Data Collection Requirements and Technologies for Digital Twin Network

draft-zcz-nmrg-digitaltwin-data-collection-03

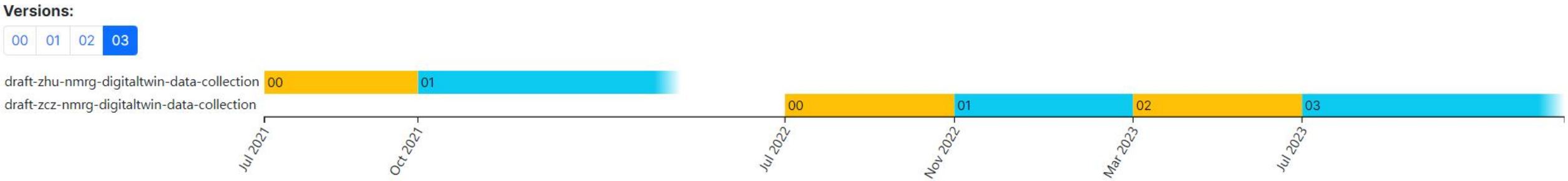
Cheng Zhou (zhouchengyjy@chinamobile.com, **Presenter**)
Danyang Chen (chendanyang@chinamobile.com)
Pdero Martinez-Julia (pedro@nict.go.jp)
Qiufang Ma (maqiufang1@huawei.com)

Scope & Objectives

- **Scope**
  - Describe the requirements of data collection for building digital twin network (DTN) system.
  - Provide data collection methods torward building DTN data repository, including existing methods and innovative new methods.

- **Objectives**
  - Identify the data collection requirements and principles for DTN
  - Call for more efficient data collection methods suitable for DTN system
  - Reach a consensus on selecting data collection methods for various network data

Change Log



- **History**
  - Initially, the draft was just a specific data collection method for DTN
  - Promote the daft to extend the scope to general data collection requirements and methods for DTN
- **Major Changes on 01~03**
  - Refined section of data collection requirements for DTN
  - Added new section for Data Collection Technologies of existing methods and innovation directions
  - Renamed the proposed method to ‘Knowledge and Instruction Driven Data Collection’, and refined the text

Table of Contents

Data Collection Requirements for DTN

- **Target Driven and On-demand Collection**
  - Complete data is NOT affordable
  - Just collect data toward the target to build specific model for specific application
- **Diverse Tools for Various Data**
  - Various data: collecting frequency, real-time level, easy-to-get, hard to measure, etc.
  - Diverse tools are needed
- **Lightweight and Efficient Collection**
  - To improve efficiency of execution, reduce the cost of computing, storage and communication bandwidth
  - Redundant data should be avoided or minimized, with optional data data compression and aggregation

Data Collection Requirements for DTN - Cont.

- **Open and Standardized Interfaces**
  - Support configuration management, including collection protocol, frequency, period, etc.
  - Be able to provide secure and reliable information exchange mechanism
- **Naming for Caching**
  - Give a unique identifier or "name" to each data or knowledge item for references
  - The name will be used by caching mechanisms to store the data and provide it for clients
- **Efficient Multi-Destination Delivery**
  - The telemetry system [RFC9232] can be used to deliver the requested data or knowledge items to the requesters.
  - Items will be provided by the closest cache to the destination of the data.

