DOI: 10.14529/jsfi230308

Digital Twins in Large-Scale Scientific Infrastructure Projects

***Denis V. Kosyakov*[1,2], *Mikhail A. Marchenko*[2]**

The article provides an overview of publications on the topic of Digital Twins of large-scale scientific infrastructure. History, basic concepts and definition of Digital Twins are given. Main terminology in the field of big science and large-scale scientific infrastructure is also described. In Russian practice, the large-scale scientific infrastructure projects are often referred to as "megascience installations". Such installations usually include facilities for research in areas such as astronomy and high-energy physics. The research infrastructure is a complex of construction facilities, engineering systems, precise control and measuring equipment, characterized by high complexity and strict requirements for all operational parameters. In addition, these facilities are associated with high operating costs, are sensitive to minor changes in their condition and environmental conditions, and carry the risk of data loss during long-term and unique experiments. Then, information about the use of Digital Twins in large scale astrophysical projects and also for particle accelerators control and tuning is provided. Potential areas of application of Digital Twins in large projects of scientific infrastructure are summarized. Necessary information about the Siberian Circular Photon Source (SKIF, in Russian) is given. On the basis of the review, and goals and objectives for the Digital Twin of the SKIF are determined. An analysis of the necessary computing resources and data storage volume is also carried out.

Keywords: digital twins, neural networks, supercomputing modelling, large-scale scientific infrastructure, Siberian Circular Photon Source.

Introduction

The term of "Digital Twin" (DT) was coined by Michael Grieves at the beginning of the 21st century. He originally introduced the concept during a presentation at University of Michigan in 2002, at an industry event dedicated to the creation of a Product Lifecycle Management (PLM) center. Grieves later expanded this idea into a course of lectures and in a white paper [1], as well as in a follow-up 2016 paper co-authored with John Vickers [2].

Already in 2011, the first journal article appeared on this topic [3]. It explored how digital twins can be effectively used to predict the lifespan of an aircraft structure. In 2012, NASA formalized the definition of DTs and highlighted their potential applications in the aerospace industry [4]. The period up to 2014 is considered to be the incubation stage of DT research [5]. At this stage of development, researchers began to explore the broader implications and potential applications of DTs, laying the groundwork for further advances in the field.

In his initial presentation, Grieves created the concept of "The Conceptual Ideal for PLM". This early description already included all basic elements of a digital twin, such as real space, virtual space, flow of data from the real to the virtual space, flow of information from the virtual to the real space, and virtual subspaces. The driving principle of the model was the existence of two interconnected systems: a physical system that has always existed, and a new virtual system that contains all the information about the physical system. This connection created a mirror effect between the real and virtual spaces. The inclusion of PLM in the title emphasized that the concept was not limited to a static view, but rather encompassed the dynamic relationship between the two systems throughout the product lifecycle. This connection persists as the

[1]Russian Research Institute of Economics, Politics and Law in Science and Technology, Moscow, Russian Federation
[2]Institute of Computational Mathematics and Mathematical Geophysics SB RAS, Novosibirsk, Russian Federation

system goes through four main phases: design, production, operation (maintenance/support) and disposal.

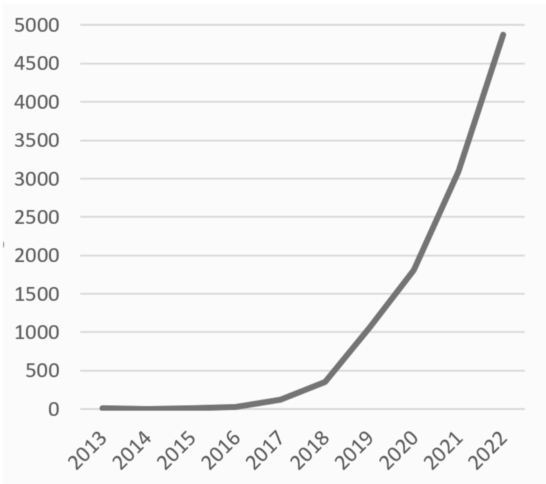
The concept was based on rapid advances in the performance and availability of computer systems, the development of computer-aided design (CAD) systems [6] and computer simulation systems [7]. In the first phase, the focus was on creating an accurate digital representation of physical objects, primarily through the use of CAD. This approach allowed to simulate the behavior of a real object using its DT at all stages of its life cycle. These simulations were based on or compared to the real characteristics of a physical object under real conditions [8]. This emphasis on accuracy and reproducibility has allowed DTs to serve as tools for studying and optimizing an object's performance throughout its lifetime.

The further evolution of digital twin technology is closely linked to advances in various related fields, including sensors and monitoring systems, the Internet of Things (IoT), industrial control systems (ICS), and the development of Industry 4.0 [9]. Advances in sensor technology and monitoring systems have made it possible to collect real-time data on physical objects that can be used to update and optimize their digital counterparts. The development of the concept of the Internet of Things [10], especially in the context of industrial production [11], has facilitated the integration of DTs into connected ecosystems, where different devices and systems can communicate and exchange information with each other. Advances in ICS have contributed to more efficient and automated production management, which has further facilitated the adoption of DTs in manufacturing processes. The concept of Industry 4.0, with its emphasis on cyber-physical systems [12], highlighted the importance of DTs as a means of bridging the gap between the physical and virtual worlds. In addition, significant advances in data collection, storage, and processing technologies have played an important role in processing large amounts of data generated by DTs [13].

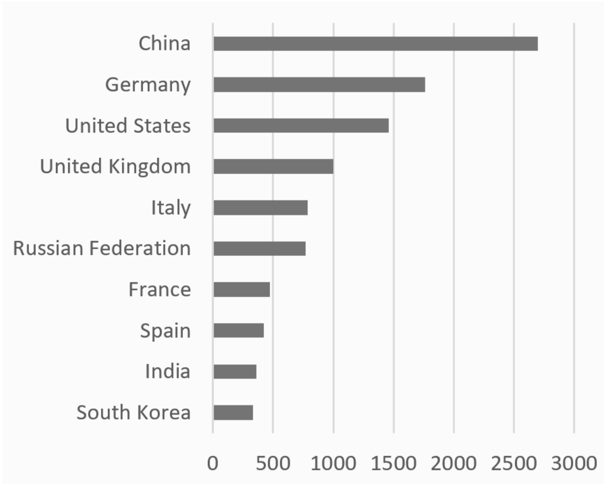
Recent advances in the field of artificial intelligence, namely in neural network technologies, have given another impulse to the development of the DT concept [14, 15]. Neural networks make it possible to build computer simulations of complex physical systems on the basis of training on a stream of real data according to the "black box" principle, which is especially useful if it is impossible or highly complex to use more traditional numerical simulations [16]. Thus, artificial intelligence technologies are one of the main components influencing the development of the DT concept [17].

In the graph showing the dynamics of the number of publications according to the Scopus database data (Fig. 1a), you can see several inflection points, a noticeable acceleration occurred in 2018, an explosive growth in the number of publications followed after 2020. Currently, this aspect is on the rise, coming out on top among other technologies. More than 17% of DT-related publications also contain one of the AI-related terms in keywords. Both leading developed countries and new industrial leaders, such as China, India, and South Korea, are interested in the topic – number of publications for the entire period is given on Fig. 1b.

