Akraino White Paper

Computing infrastructure in 2030s

~LF Edge and IOWN GF Joint PoC~

<1st edition – February 2024>

1. Abstract

Edge computing is an important technology for a data-driven society that makes decisions in real time from data, because it can reduce latency by processing at the edge near the source of the data. LF Edge, a project of The Linux Foundation, is working to establish an open and interoperable edge computing framework that is independent of hardware, processor, cloud, and operating system, which is a challenge for the practical application of edge computing. LF Edge is also focused on enabling edge AI because of the growing need for real-time AI processing of huge amounts of data, such as factory automation and autonomous driving. However, there are challenges to achieving an infrastructure that is both energy efficient and flexible, such as increasing data volumes, increasing the computational complexity of AI models, and meeting application-specific requirements such as latency and memory capacity. So we need innovation in infrastructure technology. Therefore, The Linux Foundation and IOWN GF(Innovative Optical and Wireless Network Global Forum) signed a basic agreement in June 2023 to integrate the Linux Foundation's software on the infrastructure proposed by IOWN GF to develop a common infrastructure that improves performance, reduces latency, and improves energy efficiency. Based on this agreement, we planned the IOWN GF/LF Edge Joint PoC(Proof of Concept) to demonstrate the convergence of the two technologies and the performance improvements. The PoC will build an end-to-end environment from edge to cloud using IOWN GF's infrastructure and LF Edge's platform and software, and run actual application in the environment for demonstration. This paper describes the content of the IOWN GF/LF Edge Joint PoC.

2. Background

In addition to establishing an open and interoperable edge computing framework, LF Edge is also focused on enabling edge AI in response to the growing need for real-time AI processing of large amounts of data, such as factory automation and autonomous driving.

In the AI field, as shown in Figure 1, the computational performance required for computing infrastructures is rapidly increasing due to the increase in the amount of learning data and the scale of generative AI models [1] [2]. This is

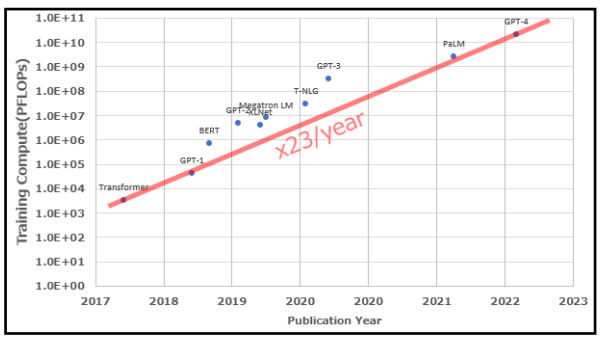

Figure 1: Computational Performance Required for AI Models

faster than the performance evolution "Moore's Law" of CPU and GPU which are responsible for the computation processing of AI. Computing infrastructures are following this speed by increasing the number of parallel servers with CPUs and GPUs. But the power consumption of data centers will be about 15 times higher in 2030 than today, leading to power shortages at this rate[3] [4] [5].

In addition, as shown in Figure 2, the performance required for each application of AI has diversified [6]. For example, prefill (LLM Inf prefill) for large language model inference emphasizes computational performance over memory bandwidth, while decode (LLM Decode) for large language model inference is the opposite. Currently, the infrastructure is built on a per-server basis, based on the performance that applications value most among compute, memory, and network resources. Therefore Some resources are underutilized, causing the infrastructure to scale.

