On the Provision of Prioritization and Soft QoS in Dynamically Reconfigurable Shared Data-Centers over InfiniBand

P. Balaji, S. Narravula, K. Vaidyanathan, H. –W. Jin and D. K. Panda

Network Based Computing Laboratory (NBCL)

The Ohio State University

Presentation Outline

- **Introduction and Motivation**
- Overview of Dynamic Reconfigurability over InfiniBand
- Issues with Basic Dynamic Reconfigurability
- Dynamic Reconfigurability with Prioritization and Soft QoS
- Experimental Results
- Conclusions and Future Work

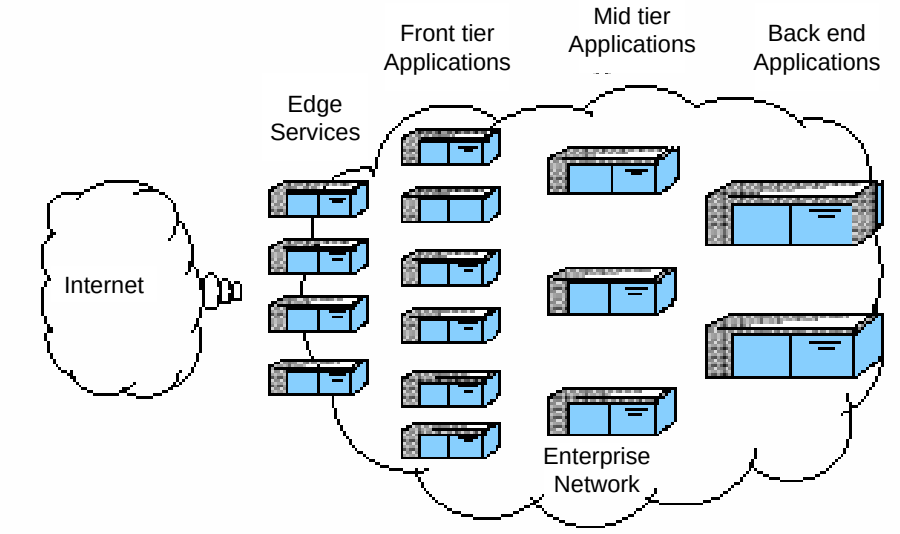
COTS Clusters

- Commodity-Off-the-Shelf (COTS) Clusters
  - High Performance-to-Cost Ratio
  - Enabled through High Performance Networks
- Advent of High Performance Networks
  - Ex: InfiniBand, Myrinet, Quadrics, 10-Gigabit Ethernet
  - High Performance Protocols: VAPI / IBAL, GM, EMP
    - Provide applications direct and protected access to the network
- InfiniBand: An Industry Standard High Performance Network Architecture
  - Low latency (< 4us) and high throughput (near wire speed = 10Gbps)
  - Offloaded Protocol Stack, Zero-copy data transfer, One-sided communication (RDMA read/write, atomics, etc)

InfiniBand-based COTS Clusters are becoming extremely popular !

Cluster-based Data-Centers

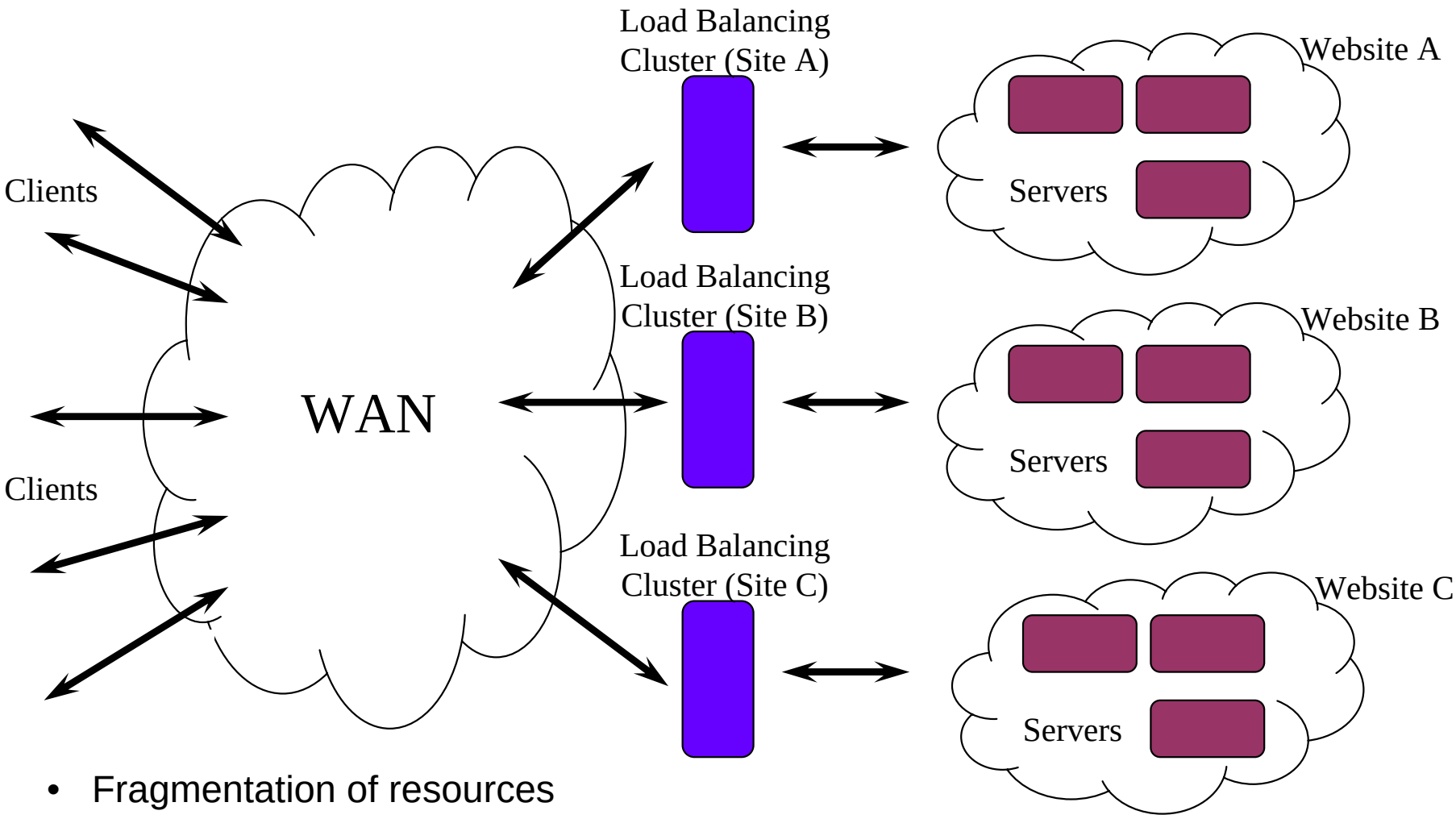
- Increasing adoption of Internet
  - Primary means of electronic interaction
  - Highly Scalable and Available Web-Servers: Critical !
- Utilizing Clusters for Data-Center environments?
  - Studied and Proposed by the Industry and Research communities



(Courtesy CSP Architecture Design)

- Nodes are logically partitioned
  - Interact depending on the query
  - Provide services requested
- Load increases in the inner tiers

