Timestamp Snooping:
An Approach for Extending SMPs

Milo M. K. Martin, Daniel J. Sorin, Anastassia Ailamaki,
Alaa R. Alameldeen, Ross M. Dickson, Carl J. Mauer,
Kevin E. Moore, Manoj Plakal, Mark D. Hill, David A. Wood

University of Wisconsin-Madison

http://www.cs.wisc.edu/multifacet/

Ninth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS-IX)

November 2000

Overview

- Problem: multiprocessors for commercial workloads
- Snooping (SMPs)
  + Finds data directly - no indirection
  - Constrains interconnect
- Goal: Free snooping from interconnect constraints
- Timestamps provide logical global order
- Evaluation vs directory protocol (CC-NUMA)
  - Commercial workloads on 16 processors
  - 6-23% faster
  - Directories use17-37% less bandwidth

EXTENDING SMPS TO GENERAL INTERCONNECTS

Outline

- Overview
- Commercial Workloads
- Traditional Coherence
- Timestamp Snooping
    - Interconnect
    - Protocol
- Evaluation
- Conclusion

Commercial Workloads

- Dominant use of multiprocessors

- Moderate processor count

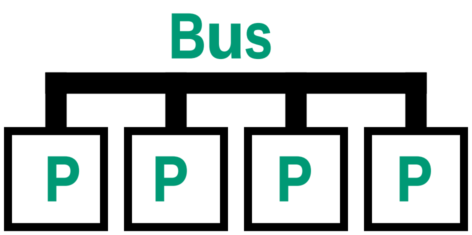
  2-8, then 16-64, but not 1024

- Many cache-to-cache transfers (3-hop or dirty misses)
  - 55-62% for OLTP [Barroso et al. ISCA '98]
  - 40-60% for our commercial workloads

DESIGN MULTIPROCESSORS FOR COMMERCIAL WORKLOADS

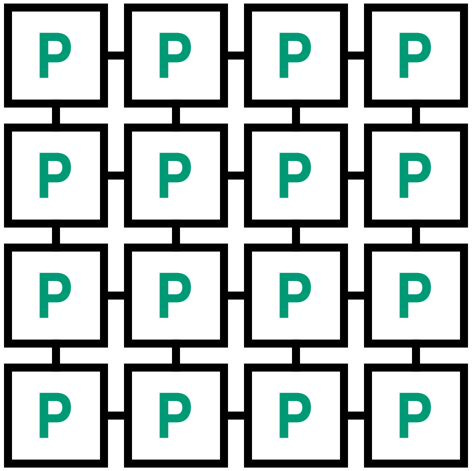
Traditional Snooping (SMPs)

- Operation
  - Requests sent on physical bus
  - Processors & memory *snoop* requests
  - Snoop responses
  - Owner responds
- Advantages
  + Fast cache-to-cache transfers
- Disadvantages
  - Bus bottleneck
  - Signaling limitations
- Agarwal et al. (1988) predicted the demise of SMPs



Directory Protocols (CC-NUMA)

- Add a level of indirection (for some requests)
  - Send requests to *a directory*
  - Directory redirects request

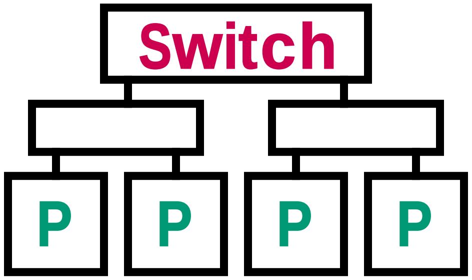


- Advantages
  - \+ Avoids broadcast → scalable
  - \+ Few interconnect restrictions
- Disadvantage
  - \- Directory state
  - \- Slow cache-to-cache transfers (3-hops)
- Example: Alpha 21364 - directory protocol with 2D torus

GAINS SCALABILITY AT THE COST OF SLOW 3-HOP TRANSACTIONS

Modern SMPs

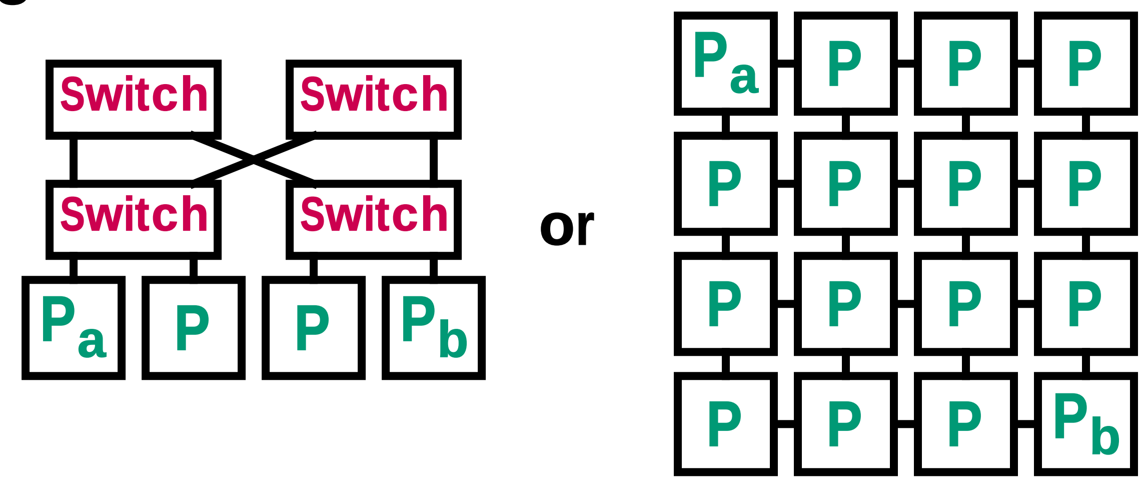
- Many enhancements
  - Multiple buses
  - Pipelined broadcast tree with point-to-point links
- Commercially successful, few academic papers
- Challenges
  - 'Logical bus' → synchronous broadcast
  - Global snoop responses
  - Arbitration & flow control
- Example: Sun E10000 - 64 processors
  130 ASICs for interconnect



SMPs IMPOSE INTERCONNECT RESTRICTIONS

Extending Snooping

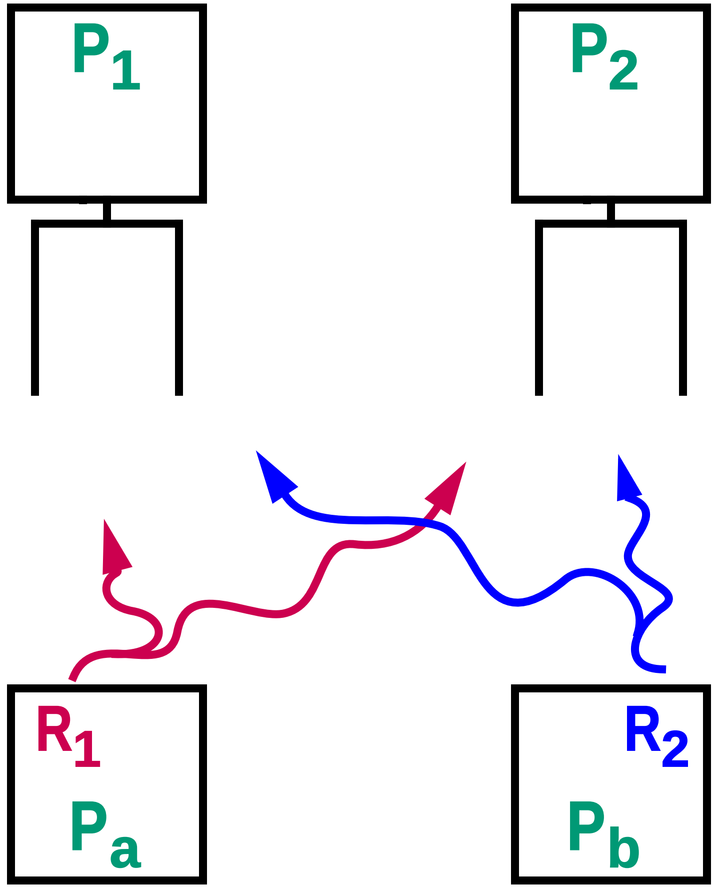
- Key requirements
  - **Total order**
  - Broadcast
- Relax other requirements
  - No synchronous interconnect
  - Arbitrary topology (direct or indirect)
  - No snoop responses
  - No global arbitration



