Unified Communication X (UCX)

API Standard

Version 1.5

Contents

Chapter 1

Preface

1.1 Scope of the Document

This document describes the UCX programming interface. The programming interface exposes a high performance communication API, which provides basic building blocks for PGAS, Message Passing Interface (MPI), Big-Data, Analytics, File I/O, and storage library developers.

1.2 Audience

This manual is intended for programmers who want to develop parallel programming models like OpenSHMEM, MPI, UPC, Chapel, etc. The manual assumes that the reader is familiar with the following:

- Basic concepts of two-sided, one-sided, atomic, and collective operations
- C programming language

1.3 Document Status

This section briefly describes a list of open issues in the UCX specification.

- UCP API - work in progress
- UCT API - work in progress

1.4 License

UCX project follows open source development model and the software is licensed under BSD-3 license.

Chapter 2

Introduction

2.1 Motivation

A communication middleware abstracts the vendor-specific software and hardware interfaces. They bridge the semantic and functionality gap between the programming models and the software and hardware network interfaces by providing data transfer interfaces and implementation, optimized protocols for data transfer between various memories, and managing network resources. There are many communication middleware APIs and libraries to support parallel programming models such as MPI, OpenSHMEM, and task-based models.

Current communication middleware designs typically take two approaches. First, communication middleware such as Intel's PSM (previously Qlogic), Mellanox's MXM, and IBM's PAMI provide high-performance implementations for specific network hardware. Second, communication middleware such as VMI, Cactus, ARMCI, GASNet, and Open MPI are tightly coupled to a specific programming model. Communication middleware designed with either of this design approach requires significant porting effort to move a new network interface or programming model.

To achieve functional and performance portability across architectures and programming models, we introduce Unified Communication X (UCX).

2.2 UCX

Unified Communication X (UCX) is a set of network APIs and their implementations for high throughput computing. UCX is a combined effort of national laboratories, industry, and academia to design and implement a high-performing and highly-scalable network stack for next generation applications and systems. UCX design provides the ability to tailor its APIs and network functionality to suit a wide variety of application domains. We envision that these APIs will satisfy the networking needs of many programming models such as the Message Passing Interface (MPI), OpenSHMEM, Partitioned Global Address Space (PGAS) languages, task-based paradigms, and I/O bound applications.

The initial focus is on supporting semantics such as point-to-point communications (one-sided and two-sided), collective communication, and remote atomic operations required for popular parallel programming models. Also, the initial UCX reference implementation is targeted to support current network technologies such as:

- Open Fabrics - InfiniBand (Mellanox, Qlogic, IBM), libfabrics, iWARP, RoCE
- Cray GEMINI & ARIES
- Shared memory (MMAP, Posix, CMA, KNEM, XPMEM, etc.)
- Ethernet (TCP/UDP)

UCX design goals are focused on performance and scalability, while efficiently supporting popular and emerging programming models.

UCX's API and design do not impose architectural constraints on the network hardware nor require any specific capabilities to the support the programming model functionality. This is achieved by keeping the API flexible and ability to support the missing functionality efficiently in the software.

Extreme scalability is an important design goal for UCX. To achieve this, UCX follows these design principles:

- Minimal memory consumption : Design avoids data-structures that scale with the number of processing elements (i.e., order N data structures), and share resources among multiple programming models.
- Low-latency Interfaces: Design provides at least two sets of APIs with one set focused on the performance, and the other focused on functionality.
- High bandwidth - With minimal software overhead combined and support for multi-rail and multi-device capabilities, the design provides all the hooks that are necessary for exploiting hardware bandwidth capabilities.
- Asynchronous Progress: API provides non-blocking communication interfaces and design supports asynchronous progress required for communication and computation overlap
- Resilience - the API exposes communication control hooks required for fault tolerant communication library implementation.

UCX design provides native support for hybrid programming models. The design enables resource sharing, optimal memory usage, and progress engine coordination to efficiently implement hybrid programming models. For example, hybrid applications that use both OpenSHMEM and MPI programming models will be able to select between a single-shared UCX network context or a stand alone UCX network context for each one of them. Such flexibility, optimized resource sharing, and reduced memory consumption, improve network and application performance.

Chapter 3

Design

The UCX framework consists of the three main components: UC-Services (UCS), UC-Transports (UCT), and UC-Protocols (UCP). Each one of these components exports a public API, and can be used as a stand-alone library.

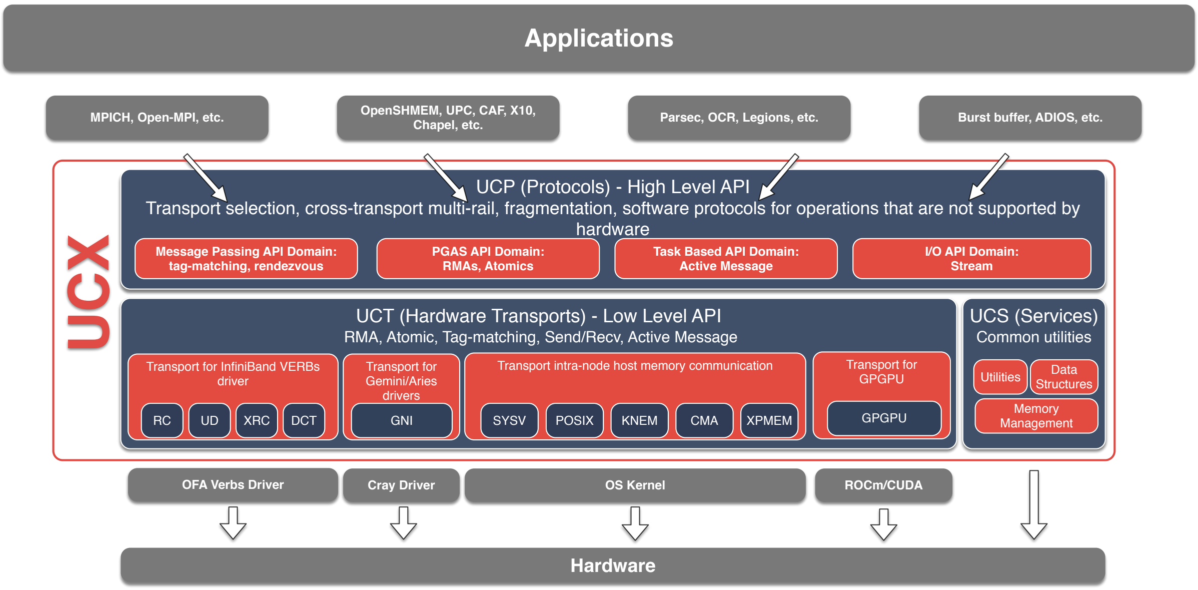

Figure 3.1: UCX Framework Architecture

3.1 UCS

UCS is a service layer that provides the necessary functionality for implementing portable and efficient utilities. This layer includes the following services:

- an abstraction for accessing platform specific functionality (atomic operations, thread safety, etc.),
- tools for efficient memory management (memory pools, memory allocators, and memory allocators hooks),
- commonly used data structures (hashes, trees, lists).

3.2 UCT

UCT is a transport layer that abstracts the differences across various hardware architectures and provides a low-level API that enables the implementation of communication protocols. The primary goal of the layer is to provide direct and efficient access to hardware network functionality. For this purpose, UCT relies on vendor provided low-level drivers such as InfiniBand Verbs, Cray's uGNI, libfabrics, etc. In addition, the layer provides constructs for communication context management (thread-based and application level), and allocation and management of device-specific memories including those found in accelerators. In terms of communication APIs, UCT defines interfaces for immediate (short), buffered copy-and-send (bcopy), and zero-copy (zcopy) communication operations.

Short: This type of operation is optimized for small messages that can be posted and completed in place.

Bcopy: This type of operation is optimized for medium size messages that are typically sent through a so-called bouncing-buffer. This auxiliary buffer is typically allocated given network constraints and ready for immediate utilization by the hardware. Since a custom data packing routine could be provided, this method can be used for non-contiguos i/o.

Zcopy: This type of operation exposes zero-copy memory-to-memory communication semantics, which means that message is sent directly from user buffer, or received directly to user buffer, without being copied between the network layers.

3.3 UCP

UCP implements higher-level protocols that are typically used by message passing (MPI) and PGAS programming models by using lower-level capabilities exposed through the UCT layer. UCP is provides the following functionality: ability to select different transports for communication, message fragmentation, multi-rail communication, and initializing and finalizing the library. Currently, the API has the following classes of interfaces: Initialization, Remote Memory Access (RMA) communication, Atomic Memory Operations (AMO), Active Message, Tag-Matching, and Collectives.

Initialization: This subset of interfaces defines the communication context setup, queries the network capabilities, and initializes the local communication endpoints. The context represented by the UCX context is an abstraction of the network transport resources. The communication endpoint setup interfaces initialize the UCP endpoint, which is an abstraction of all the necessary resources associated with a particular connection. The communication endpoints are used as input to all communication operations to describe the source and destination of the communication.

RMA: This subset of interfaces defines one-sided communication operations such as PUT and GET, required for implementing low overhead, direct memory access communications constructs needed by both distributed and shared memory programming models. UCP includes a separate set of interfaces for communicating non-contiguous data. This functionality was included to support various programming models' communication requirements and leverage the scatter/gather capabilities of modern network hardware.

AMO: This subset of interfaces provides support for atomically performing operations on the remote memory, an important class of operations for PGAS programming models, particularly OpenSHMEM.

Tag Matching: This interface supports tag-matching for send-receive semantics which is a key communication semantic defined by the MPI specification.

Stream : The API provides order and reliable communication semantics. Data is treated as an ordered sequence of bytes pushed through the connection. In contrast of tag-matching interface, the size of each individual send does not necessarily have to match the size of each individual receive, as long as the total number of bytes is the same. This API is designed to match widely used BSD-socket based programming models.

Active Message: A subset of functionality where the incoming packet invokes a sender-specified callback in order to be processed by the receiving process. As an example, the two-sided MPI interface can easily be implemented on top of such a concept (TBD: cite openmpi). However, these interfaces are more general and suited for other programming paradigms where the receiver process does not prepost receives, but expects to react to incoming packets directly. Like RMA and tag-matching interfaces, the active message interface provides separate APIs for different message types and non-contiguous data.

Collectives: This subset of interfaces defines group com- munication and synchronization operations. The collective operations include Barrier, All-to-one, All-to-all, and reduction operations. When possible, we will take advantage of hardware acceleration for collectives (e.g., InfiniBand Switch collective acceleration).

Chapter 4

Conventions and Notations

This section describes the conventions and notations in the UCX specification.

4.1 Blocking Behavior

The blocking UCX routines return only when an UCX operation is complete. After the return, the resources used in the UCX routine are available for reuse.

4.2 Non-blocking Behavior

The non-blocking UCX routines return immediately, independent of operation completion. After the return, the resources used for the routines are not necessarily available for reuse.

4.3 Fairness

UCX routines do not guarantee fairness. However, the routines enable UCX consumers to write efficient and fair programs.

4.4 with Signal Handler Functions

If UCX routines are invoked from signal a handler function, the behavior of the program is undefined.

Chapter 5

Deprecated List

globalScope> Global ucp_atomic_add32 (ucp_ep_h ep, uint32_t add, uint64_t remote_addr, ucp_rkey_h rkey)

Replaced by ucp_atomic_post with opcode UCP_ATOMIC_POST_OP_ADD.

See also

ucp_put.

globalScope> Global ucp_atomic_add64 (ucp_ep_h ep, uint64_t add, uint64_t remote_addr, ucp_rkey_h rkey)

Replaced by ucp_atomic_post with opcode UCP_ATOMIC_POST_OP_ADD.

See also

ucp_put.

globalScope> Global ucp_atomic_cswap32 (ucp_ep_h ep, uint32_t compare, uint32_t swap, uint64_t remote_addr, ucp_rkey_h rkey, uint32_t ∗result)

Replaced by ucp_atomic_fetch_nb with opcode UCP_ATOMIC_FETCH_OP_CSWAP.

See also

ucp_put.

globalScope> Global ucp_atomic_cswap64 (ucp_ep_h ep, uint64_t compare, uint64_t swap, uint64_t remote_addr, ucp_rkey_h rkey, uint64_t ∗result)

Replaced by ucp_atomic_fetch_nb with opcode UCP_ATOMIC_FETCH_OP_CSWAP.

See also

ucp_put.

globalScope> Global ucp_atomic_fadd32 (ucp_ep_h ep, uint32_t add, uint64_t remote_addr, ucp_rkey_h rkey, uint32_t ∗result)

Replaced by ucp_atomic_fetch_nb with opcode UCP_ATOMIC_FETCH_OP_FADD.

See also

ucp_put.

globalScope> Global ucp_atomic_fadd64 (ucp_ep_h ep, uint64_t add, uint64_t remote_addr, ucp_rkey_h rkey, uint64_t ∗result)

Replaced by ucp_atomic_fetch_nb with opcode UCP_ATOMIC_FETCH_OP_FADD.

See also

ucp_put.

globalScope> Global ucp_atomic_swap32 (ucp_ep_h ep, uint32_t swap, uint64_t remote_addr, ucp_rkey_h rkey, uint32_t *result)

Replaced by ucp_atomic_fetch_nb with opcode UCP_ATOMIC_FETCH_OP_SWAP.

See also

ucp_put.

globalScope> Global ucp_atomic_swap64 (ucp_ep_h ep, uint64_t swap, uint64_t remote_addr, ucp_rkey_h rkey, uint64_t *result)

Replaced by ucp_atomic_fetch_nb with opcode UCP_ATOMIC_FETCH_OP_SWAP.

See also

ucp_put.

globalScope> Global ucp_disconnect_nb (ucp_ep_h ep)

Replaced by ucp_ep_close_nb.

globalScope> Global ucp_ep_destroy (ucp_ep_h ep)

Replaced by ucp_ep_close_nb.

globalScope> Global ucp_ep_flush (ucp_ep_h ep)

Replaced by ucp_ep_flush_nb.

globalScope> Global ucp_ep_modify_nb (ucp_ep_h ep, const ucp_ep_params_t *params)

Use ucp_listener_conn_handler_t instead of ucp_listener_accept_handler_t, if you have other use case please submit an issue on `https://github.com/openucx/ucx` or report to `ucx-group@elist.ornl.gov`

globalScope> Global ucp_get (ucp_ep_h ep, void *buffer, size_t length, uint64_t remote_addr, ucp_rkey_h rkey)

Replaced by ucp_get_nb.

See also

ucp_put.

globalScope> Global ucp_listener_accept_handler_t

Replaced by ucp_listener_conn_handler_t.

globalScope> Global ucp_listener_accept_handler_t

Replaced by ucp_listener_conn_handler_t.

globalScope> Global ucp_put (ucp_ep_h ep, const void *buffer, size_t length, uint64_t remote_addr, ucp_rkey_h rkey)

Replaced by ucp_put_nb. The following example implements the same functionality using ucp_put_nb :

```
void empty_callback(void *request, ucs_status_t status)
{
}

ucs_status_t put(ucp_ep_h ep, const void *buffer, size_t length,
                  uint64_t remote_addr, ucp_rkey_h rkey)
{
    void *request = ucp_put_nb(ep, buffer, length, remote_addr, rkey,
                               empty_callback),
    if (request == NULL) {
        return UCS_OK;
    } else if (UCS_PTR_IS_ERR(request)) {
        return UCS_PTR_STATUS(request);
    } else {
        ucs_status_t status;
        do {
            ucp_worker_progress(worker);
            status = ucp_request_check_status(request);
        } while (status == UCS_INPROGRESS);
        ucp_request_release(request);
        return status;
    }
}
```

globalScope> Global ucp_request_is_completed (void *request)

Replaced by ucp_request_test.

globalScope> Global ucp_request_release (void *request)

Replaced by ucp_request_free.

globalScope> Global ucp_request_test (void *request, ucp_tag_recv_info_t *info)

Replaced by ucp_tag_recv_request_test and ucp_request_check_status depends on use case.

globalScope> Global ucp_worker_flush (ucp_worker_h worker)

Replaced by ucp_worker_flush_nb. The following example implements the same functionality using ucp_worker_flush_nb :

```
ucs_status_t worker_flush(ucp_worker_h worker)
{
    void *request = ucp_worker_flush_nb(worker);
    if (request == NULL) {
        return UCS_OK;
    } else if (UCS_PTR_IS_ERR(request)) {
        return UCS_PTR_STATUS(request);
    } else {
        ucs_status_t status;
         do {
             ucp_worker_progress(worker);
             status = ucp_request_check_status(request);
         } while (status == UCS_INPROGRESS);
         ucp_request_release(request);
         return status;
     }
}
```

Chapter 6

Module Documentation

6.1 Unified Communication Protocol (UCP) API

Modules

- UCP Application Context
- UCP Worker
- UCP Memory routines
- UCP Wake-up routines
- UCP Endpoint
- UCP Communication routines
- UCP Configuration
- UCP Data type routines

6.1.1 Detailed Description

This section describes UCP API.

6.2 UCP Application Context

Data Structures

- struct ucp_context_attr

    Context attributes. More...
- struct ucp_tag_recv_info

    UCP receive information descriptor. More...

Typedefs

- typedef struct ucp_context_attr ucp_context_attr_t

    Context attributes.
- typedef struct ucp_tag_recv_info ucp_tag_recv_info_t

    UCP receive information descriptor.
- typedef struct ucp_context ∗ ucp_context_h

    UCP Application Context.
- typedef void(∗ ucp_request_init_callback_t) (void ∗request)

    Request initialization callback.
- typedef void(∗ ucp_request_cleanup_callback_t) (void ∗request)

    Request cleanup callback.

Enumerations

- enum ucp_params_field {
  UCP_PARAM_FIELD_FEATURES = UCS_BIT(0), UCP_PARAM_FIELD_REQUEST_SIZE = UCS_BIT(1), UCP_PARAM_FIELD_REQUEST_INIT = UCS_BIT(2), UCP_PARAM_FIELD_REQUEST_CLEANUP = U↩CS_BIT(3),
  UCP_PARAM_FIELD_TAG_SENDER_MASK = UCS_BIT(4), UCP_PARAM_FIELD_MT_WORKERS_SH↩ARED = UCS_BIT(5), UCP_PARAM_FIELD_ESTIMATED_NUM_EPS = UCS_BIT(6) }

    UCP context parameters field mask.
- enum ucp_feature {
  UCP_FEATURE_TAG = UCS_BIT(0), UCP_FEATURE_RMA = UCS_BIT(1), UCP_FEATURE_AMO32 = UCS_BIT(2), UCP_FEATURE_AMO64 = UCS_BIT(3),
  UCP_FEATURE_WAKEUP = UCS_BIT(4), UCP_FEATURE_STREAM = UCS_BIT(5) }

    UCP configuration features.
- enum ucp_context_attr_field { UCP_ATTR_FIELD_REQUEST_SIZE = UCS_BIT(0), UCP_ATTR_FIELD_↩THREAD_MODE = UCS_BIT(1) }

    UCP context attributes field mask.

Functions

- void ucp_get_version (unsigned ∗major_version, unsigned ∗minor_version, unsigned ∗release_number)

    Get UCP library version.
- const char ∗ ucp_get_version_string (void)

    Get UCP library version as a string.
- static ucs_status_t ucp_init (const ucp_params_t ∗params, const ucp_config_t ∗config, ucp_context_↩h ∗context_p)

    UCP context initialization.
- void ucp_cleanup (ucp_context_h context_p)

    Release UCP application context.

- ucs_status_t ucp_context_query (ucp_context_h context_p, ucp_context_attr_t ∗attr)

    Get attributes specific to a particular context.
- void ucp_context_print_info (ucp_context_h context, FILE ∗stream)

    Print context information.

6.2.1 Detailed Description

Application context is a primary concept of UCP design which provides an isolation mechanism, allowing resources associated with the context to separate or share network communication context across multiple instances of applications.

This section provides a detailed description of this concept and routines associated with it.

6.2.2 Data Structure Documentation

6.2.2.1 struct ucp_context_attr

The structure defines the attributes which characterize the particular context.

Data Fields

uint64_t	field_mask	Mask of valid fields in this structure, using bits from ucp_context_attr↩_field. Fields not specified in this mask will be ignored. Provides ABI compatibility with respect to adding new fields.
size_t	request_size	Size of UCP non-blocking request. When pre-allocated request is used (e.g. in ucp_tag_recv_nbr) it should have enough space to fit UCP request data, which is defined by this value.
ucs_thread_↩mode_t	thread_mode	Thread safe level of the context. For supported thread levels please see ucs_thread_mode_t.

6.2.2.2 struct ucp_tag_recv_info

The UCP receive information descriptor is allocated by application and filled in with the information about the received message by ucp_tag_probe_nb or ucp_tag_recv_request_test routines or ucp_tag_recv_callback_t callback argument.

Examples:

ucp_hello_world.c.

Data Fields

ucp_tag_t	sender_tag	Sender tag
size_t	length	The size of the received data

6.2.3 Typedef Documentation

6.2.3.1 typedef struct ucp_context_attr ucp_context_attr_t

The structure defines the attributes which characterize the particular context.

6.2.3.2 typedef struct ucp_tag_recv_info ucp_tag_recv_info_t

The UCP receive information descriptor is allocated by application and filled in with the information about the received message by ucp_tag_probe_nb or ucp_tag_recv_request_test routines or ucp_tag_recv_callback_t callback argument.

6.2.3.3 typedef struct ucp_context∗ ucp_context_h

UCP application context (or just a context) is an opaque handle that holds a UCP communication instance's global information. It represents a single UCP communication instance. The communication instance could be an OS process (an application) that uses UCP library. This global information includes communication resources, endpoints, memory, temporary file storage, and other communication information directly associated with a specific UCP instance. The context also acts as an isolation mechanism, allowing resources associated with the context to manage multiple concurrent communication instances. For example, users using both MPI and OpenSHMEM sessions simultaneously can isolate their communication by allocating and using separate contexts for each of them. Alternatively, users can share the communication resources (memory, network resource context, etc.) between them by using the same application context. A message sent or a RMA operation performed in one application context cannot be received in any other application context.

6.2.3.4 typedef void(∗ ucp_request_init_callback_t) (void ∗request)

This callback routine is responsible for the request initialization.

Parameters

`in`	*request*	Request handle to initialize.

6.2.3.5 typedef void(∗ ucp_request_cleanup_callback_t) (void ∗request)

This callback routine is responsible for cleanup of the memory associated with the request.

Parameters

`in`	*request*	Request handle to cleanup.

6.2.4 Enumeration Type Documentation

6.2.4.1 enum ucp_params_field

The enumeration allows specifying which fields in ucp_params_t are present. It is used for the enablement of backward compatibility support.

Enumerator

UCP_PARAM_FIELD_FEATURES features

UCP_PARAM_FIELD_REQUEST_SIZE request_size

UCP_PARAM_FIELD_REQUEST_INIT request_init

UCP_PARAM_FIELD_REQUEST_CLEANUP request_cleanup

UCP_PARAM_FIELD_TAG_SENDER_MASK tag_sender_mask

UCP_PARAM_FIELD_MT_WORKERS_SHARED mt_workers_shared

UCP_PARAM_FIELD_ESTIMATED_NUM_EPS estimated_num_eps

6.2.4.2 enum ucp_feature

The enumeration list describes the features supported by UCP. An application can request the features using UCP parameters during UCP initialization process.

Enumerator

UCP_FEATURE_TAG Request tag matching support
UCP_FEATURE_RMA Request remote memory access support
UCP_FEATURE_AMO32 Request 32-bit atomic operations support
UCP_FEATURE_AMO64 Request 64-bit atomic operations support
UCP_FEATURE_WAKEUP Request interrupt notification support
UCP_FEATURE_STREAM Request stream support

6.2.4.3 enum ucp_context_attr_field

The enumeration allows specifying which fields in ucp_context_attr_t are present. It is used for the enablement of backward compatibility support.

Enumerator

UCP_ATTR_FIELD_REQUEST_SIZE UCP request size
UCP_ATTR_FIELD_THREAD_MODE UCP context thread flag

6.2.5 Function Documentation

6.2.5.1 void ucp_get_version (unsigned * *major_version,* unsigned * *minor_version,* unsigned * *release_number*)

This routine returns the UCP library version.

Parameters

out	*major_version*	Filled with library major version.
out	*minor_version*	Filled with library minor version.
out	*release_number*	Filled with library release number.

6.2.5.2 const char* ucp_get_version_string (void)

This routine returns the UCP library version as a string which consists of: "major.minor.release".

6.2.5.3 static ucs_status_t ucp_init (const ucp_params_t * *params,* const ucp_config_t * *config,* ucp_context_h * *context_p*) [inline], [static]

This routine creates and initializes a UCP application context.

Warning

This routine must be called before any other UCP function call in the application.

This routine checks API version compatibility, then discovers the available network interfaces, and initializes the network resources required for discovering of the network and memory related devices. This routine is responsible for initialization all information required for a particular application scope, for example, MPI application, OpenSH↵ MEM application, etc.

Note

- Higher level protocols can add additional communication isolation, as MPI does with it's communicator object. A single communication context may be used to support multiple MPI communicators.
- The context can be used to isolate the communication that corresponds to different protocols. For example, if MPI and OpenSHMEM are using UCP to isolate the MPI communication from the OpenSHMEM communication, users should use different application context for each of the communication libraries.

Parameters

in	*config*	UCP configuration descriptor allocated through ucp_config_read() routine.
in	*params*	User defined ucp_params_t configurations for the UCP application context.
out	*context_p*	Initialized UCP application context.

Returns

Error code as defined by ucs_status_t

Examples:

ucp_hello_world.c.

6.2.5.4 void ucp_cleanup (ucp_context_h *context_p*)

This routine finalizes and releases the resources associated with a UCP application context.

Warning

An application cannot call any UCP routine once the UCP application context released.

The cleanup process releases and shuts down all resources associated with the application context. After calling this routine, calling any UCP routine without calling UCP initialization routine is invalid.

Parameters

in	*context_p*	Handle to UCP application context.

Examples:

ucp_hello_world.c.

6.2.5.5 ucs_status_t ucp_context_query (ucp_context_h *context_p*, ucp_context_attr_t * *attr*)

This routine fetches an information about the context.

Parameters

in	*context_p*	Handle to UCP application context.
out	*attr*	Filled with attributes of `context_p` context.

Returns

Error code as defined by ucs_status_t

6.2.5.6 void ucp_context_print_info (ucp_context_h *context*, FILE * *stream*)

This routine prints information about the context configuration, including memory domains, transport resources, and other useful information associated with the context.

Parameters

in	*context*	Context object whose configuration to print.
in	*stream*	Output stream to print the information to.

6.3 UCP Worker

Data Structures

- struct ucp_worker_attr

  UCP worker attributes. More...
- struct ucp_worker_params

  Tuning parameters for the UCP worker. More...
- struct ucp_listener_params

  Parameters for a UCP listener object. More...
- struct ucp_listener_accept_handler
- struct ucp_listener_conn_handler

  UCP callback to handle the connection request in a client-server connection establishment flow. More...

Typedefs

- typedef struct ucp_worker_attr ucp_worker_attr_t

  UCP worker attributes.
- typedef struct ucp_worker_params ucp_worker_params_t

  Tuning parameters for the UCP worker.
- typedef struct ucp_listener_params ucp_listener_params_t

  Parameters for a UCP listener object.
- typedef BEGIN_C_DECLS struct ucp_listener_accept_handler ucp_listener_accept_handler_t
- typedef struct ucp_address ucp_address_t

  UCP worker address.
- typedef struct ucp_listener ∗ ucp_listener_h

  UCP listen handle.
- typedef struct ucp_worker ∗ ucp_worker_h

  UCP Worker.
- typedef void(∗ ucp_listener_accept_callback_t) (ucp_ep_h ep, void ∗arg)

  A callback for accepting client/server connections on a listener ucp_listener_h.
- typedef void(∗ ucp_listener_conn_callback_t) (ucp_conn_request_h conn_request, void ∗arg)

  A callback for handling of incoming connection request conn_request from a client.
- typedef struct ucp_listener_conn_handler ucp_listener_conn_handler_t

  UCP callback to handle the connection request in a client-server connection establishment flow.
- typedef enum ucp_wakeup_event_types ucp_wakeup_event_t

  UCP worker wakeup events mask.

Enumerations

- enum ucp_worker_params_field {
  UCP_WORKER_PARAM_FIELD_THREAD_MODE = UCS_BIT(0), UCP_WORKER_PARAM_FIELD_CPU_MASK = UCS_BIT(1), UCP_WORKER_PARAM_FIELD_EVENTS = UCS_BIT(2), UCP_WORKER_PARAM_FIELD_USER_DATA = UCS_BIT(3),
  UCP_WORKER_PARAM_FIELD_EVENT_FD = UCS_BIT(4) }

  UCP worker parameters field mask.
- enum ucp_listener_params_field { UCP_LISTENER_PARAM_FIELD_SOCK_ADDR = UCS_BIT(0), UCP_LISTENER_PARAM_FIELD_ACCEPT_HANDLER = UCS_BIT(1), UCP_LISTENER_PARAM_FIELD_CONN_HANDLER = UCS_BIT(2) }

  UCP listener parameters field mask.
- enum ucp_worker_attr_field { UCP_WORKER_ATTR_FIELD_THREAD_MODE = UCS_BIT(0) }

UCP worker attributes field mask.

- enum ucp_wakeup_event_types {
UCP_WAKEUP_RMA = UCS_BIT(0), UCP_WAKEUP_AMO = UCS_BIT(1), UCP_WAKEUP_TAG_SEND = UCS_BIT(2), UCP_WAKEUP_TAG_RECV = UCS_BIT(3),
UCP_WAKEUP_TX = UCS_BIT(10), UCP_WAKEUP_RX = UCS_BIT(11), UCP_WAKEUP_EDGE = UCS↩
_BIT(16) }

  UCP worker wakeup events mask.

Functions

- ucs_status_t ucp_worker_create (ucp_context_h context, const ucp_worker_params_t *params, ucp_↩
worker_h *worker_p)

  Create a worker object.
- void ucp_worker_destroy (ucp_worker_h worker)

  Destroy a worker object.
- ucs_status_t ucp_worker_query (ucp_worker_h worker, ucp_worker_attr_t *attr)

  Get attributes specific to a particular worker.
- void ucp_worker_print_info (ucp_worker_h worker, FILE *stream)

  Print information about the worker.
- ucs_status_t ucp_worker_get_address (ucp_worker_h worker, ucp_address_t **address_p, size_↩
t *address_length_p)

  Get the address of the worker object.
- void ucp_worker_release_address (ucp_worker_h worker, ucp_address_t *address)

  Release an address of the worker object.
- unsigned ucp_worker_progress (ucp_worker_h worker)

  Progress all communications on a specific worker.
- ssize_t ucp_stream_worker_poll (ucp_worker_h worker, ucp_stream_poll_ep_t *poll_eps, size_t max_eps, unsigned flags)

  Poll for endpoints that are ready to consume streaming data.
- ucs_status_t ucp_listener_create (ucp_worker_h worker, const ucp_listener_params_t *params, ucp_↩
listener_h *listener_p)

  Accept connections on a local address of the worker object.
- void ucp_listener_destroy (ucp_listener_h listener)

  Stop accepting connections on a local address of the worker object.
- ucs_status_t ucp_listener_reject (ucp_listener_h listener, ucp_conn_request_h conn_request)

  Reject an incoming connection request.
- ucs_status_t ucp_worker_fence (ucp_worker_h worker)

  Assures ordering between non-blocking operations.
- ucs_status_ptr_t ucp_worker_flush_nb (ucp_worker_h worker, unsigned flags, ucp_send_callback_t cb)

  Flush outstanding AMO and RMA operations on the worker.
- ucs_status_t ucp_worker_flush (ucp_worker_h worker)

  Flush outstanding AMO and RMA operations on the worker.

6.3.1 Detailed Description

UCP Worker routines

6.3.2 Data Structure Documentation

6.3.2.1 struct ucp_worker_attr

The structure defines the attributes which characterize the particular worker.

Data Fields

uint64_t	field_mask	Mask of valid fields in this structure, using bits from ucp_worker_attr_field. Fields not specified in this mask will be ignored. Provides ABI compatibility with respect to adding new fields.
ucs_thread_mode_t	thread_mode	Thread safe level of the worker.

6.3.2.2 struct ucp_worker_params

The structure defines the parameters that are used for the UCP worker tuning during the UCP worker creation.

Examples:

ucp_hello_world.c.

Data Fields

uint64_t	field_mask	Mask of valid fields in this structure, using bits from ucp_worker_params_field. Fields not specified in this mask would be ignored. Provides ABI compatibility with respect to adding new fields.
ucs_thread_mode_t	thread_mode	The parameter thread_mode suggests the thread safety mode which worker and the associated resources should be created with. This is an optional parameter. The default value is UCS_THREAD_MODE_SINGLE and it is used when the value of the parameter is not set. When this parameter along with its corresponding bit in the field_mask - UCP_WORKER_PARAM_FIELD_THREAD_MODE is set, the ucp_worker_create attempts to create worker with this thread mode. The thread mode with which worker is created can differ from the suggested mode. The actual thread mode of the worker should be obtained using the query interface ucp_worker_query.
ucs_cpu_set_t	cpu_mask	Mask of which CPUs worker resources should preferably be allocated on. This value is optional. If it's not set (along with its corresponding bit in the field_mask - UCP_WORKER_PARAM_FIELD_CPU_MASK), resources are allocated according to system's default policy.
unsigned	events	Mask of events (ucp_wakeup_event_t) which are expected on wakeup. This value is optional. If it's not set (along with its corresponding bit in the field_mask - UCP_WORKER_PARAM_FIELD_EVENTS), all types of events will trigger on wakeup.
void *	user_data	User data associated with the current worker. This value is optional. If it's not set (along with its corresponding bit in the field_mask - UCP_WORKER_PARAM_FIELD_USER_DATA), it will default to NULL.
int	event_fd	External event file descriptor. This value is optional. If UCP_WORKER_PARAM_FIELD_EVENT_FD is set in the field_mask, events on the worker will be reported on the provided event file descriptor. In this case, calling ucp_worker_get_efd will result in an error. The provided file descriptor must be capable of aggregating notifications for arbitrary events, for example `epoll(7)` on Linux systems. user_data will be used as the event user-data on systems which support it. For example, on Linux, it will be placed in `epoll_data_t::ptr`, when returned from `epoll_wait(2)`. Otherwise, events would be reported to the event file descriptor returned from ucp_worker_get_efd().

6.3.2.3 struct ucp_listener_params

This structure defines parameters for ucp_listener_create, which is used to listen for incoming client/server connections.

Data Fields

uint64_t	field_mask	Mask of valid fields in this structure, using bits from ucp_listener_params_field. Fields not specified in this mask would be ignored. Provides ABI compatibility with respect to adding new fields.
ucs_sock_addr_t	sockaddr	An address in the form of a sockaddr. This field is mandatory for filling (along with its corresponding bit in the field_mask - UCP_LISTENER_PARAM_FIELD_SOCK_ADDR). The ucp_listener_create routine will return with an error if sockaddr is not specified.
ucp_listener_accept_handler_t	accept_handler	Handler to endpoint creation in a client-server connection flow. In order for the callback inside this handler to be invoked, the UCP_LISTENER_PARAM_FIELD_ACCEPT_HANDLER needs to be set in the field_mask.
ucp_listener_conn_handler_t	conn_handler	Handler of an incoming connection request in a client-server connection flow. In order for the callback inside this handler to be invoked, the UCP_LISTENER_PARAM_FIELD_CONN_HANDLER needs to be set in the field_mask.

6.3.2.4 struct ucp_listener_accept_handler

Deprecated Replaced by ucp_listener_conn_handler_t.

Data Fields

ucp_listener_accept_callback_t	cb	Endpoint creation callback
void *	arg	User defined argument for the callback

6.3.2.5 struct ucp_listener_conn_handler

This structure is used for handling an incoming connection request on the listener. Setting this type of handler allows creating an endpoint on any other worker and not limited to the worker on which the listener was created.

Note

- Other than communication progress routines, it is allowed to call all other communication routines from the callback in the struct.
- The callback is thread safe with respect to the worker it is invoked on.
- It is the user's responsibility to avoid potential dead lock accessing different worker.

Data Fields

ucp_listener_conn_callback_t	cb	Connection request callback
void *	arg	User defined argument for the callback

6.3.3 Typedef Documentation

6.3.3.1 typedef struct ucp_worker_attr ucp_worker_attr_t

The structure defines the attributes which characterize the particular worker.

6.3.3.2 typedef struct ucp_worker_params ucp_worker_params_t

The structure defines the parameters that are used for the UCP worker tuning during the UCP worker creation.

6.3.3.3 typedef struct ucp_listener_params ucp_listener_params_t

This structure defines parameters for ucp_listener_create, which is used to listen for incoming client/server connections.

6.3.3.4 typedef BEGIN_C_DECLS struct ucp_listener_accept_handler ucp_listener_accept_handler_t

Deprecated Replaced by ucp_listener_conn_handler_t.

6.3.3.5 typedef struct ucp_address ucp_address_t

The address handle is an opaque object that is used as an identifier for a worker instance.

6.3.3.6 typedef struct ucp_listener* ucp_listener_h

The listener handle is an opaque object that is used for listening on a specific address and accepting connections from clients.

6.3.3.7 typedef struct ucp_worker* ucp_worker_h

UCP worker is an opaque object representing the communication context. The worker represents an instance of a local communication resource and progress engine associated with it. Progress engine is a construct that is responsible for asynchronous and independent progress of communication directives. The progress engine could be implement in hardware or software. The worker object abstract an instance of network resources such as a host channel adapter port, network interface, or multiple resources such as multiple network interfaces or communication ports. It could also represent virtual communication resources that are defined across multiple devices. Although the worker can represent multiple network resources, it is associated with a single UCX application context. All communication functions require a context to perform the operation on the dedicated hardware resource(s) and an endpoint to address the destination.

Note

Worker are parallel "threading points" that an upper layer may use to optimize concurrent communications.

6.3.3.8 typedef void(* ucp_listener_accept_callback_t) (ucp_ep_h ep, void *arg)

This callback routine is invoked on the server side upon creating a connection to a remote client. The user can pass an argument to this callback. The user is responsible for releasing the *ep* handle using the ucp_ep_destroy() routine.

Parameters

in	*ep*	Handle to a newly created endpoint which is connected to the remote peer which has initiated the connection.

in	*arg*	User's argument for the callback.

6.3.3.9 typedef void(∗ ucp_listener_conn_callback_t) (ucp_conn_request_h conn_request, void ∗arg)

This callback routine is invoked on the server side to handle incoming connections from remote clients. The user can pass an argument to this callback. The *conn_request* handle has to be released, either by ucp_ep_create or ucp_listener_reject routine.

Parameters

in	*conn_request*	Connection request handle.
in	*arg*	User's argument for the callback.

6.3.3.10 typedef struct ucp_listener_conn_handler ucp_listener_conn_handler_t

This structure is used for handling an incoming connection request on the listener. Setting this type of handler allows creating an endpoint on any other worker and not limited to the worker on which the listener was created.

Note

- Other than communication progress routines, it is allowed to call all other communication routines from the callback in the struct.
- The callback is thread safe with respect to the worker it is invoked on.
- It is the user's responsibility to avoid potential dead lock accessing different worker.

6.3.3.11 typedef enum ucp_wakeup_event_types ucp_wakeup_event_t

The enumeration allows specifying which events are expected on wakeup. Empty events are possible for any type of event except for UCP_WAKEUP_TX and UCP_WAKEUP_RX.

Note

Send completions are reported by POLLIN-like events (see poll man page). Since outgoing operations can be initiated at any time, UCP does not generate POLLOUT-like events, although it must be noted that outgoing operations may be queued depending upon resource availability.

6.3.4 Enumeration Type Documentation

6.3.4.1 enum ucp_worker_params_field

The enumeration allows specifying which fields in ucp_worker_params_t are present. It is used for the enablement of backward compatibility support.

Enumerator

UCP_WORKER_PARAM_FIELD_THREAD_MODE UCP thread mode

UCP_WORKER_PARAM_FIELD_CPU_MASK Worker's CPU bitmap

UCP_WORKER_PARAM_FIELD_EVENTS Worker's events bitmap

UCP_WORKER_PARAM_FIELD_USER_DATA User data

UCP_WORKER_PARAM_FIELD_EVENT_FD External event file descriptor

6.3.4.2 enum ucp_listener_params_field

The enumeration allows specifying which fields in ucp_listener_params_t are present. It is used for the enablement of backward compatibility support.

