OEC Valedictory Event

April 22, 2022

Mahadev Satyanarayanan
School of Computer Science
Carnegie Mellon University

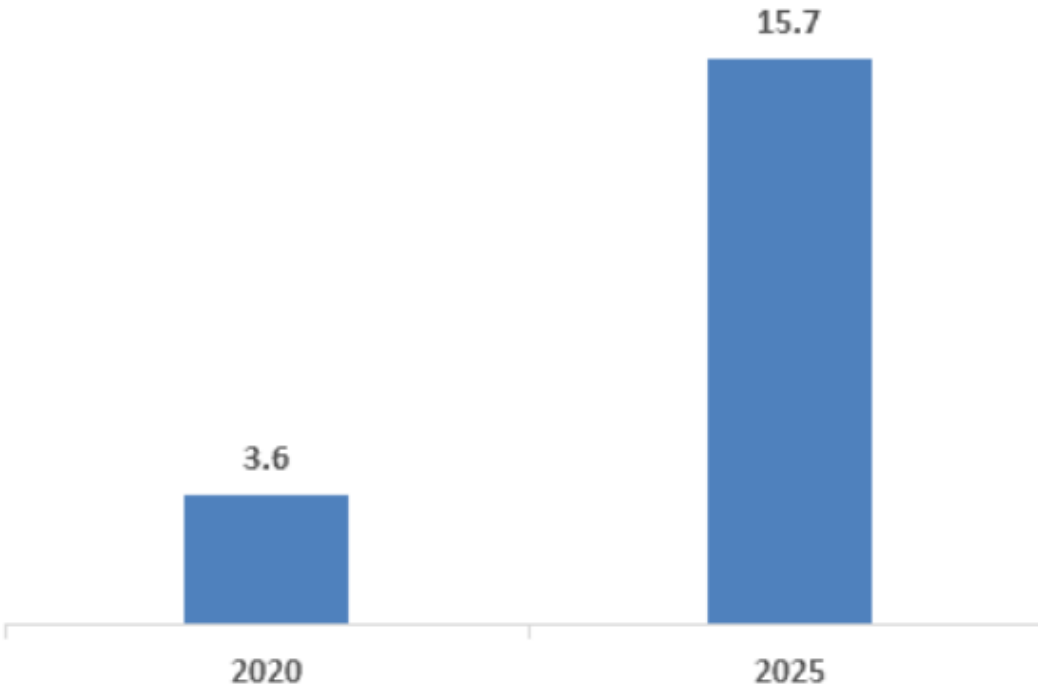
Growth of Edge Computing

Report Attribute	Details
Market size value in 2021	USD 6.29 billion
Revenue forecast in 2028	USD 61.14 billion
Growth rate	CAGR of 38.4% from 2021 to 2028
Base year for estimation	2020
Historical data	2017 - 2019
Forecast period	2021 - 2028

Source: GrandView Research, April 2021 (https://www.grandviewresearch.com/industry-analysis/edge-computing-market)

Edge Computing Market Size and Growth

According to MarketsandMarkets, the global edge computing market was valued at approximately $3.6 billion in 2020 and is expected to increase and reach $15.7 billion by 2025 at a Compound Annual Growth Rate (CAGR) of 34.1% during the forecasted period.



The global cloud computing market size was valued at USD 368.97 billion in 2021 and is expected to expand at a compound annual growth rate (CAGR) of 15.7% from 2022 to 2030.

Source: Grand View Research

How Did We Get Here?

