Optimized Collectives for PGAS Languages with One-Sided Communication

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

Partitioned Global Address Space Languages
- Partitioned Global Address Space (PGAS) Languages
  - Global pointers and distributed arrays
  - User-controlled array of data access codes
  - Communicate using impromptu reads & writes of remote memory
- Languages: UPC, Titanium, DooRay Fathar
  - Productivity benefits of shared-memory programming
  - Competitive performance on distributed memory
- Use Single Program Multiple Data (SPMD) control
  - Fixed number of compute threads
  - Global synchronization, barriers, collectives
- Explicit fast one-sided communication
  - Individual accesses and bulk copies
  - Berkeley implementations use GASNet

GASNet Portability
- Native network hardware support:
  - Quadrics/Qinu (Ili) (Ellen/Elia4)
  - Cray X1 - Gray shmm
  - SGI Altix - SG1 shmm
  - Cray XT3 - Cray Portals (Jason)
  - Dophin - SCI
  - InfinitBand - Mellanox VAPI
  - Myrinet Myrinet - CM-1 and CM-2
  - IBM Colony and Federation - LAPI
- Portable network support:
  - Ethernet - UDP: works with any TCP/IP
  - MPI 1: portable impl. 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, Cray, Altix, Alpha, PowerPC, MPSoC, PA-RISC, SPARC, T3E, X1, 3K-6
- Other Unix (FreeBSD, NetBSD, Tru64, AIX, IRIX, HPUX), Solaris, MS-Windows/Cygwin, Mac OS, Unix, SuperC, Catamount, BlueGene

GASNet on the Cray XT3
- GASNet Put/Get operations implemented over Portals Put/Get
  - Remote access region covered by Portals Memory Descriptor
- Portals Events used for GASNet operation completion
- Portals injection thread to prevent local event queue overflow
- No remote event generation
- Local Put source and Get destination regions:
  - Copies through pre-joined bounce buffers for small messages
  - Interlacing of local region for messages > 1 KB
- GASNet Active Message layer currently prototyped over MPI
  - Port to native Portals-based AM is underway
- UPC/LG Application
  - Compliant with but less than 1/2 code size of MPI-based HPL
  - System tuned DGBMM for high floating point performance
  - Moved multi-threading user-based contours (u) 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, slab-based and portlet-based implementations - half best on XT3

http://gasnet.cs.berkeley.edu
http://upc.lbl.gov