International Journal of Information Technology Vol. 28 No. 1 2022

DCverse – A Cloud Metaverse System for Carbon-neutral Data Center

Xin Zhou[a], Ruihang Wang[a] and Yonggang Wen[a]

[a]Nanyang Technological University, Singapore
{yyran, zhouxin, ygwen}@ntu.edu.sg

Abstract

Data centers have been challenged to achieve carbon neutrality as computing demands continue to increase in energy consumption. Meanwhile, a data center relies on the domain expertise of human managers, lacks automation capabilities, and is difficult to manage efficiently. All these factors have resulted in problems for the sustainable development of data centers. To address the challenges of sustainability and cost in data centers, we explored the industrial metaverse solution, and developed a Cloud Metaverse System, DCverse, which integrates digital twin and AI technology. We designed a triple-intelligent DCverse system and proposed dual-cycle control loop for optimisation strategy. With the help of this solution, the data center can not only maintain the optimal control configurations to maintain the constant temperature of the data center, but also improve the power usage effectiveness.

Keywords: Carbon Neutral, Cloud Metaverse System, Management Control for Data Centers, Energy Conservation.

I. Introduction

The data centre, as a critical infrastructure, is crucial to the global digital economy. While the global economy was in a deep slump as a result of the pandemic, global business demand for

intensive data centre services grew explosively, and internet traffic increased by nearly 40% during the quarantine period of the epidemic[3].

Data centre electricity consumption has sky-rocketed in recent years, driven by rising demand for this mission-critical ICT infrastructure. Singapore's data centre industry accounts for 5.3% of total annual electricity consumption and increased societal digitization as a data hub in Southeast Asia[10]. High energy consumption results in a high carbon footprint. For example, carbon emissions attributed to the data centre industry are approximately 2% and will continue to rise over the next decade[1], and the data centre industry will account for 8% of global carbon emissions by 2030[4]. In order to reduce the carbon footprint, the Singapore government banned the construction of new data centres in 2019[2]. Therefore, optimizing energy efficiency and decarbonizing data centres are now pressing demands.

Today, data centre management is still heavily reliant on the domain knowledge of human experts and is largely a best practice-based manual process with drawbacks of a risk-averse mindset and high staff turnover. This burdens the industry due to poor energy efficiency, a high carbon footprint, and high labour costs. Consequently, an automated, safe, and cost-effective solution for dramatic efficiency improvement and decarbonization has become highly sought after.

To address the challenges of sustainability and cost in data centers, we aim to develop and deploy an industrial metaverse solution, DCverse, over a cloud environment, for energy efficiency and decarbonization in Alicloud data centers. This solution leverages NTU's award-winning digital twin (DT) and AI technology to automatically govern and calibrate Alicloud's data centre cooling system and improve the PUE (Power Usage Effectiveness), with the ultimate objective of achieving carbon-neutral operations. This solution generates optimal configuration and control commands to address the temperature demand in data center. The solution visualizes the performance of the

physical data center in it's DT, providing the operator with a direct sense of temperature distribution in the data hall, and energy flow in chiller plant systems. Advanced control policies and novel design concepts can be tested and validated in the industrial metaverse environment. Furthermore, the industrial metaverse recommends the optimal control policy for increased energy efficiency and a lower carbon footprint.

The rest of this paper is organized as follows. In Section 2, we briefly present some related work regarding industrial metaverse and energy utilization strategies. Section 3 examines the problem statement, while in section 4 we design a triple-intelligent DCverse system and propose a dual-cycle control loop for the optimisation strategy as our solution. In Section 5, we introduce four work packages for implementing DCverse functionalities and finally we conclude this paper in section 6.

II. Related Works

A. Industrial Metaverse

The industrial metaverse[7] uses virtual and augmented reality to blend the physical and digital worlds to transform how businesses design, manufacture and interact with objects, and it is recognized as the new fundamental infrastructure for smart society, which demonstrates the rebuilding and re-creation capability of the physical world, in the cyber world. The industrial metaverse technology is composed of three phases: 1) digital twin, 2) digital thread, and 3) born-digital, as shown in Figure 1.

