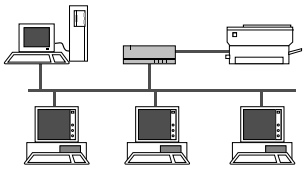


Distributed Systems
600.437
Replication

Department of Computer Science
The Johns Hopkins University

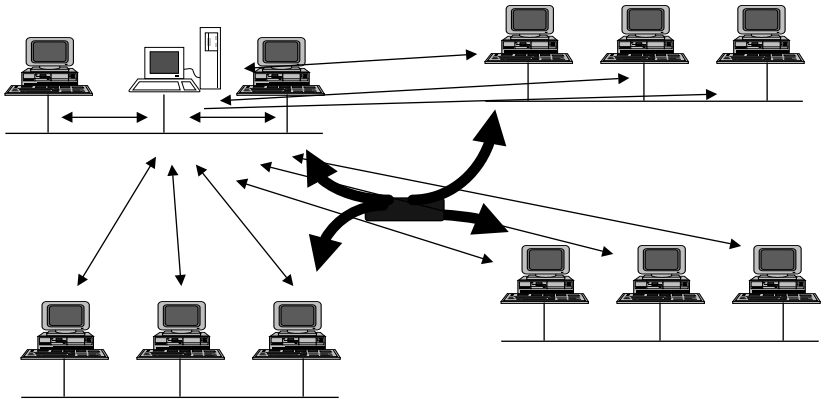
Replication

Lecture 9

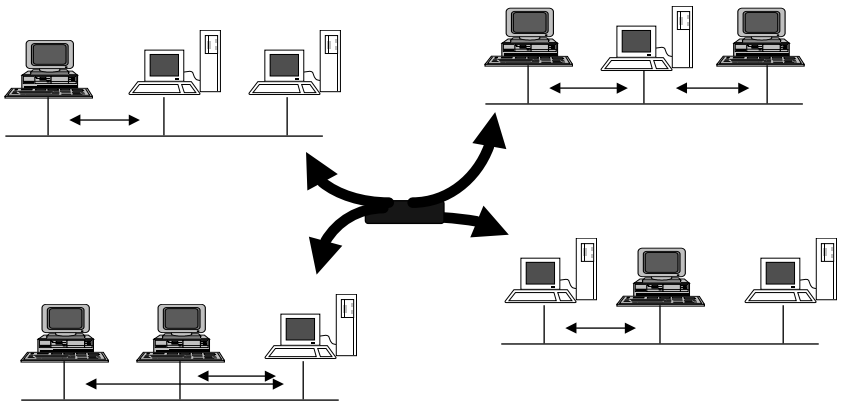
Further readings:

* Distributed Systems (Second edition) Sape Mullender, chapters 7 and 8. (Addison-Wesley) 1994.
* Concurrency control and recovery in Distributed Database Systems Bernstein, Hadzilacos and Goodman (Addison Wesley) 1987.
* Yair's PhD thesis.
* Papers from DISC98 and ICDCS99 on our www.cnds.jhu.edu/publications web page.

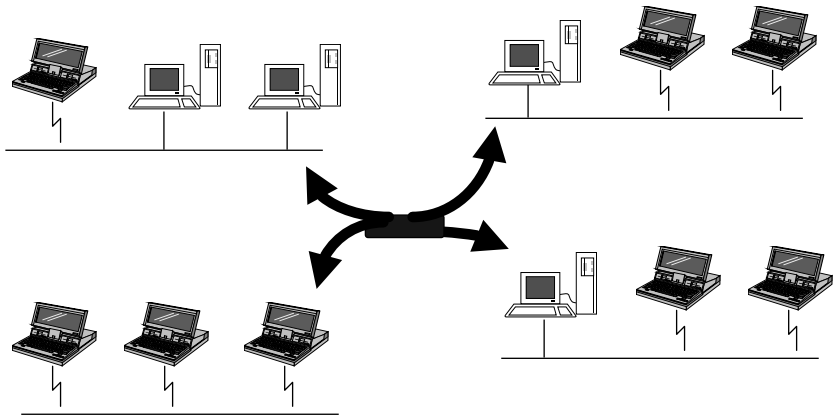
Centralized Server



Replicated Server



Replicated Server !