Enumerator

UCP_LISTENER_PARAM_FIELD_SOCK_ADDR Sock address and length.

UCP_LISTENER_PARAM_FIELD_ACCEPT_HANDLER User's callback and argument for handling the creation of an endpoint.User's callback and argument for handling the incoming connection request.

UCP_LISTENER_PARAM_FIELD_CONN_HANDLER

6.3.4.3 enum ucp_worker_attr_field

The enumeration allows specifying which fields in ucp_worker_attr_t are present. It is used for the enablement of backward compatibility support.

Enumerator

UCP_WORKER_ATTR_FIELD_THREAD_MODE UCP thread mode

6.3.4.4 enum ucp_wakeup_event_types

The enumeration allows specifying which events are expected on wakeup. Empty events are possible for any type of event except for UCP_WAKEUP_TX and UCP_WAKEUP_RX.

Note

Send completions are reported by POLLIN-like events (see poll man page). Since outgoing operations can be initiated at any time, UCP does not generate POLLOUT-like events, although it must be noted that outgoing operations may be queued depending upon resource availability.

Enumerator

UCP_WAKEUP_RMA Remote memory access send completion

UCP_WAKEUP_AMO Atomic operation send completion

UCP_WAKEUP_TAG_SEND Tag send completion

UCP_WAKEUP_TAG_RECV Tag receive completion

UCP_WAKEUP_TX This event type will generate an event on completion of any outgoing operation (complete or partial, according to the underlying protocol) for any type of transfer (send, atomic, or RMA).

UCP_WAKEUP_RX This event type will generate an event on completion of any receive operation (complete or partial, according to the underlying protocol).

UCP_WAKEUP_EDGE Use edge-triggered wakeup. The event file descriptor will be signaled only for new events, rather than existing ones.

6.3.5 Function Documentation

6.3.5.1 ucs_status_t ucp_worker_create (ucp_context_h *context,* const ucp_worker_params_t ∗ *params,* ucp_worker_h ∗ *worker_p*)

This routine allocates and initializes a worker object. Each worker is associated with one and only one application context. In the same time, an application context can create multiple workers in order to enable concurrent access to communication resources. For example, application can allocate a dedicated worker for each application thread, where every worker can be progressed independently of others.

Note

The worker object is allocated within context of the calling thread

Parameters

in	*context*	Handle to UCP application context.
in	*params*	User defined ucp_worker_params_t configurations for the UCP worker.
out	*worker_p*	A pointer to the worker object allocated by the UCP library

Returns

Error code as defined by ucs_status_t

Examples:

ucp_hello_world.c.

6.3.5.2 void ucp_worker_destroy (ucp_worker_h *worker*)

This routine releases the resources associated with a UCP worker.

Warning

Once the UCP worker destroy the worker handle cannot be used with any UCP routine.

The destroy process releases and shuts down all resources associated with the worker.

Parameters

in	*worker*	Worker object to destroy.

Examples:

ucp_hello_world.c.

6.3.5.3 ucs_status_t ucp_worker_query (ucp_worker_h *worker,* ucp_worker_attr_t ∗ *attr*)

This routine fetches information about the worker.

Parameters

in	*worker*	Worker object to query.
out	*attr*	Filled with attributes of worker.

Returns

Error code as defined by ucs_status_t

6.3.5.4 void ucp_worker_print_info (ucp_worker_h *worker,* FILE ∗ *stream*)

This routine prints information about the protocols being used, thresholds, UCT transport methods, and other useful information associated with the worker.

Parameters

in	*worker*	Worker object to print information for.
in	*stream*	Output stream to print the information to.

6.3.5.5 ucs_status_t ucp_worker_get_address (ucp_worker_h *worker*, ucp_address_t ** *address_p*, size_t * *address_length_p*)

This routine returns the address of the worker object. This address can be passed to remote instances of the UCP library in order to to connect to this worker. The memory for the address handle is allocated by this function, and must be released by using ucp_worker_release_address() routine.

Parameters

in	*worker*	Worker object whose address to return.
out	*address_p*	A pointer to the worker address.
out	*address_length_p*	The size in bytes of the address.

Returns

Error code as defined by ucs_status_t

Examples:

ucp_hello_world.c.

6.3.5.6 void ucp_worker_release_address (ucp_worker_h *worker*, ucp_address_t * *address*)

This routine release an address handle associated within the worker object.

Warning

Once the address released the address handle cannot be used with any UCP routine.

Parameters

in	*worker*	Worker object that is associated with the address object.
in	*address*	Address to release; the address object has to be allocated using ucp_worker_get_address() routine.

Examples:

ucp_hello_world.c.

6.3.5.7 unsigned ucp_worker_progress (ucp_worker_h *worker*)

This routine explicitly progresses all communication operations on a worker.

Note

- Typically, request wait and test routines call this routine to progress any outstanding operations.
- Transport layers, implementing asynchronous progress using threads, require callbacks and other user code to be thread safe.
- The state of communication can be advanced (progressed) by blocking routines. Nevertheless, the non-blocking routines can not be used for communication progress.

Parameters

in	*worker*	Worker to progress.

Returns

Non-zero if any communication was progressed, zero otherwise.

Examples:

ucp_hello_world.c.

6.3.5.8 ssize_t ucp_stream_worker_poll (ucp_worker_h *worker,* ucp_stream_poll_ep_t * *poll_eps,* size_t *max_eps,* unsigned *flags*)

This non-blocking routine returns endpoints on a worker which are ready to consume streaming data. The ready endpoints are placed in *poll_eps* array, and the function return value indicates how many are there.

Parameters

in	*worker*	Worker to poll.
out	*poll_eps*	Pointer to array of endpoints, should be allocated by user.
in	*max_eps*	Maximal number of endpoints which should be filled in *poll_eps*.
in	*flags*	Reserved for future use.

Returns

Negative value indicates an error according to ucs_status_t. On success, non-negative value (less or equal *max_eps*) indicates actual number of endpoints filled in *poll_eps* array.

6.3.5.9 ucs_status_t ucp_listener_create (ucp_worker_h *worker,* const ucp_listener_params_t * *params,* ucp_listener_h * *listener_p*)

This routine binds the worker object to a ucs_sock_addr_t sockaddr which is set by the user. The worker will listen to incoming connection requests and upon receiving such a request from the remote peer, an endpoint to it will be created. The user's call-back will be invoked once the endpoint is created.

Parameters

in	*worker*	Worker object that is associated with the params object.
in	*params*	User defined ucp_listener_params_t configurations for the ucp_listener_h.
out	*listener_p*	A handle to the created listener, can be released by calling ucp_listener_destroy

Returns

Error code as defined by ucs_status_t

6.3.5.10 void ucp_listener_destroy (ucp_listener_h *listener*)

This routine unbinds the worker from the given handle and stops listening for incoming connection requests on it.

Parameters

in	*listener*	A handle to the listener to stop listening on.

6.3.5.11 ucs_status_t ucp_listener_reject (ucp_listener_h *listener,* ucp_conn_request_h *conn_request*)

Reject the incoming connection request and release associated resources. If the remote initiator endpoint has set an ucp_ep_params_t::err_handler, it will be invoked with status UCS_ERR_REJECTED.

Parameters

in	*listener*	Handle to the listener on which the connection request was received.
in	*conn_request*	Handle to the connection request to reject.

Returns

Error code as defined by ucs_status_t

6.3.5.12 ucs_status_t ucp_worker_fence (ucp_worker_h *worker*)

This routine ensures ordering of non-blocking communication operations on the UCP worker. Communication operations issued on the *worker* prior to this call are guaranteed to be completed before any subsequent communication operations to the same worker which follow the call to fence.

Note

The primary difference between ucp_worker_fence() and the ucp_worker_flush_nb() is the fact the fence routine does not guarantee completion of the operations on the call return but only ensures the order between communication operations. The flush operation on return guarantees that all operations are completed and corresponding memory regions were updated.

Parameters

in	*worker*	UCP worker.

Returns

Error code as defined by ucs_status_t

6.3.5.13 ucs_status_ptr_t ucp_worker_flush_nb (ucp_worker_h *worker,* unsigned *flags,* ucp_send_callback_t *cb*)

This routine flushes all outstanding AMO and RMA communications on the worker. All the AMO and RMA operations issued on the *worker* prior to this call are completed both at the origin and at the target when this call returns.

Note

For description of the differences between flush and fence operations please see ucp_worker_fence()

Parameters

in	*worker*	UCP worker.

in	*flags*	Flags for flush operation. Reserved for future use.
in	*cb*	Callback which will be called when the flush operation completes.

Returns

UCS_OK - The flush operation was completed immediately.
UCS_PTR_IS_ERR(_ptr) - The flush operation failed.
otherwise - Flush operation was scheduled and can be completed in any point in time. The request handle is returned to the application in order to track progress. The application is responsible to release the handle using ucp_request_free() routine.

6.3.5.14 ucs_status_t ucp_worker_flush (ucp_worker_h *worker*)

Deprecated Replaced by ucp_worker_flush_nb. The following example implements the same functionality using ucp_worker_flush_nb :

```
ucs_status_t worker_flush(ucp_worker_h worker)
{
    void *request = ucp_worker_flush_nb(worker);
    if (request == NULL) {
        return UCS_OK;
    } else if (UCS_PTR_IS_ERR(request)) {
        return UCS_PTR_STATUS(request);
    } else {
        ucs_status_t status;
         do {
             ucp_worker_progress(worker);
             status = ucp_request_check_status(request);
         } while (status == UCS_INPROGRESS);
         ucp_request_release(request);
         return status;
     }
}
```

This routine flushes all outstanding AMO and RMA communications on the worker. All the AMO and RMA operations issued on the *worker* prior to this call are completed both at the origin and at the target when this call returns.

Note

For description of the differences between flush and fence operations please see ucp_worker_fence()

Parameters

in	*worker*	UCP worker.

Returns

Error code as defined by ucs_status_t

6.4 UCP Memory routines

Data Structures

- struct ucp_mem_map_params

  Tuning parameters for the UCP memory mapping. More...
- struct ucp_mem_advise_params

  Tuning parameters for the UCP memory advice. More...
- struct ucp_mem_attr

  Attributes of the UCP Memory handle, filled by ucp_mem_query function. More...

Typedefs

- typedef struct ucp_mem_map_params ucp_mem_map_params_t

  Tuning parameters for the UCP memory mapping.
- typedef enum ucp_mem_advice ucp_mem_advice_t

  list of UCP memory use advice.
- typedef struct ucp_mem_advise_params ucp_mem_advise_params_t

  Tuning parameters for the UCP memory advice.
- typedef struct ucp_rkey ∗ ucp_rkey_h

  UCP Remote memory handle.
- typedef struct ucp_mem ∗ ucp_mem_h

  UCP Memory handle.
- typedef struct ucp_mem_attr ucp_mem_attr_t

  Attributes of the UCP Memory handle, filled by ucp_mem_query function.

Enumerations

- enum ucp_mem_map_params_field { UCP_MEM_MAP_PARAM_FIELD_ADDRESS = UCS_BIT(0), UCP↩_MEM_MAP_PARAM_FIELD_LENGTH = UCS_BIT(1), UCP_MEM_MAP_PARAM_FIELD_FLAGS = UC↩S_BIT(2) }

  UCP memory mapping parameters field mask.
- enum ucp_mem_advise_params_field { UCP_MEM_ADVISE_PARAM_FIELD_ADDRESS = UCS_BIT(0), UCP_MEM_ADVISE_PARAM_FIELD_LENGTH = UCS_BIT(1), UCP_MEM_ADVISE_PARAM_FIELD_A↩DVICE = UCS_BIT(2) }

  UCP memory advice parameters field mask.
- enum { UCP_MEM_MAP_NONBLOCK = UCS_BIT(0), UCP_MEM_MAP_ALLOCATE = UCS_BIT(1), UC↩P_MEM_MAP_FIXED = UCS_BIT(2) }

  UCP memory mapping flags.
- enum ucp_mem_advice { UCP_MADV_NORMAL = 0, UCP_MADV_WILLNEED }

  list of UCP memory use advice.
- enum ucp_mem_attr_field { UCP_MEM_ATTR_FIELD_ADDRESS = UCS_BIT(0), UCP_MEM_ATTR_FIE↩LD_LENGTH = UCS_BIT(1) }

  UCP Memory handle attributes field mask.

Functions

- ucs_status_t ucp_mem_map (ucp_context_h context, const ucp_mem_map_params_t ∗params, ucp_↩mem_h ∗memh_p)

  Map or allocate memory for zero-copy operations.
- ucs_status_t ucp_mem_unmap (ucp_context_h context, ucp_mem_h memh)

Unmap memory segment.
- ucs_status_t ucp_mem_query (const ucp_mem_h memh, ucp_mem_attr_t *attr)

  query mapped memory segment
- ucs_status_t ucp_mem_advise (ucp_context_h context, ucp_mem_h memh, ucp_mem_advise_params_←
t *params)

  give advice about the use of memory
- ucs_status_t ucp_rkey_pack (ucp_context_h context, ucp_mem_h memh, void **rkey_buffer_p, size_←
t *size_p)

  Pack memory region remote access key.
- void ucp_rkey_buffer_release (void *rkey_buffer)

  Release packed remote key buffer.
- ucs_status_t ucp_ep_rkey_unpack (ucp_ep_h ep, const void *rkey_buffer, ucp_rkey_h *rkey_p)

  Create remote access key from packed buffer.
- ucs_status_t ucp_rkey_ptr (ucp_rkey_h rkey, uint64_t raddr, void **addr_p)

  Get a local pointer to remote memory.
- void ucp_rkey_destroy (ucp_rkey_h rkey)

  Destroy the remote key.

6.4.1 Detailed Description

UCP Memory routines

6.4.2 Data Structure Documentation

6.4.2.1 struct ucp_mem_map_params

The structure defines the parameters that are used for the UCP memory mapping tuning during the ucp_mem_map routine.

Data Fields

uint64_t	field_mask	Mask of valid fields in this structure, using bits from ucp_mem_map_← params_field. Fields not specified in this mask would be ignored. Provides ABI compatibility with respect to adding new fields.
void *	address	If the address is not NULL, the routine maps (registers) the memory segment pointed to by this address. If the pointer is NULL, the library allocates mapped (registered) memory segment and returns its address in this argument. Therefore, this value is optional. If it's not set (along with its corresponding bit in the field_mask - UCP_MEM_MAP_PARAM_F← IELD_ADDRESS), the ucp_mem_map routine will consider address as set to NULL and will allocate memory.
size_t	length	Length (in bytes) to allocate or map (register). This field is mandatory for filling (along with its corresponding bit in the field_mask - UCP_MEM_← MAP_PARAM_FIELD_LENGTH). The ucp_mem_map routine will return with an error if the length isn't specified.
unsigned	flags	Allocation flags, e.g. UCP_MEM_MAP_NONBLOCK. This value is optional. If it's not set (along with its corresponding bit in the field_mask - UCP_MEM_MAP_PARAM_FIELD_FLAGS), the ucp_mem_map routine will consider the flags as set to zero.

6.4.2.2 struct ucp_mem_advise_params

This structure defines the parameters that are used for the UCP memory advice tuning during the ucp_mem_advise routine.

Data Fields

uint64_t	field_mask	Mask of valid fields in this structure, using bits from ucp_mem_advise←_params_field. All fields are mandatory. Provides ABI compatibility with respect to adding new fields.
void *	address	Memory base address.
size_t	length	Length (in bytes) to allocate or map (register).
ucp_mem_←advice_t	advice	Memory use advice ucp_mem_advice

6.4.2.3 struct ucp_mem_attr

Data Fields

uint64_t	field_mask	Mask of valid fields in this structure, using bits from ucp_mem_attr_←field. Fields not specified in this mask would be ignored. Provides ABI compatibility with respect to adding new fields.
void *	address	Address of the memory segment.
size_t	length	Size of the memory segment.

6.4.3 Typedef Documentation

6.4.3.1 typedef struct ucp_mem_map_params ucp_mem_map_params_t

The structure defines the parameters that are used for the UCP memory mapping tuning during the ucp_mem_map routine.

6.4.3.2 typedef enum ucp_mem_advice ucp_mem_advice_t

The enumeration list describes memory advice supported by ucp_mem_advise() function.

6.4.3.3 typedef struct ucp_mem_advise_params ucp_mem_advise_params_t

This structure defines the parameters that are used for the UCP memory advice tuning during the ucp_mem_advise routine.

6.4.3.4 typedef struct ucp_rkey* ucp_rkey_h

Remote memory handle is an opaque object representing remote memory access information. Typically, the handle includes a memory access key and other network hardware specific information, which are input to remote memory access operations, such as PUT, GET, and ATOMIC. The object is communicated to remote peers to enable an access to the memory region.

6.4.3.5 typedef struct ucp_mem* ucp_mem_h

Memory handle is an opaque object representing a memory region allocated through UCP library, which is optimized for remote memory access operations (zero-copy operations). The memory handle is a self-contained object, which includes the information required to access the memory region locally, while remote key is used to access it remotely. The memory could be registered to one or multiple network resources that are supported by UCP, such as Infini←Band, Gemini, and others.

6.4.3.6 typedef struct ucp_mem_attr ucp_mem_attr_t

6.4.4 Enumeration Type Documentation

6.4.4.1 enum ucp_mem_map_params_field

The enumeration allows specifying which fields in ucp_mem_map_params_t are present. It is used for the enablement of backward compatibility support.

Enumerator

UCP_MEM_MAP_PARAM_FIELD_ADDRESS Address of the memory that would be used in the ucp_mem_map routine.

UCP_MEM_MAP_PARAM_FIELD_LENGTH The size of memory that would be allocated or registered in the ucp_mem_map routine.

UCP_MEM_MAP_PARAM_FIELD_FLAGS Allocation flags.

6.4.4.2 enum ucp_mem_advise_params_field

The enumeration allows specifying which fields in ucp_mem_advise_params_t are present. It is used for the enablement of backward compatibility support.

Enumerator

UCP_MEM_ADVISE_PARAM_FIELD_ADDRESS Address of the memory

UCP_MEM_ADVISE_PARAM_FIELD_LENGTH The size of memory

UCP_MEM_ADVISE_PARAM_FIELD_ADVICE Advice on memory usage

6.4.4.3 anonymous enum

The enumeration list describes the memory mapping flags supported by ucp_mem_map() function.

Enumerator

UCP_MEM_MAP_NONBLOCK Complete the mapping faster, possibly by not populating the pages in the mapping up-front, and mapping them later when they are accessed by communication routines.

UCP_MEM_MAP_ALLOCATE Identify requirement for allocation, if passed address is not a null-pointer then it will be used as a hint or direct address for allocation.

UCP_MEM_MAP_FIXED Don't interpret address as a hint: place the mapping at exactly that address. The address must be a multiple of the page size.

6.4.4.4 enum ucp_mem_advice

The enumeration list describes memory advice supported by ucp_mem_advise() function.

Enumerator

UCP_MADV_NORMAL No special treatment

UCP_MADV_WILLNEED can be used on the memory mapped with UCP_MEM_MAP_NONBLOCK to speed up memory mapping and to avoid page faults when the memory is accessed for the first time.

6.4.4.5 enum ucp_mem_attr_field

The enumeration allows specifying which fields in ucp_mem_attr_t are present. It is used for the enablement of backward compatibility support.

Enumerator

UCP_MEM_ATTR_FIELD_ADDRESS Virtual address

UCP_MEM_ATTR_FIELD_LENGTH The size of memory region

6.4.5 Function Documentation

6.4.5.1 ucs_status_t ucp_mem_map (ucp_context_h *context*, const ucp_mem_map_params_t ∗ *params*, ucp_mem_h ∗ *memh_p*)

This routine maps or/and allocates a user-specified memory segment with UCP application context and the network resources associated with it. If the application specifies NULL as an address for the memory segment, the routine allocates a mapped memory segment and returns its address in the *address_p* argument. The network stack associated with an application context can typically send and receive data from the mapped memory without CPU intervention; some devices and associated network stacks require the memory to be mapped to send and receive data. The memory handle includes all information required to access the memory locally using UCP routines, while remote registration handle provides an information that is necessary for remote memory access.

Note

Another well know terminology for the "map" operation that is typically used in the context of networking is memory "registration" or "pinning". The UCP library registers the memory the available hardware so it can be assessed directly by the hardware.

Memory mapping assumptions:

- A given memory segment can be mapped by several different communication stacks, if these are compatible.
- The *memh_p* handle returned may be used with any sub-region of the mapped memory.
- If a large segment is registered, and then segmented for subsequent use by a user, then the user is responsible for segmentation and subsequent management.

parameter/flag	NONBLOCK	ALLOCATE	FIXED	address	result
value	0/1 - the value only affects the register/map phase	0	0	0	error if length $>$ 0
		1	0	0	alloc+register
		0	1	0	error
		0	0	defined	register
		1	1	0	error
		1	0	defined	alloc+register,hint
		0	1	defined	error
		1	1	defined	alloc+register,fixed

Table 6.1: Matrix of behavior

Note

- **register** means that the memory will be registered in corresponding transports for RMA/AMO operations. This case intends that the memory was allocated by user before.
- **alloc+register** means that the memory will be allocated in the memory provided by the system and registered in corresponding transports for RMA/AMO operations.

- **alloc+register,hint** means that the memory will be allocated with using ucp_mem_map_params::address as a hint and registered in corresponding transports for RMA/AMO operations.
- **alloc+register,fixed** means that the memory will be allocated and registered in corresponding transports for RMA/AMO operations.
- **error** is an erroneous combination of the parameters.

Parameters

in	*context*	Application context to map (register) and allocate the memory on.
in	*params*	User defined ucp_mem_map_params_t configurations for the UCP memory handle.
out	*memh_p*	UCP handle for the allocated segment.

Returns

Error code as defined by ucs_status_t

6.4.5.2 ucs_status_t ucp_mem_unmap (ucp_context_h *context,* ucp_mem_h *memh*)

This routine unmaps a user specified memory segment, that was previously mapped using the ucp_mem_map() routine. The unmap routine will also release the resources associated with the memory handle. When the function returns, the ucp_mem_h and associated remote key will be invalid and cannot be used with any UCP routine.

Note

Another well know terminology for the "unmap" operation that is typically used in the context of networking is memory "de-registration". The UCP library de-registers the memory the available hardware so it can be returned back to the operation system.

Error cases:

- Once memory is unmapped a network access to the region may cause a failure.

Parameters

in	*context*	Application context which was used to allocate/map the memory.
in	*memh*	Handle to memory region.

Returns

Error code as defined by ucs_status_t

6.4.5.3 ucs_status_t ucp_mem_query (const ucp_mem_h *memh,* ucp_mem_attr_t * *attr*)

This routine returns address and length of memory segment mapped with ucp_mem_map() routine.

Parameters

in	*memh*	Handle to memory region.
out	*attr*	Filled with attributes of the UCP memory handle.

Returns

Error code as defined by ucs_status_t

6.4.5.4 ucs_status_t ucp_mem_advise (ucp_context_h *context,* ucp_mem_h *memh,* ucp_mem_advise_params_t ∗ *params*)

This routine advises the UCP about how to handle memory range beginning at address and size of length bytes. This call does not influence the semantics of the application, but may influence its performance. The UCP may ignore the advice.

Parameters

in	*context*	Application context which was used to allocate/map the memory.
in	*memh*	Handle to memory region.
in	*params*	Memory base address and length. The advice field is used to pass memory use advice as defined in the ucp_mem_advice list The memory range must belong to the *memh*

Returns

Error code as defined by ucs_status_t

6.4.5.5 ucs_status_t ucp_rkey_pack (ucp_context_h *context,* ucp_mem_h *memh,* void ∗∗ *rkey_buffer_p,* size_t ∗ *size_p*)

This routine allocates memory buffer and packs into the buffer a remote access key (RKEY) object. RKEY is an opaque object that provides the information that is necessary for remote memory access. This routine packs the RKEY object in a portable format such that the object can be unpacked on any platform supported by the UCP library. In order to release the memory buffer allocated by this routine the application is responsible to call the ucp_rkey_buffer_release() routine.

Note

- RKEYs for InfiniBand and Cray Aries networks typically includes InifiniBand and Aries key.
- In order to enable remote direct memory access to the memory associated with the memory handle the application is responsible to share the RKEY with the peers that will initiate the access.

Parameters

in	*context*	Application context which was used to allocate/map the memory.
in	*memh*	Handle to memory region.
out	*rkey_buffer_p*	Memory buffer allocated by the library. The buffer contains packed RKEY.
out	*size_p*	Size (in bytes) of the packed RKEY.

Returns

Error code as defined by ucs_status_t

6.4.5.6 void ucp_rkey_buffer_release (void ∗ *rkey_buffer*)

This routine releases the buffer that was allocated using ucp_rkey_pack().

Warning

- Once memory is released an access to the memory may cause a failure.
- If the input memory address was not allocated using ucp_rkey_pack() routine the behaviour of this routine is undefined.

Parameters

`in`	*rkey_buffer*	Buffer to release.

6.4.5.7 ucs_status_t ucp_ep_rkey_unpack (ucp_ep_h *ep*, const void ∗ *rkey_buffer*, ucp_rkey_h ∗ *rkey_p*)

This routine unpacks the remote key (RKEY) object into the local memory such that it can be accesses and used by UCP routines. The RKEY object has to be packed using the ucp_rkey_pack() routine. Application code should not make any alternations to the content of the RKEY buffer.

Parameters

`in`	*ep*	Endpoint to access using the remote key.
`in`	*rkey_buffer*	Packed rkey.
`out`	*rkey_p*	Remote key handle.

Returns

Error code as defined by ucs_status_t

6.4.5.8 ucs_status_t ucp_rkey_ptr (ucp_rkey_h *rkey*, uint64_t *raddr*, void ∗∗ *addr_p*)

This routine returns a local pointer to the remote memory described by the rkey.

Note

This routine can return a valid pointer only for the endpoints that are reacheble via shared memory.

Parameters

`in`	*rkey*	A remote key handle.
`in`	*raddr*	A remote memory address within the memory area described by the rkey.
`out`	*addr_p*	A pointer that can be used for direct access to the remote memory.

Returns

Error code as defined by ucs_status_t if the remote memory cannot be accessed directly or the remote memory address is not valid.

6.4.5.9 void ucp_rkey_destroy (ucp_rkey_h *rkey*)

This routine destroys the RKEY object and the memory that was allocated using the ucp_ep_rkey_unpack() routine. This routine also releases any resources that are associated with the RKEY object.

Warning

- Once the RKEY object is released an access to the memory will cause an undefined failure.
- If the RKEY object was not created using ucp_ep_rkey_unpack() routine the behaviour of this routine is undefined.

Parameters

in	*rkey*	Remote key to destroy.

6.5 UCP Wake-up routines

Functions

- ucs_status_t ucp_worker_get_efd (ucp_worker_h worker, int ∗fd)

  Obtain an event file descriptor for event notification.
- ucs_status_t ucp_worker_wait (ucp_worker_h worker)

  Wait for an event of the worker.
- void ucp_worker_wait_mem (ucp_worker_h worker, void ∗address)

  Wait for memory update on the address.
- ucs_status_t ucp_worker_arm (ucp_worker_h worker)

  Turn on event notification for the next event.
- ucs_status_t ucp_worker_signal (ucp_worker_h worker)

  Cause an event of the worker.

6.5.1 Detailed Description

UCP Wake-up routines

6.5.2 Function Documentation

6.5.2.1 ucs_status_t ucp_worker_get_efd (ucp_worker_h *worker,* int ∗ *fd*)

This routine returns a valid file descriptor for polling functions. The file descriptor will get signaled when an event occurs, as part of the wake-up mechanism. Signaling means a call to poll() or select() with this file descriptor will return at this point, with this descriptor marked as the reason (or one of the reasons) the function has returned. The user does not need to release the obtained file descriptor.

The wake-up mechanism exists to allow for the user process to register for notifications on events of the underlying interfaces, and wait until such occur. This is an alternative to repeated polling for request completion. The goal is to allow for waiting while consuming minimal resources from the system. This is recommended for cases where traffic is infrequent, and latency can be traded for lower resource consumption while waiting for it.

There are two alternative ways to use the wakeup mechanism: the first is the file descriptor obtained per worker (this function) and the second is the ucp_worker_wait function for waiting on the next event internally.

Note

UCP features have to be triggered with UCP_FEATURE_WAKEUP to select proper transport

Parameters

in	*worker*	Worker of notified events.
out	*fd*	File descriptor.

Returns

Error code as defined by ucs_status_t

Examples:

ucp_hello_world.c.

6.5.2.2 ucs_status_t ucp_worker_wait (ucp_worker_h *worker*)

This routine waits (blocking) until an event has happened, as part of the wake-up mechanism.

This function is guaranteed to return only if new communication events occur on the *worker*. Therefore one must drain all existing events before waiting on the file descriptor. This can be achieved by calling ucp_worker_progress repeatedly until it returns 0.

There are two alternative ways to use the wakeup mechanism. The first is by polling on a per-worker file descriptor obtained from ucp_worker_get_efd. The second is by using this function to perform an internal wait for the next event associated with the specified worker.

Note

During the blocking call the wake-up mechanism relies on other means of notification and may not progress some of the requests as it would when calling ucp_worker_progress (which is not invoked in that duration). UCP features have to be triggered with UCP_FEATURE_WAKEUP to select proper transport

Parameters

in	*worker*	Worker to wait for events on.

Returns

Error code as defined by ucs_status_t

Examples:

ucp_hello_world.c.

6.5.2.3 void ucp_worker_wait_mem (ucp_worker_h *worker*, void * *address*)

This routine waits for a memory update at the local memory *address*. This is a blocking routine. The routine returns when the memory address is updated ("write") or an event occurs in the system.

This function is guaranteed to return only if new communication events occur on the worker or *address* is modified. Therefore one must drain all existing events before waiting on the file descriptor. This can be achieved by calling ucp_worker_progress repeatedly until it returns 0.

Note

This routine can be used by an application that executes busy-waiting loop checking for a memory update. Instead of continuous busy-waiting on an address the application can use *ucp_worker_wait_mem*, which may suspend execution until the memory is updated. The goal of the routine is to provide an opportunity for energy savings for architectures that support this functionality.

Parameters

in	*worker*	Worker to wait for updates on.
in	*address*	Local memory address

6.5.2.4 ucs_status_t ucp_worker_arm (ucp_worker_h *worker*)

This routine needs to be called before waiting on each notification on this worker, so will typically be called once the processing of the previous event is over, as part of the wake-up mechanism.

The worker must be armed before waiting on an event (must be re-armed after it has been signaled for re-use) with ucp_worker_arm. The events triggering a signal of the file descriptor from ucp_worker_get_efd depend on the interfaces used by the worker and defined in the transport layer, and typically represent a request completion or newly available resources. It can also be triggered by calling ucp_worker_signal .

The file descriptor is guaranteed to become signaled only if new communication events occur on the *worker*. Therefore one must drain all existing events before waiting on the file descriptor. This can be achieved by calling ucp_worker_progress repeatedly until it returns 0.

```
void application_initialization() {
// should be called once in application init flow and before
// process_comminucation() is used
    ...
    status = ucp_worker_get_efd(worker, &fd);
    ...
}

void process_comminucation() {
// should be called every time need to wait for some condition such as
// ucp request completion in sleep mode.

    for (;;) {
        // check for stop condition as long as progress is made
        if (check_for_events()) {
             break;
        } else if (ucp_worker_progress(worker)) {
             continue;                   // some progress happened but condition not met
        }

        // arm the worker and clean-up fd
        status = ucp_worker_arm(worker);
        if (UCS_OK == status) {
            poll(&fds, nfds, timeout);  // wait for events (sleep mode)
        } else if (UCS_ERR_BUSY == status) {
            continue;                    // could not arm, need to progress more
        } else {
            abort();
        }
    }
}
```

Note

UCP features have to be triggered with UCP_FEATURE_WAKEUP to select proper transport

Parameters

in	*worker*	Worker of notified events.

Returns

UCS_OK The operation completed successfully. File descriptor will be signaled by new events.
UCS_ERR_BUSY There are unprocessed events which prevent the file descriptor from being armed. These events should be removed by calling ucp_worker_progress(). The operation is not completed. File descriptor will not be signaled by new events.
Other different error codes in case of issues.

Examples:

ucp_hello_world.c.

6.5.2.5 ucs_status_t ucp_worker_signal (ucp_worker_h *worker*)

This routine signals that the event has happened, as part of the wake-up mechanism. This function causes a blocking call to ucp_worker_wait or waiting on a file descriptor from ucp_worker_get_efd to return, even if no event from the underlying interfaces has taken place.

Note

It's safe to use this routine from any thread, even if UCX is compiled without multi-threading support and/or initialized with any value of ucp_params_t::mt_workers_shared and ucp_worker_params_t::thread_mode parameters

Parameters

`in`	*worker*	Worker to wait for events on.

Returns

Error code as defined by ucs_status_t

6.6 UCP Endpoint

Data Structures

- struct ucp_stream_poll_ep_t

  Output parameter of ucp_stream_worker_poll function. More...
- struct ucp_ep_params

  Tuning parameters for the UCP endpoint. More...

Typedefs

- typedef struct ucp_ep ∗ ucp_ep_h

  UCP Endpoint.
- typedef struct ucp_conn_request ∗ ucp_conn_request_h

  UCP connection request.
- typedef struct ucp_ep_params ucp_ep_params_t

  Tuning parameters for the UCP endpoint.

Enumerations

- enum ucp_ep_params_field {
  UCP_EP_PARAM_FIELD_REMOTE_ADDRESS = UCS_BIT(0), UCP_EP_PARAM_FIELD_ERR_HANDLING_MODE = UCS_BIT(1), UCP_EP_PARAM_FIELD_ERR_HANDLER = UCS_BIT(2), UCP_EP_PARAM_FIELD_USER_DATA = UCS_BIT(3),
  UCP_EP_PARAM_FIELD_SOCK_ADDR = UCS_BIT(4), UCP_EP_PARAM_FIELD_FLAGS = UCS_BIT(5),
  UCP_EP_PARAM_FIELD_CONN_REQUEST = UCS_BIT(6) }

  UCP endpoint parameters field mask.
- enum ucp_ep_params_flags_field { UCP_EP_PARAMS_FLAGS_CLIENT_SERVER = UCS_BIT(0), UCP_EP_PARAMS_FLAGS_NO_LOOPBACK = UCS_BIT(1) }

  UCP endpoint parameters flags.
- enum ucp_ep_close_mode { UCP_EP_CLOSE_MODE_FORCE = 0, UCP_EP_CLOSE_MODE_FLUSH = 1 }

  Close UCP endpoint modes.
- enum ucp_err_handling_mode_t { UCP_ERR_HANDLING_MODE_NONE, UCP_ERR_HANDLING_MODE_PEER }

  Error handling mode for the UCP endpoint.

Functions

- ucs_status_t ucp_ep_create (ucp_worker_h worker, const ucp_ep_params_t ∗params, ucp_ep_h ∗ep_p)

  Create and connect an endpoint.
- ucs_status_ptr_t ucp_ep_close_nb (ucp_ep_h ep, unsigned mode)

  Non-blocking endpoint closure.
- void ucp_ep_print_info (ucp_ep_h ep, FILE ∗stream)

  Print endpoint information.
- ucs_status_ptr_t ucp_ep_flush_nb (ucp_ep_h ep, unsigned flags, ucp_send_callback_t cb)

  Non-blocking flush of outstanding AMO and RMA operations on the endpoint.
- void ucp_request_release (void ∗request)
- void ucp_ep_destroy (ucp_ep_h ep)
- ucs_status_ptr_t ucp_disconnect_nb (ucp_ep_h ep)
- ucs_status_t ucp_request_test (void ∗request, ucp_tag_recv_info_t ∗info)

- ucs_status_t ucp_ep_flush (ucp_ep_h ep)
- ucs_status_ptr_t ucp_ep_modify_nb (ucp_ep_h ep, const ucp_ep_params_t *params)

  Modify endpoint parameters.

6.6.1 Detailed Description

UCP Endpoint routines

6.6.2 Data Structure Documentation

6.6.2.1 struct ucp_stream_poll_ep_t

The structure defines the endpoint and its user data.

Data Fields

ucp_ep_h	ep	Endpoint handle.
void *	user_data	User data associated with an endpoint passed in ucp_ep_params_t::user_data.
unsigned	flags	Reserved for future use.
uint8_t	reserved[16]	Reserved for future use.

6.6.2.2 struct ucp_ep_params

The structure defines the parameters that are used for the UCP endpoint tuning during the UCP ep creation.

Examples:

ucp_hello_world.c.

Data Fields

uint64_t	field_mask	Mask of valid fields in this structure, using bits from ucp_ep_params_field. Fields not specified in this mask would be ignored. Provides ABI compatibility with respect to adding new fields.
const ucp_address_t *	address	Destination address; this field should be set along with its corresponding bit in the field_mask - UCP_EP_PARAM_FIELD_REMOTE_ADDRESS and must be obtained using ucp_worker_get_address.
ucp_err_handling_mode_t	err_mode	Desired error handling mode, optional parameter. Default value is UCP_ERR_HANDLING_MODE_NONE.
ucp_err_handler_t	err_handler	Handler to process transport level failure.
void *	user_data	User data associated with an endpoint. See ucp_stream_poll_ep_t and ucp_err_handler_t
unsigned	flags	Endpoint flags from ucp_ep_params_flags_field. This value is optional. If it's not set (along with its corresponding bit in the field_mask - UCP_EP_PARAM_FIELD_FLAGS), the ucp_ep_create() routine will consider the flags as set to zero.

ucs_sock_addr_t	sockaddr	Destination address in the form of a sockaddr; this field should be set along with its corresponding bit in the field_mask - UCP_EP_PARAM_FIELD_SOCK_ADDR and must be obtained from the user, it means that this type of the endpoint creation is possible only on client side in client-server connection establishment flow.
ucp_conn_request_h	conn_request	Connection request from client; this field should be set along with its corresponding bit in the field_mask - UCP_EP_PARAM_FIELD_CONN_REQUEST and must be obtained from ucp_listener_conn_callback_t, it means that this type of the endpoint creation is possible only on server side in client-server connection establishment flow.

6.6.3 Typedef Documentation

6.6.3.1 typedef struct ucp_ep∗ **ucp_ep_h**

The endpoint handle is an opaque object that is used to address a remote worker. It typically provides a description of source, destination, or both. All UCP communication routines address a destination with the endpoint handle. The endpoint handle is associated with only one UCP context. UCP provides the endpoint create routine to create the endpoint handle and the destroy routine to destroy the endpoint handle.

6.6.3.2 typedef struct ucp_conn_request∗ **ucp_conn_request_h**

A server-side handle to incoming connection request. Can be used to create an endpoint which connects back to the client.

6.6.3.3 typedef struct **ucp_ep_params ucp_ep_params_t**

The structure defines the parameters that are used for the UCP endpoint tuning during the UCP ep creation.

6.6.4 Enumeration Type Documentation

6.6.4.1 enum **ucp_ep_params_field**

The enumeration allows specifying which fields in ucp_ep_params_t are present. It is used for the enablement of backward compatibility support.

Enumerator

UCP_EP_PARAM_FIELD_REMOTE_ADDRESS Address of remote peer

UCP_EP_PARAM_FIELD_ERR_HANDLING_MODE Error handling mode. ucp_err_handling_mode_t

UCP_EP_PARAM_FIELD_ERR_HANDLER Handler to process transport level errors

UCP_EP_PARAM_FIELD_USER_DATA User data pointer

UCP_EP_PARAM_FIELD_SOCK_ADDR Socket address field

UCP_EP_PARAM_FIELD_FLAGS Endpoint flags

UCP_EP_PARAM_FIELD_CONN_REQUEST Connection request field

6.6.4.2 enum **ucp_ep_params_flags_field**

The enumeration list describes the endpoint's parameters flags supported by ucp_ep_create() function.

