Designing Next Generation Data-Centers with Advanced Communication Protocols and Systems Services

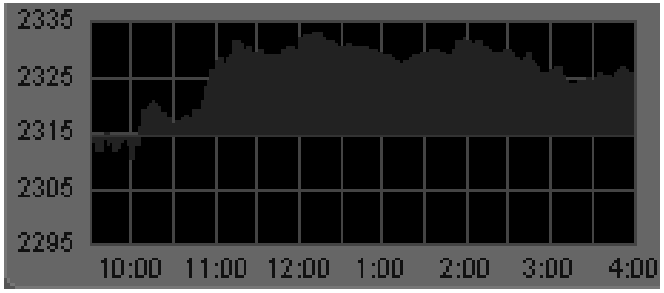
P. Balaji, K. Vaidyanathan, S. Narravula, H. –W. Jin and D. K. Panda

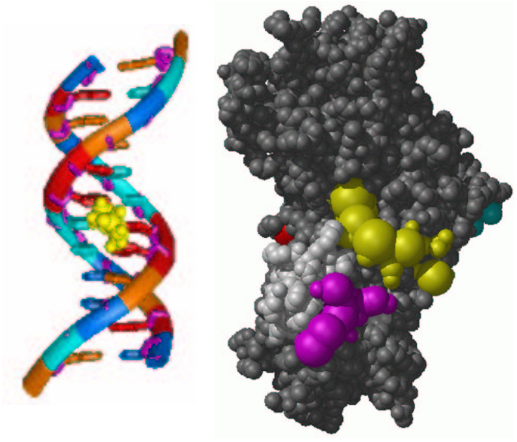
Network Based Computing Laboratory (NBCL)

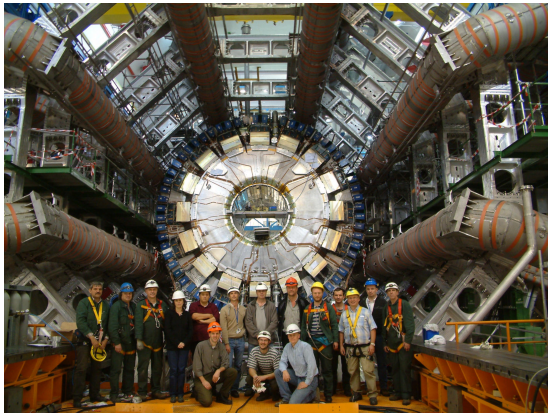
Computer Science and Engineering

Ohio State University

Introduction and Motivation

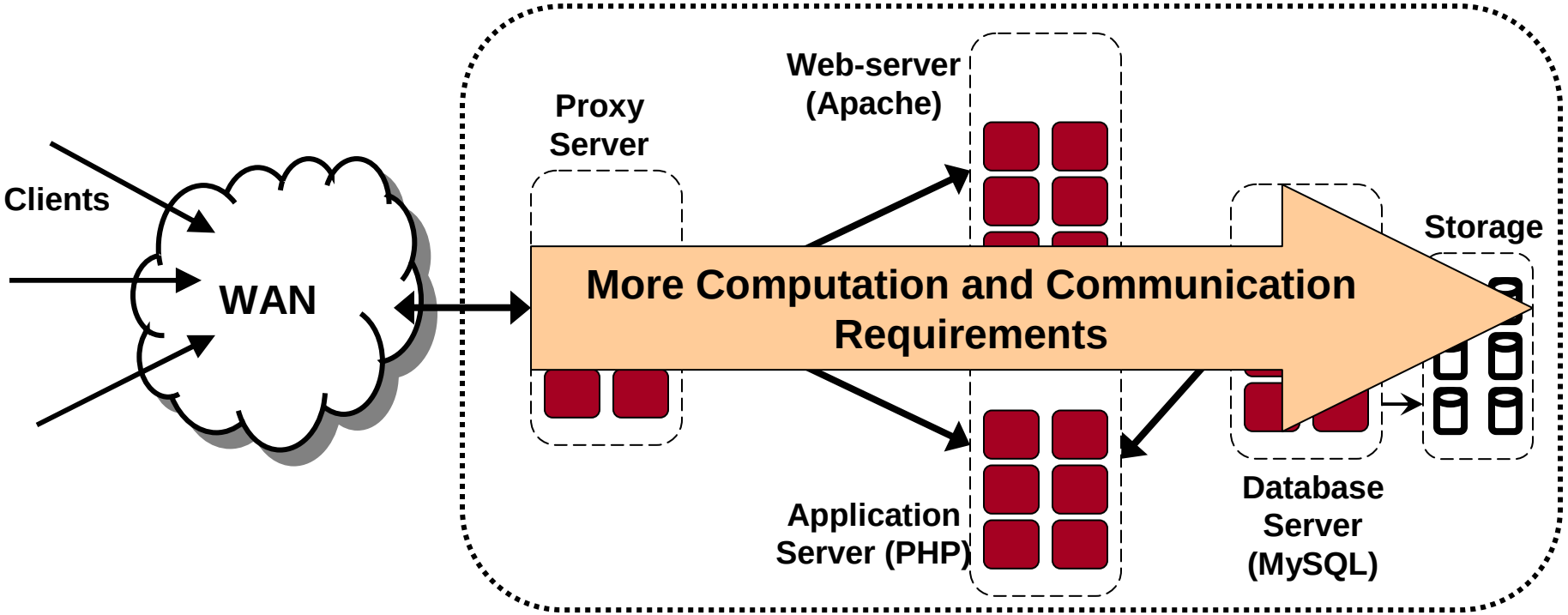






- Interactive Data-driven Applications
  - Scientific as well as Enterprise/Commercial Applications
    - Static Datasets: Medical Imaging Modalities
    - Dynamic Datasets: Stock value datasets, E-commerce, Sensors
  - E-science
  - Ability to interact with, synthesize and visualize large datasets
  - Data-centers enable such capabilities
- Clients initiate queries (over the web) to process specific datasets
  - Data-centers process data and reply to queries

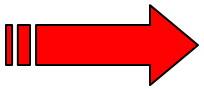
Typical Multi-Tier Data-center Environment



- Requests are received from clients over the WAN
- Proxy nodes perform caching, load balancing, resource monitoring, etc.
- If not cached, the request is forwarded to the next tiers → Application Server
- Application server performs the business logic (CGI, Java servlets, etc.)
  - Retrieves appropriate data from the database to process the requests

Limitations of Current Data-centers

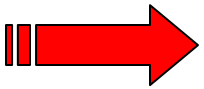
- Communication Requirements
  - TCP/IP used even in the data-center: Sub-optimal performance
    - InfiniBand and other interconnects provide more features



High Performance Sockets (e.g., SDP)

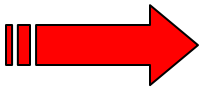
- Superior performance with no modifications

