DOI: 10. 3880/j. issn. 1004-6933. 2019. 02. 015

基于新一代大数据处理引擎 Flink 的“智慧滁河”系统

叶 枫[1,2],张 鹏[3],夏润亮[4],顾和生[5],陈 勇[2]

(1. 河海大学计算机与信息学院, 江苏 南京 211100; 2. 南京龙渊微电子科技有限公司,江苏 南京 211106;
3. 江苏省水利厅,江苏 南京 210029; 4. 黄河水利科学研究院,河南 郑州 450003;
5. 南京市江北新区环境保护与水务局,江苏 南京 210032)

摘要: 概述了水利领域大数据的特点,展示了基于 Flink 构建的“智慧滁河”系统,并以在滁河监测中获取的传感器数据为实验数据,以常用的查询操作测试了系统性能。结果表明,采用 Flink 的“智慧滁河”系统的处理能力远超传统的多层架构系统,可为水利信息化迈向“智慧”提供可行的解决方案。

关键词: 智慧水利; Flink; 大数据; 水利信息化

中图分类号: P333. 9 **文献标志码**: A **文章编号**: 1004-6933(2019)02-0090-05

“**Smart Chuhe River” system based on Flink, a new generation of big data processing engine** // YE Feng[1,2], ZHANG Peng[3], XIA Runliang[4], GU Hesheng[5], CHEN Yong[2] (1. *College of Computer and Information, Hohai University, Nanjing* 211100, *China*; 2 *Postdoctoral Centre, Nanjing Longyuan Micro-Electronic Company, Nanjing* 211106, *China*; 3. *Jiangsu Water Resources Department, Nanjing* 210029, *China*; 4. *Yellow River Institute of Hydraulic Research, Zhengzhou* 450003, *China*; 5. *Environmental Protection and Water Affairs Bureau of Nanjing Jiangbei New Area, Nanjing* 210032, *China*)

Abstract: The characteristics of big data in water conservancy were summarized, and a “Smart Chuhe River” systems based on Flink was demonstrated. Taking the sensor data obtained from Chuhe River monitoring as experimental data, the system performance is tested by common query operations. The results show that “Smart Chuhe” system based on Flink has far more processing power than the traditional multi-tier framework system, providing feasible solutions for water conservancy informatization towards “smart”.

Key words: smart water conservancy; Flink; big data; water conservancy informatization

IBM 于 2008 年率先提出了“智慧地球”,期望通过普适的数字化、网络化和智能化,提高效率、灵活性,做出更加明智的决策。“智慧”体现在 3 个方面:“更透彻的感知,更广泛的互联互通,更深入的智能化”[1]。所谓“更透彻的感知”是指利用任何可以随时随地感知、测量、捕获和传递信息的设备、系统或流程;“更广泛的互联互通”是指通过各种形式的高速宽带通信网络,将个人电子设备、组织机构等信息系统中收集和储存的分散信息及数据连接起来,进行交互及多方共享,从而更好地对环境和业务状况进行实时监控,从全局的角度分析并实时解决问题,使得工作和任务得以通过远程的、多方协作方式完成;“更深入的智能化”指使用先进技术(如数据挖掘工具、科学模型)来完成复杂的数据分析、汇总和计算,整合和分析跨地域、跨行业和职能部门的海量数据和信息,并应用到特定行业、场景以及解决方案中,以更好地支持决策和行动,如“智慧医疗”“智慧城市”和“智慧交通”等。由于水利业务的复杂性、动态性,涉及因素(天气、地形、人类活动等)的多样性及关联性,导致水利业务数据也呈现多样性、动态性、大数据规模化等特点,这为水利领域智慧化的研究带来诸多挑战,当前仍需要坚实的理论、恰当的模型、有效的范例、深入的试验和具体的实现。从信息科学和技术的角度看,主要有 2 个关键问题:①实现“智慧”,从根本上解决数据密集型科学发现的问题[2-4],这既有科研传感器数据集的海量

基金项目: 2013 年江苏水利科技项目(2013025); 国家科技支撑计划(2013BAB05B00); 2017 江苏省“六大人才高峰”项目(XYDXX-078); 2017 江苏省博士后科研资助计划(1701020C); 2018 江苏省重点研发计划(30185057012)