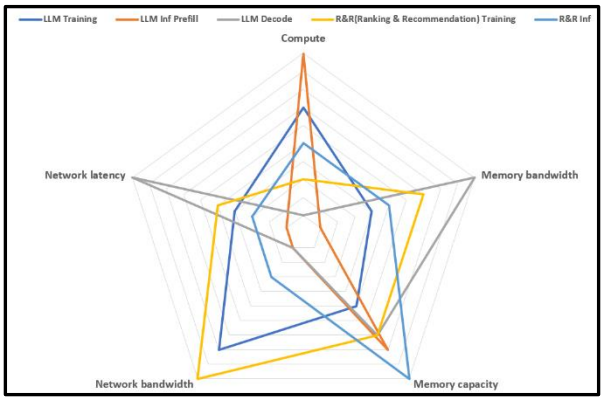

Figure 2: Performance Requirements for AI Apps

Therefore, we need a computing infrastructure that is more power efficient and more flexible in the allocation of resources to sustainably realize the growing scale and diversity of AI applications. This challenge should be more important in environments where computing infrastructure such as the edge has limitations such as power and space. We considered a change in the architecture of the computational infrastructure to be one of the effective solutions and tried it.

3. Initiatives to solve problems

The IOWN GF (Innovative Optical and Wireless Network Global Forum) proposes a network and information processing infrastructure, the Data Centric Infrastructure (DCI), which can provide high-speed, high-capacity communications and enormous computing resources that exceed the limits of existing infrastructure by utilizing innovative technologies such as optics [7] . The Linux Foundation and the IOWN GF(Innovative Optical and Wireless Network Global Forum) signed a basic agreement in June 2023 to integrate the Linux Foundation's software onto the infrastructure proposed by the IOWN GF to develop a common infrastructure that improves

performance, latency and energy efficiency [8]. LF Edge sees DCI as one of the potential solutions to these challenges in the edge computing. Therefore, we planned the IOWN GF/LF Edge Joint PoC(Proof of Concept) to demonstrate the convergence of the two technologies and the performance improvements by DCI. The PoC will build an end-to-end environment from edge to cloud using IOWN GF's infrastructure and LF Edge's platform and software, and run actual application in the environment for demonstration. The diagram below provides an overview of Joint PoC.

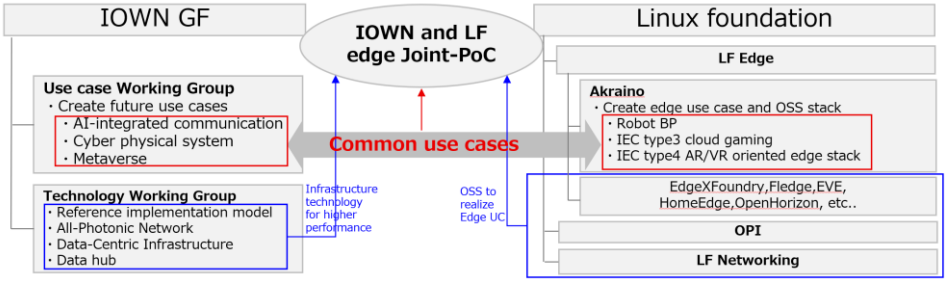

Figure 3: IOWN GF/LF Edge Joint PoC Concept

4. Infrastructure technologies proposed by IOWN GF

   IOWN GF advocates the Data Centric Infrastructure (DCI) as a new efficient and flexible computing infrastructure. There are two major features compared to current computing infrastructure.

   First, disaggregation allows for flexible device relocation. For example, if each application requires different resources, as shown in the following figure, the current computing architecture "server-oriented" allocates resources to each application on a per-server basis, which may leave some devices unused and reduce utilization efficiency. On the other hand, disaggregation can efficiently meet the requirements of various applications by selecting the required devices from the resource pool according to the application and configuring logical nodes through PCIe switches.

