# A Scalable Architecture for Cooperative Web Caching

Riccardo Lancellotti
Università di Roma
Tor Vergata

Bruno Ciciani
Università di Roma
La Sapienza

Michele Colajanni
Università di Modena e
Reggio Emilia

# Outline

Cooperative caching

Two-tier architectures

Prototype implementation

Experimental results

# Cooperative Web Caching

Cooperative lookup

Many proposed solutions:

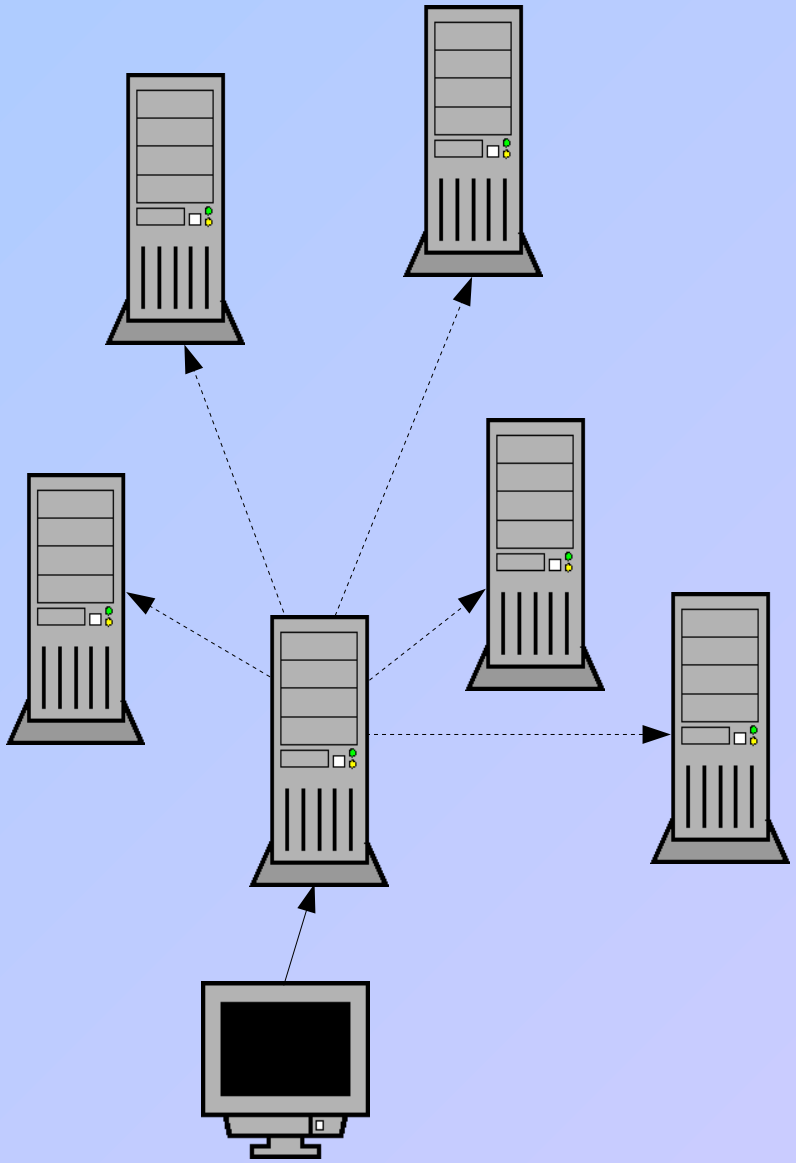
- Hierarchical architectures
- Query protocols (ICP)
- Informed protocols (CD)
- Hybrid architectures (CRISP)

Scalability issues

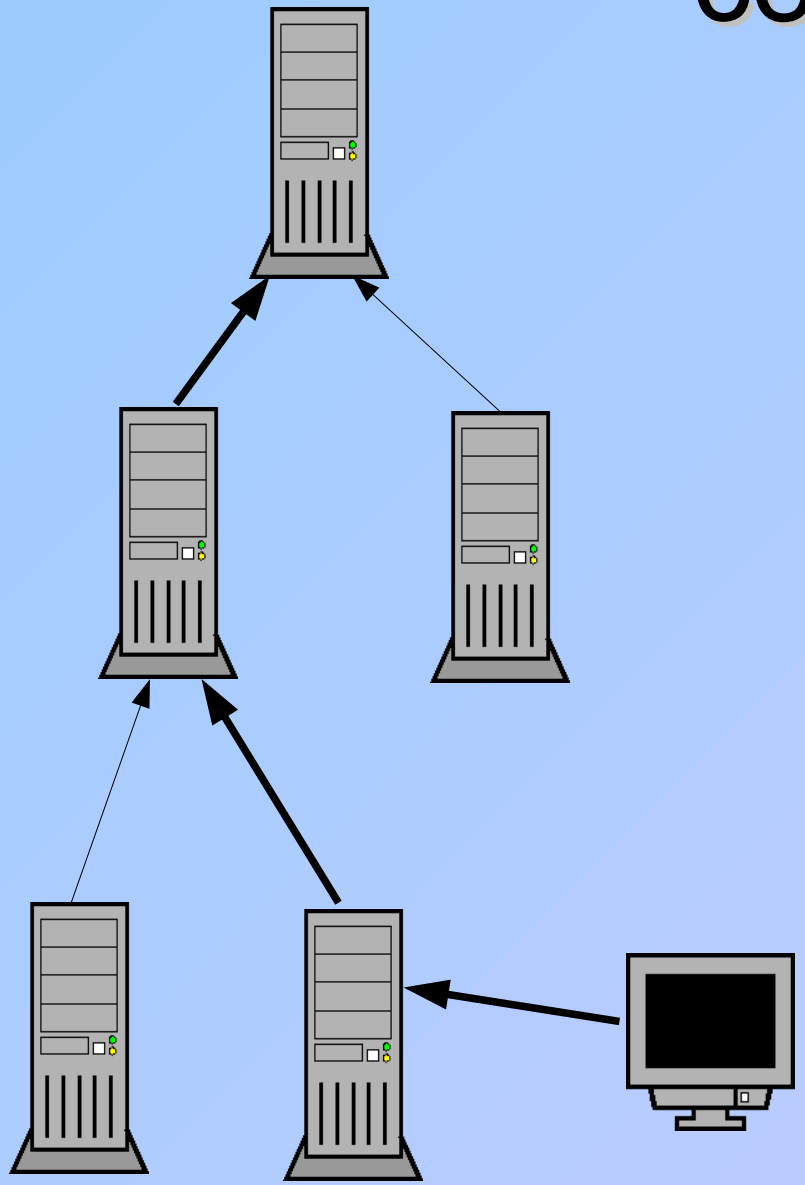
# Example: query cooperation

Network overload (load=f(n²))

Increased risk of packet loss leads to higher response time



# Example 2: hierarchical cooperation



High latency time in case of top level hit

Does not use possible hit on same level cache servers (siblings)

