Parallel and Distributed Information Retrieval

Murad Kamalov

Outline

- Introduction
- Parallel and Distributed Information Retrieval
    - Query throughput
    - Query response time
- P2P Information Retrieval
    - Chord
- Conclusions

Background

- MIMD - Multple Instruction stream Multiple Data stream
    - Inter-process communication using shared memory
- Distributed system - separate machines
    - Inter-process communication using TCP/IP
- Search Index
    - Structure to facilitate fast information retrieval

	Term 1	Term 2	Term 3
Document 1	4	5	7
Document 2	5	1	3

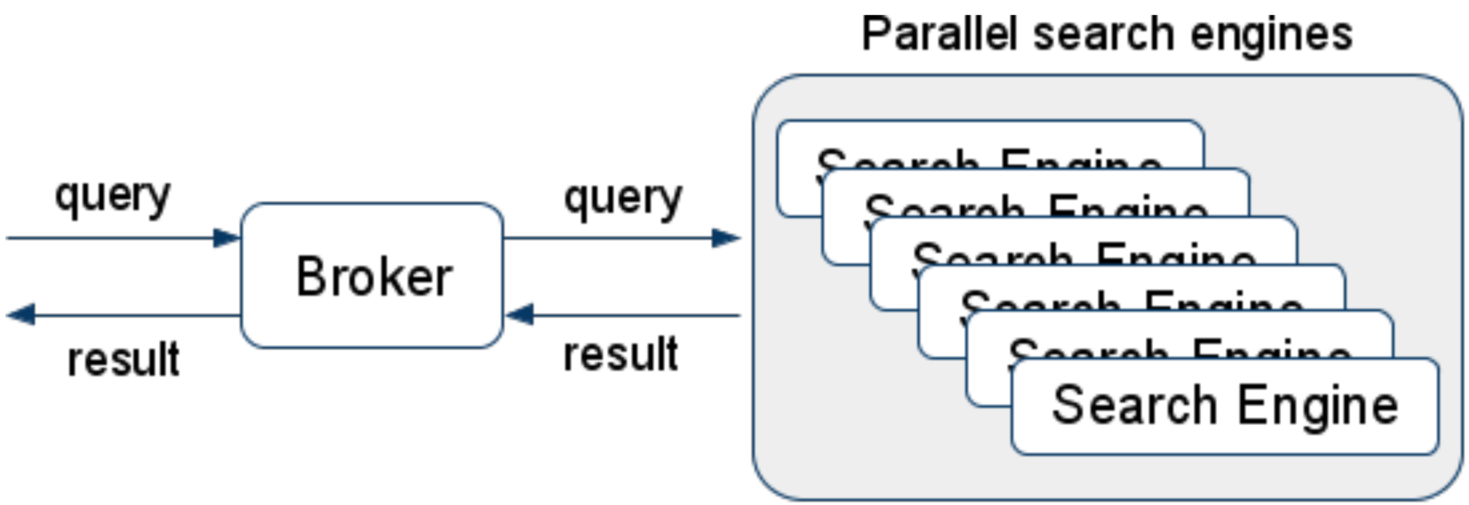
IR, Parallel and Distributed IR

- Information Retrieval (anno 1880)
  - Searching Documents
  - Searching Information within Documents
  - Searching metadata about Documents

- Parallel and Distributed IR
  - Improving query throughput
  - Improving query response time

