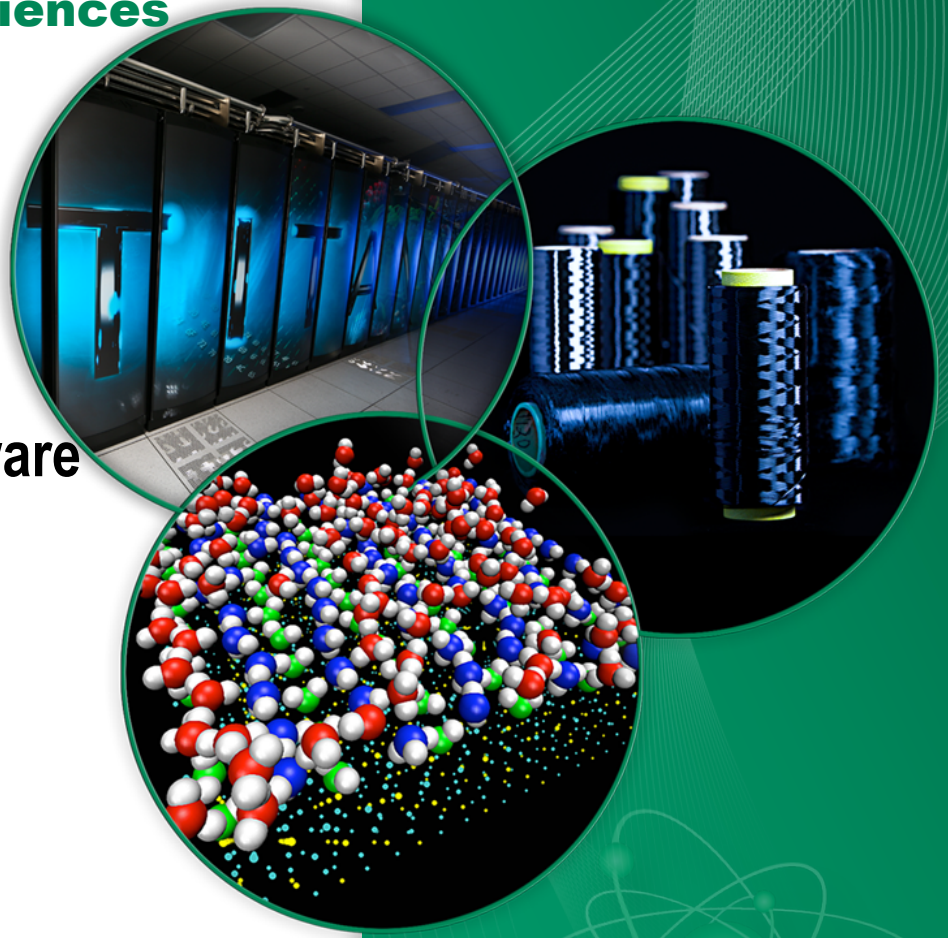
Oak Ridge National Laboratory
Computing and Computational Sciences



Designing a High Performance OpenSHMEM Implementation Using UCCS as a Communication Middleware

Pavel Shamis, Manjunath Gorentla Venkata, Stephen Poole, Aaron Welch, and Tony Curtis

Pavel Shamis

Background

- OpenSHMEM
  - Original SHMEM library designed by Cray Research
  - Merged with Silicon Graphics (SGI)
  - Multiple variants of SHMEM API by different vendors for different HW
- SGI SHMEM part of SGI's Message Passing Toolkit (MPT)
- Cray SHMEM (SeaStar, Gemini, Aries)
- Mellanox SHMEM
- HP SHMEM
- Quadrics SHMEM
- Portals SHMEM
- Despite broad availability it is not standardized
  - Application developers have to handle incompatibilities between application's of different SHMEM implementations.

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Introduction

- OpenSHMEM specification is an effort towards creating an open standard for the well-known SHMEM library.
  - PGAS based parallel programming model
  - OpenSHMEM 1.0 spec based on SGI's SHMEM API

- OpenSHMEM reference implementation (UH)
  - OpenSHMEM-GASNet
  - OpenSHMEM-UCCS aims to improve performance with a redesigned network-layer

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Abstracting the OpenSHMEM Communications Layer

- Network-specific aspects of communication handling identified and moved to separate layer
- Designed to be able to swap out one backend for another with minimal impact
- Communication calls that could be implemented with basic primitives (put/get/etc) were left in the upper layers
- Use and handling of multiple transports and/or rails left to the communications layer

