RCCL Documentation

Release 2.21.5

Advanced Micro Devices, Inc.

Dec 20, 2024

CONTENTS

The ROCm Communication Collectives Library (RCCL) is a stand-alone library that provides multi-GPU and multi-node collective communication primitives optimized for AMD GPUs. It uses PCIe and xGMI high-speed interconnects. To learn more, see *What is RCCL?*

The RCCL public repository is located at https://github.com/ROCm/rccl.

Install

- *Installing RCCL using the install script*
- *Running RCCL using Docker*
- *Building and installing RCCL from source code*

How to

- *Using the NCCL Net plugin*
- *Troubleshoot RCCL*
- *RCCL usage tips*

Examples

- RCCL Tuner plugin examples
- NCCL Net plugin examples

API reference

- *Library specification*
- *API library*

To contribute to the documentation, see Contributing to ROCm.

You can find licensing information on the Licensing page.

CHAPTER

ONE

WHAT IS RCCL?

The ROCm Communication Collectives Library (RCCL) includes multi-GPU and multi-node collective communication primitives optimized for AMD GPUs. It implements routines such as `all-reduce`, `all-gather`, `reduce`, `broadcast`, `reduce-scatter`, `gather`, `scatter`, `all-to-allv`, and `all-to-all`, as well as direct point-to-point (GPU-to-GPU) send and receive operations. It is optimized to achieve high bandwidth on platforms using PCIe and xGMI and networking using InfiniBand Verbs or TCP/IP sockets. RCCL supports an arbitrary number of GPUs installed in a single node or multiple nodes and can be used in either single- or multi-process (for example, MPI) applications.

The collective operations are implemented using ring and tree algorithms and have been optimized for throughput and latency by leveraging topology awareness, high-speed interconnects, and RDMA-based collectives. For best performance, small operations can be either batched into larger operations or aggregated through the API.

RCCL uses PCIe and xGMI high-speed interconnects for intra-node communication as well as InfiniBand, RoCE, and TCP/IP for inter-node communication. It supports an arbitrary number of GPUs installed in a single-node or multi-node platform and can easily integrate into single- or multi-process (for example, MPI) applications.

CHAPTER

TWO

INSTALLING RCCL USING THE INSTALL SCRIPT

To quickly install RCCL using the install script, follow these steps. For instructions on building RCCL from the source code, see *Building and installing RCCL from source code*. For additional tips, see *RCCL usage tips*.

2.1 Requirements

The following prerequisites are required to use RCCL:

1. ROCm-supported GPUs
2. The ROCm stack must be installed on the system, including the HIP runtime and the HIP-Clang compiler.

2.2 Quick start RCCL build

RCCL directly depends on the HIP runtime plus the HIP-Clang compiler, which are part of the ROCm software stack. For ROCm installation instructions, see Installation via native package manager.

Use the install.sh helper script, located in the root directory of the RCCL repository, to build and install RCCL with a single command. It uses hard-coded configurations that can be specified directly when using cmake. However, it's a great way to get started quickly and provides an example of how to build and install RCCL.

2.2.1 Building the library using the install script:

To build the library using the install script, use this command:

```
./install.sh
```

For more information on the build options and flags for the install script, run the following command:

```
./install.sh --help
```

The RCCL build and installation helper script options are as follows:

```
   --address-sanitizer     Build with address sanitizer enabled
-d|--dependencies          Install RCCL dependencies
   --debug                 Build debug library
   --enable_backtrace      Build with custom backtrace support
   --disable-colltrace     Build without collective trace
   --disable-msccl-kernel  Build without MSCCL kernels
   --disable-mscclpp       Build without MSCCL++ support
-f|--fast                  Quick-build RCCL (local gpu arch only, no backtrace, and␣
↪collective trace support)
```

(continues on next page)

(continued from previous page)

```
-h|--help                  Prints this help message
-i|--install               Install RCCL library (see --prefix argument below)
-j|--jobs                  Specify how many parallel compilation jobs to run ($nproc by␣
↪default)
-l|--local_gpu_only        Only compile for local GPU architecture
   --amdgpu_targets        Only compile for specified GPU architecture(s). For multiple␣
↪targets, separate by ';' (builds for all supported GPU architectures by default)
   --no_clean              Don't delete files if they already exist
   --npkit-enable          Compile with npkit enabled
   --openmp-test-enable    Enable OpenMP in rccl unit tests
   --roctx-enable          Compile with roctx enabled (example usage: rocprof --roctx-
↪trace ./rccl-program)
-p|--package_build         Build RCCL package
   --prefix                Specify custom directory to install RCCL to (default: `/opt/
↪rocm`)
   --rm-legacy-include-dir Remove legacy include dir Packaging added for file/folder␣
↪reorg backward compatibility
   --run_tests_all         Run all rccl unit tests (must be built already)
-r|--run_tests_quick       Run small subset of rccl unit tests (must be built already)
   --static                Build RCCL as a static library instead of shared library
-t|--tests_build           Build rccl unit tests, but do not run
   --time-trace            Plot the build time of RCCL (requires `ninja-build` package␣
↪installed on the system)
   --verbose               Show compile commands
```

Tip

By default, the RCCL install script builds all the GPU targets that are defined in `DEFAULT_GPUS` in CMakeLists.txt. To target specific GPUs and potentially reduce the build time, use `--amdgpu_targets` along with a semicolon (`;`) separated string list of the GPU targets.

CHAPTER

THREE

RUNNING RCCL USING DOCKER

To use Docker to run RCCL, Docker must already be installed on the system. To build the Docker image and run the container, follow these steps.

1. Build the Docker image

   By default, the Dockerfile uses `docker.io/rocm/dev-ubuntu-22.04:latest` as the base Docker image. It then installs RCCL and rccl-tests (in both cases, it uses the version from the RCCL `develop` branch).

   Use this command to build the Docker image:

   ```
   docker build -t rccl-tests -f Dockerfile.ubuntu --pull .
   ```

   The base Docker image, rccl repository, and rccl-tests repository can be modified by using `--build-args` in the `docker build` command above. For example, to use a different base Docker image, use this command:

   ```
   docker build -t rccl-tests -f Dockerfile.ubuntu --build-arg="ROCM_IMAGE_NAME=rocm/
   ↪dev-ubuntu-20.04" --build-arg="ROCM_IMAGE_TAG=6.2" --pull .
   ```

2. Launch an interactive Docker container on a system with AMD GPUs:

   ```
   docker run -it --rm --device=/dev/kfd --device=/dev/dri --group-add video --
   ↪ipc=host --network=host --cap-add=SYS_PTRACE --security-opt seccomp=unconfined␣
   ↪rccl-tests /bin/bash
   ```

To run, for example, the `all_reduce_perf` test from rccl-tests on 8 AMD GPUs from inside the Docker container, use this command:

```
mpirun --allow-run-as-root -np 8 --mca pml ucx --mca btl ^openib -x NCCL_DEBUG=VERSION /
↪workspace/rccl-tests/build/all_reduce_perf -b 1 -e 16G -f 2 -g 1
```

For more information on the rccl-tests options, see the Usage guidelines in the GitHub repository.

CHAPTER

FOUR

BUILDING AND INSTALLING RCCL FROM SOURCE CODE

To build RCCL directly from the source code, follow these steps. This guide also includes instructions explaining how to test the build. For information on using the quick start install script to build RCCL, see *Installing RCCL using the install script*.

4.1 Requirements

The following prerequisites are required to build RCCL:

1. ROCm-supported GPUs
2. Having the ROCm stack installed on the system, including the HIP runtime and the HIP-Clang compiler.

4.1.1 Building the library using CMake:

To build the library from source, follow these steps:

```
git clone --recursive https://github.com/ROCm/rccl.git
cd rccl
mkdir build
cd build
cmake ..
make -j 16      # Or some other suitable number of parallel jobs
```

If you have already cloned the repository, you can checkout the external submodules manually.

```
git submodule update --init --recursive --depth=1
```

You can substitute a different installation path by providing the path as a parameter to `CMAKE_INSTALL_PREFIX`, for example:

```
cmake -DCMAKE_INSTALL_PREFIX=$PWD/rccl-install -DCMAKE_BUILD_TYPE=Release ..
```

> **Note**
>
> Ensure ROCm CMake is installed using the command `apt install rocm-cmake`. By default, CMake builds the component in debug mode unless `DCMAKE_BUILD_TYPE` is specified.

4.1.2 Building the RCCL package and install package:

After you have cloned the repository and built the library as described in the previous section, use this command to build the package:

```
cd rccl/build
make package
sudo dpkg -i *.deb
```

> **Note**
>
> The RCCL package install process requires `sudo` or root access because it creates a directory named `rccl` in `/opt/rocm/`. This is an optional step. RCCL can be used directly by including the path containing `librccl.so`.

4.2 Testing RCCL

The RCCL unit tests are implemented using the Googletest framework in RCCL. These unit tests require Googletest 1.10 or higher to build and run (this dependency can be installed using the `-d` option for `install.sh`). To run the RCCL unit tests, go to the `build` folder and the `test` subfolder, then run the appropriate RCCL unit test executables.

The RCCL unit test names follow this format:

```
CollectiveCall.[Type of test]
```

Filtering of the RCCL unit tests can be done using environment variables and by passing the `--gtest_filter` command line flag:

```
UT_DATATYPES=ncclBfloat16 UT_REDOPS=prod ./rccl-UnitTests --gtest_filter="AllReduce.C*"
```

This command runs only the `AllReduce` correctness tests with the `float16` datatype. A list of the available environment variables for filtering appears at the top of every run. See the Googletest documentation for more information on how to form advanced filters.

There are also other performance and error-checking tests for RCCL. They are maintained separately at https://github.com/ROCm/rccl-tests.

> **Note**
>
> For more information on how to build and run rccl-tests, see the rccl-tests README file .

CHAPTER

FIVE

USING NCCL NET PLUGIN

NCCL provides a way to use external plugins to let NCCL run on many network types. This topic describes the NCCL Net plugin API and how to implement a network plugin for NCCL.

Plugins implement the NCCL network API, and decouple NCCL binary builds which are built against a particular version of the GPU stack (i.e. CUDA) from the network code which is built against a particular version of the networking stack. That way, you can easily integrate any CUDA version with any network stack version.

NCCL network plugins come as a shared library called `libnccl-net.so`. That shared library contains one or more implementations of the NCCL NET API, in the form of versioned structs, filled with pointers to all required functions.

5.1 Plugin architecture

When NCCL is initialized, it will look for a `libnccl-net.so` library and dynamically load it, then look for symbols inside the library.