- Advanced Data-center Services



Minimize the computation requirements

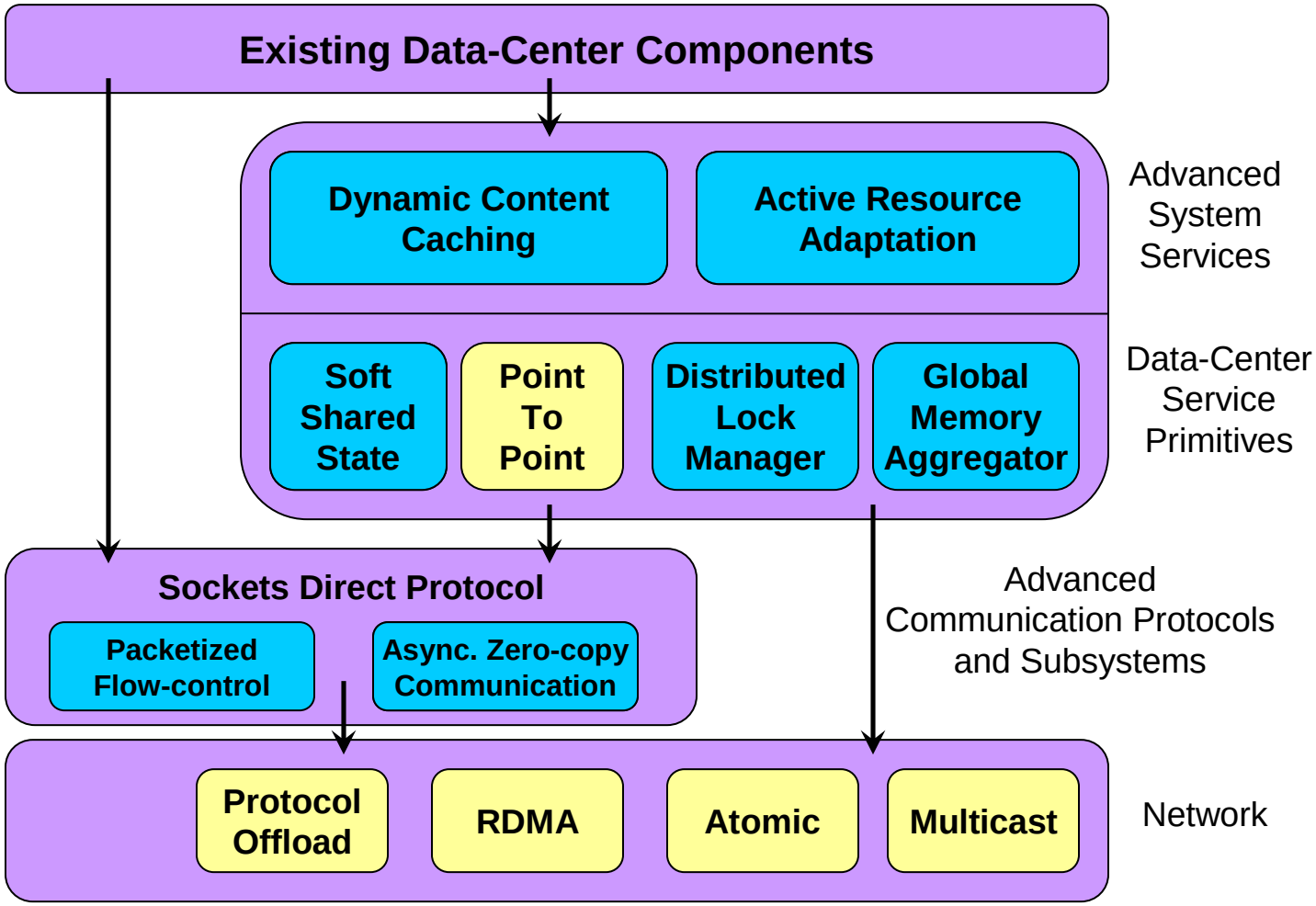
- Improved caching of documents
- Issues with caching Dynamic (or Active) Content



Maximize compute resource utilization

- Efficient resource monitoring and management
- Issues with heterogeneous load characteristics of data-centers

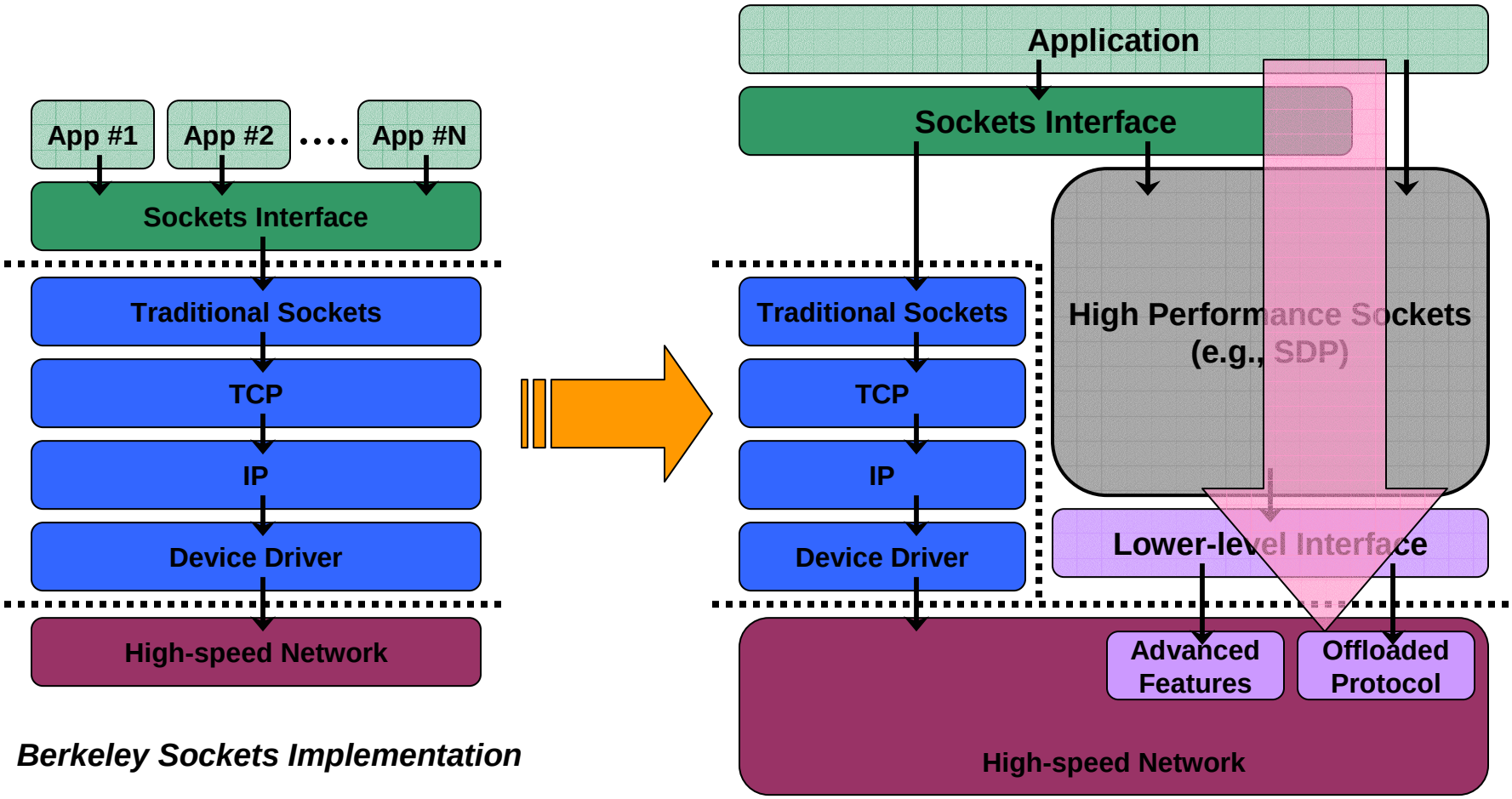
Proposed Architecture



Presentation Layout

- Introduction and Motivation
- **Advanced Communication Protocols and Subsystems**
- Data-center Service Primitives
- Dynamic Content Caching Services
- Active Resource Adaptation Services
- Conclusions and Ongoing Work