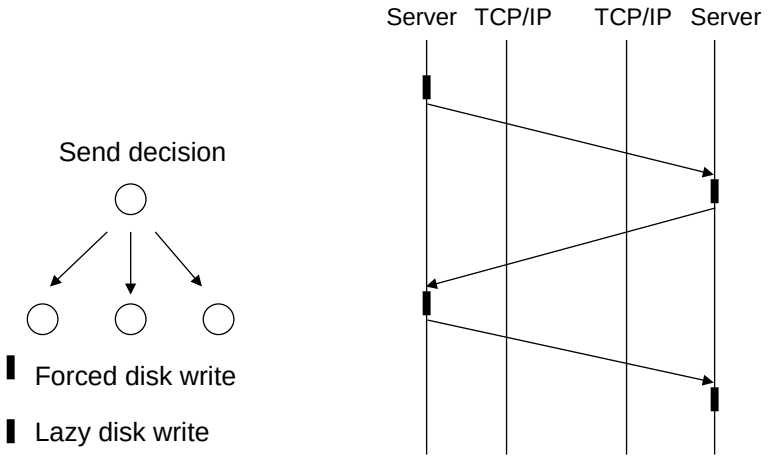
Each laptop is a (weaker?) server

Traveling salespersons

Replication Methods

- Two phase commit, three phase commit
- Primary and backups
- Primary partition
  - What happens when there is no primary partition?
- Weak consistency (weaker update semantics)
- Replication using group communication.

Two Phase Commit

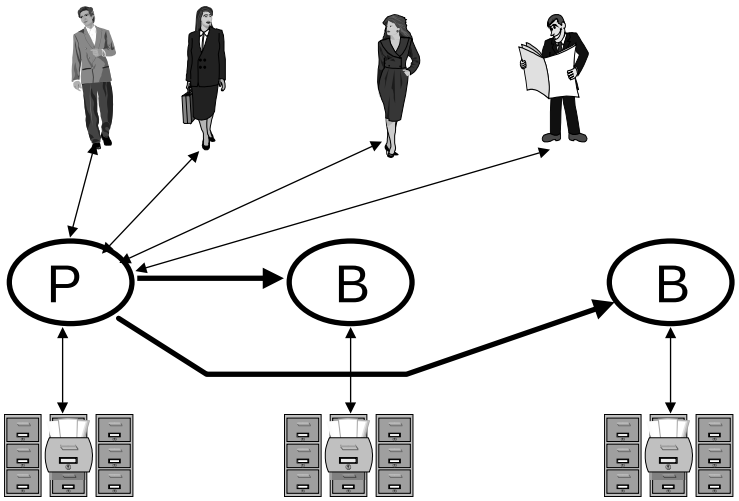


Primary and Backups

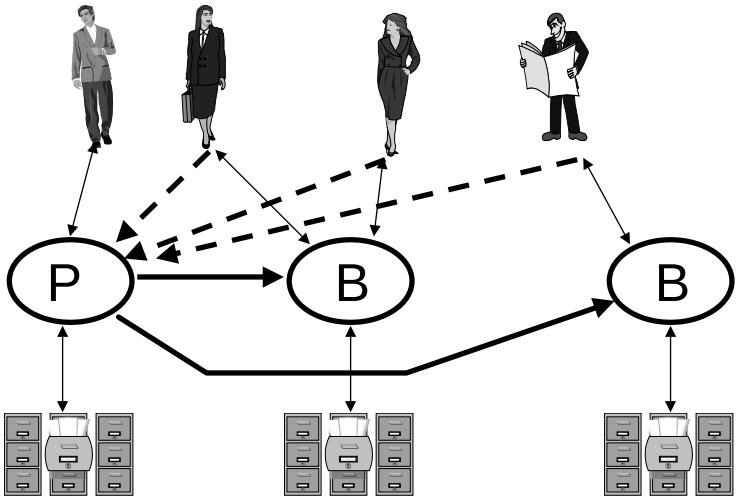
Possible options:

