Stanford WebBase Components and Applications

Junghoo Cho, Hector Garcia-Molina,
Taher Haveliwala, Wang Lam, Andreas Paepcke,
Sriram Raghavan, and Gary Wesley

9 September 2004

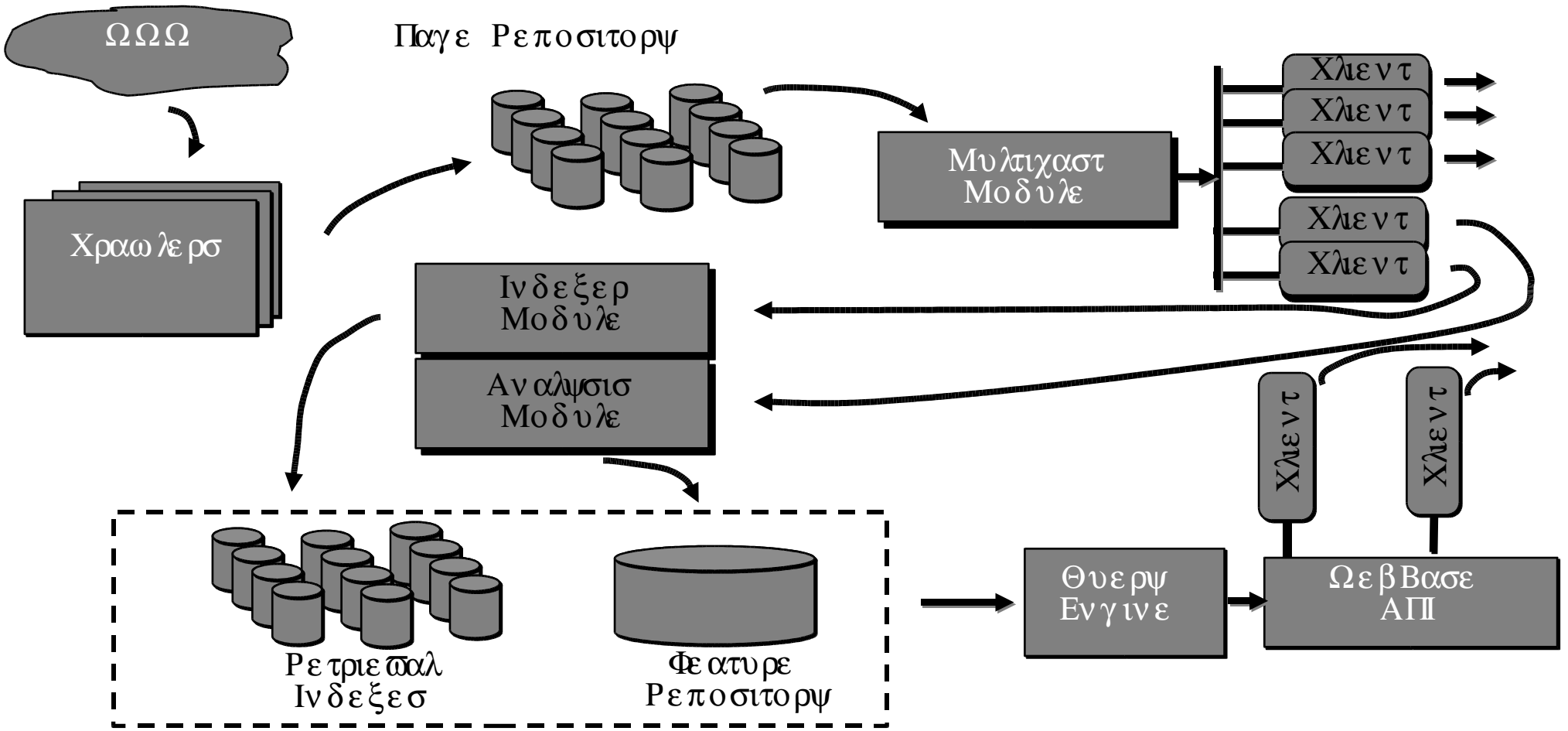
WebBase Goals

- Enable large-scale Web-related research
- Crawl and store significant pieces of the Web over time
- Index Web data easily and efficiently
- Clean, simple data retrieval over the network

