Esperanto's challenge to drive the evolution of RISC-V . A Simple Idea, Hard to Implement - An Interview with Dave Ditzel

Yutaka Yasuda

2022-09-30 36 share

In August 2022, when I returned to the United States for the first time in two and a half years, I visited two companies developing AI accelerators. One is Cerebras Systems, which was mentioned in the previous article , and the other is Esperanto Technologies (Photo 1) .

Photo 1 A building with offices. The blue sky of California fits well

Esperanto is one of the leading companies in RISC-V. Its founder, Dave Ditzel, co-authored the famous paper [1] in which David Patterson proposed the idea of RISC in 1980, and is an engineer who has been at the forefront of the processor industry for a long time. I first interviewed Dave in 2004, and since then [2] I've met him every few years to hear about his occasional work.

During my visit to Esperanto this time, I was able to ask about the current situation of Esperanto and talk about future prospects for semiconductor technology.

This visit was arranged by Mr. Eiji Kasahara, the company's senior architect. appreciate.

[1] "The case for the reduced instruction set computer", David A. Patterson, David R. Ditzel, 1980

recent situation

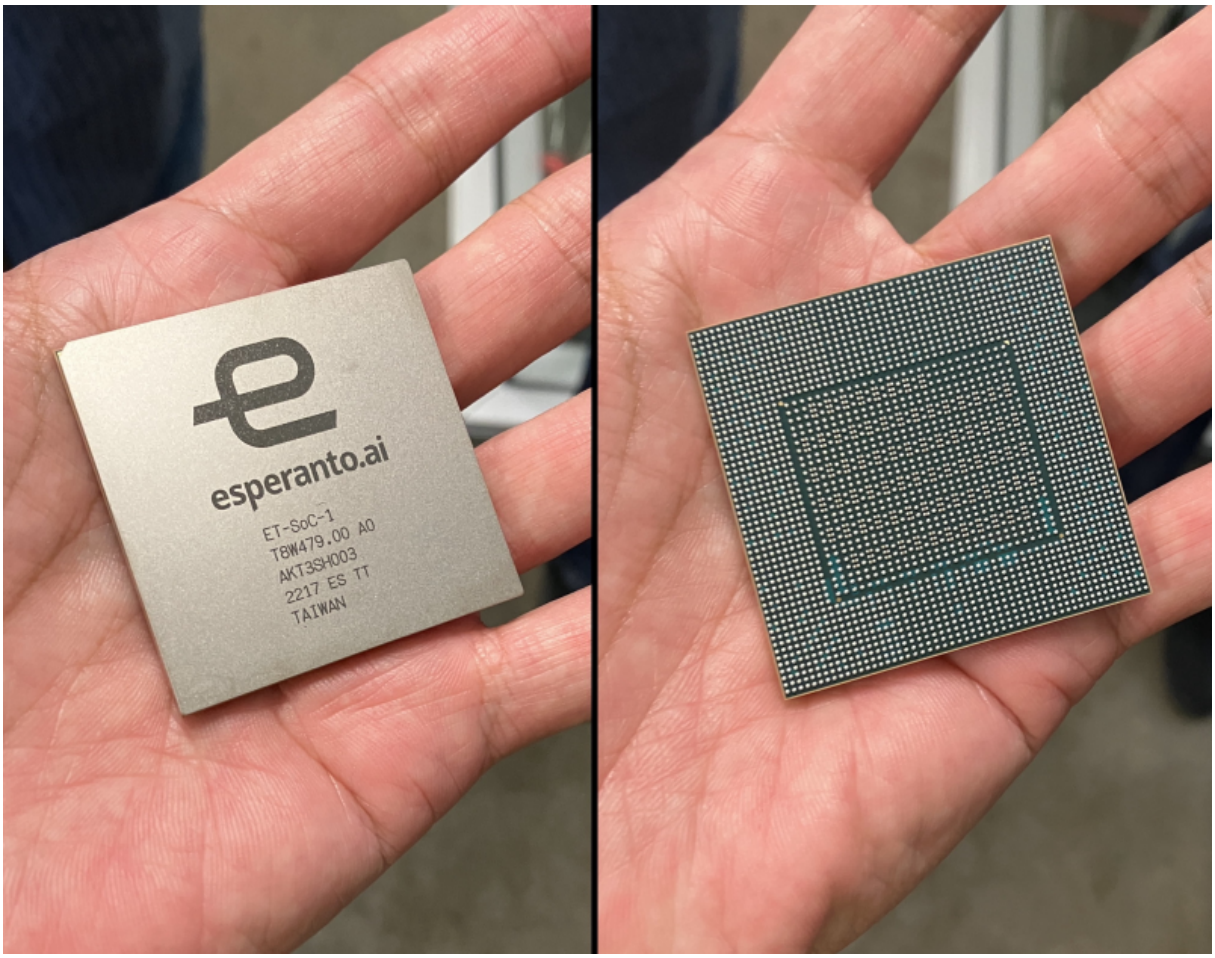
First, we asked Mr. Craig Cochran, VP of Corporate & IP Marketing, about the current situation of Esperanto. Since my last visit three years ago, they have moved to a larger office (Photo 2) .

