SAND2024-02732C

Sandia National Laboratories

Exceptional service in the national interest

PARTITIONED COMMUNICATION

And the future of application design

Matthew G. F. Dosanjh

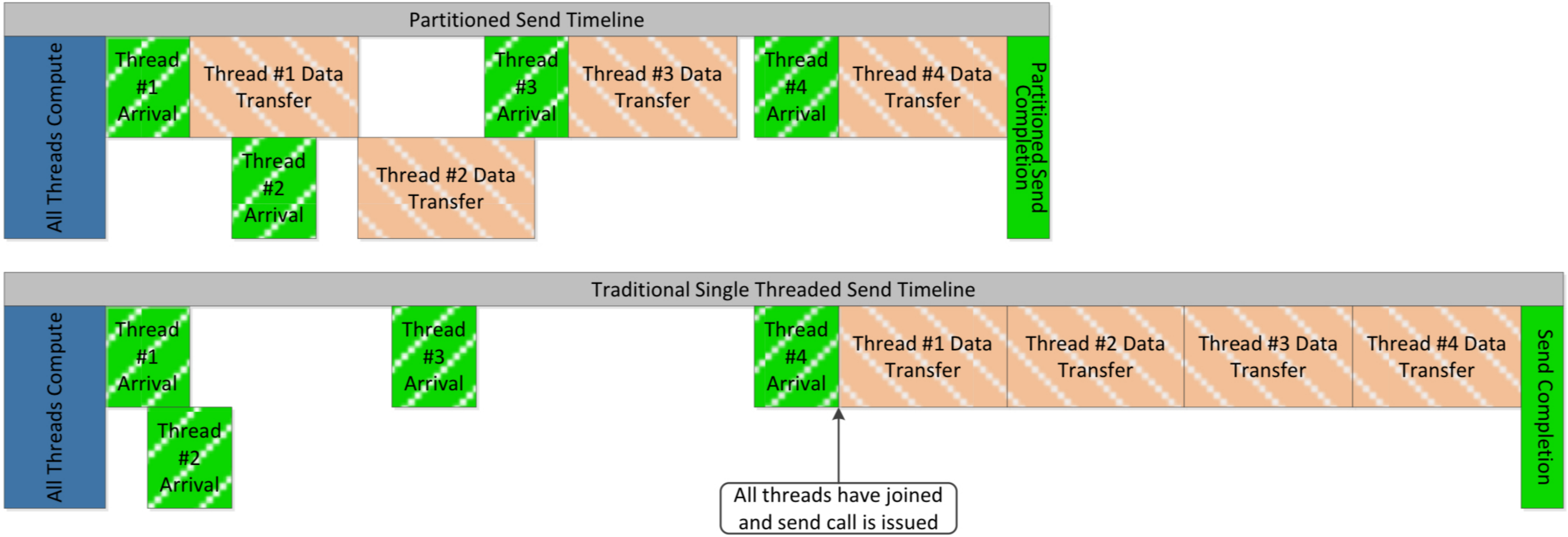
U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

