷却功耗模型.第 2 节分析模型预测温控的基本原理、总体架构、稳定性和控制器开销.第 3 节运行模拟实验,对模型预测温控算法的控制准确度、性能分析、扰动抑制和冷却能效分别设计不同的实验,以证明该算法的有效性.第 4 节讨论模型预测温控对系统能效的负面影响,并给出负载均衡的解决思路.第 5 节对本文加以总结,并指出进一步的工作方向.

1 数据中心功耗模型

数据中心普遍采用刀片服务器,采用机架方式将刀片服务器进行固定,通过空调进行制冷.从图 1 可以看出,同一机架中,空调供应的冷空气从底部往上部流动,从机架顶部散发的热空气被循环到空调处.因此,机架上部的服务器温度较高.而冷空气从刀片服务器的进风口流入,带走服务器内部工作部件(CPU、内存、硬盘、主板和网络设备等)的热量,从出风口流出.

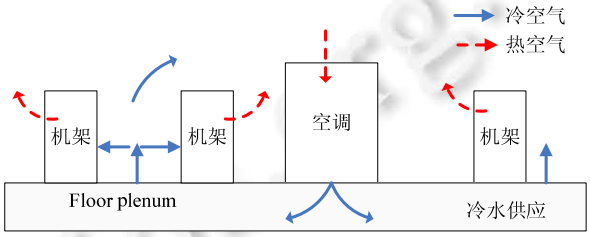


Fig.1 Air flow in data centers

图 1 数据中心空气流动

因此,研究数据中心功耗模型的主要工作是要计算机架中节点的功耗和带走节点产生热量的制冷设备功耗.

1.1 服务器功耗模型

许多研究表明服务器节点功耗与服务器频率成正比[23,25],因此服务器在 kT 时刻的功耗模型可以表示成

$$P_i(kT)=a_i f_i(kT)+c_i \tag{1}$$

a_i,c_i 为与机器硬件相关的常量,T 为获取功耗的时间间隔.根据上式可以得到节点功耗动态模型:

$$P_i((k+1)T)=P_i(kT)+a_i\Delta f_i((k+1)T) \tag{2}$$

$P_i((k+1)T)$为 i 节点在$(k+1)T$ 时刻的功耗,$\Delta f_i((k+1)T)$为$(k+1)T$ 时刻与 kT 时刻 i 节点 CPU 上的频率差异,商用计算机 CPU 频率不能连续变化,存在若干个 P 状态.

1.2 热能交换模型

图 2 描述了数据中心内服务器之间存在的热循环.研究表明,一台服务器的热量输出会对它周围所有服务器的输入温度产生影响.研究人员提出了很多方法来对数据中心内的热循环进行建模.这些模型的有效性已经由真实数据中心内使用传感器测得的温度数据进行了验证.与复杂且耗时的传统 CFD(computational fluid dynamics)仿真模型相比,这些温度预测模型的准确度和快速性都很高[22].从图 2 中得出服务器节点 i 的温度模型为

$$T_{\text{out}}^i=T_{\text{in}}^i+K_iP_i \tag{3}$$

其中,T_{out}^i 为服务器 i 的输出温度,T_{in}^i 为服务器 i 的输入温度,K_i 为热力学常量.

节点输入温度所包含的热量和节点功耗产生的热量之和就是节点的输出热量[22],则 T 时刻节点 i 所产生的热量如式(4)所示.

$$Q_{\text{out}}^i(kT)=Q_{\text{in}}^i(kT)+P_i(kT) \tag{4}$$

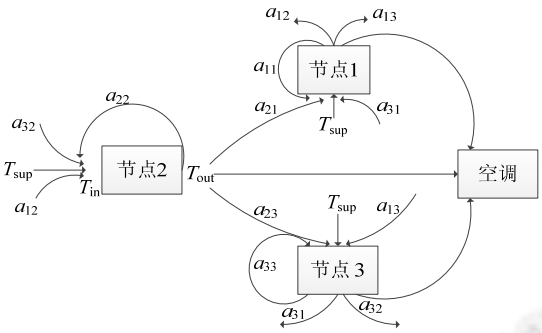


Fig.2 Heat recirculation among nodes in one rack

图 2 同一机架内节点热量交互

根据热力学公式,节点 i 输入热量可以表示为

$$Q_{in}^{i}(kT)=\rho c_p f_i T_{in}^{i}(kT) \tag{5}$$

这里,ρ,c_p,f_i 为热力学常量.其中,ρ为空气密度(典型值为 1.19kg/m^3),c_p 单位空气携带的热量(典型值为 1005Jkg^{-1}K^{-1}),f_i为节点 i 内空气流速(典型值为 520CMF=0.2454m^3/s).这里,我们采用文献[22]中采用的交互系数矩阵 $A_{n\times n}=\{\alpha_{ik}\}$来描述节点间的热量传播,如图 2 所示,其中,$\alpha_{ik}$ 表示第 i 个节点输出热量中传输到第 k 个节点的比例.因此,节点 i 在时刻 kT 时的输入热量为所有节点输出热量对 i 节点的影响加上空气循环设备的供应温度所带来的热量.

$$Q_{\text{in}}^{i}(kT)=\sum_{k=1}^{n}\alpha_{ki}(Q_{\text{in}}^{k}(kT)+P_k(kT))+Q_{\text{sup}}^{i}(kT) \tag{6}$$

通过运用式(5)中的热常量,根据公式(6)我们能够得到计算节点的输入温度模型.令 K_i 表示常量 $\rho c_p f_i$, $K_{n\times n}=diag\{K_1,K_2,...,K_n\}$, 则数据中心所有节点的输入温度向量为[22]

$$t_{\text{in}}=t_{\text{sup}}+Dp,\ D=\left[(K-A^TK)^{-1}-K^{-1}\right] \tag{7}$$

其中,D 为热循环系数矩阵,利用 D 可以将功率消耗转化为温度.节点温度向量时域模型为

$$t_{\text{in}}(kT)=t_{\text{sup}}(kT)+Dp(kT) \tag{8}$$

对于数据中心来说,通常需要保持较低的输入温度,否则机器内部的器件如 CPU、硬盘、内存等会由于温度过高而降低寿命.也就是说,节点的输入温度不能超过某一温度阈值 t_{red}.