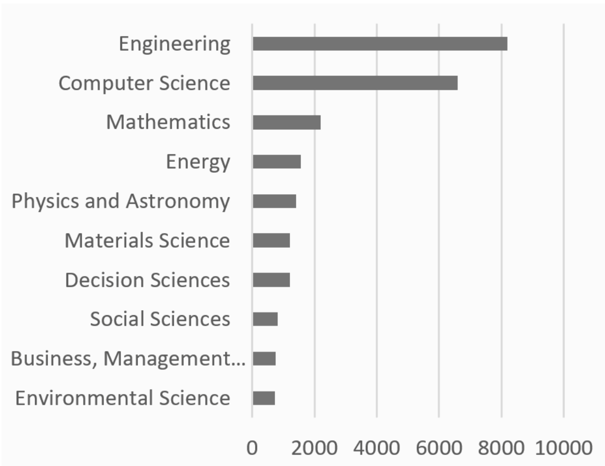
Significant research interest in DT is noted in the field of product lifecycle management [18], industrial production [5, 19]. Kritzinger and colleagues note that at the time of writing the review, a significant number of studies were conceptual in nature with little addition to analysis or planning of real use cases. Particular attention is paid to the use of DT at the operation stage [20]. Applications of data centers in Smart City systems are developing [21, 22]. As expected, there is a convergence of concepts and technologies for Building Information Modeling (BIM) and DT [23–25] (Fig. 1c).



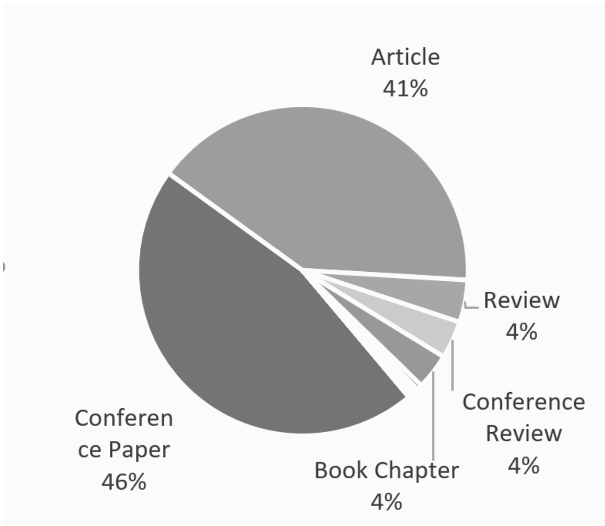
(a) Number of publications about DT summarized by all categories



(b) Contribution of the top 10 countries to the publication flow about DT



(c) Number of publications about DT in the top 10 thematic categories



(d) Distribution of the number of publications about DT by document type

Figure 1. Analysis of the topic of DTs based on publications indexed in Scopus

The novelty of the topic and the activity of researchers in this field, the high rates of scientific communication form a demand for holding thematic conferences and publishing relevant materials, which is why a significant part of publications on the topic is published in conference proceedings (46%), which is significantly higher than the average level (14.5%) (Fig. 1d). The share of reviews (4%) is slightly below the average for all topics (5.7%), which, probably due to relative novelty.

The subject of this review is the application of DT concepts and technologies in large-scale scientific infrastructure projects, which are often referred to as "megascience installations" in Russian practice. Such installations usually include facilities for research in areas such as astronomy and high-energy physics. The research infrastructure is a complex of construction facilities, engineering systems, precise control and measuring equipment, characterized by high complexity and strict requirements for all operational parameters. In addition, these facilities are associated with high operating costs, are sensitive to minor changes in their condition and environmental conditions, and carry the risk of data loss during long-term and unique experiments.

Given the rapidly expanding application of DTs in the industries listed above, it can be assumed that these concepts and technologies should be widely used in the design, monitoring, control, management, and maintenance of large-scale scientific infrastructure. Integrating DTs into such facilities has the potential to improve the efficiency and reliability of these infrastructure while reducing operating costs and minimizing the risks associated with data loss and environmental factors.

The article is organized as follows. Section 1 is devoted to history, basic concepts and definition of Digital Twins. In Section 2 we describe main terminology in the field of big science and large-scale scientific infrastructure. Section 3 contains information about the use of Digital Twins in large scale astrophysical projects. In Section 4 we provide information on application of Digital Twins for particle accelerators control and tuning. Section 5 gives brief information about the Siberian Circular Photon Source (SKIF, in Russian). In Section 6 one can find goals and objectives for Digital Twin of the SKIF. Section 7 provides analysis of necessary computing resources and data storage volume for the Digital Twins of the SKIF. Conclusion summarizes the potential areas of application of Digital Twins in large projects of scientific infrastructure.

1. Basic Concepts and Definitions

According to the Russian State Standard [53], a digital twin is a system comprising a digital model of a product and bidirectional information connections with the product or its components. Let us investigate the previous publications on this topic that by many ways have made technological basis for the Standard.

Grieves and Vickers gave the following definition of DT: "The DT is a set of virtual information constructs that fully describes a potential or actual physical manufactured product from the micro atomic level to the macro geometrical level" [2]. As the main focus of this concept was on production and the product life cycle, they also introduced the concepts of prototype DT and instance DT with corresponding requirements. There are many attempts to supplement or expand this definition [5, 9, 16, 18–20]. The authors note that due to the multiplicity of concepts and solutions in different application areas, as well as the substitution or inclusion of other concepts in the DT concept, there is a significant diversity in the understanding and interpretation of this concept. In the definition given by Glaessgen and Stragel in 2012 [4], a DT consists of three main elements: a physical product, a virtual product, and flows data linking physical and virtual products.

According to the monograph [59], DT is a set of approaches and solutions designed to solve the problem that the increasing complexity of modern systems, as well as multi-component and multi-functional products, is outpacing the growth in the capabilities of tools for their design, manufacture and safe maintenance. And the solution to this problem lies in combining a range of digital technologies to offer more efficient means of modeling, designing, building and maintaining such complex systems. These tools must comprehensively describe the object, seamlessly integrate with each other, ensuring digital continuity of the product creation environment, work in close information connection with the subject of modeling and track all phases of its development (phases of the life cycle). The main element in this definition is integration: integration of digital technologies, integration of modeling tools based on continuous information exchange at the data level, integration of the model and modeling object based on information exchange in near real time, integration of modeling at all stages of its life cycle.

From this definition and other interpretations, the following key attributes of DTs can be distinguished [18]:

- real-time reflection: DTs include both physical and virtual spaces, with the virtual space serving as a precisely synchronized and accurate reflection of the physical space;
- interoperability and convergence: this function can be viewed from three perspectives:
  - interaction and convergence in physical space: DTs provide comprehensive integration of all flows, elements, and services, ensuring the interconnection of data generated at different stages;
  - interoperability and convergence between log/archive and real-time data: DTs depend on a variety of data sources, including expertise and real-time information from all implemented systems, enabling deeper analysis and more efficient use of data;
  - interaction and convergence between physical and virtual spaces: in DTs, physical and virtual spaces are interconnected by seamless channels that allow for easy interaction between the two spaces;
- self-development: DTs support real-time data updates, allowing for continuous improvement of virtual models by comparing them in parallel with their physical counterparts.

