文章编号:1671-4598(2019)04-0198-05 DOI:10.16526/j.cnki.11—4762/tp.2019.04.044 中图分类号:TP309.2 文献标识码:A

大型计算机集群数据持续保护研究与设计

曾 发[1],麻雨欣[2],曾贵明[1],梁 君[1],荣 刚[1]
(1.中国运载火箭技术研究院 研发中心,北京 100076;2.航天材料及工艺研究所,北京 100076)

摘要:随军工集团业务应用在线协同发展,数据存储对持续保护、容灾安全、读写速度要求越来越高;针对这些要求,设计一套大数据持续保护系统,简单增加一套存储硬件,部署相关软件,不改变现有存储方案硬件结构,在块级层面实现对数据定时备份、持续数据保护、查询统计,提高数据存储安全性、容灾性能和读写速度。

关键词:大数据;块级;持续数据保护;容灾

Research and Design of Data Continuous Protection for Scale Computer Cluster

Zeng Fa[1], Ma Yuxin[2], Zeng Guiming[1], Liangjun[1], Ronggang[1]
(1. Research and Development Center, China Academy of Launch Vehicle Technology, Beijing 100076, China;
2. Aerospace Research Institute of Materials and Processing Technology, Beijing 100076, China)

Abstract: As the development of vocational work applications online in military group, the requirements for continuous protection, disaster-tolerant security, read/write speed of data storage, are getting higher and higher. To meet these requirements, a big data continuous protection system (BDCPS) is designed, by simply adding a storage device and deploying related software, without changing the existing storage scheme architecture, in order to achieve the backup at regular time, continuous data protection, inquiry and statistics of data at block level, and to improve storage security, disaster tolerance and read/write speed of data storage.

Keywords: big data; block level; continuous data protection; disaster tolerant

0 应用背景

根据IDC(internet data center)的调查,在2000年以前的十年间发生过数据灾难的公司中,有55%的公司当即倒闭,剩下的45%中,也因数据丢失,有29%在两年内倒闭,生存下来的仅占16%[1-2]。地震、火灾等不可抗因素,硬盘划伤、网络故障、硬件损坏等硬件故障,Bug、内存溢出、崩溃等软件故障,误操作、误删除、恶意攻击等人为因素[3-4],都将造成数据故障或丢失且不可避免,软硬件越多,应用越多,发生概率越大,损失越大。在此现实下,2010年前,有80%的大企业已采取数据存储持续保护设备[5-7]。由此可知数据存储对企业的重要。

目前军工型号产品在线协同设计、数字化制造的应用不断扩大和深入,各分系统和单机单位间业务协同越来越多,由此带来军工集团的计算机集群规模越来越大,业务应用数据越来越大,数据故障或丢失对业务应用系统损失越来越大,对数据存储持续保护、容灾安全、读写速度的要求越来越高,急需对现有存储系统进行升级或开发新型存储方案。

1 现有方案

现在某军工集团下辖十余个厂所,计算机集群中有上万个客户端、几百台数据库服务器和应用服务器集群,数据达PB量级,其数据存储方案采用将所有服务器通过FC(Fibre Channel)光纤交换机与SAN(Storage Area Network)统一存储相连,对外提供服务,其总体方案如图1所示。

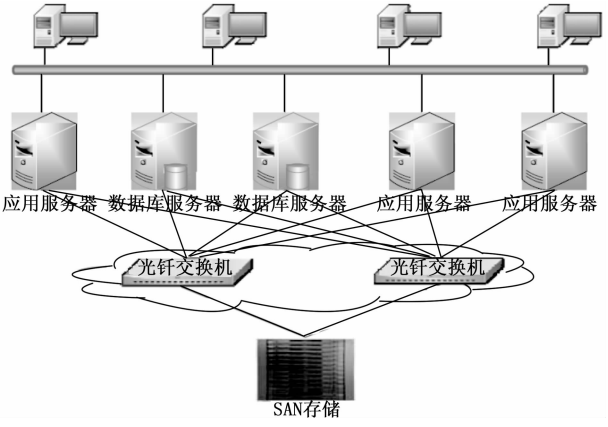

图1 现有存储方案

所有数据库服务器和应用服务器均配置以太网卡和光纤HBA(Host Bus Adapter)卡,以太网卡接入以太网络,为前端用户提供服务,光纤HBA卡通过FC光纤通道交换机与后端SAN统一存储相连,所有数据均存放在后端统一存储中。

