Power-Performance Modeling of Collectives for Exascale Systems with Heterogeneous Architectures

Presentation at MODSIM ‘14

by

Dhabaleswar K. (DK) Panda

The Ohio State University

E-mail: panda@cse.ohio-state.edu

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

Presentation Outline

- Overview of Emerging Systems
- Challenges in Power-Performance Modeling
- Proposed Research Directions
- Answers to Specific Questions

Drivers of Modern HPC Cluster Architectures

Multi-core Processors

High Performance Interconnects - InfiniBand
<1usec latency, >100Gbps Bandwidth

Accelerators / Coprocessors
high compute density, high performance/watt
>1 TFlop DP on a chip

- Multi-core processors are ubiquitous
- RDMA-Enabled Interconnects very popular in HPC clusters
- Accelerators/Coprocessors becoming common in high-end systems
- Pushing the envelope for Exascale computing

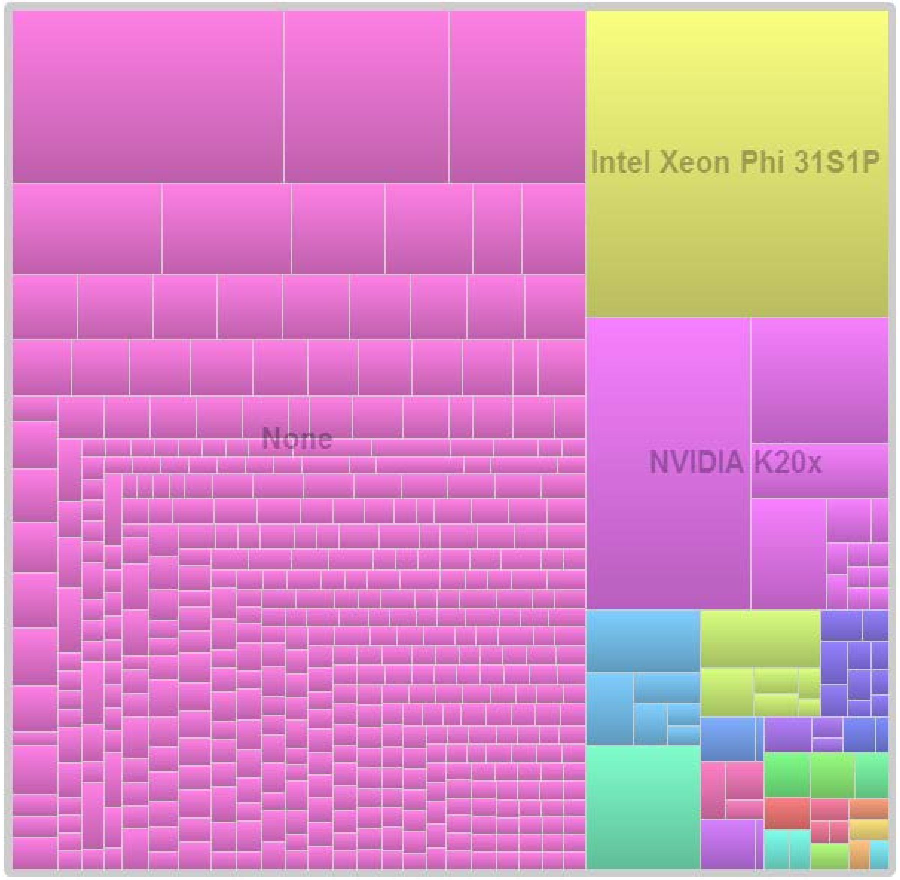
Tianhe – 2 (1)

Titan (2)

Stampede (6)

Tianhe – 1A (10)

Use of Accelerators on TOP500 Systems (June 2014)

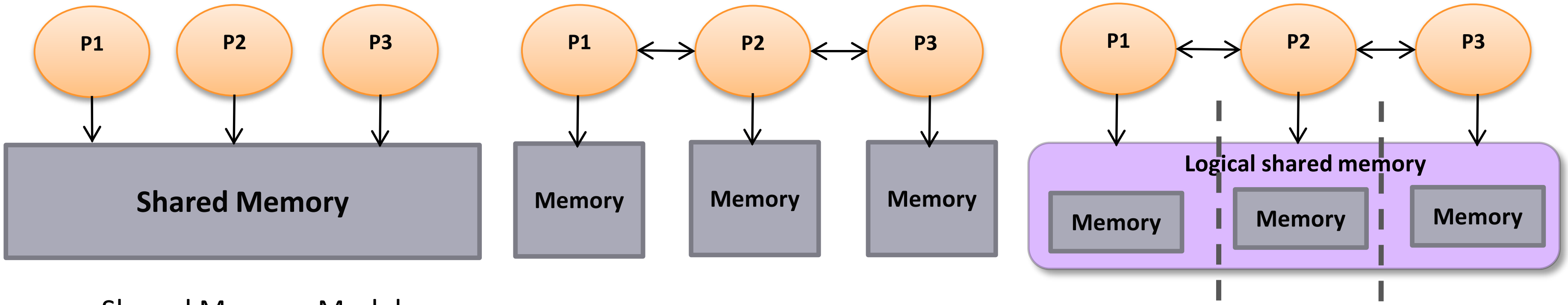


Performance share of accelerators in the Top500 systems

17.5%

17%

Parallel Programming Models Overview



Shared Memory Model

SHMEM, DSM

Distributed Memory Model

MPI (Message Passing Interface)

Partitioned Global Address Space (PGAS)

Global Arrays, UPC, Chapel, X10, CAF, …

- Programming models provide abstract machine models
- Models can be mapped on different types of systems
  - Distributed Shared Memory (DSM), MPI within a node, etc.
  - Logical shared memory across nodes (PGAS)

