Editor: Yih-Farn Chen • *chen@research.att.com*

I'm Cloud 2.0, and I'm Not Just a Data Center

Emiliano Miluzzo • *AT&T Labs Research*

Cloud service providers invest significant effort into designing, building, and empowering cloud infrastructures. At the same time, technological advances are commoditizing small devices with powerful compute, storage, and communication capabilities at unprecedented scale. What if such devices could extend the boundaries of the traditional cloud model to form an even more flexible, resource-aware, and better-performing cloud?

Big and small corporations, startups, and academic labs share a common feature: their services are often deployed on top of cloud infrastructures rather than on dedicated in-house servers. Desirable properties such as scalability, elasticity, resilience, and infrastructure management delegation fuel today's cloud computing. Despite its distributed nature, the cloud remains a remote entity, accepting requests from clients and accomplishing tasks in return — including storing and processing data. This distant cloud-centric approach makes perfect sense for many scenarios. Web services, big data storage, long-term backup, large computing jobs, device synchronization, and data brokering platforms are all examples that can exist only within a cloud environment. But what if a complimentary approach to this model is possible? How would the cloud computing landscape change? Does it even make sense to think of an alternative model?

A New Model for the Cloud

Although the consensus is that the cloud should not and cannot be replaced, let's briefly review some current computing trends. Hardware miniaturization and technological advances have made unthinkable progress: smartphones and tablets offer compute and storage capabilities comparable to the desktop computers of just a few years ago. Wearable devices such as smart goggles, smart watches, and health monitors are on the rise; many will soon be integral to our lives. Our homes are becoming smarter and even more connected, with home automation systems receiving growing attention. Our cars are being linked to the Web at an incredibly fast pace and empowered with capabilities that extend the home and office experience on the go. Networked storage devices are getting smaller, more capable, and cheaper. In the meantime, services on these platforms are often designed to follow what I believe is a rather simplistic model: a thin-client approach, with most of the data analysis, processing, and storage largely delegated to the distant cloud.

I argue that this thin-client model is no longer adequate to describe an emerging computing landscape, where abundant resources could potentially be found not only within the cloud, but also at locations much closer to users. Some of these resources could reasonably be drawn from the ensemble of personal devices that commonly surround us, such as our own smartphones, tablets, set-top boxes, connected cars, home and portable storage devices, wearables, and special-purpose computing hardware.

I call this model Cloud 2.0, a new computing paradigm that expands the boundaries of the traditional cloud computing ecosystem to provide more advanced and efficient compute, storage, and media distribution channels (see Figure 1). Cloud 2.0 augments the traditional cloud computing framework with a more flexible and resource-aware design and revisits the thin-client approach that shapes many services that run on our personal devices. Such devices are often

