Mobile Computing: the Next Decade

Mahadev Satyanarayanan
School of Computer Science
Carnegie Mellon University

1 Introduction

"Information at your fingertips anywhere, anytime" has been the driving vision of mobile computing for the past two decades. Through relentless pursuit of this vision, spurring innovations in wireless technology, energy-efficient portable hardware and adaptive software, we have now largely attained this goal. Ubiquitous email and Web access is a reality that is experienced by millions of users worldwide through their BlackBerries, iPhones, Windows Mobile, and other portable devices. Continuing on this road, mobile Web-based services and location-aware advertising opportunities have begun to appear, triggering large commercial investments. Mobile computing has arrived as a lucrative business proposition.

What will inspire our research in mobile computing over the next decade and beyond? We begin by considering two hypothetical mobile computing scenarios from the future. We then extract the deep assumptions implicit in these scenarios, and use them to speculate on the future trajectory of mobile computing. We conclude that there are really *two* fundamentally distinct strategies at play, and that the dialectic between these strategies will largely shape the mobile computing landscape of the future.

2 Scenario 1: Lost Child

Five-year old John is having a wonderful time with his family at the Macy's Thanksgiving Day parade in Manhattan. Mid-way through the parade, John sees a group of friends in the crowd nearby. He shows his parents where his friends are, and tells them he is going over to meet them. Since his parents see responsible adults in the group, they are fine with John walking over to see his friends. An hour later, John's parents walk over to where they expect to find him. To their shock, they discover that the friends have not seen John at all. He has been missing for an entire hour now, and John's parents are very concerned. Searching for a lost child in a Manhanttan crowd is a daunting task.

Fortunately, a police officer nearby is able to send out an amber alert via text message to all smartphone users within two miles. He requests them to upload all photographs they may have taken in the past hour to a secure web site that only the police can view. In a matter of minutes, the web site is populated with many photographs. New photographs continue to arrive as more people respond to the amber alert.

With John's parents helping him, the police officer searches these photographs with an application on his smartphone. His search is for the red plaid shirt that John was wearing. After a few pictures of Scottish kilts in the parade, a picture appears that thrills John's parents. In a corner of that picture, barely visible, is a small boy in a red shirt sitting on the steps of a building. The police officer recognizes the building as being just two blocks further down the parade route, and contacts one of his fellow officers who is closer to that location. Within moments, the officer is with the boy. John is safe now, but he has a lot of explaining to do . . .

3 Scenario 2: Disaster Relief

The Big One, measuring 9.1 on the Richter scale, has just hit Northern California. The entire Bay Area is one seething mass of humanity in anguish. Many highways, power cables and communication lines are severely damaged. Disaster on such a scale has not been seen since World War II. With limited manpower, unreliable communication and marginal transportation, disaster relief personnel are stretched to the limit. Internet infrastructure, in-

cluding many key data centers, have been destroyed. The Googleplex has been reduced to a smoking hulk. In spite of heroic efforts, disaster relief is painfully slow and hopelessly inadequate relative to the scale of destruction.

Sudden obsolescence of information regarding terrain and buildings is a major contributor to slow response. Vital sources of knowledge such as maps, surveys, photographs, building floor plans, and so on are no longer valid. Major highways on a map are no longer usable. Bridges, buildings, and landmarks have collapsed. GoogleEarth and GoogleMaps are now useless for this reqion. Even the physical topography of an affected area may be severely changed. Conducting search and rescue missions in the face of obsolete information is difficult and dangerous. New knowledge of terrain and buildings has to be reconstructed from scratch at sufficient resolution to make important life and death decisions in search and rescue missions.

In desperation, the rescue effort turns to an emerging technology: camera-based GigaPan sensing. Using off-the-shelf consumer-grade cameras in smartphones, local citizens take hundreds of close-up images of disaster scenes. Transmission of these images sometimes occurs via spotty low-grade wireless communication; more often, the images are physically transported by citizens or rescue workers. The captured images are then stitched together into a zoomable panorama using compute-intensive vision algorithms. To speed up the process, small GigaPan robots that can systematically photograph a scene with hundreds of close-up images are air-dropped over the area for use by citizens.

