A Database Caching Scheme for Wireless and Mobile Clients[a]

Sandeep K. S. Gupta

CSE494/598 Mobile Health and Social Networking
Spring 2009
Arizona State University

[a] Joint work with A. Kahol, S. Khurana, and P. K. Srimani

Talk Outline

- Overview of Mobile Computing
- Data management issues in mobile environment
- Database caching
- Related work
- A new scheme
- Experimental results
- Conclusions

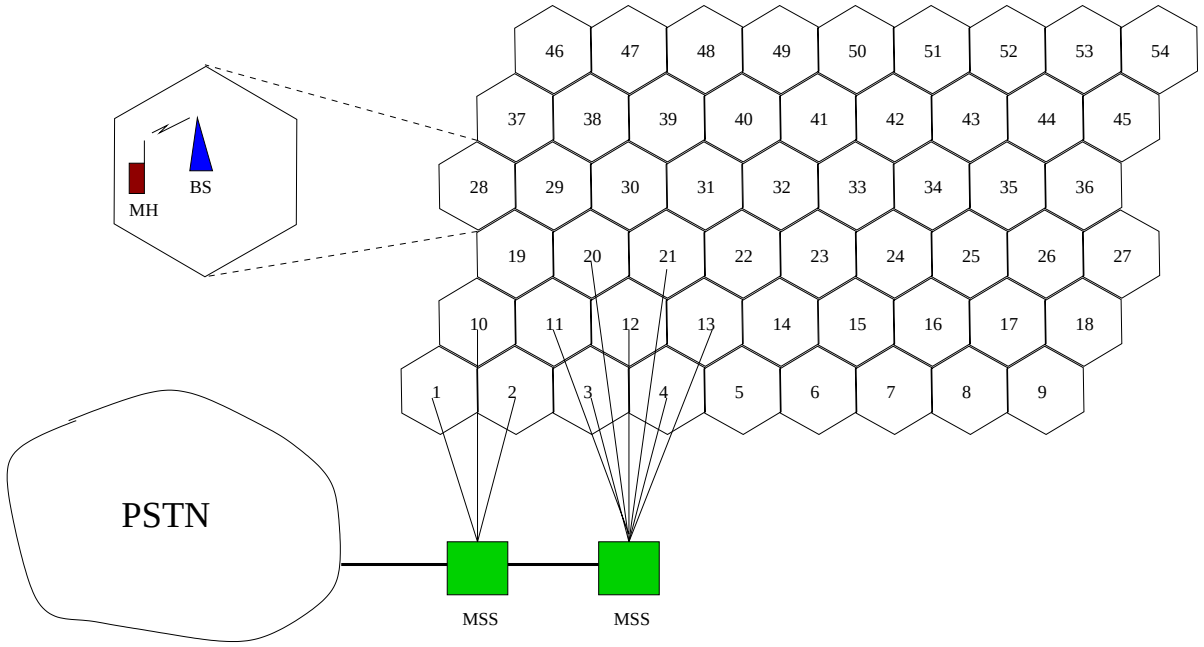
Wireless Communication

- One global bandwidth shared by all users
  - new multiple access techniques (CDMA, CSMA/CS)
  - spatial reuse (cellular networks)
- Signal fading problems
  - short-term multipath fading (Rayleigh effect): due to same signal taking different paths and arriving at the receiver
  - long-term fading (radio shadow) due to topology of the terrain
- High-bit error rates
- Low bandwidth radio links

Mobile Networks

- **Cellular Phone Networks**
  - *Analog:* Advance Mobile Phone Systems (AMPS), Total Access Communication System (TACS)
  - *Digital:* Global System for Mobile Communication (GSM), Personal Digital Cellular (PDC)
- **Satellite Networks:** Immarsat C, Motorola's Iridium
- **Wide Area Wireless Data Systems:** MOBITEX (8Kpbs, 900MHz), Cellular Digital Packet Data (CDPD) (19.2Kpbs, 800MHz), Metricom (76Kbps, 915MHz)
- **High Speed Wireless LANs:** (IEEE 802.11(1Mbps), High Performance Radio LAN (HIPERLAN) (20Mbps)
- **Overlay Networks**: Provide Global Mobility

Cellular Network Architecture



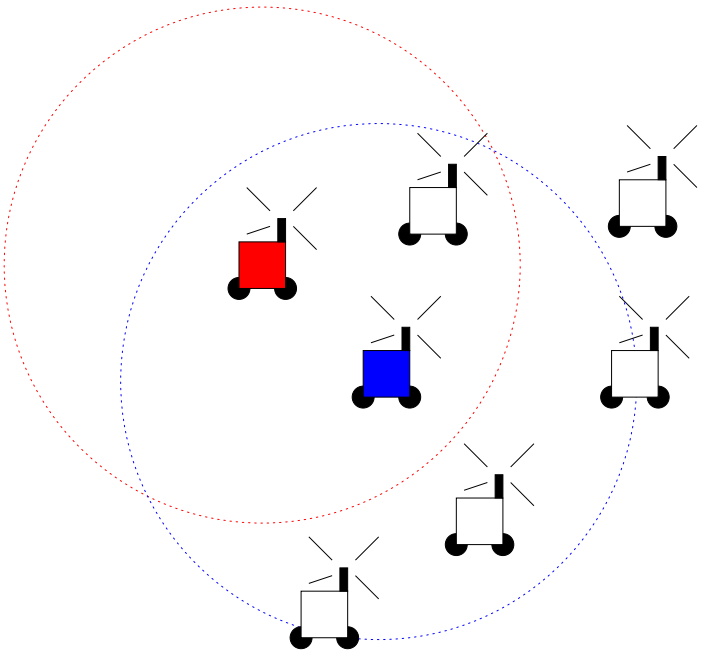
PSTN: Public Switched Telephone Network

MH: Mobile Host

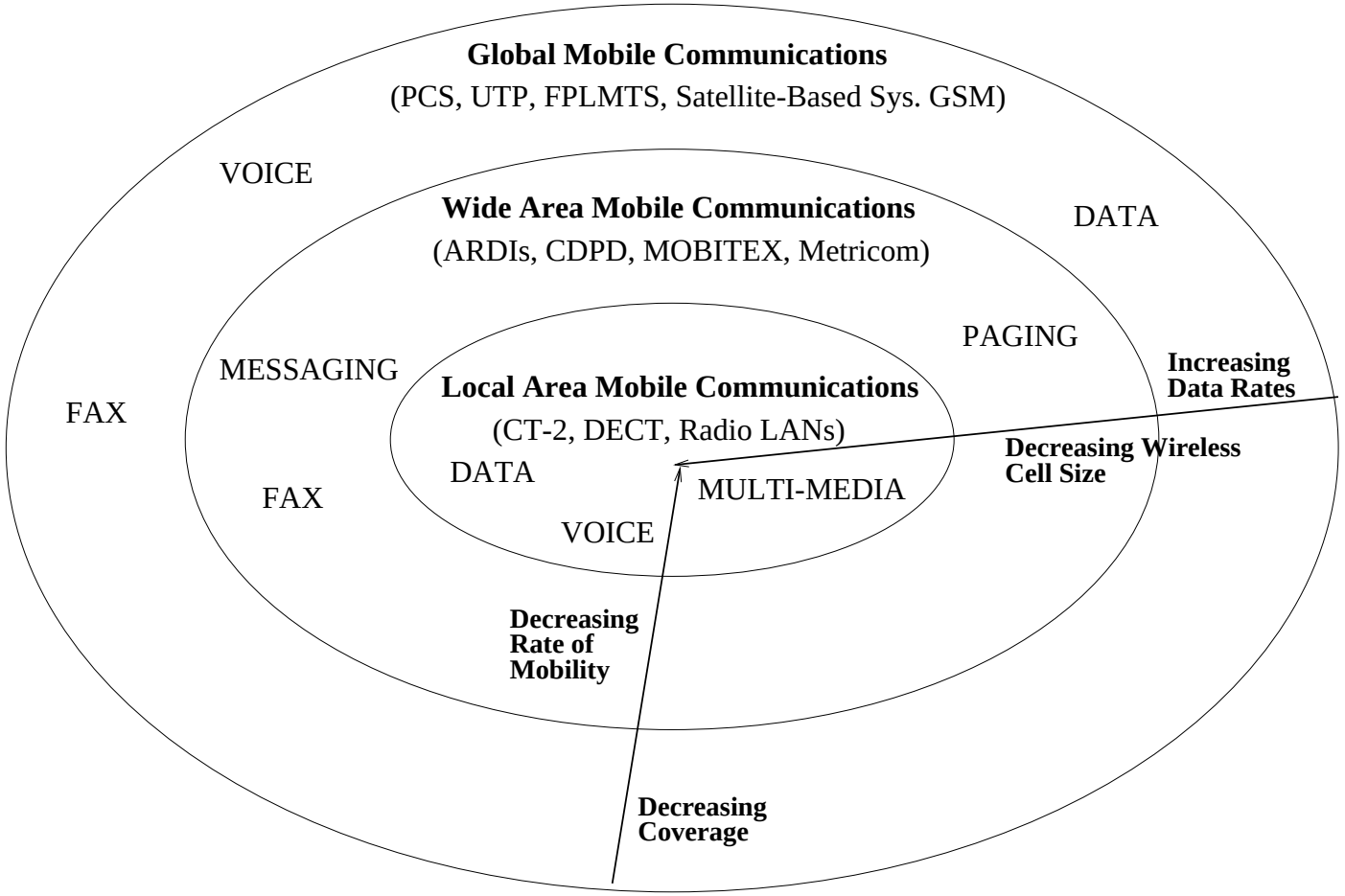
BS: Base Station

MSS: Mobile Switching Center

Mobile Ad Hoc Networks (MANET)