$$t_{\text{in}}^{i}(kT)\leqslant t_{\text{red}},\ \forall i=1,2,...,n \tag{9}$$

1.3 冷却功耗模型

数据中心内必须配置冷却系统(computer room air-conditioning,简称 CRAC).其实从图 2 中就可以看出冷却系统的作用,数据中心内的节点部件必须工作在较低温度,因此,必须要用冷却系统将热量带出去,使得所有节点工作在温度阈值 t_{red} 以下.冷却系统的设定供应温度会影响数据中心的能耗效率.冷却系统效率是由性能系数(coefficient of performance,简称 CoP)[24]来衡量的.

$$CoP=0.006\,8t_{\text{sup}}^2+0.000\,8t_{\text{sup}}+0.458 \tag{10}$$

而冷却功耗直接与 CoP 相关:

$$P_{AC}=\frac{P_c}{CoP(t_{\text{sup}})} \tag{11}$$

其中,P_c 为数据中心的机器功耗,P_{AC} 为 CRAC 功耗.当供应温度升高时,CoP 的值也会增大.从式(11)可以看出,CoP 越大,P_{AC} 会越小.因此,减小制冷功耗的直接方法就是增大 T_{sup}.但数据中心内所有机器要正常工作,输入温度 t_{in} 不能超过阈值 t_{red}.因此,最高输入温度和阈值之间差异为

$$\Delta = t_{\text{red}} - \max_{i=1,\ldots,n}\{t_{\text{in}}^i\} \tag{12}$$

从公式(12)可以看出,如果可以合理进行作业调度,使得所有机器的输入温度 t_{in} 的最大值达到最小化,就可以获得最小的冷却能耗.

2 温度敏感的调度算法设计

本节主要描述温度敏感的反馈控制调度系统的总体架构.在该架构中最重要的是一个多输入多输出温度控制器,该控制器能够调整各节点 CPU 频率来强化系统温度阈值限制.

2.1 调度算法总体架构

假设集群包含 n 个计算节点.此外,还有一个作为温度控制器的主节点,根据反馈的温度来管理这 n 个节点的频率.每个节点上有一个功率计算器,负责计算节点功率消耗;t_{ref} 为参考温度.图 3 中的反馈环路工作过程如下.

(1) 温度控制器工作后,输出各节点目标工作频率 $\boldsymbol{f}(t)$ $(f_1^j, f_2^j, \ldots, f_n^j)$.

(2) 根据节点功耗模型计算出集群内节点功耗向量 $\boldsymbol{p}(t)(P_1, P_2, \ldots, P_n)$.

(3) 根据热循环矩阵 D,可以得到所有节点输入温度向量 $\boldsymbol{t}_{\text{in}}(t)$ $(t_{\text{in}}^1, t_{\text{in}}^2, \ldots, t_{\text{in}}^n)$, 该向量输出到温度控制器.

(4) 温度控制器根据参考温度(t_{ref})和目前温度(被控量)的差异确定下一时刻频率 $\boldsymbol{f}(t+1)$(操纵变量).

从公式(8)中我们可以得到节点在 kT 时刻的温度动态模型,其中,我们认为在相当长一段时间内冷却系统供应温度不会变化.

$$t_{\text{in}}(kT) = t_{\text{in}}((k-1)T) + D\Delta p(kT) \tag{13}$$

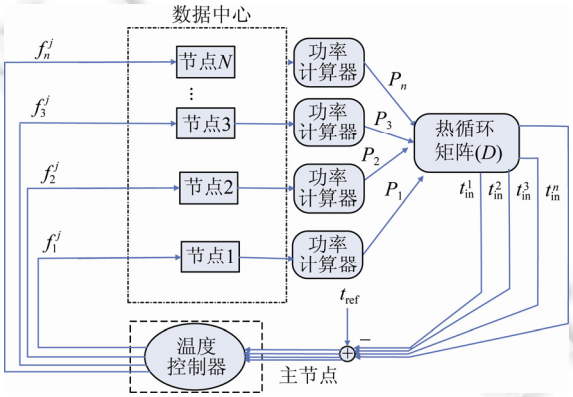


Fig.3 Architecture of thermal aware control system in data centers

图 3 温度敏感数据中心控制系统架构

根据公式(2)和公式(13),输入温度的动态模型可以由节点频率推导得出.

$$t_{\text{in}}(kT) = t_{\text{in}}((k-1)T) + DA\Delta f(kT) \tag{14}$$

其中, $A = \begin{bmatrix} a_1 & & & \\ & a_2 & & \\ & & \cdots & \\ & & & a_n \end{bmatrix}$.

从公式(14)中我们可以看到,通过调节节点的频率可以对节点温度进行控制,当节点频率确定后,节点的功耗和最大可以接受的负载量也可以确定下来.因此,对节点频率的调节实际上也是对节点负载的调度.这种调度是建立在系统级别,即在集群总体温度不超过温度阈值的情况下对集群内各节点频率的调度,而不是单节点内

操作系统根据 CPU 负载来调节自身频率.频率的调节在一定程度上会影响到任务的响应时间,但这里我们假定任务可以容忍一定的时延,关注系统的总体能耗的降低.

在温度敏感的控制系统中有一种形式的扰动为系统模型误差,称为内部扰动.在云计算平台这样的动态计算环境中,由于不同的负载类型很难对服务器节点功耗进行准确的估计[23],因为即使在同样的频率下,不同应用在单位时间内消耗的功率也是不一样的(CPU 密集或 I/O 密集).因此,节点功耗模型(2)只是一个对节点功耗的粗略估计.实际的温度向量可以表示为

$$t_{\text{in}}(kT)=t_{\text{in}}((k-1)T)+DgA\Delta f(kT) \tag{15}$$

其中, $g=\begin{bmatrix} g_1 & & & \\ & g_2 & & \\ & & \cdots & \\ & & & g_n \end{bmatrix}$.与公式(2)中仅用 a_i 作为频率系数相比,g_ia_i 才是真正的频率变化系数.g_i 为第 i 个节点估计频率与实际频率之间的差异.

理论上对内部扰动引起的温度波动可以通过温度反馈控制来抑制,并且抑制的指导原则是,在最小化冷却能耗的同时最大化系统利用率,同时,最好不要改变原有系统架构.因此,当扰动发生时,我们不采用降低供应温度或减少活动节点数目的方法.这里,我们只通过频率的调整来保持温度的阈值.

