

Design and Implementation of a Caching System for Streaming Media over the Internet

Ethen Bommaiah (Bell Labs)

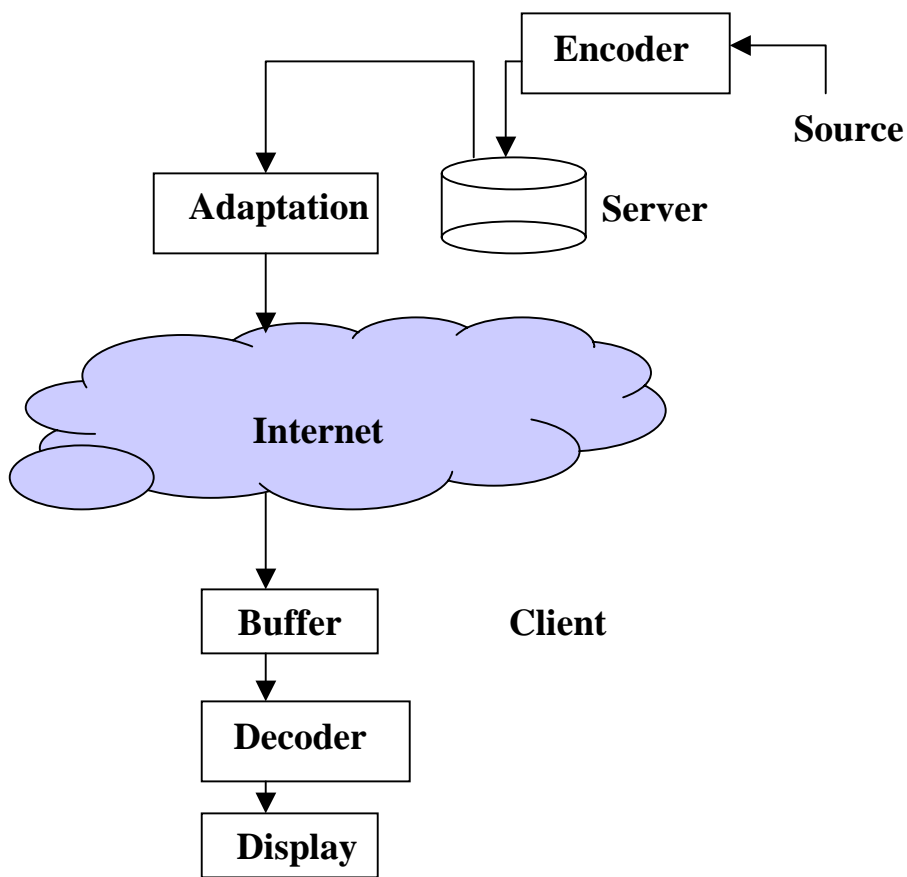
Katherine Guo (Bell Labs)

Markus Hofmann (Bell Labs)

Sanjoy Paul (Edgix Corporation)

Multimedia Streaming

- Different from download and play.
- Fill the client's play-out buffer then play from the buffer.

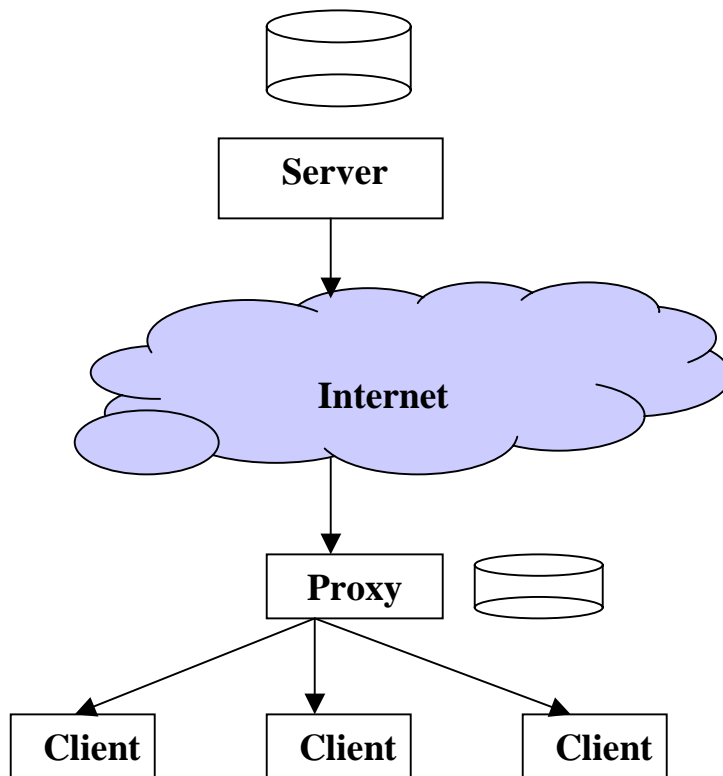


Multimedia Streaming

- By 2001, 50% of the web sites are going to have streaming capability.
- Today's streaming quality is still poor.
 - High start-up latency.
 - Unpredictable playback quality.
 - Poor performance with VCR operations.

