Code Together Podcast Episode #35
Title: Synthetic DNA & oneAPI :: Revolutionizing the Digital Storage
Industry Guest: Raja Appuswamy
Intel Guest: Sujata Tibrewala

Radhika Sarin (00:05):

Welcome to Code Together, a discussion series exploring the possibilities of cross architecture development with those who live it. I'm your host, Radhika Sarin.

Radhika Sarin (00:16):

The exponential growth of, data generation and its storage, is driving the next phase of innovation. Scientists are now able to store information as molecules of DNA, which are scalable and accurate. With the right set of open source tools and optimizations, synthetic DNA is paving the way for the next generation of advancements. Let's talk to our guests today about DNA data storage and how it's impacting the future, of data storage industry.

Our first guest today is Raja Appuswamy. Raja is an assistant professor in the data science department at EUROCOM, a French research Institute located in the sunny Sophia Antipolis tech valley of Southern France. He is a principal investigator of EU Future and Emerging Technologies project OligoArchive, which focuses on DNA data storage. Welcome Raja, it's a pleasure to have you here.

Raja Appuswamy (01:25):

Thank you, Radhika. It's really great to be here.

Radhika Sarin (01:28):

Our next guest is our very own Sujata Tibrewala. She leads a cross-functional forum of open source leaders from around the community, including Intel. She's also a cherished guest at the Code Together Podcast channel. Welcome Sujata, it's a pleasure to have you back for yet another interesting topic.

Raja Appuswamy (01:51):

Thank you so much, Radhika, pleasure to be here.

Radhika Sarin (01:54):

Great. Let's get started. Raja, can you tell us a little bit more about your project and what problems are you trying to solve with this?

Raja Appuswamy (02:03):

Absolutely. Yeah, so I can summarize the problem in a very simple way. So every single day, especially given that it's summer now, all of us want to go on vacations, all of us want to tour. We all take a lot of photographs when we travel. So a thought exercise that many of us actually can do is suppose I want to pass a photograph that I have taken now, I want to pass it down to people who are two generations after me or three generations after me. So people who are going to be living a hundred years into the

future. If I want to pass down a family heirloom, a photograph down to those people, how would I actually pass it down? I think this is the fundamental question.

Raja Appuswamy (02:43):

Nowadays that we are storing all this data, whether it's photographs, videos, files, we are sorting all this information, all this data in the form of digital data in one way or the other, right? So all these are actually sitting somewhere on a hard disk or sitting somewhere on our computers. And we never actually think about how long is this data going to be with us and what happens when that hard disk or when the computer fails, right? So in a nutshell, the real problem that we are looking at, that we are trying to solve is one that has not been solved till date, and the problem is long-term data storage. So how do we keep data for a very, very long time? So that's exactly the question that we are actually looking at. And in this particular context, we are actually looking at new types of storage media. So not existing storage media like hard disk drives or tape or solid state storage, but we are actually looking at new types of storage media. And one that I'm particularly excited about is synthetic DNA.

Sujata Tibrewala (03:41):

That's exciting, Raja. And as you were talking, I was just thinking that one of the criteria for this kind of storage is that it should survive, it should be robust and maybe even able to store and retrieve data without much power, like small footprint. Otherwise, it will not be able to pass on through generations because what is going on in my head is all those massive data centers of today, which are running Netflix and Amazon Prime and all that with their massive cooling capacity required and the power hungry servers that is required to run them.

Raja Appuswamy (04:20):

Yes, that's actually a very good point, Sujata. In fact, that's one of the key problems that we have today, right? Is the fact that storage actually consumes quite a bit of power. Power is one aspect of it. But in general, storage is actually quite expensive based on the storage devices that we have. So if we look at a simple example of Facebook or Instagram, for example, where people are constantly uploading millions and millions of photographs every minute, Instagram and Facebook have the responsibility to keep these photographs essentially forever, right? I mean, people who have a Facebook account, they would be pretty upset if Facebook went ahead and told them that, "I'm going to delete your photographs five years from now."

