

Investigation of Required Network-Storage System toward Fusion Information Science Center in Rokkasho^{*)}

Shinsuke TOKUNAGA¹⁾, Hideya NAKANISHI^{1,2)}, Kenjiro YAMANAKA³⁾, Takahisa OZEKI¹⁾,
Yuki HOMMA¹⁾ and Yasutomo ISHII¹⁾

¹⁾National Institute for Quantum Science and Technology, 2-166 Omotedate, Obuchi, Rokkasho, Aomori 039-3212, Japan

²⁾National Institute for Fusion Science, 322-6 Oroshi-cho, Toki, Gifu 509-5292, Japan

³⁾National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

(Received 10 January 2022 / Accepted 26 April 2022)

Fusion Information Science Centre is being planned as a research infrastructure integrating functions of ITER Remote Experimentation Centre, Computational Simulation Centre and replicated ITER DB. Conceptual design of storage system for the FISC which consists of short-term storage, long-term storage and data warehouse is proposed to satisfy the requirements for fast data transfer, intershot analysis, conventional offline analysis, and machine learning has been studied based on assessment of characteristics of each data access.

© 2022 The Japan Society of Plasma Science and Nuclear Fusion Research

Keywords: remote participation, storage system, ITERDB, data analysis, machine learning

DOI: 10.1585/pfr.17.2405091

1. Introduction

To demonstrate the scientific and technological feasibility of controlled fusion energy, ITER is under construction in Saint-Paul-lez-Durance, France. ITER is a large-scale scientific international collaboration project with seven parties in the world. In order to enable active experimental study using ITER, effective remote participation is essentially important for distant parties from ITER. The ITER Remote Experimentation Centre (REC) has been constructed in Rokkasho, Japan as part of the Broader Approach (BA) activities between JA and EU in order to enable effective remote participation in ITER experiment from Japan [1–6]. Targets of REC project are 1) construction of remote participation environment equivalent to the local ITER main control room aided by state-of-the-art IT technologies and 2) replication of whole data generated in ITER to REC in order to fully utilize ITER results by using its data for data-driven modeling, examination of simulation models, etc. toward DEMO.

Expected function of the REC is shown in Fig. 1. Functions for remote participation in ITER experiment, e.g., capability of shot proposal, live monitoring, and communication between remote sites and the main control room in ITER. It is noted that capability of intershot analysis toward the next discharge is also essential for remote participation. Distance between ITER and Rokkasho is about 10000 km which causes ~200 ms of latency (round trip time). Considerable degradation of the network throughput resulted by such a large latency causes unsmooth response of the analysis software running in the

local site [7]. Although it might be regarded as small inconvenience, it can seriously depress researcher's productivity. Full data replication enables smooth data access without delay for domestic researchers and also reduces onsite computation load and network congestion. Replicated ITERDB in REC will be also conveniently used for offline analyses using computation resources in Rokkasho site including the supercomputer. Domestic researchers can perform active research activities by accessing these computation resources.

Integrating these functions of REC, Computational Simulation Centre (CSC), and replicated ITERDB, “Fusion Information Science Centre (FISC)” in Rokkasho is being planned as a domestic project (Fig. 2). In the FISC, new research activities such as

- integration of empirical modeling (machine learning, artificial intelligence, and so on) and theoretical simulation to develop real-time reliable prediction model of burning plasma,
- design of shot scenarios based on simulation and examination of the simulation using the experimental result,
- active construction of datasets for machine learning using remote experiment based on feedback from the data science

would be enabled by the intimate integration of the three components of the FISC. Key of such intimate integration will be the data storage shared among these components. In the present study, current status of the conceptual design of this FISC storage is reported in the following.

author's e-mail: tokunaga.shinsuke@qst.go.jp

^{*)} This article is based on the presentation at the 30th International Toki Conference on Plasma and Fusion Research (ITC30).

