Unclouded Vision

Jon Crowcroft, Anil Madhavapeddy,
Malte Schwarzkopf, Theodore Hong
Cambridge University Computer Laboratory
15, JJ Thomson Avenue
Cambridge CB3 0FD, UK
firstname.lastname@cl.cam.ac.uk

Richard Mortier
University of Nottingham
Jubilee Campus
Nottingham NG7 2TU, UK
firstname.lastname@nottingham.ac.uk

Abstract

The commercial reality of the Internet and mobile access to it is muddy. Generalising, we have a set of cloud service providers (e.g., Amazon, Facebook, Flickr, Google, Twitter, to choose a representative few), and a set of devices that many, and soon most, people use to access these resources (i.e., so-called smartphones such as Android, Blackberry, iPhone, Maemo). This combination of hosted services and smart access devices is what many people refer to as "The Cloud" and is what makes it so pervasive.

But this situation is not entirely new. Once upon a time, as far back as the 1970s, we had 'thin clients' such as ultra-thin glass `tty`s accessing timesharing systems. Subsequently, the notion of thin client has resurfaced in various guises such as the X-Terminal, and Virtual Networked Computing (VNC) [12]. Although the world is not quite the same now as back in those thin client days, it does seem quite similar in economic terms.

But why is it not the same? Why should it not be the same? The short answer is that the end user, whether in their home or on the top of the Clapham Omnibus, has in their pocket a device with vastly more resource than the mainframe of the 1970s by any measure, whether processing speed, storage capacity or network access rate. With this much power at our fingertips, we should be able to do something smarter than simply using our devices as vastly over-specified dumb terminals.

Meanwhile, the academic reality is that many people have been working at the opposite extreme from this commercial reality, trying to build "ultra-distributed" systems, such as peer-to-peer file sharing, swarms[1], ad hoc mesh networks, mobile decentralised social networks,[2] in complete contrast to the centralisation trends of the commercial world. We choose to coin the name "The Mist" for these latter systems.

The defining characteristic of the Mist is that data is dispersed among a multitude of responsible entities (typically though not exclusively ordinary users), rather than being under the control of a single monolithic provider. Haggle[15], Mirage[10] and Nimbus[13] are examples of architectures for, respectively, networking, operating system and storage components of the Mist.

[1] `http://bittorrent.com/`
[2] `http://joindiaspora.com`, `http://peerson.net/`

These approaches are extreme points in a spectrum, each with its upsides and downsides. We will expand on the relevant capabilities of two instances of these ends subsequently; Table 1 summarises them.

For the average user, accustomed to doing plain old storage and computation on his/her own personal computer or mobile (what we might term "The Puddle"), there are multiple competing incentives pushing in many directions, both towards and away from the Cloud, and towards and away from the Mist.

Risks that the user needs to consider include:

- Losing or breaking a personal device
- Cloud provider going bust
- Virus infection
- Directed hacking attack
- Incompetence / human error
- Network failure / disconnection
- Data getting out of sync in different locatinos
- Privacy
- Scalability / flash crowd

In all of this, there is a tension between what the user needs and what the various providers need in order to make the system viable. For example, the user would like to keep her personal data completely private, but the cloud provider wants to sell her advertising against her personal data. Even nominally altruistic mist networks need incentives to keep them going. In BitTorrent, for instance, it was recently shown that a large fraction of the published content is driven by profit-making companies, rather than altruistic amateur filesharers[2].

Rather than viewing this as a zero-sum conflict, however, we seek to leverage the smart capabilities of our devices to provide happy compromises that can satisfy the needs of all parties. By looking more closely at the true underlying interests of the different sides, we can often discover solutions that achieve seemingly incompatible goals[6].

In the case of advertising, the underlying interest of the cloud provider is to be able to sell targeted ads, not to know everything about its users. Privacy-preserving query techniques can permit ads to be delivered to users matching certain criteria without the provider actually knowing which users they were[8; 9].

Another area of cloud provider interest is data mining on data about the locations or transactions of users. The recent development of differential privacy allows providers to make queries on aggregate data without being able to determine information about specific users[5].

The Cloud: Benefits

Centralising resources brings several significant benefits, specifically:

- economies of scale,
- reduction in operational complexity, and
- commercial gain.

Perhaps the most significant of these is the offloading of the configuration and management burden traditionally imposed by computer systems of all kinds. Additionally, cloud services are commonly implemented using virtualisation technology which allows such efficiencies of scale while still retaining "chinese walls", isolating users with no right to see each other.