Enumerator

UCP_EP_PARAMS_FLAGS_CLIENT_SERVER Using a client-server connection establishment mechanism. ucs_sock_addr_t sockaddr field must be provided and contain the address of the remote peer

UCP_EP_PARAMS_FLAGS_NO_LOOPBACK Avoid connecting the endpoint to itself when connecting the endpoint to the same worker it was created on. Affects protocols which send to a particular remote endpoint, for example stream

6.6.4.3 enum ucp_ep_close_mode

The enumeration is used to specify the behavior of ucp_ep_close_nb.

Enumerator

UCP_EP_CLOSE_MODE_FORCE ucp_ep_close_nb releases the endpoint without any confirmation from the peer. All outstanding requests will be completed with UCS_ERR_CANCELED error.

Note

This mode may cause transport level errors on remote side, so it requires set UCP_ERR_HANDLING_MODE_PEER for all endpoints created on both (local and remote) sides to avoid undefined behavior.

UCP_EP_CLOSE_MODE_FLUSH ucp_ep_close_nb schedules flushes on all outstanding operations.

6.6.4.4 enum ucp_err_handling_mode_t

Specifies error handling mode for the UCP endpoint.

Enumerator

UCP_ERR_HANDLING_MODE_NONE No guarantees about error reporting, imposes minimal overhead from a performance perspective.

Note

In this mode, any error reporting will not generate calls to ucp_ep_params_t::err_handler.

UCP_ERR_HANDLING_MODE_PEER Guarantees that send requests are always completed (successfully or error) even in case of remote failure, disables protocols and APIs which may cause a hang or undefined behavior in case of peer failure, may affect performance and memory footprint

6.6.5 Function Documentation

6.6.5.1 ucs_status_t ucp_ep_create (ucp_worker_h *worker,* const ucp_ep_params_t ∗ *params,* ucp_ep_h ∗ *ep_p*)

This routine creates and connects an endpoint on a local worker for a destination address that identifies the remote worker. This function is non-blocking, and communications may begin immediately after it returns. If the connection process is not completed, communications may be delayed. The created endpoint is associated with one and only one worker.

Parameters

in	*worker*	Handle to the worker; the endpoint is associated with the worker.
in	*params*	User defined ucp_ep_params_t configurations for the UCP endpoint.
out	*ep_p*	A handle to the created endpoint.

Returns

Error code as defined by ucs_status_t

Note

One of the following fields has to be specified:

- ucp_ep_params_t::address
- ucp_ep_params_t::sockaddr
- ucp_ep_params_t::conn_request

By default, ucp_ep_create() will connect an endpoint to itself if the endpoint is destined to the same *worker* on which it was created, i.e. *params.address* belongs to *worker*. This behavior can be changed by passing the UCP_EP_PARAMS_FLAGS_NO_LOOPBACK flag in *params.flags*. In that case, the endpoint will be connected to the *next* endpoint created in the same way on the same *worker*.

Examples:

ucp_hello_world.c.

6.6.5.2 ucs_status_ptr_t ucp_ep_close_nb (ucp_ep_h *ep*, unsigned *mode*)

This routine releases the endpoint. The endpoint closure process depends on the selected *mode*.

Parameters

in	*ep*	Handle to the endpoint to close.
in	*mode*	One from ucp_ep_close_mode value.

Returns

UCS_OK - The endpoint is closed successfully.
UCS_PTR_IS_ERR(_ptr) - The closure failed and an error code indicates the transport level status. However, resources are released and the *endpoint* can no longer be used.
otherwise - The closure process is started, and can be completed at any point in time. A request handle is returned to the application in order to track progress of the endpoint closure. The application is responsible for releasing the handle using the ucp_request_free routine.

Note

ucp_ep_close_nb replaces deprecated ucp_disconnect_nb and ucp_ep_destroy

6.6.5.3 void ucp_ep_print_info (ucp_ep_h *ep*, FILE ∗ *stream*)

This routine prints information about the endpoint transport methods, their thresholds, and other useful information associated with the endpoint.

Parameters

in	*ep*	Endpoint object whose configuration to print.
in	*stream*	Output stream to print the information to.

6.6.5.4 ucs_status_ptr_t ucp_ep_flush_nb (ucp_ep_h *ep*, unsigned *flags*, ucp_send_callback_t *cb*)

This routine flushes all outstanding AMO and RMA communications on the endpoint. All the AMO and RM↩A operations issued on the *ep* prior to this call are completed both at the origin and at the target endpoint when this call returns.

Parameters

in	*ep*	UCP endpoint.
in	*flags*	Flags for flush operation. Reserved for future use.
in	*cb*	Callback which will be called when the flush operation completes.

Returns

UCS_OK - The flush operation was completed immediately.
UCS_PTR_IS_ERR(_ptr) - The flush operation failed.
otherwise - Flush operation was scheduled and can be completed in any point in time. The request handle is returned to the application in order to track progress. The application is responsible to release the handle using ucp_request_free() routine.

The following example demonstrates how blocking flush can be implemented using non-blocking flush:

```
void empty_function(void *request, ucs_status_t status)
{
}

ucs_status_t blocking_ep_flush(ucp_ep_h ep, ucp_worker_h worker)
{
    void *request;

    request = ucp_ep_flush_nb(ep, 0, empty_function);
    if (request == NULL) {
        return UCS_OK;
    } else if (UCS_PTR_IS_ERR(request)) {
        return UCS_PTR_STATUS(request);
    } else {
        ucs_status_t status;
        do {
            ucp_worker_progress(worker);
            status = ucp_request_check_status(request);
        } while (status == UCS_INPROGRESS);
        ucp_request_free(request);
        return status;
    }
}
```

Examples:

ucp_hello_world.c.

6.6.5.5 void ucp_request_release (void * *request*)

Deprecated Replaced by ucp_request_free.

Examples:

ucp_hello_world.c.

6.6.5.6 void ucp_ep_destroy (ucp_ep_h *ep*)

Deprecated Replaced by ucp_ep_close_nb.

Examples:

ucp_hello_world.c.

6.6.5.7 ucs_status_ptr_t ucp_disconnect_nb (ucp_ep_h *ep*)

Deprecated Replaced by ucp_ep_close_nb.

6.6.5.8 ucs_status_t ucp_request_test (void * *request,* ucp_tag_recv_info_t * *info*)

Deprecated Replaced by ucp_tag_recv_request_test and ucp_request_check_status depends on use case.

Note

Please use ucp_request_check_status for cases that only need to check the completion status of an outstanding request. ucp_request_check_status can be used for any type of request. ucp_tag_recv_request_test should only be used for requests returned by ucp_tag_recv_nb (or request allocated by user for ucp_tag_↩
recv_nbr) for which additional information (returned via the *info* pointer) is needed.

6.6.5.9 ucs_status_t ucp_ep_flush (ucp_ep_h *ep*)

Deprecated Replaced by ucp_ep_flush_nb.

6.6.5.10 ucs_status_ptr_t ucp_ep_modify_nb (ucp_ep_h *ep,* const ucp_ep_params_t * *params*)

Deprecated Use ucp_listener_conn_handler_t instead of ucp_listener_accept_handler_t, if you have other use case please submit an issue on `https://github.com/openucx/ucx` or report to `ucx-group@elist.ornl.gov`

This routine modifies endpoint created by ucp_ep_create or ucp_listener_accept_callback_t. For example, this API can be used to setup custom parameters like ucp_ep_params_t::user_data or ucp_ep_params_t::err_handler to endpoint created by ucp_listener_accept_callback_t.

Parameters

in	*ep*	A handle to the endpoint.
in	*params*	User defined ucp_ep_params_t configurations for the UCP endpoint.

Returns

NULL - The endpoint is modified successfully.
UCS_PTR_IS_ERR(_ptr) - The reconfiguration failed and an error code indicates the status. However, the *endpoint* is not modified and can be used further.
otherwise - The reconfiguration process is started, and can be completed at any point in time. A request handle is returned to the application in order to track progress of the endpoint modification. The application is responsible for releasing the handle using the ucp_request_free routine.

Note

See the documentation of ucp_ep_params_t for details, only some of the parameters can be modified.

6.7 UCP Communication routines

Data Structures

- struct ucp_err_handler

  UCP endpoint error handling context. More...

Typedefs

- typedef uint64_t ucp_tag_t

  UCP Tag Identifier.
- typedef struct ucp_recv_desc ∗ ucp_tag_message_h

  UCP Message descriptor.
- typedef uint64_t ucp_datatype_t

  UCP Datatype Identifier.
- typedef void(∗ ucp_send_callback_t) (void ∗request, ucs_status_t status)

  Completion callback for non-blocking sends.
- typedef void(∗ ucp_err_handler_cb_t) (void ∗arg, ucp_ep_h ep, ucs_status_t status)

  Callback to process peer failure.
- typedef struct ucp_err_handler ucp_err_handler_t

  UCP endpoint error handling context.
- typedef void(∗ ucp_stream_recv_callback_t) (void ∗request, ucs_status_t status, size_t length)

  Completion callback for non-blocking stream oriented receives.
- typedef void(∗ ucp_tag_recv_callback_t) (void ∗request, ucs_status_t status, ucp_tag_recv_info_t ∗info)

  Completion callback for non-blocking tag receives.

Enumerations

- enum ucp_atomic_post_op_t {
  UCP_ATOMIC_POST_OP_ADD, UCP_ATOMIC_POST_OP_AND, UCP_ATOMIC_POST_OP_OR, UCP↩
  _ATOMIC_POST_OP_XOR,
  UCP_ATOMIC_POST_OP_LAST }

  Atomic operation requested for ucp_atomic_post.
- enum ucp_atomic_fetch_op_t {
  UCP_ATOMIC_FETCH_OP_FADD, UCP_ATOMIC_FETCH_OP_SWAP, UCP_ATOMIC_FETCH_OP_C↩
  SWAP, UCP_ATOMIC_FETCH_OP_FAND,
  UCP_ATOMIC_FETCH_OP_FOR, UCP_ATOMIC_FETCH_OP_FXOR, UCP_ATOMIC_FETCH_OP_LAST
  }

  Atomic operation requested for ucp_atomic_fetch.
- enum ucp_stream_recv_flags_t { UCP_STREAM_RECV_FLAG_WAITALL = UCS_BIT(0) }

  Flags to define behavior of ucp_stream_recv_nb function.

Functions

- ucs_status_ptr_t ucp_stream_send_nb (ucp_ep_h ep, const void ∗buffer, size_t count, ucp_datatype_↩
  t datatype, ucp_send_callback_t cb, unsigned flags)

  Non-blocking stream send operation.
- ucs_status_ptr_t ucp_tag_send_nb (ucp_ep_h ep, const void ∗buffer, size_t count, ucp_datatype_t datatype, ucp_tag_t tag, ucp_send_callback_t cb)

  Non-blocking tagged-send operations.

- ucs_status_t ucp_tag_send_nbr (ucp_ep_h ep, const void *buffer, size_t count, ucp_datatype_t datatype, ucp_tag_t tag, void *req)

  Non-blocking tagged-send operations with user provided request.
- ucs_status_ptr_t ucp_tag_send_sync_nb (ucp_ep_h ep, const void *buffer, size_t count, ucp_datatype_t datatype, ucp_tag_t tag, ucp_send_callback_t cb)

  Non-blocking synchronous tagged-send operation.
- ucs_status_ptr_t ucp_stream_recv_nb (ucp_ep_h ep, void *buffer, size_t count, ucp_datatype_t datatype, ucp_stream_recv_callback_t cb, size_t *length, unsigned flags)

  Non-blocking stream receive operation of structured data into a user-supplied buffer.
- ucs_status_ptr_t ucp_stream_recv_data_nb (ucp_ep_h ep, size_t *length)

  Non-blocking stream receive operation of unstructured data into a UCP-supplied buffer.
- ucs_status_ptr_t ucp_tag_recv_nb (ucp_worker_h worker, void *buffer, size_t count, ucp_datatype_t datatype, ucp_tag_t tag, ucp_tag_t tag_mask, ucp_tag_recv_callback_t cb)

  Non-blocking tagged-receive operation.
- ucs_status_t ucp_tag_recv_nbr (ucp_worker_h worker, void *buffer, size_t count, ucp_datatype_t datatype, ucp_tag_t tag, ucp_tag_t tag_mask, void *req)

  Non-blocking tagged-receive operation.
- ucp_tag_message_h ucp_tag_probe_nb (ucp_worker_h worker, ucp_tag_t tag, ucp_tag_t tag_mask, int remove, ucp_tag_recv_info_t *info)

  Non-blocking probe and return a message.
- ucs_status_ptr_t ucp_tag_msg_recv_nb (ucp_worker_h worker, void *buffer, size_t count, ucp_datatype_t datatype, ucp_tag_message_h message, ucp_tag_recv_callback_t cb)

  Non-blocking receive operation for a probed message.
- ucs_status_t ucp_put_nbi (ucp_ep_h ep, const void *buffer, size_t length, uint64_t remote_addr, ucp_rkey_h rkey)

  Non-blocking implicit remote memory put operation.
- ucs_status_ptr_t ucp_put_nb (ucp_ep_h ep, const void *buffer, size_t length, uint64_t remote_addr, ucp_rkey_h rkey, ucp_send_callback_t cb)

  Non-blocking remote memory put operation.
- ucs_status_t ucp_get_nbi (ucp_ep_h ep, void *buffer, size_t length, uint64_t remote_addr, ucp_rkey_h rkey)

  Non-blocking implicit remote memory get operation.
- ucs_status_ptr_t ucp_get_nb (ucp_ep_h ep, void *buffer, size_t length, uint64_t remote_addr, ucp_rkey_h rkey, ucp_send_callback_t cb)

  Non-blocking remote memory get operation.
- ucs_status_t ucp_atomic_post (ucp_ep_h ep, ucp_atomic_post_op_t opcode, uint64_t value, size_t op_size, uint64_t remote_addr, ucp_rkey_h rkey)

  Post an atomic memory operation.
- ucs_status_ptr_t ucp_atomic_fetch_nb (ucp_ep_h ep, ucp_atomic_fetch_op_t opcode, uint64_t value, void *result, size_t op_size, uint64_t remote_addr, ucp_rkey_h rkey, ucp_send_callback_t cb)

  Post an atomic fetch operation.
- ucs_status_t ucp_request_check_status (void *request)

  Check the status of non-blocking request.
- ucs_status_t ucp_tag_recv_request_test (void *request, ucp_tag_recv_info_t *info)

  Check the status and currently available state of non-blocking request returned from ucp_tag_recv_nb routine.
- ucs_status_t ucp_stream_recv_request_test (void *request, size_t *length_p)

  Check the status and currently available state of non-blocking request returned from ucp_stream_recv_nb routine.
- void ucp_request_cancel (ucp_worker_h worker, void *request)

  Cancel an outstanding communications request.
- void ucp_stream_data_release (ucp_ep_h ep, void *data)

  Release UCP data buffer returned by ucp_stream_recv_data_nb.
- void ucp_request_free (void *request)

  Release a communications request.

- int ucp_request_is_completed (void ∗request)
- ucs_status_t ucp_put (ucp_ep_h ep, const void ∗buffer, size_t length, uint64_t remote_addr, ucp_rkey_h rkey)

    Blocking remote memory put operation.
- ucs_status_t ucp_get (ucp_ep_h ep, void ∗buffer, size_t length, uint64_t remote_addr, ucp_rkey_h rkey)

    Blocking remote memory get operation.
- ucs_status_t ucp_atomic_add32 (ucp_ep_h ep, uint32_t add, uint64_t remote_addr, ucp_rkey_h rkey)

    Blocking atomic add operation for 32 bit integers.
- ucs_status_t ucp_atomic_add64 (ucp_ep_h ep, uint64_t add, uint64_t remote_addr, ucp_rkey_h rkey)

    Blocking atomic add operation for 64 bit integers.
- ucs_status_t ucp_atomic_fadd32 (ucp_ep_h ep, uint32_t add, uint64_t remote_addr, ucp_rkey_h rkey, uint32_t ∗result)

    Blocking atomic fetch and add operation for 32 bit integers.
- ucs_status_t ucp_atomic_fadd64 (ucp_ep_h ep, uint64_t add, uint64_t remote_addr, ucp_rkey_h rkey, uint64_t ∗result)

    Blocking atomic fetch and add operation for 64 bit integers.
- ucs_status_t ucp_atomic_swap32 (ucp_ep_h ep, uint32_t swap, uint64_t remote_addr, ucp_rkey_h rkey, uint32_t ∗result)

    Blocking atomic swap operation for 32 bit values.
- ucs_status_t ucp_atomic_swap64 (ucp_ep_h ep, uint64_t swap, uint64_t remote_addr, ucp_rkey_h rkey, uint64_t ∗result)

    Blocking atomic swap operation for 64 bit values.
- ucs_status_t ucp_atomic_cswap32 (ucp_ep_h ep, uint32_t compare, uint32_t swap, uint64_t remote_addr, ucp_rkey_h rkey, uint32_t ∗result)

    Blocking atomic conditional swap (cswap) operation for 32 bit values.
- ucs_status_t ucp_atomic_cswap64 (ucp_ep_h ep, uint64_t compare, uint64_t swap, uint64_t remote_addr, ucp_rkey_h rkey, uint64_t ∗result)

    Blocking atomic conditional swap (cswap) operation for 64 bit values.

6.7.1 Detailed Description

UCP Communication routines

6.7.2 Data Structure Documentation

6.7.2.1 struct ucp_err_handler

This structure should be initialized in ucp_ep_params_t to handle peer failure

Data Fields

ucp_err_handler_cb_t	cb	Error handler callback, if NULL, will not be called.
void ∗	arg	User defined argument associated with an endpoint, it will be overridden by ucp_ep_params_t::user_data if both are set.

6.7.3 Typedef Documentation

6.7.3.1 typedef uint64_t ucp_tag_t

UCP tag identifier is a 64bit object used for message identification. UCP tag send and receive operations use the object for an implementation tag matching semantics (derivative of MPI tag matching semantics).

6.7.3.2 typedef struct ucp_recv_desc∗ ucp_tag_message_h

UCP Message descriptor is an opaque handle for a message returned by ucp_tag_probe_nb. This handle can be passed to ucp_tag_msg_recv_nb in order to receive the message data to a specific buffer.

6.7.3.3 typedef uint64_t ucp_datatype_t

UCP datatype identifier is a 64bit object used for datatype identification. Predefined UCP identifiers are defined by ucp_dt_type.

6.7.3.4 typedef void(∗ ucp_send_callback_t) (void ∗request, ucs_status_t status)

This callback routine is invoked whenever the send operation is completed. It is important to note that the call-back is only invoked in a case when the operation cannot be completed in place.

Parameters

in	*request*	The completed send request.
in	*status*	Completion status. If the send operation was completed successfully UCX_OK is returned. If send operation was canceled UCS_ERR_CANCELED is returned. Otherwise, an error status is returned.

6.7.3.5 typedef void(∗ ucp_err_handler_cb_t) (void ∗arg, ucp_ep_h ep, ucs_status_t status)

This callback routine is invoked when transport level error detected.

Parameters

in	*arg*	User argument to be passed to the callback.
in	*ep*	Endpoint to handle transport level error. Upon return from the callback, this *ep* is no longer usable and all subsequent operations on this *ep* will fail with the error code passed in *status*.
in	*status*	error status.

6.7.3.6 typedef struct ucp_err_handler ucp_err_handler_t

This structure should be initialized in ucp_ep_params_t to handle peer failure

6.7.3.7 typedef void(∗ ucp_stream_recv_callback_t) (void ∗request, ucs_status_t status, size_t length)

This callback routine is invoked whenever the receive operation is completed and the data is ready in the receive buffer.

Parameters

in	*request*	The completed receive request.
in	*status*	Completion status. If the send operation was completed successfully UCX_OK is returned. Otherwise, an error status is returned.
in	*length*	The size of the received data in bytes, always boundary of base datatype size. The value is valid only if the status is UCS_OK.

6.7.3.8 typedef void(∗ ucp_tag_recv_callback_t) (void ∗request, ucs_status_t status, ucp_tag_recv_info_t ∗info)

This callback routine is invoked whenever the receive operation is completed and the data is ready in the receive buffer.

Parameters

in	*request*	The completed receive request.
in	*status*	Completion status. If the send operation was completed successfully UCX_OK is returned. If send operation was canceled UCS_ERR_CANCELED is returned. If the data can not fit into the receive buffer the UCS_ERR_MESSAGE_TRUNCATED error code is returned. Otherwise, an error status is returned.
in	*info*	Completion information The *info* descriptor is Valid only if the status is UCS_OK.

6.7.4 Enumeration Type Documentation

6.7.4.1 enum ucp_atomic_post_op_t

This enumeration defines which atomic memory operation should be performed by the ucp_atomic_post family of fuctions. All of these are non-fetching atomics and will not result in a request handle.

Enumerator

UCP_ATOMIC_POST_OP_ADD Atomic add

UCP_ATOMIC_POST_OP_AND Atomic and

UCP_ATOMIC_POST_OP_OR Atomic or

UCP_ATOMIC_POST_OP_XOR Atomic xor

UCP_ATOMIC_POST_OP_LAST

6.7.4.2 enum ucp_atomic_fetch_op_t

This enumeration defines which atomic memory operation should be performed by the ucp_atomic_fetch family of functions. All of these functions will fetch data from the remote node.

Enumerator

UCP_ATOMIC_FETCH_OP_FADD Atomic Fetch and add

UCP_ATOMIC_FETCH_OP_SWAP Atomic swap

UCP_ATOMIC_FETCH_OP_CSWAP Atomic conditional swap

UCP_ATOMIC_FETCH_OP_FAND Atomic Fetch and and

UCP_ATOMIC_FETCH_OP_FOR Atomic Fetch and or

UCP_ATOMIC_FETCH_OP_FXOR Atomic Fetch and xor

UCP_ATOMIC_FETCH_OP_LAST

6.7.4.3 enum ucp_stream_recv_flags_t

This enumeration defines behavior of ucp_stream_recv_nb function.

Enumerator

UCP_STREAM_RECV_FLAG_WAITALL This flag requests that operation will not be completed untill all amout of requested data is received and placed in the user buffer.

6.7.5 Function Documentation

6.7.5.1 ucs_status_ptr_t ucp_stream_send_nb (ucp_ep_h *ep*, const void * *buffer*, size_t *count*, ucp_datatype_t *datatype*, ucp_send_callback_t *cb*, unsigned *flags*)

This routine sends data that is described by the local address *buffer*, size *count*, and *datatype* object to the destination endpoint *ep*. The routine is non-blocking and therefore returns immediately, however the actual send operation may be delayed. The send operation is considered completed when it is safe to reuse the source *buffer*. If the send operation is completed immediately the routine returns UCS_OK and the call-back function *cb* is **not** invoked. If the operation is **not** completed immediately and no error reported, then the UCP library will schedule invocation of the call-back *cb* upon completion of the send operation. In other words, the completion of the operation will be signaled either by the return code or by the call-back.

Note

The user should not modify any part of the *buffer* after this operation is called, until the operation completes.

Parameters

in	*ep*	Destination endpoint handle.
in	*buffer*	Pointer to the message buffer (payload).
in	*count*	Number of elements to send.
in	*datatype*	Datatype descriptor for the elements in the buffer.
in	*cb*	Callback function that is invoked whenever the send operation is completed. It is important to note that the call-back is only invoked in a case when the operation cannot be completed in place.
in	*flags*	Reserved for future use.

Returns

UCS_OK - The send operation was completed immediately.
UCS_PTR_IS_ERR(_ptr) - The send operation failed.
otherwise - Operation was scheduled for send and can be completed in any point in time. The request handle is returned to the application in order to track progress of the message. The application is responsible to release the handle using ucp_request_free routine.

6.7.5.2 ucs_status_ptr_t ucp_tag_send_nb (ucp_ep_h *ep*, const void * *buffer*, size_t *count*, ucp_datatype_t *datatype*, ucp_tag_t *tag*, ucp_send_callback_t *cb*)

This routine sends a messages that is described by the local address *buffer*, size *count*, and *datatype* object to the destination endpoint *ep*. Each message is associated with a *tag* value that is used for message matching on the receiver. The routine is non-blocking and therefore returns immediately, however the actual send operation may be delayed. The send operation is considered completed when it is safe to reuse the source *buffer*. If the send operation is completed immediately the routine return UCS_OK and the call-back function *cb* is **not** invoked. If the operation is **not** completed immediately and no error reported then the UCP library will schedule to invoke the call-back *cb* whenever the send operation will be completed. In other words, the completion of a message can be signaled by the return code or the call-back.

Note

The user should not modify any part of the *buffer* after this operation is called, until the operation completes.

Parameters

in	*ep*	Destination endpoint handle.
in	*buffer*	Pointer to the message buffer (payload).
in	*count*	Number of elements to send
in	*datatype*	Datatype descriptor for the elements in the buffer.
in	*tag*	Message tag.
in	*cb*	Callback function that is invoked whenever the send operation is completed. It is important to note that the call-back is only invoked in a case when the operation cannot be completed in place.

Returns

UCS_OK - The send operation was completed immediately.
UCS_PTR_IS_ERR(_ptr) - The send operation failed.
otherwise - Operation was scheduled for send and can be completed in any point in time. The request handle is returned to the application in order to track progress of the message. The application is responsible to released the handle using ucp_request_free() routine.

Examples:

ucp_hello_world.c.

6.7.5.3 ucs_status_t ucp_tag_send_nbr (ucp_ep_h *ep,* const void ∗ *buffer,* size_t *count,* ucp_datatype_t *datatype,* ucp_tag_t *tag,* void ∗ *req*)

This routine provides a convenient and efficient way to implement a blocking send pattern. It also completes requests faster than ucp_tag_send_nbr() because:

- it always uses uct_ep_am_bcopy() to send data up to the rendezvous threshold.
- its rendezvous threshold is higher than the one used by the ucp_tag_send_nb(). The threshold is controlled by the **UCX_SEND_NBR_RNDV_THRESH** environment variable.
- its request handling is simpler. There is no callback and no need to allocate and free requests. In fact request can be allocated by caller on the stack.

This routine sends a messages that is described by the local address *buffer*, size *count*, and *datatype* object to the destination endpoint *ep*. Each message is associated with a *tag* value that is used for message matching on the receiver.

The routine is non-blocking and therefore returns immediately, however the actual send operation may be delayed. The send operation is considered completed when it is safe to reuse the source *buffer*. If the send operation is completed immediately the routine returns UCS_OK.

If the operation is **not** completed immediately and no error reported then the UCP library will fill a user provided *req* and return UCS_INPROGRESS status. In order to monitor completion of the operation ucp_request_check_status() should be used.

Following pseudo code implements a blocking send function:

```
MPI_send(...)
{
    char *request;
    ucs_status_t status;

    // allocate request on the stack
    // ucp_context_query() was used to get ucp_request_size
    request = alloca(ucp_request_size);

     // note: make sure that there is enough memory before the
     // request handle
     status = ucp_tag_send_nbr(ep, ..., request + ucp_request_size);
     if (status != UCS_INPROGRESS) {
```

```
        return status;
    }

    do {
        ucp_worker_progress(worker);
        status = ucp_request_check_status(request + ucp_request_size);
    } while (status == UCS_INPROGRESS);

    return status;
}
```

Note

The user should not modify any part of the *buffer* after this operation is called, until the operation completes.

Parameters

in	*ep*	Destination endpoint handle.
in	*buffer*	Pointer to the message buffer (payload).
in	*count*	Number of elements to send
in	*datatype*	Datatype descriptor for the elements in the buffer.
in	*tag*	Message tag.
in	*req*	Request handle allocated by the user. There should be at least UCP request size bytes of available space before the *req*. The size of UCP request can be obtained by ucp_context_query function.

Returns

UCS_OK - The send operation was completed immediately.
UCS_INPROGRESS - The send was not completed and is in progress. ucp_request_check_status() should be used to monitor *req* status.
Error code as defined by ucs_status_t

6.7.5.4 ucs_status_ptr_t ucp_tag_send_sync_nb (ucp_ep_h *ep*, const void ∗ *buffer*, size_t *count*, ucp_datatype_t *datatype*, ucp_tag_t *tag*, ucp_send_callback_t *cb*)

Same as ucp_tag_send_nb, except the request completes only after there is a remote tag match on the message (which does not always mean the remote receive has been completed). This function never completes "in-place", and always returns a request handle.

Note

The user should not modify any part of the *buffer* after this operation is called, until the operation completes.
Returns UCS_ERR_UNSUPPORTED if UCP_ERR_HANDLING_MODE_PEER is enabled. This is a temporary implementation-related constraint that will be addressed in future releases.

Parameters

in	*ep*	Destination endpoint handle.
in	*buffer*	Pointer to the message buffer (payload).
in	*count*	Number of elements to send
in	*datatype*	Datatype descriptor for the elements in the buffer.
in	*tag*	Message tag.
in	*cb*	Callback function that is invoked whenever the send operation is completed.

Returns

UCS_PTR_IS_ERR(_ptr) - The send operation failed.
otherwise - Operation was scheduled for send and can be completed in any point in time. The request handle is returned to the application in order to track progress of the message. The application is responsible to release the handle using ucp_request_free() routine.

6.7.5.5 ucs_status_ptr_t ucp_stream_recv_nb (ucp_ep_h *ep*, void * *buffer*, size_t *count*, ucp_datatype_t *datatype*, ucp_stream_recv_callback_t *cb*, size_t * *length*, unsigned *flags*)

This routine receives data that is described by the local address *buffer*, size *count*, and *datatype* object on the endpoint *ep*. The routine is non-blocking and therefore returns immediately. The receive operation is considered complete when the message is delivered to the buffer. If data is not immediately available, the operation will be scheduled for receive and a request handle will be returned. In order to notify the application about completion of a scheduled receive operation, the UCP library will invoke the call-back *cb* when data is in the receive buffer and ready for application access. If the receive operation cannot be started, the routine returns an error.

Parameters

in	*ep*	UCP endpoint that is used for the receive operation.
in	*buffer*	Pointer to the buffer to receive the data to.
in	*count*	Number of elements to receive into *buffer*.
in	*datatype*	Datatype descriptor for the elements in the buffer.
in	*cb*	Callback function that is invoked whenever the receive operation is completed and the data is ready in the receive *buffer*. It is important to note that the call-back is only invoked in a case when the operation cannot be completed immediately.
out	*length*	Size of the received data in bytes. The value is valid only if return code is UCS_OK.

Note

The amount of data received, in bytes, is always an integral multiple of the *datatype* size.

Parameters

in	*flags*	Flags defined in ucp_stream_recv_flags_t.

Returns

UCS_OK - The receive operation was completed immediately.
UCS_PTR_IS_ERR(_ptr) - The receive operation failed.
otherwise - Operation was scheduled for receive. A request handle is returned to the application in order to track progress of the operation. The application is responsible for releasing the handle by calling the ucp_request_free routine.

6.7.5.6 ucs_status_ptr_t ucp_stream_recv_data_nb (ucp_ep_h *ep*, size_t * *length*)

This routine receives any available data from endpoint *ep*. Unlike ucp_stream_recv_nb, the returned data is unstructured and is treated as an array of bytes. If data is immediately available, UCS_STATUS_PTR(_ptr) is returned as a pointer to the data, and *length* is set to the size of the returned data buffer. The routine is non-blocking and therefore returns immediately.

Parameters

in	*ep*	UCP endpoint that is used for the receive operation.
out	*length*	Length of received data.

Returns

UCS_OK - No received data available on the *ep*.
UCS_PTR_IS_ERR(_ptr) - the receive operation failed and UCS_PTR_STATUS(_ptr) indicates an error.
otherwise - The pointer to the data UCS_STATUS_PTR(_ptr) is returned to the application. After the data is processed, the application is responsible for releasing the data buffer by calling the ucp_stream_data_release routine.

Note

This function returns packed data (equivalent to ucp_dt_make_contig(1)).
This function returns a pointer to a UCP-supplied buffer, whereas ucp_stream_recv_nb places the data into a user-provided buffer. In some cases, receiving data directly into a UCP-supplied buffer can be more optimal, for example by processing the incoming data in-place and thus avoiding extra memory copy operations.

6.7.5.7 ucs_status_ptr_t ucp_tag_recv_nb (ucp_worker_h *worker*, void ∗ *buffer*, size_t *count*, ucp_datatype_t *datatype*, ucp_tag_t *tag*, ucp_tag_t *tag_mask*, ucp_tag_recv_callback_t *cb*)

This routine receives a messages that is described by the local address *buffer*, size *count*, and *datatype* object on the *worker*. The tag value of the receive message has to match the *tag* and *tag_mask* values, where the *tag_mask* indicates what bits of the tag have to be matched. The routine is a non-blocking and therefore returns immediately. The receive operation is considered completed when the message is delivered to the *buffer*. In order to notify the application about completion of the receive operation the UCP library will invoke the call-back *cb* when the received message is in the receive buffer and ready for application access. If the receive operation cannot be stated the routine returns an error.

Note

This routine cannot return UCS_OK. It always returns a request handle or an error.

Parameters

in	*worker*	UCP worker that is used for the receive operation.
in	*buffer*	Pointer to the buffer to receive the data to.
in	*count*	Number of elements to receive
in	*datatype*	Datatype descriptor for the elements in the buffer.
in	*tag*	Message tag to expect.
in	*tag_mask*	Bit mask that indicates the bits that are used for the matching of the incoming tag against the expected tag.
in	*cb*	Callback function that is invoked whenever the receive operation is completed and the data is ready in the receive *buffer*.

Returns

UCS_PTR_IS_ERR(_ptr) - The receive operation failed.
otherwise - Operation was scheduled for receive. The request handle is returned to the application in order to track progress of the operation. The application is responsible to released the handle using ucp_request_free() routine.

6.7.5.8 ucs_status_t ucp_tag_recv_nbr (ucp_worker_h *worker*, void ∗ *buffer*, size_t *count*, ucp_datatype_t *datatype*, ucp_tag_t *tag*, ucp_tag_t *tag_mask*, void ∗ *req*)

This routine receives a message that is described by the local address *buffer*, size *count*, and *datatype* object on the *worker*. The tag value of the receive message has to match the *tag* and *tag_mask* values, where the *tag_mask* indicates what bits of the tag have to be matched. The routine is a non-blocking and therefore returns immediately. The receive operation is considered completed when the message is delivered to the *buffer*. In order to monitor completion of the operation ucp_request_check_status or ucp_tag_recv_request_test should be used.

Parameters

in	*worker*	UCP worker that is used for the receive operation.

in	*buffer*	Pointer to the buffer to receive the data to.
in	*count*	Number of elements to receive
in	*datatype*	Datatype descriptor for the elements in the buffer.
in	*tag*	Message tag to expect.
in	*tag_mask*	Bit mask that indicates the bits that are used for the matching of the incoming tag against the expected tag.
in	*req*	Request handle allocated by the user. There should be at least UCP request size bytes of available space before the *req*. The size of UCP request can be obtained by ucp_context_query function.

Returns

Error code as defined by ucs_status_t

6.7.5.9 ucp_tag_message_h ucp_tag_probe_nb (ucp_worker_h *worker,* ucp_tag_t *tag,* ucp_tag_t *tag_mask,* int *remove,* ucp_tag_recv_info_t * *info*)

This routine probes (checks) if a messages described by the *tag* and *tag_mask* was received (fully or partially) on the *worker*. The tag value of the received message has to match the *tag* and *tag_mask* values, where the *tag_mask* indicates what bits of the tag have to be matched. The function returns immediately and if the message is matched it returns a handle for the message.

Parameters

in	*worker*	UCP worker that is used for the probe operation.
in	*tag*	Message tag to probe for.
in	*tag_mask*	Bit mask that indicates the bits that are used for the matching of the incoming tag against the expected tag.
in	*remove*	The flag indicates if the matched message has to be removed from UCP library. If true (1), the message handle is removed from the UCP library and the application is responsible to call ucp_tag_msg_recv_nb() in order to receive the data and release the resources associated with the message handle. If false (0), the return value is merely an indication to whether a matching message is present, and it cannot be used in any other way, and in particular it cannot be passed to ucp_tag_msg_recv_nb().
out	*info*	If the matching message is found the descriptor is filled with the details about the message.

Returns

NULL - No match found.
Message handle (not NULL) - If message is matched the message handle is returned.

Note

This function does not advance the communication state of the network. If this routine is used in busy-poll mode, need to make sure ucp_worker_progress() is called periodically to extract messages from the transport.

Examples:

ucp_hello_world.c.

6.7.5.10 ucs_status_ptr_t ucp_tag_msg_recv_nb (ucp_worker_h *worker,* void * *buffer,* size_t *count,* ucp_datatype_t *datatype,* ucp_tag_message_h *message,* ucp_tag_recv_callback_t *cb*)

This routine receives a messages that is described by the local address *buffer*, size *count*, *message* handle, and *datatype* object on the *worker*. The *message* handle can be obtain by calling the ucp_tag_probe_nb() routine.

ucp_tag_msg_recv_nb() routine is a non-blocking and therefore returns immediately. The receive operation is considered completed when the message is delivered to the *buffer*. In order to notify the application about completion of the receive operation the UCP library will invoke the call-back *cb* when the received message is in the receive buffer and ready for application access. If the receive operation cannot be stated the routine returns an error.

Parameters

in	*worker*	UCP worker that is used for the receive operation.
in	*buffer*	Pointer to the buffer to receive the data to.
in	*count*	Number of elements to receive
in	*datatype*	Datatype descriptor for the elements in the buffer.
in	*message*	Message handle.
in	*cb*	Callback function that is invoked whenever the receive operation is completed and the data is ready in the receive *buffer*.

Returns

UCS_PTR_IS_ERR(_ptr) - The receive operation failed.
otherwise - Operation was scheduled for receive. The request handle is returned to the application in order to track progress of the operation. The application is responsible to released the handle using ucp_request_free() routine.

Examples:

ucp_hello_world.c.

6.7.5.11 ucs_status_t ucp_put_nbi (ucp_ep_h *ep,* const void ∗ *buffer,* size_t *length,* uint64_t *remote_addr,* ucp_rkey_h *rkey*)

This routine initiates a storage of contiguous block of data that is described by the local address *buffer* in the remote contiguous memory region described by *remote_addr* address and the memoryhandle" *rkey*. The routine returns immediately and **does not** guarantee re-usability of the source address *buffer*. If the operation is completed immediately the routine return UCS_OK, otherwise UCS_INPROGRESS or an error is returned to user.

Note

A user can use ucp_worker_flush_nb() in order to guarantee re-usability of the source address *buffer*.

Parameters

in	*ep*	Remote endpoint handle.
in	*buffer*	Pointer to the local source address.
in	*length*	Length of the data (in bytes) stored under the source address.
in	*remote_addr*	Pointer to the destination remote memory address to write to.
in	*rkey*	Remote memory key associated with the remote memory address.

Returns

Error code as defined by ucs_status_t

6.7.5.12 ucs_status_ptr_t ucp_put_nb (ucp_ep_h *ep,* const void ∗ *buffer,* size_t *length,* uint64_t *remote_addr,* ucp_rkey_h *rkey,* ucp_send_callback_t *cb*)

This routine initiates a storage of contiguous block of data that is described by the local address *buffer* in the remote contiguous memory region described by *remote_addr* address and the memoryhandle" *rkey*. The routine returns immediately and **does not** guarantee re-usability of the source address *buffer*. If the operation is completed immediately the routine return UCS_OK, otherwise UCS_INPROGRESS or an error is returned to user. If the put operation completes immediately, the routine returns UCS_OK and the call-back routine *cb* is **not** invoked. If the operation is **not** completed immediately and no error is reported, then the UCP library will schedule invocation of the call-back routine *cb* upon completion of the put operation. In other words, the completion of a put operation can be signaled by the return code or execution of the call-back.

Note

A user can use ucp_worker_flush_nb() in order to guarantee re-usability of the source address *buffer*.

Parameters

in	*ep*	Remote endpoint handle.
in	*buffer*	Pointer to the local source address.
in	*length*	Length of the data (in bytes) stored under the source address.
in	*remote_addr*	Pointer to the destination remote memory address to write to.
in	*rkey*	Remote memory key associated with the remote memory address.
in	*cb*	Call-back function that is invoked whenever the put operation is completed and the local buffer can be modified. Does not guarantee remote completion.

Returns

UCS_OK - The operation was completed immediately.
UCS_PTR_IS_ERR(_ptr) - The operation failed.
otherwise - Operation was scheduled and can be completed at any point in time. The request handle is returned to the application in order to track progress of the operation. The application is responsible for releasing the handle using ucp_request_free() routine.