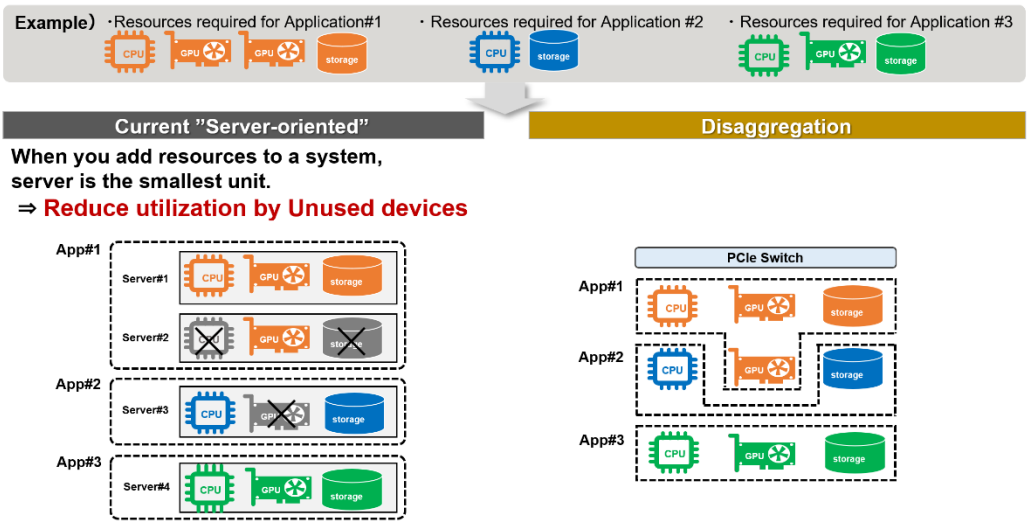

Figure 4: Disaggregation

Second is direct connection between accelerator device through optical connections or shared memory. In the current computing architecture, CPU centric, data is transferred between accelerator devices by the CPU through CPU memory. On the other hand, in the case of direct connection, data is directly transferred between accelerator devices via shared memory or optical switches. This reduces the utilization of CPU cores used for data transfer, thereby reducing cost and power consumption. The processing delay can also be reduced.

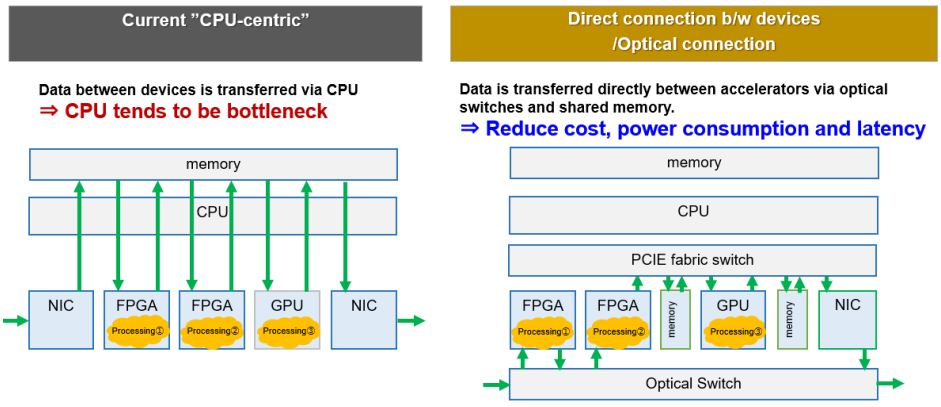

Figure 5: direct transfer between accelerator devices

The following figure shows the outline and mechanism of DCI. First, the user requests DCI the resources required to run the application, and the controller of the DCI responds by issuing instructions to configure logical nodes to the Composable Disaggregated

Infrastructure (CDI) management software that manages the resource pool and logical nodes. The controller also configures an optical switch that is responsible for the direct connection between the accelerator devices.