Streaming Cache

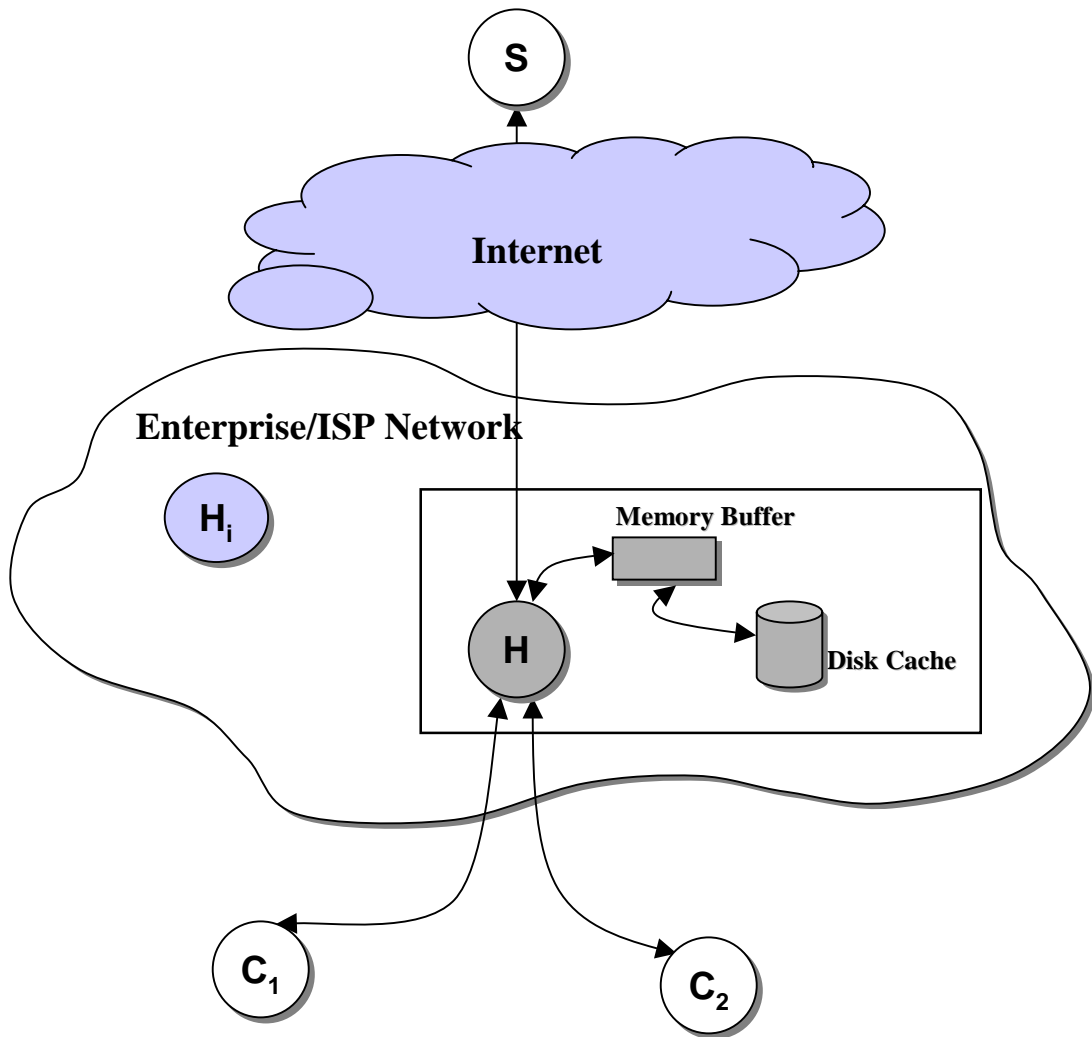
- Reduce server load for content providers.
- Reduce network load for ISPs.
- Improve client playback quality.



