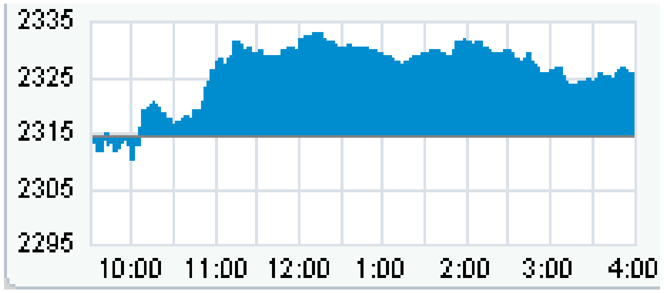
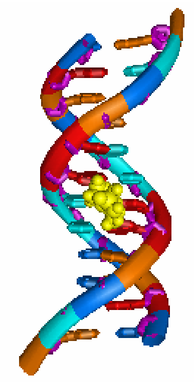
Designing Efficient Systems Services and Primitives for Next-Generation Data-Centers

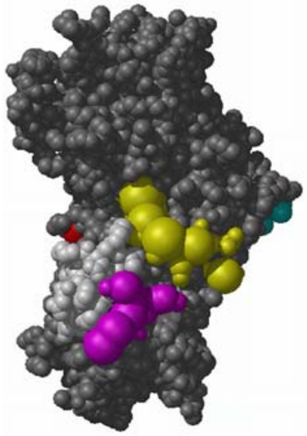
K. Vaidyanathan, S. Narravula, P. Balaji and D. K. Panda
Network Based Computing Laboratory (NBCL)
Computer Science and Engineering
Ohio State University