Raja Appuswamy (04:57):

So basically this has created a push, especially in the cloud as you mentioned, for companies which operate a cloud scale, particularly Facebook or any of the large hyperscalers, right? It's created a push for them to innovate in terms of long term storage. And one of the key aspects they're looking at, as you mentioned really, is the price point. Now on the price aspect, there is of course the power aspect, which basically is how much power that's the storage consume. And if it consumes more power, you need to bring in cooling facilities, you need to pay your electricity bills, all of that. That's one aspect of it.

Raja Appuswamy (05:29):

There is also the aspect of how long is my storage device going to last. That's also something you touched upon. In fact, of the storage devices that we actually have today, so we can say we have hard disk, solid state storage and tape, these are the most popular storage devices, if you ask the question,

"How long can these popular storage devices last?", it turns out that most storage devices we have almost, all storage devices, the popular ones we have available today do not last longer than let's say 10 or 20 years.

Raja Appuswamy (05:58):

What this really means for hyperscalers is every few years they actually need to change, right? They need to go from one generation of storage to the next generation of storage in order to deal with this kind of media decay, right? So when media fails, they need to replace the media. Now, what's interesting here is that this problem is not something that affects only the hyperscalers. This is something, as I mentioned in the thought experiment, if you think of who else is storing data for a very long time, think of museums. Whenever we go to the Louvre, for example. So I'm based in France. So if we go to the Louvre here in France, most of us actually see paintings and pictures. But today art is increasingly digital, right? So there are a lot of artists around the world who are creating digital art. And all this digital art need to be preserved essentially forever, right? This is part of our collective cultural heritage.

Raja Appuswamy (06:47):

And in many cases, there were also museums which are taking not digital art, but things that were actually analog, so things like paintings that were done 400 years back, 500 years back. A very good example of this is our collaboration with the Danish National Archive. So this is the national archive in Denmark. These guys, they actually have the responsibility of preserving paintings, key culturally significant documents over several years, right? Essentially over their lifetime. One of the key things that we are looking at in this context, for example, is a painting from a very famous king in Denmark, which was done in 1500s or 1600s. So it's been 500 years this painting has been preserved. And now they have a digital version of the painting. So the physical painting has been preserved for 500 years, but now they don't know how to preserve this digital painting longer than 10 years or 20 years, right?

Raja Appuswamy (07:37):

And so we are actually working with museums. We are trying to work together with archives. We are trying to work together with memory institutions, that's what you call them. And this is another class of people who actually, similar to the hyperscalers, they want to preserve data for a very long time, but they don't have the scale. They don't have the ability or essentially the monetary might of the hyperscalers. So they want a solution that can actually work at a much cheaper cost, much more scalable at a smaller footprint that's robust and green, as you mentioned.

Sujata Tibrewala (08:06):

Wow, this is a cross collaboration project working with artists, with museum, with libraries. And of course this is technology, right? So digital data like how to retrieve, how to store. And of course, you're working with synthetic DNA, which is an organic material. This is like a perfect example of what today's modern technology is about innovation. Not innovating in a silo, but innovating cross technology domain, which traditionally we used to think, "Okay, this thing can happen only in the silo of DNA specialists or storage specialists or museums and artists." And it's like, wow.

Raja Appuswamy (08:48):

Indeed. Indeed.

Sujata Tibrewala (08:49):

Can you talk a little bit more about how do you collaborate with all these people? Was there any methodology or something on how do you work with all of these different type of people?

Raja Appuswamy (09:01):

Indeed, that's a very good observation. This is a very interdisciplinary problem, both from the problem point of view, right? Because it comes from all these different domains, but also from the solution point of view. Essentially you need to bring together biologists who actually know how to do wet lab work on the DNA. You need to bring them together with computer scientists, together with chemists who can actually manufacture DNA, together with material science people and people who work on microfluidics and robotics who know how to build machinery that can automate the process of storing and retaining data on DNA. So it's a very, very large scale interdisciplinary initiative. And your question is absolutely relevant here. In fact, no single person can go at this alone, right?

