Summary Translation of Question & Answer Session at
Fujitsu Research Strategy Briefing Session for Media, Analysts, and Investors

Date:	June 4, 2024
Location:	Station Conference Kawasaki
Presenters:	Seishi Okamoto, Corporate Executive Officer, EVP, Head of Fujitsu Research Toshihiro Sonoda, Head of Artificial Intelligence Laboratory, Fujitsu Research Naoki Shinjo, SVP, Head of Advanced Technology Development Unit, Fujitsu Research

Questioner A

Q1: As companies are developing generative AI initiatives, such as NTT's tsuzumi and NEC's cotomi, if there is something distinctive about Fujitsu's AI, or if it has superior competitiveness, please tell us what it is.

A1 (Sonoda): The retrieval-augmented generation (RAG) part of the framework we presented today has the world's highest level of accuracy in terms of handling extremely large amounts of data. We believe that is one important issue when companies use generative AI, and that is one point of differentiation for us.

Fujitsu's generative AI amalgamation technology is also unique. It is generative AI amalgamation technology that, by selecting the best generative AI using the specialized data of a company, is able to meet the variety of use cases for a company. Its unique feature is that it is able to be provided without the need to customize generative AI for detailed fine-tuning on the part of the company.

Q2: In building your specialized AI business, please tell us if there are any industry-based or business area-based use cases that you are envisioning for clients. Please also tell us if you have any targets for the scale of this business.

A2 (Sonoda): In terms of specific use cases, we are thinking of software engineering and document-related analysis. The demo we showed was for checking contracts, but it can be for all of the documents a company uses. We also want it to be used for analyzing network log data. Another unique feature is in the area of imaging data. Fujitsu has long been involved in the area of vision-based AI, and we want it to be able to handle vision-based data. For example, we are working to provide a solution that can take monitoring-related imaging data from long-term monitoring and analyze it with generative AI. We also showed an exhibit on this today. The demos we exhibited today are all areas in which we are focusing our efforts.

(Okamoto): In terms of generative AI, we have had around 300 to 400 discussions with customers, and, even within AI, we want to build a large-scale business.

Questioner B

Q1: You mentioned that you are developing the world's first integrated analysis system for authenticity judgement, but has this been completed? If it has not yet been completed, when do you expect to complete it?

A1 (Okamoto): It is still under development. Rather than developing this system on our own, we would like to build it in collaboration with many other stakeholders. As for the timing, we could not prepare an announcement in time for today's briefing, but we plan to prepare a detailed announcement on the timing in the near future.

Questioner C

Q1: I understand that AI Computing Broker dramatically reduces power consumption by increasing GPU utilization. Am I correct in understanding that, in terms of the power consumed when the GPUs are idle, by reducing idle times by efficiently allocating jobs, it enables the power that would otherwise be wasted to instead be conserved?

A1 (Okamoto): Yes, that is correct. In relation to the jobs, it takes fairly complex technology to see the extent to which idle GPUs can be reduced, but using AI computing Broker efficiently allocates resources. We also have a demo exhibit on this, so please take a look at it.

Q2: What is the process for calculating a reduction of over 10 terawatt hours per year using AI Computing Broker? For example, if we took the case applying this technology to the existing GPUs in Japan, what would be the calculation process?

A2 (Okamoto): Because there are multiple sources of information, please allow me to provide you with the calculation process afterwards.

*Public and Investor Relations Division has provided the following answer:
Of the GPUs from Nvidia in fiscal 2023, shipments for generative AI were 550,000 units. If the latest GPUs consume roughly 1,000 watts of power per GPU, then that would result in roughly 5 terawatts per hour of power consumption. If we assume the compound annual growth rate in unit shipments is 50% per year (from TrendForce), in four years that amounts to 2 million units, and if this technology is widely deployed, we estimate that it could reduce the roughly 20 terawatt hours of power consumption by half.

Questioner D

Q1. In generative AI, compared to general-purpose interactive large language models (LLMs), is the market for specialized generative AI models larger? As a backdrop to the question, in terms of the market structure for generative AI, how do you see it evolving, including in terms of scale, and why has Fujitsu chosen to focus on specialized models?

A1 (Okamoto): In relation to ordinary LLMs as general-purpose platforms, we think the strategy for expanding users will be through internet use. As Sonoda explained earlier, we are developing specialized LLMs for enterprise use. Our view is that enterprise use of LLMs will be significant, and we think the market for enterprise use will be very sizeable.

Questioner E

Q1: The main theme of today's presentation was "Combining Technologies," but what are some of the new initiatives at Fujitsu Research to pursue this theme? Please also tell us about the collaboration among your locations around the world.

A1 (Okamoto): Currently, with our Global OneTeam organization, one of our very unique features is that each research topic is being pursued through collaborations with our other locations. In doing so, we want to leverage the distinct strengths of each location. For example, in our location in India, which we established two years ago, we recruited people with strengths in science and mathematics to work on the fields of AI and quantum computing. We also have a team of software researchers there working on Fujitsu-MONAKA. In combining technologies, it is very important to efficiently advance research work through a matrix structure, so, for example, we shift AI researchers to computing, or we create virtual organizations to work on projects. We use a variety of structures to drive our work.

Q2: How many employees currently work at your location in India, and do you have plans to further increase your employees there?

A2 (Okamoto): We have about 90 employees at Fujitsu Research in India, but Fujitsu as a whole, including the network unit, has nearly 400 employees in India. We are thinking of expanding the number of employees at Fujitsu Research in India.

Questioner F

Q1: You mentioned that the factors that differentiate your use of generative AI for enterprises are your knowledge graph extended RAD and generative AI amalgamation technology. Who are your rivals that are developing similar technologies in and outside of Japan? Please tell us what difficulties you have had catching up with these competing technologies.