NSF #CNS - 0509452

Introduction and Motivation

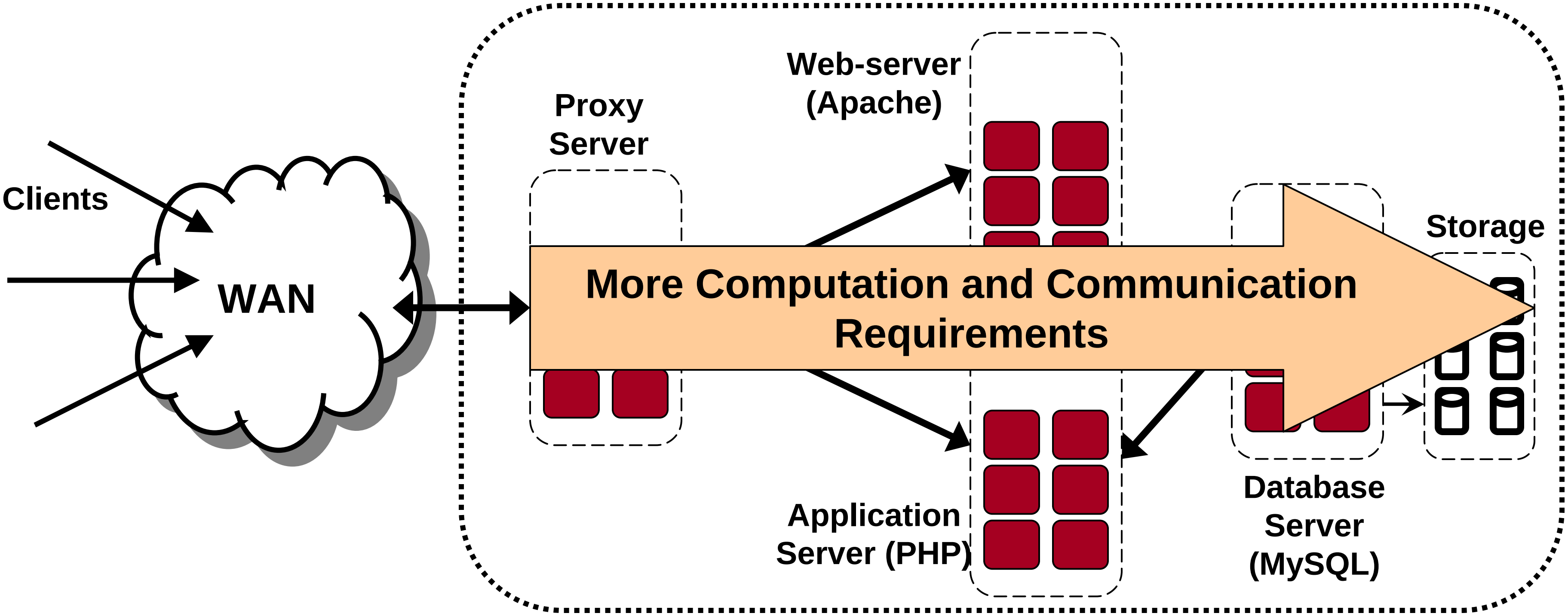






- Interactive Data-driven Applications
  - Scientific as well as Enterprise/Commercial Applications
    - Static Datasets: Medical Imaging Modalities
    - Dynamic Datasets: Stock value datasets, E-commerce, Sensors
  - Need for interacting, synthesizing and visualizing large datasets
  - Data-centers 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

