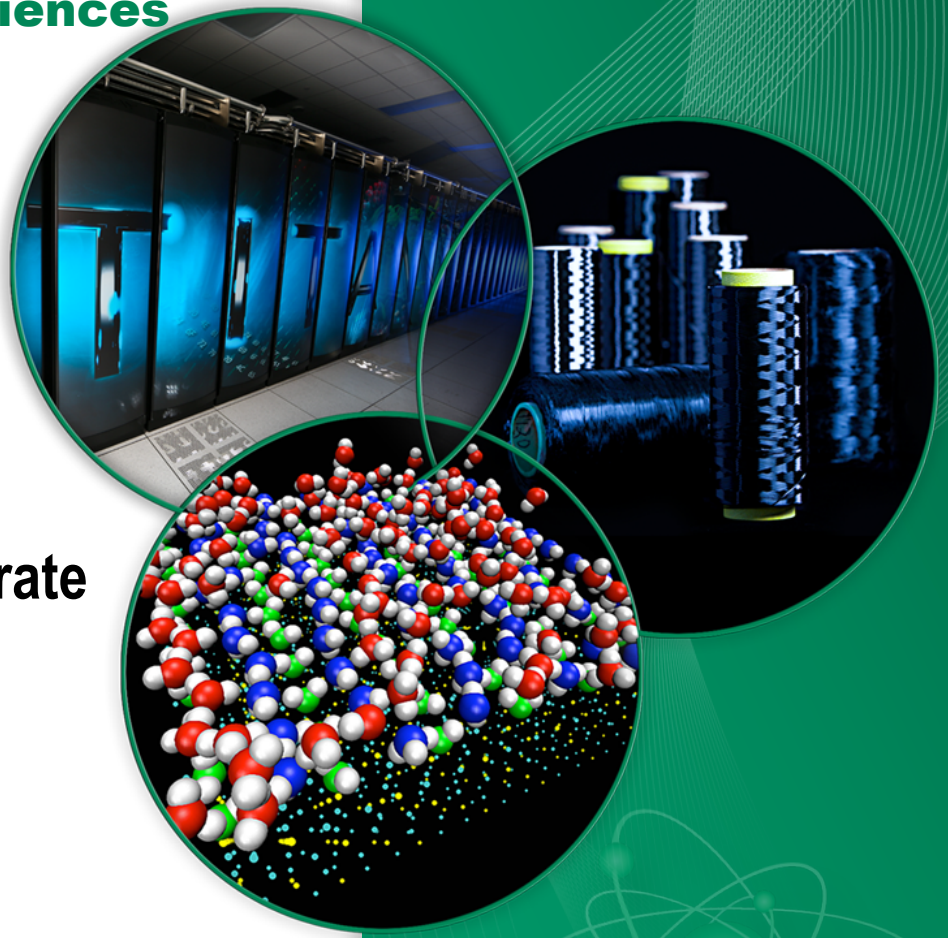
Oak Ridge National Laboratory
Computing and Computational Sciences



UCCS

Universal Common Communication Substrate

Presented by:

Pavel Shamis (Pasha)

Thomas Herault

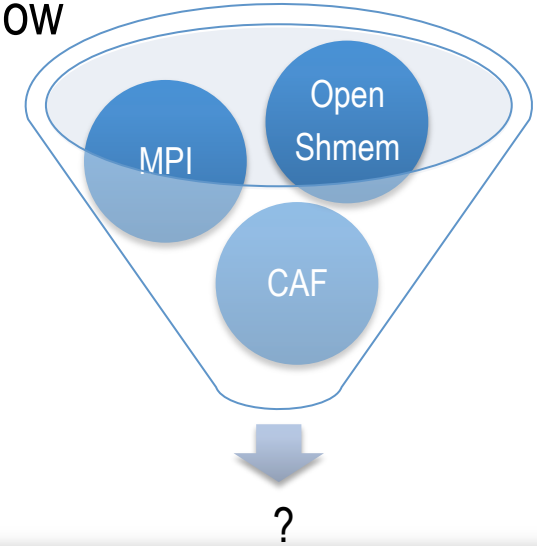
The Team

- **ORNL**
  - Pavel Shamis / Pasha
  - Manjunath Gorentla Venkata / Manju
  - Oscar Hernandez
  - Stephen Poole
  - Tomislav Janjusic / Tommy
  - Swen Boehm
  - Douglas Fuller
- **UH**
  - Tony Curtis
  - Donald Aaron Welch
  - Swaroop Pophale
  - Siddharta Jana
- **UTK**
  - George Bosilca
  - Thomas Herault
  - Aurélien Bouteiller
- **LANL**
  - Ginger Young
- **DoD**
  - Nick P.
  - Kevin B

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

History

- "*Déjà vu*" of OpenSHMEM implementation developer
  - We have seen this network code somewhere ?!
- Upper Layer Protocols can have a high degree of overlap in the requirements they place on the lower level network layers
  - Communication interface can have a high degree of overlap in communication semantics
    - Send/Recv, AM, RDMA, Collectives, etc
  - A lot of similarity in initialization and communication flow
  - Critical-path flow are similar but not identical



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

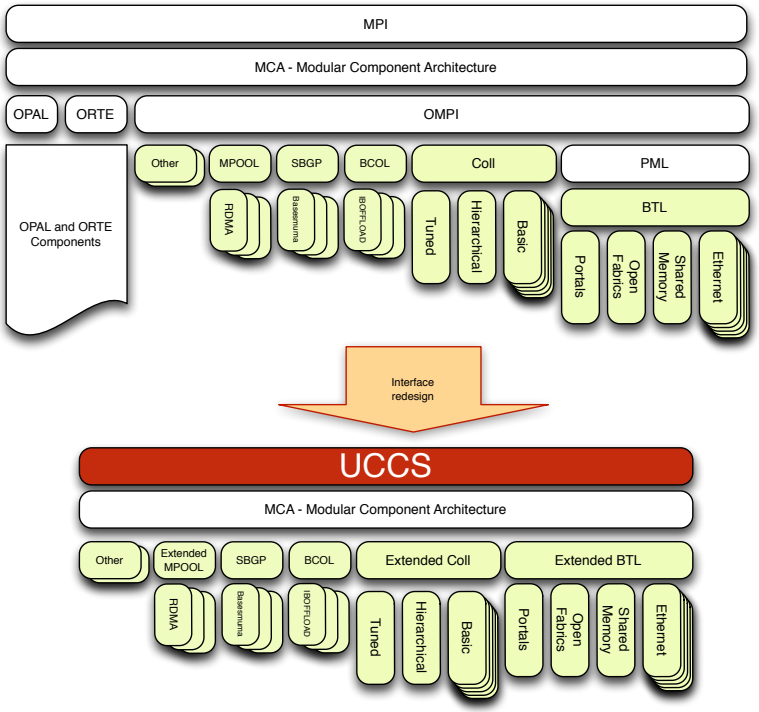
History - continued

- Idea of re-using high performance communication codes has been around for while:
  - ONET (~2009) – Rich Graham & Steve Poole
  - OpenSHMEM / “Yoda” (~2010) - Mellanox & ORNL collaboration