Some Key Events

1993	Recognition of the “Mobility Penalty”
1997	Value of Offloading to overcome Mobility Penalty
2004	Publication of “Augmenting Cognition”
2006	Creation of AWS and Cloud Computing
2009	Publication of the case for edge computing
2015	Founding of the OEC Creation of ETSI MEC standards group
2018	Microsoft (Satya Nadella) embraces the “Intelligent Edge”
2019	Creation of AWS Wavelength
2021	Vodafone offers Edge Computing as a service

Nearly 30 Years Ago

IEEE Computer, September 1993

HOT TOPICS

Editor: Ronald D. Williams, University of Virginia, Thornton Hall/Electrical Engineering, Charlottesville, VA 22903; phone, (804) 924-7960; fax, (804) 924-8818; Internet, rdw@virginia.edu

Mobile computing

M. Satyanarayanan, Carnegie Mellon University

In his book *Mind Children*, my colleague Hans Moravec draws an analogy between the seminal role of mobility in the evolution of biological species and its similar impact upon the development of computing systems.[1] Although Hans' comments are directed at robotic systems, his observation applies equally well to a much broader class of distributed computing systems involving mobile elements. Mobility will influence the evolution of distributed systems in ways that we can only dimly perceive at present.

The recent proliferation of portable computers, in conjunction with nascent high- and low-bandwidth cordless networking technology, will soon provide a pervasive hardware base for mobile computing. The stage is set for portable computing devices of every variety imaginable, from the handheld to the wearable. Telecommuting will be truly ubiquitous.

This exciting possibility brings with it new technical challenges. Mobile computing systems are constrained in important ways relative to static systems. Moreover, these constraints are intrinsic to mobility and are not just the shortcomings of current technology:

- *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.
- *Mobile elements are more prone to loss, destruction, and theft than static elements.* A Wall Street stockbroker is more likely to be mugged on the streets of Manhattan and have his or her laptop stolen than to suffer the loss of a workstation in a locked office through theft or tampering. Even if security isn't a problem, portable computers are more vulnerable to loss or damage.
- *Mobile elements must operate under a much broader range of networking conditions.* A desktop workstation can typically rely on local or wide area connectivity (LAN/WAN). A laptop in a hotel room has only modem or ISDN (Integrated Services Digital Network) connectivity. Outdoors, a laptop with a cellular modem may experience only intermittent contact with its nearest cell.

These constraints violate many of the assumptions upon which today's distributed systems are based. Further, the future ubiquity of portable computers will result in mobile computing systems that are much larger than the distributed systems of today. Scalability will thus be a continuing concern. Past experience has shown that scale has to be treated as a primary influence rather than an afterthought in the design of distributed systems.[2]

A key requirement of mobile computing systems will be the ability to access critical data regardless of location. Data from shared file systems and databases must be made available to programs running on mobile computers. For example, a technician servicing a jet engine needs access to the engineering details of that engine as well as to its past repair records; similarly, a businessman who is continuing his work on the train home needs access to his business records. For an even more complex and potentially life-threatening problem, imagine an emergency medical team responding to a case of poisoning: it would need rapid access to a medical database describing poison symptoms and antidotes, as well as to the patient's medical records to determine drug sensitivity.

The need to access shared data implies interdependence between the elements of a mobile computing system. At the same time, the need for robustness when encountering network and remote site failures requires clients to be as autonomous as possible. By its very nature, then, mobility exacerbates the tension between autonomy and interdependence so characteristic of distributed computing.

In principal, mobility should be completely transparent to users. Transparency eliminates the need to be constantly aware of one's computing environment, thus allowing the user to focus on the real tasks at hand. Adaptation to a changing computing environment should be initiated by the system rather than by the user. Although perfect transparency is unattainable, that should not deter us from striving to come as close as possible to that ideal.

The Coda file system, built by my research group at Carnegie Mellon University, represents an initial step in providing such transparency.[3] Coda facilitates the use of shared data in mobile computers by simplifying prefetching of files, allowing autonomous operation while disconnected, and transparently reintegrating changes upon reconnection. Work is under way to extend the system to exploit weak connectivity (low-bandwidth, intermittent, or both) when available. Since disconnected operation allows Coda to operate at zero bandwidth, we view weak connectivity as an opportunity rather than a problem. Our approach to exploiting weak connectivity is to use it for trickle-charging or discharging a mobile client's cache.

