The Content and Access Dynamics of a Busy Web Server

Venkat Padmanabhan and Lili Qiu

Microsoft Research *Cornell University*

Networking Workshop

MSR Cambridge, U.K.

Jan 2000

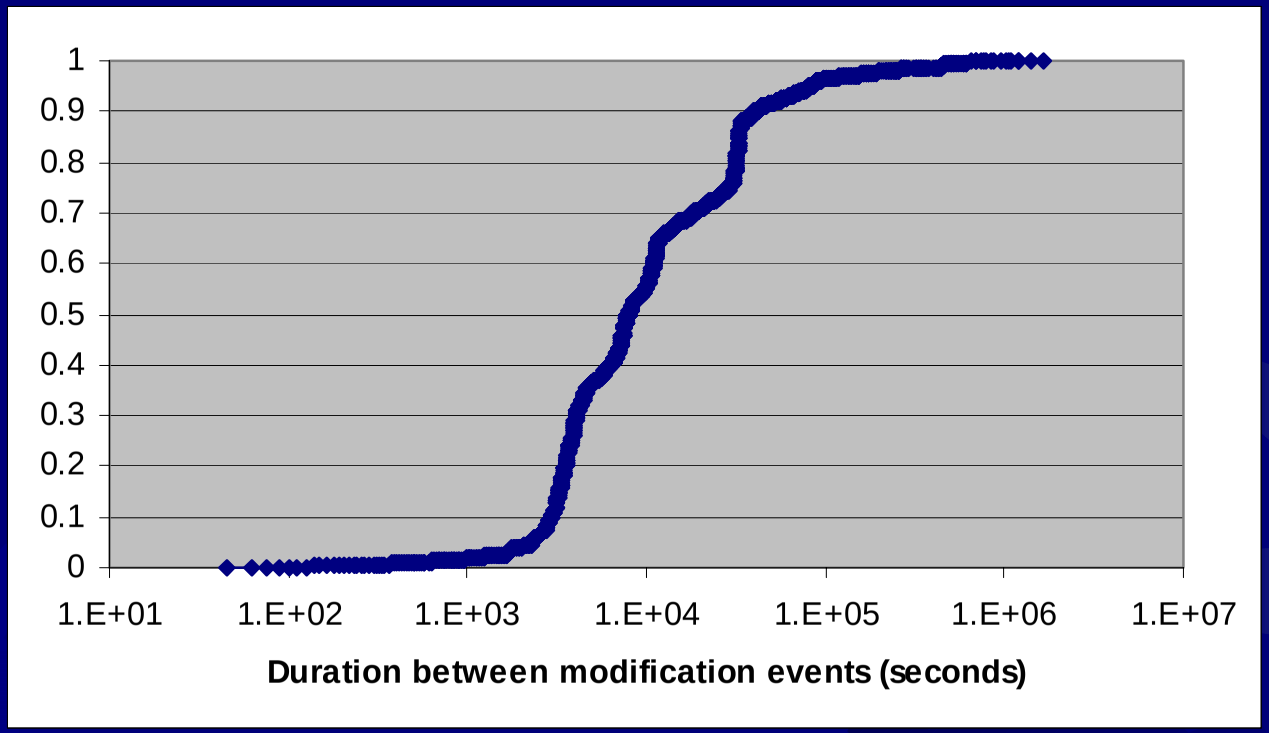
Overview

- MSNBC server site
  - server cluster with 40 nodes
  - 25 million access per day (HTML content alone)
- Server logs
  - HTTP access logs
  - CRS logs
  - HTML content logs
- Data analysis
  - content dynamics
  - access dynamics

