Episode 11: Optimizing HPC Clusters: Meeting the Challenges of Today and Tomorrow
Host: Nicole Huesman, Intel
Guests: Mike Lee, Intel; Nico Mittenzwey, Megware Computer

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

High performance computing, or HPC, has grown far beyond its traditional users, permeating a myriad of disciplines, like AI, machine learning, enterprises and the cloud. With the ubiquity of HPC comes an insatiable demand for compute across heterogeneous environments to solve some of the world's biggest challenges. In this complex landscape, the ability to ensure HPC clusters and applications are tuned to deliver optimum efficiency is key. The full stack—spanning hardware and software—matters.

Two guests are with us today to share their real-world insights into tuning HPC clusters. Nico Mittenzwey, senior HPC engineer at Megware Computer, who leads the benchmarking and training team at the company. Thanks for joining us, Nico.

Nico: Hi Nicole. Thanks for having me.

Nicole: And Mike Lee, technical, enterprise and cloud segment marketing manager at Intel. Great to have you on the program, Mike.

Mike: Thanks, Nicole. Good to be here.

Nicole: We're excited to hear about Nico's experiences at Megware and some of the challenges he's seeing among end users. Mike, Nico, take it away.

Mike: Thanks, Nicole. Hey, Nico, good to have you here today, talking on a topic that I know we both love and have been working on for the past number of years. I wanted to kind of talk about tuning for a cluster, what you've seen as concerns from your customers, you know, moving into this exascale world. Exascale is right on the horizon. What do you see that are things that are most important for users that you guys work with in terms of MPI and tuning for a cluster and optimization and so forth?

Nico: Yeah, thanks Mike. So, what we see with the exascale clusters on the horizon, especially fault tolerance becomes mission critical to those projects, and all our customers, they request this even in the current pre-exascale projects. So, this is a very important feature we need to have in any MPI implementation. Of course, also a very important feature is that the user experience or the usability of the usage of an MPI implementation needs to be there, so users are able to start their jobs without any problems without having to tune their applications by hand so they can focus on their science and run the applications without any problem.

Mike: Do you see more of your customers being more along the lines of end users, more so than they are a classical HPC user, which is, you know, somebody who is both a domain specialist and somebody who had to dip their toe into the world of HPC to get their work done.

Are you seeing more domain specialists without the experience, so therefore, like you said, they just want to get the work done. Are you seeing more of that?

Nico: Totally. I would say 80% of our customers are universities. And within those universities, again, about 80 percent are not computer scientists and they want to get their projects done with these, and they use clusters as a tool and so, they also use MPI as a tool and this has to work and it has to work perfectly and it has to be tuned to the specific hardware of the cluster and do this automatically. But of course, we also have customers who are very into the code and try to optimize their applications and even try to optimize the MPI implementations. So, there are both worlds, but the majority definitely are users and domain experts.

Mike: The Intel MPI library is one of the top three libraries that are used worldwide for MPI implementations and applications. The Intel architecture x86 basically plays a critical role in so many clusters around the world.

With that our MPI implementation is based on the MPICH implementation of Argonne National Lab. And, you know, basically we deliver that to our end customers and, you know, it scales really well, performs very well, and therefore benefits customers like you're talking about, Nico. Can you basically help us describe Megware's experience and your experiences with, you know, how the MPI library, in this case, the Intel MPI library really, you know, helps solve customer challenges and meets their demands and requirements?

Nico: Sure. So, one of our customers is the German Federal Waterways Engineering and Research Institute, and they use a lot of OpenFOAM simulations to simulate the waterways. And when they bought a cluster from us, they were astonished that they can just type MPI run and the number of processes ("mpirun-np <number of processes>)"), and then their simulations run perfectly on the cluster because before they got our cluster, they had another one from a competitor and there, they had to tune very, very many specific options of the MPI run command to get the best performance. So, this is definitely something that helps us in the market of those customers.

And of course, we are also using Intel MPI to test all of our clusters [to determine] if they can get the best performance or if they can get the performance we promised to our customers. And for example, just two month ago, we deployed a cluster of a roundabout 600 water cooled Cascade Lake AP systems. And those Cascade Lake AP systems have 96 cores per node. And in total they delivered 3 PFLOPs in the High Performance Linpack benchmark, and this system made it to the 92nd place in the Top500 where we are very proud of.