Outline

- Streaming Cache Architecture
- Key Techniques
 - segmentation and prefix caching
 - client request aggregation
 - data transfer rate control
- Implementation with RTP/RTSP
- Performance Results

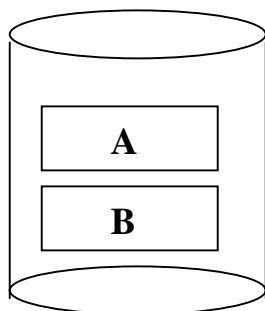
Streaming Cache Architecture



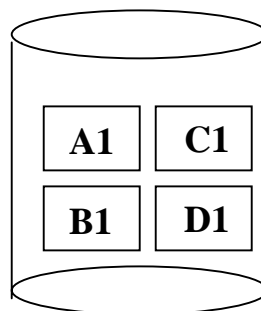
Segmentation and Prefix Caching (using disk cache)

- **Media Object Characteristics:**
 - Large size (2-hour long MPEG-I movie 1.4GB disk space)
 - Timing requirement
- **Scalable Solution:**
 - Division into smaller segments
 - Segments can be cached and replaced independently.

Without Segmentation

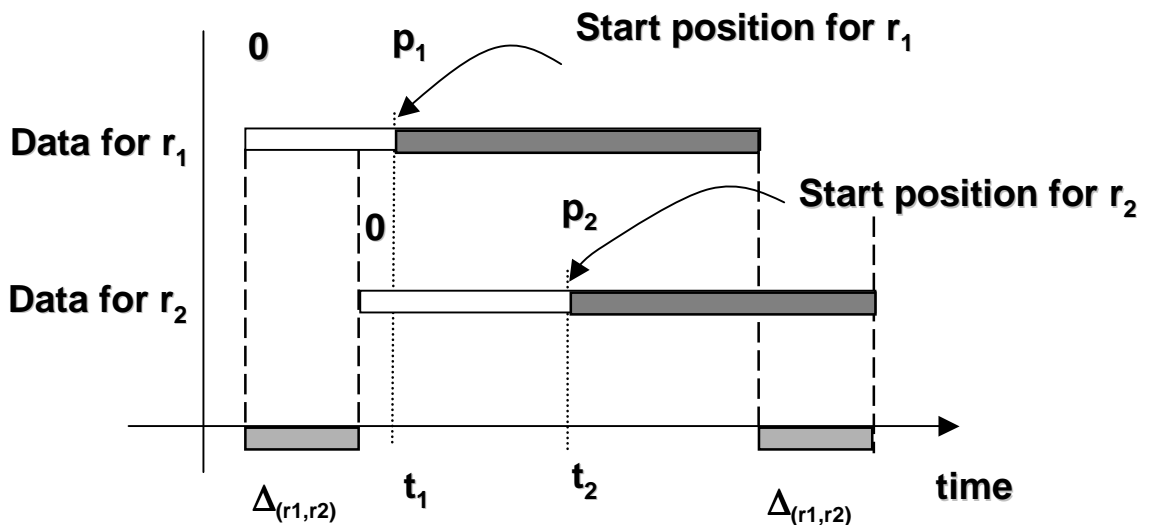


With Segmentation



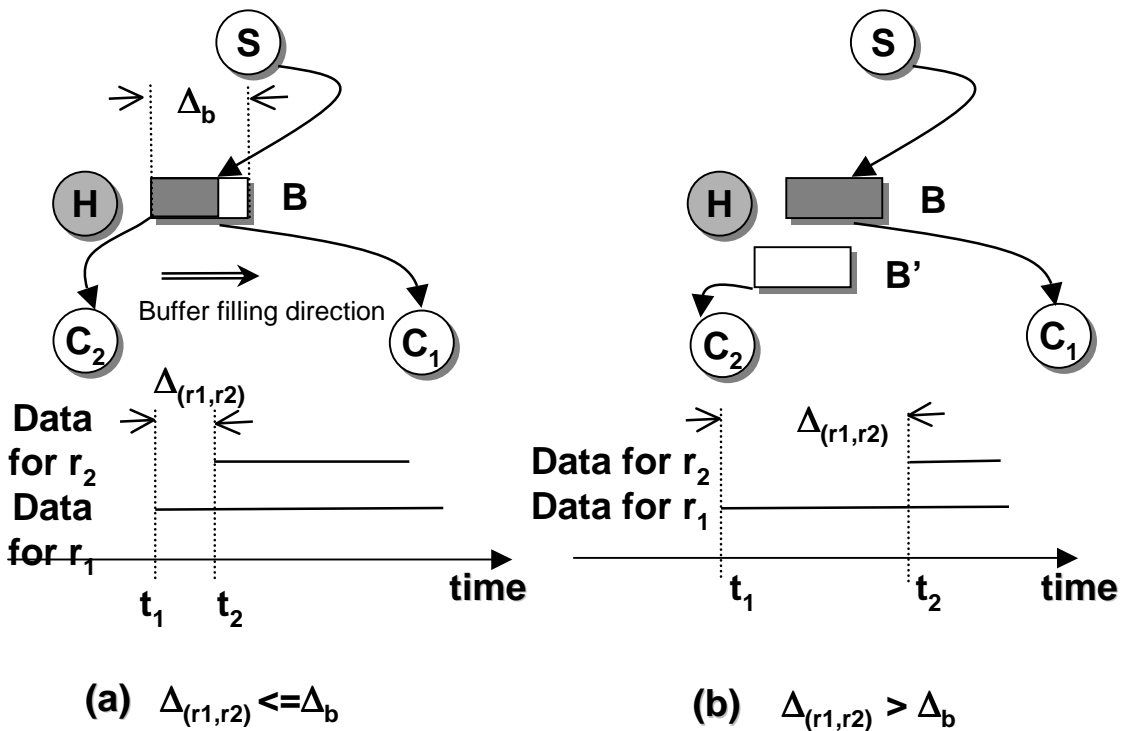
Client Request Aggregation (using memory buffer)

