Analysis of Energy and Power Consumption by Remote Data Transfers in Distributed Programming Models

Siddhartha Jana (sidjana@cs.uh.edu), Barbara Chapman (Advisor) (chapman@cs.uh.edu)
University of Houston, Texas, 77204

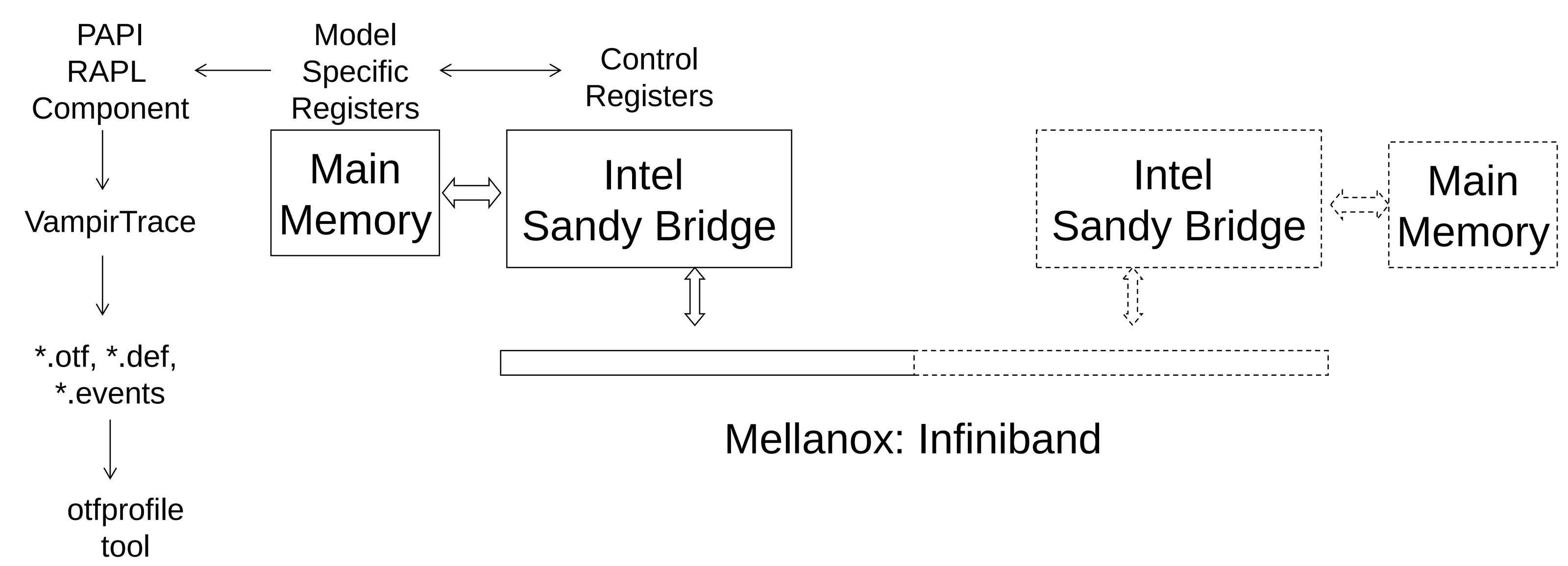
Abstract:

Recent studies of the challenges facing the Exascale era express a need for understanding the impact on energy profile of applications due to inter-process communication on large-scale systems. Programming models like MPI provide the user with explicit interfaces to initiate data-transfers among distributed processes.