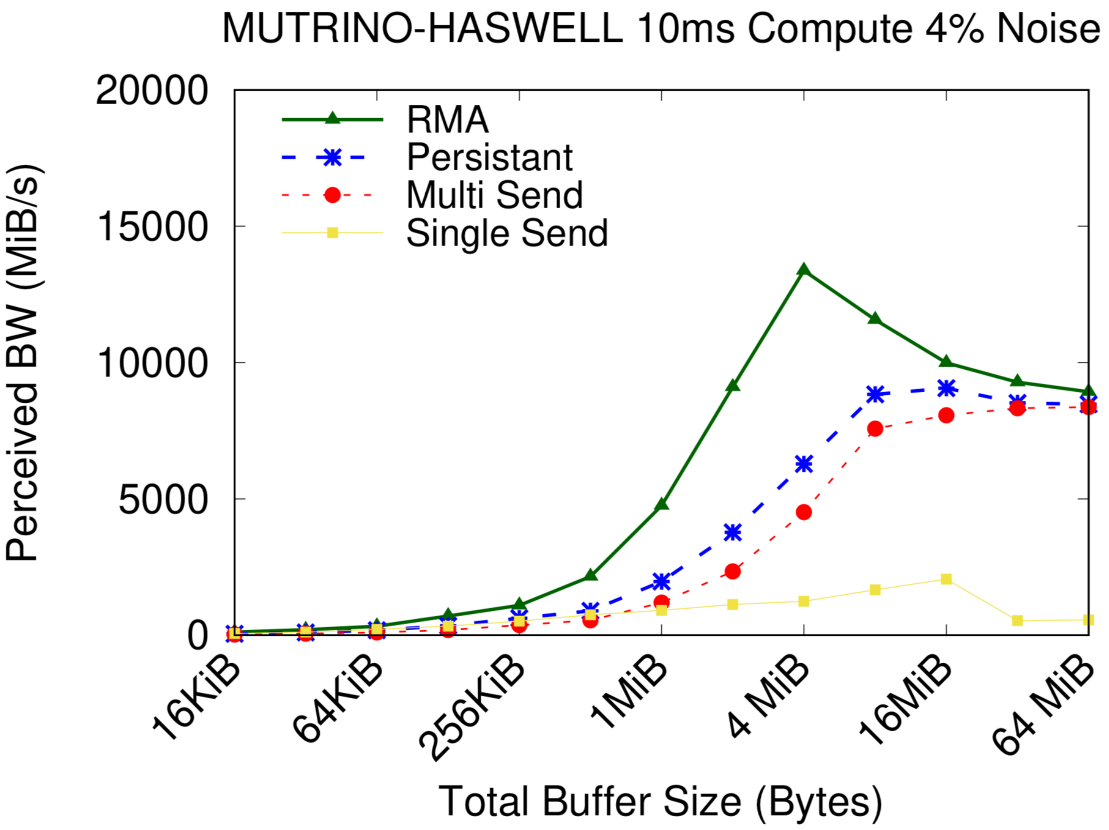
THE NEED FOR FINE GRAINED COMMUNICATION

- Overlapping through earlybird communication
- Messages are partitioned into smaller “sub messages”
- The underlying mechanism can start sending data as it’s ready



PARTITIONED COMMUNICATION PERFORMANCE

- Partitioned communication allows for greater communication performance by allowing early bird communication while minimizing the overhead associated with sending more messages.
- In the graph on the right, we show the bandwidth performance of three different underlying communication methods (Multi send, MPI Persistent, and MPI RMA).
- With reduced overheads perceived bandwidth can be greatly improved.

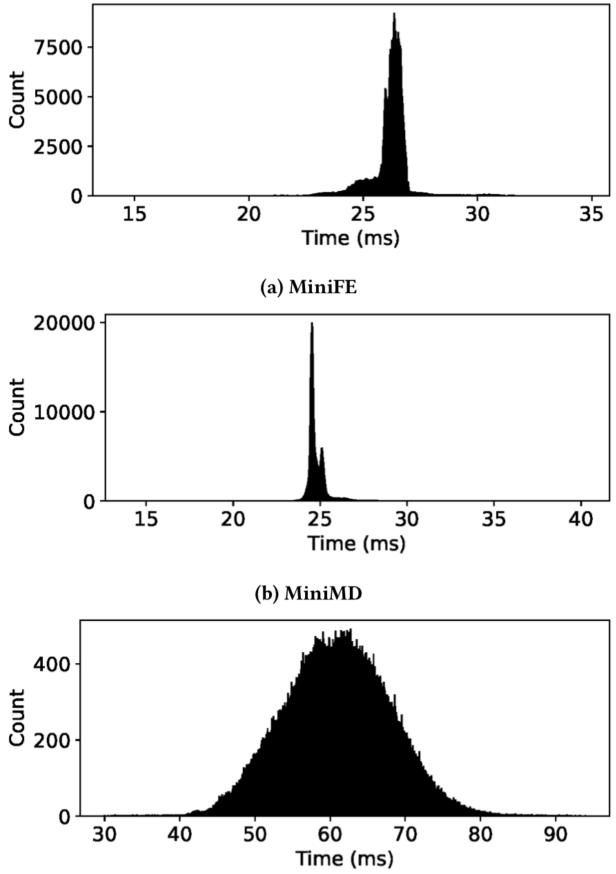


BUT THERE IS A CATCH!