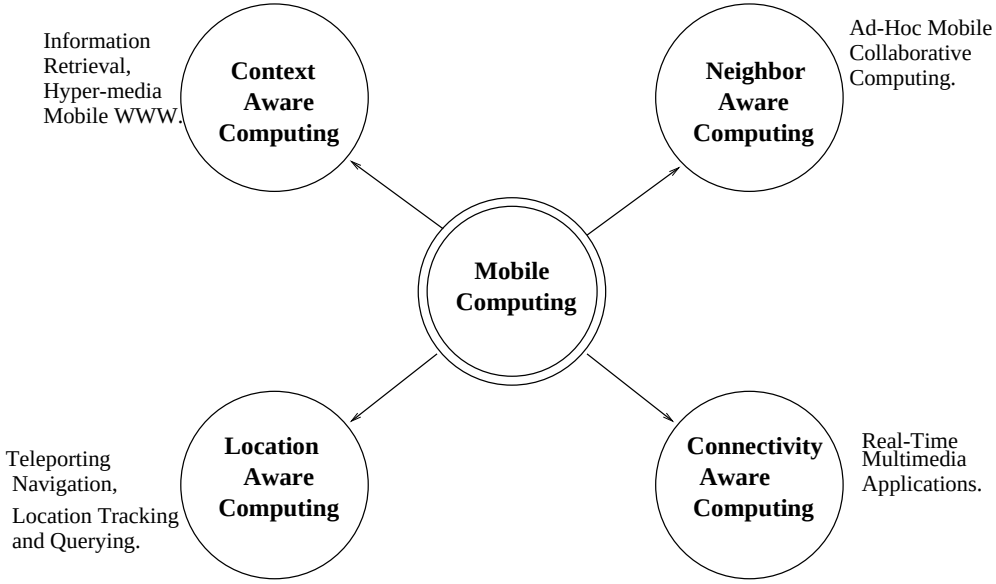


Wireless Systems, Services, and Characteristics



Mobile Computing

- Low-cost, low-power portable computing devices (laptops, PDAs).
- Computing while on move, Anytime and anywhere computing.
- Computing = word processing, database retrieval, mathematical calculation ...



Challenges in Mobile Computing

Application Layer

Resource Discovery - File servers - Print servers	Profile Management New Multimedia Applications

Transport Layer

Flow and Congestion Control

End-to-end Quality-of-Service (QoS)

Network Layer

Addressing and Routing	QoS Cotrolled Handoff
Location Mangement	Authentication

Physical/ Link Layer

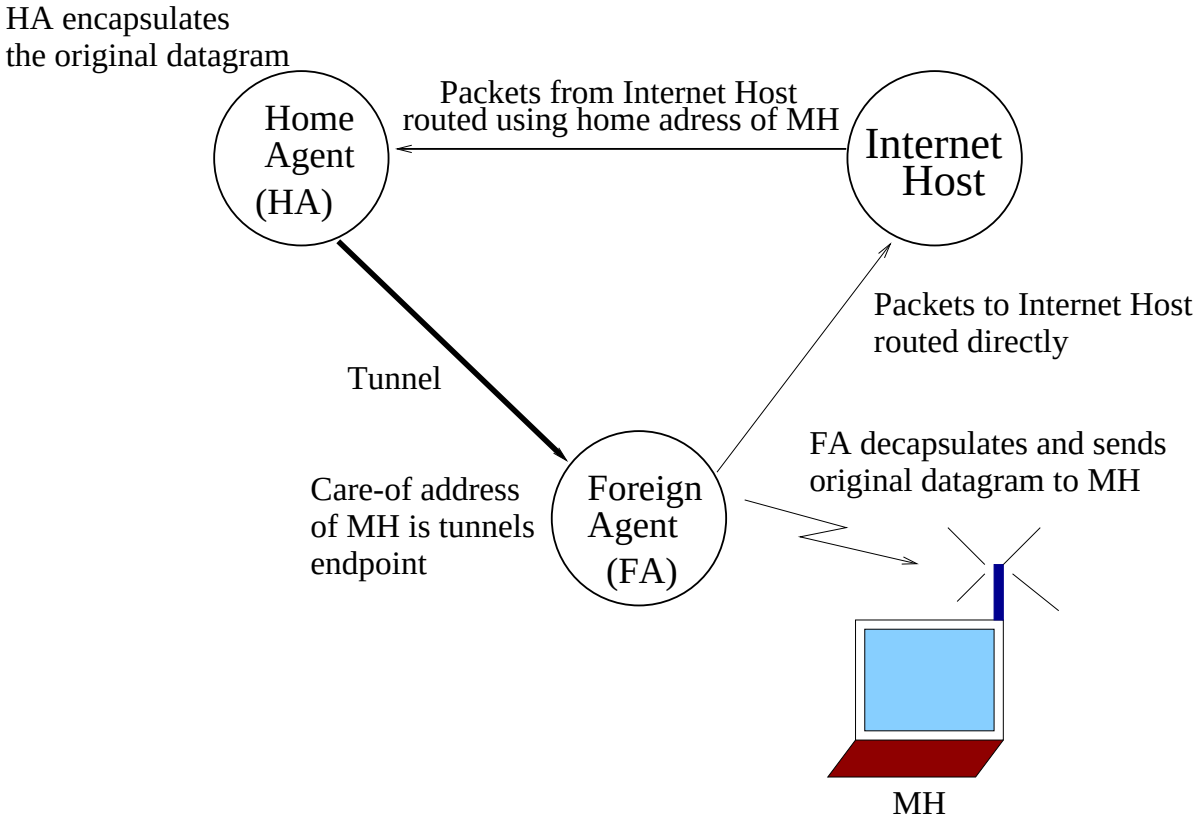
Signal Modulation	Power Control
Encryption/ Compression	Channel Access - TDMA, CDMA, FDMA

Mobility Management

- Enables networks to support mobile users, allowing them to move, while simultaneously offering them incoming calls, data packets and/or other services.
- In connection-oriented networks, mobility management consists of:
  - location management: tracking mobiles and locating them prior to establishing an incoming call.
  - handoff management (a.k.a. automatic link transfer): rerouting connections, on which the mobile user was communication while moving, with minimal loss of user data.

Mobile IP

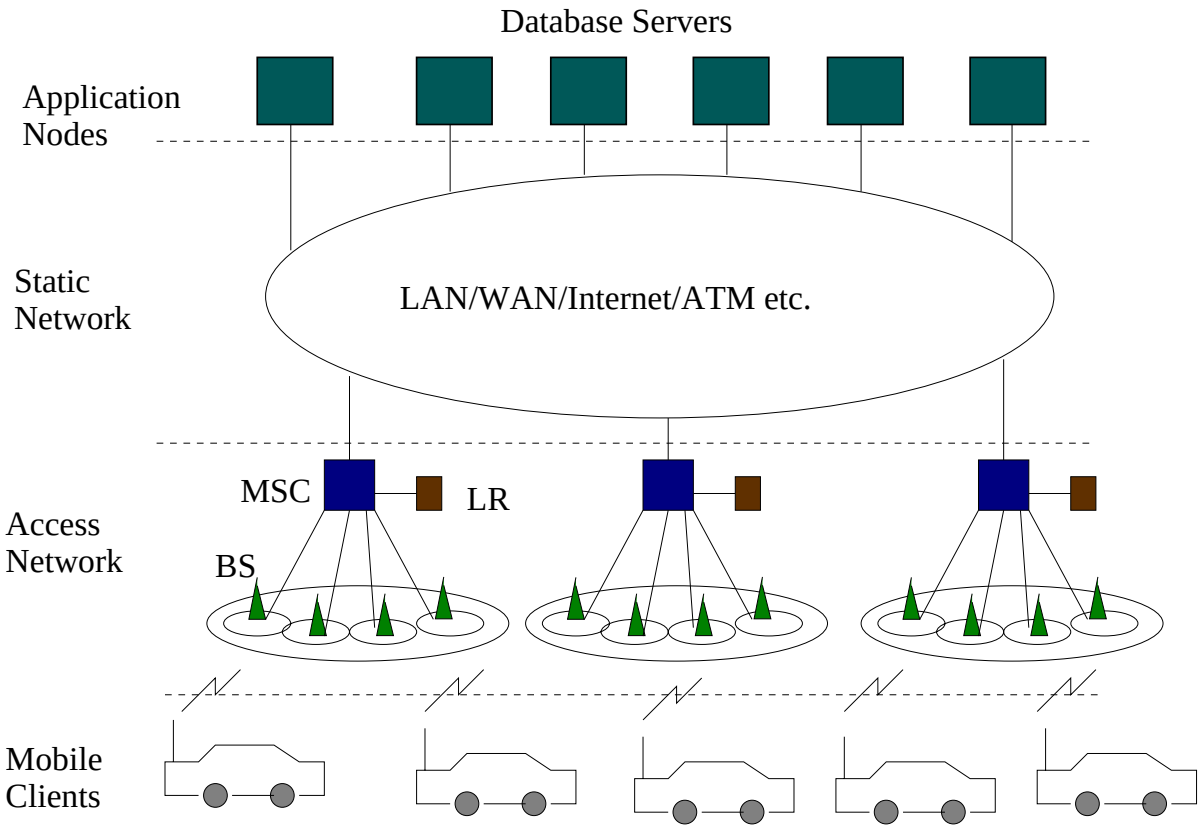
Protocol for delivering datagrams to Internet mobile users.



