NETWORK-BASED COMPUTING LABORATORY

High Performance Distributed Lock Management Services using Network-based Remote Atomic Operations

S. Narravula, A. Mamidala, A. Vishnu, K. Vaidyanathan, and D. K. Panda

Presented by Lei Chai

Network Based Computing Laboratory (NBCL)

Computer Science and Engineering

Ohio State University

OHIO STATE

Introduction

- Massive growth of parallel computing requirements
- Compute Clusters – A popular computing platform
  - Both for traditional scientific applications and data-centers
- Sharing of resources very common
  - Coordination/Synchronization of the applications
    - HPC
    - Multi-Tier Data-Centers
  - Sharing files, caches, data, etc.
- Typically managed by Lock Managers
  - Performance, Scalability and Load Resiliency – Very important!!

Presentation Roadmap

- Introduction
- Background
  - InfiniBand
  - Lock Management
- Problem Statement
- Design and Implementation
- Experimental Results
- Conclusions

InfiniBand

- Open Industry Standard based
- High Performance
  - High Bandwidth
  - Low Latencies
- Remote Direct Memory Access (RDMA) Capability
- Remote Atomic Operations
  - *Fetch and add*
  - *Compare and swap*
- *Scope for novel network based protocols and services!!*

Lock Management

- Advisory locking services
  - Logical mapping between the resources and locks
  - Application's responsibility to adhere to access restrictions
- Different lock modes
  - Shared mode locking
  - Exclusive mode locking
- Current approaches
  - Centralized Lock Managers
  - Distributed Lock Managers

Distributed Lock Manager

- Multiple nodes share the lock management responsibility
- Different dimensions of work distribution possible
  - Each server manages a set of locks
  - Multiple servers manage the work related to a single lock
  - Both
- Two-sided communication based approaches (SRSL)
  - Typically incur higher number of interrupts
    - Impact latency
- On-sided communication based approaches (DQNL) *
  - Better CPU load resiliency
  - Support for shared mode locking limited

* Distributed Queue-based Locking using Advanced Network Features, Ananth Devulapalli, Pete Wyckoff, ICPP 2005.

Problem Statement

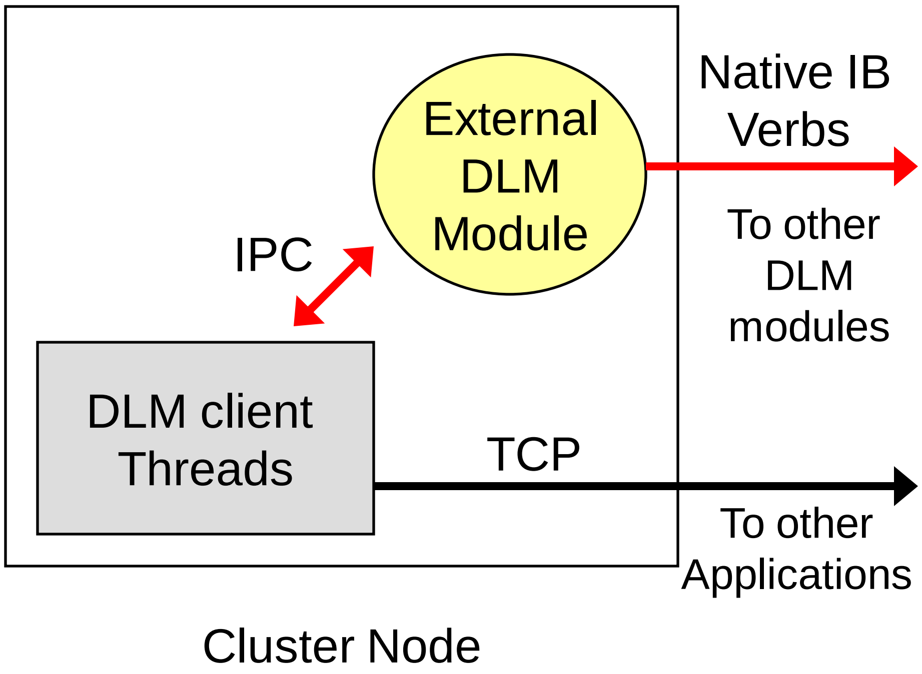
Can we design a high performance distributed lock management protocol providing efficient support for both shared mode and exclusive mode locking utilizing the one-sided network based atomic operations provided by InfiniBand in the critical path?