Major New Features in MPI-3

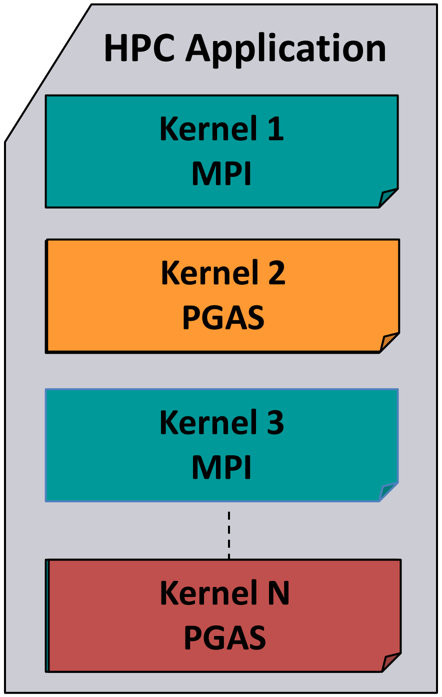
- MPI-3 introduced during Nov ‘12
- Major features
  - Non-blocking Collectives
    - Overlapping computation with collectives
  - Improved One-Sided (RMA) Model
    - Improvements from MPI-2 one-sided models
    - Flexible communication and synchronization schemes
  - MPI Tools Interface
- Specification is available from: http://www.mpi-forum.org/docs/mpi-3.0/mpi30-report.pdf

Partitioned Global Address Space (PGAS) Models

- Key features
  - Simple shared memory abstractions
  - Light weight one-sided communication
  - Easier to express irregular communication
- Different approaches to PGAS
  - Languages
    - Unified Parallel C (UPC)
    - Co-Array Fortran (CAF)
    - X10
  - Libraries
    - OpenSHMEM
    - Global Arrays
    - Chapel

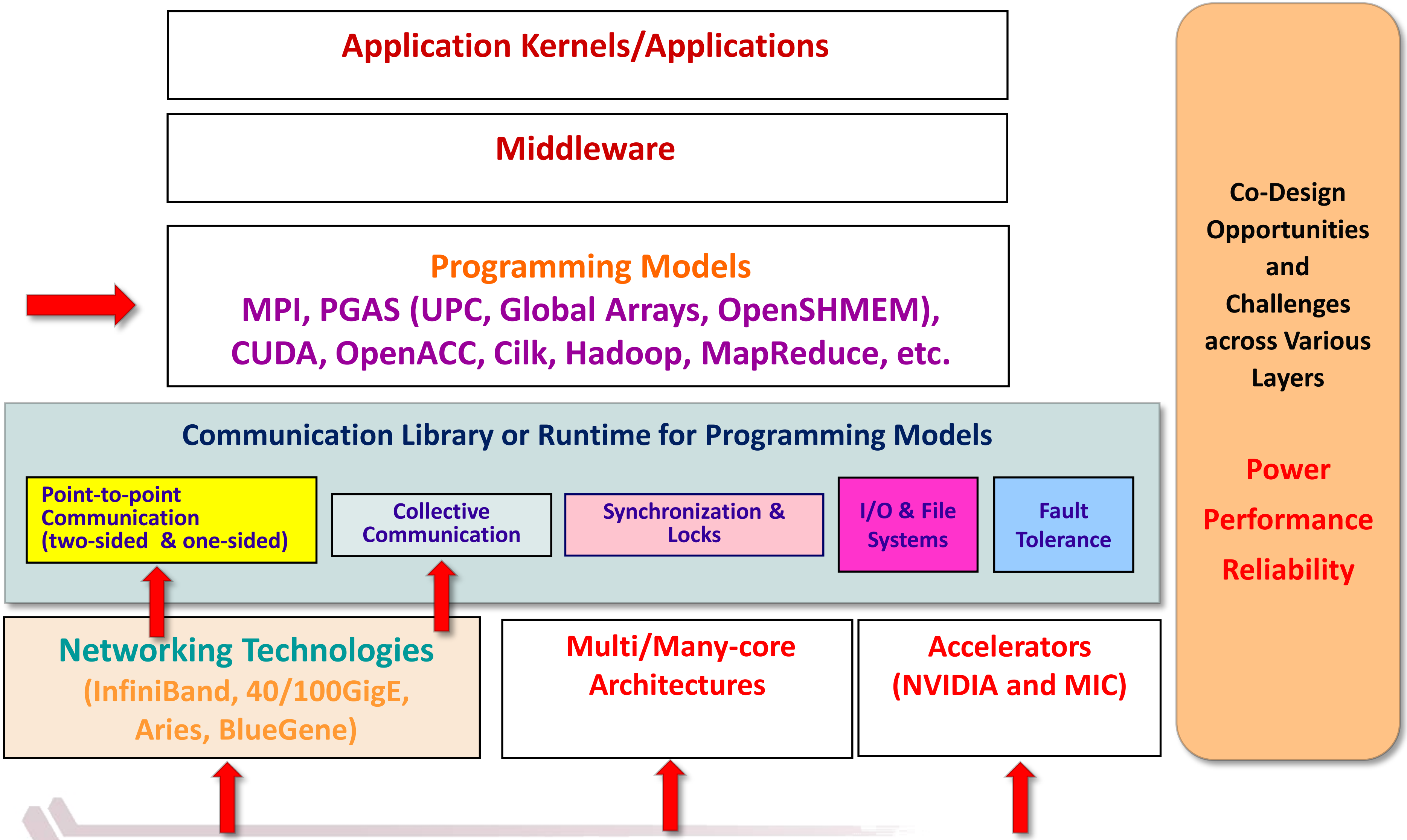
Hybrid (MPI+PGAS) Programming

- Application sub-kernels can be re-written in MPI/PGAS based on communication characteristics
- Benefits:
  - Best of Distributed Computing Model
  - Best of Shared Memory Computing Model
- Exascale Roadmap*:
  - “Hybrid Programming is a practical way to program exascale systems”



