UC SANTA BARBARA
THE Current

April 3, 2023
James Badham

Jonathan Balkind receives NSF Early CAREER Award to improve cloud-based computing

Employing a technique called microarchitectural checkpointing to redesign computer processors for cloud-based serverless computing — a new paradigm favored by cloud developers — Jonathan Balkind is developing a new application for cloud computing. An assistant professor of computer science at UC Santa Barbara, he is doing so with funding from the National Science Foundation, by way of a five-year, $630,000 NSF Early CAREER Award.

“It’s really an honor to receive the CAREER award,” Balkind said. “This is my first funded NSF proposal and was the first time I made a submission for the CAREER. I’ve had to pinch myself at least once to believe that it really happened. I’m looking forward to driving this project over the next five years.”

The long and short of application run times

While applications created for servers run for up to weeks at a time, the new serverless apps run for as little as a hundredth of a millisecond — meaning many existing processor technologies cannot keep up.

“We have spent several decades optimizing processors for long-lived applications, so that the processor could learn their behavior over time in order to predict future

behavior and, thus, operate more efficiently," Balkind explained. "With today's very short-running applications, like those in serverless, there simply isn't enough time for our processors to learn the behavior. This makes it inefficient to run serverless applications on existing servers.

"But with microarchitectural checkpointing," he continued, "you save what you learn each time the application runs, and then when you run it again later, you retrieve what you saved and then save at the end of that step, and so on. The result is that the processor learns only what it needs to, across instances of the application over time. The checkpointed information from each application is siloed, so you avoid polluting the information from one application with that of another. We will use microarchitectural checkpointing to improve the efficiency of serverless apps."

Open-sourcing a customized serverless processor

Key to Balkind's CAREER award research is the OpenPiton platform, which he will use to enable prototyping for his open-source framework for building processors. "We have a design for a processor," he said, "and people can make modifications to it, either to add a feature they want or to test things as they change the parameters of that processor, such as the number of cores or the amount of cache."

The system has evolved from work that began in 2013, when Balkind and fellow Princeton University Ph.D. students designed it to serve as a research platform that would enable users to add their own components to validate particular research ideas. "We give out just about every component that's needed to design a new processor, and as a result, we've seen users be very productive — more than 60 research projects have used the platform," he said. Additionally, a number of companies have adopted OpenPiton, including Intel, which used the platform to develop an 8-core processor chip to demonstrate the effectiveness of its new fabrication facilities.

Balkind is a significant contributor to the open-source hardware space, where, he said, "We're providing these designs and trying to build a community and make better products in the future." He received an Open Source Hardware Association Trailblazer Fellowship for his work in that field.

Making and customizing processors for specific applications is an important part of the evolution of computing. "In industry, companies routinely customize their processors for new applications as they emerge," Balkind said. His proposal for

microarchitectural checkpointing will be demonstrated as a customization of OpenPiton, which can benefit serverless applications. By open-sourcing this processor design and providing a concrete implementation of the idea, he and his team hope that it will see easier adoption into other industrial processors.

On-demand cloud computing

“If you’re a developer, there are lots of ways for you to write an app,” Balkind noted. “But if your app suddenly gets discovered, and you have, almost instantaneously, a million users a week, you need to have the flexibility to go from one server to 10,000 servers handling your requests. Serverless computing is specifically designed to do this for you.”

Around 2016, Amazon and other companies discovered that they were using only about 65% of their data-center capacity, leaving about one-third of the resource unused. Amazon responded by inventing a paradigm that would be easy to program and make it possible to scale up and down at a moment’s notice. “So, the NFL moved a bunch of their web serving to this paradigm, because they have two or three days a week when everyone uses the website, and the rest of the time it’s much quieter,” Balkind said. “It’s the same with banks. At the end of the month, customers scan their paycheck to deposit with their phone, causing a huge spike in demand that lasts three days a month. Spikes also occur that can’t be predicted.”

The hope is that not too many demand spikes occur at once, so that there is an even distribution of usage over time.

The system makes sense, but there’s a catch: the additional, previously unused 35% of capacity that Amazon and other cloud providers had available isn’t as reliable as the heavily used 65%.

“They can’t guarantee you’ll get good performance when your app runs,” Balkind said. “To make up for that, they sell a plan that allows you to pay for only the time when your application is running, whereas, normally, you pay even when it’s idle. If you’re a small-scale start-up developer and you have no demand, it’s OK; you pay nothing. As people start to use your service, you pay in a way exactly commensurate with your usage.”

Amazon was the first to do this, with its Lambda platform. “Once they did it, everyone else followed,” Balkind said. “The problem, however, is that for each individual request, it turns out they’re not getting great service — one command

might run instantaneously, and the next might take 30 seconds. That’s what we’re trying to improve.”

Media Contact

Shelly Leachman

(805) 893-2191

shelly.leachman@ucsb.edu

About UC Santa Barbara

The University of California, Santa Barbara is a leading research institution that also provides a comprehensive liberal arts learning experience. Our academic community of faculty, students, and staff is characterized by a culture of interdisciplinary collaboration that is responsive to the needs of our multicultural and global society. All of this takes place within a living and learning environment like no other, as we draw inspiration from the beauty and resources of our extraordinary location at the edge of the Pacific Ocean.