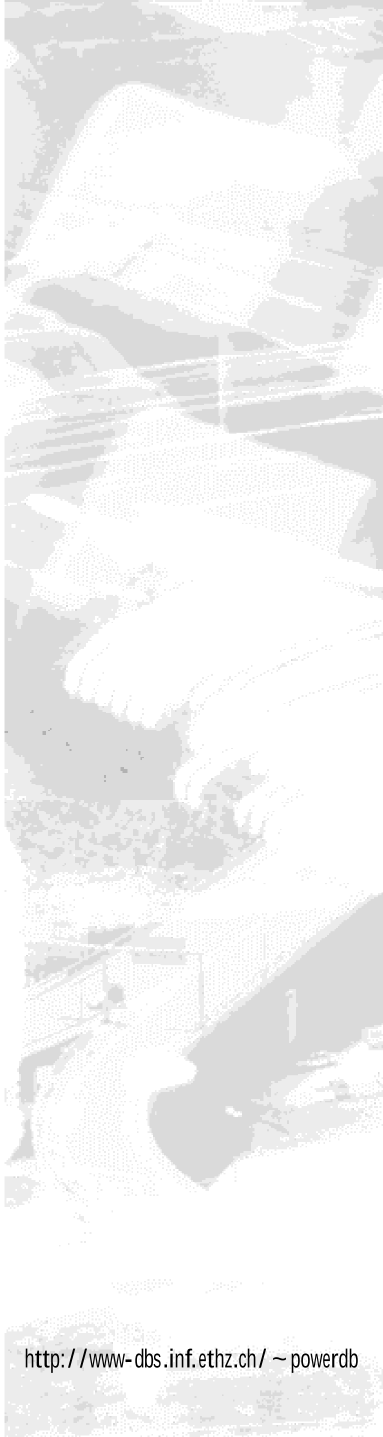


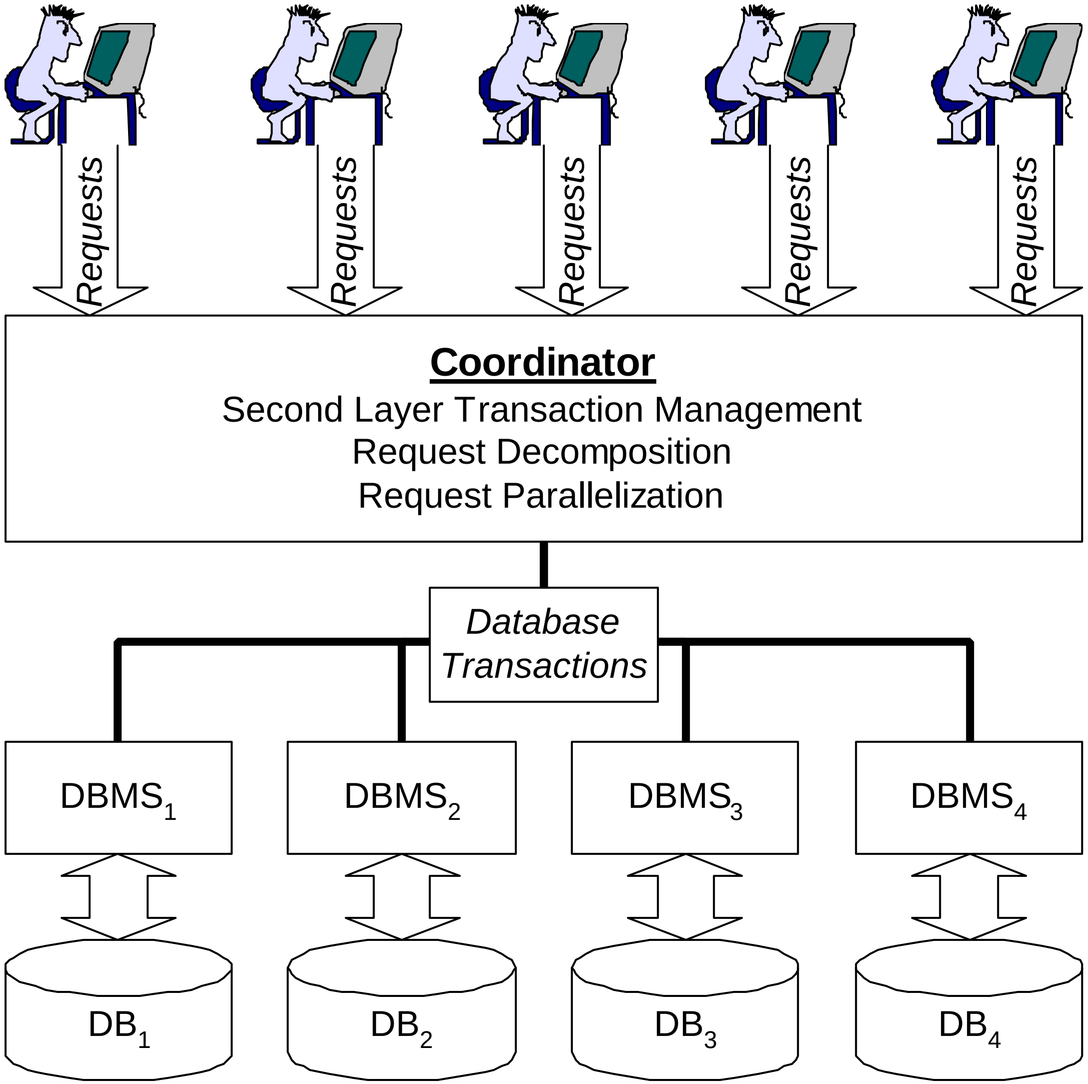
http://www-dbs.inf.ethz.ch/~powerdb

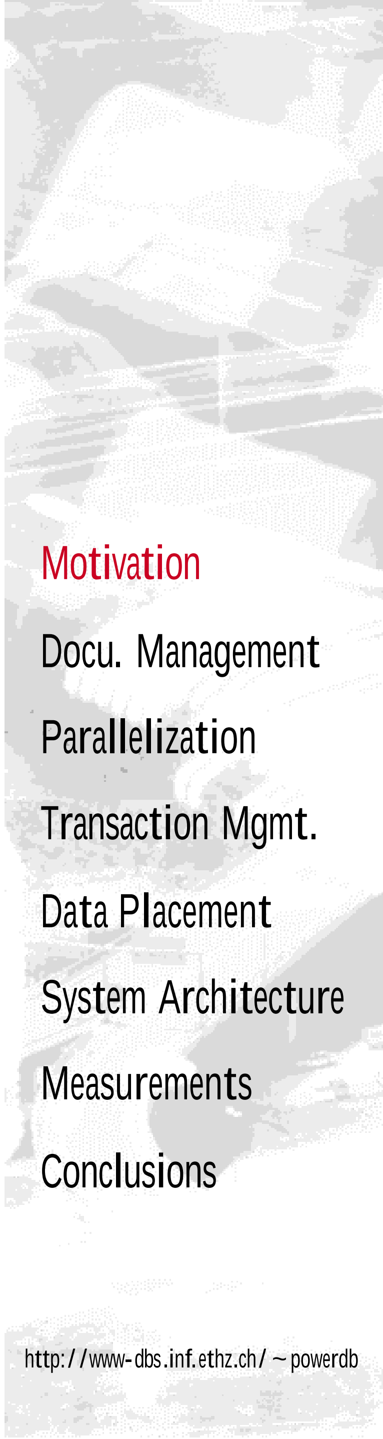
PowerDB

A Document Engine on a DB Cluster

Torsten Grabs, Klemens Böhm, Hans-Jörg Schek
Institute of Information Systems
Swiss Federal Institute of Technology, Zurich

Our Vision


Requests
Requests
Requests
Requests
Requests
Coordinator
Second Layer Transaction Management
Request Decomposition
Request Parallelization
Database Transactions
$DBMS_1$
$DBMS_2$
$DBMS_3$
$DBMS_4$
DB_1
DB_2
DB_3
DB_4



Motivation

Docu. Management

Parallelization

Transaction Mgmt.

