Phygital-data-centric Computing

Masahisa Kawashima

Abstract

The artificial intelligence (AI)/Internet of Things initiatives being undertaken in many countries will lead to a new computing paradigm called *phygital-data-centric computing*, which will create data servers near the physical world and their clients on the cloud. NTT Software Innovation Center is developing technologies necessary for value-generating, cost-effective, and operable phygital-data-centric computing. In particular, it is conducting research and development in three focus areas, 1) AI computing infrastructure, 2) data hub/pipelines, and 3) advanced analytics.

Keywords: AI, IoT, data management, data analytics, Post Moore, Society 5.0

1. Evolution of the information society

As advocated in Society 5.0 by the Japanese government in 2016, many countries are aiming to innovate human activities in the physical world with artificial intelligence (AI)/Internet of Things (IoT) technologies. This article introduces some of the technologies that NTT Software Innovation Center (SIC) is working on to achieve such aims. This section describes how information technology (IT) systems should evolve.

1.1 From individual optimization to overall optimization

Many IT systems achieve only individual optimization, creating a non-optimal condition for the overall benefits to society. One example is a car navigation system that selects the shortest route for a car without taking into account the current traffic of the route. If self-driving vehicles and mobility as a service (MaaS) are developed with this approach, roads that are prone to traffic jams, such as those in Japanese and other world cities, would become even more congested. Self-driving cars and MaaS would not be practical without a mechanism for overall traffic optimization for avoiding traffic jams. One such mechanism may collect the destinations of all people and vehicles on the road, reduce total traffic by arranging ride-sharing groups, and avoid traffic congestion by selecting second-best routes for some cars. Therefore, it will be important to optimize entire systems that coordinate multiple people and things.

1.2 Creating business opportunities and avoiding disasters

There are many data-lake products and business intelligence (BI) tools on the market, and many companies have built so-called systems-of-insight (SoI) to achieve BI and optimize business activities. However, typical SoI implementations update data in a data lake with daily batch processes. This implies that whatever advanced AI algorithms we develop, they will be applied only to non-fresh data collected the previous day.

If we can accelerate data flows so that data are aggregated for analysis in several minutes or seconds, it would enable us to generate new value. For example, the retail industry would be able to operate "mobile" stores, dynamically moving them to gathering points such as sporting events. Since the loadable capacity would be limited, selecting merchandise to be loaded with hourly demand prediction would be key. Another example would be emergency evacuation. If we can run evacuation buses based on the current and precise whereabouts of people and road conditions, it would greatly assist people in evacuating disaster-hit areas. Therefore, increasing data velocity for responsive human activities would generate significant value, enabling us to create business opportunities and avoid disasters.
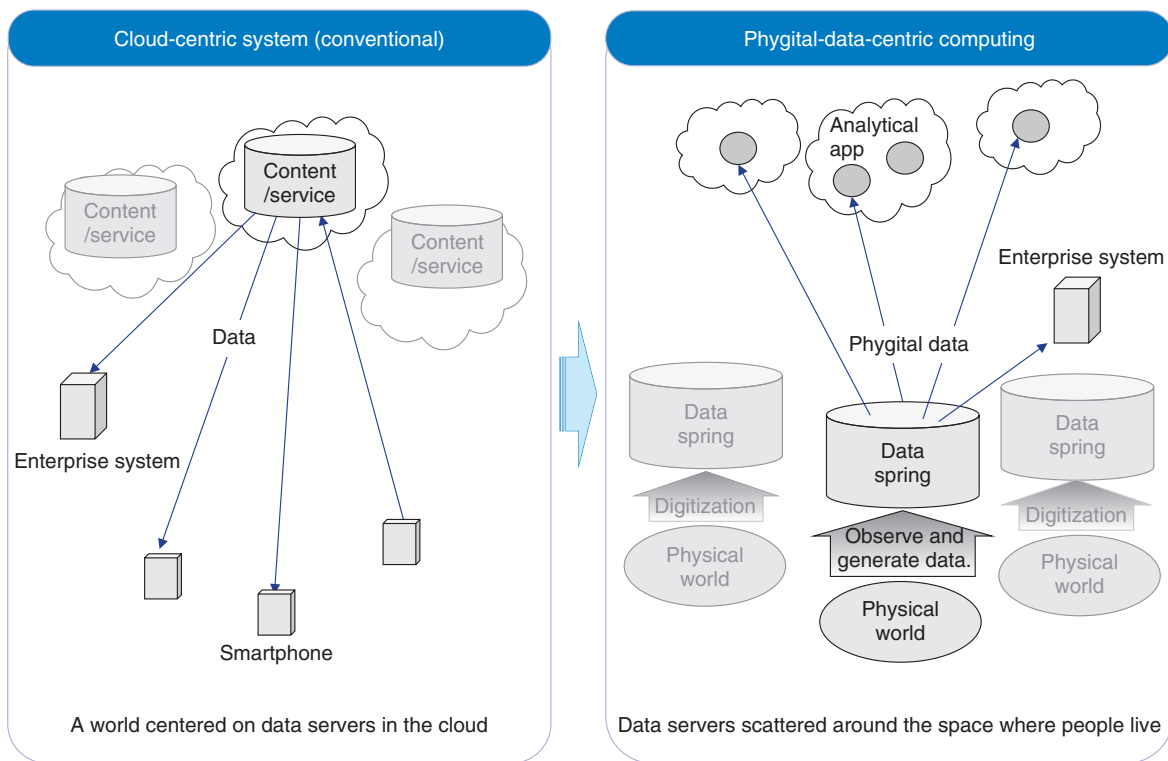
Fig. 1.   Phygital-data-centric computing.