PROVIDE TOTAL ORDER WITH FEWER INTERCONNECT RESTRICTIONS

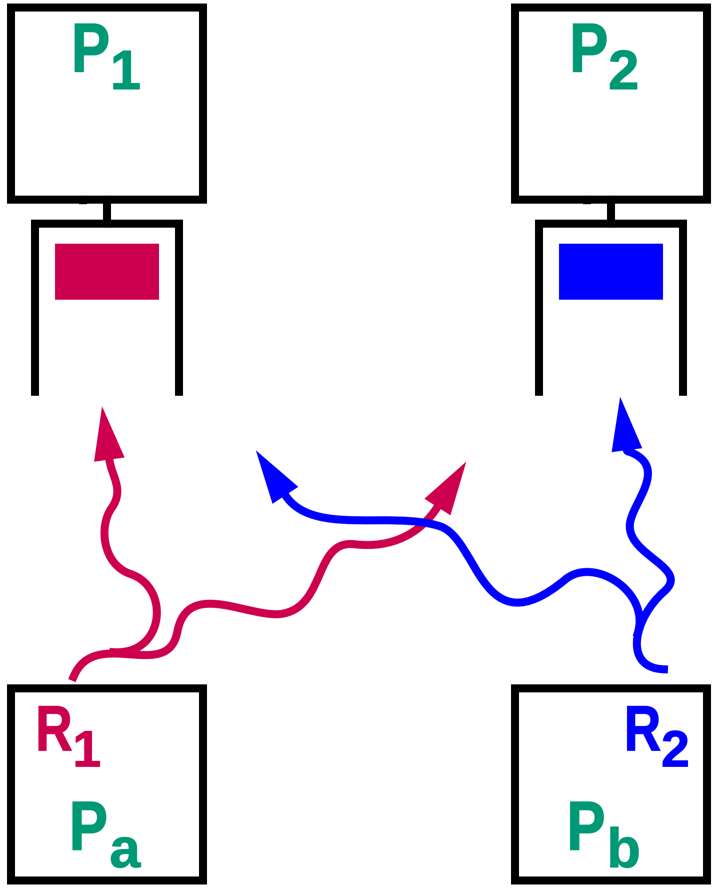
Timestamp Snooping

- **Goal:** Create a logical total order



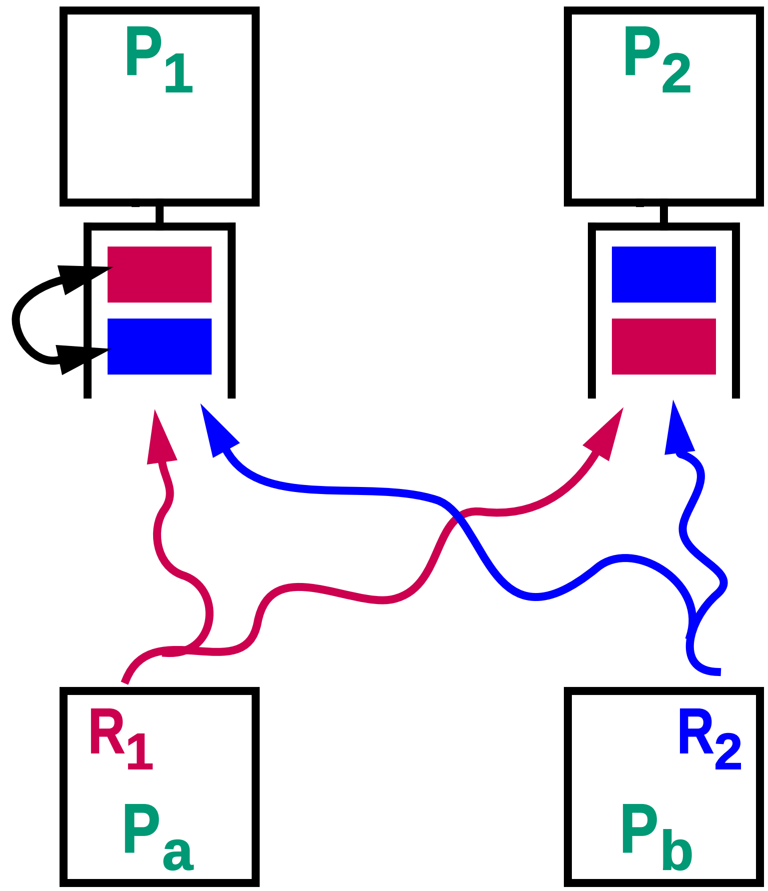
Timestamp Snooping

- **Goal:** Create a logical total order



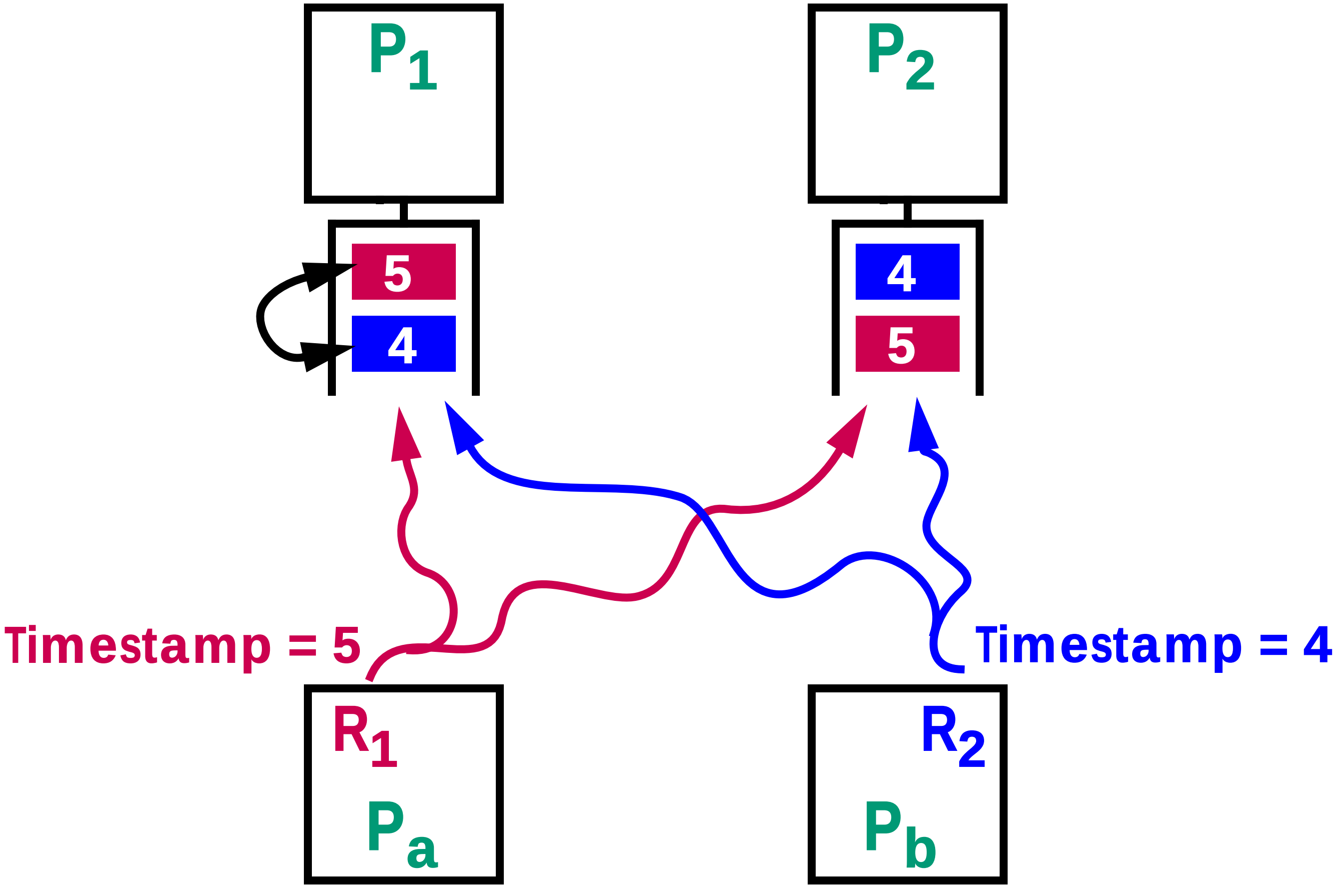
Timestamp Snooping

- **Goal:** Create a logical total order



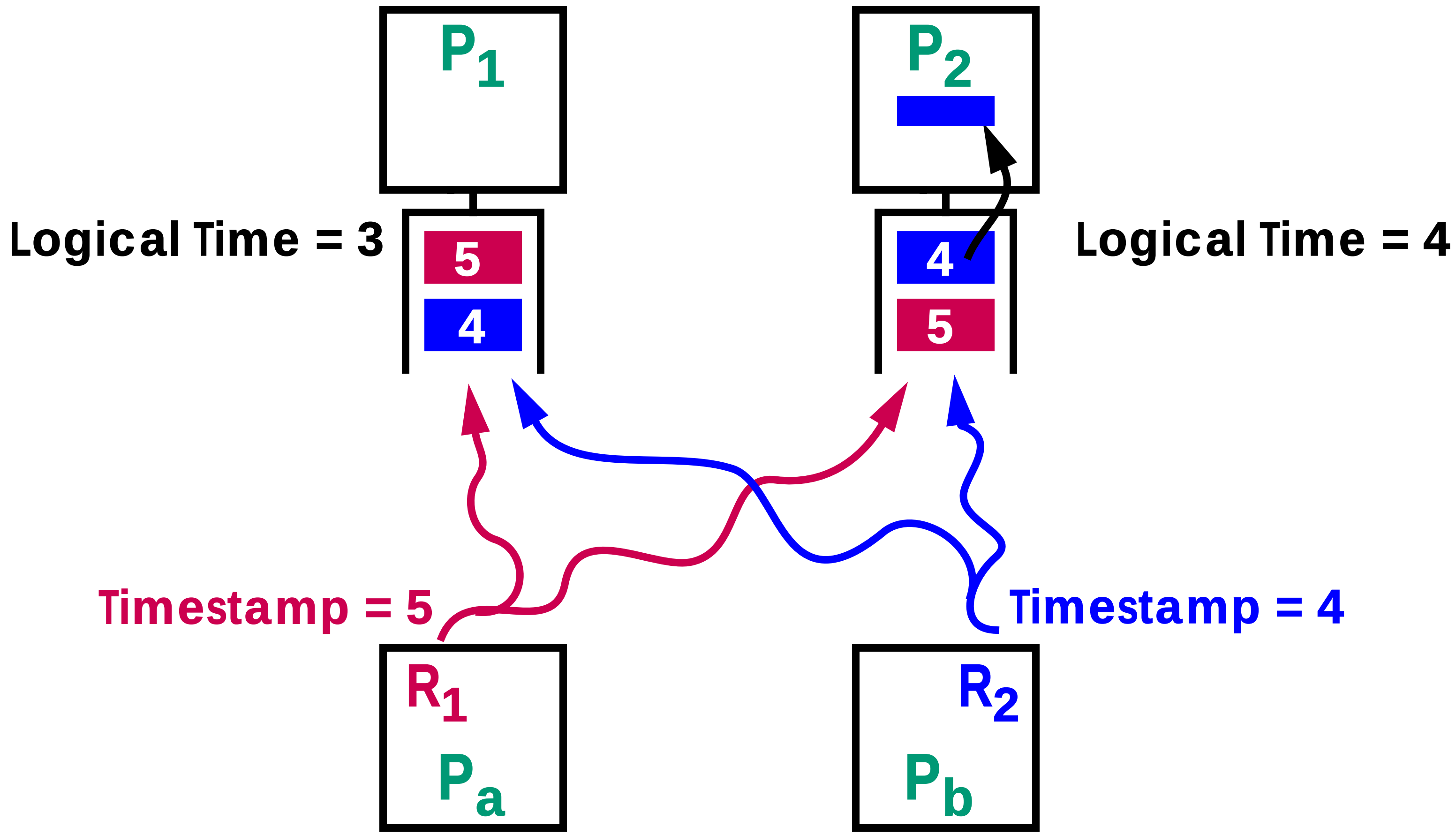
Timestamp Snooping

- **Goal:** Create a logical total order



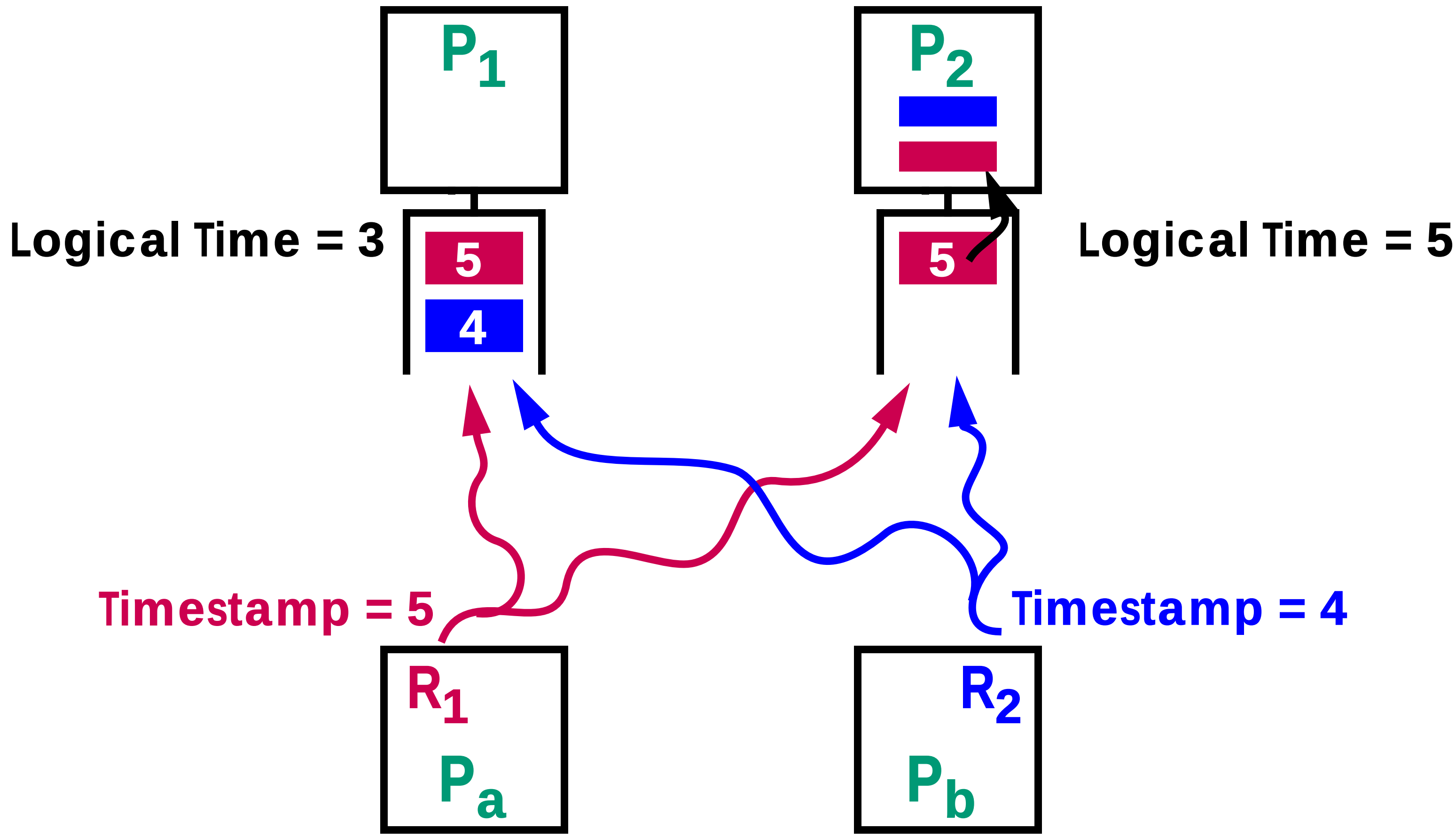
Timestamp Snooping

- **Goal:** Create a logical total order



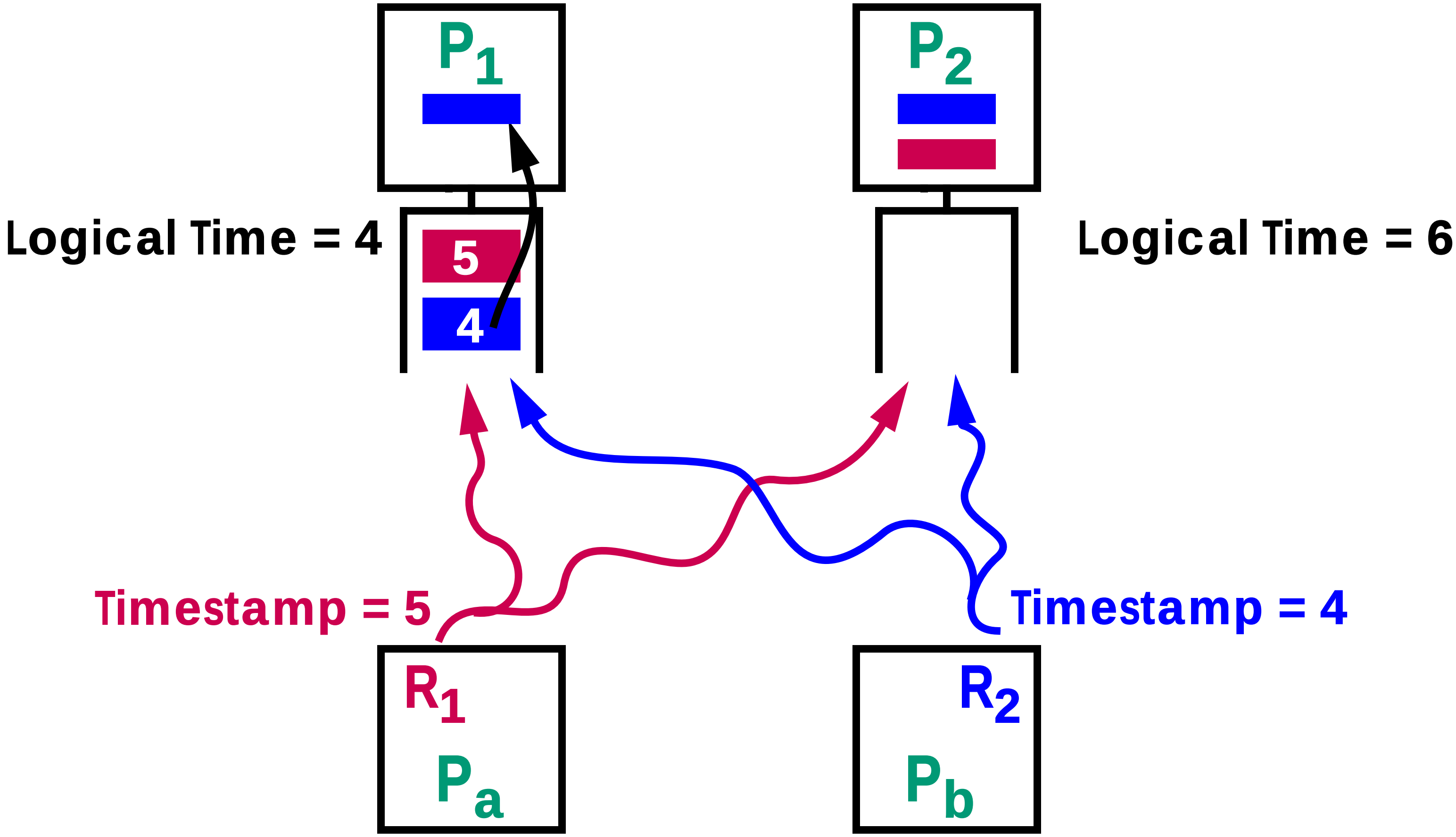
Timestamp Snooping

- **Goal: Create a logical total order**