作者简介: 叶枫(1980—),男,讲师,博士,主要从事云计算、水利领域大数据研究. E-mail: yefeng1022@hhu.edu.cn

性、多样性、丰富性、不确定性、实时性带来的挑战，也有对水利领域数据内在的时空特性和特定属性理解和处理的困难；②实现“智慧化”的河流、流域，验证所提出的“智慧河流”等的可行性、有效性，需要选型并构建面向水利领域大数据的获取、处理、分析的综合平台，以有效应对大数据多样化的处理场景。本文提出了基于新一代大数据处理引擎 Flink 的“智慧滁河”系统，旨在为水利信息化迈向“智慧”提供可行的解决方案。

1 相关概念

“智慧河流”“智慧流域”和“智慧水利”等，实质是“智慧地球”理念在水利领域的延伸。当前，对于水利领域智慧化的研究，代表性的工作有蒋云钟等[5-6]提出的“智慧流域”和王忠静等[7]提出的水联网和智慧水利的概念。蒋云钟等[5-6]提出，智慧流域是指把新一代 IT 技术充分运用于流域综合管理，把传感器嵌入和装备到流域各个角落的自然系统和人工系统中，通过普遍连接形成“流域物联网”；而后通过超级计算机和云计算将“流域物联网”整合起来，以多源耦合的气象水文信息保障平台、二元水循环及伴生过程数值模拟平台等为支撑，将其与数字流域耦合起来，完成数字流域与物理流域的无缝集成，使人类能以更加精细和动态的方式对流域进行规划、设计和管理，从而达到流域的“智慧”状态。王忠静等[7]认为，水联网的总体架构是集物理水网、虚拟水网和市场水网为一体的现代化水资源系统。在实现层面，需要物联网、云平台、服务体系来探索智慧化水利之路，已成为当前的研究共识，并已做了一些具体工作[8-9]。但是，从信息科学和技术的角度看，实现“智慧”的关键标准是从根本上是解决数据密集型科学发现的问题，目前缺乏具体完整的系统架构方案或典型案例。由于水利业务的复杂性、动态性等特点，对大规模的动态、实时、多样化水利业务数据的处理存在诸多挑战。因此，应把握“智慧”的特征，根据水利特定领域大数据驱动的主线，选择合适的平台和技术进行集成，构建智慧化应用系统。

2 水利领域大数据

实现“智慧”，从根本上说是解决数据密集型科学发现的问题，数据本身是核心，涉及数据的获取、传递、存储、处理和展现等流程。水利领域大数据最显著的特点可以概况为规模大、主题多样、数据本身的信息丰富、处理和利用难度大。根据冯钧等[10]的研究，经过长期的业务实践，水利领域已经积累了大量分布异构独立的业务数据，如，实测信息就包括水文观测信息(地表地下水量水质状态等信息)、水利设施在线运行状态信息、用水户用水排水信息等，截至 2012 年，单是水文数据，全国已超 100TB。随着各类水文、水质传感器和摄像头等的广泛应用，所获取数据的规模和增长速度都是空前的。从主题角度，有水文、水质[11]、水资源、水利设施(空间)、土壤侵蚀、灌溉、水能资源调查、农村水电等专题，以及第一次水利普查工程对河流湖泊、水利工程、重点经济社会取用水户及水利单位等对象进行普查和清查汇总形成的普查成果数据。这些主题不断丰富着水利领域的大数据集。以水资源数据为例，水利信息往往具有以下特征[12]：存在异常数值，连续观测的值往往是彼此密切相关的，依赖于其他变量等。水利领域数据的处理和利用难度很大，主要是因为数据获取过程中对位置、环境、天气等相关因素的数据系统标注缺失，丢失了数据之间存在的丰富时空关联信息。水利领域的业务数据迥异于商业数据，其数据类型丰富、规模大，属性和关联关系也更加复杂[13]，因此，研究并选择合适的大数据处理平台和技术成为构建智慧化系统的关键。

3 基于 Flink 的体系架构

3.1 新一代大数据处理引擎 Flink

Apache Flink[14-16]是由欧洲的多名研究者和多家资助单位联合研发的一款开源的并行化数据分析软件，现已成为 Apache Software Foundation 的顶级项目。Flink 本质是一个流式计算引擎，在同一个运行时(Runtime)，分别搭建了流式计算和批处理的编程接口和相配套的生态系统，其体系结构图见图 1。

