Module 3
Distributed Databases

Dr. Geetha Mary A
Associate Professor ,
SCOPE, Vellore Institute of Technology,
Vellore

Source:
Mc Graw Hill, Inc. 2017, Third Edition

RamaKrishnan and Gehrke, Database Management Systems,

Distributed Databases

I. Introduction to DDBMS
II. Architecture of DDBs
III. Storing data in DDBs
IV. Distributed catalog management
V. Distributed query processing
VI. Transaction Processing
VII. Distributed concurrency control and recovery

Dr. Geetha Mary A, Asso. Prof., SCOPE, VIT, Vellore

I.Introduction to DDBMS

- Data in a distributed database system is stored across several sites.
- Each site is typically managed by a DBMS that can run independently of the other sites that co-operate in a transparent manner.
  - ***Transparent*** implies that each user within the system may access all of the data within all of the databases as if they were a single database
- There should be 'location independence' i.e.- as the user is unaware of where the data is located it is possible to move the data from one physical location to another without affecting the user.

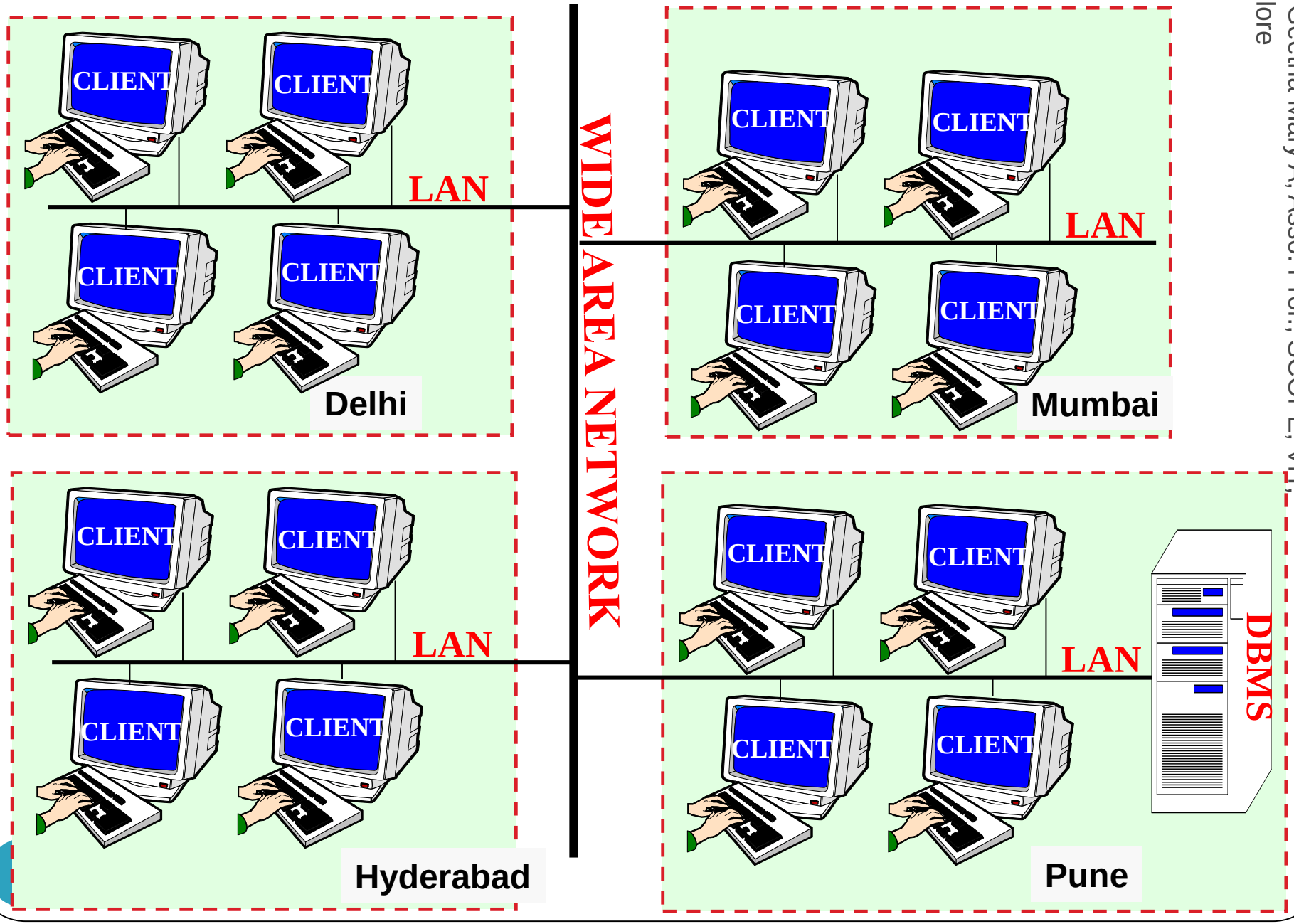
Dr. Geetha Mary A, Asso. Prof., SCOPE, VIT, Vellore

DDBMS properties

- **Distributed data independence**: Users should be able to ask queries without specifying where the referenced relations, or copies or fragments of the relations, are located.

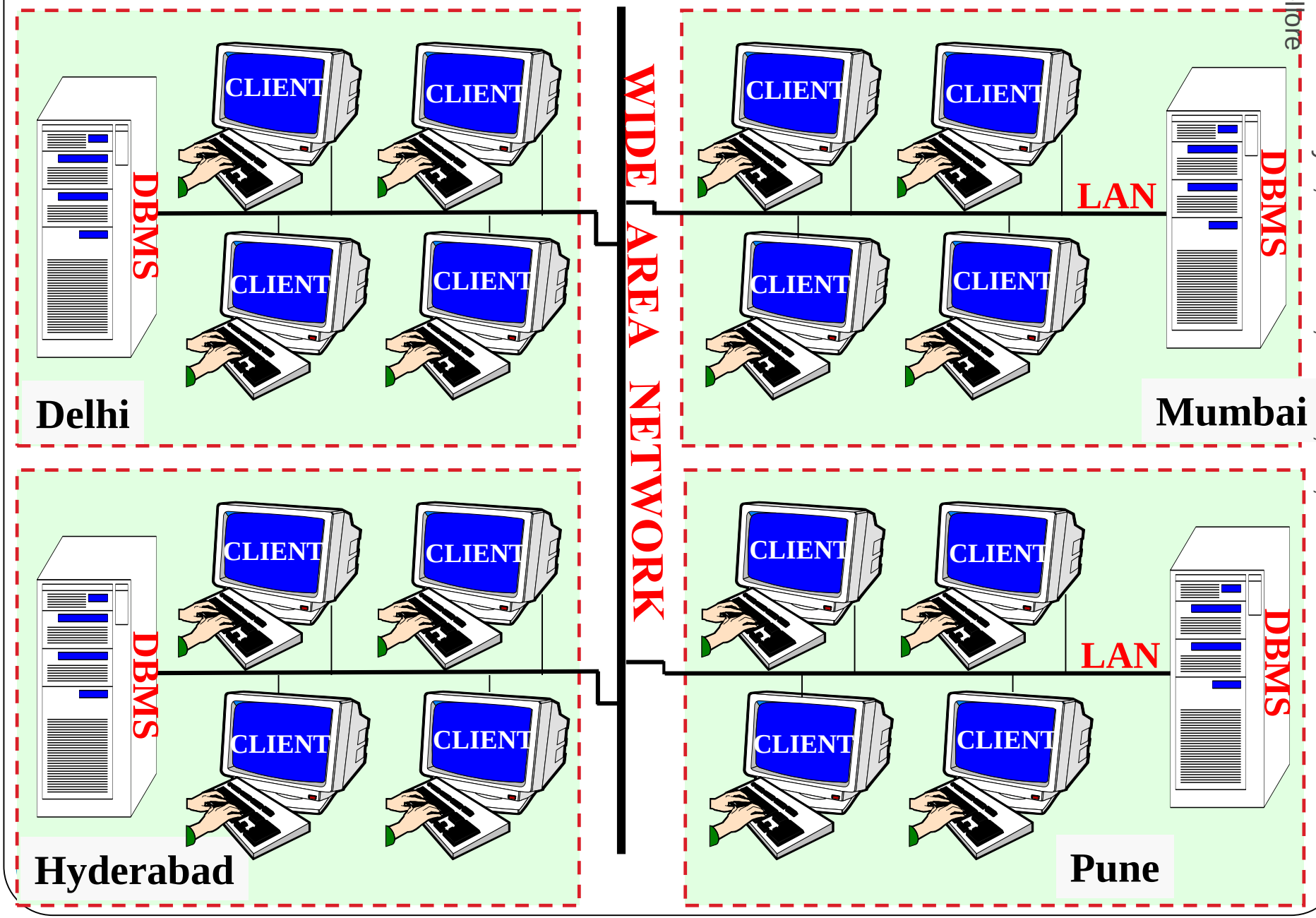
- **Distributed transaction atomicity:** Users should be able to write transactions that access and update data at several sites just as they would write transactions over purely local data.

