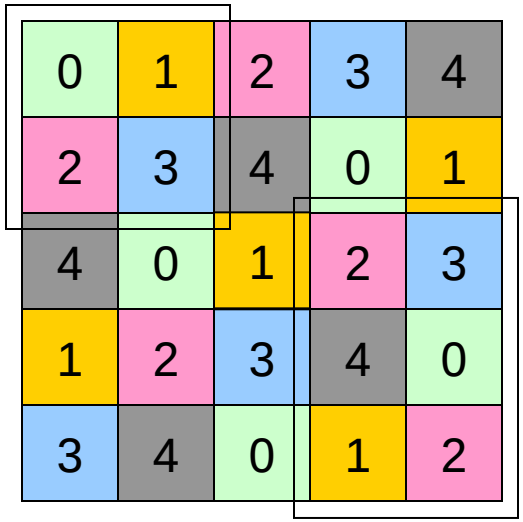
Multi-site Retrieval of Declustered Data

Ali Saman Tosun
Department of Computer Science
University of Texas at San Antonio

Introduction

- Spatial Databases
  - Cartography, transportation, GIS
  - Large data repositories
  - Efficient retrieval and storage
- Index structures developed for single disk, single processor environments are *ineffective*
- Multiple disk architectures
  - Exploit I/O parallelism
  - *Declustering*

Declustering

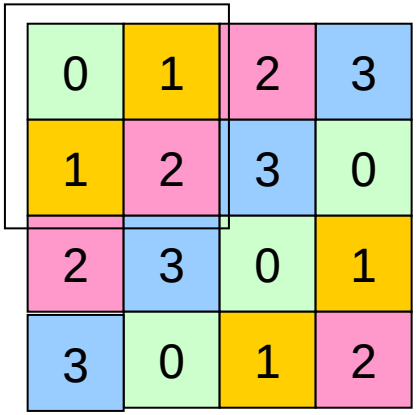


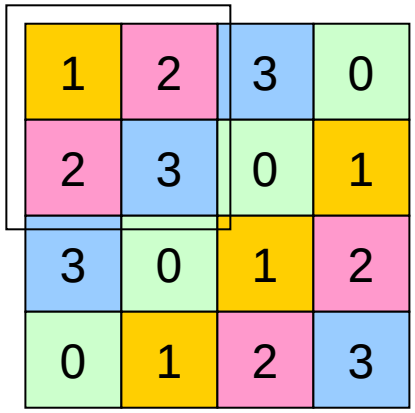
5-by-5 grid using 5 disks

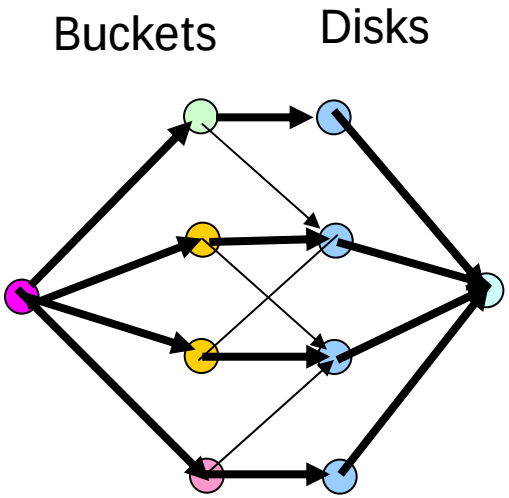
$$OPT = \left\lceil \frac{ij}{N} \right\rceil$$

- Optimality for all range queries not possible when N>5
- Worst-case additive error is O(log N)
- Schemes proposed to achieve good performance in *average case* or *worst-case*