Shared Multi-Tier Data-Centers



- Fragmentation of resources
- Service differentiation
- QoS guarantees

Objective

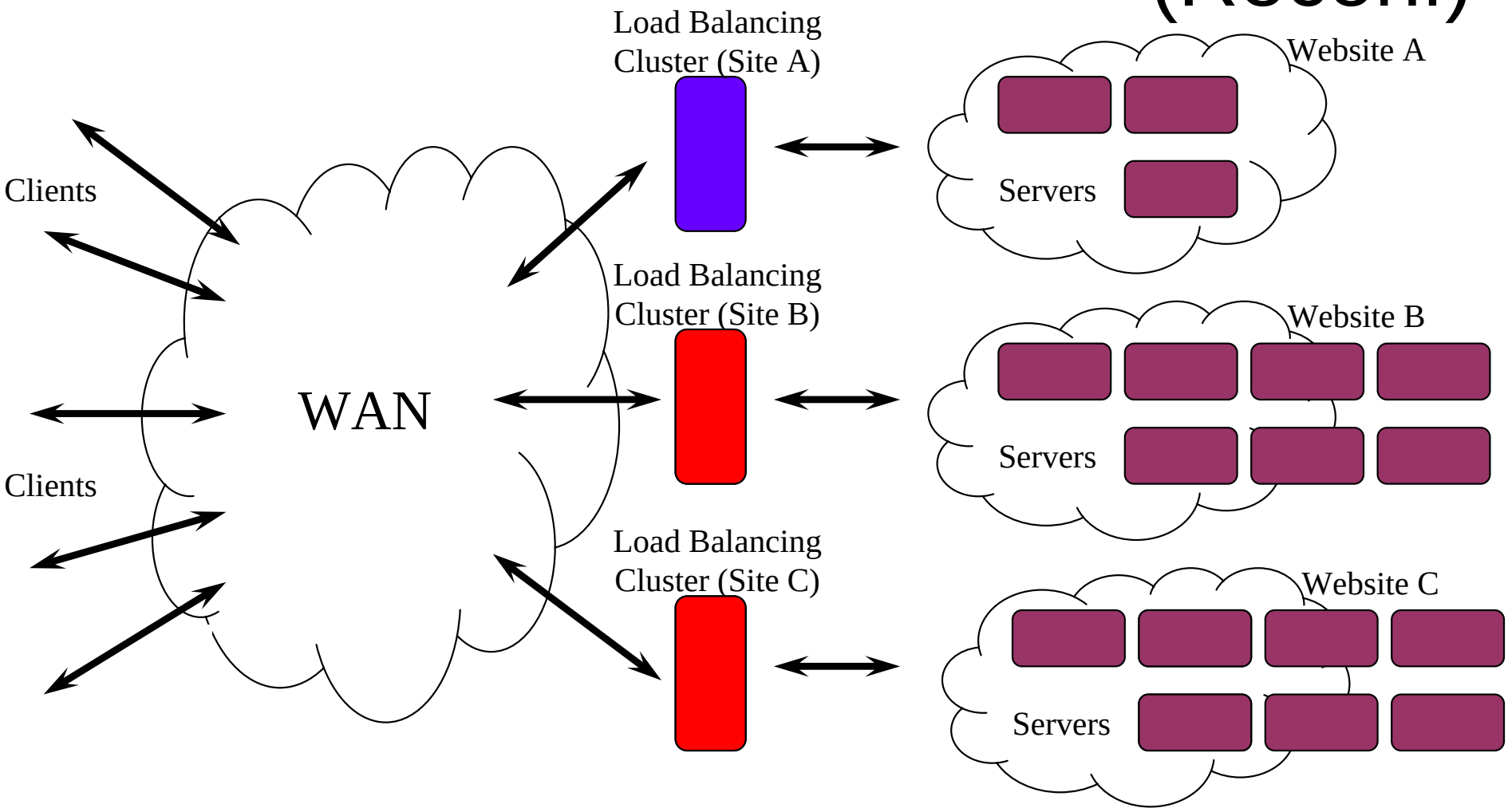
- Fragmentation of resources needs to be curbed [balaji04_reconf]
  - Dynamically configuring nodes allotted to each service
- Service differentiation for different websites hosted
  - Intelligent dynamic reconfiguration based on pre-defined prioritization rules
- QoS guarantees for low-priority requests
  - Ensure that low priority websites are given certain minimal resources at all times
- *Can the advanced features provided by InfiniBand help in providing dynamic reconfigurability with QoS and prioritization for different websites?*

balaji04_reconf: "Exploiting Remote Memory Operations to Design Efficient Reconfiguration for Shared Data-Centers over InfiniBand". P. Balaji, K. Vaidyanathan, S. Narravula, S. Krishnamoorthy, H. –W. Jin and D. K. Panda. In the RAIT workshop, held in conjunction with Cluster 2004.

Presentation Outline

- Introduction and Motivation
- **Overview of Dynamic Reconfigurability over InfiniBand**
- Issues with Basic Dynamic Reconfigurability
- Dynamic Reconfigurability with Prioritization and Soft QoS
- Experimental Results
- Conclusions and Future Work

Basic Dynamic Reconfigurability (Reconf)



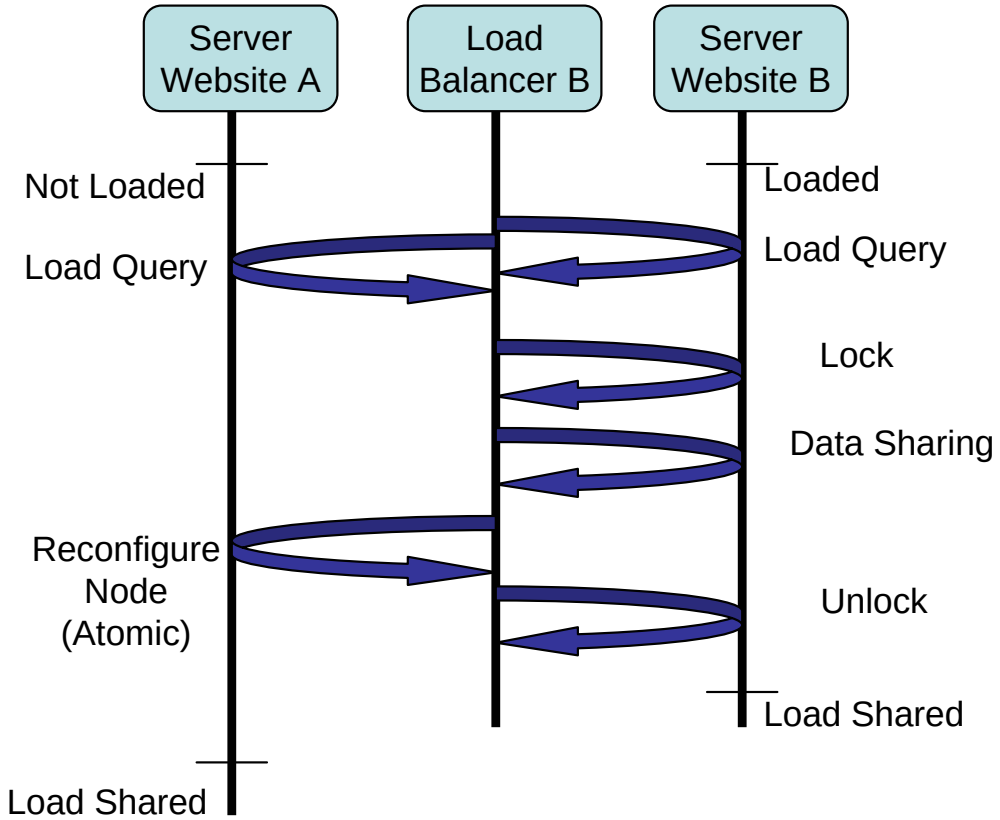
Nodes reconfigure themselves to highly loaded websites at run-time

