Hewlett Packard
Enterprise

OVERVIEW OF OPENSHMEM

Barbara Chapman
HPE and Stony Brook University

With thanks to Tony Curtis, Stony Brook University and Naveen Ravi, HPE

OVERVIEW

www.openshmem.org

- OpenSHMEM
  - PGAS-based library interface specification
  - Semantics and syntax of the model defined by the OpenSHMEM specification committee
  - Members include vendors (AMD, HPE, Intel, Nvidia,), labs (ANL, LANL, ORNL,), and universities (SBU, TU Dresden), individual contributors are welcome to join (it’s free!)

- Brief history
  - SHMEM first introduced by Cray for their T3D in 1993
    - Roughly the same time as first MPI standard
    - Lack of standard led to divergent implementations (HP SHMEM, IBM SHMEM, ..)
  - Work began on OpenSHMEM version 1.0 for C and Fortran in 2010
    - First reference implementation 2011

OPENSHMEM IN A NUTSHELL

- SPMD parallel programming library
  - Many functions similar to MPI (point-to-point, collectives)
  - Broadcast, reduction, barrier synchronization, ,..
- Partitioned Global Address Space = PGAS
  - Private and shared data (symmetric arrays, variables)
  - Shared data objects remotely accessible (put/get/collectives/atomics)
- Focus on performance via fast one-sided and collective communication
  - Remote data transfer: Remote Direct Memory Access (RMA, or RDMA)
    –MPI RMA is more cumbersome, has different address semantics
  - Direct exploitation of low-level network APIs, maximal asynchrony

```
shmem_init()

shmem_malloc(...)
shmem_free(...)

shmem_int_put(...)
shmem_int_atomic_fetch_add(..)
shmem_long_get_nbi(...)

shmem_barrier_all()

shmem_finalize()
```

NEED FOR OPENSHMEM

OpenSHMEM

Library interface specification

- MPI provides generality, sometimes at the cost of efficiency

- Ineffective for certain specific use cases
  - Applications with single-word or small-message communications
  - Applications with compute-defined communications with irregular and random-access patterns
  - Dynamic work stealing models

- Efficiently implemented PGAS programming scheme can satisfy these application requirements
  - OpenSHMEM RMA provides high performance, enables high levels of asynchrony

- Library approach has benefits with respect to maintainability and adaptability
  - Facilitates implementation of language-based PGAS model (e.g. Fortran coarrays, Chapel, UPC++)

- Active community with multiple implementations of the specification, including:
  - HPE Cray OpenSHMEMX, IBM Spectrum
  - Intel/Sandia SHMEM (SOS), Open MPI OSHMEM; Ohio State MVAPICH2-X; ANL OSHMPI
  - Reference implementation: Stony Brook University / OSSS

HPE CRAY OPENSHMEMX

HPE Cray OpenSHMEMX

Implementation of the specification

- What is HPE Cray OpenSHMEMX?
  - A key implementation of the OpenSHMEM specification
  - Implementation is over 25 years old
  - HPE proprietary library distributed as part of the HPE Cray Programming Environment

- Provides a scalable and performant implementation of the OpenSHMEM specification
- Introduces experimental features to allow users to evaluate capabilities for standardization

- Default OpenSHMEM implementation for various mission critical large-scale systems
  - Cray OpenSHMEMX = OpenSHMEM + X = specification-defined APIs + implementation-specific APIs

- Supports x86 and aarch64 platforms with specific optimizations for HPE Slingshot NIC
  - GPU awareness coming in 2024

- Multiple experimental features standardized over time – e.g. teams, multi-threading, non-blocking AMOs
- Some features under consideration for the specification are already being experimented with
  - Available to customers as SHMEMX-prefixed APIs

THANK YOU