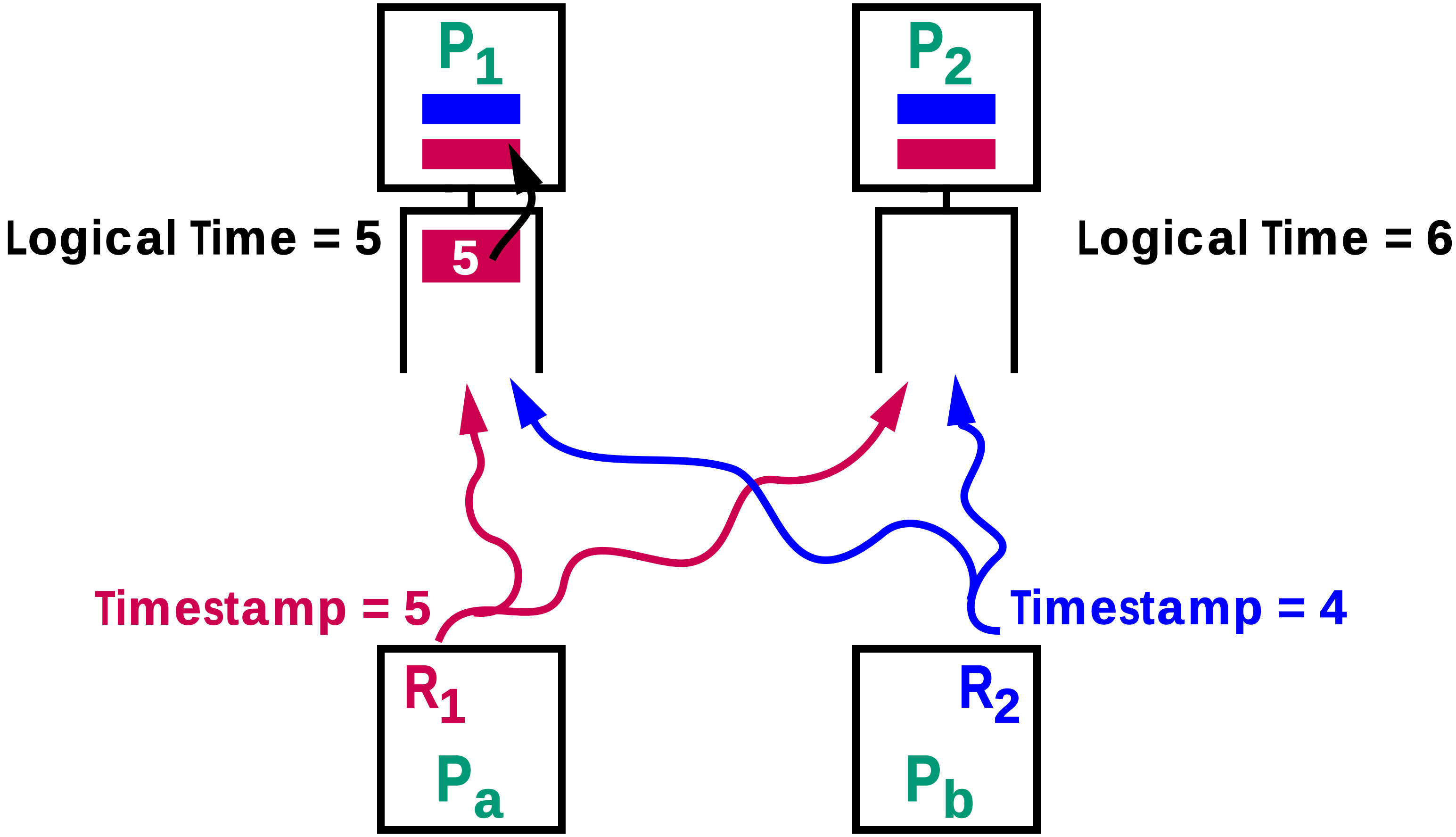
Timestamp Snooping

- **Goal:** Create a logical total order



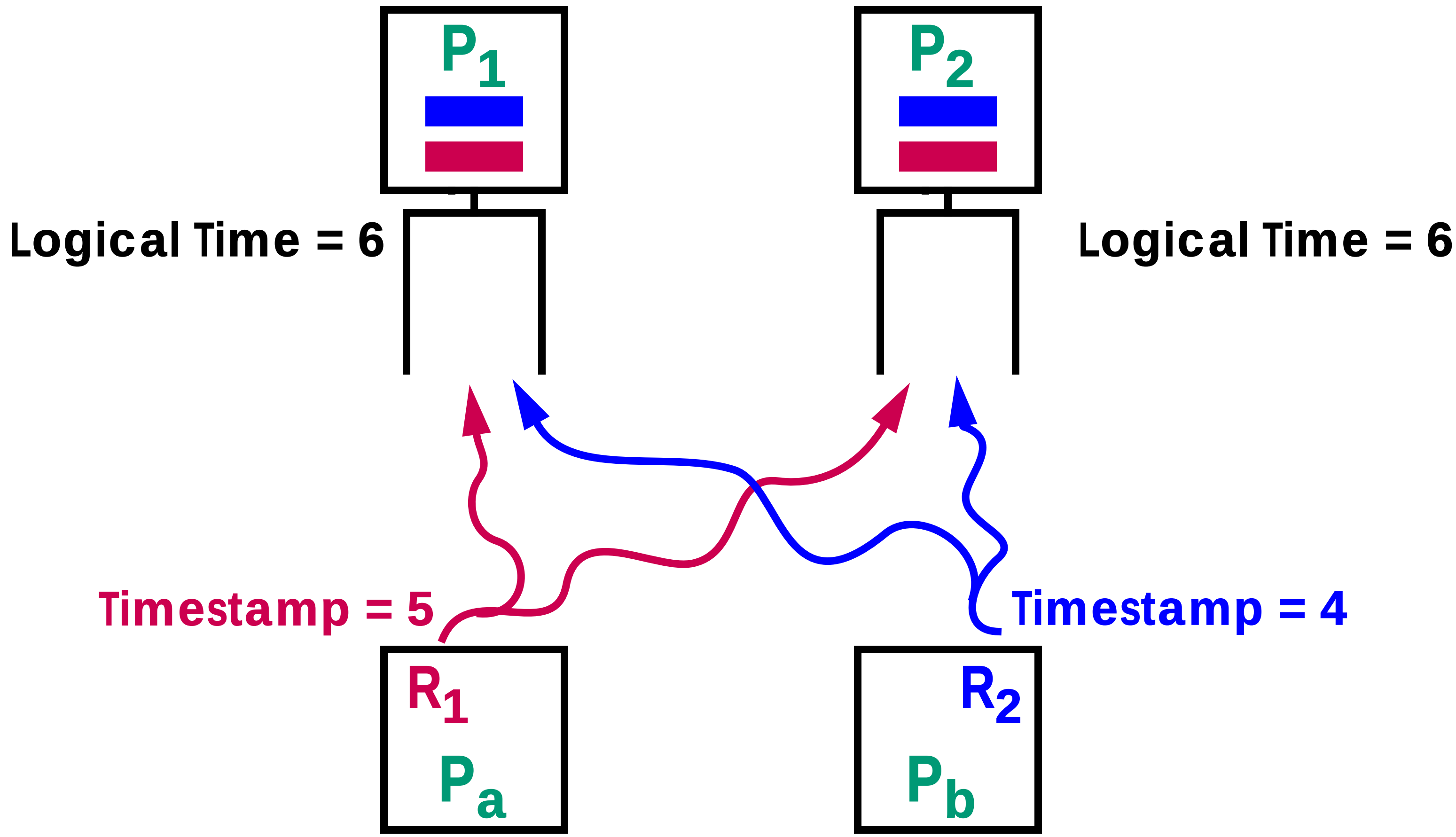
Timestamp Snooping

- **Goal: Create a logical total order**



Timestamp Snooping

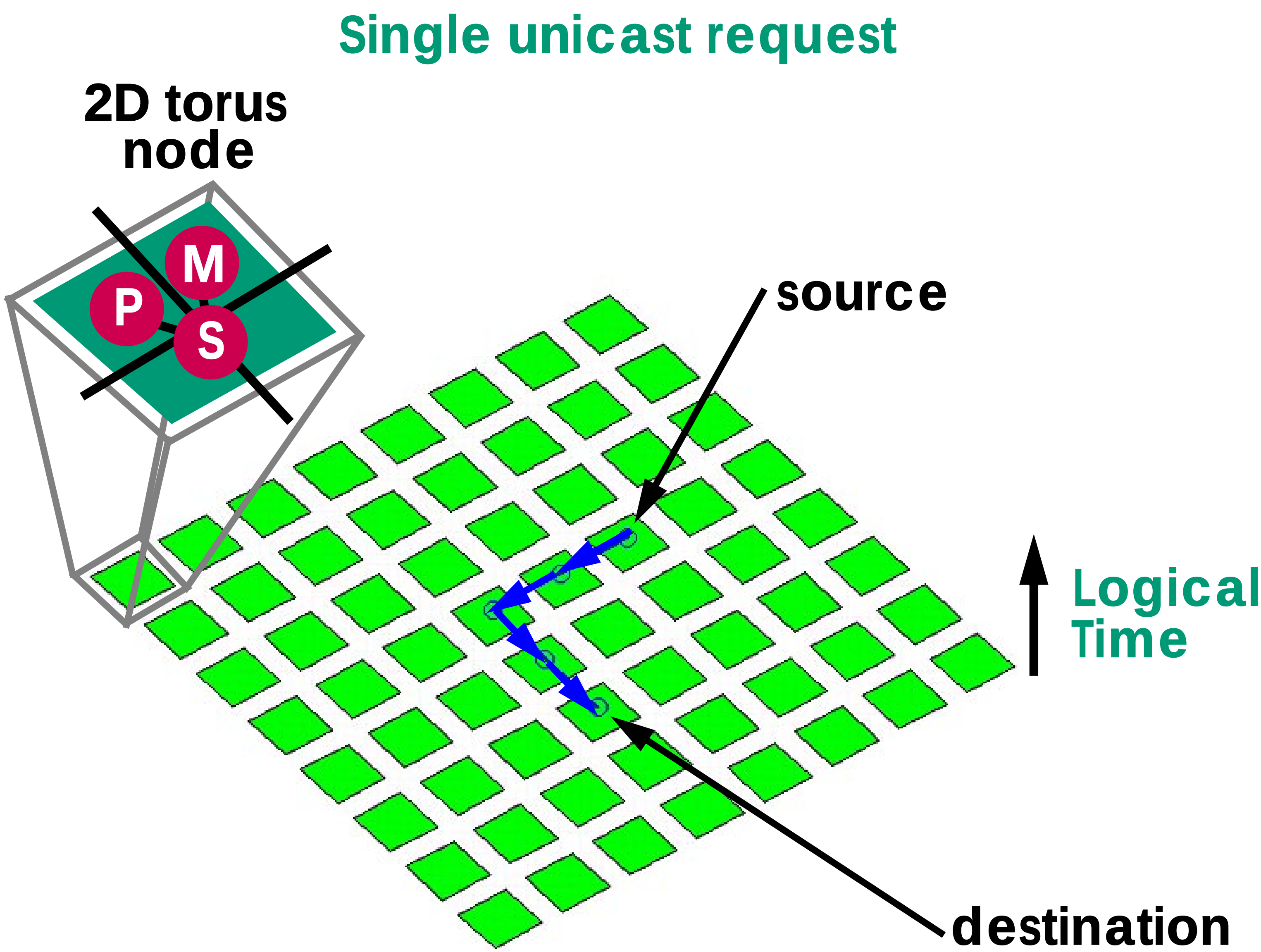
- **Goal: Create a logical total order**



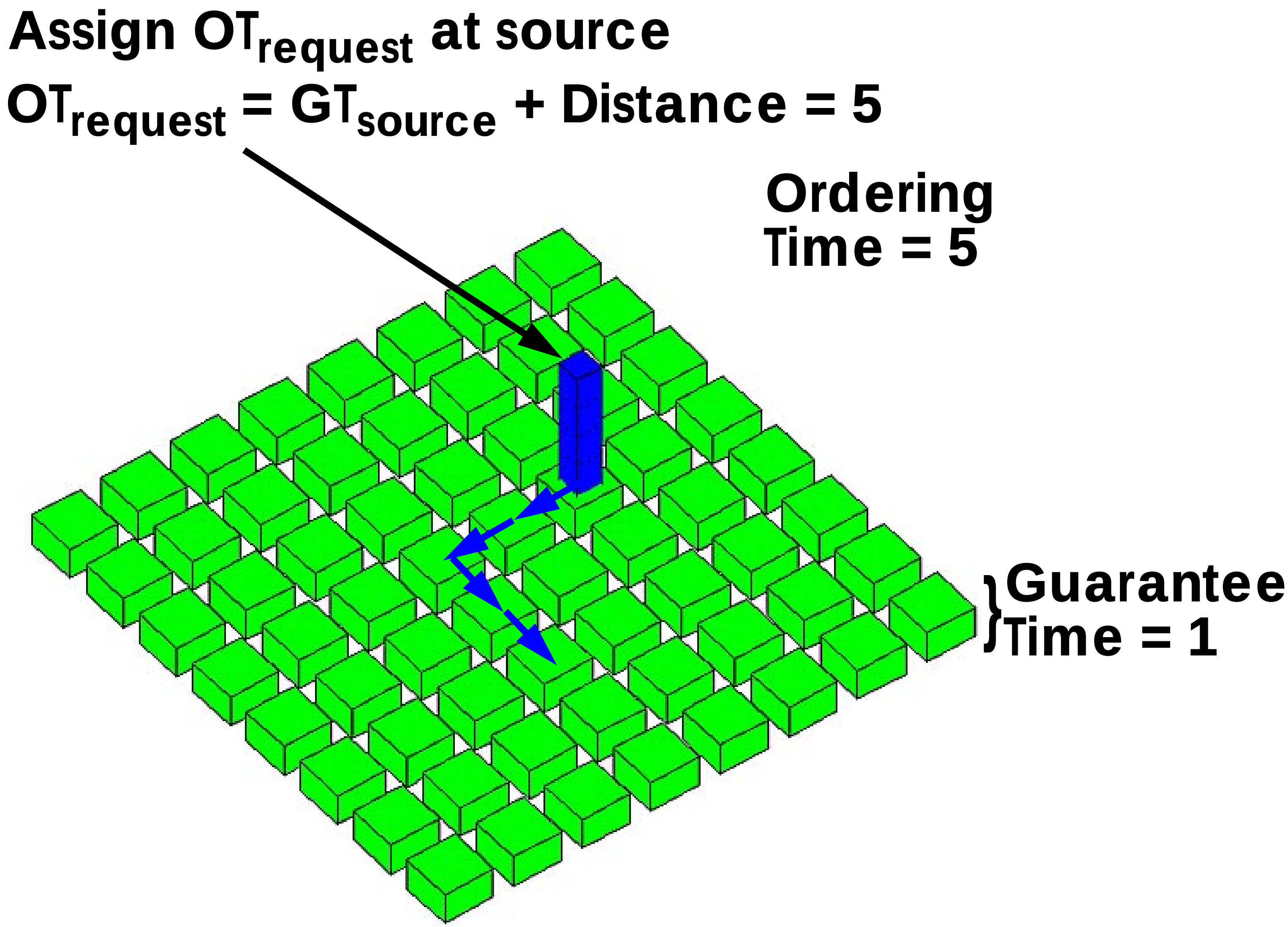
Logical Time

- Ordering Time (OT)
  - Arrival timestamp of request
  - Assign at source
  - Broadcast without regard to order
  - Re-order at the end-points
- Guarantee Time (GT)
  - Logical time base
  - Recursively maintained at switches
- Invariant
  - Messages delivered while $OT_{request} \geq GT_{destination}$

