SAND2007-2825P

# Advanced Interconnect Architectures

**May 3, 2007**

**Keith D. Underwood**
**Principal Member of Technical Staff**

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

# Interconnect Design Philosophy

- Overlap through offload and independent progress
  - The processors should be processing
  - Achieving scalability is the primary goal – want to be able to hide all of the communication time possible
  - Overlap is optimized when independent progress and offload are available
- One API for networking
  - MPI is the core of making the machine run fast, but…
  - Everything else should not be a second class citizen
- Evolution where possible, but always push the envelope for bandwidth, latency, and message throughput

# Striving for Balance

- MPI latency
  - A persistent challenge – finally approaching speed of light limitations
  - Currently requires at least 3 memory accesses (likely more)
  - Target: 500 ns
- MPI bandwidth
  - 0.5 bytes per real FLOP
  - Just add money…
  - Target: 50 GB/s
- MPI message throughput
  - Critical to increase this metric to at least keep pace with bandwidth
  - Should be sustainable in “real” scenarios
  - Target: 20 Mmsgs/s/direction

# A Contrarian View

- Offload, not Onload
  - Current trends toward onload because of core count
  - Lots of “interesting benchmarks” to say that is a GOOD thing
- Motivations for offload
  - Increased performance
    - Can match message rate or better
    - Eliminate a lot of work for the processor
  - Reduced power per operation
  - Silicon is cheap

# A Few Oddities

- Intra-Core messaging: why not use the NIC?
  - The NIC is on the processor bus
  - The system should have roughly as much NIC bandwidth as memory bandwidth
- You need how many links?
  - Disparity between HyperTransport bandwidth and network/memory bandwidth is concerning
  - Multiple processor links needed to drive a single NIC
  - Don’t expect PCI-Express to ever give the best performance