A1 (Sonoda): In regard to knowledge graph extended RAG, we believe that, for example, Microsoft and Neos could team up and use something like a knowledge graph to create RAG. Fujitsu believes that knowledge graphs and LLMs are very compatible. We have been researching knowledge graphs for a long time, and have also been developing technology related to graphs, mainly in India, and have technologies to analyze large-scale graphs. When knowledge graphs become large, there is a need to analyze them and search the graph against the schema, but we have been at the forefront of these technologies, so we would like to be a leader in them rather than play catch up.

As for selecting a model that is the best comparison to amalgamation technology routing, there are similar technologies out there, such as SambaNova and Sakana AI, but I do not know the specifics of them. We do know that combining several models, such as a mixture of experts model, will increase accuracy, and have devised a way to smoothly route them, as we specialize in enterprises. We would like to be the leader in specialized AI for enterprises. There are various different use cases, so in particular we would like to sharpen our amalgamation technology in the area of smoothly extracting what should be chosen when creating specialized models for enterprises.

Questioner G

Q1: Regarding your generative AI framework for enterprises, it is my understanding that you are not creating models from scratch, but rather amalgamating existing models to create specialized models. I believe there is a trend among other Japanese companies of creating specialized models for Japanese enterprises, such as tsuzumi and cotomi, for example. Please tell us about the background behind why you chose this strategy, instead of creating a single core model from scratch.

A1 (Sonoda): As I mentioned previously, it has been confirmed that combining several models will increase accuracy. In addition, we believe that the concept for specialized models is not to have a large-scale general-purpose model, but instead to have models specialized in one area and improve their costs and responses. It is for this reason that we chose specialized models. We believe that, through these specialized models, we can achieve various functions, such as merging various things and running pipelines simultaneously, and we will be able to meet customers' needs and use cases. Another feature of these specialized models is that they also amalgamate not only generative AI, but also existing AI, such as, for example, optimization and predictive AI, that generative AI is said to be bad at. In the case studies that we presented as examples in the demo exhibition, we combined optimization and forecasting AI and then selected amalgamation to handle a variety of functions. We believe this will create a variety of services and value.

Q2: In your automatic generation of AI models, you mentioned that if the proper AI model does not exist, your technology will automatically generate the required AI model. How do you achieve this?

A2 (Sonoda): There are several patterns. We have technology that automates fine-tuning and prompts to generate specialized generative AI, as well as a technology called AutoML that is not generative AI, but automatically generates predictive models from data. In addition, our technology that was announced in the fall of last year is a technology that automatically generates a proper model through the interactive input of proper requirements. These technologies are the mechanism that automatically generates models through things such as the use of data to make predictions, the input of requirements and creation of proper models, and the input of information that you want the model to specialize in to create a specialized model.

Questioner H

Q1: You mentioned that you are advancing specialized models for generative AI. Please tell us whether the environment in which the specialized models will be used will be a mega-cloud infrastructure, or if you envision the models being used in a private environment, as they will be specialized for enterprises.

A1 (Sonoda): We, of course, believe that the models will also be used on-premises. Customers have requested for the models to be on-premises, and we are also considering using the models in a virtual private cloud environment. If a customer prefers to use the cloud, then we will set up the model there. Essentially, we would like to make it so that the models can be used anywhere.

Q2: Regarding Fujitsu-MONAKA, you mentioned about post-Fugaku, but I believe that this will require a substantial investment in R&D. From the perspective of return on investment, I believe that you are also looking at overseas markets, but could you please tell us if there are any differences you will need to consider in the R&D of Fujitsu-MONAKA for HPC and AI overseas compared to Japan?

A2 (Shinjo): We are, of course, not only thinking of the Japanese marketplace, but marketplaces overseas as well. One thing we are considering is the open-source community, which I believe I should say is, of course, mostly overseas. We would like to enter such areas to find out what sort of a demand there is and what sort of use cases are needed, and to be able to demonstrate our capabilities in these areas.

Questioner I

Q1: You mentioned rulemaking, but I believe that AI regulatory laws have been established in Europe, and legislation for AI is underway in the US and Japan. I believe that the points of interest and points emphasized in these laws vary by country. Could you please tell us how you will handle the difficulties regarding this?

A1 (Okamoto): Regarding how we will absorb the differences in regulations in each country, we believe that now is the time that we should make preparations for these regulations. Fujitsu does business globally, and we also have global locations for Fujitsu Research. Research is being done by the Global One Team, but we must understand what is happening in each country to handle AI regulations in each region. We have also released an AI Ethics Impact Assessment toolkit on Fujitsu Kozuchi to determine what sort of rules could have what sort of impact.

Q2: There was an explanation about your integrated analysis system for determining authenticity. Could you please tell us about the difficulties in handling generative AI as it rapidly advances and risks pop up one after another?

A2 (Okamoto): Various stakeholders are involved with the integrated analysis system for determining authenticity. We believe that there are areas that can only be handled with Fujitsu’s technology, and some areas where that is not the case. It is for this reason that we believe it will be very important to have various people and accredited organizations join in and create an ecosystem. I believe that we will have the opportunity to announce more information regarding this on a later date.

Questioner J

Q1: You mentioned that you want to be a global leader in such areas as knowledge graph technology and amalgamation technology, but are these technologies customized for the Japanese language? Or can they be used for any language? Please tell us about whether the focus of your business will be inside Japan or whether it will be global.

A1 (Okamoto): We plan to compete globally. We have a global research team, and, while we will also develop Japanese language solutions, we plan to support English and other necessary languages.