工业互联网高并发流式数据处理技术及应用

朱　抗，何成昭，梁文超，王　超，胡卫民
（株洲中车时代电气股份有限公司，湖南 株洲　412001）

摘　要：随着工业互联网平台接入的设备数量日益增多，大量设备的运行数据以及环境数据源源不断地向远程服务端传输，从而使远程服务端面临着高并发压力以及大量数据实时处理响应时间长的问题。为此，文章提出一种工业互联网高并发流式数据处理方案，其通过构建基于异步网络通信的 Netty 服务器框架，将接收到的具有多源异构和高并发特点的数据流推送至消息处理系统 Kafka 中，再由 Kafka 消费者将协议解析、数据持久化、实时分析等耗时操作放在服务器集群进行分布式计算。实际数据处理应用结果表明，该方法在处理大量设备回传的高并发流式数据时能保持每帧毫秒级的速度，能满足系统对数据处理的需求。

关键词：工业互联网；高并发；数据流；Netty；Kafka
中图分类号：TP393.09　　**文献标识码：**A　　**文章编号：**2096–5427(2021)05–0020–06
doi:10.13889/j.issn.2096–5427.2021.05.200

High Concurrent Streaming Data Processing Technology and Its Application in Industrial Internet

ZHU Kang, HE Chengzhao, LINAG Wenchao, WANG Chao, HU Weiming
(Zhuzhou CRRC Times Electric Co., Ltd., Zhuzhou, Hunan 412001, China)

Abstract: With the increasing number of devices connected to the industrial internet platform and a large number of equipment operation data and environment data are continuously transmitted to remote servers, remote servers are facing the problem of high concurrency pressure and long response time of real-time processing of a large amount of data. Therefore, this paper proposes a high concurrency streaming data processing scheme for industrial internet. By constructing Netty server framework based on asynchronous network communication, it pushes the received data stream with the characteristics of multi-source heterogeneous and high concurrency to the message processing system Kafka, then Kafka consumers put the time-consuming operations such as protocol parsing, data persistence and real-time analysis on the server cluster for distributed computing. Actual data processing application results show that this method can maintain the speed of milliseconds when dealing with a large number of high concurrent streaming data returned by devices, and can meet the needs of the system for data processing.

Keywords: industrial internet; high concurrency; data stream; Netty; Kafka

0　引言

工业互联网是基于工业设备的实时、高效数据采集体系，其通过端到端的数据集成与建模分析，最终实现设备的智能化控制[1]。自 2012 年国际工业巨头 GE 公司提出工业互联网的概念并发表了工业互联网白皮书以来，工业互联网相关的概念和技术开始快速普及。国外，GE 公司推出工业互联网操作系统 Predix，西门子公司推出工业互联网平台 MindSphere，ABB 公司推出了 ABB Ability 数字化解

收稿日期：2021–05–19
作者简介：朱抗（1983—），男，工程师，主要从事新能源信息化系统技术研究。

决方案 [2-3]；国内，华为技术有限公司、三一集团、海尔集团、中国航天科工集团有限公司等分别推出了 FusionPlant、树根互联、COSMOPlat 及 CMSS 等工业互联网平台，这使大量工业设备具备了与工业互联网平台互联的基础 [4-5]。

随着工业互联网上大量设备的接入，传回系统后台的数据呈现爆发式增长的态势，不仅数据体量大，而且数据种类多 [6]。工业数据的爆发式增长需要高效的平台来管理数据以及大容量的数据存储介质来实现海量数据高可靠的管理和存储 [7]。大量设备的运行数据以及环境数据源源不断地向远程服务端传输，给远程服务端带来高并发的压力，会引发服务器资源耗尽、数据丢失等问题；同时远程服务端在进行数据收集、多种协议解析及实时数据处理的过程中，对数据处理的实时性要求高，否则将影响数据的持久化以及应用层的用户体验。为了满足大量数据实时处理的需要，研究具备高并发、低时延特性的网络服务程序就显得格外重要 [8]。为此，本文提出一种工业互联网高并发流式数据处理方法及应用方案，其首先利用 Netty 通信框架来解决海量数据回传到远程服务端的高并发问题，然后利用 Kafka 消息队列将数据解析、实时分析等耗时操作交由集群进行分布式运算，从而保证数据的持久化、实时处理操作的短响应时间以及高吞吐量。