存储系统为基于FC光纤通道的SAN统一存储,所有数据均存放在同一台存储系统中,严重拖慢存储系统性能,并进一步影响到前端业务系统。大量非核心数据占用大量存储空间,消耗大量存储系统性能,存储系统无法为前端业务系统提供足够存储空间和存储性能;加上利用数据库自身功能备份数据对存储资源的消耗,无法为前端业务系

收稿日期:2018-10-16; 修回日期:2018-10-30。

作者简介:曾 发(1985-),男,江西萍乡人,硕士,工程师,主要从事数据管理总体设计方向的研究。

统提供足够 I/O 支撑保障。存储系统未对数据保护，业务系统发生损坏、数据丢失、数据逻辑错误等问题，都将给业务系统带来极大影响，甚至整个系统崩溃。目前，其厂所级应用系统瘫痪频率已达到每月数次，故障修复长达数小时甚至一天，期间很多工作只能中断，给集团业务带来极大损失。

2　总体方案

针对现有存储方案存在问题和未来大数据业务应用需求，本文面向大型计算机集群业务应用，充分利用现有存储设备，增加一套备份存储硬件，设计一套数据块级的大数据持续保护系统（Big Data Continuous Protection System，简称 BDCPS），在不影响前端应用前提下，捕获、跟踪数据变化，进行实时备份并能恢复到此前任意时间点，实现数据保护，防止数据发生损坏或丢失，并剥离出部分非核心数据到新存储平台，以减轻业务系统存储压力，释放存储资源，保障前端应用高效运行，其总体方案如图 2 所示，由图可见系统配置简单、结构清晰，各服务器、现有 SAN 存储系统、BDCPS 均接入机房 LAN（Local Area Network）网络与 SAN 网络中。

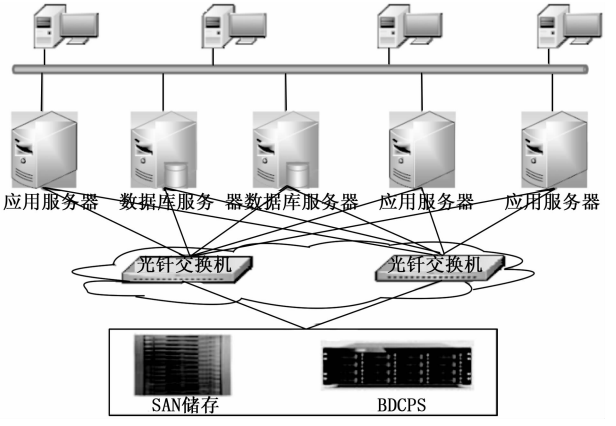

图 2　BDCPS 总体方案

BDCPS 同时支持 Unix、Linux 和 Windows 平台，通过部署其中的模块，为整个计算机集群中的操作系统、数据库、正运行关键业务的服务器提供快速、可靠、完全的数据备份和恢复服务。系统保护业务数据同时持续保护数据库服务器数据，保证备份数据与业务应用主机数据完全一致，当业务应用出现故障，通过数据挂载恢复业务。系统定时将非核心业务应用数据备份至备份存储设备中，保证重要数据丢失或损坏后能及时恢复数据。新增存储存放非核心数据，并承载一些测试、统计、数据分析，避免所有工作都由业务系统主存储负载。根据上述功能需求，将 BDCPS 划分为数据定时备份、持续数据保护和数据查询统计三个主要功能模块。