Raja Appuswamy (09:41):

And so you asked about the methodology of how we do it, and this is precisely why the European Union has a class of projects that's called the Future and Emerging Technologies initiative. So European Union offers grants for many different projects. And this particular type of project, which is called the FET, is really targeted for this kind of interdisciplinary work. The goal of FET is to bring together experts in these different disciplines under a single umbrella to solve a major societal challenge. And that's exactly what we are doing here. So there's a clear system in place, a clear method in place for people to collaborate together and the European Union facilitates it. We in 2019 put together one such FET project called OligoArchive, which spans several countries and several groups working on all these different domains. And that's kind of how we do the cross collaboration.

Raja Appuswamy (10:31):

And so throwing back in terms of the cross collaboration, I actually wanted to also ask you because I know you have been on the oneAPI side of things, and now you're on the open source side of things. Also, of course, you've worked with many different teams and many different technologies. So I wanted to ask you both in terms of Intel and from your point of view, is there any specific methodology that you follow in terms of coordinating activities across these different people?

Sujata Tibrewala (10:52):

You know, I have been running the Intel Software Innovator Program. Even before I took up oneAPI as a technology, I was running this Intel Software Innovator Program for NFV and SDN, so networking technology. I can't even count how many open source projects were under that umbrella. And oneAPI itself is a beast. The name is oneAPI, but under the hood, it is a full stack starting from firmware to the orchestration layer, to the applications, and AI/ML, HPC, IOT, et cetera. So it's a beast in itself. And finding people who are working in all of these domains... So because of oneAPI has so many components, it can be used by many practitioners, right? So an innovator who's working in AI/ML can use oneAPI. An innovator who is working with an application which uses SQL for an HPC application can use it. Your application comes in that aspect, but then your application is OneOligo, a synthetic DNA. But then there are many others. NAMD's another very good example, which is like molecular simulation.

Sujata Tibrewala (12:05):

Then there's epistasis detection. That innovator, you're working with professor Alexander, he has been a guest in this podcast before. And then also something completely diametrically opposite like face detection, for example, right? Like foreign banking application. So basically what it boils down to is how that workload can be boiled down to the simple set of instructions, procedures. And that's what oneAPI does, gives a language to optimize these simple set of instructions to do it parallelly. And that's what all of these innovators are doing.

Sujata Tibrewala (12:45):

So to work with all of them, basically I have to understand what their problem is and how we can solve that problem, how as Intel, we can enable them but also not dictating them what to do. Like, just help them use us as tools. So we have an engineering force behind all of this mechanism and also open source community. Don't forget the open source community. So my role was basically just a facilitator, problem solver, connecting people. Just helping. So I think that is how I build the community and work with them. And I think working with such a diverse set of people, only that is the thing that can work. Just be open to listen, help them, then step back and let them do the job.

Raja Appuswamy (13:37):

Indeed. Indeed. Yes. In general, I think that's excellent advice, right? So that's kind of the similar approach that I think that any interdisciplinary issue should follow, right? Because when you look at synthetic DNA storage, or in general any problem that requires expertise from multiple domains, it's extremely unlikely for one person to have the knowledge of everything. And so it's really important for people in their domains to bring their skill set. It's important to establish these kind of interfaces that people can actually have so that they understand what each person is talking about without having to know too much about the individual domain. I think that's one of the key challenges in any interdisciplinary project.

Sujata Tibrewala (14:15):
Yeah. That's another thing, right? Like abstraction.

Raja Appuswamy (14:17):
Indeed.

Sujata Tibrewala (14:18):

So you can know how much you need to know and then let the experts deal with the depth of it. Just let them go. Give them the freedom to work. But then when we are talking about abstraction and standardizing the processes, do you have any standardization when you are working with these different groups? Like if there's some common language that everybody can talk while they're also working on their individual work?