2.2 温度控制器设计

温度敏感数据中心控制系统架构中最核心的就是一个温度控制器,该温度控制器的设计基于模型预测控制理论.MPC 是一种高级控制技术.该控制技术能够有效处理变量耦合性较强的多输入多输出(multi-input multi-output,简称 MIMO)问题.这里的多输入多输出是指控制器必须要对集群内所有节点的频率进行控制,使得所有节点的输入温度低于温度阈值.耦合性强是指每个节点的功耗变化都会对其他节点的温度产生影响.MPC 通过将优化理论和过程控制结合起来,对输入向量 $\boldsymbol{f}(t+1)$进行统一求解,能够较好地应用于数据中心温度控制.

从第 2.1 节的分析可以得出,温度敏感数据中心要解决以下 3 个问题.

(1) 跟踪:所有节点的温度要逼近温度阈值.

(2) 封顶:所有节点在任何情况下,最大温度都不能超过温度阈值,节点频率不要超过可用频率.

(3) 优化:在温度调节过程中,系统的利用率要最大化;即使在温度异常情况下,系统性能也不能降低太多.

这 3 个问题 MPC 控制器都能够较好地解决.模型预测控制可以使用系统模型来预测系统将来的输出,然后使用该预测来优化控制信号.控制器通过有限 N 步预测对一个优化控制问题求解,该 N 步被称为预测维;控制目标是在限制条件下选择一个输入轨迹来最小化代价函数.输入轨迹包含在将来 M 个控制周期内的控制输入:$\Delta f(k),\Delta f(k+1|k),\ldots,\Delta f(k+M-1|k)$,这里,$M$ 被称为控制维.$\Delta f(k+i|k)$ 是指变量 Δf 在时刻 $k+i$ 的值依赖于 k 时刻的值,该值为 MPC 控制器对未来输入变量的预测.一旦输入轨迹计算出来,则只有第 1 个计算值$\Delta f(k)$被用来作为下一时刻的控制输入.在下一个控制周期,预测维度向前滑动一个周期,然后根据实际反馈输出 $\boldsymbol{t}_{\text{in}}$ 来再次计算控制输入.由于系统模型的不准确性,初始估计可能会存在偏差,因此,控制输入在每个周期都会重新计算.

在每个控制周期最后,控制器会计算控制输入$\Delta f(k)$,该控制输入可以最小化如下的代价函数:

$$V(k)=\sum_{i=1}^{P}\left\|t_{\text{in}}(t+i|i)-t_{\text{ref}}(k+i|k)\right\|_{Q(i)}^2+\sum_{i=1}^{M-1}\left\|\Delta f(t+i|i)+f(t+i|i)-f_{\max}\right\|_{R(i)}^2 \tag{16}$$

在式(16)中,代价函数的第 1 项代表跟踪误差,该误差代表 $t+i$ 时刻的节点温度向量与参考温度之间的差异.参考温度定义了一条理想的输出曲线,根据该曲线,所有节点温度应该从目前的温度 $t_{\text{in}}(t+i|i)$朝目标温度 $t_{\text{ref}}(k+i|k)$(比如温度阈值)变化.如果系统是稳定的,那么通过最小化温度误差,输入温度会趋向于目标温度.假如所有节点的温度都接近参考温度,也就是说,所有节点温度都接近,那么数据中心内节点温度分布平均,节点间热循环就会最小,系统冷却能耗就可以达到最小化.

式(16)中的第 2 项是控制惩罚.温度反馈控制系统在优化系统制冷效率时(第 1 项),同时要降低控制惩罚.第

2 项的减小意味着必须缩小系统最高频率($f_{\max}$)与目前的系统频率($f(t+i+1|i)=\Delta f(t+i|i)+f(t+i|i)$)的差异.当所有节点的频率都接近最高频率时,系统的利用率最高.

温度敏感的反馈控制器必须满足两个条件:首先,每个节点的频率必须在一个合理范围内,该范围可以由系统性能要求进行配置;其次,所有节点输入温度不能超过温度阈值.这两个条件可以表示为

$$\begin{cases} f_{i,\min} \leqslant \Delta f_i(k) + f_i(k) \leqslant f_{i,\max} (1 \leqslant i \leqslant N) \\ t_{\text{in}}^i \leqslant t_{\text{ref}} \end{cases} \tag{17}$$

式(17)通过使所有节点温度逼近阈值温度,避免了集群利用率低下的问题:利用率低下本身就会造成能耗浪费;而节点负载过高又会带来系统性能的下降(如响应时间过长).因此,为了系统能耗和性能的均衡,条件(17)将系统频率限定在一个较合理的区间$[f_{i,\min}, f_{i,\max}]$.综合式(16)和式(17),我们可以得到一个在输入、输出受限条件下的系统跟踪、封顶和优化方案,该方案可以和其他功耗管理机制同时使用,实现多种温度下,功耗管理方案的无缝集成.

2.3 稳定性分析

基于控制理论设计温度控制算法的优点在于,当系统模型(14)不确定时,控制理论提供了保持系统稳定性的理论依据.由于云计算平台上应用的多样性,不同应用在相同系统频率下的功耗也是不一样的,因此,理论功耗模型和实际功耗模型会存在差异.具体差异体现在式(15)中参数 g 的大小.模型的不确定性会造成系统输出、输入温度的波动.我们定义一个计算集群是温度稳定的,当且仅当输入温度 $t_{\text{in}}(k)$向量中的最大输入温度趋近参考点 t_{ref},即 $\lim_{k\to\infty}\max(t_{\text{in}}(k))=t_{\text{ref}}$.操作变量 $f(k)$可以通过当前温度 $t_{\text{in}}(k)$、温度参考点 t_{ref}和上一时刻的系统频率 $f(k-1)$计算得到.

根据控制理论,如果闭环系统的所有极点都处于复平面的单位圆内,而且从控制输入 u 到系统状态变量 x 之间的 DC 增益是一个单位矩阵,则系统的输出,如数据中心内节点的输入温度,会收敛到参考温度.

2.4 控制器开销

我们使用 Matlab 2010b 中的 Lsqlin 函数完成了基于模型预测的温度控制器.Lsqlin 的计算复杂度是与节点数目和控制/预测维度有关的.该仿真是在一个 PC 工作站上实现的.该工作站的配置为:双核 3.00GHZ Core 2 CPU,4GB RAM.在 Matlab 中每调用一次该函数,在 50 台服务器间搜索优化结果花费的时间少于 6ms.从上面的分析可知,当采样周期为 1s 时,温度控制器的计算开销只用了不到 1%的 CPU.