DISTRIBUTED PROCESSING ARCHITECTURE



Dr. Geetha Mary A, Asso. Prof., SCOPE, VIT, Vellore

DISTRIBUTED DATABASE ARCHITECTURE



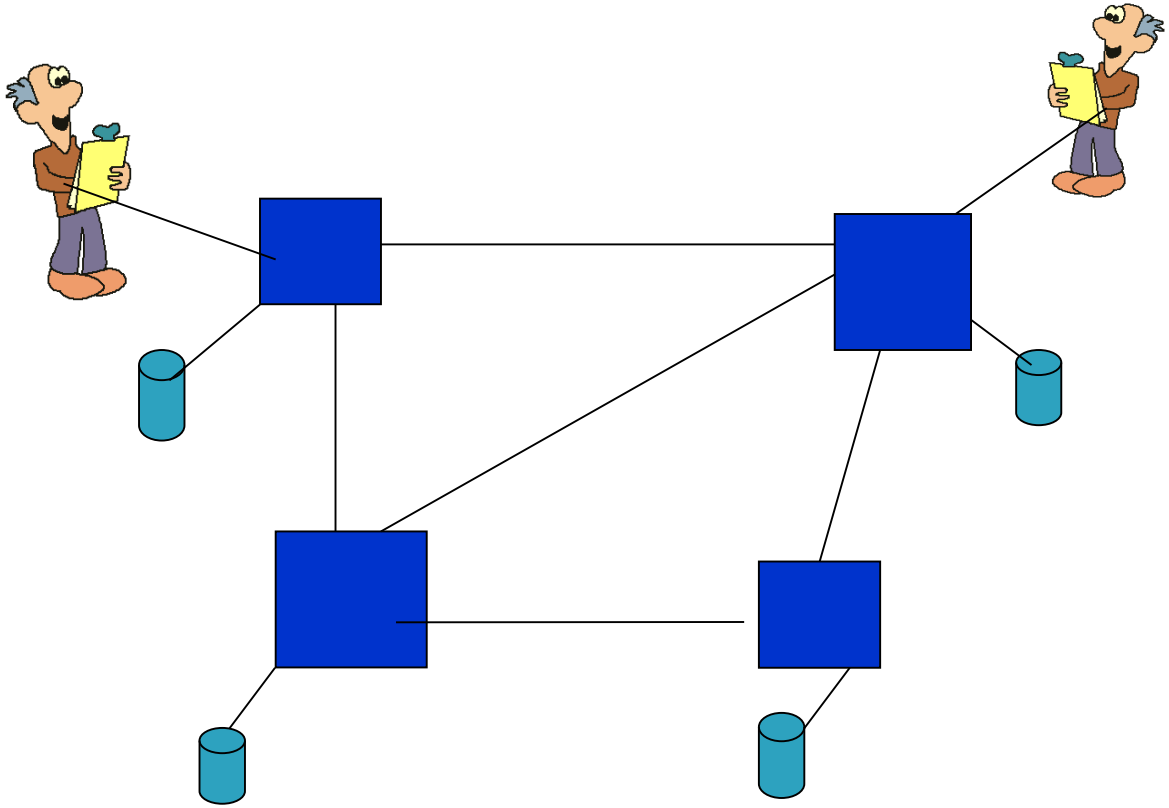
Dr. Geetha Mary A, Asso. Prof., SCOPE, VIT, Vellore

Distributed database

- Communication Network- DBMS and Data at each node

•Users are unaware of the distribution of the data

= Location transparency



Types of Distributed Databases

- **Homogeneous distributed database system** :
  - If data is distributed but all servers run the same DBMS software.

- **Heterogeneous distributed database** :
  - If different sites run under the control of different DBMSs, essentially autonomously, are connected to enable access to data from multiple sites.

- The key to building heterogeneous systems is to have well-accepted standards for gateway protocols.
- **A gateway protocol** is an API that exposes DBMS functionality to external applications.

Dr. Geetha Mary A, Asso. Prof., SCOPE, VIT, Vellore

I. Introduction to DDBMS
II. Architecture of DDBs
III. Storing data in DDBs
IV. Distributed catalog management
V. Distributed query processing
VI. Transaction Processing
VII. Distributed concurrency control and recovery

2.DISTRIBUTED DBMS ARCHITECTURES

1. Client-Server Systems
2. Collaborating Server Systems
3. Middleware Systems

1.Client-Server Systems:

1. A Client-Server system has one or more client processes and one or more server processes,
2. A client process can send a query to any one server process.
3. Clients are responsible for user-interface issues,
4. Servers manage data and execute transactions.
5. A client process could run on a personal computer and send queries to a server running on a mainframe.
6. The Client-Server architecture does not allow a single query to span multiple servers

