Taming Big Data for University Research

This Virginia Tech Hume Center for National Security and Technology research project into the impact of machine learning on radio frequency management lacked the computing power required to manage its vast quantities of data. Fortunately, Pure Storage FlashBlade eliminated the interference.

WILLIAM CLARK HAD A BIG DATA PROBLEM. Clark is a research associate and Ph.D. student working out of **Virginia Tech's Ted and Karyn Hume Center for National Security and Technology**, and the computing resources he used in his work couldn't keep up with his research projects in radio frequency machine learning (RFML). The research topic he was tackling required such vast quantities of radio frequency data, the five servers he was allotted for the job had to commit most of their processing power just to contain it, leaving little computing power available for churning through the data.

While Clark's research is focused on RFML, one aspect of what he's studying is how much data is enough to say that a given machine learning model is being trained well. That could require anywhere from 2,000 observations, or examples of radio frequency data, to between 2 million and 14 million. "The more data you use, the slower the training because the model has to go

through all of that data each time it updates its learned features," he noted.

The mission of the Hume Center is to get students involved with research so compelling, they'll be persuaded that careers with the U.S. government could be as satisfying as jobs in industry. Clark had eight undergraduates doing the coding needed to transmit and collect all of that RFML data — and the limited computing power they had to work with was unlikely to inspire them to pursue public sector work.

"Students are very interested in working in machine learning and artificial intelligence," said Dr. William "Chris" Headley, associate director of the Electronic Systems Lab at Hume. "But when they think of ML or AI, they think about commercial entities like Google or Facebook or Amazon. We want students to know, yes, you can do cutting-edge stuff, but you can do it while helping our national security too."

The Importance of Spectrum Sensing

Radio frequency is something we're all familiar with. Anybody who's ever used a cellphone, cordless phone, Bluetooth device or remote-control car has taken advantage of wireless signals. The Electronic Systems Lab at Hume focuses a lot of its attention on "spectrum sensing."

"There are a whole bunch of signals in the air all around us, different formats, different frequencies. And what we're trying to do is figure out where they are, when they are and what they are," Headley explained. Spectrum sensing helps engineers understand where the signals are so they can design systems that avoid having those signals interfere with each other. The less interference there is, the faster the performance. Plus, by making spectrum sensing faster and more effective, the greater the number of signals that can be packed into an area and the faster the speeds — a set of techniques known

Spectrum sensing helps engineers understand where the signals are so they can design systems that avoid having those signals interfere with each other. The less interference there is, the faster the performance.

About the Hume Center

Virginia Tech's Ted and Karyn Hume Center for National Security and Technology has a core mission embedded into all of its programs: to engage students in research that will help prepare them as next-generation national security leaders.

The Center has three primary divisions:

The Electronic Systems Lab, where Headley and Clark work, which covers the electromagnetic spectrum, spectrum dominance, RF machine learning, embedded learning and related topics, tends to draw in students studying electrical engineering and computer engineering.

The Intelligent Systems Lab, focused on cybersecurity, data science and AI work, pulls in students majoring in data analytics and computer science.

The Aerospace and Ocean Systems Lab, for aerospace engineering students, concentrates on aerospace and satellite development and drone work, with an AI influence.

Undergraduates are involved in experiential learning and internship and career opportunities in the national security sector. At the graduate level, research programs provide sponsored research with defense and intelligence organizations. As the mission statement for Hume Center clarifies, if student involvement isn't part of the research program, it doesn't belong there.

as "dynamic spectrum access."

From a national security perspective, spectrum sensing is used to figure out where an adversary is communicating, to eavesdrop on it or jam it and to come up with ways to avoid having the enemy detect or interfere with signals. This military application is an endeavor Headley called "spectrum dominance: "How do we dominate the spectrum? How do we make sure that our communications are successful, and our adversary's communications are unsuccessful?"

Machine Learning in Radio Frequency

Traditionally, the U.S. government has used a brute force approach in allocating spectrum for various uses, which, as Headley put it, is "very fixed, rigid and underutilized in a lot of ways." As an example, the U.S. Navy has a dedicated spectrum band that applies throughout the entire country. "But it's a broad stroke," he said. "You don't really need naval radar in Kansas."

On the consumer side, WiFi and Bluetooth radios built into the access points and other devices people use today are already designed to sense the spectrum and figure out how well they're performing. If they perform poorly, they'll adapt automatically by moving to a different part of the spectrum. RFML is trying to make what's called "cognitive" or "intelligent" radio, where instead of just adapting, these devices can learn to predict.

