

# Towards Collaborative Search in Digital Libraries Using Peer-to-Peer Technology

---



MAX-PLANCK-GESELLSCHAFT

## 6th DELOS Workshop on Digital Library Architectures

S. Margherita di Pula (Cagliari), Italy

24-25 June 2004

## Matthias Bender & Christian Zimmer

Max-Planck-Institut für Informatik, Saarbrücken, Germany

Databases and Information Systems Group

# Talk Outline

---

- | Motivation
- | P2P Architectures
- | Design Fundamentals
- | Implementation
- | Open Questions
- | Conclusion & Future Work
  
- | Questions & Discussion





# Motivation (1)

## What is a “Peer-to-Peer System”?

- | Decentralized, self-organizing, highly dynamic loose coupling of many autonomous computers

## Main advantages:

- | High Scalability
- | Load Balancing
- | No Single Point of Failure

**„Why ask one, if you can ask thousands?“**

## Main problem

- | Efficient location of nodes in a distributed P2P architecture



# Motivation (2)

---

## Digital Libraries & Peer-to-Peer

### Peers can be:

- | Traditional Digital Libraries (DBLP, ACM...)
- | Web Portals (Amazon, IMDB...)
- | Web Encyclopedias (Wikipedia...)
- | Web Users with bookmarks
- | ....

➡ Concept of Digital Libraries is broadening

# P2P Architectures (1)

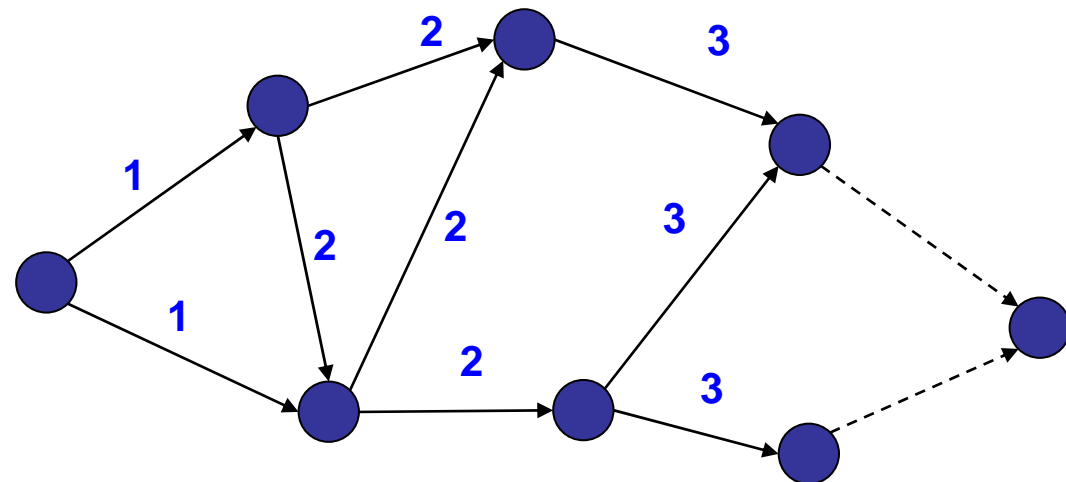
## Unstructured Architectures

Example: GNUTELLA

- | Learn and remember neighbors
- | Message flooding: forward messages (with „Time-to-Live“ value) to all neighbors

### Disadvantages

- | Unnecessary messages
- | Messages do not necessarily reach all peers – no guarantees





# P2P Architectures (2)

## Structured Architectures

### Example: CHORD

- | Efficiently maps keys on peers in a distributed manner
- | Supports exactly one operation: given key, returns the peer currently responsible – **lookup(key)**

### Intuition:

- | Peers and keys are mapped to the same cyclic ID space using hash functions – peers form the so-called **Chord-Ring**
- | Every key is assigned to its closest successor peer on the ring