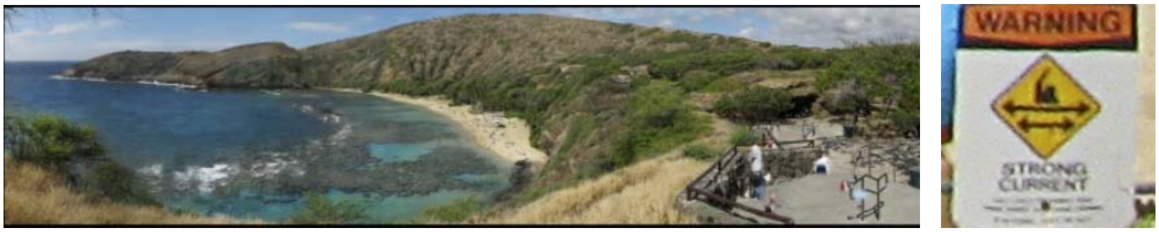
Slowly and painstakingly, detailed maps and topographical overlays are constructed bottom-up. As they become available, rescue efforts for those areas are sped up and become more effective. Rescuing trapped people is still dangerous, but at least the search teams are now armed with accurate information that gives them a fighting chance ...

4 Reflecting on these Scenarios

These scenarios embody a number of themes that will be central to the evolution of mobile computing over the next decade. We explore these themes next.

Common to both scenarios is the prominent role of mobile devices as *rich sensors.* While their computing and communicaton roles continue to be important, it is their rich sensing role (image capture) that stands out most prominently in these scenarios. We use the term "rich" to connote the depth and complexity about the real world that is being captured. This is in contrast to simple scalar data such as temperature, time and location that are involved in typical sensor network applications. When cell phones with integrated cameras first appeared, people wondered if they represented a solution in search of a problem. Would mobile users take so many photographs that this capability was worth supporting? Today, the value of this functionality is no longer questioned. Tomorrow the roles will be reversed: people will wonder why any digital camera lacks the wireless capability to transmit its images. Video capture, leading to even richer sensing and recording of the real world is also likely to gain traction.

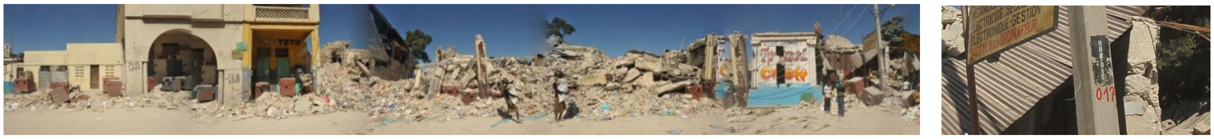
A second emergent theme is that of *near-real-time data consistency.* This is most apparent in the lost child scenario, where the only useful images are very recent ones. Pictures taken before the child was lost are useless in this context. Recency of data is also important in the disaster relief scenario. A major earthquake is often followed by aftershocks for hours or possibly days. These aftershocks can add to the damage caused by the original quake, and in some cases be the "tipping point" that triggers major structural and topographical changes. Regions that have already been mapped after the original quake may need to be remapped. The need for near-real-time data consistency forces rethinking of a long line of work in mobile computing that relates to the use of prefetching and hoarding for failure resiliency. The core concepts behind those techniques may still be valuable, but major changes in their implementations may have to be developed in order to apply them to the new context. In the disaster relief scenario, for example, many old maps and photographs may still be valid if the buildings and terrain involved have only been minimally affected. However, discovering whether it is acceptable to use hoarded information about them is a challenge. No central authority (e.g. a server) can answer this question with confidence. Only an on-the-spot entity (e.g. a user with a mobile device) can assess whether current reality is close enough to old data for safe reuse. That de-



(a) Panorama

(b) Full Zoom

Figure 1: GigaPan Image of Hanuama Bay, Hawaii (May 19, 2008)