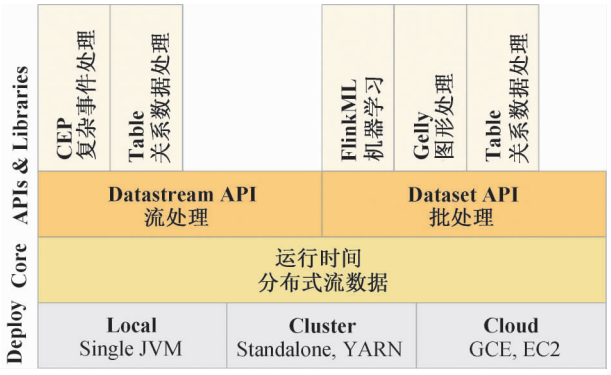

图 1 Apache Flink 的体系架构

Apache Flink 具有的特性是：①既能支持高吞吐、低延迟、高性能的流处理操作，也可以用于批数据的处理；②支持有状态计算的 Exactly-once 语义；③具备基于轻量级、分布式快照(Snapshot)实现的

容错机制;④支持高度灵活的窗口(Window)操作,包括 Time、Count、Session 以及 Data-driven 的窗口操作;⑤在 JVM 内部实现了自身的内存管理;⑥支持迭代计算;⑦支持机器学习(FlinkML)、图分析(Gelly)、关系数据处理(Table)和复杂事件处理(CEP);⑧支持 Flink on YARN、HDFS、Kafka、Apache HBase、Hadoop、RabbitMQ、S3 以及 XtreemFS 等大数据相关软件。

相比于 Spark、Storm 等大数据平台,Apache Flink 被认为是第四代,也是最新一代的大数据处理引擎。文献[17]通过 2 个试验评估吞吐效率和节点失效下的弹性(Resilience),结果表明,Flink 比 Spark Streaming 快 15 倍。Chintapalli 等[18]开发了一个流处理 Benchmark,用于测评 Flink、Storm 和 Spark Streaming。通过构造数据管道机制,最大程度地模拟了真实世界的数据流场景。结果表明,Flink、Storm 性能近似,对于流数据的响应近似线性;而 Spark Streaming 虽吞吐量较大,但延迟较高。基于 3 个不同数据集和不同的算法,文献[19]提出了一个用于比较 Apache Flink 和 Apache Spark 的 Benchmark,结果表明 Apache Flink 在数据挖掘和图处理方面比 Apache Spark 更优。可以看出,Flink 的优势在于其从上到下提供了一整套完整的、针对大数据的栈式解决方案,并为用户提供了易于使用的数据分析系统。

3.2 “智慧滁河”系统的体系架构

“智慧滁河”系统的体系结构分为 5 层:感知识别层、网络构建层、基础设施层、平台与服务层以及应用层,见图 2。感知识别层由 3 部分组成,一是多源数据的感知、采集机制,如:RFID、视频探头、全球

