分　数：	
评卷人：	

研究生（数据中心技术）课程论文（报告）

题　目：跨地理分布的数据中心能耗优化方法

学　　号　M202073386

姓　　名　邹聪敏

专　　业　电子信息

课程指导教师　施展 童薇

院（系、所）　计算机科学与技术学院

2020 年 12 月 28 日

跨地理分布的数据中心能耗优化方法

邹聪敏 M202073386

华中科技大学计算机学院

摘 要 随着云计算的快速发展，主要的云服务提供商，如亚马逊、谷歌、Facebook、阿里巴巴等，已经部署了越来越多的跨地理分布的数据中心，为客户提供更高的可靠性和服务质量。这种跨地理分布数据中心系统的基本需求是将大量数据从一个数据中心传输到另一个数据中心。这种数据中心间批量数据传输的跨地理分布和大延迟容忍度为云服务提供商提供了优化运营成本的空间。现有的关于数据中心间海量数据传输的研究大多集中在最小化网络带宽成本上。然而，大容量数据传输的能源成本也占数据中心运营成本的很大一部分，这一问题仍有待探讨。本文研究了三篇顶会论文，从多个角度对跨地理分布的数据中心能耗优化方法进行了综述，其中最为重要的三个方向为：最小化时变系统动态下具有保证服务质量（即服务延迟）的跨地理分布数据中心的能量成本、设计并实现一种新的任务调度算法 Flutter 以减少跨地理分布数据中心的大数据处理任务的完成时间和网络开销、系统研究如何路由和调度数据中心间的批量数据传输以最小化跨地理分布数据中心的能源成本。

关键词 跨地理分布；数据中心；能耗优化；任务调度；多角度

Energy consumption optimization methods of Geographically Distributed Data Centers

ZOU Cong-Min

[1)](School of computer science, Huazhong University of science and technology)

Abstract With the rapid development of cloud computing, major cloud service providers, such as Amazon, Google, Facebook, Alibaba, have deployed more and more cross geographical data centers to provide customers with higher reliability and service quality. The basic requirement of this cross geographic data center system is to transfer a large amount of data from one data center to another. The delay tolerance between cloud service providers and data centers is optimized. Most of the existing researches on massive data transmission between data centers focus on minimizing the cost of network bandwidth. However, the energy cost of large capacity data transmission also accounts for a large part of the operation cost of the data center, which remains to be discussed. This paper studies three top papers and summarizes the energy consumption optimization methods of cross geographical distribution data center from multiple perspectives. The most important three directions are: minimizing the energy cost of cross geographical distribution data center with guaranteed quality of service (i.e. service delay) under time-varying system dynamics, designing and implementing a new task scheduling algorithm flutter to reduce cross geographical distribution In order to minimize the energy cost of cross geographic data centers, the completion time and network overhead of big data processing tasks in distributed data centers are studied.

Key words Cross geographic distribution; data center; energy consumption optimization; task scheduling; multi perspective

———————————————

1 引言

在云计算服务的发展过程中，出现了向大型数据中心（数据中心）发展的趋势。这些分布式控制系统的关键作用是提供互联网服务，如音频和视频分发、网络搜索和数据分析。亚马逊、谷歌、Facebook、阿里巴巴等主要云服务提供商（CSP）一直在部署数十个甚至数百个地理分布（geo distributed）数据中心，以提供更好的可靠性和服务质量（QoS）。

每个地理上分布的数据中心由成千上万的服务器、网络设备和冷却系统组成。每个数据中心都会消耗大量的布朗能量来处理用户请求和满足数据中心的其他要求。根据美国能源部的数据，2014 年，美国所有的分布式控制系统消耗了大约 700 亿千瓦时，占美国总发电量的 18%。据预测，到 2030 年，信息和通讯技术（ICT）每年的总能耗将超过全球总用电量的 20%，其中，数据中心能耗将超过 ICT 总能耗的三分之一，如图 1 所示。

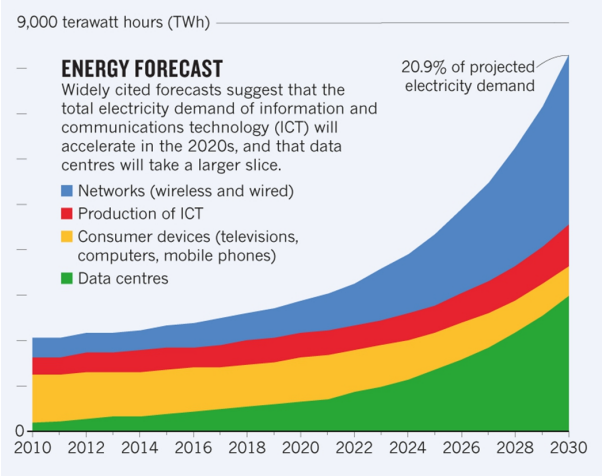

图 1　信息和通讯技术能耗预测

数据中心的功耗主要由 IT 设备和冷却系统组成。数据中心在冷却系统上消耗了相当大一部分的电力，以平衡数据中心内部的温度并提高服务器的效率。冷却系统的功耗通过功率使用效率（PUE）进行总结。PUE 是设施总功率与 IT 设备总负荷之比。IT 设备也消耗了总功率的很大一部分，是衡量数据中心能效的关键指标。因此，为了减少 PUE，许多工程优化技术被开发出来，如先进的冷却机制、虚拟化和数据中心电源转换架构。这些研究工作主要集中在降低数据中心的能耗。像谷歌、微软和苹果这样的大公司正在使用可再生能源（如太阳能电池板和风力涡轮机），并在现场可再生能源发电方面进行直接投资，以减少棕色能源的消耗。然而，由于太阳能和风能的高度间歇性和波动性，完全使用可再生能源为数据中心供电存在挑战。为了克服这些挑战，除了现场可再生能源外，数据中心还使用棕色能源作为主要能源。

学术界和工业界都提出了各种优化技术来解决不同方面的问题，例如降低数据中心的电力成本，同时保证终端用户的服务质量（即服务延迟）。然而，为了降低电力成本，降低电价和节能服务器是很重要的。服务器在能源效率方面的同质性和地理分布数据中心的动态电价是将用户请求转移到数据中心以降低总体能源成本的关键思想。研究表明，使用地理负载平衡（GLB）可以最大限度地利用可再生能源，减少棕色能源的消耗。在 GLB 中，最初用户请求被前端代理接受，然后这些请求被转发到地理上分布的数据中心。因此，服务提供商的关键作用是开发和评估节能解决方案，以动态降低电力成本。

为了处理地理上分布的大量数据，我们传统上需要将所有要处理的数据传输到单个数据中心，以便以集中的方式处理这些数据。然而，有时，这种传统智慧在实践中可能并不可行。首先，由于法律原因或隐私问题，跨国界移动用户数据可能不实际。其次，随着数据量呈指数级增长，跨地理分布的数据中心移动大量数据的成本（考虑到带宽成本和时间）可能变得令人望而却步。

