

Component PAPI: Performance Measurement Beyond the CPU



Presented by

Jack Dongarra

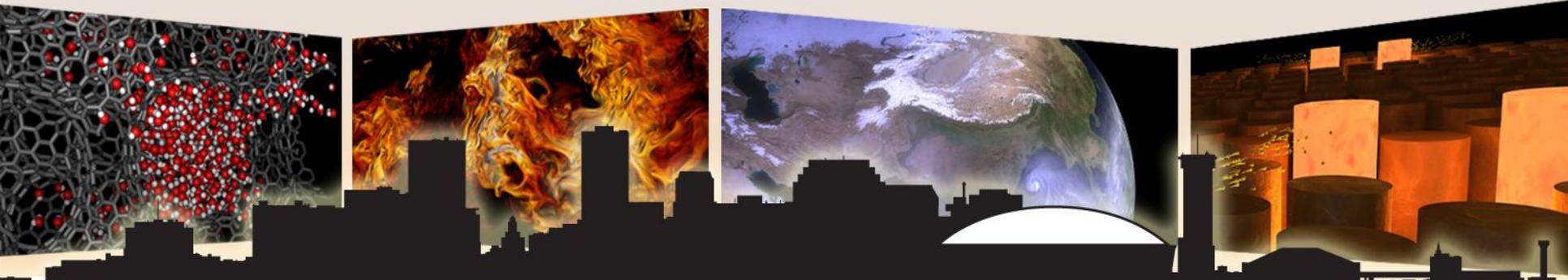
Heike Jagode

Shirley Moore

Dan Terpstra

University of Tennessee

Oak Ridge National Laboratory



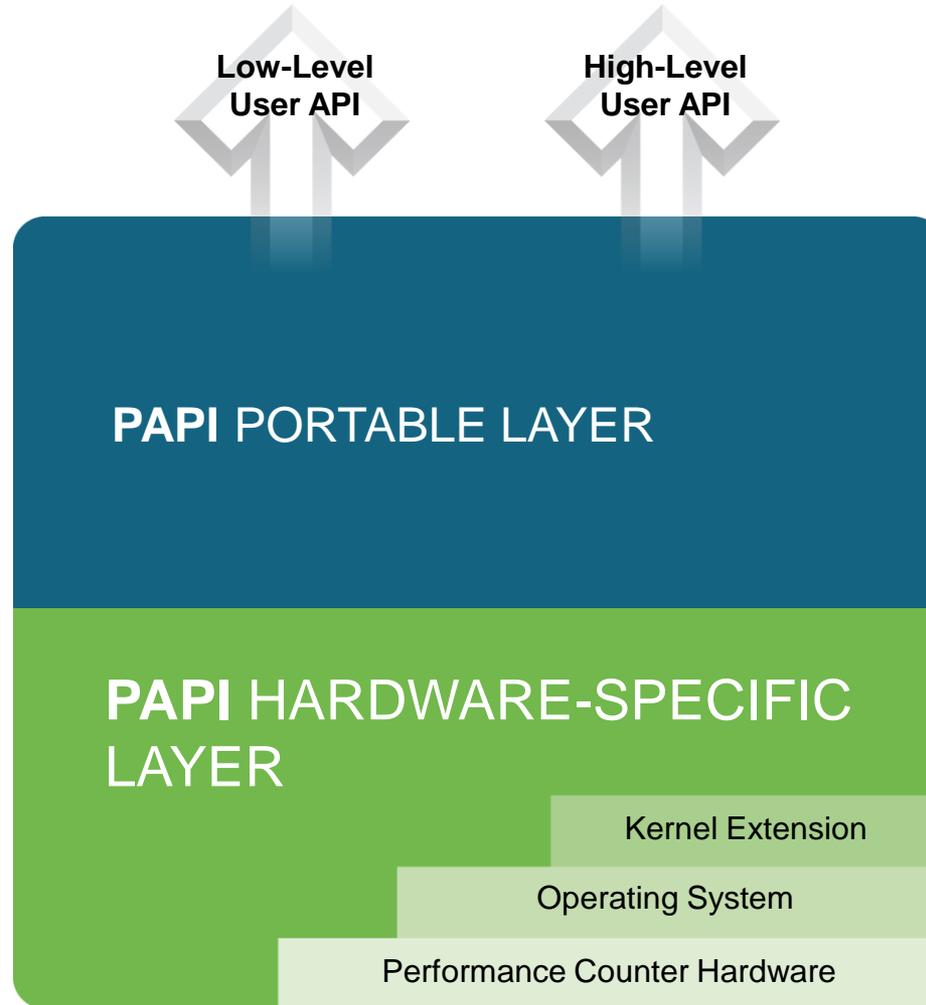
THE UNIVERSITY of
TENNESSEE
Department of Electrical Engineering
and Computer Science