BDCPS 内部组成结构见图 3，其中：文件数据捕获、跟踪子模块负责截获业务应用层或文件层传输过来的写操作及修改信息，再把截获的信息传递给标记子模块；磁盘块数据捕获、跟踪子模块负责截获物理磁盘层上数据块的写操作及修改信息，再把截获的信息传递给标记子模块；标记子模块包括应用标记和事件标记，前者标记此次截获的信息属于哪个文件以及属于哪个被保护的应用，后者使用事件对数据进行划分并将事件流与应用标记之后的数据合并，标记后的数据传递给缓存区；缓存区子模块缓存业务系统主机应用层数据或内核层数据，保证截获的数据传递到 BDCPS 服务器前不会丢失；网络子模块采用 SAN 网络双冗余数据传输，能够处理网络故障和数据重发，其数据迁移器组件快速高效将数据传输到 BDCPS 服务器的缓存区；I/O 切换子模块，用于发生故障时，实现存储切换，将业务系统主机磁盘读写重定向到虚拟磁盘，即通过网络向 BDCPS 服务器读写数据，直到业务系统主机数据恢复完成，虚拟卷中数据与磁盘数据完全同步，I/O 切换为原来的状态；多级缓存子模块，BDCPS 服务器采用大容量内存和 SSD（Solid State Drives）固态硬盘相结合的多级缓存方式以满足大量业务系统主机数据缓存；存储子模块，用于数据存储、管理、查询、统计等以及存储池中物理磁盘管理，并对冗余数据进行去重删除，减少数据存储所需空间；存储池，数据存储的物理磁盘，用于业务系统主机数据备份存储，包括定时备份和持续保护实时备份，并形成影子副本（Shadow Copy）；远程备份子模块，将数据定时或实时备份到异地的 BDCPS 服务器，可根据需要部署，本文没有进行远程备份。文件数据捕获、跟踪子模块和磁盘块数据捕获、跟踪子模块是 BDCPS 的关键子模块，提供业务系统主机中的数据变化部分，下文提到的数据分流器组件即属于这两个子模块，这两个子模块与标记子模块组合，根据不同数据恢复算法，形成快照模块，获得被保护数据的一致性快照，并以虚拟磁带库形式存储于 BDCPS 服务器，发生故障时，其逻辑卷通过网络子模块，直接挂载到业务系统主机上的虚拟磁盘。

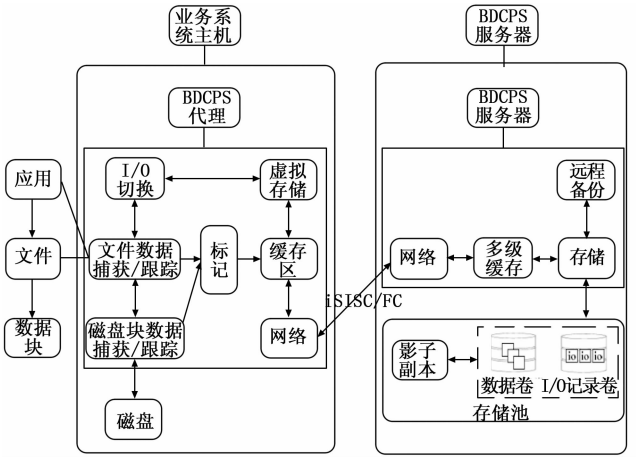

图 3　BDCPS 内部组成结构

数据定时备份、持续数据保护和数据查询统计三个主要功能模块，则由上述子模块的不同组合来实现，其部署设置和工作机理见下文详述。

根据部署位置划分，BDCPS 分为 BDCPS 代理和 BDCPS 服务器两大部分。BDCPS 代理部署在需保护的业务系统服务器上，实时监控服务器上所存数据变化，并把监控信息、标记信息发送给 BDCPS 服务器存储起来。当故障发生时，BDCPS 代理负责存储切换和数据恢复。BDCPS 服务器主要用来存储数据，并随时准备提供一个完整的数据版本给 BDCPS 代理使用。