The digital twin is a virtual replica of a real-world entity. It starts as a multi-scale and multi-physics simulation tool, can be extended to include neural network approximation[12], and is thus valuable for the lifecycle management of complex industrial systems like a data centre. The digital

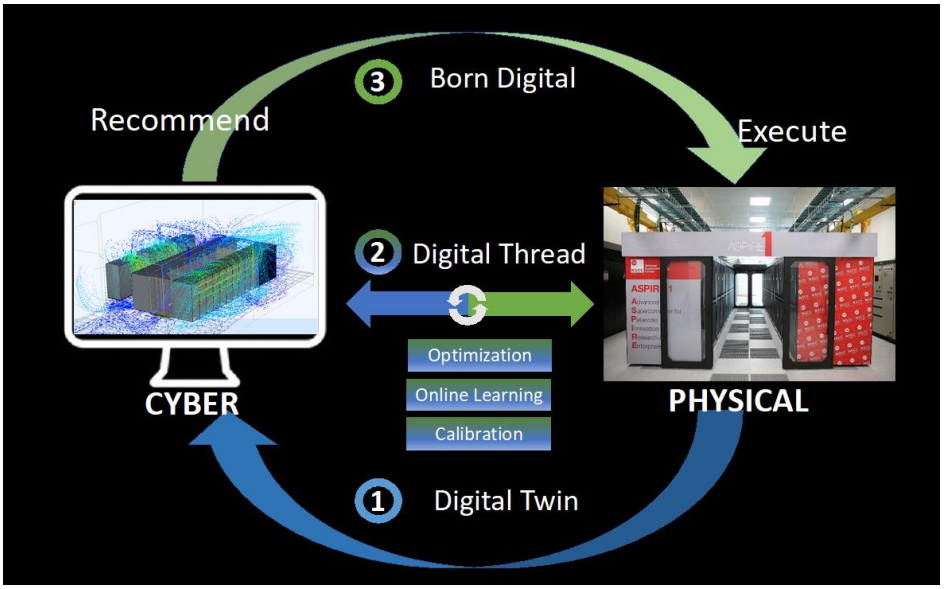

Fig 1 Three progressive phases for Industrial metaverse structure

thread is the deep fusion of the industrial scenario with AI technology, which enables a two-way connection between the physical world and the cyber world.[9] The physical world problems are mapped to the cyber world, where optimization is conducted and tested with a low cost and high efficiency. In return, the optimal solution generated by AI in the cyber world is feedback which is used to solve the physical world problems. Born digital is the ultimate form of the industrial metaverse, with the highest intelligence. It is the "Oasis" for various advanced AI technologies to be implemented and functioned. In this phase, new patterns, new solutions, and new products are created in the cyber world, even without any reference in the physical world, and achieve a syncretic implementation in the physical world.

B. Energy Utilization for Data Centre

In order to improve energy utilization efficiency, existing AI-empowered energy utilization research usually aims to optimize either the IT system or the cooling system and starts from three perspectives: 1) Enabling Techniques, 2)Energy-Efficient Computing, and 3)Energy-Efficient Cool-

ing.

Some companies and researchers start by investigating efficiency technologies from the enabling techniques of the data centers. In the cooling management of data centers, traditional air-cooling systems with a raised floor layout, due to not reaching the best cooling efficiency, have been replaced by these new layouts which are based on ideas such as preventing dissipated hot air from intermingling with cold supply air via hot aisle containment[11] and in-rack air cooling[5]. In terms of energy-efficient computing, the authors in [8] proposed a hierarchical framework for resource allocation and server power management based on a Deep Q network with LSTM workload predictor, which realized a 53.97% energy saving. Similarly, another work[14] also formalized the problem as a Markov Decision Process, and devised a Deep Q network with LSTM state predictor. The experimental results showed significant cooling energy saving. As for the energy-efficient cooling technologies, in [6], a data-driven, model-based MPC controller with random walk exploration was proposed for dynamic control of the Computer Room Air Conditioning(CRAC) blower rotational speed and the valve opening for chilled water. In addition, the authors proposed a DQN controller with reward sharing and fingerprint, which realized smaller rack inlet temperature variance in [13].