6.7.5.13 ucs_status_t ucp_get_nbi (ucp_ep_h *ep*, void ∗ *buffer*, size_t *length*, uint64_t *remote_addr*, ucp_rkey_h *rkey*)

This routine initiate a load of contiguous block of data that is described by the remote memory address *remote_addr* and the memory handle *rkey* in the local contiguous memory region described by *buffer* address. The routine returns immediately and **does not** guarantee that remote data is loaded and stored under the local address *buffer*.

Note

A user can use ucp_worker_flush_nb() in order guarantee that remote data is loaded and stored under the local address *buffer*.

Parameters

in	*ep*	Remote endpoint handle.
in	*buffer*	Pointer to the local source address.
in	*length*	Length of the data (in bytes) stored under the source address.
in	*remote_addr*	Pointer to the destination remote memory address to write to.
in	*rkey*	Remote memory key associated with the remote memory address.

Returns

Error code as defined by ucs_status_t

6.7.5.14 ucs_status_ptr_t ucp_get_nb (ucp_ep_h *ep*, void ∗ *buffer*, size_t *length*, uint64_t *remote_addr*, ucp_rkey_h *rkey*, ucp_send_callback_t *cb*)

This routine initiates a load of a contiguous block of data that is described by the remote memory address *remote_addr* and the memory handle *rkey* in the local contiguous memory region described by *buffer* address. The routine returns immediately and **does not** guarantee that remote data is loaded and stored under the local address *buffer*. If the operation is completed immediately the routine return UCS_OK, otherwise UCS_INPROGRESS or an error is returned to user. If the get operation completes immediately, the routine returns UCS_OK and the call-back routine *cb* is **not** invoked. If the operation is **not** completed immediately and no error is reported, then the UCP library will schedule invocation of the call-back routine *cb* upon completion of the get operation. In other words, the completion of a get operation can be signaled by the return code or execution of the call-back.

Note

A user can use ucp_worker_flush_nb() in order to guarantee re-usability of the source address *buffer*.

Parameters

in	*ep*	Remote endpoint handle.
in	*buffer*	Pointer to the local source address.
in	*length*	Length of the data (in bytes) stored under the source address.
in	*remote_addr*	Pointer to the destination remote memory address to write to.
in	*rkey*	Remote memory key associated with the remote memory address.
in	*cb*	Call-back function that is invoked whenever the get operation is completed and the data is visible to the local process.

Returns

UCS_OK - The operation was completed immediately.
UCS_PTR_IS_ERR(_ptr) - The operation failed.
otherwise - Operation was scheduled and can be completed at any point in time. The request handle is returned to the application in order to track progress of the operation. The application is responsible for releasing the handle using ucp_request_free() routine.

6.7.5.15 ucs_status_t ucp_atomic_post (ucp_ep_h *ep,* ucp_atomic_post_op_t *opcode,* uint64_t *value,* size_t *op_size,* uint64_t *remote_addr,* ucp_rkey_h *rkey*)

This routine posts an atomic memory operation to a remote value. The remote value is described by the combination of the remote memory address *remote_addr* and the remote memory handle *rkey*. Return from the function does not guarantee completion. A user must call ucp_ep_flush_nb or ucp_worker_flush_nb to guarantee that the remote value has been updated.

Parameters

in	*ep*	UCP endpoint.
in	*opcode*	One of ucp_atomic_post_op_t.
in	*value*	Source operand for the atomic operation.
in	*op_size*	Size of value in bytes
in	*remote_addr*	Remote address to operate on.
in	*rkey*	Remote key handle for the remote memory address.

Returns

Error code as defined by ucs_status_t

6.7.5.16 ucs_status_ptr_t ucp_atomic_fetch_nb (ucp_ep_h *ep,* ucp_atomic_fetch_op_t *opcode,* uint64_t *value,* void ∗ *result,* size_t *op_size,* uint64_t *remote_addr,* ucp_rkey_h *rkey,* ucp_send_callback_t *cb*)

This routine will post an atomic fetch operation to remote memory. The remote value is described by the combination of the remote memory address *remote_addr* and the remote memory handle *rkey*. The routine is non-blocking and therefore returns immediately. However the actual atomic operation may be delayed. The atomic operation is not considered complete until the values in remote and local memory are completed. If the atomic operation completes immediately, the routine returns UCS_OK and the call-back routine *cb* is **not** invoked. If the operation is **not** completed immediately and no error is reported, then the UCP library will schedule invocation of the call-back routine *cb* upon completion of the atomic operation. In other words, the completion of an atomic operation can be signaled by the return code or execution of the call-back.

Note

The user should not modify any part of the *result* after this operation is called, until the operation completes.

Parameters

in	*ep*	UCP endpoint.
in	*opcode*	One of ucp_atomic_fetch_op_t.
in	*value*	Source operand for atomic operation. In the case of CSWAP this is the conditional for the swap. For SWAP this is the value to be placed in remote memory.
in,out	*result*	Local memory address to store resulting fetch to. In the case of CSWAP the value in result will be swapped into the *remote_addr* if the condition is true.
in	*op_size*	Size of value in bytes and pointer type for result
in	*remote_addr*	Remote address to operate on.
in	*rkey*	Remote key handle for the remote memory address.
in	*cb*	Call-back function that is invoked whenever the send operation is completed. It is important to note that the call-back function is only invoked in a case when the operation cannot be completed in place.

Returns

UCS_OK - The operation was completed immediately.
UCS_PTR_IS_ERR(_ptr) - The operation failed.
otherwise - Operation was scheduled and can be completed at any point in time. The request handle is returned to the application in order to track progress of the operation. The application is responsible for releasing the handle using ucp_request_free() routine.

6.7.5.17 ucs_status_t ucp_request_check_status (void ∗ *request*)

This routine checks the state of the request and returns its current status. Any value different from UCS_INPROGRESS means that request is in a completed state.

Parameters

in	*request*	Non-blocking request to check.

Returns

Error code as defined by ucs_status_t

Examples:

ucp_hello_world.c.

6.7.5.18 ucs_status_t ucp_tag_recv_request_test (void ∗ *request*, ucp_tag_recv_info_t ∗ *info*)

This routine checks the state and returns current status of the request returned from ucp_tag_recv_nb routine or the user allocated request for ucp_tag_recv_nbr. Any value different from UCS_INPROGRESS means that the request is in a completed state.

Parameters

in	*request*	Non-blocking request to check.
out	*info*	It is filled with the details about the message available at the moment of calling.

Returns

Error code as defined by ucs_status_t

6.7.5.19 ucs_status_t ucp_stream_recv_request_test (void * *request,* size_t * *length_p*)

This routine checks the state and returns current status of the request returned from ucp_stream_recv_nb routine. Any value different from UCS_INPROGRESS means that the request is in a completed state.

Parameters

in	*request*	Non-blocking request to check.
out	*length_p*	The size of the received data in bytes. This value is only valid if the status is UCS_OK. If valid, it is always an integral multiple of the datatype size associated with the request.

Returns

Error code as defined by ucs_status_t

6.7.5.20 void ucp_request_cancel (ucp_worker_h *worker,* void * *request*)

Parameters

in	*worker*	UCP worker.
in	*request*	Non-blocking request to cancel.

This routine tries to cancels an outstanding communication request. After calling this routine, the *request* will be in completed or canceled (but not both) state regardless of the status of the target endpoint associated with the communication request. If the request is completed successfully, the send or receive completion callbacks (based on the type of the request) will be called with the *status* argument of the callback set to UCS_OK, and in a case it is canceled the *status* argument is set to UCS_ERR_CANCELED. It is important to note that in order to release the request back to the library the application is responsible to call ucp_request_free().

6.7.5.21 void ucp_stream_data_release (ucp_ep_h *ep,* void * *data*)

Parameters

in	*ep*	Endpoint *data* received from.
in	*data*	Data pointer to release, which was returned from ucp_stream_recv_data_nb.

This routine releases internal UCP data buffer returned by ucp_stream_recv_data_nb when *data* is processed, the application can't use this buffer after calling this function.

6.7.5.22 void ucp_request_free (void * *request*)

Parameters

in	*request*	Non-blocking request to release.

This routine releases the non-blocking request back to the library, regardless of its current state. Communications operations associated with this request will make progress internally, however no further notifications or callbacks would be invoked for this request.

6.7.5.23 int ucp_request_is_completed (void ∗ *request*)

Deprecated Replaced by ucp_request_test.

6.7.5.24 ucs_status_t ucp_put (ucp_ep_h *ep*, const void ∗ *buffer*, size_t *length*, uint64_t *remote_addr*, ucp_rkey_h *rkey*)

Deprecated Replaced by ucp_put_nb. The following example implements the same functionality using ucp_put_nb :

```
void empty_callback(void *request, ucs_status_t status)
{
}

ucs_status_t put(ucp_ep_h ep, const void *buffer, size_t length,
                  uint64_t remote_addr, ucp_rkey_h rkey)
{
    void *request = ucp_put_nb(ep, buffer, length, remote_addr, rkey,
                               empty_callback),
    if (request == NULL) {
        return UCS_OK;
    } else if (UCS_PTR_IS_ERR(request)) {
        return UCS_PTR_STATUS(request);
    } else {
        ucs_status_t status;
        do {
            ucp_worker_progress(worker);
            status = ucp_request_check_status(request);
        } while (status == UCS_INPROGRESS);
        ucp_request_release(request);
        return status;
    }
}
```

This routine stores contiguous block of data that is described by the local address *buffer* in the remote contiguous memory region described by *remote_addr* address and the memory handle *rkey*. The routine returns when it is safe to reuse the source address *buffer*.

Parameters

in	*ep*	Remote endpoint handle.
in	*buffer*	Pointer to the local source address.
in	*length*	Length of the data (in bytes) stored under the source address.
in	*remote_addr*	Pointer to the destination remote address to write to.
in	*rkey*	Remote memory key associated with the remote address.

Returns

Error code as defined by ucs_status_t

6.7.5.25 ucs_status_t ucp_get (ucp_ep_h *ep*, void ∗ *buffer*, size_t *length*, uint64_t *remote_addr*, ucp_rkey_h *rkey*)

Deprecated Replaced by ucp_get_nb.

See also

ucp_put.

This routine loads contiguous block of data that is described by the remote address *remote_addr* and the memory handle *rkey* in the local contiguous memory region described by *buffer* address. The routine returns when remote data is loaded and stored under the local address *buffer*.

Parameters

in	*ep*	Remote endpoint handle.
in	*buffer*	Pointer to the local source address.
in	*length*	Length of the data (in bytes) stored under the source address.
in	*remote_addr*	Pointer to the destination remote address to write to.
in	*rkey*	Remote memory key associated with the remote address.

Returns

Error code as defined by ucs_status_t

6.7.5.26 ucs_status_t ucp_atomic_add32 (ucp_ep_h *ep*, uint32_t *add*, uint64_t *remote_addr*, ucp_rkey_h *rkey*)

Deprecated Replaced by ucp_atomic_post with opcode UCP_ATOMIC_POST_OP_ADD.

See also

ucp_put.

This routine performs an add operation on a 32 bit integer value atomically. The remote integer value is described by the combination of the remote memory address *remote_addr* and the remote memory handle *rkey*. The *add* value is the value that is used for the add operation. When the operation completes the sum of the original remote value and the operand value (*add*) is stored in remote memory. The call to the routine returns immediately, independent of operation completion.

Note

The remote address must be aligned to 32 bit.

Parameters

in	*ep*	Remote endpoint handle.
in	*add*	Value to add.
in	*remote_addr*	Pointer to the destination remote address of the atomic variable.
in	*rkey*	Remote memory key associated with the remote address.

Returns

Error code as defined by ucs_status_t

6.7.5.27 ucs_status_t ucp_atomic_add64 (ucp_ep_h *ep*, uint64_t *add*, uint64_t *remote_addr*, ucp_rkey_h *rkey*)

Deprecated Replaced by ucp_atomic_post with opcode UCP_ATOMIC_POST_OP_ADD.

See also

ucp_put.

This routine performs an add operation on a 64 bit integer value atomically. The remote integer value is described by the combination of the remote memory address *remote_addr* and the remote memory handle *rkey*. The *add* value is the value that is used for the add operation. When the operation completes the sum of the original remote value and the operand value (*add*) is stored in remote memory. The call to the routine returns immediately, independent of operation completion.

Note

The remote address must be aligned to 64 bit.

Parameters

in	*ep*	Remote endpoint handle.
in	*add*	Value to add.
in	*remote_addr*	Pointer to the destination remote address of the atomic variable.
in	*rkey*	Remote memory key associated with the remote address.

Returns

Error code as defined by ucs_status_t

6.7.5.28 ucs_status_t ucp_atomic_fadd32 (ucp_ep_h *ep*, uint32_t *add*, uint64_t *remote_addr*, ucp_rkey_h *rkey*, uint32_t ∗ *result*)

Deprecated Replaced by ucp_atomic_fetch_nb with opcode UCP_ATOMIC_FETCH_OP_FADD.

See also

ucp_put.

This routine performs an add operation on a 32 bit integer value atomically. The remote integer value is described by the combination of the remote memory address *remote_addr* and the remote memory handle *rkey*. The *add* value is the value that is used for the add operation. When the operation completes, the original remote value is stored in the local memory *result*, and the sum of the original remote value and the operand value is stored in remote memory. The call to the routine returns when the operation is completed and the *result* value is updated.

Note

The remote address must be aligned to 32 bit.

Parameters

in	*ep*	Remote endpoint handle.
in	*add*	Value to add.
in	*remote_addr*	Pointer to the destination remote address of the atomic variable.
in	*rkey*	Remote memory key associated with the remote address.
out	*result*	Pointer to the address that is used to store the previous value of the atomic variable described by the *remote_addr*

Returns

Error code as defined by ucs_status_t

6.7.5.29 ucs_status_t ucp_atomic_fadd64 (ucp_ep_h *ep*, uint64_t *add*, uint64_t *remote_addr*, ucp_rkey_h *rkey*, uint64_t ∗ *result*)

Deprecated Replaced by ucp_atomic_fetch_nb with opcode UCP_ATOMIC_FETCH_OP_FADD.

See also

ucp_put.

This routine performs an add operation on a 64 bit integer value atomically. The remote integer value is described by the combination of the remote memory address *remote_addr* and the remote memory handle *rkey*. The *add* value is the value that is used for the add operation. When the operation completes, the original remote value is stored in the local memory *result*, and the sum of the original remote value and the operand value is stored in remote memory. The call to the routine returns when the operation is completed and the *result* value is updated.

Note

The remote address must be aligned to 64 bit.

Parameters

`in`	*ep*	Remote endpoint handle.
`in`	*add*	Value to add.
`in`	*remote_addr*	Pointer to the destination remote address of the atomic variable.
`in`	*rkey*	Remote memory key associated with the remote address.
`out`	*result*	Pointer to the address that is used to store the previous value of the atomic variable described by the *remote_addr*

Returns

Error code as defined by ucs_status_t

6.7.5.30 ucs_status_t ucp_atomic_swap32 (ucp_ep_h *ep*, uint32_t *swap*, uint64_t *remote_addr*, ucp_rkey_h *rkey*, uint32_t ∗ *result*)

Deprecated Replaced by ucp_atomic_fetch_nb with opcode UCP_ATOMIC_FETCH_OP_SWAP.

See also

ucp_put.

This routine swaps a 32 bit value between local and remote memory. The remote value is described by the combination of the remote memory address *remote_addr* and the remote memory handle *rkey*. The *swap* value is the value that is used for the swap operation. When the operation completes, the remote value is stored in the local memory *result*, and the operand value (*swap*) is stored in remote memory. The call to the routine returns when the operation is completed and the *result* value is updated.

Note

The remote address must be aligned to 32 bit.

Parameters

`in`	*ep*	Remote endpoint handle.
`in`	*swap*	Value to swap.
`in`	*remote_addr*	Pointer to the destination remote address of the atomic variable.
`in`	*rkey*	Remote memory key associated with the remote address.
`out`	*result*	Pointer to the address that is used to store the previous value of the atomic variable described by the *remote_addr*

Returns

Error code as defined by ucs_status_t

6.7.5.31 ucs_status_t ucp_atomic_swap64 (ucp_ep_h *ep*, uint64_t *swap*, uint64_t *remote_addr*, ucp_rkey_h *rkey*, uint64_t ∗ *result*)

Deprecated Replaced by ucp_atomic_fetch_nb with opcode UCP_ATOMIC_FETCH_OP_SWAP.

See also

ucp_put.

This routine swaps a 64 bit value between local and remote memory. The remote value is described by the combination of the remote memory address *remote_addr* and the remote memory handle *rkey*. The *swap* value is the value that is used for the swap operation. When the operation completes, the remote value is stored in the local memory *result*, and the operand value (*swap*) is stored in remote memory. The call to the routine returns when the operation is completed and the *result* value is updated.

Note

The remote address must be aligned to 64 bit.

Parameters

`in`	*ep*	Remote endpoint handle.
`in`	*swap*	Value to swap.
`in`	*remote_addr*	Pointer to the destination remote address of the atomic variable.
`in`	*rkey*	Remote memory key associated with the remote address.
`out`	*result*	Pointer to the address that is used to store the previous value of the atomic variable described by the *remote_addr*

Returns

Error code as defined by ucs_status_t

6.7.5.32 ucs_status_t ucp_atomic_cswap32 (ucp_ep_h *ep*, uint32_t *compare*, uint32_t *swap*, uint64_t *remote_addr*, ucp_rkey_h *rkey*, uint32_t ∗ *result*)

Deprecated Replaced by ucp_atomic_fetch_nb with opcode UCP_ATOMIC_FETCH_OP_CSWAP.

See also

ucp_put.

This routine conditionally swaps a 32 bit value between local and remote memory. The swap occurs only if the condition value (*continue*) is equal to the remote value, otherwise the remote memory is not modified. The remote value is described by the combination of the remote memory address `remote_addr` and the remote memory handle *rkey*. The `swap` value is the value that is used to update the remote memory if the condition is true. The call to the routine returns when the operation is completed and the *result* value is updated.

Note

The remote address must be aligned to 32 bit.

Parameters

`in`	*ep*	Remote endpoint handle.
`in`	*compare*	Value to compare to.
`in`	*swap*	Value to swap.
`in`	*remote_addr*	Pointer to the destination remote address of the atomic variable.
`in`	*rkey*	Remote memory key associated with the remote address.
`out`	*result*	Pointer to the address that is used to store the previous value of the atomic variable described by the *remote_addr*

Returns

Error code as defined by ucs_status_t

6.7.5.33 ucs_status_t ucp_atomic_cswap64 (ucp_ep_h *ep*, uint64_t *compare*, uint64_t *swap*, uint64_t *remote_addr*, ucp_rkey_h *rkey*, uint64_t ∗ *result*)

Deprecated Replaced by ucp_atomic_fetch_nb with opcode UCP_ATOMIC_FETCH_OP_CSWAP.

See also

ucp_put.

This routine conditionally swaps a 64 bit value between local and remote memory. The swap occurs only if the condition value (*continue*) is equal to the remote value, otherwise the remote memory is not modified. The remote value is described by the combination of the remote memory address `remote_addr` and the remote memory handle *rkey*. The `swap` value is the value that is used to update the remote memory if the condition is true. The call to the routine returns when the operation is completed and the *result* value is updated.

Note

The remote address must be aligned to 64 bit.

Parameters

`in`	*ep*	Remote endpoint handle.
`in`	*compare*	Value to compare to.
`in`	*swap*	Value to swap.
`in`	*remote_addr*	Pointer to the destination remote address of the atomic variable.
`in`	*rkey*	Remote memory key associated with the remote address.
`out`	*result*	Pointer to the address that is used to store the previous value of the atomic variable described by the *remote_addr*

Returns

Error code as defined by ucs_status_t

6.8 UCP Configuration

Data Structures

- struct ucp_params

  Tuning parameters for UCP library. More...

Typedefs

- typedef struct ucp_params ucp_params_t

  Tuning parameters for UCP library.
- typedef struct ucp_config ucp_config_t

  UCP configuration descriptor.

Functions

- ucs_status_t ucp_config_read (const char ∗env_prefix, const char ∗filename, ucp_config_t ∗∗config_p)

  Read UCP configuration descriptor.
- void ucp_config_release (ucp_config_t ∗config)

  Release configuration descriptor.
- ucs_status_t ucp_config_modify (ucp_config_t ∗config, const char ∗name, const char ∗value)

  Modify context configuration.
- void ucp_config_print (const ucp_config_t ∗config, FILE ∗stream, const char ∗title, ucs_config_print_flags_t print_flags)

  Print configuration information.

6.8.1 Detailed Description

This section describes routines for configuration of the UCP network layer

6.8.2 Data Structure Documentation

6.8.2.1 struct ucp_params

The structure defines the parameters that are used for UCP library tuning during UCP library initialization.

Note

UCP library implementation uses the features parameter to optimize the library functionality that minimize memory footprint. For example, if the application does not require send/receive semantics UCP library may avoid allocation of expensive resources associated with send/receive queues.

Examples:

ucp_hello_world.c.

Data Fields

uint64_t	field_mask	Mask of valid fields in this structure, using bits from ucp_params_field. Fields not specified in this mask would be ignored. Provides ABI compatibility with respect to adding new fields.
uint64_t	features	UCP features that are used for library initialization. It is recommended for applications only to request the features that are required for an optimal functionality This field must be specified.
size_t	request_size	The size of a reserved space in a non-blocking requests. Typically applications use this space for caching own structures in order to avoid costly memory allocations, pointer dereferences, and cache misses. For example, MPI implementation can use this memory for caching MPI descriptors This field defaults to 0 if not specified.
ucp_request_↩ init_callback_t	request_init	Pointer to a routine that is used for the request initialization. This function will be called only on the very first time a request memory is initialized, and may not be called again if a request is reused. If a request should be reset before the next reuse, it can be done before calling ucp_request↩ _free. *NULL* can be used if no such is function required, which is also the default if this field is not specified by field_mask.
ucp_request_↩ cleanup_↩ callback_t	request_cleanup	Pointer to a routine that is responsible for final cleanup of the memory associated with the request. This routine may not be called every time a request is released. For some implementations, the cleanup call may be delayed and only invoked at ucp_worker_destroy. *NULL* can be used if no such function is required, which is also the default if this field is not specified by field_mask.
uint64_t	tag_sender_↩ mask	Mask which specifies particular bits of the tag which can uniquely identify the sender (UCP endpoint) in tagged operations. This field defaults to 0 if not specified.
int	mt_workers_↩ shared	This flag indicates if this context is shared by multiple workers from different threads. If so, this context needs thread safety support; otherwise, the context does not need to provide thread safety. For example, if the context is used by single worker, and that worker is shared by multiple threads, this context does not need thread safety; if the context is used by worker 1 and worker 2, and worker 1 is used by thread 1 and worker 2 is used by thread 2, then this context needs thread safety. Note that actual thread mode may be different from mode passed to ucp_init. To get actual thread mode use ucp_context_query.
size_t	estimated_↩ num_eps	An optimization hint of how many endpoints would be created on this context. For example, when used from MPI or SHMEM libraries, this number would specify the number of ranks (or processing elements) in the job. Does not affect semantics, but only transport selection criteria and the resulting performance. The value can be also set by UCX_NU↩ M_EPS environment variable. In such case it will override the number of endpoints set by *estimated_num_eps*

6.8.3 Typedef Documentation

6.8.3.1 typedef struct ucp_params ucp_params_t

The structure defines the parameters that are used for UCP library tuning during UCP library initialization.

Note

UCP library implementation uses the features parameter to optimize the library functionality that minimize memory footprint. For example, if the application does not require send/receive semantics UCP library may avoid allocation of expensive resources associated with send/receive queues.

6.8.3.2 typedef struct ucp_config ucp_config_t

This descriptor defines the configuration for UCP application context. The configuration is loaded from the run-time environment (using configuration files of environment variables) using ucp_config_read routine and can be printed using ucp_config_print routine. In addition, application is responsible to release the descriptor using ucp_config_↩ release routine.

6.8.4 Function Documentation

6.8.4.1 ucs_status_t ucp_config_read (const char ∗ *env_prefix*, const char ∗ *filename*, ucp_config_t ∗∗ *config_p*)

The routine fetches the information about UCP library configuration from the run-time environment. Then, the fetched descriptor is used for UCP library initialization. The Application can print out the descriptor using print routine. In addition the application is responsible to release the descriptor back to UCP library.

Parameters

in	*env_prefix*	If non-NULL, the routine searches for the environment variables that start with *UCX_<env_prefix>_* prefix. Otherwise, the routine searches for the environment variables that start with *UCX_* prefix.
in	*filename*	If non-NULL, read configuration from the file defined by *filename*. If the file does not exist, it will be ignored and no error reported to the application.
out	*config_p*	Pointer to configuration descriptor as defined by ucp_config_t.

Returns

Error code as defined by ucs_status_t

Examples:

ucp_hello_world.c.

6.8.4.2 void ucp_config_release (ucp_config_t ∗ *config*)

The routine releases the configuration descriptor that was allocated through ucp_config_read() routine.

Parameters

out	*config*	Configuration descriptor as defined by ucp_config_t.

Examples:

ucp_hello_world.c.

6.8.4.3 ucs_status_t ucp_config_modify (ucp_config_t * *config*, const char * *name*, const char * *value*)

The routine changes one configuration setting stored in configuration descriptor.

Parameters

in	*config*	Configuration to modify.
in	*name*	Configuration variable name.
in	*value*	Value to set.

Returns

Error code.

6.8.4.4 void ucp_config_print (const ucp_config_t * *config*, FILE * *stream*, const char * *title*, ucs_config_print_flags_t *print_flags*)

The routine prints the configuration information that is stored in configuration descriptor.

Parameters

in	*config*	Configuration descriptor to print.
in	*stream*	Output stream to print the configuration to.
in	*title*	Configuration title to print.
in	*print_flags*	Flags that control various printing options.

Examples:

ucp_hello_world.c.

6.9 UCP Data type routines

Data Structures

- struct ucp_dt_iov

  Structure for scatter-gather I/O. More...
- struct ucp_generic_dt_ops

  UCP generic data type descriptor.

Macros

- #define ucp_dt_make_contig(_elem_size) (((ucp_datatype_t)(_elem_size) << UCP_DATATYPE_SHIFT) | UCP_DATATYPE_CONTIG)

  Generate an identifier for contiguous data type.
- #define ucp_dt_make_iov() (UCP_DATATYPE_IOV)

  Generate an identifier for Scatter-gather IOV data type.

Typedefs

- typedef struct ucp_dt_iov ucp_dt_iov_t

  Structure for scatter-gather I/O.
- typedef struct ucp_generic_dt_ops ucp_generic_dt_ops_t

  UCP generic data type descriptor.

Enumerations

- enum ucp_dt_type {
  UCP_DATATYPE_CONTIG = 0, UCP_DATATYPE_STRIDED = 1, UCP_DATATYPE_IOV = 2, UCP_DATATYPE_GENERIC = 7,
  UCP_DATATYPE_SHIFT = 3, UCP_DATATYPE_CLASS_MASK = UCS_MASK(UCP_DATATYPE_SHIFT)
  }

  UCP data type classification.

Functions

- ucs_status_t ucp_dt_create_generic (const ucp_generic_dt_ops_t *ops, void *context, ucp_datatype_t *datatype_p)

  Create a generic datatype.
- void ucp_dt_destroy (ucp_datatype_t datatype)

  Destroy a datatype and release its resources.

Variables

- void *(* ucp_generic_dt_ops::start_pack)(void *context, const void *buffer, size_t count)

  Start a packing request.
- void *(* ucp_generic_dt_ops::start_unpack)(void *context, void *buffer, size_t count)

  Start an unpacking request.
- size_t(* ucp_generic_dt_ops::packed_size)(void *state)

  Get the total size of packed data.
- size_t(* ucp_generic_dt_ops::pack)(void *state, size_t offset, void *dest, size_t max_length)

Pack data.

- ucs_status_t(∗ ucp_generic_dt_ops::unpack)(void ∗state, size_t offset, const void ∗src, size_t length)

  Unpack data.

- void(∗ ucp_generic_dt_ops::finish)(void ∗state)

  Finish packing/unpacking.

6.9.1 Detailed Description

UCP Data type routines

6.9.2 Data Structure Documentation

6.9.2.1 struct ucp_dt_iov

This structure is used to specify a list of buffers which can be used within a single data transfer function call.

Note

If *length* is zero, the memory pointed to by *buffer* will not be accessed. Otherwise, *buffer* must point to valid memory.

Data Fields

void ∗	buffer	Pointer to a data buffer
size_t	length	Length of the *buffer* in bytes

6.9.3 Macro Definition Documentation

6.9.3.1 #define ucp_dt_make_contig(*_elem_size*) (((ucp_datatype_t)(_elem_size) << UCP_DATATYPE_SHIFT) | UCP_DATATYPE_CONTIG)

This macro creates an identifier for contiguous datatype that is defined by the size of the basic element.

Parameters

in	*_elem_size*	Size of the basic element of the type.

Returns

Data-type identifier.

Note

In case of partial receive, the buffer will be filled with integral count of elements.

Examples:

ucp_hello_world.c.

6.9.3.2 #define ucp_dt_make_iov( ) (UCP_DATATYPE_IOV)

This macro creates an identifier for datatype of scatter-gather list with multiple pointers

Returns

Data-type identifier.

Note

In case of partial receive, ucp_dt_iov_t::buffer can be filled with any number of bytes according to its ucp_dt↩_iov_t::length.

6.9.4 Typedef Documentation

6.9.4.1 typedef struct ucp_dt_iov ucp_dt_iov_t

This structure is used to specify a list of buffers which can be used within a single data transfer function call.

Note

If *length* is zero, the memory pointed to by *buffer* will not be accessed. Otherwise, *buffer* must point to valid memory.

6.9.4.2 typedef struct ucp_generic_dt_ops ucp_generic_dt_ops_t

This structure provides a generic datatype descriptor that is used for definition of application defined datatypes.

Typically, the descriptor is used for an integration with datatype engines implemented within MPI and SHMEM implementations.

Note

In case of partial receive, any amount of received data is acceptable which matches buffer size.

6.9.5 Enumeration Type Documentation

6.9.5.1 enum ucp_dt_type

The enumeration list describes the datatypes supported by UCP.

Enumerator

UCP_DATATYPE_CONTIG Contiguous datatype

UCP_DATATYPE_STRIDED Strided datatype

UCP_DATATYPE_IOV Scatter-gather list with multiple pointers

UCP_DATATYPE_GENERIC Generic datatype with user-defined pack/unpack routines

UCP_DATATYPE_SHIFT Number of bits defining the datatype classification

UCP_DATATYPE_CLASS_MASK Data-type class mask

6.9.6 Function Documentation

6.9.6.1 ucs_status_t ucp_dt_create_generic (const ucp_generic_dt_ops_t ∗ *ops,* void ∗ *context,* ucp_datatype_t ∗ *datatype_p*)

This routine create a generic datatype object. The generic datatype is described by the *ops* object which provides a table of routines defining the operations for generic datatype manipulation. Typically, generic datatypes are used for integration with datatype engines provided with MPI implementations (MPICH, Open MPI, etc). The application is responsible to release the *datatype_p* object using ucp_dt_destroy() routine.

Parameters

in	*ops*	Generic datatype function table as defined by ucp_generic_dt_ops_t .
in	*context*	Application defined context passed to this routine. The context is passed as a parameter to the routines in the *ops* table.
out	*datatype_p*	A pointer to datatype object.

Returns

Error code as defined by ucs_status_t

6.9.6.2 void ucp_dt_destroy (ucp_datatype_t *datatype*)

This routine destroys the *datatype* object and releases any resources that are associated with the object. The *datatype* object must be allocated using ucp_dt_create_generic() routine.

Warning

- Once the *datatype* object is released an access to this object may cause an undefined failure.

Parameters

in	*datatype*	Datatype object to destroy.

6.9.7 Variable Documentation

6.9.7.1 void∗(∗ ucp_generic_dt_ops::start_pack) (void ∗context, const void ∗buffer, size_t count)

The pointer refers to application defined start-to-pack routine. It will be called from the ucp_tag_send_nb routine.

Parameters

in	*context*	User-defined context.
in	*buffer*	Buffer to pack.
in	*count*	Number of elements to pack into the buffer.

Returns

A custom state that is passed to the following pack() routine.

6.9.7.2 void∗(∗ ucp_generic_dt_ops::start_unpack) (void ∗context, void ∗buffer, size_t count)

The pointer refers to application defined start-to-unpack routine. It will be called from the ucp_tag_recv_nb routine.

Parameters

in	*context*	User-defined context.
in	*buffer*	Buffer to unpack to.
in	*count*	Number of elements to unpack in the buffer.

Returns

A custom state that is passed later to the following unpack() routine.

6.9.7.3 size_t(∗ ucp_generic_dt_ops::packed_size) (void ∗state)

The pointer refers to user defined routine that returns the size of data in a packed format.

Parameters

in	*state*	State as returned by start_pack() routine.

Returns

The size of the data in a packed form.

6.9.7.4 size_t(∗ ucp_generic_dt_ops::pack) (void ∗state, size_t offset, void ∗dest, size_t max_length)

The pointer refers to application defined pack routine.

Parameters

in	*state*	State as returned by start_pack() routine.
in	*offset*	Virtual offset in the output stream.
in	*dest*	Destination to pack the data to.
in	*max_length*	Maximal length to pack.

Returns

The size of the data that was written to the destination buffer. Must be less than or equal to *max_length*.

6.9.7.5 ucs_status_t(∗ ucp_generic_dt_ops::unpack) (void ∗state, size_t offset, const void ∗src, size_t length)

The pointer refers to application defined unpack routine.

Parameters

in	*state*	State as returned by start_unpack() routine.
in	*offset*	Virtual offset in the input stream.
in	*src*	Source to unpack the data from.
in	*length*	Length to unpack.

Returns

UCS_OK or an error if unpacking failed.

6.9.7.6 void(∗ ucp_generic_dt_ops::finish) (void ∗state)

The pointer refers to application defined finish routine.

Parameters

in	*state*	State as returned by start_pack() and start_unpack() routines.

6.10 Unified Communication Transport (UCT) API

Modules

- UCT Communication Resource
- UCT Communication Context
- UCT Memory Domain
- UCT Active messages
- UCT Remote memory access operations
- UCT Atomic operations
- UCT Tag matching operations

6.10.1 Detailed Description

This section describes UCT API.

6.11 UCT Communication Resource

Modules

- UCT interface operations and capabilities

    List of capabilities supported by UCX API.

Data Structures

- struct uct_md_resource_desc

    Memory domain resource descriptor. More...
- struct uct_tl_resource_desc

    Communication resource descriptor. More...
- struct uct_iface_attr

    Interface attributes: capabilities and limitations. More...
- struct uct_iface_attr.cap
- struct uct_iface_attr.cap.put
- struct uct_iface_attr.cap.get
- struct uct_iface_attr.cap.am
- struct uct_iface_attr.cap.tag
- struct uct_iface_attr.cap.tag.recv
- struct uct_iface_attr.cap.tag.eager
- struct uct_iface_attr.cap.tag.rndv
- struct uct_iface_attr.cap.atomic32
- struct uct_iface_attr.cap.atomic64
- struct uct_iface_params

    Parameters used for interface creation. More...
- union uct_iface_params.mode
- struct uct_iface_params.mode.device
- struct uct_iface_params.mode.sockaddr
- struct uct_completion

    Completion handle. More...
- struct uct_pending_req

    Pending request. More...
- struct uct_iov

    Structure for scatter-gather I/O. More...

Typedefs

- typedef struct uct_md_resource_desc uct_md_resource_desc_t

    Memory domain resource descriptor.
- typedef struct uct_tl_resource_desc uct_tl_resource_desc_t

    Communication resource descriptor.
- typedef struct uct_iface ∗ uct_iface_h
- typedef struct uct_iface_config uct_iface_config_t
- typedef struct uct_md_config uct_md_config_t
- typedef struct uct_ep ∗ uct_ep_h
- typedef void ∗ uct_mem_h
- typedef uintptr_t uct_rkey_t
- typedef struct uct_md ∗ uct_md_h

    Memory domain handler.
- typedef struct uct_md_ops uct_md_ops_t

- typedef void ∗ uct_rkey_ctx_h
- typedef struct uct_iface_attr uct_iface_attr_t
- typedef struct uct_iface_params uct_iface_params_t
- typedef struct uct_md_attr uct_md_attr_t
- typedef struct uct_completion uct_completion_t
- typedef struct uct_pending_req uct_pending_req_t
- typedef struct uct_worker ∗ uct_worker_h
- typedef struct uct_md uct_md_t
- typedef enum uct_am_trace_type uct_am_trace_type_t
- typedef struct uct_device_addr uct_device_addr_t
- typedef struct uct_iface_addr uct_iface_addr_t
- typedef struct uct_ep_addr uct_ep_addr_t
- typedef struct uct_tag_context uct_tag_context_t
- typedef uint64_t uct_tag_t
- typedef int uct_worker_cb_id_t
- typedef void ∗ uct_conn_request_h
- typedef struct uct_iov uct_iov_t

  Structure for scatter-gather I/O.
- typedef void(∗ uct_completion_callback_t) (uct_completion_t ∗self, ucs_status_t status)

  Callback to process send completion.
- typedef ucs_status_t(∗ uct_pending_callback_t) (uct_pending_req_t ∗self)

  Callback to process pending requests.
- typedef ucs_status_t(∗ uct_error_handler_t) (void ∗arg, uct_ep_h ep, ucs_status_t status)

  Callback to process peer failure.
- typedef void(∗ uct_pending_purge_callback_t) (uct_pending_req_t ∗self, void ∗arg)

  Callback to purge pending requests.
- typedef size_t(∗ uct_pack_callback_t) (void ∗dest, void ∗arg)

  Callback for producing data.
- typedef void(∗ uct_unpack_callback_t) (void ∗arg, const void ∗data, size_t length)

  Callback for consuming data.
- typedef void(∗ uct_sockaddr_conn_request_callback_t) (uct_iface_h iface, void ∗arg, uct_conn_request_h conn_request, const void ∗conn_priv_data, size_t length)

  Callback to process an incoming connection request message on the server side.
- typedef ssize_t(∗ uct_sockaddr_priv_pack_callback_t) (void ∗arg, const char ∗dev_name, void ∗priv_data)

  Callback to fill the user's private data on the client side.

Enumerations

- enum uct_device_type_t {
  UCT_DEVICE_TYPE_NET, UCT_DEVICE_TYPE_SHM, UCT_DEVICE_TYPE_ACC, UCT_DEVICE_TY↩
  PE_SELF,
  UCT_DEVICE_TYPE_LAST }

  List of UCX device types.
- enum uct_iface_event_types { UCT_EVENT_SEND_COMP = UCS_BIT(0), UCT_EVENT_RECV = UCS_↩
  BIT(1), UCT_EVENT_RECV_SIG = UCS_BIT(2) }

  Asynchronous event types.
- enum uct_flush_flags { UCT_FLUSH_FLAG_LOCAL = 0, UCT_FLUSH_FLAG_CANCEL = UCS_BIT(0) }

  Flush modifiers.
- enum uct_progress_types { UCT_PROGRESS_SEND = UCS_BIT(0), UCT_PROGRESS_RECV = UCS_↩
  BIT(1), UCT_PROGRESS_THREAD_SAFE = UCS_BIT(7) }

  UCT progress types.
- enum uct_cb_flags { UCT_CB_FLAG_RESERVED = UCS_BIT(1), UCT_CB_FLAG_ASYNC = UCS_BIT(2) }

Callback flags.
- enum uct_iface_open_mode { UCT_IFACE_OPEN_MODE_DEVICE = UCS_BIT(0), UCT_IFACE_OPEN_MODE_SOCKADDR_SERVER = UCS_BIT(1), UCT_IFACE_OPEN_MODE_SOCKADDR_CLIENT = UCS_BIT(2) }

  Mode in which to open the interface.
- enum uct_cb_param_flags { UCT_CB_PARAM_FLAG_DESC = UCS_BIT(0) }

  Flags for active message and tag-matching offload callbacks (callback's parameters).

