OverCite:

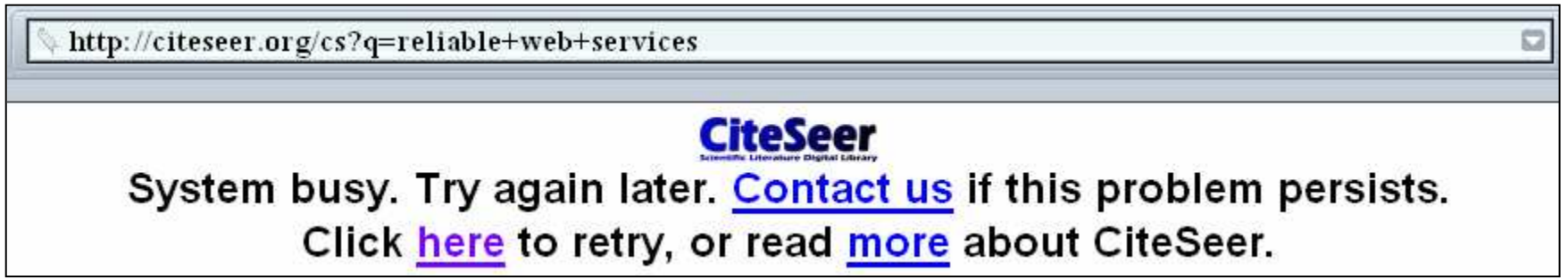
A Cooperative Digital Research Library

Jeremy Stribling, Isaac G. Councill, Jinyang Li,
M. Frans Kaashoek, David Karger,
Robert Morris, Scott Shenker

Everyone Loves CiteSeer

- Online repository of academic papers
- Crawls, indexes, links, and ranks papers
- Important resource for CS community

Everyone Hates CiteSeer



- Burden of running the system forced on one site
- New resource-heavy features difficult to support
- Scalability to large document sets uncertain

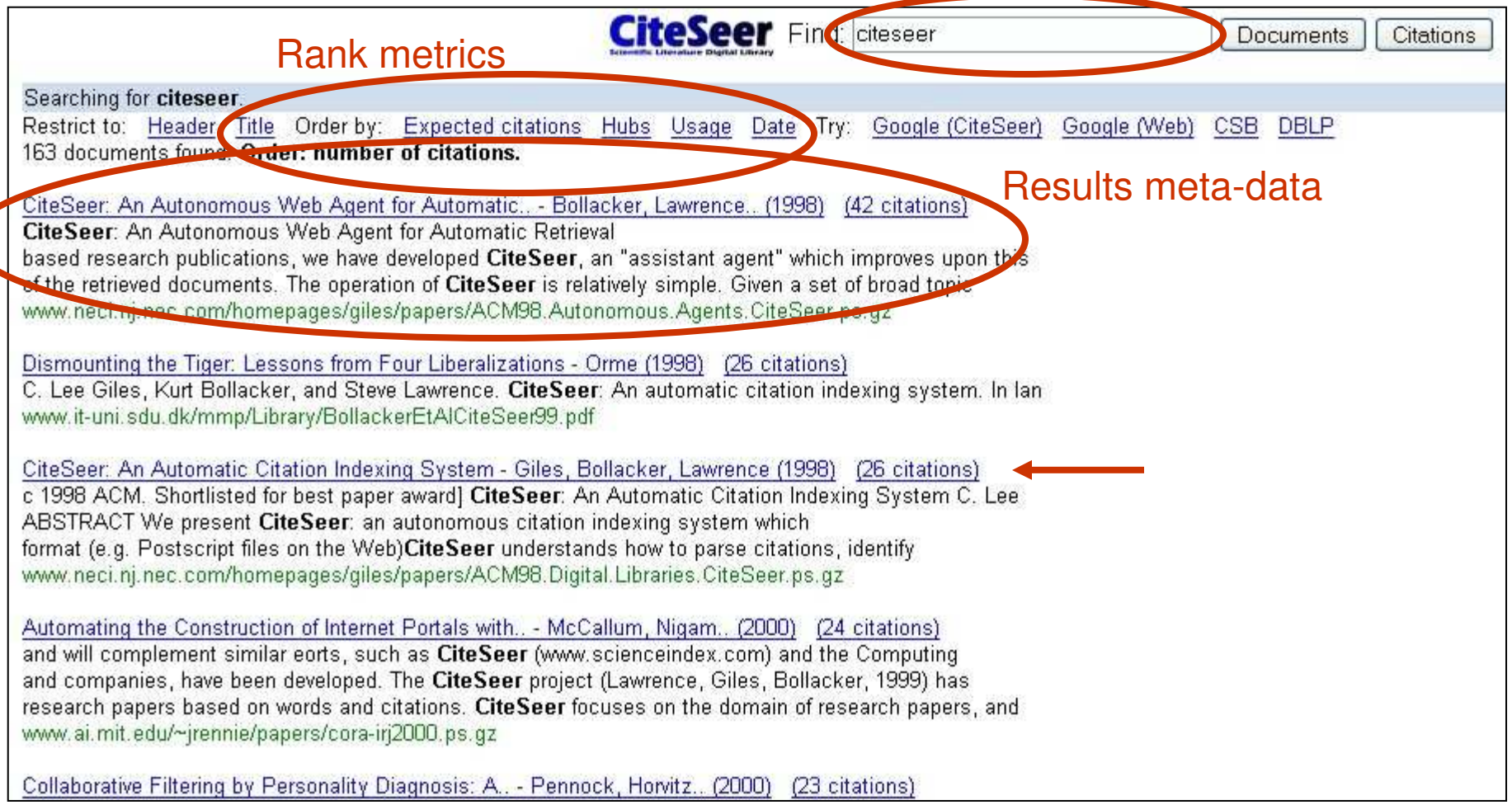
What Can We Do?