BDCPS 将每个磁盘即逻辑单位划分为固定大小的数据块，根据数据保护的恢复粒度大小，数据块大小可以设置为 4 KB、8 KB、16 KB、32 KB、64 KB，并以数据块为单位记录磁盘中的数据变化，块大小设置越小，数据恢复的粒度越小，但相应块数量越多，需读写的块总数越多，效率降低。BDCPS 自动将业务系统主机中磁盘更改过的所有数据按时间顺序保存下来，每次写操作都会生成带有时间戳的数据块版本，并形成 I/O 记录，在业务系统主机恢复数据时，能够获取任意一个时间的数据状态。

BDCPS 采用虚拟化技术统一管理物理磁盘阵列，将其虚拟化成逻辑存储池，根据制定的存储池配置方案来对存储池进行划分，形成一块块指定大小的逻辑卷，以便统一管理分配，更大化地利用存储空间。

2.1 数据定时备份

为保证系统数据安全，对各重要和核心业务应用服务器都进行数据备份，按如下步骤设置：不改变现有方案结构，将 BDCPS 接入 LAN 网络与 SAN 网络中；在需备份服务器中安装相应系统版本的代理模块；备份数据通过虚拟磁带库存储至 BDCPS 中。

部署完 BDCPS 后，设定备份策略自动完成备份任务，其工作流程如下：定时备份模块根据备份策略，向各备份源发送备份任务，备份源上的相应代理接到任务后，抓取备份数据，移交到数据迁移器，迁移器将数据通过 SAN 网络以虚拟磁带方式保存至 BDCPS 自身存储磁盘中。当备份源数据损坏时，可通过 BDCPS 选择已备份的任意时间点进行恢复。

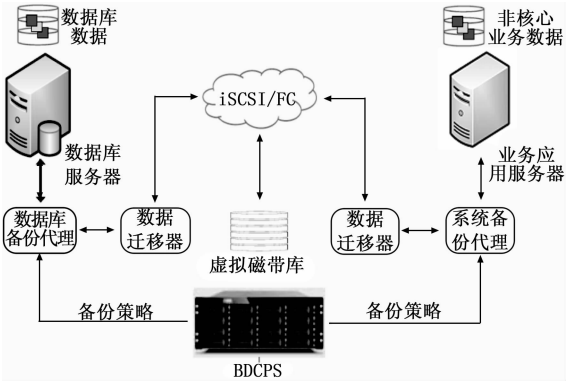

图 4 数据定时备份结构

对于现有数据库服务器和应用服务器集群中前期存储的部分非核心数据，数据迁移备份到 BDCPS 自身存储磁盘后，即根据策略删除现有服务器集群中的原数据，以释放现有业务系统的存储资源，可以在不停止现有业务处理的情况下，有效增加用于处理核心业务数据的存储资源，减轻服务器存储压力，提升其运行速度，重新恢复前端应用的高效运行，需要使用这些数据时，再从 BDCPS 中迁回业务系统服务器。

2.2 持续数据保护

对集群中各在线业务应用服务器及数据库服务器进行持续数据保护，通过以下方式实现：不改变现有方案结构，将 BDCPS 接入 LAN 网络与 SAN 网络中，激活持续数据保护模块、快照模块；在需持续数据保护的服务器中安装持续数据保护数据分流器，通过持续数据保护模块实现核心业务的持续数据保护功能；持续数据保护的数据存储于 BDCPS 自身存储磁盘中。

如图 5 所示，BDCPS 部署完成后，系统自动将需保护服务器进行有效持续数据保护，在数据写入被保护服务器自身存储设备同时，写入 BDCPS 中，保证 BDCPS 连接中的数据与被保护数据完全一致，并生成每个 I/O 记录点和一致性快照。利用快照功能，可进行连续的或基于时间点的快照工作，当被保护服务器发生逻辑错误时，快速有效挂载快照点，避免逻辑错误造成数据损坏。

