Episode 20: Blazing a Trail in DPC++
Host: Nicole Huesman, Intel
Guests: Aleksander Ilic & Rafael Torres Campos, INESC-ID & IST

Nicole Huesman: Welcome to *Code Together*, an interview series exploring the possibilities of cross-architecture development with those at the forefront. I'm your host, Nicole Huesman.

We've often talked about the challenge that developers face in programming for heterogeneous computing environments as data-centric workloads become ever more pervasive. Today, we'll talk with colleagues at INESC-ID in Lisbon, Portugal, about their experience in building a cross-architecture application focused on disease detection and prevention.

Aleksandar Ilic is an Assistant Professor with the Department of Electrical and Computer Engineering at the IST University of Lisbon, and a Senior Researcher in the HPC Architectures & Systems Group at INESC-ID. He's currently focused on research primarily in high-performance and energy-efficient computing, highly heterogenous systems, parallel applications and computer architectures. Great to have you with us, Aleksander!

Aleksander Ilic: Thanks Nicole. I'm also happy to be here.

Nicole Huesman: Rafael Torres Campos also joins us. Rafael works as a Researcher at INESC-ID as a part of the High Performance Computing Architectures & Systems group, after earning his Masters of Science degree in Electrical and Computer Engineering at IST University of Lisbon. His interests include performance modeling of heterogeneous systems and GPUs. Thanks so much for being here with us, Rafael!

Rafael Torres Campos: Hi Nicole. Thank you for having me.

Nicole Huesman: I'd also like to welcome Sujata Tibrewala, Intel's oneAPI Developer Community Manager, to the program!

Sujata Tibrewala: Thank you, Nicole. Glad to be back!

Nicole Huesman: Aleksander, Rafael—can you tell us a little bit about the work you're doing at INESC-ID, and what inspires you about this work?

Aleksander: Sure, Nicole! So, INESC-ID and IST are the main research and education centers in Portugal. Here we've actually gathered up a great team of researchers that glean the topics from many different scientific areas where we have a strong focus on the ones with high impact to the society.

So, in our particular case, in the High Performance Computing Architectures and Systems Group, we are working for many years on efficient solutions for high performance and energy-efficient computing in highly heterogeneous environments. For example, the ones that are combining multi-core CPUs, GPUs, FPGAs, and very recently, even TPU tensor processing units. So, this research actually covers many different tasks, such as application optimization, parallelization, programmability—that's scheduling, modeling and so on.

So recently, we came to an idea to actually incorporate these expert eyes that have all this knowledge and tackle one important application in bioinformatics, which is epistasis detection.

However, detecting epistasis is a challenging problem because it may require a huge amount of time to actually process the data side that is given to us. And these data sets can actually have hundreds of thousands of data from the patients. So, basically, what is this epistasis detection and why it is so important is, in a nutshell, those clever folks from bioinformatics, they actually discovered that even very small changes in your gene structure can have significant impact to the observable trait. In very simple terms, we can think of it that observable trait is a kind of a disease. So just the small change can actually mean that you can have a certain disease or not have a certain disease. So, when we look to this one, actually the small changes can occur even at the level of a single nucleotide, and this is called single nucleotide polymorphism, or SNPs. So, we work at the level of those SNPs. I'm not going to actually shoot you with a lot of technical information, but this is just to give you a high-level overview.

So, the problem here is that for some complex traits or diseases, multiple SNPs can interact with one another, and then the problem comes because figuring out which ones interact and how many of them actually interact is far from being a trivial task. For example, for some diseases such as Alzheimer's, diabetes or breast cancer, it has already been proven that there is some SNPs that are interacting between each other, and from there, you can actually evaluate if, at a certain age, you're going to get this disease or not. However, there are many of those who are yet to be discovered, especially the complex and more problematic diseases that you have nowadays. So for example, by knowing this information, the patient data can be processed at early stages and you can evaluate its potential to get the diseases at later stages, like these, for example, we can do early diagnostics or we can do disease prevention or even in the future, make it part of personalized medicine.

So, for this part of our investigation and for many different diseases for which we have yet to find those interactions, our job was actually to investigate this novel, high-performance and energy-efficient algorithms in order to do epistasis detection as precisely as possible. And this research is conducted within our HiPErBio project, which is a national Portuguese project. And it involves more than a dozen of team members from INESC-ID and some clever folks from Spain as well, and they include Rafael Campos, one of the main developers in this project. Some developments were recently are also conducted in close collaboration with the Intel oneAPI team and they're greatly supported, endlessly, like Sujata who you also have today here.