Functions

- ucs_status_t uct_query_md_resources (uct_md_resource_desc_t **resources_p, unsigned *num_resources_p)

  Query for memory resources.
- void uct_release_md_resource_list (uct_md_resource_desc_t *resources)

  Release the list of resources returned from uct_query_md_resources.
- ucs_status_t uct_md_open (const char *md_name, const uct_md_config_t *config, uct_md_h *md_p)

  Open a memory domain.
- void uct_md_close (uct_md_h md)

  Close a memory domain.
- ucs_status_t uct_md_query_tl_resources (uct_md_h md, uct_tl_resource_desc_t **resources_p, unsigned *num_resources_p)

  Query for transport resources.
- void uct_release_tl_resource_list (uct_tl_resource_desc_t *resources)

  Release the list of resources returned from uct_md_query_tl_resources.
- ucs_status_t uct_md_iface_config_read (uct_md_h md, const char *tl_name, const char *env_prefix, const char *filename, uct_iface_config_t **config_p)

  Read transport-specific interface configuration.
- void uct_config_release (void *config)

  Release configuration memory returned from uct_md_iface_config_read() or from uct_md_config_read().
- ucs_status_t uct_iface_open (uct_md_h md, uct_worker_h worker, const uct_iface_params_t *params, const uct_iface_config_t *config, uct_iface_h *iface_p)

  Open a communication interface.
- void uct_iface_close (uct_iface_h iface)

  Close and destroy an interface.
- ucs_status_t uct_iface_query (uct_iface_h iface, uct_iface_attr_t *iface_attr)

  Get interface attributes.
- ucs_status_t uct_iface_get_device_address (uct_iface_h iface, uct_device_addr_t *addr)

  Get address of the device the interface is using.
- ucs_status_t uct_iface_get_address (uct_iface_h iface, uct_iface_addr_t *addr)

  Get interface address.
- int uct_iface_is_reachable (const uct_iface_h iface, const uct_device_addr_t *dev_addr, const uct_iface_addr_t *iface_addr)

  Check if remote iface address is reachable.
- ucs_status_t uct_ep_check (const uct_ep_h ep, unsigned flags, uct_completion_t *comp)

  check if the destination endpoint is alive in respect to UCT library
- ucs_status_t uct_iface_event_fd_get (uct_iface_h iface, int *fd_p)

  Obtain a notification file descriptor for polling.
- ucs_status_t uct_iface_event_arm (uct_iface_h iface, unsigned events)

  Turn on event notification for the next event.
- ucs_status_t uct_iface_mem_alloc (uct_iface_h iface, size_t length, unsigned flags, const char *name, uct_allocated_memory_t *mem)

Allocate memory which can be used for zero-copy communications.

- void uct_iface_mem_free (const uct_allocated_memory_t ∗mem)

  Release memory allocated with uct_iface_mem_alloc().

- ucs_status_t uct_iface_accept (uct_iface_h iface, uct_conn_request_h conn_request)

  Accept connection request.

- ucs_status_t uct_iface_reject (uct_iface_h iface, uct_conn_request_h conn_request)

  Reject connection request. Will invoke an error handler uct_error_handler_t on the remote transport interface, if set.

- ucs_status_t uct_ep_create (uct_iface_h iface, uct_ep_h ∗ep_p)

  Create new endpoint.

- ucs_status_t uct_ep_create_connected (uct_iface_h iface, const uct_device_addr_t ∗dev_addr, const uct_iface_addr_t ∗iface_addr, uct_ep_h ∗ep_p)

  Create an endpoint which is connected to remote interface.

- void uct_ep_destroy (uct_ep_h ep)

  Destroy an endpoint.

- ucs_status_t uct_ep_get_address (uct_ep_h ep, uct_ep_addr_t ∗addr)

  Get endpoint address.

- ucs_status_t uct_ep_connect_to_ep (uct_ep_h ep, const uct_device_addr_t ∗dev_addr, const uct_ep_addr_t ∗ep_addr)

  Connect endpoint to a remote endpoint.

- ucs_status_t uct_ep_create_sockaddr (uct_iface_h iface, const ucs_sock_addr_t ∗sockaddr, uct_sockaddr_priv_pack_callback_t pack_cb, void ∗arg, uint32_t cb_flags, uct_ep_h ∗ep_p)

  Initiate a client-server connection to a remote peer.

- ucs_status_t uct_iface_flush (uct_iface_h iface, unsigned flags, uct_completion_t ∗comp)

  Flush outstanding communication operations on an interface.

- ucs_status_t uct_iface_fence (uct_iface_h iface, unsigned flags)

  Ensures ordering of outstanding communications on the interface. Operations issued on the interface prior to this call are guaranteed to be completed before any subsequent communication operations to the same interface which follow the call to fence.

- ucs_status_t uct_ep_pending_add (uct_ep_h ep, uct_pending_req_t ∗req, unsigned flags)

  Add a pending request to an endpoint.

- void uct_ep_pending_purge (uct_ep_h ep, uct_pending_purge_callback_t cb, void ∗arg)

  Remove all pending requests from an endpoint.

- ucs_status_t uct_ep_flush (uct_ep_h ep, unsigned flags, uct_completion_t ∗comp)

  Flush outstanding communication operations on an endpoint.

- ucs_status_t uct_ep_fence (uct_ep_h ep, unsigned flags)

  Ensures ordering of outstanding communications on the endpoint. Operations issued on the endpoint prior to this call are guaranteed to be completed before any subsequent communication operations to the same endpoint which follow the call to fence.

- void uct_iface_progress_enable (uct_iface_h iface, unsigned flags)

  Enable synchronous progress for the interface.

- void uct_iface_progress_disable (uct_iface_h iface, unsigned flags)

  Disable synchronous progress for the interface.

- unsigned uct_iface_progress (uct_iface_h iface)

  Perform a progress on an interface.

6.11.1 Detailed Description

This section describes a concept of the Communication Resource and routines associated with the concept.

6.11.2 Data Structure Documentation

6.11.2.1 struct uct_md_resource_desc

This structure describes a memory domain resource.

Examples:

uct_hello_world.c.

Data Fields

char	md_name[UCT_MD_NAME_MAX]	Memory domain name

6.11.2.2 struct uct_tl_resource_desc

Resource descriptor is an object representing the network resource. Resource descriptor could represent a stand-alone communication resource such as an HCA port, network interface, or multiple resources such as multiple network interfaces or communication ports. It could also represent virtual communication resources that are defined over a single physical network interface.

Examples:

uct_hello_world.c.

Data Fields

char	tl_name[UCT_TL_NAME_MAX]	Transport name
char	dev_name[UCT_DEVICE_NAME_MAX]	Hardware device name
uct_device_type_t	dev_type	Device type. To which UCT group it belongs to

6.11.2.3 struct uct_iface_attr

Examples:

uct_hello_world.c.

Data Fields

struct uct_iface_attr	cap	Interface capabilities
size_t	device_addr_len	Size of device address
size_t	iface_addr_len	Size of interface address
size_t	ep_addr_len	Size of endpoint address
size_t	max_conn_priv	Max size of the iface's private data. used for connection establishment with sockaddr

double	overhead	Message overhead, seconds
double	bandwidth	Maximal bandwidth, bytes/second
uct_linear_↩growth_t	latency	Latency model
uint8_t	priority	Priority of device

6.11.2.4 struct uct_iface_attr.cap

Data Fields

cap	put	Attributes for PUT operations
cap	get	Attributes for GET operations
cap	am	Attributes for AM operations
cap	tag	Attributes for TAG operations
cap	atomic32	
cap	atomic64	Attributes for atomic operations
uint64_t	flags	Flags from UCT interface operations and capabilities

6.11.2.5 struct uct_iface_attr.cap.put

Data Fields

size_t	max_short	Maximal size for put_short
size_t	max_bcopy	Maximal size for put_bcopy
size_t	min_zcopy	Minimal size for put_zcopy (total of uct_iov_t::length of the *iov* parameter)
size_t	max_zcopy	Maximal size for put_zcopy (total of uct_iov_t::length of the *iov* parameter)
size_t	opt_zcopy_align	Optimal alignment for zero-copy buffer address
size_t	align_mtu	MTU used for alignment
size_t	max_iov	Maximal *iovcnt* parameter in uct_ep_put_zcopy

6.11.2.6 struct uct_iface_attr.cap.get

Data Fields

size_t	max_short	Maximal size for get_short
size_t	max_bcopy	Maximal size for get_bcopy
size_t	min_zcopy	Minimal size for get_zcopy (total of uct_iov_t::length of the *iov* parameter)
size_t	max_zcopy	Maximal size for get_zcopy (total of uct_iov_t::length of the *iov* parameter)
size_t	opt_zcopy_align	Optimal alignment for zero-copy buffer address
size_t	align_mtu	MTU used for alignment
size_t	max_iov	Maximal *iovcnt* parameter in uct_ep_get_zcopy

6.11.2.7 struct uct_iface_attr.cap.am

Data Fields

size_t	max_short	Total max. size (incl. the header)

size_t	max_bcopy	Total max. size (incl. the header)
size_t	min_zcopy	Minimal size for am_zcopy (incl. the header and total of uct_iov_t::length of the *iov* parameter)
size_t	max_zcopy	Total max. size (incl. the header and total of uct_iov_t::length of the *iov* parameter)
size_t	opt_zcopy_align	Optimal alignment for zero-copy buffer address
size_t	align_mtu	MTU used for alignment
size_t	max_hdr	Max. header size for zcopy
size_t	max_iov	Maximal *iovcnt* parameter in uct_ep_am_zcopy

6.11.2.8 struct uct_iface_attr.cap.tag

Data Fields

tag	recv	
tag	eager	Attributes related to eager protocol
tag	rndv	Attributes related to rendezvous protocol

6.11.2.9 struct uct_iface_attr.cap.tag.recv

Data Fields

size_t	min_recv	Minimal allowed length of posted receive buffer
size_t	max_zcopy	Maximal allowed data length in uct_iface_tag_recv_zcopy
size_t	max_iov	Maximal *iovcnt* parameter in uct_iface_tag_recv_zcopy
size_t	max_outstanding	Maximal number of simultaneous receive operations

6.11.2.10 struct uct_iface_attr.cap.tag.eager

Data Fields

size_t	max_short	Maximal allowed data length in uct_ep_tag_eager_short
size_t	max_bcopy	Maximal allowed data length in uct_ep_tag_eager_bcopy
size_t	max_zcopy	Maximal allowed data length in uct_ep_tag_eager_zcopy
size_t	max_iov	Maximal *iovcnt* parameter in uct_ep_tag_eager_zcopy

6.11.2.11 struct uct_iface_attr.cap.tag.rndv

Data Fields

size_t	max_zcopy	Maximal allowed data length in uct_ep_tag_rndv_zcopy
size_t	max_hdr	Maximal allowed header length in uct_ep_tag_rndv_zcopy and uct_ep_tag_rndv_request
size_t	max_iov	Maximal *iovcnt* parameter in uct_ep_tag_rndv_zcopy

6.11.2.12 struct uct_iface_attr.cap.atomic32

Data Fields

uint64_t	op_flags	Attributes for atomic-post operations
uint64_t	fop_flags	Attributes for atomic-fetch operations

6.11.2.13 struct uct_iface_attr.cap.atomic64

Data Fields

uint64_t	op_flags	Attributes for atomic-post operations
uint64_t	fop_flags	Attributes for atomic-fetch operations

6.11.2.14 struct uct_iface_params

This structure should be allocated by the user and should be passed to uct_iface_open. User has to initialize all fields of this structure.

Examples:

uct_hello_world.c.

Data Fields

ucs_cpu_set_t	cpu_mask	Mask of CPUs to use for resources
uint64_t	open_mode	Interface open mode bitmap. uct_iface_open_mode
union uct_iface_params	mode	Mode-specific parameters
ucs_stats_node_t *	stats_root	Root in the statistics tree. Can be NULL. If non NULL, it will be a root of *uct_iface* object in the statistics tree.
size_t	rx_headroom	How much bytes to reserve before the receive segment.
void *	err_handler_arg	Custom argument of *err_handler*.
uct_error_handler_t	err_handler	The callback to handle transport level error.
uint32_t	err_handler_flags	Callback flags to indicate where the *err_handler* callback can be invoked from. uct_cb_flags
void *	eager_arg	These callbacks are only relevant for HW Tag Matching
uct_tag_unexp_eager_cb_t	eager_cb	Callback for tag matching unexpected eager messages
void *	rndv_arg	
uct_tag_unexp_rndv_cb_t	rndv_cb	Callback for tag matching unexpected rndv messages

6.11.2.15 union uct_iface_params.mode

Mode-specific parameters

Data Fields

mode	device	The fields in this structure (tl_name and dev_name) need to be set only when the UCT_IFACE_OPEN_MODE_DEVICE bit is set in uct_iface_params_t::open_mode This will make uct_iface_open open the interface on the specified device.

mode	sockaddr	These callbacks and address are only relevant for client-server connection establishment with sockaddr and are needed on the server side. The callbacks and address need to be set when the UCT_IFACE_OPEN_MODE_SOCKADDR_SERVER bit is set in uct_iface_params_t::open_mode. This will make uct_iface_open open the interface on the specified address as a server.

6.11.2.16 struct uct_iface_params.mode.device

The fields in this structure (tl_name and dev_name) need to be set only when the UCT_IFACE_OPEN_MODE_DEVICE bit is set in uct_iface_params_t::open_mode This will make uct_iface_open open the interface on the specified device.

Data Fields

const char *	tl_name	Transport name
const char *	dev_name	Device Name

6.11.2.17 struct uct_iface_params.mode.sockaddr

These callbacks and address are only relevant for client-server connection establishment with sockaddr and are needed on the server side. The callbacks and address need to be set when the UCT_IFACE_OPEN_MODE_SOCKADDR_SERVER bit is set in uct_iface_params_t::open_mode. This will make uct_iface_open open the interface on the specified address as a server.

Data Fields

ucs_sock_addr_t	listen_sockaddr	
void *	conn_request_arg	Argument for connection request callback
uct_sockaddr_conn_request_callback_t	conn_request_cb	Callback for an incoming connection request on the server
uint32_t	cb_flags	Callback flags to indicate where the callback can be invoked from. uct_cb_flags

6.11.2.18 struct uct_completion

This structure should be allocated by the user and can be passed to communication primitives. User has to initializes both fields of the structure. If the operation returns UCS_INPROGRESS, this structure will be in use by the transport until the operation completes. When the operation completes, "count" field is decremented by 1, and whenever it reaches 0 - the callback is called.

Notes:

- The same structure can be passed multiple times to communication functions without the need to wait for completion.
- If the number of operations is smaller than the initial value of the counter, the callback will not be called at all, so it may be left undefined.

Examples:

uct_hello_world.c.

Data Fields

uct_↩completion_↩callback_t	func	User callback function
int	count	Completion counter

6.11.2.19 struct uct_pending_req

This structure should be passed to uct_pending_add() and is used to signal new available resources back to user.

Data Fields

uct_pending_↩callback_t	func	User callback function
char	priv[UCT_PEN↩DING_REQ_P↩RIV_LEN]	Used internally by UCT

6.11.2.20 struct uct_iov

Specifies a list of buffers which can be used within a single data transfer function call.

```
buffer
|
+-----------+-------+-----------+-------+-----------+
|  payload  | empty |  payload  | empty |  payload  |
+-----------+-------+-----------+-------+-----------+
|<-length-->|       |<-length-->|       |<-length-->|
|<---- stride ----->|<---- stride ----->|
```

Note

The sum of lengths in all iov list must be less or equal to max_zcopy of the respective communication operation.
If *length* or *count* are zero, the memory pointed to by *buffer* will not be accessed. Otherwise, *buffer* must point to valid memory.
If *count* is one, every iov entry specifies a single contiguous data block
If *count* $>$ 1, each iov entry specifies a strided block of *count* elements and distance of *stride* byte between consecutive elements

Examples:

uct_hello_world.c.

Data Fields

void ∗	buffer	Data buffer
size_t	length	Length of the payload in bytes
uct_mem_h	memh	Local memory key descriptor for the data
size_t	stride	Stride between beginnings of payload elements in the buffer in bytes
unsigned	count	Number of payload elements in the buffer

6.11.3 Typedef Documentation

6.11.3.1 typedef struct uct_md_resource_desc uct_md_resource_desc_t

This structure describes a memory domain resource.

6.11.3.2 typedef struct uct_tl_resource_desc uct_tl_resource_desc_t

Resource descriptor is an object representing the network resource. Resource descriptor could represent a stand-alone communication resource such as an HCA port, network interface, or multiple resources such as multiple network interfaces or communication ports. It could also represent virtual communication resources that are defined over a single physical network interface.

6.11.3.3 typedef struct uct_iface* uct_iface_h

6.11.3.4 typedef struct uct_iface_config uct_iface_config_t

6.11.3.5 typedef struct uct_md_config uct_md_config_t

6.11.3.6 typedef struct uct_ep* uct_ep_h

6.11.3.7 typedef void* uct_mem_h

6.11.3.8 typedef uintptr_t uct_rkey_t

6.11.3.9 typedef struct uct_md* uct_md_h

6.11.3.10 typedef struct uct_md_ops uct_md_ops_t

6.11.3.11 typedef void* uct_rkey_ctx_h

6.11.3.12 typedef struct uct_iface_attr uct_iface_attr_t

6.11.3.13 typedef struct uct_iface_params uct_iface_params_t

6.11.3.14 typedef struct uct_md_attr uct_md_attr_t

6.11.3.15 typedef struct uct_completion uct_completion_t

6.11.3.16 typedef struct uct_pending_req uct_pending_req_t

6.11.3.17 typedef struct uct_worker* uct_worker_h

6.11.3.18 typedef struct uct_md uct_md_t

6.11.3.19 typedef enum uct_am_trace_type uct_am_trace_type_t

6.11.3.20 typedef struct uct_device_addr uct_device_addr_t

6.11.3.21 typedef struct uct_iface_addr uct_iface_addr_t

6.11.3.22 typedef struct uct_ep_addr uct_ep_addr_t

6.11.3.23 typedef struct uct_tag_context uct_tag_context_t

6.11.3.24 typedef uint64_t uct_tag_t

6.11.3.25 typedef int uct_worker_cb_id_t

6.11.3.26 typedef void* uct_conn_request_h

6.11.3.27 typedef struct uct_iov uct_iov_t

Specifies a list of buffers which can be used within a single data transfer function call.

```
buffer
|
+-----------+-------+-----------+-------+-----------+
|  payload  | empty |  payload  | empty |  payload  |
+-----------+-------+-----------+-------+-----------+
|<-length-->|       |<-length-->|       |<-length-->|
|<---- stride ----->|<---- stride ----->|
```

Note

The sum of lengths in all iov list must be less or equal to max_zcopy of the respective communication operation.
If *length* or *count* are zero, the memory pointed to by *buffer* will not be accessed. Otherwise, *buffer* must point to valid memory.
If *count* is one, every iov entry specifies a single contiguous data block
If *count* $>$ 1, each iov entry specifies a strided block of *count* elements and distance of *stride* byte between consecutive elements

6.11.3.28 typedef void(∗ uct_completion_callback_t) (uct_completion_t ∗self, ucs_status_t status)

Parameters

in	*self*	Pointer to relevant completion structure, which was initially passed to the operation.
in	*status*	Status of send action, possibly indicating an error.

6.11.3.29 typedef ucs_status_t(∗ uct_pending_callback_t) (uct_pending_req_t ∗self)

Parameters

in	*self*	Pointer to relevant pending structure, which was initially passed to the operation.

Returns

UCS_OK - This pending request has completed and should be removed. UCS_INPROGRESS - Some progress was made, but not completed. Keep this request and keep processing the queue. Otherwise - Could not make any progress. Keep this pending request on the queue, and stop processing the queue.

6.11.3.30 typedef ucs_status_t(∗ uct_error_handler_t) (void ∗arg, uct_ep_h ep, ucs_status_t status)

Parameters

in	*arg*	User argument to be passed to the callback.
in	*ep*	Endpoint which has failed. Upon return from the callback, this *ep* is no longer usable and all subsequent operations on this *ep* will fail with the error code passed in *status*.

in	*status*	Status indicating error.

Returns

UCS_OK - The error was handled successfully. Otherwise - The error was not handled and is returned back to the transport.

6.11.3.31 typedef void(∗ uct_pending_purge_callback_t) (uct_pending_req_t ∗self, void ∗arg)

Parameters

in	*self*	Pointer to relevant pending structure, which was initially passed to the operation.
in	*arg*	User argument to be passed to the callback.

6.11.3.32 typedef size_t(∗ uct_pack_callback_t) (void ∗dest, void ∗arg)

Parameters

in	*dest*	Memory buffer to pack the data to.
in	*arg*	Custom user-argument.

Returns

Size of the data was actually produced.

6.11.3.33 typedef void(∗ uct_unpack_callback_t) (void ∗arg, const void ∗data, size_t length)

Parameters

in	*arg*	Custom user-argument.
in	*data*	Memory buffer to unpack the data from.
in	*length*	How much data to consume (size of "data")

Note

The arguments for this callback are in the same order as libc's memcpy().

6.11.3.34 typedef void(∗ uct_sockaddr_conn_request_callback_t) (uct_iface_h iface, void ∗arg, uct_conn_request_h conn_request, const void ∗conn_priv_data, size_t length)

This callback routine will be invoked on the server side upon receiving an incoming connection request. It should be set by the server side while initializing an interface. Incoming data is placed inside the conn_priv_data buffer. This callback has to be thread safe. Other than communication progress routines, it is allowed to call other UCT communication routines from this callback.

Parameters

in	*iface*	Transport interface.

in	*arg*	User defined argument for this callback.
in	*conn_request*	Transport level connection request. The user should accept or reject the request by calling uct_iface_accept or uct_iface_reject routines respectively.
in	*conn_priv_data*	Points to the received data. This is the private data that was passed to the uct_ep_create_sockaddr function on the client side.
in	*length*	Length of the received data.

6.11.3.35 typedef ssize_t(∗ uct_sockaddr_priv_pack_callback_t) (void ∗arg, const char ∗dev_name, void ∗priv_data)

This callback routine will be invoked on the client side before sending the transport's connection request to the server. The callback routine must be set by the client when creating an endpoint. The user's private data should be placed inside the priv_data buffer to be sent to the server side. The maximal allowed length of the private data is indicated by the field max_conn_priv inside uct_iface_attr. Communication progress routines should not be called from this callback. It is allowed to call other UCT communication routines from this callback.

Parameters

in	*arg*	User defined argument for this callback.
in	*dev_name*	Device name. This routine may fill the user's private data according to the given device name. The device name that is passed to this routine, corresponds to the dev_name field inside uct_tl_resource_desc_t as returned from uct_md↩_query_tl_resources.
out	*priv_data*	User's private data to be passed to the server side.

Returns

Negative value indicates an error according to ucs_status_t. On success, non-negative value indicates actual number of bytes written to the *priv_data* buffer.

6.11.4 Enumeration Type Documentation

6.11.4.1 enum uct_device_type_t

Enumerator

UCT_DEVICE_TYPE_NET Network devices

UCT_DEVICE_TYPE_SHM Shared memory devices

UCT_DEVICE_TYPE_ACC Acceleration devices

UCT_DEVICE_TYPE_SELF Loop-back device

UCT_DEVICE_TYPE_LAST

6.11.4.2 enum uct_iface_event_types

Note

The UCT_EVENT_RECV and UCT_EVENT_RECV_SIG event types are used to indicate receive-side completions for both tag matching and active messages. If the interface supports signaled receives (UCT_↩IFACE_FLAG_EVENT_RECV_SIG), then for the messages sent with UCT_SEND_FLAG_SIGNALED flag, UCT_EVENT_RECV_SIG should be triggered on the receiver. Otherwise, UCT_EVENT_RECV should be triggered.

Enumerator

UCT_EVENT_SEND_COMP Send completion event

UCT_EVENT_RECV Tag or active message received

UCT_EVENT_RECV_SIG Signaled tag or active message received

6.11.4.3 enum uct_flush_flags

Enumerator

UCT_FLUSH_FLAG_LOCAL Guarantees that the data transfer is completed but the target buffer may not be updated yet.

UCT_FLUSH_FLAG_CANCEL The library will make a best effort attempt to cancel all uncompleted operations. However, there is a chance that some operations will not be canceled in which case the user will need to handle their completions through the relevant callbacks.

6.11.4.4 enum uct_progress_types

Enumerator

UCT_PROGRESS_SEND Progress send operations

UCT_PROGRESS_RECV Progress receive operations

UCT_PROGRESS_THREAD_SAFE Enable/disable progress while another thread may be calling ucp_↩ worker_progress().

6.11.4.5 enum uct_cb_flags

List of flags for a callback.

Enumerator

UCT_CB_FLAG_RESERVED Reserved for future use.

UCT_CB_FLAG_ASYNC Callback may be invoked from any context (thread, process). For example, it may be called from a transport async progress thread. To guarantee async invocation, the interface must have the UCT_IFACE_FLAG_CB_ASYNC flag set. If async callback is requested on an interface which only supports sync callback (i.e., only the UCT_IFACE_FLAG_CB_SYNC flag is set), the callback may be invoked only from the context that called uct_iface_progress).

6.11.4.6 enum uct_iface_open_mode

Enumerator

UCT_IFACE_OPEN_MODE_DEVICE Interface is opened on a specific device

UCT_IFACE_OPEN_MODE_SOCKADDR_SERVER Interface is opened on a specific address on the server side

UCT_IFACE_OPEN_MODE_SOCKADDR_CLIENT Interface is opened on a specific address on the client side

6.11.4.7 enum uct_cb_param_flags

Enumerator

UCT_CB_PARAM_FLAG_DESC If this flag is enabled, then data is part of a descriptor which includes the user-defined rx_headroom, and the callback may return UCS_INPROGRESS and hold on to that descriptor. Otherwise, the data can't be used outside the callback. If needed, the data must be copied-out.

```
descriptor    data
|             |
+-------------+-------------------------+
| rx_headroom | payload                 |
+-------------+-------------------------+
```

6.11.5 Function Documentation

6.11.5.1 ucs_status_t uct_query_md_resources (uct_md_resource_desc_t ** *resources_p*, unsigned * *num_resources_p*)

Obtain the list of memory domain resources available on the current system.

Parameters

out	*resources_p*	Filled with a pointer to an array of resource descriptors.
out	*num_↩ resources_p*	Filled with the number of resources in the array.

Returns

Error code.

Examples:

uct_hello_world.c.

6.11.5.2 void uct_release_md_resource_list (uct_md_resource_desc_t ∗ *resources*)

This routine releases the memory associated with the list of resources allocated by uct_query_md_resources.

Parameters

in	*resources*	Array of resource descriptors to release.

Examples:

uct_hello_world.c.

6.11.5.3 ucs_status_t uct_md_open (const char ∗ *md_name,* const uct_md_config_t ∗ *config,* uct_md_h ∗ *md_p*)

Open a specific memory domain. All communications and memory operations are performed in the context of a specific memory domain. Therefore it must be created before communication resources.

Parameters

in	*md_name*	Memory domain name, as returned from uct_query_md_resources.
in	*config*	MD configuration options. Should be obtained from uct_md_config_read() function, or point to MD-specific structure which extends uct_md_config_t.
out	*md_p*	Filled with a handle to the memory domain.

Returns

Error code.

Examples:

uct_hello_world.c.

6.11.5.4 void uct_md_close (uct_md_h *md*)

Parameters

in	*md*	Memory domain to close.

Examples:

uct_hello_world.c.

6.11.5.5 ucs_status_t uct_md_query_tl_resources (uct_md_h *md,* uct_tl_resource_desc_t ∗∗ *resources_p,* unsigned ∗ *num_resources_p*)

This routine queries the memory domain for communication resources that are available for it.

Parameters

in	*md*	Handle to memory domain.
out	*resources_p*	Filled with a pointer to an array of resource descriptors.
out	*num_↩ resources_p*	Filled with the number of resources in the array.

Returns

Error code.

Examples:

uct_hello_world.c.

6.11.5.6 void uct_release_tl_resource_list (uct_tl_resource_desc_t * *resources*)

This routine releases the memory associated with the list of resources allocated by uct_md_query_tl_resources.

Parameters

in	*resources*	Array of resource descriptors to release.

Examples:

uct_hello_world.c.

6.11.5.7 ucs_status_t uct_md_iface_config_read (uct_md_h *md,* const char * *tl_name,* const char * *env_prefix,* const char * *filename,* uct_iface_config_t ** *config_p*)

Parameters

in	*md*	Memory domain on which the transport's interface was registered.
in	*tl_name*	Transport name. If *md* supports UCT_MD_FLAG_SOCKADDR, the transport name is allowed to be NULL. In this case, the configuration returned from this routine should be passed to uct_iface_open with UCT_IFACE_OPEN_↩ MODE_SOCKADDR_SERVER or UCT_IFACE_OPEN_MODE_SOCKADD↩ R_CLIENT set in uct_iface_params_t::open_mode. In addition, if tl_name is not NULL, the configuration returned from this routine should be passed to uct_iface_open with UCT_IFACE_OPEN_MODE_DEVICE set in uct_iface_↩ params_t::open_mode.
in	*env_prefix*	If non-NULL, search for environment variables starting with this UCT_↩ <prefix>_. Otherwise, search for environment variables starting with just U↩ CT_.
in	*filename*	If non-NULL, read configuration from this file. If the file does not exist, it will be ignored.
out	*config_p*	Filled with a pointer to configuration.

Returns

Error code.

Examples:

uct_hello_world.c.

6.11.5.8 void uct_config_release (void * *config*)

Parameters

in	*config*	Configuration to release.

Examples:

uct_hello_world.c.

6.11.5.9 ucs_status_t uct_iface_open (uct_md_h *md*, uct_worker_h *worker*, const uct_iface_params_t * *params*, const uct_iface_config_t * *config*, uct_iface_h * *iface_p*)

Parameters

in	*md*	Memory domain to create the interface on.
in	*worker*	Handle to worker which will be used to progress communications on this interface.
in	*params*	User defined uct_iface_params_t parameters.
in	*config*	Interface configuration options. Should be obtained from uct_md_iface_config_read() function, or point to transport-specific structure which extends uct_iface_config_t.
out	*iface_p*	Filled with a handle to opened communication interface.

Returns

Error code.

Examples:

uct_hello_world.c.

6.11.5.10 void uct_iface_close (uct_iface_h *iface*)

Parameters

in	*iface*	Interface to close.

Examples:

uct_hello_world.c.

6.11.5.11 ucs_status_t uct_iface_query (uct_iface_h *iface*, uct_iface_attr_t * *iface_attr*)

Parameters

in	*iface*	Interface to query.
out	*iface_attr*	Filled with interface attributes.

Examples:

uct_hello_world.c.

6.11.5.12 ucs_status_t uct_iface_get_device_address (uct_iface_h *iface*, uct_device_addr_t * *addr*)

Get underlying device address of the interface. All interfaces using the same device would return the same address.

Parameters

in	*iface*	Interface to query.
out	*addr*	Filled with device address. The size of the buffer provided must be at least uct_iface_attr_t::device_addr_len.

Examples:

uct_hello_world.c.

6.11.5.13 ucs_status_t uct_iface_get_address (uct_iface_h *iface*, uct_iface_addr_t * *addr*)

requires UCT_IFACE_FLAG_CONNECT_TO_IFACE.

Parameters

in	*iface*	Interface to query.
out	*addr*	Filled with interface address. The size of the buffer provided must be at least uct_iface_attr_t::iface_addr_len.

Examples:

uct_hello_world.c.

6.11.5.14 int uct_iface_is_reachable (const uct_iface_h *iface*, const uct_device_addr_t * *dev_addr*, const uct_iface_addr_t * *iface_addr*)

This function checks if a remote address can be reached from a local interface. If the function returns true, it does not necessarily mean a connection and/or data transfer would succeed, since the reachability check is a local operation it does not detect issues such as network mis-configuration or lack of connectivity.

Parameters

in	*iface*	Interface to check reachability from.
in	*dev_addr*	Device address to check reachability to. It is NULL if iface_attr.dev_addr_len == 0, and must be non-NULL otherwise.
in	*iface_addr*	Interface address to check reachability to. It is NULL if iface_attr.iface_addr↩_len == 0, and must be non-NULL otherwise.

Returns

Nonzero if reachable, 0 if not.

Examples:

uct_hello_world.c.

6.11.5.15 ucs_status_t uct_ep_check (const uct_ep_h *ep*, unsigned *flags*, uct_completion_t * *comp*)

This function checks if the destination endpoint is alive with respect to the UCT library. If the status of *ep* is known, either UCS_OK or an error is returned immediately. Otherwise, UCS_INPROGRESS is returned, indicating that synchronization on the status is needed. In this case, the status will be be propagated by *comp* callback.

Parameters

in	*ep*	Endpoint to check
in	*flags*	Flags that define level of check (currently unsupported - set to 0).
in	*comp*	Handler to process status of *ep*

Returns

Error code.

6.11.5.16 ucs_status_t uct_iface_event_fd_get (uct_iface_h *iface,* int ∗ *fd_p*)

Only interfaces that support at least one of the UCT_IFACE_FLAG_EVENT∗ flags will implement this function.

Parameters

in	*iface*	Interface to get the notification descriptor.
out	*fd_p*	Location to write the notification file descriptor.

Returns

Error code.

6.11.5.17 ucs_status_t uct_iface_event_arm (uct_iface_h *iface,* unsigned *events*)

This routine needs to be called before waiting on each notification on this interface, so will typically be called once the processing of the previous event is over.

Parameters

in	*iface*	Interface to arm.
in	*events*	Events to wakeup on. See uct_iface_event_types

Returns

UCS_OK The operation completed successfully. File descriptor will be signaled by new events.
UCS_ERR_BUSY There are unprocessed events which prevent the file descriptor from being armed. The operation is not completed. File descriptor will not be signaled by new events.
Other different error codes in case of issues.

6.11.5.18 ucs_status_t uct_iface_mem_alloc (uct_iface_h *iface,* size_t *length,* unsigned *flags,* const char ∗ *name,* uct_allocated_memory_t ∗ *mem*)

Allocate a region of memory which can be used for zero-copy data transfer or remote access on a particular transport interface.

Parameters

in	*iface*	Interface to allocate memory on.
in	*length*	Size of memory region to allocate.
in	*flags*	Memory allocation flags, see uct_md_mem_flags.
in	*name*	Allocation name, for debug purposes.

out	*mem*	Descriptor of allocated memory.

Returns

UCS_OK if allocation was successful, error code otherwise.

6.11.5.19 void uct_iface_mem_free (const uct_allocated_memory_t * *mem*)

Parameters

in	*mem*	Descriptor of memory to release.

6.11.5.20 ucs_status_t uct_iface_accept (uct_iface_h *iface*, uct_conn_request_h *conn_request*)

Parameters

in	*iface*	Transport interface which generated connection request *conn_request*.
in	*conn_request*	Connection establishment request passed as parameter of uct_sockaddr_conn_request_callback_t.

Returns

Error code as defined by ucs_status_t

6.11.5.21 ucs_status_t uct_iface_reject (uct_iface_h *iface*, uct_conn_request_h *conn_request*)

Parameters

in	*iface*	Interface which generated connection establishment request *conn_request*.
in	*conn_request*	Connection establishment request passed as parameter of uct_sockaddr_conn_request_callback_t.

Returns

Error code as defined by ucs_status_t

6.11.5.22 ucs_status_t uct_ep_create (uct_iface_h *iface*, uct_ep_h * *ep_p*)

Parameters

in	*iface*	Interface to create the endpoint on.
out	*ep_p*	Filled with handle to the new endpoint.

Examples:

uct_hello_world.c.

6.11.5.23 ucs_status_t uct_ep_create_connected (uct_iface_h *iface*, const uct_device_addr_t * *dev_addr*, const uct_iface_addr_t * *iface_addr*, uct_ep_h * *ep_p*)

requires UCT_IFACE_FLAG_CONNECT_TO_IFACE capability.

Parameters

in	*iface*	Interface to create the endpoint on.
in	*dev_addr*	Remote device address to connect to.
in	*iface_addr*	Remote interface address to connect to.
out	*ep_p*	Filled with handle to the new endpoint.

Examples:

uct_hello_world.c.

6.11.5.24 void uct_ep_destroy (uct_ep_h *ep*)

Parameters

in	*ep*	Endpoint to destroy.

Examples:

uct_hello_world.c.

6.11.5.25 ucs_status_t uct_ep_get_address (uct_ep_h *ep,* uct_ep_addr_t ∗ *addr*)

Parameters

in	*ep*	Endpoint to query.
out	*addr*	Filled with endpoint address. The size of the buffer provided must be at least uct_iface_attr_t::ep_addr_len.

Examples:

uct_hello_world.c.

6.11.5.26 ucs_status_t uct_ep_connect_to_ep (uct_ep_h *ep,* const uct_device_addr_t ∗ *dev_addr,* const uct_ep_addr_t ∗ *ep_addr*)

requires UCT_IFACE_FLAG_CONNECT_TO_EP capability.

Parameters

in	*ep*	Endpoint to connect.
in	*dev_addr*	Remote device address.
in	*ep_addr*	Remote endpoint address.

Examples:

uct_hello_world.c.

6.11.5.27 ucs_status_t uct_ep_create_sockaddr (uct_iface_h *iface,* const ucs_sock_addr_t ∗ *sockaddr,* uct_sockaddr_priv_pack_callback_t *pack_cb,* void ∗ *arg,* uint32_t *cb_flags,* uct_ep_h ∗ *ep_p*)

This routine will create an endpoint for a connection to the remote peer, specified by its socket address. The user may provide a callback function which will be used to fill the private data that will be sent on a connection request to the remote peer.

Note

It is never guaranteed that the callaback will be called. If, for example, the endpoint goes into error state before issuing the connection request, the callback will not be invoked.
The interface in this routine requires the UCT_IFACE_FLAG_CONNECT_TO_SOCKADDR capability.

Parameters

in	*iface*	Interface to create the endpoint on.
in	*sockaddr*	The sockaddr to connect to on the remote peer.
in	*pack_cb*	Callback for filling the user's private data.
in	*arg*	User defined argument for the callback.
in	*cb_flags*	Required callback flags to indicate where the uct_sockaddr_priv_pack_↩ callback_t callback can be invoked from.
out	*ep_p*	Handle to the created endpoint.

Returns

UCS_OK Connection request was sent to the server. This does not guarantee that the server has received the message; in case of failure, the error will be reported to the interface error handler callback provided to uct_iface_open via uct_iface_params_t::err_handler.
error code In case of an error. (ucs_status_t)

6.11.5.28 ucs_status_t uct_iface_flush (uct_iface_h *iface*, unsigned *flags*, uct_completion_t ∗ *comp*)

Flushes all outstanding communications issued on the interface prior to this call. The operations are completed at the origin or at the target as well. The exact completion semantic depends on *flags* parameter.

Note

Currently only one completion type is supported. It guarantees that the data transfer is completed but the target buffer may not be updated yet.

Parameters

in	*iface*	Interface to flush communications from.
in	*flags*	Flags that control completion semantic (currently only UCT_FLUSH_FLAG↩ _LOCAL is supported).
in,out	*comp*	Completion handle as defined by uct_completion_t. Can be NULL, which means that the call will return the current state of the interface and no completion will be generated in case of outstanding communications. If it is not NULL completion counter is decremented by 1 when the call completes. Completion callback is called when the counter reaches 0.

Returns

UCS_OK - No outstanding communications left. UCS_INPROGRESS - Some communication operations are still in progress. If non-NULL 'comp' is provided, it will be updated upon completion of these operations.

6.11.5.29 ucs_status_t uct_iface_fence (uct_iface_h *iface*, unsigned *flags*)

Parameters

in	*iface*	Interface to issue communications from.
in	*flags*	Flags that control ordering semantic (currently unsupported - set to 0).

Returns

UCS_OK - Ordering is inserted.

6.11.5.30 ucs_status_t uct_ep_pending_add (uct_ep_h *ep*, uct_pending_req_t ∗ *req*, unsigned *flags*)

Add a pending request to the endpoint pending queue. The request will be dispatched when the endpoint could potentially have additional send resources.

Parameters

in	*ep*	Endpoint to add the pending request to.
in	*req*	Pending request, which would be dispatched when more resources become available. The user is expected to initialize the "func" field. After passed to the function, the request is owned by UCT, until the callback is called and returns UCS_OK.
in	*flags*	Reserved for future use.