 1089-7801/14/$31.00 © 2014 IEEE Published by the IEEE Computer Society

IC-18-03-BWRS.indd 73 10/04/14 7:39 PM

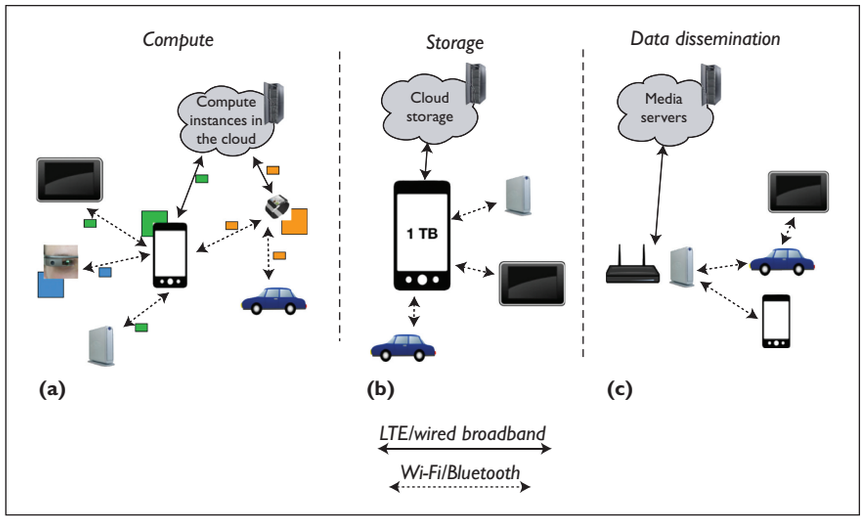

Figure 1. The Cloud 2.0 vision. (a) Compute tasks are distributed to nearby devices and the cloud. (b) The storage on a mobile device expands seamlessly by including local and cloud storage under a common namespace. (c) Media distribution occurs over home and public Wi-Fi networks, with media precaching on Wi-Fi routers.

underutilized in favor of more cloud-centric architectures. In reality, they are perfectly capable of embracing some of the compute and storage jobs that currently require constant cloud involvement — either on their own or via local inter-device cooperation.

This approach would help mitigate some of the shortcomings of an exclusively cloud-centric model. Local processing can reduce latency and boost quality of service (QoS) guarantees for many applications. Edge device participation in personal data storage can increase data privacy, allow faster data storage and retrieval, and reduce the monetary cost for storing data in the cloud by reducing traffic in the backbone network. At the same time, less traffic in the network promotes bandwidth savings and could reduce congestion in a scenario where a growing number of devices and sensors are network-enabled. Local device interaction for compute and storage jobs occurs over short-range radio technology, such as Wi-Fi and Bluetooth, using single-hop connections, whereas communication with the cloud is enabled via high-speed wireless data connections (such as LTE) or wired broadband (see Figure 1).

Here, I discuss how Cloud 2.0 can support some application domains in novel ways and allow applications to be resource-aware by balancing local versus remote computation. I also propose alternatives to data storage and data distribution services.

Mobile Computing

Smartphones and tablets are becoming powerful computers. Moreover, they support suites of sensors that can learn our activities, monitor our behavior, receive input, and provide suggestions for improved wellbeing (www.att.com/gen/press-room?pid=23974).[1,2] This trend will likely continue, with air quality, medical sensing (pulse oximeter, heart monitor, and so on), 3D or stereo cameras, electroencephalography (EEG) headsets, and radar/sonar being some potential sensors future mobile devices could support. With air quality sensing, municipalities could rely on 24/7 crowdsourced pollution-level measurements without needing specialized and costly measurement stations across town. Cameras with 3D or stereo capabilities will enable advanced, touch-free gesture recognition on mobile devices. Radar and sonar sensing will make people more aware of their surroundings — for example, incoming cars when walking and cycling — and issue alerts to possible perils. People with visual impairments could rely on radar or sonar sensing on their mobile devices to navigate in any environment. Smart glasses, with their onboard cameras and microphones, are expected to become mainstream along with other wearables to monitor, for example, vital signs on the move. Cars will be connected to the Internet, able to interact with each other (vehicle-to-vehicle, or V2V) and with road-side infrastructures (vehicle-to-infrastructure, or V2I) for safety purposes.[3]

By applying intelligence and machine learning reasoning to such large volumes of sensor data, we can interpret and infer patterns to actuate components, realize higher-level applications, and offer new forms of interaction, such as complex activity and context recognition, situation awareness, augmented reality, or voice commands. A common approach is to offload such data processing to the cloud, with consequent high traffic volumes in the cellular and backbone network and possible data latency increases. Given personal devices' growing computation capabilities, researchers have proposed new hybrid architectures aimed at balancing local computation versus remote offloading toward more flexible solutions.[4–7]

Cloud 2.0 elaborates on this principle by extending traditional cloud capabilities to promote local computation when possible. Within the Cloud 2.0 framework, the cloud's role remains central to all those services and applications that can't rely only on local resources. Local computation might imply the involvement of a single device — usually where the data is generated, such as a smartphone — or could require a federation of devices when available, composed

IC-18-03-BWRS.indd 74 10/04/14 7:39 PM

from a combination of mobile and static platforms — for example, a set-top box, a laptop computer, and a tablet.[6]