Data Placement

System Architecture

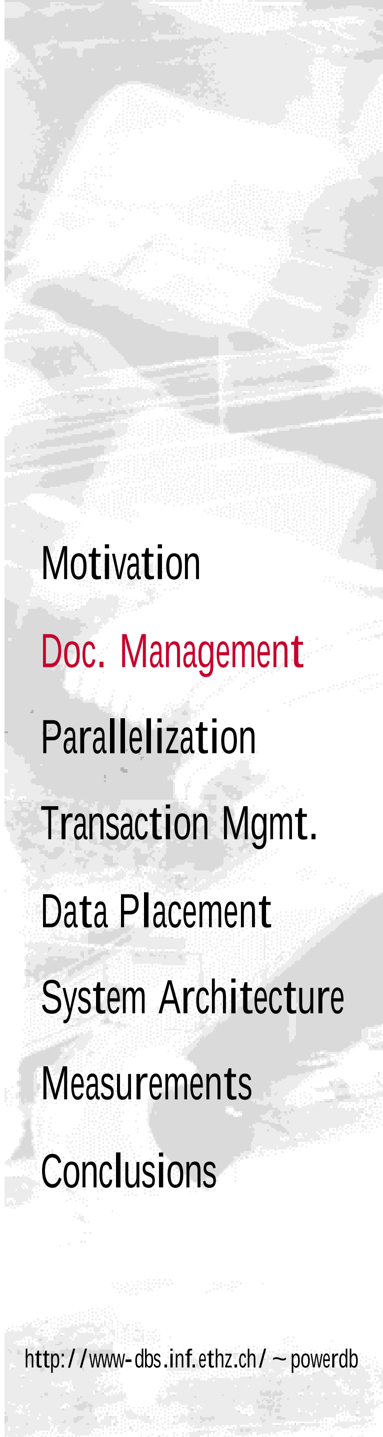
Measurements

Conclusions

Motivation

- Current Problems for Search Engines [Kirsch 1998]:
  - query response time
  - size of index
  - cost of hardware
  - freshness of the data in the index
- Documents on the Intra-Nets: Contracts as XML-Documents
  - immediate availability of new documents
- PowerDB technology:
  - high parallelism: reduce response times
  - commodity HW/SW approach: reduce cost
  - semantic transaction management: decrease update window

→ Case Study News Search Engine



Motivation

Doc. Management

Parallelization

Transaction Mgmt.

Data Placement

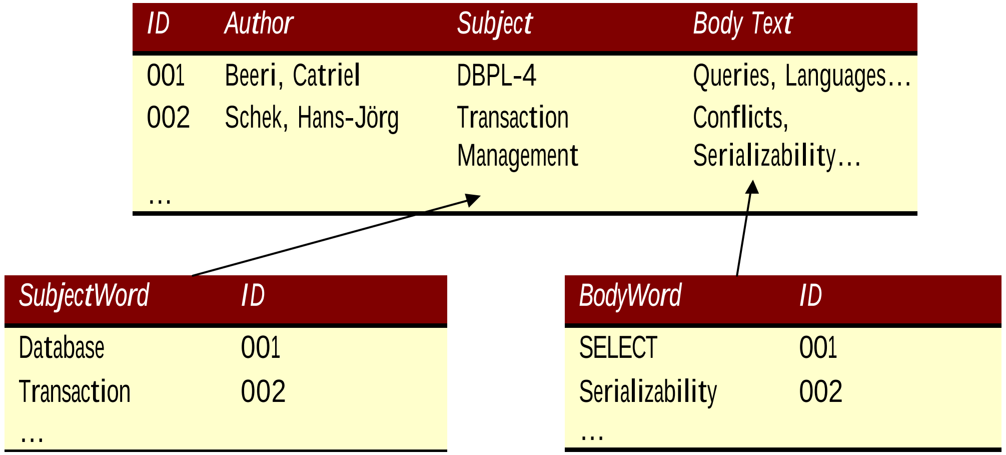
System Architecture

Measurements

Conclusions

Document Management

- Documents from discussion groups:
  - author
  - subject
  - news body text
- Services offered:
  - insertion of a new document
  - retrieval of documents that qualify for (some) words



ID	*Author*	*Subject*	*Body Text*
001	Beeri, Catriel	DBPL-4	Queries, Languages…
002	Schek, Hans-Jörg	Transaction Management	Conflicts, Serializability…
…			

SubjectWord	*ID*
Database	001
Transaction	002
…	

BodyWord	*ID*
SELECT	001
Serializability	002
…	

Motivation

Doc. Management

Parallelization

Transaction Mgmt.

Data Placement

System Architecture

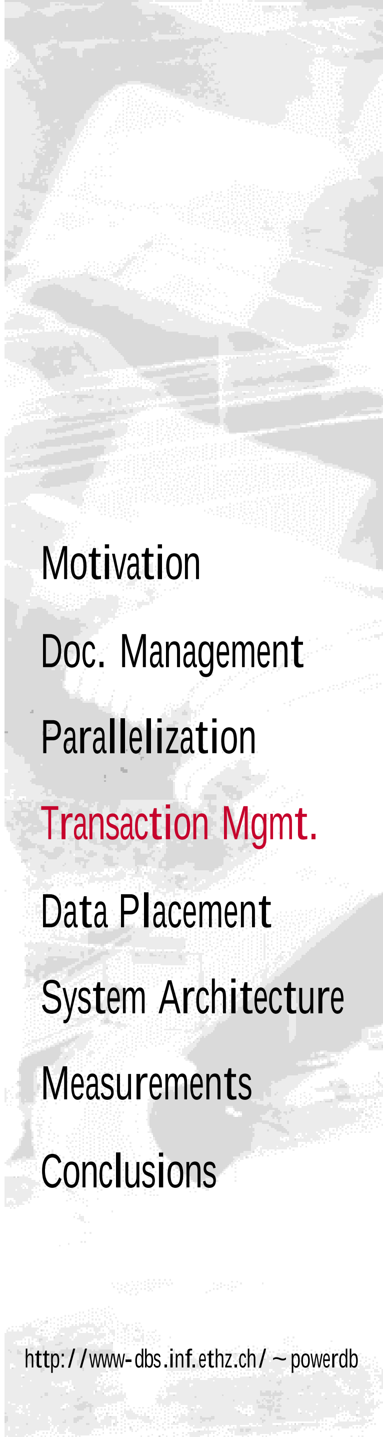
Measurements

Conclusions

Parallelization

```
BOT
Insert Into DOCS
   Values(…)



Insert Into SUBJIDX
   Values(…)



Insert Into BODYIDX
   Values(…)
EOT
```

```
BOT
Insert Into DOCS
   Values(…)
EOT
BOT
Insert Into SUBJIDX
   Values(…)
EOT
BOT
Insert Into BODYIDX
   Values(…)
EOT
```



Motivation

Doc. Management

Parallelization

Transaction Mgmt.