Raja Appuswamy (14:44):

Yeah. That's really another interesting point, right? So broadly speaking, at the storage level you can see there are many standards. So for traditional storage devices, you have the SATA standard or the SAS standard. Now we have NVMe devices. So all these interface standards are there for storage devices. So starting from that point, we don't have a standard for DNA storage yet. That's definitely because DNA storage is very new. So it's still an emerging technology. It's not a fully mature technology compared to

other storage media. So at the DNA storage level, there is no standard yet. But things are moving very fast. So you might be aware that there is an alliance of people who are working on DNA storage, it's called the DNA Storage Alliance, aptly named. And we have many leaders in DNA storage including companies that actually manufacture DNA. Microsoft Research is a part of it. Many, many universities are a part of it. Some traditional storage companies are a part of it.

Raja Appuswamy ([15:36](#)):

And so this alliance is trying to standardize various aspects of DNA storage and from OligoArchive, we are also a part of this DNA storage Alliance. We have some key members in OligoArchive playing a part of this. So the goal is hopefully in the near future, people work on standardizing the interface to DNA as a storage device also, right? So that's at the storage level.

Raja Appuswamy ([15:56](#)):

In terms of, what you mentioned, in terms of interfaces and standards, when it comes to talking to people or when it comes to collaborating with people, I think that's another pretty interesting problem. One of the key aspects of the project also is this, right? So we have all these different components as I told you and what we want to build is a stack. So we actually want to build multiple layers in the DNA storage stack. Each layer will actually have a responsibility. So the bottom most layer, you can think of it as a box or a disk, right? A DNA disk. And inside this DNA disk, whatever is there inside this DNA disk is something that involves chemistry for manufacturing the DNA, biochemistry. And then you actually have sequencing technologies for reading the DNA back.

Raja Appuswamy ([16:38](#)):

Now, anything inside this DNA disk is something that the layers above are not going to see. So we are going to put an abstraction over it. Now, a layer above that, we can actually have methods to actually anchor and decode data on DNA, right? So I'll get into details of all this, but essentially what I'm trying to say here is one of the key objectives of this project, OligoArchive, is to identify how many layers should we have, what are those layers, and what should the interfaces be. And once we identify this, we can actually assign different people to different layers, right? And because we have standard interfaces in these layers, people can actually work within their layer while making sure that everybody's work is going to be compatible. So that's something that we are actually doing.

Sujata Tibrewala ([17:22](#)):

Awesome. So I just wanted to understand coming from networking, for example, and also at oneAPI and CC++ standards, cetera, there's general push in the industry today going towards open source as de facto standards versus actual standardization because standardization has a notorious reputation of being very slow. What is your thoughts on that?

Raja Appuswamy ([17:50](#)):

I think open source really promotes innovation, right? We have seen a major shift especially in the last decade. So one of the other areas that I work on is data management. So I work on databases and data analytics. You can really see most popular database solutions right now these days, many of them are open source in one way or the other. It really will spur innovation. From the point of view of DNA storage particularly, there are many components that we are actually making open source. In fact, the European Union requires that whatever work that we actually do in the project, because we do have a

startup in the project which works on manufacturing DNA, there are certain aspects of the work that can be critical for their growth, the startup growth. So those aspects are protected. But otherwise, anything that's not really critical for the project, we have to make open source and we have to make all the data and the code available to the public.

Raja Appuswamy (18:41):

So this is something that's a part of European Union. I think European projects are fantastic in this aspect. They really try to push the researchers as far as possible to make sure that things are open, things are reproducible. And so I think it's a very, very good thing in general.

Sujata Tibrewala (18:56):

Wow. Yeah, that's really awesome. So earlier I was advocating for open source and oneAPI or networking or NFE, but now in my new role, advocating for open source across the board, and maybe this is something that if audience is not aware, Intel is contributing to more than 800 plus projects. In many of these open source projects, we are leading contributors. Linux is one good example. oneAPI is another example as well. But then my point is that across the industry, across the software stack, we are promoting open source. One of the reasons like you said, Raja, is propagating innovation, letting community do, because even if we have a huge set of engineers, even then we are just one company. And our vision can sometimes be restricted, but when we put it out in the open, the kind of innovation that the community can bring is awesome and unparalleled. We can never even imagine sometimes.