(a) Panorama

(b) Full Zoom

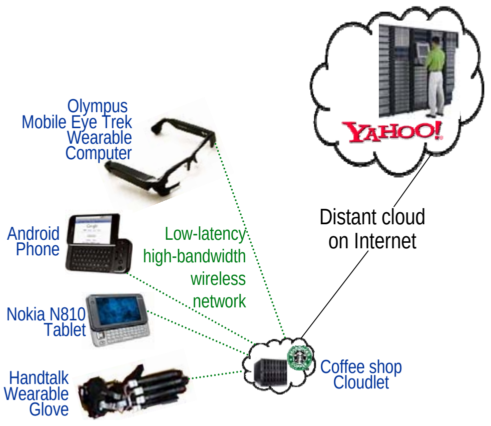
Figure 2: GigaPan Image of Downtown Port Au Prince, Haiti (January 29, 2010)

termination may involve human judgement, possibly assisted by software (e.g. a program that compares two images to estimate disruption).

A third emergent theme is that of *opportunism.* This is most evident in the lost child scenario. The users who contribute pictures were completely unaware of their potential use in searching for the lost child. They took the pictures for some other reason, such as a funny float in the parade. But because of the richness of the sensed data, there are potentially "uninteresting" aspects of the image (e.g. small child in the corner of the picture) that prove to be very important in hindsight — it is context that determines importance. Although the theme of opportunism also applies to simpler sensed data (e.g., anti-lock braking devices on cars transmit their GPS coordinates on each activation, enabling a dynamic picture of slick spots on roads to be obtained by maintenance crews), the richness of captured data greatly increases the chances for opportunistic reuse. An airport video image that was deemed uninteresting on 9/10/2001 may prove to be of high interest two days later because it includes the face of a 9/11 hijacker. With such opportunism comes, of course, many deep and difficult questions pertaining to privacy. While these questions already exist today with mining data from surveillance cameras, they will grow in frequency and significance as mobile users increasingly contribute their rich sensed data. One can easily imagine a business model that provides small rewards for contributors of such data, while reaping large profits by mining aggregated data for customers.

A final emergent theme is the need to broaden our definition of "mobile computing" to embrace developments that lie well outside our narrow historical concerns. Examples include non-indexed image search in the lost child scenario and GigaPan technology in the disaster relief scenario. These may feel like science fiction, but they are reality today.

For example, consider GigaPan technology. Figure 1(a) shows a 5.6 gigapixel panorama that has been stitched together from 378 individual images captured with a consumer-grade digital camera. The software available for navigating such an image allows a user to probe the panorama at very high zoom levels, much like GoogleEarth. This image, and many others, can be explored at the GigaPan web site (http://www.gigapan.org) [5]. The level of detail can be astonishing. For example, Figure 1(b) shows a legible warning sign at a lifeguard station. In Figure 1(a), the entire lifeguard station is barely visible as a speck on the distant beach. Figure 2(a) is relevant to the disaster relief scenario. It shows a panorama stitched together from 225 individual images of downtown Port Au Prince, Haiti that were taken by a news reporter who was covering the earthquake relief effort. These images were stitched together after the reporter's return to the United States, since the stitching capability was not available at the disaster site. Figure 2(b) shows a zoomed-in view of damaged electrical infrastructure, including the ID number of the tower that has been destroyed. Imagine how valuable this sensing and mapping capability would be if it were available at large scale at a disaster site, very soon after the disaster strikes.



(a) Cloudlet Concept

	Cloudlet	**Cloud**
State	Only soft state	Hard and soft state
Management	Self-managed; little to no professional attention	Professionally administered, 24x7 operator
Environment	“Datacenter in a box” at business premises	Machine room with power conditioning and cooling
Ownership	Decentralized ownership by local business	Centralized ownership by Amazon, Yahoo!, etc.
Network	LAN latency/bandwidth	Internet latency/bandwidth
Sharing	Few users at a time	100s-1000s of users at a time

(b) Key Differences: Cloudlet vs. Cloud

Figure 3: Extending the Classic 2-level Mobile Computing Architecture to 3 Levels

5 Transient Infrastructure

Since birth, mobile computing has implicitly assumed a 2-level hierarchy. Originally, the two levels were identified as “servers” and “clients.” More recent terminology uses “cloud” to connote the computational and information resources represented by a collection of servers. Regardless of terminology, however, the 2-level concept is woven quite deeply into our thinking about mobile computing. The upper layer (“cloud” or “server”) is assumed to be well-managed, trusted by the lower layer, and free from concerns that are specific to mobility such as battery life and size/weight constraints.