Data Mangement Issues in Mobile Environment

- Data management concerns with:
  1. modeling and efficient storage of information
  2. retrieval and manipulation of information
- Need for new modeling concepts and techniques for mobile environments, e.g.
  - Tracking of moving objects: new location management scheme to deal with rapidly updated location information
  - Cache maintenance schemes for mobile environment
  - Concurrency control schemes for transaction processing

Mobile Computing Environment



MSC: Mobile Switching Center
BS: Base Station
LR: Location Registrar

Database Querying in Mobile Environment

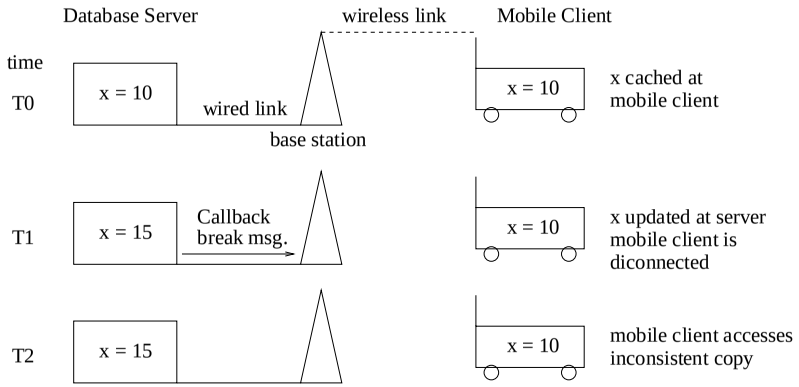
- A mobile client may query various databases for
  - location dependent information
  - time dependent information
- Requirements:
  - minimize query delay (response time)
  - maximize number of queries answered per unit time (system throughput)
  - handle client disconnection
  - conserve wireless bandwidth and battery power
  - minimize server load
  - handle mobility

Caching in Mobile Environments: Advantages

- Helps reduce latency caused by slow wireless links
- Enables limited functionality in mobile hosts even in disconnected mode
- Helps conserve battery power by reducing the number of uplink queries
- Conserves bandwidth

Caching in Mobile Environments: Problems

- Classic solutions do not work



- Need new caching schemes

Caching in Mobile Environments: Requirements

Efficient caching scheme should take into account:

- Data access pattern
- Data update rate
- Communication/access cost
- Mobility pattern of the client
- Connectivity characteristics (disconnection frequency, available bandwidth)
- Data currency requirements of the user (user expectations)
- Location-dependence of the information

General Issues in Designing Caching Scheme

- Where to cache? How many levels of caching to use?
- What to cache (when to cache a data item and how long)?
- How to invalidate cached items? Who is responsible for invalidation? What is the granularity at which the invalidation is done?
- What data currency guarantees the system can provide to the user? What are the costs involved? How to charge the user?
- What is the effect on query delay (response time) and system throughput (query completion rate)?

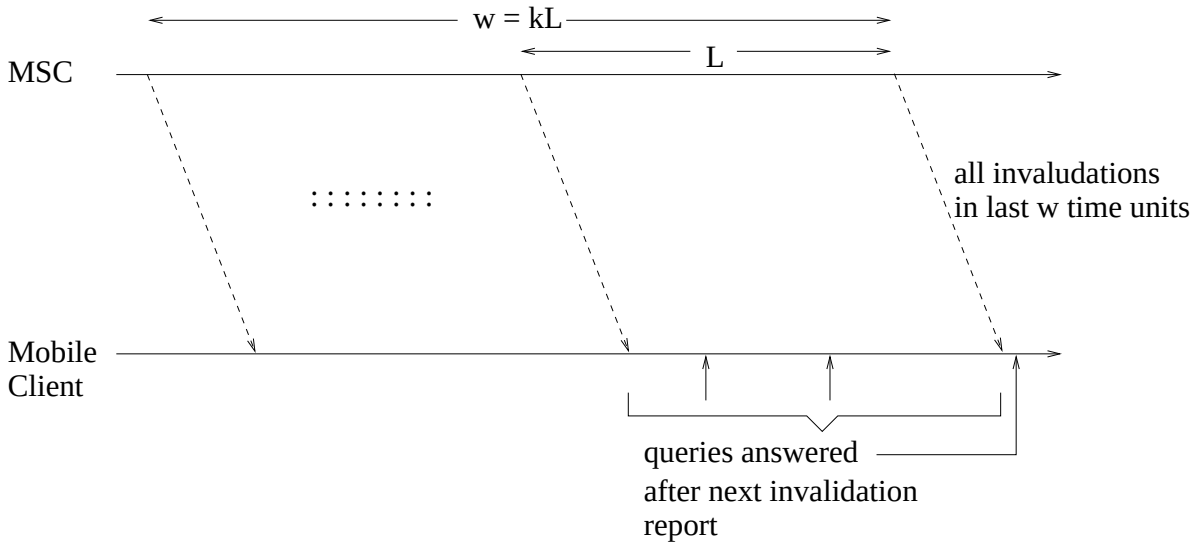
Classification of Cache Invalidation Schemes

- Who is incharge of invalidation
  - Server or Client (Push or Pull): Callback/Validation Check
- Whether or not server maintains per client state information:
  - Stateless or Stateful server
- How server sends invalidation reports:
  - Synchronous or Asynchronous
- What kind of information is sent in the invalidation report
  - State or History based
- How information is organized in invalidation reports:
  - Uncompresses or Compressed

Related Work

- Broadcasting Invalidation Reports (Barbara-Imilienski at Rutgers)
- Disconnected operation in CODA (Satanarayan et. al. at CMU)

Broadcasting Invalidation Reports



- uses *stateless* server and *synchronous* broadcast
- A query is satisfied after receiving the next invalidation report
- Each invalidation report carries all items that changed in a window of time w. *Client's entire cache is invalid if it is disconnected for more than w time units.*

Disconnected Operation in CODA File System

- goal: **CO**nstant **D**ata **A**vailability
- mechanisms: *server replication* and *disconnected operation*
- caching scheme (*asynchronous*, *stateful*):
  - Uses *callbacks* while client is reachable from a server
  - During disconnection permits access to possibly stale data
  - Upon reconnection, the client does validity check on each volume cached.
- uses *prioritized* cache replacement scheme.

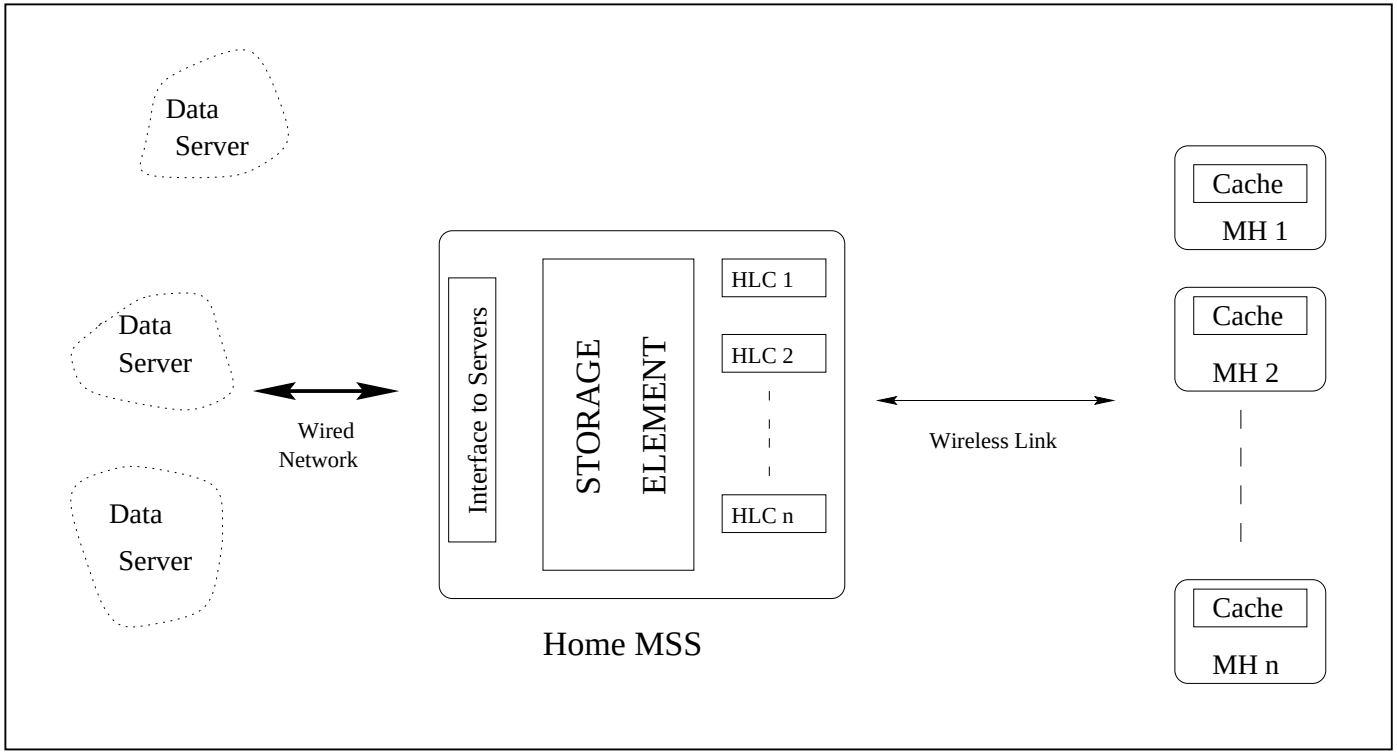
Motivation for the Proposed Scheme

- Drawbacks of Barbara-Imielinski' scheme
  - poor delay characteristics due to the waiting involved before answering a query.
  - poor network utilization characteristics due to the answering of queries in bursts.
  - does not support arbitrary sleep duration.
- Drawbacks of CODA caching scheme:
  - Server has to keep cache state of each client
  - Client has to perform volume-by-volume validation check after each reconnection