The `NCCL_NET_PLUGIN` environment variable allows multiple plugins to coexist. If set, NCCL will look for a library with a name of `libnccl-net-${NCCL_NET_PLUGIN}.so`. It is therefore advised to name the library following that pattern, with a symlink pointing `libnccl-net.so` to `libnccl-net-${NCCL_NET_PLUGIN}.so`. That way, if there are multiple plugins in the path, setting `NCCL_NET_PLUGIN` will allow users to select the right plugin.

5.1.1 Struct versioning

Once a library is found, NCCL will look for a symbol named `ncclNet_vX`, with `X` increasing over time. The versioning ensures that the plugin and the NCCL core are compatible.

Plugins are encouraged to provide multiple of those symbols, implementing multiple versions of the NCCL NET API, so that the same plugin can be compiled and support a wide range of NCCL versions.

Conversely, and to ease transition, NCCL can choose to support different plugin versions, looking for the latest ncclNet struct version, but also looking for older ones so that older plugins would still work.

5.1.2 In-network collective operations (`collNet`)

Additionally to the `ncclNet` structure, network plugins can provide a `collNet` structure which implements in-network collective operations, if supported. That can be used by the NCCL `collNet` algorithm to accelerate inter-node reductions in allReduce.

The `collNet` struct is a different, optional struct provided by the network plugin, but its versioning is tied to the ncclNet struct and many functions are common between the two to ease the implementation.

5.1.3 Headers management

To help users build plugins effortlessly, plugins should copy the `ncclNet_vX` definitions they support to their internal includes. An example is shown in `ext-net/example/` where we keep all headers in the `nccl/` directory and provide thin layers to implement old versions on top of newer ones.

The `nccl/` directory is populated with `net_vX.h` files extracting all relevant definitions from old API versions. It also provides error codes in `err.h`.

5.2 API (v6)

Below is the main `ncclNet_v6` struct. Each function is explained in later sections.

```
typedef struct {
// Name of the network (mainly for logs)
const char* name;
// Initialize the network.
ncclResult_t (*init)(ncclDebugLogger_t logFunction);
// Return the number of adapters.
ncclResult_t (*devices)(int* ndev);
// Get various device properties.
ncclResult_t (*getProperties)(int dev, ncclNetProperties_v6_t* props);
// Create a receiving object and provide a handle to connect to it. The
// handle can be up to NCCL_NET_HANDLE_MAXSIZE bytes and will be exchanged
// between ranks to create a connection.
ncclResult_t (*listen)(int dev, void* handle, void** listenComm);
// Connect to a handle and return a sending comm object for that peer.
// This call must not block for the connection to be established, and instead
// should return successfully with sendComm == NULL with the expectation that
// it will be called again until sendComm != NULL.
ncclResult_t (*connect)(int dev, void* handle, void** sendComm);
// Finalize connection establishment after remote peer has called connect.
// This call must not block for the connection to be established, and instead
// should return successfully with recvComm == NULL with the expectation that
// it will be called again until recvComm != NULL.
ncclResult_t (*accept)(void* listenComm, void** recvComm);
// Register/Deregister memory. Comm can be either a sendComm or a recvComm.
// Type is either NCCL_PTR_HOST or NCCL_PTR_CUDA.
ncclResult_t (*regMr)(void* comm, void* data, int size, int type, void** mhandle);
/* DMA-BUF support */
ncclResult_t (*regMrDmaBuf)(void* comm, void* data, size_t size, int type, uint64_t␣
↪offset, int fd, void** mhandle);
ncclResult_t (*deregMr)(void* comm, void* mhandle);
// Asynchronous send to a peer.
// May return request == NULL if the call cannot be performed (or would block)
ncclResult_t (*isend)(void* sendComm, void* data, int size, int tag, void* mhandle,␣
↪void** request);
// Asynchronous recv from a peer.
// May return request == NULL if the call cannot be performed (or would block)
ncclResult_t (*irecv)(void* recvComm, int n, void** data, int* sizes, int* tags, void**␣
↪mhandles, void** request);
// Perform a flush/fence to make sure all data received with NCCL_PTR_CUDA is
// visible to the GPU
ncclResult_t (*iflush)(void* recvComm, int n, void** data, int* sizes, void** mhandles,␣
```

(continues on next page)

(continued from previous page)

```
↪void** request);
// Test whether a request is complete. If size is not NULL, it returns the
// number of bytes sent/received.
ncclResult_t (*test)(void* request, int* done, int* sizes);
// Close and free send/recv comm objects
ncclResult_t (*closeSend)(void* sendComm);
ncclResult_t (*closeRecv)(void* recvComm);
ncclResult_t (*closeListen)(void* listenComm);
} ncclNet_v6_t;
```

5.2.1 Error codes

All plugins functions use NCCL error codes as return value. *ncclSuccess* should be returned upon success.

Otherwise, plugins can return one of the following:

- `ncclSystemError` is the most common error for network plugins, when a call to the linux kernel or a system library fails. This typically includes all network/hardware errors.
- `ncclInternalError` is returned when the NCCL core code is using the network plugin in an incorrect way, for example allocating more requests than it should, or passing an invalid argument to calls.
- `ncclInvalidUsage` should be returned when the error is most likely a user error. This can include misconfiguration, but also sizes mismatch.
- `ncclInvalidArgument` should usually not be used by plugins since arguments should be checked by the NCCL core layer.
- `ncclUnhandledCudaError` is returned when an error comes from CUDA. Since network plugins should not need to rely on CUDA, this should not be common.

5.2.2 Operation overview

NCCL will call the `init` function first, then query the number of network devices with the `devices` function, getting each network device properties with `getProperties`.

To establish a connection between two network devices, NCCL will first call `listen` on the receiving side, pass the returned handle to the sender side of the connection, and call `connect` with that handle. Finally, *accept* will be called on the receiving side to finalize the connection establishment.

Once the connection is established, communication will be done using the functions `isend`, `irecv` and `test`. Prior to calling `isend` or `irecv`, NCCL will call the `regMr` function on all buffers to allow RDMA NICs to prepare buffers. `deregMr` will be used to unregister buffers.

In certain conditions, `iflush` will be called after a receive calls completes to allow the network plugin to flush data and ensure the GPU will observe the newly written data.

To close the connections NCCL will call `closeListen` to close the object returned by `listen`, `closeSend` to close the object returned by `connect` and `closeRecv` to close the object returned by `accept`.

5.2.3 API Functions

Initialization

- `name` - The `name` field should point to a character string with the name of the network plugin. This will be used for all logging, especially when `NCCL_DEBUG=INFO` is set.

> **Note**
>
> Setting NCCL_NET=<plugin name> will ensure a specific network implementation is used, with a matching name. This is not to be confused with NCCL_NET_PLUGIN which defines a suffix to the libnccl-net.so library name to load.

- init - As soon as NCCL finds the plugin and the correct ncclNet symbol, it will call the init function. This will allow the plugin to discover network devices and make sure they are usable. If the init function does not return ncclSuccess, then NCCL will not use the plugin and fall back on internal ones.

  To allow the plugin logs to integrate into the NCCL logs seemlessly, NCCL provides a logging function to init. This function is typically used to allow for INFO and WARN macros within the plugin code adding the following definitions:

```
#define WARN(...) logFunction(NCCL_LOG_WARN, NCCL_ALL, __FILE__, __LINE__, __VA_ARGS__)
#define INFO(FLAGS, ...) logFunction(NCCL_LOG_INFO, (FLAGS), __func__, __LINE__, __VA_
↪ARGS__)
```

- devices - Once the plugin is initialized, NCCL will query the number of devices available. It should not be zero, otherwise NCCL initialization will fail. If no device is present or usable, the init function should not return ncclSuccess.
- getProperties - Right after getting the number of devices, NCCL will query properties for each available network device. These properties are critical when multiple adapters are present to ensure NCCL uses each adapter in the most optimized way.

The name is only used for logging.

The pciPath is the base for all topology detection and should point to the PCI device directory in /sys. This is typically the directory pointed by /sys/class/net/eth0/device or /sys/class/infiniband/mlx5_0/device. If the network interface is virtual, then pciPath should be NULL.

The guid field is used to determine when network adapters are connected to multiple PCI endpoints. For normal cases, it can be set to the device number. If multiple network devices have the same guid, then NCCL will consider the are sharing the same network port to the fabric, hence it will not use the port multiple times.

The ptrSupport field indicates whether or not CUDA pointers are supported. If so, it should be set to NCCL_PTR_HOST``|``NCCL_PTR_CUDA, otherwise it should be set to NCCL_PTR_HOST. If the plugin supports dmabuf, it should set ptrSupport to NCCL_PTR_HOST``|``NCCL_PTR_CUDA``|``NCCL_PTR_DMABUF and provide a regMrDmaBuf function.

The speed field indicates the speed of the network port in Mbps (10^6 bits per second). This is important to ensure proper optimization of flows within the node.

The port field indicates the port number. This is important again for topology detection and flow optimization within the node when a NIC with a single PCI connection is connected to the fabric with multiple ports.

The latency field indicates the network latency in microseconds. This can be useful to improve the NCCL tuning and make sure NCCL switches from tree to ring at the right size.

The maxComms field indicates the maximum number of connections we can create.

The maxRecvs field indicates the maximum number for grouped receive operations (see grouped receive).

Connection establishment

Connections are used in an unidirectional manner. There is therefore a sender side and a receiver side.

- `listen` - To create a connection, NCCL will start by calling `listen` on the receiver side. This function takes a device number as input argument, and should return a local `listenComm` object, and a `handle` to pass to the other side, so that the sender side can connect to the receiver.

  The `handle` is a buffer of size `NCCL_NET_HANDLE_MAXSIZE` and is provided by NCCL.

  This call should never block, but contrary to `connect` and `accept`, `listenComm` should never be `NULL` if the call succeeds.

- `connect` - NCCL will use its bootstrap infrastructure to provide the `handle` to the sender side, then call `connect` on the sender side on a given device index `dev`, providing the `handle`. `connect` should not block either, and instead set `sendComm` to `NULL` and return `ncclSuccess`. In that case, NCCL will call `accept` again until it succeeds.

- `accept` - To finalize the connection, the receiver side will call `accept` on the `listenComm` returned by the `listen` call previously. If the sender did not connect yet, `accept` should not block. It should return `ncclSuccess`, setting `recvComm` to `NULL`. NCCL will call `accept` again until it succeeds.