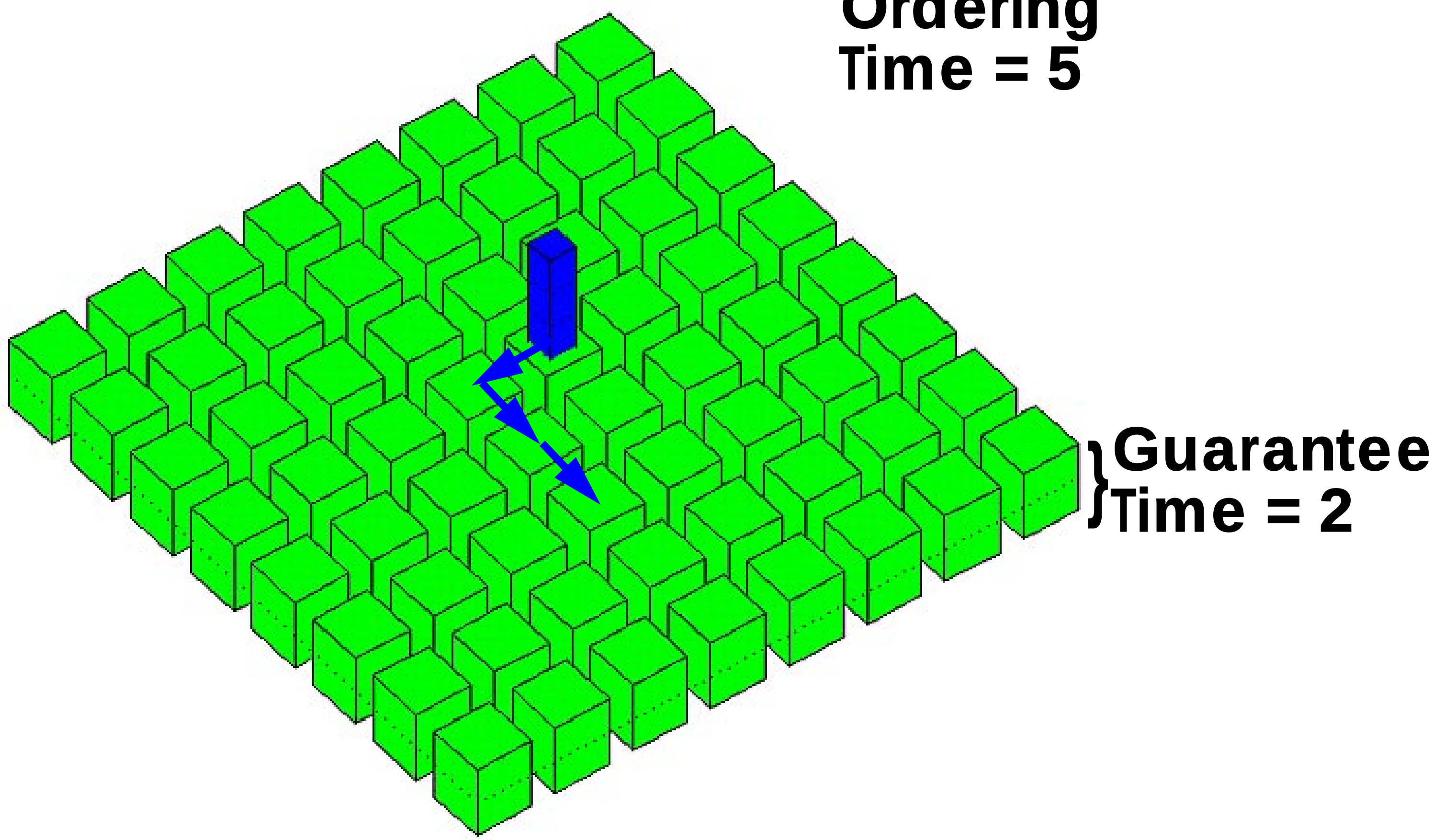
Uncontended Example



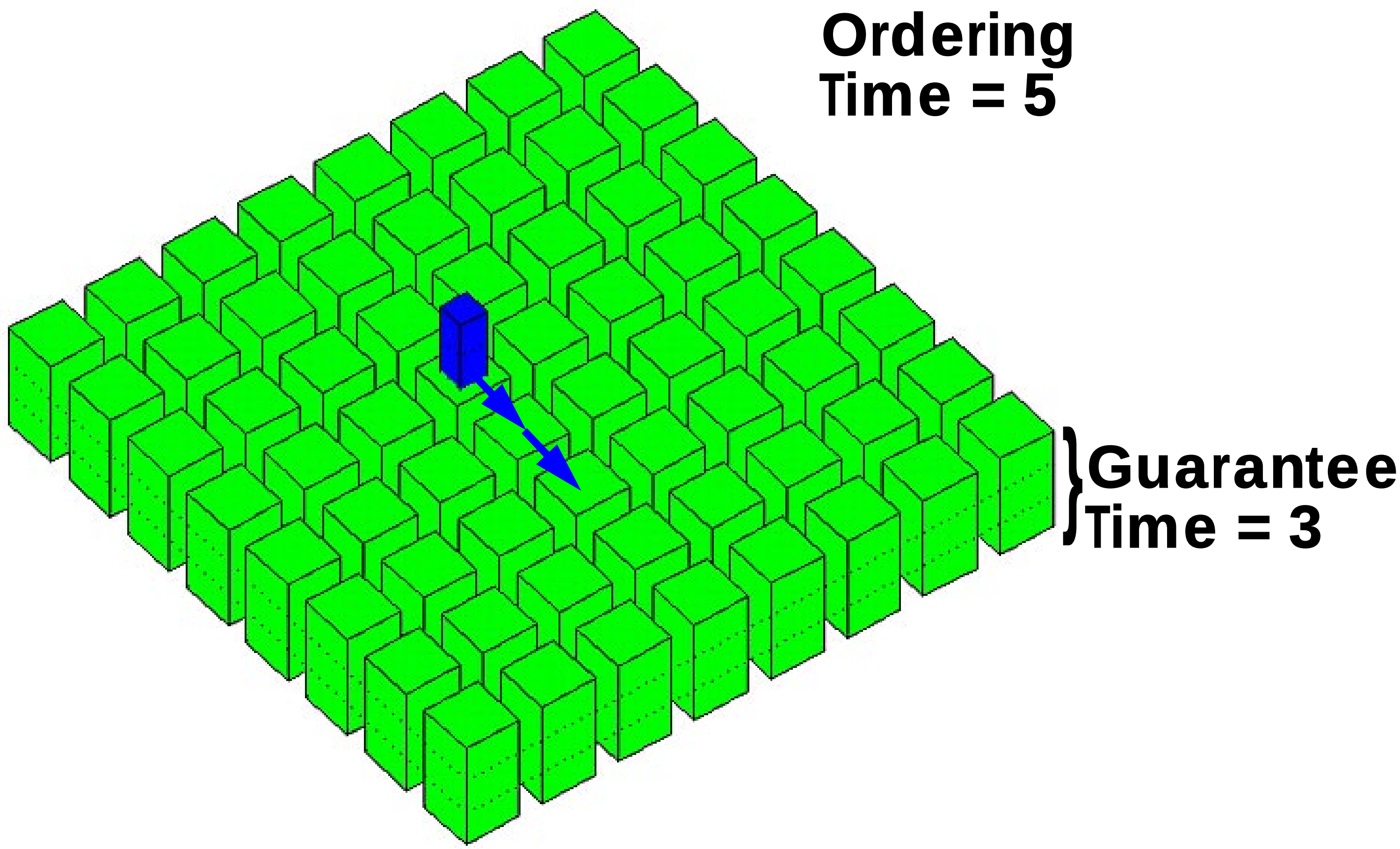
Uncontended Example



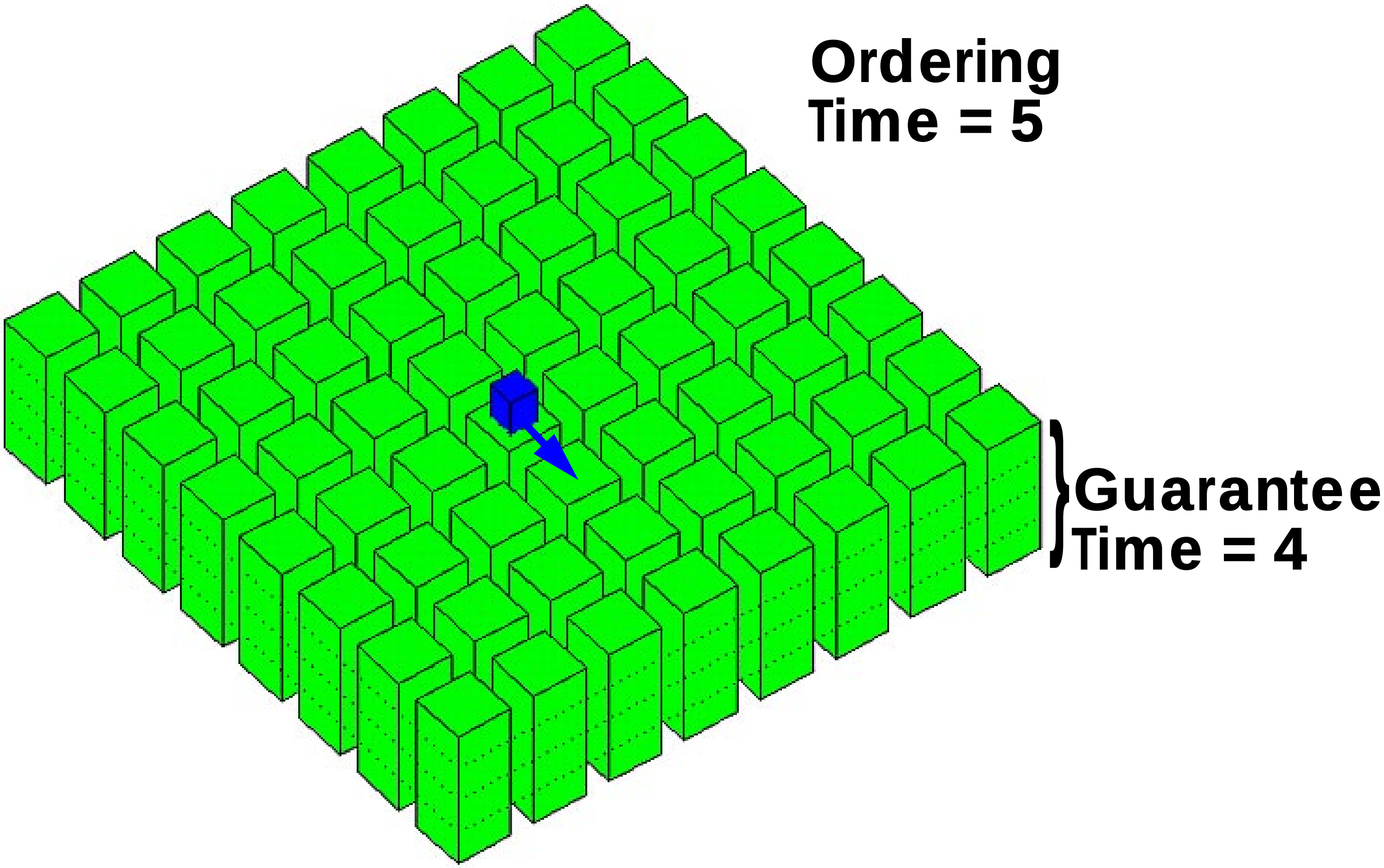
Uncontended Example



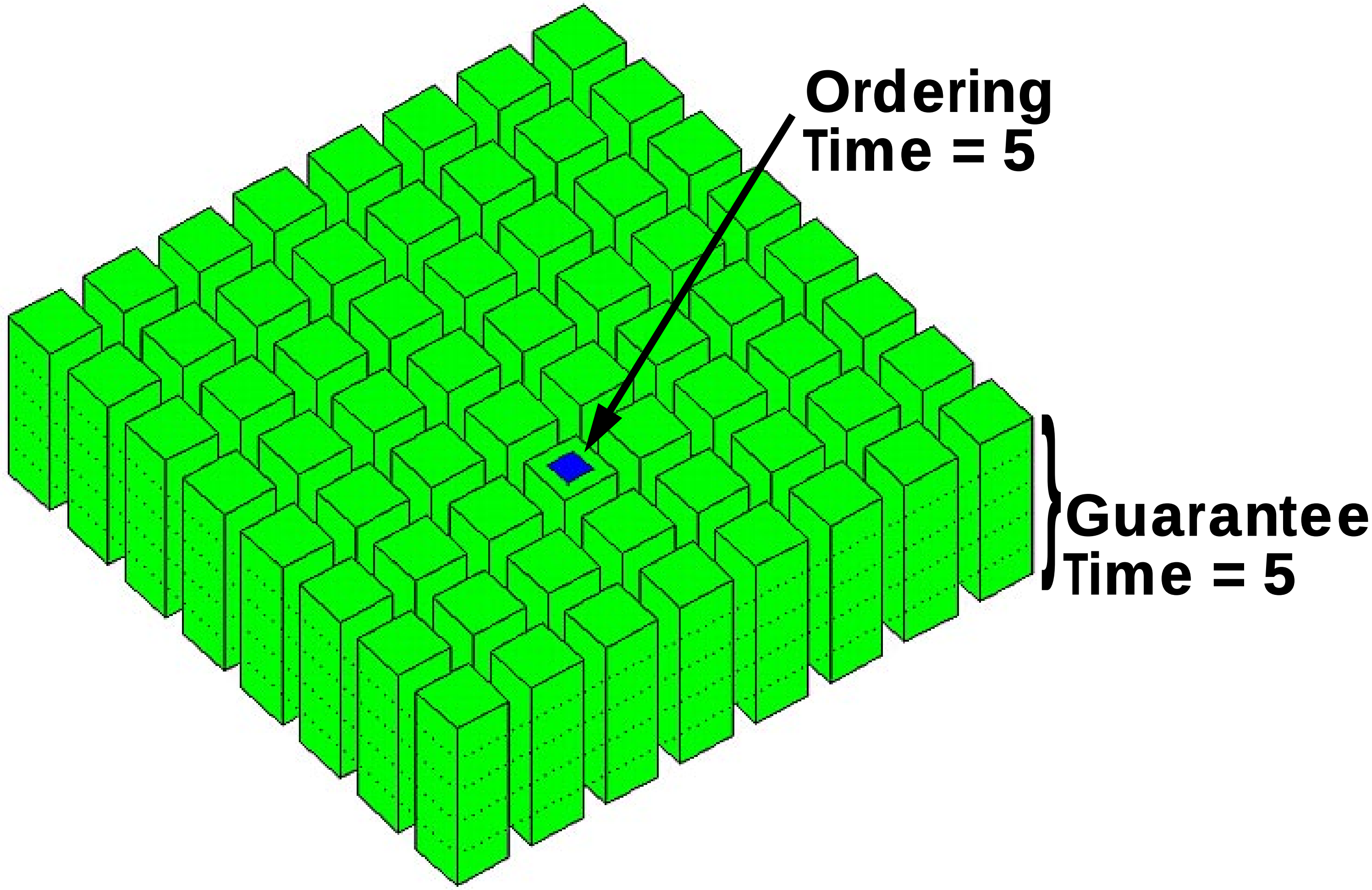
Uncontended Example



Uncontended Example



Uncontended Example

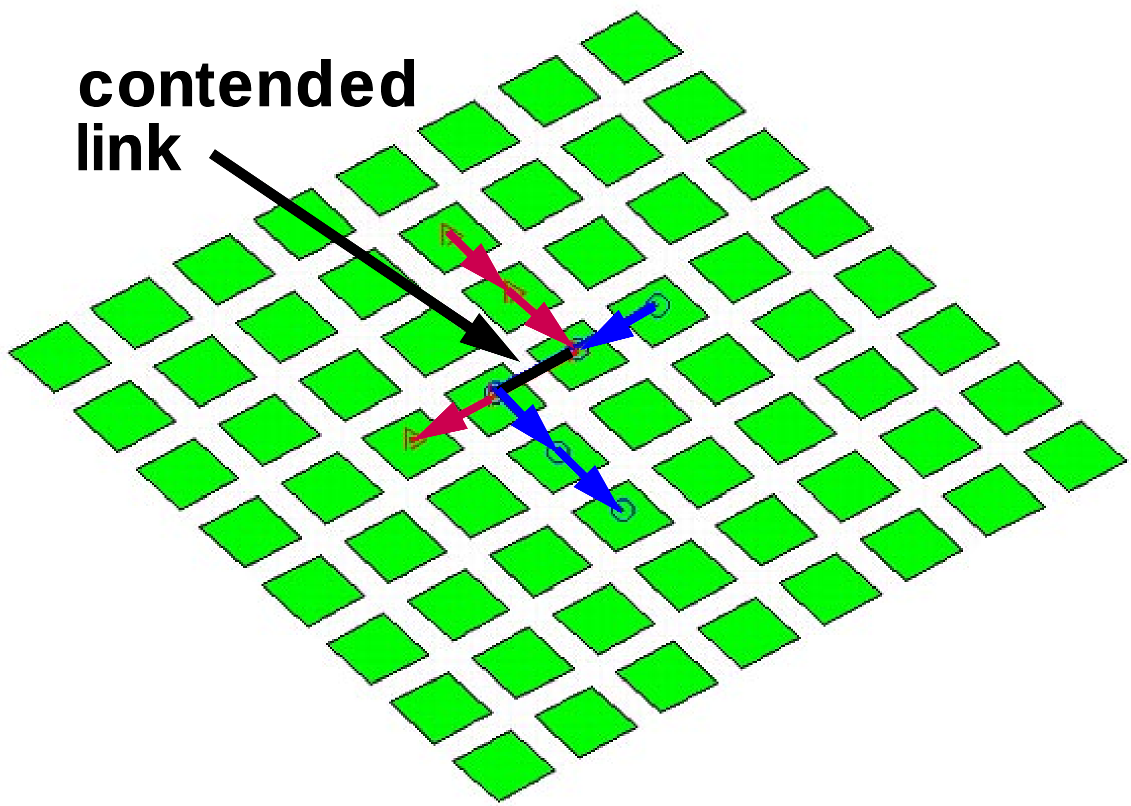


Interconnect Contention

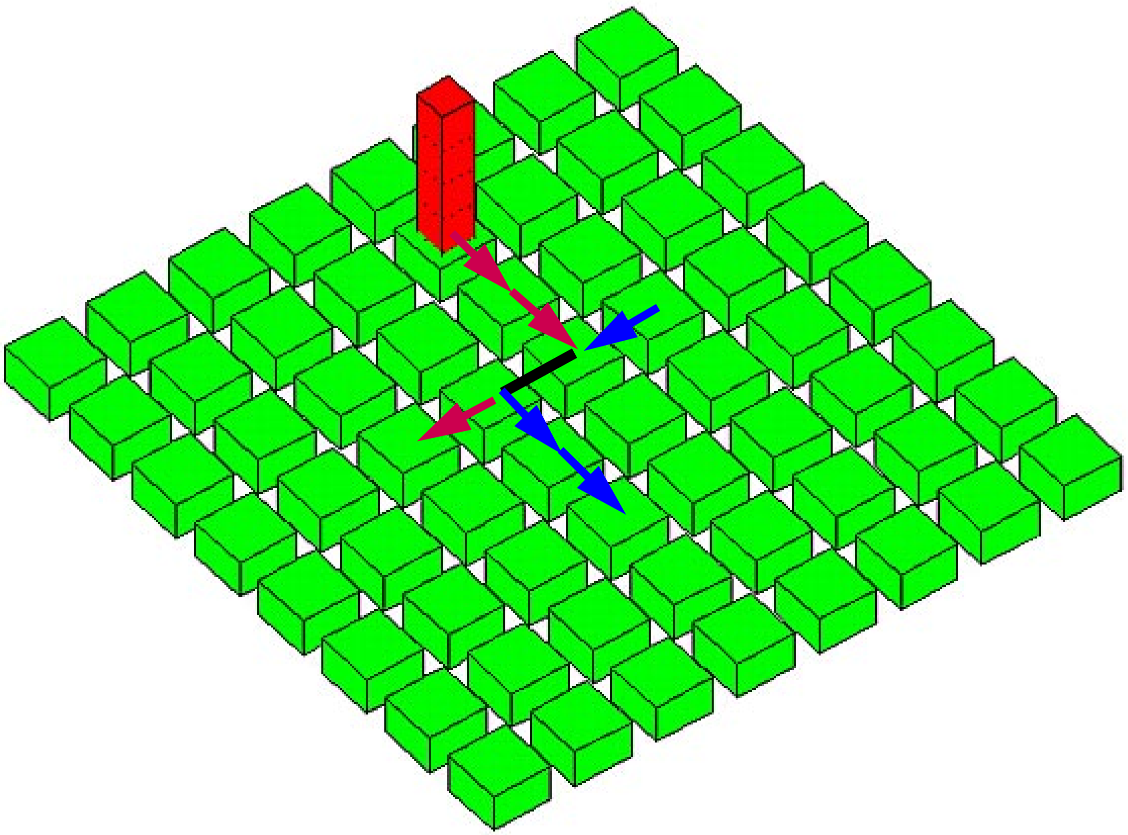
- **Invariant**
  - **Requests delivered while $OT_{request} \geq GT_{destination}$**
- **No contention**
  - **GTs always advance**
- **Contention**
  - **Recursively delay GTs to ‘warp time’**
  - **Prevent requests from being *late***

