

面向智能电网的数据密集型云存储策略研究

丁 杰，奚后玮，韩海韵，周爱华

（中国电力科学研究院，江苏 南京 210003）

摘 要：智能电网环境下数据密集型应用往往涉及跨数据中心的数据传输和数据中心内的数据迁移，这对数据分布提出了新的挑战。为充分利用计算存储资源，满足智能电网大规模数据可靠存储和高效处理的实际需求，提出了基于云计算的数据密集型存储方法，该方法将数据集映射成数据空间的点集。设计了两阶段分类过程：第1阶段基于传统的K均值算法实现点集的初始分类；第2阶段针对各数据集与初始聚类的隶属关系，引入数据迁移的代价函数，对初始分类进行调节，实现数据集到数据中心的布局方案。实验结果表明，该算法能够有效地提高数据存取效率和兼顾全局负载均衡。

关键词：智能电网；云计算；数据分布；数据迁移；一致性哈希

0 引言

伴随着互联网的深度普及以及IT应用模式转变^[1]，数据量开始爆炸性增长，数据价值逐渐提高，海量数据处理的需求越来越迫切。基于资源管理效能和信息处理能力的优势，云计算在大规模科学计算等领域已取得成功案例^[2-4]，但在电力系统中的应用研究还处于起步阶段。目前电力云计算在平台设计、系统实现和前景展望等方面提出了初步设想，例如：文献[5]针对传统电力系统计算平台在计算、存储、信息集成和分析等方面的不足，提出建立基于云计算的电力系统计算平台，展望了云计算在电力系统安全分析、潮流与优化潮流计算、系统恢复、监控、调度、可靠性分析等领域的应用前景；文献[6]研究了云计算中的虚拟化、分布式存储与并行编程模型等问题，提出基于云计算的智能电网信息平台的体系结构，实现智能电网海量信息的可靠存储与快速并行处理；文献[7]针对电力云计算应用的可行性和必要性问题，阐述了电力系统云计算中心的建设目标和系统特点，将仿真云计算中心系统架构划分为基础设施云、数据管理云、仿真计算云、协同工作云和咨询服务云等多个层次，并给出了实际应用场景。

云存储是在云计算概念上延伸和发展出来的，是通过集群应用、网格技术或分布式文件系统等功能，将集群内的物理存储资源无缝整合为统一的虚拟存储资源，为用户提供透明虚拟存储资源，共同对外提供数据存储和业务访问功能的系统。目前，云存储发展呈现分布式数据密集型趋势，广泛应用于天文学^[8]、物理学^[9]和生物信息学^[10]领域，这类应用的部署和执行所涉及的TB，甚至PB级的数据往往存储于分布式的数据中心，需要多数据中心的有机协同。而在电力行业，随着电网建设规模的不断扩大，数字化电网、数字化变电站等研究应用的不断深入，系统面对的采集点越来越多。一个中等规模地区的采集量可以达到2万至10万，而一个大型地调未来可能面临50-100万的数据采集规模，一年的数据存储规模将从目前的GB级转向TB级。此外，随着调度自动化水平的不断提高，提出了实时运行数据不采用周期性采样存储而是按照实际时间序列连续存储的更高的要求，以满足更多的应用需求，这也将导致数据存储规模数十倍的增长，同时，历史数据的存储组织策略以及查询检索策略也将变得相当复杂。如此海量规模的电力信息能否实现有效存储并进行高效处理将是一个很大的问题。

针对上述问题，常见的数据密集型管理策略主要针对分布式环境下的大规模数据建模和基础设施服务展开研究，例如：文献[10]使用一种面向角色的数据建模方法，用于网格环境下的数据建模，并使用数据网格来对数据进行管理；文献[9]采用流程定义语言表示其数据流，实现数据资源的优化管理；文献[11]在数据流定义的基础上使用P2P模式实现分布存储资源中海量数据集的访问、移动和修改。然而，现有的系统的数据管理策略没有关注数据的存放分布和数据间依赖性的分析，无法减少数据迁移所带来的