So, our current research efforts in the epistasis detection—due to Sujata's support, in fact—are mainly focused on state-of-the-art Intel CPU plus GPU platforms and they actually now involve a wide variety of oneAPI tools. So, I believe that now Rafael, since he's one of the main developers on this topic, can give us more insights of these developments and what you're actually doing and with your team.

Rafael: Yeah, sure. Well, I can start by giving an overview of the work we have been developing in this area. So, I mean, in the beginning, we used to have this applications developed mostly for CPU, and therefore we have the challenge because we wanted to adapt this to the integrated GPUs that you see on most of these system on a chip (SoC) of Intel. We had this small GPU there that was not being used, and the challenge was to try to make use of this GPU, and later, to try to make use of the CPU and the GPU at the same time. So, the interesting part of these GPUs is that they actually share parts of the memory subsystem with the CPU, and this way we make use of this architectures to system to obtain a more efficient heterogeneous implementation of our bioinformatics applications.

So, this had some challenges, obviously. We had to spread work across these different devices and load balancing is a great example of something that a lot of times goes really wrong and we have to really focus to get it right. We have tried to explore the performance benefits, and we have been able to develop some frameworks to make use of the best performance of each device. And also, we explore the power and the energy efficiency when we perform the epistasis detection. We have also published this mainly last year at the Euro-Par 2020 conference. We have published some work using integrated GPUs and CPUs by Intel to obtain efficient epistasis detection. And lately with the collaboration we've been having thanks to Sujata, we have also been able to test this newest GPUs, you know, that Intel has been releasing, we've been using the oneAPI DevCloud, and we've been accessing some GPUs like the Intel® Iris® X^e MAX. And we have been focusing our development and our research in those areas. And we expect to have great results in terms of exploring these devices to obtain efficient and high-performance epistasis detection.

Sujata: Thank you so much, Rafael, for that overview. I would like to mention that Rafael is one of the winners in the oneAPI contest, you know, the Great Cross-Architecture Challenge that we held recently. So, what I would like either Professor Aleksander or Rafael to touch upon is how particularly was your experience in using oneAPI and DPC++ and what actually that brought into your project specifically?

Rafael: Well, it was really a big happy surprise to actually get that award in this Cross-Architecture Challenge. Everything we do ends up being a consequence of little developments over time, right? And it's good when they amount to this kind of recognition. Actually, before DPC++, we began by using OpenCL to target these GPUs, and starting from the C++ applications, we implemented OpenCL, and since OpenCL is a very well-established programming model, we were able to attain really high-performance values in these applications. And eventually, we ended up moving to exploring this other programming models like OpenMP. And finally, we landed in DPC++, which ended up being the most interesting alternative we found. And immediately, we could see some benefits comparing to OpenCL.

For example, in DPC++, we're able to have every code—device codes and the host codes—all in the same file. In that case, for example, when we are developing in a team switching files back and forth, we immediately halve the number of files that we have available, and it makes it a lot easier. There are other advantages, of course. We can also use unified shared memory, for example, which is not available in OpenCL. That's an advantage that we hope to explore more as we are starting this development phase in DPC++. Especially for heterogeneous applications, we think that sharing the common data structure for CPU and GPU and even incorporating other devices in this frameworks, like FPGAs, we should be able to get big advantages from using oneAPI and DPC++ in this aspect. And that's really, we're still trying to figure some things out about the language. And we, I think got some good results, for example, as Sujata said in the award-winning submission, you may call it, we ported some designs we had for OpenCL into Data Parallel C++, and we compare them to this highly optimized C++ implementation for CPU and were able to obtain speedups even with the Gen9 integrated GPU that has 24 execution units and the speedups for comparing to the CPU. And even when we tested in these new Intel® Iris® X^e MAX, we were able to obtain speedups of six times, and it was quite powerful compared to the CPU. Moreover, oneAPI has also integrated—in the base toolkit—these analysis tools that we often use in our work. [I'm] talking about Intel® Advisor and Intel® VTune™. Those are a great help because the applications allowed us to immediately see what is bottlenecking our execution, and where we can improve. And I think they are also great help for anyone that's looking to optimize their GPU applications.

Aleksander: As of yesterday, I think you also have a DPC++ version that surpasses the OpenCL, right?