- Solution #1: All your © are belong to ACM
- Solution #2: Donate money to PSU
- Solution #3: Run your own mirror
- Solution #4: Aggregate donated resources

Solution: OverCite

- Distribute crawling, storage, queries via a DHT
- Goal: Distribute work *evenly* among nodes
- 100 nodes → 30x performance benefit

CiteSeer Today



CiteSeer Today

CiteSeer: An Automatic Citation Indexing System (1998) (Make Corrections) (34 citations)
C. Lee Giles, Kurt D. Bollacker, Steve Lawrence
Digital Libraries 98 - The Third ACM Conference on Digital Libraries

CiteSeer Home/Search Context Related

View or download:
nec.com/homepages/ ies.CiteSeer.ps.gz
nec.com/homepages/ sdl98letter.ps.Z
Cached: PS.gz PS PDF Image Update Help

From: nec.com/homepages/giles/papers (more)
From: nec.com/homepages/lawren papers
Homepages: C.Giles HPSearch (Update Links)

Cached doc

(Enter summary)

Rate this article: 1 2 3 4 5 (bes
Comment on this article

Abstract: We present CiteSeer: an autonomous citation indexing system which indexes academic literature in electronic format (e.g. Postscript files on the Web). CiteSeer understands how to parse citations, identify citations to the same paper in different formats, and identify the context of citations in the body of articles. CiteSeer provides most of the advantages of traditional (manuall constructed) citation indexes (e.g. the ISI citation indexes) including: literature retrieval by following... (Update)

Cited by: More
Search Engine-Crawler Symbiosis: Adapting to - Community Interests Gautam
eBizSearch: A Niche Search Engine for e-Business - Lee Giles Yves
Natural Communities in Large Linked Networks - John Hopcroft Omar (2003)

Cited by

Similar documents (at the sentence level):
9.6%: CiteSeer: An Autonomous Web Agent for Automatic.. - Bollacker, Lawrence.. (1998)

Active bibliography (related documents): More All
2.6: Dismounting the Tiger: Lessons from Four Liberalizations - Orme (1998)
0.6: Essays of an Information Scientist: Creativity, Delayed.. - Vo Curre Nt (1989)
0.6: CitEc: an Autonomous Citation Index for Economics - Krichel, Lawrence (1999)

Similar documents based on text: More All
1.6: A System For Automatic Personalized Tracking of.. - Bollacker, Lawrence.. (1999)
1.4: Distributed Error Correction - Lawrence, Bollacker, Giles (1999)
1.4: Autonomous Citation Matching - Lawrence, Giles, Bollacker (1999)