Although "digital twin" is the most commonly used term, "digital model" and "digital shadow" are often used interchangeably to refer to digital representations of physical objects. This trio of terms corresponds to the evolution of the concept: "digital model" represents a virtual copy of a real physical object, "digital shadow" is its synchronous "reflection", updating its state based on data about a real object in real time, while a full-fledged "digital twin" also provides data transfer to control systems of a real object, allowing to correct its state. Some authors use the term "digital twin" to refer to "digital model" or "digital shadow".

As we can see, the most important element of the DT concept is the digital representation of a real or projected object in cyberspace. At the same time, if the initial concept assumed a high degree of detail and completeness of this model, later supplemented by the requirement of full-fledged behavior modeling, then there were disagreements about the information relationships between the real object and its DT almost immediately. Later, as the concept extended to a wide range of applications, the idea of the required degree of detail and completeness of the digital model also changes. Taking into account the specifics of the tasks to be solved, it is often sufficient to model only specific aspects of physical objects and systems in one way or another, including on the basis of statistical models [26] or machine learning methods [14, 27].

2. On Terminology in the Field of Big Science and Large-Scale Scientific Infrastructure

The term "megascience" was legitimized in 1992, when the OECD created the Megascience Forum [28]. This forum was then renamed the Global Science Forum. The term "megascience" is actively used in Russian practice, both in official documents and in scientific literature. Of the 127 results that Scopus provides for the request TITLE-ABS-KEY ("megascience"), 76 belong to Russian authors.

In foreign literature, the umbrella term "Big Science" has become more widespread (987 publications in Scopus), and this term refers to any large-scale projects, including those not related to unique scientific facilities [29, 30]. In the context of large scientific facilities or other examples of research infrastructure and related research projects, the term Large Scale Research Infrastructure is used [31]. At the same time, this term is rarely used in the case of discussing specific projects of Big Science, which somewhat complicates the task of finding publications related to such projects and the use of DT technologies in them.

In this regard, the search was organized considering the main scientific areas in which research infrastructure of the corresponding class exist, are being built or designed. Such areas primarily include astronomy, particle accelerators, colliders, fusion reactors, and neutrino detectors. The search was also carried out by the names of the most well-known projects in these areas. The search results showed an extremely small number of publications about DTs in large-scale scientific infrastructure projects.

3. Digital Twins for Astronomical Complexes

DT technologies are used or planned for use in three major astrophysical projects. The Australian Square Kilometre Array Pathfinder (ASKAP) project is a radio telescope consisting of 36 parabolic antennas, each 12 metres in diameter, distributed in two dimensions with baselines up to 6 kilometres [32]. Each antenna is equipped with a monitoring system that measures parameters such as temperature, voltage, orientation and system status. DT technologies are used to create a realistic virtual representation of everything a complex that reflects the current state, allows displaying log/archive and current data on monitoring indicators. The developed DT can be used both for controlling the complex and scientific experiments, as well as for creating realistic visualizations, including for the purposes of public communication [33]. This project is part of the international Square Kilometre Array (SKA) project, the other part of which is the MeerKAT telescope [34] is based in South Africa. The use of DT technologies for the effective operation of these complexes is discussed in the report [35].

China's Five-hundred-meter Aperture Spherical radio Telescope (FAST), also known as the Tianyang (Celestial Eye), is the largest filled-aperture radio telescope in the world, located in southern China, officially commissioned in 2020 [36]. An important structural element of the FAST telescope is a flexible network of cables that supports the structure of the active reflector and allows the geometry of the active reflector to be changed. This network is equipped with more than 500 sensors for condition monitoring. Structural changes in the network associated with material fatigue are a key technical problem that determines the condition of the entire telescope. A DT of this cable network, based on data received from highly sensitive sensors and other measuring equipment, is designed to monitor and predict the state of the system [37]. The DT uses a physical model made using the ANSYS software package and provides an increase in efficiency and a reduction in the cost of supporting the operation of this complex.

The overall design of the FAST telescope includes other equally important components, respectively, there is a common task of monitoring the technical condition, supporting the operation of other units and components, as well as presenting information about the project in the media and other areas of scientific communication. In the article [38], the method of rapid prototyping used in the creation of a DT of the entire complex, which is a detailed 3D model, as well as a conceptual technical scheme for the use of the data center for the above purposes, is presented. The authors note that for the safe and efficient operation of a plant of this class, along with operational data, it is necessary to collect information about the state of the environment, such as temperature, humidity, atmospheric pressure, wind speed and direction, light intensity, visibility, cloud cover and precipitation. This data can help optimize the operation of the complex in various weather conditions and prevent the negative consequences of dangerous natural phenomena. A prototype of the FAST's DT for the purpose of automated control of the telescope is also discussed in [39].

The extended ROentgen Survey with an Imaging Telescope Array (eROSITA) X-ray telescope project, which is the main instrument on board the Spectrum-Roentgen-Gamma (SRG) mission [40], uses DT technology in a completely different context. In [41] they describe the creation of a data center based on real data obtained during the mission. This DT is used to tune and test the algorithm for detecting clusters of galaxies and active galaxies galactic nuclei. This high-level simulation dramatically increases the understanding of real-world data and enhances the analysis capabilities of the complex arrays of signal sources observed by eROSITA.

4. Application of Digital Twins for Particle Accelerators

The booster is an integral part of the Fermi National Laboratory (Fermilab) accelerator complex in the USA and provides a flux of low-energy neutrinos for the MicroBooNE experiment [42]. The Gradient Magnet Power Supply (GMPS) is an important subsystem of this accelerator complex and is implemented in the form of four power supplies evenly distributed across the Fermilab Booster. Each feeds one of four complete gradient magnets, which are responsible for controlling and accelerating the 400 MeV proton beam that Booster receives from the linear accelerator to an energy of 8 GeV. GMPS operates in a 15 Hz cycle between injection and beam output. Currently, there is an operator in the control loop setting targets for a proportional-integral-differential control circuit that applies compensating offsets for GMPS. The goal is to create a reinforcement learning-based control circuit to optimize the control process [43]. A DT of the Booster-GMPS system was used to train, calibrate, and validate the model, created on the basis of real data collected during 6 months of operation of the complex based on a long chain of elements of short-term memory (LSTM): one of the architectures of recurrent neural networks.

The Compact Linear Collider (CLIC) is a projected accelerator that is being developed as a complement to CERN's accelerator complex. Its goal is to collide electrons and positrons head-on at energies of up to a few teraelectronvolts (TeV). In order to optimally use its physical potential, the CLIC is supposed to be built and operated in three stages at collision energies of 380 GeV, 1.5 TeV and 3 TeV, respectively, on a section with a length of 11 to 50 kilometres [44]. In the design, construction and operation of such equipment, the spatial alignment of the focusing beams of charged particles of magnetic assemblies is critical. Small deviations of the electromagnetic axes lead to significant errors in the due to the large size of the structures and the decrease in quality or failure of experiments. In the CLIC project, it is necessary to provide spatial alignment of several thousand large assemblies larger than a meter within the target combined standard uncertainty of 12 metres.