The design of Coda extends the file access paradigm of the Andrew file system to mobile environments.[4] Coda completely hides mobility, so an application cannot tell whether it is connected or disconnected from servers. Coda has the further advantage of being binary compatible with applications written for the Berkeley Software Distribution (BSD) Unix interface. Thus, for an important class of common applications, Coda achieves full transparency.

But the strategy of insulating applications from the limitations inherent

September 1993 81

Mobile computing

M. Satyanarayanan, Carnegie Mellon University

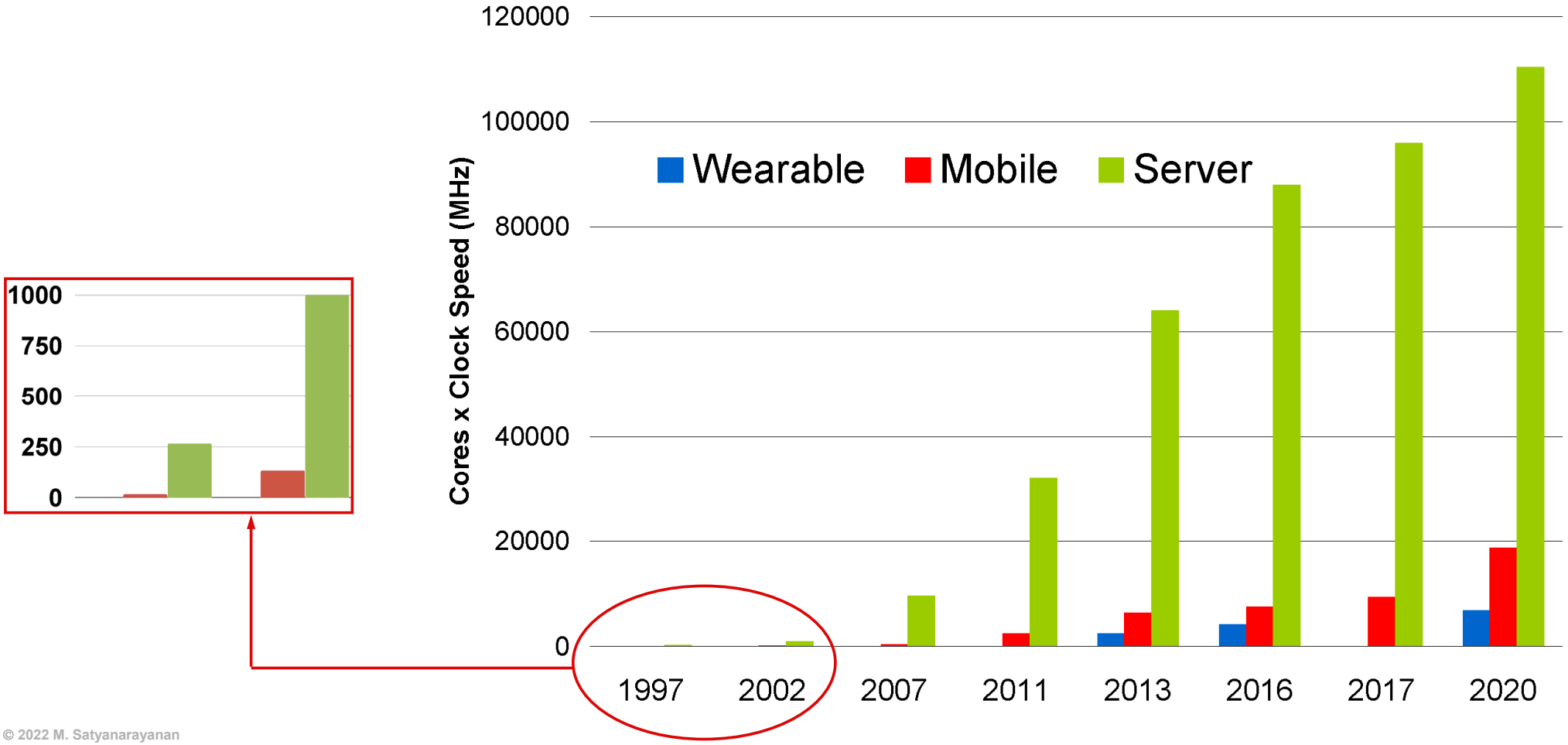
• *Mobile elements are resource-poor relative to static elements.* Regardless of future technological advances, a mobile unit's weight, power, size, and ergonomics will always render it less computationally capable than its static counterpart. While mobile elements will undoubtedly improve in absolute ability, they will always be at a relative disadvantage.