Photo 2 Entrance of the new office. It's spacious, and you can see the greenery from the back window, creating a nice atmosphere.

They have already completed the first silicon. The world's largest RISC-V chip, ET-SoC-1 (Photo 3) . The chip has four high performance RISC-V cores called ET-Maxion and 1088 low power RISC-V cores with vector units called

processors.

Photo 3 ET-SoC-1 chip. big



The strength of their engine lies in the parallel computing power of Minions and how they set their power budgets. Between 10 Watts and 60 Watts, the voltage and frequency can be adjusted via software to meet the power requirements of a particular application.

> **Craig:** A lot of other companies doing AI processing acceleration are taking the approach of building systems on big chips that use your entire power budget. Systolic arraysCNNs. For dense networks like ResNet 50, it scores

include recommender systems and transformer models for natural language processing. Limited parallelism and operating at standard voltages lead to inefficiencies. Our approach is to run many chips in parallel and at low voltages.

A good trade-off between power and performance for our chip, what we call the sweet spot, is around 20 watts. For example, running at 20 watts per chip allows 6 chips to fit into a 120 watt power budget, which is more than four times better performance than other 1-chip systems running at 120 watts.

Cerebras, another AI accelerator company that I visited during my trip to the United States, also uses dataflow machines to handle sparse matrix operations. Since 2016, when Google announced the TPU, the processing of sparse matrices has been devised for each. Esperanto is trying to optimize this with memory placement and efficiency (processing power per power) .

In the interview, I asked about the purpose of grouping more than 1,000 minions (they call them shires) for that purpose, and about the ingenuity of local memory configuration using SRAM, but I will not discuss this in this article. For those who are interested, the announcement [3] at RISC-V Days Tokyo 2021 Autumn will be helpful.

[3] “ Esperanto Technologies @ RISC-V Tokyo Days 2021: Japanese Full Presentation ” , Eiji Kasahara, 2021

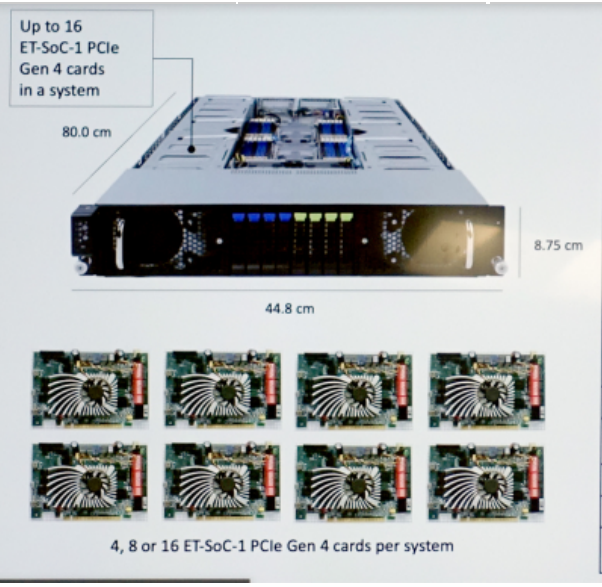
Photo 4 Building with offices. The blue sky of California fits well

Pre-production server

The ET-SoC-1 chip is currently undergoing initial customer evaluation and a server is being developed for it.

> **Craig:** Right now we're working with our first customer to develop a system that they can buy(picture 5) . The system is powered by two high-end Xeon processors and can accommodate 4, 8 or 16 of our cards. This is a high density solution, stacking up to 20 units in one rack.

