Cray SHMEM on XT3

Monika ten Bruggencate
monikatb@cray.com

Outline

1. Introduction to SHMEM
2. Cray SHMEM implementation on XT3
3. Cray SHMEM 1.0 Release on XT3
4. Future Work
5. References

1. Introduction to SHMEM

- Programming Model
  - Memory is private to each process:
    Remotely accessible, not shared
  - SHMEM is one-sided message passing model:
    Put and get operations

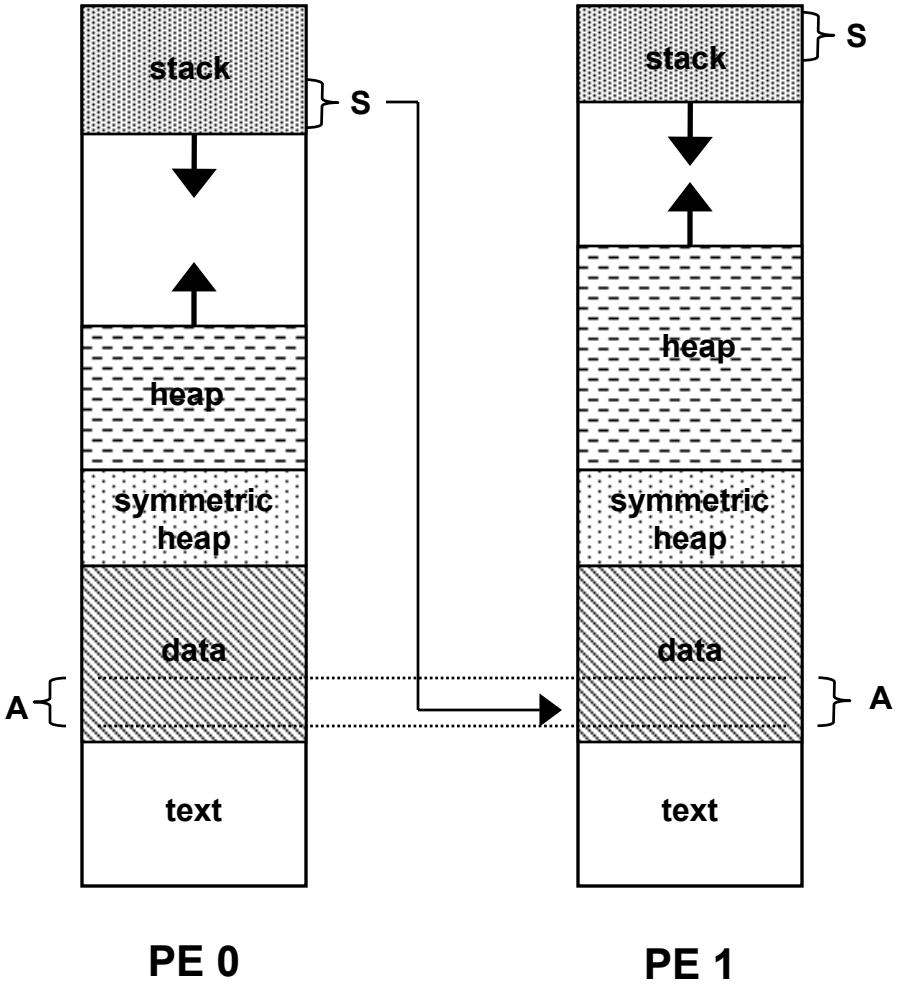
Introduction(cont.)

- Programming Model (cont.)
  - SHMEM is SPMD programming model
    - Processes run in parallel from launch to termination
    - No processes can be added or removed
    - All processes execute same application asynchronously
    - Synchronization for data exchanges
    - SHMEM application can be part of MPMD type MPI job

Introduction(cont.)

- Symmetric Data Objects
  - Primary concept in SHMEM
  - Virtual addresses of symmetric data object on different processes have definite, known relationship
    - ⇒ Access remote symmetric data objects by using address of corresponding local data object
  - C: global, static or shmalloc'd data
  - Fortran: common block, SAVE attribute or shpalloc'd data

Introduction (cont.)




```
long A[10];

void foobar(void){
     long S[10];
     if(my_pe == 0) {

        shmem_put64(A,S,10,1);

     }
}
```

Introduction(cont.)