- Heterogeneity
 - Media object
 - Arrival time
 - Request range
- Temporal Distance btwn requests:
$$\Delta_{(r_1,r_2)} = (t_2 - p_2) - (t_1 - p_1) = (t_2 - t_1) - (p_2 - p_1)$$

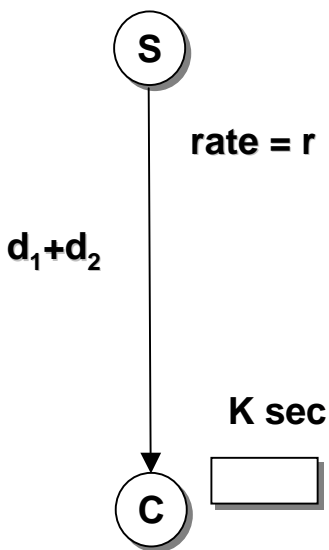


Client Request Aggregation -- Ring Buffer

Buffer Temporal Distance: Δ_b

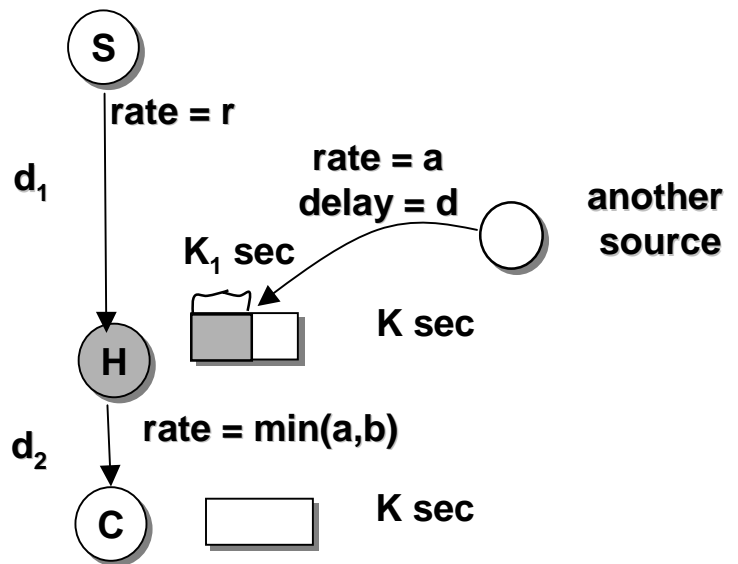