Data Collection - Existing Methods

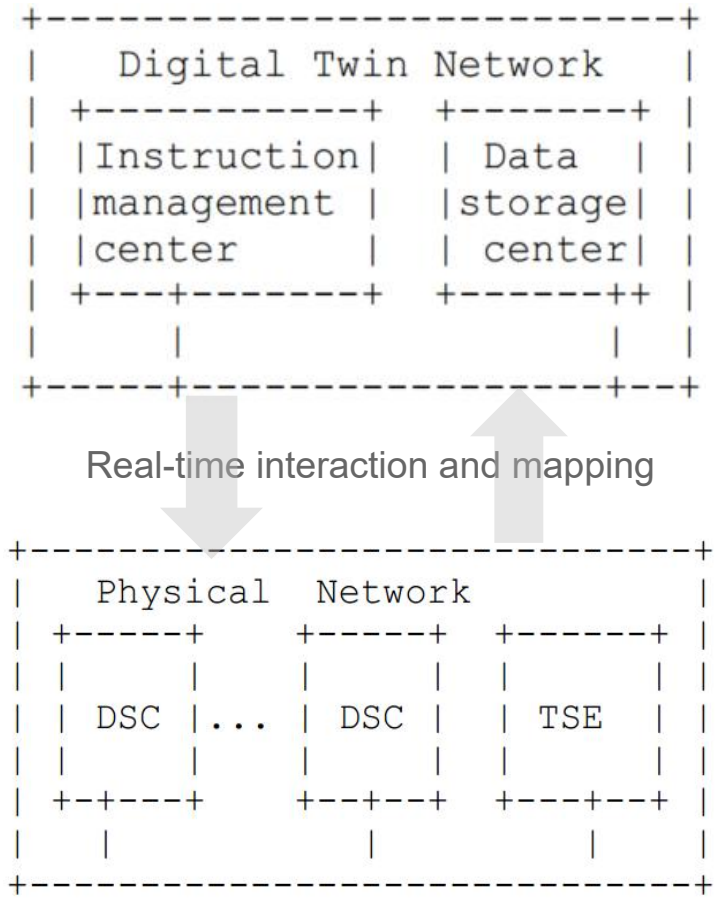
- SNMP, RESTCONF, NETCONF, INT (In-band Network Telemetry), DPI (Deep Packet Inspection), IPFIX, etc. can be candidate tools
- YANG data model and associated mechanisms([RFC8639], [RFC8641], etc.) can help subscriber applications to request a continuous and customized stream of updates from a YANG datastore.
- Appendix-A in [RFC9232] gives a survey on existing network telemetry techniques, exploring an overview of management plane, control plane and data plane telemetry techniques and standards.
- **Some innovation methods are on the way...**
  - [I-D.claise-opsawg-collected-data-manifest] proposes a YANG model to store contextual information along with the collected data in order to keep the collected data exploitable;
  - [I-D.ietf-ippm-explicit-flow-measurements] addresses the network performance measurement problem under encrypted transport protocols, via proposing some hybrid measurement methods based on marking bits in packet headers.

Data Collection - Future Innovation Directions

- High-performance data collection technology based on programmable circuits
- Measurement methods for complex monitoring information such as network performance and network traffic
- Distributed and collaborative data collection techniques for integrating and fusing data from multiple data sources
- Assessment of federation policies in data provisioning to enable cross-domain data provision and implement multi-domain digital twin scenarios
- Investigating self-adaptive and self-learning data collection techniques that can dynamically adjust data collection parameters, methods, and priorities based on network conditions and user requirements
- Exploring machine learning and AI techniques to enhance the efficiency and accuracy of data collection processes by identifying patterns, correlations and anomalies in network data

Knowledge and Instruction Driven Data Collection for DTN - New method proposed

- Current collection methods collect raw and full data from Physical Network
  - Time-cost
  - Insufficient storage resources
  - Low computational efficiency
  - Waste of bandwidth resources caused by data transmission
- Main idea of the method
  - Send instructions to the elements of the PN to pre-process the data (data cleaning or knowledge representation) before sending it back to be applied to the DTN.
  - Efficient and lightweight data collection, aggregation and correlation.