The authors in [45] found several gaps in knowledge that limit this possibility. Among them was the lack of uncertainty definitions to compensate for the thermal error applied to correct the instability of assembly dimensions, after metrology, as well as in assembly and alignment time. A new methodology was developed that used a combination of Monte Carlo modelling and high-precision traceable reference measurements to quantify the uncertainty of the various thermal expansion models used. The authors believe that this methodology can be used to create high-precision DTs with known uncertainty not only for accelerators, but also for other large infrastructure facilities with high accuracy requirements even in the case of complex transient modes of operation. Such data centers can help in optimizing the operation of complex equipment and making important decisions.

In [46] authors discuss the prospects for the development of the bERLinPro accelerator recuperator project at Helmholtz-Zentrum Berlin (HZB), which was officially completed in 2020. This accelerator complex will get a new life as part of the Sealab (Superconducting RF Electron Accelerator Laboratory) project. It is assumed that Sealab will become a test site not only for physical experiments and development of technical solutions, but also an ideal object for research in the field of new control schemes for such complexes based on DT technologies and machine learning.

At the IPAC'23 – 14th International Particle Accelerator Conference, held in May 2023 in Venice, three reports were announced, touching on the topic of DTs in solving various problems in high-energy physics. The first report concerns the creation of a DT of the Karlsruhe Research Accelerator (KARA) with information about the energy system within the framework of the ACCESS (ACCelerator Energy System Stability) project in Energy Lab 2.0 in a simulation environment in real time. The goal of the project is to test energy solutions that can be applied to accelerators in a secure and flexible virtual environment, without interfering with the experiments conducted at KARA, while maintaining high accuracy [47].

The second report [48] discusses the creation of software tools for automatic tuning and alignment of the electron source and beam transport line based on machine learning methods and special simulation in the Test Brookhaven National Laboratory Accelerator Center. In this case, it is assumed that the model will be applied to DTs of electron beams in the Sirepo simulation environment, replicating the characteristics of real beams collected through the Bluesky system.

The third report [49] focuses on the implementation of a new product lifecycle management platform at CERN, combining log/archive data and data from different systems and external partners into a common coherent framework. One of the main goals of this platform is to create the foundation for DTs of current and future accelerators with the goal of radically reducing development time, operation and maintenance costs.

In a presentation at the ICALEPCS Data Science and ML Workshop in 2021, Orali Edelen and her colleagues presented a conceptual vision of the capabilities of artificial intelligence, machine learning, and DT technologies at the SLAC National Accelerator Laboratory (Stanford Linear Accelerator Center) in USA [50]. It notes that in the complex of the X-ray free-electron laser Linac Coherent Light Source (LCLS), about 400 hours a year are spent on setting up the equipment, for different experiments the configuration changes several times a day, the tuning cycle takes about 30 minutes. The cost of one hour of the experiment is estimated at 30 thousand US dollars, so the losses associated with long, mainly manual cycles of equipment tuning can be estimated at 12 million dollars.

At the same time, the creation of fast and accurate models that could provide automation of tuning processes is complicated by the need to use computationally complex models, many insignificant but tending to accumulate uncertainties, various kinds of fluctuations and noise, implicit dependencies, nonlinear effects and instability. In this regard, the functioning of the complex depends on the daily and constant work of operators, who ensure the performance of tasks for monitoring and correcting the operation of equipment. The authors of the report see a way out in the active use of machine learning technologies, real-time modeling, including the creation of DTs in the tasks of anomaly detection and fault prediction, diagnostics, automated control and optimization.

5. Siberian Circular Photon Source (SKIF)

The Siberian Circular Photon Source (SKIF, in Russian), a 4+ generation synchrotron radiation source, is under development near Novosibirsk, in the Koltsovo settlement. The SKIF is comprised of an complicated accelerator complex, including a 200 MeV linear electron accelerator, a full-energy synchrotron booster, and a storage ring. This facility will feature a 3 GeV relativistic electron storage with a 476 metres perimeter and an ultra-small calculated horizontal natural emittance of 73.2 pm·rad, enabling it to generate synchrotron radiation beams of maximum brightness in the 100 eV to 100 keV range at 30 experimental stations. Particularly, for photon energies around ∼1–5 keV, the emittance of the source reaches the wave diffraction limit, enhancing the spatial coherence of the Synchrotron Radiation (SR) and broadening the complex's research capabilities [51, 52].

It is known that SR, the electromagnetic radiation emitted by relativistic charged particles on curved trajectories, serves as a versatile tool for cutting-edge interdisciplinary research and technology applications in various critical economic sectors, contributing to technological sovereignty. SR sources, cyclic electron accumulators with several GeV of energy and orbital lengths from several hundred meters to kilometers, generate intense beams of charged particles with minimal phase volume emittance. These beams, moving in a transverse magnetic field, produce powerful and bright radiation, channeled to experimental stations for diverse research purposes.

Globally, dozens of SR sources are utilized for research in physics, chemistry, biology, medicine, geology, archaeology, materials science, and applied radiation applications, representing the most numerous class of ultra-relativistic energy electron beam accumulators. These facilities, operating in a collective access mode, provide infrastructure to various user organizations based on open competition outcomes. The key user characteristics of SR sources hinge prominently on the brightness and coherence of the emitted radiation. The brightness of an SR source refers to the intensity of the photon flux it produces, while coherence pertains to the uniformity and phase correlation of the radiation waves. The greater the brightness and the higher the coherent fraction of the photon flux, the superior is the "quality" of the facility. These attributes significantly enhance the facility's appeal to researchers from various external organizations. High brightness enables the study of faster processes and finer structures, while high coherence is crucial for techniques like coherent diffraction imaging and holography. Therefore, facilities that exhibit higher levels of brightness and coherence are more desirable for cutting-edge research, attracting a broader spectrum of scientific inquiries and explorations.

6. Goals and Objectives for Digital Twin of the SKIF

The DT for the SKIF aims to unify diverse mathematical and machine learning models, assimilate Big Data telemetry for identifying SKIF's structures and parameters, optimize product tuning using specific criteria, and integrate infrastructure components like software interfaces, digital test sites, and visualization tools.

The primary principles of SKIF's DT, as outlined [5, 18], include real-time reflection encompassing both physical and virtual spaces, with the virtual space accurately mirroring the physical space. Furthermore, it supports self-improvement by updating models parameters in real time, allowing continuous enhancement of mathematical models through parallel comparison with their physical counterparts. Figure 2 illustrates the interaction between the DT, SKIF, the automated control system (ACS) and the human operator.