有人指出，将计算任务转移到数据所在的位置，而不是跨数据中心传输数据，以便在同一数据中心内对数据进行本地处理，这可能是一个更好的设计。当然，这种处理后的中间结果可能仍然需要跨数据中心传输，但它们通常体积小得多，大大降低了数据传输的成本。一般来说，基本目标是将任务放在各自最好的数据中心，从而最大限度地缩短大数据处理应用程序中的作业完成时间。然而，以前的工作是在假设的基础上设计的，这些假设通常是不现实的，例如数据中心间链路上不会出现瓶颈。

这种离线优化不可避免地依赖于对任务执行时间和中间结果传输时间的先验知识，如果没有先进的预测算法，这两种方法都不容易获得。即使这些知识是可用的，Spark 中的一个大型数据处理任

务也可能涉及一个包含数百个任务的有向无环图（DAG），而调度这种 DAG 的最优解一般是 NP 完全的。

除了作业完成时间外，不同数据中心之间的数据传输成本也是一些工作负载的一个重要问题，因为成本可能非常高。更具体地说，对于没有网络预算约束的工作负载，作业完成时间是唯一的目标，我们可以直接对其进行优化。而对于具有网络预算约束的工作负载，达到最佳作业完成时间并不一定会导致数据传输的最优成本。这是因为作业完成时间只关心瓶颈环节，而数据传输的成本节约只能通过减少跨数据中心所有链路上的数据传输总量来实现。因此，在这种情况下，我们的目标是以合理的带宽成本优化作业完成时间，并为这些工作负载设计可调的网络预算。

本文调研的第一个方向考虑多个地理分布的分布式控制系统。用户请求到达全局负载均衡器（global-LB），该负载均衡器根据当前电价、可再生能源水平、服务器利用率水平和延迟等多个因素分配给一个唯一的数据中心。目标是使系统的总能量成本最小化。在动态电力市场条件下，以可再生能源的可用性为约束混合整数线性规划问题，建立了数据中心系统总成本最小化的优化问题。然后我们开发了一个贪婪的在线算法来解决优化问题。我们基于实际工作负载轨迹和每小时电价进行了大量的实验评估，以证明我们提出的算法的有效性。

在本文的第二个调研方向中，论文开发并实现了一个新的系统 Flutter，它可以在广域范围内跨数据中心调度任务，同时兼顾时间和成本效益。在设计 Flutter 时，我们主要关注的是实用性和真实世界的实现，而不是我们的结果的最佳性。为了实用，Flutter 首先被设计成一种在线调度算法，根据当前的作业进度进行动态调整。Flutter 还具有阶段意识：它使作业中每个阶段的完成时间最小化，这对应于阶段中组成任务的最慢完成时间。

数据中心间批量数据传输的地理分布和大延迟容忍度为云服务提供商提供了降低运营成本的机会。在空间上，不同位置的数据中心具有不同的运行特性，例如网络带宽和单位电价。因此云服务提供商可以通过为不同的数据中心间批量数据传输分配不同的路由来降低运营成本。在时间上，只要能在截止日期前完成，数据中心间批量数据传输可以在到达后的任何时间启动。因此，云服务提供商可以对数据中心间批量数据传输进行灵活的调度，以降低运营成本。

本文调研的第三个方向研究了多电力市场环境下地理分布数据中心间直流无刷直流输电系统的能量成本最小化问题，并将此优化问题（最小化批量数据传输的能源消耗，MIN EC-BDT）建立在最小成本多商品流模型中。

论文提出了一个两阶段的方法来快速地求解最小化批量数据传输的能源消耗问题的最优解。对于每个批量数据传输，我们的方法在第一阶段沿可用时隙搜索最优需求分配。然后分别计算出最优需求划分中各部分的最优路径和调度。

对实际的直流互联网络和电价的广泛评估表明，我们的两阶段优化方法可以比现有的跨直流批量数据传输方法节省大量的能源成本。计算结果还表明，该方法在地理分布数据中心间的数据中心间批量数据传输节能方面具有较高的计算效率。

2 问题与挑战

为了更好地理解，本文总结了表 1 中的符号及其定义，这些符号将在本文中使用。

表 1 本文要使用到的符号

Notations	Definitions
$t \in \{1, T\}$	Index of time slot
$i \in \{1, N\}$	Index of data center
$W(t)$	Total workload arrived at global load balancer at time t
$w_i(t)$	Workload assigned to DC i at time t
$L_i(t)$	Network delay experienced from global load balancer to DC i
D^{max}	Maximum delay threshold
$d_i(t)$	Average delay at DC i at time t
$d_i^Q(t)$	Queuing delay at DC i at time t
R_i^{max}	Maximum renewable power level at DC i
$R_i(t)$	Renewable power level at DC i at time t
$q_i(t)$	Price of brown energy at DC i at time t
$S_i^{max}(t)$	Maximum number of servers at DC i at time t
$S_i^{ac}(t)$	Total number of active servers at DC i at time t
$S_i^{in}(t)$	Total number of inactive servers at DC i at time t
μ_i	Service rate at DC i
$PUE_i(t)$	Power usage effectiveness factor at DC i at time t
$P_i^{IT}(t)$	Power consumption of IT equipment in DC i at time t
$P_i(t)$	Total power consumption of DC i at time t
$P_i^{ac}(t)$	Power consumption of active servers in DC i at time t
$P_i^{in}(t)$	Power consumption of inactive servers in DC i at time t
$C_i(t)$	Electricity cost of DC i at time t
$C(t)$	Total electricity cost at time t

2.1 问题表述

我们考虑 N 个地理分布的数据中心，每个数据

中心有固定数量的服务器。我们假设所有的服务器都是同质的。每个数据中心 i∈{12}都由棕色能源和可再生能源（如太阳能）供电。在用 t∈{12}表示的每个离散时间中，用户请求 W（t）到达全局负载平衡器。全局负载均衡器（global-loadbalancer，global-LB）的功能是根据最小的功耗和减少工作负载的平均延迟，在线决定 W（t）到适当的数据中心 i，从而使系统的总成本最小。在选择数据中心之后，传入的工作负载到达本地负载平衡器（locallb）。数据中心的系统模型如图 2 所示。每个数据中心根据服务器的当前利用率级别将分配的工作负载分配给相应的服务器。

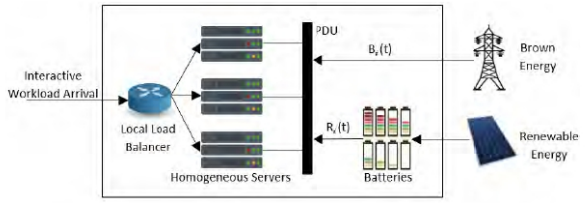

图 2 数据中心的系统模型

2.2 工作量模型

数据中心中的工作负载大致分为两类：对延迟敏感的交互工作负载和延迟容忍的工作负载批处理工作负载。在这项工作中，我们假设工作负载是交互式的，因此由于其延迟敏感特性，因此是不可分割的。

2.3 跨数据中心的带宽

为了激励我们的工作，我们从一个真实世界的实验开始，虚拟机（VMs）在 Amazon EC2 的四个代表性地区发起并分布：欧盟（法兰克福）、美国东部（弗吉尼亚州北部）、美国西部（俄勒冈州）和亚太地区（新加坡）。我们使用的所有 VM 实例都是 m3。xlarge，有四个核心和 15GB 的主内存。为了说明数据中心间链路的可用容量，我们使用 iperf 实用程序测量了数据中心之间虚拟机之间的可用带宽，结果如表 2 所示。