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

History - continued

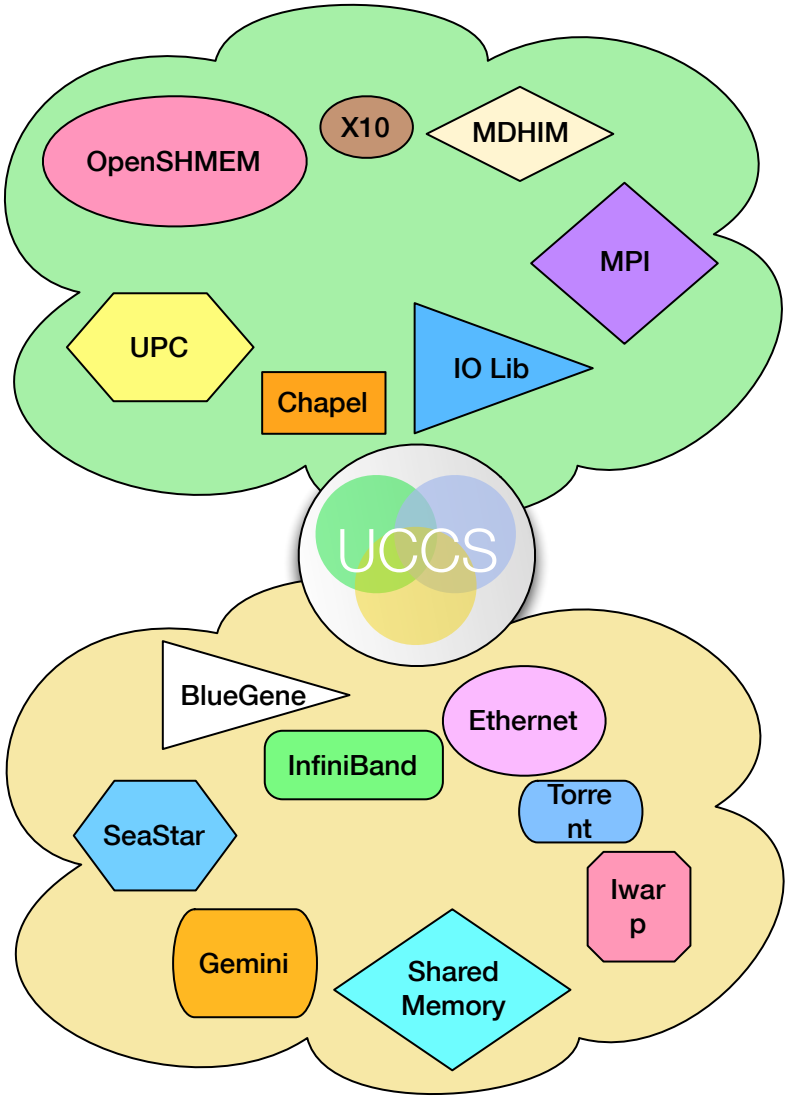
- Universal Common Communication Substrate (UCCS) – Beginning…
  - Let's re-use internal MPI network codes and expertise to design a *standalone* communication middleware that serves broader HPC community with an initial focus OpenSHMEM/PGAS (but not only…)
- In addition to high-performance implementation we would like to standardize the API



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

Goals

- Provide a common low-level scalable, robust, portable, simple and performance driven communication API for multiple parallel programming models over modern network interfaces
- Increasing code reusability and reducing development effort
- Include performance/power measurement capabilities in a central location



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

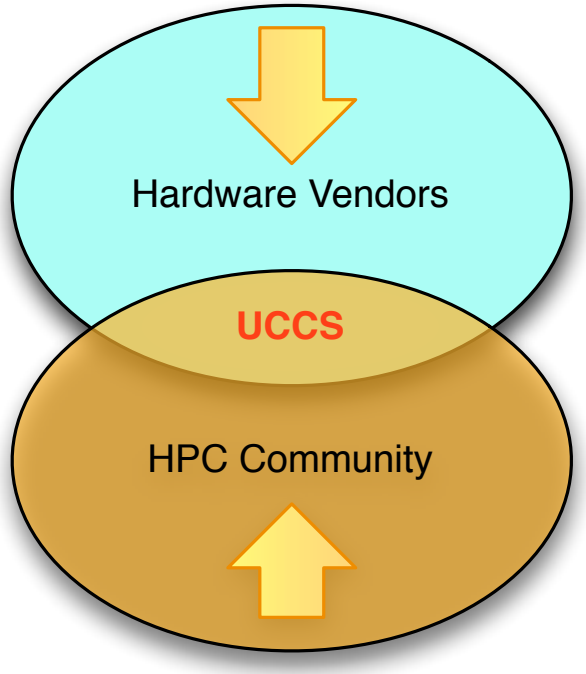
Goals - Continued

- Support hybrid programming environments efficiently
- Provide flexible API to accommodate requirements of I/O systems, Big Data applications, and Languages
- Runtime support for multiple network technologies (when possible)
- Provide an interface for code translation (CAF, UPC, etc.)
- Performance
- Define specification describing the communication middleware