Dr. Geetha Mary A, Asso. Prof., SCOPE, VIT, Vellore

TERMINALS

DUMB

DUMB

DUMB

SPECIALISED NETWORK CONNECTION

MAINFRAME COMPUTER

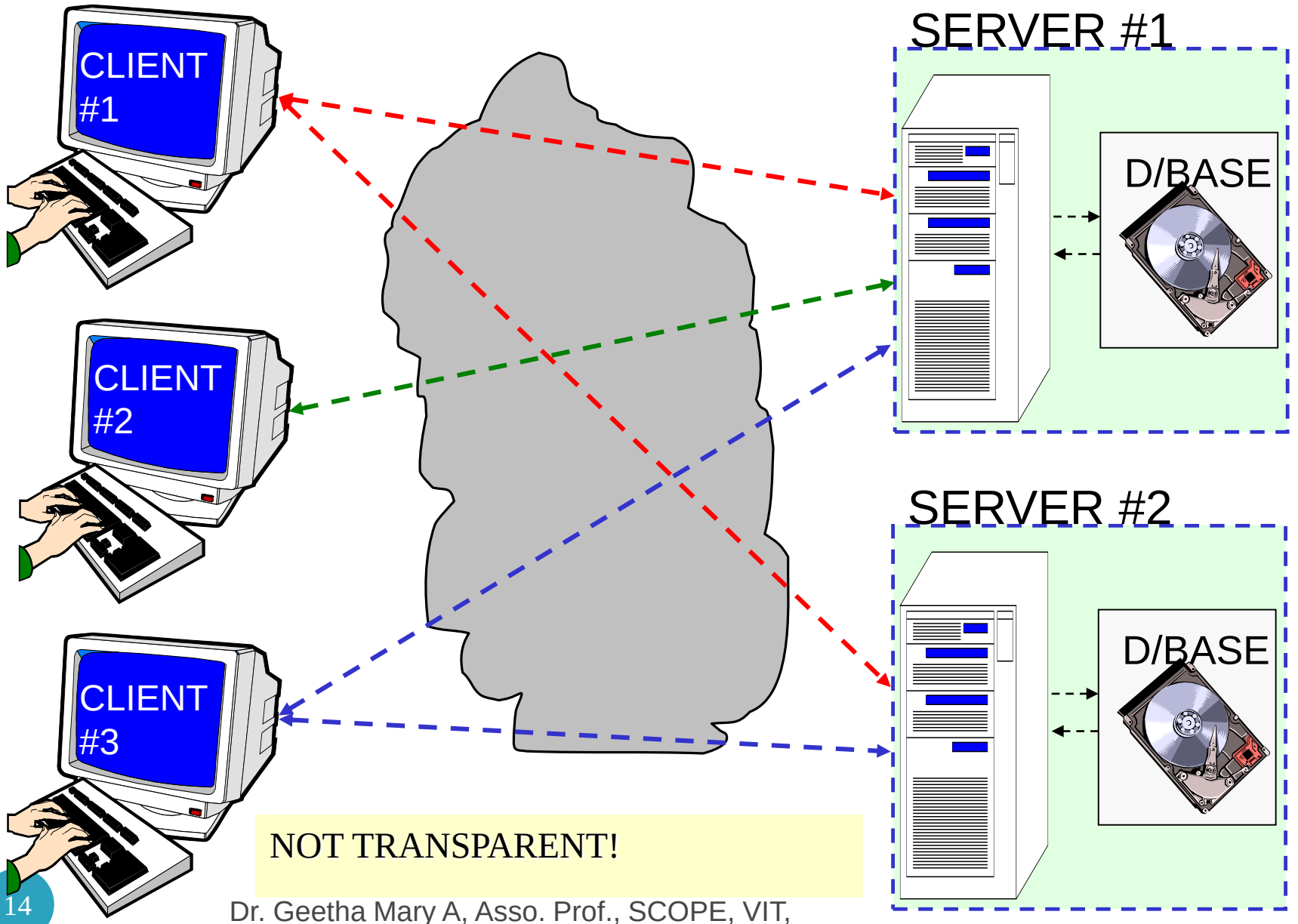
PRESENTATION LOGIC
BUSINESS LOGIC
DATA LOGIC

2.Collaborating Server Systems

1. The client process would have to be capable of breaking such a query into appropriate subqueries.
2. **A Collaborating Server** system can have a collection of database servers, each capable of running transactions against local data, which cooperatively execute transactions spanning multiple servers.
3. When a server receives a query that requires access to data at other servers, it generates appropriate subqueries to be executed by other servers .
4. puts the results together to compute answers to the original query.

Dr. Geetha Mary A, Asso. Prof., SCOPE, VIT, Vellore

M:N CLIENT/SERVER DBMS ARCHITECTURE



NOT TRANSPARENT!

3.Middleware Systems:

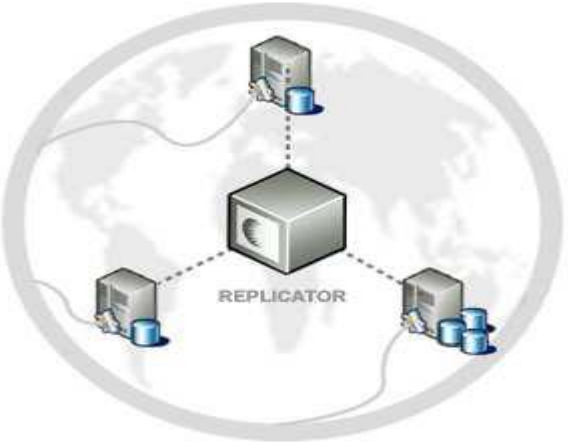
- The Middleware architecture is designed to allow a single query to span multiple servers, without requiring all database servers to be capable of managing such multisite execution strategies.

- It is especially attractive when trying to integrate several legacy systems, whose basic capabilities cannot be extended.

I. Introduction to DDBMS
II. Architecture of DDBs
III. Storing data in DDBs
IV. Distributed catalog management
V. Distributed query processing
VI. Transaction Processing
VII. Distributed concurrency control and recovery