Replication





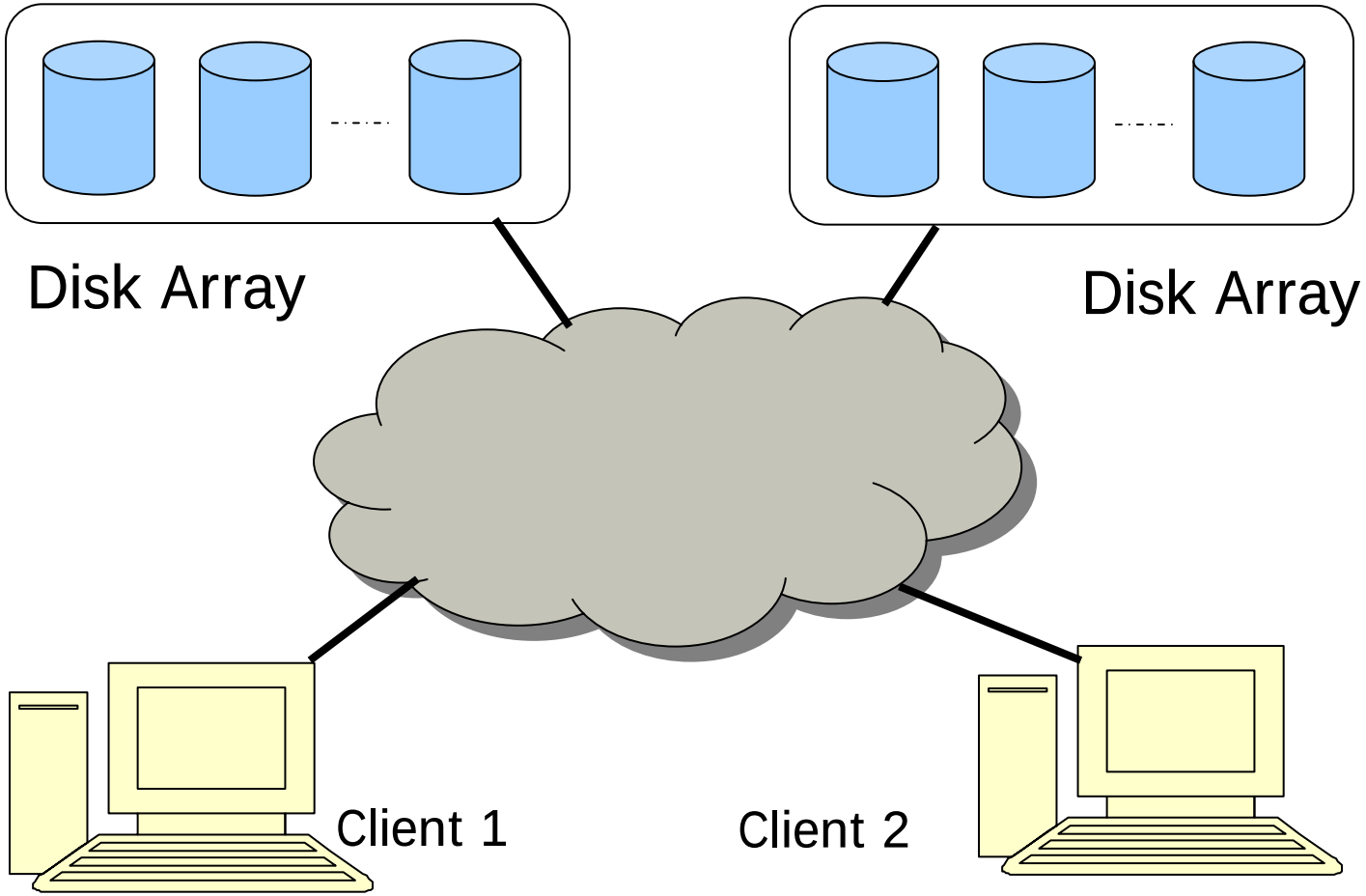


- Reduce worst case additive error
- Maxflow computation needed to check for optimality
- Spatial range queries are targeted

Multi-site Retrieval Problem

- Replication of data at different sites
  - Declustering typically uses single site replication
- Formulate the problem as a maximum flow problem
- Propose a low complexity online algorithm
- Evaluate various replication schemes, query types, query loads.
- Easy to extent to nonuniform grid, more than 2 sites and nonidentical disk arrays