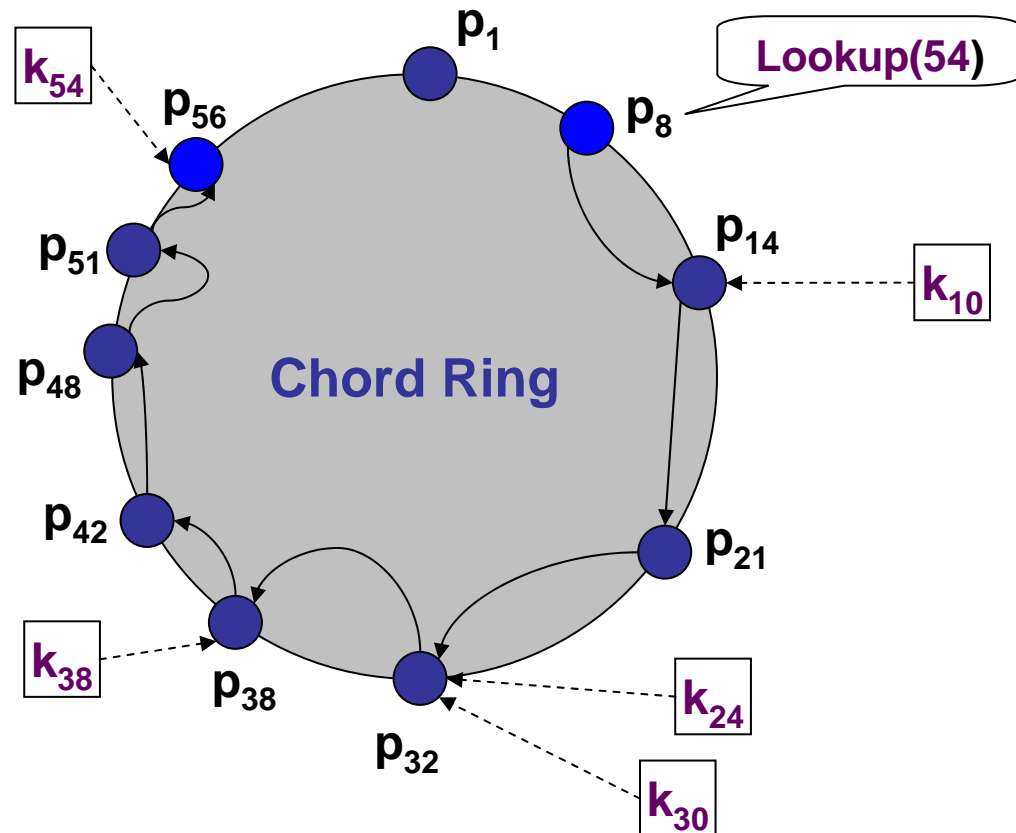
# P2P Architectures (3)

## Example:

- All  $n$  peers  $p_i$  are arranged on the ring and all keys  $k_j$  are assigned to their closest successor peers