The Sockets Protocol Stack



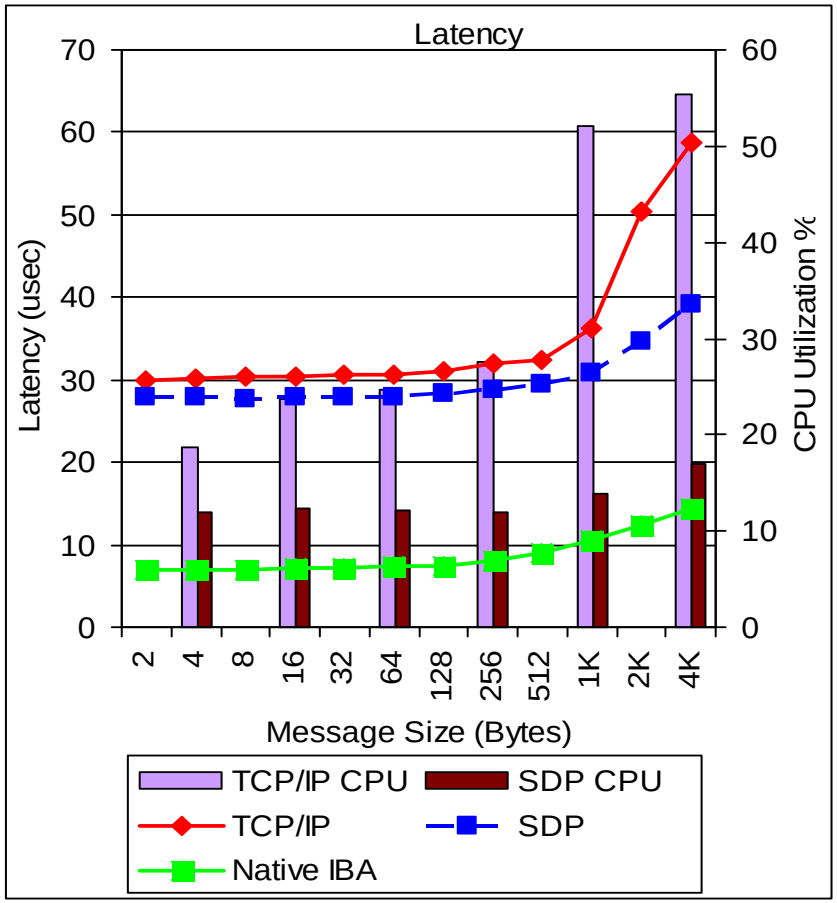
Berkeley Sockets Implementation

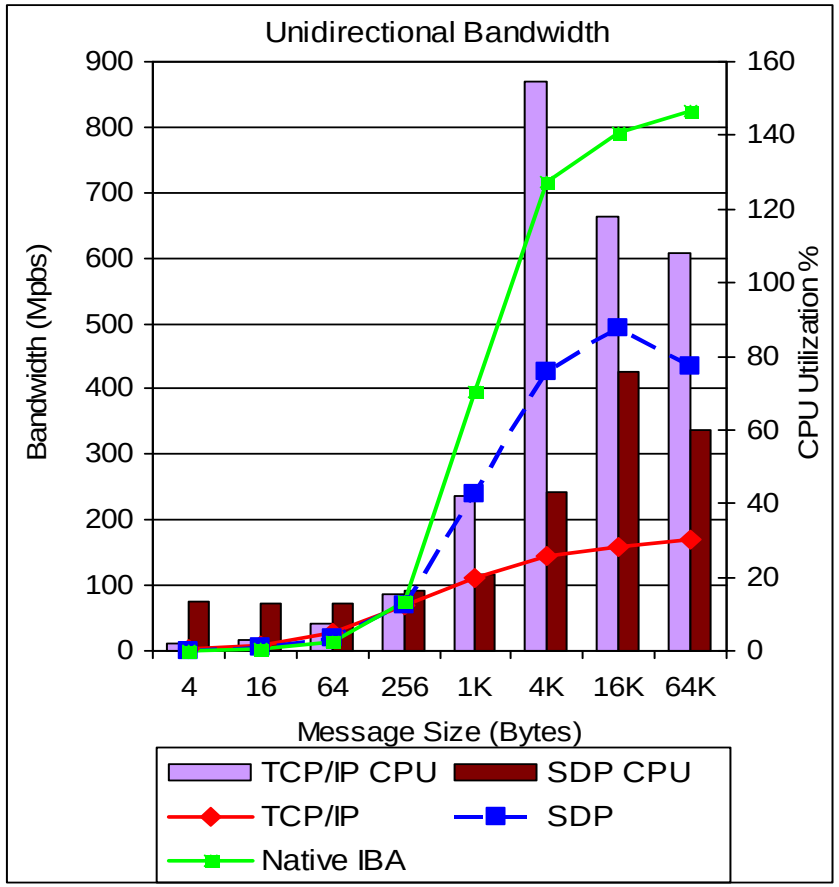
The Sockets Protocol Stack allows applications to utilize the network performance and capabilities with NO or MINIMAL modifications

InfiniBand and Features

- An emerging open standard high performance interconnect
- High Performance Data Transfer
  - Interprocessor communication and I/O
  - Low latency (~1.0-3.0 microsec), High bandwidth (~10-20 Gbps) and low CPU utilization (5-10%)
- Flexibility for WAN communication
- Multiple Operations
  - Send/Recv
  - RDMA Read/Write
  - Atomic Operations (very unique)
    - high performance and scalable implementations of distributed locks, semaphores, collective communication operations
- **Range of Network Features and QoS Mechanisms**
  - Service Levels (priorities)
  - Virtual lanes
  - Partitioning
  - Multicast
    - allows to design a new generation of scalable communication and I/O subsystem with QoS