Photo 5 Evaluation server



Server configuration	Standard 2U chassis
System host processor	2 x 3rd gen Intel® Xeon processor
System memory and storage	8-Channel RDIMM / LRDIMM DDR4 per processor, 24 x DIMMs
ET-SoC-1 PCIe cards	8 for standard system configurations; upgradable
ET-SoC-1 PCIe card details	Gen 4; 8 lanes; FH/HL
ET-SoC-1 performance	600-800 MHz
ET-SoC-1 power consumption	10W to 60W (workload dependent)
ET-SoC-1 RISC-V CPUs	8,448 (high performance, low power)
Pre-installed AI models	DLRM RMC 1, DLRM RMC 2, DLRM RMC 3, ResNet50, BERT small; additional models to be supported throughout 2H 2022
User interface programs	Jupyter Notebook / Lab & CLI
Performance, power, and trace analysis tools	Included (ET-PowerTop, Perfetto)
Training and documentation	Included
System Power	2 x 2,200W (C19 cable)
Connectivity	2 x 10Gb/s BASE-T LAN ports (Intel® X710-AT2)

The card with ET-SoC-1 has PCIe gen4 connection. Equipping a maximum of 16 cards in a 2U size is quite an amazing number. It's unlikely that everything will run at 60W, but the balance between the chassis and the number of cards that gives the optimum cost performance according to the load and processing speed requirements, and the power consumption (= processing capacity) of each card is flexible. can be found in

This high degree of freedom is important for large-scale cluster systems, including HPC, because optimizing cost efficiency is a top priority.

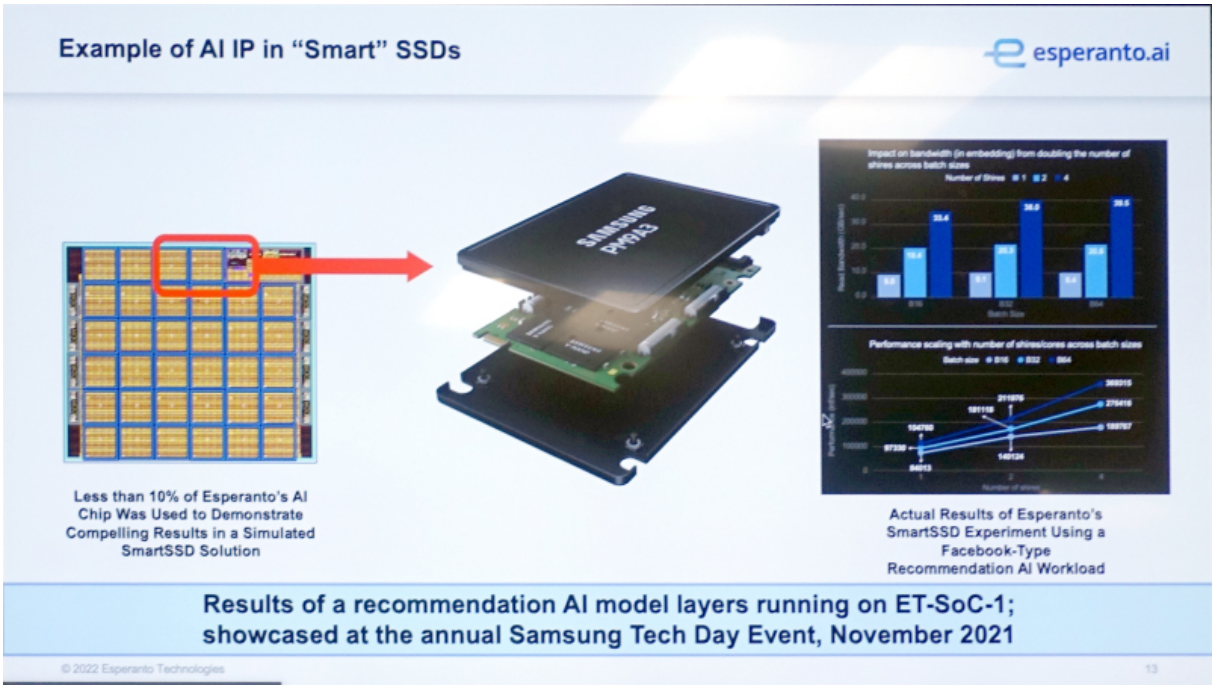
> **Craig:** We also offer pre-installed AI models. A DLR(Deep Learning Runtime)is a recommender model. There is also BERT, a transformer model for natural language processing. There is also ResNet50. Of course, you can also import your own models. It supports Jupyter Notebook as user interface. It's an open-source standard and good for experimentation.