- `closeListen/closeSend/closeRecv` - Once a `listenComm/sendComm/recvComm` is no longer needed, NCCL will call `closeListen/closeSend/closeRecv` to free the associated resources.

Communication

Communication is done using asynchronous send and receive operations: `isend`, `irecv` and `test`. To support RDMA capabilities, buffer registration and flush functions are provided.

To keep track of asynchronous send, receive and flush operations, requests are returned to NCCL, then queried with `test`. Each `sendComm` or `recvComm` must be able to handle `NCCL_NET_MAX_REQUESTS` requests in parallel.

> **Note**
>
> That value should be multiplied by the multi-receive capability of the plugin for the sender side, so that we can effectively have `NCCL_NET_MAX_REQUESTS` multi-receive operations happening in parallel. So, if we have a *maxRecvs`value of 8 and ``NCCL_NET_MAX_REQUESTS`* is 8, then each `sendComm` must be able to handle up to 8x8=64 concurrent `isend` operations.

- `regMr` - Prior to sending or receiving data, NCCL will call `regMr` with any buffers later used for communication. It will provide a `sendComm` or `recvComm` as `comm` argument, then the buffer pointer `data`, `size`, and `type` being either `NCCL_PTR_HOST`, or `NCCL_PTR_CUDA` if the network supports CUDA pointers.

  The network plugin can use the output argument *mhandle* to keep any reference to that memory registration, as this `mhandle` will be passed back for all `isend`, `irecv`, `iflush` and `deregMr` calls.

- `regMrDmaBuf` - If the plugin has set the `NCCL_PTR_DMABUF` property in `ptrSupport`, NCCL will use `regMrDmaBuf` instead of `regMr`. If the property was not set, `regMrDmaBuf` can be set to `NULL`.

- `deregMr` - When buffers will no longer be used for communication, NCCL will call `deregMr` to let the plugin free resources. This function is used to deregister handles returned by both `regMr` and `regMrDmaBuf`.

- `isend` - Data will be sent through the connection using `isend`, passing the `sendComm` previously created by `connect`, and the buffer described by `data`, `size`, and `mhandle`. A `tag` must be used if the network supports multi-receive operations (see `irecv`) to distinguish between different sends matching the same multi-receive. Otherwise it can be set to 0.

  The `isend` operation returns a handle in the `request` argument for further calls to `test`. If the `isend` operation cannot be initiated, `request` can be set to `NULL` and NCCL will call `isend` again later.

- `irecv` - To receive data, NCCL will call `irecv` with the `recvComm` returned by `accept`. The argument `n` will allow NCCL to perform a multi-receive, to allow grouping of multiple sends through a single network connection. Each buffer will be described by the `data`, `sizes`, and `mhandles` arrays. `tags` will specify a tag for each receive so that each of the `n` independent `isend` operations is received into the right buffer.

  If all receive operations can be initiated, `irecv` will return a handle in the `request` pointer, otherwise it will set it to `NULL`. In the case of multi-receive, all `n` receive operations are handled by a single request handle.

  The sizes provided to `irecv` can (and will) be larger than the size of the `isend` operation. However, if the receive size is smaller than the send size this is an error.

Note

For a given connection, send/receive operations should always match in the order they were posted. Tags provided for receive operations are only used to assign a given send operation to one of the buffers of the first (multi-)receive in the queue, not to allow for out-of-order tag matching on any receive operation posted.

- `test` - After an `isend` or `irecv` operation is initiated, NCCL will call *test* on the request handles until they complete. When that happens, `done` will be set to 1 and `sizes` will be set to the real size sent or received, the latter being potentially lower than the size passed to `irecv`.

  In the case of a multi-receive, all receives will be considered as done as a single operation (the goal being to allow aggregation), hence they share a single request and a single `done` status. However, they can have different sizes, so when `done` is non-zero, the `sizes` array should contain the `n` sizes corresponding to the buffers passed to `irecv`.

  Once `test` returns 1 in `done`, the request handle can be freed, meaning that NCCL will never call `test` again on that request (until it is reallocated by another call to `isend` or `irecv`).

- `iflush` - After a receive operation completes, if the operation was targeting GPU memory and received a non-zero number of bytes, NCCL will call `iflush` to let the network flush any buffer and ensure the GPU can read it right after without seeing stale data. This flush operation is decoupled from the `test` code to improve latency of `LL*` protocols, as those are capable of determining when data is valid or not.

  `iflush` returns a request which needs to be queried with `test` until it completes.

CHAPTER

SIX

TROUBLESHOOTING RCCL

This topic explains the steps to troubleshoot functional and performance issues with RCCL. While debugging, collect the output from the commands in this guide. This data can be used as supporting information when submitting an issue report to AMD.

6.1 Collecting system information

Collect this information about the ROCm version, GPU/accelerator, platform, and configuration.

- Verify the ROCm version. This might be a release version or a mainline or staging version. Use this command to display the version:

```
cat /opt/rocm/.info/version
```

  Run the following command and collect the output:

```
rocm_agent_enumerator
```

  Also, collect the name of the GPU or accelerator:

```
rocminfo
```

- Run these `rocm-smi` commands to display the system topology.

```
rocm-smi
rocm-smi --showtopo
rocm-smi --showdriverversion
```

- Determine the values of the `PATH` and `LD_LIBRARY_PATH` environment variables.

```
echo $PATH
echo $LD_LIBRARY_PATH
```

- Collect the HIP configuration.

```
/opt/rocm/bin/hipconfig --full
```

- Verify the network settings and setup. Use the `ibv_devinfo` command to display information about the available RDMA devices and determine whether they are installed and functioning properly. Run `rdma link` to print a summary of the network links.

```
ibv_devinfo
rdma link
```

6.1.1 Isolating the issue

The problem might be a general issue or specific to the architecture or system. To narrow down the issue, collect information about the GPU or accelerator and other details about the platform and system. Some issues to consider include:

- Is ROCm running on:
  - A bare-metal setup
  - In a Docker container (determine the name of the Docker image)
  - In an SR-IOV virtualized
  - Some combination of these configurations
- Is the problem only seen on a specific GPU architecture?
- Is it only seen on a specific system type?
- Is it happening on a single node or multinode setup?
- Use the following troubleshooting techniques to attempt to isolate the issue.
  - Build or run the develop branch version of RCCL and see if the problem persists.
  - Try an earlier RCCL version (minor or major).
  - If you recently changed the ROCm runtime configuration, AMD Kernel-mode GPU Driver (KMD), or compiler, rerun the test with the previous configuration.

6.2 Collecting RCCL information

Collect the following information about the RCCL installation and configuration.

- Run the `ldd` command to list any dynamic dependencies for RCCL.

```
ldd <specify-path-to-librccl.so>
```

- Determine the RCCL version. This might be the pre-packaged component in `/opt/rocm/lib` or a version that was built from source. To verify the RCCL version, enter the following command, then run either rccl-tests or an e2e application.

```
export NCCL_DEBUG=VERSION
```

- Run rccl-tests and collect the results. For information on how to build and run rccl-tests, see the rccl-tests GitHub.
- Collect the RCCL logging information. Enable the debug logs, then run rccl-tests or any e2e workload to collect the logs. Use the following command to enable the logs.

```
export NCCL_DEBUG=INFO
```

6.2.1 Using the RCCL Replayer

The RCCL Replayer is a debugging tool designed to analyze and replay the collective logs obtained from RCCL runs. It can be helpful when trying to reproduce problems, because it uses dummy data and doesn't have any dependencies on non-RCCL calls. For more information, see RCCL Replayer GitHub documentation.

You must build the RCCL Replayer before you can use it. To build it, run these commands. Ensure `MPI_DIR` is set to the path where MPI is installed.

```
cd rccl/tools/rccl_replayer
MPI_DIR=/path/to/mpi make
```

To use the RCCL Replayer, follow these steps:

1. Collect the per-rank logs from the RCCL run by adding the following environment variables. This prevents any race conditions that might cause ranks to interrupt the output from other ranks.

   ```
   NCCL_DEBUG=INFO NCCL_DEBUG_SUBSYS=COLL NCCL_DEBUG_FILE=some_name_here.%h.%p.log
   ```

2. Combine all the logs into a single file. This will become the input to the RCCL Replayer.

   ```
   cat some_name_here_*.log > some_name_here.log
   ```

3. Run the RCCL Replayer using the following command. Replace <numProcesses> with the number of MPI processes to run, </path/to/logfile> with the path to the collective log file generated during the RCCL runs, and <numGpusPerMpiRank> with the number of GPUs per MPI rank used in the application.

   ```
   mpirun -np <numProcesses> ./rcclReplayer </path/to/logfile> <numGpusPerMpiRank>
   ```

   In a multi-node application environment, you can replay the collective logs on multiple nodes using the following command:

   ```
   mpirun --hostfile <path/to/hostfile.txt> -np <numProcesses> ./rcclReplayer </path/
   ↪to/logfile> <numGpusPerMpiRank>
   ```

   Note

   Depending on the MPI library you're using, you might need to modify the `mpirun` command.

6.3 Analyzing performance issues

If the issues involve performance issues in an e2e workload, try the following microbenchmarks and collect the results. Follow the instructions in the subsequent sections to run these benchmarks and provide the results to the support team.

- TransferBench
- RCCL Unit Tests
- rccl-tests

6.3.1 Collect the TransferBench data

TransferBench allows you to benchmark simultaneous copies between user-specified devices. For more information, see the TransferBench documentation.

To collect the TransferBench data, follow these steps:

1. Clone the TransferBench Git repository.

   ```
   git clone https://github.com/ROCm/TransferBench.git
   ```

2. Change to the new directory and build the component.

```
cd TransferBench
make
```

3. Run the TransferBench utility with the following parameters and save the results.

```
USE_FINE_GRAIN=1 GFX_UNROLL=2 ./TransferBench a2a 64M 8
```

6.3.2 Collect the RCCL microbenchmark data

To use the RCCL tests to collect the RCCL benchmark data, follow these steps:

1. Disable NUMA auto-balancing using the following command:

```
sudo sysctl kernel.numa_balancing=0
```

   Run the following command to verify the setting. The expected output is `0`.

```
cat /proc/sys/kernel/numa_balancing
```

2. Build MPI, RCCL, and rccl-tests. To download and install MPI, see either OpenMPI or MPICH. To learn how to build and run rccl-tests, see the rccl-tests GitHub.
3. Run rccl-tests with MPI and collect the performance numbers.

6.4 RCCL and NCCL comparisons

If you are also using NVIDIA hardware or NCCL and notice a performance gap between the two systems, collect the system and performance data on the NVIDIA platform. Provide both sets of data to the support team.

