Workgroup: Internet Research Task Force
Internet-Draft:
draft-zcz-nmrg-digitaltwin-data-collection-01
Published: 7 November 2022
Intended Status: Informational
Expires: 11 May 2023

Authors: C. Zhou D. Chen P. Martinez-Julia
China Mobile China Mobile NICT
Q. Ma
Huawei

Data Collection Requirements and Technologies for Digital Twin Network

Abstract

A Digital Twin Network is a virtual representation of a physical network, which is meant to be used by a management system to analyze, diagnose, emulate and control the physical network based on monitoring information, data, models, and interfaces. The construction and state update of a Digital Twin Network require obtaining real-time information of the physical network it represents (i.e., telemetry data). This document aims to describe the data collection requirements and provide data collection methods or tools to build the data repository for building and updating a digital twin network.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

This Internet-Draft will expire on 11 May 2023.

Copyright Notice

Table of Contents

1. Introduction

With the deployment of Internet of Things (IoT), cloud computing and data center, etc., the scale of the current network is expanded gradually. However, the increase of network scale leads to also increasing the complexity of the current network, and it induces plenty of problems. In order to improve the autonomy ability of network and reduce potential negative effects on physical and virtual networks, we consider that an endogenous intelligent and autonomous network architecture which achieves self-optimization and decision is indispensable (in general, self-management and self-operation). The digital twin technology answers to the challenge of building self-management systems because it can optimize and validate policies through real-time and interactive mapping with physical entities.[I-D.irtf-nmrg-network-digital-twin-arch]

Data is the cornerstone required for constructing a digital twin for a network, namely a Digital Twin Network (DTN). In the face of large network scale, data collection, storage and management are faced with great challenges. So, data collection methods and tools should meet the requirements of target-driven, diversity, lightweight and efficiency, while being open and standardized. Among all the requirements, achieving a lightweight and efficient data collection method is of the most importance. If the full-data collection method is adopted, huge storage space and bandwidth resource is needed, especially for complex scenarios that require real-time data and traffic from multi-source and heterogeneous devices. Therefore, it is extremely important to agree on lightweight and efficient data collection, aggregation, and correlation methods, toward building the transmission of monitoring information (telemetry data), processing, and storage required to build a DTN system.

This document aims to describe the data collection requirements and proposes efficient data collection methods or tools to build the data repository for digital twin network.

2. Definitions and Acronyms

PN: Physical Network

IMC: Instruction Management Center

DSC: Data Storage Center

DTN: Digital Twin Network

TSE: Telemetry Streaming Element

RDF: Resource Description Framework

CEP: Complex Event Processing

3. Data Collection Requirements for Digital Twin Network

3.1. Target-driven and On-demand Collection

The monitoring data of a network is the basis to build a DTN system. Such data is collected from physical and virtual networks. It includes, but is not limited to, the following types:

*Provisional and operational status of physical or virtual devices, as well as the network topology with all network elements.

*Configuration data that is required to transform a network system from its initial default state into its current state.

*Running status of physical, logical, or virtual ports and links.

*Logs and events records of all the network elements.

*Statistics (packet loss, traffic throughput, latency, etc.) of flows and ports.

*Various data regarding users and services.

*Life-cycle operation data of all network elements.

*All above data in time series.

The collection of the monitoring information from a network required for maintaining a DTN (telemetry data) should be in target-driven and on-demand mode. It is not always necessary to collect all monitoring information from the network (telemetry data) listed above because of the high cost of resources (CPU, memory, bandwidth etc.). The type, frequency and method of data collection aim to meet the application of a DTN depends on the specific network topology and application requirements.

3.2. Diverse Tools for Various Data Collection

The different types of monitoring information required to maintain a DTN (telemetry data) have several characteristics. Some data (e.g. hardware status, environmental data, etc.) requires lower collecting frequency, and some data (e.g. flow status, link fault, etc.) needs to be of higher level of real-time. Some data (e.g. device status, port statistics, etc.) can be collected directly and simply via normal tools, while some data (e.g. per-flow latency, traffic matrix, etc.) can only be acquired through complex network measurement. Therefore, multiple tools or methods are needed to collect the massive data required to build the DTN entity.