** The International Exascale Software Roadmap, Dongarra, J., Beckman, P. et al., Volume 25, Number 1, 2011, International Journal of High Performance Computer Applications, ISSN 1094-3420*

Designing Software Libraries for Multi-Petaflop and Exaflop Systems with MPI+X: Challenges



Application Kernels/Applications

Middleware

Programming Models
MPI, PGAS (UPC, Global Arrays, OpenSHMEM), CUDA, OpenACC, Cilk, Hadoop, MapReduce, etc.

Communication Library or Runtime for Programming Models

Point-to-point Communication (two-sided & one-sided)

Collective Communication

Synchronization & Locks

I/O & File Systems

Fault Tolerance

Networking Technologies (InfiniBand, 40/100GigE, Aries, BlueGene)

Multi/Many-core Architectures

Accelerators (NVIDIA and MIC)

Co-Design Opportunities and Challenges across Various Layers

Power

Performance

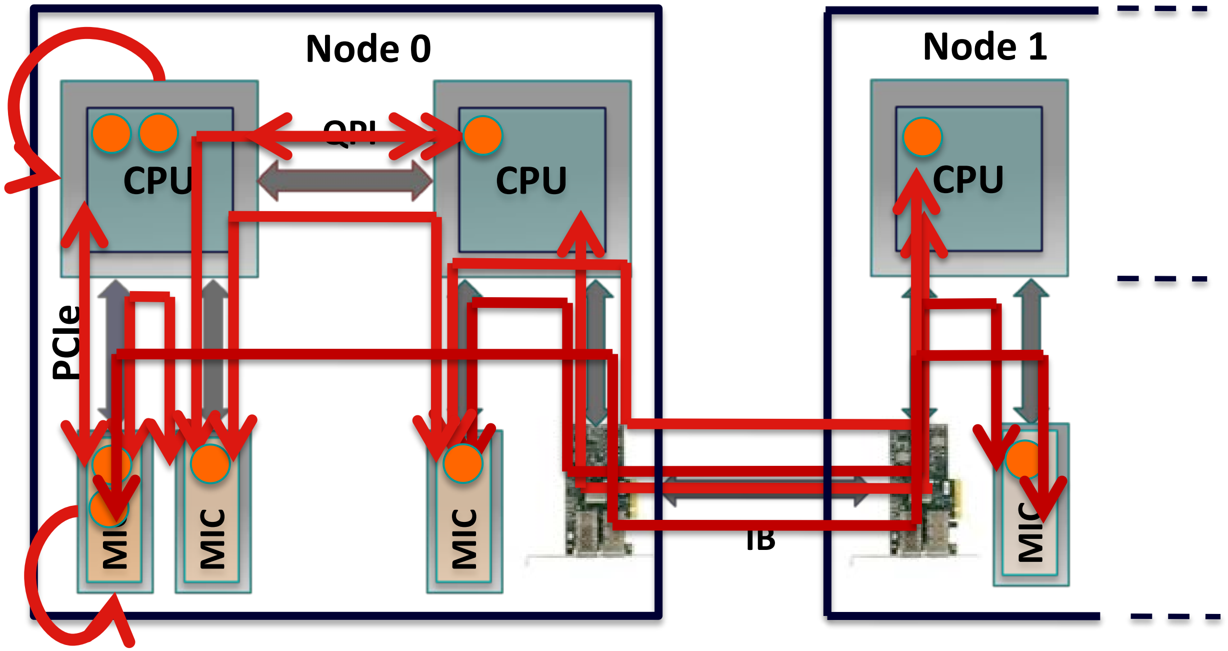
Reliability

Presentation Outline

- Overview of Emerging Systems
- Challenges in Power-Performance Modeling
- Proposed Research Directions
- Answers to Specific Questions

Data Movement Paths on Intel Xeon Phi Clusters

- Connected as PCIe devices – Flexibility but Complexity



● MPI Process

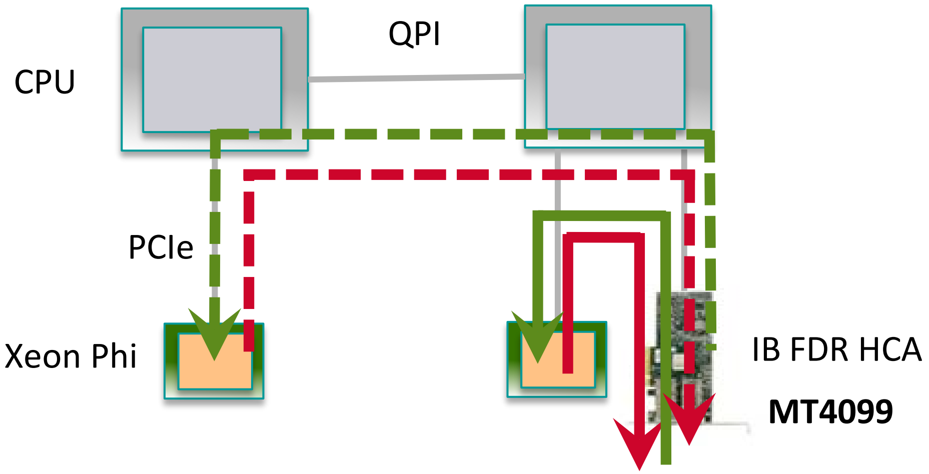
1. Intra-Socket
2. Inter-Socket
3. Inter-Node
4. Intra-MIC
5. Intra-Socket MIC-MIC
6. Inter-Socket MIC-MIC
7. Inter-Node MIC-MIC
8. Intra-Socket MIC-Host
9. Inter-Socket MIC-Host
10. Inter-Node MIC-Host
11. Inter-Node MIC-MIC with IB adapter on remote socket

and more . . .