CHAPTER

SEVEN

RCCL USAGE TIPS

This topic describes some of the more common RCCL extensions, such as NPKit and MSCCL, and provides tips on how to configure and customize the application.

7.1 NPKit

RCCL integrates NPKit, a profiler framework that enables the collection of fine-grained trace events in RCCL components, especially in giant collective GPU kernels. See the NPKit sample workflow for RCCL for a fully-automated usage example. It also provides useful templates for the following manual instructions.

To manually build RCCL with NPKit enabled, pass `-DNPKIT_FLAGS="-DENABLE_NPKIT -DENABLE_NPKIT_... (other NPKit compile-time switches)"` to the `cmake` command. All NPKit compile-time switches are declared in the RCCL code base as macros with the prefix `ENABLE_NPKIT_`. These switches control the information that is collected.

> **Note**
>
> NPKit only supports the collection of non-overlapped events on the GPU. The `-DNPKIT_FLAGS` settings must follow this rule.

To manually run RCCL with NPKit enabled, set the environment variable `NPKIT_DUMP_DIR` to the NPKit event dump directory. NPKit only supports one GPU per process. To manually analyze the NPKit dump results, use npkit_trace_generator.py.

7.2 MSCCL/MSCCL++

RCCL integrates MSCCL and MSCCL++ to leverage these highly efficient GPU-GPU communication primitives for collective operations. Microsoft Corporation collaborated with AMD for this project.

MSCCL uses XMLs for different collective algorithms on different architectures. RCCL collectives can leverage these algorithms after the user provides the corresponding XML. The XML files contain sequences of send-recv and reduction operations for the kernel to run.

MSCCL is enabled by default on the AMD Instinct™ MI300X accelerator. On other platforms, users might have to enable it using the setting `RCCL_MSCCL_FORCE_ENABLE=1`. By default, MSCCL is only used if every rank belongs to a unique process. To disable this restriction for multi-threaded or single-threaded configurations, use the setting `RCCL_MSCCL_ENABLE_SINGLE_PROCESS=1`.

RCCL allreduce and allgather collectives can leverage the efficient MSCCL++ communication kernels for certain message sizes. MSCCL++ support is available whenever MSCCL support is available. To run a RCCL workload with MSCCL++ support, set the following RCCL environment variable:

```
RCCL_MSCCLPP_ENABLE=1
```

To set the message size threshold for using MSCCL++, use the environment variable `RCCL_MSCCLPP_THRESHOLD`, which has a default value of 1MB. After `RCCL_MSCCLPP_THRESHOLD` has been set, RCCL invokes MSCCL++ kernels for all message sizes less than or equal to the specified threshold.

The following restrictions apply when using MSCCL++. If these restrictions are not met, operations fall back to using MSCCL or RCCL.

- The message size must be a non-zero multiple of 32 bytes
- It does not support `hipMallocManaged` buffers
- Allreduce only supports the `float16`, `int32`, `uint32`, `float32`, and `bfloat16` data types
- Allreduce only supports the sum operation

7.3 Enabling peer-to-peer transport

To enable peer-to-peer access on machines with PCIe-connected GPUs, set the HSA environment variable as follows:

```
HSA_FORCE_FINE_GRAIN_PCIE=1
```

This feature requires GPUs that support peer-to-peer access along with proper large BAR addressing support.

7.4 Improving performance on the MI300X accelerator when using fewer than 8 GPUs

On a system with 8*MI300X accelerators, each pair of accelerators is connected with dedicated XGMI links in a fully-connected topology. For collective operations, this can achieve good performance when all 8 accelerators (and all XGMI links) are used. When fewer than 8 GPUs are used, however, this can only achieve a fraction of the potential bandwidth on the system. However, if your workload warrants using fewer than 8 MI300X accelerators on a system, you can set the run-time variable `NCCL_MIN_NCHANNELS` to increase the number of channels. For example:

```
export NCCL_MIN_NCHANNELS=32
```

Increasing the number of channels can benefit performance, but it also increases GPU utilization for collective operations. Additionally, RCCL pre-defines a higher number of channels when only 2 or 4 accelerators are in use on a 8*MI300X system. In this situation, RCCL uses 32 channels with two MI300X accelerators and 24 channels for four MI300X accelerators.

CHAPTER

EIGHT

RCCL LIBRARY SPECIFICATION

This document provides details of the API library.

8.1 Communicator functions

ncclResult_t **ncclGetUniqueId**(*ncclUniqueId* *uniqueId)

Generates an ID for ncclCommInitRank.

Generates an ID to be used in ncclCommInitRank. ncclGetUniqueId should be called once by a single rank and the ID should be distributed to all ranks in the communicator before using it as a parameter for ncclCommInitRank.

Parameters
uniqueId – **[out]** Pointer to where uniqueId will be stored

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclCommInitRank**(*ncclComm_t* *comm, int nranks, *ncclUniqueId* commId, int rank)

Creates a new communicator (multi thread/process version).

Rank must be between 0 and nranks-1 and unique within a communicator clique. Each rank is associated to a CUDA device, which has to be set before calling ncclCommInitRank. ncclCommInitRank implicitly syncronizes with other ranks, so it must be called by different threads/processes or use ncclGroupStart/ncclGroupEnd.

Parameters

- **comm** – **[out]** Pointer to created communicator
- **nranks** – **[in]** Total number of ranks participating in this communicator
- **commId** – **[in]** UniqueId required for initialization
- **rank** – **[in]** Current rank to create communicator for

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclCommInitAll**(*ncclComm_t* *comm, int ndev, const int *devlist)

Creates a clique of communicators (single process version).

This is a convenience function to create a single-process communicator clique. Returns an array of ndev newly initialized communicators in comm. comm should be pre-allocated with size at least ndev*sizeof(ncclComm_t). If devlist is NULL, the first ndev HIP devices are used. Order of devlist defines user-order of processors within the communicator.

Parameters

- **comm** – **[out]** Pointer to array of created communicators
- **ndev** – **[in]** Total number of ranks participating in this communicator
- **devlist** – **[in]** Array of GPU device indices to create for

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclCommDestroy**(*ncclComm_t* comm)

Frees local resources associated with communicator object.

Destroy all local resources associated with the passed in communicator object

Parameters
comm – **[in]** Communicator to destroy

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclCommAbort**(*ncclComm_t* comm)

Abort any in-progress calls and destroy the communicator object.

Frees resources associated with communicator object and aborts any operations that might still be running on the device.

Parameters
comm – **[in]** Communicator to abort and destroy

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclCommCount**(const *ncclComm_t* comm, int *count)

Gets the number of ranks in the communicator clique.

Returns the number of ranks in the communicator clique (as set during initialization)

Parameters

- **comm** – **[in]** Communicator to query
- **count** – **[out]** Pointer to where number of ranks will be stored

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclCommCuDevice**(const *ncclComm_t* comm, int *device)

Get the ROCm device index associated with a communicator.

Returns the ROCm device number associated with the provided communicator.

Parameters

- **comm** – **[in]** Communicator to query
- **device** – **[out]** Pointer to where the associated ROCm device index will be stored

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclCommUserRank**(const *ncclComm_t* comm, int *rank)

Get the rank associated with a communicator.

Returns the user-ordered "rank" associated with the provided communicator.

Parameters

- **comm** – **[in]** Communicator to query
- **rank** – **[out]** Pointer to where the associated rank will be stored

Returns
Result code. See *Result Codes* for more details.

8.2 Collective communication operations

Collective communication operations must be called separately for each communicator in a communicator clique.

They return when operations have been enqueued on the hipstream.

Since they may perform inter-CPU synchronization, each call has to be done from a different thread or process, or need to use Group Semantics (see below).

ncclResult_t **ncclReduce**(const void *sendbuff, void *recvbuff, size_t count, *ncclDataType_t* datatype, *ncclRedOp_t* op, int root, *ncclComm_t* comm, hipStream_t stream)

Reduce.

Reduces data arrays of length *count* in *sendbuff* into *recvbuff* using *op* operation. recvbuff* may be NULL on all calls except for root device. root* is the rank (not the HIP device) where data will reside after the operation is complete. In-place operation will happen if sendbuff == recvbuff.

Parameters

- **sendbuff** – **[in]** Local device data buffer to be reduced
- **recvbuff** – **[out]** Data buffer where result is stored (only for *root* rank). May be null for other ranks.
- **count** – **[in]** Number of elements in every send buffer
- **datatype** – **[in]** Data buffer element datatype
- **op** – **[in]** Reduction operator type
- **root** – **[in]** Rank where result data array will be stored
- **comm** – **[in]** Communicator group object to execute on
- **stream** – **[in]** HIP stream to execute collective on

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclBcast**(void *buff, size_t count, *ncclDataType_t* datatype, int root, *ncclComm_t* comm, hipStream_t stream)

(Deprecated) Broadcast (in-place)

Copies *count* values from *root* to all other devices. root is the rank (not the CUDA device) where data resides before the operation is started. This operation is implicitly in-place.

Parameters

- **buff** – **[inout]** Input array on *root* to be copied to other ranks. Output array for all ranks.
- **count** – **[in]** Number of elements in data buffer
- **datatype** – **[in]** Data buffer element datatype
- **root** – **[in]** Rank owning buffer to be copied to others
- **comm** – **[in]** Communicator group object to execute on

- **stream** – **[in]** HIP stream to execute collective on

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclBroadcast**(const void *sendbuff, void *recvbuff, size_t count, *ncclDataType_t* datatype, int root, *ncclComm_t* comm, hipStream_t stream)

Broadcast.

Copies *count* values from *sendbuff* on *root* to *recvbuff* on all devices. root* is the rank (not the HIP device) where data resides before the operation is started. sendbuff* may be NULL on ranks other than *root*. In-place operation will happen if *sendbuff* == *recvbuff*.

Parameters

- **sendbuff** – **[in]** Data array to copy (if *root*). May be NULL for other ranks
- **recvbuff** – **[in]** Data array to store received array
- **count** – **[in]** Number of elements in data buffer
- **datatype** – **[in]** Data buffer element datatype
- **root** – **[in]** Rank of broadcast root
- **comm** – **[in]** Communicator group object to execute on
- **stream** – **[in]** HIP stream to execute collective on

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclAllReduce**(const void *sendbuff, void *recvbuff, size_t count, *ncclDataType_t* datatype, *ncclRedOp_t* op, *ncclComm_t* comm, hipStream_t stream)

All-Reduce.