Currently, some widely-used tools, such as SNMP, NetConf, Telemetry, INT (In-band Network Telemetry), DPI (Deep Packet Inspection), etc. can be candidate tools to collect data for digital twin network. Yang data model and associated mechanisms defined in [RFC8639] [RFC8641] enable subscriber-specific subscriptions to a publisher's event streams, and can help subscriber applications to request a continuous and customized stream of updates from a YANG datastore. [RFC9232]'s Appendix-A gives a survey on existing network telemetry techniques, which explores an overview of management plane, control plane and data plane telemetry techniques and standards.

Going forward, it is necessary to study new data collection technology in the following aspects in combination with the data requirements of network application for DTN:

*High-performance data collection technology based on programmable circuits.

*Measurement methods for complex monitoring information such as network performance and network traffic.

*Collaborative data collection technology for multiple data sources.

*Distributed and collaborative data collection technology for complex network, and the time synchronization problem of data acquisition.

3.3. Lightweight and Efficient Collection

Data collection tools and methods should be as lightweight as possible, so as to reduce the occupation of network equipment resources and ensure that data collection does not affect the normal operation of the network. The major requirements are list as below.

*Data collection tools and methods need to improve efficiency of execution, reduce the cost of computing, storage and communication bandwidth.

*The collection of redundant data should be avoided or minimized.

*For the data set that needs to be collected, make full use of the data compression technology, to reduce the resource cost in the collection phase.

3.4. Open and Standardized Interfaces

Data collection interface used to build the DTN should be open and standardized to help avoid either hardware or software vendor lock, and achieve inter-operability. The major requirements of data collection interfaces are:

- Support configuration management, including the data collection protocol, frequency or period, etc.

- Support several rate options (e.g. minute-level, 10-second level, second level (near real time), and real time level) to accommodate different data requirements from applications.

- Be extensible so that more features can be added with limited parameter changes and with backward compatibility.

- Be able to provide secure and reliable information exchange mechanism.

3.5. Naming for Caching

Both raw monitoring information (telemetry data) and knowledge items obtained from monitoring must be able to be addressed uniquely. This means to give a unique identifier or "name" to each data or knowledge item that references it. This name will be used by caching mechanisms to store the data and provide it for clients that request it, which will also use such name.

3.6. Efficient Multi-Destination Delivery

The maintenance of DTN systems will not be the sole purpose of monitoring information and knowledge communication. Other applications would also request raw monitoring information (telemetry data) or knowledge items. They can use the name to identify it. The monitoring system (telemetry system), following the recommendations of RFC 9232 [RFC9232], will deliver the requested data or knowledge items to the requesters as much efficiently as possible. On the one hand, items will be provided by the closest cache to the destination of the data. On the other hand, items will be replicated in the best nodes, following an efficient multi-cast spanning tree. Different underlying protocols can be used to achieve this mechanism.

4. An Efficient Data Collection Method for Digital Twin Network

4.1. Overview

The DTN's data repository sub-system manages all network data, in real time, from the PN to the DTN. Sufficient and timely data are always required to construct the twin entity and various data models. However the existing methods collect the full data from the PN for modeling, and do not consider problems like time-lag, insufficient storage resources, low computational efficiency and waste of bandwidth resources caused by data transmission. In order to solve these problems, this section introduces an efficient data collection method, named 'knowledge and instruction driven data collection'. This data collection method is based on sending instructions to the elements of the PN for them to pre-process the data (data cleaning or knowledge representation) before sending it back to be applied to the DTN.