Multi-site Retrieval Problem

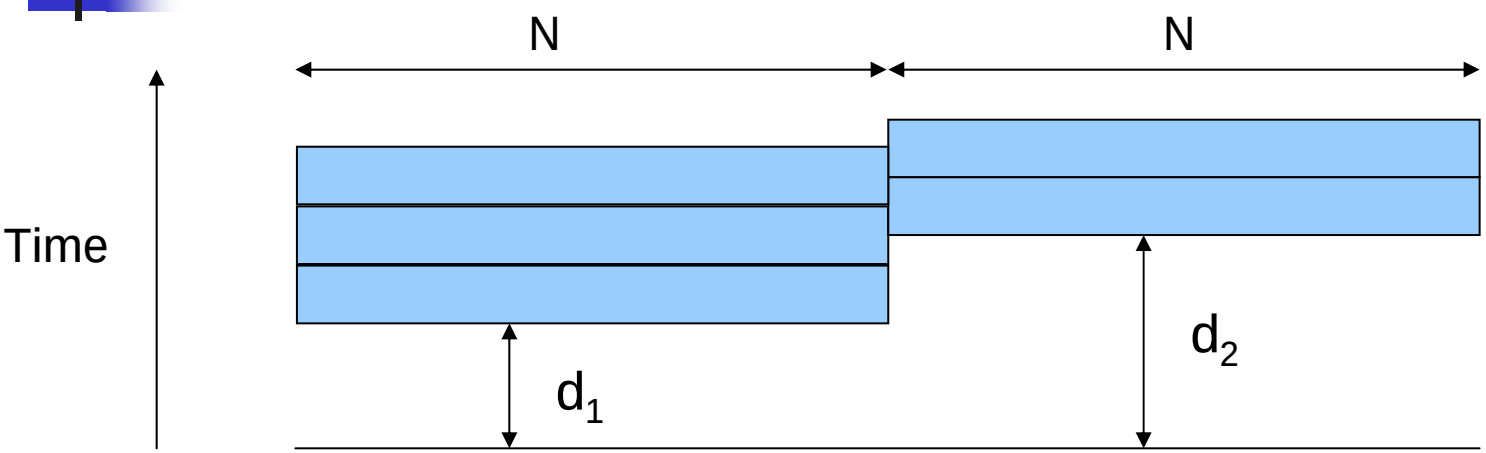


System Model and Goals

- Data is stored using a NxN declustering system with N disks at each site
- Each client has a constant delay to each of the copies

- What is the optimal retrieval time?
- How can we compute best retrieval schedule?