Returns

UCS_OK - request added to pending queue UCS_ERR_BUSY - request was not added to pending queue, because send resources are available now. The user is advised to retry.

6.11.5.31 void uct_ep_pending_purge (uct_ep_h *ep*, uct_pending_purge_callback_t *cb*, void ∗ *arg*)

Remove pending requests from the given endpoint and pass them to the provided callback function. The callback return value is ignored.

Parameters

in	*ep*	Endpoint to remove pending requests from.
in	*cb*	Callback to pass the removed requests to.
in	*arg*	Argument to pass to the *cb* callback.

6.11.5.32 ucs_status_t uct_ep_flush (uct_ep_h *ep*, unsigned *flags*, uct_completion_t ∗ *comp*)

Flushes all outstanding communications issued on the endpoint prior to this call. The operations are completed at the origin or at the target as well. The exact completion semantic depends on *flags* parameter.

Parameters

in	*ep*	Endpoint to flush communications from.
in	*flags*	Flags uct_flush_flags that control completion semantic.
in,out	*comp*	Completion handle as defined by uct_completion_t. Can be NULL, which means that the call will return the current state of the endpoint and no completion will be generated in case of outstanding communications. If it is not NULL completion counter is decremented by 1 when the call completes. Completion callback is called when the counter reaches 0.

Returns

UCS_OK - No outstanding communications left. UCS_ERR_NO_RESOURCE - Flush operation could not be initiated. A subsequent call to uct_ep_pending_add would add a pending operation, which provides an opportunity to retry the flush. UCS_INPROGRESS - Some communication operations are still in progress. If non-NULL 'comp' is provided, it will be updated upon completion of these operations.

6.11.5.33 ucs_status_t uct_ep_fence (uct_ep_h *ep*, unsigned *flags*)

Parameters

in	*ep*	Endpoint to issue communications from.
in	*flags*	Flags that control ordering semantic (currently unsupported - set to 0).

Returns

UCS_OK - Ordering is inserted.

6.11.5.34 void uct_iface_progress_enable (uct_iface_h *iface*, unsigned *flags*)

Notify the transport that it should actively progress communications during uct_worker_progress().

When the interface is created, its progress is initially disabled.

Parameters

in	*iface*	The interface to enable progress.
in	*flags*	The type of progress to enable as defined by uct_progress_types

Note

This function is not thread safe with respect to ucp_worker_progress(), unless the flag UCT_PROGRESS_THREAD_SAFE is specified.

Examples:

uct_hello_world.c.

6.11.5.35 void uct_iface_progress_disable (uct_iface_h *iface*, unsigned *flags*)

Notify the transport that it should not progress its communications during uct_worker_progress(). Thus the latency of other transports may be improved.

By default, progress is disabled when the interface is created.

Parameters

in	*iface*	The interface to disable progress.
in	*flags*	The type of progress to disable as defined by uct_progress_types.

Note

This function is not thread safe with respect to ucp_worker_progress(), unless the flag UCT_PROGRESS_THREAD_SAFE is specified.

6.11.5.36 unsigned uct_iface_progress (uct_iface_h *iface*)

6.12 UCT Communication Context

Enumerations

- enum uct_alloc_method_t {
  UCT_ALLOC_METHOD_THP, UCT_ALLOC_METHOD_MD, UCT_ALLOC_METHOD_HEAP, UCT_ALL↩
  OC_METHOD_MMAP,
  UCT_ALLOC_METHOD_HUGE, UCT_ALLOC_METHOD_LAST, UCT_ALLOC_METHOD_DEFAULT = U↩
  CT_ALLOC_METHOD_LAST }

  Memory allocation methods.

Functions

- ucs_status_t uct_worker_create (ucs_async_context_t ∗async, ucs_thread_mode_t thread_mode, uct_↩
  worker_h ∗worker_p)

  Create a worker object.
- void uct_worker_destroy (uct_worker_h worker)

  Destroy a worker object.
- void uct_worker_progress_register_safe (uct_worker_h worker, ucs_callback_t func, void ∗arg, unsigned flags, uct_worker_cb_id_t ∗id_p)

  Add a slow path callback function to a worker progress.
- void uct_worker_progress_unregister_safe (uct_worker_h worker, uct_worker_cb_id_t ∗id_p)

  Remove a slow path callback function from worker's progress.
- ucs_status_t uct_config_get (void ∗config, const char ∗name, char ∗value, size_t max)

  Get value by name from interface/MD configuration.
- ucs_status_t uct_config_modify (void ∗config, const char ∗name, const char ∗value)

  Modify interface/MD configuration.
- unsigned uct_worker_progress (uct_worker_h worker)

  Explicit progress for UCT worker.

6.12.1 Detailed Description

UCT context abstracts all the resources required for network communication. It is designed to enable either share or isolate resources for multiple programming models used by an application.

This section provides a detailed description of this concept and routines associated with it.

6.12.2 Enumeration Type Documentation

6.12.2.1 enum **uct_alloc_method_t**

Enumerator

UCT_ALLOC_METHOD_THP Allocate from OS using libc allocator with Transparent Huge Pages enabled

UCT_ALLOC_METHOD_MD Allocate using memory domain

UCT_ALLOC_METHOD_HEAP Allocate from heap using libc allocator

UCT_ALLOC_METHOD_MMAP Allocate from OS using mmap() syscall

UCT_ALLOC_METHOD_HUGE Allocate huge pages

UCT_ALLOC_METHOD_LAST

UCT_ALLOC_METHOD_DEFAULT Use default method

6.12.3 Function Documentation

6.12.3.1 ucs_status_t uct_worker_create (ucs_async_context_t * *async*, ucs_thread_mode_t *thread_mode*, uct_worker_h * *worker_p*)

The worker represents a progress engine. Multiple progress engines can be created in an application, for example to be used by multiple threads. Transports can allocate separate communication resources for every worker, so that every worker can be progressed independently of others.

Parameters

in	*async*	Context for async event handlers. Must not be NULL.
in	*thread_mode*	Thread access mode to the worker and all interfaces and endpoints associated with it.
out	*worker_p*	Filled with a pointer to the worker object.

Examples:

uct_hello_world.c.

6.12.3.2 void uct_worker_destroy (uct_worker_h *worker*)

Parameters

in	*worker*	Worker object to destroy.

Examples:

uct_hello_world.c.

6.12.3.3 void uct_worker_progress_register_safe (uct_worker_h *worker*, ucs_callback_t *func*, void * *arg*, unsigned *flags*, uct_worker_cb_id_t * *id_p*)

If *id_p is equal to UCS_CALLBACKQ_ID_NULL, this function will add a callback which will be invoked every time progress is made on the worker. *id_p will be updated with an id which refers to this callback and can be used in uct_worker_progress_unregister_safe to remove it from the progress path.

Parameters

in	*worker*	Handle to the worker whose progress should invoke the callback.
in	*func*	Pointer to the callback function.
in	*arg*	Argument for the callback function.
in	*flags*	Callback flags, see ucs_callbackq_flags.
in,out	*id_p*	Points to a location to store a callback identifier. If *id_p is equal to UCS_CA↩ LLBACKQ_ID_NULL, a callback will be added and *id_p will be replaced with a callback identifier which can be subsequently used to remove the callback. Otherwise, no callback will be added and *id_p will be left unchanged.

Note

This function is thread safe.

6.12.3.4 void uct_worker_progress_unregister_safe (uct_worker_h *worker*, uct_worker_cb_id_t * *id_p*)

If *id_p is not equal to UCS_CALLBACKQ_ID_NULL, remove a callback which was previously added by uct_↩ worker_progress_register_safe. *id_p will be reset to UCS_CALLBACKQ_ID_NULL.

Parameters

`in`	*worker*	Handle to the worker whose progress should invoke the callback.
`in,out`	*id_p*	Points to a callback identifier which indicates the callback to remove. If ∗id← _p is not equal to UCS_CALLBACKQ_ID_NULL, the callback will be removed and ∗id_p will be reset to UCS_CALLBACKQ_ID_NULL. If ∗id_p is equal to UCS_CALLBACKQ_ID_NULL, no operation will be performed and ∗id_p will be left unchanged.

Note

This function is thread safe.

6.12.3.5 ucs_status_t uct_config_get (void ∗ *config,* const char ∗ *name,* char ∗ *value,* size_t *max*)

Parameters

`in`	*config*	Configuration to get from.
`in`	*name*	Configuration variable name.
`out`	*value*	Pointer to get value. Should be allocated/freed by caller.
`in`	*max*	Available memory space at *value* pointer.

Returns

UCS_OK if found, otherwise UCS_ERR_INVALID_PARAM or UCS_ERR_NO_ELEM if error.

6.12.3.6 ucs_status_t uct_config_modify (void ∗ *config,* const char ∗ *name,* const char ∗ *value*)

Parameters

`in`	*config*	Configuration to modify.
`in`	*name*	Configuration variable name.
`in`	*value*	Value to set.

Returns

Error code.

6.12.3.7 unsigned uct_worker_progress (uct_worker_h *worker*)

This routine explicitly progresses any outstanding communication operations and active message requests.

Note

- In the current implementation, users **MUST** call this routine to receive the active message requests.

Parameters

`in`	*worker*	Handle to worker.

Returns

Non-zero if any communication was progressed, zero otherwise.

Examples:

uct_hello_world.c.

6.13 UCT Memory Domain

Data Structures

- struct uct_md_attr

  Memory domain attributes. More...
- struct uct_md_attr.cap
- struct uct_allocated_memory

  Describes a memory allocated by UCT. More...
- struct uct_rkey_bundle

  Remote key with its type. More...

Typedefs

- typedef struct uct_allocated_memory uct_allocated_memory_t

  Describes a memory allocated by UCT.
- typedef struct uct_rkey_bundle uct_rkey_bundle_t

  Remote key with its type.

Enumerations

- enum uct_sockaddr_accessibility_t { UCT_SOCKADDR_ACC_LOCAL, UCT_SOCKADDR_ACC_REMOTE }

  Socket address accessibility type.
- enum {
  UCT_MD_FLAG_ALLOC = UCS_BIT(0), UCT_MD_FLAG_REG = UCS_BIT(1), UCT_MD_FLAG_NEED_↩
  MEMH = UCS_BIT(2), UCT_MD_FLAG_NEED_RKEY = UCS_BIT(3),
  UCT_MD_FLAG_ADVISE = UCS_BIT(4), UCT_MD_FLAG_FIXED = UCS_BIT(5), UCT_MD_FLAG_RKE↩
  Y_PTR = UCS_BIT(6), UCT_MD_FLAG_SOCKADDR = UCS_BIT(7) }

  Memory domain capability flags.
- enum uct_md_mem_flags {
  UCT_MD_MEM_FLAG_NONBLOCK = UCS_BIT(0), UCT_MD_MEM_FLAG_FIXED = UCS_BIT(1), UCT↩
  _MD_MEM_FLAG_LOCK = UCS_BIT(2), UCT_MD_MEM_FLAG_HIDE_ERRORS = UCS_BIT(3),
  UCT_MD_MEM_ACCESS_REMOTE_PUT = UCS_BIT(5), UCT_MD_MEM_ACCESS_REMOTE_GET =
  UCS_BIT(6), UCT_MD_MEM_ACCESS_REMOTE_ATOMIC = UCS_BIT(7), UCT_MD_MEM_ACCESS_↩
  ALL,
  UCT_MD_MEM_ACCESS_RMA }

  Memory allocation/registration flags.
- enum uct_mem_advice_t { UCT_MADV_NORMAL = 0, UCT_MADV_WILLNEED }

  list of UCT memory use advice

Functions

- ucs_status_t uct_md_query (uct_md_h md, uct_md_attr_t ∗md_attr)

  Query for memory domain attributes.
- ucs_status_t uct_md_mem_alloc (uct_md_h md, size_t ∗length_p, void ∗∗address_p, unsigned flags, const char ∗name, uct_mem_h ∗memh_p)

  Allocate memory for zero-copy sends and remote access.
- ucs_status_t uct_md_mem_free (uct_md_h md, uct_mem_h memh)

  Release memory allocated by uct_md_mem_alloc.
- ucs_status_t uct_md_mem_advise (uct_md_h md, uct_mem_h memh, void ∗addr, size_t length, uct_mem↩
  _advice_t advice)

Give advice about the use of memory.
- ucs_status_t uct_md_mem_reg (uct_md_h md, void ∗address, size_t length, unsigned flags, uct_mem_h ∗memh_p)

  Register memory for zero-copy sends and remote access.
- ucs_status_t uct_md_mem_dereg (uct_md_h md, uct_mem_h memh)

  Undo the operation of uct_md_mem_reg().
- int uct_md_is_mem_type_owned (uct_md_h md, void ∗addr, size_t length)

  Check if memory type is owned by MD.
- ucs_status_t uct_mem_alloc (void ∗addr, size_t min_length, unsigned flags, uct_alloc_method_t ∗methods, unsigned num_methods, uct_md_h ∗mds, unsigned num_mds, const char ∗name, uct_allocated_memory_t ∗mem)

  Allocate memory for zero-copy communications and remote access.
- ucs_status_t uct_mem_free (const uct_allocated_memory_t ∗mem)

  Release allocated memory.
- ucs_status_t uct_md_config_read (const char ∗name, const char ∗env_prefix, const char ∗filename, uct_md_config_t ∗∗config_p)

  Read the configuration of the MD component.
- int uct_md_is_sockaddr_accessible (uct_md_h md, const ucs_sock_addr_t ∗sockaddr, uct_sockaddr_accessibility_t mode)

  Check if remote sock address is accessible from the memory domain.
- ucs_status_t uct_md_mkey_pack (uct_md_h md, uct_mem_h memh, void ∗rkey_buffer)

  Pack a remote key.
- ucs_status_t uct_rkey_unpack (const void ∗rkey_buffer, uct_rkey_bundle_t ∗rkey_ob)

  Unpack a remote key.
- ucs_status_t uct_rkey_ptr (uct_rkey_bundle_t ∗rkey_ob, uint64_t remote_addr, void ∗∗addr_p)

  Get a local pointer to remote memory.
- ucs_status_t uct_rkey_release (const uct_rkey_bundle_t ∗rkey_ob)

  Release a remote key.

6.13.1 Detailed Description

The Memory Domain abstracts resources required for network communication, which typically includes memory, transport mechanisms, compute and network resources. It is an isolation mechanism that can be employed by the applications for isolating resources between multiple programming models. The attributes of the Memory Domain are defined by the structure uct_md_attr(). The communication and memory operations are defined in the context of Memory Domain.

6.13.2 Data Structure Documentation

6.13.2.1 struct uct_md_attr

This structure defines the attributes of a Memory Domain which includes maximum memory that can be allocated, credentials required for accessing the memory, and CPU mask indicating the proximity of CPUs.

Data Fields

struct uct_md_attr	cap	
uct_linear_growth_t	reg_cost	Memory registration cost estimation (time,seconds) as a linear function of the buffer size.

char	component_↩ name[UCT_M↩ D_COMPONE↩ NT_NAME_M↩ AX]	MD component name
size_t	rkey_packed_↩ size	Size of buffer needed for packed rkey
cpu_set_t	local_cpus	Mask of CPUs near the resource

6.13.2.2 struct uct_md_attr.cap

Data Fields

size_t	max_alloc	Maximal allocation size
size_t	max_reg	Maximal registration size
uint64_t	flags	UCT_MD_FLAG_xx
uint64_t	reg_mem_types	
uct_memory_↩ type_t	mem_type	UCS_BIT(uct_memory_type_t) Supported(owned) memory type

6.13.2.3 struct uct_allocated_memory

This structure describes the memory block which includes the address, size, and Memory Domain used for allocation. This structure is passed to interface and the memory is allocated by memory allocation functions uct_mem↩ _alloc.

Data Fields

void *	address	Address of allocated memory
size_t	length	Real size of allocated memory
uct_alloc_↩ method_t	method	Method used to allocate the memory
uct_memory_↩ type_t	mem_type	type of allocated memory
uct_md_h	md	if method==MD: MD used to allocate the memory
uct_mem_h	memh	if method==MD: MD memory handle

6.13.2.4 struct uct_rkey_bundle

This structure describes the credentials (typically key) and information required to access the remote memory by the communication interfaces.

Data Fields

uct_rkey_t	rkey	Remote key descriptor, passed to RMA functions
void *	handle	Handle, used internally for releasing the key
void *	type	Remote key type

6.13.3 Typedef Documentation

6.13.3.1 typedef struct uct_allocated_memory uct_allocated_memory_t

This structure describes the memory block which includes the address, size, and Memory Domain used for allocation. This structure is passed to interface and the memory is allocated by memory allocation functions uct_mem↩ _alloc.

6.13.3.2 typedef struct uct_rkey_bundle uct_rkey_bundle_t

This structure describes the credentials (typically key) and information required to access the remote memory by the communication interfaces.

6.13.4 Enumeration Type Documentation

6.13.4.1 enum uct_sockaddr_accessibility_t

Enumerator

UCT_SOCKADDR_ACC_LOCAL Check if local address exists. Address should belong to a local network interface

UCT_SOCKADDR_ACC_REMOTE Check if remote address can be reached. Address is routable from one of the local network interfaces

6.13.4.2 anonymous enum

Enumerator

UCT_MD_FLAG_ALLOC MD supports memory allocation

UCT_MD_FLAG_REG MD supports memory registration

UCT_MD_FLAG_NEED_MEMH The transport needs a valid local memory handle for zero-copy operations

UCT_MD_FLAG_NEED_RKEY The transport needs a valid remote memory key for remote memory operations

UCT_MD_FLAG_ADVISE MD supports memory advice

UCT_MD_FLAG_FIXED MD supports memory allocation with fixed address

UCT_MD_FLAG_RKEY_PTR MD supports direct access to remote memory via a pointer that is returned by uct_rkey_ptr

UCT_MD_FLAG_SOCKADDR MD support for client-server connection establishment via sockaddr

6.13.4.3 enum uct_md_mem_flags

Enumerator

UCT_MD_MEM_FLAG_NONBLOCK Hint to perform non-blocking allocation/registration: page mapping may be deferred until it is accessed by the CPU or a transport.

UCT_MD_MEM_FLAG_FIXED Place the mapping at exactly defined address

UCT_MD_MEM_FLAG_LOCK Registered memory should be locked. May incur extra cost for registration, but memory access is usually faster.

UCT_MD_MEM_FLAG_HIDE_ERRORS Hide errors on memory registration. In some cases registration failure is not an error (e. g. for merged memory regions).

UCT_MD_MEM_ACCESS_REMOTE_PUT enable remote put access

UCT_MD_MEM_ACCESS_REMOTE_GET enable remote get access

UCT_MD_MEM_ACCESS_REMOTE_ATOMIC enable remote atomic access

UCT_MD_MEM_ACCESS_ALL enable local and remote access for all operations

UCT_MD_MEM_ACCESS_RMA enable local and remote access for put and get operations

6.13.4.4 enum uct_mem_advice_t

Enumerator

UCT_MADV_NORMAL No special treatment

UCT_MADV_WILLNEED can be used on the memory mapped with UCT_MD_MEM_FLAG_NONBLOCK to speed up memory mapping and to avoid page faults when the memory is accessed for the first time.

6.13.5 Function Documentation

6.13.5.1 ucs_status_t uct_md_query (uct_md_h *md*, uct_md_attr_t ∗ *md_attr*)

Parameters

in	*md*	Memory domain to query.
out	*md_attr*	Filled with memory domain attributes.

6.13.5.2 ucs_status_t uct_md_mem_alloc (uct_md_h *md*, size_t ∗ *length_p*, void ∗∗ *address_p*, unsigned *flags*, const char ∗ *name*, uct_mem_h ∗ *memh_p*)

Allocate memory on the memory domain. In order to use this function, MD must support UCT_MD_FLAG_ALLOC flag.

Parameters

in	*md*	Memory domain to allocate memory on.
in,out	*length_p*	Points to the size of memory to allocate. Upon successful return, filled with the actual size that was allocated, which may be larger than the one requested. Must be >0.
in,out	*address_p*	The address
in	*flags*	Memory allocation flags, see uct_md_mem_flags.
in	*name*	Name of the allocated region, used to track memory usage for debugging and profiling.
out	*memh_p*	Filled with handle for allocated region.

6.13.5.3 ucs_status_t uct_md_mem_free (uct_md_h *md*, uct_mem_h *memh*)

Parameters

in	*md*	Memory domain memory was allocated on.
in	*memh*	Memory handle, as returned from uct_md_mem_alloc.

6.13.5.4 ucs_status_t uct_md_mem_advise (uct_md_h *md*, uct_mem_h *memh*, void ∗ *addr*, size_t *length*, uct_mem_advice_t *advice*)

This routine advises the UCT about how to handle memory range beginning at address and size of length bytes. This call does not influence the semantics of the application, but may influence its performance. The advice may be ignored.

Parameters

in	*md*	Memory domain memory was allocated or registered on.
in	*memh*	Memory handle, as returned from uct_md_mem_alloc
in	*addr*	Memory base address. Memory range must belong to the *memh*
in	*length*	Length of memory to advise. Must be >0.
in	*advice*	Memory use advice as defined in the uct_mem_advice_t list

6.13.5.5 ucs_status_t uct_md_mem_reg (uct_md_h *md,* void ∗ *address,* size_t *length,* unsigned *flags,* uct_mem_h ∗ *memh_p*)

Register memory on the memory domain. In order to use this function, MD must support UCT_MD_FLAG_REG flag.

Parameters

in	*md*	Memory domain to register memory on.
out	*address*	Memory to register.
in	*length*	Size of memory to register. Must be >0.
in	*flags*	Memory allocation flags, see uct_md_mem_flags.
out	*memh_p*	Filled with handle for allocated region.

Examples:

uct_hello_world.c.

6.13.5.6 ucs_status_t uct_md_mem_dereg (uct_md_h *md,* uct_mem_h *memh*)

Parameters

in	*md*	Memory domain which was used to register the memory.
in	*memh*	Local access key to memory region.

Examples:

uct_hello_world.c.

6.13.5.7 int uct_md_is_mem_type_owned (uct_md_h *md,* void ∗ *addr,* size_t *length*)

Check memory type.

Returns

Nonzero if memory is owned, 0 if not owned

Parameters

in	*md*	Memory domain to detect if memory belongs to.
in	*addr*	Memory address to detect.
in	*length*	Size of memory

6.13.5.8 ucs_status_t uct_mem_alloc (void ∗ *addr,* size_t *min_length,* unsigned *flags,* uct_alloc_method_t ∗ *methods,* unsigned *num_methods,* uct_md_h ∗ *mds,* unsigned *num_mds,* const char ∗ *name,* uct_allocated_memory_t ∗ *mem*)

Allocate potentially registered memory. Every one of the provided allocation methods will be used, in turn, to perform the allocation, until one succeeds. Whenever the MD method is encountered, every one of the provided MDs will be used, in turn, to allocate the memory, until one succeeds, or they are exhausted. In this case the next allocation method from the initial list will be attempted.

Parameters

in	*addr*	If *addr* is NULL, the underlying allocation routine will choose the address at which to create the mapping. If *addr* is non-NULL but UCT_MD_MEM_F↩ LAG_FIXED is not set, the address will be interpreted as a hint as to where to establish the mapping. If *addr* is non-NULL and UCT_MD_MEM_FLAG↩ _FIXED is set, then the specified address is interpreted as a requirement. In this case, if the mapping to the exact address cannot be made, the allocation request fails.
in	*min_length*	Minimal size to allocate. The actual size may be larger, for example because of alignment restrictions.
in	*flags*	Memory allocation flags, see uct_md_mem_flags.
in	*methods*	Array of memory allocation methods to attempt.
in	*num_methods*	Length of 'methods' array.
in	*mds*	Array of memory domains to attempt to allocate the memory with, for M↩ D allocation method.
in	*num_mds*	Length of 'mds' array. May be empty, in such case 'mds' may be NULL, and MD allocation method will be skipped.
in	*name*	Name of the allocation. Used for memory statistics.
out	*mem*	In case of success, filled with information about the allocated memory. uct_↩ allocated_memory_t.

6.13.5.9 ucs_status_t uct_mem_free (const uct_allocated_memory_t ∗ *mem*)

Release the memory allocated by uct_mem_alloc.

Parameters

in	*mem*	Description of allocated memory, as returned from uct_mem_alloc.

6.13.5.10 ucs_status_t uct_md_config_read (const char ∗ *name,* const char ∗ *env_prefix,* const char ∗ *filename,* uct_md_config_t ∗∗ *config_p*)

Parameters

in	*name*	Name of the MD or the MD component.
in	*env_prefix*	If non-NULL, search for environment variables starting with this UCT_↩ <prefix>_. Otherwise, search for environment variables starting with just U↩ CT_.
in	*filename*	If non-NULL, read configuration from this file. If the file does not exist, it will be ignored.
out	*config_p*	Filled with a pointer to the configuration.

Returns

Error code.

Examples:

uct_hello_world.c.

6.13.5.11 int uct_md_is_sockaddr_accessible (uct_md_h *md,* const ucs_sock_addr_t ∗ *sockaddr,* uct_sockaddr_accessibility_t *mode*)

This function checks if a remote sock address can be accessed from a local memory domain. Accessibility can be checked in local or remote mode.

Parameters

in	*md*	Memory domain to check accessibility from. This memory domain must support the UCT_MD_FLAG_SOCKADDR flag.
in	*sockaddr*	Socket address to check accessibility to.
in	*mode*	Mode for checking accessibility, as defined in uct_sockaddr_accessibility_t. Indicates if accessibility is tested on the server side - for binding to the given sockaddr, or on the client side - for connecting to the given remote peer's sockaddr.

Returns

Nonzero if accessible, 0 if inaccessible.

6.13.5.12 ucs_status_t uct_md_mkey_pack (uct_md_h *md*, uct_mem_h *memh*, void ∗ *rkey_buffer*)

Parameters

in	*md*	Handle to memory domain.
in	*memh*	Local key, whose remote key should be packed.
out	*rkey_buffer*	Filled with packed remote key.

Returns

Error code.

6.13.5.13 ucs_status_t uct_rkey_unpack (const void ∗ *rkey_buffer*, uct_rkey_bundle_t ∗ *rkey_ob*)

Parameters

in	*rkey_buffer*	Packed remote key buffer.
out	*rkey_ob*	Filled with the unpacked remote key and its type.

Returns

Error code.

6.13.5.14 ucs_status_t uct_rkey_ptr (uct_rkey_bundle_t ∗ *rkey_ob*, uint64_t *remote_addr*, void ∗∗ *addr_p*)

This routine returns a local pointer to the remote memory described by the rkey bundle. The MD must support UCT_MD_FLAG_RKEY_PTR flag.

Parameters

in	*rkey_ob*	A remote key bundle as returned by the uct_rkey_unpack function.
in	*remote_addr*	A remote address within the memory area described by the rkey_ob.
out	*addr_p*	A pointer that can be used for direct access to the remote memory.

Returns

Error code if the remote memory cannot be accessed directly or the remote address is not valid.

6.13.5.15 ucs_status_t uct_rkey_release (const uct_rkey_bundle_t ∗ *rkey_ob*)

Parameters

in	*rkey_ob*	Remote key to release.

6.14 UCT Active messages

Typedefs

- typedef ucs_status_t(∗ uct_am_callback_t) (void ∗arg, void ∗data, size_t length, unsigned flags)

  Callback to process incoming active message.
- typedef void(∗ uct_am_tracer_t) (void ∗arg, uct_am_trace_type_t type, uint8_t id, const void ∗data, size_t length, char ∗buffer, size_t max)

  Callback to trace active messages.

Enumerations

- enum uct_msg_flags { UCT_SEND_FLAG_SIGNALED = UCS_BIT(0) }

  Flags for active message send operation.
- enum uct_am_trace_type {
  UCT_AM_TRACE_TYPE_SEND, UCT_AM_TRACE_TYPE_RECV, UCT_AM_TRACE_TYPE_SEND_DR↩
  OP, UCT_AM_TRACE_TYPE_RECV_DROP,
  UCT_AM_TRACE_TYPE_LAST }

  Trace types for active message tracer.

Functions

- ucs_status_t uct_iface_set_am_handler (uct_iface_h iface, uint8_t id, uct_am_callback_t cb, void ∗arg, uint32_t flags)

  Set active message handler for the interface.
- ucs_status_t uct_iface_set_am_tracer (uct_iface_h iface, uct_am_tracer_t tracer, void ∗arg)

  Set active message tracer for the interface.
- void uct_iface_release_desc (void ∗desc)

  Release AM descriptor.
- ucs_status_t uct_ep_am_short (uct_ep_h ep, uint8_t id, uint64_t header, const void ∗payload, unsigned length)
- ssize_t uct_ep_am_bcopy (uct_ep_h ep, uint8_t id, uct_pack_callback_t pack_cb, void ∗arg, unsigned flags)
- ucs_status_t uct_ep_am_zcopy (uct_ep_h ep, uint8_t id, const void ∗header, unsigned header_length, const uct_iov_t ∗iov, size_t iovcnt, unsigned flags, uct_completion_t ∗comp)

  Send active message while avoiding local memory copy.

6.14.1 Detailed Description

Defines active message functions.

6.14.2 Typedef Documentation

6.14.2.1 typedef ucs_status_t(∗ uct_am_callback_t) (void ∗arg, void ∗data, size_t length, unsigned flags)

When the callback is called, *flags* indicates how *data* should be handled. If *flags* contain UCT_CB_PARAM_FL↩
AG_DESC value, it means *data* is part of a descriptor which must be released later by uct_iface_release_desc by the user if the callback returns UCS_INPROGRESS.

Parameters

in	*arg*	User-defined argument.
in	*data*	Points to the received data. This may be a part of a descriptor which may be released later.
in	*length*	Length of data.
in	*flags*	Mask with uct_cb_param_flags

Note

This callback could be set and released by uct_iface_set_am_handler function.

Return values

UCS_OK	- descriptor was consumed, and can be released by the caller.
UCS_INPROGRESS	- descriptor is owned by the callee, and would be released later. Supported only if *flags* contain UCT_CB_PARAM_FLAG_DESC value. Otherwise, this is an error.

6.14.2.2 typedef void(∗ uct_am_tracer_t) (void ∗arg, uct_am_trace_type_t type, uint8_t id, const void ∗data, size_t length, char ∗buffer, size_t max)

Writes a string which represents active message contents into 'buffer'.

Parameters

in	*arg*	User-defined argument.
in	*type*	Message type.
in	*id*	Active message id.
in	*data*	Points to the received data.
in	*length*	Length of data.
out	*buffer*	Filled with a debug information string.
in	*max*	Maximal length of the string.

6.14.3 Enumeration Type Documentation

6.14.3.1 enum uct_msg_flags

Enumerator

UCT_SEND_FLAG_SIGNALED Trigger UCT_EVENT_RECV_SIG event on remote side. Make best effort attempt to avoid triggering UCT_EVENT_RECV event. Ignored if not supported by interface.

6.14.3.2 enum uct_am_trace_type

Enumerator

UCT_AM_TRACE_TYPE_SEND

UCT_AM_TRACE_TYPE_RECV

UCT_AM_TRACE_TYPE_SEND_DROP

UCT_AM_TRACE_TYPE_RECV_DROP

UCT_AM_TRACE_TYPE_LAST

6.14.4 Function Documentation

6.14.4.1 ucs_status_t uct_iface_set_am_handler (uct_iface_h *iface*, uint8_t *id*, uct_am_callback_t *cb*, void * *arg*, uint32_t *flags*)

Only one handler can be set of each active message ID, and setting a handler replaces the previous value. If cb == NULL, the current handler is removed.

Parameters

in	*iface*	Interface to set the active message handler for.
in	*id*	Active message id. Must be 0..UCT_AM_ID_MAX-1.
in	*cb*	Active message callback. NULL to clear.
in	*arg*	Active message argument.
in	*flags*	Required callback flags

Returns

error code if the interface does not support active messages or requested callback flags

Examples:

uct_hello_world.c.

6.14.4.2 ucs_status_t uct_iface_set_am_tracer (uct_iface_h *iface*, uct_am_tracer_t *tracer*, void * *arg*)

Sets a function which dumps active message debug information to a buffer, which is printed every time an active message is sent or received, when data tracing is on. Without the tracer, only transport-level information is printed.

Parameters

in	*iface*	Interface to set the active message tracer for.
in	*tracer*	Active message tracer. NULL to clear.
in	*arg*	Tracer custom argument.

6.14.4.3 void uct_iface_release_desc (void * *desc*)

Release active message descriptor *desc*, which was passed to the active message callback, and owned by the callee.

Parameters

in	*desc*	Descriptor to release.

Examples:

uct_hello_world.c.

6.14.4.4 ucs_status_t uct_ep_am_short (uct_ep_h *ep*, uint8_t *id*, uint64_t *header*, const void * *payload*, unsigned *length*)

Examples:

uct_hello_world.c.

6.14.4.5 ssize_t uct_ep_am_bcopy (uct_ep_h *ep*, uint8_t *id*, uct_pack_callback_t *pack_cb*, void * *arg*, unsigned *flags*)

Examples:

uct_hello_world.c.

6.14.4.6 ucs_status_t uct_ep_am_zcopy (uct_ep_h *ep,* uint8_t *id,* const void * *header,* unsigned *header_length,* const uct_iov_t * *iov,* size_t *iovcnt,* unsigned *flags,* uct_completion_t * *comp*)

The input data in *iov* array of uct_iov_t structures sent to remote side ("gather output"). Buffers in *iov* are processed in array order. This means that the function complete iov[0] before proceeding to iov[1], and so on.

Parameters

`in`	*ep*	Destination endpoint handle.
`in`	*id*	Active message id. Must be in range 0..UCT_AM_ID_MAX-1.
`in`	*header*	Active message header.
`in`	*header_length*	Active message header length in bytes.
`in`	*iov*	Points to an array of uct_iov_t structures. The *iov* pointer must be valid address of an array of uct_iov_t structures. A particular structure pointer must be valid address. NULL terminated pointer is not required.
`in`	*iovcnt*	Size of the *iov* data uct_iov_t structures array. If *iovcnt* is zero, the data is considered empty. *iovcnt* is limited by uct_iface_attr::cap::am::max_iov
`in`	*flags*	Active message flags, see uct_msg_flags.
`in`	*comp*	Completion handle as defined by uct_completion_t.

Returns

UCS_INPROGRESS Some communication operations are still in progress. If non-NULL *comp* is provided, it will be updated upon completion of these operations.

Examples:

uct_hello_world.c.

6.15 UCT Remote memory access operations

Functions

- ucs_status_t uct_ep_put_short (uct_ep_h ep, const void ∗buffer, unsigned length, uint64_t remote_addr, uct_rkey_t rkey)
- ssize_t uct_ep_put_bcopy (uct_ep_h ep, uct_pack_callback_t pack_cb, void ∗arg, uint64_t remote_addr, uct_rkey_t rkey)
- ucs_status_t uct_ep_put_zcopy (uct_ep_h ep, const uct_iov_t ∗iov, size_t iovcnt, uint64_t remote_addr, uct←_rkey_t rkey, uct_completion_t ∗comp)

    Write data to remote memory while avoiding local memory copy.
- ucs_status_t uct_ep_get_short (uct_ep_h ep, void ∗buffer, unsigned length, uint64_t remote_addr, uct_←rkey_t rkey)
- ucs_status_t uct_ep_get_bcopy (uct_ep_h ep, uct_unpack_callback_t unpack_cb, void ∗arg, size_t length, uint64_t remote_addr, uct_rkey_t rkey, uct_completion_t ∗comp)
- ucs_status_t uct_ep_get_zcopy (uct_ep_h ep, const uct_iov_t ∗iov, size_t iovcnt, uint64_t remote_addr, uct←_rkey_t rkey, uct_completion_t ∗comp)

    Read data from remote memory while avoiding local memory copy.

6.15.1 Detailed Description

Defines remote memory access operations.

6.15.2 Function Documentation

6.15.2.1 ucs_status_t uct_ep_put_short (uct_ep_h *ep*, const void ∗ *buffer*, unsigned *length*, uint64_t *remote_addr*, uct_rkey_t *rkey*)

6.15.2.2 ssize_t uct_ep_put_bcopy (uct_ep_h *ep*, uct_pack_callback_t *pack_cb*, void ∗ *arg*, uint64_t *remote_addr*, uct_rkey_t *rkey*)

6.15.2.3 ucs_status_t uct_ep_put_zcopy (uct_ep_h *ep*, const uct_iov_t ∗ *iov*, size_t *iovcnt*, uint64_t *remote_addr*, uct_rkey_t *rkey*, uct_completion_t ∗ *comp*)

The input data in *iov* array of uct_iov_t structures sent to remote address ("gather output"). Buffers in *iov* are processed in array order. This means that the function complete iov[0] before proceeding to iov[1], and so on.

Parameters

in	*ep*	Destination endpoint handle.
in	*iov*	Points to an array of uct_iov_t structures. The *iov* pointer must be valid address of an array of uct_iov_t structures. A particular structure pointer must be valid address. NULL terminated pointer is not required.
in	*iovcnt*	Size of the *iov* data uct_iov_t structures array. If *iovcnt* is zero, the data is considered empty. *iovcnt* is limited by uct_iface_attr::cap::put::max_iov
in	*remote_addr*	Remote address to place the *iov* data.
in	*rkey*	Remote key descriptor provided by uct_rkey_unpack
in	*comp*	Completion handle as defined by uct_completion_t.

Returns

UCS_INPROGRESS Some communication operations are still in progress. If non-NULL *comp* is provided, it will be updated upon completion of these operations.

6.15.2.4 ucs_status_t uct_ep_get_short (uct_ep_h *ep,* void ∗ *buffer,* unsigned *length,* uint64_t *remote_addr,* uct_rkey_t *rkey*)

6.15.2.5 ucs_status_t uct_ep_get_bcopy (uct_ep_h *ep,* uct_unpack_callback_t *unpack_cb,* void ∗ *arg,* size_t *length,* uint64_t *remote_addr,* uct_rkey_t *rkey,* uct_completion_t ∗ *comp*)

6.15.2.6 ucs_status_t uct_ep_get_zcopy (uct_ep_h *ep,* const uct_iov_t ∗ *iov,* size_t *iovcnt,* uint64_t *remote_addr,* uct_rkey_t *rkey,* uct_completion_t ∗ *comp*)

The output data in *iov* array of uct_iov_t structures received from remote address ("scatter input"). Buffers in *iov* are processed in array order. This means that the function complete iov[0] before proceeding to iov[1], and so on.

Parameters

in	*ep*	Destination endpoint handle.
in	*iov*	Points to an array of uct_iov_t structures. The *iov* pointer must be valid address of an array of uct_iov_t structures. A particular structure pointer must be valid address. NULL terminated pointer is not required.
in	*iovcnt*	Size of the *iov* data uct_iov_t structures array. If *iovcnt* is zero, the data is considered empty. *iovcnt* is limited by uct_iface_attr::cap::get::max_iov
in	*remote_addr*	Remote address of the data placed to the *iov*.
in	*rkey*	Remote key descriptor provided by uct_rkey_unpack
in	*comp*	Completion handle as defined by uct_completion_t.

Returns

UCS_INPROGRESS Some communication operations are still in progress. If non-NULL *comp* is provided, it will be updated upon completion of these operations.

6.16 UCT Atomic operations

Functions

- ucs_status_t uct_ep_atomic_cswap64 (uct_ep_h ep, uint64_t compare, uint64_t swap, uint64_t remote_addr, uct_rkey_t rkey, uint64_t ∗result, uct_completion_t ∗comp)
- ucs_status_t uct_ep_atomic_cswap32 (uct_ep_h ep, uint32_t compare, uint32_t swap, uint64_t remote_addr, uct_rkey_t rkey, uint32_t ∗result, uct_completion_t ∗comp)
- ucs_status_t uct_ep_atomic32_post (uct_ep_h ep, uct_atomic_op_t opcode, uint32_t value, uint64_↩ t remote_addr, uct_rkey_t rkey)
- ucs_status_t uct_ep_atomic64_post (uct_ep_h ep, uct_atomic_op_t opcode, uint64_t value, uint64_↩ t remote_addr, uct_rkey_t rkey)
- ucs_status_t uct_ep_atomic32_fetch (uct_ep_h ep, uct_atomic_op_t opcode, uint32_t value, uint32_t ∗result, uint64_t remote_addr, uct_rkey_t rkey, uct_completion_t ∗comp)
- ucs_status_t uct_ep_atomic64_fetch (uct_ep_h ep, uct_atomic_op_t opcode, uint64_t value, uint64_t ∗result, uint64_t remote_addr, uct_rkey_t rkey, uct_completion_t ∗comp)

6.16.1 Detailed Description

Defines atomic operations.