Reconf Design

- Support for Existing Applications
  - Utilizing *External Helper Modules* (external programs running on each node) to take care of load monitoring, reconfiguration, etc.
- Load-Balancer based vs. Server based Reconfiguration
- Remote Memory Operations based Design
  - Locking and Data Sharing are based on InfiniBand one-sided operations and atomics
  - Load-balancers remotely monitor and reconfigure the system

Utilizing InfiniBand Features

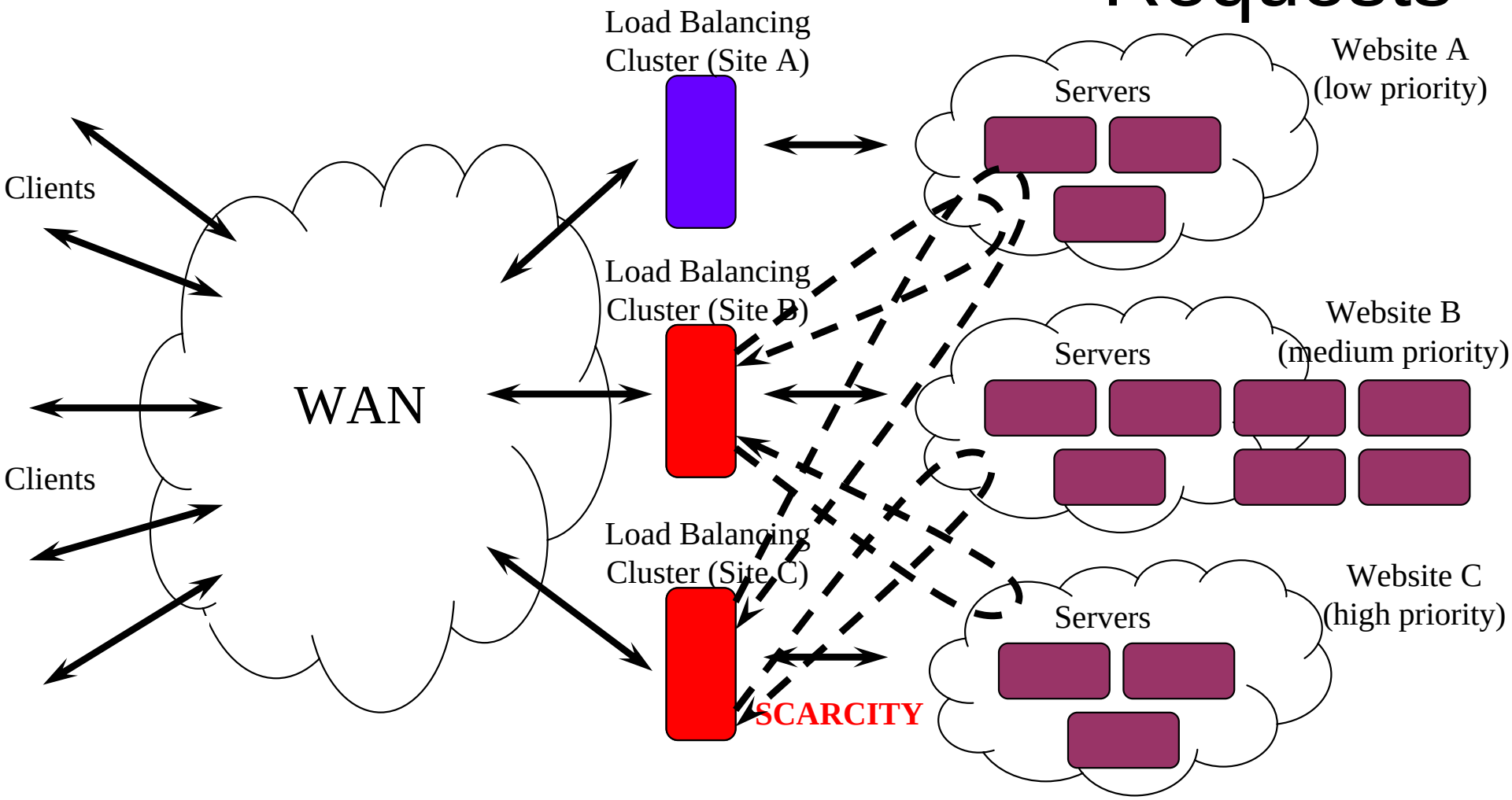
- Two-level hierarchical locking mechanism
  - Both locks performed remotely using InfiniBand Atomic Operations
- Completely load resilient design



Presentation Outline

- Introduction and Motivation
- Overview of Dynamic Reconfigurability over InfiniBand
- **Issues with Basic Dynamic Reconfigurability**
- Dynamic Reconfigurability for Prioritization and Soft QoS
- Experimental Results
- Conclusions and Future Work

Issues with Reconf on High Priority Requests



High Priority website may get lesser number of servers compared to medium/low priority websites since Reconf does not have any idea about Prioritization between websites