Presentation Roadmap

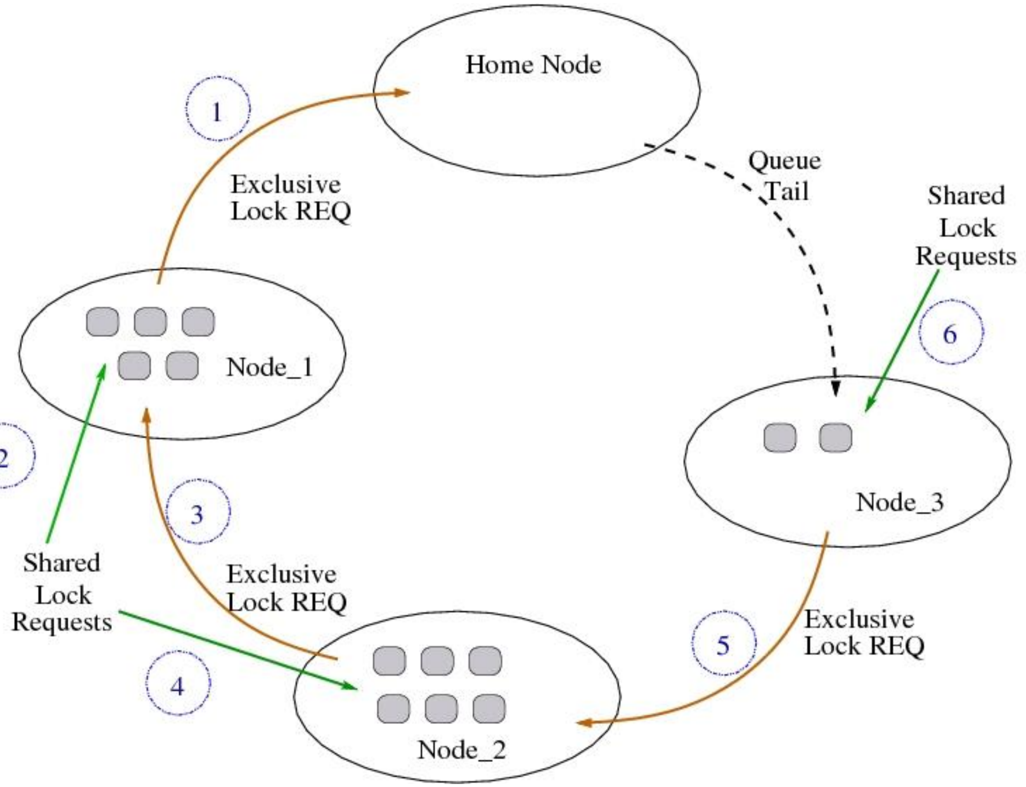
- Introduction
- Background
  - InfiniBand
  - Lock Management
- Problem Statement
- Design and Implementation
- Experimental Results
- Conclusions

Design of the Distributed Lock Manager

- Advisory locking support
  - Logical Lock -> Key
- Three possible lock states
  - Unlocked
  - Shared lock acquired
  - Exclusive lock acquired
- Distribution
  - All keys distributed evenly
- External module based design



Distributed Queue for Shared/Exclusive Locks



- Distributed Queue maintained for exclusive locks
- Shared locks queued on the nodes in the distributed queue

Basic Idea

- Use InfiniBand's Remote Atomic Operations
- Each key assigned to a "Home node"
- The home node exposes a 64 bit window for each key
  - Split into two 32-bit fields
  - Left Field -> Node representing the tail of exclusive requests
  - Right Field -> # Shared requests at the end of the queue