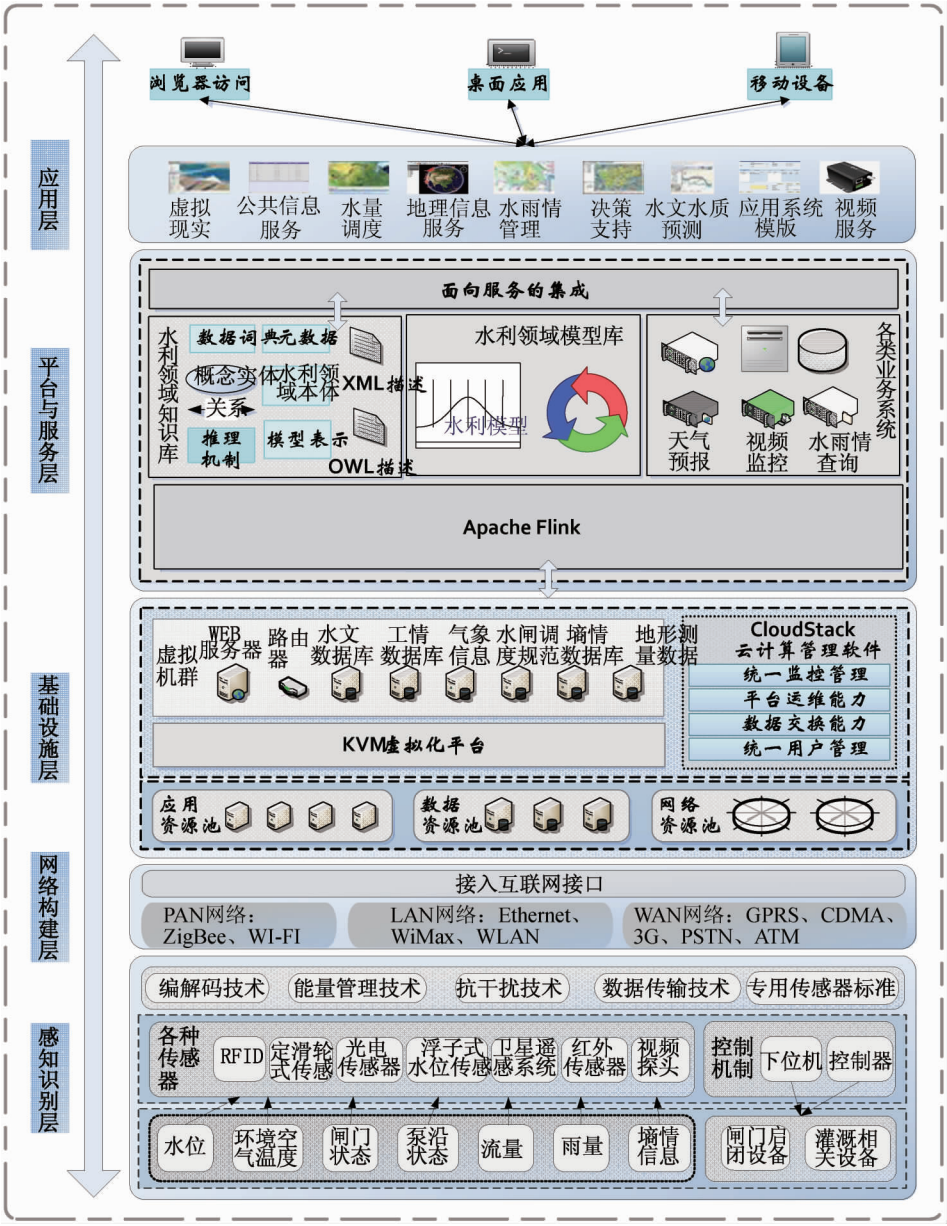

图 2 “智慧滁河”的体系结构

定位系统、遥感、各类传感器(水位、水质、雨量等);二是控制机制,用于交互下位机和控制器,并与网络构建层通信;三是能量管理、抗干扰、编解码等机制。如,将所监测的河道沿程的水位、流速、水质以及视频监控摄像头自动获取监测闸门的开度等传给网络构建层,也会接受反馈的指令,对传感器、下位机和控制器进行操作,进而调整传感器,操控闸门、灌溉设备等。网络构建层是实现物-物互联的重要基础,主要由各类网络组成,包括各种通信网络、互联网形成的融合网络。该层的作用是把感知识别层所采集的数据接入网络。以传感器数据为例,各种监测数据通过无线传感器网络形成局部区域网络,将覆盖区域的信息收集起来,然后通过网络构建层自动向位于基础设施层的数据中心传输数据流。不难看出,网络构建层是"更广泛的互联互通"的实现基础。基础设施层、平台与服务层以及应用服务层均基于以云计算[20]为代表的大规模分布式处理平台之上。基础设施层与云计算模型的 IaaS(基础设施即服务,Infrastructure as a Service)层是相对应的,主要包括:用于管理、存储水利领域大数据的存储机制(关系数据库,NoSQL[21-22]数据库等)、虚拟机群、网络资源,该层可由开源或商业的云计算解决方案实现,如 Apache CloudStack、OpenStack 等。用户可以按需整合资源以适于不同的业务处理场景。