Moore's Law works differently for mobile and static computers

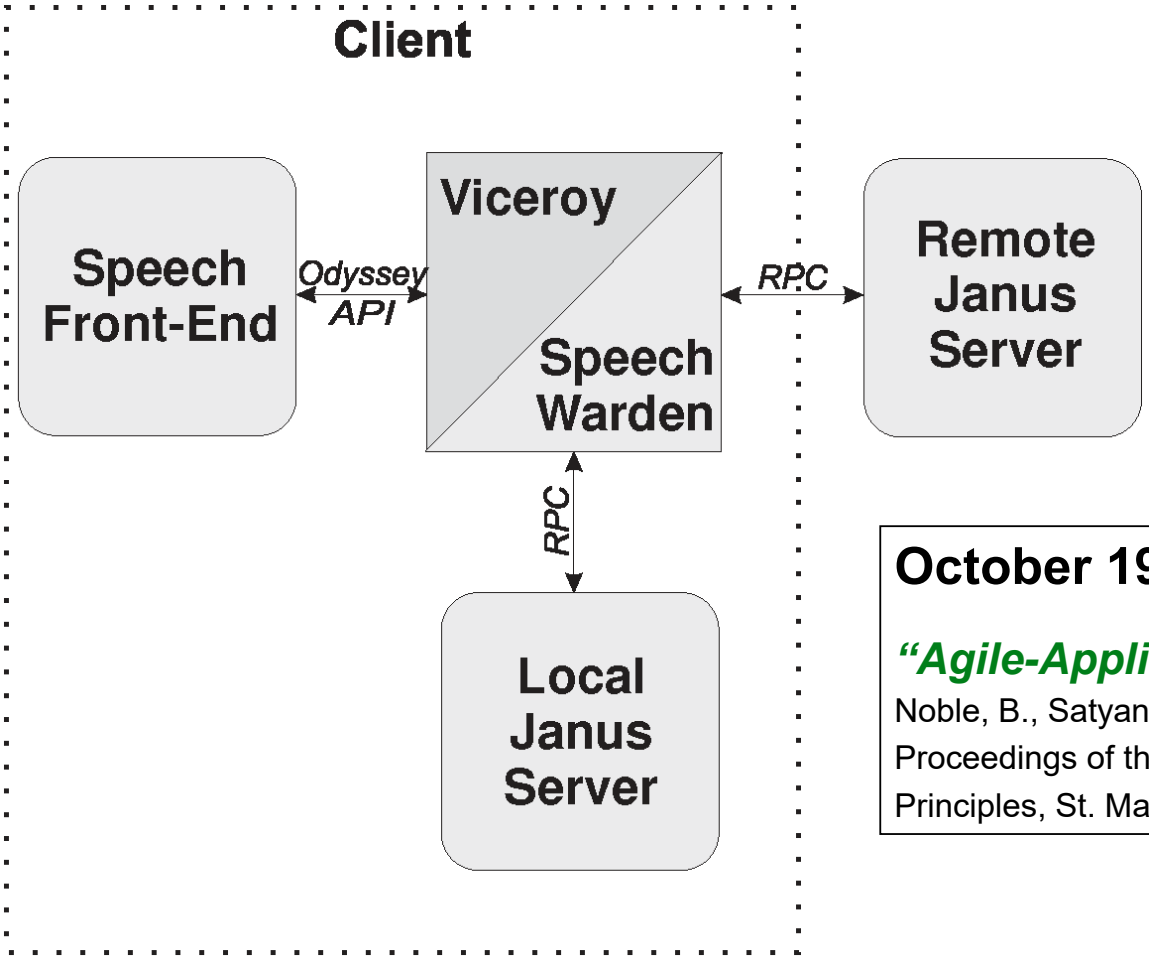
September 1993

Is This Still True ~30 Years Later?

