Architecting Systems of the Future

featuring Eric Werner interviewed by Suzanne Miller

Suzanne Miller: Welcome to the SEI podcast series, a production of the Carnegie Mellon University Software Engineering Institute. The SEI is a federally funded research and development center at Carnegie Mellon University in Pittsburgh, Pennsylvania. A transcript of today's podcast is posted on the SEI website at sei.cmu.edu/podcasts.

My name is Suzanne Miller. I am a principal researcher here at the SEI, and today I am very pleased to introduce you to Eric Werner, a chief architect at one of our newer programs, the Emerging Technology Center. In his role at the center, Eric establishes the technical direction of the center in the areas of software development, high-performance computing, cloud computing, and data analytics. But Eric also conducts research, and that's what we're here to talk about today. So, we're going to be talking about research that Eric and several of his colleagues are conducting to help software developers create systems for the many-core central processing units in massively parallel computing environments. Those are the environments that seem to be the future for software computing. So, this is really important get-ahead-of-the-power-curve kind of research. Welcome, Eric.

Eric Werner: Thanks for having me, Suzanne.

Suzanne: So let's start off by having you explain just what do you mean by systems of the future? How are those different from systems of today, and how do they present challenges for our software developers?

Eric: Our research is inspired by Moore's Law, which states that the number of transistors on a chip doubles every 18 months. That's classically interpreted as the speed of chips doubles every 18 months or two years, depending on the source. That's not always the case. In recent times chip manufacturers haven't been focusing on upping the clock speed of processors, because they've hit some sorts of limits in doing so. They have been focusing on putting more transistors on the chips, and that gives us parallel chips. So, we have multicore CPUs, which is to say a central processing unit can do many things at once.

Classically, programmers are trained to develop on single core, single processors, which is to say the computer does one thing at a time. In doing so you're not taking advantage of all the hardware that you have available.

In recent times, chip manufacturers and people designing computers have been focusing on graphical processing units. These are custom chips that have lots of little tiny cores that can do a lot of things at once. Historically, they have been used for rendering graphic scenes. It turns out you can do general purpose computation on these chips as well. It's a completely different programming paradigm, and we're trying to bring that to the developers.

Suzanne: So, you have to do things differently as a programmer. You can't just do what we call single-threaded kind of programming where you're just doing a serial process. You have to think about how you're going to branch off and how you're going to take advantage of those parallel computing CPU's that you're going to be using. And, that changes the way you architect things—that's why we're here—and it changes the way you implement as well, right?

Eric: That's right. There are lots of different programming paradigms to take advantage of multicore or multichip computers. You can do threads. You can do message passing. You can do things from functional programming. People have heard of map reduce, that's another example. When you bring GPUs [graphical processing units] into the mix, there are special purpose libraries for taking advantage of that kind of hardware, but it's different. It's unique, and frankly it's hard to use. So, the general thrust of our research is to make that kind of hardware available to your average programmer

Suzanne: I can see lots of different directions that research could take. What are the aspects that you're focusing on now? What do you think are the highest leverage things that the SEI can contribute in terms of helping developers to architect and to implement in these kinds of systems?

Eric: We decided to focus on two dimensions. One is limit the set of hardware that we're going to target for our early research, which is to say we're going to focus on GPUs.

Suzanne: Okay, the graphical processing units.

Eric: Correct. And then, in terms of the domain that we're going to focus on, we're focusing on graph analytics. Graphs are a data structure in computer science that are used all over the place. You can model social networks. You can model network flow. You can model all sorts of different things. But, they're hard to optimize because graphs don't have what's called "locality of reference." When you're at a node, and you're going to go to another node, you have no idea which node you're going to go to. So, it's hard for the computer. It's hard for the compiler. It's hard for the programmer to kind of line that up so that all those accesses [the computer accessing the information about the different nodes in the graph] are faster.

Suzanne: Right, so they get optimized.

Eric: Yes. So, what we're trying to do is, we're exploring using these different kinds of hardware architectures to attack this very important problem in computer science, and it's used all over the world in all sorts of industry, government, and in academia.

Suzanne: So, typically the SEI tends to publish reports. That's our main mode of disseminating information. In the recent past—and I know over at the ETC—we've also done some work where we've provided prototypes and things for people to use. Is that part of your vision for this? Or, is this more of a traditional report type?

Eric: It is. For the initial phases of our research we're trying to get our arms around the problem. So, our initial output will be a technical report describing the problems facing our approach. The next phase of our research is—we're actually looking to release a library that developers can use. It's a software library that developers can use for doing graph analytics targeting different hardware architectures.

Suzanne: And, a couple of exemplars of how that library's been used to do some tasks, as an example?

Eric: Absolutely. One of the key things we're trying to do is separate the concerns, or understand how much we can separate the concerns from the graph analytics part to the.

Suzanne: The domain aspect.

Eric: The domain aspect from the hardware part.

Suzanne: Okay, very cool. So, this research is a little different from some of the other things that are going on at the Emerging Technology Center. How do you see this fitting in with some of the other things you're doing there?

Eric: Right. So, the Emerging Technology Center, it's a relatively new program at the SEI. If I had to say what we focus on, it's a data-intensive, scalable computing. We help the government stay on top of the edge of technology.