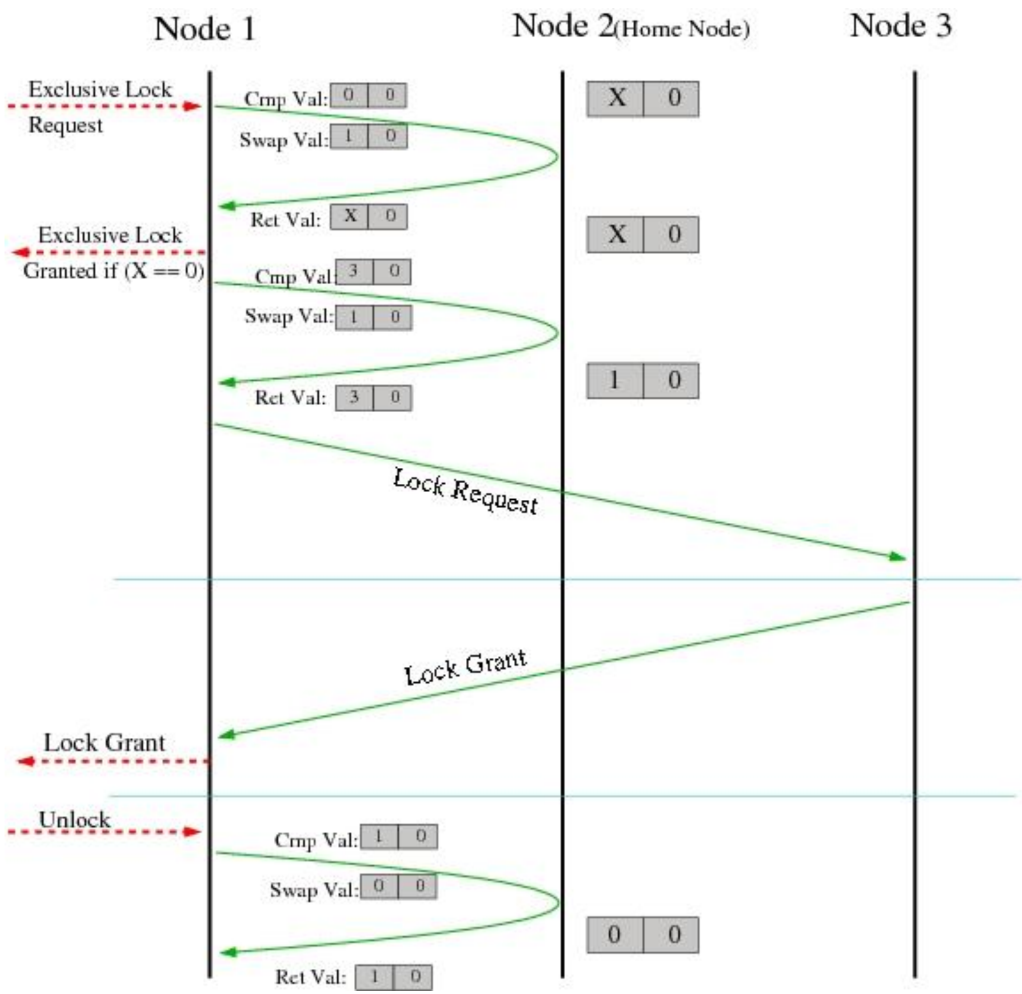
Excl	Shrd
32-bit	32-bit

- To acquire a lock the nodes perform a remote atomic operation on this 64-bit field