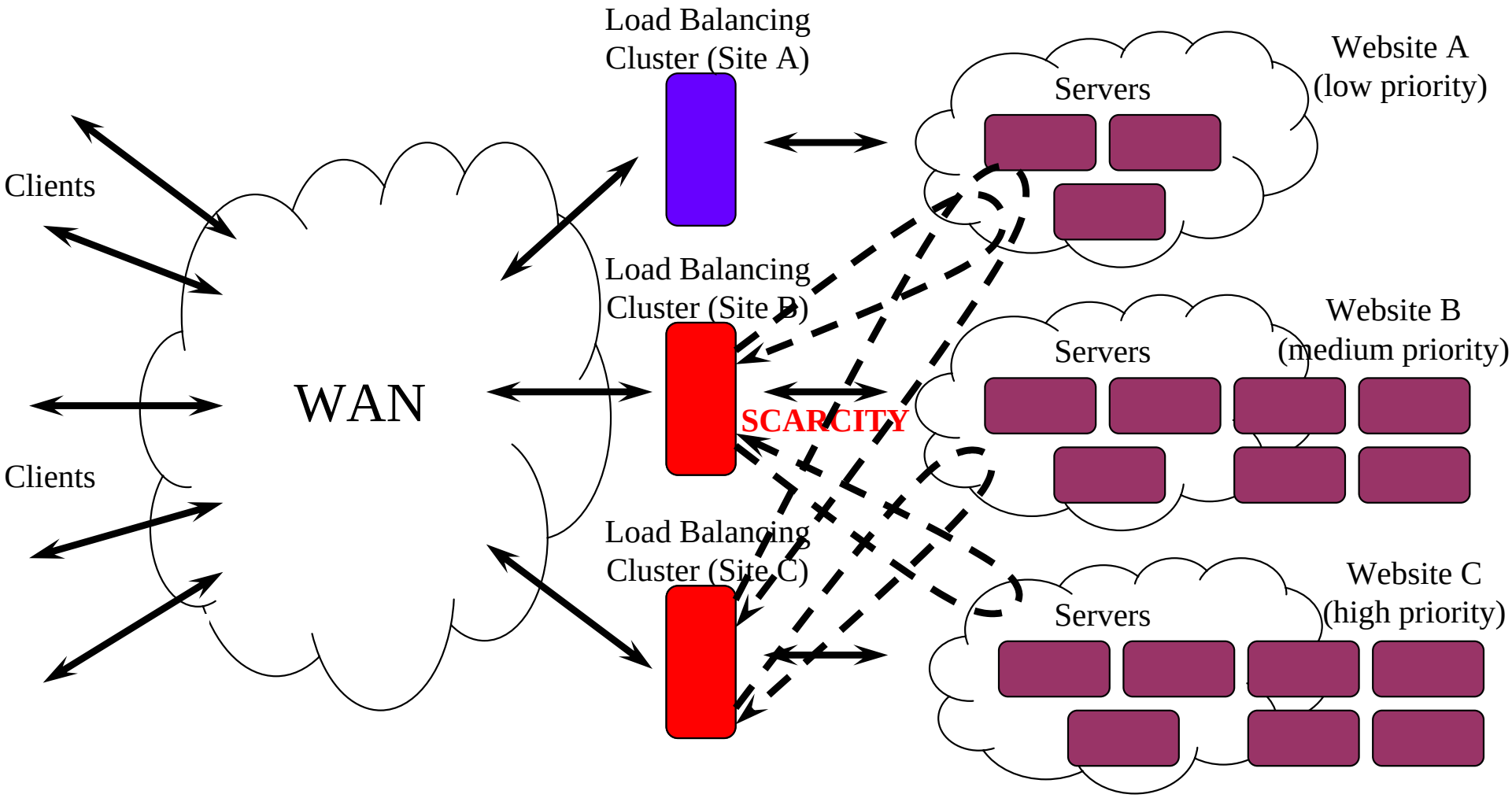
Presentation Outline

- Introduction and Motivation
- Overview of Dynamic Reconfigurability over InfiniBand
- Issues with Basic Dynamic Reconfigurability
- **Dynamic Reconfigurability for Prioritization and Soft QoS**
- Experimental Results
- Conclusions and Future Work

Dynamic Reconfigurability with Prioritization (Reconf-P)

- Prioritization support for Reconf
  - Reconf requires additional logic to be priority aware
  - Pre-defined rules for prioritization amongst various websites
- Reconfiguration is website priority aware
  - A node is said to be a free node if one of the following is true:
    - It is lightly loaded
    - It is serving a website with a lower priority than the requesting website

Reconf with Prioritization

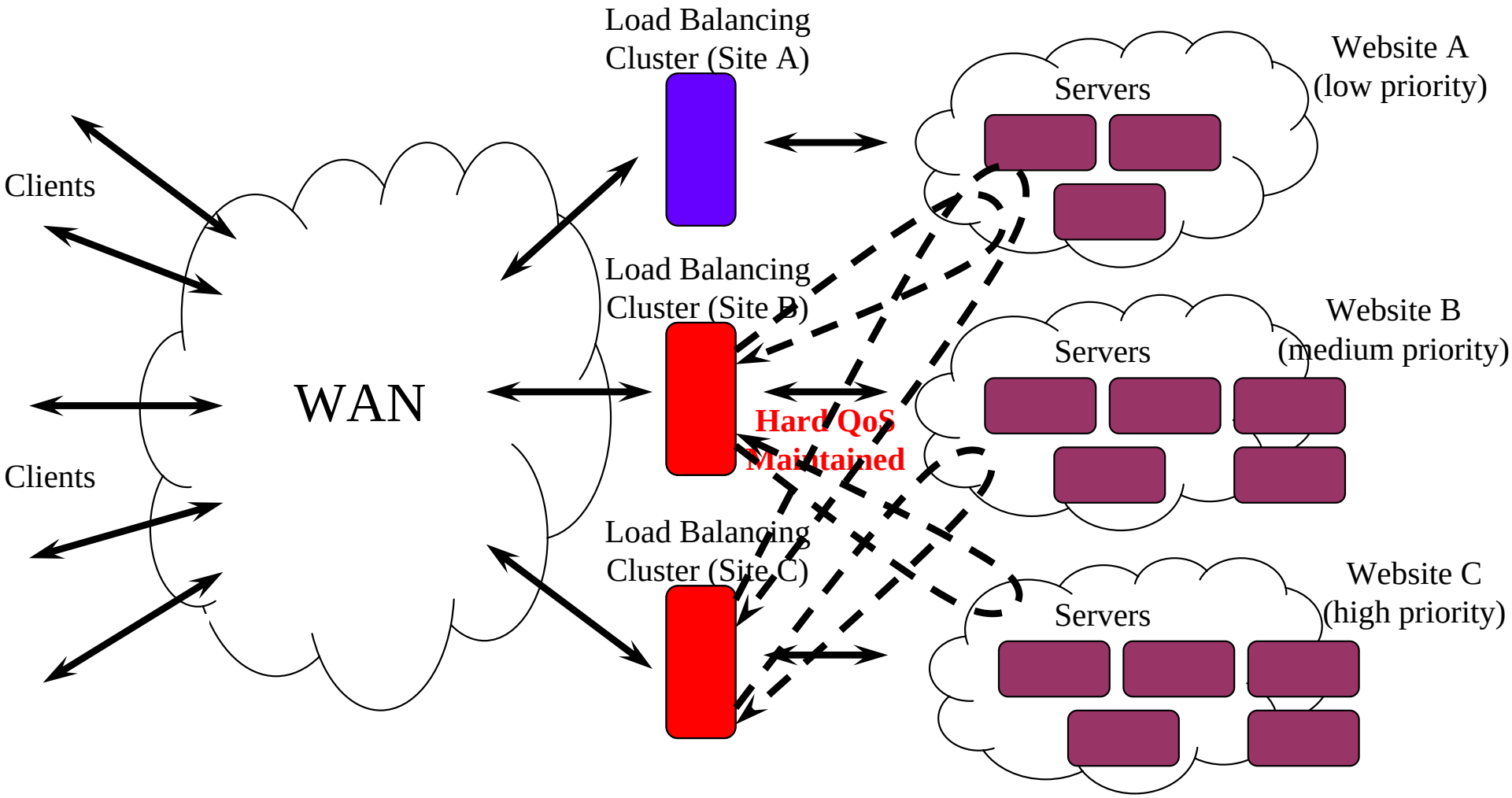


Low Priority websites may never get guaranteed number of servers since Reconf-P does not have any idea about QoS guarantees for websites

Dynamic Reconfigurability with Prioritization and Soft QoS Guarantees (Reconf-PQ)

- Prioritization based Dynamic Reconfigurability
  - Allows high paying websites to achieve a better performance
  - Can result in scarcity of resources for low priority websites
- QoS guarantees required to ensure scarcity-free reconfiguration
  - Static allocation always provides QoS guarantees
    - Low priority requests are given resources statically and never taken away
    - QoS provided based on the resources available
  - Reconf-PQ based design
    - We want to ensure that low priority requests have some guaranteed resources (Hard QoS)
    - We also want to achieve greater revenue by over-selling our resources
    - *Soft QoS Guarantees:* Maximum resources we can allot based on other requests !
    - Soft QoS ensures that a websites allocation does not deny other websites of their Hard QoS

Reconf with Prioritization and QoS



Reconf-PQ reconfigures nodes for different websites but also guarantees fixed number of nodes to low priority requests