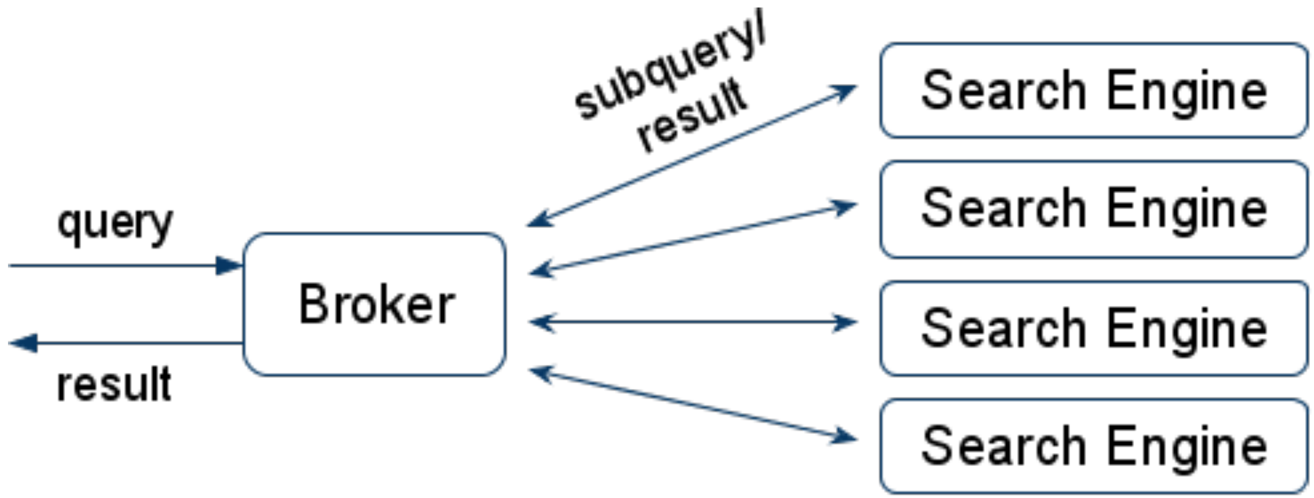
INFORMATION RETRIEVAL

Improving query throughput



- Broker == Load Balancer
- **Search Engines** either share search index or have their own copy of one
- Applicable on both MIMD architectures and distributed systems

Improving Query response time

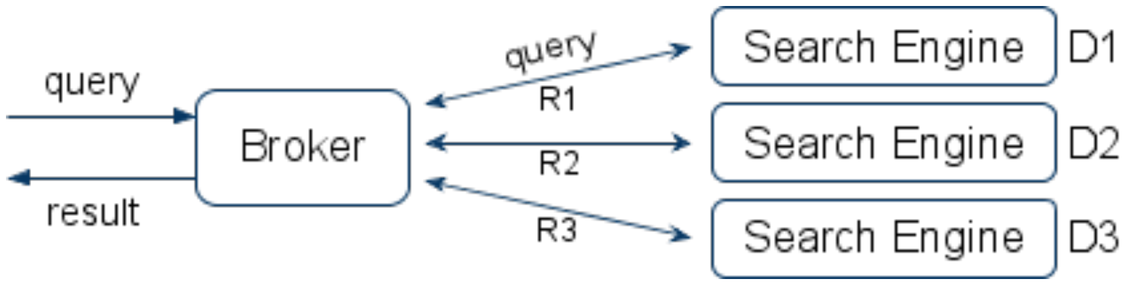


- Broker
  - Splits query to sub-queries and sends them to multiple search engines
  - Aggregates search results

Improving Query response time (2)

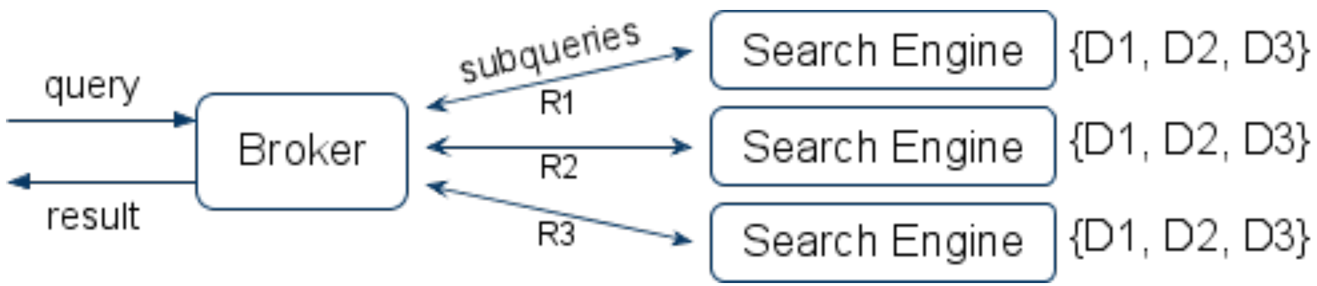
- Two ways to improve query response time
- Documents: {D1, D2, D3}

