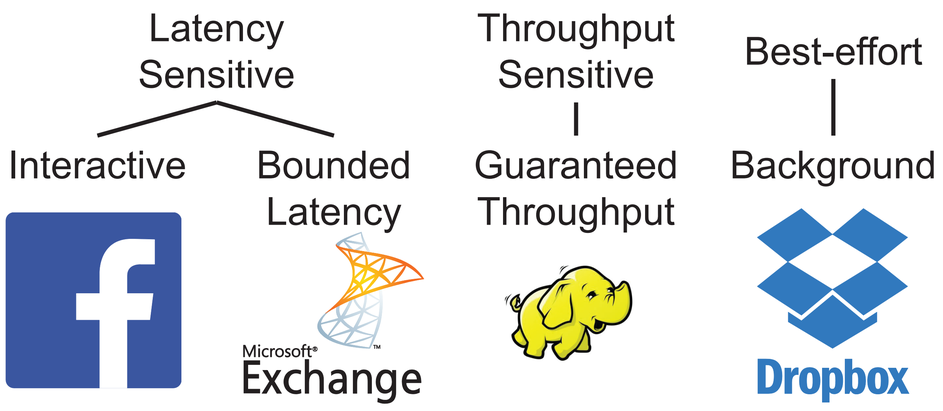
Enabling End-to-End Latency & Throughput SLOs on Shared Storage

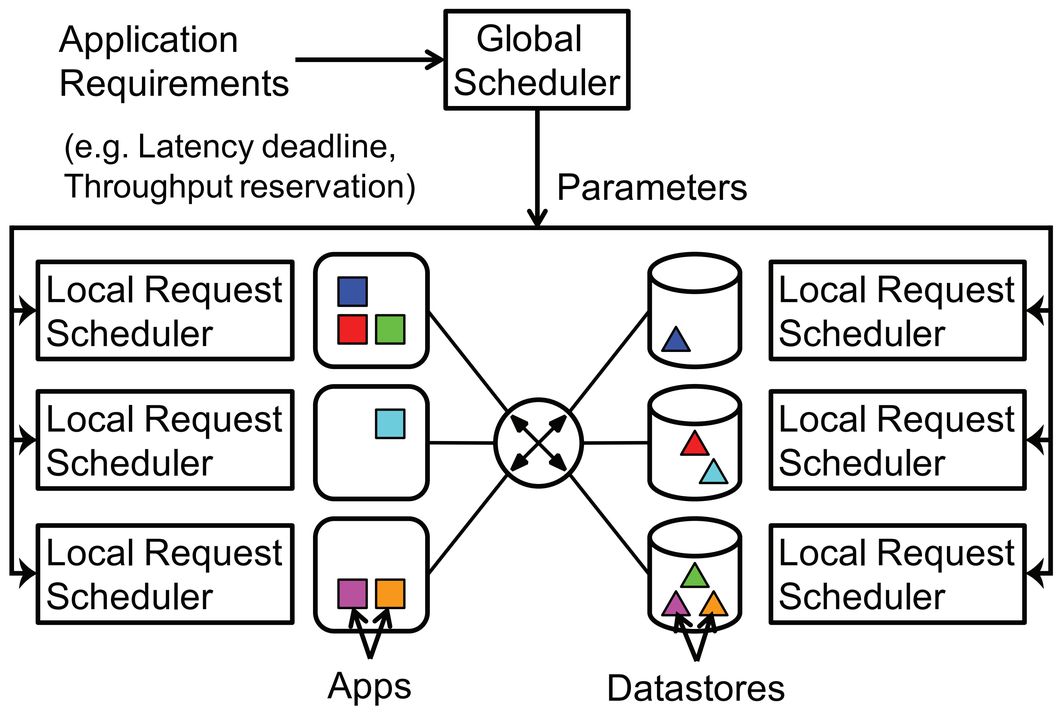
Timothy Zhu, Alexey Tumanov, Michael A. Kozuch*, Mor Harchol-Balter, Gregory R. Ganger
(Carnegie Mellon University, *Intel Labs)

PROBLEM STATEMENT

- **Share storage while satisfying a mix of latency and throughput objectives**
- **Challenges:**
  - **End-to-end (network + storage) latency**
  - **Automatic system parameter configuration**
  - **Diverse workload requirements**