Reduces data arrays of length *count* in *sendbuff* using *op* operation, and leaves identical copies of result on each *recvbuff*. In-place operation will happen if sendbuff == recvbuff.

Parameters

- **sendbuff** – **[in]** Input data array to reduce
- **recvbuff** – **[out]** Data array to store reduced result array
- **count** – **[in]** Number of elements in data buffer
- **datatype** – **[in]** Data buffer element datatype
- **op** – **[in]** Reduction operator
- **comm** – **[in]** Communicator group object to execute on
- **stream** – **[in]** HIP stream to execute collective on

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclReduceScatter**(const void *sendbuff, void *recvbuff, size_t recvcount, *ncclDataType_t* datatype, *ncclRedOp_t* op, *ncclComm_t* comm, hipStream_t stream)

Reduce-Scatter.

Reduces data in *sendbuff* using *op* operation and leaves reduced result scattered over the devices so that *recvbuff* on rank i will contain the i-th block of the result. Assumes sendcount is equal to nranks*recvcount, which means that *sendbuff* should have a size of at least nranks*recvcount elements. In-place operations will happen if recvbuff == sendbuff + rank * recvcount.

Parameters

- **sendbuff** – **[in]** Input data array to reduce
- **recvbuff** – **[out]** Data array to store reduced result subarray
- **recvcount** – **[in]** Number of elements each rank receives
- **datatype** – **[in]** Data buffer element datatype
- **op** – **[in]** Reduction operator
- **comm** – **[in]** Communicator group object to execute on
- **stream** – **[in]** HIP stream to execute collective on

Returns

Result code. See *Result Codes* for more details.

ncclResult_t **ncclAllGather**(const void *sendbuff, void *recvbuff, size_t sendcount, *ncclDataType_t* datatype, *ncclComm_t* comm, hipStream_t stream)

All-Gather.

Each device gathers *sendcount* values from other GPUs into *recvbuff*, receiving data from rank i at offset i*sendcount. Assumes recvcount is equal to nranks*sendcount, which means that recvbuff should have a size of at least nranks*sendcount elements. In-place operations will happen if sendbuff == recvbuff + rank * sendcount.

Parameters

- **sendbuff** – **[in]** Input data array to send
- **recvbuff** – **[out]** Data array to store the gathered result
- **sendcount** – **[in]** Number of elements each rank sends
- **datatype** – **[in]** Data buffer element datatype
- **comm** – **[in]** Communicator group object to execute on
- **stream** – **[in]** HIP stream to execute collective on

Returns

Result code. See *Result Codes* for more details.

ncclResult_t **ncclSend**(const void *sendbuff, size_t count, *ncclDataType_t* datatype, int peer, *ncclComm_t* comm, hipStream_t stream)

Send.

Send data from *sendbuff* to rank *peer*. Rank *peer* needs to call ncclRecv with the same *datatype* and the same *count* as this rank. This operation is blocking for the GPU. If multiple ncclSend and ncclRecv operations need to progress concurrently to complete, they must be fused within a ncclGroupStart / ncclGroupEnd section.

Parameters

- **sendbuff** – **[in]** Data array to send
- **count** – **[in]** Number of elements to send
- **datatype** – **[in]** Data buffer element datatype
- **peer** – **[in]** Peer rank to send to
- **comm** – **[in]** Communicator group object to execute on
- **stream** – **[in]** HIP stream to execute collective on

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclRecv**(void *recvbuff, size_t count, *ncclDataType_t* datatype, int peer, *ncclComm_t* comm, hipStream_t stream)

Receive.

Receive data from rank *peer* into *recvbuff*. Rank *peer* needs to call ncclSend with the same datatype and the same count as this rank. This operation is blocking for the GPU. If multiple ncclSend and ncclRecv operations need to progress concurrently to complete, they must be fused within a ncclGroupStart/ ncclGroupEnd section.

Parameters

- **recvbuff** – **[out]** Data array to receive
- **count** – **[in]** Number of elements to receive
- **datatype** – **[in]** Data buffer element datatype
- **peer** – **[in]** Peer rank to send to
- **comm** – **[in]** Communicator group object to execute on
- **stream** – **[in]** HIP stream to execute collective on

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclGather**(const void *sendbuff, void *recvbuff, size_t sendcount, *ncclDataType_t* datatype, int root, *ncclComm_t* comm, hipStream_t stream)

Gather.

Root device gathers *sendcount* values from other GPUs into *recvbuff*, receiving data from rank i at offset i*sendcount. Assumes recvcount is equal to nranks*sendcount, which means that *recvbuff* should have a size of at least nranks*sendcount elements. In-place operations will happen if sendbuff == recvbuff + rank * sendcount. recvbuff* may be NULL on ranks other than *root*.

Parameters

- **sendbuff** – **[in]** Data array to send
- **recvbuff** – **[out]** Data array to receive into on *root*.
- **sendcount** – **[in]** Number of elements to send per rank
- **datatype** – **[in]** Data buffer element datatype
- **root** – **[in]** Rank that receives data from all other ranks
- **comm** – **[in]** Communicator group object to execute on
- **stream** – **[in]** HIP stream to execute collective on

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclScatter**(const void *sendbuff, void *recvbuff, size_t recvcount, *ncclDataType_t* datatype, int root, *ncclComm_t* comm, hipStream_t stream)

Scatter.

Scattered over the devices so that recvbuff on rank i will contain the i-th block of the data on root. Assumes sendcount is equal to nranks*recvcount, which means that *sendbuff* should have a size of at least nranks*recvcount elements. In-place operations will happen if recvbuff == sendbuff + rank * recvcount.

Parameters

- **sendbuff** – **[in]** Data array to send (on *root* rank). May be NULL on other ranks.
- **recvbuff** – **[out]** Data array to receive partial subarray into
- **recvcount** – **[in]** Number of elements to receive per rank
- **datatype** – **[in]** Data buffer element datatype
- **root** – **[in]** Rank that scatters data to all other ranks
- **comm** – **[in]** Communicator group object to execute on
- **stream** – **[in]** HIP stream to execute collective on

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclAllToAll**(const void *sendbuff, void *recvbuff, size_t count, *ncclDataType_t* datatype, *ncclComm_t* comm, hipStream_t stream)

All-To-All.

Device (i) send (j)th block of data to device (j) and be placed as (i)th block. Each block for sending/receiving has *count* elements, which means that *recvbuff* and *sendbuff* should have a size of nranks*count elements. In-place operation is NOT supported. It is the user's responsibility to ensure that sendbuff and recvbuff are distinct.

Parameters

- **sendbuff** – **[in]** Data array to send (contains blocks for each other rank)
- **recvbuff** – **[out]** Data array to receive (contains blocks from each other rank)
- **count** – **[in]** Number of elements to send between each pair of ranks
- **datatype** – **[in]** Data buffer element datatype
- **comm** – **[in]** Communicator group object to execute on
- **stream** – **[in]** HIP stream to execute collective on

Returns
Result code. See *Result Codes* for more details.

8.3 Group semantics

When managing multiple GPUs from a single thread, and since NCCL collective calls may perform inter-CPU synchronization, we need to "group" calls for different ranks/devices into a single call.

Grouping NCCL calls as being part of the same collective operation is done using ncclGroupStart and ncclGroupEnd. ncclGroupStart will enqueue all collective calls until the ncclGroupEnd call, which will wait for all calls to be complete. Note that for collective communication, ncclGroupEnd only guarantees that the operations are enqueued on the streams, not that the operation is effectively done.

Both collective communication and ncclCommInitRank can be used in conjunction of ncclGroupStart/ncclGroupEnd.

ncclResult_t **ncclGroupStart**()

Group Start.

Start a group call. All calls to RCCL until ncclGroupEnd will be fused into a single RCCL operation. Nothing will be started on the HIP stream until ncclGroupEnd.

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclGroupEnd**()

Group End.

End a group call. Start a fused RCCL operation consisting of all calls since ncclGroupStart. Operations on the HIP stream depending on the RCCL operations need to be called after ncclGroupEnd.

Returns
Result code. See *Result Codes* for more details.

8.4 Library functions

ncclResult_t **ncclGetVersion**(int *version)

Return the RCCL_VERSION_CODE of RCCL in the supplied integer.

This integer is coded with the MAJOR, MINOR and PATCH level of RCCL.

Parameters
version – **[out]** Pointer to where version will be stored

Returns
Result code. See *Result Codes* for more details.

const char ***ncclGetErrorString**(*ncclResult_t* result)

Returns a string for each result code.

Returns a human-readable string describing the given result code.

Parameters
result – **[in]** Result code to get description for

Returns
String containing description of result code.

8.5 Types

There are few data structures that are internal to the library. The pointer types to these structures are given below. The user would need to use these types to create handles and pass them between different library functions.

typedef struct ncclComm ***ncclComm_t**

Opaque handle to communicator.

A communicator contains information required to facilitate collective communications calls

struct **ncclUniqueId**

Opaque unique id used to initialize communicators.

The *ncclUniqueId* must be passed to all participating ranks

8.6 Enumerations

This section provides all the enumerations used.

enum **ncclResult_t**

Result type.

Return codes aside from ncclSuccess indicate that a call has failed

Values:

enumerator **ncclSuccess**

No error

enumerator **ncclUnhandledCudaError**

Unhandled HIP error

enumerator **ncclSystemError**

Unhandled system error

enumerator **ncclInternalError**

Internal Error - Please report to RCCL developers

enumerator **ncclInvalidArgument**

Invalid argument

enumerator **ncclInvalidUsage**

Invalid usage

enumerator **ncclRemoteError**

Remote process exited or there was a network error

enumerator **ncclInProgress**

RCCL operation in progress

enumerator **ncclNumResults**

Number of result types

enum **ncclRedOp_t**

Reduction operation selector.

Enumeration used to specify the various reduction operations ncclNumOps is the number of built-in ncclRedOp_t values and serves as the least possible value for dynamic ncclRedOp_t values constructed by ncclRedOpCreate functions.

ncclMaxRedOp is the largest valid value for ncclRedOp_t and is defined to be the largest signed value (since compilers are permitted to use signed enums) that won't grow sizeof(ncclRedOp_t) when compared to previous RCCL versions to maintain ABI compatibility.

Values:

enumerator **ncclSum**

Sum

enumerator **ncclProd**

Product

enumerator **ncclMax**

Max

enumerator **ncclMin**

Min

enumerator **ncclAvg**

Average

enumerator **ncclNumOps**

Number of built-in reduction ops

enumerator **ncclMaxRedOp**

Largest value for ncclRedOp_t

enum **ncclDataType_t**

Data types.