Mike: That's really great. Nico, have you guys given a try to the latest auto tuner? For those in our audience, if you haven't given the auto tuner a try, it is basically a utility that we've included with the Intel MPI library for quite some time, and it's been improved over time. And the auto tuner allows you to tune either the application to the platform or the platform itself. So Nico, have you guys had a chance to try the latest auto tuner from the Intel MPI library utilities?

Nico: So, to be honest, we didn't have time to take a look at the latest version, but we used the MPI tune in the past already and saw great results with it.

Mike: Yeah, again, you know, it does help you find somewhere between five and maybe even up to 10% better performance and in a large cluster, that makes a huge difference. If you can get your job done 10% faster, that just means, you know, faster time to solution or faster time to market, if you will.

Nico: Yes, of course.

Mike: You know, I know in the latest releases, we've done a lot of work in terms of optimizing for InfiniBand, and I'd like to hear if you've seen those results in the work that you've done as well.

Nico: Yes, definitely. So like I said, we used Intel MPI in the past as our default go-to MPI implementation on all of our clusters, but with the rise of Omni-Path, we had a feeling that Intel was concentrating on optimizing the Intel MPI software only for the Omni-Path hardware, and we saw some performance drops compared to OpenMPI on the InfiniBand network.

But in recent months with written updates, there are no more any difference to OpenMPI, and like I said before, we prefer Intel MPI to OpenMPI because it's much easier to use for the end customer, and so, we are back to Intel MPI as our default MPI implementation.

Mike: Yeah, that's great. Thank you for bringing that up to us as well as other customers have called that out. Let me just address a couple of things there.

So as far as Omni-Path goes, basically Intel has refocused our efforts in other areas. And as such, you know, that combined with the fact that we actually moved to the new implementation out of Argonne National Lab called the CH4 implementation, the Intel MPI team has, you know, had its hands full and as such, you know, there were some things that needed to be optimized. And that's why you saw some of the issues that we had with the performance with InfiniBand. But as you said, you know, you guys have gotten the latest releases and the team has done a really great job bringing that back up to what's been expected and what we're delivering now to the marketplace, which is, you know, performance across that fabric as well. So, to that end, you know, we are always keeping our eye, of course, on supporting our customers and ensuring that they get the best possible performance on the fabrics with the Intel architectures.

So, let me just touch on a few things on the evolution of the Intel MPI library. As I said, you know, we had moved to the new CH4 for implementation from Argonne National Lab, and that was really due to the fact that it is the next generation and you know, as we are on the dawn of exascale, the CH4 implementation out of Argonne basically, you know, puts us on to that path. And then of course, the market adoption has been really great and we continue to get good results and get good feedback from our customers—so, say, your company, Nico.

The other things that we are doing as the evolution goes is, as on-prem is still a very large market, we are seeing some use of Intel MPI in the cloud service providers too, such as Amazon and others. So, we've done work there as well.

I'd like to touch upon, Nico, some of the work that you guys have done, or your experiences really, with some of the other Intel technologies, such as Intel® Optane, and you know, what are your thoughts about fault tolerance? I know we touched on that earlier, but those seem to be, you know, again, given exascale, you know, 10 to the 18th, right? There's no doubt that things will go wrong and when they do go wrong, how do you fix them? Do you have any thoughts around how that will evolve from the perspective of Megware and yourself?

Nico: Yeah, sure. Like I said before, most of our customers, which are currently installing pre-exascale projects, they request for fault tolerance for their software and their hardware and one of the solutions that you can use there is, of course, the Intel® Optane, for example, to use the checkpoint restart mechanism in the Intel MPI implementation. But, of course, you can also use the Intel® Optane DIMMS in this regard and yeah, there's a research project we are undertaking together with the Jülich Supercomputing Center. They are looking into this and trying to find out how they can use this efficiently to maybe not even have to restart an MPI application, but also, yeah, just continue and move the memory to another node.