Case 1: execute query on different documents



result = R1 ∪ R2 ∪ R3

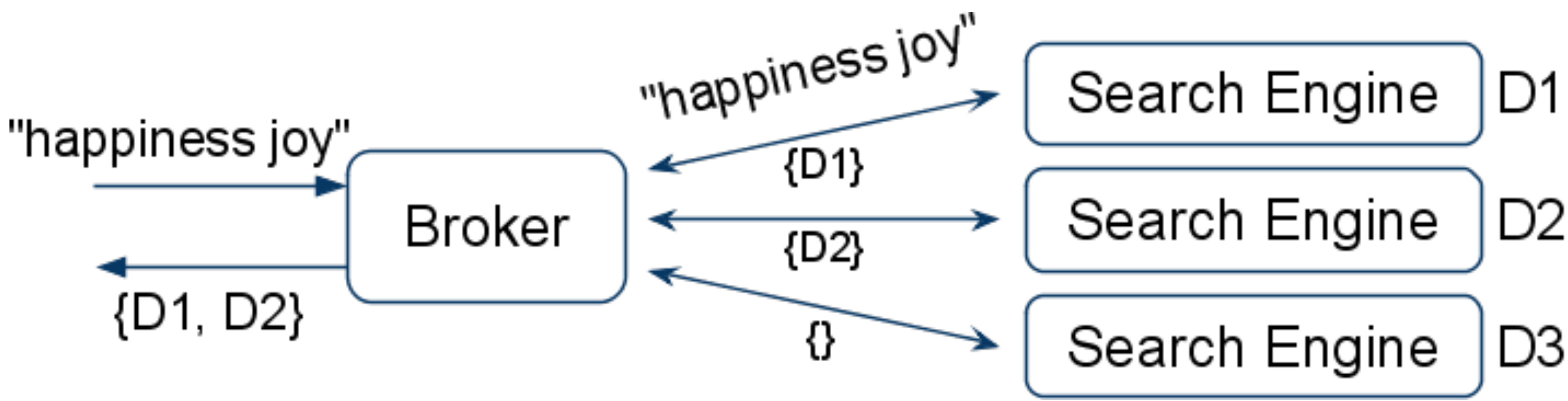
Case 2: split query



query = Q1 ∪ Q2 ∪ Q3   result = R1 ∪ R2 ∪ R3

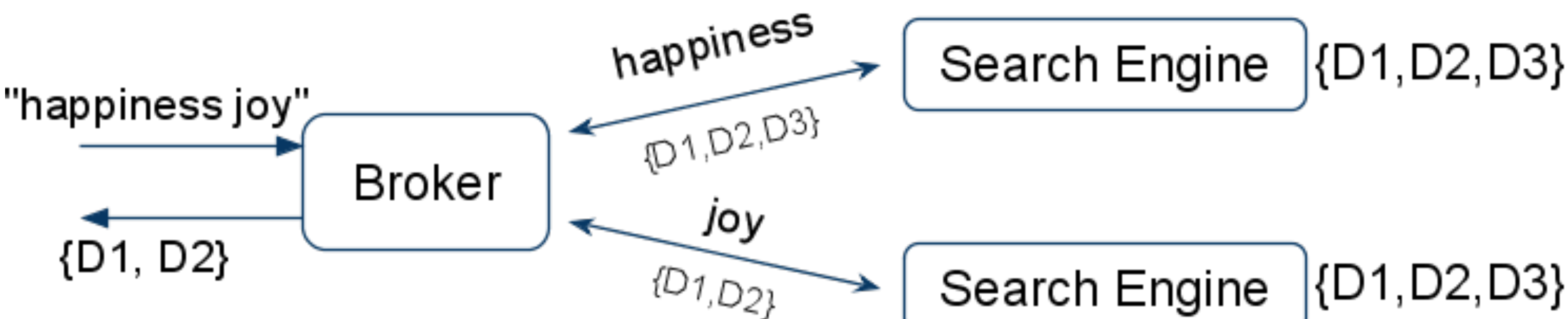
Case 1: Example

- D1 = ***Happiness*** is a feeling characterised by stupidness, weird, well being, or ***joy***
- D2 = ***Happiness*** is feeling of ***joy*** .
- D3 = The pursuit of ***Happiness***
- QUERY = "happiness joy"



Case 2: Example

- D1 = ***Happiness*** is a feeling characterised by stupidness, weird, well being, or ***joy***
- D2 = ***Happiness*** is feeling of ***joy*** .
- D3 = The pursuit of ***Happiness***
- QUERY = "happiness joy"