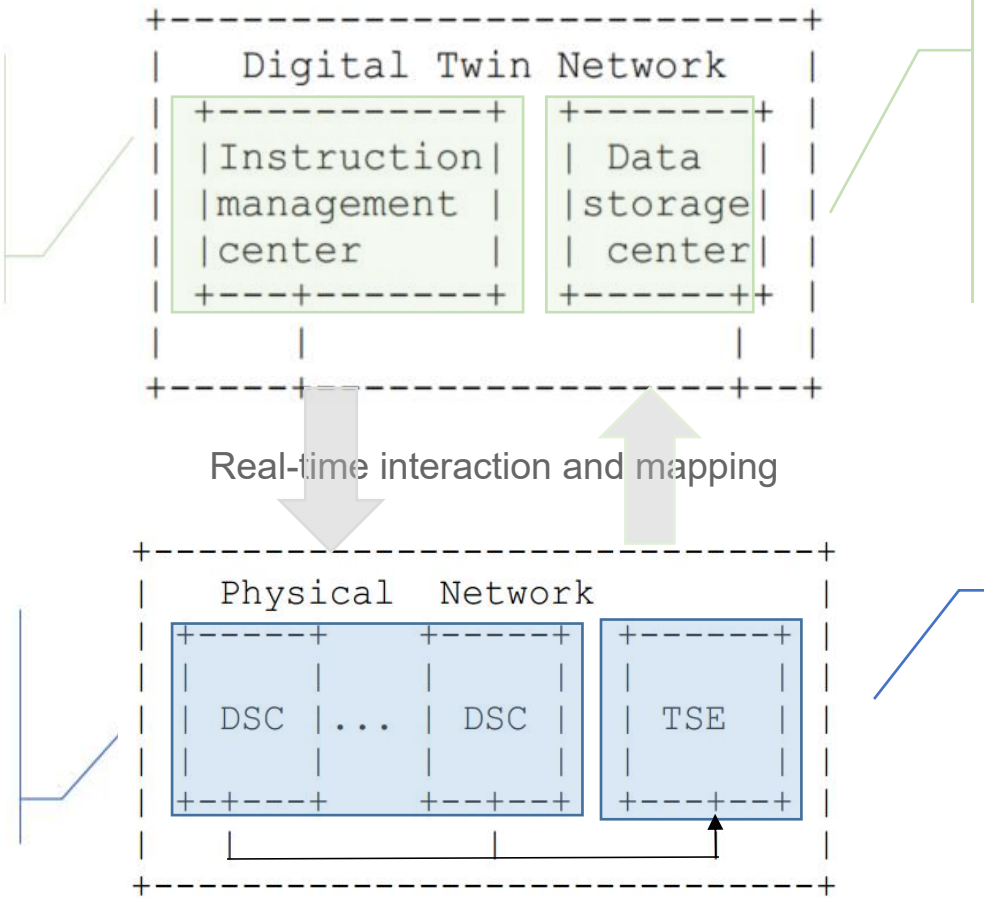
Knowledge and Instruction Driven Data Collection for DTN - Components and Mechanism

IMC: Instruction Management Center

- Manage the registration of the TSE in PN.
- Adaptively configure data collection instructions according to the requirements of the DSC in TN

DSC: Data Storage Center in PN

- Store data, such as performance indicators, operational status, logs, business requirements, etc.
- Automatically parse the instructions sent by the IMC in TN



