Bruke Kifle: This is ACM ByteCast, a podcast series from the Association for Computing Machinery, the world's largest education and scientific computing society. We talk to researchers, practitioners and innovators who are at the intersection of computing research and practice. They share their experiences, the lessons they've learned and their own visions for the future of computing. I'm your host, Bruke Kifle.

Today, we're embarking on a journey through the cutting-edge domains of systems and networks, mobile and edge computing, security and privacy, all integral components shaping the landscape of our digital future. Systems and networks facilitate seamless communication and collaboration in our increasingly connected world. Mobile and edge computing redefine how we access and process information empowering real-time decision-making on an unprecedented scale. At the same rate, security and privacy are essential for safeguarding digital identities and sensitive information from potential threats. These pillars are the backbone of our digital world, shaping how we communicate, innovate and protect our digital infrastructure.

Ramon Caceres is a computer science researcher and software engineer born and raised in the Dominican Republic. His areas of focus have included systems and networks, mobile and edge computing, mobility modeling, security and privacy. Most recently he worked at Google where he built large-scale privacy infrastructure and was previously a researcher at Bell Labs, AT&T Labs, IBM Research and has also held leadership positions in several startup companies. Ramon is an ACM fellow and IEEE fellow, he serves on the board of the CRA committee on widening participation in computing research and holds a PhD and MS degrees in computer science from the University of California at Berkeley and a bachelor's of engineering from McGill University. Ramon, welcome to ByteCast.

Ramon Caceres: Hi, thank you very much for having me.

Bruke Kifle: You have such a remarkable and interesting career that spans research, products, academia, industry. Can you describe some of the key points across your personal and professional career that have led you into the field of computing but, more specifically, motivate you to pursue your field of study?

Ramon Caceres: Yeah. So, it is been a very indirect path and I find that, often, maybe young people feel that they look at someone who's near the end of their career with some achievements and feel that people got there in a very direct purposeful path but I think, often, as in my case, it took a little bit of exploration and different paths. So, let just give you a sense. So, I was interested in engineering since I was a boy. I grew up in a community associated with a remote mining operation in the Dominican Republic and it was very, very small, under a hundred people. But the engineers, in my mind as a boy, they built the place, literally, they built all the infrastructure from scratch, the roads, the buildings, the phone network, the electrical network, the plumbing and, whenever there

was a problem to be solved, they often just built something new to solve it and so I was always interested in that ability to build infrastructure that serve people.

So, that was an interest I had from very young but as I was finished ... Getting into my senior year of high school, I was still undecided as to which type of engineering I wanted to pursue in university, civil or mechanical or electrical, I just didn't know. And just by chance, a classmate lent me an issue of *Scientific American* in the beginning of my senior year of high school, I remember very clearly, and that issue was focused on integrated circuits and their uses in computers. And I had never touched a computer before, I probably had never seen one other than a calculator but I read the issue with interest and the topic just fascinated me. So, largely on the strength of reading up that whole magazine cover a cover, I decided to pursue computer engineering for my university career.

Back then, it was often considered a branch of electrical engineering so, formally, I joined an electrical engineering program at McGill University. And as my undergraduate studies advanced, I came to realize that I liked computer science as a whole, including specifically software, not only the hardware topics that had been the focus of my undergraduate curriculum in computer engineering. And so, I went ahead and finished the electrical engineering degree, I was very far along when I realized these other interests and so I decided to pursue a masters in computer science to learn more about computers and computer science. And when I went to Berkeley for my master's, I fairly quickly switched my focus from hardware to software after taking some courses in operating systems which I loved immediately.

So, another type of gradual exposure was to research. So, I had no idea what research was when I started my undergraduate career, certainly. And even when I finished it, I'd never really done any research, I hadn't had the opportunity to work at a research lab at the university. But certainly, during my masters at Berkeley, I was surrounded by great research and did a little bit of it for my masters project and I was very intrigued by doing more research. But I went ahead and finished my plan to go to work after my masters partly for financial reasons. So, I worked for three years as a software engineer building product, a lot of fun, I learned a lot but in the back of my mind I had this nagging feeling that I wanted to learn more about research.