表 2 截至 2015 年 7 月，跨地理分布数据中心的虚拟机之间的可用带宽

	EU	US-East	US-West	Singapore
EU	946 Mbps	136 Mpbs	76.3 Mbps	**49.3 Mbps**
US-East	-	1.01 Gbps	**175 Mbps**	52.6 Mbps
US-West	-	-	945 Mbps	76.9 Mbps
Singapore	-	-	-	945 Mbps

从这张表中，我们可以得出两个有说服力的证据。一方面，当同一数据中心中的虚拟机通过内部数据中心网络相互通信时，可用带宽始终很高，约为 1 Gbps，这对于典型的基于 Spark 的数据并行应用程序来说是足够的。另一方面，跨数据中心的虚拟机之间的带宽要低一个数量级，并且对于虚拟机之间不同的数据中心间链路，带宽差别很大。例如，表中最高带宽为 175Mbps，而最低带宽仅为 49Mbps。

我们的观察清楚地表明，如果我们在不同的数据中心运行相同的数据并行应用程序，那么在作业完成时间方面，跨数据中心的中间结果的传输时间可能很快成为瓶颈。因此，仔细地将任务调度到尽可能好的数据中心对于更好地利用可用的数据中心间带宽非常重要；当数据中心间带宽较低且更分散时，更是如此。Flutter 是首先也是最重要的是设计为网络感知，因为任务可以在地理分布的数据中心之间调度，同时感知可用的数据中心间带宽。

2.4 跨数据中心的带宽成本

除了不同数据中心的带宽差异外，带宽定价也是网络预算约束下工作负载的一个重要因素。让我们先看一下不同数据中心之间数据传输的定价。我们分别在表 3 和表 4 中展示了 Microsoft Azure 和 Amazon Web Service（AWS）的数据传输定价。定价单位均为美元/GB。我们还应该注意到，定价仅针对到其他数据中心的出站流量。入站流量对几乎所有公共云提供商都是免费的，比如 AWS、Azure 和 Google 计算引擎。

表 3 Azure 中跨地理分布数据中心的带宽成本

US and EU	Asia Pacific, Japan and Australia	Brazil
0.087	0.138	0.181

根据这些表格，我们可以清楚地看到两个云平

台的网络定价会有很大的不同。更具体地说，在 Azure 中，每 GB 数据中心间数据传输的最昂贵价格是最便宜价格的两倍以上。此外，最昂贵的定价是 AWS 最便宜定价的 8 倍。这两种情况都强烈暗示，我们也应该避免将任务调度到数据中心，因为这样会导致数据传输的高成本。也就是说，在网络预算受限的情况下，一个明智的调度决策除了要考虑数据中心之间带宽的差异外，还应考虑数据中心间数据传输的定价差异和定价策略。

表 4　跨地理分布数据的亚马逊网络服务中心带宽成本

US and EU	Singapore and Tokyo	Sydney	San Paulo
0.02	0.09	0.14	0.16

为了说明任务/作业完成时间的最佳调度选择不一定会导致数据传输的最低成本，我们在图 3 中提供了一个简单的激励示例。在这个例子中，我们可以看到，如果我们只考虑任务完成时间，那么获得所有输入所需的时间是相同的。然后它将获得相同的任务完成时间，因为无论我们在哪里安排任务，我们都需要从其他数据中心传输其他输入，并且输入的大小是相同的。然而，我们可以很容易地看到，将输入从数据中心 1 传输到数据中心 2 要比另一种方式昂贵得多。因此，如果我们同时考虑到数据中心之间的数据传输成本，那么将任务调度到数据中心 1 显然是一个更好的解决方案。这个动机的例子强烈地暗示了在任务调度中需要考虑网络预算和定价策略，特别是对于具有网络预算约束的工作负载。

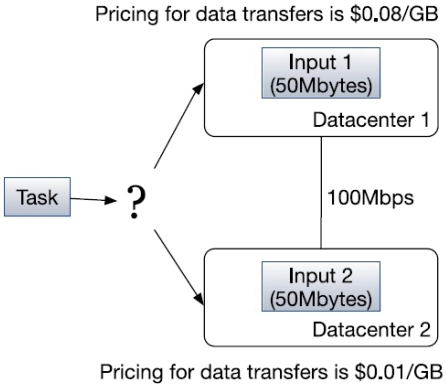

图 3　当同时考虑传输数据的带宽和成本时的任务调度举例

2.5　模型讨论

我们现在讨论上述模型中的假设和一些实际考虑。

以上，我们假设系统在时隙时间内运行，而时隙时间是将复杂连续时间模型转化为离散最优决策过程的一种合理方法。跨数据中心间链路的大容量数据流可以沿多个多跳路径分割和传输，每个多跳路径可以随时间最优地计算。这是一个常见的假设，由提出。

在每个时间间隔进行调度决策时，假设区域电价是时变的。在文献，中，通常会做出这样的假设。一些统计机器学习技术，可以用来实现这种预测。此外，由于电力需求巨大，数据中心在日前市场使用长期合同从电网获取电力，因为长期合同成本低于实时市场电价。这种日前每小时或 15 分钟的价格信息也可以作为或指导电价预测。

我们只从数据流量负载的角度来研究数据中心的功率，而没有考虑其他因素，例如温度。因为业务负载是影响内部数据中心-批量数据传输功率的最重要（我们可以控制）运行特性。虽然我们从数据业务负载的角度关注于数据中心间批量数据传输功率建模，但需要注意的是，这种考虑对所提出算法的本质没有影响。

最后，在本文中，为了突出我们方法的关键点，我们在对跨数据中心批量数据传输进行最优路由和调度决策时只考虑了能量成本。然而，我们的模型和算法可以很容易地扩展，以适应其他类型的操作成本。结果表明，我们的方法也可以应用于其它类型的数据中心间批量数据传输优化。

3　能耗优化问题求解

3.1　MILP 优化技术

在本节中，我们将给出第二节中讨论的优化问题的解决方案。采用混合整数线性规划（mixed integer linear programming，MILP）优化技术求解该问题的离线解。

对于上述中给出的优化问题，有两个主要的求解问题。首先是未来数据的不可用性，比如未来电价和用户请求数量。这意味着我们无法为优化问题提供未来的数据，而优化问题是获得最优解所必需的。其次，即使所述数据是可用的，对于大量的时

隙和数据中心，获得最优解的计算成本也太高。例如，分支定界算法在最坏情况下将遍历 22M 个节点。一个简单的暴力算法将需要 NT 的计算次数，这对于 N 和 T 的较高值是不可行的。因此，在现实中（由于未来数据不可用）和计算（由于计算步骤非常多）都不可能获得最优解。

为了解决上面中的优化问题，我们提出了一种基于贪婪算法设计技术的绿色地理负载平衡（GreenGLB）（算法 1）。我们提出的算法的基本思想是在考虑到当前的电价、可再生能源水平以及尊重给定的一组约束条件的情况下，分配每个时刻的输入工作量。算法 1 描述了该算法的工作原理。