平台与服务层包括诸多相关的业务系统、模型库和知识库等,如各类数据存储机制,知识库包含灌区相关的专家知识,模型库包括系统涉及的计算模型(如需水预报模型)。通过对各类业务系统、模型库和知识库进行集成,实现具体的业务服务、中间件或服务工作流,以供上层应用软件调用。Apache Flink 为构建平台层提供最为直接的支持,它能为业务逻辑提供并行化的处理机制,也能访问基础设施层中的数据资源等。在正确部署 Apache Flink 的基础上,基于 Apache Flink 提供的流数据 Datastream API 或 Batch Processing API,编程先获取执行环境,然后通过各种连接器(connectors)获取相应的数据源,接下来利用各种转换函数(Map、KeyBy、Reduce、Window 等)对数据进行处理,最后将处理的数据保存。

处于整个体系结构最上层的是应用服务层。该层主要访问界面或接口,承担计算和分析结果展示的任务。它既可以通过各种应用软件调用平台与中间件层的服务,也可以调用来自第三方提供的各类服务(百度路径导航服务、地震信息查询服务、天气预报服务等),将结果和交互界面展示给用户。

4 系统验证

选取 2015 年 1 月 1 日至 2017 年 6 月 30 日的滁河实时水位数据集,一共有 18 910 865 条数据记录。运行环境是是由 3 台同型号 PC 组成的集群,配置为:处理器为 AMD Ryzen 7 1700X(八核),内存是 32G 海盗船 DDR4 3000Mhz,硬盘是三星 sm961 的 128G 固态硬盘,显卡为华硕的 GeForce GTX 1060。试验一是查找特定水位值,如查找河道水位高度在"5.5"以上的记录。通过逻辑访问 MySQL 库表的检索方式,需要 8.41 s;通过逻辑访问 MongoDB 的检索方式,需要 7.46 s;而通过 Flink 实现的逻辑访问 MongoDB,检索完成时间只需要 0.03 s左右。试验二是查找最小值,如查找 70 余个监测站点出现最低水位值的记录。通过逻辑访问 MySQL 库表的检索方式,需要 16.2 s;通过逻辑访问 MongoDB 的检索方式,需要 10.3 s;而通过 Flink 实现的逻辑访问 MySQL 或 MongoDB,检索完成时间只需要 0.03 s 左右。试验三是删除数据操作。以 500 万条记录为例,效果更为明显,只需要 3.22 s,远远小于使用 MySQL 的逻辑所需的 122 s。不难看出,通过 Flink 平台提供的 API 实现机制,要比传统的 Java EE 技术所开发的多层架构有更优的性能表现,充分利用了对于集群的并发机制,让大数据处理的方式简单高效。

5 结 语

基于 Flink 的"智慧滁河"系统的计算能力远超传统的多层架构系统,让大数据处理的方式变得简单高效,也为水利信息化迈向"智慧"提供了可行解决方案。后续的研究将集中于研究 Flink 平台上的机器学习算法与水文数据分析的结合,特别是针对传感器流数据的实时分析工作,进一步完善"智慧滁河"系统,为防汛防旱提供更加迅捷的决策支持,丰富水利信息化领域的数据智能化处理的完整案例。

参考文献:

[1] IBM 商业价值研究院. 智慧地球赢在中国[EB/OL]. [2014-10-30]. http://www-31.ibm.com/innovation/cn/think/downloads/smart_China.pdf

[2] HEY T, TANSLEY S, TOLLE K. The fourth paradigm: data-intensive scientific discovery [M]. Redmond Washington: Microsoft Research, 2009.

[3] 孟小峰,慈祥. 大数据管理:概念、技术与挑战[J]. 计算机研究与发展, 2013, 50(1): 146-169. (MENG

Xiaofeng, CI Xiang. Big data management: concepts, techniques and challenges [J]. Journal of Computer Research and Development, 2013, 50(1): 146-169. (in Chinese))

[4] 李国杰,程学旗. 大数据研究: 未来科技及经济社会发展的重大战略领域: 大数据的研究现状与科学思考[J]. 中国科学院院刊,2012(6): 647-657. (LI Guojie, CHENG Xueqi. Research status and scientific thinking of big data [J]. Bulletin of Chinese Academy of Sciences, 2012(6): 647-657. (in Chinese))

[5] 蒋云钟,冶运涛,王浩. 基于物联网理念的流域智能调度技术体系刍议[J]. 水利信息化,2010(4): 1-10. (JIANG Yunzhong, YE Yuntao, WANG Hao. Discussion on intelligent regulation technology architecture for river basin based on interact of things [J]. Water Resources Informatization, 2010(4): 1-10. (in Chinese))