Top level cache servers can became bottleneck

# Two-tier Architectures

Organization of cache servers in clusters

- Physical clustering based on networking characteristics
- Logical clustering for cooperative lookup

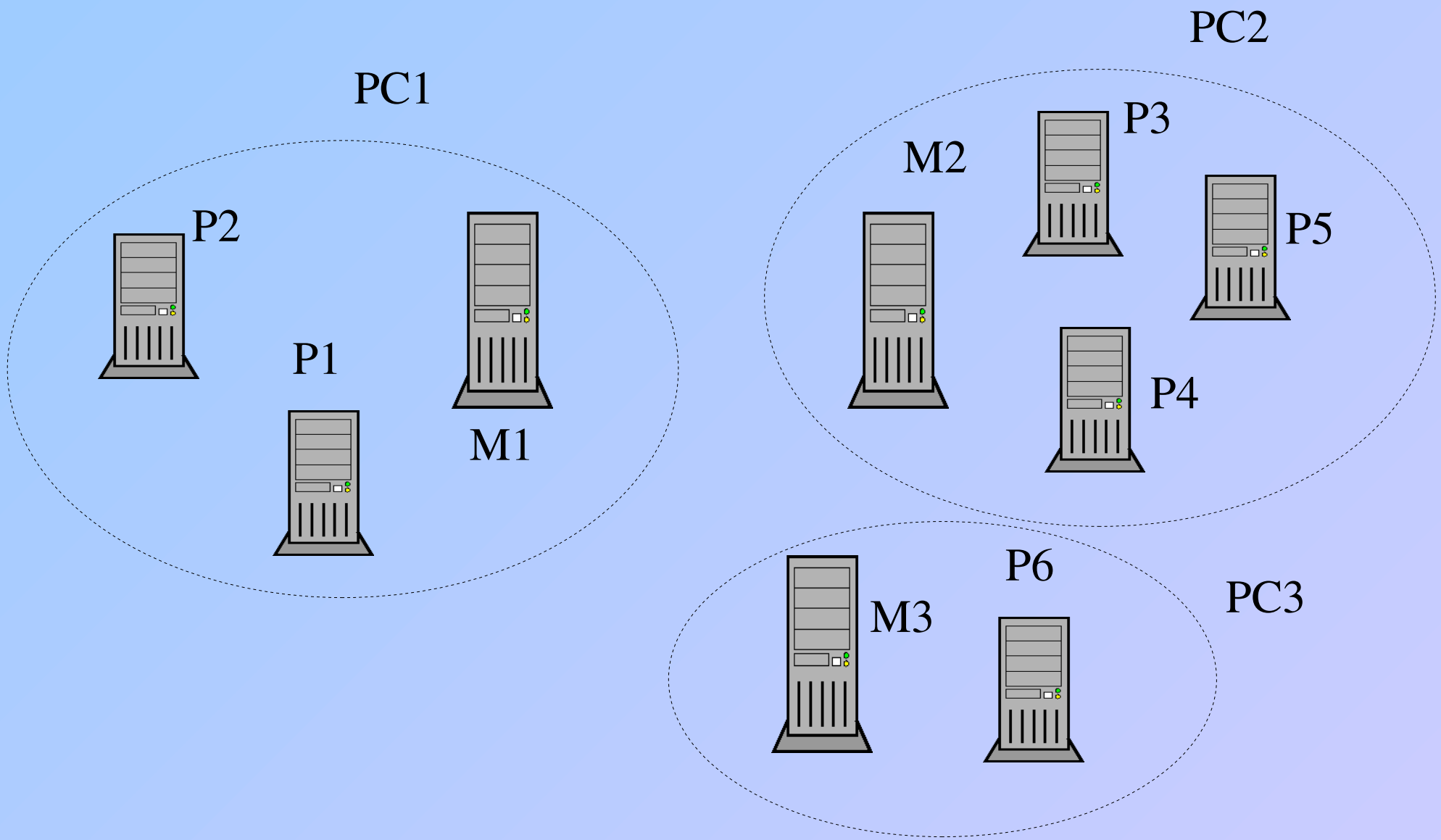
Motivation:

- Hit Rate > informed protocol
- Overhead < query protocol

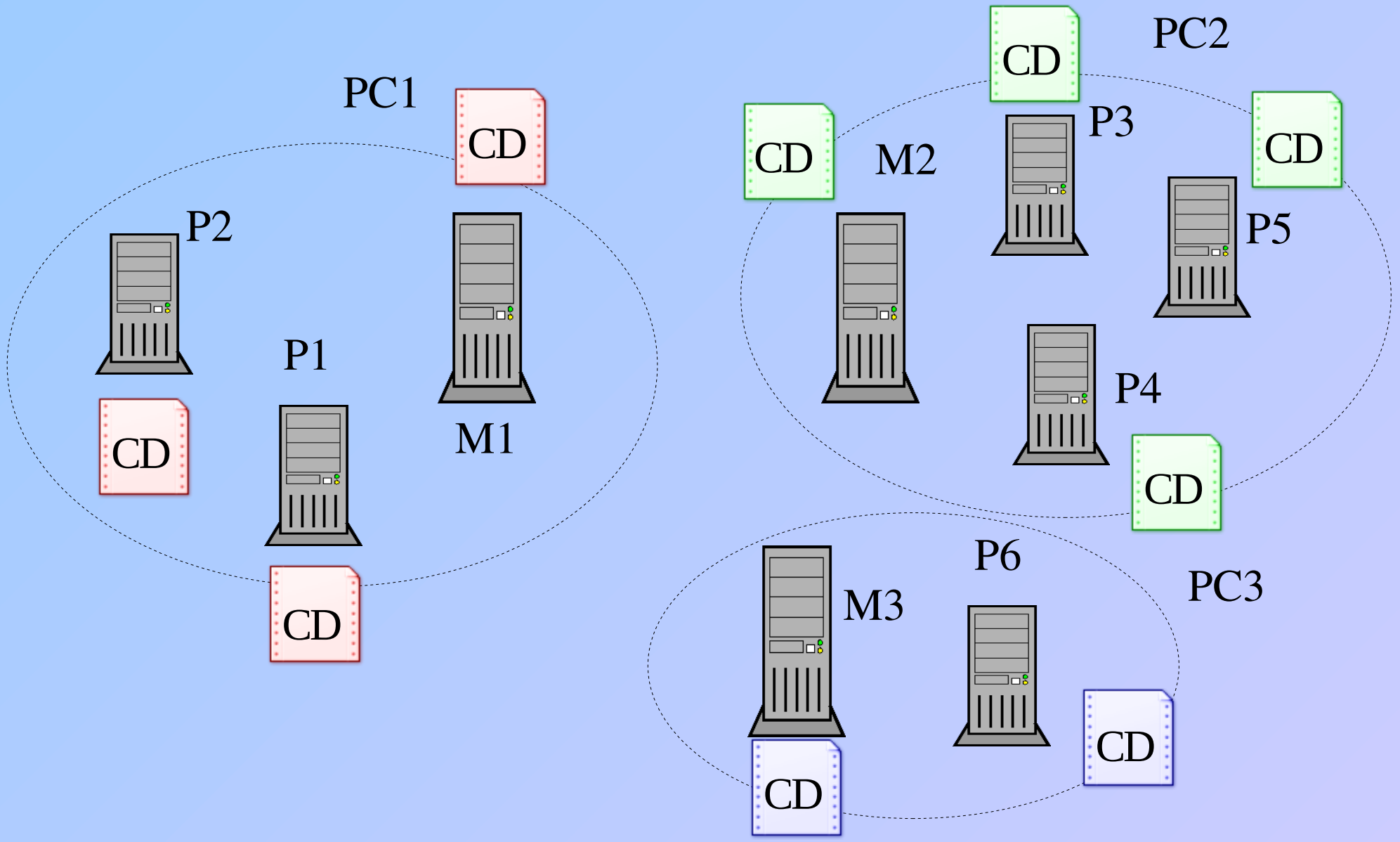
Proposals:

- 2TC with hierarchical logical clustering (article)
- 2TC with flat logical clustering

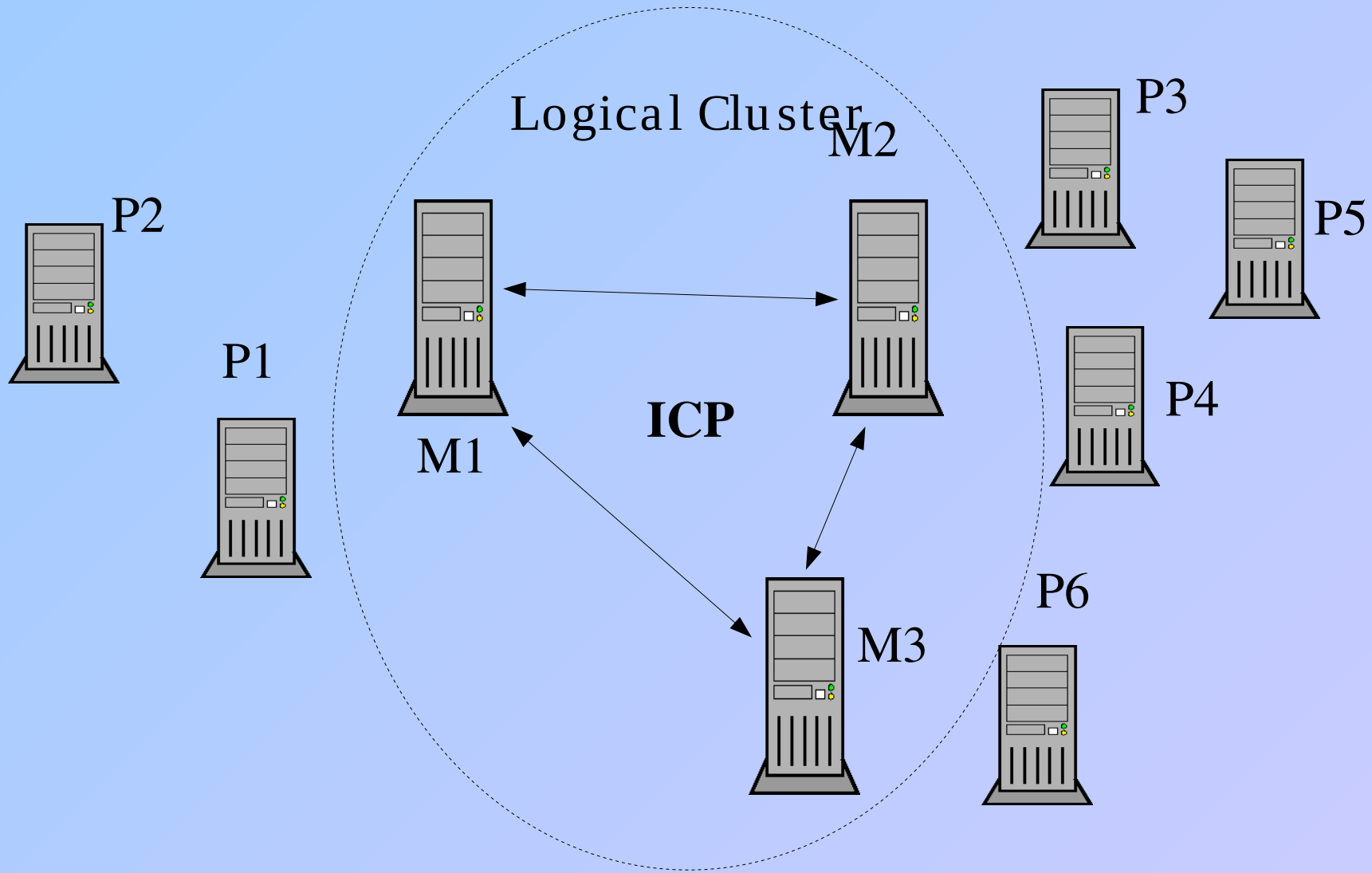
# 2TC with Hierarchical Logical Clustering (2TC_HC)