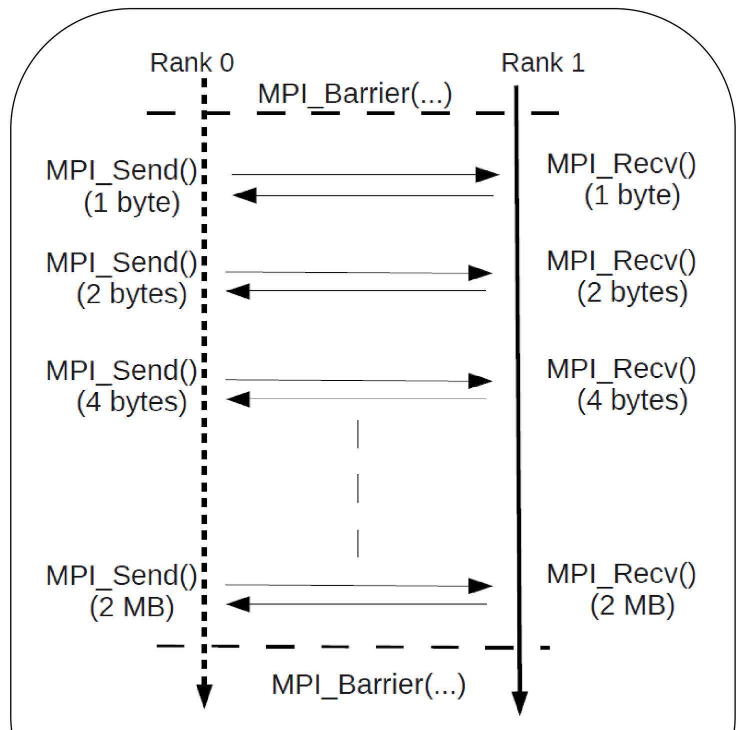
This poster presents empirical evidence that: controllable factors like the size of the data-payload to be transferred and the number of explicit calls used to service such transfers have a direct impact on the power signatures of communication kernels. Moreover, the choice of the transport layer (along with the associated interconnect) and the design of the inter-process communication protocol are responsible for these signatures. Results discussed in this poster motivate the incorporation of energy-based metrics for fine tuning middleware that targets exascale machines.

Experimental Setup

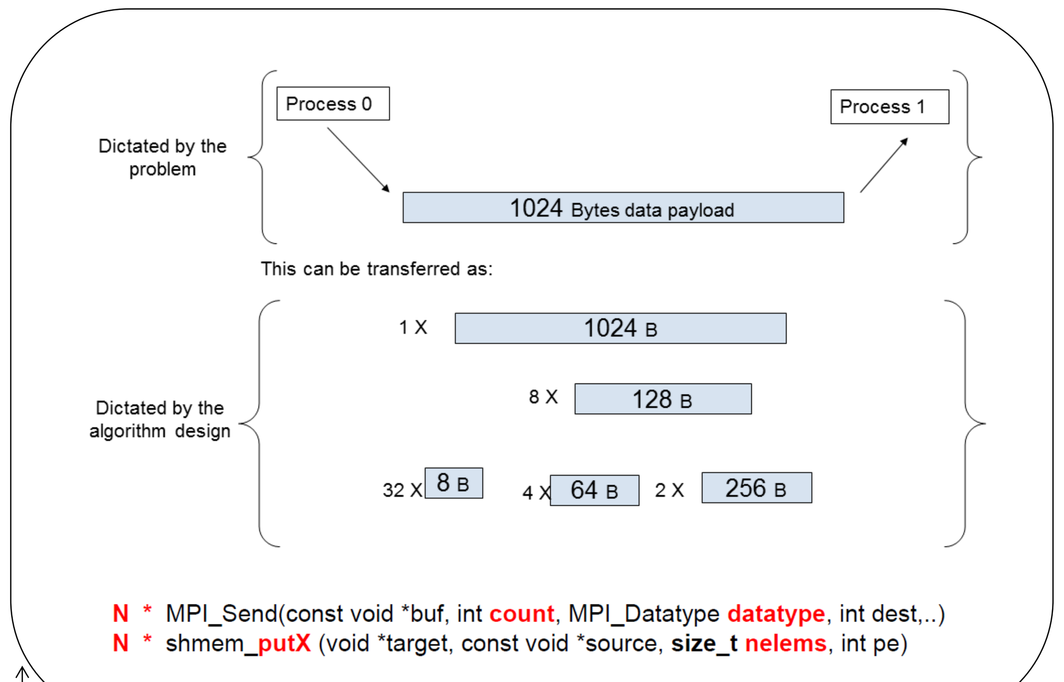
- CPU: Intel Sandy Bridge E5-2690
- Operating Frequency: 2.8GHz
- NIC: Mellanox MT27500 : Connect-X
- Mellanox Scalable SHMEM (ver. 2.2)
- One MPI-rank per process, bound to the socket closest to the NIC