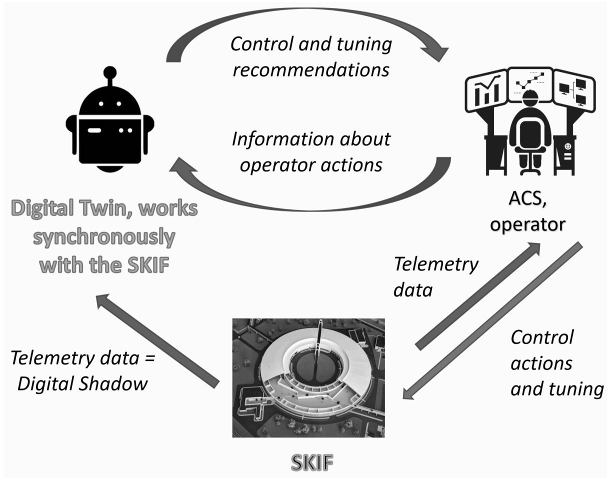


Figure 2. Scheme of interaction between the DT, SKIF, ACS and human operator

A dedicated computing environment for its DT is necessary, consisting of a computing platform for mathematical models and data integration, HPC servers and data storage. The computing platform capabilities are as follows: collecting and assimilating telemetry data, real-time control and tuning of the SKIF, making virtual experiments simulating critical operating modes, storing log/archive data archives, making experiment and data management, providing information security, making public demonstrations and users training.

The platform contains of a set of detailed mathematical models for precise virtual experiments and parametric analysis of phenomena. Purpose of the models: simulation of heat transfer, electromagnetism, radiation-sample-detector interactions, thermomechanical stresses and deformations, and stability and seismic resistance, etc. To make computations using detailed mathematical models, a HPC cluster is needed.

The set of models is central approach to conduct virtual experiments allowing for the simulation of various scenarios in a controlled virtual environment. This aspect is especially beneficial for tests that are impractical or risky to perform physically. Predictive analytics form a crucial part of these models, encompassing performance forecasting and anomaly prediction. These analytics aid in preemptively identifying potential operational challenges or deviations, facilitating timely maintenance and upgrades. This foresight is instrumental in scheduling equipment modernization and repairs, ensuring the facility's state-of-the-art status and optimal functionality.

The planning of experiments is also refined through these models. By simulating different parameters and conditions, they aid in optimizing experimental procedures, leading to more effective and efficient research outcomes. Additionally, the models play a pivotal role in tuning experimental data, a process vital for the calibration of instruments and validation of experimental results, ensuring their accuracy and reliability.

The platform includes Deep Machine Learning models and surrogate methods for fast computations. The surrogate solver, a blend of a Deep Machine Learning solver and a set of rapid mathematical models, will operate online in real physical time, using detailed mathematical models results as synthetic datasets. It has a property of self-teaching through telemetry data and synthetic dataset assimilation. This solver will work primarily for SKIF tuning and control.

Data assimilation is a critical process, integrating and analyzing information through detailed mathematical models that run offline on a supercomputer cluster. Figure 3 presents the data assimilation scheme in the SKIF's DT.

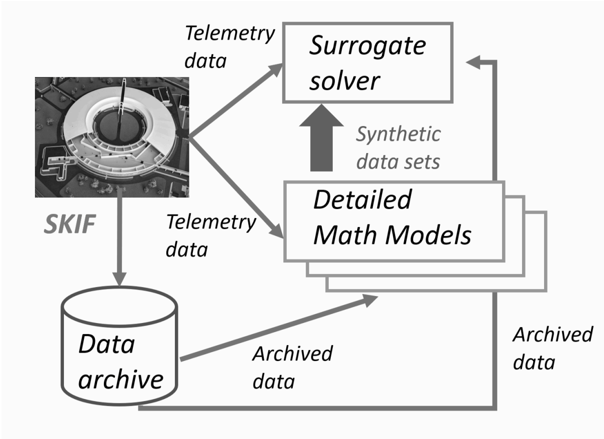


Figure 3. Scheme of data assimilation in SKIF's DT

The platform also comprises a 3D Building Information Model (BIM) of the SKIF, SKIF's documentation and research method descriptions.

The control scheme of the SKIF, as shown in Fig. 4, demonstrates how the DT technology integrates with the facility's operational management. This integration facilitates real-time monitoring and adjustment of the SKIF's parameters, ensuring its performance is maintained at an optimal level. The DT acts as an intellectual decision-support tool, enhancing the precision and efficiency of the facility's operations, and allowing for swift responses to operational changes or anomalies. This holistic approach to managing the SKIF underscores the DT's role in elevating the facility's operational capabilities.

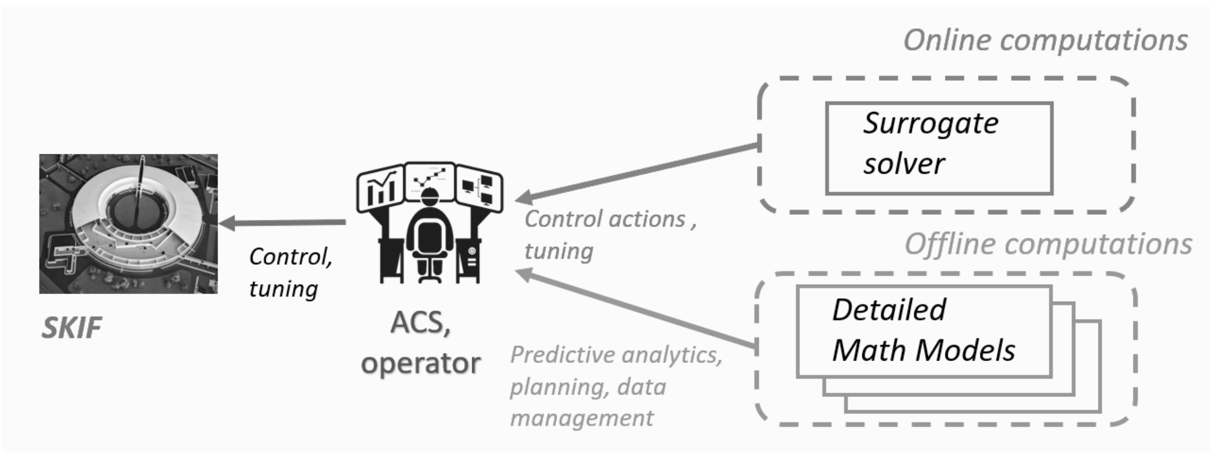


Figure 4. Scheme of controlling the SKIF by the DT

7. Analysis of Necessary Computing Resources and Data Storage Volume

Key computing tasks for SKIF encompass on-the-fly data processing through pipelines for sampling and optimization of data acquisition parameters, deciphering crystal structures from X-ray diffraction data, addressing the inverse problem of spectroscopy, and engaging in pattern recognition, classification, and image processing [51, 52]. The tasks also include applying deep learning neural network technologies, conducting mathematical modeling in quantum mechanics and engineering calculations, developing customized algorithms and software systems for the local community, and employing DT technologies for precise control and tuning.

The solution of the tasks assigned to the SKIF research team will be impossible without the utmost level of informatization and intellectualization of the processes of obtaining new scientific knowledge at this facility with the widespread use of AI technologies in data management chains in order to quickly obtain the final results of the study.

One of the key aspects that ensure high economic efficiency and scientific effectiveness of Large-Scale Scientific Infrastructure is a correctly selected and implemented data flow management policy based on modern information technologies. According to statistics, it is the Large-Scale Scientific Infrastructure that are among the most demanding users of HPC resources and structured storage of large amounts of information. Examples of such infrastructure are nuclear-physical facilities such as the Large Hadron Collider (LHC) or the International Thermonuclear Reactor (ITER), which is currently under construction. This should come as no surprise: Large-Scale Scientific Infrastructure generates Big Data.