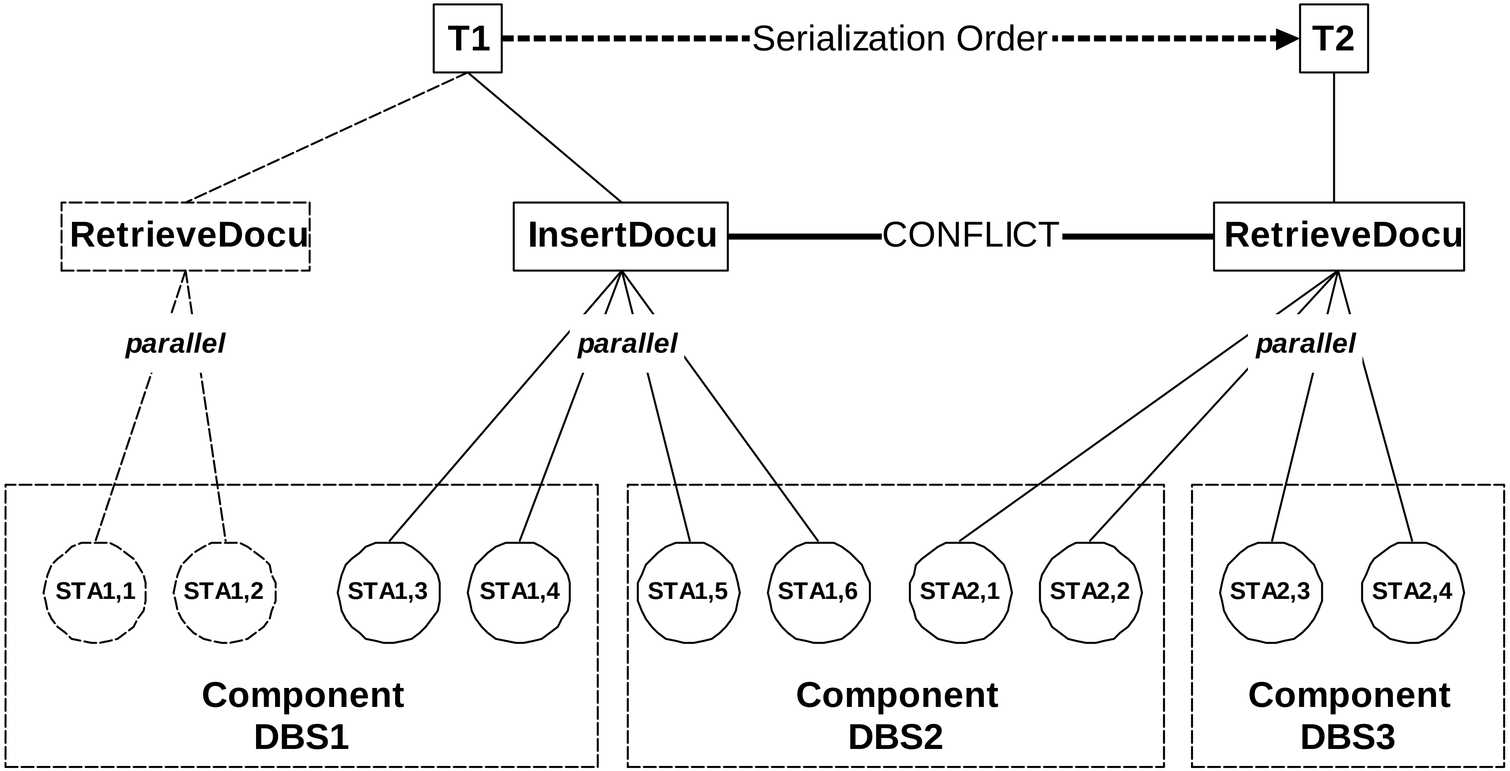
Data Placement

System Architecture

Measurements

Conclusions

Transaction Management



Motivation

Doc. Management

Parallelization

Transaction Mgmt.

Data Placement

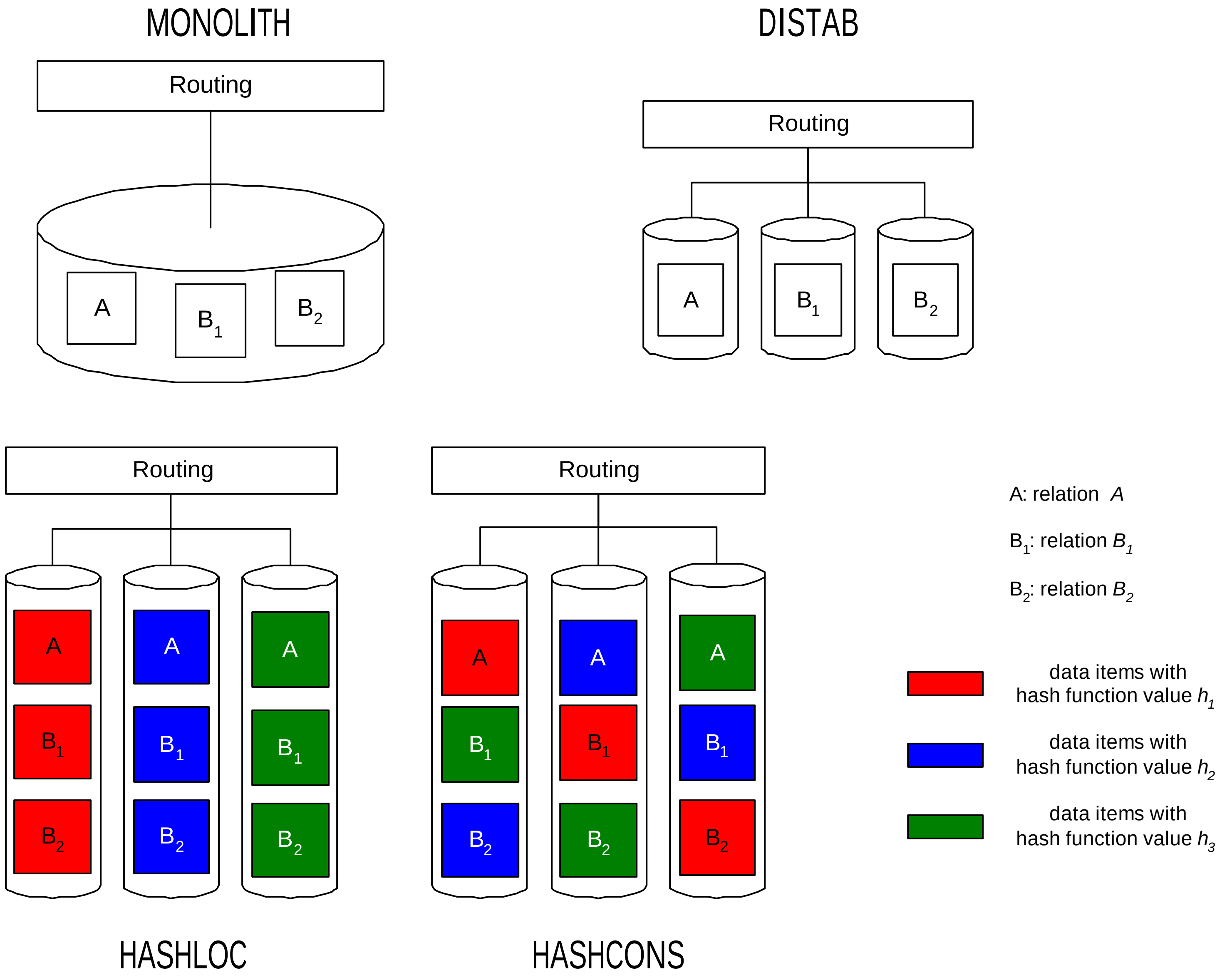
System Architecture

Measurements

Conclusions

http://www-dbs.inf.ethz.ch/~powerdb

Data Placement



Motivation

Doc. Management

Parallelization

Transaction Mgmt.