Related documents from co-citation: More All
8: The anatomy of a large-scale hypertextual Web search engine - Brin, Page
6: Citation Indexing: Its Theory and Application in Science (context) - Garfield - 1979
6: Identifying and merging related bibliographic records - Hylton - 1996

BibTeX entry: (Update)

C. Lee Giles, Kurt Bollacker, and Steve Lawrence. CiteSeer: An automatic citation indexing system. In Ian Witten, Rob Akscyn, and Frank M. Shipman III, editors, Digital Libraries 98 - The Third ACM Conference on Digital Libraries, pages 89--98, Pittsburgh, PA, June 23--26 1998. ACM Press. http://citeseer.csail.mit.edu/article/giles98citeseer.html More

```
@inproceedings{ giles98citeseer,
    author = "C. Lee Giles and Kurt Bollacker and Steve Lawrence",
    title = "{CiteSeer}: An Automatic Citation Indexing System",
```

CiteSeer Today: Local Resources

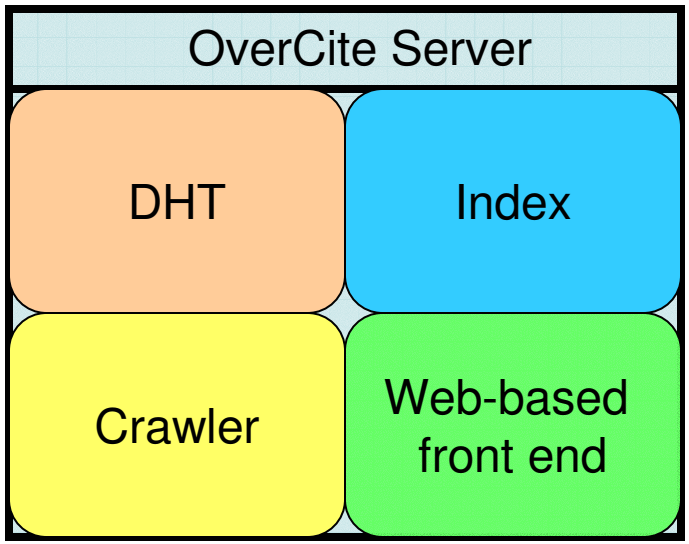
# documents	715,000 ←
Document storage	767 GB
Meta-data storage	44 GB
Index size	18 GB
Total storage	829 GB ←
Searches	250,000/day ←
Document traffic	21 GB/day
Total traffic	34.4 GB/day ←

Challenges

- Distribute storage for parallel speedup
- Replicate storage for availability
- Parallelize query load for load-balancing
- Distribute crawling for improved throughput

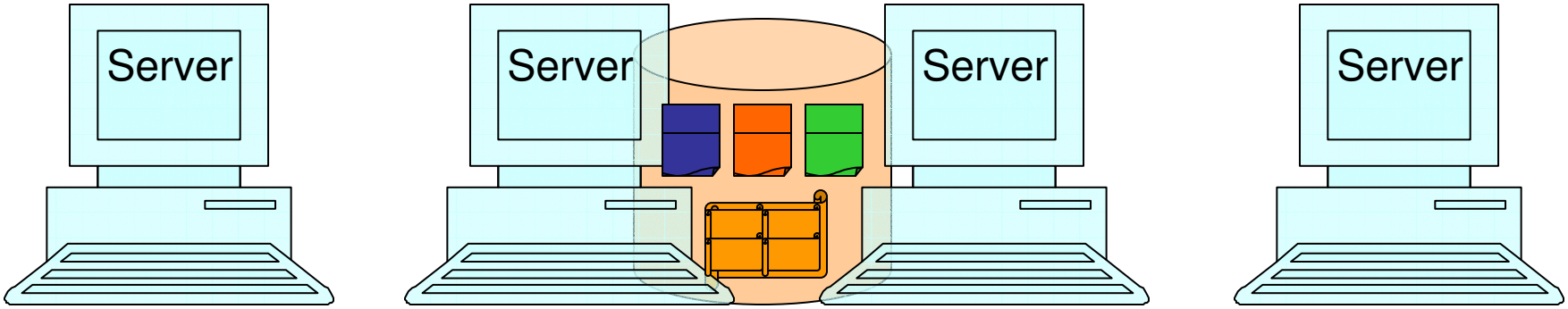
OverCite Architecture

- Several hundred relatively stable nodes
- Each node runs four separate components