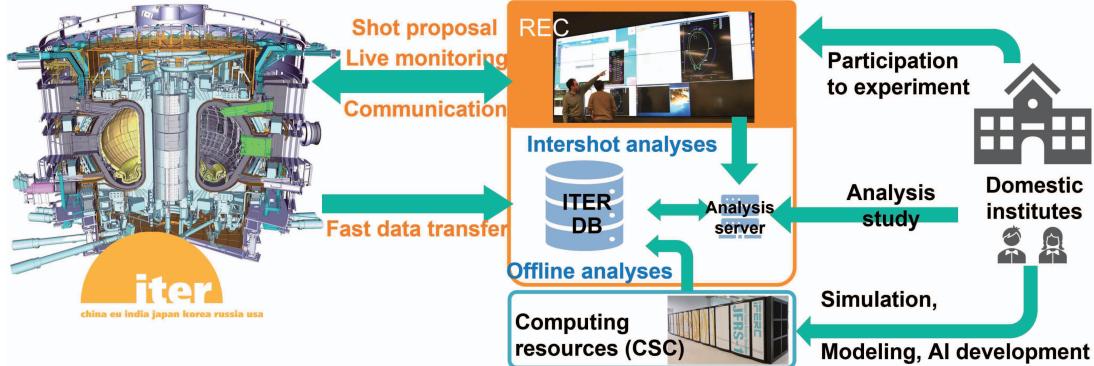


Fig. 1 Remote participation in ITER Experiment via REC.

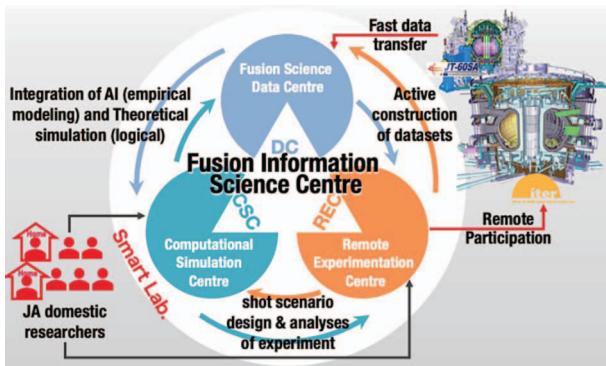


Fig. 2 Concept of the Fusion Information Science Centre.

2. Requirements for the FISC Network-Storage System

The FISC network-storage system needs to satisfy several requirements, such as,

- enough capacity of the storage to store transferred whole ITER data,
- sufficient throughput to receive massive ITER data,
- quick and smooth access to the latest shot data by remote participants in REC for intershot analysis,
- usability of the ITER data to perform offline analyses activities by domestic researchers including data-driven modeling, development of AI and so on based on the ITER data.

In order to realize above requirements, following functions or systems have to be prepared.

- 1) Fast data transfer method to enable replication of the whole data.
- 2) Secure and stable broadband connection between ITER Organization (IO) and REC.
- 3) Quick and smooth access to essential data demanded in intershot analysis for remote participants.
- 4) Intimate linkage between ITER database and computation resources in Rokkasho for efficient data-driven modeling studies.

As an important infrastructure to promote research activities based on ITER data toward DEMO, all of these functions have to be implemented in the FISC network-storage system. Current status of preparation of each function is explained in the following sections.

1) Fast data transfer method

Fast data transfer method to enable whole data replication from ITER to REC is demanded. Distance between ITER and REC (10,000 km) causes ~200 ms of latency (round trip time). It causes significant degradation of data throughput of communications based on TCP/IP. Thus, replication of huge data such as whole ITER data over the high-latency network requires special technique to overcome the problem. A demonstration using MMCFTP has been carried out to prove capability of the data replication [8,9]. Pulse interval of the ITER experiment is defined as longer than 30 minutes while generated data size is estimated as ~1TB/shot. Therefore, 1 TB of data was transmitted every 30 minutes, and 105TB was stably transferred in 50 hours in total. Average data throughput ~7.9 Gbps (target speed was 8 Gbps in 10 Gbps connection route) was observed which successfully proved the feasibility of whole data replication for the earlier phase of ITER experiment. Hence the MMCFTP recorded 416Gbps of data transfer in 2019 between Japan and US while the maximum sustained data flow on DAN (Data Archiving Network) in IO in the latter phase of the ITER experiment is 50 GB/s (400 Gbps) [10], fast data transfer method for transcontinental replication of the whole ITER data is ready.