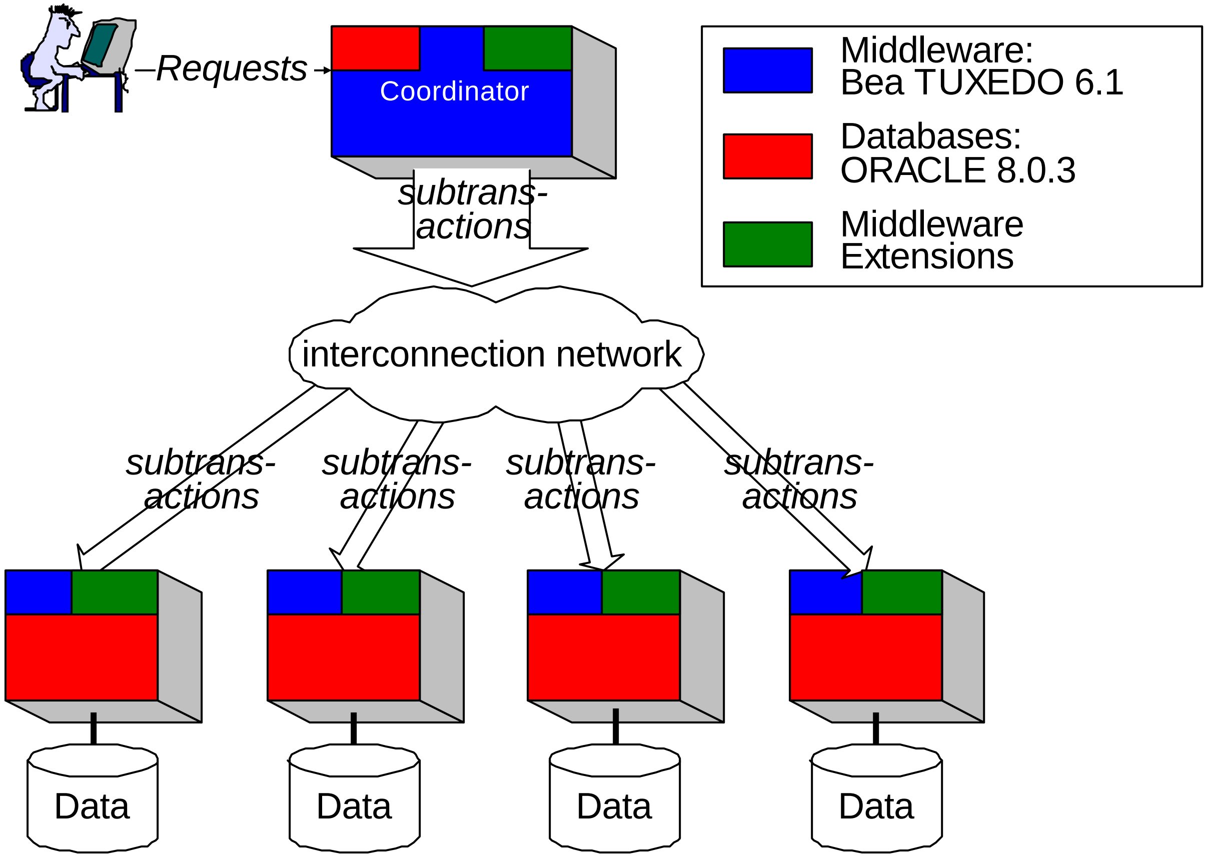
Data Placement

System Architecture

Measurements

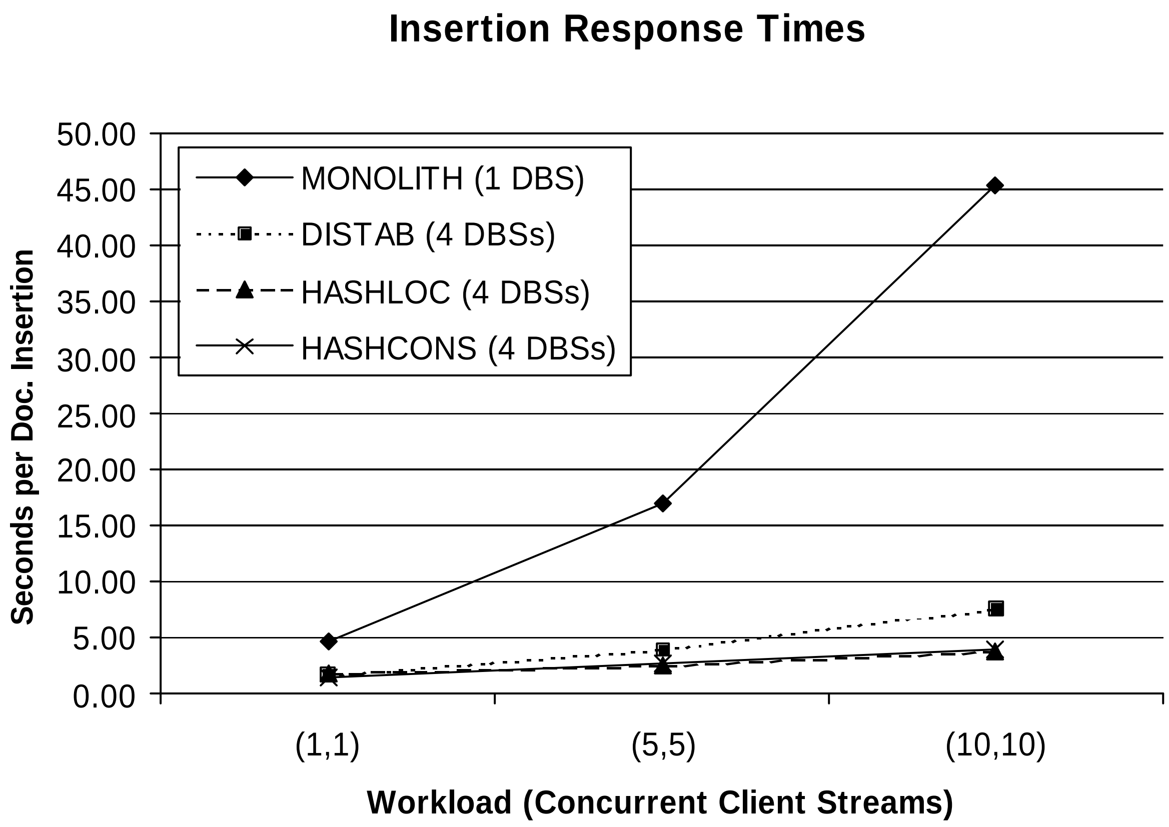
Conclusions

System Architecture



Measurements

From 1 component DBS to 4 component DBSs



- Retrieval response times behave alike!

Motivation

Doc. Management

Parallelization

Transaction Mgmt.

Data Placement

System Architecture

Measurements

Conclusions

http://www-dbs.inf.ethz.ch/~powerdb

Motivation

Doc. Management

Parallelization

Transaction Mgmt.