如今大型数据中心可以容纳超过 10 万台左右的服务器,也就是说,我们设计的温度控制器需要计算超过 1 万个变量(某些节点包含 10 个刀片).为了增强我们的温度控制器的扩展性,我们可以采用分布式的温度控制策略,即将这些服务器分成多个较小的簇(如每个簇包含 80 台~300 台左右的服务器),每个簇可以由一个温度控制器控制.这个分割可以基于空间局部性,即空间位置相近的簇可以分成一簇(这与热循环的空间局部性是一致的).这样的集群分布式策略被广泛应用于实际数据中心(如微软的 WAS[26]和谷歌的 GFS).

3 实验结果与分析

3.1 模拟实验环境

在实验中需要周期性地调度 Matlab 实现的基于模型预测控制的温度控制器.每个节点的频率统一由温度控制器来决定.我们使用功率计算器来完成对功率向量的计算,将收集到的温度向量反馈给温度控制器,控制器计算控制输入$\Delta f(k)$,将最后的频率向量 $f(k)$传递给数据中心各节点.

实验中采用的负载纪录是按照 TPC-W 标准生成的.TPC-W 是一个国际公认的对服务器性能进行测试的标准,其依据是看在电子商务环境下,一个服务器每秒能够完成多少次标准商务事务.TPC-W 中给出了 3 种典型的事务类型:浏览、购买和预定.这 3 种事务其实也是多种请求混合组成.通过在武汉大学国际软件学院计算中心的多台 WEB 服务器上部署一个遵循 TPC-W 标准的在线书店仿真程序,并在这些节点上调用命令来采集各节

点的吞吐率、CPU 利用率和频率.这种在线书店事务负载可以代表目前主流的 B2C 电子商务网站的数据中心的日常业务活动,而目前的数据中心中很大比例都是处理此类业务,因此,这类负载具有一定的代表性.

热循环矩阵 D 是通过一个开源的 CFD 仿真软件 OpenFoam 模拟数据中心内节点的热量的交互过程来获取的.图 4 描述了我们采用 OpenFoam 模拟的数据中心的物理架构.我们仿真了 50 个节点的数据中心,这 50 个节点被分布成两行、多个 42U 工业标准机架.数据中心的空间大小为 7.6m×7.6m×7.6m.冷空气是通过一个空调来供应的,空气流速为 8.5m^3/s.空调处于天花板上,离 x 轴大约 2.79m.中心共有 10 个机架,每个机架包含 5 个 7U 大小的底盘.第 1 行离 x 轴大约 2.5m,第 2 行距 x 轴为 4.2m.在后续所有的实验中,数据中心的物理结构和冷却配置没有变化,而所有这些配置获得的热循环矩阵($D4$,$D50$)都是通过这样的仿真得到的.

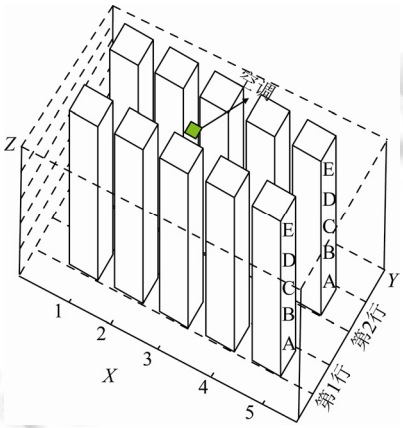
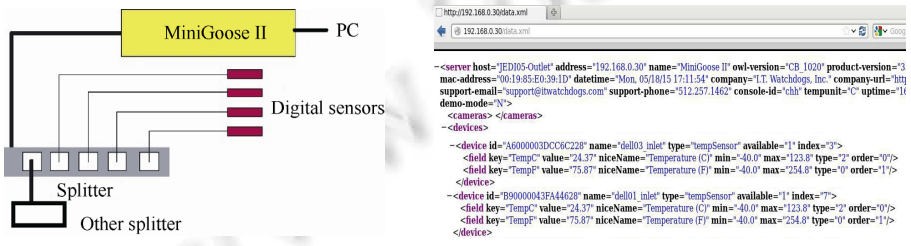


Fig.4 Physical layout of a data center

图 4 数据中心物理架构

3.2 真实硬件环境

武汉大学国际软件学院计算中心拥有 50 台机架型服务器,每台服务器大小为 1U[2],服务器型号为 Dell PowerEdge R220(20 台服务器的 CPU 为双核 Pentium 3.4G,另外 30 台 CPU 为四核 Xeon 3.4G).计算中心机房的物理结构和空调位置与模拟数据中心环境基本一致.我们将这些服务器放入两个机架中,其位置为第 1 行的第 4 列和第 5 列.机架上的每个底盘 A~E 都容纳 5 台服务器(每个机架大小为 7U,每台服务器大小为 1U,服务器间空隙为 0.5U).我们在每台服务器的空气入口处(服务器的前面板)放置温度传感器来实时地监控节点的输入温度.此处采用美国 ITWatchDogs 公司的温度传感器.这些传感器带有 RJ45 接口,可以方便与该公司生产的温度数据采集器 MiniGooseII 互连.MiniGooseII 可以直接读取温度传感器获取的节点输入温度数据.通过网络配置,PC 端可以通过 Web 页面直接监控从 MIniGooseII 传递过来的各节点温度.图 5(a)是 MiniGoose II 和 4 个温度传感器的逻辑连接示意图,图 5(b)是在 PC 机上显示的温度传感器获取的节点输入温度的 XML 文件.



(a) 温度传感器配置 (b) 温度数据的网页显示

Fig.5 Temperature sensors configuration and XML temperature data which generated by MiniGoose II

图 5 温度传感器配置和 MiniGoose II 生成的 XML 温度数据

初始实验中,我们使用了一个包含 4 个节点(2 台服务器 CPU 为 Pentium,另 2 台为 Xeon)的集群,放入第 1 行第 1 列的底盘 A,来观察基于模型预测的控制器对节点温度的影响.分别采用 4 个温度传感器安置在服务器的输入端.4 台服务器上均安装 Linux 系统,内核版本为 2.6,其中 3 台机器需安装 Apache 服务器作为 WEB 服务器,另有 1 台机器安装 MySql 作为数据库服务器.在机房外独立安装一台服务器(与机房内机器属于同一个网络),在此服务器上安装 LVS(Linux virtual server)模块来充当负载分配器,与此同时,MiniGoose II 也直接与该服务器连接.该服务器负责收集各节点温度,同时按照实时获得的温度进行频率的调节,也就是说,调度算法也在该服务器上实现.在 4 个节点的服务器集群上部署按 TPC-W 标准生成的在线书店程序.