时间开销和提升整体执行效率。因此，本文结合智能电网的数据特性，分析云计算环境下存储应用系统的特点：①分布式应用所需的数据集通常位于多个数据中心，数据中心间的数据传输无可避免，需要考虑网络带宽资源和传输的时间开销；②是数据依赖性能够有效地提高数据管理和使用的效率，需要设计合理的数据分布策略来保持数据间的依赖关系；③数据中心内数据分布的可扩展性是提升整体性能的重要因素，需要对不同的数据分布模式进行量化分析，兼顾数据的均匀分布和全局的负载均衡。在此基础上全面分析数据传输次数、数据集大小以及数据中心间网络带宽等因素，通过聚类分析、依赖性分析和哈希算法对多数据中心的数据集分布进行统筹规划，并引入系统执行效能的代价函数对数据分布方案进行评价和调整，从而在降低系统开销的同时最大限度地兼顾数据集间的依赖关系。

1 多数据中心的数据分布

1.1 云存储模型

随着电网调度“大运行”体系的提出，电网特性呈现区域模式主导转向总体模式的发展趋势，电网耦合也越来越紧密，对一体化运行提出新的要求。为紧密结合一体化运行的实际需求，在广域分布的智能电网调度技术支持系统基础上，探索建立在云计算技术基础上的调度系统数据存储架构，支撑电力流、信息流、业务流的高度一体化。本文涉及的数据中心专指广域范围的调度中心，类似“三华”互备的模式，采用广域组网技术，在满足管辖范围内的调控功能需求的同时具有大电网数据级互备能力。在此基础上对电力云存储进行初步设想，将存储包括数据分布和数据灾备2个重要部分。针对多数据中心的存储备份问题，目前已开展了相关的实践，如国家电网公司已建成的北京、上海和西安三地数据容灾中心，采用就近灾备的策略，利用异步镜像技术实现网省公司及直属单位生产中心数据在容灾中心的异地备份功能，已基本具备了公司生产数据的容灾能力。但受网络通信开销等因素限制，暂时无法实现实时灾备以及三地互备的能力。因此，本文提出的云存储模型主要关注于多数据中心的数据分布策略的研究，而对于数据灾备问题，暂不考虑多数据中心的数据级容灾，仅在数据中心内部基于云计算的容错技术或借助于业务系统本身所提供的灾备方案进行数据备份。

本节将智能电网信息处理抽象成工作流和数据流，结合云存储下数据分布的相关概念进行建模，包括数据模型、应用模型和依赖关系等。智能电网信息平台可以表示为扁平化的多个分布式数据中心（不考虑目前调度领域多级数据中心的纵向贯通）组成的集合 $C=\{C_1, C_2, \dots, C_m\}$ ， C_i 表示第 i 个数据中心。由于云计算的数据类型具有复杂性和多样性特点，因此本文屏蔽了智能电网环境下数据的结构特性，数据被视为数据密集型应用环境下面向多任务的数据集。定义数据流关联的数据集的全集为 D ，相应工作流的任务集为 $T=\{T_1, T_2, \dots, T_n\}$ 。对于任意的数据集 $d_i \in D$ ，定义描述数据集属性的二元组为 $\langle T_i, s_i \rangle$ ，其中 $T_i \subseteq T$ 表示调用数据集 d_i 的所有任务的集合， s_i 是数据集 d_i 的大小。对任意的数据集 d_i 和 d_j ，相应的依赖关系定义为

$$dep_{i,j} = |T_i \cap T_j| \quad (1 \leq i, j \leq n) \quad (1)$$

其中 $|T_i \cap T_j|$ 表示任务集中同时调用了 d_i 和 d_j 的任务数量，由于电力系统大多业务应用与数据源间的映射关系相对比较稳定，如调度管理早会包的报表设计，同一报表内容往往关联多级调度机构的不同数据库，但报表字段通常是固定的，且对应于指定的数据库。因此，可以统计一段时间内任务集对数据集的调用次数，来确定数据集之间的依赖关系。