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

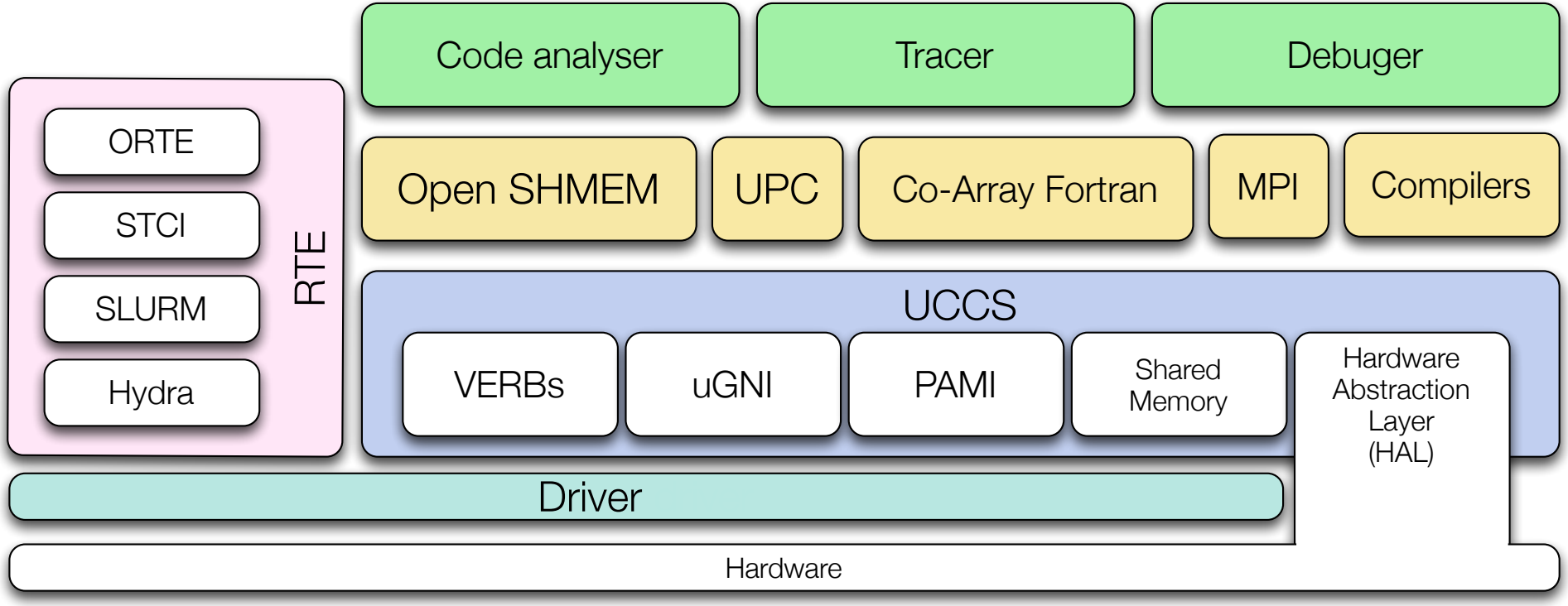
Long Term Goals

- Direct network hardware support
- Co-design
  - Hardware
  - Compilers
- Community support



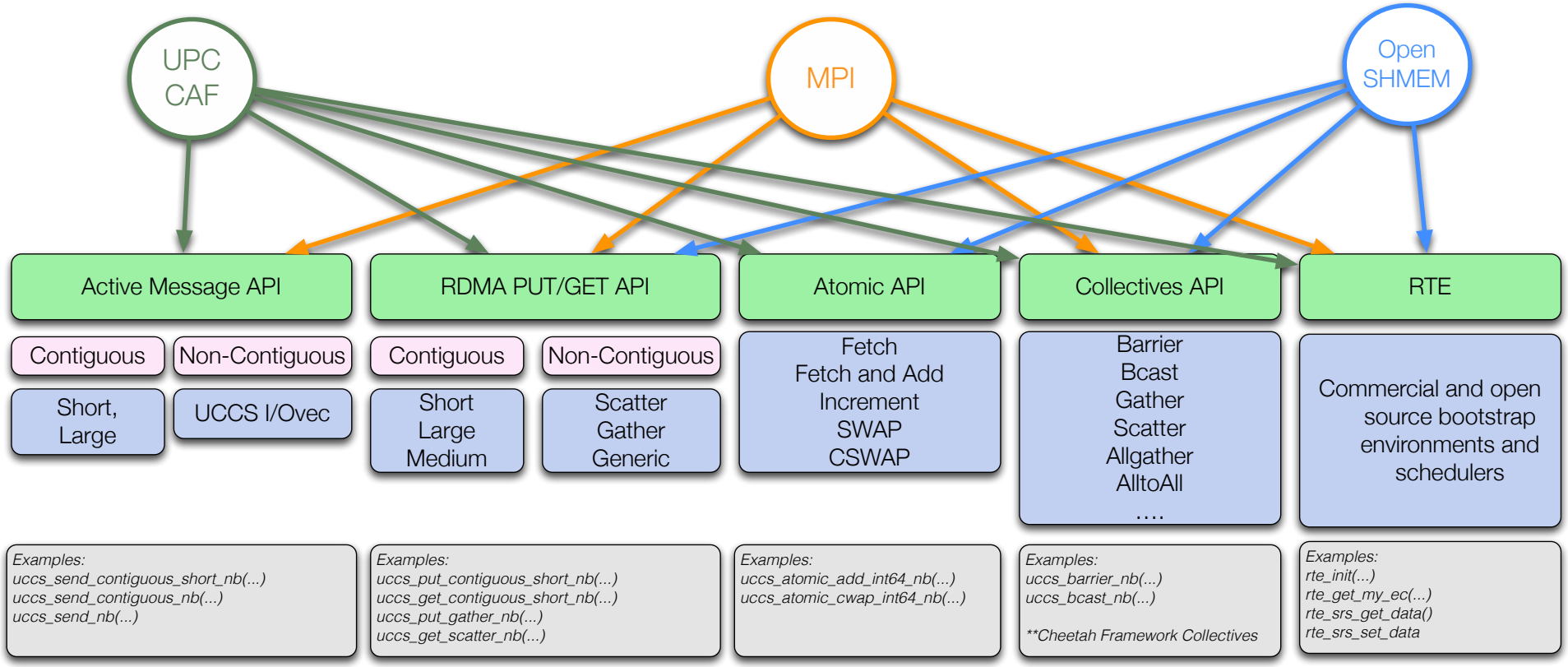
OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Overview



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

UCCS API



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

Implementation Details

- Initial code was based on Byte Transfer Layer (BTL)….
  - But we had to rewrite most of the critical path

- Modular Component Architecture (MCA)
  - Based on Open MPI MCA, which is essentially dynamically loaded libraries/components
  - Available as a standalone library: http://uccs.github.io/libocoms/
  - OCOMS – Open Component Module Service