Presentation Outline

- Introduction and Motivation
- Overview of Dynamic Reconfigurability over InfiniBand
- Issues with Basic Dynamic Reconfigurability
- Dynamic Reconfigurability for Prioritization and Soft QoS
- **Experimental Results**
- Conclusions and Future Work

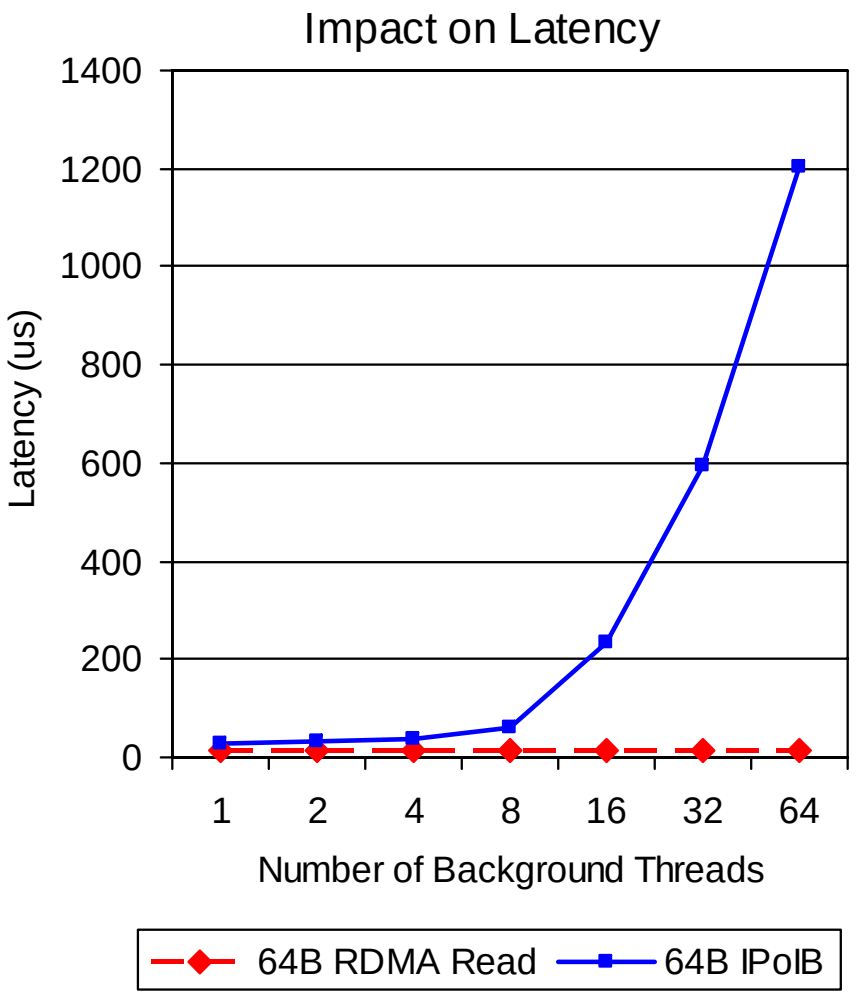
Experimental Test-bed

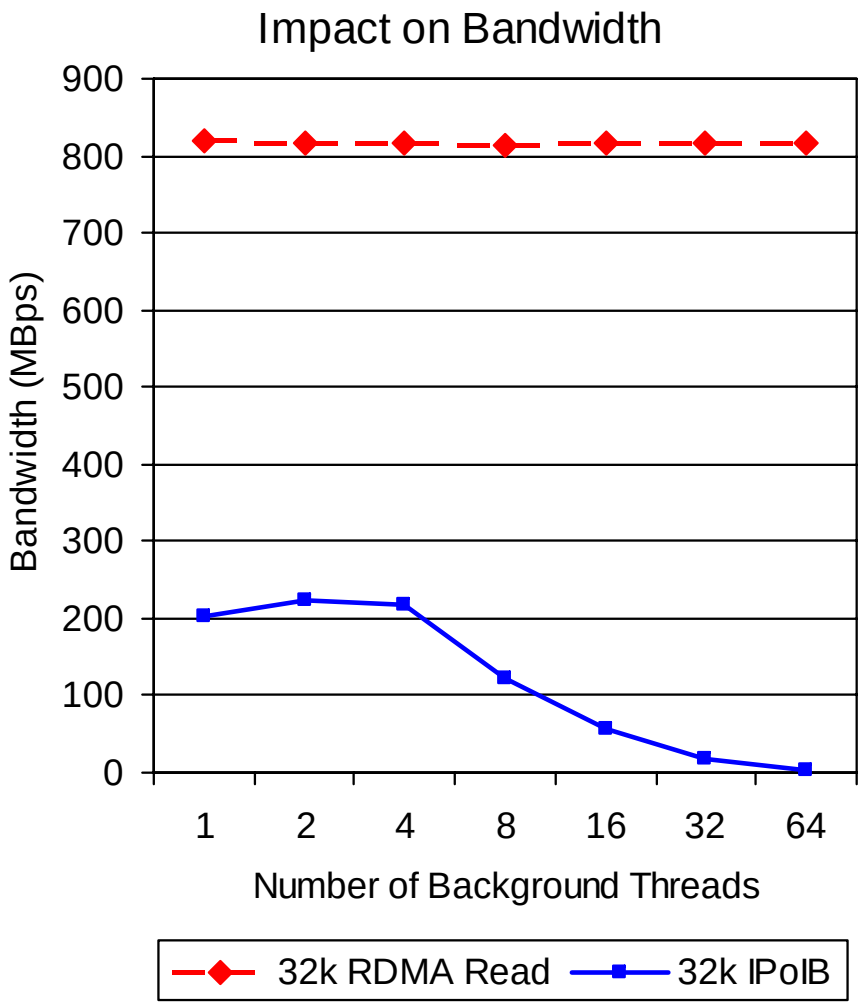
- Cluster 1 with:
  - 8 SuperMicro SUPER X5DL8-GG nodes; Dual Intel Xeon 3.0 GHz processors
  - 512 KB L2 cache, 2 GB memory; PCI-X 64-bit 133 MHz
- Cluster 2 with:
  - 8 SuperMicro SUPER P4DL6 nodes; Dual Intel Xeon 2.4 GHz processors
  - 512 KB L2 cache, 512 MB memory; PCI-X 64-bit 133 MHz
- InfiniBand Interconnect with:
  - Mellanox MT23108 Dual Port 4x HCAs; MT43132 24-port switch
- Apache 2.0.50 Web and PHP 4.3.7 servers; MySQL 4.0.12 Database server

Experimental Outline

- Load resilience capabilities of InfiniBand in the data-center environment
- Performance of Reconf comparing with static allocation schemes
- Performance of Reconf, Reconf-P, Reconf-PQ
- QoS meeting capabilities for Reconf, Reconf-P, Reconf-PQ