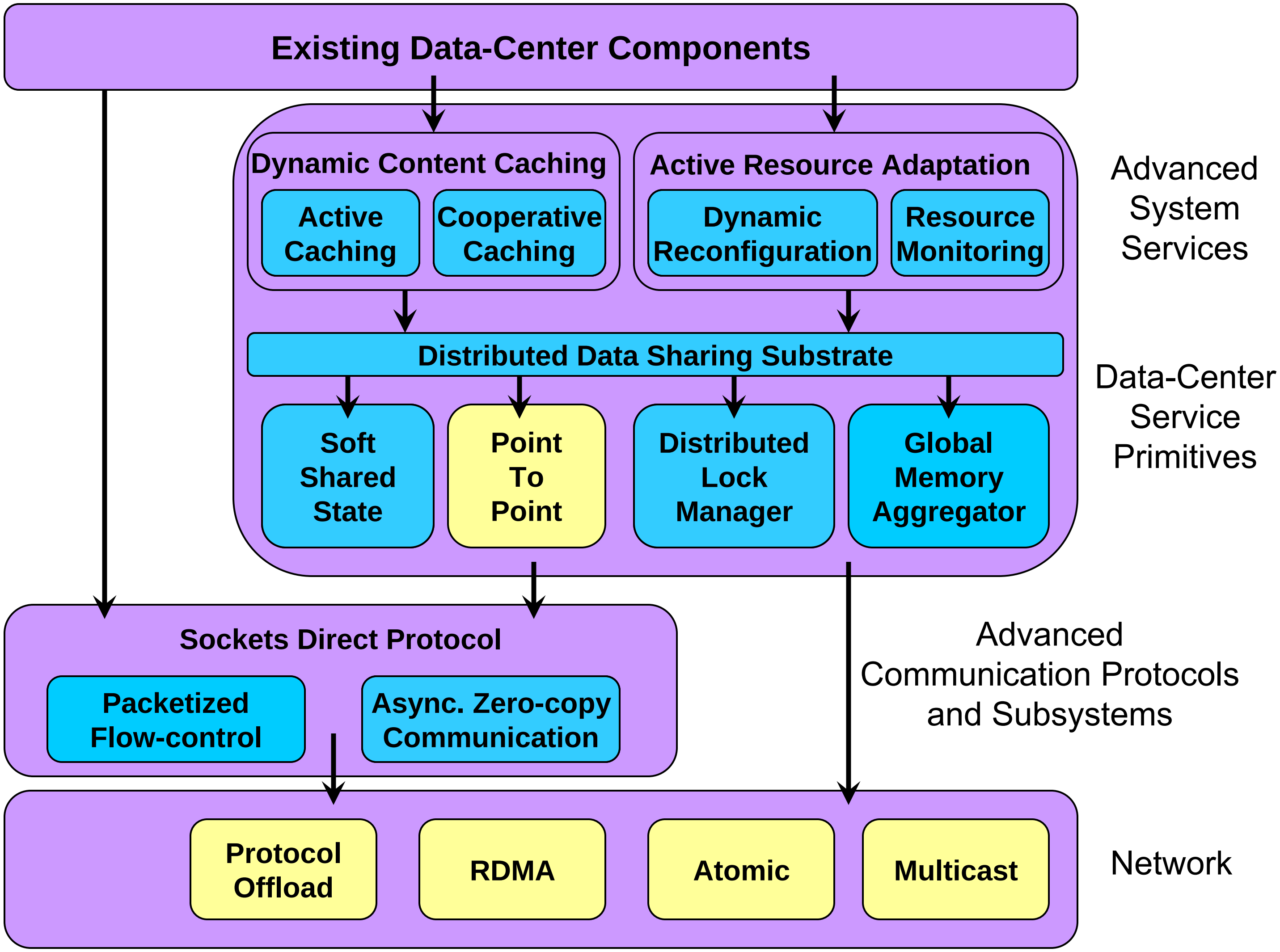


- Requests are received from clients over the WAN
- Proxy nodes perform caching, load balancing, resource monitoring, etc.
- If not cached, the request is forwarded to the next tiers → Application Server
- Application server performs the business logic (CGI, Java servlets, etc.)
  - Retrieves appropriate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

Overview of Research

- **Propose a novel framework for next generation data-centers**
  - Delivering performance and scalability
  - Providing advanced features such as active caching, fine-grain resource monitoring, dynamic resource adaptation, etc
- **Novel approaches using the advanced features of InfiniBand and other RDMA-enabled Networks**
  - Resilient to the load on the back-end servers
  - Order of magnitude performance gain for several scenarios
  - Exploit features like RDMA and remote atomic operations for new primitives and services
- **Three-layer Architecture**
  - Advanced Communication Protocol Support
  - Data-Center Primitives
  - Data-Center Services