III. Problem Statement

In a data center, the data hall temperature is coupled to the chilled water through the Computer Room Air Handler (CRAH), shown in Figure 2. Given the air supply temperature, air flow rate, and IT load (i.e., server power consumption), the chiller water flow rate is regulated to ensure the water matches the heat dissipation requirements. In this process, it is critical to regulate the chilled water flow rate whilst keeping the temperature of the water flowing through the chiller

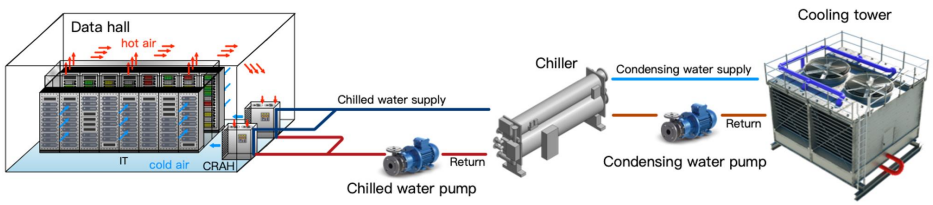

Fig 2 A reference data center architecture with data hall + chiller plant

within a specific range. On the chiller plant side, there is a trade-off between the water flow rate and temperature, which affects the power consumption of the pumps and chillers. In the data hall, the temperature set-points of the air side (e.g., air supply temperature) will also affect the holistic power performance of the data center and result in a trade-off between the power consumption of the CRAH and the chiller plant.

In this paper, the work addresses the above challenges for optimal operation in the Alicloud data center. The specific problem includes two aspects: 1) Finding the optimal control configurations to maintain the constant temperature of the data center's hall; 2) Optimizing the PUE performance.

DT and AI-based technologies are adopted to optimize the control policy, which guarantees the volume of cooling load for heat removal and improves the PUE of the entire data centre, which results in significant reductions in the carbon footprint.

IV. Framework and Methodology

To tackle the optimal operation challenges in the Alicloud data center, we propose integrating AI technologies with DTs to develop a unified platform as an instance of the industrial metaverse, to help the operator improve the manageability of the data centers. The designed industrial metaverse for the data center, DCverse, provides system visualization, simulation, validation, and optimization capabilities.

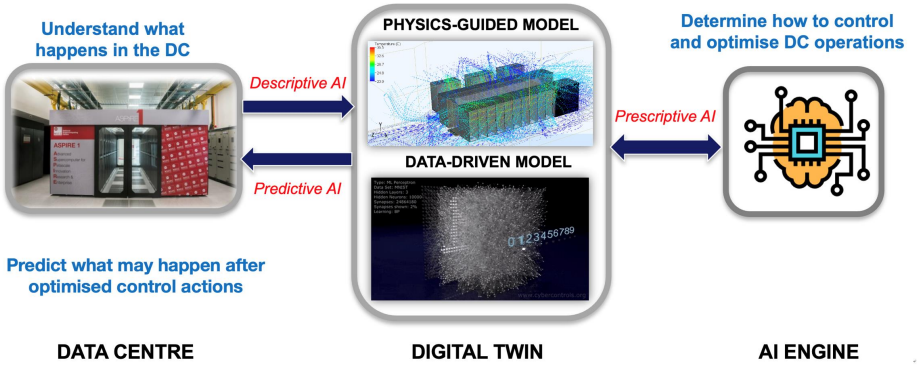

Fig 3 High-level architecture for AI and DT

A. Triple-intelligent DCverse System

Most of the current solutions are in the digital twin phase. Our proposed DCverse system is functional in the digital thread phase. Figure 3 shows the high-level architecture for AI and DT integration in DCverse. It offers 3 tiers of intelligence. The first tier is descriptive intelligence, where the internal behaviors of the system can be accurately modelled through the collection and analysis of historical and online data. The second tier is predictive intelligence, where system behaviors under hypothetical inputs can be predicted to anticipate anomalies/failures in the data centre. The third tier is prescriptive intelligence, where actions to improve system management and efficiency can be proposed and subsequently verified and validated on the cyber-system before implementing the optimized control policies.

B. Dual-cycle Control Loop for Optimisation

Another novelty of our proposed work is the use of the dual-cycle control loop for optimisation (shown in Figure 4). The workflow of this control loop is described as follows:

1) Digital twin construction: Raw data (e.g., room layout, device specs, setpoints, etc.) collected from the physical data center is transformed (via Cloud3DView) to construct the physical

