Moving Clouds Brian Noble & Azarias Reda {bnoble,azarias}@umich.edu

Latency presents an enduring—and worsening—challenge to mobile systems designers. The problem of latency-dominated performance problems date back to the earliest days of gigabit networking [2]. Unfortunately, the "reach" of latency as a first-class concern is extending. Bandwith grows roughly as the square of latency, while storage capacity is growing faster still [4]

There are several reasons for this. The instantaneous nature of latency makes it hard to improve the metric by simply packing more bits on the wire. To make things worse, additional devices on the network, such as firewalls and switches, add more delay to network packets while minimally impacting the aggregate bandwidth. Finally (and perhaps most importantly), bandwidth is simply easier to sell in the marketplace.

Unfortunately, even in systems with an adequate balance between latency and bandwidth—and such systems will become increasingly rare—humans are acutely sensitive to delay and jitter. Performance analyses of interactive applications, ranging from video streaming to augmented reality, show a modest increase in latency can make a session noticeably annoying or unusable [3]. Even at a few hundred milliseconds of latency, all too common in cellular networks, users reported many applications start to become noticeably annoying [10]. For highly interactive applications, user experience degrades significantly much sooner. Although running all applications locally would result in a more crisp experience, this is not often feasible in the pervasive context as many tasks operate on computationally constrained devices.

The latency problem is even more pronounced in challenged network environments, where resources are limited to begin with. With shared dial up connections in internet kiosks as the typical way to connect to the internet [5], developing countries face significant challenges in network access. This makes even simple network tasks unpleasant, and rich media prohibitively difficult. Working through an interactive session in one of these kiosks can be charitably described as frustrating [6]. Provisioning data and computational support as close to demand as possible is the key to solving this problem [7].

To address this provisioning problem, we propose *the moving cloud*, a proactive data delivery framework that leverages route fingerprints in individual mobility with users' contextualized behavior of data access for predictive data placement. Essentially, we are trading bandwidth and storage for latency, exchanging resources that grow more quickly for the one that grows most slowly. The moving cloud alleviates the latency problem by proactively placing content where it needs to be in the near future, so that resources are closely and readily available when requested by the user. This paradigm enables a number of networking scenarios including mobile resource augmentation, on-demand social networks, personal content distribution and vehicular applications.

People are creatures of habit, and move in repeated patterns that can be probabilistically learned. As such, several models have been suggested for predicting human mobility [1, 8, 9]. Given the history of locations visited, these models typically predict the next location for a user. Our framework employs a new approach for augmenting these predictions with time bounds, producing actionable information for data placement. This temporal component of mobility is crucial as data delivery often has some freshness constraint. Our approach can be combined with most existing location predictors, and enhance them with an expected-time dimension. For example, coupled with a second order Markov model [9], our system can predict time of arrival within an hour in more than 90% of the hits in the dataset.

Another important observation is the contextual nature of data use. Not only do people move in repeated patterns, they also access data in a habitual manner. Data accessed in one context, say during work hours, is often different from data accessed in another, say while at a restaurant. As a result, individual mobility can also provide an informed data selection policy in content placement. We combine these complementary observations in building a secure, generic, proactive and predictive data delivery framework.

References

[1] GONZALEZ, M. C., HIDALGO, C. A., AND BARABASI, A.-L. Understanding individual human mobility patterns. *Nature 453*, 7196 (June 2008), 779–782.

[2] KLEINROCK, L. The latency/bandwidth tradeoff in gigabit networks. *Communications Magazine, IEEE 30*, 4 (apr. 1992), 36 –40.

[3] LAGAR-CAVILLA, H. A., TOLIA, N., DE LARA, E., SATYANARAYANAN, M., AND O'HALLARON, D. Interactive resource-intensive applications made easy. In *Middleware '07: Proceedings of the ACM/IFIP/USENIX 2007 International Conference on Middleware*, pp. 143–163.

[4] PATTERSON, D. A. Latency lags bandwith. *Commun. ACM 47*, 10 (2004), 71–75.

[5] PETRAZZINI, B., AND KIBATI, M. The Internet in developing countries. *Commun. ACM 42*, 6 (1999), 31–36.

[6] REDA, A., NOBLE, B., AND HAILE, Y. Distributing private data in challenged network environments. In *WWW '10*, pp. 801–810.

[7] SATYANARAYANAN, M., BAHL, P., CACERES, R., AND DAVIES, N. The case for vm-based cloudlets in mobile computing. *IEEE Pervasive Computing 8* (2009), 14–23.

[8] SONG, C., QU, Z., BLUMM, N., AND BARABASI, A.-L. Limits of Predictability in Human Mobility. *Science 327*, 5968 (2010), 1018–1021.

[9] SONG, L., KOTZ, D., JAIN, R., AND HE, X. Evaluating location predictors with extensive wi-fi mobility data. *SIGMOBILE Mob. Comput. Commun. Rev. 7*, 4 (2003), 64–65.

[10] TOLIA, N., ANDERSEN, D. G., AND SATYANARAYANAN, M. Quantifying interactive user experience on thin clients. *Computer 39* (2006), 46–52.