Proposed Architecture

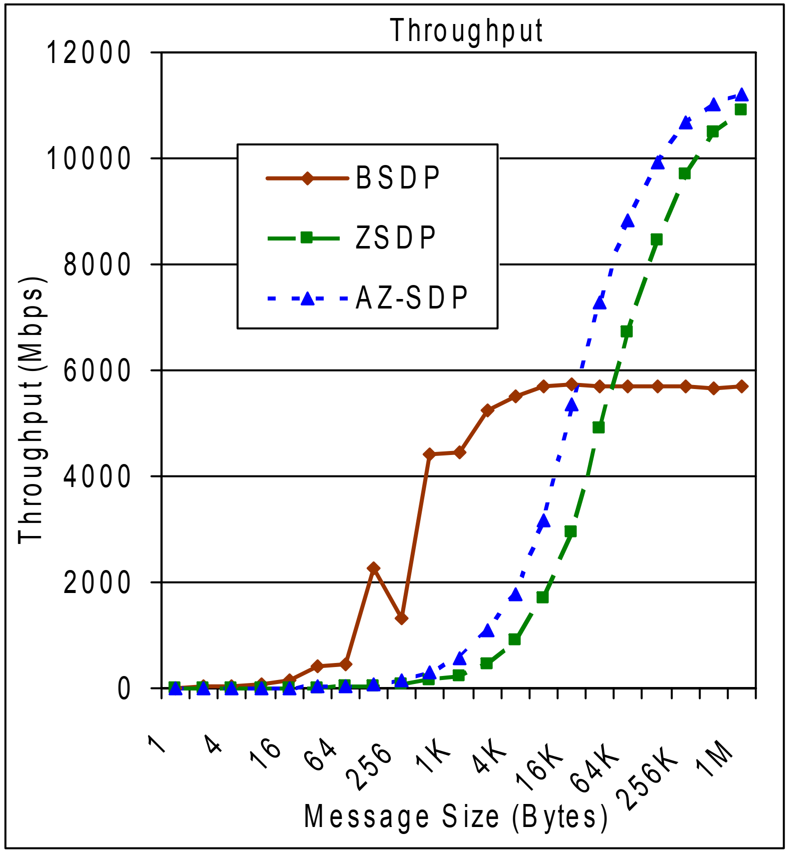


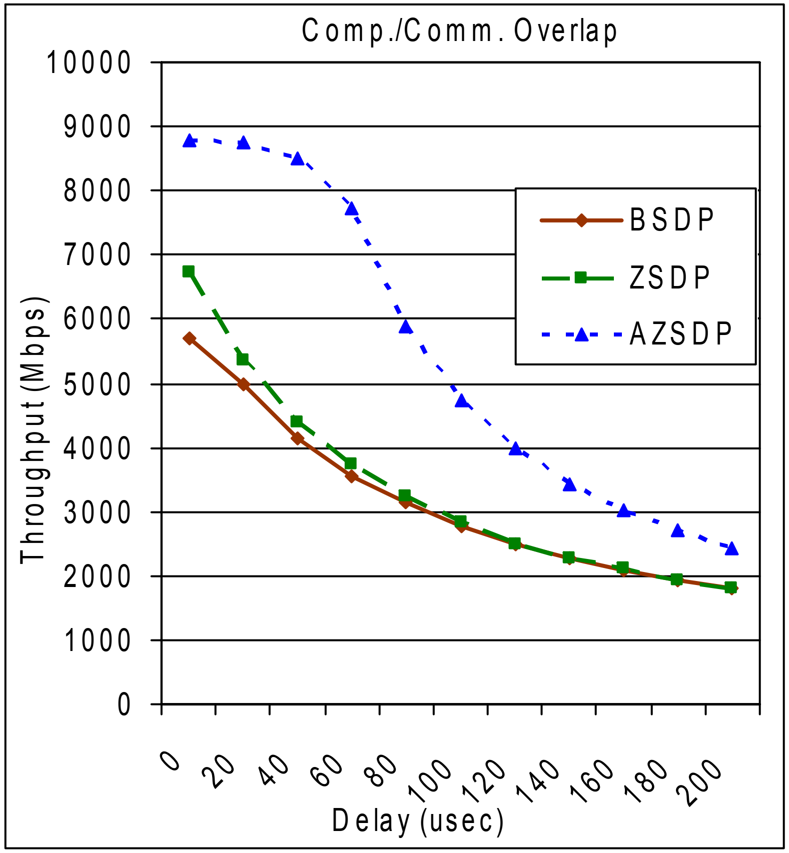
Publications (So Far)

- Architecture for Caching Responses with Multiple Dynamic Dependencies in Multi-Tier Data-Centers over InfiniBand, CCGrid 2005
- On the Provision of Prioritization and Soft QoS in Dynamically Reconfigurable Shared Data-Centers over InfiniBand, ISPASS 2005
- Asynchronous Zero-copy Communication for Synchronous Sockets in the Sockets Direct Protocol (SDP) over InfiniBand, CAC 2006
- Designing Efficient Cooperative Caching Schemes for Multi-Tier Data-Centers over RDMA-enabled Networks, CCGrid 2006
- Exploiting RDMA operations for Providing Efficient Fine-Grained Resource Monitoring in Cluster-Based Servers, RAIT 2006
- DDSS: A Low-Overhead Distributed Data Sharing Substrate for Cluster-Based Data-Centers over Modern Interconnects, HiPC 2006
- High Performance Distributed Lock Management Services using Network-based Remote Atomic Operations, CCGrid 2007

http://nowlab.cse.ohio-state.edu/projects/data-centers/index.html

Sockets Direct Protocol: Throughput and 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