Craig: We're focused from the cloud to the edge, starting with the data center. So our first product is for the data center, but we plan to expand to something called Fog, which sits between the cloud and the edge. Fogis perfect for San Francisco companies. (laughter)

We're working towards the edge of the network, not something like an IoT device.

Here are some such examples. Our target is smart SSD. Take Samsung as an example. We believe that by introducing AI functions into SSDs, computing can be done closer to storage.

Picture 6 Samsung Intelligent SSD Case

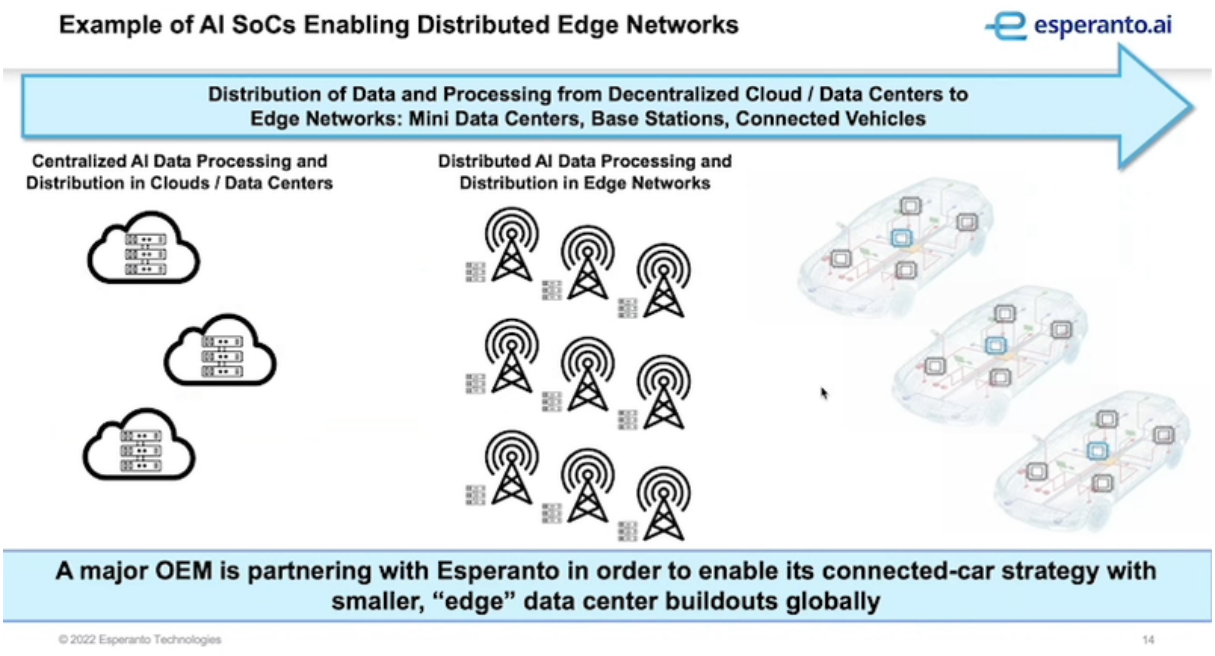


Author (Yasu) : I've seen intelligent SSDs and composabledata centers, but didn't know Samsung was using your SoC.

Yes, I knew that)We demonstrated at Samsung Tech Day in November 2021 that a small number of cores in our chips could handle it. Soeven DRAM may be unnecessary(by using SRAM built into the chip)

We are also committed to moving to a distributed edge network. I'm talking to a connected car company. They connect millions of cars with 5G modems to data centers, collect all kinds of sensor data, and put them in data centers, but they don't store all of the data, just the inferred results. and Therefore, we are considering installing our SoC in a car for inference.