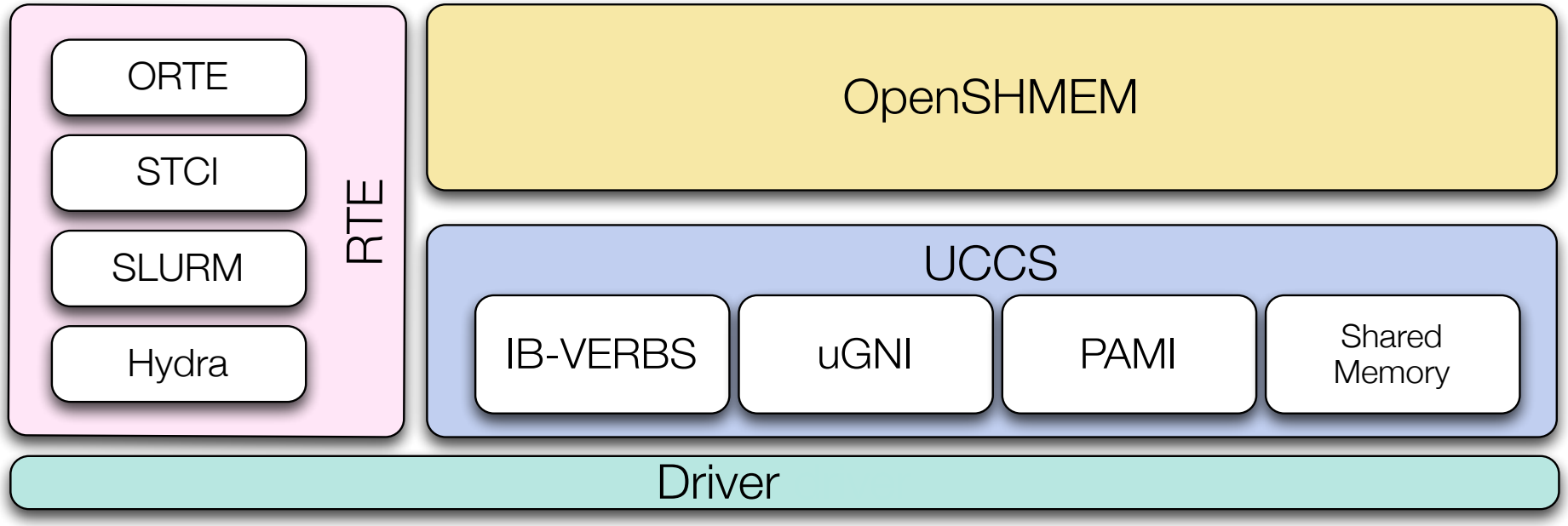
OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Background

- UCCS (Universal Common Communication Substrate)
  - Aims to provide a high performance low-level communication interface for implementing parallel programming models
- Broad range of communication semantics
  - Active messages, collective operations, puts, gets, and atomic operations for one-sided and two-sided communication
- Design goals:
  - Minimize software overhead
  - Provide direct access to network hardware

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

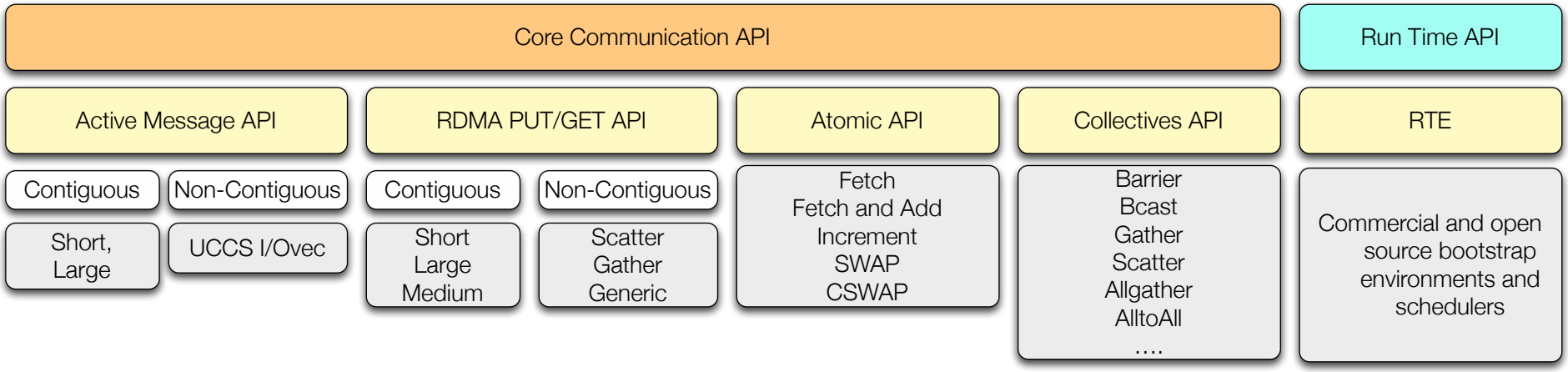
Design (OpenSHMEM UCCS integration)



OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Design (UCCS API)

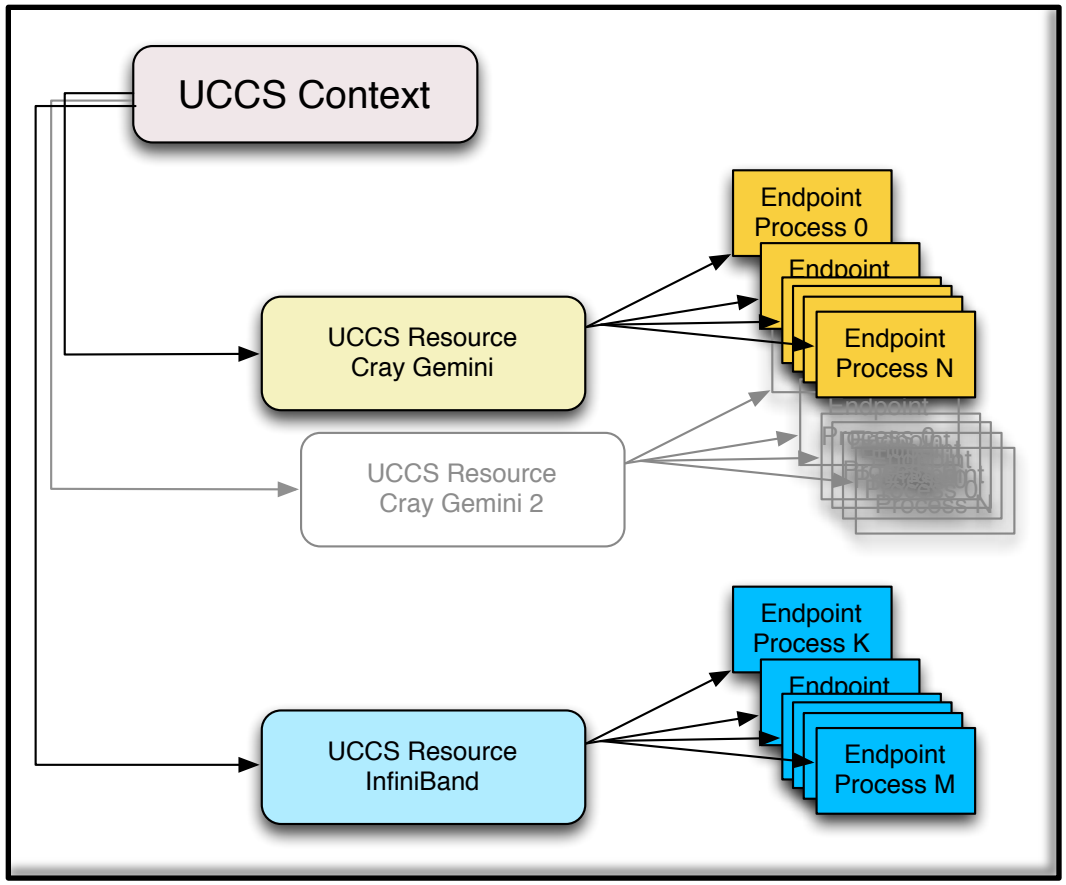
- UCCS interface divided between core API and run-time environment
  - Relies on external libraries: (ORTE, SLURM, STCI)



OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Design (UCCS API)

- UCCS API: Generic, hardware agnostic interface to isolate programmer from hardware specific details.
  - Key concepts:
    1. Communication contexts
    2. Resources
    3. Endpoints