And so, after three years as an engineer, I went back for my PhD and that sent me on the path to a research career. But as you mentioned, I've gone back and forth between research and engineering my whole career and it's been a fun and indirect path which, yeah, it's kept me happy.

Bruke Kifle: That's quite exciting, it seems like it's a combination of a journey of both exploration but also, I think, as you described it, gradual exposure. But also I think you're able to blend such an interesting foundation in the hardware in

addition to some of your studies and experience in software, right? So, I'm curious how that breadth of exposure, both from a research to practice but also from a hardware and software end-to-end view has enriched your contributions as a scientist and as an engineer.

Ramon Caceres: Yeah, that's a good question. I've often thought that having that strong base in hardware which I did from the computer engineering focus of my undergrad has really helped me throughout my career. I can't speak for how other folks think but working, certainly, in systems and networking, which is a fairly low-level software, it's often been useful to me to, I felt, to have a really strong sense, concrete sense, of what the hardware was doing underneath the software that I was writing. Whether it be the CPU and the caches or a network switch or other parts of hardware, being able to visualize what they were doing concretely, I think, helps a systems and networking person make sound design decision grounded in reality. So, I think that that breadth has been very helpful to me so that's ... Yeah, I believe so.

Bruke Kifle: Certainly a strong argument to be made for a lot of modern-day computer scientists who may potentially have a strong foundation in the software to also ensure equal grounding and understanding what happens in the back, right, understanding the fundamentals behind low-level software but also how the actual hardware enables some of these capabilities. So, I think that's quite interesting. I would love to touch on one of your primary areas of focus which is this idea of privacy and security. Obviously, safeguarding privacy and security and systems that handle vast amount of data, of user data is actually very important and has been a pervasive topic of discussion for many years, possibly many decades.

So, what do you see, having your breadth of experiences working at Google, working at IBM Research, what do you see as some of the most significant challenges in achieving this goal of managing large amounts of user data but still upholding and safeguarding these principles of privacy and security?

Ramon Caceres: I think the most crucial thing is to prioritize and build in privacy and security from the beginning of designing and building any system or an application or product or service. This is well known, I'm not the first person to say this, but, unfortunately, there are often commercial interests that get in the way of that, not maliciously, but simply because it's a matter of priorities, where to put resources, where to put the engineers and so on. And there's often a lot of pressure to get to market quickly, you always have to make choices in engineering as to where you put your resources in.

It's surely been the case many times where privacy and security have not been the top priority and, therefore, there's been a certain loss of privacy and security in systems that have come out to the world. And it is very difficult, in my experience, to put privacy and security back or to add it to something. Once it's been designed and built and especially once it's out in the world, it takes a

lot of work and it's just quite difficult. Sometimes the horses left the barn kind of thing sometimes in terms of data and after several years of data being collected without too many precautions and it's very hard to get control of that data afterwards.

So, I think that's really the challenge is to, for us as a community and as a society even to prioritize these two goals of privacy and security from earlier on in the design process for any product, that's going to be an ongoing challenge forever because those trade-offs will always be there and it's possible to do it. I've seen good examples and, unfortunately, we've seen a lot of bad examples in the past as well.

Bruke Kifle: Yeah. Maybe grounding this in one of the projects that you've been involved in during your long and fruitful tenure at Google, you were a key member of the team that developed and operated Zanzibar which I understand is Google's global authorization system. For the audience who maybe may not have an understanding and myself, could you shed light on what is the problem that this authorization system addresses and why is it critical for an organization like Google that manages such large number of user data and user requests? Why is it so important to this digital infrastructure of a company like Google?

Ramon Caceres: Sure. So, the central issue addressed here is that determine whether an online user is authorized to access a digital object that's out there is central to preserving privacy. So, that decision of this can this user access this object when they attempt to access it is an authorization decision, that's what an authorization system does is make that decision, does this user have permission to access this object. And Zanzibar is such an authorization system and the main challenge in building and operating Zanzibar is one of scale. So, that authorization decision is fairly easy to implement if you have, say, 10,000 users and a million objects, you can do that on a laptop these days, probably, a database.