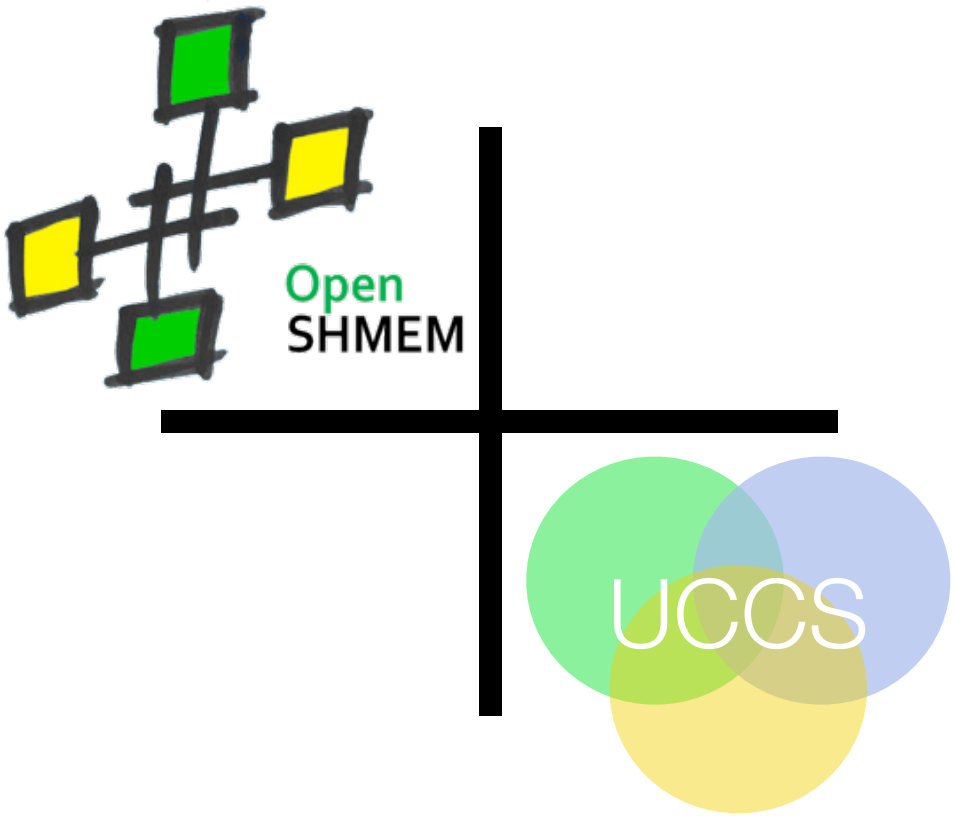
OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Implementation detail - continued

- Runtime Environment Abstraction – libRTE
  - A standalone Abstraction for Runtime environments
    - STCI, ORTE, ALPS, SLURM
  - Will be available soon: http://uccs.github.io/librte/
- Collectives - http://www.csm.ornl.gov/cheetah/
  - Work in progress
- We don't like "bundles"
- UCCS Specifications v0.1, v0.2, and v0.3 work in progress
- Supported communication methods (pre-production):
  - Infiniband , Cray, shared memory

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

OpenSHMEM and UCCS



UNIVERSITY of HOUSTON

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

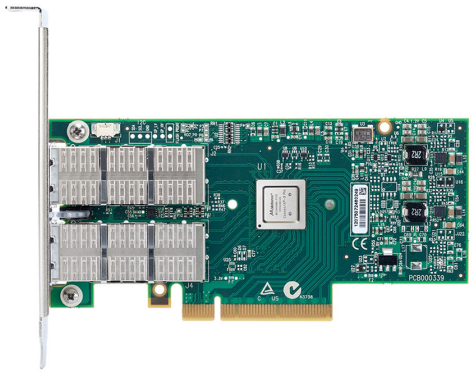
Update

- OpenSHMEM Reference Implementation
  - Internal interfaces were extended to support UCCS and libRTE and continue to support GASnet
  - About 90% complete
    - More testing required
- UCCS is a used as development platform for future OpenSHMEM extensions and research
    - Non-blocking communication
    - Extended network operations
    - Collectives, etc.
- OpenSHMEM-UCCS pre-production version is used internally for extensions evaluation and application development

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

InfiniBand

- Mellanox Connect-X and Connect-IB HCAs provide technology enabling efficient and high-performance implementation of OpenSHMEM and UCCS
  - Low software overhead
  - Hardware Offload

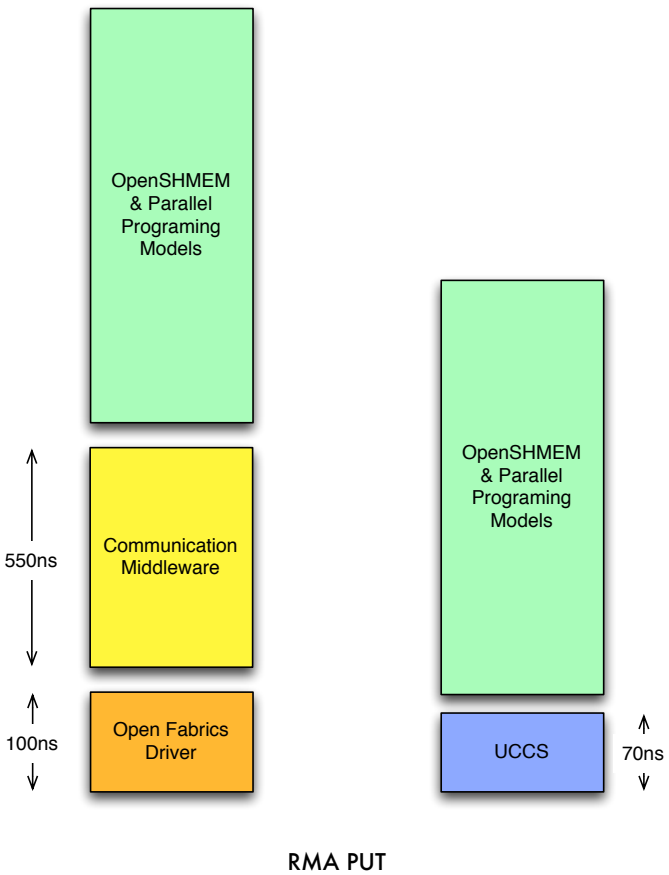


Mellanox InfiniBand	UCCS	OpenSHMEM
RDMA	V	V
AMO	V	V
Collectives/CORE-Direct	V	V

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

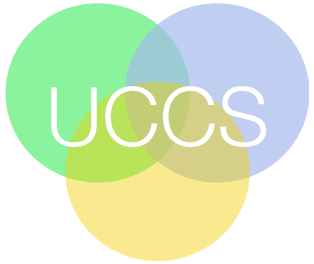
InfiniBand - Continued

- UCCS implements experimental user-level VERBS bypass mode for Mellanox Connect-X devices
  - Direct HAL access
    - Doorbell on PCIex from UCCS
  - Avoiding replicated decision logic
- We have no goal to replace OFED !
  - Investigation into software overheads
  - A standalone ULP, similar to RDS, SDP, etc.