Photo 7 Application example of AI SoC to distributed edge



Craig: So energy efficiency is very important not only for data centers but also for edge deployments. Our design is also highly scalable. This means that the architecture is orderly and can be easily scaled up or down depending on the application.

in data centers. Now that the SoC for that purpose has been completed, we are proceeding with expansion to the edge area. In addition, the pre-production server mentioned above uses Xeon as an external host processor, but it is said that Self-Hosted Mode, which does not require an external processor for edge use and uses Maxion as a host, is under development.

what was hard

I wanted to ask Dave a little bit about his journey so far.

> **Yasu:** It's been about four years since I first visited Esperanto. At that time, it was still a small office with only a few rooms. Now, the first chip, ET-SoC-1, has been released, and systems equipped with it are in the process of being created. What has been the hardest part so far?
>
> **Dave:** Right. Is it about allowing all the pieces to interact with each other? So the challenge was not the individual CPUs, but how they were interconnected. At first, I was worried that it would take too long to connect all 1000 CPUs correctly. We used Synopsys' ZeBu to run RTL simulations on a large FPGA system. I don't know how many thousands of FPGAs were used internally, but all 1000 CPUs were simulated. (Oh)
>
> And then we ran our software on it. It was slow, around 100 kilohertz instead of 1 gigahertz. Still, it's enough to find problems when a CPU can't communicate with (locally placed) SRAM memory on the die, or from SRAM memory to DRAM.

validate and debug, test all possible combinations, and write the tests for it.

(I see) I spent so much time on this. This was the hardest part, verifying all 1,000 CPUs.

Yasu: Sorry for the silly question. If you could do an RTL-level simulation, wouldn't that system be able to synthesize 100% correct wiring from the logical definition information? How could something like a wiring error occur?

Dave: That's because there was something missing in the original specification.

Even if the FPGA is prepared (on the simulator) , the missing interconnect information will be copied as it is. Also, even if the wire is connected, it may be connected to the wrong partner. But usually it's not such a simple wiring problem. There is something in the signal processing protocol and control algorithm.

Here's a case Dave gave us: In other words, if a cache fault occurs when loading an instruction, and at the same time a division by zero occurs in a floating operation.

However, there was only one register to hold who had interrupted, and "a situation where control could not be performed correctly due to reasons that were not initially assumed" would occur. While this example is too simplistic to be called "that's not enough thinking" , it's no surprise that complex protocol processing interconnecting so many processors and memories can lead to the unthinkable [4].] .

Yasu: By the way, if a test shows a problem like that, how do you determine what caused it? (Is it possible?)

Dave: Now that's writing a lot of tests. (meaning that you can only do that)

Yasu: Hmm,need a lot of test cases (too many objects)

Dave: I make a lot. Test coverage is also an issue. If you say"This gate never changed statethinking, " oh, am I testing this properly?" Then we have to create a special test case to make this gate False.

It took a lot of work, but the FPGA (simulator) system was very useful. The first silicon works fine and is shipping. Verification took longer than expected, but it was only a one-time tape-out.

Yasu: That sounds great.

At a conceptual level, Dave says it's easy. “ Put 1,000 CPUs on a chip? OK, super simple! ” But to actually implement it, it's just a matter of accumulating small details. " Everything in 25 billion transistors has to be perfect! "

Photo 8 Mr. Dave Ditzel

Inter-die connection

Next, I asked him about the future.

> **Yasu:** So what are you most looking forward to in the future?
>
> **Dave:** A change to a new implementation style, chiplet-based. we are going there (Yes, that's right.)Recently, there was a [5] big announcement about UCIe. So now that we have the busses to connect the individual dies into chips, a lot of our thinking now is how to partition them and what to do with them.
>
> We also have plans to improve minions. This is because the RISC-V community has completed a standard specification for vector units [6] . Esperanto will also use it in the next generation, but it will be compatible

we could use wasn't fixed, so we had no choice but to design them ourselves.

Yasu: Yes. You've talked about it before [7] .

Dave: And what happens as the process shrinks over generations? Chiplets can make really big chips, so today it would be like starting with 4,000 CPUs. Then there are 8,000 5 nanometers. If 3 nano is 16,000, 2 nano is 32,000, and 1.5 nanometers in 2030,there will be 64,000 CPUs in one package.

Yasu: It's a very big package, isn't it?

Dave: I think it's very likely that it will be a big package, and it will really change the future.