2) Secure and stable broadband connection between IO and REC

In order to enable secure and fast data transfer using MMCFTP, high bandwidth and encrypted transfer route has to be prepared. A layer-2 VPN should be the most preferable and cost-effective option for this purpose. Such a L2 VPN between IO and REC has been established in 2020. In the ITER site, the most important network segment around the ITER plant, so-called POZ (Plant Operation Zone) is strictly isolated and cannot be accessed



Fig. 3 Network segments in ITER and REC.

from the outside. External network to POZ is called as XPOZ (External to POZ). The XPOZ is interface between the POZ and outer networks. A dedicated isolated network in REC is connected to the XPOZ via L2 VPN (Fig. 3).

IO connects to the French provider RENATER and, via it, to the pan-European network GÉANT. Currently, IO has two redundant 10 Gbps physical connection to RENATER. GÉANT provides connection from Marseilles to Amsterdam with bandwidth more than 100 Gbps. GÉANT is connected to the Japanese provider SINET at Amsterdam. EU-JA connection between Amsterdam and Tokyo is 100 Gbps at this moment. The domestic line of the SINET from Tokyo to Aomori is also 100 Gbps. Connection between the SINET Aomori DC and REC in Rokkasho is currently 10 Gbps. These networks are continuously being upgraded. A project to provide 200 - 400 Gbps optic fiber connectivity from ITER to a data center in Marseilles is under study. GÉANT is planning to upgrade the major route to 400 Gbps. Domestic lines of the SINET will be upgraded to 400 Gbps in 2022. The connection between REC in Rokkasho and its nearest SINET access point (Kamikita DC) will be also upgraded to 100 Gbps in 2022. Thus, the L2 VPN connection between IO and REC will have sufficient bandwidth of 100 Gbps for the ITER first plasma.

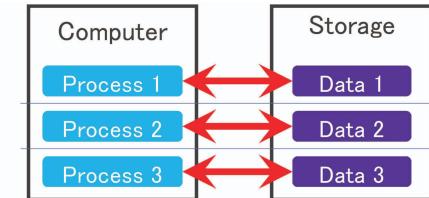
Therefore, requirements 1) and 2) are expected to be satisfied by the MMCFTP and the L2 VPN connection with sufficient bandwidth, respectively. Then, conceptual design study of the FISC network-storage system to satisfy the requirements, 3) quick and smooth access to essential data for inter-shot analysis and 4) intimate linkage between ITER database and computation resource in Rokkasho are considered in the following sections.

3. Conceptual Design Study of FISC Storage

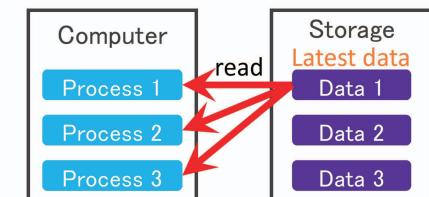
3.1 Difference of access pattern for several data analysis

As mentioned in the introduction, the FISC storage will serve the data for several different use cases of analyses, i.e., intershot analysis, conventional offline analysis and statistical analysis including data-driven modeling or developments of AIs based on machine learning. Thus, four kinds of access are expected to the FISC storage as below.

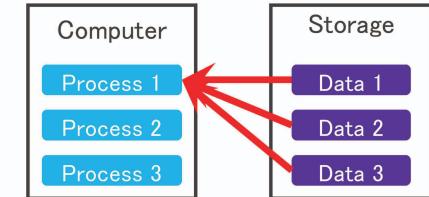
- I. Fast Data Transfer
- II. Intershot Analysis



(a) Typical access pattern of supercomputer



(b) Typical access pattern of intershot analysis



(c) Typical access pattern of machine learning

Fig. 4 Assessment of data access.

III. (Conventional) Offline Analysis

IV. Machine Learning

In order to enable smooth and quick data accesses, sufficient data throughput is demanded for the FISC storage.

One of the fastest storage systems is one used with supercomputer. For example, data throughput of the *Lustre* storage of a supercomputer in Rokkasho (JFRS-1) is 135 GB/s. Lustre allows clients (processes) to access multiple OSS (Object Storage Server) simultaneously and independently each other. Thus, total throughput of the file system can simply scale as the number of hardware grows. However, "data access pattern" for supercomputer storage is significantly different from expected access patterns for inter-shot analysis or machine learning (Fig. 4). In a supercomputer, each process submitted by users makes disk I/O basically only to the corresponding working directory mainly to WRITE the calculation results (Fig. 4(a)). It does not access to the other disk assigned to the other users. Storage system for supercomputer is optimized for such purely independent process-data relationships. On the other hand, in case of the intershot analysis (Fig. 4(b)), multiple pro-

Table 1 Characteristics of data I/O related to four kinds of data access.