Enumeration of the various supported datatype

Values:

enumerator **ncclInt8**

enumerator **ncclChar**

enumerator **ncclUint8**

enumerator **ncclInt32**

enumerator **ncclInt**

enumerator **ncclUint32**

enumerator **ncclInt64**

enumerator **ncclUint64**

enumerator **ncclFloat16**

enumerator **ncclHalf**

enumerator **ncclFloat32**

enumerator **ncclFloat**

enumerator **ncclFloat64**

enumerator **ncclDouble**

enumerator **ncclBfloat16**

enumerator **ncclFp8E4M3**

enumerator **ncclFp8E5M2**

enumerator **ncclNumTypes**

CHAPTER

NINE

API LIBRARY

struct **ncclConfig_t**

Communicator configuration.

Users can assign value to attributes to specify the behavior of a communicator

Public Members

size_t **size**

Should not be touched

unsigned int **magic**

Should not be touched

unsigned int **version**

Should not be touched

int **blocking**

Whether or not calls should block or not

int **cgaClusterSize**

Cooperative group array cluster size

int **minCTAs**

Minimum number of cooperative thread arrays (blocks)

int **maxCTAs**

Maximum number of cooperative thread arrays (blocks)

const char ***netName**

Force NCCL to use a specfic network

int **splitShare**

Allow communicators to share resources

struct **ncclUniqueId**

Opaque unique id used to initialize communicators.

The *ncclUniqueId* must be passed to all participating ranks

Public Members

char **internal**[NCCL_UNIQUE_ID_BYTES]

Opaque array>

file **mainpage.txt**

file **nccl.h.in**

#include <hip/hip_runtime.h>#include <hip/hip_fp16.h>#include <limits.h>

Defines

NCCL_H_

NCCL_MAJOR

NCCL_MINOR

NCCL_PATCH

NCCL_SUFFIX

NCCL_VERSION_CODE

NCCL_VERSION(X, Y, Z)

RCCL_BFLOAT16

RCCL_FLOAT8

RCCL_GATHER_SCATTER

RCCL_ALLTOALLV

NCCL_COMM_NULL

NCCL_UNIQUE_ID_BYTES

NCCL_CONFIG_UNDEF_INT

NCCL_CONFIG_UNDEF_PTR

NCCL_SPLIT_NOCOLOR

NCCL_CONFIG_INITIALIZER

Typedefs

typedef struct ncclComm *ncclComm_t

Opaque handle to communicator.

A communicator contains information required to facilitate collective communications calls

typedef int msccl_algo_handle_t

Opaque handle to MSCCL algorithm.

Enums

enum ncclResult_t

Result type.

Return codes aside from ncclSuccess indicate that a call has failed

Values:

enumerator ncclSuccess

No error

enumerator ncclUnhandledCudaError

Unhandled HIP error

enumerator ncclSystemError

Unhandled system error

enumerator ncclInternalError

Internal Error - Please report to RCCL developers

enumerator ncclInvalidArgument

Invalid argument

enumerator ncclInvalidUsage

Invalid usage

enumerator ncclRemoteError

Remote process exited or there was a network error

enumerator **ncclInProgress**

RCCL operation in progress

enumerator **ncclNumResults**

Number of result types

enum **ncclRedOp_dummy_t**

Dummy reduction enumeration.

Dummy reduction enumeration used to determine value for ncclMaxRedOp

Values:

enumerator **ncclNumOps_dummy**

enum **ncclRedOp_t**

Reduction operation selector.

Enumeration used to specify the various reduction operations ncclNumOps is the number of built-in ncclRedOp_t values and serves as the least possible value for dynamic ncclRedOp_t values constructed by ncclRedOpCreate functions.

ncclMaxRedOp is the largest valid value for ncclRedOp_t and is defined to be the largest signed value (since compilers are permitted to use signed enums) that won't grow sizeof(ncclRedOp_t) when compared to previous RCCL versions to maintain ABI compatibility.

Values:

enumerator **ncclSum**

Sum

enumerator **ncclProd**

Product

enumerator **ncclMax**

Max

enumerator **ncclMin**

Min

enumerator **ncclAvg**

Average

enumerator **ncclNumOps**

Number of built-in reduction ops

enumerator **ncclMaxRedOp**

Largest value for ncclRedOp_t

enum **ncclDataType_t**

Data types.

Enumeration of the various supported datatype

Values:

enumerator **ncclInt8**

enumerator **ncclChar**

enumerator **ncclUint8**

enumerator **ncclInt32**

enumerator **ncclInt**

enumerator **ncclUint32**

enumerator **ncclInt64**

enumerator **ncclUint64**

enumerator **ncclFloat16**

enumerator **ncclHalf**

enumerator **ncclFloat32**

enumerator **ncclFloat**

enumerator **ncclFloat64**

enumerator **ncclDouble**

enumerator **ncclBfloat16**

enumerator **ncclFp8E4M3**

enumerator **ncclFp8E5M2**

enumerator **ncclNumTypes**

enum **ncclScalarResidence_t**

Location and dereferencing logic for scalar arguments.

Enumeration specifying memory location of the scalar argument. Based on where the value is stored, the argument will be dereferenced either while the collective is running (if in device memory), or before the ncclRedOpCreate() function returns (if in host memory).

Values:

enumerator **ncclScalarDevice**

Scalar is in device-visible memory

enumerator **ncclScalarHostImmediate**

Scalar is in host-visible memory

Functions

ncclResult_t **ncclMemAlloc**(void **ptr, size_t size)

ncclResult_t **pncclMemAlloc**(void **ptr, size_t size)

ncclResult_t **ncclMemFree**(void *ptr)

ncclResult_t **pncclMemFree**(void *ptr)

ncclResult_t **ncclGetVersion**(int *version)

Return the RCCL_VERSION_CODE of RCCL in the supplied integer.

This integer is coded with the MAJOR, MINOR and PATCH level of RCCL.

Parameters
version – **[out]** Pointer to where version will be stored

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclGetUniqueId**(*ncclUniqueId* *uniqueId)

Generates an ID for ncclCommInitRank.

Generates an ID to be used in ncclCommInitRank. ncclGetUniqueId should be called once by a single rank and the ID should be distributed to all ranks in the communicator before using it as a parameter for ncclCommInitRank.

Parameters
uniqueId – **[out]** Pointer to where uniqueId will be stored

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclCommInitRankConfig**(*ncclComm_t* *comm, int nranks, *ncclUniqueId* commId, int rank, *ncclConfig_t* *config)

Create a new communicator with config.

Create a new communicator (multi thread/process version) with a configuration set by users. See *Communicator Configuration* for more details. Each rank is associated to a CUDA device, which has to be set before calling ncclCommInitRank.

Parameters

- **comm** – **[out]** Pointer to created communicator
- **nranks** – **[in]** Total number of ranks participating in this communicator
- **commId** – **[in]** UniqueId required for initialization
- **rank** – **[in]** Current rank to create communicator for. [0 to nranks-1]
- **config** – **[in]** Pointer to communicator configuration

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclCommInitRank**(*ncclComm_t* *comm, int nranks, *ncclUniqueId* commId, int rank)

Creates a new communicator (multi thread/process version).

Rank must be between 0 and nranks-1 and unique within a communicator clique. Each rank is associated to a CUDA device, which has to be set before calling ncclCommInitRank. ncclCommInitRank implicitly syncronizes with other ranks, so it must be called by different threads/processes or use ncclGroupStart/ncclGroupEnd.

Parameters

- **comm** – **[out]** Pointer to created communicator
- **nranks** – **[in]** Total number of ranks participating in this communicator
- **commId** – **[in]** UniqueId required for initialization
- **rank** – **[in]** Current rank to create communicator for

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclCommInitAll**(*ncclComm_t* *comm, int ndev, const int *devlist)

Creates a clique of communicators (single process version).

This is a convenience function to create a single-process communicator clique. Returns an array of ndev newly initialized communicators in comm. comm should be pre-allocated with size at least ndev*sizeof(ncclComm_t). If devlist is NULL, the first ndev HIP devices are used. Order of devlist defines user-order of processors within the communicator.

Parameters

- **comm** – **[out]** Pointer to array of created communicators
- **ndev** – **[in]** Total number of ranks participating in this communicator
- **devlist** – **[in]** Array of GPU device indices to create for

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclCommFinalize**(*ncclComm_t* comm)

Finalize a communicator.

ncclCommFinalize flushes all issued communications and marks communicator state as ncclInProgress. The state will change to ncclSuccess when the communicator is globally quiescent and related resources are freed; then, calling ncclCommDestroy can locally free the rest of the resources (e.g. communicator itself) without blocking.

Parameters
comm – **[in]** Communicator to finalize

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclCommDestroy**(*ncclComm_t* comm)

Frees local resources associated with communicator object.

Destroy all local resources associated with the passed in communicator object

Parameters
comm – **[in]** Communicator to destroy

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclCommAbort**(*ncclComm_t* comm)

Abort any in-progress calls and destroy the communicator object.

Frees resources associated with communicator object and aborts any operations that might still be running on the device.

Parameters
comm – **[in]** Communicator to abort and destroy

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclCommSplit**(*ncclComm_t* comm, int color, int key, *ncclComm_t* *newcomm, *ncclConfig_t* *config)

Create one or more communicators from an existing one.

Creates one or more communicators from an existing one. Ranks with the same color will end up in the same communicator. Within the new communicator, key will be used to order ranks. NCCL_SPLIT_NOCOLOR as color will indicate the rank will not be part of any group and will therefore return a NULL communicator. If config is NULL, the new communicator will inherit the original communicator's configuration

Parameters

- **comm** – **[in]** Original communicator object for this rank
- **color** – **[in]** Color to assign this rank
- **key** – **[in]** Key used to order ranks within the same new communicator
- **newcomm** – **[out]** Pointer to new communicator
- **config** – **[in]** Config file for new communicator. May be NULL to inherit from comm

Returns
Result code. See *Result Codes* for more details.

const char ***ncclGetErrorString**(*ncclResult_t* result)

Returns a string for each result code.

Returns a human-readable string describing the given result code.

Parameters
result – **[in]** Result code to get description for

Returns
String containing description of result code.

const char ***ncclGetLastError**(*ncclComm_t* comm)

ncclResult_t **ncclCommGetAsyncError**(*ncclComm_t* comm, *ncclResult_t* *asyncError)

Checks whether the comm has encountered any asynchronous errors.

