GASNet-EX at Lawrence Berkeley National Lab (http://gasnet.lbl.gov)

- GASNet-EX: communications middleware to support exascale clients
  - One-sided communication - Remote Memory Access (RMA)
  - Active Messages - remote procedure call
  - Implemented over the native APIs for all networks of interest to DOE
- GASNet-EX is an evolution of GASNet-1 for exascale
  - Retains GASNet-1's wide portability (laptops to supercomputers)
  - Provides backwards compatibility for the dozens of GASNet-1 clients, including multiple UPC and CAF/Fortran08 compilers
  - Focus remains on one-sided RMA and Active Messages
  - Reduces CPU and memory overheads
  - Improves many-core and multi-threading support
  - Implements the Atomic Domains concept (first introduced by UPC 1.3)
  - Domains permit use of NIC offload even when not coherent with CPU
  - Domains are created collectively outside the critical path
  - A Domain has an associated data type and set of allowed operations
  - Domains select the best implementation for the data type and operations
  - e.g. use offload if and only if NIC implements all the requested ops
- Example: non-blocking atomic fetch-and-add (FADD) on unsigned 64-bit integer
  - gex_AD_OpNB_U64(domain, &result, \*target += addend, flags)
  - flags includes optional behaviors and assertions, such as memory fences
- GASNet-EX provides a network-independent "reference implementation"
  - Uses Active Messages to perform operations using the target CPU
  - Uses GASNet-Tools for atomicity (inline assembly for numerous CPUs)
  - For Cray Aries improves performance vs. reference implementation
- Current enhancements:
  - "Immediate mode" injection to avoid stalls due to back-pressure
  - Explicit handling of local-completion (source buffer lifetime)
  - New AM interfaces, e.g. to reduce buffer copies between layers
  - Vector-Index-Strided for non-contiguous point-to-point RMA
  - Remote Atomics, implemented with NIC offload where available
  - Subsets teams and non-blocking collectives
- Future enhancements may include:
  - Offset-based addressing
  - Multiple endpoints/segments, e.g. to enhance multithreading support
  - Communication directly to/from device memory (e.g. GPUDirect)

Remote Atomics with Cray Aries NIC Offload

- Three different MPI implementations
- Two distinct network hardware types
- On four systems the performance of GASNet-EX matches or exceeds that of MPI RMA and message-passing:
  - 8-byte Put latency 6% to 55% better
  - 8-byte Get latency 5% to 45% better
  - Better flood bandwidth efficiency, typically saturating at ⅓ or ⅔ the transfer size

GASNet-EX RMA Performance versus MPI RMA and Isend/Irecv

Exascale Scientific Applications

- Exascale Scientific Applications
- GASNet-EX
- InfiniBand
- Cray XC
- IBM BG/Q
- Ethernet
- MPI

This work was supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration. This work was supported by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported under Contract No. DE-AC02-05CH11231. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. This research was supported in part by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.