6.16.2 Function Documentation

6.16.2.1 ucs_status_t uct_ep_atomic_cswap64 (uct_ep_h *ep*, uint64_t *compare*, uint64_t *swap*, uint64_t *remote_addr*, uct_rkey_t *rkey*, uint64_t ∗ *result*, uct_completion_t ∗ *comp*)

6.16.2.2 ucs_status_t uct_ep_atomic_cswap32 (uct_ep_h *ep*, uint32_t *compare*, uint32_t *swap*, uint64_t *remote_addr*, uct_rkey_t *rkey*, uint32_t ∗ *result*, uct_completion_t ∗ *comp*)

6.16.2.3 ucs_status_t uct_ep_atomic32_post (uct_ep_h *ep*, uct_atomic_op_t *opcode*, uint32_t *value*, uint64_t *remote_addr*, uct_rkey_t *rkey*)

6.16.2.4 ucs_status_t uct_ep_atomic64_post (uct_ep_h *ep*, uct_atomic_op_t *opcode*, uint64_t *value*, uint64_t *remote_addr*, uct_rkey_t *rkey*)

6.16.2.5 ucs_status_t uct_ep_atomic32_fetch (uct_ep_h *ep*, uct_atomic_op_t *opcode*, uint32_t *value*, uint32_t ∗ *result*, uint64_t *remote_addr*, uct_rkey_t *rkey*, uct_completion_t ∗ *comp*)

6.16.2.6 ucs_status_t uct_ep_atomic64_fetch (uct_ep_h *ep*, uct_atomic_op_t *opcode*, uint64_t *value*, uint64_t ∗ *result*, uint64_t *remote_addr*, uct_rkey_t *rkey*, uct_completion_t ∗ *comp*)

6.17 UCT Tag matching operations

Data Structures

- struct uct_tag_context

    Posted tag context.

Typedefs

- typedef ucs_status_t(∗ uct_tag_unexp_eager_cb_t) (void ∗arg, void ∗data, size_t length, unsigned flags, uct_tag_t stag, uint64_t imm)

    Callback to process unexpected eager tagged message.
- typedef ucs_status_t(∗ uct_tag_unexp_rndv_cb_t) (void ∗arg, unsigned flags, uint64_t stag, const void ∗header, unsigned header_length, uint64_t remote_addr, size_t length, const void ∗rkey_buf)

    Callback to process unexpected rendezvous tagged message.

Functions

- ucs_status_t uct_ep_tag_eager_short (uct_ep_h ep, uct_tag_t tag, const void ∗data, size_t length)

    Short eager tagged-send operation.
- ssize_t uct_ep_tag_eager_bcopy (uct_ep_h ep, uct_tag_t tag, uint64_t imm, uct_pack_callback_t pack_cb, void ∗arg, unsigned flags)

    Bcopy eager tagged-send operation.
- ucs_status_t uct_ep_tag_eager_zcopy (uct_ep_h ep, uct_tag_t tag, uint64_t imm, const uct_iov_t ∗iov, size↩_t iovcnt, unsigned flags, uct_completion_t ∗comp)

    Zcopy eager tagged-send operation.
- ucs_status_ptr_t uct_ep_tag_rndv_zcopy (uct_ep_h ep, uct_tag_t tag, const void ∗header, unsigned header_length, const uct_iov_t ∗iov, size_t iovcnt, unsigned flags, uct_completion_t ∗comp)

    Rendezvous tagged-send operation.
- ucs_status_t uct_ep_tag_rndv_cancel (uct_ep_h ep, void ∗op)

    Cancel outstanding rendezvous operation.
- ucs_status_t uct_ep_tag_rndv_request (uct_ep_h ep, uct_tag_t tag, const void ∗header, unsigned header↩_length, unsigned flags)

    Send software rendezvous request.
- ucs_status_t uct_iface_tag_recv_zcopy (uct_iface_h iface, uct_tag_t tag, uct_tag_t tag_mask, const uct_↩iov_t ∗iov, size_t iovcnt, uct_tag_context_t ∗ctx)

    Post a tag to a transport interface.
- ucs_status_t uct_iface_tag_recv_cancel (uct_iface_h iface, uct_tag_context_t ∗ctx, int force)

    Cancel a posted tag.

6.17.1 Detailed Description

Defines tag matching operations.

6.17.2 Typedef Documentation

6.17.2.1 typedef ucs_status_t(∗ uct_tag_unexp_eager_cb_t) (void ∗arg, void ∗data, size_t length, unsigned flags, uct_tag_t stag, uint64_t imm)

This callback is invoked when tagged message sent by eager protocol has arrived and no corresponding tag has been posted.

Note

The callback is always invoked from the context (thread, process) that called *uct_iface_progress()*.
It is allowed to call other communication routines from the callback.

Parameters

in	*arg*	User-defined argument
in	*data*	Points to the received unexpected data.
in	*length*	Length of data.
in	*desc*	Points to the received descriptor, at the beginning of the user-defined rx_↩ headroom.
in	*stag*	Tag from sender.
in	*imm*	Immediate data from sender.

Warning

If the user became the owner of the *desc* (by returning UCS_INPROGRESS) the descriptor must be released later by uct_iface_release_desc by the user.

Return values

UCS_OK	- descriptor was consumed, and can be released by the caller.
UCS_INPROGRESS	- descriptor is owned by the callee, and would be released later.

6.17.2.2 typedef ucs_status_t(∗ uct_tag_unexp_rndv_cb_t) (void ∗arg, unsigned flags, uint64_t stag, const void ∗header, unsigned header_length, uint64_t remote_addr, size_t length, const void ∗rkey_buf)

This callback is invoked when rendezvous send notification has arrived and no corresponding tag has been posted.

Note

The callback is always invoked from the context (thread, process) that called *uct_iface_progress()*.
It is allowed to call other communication routines from the callback.

Parameters

in	*arg*	User-defined argument
in	*flags*	Mask with uct_cb_param_flags
in	*stag*	Tag from sender.
in	*header*	User defined header.
in	*header_length*	User defined header length in bytes.
in	*remote_addr*	Sender's buffer virtual address.
in	*length*	Sender's buffer length.
in	*rkey_buf*	Sender's buffer packed remote key. It can be passed to uct_rkey_unpack() to create uct_rkey_t.

Warning

If the user became the owner of the *desc* (by returning UCS_INPROGRESS) the descriptor must be released later by uct_iface_release_desc by the user.

Return values

UCS_OK	- descriptor was consumed, and can be released by the caller.
UCS_INPROGRESS	- descriptor is owned by the callee, and would be released later.

6.17.3 Function Documentation

6.17.3.1 ucs_status_t uct_ep_tag_eager_short (uct_ep_h *ep*, uct_tag_t *tag*, const void * *data*, size_t *length*)

This routine sends a message using short eager protocol. Eager protocol means that the whole data is sent to the peer immediately without any preceding notification. The data is provided as buffer and its length,and must not be larger than the corresponding *max_short* value in uct_iface_attr. The immediate value delivered to the receiver is implicitly equal to 0. If it's required to pass non-zero imm value, uct_ep_tag_eager_bcopy should be used.

Parameters

in	*ep*	Destination endpoint handle.
in	*tag*	Tag to use for the eager message.
in	*data*	Data to send.
in	*length*	Data length.

Returns

UCS_OK - operation completed successfully.
UCS_ERR_NO_RESOURCE - could not start the operation now due to lack of send resources.

6.17.3.2 ssize_t uct_ep_tag_eager_bcopy (uct_ep_h *ep*, uct_tag_t *tag*, uint64_t *imm*, uct_pack_callback_t *pack_cb*, void * *arg*, unsigned *flags*)

This routine sends a message using bcopy eager protocol. Eager protocol means that the whole data is sent to the peer immediately without any preceding notification. Custom data callback is used to copy the data to the network buffers.

Note

The resulted data length must not be larger than the corresponding *max_bcopy* value in uct_iface_attr.

Parameters

in	*ep*	Destination endpoint handle.
in	*tag*	Tag to use for the eager message.
in	*imm*	Immediate value which will be available to the receiver.
in	*pack_cb*	User callback to pack the data.
in	*arg*	Custom argument to *pack_cb*.
in	*flags*	Tag message flags, see uct_msg_flags.

Returns

$>=$0 - The size of the data packed by *pack_cb*.
otherwise - Error code.

6.17.3.3 ucs_status_t uct_ep_tag_eager_zcopy (uct_ep_h *ep*, uct_tag_t *tag*, uint64_t *imm*, const uct_iov_t * *iov*, size_t *iovcnt*, unsigned *flags*, uct_completion_t * *comp*)

This routine sends a message using zcopy eager protocol. Eager protocol means that the whole data is sent to the peer immediately without any preceding notification. The input data (which has to be previously registered) in *iov* array of uct_iov_t structures sent to remote side ("gather output"). Buffers in *iov* are processed in array order, so the function complete *iov*[0] before proceeding to *iov*[1], and so on.

Note

The resulted data length must not be larger than the corresponding *max_zcopy* value in uct_iface_attr.

Parameters

in	*ep*	Destination endpoint handle.
in	*tag*	Tag to use for the eager message.
in	*imm*	Immediate value which will be available to the receiver.
in	*iov*	Points to an array of uct_iov_t structures. A particular structure pointer must be valid address. NULL terminated pointer is not required.
in	*iovcnt*	Size of the *iov* array. If *iovcnt* is zero, the data is considered empty. Note that *iovcnt* is limited by the corresponding *max_iov* value in uct_iface_attr.
in	*flags*	Tag message flags, see uct_msg_flags.
in	*comp*	Completion callback which will be called when the data is reliably received by the peer, and the buffer can be reused or invalidated.

Returns

UCS_OK - operation completed successfully.
UCS_ERR_NO_RESOURCE - could not start the operation now due to lack of send resources.
UCS_INPROGRESS - operation started, and *comp* will be used to notify when it's completed.

6.17.3.4 ucs_status_ptr_t uct_ep_tag_rndv_zcopy (uct_ep_h *ep,* uct_tag_t *tag,* const void ∗ *header,* unsigned *header_length,* const uct_iov_t ∗ *iov,* size_t *iovcnt,* unsigned *flags,* uct_completion_t ∗ *comp*)

This routine sends a message using rendezvous protocol. Rendezvous protocol means that only a small notification is sent at first, and the data itself is transferred later (when there is a match) to avoid extra memory copy.

Note

The header will be available to the receiver in case of unexpected rendezvous operation only, i.e. the peer has not posted tag for this message yet (by means of uct_iface_tag_recv_zcopy), when it is arrived.

Parameters

in	*ep*	Destination endpoint handle.
in	*tag*	Tag to use for the eager message.
in	*header*	User defined header.
in	*header_length*	User defined header length in bytes. Note that it is limited by the corresponding *max_hdr* value in uct_iface_attr.
in	*iov*	Points to an array of uct_iov_t structures. A particular structure pointer must be valid address. NULL terminated pointer is not required.
in	*iovcnt*	Size of the *iov* array. If *iovcnt* is zero, the data is considered empty. Note that *iovcnt* is limited by the corresponding *max_iov* value in uct_iface_attr.
in	*flags*	Tag message flags, see uct_msg_flags.
in	*comp*	Completion callback which will be called when the data is reliably received by the peer, and the buffer can be reused or invalidated.

Returns

$>=0$ - The operation is in progress and the return value is a handle which can be used to cancel the outstanding rendezvous operation.
otherwise - Error code.

6.17.3.5 ucs_status_t uct_ep_tag_rndv_cancel (uct_ep_h *ep*, void ∗ *op*)

This routine signals the underlying transport disregard the outstanding operation without calling completion callback provided in uct_ep_tag_rndv_zcopy.

Note

The operation handle should be valid at the time the routine is invoked. I.e. it should be a handle of the real operation which is not completed yet.

Parameters

in	*ep*	Destination endpoint handle.
in	*op*	Rendezvous operation handle, as returned from uct_ep_tag_rndv_zcopy.

Returns

UCS_OK - The operation has been canceled.

6.17.3.6 ucs_status_t uct_ep_tag_rndv_request (uct_ep_h *ep*, uct_tag_t *tag*, const void ∗ *header*, unsigned *header_length*, unsigned *flags*)

This routine sends a rendezvous request only, which indicates that the data transfer should be completed in software.

Parameters

in	*ep*	Destination endpoint handle.
in	*tag*	Tag to use for matching.
in	*header*	User defined header
in	*header_length*	User defined header length in bytes. Note that it is limited by the corresponding *max_hdr* value in uct_iface_attr.
in	*flags*	Tag message flags, see uct_msg_flags.

Returns

UCS_OK - operation completed successfully.
UCS_ERR_NO_RESOURCE - could not start the operation now due to lack of send resources.

6.17.3.7 ucs_status_t uct_iface_tag_recv_zcopy (uct_iface_h *iface*, uct_tag_t *tag*, uct_tag_t *tag_mask*, const uct_iov_t ∗ *iov*, size_t *iovcnt*, uct_tag_context_t ∗ *ctx*)

This routine posts a tag to be matched on a transport interface. When a message with the corresponding tag arrives it is stored in the user buffer (described by *iov* and *iovcnt*) directly. The operation completion is reported using callbacks on the *ctx* structure.

Parameters

in	*iface*	Interface to post the tag on.
in	*tag*	Tag to expect.
in	*tag_mask*	Mask which specifies what bits of the tag to compare.
in	*iov*	Points to an array of uct_iov_t structures. The *iov* pointer must be valid address of an array of uct_iov_t structures. A particular structure pointer must be valid address. NULL terminated pointer is not required.

in	*iovcnt*	Size of the *iov* data uct_iov_t structures array. If *iovcnt* is zero, the data is considered empty. *iovcnt* is limited by uct_iface_attr::cap::tag::max_iov
in,out	*ctx*	Context associated with this particular tag, "priv" field in this structure is used to track the state internally.

Returns

UCS_OK - The tag is posted to the transport.
UCS_ERR_NO_RESOURCE - Could not start the operation due to lack of resources.
UCS_ERR_EXCEEDS_LIMIT - No more room for tags in the transport.

6.17.3.8 ucs_status_t uct_iface_tag_recv_cancel (uct_iface_h *iface,* uct_tag_context_t * *ctx,* int *force*)

This routine cancels a tag, which was previously posted by uct_iface_tag_recv_zcopy. The tag would be either matched or canceled, in a bounded time, regardless of the peer actions. The original completion callback of the tag would be called with the status if *force* is not set.

Parameters

in	*iface*	Interface to cancel the tag on.
in	*ctx*	Tag context which was used for posting the tag. If force is 0, *ctx->completed_cb* will be called with either UCS_OK which means the tag was matched and data received despite the cancel request, or UCS_ERR_CANCELED which means the tag was successfully canceled before it was matched.
in	*force*	Whether to report completions to *ctx->completed_cb*. If nonzero, the cancel is assumed to be successful, and the callback is not called.

Returns

UCS_OK - The tag is canceled in the transport.

6.18 UCT interface operations and capabilities

List of capabilities supported by UCX API.

Macros

- #define UCT_IFACE_FLAG_AM_SHORT UCS_BIT(0)
- #define UCT_IFACE_FLAG_AM_BCOPY UCS_BIT(1)
- #define UCT_IFACE_FLAG_AM_ZCOPY UCS_BIT(2)
- #define UCT_IFACE_FLAG_PENDING UCS_BIT(3)
- #define UCT_IFACE_FLAG_PUT_SHORT UCS_BIT(4)
- #define UCT_IFACE_FLAG_PUT_BCOPY UCS_BIT(5)
- #define UCT_IFACE_FLAG_PUT_ZCOPY UCS_BIT(6)
- #define UCT_IFACE_FLAG_GET_SHORT UCS_BIT(8)
- #define UCT_IFACE_FLAG_GET_BCOPY UCS_BIT(9)
- #define UCT_IFACE_FLAG_GET_ZCOPY UCS_BIT(10)
- #define UCT_IFACE_FLAG_ATOMIC_CPU UCS_BIT(30)
- #define UCT_IFACE_FLAG_ATOMIC_DEVICE UCS_BIT(31)
- #define UCT_IFACE_FLAG_ERRHANDLE_SHORT_BUF UCS_BIT(32)
- #define UCT_IFACE_FLAG_ERRHANDLE_BCOPY_BUF UCS_BIT(33)
- #define UCT_IFACE_FLAG_ERRHANDLE_ZCOPY_BUF UCS_BIT(34)
- #define UCT_IFACE_FLAG_ERRHANDLE_AM_ID UCS_BIT(35)
- #define UCT_IFACE_FLAG_ERRHANDLE_REMOTE_MEM UCS_BIT(36)
- #define UCT_IFACE_FLAG_ERRHANDLE_BCOPY_LEN UCS_BIT(37)
- #define UCT_IFACE_FLAG_ERRHANDLE_PEER_FAILURE UCS_BIT(38)
- #define UCT_IFACE_FLAG_EP_CHECK UCS_BIT(39)
- #define UCT_IFACE_FLAG_CONNECT_TO_IFACE UCS_BIT(40)
- #define UCT_IFACE_FLAG_CONNECT_TO_EP UCS_BIT(41)
- #define UCT_IFACE_FLAG_CONNECT_TO_SOCKADDR UCS_BIT(42)
- #define UCT_IFACE_FLAG_AM_DUP UCS_BIT(43)
- #define UCT_IFACE_FLAG_CB_SYNC UCS_BIT(44)
- #define UCT_IFACE_FLAG_CB_ASYNC UCS_BIT(45)
- #define UCT_IFACE_FLAG_EVENT_SEND_COMP UCS_BIT(46)
- #define UCT_IFACE_FLAG_EVENT_RECV UCS_BIT(47)
- #define UCT_IFACE_FLAG_EVENT_RECV_SIG UCS_BIT(48)
- #define UCT_IFACE_FLAG_TAG_EAGER_SHORT UCS_BIT(50)
- #define UCT_IFACE_FLAG_TAG_EAGER_BCOPY UCS_BIT(51)
- #define UCT_IFACE_FLAG_TAG_EAGER_ZCOPY UCS_BIT(52)
- #define UCT_IFACE_FLAG_TAG_RNDV_ZCOPY UCS_BIT(53)

6.18.1 Detailed Description

The definition list presents a full list of operations and capabilities exposed by UCX API.

6.18.2 Macro Definition Documentation

6.18.2.1 #define UCT_IFACE_FLAG_AM_SHORT UCS_BIT(0)

Short active message

Examples:

uct_hello_world.c.

6.18.2.2 #define UCT_IFACE_FLAG_AM_BCOPY UCS_BIT(1)

Buffered active message

Examples:

uct_hello_world.c.

6.18.2.3 #define UCT_IFACE_FLAG_AM_ZCOPY UCS_BIT(2)

Zero-copy active message

Examples:

uct_hello_world.c.

6.18.2.4 #define UCT_IFACE_FLAG_PENDING UCS_BIT(3)

Pending operations

6.18.2.5 #define UCT_IFACE_FLAG_PUT_SHORT UCS_BIT(4)

Short put

6.18.2.6 #define UCT_IFACE_FLAG_PUT_BCOPY UCS_BIT(5)

Buffered put

6.18.2.7 #define UCT_IFACE_FLAG_PUT_ZCOPY UCS_BIT(6)

Zero-copy put

6.18.2.8 #define UCT_IFACE_FLAG_GET_SHORT UCS_BIT(8)

Short get

6.18.2.9 #define UCT_IFACE_FLAG_GET_BCOPY UCS_BIT(9)

Buffered get

6.18.2.10 #define UCT_IFACE_FLAG_GET_ZCOPY UCS_BIT(10)

Zero-copy get

6.18.2.11 #define UCT_IFACE_FLAG_ATOMIC_CPU UCS_BIT(30)

Atomic communications are consistent with respect to CPU operations.

6.18.2.12 #define UCT_IFACE_FLAG_ATOMIC_DEVICE UCS_BIT(31)

Atomic communications are consistent only with respect to other atomics on the same device.

6.18.2.13 #define UCT_IFACE_FLAG_ERRHANDLE_SHORT_BUF UCS_BIT(32)

Invalid buffer for short operation

6.18.2.14 #define UCT_IFACE_FLAG_ERRHANDLE_BCOPY_BUF UCS_BIT(33)

Invalid buffer for buffered operation

6.18.2.15 #define UCT_IFACE_FLAG_ERRHANDLE_ZCOPY_BUF UCS_BIT(34)

Invalid buffer for zero copy operation

6.18.2.16 #define UCT_IFACE_FLAG_ERRHANDLE_AM_ID UCS_BIT(35)

Invalid AM id on remote

6.18.2.17 #define UCT_IFACE_FLAG_ERRHANDLE_REMOTE_MEM UCS_BIT(36)

Remote memory access

6.18.2.18 #define UCT_IFACE_FLAG_ERRHANDLE_BCOPY_LEN UCS_BIT(37)

Invalid length for buffered operation

6.18.2.19 #define UCT_IFACE_FLAG_ERRHANDLE_PEER_FAILURE UCS_BIT(38)

Remote peer failures/outage

6.18.2.20 #define UCT_IFACE_FLAG_EP_CHECK UCS_BIT(39)

Endpoint check

6.18.2.21 #define UCT_IFACE_FLAG_CONNECT_TO_IFACE UCS_BIT(40)

Supports connecting to interface

Examples:

uct_hello_world.c.

6.18.2.22 #define UCT_IFACE_FLAG_CONNECT_TO_EP UCS_BIT(41)

Supports connecting to specific endpoint

Examples:

uct_hello_world.c.

6.18.2.23 #define UCT_IFACE_FLAG_CONNECT_TO_SOCKADDR UCS_BIT(42)

Supports connecting to sockaddr

6.18.2.24 #define UCT_IFACE_FLAG_AM_DUP UCS_BIT(43)

Active messages may be received with duplicates This happens if the transport does not keep enough information to detect retransmissions

6.18.2.25 #define UCT_IFACE_FLAG_CB_SYNC UCS_BIT(44)

Interface supports setting a callback which is invoked only from the calling context of uct_worker_progress()

6.18.2.26 #define UCT_IFACE_FLAG_CB_ASYNC UCS_BIT(45)

Interface supports setting a callback which will be invoked within a reasonable amount of time if uct_worker_progress() is not being called. The callback can be invoked from any progress context and it may also be invoked when uct_worker_progress() is called.

6.18.2.27 #define UCT_IFACE_FLAG_EVENT_SEND_COMP UCS_BIT(46)

Event notification of send completion is supported

6.18.2.28 #define UCT_IFACE_FLAG_EVENT_RECV UCS_BIT(47)

Event notification of tag and active message receive is supported

6.18.2.29 #define UCT_IFACE_FLAG_EVENT_RECV_SIG UCS_BIT(48)

Event notification of signaled tag and active message is supported

6.18.2.30 #define UCT_IFACE_FLAG_TAG_EAGER_SHORT UCS_BIT(50)

Hardware tag matching short eager support

6.18.2.31 #define UCT_IFACE_FLAG_TAG_EAGER_BCOPY UCS_BIT(51)

Hardware tag matching bcopy eager support

6.18.2.32 #define UCT_IFACE_FLAG_TAG_EAGER_ZCOPY UCS_BIT(52)

Hardware tag matching zcopy eager support

6.18.2.33 #define UCT_IFACE_FLAG_TAG_RNDV_ZCOPY UCS_BIT(53)

Hardware tag matching rendezvous zcopy support

6.19 Unified Communication Services (UCS) API

Modules

- UCS Communication Resource

6.19.1 Detailed Description

This section describes UCS API.

6.20 UCS Communication Resource

Data Structures

- struct ucs_sock_addr

Typedefs

- typedef void(∗ ucs_async_event_cb_t) (int id, void ∗arg)
- typedef struct ucs_sock_addr ucs_sock_addr_t
- typedef unsigned long ucs_time_t
- typedef void ∗ ucs_status_ptr_t

    Status pointer.

Enumerations

- enum ucs_callbackq_flags { UCS_CALLBACKQ_FLAG_FAST = UCS_BIT(0), UCS_CALLBACKQ_FLAG↩ _ONESHOT = UCS_BIT(1) }
- enum ucs_status_t {
  UCS_OK = 0, UCS_INPROGRESS = 1, UCS_ERR_NO_MESSAGE = -1, UCS_ERR_NO_RESOURCE = -2,
  UCS_ERR_IO_ERROR = -3, UCS_ERR_NO_MEMORY = -4, UCS_ERR_INVALID_PARAM = -5, UCS_E↩ RR_UNREACHABLE = -6,
  UCS_ERR_INVALID_ADDR = -7, UCS_ERR_NOT_IMPLEMENTED = -8, UCS_ERR_MESSAGE_TRUN↩ CATED = -9, UCS_ERR_NO_PROGRESS = -10,
  UCS_ERR_BUFFER_TOO_SMALL = -11, UCS_ERR_NO_ELEM = -12, UCS_ERR_SOME_CONNECTS↩ _FAILED = -13, UCS_ERR_NO_DEVICE = -14,
  UCS_ERR_BUSY = -15, UCS_ERR_CANCELED = -16, UCS_ERR_SHMEM_SEGMENT = -17, UCS_ER↩ R_ALREADY_EXISTS = -18,
  UCS_ERR_OUT_OF_RANGE = -19, UCS_ERR_TIMED_OUT = -20, UCS_ERR_EXCEEDS_LIMIT = -21, UCS_ERR_UNSUPPORTED = -22,
  UCS_ERR_REJECTED = -23, UCS_ERR_FIRST_LINK_FAILURE = -40, UCS_ERR_LAST_LINK_FAILU↩ RE = -59, UCS_ERR_FIRST_ENDPOINT_FAILURE = -60,
  UCS_ERR_LAST_ENDPOINT_FAILURE = -79, UCS_ERR_ENDPOINT_TIMEOUT = -80, UCS_ERR_LA↩ ST = -100 }

    Status codes.
- enum ucs_thread_mode_t { UCS_THREAD_MODE_SINGLE, UCS_THREAD_MODE_SERIALIZED, UC↩ S_THREAD_MODE_MULTI, UCS_THREAD_MODE_LAST }

    Thread sharing mode.

Functions

- ucs_status_t ucs_async_set_event_handler (ucs_async_mode_t mode, int event_fd, int events, ucs_async↩ _event_cb_t cb, void ∗arg, ucs_async_context_t ∗async)
- ucs_status_t ucs_async_add_timer (ucs_async_mode_t mode, ucs_time_t interval, ucs_async_event_cb_t cb, void ∗arg, ucs_async_context_t ∗async, int ∗timer_id_p)
- ucs_status_t ucs_async_remove_handler (int id, int sync)
- ucs_status_t ucs_async_modify_handler (int fd, int events)
- ucs_status_t ucs_async_context_create (ucs_async_mode_t mode, ucs_async_context_t ∗∗async_p)

    Create an asynchronous execution context.
- void ucs_async_context_destroy (ucs_async_context_t ∗async)

    Destroy the asynchronous execution context.
- void ucs_async_poll (ucs_async_context_t ∗async)

6.20.1 Detailed Description

This section describes a concept of the Communication Resource and routines associated with the concept.

6.20.2 Data Structure Documentation

6.20.2.1 struct ucs_sock_addr

BSD socket address specification.

Data Fields

const struct sockaddr *	addr	Pointer to socket address
socklen_t	addrlen	Address length

6.20.3 Typedef Documentation

6.20.3.1 typedef void(* ucs_async_event_cb_t) (int id, void *arg)

Async event callback.

Parameters

id	Event id (timer or file descriptor).
arg	User-defined argument.

6.20.3.2 typedef struct ucs_sock_addr ucs_sock_addr_t

BSD socket address specification.

6.20.3.3 typedef unsigned long ucs_time_t

Copyright (C) Mellanox Technologies Ltd. 2001-2017. ALL RIGHTS RESERVED.

See file LICENSE for terms.

UCS time units. These are not necessarily aligned with metric time units. MUST compare short time values with UCS_SHORT_TIME_CMP to handle wrap-around.

6.20.3.4 typedef void* ucs_status_ptr_t

A pointer can represent one of these values:

- NULL / UCS_OK
- Error code pointer (UCS_ERR_xx)
- Valid pointer

6.20.4 Enumeration Type Documentation

6.20.4.1 enum ucs_callbackq_flags

Callback flags

Enumerator

UCS_CALLBACKQ_FLAG_FAST Fast-path (best effort)

UCS_CALLBACKQ_FLAG_ONESHOT Call the callback only once (cannot be used with FAST)

6.20.4.2 enum ucs_status_t

Note

In order to evaluate the necessary steps to recover from a certain error, all error codes which can be returned by the external API are grouped by the largest entity permanently effected by the error. Each group ranges between its UCS_ERR_FIRST_<name> and UCS_ERR_LAST_<name> enum values. For example, if a link fails it may be sufficient to destroy (and possibly replace) it, in contrast to an endpoint-level error.

Enumerator

UCS_OK

UCS_INPROGRESS

UCS_ERR_NO_MESSAGE

UCS_ERR_NO_RESOURCE

UCS_ERR_IO_ERROR

UCS_ERR_NO_MEMORY

UCS_ERR_INVALID_PARAM

UCS_ERR_UNREACHABLE

UCS_ERR_INVALID_ADDR

UCS_ERR_NOT_IMPLEMENTED

UCS_ERR_MESSAGE_TRUNCATED

UCS_ERR_NO_PROGRESS

UCS_ERR_BUFFER_TOO_SMALL

UCS_ERR_NO_ELEM

UCS_ERR_SOME_CONNECTS_FAILED

UCS_ERR_NO_DEVICE

UCS_ERR_BUSY

UCS_ERR_CANCELED

UCS_ERR_SHMEM_SEGMENT

UCS_ERR_ALREADY_EXISTS

UCS_ERR_OUT_OF_RANGE

UCS_ERR_TIMED_OUT

UCS_ERR_EXCEEDS_LIMIT

UCS_ERR_UNSUPPORTED

UCS_ERR_REJECTED

UCS_ERR_FIRST_LINK_FAILURE

UCS_ERR_LAST_LINK_FAILURE

UCS_ERR_FIRST_ENDPOINT_FAILURE

UCS_ERR_LAST_ENDPOINT_FAILURE

UCS_ERR_ENDPOINT_TIMEOUT

UCS_ERR_LAST

6.20.4.3 enum ucs_thread_mode_t

Specifies thread sharing mode of an object.

Enumerator

UCS_THREAD_MODE_SINGLE Only the master thread can access (i.e. the thread that initialized the context; multiple threads may exist and never access)

UCS_THREAD_MODE_SERIALIZED Multiple threads can access, but only one at a time

UCS_THREAD_MODE_MULTI Multiple threads can access concurrently

UCS_THREAD_MODE_LAST

6.20.5 Function Documentation

6.20.5.1 ucs_status_t ucs_async_set_event_handler (ucs_async_mode_t *mode*, int *event_fd*, int *events*, ucs_async_event_cb_t *cb*, void * *arg*, ucs_async_context_t * *async*)

Register a file descriptor for monitoring (call handler upon events). Every fd can have only one handler.

Parameters

mode	Thread or signal.
event_fd	File descriptor to set handler for.
events	Events to wait on (POLLxx/EPOLLxx bits).
cb	Callback function to execute.
arg	Argument to callback.
async	Async context to which events are delivered. If NULL, safety is up to the user.

Returns

Error code as defined by ucs_status_t.

6.20.5.2 ucs_status_t ucs_async_add_timer (ucs_async_mode_t *mode*, ucs_time_t *interval*, ucs_async_event_cb_t *cb*, void * *arg*, ucs_async_context_t * *async*, int * *timer_id_p*)

Add timer handler.

Parameters

mode	Thread or signal.
interval	Timer interval.
cb	Callback function to execute.
arg	Argument to callback.
async	Async context to which events are delivered. If NULL, safety is up to the user.
timer_id_p	Filled with timer id.

Returns

Error code as defined by ucs_status_t.

6.20.5.3 ucs_status_t ucs_async_remove_handler (int *id*, int *sync*)

Remove an event handler (Timer or event file).

Parameters

id	Timer/FD to remove.
sync	If nonzero, wait until the handler for this event is not running anymore. Cannot be used in the context of the event handler itself because it would deadlock.

Returns

Error code as defined by ucs_status_t.

6.20.5.4 ucs_status_t ucs_async_modify_handler (int *fd*, int *events*)

Modify events mask for an existing event handler (event file).

Parameters

fd	File descriptor modify events for.
events	New set of events to wait on (POLLxx/EPOLLxx bits).

Returns

Error code as defined by ucs_status_t.

6.20.5.5 ucs_status_t ucs_async_context_create (ucs_async_mode_t *mode*, ucs_async_context_t ∗∗ *async_p*)

Allocate and initialize an asynchronous execution context. This can be used to ensure safe event delivery.

Parameters

mode	Either to use signals or epoll threads to wait.
async_p	Event context pointer to initialize.

Returns

Error code as defined by ucs_status_t.

Examples:

uct_hello_world.c.

6.20.5.6 void ucs_async_context_destroy (ucs_async_context_t ∗ *async*)

Clean up the async context, and release system resources if possible. The context memory released.

Parameters

async	Asynchronous context to clean up.

Examples:

uct_hello_world.c.

6.20.5.7 void ucs_async_poll (ucs_async_context_t ∗ *async*)

Poll on async context.

Parameters

async	Async context to poll on. NULL polls on all.

Chapter 7

Data Structure Documentation

7.1 ucp_generic_dt_ops Struct Reference

UCP generic data type descriptor.

Data Fields

- void ∗(∗ start_pack)(void ∗context, const void ∗buffer, size_t count)

  Start a packing request.
- void ∗(∗ start_unpack)(void ∗context, void ∗buffer, size_t count)

  Start an unpacking request.
- size_t(∗ packed_size)(void ∗state)

  Get the total size of packed data.
- size_t(∗ pack)(void ∗state, size_t offset, void ∗dest, size_t max_length)

  Pack data.
- ucs_status_t(∗ unpack)(void ∗state, size_t offset, const void ∗src, size_t length)

  Unpack data.
- void(∗ finish)(void ∗state)

  Finish packing/unpacking.

7.1.1 Detailed Description

This structure provides a generic datatype descriptor that is used for definition of application defined datatypes.

Typically, the descriptor is used for an integration with datatype engines implemented within MPI and SHMEM implementations.

Note

In case of partial receive, any amount of received data is acceptable which matches buffer size.

The documentation for this struct was generated from the following file:

- ucp.h

7.2 uct_tag_context Struct Reference

Posted tag context.

Data Fields

- void(* tag_consumed_cb)(uct_tag_context_t *self)
- void(* completed_cb)(uct_tag_context_t *self, uct_tag_t stag, uint64_t imm, size_t length, ucs_status_t status)
- void(* rndv_cb)(uct_tag_context_t *self, uct_tag_t stag, const void *header, unsigned header_length, ucs_status_t status)
- char priv [UCT_TAG_PRIV_LEN]

7.2.1 Detailed Description

Tag context is an object which tracks a tag posted to the transport. It contains callbacks for matching events on this tag.

7.2.2 Field Documentation

7.2.2.1 void(* uct_tag_context::tag_consumed_cb) (uct_tag_context_t *self)

Tag is consumed by the transport and should not be matched in software.

Parameters

in	*self*	Pointer to relevant context structure, which was initially passed to uct_iface_tag_recv_zcopy.

7.2.2.2 void(* uct_tag_context::completed_cb) (uct_tag_context_t *self, uct_tag_t stag, uint64_t imm, size_t length, ucs_status_t status)

Tag processing is completed by the transport.

Parameters

in	*self*	Pointer to relevant context structure, which was initially passed to uct_iface_tag_recv_zcopy.
in	*stag*	Tag from sender.
in	*imm*	Immediate data from sender. For rendezvous, it's always 0.
in	*length*	Completed length.
in	*status*	Completion status: (a) UCS_OK - Success, data placed in provided buffer. (b) UCS_ERR_TRUNCATED - Sender's length exceed posted buffer, no data is copied. (c) UCS_ERR_CANCELED - Canceled by user.

7.2.2.3 void(* uct_tag_context::rndv_cb) (uct_tag_context_t *self, uct_tag_t stag, const void *header, unsigned header_length, ucs_status_t status)

Tag was matched by a rendezvous request, which should be completed by the protocol layer.

Parameters

in	*self*	Pointer to relevant context structure, which was initially passed to uct_iface_tag_recv_zcopy.
in	*stag*	Tag from sender.

in	*header*	User defined header.
in	*header_length*	User defined header length in bytes.
in	*status*	Completion status.