我们的问题设置与其他文献中提出的其他模型有显著不同。我们已经考虑到工作量是不可分割的，此外，我们还整合了可再生能源和棕色能源。据我们所知，我们是第一个提出一个不可分割的工作负载分配方案，沿着 GreenGLB 算法 1 来解决优化问题，如图 4 所示。

Algorithm 1 GreenGLB to Solve the Optimization Problem

1: **for** $t \leftarrow 1$ to T **do**
2: Read $W(t), q_i(t), \forall i \in [1, N]$
3: Compute $S_i^{ac}(t), S_i^{in}(t), P_i^{ac}(t), P_i^{in}(t), d_i(t), R_i(t)$ and $P_i(t)$ to solve the optimization problem
4: $i^* = \underset{arg\ i\in[1,N]}{\min} \sum_{i=1}^{N} [C_i(t) + d_i(t)]$
5: Subject to constraint (7a)- (7e)
6: Assign the incoming workload $W(t)$ to data center i^*
7: Update $S_i^{ac}(t), S_i^{in}(t)$ and $w_i(t) \quad \forall i \in [1, N]$
8: **end for**

图 4 GreenGLB 算法 1 流程图

3.2 基于 spark 的 flutter 算法设计

在我们讨论了如何有效地解决任务调度问题之后，我们现在可以看看如何在 Spark 中实现它，Spark 是一个流行于大数据处理的现代框架。

Spark 是一个快速通用的分布式数据分析框架。与基于磁盘的 Hadoop 不同，Spark 会将中间结果的一部分缓存在内存中。因此，它将大大加快迭代作业的速度，因为它可以直接从主存而不是磁盘获得前一阶段的输出。现在随着 Spark 越来越成熟，为不同应用设计的几个项目都是基于 Spark 构建的，比如 MLlib、Spark Streaming 和 Spark SQL。所有这些项目都依赖于 Spark 的核心模块，包含 Spark 的几个基本功能，包括弹性分布式数据集（RDDs）和调度。

为了将我们的调度算法整合到 Spark 中，我们重写了调度模块来实现我们的算法。从视图顶部看，在 Spark 中提交作业后，该作业将被转换为任务的 DAG，并由 DAG 调度器处理。然后，DAG 调度器将首先检查最终阶段的父阶段是否完成。如果是，则最终阶段将直接提交给任务调度器以进行任务调度。如果没有，则递归提交最后阶段的父阶段，直到 DAG 调度程序找到就绪阶段为止。

我们的实现的详细架构如图 5 所示。从图中我们可以看到，在 DAG 调度程序找到一个就绪阶段之后，它将为该就绪阶段创建一个新的任务集。在这里，如果任务集是一组 reduce 任务，我们将首先从 mapOutputRacker 获取映射任务的输出信息，然后保存到这个任务集。然后，该任务集将被提交到任务计划程序并添加到挂起的任务集列表中。当任务集等待资源时，SchedulerBackend（也是集群管理器）将在集群中提供一些免费资源。在接收到资源之后，Flutter 将在队列中选择一个任务集，并确定应该将哪个任务分配给哪个执行者。它还需要与 TaskSetManager 交互以获取任务的描述，然后将这些任务描述返回给 SchedulerBackend 以启动任务。在整个过程中，获取 map 任务的输出和调度过程是两个关键步骤。

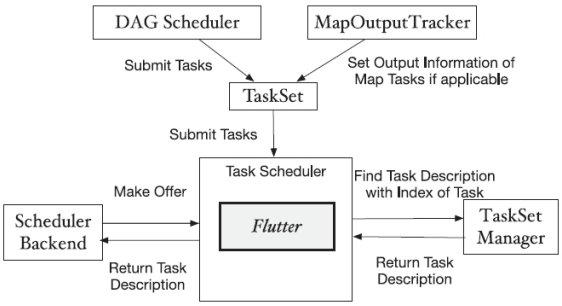

图 5 Spark 中 flutter 算法的设计

3.3 两阶段优化方法

MIN-EC-BDT（最小化海量数据传输的能源消耗）是一个典型的动态网络流优化调度问题。在本节中，我们首先描述和分析基于时间展开的方法用于解决 MIN-EC-BDT 问题。在此基础上，提出了一种求解该问题的两阶段优化方法。

3.4 基于时间展开的方法

解决动态网络上的流问题最直接的方法之一

是将其简化为静态时间扩展网络上的类似问题，。通过在每个时间间隔内引入所有节点的虚拟副本，可以将动态问题转化为时间扩展图中的等效静态问题。

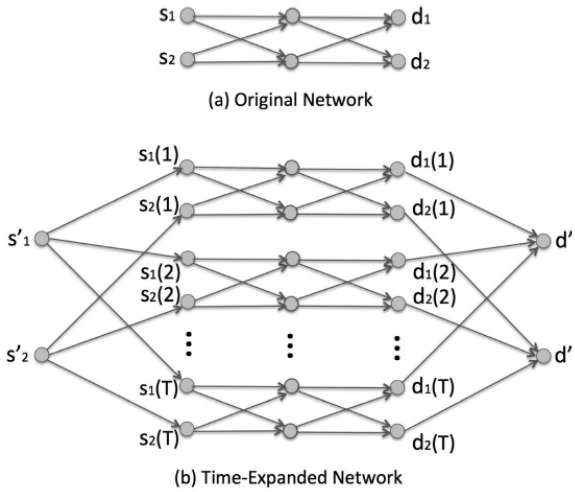

图 6　原始网络和时间扩展网络之间的转换

4 能耗优化方法实验效果

4.1 基准算法

我们将我们提出的算法（GreenGLB）与三种工作负载管理方案进行比较，如下所述。

（1）最小工作负载地理负载平衡（MGLB）：在该方案中，传入的工作负载被路由到地理上分布的数据中心，该数据中心具有最小的现有工作负载。

（2）随机地理负载平衡（RGLB）：在 RGLB 中，传入的工作负载被随机地分配到一个地理分布的数据中心，而不需要任何优化技术。该方案类似于中提出的算法。

（3）循环（RR）：RR 工作负载分配策略为全局负载平衡器设置了相等的权重，其中工作负载在所有地理分布数据中心之间均匀分布。该策略不考虑任何有关可再生能源可用性或动态电价的信息。

4.2　GreenGLB 算法实验结果

在本节中，我们展示了基于实际电价数据和工作负载跟踪的算法的性能。

图 6 显示了四种不同输入工作负载分配方案的数据中心的平均能量成本。由于 RR 没有考虑动态电价和可再生能源分配给特定地理分布数据中心的工作量，因此它产生了最高的能源成本。由于传入的工作负载随机分布到地理上分布的数据中心，RGLB 的成本非常高。与 RGLB 和 RR 相比，MGLB 获得了更低的能源成本。性能可以归因于算法的工作，在这种情况下，所有的工作负载都被路由到一个地理上分布的数据中心，该数据中心具有最小的现有工作负载，而没有进行任何优化。