Contention Example

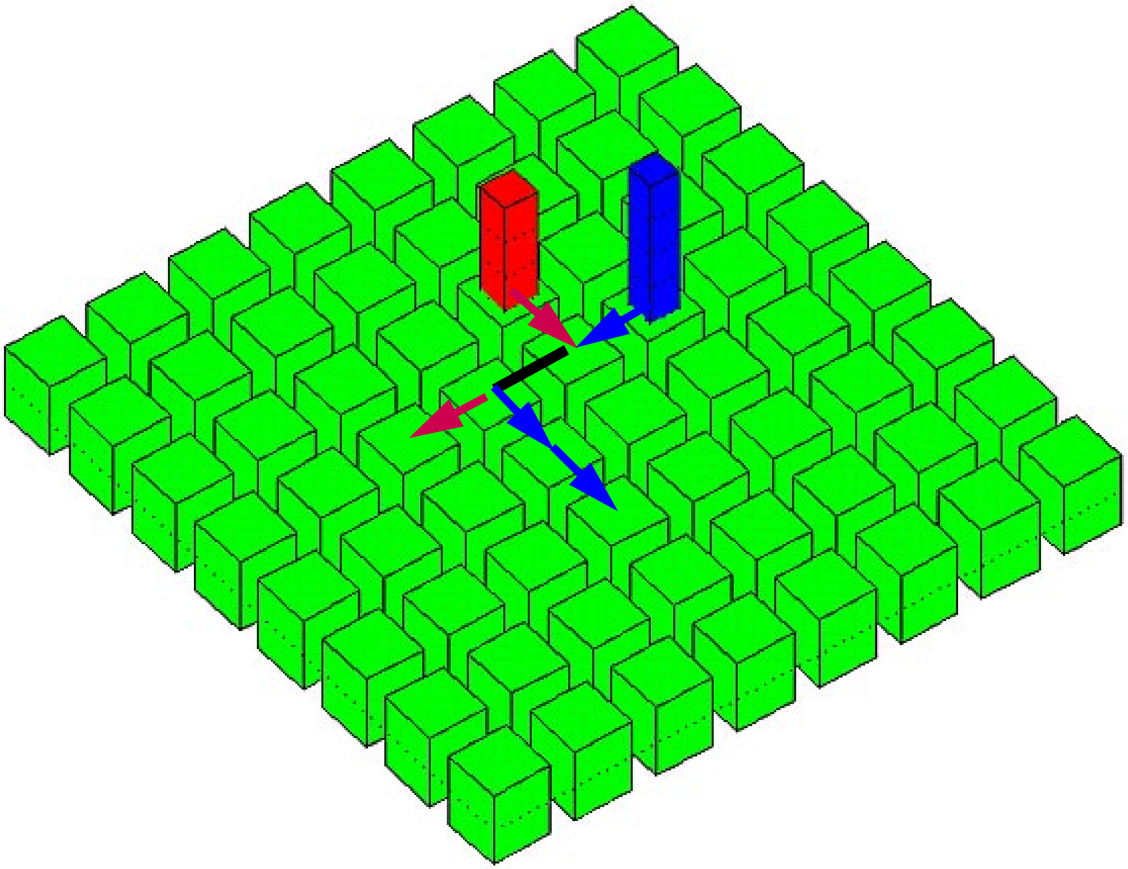
Two request example



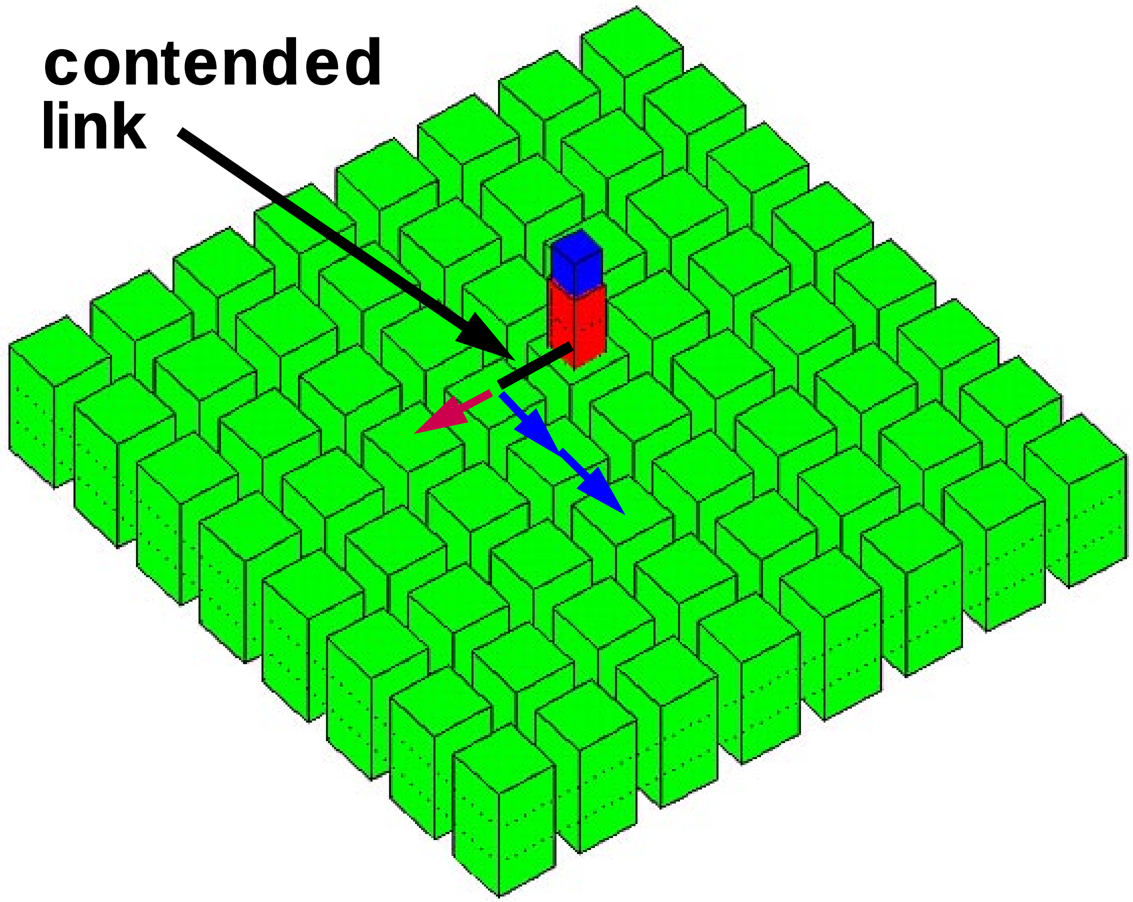
Contention Example



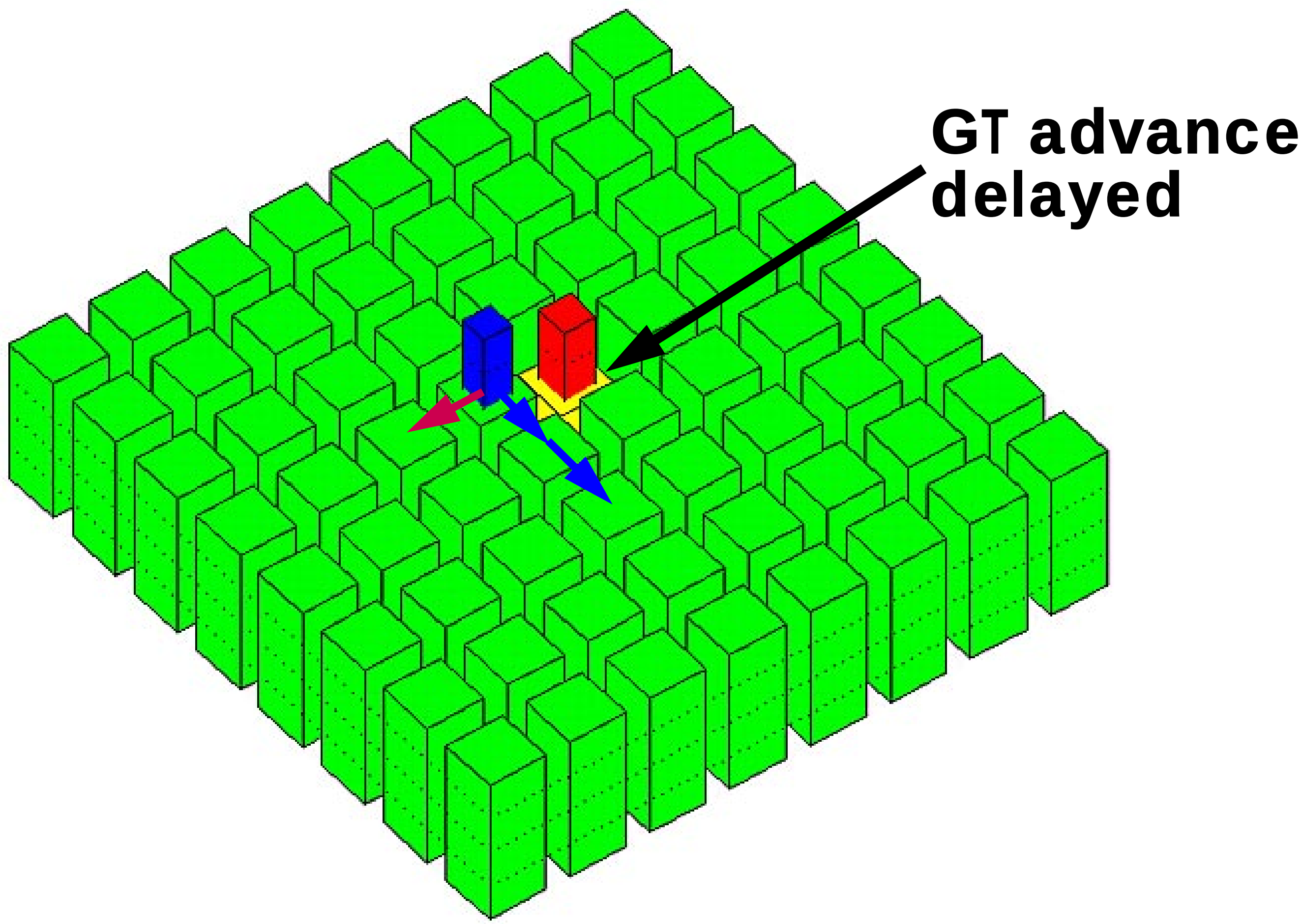
Contention Example



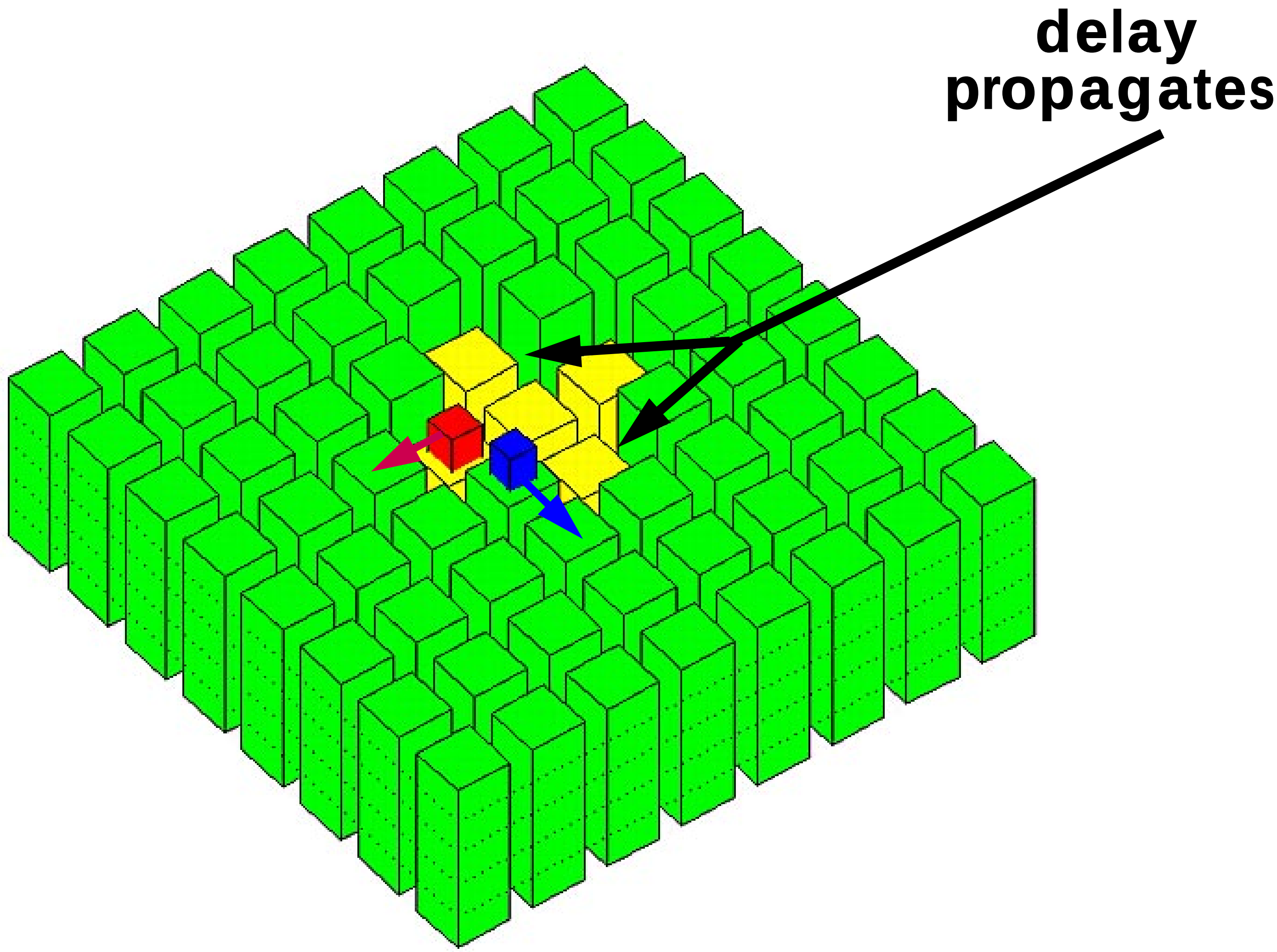
Contention Example



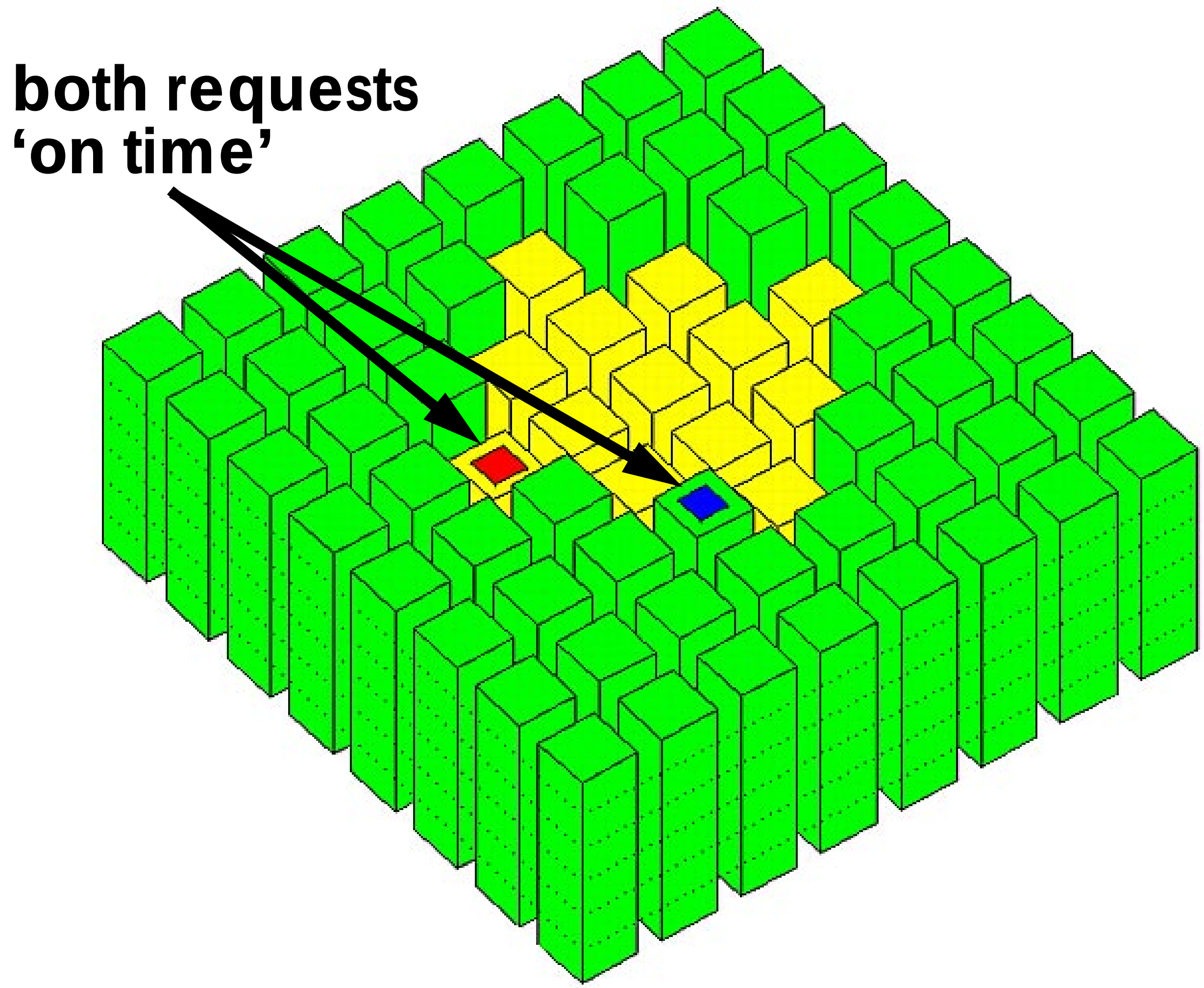
Contention Example



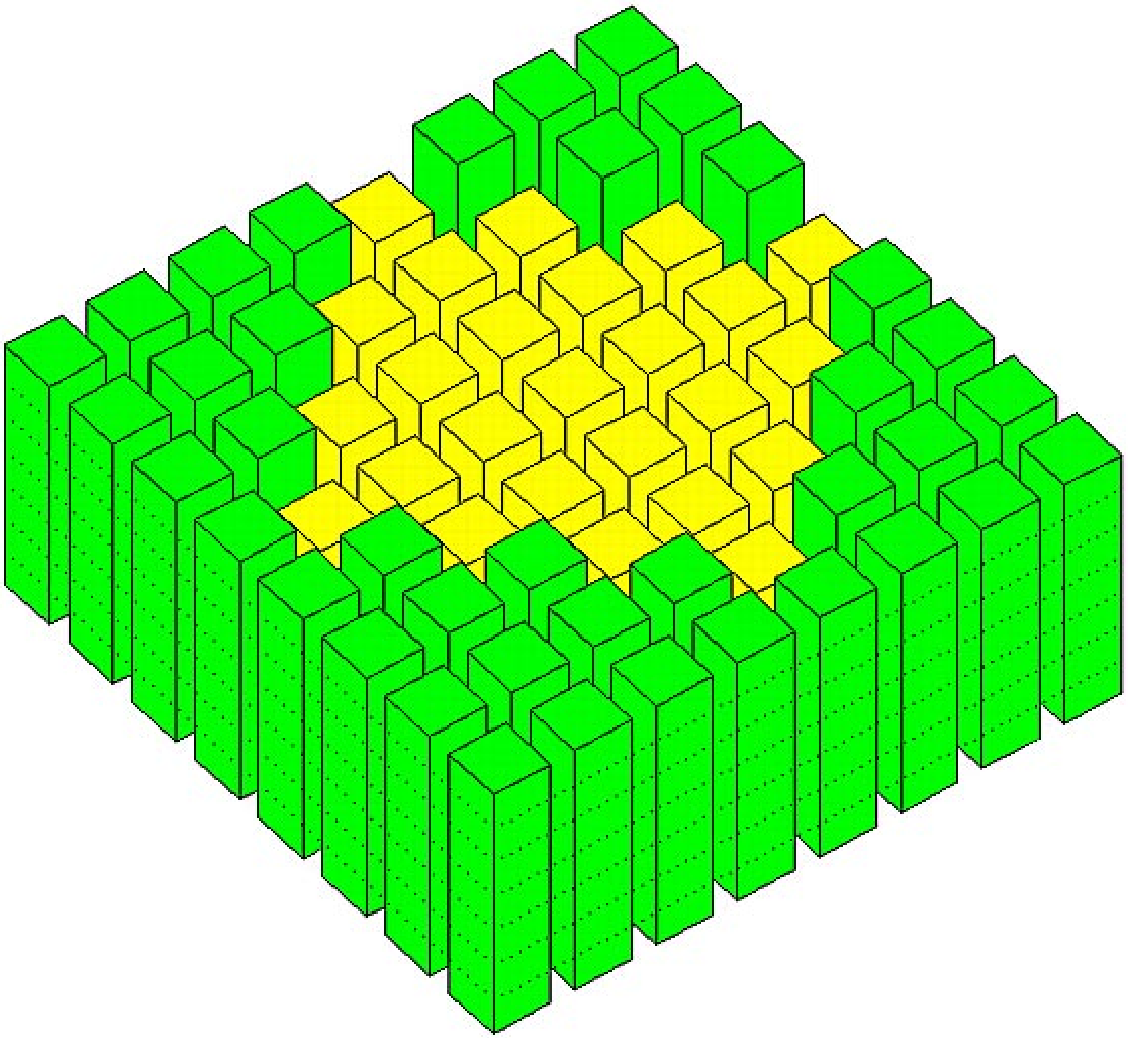