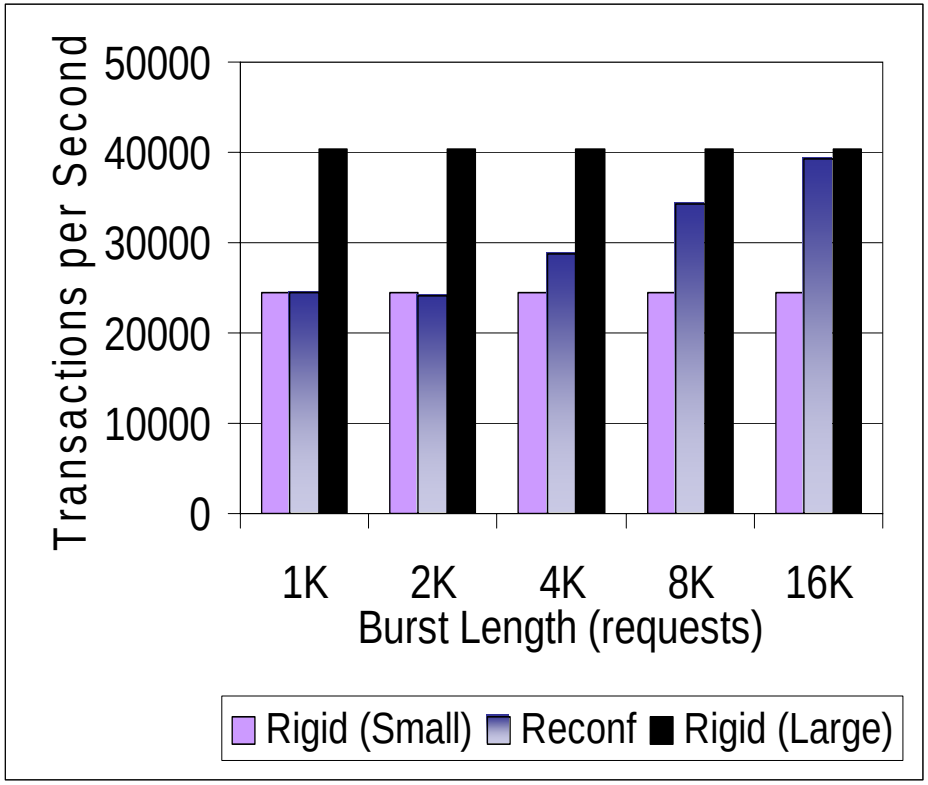
Load resilience capabilities of InfiniBand