[6] 蒋云钟,冶运涛,王浩. 智慧流域及其应用前景[J]. 系统工程理论与实践,2011,31(6): 1174-1181. (JIANG Yunzhong, YE Yuntao, WANG Hao. Smart basin and its prospects for application [J]. Systems Engineering: Theory & Practice, 2011, 31(6): 1174-1181. (in Chinese))

[7] 王忠静,王光谦,王建华,等. 基于水联网及智慧水利提高水资源效能[J]. 水利水电技术,2013,44(1): 1-6. (WANG Zhongjing, WANG Guangqian, WANG Jianhua, et al. Developing the internet of water to prompt water utilization efficiency [J]. Water Resources and Hydropower Engineering, 2013, 44(1): 1-6. (in Chinese))

[8] 许峰,朱跃龙,叶枫. 水利云平台的研究与应用[C]// 电子政务理事会. 中国电子政务年鉴(2012). 北京: 社会科学文献出版社,2012.

[9] 芮晓玲,吴一凡. 基于物联网技术的智慧水利系统[J]. 计算机系统应用,2012,21(6): 161-163. (RUI Xiaoling, WU Yifan. Intelligent system of water conservancy based on internet of things [J]. Computer Systems & Applications, 2012, 21(6): 161-163. (in Chinese))

[10] 冯钧,许潇,唐志贤等. 水利大数据及其资源化关键技术研究[J]. 水利信息化,2013(8): 6-9. (FENG Jun, XU Xiao, TANG Zhixian, et al. Research on key technology of water big data and resource utilization [J]. Water Resources Informatization, 2013(8): 6-9. (in Chinese))

[11] 于嘉骥,张慧妍,王小艺,等. 基于改进的投影寻踪-云模型的农业灌溉水质综合评价[J]. 水资源保护,2017,33(6): 142-146. (YU Jiaji, ZHANG Huiyan, WANG Xiaoyi, et al. Comprehensive evaluation of agricultural irrigation water quality based on modified projection pursuit - cloud model [J]. Water Resources Protection, 2017, 33(6): 142-146. (in Chinese))

[12] HELSEL D R, HIRSCH R M. Statistical methods in water resources [EB/OL]. [2018-05-20]. http://water.usgs.gov/pubs/twri/twri4a3/.

[13] 黄黎明,张可,龚寻,等. 大数据视角下水利工程质量风险管理[J]. 水利经济,2017,35(6): 66-70. (HUANG Liming, ZHANG Ke, GONG Xun, et al. Management of quality risk of water conservancy projects from perspective of big data [J]. Journal of Economics of Water Resources, 2017, 35(6): 66-70. (in Chinese))

[14] FRIEDMAN E, TZOUMAS K. Introduction to apache flink: stream processing for real time and beyond [M]. Sebastopol: O'Reilly Media, 2016.

[15] DESHPANDE T. Learning apache flink [M]. Birmingham: Packt Publishing, 2017.

[16] CARBONE P, GÁBOR E. HERMANN G, et al. Large - scale data stream processing systems [C]// Handbook of Big Data Technologies. Berlin: Springer, 2017.

[17] LOPEZ M A, LOCATOR A G P, CARLOS O M B. A performance comparison of open - source stream processing platforms [C]//Proceedings of 2016 IEEE Global Communications Conference. New York: IEEE Computer Society, 2016.

[18] CHINTAPALLI S, DAGIT D, EVANS B, et al. Benchmarking streaming computation engines: storm, flink and spark streaming [C]//Proceedings of IEEE 28th International Parallel and Distributed Processing Symposium Workshops. New York: IEEE Computer Society, 2016.

[19] SPANGENBERG N, ROTH M, FRANCZYK B. Evaluating new approaches of big data analytics frameworks [C]// Proceedings of 18th International Conference on Business Information Systems. Berlin: Springer Verlag, 2015.

[20] WANG Lizhe, RAJIV R, CHEN Jinjun, et a1. Cloud computing: methodology, systems and applications [M]. Boca Raton: CRC Press, 2012.

[21] TIWARI S. Professional NoSQL [M]. Indianapolis: John Wiley & Son, 2011.

[22] 陆嘉恒. 大数据挑战与 NoSQL 数据库技术[M]. 北京: 电子工业出版社,2013.

(收稿日期: 2018-05-28　编辑: 彭桃英)