Sujata Tibrewala (20:01):

For example, Raja, you are working on synthetic DNA storage. I mean, from mental, we don't know if we could have come up with that.

Raja Appuswamy (20:08):

Indeed. Yes.

Sujata Tibrewala (20:09):

So that is innovation. I mean, again, coming back to my previous point, we are just tools. It's like what the community does with what we are developing. We are just giving the industry tools. So in that spirit, we want to make everything open source.

Sujata Tibrewala (20:26):

So anyway, that brings me to the next question. We have been talking about synthetic DNA and I'm dying to hear what actually it is. So are we ready to dive a little bit deeper into what it is?

Raja Appuswamy (20:38):

Yes, absolutely. So basically the DNA that I am talking about when I say synthetic DNA is the exact same DNA that we all have in our bodies, right? So pretty much every living thing. As we all know, DNA is kind of the hereditary source of information. So the information passes down from one generation to another through DNA. So DNA, for Biology 101, it's a macromolecule. What's called a macromolecule? It's a molecule that's composed of sub molecules. The sub molecules are really four types of sub molecules. So we have adenine, cytosine, guanine and thymine. So for all practical purposes, for this

discussion, all we can think of is DNA is basically a long stretch of sub molecules that are ACGT, right? So in terms of viewing DNA, we can actually view it as a string to simplify things. And it's a long string of ACGT.

Raja Appuswamy (21:27):

Now, the human DNA is about 3 billion bases long or 3 billion of these characters, right? So ACGT. When we refer to synthetic DNA, what we refer to is not the human DNA or not any biological DNA. Due to advances over the last few decades in chemistry, we are now able to manufacture DNA outside using well-established biochemical technique. And there are several companies that are actually working on manufacturing DNA. And manufacturing DNA is a very, very key application for many biological applications. So we actually need DNA as a source for a variety of different applications. So we are now able to manufacture DNA outside the body, essentially. And this is what we refer to as synthetic DNA, right? To differentiate it from biological DNA.

Raja Appuswamy (22:14):

Traditionally, the biological DNA is double stranded helix. So it's well famous, the Watson-Crick double helix structure of DNA. The synthetic DNA does not have to be a double stranded helix typically. So normally when we talk about synthetic DNA, we're referring to one single strand of DNA, right? And that's what we refer to as the synthetic DNA that's relevant to us for the point of view of DNA storage.

Raja Appuswamy (22:36):

Now, the first question that comes to mind is why are we talking about DNA? Why do we want to store data on DNA? There are some key properties of DNA that make it very relevant for data storage. The first property is basically the fact that it's very, very durable, right? So there is a work from professor George Church at Harvard and his group where they actually extracted the DNA from an extinct animal species called the woolly mammoth that lived about 5,000 years ago in Siberia. So they went to Siberia, dug up the fossils, extracted the DNA. They're basically using the DNA in combination with variety of other technologies, particularly gene editing technologies, to try to see if they can bring the woolly mammoth back to life, right? And there is in fact a company that has been established that is actually trying to do what's called de-extinction, right?

Raja Appuswamy (23:23):

So the reason why I'm giving you the example, is DNA can last several thousands of years in even such a damaging environment, right? It's not really an ideal environment to store the DNA in. And even there, it actually lasts quite long. Now, if you can store DNA without any water in a profoundly protected environment, in a neutral environment, it can actually survive. It can last for many, many, many generations. Many, many, many millennia. This is the first reason why we want to store data on DNA. Because if you remember what I said in the beginning of the podcast, all the storage devices we have available today, they last about 10 or 20 years, right? And DNA can last thousands of years.

Raja Appuswamy (24:02):

The second reason why we want to store data on DNA is because it's incredibly dense. So looking at DNA, the amount of data that you can store in DNA and if I compare it to the amount of data that you can store today on tape, DNA is about seven orders of magnitude. 10 to the seven times denser than any future projections of tape that we have available today.

Sujata Tibrewala (24:23):

Wow.