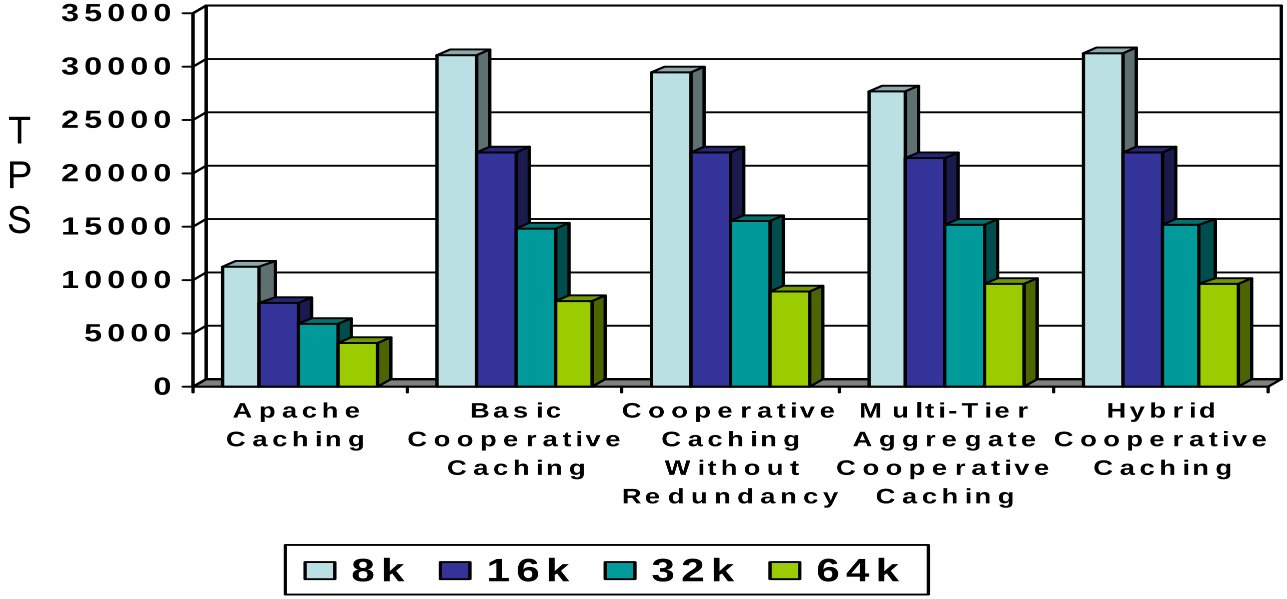
- Introduction and Motivation
- **Cooperative Caching Services**
- Resource Monitoring Services
- Conclusions and Ongoing Work

Cooperative Caching Services

- Aggregate cache benefits – well known!!
- Performance considerations
  - Two-sided operation vs. One-sided RDMA operations
  - Placement of data (Local Vs. Remote)
  - Controlling data redundancy
  - Utilize available remote memory
  - Load sensitive Protocols
- Objective
  - Can we design efficient cooperative caching schemes utilizing the idle resources in the Data-Centers and the RDMA capabilities in networks and eliminate redundancy to optimize available system cache size?

Data-Center Throughput with Cooperative Caching

8-Proxy nodes



- Our schemes achieve significant performance gain over basic Apache Caching (AC)

Designing Efficient Cooperative Caching Schemes for Multi-Tier Data-Centers over RDMA-enabled Networks, S. Narravula, H. -W. Jin, K. Vaidyanathanand D. K. Panda. In International Symposium on Cluster Computing and the Grid (CCGrid), 2006