- RESULT = {D1, D2, D3} ∪ {D1,D2} = **{D1, D2}**

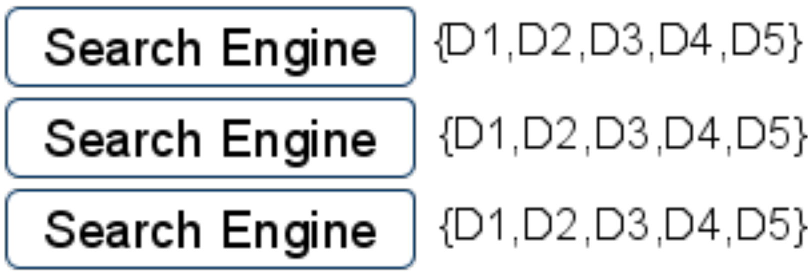
Query Processing

- Select collections to search
- Distribute query to selected collections
- Evaluate query at distributed collections in parallel
- Combine results from distributed collections into final result

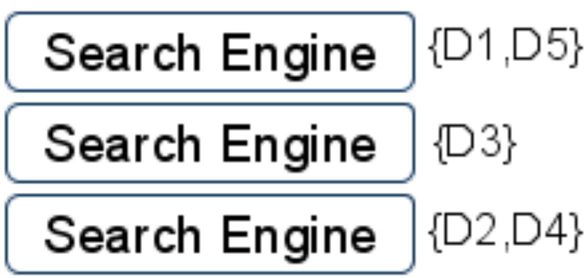
How to partition documents into collections?
How to select collections to search?

Collection Partitioning

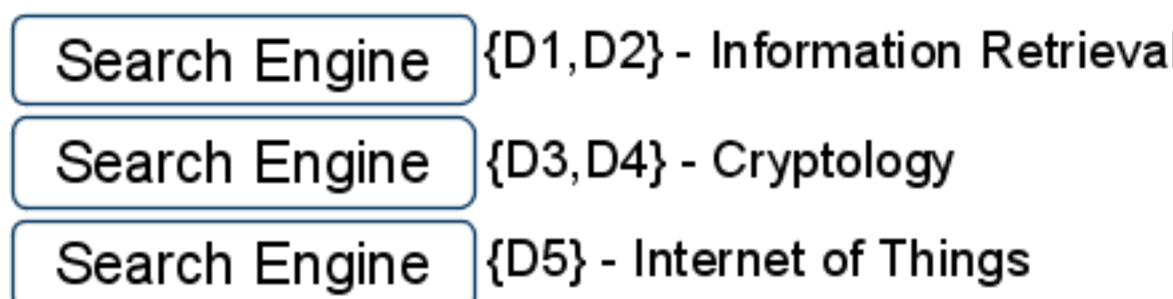
- Replicate collections across all search servers



- Random distribution



- Semantic distribution
  - Subject specific
  - Alphabetical



Source Selection

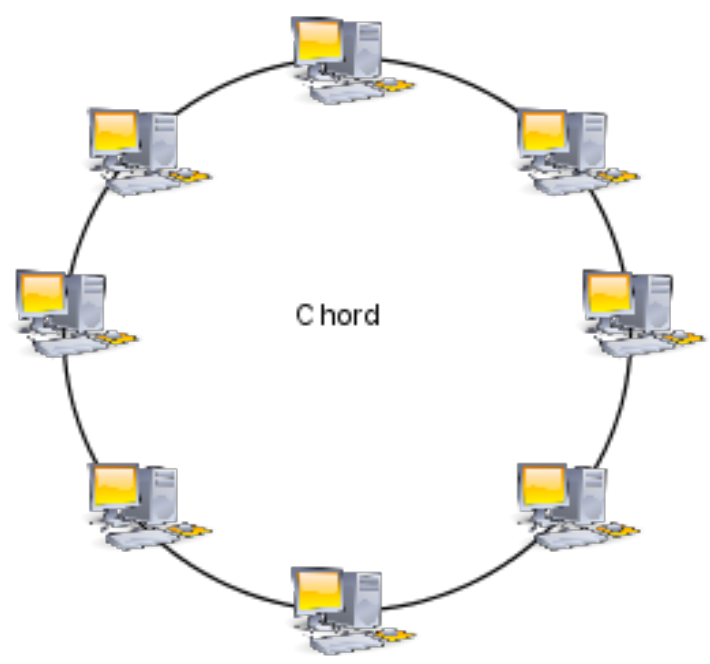
- Specifies, how broker selects search engines to which query will be sent