云计算应用于智能电网需要整合电力系统现有的业务数据信息和计算存储资源，业务应用往往涉及分处于不同数据中心的多个数据集，需要移动计算或移动相关的数据集到任务调度的数据中心，数据迁移的时间开销无可避免。对于云环境中的 m 个数据中心 $C=\{C_1, C_2, \dots, C_m\}$ ， C_i 与 C_j 间的带宽表示为 b_{ij} ，相应的带宽矩阵为

$$B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ b_{m1} & b_{m2} & \cdots & b_{mm} \end{bmatrix} \quad (2)$$

从而，数据集 d_k 在数据中心 C_i 与 C_j 间迁移的时间开销计算如下：

$$TC(d_k, C_i, C_j) = \frac{s_k}{b_{i,j}} + \varepsilon_{i,j}^k \quad (1 \leq k \leq n, 1 \leq i, j \leq m) \quad (3)$$

其中 $\varepsilon_{i,j}^k$ 表示数据传输过程中请求、响应等造成的额外开销，由于云计算环境下的数据规模较大， $\varepsilon_{i,j}^k$ 相对较小，故忽略上述额外开销后，相应的时间开销可以简化为

$$TC(d_k, C_i, C_j) \approx \frac{s_k}{b_{i,j}} \quad (1 \leq k \leq n, 1 \leq i, j \leq m) \quad (4)$$

1.2 两阶段数据分布策略

本节引入数据集的聚类分析，设计了数据分布的两阶段策略：第一阶段基于K均值分析[12]对数据进行迭代计算，生成数据集的初始分类；第二阶段详细分析数据集和数据中心间的隶属关系，引入数据传输的时间开销评估，形成数据集地最优分布。为形成数据集到各数据中心的映射视图，需要将数据集视为数据空间的特征点集，对于给定的数据集 d_i 和 d_j ，两者在数据空间的距离可以根据依赖关系计算为

$$dist(d_i, d_j) = \begin{cases} \frac{1}{dep_{i,j}} & (dep_{i,j} \neq 0, 1 \leq i, j \leq n) \\ 0 & else \end{cases} \quad (5)$$

具体分类由迭代过程完成，首先计算当前隶属于各数据中心的数据集合的几何中心，从而待分类数据集到数据中心的距离可以表征为与中心几何中心的距离，并将该数据集归并到距离最小的数据中心，迭代的终止条件为各数据中心的数据组成不再变化，从而各数据中心可以映射为空间的 m 个数据集类，记为 $\omega = \{\omega_1, \omega_2, \dots, \omega_m\}$ 。其次，对数据集类抽取代表元形成，分析各数据集与相应代表元的距离关系，实现数据的精确分布。本文使用AP聚类算法[13]进行代表元的抽取，其输入为数据集间的实值相似度。对于数据集 d_i 和 d_k ，相似度由数据空间中两者间的负欧式距离给出：

$$sim(i, k) = -\|dist(d_i, d_k)\|^2 \quad (6)$$

对所有的数据集对进行相似度计算，依据样本间相似度构建相似度矩阵并设定参考值之后，数据集间的两类消息（可靠度和有效度）分别采用不同的机制不断地被更新，二者可认为是对数似然比。可靠度 $r(i, k)$ 反映了数据空间中 d_k 作为 d_i 的代表特征点的可信度，更新规则为

$$r(i, k) \leftarrow \leftarrow sim(i, k) - \max_{k' \neq k} \{a(i, k') + sim(i, k')\} \quad (7)$$