Challenges

- Crawling
  - Minimize load on Web servers; honor robots.txt
  - Maximize throughput (parallelization)
  - Simplify operation (easy to start/stop/resume)
- Indexing
  - Scale to large datasets (storage, time)
- Distribution
  - Maximize throughput
  - Simplify retrieval (fetch and use of data)

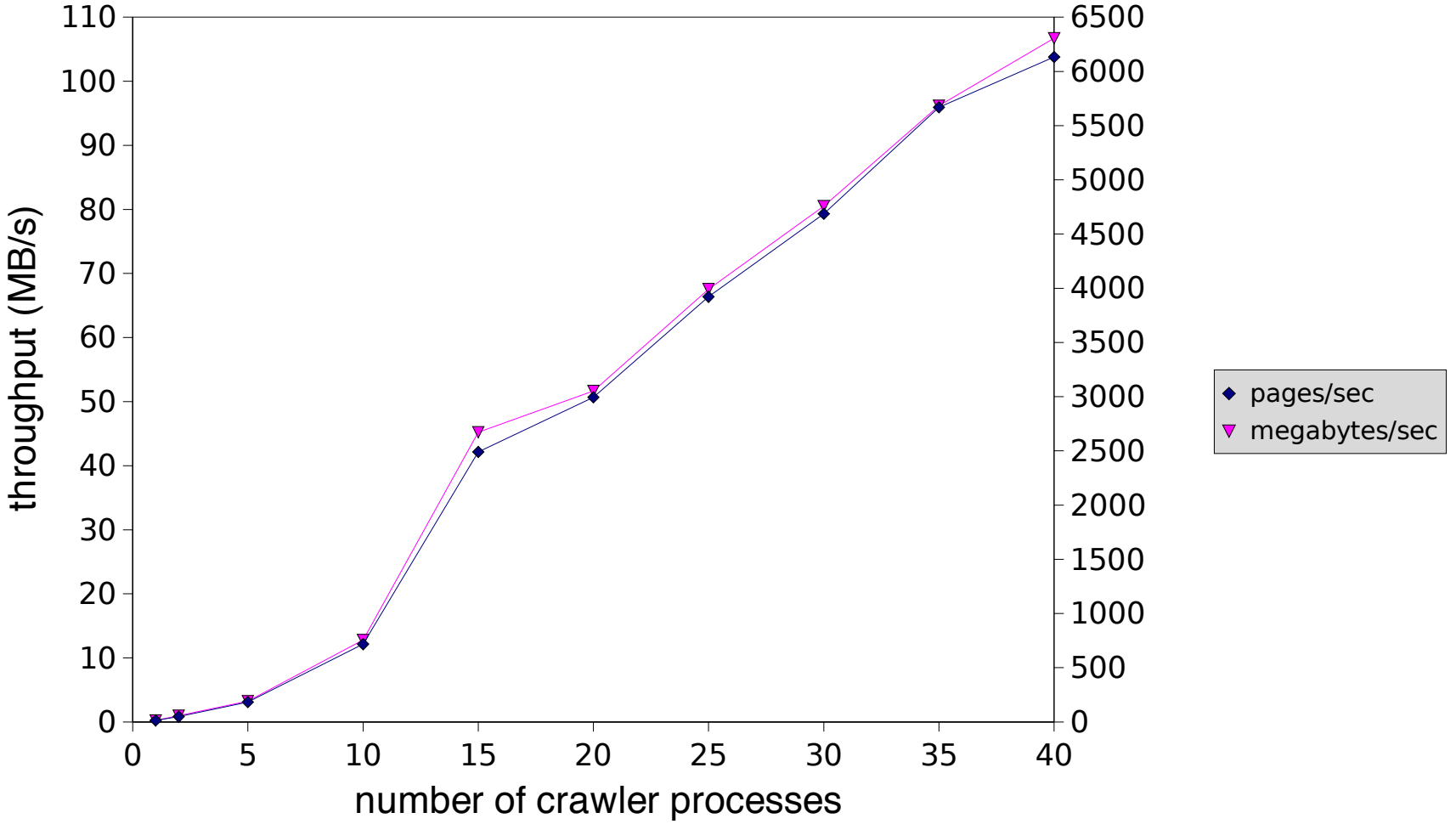
Architecture



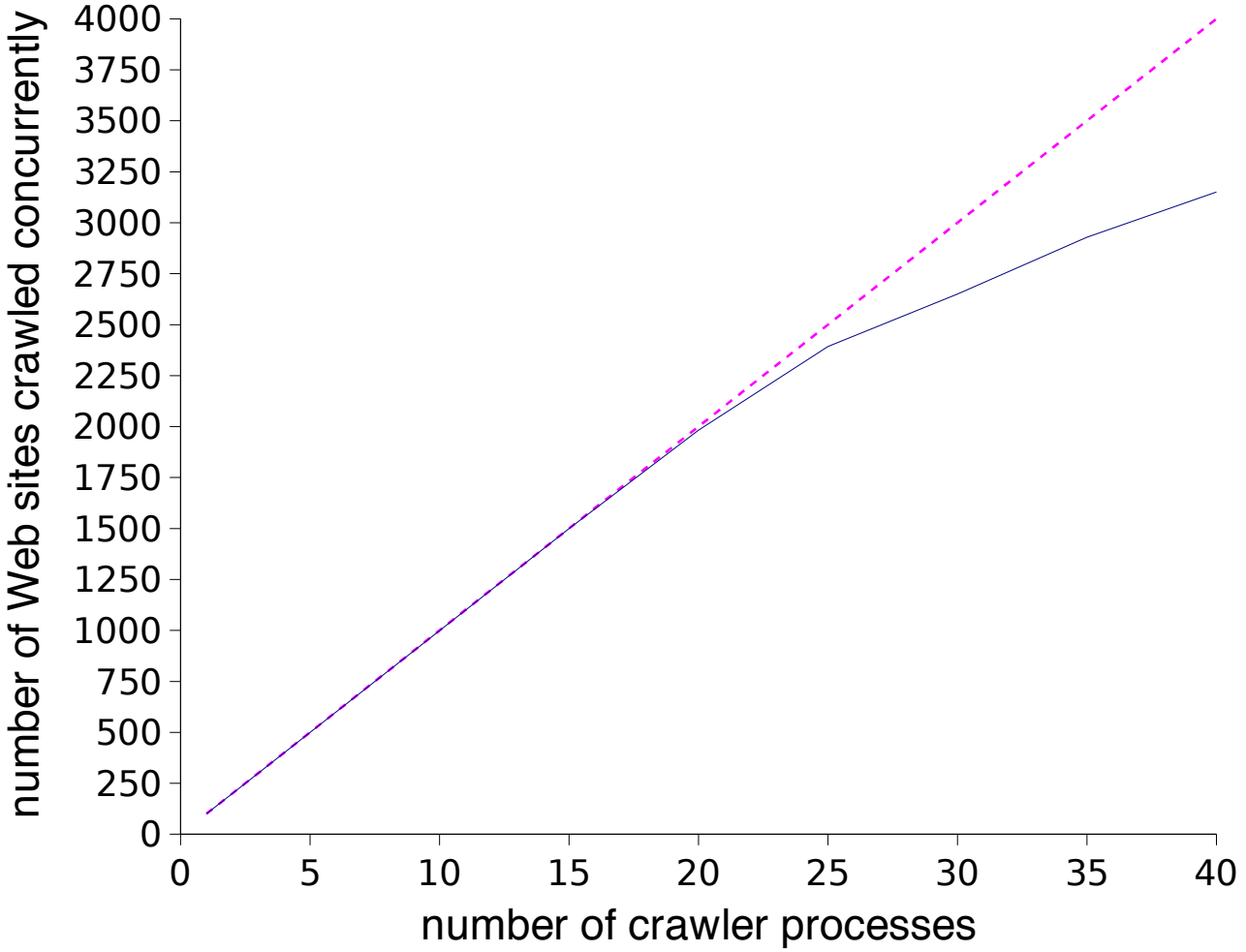
Site Crawler

- Crawl by site (FQDN)
- Independent units of work
- Coverage
- Database-backed controller
- Prevents excess crawl
- Site-specific crawl paramaters:
  pages, depth, pause, types, ...