The Cloud: Costs

Why should we trust a cloud provider with our personal data? There are many ways that they might abuse that trust, notwithstanding that most operate within jurisdictions implementing various forms of data protection legislation. The waters are further muddied by the various commercial terms and conditions to which users initially sign up, but which providers often evolve over time. When was the last time *you* checked the URL to which your providers will post alterations to their terms and conditions, privacy policies, etc? In such cases, how can we get our data back and move it to another provider, also making sure that they have really really deleted it?

The Mist: Benefits

Accessing the Cloud can be financially costly due to the need for constant high-bandwidth access. Using the Mist, we can reduce our access costs because data is stored locally and need only be uploaded to others selectively and intermittently. We keep control over privacy, choosing exactly what to share with whom and when. We also have better access to our data: we retain control over the interfaces used to access it, we are immune to service disruptions which might affect the network or cloud provider, and we cannot be locked out from our own data by a cloud provider.

The Mist: Costs

Ensuring reliability and availability in such a distributed decentralised system is extremely complex. In particular, a new vector for breach of personal data is introduced: we might leave our fancy device on top of the aforesaid Clapham Omnibus with our data on! We have to manage the operation of the system ourselves, and need to be connected often enough for others to be able to contact us.

Droplets: A Happy Compromise?

In between these two extremes should lie the makings of a design that has all the positives and none of the negatives. In fact, a hint of a way forward is contained in the comments above.

If data is encrypted both on our personal computer/device and in the cloud, then we don't really care where it is stored for privacy reasons. However, as a user, we do care where it is stored for performance reasons. Hence we'd like to carry information of immediate value close to us. We would also like it replicated elsewhere for reliability reasons. Further, we observe that interest/popularity in objects is Zipf-distributed. We also observe that the vast majority of user generated content is of interest *only* within the small social circle of the content subject/creator/producer/owner.

In the last paragraph, it might be unclear who "we" are: "we" refers to Joe Public, whether sitting at home or on the top of that bus. However, there are two other important stakeholders: The Cloud, and The Net. Both need to make money lest all of this fail.

The service provider needs revenue to cover opex and to make a profit but is loathe to charge the user directly. Even in the network case, ISPs (and 3G providers) are mostly heading toward flat data rates. As well as targeted advertisements and associated "click-through" revenue, service providers also want to carry out data mining to do market research of a more general kind. Here, differential privacy[5] and techniques such as k-anonymity[16] come to our aid.

Fortunately, recent advances in security, e.g., Shikfa *et al.* matching interests in the crypto domain[14], or Saikat *et al.* and Haddadi *et al.* with their schemes for privacy preserving advertising[7] and mobile advertising [9], hint at ways to continue to support the two-sided business models that abound in today's Internet.

So we propose *Droplets*, half way betweewn the cloud and the Mist. Droplets make use of Mirage[10], Nimbus[13] and Haggle[15]. They float between the personal device and the cloud using technologies such as social networks, virtualisation and migration[1; 3], and they provide the basic components of a Personal Container[11]. They condense within social networks, where privacy is assured by society, but in the great unwashed Internet, they stay opaque. Techniques alluded to above allow the service providers to continue to provide the storage, computation, indexing, search and transmission they do today, with the same wide range of business models.

By way of example, consider the following. As part of the instantiation of their Personal Container, Joe Public runs

an instance of a Nimbus "trust fountain". When creating a droplet from some data stored in his Personal Container, this trust fountain creates a cryptographic attestation proving Joe's ownership of the data at that time in the form of a time-dependent hash token.

The droplet is then encrypted under this hash token using a fast, medium strength cipher[3] and pushed out to the cloud. By selectively publishing the token, Joe can grant access to the published droplet allowing, e.g., a provider offering free data storage and hosting in exchange for data mining access. Alternatively, the token might only be shared with a few friends via a wireless network in a coffee shop, granting only them access to the data at that time.

A secondary purpose of the attestation is to enable "backwards provenance", i.e., a way to prove ownership. Imagine that Joe publishes a picture of some event that he took using his smartphone while driving past it on that oft-considered bus. A large news agency picks up and uses that picture after Joe publishes it to his Twitter stream using a droplet. The attached attestations then enables the news agency to compensate both the owner and potentially the owner's access provider, who takes a share in all profits made of Joe's digital assets in exchange for serving them.

Furthermore, Joe is given a tool to counter "hijacking" of his creation even if the access token becomes publicly known: Using the cryptographic properties of the token, the issue log of his trust fountain and his provider's confirmation of receiving the attested droplet together form sufficient evidence to prove ownership and take appropriate legal action. However, note that Joe Public can always chose to deny ownership, as only his trust fountain holds the crucial information necessary to regenerate the hash token and thus prove the attestation's origin.

