System Software for High Performance Communication

Protocols, Tools, and Techniques for Commodity Systems

Patrick Bridges and Arthur Maccabe

Department of Computer Science

University of New Mexico

http://lacsi.rice.edu/review/2004/slides/sys-comm.pdf

LACSI

Ultra-scale Systems Software

- System software a limiting factor for current ASC systems
  - Basic system services can already limit ASC system scalability
  - ~15% of ASC Q processors dedicated to system software
  - Custom system software solutions limit system programmability
- Life will keep getting harder for system software
  - Petaflops systems may have 100,000+ processing elements
  - Heterogeneous processing elements (e.g. PIM systems)
  - Applications getting more diverse
    - Multi-physics codes
    - Static system software configuration insufficient
- Long-term goal: Address long-term system software challenges for ultra-scale systems

LACSI Research at UNM

- Commodity Systems Software
  - Freely available, widely used, familiar to application scientists
  - Well studied, understood and optimized for its environment
  - Not designed for use in large-scale scientific environments
- Scaling up Commodity
  - Long-term: Commodity system software for ultra-scale systems
  - Short-term: Measure and improve ASC application performance and scalability with improved system software for ASC systems
  - Method: analyze, measure, and modify existing commodity system software in consultation with LANL ACL
- Remainder of talk
  - Brief overview of past work
  - Snapshot of current work
  - Interactions with LANL, other tri-labs

LACSI

Past Research

- Splintering and offloading of TCP functionality to programmable NICs (Myrinet, Intel ACENICs)
  - Understand impact of TCP offloading on communication primitives used by ASC applications
  - Results in improved communication primitive performance
- Use a dedicated CPU for Linux OS/networking processing
  - Increase predictability of OS processing costs to improve application scalability and performance predictability
  - Possible with only minor changes to Linux kernel
- Careful changes to commodity system software can have substantial performance improvements to the system software that manages ASC machines

Current LACSI Research

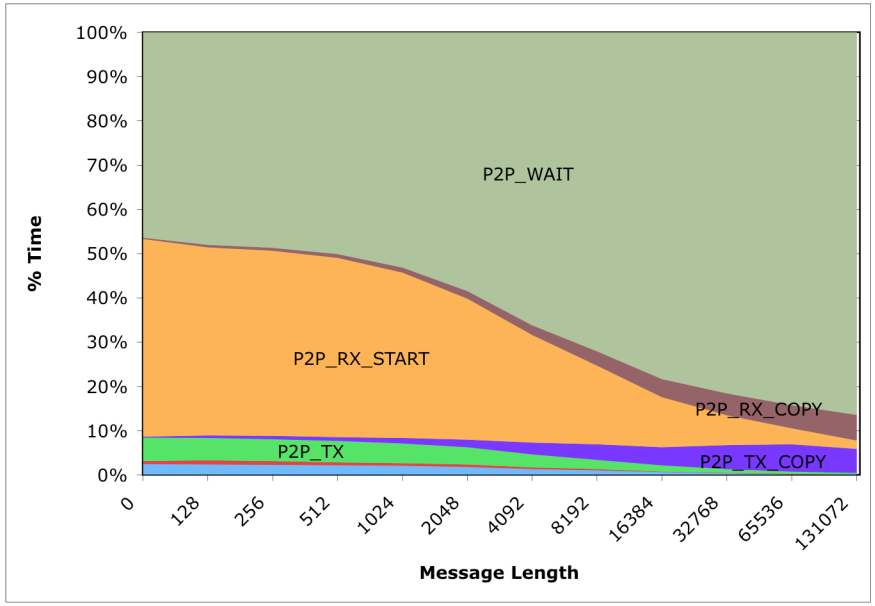
- System software monitoring techniques
  - Sophisticated tools to diagnose and measure system software effects in large-scale systems
  - Monitoring system software effects on ASC-related applications to guide future system software research
- LA-MPI/Open MPI Reliability protocols
  - Understand inherent and implementation-dependent performance/reliability tradeoffs in application runtime libraries
  - Examine improvements to enhance MPI runtime performance in the common case without sacrificing reliability in the uncommon case
- TCP scalability enhancements
  - Enhance TCP scalability in large-scale commodity clusters
  - Take advantage of homogeneous nature of cluster networks