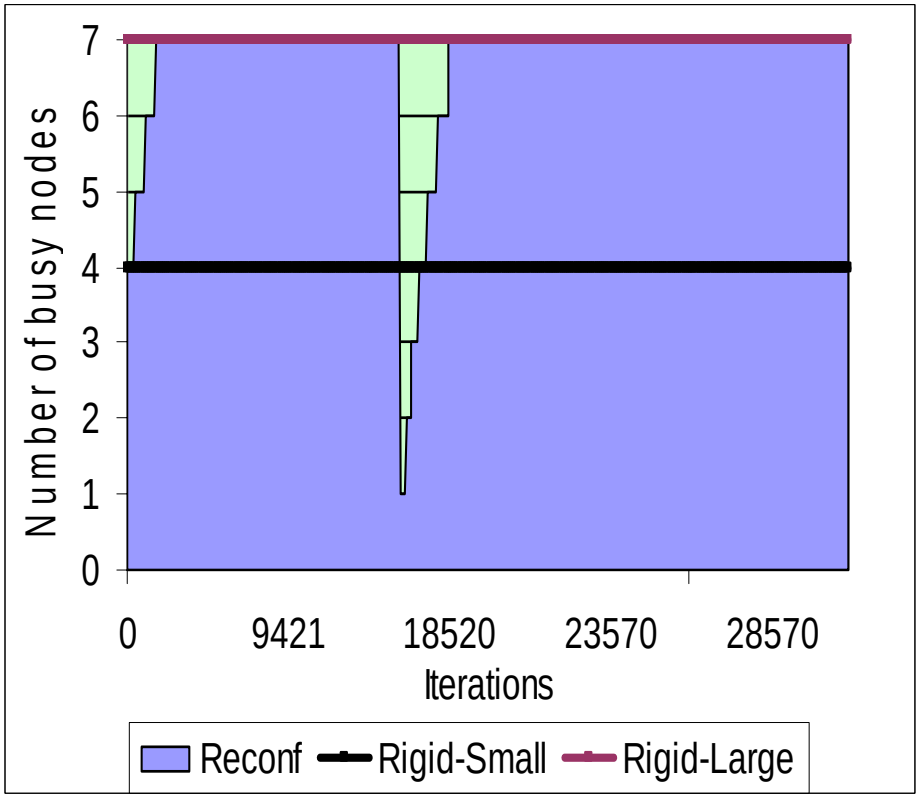




- Remote memory operations are not affected AT ALL with remote server load

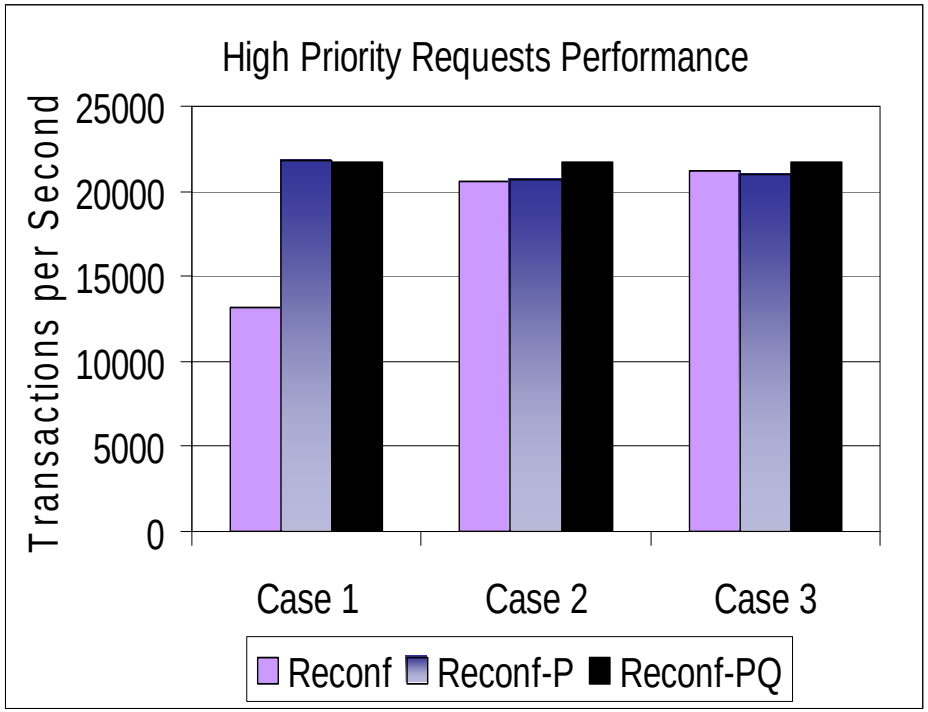
Basic Reconfigurability Performance

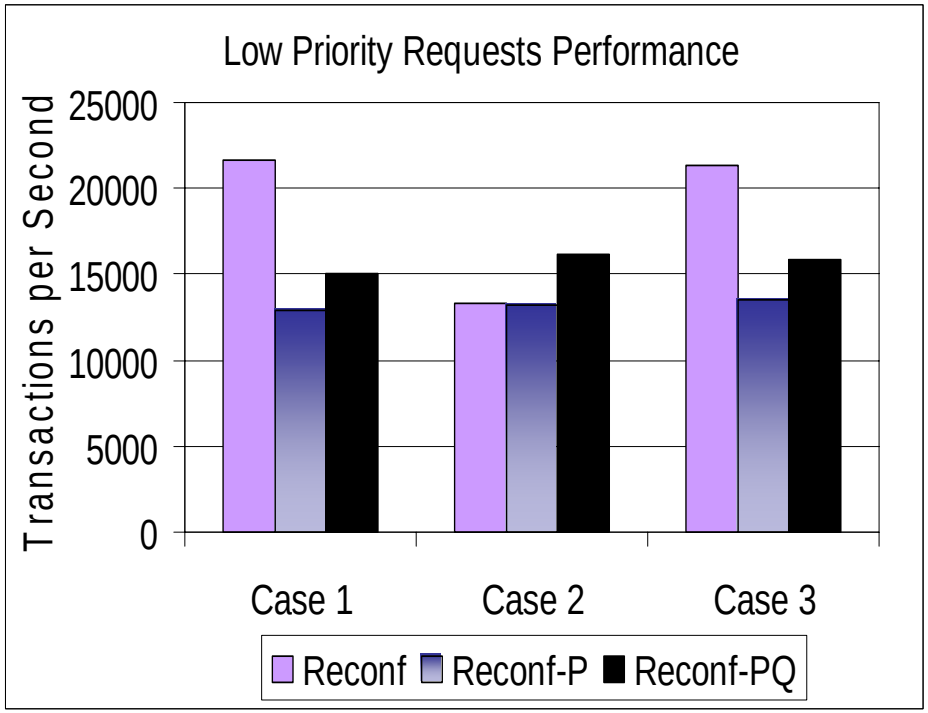




- Large Burst Length allows reconfiguration of the system closer to the best case; reconfiguration time is negligible;
- Performs comparably with the static scheme for small burst sizes

Reconfigurability Performance with QoS and Prioritization

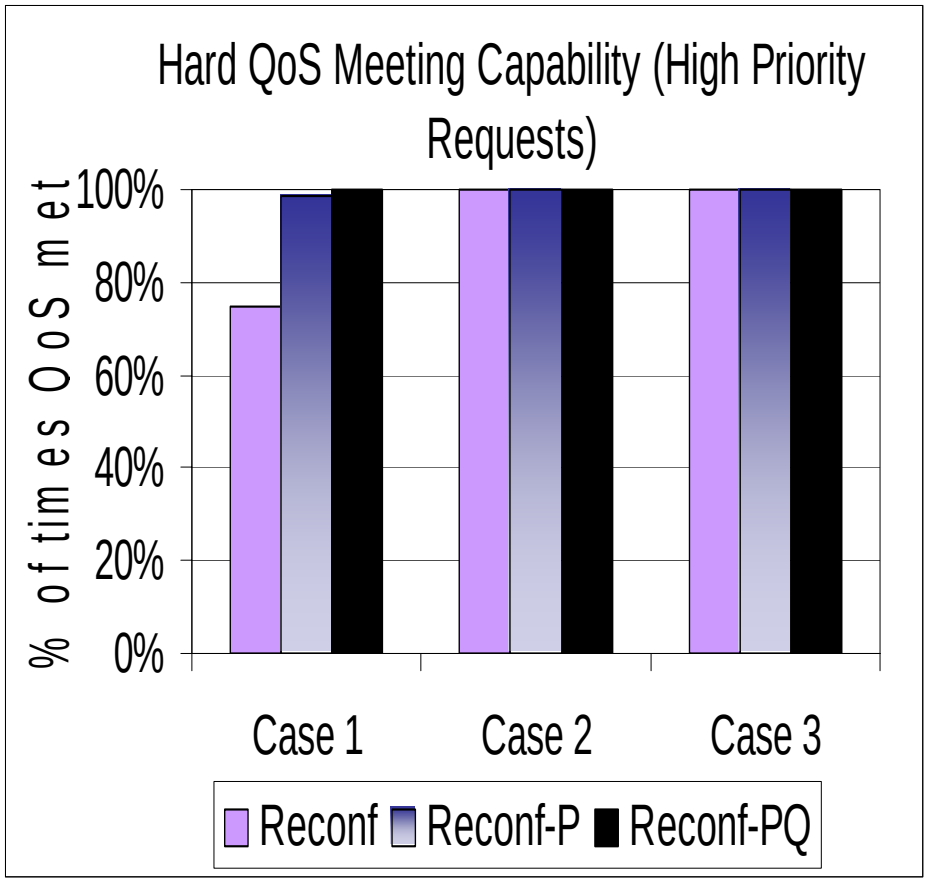


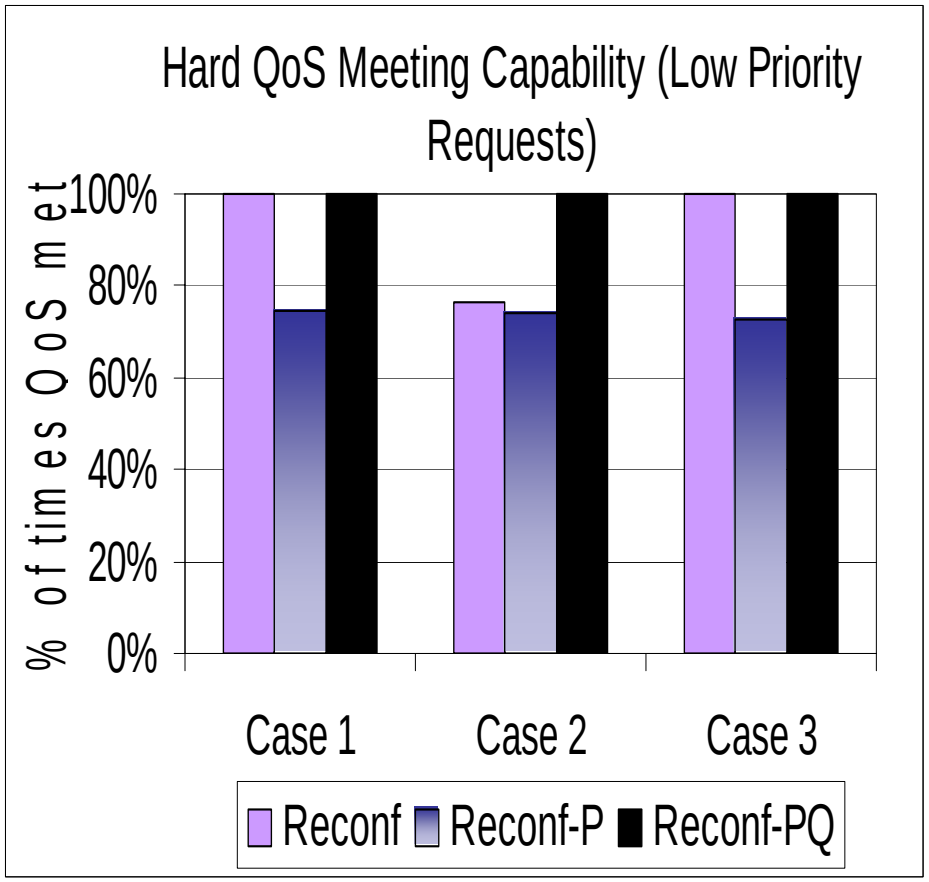


- **Case 1:** A load of high priority requests arrives when a load of low priority requests already exists
- **Case 2:** A load of low priority requests arrives when a load of high priority requests already exists
- **Case 3:** Both high and low priority requests arrive simultaneously

- Reconf does not perform any additional reconfiguration
- Reconf and Reconf-P allocate maximum number of nodes to the low-priority website whereas Reconf-PQ allocates nodes to the QoS guaranteed to that website.