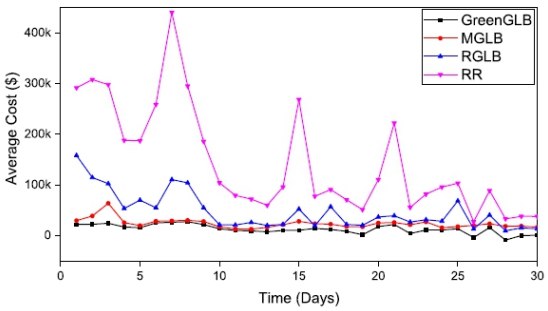

图 6　四种算法的平均每日能量消耗

GreenGLB 的结果是 4 种策略中最好的。GreenGLB 通过采用布朗能源的最低电价、最大可用可再生能源水平和所有地理分布分布式控制系统中的活动服务器总数，动态最小化所有工作负载的总体能源成本。我们观察到，在某些时隙中，总能量消耗为负（见图 7）。能源成本为负的原因是加拿大安大略省的负电价。回想一下，当一个高而不灵活的发电同时出现低电力需求和高可再生能源供应时，就会出现负电价。

在表 5 中，我们比较和总结了 GreenGLB 相对于 MGLB、RGLB 和 RR 的性能改进。条目显示，GreenGLB 的性能比 MGLB 提高了 46%，比 RGLB 提高了 73%，比 MGLB 提高了 83%。因此，GreenGLB 极大地降低了地理分布分布式控制系统的总能源成本。

表 5 GreenGLB 算法的能量消耗改进

Comparison Factor	Improvement of GreenGLB over MGLB	Improvement of GreenGLB over RGLB	Improvement of GreenGLB over RR
Energy Cost	46%	73%	83%

图 7 显示了使用 GreenGLB 来满足传入工作负载需求和延迟约束的跨地理分布的数据中心的平均活动服务器数量。每次 GreenGLB 计算活动服务

器的数量。我们观察到，活动服务器的数量根据传入的工作量、电价和可用的可再生能源水平而变化。

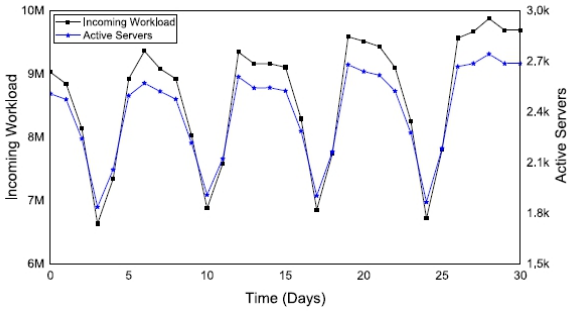

图 7 工作负载分配和活动服务器

图 8 展示了总工作负载的处理时间（网络和队列延迟+服务器处理时间）。回想一下，在我们的实验中，最大处理时间被设置为 11ms。从图中我们可以看到，响应时间总是低于最大限制，这意味着我们提出的算法达到了要求的服务质量。

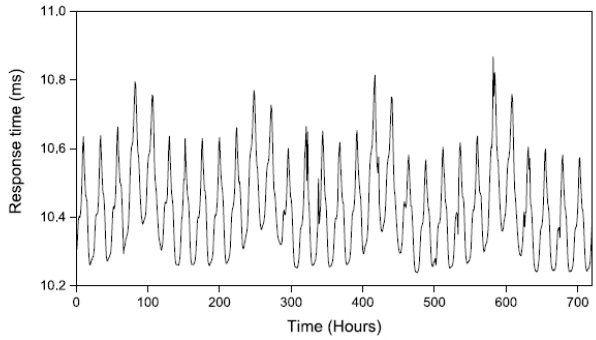

图 8 总工作负载的处理时间

4.3 实验装置

我们首先描述我们在实验中使用的试验台，然后简要介绍整个评估过程中使用的应用程序、基线和度量。

我们的实验在 6 个数据中心进行，共有 25 个实例，其中两个数据中心在多伦多。其他数据中心位于不同的学术机构：维多利亚、卡尔顿、卡尔加里和约克。实验中使用的所有实例都是 m.large，有 4 个内核和 8gb 的主内存。这些区域的虚拟机之间的带宽容量由 iperf 测量，如表 6 所示。安大略省的数据中心通过专用的 1GE 链路相互连接。因此，从表中可以看出，多伦多、卡尔顿和约克的数据中心之间的带宽容量相对较高，但仍低于同一数据中心内的带宽容量。

表 6 跨地域分布的虚拟机之间的数据中心的可用带宽（Mbps）

	Tor-1	Tor-2	Victoria	Carleton	Calgary	York
Tor-1	1000	931	376	822	99.5	677
Tor-2	-	1000	389	935	97.1	672
Victoria	-	-	1000	381	82.5	408
Carleton	-	-	-	1000	93.7	628
Calgary	-	-	-	-	1000	95.6
York	-	-	-	-	-	1000

Note: "Tor" is short for Toronto. Tor-1 and Tor-2 are two data centers located at Toronto.

4.3.1 应用程序设定

我们在 Spark 上部署了三个应用程序。它们是 Wor 数据中心 ount、PageRank 和 GraphX。

Wor 数据中心 ount：Wor 数据中心 ount 计算单个或一批文件中每个单词出现的频率。它将首先计算每个分区中单词的出现频率，然后将前一步的结果进行聚合以获得最终结果。我们选择 Wor 数据中心 ount 是因为它是一个基本的应用程序在分布式数据处理中，可以处理真实世界的数据痕迹，如 Wikipedia 转储。

PageRank：它根据指向网站的链接的数量和质量计算网站的权重。这种方法依赖于这样一个假设：如果有许多其他重要的网站链接到一个网站，那么这个网站是重要的。它是一个典型的具有多次迭代的数据处理应用程序。我们用它来计算网站排名和用户在社交网络中的影响力。

GraphX:GraphX 是基于 Spark 构建的用于并行图处理的模块。我们运行应用程序 livejournalagerank 作为 GraphX 之上的代表性应用程序。即使应用程序也被命名为“PageRank”，计算模块在 GraphX 上也是完全不同的。我们选择它是因为我们也希望评估基于 Spark 的系统的 Flutter。

4.3.2 输入

对于 Wor 数据中心 ount，我们使用 10gb 的维基百科作为输入。对于 PageRank，我们使用 Google 在私有测试平台上发布的 875713 个节点和 5105039 个边的非结构化图，以及 Pokec 在线社交网络发布的 1632803 个节点和 30622564 条边的有向图。对于 GraphX，我们在 LiveJournal 在线社交网络中采用有向图，有 4847571 个节点和 68993773 个边，

其中 LiveJournal 是一个免费的在线社区。

4.4 工作完成时间

我们在图 9 中绘制了三个应用程序的作业完成时间。如我们所见，这三种 Flutter 应用的完成时间都缩短了。更具体地说，Flutter 将 Wor 数据中心 ount 和 PageRank 的作业完成时间分别减少了 22.1%和 25%。GraphX 的完成时间也减少了 20 秒以上。改进主要有两个原因。首先，Flutter 能够自适应地将 reduce 任务调度到一个数据中心，以最少的传输时间获得所有的中间结果。这样就可以尽快开始任务。第二种是 Flutter 会把阶段中的任务作为一个整体来安排。因此，它可以显著地减轻掉队者在该阶段运行缓慢的任务，并进一步提高整体性能。