- Critical for runtimes to optimize data movement, hiding the complexity
- Significant trade-off in Power and Performance for each of the data path

Communication Performance on MIC Clusters

- Different paths have different performance
- Asymmetric communication performance



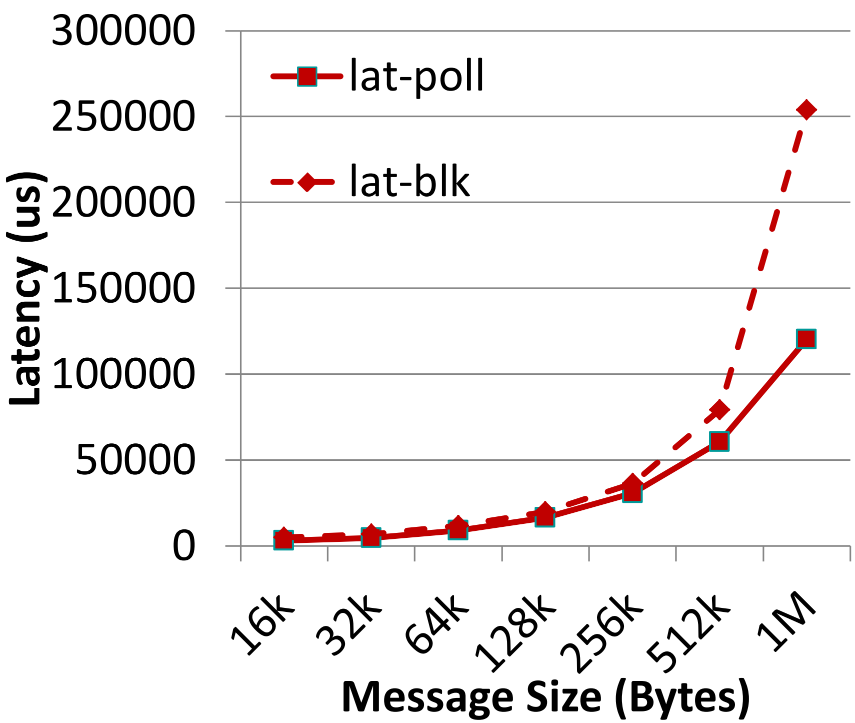
Peak IB FDR Bandwidth: 6397 MB/s

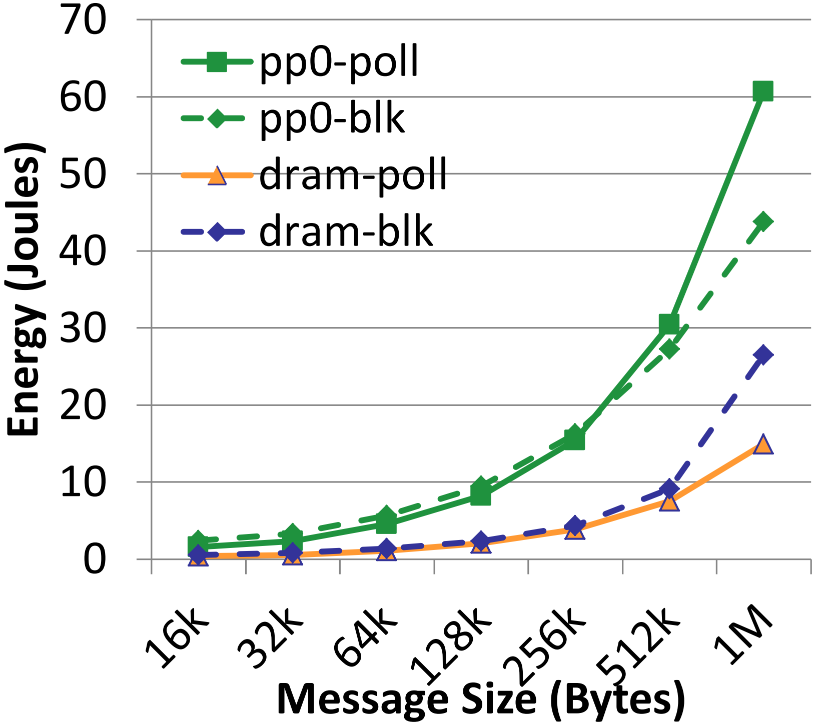
			E5-2670 (SandyBridge)	**E5-2680 v2 (IvyBridge)**
Intra-socket	→	IB Read from Xeon Phi (P2P Read)	962 MB/s (15%)	3421 MB/s (54%)
	→	IB Write to Xeon Phi (P2P Write)	5280 MB/s (83%)	6396 MB/s (100%)
Inter-socket	⇢	IB Read from Xeon Phi (P2P Read)	370 MB/s (6%)	247 MB/s (4%)
	⇢	IB Write to Xeon Phi (P2P Write)	1075 MB/s (17%)	1179 MB/s (19%)