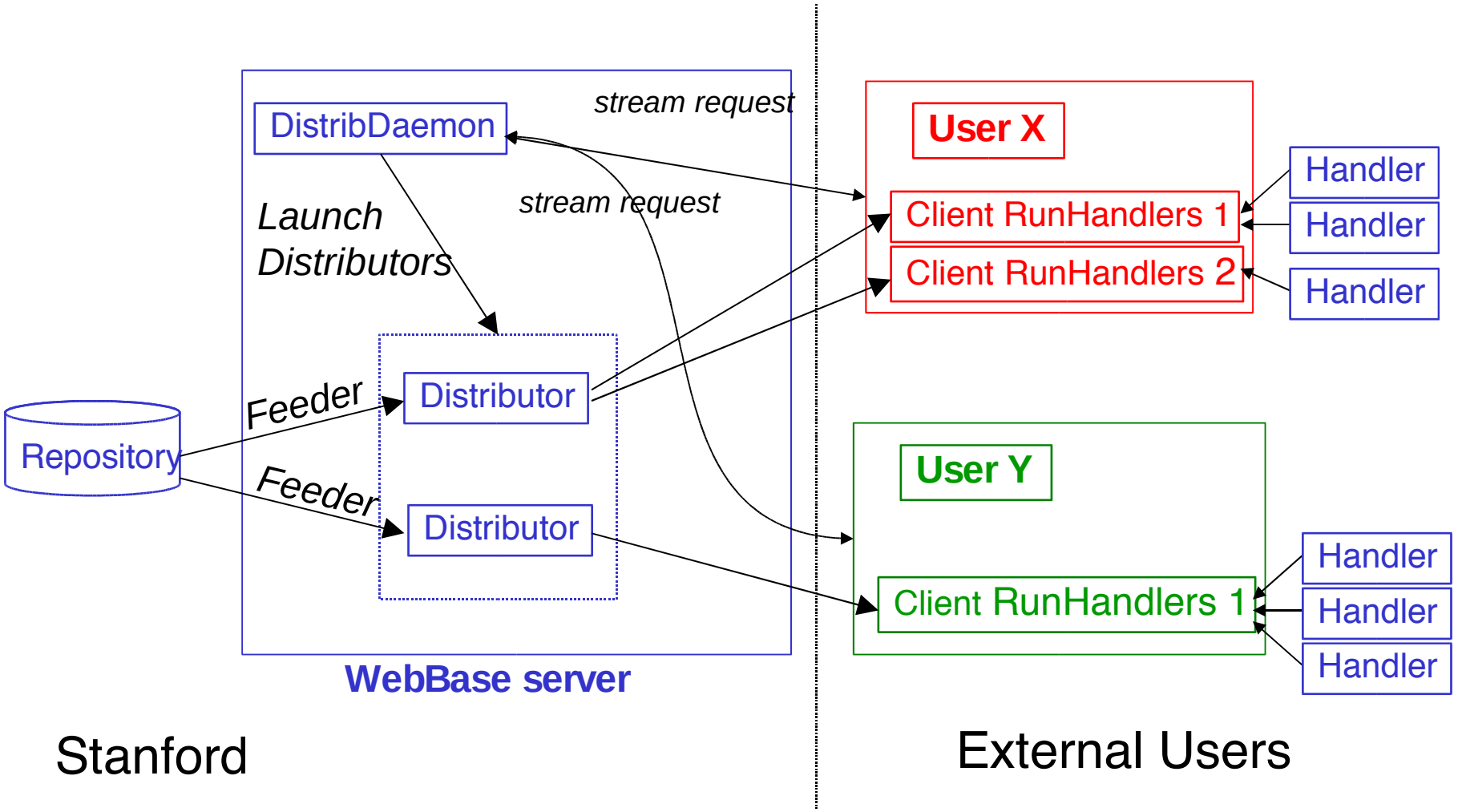
Crawler Performance



Crawler Performance



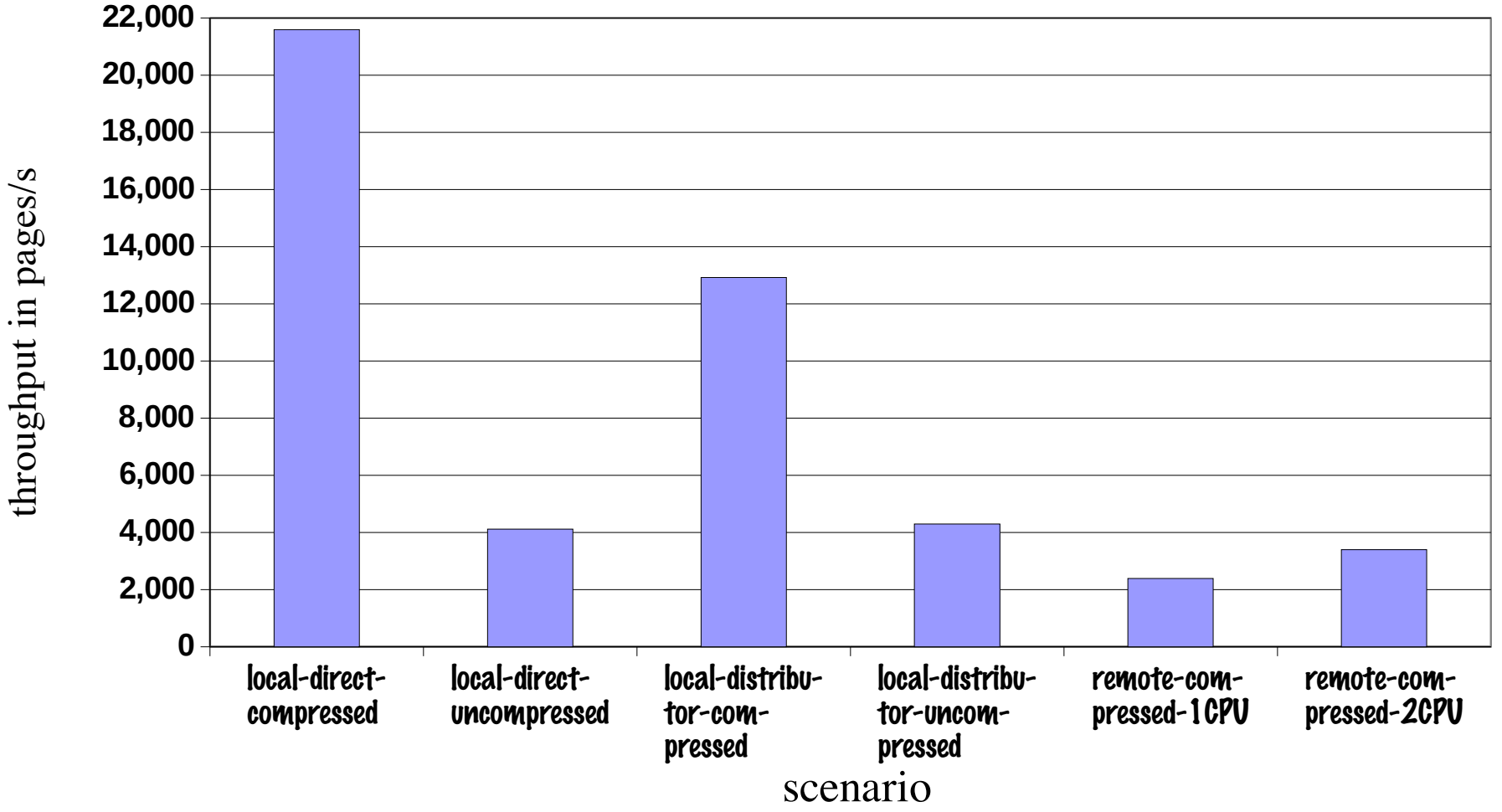
Distributor



Request Interface

- Crawl
  - Choose a specific crawl space and date of interest
- Site or sequence of Web sites
  - Request specific sites, or all of them in crawl order
- Number of pages
  - Restrict fetch of site to a shallow crawl
- Opaque page offsets
  - Restart from a known point

Distributor Performance



For More Information

- In submission:
  http://dbpubs.stanford.edu/pub/2004-34
- On the Web:
  http://www-diglib.stanford.edu/~testbed/doc2/WebBase/
- Source code:
  ftp://db.stanford.edu/pub/digital_library/