3.Storing Data in DDBs

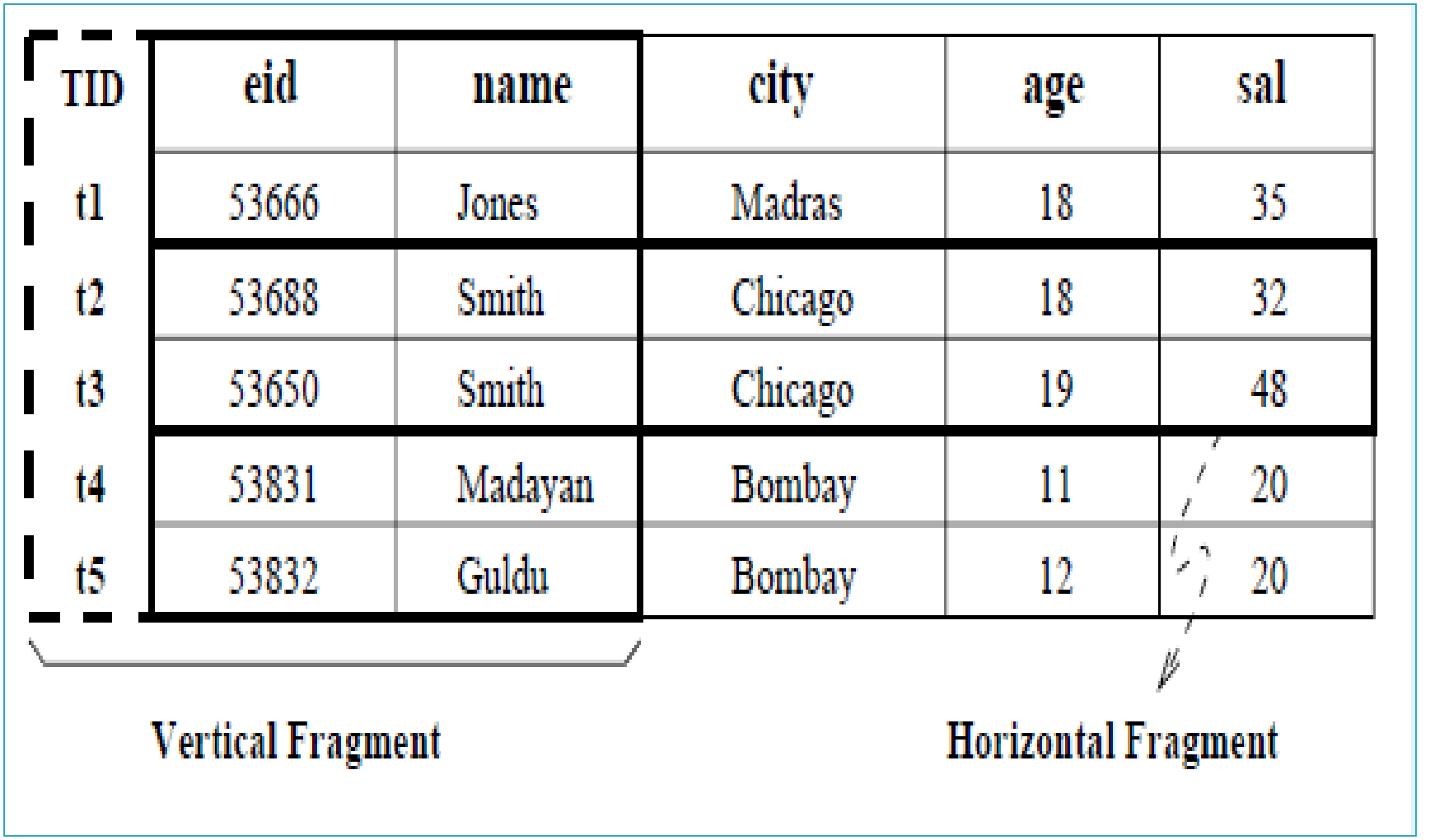
- ✓ In a distributed DBMS, relations are stored across several sites.
- ✓ Accessing a relation that is stored at a remote site includes message-passing costs.
- ✓ A single relation may be *partitioned* or *fragmented* across several sites.



Types of Fragmentation:

- **Horizontal fragmentation:** The union of the horizontal fragments must be equal to the original relation. Fragments are usually also required to be disjoint.

- **Vertical fragmentation:** The collection of vertical fragments should be a lossless-join decomposition.


TID
eid
name
city
age
sal
t1
53666
Jones
Madras
18
35
t2
53688
Smith
Chicago
18
32
t3
53650
Smith
Chicago
19
48
t4
53831
Madayan
Bombay
11
20
t5
53832
Guldu
Bombay
12
20
Vertical Fragment
Horizontal Fragment

Replication

- **Replication** means that we store several copies of a relation or relation fragment.

- The motivation for replication is two fold:
  1. **Increased availability of data:**
  2. **Faster query evaluation:**

- Two kinds of replications
  1. **synchronous replication**
  2. **asynchronous replication**

I. Introduction to DDBMS
II. Architecture of DDBs
III. Storing data in DDBs
IV. Distributed Catalog Management
V. Distributed Query Processing
VI. Transaction Processing
VII. Distributed concurrency control and recovery

4.Distributed Catalog Management

1. Naming Objects

- If a relation is fragmented and replicated, we must be able to uniquely identify each replica of each fragment.
  1. **A *local name field***
  2. **A *birth site field***

2. Catalog Structure

- A centralized system catalog can be used. It is vulnerable to failure of the site containing the catalog.
- An alternative is to maintain a copy of a global system catalog. compromises site autonomy.

4.Distributed Catalog Management

- **A better approach:**
- Each site maintains a local catalog that describes all copies of data stored at that site.

- In addition, the catalog at the birth site for a relation is responsible for keeping track of where replicas of the relation are stored.

I. Introduction to DDBMS
II. Architecture of DDBs
III. Storing data in DDBs
IV. Distributed Catalog management
V. Distributed Query Processing
VI. Transaction Processing
VII. Distributed concurrency control and recovery