MVAPICH2/MVAPICH2-X Software

- High Performance open-source MPI Library for InfiniBand, 10Gig/iWARP, and RDMA over Converged Enhanced Ethernet (RoCE)
  - MVAPICH (MPI-1), MVAPICH2 (MPI-2.2 and MPI-3.0), Available since 2002
  - MVAPICH2-X (MPI + PGAS), Available since 2012
  - Support for GPGPUs and MIC
  - **Used by more than 2,175 organizations in 72 countries**
  - **More than 220,000 downloads from OSU site directly**
  - Empowering many TOP500 clusters
    - 7th ranked 519,640-core cluster (Stampede) at TACC
    - 13th ranked 74,358-core cluster (Tsubame 2.5) at Tokyo Institute of Technology
    - 23rd ranked 96,192-core cluster (Pleiades) at NASA
  - Available with software stacks of many IB, HSE, and server vendors including Linux Distros (RedHat and SuSE)
  - http://mvapich.cse.ohio-state.edu
- Partner in the U.S. NSF-TACC Stampede System

Energy-Performance Tradeoff of Collectives: 64-process Alltoall (Polling Vs Blocking)

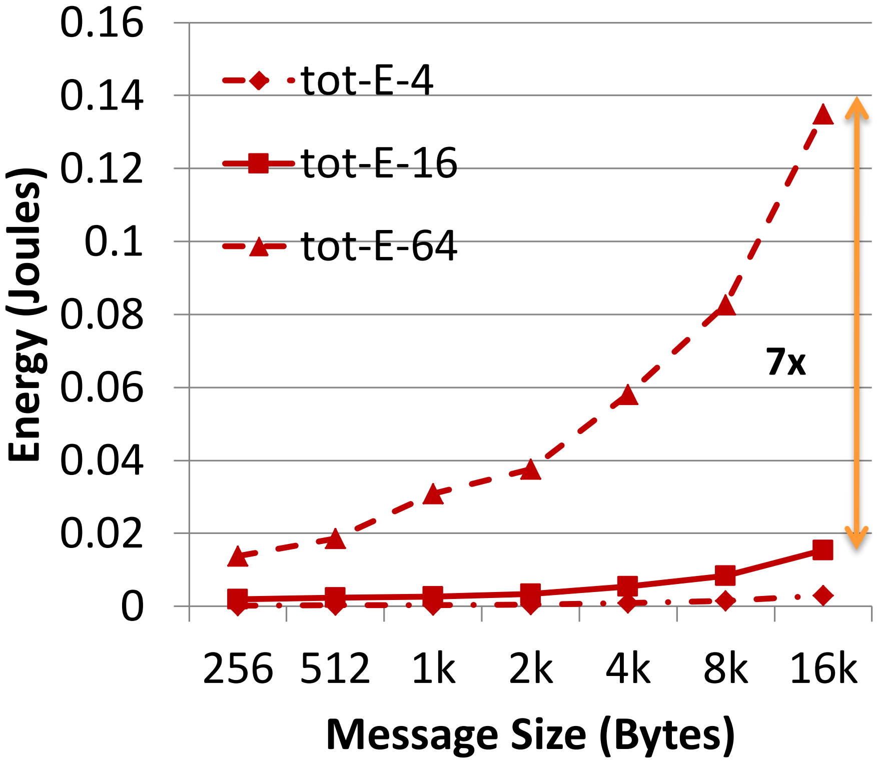


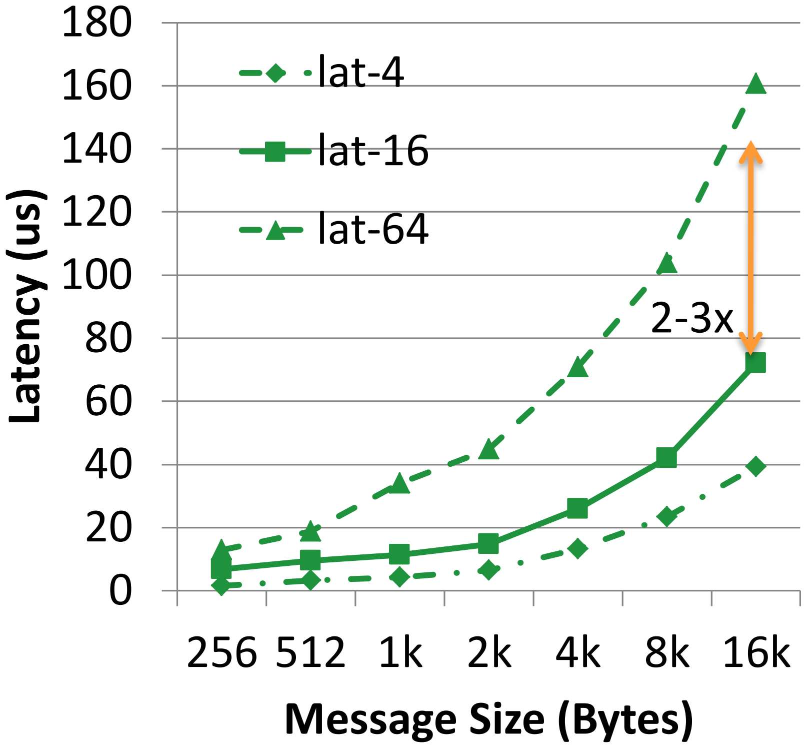


- Blocking mode involves suspension => lesser processor energy
- Will involve greater cache misses => greater DRAM energy usage
- Latency is higher with blocking

A. Venkatesh, K. Kandalla and D. K. Panda, Evaluation of Energy Characteristics of MPI Communication Primitives with RAPL, Int'l Workshop on High Performance, Power-Aware Computing (HPPAC), held in conjunction with Int'l Parallel and Distributed Processing Symposium (IPDPS '13), May 2013.