But when you tackle that problem for the scale of the planet, which is at the scale that Google often operates, billions of users, trillions of objects, it becomes actually quite difficult to build a consistent, correct high-performant, highly available system that does authorization for all of those users and all of those objects, all of those permissions, also trillions of permissions. That's what Zanzibar solves and, again, the requirements are quite stringent. So, the system has to be very highly available, very reliable because, if the system like Zanzibar goes down, then all its clients that rely on it for that authorization decision need to assume the answer to every authorization question is no while the system is down because privacy is at stake. And so, if the system doesn't respond, then it's essentially a huge denial of service until the system comes back so availability is very, very important.

And in addition, the system, of course, has to be correct and consistent because, again, privacy is at stake. So, if someone removes a user from a permission list,

from an access control list, it is really important that other person is really removed. So, if somebody doesn't want someone else to access their personal physical locations which are relevant to physical security in some people, then you really need to make sure that the system consistently and correctly removes that other person before reporting that it's removed. So, correctness is also very important. And then performance is important because the authorization decision is really just the beginning of the task for the product that's involved. Whether it be a photo serving product or a video serving product, it's going to ask that authorization decision early on when a user requests access to an object and the answer is, yes, you still have to do all the work of actually presenting the video or the photo to the user so the latency budget for the authorization system is quite small.

So, the system has to be fast, it has to be correct and highly available, and it has to handle a massive load. So, at Google, Zanzibar handled millions of authorization requests for that coming from all over the world. So, it turns into a very large scale reliable system that is not easy to build. So, that's where we put our effort and it was a lot of fun working on it. And yeah, there's a lot writing on it, as you said, at the heart it is protecting privacy by ensuring that only people who have permission to access digital objects are able to do so.

Bruke Kifle: So, scalability, availability, correctness performance, it seems like a stressful and significant undertaking with, as you described it, high consequence of action because this authorization system is what enables access to these core products, these core services that people rely on Google authorization mechanisms in place to prevent undesired access of their data. So, it seems like quite the stressful undertaking but, hopefully, it was an enjoyable journey and a great sense of satisfaction and a pretty meaningful contribution as well.

Ramon Caceres: Yeah, you hit it on the head. As a team, we took our responsibility very, very seriously and the engineering team was relatively small, never more than 10 people, it was four or five when I started on it but we also rely very much on the operations side on what Google called the system reliability engineers. There's a whole team of people keeping an eye on the system's operation 24 hours a day and responding within minutes, within five minutes, if there's any problem, even just an early warning of a problem. And so, it's a constant attention to detail and attention to make sure the system is healthy and attacking problems early, early on before they bring down the system which it never did come down so it's very highly available. We were very proud of the availability among all the other objectives that you mentioned.

Bruke Kifle: Wow.

Ramon Caceres: So, [inaudible 00:16:04] availability over years of operation. But it does take a realization that it's important to people's lives and we just build the engineering team and the site reliability engineering team around those goals and making sure that it's not overlooked again.

Bruke Kifle: Very interesting. I would love to talk about one of your other areas of focus. I know, some time back, your work in introducing cloudlets to support this idea of latency-sensitive or resource-intensive mobile applications was impactful and, obviously, still, in some form, relevant today. Considering the constantly evolving computing paradigm and the emergence of various computing architectures, edge computing architectures, technologies, what are your thoughts or how do you perceive the future as it relates to mobile and edge computing in the next, say, five or 10 years?

Ramon Caceres: Yeah. So, I think that the union of artificial intelligence with mobile and edge computing is growing and important. So, the ability to do significant machine learning at the edge of a network or even in the mobile devices themselves is going to become increasingly important. So, this is something that's already becoming possible, just in the last few years, our phones can now do speech recognition of your speech and interpretation of that speech and act on commands and do all of that processing locally on your phone, this is something that wasn't possible 10 years ago. I think there are many, many interesting and important applications that will be enabled by this marriage of machine learning with mobile and edge computing so that's something that I look forward to seeing evolve in the next few years.

Bruke Kifle: And where do you see us, I think we discussed some of these, but joining these two areas of focus, right, on the privacy and security piece but also the capabilities of what we can unlock with mobile and edge computing? Where do you see these as core enablers for interesting directions in AI? Privacy preserving, I think you described some of the capabilities that now we're able to harness at the edge, but do you see some of these benefits bringing these two areas of focus together as well?