## Naive routing approach:

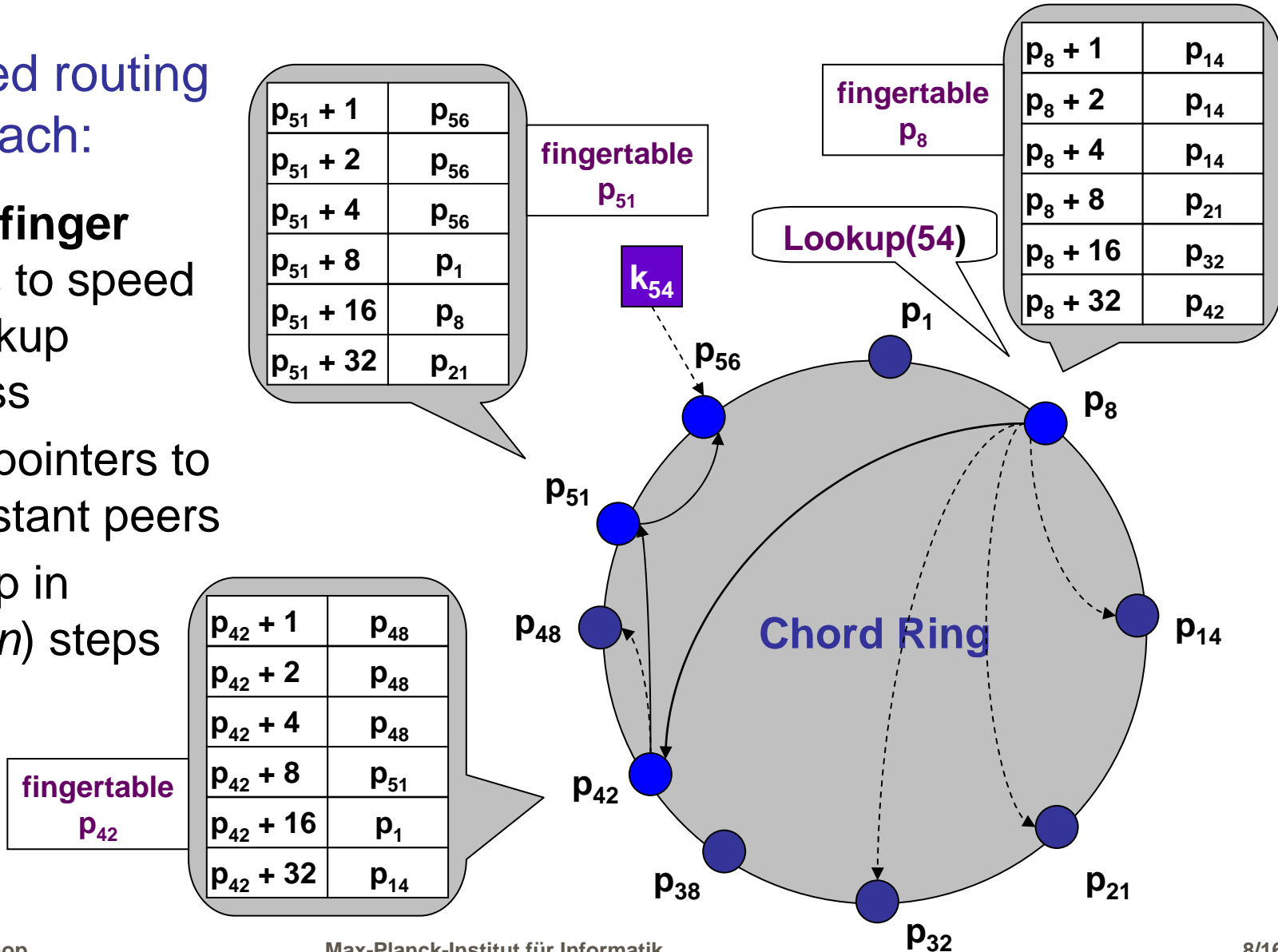
- Lookup requests are forwarded on the ring until the responsible peer is found
- Causes up to  $n$  message forwards



# P2P Architectures (4)

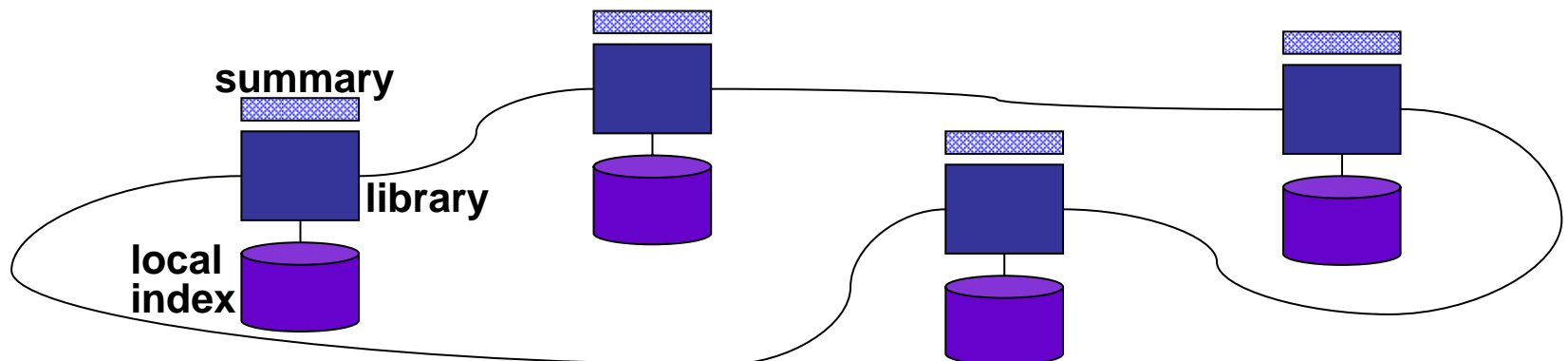
Enhanced routing approach:

- I Using **finger tables** to speed up lookup process
- I Store pointers to few distant peers
- I Lookup in  $O(\log n)$  steps