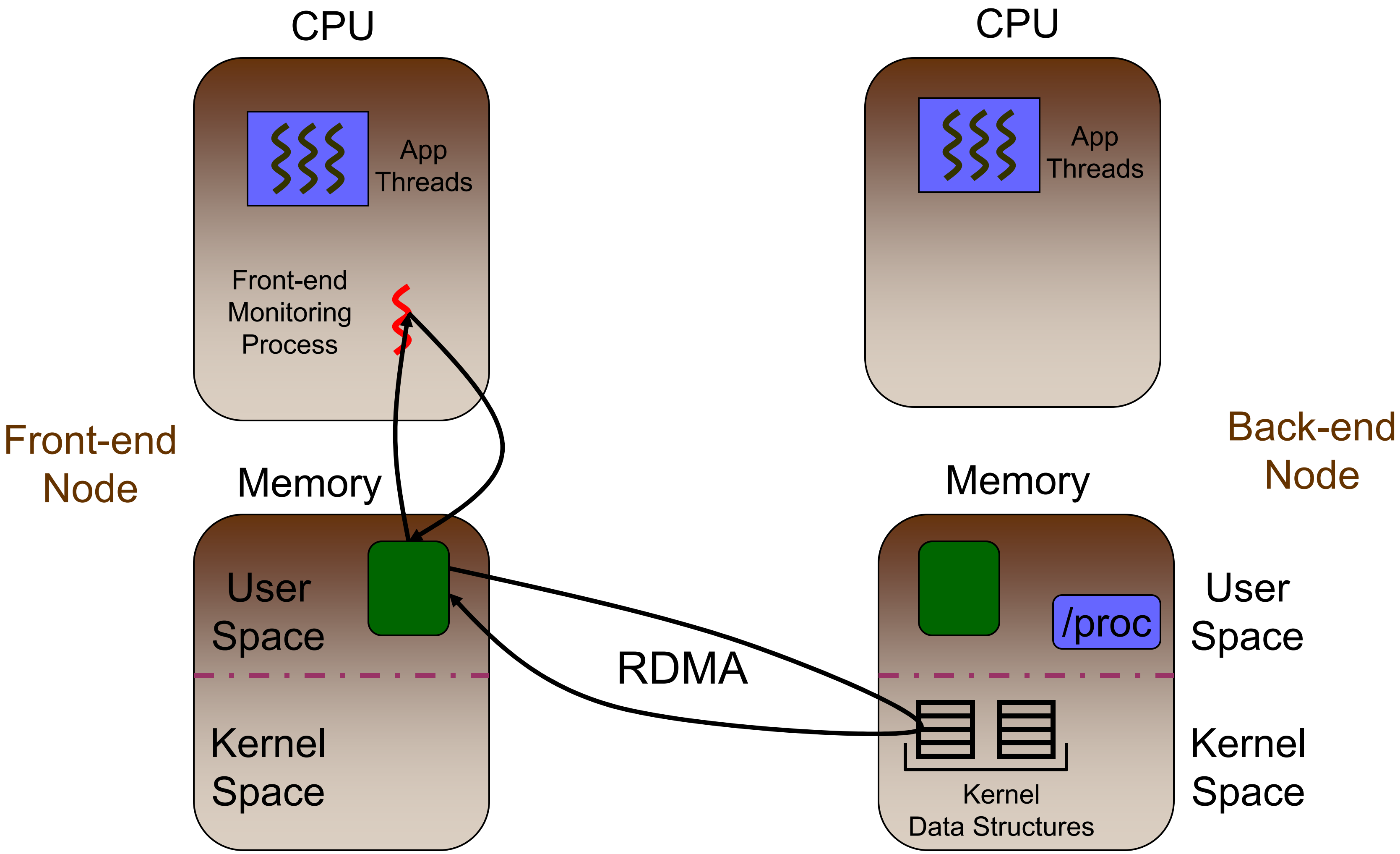
Presentation Layout

- Introduction and Motivation
- Data-center Service Primitives
- Cooperative Caching Services
- **Resource Monitoring Services**
- Conclusions and Ongoing Work