- The previous experiment made a number of assumptions
    - 10ms compute time
    - A single thread delayed by 4%
    - A wide range of message sizes
- But what do applications actually do?
    - It’s easy to measure certain things
        - Total communication volume
        - Total computation time
    - Others are harder
        - Mean and distribution of thread arrival times.

EXPLORATION OF THREAD TIMINGS

- We’ve measure the thread arrival times of three different proxy applications.
- These proxy apps are representative of different code classes
  - Finite Elements
  - Molecular Dynamics
  - Monte Carlo
- Each has unique behaviors
- Application behavior can change over time
  - MiniMD sees different behavior for roughly the first 10 iterations



(a) MiniFE

(b) MiniMD

(c) MiniQMC

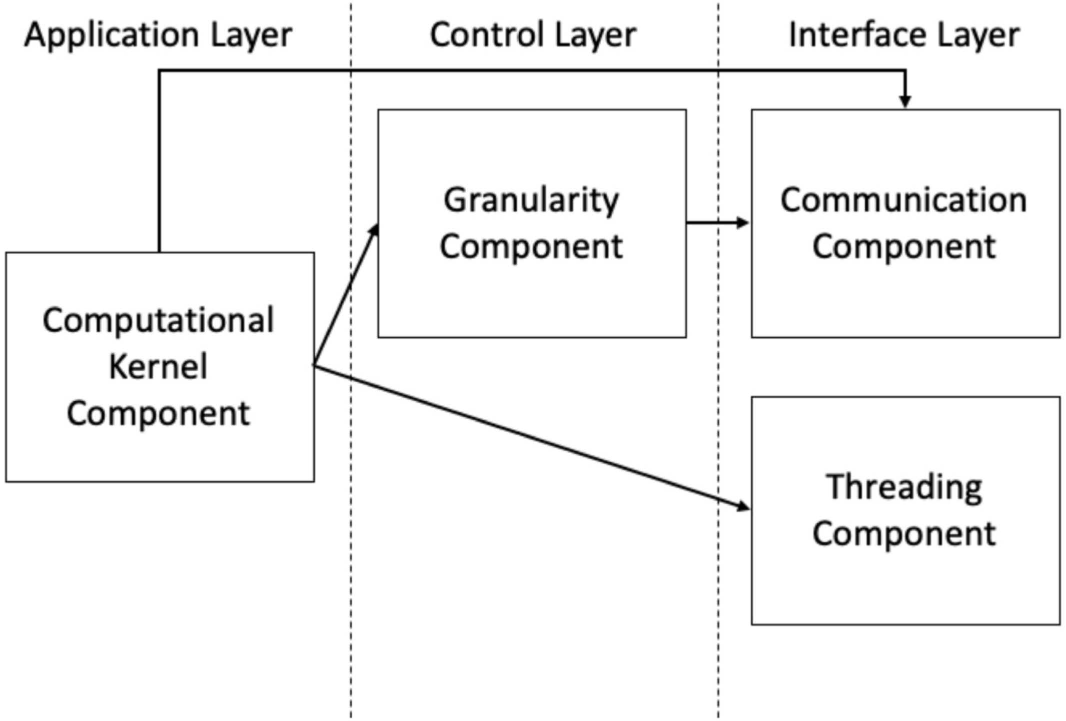
Figure 3: Application thread arrival time histograms for each of our three applications. Each has a bin width of 10 microseconds.