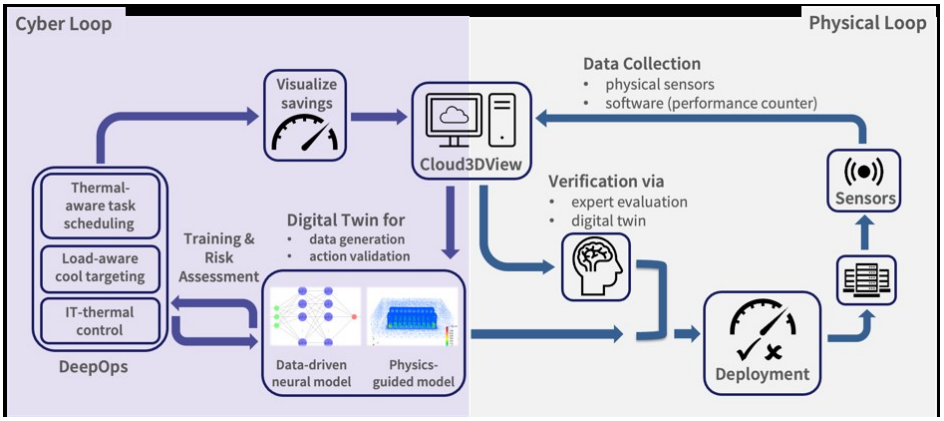

Fig 4 Dual-cycle control loop

rule-based and data-driven digital twins. To ensure the accuracy of digital twins, auto-calibration is performed to ensure the digital twins constructed are of industry-grade accuracy.

2) DeepOps training: The well-calibrated digital twins can synthesize massive amounts of training data, including emergencies and anomalies, in a relatively short time. The DRL-based agent can directly interact with the digital twins to learn system behaviours. The trained agent can then make optimal decisions to implement operations such as PUE optimisation etc. With this approach, the well-trained intelligent agent will be able to exhibit robustness and stability under different circumstances.

3) Verification: The control policies/actions derived by DeepOps are validated by the digital twins in advance to understand the risk/return of adoption. If needed, it is also possible for experienced human experts to be involved in this loop to ensure the safe deployment of AI-based optimisation.

4) Deployment: Based on the estimated results of digital twins and/or human experts in the loop, the on-site operator will decide whether to adopt the recommended actions. To avoid the uncertainty of AI-based optimisation, a fallback mechanism is available (e.g. over-provisioning) in the deployment phase to prevent fatal system errors.

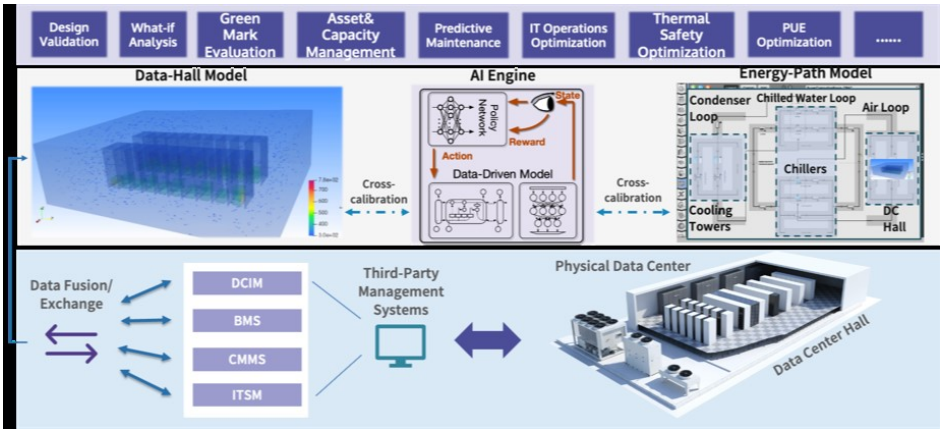

Fig 5 Cognitive Digital Twin Platform structure

The cognitive digital twin platform structure is illustrated in Figure 5. The integrated solution will be developed and perform co-simulation for both the data hall (e.g., thermodynamics) and chiller plant (e.g., power status), where there is currently no proven, usable, credible co-simulation software package available. This integrated platform can be used to statistically validate the control handlers of the data hall and chiller plant for stable and safe operations. In the solution, the AI engine interacts with the digital twin which provides huge, high-quality and high-diversity operation data for the AI to train and achieve an optimal solution. Much of the unbounded data is essential for the performance of AI, but usually cannot be provided by the physical data center. The AI engine derives the optimized safety-aware control policy by dynamically recommending proper control actions to satisfy the constraints of the control handlers (e.g., airside and/or water side) while improving the PUE. The training process does not involve tedious hyper-parameter tuning, and its robustness makes it a strong candidate for deriving optimal control policies for complex data center cooling facilities.