Content Dynamics

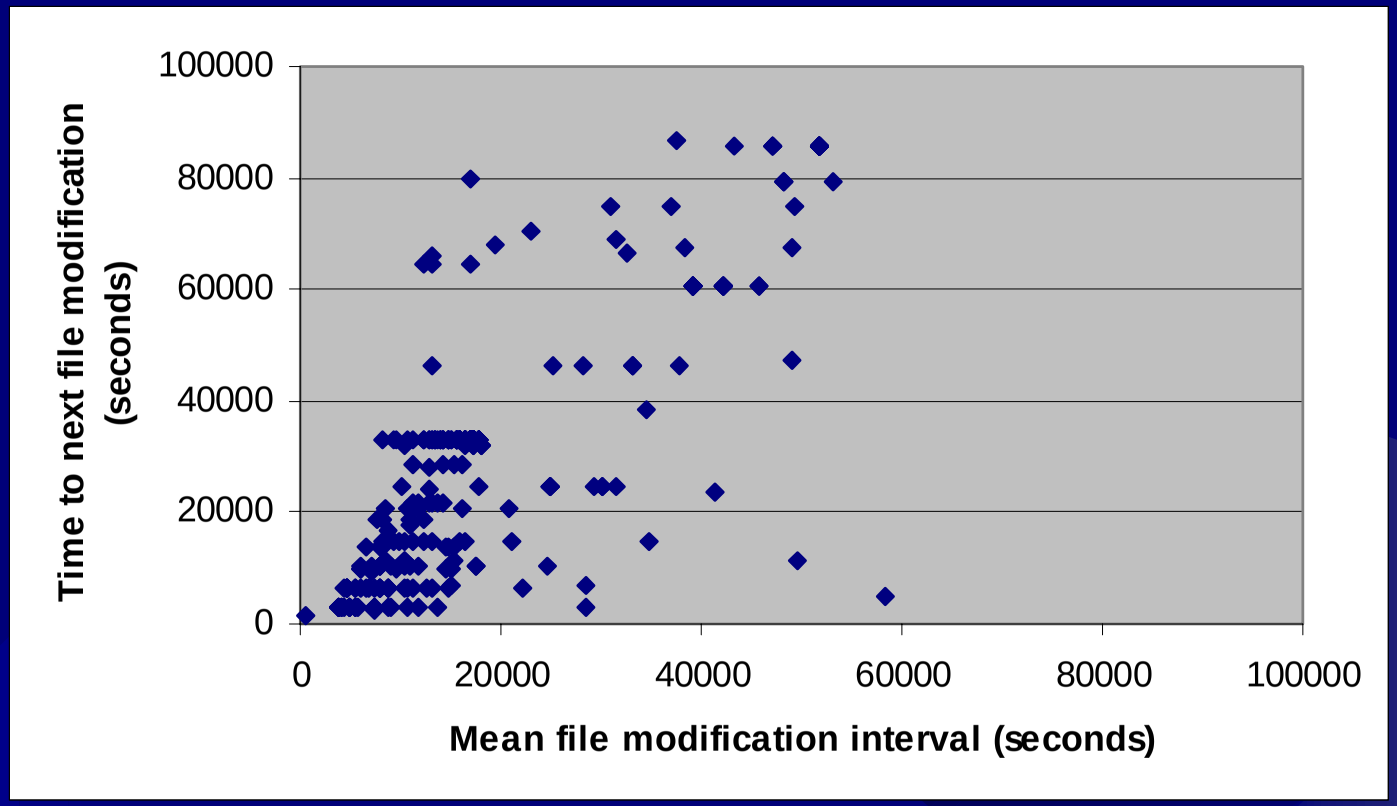
- Over a one-week period:
  - 6000 file creations, 24000 file modifications
    - a small number of image modifications as well
  - only 2500 distinct files modified
  - makes caching, prefetching challenging
- Most file modifications are minimal
  - little change in size or content
    - cosine text similarity metric
  - delta encoding would be useful