1　关键技术

Netty 由于具备并发性能高、传输速度快及便于封装的优点，已成为高性能通信领域的首选框架 [9]。为了实时处理 Netty 通信框架（图 1）中的高并发数据，可以使用 Kafka 消息中间件以异步的方式将消息接收和写入数据库这种耗时操作进线解耦，从而避免程序阻塞情况的发生 [10]。

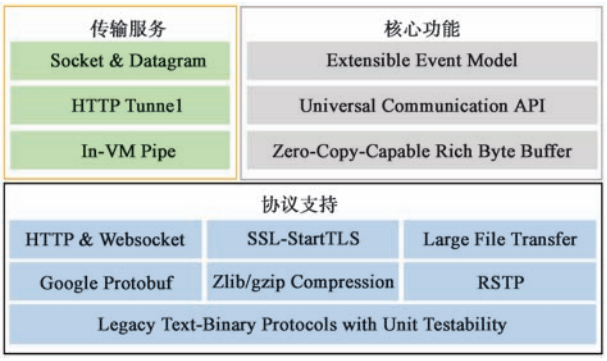


图 1　Netty 整体框架
Fig. 1　Overall framework of Netty

1.1　Netty 网络框架

Netty 是基于 NIO 模型和 epoll 模型的异步通信框架，在开发具备良好可维护性、高性能、高扩展性的网络协议服务器方面具有很大的优势 [11]。

NIO 以块的方式在缓冲区中处理数据，极大地增强了数据操作的灵活性。NIO 的非阻塞 IO 特性决定其允许线程利用操作间隙进行其他事务处理，从而减少了服务器创建线程的数量，节省了服务器性能开销。利用一个线程处理多个客户端连接时，会用到选择器 (selector)。选择器能够监听多个注册的通道上是否有事件发生，如有事件发生，便获取事件，然后针对每个事件进行相应的处理，这样就不必为每个连接都创建一个线程，且只有真正有读写事件发生时才会进行读写，极大地减少了系统开销。

当文件描述符的内核缓冲区非空的时候，epoll 模型发出可读信号进行通知；当写缓冲区不满的时候，则发出可写信号通知。这种主动通知机制避免了对大量并发连接进行轮询，减少了不必要的系统开销。

1.2　Kafka 消息队列

Kafka 是一种高吞吐量、持久化、分布式的消息队列系统，主要架构涉及工作流程、存储机制，以及生产者和消费者，是当前处理大数据计算的一个非常重要的组件，用于解决应用解耦、异步通信及流量控制等问题（图 2）[12]。

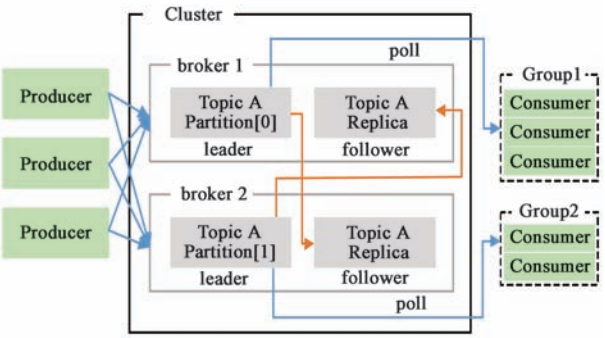


图 2　Kafka 整体框架
Fig. 2　Overall framework of Kafka

Kafka 本质上是一个消息队列。消息队列允许程序独立地扩展或修改队列两边的处理过程，从而实现解耦。系统可恢复性强体现在，即使一个处理消息的进程挂掉，加入队列中的消息仍然可以在系统恢复后被处理。消息队列的缓冲机制有助于解决生产消息和消费消息的处理速度不一致的问题。异步通信机制允许程序将消息放入队列，但是不必立即处理它。

Kafka 处理的消息来自任意多个被称为生产者(producer) 的进程，消息数据通过生产者被发布到不同主题 (topic) 下的分区 (partition)。在分区内部，消息被加上索引和时间戳存储在一起，而其他被称为消费者(Consumer) 的进程则可以从不同分区订阅消息。

2 总体设计

对于涉及众多不同行业工业设备且设备数据回传方式存在多样化的工业互联网平台，其用于海量回传数据的多种通信协议需兼容性良好，网络层应具备高并发能力，可通过消息队列灵活、高效地处理数据；平台需结合大数据技术对设备运行工况进行分析，了解关键器件在不同设备、不同地域、不同时间维度上的运行状况，以指导后续产品的系统及控制策略设计等。