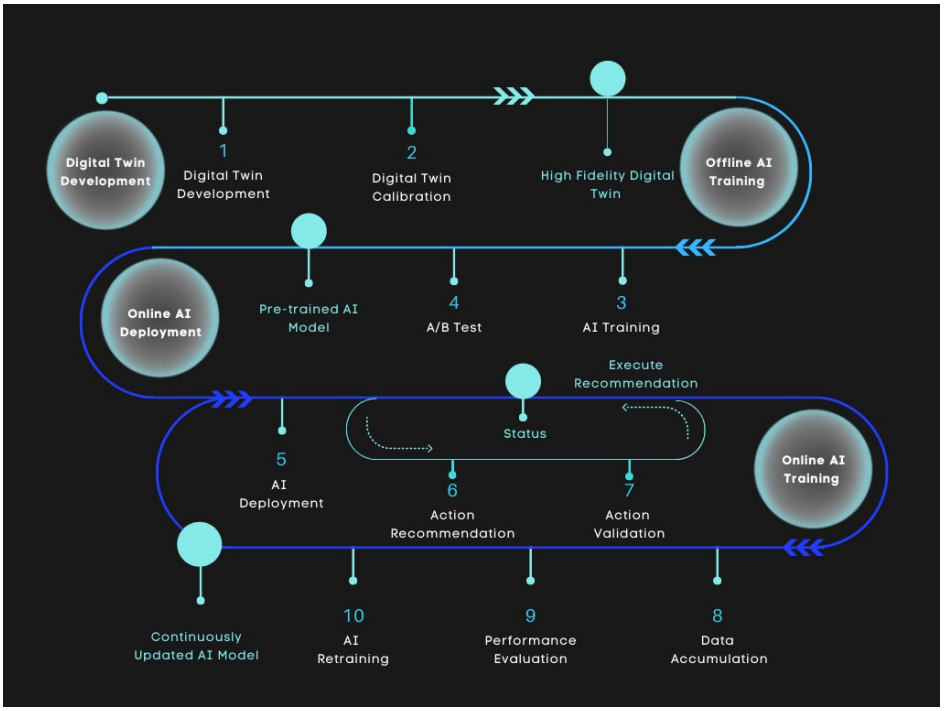

Fig 6 AI journey of DC Digital Transformation toward Sustainability

V. Work Package (WP)

The DCverse solution aims to develop our AI and digital twin technologies for data center, and emerging transformative technological trends (e.g., Robotic Processing Automation – RPA, No Code Low Code – NCLC, etc), into an innovation platform. The journey of AI in data center sustainability optimization is illustrated in Figure 6.

To realize DCverse functionalities in detail, we decompose this model into the following 4 work packages (WPs):

1) Data Collection

The data drives the proposed DT and AI-based industrial metaverse platform. The required data can be categorized into two classes: the design data (e.g., geometry, layout, dimensionality, specification, configuration, etc.) and operational data (e.g., air/water temperature, power, air/water flow rate, etc.). The former data set is used to construct the fundamental DTs for the data centre,

including the data hall and chiller plant. The second dataset calibrates the DTs and provides real-time system dynamics to the AI agent for control optimization. Thus, this WP aims to develop the data engine, including the gateway and storage modules. The gateway module can seamlessly plug-in third-party monitoring systems (e.g., DCIM, PMS, BMS, etc.) to retrieve operational data, and the storage module stores the data after pre-processing.

2) Digital Twin Development

This WP aims to develop an integrated DT that mimics dynamic transitions of the entire data centre, conducting what-if analysis for validating various water flow rate settings and temperature settings, and guaranteeing the operations within a safe range. We combine the CFD engine of the data hall (thermodynamic analysis) and the energy engine of the chiller plant (energy performance analysis), forming a holistic DT that can concurrently predict thermal and energy status under certain boundary conditions. The digital twin will be calibrated based on the operational data to achieve industry-grade prediction accuracy. Subsequently, the integrated DT will interact with the AI engine (AI-based control algorithms), providing large amounts of diverse training data for PUE optimization, without bringing any risk to the physical system.

3) Offline AI Training

Due to the increased data centre scale, the high-dimensional state space adversely challenges the decision-making of human experts, resulting in human error, safety violations, wastage of cooling capacity, and a high carbon footprint. AI technologies offer industry another way of achieving automatic control and recommendation. The reinforcement learning (RL) approach learns system behaviors by interacting with the target system in a trial-and-error manner, which may cause violations in the training and operational phases. In this WP, the team aims to develop a safety-oriented AI engine with RL algorithms. The RL algorithm jointly controls the data hall and chiller plant, to

save power consumption and improve the PUE while guaranteeing constraints such as water flow rate and temperature within a safe range, resulting in significant reductions in carbon emissions.

