Lawrence Berkeley National Laboratory
Recent Work

**Title**
Optimized collectives for PGAS languages with one-sided communication

**Permalink**
https://escholarship.org/uc/item/2p00516t

**Authors**
Bonachea, D
Hargrove, P
Nishtala, R
et al.

**Publication Date**
2006-12-01

**DOI**
10.1145/1188455.1188604

Peer reviewed
Optimized Collectives for PGAS Languages with One-Sided Communication

Dan Bonachea, Rajesh Nishtala, Paul Hargrove, Mike Welcome, Kathy Yelick

Partitioned Global Address Space Languages

- Collective interface specifically designed for PGAS Languages
  - Data movement: Broadcast, Scatter, Gather, Gather-All, Transpose
  - Supported of collective support in UPC and Titanium languages
  - Extensible to variable-contribution and teams-based subset collectives
  - Achieves performance not obtainable from language-level implementations
- Interface includes many collectives features not found in MPI-2
  - Fully non-blocking collectives
  - Allows overlap of latency with computation and other communication
  - Exploit global address knowledge when available
  - Avoids RDMA-based impl - no rendezvous or eager buffering costs
  - Explicit consistency flags for detailed control over data sync.
  - Syntactic subset collectives: data not produced/consumed in current phase
  - Per-thread sync: data has affinity to producer or consumer (MPI style)
  - Global sync: barrier-like data sync (more efficient than full barrier)

High Performance

- GASNet Portability
  - Native network hardware support:
    - Quadrant QNet III (BlueGene/L)
    - Cray X1 - Gray kshmen
    - Intel X1 - XG kshmen
    - Cray XT3 - Cray Portals (Hydid)
    - Daphne - SCI
    - InfinitBand - Mellanox VAPI
    - Myrinet Myrinet - CM-1 and CM-2
    - IBM Colony and Federation - CMPI
  - Portable network support:
    - Ethernet, UDP, works with any TCP/IP
    - MPI 1: portable mpi for other HPC systems
    - Berkeley UPC, Titanium & GASNet highly portable
    - Runtime and generated code all ANSI C
      - New platform ports in 2-3 days
      - New network hardware 2-3 weeks
    - CPUs: Intel, Itanium, Opteron, Athlon, Alpha, PowerPC, MIPS, PA-RISC, SPARC, S3E, X-1, IX-1, IX-2
    - Other Unix hosts FreeBSD, NetBSD, Tru64, AIX, IRIX, HPUX, Solaris, MS-Windows/Cygnus
    - Mac OSx, Unix, SunOS, Ultrix, Cattamount, BlueGene,

- GASNet on the Cray XT3
  - GASNet PortDel operations implemented over Portals PortDel
  - Remote access region divesed by Portals Memory Descriptor
  - Portals Events used for GASNet operation completion
  - PortDel injection throttled to prevent local event queue overflow
  - No remote stream generation
  - Local put source and Get destination regions:
    - copied through pre-pinned bounce buffers for small messages
    - pre-pinned of local region for messages >= 1 KB
  - GASNet Active Message layer currently prootyped over Mpi
    - Port to native Portals-based AM is underway
  - UPC LI Application
    - Compliant with but less than 1/2 code size of MPI based HPL
    - System tuned DTBMM for high floating point performance
    - Move multi-threading (user-based coroutines) that hold on long latency communication, using a highly asynchronous algorithm
  - UPC FT benchmark:
    - Each thread uses non-blocking Put to send message as soon as local FFT completes - overlap computation with communication
    - Bulk, multi-based and permuted-implementations - best on XT3