The importance of this problem is becoming more and more apparent, which is reflected in the regular appearance of specialized articles, reviews, and entire thematic journal issues. For example, in a recently published analytical review [55], prepared by the IT managers of the four leading US synchrotron centers (APS, NSLS II, ALS and SLAC), forecasts the pace of development of the needs of the US national network of Large-Scale Scientific Infrastructure over the next few years. According to this forecast, by 2028, U.S. synchrotron sources will exceed the exabyte threshold (quintillion, or 10^{18} bytes) in terms of the amount of experimental data generated per year, and the requirements for peak performance of local computing systems will approach, respectively, exascale (quintillion floating-point operations per second).

According to the authors of another analytical review [56, 57], synchrotron crystallography and, in particular, macromolecular crystallography for biomedicine and pharmaceuticals, as well as computational tomography and other X-ray imaging techniques will be the key scientific areas shaping trends for ultra-high-performance IT infrastructure. An extremely resource-intensive experimental method is the recently developing resolving serial crystallography – microsecond for 4th generation synchrotron radiation sources and nanosecond for X-ray free electron lasers [58].

The IT infrastructure of the SKIF should include at least two layers. First layer combines automation and computing resources localized at experimental stations directly connected to information generators e.g., a high-speed two-dimensional detectors during a synchrotron experiments. The elements of the automation complex distributed among the experimental stations of the synchrotron radiation are entrusted with the role of experiment management, data collection in accordance with the experiment configuration, and transmission of primary "raw" experimental data to the DPC for subsequent processing and storage.

Second layer combines centralized computing facilities: resources hosted in the Data Processing Center (DPC) of the experimental stations.These resources are entrusted with the function of numerical mathematical processing of data obtained during the synchrotron experiment, placing these data, as well as the results of their processing in file storage, as well as providing access to them to the participants of the research collaboration.

The core switch of the SKIF's DPC should provide for the presence of 100–200 Gigabit ports to maintain high-speed communication with specialized or external computing resources, such as the existing Siberian Supercomputer Center of the Siberian Branch of the RAS at the Institute of Computational Mathematics and Mathematical Geophysics SB RAS or the promising Lavrentiev Supercomputer Center in Novosibirsk Akademgorodok.

The detailed status of the SKIF project at the current stage allows us to carry out initial assessments of the needs of experimental stations in data center computing resources and data storage volume:

- the rate of "hot" experimental data from one critical detector is 100–200 Gigabits per second;
- volumes of "hot" detector experimental data for short-term storage 200–240 Terabytes per day;
- the total bandwidth of the "hot" detector data storage system is 30–40 Gigabytes per second;
- the speed of experimental data from the "slow" detector equipment of synchrotron radiation stations is 10–20 Gigabits per second;
- the total peak rate of "slow" data from all six experimental stations is 200–250 Gigabits per second;
- the volume of "slow" experimental data for all six stations is 20–30 Terabytes per day;
- the volume of long-term (year) "warm" storage of experimental data is 4–8 Petabytes per year;
- universal multicore CPUs 275–350 Teraflops with double precision FP64;
- GPU accelerators 750–850 Teraflops with double precision FP64.

Besides, all experimental data, as well as telemetry data, must be stored eternally in accordance with the FAIR principle [54].

In the above values, a portion of the computing resources is allocated to operate the SKIF's DT. According to preliminary estimates of the SKIF's information complexity and based on the experience of developing DTs at Peter the Great St. Petersburg Polytechnic University by the team of A.I. Borovkov (see, for example, [59]), to carry out offline computations using a DT requires about 100 Teraflops of computing power on universal multi-core CPUs. To conduct online simulations primarily using neural network models, the digital twin requires about 10–30 Teraflops on GPU accelerators. At the same time, to implement a continuous mode of additional training of neural network models, about 100 Teraflops are required on GPU accelerators.

Conclusion

The review showed a slight but growing interest in the concept of DTs in the field of large-scale scientific infrastructure. There is a reason to believe that, at least in a number of cases, digital models with different bases are created in these tasks, connected by data flows with physical objects, which is quite consistent with the concept of a DT, but the appropriate terminology is not used. There is a growing interest in using machine learning technologies to create such digital models in these areas. For example, a query like TITLE-ABS-KEY ("particle accelerator*" AND ("neural network*" OR "machine learning")) yields 415 results in Scopus, most of which have been in recent years. It can be assumed that as the concept of DTs is accepted in the scientific and engineering communities associated with big science and related installations, the number of studies and cases of real application of DT technology will grow rapidly, which has already been observed in other applied areas.

Analysis of the literature shows that potential areas of application of DT in large projects of scientific infrastructure include:

- optimization of general management and reporting, increasing the efficiency of external communications with management and financing organizations, society through the creation of realistic three-dimensional models of scientific facilities with the accumulation of data on the state of the facilities themselves and the environment, experiments and main results;
- monitoring, maintenance, anomaly detection, prediction of equipment failures based on the collection and processing of data from sensor networks, comparison with the results of simulations using physical models or neural networks trained on real data;
- management of the life cycle of complex scientific equipment through the creation and constant updating of its DT, which provides consistent storage and processing of log/archive and current information about its components and their condition, routine and urgent work on configuration, repair, and updating of equipment;
- development, verification, fine-tuning of equipment and/or software for event detection based on DTs of experimental spaces or observed objects and phenomena;
- creation and support of automatic control systems that provide optimization of operating modes, reduction of time for setting up and changing configurations, reduction of delays in the generation of control signals.

Conceptualizing these applications will accelerate and optimize the adoption of DT technologies, and provide attention to the potential benefits and possible drawbacks of the technology, and the risks associated with its application. An essential issue related to this conceptualization is the correct definition of the object, the virtual embodiment of which is the DT. We have already seen examples of DTs of the systems and objects under study, which is markedly different from the classical interpretations of this technology. Probably, in the case of using a data center for automated control of complex objects, we can say that a virtual copy of the “operator control and control system control object control object” system is being formed, which also requires a significant rethinking of classical approaches.

Thus, DT technologies have significant potential in the design, construction and operation of large scientific infrastructure projects, providing the ability to virtually test various concepts and configurations, accelerate and improve the processes of setting up and optimizing equipment operation, monitoring the performance of critical components, predicting faults, which will ultimately lead to more scientific results at the same or lower cost.

As a perspective, the experience and results of making the DT of the SKIF may be used in the development of a universal platform for digital twins of other Large-Scale Research Infrastructure.

Acknowledgements

Authors thank Yan Zubavichus, Vladimir Poteryaev, Stanislav Shakirov, Igor Marinin, Maxim Gorodnichev, Yuriy Medvedev and Ruslan Permyakov for fruitful discussion.

The work was supported by the state project of ICMMG SB RAS 0251-2022-0005.

