How we created edge computing

Edge computing processes data on infrastructure that is located close to the point of data creation. Mahadev Satyanarayanan recounts how recognition of the potential limitations of centralized, cloud-based processing led to this new approach to computing.

Mahadev Satyanarayanan

In 2009, together with Victor Bahl, Ramón Cáceres and Nigel Davies, I published a paper entitled 'The case for VM-based cloudlets in mobile computing'. Today, the work is generally viewed as the founding manifesto of edge computing. At a time when there was widespread euphoria about the limitless possibilities of cloud computing, the paper put forward an alternative viewpoint. It argued that the extreme consolidation (the concentration of computing resources into a few large data centres) implicit in cloud computing would fundamentally limit its ability to sustain latency-sensitive and bandwidth-hungry applications that would emerge in the future. It also identified bandwidth scalability of cloud-based applications based on video data from sensors as a key concern. To support these future applications, the paper argued for a dispersed infrastructure of micro-datacentres called cloudlets, which avoids extreme consolidation while preserving cloud computing attributes such as multi-tenancy with strong isolation (in other words, the ability to concurrently run code from different parties that do not trust each other). With the recent emergence of edge computing, our proposal has become mainstream. But what led us to this idea?

In 1993, I wrote a short thought piece on the topic of mobile computing, a field that was just emerging at the time. This, I believe, is the first paper where the inherent resource poverty of mobile devices was identified as a key long-term constraint of mobile computing. As I wrote in the article: "Mobile elements are resource-poor relative to static elements. Regardless of future technological advances, a mobile unit's weight, power, size and ergonomics will always render it less computationally capable than its static counterpart. While mobile elements will undoubtedly improve in absolute ability, they will always be at a relative disadvantage." In the 25 years or so since that article, the prediction has remained consistently true.

To overcome this fundamental limitation, my colleagues and I at Carnegie Mellon University proposed a technique that is now widely used for many compute-intensive applications: a mobile device offloads heavy computations over a wireless network to a server that is much more powerful than the mobile device. The approach was first demonstrated in the Odyssey system (Fig. 1) in 1997. An important aspect of this implementation was Odyssey's ability to select the optimal execution mode (local, remote or hybrid) based on runtime factors such as current network bandwidth. Odyssey was thus the technical forerunner of today's mobile speech-to-text systems, as well as modern mechanisms for adaptive offload. Since 2001, this approach has also been known by the term cyber foraging and has been a key area of mobile computing research.

The emergence of cloud computing around 2007 both simplified and complicated offloading. On the one hand, the cloud was the natural answer to where offloaded execution should be performed. On the other hand, the likely distances to the cloud, necessary for the consolidation implicit in cloud computing, were problematic. End-to-end communication over a wide area network to a distant cloud involves many network hops and results in high round-trip times. Bottlenecks for network bandwidth are also likely.

By 2008, I was convinced that embracing cloud computing for offloading was a sterile strategy: it would never be able to sustain applications such as augmented reality, which were starting to emerge. I shared these concerns with researchers in mobile computing including Bahl of Microsoft Research, Roy Want who was then at Intel, Cáceres who was then at AT&T Research and Davies of Lancaster University. They agreed with my concerns, and expressed interest in exploring this topic more deeply. We met for a day and a half of brainstorming in October 2008, hosted by Bahl at the Microsoft Research office in Redmond, Washington. Many of the key themes of edge computing emerged from this meeting, including the concept of a cloudlet as a 'data centre in a box' and the concept of a tiered architecture with the cloud at tier 1, cloudlets at tier 2 and cloudlet-associated mobile devices at tier 3. To capture the ideas that emerged from the meeting, we wrote the 2009 paper.

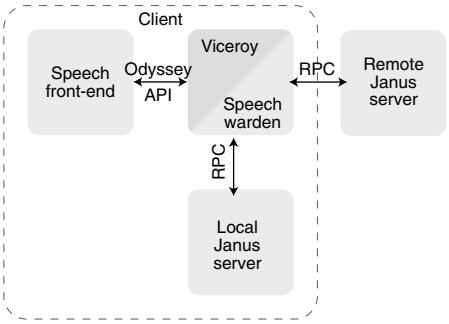

Fig. 1 | Offloading from a mobile device. The Janus speech-recognition application was modified to operate in one of three modes in Odyssey: local, remote and hybrid. Image reproduced from B. D. Noble et al., *Proc. 16th ACM Symp. Operating Systems Principles* 276–287 (Association for Computing Machinery, 1997). Viceroy is a system module described by B. D. Noble et al.

As we learned over the next few years, just writing a paper is, alas, not sufficient to convince many sceptics. It took numerous years of empirical measurements, and the implementation of applications that are critically dependent on low latency or bandwidth scalability, to overcome scepticism about the need for edge computing. By 2018, the seed that we had planted in 2009 had grown into industry-wide interest and activity. While edge computing is still in its infancy, there is no question that it is here to stay. ❐

Mahadev Satyanarayanan
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA.
e-mail: satya@cs.cmu.edu

Published online: 16 January 2019
https://doi.org/10.1038/s41928-018-0194-x