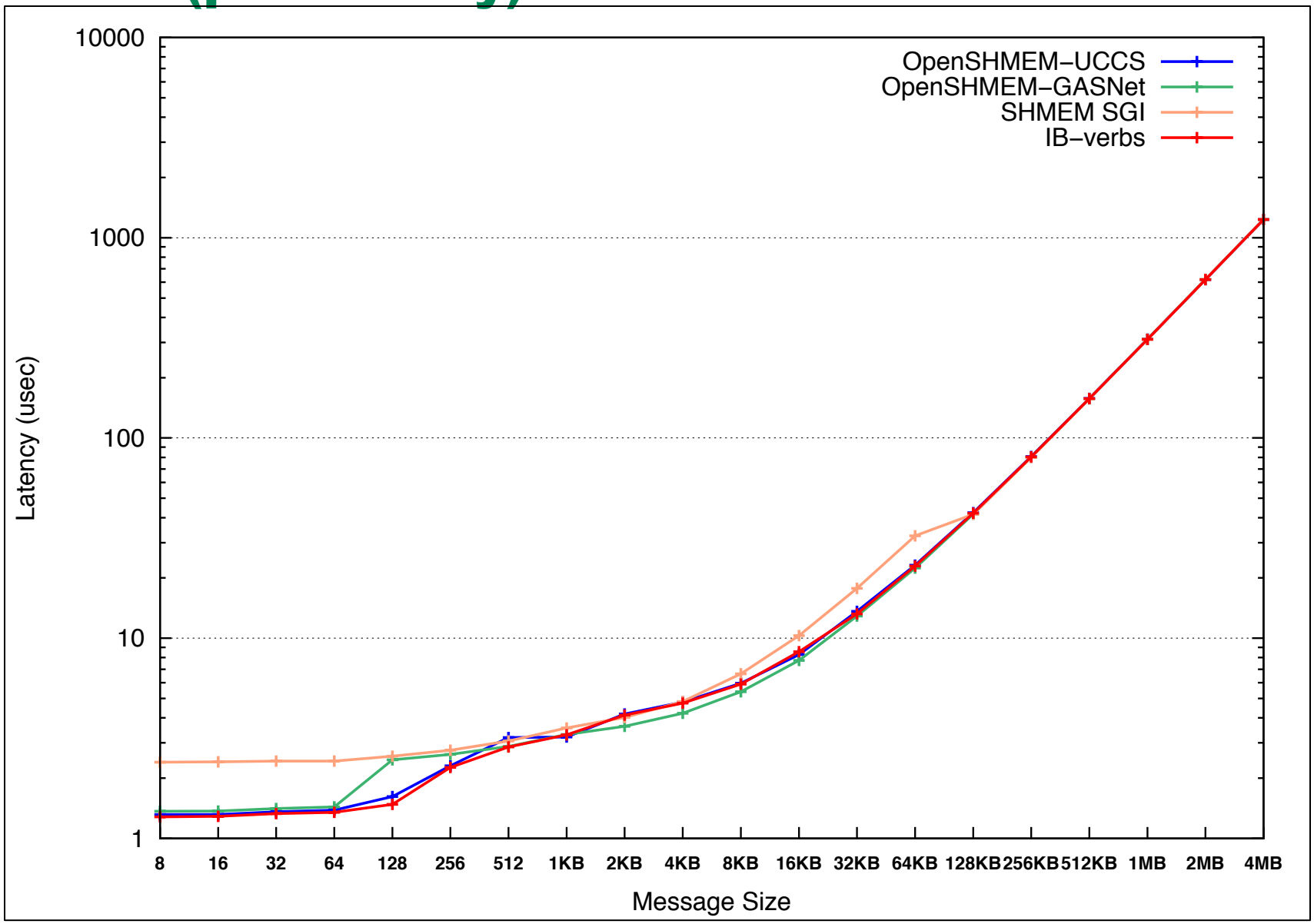
OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Test Platform

- OpenSHMEM reference implementation 1.0e
- GASNet v1.20.2
- Pre-production version of UCCS based on v0.2 of UCCS specification
- SGI Altix XE1300 system with 12 nodes with two Intel Xeon X5660 CPUs
- InfiniBand interconnect using Mellanox ConnectX-2 QDR HCA
- SGI MPT v2.03