Speech recognition engines and image-processing algorithms for, respectively, voice commands and augmented reality applications on smartphones and smart glasses could complete their compute jobs on their own or through interactions with each other. Because of these devices' powerful processing power, less engagement of the cloud might be necessary, resulting in less bandwidth usage and possibly a smoother user experience owing to faster application responsiveness. If sensor data is needed in the cloud to improve a machine-learning algorithm, the local device could send this data to the cloud opportunistically at a later time — for example, during off-peak hours. If the device generating the data is a mobile platform, it can send data when it's in Wi-Fi range or during battery charging sessions.

Although small personal devices play a major role in Cloud 2.0, other classes of hardware and computing platforms can be active parts of the framework. For instance, experimentation and pilot studies are under way to assess the feasibility and utility of V2V and V2I technologies for road safety improvement (www.its.dot.gov/safety_pilot/).[3] In this environment, cars will generate large amounts of sensor data to determine driving conditions, vehicles' relative positioning, and dangerous events that require prompt alerts to drivers and pedestrians. One possibility is, again, to always rely on the cloud to collect cars' and pedestrians' sensor data, process it, and send the results of the computation back to the clients on the road. However, the sheer volume of data these clients could generate via V2V and V2I applications — on the order of 10 messages per second per car — could drastically limit the scalability of a cloud-only solution, both for data communication and processing.

Cloud 2.0 would let application developers design more flexible and efficient V2V and V2I services by offloading parts of the work to nearby computing elements. These computing platforms could be other cars or the road-side infrastructure nodes themselves. Delegating processing and data management to edge devices can siphon data from the network and hence reduce potential traffic congestion, while fostering enhanced QoS by reducing end-to-end delay. This is the case, for example, if cars and pedestrian devices are to cooperate to reduce accidents between pedestrians and vehicles, where real-time alerts must be delivered to the pedestrian (with less than a 100 ms delay) for the service to be effective.

These are only a fraction of the possible application domains where local computation is a viable alternative to boost system performance. However, Cloud 2.0 is applicable to a much broader spectrum of scenarios in which we could exploit local data processing rather than invoke distant cloud intervention, while guaranteeing solid application and service performance.

Personal Data Storage

Although big data is maturing as an important business and research driver in industry and academia, an emerging concept is gaining traction as regards collecting and providing users with their own personal traces from common daily interactions with the digital world. Researchers refer to this new paradigm as *small data*,[8] which lets users draw interesting inferences, extract hidden patterns related to a person's own wellbeing, and enable applications to mine life-logging data. Small data storage requirements, along with users' growing digital footprint — including the pictures and videos from mobile devices that account for most personal storage needs — are driving the demand for more capable personal storage solutions.

Unsurprisingly, when it comes to cloud services, one of the most popular is cloud storage. For one thing, it frees users from having to maintain and manage personal storage hardware (for instance, network-attached storage [NAS] or PCs) while providing a seamless user experience through transparent cross-device synchronization.

With the average household digital footprint skyrocketing (www.gartner.com/newsroom/id/2060215) and the high costs of cloud storage services (hundreds of US dollars per year for roughly 200 gigabytes), Cloud 2.0 could be a viable solution to expand cloud storage capabilities and meet users' storage requirements at a lesser cost. To achieve this goal, Cloud 2.0 can intelligently combine traditional cloud storage with available storage from personal edge devices — smartphones, tablets, PCs, set-top boxes, wearables, car storage, Wi-Fi, and Bluetooth-enabled flash storage. This free space — on the order of tens of gigabytes today, but likely increasing up to terabytes in the future — could be well suited to transparently expanding the storage available to users beyond the boundaries of centralized cloud offerings.

Cloud 2.0 storage services can provide a flexible and scalable platform for personal data storage, including backup, intelligent file placement and retrieval, and data sharing. By applying smart data-replication techniques across different personal devices, this solution provides not only flexibility but also resiliency to failure. In addition, the new Cloud 2.0 storage model offers enhanced scalability properties: a user can add more storage as needed by provisioning a new device (such as a NAS or a cheap network-enabled flash storage device) to their existing personal storage system, which transparently reconfigures with no need for manual data migration or painful data synchronization. Distributed storage solutions — such as Space Monkey (www.spacemonkey.