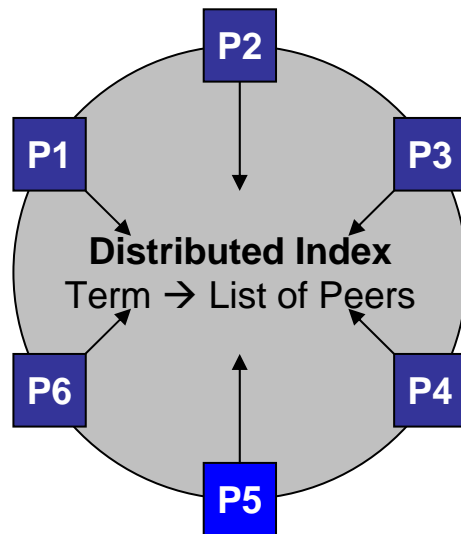
# Design Fundamentals (1)

- | Peers form a conceptually global, but physically distributed directory
- | Each library posts meta-information (**Posts**), e.g. per-term summaries of its local index
- | For every term the responsible peer maintains a **PeerList** containing all known Posts
- | Queries can be forwarded to suitable remote libraries using the meta-information in the global directory

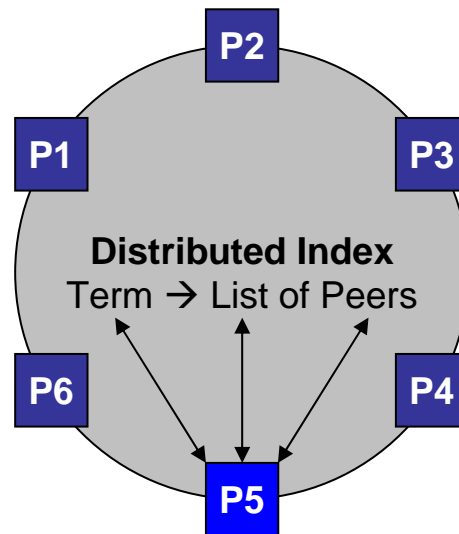


# Design Fundamentals (2)

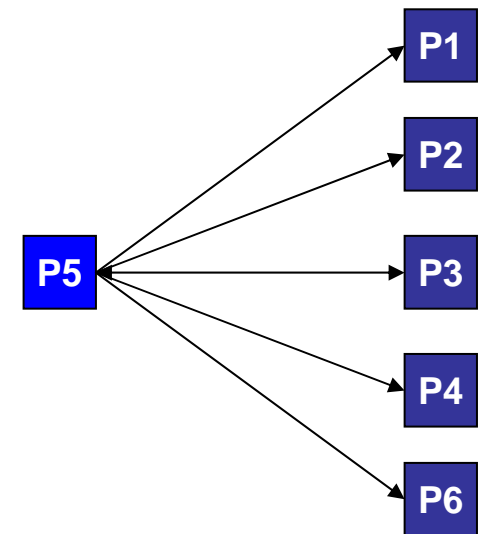
## Query Execution:



**Step 0:**  
Post per-term  
summaries of local  
indexes



**Step 1:**  
Retrieve list of peers  
for each query term



**Step 2:**  
Retrieve and  
combine local query  
results from peers



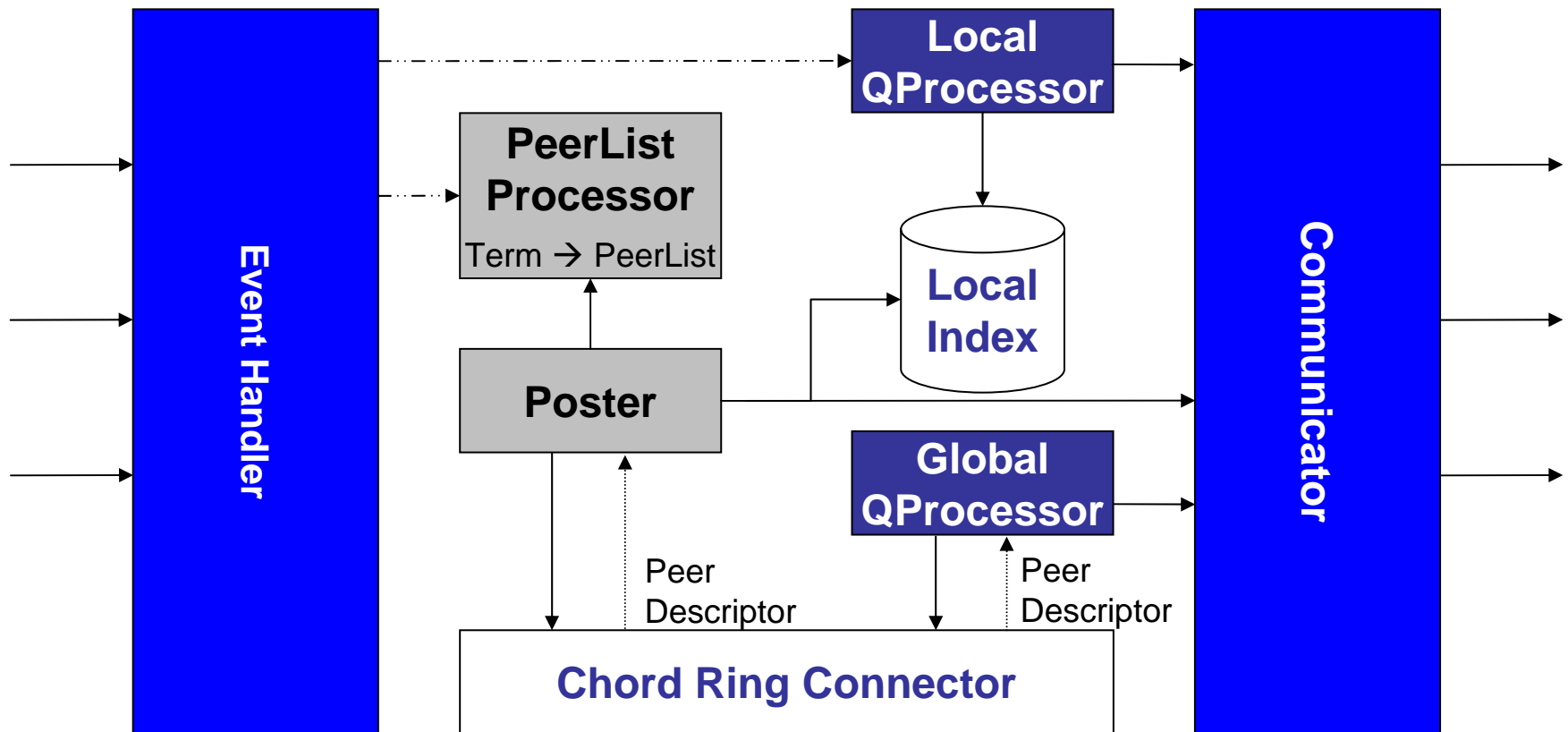
# Design Fundamentals (3)

## Why this approach?

- | Posting only meta-information about peers (libraries) instead of information about all documents saves network bandwidth (**Odissea**)
- | Gossiping algorithms or global directories replicated at every peer cause high data transfer in combination with update problems (**PlanetP**)
- | This approach can easily be extended by creating and combining multiple global directories, e.g. for index data or relevance feedback data

# Implementation (1)

## Architecture of a single library peer



# Implementation (2)

## Prototype Implementation

- | IR measures ( $tf$ ,  $idf$ ) to select suitable peers and to rank results

**P2P Search**

**Chord Component**

Local Chord Port: 9001  
Local Application Port: 9002  
Remote IP: localhost  
Remote Chord Port:

name	value
chord id	36155
ip	139.19.54.20
ring exponent	65536
ring exponent	16
succ id	36155

**Posts**

Post:

d (59093)  
a (63874)  
c (38842)  
b (18590)

**Queries**

IP:Port	URL	Title	Score
Peer 127.0.0.1:9002	DOC 1	DOC 1	1.0986132886686...
Peer 127.0.0.1:9002	DOC 2	DOC 2	0.4785563631972...

# Open Questions

- | QoS measures to improve peer selection?
- | Global computation of measures like *tf* or *idf*?
- | Enhancement of collaborative search by library-specific properties (e.g. topics)?
- | Estimation of benefit/cost ratio when deciding to contact specific peers?
- | Controlling the dynamics of Peer-to-Peer Systems?

?

?

?

?

?



# Conclusions & Ongoing Work

---

- | Peer-to-Peer approach for collaborative search across a large number of digital libraries
- | Scalable Search engine combining local index structures of autonomous peers with a distributed global directory
- | Extendable system architecture to combine peer-specific data and user-specific behavior
- | Detailed evaluation scenarios needed with a high number of collaborating libraries for testing different query routing and execution strategies

# Questions & Discussion

---

**Thank you for  
your attention!**

**Any questions?**