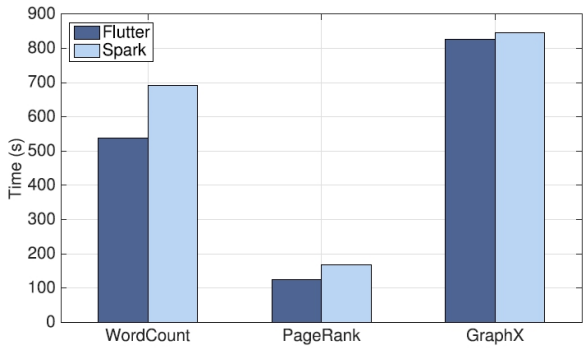

图 9 不同作业的完成时间

在这种情况下，GraphX 上的作业完成时间的改进似乎很小，这是因为有三个彼此具有高带宽的数据中心，并且延迟调度也可以调度这三个数据中心中的任务。尽管 GraphX 应用程序的作业完成时间没有显著减少，但我们将展示 Flutter 将显著减少 GraphX 应用程序在不同数据中心之间传输的通信量。

4.5 两阶段优化方法评估

在这一节中，我们评估了所提出的两阶段优化方法，该方法适用于实际的内部直流网络和电价的最小化批量数据传输的能源消耗。

4.5.1 评估设置

为了描述数据中心间批量数据传输的空间和时间灵活性带来的能源成本节约效益，我们对一个大型数据中心间网络进行了模拟，该网络由 11 个地理分布数据中心组成，具有真实的网络拓扑结构，如图 10 所示。 所有地理分布的数据中心都由同一个云服务提供商运营。任何两个相邻数据中心之间的主干链路是双向的，每个主干链路的带宽是不同的（从 1 到 3 Gbps 不等），这取决于所使用的网络服务提供商。这些数据中心位于不同的区域电力市场，如加利福尼亚州、中西部等。我们使用这些地理分布数据中心 48 小时的日前每小时电价（美元/兆瓦时），即从 2012 年 1 月 31 日凌晨 0 点到 2012 年 2 月 2 日凌晨进行模拟。对于每个数据中心，内部直流批量数据传输的单位输入和输出能耗 eini 和 eouti 在[10, 50]KWh 之间均匀随机。为了匹配每小时的电价数据，我们将 inter 数据中心批量数据传输的调度间隔（时隙）设置为 1 小时，即我们可以在每个时隙带宽为 2.5 Gbps 的链路上传输 9000 G 数据。

图 10 真实的网络拓扑结构

4.5.2 比较方法

我们比较研究了以下数据中心间批量数据传输调度方法：

FAST, 该方法旨在通过确保在每个最近的时隙中传输的最大数据量来实现最快的 inter 数据中心-批量数据传输传输。

FAST_MIN，此方法在保持最快传输的同时最大限度地降低了能量消耗。文献采用了相似模型（最大并发流和最小代价多商品流）来最小化数据中心间视频业务的带宽成本。

AVG_Demand, 该方法将批量数据传输需求平均分配到所有可用的时隙中，然后为每个时隙寻找最优的路由和调度，以使能耗最小。

EXPANSION, 这种方法在上文中讨论，通过将 MIN-EBC-批量数据传输转换为时间扩展网络，并解决时间展开图中的大规模转换问题。

2Stage_MinEC, 我们在第 4.2 节中提出的两阶段优化方法，用于快速求解 Min EC-BDT。

2State_MinE，这是我们两阶段优化方法的一个

变种，它使直流间无刷直流输电的能源消耗最小化，而不是能源成本最小化。

4.5.3 节约能耗成本效果

首先，我们为每个数据中心间批量数据传输设置批量数据需求 d=4500 G，并让服务提供商分别安排 3 个、5 个和 10 个数据中心间批量数据传输。1 对于每个数据中心间批量数据传输数量，截止时间 T 从能够完成数据需求的最小值到本评估中使用的最大截止日期 24 不等。调度从时隙 t0=1 开始，即 1 月 31 日凌晨 0 点到凌晨 1 点。通过将最大能量成本设置为 1，不同方法的归一化能量成本如图 11 所示。

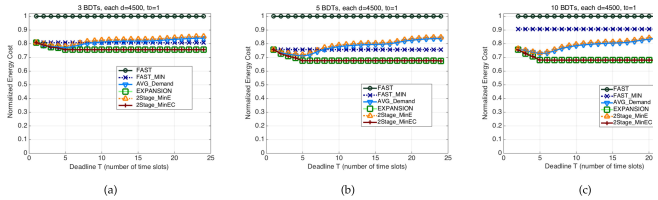

图 11 当 d=10；t0=1，且批量数据传输数量分别为 3、5 和 10 时，所有比较方法的归一化能量成本

总的来说，我们的方法（2Stage_Min_EC）和扩展方法在所有实验中都达到了最小的能量消耗。这两种方法在本质上是等价的，它们的唯一区别在于，第 2 阶段的优化解是在原始（底层）网络上求解的，而展开式在时间扩展的网络上求解。这表明了我们的方法在降低直流间无刷直流输电的能源成本方面有很大的潜力。该方法的能量消耗随着 T 的增加而降低，直至收敛。这是一个理想的属性，并证明了利用时间灵活性来安排数据中心间批量数据传输的必要性和重要性。当 T 增大时，我们的方法具有更大的时间灵活性，使得数据中心间的批量数据传输在较低的电价下调度。因此，它们的能源成本降低了。然而，当 T 增加到某个值时，持续放松它不会带来更低的值，因为在增加的时隙中不存在更低的电价。

4.5.4 计算时间

我们还记录了每种方法的计算时间（使用 matlab2011 和 CVX 解算器），以便在所有实验中得到解。我们在图 12 中绘制了 5 个批量数据传输和每个 d=13500 的平均计算时间，作为比较的代表性结果。

从这个图中我们可以看出，快速方法和快速最小法花费的时间最少，还没有考虑到我们的方法（P2Stage_MinEBC）的并行实现。这两种方法都是简单地利用最近的时隙来传递批量数据传输需求，求解所需的计算时间较少，且计算时间与 T 无关（T 的变化没有明显的变化）。快速方法比快速方法慢一些，因为从快速方法求解的可行解中找出能量消耗最小的解还需要一定的时间。AVG_需求和两种两阶段优化方法（2Stage_MinE 和 2Stage_MinEC）的计算时间与 T 值几乎呈线性关系。

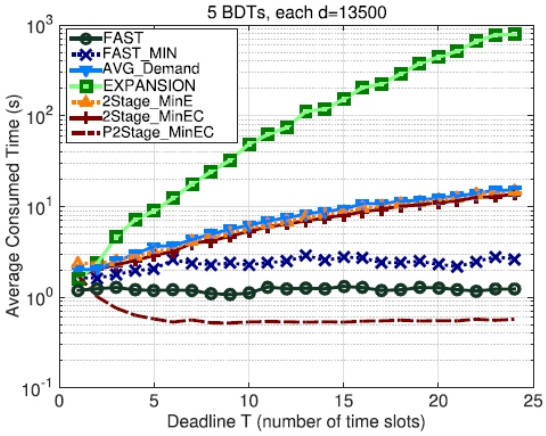

图 12 每种方法的平均计算时间

虽然这三种方法在计算时间上的差异很小，但是我们的方法（2Stage_MinEBC）比 AVG_Demand 和 2Stage_MinE 方法花费的时间更少。在所有的方法中，展开法的计算时间最大（几乎是指数到 T），因为它解决的是时间扩展网络上的优化问题，比原网络大得多。