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

Multithreading and Fault Tolerance in UCCS

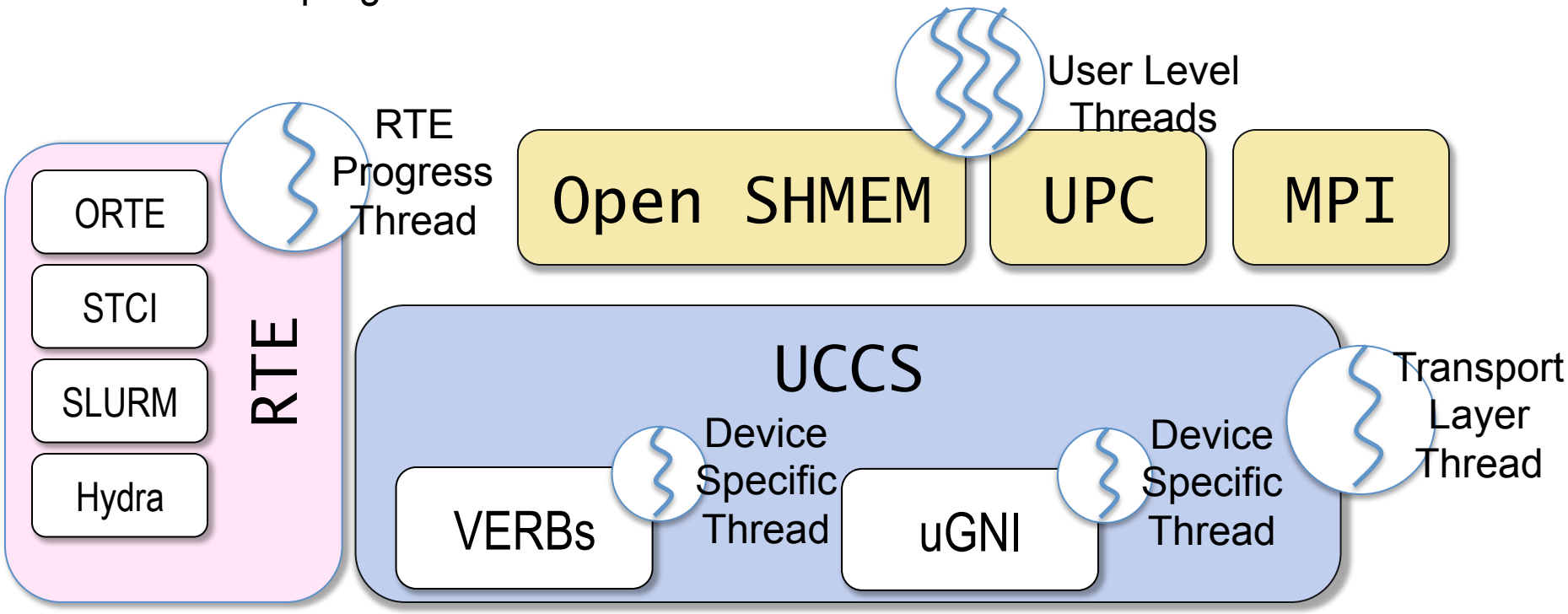


Universal Common Communication Substrate

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

Multithreading Support

- Multiple user level threads
  - Simultaneous `uccs_progress / uccs_*` calls
  - Can execute USER or NETWORK active message callbacks
  - Provides progress for all communications

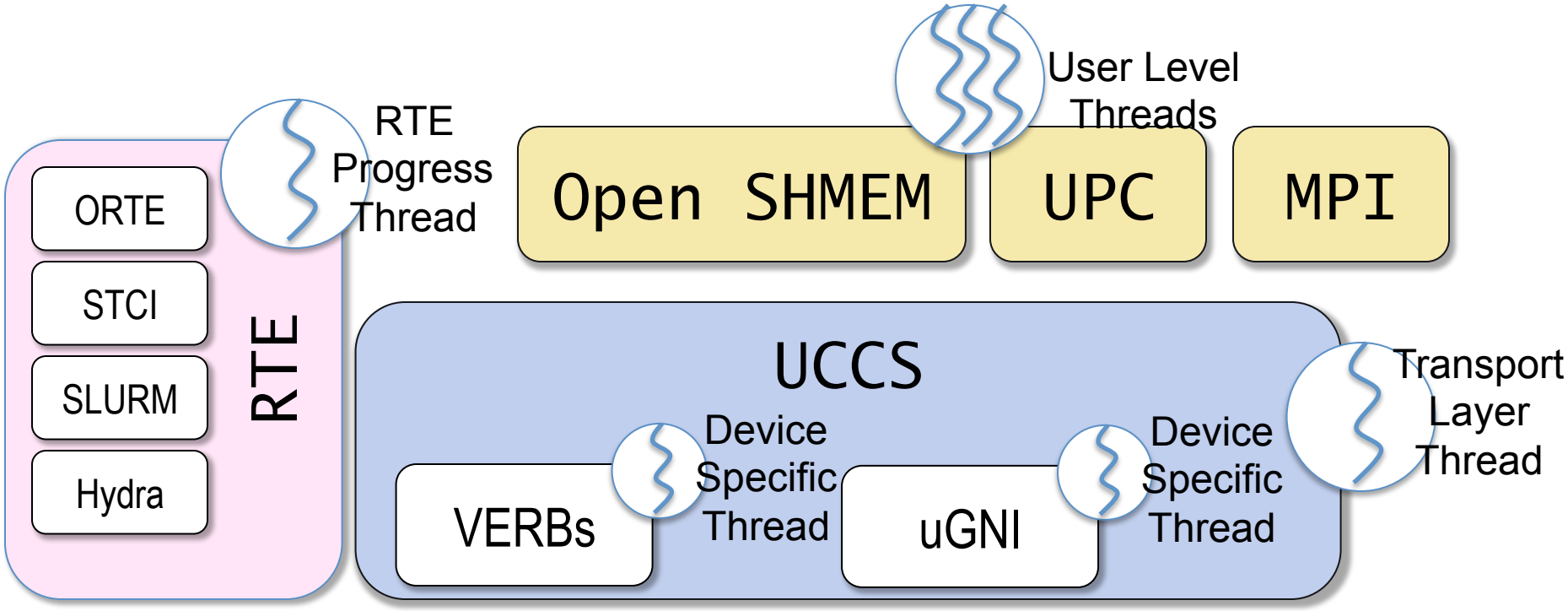


OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Multithreading Support

Creation controlled through MCA parameter

- Device Specific Threads
  - Goal: avoid active polling loop on each device
  - Blocking call to detect network activity / availability ➔ Device Thread
    - (devices that share a `select/poll` interface may share the device thread)