Data Placement

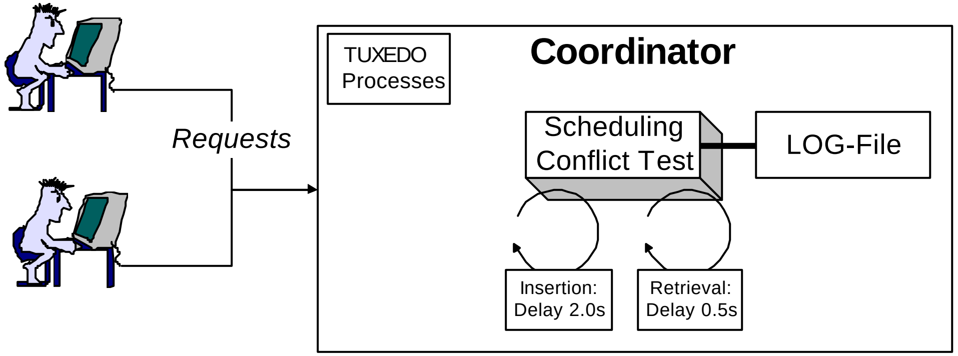
System Architecture

Measurements

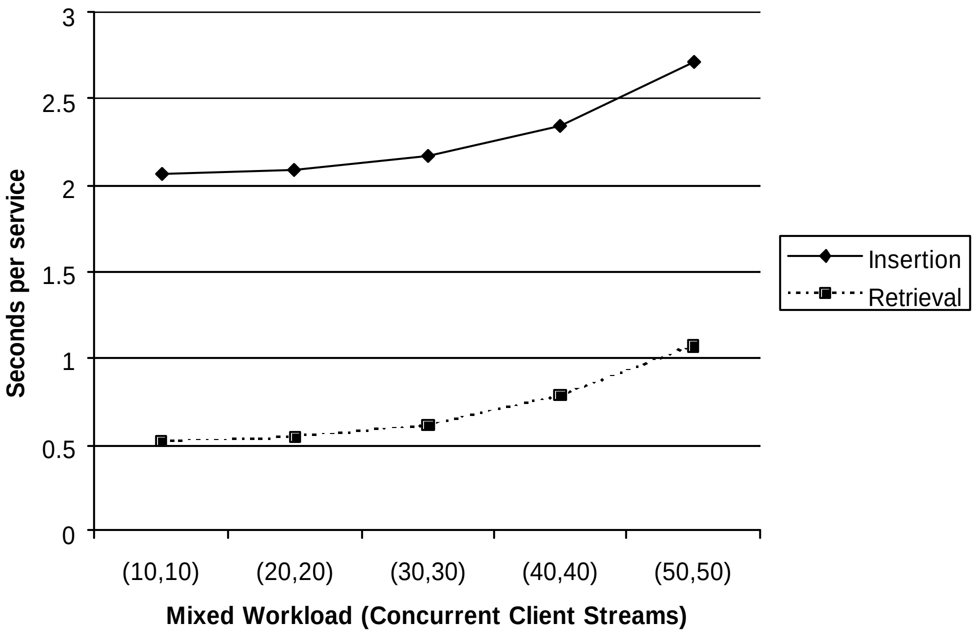
Conclusions

http://www-dbs.inf.ethz.ch/~powerdb

Measurements: Coordinator Scalability



Coordinator Response Times - Rudimentary Setup



Motivation

Doc. Management

Parallelization

Transaction Mgmt.

Data Placement

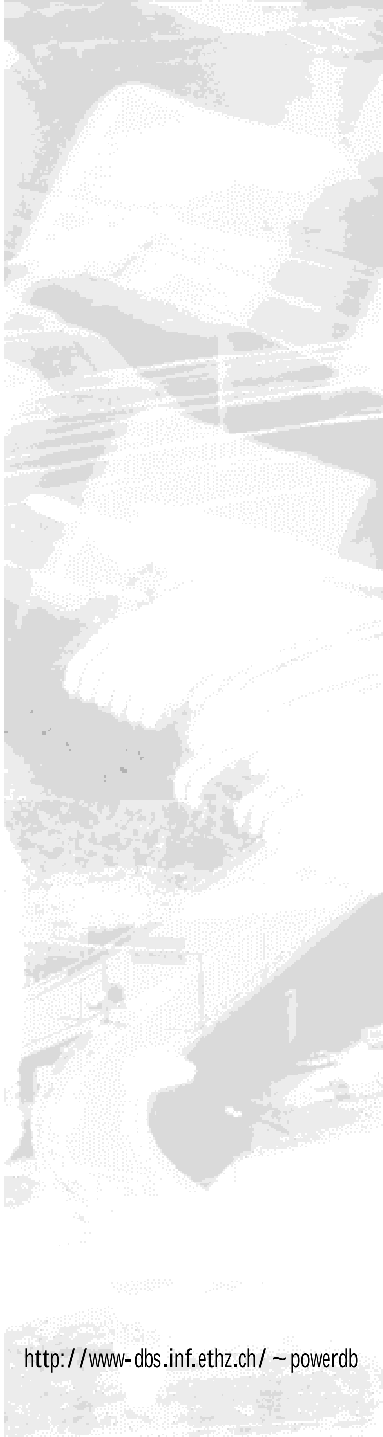
System Architecture

Measurements

Conclusions

Conclusions

- Document management with PC clusters is a good idea
  - speed-up by an order of magnitude from 1 to 4 component DBSs
- Problems solved:
  - response time speed-up
  - hardware cost
  - freshness of the data in the index
- Architectural propositions:
  - HASHLOC or HASHCONS have provided best response time results
- Concurrency control and logging at the coordinator is not a bottleneck



Bibliography

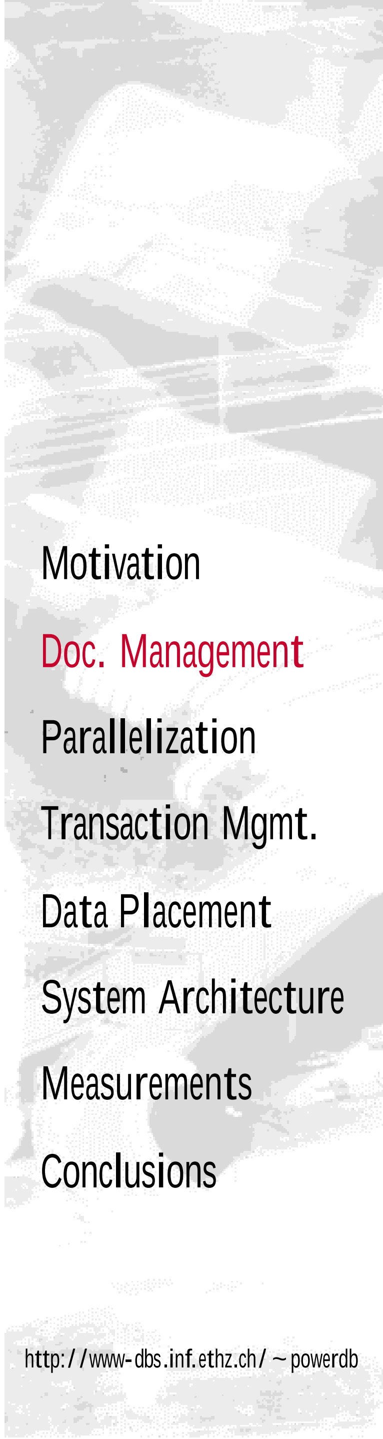
J. Gray: Super-Servers - Commodity Computer Clusters Pose a Software Challenge, BTW 1995.

Inktomi Corp.: The Inktomi Technology behind HotBot. Technical Report of Inktomi Corp., 1996.

M. Kamath and K. Ramamritham: Efficient Transaction Support for Dynamic Information Retrieval Systems, SIGIR 1996.

H. Kaufmann and H.-J. Schek: Extending TP-Monitors for Intra-Transaction Parallelism, PDIS 1996.

S. Kirsch: Infoseek’s Experiences Searching the Internet. SIGIR Forum 32(2), 1998.



Motivation

Doc. Management

Parallelization

Transaction Mgmt.