2.1 数据回传方式

工业互联网平台接入的设备类型有轧机、矿用卡车、铁路供电装置、风场设备及暖通设备等，实时回传的行业数据类型包括 UDP 协议数据、json 格式数据等，通过这些数据可以实时监控设备的运行状态。图 3 示出一种多行业的数据回传方式。由于设备安装地点广域、分散，为了在一套系统中汇总整合不同地域的设备运行数据，需将大量设备运行数据、环境数据等汇总到一个数据中心，数据传输方式可借助无线网络、VPN 等多种方式进行传输 [13]。

2.2 系统数据流程

基于Netty的工业大数据平台主要包含数据采集、

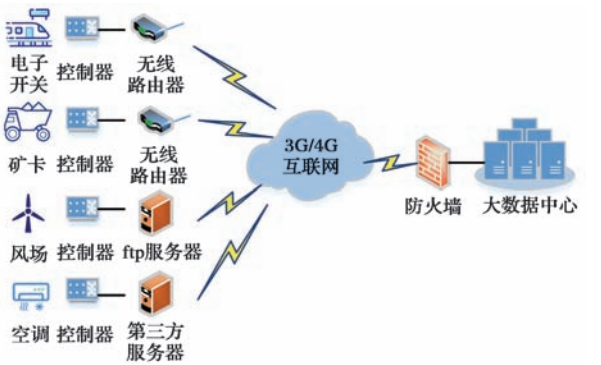


图 3 多行业的数据回传方式

Fig. 3 Data return mode of multi-industry

数据传输、消息队列、数据处理、数据推送及存储等功能。其中，数据采集对象涉及风场数据、轧机数据、空调数据及铁路供电数据等，也可根据需要扩展到更多的行业领域数据；数据类型包括实时数据、文件数据及网络数据等；数据通信部分涵盖 VPN 通道传输、公网传输和 4G 传输等多种传输方式；数据处理部分采用分布式消息队列 Kafka 结合 Spark Streaming 流式处理技术保证数据的高并发处理 [14]；数据存储部分采用分布式存储系统 Hbase 进行海量数据的存储；数据展示部分采用 rest 接口和 WebSocket 技术实现历史数据的请求和实时数据推送。图 4 示出数据处理流程设计。

2.3 数据传输和并发处理

对于数据流的传输，Netty 提供了 ChannelHandler 组件，它基于网络事件触发，作为处理所有入站和出站数据的应用程序逻辑的容器。应用程序中的数据在 Netty 中的传输如图 5 所示。对于应用程