Ramon Caceres: Yes, I think that one of the areas of work that I've been very interested in following, although I haven't done any work personally on it, is federated machine learning which is the idea of, instead of sending training data to a central location like a data center, you keep the training data local to where it's being generated, a mobile device, for example, and federating or breaking up the machine learning task into a distributed task where each mobile device does a piece of the learning task with the data that's local to it and belongs to the user of that mobile device, presumably, and this architecture has very strong privacy preserving properties and it's been worked on for years now.

I think it's coming to a head, it's being used and it has many practical applications and, again, I really like the privacy preserving property of it. So, that brings together artificial intelligence, mobile and edge computing and privacy. And so, I think that's an important development that we'll see much more of in the coming years.

Bruke Kifle: Certainly, yeah. I think, alongside some very core tenets, I know privacy and security are foundational pillars of how we think about AI development and

deployment. And so, I think a lot of the work around federated learning has been huge advancement to leveraging some of these core capabilities that you described but also ensuring that we uphold these core tenets or principles of ensuring privacy and safeguarding user data.

ACM ByteCast is available on Apple Podcasts, Google Podcasts, Podbean, Spotify, Stitcher and TuneIn. If you're enjoying this episode, please subscribe and leave us a review on your favorite platform.

I think these have been very interesting discussions on core technical areas of contribution over the course of your career, I would like to turn to something quite interesting that I actually saw on your website which is your very impressive sailing passion and background and competitive experiences as well which is quite interesting. I'm curious, could you share a bit more about your passion for sailing and, perhaps, how this pursuit has influenced your life or even informed your scientific interests and contributions?

Ramon Caceres: Yeah, thanks for asking. So, I've always loved the sea and boats. I grew up right on the coast of the Dominican Republic with the Caribbean Sea as my backyard, quite literally, maybe 30 yards from my home. And so, I did a lot of fishing growing up from the shore but also on boats and spearfishing and so I've always loved boats and I found sailing relatively late in life compared to many people. In my 20s, there were no sailboats where I grew up, there were no pleasure craft of any kind, it was all working fishing boats. But a colleague in my first job in Silicon Valley had a beautiful 35-foot sailboat in the San Francisco Bay, which is a wonderful place to sail, really good wind there, and he brought me onto his boat and taught me a few things and I've been very passionate about sailing since then, that was in my mid-20s.

And I love sailing especially because of how peaceful it is. You bring up the sails, you turn off the engine and it's very, very quiet. You're just relying on the wind to drive you and that quiet really brings a peacefulness and I think it's very conducive to creative thought, actually. You're asking me about connections to work, it's not so much that I do it for work but I do get a lot of good ideas when my mind clears out there and so I can get a lot out of that that way as well. Just not thinking about anything else but sailing and, eventually, it's very nice to just clear your mind that way.

I like the competitive nature of it, I do a lot of racing on sailboats and that is not as relaxing but it actually focuses your mind even more. So, it does clear your mind of everything else but very restful that way. If you've been working hard on something and pursuing a problem for a long time, debugging something difficult and you go for a sailboat race, you can't afford to be thinking about work when you're racing and so that does really clear your mind of everything. Again, I think, in the end, very conducive to creative thought because you're no longer ... Your mind just gets a rest.

Bruke Kifle: I think that-

Ramon Caceres: So, yeah, I think that may be one small connection.

Bruke Kifle: Yeah, no, I think it's great. I think I'm generally also an advocate that we should all have, outside of our core pursuits, professional pursuits, career pursuits, academic pursuits, have these passions that we're able to pursue. And so, it's great to see that it's an opportunity for you to get that sense of peace and relaxation but it also serves as an opportunity for creative thoughts. So, it's great that it's achieving multiple objectives but that's quite exciting. I think the next thing that I actually wanted to get your thoughts on is you have this very rich experience, of course, in academia, in industry, spanning both research but also large scale, highly scalable production products that are used by real users but you've also navigated these various roles and responsibilities as an individual contributor, as a leader.

How do you approach or how do you think about approaching this gap between theory and practice? You described early in your career that you worked as a software engineer, went back to graduate school, pursued multiple hats as a researcher, as a startup leader, as an engineer. So, what lessons have you learned from these rich experiences and how would you offer this as advice to aspiring researchers and innovators?