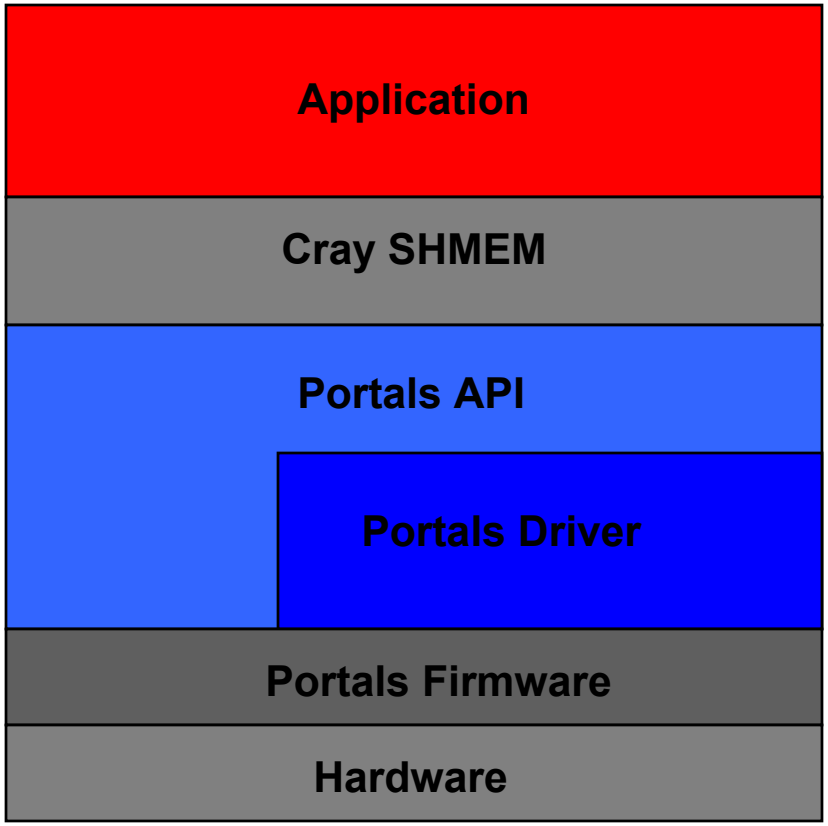
- Goal

  Deliver best possible communication performance by minimizing overhead associated with data transfer

2. Cray SHMEM Implementation on XT3

- XT3 uses Portals Networking Protocol
  - One-sided RMA protocol
  - Guarantees reliable, ordered message delivery between pairs of processes
  - Connection-less
  - Designed specifically for scalability
- Cray SHMEM layered on top of Portals 3.3

Cray SHMEM on XT3 (cont.)



Cray SHMEM on XT3 (cont.)

- Portals resources
  - Memory Descriptor (MD) identifies a memory region to be used in operation
  - Event Queue (EQ) used to record information about operation
- SHMEM start-up
  - Set up Portals resources
    - MDs to describe four memory regions
    - EQ to monitor transfer completions

Cray SHMEM on XT3 (cont.)

- SHMEM data transfer
  - Source and target addresses determine which MDs and EQs to supply to Portals call
  - Execute Portals put or get command
  - Monitor EQ for completion event if necessary
  - Persistent Portals resources => low overhead on transmit path

3. Cray SHMEM 1.0 Release

- Functionality Supported
  - Initialization and Clean up
    - shmem_init or start_pes
    - shmem_finalize
  - Queries
    - shmem_my_pe, shmem_n_pes
  - Puts and Gets
    - shmem_xxx_{put,get} (generic & different types)
    - shmem_{put,get}xxx (different bit counts)
    - shmem_{put,get}mem

Cray SHMEM 1.0 Release (cont.)

- Functionality Supported (cont.)
  - Synchronization
    - shmem_fence
    - shmem_quiet
    - shmem_barrier_all
    - shmem_barrier
  - Wait
    - shmem_xxx_wait (generic & different integer types)
    - shmem_xxx_wait_until (generic & different integer types)

Cray SHMEM 1.0 Release (cont.)

- Functionality Supported (cont.)
  - Broadcast
    - shmem_broadcastxxx (generic & different bit counts)
  - Reductions
    - shmem_xxx_yyy_to_all for operations sum, prod, max, min, and, or, xor (different types)
    - Currently supported on all PEs only

Cray SHMEM 1.0 Release (cont.)

- Functionality Supported (cont.)
  - Events
    - shmem_{clear,set,test,wait}_event
  - Strided Puts and Gets
    - shmem_xxx_i{put,get} (generic & different types)
    - shmem_i{put,get}xxx (different bit counts)

Cray SHMEM 1.0 Release (cont.)