Assumptions

- Wired links are considerably faster than the wireless links
  - For all practical purposes we assume data is present on every MSS
- *Home Agent*(HA): Each mobile client has a node on the static network, which maintains it's state information in the form of a *Home Location Cache*(HLC).
- Handoff protocol ensures proper forwarding of messages
- Messages are delivered in FIFO order

System Architecture



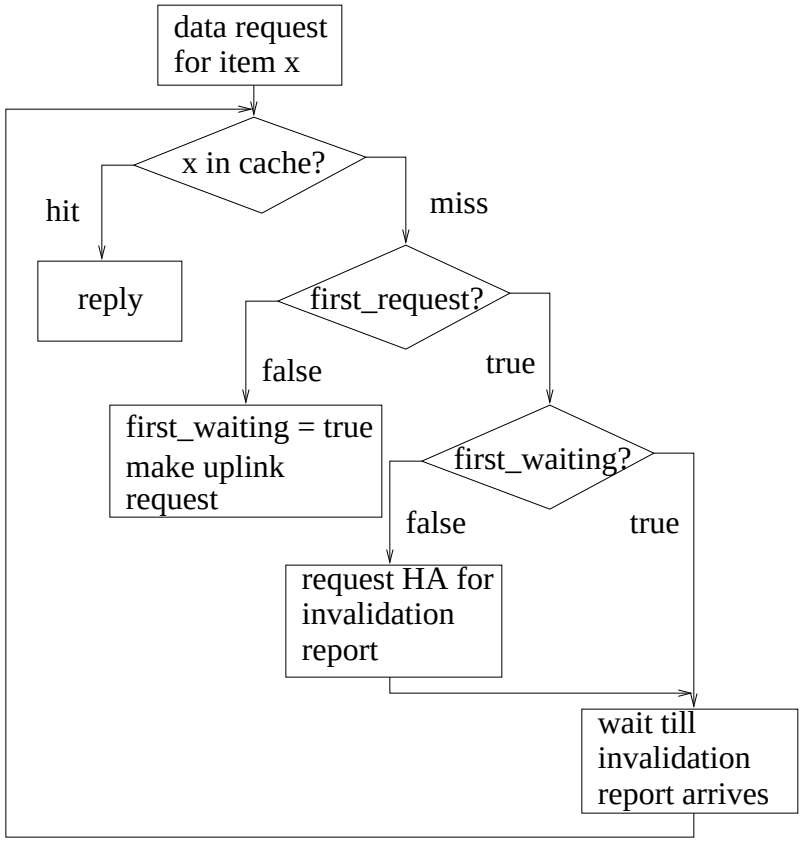
Salient Features

- Asynchronous transmission of invalidation reports
- Maintaining the state information of the data cached by each MH.
- Support for arbitrary sleep by maintaining the timestamp of the last invalidation report destined for an MH at it's HLC.
- Transparently supports mobility by assuming an underlying mobility management scheme eg. *mobile IP*.

Actions at Mobile Agent (MA)

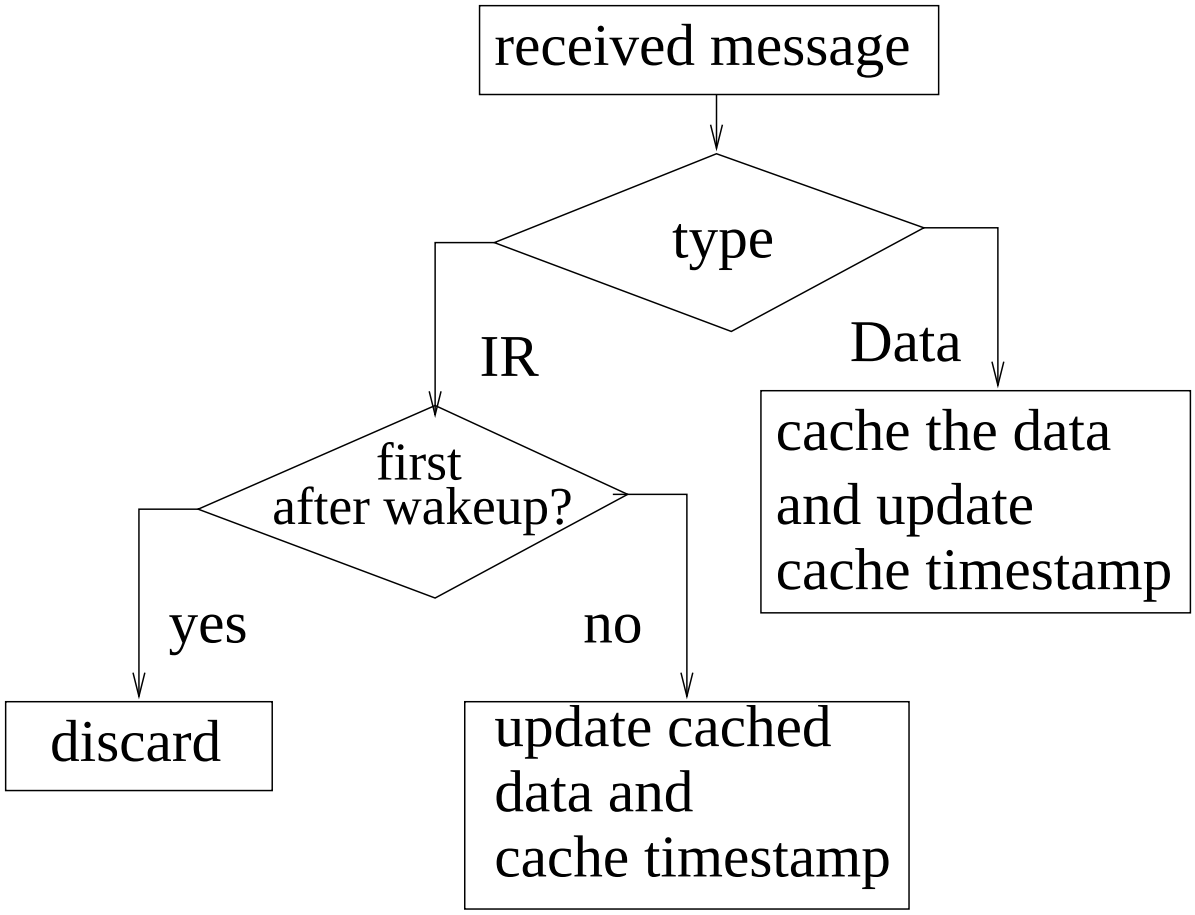
- **MA receives a request for data from a mobile client**
  - MA sends the data and deletes entries in the HLC of invalidations already received by the mobile as indicated by the request timestamp.
  - If the data request is the first after a sleep, the HA sends an invalidation report consisting of all the items indicated as *changed* in the HLC.
- **Some DATA changes on the servers**
  HA sends an invalidation report to all MHs currently caching the data and adds the current time to their HLCs.

Satisfying Query at a Client

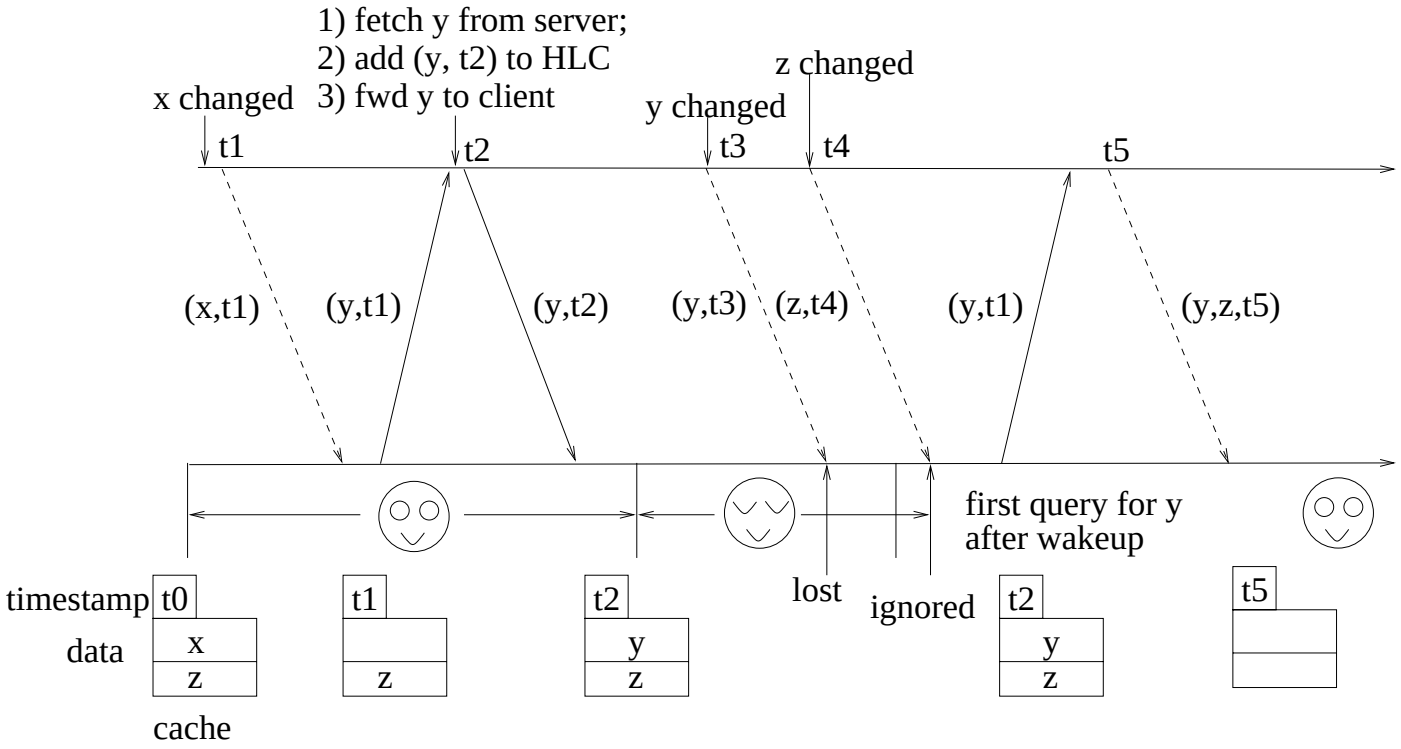


- *first_request* is set to true on reconnection.