- All collections equally likely to contain document
    - random partitioning
    - significant semantic overlap
- Collection ranking
    - semantic distribution

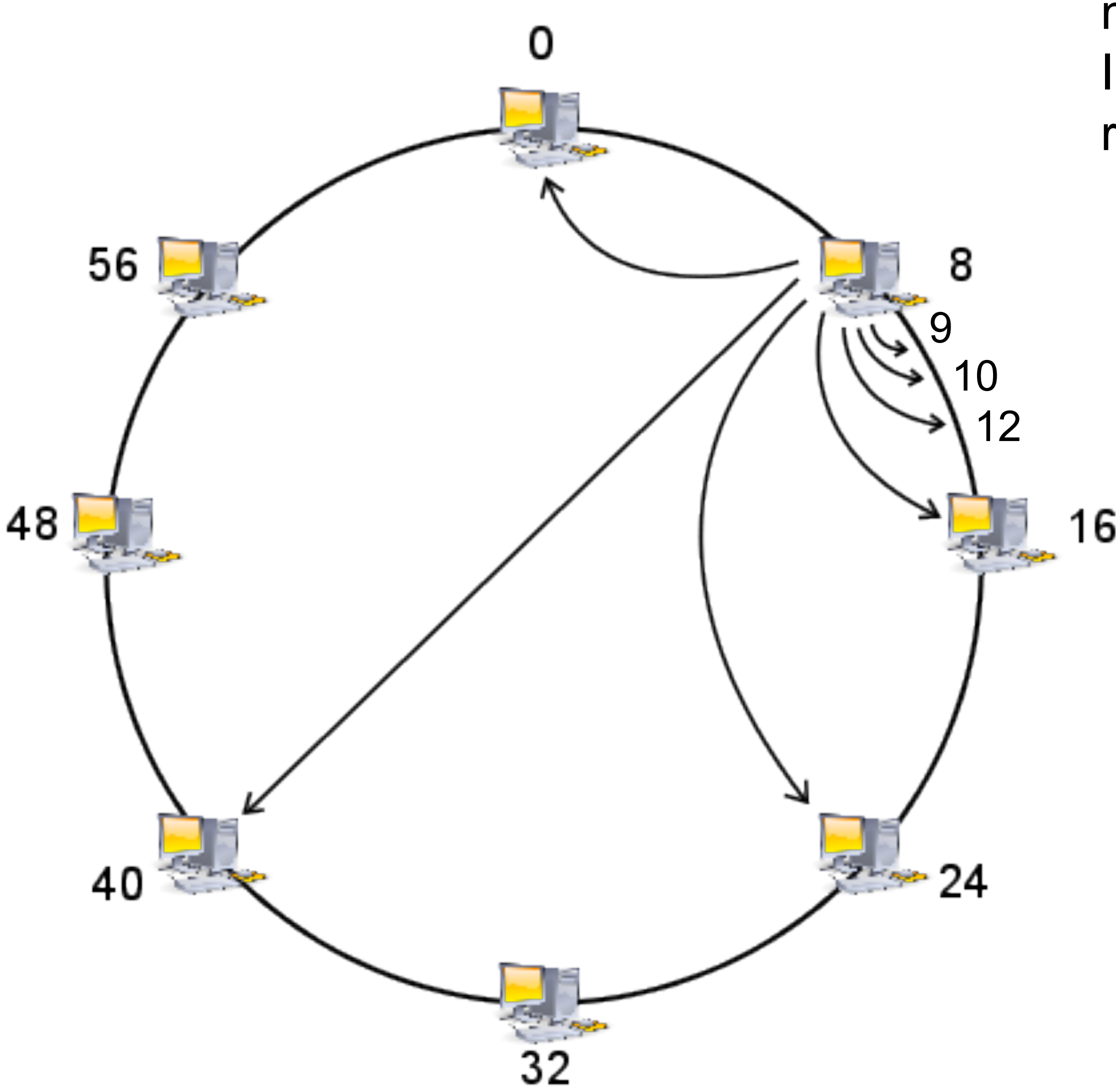
P2P Information Retrieval

- Network consists out of peers.
- Documents are distributed among the peers

- Chord
  - P2P protocol for retrieving documents from ring of equal peers
  - Fully decentralised
  - Widely use in DHT implementations
  - Ring represents ID space of length 2^m
  - Each node has m-bit ID = SHA1(IP address)
  - Each document has m-bit ID (SHA1 based)
  - Each node maintains routing table with at most m entities
  - IDs of documents and nodes are in in the same namespace