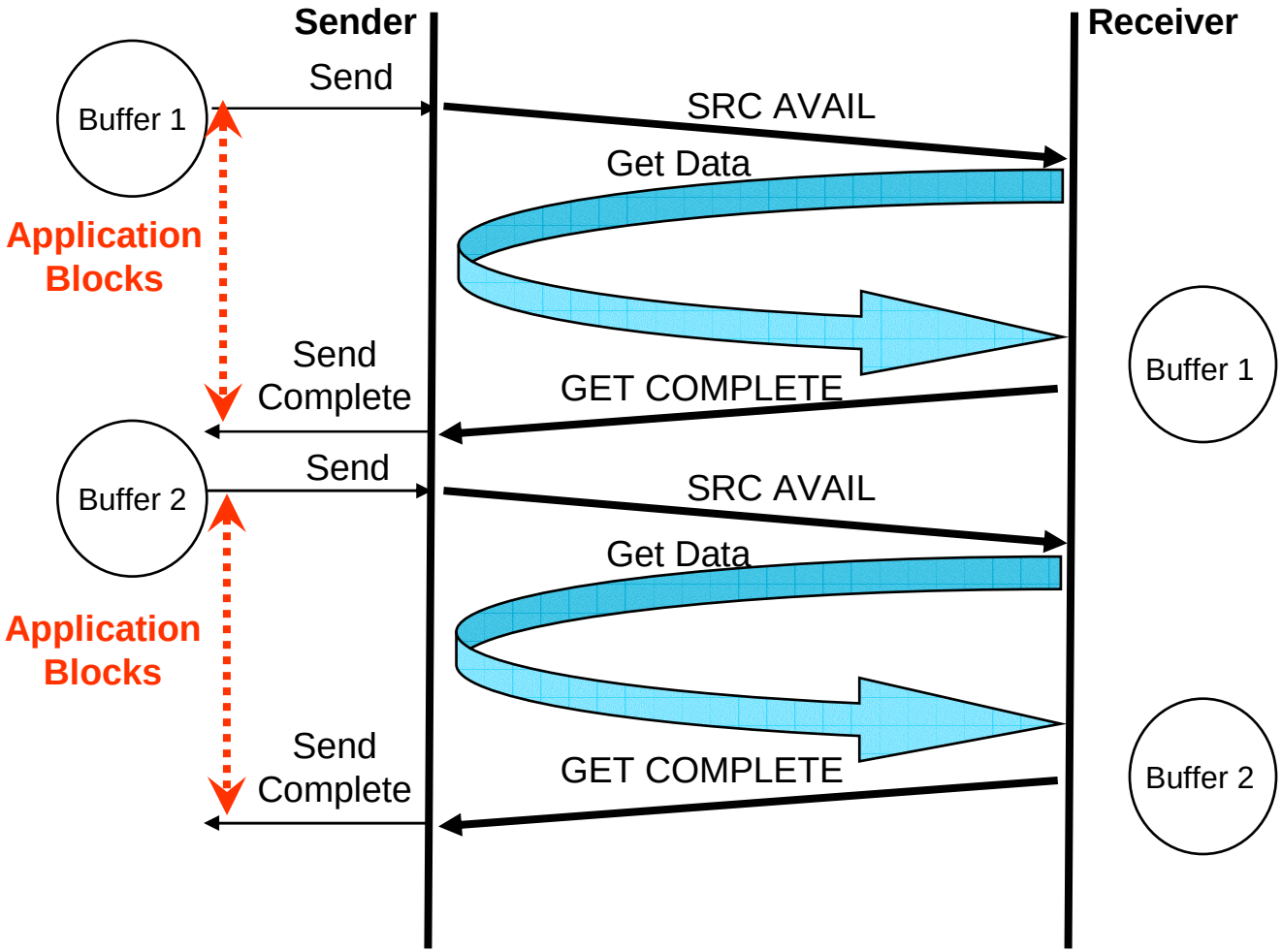
SDP Latency and Bandwidth



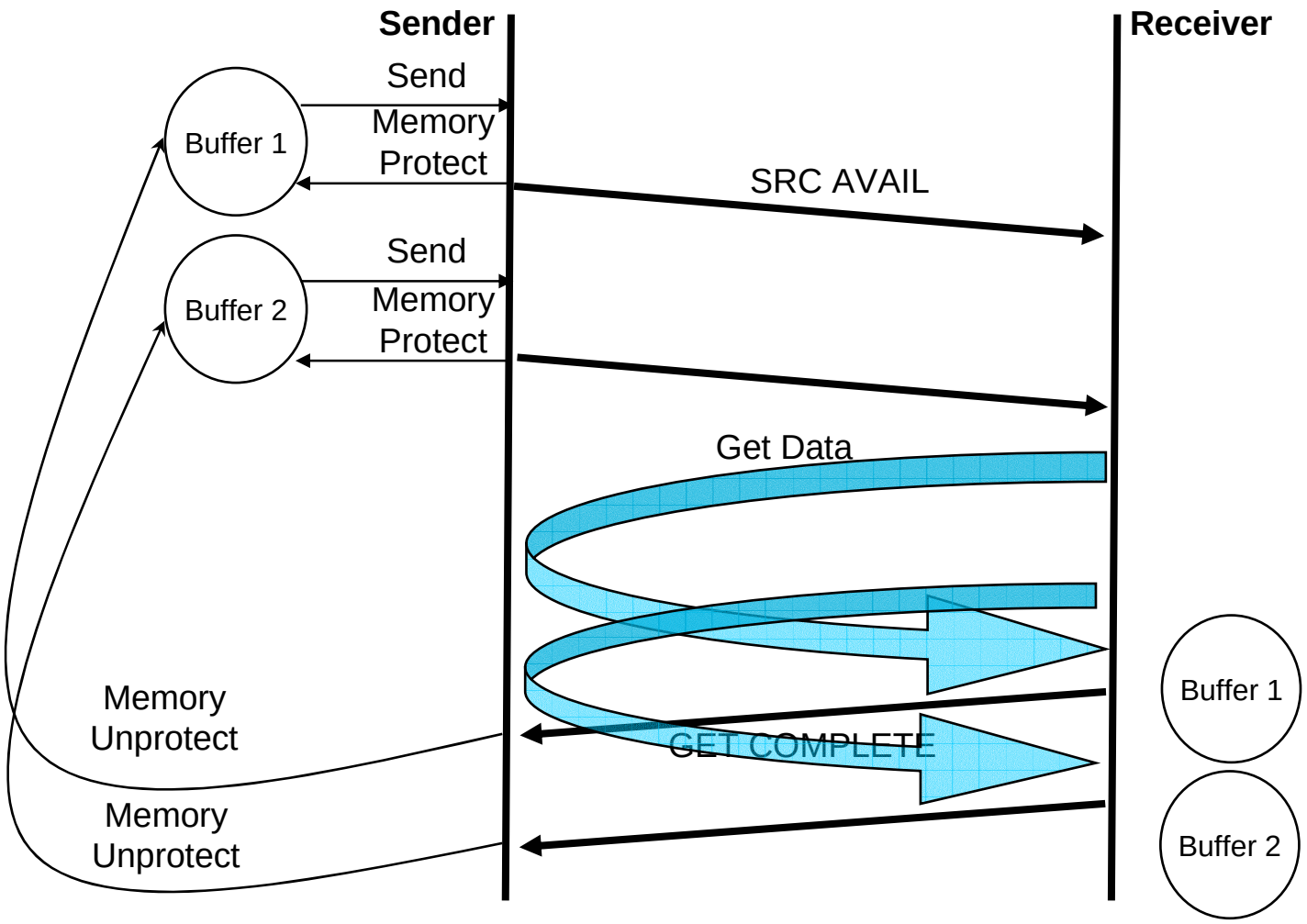


"Sockets Direct Protocol over InfiniBand in Clusters: Is it Beneficial?", P. Balaji, S. Narravula, K. Vaidyanathan, K. Savitha, D. K. Panda. IEEE International Symposium on Performance Analysis and Systems (ISPASS), 04.

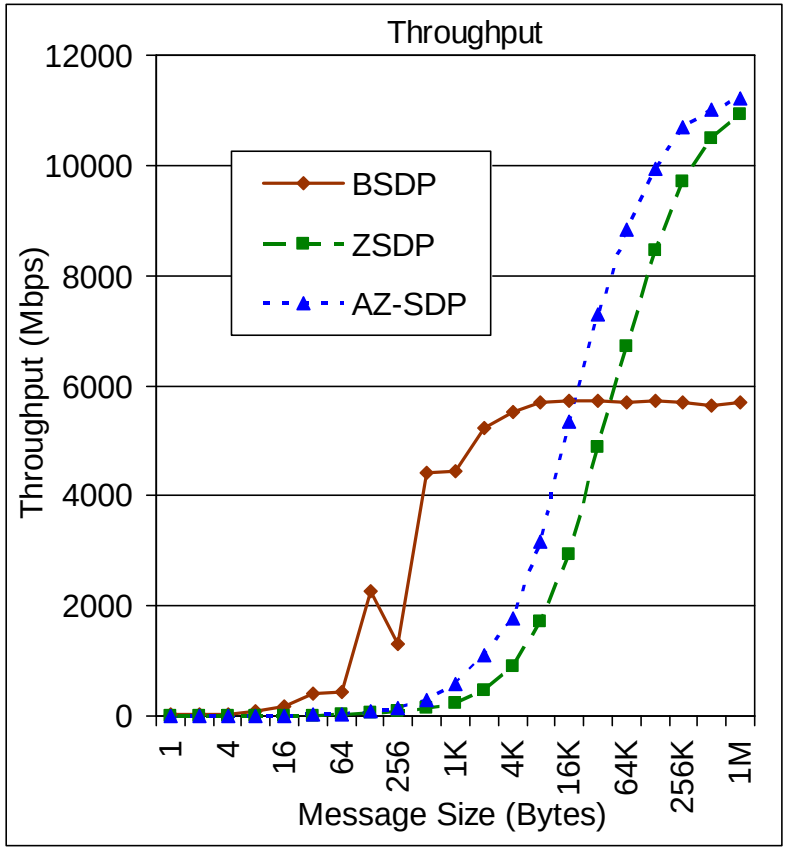
Zero-Copy Communication for Sockets

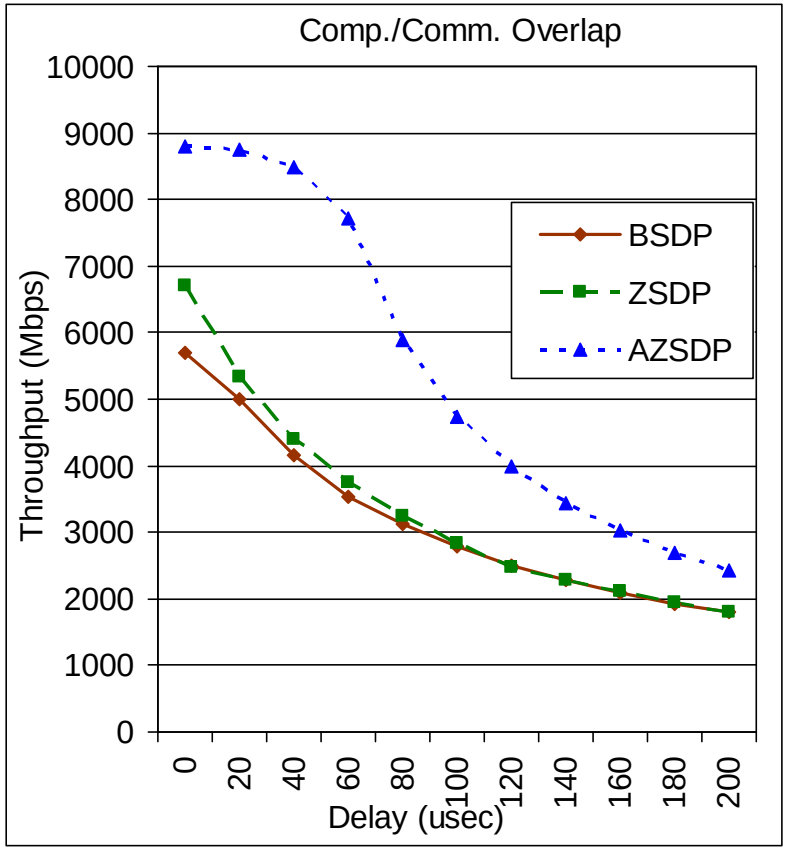


Asynchronous Zero-Copy SDP



Throughput and Comp./Comm. Overlap





"Asynchronous Zero-copy Communication for Synchronous Sockets in the Sockets Direct Protocol (SDP) over InfiniBand". P. Balaji, S. Bhagvat, H. –W. Jin and D. K. Panda. Workshop on Communication Architecture for Clusters (CAC); with IPDPS '06.