	Fast Data Transfer	Intershot Analysis	Offline Analysis (Conventional)	Machine Learning
Related Data Size	Middle (~TB)	Small (~GB)	Huge (~PB)	Large (?)
Needed Throughput	High	High (rush to the latest data)	Middle	High
Storage Period of Related Data	short	short	long	long
Needs of Extensibility (Scale-out)	low	low	high	high

The diagram shows two arrows pointing downwards from the 'Machine Learning' row of the table to two separate text boxes. The left arrow points to a yellow box containing the text 'High-speed, short term, rather small storage'. The right arrow points to a green box containing the text 'Huge and High speed?'

cesses induce READ access rushing only to the latest data, simultaneously. In such a use case, obviously “I/O speed of a corresponding disk” will become the bottleneck. In another case for machine learning (Fig. 4(c)), a learning process will demand access to various and wide range of shot data. In such a case, it is prospected that “performance of the data locator” could be one of the determinant factors of the throughput. Preparing a huge homogenous storage serving data with sufficient throughput for any of such various access pattern would be very costly and inefficient. Thus, a distributed storage system which consists of various adequate storages is considered in the following.

3.2 Analysis of accesses to the storage

In order to consider the best mix and composition of the storages for the FISC storage, characteristics of the data I/O related to four kinds of data access, i.e., Fast Data Transfer (FDT), Intershot Analysis (ISA), Conventional Offline Analysis (COA) and Machine Learning (ML) were analyzed. Assessment of several important characteristics of the data access is summarized in the Table 1.

At first, the FDT is “WRITE” access to the storage in contrast to the other three data accesses to “READ”. The FDT will import experimental data generated in ITER after every discharge. It works as single process to make massive write access to the storage system intermittently. The related data size with this I/O will be the size of one-shot data in ITER. The one-shot data size in ITER is expected to be 1 TB in the early phase and could be ~100 TB in the latter DT phase. As is mentioned in the section 2 (1), the FDT will be carried out using highspeed data transfer method: MMCFTP, and its throughput can reach several hundred Gbps. Once the data has been received by this temporal storage from ITER to Rokkasho, the corresponding data will be moved to a long-term storage sooner or later. Therefore, the storage period of this temporal storage for the FDT can be short. Since the data will be accumulated in the long-term storage and will not stay long on this temporal storage, this temporal storage would not be intensely scaled-out.

Considering the data access for ISA, subject data which can be analyzed in the ISA for next shot during the shot-interval (30 min or longer) is limited to essential dataset. There should be no enough time to perform de-

tailed analyses with comprehensive data. Thus, the related data size of the ISA is rather small. However, a lot of read access will rush to the latest data as explained in the previous subsection, and it will require high throughput. Once the relevant experimental campaign was finished, the corresponding data would be no longer subject of the ISA. The subject data of the ISA will be replaced by newer data, and old data will be moved to the long-term storage. Thus, the storage period of the ISA related data can be short. It will not be strongly demanded to be scaled-out either.

Characteristics of the data access related to FDT and ISA is similar. These demands can be satisfied by a high-speed, short-term, rather small storage. An all-flash storage can be a good candidate for this part. It also meets the requirement of “I/O speed of a corresponding disk” for Intershot Analysis explained in the Fig. 4 (b).

Considering the data access related to the COA, it is impossible to scope or limit the subject data. Smooth data access to all of the data must be assured. Thus, this data access should subject the long-term storage including archived data. The data size will reach several ten petabytes even in the first plasma phase and would reach exabyte in the DT phase. Preparing such a massive storage with ultra-fast I/O speed will need unrealistic cost. In the conventional analysis (here it means non-statistical analysis), of course the data access speed is faster is better. However, such activities are often carried out based on interactive manual iteration. In other word, a few seconds of data reading time might not be so critical bottleneck compared to the human actions. Therefore, demand of the throughput for this data access could be regarded as moderate. Since the subject data used in COA is all ITER data including archives, its storage period is longer than lifetime of the ITER itself. It has to be rapidly scaled out to accumulate all of the data generated in ITER as the experiment progresses.