Introduction

- **PAPI** has provided a consistent programming interface for performance counter hardware
- **PAPI-C** extends that interface to multiple performance counter domains
- Interesting performance phenomenon can be measured throughout high performance computing systems:
 - File systems
 - Network fabrics
 - System health characteristics
- **PAPI-C** can provide the interface between user level performance tools and low level performance measurements
- Third parties can develop **PAPI-C** components for specialized hardware

Monolithic “PAPI Classic”

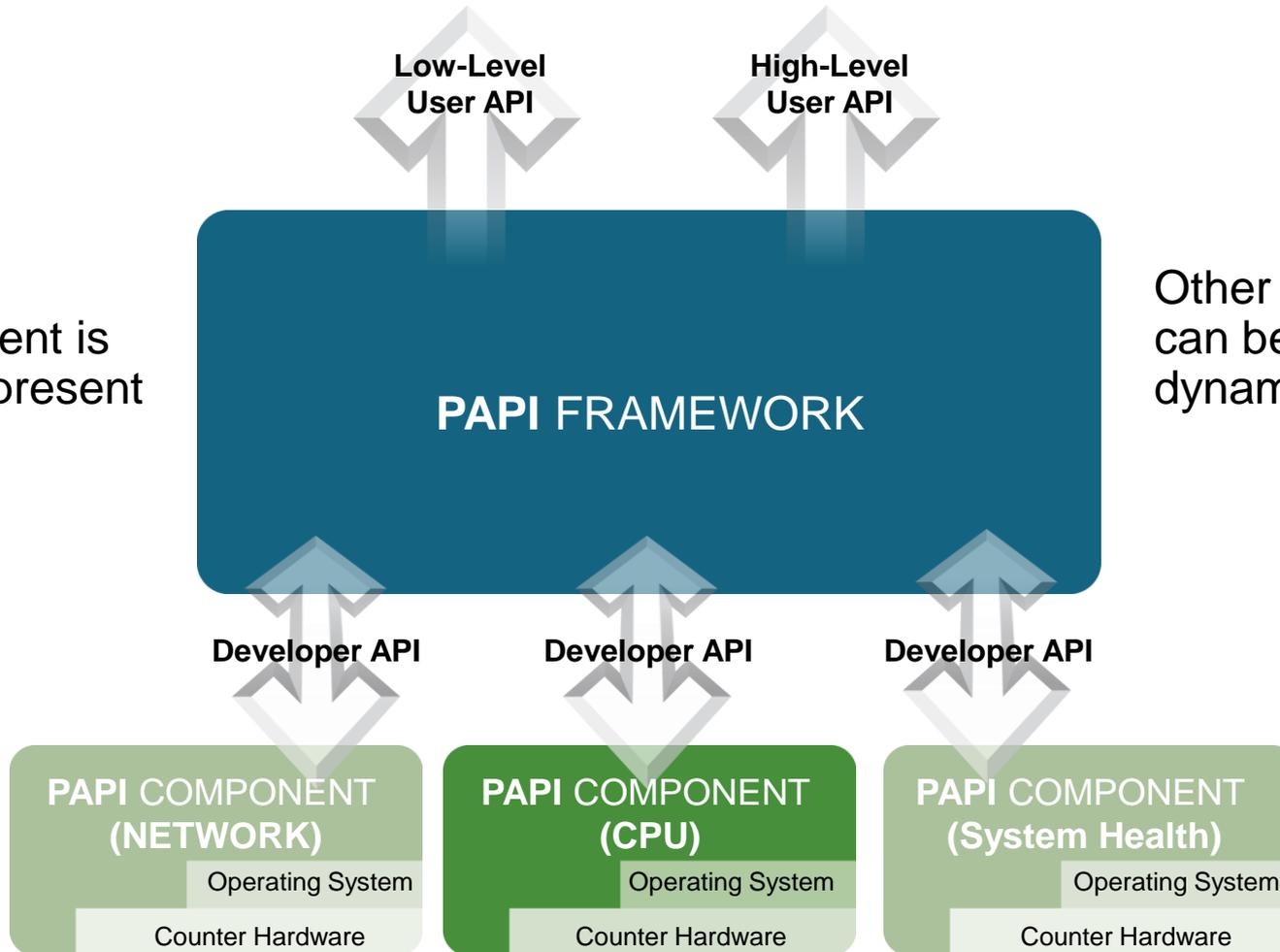
Portable hardware-independent code and hardware-specific code...