Future architectures for mobile computing are likely to extend this 2-level hierarchy to at least one additional layer, possibly more. The case for an intermediate layer called a *cloudlet* was articulated in a recent paper [3]. In that work, the rationale offered for the architectural extension is low latency network communication to computational resources in order to enable a new genre of immersive mobile applications. Cloudlets are viewed as decentralized and widely-dispersed Internet infrastructure whose compute cycles and storage resources can be leveraged by nearby mobile computers. A natural implementation is to extend Wi-Fi access points to include substantial processing, memory and persistent storage for use by associated mobile devices.

A cloudlet can be viewed as a “data center in a box.” It is self-managing, requiring little more than power, Internet connectivity, and access control for setup. This simplicity of management corresponds to an appliance model of computing resources, and makes it trivial to deploy. For safe deployment in unmonitored areas, the cloudlet may be packaged in a tamper-resistant or tamper-evident enclosure with third-party remote monitoring of hardware integrity.

Figure 3(b) summarizes some of the key differences between cloudlets and clouds. Most importantly, a cloudlet only contains *soft state* such as cache copies of data or code that is available elsewhere. Loss or destruction of a cloudlet is hence not catastrophic. This stateless model leads to an important research challenge: how can a mobile device rapidly and safely customize a cloudlet for its specific use? A possible solution, based on dynamic virtual machine synthesis, is sketched in [3]. Other approaches may also need to be explored.

Although originally motivated by considerations of network latency, cloudlets have much broader relevance. In particular, they are relevant to both the scenarios presented earlier. The GigaPan approach relies on compute-intensive vision algorithms to stitch together a zoomable panorama from individual images. Under normal conditions, these algorithms can be executed in the cloud. However, cloud computing may be compromised in the aftermath of a disaster. The physical infrastructure necessary for good Internet connectivity may have been destroyed and it may be many days or weeks before these can be repaired. Limited Internet connectivity may be re-established soon after the catastrophic event, but there will be very high demand on this scarce resource from diverse sources: families trying to desperately learn and share information about the fate of loved ones, citizen reporters and professional journalists sharing videos, images, blogs, and tweets of the disaster area with the outside world, and disaster relief agencies coordinating their efforts with their home bases. Under these conditions, cloudlets may be needed to support cloud computing.

We envision opportunistic deployment of cloudlets in disaster relief. In the immediate aftermath of a disaster, before external IT supplies have arrived, any available hardware such as an undamaged desktop can be pressed into service as a cloudlet. A cloudlet can even be built around a high-end laptop, with its few hours of battery life being priceless prior to the arrival of emergency electrical generators. As IT supplies arrive, temporary cloudlets may be replaced by purpose-designed equipment.

Cloudlets also have relevance to the lost child scenario. In that scenario, the near-real-time image search will require extensive computation since pre-computed indexes are not available for the contributed images. Cloud computing is the obvious answer for this, but exactly where in the cloud to compute is an open question. The task involves submission of images from a lot of people in the immediate neighborhood of the lost child; the search results will also be viewed there. This suggests use of local infrastructure (i.e., a cloudlet) rather than distant infrastructure. Once the search is completed (successfully or unsuccessfully) the contributed images can be discarded. This fits well with the stateless model of cloudlets and their use as transient infrastructure.

6 Competing Design Strategies

So far, this paper has focused on how mobile computing today and tomorrow differs from the past. Amidst all this change, however, certain fundamental challenges of mobility have remained invariant since they were articulated over 15 years ago [2].

First, wireless connectivity is highly variable in performance and reliability. Many real-world factors hinder ubiquitous high-bandwidth wireless connectivity. For example, Wi-Fi connectivity in public spaces often requires a subscription or one-time payment to the service provider. In private spaces (such as inside a customer's premises), there may be organizational or access control reasons that prevent Wi-Fi connectivity. 3G or 4G connectivity has wider coverage, but offers signifiantly poorer bandwidth. Even these lower-bandwidth alternatives are sometimes unavailable within buildings. Finally, there are situations where wireless transmissions are forbidden: for example, during air travel.