Data Transfer Rate Control



(a) client-server

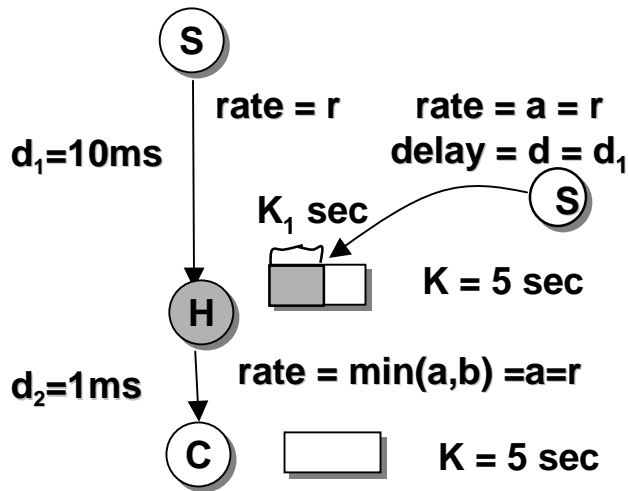
$$L_0 = 2(d_1 + d_2) + K$$



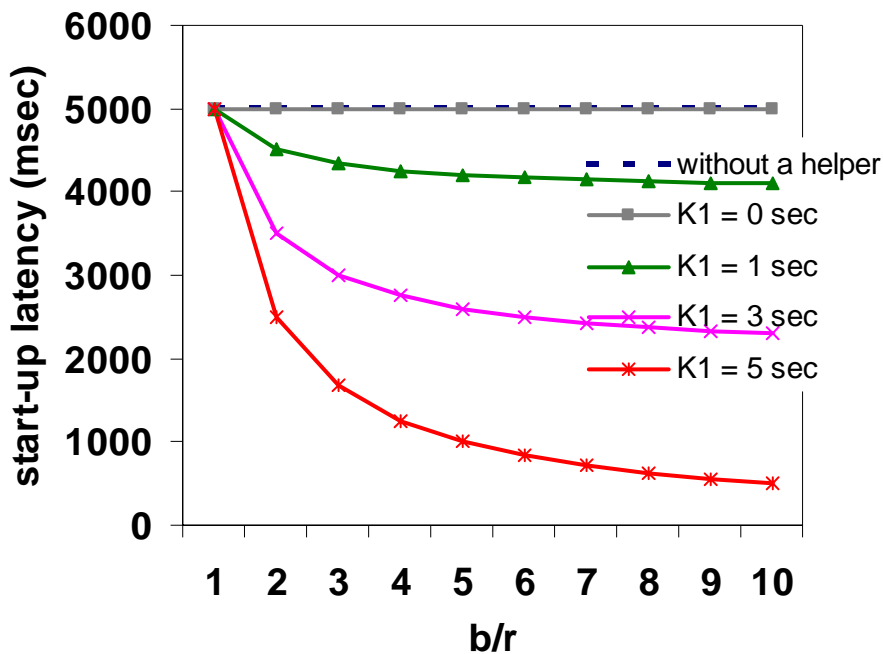
(b) client-helper-server

$$L_1 = d_2 + \max(K_1 r / b, 2d) + d_2 + (K - K_1) r / \min(a, b)$$

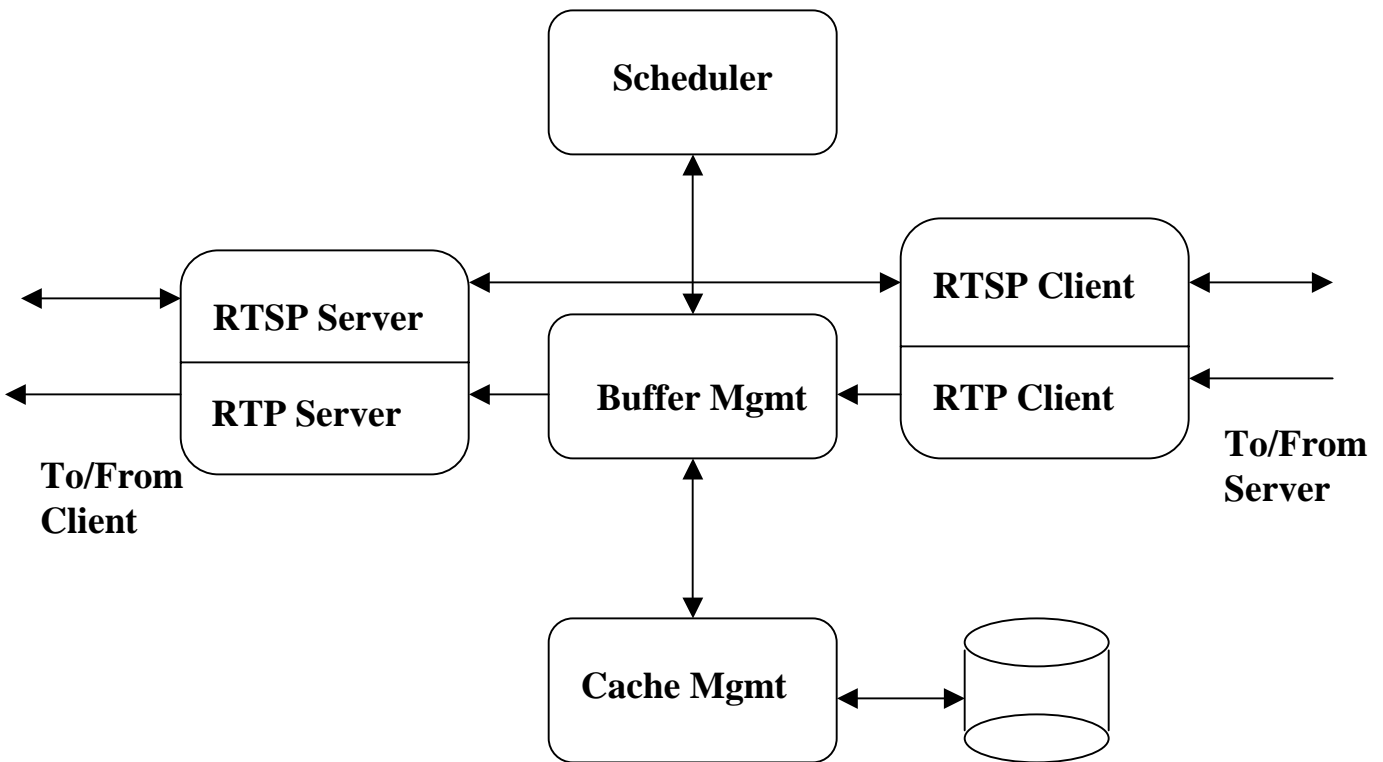
Data Transfer Rate Control -- Start-up latency