2.   Phygital-data-centric computing

The IT evolution mentioned above will reverse data flows. In today's Internet, data flow from servers on the cloud to devices in the physical world. The IT infrastructure has evolved to serve more data in this direction. However, the new AI/IoT technologies like those in the examples mentioned above will be implemented with a new computing paradigm called *phygital-data-centric computing* (**Fig. 1**) with the following design principles.

(1)   Observe people or things in the physical world and generate data reflecting that observation. We call such data *phygital data*.

(2)   Create servers of phygital data, i.e., *data springs*, in the proximity of observed people or things.

(3)   Develop on the cloud analytical applications that access data springs and generate value out of data.

3.   Research and development (R&D) areas for phygital-data-centric computing

NTT SIC is developing technologies needed for

value-generating, cost-effective, and operable implementation of phygital-data-centric computing. In particular, we are conducting R&D in the following three areas.

3.1   AI computing infrastructure for generating phygital data

Flexibility is critical for phygital data generation. For example, in typical camera-based shopper-behavior analysis for retail stores, a shopper's age and sex are estimated and logged. However, some may want to log the shopper's body shape as well. Some may even want to log his/her actions such as comparing products, reading product displays, and picking up a product. Accommodating such varying demands will inevitably require the deployment of software-based AI-inference runtimes, which are computationally intensive.

Even a small convenience store would have to install six to eight cameras, and shop owners may want to apply multiple phygital-data-generation methods to their cameras. For example, they may want to generate population-density heat maps, shoplifting-prevention alerts, and notifications for

disabled shoppers. Even a small convenience store would have to deploy 20–50 AI-inference runtimes, which incurs a huge computational load.

In response to this issue, we developed Carrier Cloud for Deep Learning, which accommodates a massive amount of deep learning inference tasks, leveraging our deep-learning runtime-optimization technology. This deep-learning inference-processing platform is explained in "Carrier Cloud for Deep Learning to Enable Highly Efficient Inference Processing—R&D Technologies as a Source of Competitive Power in Company Activities" in this issue [1].

3.2 Data hub/pipelines for data flowing from data springs to analytical applications

Enabling an analytical application to access fresh data is not straightforward due to issues with data volume and the scattered nature of data springs and analytical applications.

Data volume is huge and unpredictable in many cases. For example, we may want to implement real-time monitoring of a car's geolocation, which will be very helpful in achieving the above-mentioned smart traffic distribution. Provided that the number of connected cars on the road is probably in the millions and that car density varies among regions and is very unpredictable, a cost-effective implementation of a real-time car-location database is not straightforward. Moreover, if geolocation data become more precise with a global navigation satellite system, we will have to increase the location-update frequency accordingly, which will significantly increase the database workload. In response to this issue, we are developing a high-speed spatio-temporal database data-management system called *Axispot*™, which is introduced in "Introduction to Axispot™, Real-time Spatio-temporal Data-management System, and Its High-speed Spatio-temporal Data-search Technology" in this issue [2].

Both data springs and analytical applications are geographically scattered by nature. Let us consider real-time population density heat maps and temperature/humidity maps of commercial buildings. Such heat maps will be useful for various business purposes such as sales forecasting, dynamic marketing, and public safety. Data springs, i.e., data sources for heat maps, would be created in the proximity of buildings, and analytical applications would run in various places including public clouds and corporate datacenters. How could data springs responsively notify analytical applications of events while avoid-

ing unnecessary communication traffic? In many cases, as data are privacy sensitive, data owners would not be able to share their data without proper privacy concealment. However, privacy concealment would make generating value from data difficult. How can we address this dilemma? We are developing the *iChie* data hub, which is described in "iChie: Speeding up Data Collaboration between Companies" in this issue [3].

3.3 Advanced analytics for adding value to phygital data

Last but most important is analytics. The AI computing infrastructure and data hub/pipelines mentioned above will prepare various phygital data for analytics. However, without further advancement in data-analytics algorithms and tools, people would not be able to generate much value out of phygital data.