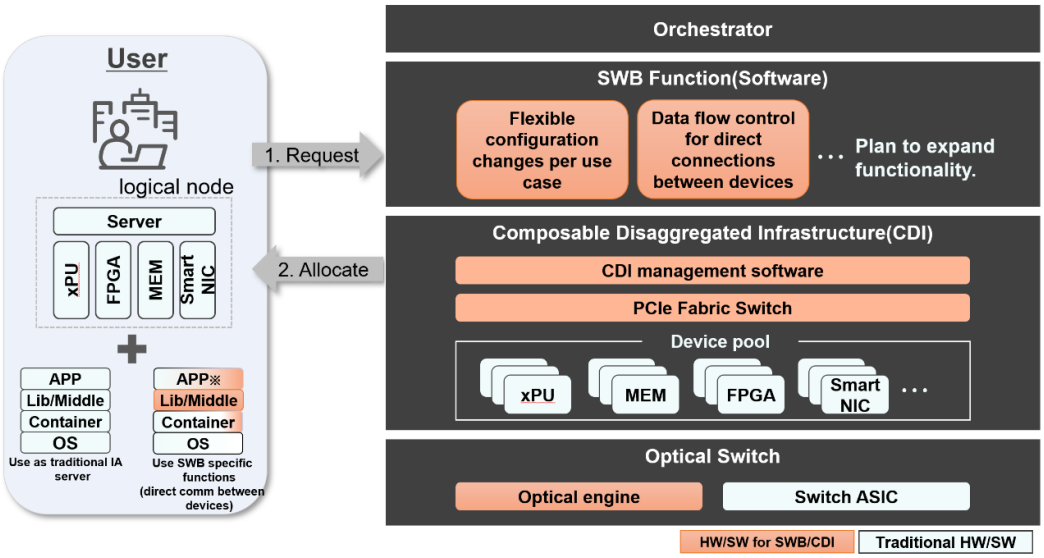

Figure 6: Data Centric Infrastructure (DCI)

5. About IOWN GF/LF Edge Joint PoC

This chapter describes the details of the IOWN GF/LF Edge Joint PoC. In the PoC, we will demonstrate the following two points. The first point is the convergence of the IOWN GF and LF Edge technologies. We will confirm that LF Edge platform and software can be integrated on DCI, and we will extract the mechanisms necessary to manage and control it as a resource by changing from a current server architecture to DCI through actual environment construction. The second point is the demonstration of improved processing performance and reduced power consumption by applying DCI. For this demonstration, we will implement video inference by AI, a typical processing of edge AI, as an end-to-end application to access, edge, and cloud areas as shown in the figure below. The video is input from the access area, sent to the edge cloud via 5GRAN and the core network, and inference processing by AI is performed. The processed video is sent to the cloud. Here, we will emulate and use DCI using commercially available composable disaggregation products for computing resources in the edge cloud. AI inference processing is divided into three phases: decoding, filtering/resizing, and inference, which are implemented in accelerator devices such as FPGAs and GPUs, and data is transmitted and received by direct connection between accelerator device, one of the features of DCI. We also measure the improvement in processing performance and reduction in power consumption by applying DCI compared with

using a current server architecture as a computing resource. In addition, we will implement LF Edge Akraino IEC (Integrated Edge Cloud) type2 Blueprint on DCI as middleware for building edge cloud environments to verify that DCI and LF Edge edge computing platforms can be integrated.