IC-18-03-BWRS.indd 75 10/04/14 7:39 PM

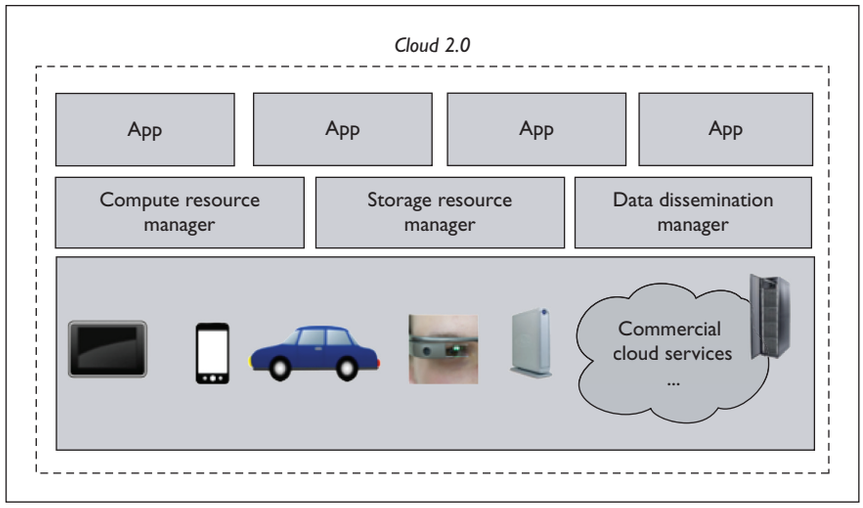

Figure 2. Cloud 2.0 functional requirements. Compute, storage, and data dissemination manager abstractions hide the complexity of the underlying physical systems, while developers can use APIs to build apps that fully exploit the Cloud 2.0 fabric.

com) or My Cloud (www.wdc.com) with static home storage boxes, or Younity (http://getyounity.com), Ori,[9] and PSCloud[10] in the mobile space — all highlight a growing interest in solutions that move away from the purely centralized cloud storage model.

Data Dissemination

Smartphones and tablets are fueling the rapid growth of mobile video streaming traffic, surpassing media consumption on PCs and desktops (www.streamingmedia.com/Articles/Editorial/Featured-Articles/The-State-of-Mobile-Video-2013-87931.aspx). Connected cars will soon accelerate this phenomenon, with thousands of vehicles simultaneously fetching media on the go. Content delivery networks (CDNs) are a solution introduced to meet increasing traffic demand and deliver high-quality customer experience by decentralizing media storage and caching infrastructures (www.business.att.com/enterprise/Service/hosting-services/content-delivery/distribution/).

Cloud 2.0 can enable new generations of CDNs by pushing the ability to store and cache content further out to edge nodes to enhance the user experience and reduce the backbone traffic load. Less traffic in the backbone network could promote less costly infrastructure upgrades. By storing media content on the road-side infrastructure used in the V2I system, a data download session from moving cars could complete faster because of the high-speed, short-range radio connection (typically Wi-Fi). When, in a few years, autonomously driving cars make their appearance, and passengers are busier consuming data than driving, such a design could relieve the pressure of high bandwidth usage off the cellular wireless channel and backbone segments that a centralized cloud solution would otherwise potentially introduce.

Similarly, smart caching on a set-top box based on television users' watching patterns[11] could promote an enhanced user experience by making content readily available, and reduce backbone data traffic during peak hours. For example, if Bob watches the first episode of a 20-episode TV show, the next time, he would most likely select the second episode. This prediction can allow intelligent pre-fetching of the episode on Bob's set-top box at a prior time during off-peak hours.