Chord: routing table example

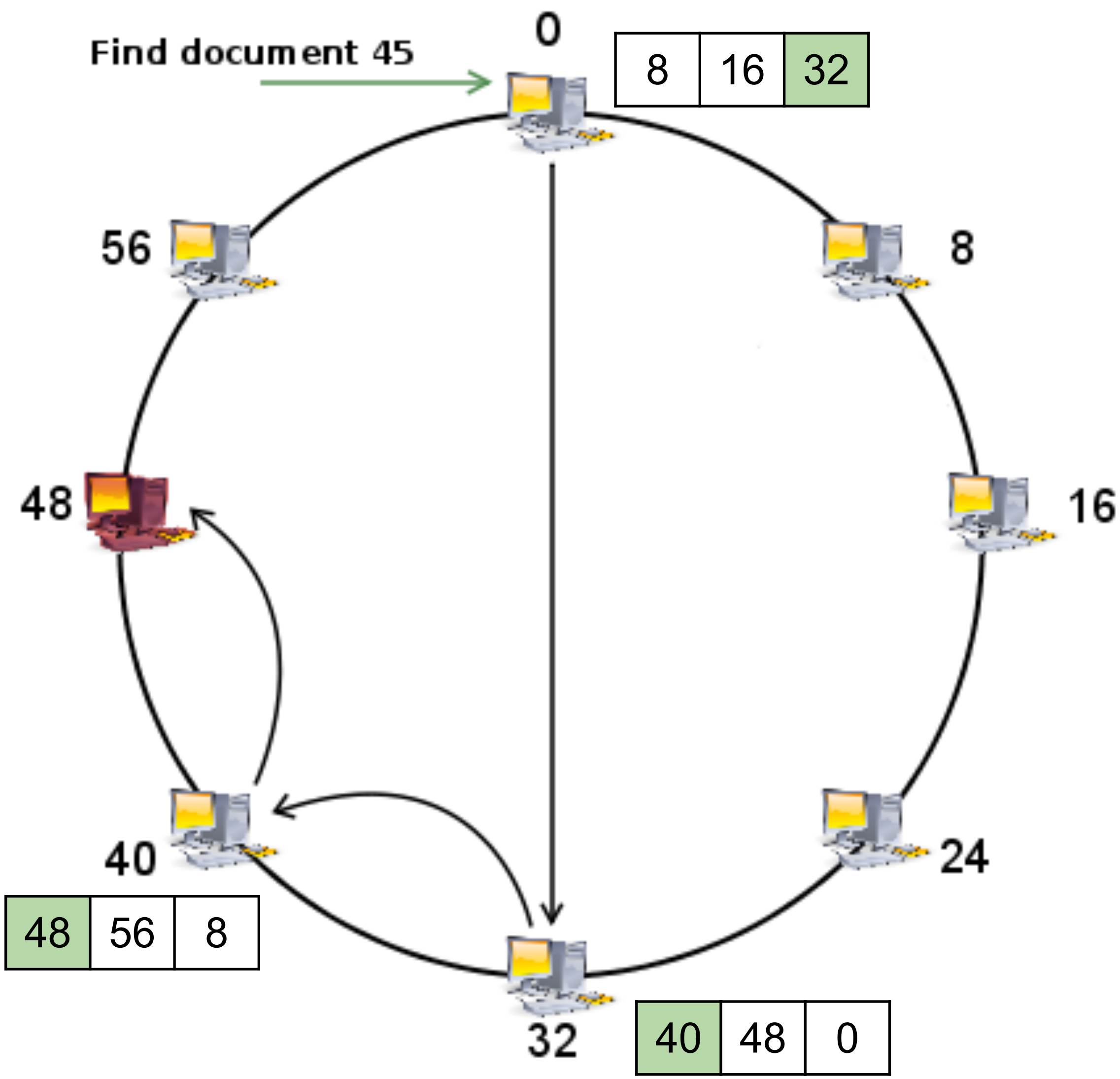


m = 6
ID space = $2^6 = 64$
routing entry = $N + 2^{j-1}$, for j = 1..m

- In our example N = 8
- Thus, we get following routing table

j	Eq.	ID	Node
1	$8 + 2^0$	9	16
2	$8 + 2^1$	10	16
3	$8 + 2^2$	12	16
4	$8 + 2^3$	16	16
5	$8 + 2^4$	24	24
6	$8 + 2^5$	40	40