虽然我们的方法（2Stage_MinEC）比 FAST 方法稍慢，但它比其他方法更快，并且实现了最大的能源成本节约。扩展方法也可以实现与我们的方法相同的节能效果，但计算时间要大得多，特别是当 T 较大时。实验结果表明，该方法能在节省能量和计算时间之间取得较好的平衡。此外，我们的方法的并行实现（P2Stage_-MinEC）使用 T 个并行线程并行求解最优解，进一步大大减少了计算时间（甚至低于快速方法）。

4.5.5 对数据传输量的鲁棒性

在下面，我们考虑数据中心间批量数据传输 3、5 和 10 数量的不同组合，每个数据中心间批量数据传输需要 d=4500，13500，27000 G，以表示

轻、中、高数据传输量，并评估我们的方法对数据传输量的鲁棒性。

通过使用不同组合重复上述试验，可在第 5.3.1 节中找到与 d=4500 相似的结果。我们的方法的能源成本总是最低的。证明了该方法对数据传输量的鲁棒性。我们还绘制了当 T=24 时，与图 6 中的快速方法相比，对于每组 d 和内部数据中心 批量数据传输的能量成本降低率。结果表明，该方法的能耗降低率明显高于其他方法。当数据传输量较低（5 批量数据传输 s 和 d=4500g）时，该方法的压缩比可达 32.5%。即使对于高水平的数据传输量（10 批量数据传输和 d=27000 G），与快速方法相比，我们的方法也可以实现大约 19.2%的能源成本降低。

图 13 还显示了一个轻微的趋势，即随着数据传输量的增加，通过我们的方法实现的节能降低，例如，对于固定数量的数据中心间批量数据传输（或 d），能量成本降低率随着 d 的值（或数据中心间批量数据传输的数量）的增加而减小。究其原因，是直流互联网络的链路容量有限，较大的数据传输量一般需要更多的时隙来传输数据。虽然我们的方法已经优先考虑了电价较低的时隙，但是随着数据传输量的增大，每批量数据传输数据的能量成本也会增加。这就是为什么当数据传输量较大时，我们的方法节省的能量成本会降低，即使对于每个给定的数据中心-批量数据传输数据传输量，它已经达到了最小的能量成本。

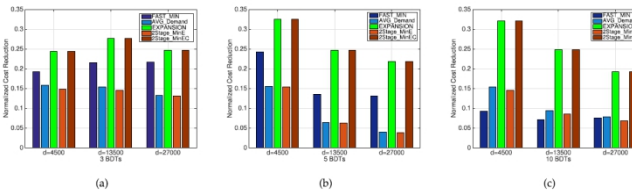

图 13 当 T=24 和 t0=1 时，不同批量数据传输数量的所有比较方法的能源成本降低率。

4.5.6 对调度开始时间的鲁棒性

在下面，我们首先将 d 和 T 分别固定为 4500g 和 6，并在图 14 中绘制了当 t0 从 1 到 24 变化时的归一化能量成本降低率曲线。该图直观地表明，不同的调度开始时间 t0 对数据中心间批量数据传输的能量消耗有很大的影响。即使是快速和快速的方法，它们的能量消耗也不再是常数。为了验证我们的方法在不同的调度开始时间下的节能鲁棒性，我们重复上面的实验，如图所示。4 和 6，将 t0 从 1 设置为 24。对于 t0 的每一个值，我们的方法的能量成本仍然是最低的，这表明我们的方法对于调度开始时间 t0 的能量成本节约是稳健的。所有方法的总体趋势与前面描述的 t0=1 的情况类似。为了与 t0=1（自由小时和电价较低）的情况进行比较，我们在图中绘制了能源成本和相应的降低率。8 和 9 分别为 t0=18（繁忙时间和电价高）。

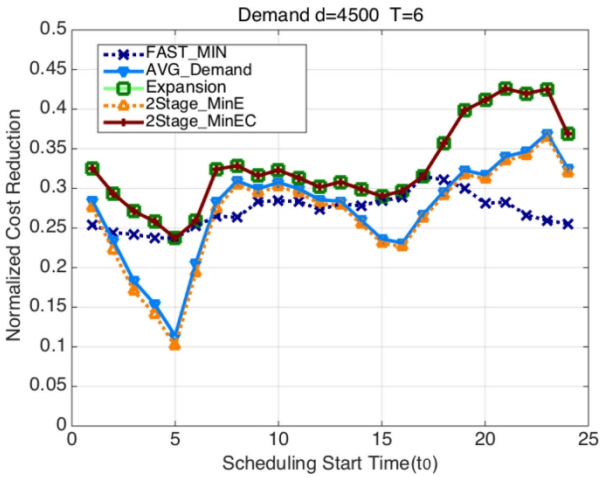

图 14 当 t0 从 1 到 24 变化时的归一化能量成本降低率曲线

比较以上的图，我们可以看到，当 t0=18 时，我们的方法的能源成本进一步降低（比率达到 48%）。这意味着当调度开始时间 t0 在繁忙时间时，我们的方法在节省能源成本方面获得了更多的好处。这种进一步的减少也通过比较图来确认。6 和 9。它们都显示了当 T=24 时，需求 d 和直流间批量数据传输数量的不同组合的能量成本降低率。图 9 还验证了我们的方法在数据传输量上的上述趋势。值得注意的是，当 t0=18 时，与 t0=1 相比，平均需求和 2 阶段采矿法的能源成本呈现出不同的变化趋势。当 T 从 1 变化到 24 时，它们的能量消耗先增加（只有几个时隙），然后降低。这也是由于现实生活中的电价变化，如前所述（繁忙时间过后，电价一般呈下降趋势）。

当 t0 处于其他时隙时，我们也测试了所有的方法。结果表明，我们的方法节省能源成本的性能介于免费（t0=1）和繁忙时间（t0=18）之间，但在所有方法中总能达到最低能耗。由于篇幅的限制，我们跳过了细节。

4.5.7 大规模模拟

由于实际直流互联网络的规模局限于实际情况。接下来，我们将通过大规模仿真来评估我们的方法的性能。在这个模拟中，我们使用随机图的演算法来产生不同节点数的网路拓扑。每个节点（数据中心）位于随机选择的区域电力市场中。如果多个数据中心位于同一电力市场，我们将使用不同的集线器来区分它们。其他参数的值设置与之前相同。

首先，我们让数据中心的数量在 10 到 100 之间变化。对于每个数据中心间网络，我们将数据中心间 批量数据传输的数量固定为节点数量的一半，d=4500，t0=18（忙时），T=24。请注意，由于在大规模环境中求解优化解的速度太慢，因此本模拟不包括展开法。图 15 示出了与快速方法相比的能量成本降低率。如图所示，我们的方法（2Stage_MinEC）在所有情况下实现了最大的能源成本降低率。降低能源成本方面我们的方法比任何其他方法都要高得多。结果表明，该方法在大规模环境下也能显著节约能源成本。此外，我们的方法节省的能量不会随着节点数的增加而恶化。这是一种理想的特性，特别是在未来直流网络规模可能大幅增加的情况下。主要原因是对于规模更大的跨直流网络，可以使用更多的地理分布数据中心（可能在较低的电价下）来执行跨直流批量数据传输，这带来了更大的时间和空间灵活性，以最小化能源成本。