Rafael: Oh yes. Everything is very recent. And sometimes we get these amazing results. It takes a bit of time to process them, we are a bit slower than the machines! And we are looking forward to using Data Parallel C++ in the future, and we have some really positive results.

Sujata: So, one of the things I wanted to understand from you is, you mentioned all the good results that you have gotten, like, you know, moving to Data Parallel C++ from OpenCL. Were there any challenges or anything that your listeners should look out for?

Aleksander: Yeah, I can jump in here, and then Rafael can go technical in there! You know, it's like, I know that we really pushed Rafael here in the team. Okay, let's go and make the DPC++, because as you know, also, you know, with these collaborations, we do many tutorials together and things like that. And then we really wanted to port OpenCL to DPC++, and then Rafael started doing it, and you know, for several weeks he does everything as he should, he writes the same code, but he gets very poor performance. You know, we were like, c'mon, let's do it, let's do it. And at the end, I think it was a very simple tweak, like he needed to move two lines above, and then he gets a little bit better performance. So, this is probably the part where users need to get accustomed, you know, to a little bit newer frameworks and newer languages. Right, Rafael?!

Rafael: Yeah, so there have been a few results that were quite, well, different during the development of this application in DPC++. For example, I was used to using OpenCL, and one thing that many developers using OpenCL may know is that the execution queues where you send the work to in OpenCL, they're always ordered, so, whatever you send there first is the first thing to come out. That is not the case for DPC++, which led to some really weird afternoon, where I was just getting just the most bizarre results and I could not figure out why everything was coming out wrong. And in the end, I ended up learning that DPC++ queues were not ordered. The run time just chooses what is best for execution first and does it. And that was interesting to learn. I ended up dealing with that quite easily once I learned what the mistake was. And well, there's more stories obviously. Now there's a lot of documentation and even this new book that came out for DPC++. I think, take it slowly if you're a developer in this, and try to figure out what you're doing, try to check the sources, the Wiki, and all the examples you have online. And eventually you're going to get it!

Sujata: Thank you, Rafael, for that valuable feedback.

Rafael: Nah, thank you for all the help. I mean, every time we had an obstacle and we reached out for help to Sujata's team, immediately we had great help and great feedback, and that really, really helped us a lot in our research and all the problems we had.

Sujata: Yeah, of course, we look forward to working with you guys. You guys are doing such great work. I'm always happy to help you. And I hope you continue to do this work and get more and more excellent results.

Aleksander: Thank you!

Nicole: Such a great discussion. And I think, you know, our listeners are going to be able to learn so much through the experiences that you're sharing here, both Aleksander and Raphael. So, thank you so much for being with us today.

As we wrap up, Aleksander, where can listeners go to learn more?

Aleksander: Typically, the best place to get to know a little bit more about research is actually to take a look at our publications such as our Euro-Par 2020 paper, as Rafael mentioned, and the others can be found on the INESC-ID website. So, of course, you can also check our very recent talks on the oneAPI Dev Summit, like in the last year, and …

Sujata: … Yeah, I was going to mention that you guys are doing a hands-on tutorial [at oneAPI Dev Summit 2021], actually. So, you know, developers will be able to get a deep dive on your work and also get to know about your experiences and they can ask whatever questions they may have. Yeah.

Nicole: Excellent. And Rafael, any pointers you'd like to leave folks with?

Rafael: Well, if I can point anyone to look at our research, as Professor Aleksander said, INESC-ID's websites, and also we have this DevMesh project that we posted on Intel DevMesh. This was actually the first thing that led us to this collaboration with Sujata. We posted it there basically because I needed to have my DevCloud account renewed, and this collaboration just started, they noticed that the project was there. So, it's devmesh.intel.com/project/igpu-epistasis.

Nicole: Excellent. We are so looking forward to continued collaboration, looking forward to seeing what you do in the future and having you back on the program. Rafael, it's been great to have you here. Thanks so much for sharing your deep technical insights.

Rafael: Thank you so much, Nicole. It's been an absolute pleasure.

Nicole: And Aleksander, always so wonderful to speak with you. Thanks so much for being here.

Aleksander: Thanks, Nicole. It's been a great pleasure to join the discussion today. Thanks also to Sujata as well for all the help.

Nicole: Absolutely. Sujata, great to have you on the program and we're going to love having you back.

Sujata: Thank you so much, Nicole, for inviting me. Thank you, Aleksander and Rafael, for all the wonderful work that you've been doing.

Nicole: And a big thank you to all of our listeners for joining us. Let's continue the conversation at oneapi.com.