DSC: Data Storage Center in TN

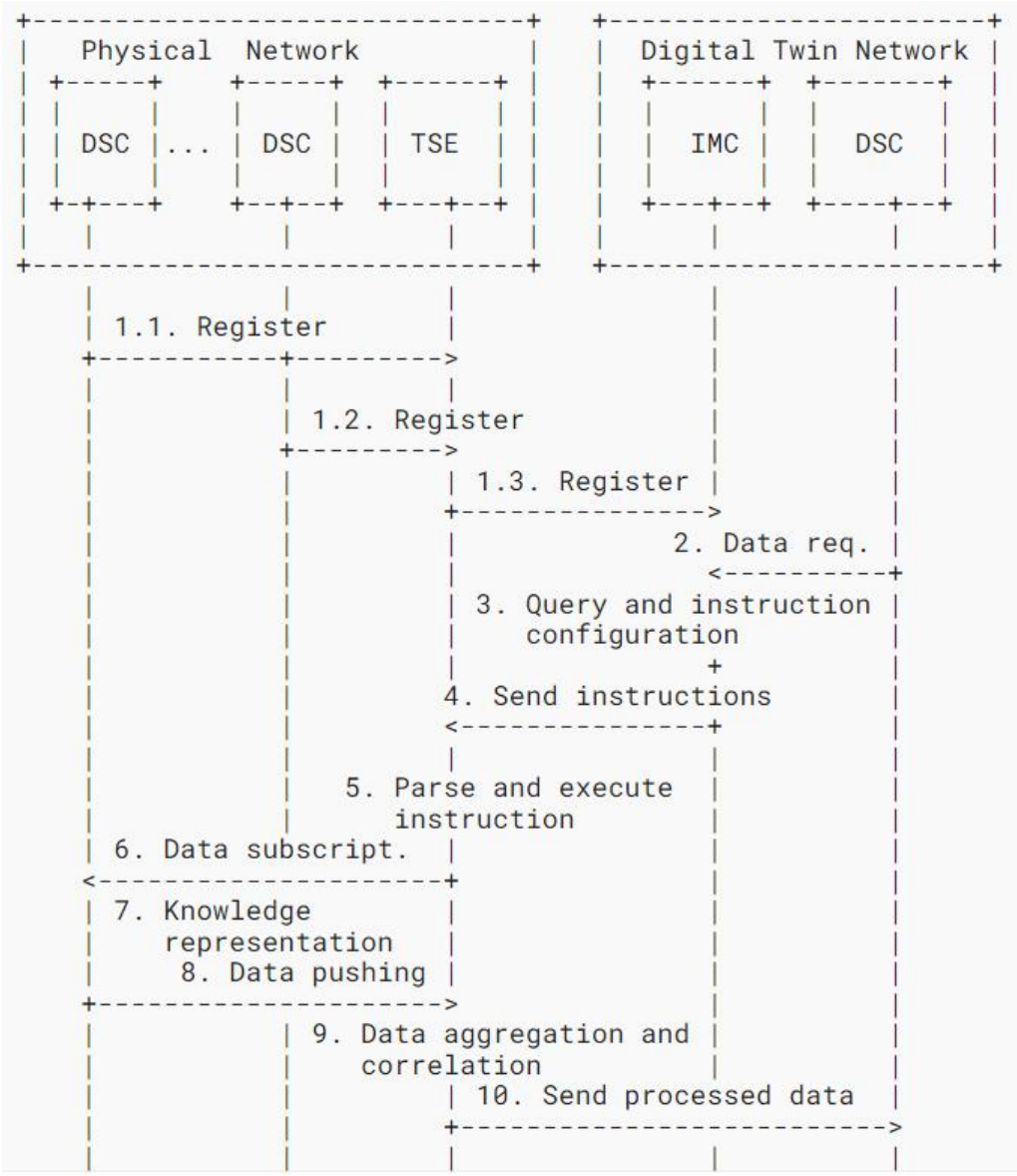
- Store the effective information after data processing and knowledge representation returned by the TSE in PN

TSE: Telemetry Streaming Element **in PN**

- data collection
- Data aggregation
- Data correlation
- Data query

Knowledge and Instruction Driven Data Collection for DTN - Detailed Procedures

1. Register: DSC in the PN side registers into the TSE. The TSE registers into the IMC.
2. DSC in the DTN: send data Request
3. IMC: query and config
4. IMC: send Instructions
5. TSE: Parse and execute instructions
6. TSE: send data subscription
7. DSC in PN: represents the data semantically in RDF form or sends the data in raw form to the TSE for it to make the semantic representation.
8. DSC in PN: pushes the data or knowledge item to the TSE
9. TSE: aggregates and correlates the collected data or knowledge.
10. TSE: sends the processed data or knowledge to DSC in TN