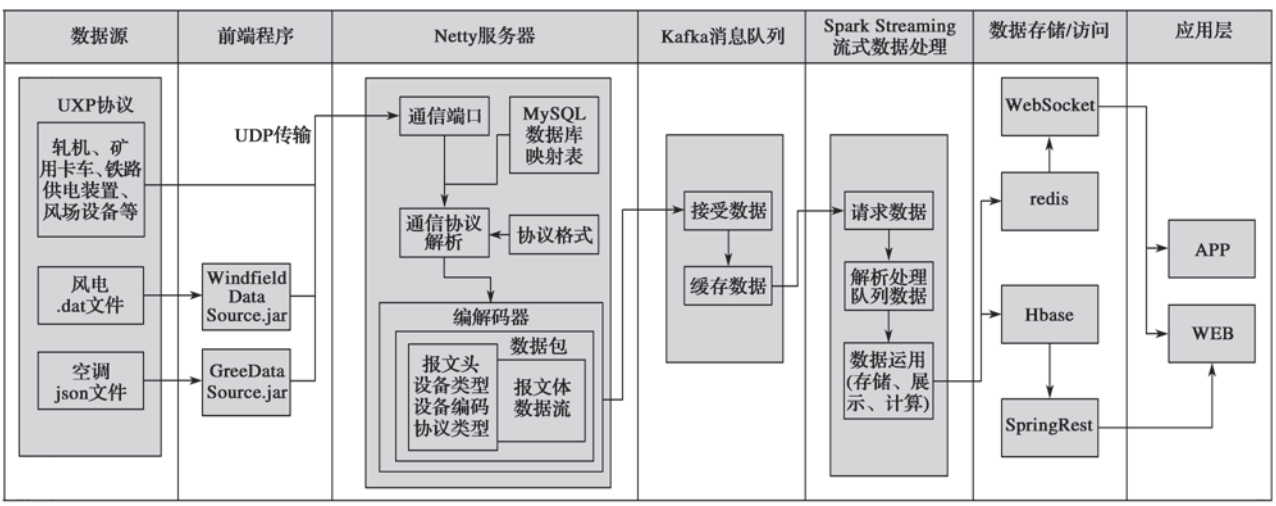


图 4 数据处理流程设计

Fig. 4 Design of data processing flow

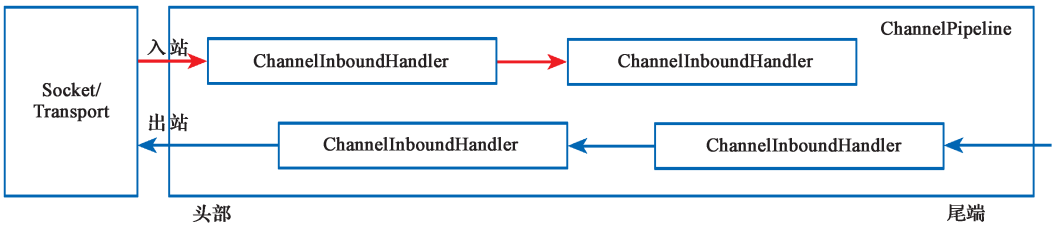


图 5　包含入站和出站 ChannelHandler 的 ChannelPipeline
Fig. 5　ChannelPipeline containing inbound and outbound Channelhandlers

序而言，接收并处理传输到 Netty 框架的数据需要 ChannelInboundHandler 子接口，应用程序的业务逻辑通常要驻留在一个或者多个 ChannelInboundHandler 中，接收事件并执行它们所实现的处理逻辑，并将数据传递给链中的下一个 ChannelHandler。

ChannelPipeline 为 ChannelHandler 链提供了容器，并定义用于在该链上传播的入站和出站事件流的应用程序编程接口 (application programming interface, API)。当通道 (channel) 被创建时，它会被自动地分配到其专属的 ChannelPipeline。Netty 的 EventLoop 是协同设计的一部分，它采用了 2 个基本的 API：并发和网络编程，而服务于 Channel 的 I/O 和事件的 EventLoop 包含在 EventLoopGroup 中。图 6 示出 EventLoopGroup 的工作模式，它负责为每个新创建的 Channel 创建一个 EventLoop，使用事件循环的方式进行分配以获取一个均衡的分布。

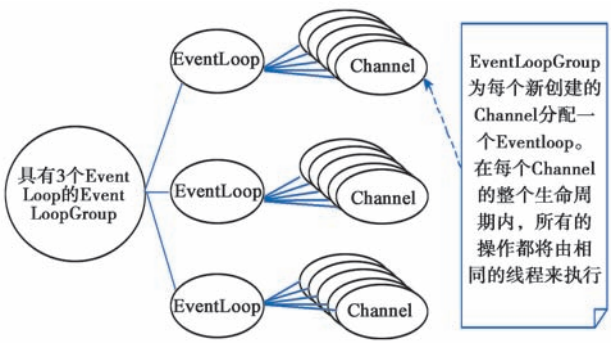


图 6　非阻塞传输的线程分配方式
Fig. 6　Thread allocation for non-blocking transmission

2.4　Netty 结合 Kafka 的数据处理流程

通过 Netty 服务器，以 UXP 协议对数据进行解析，并最终将各行业数据归一化为相同结构的数据包，流程如图 7 所示。经过数据归一化流程后，每个行业数据都将以相同结构的 Kafka 实例发送到消息队列中，然后在 Spark 程序中进行解析，实现了行业数据与后台程序之间的解耦。这将极大减小后台程序之间的行业区别，给数据持久化和应用层程序的开发带来便利。