are linked with low-level libraries and drivers to form a monolithic library

Component PAPI

A CPU component is always present



Other components can be added dynamically

File system components: Lustre

- **Measures data collected in: /proc/.../stats
and: /proc/.../read_ahead_stats**

Hits	631592284
misses	9467662
readpage not consecutive	931757
miss inside window	81301
failed grab_cache_page	5621647
failed lock match	2135855
read but discarded	2089608
zero size window	6136494
read-ahead to EOF	160554
hit max r-a issue	25610

- **Snippet of available native events for Lustre:**

0x44000002	fastfs_llread	bytes read on this lustre client
0x44000003	fastfs_llwrite	bytes written on this lustre client
0x44000004	fastfs_wrong_readahead	bytes read but discarded due to readahead
0x44000005	work_llread	bytes read on this lustre client
0x44000006	work_llwrite	bytes written on this lustre client
0x44000007	work_wrong_readahead	bytes read but discarded due to readahead

System health components: Im-sensors

- Access computer health monitoring sensors, exposed by Im_sensors library
- User can closely monitor the system's hardware health
 - Observe feedback between performance and environmental conditions
- Available features and monitored events depend on hardware setup
- Snippet of available native events for Im-sensors:

...

0x4c000000 LM_SENSORS.max1617-i2c-0-18.temp1.temp1_input

0x4c000001 LM_SENSORS.max1617-i2c-0-18.temp1.temp1_max

0x4c000002 LM_SENSORS.max1617-i2c-0-18.temp1.temp1_min

...