Contention Example



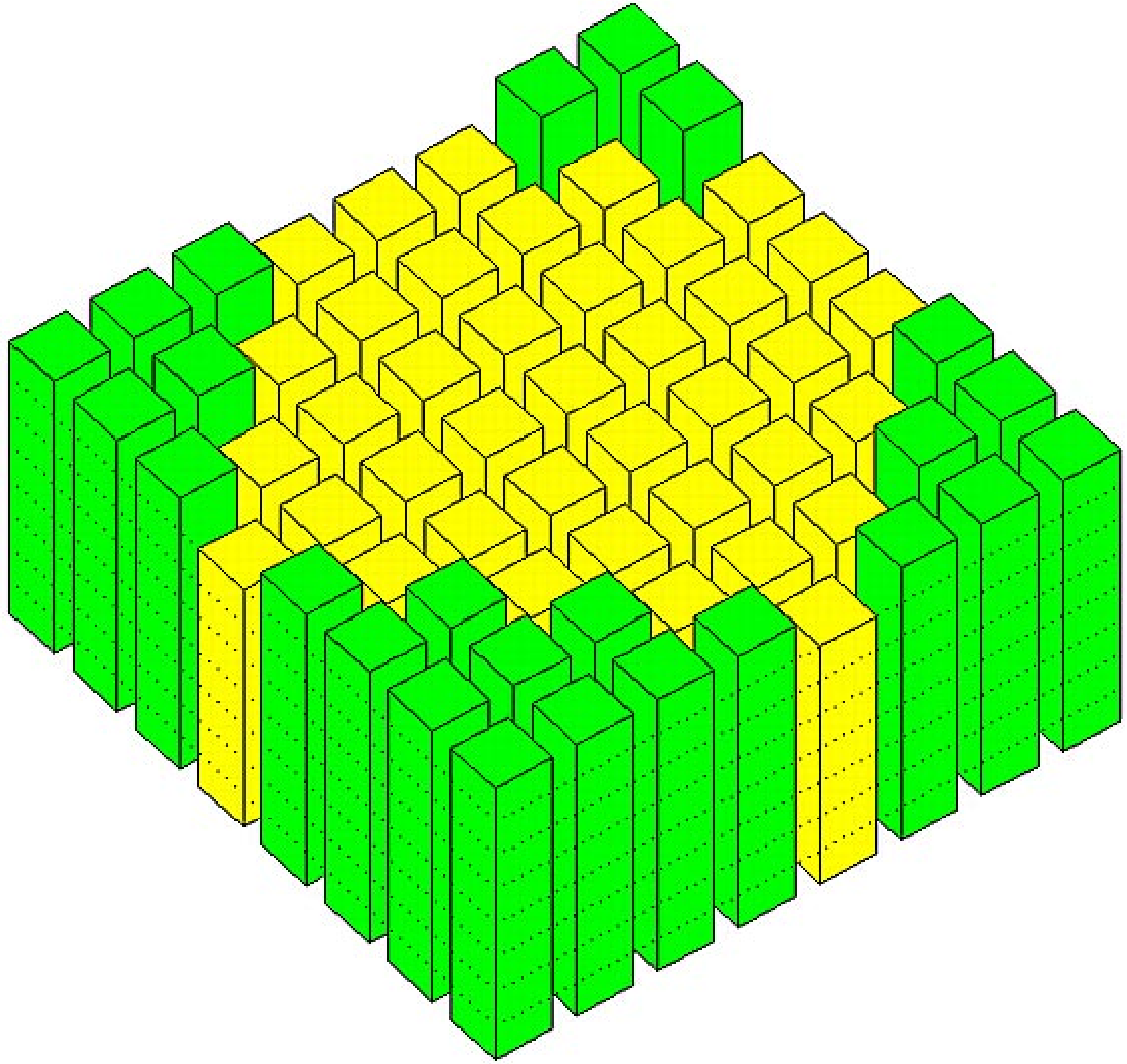
Contention Example



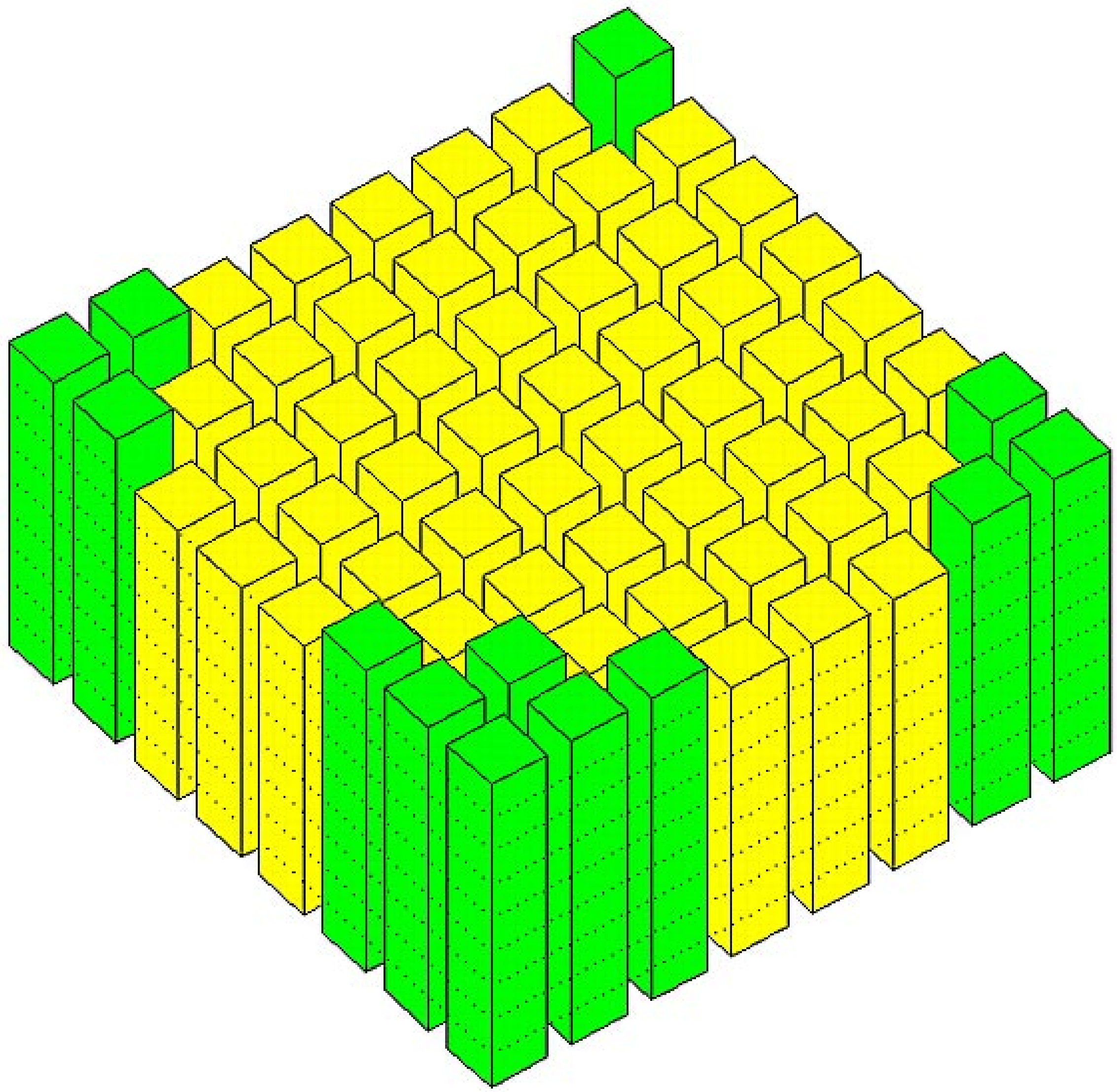
Contention Example



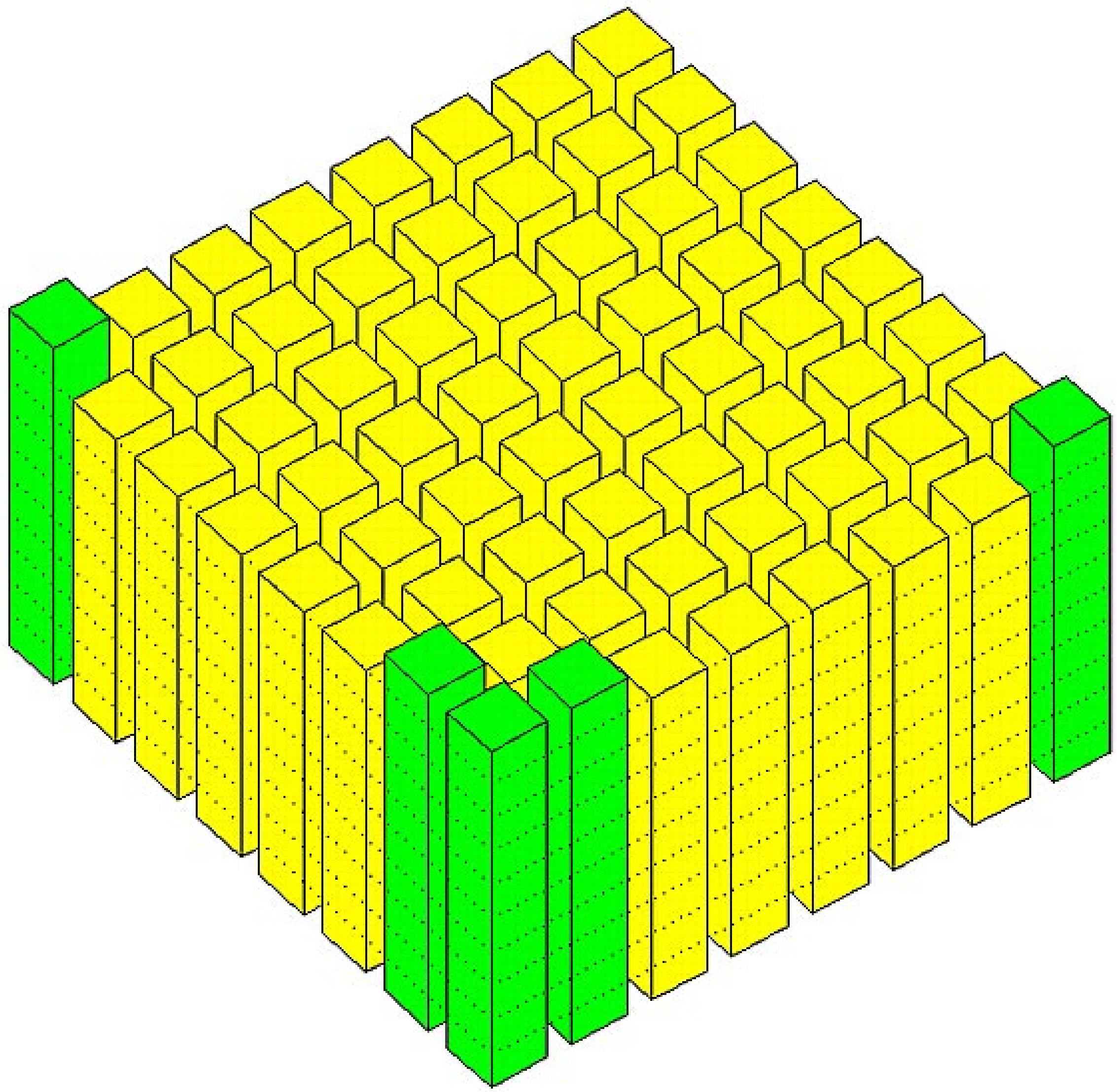
Contention Example



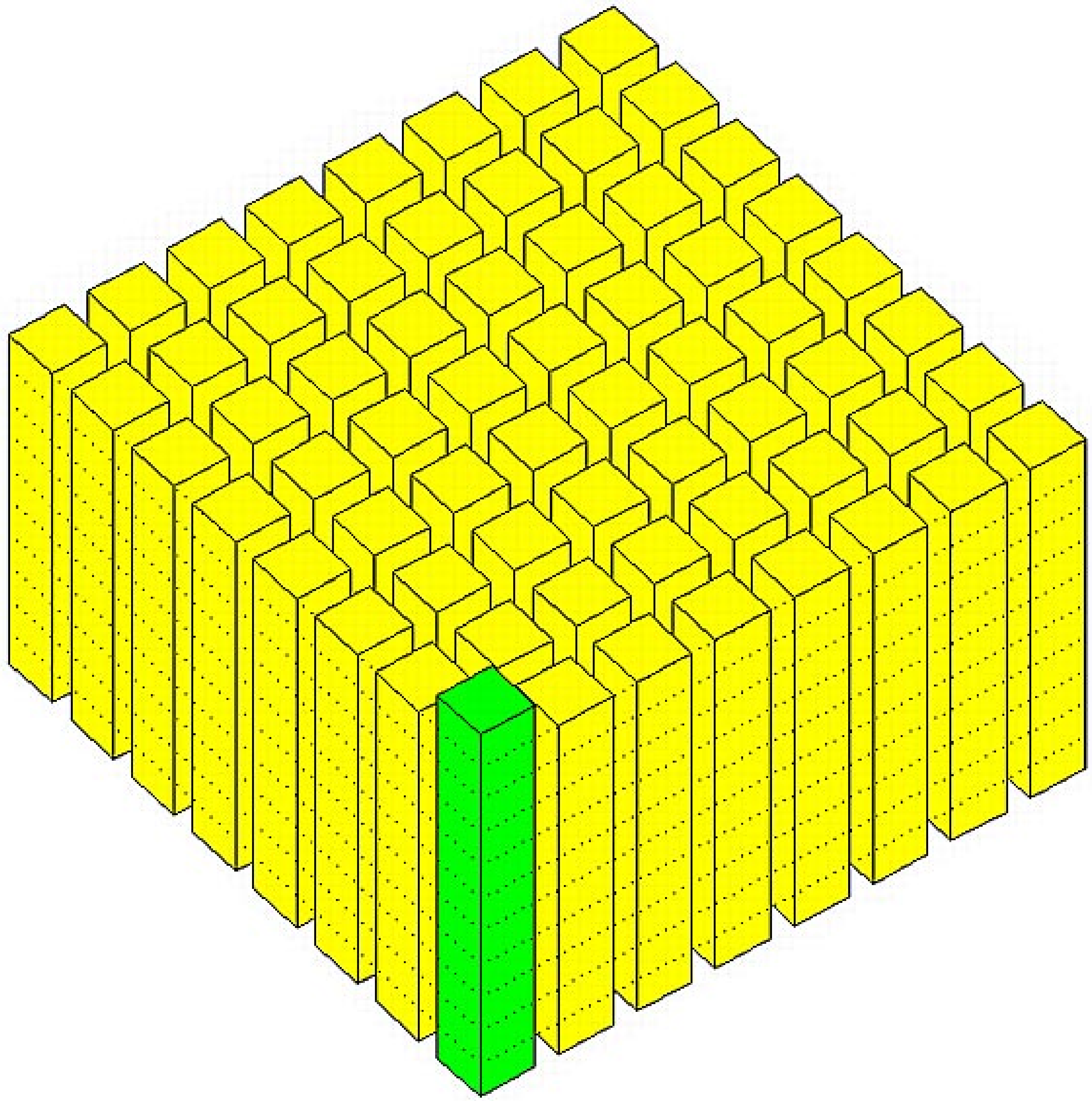
Contention Example



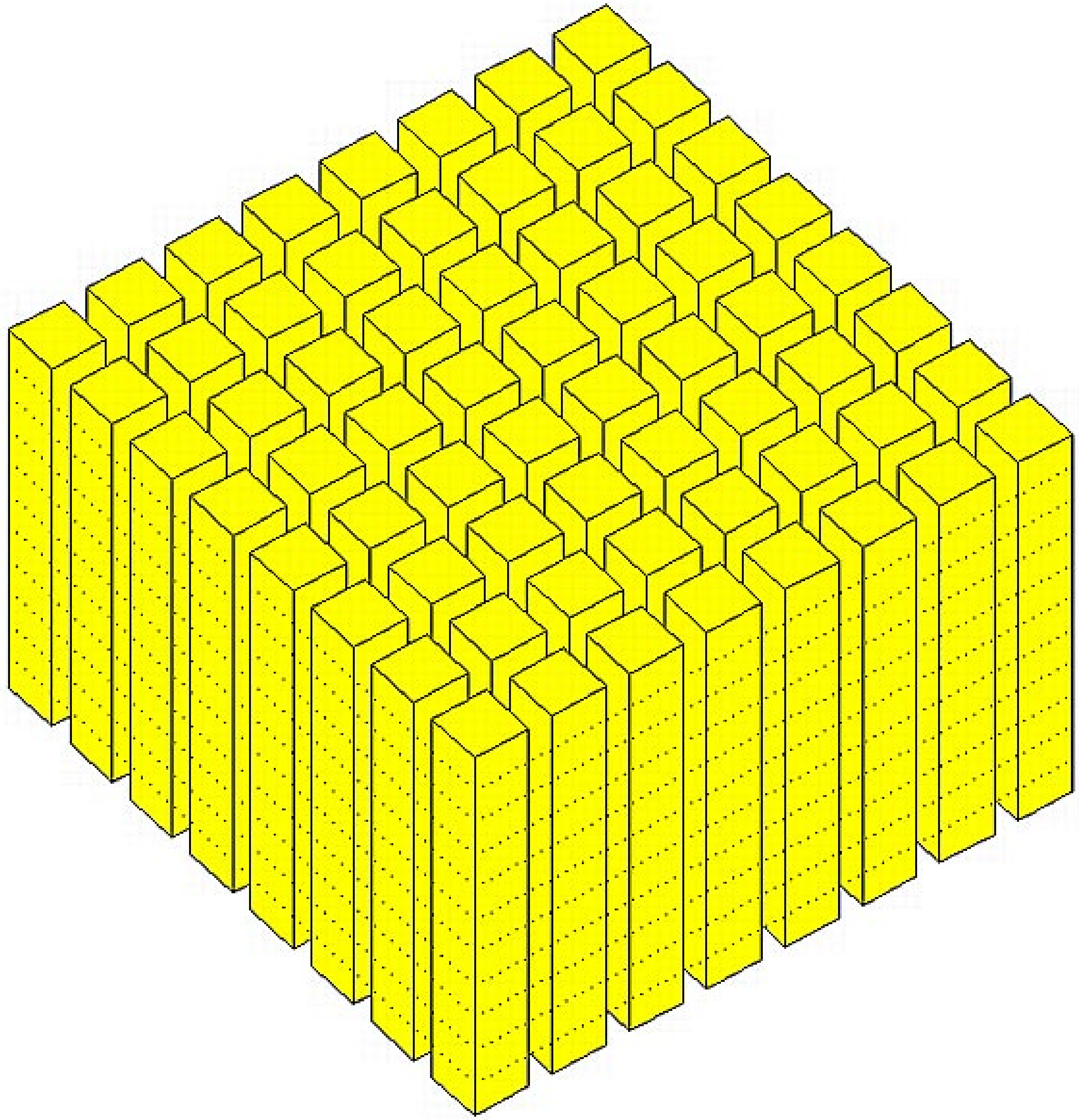
Contention Example



Contention Example