We do this through a number of different mechanisms. We do technical evaluation, which is we understand what's out in the world, and we can help talk about it, and we can help map the technology that's out there to important government problems. We do technology demonstrations. We do prototyping. We bring stuff into our lab and, in a hands-on way, we put it through its paces. We help with technology transition, which is we get it out of the lab and into the hands of the folks doing the mission.

Suzanne: That's really what some of this is.

Eric: Absolutely.

Suzanne: These GPUs are out there, but it's how do we use them in a better fashion for doing tasks that we may not have thought about using them for before and using them in ways that are going to be focused on operational problems, not just video games.

Eric: Exactly. And, we're trying to make it less of a niche skill. Right now you need very special-purpose training and understanding to be able to take advantage of these types of hardware architectures.

The final thing I wanted to touch on with what the ETC, the Emerging Technology Center, does is reality transfer. There are a lot of interesting problems in the world. There are a lot of hard problems in the world. We have knowledge in the government space to transfer our understanding of *What are important problems to be working on*? *What are relevant problems to be working on?* for the government. So, when we go and talk to folks in academia or folks in the industry who are looking at new things, we can help inform them about priorities that our government…

Suzanne: So, you're acting as a proxy for all of those problems that you're seeing in government and filtering those out to the industry to help them understand what actually the reality is in the world, in the government world.

Eric: Absolutely. Because people who are doing world class research aren't necessarily tracking all the ins and the outs of the mission, and so we help do that translation.

Suzanne: And that mission is changing.

Eric: Always.

Suzanne: I know one of the areas that ETC works in is the cyber intelligence, cybersecurity side. That's a very dynamic kind of space, so that reality transfer is an important part of what you guys do.

Eric: Yes, and even the definitions of some of those things are changing.

Suzanne: Yes, indeed. This kind of work has got to involve collaboration with hardware manufacturers, with researchers that are doing advanced kinds of software development techniques. Who are you working with in this area, collaborating with, to make this architectures-of-the-future really happen?

Eric: So, we wanted to start off with a platform and an environment that's well known. There's an organization called Graph 500—which is an international organization of researchers and folks who are working the space of graph analytics—where they define a standard set of tests

and a standard set of algorithms where you can test your implementation of graph algorithms against their framework.

So, we wanted to fall in on something that was already out there in the world. We're working with Andrew Lumsdaine. He is from Indiana University's Extreme Scale Computing Lab. He is considered a world leader in the area of graph analytics. He's on the executive committee of the Graph 500, and he's one of the collaborators on our project.

Suzanne: Nice. You said you've also got some hardware collaborators, or are you mostly using commodity hardware?

Eric: We're trying to use stuff that is out in the field.

Suzanne: Okay.

Eric: So, we have GPU servers in our server cluster that we're using.

Suzanne: You have quite a lab over there from what I understand. He is smiling in case you can't see him on the podcast. So, let's say that you succeed in this in terms of this first phase—and you start getting your prototypes out—once you've got some prototypes out that demonstrate how to use this library, what do you see as being the next big problem in moving towards making multicore computing accessible to software developers?

Eric: So again, we restricted our research in these early phases to a domain, which is graph analytics, and to a specific type of hardware, which is GPUs. I think as we look at the curve of technology, GPUs are only the start. People are building systems that have multiple GPUs in the same system. If you look at recent mobile phone releases, not only do you have a CPU and a GPU, but you have extra processing units; auxiliary processing units that might do special-purpose stuff; for example, understanding motion sensors inside of a phone. There're all these different types of…

Suzanne: Or voice processors.

Eric: Exactly. There are all these types of architectures that are showing up in phones and computers. We want to make sure that people can take advantage of them. So, for the first part of our research we limited ourselves to GPUs.

Looking forward, we can look at other systems. For example, people are putting multiple GPUs in the same system. You have auxiliary processors in the same system. We can also look at FPGAs, which are field-programmable gate arrays.

Suzanne: You're looking at multiple domains. You're looking at mobile, computing and other places besides traditional systems and these are showing up as well, right?

Eric: Absolutely. Many people work on desktop systems or laptops. We're also looking at making this type of programming applicable to datacenters but also on the edge in mobile devices.

Suzanne: We've got other research that we talked about in the podcast about some of the work we're doing in that, so there's probably some synergy between you and some of the folks over at Ed Morris's group, the mobile-at-the-edge kinds of stuff.

Eric: Absolutely. We have a great relationship with his group, and we're really excited about the work they're doing as well.

Suzanne: Excellent. Alright, well thank you very much for joining us today. I'm looking forward to results in some of these other areas. I especially want to look at some of the prototypes when they become available.

For more information on the research that Eric's team is doing at the Emerging Technology Center, please visit, http://www.sei.cmu.edu/about/organization/etc/ and that's ETC, Emerging Technology Center, not Et cetera.

You can also check out our podcast interview earlier this year with Eric's ETC colleagues, Jay McAllister and Troy Townsend, who discuss their work on cyber-intelligence tradecraft. This podcast is available on the SEI website at sei.cmu.edu/podcasts and on Carnegie Mellon University's iTunes U site. As always, if you have any question, please don't hesitate to email us at info@sei.cmu.edu. Thank you.