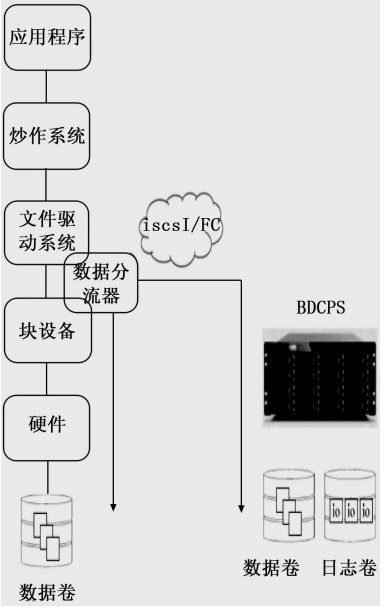

图 5 数据持续保护结构

BDCPS 针对大型计算机集群业务应用中大数据、高并发特点，数据持续保护特点，采用大容量高速缓存和 SSD 多级缓存架构提高写入性能，利用多核处理器技术和并行队列处理算法，提高数据持续保护的速度，降低其对业务系统存储性能的影响。

2.3 数据查询统计

数据查询统计主要基于 BDCPS 内置的影子副本挂载功能实现，利用影子副本即时挂载技术，可不停止数据持续保护，直接将多个历史数据状态点瞬间挂载到不同主机或虚拟机，与 BDCPS 磁盘组同时读写访问、分别回滚，其工

作方式见图 6。

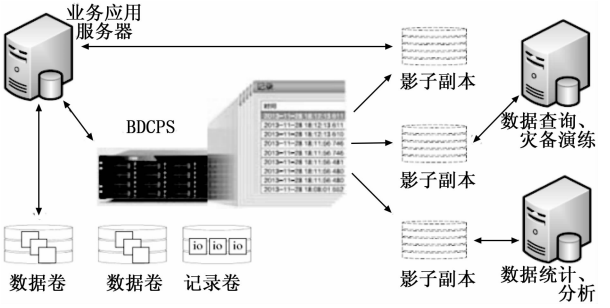

图 6　数据查询统计结构

影子副本可极大改善数据持续保护的易用性，包括灾备演练、恢复、测试和数据统计、分析、再利用体验，并简化操作步骤，相当于将不同时间点的快照数据，虚拟生成若干个磁盘组，以方便查询最符合要求的数据用于恢复。当不需要这些数据挂载点时，可随时删除，不影响原有数据持续保护任务保护的数据。影子卷产生的增量变化数据不被保留，占用磁盘空间将释放回存储池。

若选择将数据持续保护卷组先进行设备上或设备间的复制，当副本数据独立存放于单独 RAID 磁盘，此时再使用影子副本功能做读写分离或查询等，适用于高 I/O 负载的业务，同时保持不停止保护和复制。

3　性能测试

3.1　试验配置

为方便测试，本文对系统进行简化，选用 4 台客户端计算机作为业务系统主机、1 台服务器作为 BDCPS 服务器，组成测试网络，分别安装 BDCPS 代理、BDCPS 服务，主要硬件配置见表 1～表 2。

表 1　业务端硬件配置

名称	规格/性能
CPU	Inter CPU I5 2400 @3.1GHz 3.1GHz
内存	Kingston DDR4 4GB
硬盘	Seagate SATA 6Gb/s 7200r/min 300GB
网络带宽	1000Mbps

表 2　BDCPS 服务器硬件配置

名称	规格/性能
CPU	Inter CPU E5 2667 v4 @3.2GHz 3.2GHz
内存	Kingston DDR4 32GB
硬盘	Inter S4510 SSD 240GB＋Seagate SAS 6Gb/s 7200r/min 4×1.8TB
网络带宽	1000Mbps

3.2　测试试验

限于篇幅，本文选取数据持续保护中数据恢复性能这一重要指标作为考核对象，试验主要测试 BDCPS 在不同数据量和不同业务主机数从 BDCPS 服务器获取数据的数据恢复时间，并设定被保护块设备大小为 16 KB。恢复模式分别采用全量恢复、增量恢复[8]，全量恢复算法基本思路：设数据保护初始时刻为 t_0，指定目标恢复时间点为 t_T，找出从 t_0 到 t_T 分支上所有写操作记录，最后在相同源地址的多个写操作记录中，取时间戳最接近 t_T 的那个记录；增量恢复算法基本思路：设数据保护初始时刻为 t_0，指定目标恢复时间点为 t_T，当前时刻 t_c，先查找出 t_T 时刻到 t_c 时刻发生改变的数据扇区，再查找这些扇区从 t_0 到 t_T 分支上所有写操作记录，最后在相同扇区号的多个写操作记录中，取时间戳最接近 t_T 的那个记录。对单台主机，测试恢复数据量分别为 10 MB、50 MB、200 MB，对比两种恢复模式从 BDCPS 服务器获取数据的数据恢复时间，结果见表 3。对多台主机，测试 4 台主机并发恢复性能，结果见表 4，以及 4 台主机平均恢复时间和恢复延迟率，结果见表 5。