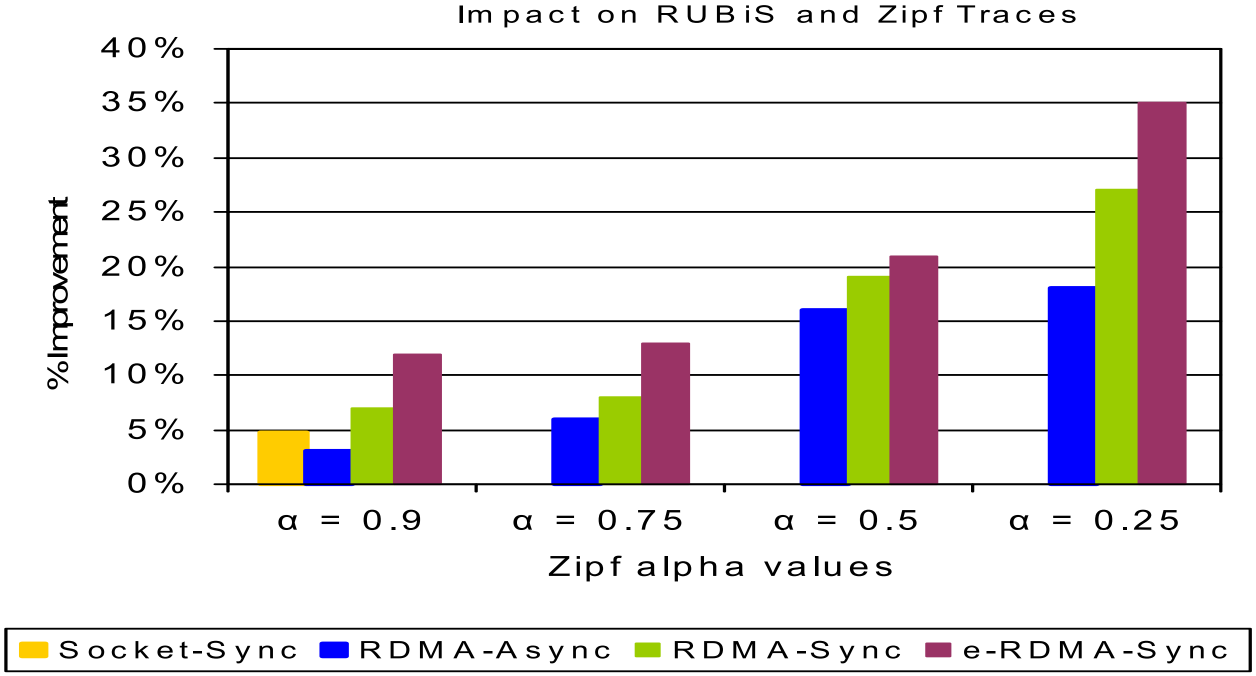
Resource Monitoring Services

- Traditional approaches
  - Coarse-grained in nature
  - Assume resource usage is consistent throughout the monitoring granularity (in the order of seconds)
- This assumption is no longer valid
  - Resource usage is becoming increasingly divergent
- Fine-grained monitoring is desired but has additional overheads
  - High overheads, less accurate, slow in response
- Can we design fine-grained resource monitoring scheme with low overhead and accurate resource usage?

Synchronous Resource Monitoring using RDMA (RDMA-Sync)



Impact of Fine-grained Monitoring with Applications



- Our schemes (RDMA-Sync and e-RDMA-Sync) achieve significant performance gain over existing schemes

Exploiting RDMA operations for Providing Efficient Fine-Grained Resource Monitoring in Cluster-Based Servers, K. Vaidyanathan, H. –W. Jin and D. K. Panda. Workshop on Remote Direct Memory Access (RDMA): Applications, Implementations and Technologies, 2006

Work-in-Progress

- Data-Center Primitives
  - Efficient Global Memory Aggregator Mechanisms
- Advanced Communication Protocol Mechanisms
  - Efficient Packetized Flow-Control
- Detailed Data-Center Evaluation with the proposed framework
- Software release of several data-center components
  - Have received multiple requests from organizations for such a release including a large financial company

Conclusions

- Proposed new protocols, primitives and services for next generation data-centers
  - Use advanced features of InfiniBand and other RDMA-Enabled interconnects
  - Significant performance gains and scalability for several scenarios
- Potential for designing next generation scalable and high performance data-center architectures

Challenges and Discussion Bullet

- Challenges
  - Benefits of all these components and services in an integrated manner for handling
    - Petabytes of data and Multi-thousand users
  - Redesigning middleware and applications on next generation data-centers
- Significance to the SMA and PDOS components of the CSR program
- Discussion Bullet
  - How to re-architect next generation data-center architectures, software services, middleware and applications with advances in modern networking technologies and capabilities?

Web Pointers

Network Based Computing Laboratory

NBCL

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

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

Email: panda@cse.ohio-state.edu