- Backups are maintained for availability only.
- Backups can improve performance for reads, updates are sent to the primary by the user.
  - **What is the query semantics? How can one copy serializability be achieved?**
- The user interacts with one copy, and if it is a backup, the updates are sent to the primary
  - **What is the query semantics with regards to our own updates?**

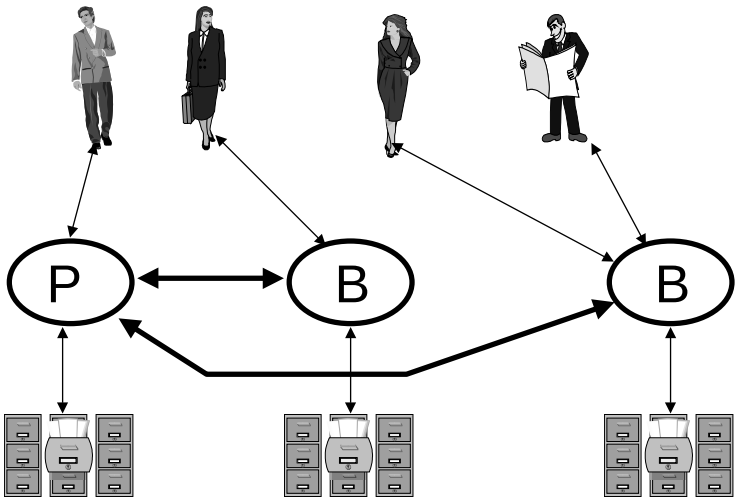
Primary and Backups (1)



Primary and Backups (2)



Primary and Backups (3)



Weak Consistency (weaker update semantics)

The Anti-Entropy method: Golding 92

- State kept by the replication servers can be weakly consistent. i.e. copies are allowed to diverge temporarily. They will eventually come to agreement.
- From time to time, a server picks another server and these two servers exchange updates and converge to the same state.
- Total ordering is obtained after getting one message from every server (directly).
- Lamport time stamps are used to order messages.

The Anti-Entropy method

Knowledge at Server A

A	1	3	5	12
B	2			
C	2	3	4	

Knowledge at Server B

A	1	3			
B	2	5	6	9	11
C	2				

Summary A: 12, 2, 4

Summary B: 3, 11, 2

Numbers refer to Lamport time stamps.

The Anti-Entropy Method (cont.)