有效度 $a(i, k) + r(i, k)$ 反映了数据空间中 d_k 作为 d_i 的代表特征点的累积可信度，通过收集来自样本数据的信息而确定是否每一个候选代表样本是个合适的代表样本。，更新规则为

$$a(i, k) \leftarrow \leftarrow \min \left\{ 0, r(k, k) + \sum_{i' \in \{i, k\}} \min \{0, r(i', k)\} \right\} \quad (8)$$

通过组合可靠度和有效度来确定代表元，二者仅需要在数据集对之间传输。对于数据集 d_i 而言，认为使得 $a(i, k) + r(i, k)$ 最大化的数据集 d_k 作为 d_i 的代表元是合理的，而当局部 $a(i, k) + r(i, k)$ 值保持不变时，消息传递过程将停止。

通过上述迭代机制对 m 个数据集类均选取一个代表元，生成的代表元记为 $E = \{e_1, e_2, \dots, e_m\}$ ，其中 e_i 表示类 ω_i 的代表元。为生成数据的最优分布，需要在数据集的初始分布基础上引入数据传输的时间开销进一步迭代分析，基于2点假设：①数据集是低耦合高内聚的，即数据集的划分具有原子性；②数据传输

是以数据集为单位的，即每次数据集的迁移都需要传输该数据集的全部数据。以下以数据集 dt 为例计算时间开销。假设目标数据集 dt 分布于数据中心 C_{des} ，由 dt 的属性集 $\langle T_t, st \rangle$ 可知集合 T_t 包含了所有调用数据集 dt 的任务。将任务集 T_t 划分为 m 个子集 $\{T_{t1}, T_{t2}, \dots, T_{tm}\}$ ，子集 T_{ti} 为 T_t 中调度至数据中心 C_i 运行的任务。对于任务集 T_{ti} 中的各个任务均需要将数据集 dt 由中心 C_{des} 传输至 C_i 来进行数据调用，则对于传输 dt 而言的总体时间开销可以计算如下：

$$TotalTC_t = \sum_{i=1}^m |T_{ti}| TC(d_t, C_i, C_{des}) \quad (9)$$

其中 $|T_{ti}|$ 表示任务子集 T_{ti} 中所包含任务的数量。

将 dt 与 E 中各元素进行比较，选取具有如下条件 C_ξ 作为 dt 隶属的数据中心，式中 λ 为实验的经验参数。

$$\xi = \arg \left(\min_{1 \leq i \leq m} \left(dist(d_t, e_i) \times \left(1 - \lambda \times \frac{\min\{TotalTC_k\}}{TotalTC_i} \right) \right) \right) \quad (10)$$

其中 $dist(dt, e_i)$ 表示 dt 与代表元 e_i 之间的空间距离，通过最小化空间距离来确定 dt 所隶属的数据中心，为兼顾数据传输的时间开销问题，通过计算传输 dt 到候选数据中心的时间开销作为修正参数，即传输数据的时间开销越大，则修正参数也越大，从而对目标函数起到约束作用。

2 本地数据中心存储模型

目前电力系统存储网络已经具有比较完整的物理架构，以及整合分布在各级的存储资源，如容灾中心已经建成包括同构企业级存储设备和光纤交换机在内的基于存储区域网（SAN）的基础架构平台，依照存储同构原则进行存储设备的逻辑划分，通过链路上的存储设备冗余来提高存储效能。而调度领域由于数据专业性较强，一般都是使用独立的存储系统，没有组成统一的存储网络，其数据或模型的交换主要基于接口开发实现。随着三华互备的建设，电力调度也逐渐形成统一的数据中心，但智能调度支持系统主要还是基于应用层接口封装来实现对不同模块数据的统一访问。因此，现阶段电力系统的数据存储架构需要整合形成扁平化的对等存储网络，数据存储的效率和扩展性有待进一步提高。