- Functionality Supported (cont.)
  - Symmetric Heap management
    - shmalloc
    - shfree
    - shrealloc
  - Fortran Interface
    - Functions corresponding to C interface
    - include ‘mpp/shmem.fh’

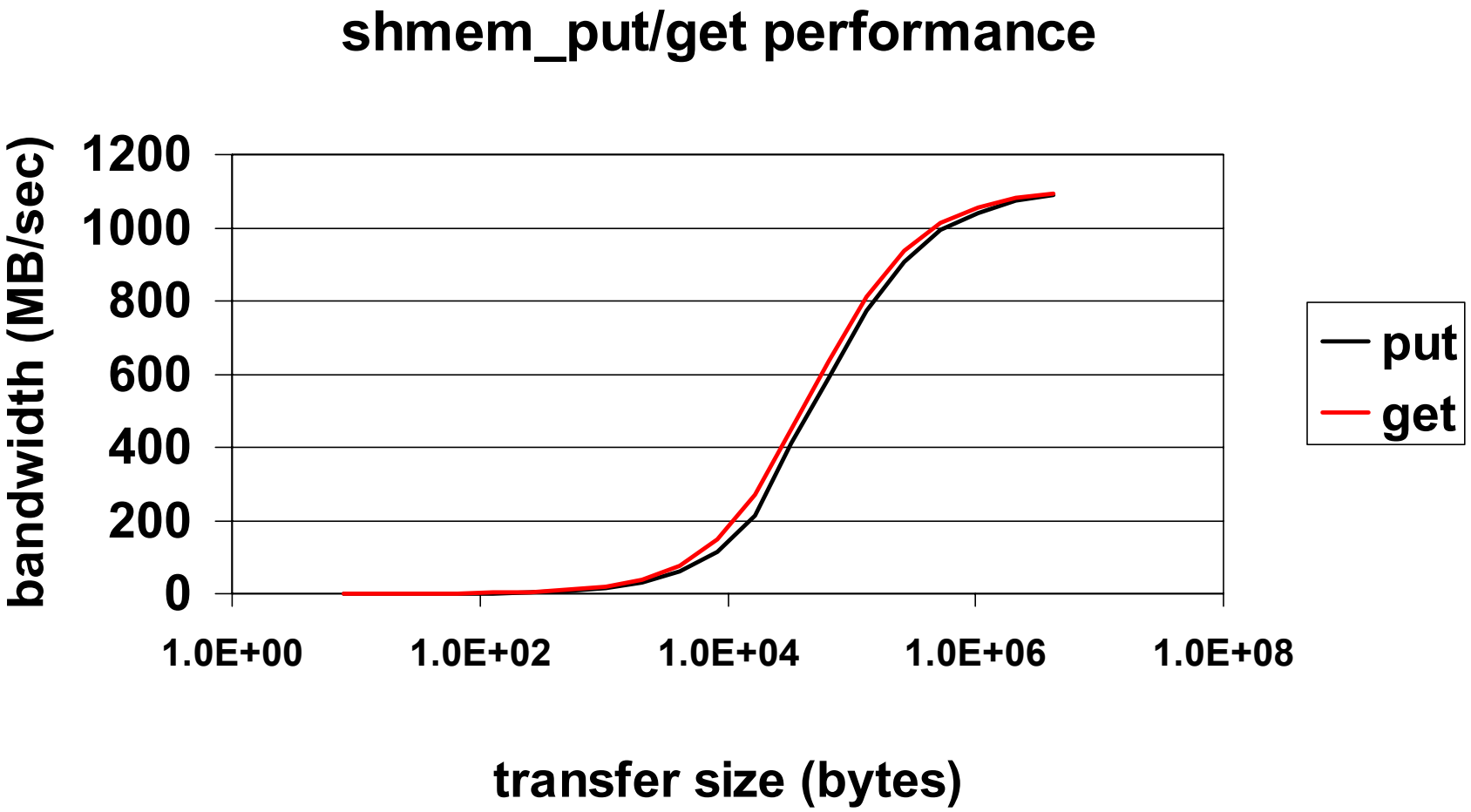
Cray SHMEM 1.0 Release (cont.)