One reason is the shortage of data scientists. In many cases, value-generating data models should handle many variables as input, which increases the obstacles for application developers. Without data scientists with advanced skills, such data models cannot be properly developed. In response, we are developing a data modeling automation technology called *RakuDA* [4] and a technology called *t-VAE* [5] for suppressing instability in AI model training for anomaly detection.

Another reason lies in the fundamental difference between calculation and analytical computing. Conventional computers were designed to execute many calculations or comparisons, i.e., $+$, $-$, $\times$, $\div$, $<$, $>$, $=$, and $\neq$ tasks. However, many optimization problems cannot be translated into a sequence of calculation/comparison tasks. We have to compare all possible combinations with such computers, which is not very practical. Let us consider the above-mentioned smart traffic distribution for self-driving cars and MaaS. We would have to optimize each car's route and ride-sharing groups, taking into account various data such as travel demands, current and predicted road traffic, ride-sharing car availability, and special needs of passengers. It is difficult for conventional computers to solve optimization problems in real time under conditions in which the number of combinations increases exponentially. In response, we are developing a hybrid computation platform for combinational optimization, which is featured in "LASOLV™ Computing System: Hybrid Platform for Efficient Combinatorial Optimization" in this issue [6].

Some analytical applications, such as those for trading, may involve transactional database updates.

Transactional databases have been implemented using a scale-up approach, i.e., improving a single server's performance in accordance with an increase in workload. However, as Moore's Law is reaching its limit, we cannot expect further performance improvement. Therefore, we developed high-speed transaction-processing technology, which is introduced in "A Method for High-speed Transaction Processing on Many-core CPU" in this issue [7].

4. Future directions

As described above, NTT SIC is putting great effort into phygital-data-centric computing. We will make our technologies ready for beta evaluation from the early stages of development and improve upon them after receiving feedback from NTT's partners and customers.

References

[1] D. Hamuro, K. Iida, K. Usami, S. Yura, Y. Matsuo, T. Eda, A. Saka-moto, M. Toyama, K. Mikami, N. Inoue, R. Nakayama, S. Enomoto, T. Sasaki, X. Shi, Y. Hirokawa, and K. Inaya, "Carrier Cloud for Deep Learning to Enable Highly Efficient Inference Processing—R&D Technologies as a Source of Competitive Power in Company Activities," NTT Technical Review, Vol. 18, No. 1, pp. 15–21, 2020.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr202001fa2.html
[2] M. Hanadate, T. Kimura, N. Oki, N. Shigematsu, I. Ueno, I. Naito, T. Kubo, K. Miyahara, and A. Isomura, "Introduction to Axispot™, Real-time Spatio-temporal Data-management System, and Its High-speed Spatio-temporal Data-search Technology," NTT Technical Review, Vol. 18, No. 1, pp. 22–29, 2020.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr202001fa3.html
[3] N. Yamamoto, D. Tokunaga, and S. Mochida, "iChie: Speeding up Data Collaboration between Companies," NTT Technical Review, Vol. 18, No. 1, pp. 30–34, 2020.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr202001fa4.html
[4] NTT, "RakuDA: Automated Data Analysis," NTT R&D Forum 2018 Autumn, H13, 2018.
https://labevent.ecl.ntt.co.jp/forum2018a/elements/pdf_eng/H13_e.pdf
[5] H. Takahashi, T. Iwata, Y. Yamanaka, M. Yamada, and S. Yagi, "Student-t Variational Autoencoder for Robust Density Estimation," Proc. of International Joint Conferences on Artificial Intelligence Organization (IJCAI) 2018, pp. 2696–2702, Stockholm, Sweden, July 2018.
[6] J. Arai, S. Yagi, H. Uchiyama, K. Tomita, K. Miyahara, T. Tomoe, and K. Horikawa, "LASOLV™ Computing System: Hybrid Platform for Efficient Combinatorial Optimization," NTT Technical Review, Vol. 18, No. 1, pp. 35–40, 2020.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr202001fa5.html
[7] S. Nakazono and H. Uchiyama, "A Method for High-speed Transaction Processing on Many-core CPU," NTT Technical Review, Vol. 18, No. 1, pp. 41–44, 2020.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr202001fa6.html

Masahisa Kawashima
Vice President, Head of NTT Software Innovation Center.
He received a Ph.D. in electrical engineering from Waseda University, Tokyo, in 1994 and an M.Sc. in technology management from MIT's Sloan School of Management, USA, in 2002. Dr. Kawashima is the head of NTT Software Innovation Center. He has been engaged in technology and business development at NTT since joining the organization in 1994. With his enthusiasm for technology strategies, he has accomplished several initiatives to update NTT's business operations in line with new technology trends. He played a leading role in the organization of the NTT Open Source Software Center in 2006, the strategic Wi-Fi service renewal for NTT WEST in 2011, and the release of the network-functions-virtualization-enabled software-defined wide area network service platform called CLOUDWAN for NTT's global businesses in 2017.