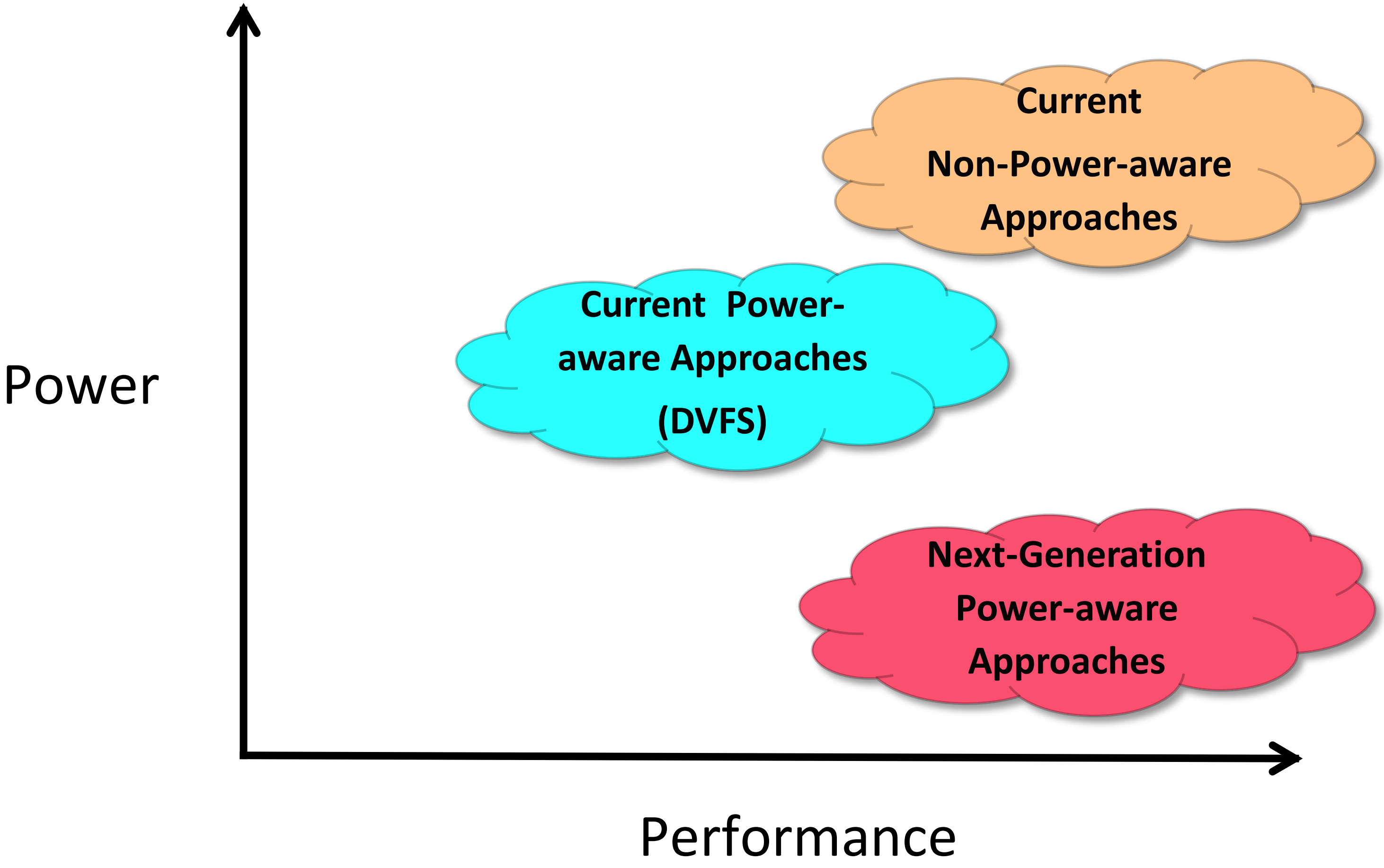
Latency Vs Energy Scaling: Collectives: Allreduce





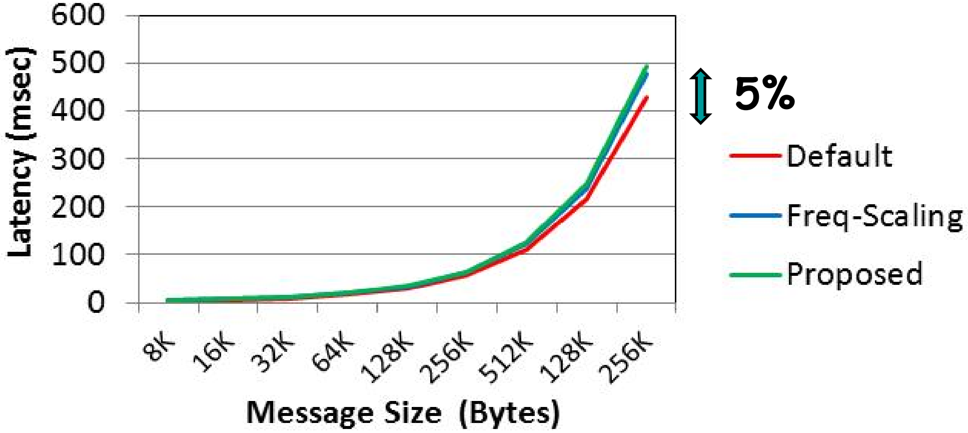
- Tot-E-#N is energy spent by the collective across nodes
- 2x-3x growth in latency from 16 -> 64
- 7x-8x growth in Energy!
- **Can we model this power-performance tradeoff?**