Yasu: By the way, Japan lost in the front-end process(competition), but it is still very strong in the back-end process.

Dave: Yes. I am currently writing another paper that looks ahead to 2030, and I have a message for you. " Hey!We at Esperanto and other companies are going to put a lot of die out there!So be prepared to make a really big package!"

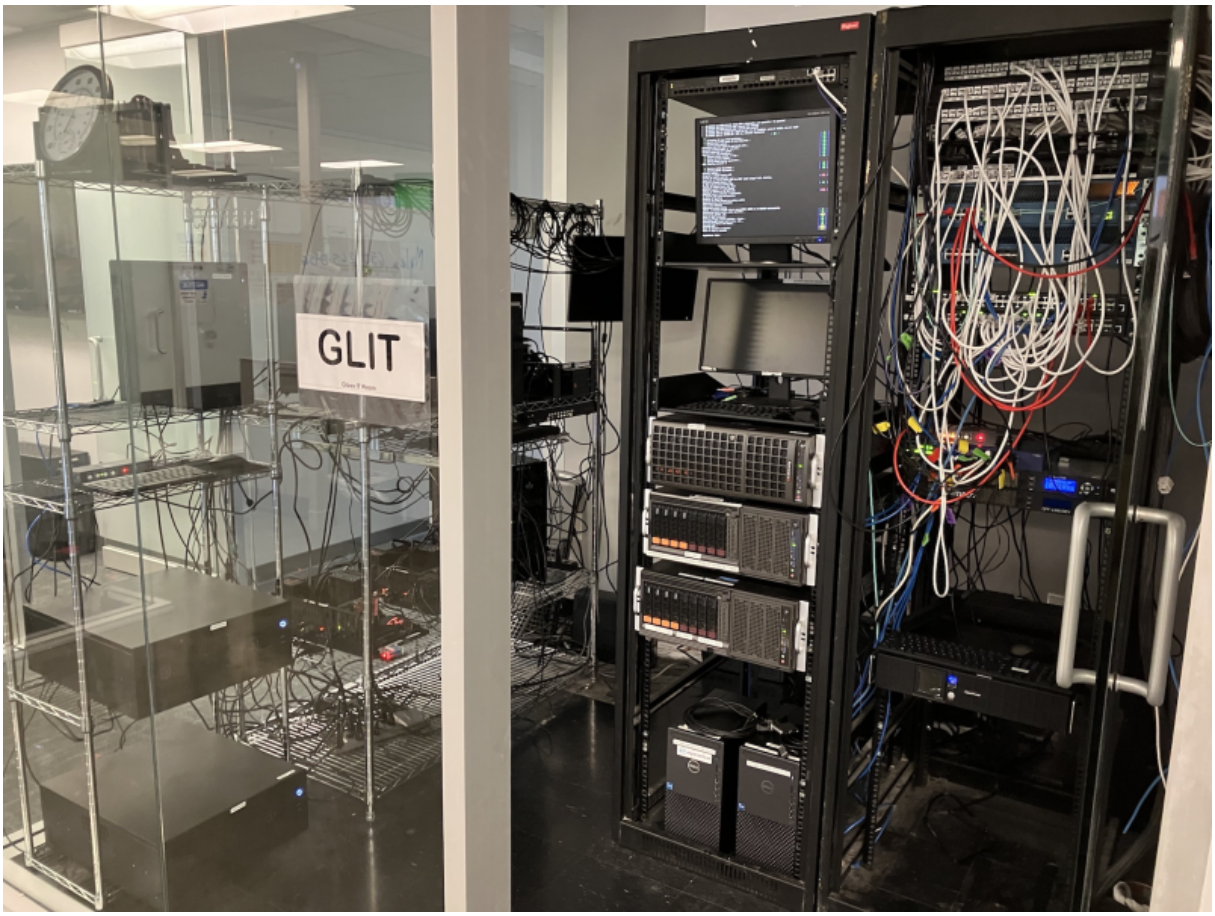
Yasu: That's true.

Dave: It will be from now on. Considering how many chips(packages)you can put in one rack, you can probably put 300 to 1,000 in one rack. 1000 means that 64 million RISC-V cores, which is far more than the number of CPUs in current supercomputers, can be put on one rack of a small company. People will reach that stage in a few years.

