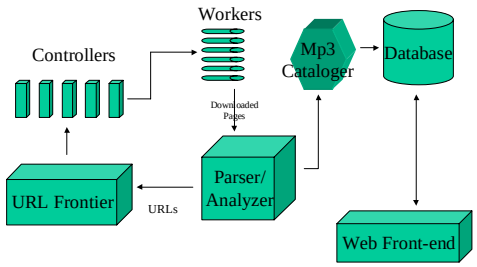
CWJ^{3}P

Chris Waterman
Jeff Phillips
Jason Jenks

Overall Crawler Design



Problems Discovered

- 64 Megabyte heap
- Tendency to get “slammed” with links to the same server
- Memory Leaks ???
- HTTP Timeouts

Solutions

- Disk-based data-structures w/ caches
- Limit worker queue’s size
- Asynchronous download of URLs

Things To Do Differently

- Different language (c++)
- Better way to keep workers working
- Distributed computing
- Phrase searching support