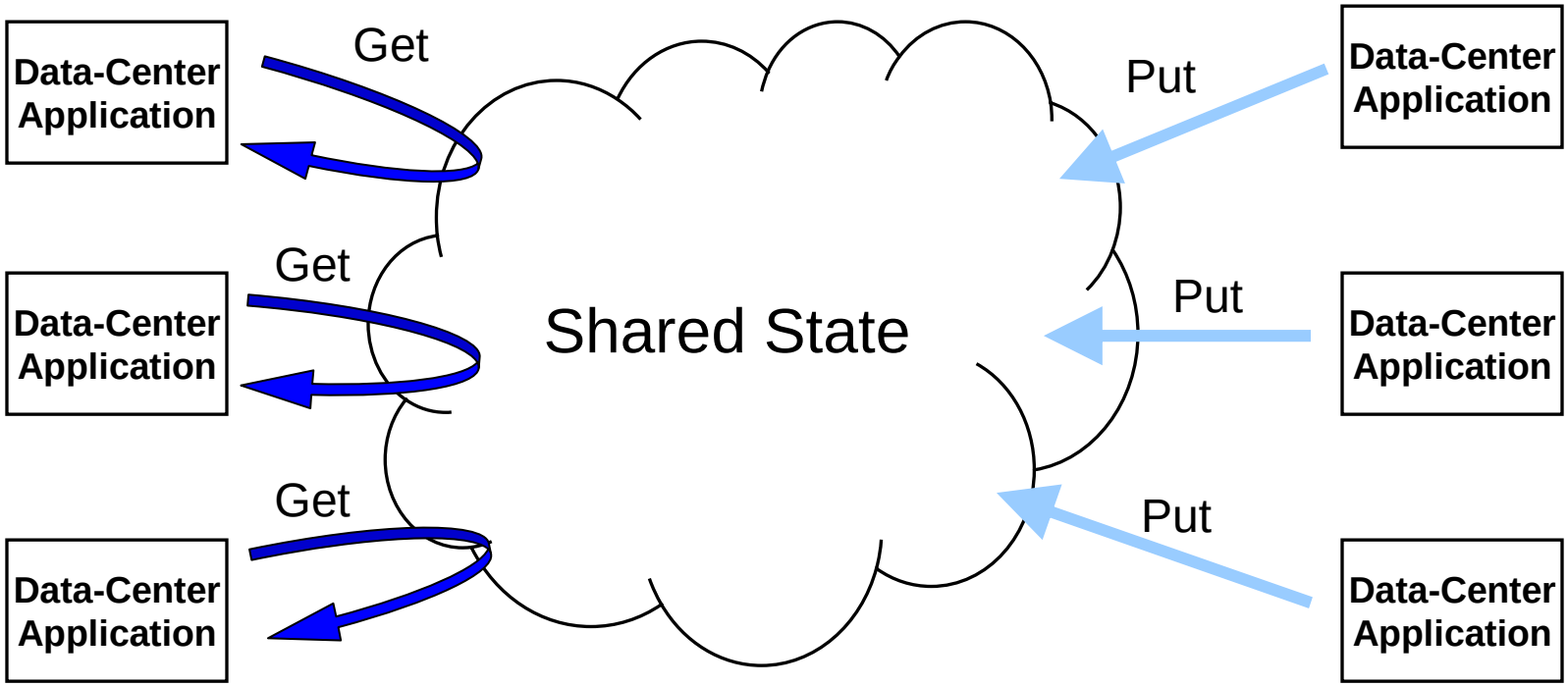
Presentation Layout

- Introduction and Motivation
- Advanced Communication Protocols and Subsystems
- **Data-center Service Primitives**
- Dynamic Content Caching Services
- Active Resource Adaptation Services
- Conclusions and Ongoing Work

Data-Center Service Primitives

- Common Services needed by Data-Centers
  - Better resource management
  - Higher performance provided to higher layers
- Service Primitives
  - Soft Shared State
  - Distributed Lock Management
  - Global Memory Aggregator
- Network Based Designs
  - RDMA, Remote Atomic Operations

Soft Shared State

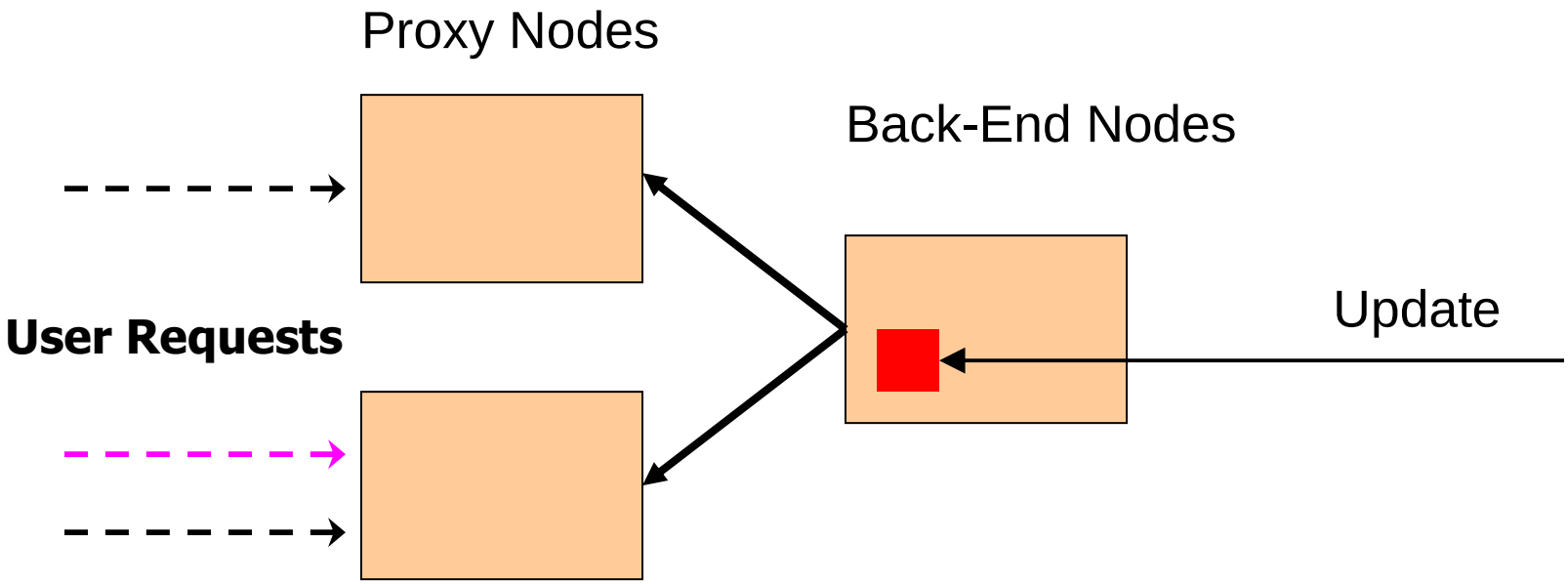


Presentation Layout

- Introduction and Motivation
- Advanced Communication Protocols and Subsystems
- Data-center Service Primitives
- **Dynamic Content Caching Services**
- Active Resource Adaptation Services
- Conclusions and Ongoing Work