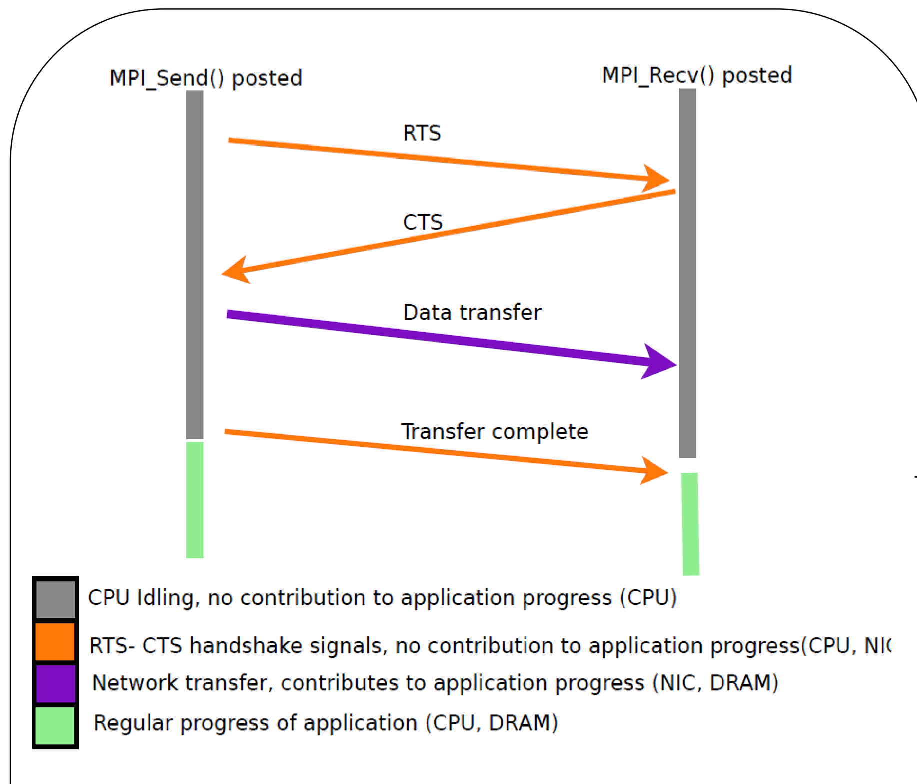
Microbenchmark Design



Characteristics of communication intensive kernels



Rendezvous Protocol



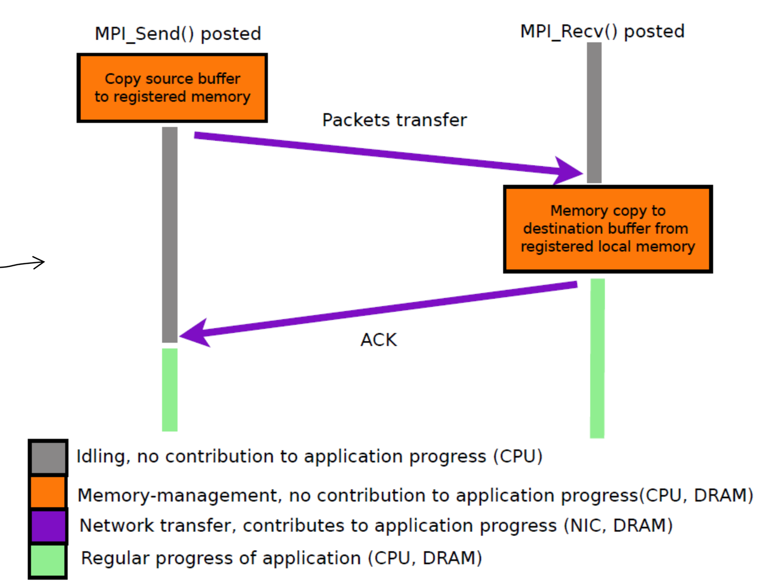
- Additional handshake signals
- No contribution made to application progress by the CPU

Factors impacting energy and power consumption of data transfers

Choice of programming model constructs	
Communication kernel characteristics e.g. total size of the data-payload transferred, the number of calls initiated to service the transfers	
Choice of Transport Layer e.g. TCP, OpenFabrics, shared memory	Communication protocol e.g. Message passing (Eager, Rendezvous) or Direct access
Implementation details Polling, registration of memory, reliability, reusability of memory, caching, memory management, fault-tolerance	Flow/congestion control, routing protocols, deadlock handling, load-balancing, quality-of-service
Cache sizes, set-associativity, cache-coherency protocol memory bandwidth, Hyperthreading, page-replacement	router-switch, organization, network topology, reliability, latency, peak-bandwidth