Mike: Hmm. Very good. Very good. Intel is also working on an open source project called DAOS. Have you seen any customers interested in DAOS, or had any experience with DAOS at all?

Nico: Yes, actually we tested this some months ago. We were in the early beta, and tested it in our benchmark center, and we also have two or three bigger customers who are testing this in the university cluster environment, and now we are seeing really good results on the performance. However, I have to say most of our customers out there are still using HDDs to store most of their application data because it's still the most cheap option you have for storing, but this is definitely an option we are looking into. And also our customers are looking into it for their future clusters.

Mike: Yeah. This is pretty exciting times in the industry with so many new technologies coming on board and you know, these initiatives. And one of them I'd like to talk about is heterogeneous computing.

Intel basically announced at last Supercomputing, at SC19, the oneAPI initiative. I'd like to get your take on what you think about heterogeneous computing in the cluster environment as well as how oneAPI can impact that.

Nico: Yes, so, heterogeneous computing will be, I think, the default or the standard way to build clusters in the future. We are very, very convinced about this because you will need to have some kind of accelerators to accelerate your applications, and there, we still see a lot of fragmentation in the market, where, you know, you have NVIDIA, you have AMD, and Intel GPUs are also coming up, and we are very excited about those. To have one

programming model which can talk to all of this hardware and use all of those accelerators in a very efficient way would be really, really great for the HPC community. Because currently you cannot write a program for all the accelerators which are in the market. You have to really decide on which accelerator you use, and then you have to write your code for this accelerator. And of course, this is really hindering the advancement of the whole field. So, we are really looking forward to seeing oneAPI in the market, and then seeing our customers adopt it for their applications.

Mike: That sounds terrific. So, you know, let's talk about what's next. I think Nico put it best. We have a variety of accelerator hardware out there with Intel basically coming into the game soon, and as such, you know, there is no one common programming model to address that.

I can tell you some of the things that we are doing with Intel MPI is basically figuring out how best to satisfy this demand for a common programming model in terms of the hardware. So, from that perspective, you know, I think you can look forward to us having support for these accelerators, or XPUs as we also call them, for the Intel MPI library moving forward.

So given that, what are the things that you're most looking forward to in the near future, and also maybe let's say three to five years out?

Nico: Yeah, like you said, with Intel GPUs are the most important hardware which will come in the market in the next years, and we are very, very much looking forward to have one in our benchmark center, and then of course, show to our customers and show them how they can port their applications via oneAPI to those accelerators.

Of course, we also still see Intel MPI as a very important software package on our clusters, and we would like to see even more performance on InfiniBand, Omni-Path and Ethernet, for that matter.

But on the other hand, we like to still have the user friendliness, which is currently very good and is currently the machine learnings and other software coming to the cluster. We see a shift in the usage of clusters. For example, some of our users are using the Jupyter notebook to run jobs on clusters, and if somehow we can abstract the Intel MPI to be used within such a Jupiter notebook, this would be really, really helpful for most of our customers because this will increase the use of friendliness a lot.

Nicole: Thanks guys so much for such a great discussion today. For listeners who want to learn more, where can they go for more information?

Mike: You can visit us at software.intel.com. Click on the data center link, and you'll find all the latest tools for developing for Intel MPI, as well as the rest of our software development tools.

Nicole: Nico, where can listeners go to learn more about Megware Computer?

Episode 11: Optimizing HPC Clusters: Meeting the Challenges of Today and Tomorrow
Host: Nicole Huesman, Intel
Guests: Mike Lee, Intel; Nico Mittenzwey, Megware Computer

Nico: You can go to Megware.com, where you can learn more about our products and cluster computing in general.

Nicole: Fantastic. Nico, it's been so great to have you on today's program. Thanks so much for your insights.

Nico: Yeah. Thank you too. It was really great.

Nicole: And Mike, thank you so much for joining us.

Mike: Thanks for having me.

Nicole: As HPC becomes even more ubiquitous and heterogeneous computing becomes, really, the default, it'll be exciting to see the evolution in this space. For all of you listening, thanks so much for joining us. Let's continue the conversation at oneapi.com. Until next time!