Contention Example

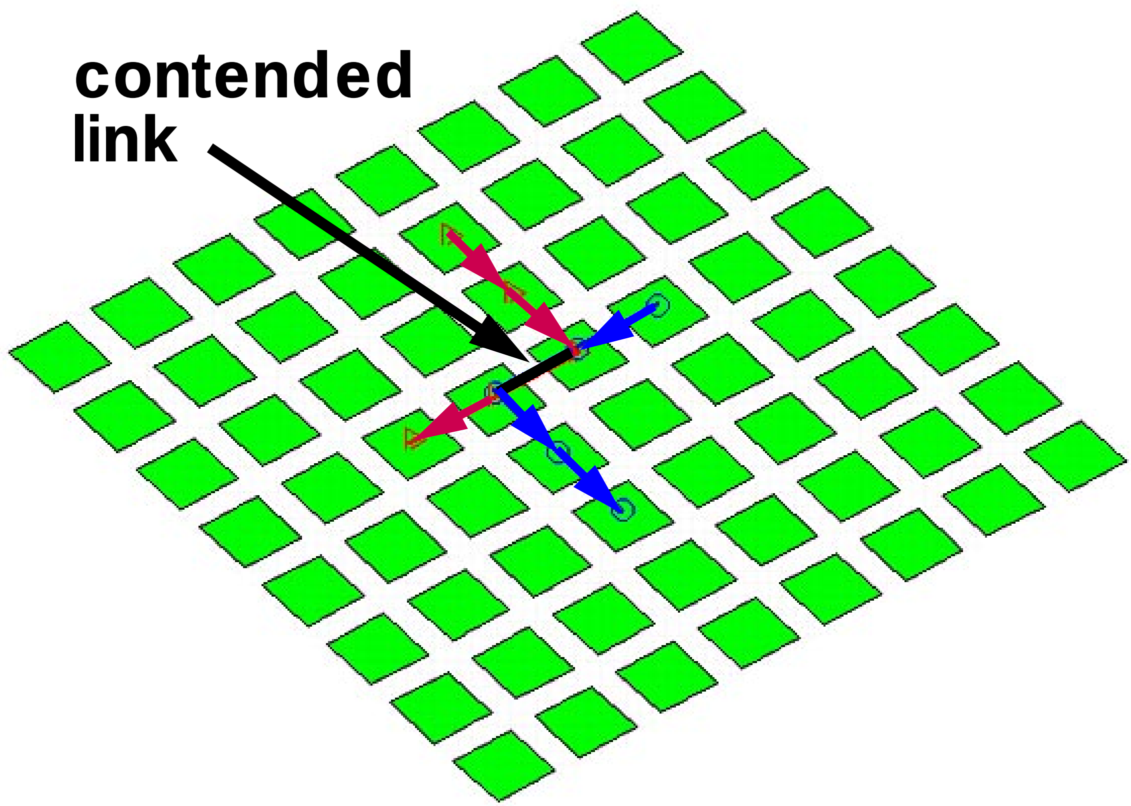


Adding Slack

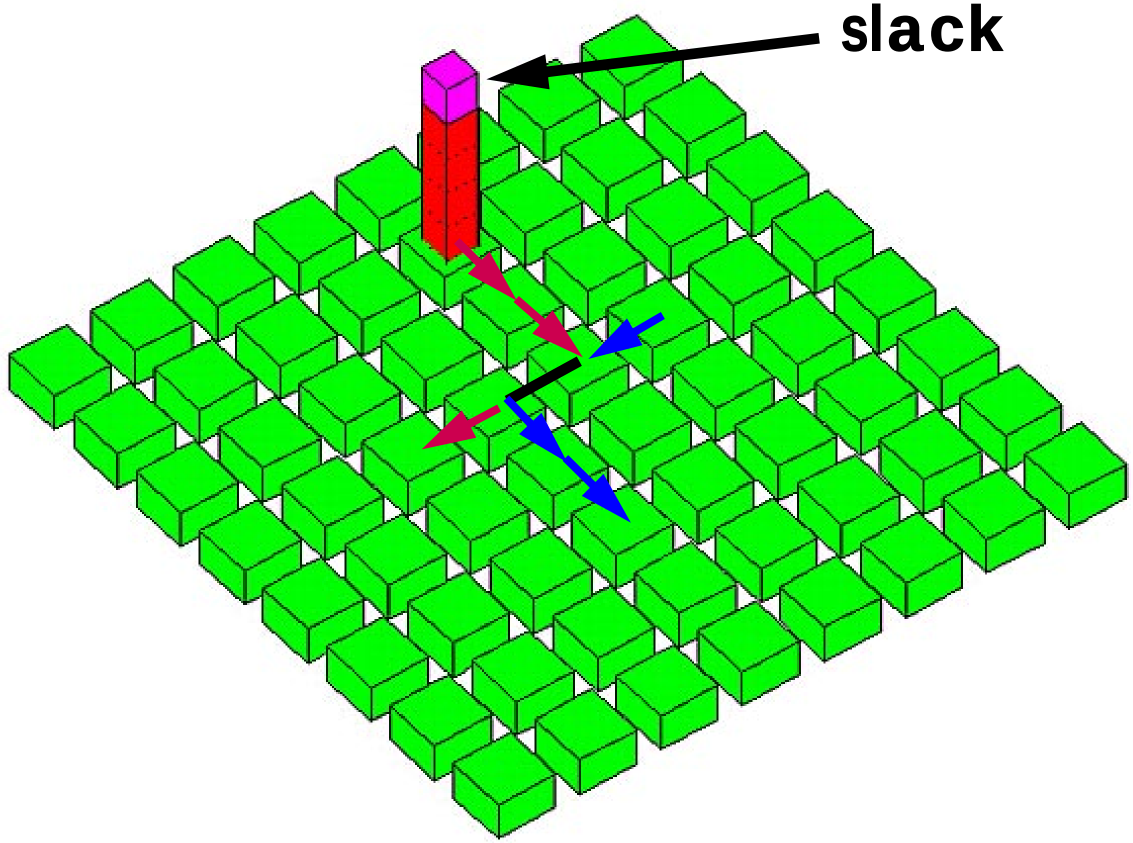
- Contention
  - GTs delayed
  - Can delay processing of other requests
  - Recursively propagates
- Contention is common
  - Avoid delaying GTs in moderate contention
  - Add *slack* to initial OTs
  - Slack: extra logical time to reach destination

