# Robust Caching for Rapidly-growing Repositories

***Tanu Malik***
***(Purdue U.)***

***Abstract***

Modern scientific repositories are growing rapidly. Using the latest data to answer user queries over the network can incur exorbitant network costs, unless an effective dynamic data caching system is used. Compared to known systems, however, such a caching system has to be more general-purpose and meet harder challenges. It needs to manage not only the network latency costs, but also the network traffic costs. It needs to work with scientific queries which include complex SQL constructs. Furthermore, it needs to adapt to evolving workloads. Taking the "no-approximations" hotspot decoupling approach (keep query hotspots in proxycaches close to users, and update hotspots away), we design RIVER, a caching system that meets these needs.We present cache management algorithms for RIVER, which adaptively choose between three data communication options—loading objects, shipping queries, and shipping updates—to optimize network costs while respecting data freshness and cache size constraints. Our techniques combine elements from graph theory and network flow. We evaluate their performance on real queries and updates from the SDSS astronomy workloads.