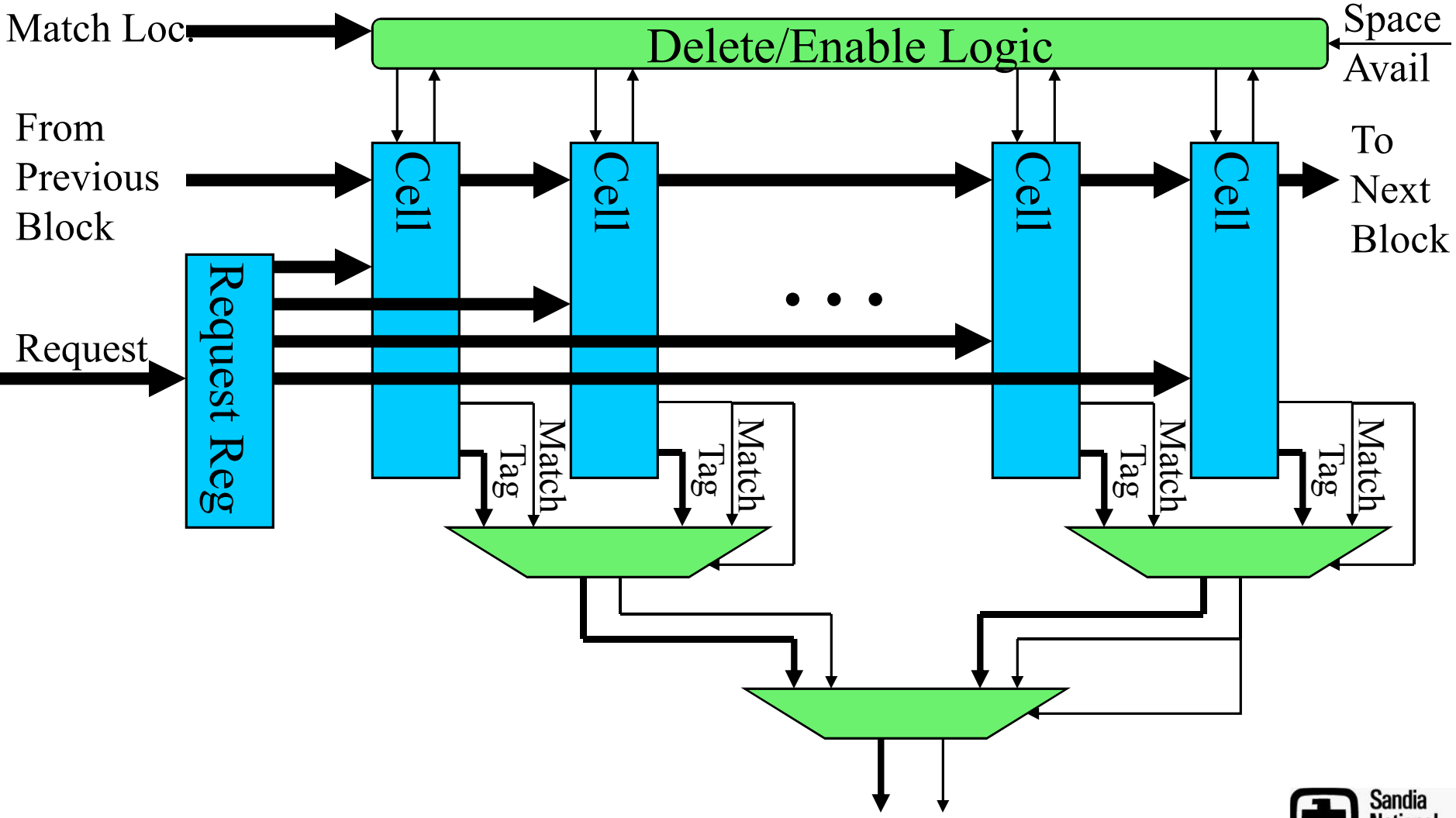
# Brief MPI Implementation Tutorial

- To enable overlap between computation and communication, non-blocking receives are used
- Receives must be matched in the order they were posted
  - Typical implementations maintain “posted receives” in a linked list
  - MPI ordering and “wildcard” semantics make other data structures challenging
- Typical benchmarks lie about message rate
  - Data access patterns allow prefetchers to work for linked list
  - Data access patterns are cache friendly