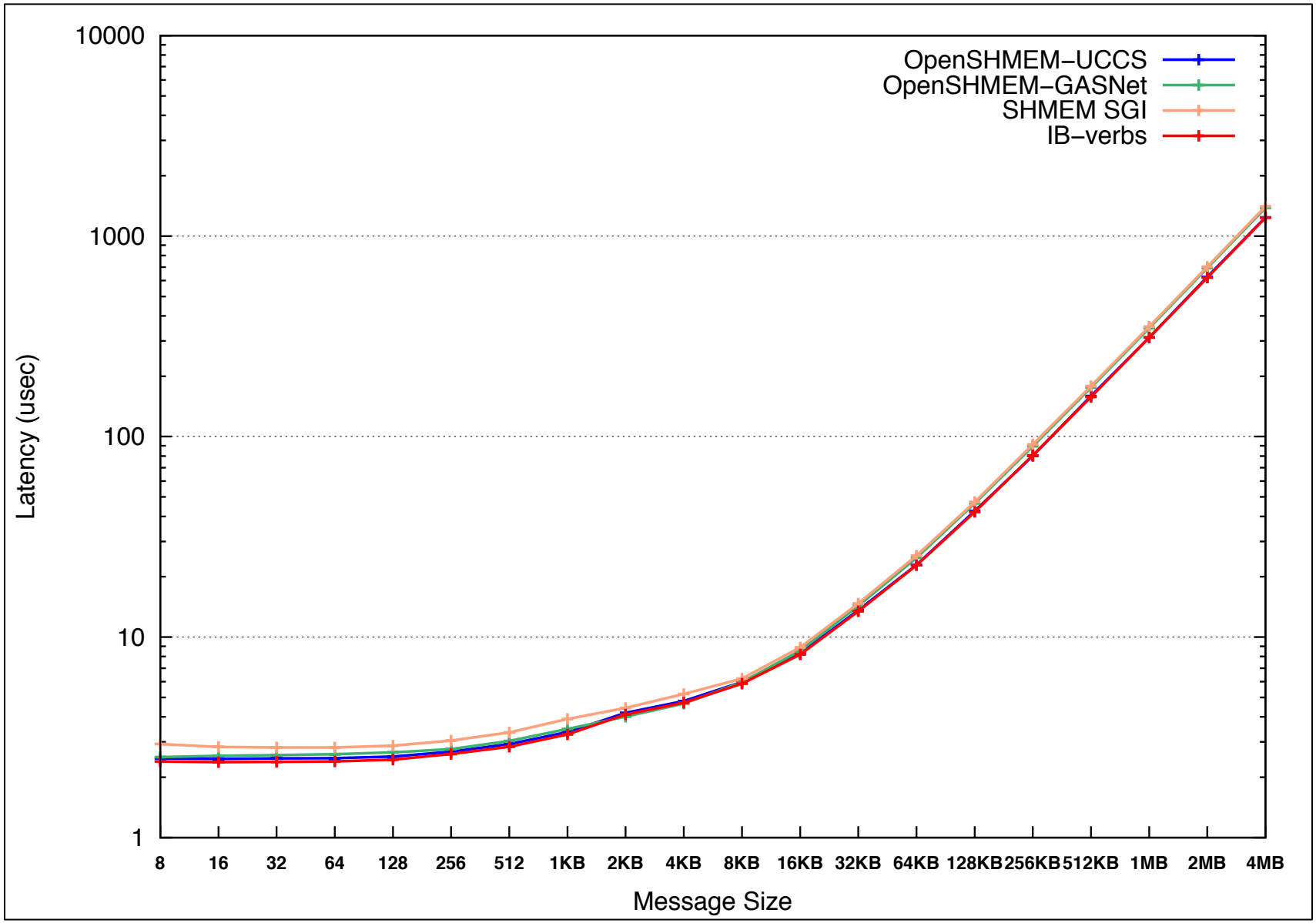
OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Results (put latency)