Dr. Geetha Mary A, Asso. Prof., SCOPE, VIT, Vellore

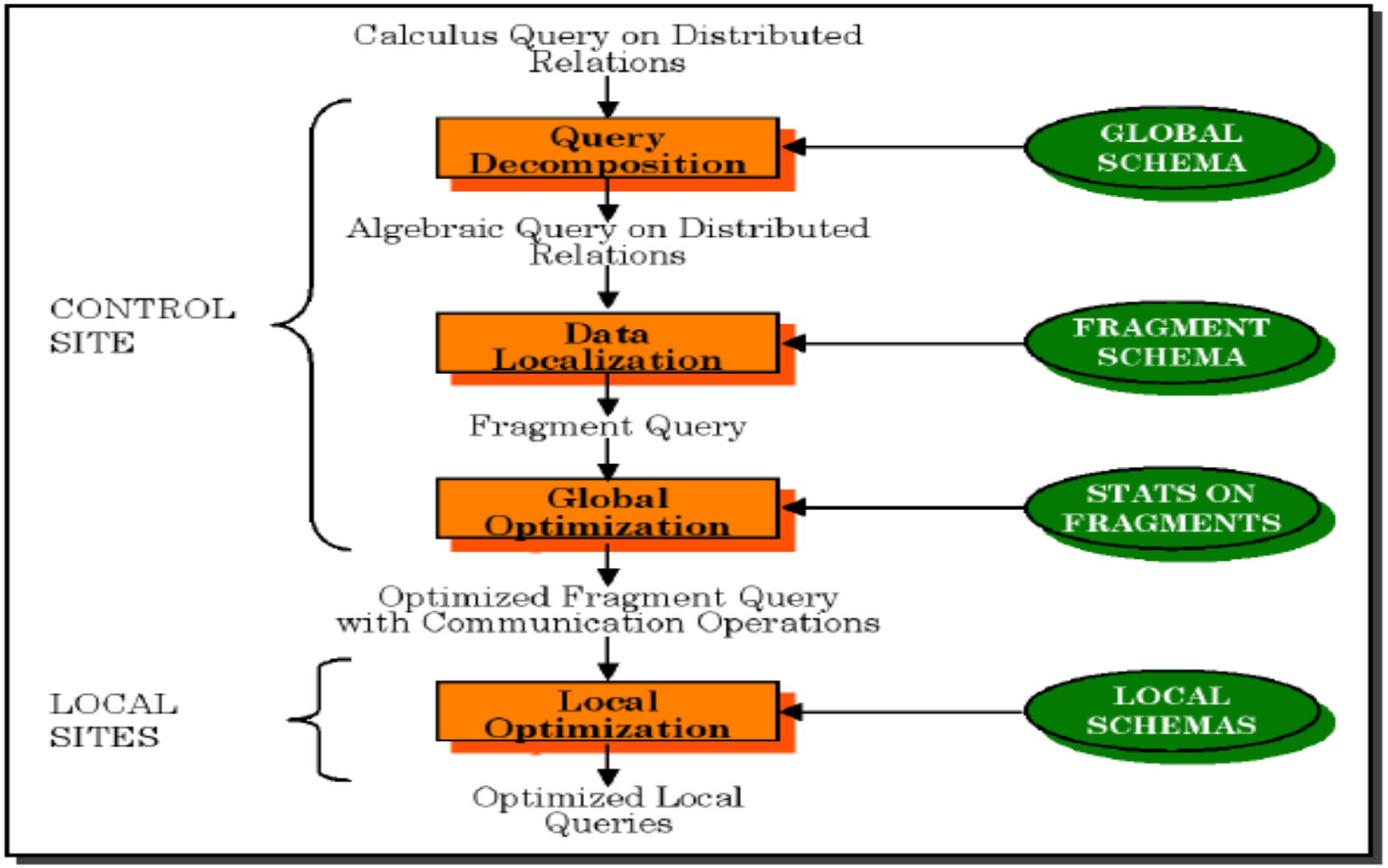
5.Distributed Query Processing

- **Distributed query processing:** Transform a high-level query (of relational calculus/SQL) on a distributed database (i.e., a set of global relations) into an equivalent and efficient lower-level query (of relational algebra) on relation fragments.

- Distributed query processing is more complex

1. **– Fragmentation/replication of relations**
2. **– Additional communication costs**
3. **– Parallel execution**

Distributed Query Processing Steps



5.Distributed Query Processing

- **Sailors(*sid: integer, sname: string, rating: integer, age: real)***
- **Reserves(*sid: integer, bid: integer, day: date, rname: string)***

- Assume Reserves and Sailors relations
  - each tuple of Reserves is 40 bytes long
    - a page can hold 100 Reserves tuples
    - 1,000 pages of such tuples.
  - each tuple of Sailors is 50 bytes long
    - a page can hold 80 Sailors Tuples
  - 500 pages of such tuples
- ***How to estimate the cost?***

5.Distributed Query Processing

- Criteria for measuring the cost of a query evaluation strategy
  - For centralized DBMSs number of disk accesses (# blocks read / written)
  - For distributed databases, additionally
    - The cost of data transmission over the network
    - Potential gain in performance from having several sites processing parts of the query in parallel

5.Distributed query processing

- To estimate the cost of an evaluation strategy, in addition to counting the number of page I/Os.
- we must count the number of pages that are shipped is a communication costs.
- Communication costs is a significant component of overall cost in a distributed database.

1. **Nonjoin Queries in a Distributed DBMS**
2. **Joins in a Distributed DBMS**
3. **Cost-Based Query Optimization**

Dr. Geetha Mary A, Asso. Prof., SCOPE, VIT, Vellore

5.Distributed Query Processing

1.Nonjoin Queries in a Distributed DBMS

- Even simple operations such as scanning a relation, selection, and projection are affected by fragmentation and replication.

```
SELECT S.age
FROM Sailors S
WHERE S.rating > 3 AND S.rating < 7
```

- Suppose that the Sailors relation is horizontally fragmented, with all tuples having a rating less than 5 at **Mumbai** and all tuples having a rating greater than 5 at **Delhi.**

The DBMS must answer this query by evaluating it at both sites and taking the union of the answers.

5.Distributed Query Processing

Eg 1:

```
SELECT avg(age)
FROM Sailors S
WHERE S.rating > 3 AND S.rating < 7
```

- ***taking the union of the answers is not enough***

Eg 2:

```
SELECT S.age
FROM Sailors S
WHERE S.rating > 6
```

- **Eg 3:** suppose that the Sailors relation is vertically fragmented, with the *sid* and *rating* fields at MUMBAI and the *sname* and *age* fields at DELHI
- This vertical fragmentation would be a lossy decomposition

5.Distributed Query Processing

Eg 4:the entire Sailors relation is stored at both MUMBAI and DELHI sites.

- ***Where should the query be executed?***

Joins in a Distributed DBMS

- **Eg:** the Sailors relation is stored at MUMBAI, and that the Reserves relation is stored at DELHI.
- Joins of relations at different sites can be very expensive.
- JOIN STRATEGY
  1. **Fetch as needed**
  2. **Ship whole**
  3. **Semijoins**
  4. **Bloomjoins**

Which strategy is better for me?

5.Distributed Query Processing

- **1.Fetch As Needed**
- **Page-oriented Nested Loops join:** For each *page* of R, get each *page* of S, and write out matching pairs of tuples <r, s>, where r is in R-page and s is in S-page.

- We could do a page-oriented nested loops join in **London** with Sailors as the outer, and for each Sailors page, fetch all Reserves pages from **Paris**.

- If we cache the fetched Reserves pages in **Paris** until the join is complete, pages are fetched only once

Dr. Geetha Mary A, Asso. Prof., SCOPE, VIT, Vellore

Fetch As Needed: Transferring the relation piecewise

R

A	B
3	7
1	1
4	6
7	7
4	5
6	2
5	7

S

B	C	D
9	8	8
1	5	1
9	4	2
4	3	3
4	2	6
5	7	8

R ⋈ S

A	B	C	D
1	1	5	1
4	5	7	8

QUERY: The query asks for R ⋈ S

5.Distributed Query Processing

- Assume Reserves and Sailors relations
  - **each tuple of Reserves is 40 bytes long**
  - **a page can hold 100 Reserves tuples**
  - **1,000 pages of such tuples.**
  - each tuple of Sailors is 50 bytes long
  - a page can hold 80 Sailors Tuples
  - 500 pages of such tuples
- *td* is cost to read/write page; ts is cost to ship page.
- The cost is = 500*td* to scan Sailors
- for each Sailors page, the cost of scanning shipping all of Reserves, which is =1000(*td* + *ts*).
- The **total cost** is = 500*td* + 500, 000(*td* + *ts*).

5.Distributed Query Processing

- Assume Reserves (Paris)and Sailors (London) relations
  - **each tuple of Reserves is 40 bytes long**
  - **a page can hold 100 Reserves tuples**
  - **1,000 pages of such tuples.**
  - each tuple of Sailors is 50 bytes long
  - a page can hold 80 Sailors Tuples
  - 500 pages of such tuples
- *If the query site is different. Consider join at London and result is shipped to the query site.*
- *Number of tuples in result = 100,000, each tuple =90 bytes long (40+50)*
- *Page size is 4000 bytes (80 sailors tuple fit in a page and 50 bytes long, so 80*50 = 4000 bytes)*
- *4000/90 = 44 result tuples fit in a page.*
- *Result size is 100,000 /44 = 2273 pages.*
- *Cost of shipping the result to another site is 2273 ts*

5.Distributed Query Processing

- This cost depends on the size of the result.
- The cost of shipping the result is greater than the cost of shipping both Sailors and Reserves to the query site.

- **2.Ship to One Site (sort-merge-join)**
- The cost of scanning and shipping Sailors, saving it at Paris, and then doing the join at Paris is

  =500(2*td* + *ts*) + (Result)*td*

  =500(2td+ts)+4500 td

  [Cost of sort merge join is 3*(m+n), so 3*(1000+500) = 4500]

- The cost of shipping Reserves and doing the join at London is

  =1000(2*td*+*ts*)+(result)*td*.