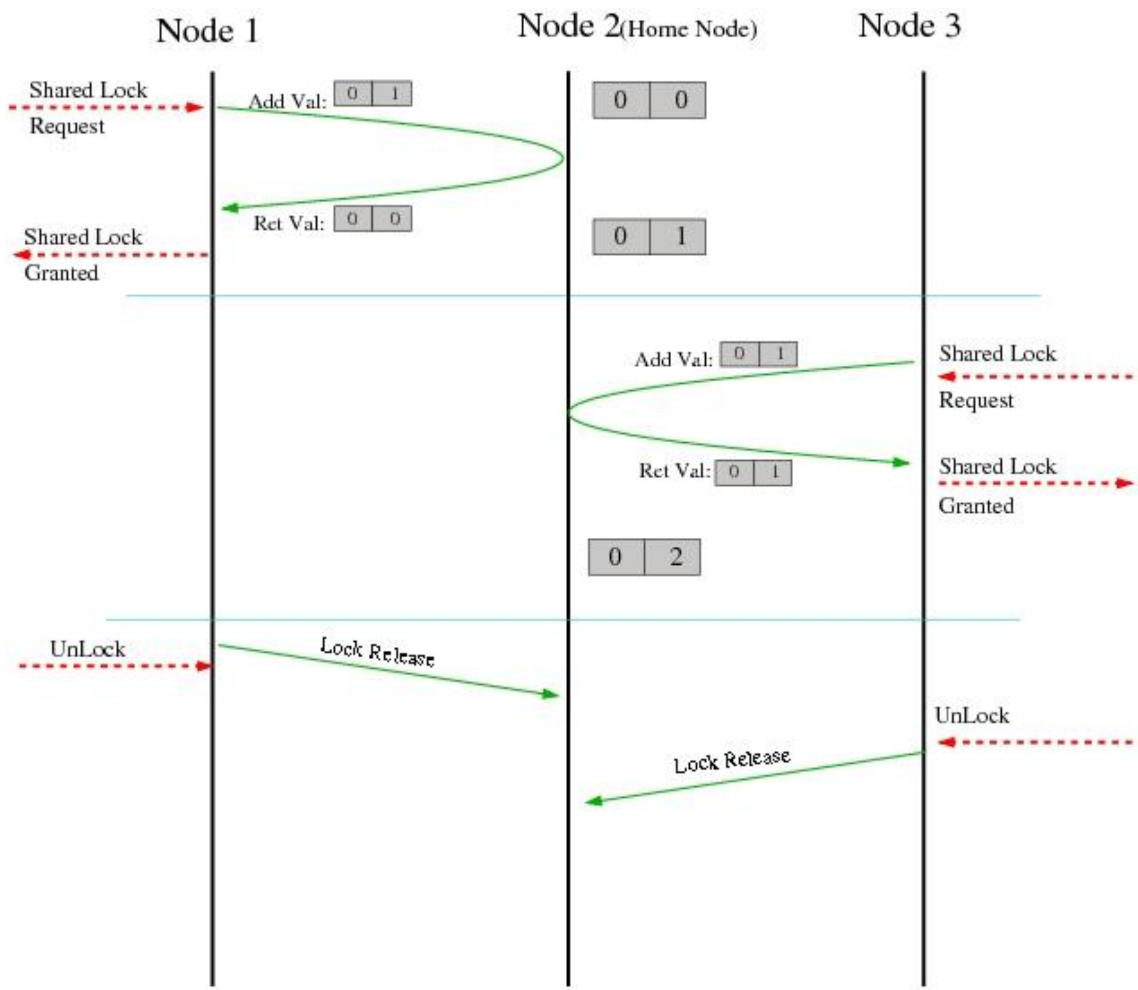
Detailed operations

- Four possible operations
  - Lock (SHRD)
  - Unlock (SHRD)
  - Lock (EXCL)
  - Unlock (SHRD)
- Possible scenarios
  - Exclusive Locking Protocol
  - Shared Locking Protocol
  - Shared Locking followed by Exclusive Locking
  - Exclusive Locking followed by Shared Locking