表 3　单台主机不同恢复模式性能对比

恢复数据大小/MB	全量恢复时间/s	增量恢复时间/s
10	2	1.2
50	3.5	2
200	6	3.6

表 4　多主机不同恢复模式性能对比

恢复数据量	全量恢复时间/s			增量恢复时间/s		
/MB	10	50	200	10	50	200
主机 1	3	4.8	8.3	1.8	2.4	4.5
主机 2	2.4	3.6	7.6	3.5	4.2	5.8
主机 3	3.6	4.8	8.9	3	3.2	4.2
主机 4	3	5.4	8.8	3	4.2	5.3

表 5　多主机平均恢复性能及服务延迟率

恢复数据量	平均恢复时间/s			服务延迟率/s		
/MB	10	50	200	10	50	200
全量恢复	3	4.8	7.2	32	29	23
增量恢复	3	3.6	4.8	46	42	32

3.3　试验分析

由表 3 可知，增量恢复快过全量恢复，因为全量恢复需恢复整个被保护块设备，因此恢复时间随恢复数据增加而延长。而增量恢复只是在当前状态下往前回滚，与恢复数据量无直接关系，所以恢复时间随恢复数据增加而变化不大。由表 4 和表 5 可知，多主机并发恢复比单主机恢复时间稍长，主要是 BDCPS 服务器并发检索各主机数据需更多读磁盘开销，并需占用较多 CPU 资源，而每个主机获得的 CPU 时间片少了，对恢复时间也有一定影响。定义每次并发恢复中最慢和最快的时间差占整个恢复时间的比率为服务延迟率，由表 5 还可知，服务延迟率都低于 50%，并随恢复数据增加，偶然因素的影响相对变小，因此延迟率进

一步减小。

4 结论

本文面向大型计算机集群业务应用，针对现有数据存储方案面临的问题，设计一套大数据持续保护系统，通过简单增加一套多功能一体化设备，不改变现有方案硬件结构，就实现对数据定时备份、持续数据保护、查询统计，具有如下优势：

1）能把任何造成数据损坏的问题得到妥善解决，对硬件故障、误操作造成的逻辑错误、单个文件丢失、站点级灾难皆有完备保护能力，并将数据定时备份，使其坚如磐石；

2）无须人工关注数据保护过程，自动实现数据保护，持续捕捉和跟踪数据发生的任何改变，能将数据恢复到任意时间点，RPO[9]可接近 0，RTO 可达秒级，几乎没有数据丢失，灾难发生后数据恢复无需中断业务；

3）集成查询统计功能，通过影子副本将数据挂载到前端查询、统计服务器，大大提高查询检索效率，且对业务应用主存储不带来性能上损耗，查询时无需中断数据保护操作；

4）迁移集群中部分非核心数据，释放现有存储资源，减轻服务器压力，提升其运行速度，并使用 SSD 缓存加速存储，构建灵活缓存，加速读写数据，提高集群的数据存储读写速度。

参考文献：

[1] 张妤芝. 基于 Windows 的持续数据保护系统的研究与实现 [D]. 上海：上海交通大学，2010.

[2] Wallis J. Common causes for data loss [EB/OL]. 2008. Http：//www. isnare. com.

[3] 赵瑞君. 基于网络的连续数据保护系统设计与实现 [D]. 武汉：华中科技大学，2012.

[4] 李祚衡. 块级连续数据保护技术的研究 [D]. 华中科技大学，2008.