References

1. Grieves, M.: Digital Twin: Manufacturing Excellence Through Virtual Factory Replication. White Paper, pp. 1–7 (2014).

2. Grieves, M., Vickers, J.: Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems. In: Transdisciplinary Perspectives on Complex Systems: New Findings and Approaches. pp. 85–113. Springer (2016). https://doi.org/10.1007/978-3-319-38756-7_4

3. Tuegel, E.J., *et al.*: Reengineering Aircraft Structural Life Prediction Using a Digital Twin. Int. J. Aerospace Eng. Article 154798. (2011). https://doi.org/10.1155/2011/154798

4. Glaessgen, E.H., Stargel, D.S.: The Digital Twin Paradigm for Future NASA and U.S. Air Force Vehicles (2012).

5. Tao, F., *et al.*: Digital Twin in Industry: State-of-the-Art. IEEE Trans. Ind. Inform. 15(4), 2405–2415 (2019). https://doi.org/10.1109/TII.2018.2873186

6. O'Connell, C.: CAD/CAM (Computer-Aided Design/Computer-Aided Manufacturing). Sci. & Tech. Libraries, Routledge 7(4), 127–154 (1987). https://doi.org/10.1300/J122v07n04_13

7. Borrelli, A., Wellmann, J.: Computer Simulations Then and Now: An Introduction and Historical Reassessment. N.T.M. 27(4), 407–417 (2019). https://doi.org/10.1007/s00048-019-00227-6

8. Boschert, S., Rosen, R.: Digital Twin-The Simulation Aspect. In: Mechatronic Futures: Challenges and Solutions for Mechatronic Systems and Their Designers. pp. 59–74. Springer (2016). https://doi.org/10.1007/978-3-319-32156-1_5

9. Fuller, A., *et al.*: Digital Twin: Enabling Technologies, Challenges and Open Research. IEEE Access 8, 108952–108971 (2020). https://doi.org/10.1109/ACCESS.2020.2998358

10. Atzori, L., Iera, A., Morabito, G.: The Internet of Things: A Survey. Comput. Netw. 54(15), 2787–2805 (2010). https://doi.org/10.1016/j.comnet.2010.05.010

11. Sisinni, E., *et al.*: Industrial Internet of Things: Challenges, Opportunities, and Directions. IEEE Trans. Ind. Inform. 14(11), 4724–4734 (2018). https://doi.org/10.1109/TII.2018.2852491

12. Lee, J., Bagheri, B., Kao, H.-A.: A Cyber-Physical Systems Architecture for Industry 4.0-based Manufacturing Systems. Manuf. Lett. 3, 18–23 (2015). https://doi.org/10.1016/j.mfglet.2014.12.001

13. Philip Chen, C.L., Zhang, C.-Y.: Data-intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data. Inf. Sci. 275, 314–347 (2014). https://doi.org/10.1016/j.ins.2014.01.015

14. Lermer, M., Reich, C.: Creation of Digital Twins by Combining Fuzzy Rules with Artificial Neural Networks. In: IECON 2019 - 45th Annual Conference of the IEEE Industrial Electronics Society, Lisbon, Portugal, October 14-17, 2019. pp. 5849–5854. IEEE (2019). https://doi.org/10.1109/IECON.2019.8926914

15. Tarkhov, D.A., Malykhina, G.F.: Neural Network Modelling Methods for Creating Digital Twins of Real Objects. Journal of Physics: Conference Series 1236(1), 012056 (2019). https://doi.org/10.1088/1742-6596/1236/1/012056

16. Kaur, M.J., Mishra, V.P., Maheshwari, P.: The Convergence of Digital Twin, IoT, and Machine Learning: Transforming Data into Action. In: Farsi, M., Daneshkhah, A., Hosseinian-Far, A., Jahankhani, H. (eds) Digital Twin Technologies and Smart Cities. Internet of Things. pp. 3–17. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-18732-3_1

17. Rathore, M.M., *et al.*: The Role of AI, Machine Learning, and Big Data in Digital Twinning: A Systematic Literature Review, Challenges, and Opportunities. IEEE Access 9, 32030–32052 (2021). https://doi.org/10.1109/ACCESS.2021.3060863

18. Tao, F., *et al.*: Digital Twin-driven Product Design, Manufacturing and Service with Big Data. Int. J. Adv. Manuf. Tech. 94(9-12), 3563–3576 (2018). https://doi.org/10.1007/s00170-017-0233-1

19. Kritzinger, W., *et al.*: Digital Twin in Manufacturing: A Categorical Literature Review and Classification. IFAC-PapersOnLine 51(11), 1016–1022 (2018). https://doi.org/10.1016/j.ifacol.2018.08.474

20. Errandonea, I., Beltran, S., Arrizabalaga, S.: Digital Twin for Maintenance: A Literature Review. Comput. Ind. 123, 103316 (2020). https://doi.org/10.1016/j.compind.2020.103316

21. Mohammadi, N., Taylor, J.E.: Smart City Digital Twins. In: 2017 IEEE Symposium Series on Computational Intelligence, SSCI 2017, Honolulu, HI, USA, November 27 – Dec. 1, 2017. pp. 1–5. IEEE (2018). https://doi.org/10.1109/SSCI.2017.8285439

22. Deng, T., Zhang, K., Shen, Z.-J.M.: A Systematic Review of a Digital Twin City: A New Pattern of Urban Governance Toward Smart Cities. J. Manag. Sci. Eng. 6(2), 125–134 (2021). https://doi.org/10.1016/j.jmse.2021.03.003

23. Deng, M., Menassa, C.C., Kamat, V.R.: From BIM to Digital Twins: A Systematic Review of the Evolution of Intelligent Building Representations in the AEC-FM Industry. J. Inf. Technol. Constr. 26, 58–83 (2021). https://doi.org/10.36680/J.ITCON.2021.005

24. Lu, Q., *et al.*: From BIM Towards Digital Twin: Strategy and Future Development for Smart Asset Management. Stud. Comput. Intell. 853, 392–404 (2020). https://doi.org/10.1007/978-3-030-27477-1_30

25. Sepasgozar, S.M.E., *et al.*: Lean Practices Using Building Information Modeling (BIM) and Digital Twinning for Sustainable Construction. Sustainability (Switzerland) 13(1), 1–22 (2021). https://doi.org/10.3390/su13010161

26. Regis, A., *et al.*: Physic-based vs Data-based Digital Twins for Bush Bearing Wear Diagnostic. Wear 526-527, 204888 (2023). https://doi.org/10.1016/j.wear.2023.204888

27. Sarkar, P.: Digital Twin Modeling Using Machine Learning and Constrained Optimization. Medium (2022). https://towardsdatascience.com/digital-twin-modeling-using-machine-learning-and-constrained-optimization-401187f2a382, accessed: 2023-04-10

28. Jacob, M., Hallonsten, O.: The Persistence of Big Science and Megascience in Research and Innovation Policy. Sci. Public Policy 39(4), 411–415 (2012). https://doi.org/10.1093/scipol/scs056

29. Borner, K., Silva, F.N., Milojevic, S.: Visualizing Big Science Projects. Nat. Rev. Phys. 3(11), 753–761 (2021). https://doi.org/10.1038/s42254-021-00374-7

30. Cramer, K.C., Hallonsten, O.: Big Science and Research Infrastructures in Europe. 2020, pp. 1–264. Edward Elgar Publishing, Cheltenham, UK (2020). https://doi.org/10.4337/9781839100017