YES! The *"mobility penalty"* is very real and persistent



Offload Computation from Device to Server

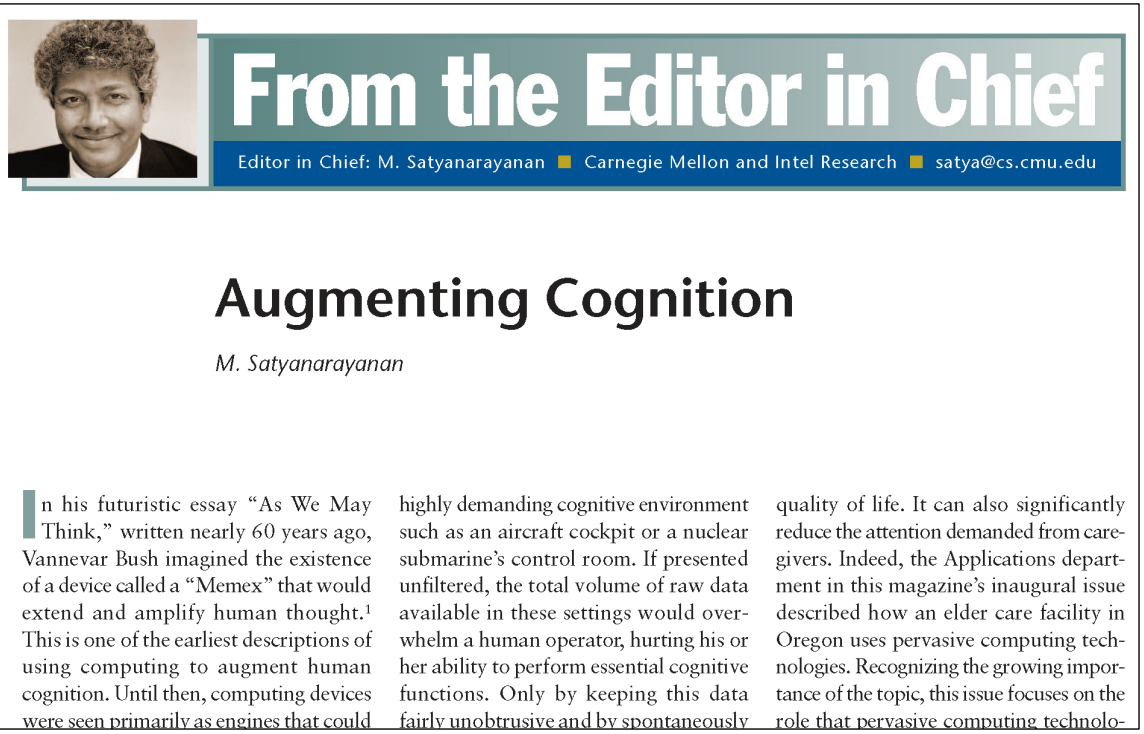


October 1997 (10 years before iPhone and Siri)

"Agile-Application Aware Adaptation for Mobility"

Noble, B., Satyanarayanan, M., Narayanan, D., Tilton, E., Flinn, J., Walker, K. Proceedings of the 16th ACM Symposium on Operating Systems Principles, St. Malo, France, October 1997

2004 Thought Piece in IEEE Pervasive Computing



From the Editor in Chief

Editor in Chief: M. Satyanarayanan ■ Carnegie Mellon and Intel Research ■ satya@cs.cmu.edu