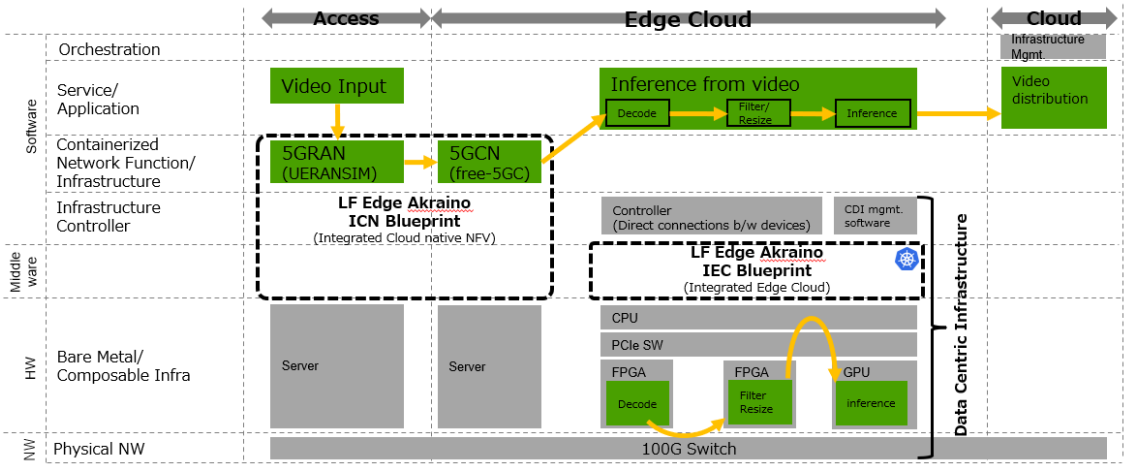

Figure 7: IOWN GF/LF Edge Joint PoC Details

6. Conclusion

We will built the end-to-end environment described above as a base for future demonstrations and report on it in April 2024 as the second edition of this paper. In the second edition, we will report the convergence results of DCI proposed by IOWN GF and LF Edge platform and software, the first demonstration point of PoC. We will continue to demonstrate the improvement of processing performance and reduction of power consumption by applying DCI, which is another demonstration point, by comparing before and after the application of DCI. In addition, we plan to examine the content of the demonstration experiment in the future regarding flexibility among the challenges of computing mentioned at the beginning. For example, an all-photonic network (APN) that connects edge sites directly by light and disaggregation proposed by IOWN may combine resources from distant edge sites to more flexibly construct logical nodes. We will continue to demonstrate how LF Edge's platform and software can be used in such example. We would also like to consider publishing this activity as an IOWN GF PoC report. As a result, the collaboration between IOWN GF and LF Edge will accelerate the evolution of technology and the realization of synergies, which will contribute to the development of the industry.

7. Abbreviation

1	IOWN GF	Innovative Optical and Wireless Network Global Forum
2	CPU	Central Processing Unit
3	GPU	Graphics Processing Unit
4	DCI	Data-Centric Infrastructure
5	PCIe	Peripheral Component Interconnect-Express
6	FPGA	Field Programmable Gate Array

References

[1] A. Gholami, "AI and Memory Wall," 30 3 2021. [Online]. Available: https://medium.com/riselab/ai-and-memory-wall-2cb4265cb0b8.

[2] S. McAleese, "Retrospective on 'GPT-4 Predictions' After the Release of GPT-4," 18 3 2023. [Online]. Available: https://www.lesswrong.com/posts/iQx2eeHKLwgBYdWPZ/retrospective-on-gpt-4-predictions-after-the-release-of-gpt.

[3] S. R. a. F. Boshell, "The nexus between data centres, efficiency and renewables: a role model for the energy transition," 26 6 2020. [Online]. Available: https://energypost.eu/the-nexus-between-data-centres-efficiency-and-renewables-a-role-model-for-the-energy-transition/.

[4] Japan Science and Technology Agency, "Impact of Progress of Information Society on Energy Consumption," 2 2020. [Online]. Available: https://www.jst.go.jp/lcs/pdf/fy2020-pp-03.pdf.

[5] N. Willing, "New Technologies Are Needed to Curb Data Center Energy Use, Says the IEA," 3 8 2023. [Online]. Available: https://www.techopedia.com/new-technologies-are-needed-to-curb-data-center-energy-use-says-the-iea.

[6] M. Omar Baldonado, "Meta's evolution of network for AI," 17 10 2023. [Online]. Available: https://www.youtube.com/watch?v=5gOOtFySrqA.

[7] IOWN GLOBAL FORUM, "DCI Product Concept Paper," 19 10 2023. [Online]. Available: https://iowngf.org/wp-content/uploads/formidable/21/IOWN-GF-RD-DCI_PCP-1.1.pdf. [Accessed 9 2 2024].

[8] IOWN GLOBAL FORUM, "Linux Foundation and IOWN Global Forum to Collaborate for Future Smart Connected World," IOWN GLOBAL FORUM, 14 6 2023. [Online]. Available: https://iowngf.org/press-releases/linux-foundation-and-iown-global-forum-to-collaborate-for-future-smart-connected-world/. [Accessed 9 2 2024].

Authors

- Haruhisa Fukano(Fujitsu)
- Toshimichi Fukuda(Fujitsu)
- Reo Inoue(Fujitsu)