Optimal Retrieval Time

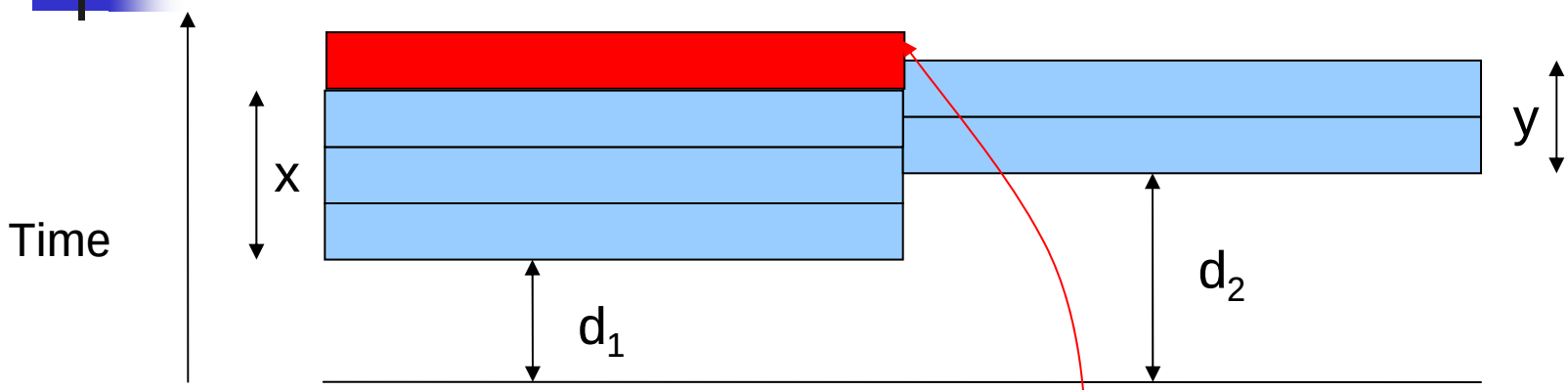


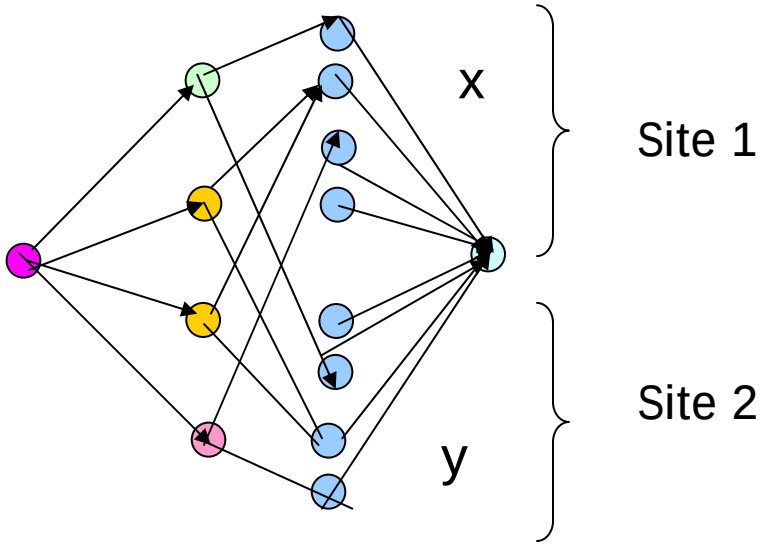
Theorem: Consider a query Q and Let A= $\left\lceil \frac{|d_2 - d_1|}{t_{seek}} \right\rceil$ and B = $\left\lceil \frac{|Q|}{N} \right\rceil$

Optimal Retrieval Time for query Q is

$$\max\{\min(d_1, d_2) + (A + \left\lfloor \frac{B-A}{2} \right\rfloor) tseek, \max(d_1, d_2) + \left\lceil \frac{B-A}{2} \right\rceil tseek\}$$

Maximum Flow Formulation





If maximum flow <> |Q|
increment x and try again

Maximum Flow Representation

- *Theorem*: There exists a flow of size |Q| from the source to the sink in the flow graph if and only if there exists a solution to the multi-site retrieval problem with response time <=max($d_1+c_1*t_{seek}$,$d_2+c_2*t_{seek}$)

- *Theorem*: If a query Q where $\left\lceil \frac{|d_1-d_2|}{t_{seek}} \right\rceil N \geq |Q|$ can be retrieved in $\min(d_1, d_2) + \left\lceil \frac{|d_1-d_2|}{t_{seek}} \right\rceil t_{seek}$ time from the closer (minimum delay) site, then optimal retrieval uses a single site.

Online Algorithm

- Online algorithms are necessary for real world settings
  - Make a decision with the currently available information
  - Combine multiple queries into a single query and use online algorithm.
- Makes a decision for retrieval for each bucket
  - Decision is whether to retrieve it from site1 or site2
- Algorithm is efficient and has O(|Q|) complexity

- Idea: Consider retrieval from site 1 and site 2. Schedule it for retrieval from the site that minimizes retrieval time.

Replication Scheme

- *Random Duplicate Allocation*: store each bucket on 2 randomly chosen disks
- *Orthogonal Allocation*: when the disks that a bucket is stored at is considered as a pair, each pair appears only once
- *Partitioned Replication*: set of disks are divided into groups and disks in a group are replicated on other disks in that group.
- *Dependent periodic allocation*: allocate a shifted version of first copy as second copy.