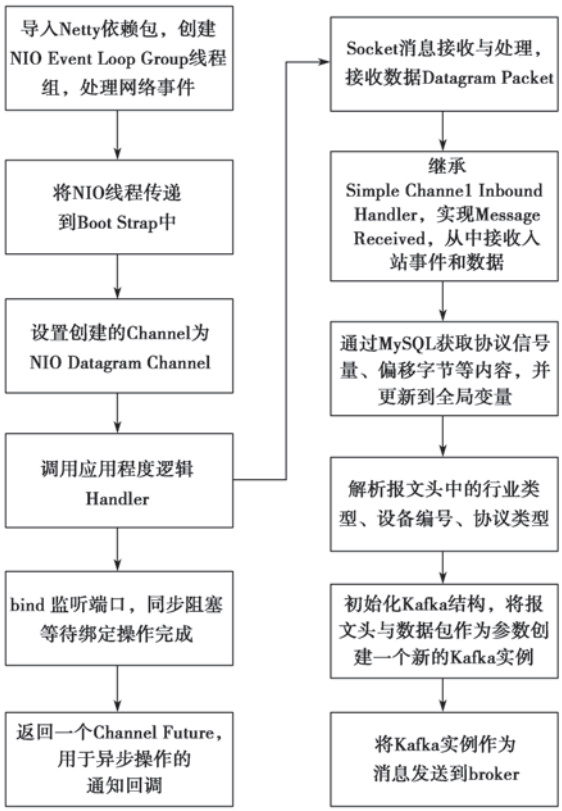


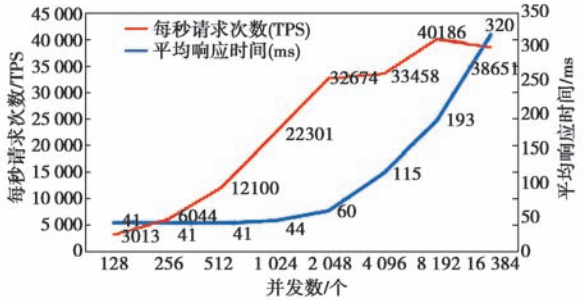
图 7　基于 Netty 和 Kafka 的网络处理流程
Fig. 7　Network processing flow based on Netty and Kafka

3　实验验证及分析

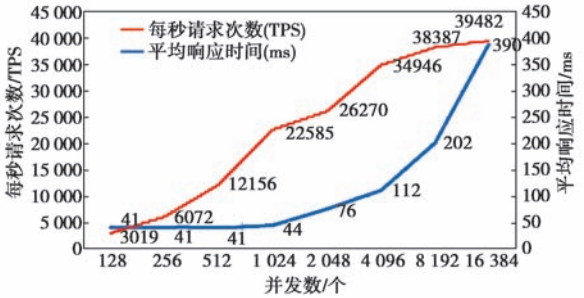
为了验证所设计系统的性能，挑选了运行在全国各地 3 个行业的 528 台设备，以 3 种不同数据回传方式向测试集群上传需要处理的数据，测试集群总共 4 个节点。测试场景如下：风场接入 25 台设备，回传设备运行数据、风场遥信数据和风场遥测数据，每 10 min 将数据存为 .dat 文件并通过 ftp 回传到测试集群，然后由前端程序进行解析并发往 Netty 服务器；接入空调 500 台，数据以 json 格式提供，由前端程序进行请求并转发给 Netty 服务器；接入轧机 3 台，以公网传输的方式通过上位机程序将数据包直接发往 Netty 服务器，数据回传时间为 3 s。

3.1 网络通信性能测试

通过 Netty 服务器程序，分别在 1 次连接 1 000 次数据请求和 1 次只进行 1 次数据请求但是循环 1 000 次两种工况下进行通信效率测试，设置并发数分别为 128 和 256，得到的结果如图 8 所示。



(a) 1 次连接 1 000 次请求



(b) 1 次连接 1 次请求，循环 1 000 次

图 8 网络通信性能测试

Fig. 8 Network communication performance test

从测试结果可以看出，Netty 服务器对于高并发数据处理的效率，在 4 万次请求解析过程中响应时间不超过 400 ms。每次创建连接和重用连接的平均响应时间最大差值也仅为 70 ms，这就给工业领域多行业设备的接入带来极大优势，可以不断地创建新连接进行数据回传，性能损耗对应用层几乎没有影响。

3.2 数据处理性能测试

为了测试大量数据通过消息队列进行数据解析、数据持久化等耗时操作的性能，分别对消息队列每秒所能处理的数据帧数、平均耗时及处理延迟时间等指标进行验证。当前接入 Kafka broker 节点数为 4 个，实验将对 Kafka 消息队列的数据处理情况进行监控。由图 9 可以得知，平均每秒处理的数据流为 92.64 帧，即可完成所有数据的消费。由图 10 可以得知，每笔数据从接入到处理再到存入数据库平均所耗费的时间为 169 ms。由图 11 可以得知，数据处理的平均延迟时间为 171 ms。可见所设计系统在处理高并发流式数据的过程中，数据处理平均耗时和平均延迟时间均能满足系统需求。