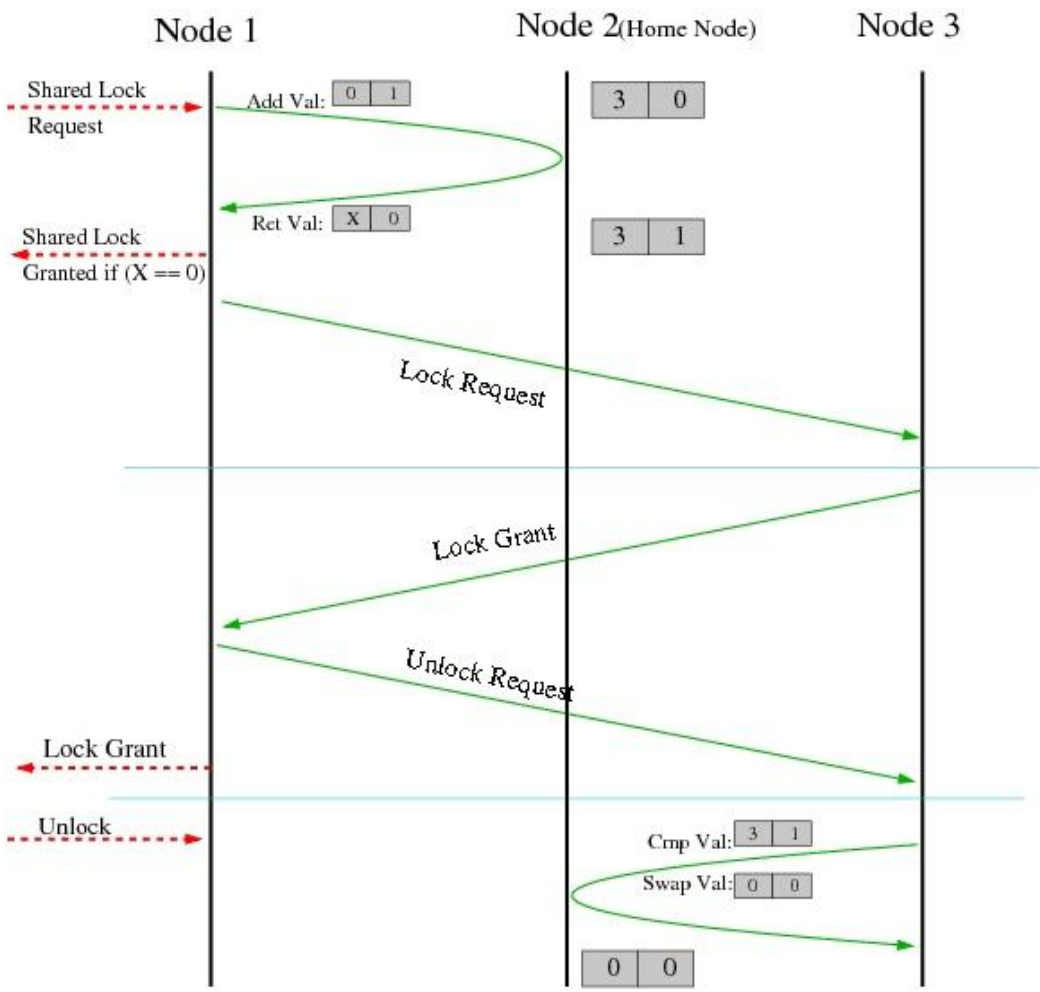
Exclusive Locking Protocol



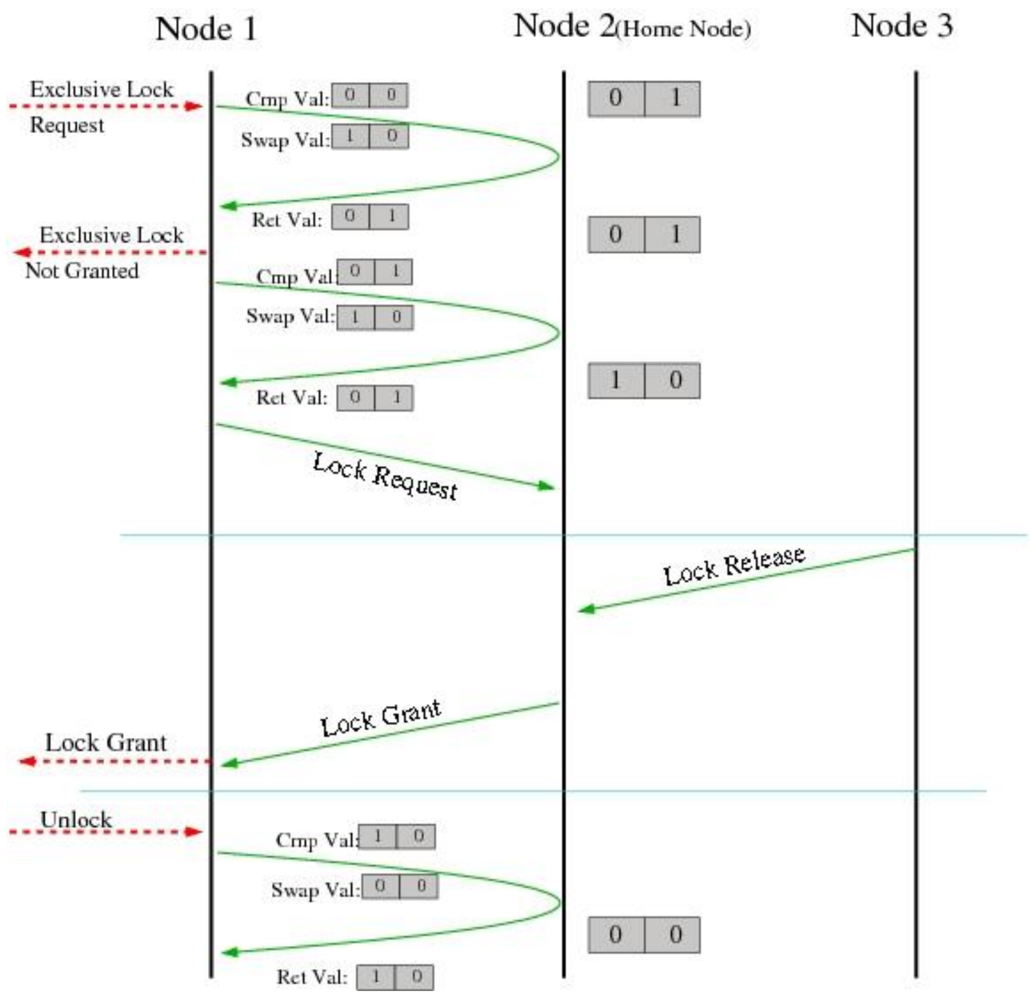
Shared Locking Protocol



Exclusive Locking followed by Shared Locking



Shared Locking followed by Exclusive Locking



Cost Models

	Lock	Unlock
SRSL	$2 * T_{Send} + 2 * T_{IPC}$	$T_{IPC\text{-}Initiate}$
DQNL	$T_{RDMAAtomic} + 2 * T_{IPC}$	$T_{IPC\text{-}Initiate}$
N-CoSED	$T_{RDMAAtomic} + 2 * T_{IPC}$	$T_{IPC\text{-}Initiate}$

Unlock latency is hidden from the process initiating the unlock and is hence constant

Presentation Roadmap

- Introduction
- Background
  - InfiniBand
  - Lock Management
- Problem Statement
- Design and Implementation
- Experimental Results
- Conclusions

Experimental Results

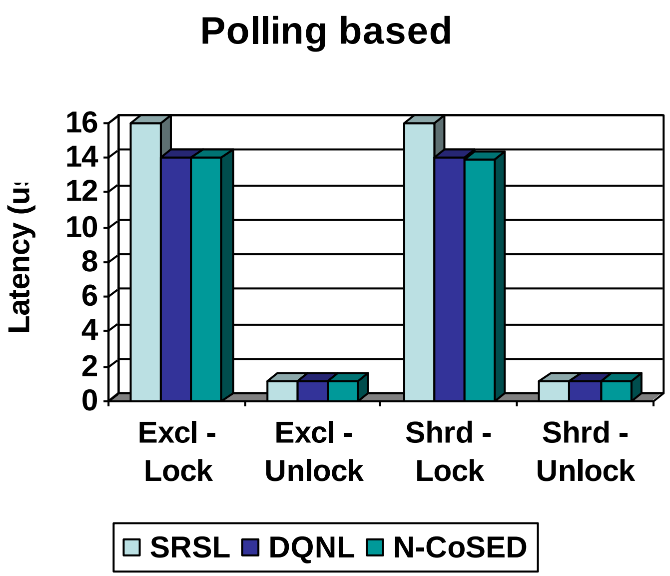
- Experimental test bed used
  - 32 node Intel Xeon (dual 3.6Ghz) Cluster
  - MT25208 HCA's
  - Flextronics 144 port DDR switch
  - OFED 1.1.1 Software Stack
- Overview
  - Network-level micro-benchmarks
  - Basic performance
  - Timing breakup of basic operations
  - Lock cascading effect