Handling of Messages at Client



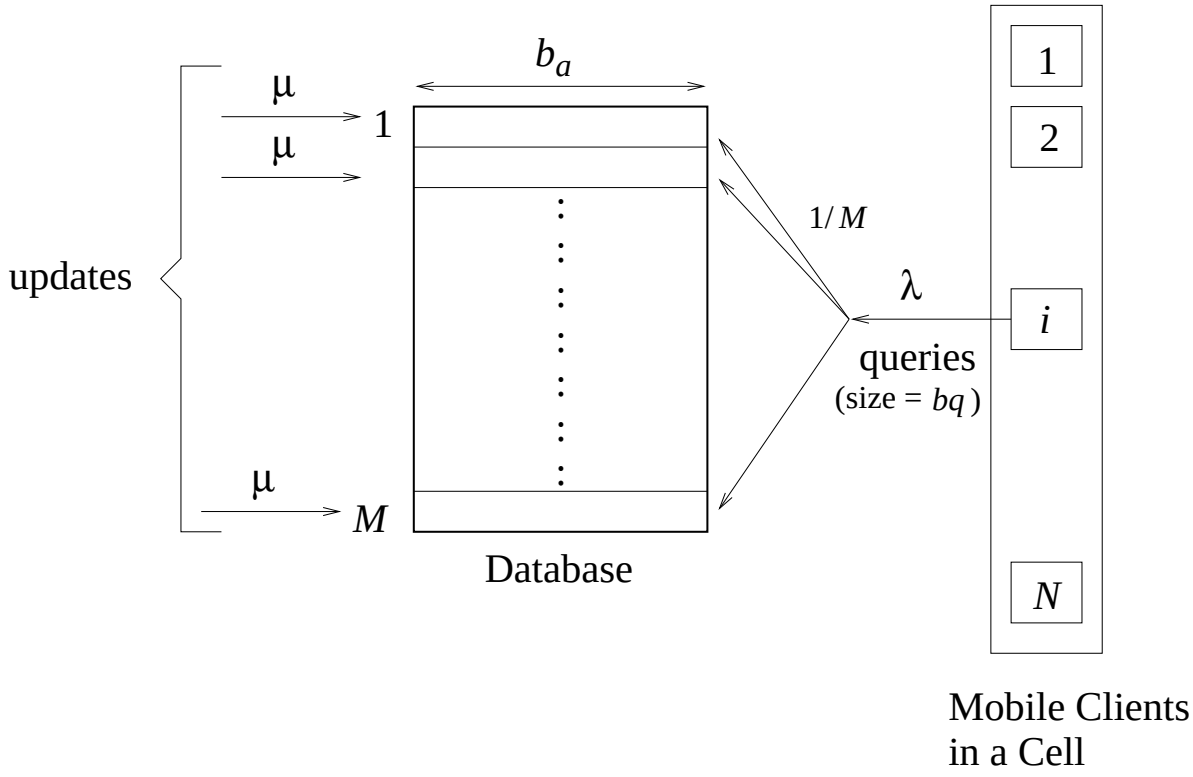
An Example Scenario



Performance Analysis

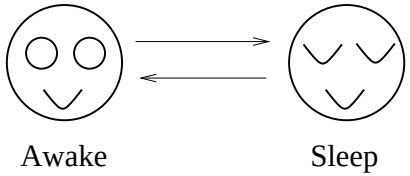
- Query-Update Model
- Client Sleep Model
- Some assumptions
- Estimation of miss probability
- Estimation of mean query delay

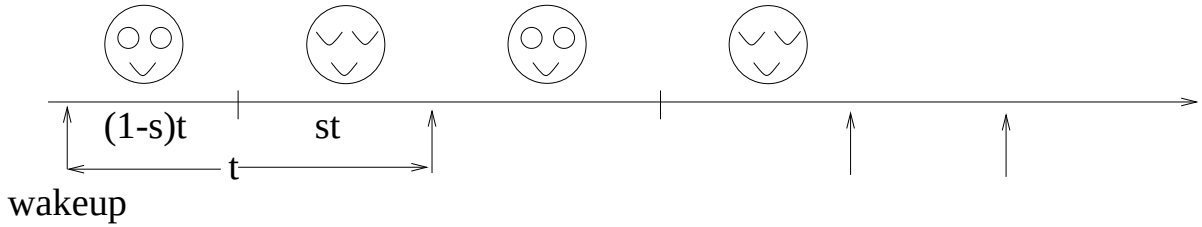
Query-Update Model



Client Sleep Model

- Sleep = The client is unreachable e.g. due to link failure
- Awake = The client can receive messages (includes both active and doze CPU modes)





- s is the fraction of time a client sleeps.
- Inter-wakeup time: exponential distribution with mean $1/\omega$.

Some Assumptions

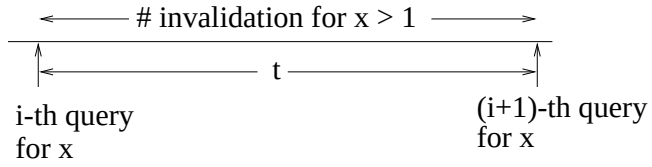
- A single wireless channel of bandwidth C; shared by all the mobile clients and base station.
- Queries during sleep are lost i.e.

$$\text{Effective query rate} : \lambda_e = (1 - s)\lambda.$$

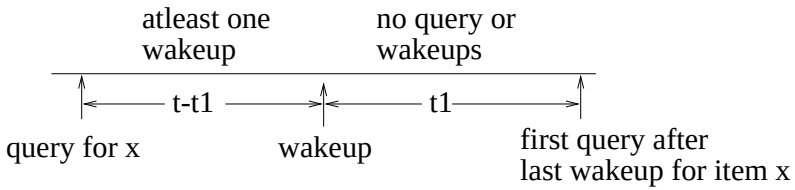
- Each invalidation is of size b_i bits.
- Delays in the wired network are ignored.
- Any processing overheads are ignored.

Estimating Miss Probability

- Scenarios for miss (need for uplink request):



(a) Miss due to absence of valid data item in cache



(b) Miss due to diconnections.

- $P_{miss} = P[\text{Event a}] + P[\text{Event b}]$.

Estimating Miss Probability (Cont.)

- query rate for a fixed data item $= \lambda_x = \lambda_e / M$.
- Probability of miss due to absence of valid data item in cache:

$$P[\text{Event a}] = \int_0^\infty (\lambda_x e^{-\lambda_x t})(1 - e^{-\mu t}) dt$$

$$= \frac{\mu}{\lambda_x + \mu} = \frac{M\mu}{(1-s)\lambda + M\mu}.$$

- Probability of miss due to disconnection:

$$P[\text{Event b}] = \int_0^\infty (P[\text{no IVs and Qs for x during time t}]$$

$$\times P[\text{the query (for x) is 1st after wakeup}) dt$$

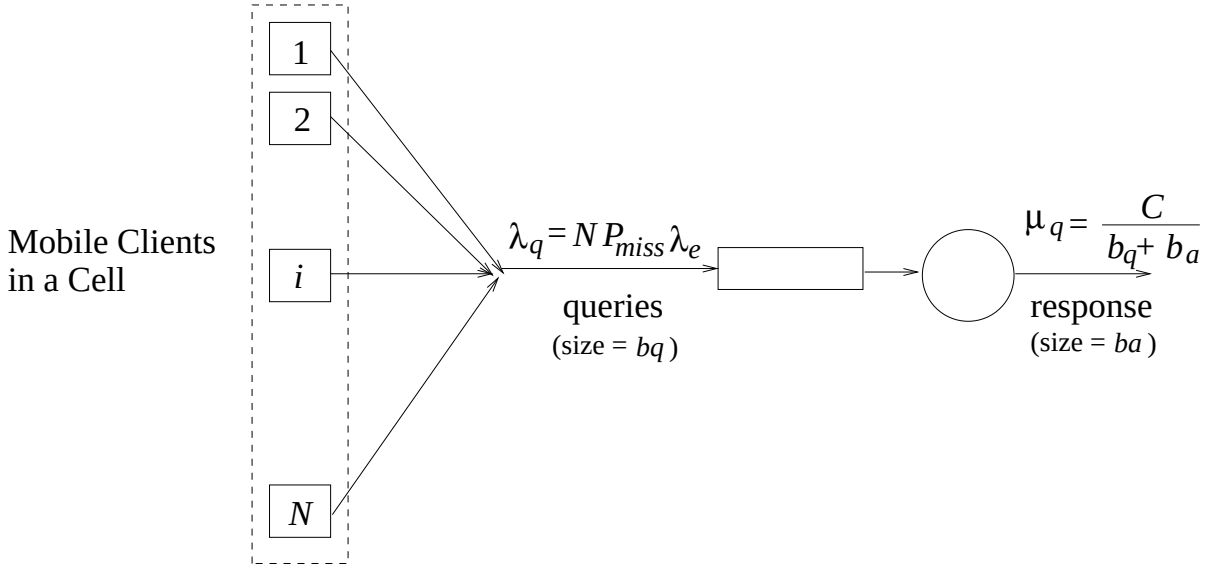
$$= \int_0^\infty \lambda_x e^{-\lambda_x t} e^{-\mu t} (\int_0^t e^{-\lambda_e t_1} \omega e^{-\omega t_1} (1 - e^{-\omega(t-t_1)}) dt_1) dt$$

Estimation of Delay

- Estimation of arrival and service rate of queries requiring uplinks
- Estimation of arrival and service rate of invalidations
- Estimating the average delay T_q for a query which requires uplink
- Average query processing delay $= P_{miss} T_q$.