针对上述问题，借鉴产、学、研各界存储相关的数据管理研究^[14-16]（主要是将数据切分为相同大小的数据块，并随机地分配到不同的物理磁盘）。本文研究的数据存储建立在底层分布式文件系统的开源架构基础上，基于HDFS提供的容错机制实现数据的容灾备份。在数据存储机制的设计上，本节不考虑存储网络（如SAN）复杂的路由结构，将存储系统简化为1个磁盘的集合，表示为 $Disk = \{Disk1, Disk2, \dots, Diskl\}$ ，同时回避了数据的切分过程，围绕数据存储的有效性和扩展性进行分析，达到2个方面的设计要求：①数据集平均分布于不同的磁盘；②最小化物理磁盘增加或删除所导致的数据重新分布的开销。对此，业界较为成功的模型是Amazon Dynamo架构^[17]。本文在此基础上通过一致性哈希算法^[18]计算数据集的键值对（key-value），形成数据布局的环形拓扑，实现了系统中不同磁盘间数据的动态分布。为降低算法设计的复杂性，假设所有的物理磁盘均具有相同的空间大小，选取合适的哈希函数 h 将数据集映射到实值区间 $[0,1]$ ，同时将区间 $[0,1]$ 划分为 l 个子区间，分别对应于 l 个磁盘。例如，

磁盘 i 对应的子区间表示为 $\left[\frac{i-1}{l}, \frac{i}{l}\right]$ 。考虑第 $l+1$ 个磁盘被添加的情况，对所有的子区间进行对等划分，如 $\left[\frac{i-1}{l}, \frac{i}{l}\right]$ ($1 \leq i \leq l$) 划分出的子块计算为 $\left[\frac{i(l+1)-1}{l(l+1)}, \frac{i}{l}\right]$ ($1 \leq i \leq l$)。所有的子块均被集中统计和映射为区间 $\left[\frac{l-i-1}{l(l+1)}, \frac{l-i}{l(l+1)}\right]$ ($1 \leq i \leq l$)，并分配至磁盘 $l+1$ ，如图1所示。

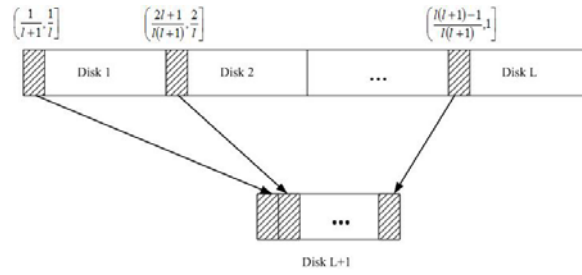


图1 磁盘数据的区间划分

3 实验结果及分析

中国电力科学研究院信息与通信研究所拥有先进的云计算研发实验室环境，包括：完整的计算机网络环境、20台刀片服务器（2路4核CPU，4G内存）和相应的操作系统环境、Oracle、DB2、SQL Server、Sybase等大型数据库系统环境、各种开发工具套件和丰富的组件库、门类齐全的技术文档资料库。在此基础上构建云存储原型系统，对文中提出的数据分布策略进行了实验的仿真执行和分析，并记录了任务执行的数据传输次数和时间开销。

原型系统环境的物理网络拓扑如图2所示，共划分虚拟数据中心数目为10，通过三层交换机互连，数据中心连接的网络带宽为1000M。数据中心内通过虚拟存储节点形成环形的逻辑结构，虚拟实例配置为CPU 3C、RAM 1.5G、硬盘120G。