"If we can learn the habits of users in a spectrum — oh, the spectrum is used here at this part of the day more often — then we can predict to avoid that part of the band at that time of day and avoid the interference in the first place," said Headley.

But as he observed, "There's a lot going on in wireless communications. You've got all of the parameters of the actual signal that you're trying to transmit, there are the imperfections in the transmit hardware, the receiver hardware, you have the channel propagation effects. And when you combine all those things in a real system, it's complex. We can model different pieces of it, but

Machine learning isn't some "godlike entity that will solve problems that we already know the answer to," Headley emphasized. "Where machine learning shines is where we don't currently, as of today, have good models of the underlying systems but have access to lots of data."

putting it all together and making it work for different environments — urban, rural, mountainous, desert — it's very hard to model."

Machine learning isn't some "godlike entity that will solve problems that we already know the answer to," Headley emphasized. "Where machine learning shines is where we don't currently, as of today, have good models of the underlying systems but have access to lots of data."

That's where Clark's work comes in. By gathering — or synthetically creating — massive data sets, machine learning can try to solve the various spectrum problems and develop what Headley referred to as "intuition" about them.

The Challenges of Wrangling Wireless Data

There are a multitude of "canonical" data sets online — just not for RFML. If somebody is researching machine learning for audio, speech or video, there's plenty of that to download because people are constantly uploading it to numerous social sites. But who bothers recording their wireless traffic?

That means the data to load into the RFML models needs to be generated. There are two primary ways to do so. As Headley explained, researchers can "stick an antenna out the window and capture data" during live exercises. Or, as Clark has done, they can combine synthetic data and real-world data to produce data sets that combine the benefits of both.

"The nice thing about real data is it's real," noted Headley. "You get the hardware effects; you get channel propagation effects." At the same time, he added, that's a time-intensive way to get data for testing various models.

Also, there's an added wrinkle: privacy concerns. "When people upload their pictures to Facebook, they've signed a term of agreement that Facebook can use that picture," Headley explained. "If the government wants to create a canonical data set of wireless traffic, they can't just upload people's cellular

communications or wireless traffic. What they have to do is scrub the personal identifiers from that data, which is extremely cost-intensive."

On top of that, the data has to be "labeled." "If I want to know where signals are, and I want to train machine learning to do it, I have to know where they are in the first place to get the network going. And that's very hard to do on real data," he said. "That's a constant problem we have — data is the dominant part of the problem."

"You need to have as much information as you possibly can collect — terabytes of information," said Clark. "So, figuring out the database structure, how to store it, how to get it on disk efficiently and back off disk efficiently is a significant challenge."

As an example, a **DARPA challenge** run in 2019 by the Defense Advanced Research Projects Agency to demonstrate a radio protocol that could best use a given communication channel in the presence of other dynamic users and interfering signals required 23 packed server racks in a custom-built testbed environment to generate mere seconds of simulation. (Virginia Tech's own team came in 11th out of 30 teams.)

Or consider WiFi. According to Clark, WiFi uses up to 20 megahertz of spectrum. "That's 20 million samples per second. If I want to look over 10 minutes [of active WiFi data], that's basically a full hard drive when including the idle time as well. Collecting it and storing it and reusing it is extremely expensive and time-consuming."

Then, with machine learning, the work consists of training models that replicate real life. As Clark explained, "Everything is in a random configuration. And then you put your data through and update your model to learn. If you do it once and you get lucky, you're off! But chances are you need to do it hundreds or thousands of times to even know whether the model you're using is well suited to the problem."

"You need to have as much information as you possibly can collect — terabytes of information," said Clark. "So, figuring out the database structure, how to store it, how to get it on disk efficiently and back off disk efficiently is a significant challenge."

Pure Storage FlashBlade Explained

Pure Storage FlashBlade is an all-flash storage system for consolidating both file and object data. To increase capacity and performance, the institutional customer simply has to add more blades, starting at seven blades and scaling up to 150. The object storage used by FlashBlade is the same format used in public cloud storage services, such as Amazon S3. However, FlashBlade adds the advantage of unifying traditional file storage with object storage into a single solution.

According to the company, Pure Storage's approach to flash design provides better performance, longer media endurance and higher storage device density than off-the-shelf solid-state drives.