31. D'Ippolito, B., Ruling, C.-C.: Research Collaboration in Large Scale Research Infrastructures: Collaboration Types and Policy Implications. Res. Policy 48(5), 1282–1296 (2019). https://doi.org/10.1016/j.respol.2019.01.011

32. Johnston, S., *et al.*: Science with ASKAP: The Australian Square-Kilometre-Array Pathfinder. Exp. Astron. 22(3), 151–273 (2008). https://doi.org/10.1007/s10686-008-9124-7

33. Bednarz, T., *et al.*: Digital Twin of the Australian Square Kilometre Array (ASKAP). In: SIGGRAPH Asia 2020 Posters. Article 15. ACM (2020). https://doi.org/10.1145/3415264.3425462

34. Jonas, J.L.: MeerKAT - The South African Array with Composite Dishes and Wide-Band Single Pixel Feeds. Proc. IEEE 97(8), 1522–1530 (2009). https://doi.org/10.1109/JPROC.2009.2020713

35. Taljaard, C., Chrysostomou, A., Van Zyl, A.N.: Sculpting a Maintenance Twin for SKA. In: Modeling, Systems Engineering, and Project Management for Astronomy IX, vol. 11450, pp. 114500F. SPIE (2020). https://doi.org/10.1117/12.2562337

36. Nan, R., *et al.*: The Five-Hundred-Meter Aperture Spherical Radio Telescope (FAST) Project. Int. J. Mod. Phys. D 20(6), 989–1024 (2011). https://doi.org/10.1142/S0218271811019335

37. Li, Q.-W., *et al.*: Prognostics and Health Management of FAST Cable-Net Structure Based on Digital Twin Technology. Res. Astron. Astrophys. 20(5) (2020). https://doi.org/10.1088/1674-4527/20/5/67

38. Wen, J., *et al.*: Rapid Modeling Method for The Digital Twin of Five-Hundred-Meter Aperture Spherical Radio Telescope. IAENG Int. J. Comput. Sci. 49(2) (2022).

39. Zhang, Q., Wu, P., Zhao, Z.: Design and Application of Digital Twin System Architecture for Large Radio Telescope. Jisuanji Jicheng Zhizao Xitong/Comput. Integr. Manuf. Syst. CIMS 27(2), 364–373 (2021). https://doi.org/10.13196/j.cims.2021.02.005

40. Predehl, P., *et al.*: The eROSITA X-ray Telescope on SRG. Astron. Astrophys. 647 (2021). https://doi.org/10.1051/0004-6361/202039313

41. Seppi, R., *et al.*: Detecting Clusters of Galaxies and Active Galactic Nuclei in an eROSITA All-Sky Survey Digital Twin. Astron. Astrophys. 665 (2022). https://doi.org/10.1051/0004-6361/202243824

42. Jones, B.J.P.: The Status of the MicroBooNE Experiment. J. Phys.: Conf. Ser. 408 (2013). https://doi.org/10.1088/1742-6596/408/1/012028

43. Kafkes, D., Schram, M.: Developing Robust Digital Twins and Reinforcement Learning for Accelerator Control Systems at the Fermilab Booster. In: Proc. of the 12th International Particle Accelerator Conference. pp. 2268–2271. JACoW Publishing, Geneva, Switzerland (2021). https://doi.org/10.18429/JACoW-IPAC2021-TUPAB327

44. Abramowicz, H., *et al.*: Higgs Physics at the CLIC Electron-Positron Linear Collider. Eur. Phys. J. C 77(7), article 475 (2017). https://doi.org/10.1140/epjc/s10052-017-4968-5

45. Doytchinov, I., *et al.*: Thermal Effects Compensation and Associated Uncertainty for Large Magnet Assembly Precision Alignment. Precis. Eng. 59, 134–149 (2019). https://doi.org/10.1016/j.precisioneng.2019.06.005

46. Neumann, A., *et al.*: bERLinPro Becomes SEALab: Status and Perspective of the Energy Recovery Linac at HZB. In: Proc. 13th International Particle Accelerator Conference. pp. 1110–1113. JACoW Publishing, Geneva, Switzerland (2022). https://doi.org/10.18429/JACoW-IPAC2022-TUPOPT048

47. Mahshid, M.Z., *et al.*: Realization of an Energy System-Informed Digital Twin of the KARA Accelerator at KIT in a Real-Time Simulation Environment: the ACCESS Project. In: 14th International Particle Accelerator Conference (2023).

48. Giles, A., *et al.*: Operation the Accelerator Test Facility Linac Transport Beamline by Using Artificial Intelligence and Machine Learning Methods. In: 14th International Particle Accelerator Conference (2023).

49. Ansel, A., *et al.*: A New Product Lifecycle Management Platform for CERN's Accelerator Complex and Beyond. In: 14th International Particle Accelerator Conference (2023).

50. Edelen, A.: AI/ML and Its Operational Challenges at SLAC's Accelerators and Collaborating Facilities (2021).

51. Schaefer, K.I. (ed.): Technological Infrastructure of the SKIF Center for Common Use. Novosibirsk (2022). https://disk.yandex.ru/d/1SBhHph2rgbeBg, accessed: 2023-04-10

52. Scientific Program of the SKIF Center for Common Use: Key Areas of Research at the First Stage Experimental Stations and the Concept of Infrastructure Development until 2035. Novosibirsk, 439 p. (2023). https://disk.yandex.ru/d/gxEIdsjIalIvHw, accessed: 2023-04-10

53. GOST R 57700.37-2021: Computer Models and Modeling. Digital Twins of Products. General Provisions.

54. Wilkinson, M., Dumontier, M., Aalbersberg, I., *et al.*: The FAIR Guiding Principles for Scientific Data Management and Stewardship. Sci. Data 3, 160018 (2016). `https://doi.org/10.1038/sdata.2016.18`

55. Schwarz, N., Campbell, S., Hexemer, A., *et al.*: Enabling Scientific Discovery at Next-Generation Light Sources with Advanced AI and HPC. In: Driving Scientific and Engineering Discoveries Through the Convergence of HPC, Big Data and AI, SMC 2020, Comm. Comp. Inf. Sci., vol. 1315, pp. 145–156. Springer, Cham (2020). `https://doi.org/10.1007/978-3-030-63393-6_10`

56. Wang, C., Steiner, U., Sepe, A.: Synchrotron Big Data Science. Small 14(46), 1802291 (2018). `https://doi.org/10.1002/smll.201802291`

57. Wagner, H.: Data Handling and Storage. Synchrotron Radiation News 32(3), 2–3 (2019). `https://doi.org/10.1080/08940886.2019.1618682`

58. Ponsard, R., Janvier, N., Kieffer, J., *et al.*: RDMA Data Transfer and GPU Acceleration Methods for High-Throughput Online Processing of Serial Crystallography Images. J. Synchrotron Radiat. 27(5), 1297–1306 (2020). `https://doi.org/10.1107/s1600577520008140`

59. Prokhorov, A., Lysachev, M., Borovkov, A. (eds.): Digital Twin. Analysis, Trends, World Experience. First Edition. Moscow, Alliance Print, 401 p. (2020).