# Bistro: A Scalable and Secure Data Transfer Service for Digital Government Applications*

[ Appeared in Communications of the ACM (CACM), January 2003 ]

Leana Golubchik† leana@cs.usc.edu
William C. Cheng‡ bill.cheng@acm.org
Cheng-Fu Chou§ chengfu@cs.umd.edu
Samir Khuller§ samir@cs.umd.edu
Hanan Samet§ hjs@cs.umd.edu
Y.C. Justin Wan§ ycwan@cs.umd.edu

http://bourbon.usc.edu/iml/bistro/

Government at all levels is a major collector and provider of data.

In this project we focus on the *collection* of data over wide-area networks and address the scalability issues which arise in the context of Internet-based massive data collection applications. Furthermore, security, due to the need for privacy and integrity of the data, is a central issue for data *collection* applications which use a public infrastructure such as the Internet. Numerous digital government applications require collection of data over wide-area networks [5]. One compelling example of such an application is IRS' electronic submission of income tax forms. Other digital government applications include collecting census data, federal statistics, and surveys; gathering and tallying of electronic votes; collecting crime data for the Justice department; collecting data from sensors for disaster response applications; collecting data from geological surveys; collecting electronic filings of patents, permits, and securites (for SEC) applications; grant proposals and contract bids submissions; and so on. All these applications have scalability and security needs in-common.

The poor performance that may currently be experienced by digital government users, given the existing state of technology (as in Figure 1(a)), is largely due to how (independent) data transfers using TCP/IP work over the Internet. TCP/IP is good at equally sharing bandwidth between data streams, which in large-scale applications can lead to poor performance for individual clients (as they receive only a very small share of this bandwidth). Given that TCP/IP is here to stay for the foreseeable future, what is needed is a scalable yet cost-effective solution which can be easily deployed over the existing Internet technology.

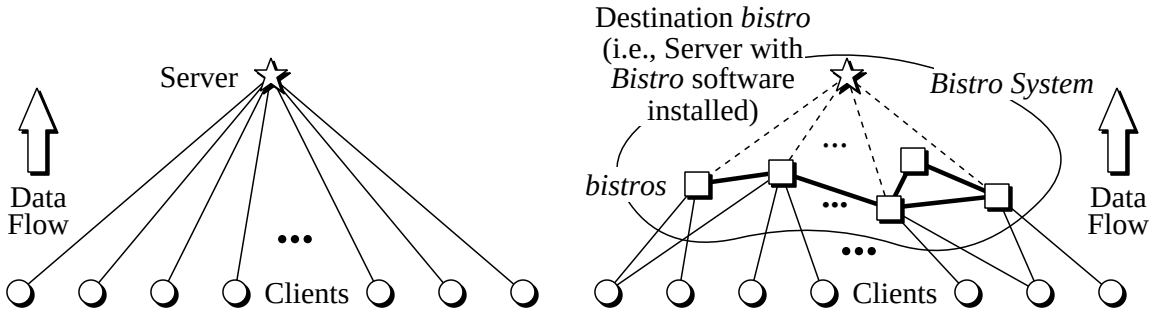
We are designing and developing a system, termed *Bistro*, which addresses scalability needs of digital government data collection applications while allowing them to share the same infrastructure and resources efficiently, cost-effectively, and securely [1]. *Bistro*'s basic approach is to introduce intermediate hosts, termed *bistro*s, which allow replacement of a traditionally "synchronized client push" approach with a "non-synchronized combination of client-push and server-pull" approach (as depicted in Figure 1(b)). This in turn allows spreading of the workload on the destination server and the network over time, with subsequent elim-

*This work is supported in part by the NSF Digital Government 0091474 grant.

†Computer Science Department, University of Southern California, Los Angeles, CA 90089. This work was partly done while the author was with the Department of Computer Science and UMIACS at the University of Maryland.

‡TeleGIF, Marina del Rey, California. This work was partly done while the author was with the Department of Computer Science and UMIACS at the University of Maryland.

§Department of Computer Science and UMIACS, University of Maryland, College Park, MD 20742.




(a) Upload for a single event (current state) (b) Upload with the *Bistro System*

Figure 1: Data collection for digital government applications.

ination of hot-spots as well as significant improvements in performance for both clients and servers. Our on-going research [2, 4] indicates that orders of magnitude improvement can be achieved with our *Bistro* architecture and the corresponding data collection algorithms which it affords.

*Bistro*'s design allows for a gradual deployment and experimentation over the public Internet (by simply downloading *Bistro* server software and installing it on public servers). *Bistro*'s security protocol and trust structure [3] are designed such that only encrypted data travels through (not necessarily trusted) *bistro*s. This means that a government agency does not have to trust *bistro*s installed by other agencies or commercial institutions; at the same time these (untrusted) *bistro*s can significantly improve the agency's data collection performance. Each application (within each agency) can have its own scalability, security, fault tolerance, and other data collection needs, and these applications and agencies can still share available resources, if so desired, across all *Bistro* servers.

We believe that an appropriately designed *single* infrastructure, such as *Bistro*, can address all digital government wide-area data collection needs in a scalable, secure, and cost-effective manner.

## References

[1] S. Bhattacharjee, W. C. Cheng, C.-F. Chou, L. Golubchik, and S. Khuller. Bistro: a platform for building scalable wide-area upload applications. *ACM SIGMETRICS Performance Evaluation Review (also presented at the Workshop on Performance and Architecture of Web Servers (PAWS) in June 2000)*, 28(2):29–35, September 2000.

[2] W. C. Cheng, C.-F. Chou, and L. Golubchik. Performance of online batch-based digital signatures. *Submitted for publication.*

[3] W. C. Cheng, C.-F. Chou, L. Golubchik, and S. Khuller. A secure and scalable wide-area upload service. In *Proceedings of the 2nd International Conference on Internet Computing, Volume 2*, pages 733–739, June 2001.

[4] W. C. Cheng, C.-F. Chou, L. Golubchik, S. Khuller, and Y.C. Wan. On a graph-theoretic approach to scheduling large-scale data transfers. *Submitted for publication.*

[5] W.C. Cheng, C.F. Chou, L. Golubchik, S. Khuller, and H. Samet. Scalable data collection for internet-based digital government applications. In *1st National Conference on Digital Government Research*, pages 108–113, Los Angeles, CA, May 2001.