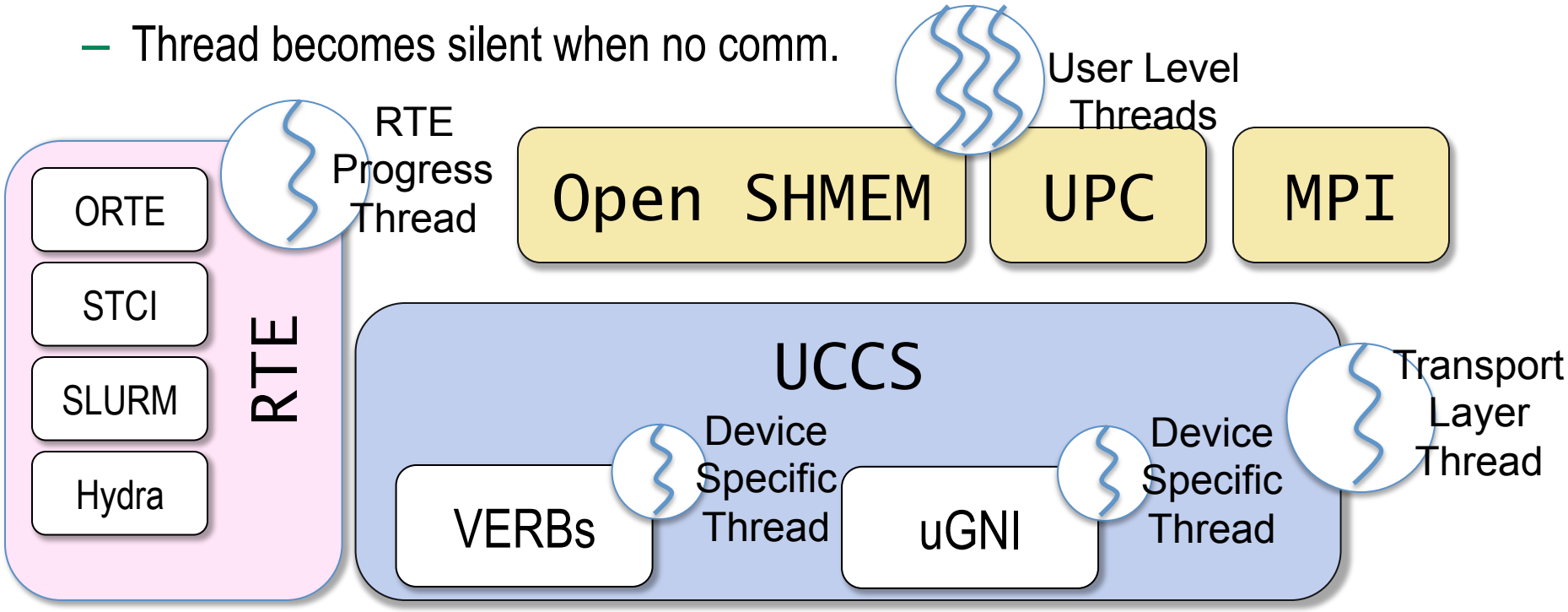


OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Multithreading Support

Creation controlled through MCA parameter

- Device Specific Threads
  - Can execute NETWORK AM callbacks (internal)
    - (may delegate USER AM callbacks to the UCCS Library thread)
  - Device-specific asynchronous progress
  - Thread becomes silent when no comm.

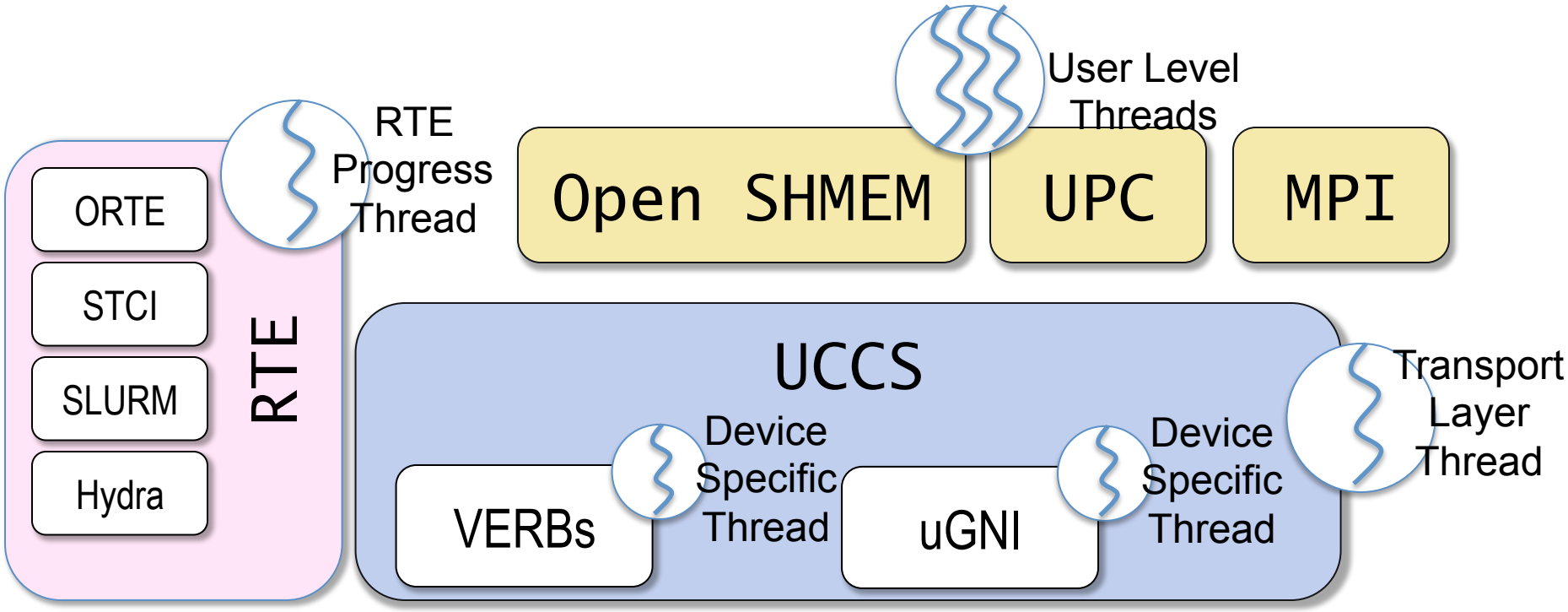


OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Multithreading Support

- Transport Layer Thread
  - Serves as delegate for USER callbacks When triggered by a Device Specific Thread