4.2. Efficient Data Collection Mechanism

The management system structure consists of the PN and the DTN. The PN includes multiple Data Storage Centers (DSC) and Telemetry Streaming Element (TSE), and the DTN includes the Instruction Management Center (IMC) and Data Storage Center (DSC). The TSE has multiple functions, including data collection, data aggregation, data correlation, knowledge representation and query, etc. In addition, a Complex Event Processing (CEP) engine is integrated into TSE to perform queries to the streamed data. The IMC has two functions. On the one hand, it is used to manage the registration of the DSC in the PN side, and its registration information can include various key information such as the IP address of the DSC in the PN side, chosen data type, and various index names in the data, data source name and data size, etc. On the other hand, it is used to adaptively configure data collection instructions according to the collection requirements of the DSC in the DTN side and search for IP addresses to send instructions. The instruction-carrying information includes rule-based mathematical expressions, executable models in .exe format, dynamic collection frequency, parameter lists, program text files in .m format, text files with parameter configuration, and other types of files. Instructions are flexible and programmable, and can be created, modified, combined, and deleted at any time according to requirements. When the DSC of the DTN side requests data to the IMC, the IMC searches the IP address of the DSC in the database with the registration information, which is built according to critical information, such as data type and data name, and functional instructions for data processing or knowledge representation can be implemented depending on the demand configuration. The DSC of the DTN side stores the effective information after data processing and knowledge representation returned by the TSE.

The DSC in the PN side has two functions. On the one hand, it stores data of various types, such as performance indicators, operational status, log, traffic scheduling, business requirements, etc. On the other hand, it has the function of automatically parsing the instructions sent by the TSE. Then the operating environment of the instruction is configured according to the instruction needs, and data processing or knowledge representation is performed based on the instruction. Data processing mainly includes data cleaning, filling missing data, normalization, conflict verification, etc. Knowledge representation refers to the representation of the original data as a data structure that can be used for efficient computation. Such representation results are closer to machine language, which is conducive to the rapid and accurate construction of the model. The role of knowledge representation is to represent the original data as a data structure that can be used to efficiently calculate.

```
   +------------------------------+   +-----------------------+
   |   Physical  Network          |   |  Digital Twin Network |
   | +-----+    +-----+  +------+ |   |  +------+  +-------+  |
   | |     |    |     |  |      | |   |  |      |  |       |  |
   | | DSC |... | DSC |  | TSE  | |   |  |  IMC |  |  DSC  |  |
   | |     |    |     |  |      | |   |  |      |  |       |  |
   | +-+---+    +--+--+  +---+--+ |   |  +---+--+  +----+--+  |
   |   |           |         |    |   |      |          |     |
   +------------------------------+   +-----------------------+
       |           |         |               |          |
       | 1.1. Register       |               |          |
       +-----------+--------->               |          |
       |           |         |               |          |
       |           | 1.2. Register           |          |
       |           +--------->               |          |
       |           |         | 1.3. Register |          |
       |           |         +--------------->          |
       |           |         |             2. Data req. |
       |           |         |               <----------+
       |           |         | 3. Query and instruction |
       |           |         |    configuration         |
       |           |         |               +          |
       |           |         4. Send instructions       |
       |           |         <---------------+          |
       |           |         |               |          |
       |           |   5. Parse and execute  |          |
       |           |      instruction        |          |
       | 6. Data subscript.  |               |          |
       <---------------------+               |          |
       | 7. Knowledge        |               |          |
       |    representation   |               |          |
       |     8. Data pushing |               |          |
       +--------------------->               |          |
       |           | 9. Data aggregation and |          |
       |           |    correlation          |          |
       |           |         | 10. Send processed data  |
       |           |         +-------------------------->
       |           |         |               |          |
```

Figure 1: Data Collection Process

4.3. Data Collection Process

The specific process is as follows:

- The DSC in the PN side registers into the TSE. The TSE registers into the IMC. Both provide their IP addresses, the data type, the data source, the data size, etc.

- The DSC in the DTN side sends the data collection request to the IMC.

- According to the data collection request, the IMC intelligently queries the registration addressing information and configures the data processing instruction.

- The IMC in the DTN side sends the corresponding instruction according to the query result to the TSE.

- After receiving the instructions, the TSE parses them and executes them. The query function can be performed by the CEP engine, which receives all monitoring information (telemetry data) and processes it with all queries provided.

- The TSE sends data subscription to DSC in the PN side.

- The DSC in the PN side represents the data semantically in RDF form or sends the data in raw form to the TSE for it to make the semantic representation.

- The DSC in the PN side pushes the data or knowledge item to the TSE.

- The TSE aggregates and correlates the collected data or knowledge items. Then, according to the actual needs, generates aggregated data or knowledge items.