Knowledge at Server A

A	1	3	5	12
B	2			
C	2	3	4	

Knowledge at Server B

A	1	3			
B	2	5	6	9	11
C	2				

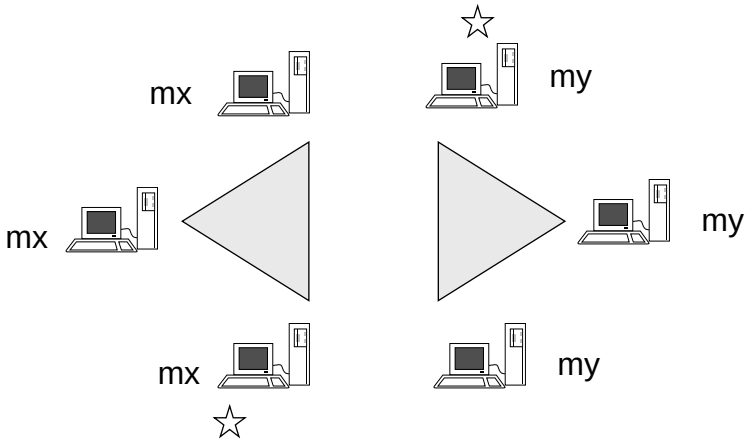
Summary A: 12, 2, 4

Summary B: 3, 11, 2

Summary After merge: 12, 11, 4

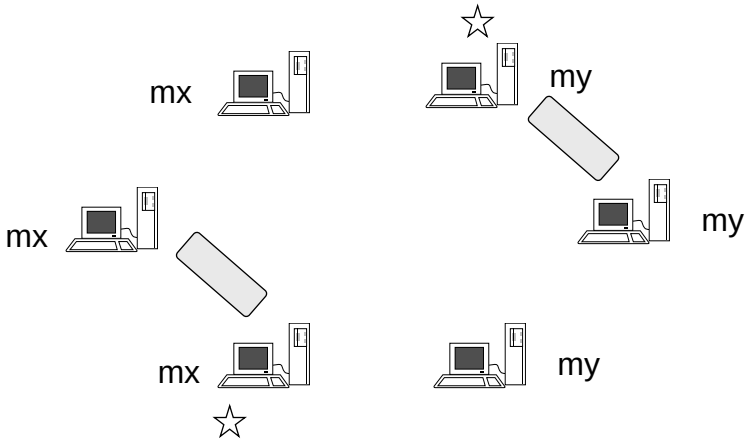
Eventual Path Propagation

Partitioned system



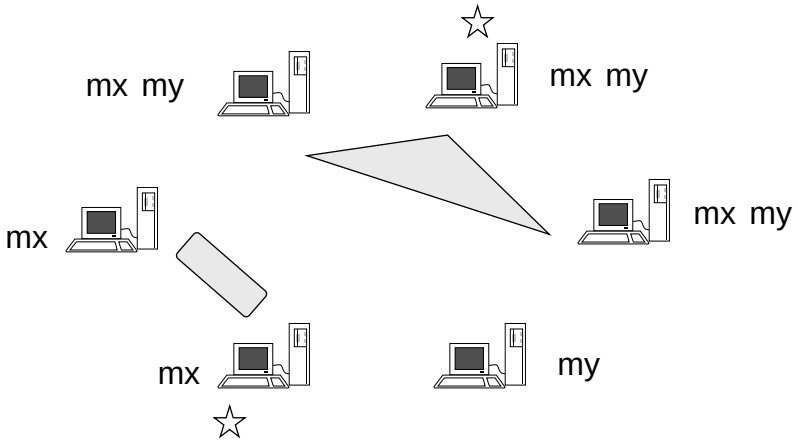
Eventual Path Propagation (cont.)

Further partitioning



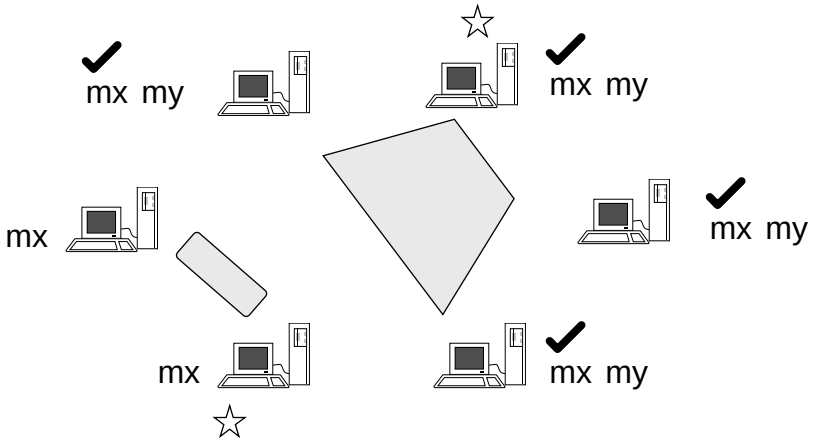
Eventual Path Propagation (cont.)