0x4c000049 LM_SENSORS.w83793-i2c-0-2f.fan1.fan1_input

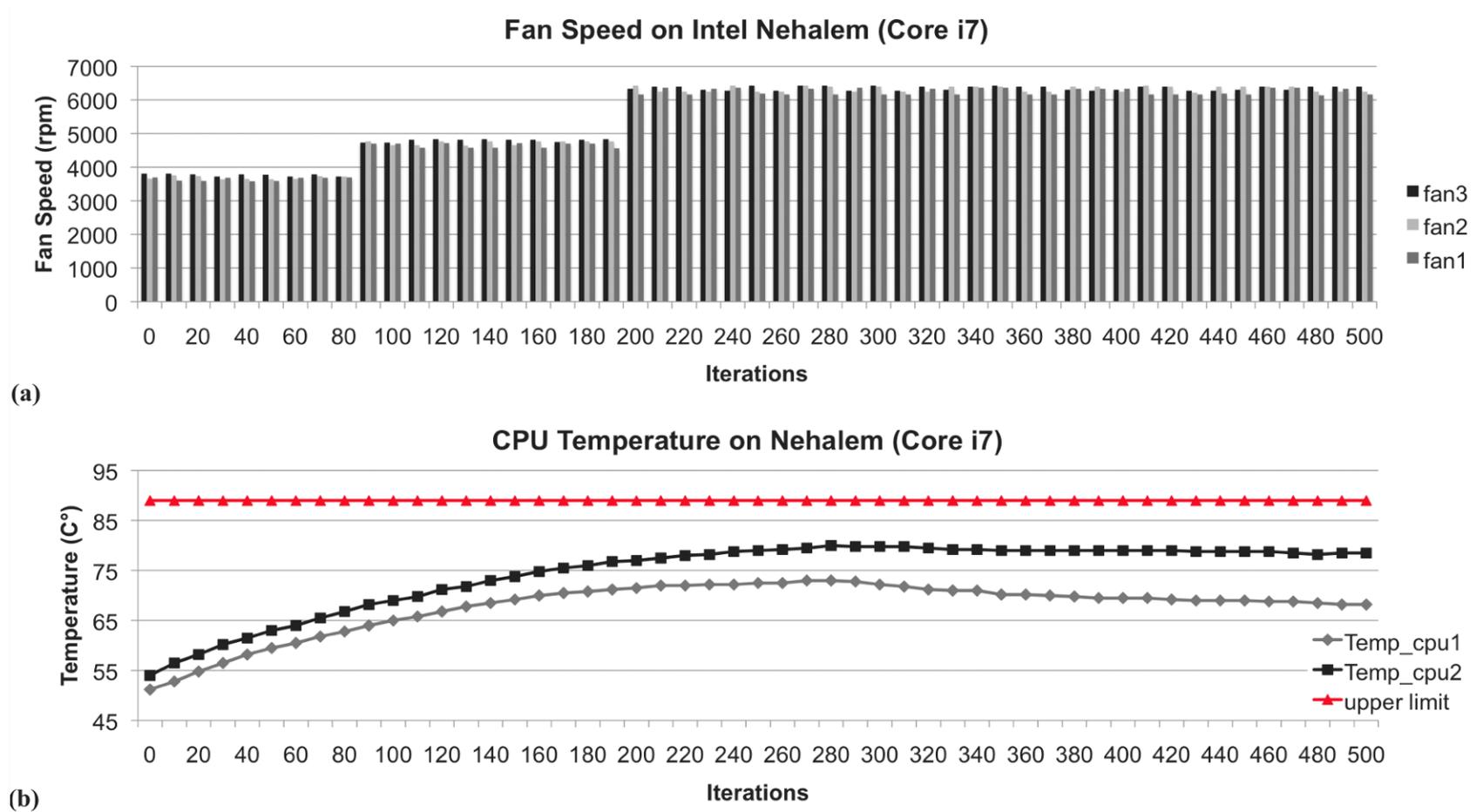
0x4c00004a LM_SENSORS.w83793-i2c-0-2f.fan1.fan1_min

0x4c00004b LM_SENSORS.w83793-i2c-0-2f.fan1.fan1_alarm

...