System Measurement Techniques

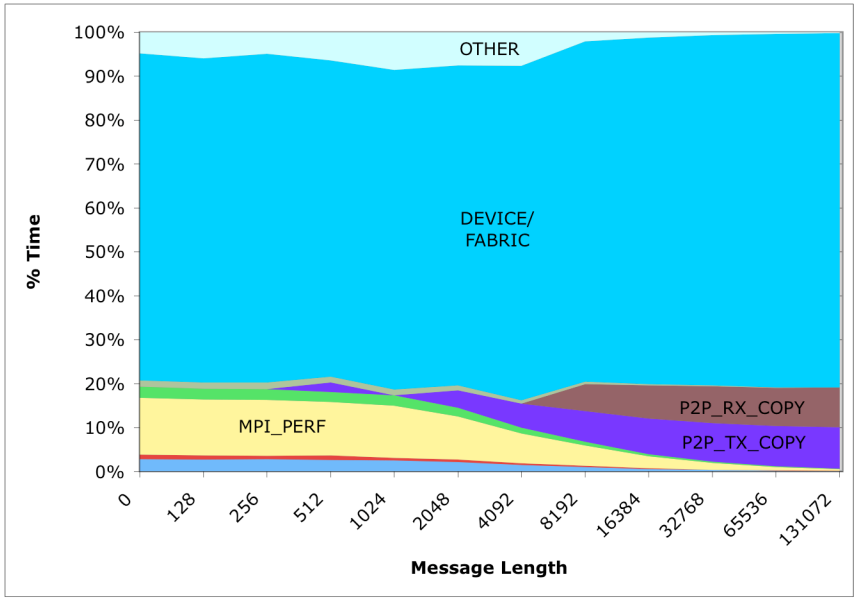
- Goal: accurately measure system software performance in large-scale systems
  - Cross-host operating system and networking interactions effect application performance (e.g. SAGE on ASC Q)
  - Online data availability for system software adaptation research
- Problem: Existing monitoring systems too heavyweight, no online availability, or do not measure cross-host interactions
- Solution: IMPuLSE - Integrated Monitoring and Profiling for Large-Scale Environments
  - Message-centric monitoring approach that associates a *performance summary* with each transmitted message
  - Propagate data from incoming summaries to outgoing summaries
  - Use blocking MPI calls to determine performance causality

IMPuLSE Proof of Concept

- Proof-of-concept host/message-centric monitoring comparison
- MPI ping-pong test over Myrinet using 2GHz PIII Xeons



(a) Host-based Monitoring



(b) Message-based Monitoring

LACSI

LA-MPI/Open MPI Reliability Protocols

- Goal: Understand and decrease the reliability costs in LA-MPI
  - Reliability protocols needed for large-scale systems
  - Existing LA-MPI protocols implement reliability
  - Reliability costs may be on the critical path, but packet loss and corruption is relatively uncommon in most HPC systems
- Results: Studied the LA-MPI reliability protocol, discussing and prototyping improvements
  - LA-MPI lacks well-understood optimizations (e.g. piggybacking of acks) that improve networking performance in the common case
  - More sophisticated protocols (e.g. the BLAST RPC protocol) may be more suited to LA-MPI and Open MPI long-message transfer needs
  - Working with Rich Graham et al. in CCS-1 on Open MPI reliability issues