[5] 刘正伟. 海量数据持续数据保护技术研究与应用 [J]. 济南：山东大学，2011.

[6] Kuhn D R，Reilly M J. An investigation of the applicability of design of experiments to software testing [A]. Proceedings of the 27th NASA/IEEE software Engineering Workshop [C]. NASA Goddard Space Flight Center 2002.

[7] Wang SP，etc. Continuous data protection technology overview. Information technology letter，2008，6 (6)：24－33.

[8] Xu L，Xie C S，Yang Q. Optimal implementation of continuous data protection in Linux Kernel [A]. NAS 08th International Conference on Networking，Architecture，and Storage [C]. 2008. 28－35.

[9] Damoulakis J. Continuous protection [J]. Storage Magazine，2006，3 (4)：33－39.

(上接第 158 页)

4 结束语

随着无人机技术的进步，全自动化的无人机精细化巡检杆塔逐渐变得可行，而巡检线路规划是无人机自动化巡检过程中重要的一环，关乎巡检的效率和安全性。本文提出了基于两阶段优化确定无人机巡视杆塔考虑避障的路径优化方法。该方法通过动态规划算法优化无人机经过杆塔的次序，通过人工势场法规划无人机在两个杆塔之间避开障碍物的飞行路径。该方法确定的巡视路径具有巡视路径短、无人机在巡视过程中转向少且转向角度小、能按照安全距离要求有效避开障碍点的特点。

参考文献：

[1] 邓荣军，王 斌，熊 典，等. 基于遗传算法的输电线路无人机巡检路径规划 [J]. 计算机测量与控制，2015，23 (4)：1299－1301.

[2] 张 剑，王世勇，陈 玺，等. 基于柱状空间和支持向量机的无人机巡线避障方法 [J]. 中国电力，2015，48 (3)：56－60.

[3] 熊 典. 输电线路无人机巡检路径规划研究及应用 [D]. 武汉：武汉科技大学，2014.

[4] 黄俊璞，林 韩，宋福根，等. 输电线路上方无人机巡检避障策略 [J]. 电器应用，2015 (23)：32－34.

[5] 施孟佶，秦开宇，李 凯，等. 高压输电线路多无人机自主协同巡线设计与测试 [J]. 电力系统自动化，2017，41 (10)：117－122.

[6] Qu Y，Zhang Y，Zhang Y. A UAV solution of regional surveillance based on pheromones and artificial potential field theory [A]. International Conference on Unmanned Aircraft Systems [C]. IEEE，2015.

[7] Oyekan J，Hu H. Towards Autonomous Patrol Behaviours for UAVs [A]. Proceedings of Uk Epsrc Workshop on Human Adaptive Mechatronics [C]. Staffordshire，2010.

[8] Wei W L，Ding J，Huang J F. Brief analysis on the airline planning of fixed wing UAV power line patrol [J]. China Electric Power，2016.

[9] Sujit P B，Saripalli S，Sousa J B. Unmanned aerial vehicle path following：a survey and analysis of algorithms for fixed－wing unmanned aerial vehicles [J]. IEEE Control Systems，2014，34 (1)：42－59.

[10] Bortoff S A. Path planning for UAVs [A]. 2000. proceedings of the American control conference [C]. IEEE，2002 (1)：364－368.

[11] Tisdale J，Kim Z，Hedrick J K. Autonomous UAV path planning and estimation [J]. IEEE Robotics & Automation Magazine，2009，16 (2)：35－42.

[12] Bellingham J，Tillerson M，Richards A，et al. multi－task allocation and path planning for cooperating UAVs [J]. Cooperative Systems，2003，1：23－41.

[13] Pehlivanoglu Y V. A new vibrational genetic algorithm enhanced with a Voronoi diagram for path planning of autonomous UAV [J]. Aerospace Science & Technology，2012，16 (1)：47－55.

[14] 丁家如，杜昌平，赵 耀，等. 基于改进人工势场法的无人机路径规划算法 [J]. 计算机应用，2016，36 (1)：287－290.