Im-sensors component example

libsensors version 3.1.1



Network components: InfiniBand

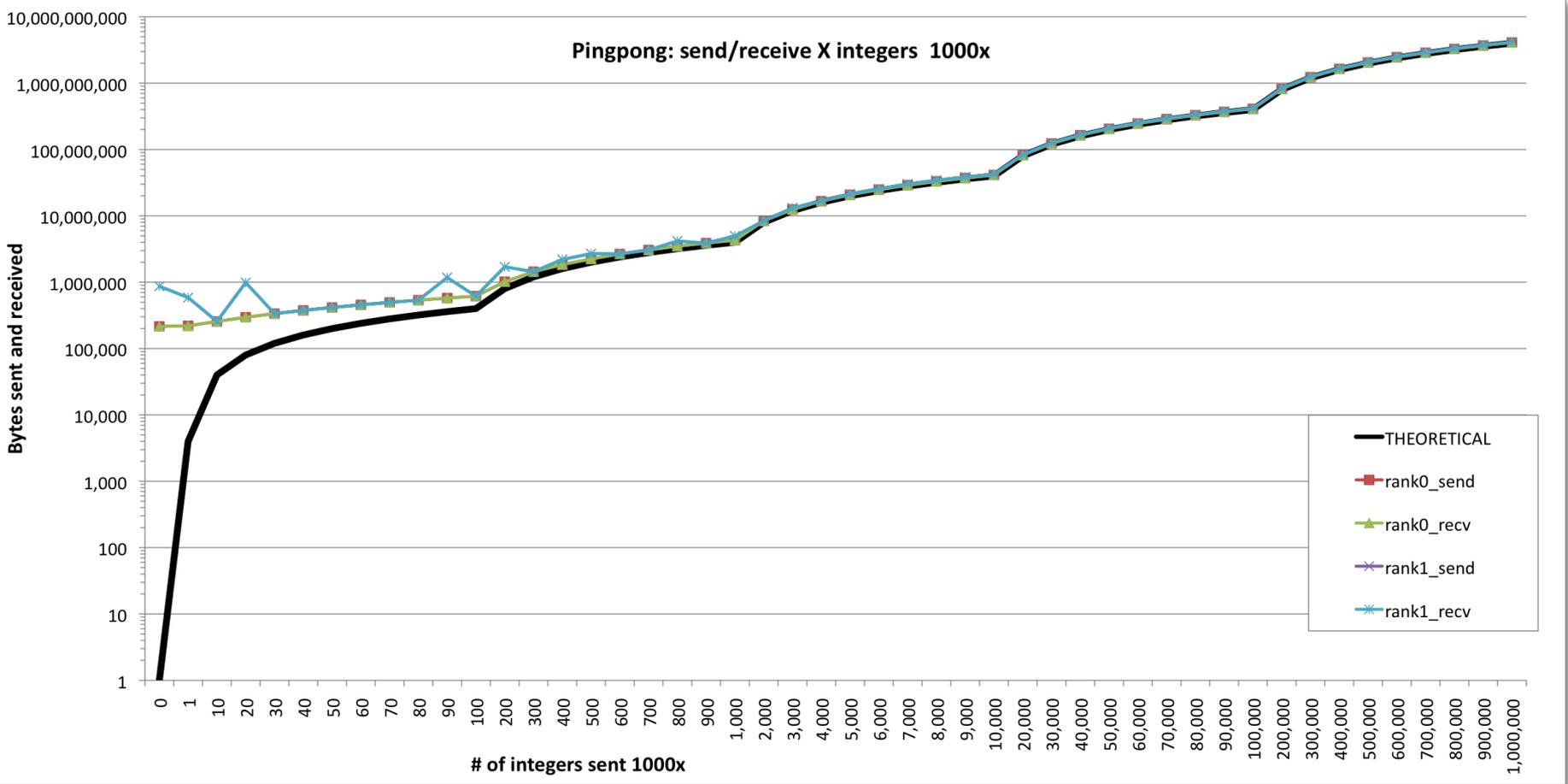
- Measures everything that is provided by the libibmad:
- Errors, bytes, packets, local IDs (LID), global IDs (GID), etc.
- ibmad library provides low-layer IB functions for use by the IB diagnostic and management programs, including MAD, SA, SMP, and other basic IB functions
- Snippet of available native events on a machine with 2 IB devices, mthca0 and mthca1:

...