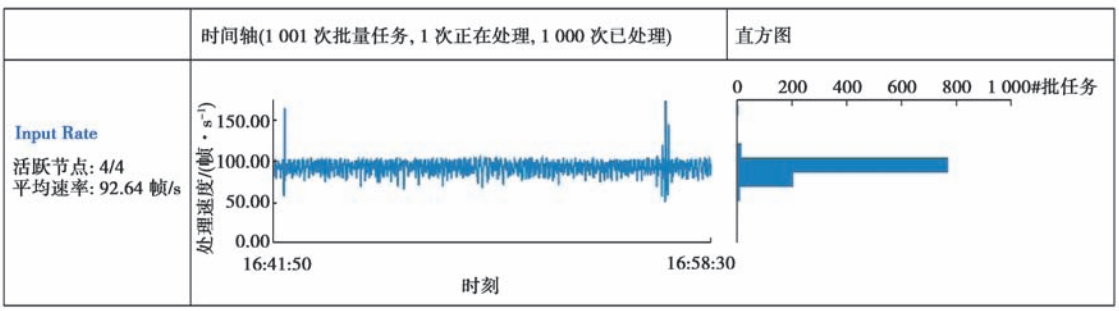


图 9 数据后台平均处理数据流帧数

Fig. 9 Average frame number of data stream processed on server

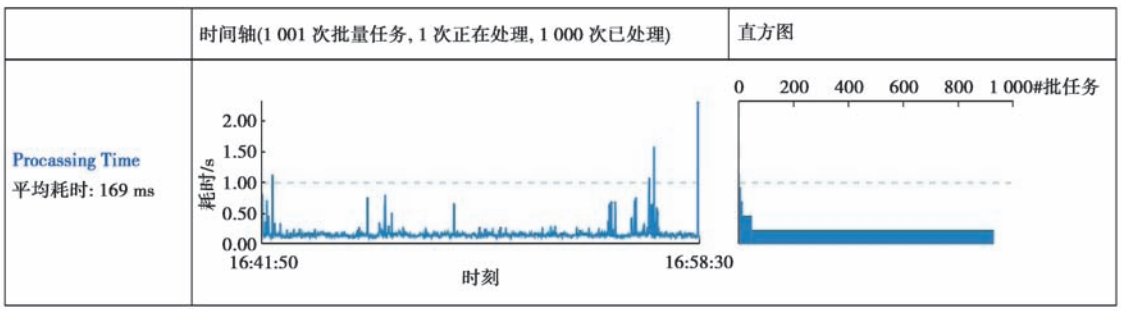


图 10 数据后台平均处理数据流耗时

Fig. 10 Average processing time of data stream on server

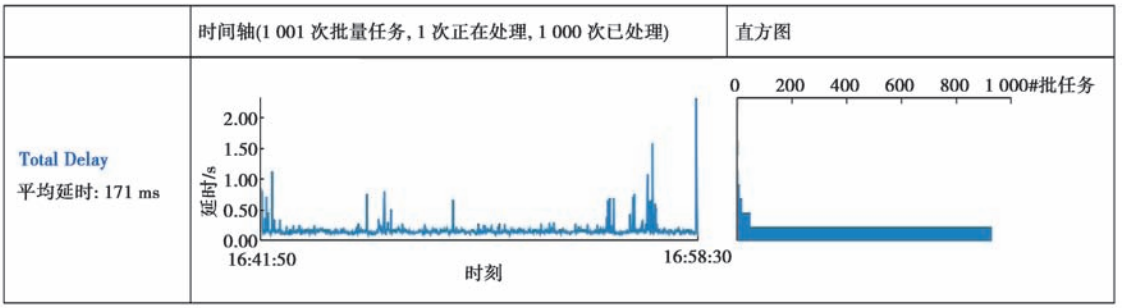


图 11　数据后台平均处理数据流延时

Fig. 11　Average processing delay of data stream on server

4　结语

本文提出一种工业互联网高并发流式数据处理技术及应用方案，并详细介绍了多行业数据接入平台的数据接入方式和数据处理方式、基于 Netty 的高并发数据处理机制和 Kafka 消息队列。使用该方案的工业互联网大数据平台能满足广域、分散的大量设备的数据接入，并且支持多种数据格式以及多种传输协议，解决了不同行业数据的一致性问题，在高并发的情况下依然能保持毫秒级的处理速度。

后续可以考虑优化消息中间件中 Consumer 和 Partition 的数量配置，一方面降低对系统硬件资源的需求，另一方面可以给应用层软件提供更加灵活的数据接口。

参考文献：

[1] JI C, SHAO Q, SUN J, et al. Device Data Ingestion for Industrial Big Data Platforms with a Case Study[J]. Sensors, 2016, 16(3): 279.