- The TSE sends the resulting data or knowledge items to the DSC in the DTN side.

4.4. Query and Aggregation Functions

The TSE supports an arbitrary number of queries and aggregation functions. As a minimum, it will support:

- *A function to apply a particular calculation to the values retrievied from a specified metric for a specified period of time. The basically supported calculations must be:

  - -Average: Returns the single number resulting from averaging all values in the period.

  - -Maximum: Returns the single number that represents the highest value in the period.

  - -Minimum: Returns the single number that represents the lowest value in the period.

  - -Percentile X: Returns the percentile of calculated at position X (from 0, which is the minimum, to 100, which is the maximum).

  - -Moving Average X: Transforms all values of the specified period by calculating every value as the average of the previous X values (or less if there are not enough).

  - -Filter Previous X: Removes the values that change less than X percent from the previous value.

  - -Filter Average X: Removes the values that change less than X percent from the average value.

  - -Filter Moving Average X Y: Removes the values that change less than Y percent from the value of the moving average for X previous values.

- *A function to represent the collected values in a semanting structure following some ontology, information model, and data format (YANG). This will enforce semantic constraints to the values, such as avoiding negative measures of some parameters (e.g., bandwidth usage).

- *A function to analyze the collected values to detect some pattern (provided) and, if so, trigger some notification that other module can use to execute some action.

The particular behavior of the three functions will be described in a high-level language that is transformed to the specific code used by the device, such as [P4].

5. Summary

This draft describes the requirements for data collection and provides the data collection methods or tools required to build the data repository for maintaining DTN systems. These data collection methods or tools should meet the requirement of target-driven, diversity, lightweight and efficiency, while being open and standardized. Among all the requirements, lightweight and efficiency requirements are the most important. Thus, this draft provides a lightweight and efficient method for data collection that is particularly optimized for maintaining DTN systems. Going forward, more methods (transformation and aggregation functions) and tools (solutions) shall be studied to extend the contents of this draft.

6. Security Considerations

TBD.

7. IANA Considerations

This document has no requests to IANA.

8. References

8.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC8639] Voit, E., Clemm, A., Gonzalez Prieto, A., Nilsen-Nygaard, E., and A. Tripathy, "Subscription to YANG Notifications", RFC 8639, DOI 10.17487/RFC8639, September 2019, <https://www.rfc-editor.org/info/rfc8639>.

[RFC8641] Clemm, A. and E. Voit, "Subscription to YANG Notifications for Datastore Updates", RFC 8641, DOI 10.17487/RFC8641, September 2019, <https://www.rfc-editor.org/info/rfc8641>.

[RFC9232] Song, H., Qin, F., Martinez-Julia, P., Ciavaglia, L., and A. Wang, "Network Telemetry Framework", RFC 9232, DOI 10.17487/RFC9232, May 2022, <https://www.rfc-editor.org/info/rfc9232>.

8.2. Informative References

[I-D.irtf-nmrg-network-digital-twin-arch]

Zhou, C., Yang, H., Duan, X., Lopez, D., Pastor, A., Wu, Q., Boucadair, M., and C. Jacquenet, "Digital Twin Network: Concepts and Reference Architecture", Work in Progress, Internet-Draft, draft-irtf-nmrg-network-digital-twin-arch-02, 24 October 2022, <https://www.ietf.org/archive/id/draft-irtf-nmrg-network-digital-twin-arch-02.txt>.

[P4] The P4 Language Consortium, "P4 Language Specification (https://p4.org/p4-spec/docs/P4-16-v-1.2.3.html)", 11 July 2022.

Authors' Addresses

Cheng Zhou
China Mobile
Beijing
100053
China

Email: zhouchengyjy@chinamobile.com

Danyang Chen
China Mobile
Beijing
100053
China

Email: chendanyang@chinamobile.com

Pedro Martinez-Julia
NICT
4-2-1, Nukui-Kitamachi, Koganei, Tokyo
184-8795
Japan

Email: pedro@nict.go.jp

Qiufang Ma
Huawei
Nanjing
210012
China

Email: maqiufang1@huawei.com