  =1000(2td+ts)+4500td

Ship Whole: Transferring the complete relation

R

A	B
3	7
1	1
4	6
7	7
4	5
6	2
5	7

S

B	C	D
9	8	8
1	5	1
9	4	2
4	3	3
4	2	6
5	7	8

R ⋈ S

A	B	C	D
1	1	5	1
4	5	7	8

QUERY: The query asks for R ⋈ S

5.Distributed Query Processing

- Ship Whole vs Fetch as needed:
- Fetch as needed results in a high number of messages
- Ship whole results in high amounts of transferred data
- *Note: Some tuples in Reserves do not join with any tuple in Sailors, we could avoid shipping them.*

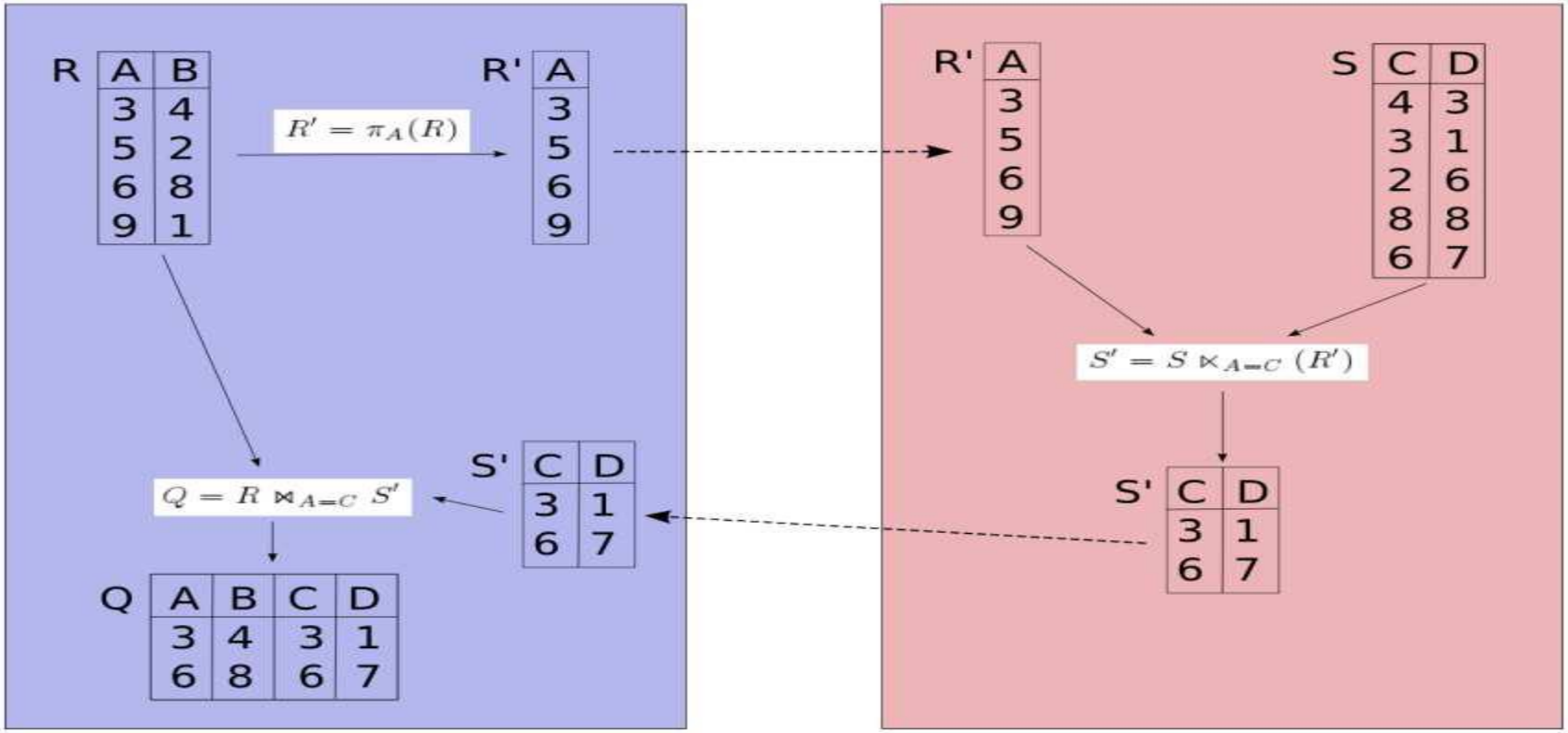
3.Semijoins and Bloomjoins

Semijoins: 3 steps:

1. At London, compute the **projection** of Sailors onto the join columns i.e *sid* and ship this projection to Paris.
2. 2. At Paris, compute the **natural join** of the projection received from the first site with the Reserves relation.
3. The result of this join is called the **reduction** of Reserves with respect to Sailors. ship the reduction of Reserves to London.
4. 3. At London, compute the join of the reduction of Reserves with Sailors.

Semijoin:

- Semijoin: Requesting all join partners in just one step.



5.Distributed query processing

- **Bloomjoins:** 3 steps:

1. At London, A bit-vector of (some chosen) size k is computed by hashing each tuple of Sailors into the range 0 to $k-1$ and setting bit i to 1. if some tuple hashes to i, and 0 otherwise then ship this to DELHI

2. At Paris, the reduction of Reserves is computed by hashing each tuple of Reserves (using the *sid* field) into the range 0 to $k-1$, using the same hash function used to construct the bit-vector, and discarding tuples whose hash value i corresponds to a 0 bit.ship the reduction of Reserves to London.

3. At London, compute the join of the reduction of Reserves with Sailors.

Bloom join:

- Bloom join:
- Also known as bit-vector join
- Avoiding to transfer all join attribute values to the other node
- Instead transferring a bitvector B[1 : : : n]
- Transformation
  - Choose an appropriate hash function h
  - Apply h to transform attribute values to the range [1 : : : n]
  - Set the corresponding bits in the bitvector B[1 : : : n] to

Dr. Geetha Mary A, Asso. Prof., SCOPE, VIT, Vellore

Cost-Based Query Optimization

- optimizing queries in a distributed database poses the following additional challenges:
  - Communication costs must be considered. If we have several copies of a relation, we must also decide which copy to use.
  - If individual sites are run under the control of different DBMSs, the autonomy of each site must be respected while doing global query planning.