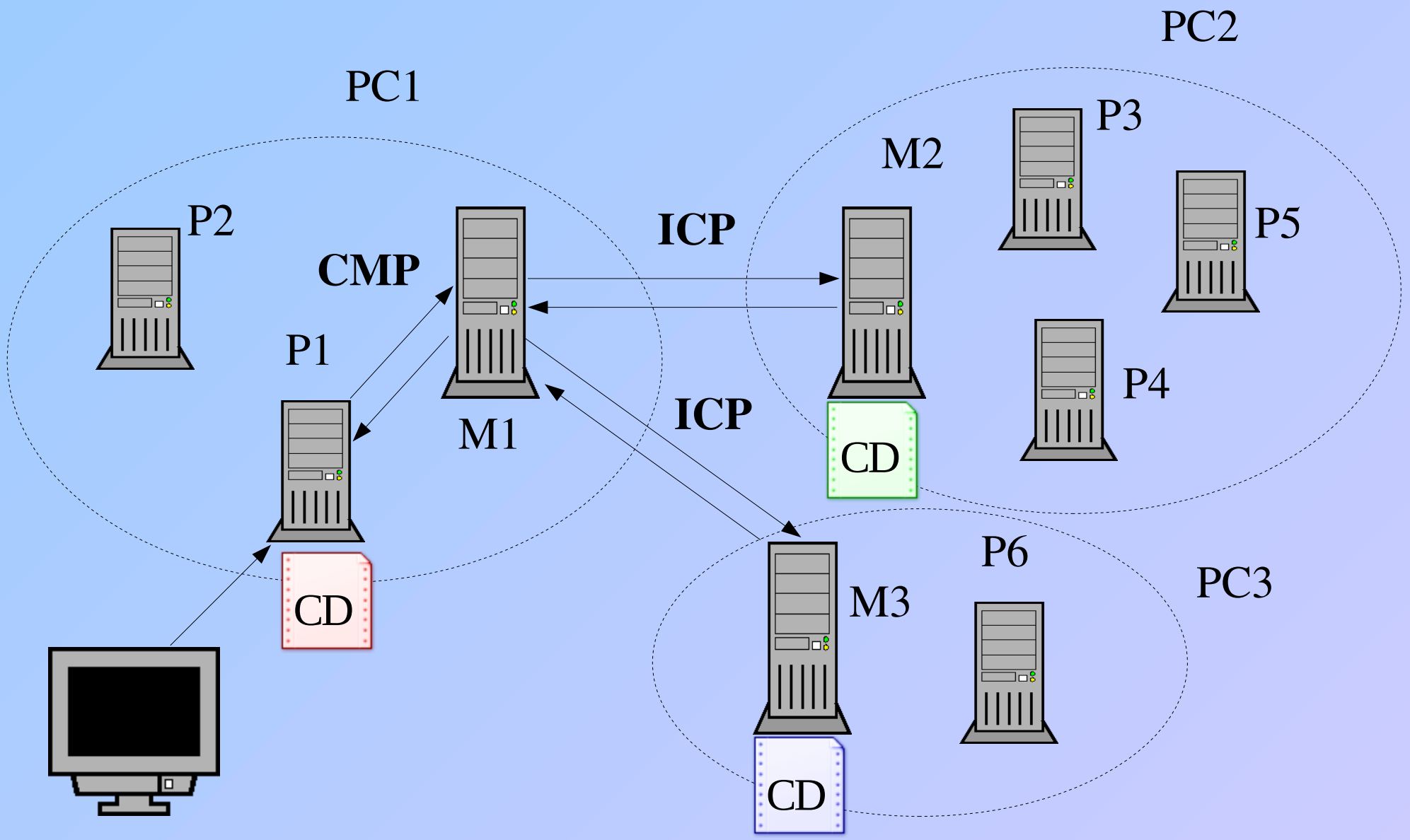
# 2TC with Hierarchical Logical Clustering (2TC_HC)



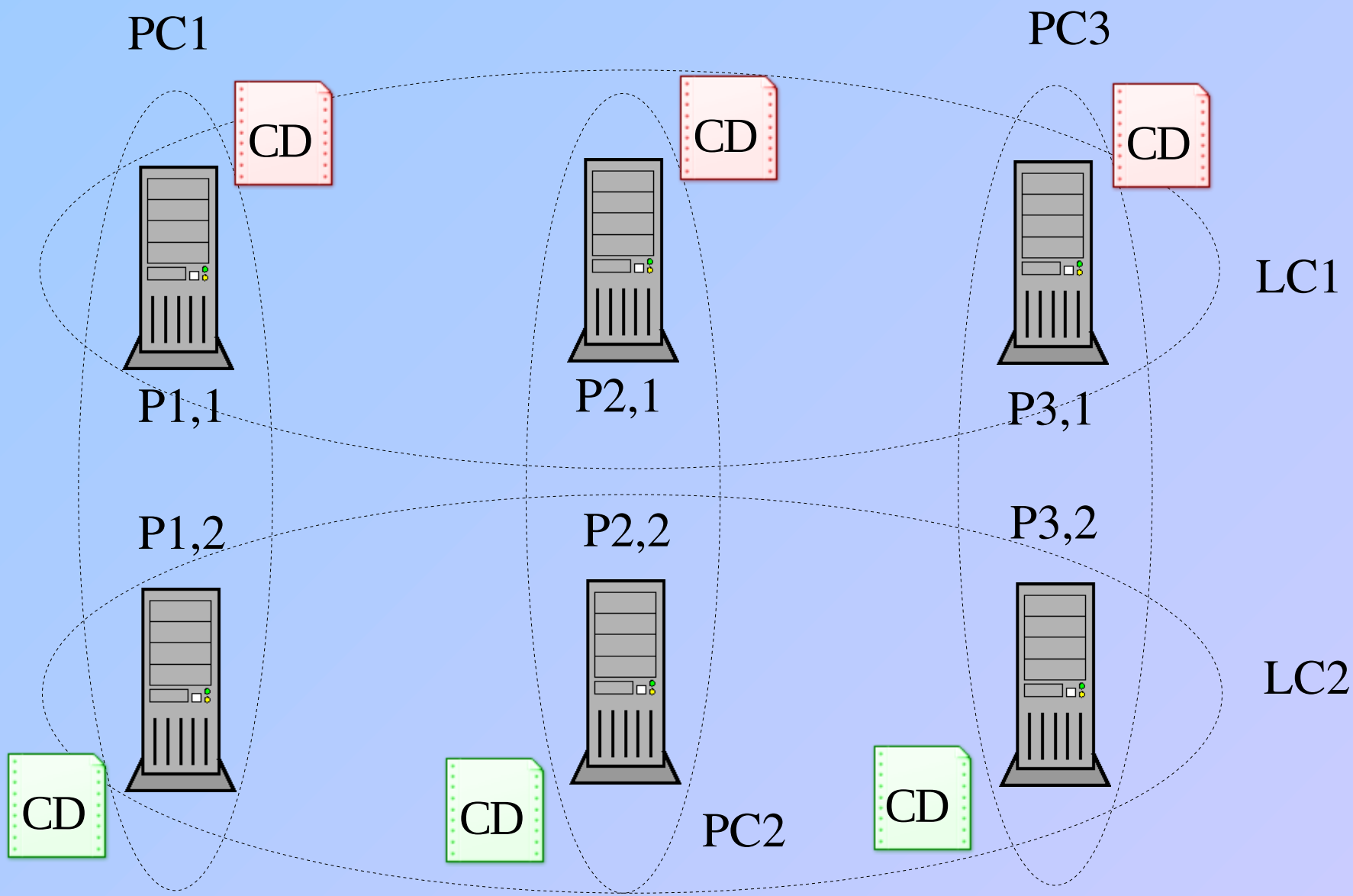
# 2TC with Hierarchical Logical Clustering (2TC_HC)