Merging

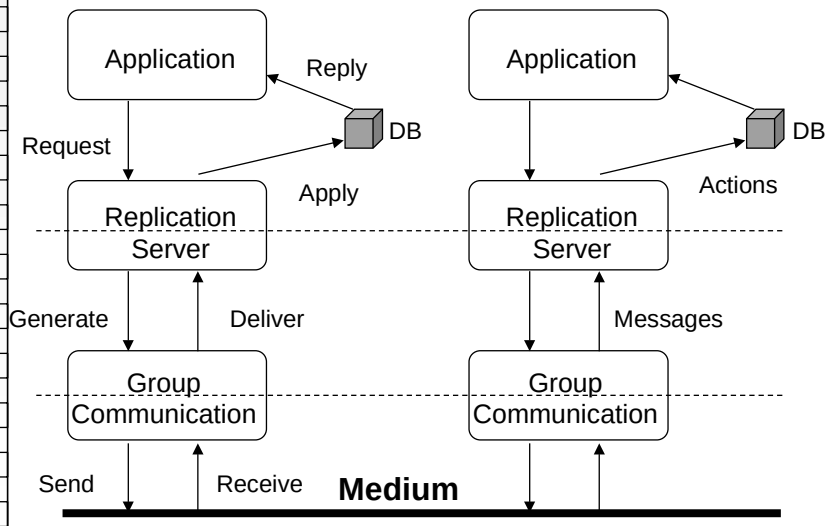


Eventual Path Propagation (cont.)

Further merging



Replication using Group Communication



Replication Engine

- Knowledge propagation
  - Eventual Path Propagation.
- Amortizing end-to-end acknowledgments
  - Low level Ack derived from Safe Delivery of group communication.
  - End-to-end Ack upon membership changes.
- Primary component selection
  - Dynamic Linear Voting.

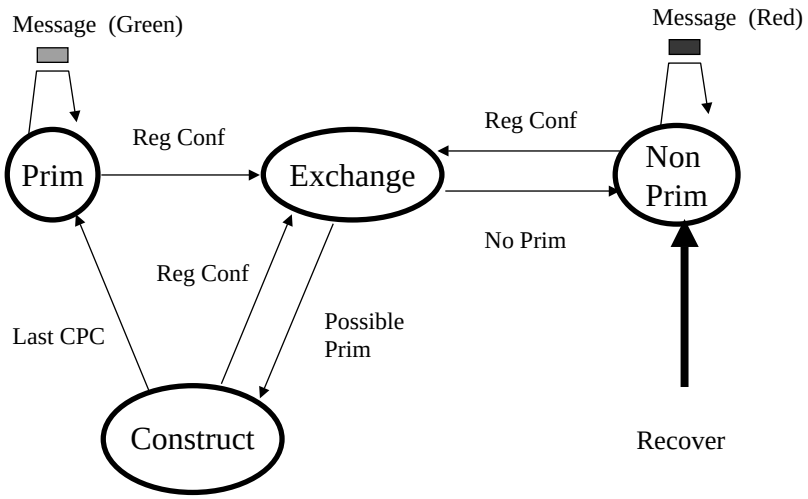
Action Ordering

(Red) Order is unknown

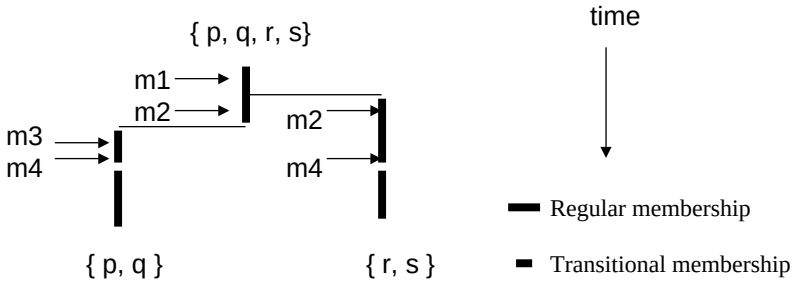
(Green) Order is known

(White) (I know that)
Order is known to all

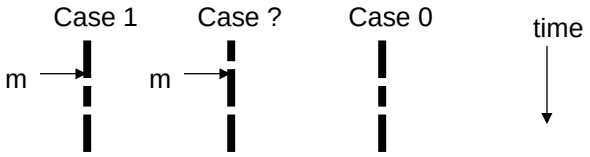
Conceptual State Diagram



Network Partitions and Uncertainty

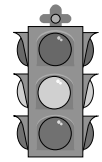


Consensus is not possible but...