Active Caching

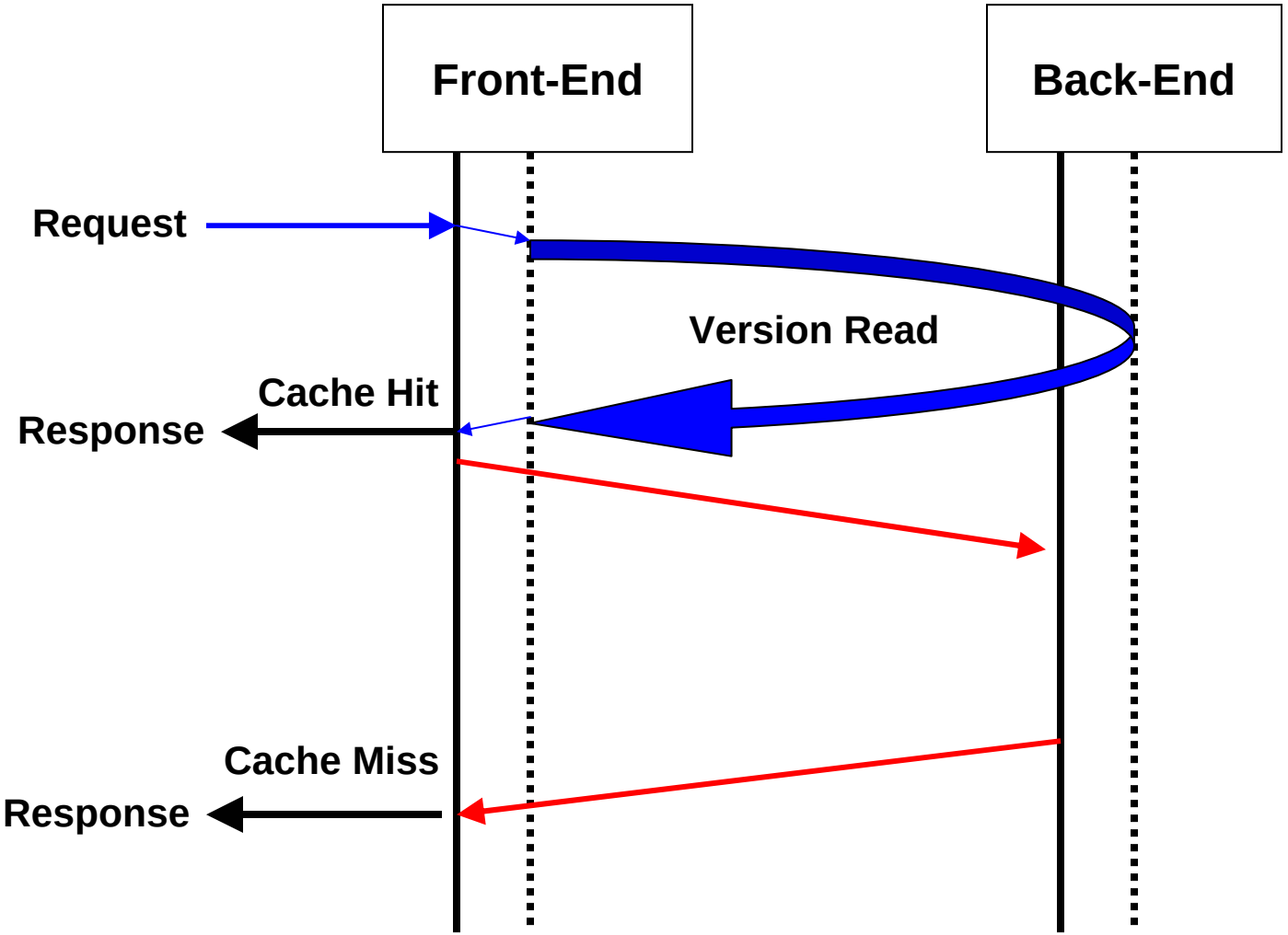
- Dynamic data caching – challenging!
- Cache Consistency and Coherence
  - Become more important than in static case



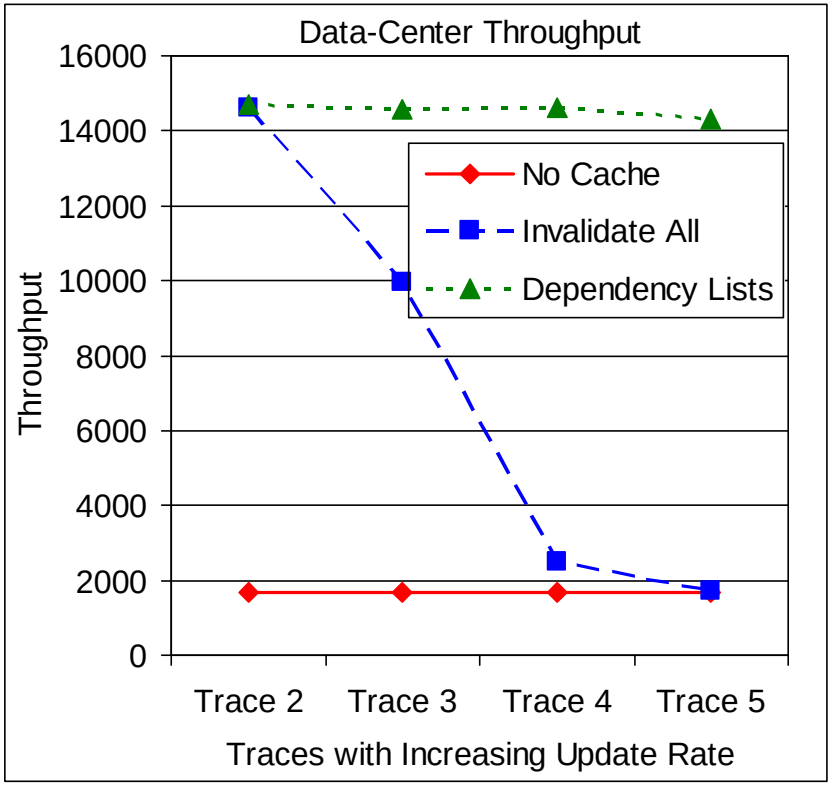
Active Cache Design

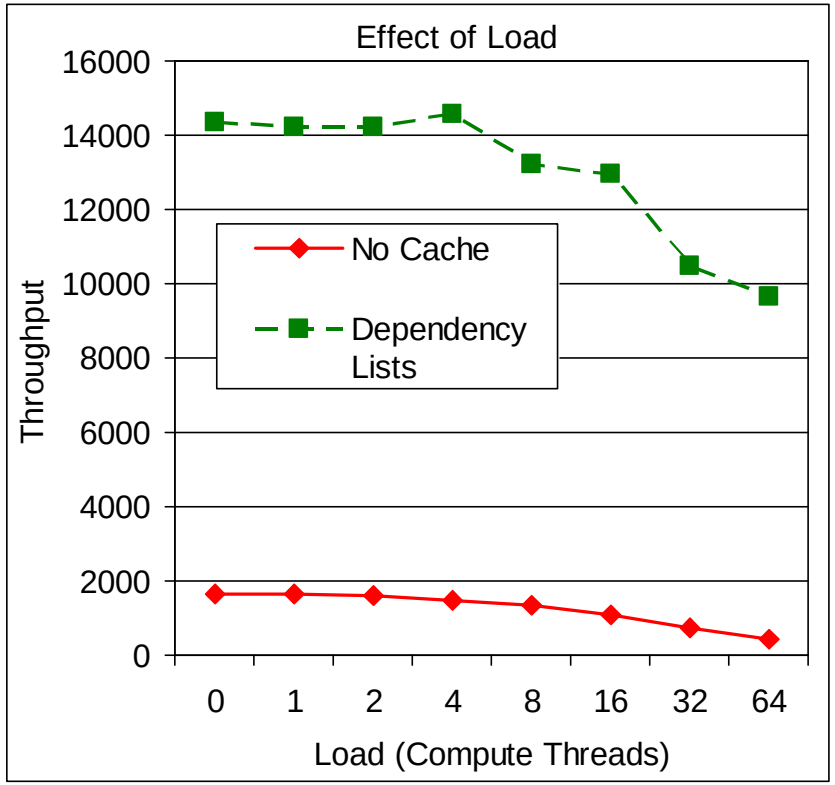
- Efficient mechanisms needed
  - RDMA based design
  - Load resiliency
- Our cooperation protocols
  - *No-Dependency*
  - *Invalidate-All*
- Client Polling based design