[2] 张平 , 刘会永 , 李文璟 , 等 . 工业智能网——工业互联网的深化与升级 [J]. 通信学报 , 2018, 39(12): 134-140.
ZHANG P, LIU H Y, LI W J, et al. Industrial intelligent network: deepening and upgrading of industrial Internet[J]. Journal on Communications, 2018, 39(12):134-140.

[3] 余晓晖 , 刘默 , 蒋昕昊 , 等 . 工业互联网体系架构 2.0[J]. 计算机集成制造系统 , 2019, 25(12): 2983-2996.
YU X H, LIU M, JIANG X H, et al. Industrial Internet Architecture 2.0[J]. Computer Integrated Manufacturing Systems, 2019, 25(12): 2983-2996.

[4] 郑树泉 , 覃海焕 , 王倩 . 工业大数据技术与架构 [J]. 大数据 , 2017, 3(4): 67-80.
ZHENG S Q, QIN H H, WANG Q. Industrial Big Data Technologies and Architecture[J].Big data, 2017, 3(4): 67-80.

[5] 工业互联网产业联盟 . 工业互联网平台白皮书 [R]. 北京 : 中国信息通信研究院 , 2017.

[6] 王宏志 , 梁志宇 , 李建中 , 等 . 工业大数据分析综述 : 模型与算法 [J]. 大数据 , 2018, 4(5): 62-79.
WANG H Z, LIANG Z Y, LI J Z, et al. Survey on industrial big data analysis: models and algorithms[J]. Big data, 2018, 4(5): 62-79.

[7] XIE Z, CAO Z, WANG Z, et al. Modeling traffic of big data platform for large scale datacenter networks[C]//2016 IEEE 22nd International Conference on Parallel and Distributed Systems (ICPADS). Wuhan, China: IEEE, 2017: 224-231.

[8] 陈超录 , 罗煌 , 董其爱 , 等 . 智轨电车低压配电与数据采集系统研究 [J]. 控制与信息技术 , 2020(1): 58-61, 65.
CHENG C L, LUO H, DONG Q A, et al. Research on the Low-voltage Distribution and Data Acquisition System of Autonomous-rail Rapid Tram[J]. Control and Information Technology, 2020(1): 58-61, 65.

[9] 彭勤 , 李勋 , 黄舟 . 有轨电车调度管理系统通信集成技术 [J]. 控制与信息技术 , 2018(4): 64-68.
PENG Q, LI X, HUANG Z, et al. Communication Integration Technology of Tramcar Dispatching Management System[J]. Control and Information Technology, 2018(4): 64-68.

[10] 夏斐 . 基于 Netty 的消息中间件的研究与实现 [D]. 成都 : 电子科技大学 , 2018.

[11] 甄凯成 , 黄河 , 宋良图 . 基于 Netty 和 Kafka 的物联网数据接入系统 [J]. 计算机工程与应用 , 2020, 56(5): 135-140.
ZHEN K C, HUANG H, SONG L T. Internet of Things Data Injection System Based on Netty and Kafka[J].Computer Engineering and Applications, 2020, 56(5): 135-140.

[12] 孙龙龙 . 面向 Kafka 消息中间件的监控管理系统设计与实现 [D]. 武汉 : 华中科技大学 , 2019.

[13] 梁涛 , 许琰 , 李燕超 , 等 . 大数据处理技术在风电机组状态监测中的应用 [J]. 水力发电 , 2019, 45(8): 107-110.
LIANG T, XU Y, LI Y C, et al.Application of Big Data Processing Technology in Condition Monitoring of Wind Turbines[J].Water Power, 2019, 45(8): 107-110.

[14] 田璐 , 齐林海 , 李青 , 等 . 基于 Spark Streaming 的电力流式大数据分析架构及应用 [J]. 电力信息与通信技术 , 2019, 17(2): 23-29.
TIAN L, QI L H, LI Q, et al. Power Streaming Big Data Analysis Architecture and Application Based on Spark Streaming[J]. Electric Power Information and Communication Technology, 2019, 17(2): 23-29.