Of course, whenever a droplet becomes sufficiently popular to merit condensation into a cloud burst of marketing, then we have the means to support this transition, and we have the motivation and incentives to make sure the right parties are rewarded. In this last paragraph, "we" refers to all stakeholders: users, government and business. It seems clear that the always-on, everywhere-logged, ubiquitously-connected vision will continue to be built, while real people become increasingly concerned about their privacy [4]. Without such features, it is unclear how long commercial exploitation of personal data will continue to be acceptable to the public; but without such exploitation, it is unclear how service providers can continue to provide the many "free" Internet services on which we have come to rely.

References

[1] C. Clark, K. Fraser, S. Hand, J. G. Hansen, E. Jul, C. Limpach, I. Pratt, and A. Warfield. Live migration of virtual machines. In *USENIX Symposium on Networked Systems Design & Implementation (NSDI)*, pages 273–286, Berkeley, CA, USA, 2005. USENIX Association.

[2] R. Cuevas, M. Kryczka, A. Cuevas, S. Kaune, C. Guerrero, and R. Rejaie. Is content publishing in BitTorrent altruistic or profit-driven, Jul 2010.

[3] B. Cully, G. Lefebvre, D. T. Meyer, A. Karollil, M. J. Feeley, N. C. Hutchinson, and A. Warfield. Remus: High availability via asynchronous virtual machine replication. In *USENIX Symposium on Networked Systems Design & Implementation (NSDI)*, Berkeley, CA, USA, April 2008. USENIX Association.

[4] C. Doctorow. *The Things that Make Me Weak and Strange Get Engineered Away.* Tor.com, August 2008. `http://www.tor.com/stories/2008/08/weak-and-strange`.

[5] C. Dwork. Differential privacy. In *International Colloquium on Automata, Languages and Programming (ICALP)*, pages 1–12, 2006.

[6] R. Fisher, B. M. Patton, and W. L. Ury. *Getting to Yes: Negotiating Agreement Without Giving In.* Houghton Mifflin, April 1992.

[7] S. Guha, B. Cheng, A. Reznichenko, H. Haddadi, and P. Francis. Privad: Rearchitecting online advertising for privacy. Technical Report MPI-SWS-2009-004, Max Planck Institute for Software Systems, 2009.

[8] S. Guha, A. Reznichenko, K. Tang, H. Haddadi, and P. Francis. Serving Ads from localhost for Performance, Privacy, and Profit. In *Proceedings of Hot Topics in Networking (HotNets)*, New York, NY, October 2009.

[9] H. Haddadi, P. Hui, and I. Brown. MobiAd: Private and scalable mobile advertising. In *Proceedings of MobiArch 2010*, 2010. To appear.

[10] A. Madhavapeddy, R. Mortier, R. Sohan, T. Gazagnaire, S. Hand, T. Deegan, D. McAuley, and J. Crowcroft. Turning down the lamp: Software specialisation for the cloud. In *Proceedings of the 2nd USENIX Workshop on Hot Topics in Cloud Computing (HotCloud)*, Boston, MA, USA, June 2010. USENIX.

[11] R. Mortier, C. Greenhalgh, D. McAuley, A. Spence, A. Madhavapeddy, J. Crowcroft, and S. Hand. The Personal Container, or Your Life In Bits. In submission to Digital Futures 2010, October.

[12] T. Richardson, Q. Stafford-Fraser, K. R. Wood, and A. Hopper. Virtual network computing. *IEEE Internet Computing*, 2(1):33–38, January 1998.

[13] M. Schwarzkopf and S. Hand. Nimbus: Intelligent Personal Storage, 2010. Poster at the Microsoft Research Summer School 2010, Cambridge, UK.

[14] A. Shikfa, M. Önen, and R. Molva. Privacy in content-based opportunistic networks. In *AINA Workshops*, pages 832–837, 2009.

[15] J. Su, J. Scott, P. Hui, J. Crowcroft, E. De Lara, C. Diot, A. Goel, M. H. Lim, and E. Upton. Haggle: seamless networking for mobile applications. In *UbiComp'07: Proceedings of the 9th international conference on Ubiquitous computing*, pages 391–408, Berlin, Heidelberg, 2007. Springer-Verlag.

[3] Strong encryption is not required as the attestations are unique for each droplet publication and breaking one does not grant an attacker access to any other droplets.

Platform	*Google AppEngine*	*VM (e.g., on EC2)*	*Home Computer*	*Mobile Phone*
Storage	moderate	moderate	high	low
Bandwidth	high	high	limited	low
Accessibility	always on	always on	variable	variable
Computation	limited	flexible, plentiful	flexible, limited	limited
Cost	free	expensive	cheap	cheap
Reliability	high	high	medium (failure)	low (loss)

Table 1: Comparison of different platforms to store and handle personal data.

[16] L. Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, 2002.