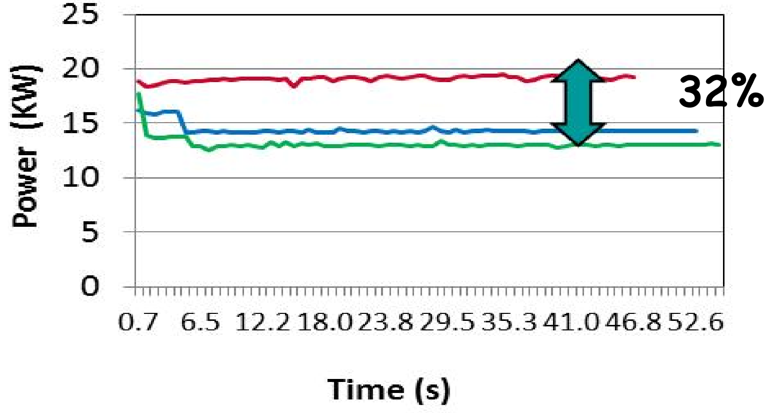
Designing Collectives with Power-Performance Tradeoff



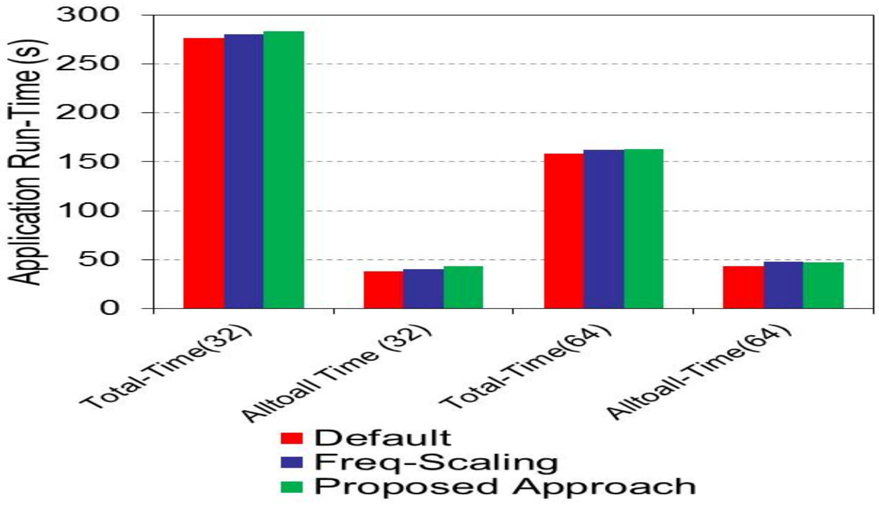
Designing Power-Aware Collectives

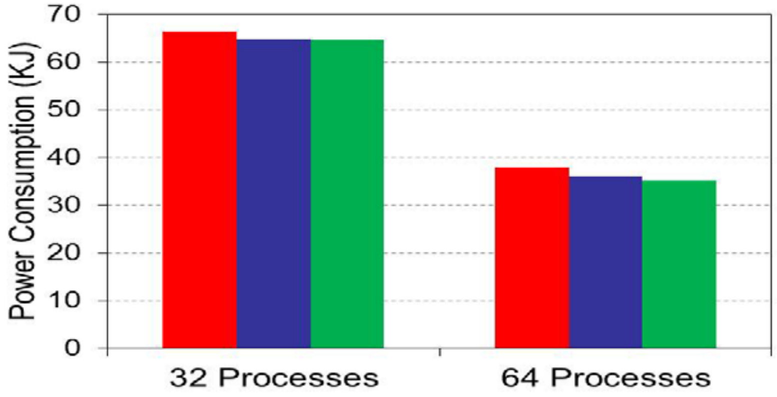
Performance and Power Comparison : MPI_Alltoall with 64 processes on 8 nodes





CPMD Application Performance and Power Savings





K. Kandalla, E. P. Mancini, S. Sur and D. K. Panda, "Designing Power Aware Collective Communication Algorithms for Infiniband Clusters", ICPP '10