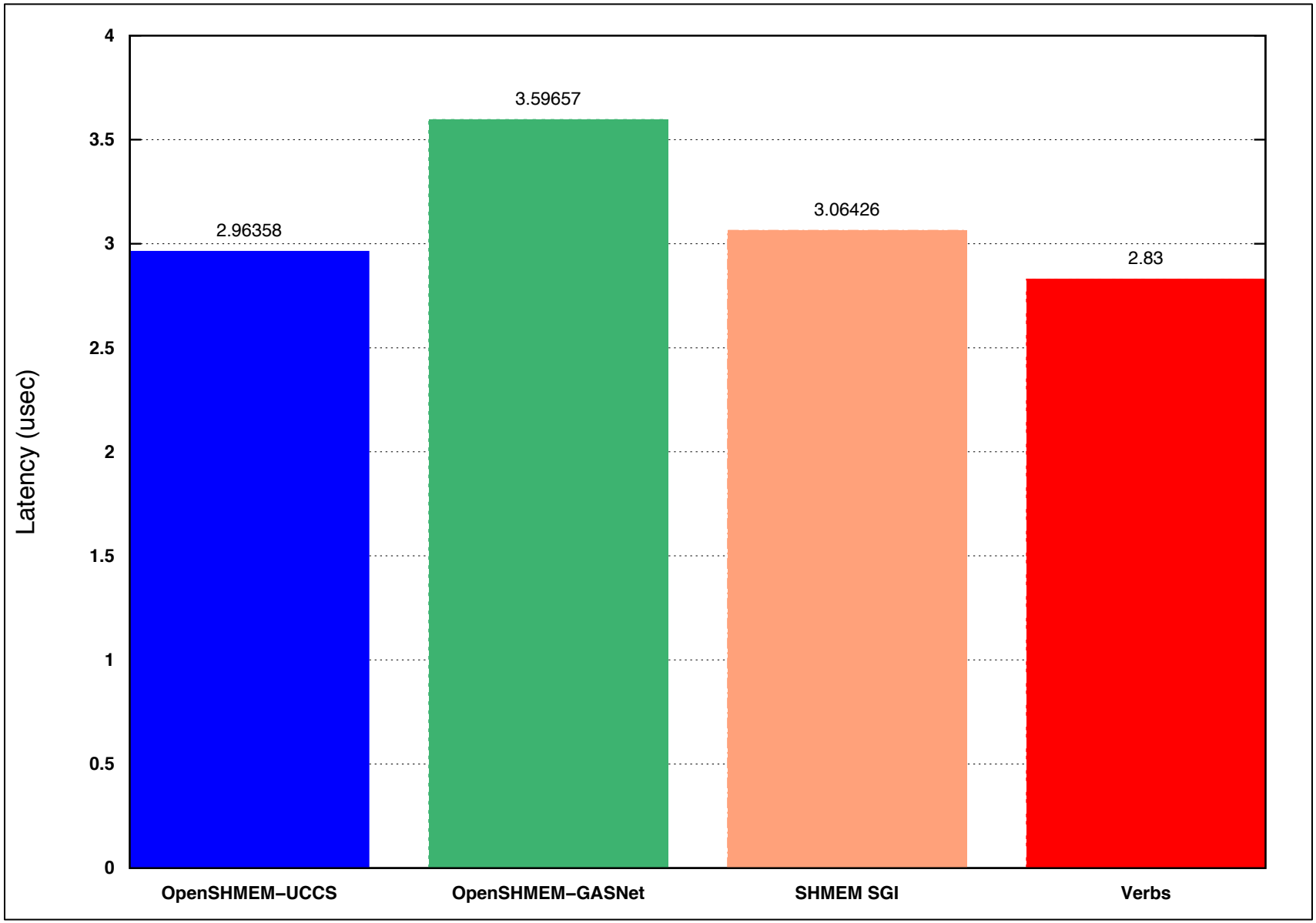
OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Results (get latency)