The Cloud 2.0 approach combines the availability, flexibility, and scalability of edge devices with the resiliency and large computation capabilities of the cloud to offer even more efficient, scalable, and resilient cloud services. Figure 2 shows Cloud 2.0's functional requirement architecture. Compute, storage, and data dissemination managers are tasked with abstracting away the lower-level details of the underlying platforms' compute, storage, and physical characteristics to present a unified, holistic view to application developers and users.

While this architecture is certainly a starting point, many challenges remain before we can successfully progress toward a functional Cloud 2.0 design. A nontrivial effort is needed to design the compute, storage, and data dissemination modules to take into account the complexity and dynamics of the underlying heterogeneous platform ecosystem. Even more pressing is the need to find ways to deal with some devices' mobility and battery limitations while still guaranteeing support for an adequate user experience. Finally, proper privacy and security measures are needed to preserve users' data privacy and platform integrity and protect them from possible attacks aimed at compromising the system during task cooperation.

Acknowledgments

Thanks to my colleagues Yih-Farn Chen, Brian Amento, Hal Purdy, Eric Cheung, Rajesh Panta, Greg Bond, and Tom Smith from AT&T Labs Research for their valuable feedback.

References

1. E. Miluzzo, "Smartphone Sensing," doctoral dissertation, Computer Science Department, Dartmouth College, June 2011.

IC-18-03-BWRS.indd 76 10/04/14 7:39 PM

2. N.D. Lane et al., "BeWell: A Smartphone Application to Monitor, Model and Promote Wellbeing," *Proc. 5th Int'l Conf. Pervasive Computing Technologies for Healthcare*, 2011.
3. K. Robillard, "DOT Plans to Mandate 'Talking' Cars," *Politico*, 3 Feb. 2014; www.politico.com/story/2014/02/sources-dot-to-announce-mandate-on-talking-cars-103022.html.
4. E. Miluzzo et al., "Darwin Phones: The Evolution of Sensing and Inference on Mobile Phones," *Proc. 8th Int'l ACM Conf. Mobile Systems, Applications, and Services* (MobiSys 10), 2010, pp. 5–20.
5. E. Cuervo et al., "MAUI: Making Smartphones Last Longer with Code Offload," *Proc. 8th Int'l ACM Conf. Mobile Systems, Applications, and Services* (MobiSys 10), 2010, pp. 49–62.
6. E. Miluzzo, R. Cáceres, and Y.-F. Chen, "mClouds – Computing on Clouds of Mobile Devices," *Proc. 3rd ACM Workshop Mobile Cloud Computing and Services* (MCS 12), 2012, pp. 9–14.
7. M.-R. Ra et al., "Odessa: Enabling Interactive Perception Applications on Mobile Devices," *Proc. 9th Int'l ACM Conf. Mobile Systems, Applications, and Services* (MobiSys 11), 2011, pp. 43–56.
8. D. Estrin, "Small Data, Where *N* = Me," *Comm. ACM*, vol. 57, no. 4, 2014, pp. 32–34.
9. A.J. Mashtizadeh et al., "Replication, History, and Grafting in the Ori File System," *Proc. 24th Symp. Operating Systems Principles*, 2013, pp. 151–156.
10. S. Bazarbaye et al., "PSCloud: A Durable Context-Aware Personal Storage Cloud," *Proc. Workshop on Hot Topics in Dependable Systems* (HotDep 13), 2013, article no. 9.
11. Y.-F. Chen et al., "Zebroid: Using IPTV Data to Support Peer-Assisted VoD Content Delivery," *Proc. 18th Int'l Workshop Network and Operating Systems Support for Digital Audio and Video*, 2009, pp. 115–120.

Emiliano Miluzzo is a Senior Member of Technical Staff at AT&T Labs Research. His research interests include mobile, pervasive, and distributed computing, mobile sensing systems, and big data analysis. Miluzzo has a PhD in computer science from Dartmouth College. Contact him at miluzzo@research.att.com.

Selected CS articles and columns are also available for free at http://ComputingNow.computer.org.

IC-18-03-BWRS.indd 77 10/04/14 7:39 PM