- Case 0 and Case 1 can not both have processes.
- Case 1 process “commits” knowing that all the partitioned processes have m.
- Case 0 process knows that no process “commits”.
- Even Case ? process knows valuable information

Action Ordering



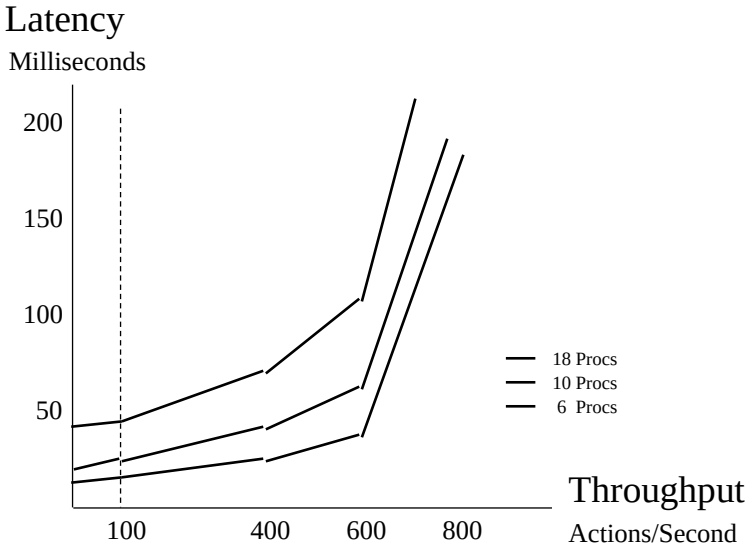
(Red) Order is unknown

(Yellow) Transitional membership

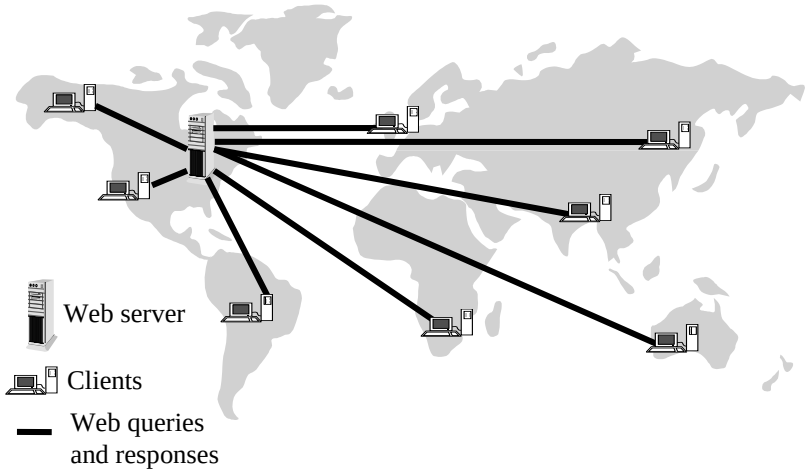
(Green) Order is known

(White) (I know that)
Order is known to all

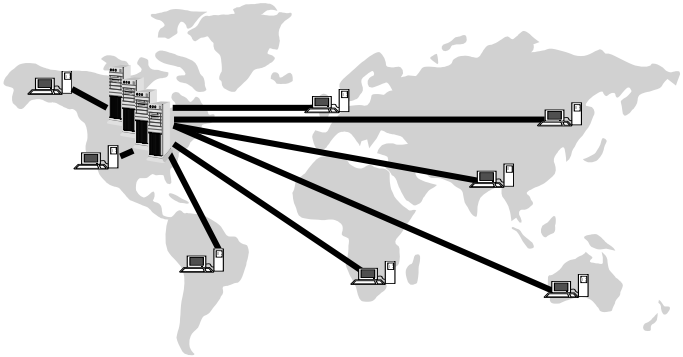
The Performance Difference



The Web Model - A Test Case for Replication

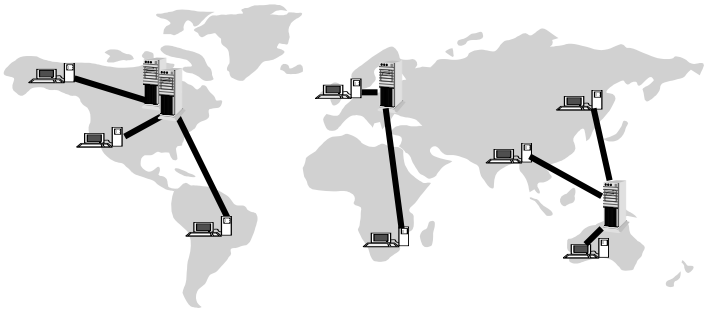


Cluster Replication



- Improves performance - the load is shared by several servers.
- Improves the availability of the web server.
- Moderate effort is required to set up and maintain.
- Single point of failure in the network.
- High latency for clients that are distant (network-wise).