RDMA based Client Polling Design



Active Caching - Performance



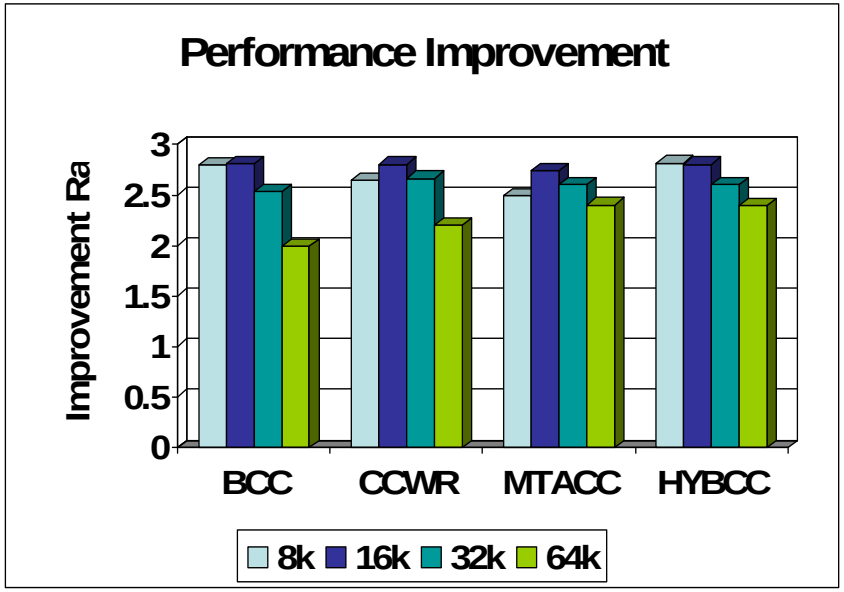


- Higher overall performance – Up to an order of magnitude
- Performance is sustained under loaded conditions

Architecture for Caching Responses with Multiple Dynamic Dependencies in Multi-Tier Data-Centers over InfiniBand. S. Narravula, P. Balaji, K. Vaidyanathan, H. -W. Jin and D. K. Panda. CCGrid-2005

Multi-tier Cooperative Caching

- RDMA based schemes
- Effective use of system-wide memory from across multiple tiers
- Significant performance benefits
  - Our Schemes
    - BCC, CCWR, MTACC and HYBCC
  - Up to 2-3 times compared to the base case



S. Narravula, H. -W. Jin, K. Vaidyanathan and D. K. Panda, Designing Efficient Cooperative Caching Schemes for Multi-Tier Data-Centers over RDMA-enabled Networks. IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid 06).

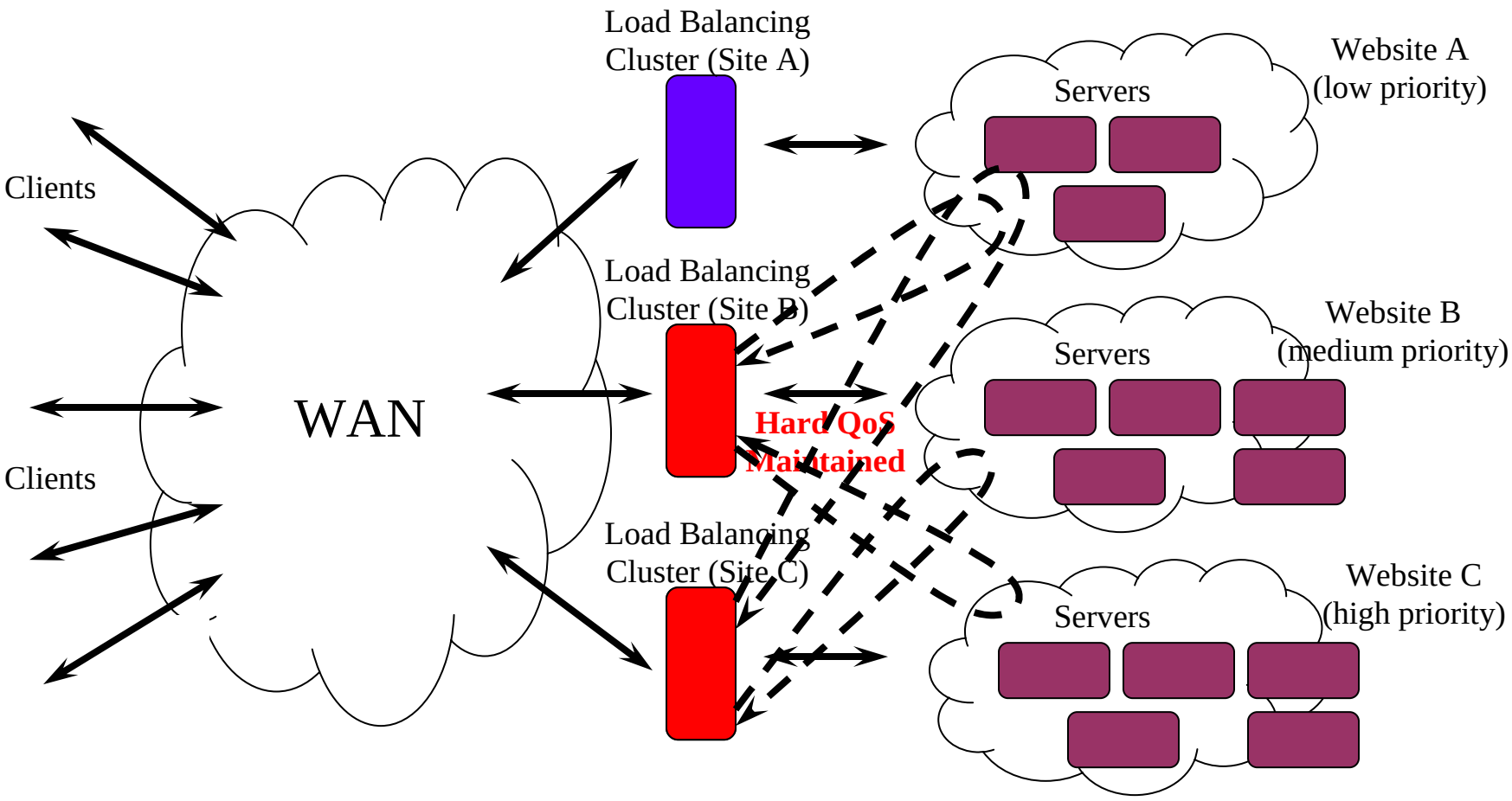
Presentation Layout

- Introduction and Motivation
- Advanced Communication Protocols and Subsystems
- Data-center Service Primitives
- Dynamic Content Caching Services
- **Active Resource Adaptation Services**
- Conclusions and Ongoing Work