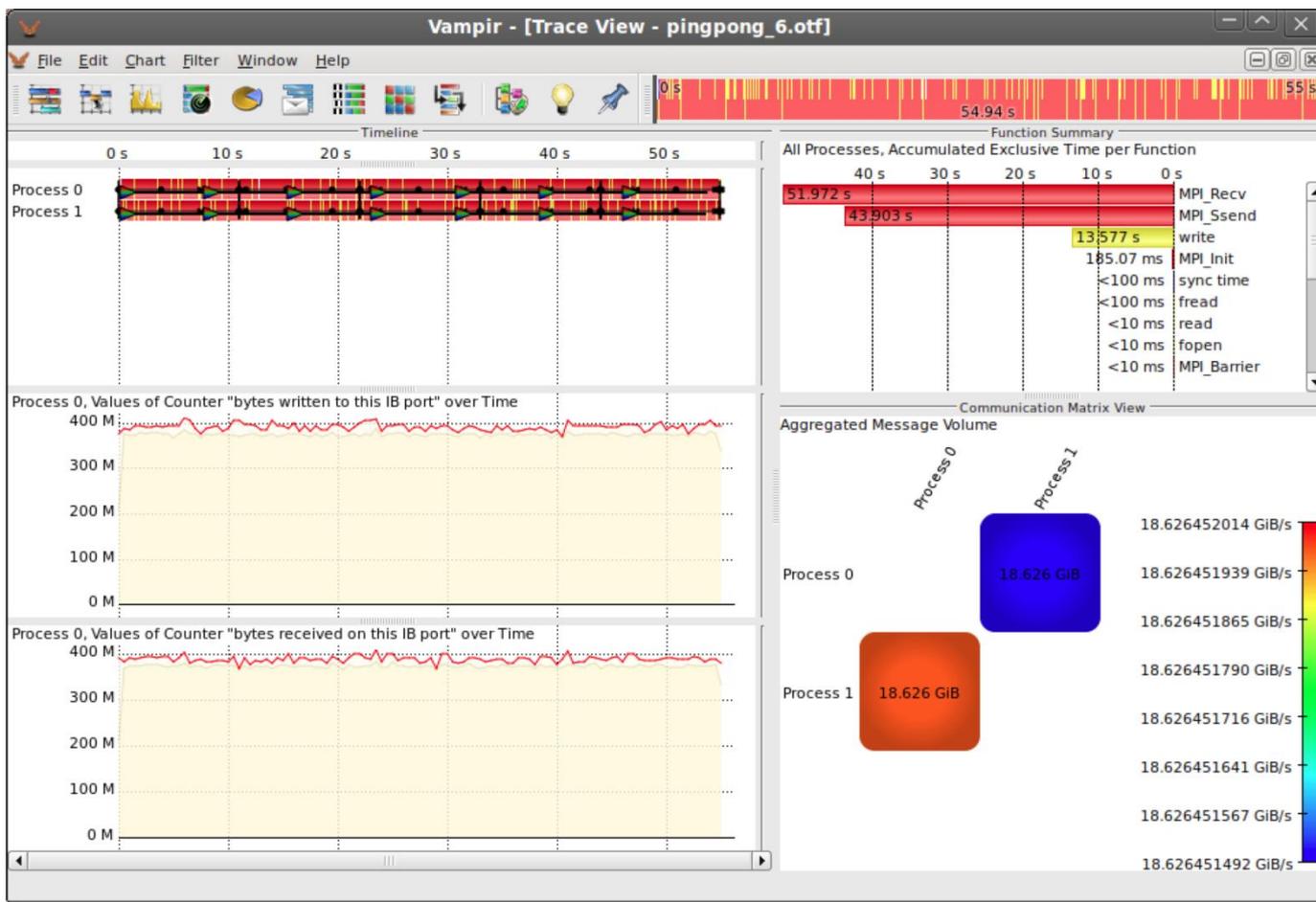
0x44000000	mthca0_1_recv	bytes received on this IB port
0x44000001	mthca0_1_send	bytes written to this IB port
0x44000002	mthca1_1_recv	bytes received on this IB port
0x44000003	mthca1_1_send	bytes written to this IB port

...

InfiniBand component results



InfiniBand events measured over time (via Vampir linked with PAPI)



Run Pingpong 5x: send 1,000,000 integers 1000x (theor: ~19 GB)

Conclusion and future directions

- **Component PAPI** provides performance measurement beyond the CPU
- Increasing CPU speeds and densities places greater importance on
 - Thermal health and management
 - Power consumption
- Higher processor counts make communication metrics more critical (bandwidth, latency, how many bytes transferred)
- Third parties can develop and contribute specialized components
- User-level performance tools can access multiple components with a common interface

Contact

Dan Terpstra

UTK Innovative Computing Laboratory
terpstra@eecs.utk.edu

For more information

<http://icl.cs.utk.edu/papi/>

- **Software and documentation**
- **Component interface details**
- **Reference materials**
- **Papers and presentations**
- **Third-party tools**
- **Mailing lists and User Forum**

Acknowledgments

This work used resources of the National Center for Computational Sciences at Oak Ridge National Laboratory, which is supported by the Office of Science of the Department of Energy under Contract DE-AC05-00OR22725.

These resources were made available via the Performance Evaluation and Analysis Consortium End Station, a Department of Energy INCITE project.

This work was also supported in part by the U.S. Department of Energy Office of Science under contract DE-FC02-06ER25761; by the National Science Foundation under Grant No. 0910899, as well as Software Development for Cyberinfrastructure (SDCI) Grant No. NSF OCI-0722072 Subcontract No. 207401; and by the Department of Defense, using resources at the Extreme Scale Systems Center.