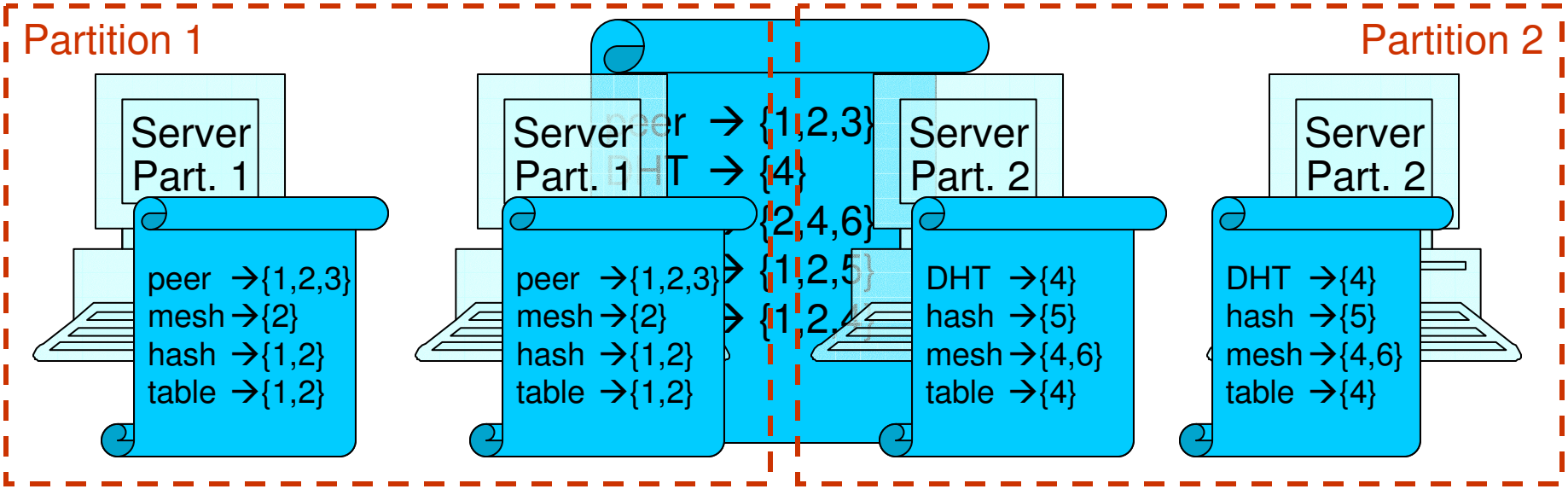
Documents and Meta-data

- DHT stores papers for parallelism/availability
- DHT stores meta-data tables
  - e.g., document IDs → {title, author, year, etc.}
- Use large-state DHT like OneHop [Gupta et al. NSDI '04] or Accordion [Li et al. NSDI '05]