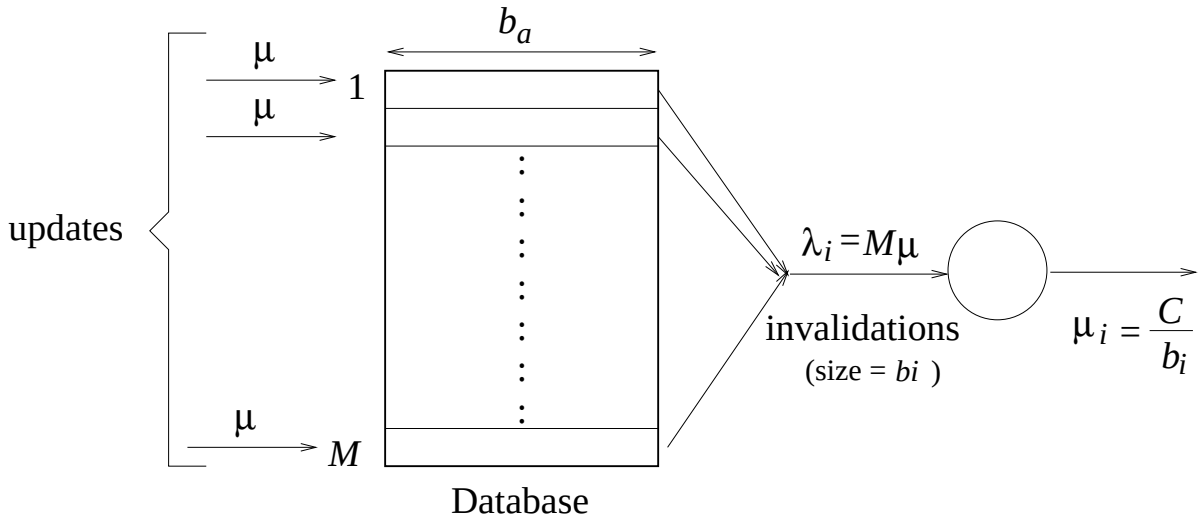
Arrival and service rate of queries experiencing cache miss

- λ_q: average arrival rate of queries requiring uplinks
- μ_q: service rate of queries requiring uplinks (deterministic)



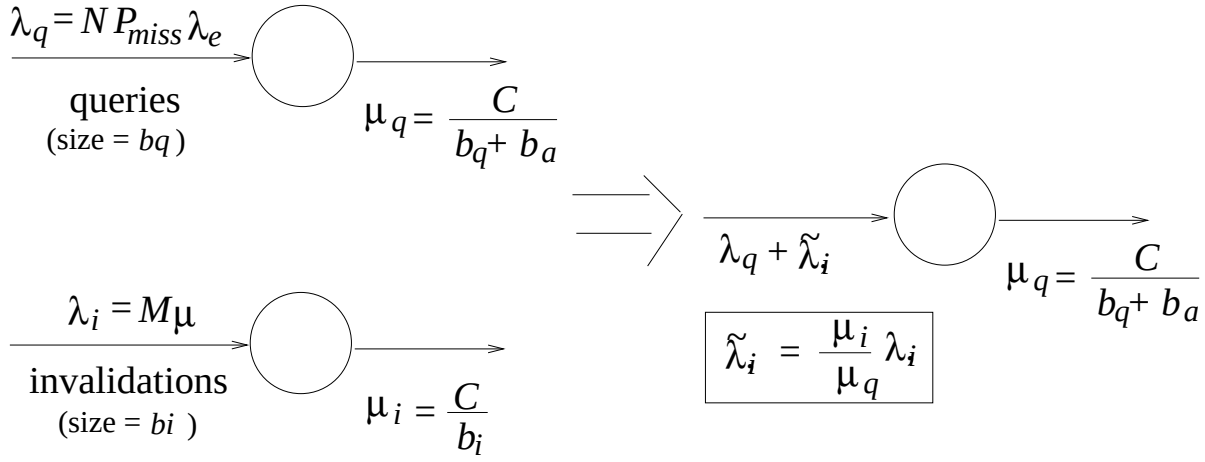
Arrival and service rate of invalidations

- λ_i: average arrival rate of invalidations
- μ_i: service rate of invalidations (deterministic)



Modeling the Effect of A Single Wireless Channel

- Combine the previous two M/D/1 queues into a single M/D/1 queue
- $\tilde{\lambda}_i$: Adjusted arrival rate of invalidations