Query Types

- Range Query: Rectangular in shape. Most common query type
- Arbitrary Query: Any subset of set of buckets
- Connected Query: buckets form a connected graph when neighboring nodes are connected

0	1	2	3
1	2	3	0
2	3	0	1
3	0	1	2

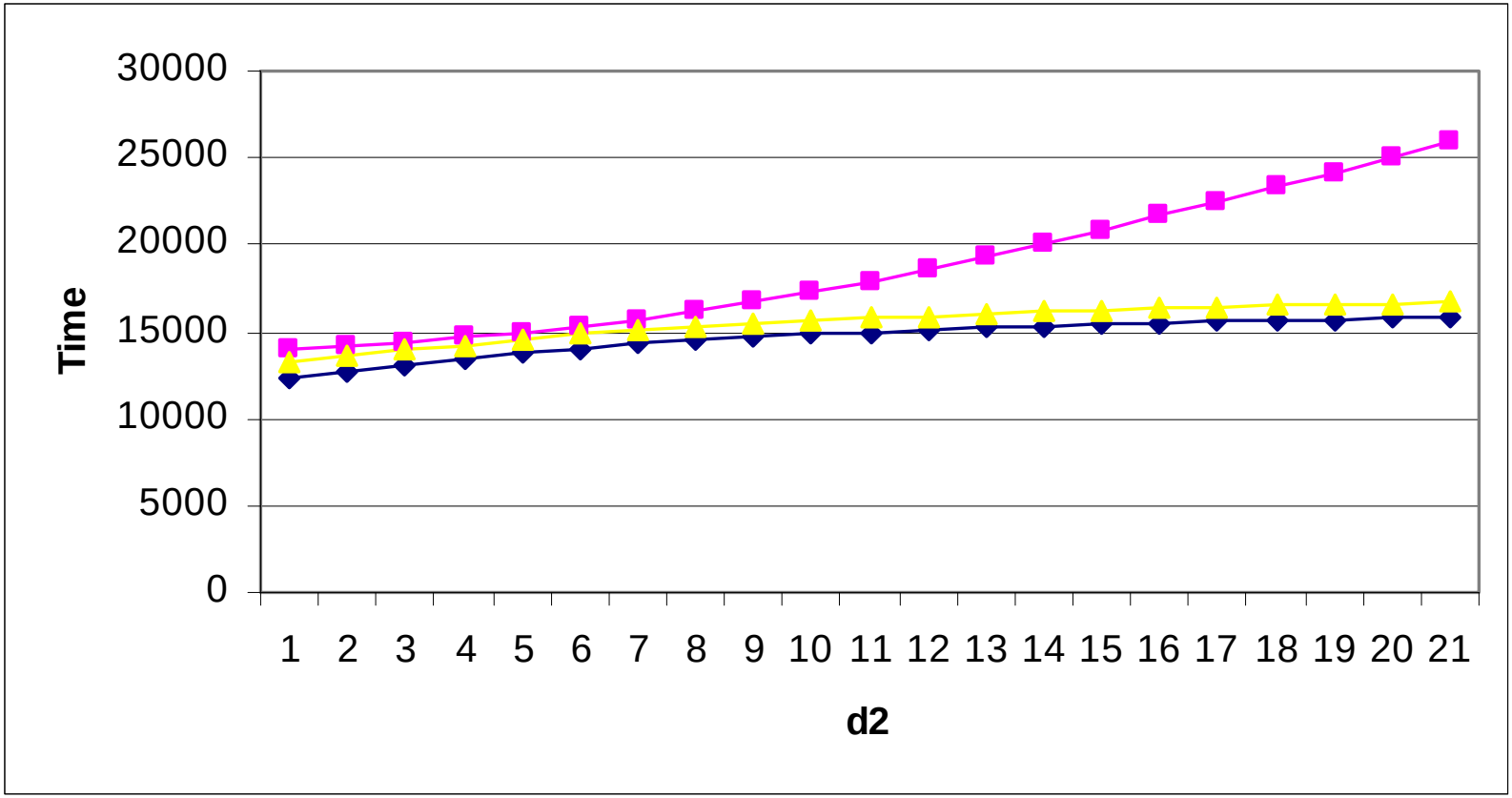
0	1	2	3
1	2	3	0
2	3	0	1
3	0	1	2

0	1	2	3
1	2	3	0
2	3	0	1
3	0	1	2

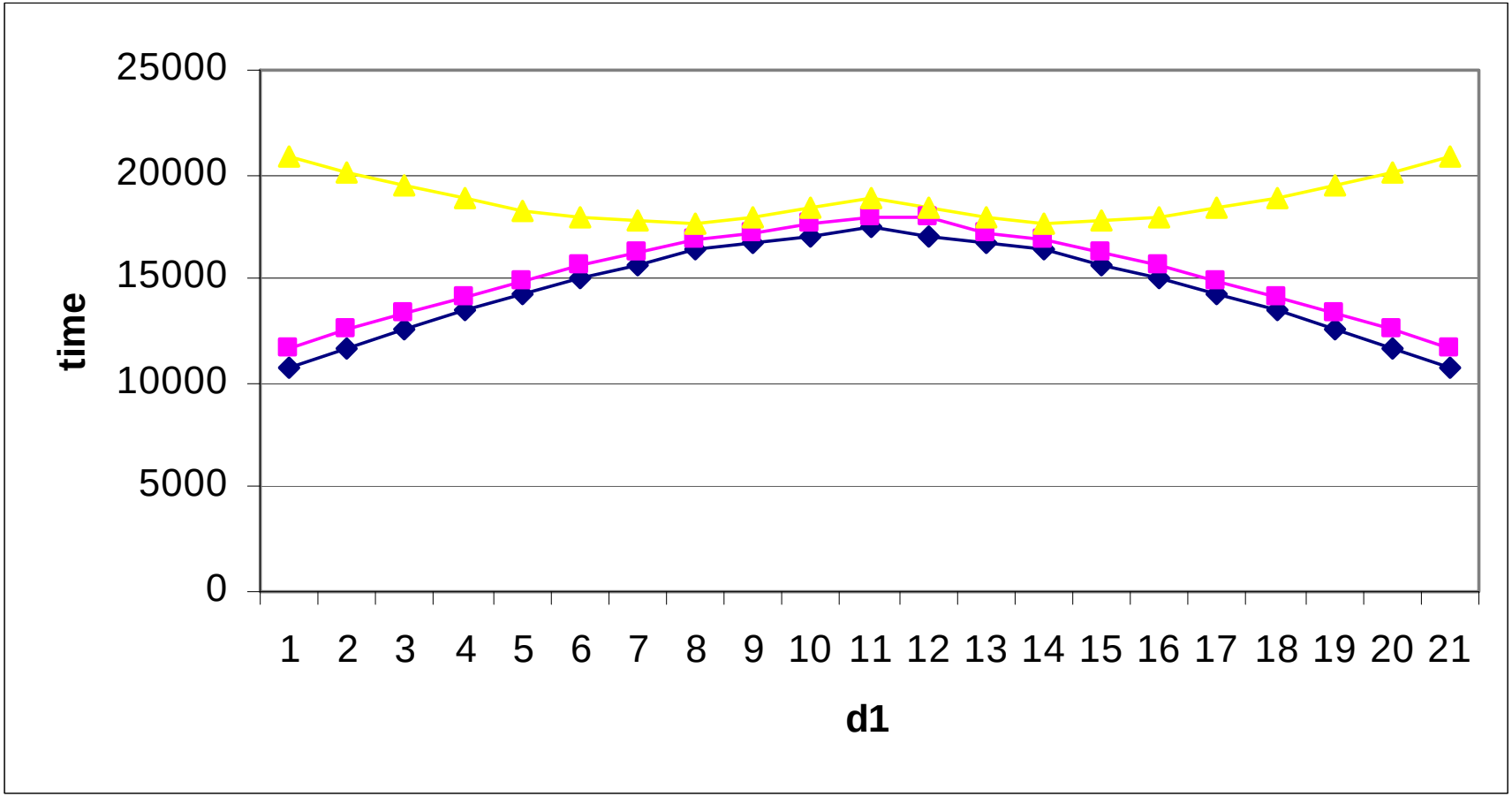
Load Generation

- Query Load
  - *Load1*: Distribution of queries reflect the distribution of queries of the particular query type.
  - *Load2*: Distribution of queries is uniform
  - *Load3*: Smaller queries more likely
- Approach: Pick k (number of disk accesses) and then pick a query which requires k disk accesses.

50 disks, Orthogonal, Range, Load 3



N=50, Orthogonal, Range, Load 3



Summary of Results

- Behavior for replication scheme similar to single site case.
- Best performance achieved when delays are close to each other.
- Online algorithm performs best when delays are close to each other.