Scope of this poster

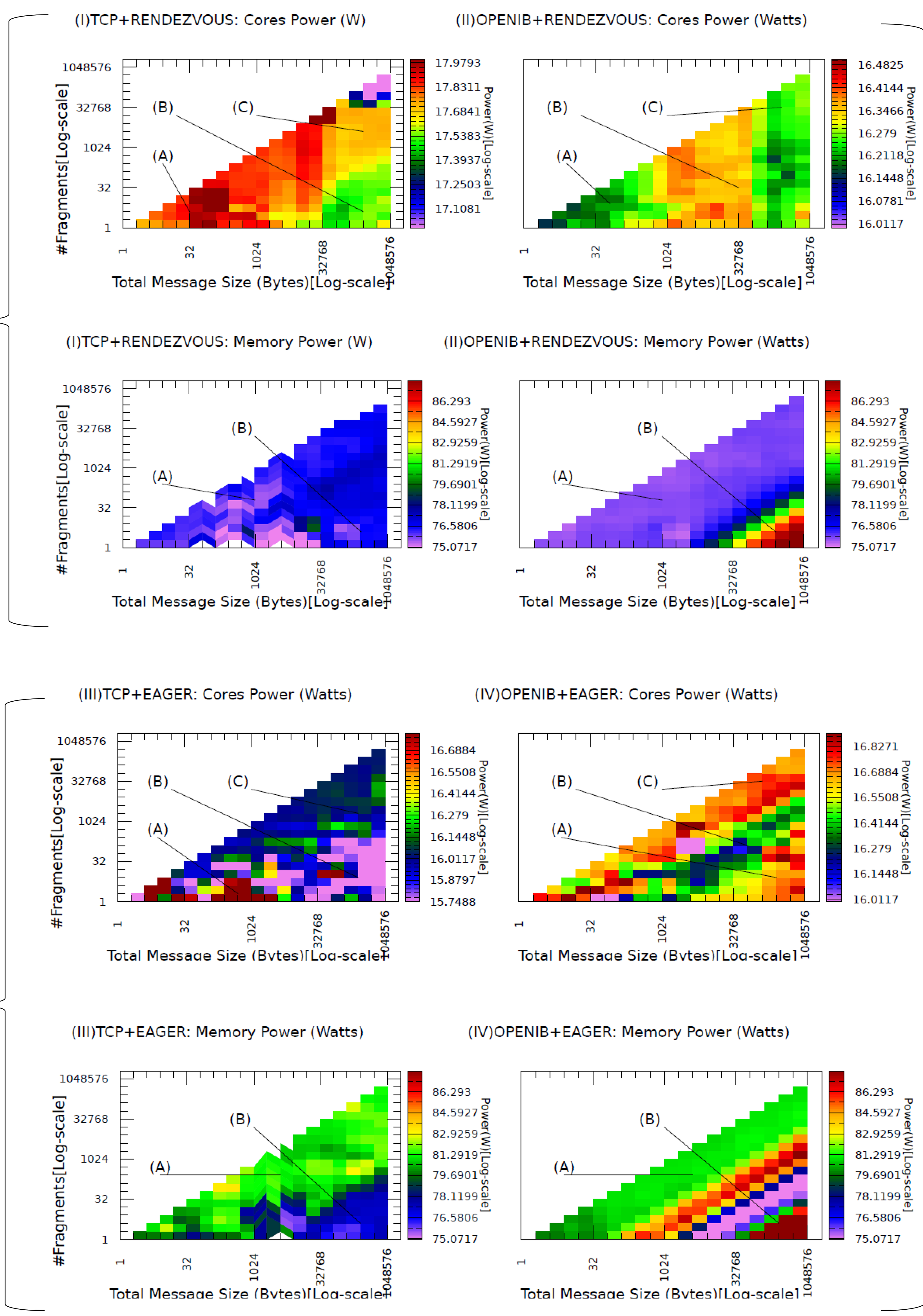
Eager Protocol



- Typically involves registration of memory by the NIC
- OS-kernel bypass
- RDMA capable NICs