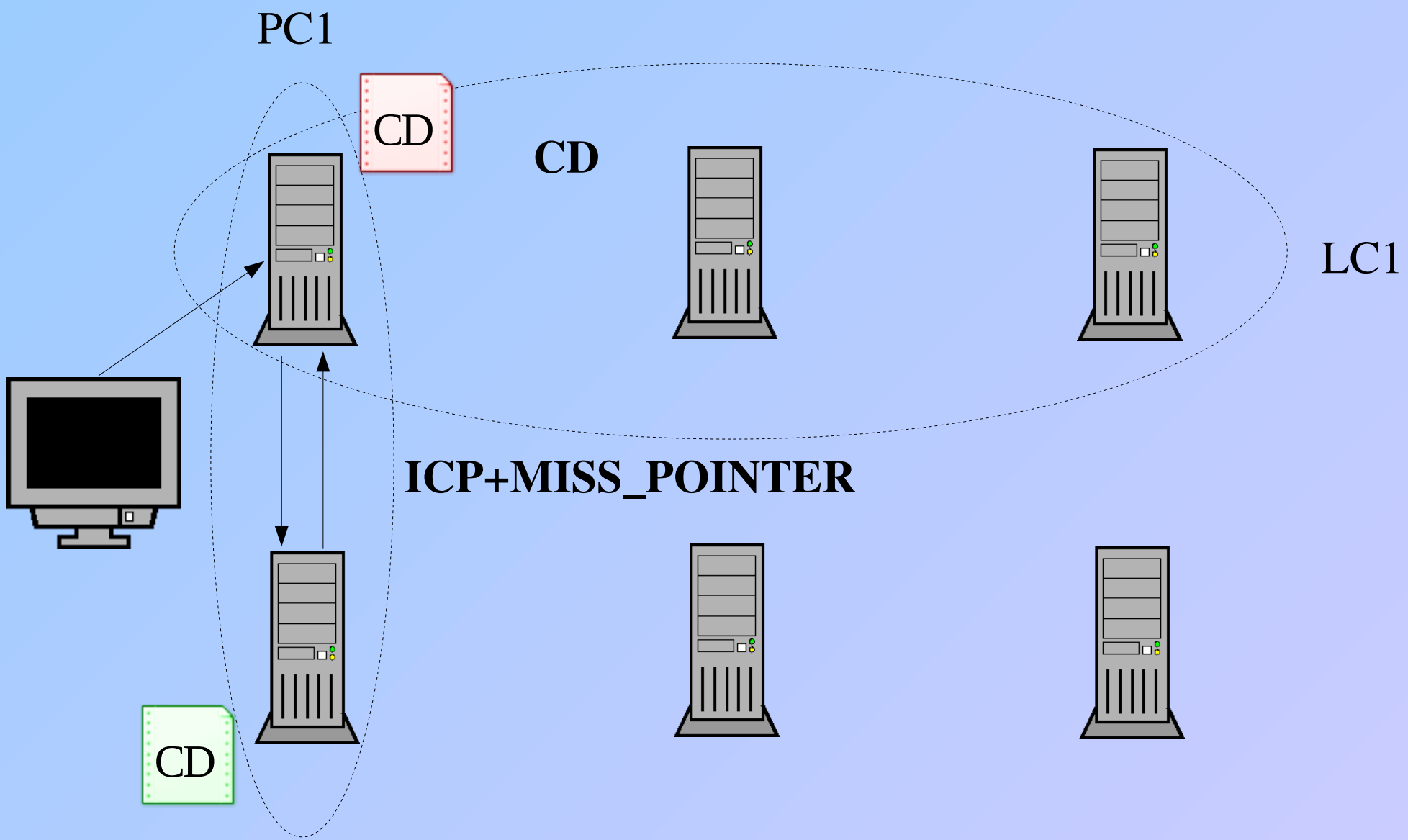
# 2TC with Hierarchical Logical Clustering (2TC_HC)



# 2TC with Flat Logical Clustering (2TC_FC)



# 2TC with Flat Logical Clustering (2TC_FC)



# Implementation

Two working prototypes based on Squid 2.4

2TC_HC

- Modified peer selection algorithm
- Added CMP support

2TC_FC

- Modified peer selection algorithm
- Added ICP_MISS_POINTER support in ICP module