4) Online AI Deployment and Online AI Training

The trained AI model will be integrated with the DCWiz platform by Red Dot Anlaytics Pte Ltd and then deployed to the Alicloud data center to provide recommendations and achieve the desired demand. The changes in system dynamic distribution (e.g., server installation, weather conditions, aging of equipment, etc.) can make AI policies degrade or even fail over time. Therefore, after the AI policy is deployed, this WP will develop a continuous learning method to adapt the policy to the changing system environment. Specifically, the team will first design a metric to monitor the performance of the AI policy during the operational phase. If the states deviate from the control target, it will trigger a policy improvement. Also, this WP aims to develop algorithms that can quickly improve the AI policy with fewer sample shots. The continuous learning method will ensure the performance of the deployed AI policy throughout the data center's lifecycle.

VI. Conclusion

As energy usage for computers continues to rise, data centers are finding it difficult to become carbon neutral. A data center, on the other hand, is challenging to run efficiently and depends on the subject-matter expertise of human managers. The sustainable development of data centers has been hampered by both of these factors. We have investigated the industrial metaverse solution and created the Cloud Metaverse System, DCverse, which combines digital twin and AI technology, to address the problems of sustainability and affordability in data centers. With the aid of DCverse, the data center is able to increase the efficiency of its power usage in addition to maintaining the best control configurations to maintain the data center's temperature at a constant level.

References

[1] *Ushering in a net-zero emissions Akamai Edge by 2030.* Akamai Technologies, 2021.

[2] *Building Green Data Centres- Singapore Lifts Moratorium on New Data Centres, Introduces Environmental Sustainability Standards*. Lexology, Singapore, 2022.

[3] Global internet phenomena report. Technical report, Sandvine, 2022.

[4] Zhiwei Cao, Xin Zhou, Han Hu, Zhi Wang, and Yonggang Wen. Towards a systematic survey for carbon neutral data centers. *IEEE Communications Surveys & Tutorials*, 2022.

[5] Kevin Dunlap and Neil Rasmussen. Choosing between room, row, and rack-based cooling for data centers. *APC White Paper*, 130, 2012.

[6] Nevena Lazic, Craig Boutilier, Tyler Lu, Eehern Wong, Binz Roy, MK Ryu, and Greg Imwalle. Data center cooling using model-predictive control. *Advances in Neural Information Processing Systems*, 31, 2018.

[7] Jay Lee and Pradeep Kundu. Integrated cyber-physical systems and industrial metaverse for remote manufacturing. *Manufacturing Letters*, 34:12–15, 2022.

[8] Ning Liu, Zhe Li, Jielong Xu, Zhiyuan Xu, Sheng Lin, Qinru Qiu, Jian Tang, and Yanzhi Wang. A hierarchical framework of cloud resource allocation and power management using deep reinforcement learning. In *2017 IEEE 37th international conference on distributed computing systems (ICDCS)*, pages 372–382. IEEE, 2017.

[9] Tiziana Margaria and Alexander Schieweck. The digital thread in industry 4.0. In *Integrated Formal Methods: 15th International Conference, IFM 2019, Bergen, Norway, December 2–6, 2019, Proceedings 15*, pages 3–24. Springer, 2019.

[10] Mordor Intelligence. *Asia Pacific Green Data Center Market*, 2022.

[11] Robert Bob Sullivan, Guoqiang Li, and Xiaofei Zhang. Cold aisle or hot aisle containment-is one better than the other? In *2018 IEEE International Telecommunications Energy Conference (INTELEC)*, pages 1–4. IEEE, 2018.

[12] Fei Tao and Qinglin Qi. Make more digital twins. *Nature*, 573(7775):490–491, 2019.

[13] Jianxiong Wan, Jie Zhou, and Xiang Gui. Intelligent rack-level cooling management in data centers with active ventilation tiles: A deep reinforcement learning approach. *IEEE Intelligent Systems*, 36(6):42–52, 2021.

[14] Deliang Yi, Xin Zhou, Yonggang Wen, and Rui Tan. Toward efficient compute-intensive job allocation for green data centers: A deep reinforcement learning approach. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pages 634–644. IEEE, 2019.