# Hardware to Address Specific Challenges

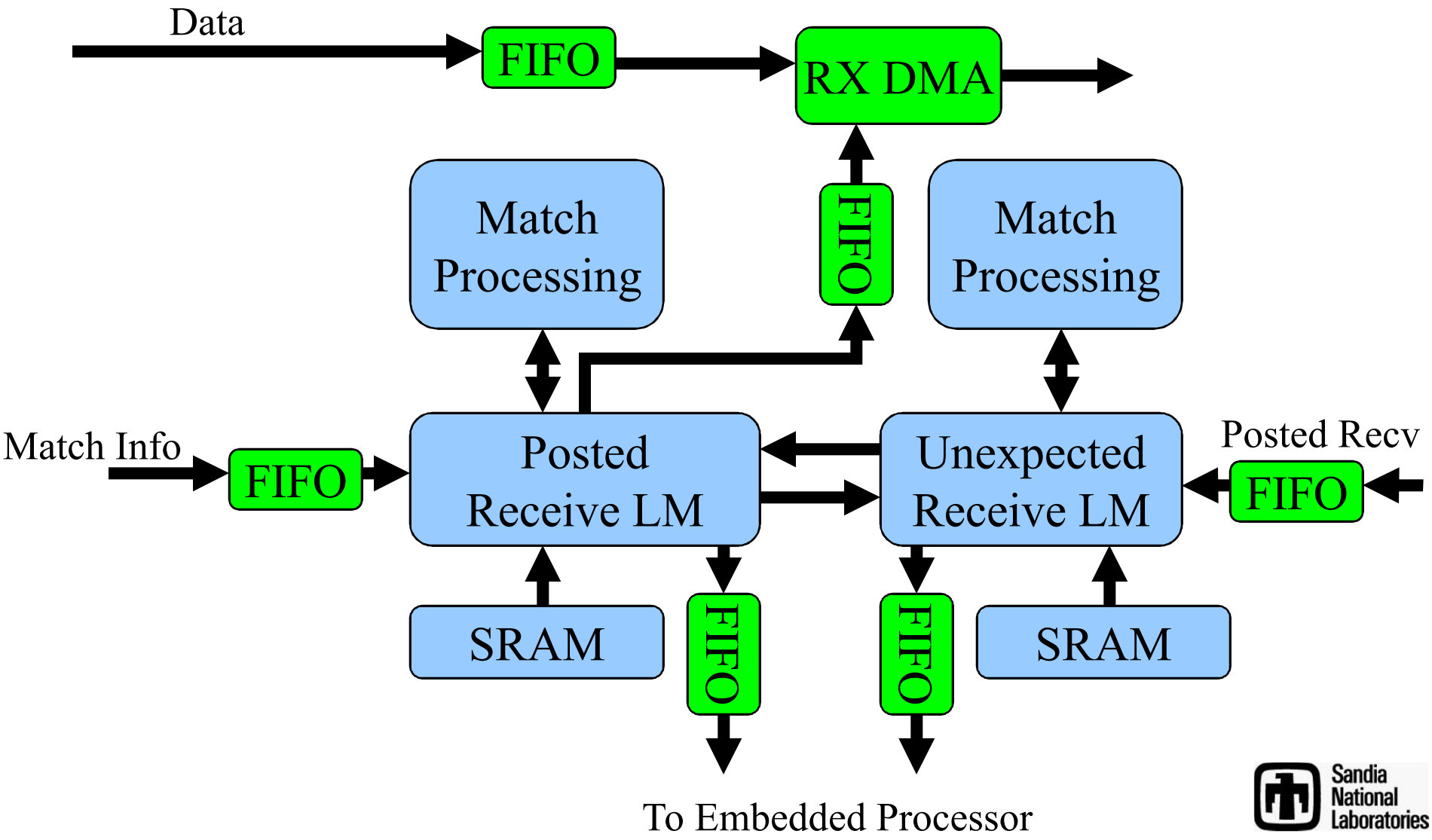
- ALPU for associative matching
  - "Quick look" at head of list
  - Traverse a short list in 7 cycles
- List management unit because an embedded processor is bad at managing lists
  - Hardware unit can overlap fetching of list items with processing of list items
  - Hardware is tightly coupled to SRAM
- High performance for SHMEM operations with minimal hardware support
  - Found relatively little need for "E-register" style accesses
  - Full bandwidth exploited when processors computed addresses