Data Placement

System Architecture

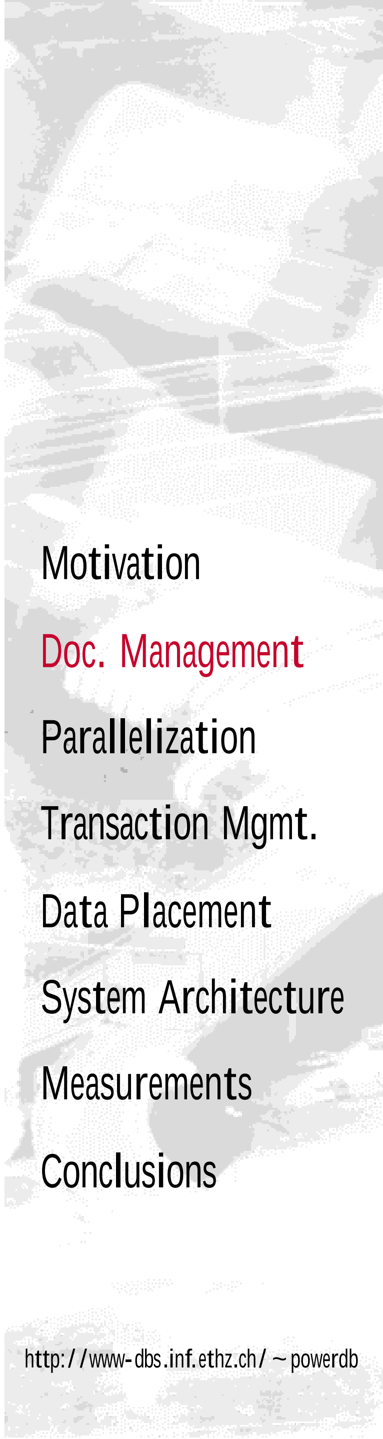
Measurements

Conclusions

RDBMS Mapping (I)

- Insertion service:
  integer InsertDocu(text)
- Single DBMS SQL transaction for insertion service

```
BOT
...
Insert into DOCS values (id1, author1, text1);
Insert into INDEX1 values (id1, t1);
Insert into INDEX2 values (id1, t2);
...
EOT
```



Motivation

Doc. Management

Parallelization

Transaction Mgmt.

Data Placement

System Architecture

Measurements

Conclusions

RDBMS Mapping (II)

- Retrieval service:
  {doctitles} RetrieveDocu({(field, term)})

- Single DBMS SQL transaction for retrieval service

```
BOT
...
Select * from DOCS where
(Select * from INDEX1 where term = t1)
Intersect
(Select * from INDEX1 where term = t2);
...
EOT
```

Transaction Management: Implementation

- Service invocation represented by bitstring signature
- Bit is set iff a term occurs in doctext or query, resp.
- Conflicts only between insertion and retrieval (and vice-versa)
- Efficient signature checking to detect conflicts:

  $$CON \Rightarrow sig_{InsertDocu} \wedge sig_{RetrieveDocu} = sig_{RetrieveDocu}$$

- „False drops “possible
- Two-phase locking protocol for signatures: not database pages locked but signatures
- Conflicts lead to sequential service execution

Motivation

Doc. Management

Parallelization

Transaction Mgmt.

Data Placement

System Architecture

Measurements

Conclusions

Transaction Management: Implementation

Motivation

Doc. Management

Parallelization

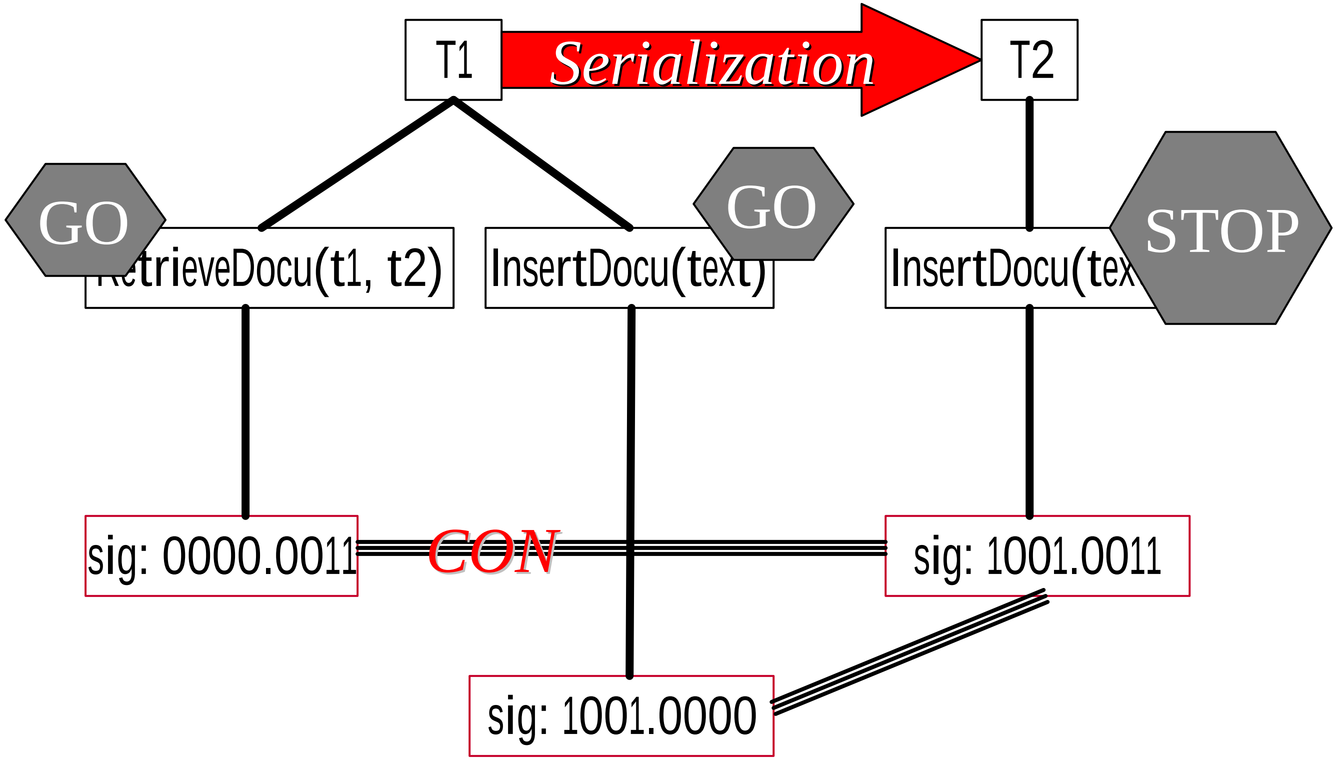
Transaction Mgmt.

Data Placement

System Architecture

Measurements

Conclusions



Motivation

Doc. Management

Parallelization

Transaction Mgmt.