Second, mobile hardware is necessarily resource-poor relative to static client and server hardware. Considerations of weight, size, battery life, ergonomics, and heat dissipation exact a severe penalty in processor speed, memory size, disk capacity, etc. For a user, a mobile device can never be too small, too light or have too long a battery life. While mobile hardware continues to evolve and improve, computation on mobile devices will always be a compromise. An additional obstacle is the slow pace of improvement in battery technology, especially when compared to Moore's Law.

The first challenge (uncertain connectivity) leads to a "Swiss Army Knife" design philosophy: try to cram as much functionality as possible into a compact design that is self-contained and as frugal as possible in resource usage. Unfortunately, this approach often compromises usability, just as the tools in a real Swiss Army Knife (such as knife, fork, can opener, and corkscrew) are poor substitutes for full-sized implementations. Miniscule displays and keyboards are especially challenging for mobile users, particularly in the context of a graying population. Unfortunately, the incentives of today's marketplace tend to reward itemizable functionality enhancements rather than improvements to more diffuse attributes such as usability.

The second challenge (resource poverty) combined with the limitations of the Swiss Army Knife approach will eventually lead to a very different design philosophy. Rather than relying exclusively on a self-contained mobile device, one can use that device to leverage other resources such as a distant cloud, a nearby cloudlet, or an interaction device such as a large display. We refer to this as a "wallet" design philosophy because it resembles the role of wallets in everyday life. A typical wallet contains things like cash, credit cards, and ID cards. None of these items are intrinsically valuable. Rather, their value lies in their ability to elicit useful goods and services on demand from the environment.

Using a large wall-mounted display to augment the small display of a mobile device is an intriguing possibility. Transient use of displays in public spaces was prophesized almost two decades ago by Weiser's seminal paper on ubiquitous computing [4]. Today, there is a convergence of hardware and software technologies that are relevant to this aspect of Weiser's vision. In the near future, we envision a typical mobile user walking up to a display and using it for tasks that benefit from substantial screen real estate (including collaborative tasks and games). Privacy-sensitive information can be presented to the user on the mobile device, augmenting the less sensitive information that is presented on the large public display. User interactions may also occur through the mobile device.

The future evolution of mobile computing systems will largely be driven by the dialectic between resource poverty and uncertain connectivity. Reconciling their contradictory demands will itself be a challenge. Only an adaptive system design that can dynamically switch between a "wallet" mode of operation and a "Swiss army knife" mode is likely to produce satisfactory results.

A counterpoint to the "resource-poor mobile device immersed in resource-rich surroundings" paradigm is the state of affairs in the developing world. There, the mobile device is often the most technologically advanced entity in its surroundings. This leads to unique opportunities for high impact, but also requires out-of-the box thinking. A good example is the CAM framework for secure document processing via mobile phones in the developing world [1]. The concept of embedding programs for processing paper documents directly on those documents as 2D bar codes, and using smartphones to decode and process these programs, is an innovation directly inspired by the challenges of the developing world. As the old saying goes, "Necessity is the mother of invention." It has never been more true than in mobile computing!

References

[1] PARIKH, T. Using Mobile Phones for Secure, Distributed Document Processing in the Developing World. *IEEE Pervasive Computing 4*, 2 (April-June 2005).

[2] SATYANARAYANAN, M. Fundamental Challenges in Mobile Computing. In *Proceedings of the ACM Symposium on Principles of Distributed Computing* (1996).

[3] SATYANARAYANAN, M., BAHL, V., CACERES, R., AND DAVIES, N. The Case for VM-based Cloudlets in Mobile Computing. *IEEE Pervasive Computing 8*, 4 (Oct-Dec 2009).

[4] WEISER, M. The Computer for the 21st Century. *Scientific American* (September 1991).

[5] WIKIPEDIA. Gigapan. `http://en.wikipedia.org/wiki/Gigapan` (Online, accessed 2010-03-04).