Next Steps

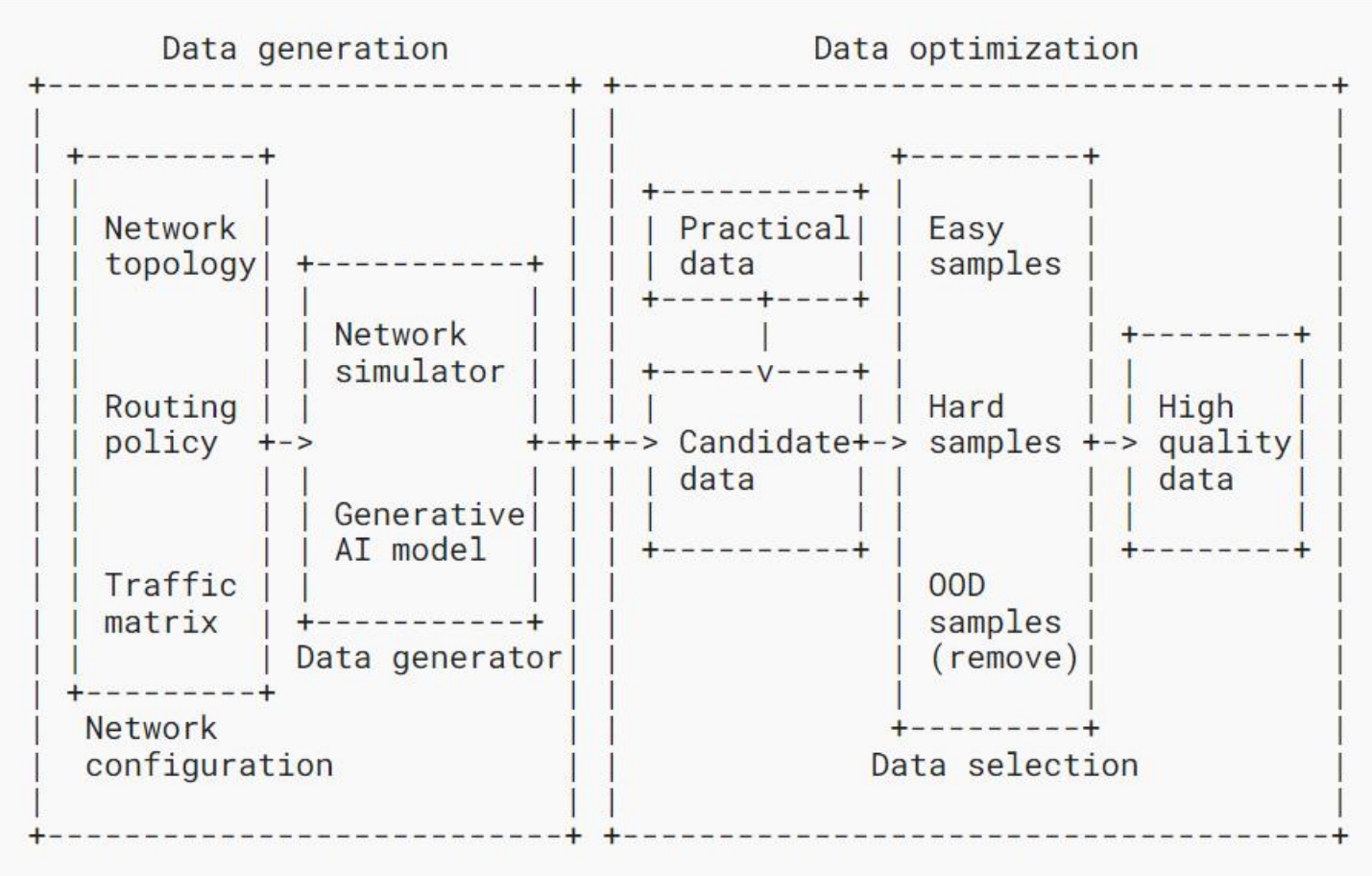
- Call for more efficient data collection methods suitable for DTN system to enrich the draft.
- Verify the data collection methods in a DTN demo system.

- **Looking forward to the comments and suggestions**.

One More Thing
- Research on data generation and data optimization

- Motivation
  - Practical data shortage
  - Data privacy protection
  - Various data sources with varying quality
- Goals
  - Generate simulated data
  - Select high-quality data from multi-source data
- Initial draft:

  https://www.ietf.org/archive/id/draft-li-nmrg-dtn-data-generation-optimization-00.html
- Call for inputs
  - Data generation methods
  - Data optimization methods
  - Deployment



Framework of Data Generation and Optimization for DTN Performance Modeling