Application field of the ML studies are now rapidly broadening. It is not so easy to predict the data scope to be interested in the context of the ML at present. So, the corresponding storage system has to provide flexibility to adapt the new research activities in future. Usually, the ML is applied to some prepared relevant dataset extracted from the whole database in advance. Therefore, the subject dataset of the ML will be distilled from all of the ITER

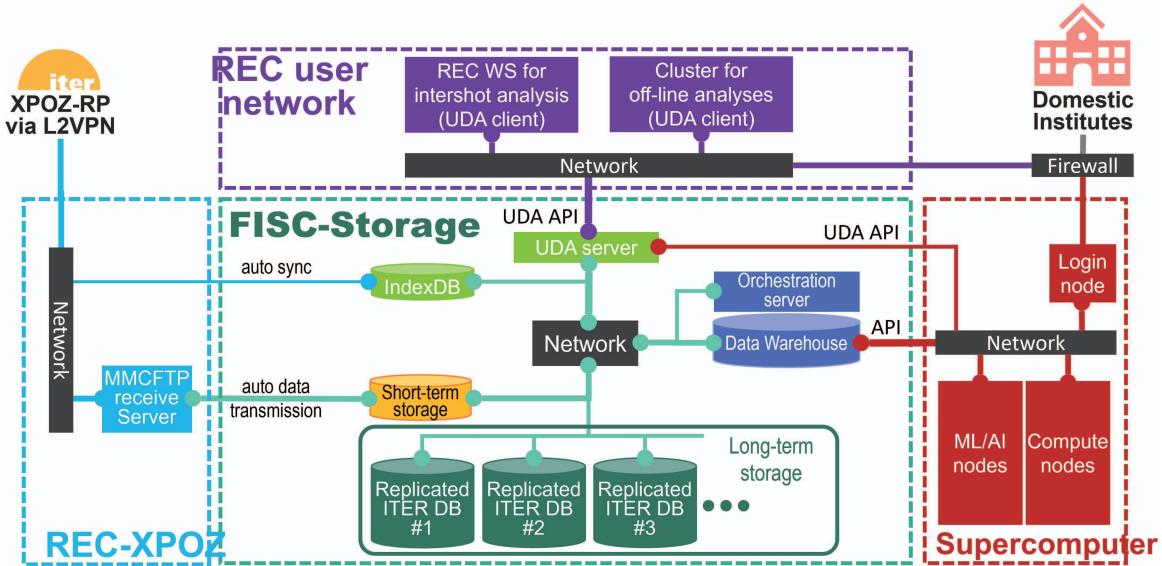


Fig. 5 Conceptual design of FISC storage consists of short-term storage, long-term storage, and data warehouse.

database, and it would be smaller than the whole database. Since the ML study will statistically process large number of data on powerful GPU nodes, requirement of the data throughput is intensive. As is discussed in the section 3.1, performance of the data locator would be important for such kind of access pattern. Time series data of essential plasma physics and plant operation will be subjected by the ML studies over the long period. And such storage with time series data will also need to be scaled-out as the ITER experiment progresses.

It is often said that data scientists spend most of their time for cleaning and organizing dataset for machine learning. Such pre-processing workload is necessary in order to prepare relevant dataset before the ML studies. In order to promote active data-driven modeling studies based on the ML, reducing this researcher's workload by automated “Data Orchestration” of the replicated data is very important. Such automated pre-processing is crucial especially to handle huge data like the ITER data. Therefore, the related data to the ML activities should be extracted data and different from the whole database for the COA. The dataset for the ML activity would be generated by orchestration process and might be stored with some DBMS so as to enable fast response of the data locator. This kind of storage for pre-processed data can be regarded as so-called Data Warehouse (DWH). This DWH for serving pre-processed data for data access related to the ML should be separately prepared beside the general long-term storage for the COA.