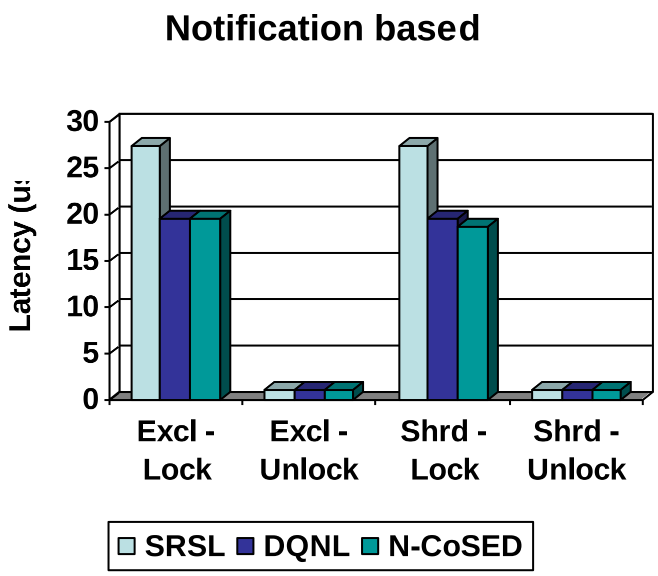
Network-level Operations Latency

	Polling (us)	Notification (us)
Send/Recv (128 B)	4.07	11.18
RDMA CS	5.78	12.97
RDMA FA	5.77	12.96

- Polling Mechanism
  - Scenarios requiring very low latencies
  - Scenarios that can afford to spend CPU time polling
- Notification Mechanism
  - Typical data-center scenarios