# Experimental Results

Web Polygraph as a workload generator

- Workload from IRCACHE 2nd Cacheoff
- Data collected from Squid logs

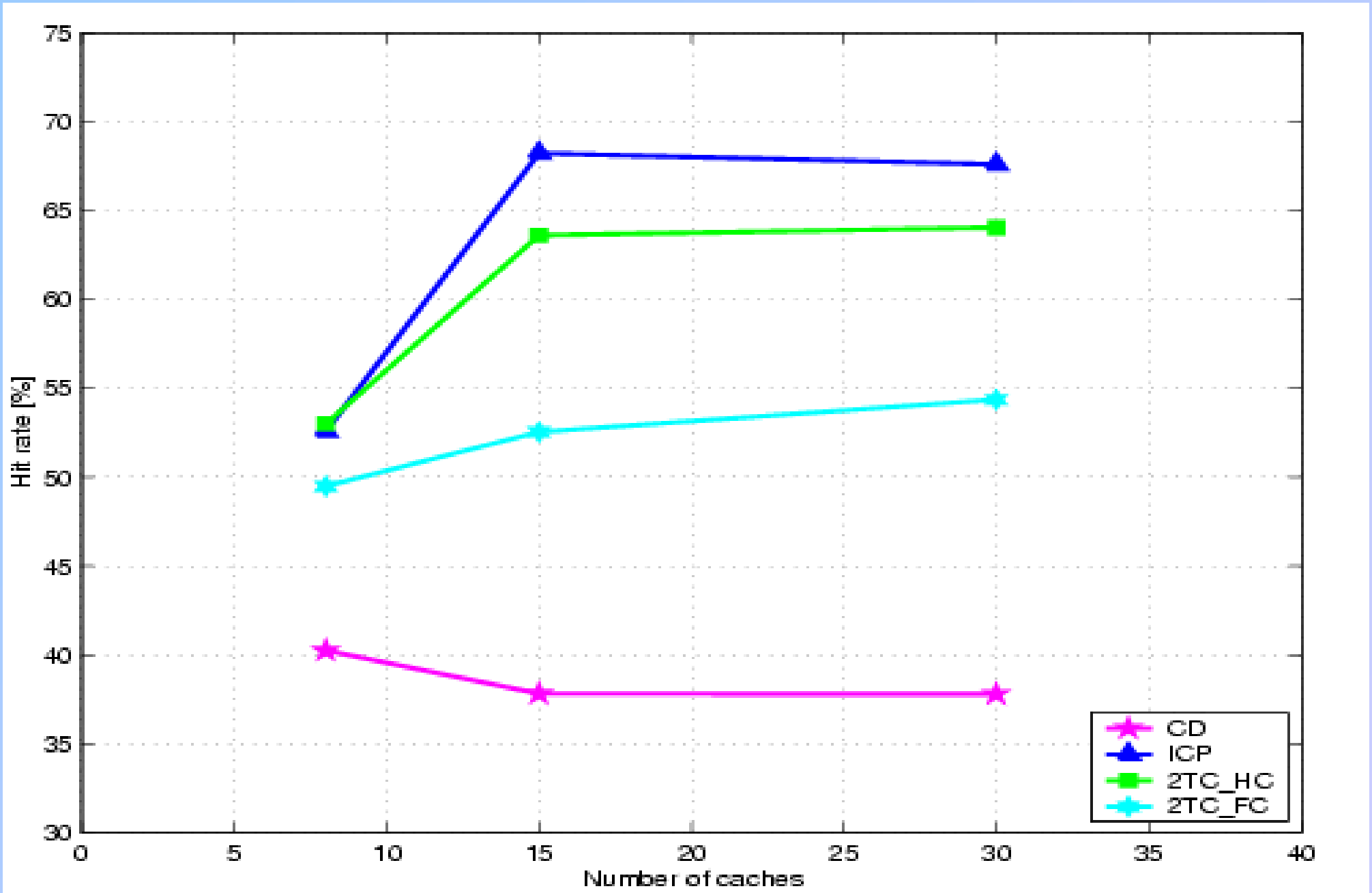
Local testbed

- Clusters of 8, 15, 30 cache servers, 1 web server

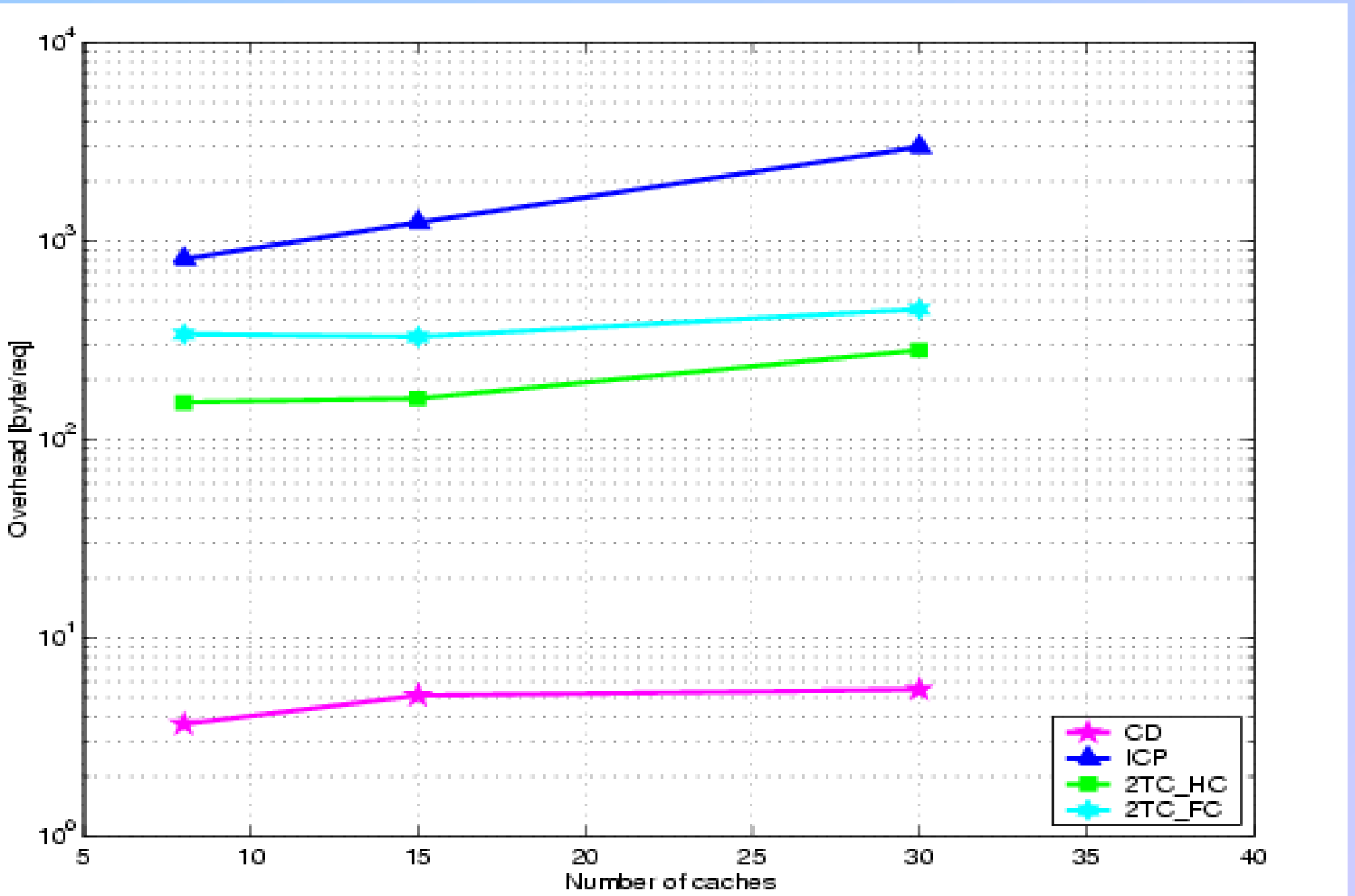
Geographic testbed

- GARR Network
- Sunday (reduced network load)
- 5 cache servers in Modena, 5 in Rome, 2 web servers

# Scalability Test (HR)



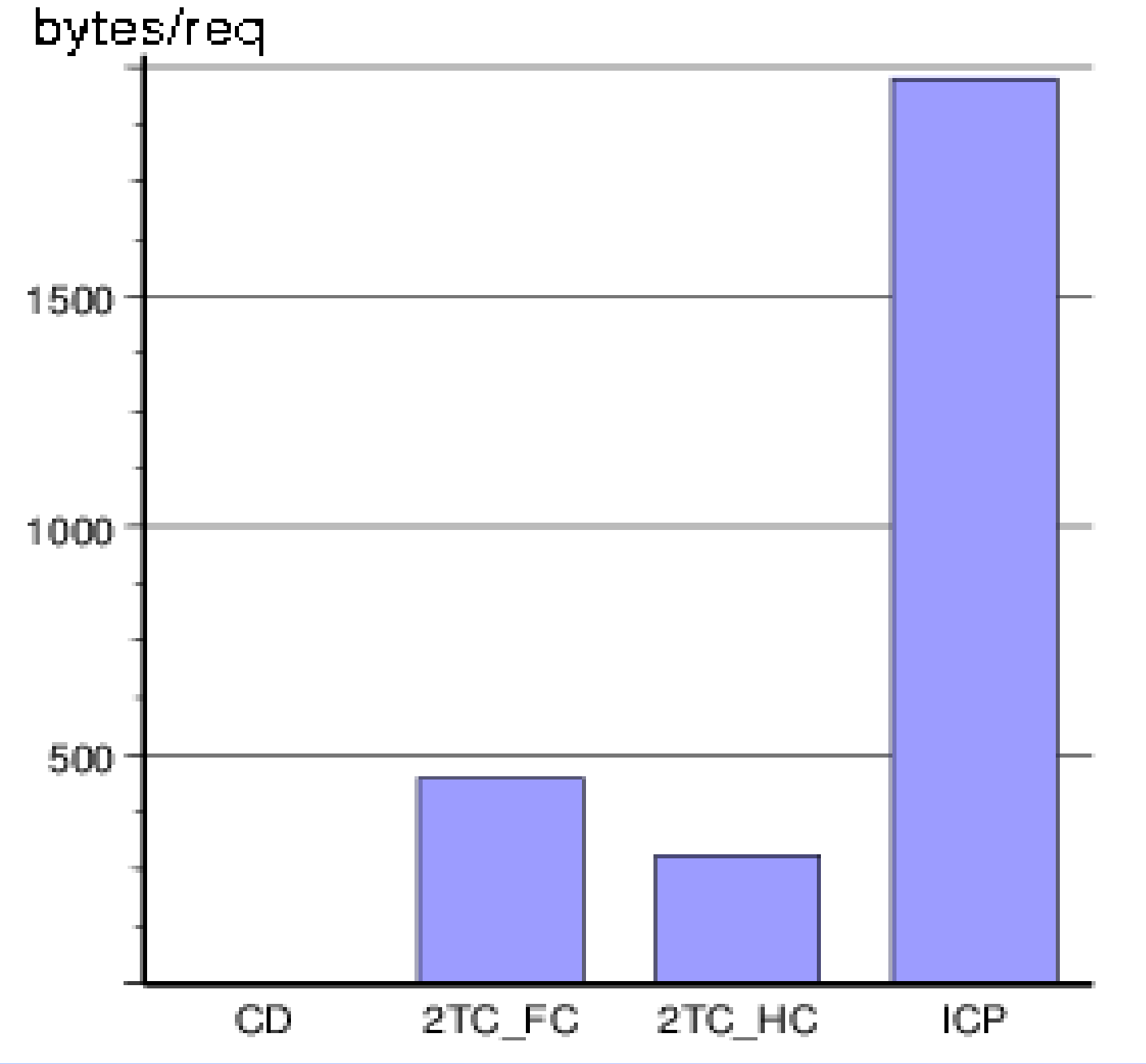
# Scalability Test (log(overhead))