# Associative List Processing Unit Details

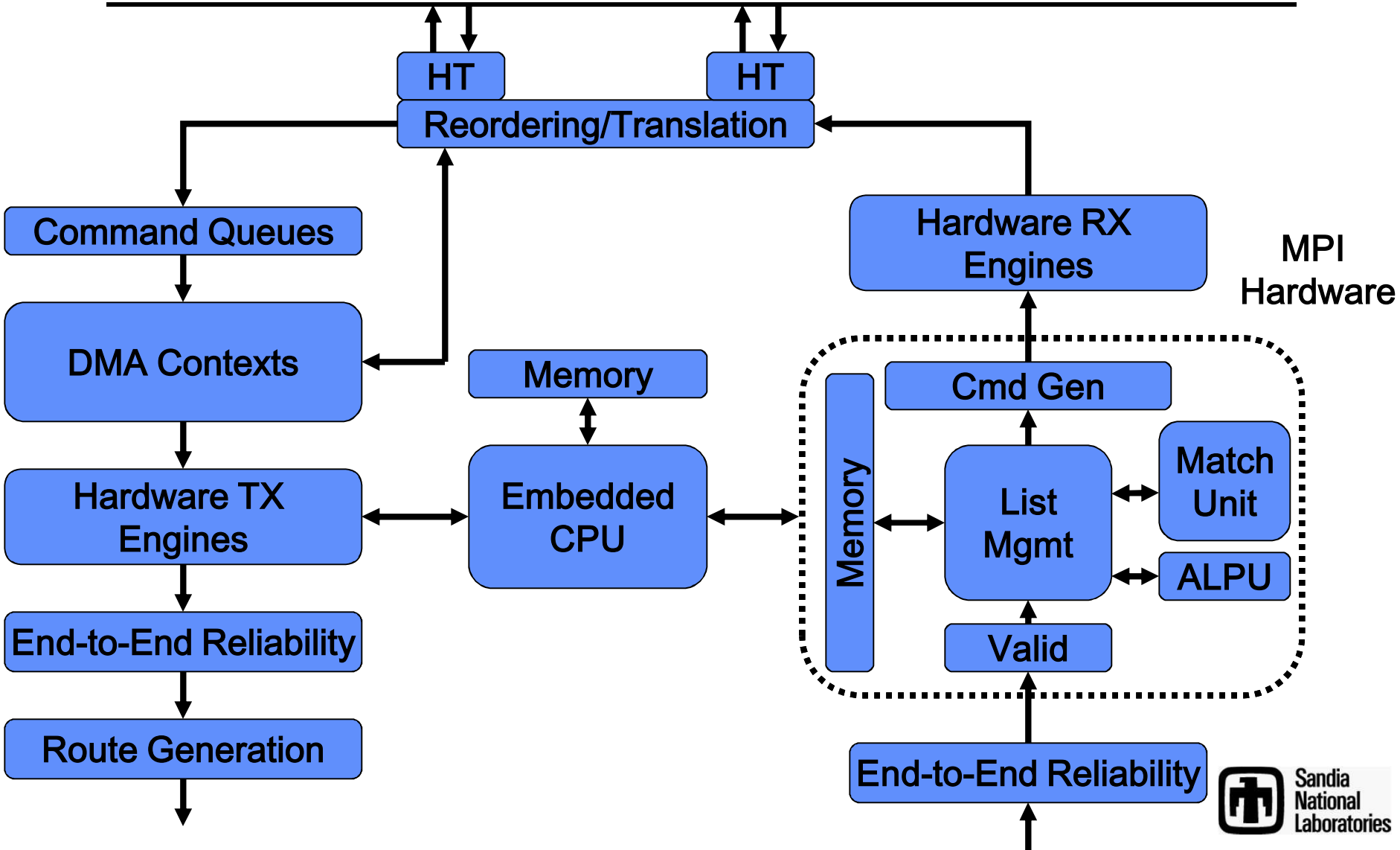



Sandia National Laboratories

# More Details

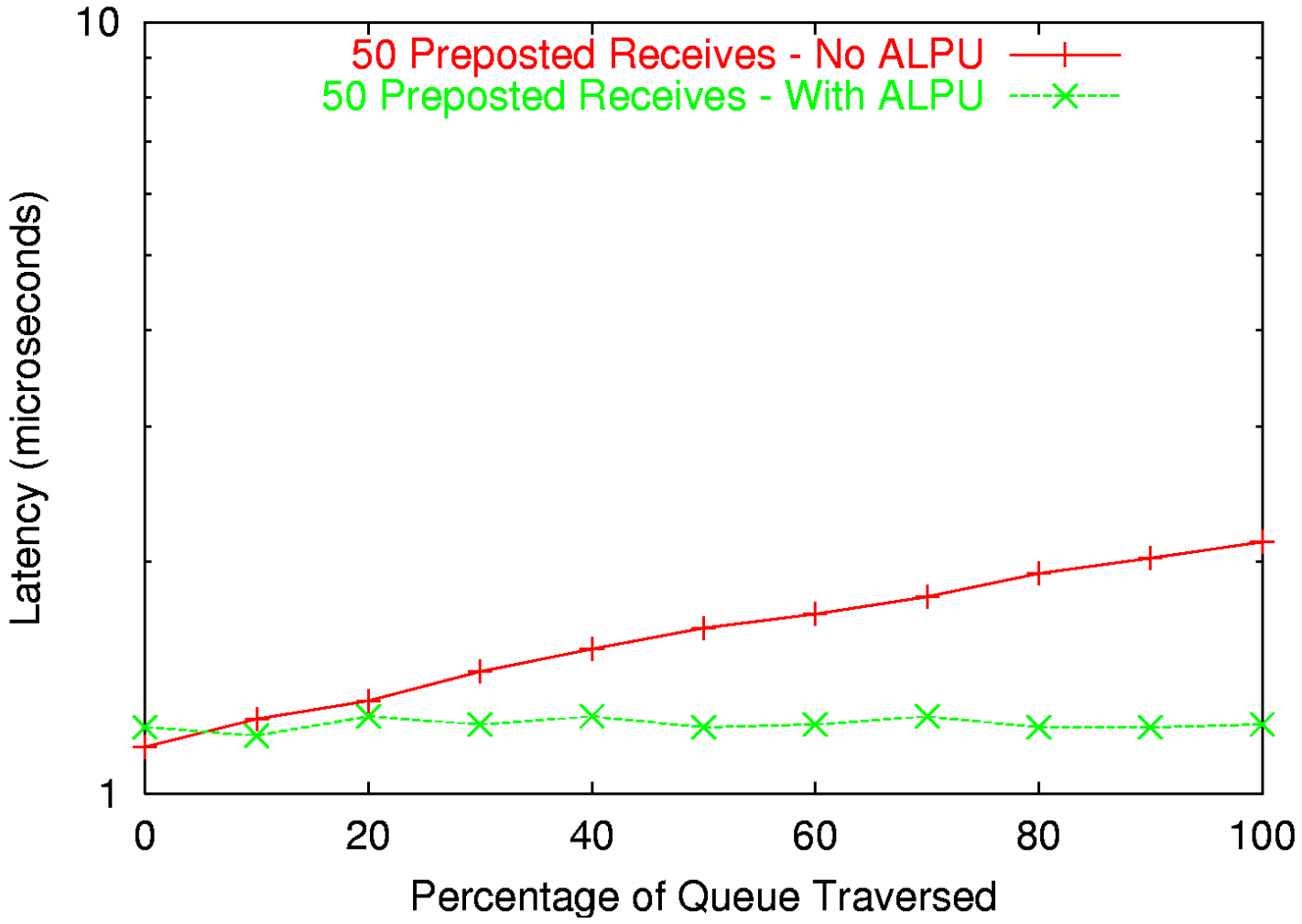


# Overall Block Diagram



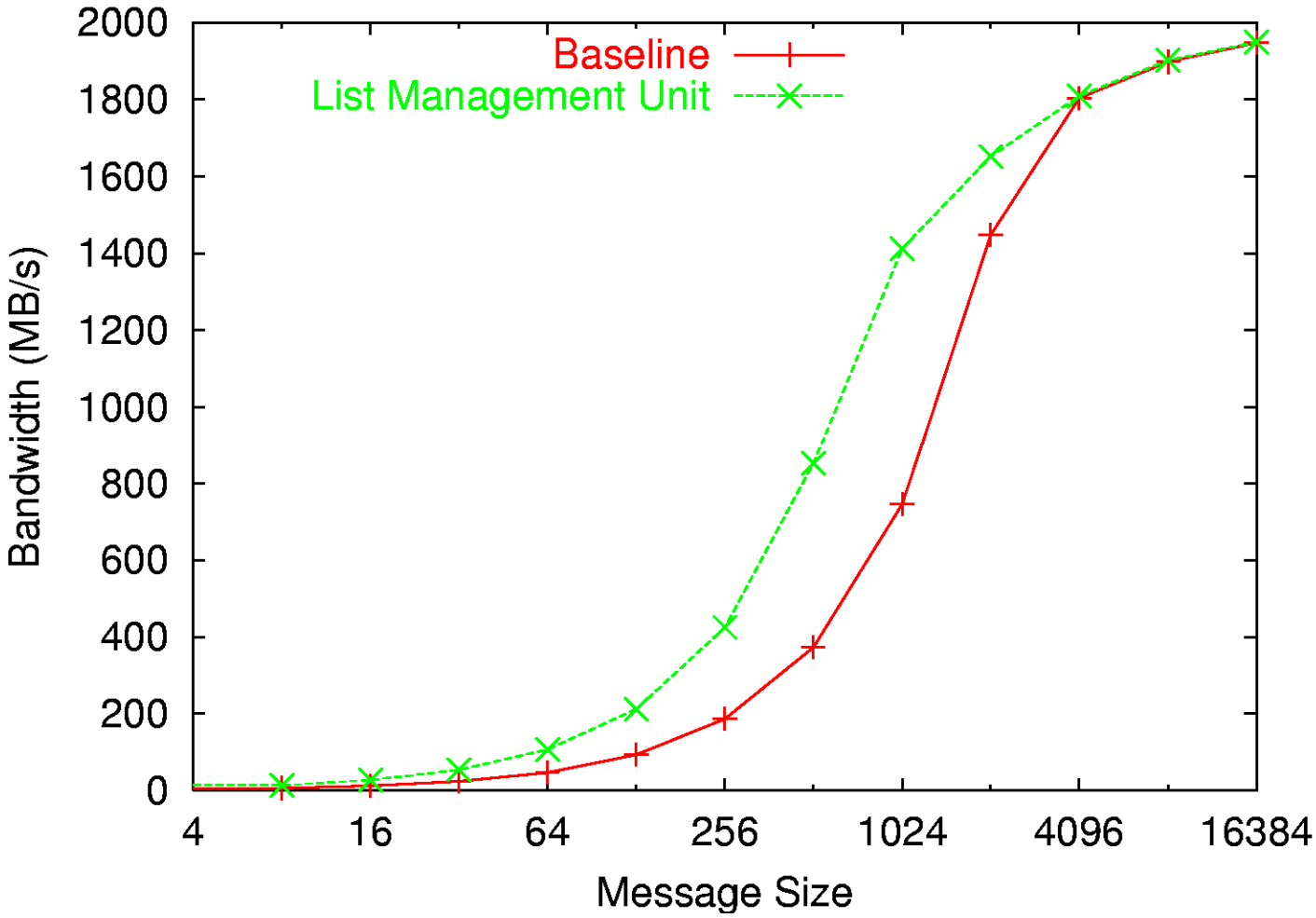

Sandia National Laboratories

