System Software for High-Performance Communication

Arthur Maccabe
University of New Mexico

http://lacsi.rice.edu/reviews/slides_2006

LACSI

Overview

- Goal
  - —Building scalable systems software
  - —Scaling requires that we understand how resources are used and the benefits of adding resources
- Approach
  - —Modeling to understand how resources might be used
    - – Modeling protocol offload
  - —Monitoring to understand how applications use resources
    - – System call monitoring
    - – Message-centric monitoring
  - —Implementation to illustrate scaling
    - – Open MPI on Infiniband
- Faculty
  - —Bridges & Maccabe

LACSI

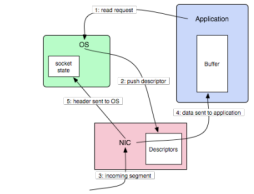
Modeling Protocol Offload

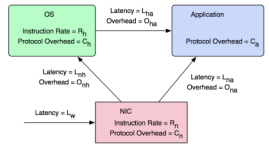
- Publications
  - —"Modeling Protocol Offload for Message-oriented Communication," IEEE Cluster 2005
  - —"An Extensible Message-Oriented Offload Model for High-Performance Applications," LACSI Symposium 2005
- Patricia Gilfeather, PhD, Fall 2005
- Basic Idea
  - —Develop a model to explore the benefits of *partial* protocol offload.
    - – Leave error processing on host processor
    - – Minimize NIC resources while maximizing performance
  - —Specific interest in partial offload of TCP/IP (commodity protocol).
- Highlight
  - —Working with SeaFire to explore commercial product based on partial offload.
  - —Specific application Grid FTP at 40Gbps.

LACSI

Modeling Protocol Offload

- Partial Offload (Splintered Implementations)
  - —Isolate functionality
  - —Distribute functionality
  - —Constrain NIC resources
- EMO (Extensible Message-Oriented Model)
  - —An extension of LogP
  - —Focus on NIC design – "what if designs"
    - – Where to add resources
    - – Benefits to additional offload
- Comparison to LAWS and LogP
- Initial Validation



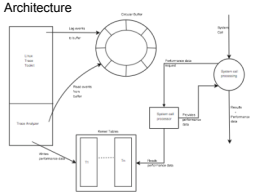


LACSI

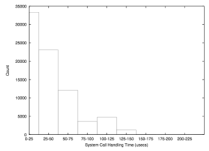
System Call Monitoring

- **Publication**
  - —"A Framework for Analyzing Linux System Overheads on HPC Applications," LACSI Symposium 2005
- **Sushant Sharma, MS Summer 2005, now in CCS1**
- **Basic idea**
  - —Need to monitor system calls to better understand the interaction between operating system implementation and application performance
  - —Extend LTT (Linux Tracing Toolkit)
- **Highlight**
  - —Monitoring overhead is minimal (~3-6%)

Architecture



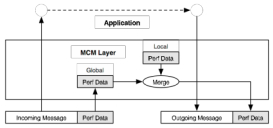
Distribution of time spent in recvfrom syscall (NAS CG-B)



LACSI

Message-Centric Monitoring

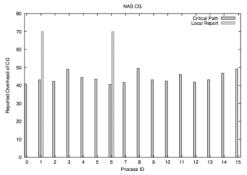
- **Publications**
  - —"Online Critical Path Profiling for Parallel Applications," IEEE Cluster 2005
- **Wenbin Zhu, PhD student**
- **Basic idea**
  - —Tag local information into application messages
  - —Merge local data with data from incoming messages to provide global view (critical path)
- **Highlight**
  - —<3% overhead
  - —Able to detect injected overhead
  - —Online analysis

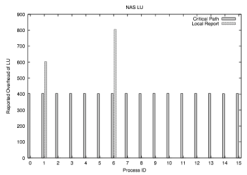


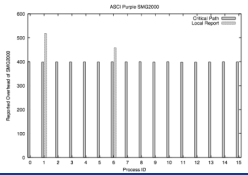
LACSI

Message-Centric Monitoring

- **Online analysis**







Nodes 1 & 6 detect additional overhead

LACSI

Open MPI on Infiniband

- **Publication**
  - —"Infiniband Scalability in OpenMPI," IPDPS 2006
- **Galen Shipman, MS Fall 2005, now in CCS1**
- **Basic idea**
  - —Develop and Open MPI implementation for Infiniband
- **Highlight**
  - —Compared to MVAPICH
    - Small message latency improved by 10%
    - Per host memory usage decreased by 300%
    - Latency is highly predictable

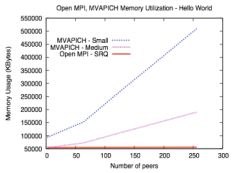
LACSI

Open MPI on Infiniband

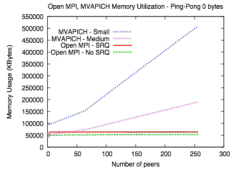
- **Approach**
  - —**MVAPICH**
    - **Uses pure RDMA protocol**
    - **Allocates a buffer per peer**
      - **Memory use**
      - **Latency depends on node index**
        - **good latency for node 0 or 1 :)**
  - —**Open MPI**
    - **Uses send/receive for small messages**
    - **Uses a Shared Request Queue (SRQ) for completions**
      - **Issue: no flow control when using SRQ!**
    - **Dynamic allocation of communication buffers reflects actual use of communication resources by application**

LACSI

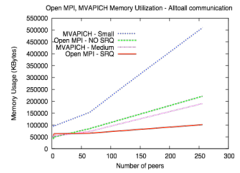
Open MPI on Infiniband



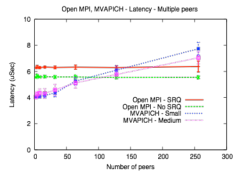
Hello world Memory Usage



Pairwise Ping-pong Memory Latency



All-to-all Memory Usage

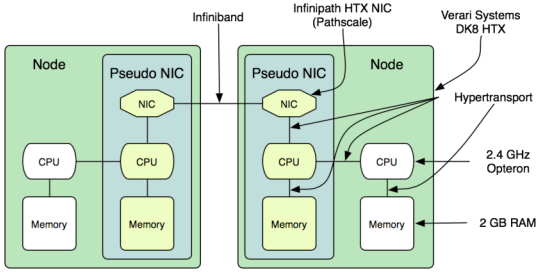


Multi-node Zero Byte Latency

LACSI

Future

- **Offload Testbed – beyond modeling, implement and measure**



LACSI

Summary

- **Focus on**
  - —**Improving communication performance**
  - —**Measuring performance**
- **Degrees**
  - —**Two MS degrees**
    - **Sushant started working for Wu Feng and is now working for Ron Minnich**
    - **Galen is working for David Daniel**
  - —**One PhD**
    - **Patricia is currently working as a postdoc at UNM**
- **Relevance to WSR**
  - —**Need to understand and manage resource usage**
    - **Monitor to understand**
    - **Manage resources to build scalable systems**

LACSI