Creation controlled through MCA parameter

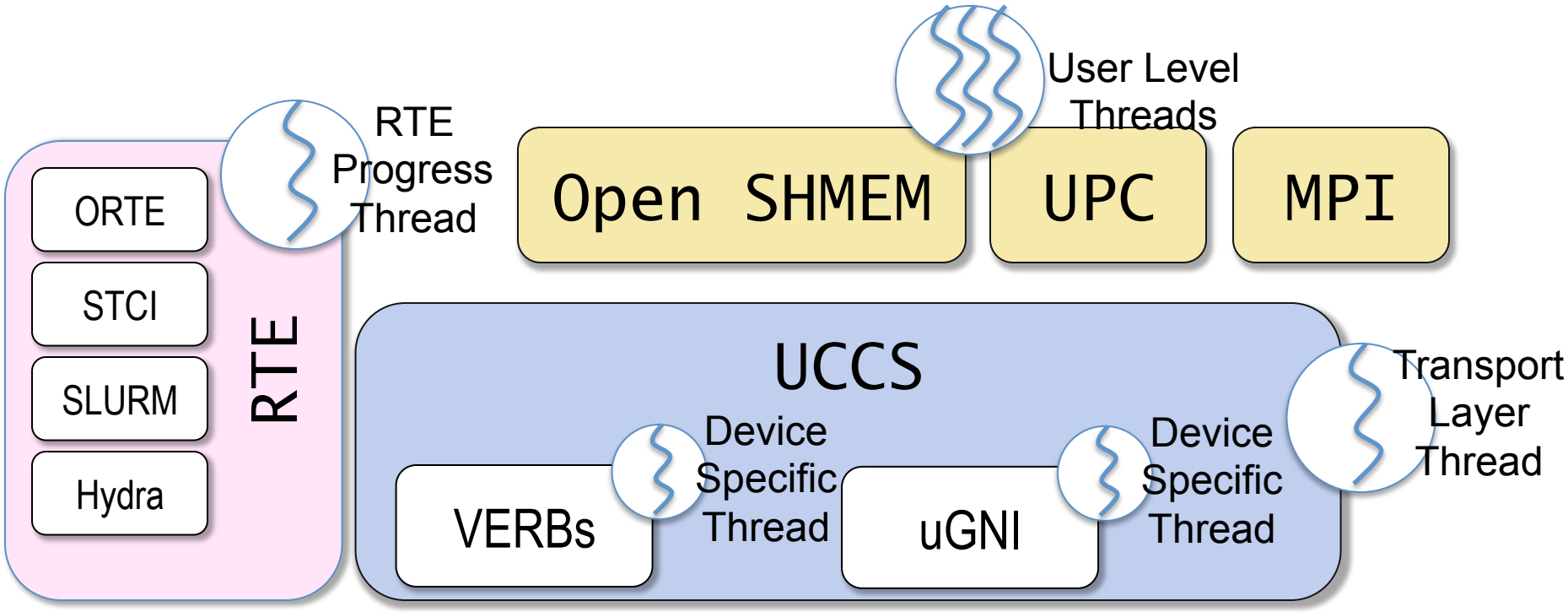


OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE U.S. DEPARTMENT OF ENERGY

Multithreading Support

Creation controlled through MCA parameter

- RTE Progress Thread
  - allows asynchronous progress of the runtime environment
  - Mostly silent thread



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

Fault-Tolerance Extension

- Goal: allow to write libraries / applications that tolerate failures
  - Focus: permanent process crash (fail-stop)
  - Low/no overhead approach when there is no failure
    - e.g. no forced strong synchronization of collectives

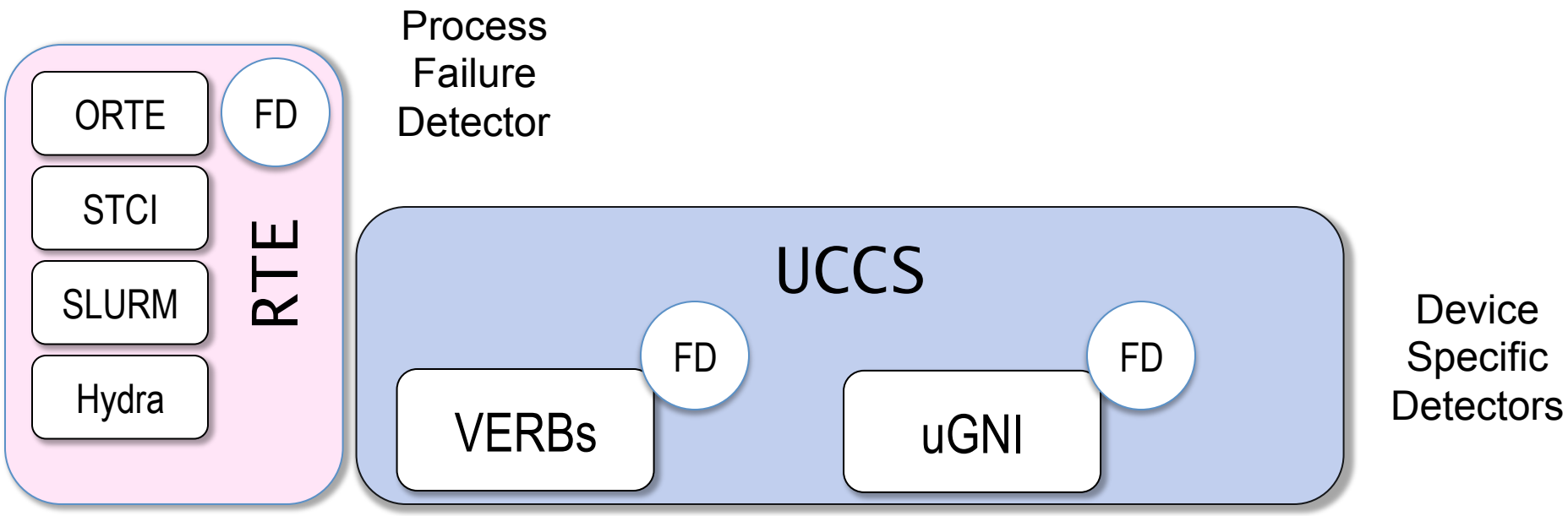
      ➔no guarantee of consistent return code for collectives if a failure hit a processor

- State of UCCS library is well defined after a failure
- Active Message integrity is preserved
- RMA operations can be interrupted by failures
  - Exposed memory may or may not have been modified by call