Data Placement

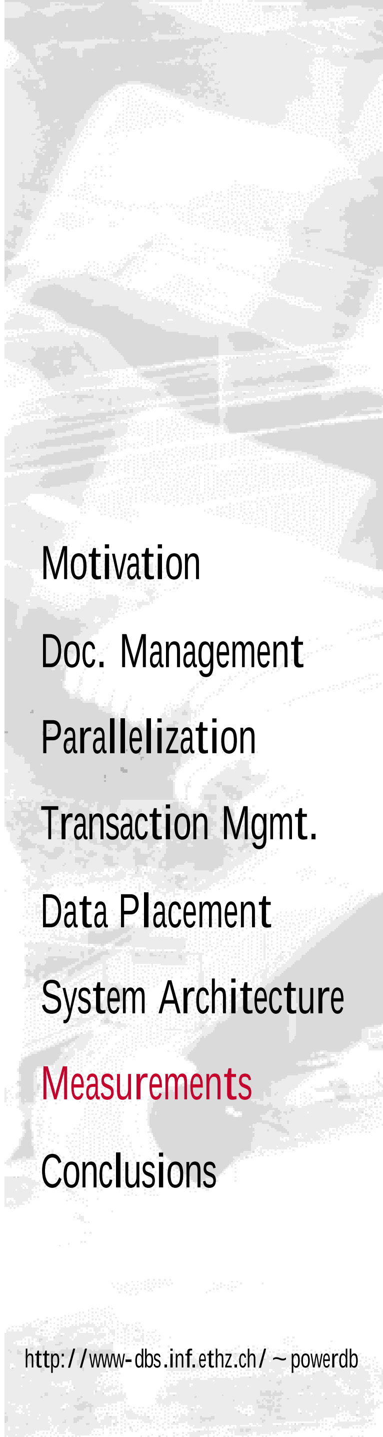
System Architecture

Measurements

Conclusions

Prototype System Components

- Software
  - „Cluster System Glue “-- Bea Sys. TUXEDO
    - remote service invocation
    - buffered data transmission
    - no XOpen/XA features used
  - Database Servers -- ORACLE 8.0.3
  - Windows NT 4.0 Server
  - Proprietary imlementations:
    - database mapping
    - service decomposition and parallelization
    - semantic transaction management
- Hardware
  - 266 MHz Pentium PCs
  - 2 Disks: IDE, SCSI



Motivation

Doc. Management

Parallelization

Transaction Mgmt.

Data Placement

System Architecture

Measurements

Conclusions

Measurements

- Prototype System: dimensions of the measurements
  - data placement
  - number of component databases: 1, 2, 4 component DBSs
  - workload: multi-programming degree (#retriever, #inserter)
    - namely: (1,1) - (5,5) - (10,10)
- Simulation Studies: bottleneck tests for coordinator
  - coordinator is centralized
  - scalability tests in different setups with up to (50,50) clients
    - Full Setup: as discussed previously
    - Reduced Setup: application spec. operations at coordinator
    - Rudimentary Setup
      - DB components switched off
      - but: no application specific operations at coordinator

Motivation

Doc. Management

Parallelization

Transaction Mgmt.

Data Placement

System Architecture

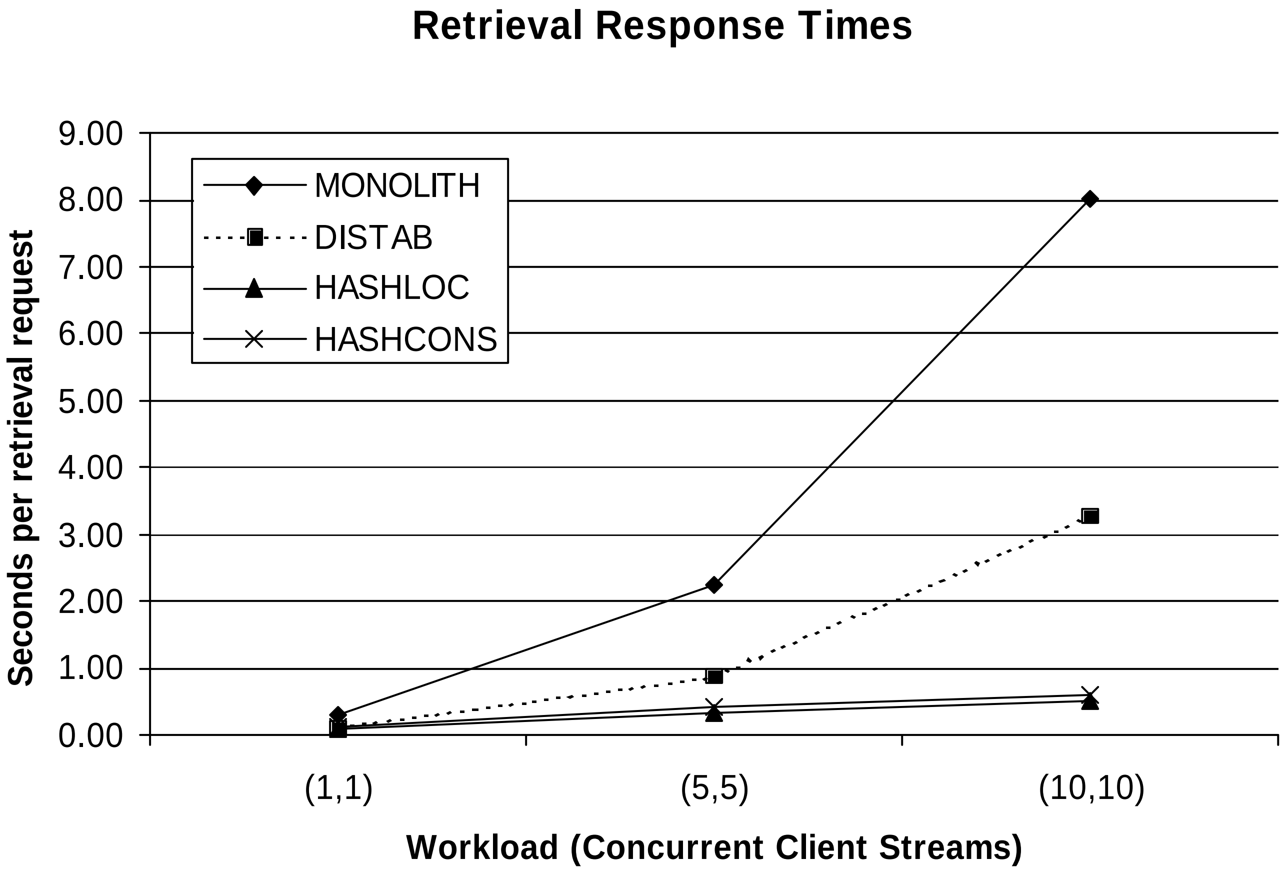
Measurements

Conclusions

http://www-dbs.inf.ethz.ch/~powerdb

Measurements

From 1 component DBS to 4 component DBSs



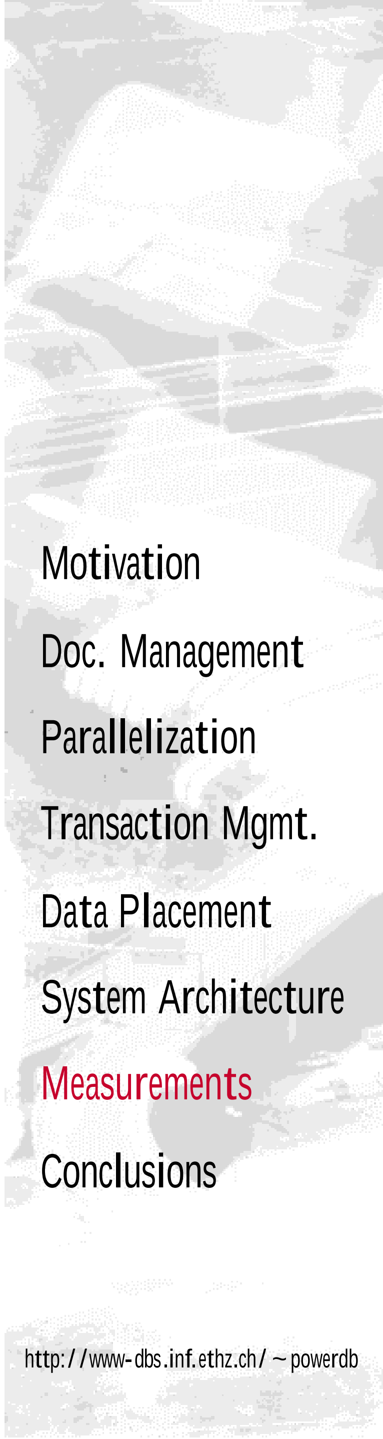
Measurements

Full configuration: insertion speed-up from 1 to 4 component DBSs

Placement / Workload	DISTAB	HASHLOC	HASHCONS
(1,1)	2.8	2.5	3.1
(5,5)	4.2	6.3	6.1
(10,10)	6	12	11.2