EVALUATING DIFFERENT APPLICATION DESIGNS

- As shown in the previous slide applications can behave very differently
- This is also dependent on other factors
  - Number of threads per process
  - Problem size
  - Network hardware
  - Network middleware
- Manually testing different configurations is time consuming
- We need a way to search the application design space

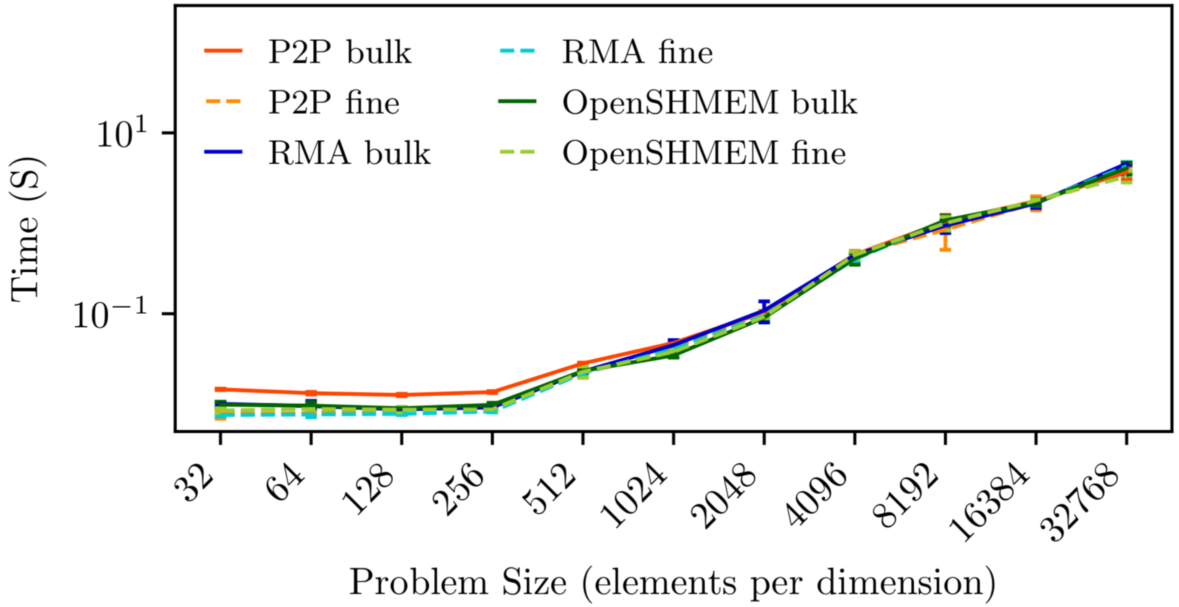
A MODULAR FRAMEWORK FOR APPLICATION DESIGN

- MiniMOD is a framework to allow for exploring different application design choices at runtime
- This includes different kernels, communication behaviors, and underlying communication libraries.
- By modularizing all of these parts, they can be combined at runtime to emulate different application designs in a fair testing enviroment