数据中心模拟器也按照这里的初始设定模拟了一个四节点的集群,节点的功耗参数参考实际硬件(PowerEdge 220)的白皮书设定.这 4 个节点处于同一个机架内.在四节点的集群情况下,t_{ref} 为 25°C,而供应温度为 24°C.通常,大型数据中心的温度阈值也是 25°C.此外,PowerEdge R220 的处理器都有 6 种功耗状态($P0$,$P1$,…,$P5$),分别对应频率 3.4 GHz,3.0 GHz,2.6 GHz,2.2 GHz,1.8 GHz 和 1.6 GHz.实验中所有采样时间均为 1s.第 3.3 节~第 3.6 节中的温度均来源于实际温度传感器.

3.3 对比方法

我们将采用了 MPC 控制理论的温度控制器称为模型预测温控.这里,我们还设计了其他两种温度控制器,分别为热传递最小温控和传统反馈温控.这 3 种方式都是在线温控方法,因为它们在每个采样周期通过频率调整方法来最小化热量传递.从文献[22]中我们已经知道,在负载一定的条件下,在数据中心内求解合理的负载分配来实现最小化输入温度是一个 NP 问题,因此,通常在线求解最小温度都是采用启发式算法.而热传递最小温控方法是由文献[21,26]中的启发式算法推导得出的.为了阐述热传递最小温控方法,我们必须在式(7)介绍的热循环系数矩阵 D 的基础上进一步讨论数据中心内节点的热效率.

节点的热效率在很大程度上依靠节点本身和其相邻节点的热循环.热传递最小温控算法严格根据数据中心节点的热向量矩阵描述的节点热效率进行负载分配.数据中心热向量矩阵可以用 $\boldsymbol{R}=(R_1,R_2,...,R_i,...,R_n)$来表示,其中,$R_i$ 为节点 i 产生的热量对所有节点的影响,v_j 为代表热循环重要性的权值,d_{ij} 代表热循环矩阵 D 中的对应元素.

$$R_i=\sum_{j=1}^{n}v_j d_{ji}p_i^{\max},\ v_j=\sum_{i=1}^{n}d_{ji}p_i^{\max} \tag{18}$$

安全热传递最小温控算法(safe-least recirculation heat extended,简称 Safe-LRHex).在每个采样周期,当集群内所有节点的最大输入温度没有超过参考温度时,该方法从可用节点中选出具有最小 R_i 的节点,将其频率增加一级;当集群内有节点的最大输入温度超过参考温度时,则从可用节点中选出具有最大 R_i 的节点,将其频率减小一级;节点具有的 R_i 值越大,说明其热效率越高.当集群内节点的最大温度低于参考温度时,增大具有最小 R_i 节点的频率对整个集群的峰值温度影响最小;当最大温度高于参考温度时,减小具有最大 R_i 节点的频率,对整个集群的峰值温度下降得最快.由于该算法为一种试凑型算法,因此我们通过设定一个安全裕度来消除正误差,称其为安全热传递最小温控算法.

传统反馈温控算法(traditional feedback).参考文献[18]提出了一种通过动态调整节点负载来降低节点内各部件温度的反馈控制算法.然而该方法只是一个单节点温度控制算法,因为它只控制节点内功能部件,如 CPU、硬盘等的温度.该方法使用单输入单输出(single input/single output,简称 SISO)比例控制器来控制负载的变化,达到对组件温度的控制,但并没有考虑节点间的温度影响.这里,我们对该方法做了改进,使其适合分布式计算系统.在每个节点上我们都部署一个 SISO PID 型温度控制器,每个温控器通过调节本节点 CPU 频率来控制自己节点的温度.为了保证整个集群内所有节点的温度不超过阈值温度,这里采用了一个保守方法:将集群内所有节点的最低频率分配到每个节点.在每个采样周期末尾,所有节点都被分配同样的频率,该频率可以保障最高温度节点的温度不超过阈值.

3.4 性能分析

在这个实验中,我们比较集群级的温度控制,即安全热传递最小温控和模型预测温控对系统性能的影响.在

4 节点的集群中,分别将参考温度设定为 24.7°C,24.8°C 和 25°C,我们把这两种温控算法各运行 3 次.从图 6 可以看出,模型预测温控算法的吞吐率要高于安全热传递最小温控算法,特别是当参考温度较低,集群负载较低时.此外,图中虚线所示为没有采用 DVFS 温度控制算法的集群的最大理想吞吐率,所有节点的频率均为最大.当参考温度在 25°C 时,模型预测温控的性能较之理想状态有 10%左右的降低.这个损失是可以接受的,因为集群温度得到了控制,减少了系统冷却功耗(第 3.6 节).事实上,如果集群不采用温度控制,在虚线所示的最大理想吞吐率下,集群的峰值温度将超过参考温度,冷却能耗增加.

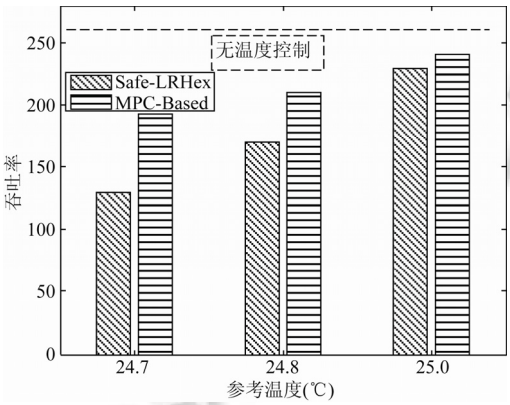


Fig.6 Throughput comparison between Safe-LRHex and MPC-based temperature controllers under different reference temperatures