As the result of the assessment of the data I/O related to four kinds of data access which is summarized in the Table 1, requirements for the FDT and the ISA will be satisfied by a high-speed, rather small, short-term storage. An all-flash storage server would be suitable candidate for this purpose. The storage system for the COA will be huge and comprehensive long-term storage. Severe requirement of

the extensibility and data protection strategy will be the key issue for this huge long-term storage, while the requirement of the throughput is moderate. So-called “object storage” system on distributed storage might be worth to consider as a candidate for this long-term storage to satisfy requirement of smooth extensibility. Data protection method is also crucial issue for such a massive storage. The “Erasure Coding (EC)” technology [11] for networked distributed storage system which is often available with object storage system would be promising solution for data protection of exabyte scale storage. EC provides data redundancy by breaking a data unit (file or object) into fragments (data blocks), which are then expanded with additional fragments (parity blocs) that can be used for data recovery and stored over distributed storage media. It can reduce the storage efficiency by approximately 50% compared to replication while maintaining the same durability of the data. The storage for the ML needs flexibility to adapt the new research activities in future, and fast data location service would be important for the throughput of the ML access pattern. This storage would be prepared as a DWH to serve the pre-processed data for the ML studies.

Consequently, it is concluded that the rational design of the FISC storage would be composition of the following storages.

- Short-term storage: to receive fast data transfer (FDT) and to serve data for intershot analysis (ISA). Fast disk I/O will be important.
- Long-term storage including archived data: to serve data for conventional offline analysis (COA). Extensibility and data protection strategy will be important.
- Data warehouse: to serve pre-processed dataset for machine learning (ML). Performance of the data locator would be important.

The conceptual design of the FISC storage consists of above parts is depicted in the Fig. 5. Further detailed design studies will be continued based on this conceptual design.

4. Summary

Toward remote participation in ITER experiment, preparation of the ITER Remote Experimentation Centre (REC) is ongoing in Rokkasho, Japan as the collaboration project between JA and EU. Targets of the REC project is construction of remote participation environment equivalent to the local ITER main control room aided by state-of-the-art IT technologies, and replication of whole data generated in ITER to REC in order to make the most of ITER results by using the data for data-driven modeling, examination of simulation model, etc. toward DEMO. Analysis system of ITER experiment data based on full data replication enables smooth data access without delay for domestic researchers and also reduces onsite computation load and network congestion.

Fusion Information Science Centre (FISC) is being planned as a JA domestic project integrating REC, Computational Simulation Centre, and ITERDB. A conceptual design study of the FISC network-storage system was carried out. Characteristics of four kinds of access to the FISC storage, fast data transfer, intershot analysis, conventional offline analysis, and machine learning were analyzed in terms of related data size, needed throughput, storage period of related data, and need of extensibility. Difference of the data access pattern was also considered. Short-term storage with sufficient data I/O speed will be demanded to receive data transferred by fast data transfer method and

to serve the data for intershot analysis. All flash storage would be a suitable candidate for this storage. Long-term storage including archived data will be demanded to serve all data for conventional offline analysis. Requirements of the throughput for the long-term storage could be moderate. Distributed object storage system is potentially a preferable candidate for this to satisfy intense requirement of smooth extensibility. As the data protection strategy of huge data storage like this long-term storage, the erasure coding technology may provide preferable solution. Data warehouse will be demanded to serve pre-processed (orchestrated) dataset for machine learning process.

Conceptual design of the FISC storage consists of short-term storage, long-term storage, and data warehouse to satisfy the requirements for four kinds of data access was presented. Reconciling such distributed composite storage system with UDA (Unified Data Access) system developed by ITER Organization will be crucial issue in the next step.

- [1] D. Stepanov *et al.*, *Fusion Eng. Des.* **86**, 1302 (2011).
- [2] T. Ozeki *et al.*, *Fusion Eng. Des.* **89**, 529 (2014).
- [3] G.D. Tommasi *et al.*, *Fusion Eng. Des.* **96-97**, 769 (2015).
- [4] T. Ozeki *et al.*, *Fusion Eng. Des.* **112**, 1055 (2016).
- [5] J. Farthing *et al.*, *Fusion Eng. Des.* **128**, 158 (2018).
- [6] S. Clement-Lorenzo, 27th IAEA Fusion Energy Conference (2018), FIP/P1-16.
- [7] S. Tokunaga *et al.*, *Fusion Eng. Des.* **154**, 111554 (2020).
- [8] K. Yamanaka *et al.*, *Fusion Eng. Des.* **138**, 202 (2019).
- [9] H. Nakanishi *et al.*, *Plasma Fusion Res.* **16**, 2405017 (2021).
- [10] ITER IDM, System Design Description for CODAC, CODAC DDD v3.0 (ITER_D_6M58M9) (2021).
- [11] A. Datta *et al.*, *ACM SIGACT News*, 44, pp. 89-105 (2013).