Slack Example

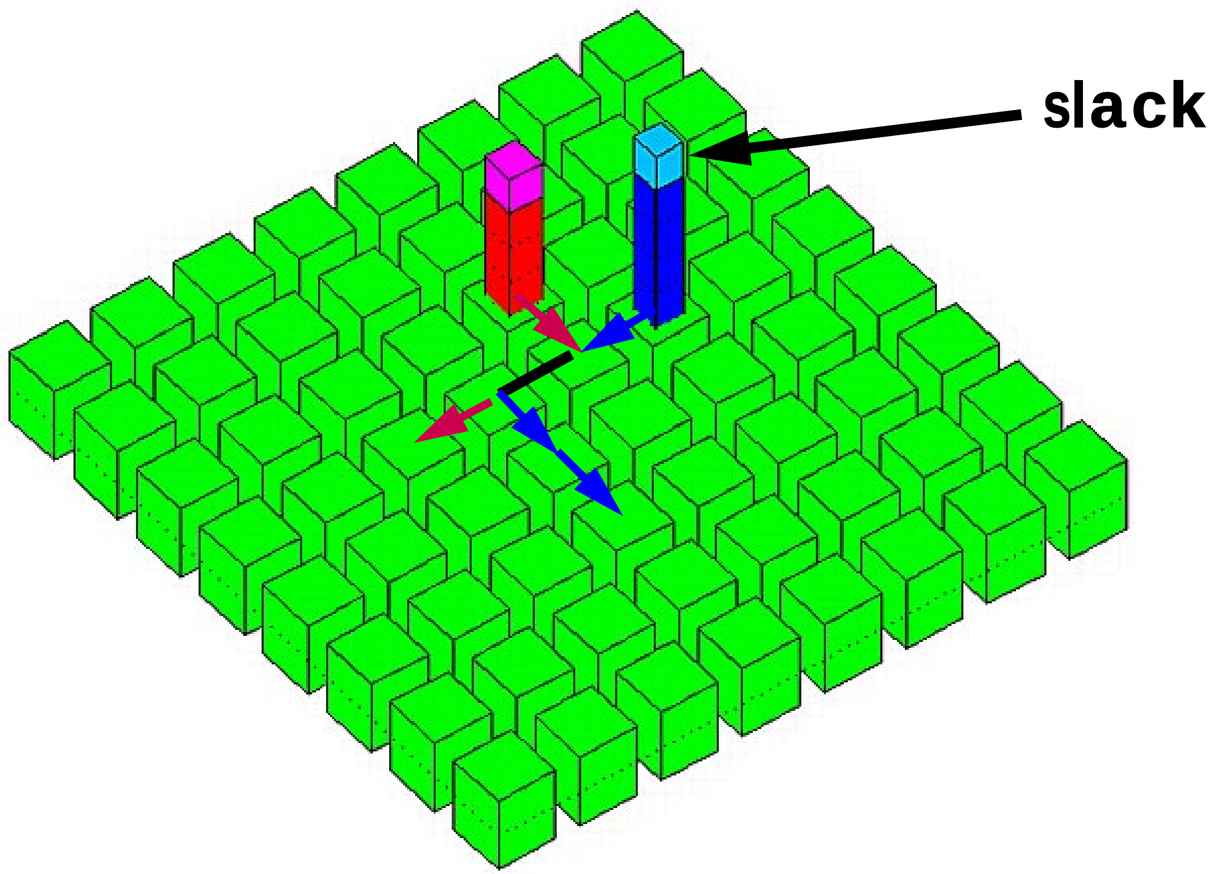
Two requests with slack



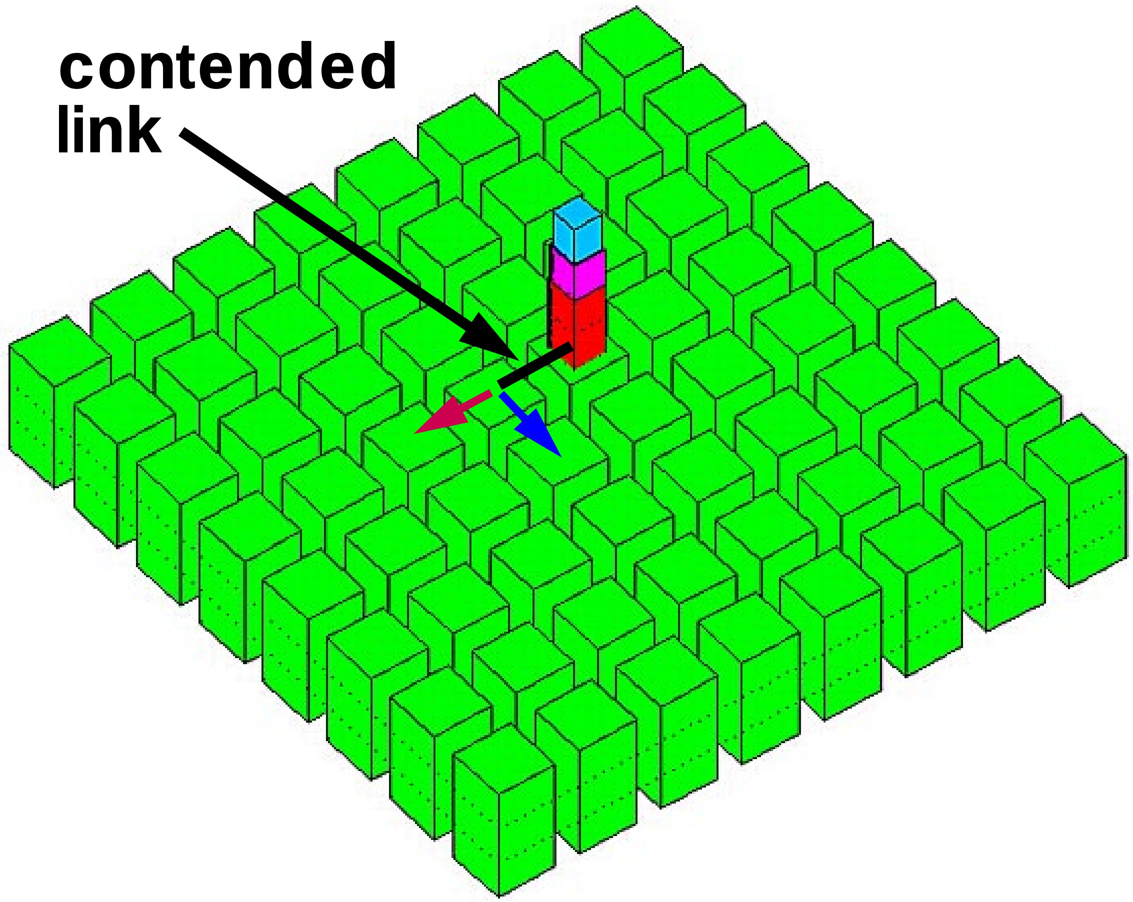
Slack Example



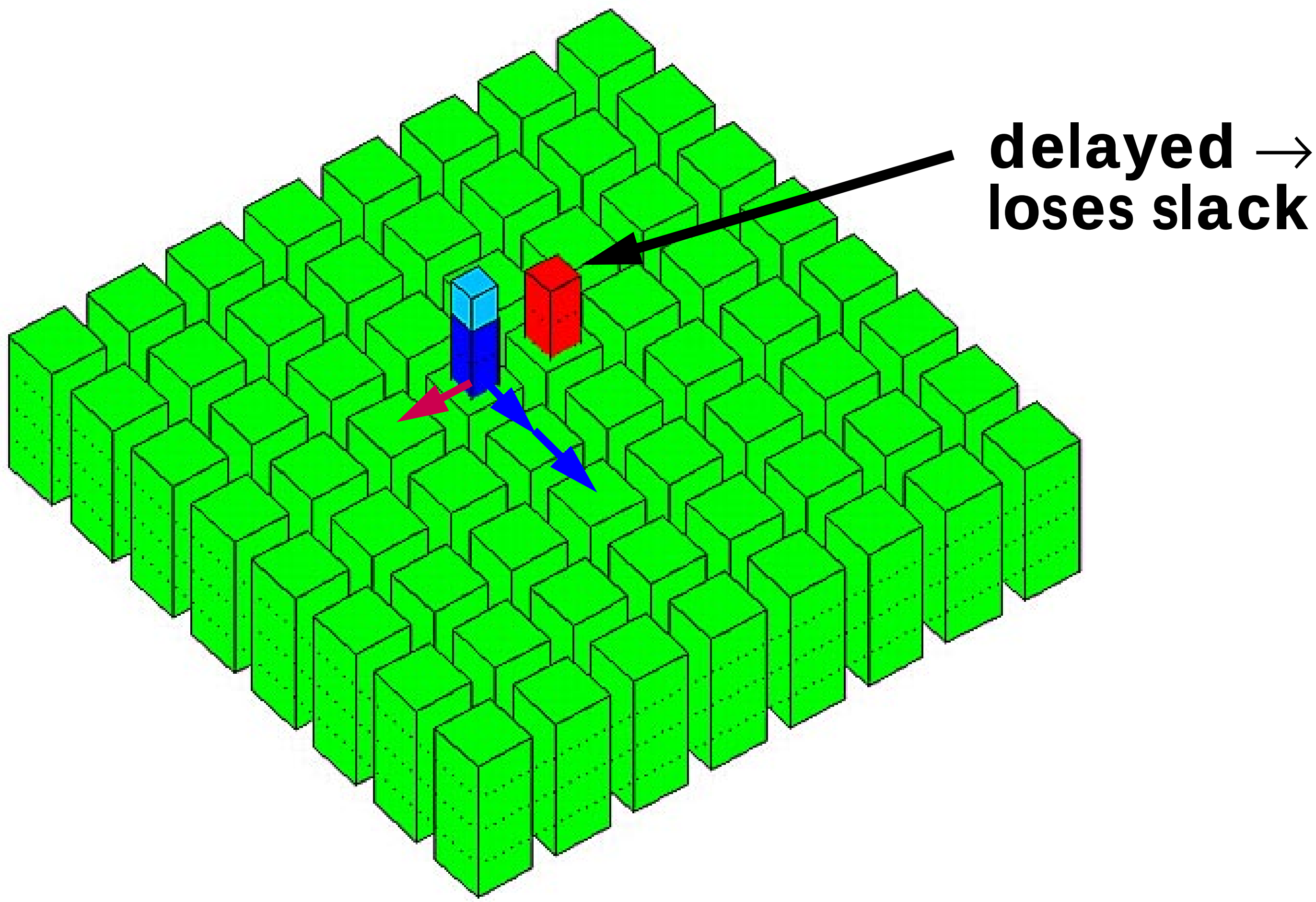
Slack Example



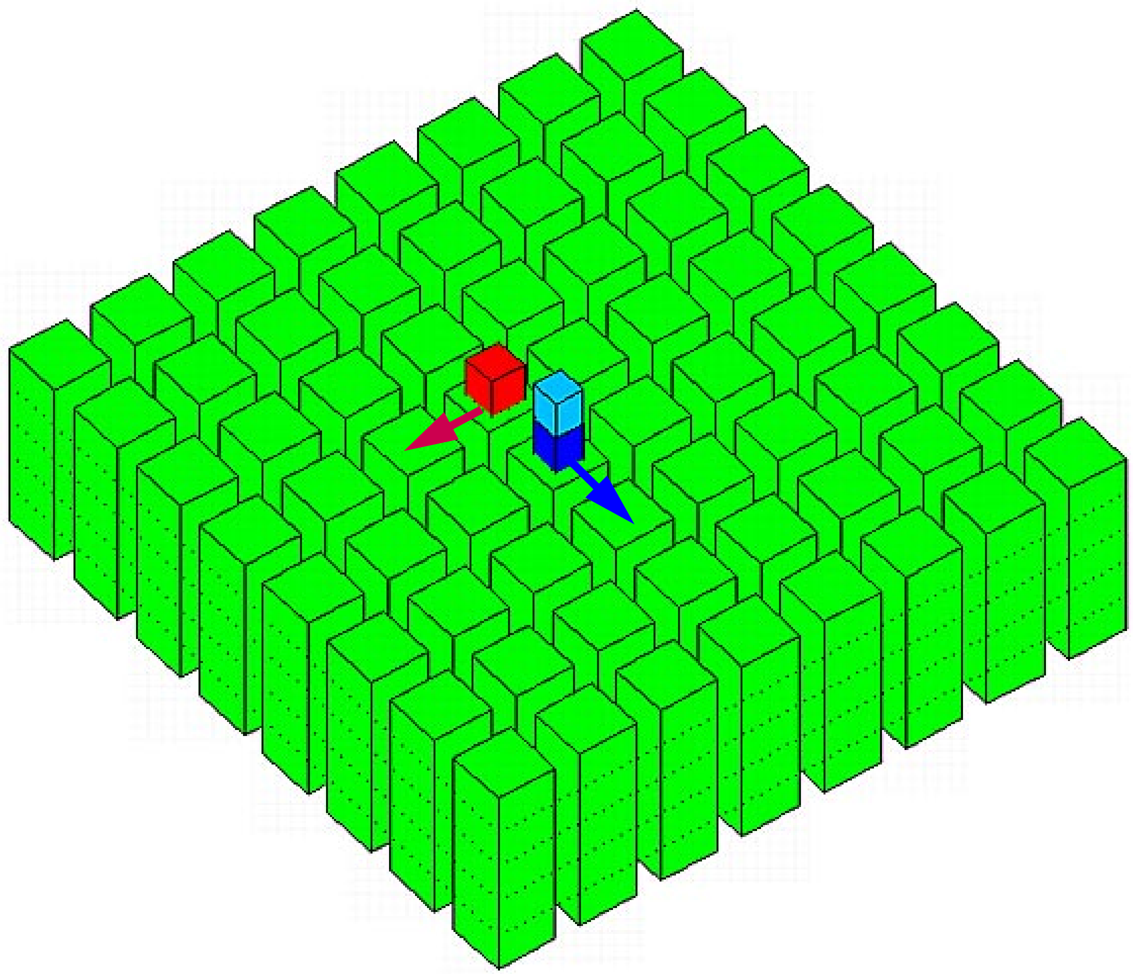
Slack Example



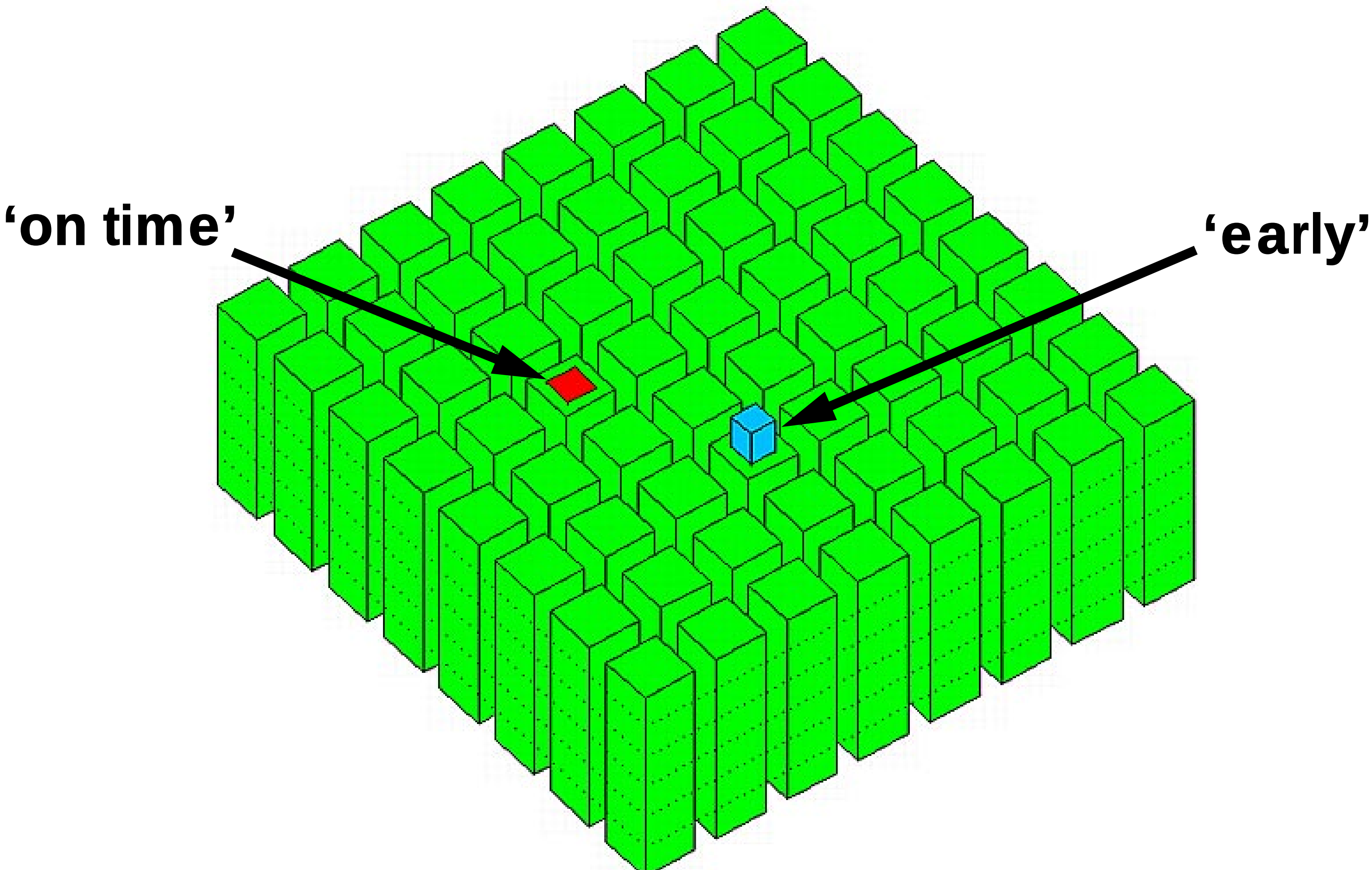
Slack Example



Slack Example



Slack Example



Avoids disruption in common cases

Implementation: Tokens

- Token passing implementation
  - Encode delta OTs and GTs implicitly
  - Extra bit per link
  - Small field per request
  - Simple algorithm in switches
- Advantages
  + Total order
  + Asynchronous
  + Variable link delay
- Disadvantages
  - Switch complexity

TOKENS ENCODE LOGICAL TIME

Timestamp Snooping Protocol

- Conventional MSI write-invalidate protocol
- Track if memory is owner
  - 1 state bit per block in memory (0.2% overhead)
  - Old idea from Synapse [Frank, 1984]
  - Avoids snoop responses
- Does not require synchronous broadcast

EXTENDS WELL-ACCEPTED SNOOPING PROTOCOLS

Outline

- Commercial Workloads
- Traditional Coherence
- Timestamp Snooping
- Evaluation
    - Workloads
    - Simulated System
    - Execution Time
    - Bandwidth
- Conclusion & Future Work

Workloads

- On-line transaction processing (OLTP)
  - IBM's DB2, TPC-C like, 400 MB in-memory DB
- Decision Support System (DSS)
  - IBM's DB2, Q12 from TPC-H, 100 MB in-memory DB
- Apache - web server
  - 8000 static files, 160 MB total
- Altavista - search engine
  - 500 MB index, 160,000 pages
- Barnes - scientific benchmark
  - 16K bodies

Simulated System

- Extended Virtutech's Simics full-system simulator
- 16 processors
- SPARC/Solaris 7
- Processor can execute 4 billion instructions/second including L1 cache misses
- Parameters
  - 4 MB, 4-way set-associative blocking L2 caches
  - 64 Byte blocks
- Vary protocol
  - Timestamp Snooping
  - DirOpt: non-blocking directory protocol
- Interconnect
  - 2D Torus (4x4)
  - Interconnect bandwidth unconstrained

Latency Assumptions

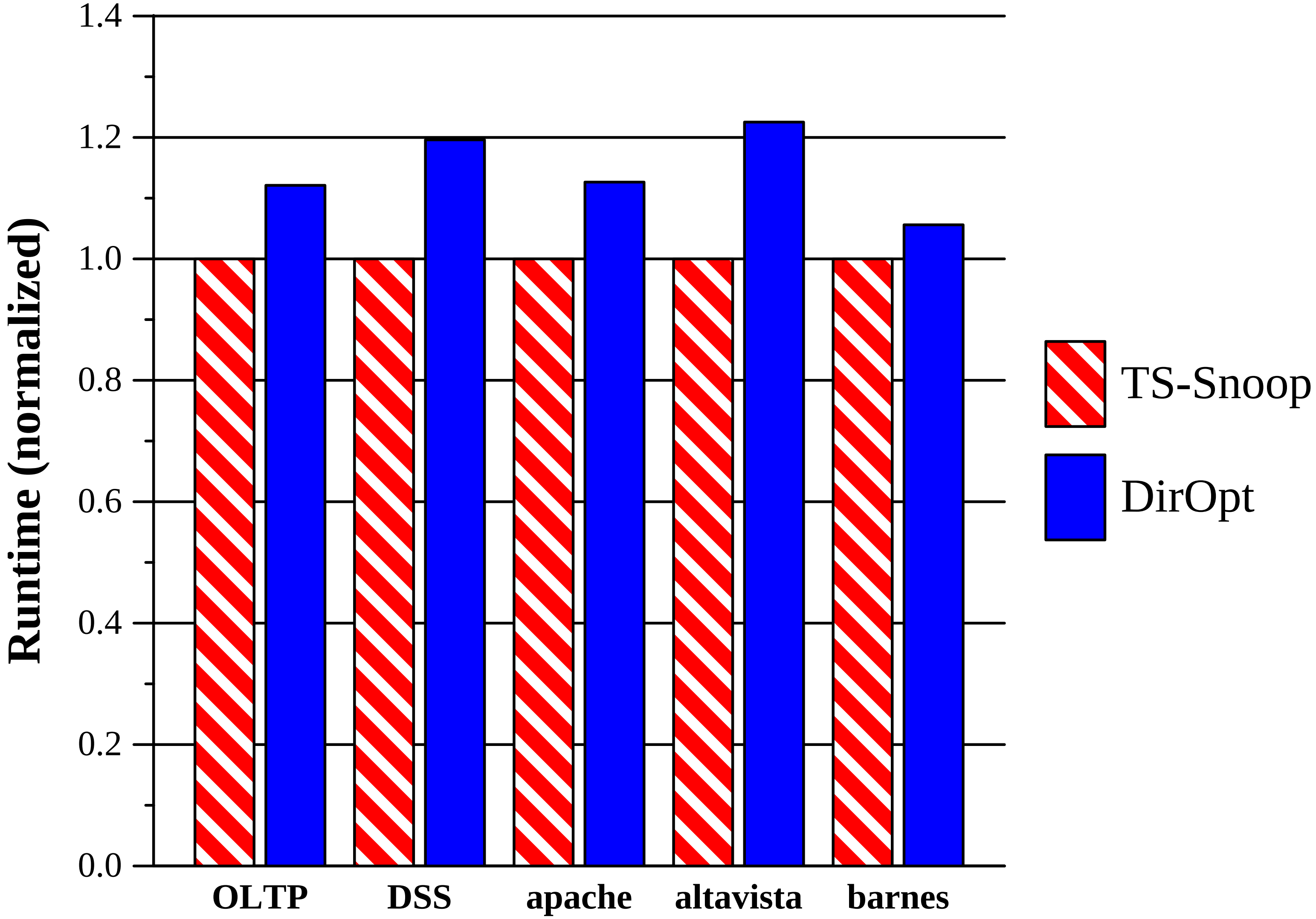
- Switch-to-switch - 15 ns
- Enter & exit network - 4 ns
- DRAM/directory access - 80 ns
- Cache SRAM access - 25 ns

	from Memory	from Cache
Directory (CC-NUMA)	2 hop + DRAM 148 ns	3 hop + directory + SRAM 207 ns
TS Snoop	2 hop + DRAM 148 ns	2 hop + SRAM 93 ns

same

2x

Execution Time Results



TIMESTAMP SNOOPING IS 6-23% FASTER THAN DIRECTORIES

Bandwidth Assumptions

- Back-of-the-envelope calculation
  - Data at memory
  - One request, one data response
  - Dependent on number of processors

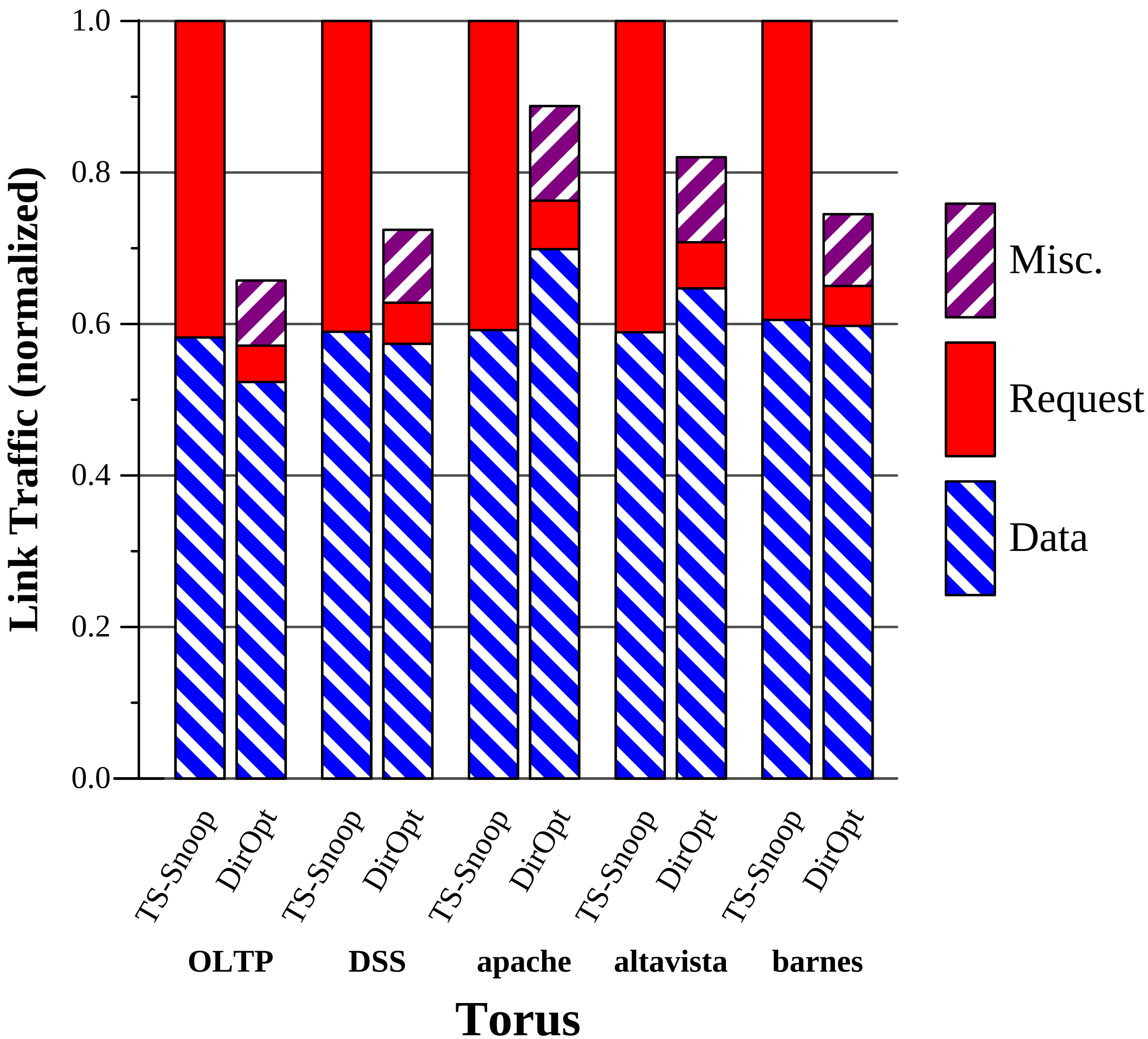
	Request	Data Response	Total
Message Size	8 Bytes	72 Bytes	
Directory (CC-NUMA)	Unicast 2 × 8 B	Unicast 2 × 72 B	= 160 B
TS Snoop	Broadcast 15 × 8 B	Unicast 2 × 72 B	= 264 B

8x

same

CONSERVATIVE ESTIMATE: DIRECTORIES 53% LESS BANDWIDTH/MISS

Bandwidth Results



DIRECTORIES USE 17-37% LESS BANDWIDTH

Conclusion

- **Comparison vs directory protocols**
  - **Efficient cache-to-cache transfers $\rightarrow$ performance advantage**
  - **Latency/bandwidth trade-off**

- **Comparison vs current SMPs**
  - **More interconnect choices**
  - **Less global communication**

- **Future work**
  - **Multicast snooping on Timestamp Snooping network**
  - **Bandwidth adaptive snooping hybrid**

http://www.cs.wisc.edu/multifacet/