Cost-Based Query Optimization

- Cost-based approach; consider all plans, pick cheapest; similar to centralized optimization.
  - ***Difference 1: Communication costs must be considered.***
  - ***Difference 2: Local site autonomy must be respected.***
  - ***Difference 3: New distributed join methods.***
- Query site constructs global plan, with suggested local plans describing processing at each site.
- ***If a site can improve suggested local plan, free to do so.***

6.DISTRIBUTED TRANSACTIONS PROCESSING

- A given transaction is submitted at some one site, but it can access data at other sites.
- When a transaction is submitted at some site, the transaction manager at that site breaks it up into a collection of one or more subtransactions that execute at different sites.
-

Dr. Geetha Mary A, Asso. Prof., SCOPE, VIT, Vellore

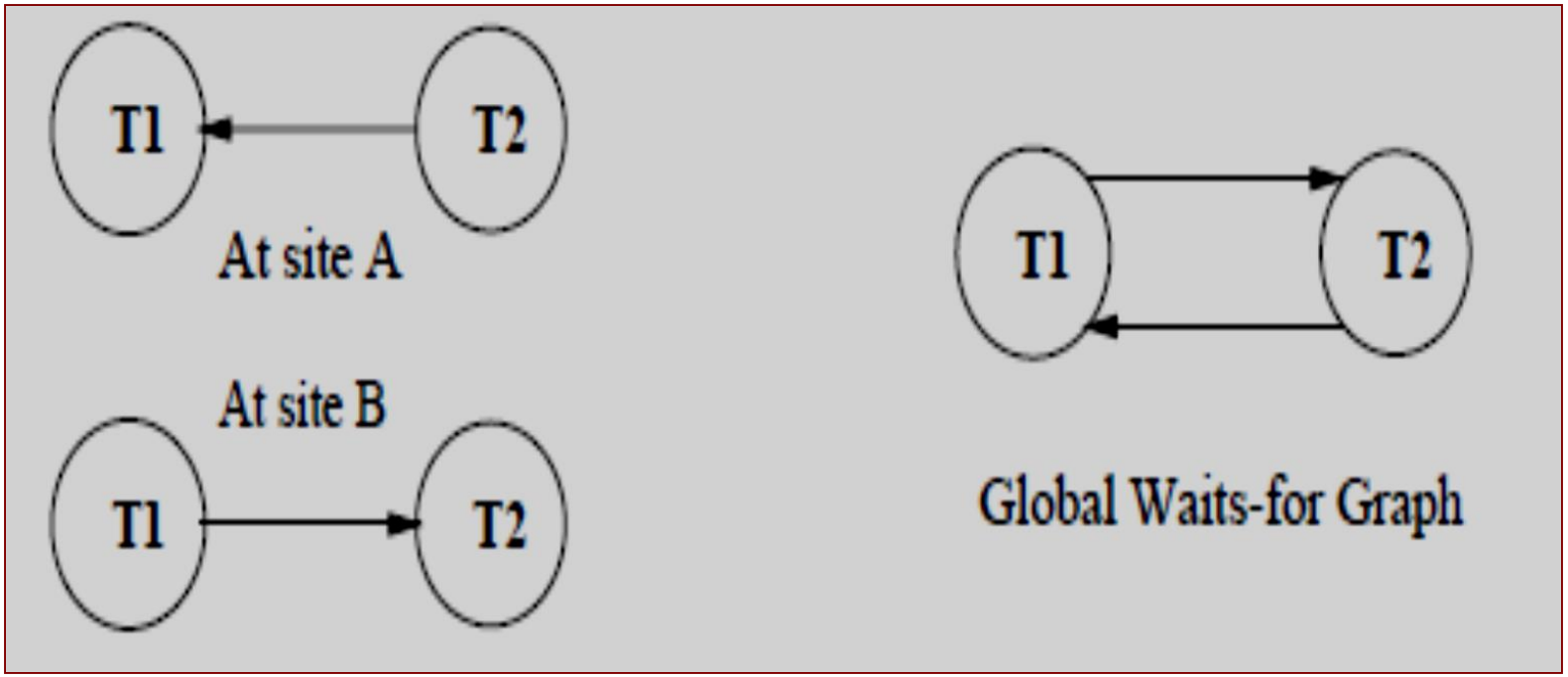
7.Distributed Concurrency Control

- **Lock management** can be distributed across sites in many ways:

1. **Centralized:** A single site is incharge of handling lock and unlock requests for all objects.
2. **Primary copy:** One copy of each object is designated as the primary copy.
   - All requests to lock or unlock a copy of this object are handled by the lock manager at the site where the primary copy is stored.
3. **Fully distributed:** Requests to lock or unlock a copy of an object stored at a site are handled by the lock manager at the site where the copy is stored.

7.Distributed Concurrency Control

- **Phantom Deadlocks:** delays in propagating local information might cause the deadlock detection algorithm to identify `**deadlocks'** that do not really exist. Such situations are called **phantom deadlocks**



Distributed deadlock detection algorithms

- Centralised
- Heirarchical
- Waiting longer!! abort

7.Distributed Recovery

- Recovery in a distributed DBMS is more complicated than in a centralized DBMS for the following reasons:

1. New kinds of failure can arise:
   - failure of communication links.
   - failure of a remote site at which a subtransaction is executing.
2. Either all subtransactions of a given transaction must commit, or none must commit.
3. This property must be guaranteed in spite of any combination of site and link failures.

Dr. Geetha Mary A, Asso. Prof., SCOPE, VIT, Vellore

7.Distributed Recovery

- **Two-Phase Commit (2PC):**

- Site at which Xact originates is **coordinator**;
- other sites at which it executes subXact are **subordinates**.
- When the user decides to commit a transaction:
  - The commit command is sent to the coordinator for the transaction.
  - This initiates the 2PC protocol:

Dr. Geetha Mary A, Asso. Prof., SCOPE, VIT, Vellore

7.Distributed Recovery(2pc)