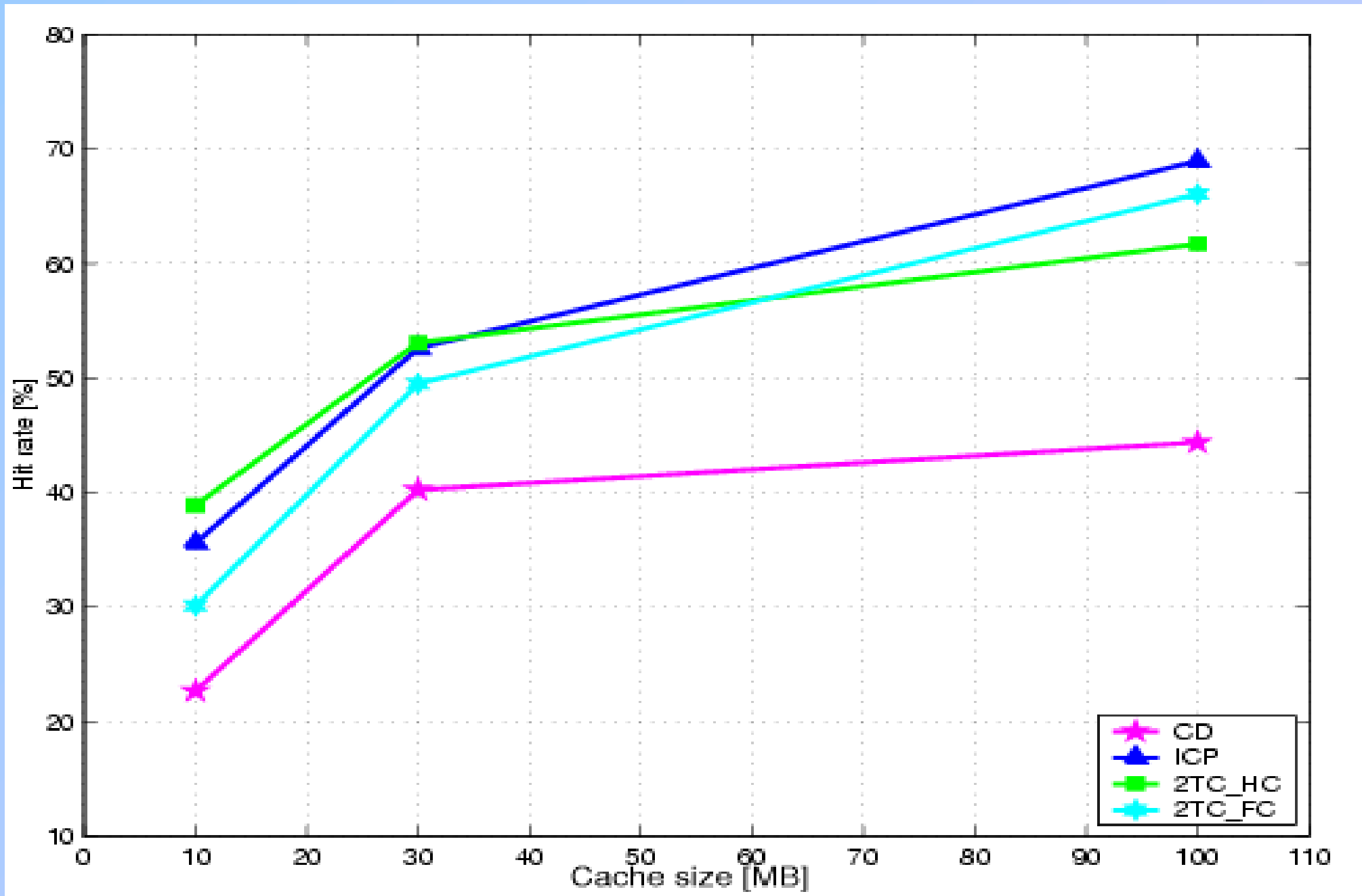
# Scalability Test (overhead)

The real look of overhead comparison!