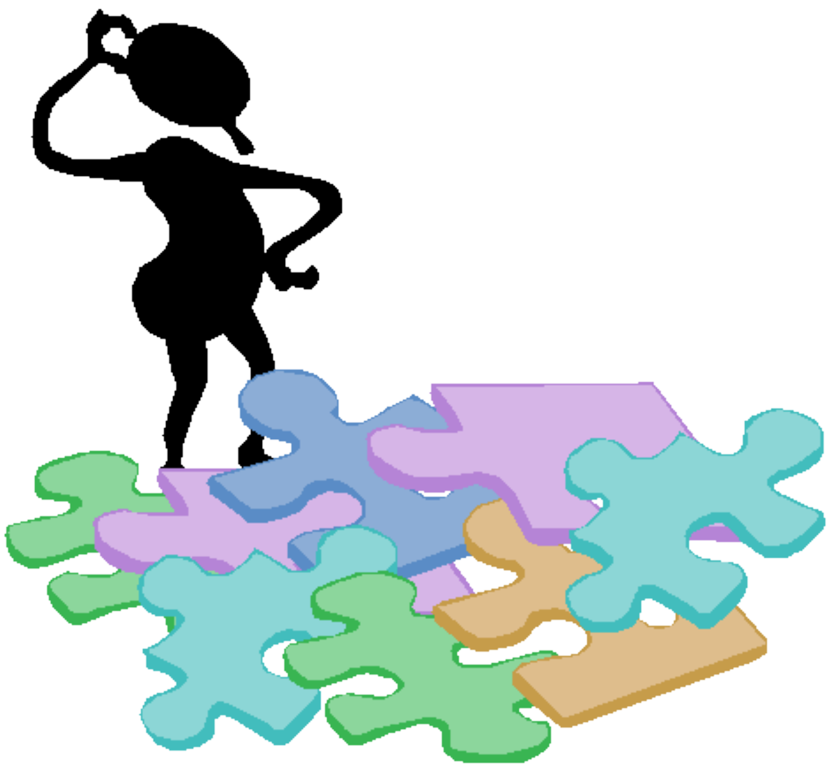
Chord: Document lookup example



- Send request for document 45, to node 0
- Each node is responsible for documents which ID's are smaller/equal to his own ID and bigger than ID of previous node
- E.g Node 48 is responsible for documents in range 41.. 48

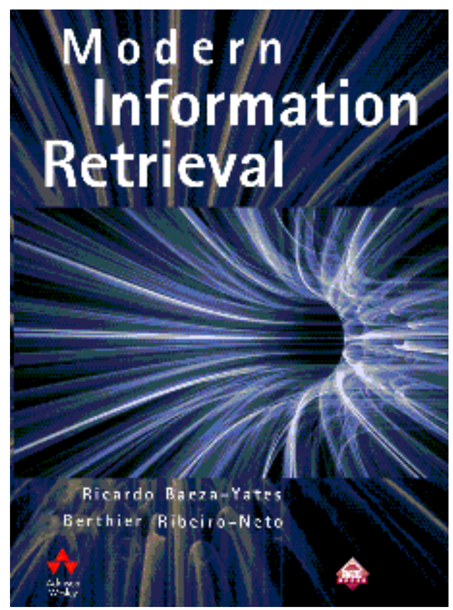
Conclusions

- IR is large area with many subareas
- Solutions based on Parallel and Distributed IR are around for a while
  - We reviewed some of them
- Recently a lot of attention was paid to P2P technologies for Information Retrieval
  - Cheaper cost of deployment



References

- Modern Information Retrieval, Chapter 9, Parallel and Distributed IR, book by Ricardo Baeza-Yates and Berthier Ribeiro-Neto
- Chord: A Scalable Peer-to-peer Lookup Protocol for Internet Applications. *Ion Stoica , Robert Morris*



Thank You!