Raja Appuswamy (24:23):

Yes. And this is based on a study by the Semiconductor Synthetic Biology Consortium. So it's a very dense three-dimensional storage media. So that's the second reason. The third reason is eternal relevance. As long as we live on this earth, we always have the need to sequence DNA. Whether it's actually for health reasons, right? Because today there's a lot of medical applications are being driven by analyzing a patient's or a human's genome. Especially for example if you look at precision oncology and cancer, or whether it's for other reasons. Whatever other reasons may be, right? We always have the need to read DNA, and we'll always have the ability to read DNA now that we have sequencing technologies, right? We have the ability to read DNA always, which means that DNA really has this internal relevance, right? So you will never be able to not read DNA in the future.

Raja Appuswamy (25:09):

And this is another key aspect. So think of whatever data that you might have stored on a floppy disk 20 years ago. Are you able to read it anymore? That's the question, right? I mean, for most devices, this is not possible anymore. Because even if the data stays on the device and even if the device is not decayed, usually the reader that's required for reading data from the device has gone extinct, and you can no longer read the data anymore, right? And DNA does not have this problem. So these are the reasons why we want to look at synthetic DNA.

Sujata Tibrewala (25:38):

Can we talk a little bit more about why we are talking right now? Our connection. I know you've used oneAPI for DNA. So how has oneAPI helped you in your project?

Raja Appuswamy (25:51):

I think that's a very unusual link, right?

Sujata Tibrewala (25:54):

Yes.

Raja Appuswamy (25:54):

Because when you talk about DNA storage... I mean, on one hand we have, yes, DNA storage. On the other hand, we have oneAPI, which is kind of this cross industry initiative to enable heterogeneous spiral programming, to simplify it, right? I'm really grossly simplifying it. But what's the link between the two? One is on storage. One is on processing. It turns out the link between the two that we found was one of the key stages in actually processing data stored on DNA. So just to give you a quick overview, we stored data on DNA. We're converting binary data, 0s and 1s, into a sequence of nucleotide SCGD. So think of a very simple mapping. It's not just simple mapping, it's more complicated. But for the practical purposes, think of we map, 00 to A, 01 to C, 10 to G, and 11 to T, right? It's much more complicated than this in practice. But once we do this mapping from binary to DNA, we can now manufacture DNA. So whatever data we have converted into DNA, we can go ahead and manufacture it.

Raja Appuswamy ([26:48](#)):

Now, when we need to read the data back from DNA, we sequence the DNA that we have manufactured. Now, the sequencing is very similar to the way, COVID for example, when you're testing in terms of RTPCR, right? So take a small sample from you. They amplify the sample using PCR, polymerase chain reaction, and then they sequence it to see if there is actually a COVID strain there. It's the same sequencing technology that we use to read data back from DNA.

Raja Appuswamy ([27:12](#)):

The thing is, when you read the data back from DNA... So you might have stored, let's say data on 10 DNA strands or 100 DNA strands, when you read the data back, you don't get what you store. What you get is a noisy copy of the original. So for every strand of DNA, you might get one copy from the sequencing, or you might actually get multiple copies. So it turns out that the key problem here that we want to actually look at is how do you recover the original data back from these noisy copies? And here it turns out we can actually model this as a database join problem. And here is where we actually came ahead with oneAPI, because we wanted to parallelized this database join across multiple processors. And oneAPI helped us there.

Sujata Tibrewala ([27:56](#)):

Yeah, that's awesome. Like you said, this is an unlikely link. So thank you for giving us that explanation on how digital to nucleotide conversion you're doing, and that processing was speeded up by oneAPI. So for the listeners who are not familiar with oneAPI, like Raja earlier said, it is a set of language and libraries which enables you to write program once and then be able to run it on CPUs, GPUs, FPGAs or some accelerators. This is based on something called sickle, which has been in the industry for quite a long time. And we adopted this as the basis for oneAPI tool chest. We, meaning Intel. And then there are other key players.