- Preliminary Performance Data
  - Simple SHMEM get/put operations map well onto XT3 architecture
  - Advanced SHMEM operations do not map well onto XT3 architecture
  - Portals not tuned yet, e.g. no OS-bypass

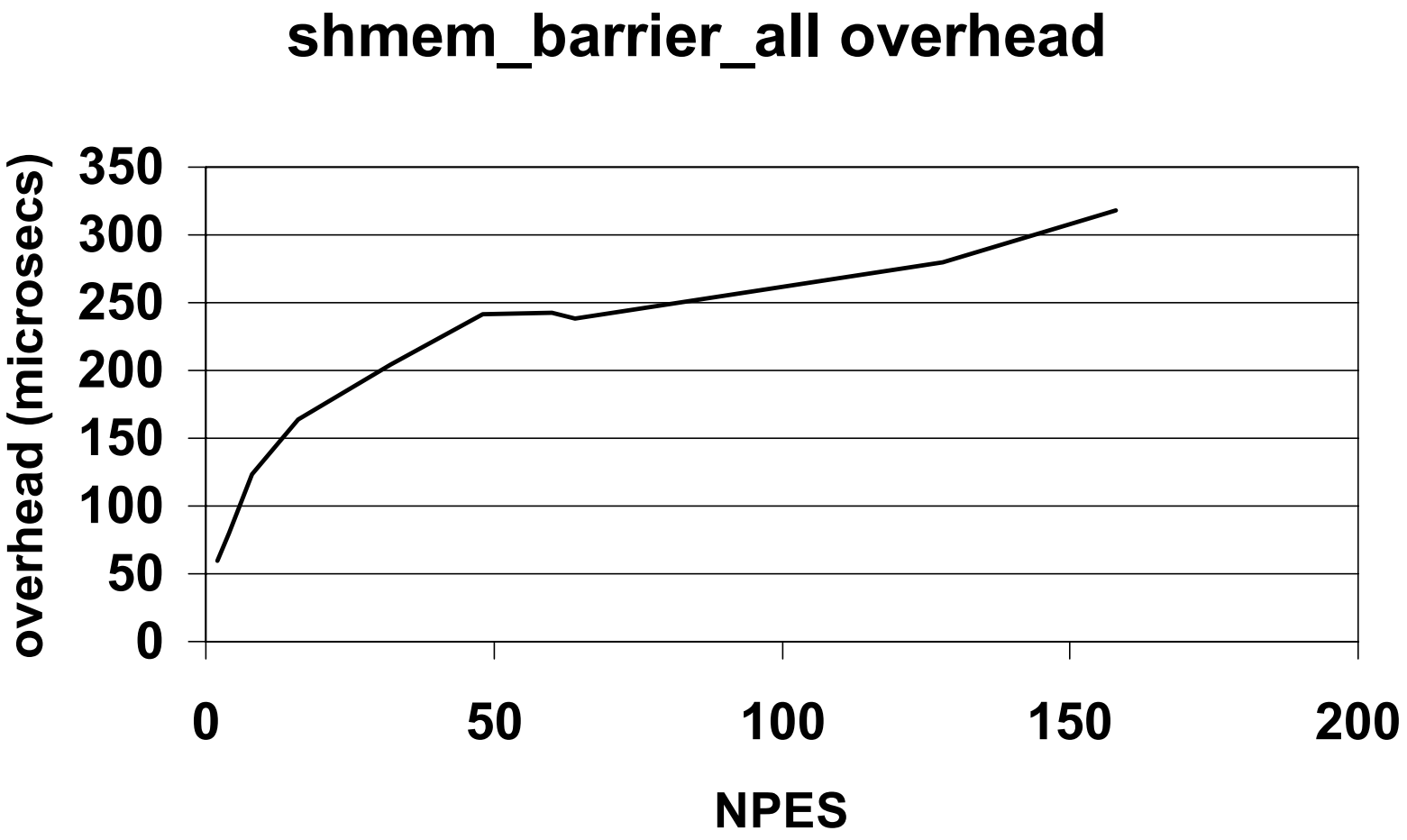
Cray SHMEM 1.0 Release (cont.)

- Preliminary Performance Data (cont.)
  - MPI latency
    25 usec for 8 bytes
  - SHMEM put latency
    22 usec for 8 bytes

Cray SHMEM 1.0 Release (cont.)



Cray SHMEM 1.0 Release (cont.)



4. Future Work

- Add Further Functionality
  - Non-blocking put and get operations
  - Atomic operations, depends on Portals work
  - Locking functions
  - Single element put and get operations
  - Your input is welcome
  - “USE SHMEM” ?
- Performance Tuning
  - Strided put and get operations
  - Collectives

5. References

- The Portals 3.3 Message Passing Interface
  R. Brightwell, R. Riesen, 2003
- Algorithms for Scalable Synchronization on Shared-Memory Multiprocessors
  J. Mellor-Crummey, M. Scott
  ACM Trans. On Computer Systems, 1991
- Networking Track
  CUG 2005