Augmenting Cognition

M. Satyanarayanan

In his futuristic essay "As We May Think," written nearly 60 years ago, Vannevar Bush imagined the existence of a device called a "Memex" that would extend and amplify human thought.[1] This is one of the earliest descriptions of using computing to augment human cognition. Until then, computing devices were seen primarily as engines that could highly demanding cognitive environment such as an aircraft cockpit or a nuclear submarine's control room. If presented unfiltered, the total volume of raw data available in these settings would overwhelm a human operator, hurting his or her ability to perform essential cognitive functions. Only by keeping this data fairly unobtrusive and by spontaneously quality of life. It can also significantly reduce the attention demanded from caregivers. Indeed, the Applications department in this magazine's inaugural issue described how an elder care facility in Oregon uses pervasive computing technologies. Recognizing the growing importance of the topic, this issue focuses on the role that pervasive computing technolo-

"For example, imagine a wearable computer with a head-up display in the form of eyeglasses and with a built-in camera for continuous face recognition. This would offer the essentials of an augmented-reality system to aid cognition. When you look at a person, his or her name could pop up, possibly with additional cues to guide your greeting. Such "magic glasses" could transform your environment."

Where to Offload?

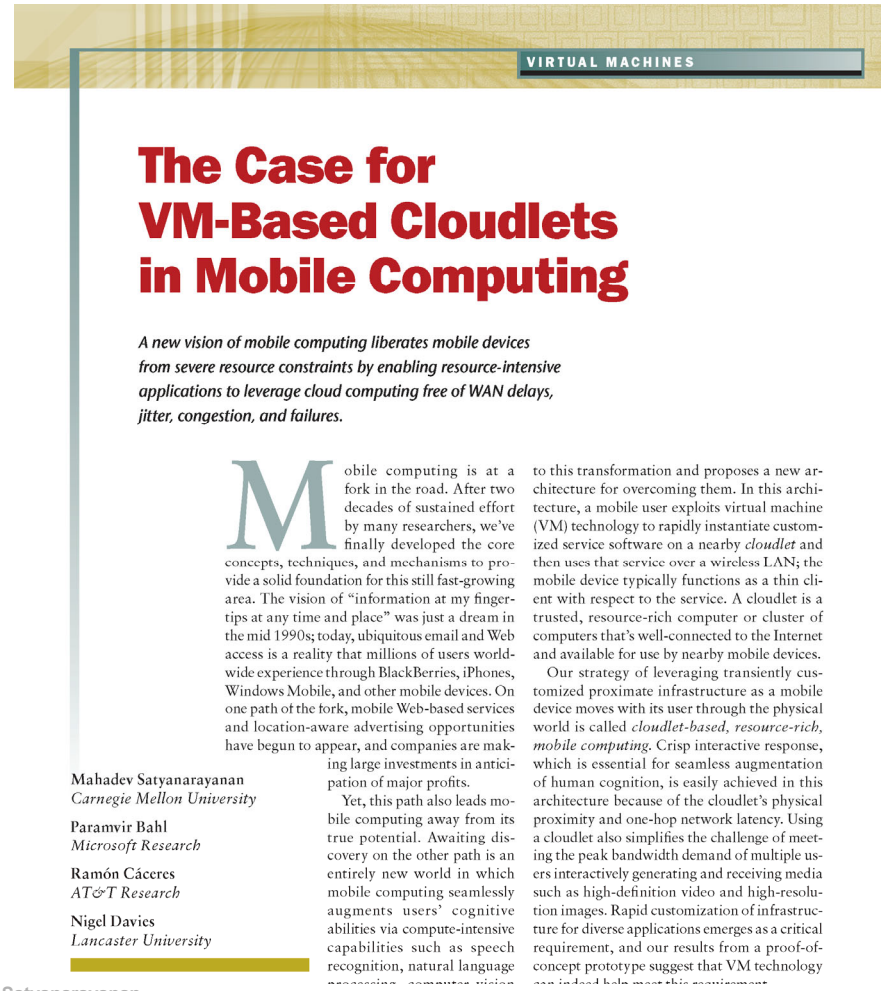
The Cloud Emerges (~2006-2010)
(along with exascale data centers & CDNs)

***Consolidation*, *Economies of Scale*, and *OpEx for Capex* are key themes**

At what price?