Ramon Caceres: Yeah, but partly it's a personal taste. I'm attracted both to research and to product work. Some people are quite happy to live in either of those worlds for their whole career, and that's great, but I got a taste for both early on and I've always ... I wanted to do both. And so, sometimes, when I am doing only one of them for too long, I miss the other one and that's led me to switch jobs and there are many situations where you're not able to do both in my experience. So, when I'm doing only one, I miss the other one and that switching back and forth has kept me very happy and engaged. It's perhaps not the fastest way to advance a career because of too much switching, there's a cost to switching jobs and focus from product to research and back, there's always a start-up cost when you join a new company.

But there are situations where you're able to exercise both. So, I should say, for example, a company like Google, it faces hard enough problems, difficult enough problems. Very often, we talked about the scaling problems in Zanzibar, those are problems where it requires research, it requires research to solve. And so, you're both working on a production system, so you get the satisfaction of impacting people on a daily basis which is what I get out of product work and I miss it when I don't have it, but you also get the challenge of working on a hard problem for which you don't know the answer when you first come across it and you have to do research, you have to explore the problem and try different approaches and finally come up with a good solution.

And so, that mix of research and product work is ideal for me and I haven't always found that in a single job which is why I've switched back and forth but it is possible to find it.

Bruke Kifle: Very interesting. I think another thing that's quite interesting from your journey that you described is your upbringing, you're born and raised in the DR. I share a similar background, I was born in Ethiopia, moved to the US at a pretty young age, so not too similar but I think the core principles are the same. And I know, in your role outside of your core engagements in industry, you serve as a board member for the CRA Committee on Widening Participation in Computing Research, so this idea of really advocating for representation, inclusion, diversity in the field.

How has your personal journey motivated your interest in these kind of efforts? But then, more generally, how can we, as a computing industry or a computing society, create more opportunities for underrepresented groups? What are the things, the initiatives, the things that we have to do to actually promote diversity within the industry and the field?

Ramon Caceres: Yes, thanks for asking. So, yeah, actually, it began in the late 1980s when I went back to Berkeley for my PhD, I became involved and have been involved since in efforts to increase participation of women and underrepresented groups in computing. I became gradually aware of it, certainly, as I advanced from undergraduate to graduate school, it became apparent that there were not very many Latinos around me, for example. And I also became aware of the gross under-representation of women in computer science, especially as you move up in seniority, problem gets even worse. So, it's an ongoing struggle, it looked like we were making progress there in the early '90s and then there's been some steps back, unfortunately. It's very frustrating that we're still struggling with this but it's important for the field to have broad representation. I think the products that get created are better, they serve more people and serve them better, we should keep working to increase that representation.

I just came back from a workshop that I helped co-chair, organized by the CRA Committee on Widening Participation in Computing Research that you mentioned, CRAWP. So, it's the seventh time we put together a yearly workshop on, it's a mentoring workshop for individuals from groups that are underrepresented in computing such as African-Americans, Latinos, people with disabilities. There's also a sibling workshop that focused on women in computing research and we bring together, all expenses paid, in this case, it was about 150 students from a broad representation of schools, not just human groups, and we bring them together with about 30 senior researchers, academics and industry, for several days of mentoring talks and one-on-one sessions and it's one way to bridge that gap that many people feel in feeling comfortable, feeling like they belong at the graduate level, in this case, these are all graduate students.

So, there are other organizations dedicated to similar goals and we need to keep working on this because we've made a tiny bit of progress in the last 30 years and we need to make a lot more.

Bruke Kifle: Certainly. I think it's a never-ending process but, of course, I think there's all the research to back that diverse teams, as you said, create diverse products, are more effective, more productive but I think, starting as early as possible in the pipeline, as early as education, higher education, secondary education, primary education, I think can be the key to widening access for underrepresented groups. So, I think it's quite normal but also very encouraging to see some of your work to contribute to these initiatives as well.

Ramon Caceres: Yeah. One of the things I've learned from these efforts over many years is that it's very important to have role models, someone who looks like you, who you feel represents you. And in that sense, I'm very gratified to have been recognized recently by the ACM as an ACM fellow. As far as I've been able to determine, it looks like I'm the first ACM fellow from the Dominican Republic and, of course, it feels good personally to be recognized by it. It's also, I think important to serve as an example to show young people that it's possible to succeed in this field and so that encourages them to pursue their career goals is to see that someone has done it as well.