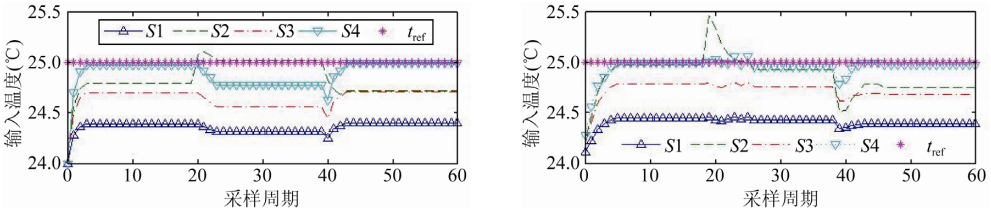
图 6 不同参考温度下的安全热传递最小温控与模型预测温控吞吐率比较

从第 2.2 节的分析过程可知,模型预测算法在相同的参考温度下会有较小的稳态误差和较大的集群利用率;安全热传递最小算法的调度策略则是先增加热效率较低的节点(节点 2 和节点 1),而这些节点的吞吐率较小(相同功率下).此外,安全热传递最小算法由于安全裕度的问题,节点利用率本身就比正常情况要低,所以这两种算法在负载较低时吞吐率差距较大.当参考温度增大时,所有节点利用率都较高,所以两种算法吞吐率 接近.

3.5 内部扰动抑制

在本节中我们分析当系统温度模型(式(15))由于功率模型的不确定性引发输入温度波动时,模型预测温控算法如何保持整个系统温度的稳定性.这里我们对比传统反馈温控算法,传统反馈温控基于经典控制理论,因此也可以获得离散频率来逼近浮点型频率.在经过 20 个周期后,节点 2 上的负载从订购模式变化到浏览模式,而浏览模式下,1 个浏览请求消耗的功率是订购模式下的 1.5 倍.从图 7 可以看出,传统反馈温控和模型预测温控对这种突变的处理过程是类似的,在节点 2 的温度升高后的一个周期,所有节点的温度都有所下降.其差别在于,在模型预测温控中剩余节点的温度下降幅度不大.经过短暂的温度调整后,传统反馈温控算法中节点 2 的温度回归到参考温度;而在模型预测温控中,节点 2 和 4 的温度都接近参考温度.

在第 40 个周期,所有节点上的访问模式由目前的浏览模式切换到购买模式,1 个购买请求消耗的功率是 1 个预定模式的 0.8 倍.因此,所有节点的温度在接下来的几个周期都有所下降,因为在相同的请求个数下会消耗更少的功率.接着,在反馈环路作用下,随着各节点请求数的增加,频率会随之上升.从图 7 可以看出,所有的节点温度都有所上升.



(a) 传统温度控制器 (b) 模型预测温度控制器

Fig.7 Temperature fluctuation containment under traditional feedback and MPC-based temperature controllers for different workload types: 1 browsing=1.5 ordering, 1 shopping=0.8 ordering

图 7 传统反馈温控和模型预测温控对不同负载类型引发温度波动的抑制,1 个浏览=1.5 预订,1 个购买=0.8 预订

图 8 展示了在访问模式变化的情况下,在两种不同的温控算法下,集群吞吐率的变化.从图 8 中更加确定了当负载类型变化时,模型预测温控的性能损失较小(公式(16),MPC 控制器在控制过程中寻找最小化代价函数的解).当节点 2 上的负载访问模式功率消耗量为原来的 1.5 倍时,根据公式(15)可知,集群内节点的峰值温度会有很大的升高,无论是节点 2 还是相邻节点.基于温度的反馈,模型预测算法通过动态调节频率使每个节点的温度都回归到峰值温度下:节点 2 的频率上升,用来容纳功耗更大的请求;其他节点的频率下降,用来降低系统峰值温度,而总的吞吐率就不会下降太多.而在传统反馈温控算法中,所有节点的频率都是同时下降到相同的大小,因此,在浏览模式下,两者吞吐率的差异最大.当所有节点的负载模式消耗的功率在第 40 个周期下降到原来的 0.8 倍时,传统反馈温控和模型预测温控都提高所有节点的频率来适应增加的负载量,在该模式下,所有节点的利用率均达到最高,两种算法的吞吐率差异很小.

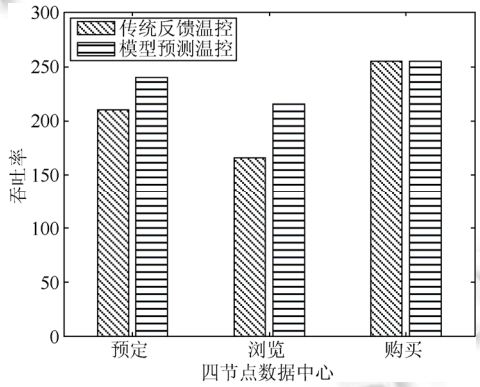


Fig.8 Throughput comparison between traditional feedback and MPC-based temperature controllers when workload types change

图 8 负载类型变化时传统反馈温控和模型预测温控的吞吐率比较

3.6 冷却功耗比较

在本节中我们比较模型预测温控、安全热传递最小温控、负载均衡这 3 种算法的功耗效率.负载均衡是指节点不进行 DVFS 频率调节,将负载平均分布到各个节点上.这个实验基于 50 个节点的数据中心,参考温度为 25°C.从图 9 中可以看出,在相同的负载下,不同调度算法下的冷却功耗大小,其中,模型预测温控可以获得最小的冷却功耗.从图 9 中可以看出,负载越轻,模型预测温控优势越大,因为此时调度优化的裕度很大;当负载较高时,各节点负载都很高,调度优化的余地很小.

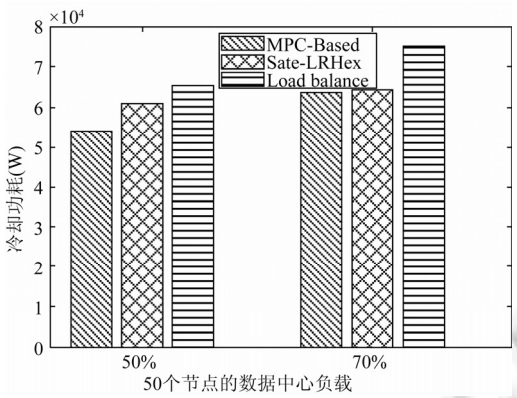


Fig. 9 Cooling power comparison among three methods

图 9 3 种方法下的冷却能耗比较

4 讨 论

节点频率的调整确实会影响某些强耦合应用,造成该应用在频率下降的某些节点上 CPU 利用率过高,进而任务的运行时间增大,这样会造成一定比例的能耗增加,降低 MPC 的控制效果.我们的解决办法是通过一定的负载均衡方式,将由于频率降低而引起过载的节点上的负载迁移到轻载节点上,进行二次调度来控制其运行时间的增加,降低系统能耗.