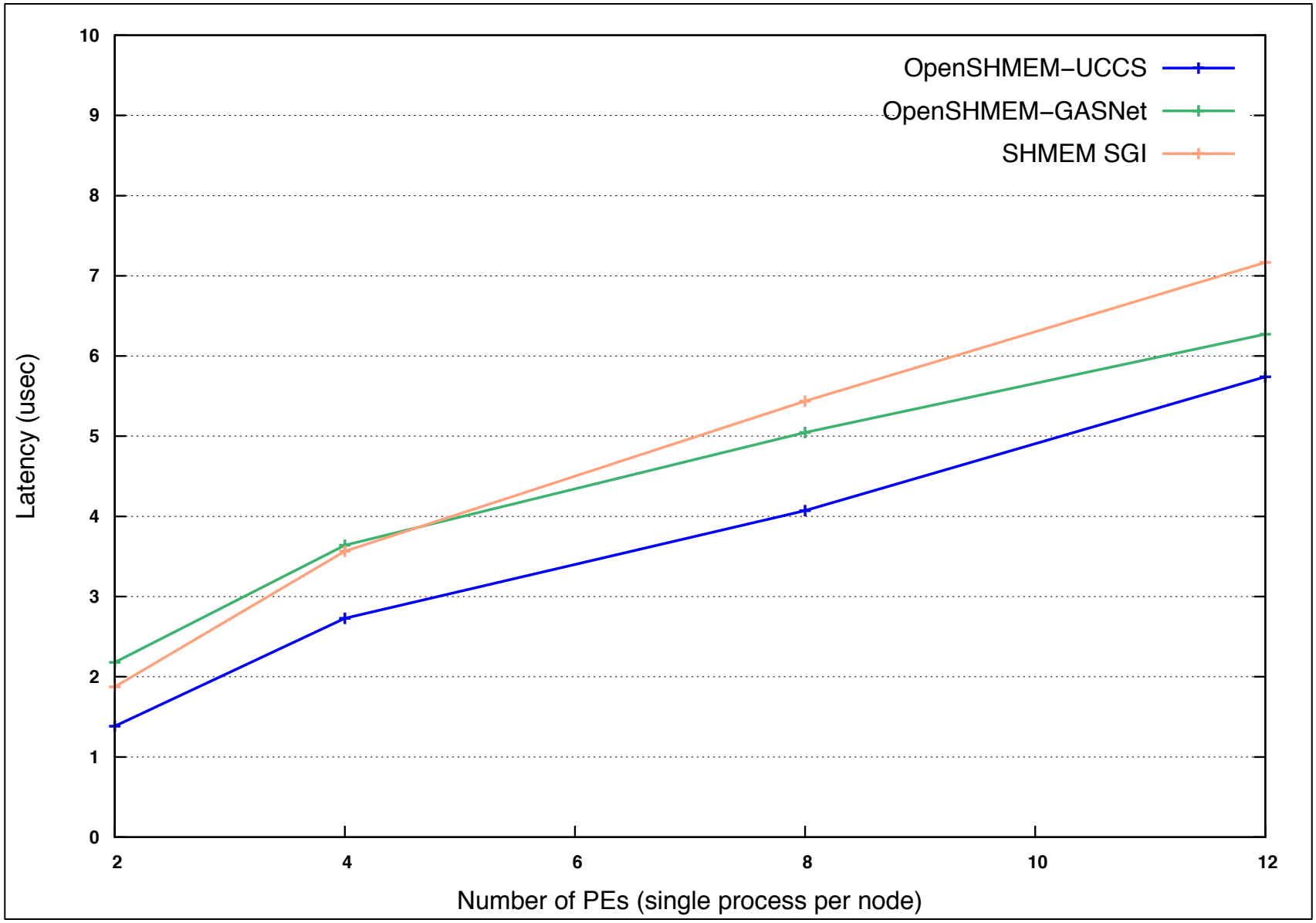
OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Results (Fetch-and-add)



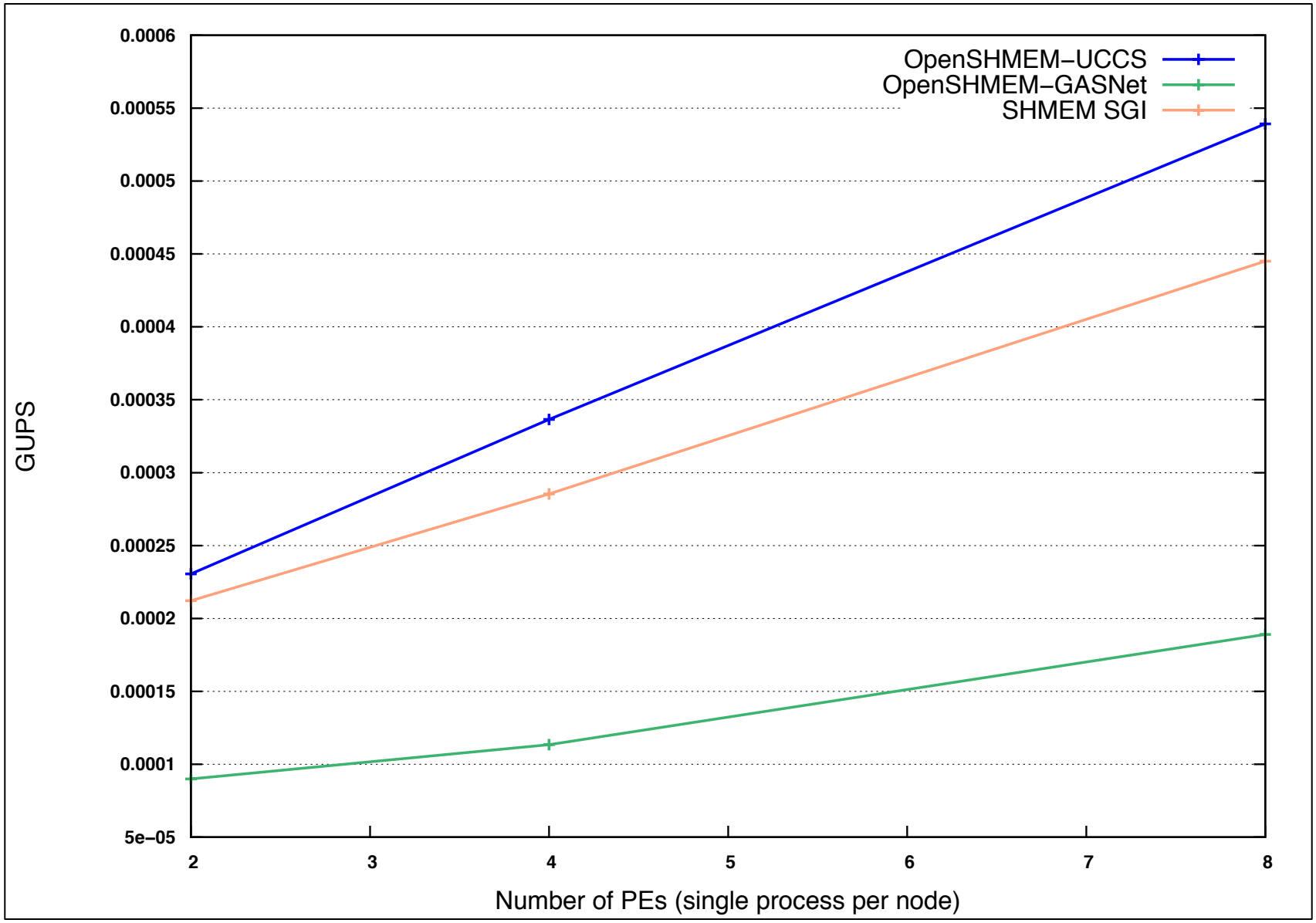
OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Results (hypercube barrier using put)



OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Results (GUPS)



OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Conclusions and Future Work

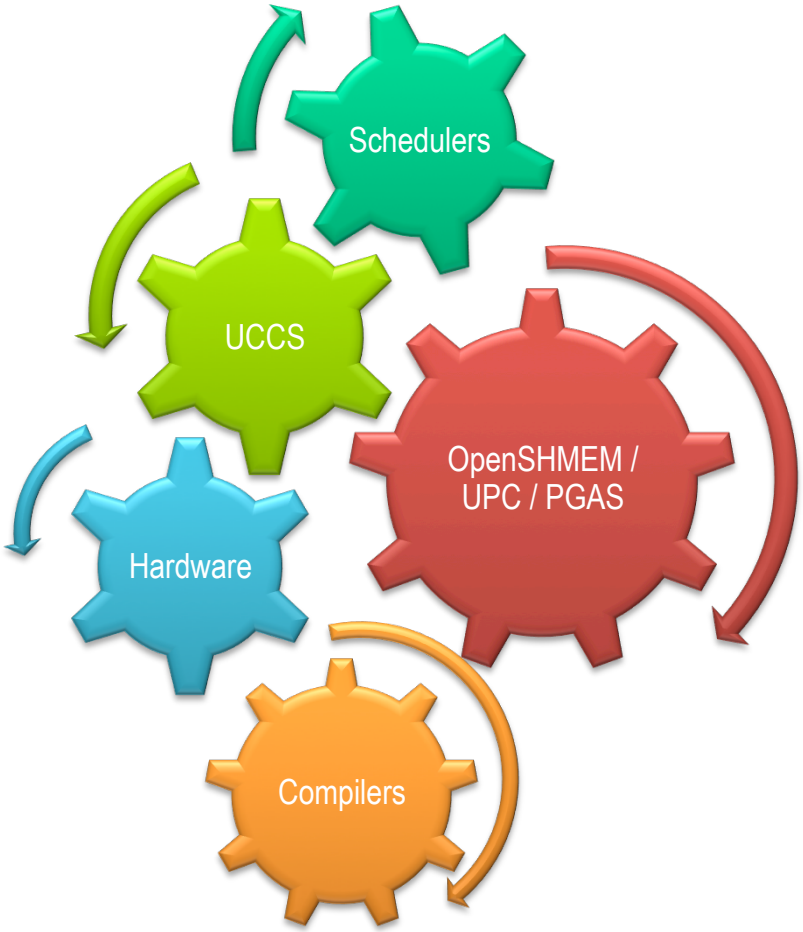
- OpenSHMEM-UCCS adds little overhead compared to native IB interface
- Extend the integration of UCCS with OpenSHMEM
- Extend UCCS to support intra-node communication
- Include additional transport layers Cray uGNI and IBM PAMI
- Extend UCCS InfiniBand transport layer to support advanced IB capabilities
- Extend AMOs, and on demand memory registration

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Acknowledgements

This work was supported by the United States Department of Defense & used resources of the Extreme Scale Systems Center at Oak Ridge National Laboratory.

Questions ?



OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY