

面向业务负载的高效能云数据中心系统设计

黄庆¹, 罗志敏², 李杰³

(1. 广州市数字政府运营中心, 广东广州 511348; 2. 中国移动通信集团广东有限公司广州分公司, 广东广州 510000; 3. 暨南大学信息科学技术学院, 广东广州 510632)

摘要: 云数据中心不合理的业务负载调度和资源分配不仅会导致硬件资源利用率低, 而且会引起 IT 设备的局部热点, 这会降低 IT 设备的可靠性, 同时增加云数据中心制冷系统的能量消耗。本文设计了一种面向业务负载的高效能云数据中心系统。该系统统筹管理 IT 设备、制冷设备及供配电设备, 构建云数据中心的能耗模型, 通过合理的业务负载调度和资源管理, 在有限的云数据中心资源下, 提高 IT 设备的资源利用率并降低云数据中心的能量消耗。

关键词: 业务负载; 云数据中心; 资源调度; 能耗优化

本文著录格式: 黄庆, 罗志敏, 李杰. 面向业务负载的高效能云数据中心系统设计[J]. 新一代信息技术, 2023, 6(15): 32-38

中图分类号: TP39

文献标识码: A

High-Efficiency Cloud Data Center System Design for Business Workloads

HUANG Qing¹, LUO Zhi-min², LI Jie³

(1. Guangzhou Digital Government Operation Center, Guangzhou, Guangdong 511348, China; 2. China Mobile Communication Group Guangdong Co., Ltd. Guangzhou Branch, Guangzhou, Guangdong 510000, China; 3. Information Science Technology College, Jinan University, Guangzhou, Guangdong 510632, China)

Abstract: Unreasonable business workload scheduling and resource allocation in cloud data centers not only lead to low hardware resource utilization but also cause local hotspots in IT equipment, which can reduce the reliability of IT equipment and increase the energy consumption of the cooling system in cloud data centers. In this paper, we design a high-efficiency cloud data center system for business workload. The system integrates the management of IT equipment, cooling equipment, and power supply and distribution equipment, builds a power consumption model for cloud data centers, improves the resource utilization of IT equipment, and reduces the energy consumption of cloud data centers under the limited cloud data center resources through reasonable business workload scheduling and resource management.

Key words: business workloads; cloud data centers; resource scheduling; optimizing energy consumption

Citation: HUANG Qing, LUO Zhi-min, LI Jie. High-Efficiency Cloud Data Center System Design for Business Workloads[J]. New Generation of Information Technology, 2023, 6(15): 32-38

1 引言

随着云计算和大数据技术的发展, 为应对数据量爆炸式的增长, 数据中心的数量和规模在不断地扩大, 这种扩大也引起了数据中心巨大的能量消耗^[1]。据报道, 全球数据中心的耗电量占据总耗电量的1%—3%。《中国数据中心可再生能源报告》公布^[2], 我国数据中心耗电量连续八年增长率超过12%。2018年的耗电量达到1 669亿kWh, 相当于1.6个三峡发电站一年的发电总量。报告预计到2030年, 全国数据中心总

用电量将占据社会总用电量的1.5%—2%。

PUE(Power Usage Effectiveness)是指数据中心总能耗与IT设备能耗的比值, PUE越接近1, 表示数据中心的能耗越低^[3]。目前规划在建的超大型、大型数据中心平均设计PUE分别为1.41和1.48。预计2023年底, 数据中心的PUE力争降到1.3以下^[4]。

数据中心硬件设备主要由数千台相互连接的IT设备、保持设备正常运行的制冷设备、供配电设备及照明设备等组成。统计表明, 典型数据中心的IT设备能耗占总能耗的50%左右, 制冷设备占40%左右, 供

配电设备占7%左右,照明设备约占3%。虽然不同数据中心的能耗占比有一定的差异,但其占比排名基本保持不变^[5-6]。

为了降低数据中心IT设备的能耗,传统方法是调度业务负载到少量IT设备中处理,然后将空闲的IT设备转入低能耗状态,从而降低IT设备的能耗^[7-8]。然而,这种不均匀的业务负载分配会在短时间内迅速提升部分工作服务器的温度,从而产生局部热点。这些热点需要制冷设备排出更多的冷气流来降温以保证其稳定性,从而引起数据中心制冷设备额外的能量消耗。因此,如何正确有效地将业务负载放置在合适的IT设备上,以降低数据中心的总能耗(IT设备能耗和制冷设备的能耗),已成为设计现代数据中心系统最重要的挑战之一^[9-10]。

本文设计了一种面向业务负载的高效能云数据中心系统。该系统统筹管理IT设备、制冷设备及供配电设备,构建云数据中心的能耗模型,通过合理的业务负载调度和资源管理,在有限的云数据中心资源下,提高IT设备的资源利用率并降低云数据中心的能量消耗。

2 系统设计

本文设计的云数据中心系统根据业务负载状态、硬件资源状态、网络带宽消耗、数据传输能源消耗、数据传输延迟等指标,构建出多目标优化的调度机制,以降低云数据中心的能量消耗。

面向业务负载的高效能云数据中心系统包括五个部分,分别是用户业务负载请求、资源管理调度策略、数据中心硬件设备、数据采集平台和系统管理。

系统管理模块是配合控制模块对系统整体进行管控,负责日志管理、统计报表、邮件告警等。数据采集平台通过对数据中心IT设备、制冷设备、供配电设备进行数据采集,构建其数据中心细粒度感知模型,实现对基础能耗的精确感知,并根据能效比、PUE、收益等多维度提供决定应用、能耗调度策略。此外,该平台通过数据中心细粒度感知模型还可以实现资源管理、指标监控、诊断告警、权限管理、服务器能耗转换、计算管理、网络管理等功能。数据中心硬件设备主要包括IT设备、制冷设备、供配电设备以及照明设备等^[5]。IT设备资源支持用户大量数据的存储、业务负载计算处理,实现网络接入和管理。制冷设备确保数据中心内部温度和湿度的稳定。供配电设备为数据中心机房内所需电源提供稳定、可靠的动力。照明设备占据数据中心总能耗约3%,相比传统照明系统方案,智能照明系统方案能够有效节约数据中心运维成本^[6]。数据中心资源管理调度策略主要目的是合理分配业务负载到数据中心IT设备资源,提高IT设备服务器利用率的同时避免热点产生,从而有效节省数据中心总体能耗。用户业务负载请求包括业务负载的数量长度、业务负载的提交率与处理业务负载服务器资源的使用率。

图1描述了面向业务负载高效能云数据中心系统的主要流程。用户提交业务负载部署在虚拟机中。资源调度管理策略负责虚拟机的放置和迁移。此外,它还可以动态调整制冷系统的冷气流以降低冷却成本。该调度决策是基于数据中心能耗模型,包括IT能耗模型、制冷能耗模型以及供配电能耗模型。



图1 面向业务负载的高效能云数据中心系统设计

2.1 系统管理

数据中心的系统管理主要负责日志管理、统计报表、邮件告警。具体功能包括:(1)日志管理:对系统事件日志与业务事件日志查看与查询管理;(2)报表统计:查看系统统计报表;(3)邮件告警:对越权、IP异常、连续登陆等事件进行邮件告警。

系统管理作为整体方案的信息汇总呈现平台,将当前数据中心采集的实时参数、控制决策后参数调整以及故障预警情况展示给运维管理人员。运维人员主要依托系统管理的总控中心大屏展示的信息来了解并分析庞大、复杂的系统和设备的运行情况。由于系统管理对象的复杂性,很多情况下,运维管理人员需要从不同维度同时了解、分析数据中心运行情况,这就需要从不同维度展示运行情况。

2.2 数据采集平台

数据采集的主要作用包括预测 PUE 值,计算热量循环矩阵与构建数据中心能耗模型。PUE 反映数据中心服务器的能量效率。通过采集数据中心 IT 设备运行负载功率、网络传输功率、运行的冷风机的数量、冷却塔数量、水泵数量、冷却器、湿球温度、空气焓值等特征值,利用机器学习中的神经网络算法预测 PUE 的值。相关研究表明这种预测误差仅为 0.4%^[11]。此外,利用采集数据中心各个 IT 设备的初始入口温度、运行入口温度、初始出口温度、运行出口温度、运行功率、空闲功率等能够得出热量分布矩阵,其目的是构建 IT 设备能耗与制冷能耗之间的关系模型^[12]。数据采集主要是对数据中心系统架构包含的设备的数据参数进行实时采集,如供电设备的电压、电流、功率采集,制冷系统空调温度采集,数据中心 IT 设备资源细粒度感知数据、精确能源能耗数据采集等等。通过构建数据中心硬件设备能耗模型实现对数据中心业务负载预测。目前数据中心常见的数据采集通信协议包括 Modbus 协议,IPMI 标准协议,SNMP 协议等^[13]。数据中心设备支持这些协议标准,能够对功率相关的数据进行很好的采集。

收集数据中心设备能耗数据是为云数据中心建立和训练/适配能耗模型的第一步。通过数据中心 IT 设备数据采集,实现对资源的精确管控。IT 设备包括服务器、存储、网络等。通过传感器、监控设备等实时监测数据中心的资源使用情况,包括资源的负载、温度、能耗等指标。例如服务器设备则收集服务器中的关键耗电组件(如 CPU、内存、磁盘)的相关数据。该过程主要通过硬件直接测量或者软件监控测量得出。将监测到的数据采集并进行分析,用以构建服务器设

备能耗模型,并通过设计的智能算法对资源的状态进行预测和评估,提前做出调整和优化。根据资源的状态和需求,通过调度资源管理策略进行资源的动态分配和管理,以保证资源的高效利用和业务负载的顺利执行。

云数据中心为了保证 IT 设备的正常运行,需要提供稳定的供电和有效的制冷。云数据中心通过搭建完善的供配电设备,包括电网供电、新能源供电与不间断电源供电(Uninterruptible Power Supply, UPS)等,以确保服务器等设备在停电或电压波动等情况下能够正常运行。云数据中心通过制冷系统对 IT 设备进行散热,以保持设备的温度在安全范围内。制冷设备通常包括空调、冷却塔、冷却液等,通过控制温度和湿度来保证设备的稳定运行。云数据中心通过监测和管理供电和制冷系统的运行状态,包括电压、温度、能耗等指标,实现对供电和制冷功能的精确管控。通过传感器、监控设备等实时监测和采集数据,用以构建供配电与制冷设备能耗模型,并通过算法进行分析和优化,以提高供电和制冷系统的效率和可靠性。

2.3 数据中心硬件资源能耗模型

数据中心的基础设施总能耗主要包括 IT 设备能耗、制冷设备能耗、供配电设备能耗。由于照明设备能耗占比较小,且为固定常数,因此可忽略不计。其中 IT 设备能耗模型包括服务器能耗与网络能耗模型。制冷能耗模型分为两种类型,分别为风冷型制冷系统能耗模型与水冷型制冷系统能耗模型。供配电设备供电能耗模型包括新能源供电能耗模型、电网能源供电能耗模型与电池 UPS 能源供电能耗模型。

2.3.1 IT 设备能耗模型

IT 设备能耗主要包括服务器能耗与网络设备能耗。具体表示如下:

$$E_{IT} = E_{servers} + E_{net}$$

其中, $E_{servers}$ 表示所有服务器处理业务负载消耗的能量, E_{net} 代表数据中心网络设备的总能耗。

服务器能耗模型:对物理服务器的功能进行建模更常见的方法是将其视为功能组件的组装。相关研究表明^[14],服务器的能耗与 CPU 使用率呈线性相关。其能耗模型表示如下:

$$E_{server}^i = E_{idle}^i + (E_{max}^i - E_{idle}^i) \times u_{CPU}^i$$

$$E_{servers} = \sum_{i=1}^n E_{server}^i$$

其中, E_{idle}^i 和 E_{max}^i 分别表示第 i 台服务器空闲和充分利用时的能耗, u_{CPU}^i 表示第 i 台服务器 CPU 利用率。该模型表示服务器的能耗随着 CPU 利用率从 0%

增加到 100% 的线性增长。

网络能耗模型:数据中心内的各个服务器设备连接以及数据中心之间站点连接构成了网络系统能耗模型。网络全局能耗主要包括 3 个部分,即网络链路、网络接口及网络设备能耗(路由器、交换机等)^[4]。由于网络接口能耗的占比较小,在此可忽略不计。网络全局能耗的具体表示如下:

$$E_{\text{net}} = \sum_{(u,v) \in E} X_{u,v} \times E(u,v) + \sum_{u \in V} Y_u \times E(u)$$

其中, $E(u,v)$ 表示链路 (u,v) 的能耗, $E(u)$ 是网络某一交换器或路由器的能耗, $X_{u,v}$ 表示链路的状态是开启或者关闭, Y_u 代表网络设备的开启或者关闭。

2.3.2 制冷设备能耗模型

制冷系统的设计是为了排出数据中心的热量。数据中心的散热可以通过不同的转换介质进行,从而形成不同的制冷基础设施,包括风冷型制冷系统与液冷型制冷系统^[15]。

风冷型制冷系统在数据中心被广泛使用。由于冷气流可以直接穿过装有 IT 设备的空间,并且易于循环,因此它可以作为一种换热介质冷却来自服务器的热量。具体来说,热量通过回流管返回到机房空调(Computer Room Air Conditioning, CRAC)单元的出风口,最终释放

到周围环境中。CRAC 单元是监控和维持数据中心温度、空气分布和湿度的关键冷却部件。液冷型制冷系统采用冷冻液作为制冷剂,将数据中心内部部件产生的热量传递到设备外部,从而为 IT 设备降温。液体制冷比风冷制冷提供更好的导热性,并且在处理日益增长的数据中心密度时更有效。虽然液体制冷系统因其更高的制冷效率和计算密度而越来越多地用于现代大型数据中心,但它也存在巨大投资成本、液体泄漏等不足^[15]。目前风冷型制冷系统仍然普遍应用在工业界。为此,本节构建基于风冷型制冷系统的能耗模型。

制冷系统工作所产生的功率消耗由其热负荷与性能系数决定。其中热负荷由 IT 设备处理业务负载所产生。制冷能耗与制冷性能系数(Coefficient of Performance, CoP)以及 IT 设备能耗之间的关系如下:

$$E_{\text{AC}} = \frac{E_{\text{IT}}}{\text{CoP}}$$

图 2 表示服务器设备能耗与制冷设备能耗之间的关系。数据中心的制冷系统性能系数 CoP 根据不同类型的数据中心而设置不同的值。这取决于数据中心的物理布局 and 热力学特性。来自惠普实验室数据中心的数据表明^[9-10,12],其性能系数与供应温度之间的关系表示如下:

$$\text{CoP} = 0.0068T_{\text{sup}}^2 + 0.0008T_{\text{sup}} + 0.458$$

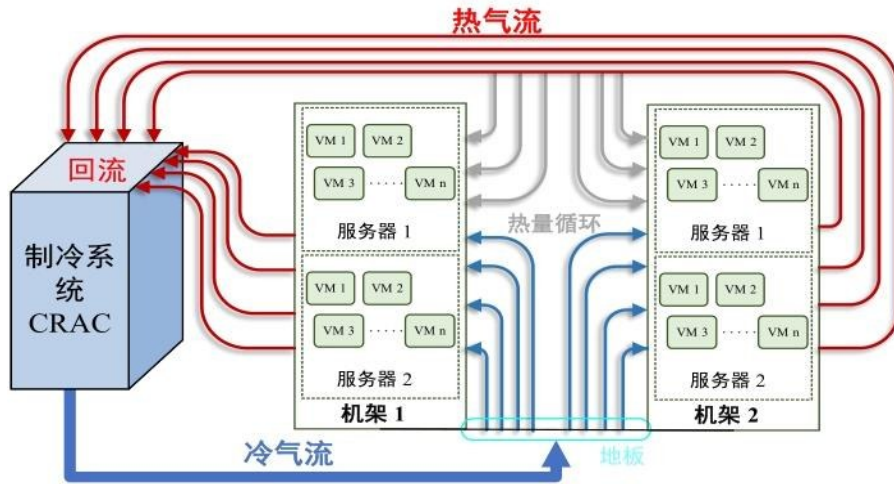


图 2 基于风冷型数据中心的热量循环示意图

其中, T_{sup} 表示制冷系统的供应温度。制冷系统的性能系数随着供应温度的增加而提高,而制冷性能系数的提高,导致制冷能耗会降低。

2.3.3 供配电设备能耗模型

供配电设备能耗一般采用多路能源供电模型。多路能源供电主要包括电力市场和新能源供电技术。目前的新能源供电方式,除了太阳能发电,风能发电

之外,还包括地热发电与生物发电等。这些发电形式的推广,不仅能有效缓解能源短缺问题,还能在一定程度上保护日益恶化的生态环境。智能化供配电系统能够根据最优的用电成本选择使用不同类型的供电方式(新能源供电,电池 UPS 能源供电、电网能源供电)。多路能源供电能耗模型如图 3 所示。

从图 3 中可以得知,新能源供应为 $r(\tau)$,电力市场的供应为 $g_{\text{bef}}(t)/T + g_{\text{rt}}(\tau)$,其中, T 为细粒度感知时间,

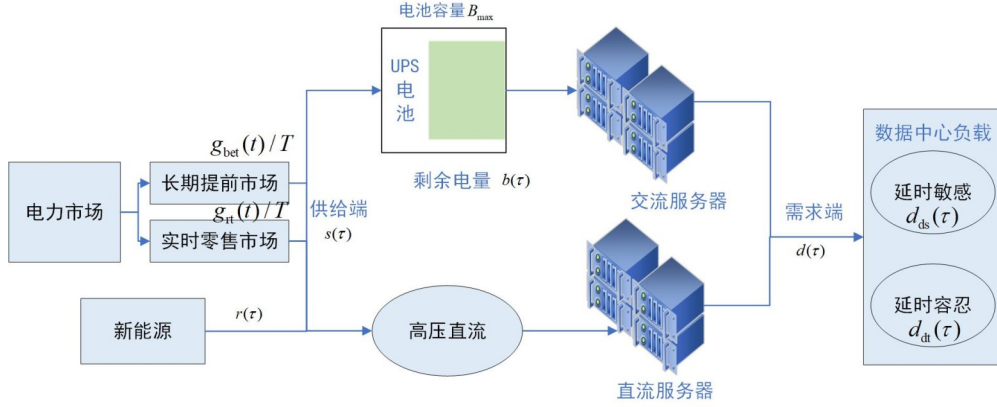


图3 多路能源供配电能耗模型

数据中心从新能源和电网中获得的能源供应为 $s(\tau)$ ，具体计算如下：

$$s(\tau) = g_{bet}(t)/T + g_{rt}(\tau) + r(\tau)$$

在多路能源供应端，一部分能源通过高压直流端传输给需求端，没有使用的电源需求端将被存储到储能设备UPS中。对于供配电需求端来说，数据中心负载需求分为需要立即获取服务的延时敏感需求 $d_{ds}(\tau)$ 与允许延时完成服务的延时容忍需求 $d_{dt}(\tau)$ 。延时容忍型需求进入调度器后会被缓存到服务队列 $Q(\tau)$ 中，当调度器需求能源供应为 $d_{dt}(\tau)$ 时，系统将产生的业务负载总量为 $\gamma Q(\tau)$ ，其中 $\gamma \in [0, 1]$ ，因此当细粒度时刻为 T 服务队列产生的业务量为

$$Q(\tau+1) = \max\{Q(\tau) - \gamma Q(\tau), 0\} + d_{dt}(\tau)$$

多路能源供应端与需求端需求决定了储能设备UPS的能源存储，当UPS处于充电时表示为 b_{rc} ，处于放电状态为 b_{dc} 。当供给量大于需求量时，多余的能源将被存储到UPS中，即

$$b_{rc}(\tau) = s(\tau) - d_{ds}(\tau) - \gamma Q(\tau)$$

当需求量大于供给量时，UPS会放电以满足需求端需要。

$$b_{dc}(\tau) = d_{ds}(\tau) + \gamma Q(\tau) - s(\tau)$$

为了避免过多的充电放电操作而影响UPS的寿命，设定在一段时间内，UPS电池允许的最大充电放电次数为 N_{max} ，则满足：

$$0 \leq \sum_{\tau=0}^{t-1} n(\tau) \leq N_{max}$$

其中，如果 $b_{rc}(\tau) > 0$ 或者 $b_{dc}(\tau) > 0$ ，那么 $n(\tau) = 1$ ，否则 $n(\tau) = 0$ 。 $n(\tau)$ 表示某一时刻UPS充放电累计次数。每一次充放电的能耗为 C_b ，则某一时刻充放电的能耗为 $n(\tau)C_b$ 。

如果多路能源端产出的电量充足，需求端需要的业务负载需求端电量很少，在运行一段时间内，会产生浪费，线路的损耗功率与额外的浪费功率开销总和

$D(\tau)$ 可以定义为供应的能源总量减去负载消耗的能源总量再减去储能电池能允许存储的最大能量量，具体表示如下：

$$D(\tau) = b(\tau) + s(\tau) - d_{ds}(\tau) - \gamma Q(\tau) - B_{max}$$

其中， B_{max} 表示电池最大储能容量。在每个细粒度的时隙内，数据中心供电设备功率消耗 E_{DS} 为电网电能购买开销、UPS操作开销和浪费的能源成本的总和。其中，新能源的主要开销是基础设施部署费用，比如太阳能电板或者风涡轮部署带来的费用，而其操作运营开销较小，可将其忽略不计。因此，供配电能耗模型为

$$E_{DS} = s(\tau) + n(\tau)C_b + D(\tau)$$

其中， $s(\tau)$ 表示为数据中心从新能源和电网中获得的能量。 $n(\tau)C_b$ 表示为电池UPS充放电能量开销。 $D(\tau)$ 为线路损耗能量与额外的能量开销总和。

2.3.4 数据中心总能耗模型

已知数据中心总能耗主要包括IT设备能耗、制冷设备能耗以及供配电设备能耗：

$$E_{total} = E_{IT} + E_{AC} + E_{DS}$$

其中， E_{total} 表示数据中心总能耗； E_{IT} 表示IT设备能耗； E_{AC} 表示制冷设备能耗； E_{DS} 表示供配电设备能耗。通过构建数据中心总能耗模型，提出最优的业务负载调度策略，实现优化数据中心总能耗最小的目的。

2.4 用户业务负载请求

用户的请求被视为处理业务负载，则执行这些业务负载需要相同数量的虚拟机。考虑业务负载直接部署在虚拟机中，且每个服务器在任何给定的时间托管一个或多个虚拟机。虚拟机的数量表示为 $VM = \{VM_1, VM_2, \dots, VM_n\}$ ，其中 N 是提交的业务负载的总数。根据虚拟机对业务负载进行建模，假设每个虚拟机执行单个业务负载，并且在执行完业务负载后终止虚拟机的进程。每个业务负载具有CPU需求 R_{CPU} 、内

存需求 R_{mem} 和业务负载长度 l 。因此,每个业务负载都具有三重属性 $\{R_{\text{CPU}}, R_{\text{mem}}, l\}$ 。在实验的开始阶段虚拟机业务负载会按照合适的调度策略,每隔一段时间进行虚拟机或者业务负载的调度。

2.5 资源调度管理策略

对数据中心性能指标的衡量,需要从数据中心关键设备参数着手,本文考虑的主要性能指标包括数据中心 IT 设备的利用效率 PUE、基础设施设备 (IT 设备、制冷设备以及供配电设备) 能耗、服务器的使用率、网络传输性能 (如时延、丢包率、吞吐率等)。这些关键设备的性能指标对数据中心进行整体综合评价,使得

面向业务负载的高效能云数据中心系统设计和评估有了可量化的依据。

在云计算环境中,业务负载部署到虚拟机中,虚拟机通过调度策略分配到服务器中运行。制冷系统 CRAC 的目的是冷却运行的服务器所产生的过高温度。调度策略主要包括虚拟机的初始化部署、动态虚拟机的迁移及动态制冷系统调节。资源调度策略通过构建数据中心能耗模型作为优化目标实现虚拟机到服务器的分配。数据中心总能耗模型包括服务器能耗模型、网络能耗模型、制冷能耗模型以及供配电能耗模型。图 4 显示了调度策略与不同设备能耗模型的关系。

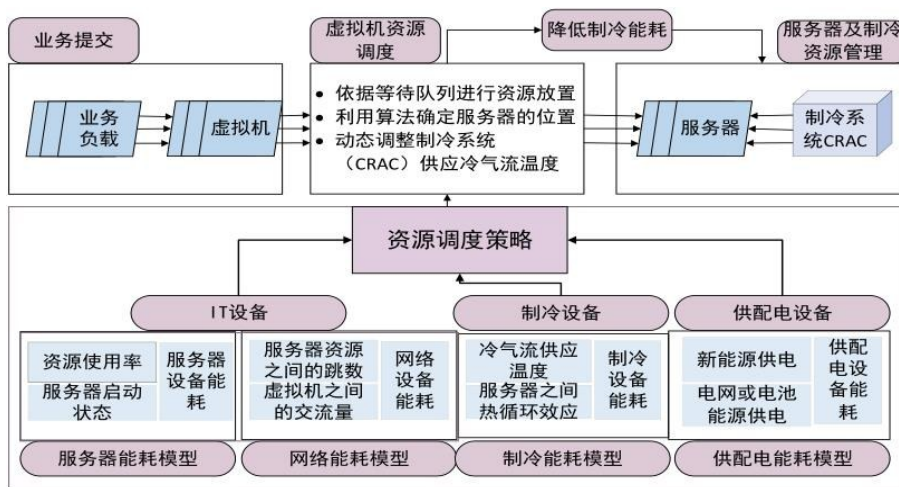


图 4 数据中心内部资源调度策略

在目前对服务器能耗的优化上,主要考虑业务负载在服务器中的整合问题,反而会引起服务器局部热点产生,从而导致数据中心制冷能耗过高的问题。在考虑优化数据中心能耗问题时,需要考虑服务器之间的热量循环效应影响,以及调度策略分配业务负载到服务器资源需要满足服务器入口最低温度约束。然而,相关研究表明^[10],服务器入口温度并不能反映服务器真实的温度,在入口气流温度不变的情况下,服务器处理的业务负载量取决于服务器中的芯片温度。因此,服务器能耗模型需要采用芯片级温度去修正。与目前普遍采用的服务器入口气流感知负载调度策略相比,基于芯片温度的能耗传导业务负载调度策略可以有效降低服务器之间的热量循环影响,提高数据中心的制冷系统效率,实现数据中心节能目的。

网络能耗占据总数据中心总能耗的 10% 左右^[16],可见,网络能耗的优化也会影响数据中心总能耗的变化。考虑优化网络能耗,最普遍的做法是优化虚拟机业务负载的放置,这虽然可以有效取得网络能耗的优化,但综合考虑服务器能耗与制冷能耗,并不能有效

取得数据中心全局优化。针对单一的网络目标优化,可以设计实现能耗最低的数据中心网络拓扑结构,满足网络能耗最低目标的网络路由算法。针对数据中心全局能耗优化,即同时考虑服务器能耗、网络能耗与制冷能耗,可以构建数据中心全局能耗模型,采用全局算法实现虚拟机业务负载的有效放置。具体实施步骤:首先构建数据中心全局制冷能耗模型需要考虑数据中心内的热量循环效应,设计出虚拟机业务负载放置算法从而提高制冷系统的能量效率。然后利用设计的虚拟机业务负载放置算法控制激活服务器的数量,目的是降低服务器能耗。最后分配具有更多流量请求的虚拟机到临近的且运行效率更高的服务器中,从而降低网络能耗。即通过设计出高效的虚拟机业务负载放置算法,实现降低数据中心全局能耗的目的。

对于多路能源供配电能耗节省提出一种策略模型,可以智能选择使用不同类型的供电方式 (新能源供电、电池能源供电或电网能源供电) 进行数据中心设备电力供应。该策略模型的目的是最大限度降低

