Interview with Swami Sundararaman

RIK FARROW

Swaminathan (Swami) Sundararaman is the Lead Architect of ParallelM, an early-stage startup focused on production machine learning and deep learning. Swami was previously at Fusion-io Inc. and Sandisk Corp. He holds a PhD from the University of Wisconsin-Madison. swaminathan.sundararaman@parallelmachines.com

Rik Farrow is the editor of *;login:*. rik@usenix.org

When I read the announcement for the HotEdge workshop [1], I was immediately intrigued. What the heck is HotEdge? I thought the "edge" consisted of network devices and content distribution networks. As I read the prospectus, I learned that edge, in this context, means something quite different from what I (and most of my friends) considered it to be.

I'll let the words of one of the co-chairs, Swami Sundararaman, do the explaining.

Rik Farrow: I thought the "edge" consisted of network devices and CDNs. But the edge in HotEdge is something different.

Swami Sundararaman: Edge computing has many definitions depending on who you ask. The one that I like the best is the following: edge computing is a new computing paradigm where server resources, ranging from a miniature computer (such as Raspberry Pi) to a small datacenter, are placed closer to data and information generation sources. Application and systems developers could use these resources to enable a new class of latency and bandwidth-sensitive applications (such as augmented reality, wearable cognitive assistance, sensor data processing, etc.) that are not realizable with current cloud computing architectures.

In most ways, edge computing is the opposite of cloud computing and therefore requires rethinking many tradeoffs that have become normal in cloud computing. Compared to its potential and the dire need for solutions for upcoming applications, we did not see workshops to foster early-stage ideas in this field as it deserves. As co-chairs, Irfan Ahmad and I wanted to help advance the field of edge computing by providing a venue where both researchers and practitioners could come together to both share their vision for building edge computing applications and systems and also discuss their nascent ideas and receive feedback well in advance of rigorous academic or industrial product treatments.

RF: So what's changed that has created the need for this new computing paradigm?

SS: There are two major trends that are driving the demand for low-latency offloading infrastructure. First, with the advent of Internet of Things (IoT), there are many more connected devices that are constantly generating tons of data (such as video, audio, sensor data, image, text, etc.). Second, the need to act or react quickly to changes provides tremendous value for businesses using sophisticated techniques (such as machine learning, deep learning, and image processing) that are both compute and memory intensive. Unfortunately, having heavy compute and/or memory demands on these sensor or other devices is not always possible for multiple reasons: power requirements, form factor, cost, development effort, etc.

Also, the latency required to move the data from these devices to the cloud (which could take multiple hops) is high and would result in a poor user experience. Even cloud providers such as Amazon and Microsoft have identified the above-mentioned trends and have started to heavily invest in edge computing (see Greengrass [2] and Azure IoT Edge [3]).

RF: Looking at your background [4], I see recent work with non-volatile memory, and older work in operating systems and storage. To me, edge computing seems very different from what you have done in the past. What drew you to create a gathering place for edge researchers, given that you have focused on other areas in your past?

SS: This is a great question.

I love working on cutting-edge technologies and was fortunate to work on many diverse interesting problems in the past. As you have correctly observed, I started with file and storage systems research and then worked on operating systems and distributed systems in addition to traditional storage systems during my PhD. When I graduated, I jumped at the opportunity to work on non-volatile memory (including flash and persistent memory technologies), the latest upcoming storage technology at that time. At my current job, I am working on automating the deployment, orchestration, and management of machine learning in production.

As a company, we were initially interested in deploying machine learning at scale in the context of IoT. Very soon we realized that the biggest challenge to deploying performant machine learning on (or near) the "things" in IoT is the lack of infrastructure and standards. We explored the possibility of leveraging edge computing to solve our problem since it was very promising and had the potential of being the vehicle to deploy machine learning (and other upcoming compute-intensive technologies) because it addressed many of the issues (such as latency, power, compute, scale, etc.). Unfortunately, we couldn't fully embrace edge computing for its lack of wide-scale adoption. On further investigation, we discovered that there are still many open problems in edge computing and also that edge computing itself is not yet well defined.

These problems motivated me to contribute and help advance the field of edge computing. We realized that there were a few full-fledged conferences (such as SEC and Edge), but there were no workshops to discuss nascent ideas similar to what we have in other fields (such as HotOS, HotCloud, HotMobile, HotStorage, etc.). This was the primary motivation for starting HotEdge (which serves as a gathering place for edge computing researchers).

RF: Edge sounds both very interesting and important when moving forward with many technologies. But the requirements you mention, such as the need for systems that can provide enough compute power or can be scaled out to do this, will also be attractive targets. The data processed on these systems will also need to be protected. While it's still early days for edge, are people starting to think about the security requirements for these systems? I find myself imagining edge systems mining digital currency or used for spying on people using augmented reality.

SS: Yes, researchers and also industry folks have already started thinking about both security and privacy in edge computing. This is one of the key pieces needed for the adoption of edge computing as multiple entities/users could be sharing the same edge infrastructure (including CPU, memory, network, and storage). There have already been a few blogs and papers that focused on addressing both security and privacy issues in the context of edge computing.

References

[1] HotEdge Workshop: https://www.usenix.org/conference/hotedge18.

[2] Amazon Greengrass: https://aws.amazon.com/greengrass.

[3] Microsoft Azure IoT Edge: https://azure.microsoft.com/en-us/resources/videos/microsoft-ignite-2017-enable-edge-computing-with-azure-iot-edge.

[4] Swami Sundararaman at University of Wisconsin: http://pages.cs.wisc.edu/~swami/.