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

Fault-Tolerance Extension: Failure Detection

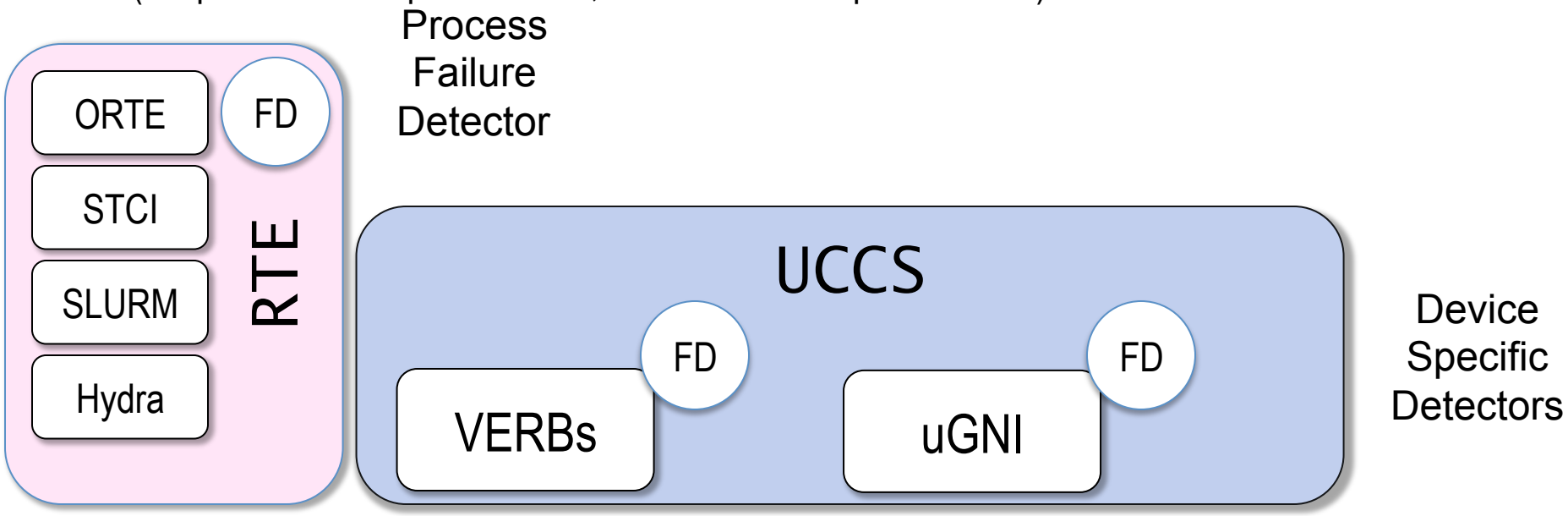
- Failure Detection is done
  - inside the UCCS library through device-specific failure detectors
    - Broken connections
    - Switch/Router feedback information (down links)
    - Low-level heartbit messages (e.g. ICMP)



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

Fault-Tolerance Extension: Failure Detection

- Failure Detection is done
  - At the runtime level: heartbeat protocol in the out-of-band network
- Local reconciliation of failure reporting
- Definitive decisions
  - (if a process is suspected dead, it will remain suspected dead)



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

Fault-Tolerance Extension: Failure Notification

- Detected failures are eventually globally notified to all UCCS processors
  - E.g. through the OOB network, by the RTE, asynchronously
  - Group Membership Service

- UCCS processes can return error codes when a call involves any failed process
  - if the operation cannot complete without participation of the dead process, an error is returned
  - if the operation can locally complete without the participation of the dead process, no error may be returned

- Active Messages:
  - an error handler can be registered to be called when a process death is detected
  - Coupled with GMS: eventual agreement on the list of failed processes

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

We are open for collaboration

http://uccs.github.io/uccs

uccs-info@ornl.gov

UNIVERSITY of HOUSTON

OAK RIDGE National Laboratory

INNOVATIVE COMPUTING LABORATORY

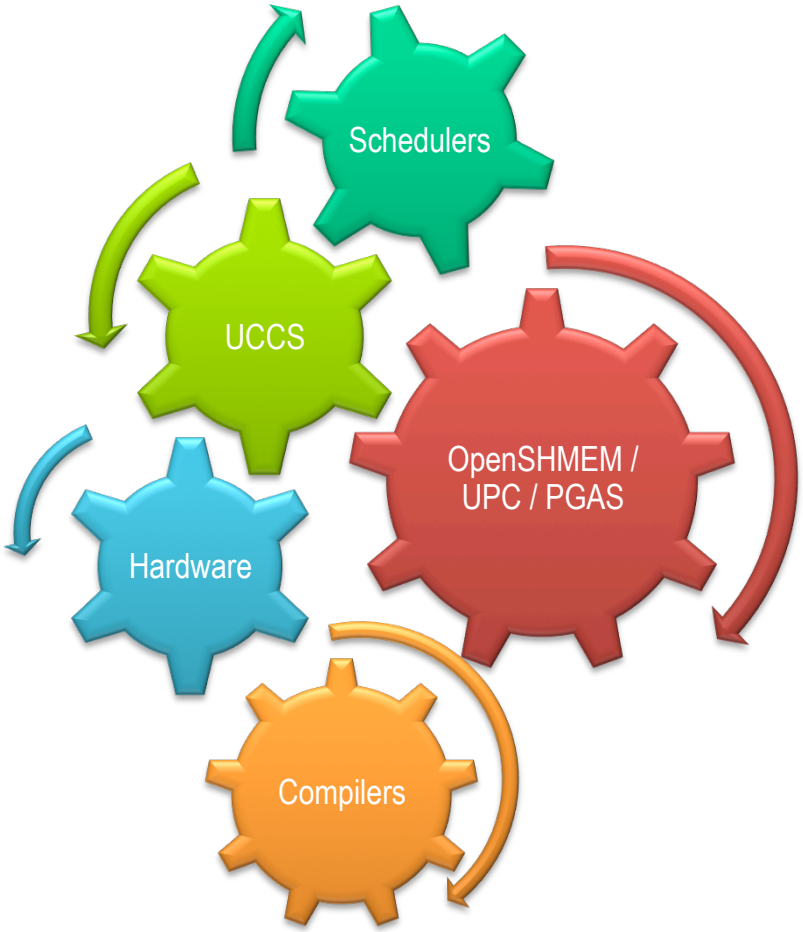
DEPARTMENT OF DEFENSE UNITED STATES OF AMERICA

Mellanox® TECHNOLOGIES

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