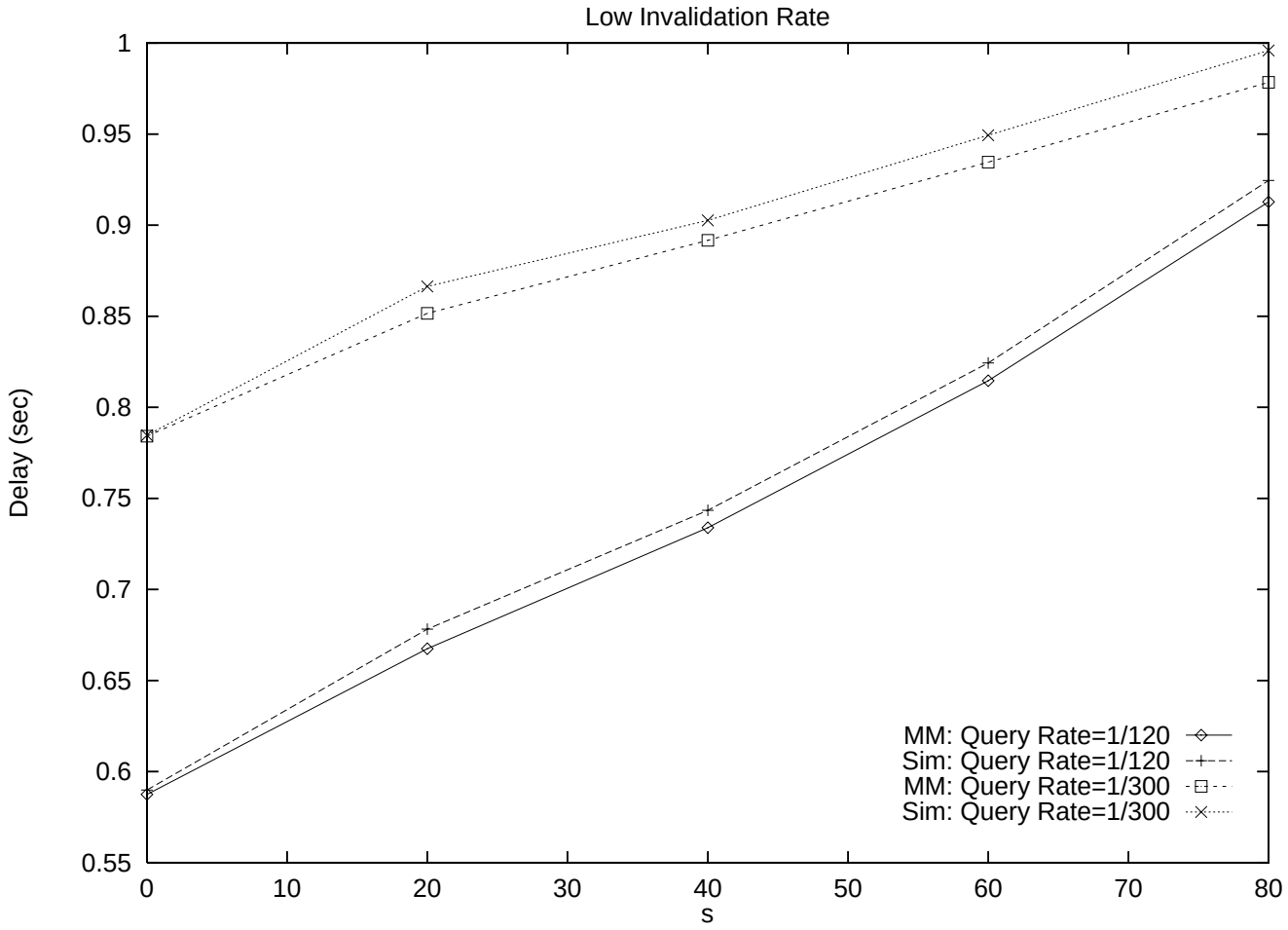


Simulation

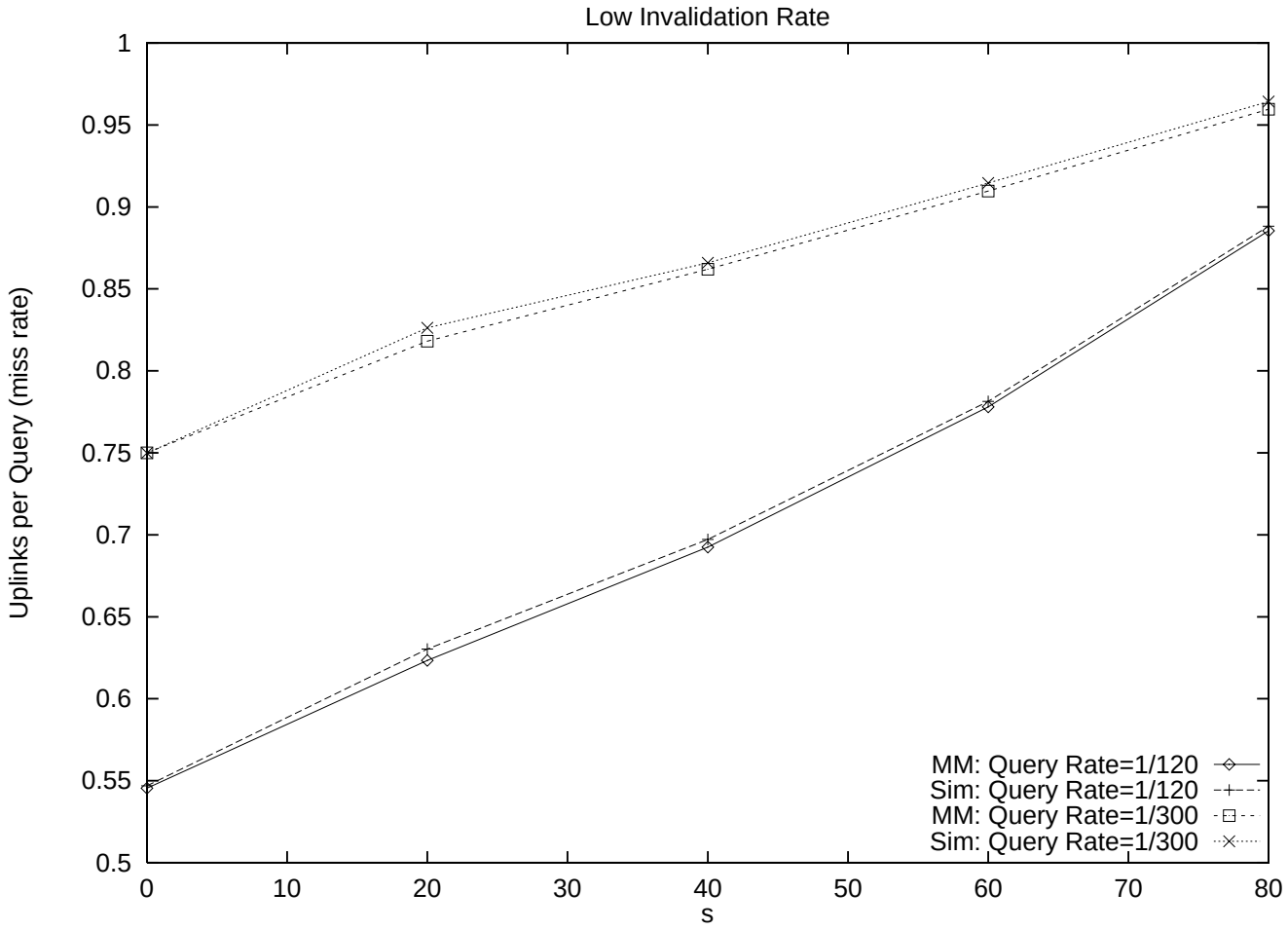
- Simulated using SES/Workbench.
- Simulated both the proposed scheme and Barbara and Imilienski's scheme.
- Comparison with analytical results.
- Default Parameter:

Parameter	Value
N	25
M	100
λ	1/120 query/s
μ	10^{-4} updates/s
b_a	1200 bytes
b_q	64 bytes
b_i	64 bytes
W	10000 bits/sec
s	20%
ω	1800 sec
L	10 sec.
w (TS)	100×L