# ALPU Results



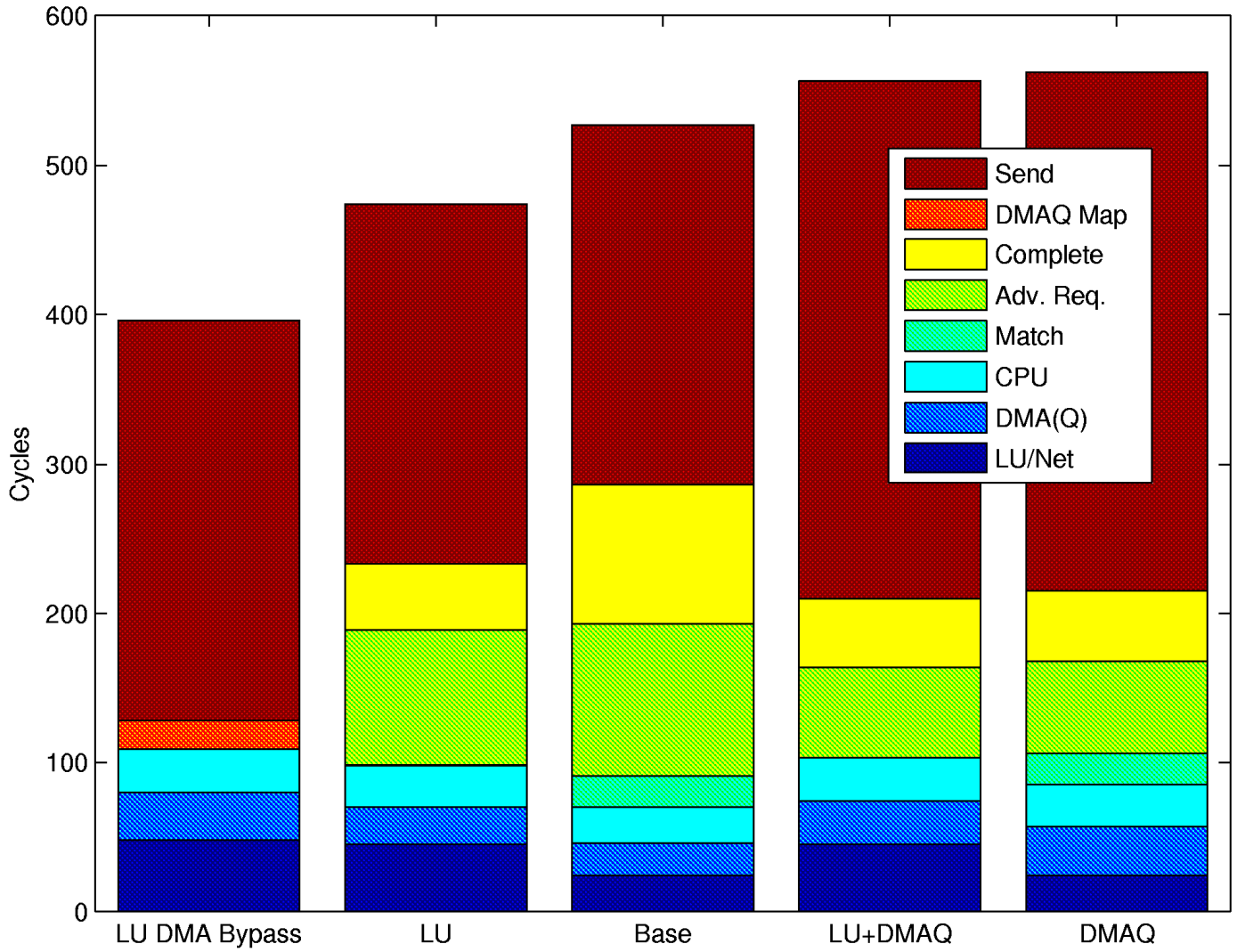

Sandia National Laboratories

# List Management Unit Results

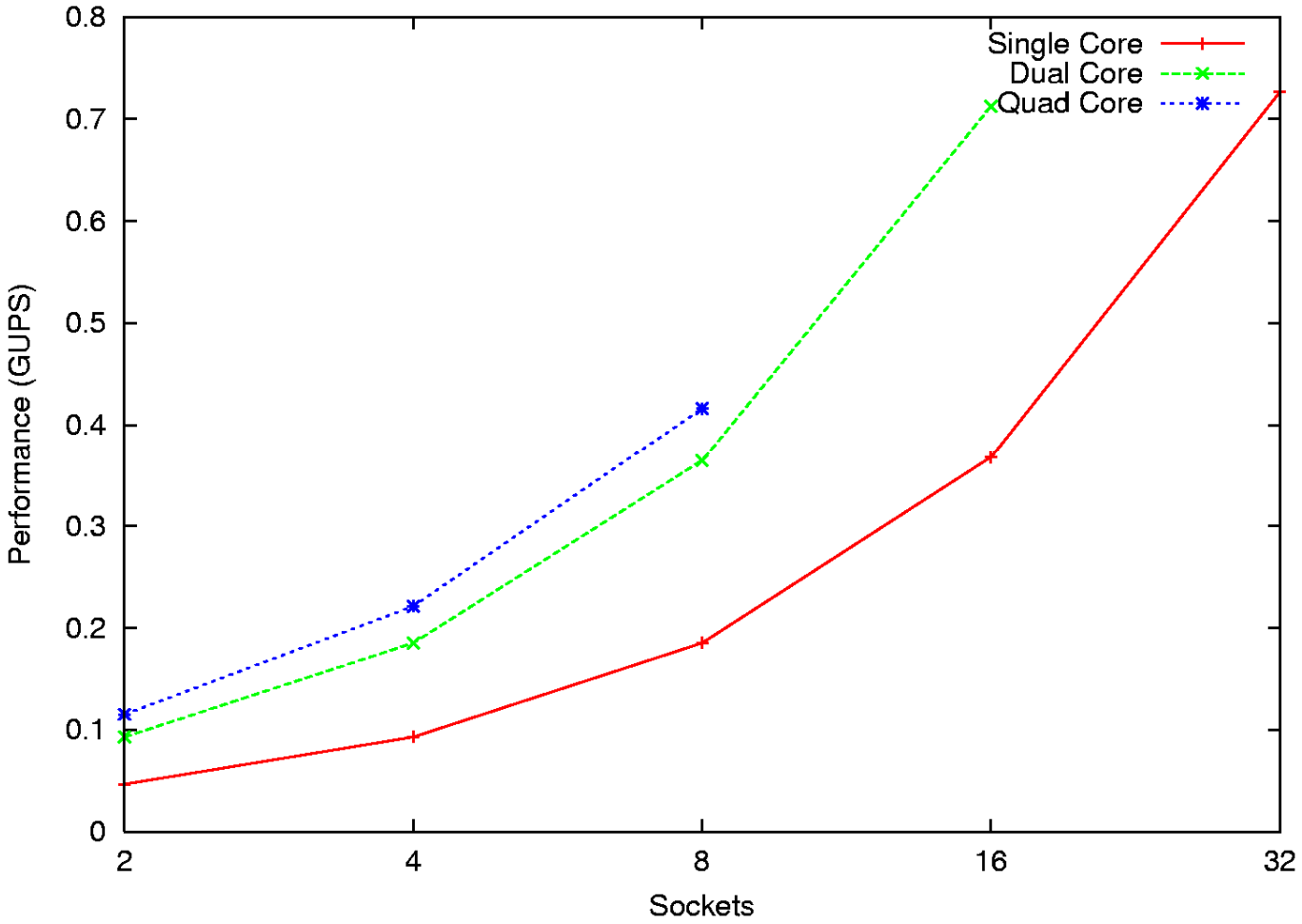



Sandia National Laboratories

# Timing Breakdown



# SHMEM GUPS Results




Sandia National Laboratories

# Conclusions

- Offload allows the processors to keep processing
  - Special hardware features are easier to implement in the NIC than in the processor
  - Simulations need to target realistic scenarios
- Special hardware can offer significant acceleration of MPI processing
  - Target common operations for optimizations
  - Build more power/area efficient structures
- SHMEM, in contrast, needs little special hardware
  - Address generation is a simple computation
  - Hardware interface is awkward for robust PGAS support