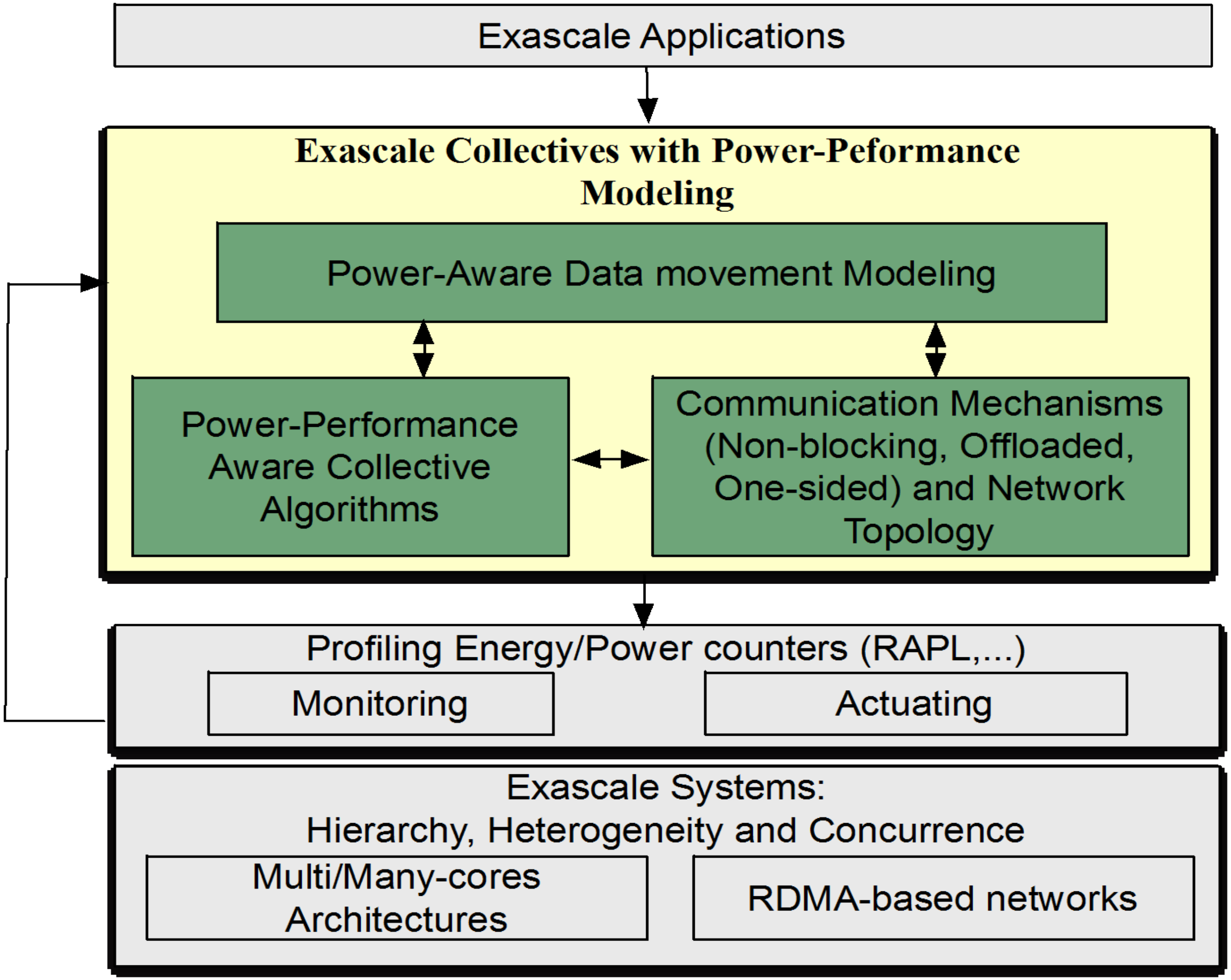
Summary of Current Limitations

- Current communication models (LogP, LogGP, PLogP, Hockney) do not take into account
  - Energy
  - Diverse communication paths
  - Communication paths involving heterogeneity
- Power-performance tradeoffs for new generation communication mechanisms are not understood
  - MPI-3 RMA operations
  - Non-blocking collectives
  - Adapters with network offload

Presentation Outline

- Overview of Emerging Systems
- Challenges in Power-Performance Modeling
- Proposed Research Directions
- Answers to Specific Questions

Framework for Designing Exascale Collectives with Power-Performance Modeling



Exascale Applications

Exascale Collectives with Power-Peformance Modeling

Power-Aware Data movement Modeling

Power-Performance Aware Collective Algorithms

Communication Mechanisms (Non-blocking, Offloaded, One-sided) and Network Topology

Profiling Energy/Power counters (RAPL,...)

Monitoring

Actuating

Exascale Systems:
Hierarchy, Heterogeneity and Concurrence

Multi/Many-cores Architectures

RDMA-based networks

New Point-to-point Communication Models for Emerging Architectures and Programming Models

- Need to capture both power and performance
- Heterogeneity of processors
  - Powerful cores (high clock frequency, out-of-order processing and multiple functional units)
  - Simplified cores (low clock frequency and in-order processing)
- Capture behavior of different communication paths
  - Over QPI or HT
  - Data movement in/out from accelerators
- Takes into account of
  - Concurrency in communication
  - Overlap of computation and communication

New Point-to-point Communication Models (Cont'd)

- Communication and synchronization primitives for emerging programming models
  - One-sided RMA model in MPI-3
  - PGAS models
  - Communication under MPI+X model
- Takes into account of advances in networking
  - Blocking/Non-blocking
  - Onloading/Offloading

(Power + Performance)-Aware Models for Collectives and Designing New Collective Algorithms

- Uses point-to-point communication models
- Explores advanced communication mechanisms and schemes
  - MPI-3 RMA communication
  - Network interfaces with collective offload
  - Hierarchical architecture
  - Network topology
- Considers factors related to both power and performance
  - Power-state dependent performance levels
  - Time overheads of power-throttling
  - Effect of differential power-states of different processes involved in a collective

Presentation Outline

- Overview of Emerging Systems
- Challenges in Power-Performance Modeling
- Proposed Research Directions
- Answers to Specific Questions

Answers to Specific Questions

- *What is the major contribution of your research?*
  - Modeling the power and performance characteristics of collectives by taking into account
    - Heterogeneity of architectures
    - Emerging MPI+X programming models
- *What are the gaps you identify in the research coverage?*
  - Addressing the following limitations of current models
    - Not power/energy-cognizant
    - Do not address new architectures and programming models
- *What is the bigger picture for your research area?*
  - Understanding the implication of communication from both performance/power angles at scale

Answers to Specific Questions (Cont'd)

- *What major opportunities do you see for cross-pollination between your projects?*
  - Modeling resource utilization and effects of contention on highly parallel architectures
  - Modeling effects of software framework and programming paradigm on power and performance
  - New low-overhead techniques to leverage power and performance throttling
- *What would make it easier to use the results of other projects to further your own research?*
  - Better tools to identify the vulnerable design spots in communication models
- *What would you like to most see addressed other than what they are working on?*
  - Negligible overhead in controlling power-state of cores on nodes
  - Accurate tools and techniques to predict dynamic network congestion

Pointers

Network Based Computing Laboratory

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

MVAPICH

http://mvapich.cse.ohio-state.edu

panda@cse.ohio-state.edu

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