Sujata Tibrewala ([28:43](#)):

So this programs and library of oneAPI is being used at many technology verticals. So it's being used for HPC, for AM/ML, for vision and for rendering, for IOT applications, and also in FPG. So a lot of FPG applications are making use of oneAPI right now. It's a really interesting set of applications and libraries, and it's awesome to see the industry making use of that. And OneOligo is a good example of that. So Raja, can you focus a little bit more on how you got the acceleration with oneAPI and how do you see it playing a role in growing adoption of analytics, AI and machine learning in your case?

Raja Appuswamy ([29:34](#)):

Yeah, sure. So in terms of the acceleration with oneAPI as I mentioned, one of the things that my group focuses on is database acceleration. So we modeled that particular problem that we had on DNA storage as a database join problem. And we actually implemented a database join using data parallel C++. We used the oneAPI toolkit for that. And by doing so, we were able to run the program on CPUs, GPUs, and potentially even FPGAs that we are actually looking at right now, right? Interestingly here, the fact that we chose to model it as a database problem, it actually shows the fact that oneAPI is applicable in a much broader scope, right? And I think that's my other area of focus, which is actually data analytics acceleration. And that's what we are actually looking at in terms of oneAPI.

Sujata Tibrewala ([30:18](#)):

Oh, yeah. That's awesome. I mean, yeah, oneAPI is a tool and it can be used. Yeah. And you're also talking about XJoin and oneAPI acceleration?

Raja Appuswamy ([30:28](#)):

Yeah. So XJoin is the solution where we model the recovery from DNA storage as a join problem and we actually call the solution XJoin there.

Sujata Tibrewala ([30:36](#)):

Oh, okay. That's awesome. And I cannot resist here because I know that we are working on something really, really very exciting and put together another group of disparate people like RISC-V. So oneAPI RISC-V. Do you want to touch up on that? I know we will probably do a full podcast on that, but maybe give a teaser to our listeners today.

Raja Appuswamy ([30:58](#)):

Yes, absolutely. Yeah. I mean, this is something that I'm very excited about and I hope to come back sometime in the near future to talk about it. But essentially, one of the things that we are looking at both in context of DNA storage and beyond is how do we design a hardware accelerator for accelerating certain key application verticals, right? So in terms of DNA storage, as I mentioned, obviously we want to accelerate the data retrieval from DNA. But as I mentioned, the problem, if we look at it in the point of view of data analytics, you can already imagine the market for hardware acceleration of data analytics and machine learning, right? And one of the key players here, one of the key upcoming standards here is RISC-V, which I'm pretty sure everybody knows about RISC-V by now. It's open standard ISA, which actually gives hardware designers the freedom to come up with custom extensions to add the ISA level. So you can actually come up with new instructions that can actually accelerate various application verticals.

Raja Appuswamy ([31:52](#)):

And because it's actually an open standard, you are able to basically build off of an existing tool chain and build your own hardware accelerator using RISC-V, which we think is a very, very powerful idea, right? And RISC-V originated in Berkeley as many other incredibly amazing architecture innovations. And what we are trying to do in the group, and this is still very new and that's why I mentioned I'm going to be really excited to come back again, what we are trying to do is basically bring together RISC-V and oneAPI. oneAPI as the software standard for parallel programming and RISC-V as the hardware standard for developing parallel accelerators.

Raja Appuswamy ([32:28](#)):

So the goal is to be able to build a customized accelerator with RISC-V and to be able to program it using a general purpose parallel programming framework like oneAPI, right? And I think together these two, oneAPI and RISC-V, are really going to really propel a massive innovation, both in machine learning but also in unconventional domains like DNA storage where we will be able to put together really customized hardware accelerators and program them very effectively. And that's exactly what we are trying to do in the group in terms of oneAPI and RISC-V.

Sujata Tibrewala ([33:01](#)):

Synthetic DNA & oneAPI :: Revolutionizing the Digital Storage

Yeah. This is something like it was just a matter of great timing because like you said, oneAPI is open source acceleration in software, right? And RISC-V is open source and open standards for hardware. We were looking on expanding oneAPI's reach to RISC-V. And just at that time, at the right time, you also approached us on that. It was awesome to connect you with other innovators in the community. And I know you have brought in some people into that collaboration group as well. So yeah, I just can't wait to see what this panel comes up with.