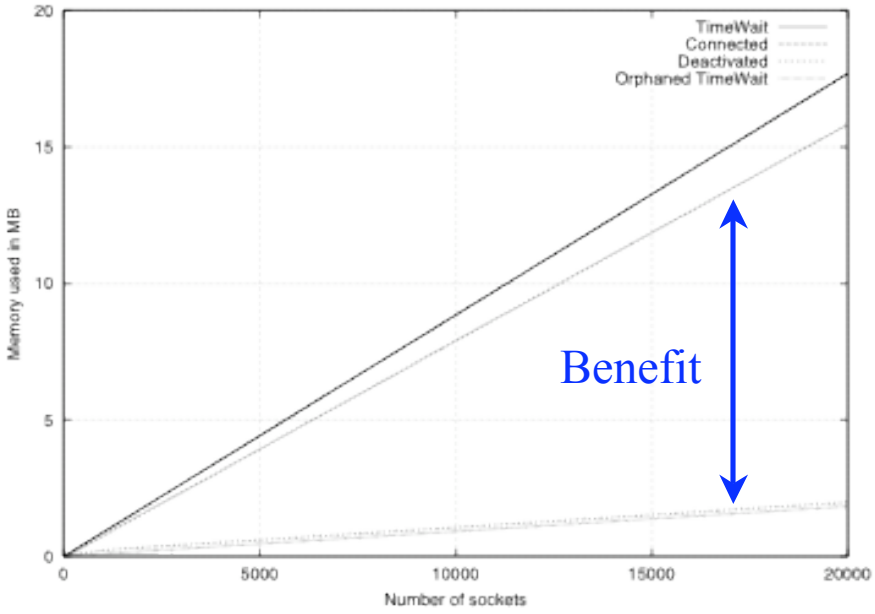
TCP Scalability Enhancements

- Goal: Increase the scalability of TCP for large-scale systems
  - Connections to visualization back-ends
  - Communication with file servers, external data sources
  - Mainline communication protocol in low-cost (e.g. Ethernet) systems
- Problem: TCP connections require a large amount of state that consumes too much memory for offloading to iNICs or TOEs
- Solution: Deactivate idle connections until needed and then reactivate without the TCP three-way-handshake
  - Keep a "working set" of active TCP connections
  - Using existing Linux minisocks infrastructure for deactivation
  - Take advantage of the homogeneous nature of HPC networks
  - Compromise point between on-demand connection establishment and opening all possible connections at program startup

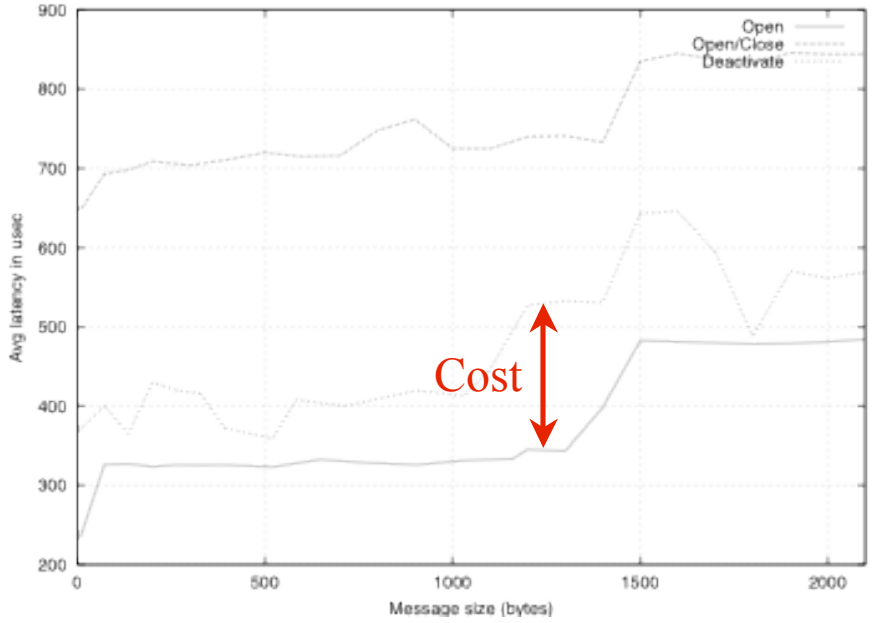
LACSI

Scalability Enhancements Results

- **Benefits of deactivation/reactivation**
  - —Order of magnitude memory savings versus full TCP connections
  - —40% reactivate savings versus on-demand connection establishment



(a) Memory Usage



(b) Latency

LACSI

Future System Software Work

- IMPuLSE
  - Port to new UNM/LACSI cluster on arrival (Opteron/Infiniband)
  - Application/OS studies on medium and large-scale systems
  - Produce data for analysis with HPCToolkit
- LA-MPI/Open MPI
  - Optimized reliability protocol support for Open MPI
  - BLAST protocol support in LA-MPI/Open MPI
  - Evaluation of Infiniband bit-error rates
- TCP/Linux
  - Design and evaluation of automatic connection deactivation policies
  - NS2-based simulation studies for deactivation/reactivation
  - Initial experimentation with offloading bundled connections

LANL and other DOE Interactions

- Los Alamos Interactions
  - Working with Ron Minnich, Sung-Eun Choi on TCP enhancements
  - Working with Rich Graham on LA-MPI, Open MPI protocols
  - Faculty and domestic students visit LANL collaborators regularly
- LACSI Benefits to UNM
  - Expose LANL problems to UNM faculty and students
  - Involve new faculty with LANL
  - Seed money for innovative research projects
- Other ASC/Tri-labs Interactions
  - Working with SNL on non-commodity operating systems
  - Working with LLNL on porting K42 research operating system to ASC Red Storm hardware to
  - Facilitate comparisons with commodity and custom approaches

LACSI