Wide-area (Cluster) Replication

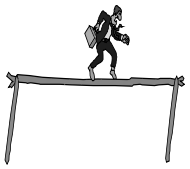


- Improves performance
  - the load is shared by several servers.
  - Clients access the “best” server.
- Improves the availability of the web service
  - network availability.
  - server availability.
- Complex to implement, deploy and maintain.

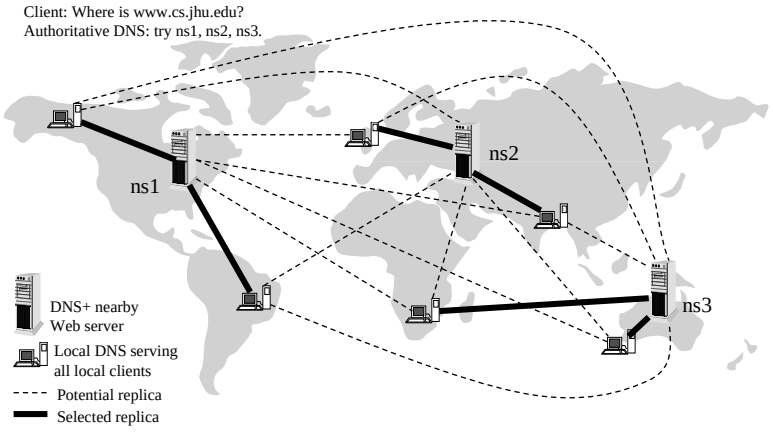
The Technical Challenge

Making wide-area cluster replication work in a Web environment.

- Get the “nearest” server:
  - Based on network topology.
- From the nearest to the best:
  - Server availability
  - Server load
- Do it automatically and seamlessly:
  - HTTP redirect method - Application layer.
  - DNS round trip method - Session layer.
  - Shared IP address method - Network layer.

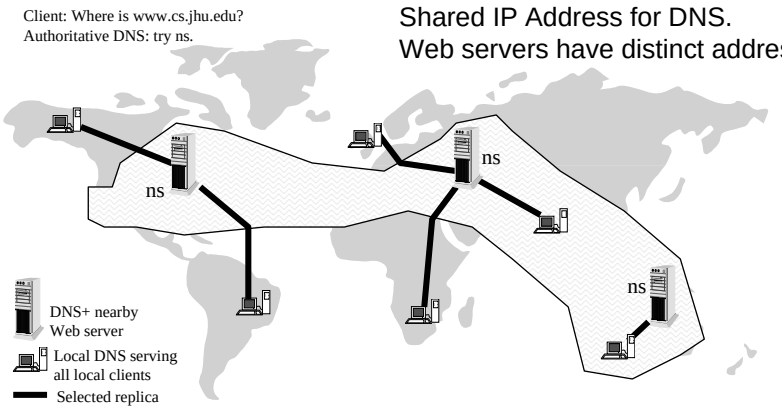


DNS Round Trip Times Method



- No special requirements.
- Convergence time is linear with the number of replicas.