[6] “ RISC-V International Ratifies 15 New Specifications, Opening Up New Possibilities for RISC-V Designs ” , Dec. 2021

[7] The movement of open source processors seen through interviews with Wave Computing (MIPS) and Esperanto Technologies (RISC-V) (gihyo.jp NEWS & REPORT 2019.9.9)

Photo 9 Lab space with test server containing ET-SoC-1 card



Moore's law

I'm talking to a visionary named Dave. Let's talk about the future a little further.

happens next? Is it the end of silicon?

Dave: No. Everyone issaying, "This is the end of Moore's LawFor another decade they have been predicting the end of Moore's Law, but it is not over.

Yasu: I agree. So I would like to ask, what will end Moore's Law?

Dave: I think the biggest challenge will be the packaging cost of the chip. When we made Transmeta, we needed a team (or a number of people)to make a single CPU. Espernto allows a slightly smaller team to build a single CPU.

However, we had to assemble 1,000 CPUs (cores) , which made the design extremely complicated. These days, it's said that it costs an average of nearly $250 million to make a chip like ours. This number increases with each generation.

So we may need another law: Dave 's Law of Design Costs . It doubles every cycle. (hahaha)

So as technology advances, it becomes very difficult for people to afford to build a completely new chip that specializes in a certain function. ( ) In that respect, chiplets will help a little, but it will be very difficult for large projects. I think we will be in that situation in 2030, at 1.5 nanometers. Everyone says that we will definitely reach 1.5 nanometers, so we have three more generations left. Arrive very quickly. I doubt the end of Moore's Law will come before 1 nanometer. (difficult) they will find their way.

Craig: Certainly. Everyone thought Fin-FET was the end of the world, but it's done.

complexity of the design and its cost will become a really big problem. Also, in order to pay the cost, the product must be very ubiquitous, such as a mobile phone. Devices like ` `Playstation'' may not produce enough in the future. Even if you come up with something new you want to do, there is no person (organization)with the financial power to actually design it. This is the problem.

Kasahara: I think it would be extremely difficult for start-up companies to tackle this issue.

Dave: It's going to be very expensive. I think this is the challenge for the next 10 years. Moore's Law doesn't stop. But if you make the chip very small, the fab business will be destroyed. If you make one wafer and get 1 million chips, you're done. I think this business and design dilemma is much more fundamental than physical materials issues.

Yasu: You need some kind of breakthrough.

Dave: Well, people talk about high-level design, etc., but it's not a complete solution.

Yasu: Hmm. Now think about what happened on the software side. So open source software development was beneficial in terms of scalability. All software is getting more complex every year, but open development keeps up. The development of such new approaches may occur in the silicon industry.

Dave: Software evolves very quickly. Hardware follows Moore's Law , so it 'sverypredictable.no one would have known. Yesterday in a meeting me and Craig were talking about the roadmap and said that we've been doing

going to happen in the next four years.

Yasu: Hahaha. I agree.

Dave: Too many things change right now. What is popular today will not be popular tomorrow. Otherwise we would be using 5th generation Napster or Friendster. (Everyone laughs)

Time ran out and Dave headed off to his next meeting.

In response to the author's question, "What will stop Moore's Law?" , Dave pointed out that design costs are rising too much as "Dave's Law of Design Costs." I explained that I would not be able to make any. When moving from 32nm to 28nm, it was predicted that the development cost of process technology would rise too much and the industry would become extremely oligopolistic, and that only products (and their companies) that could be shipped in large quantities would benefit from this. . However, the industry just "found a way forward" and solved it, and now many companies and users are benefiting from the 7nm process.

As Dave said, nobody knows how things will go in the future. However , I would like to be aware of "Dave's Law of Design Cost" . It's been a long time since I had a discussion with Dave.

meeting room name COLUMN

One of my favorite things about visiting startups is seeing the name of the conference room. At Esperanto, conference rooms were named after

CRISP (C-language Reduced Instruction Set Processor) was a low-power processor developed experimentally by AT&T for handheld machines, and later became the prototype of the Hobbit processor for PDA called EO by AT&T. Although not a well-known name, this CRISP is also one of the processors designed by Dave.

When I first met Dave for an interview in 2004, I remember hearing about this CRISP from him .

Photo 10 Dave poses in front of CRISP's room

Machine learning/AI •