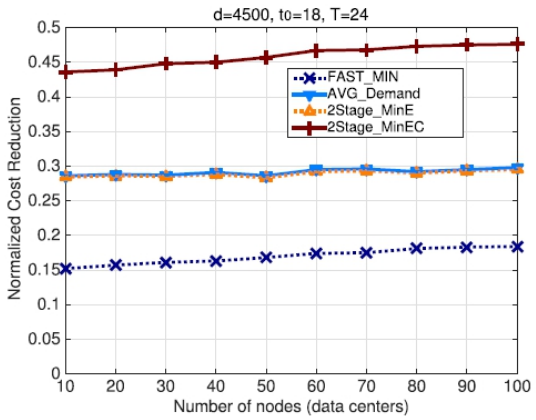

图 15 不同节点数的标准化成本降低率

然后，我们将节点数目固定为 50 个，并改变数据中心间批量数据传输的数量，以在不同的数据传输量规模上评估这些方法。如图 16 所示，当数据中心间批量数据传输的数目从 10 到 100 变化时，我们的方法的能量成本降低率总是大于任何其他方法。结果表明，与现有的批量数据传输方法相比，即使在数据中心间批量数据传输规模较大的情况下，该方法也能显著地节省能源成本。随着数据中心-批量数据传输数目的增加，快速最小法的能耗降低率比其他三种方法都要快。当数据中心间批量数据传输的数目为 100 时，仅实现 5%的成本降低。虽然我们的方法显示出随着直流间批量数据传输数量的增加而略有下降的趋势，但是它仍然比其他方法节省能源成本。这表明，即使在直流间批量数据传输规模不断扩大的情况下，我们的方法也能显著地节省能源成本。

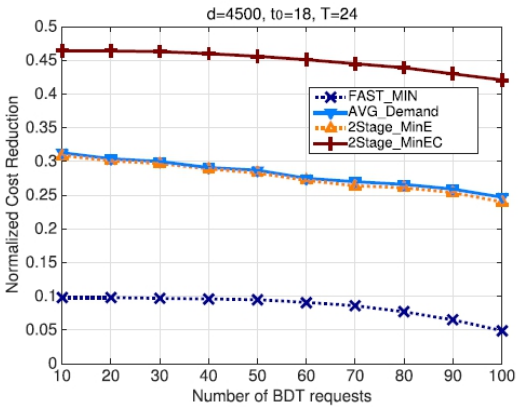

图 16 不同批量数据传输请求数下的标准化成本降低率

综上所述，对实际直流网络和实际电价的评估表明，我们的方法比现有的跨直流无刷直流输电系统方法节省了大量的能源成本。大规模仿真结果表明，该方法不仅在实际有限的案例研究中，而且在大规模的仿真环境中都能实现显著的节能效果。

5 总结与展望

地理分布的数据中心的能耗是互联网服务提供商（ISP）的一个重要成本因素。ISP 使用各种工作负载和电源管理技术来降低总体能源成本。在这篇论文中，我们研究了一个同时考虑动态电价、现场可再生能源和活动服务器数量的总能源成本最小化问题。为了解决这一问题，我们提出了一种基于贪婪算法设计技术的绿色地理负载平衡

（GreenGLB）在线算法，用于分配交互式和不可分割的传入工作负载。实验结果表明，使用本文提出的算法可以显著降低伪造分布式数据中心的总能耗。

至于将来的工作，我们希望将所提出的模型扩展为其他参数，例如批处理工作负载和带宽成本。基于异构服务器的分布式控制系统将用于跨多个地理分布数据中心的成本最小化问题建模。

另外在本文中，我们还重点讨论了在有/无网络预算约束的情况下，如何在地理分布的数据中心中更接近数据地调度任务。通过测量 amazonec2 数据中心的可用带宽，我们首先发现网络可能是地理分布大数据处理的瓶颈。因此，我们将无网络预算约束的工作负载问题描述为一个整数线性规划问题，同时考虑了网络和计算资源的约束。我们还发现，我们可以将整数线性规划问题转化为一个线性规划问题，且具有相同的最优解。然而，我们发现达到最佳完成时间并不能保证最佳的网络成本。因此，我们将网络预算约束下的工作负载问题分别描述为另一个整数线性规划问题。

基于这些理论见解，我们设计并实现了 Flutter，一种新的跨地理分布数据中心的任务调度框架，用于在有/无网络预算约束的情况下，对两种工作负载进行任务调度。通过使用 Amazon EC2 上的跨数据中心网络测试台和虚拟机进行的实际性能评估，我们已经展示了令人信服的证据，即 Flutter 不仅能够缩短作业完成时间，而且能够减少需要传输到其他数据中心的流量。我们还可以大大降低跨数据中心的数据传输成本。在我们未来的工作中，我们将研究如何在广域大数据处理环境中联合优化数据放置、复制和任务调度，以获得更好的性能。我们还计划应用 DAG 调度技术直接优化作业级性能。

云计算的快速扩散促进了大规模商业数据中心的快速增长。地理分布的数据中心经常被主要的云服务提供商用来为客户提供更好的可靠性和服务质量。在这样大规模的地理分布数据中心网络中，由于需要在这些数据中心之间传输大量数据（定期数据备份、软件分发、虚拟机克隆、分布式数据库等），数据中心间的批量数据传输成为一个重要且日益增长的需求。虽然地理分布和延迟容忍度大的数据中心间的海量数据传输已经被许多工程应用于降低地理分布数据中心的运行成本，还有很多问题有待研究。基于现有直流间大容量数据传输的研究主要集中于带宽成本的优化，本文提出了一种有效的两阶段优化方法来解决多电力市场环境下地理分布数据中心的能量成本最小化问题。对于每个批量数据传输，该方法首先沿可用时隙搜索最优需求分配，然后分别计算每个时隙的最优路由和调度。对实际直流网络和实际电价的大量评估表明，与现有的跨直流大容量数据传输调度方法相比，所提出的两阶段优化方法可以显著节省能源成本。

在未来，我们计划整合中间数据中心的储存和转发能力，以进一步降低数据中心间批量数据传输的能量成本，即中间数据中心能够暂时储存要中继的大量数据，并在稍后转发到其下游中继节点或目的地。通过适当地确定批量数据应在中间数据中心存储的时间和数量，可以为数据中心间的批量数据传输节省更多的能源成本。

参 考 文 献

[1] Lu X , Kong F , Liu X , et al. Bulk Savings for Bulk Transfers: Minimizing the Energy-Cost for Geo-Distributed Data Centers[J]. IEEE Transactions on Cloud Computing, 2017:1-1.

[2] Khalil I , Ahmad I , Almazroi A A . Energy Efficient Indivisible Workload Distribution in Geographically Distributed Data Centers[J]. IEEE Access, 2019, 7(1):82672-82680.

[3] Hu Z , Li B , Luo J . Time- and Cost- Efficient Task Scheduling across Geo-Distributed Data Centers[J]. IEEE Transactions on Parallel & Distributed Systems, 2018, PP(3):1-1.