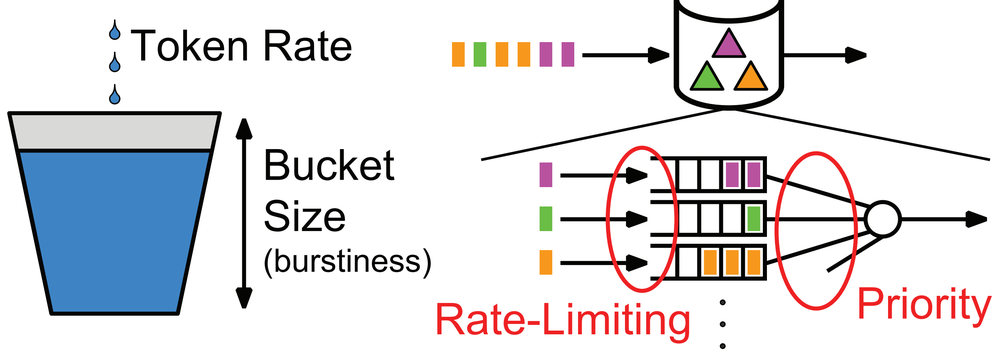


SYSTEM DESIGN



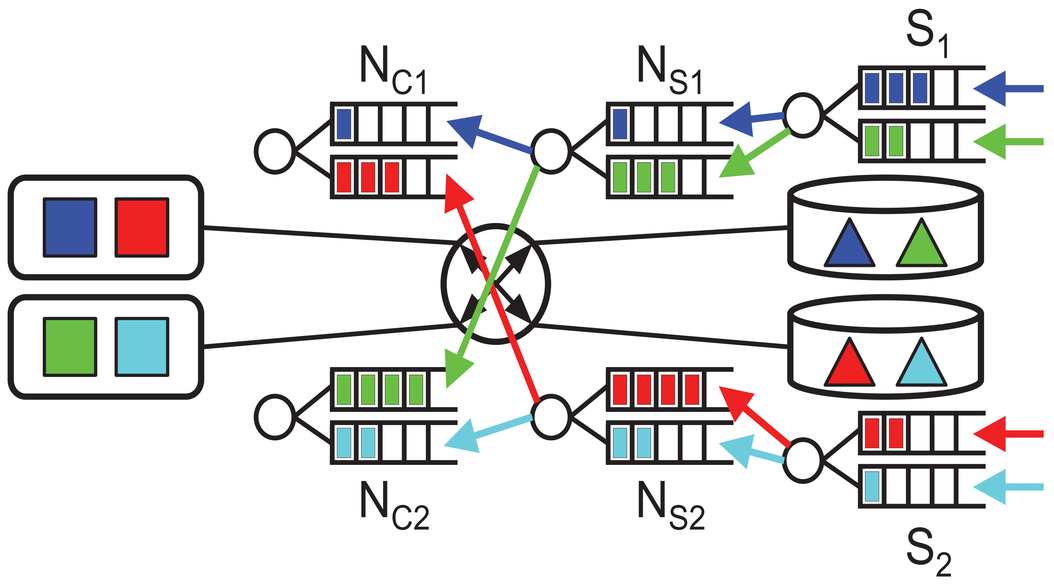
LOCAL REQUEST SCHEDULER

- **Each client app gets a FIFO queue**
- **Priority provides latency differentiation**
- **Rate-limiting avoids starvation**
  - **Token-bucket model**

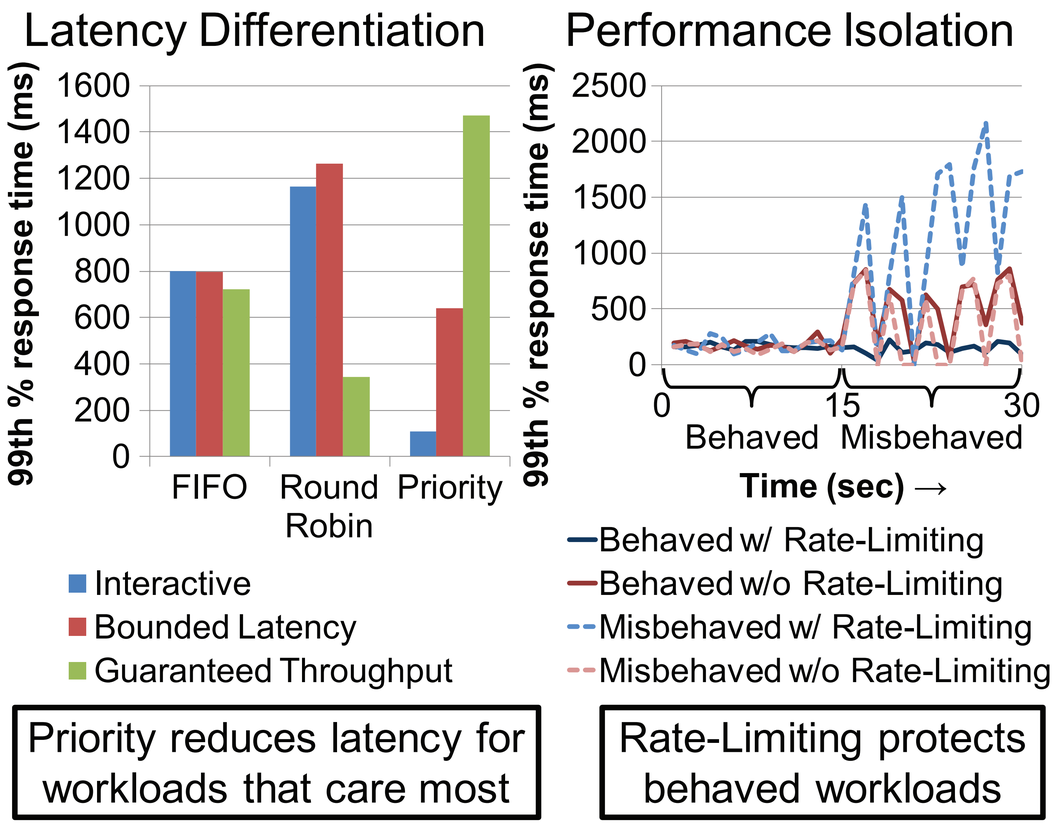


GLOBAL SCHEDULER

- **Assigning priorities to meet end-to-end deadlines is hard**
  - **Client priorities may be different between queues**
  - **Combinatorial optimization problem**



PRELIMINARY RESULTS



Priority reduces latency for workloads that care most

Rate-Limiting protects behaved workloads

POTENTIAL DIRECTIONS

- **Flexible user SLOs (e.g., soft/hard deadline)**
  - **Latency and/or throughput**
- **Automatic app/datastore placement decisions**
- **App and data migration for better placement**

Intel Science & Technology Center for Cloud Computing

Carnegie Mellon University · Georgia Tech · intel · Princeton University · UC Berkeley · University of Washington