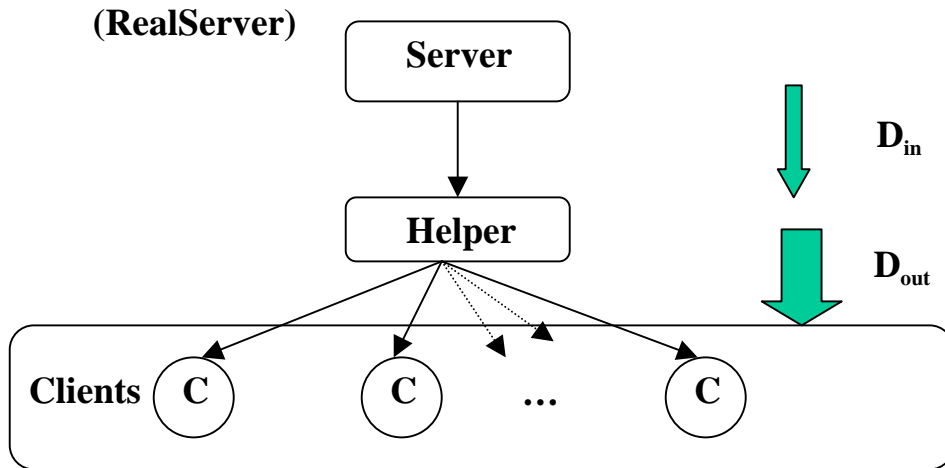
(b) client-helper-server



Implementation



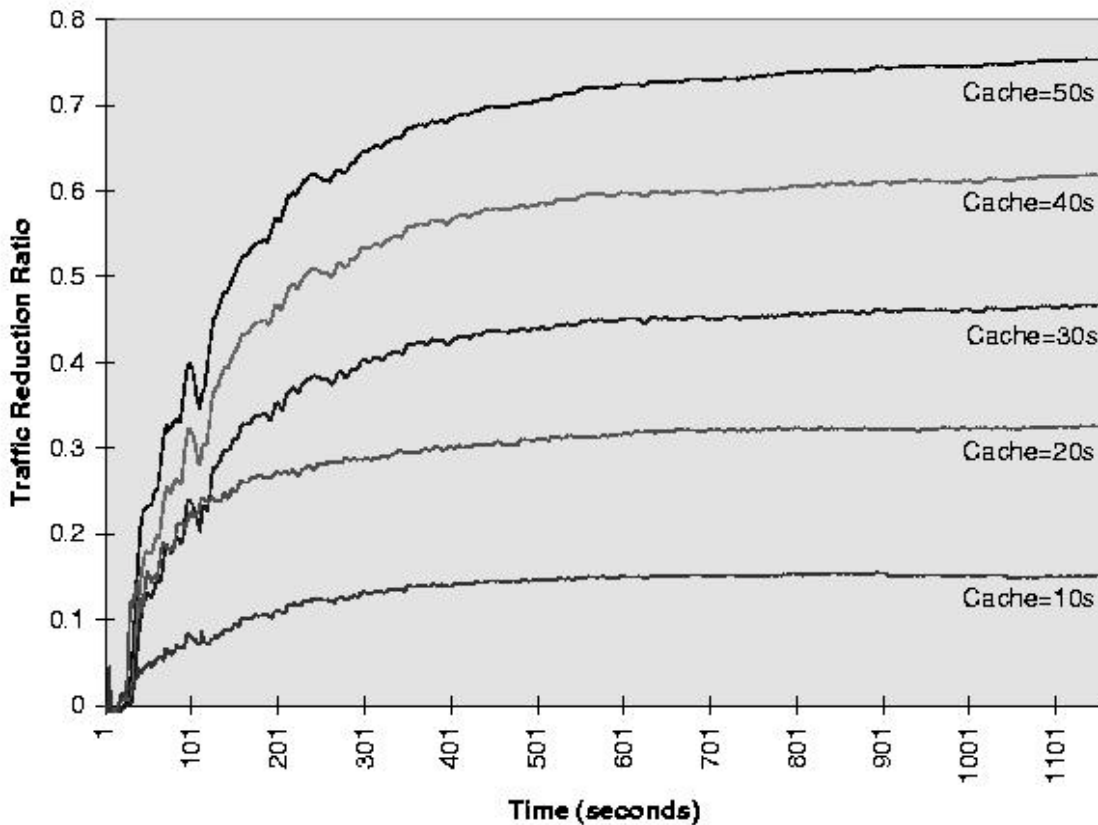
Performance Tests



- Server: Sun Ultra-4, 4 CPUs, 1GB mem, Sun OS 5.6.
- Client: 300MHz Pentium, 250MB mem, FreeBSD.
- Helper: 400MHz Pentium II, 250MB mem, FreeBSD.
- 12 MPEG Clips (40s-70s)
- Request whole clips with Zpif distribution.
- Request inter-arrival time: 15s.
- Traffic Reduction Ratio: $\mathbf{R} = (\mathbf{D}_{out} - \mathbf{D}_{in}) / \mathbf{D}_{in}$