Query whether the provided communicator has encountered any asynchronous errors

Parameters

- **comm** – **[in]** Communicator to query
- **asyncError** – **[out]** Pointer to where result code will be stored

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclCommCount**(const *ncclComm_t* comm, int *count)

Gets the number of ranks in the communicator clique.

Returns the number of ranks in the communicator clique (as set during initialization)

Parameters

- **comm** – **[in]** Communicator to query
- **count** – **[out]** Pointer to where number of ranks will be stored

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclCommCuDevice**(const *ncclComm_t* comm, int *device)

Get the ROCm device index associated with a communicator.

Returns the ROCm device number associated with the provided communicator.

Parameters

- **comm** – **[in]** Communicator to query
- **device** – **[out]** Pointer to where the associated ROCm device index will be stored

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclCommUserRank**(const *ncclComm_t* comm, int *rank)

Get the rank associated with a communicator.

Returns the user-ordered "rank" associated with the provided communicator.

Parameters

- **comm** – **[in]** Communicator to query
- **rank** – **[out]** Pointer to where the associated rank will be stored

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclCommRegister**(const *ncclComm_t* comm, void *buff, size_t size, void **handle)

ncclResult_t **ncclCommDeregister**(const *ncclComm_t* comm, void *handle)

ncclResult_t **ncclRedOpCreatePreMulSum**(*ncclRedOp_t* *op, void *scalar, *ncclDataType_t* datatype, *ncclScalarResidence_t* residence, *ncclComm_t* comm)

Create a custom pre-multiplier reduction operator.

Creates a new reduction operator which pre-multiplies input values by a given scalar locally before reducing them with peer values via summation. For use only with collectives launched against *comm* and *datatype*.

The residence* argument indicates how/when the memory pointed to by *scalar* will be dereferenced. Upon return, the newly created operator's handle is stored in *op*.

Parameters

- **op** – **[out]** Pointer to where newly created custom reduction operator is to be stored
- **scalar** – **[in]** Pointer to scalar value.
- **datatype** – **[in]** Scalar value datatype
- **residence** – **[in]** Memory type of the scalar value
- **comm** – **[in]** Communicator to associate with this custom reduction operator

Returns

Result code. See *Result Codes* for more details.

ncclResult_t **ncclRedOpDestroy**(*ncclRedOp_t* op, *ncclComm_t* comm)

Destroy custom reduction operator.

Destroys the reduction operator *op*. The operator must have been created by ncclRedOpCreatePreMul with the matching communicator *comm*. An operator may be destroyed as soon as the last RCCL function which is given that operator returns.

Parameters

- **op** – **[in]** Custom reduction operator is to be destroyed
- **comm** – **[in]** Communicator associated with this reduction operator

Returns

Result code. See *Result Codes* for more details.

ncclResult_t **ncclReduce**(const void *sendbuff, void *recvbuff, size_t count, *ncclDataType_t* datatype, *ncclRedOp_t* op, int root, *ncclComm_t* comm, hipStream_t stream)

Reduce.

Reduces data arrays of length *count* in *sendbuff* into *recvbuff* using *op* operation. recvbuff* may be NULL on all calls except for root device. root* is the rank (not the HIP device) where data will reside after the operation is complete. In-place operation will happen if sendbuff == recvbuff.

Parameters

- **sendbuff** – **[in]** Local device data buffer to be reduced
- **recvbuff** – **[out]** Data buffer where result is stored (only for *root* rank). May be null for other ranks.
- **count** – **[in]** Number of elements in every send buffer
- **datatype** – **[in]** Data buffer element datatype
- **op** – **[in]** Reduction operator type
- **root** – **[in]** Rank where result data array will be stored
- **comm** – **[in]** Communicator group object to execute on
- **stream** – **[in]** HIP stream to execute collective on

Returns

Result code. See *Result Codes* for more details.

ncclResult_t **ncclBcast**(void *buff, size_t count, *ncclDataType_t* datatype, int root, *ncclComm_t* comm, hipStream_t stream)

(Deprecated) Broadcast (in-place)

Copies *count* values from *root* to all other devices. root is the rank (not the CUDA device) where data resides before the operation is started. This operation is implicitly in-place.

Parameters

- **buff** – **[inout]** Input array on *root* to be copied to other ranks. Output array for all ranks.
- **count** – **[in]** Number of elements in data buffer
- **datatype** – **[in]** Data buffer element datatype
- **root** – **[in]** Rank owning buffer to be copied to others
- **comm** – **[in]** Communicator group object to execute on
- **stream** – **[in]** HIP stream to execute collective on

Returns

Result code. See *Result Codes* for more details.

ncclResult_t **ncclBroadcast**(const void *sendbuff, void *recvbuff, size_t count, *ncclDataType_t* datatype, int root, *ncclComm_t* comm, hipStream_t stream)

Broadcast.

Copies *count* values from *sendbuff* on *root* to *recvbuff* on all devices. root* is the rank (not the HIP device) where data resides before the operation is started. sendbuff* may be NULL on ranks other than *root*. In-place operation will happen if *sendbuff* == *recvbuff*.

Parameters

- **sendbuff** – **[in]** Data array to copy (if *root*). May be NULL for other ranks
- **recvbuff** – **[in]** Data array to store received array
- **count** – **[in]** Number of elements in data buffer
- **datatype** – **[in]** Data buffer element datatype
- **root** – **[in]** Rank of broadcast root
- **comm** – **[in]** Communicator group object to execute on
- **stream** – **[in]** HIP stream to execute collective on

Returns

Result code. See *Result Codes* for more details.

ncclResult_t **ncclAllReduce**(const void *sendbuff, void *recvbuff, size_t count, *ncclDataType_t* datatype, *ncclRedOp_t* op, *ncclComm_t* comm, hipStream_t stream)

All-Reduce.

Reduces data arrays of length *count* in *sendbuff* using *op* operation, and leaves identical copies of result on each *recvbuff*. In-place operation will happen if sendbuff == recvbuff.

Parameters

- **sendbuff** – **[in]** Input data array to reduce
- **recvbuff** – **[out]** Data array to store reduced result array
- **count** – **[in]** Number of elements in data buffer

- **datatype** – **[in]** Data buffer element datatype
- **op** – **[in]** Reduction operator
- **comm** – **[in]** Communicator group object to execute on
- **stream** – **[in]** HIP stream to execute collective on

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclReduceScatter**(const void *sendbuff, void *recvbuff, size_t recvcount, *ncclDataType_t* datatype, *ncclRedOp_t* op, *ncclComm_t* comm, hipStream_t stream)

Reduce-Scatter.

Reduces data in *sendbuff* using *op* operation and leaves reduced result scattered over the devices so that *recvbuff* on rank i will contain the i-th block of the result. Assumes sendcount is equal to nranks*recvcount, which means that *sendbuff* should have a size of at least nranks*recvcount elements. In-place operations will happen if recvbuff == sendbuff + rank * recvcount.

Parameters

- **sendbuff** – **[in]** Input data array to reduce
- **recvbuff** – **[out]** Data array to store reduced result subarray
- **recvcount** – **[in]** Number of elements each rank receives
- **datatype** – **[in]** Data buffer element datatype
- **op** – **[in]** Reduction operator
- **comm** – **[in]** Communicator group object to execute on
- **stream** – **[in]** HIP stream to execute collective on

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclAllGather**(const void *sendbuff, void *recvbuff, size_t sendcount, *ncclDataType_t* datatype, *ncclComm_t* comm, hipStream_t stream)

All-Gather.

Each device gathers *sendcount* values from other GPUs into *recvbuff*, receiving data from rank i at offset i*sendcount. Assumes recvcount is equal to nranks*sendcount, which means that recvbuff should have a size of at least nranks*sendcount elements. In-place operations will happen if sendbuff == recvbuff + rank * sendcount.

Parameters

- **sendbuff** – **[in]** Input data array to send
- **recvbuff** – **[out]** Data array to store the gathered result
- **sendcount** – **[in]** Number of elements each rank sends
- **datatype** – **[in]** Data buffer element datatype
- **comm** – **[in]** Communicator group object to execute on
- **stream** – **[in]** HIP stream to execute collective on

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclSend**(const void *sendbuff, size_t count, *ncclDataType_t* datatype, int peer, *ncclComm_t* comm, hipStream_t stream)

Send.

Send data from *sendbuff* to rank *peer*. Rank *peer* needs to call ncclRecv with the same *datatype* and the same *count* as this rank. This operation is blocking for the GPU. If multiple ncclSend and ncclRecv operations need to progress concurrently to complete, they must be fused within a ncclGroupStart / ncclGroupEnd section.

Parameters

- **sendbuff** – **[in]** Data array to send
- **count** – **[in]** Number of elements to send
- **datatype** – **[in]** Data buffer element datatype
- **peer** – **[in]** Peer rank to send to
- **comm** – **[in]** Communicator group object to execute on
- **stream** – **[in]** HIP stream to execute collective on

Returns

Result code. See *Result Codes* for more details.

ncclResult_t **ncclRecv**(void *recvbuff, size_t count, *ncclDataType_t* datatype, int peer, *ncclComm_t* comm, hipStream_t stream)

Receive.

Receive data from rank *peer* into *recvbuff*. Rank *peer* needs to call ncclSend with the same datatype and the same count as this rank. This operation is blocking for the GPU. If multiple ncclSend and ncclRecv operations need to progress concurrently to complete, they must be fused within a ncclGroupStart/ ncclGroupEnd section.

Parameters

- **recvbuff** – **[out]** Data array to receive
- **count** – **[in]** Number of elements to receive
- **datatype** – **[in]** Data buffer element datatype
- **peer** – **[in]** Peer rank to send to
- **comm** – **[in]** Communicator group object to execute on
- **stream** – **[in]** HIP stream to execute collective on

Returns

Result code. See *Result Codes* for more details.

ncclResult_t **ncclGather**(const void *sendbuff, void *recvbuff, size_t sendcount, *ncclDataType_t* datatype, int root, *ncclComm_t* comm, hipStream_t stream)

Gather.

Root device gathers *sendcount* values from other GPUs into *recvbuff*, receiving data from rank i at offset i*sendcount. Assumes recvcount is equal to nranks*sendcount, which means that *recvbuff* should have a size of at least nranks*sendcount elements. In-place operations will happen if sendbuff == recvbuff + rank * sendcount. recvbuff* may be NULL on ranks other than *root*.