供电系统总成本,并考虑到了业务负载特性和电池寿命的限制。此外,智能化供电系统还要考虑供电电网故障期间管理数据中心的工作负载和能源及制冷供应。

3 结论

面向业务负载的高效能云数据中心系统设计是用于实现业务负载均衡并高效管理数据中心硬件设备资源的总体性方案,此方案根据不同业务类型调度特性动态调节不同类型的资源调度策略。首先构建数据中心的系统设备(IT设备、制冷设备,供配电设备)能耗模型,然后设计业务负载调度策略(数据中心内部调度,多数据中心之间业务负载均衡调度),并给出不同的策略优化机制、约束条件及求解算法,实现数据中心硬件设备资源有效管理,从而提高业务负载在数据中心硬件资源分配中的均衡,降低数据中心的能耗。

参考文献

- [1] 汪昊. 能耗感知的数据中心负载均衡策略的研究与实现[D]. 南京: 东南大学, 2018.
- [2] 中国数据中心. 中国数据中心可再生能源应用发展报告[R/OL]. (2020)[2023]. <https://www.tudouwenku.com/doc/455636.html>.
- [3] 王继业, 周碧玉, 张法, 等. 数据中心能耗模型及能效算法综述[J]. 计算机研究与发展, 2019, 56(8): 1587-1603.
- [4] 工业和信息化部. 工业和信息化部关于印发《新型数据中心发展三年行动计划(2021-2023 年)》的通知[EB/OL]. (2021)[2023]. https://www.gov.cn/zhengce/zhengceku/2021-07/14/content_5624964.htm.
- [5] DAYARATHNA M, WEN Y G, FAN R. Data center energy consumption modeling: A survey[J]. IEEE Communications Surveys & Tutorials, 2015, 18(1): 732-794.
- [6] LIN W W, SHI F, WU W T, et al. A taxonomy and survey of power models and power modeling for cloud servers[J]. ACM Computing Surveys, 2021, 53(5): 1-41.
- [7] 阮顺领, 卢才武. 虚拟化数据中心全局能耗优化调度方法[J]. 计算机工程, 2013, 39(12): 303-307.
- [8] WANG J J, YU J Y, ZHAI R, et al. GMPR: A two-phase heuristic algorithm for virtual machine placement in large-scale cloud data centers[J]. IEEE Systems Journal, 2023, 17(1): 1419-1430.
- [9] FENG H, DENG Y H, ZHOU Y, et al. Towards heat-recirculation-aware virtual machine placement in data centers[J]. IEEE Transactions on Network and Service Management, 2022, 19(1): 256-270.
- [10] LI X, GARRAGHAN P, JIANG X H, et al. Holistic virtual machine scheduling in cloud datacenters towards minimizing total

energy[J]. IEEE Transactions on Parallel and Distributed Systems, 2018, 29(6): 1317-1331.

- [11] GAO J. Machine learning applications for data center optimization[EB/OL]. (2014)[2023]. <https://api.semanticscholar.org/CorpusID:64625439>.
- [12] LI J, DENG Y, ZHOU Y, et al. Towards thermal-aware workload distribution in cloud data centers based on failure models[J]. IEEE Transactions on Computers, 2023, 72(2): 586-599.
- [13] 张徐东. 高效能数据中心能耗数据实时采集系统设计与实现[D]. 北京: 中国科学院大学, 2020.
- [14] FAN X B, WEBER W D, BARROSO L A. Power provisioning for a warehouse-sized computer[C]//Proceedings of the 34th Annual International Symposium on Computer Architecture. New York: ACM, 2007: 13-23.
- [15] ZHANG W W, WEN Y G, WONG Y W, et al. Towards joint optimization over ICT and cooling systems in data centre: A survey [J]. IEEE Communications Surveys & Tutorials, 2016, 18(3): 1596-1616.
- [16] FENG H, DENG Y, LI J. A global-energy-aware virtual machine placement strategy for cloud data centers[J]. Journal of Systems Architecture, 2021, 116: 1020.

作者简介



黄庆(1975—),女,高级工程师,在广州市数字政府运营中心工作。研究方向:云计算、大数据、网络通信、数字政府。



罗志敏(1979—),男,硕士研究生,研究方向:计算机应用技术、图像处理、人工智能。



李杰(1991—),男,博士研究生,研究方向:云计算、数据管理、资源调度等。