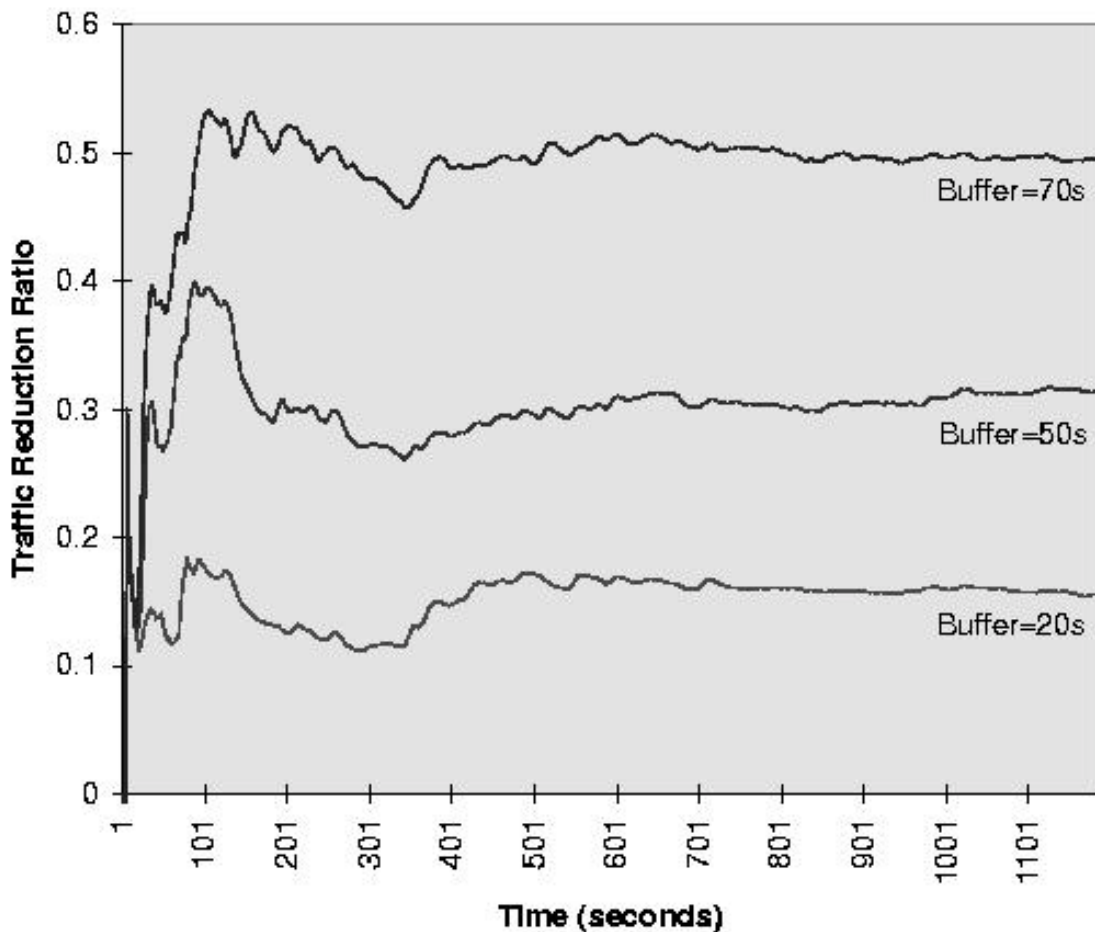
Network and Server Load Reduction

- Prefix Caching

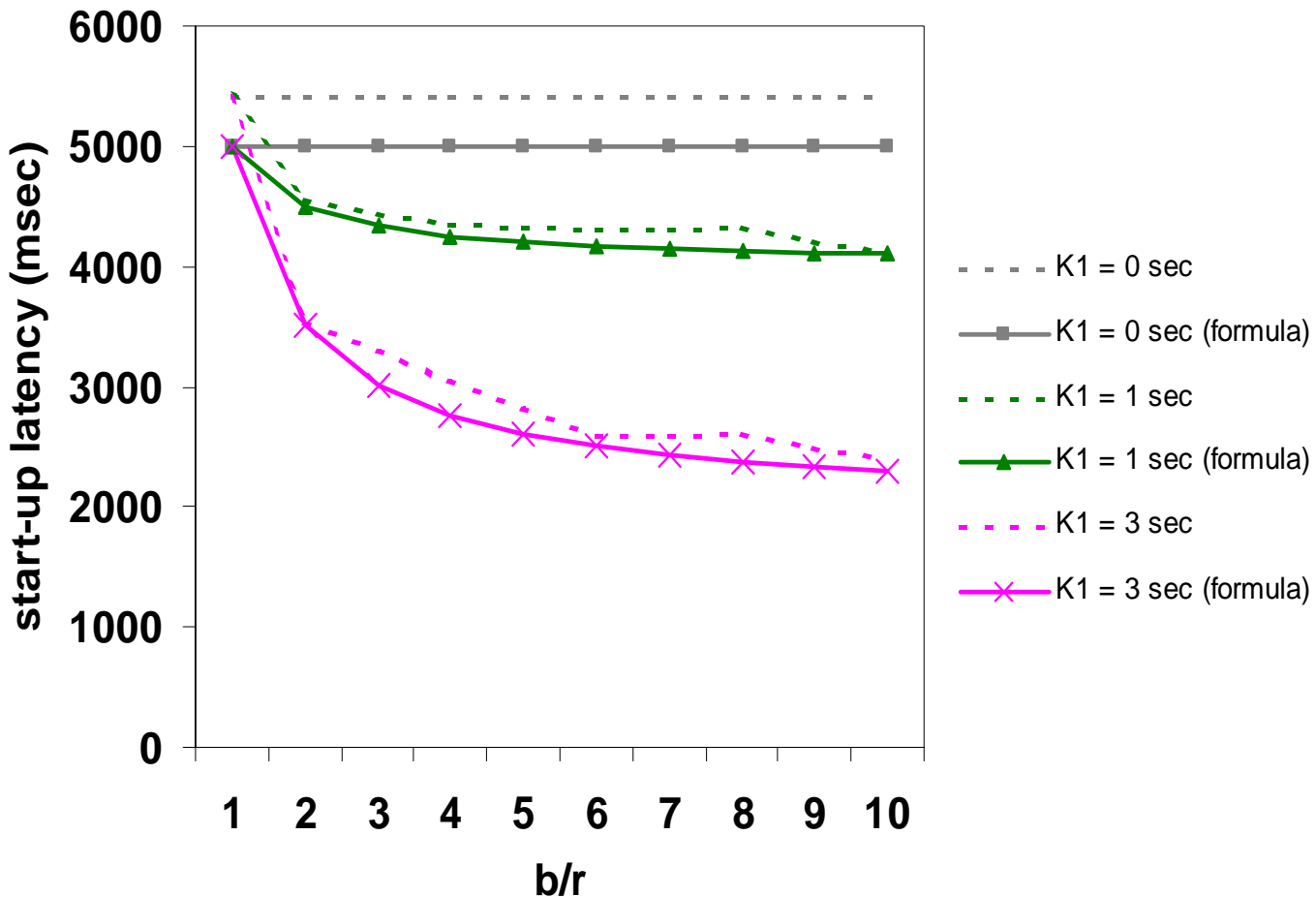
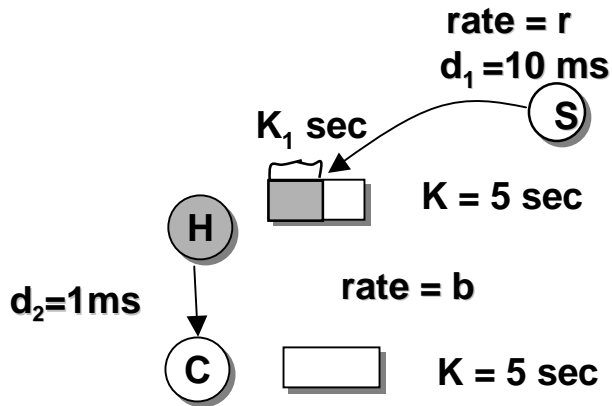


Network and Server Load Reduction

- Ring buffer request aggregation



Client Start-up Latency



Conclusion

- Benefits of helpers:
 - reduce server load
 - reduce network load
 - improve end user streaming quality
- Techniques at the helper:
 - prefix caching on disk
 - ring buffer in memory
 - data transfer rate control
- Other techniques to use:
 - video smoothing
 - layered encoding
 - distributed cooperative caching