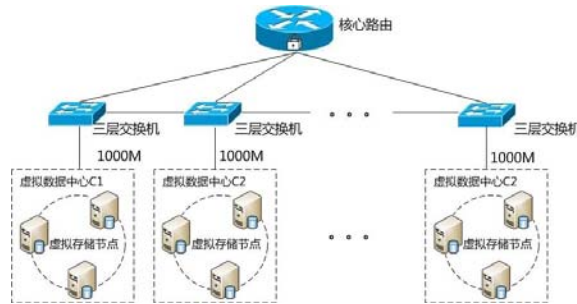


图2 原型系统的网络拓扑结构

测试数据包括50438个文档（PPT、WORD、XLS、TXT、PDF等）、14099张图片、182个视频、配置库及数据库文件，共计1.1T，将测试数据分别划分为20、40、...、120个数据集，每个数据集的大文件以文件切片的方式（最大切片为64M）进行分布式存储。实验任务集是参数相同的测试流程，随机分布于10个数据中心，且与数据集间的调用关系也是随机确定的，每个任务Tk包括两组参数，由一个二元组{Ck,Dk}描述，其中Ck表示Tk运行所在的数据中心，Dk={dk1, dk2,...}为Tk调用的数据集的集合。表1给出的测试结果是测试流程的统计均值，从统计结果可以看出，数据的传输次数和时间开销随着数据集数目的增多而增加，从实验的数据规模来看，两者的增加率逐渐呈现下降趋势。实验统计结果的多项式插值拟合图见图3和图4。

表 1 数据传输次数和时间开销统计

数据集数目	数据中心数目	传输次数	时间开销/h
20	10	37	0.126
40		142	0.203
60		226	0.374
80		291	0.552
100		322	0.591

120		335	0.638
-----	--	-----	-------

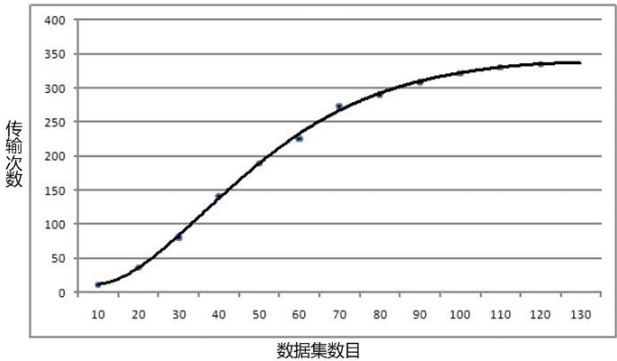


图3 传输次数统计图

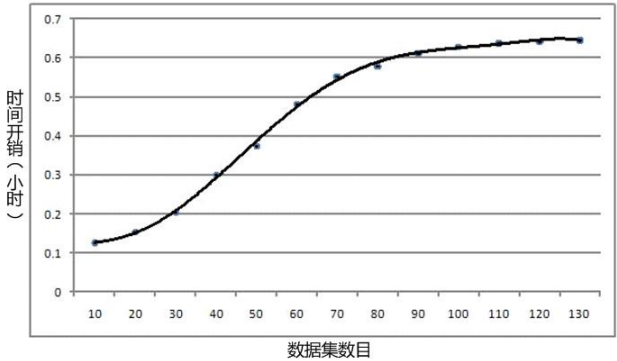


图4 时间开销统计图

4 结束语

本文针对云计算环境下面向数据密集型应用的数据管理问题和挑战，特别是跨数据中心数据传输的时间开销和数据中心存储的可扩展性，在对问题的分析、建模和综合考虑数据规模、网络带宽等因素的基础上，提出兼顾系统开销和数据依赖性的数据分布策略，并通过实验仿真测试进行了算法验证。下一步工作重点包括四个方面：一是研究现有原型向云计算分布式数据中心环境应用的演进路线规划和分析，着重改进聚类分析算法实现数据集的有效划分，以及降低时间开销和兼顾负载均衡；二是弱化算法的前提假设条件，以提高算法的适用性和可扩展性；三是在本文使用的云计算副本技术基础上，研究跨数据中心的冗余备份问题；四是进一步细化算法的实验环节，增大测试规模并在较为复杂的网络条件下进行测试，以更好地验证本文方法的有效性。

参考文献：

[1] WEISS A. Computing in the Cloud[J]. ACM Networker, 2007, 11(4): 18-25.