The company's subscription **Evergreen Storage** program, heralded by analyst firms as the secret to Pure Storage's high Net Promoter Score service ratings, has (as **one IDC report described it**) "forever changed" customer expectations about technology refresh and storage lifecycle management by delivering numerous benefits in two levels of support. At the lower-cost "silver level," support covers:

- "Love Your Storage," which formalizes the company's 30-day money-back guarantee on new system purchases
- Bundling all array software with the flash array rather than delivering it piecemeal and with a bunch of extra price tags
- "Right-Size Guarantee," which protects customers in their storage sizing efforts and higher-than-industry-norm data reduction ratios
- "Flat and Fair," a maintenance program that guarantees predictable maintenance rates on a per-device basis and free replacement of failed components over the life of the array
- "White-Glove Support," which alerts customers to potential array problems before operations are affected, through the use of artificial intelligence, machine learning and predictive analytics

The "gold" version of the subscription adds upgrades to new controllers every three years, either through "Free Every Three" or "Upgrade Flex" programs.

Learn more **at the Pure Storage website**.

Add up the sum total of the whole job, and it ends up being a lot for a small group of servers to process.

Finding a Better Solution

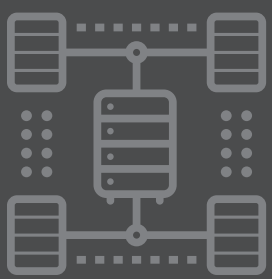
While Clark struggled to maintain control in his tiny boat adrift on a sea of data, Headley set about coming up with a better solution for controlling the seas. He knew that storage centralization held the key. He wrote an application and won a DURIP grant under the **Defense University Research Instrumentation Program**, which allows institutions to acquire equipment for supporting research projects run for the U.S. government.

After evaluating numerous vendors' solutions and conferring with the university's IT organization, Headley and his team zeroed in on **Pure Storage** and its **FlashBlade** storage system. "We liked what we heard, we liked what they had available and we developed a relationship with them," he recalled.

The two biggest draws were Pure's speed and adaptability. "We could start with a lower amount of storage, but then it was very easy to slot in more storage as the need arose. We can't buy a Pure Storage FlashBlade on every program we have, but we can buy smaller pieces to build a bigger system. We really liked the ability to upgrade in that fashion."

There was no shifting this data management work to the public cloud. The servers Clark is using reside in a chilled storage room on the Virginia Tech campus. As Headley pointed out, the Hume Center focuses on national security research, with varying levels of restriction. "Since the data is restricted, we can't just put it in the cloud," he said. "That's another reason why the FlashBlade, as opposed to a cloud-based solution for storage, is so incredibly important to us." It has to reside in a protected environment with heightened levels of security imposed on those who access the servers.

Pure Storage sent its own expert to work alongside Virginia Tech's IT team for the installation and testing. While that process went smoothly and had no implementation issues, a problem surfaced on the Virginia Tech side: There was an insufficient number



"Instead of each system having only a small portion of my data, the five servers suddenly could see the entire data set online and could train all at once, as opposed to doing piecemeal approaches," Clark recalled.

of switch ports to accommodate the full array — ten blades with 17 terabytes of storage per blade. That shortage has since been addressed. "And it's just going great," said Clark.

From Farm Road to Superhighway with Pure Storage

A major benefit of using Pure Storage FlashBlade is that placing the data into a central flash array instead of on the servers frees those machines up to do what they're best at: processing data.

The FlashBlade is simple, said Headley. It's purpose-built for situations where users — including campus researchers — can leverage their troves of data to gain the insights they're seeking.

"Spectrum is a finite resource. Being able to open the door to more intelligent, more predictive communications and get rid of these rigid spectrum allocations will allow more people to use the spectrum and drive down the costs," he suggested.

So far, Clark has trained 5,000 models over the past six months. Previously, it would have taken him three weeks to train between one and 10 models, depending on how much data the model required. Now he's training the same number in a single day.

The moment Clark's data was moved to Pure Storage's FlashBlade, his tiny collection of over-worked servers was pointed back to the data, and it was like turning off a one-lane farm road onto a superhighway. "Instead of each system having only a small portion of my data, the five servers suddenly could see the entire data set online and could train all at once, as opposed to doing piecemeal approaches," he recalled. "Instead of having five servers picking at the problem, I had five servers that could fully tackle it. They would probably still be churning if I didn't have Pure Storage as the backend." ■