CDF of Duration between Modification Events



Distinct knees in the CDF at 1 hour and 1 day

Predictive Power of Modification History

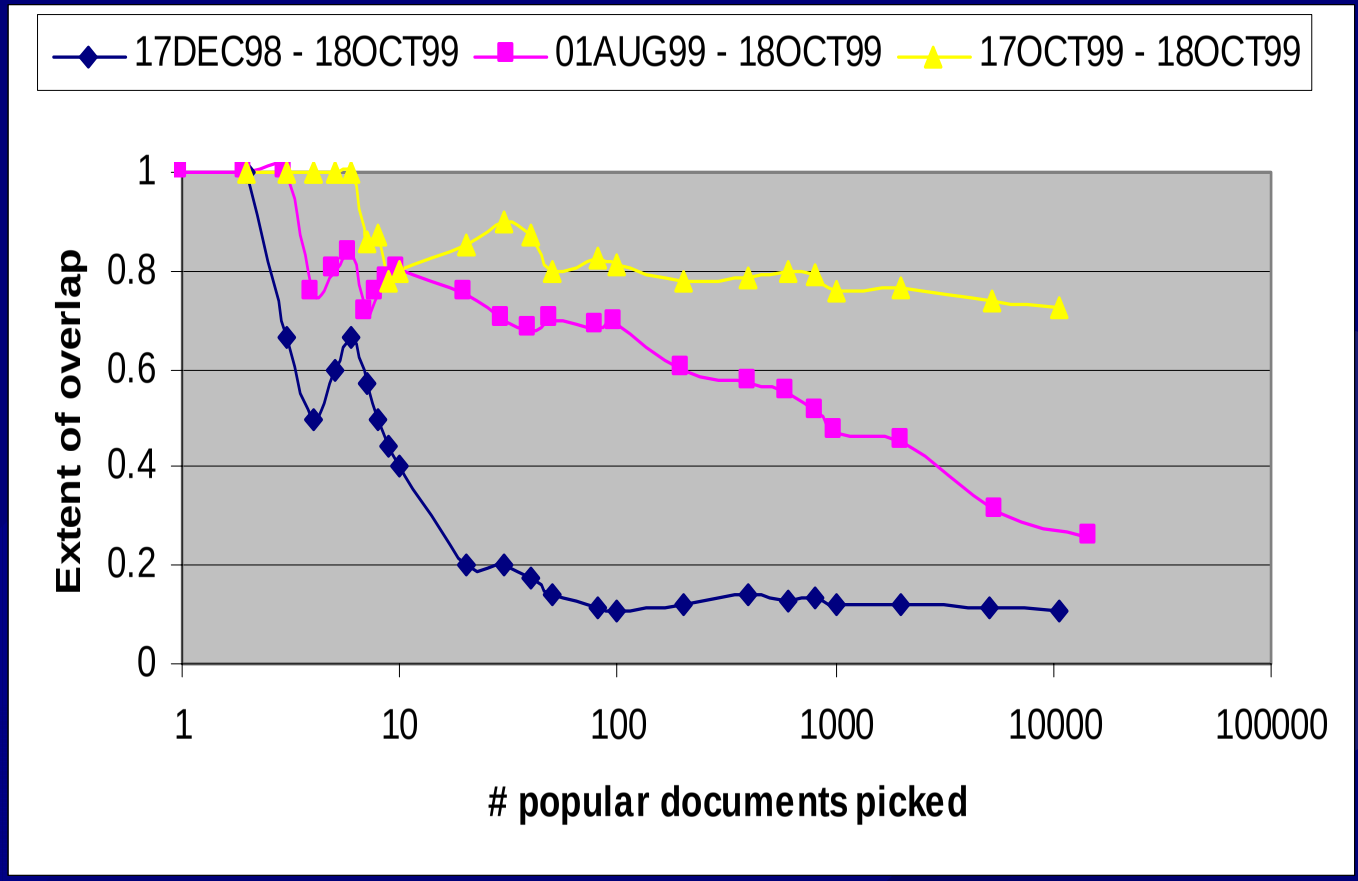


Averaging yields a good prediction ⇒ adaptive TTL cache validation may be useful

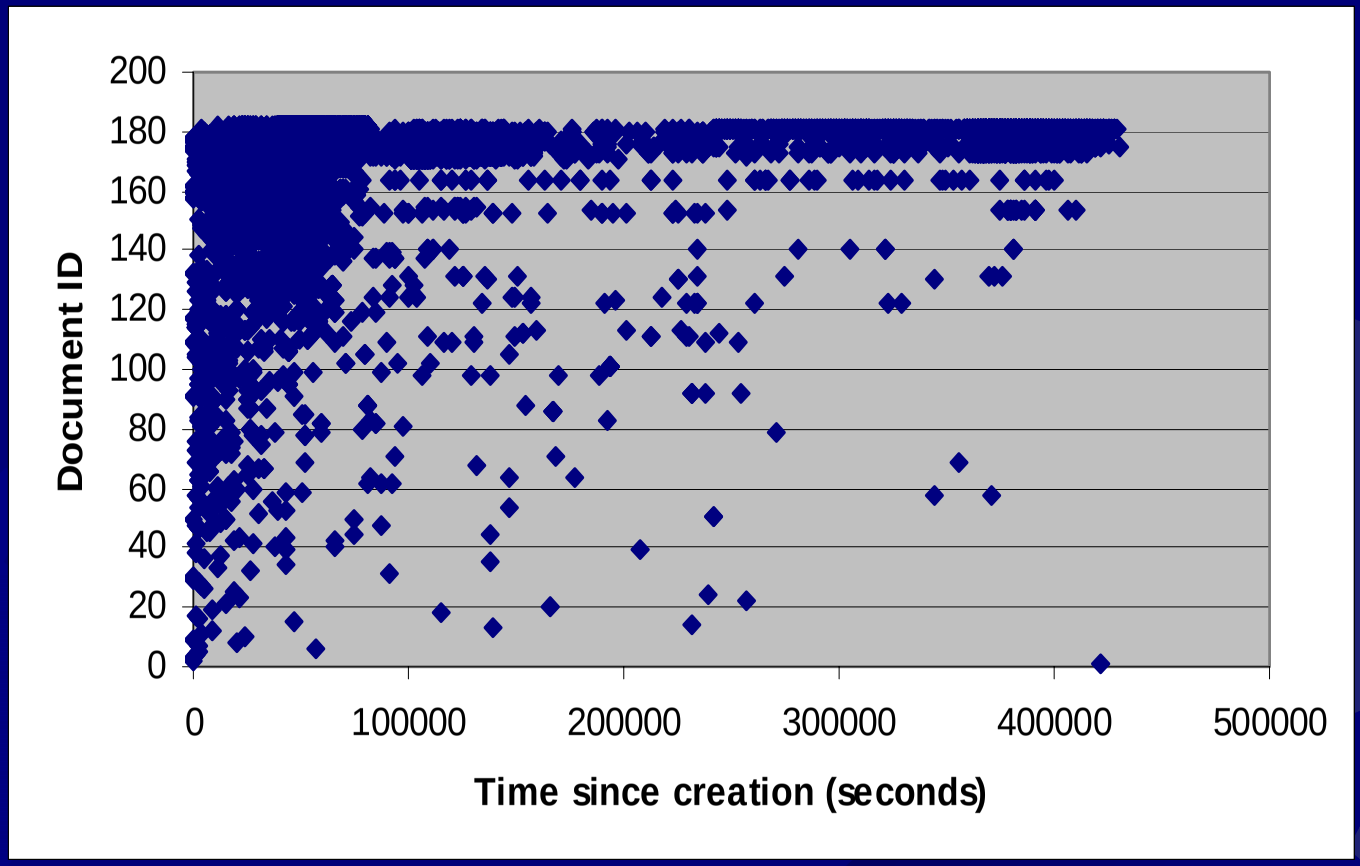
Access Dynamics

- Zipf-like distribution of file popularity
  - greater concentration of accesses than at proxies
  - alpha is 1.4-1.8 (compared to < 1.0 at proxies
  - Zipf-like distribution holds even at higher levels of a multi-level caching hierarchy
- Other analyses:
  - temporal stability of file popularity