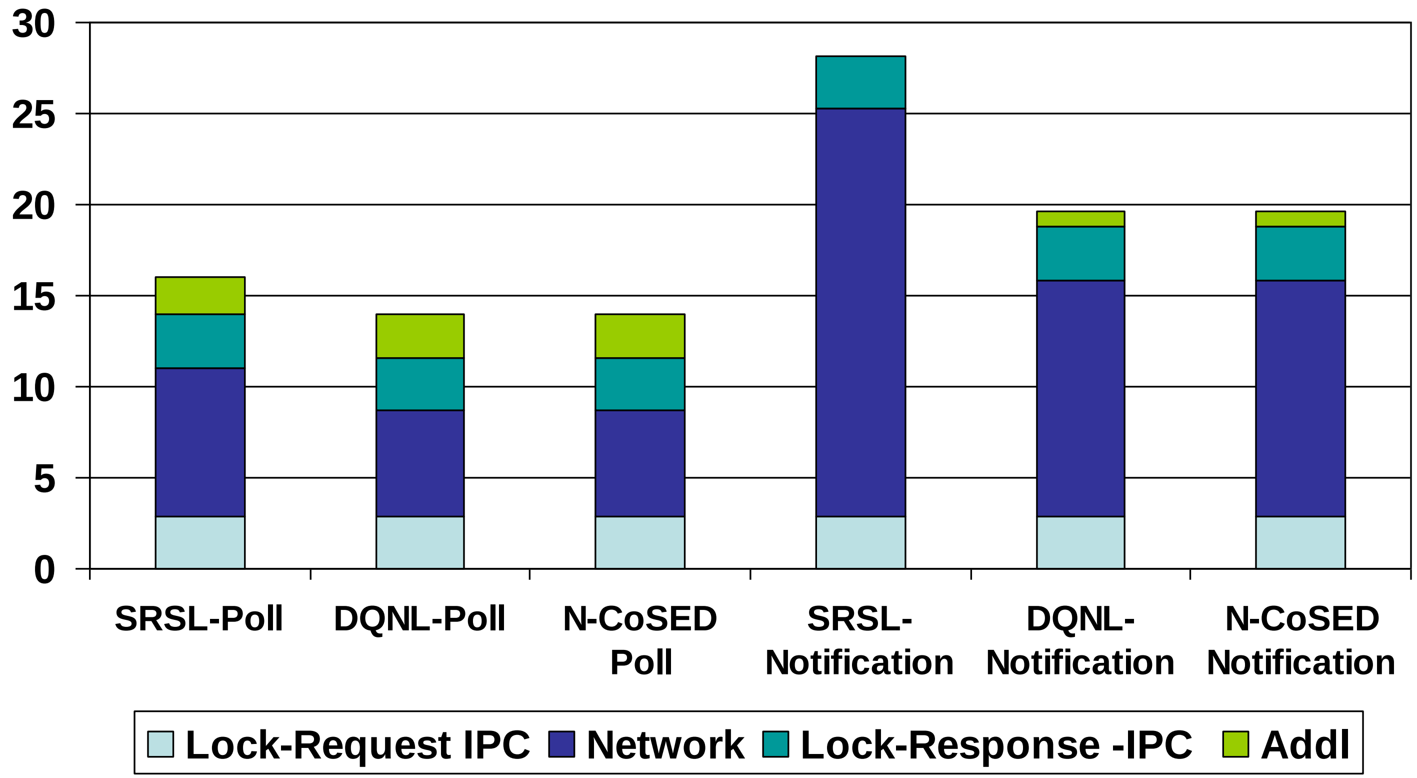
Basic Performance





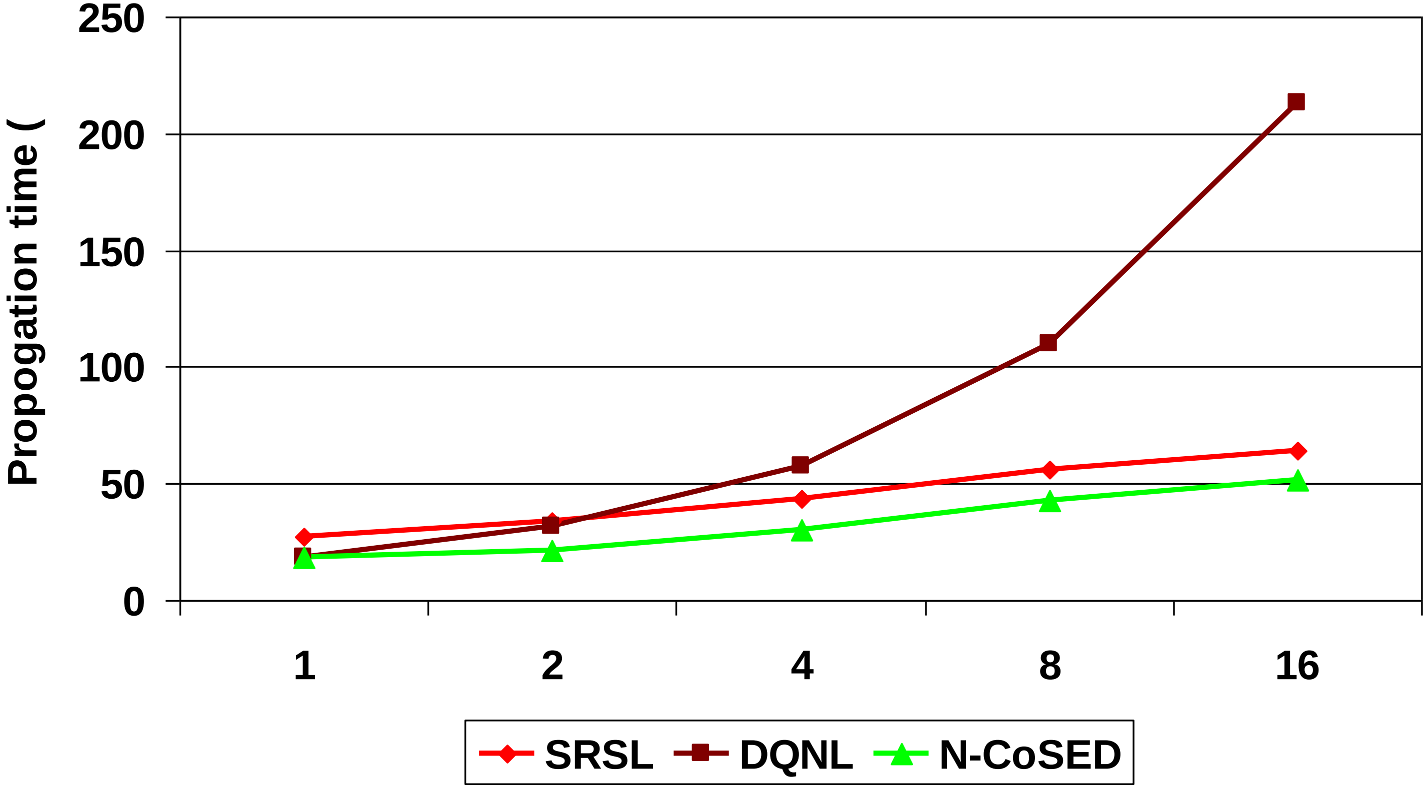
Under ideal conditions DQNL and N-CoSED lock latencies are lower than the SRSL case