7.2.2.4 char uct_tag_context::priv[UCT_TAG_PRIV_LEN]

A placeholder for the private data used by the transport

The documentation for this struct was generated from the following file:

- uct.h

Chapter 8

Example Documentation

8.1 ucp_hello_world.c

UCP hello world client / server example utility.

```
#ifndef HAVE_CONFIG_H
#  define HAVE_CONFIG_H /* Force using config.h, so test would fail if header
                           actually tries to use it */
#endif

/*
 * UCP hello world client / server example utility
 * -----------------------------------------------
 *
 * Server side:
 *
 *    ./ucp_hello_world
 *
 * Client side:
 *
 *    ./ucp_hello_world -n <server host name>
 *
 * Notes:
 *
 *    - Client acquires Server UCX address via TCP socket
 *
 *
 * Author:
 *
 *    Ilya Nelkenbaum <ilya@nelkenbaum.com>
 *    Sergey Shalnov <sergeysh@mellanox.com> 7-June-2016
 */

#include "ucx_hello_world.h"

#include <ucp/api/ucp.h>

#include <sys/socket.h>
#include <sys/types.h>
#include <sys/epoll.h>
#include <netinet/in.h>
#include <assert.h>
#include <netdb.h>
#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <unistd.h>  /* getopt */
#include <ctype.h>   /* isprint */
#include <pthread.h> /* pthread_self */
#include <errno.h>   /* errno */
#include <time.h>
#include <signal.h>  /* raise */

struct msg {
    uint64_t        data_len;
};

struct ucx_context {
    int             completed;
};
```

```
enum ucp_test_mode_t {
    TEST_MODE_PROBE,
    TEST_MODE_WAIT,
    TEST_MODE_EVENTFD
} ucp_test_mode = TEST_MODE_PROBE;

static struct err_handling {
    ucp_err_handling_mode_t ucp_err_mode;
    int                     failure;
} err_handling_opt;

static ucs_status_t client_status = UCS_OK;
static uint16_t server_port = 13337;
static long test_string_length = 16;
static const ucp_tag_t tag  = 0x1337a880u;
static const ucp_tag_t tag_mask = -1;
static ucp_address_t *local_addr;
static ucp_address_t *peer_addr;

static size_t local_addr_len;
static size_t peer_addr_len;

static int parse_cmd(int argc, char * const argv[], char **server_name);

static void request_init(void *request)
{
    struct ucx_context *ctx = (struct ucx_context *) request;
    ctx->completed = 0;
}

static void send_handler(void *request, ucs_status_t status)
{
    struct ucx_context *context = (struct ucx_context *) request;

    context->completed = 1;

    printf("[0x%x] send handler called with status %d (%s)\n",
           (unsigned int)pthread_self(), status, ucs_status_string(status));
}

static void failure_handler(void *arg, ucp_ep_h ep, ucs_status_t status)
{
    ucs_status_t *arg_status = (ucs_status_t *)arg;

    printf("[0x%x] failure handler called with status %d (%s)\n",
           (unsigned int)pthread_self(), status, ucs_status_string(status));

    *arg_status = status;
}

static void recv_handler(void *request, ucs_status_t status,
                        ucp_tag_recv_info_t *info)
{
    struct ucx_context *context = (struct ucx_context *) request;

    context->completed = 1;

    printf("[0x%x] receive handler called with status %d (%s), length %lu\n",
           (unsigned int)pthread_self(), status, ucs_status_string(status),
           info->length);
}

static void wait(ucp_worker_h ucp_worker, struct ucx_context *context)
{
    while (context->completed == 0) {
        ucp_worker_progress(ucp_worker);
    }
}

static ucs_status_t test_poll_wait(ucp_worker_h ucp_worker)
{
    int ret = -1, err = 0;
    ucs_status_t status;
    int epoll_fd_local = 0, epoll_fd = 0;
    struct epoll_event ev;
    ev.data.u64 = 0;

    status = ucp_worker_get_efd(ucp_worker, &epoll_fd);
    CHKERR_JUMP(UCS_OK != status, "ucp_worker_get_efd", err);

    /* It is recommended to copy original fd */
    epoll_fd_local = epoll_create(1);

    ev.data.fd = epoll_fd;
    ev.events = EPOLLIN;
    err = epoll_ctl(epoll_fd_local, EPOLL_CTL_ADD, epoll_fd, &ev);
```

```
    CHKERR_JUMP(err < 0, "add original socket to the new epoll\n", err_fd);

    /* Need to prepare ucp_worker before epoll_wait */
    status = ucp_worker_arm(ucp_worker);
    if (status == UCS_ERR_BUSY) { /* some events are arrived already */
        ret = UCS_OK;
        goto err_fd;
    }
    CHKERR_JUMP(status != UCS_OK, "ucp_worker_arm\n", err_fd);

    do {
        ret = epoll_wait(epoll_fd_local, &ev, 1, -1);
    } while ((ret == -1) && (errno == EINTR));

    ret = UCS_OK;

err_fd:
    close(epoll_fd_local);

err:
    return ret;
}

static int run_ucx_client(ucp_worker_h ucp_worker)
{
    ucp_tag_recv_info_t info_tag;
    ucp_tag_message_h msg_tag;
    ucs_status_t status;
    ucp_ep_h server_ep;
    ucp_ep_params_t ep_params;
    struct msg *msg = 0;
    struct ucx_context *request = 0;
    size_t msg_len = 0;
    int ret = -1;

    /* Send client UCX address to server */
    ep_params.field_mask      = UCP_EP_PARAM_FIELD_REMOTE_ADDRESS
       |
                                UCP_EP_PARAM_FIELD_ERR_HANDLING_MODE;
    ep_params.address         = peer_addr;
    ep_params.err_mode        = err_handling_opt.ucp_err_mode;

    status = ucp_ep_create(ucp_worker, &ep_params, &server_ep);
    CHKERR_JUMP(status != UCS_OK, "ucp_ep_create\n", err);

    msg_len = sizeof(*msg) + local_addr_len;
    msg = calloc(1, msg_len);
    CHKERR_JUMP(!msg, "allocate memory\n", err_ep);

    msg->data_len = local_addr_len;
    memcpy(msg + 1, local_addr, local_addr_len);

    request = ucp_tag_send_nb(server_ep, msg, msg_len,
                              ucp_dt_make_contig(1), tag,
                              send_handler);
    if (UCS_PTR_IS_ERR(request)) {
        fprintf(stderr, "unable to send UCX address message\n");
        free(msg);
        goto err_ep;
    } else if (UCS_PTR_STATUS(request) != UCS_OK) {
        wait(ucp_worker, request);
        request->completed = 0; /* Reset request state before recycling it */
        ucp_request_release(request);
    }

    free (msg);

    if (err_handling_opt.failure) {
        fprintf(stderr, "Emulating unexpected failure on client side\n");
        raise(SIGKILL);
    }

    /* Receive test string from server */
    for (;;) {

        /* Probing incoming events in non-block mode */
        msg_tag = ucp_tag_probe_nb(ucp_worker, tag, tag_mask, 1, &info_tag);
        if (msg_tag != NULL) {
            /* Message arrived */
            break;
        } else if (ucp_worker_progress(ucp_worker)) {
            /* Some events were polled; try again without going to sleep */
            continue;
        }

        /* If we got here, ucp_worker_progress() returned 0, so we can sleep.
         * Following blocked methods used to polling internal file descriptor
```

```
         * to make CPU idle and don't spin loop
         */
        if (ucp_test_mode == TEST_MODE_WAIT) {
            /* Polling incoming events*/
            status = ucp_worker_wait(ucp_worker);
            CHKERR_JUMP(status != UCS_OK, "ucp_worker_wait\n", err_ep);
        } else if (ucp_test_mode == TEST_MODE_EVENTFD) {
            status = test_poll_wait(ucp_worker);
            CHKERR_JUMP(status != UCS_OK, "test_poll_wait\n", err_ep);
        }
    }

    msg = malloc(info_tag.length);
    CHKERR_JUMP(!msg, "allocate memory\n", err_ep);

    request = ucp_tag_msg_recv_nb(ucp_worker, msg, info_tag.
      length,
                                  ucp_dt_make_contig(1), msg_tag,
                                  recv_handler);

    if (UCS_PTR_IS_ERR(request)) {
        fprintf(stderr, "unable to receive UCX data message (%u)\n",
                UCS_PTR_STATUS(request));
        free(msg);
        goto err_ep;
    } else {
        wait(ucp_worker, request);
        request->completed = 0;
        ucp_request_release(request);
        printf("UCX data message was received\n");
    }

    printf("\n\n----- UCP TEST SUCCESS ----\n\n");
    printf("%s", (char *)(msg + 1));
    printf("\n\n---------------------------\n\n");

    free(msg);

    ret = 0;

err_ep:
    ucp_ep_destroy(server_ep);

err:
    return ret;
}

static void flush_callback(void *request, ucs_status_t status)
{
}

static ucs_status_t flush_ep(ucp_worker_h worker,
      ucp_ep_h ep)
{
    void *request;

    request = ucp_ep_flush_nb(ep, 0, flush_callback);
    if (request == NULL) {
        return UCS_OK;
    } else if (UCS_PTR_IS_ERR(request)) {
        return UCS_PTR_STATUS(request);
    } else {
        ucs_status_t status;
        do {
            ucp_worker_progress(worker);
            status = ucp_request_check_status(request);
        } while (status == UCS_INPROGRESS);
        ucp_request_release(request);
        return status;
    }
}

static int run_ucx_server(ucp_worker_h ucp_worker)
{
    ucp_tag_recv_info_t info_tag;
    ucp_tag_message_h msg_tag;
    ucs_status_t status;
    ucp_ep_h client_ep;
    ucp_ep_params_t ep_params;
    struct msg *msg = 0;
    struct ucx_context *request = 0;
    size_t msg_len = 0;
    int ret = -1;

    /* Receive client UCX address */
    do {
        /* Progressing before probe to update the state */
```

```
        ucp_worker_progress(ucp_worker);

        /* Probing incoming events in non-block mode */
        msg_tag = ucp_tag_probe_nb(ucp_worker, tag, tag_mask, 1, &info_tag);
    } while (msg_tag == NULL);

    msg = malloc(info_tag.length);
    CHKERR_JUMP(!msg, "allocate memory\n", err);
    request = ucp_tag_msg_recv_nb(ucp_worker, msg, info_tag.
      length,
                                  ucp_dt_make_contig(1), msg_tag, recv_handler);

    if (UCS_PTR_IS_ERR(request)) {
        fprintf(stderr, "unable to receive UCX address message (%s)\n",
                ucs_status_string(UCS_PTR_STATUS(request)));
        free(msg);
        goto err;
    } else {
        wait(ucp_worker, request);
        request->completed = 0;
        ucp_request_release(request);
        printf("UCX address message was received\n");
    }

    peer_addr = malloc(msg->data_len);
    if (!peer_addr) {
        fprintf(stderr, "unable to allocate memory for peer address\n");
        free(msg);
        goto err;
    }

    peer_addr_len = msg->data_len;
    memcpy(peer_addr, msg + 1, peer_addr_len);

    free(msg);

    /* Send test string to client */
    ep_params.field_mask      = UCP_EP_PARAM_FIELD_REMOTE_ADDRESS
       |
                                UCP_EP_PARAM_FIELD_ERR_HANDLING_MODE |
                                UCP_EP_PARAM_FIELD_ERR_HANDLER |
                                UCP_EP_PARAM_FIELD_USER_DATA;
    ep_params.address         = peer_addr;
    ep_params.err_mode        = err_handling_opt.ucp_err_mode;
    ep_params.err_handler.cb  = failure_handler;
    ep_params.err_handler.arg = NULL;
    ep_params.user_data       = &client_status;

    status = ucp_ep_create(ucp_worker, &ep_params, &client_ep);
    CHKERR_JUMP(status != UCS_OK, "ucp_ep_create\n", err);

    msg_len = sizeof(*msg) + test_string_length;
    msg = calloc(1, msg_len);
    CHKERR_JUMP(!msg, "allocate memory\n", err_ep);

    msg->data_len = msg_len - sizeof(*msg);
    generate_random_string((char *)(msg + 1), test_string_length);

    request = ucp_tag_send_nb(client_ep, msg, msg_len,
                              ucp_dt_make_contig(1), tag,
                              send_handler);
    if (UCS_PTR_IS_ERR(request)) {
        fprintf(stderr, "unable to send UCX data message\n");
        free(msg);
        goto err_ep;
    } else if (UCS_PTR_STATUS(request) != UCS_OK) {
        printf("UCX data message was scheduled for send\n");
        wait(ucp_worker, request);
        request->completed = 0;
        ucp_request_release(request);
    }

    status = flush_ep(ucp_worker, client_ep);
    printf("flush_ep completed with status %d (%s)\n",
            status, ucs_status_string(status));

    ret = 0;
    free(msg);

err_ep:
    ucp_ep_destroy(client_ep);

err:
    return ret;
}

static int run_test(const char *client_target_name, ucp_worker_h ucp_worker)
```

```
{
    if (client_target_name != NULL) {
        return run_ucx_client(ucp_worker);
    } else {
        return run_ucx_server(ucp_worker);
    }
}

int main(int argc, char **argv)
{
    /* UCP temporary vars */
    ucp_params_t ucp_params;
    ucp_worker_params_t worker_params;
    ucp_config_t *config;
    ucs_status_t status;

    /* UCP handler objects */
    ucp_context_h ucp_context;
    ucp_worker_h ucp_worker;

    /* OOB connection vars */
    uint64_t addr_len = 0;
    char *client_target_name = NULL;
    int oob_sock = -1;
    int ret = -1;

    memset(&ucp_params, 0, sizeof(ucp_params));
    memset(&worker_params, 0, sizeof(worker_params));

    /* Parse the command line */
    status = parse_cmd(argc, argv, &client_target_name);
    CHKERR_JUMP(status != UCS_OK, "parse_cmd\n", err);

    /* UCP initialization */
    status = ucp_config_read(NULL, NULL, &config);
    CHKERR_JUMP(status != UCS_OK, "ucp_config_read\n", err);

    ucp_params.field_mask   = UCP_PARAM_FIELD_FEATURES |
                              UCP_PARAM_FIELD_REQUEST_SIZE |
                              UCP_PARAM_FIELD_REQUEST_INIT;
    ucp_params.features     = UCP_FEATURE_TAG;
    if (ucp_test_mode == TEST_MODE_WAIT || ucp_test_mode == TEST_MODE_EVENTFD) {
        ucp_params.features |= UCP_FEATURE_WAKEUP;
    }
    ucp_params.request_size    = sizeof(struct ucx_context);
    ucp_params.request_init    = request_init;

    status = ucp_init(&ucp_params, config, &ucp_context);

    ucp_config_print(config, stdout, NULL, UCS_CONFIG_PRINT_CONFIG);

    ucp_config_release(config);
    CHKERR_JUMP(status != UCS_OK, "ucp_init\n", err);

    worker_params.field_mask  = UCP_WORKER_PARAM_FIELD_THREAD_MODE
      ;
    worker_params.thread_mode = UCS_THREAD_MODE_SINGLE;

    status = ucp_worker_create(ucp_context, &worker_params, &ucp_worker);
    CHKERR_JUMP(status != UCS_OK, "ucp_worker_create\n", err_cleanup);

    status = ucp_worker_get_address(ucp_worker, &local_addr, &local_addr_len);
    CHKERR_JUMP(status != UCS_OK, "ucp_worker_get_address\n", err_worker);

    printf("[0x%x] local address length: %lu\n",
           (unsigned int)pthread_self(), local_addr_len);

    /* OOB connection establishment */
    if (client_target_name) {
        peer_addr_len = local_addr_len;

        oob_sock = client_connect(client_target_name, server_port);
        CHKERR_JUMP(oob_sock < 0, "client_connect\n", err_addr);

        ret = recv(oob_sock, &addr_len, sizeof(addr_len), 0);
        CHKERR_JUMP(ret < 0, "receive address length\n", err_addr);

        peer_addr_len = addr_len;
        peer_addr = malloc(peer_addr_len);
        CHKERR_JUMP(!peer_addr, "allocate memory\n", err_addr);

        ret = recv(oob_sock, peer_addr, peer_addr_len, 0);
        CHKERR_JUMP(ret < 0, "receive address\n", err_peer_addr);
    } else {
        oob_sock = server_connect(server_port);
        CHKERR_JUMP(oob_sock < 0, "server_connect\n", err_peer_addr);
```

```
        addr_len = local_addr_len;
        ret = send(oob_sock, &addr_len, sizeof(addr_len), 0);
        CHKERR_JUMP((ret < 0 || ret != sizeof(addr_len)),
                    "send address length\n", err_peer_addr);

        ret = send(oob_sock, local_addr, local_addr_len, 0);
        CHKERR_JUMP((ret < 0 || ret != local_addr_len),
                    "send address\n", err_peer_addr);
    }

    ret = run_test(client_target_name, ucp_worker);

    if (!err_handling_opt.failure) {
        /* Make sure remote is disconnected before destroying local worker */
        barrier(oob_sock);
    }
    close(oob_sock);

err_peer_addr:
    free(peer_addr);

err_addr:
    ucp_worker_release_address(ucp_worker, local_addr);

err_worker:
    ucp_worker_destroy(ucp_worker);

err_cleanup:
    ucp_cleanup(ucp_context);

err:
    return ret;
}

int parse_cmd(int argc, char * const argv[], char **server_name)
{
    int c = 0, index = 0;
    opterr = 0;

    err_handling_opt.ucp_err_mode   = UCP_ERR_HANDLING_MODE_NONE;
    err_handling_opt.failure        = 0;

    while ((c = getopt(argc, argv, "wfben:p:s:h")) != -1) {
        switch (c) {
        case 'w':
            ucp_test_mode = TEST_MODE_WAIT;
            break;
        case 'f':
            ucp_test_mode = TEST_MODE_EVENTFD;
            break;
        case 'b':
            ucp_test_mode = TEST_MODE_PROBE;
            break;
        case 'e':
            err_handling_opt.ucp_err_mode   = UCP_ERR_HANDLING_MODE_PEER;
            err_handling_opt.failure        = 1;
            break;
        case 'n':
            *server_name = optarg;
            break;
        case 'p':
            server_port = atoi(optarg);
            if (server_port <= 0) {
                fprintf(stderr, "Wrong server port number %d\n", server_port);
                return UCS_ERR_UNSUPPORTED;
            }
            break;
        case 's':
            test_string_length = atol(optarg);
            if (test_string_length <= 0) {
                fprintf(stderr, "Wrong string size %ld\n", test_string_length);
                return UCS_ERR_UNSUPPORTED;
            }
            break;
        case '?':
            if (optopt == 's') {
                fprintf(stderr, "Option -%c requires an argument.\n", optopt);
            } else if (isprint (optopt)) {
                fprintf(stderr, "Unknown option `-%c'.\n", optopt);
            } else {
                fprintf(stderr, "Unknown option character `\\x%x'.\n", optopt);
            }
        case 'h':
        default:
            fprintf(stderr, "Usage: ucp_hello_world [parameters]\n");
            fprintf(stderr, "UCP hello world client/server example utility\n");
            fprintf(stderr, "\nParameters are:\n");
```

```
            fprintf(stderr, "  -w      Select test mode \"wait\" to test "
                    "ucp_worker_wait function\n");
            fprintf(stderr, "  -f      Select test mode \"event fd\" to test "
                    "ucp_worker_get_efd function with later poll\n");
            fprintf(stderr, "  -b      Select test mode \"busy polling\" to test "
                    "ucp_tag_probe_nb and ucp_worker_progress (default)\n");
            fprintf(stderr, "  -e      Emulate unexpected failure on server side"
                    "and handle an error on client side with enabled "
                    "UCP_ERR_HANDLING_MODE_PEER\n");
            fprintf(stderr, "  -n name Set node name or IP address "
                    "of the server (required for client and should be ignored "
                    "for server)\n");
            fprintf(stderr, "  -p port Set alternative server port (default:13337)\n");
            fprintf(stderr, "  -s size Set test string length (default:16)\n");
            fprintf(stderr, "\n");
            return UCS_ERR_UNSUPPORTED;
        }
    }
    fprintf(stderr, "INFO: UCP_HELLO_WORLD mode = %d server = %s port = %d\n",
            ucp_test_mode, *server_name, server_port);

    for (index = optind; index < argc; index++) {
        fprintf(stderr, "WARNING: Non-option argument %s\n", argv[index]);
    }
    return UCS_OK;
}
```

8.2 uct_hello_world.c

UCT hello world client / server example utility.

```
#include "ucx_hello_world.h"

#include <uct/api/uct.h>

#include <assert.h>
#include <ctype.h>

typedef enum {
    FUNC_AM_SHORT,
    FUNC_AM_BCOPY,
    FUNC_AM_ZCOPY
} func_am_t;

typedef struct {
    int  is_uct_desc;
} recv_desc_t;

typedef struct {
    char                *server_name;
    uint16_t             server_port;
    func_am_t            func_am_type;
    const char          *dev_name;
    const char          *tl_name;
    long                 test_strlen;
} cmd_args_t;

typedef struct {
    uct_iface_attr_t    attr;   /* Interface attributes: capabilities and limitations */
    uct_iface_h         iface;  /* Communication interface context */
    uct_md_h            pd;     /* Memory domain */
    uct_worker_h        worker; /* Workers represent allocated resources in a communication
       thread */
} iface_info_t;

/* Helper data type for am_short */
typedef struct {
    uint64_t            header;
    char               *payload;
    size_t              len;
} am_short_args_t;

/* Helper data type for am_bcopy */
typedef struct {
    char               *data;
    size_t              len;
} am_bcopy_args_t;

/* Helper data type for am_zcopy */
typedef struct {
    uct_completion_t    uct_comp;
```

```
    uct_md_h            md;
    uct_mem_h           memh;
} zcopy_comp_t;

static void* desc_holder = NULL;

static char *func_am_t_str(func_am_t func_am_type)
{
    switch (func_am_type) {
    case FUNC_AM_SHORT:
        return "uct_ep_am_short";
    case FUNC_AM_BCOPY:
        return "uct_ep_am_bcopy";
    case FUNC_AM_ZCOPY:
        return "uct_ep_am_zcopy";
    }
    return NULL;
}

static size_t func_am_max_size(func_am_t func_am_type,
                               const uct_iface_attr_t *attr)
{
    switch (func_am_type) {
    case FUNC_AM_SHORT:
        return attr->cap.am.max_short;
    case FUNC_AM_BCOPY:
        return attr->cap.am.max_bcopy;
    case FUNC_AM_ZCOPY:
        return attr->cap.am.max_zcopy;
    }
    return 0;
}

/* Helper function for am_short */
void am_short_params_pack(char *buf, size_t len, am_short_args_t *args)
{
    args->header      = *(uint64_t *)buf;
    if (len > sizeof(args->header)) {
        args->payload = (buf + sizeof(args->header));
        args->len     = len - sizeof(args->header);
    } else {
        args->payload = NULL;
        args->len     = 0;
    }
}

ucs_status_t do_am_short(iface_info_t *if_info, uct_ep_h ep, uint8_t id,
                         const cmd_args_t *cmd_args, char *buf)
{
    ucs_status_t    status;
    am_short_args_t send_args;

    am_short_params_pack(buf, cmd_args->test_strlen, &send_args);

    do {
        /* Send active message to remote endpoint */
        status = uct_ep_am_short(ep, id, send_args.header, send_args.payload,
                                 send_args.len);
        uct_worker_progress(if_info->worker);
    } while (status == UCS_ERR_NO_RESOURCE);

    return status;
}

/* Pack callback for am_bcopy */
size_t am_bcopy_data_pack_cb(void *dest, void *arg)
{
    am_bcopy_args_t *bc_args = arg;
    memcpy(dest, bc_args->data, bc_args->len);
    return bc_args->len;
}

ucs_status_t do_am_bcopy(iface_info_t *if_info, uct_ep_h ep, uint8_t id,
                         const cmd_args_t *cmd_args, char *buf)
{
    am_bcopy_args_t args;
    ssize_t len;

    args.data = buf;
    args.len  = cmd_args->test_strlen;

    /* Send active message to remote endpoint */
    do {
        len = uct_ep_am_bcopy(ep, id, am_bcopy_data_pack_cb, &args, 0);
        uct_worker_progress(if_info->worker);
    } while (len == UCS_ERR_NO_RESOURCE);
    /* Negative len is an error code */
```

```
    return (len >= 0) ? UCS_OK : len;
}

/* Completion callback for am_zcopy */
void zcopy_completion_cb(uct_completion_t *self, ucs_status_t status)
{
    zcopy_comp_t *comp = (zcopy_comp_t *)self;
    assert((comp->uct_comp.count == 0) && (status == UCS_OK));
    uct_md_mem_dereg(comp->md, comp->memh);
    desc_holder = (void *)0xDEADBEEF;
}

ucs_status_t do_am_zcopy(iface_info_t *if_info, uct_ep_h ep, uint8_t id,
                         const cmd_args_t *cmd_args, char *buf)
{
    uct_mem_h memh;
    uct_iov_t iov;
    zcopy_comp_t comp;

    ucs_status_t status = uct_md_mem_reg(if_info->pd, buf, cmd_args->test_strlen,
                                         UCT_MD_MEM_ACCESS_RMA, &memh);
    iov.buffer          = buf;
    iov.length          = cmd_args->test_strlen;
    iov.memh            = memh;
    iov.stride          = 0;
    iov.count           = 1;

    comp.uct_comp.func  = zcopy_completion_cb;
    comp.uct_comp.count = 1;
    comp.md             = if_info->pd;
    comp.memh           = memh;

    if (status == UCS_OK) {
        do {
            status = uct_ep_am_zcopy(ep, id, NULL, 0, &iov, 1, 0,
                                     (uct_completion_t *)&comp);
            uct_worker_progress(if_info->worker);
        } while (status == UCS_ERR_NO_RESOURCE);

        if (status == UCS_INPROGRESS) {
            while (!desc_holder) {
                /* Explicitly progress outstanding active message request */
                uct_worker_progress(if_info->worker);
            }
            status = UCS_OK;
        }
    }
    return status;
}
static void print_strings(const char *label, const char *local_str,
                          const char *remote_str)
{
    fprintf(stdout, "\n\n----- UCT TEST SUCCESS ----\n\n");
    fprintf(stdout, "[%s] %s sent %s", label, local_str, remote_str);
    fprintf(stdout, "\n\n---------------------------\n");
    fflush(stdout);
}

/* Callback to handle receive active message */
static ucs_status_t hello_world(void *arg, void *data, size_t length, unsigned flags)
{
    recv_desc_t *rdesc;
    func_am_t func_am_type = *(func_am_t *)arg;
    print_strings("callback", func_am_t_str(func_am_type), data);

    if (flags & UCT_CB_PARAM_FLAG_DESC) {
        rdesc = (recv_desc_t *)data - 1;
        /* Hold descriptor to release later and return UCS_INPROGRESS */
        rdesc->is_uct_desc = 1;
        desc_holder = rdesc;
        return UCS_INPROGRESS;
    }

    /* We need to copy-out data and return UCS_OK if want to use the data
     * outside the callback */
    rdesc = malloc(sizeof(*rdesc) + length);
    rdesc->is_uct_desc = 0;
    memcpy(rdesc + 1, data, length);
    desc_holder = rdesc;
    return UCS_OK;
}

/* init the transport  by its name */
static ucs_status_t init_iface(char *dev_name, char *tl_name,
                               func_am_t func_am_type,
                               iface_info_t *iface_p)
{
```

```
    ucs_status_t         status;
    uct_iface_config_t  *config; /* Defines interface configuration options */
    uct_iface_params_t  params;

    params.open_mode            = UCT_IFACE_OPEN_MODE_DEVICE;
    params.mode.device.tl_name  = tl_name;
    params.mode.device.dev_name = dev_name;
    params.stats_root           = NULL;
    params.rx_headroom          = sizeof(recv_desc_t);

    UCS_CPU_ZERO(&params.cpu_mask);
    /* Read transport-specific interface configuration */
    status = uct_md_iface_config_read(iface_p->pd, tl_name, NULL, NULL, &config);
    CHKERR_JUMP(UCS_OK != status, "setup iface_config", error_ret);

    /* Open communication interface */
    status = uct_iface_open(iface_p->pd, iface_p->worker, &params, config,
                            &iface_p->iface);
    uct_config_release(config);
    CHKERR_JUMP(UCS_OK != status, "open temporary interface", error_ret);

    /* Enable progress on the interface */
    uct_iface_progress_enable(iface_p->iface,
                              UCT_PROGRESS_SEND |
      UCT_PROGRESS_RECV);

    /* Get interface attributes */
    status = uct_iface_query(iface_p->iface, &iface_p->attr);
    CHKERR_JUMP(UCS_OK != status, "query iface", error_iface);

    /* Check if current device and transport support required active messages */
    if ((func_am_type == FUNC_AM_SHORT) &&
        (iface_p->attr.cap.flags & UCT_IFACE_FLAG_AM_SHORT)) {
        return UCS_OK;
    }

    if ((func_am_type == FUNC_AM_BCOPY) &&
        (iface_p->attr.cap.flags & UCT_IFACE_FLAG_AM_BCOPY)) {
        return UCS_OK;
    }

    if ((func_am_type == FUNC_AM_ZCOPY) &&
        (iface_p->attr.cap.flags & UCT_IFACE_FLAG_AM_ZCOPY)) {
        return UCS_OK;
    }

error_iface:
    uct_iface_close(iface_p->iface);
error_ret:
    return UCS_ERR_UNSUPPORTED;
}

/* Device and transport to be used are determined by minimum latency */
static ucs_status_t dev_tl_lookup(const cmd_args_t *cmd_args,
                                  iface_info_t *iface_p)
{
    uct_md_resource_desc_t  *md_resources; /* Memory domain resource descriptor */
    uct_tl_resource_desc_t  *tl_resources; /*Communication resource descriptor */
    unsigned                num_md_resources; /* Number of protected domain */
    unsigned                num_tl_resources; /* Number of transport resources resource objects created */
    uct_md_config_t         *md_config;
    ucs_status_t            status;
    int                     i;
    int                     j;

    status = uct_query_md_resources(&md_resources, &num_md_resources);
    CHKERR_JUMP(UCS_OK != status, "query for memory domain resources", error_ret);

    /* Iterate through protected domain resources */
    for (i = 0; i < num_md_resources; ++i) {
        status = uct_md_config_read(md_resources[i].md_name, NULL, NULL, &md_config);
        CHKERR_JUMP(UCS_OK != status, "read PD config", release_pd);

        status = uct_md_open(md_resources[i].md_name, md_config, &iface_p->pd);
        uct_config_release(md_config);
        CHKERR_JUMP(UCS_OK != status, "open memory domains", release_pd);

        status = uct_md_query_tl_resources(iface_p->pd, &tl_resources, &
      num_tl_resources);
        CHKERR_JUMP(UCS_OK != status, "query transport resources", close_pd);

        /* Go through each available transport and find the proper name */
        for (j = 0; j < num_tl_resources; ++j) {
            if (!strcmp(cmd_args->dev_name, tl_resources[j].dev_name) &&
                !strcmp(cmd_args->tl_name, tl_resources[j].tl_name)) {
                status = init_iface(tl_resources[j].dev_name,
                                    tl_resources[j].tl_name,
```

```
                                    cmd_args->func_am_type, iface_p);
                if (UCS_OK == status) {
                    fprintf(stdout, "Using %s with %s.\n",
                            tl_resources[j].dev_name,
                            tl_resources[j].tl_name);
                    fflush(stdout);
                    uct_release_tl_resource_list(tl_resources);
                    goto release_pd;
                }
            }
        }
        uct_release_tl_resource_list(tl_resources);
        uct_md_close(iface_p->pd);
    }

    fprintf(stderr, "No supported (dev/tl) found (%s/%s)\n",
            cmd_args->dev_name, cmd_args->tl_name);
    status = UCS_ERR_UNSUPPORTED;

release_pd:
    uct_release_md_resource_list(md_resources);
error_ret:
    return status;
close_pd:
    uct_md_close(iface_p->pd);
    goto release_pd;
}

int print_err_usage()
{
    const char func_template[] = "  -%c      Select \"%s\" function to send the message%s\n";

    fprintf(stderr, "Usage: uct_hello_world [parameters]\n");
    fprintf(stderr, "UCT hello world client/server example utility\n");
    fprintf(stderr, "\nParameters are:\n");
    fprintf(stderr, func_template, 'i', func_am_t_str(FUNC_AM_SHORT), " (default)");
    fprintf(stderr, func_template, 'b', func_am_t_str(FUNC_AM_BCOPY), "");
    fprintf(stderr, func_template, 'z', func_am_t_str(FUNC_AM_ZCOPY), "");
    fprintf(stderr, "  -d      Select device name\n");
    fprintf(stderr, "  -t      Select transport layer\n");
    fprintf(stderr, "  -n name Set node name or IP address "
            "of the server (required for client and should be ignored "
            "for server)\n");
    fprintf(stderr, "  -p port Set alternative server port (default:13337)\n");
    fprintf(stderr, "  -s size Set test string length (default:16)\n");
    fprintf(stderr, "\n");
    return UCS_ERR_UNSUPPORTED;
}

int parse_cmd(int argc, char * const argv[], cmd_args_t *args)
{
    int c = 0, index = 0;

    assert(args);
    memset(args, 0, sizeof(*args));

    /* Defaults */
    args->server_port   = 13337;
    args->func_am_type  = FUNC_AM_SHORT;
    args->test_strlen   = 16;

    opterr = 0;
    while ((c = getopt(argc, argv, "ibzd:t:n:p:s:h")) != -1) {
        switch (c) {
        case 'i':
            args->func_am_type = FUNC_AM_SHORT;
            break;
        case 'b':
            args->func_am_type = FUNC_AM_BCOPY;
            break;
        case 'z':
            args->func_am_type = FUNC_AM_ZCOPY;
            break;
        case 'd':
            args->dev_name = optarg;
            break;
        case 't':
            args->tl_name = optarg;
            break;
        case 'n':
            args->server_name = optarg;
            break;
        case 'p':
            args->server_port = atoi(optarg);
            if (args->server_port <= 0) {
                fprintf(stderr, "Wrong server port number %d\n",
                        args->server_port);
```

```
                return UCS_ERR_UNSUPPORTED;
            }
            break;
        case 's':
            args->test_strlen = atol(optarg);
            if (args->test_strlen <= 0) {
                fprintf(stderr, "Wrong string size %ld\n", args->test_strlen);
                return UCS_ERR_UNSUPPORTED;
            }
            break;
        case '?':
            if (optopt == 's') {
                fprintf(stderr, "Option -%c requires an argument.\n", optopt);
            } else if (isprint (optopt)) {
                fprintf(stderr, "Unknown option `-%c'.\n", optopt);
            } else {
                fprintf(stderr, "Unknown option character `\\x%x'.\n", optopt);
            }
        case 'h':
        default:
            return print_err_usage();
        }
    }
    fprintf(stderr, "INFO: UCT_HELLO_WORLD AM function = %s server = %s port = %d\n",
            func_am_t_str(args->func_am_type), args->server_name,
            args->server_port);

    for (index = optind; index < argc; index++) {
        fprintf(stderr, "WARNING: Non-option argument %s\n", argv[index]);
    }

    if (args->dev_name == NULL) {
        fprintf(stderr, "WARNING: device is not set\n");
        return print_err_usage();
    }

    if (args->tl_name == NULL) {
        fprintf(stderr, "WARNING: transport layer is not set\n");
        return print_err_usage();
    }

    return UCS_OK;
}

/* The caller is responsible to free *rbuf */
int sendrecv(int sock, const void *sbuf, size_t slen, void **rbuf)
{
    int ret = 0;
    size_t rlen = 0;
    *rbuf = NULL;

    ret = send(sock, &slen, sizeof(slen), 0);
    if ((ret < 0) || (ret != sizeof(slen))) {
        fprintf(stderr, "failed to send buffer length\n");
        return -1;
    }

    ret = send(sock, sbuf, slen, 0);
    if ((ret < 0) || (ret != slen)) {
        fprintf(stderr, "failed to send buffer\n");
        return -1;
    }

    ret = recv(sock, &rlen, sizeof(rlen), 0);
    if (ret < 0) {
        fprintf(stderr, "failed to receive device address length\n");
        return -1;
    }

    *rbuf = calloc(1, rlen);
    if (!*rbuf) {
        fprintf(stderr, "failed to allocate receive buffer\n");
        return -1;
    }

    ret = recv(sock, *rbuf, rlen, 0);
    if (ret < 0) {
        fprintf(stderr, "failed to receive device address\n");
        return -1;
    }

    return 0;
}

int main(int argc, char **argv)
{
    uct_device_addr_t    *own_dev;
```

```
uct_device_addr_t   *peer_dev   = NULL;
uct_iface_addr_t    *own_iface;
uct_iface_addr_t    *peer_iface = NULL;
uct_ep_addr_t       *own_ep     = NULL;
uct_ep_addr_t       *peer_ep    = NULL;
ucs_status_t        status      = UCS_OK; /* status codes for UCS */
uct_ep_h            ep;                   /* Remote endpoint */
ucs_async_context_t *async;               /* Async event context manages
                                             times and fd notifications */
cmd_args_t          cmd_args;

iface_info_t        if_info;
uint8_t             id = 0;
int                 oob_sock = -1;  /* OOB connection socket */

/* Parse the command line */
if (parse_cmd(argc, argv, &cmd_args)) {
    status = UCS_ERR_INVALID_PARAM;
    goto out;
}

/* Initialize context
 * It is better to use different contexts for different workers
 */
status = ucs_async_context_create(UCS_ASYNC_MODE_THREAD, &async);
CHKERR_JUMP(UCS_OK != status, "init async context", out);

/* Create a worker object */
status = uct_worker_create(async, UCS_THREAD_MODE_SINGLE, &
  if_info.worker);
CHKERR_JUMP(UCS_OK != status, "create worker", out_cleanup_async);

/* Search for the desired transport */
status = dev_tl_lookup(&cmd_args, &if_info);
CHKERR_JUMP(UCS_OK != status, "find supported device and transport",
            out_destroy_worker);

own_dev = (uct_device_addr_t*)calloc(1, if_info.attr.device_addr_len);
CHKERR_JUMP(NULL == own_dev, "allocate memory for dev addr",
            out_destroy_iface);

own_iface = (uct_iface_addr_t*)calloc(1, if_info.attr.iface_addr_len);
CHKERR_JUMP(NULL == own_iface, "allocate memory for if addr",
            out_free_dev_addrs);

/* Get device address */
status = uct_iface_get_device_address(if_info.iface, own_dev);
CHKERR_JUMP(UCS_OK != status, "get device address", out_free_if_addrs);

if (cmd_args.server_name) {
    oob_sock = client_connect(cmd_args.server_name, cmd_args.server_port);
    if (oob_sock < 0) {
        goto out_free_if_addrs;
    }
} else {
    oob_sock = server_connect(cmd_args.server_port);
    if (oob_sock < 0) {
        goto out_free_if_addrs;
    }
}

status = sendrecv(oob_sock, own_dev, if_info.attr.device_addr_len,
                  (void **)&peer_dev);
CHKERR_JUMP(0 != status, "device exchange", out_free_dev_addrs);

status = uct_iface_is_reachable(if_info.iface, peer_dev, NULL);
CHKERR_JUMP(0 == status, "reach the peer", out_free_if_addrs);

/* Get interface address */
if (if_info.attr.cap.flags & UCT_IFACE_FLAG_CONNECT_TO_IFACE) {
    status = uct_iface_get_address(if_info.iface, own_iface);
    CHKERR_JUMP(UCS_OK != status, "get interface address", out_free_if_addrs);

    status = sendrecv(oob_sock, own_iface, if_info.attr.iface_addr_len,
                      (void **)&peer_iface);
    CHKERR_JUMP(0 != status, "ifaces exchange", out_free_if_addrs);
}

if (if_info.attr.cap.flags & UCT_IFACE_FLAG_CONNECT_TO_EP) {
    own_ep = (uct_ep_addr_t*)calloc(1, if_info.attr.ep_addr_len);
    CHKERR_JUMP(NULL == own_ep, "allocate memory for ep addrs", out_free_if_addrs);

    /* Create new endpoint */
    status = uct_ep_create(if_info.iface, &ep);
    CHKERR_JUMP(UCS_OK != status, "create endpoint", out_free_ep_addrs);

    /* Get endpoint address */
```

```
        status = uct_ep_get_address(ep, own_ep);
        CHKERR_JUMP(UCS_OK != status, "get endpoint address", out_free_ep);

        status = sendrecv(oob_sock, own_ep, if_info.attr.ep_addr_len,
                          (void **)&peer_ep);
        CHKERR_JUMP(0 != status, "EPs exchange", out_free_ep);

        /* Connect endpoint to a remote endpoint */
        status = uct_ep_connect_to_ep(ep, peer_dev, peer_ep);
        barrier(oob_sock);
    } else if (if_info.attr.cap.flags & UCT_IFACE_FLAG_CONNECT_TO_IFACE) {
        /* Create an endpoint which is connected to a remote interface */
        status = uct_ep_create_connected(if_info.iface, peer_dev, peer_iface, &ep);
    } else {
        status = UCS_ERR_UNSUPPORTED;
    }
    CHKERR_JUMP(UCS_OK != status, "connect endpoint", out_free_ep);

    if (cmd_args.test_strlen > func_am_max_size(cmd_args.func_am_type, &if_info.attr)) {
        status = UCS_ERR_UNSUPPORTED;
        fprintf(stderr, "Test string is too long: %ld, max supported: %lu\n",
                cmd_args.test_strlen,
                func_am_max_size(cmd_args.func_am_type, &if_info.attr));
        goto out_free_ep;
    }

    /*Set active message handler */
    status = uct_iface_set_am_handler(if_info.iface, id, hello_world,
                                      &cmd_args.func_am_type, 0);
    CHKERR_JUMP(UCS_OK != status, "set callback", out_free_ep);

    if (cmd_args.server_name) {
        char *str = (char *)malloc(cmd_args.test_strlen);
        generate_random_string(str, cmd_args.test_strlen);

        /* Send active message to remote endpoint */
        if (cmd_args.func_am_type == FUNC_AM_SHORT) {
            status = do_am_short(&if_info, ep, id, &cmd_args, str);
        } else if (cmd_args.func_am_type == FUNC_AM_BCOPY) {
            status = do_am_bcopy(&if_info, ep, id, &cmd_args, str);
        } else if (cmd_args.func_am_type == FUNC_AM_ZCOPY) {
            status = do_am_zcopy(&if_info, ep, id, &cmd_args, str);
        }

        free(str);
        CHKERR_JUMP(UCS_OK != status, "send active msg", out_free_ep);
    } else {
        recv_desc_t *rdesc;

        while (!desc_holder) {
            /* Explicitly progress any outstanding active message requests */
            uct_worker_progress(if_info.worker);
        }

        rdesc = desc_holder;
        print_strings("main", func_am_t_str(cmd_args.func_am_type),
                                       (char *)(rdesc + 1));
        if (rdesc->is_uct_desc) {
            /* Release descriptor because callback returns UCS_INPROGRESS */
            uct_iface_release_desc(rdesc);
        } else {
            free(rdesc);
        }
    }

    barrier(oob_sock);
    close(oob_sock);

out_free_ep:
    uct_ep_destroy(ep);
out_free_ep_addrs:
    free(own_ep);
    free(peer_ep);
out_free_if_addrs:
    free(own_iface);
    free(peer_iface);
out_free_dev_addrs:
    free(own_dev);
    free(peer_dev);
out_destroy_iface:
    uct_iface_close(if_info.iface);
    uct_md_close(if_info.pd);
out_destroy_worker:
    uct_worker_destroy(if_info.worker);
out_cleanup_async:
    ucs_async_context_destroy(async);
out:
```

```
    return status == UCS_ERR_UNSUPPORTED ? UCS_OK : status;
}
```

Index