1. Coordinator sends prepare msg to each subordinate.

2. Subordinate force-writes an abort or prepare log record and then sends a no or yes msg to coordinator.

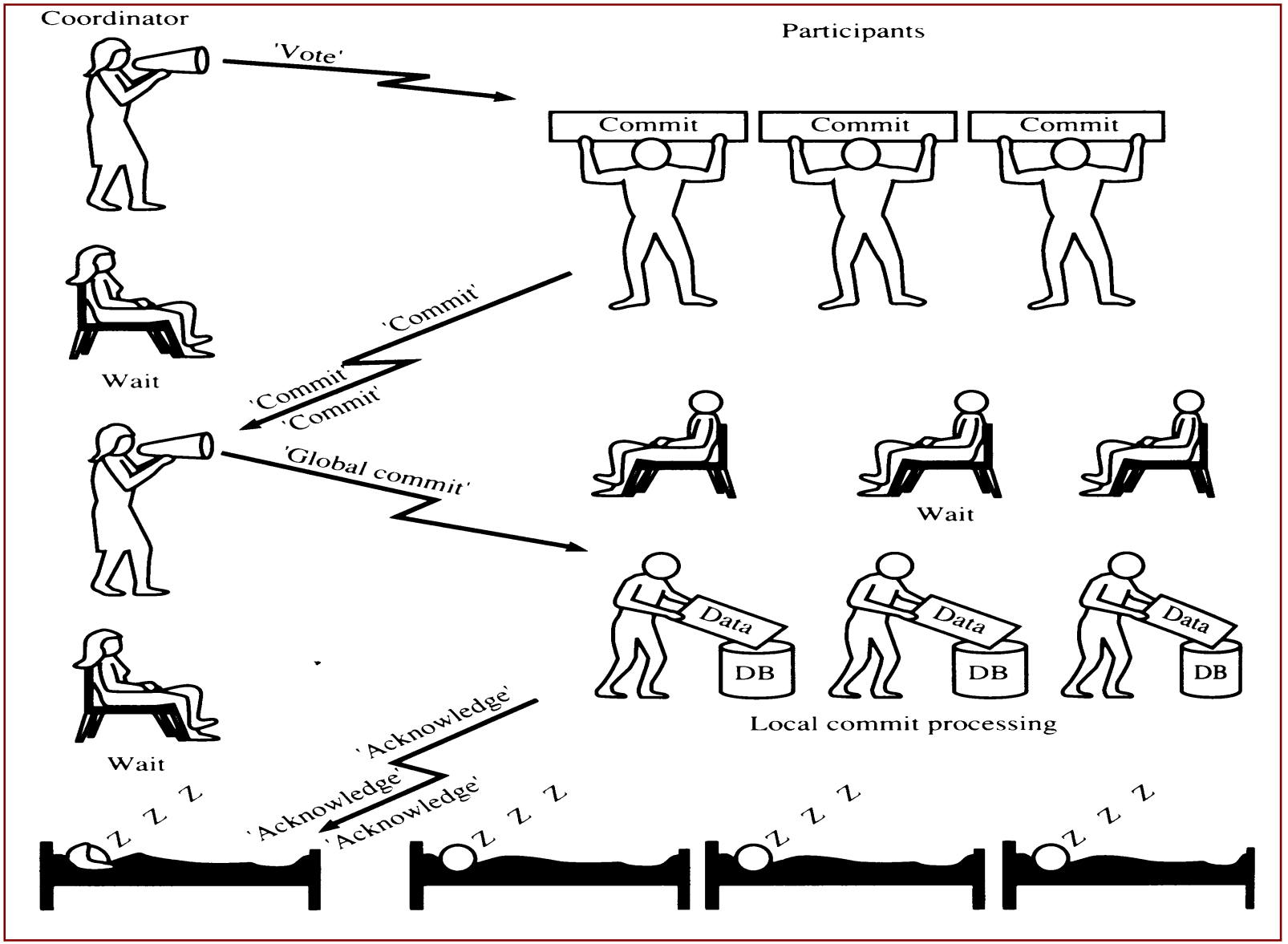
3. If coordinator gets all yes votes, force-writes a commit log record and sends commit msg to all subs. Else, force-writes abort log rec, and sends abort msg.

4. Subordinates force-write abort/commit log rec based on msg they get, then send ack msg to coordinator.

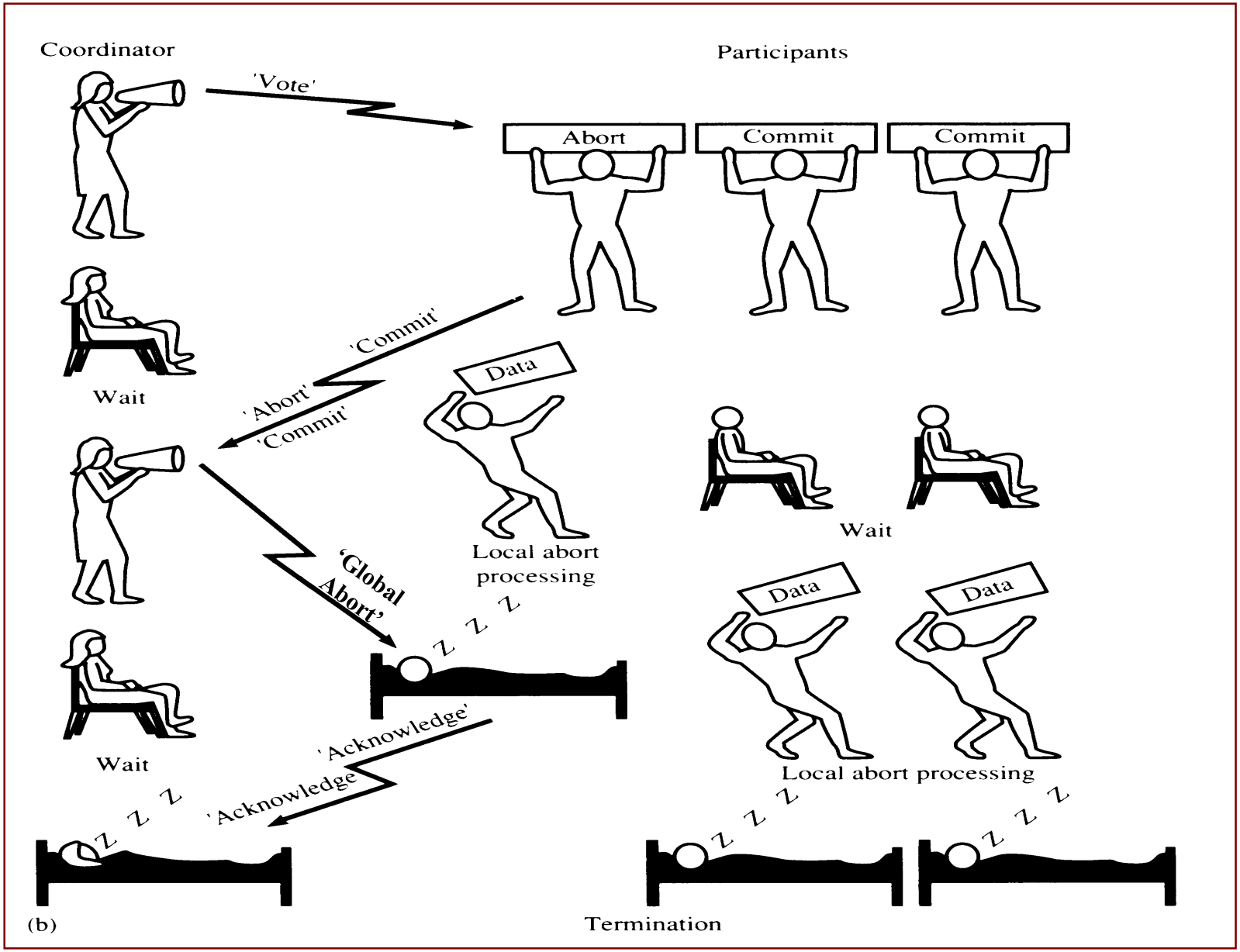
5. Coordinator writes end log rec after getting ack msg from all subs

Dr. Geetha Mary A, Asso. Prof., SCOPE, VIT, Vellore

TWO-PHASE COMMIT (2PC) - commit



TWO-PHASE COMMIT (2PC) - ABORT



7.Distributed Recovery(3pc)

- **Three-Phase Commit**

1. A commit protocol called **Three-Phase Commit (3PC)** can avoid blocking even if the coordinator site fails during recovery.

2. The basic idea is that when the coordinator sends out *prepare messages* and receives *yes* votes from all subordinates.

3. It sends all sites a *precommit* message, rather than a *commit* message.
4. When a sufficient number more than the maximum number of failures that must be handled of *acks* have been received.
5. The coordinator force-writes a *commit* log record and sends a *commit message* to all subordinates.

Dr. Geetha Mary A, Asso. Prof., SCOPE, VIT, Vellore

Distributed Database

- ***Advantages:***
  - Reliability
  - Performance
  - Growth (incremental)
  - Local control
  - Transparency
- ***Disadvantages:***
  - Complexity of Query opt.
  - Concurrency control
  - Recovery
  - Catalog management