Breakup of Basic Performance



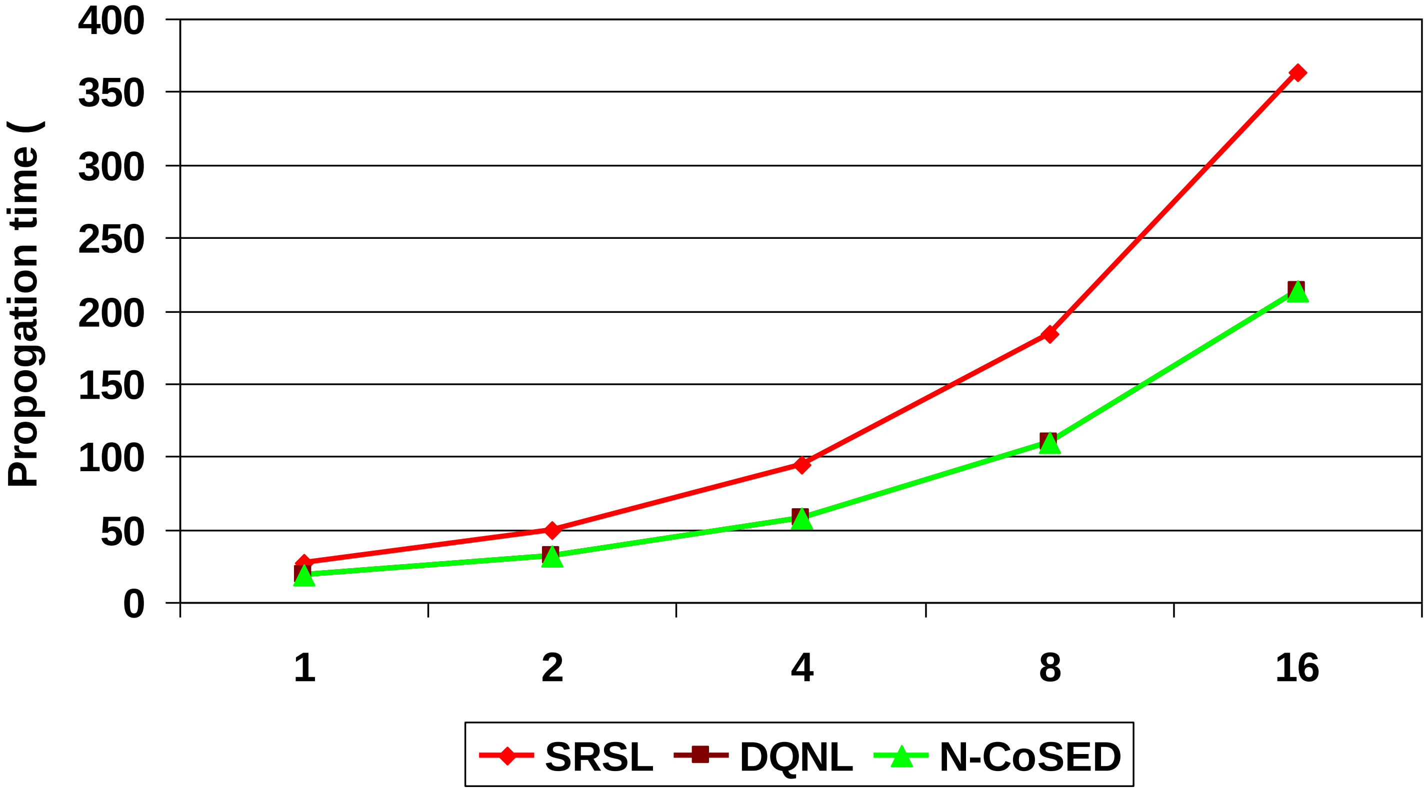
The SRSL schemes clearly show higher network times owing to the extra network message

Shared Lock Propagation



- DQNL basic queuing mechanism ends up with serial unlock operations
- SRSL incurs the constant overhead of an extra message over N-CoSED
- N-CoSED performs the best in all cases

Exclusive Lock Propagation



- DQNL and N-CoSED show identical performance
- SRSL incurs the aggregated overhead of an extra message for each unlock

NETWORK-BASED COMPUTING LABORATORY

Presentation Roadmap

- Introduction
- Background
  - InfiniBand
  - Lock Management
- Problem Statement
- Design and Implementation
- Experimental Results
- Conclusions and Future Work

OHIO STATE

Conclusions and Future Work

- One sided Distributed Locking Protocol based on InfiniBand’s RMA operations
- Performance benefits
- Good distribution of lock management work
- Future Work
  - Extend to starvation free designs
  - Investigate use of programmable NIC’s provided by other modern interconnects

Acknowledgements

Our research is supported by the following organizations

- Current Funding support by

- Current Equipment support by

Questions?

Web Pointers

http://nowlab.cse.ohio-state.edu

{narravul, mamidala, vishnu, vaidyana, panda}
@ cse.ohio-state.edu