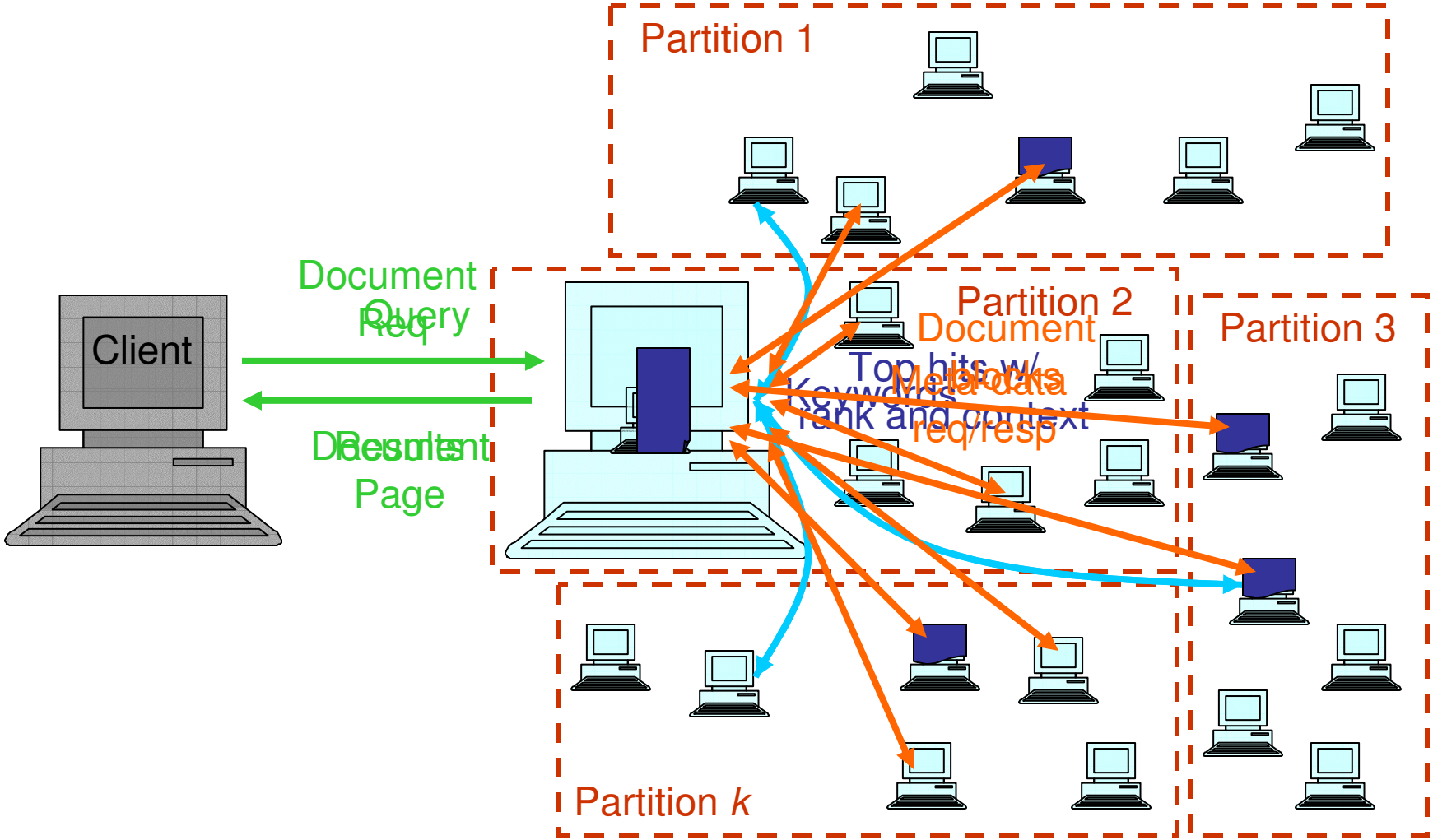


Index

- Goal: Parallelize queries
- Partition by document, not keyword
- Divide the index into *k* partitions
- Each query sent to only *k* nodes



Anatomy of a Query



OverCite, IPTPS 2005

Properties of OverCite

Operation	**CiteSeer Today**	**OverCite (per node)**
Crawling	0.735 MB/doc	**(3/*n*)x**
Storage	829 GB	**(3/*n*)x**
Query bw	--	**(3.5/*n* GB)/day**
Serving Documents	21 GB/day	**(2/*n*)x**

- Performance scales with *n*/3 system-wide

CiteSeer Extensions

- Document alerts (e.g., SmartSeer)
- Amazon-like recommendations
- Plagiarism checking
- Expand document collection
  - Larger set of disciplines
  - Preprints and public reviews

Related Work

- Search on DHTs
  - Partition by keyword

    [Li et al. IPTPS '03, Reynolds & Vadhat Middleware '03, Suel et al. IWWD '03]
  - Hybrid schemes

    [Tang & Dwarkadas NSDI '04, Loo et al. IPTPS '04, Shi et al. IPTPS '04]
- Distributed crawlers

  [Loo et al. TR '04, Cho & Garcia-Molina WWW '02, Singh et al. SIGIR '03]
- Parallel speedup

  [Dean & Ghemawat OSDI '04]

Summary

- A design for storing and coordinating a digital repository using a DHT
- Spreads load across many volunteer nodes
- Support for resource-intensive new features
- Run CiteSeer as a community