Shared IP Address Method



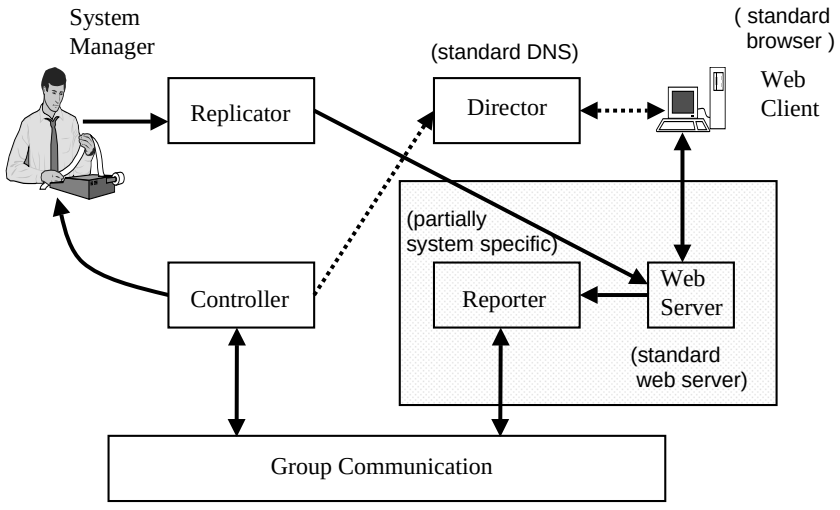
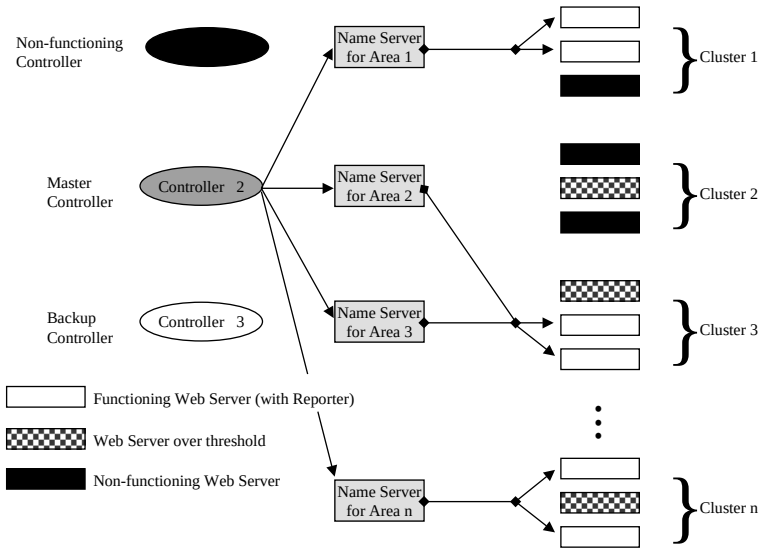
- Accurate with no convergence time.
- Requires one autonomous system to host all of the replicas.

Practical Implementation: The Walrus System

*A **W**ide **A**rea **L**oad **R**e-balancing **U**ser-transparent **S**ystem.*

- No change to the web server.
- No change to the web client.
- No change to the infrastructure (ISP, DNS, OS).
- Implemented in a Unix environment, but can be easily ported to other environments.
- The Walrus system is available - see http://www.cnds.jhu.edu/walrus

The Walrus Architecture



Walrus in Action