Raja Appuswamy (33:41):

Indeed. Yes.

Sujata Tibrewala (33:43):

Lot of exciting things to look forward to.

Radhika Sarin (33:45):

All right. Thank you. This has been very exciting. Obviously, lots to talk about. But Raja, are there any other resources you'd like to share with our audience?

Raja Appuswamy (33:55):

Absolutely. Yes. There's plenty of stuff that I couldn't share in detail because of the time constraint. I would like to share the official OligoArchive website. So it's oligoarchive.eu, where we have plenty of information about the project, about DNA storage. Also on Intel DevMesh, most of the work that we have done on oneAPI is publicly available for people to go ahead and play around with. So I think these are the two main resources that I would like to share.

Radhika Sarin (34:17):

That's great. Sujata, what resources can you recommend from Intel?

Sujata Tibrewala (34:22):

So oneAPI definitely. They can look at oneapi.com. And then at Intel Software Developer Zone they will find oneAPI's technology. So there's loads of information there, like many tutorials, et cetera. On the industry initiative side, it is oneapi.com. And at an overall level, if you want to look at what Intel is doing in open source in general, beyond oneAPI, you can look at openintel.com. And that is the website where we talk about all the different open source projects that we are collaborating on in the industry.

Radhika Sarin (34:59):

Perfect. This has been a great conversation and I want to thank you both, Raja and Sujata, for this amazing discussion.

Raja Appuswamy (35:08):

Thank you very much, Radhika. It was really a pleasure to be here. And thank you Sujata for all the exciting questions, right? And I hope to be able to come back sometime soon to share more details about our work.

Sujata Tibrewala (35:18):

Yeah. Thank you so much, Raja, for working with us. You've been, in the true sense of the word, innovator. And thank you Radhika for having me.

Radhika Sarin ([35:27](#)):

Oh, it was a pleasure. And thank you for taking the time. And a big thank you to all of our listeners for joining us today. Let's continue the conversation at oneapi.com.

PODCAST # 35

Synthetic DNA & oneAPI :: Revolutionizing the Digital Storage

Abstract

The exponential growth of data generation and the need for it's storage is driving the next phase of innovation in data storage. Scientists are now able to store information as molecules of DNA, which both scalable and accurate. With the right set of open-source tools and optimizations, synthetic DNA is paving the way for the next generation of advancements

Industry Guest: Raja Appuswamy, Assistant Professor, EURECOM - Data Science Department
https://www.linkedin.com/in/raja-appuswamy-b418425/
https://twitter.com/raja_appuswamy

Intel Guest: Sujata Tibrewala, Sr. Manager Open Source Alliance
https://www.linkedin.com/in/sujata-tibrewala-a531198/
https://twitter.com/sujatatibre

Resources

https://oligoarchive.eu/

https://www.intel.com/content/www/us/en/developer/tools/oneapi/overview.html

https://www.oneapi.io/

https://www.intel.com/content/www/ca/en/developer/overview.html

https://www.intel.com/content/www/us/en/developer/articles/success-story/dna-storage-heterogeneous-architectures-oneapi.html

https://www.intel.com/content/www/us/en/developer/videos/oneoligo-use-oneapi-to-accelerate-dna-data-storage.html

https://devmesh.intel.com/projects/oneoligo-860c37

https://devmesh.intel.com/users/raja-appuswamy

https://riscv.org/

Keywords

Synthetic DNA

Digital Storage

oneAPI

SYCL

Parallel Programming

Hardware Accelerators

RISC-V

Data Storage

Cross Architecture

Open-Source

Synthetic DNA & oneAPI :: Revolutionizing the Digital Storage

Molecular Biology

Hyperscalers

Hashtags
#oneAPI
#DNAstorage
#ParallelProgramming
#Opensource
#DigitalStorage
#HeterogeneousProgramming
#CrossArchitecture
#RISCV
#SYCL
#Developers
#OpenSource
#HardwareArchitecture