  - temporal stability of client interest group
  - spatial locality in client accesses

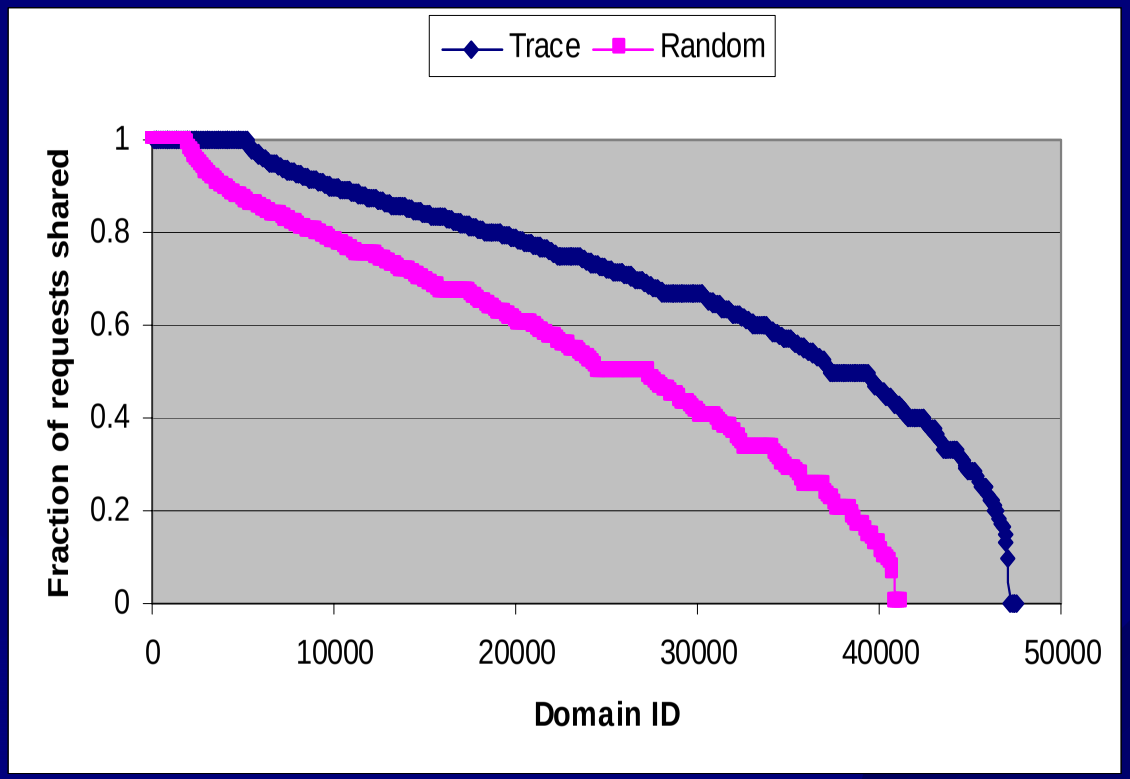
Temporal Stability of File Popularity



Impact of Age on Popularity



Spatial Locality in Client Accesses



Domain membership is significant except when there is a “hot” event of global interest

Summary

- Frequent but minimal file modifications
- Modification history is a good predictor
- Zipf-like distribution of file popularity with a large alpha
- Domain membership has a significant bearing on client interests

Paper submitted for publication