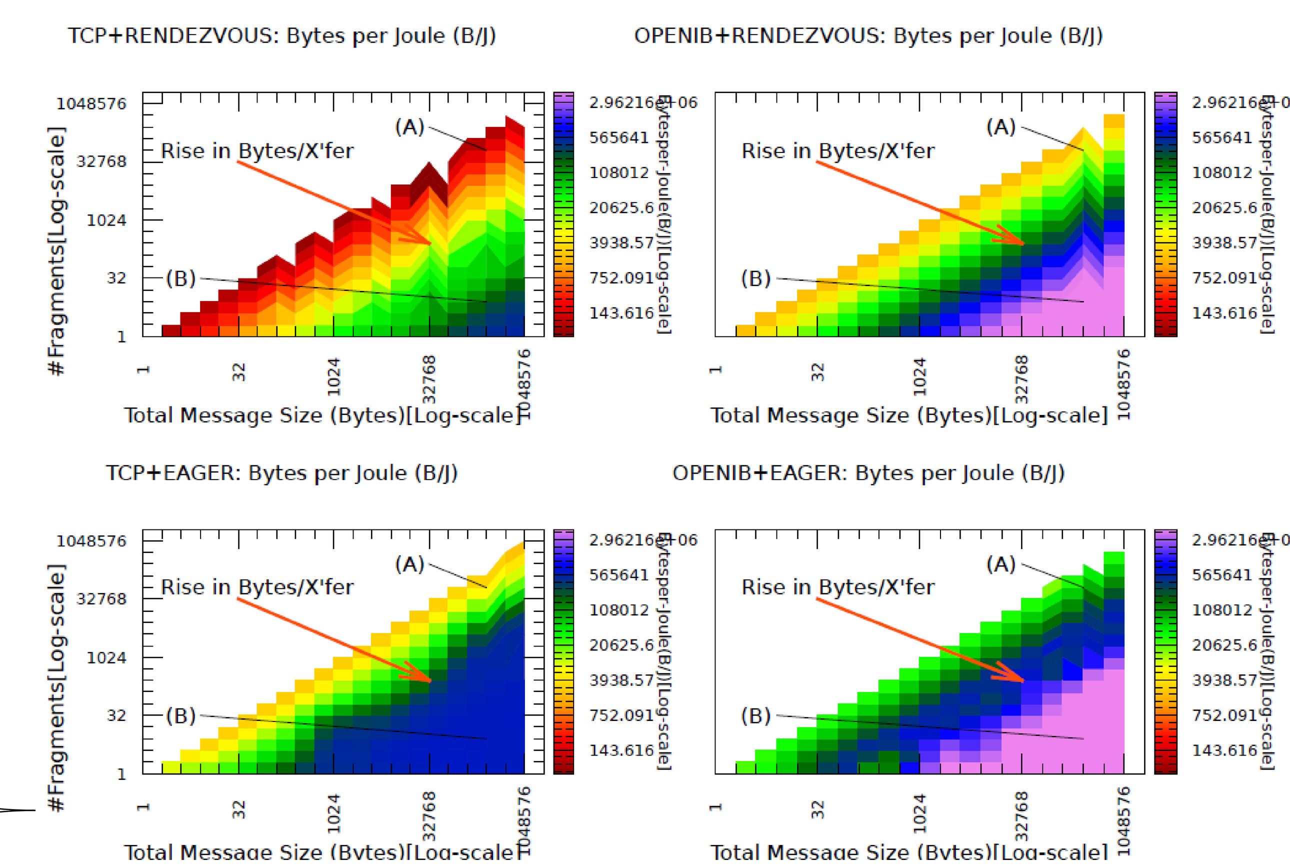
Traditional TCP over Ethernet

OpenFabrics Over InfiniBand



(A): Small data(< 1KB), (B): Medium data (1KB - 64KB), (C): Bulk data (>64KB)

Bytes transferred per Joule of energy consumed



Conclusions:

- Remote data transfers lead to significant impact on the power-signatures of CPU and the memory.
- Small aggregated messages (< 1KB) lead to better energy efficiency than large bulk transfers (> 64K).
- It is important to exploit capabilities of underlying interconnect like OS-kernel bypass, etc.
- Mapping of communication model to RDMA capable NICs is a primary determinant of power-profile of applications.
- Need for design of *energy-aware* middleware.

Publications:

- *"Analysis of Energy and Performance of Code Transformations for PGAS-based Data Access Patterns"*, S.Jana, J.Schuchart, and B.Chapman, PGAS 2014
- *"Power Consumption Due to Data Movement in Distributed Programming Models,"* S.Jana, O.Hernandez, S.Poole, and B.Chapman, Euro-Par 2014 Parallel Processing, 366-378
- *"Analyzing the Energy and Power Consumption of Remote Memory Accesses in the OpenSHMEM Model,"* S.Jana, O.Hernandez, S.Poole, C.Hsu, and B.Chapman, First OpenSHMEM Workshop: Experiences, Implementations and Tools, 2014

- Research supported by the U.S. DOD
- Resources supported by ORNL, TN
- Facility supported by U.S. DOE through Contract No. DE-AC05-00OR22725.