Full configuration: retrieval speed-up from 1 to 4 component DBSs

Placement / Workload	DISTAB	HASHLOC	HASHCONS
(1,1)	2.7	2.7	2.5
(5,5)	2.5	6.4	5.3
(10,10)	2.5	15	13



Motivation

Doc. Management

Parallelization

Transaction Mgmt.

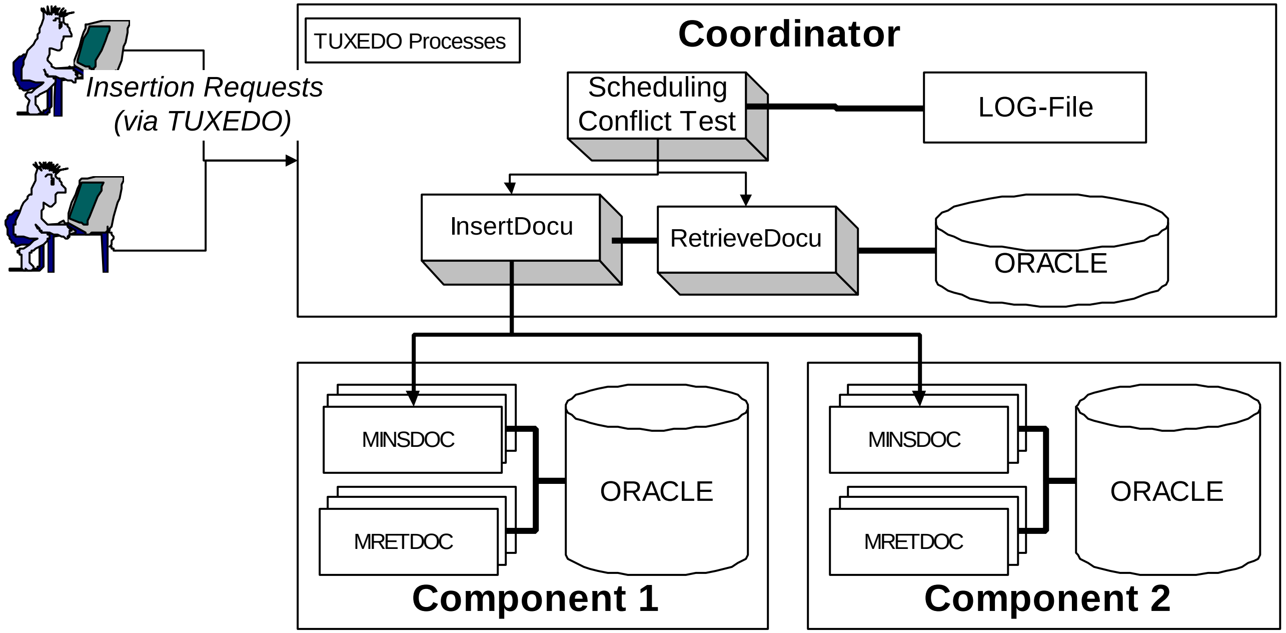
Data Placement

System Architecture

Measurements

Conclusions

Full Setup



Motivation

Doc. Management

Parallelization

Transaction Mgmt.

Data Placement

System Architecture

Measurements

Conclusions

Rudimentary Setup: Results

Motivation

Doc. Management

Parallelization

Transaction Mgmt.

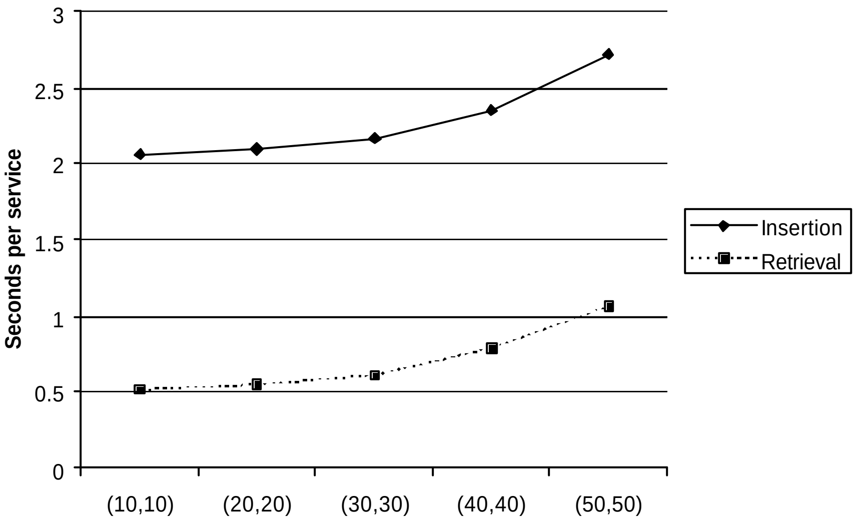
Data Placement

System Architecture

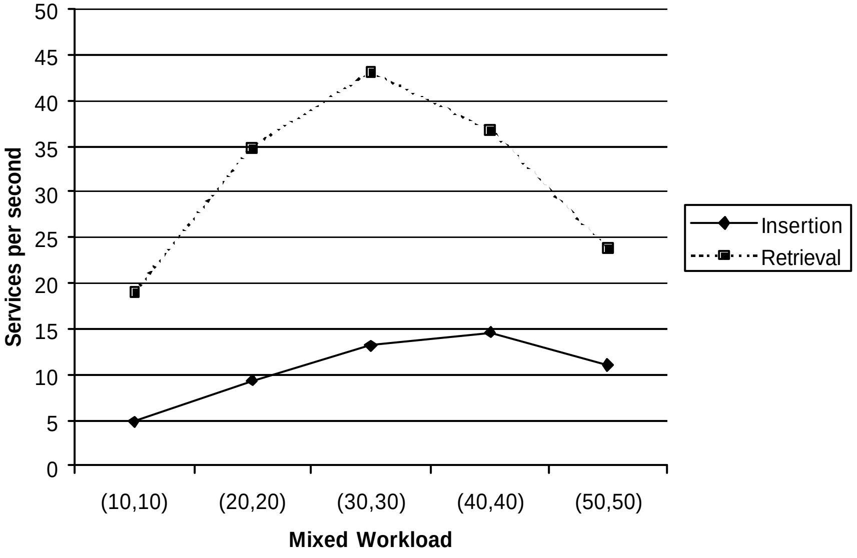
Measurements

Conclusions

Coordinator Response Times - Rudimentary Setup



Coordinator Throughput - Rudimentary Setup


Services per second
(10,10)
(20,20)
(30,30)
(40,40)
(50,50)
Mixed Workload
Insertion
Retrieval