The price is *end-to-end latency of offload*

Edge Computing: Low-Latency Offload



VIRTUAL MACHINES

The Case for VM-Based Cloudlets in Mobile Computing

A new vision of mobile computing liberates mobile devices from severe resource constraints by enabling resource-intensive applications to leverage cloud computing free of WAN delays, jitter, congestion, and failures.

Mahadev Satyanarayanan
Carnegie Mellon University

Paramvir Bahl
Microsoft Research

Ramón Cáceres
AT&T Research

Nigel Davies
Lancaster University

Mobile computing is at a fork in the road. After two decades of sustained effort by many researchers, we've finally developed the core concepts, techniques, and mechanisms to provide a solid foundation for this still fast-growing area. The vision of "information at my fingertips at any time and place" was just a dream in the mid 1990s; today, ubiquitous email and Web access is a reality that millions of users worldwide experience through BlackBerries, iPhones, Windows Mobile, and other mobile devices. On one path of the fork, mobile Web-based services and location-aware advertising opportunities have begun to appear, and companies are making large investments in anticipation of major profits.

Yet, this path also leads mobile computing away from its true potential. Awaiting discovery on the other path is an entirely new world in which mobile computing seamlessly augments users' cognitive abilities via compute-intensive capabilities such as speech recognition, natural language

to this transformation and proposes a new architecture for overcoming them. In this architecture, a mobile user exploits virtual machine (VM) technology to rapidly instantiate customized service software on a nearby *cloudlet* and then uses that service over a wireless LAN; the mobile device typically functions as a thin client with respect to the service. A cloudlet is a trusted, resource-rich computer or cluster of computers that's well-connected to the Internet and available for use by nearby mobile devices.

Our strategy of leveraging transiently customized proximate infrastructure as a mobile device moves with its user through the physical world is called *cloudlet-based, resource-rich, mobile computing*. Crisp interactive response, which is essential for seamless augmentation of human cognition, is easily achieved in this architecture because of the cloudlet's physical proximity and one-hop network latency. Using a cloudlet also simplifies the challenge of meeting the peak bandwidth demand of multiple users interactively generating and receiving media such as high-definition video and high-resolution images. Rapid customization of infrastructure for diverse applications emerges as a critical requirement, and our results from a proof-of-concept prototype suggest that VM technology

Published in October 2009

CMU, Microsoft, AT&T, Intel, Lancaster University authors

Now seen as the "founding manifesto" of Edge Computing

Rejected NSF Expeditions Proposal (2009)

"Many panelists do not agree with the premise of the proposal in which distant cloud computing incurs too high latency to be acceptable by mobile applications. They question the validity of such assumption as the proposal provides no real data to justify it."

Panel Summary

Creation of the OEC Initiative

(June 10, 2015)

Edge Computing Gains Momentum

- 2018: Satya Nadella (Microsoft) speaks of the "Intelligent Cloud and Intelligent Edge"
- 2019: AWS announces Wavelength
- 2020: Microsoft announces Azure for Operators BU
- 2021: Vodafone offers Edge Computing as a service

Over to Rolf

"The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it."

Mark Weiser, 1991

We are not there yet with Edge Computing

Your work, and ours, over the next many years will help us get there

“Success Has a Thousand Fathers”

All of the OEC companies: technical expertise, financial gifts, equipment, …

Federal funding agencies: DARPA, NSF

My research team at CMU: especially Jan Harkes & Tom Eiszler

Jim

Rolf

Thank You!