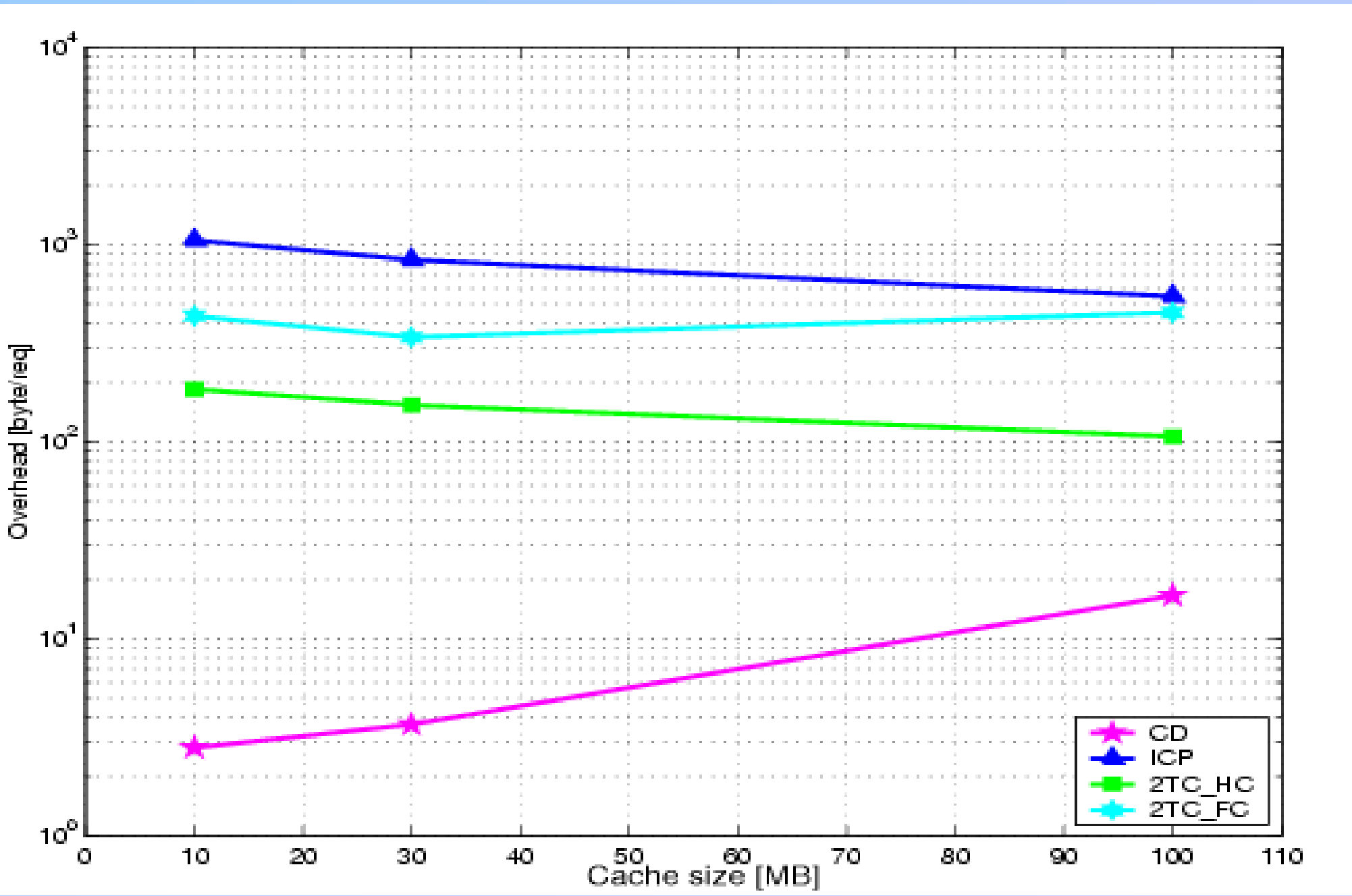
(30 nodes)



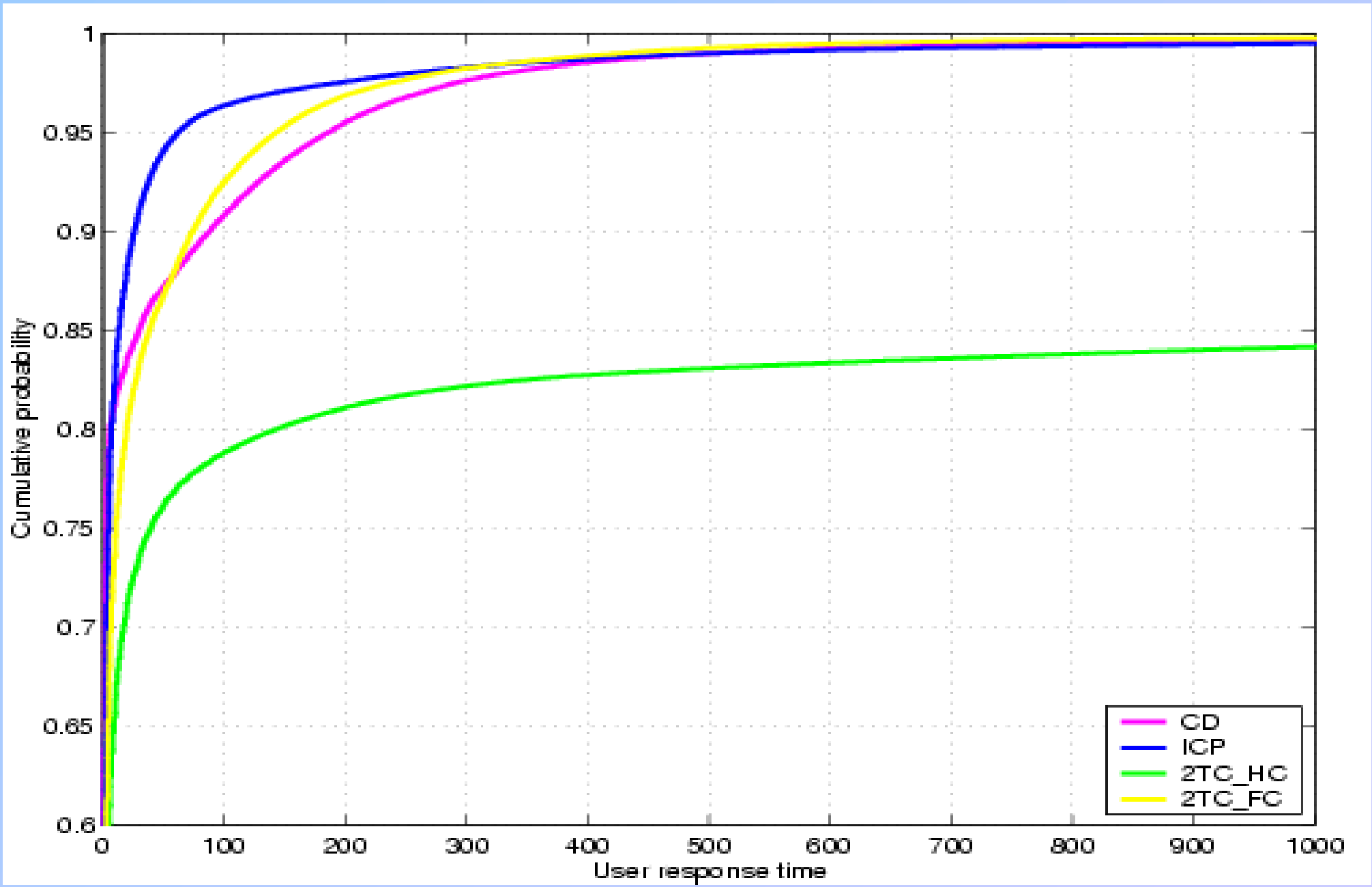
# Sensitivity to cache size



# Sensitivity to cache size



# User Response Time



# Summary

Both two-tier architectures offer:

- Cooperation overhead < ICP
- Hit rates > CD
- Better scalability than pure architectures

Flat clustering is preferable when there are:

- well connected intra-PC links and
- loose inter-PC connections

Hierarchical clustering is better when

- all caches are well connected or
- there are large cache digests to be exchanged

# A Scalable Architecture for Cooperative Web Caching

Riccardo Lancellotti
Università di Roma
Tor Vergata

Bruno Ciciani
Università di Roma
La Sapienza

Michele Colajanni
Università di Modena e
Reggio Emilia