Active Resource Adaptation

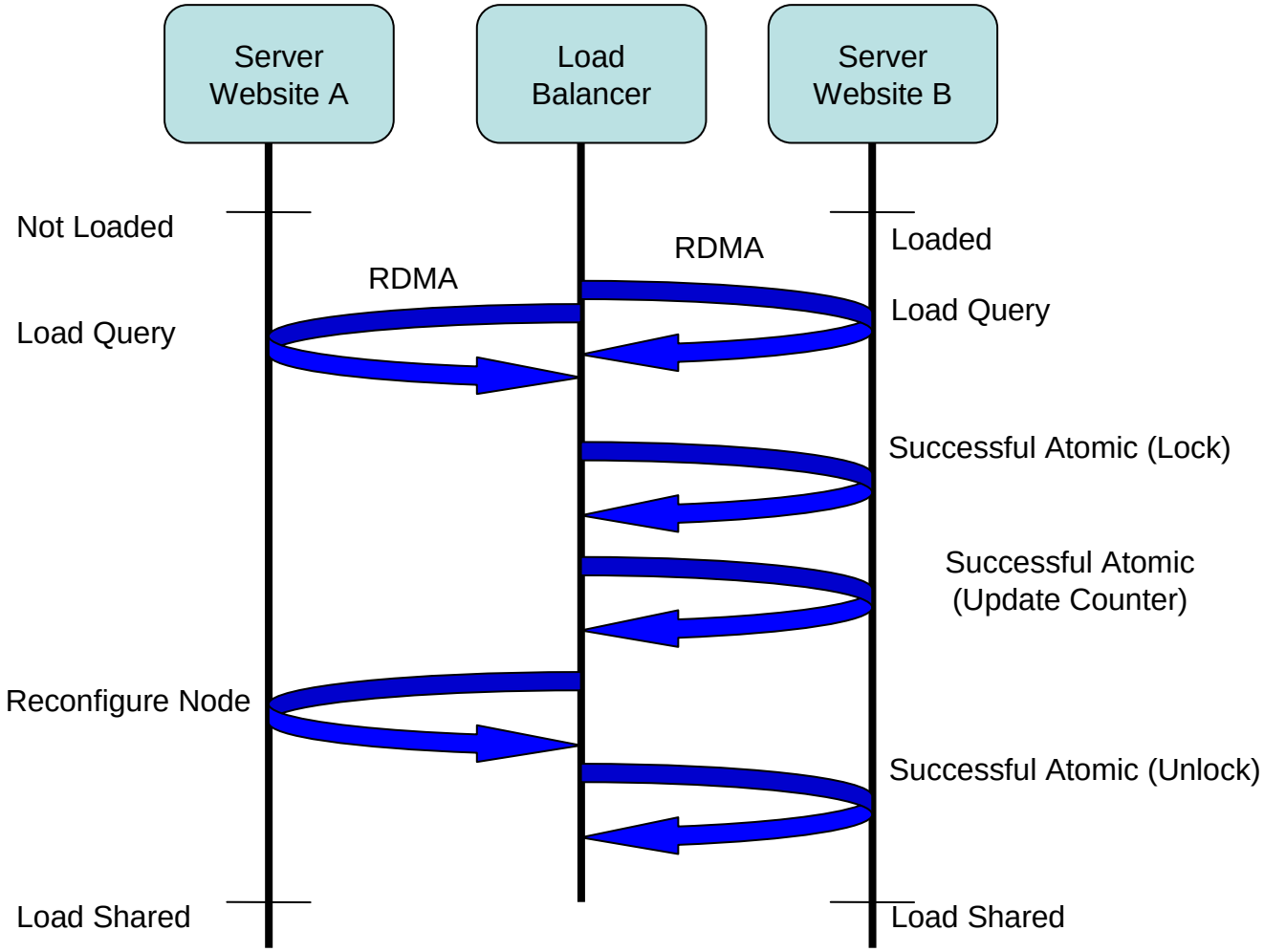
- Increasing popularity of Shared data-centers
- How to decide the number of proxy nodes vs. application servers vs. database servers
- Current approach
  - Use a rigid configuration
  - Over-Provisioning
- Active Resource Adaptation
  - Reconfigure nodes from one tier to another tier
  - Allocate resources based on system load and traffic pattern
  - Meet QoS and Prioritization constraints
  - Load Resiliency

Active Resource Adaptation in Shared Data-Centers

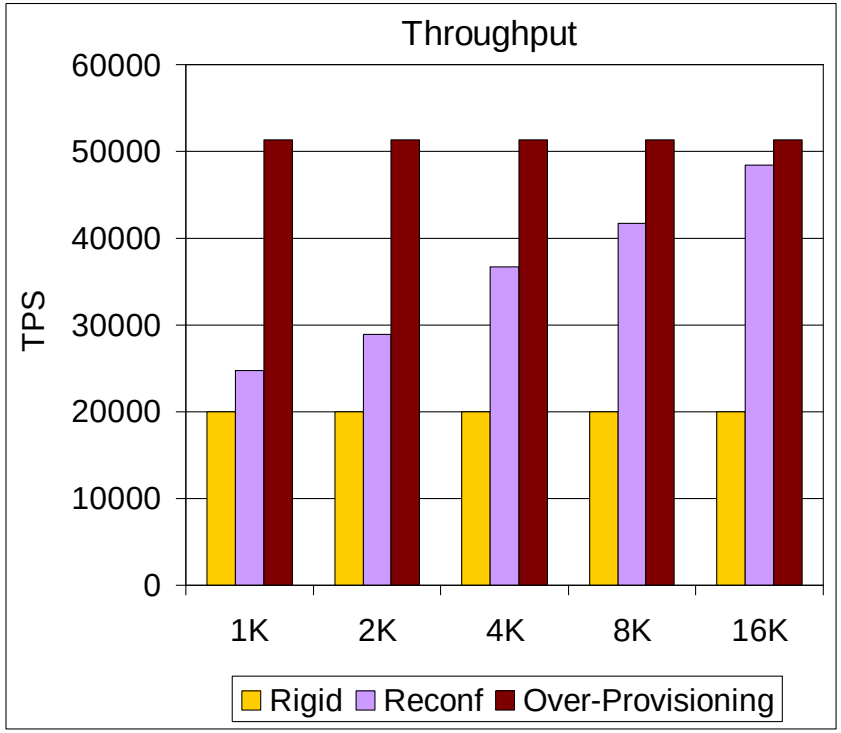


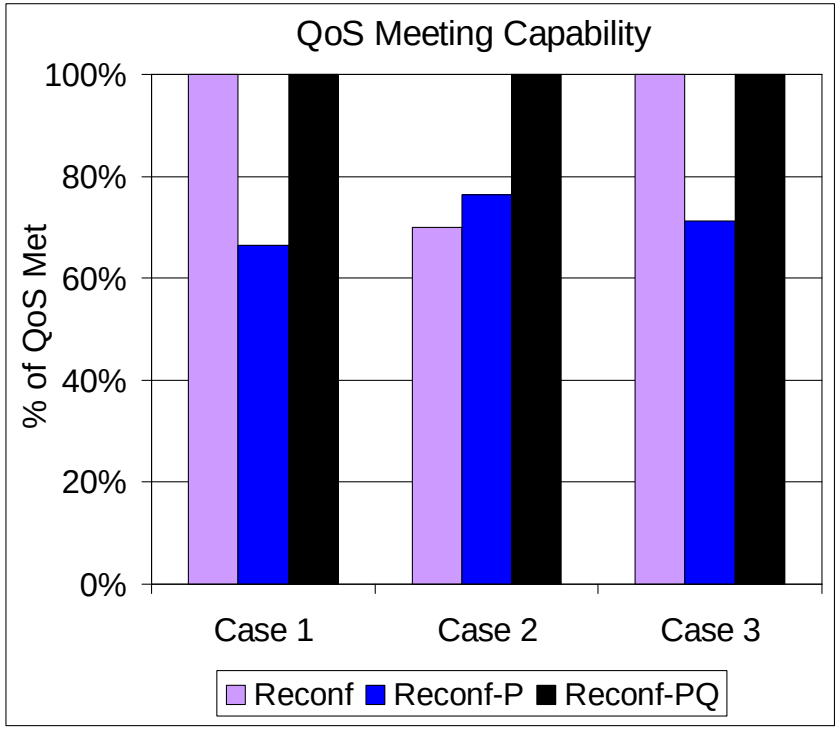
Reconf-PQ reconfigures nodes for different websites but also guarantees fixed number of nodes to low priority requests

Active Resource Adaptation Design



Dynamic Reconfigurability using RDMA operations





"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. IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS) '05.

Presentation Layout

- Introduction and Motivation
- Advanced Communication Protocols and Subsystems
- Data-center Service Primitives
- Dynamic Content Caching Services
- Active Resource Adaptation Services
- **Conclusions and Ongoing Work**

Conclusions

- Proposed a novel framework for data-centers to address the current limitations
  - Low performance due to high communication overheads
  - Lack of efficient support of advanced features such as active caching, dynamic resource adaptation, etc
- Three-layer Architecture
  - Communication Protocol Support
  - Data-Center Primitives
  - Data-Center Services
- Novel approaches using the advanced features of InfiniBand
  - Resilient to the load on the back-end servers
  - Order of magnitude performance gain for several scenarios

Work-in-Progress

- Data-Center Primitives
  - Efficient System-Wide Soft Shared State Mechanisms
  - Efficient Distributed Lock Manager Mechanisms
- Fine-Grained Active Resource Adaptation
  - Fine-grain resource monitoring
  - Resource adaptation with database servers and multi-stage reconfigurations
- Detailed Data-Center Evaluation with the proposed framework

Web Pointers

NBCL

Website: http://www.cse.ohio-state.edu/~panda

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

Email: panda@cse.ohio-state.edu