节点上的运行时间增加主要是由于 CPU 利用率过大引起的,因此,我们假设为每个节点的 CPU 利用率设定一个最大阈值 $U_{max\text{-}ref}$,如 85%.利用率超过该值则为过载,响应时间较长.同时我们定义节点利用率的最低阈值 $U_{min\text{-}ref}$,如 40%,利用率小于该值则节点轻载,能耗浪费较多.因此,我们的二次调度的思路是将集群内的节点分为 3 种类型:过载、轻载和适中.负载均衡是将过载的节点中的负载迁移到轻载节点上,这样整个执行时间就可以降低;适中节点则不进行负载迁移.这样处理的原因是,迁移会有一定的能耗成本,因此尽量减少迁移动作.我们的算法如下.

1) 将集群节点按照 $U_{min\text{-}ref}$,$U_{max\text{-}ref}$ 分为过载集合 P,轻载集合 V 和适中集合 T.

2) 对过载集合 P 中的集合按照节点的热效率 R 降序排列.

3) 对轻载节点 V 中的集合按照节点的热效率 R 升序排列.

4) 选择集合 P 中的第 1 个节点 i,在该节点中选择使节点利用率升高最多的任务 k;如果移除该任务后节点利用率低于 $U_{max\text{-}ref}$,则将该节点从 P 移动到 T;否则继续保留.

5) 选择集合 V 中的第 1 个节点,装入该任务 k.如果装入后,节点 V 的利用率高于 $U_{min\text{-}ref}$,则将该节点从 V 移动到 T;否则保留.

6) 返回步骤 4),直至 P 为空.

将 P,V 集合分别按降序排列和升序排列的目的是将热效率较高的节点上的任务迁移到热效率较低的节点上,这样对整个集群的峰值输入温度影响较小.

我们对第 3.6 节中的 50 个节点,负载为 50%的数据中心按照 MPC 温控算法调度后的场景进行二次负载均衡调度,调度算法如上所示,调度结果如图 10 所示.从图 10 可以看出,经过 MPC 温控后进行了节点频率调整,确实造成了一定的时延和节点计算能耗的增加,而经过我们的负载均衡算法后,时延和计算能耗都有接近 8%的降低.图中虚线所示的运行时间和计算能耗是经过归一化后的结果,归一化的参考是没有经过 DVFS 频率调节的集群的运行时间和能耗.

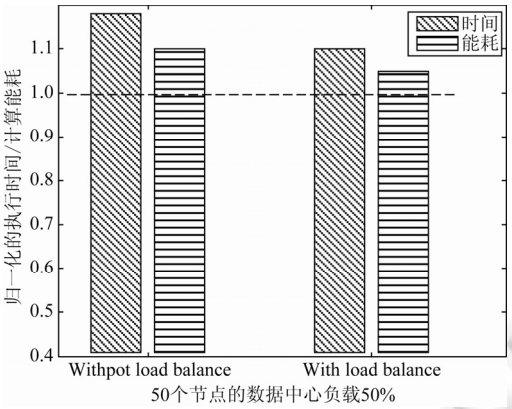


Fig.10 Execution time and computing energy comparison between load balance and no load balance after MPC-based temperature controller

图 10 MPC 温控后采用负载均衡和不采用负载均衡的执行时间与计算能耗对比

5 结束语

本文提出的模型预测温控算法是一种数据中心能耗管理应用技术.其主要应用范围为传统的风冷型数据中心,应用的场景为面向 Internet 的电子商务类型的事务型场景.但该算法对目前日益流行的液冷、风冷和液冷相结合的综合冷却技术以及高性能计算等应用场景具有一定的局限性,今后的工作可以向该方向延伸,希望能够在更广阔的场景下设计更高效的能耗管理技术.

将该方法应用于网络设备和存储设备的温度调节存在一定的难度.对网络设备和存储设备来说,没有频率可以控制,因此,这里可以将控制变量频率转为设备利用率.于是,存储设备和网络设备的能耗模型就可以与这些设备的利用率建立线性关系.这样,理论上也可以采用模型预测算法来调节各存储节点或网络节点的利用率,以此对这些设备的温度进行控制.但真正的难点在于,对存储节点来说,设备利用率的控制很难通过负载的调度来实现.因为存储设备的负载与其设备上的文件块放置策略和访问频度有关,这些难以通过任务调度来实现.对网络节点来说,其利用率的调节涉及到数据包的路由策略的选择,同样难以调度.但研究数据中心内所有设备的能耗管理,并对它们统一进行温度控制无疑是最完全的解决方案,也是我们下一步工作的重点.

References:

[1] Mukherjee T, Banerjee A, Varsamopoulos G, Gupta SKS. Spatio-Temporal thermal-aware job scheduling to minimize energy consumption in virtualized heterogeneous data centers. Computer Networks, 2009,53(17):2888−2904. [doi: 10.1016/j.comnet.2009.06.008]

[2] Reducing energy consumption and cost in the data center. 2014. http://www.datacenterknowledge.com/archives/2014/12/11/reducing-energy-consumption-cost-data-center/

[3] In the data center, power and cooling costs more than the IT equipment it supports. 2007. http://www.electronics-cooling.com/2007/02/in-the-data-center-power-and-cooling-costs-more-than-the-it-equipment-it-supports/

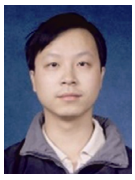
[4] Li S, Le H, Pham N, Heo J, Abdelzaher T. Joint optimization of computing and cooling energy: Analytic model and machine room case study. In: Proc. of the Int'l Conf. on Distributed Computing Systems (ICDCS 2012). IEEE, 2012. 396−405. [doi: 10.1109/ICDCS.2012.64]

[5] Parolini L, Sinopoli B, Krogh BH, Wang ZK. A cyber-physical-system approach to data center modeling and control for energy efficiency. In: Proc. of the IEEE, 2011,100(1):254−268. [doi: 10.1109/JPROC.2011.2161244]