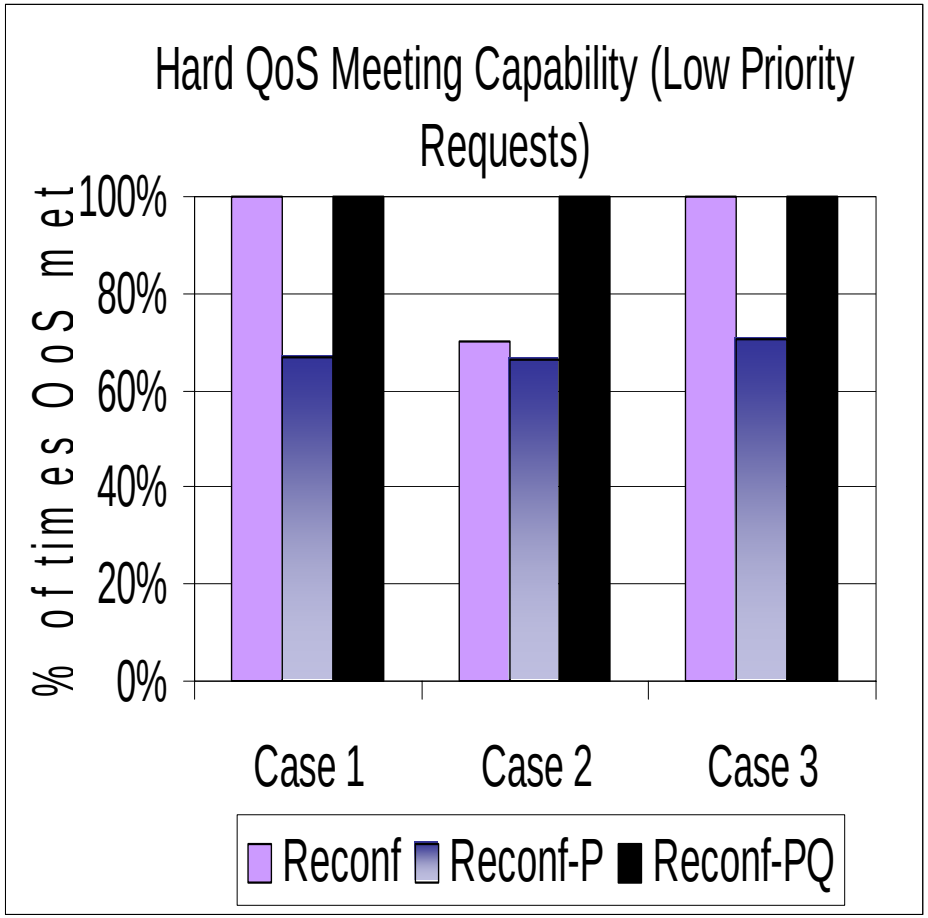
QoS Meeting Capability

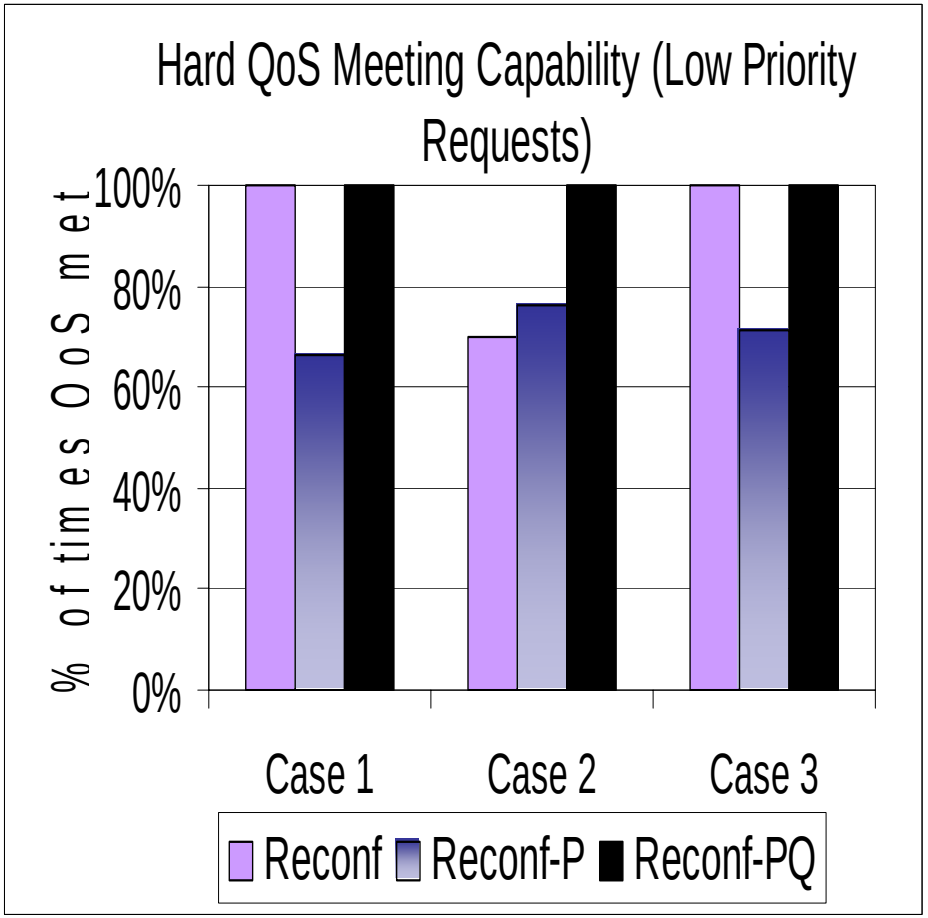




- Reconf and Reconf-P perform well only in some cases and lack consistency in providing the guaranteed QoS requirements to both websites
- Reconf-PQ meets the guaranteed QoS requirements in all cases

QoS Meeting Capability – Zipf and Worldcup Traces





- Similar trends are seen for Zipf and Worldcup traces with QoS meeting capability of nearly 100% for Reconf-PQ

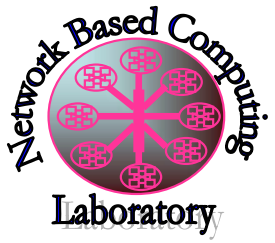
Presentation Outline

- Introduction and Motivation
- Overview of Dynamic Reconfigurability over InfiniBand
- Issues with Basic Dynamic Reconfigurability
- Dynamic Reconfigurability for Prioritization and Soft QoS
- Experimental Results
- **Conclusions and Future Work**

Concluding Remarks & Future Work

- Shared Data-Centers are commonly used by several ISPs
  - Resource Fragmentation
  - Prioritization for high paying websites
  - QoS guarantees for all websites
- Extended our previous Dynamic Reconfigurability scheme
  - Prioritization improves the performance of high priority websites
  - QoS guarantees protect the low priority websites from scarcity of resources
- Multi-Stage Reconfigurations
  - Least loaded servers might not be the best server to reconfigure, Caching constraints, Hardware heterogeneity
- Fine Grained Resource Reconfigurations
  - Have done some initial study on file system reconfigurations
  - Memory reconfiguration: utilizing remote memory in clusters as secondary cache

Web Pointers



NBC-LAB

Group Homepage: http://nowlab.cis.ohio-state.edu

Emails: {balaji, narravul, vaidyana, jinhy, panda}@cse.ohio-state.edu