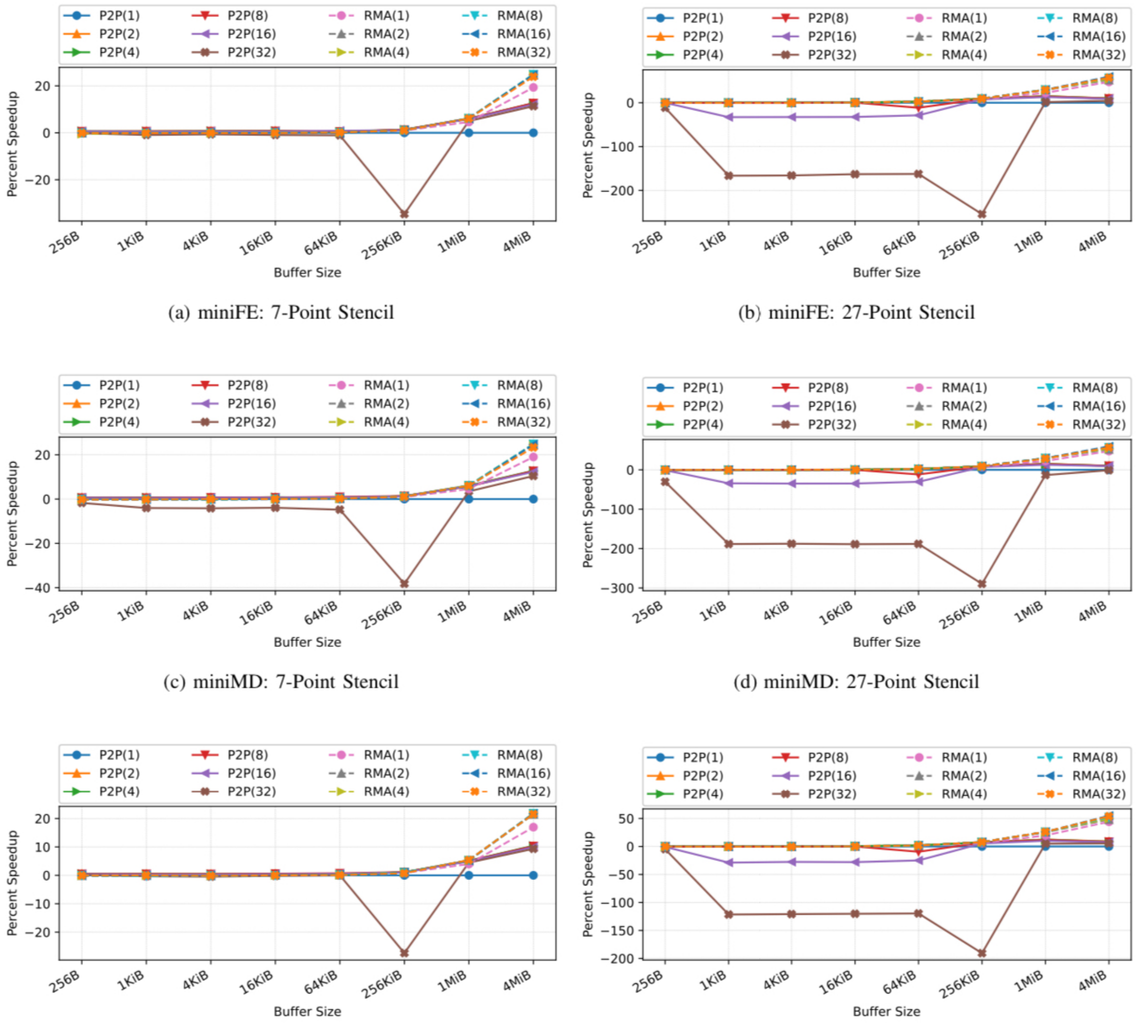
MINIMOD HEAT DIFFUSION

- Heat Diffusion shows is a 2d halo exchange code that shows interacts with 4 neighbors.
- MiniMOD allows us to directly compare different communication library choices
  - MPI Point to Point
  - MPI Remote Memory Access
  - OpenSHMEM
- It also allows us to change the application behavior
  - Traditional Bulk Synchronous
  - Fine Grained Communication
- The results of this change based on problem size



CONFIGURABLE APPLICATION PROXY

- However we may want a more general application model
- This benchmark takes in a threads timing distribution
  - Laggard thread model
  - Normal distribution
  - KDE based on the thread timings data
- It also allows us to specify how many peers each process is communicating with and how much each message is comprised of



(a) miniFE: 7-Point Stencil

(b) miniFE: 27-Point Stencil

(c) miniMD: 7-Point Stencil

(d) miniMD: 27-Point Stencil

(e) miniQMC: 7-Point Stencil

(f) miniQMC: 27-Point Stencil

CONCLUSIONS

- Partitioned communication opens the door for new optimizations
- That opens a large optimization space for application design
- We’ve created tools to help explore this space before committing to implement these designs

QUESTIONS?