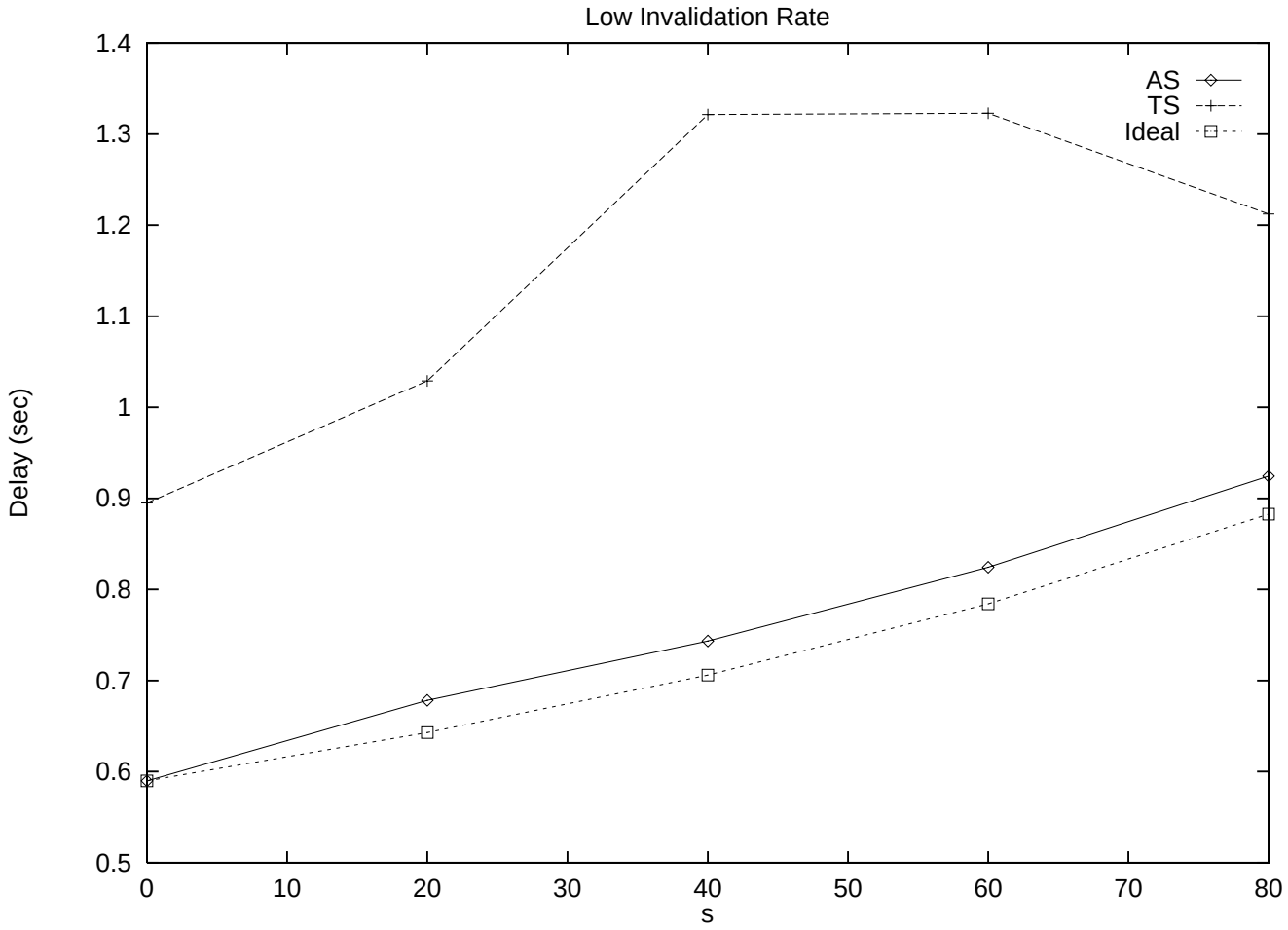
Query Delay vs Sleep



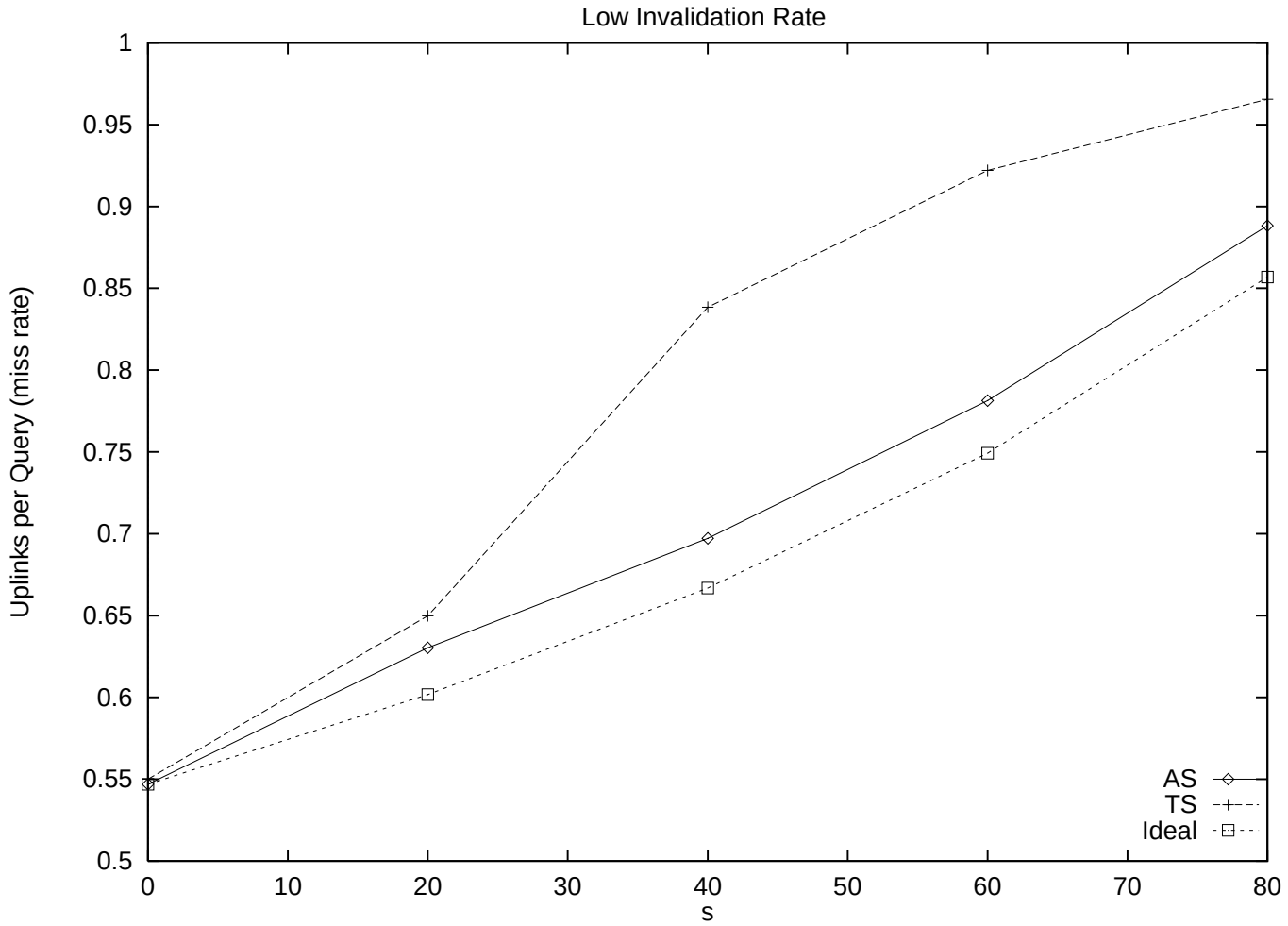
Number of Uplinks vs Sleep



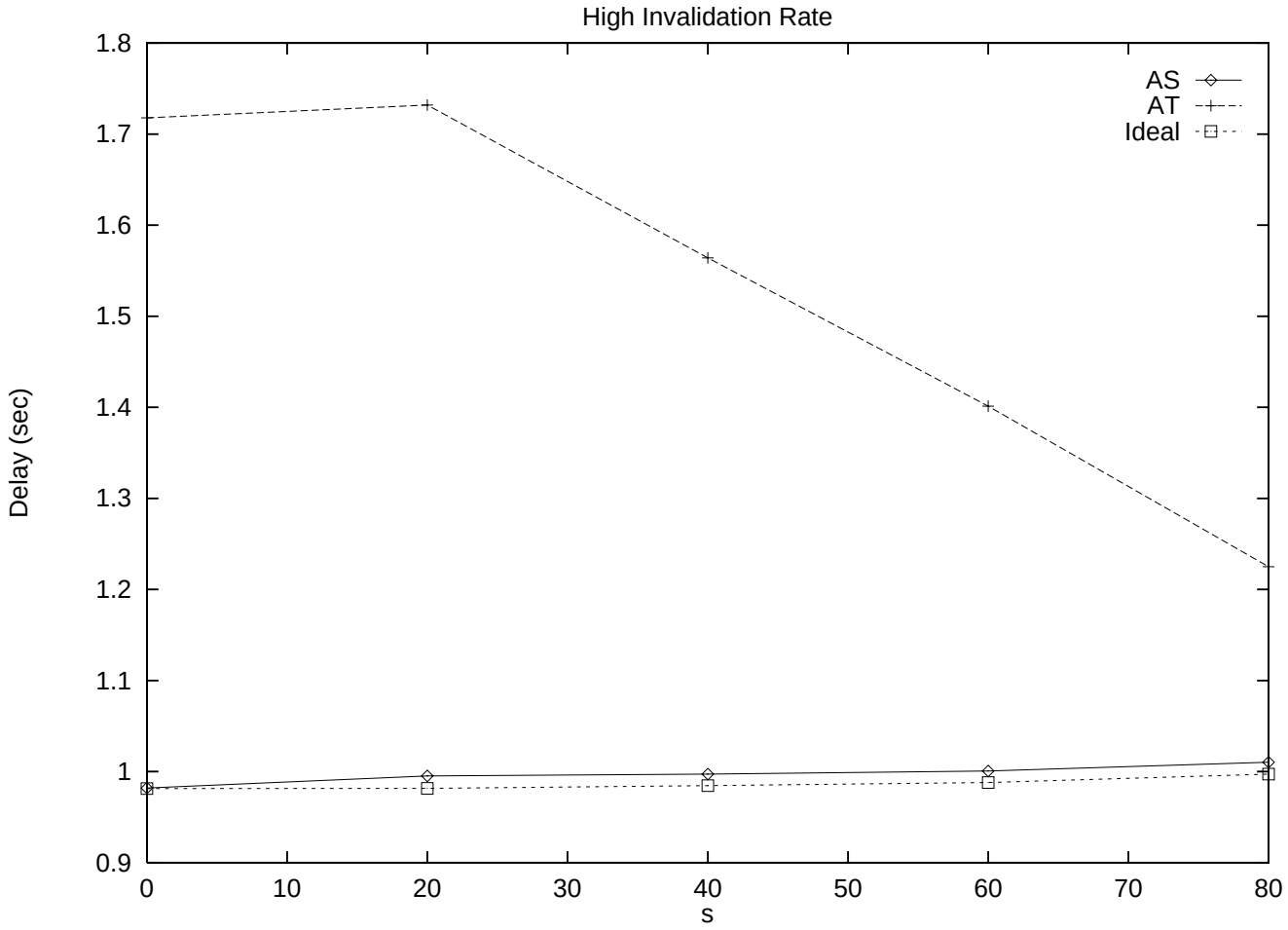
Comparison: Delay vs Sleep



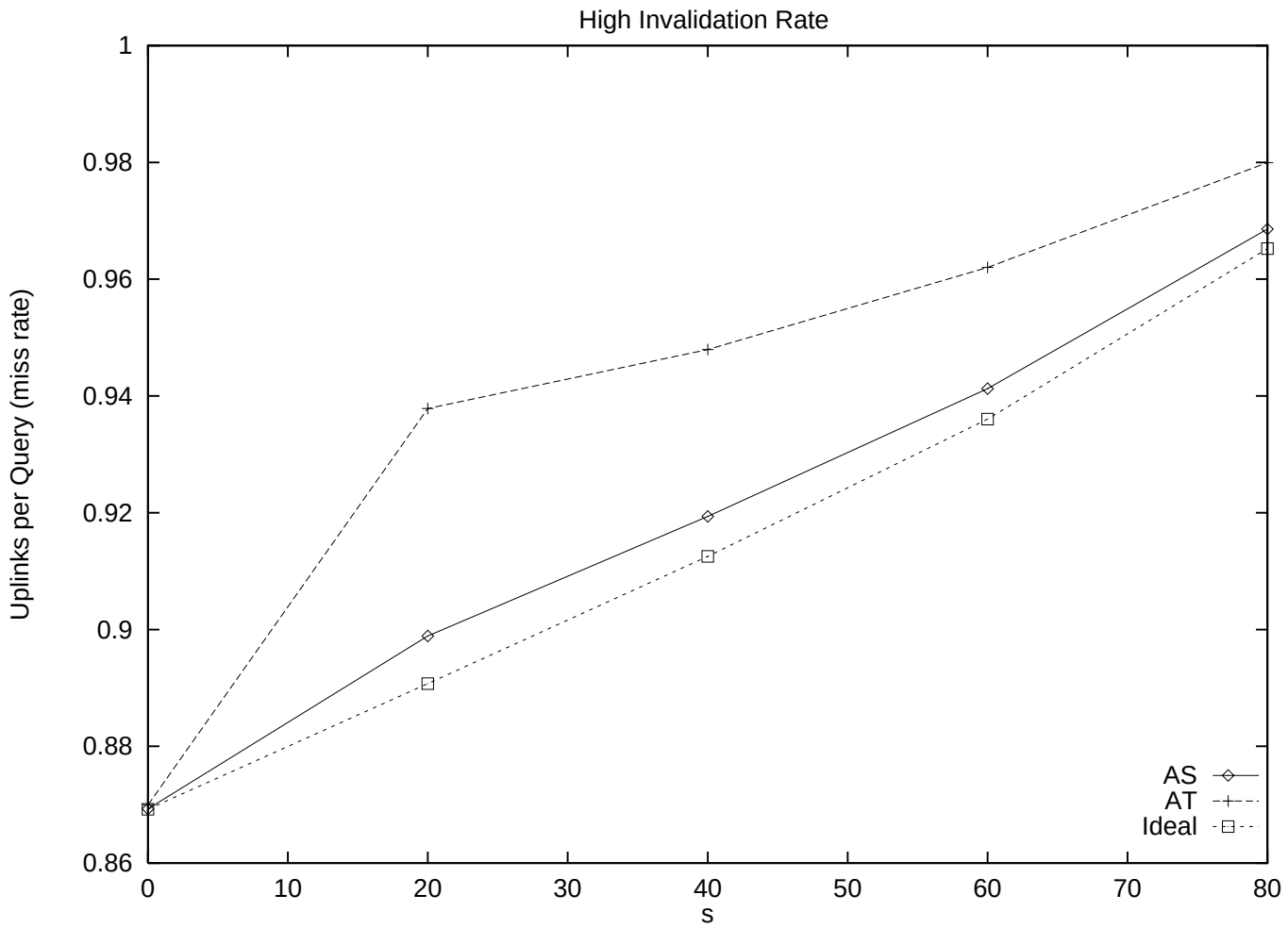
Comparison: Uplinks vs Sleep



Comparison: Dealy vs Sleep



Comparison: Uplinks vs Sleep



Conclusions

- The caching strategy is very close to an ideal scheme with regards to the hit-rate and delay.
- Simulation study along with the mathematical modeling show that the scheme
  - saves on uplinks and hence conserves battery power
  - provides support for arbitrary disconnections with minimal adverse effects.
- Drawbacks of the scheme are
  - Uses more buffer space at the Home Agent
  - Communication via Home Agent