Bruke Kifle: 100%,100%. I think, even for me as someone who's early in my career, I think having folks like yourself to look up to is certainly a point of inspiration. And so, we take great pride in having you join as an esteemed ACM fellow and very excited to see how many more future fellows we're able to inspire through your leadership in your mentorship as well. So, very exciting. Looking ahead, I'd love to get some of your thoughts. I know, as we've discussed, you've had such a rich career.

We're seeing so many rapid developments in technology, whether it be from a compute point of view, whether it be from advancements in software and AI and applications, what are some emerging trends or challenges in computer science or in software engineering that you find particularly interesting or perhaps even concerning or challenging? And what is your call to action or encouragement for those who may be interested in contributing to addressing these challenges or opportunities?

Ramon Caceres: We've touched on some of this before but let me come at it from this point of view here that you just brought up. I think that the challenge of preserving privacy has been with us for a long time but it continues to be with us and it might be exacerbated by the increasing volumes of data we collect and also the advances in artificial intelligence and its great data needs for training. And so, we mentioned before, this combination of doing machine learning at the edge of the network and in mobile devices themselves and the area known as federated learning which keeps the training data at the point of creation, for example, mobile devices, that are sending it to a faraway data center for central

processing. I think that architecture, that structure is fundamentally more privacy preserving than sending data to a central location, keeping it federated, keeping it in the source, near the source where the devices still belong to the person who generates the data, for example, and doing the learning there without letting the raw data escape.

That has a very, very strong privacy preserving properties and I think that I find that whole area of work very interesting and promising. And I think, like I said before, it's important to build in privacy into your systems and having a structure of data processing that's fundamentally more privacy preserving than others is a good step forward. So, the challenge is there, it's going to get harder, the [inaudible 00:33:54], and I see hope in this federated learning in particular at the edges of the network.

Bruke Kifle: Well, I think you captured it well, there's a challenge but also great promise and opportunity. As we round off this chat, I'd love to give you an opportunity to share some pieces of what we call bites of either advice or guidance, there are many of those in the audience who are aspiring entrepreneurs, engineers, scientists. And so, as you think about your journey, as you described, the young boy who was interested in creating solutions to serve people and over the course of your rich experience in academia and industry, what pieces of guidance or recommendation or life advice would you like to share to those who are aspiring to reach your levels?

Ramon Caceres: Well, I think a very useful thing to realize which some people may not when they're still young and being formed in their careers is what I mentioned before, that even the most successful people in our field, or any field, have faced their own challenges, and it's not always this primrose path to success. People hit challenges, they have setbacks, they have to rethink what they want to do, change direction, I certainly changed direction a number of times as we've talked about several times, and I think that knowing that that is very common, if not the norm, should give people encouragement. You should always try to persevere, there will always be setbacks in careers and your work and life in general and just having the perseverance, the patience to take a step back and see if there's some other direction that will be more fruitful that you should pursue.

In my case, I went from electrical engineering to computer science, from hardware to software and gone back and forth from research to product development and a combination of both and I think that's brought me a great enjoyment always to pursue my interests and not be discouraged by, perhaps, hitting a wall at some point and feeling like I'm not being creative anymore in that particular pursuit and that changing direction allowed me to get new inspiration and new motivation. So, yeah, stay flexible and realize that other people have also had to change direction and take indirect paths to their careers and success.

Bruke Kifle: I think those are really great points. There's no one path to success and there will always be setbacks but focus on getting back up, persevering and being patient with yourself. So, I think those are wonderful pieces of advice to part with. I think it's been a very interesting discussion and it's been a great pleasure to have you on ByteCast, Ramon, thank you so much for joining us.

Ramon Caceres: Thank you very much for having me on the podcast, I enjoyed talking to you.

Bruke Kifle: ACM ByteCast is a production of the Association for Computing Machinery's Practitioner Board. To learn more about ACM and its activities, visit acm.org. For more information about this and other episodes, please visit our website at learning.acm.O-R-G/B-Y-T-E-C-A-S-T. That's learning.acm.org/bytecast.