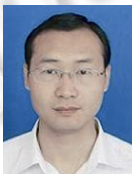
[6] Breen TJ, Walsh EJ, Punch J, Shan AJ, Bash CE. From chip to cooling tower data center modeling: Part I—Influence of server inlet temperature and temperature rise across cabinet. In: Proc. of the 12th IEEE Intersociety Conf. on Thermal and Thermo Mechanical Phenomena in Electronic Systems (ITherm). IEEE, 2010. 1−10. [doi: 10.1109/ITHERM.2010.5501421]

[7] Chen, Y, Gmach, D, Hyser C, Wang ZK, Bash CE, Hoover C, Singhal S. Integrated Management of application performance, power and cooling in data centers. In: Proc. of the Network Operations and Management Symp. (NOMS). IEEE, 2010. 615−622. [doi: 10.1109/NOMS.2010.5488433]

[8] Intel Corporation. Dual-Core Intel® Xeon® processor LV and ULV. Datasheet, 2006.
[9] Ranganathan P, Leech P, Irwin D, Chase J. Ensemble-Level power management for dense blade servers. In: Proc. of the 33rd Annual Int'l Symp. on Computer Architecture (ISCA 2006). Washington: IEEE Computer Society, 2006. 66−77. [doi: 10.1109/ISCA.2006.20]
[10] Brunschwiler T, Smith B, Ruetsche E, Michel B. Toward zero emission data centers through direct reuse of thermal energy. IBM Journal of Research and Development, 2009,53(3):1−13. [doi: 10.1147/JRD.2009.5429024]
[11] GoiriÍ, Katsak W, Le K, Nguyen TD, Bianchini R. Parasol and green switch: managing datacenters powered by renewable energy. In: Proc. of the ASPLOS 2013. New York: ACM, 2013. 51−64. [doi: 10.1145/2451116.2451123]
[12] Wang D, Ren CG, Sivasubramaniam A. Virtualizing power distribution in datacenters. In: Proc. of the ISCA 2013. New York: ACM, 2013. 595−606. [doi: 10.1145/2485922.2485973]
[13] Tsafrir D, Etsion Y, Feitelson DG. Backfilling using system generated predictions rather than user runtime estimates. IEEE Trans. on Parallel and Distributed Systems, 2007,18(6):789−803. [doi: 10.1109/TPDS.2007.70606]
[14] Manzanares A, Qin X, Ruan XJ, Yin S. PRE-BUD: Prefetching for energy-efficient parallel I/O systems with buffer disks. ACM Trans. on Storage, 2011,7(1):Article 3. [doi: 10.1145/1970343.1970346]
[15] Chen T, Yang Y, Zhang HG, Kim H, Horneman K. Network energy saving technologies for green wireless access networks. IEEE Wireless Communications, 2011,18(5):30−38. [doi: 10.1109/MWC.2011.6056690]
[16] Uptime Institute: The average PUE is 1.8. 2011. http://www.datacenterknowledge.com/archives/2011/05/10/uptime-institute-the-average-pue-is-1-8/
[17] Data centers move to cut water waste. 2009. http://www.datacenterknowledge.com/archives/2009/04/09/data-centers-move-to-cut-water-waste/
[18] Ramos L, Bianchini R. C-Oracle: Predictive thermal management for data centers. In: Proc. of IEEE the 14th Int'l Symp. on High Performance Computer Architecture (HPCA 2008). IEEE, 2008. 111−122. [doi: 10.1109/HPCA.2008.4658632]
[19] Li L, Liang CJM, Liu J. ThermoCast: A cyber-physical forecasting model for data centers. In: Proc. of the KDD 2011. New York: ACM, 2011. 1370−1378. [doi: 10.1145/2020408.2020611]
[20] Choi J, Kim Y, Sivasubranmaniam A. Modeling and managing thermal profiles of rack-mounted servers with *ThermoStat*. In: Proc. of the HPCA 2007. IEEE, 2007. 205−215. [doi: 10.1109/HPCA.2007.346198]
[21] Wang XR, Chen M. Cluster-Level feedback power control for performance optimization. In: Proc. of IEEE the 14th Int'l Symp. on High Performance Computer Architecture. IEEE, 2008. 101−110. [doi: 10.1109/HPCA.2008.4658631]
[22] Tang QH, Gupta SKS, Varsamopoulos G. Energy-Efficient thermal aware task scheduling for homogeneous high-performance computing data centers: A cyber-physical approach. IEEE Trans. on Parallel and Distributed Systems, 2008,19(11):1458−1472. [doi: 10.1109/TPDS.2008.111]
[23] Wang XR, Wang YF. Coordinating power control and performance management for virtualized server clusters. IEEE Trans on Parallel and Distributed Systems, 2011,22(2):245−259. [doi: 10.1109/TPDS.2010.91]
[24] Moore J, Chase J, Ranganathan P, Sharma R. Making scheduling "cool": Temperature-Aware resource assignment in data centers. In: Proc. of the 2005 Usenix Annual Technical Conf. Usenix, 2005. 61−75.
[25] Raghavendra R, Ranganathan P, Talwar V, Wang ZH, Zhu XY. No "power" struggles: Coordinated multi-level power management for the data center. In: Proc. of the 13th Int'l Conf. on Architectural Support for Programming Languages and Operating Systems (ASPLOS XIII). New York: ACM, 2008. 48−59. [doi: 10.1145/1346281.1346289]
[26] Abbasi Z, Varsamopoulos G, Gupta SKS. Thermal aware server provisioning and workload distribution for Internet data centers. In: Proc. of the 19th ACM Int'l Symp. on High Performance Distributed Computing (HPDC 2010). New York: ACM, 2010. 130−141. [doi: 10.1145/1851476.1851493]

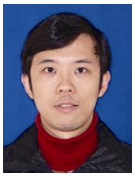


赵小刚(1979－),男,湖北咸宁人,博士,讲师,CCF 专业会员,主要研究领域为云计算,海量存储系统管理,系统能耗控制.



丁玲(1979－),男,讲师,主要研究领域为无线传感器网络,车联网.

胡启平(1963－),男,博士,教授,博士生导师,主要研究领域为软件工程,地理信息系统.



沈志东(1975－),男,博士,副教授,CCF 专业会员,主要研究领域为网络安全.