[2] BRANTNER M, FLORESCUY D, GRAF D, et al. Building a database on S3[C]// Proceedings of the 2008 ACM SIGMOD international conference on management of data, Jun 9-12, 2008, Vancouver, Canada.

[3] BUYYA R, YEO C S, VENUGOPAL S. Market-oriented cloud computing: vision, hype, and reality for delivering IT service as computing utilities[C]// Proceddings of the 10th IEEE International Conference on High Performance Computing and Communications, September 25-27, 2008, Dalian, China: 5-13.

[4] GROSSMAN R, Gu Yunhong. Data mining using high performance data clouds: experients studies using sector and sphere[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24-27, 2008, Las Vegas, NV, USA.

- [5] 赵俊华,文福拴,薛禹胜,等. 云计算:构建未来电力系统的核心计算平台[J].电力系统自动化,2010,34(15):1-8.
- [6] 王德文,宋亚奇,朱永利.基于云计算的智能电网信息平台[J].电力系统自动化,2010,34(22):7-12.
- [7] 沐连顺,崔立忠,安宁.电力系统云计算中心的研究与实践[J].电网技术,2011,35(6):171-175.
- [8] DEELMAN E, CHERVENAK A. Data management challenges of data-intensive scientific workflows[C]// Proceedings of the IEEE International Symposium on Cluster Computing and the Grid, May 19-22, 2008, Lyon, France: 687-692.
- [9] DEELMAN E, BLYTHE J, GIL Y, et al. Pegasus: Mapping scientific workflows onto the grid[J]. Grid Computing, 2004(3165): 131-140.
- [10] LUDANSCHER B, ALTINTAS I, BERKLEY C, et al. Scientific workflow management and the Kepler system[J]. Concurrency and Computation: Practice and Experience, 2005, 18(10): 1039-1065.
- [11] CHURCHES D., GROMBAS G, HARRISON A, et al. Programming scientific and distributed workflow with Triana services. Concurrency and Computation: Practice and Experience[J]. 2006, 18(10): 1021-1037.
- [12] HARTIGAN J A, WONG M A. A k-means clustering algorithm[J]. Applied Statistics, 1979, 28(1): 100-108.
- [13] FREY B J, DUECK D. Clustering by passing messages between data points. Science[J]. 2007, 315(5814): 972-976.
- [14] GHEMAWAT S, GOBIOFF H, LEUNG S T. The google file system. ACM SIGOPS Operating Systems Review[J]. 2003, 37(5): 29-43.
- [15] BLANM M, BRADY J, BRUCK J, et al. Evenodd: an optimal scheme for tolerating double disk failures in RAID architecture[J]. IEEE Trans on Computers, 1995, 44(2): 192-202.
- [16] SCHWABE E J, SUTHERLAND I M. Flexible usage of redundancy in disk arrays[J]. Mathematics Systems Theory, 1999, 32(5): 561-587.
- [17] DECANDIA G, HASTORUN D, JAMPANI M, et al. Dynamo: Amazon's highly available key-value store[J]. ACM SIGOPS Operating Systems Review, 2007, 41(6): 205-220.
- [18] KARGER D, LEHMAN E, LEIGHTON T, et al. Consistent hashing and random trees: distributed caching protocols for relieving hot spots on the World Wide Web[C]// Proceedings of the 29th Annual ACM Symposium on Theory of Computing, May 4-6, 1997, El Paso. TX, USA: 654-663.

作者简介:

丁 杰 (1983-), 男, 博士, 工程师, 主要从事电力系统信息化、模式识别和图像处理的研究工作, E-mail: ding-jie2@sgepri.sgcc.com.cn;

奚后玮 (1963-), 男, 学士, 教授级高级工程师, 主要从事电力系统分析与控制方面的研究开发和管理的工作, E-mail: xihouwei@sgepri.sgcc.com.cn;

韩海韵 (1980-), 男, 学士, 工程师, 主要从事电力信息化研究工作, E-mail: hanhaiyun@sgepri.sgcc.com.cn。