Parameters

- **sendbuff** – **[in]** Data array to send

- **recvbuff** – **[out]** Data array to receive into on *root*.
- **sendcount** – **[in]** Number of elements to send per rank
- **datatype** – **[in]** Data buffer element datatype
- **root** – **[in]** Rank that receives data from all other ranks
- **comm** – **[in]** Communicator group object to execute on
- **stream** – **[in]** HIP stream to execute collective on

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclScatter**(const void *sendbuff, void *recvbuff, size_t recvcount, *ncclDataType_t* datatype, int root, *ncclComm_t* comm, hipStream_t stream)

Scatter.

Scattered over the devices so that recvbuff on rank i will contain the i-th block of the data on root. Assumes sendcount is equal to nranks*recvcount, which means that *sendbuff* should have a size of at least nranks*recvcount elements. In-place operations will happen if recvbuff == sendbuff + rank * recvcount.

Parameters

- **sendbuff** – **[in]** Data array to send (on *root* rank). May be NULL on other ranks.
- **recvbuff** – **[out]** Data array to receive partial subarray into
- **recvcount** – **[in]** Number of elements to receive per rank
- **datatype** – **[in]** Data buffer element datatype
- **root** – **[in]** Rank that scatters data to all other ranks
- **comm** – **[in]** Communicator group object to execute on
- **stream** – **[in]** HIP stream to execute collective on

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclAllToAll**(const void *sendbuff, void *recvbuff, size_t count, *ncclDataType_t* datatype, *ncclComm_t* comm, hipStream_t stream)

All-To-All.

Device (i) send (j)th block of data to device (j) and be placed as (i)th block. Each block for sending/receiving has *count* elements, which means that *recvbuff* and *sendbuff* should have a size of nranks*count elements. In-place operation is NOT supported. It is the user's responsibility to ensure that sendbuff and recvbuff are distinct.

Parameters

- **sendbuff** – **[in]** Data array to send (contains blocks for each other rank)
- **recvbuff** – **[out]** Data array to receive (contains blocks from each other rank)
- **count** – **[in]** Number of elements to send between each pair of ranks
- **datatype** – **[in]** Data buffer element datatype
- **comm** – **[in]** Communicator group object to execute on
- **stream** – **[in]** HIP stream to execute collective on

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclAllToAllv**(const void *sendbuff, const size_t sendcounts[], const size_t sdispls[], void *recvbuff, const size_t recvcounts[], const size_t rdispls[], *ncclDataType_t* datatype, *ncclComm_t* comm, hipStream_t stream)

All-To-Allv.

Device (i) sends sendcounts[j] of data from offset sdispls[j] to device (j). At the same time, device (i) receives recvcounts[j] of data from device (j) to be placed at rdispls[j]. sendcounts, sdispls, recvcounts and rdispls are all measured in the units of datatype, not bytes. In-place operation will happen if sendbuff == recvbuff.

Parameters

- **sendbuff** – **[in]** Data array to send (contains blocks for each other rank)
- **sendcounts** – **[in]** Array containing number of elements to send to each participating rank
- **sdispls** – **[in]** Array of offsets into *sendbuff* for each participating rank
- **recvbuff** – **[out]** Data array to receive (contains blocks from each other rank)
- **recvcounts** – **[in]** Array containing number of elements to receive from each participating rank
- **rdispls** – **[in]** Array of offsets into *recvbuff* for each participating rank
- **datatype** – **[in]** Data buffer element datatype
- **comm** – **[in]** Communicator group object to execute on
- **stream** – **[in]** HIP stream to execute collective on

Returns

Result code. See *Result Codes* for more details.

ncclResult_t **mscclLoadAlgo**(const char *mscclAlgoFilePath, *mscclAlgoHandle_t* *mscclAlgoHandle, int rank)

MSCCL Load Algorithm.

Load MSCCL algorithm file specified in mscclAlgoFilePath and return its handle via mscclAlgoHandle. This API is expected to be called by MSCCL scheduler instead of end users.

Parameters

- **mscclAlgoFilePath** – **[in]** Path to MSCCL algorithm file
- **mscclAlgoHandle** – **[out]** Returned handle to MSCCL algorithm
- **rank** – **[in]** Current rank

Returns

Result code. See *Result Codes* for more details.

ncclResult_t **mscclRunAlgo**(const void *sendBuff, const size_t sendCounts[], const size_t sDisPls[], void *recvBuff, const size_t recvCounts[], const size_t rDisPls[], size_t count, *ncclDataType_t* dataType, int root, int peer, *ncclRedOp_t* op, *mscclAlgoHandle_t* mscclAlgoHandle, *ncclComm_t* comm, hipStream_t stream)

MSCCL Run Algorithm.

Run MSCCL algorithm specified by mscclAlgoHandle. The parameter list merges all possible parameters required by different operations as this is a general-purposed API. This API is expected to be called by MSCCL scheduler instead of end users.

Parameters

- **sendBuff** – **[in]** Data array to send
- **sendCounts** – **[in]** Array containing number of elements to send to each participating rank
- **sDisPls** – **[in]** Array of offsets into *sendbuff* for each participating rank
- **recvBuff** – **[out]** Data array to receive
- **recvCounts** – **[in]** Array containing number of elements to receive from each participating rank
- **rDisPls** – **[in]** Array of offsets into *recvbuff* for each participating rank
- **count** – **[in]** Number of elements
- **dataType** – **[in]** Data buffer element datatype
- **root** – **[in]** Root rank index
- **peer** – **[in]** Peer rank index
- **op** – **[in]** Reduction operator
- **mscclAlgoHandle** – **[in]** Handle to MSCCL algorithm
- **comm** – **[in]** Communicator group object to execute on
- **stream** – **[in]** HIP stream to execute collective on

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **mscclUnloadAlgo**(*mscclAlgoHandle_t* mscclAlgoHandle)

MSCCL Unload Algorithm.

Deprecated:

This function has been removed from the public API.

Unload MSCCL algorithm previous loaded using its handle. This API is expected to be called by MSCCL scheduler instead of end users.

Parameters
mscclAlgoHandle – **[in]** Handle to MSCCL algorithm to unload

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclGroupStart**()

Group Start.

Start a group call. All calls to RCCL until ncclGroupEnd will be fused into a single RCCL operation. Nothing will be started on the HIP stream until ncclGroupEnd.

Returns
Result code. See *Result Codes* for more details.

ncclResult_t **ncclGroupEnd**()

Group End.

End a group call. Start a fused RCCL operation consisting of all calls since ncclGroupStart. Operations on the HIP stream depending on the RCCL operations need to be called after ncclGroupEnd.

Returns
Result code. See *Result Codes* for more details.

group **rccl_result_code**

The various result codes that RCCL API calls may return

Enums

enum **ncclResult_t**

Result type.

Return codes aside from ncclSuccess indicate that a call has failed

Values:

enumerator **ncclSuccess**

No error

enumerator **ncclUnhandledCudaError**

Unhandled HIP error

enumerator **ncclSystemError**

Unhandled system error

enumerator **ncclInternalError**

Internal Error - Please report to RCCL developers

enumerator **ncclInvalidArgument**

Invalid argument

enumerator **ncclInvalidUsage**

Invalid usage

enumerator **ncclRemoteError**

Remote process exited or there was a network error

enumerator **ncclInProgress**

RCCL operation in progress

enumerator **ncclNumResults**

Number of result types

group **rccl_config_type**

Structure that allows for customizing Communicator behavior via ncclCommInitRankConfig

Defines

NCCL_CONFIG_INITIALIZER

group **rccl_api_version**

API call that returns RCCL version

group **rccl_api_communicator**

API calls that operate on communicators. Communicators objects are used to launch collective communication operations. Unique ranks between 0 and N-1 must be assigned to each HIP device participating in the same Communicator. Using the same HIP device for multiple ranks of the same Communicator is not supported at this time.

group **rccl_api_errcheck**

API calls that check for errors

group **rccl_api_comminfo**

API calls that query communicator information

group **rccl_api_enumerations**

Enumerations used by collective communication calls

Enums

enum **ncclRedOp_dummy_t**

Dummy reduction enumeration.

Dummy reduction enumeration used to determine value for ncclMaxRedOp

Values:

enumerator **ncclNumOps_dummy**

enum **ncclRedOp_t**

Reduction operation selector.

Enumeration used to specify the various reduction operations ncclNumOps is the number of built-in ncclRedOp_t values and serves as the least possible value for dynamic ncclRedOp_t values constructed by ncclRedOpCreate functions.

ncclMaxRedOp is the largest valid value for ncclRedOp_t and is defined to be the largest signed value (since compilers are permitted to use signed enums) that won't grow sizeof(ncclRedOp_t) when compared to previous RCCL versions to maintain ABI compatibility.

Values:

enumerator **ncclSum**

Sum

enumerator **ncclProd**

Product

enumerator **ncclMax**

Max

enumerator **ncclMin**

Min

enumerator **ncclAvg**

Average

enumerator **ncclNumOps**

Number of built-in reduction ops

enumerator **ncclMaxRedOp**

Largest value for ncclRedOp_t

enum **ncclDataType_t**

Data types.

Enumeration of the various supported datatype

Values:

enumerator **ncclInt8**

enumerator **ncclChar**

enumerator **ncclUint8**

enumerator **ncclInt32**

enumerator **ncclInt**

enumerator **ncclUint32**

enumerator **ncclInt64**

enumerator **ncclUint64**

enumerator **ncclFloat16**

enumerator **ncclHalf**

enumerator **ncclFloat32**

enumerator **ncclFloat**

enumerator **ncclFloat64**

enumerator **ncclDouble**

enumerator **ncclBfloat16**

enumerator **ncclFp8E4M3**

enumerator **ncclFp8E5M2**

enumerator **ncclNumTypes**

group **rccl_api_custom_redop**

API calls relating to creation/destroying custom reduction operator that pre-multiplies local source arrays prior to reduction

Enums

enum **ncclScalarResidence_t**

Location and dereferencing logic for scalar arguments.

Enumeration specifying memory location of the scalar argument. Based on where the value is stored, the argument will be dereferenced either while the collective is running (if in device memory), or before the ncclRedOpCreate() function returns (if in host memory).

Values:

enumerator **ncclScalarDevice**

Scalar is in device-visible memory

enumerator **ncclScalarHostImmediate**

Scalar is in host-visible memory

group **rccl_collective_api**

Collective communication operations must be called separately for each communicator in a communicator clique.

They return when operations have been enqueued on the HIP stream. Since they may perform inter-CPU synchronization, each call has to be done from a different thread or process, or need to use Group Semantics (see below).

group **msccl_api**

API calls relating to the optional MSCCL algorithm datapath

Typedefs

typedef int **mscclAlgoHandle_t**

Opaque handle to MSCCL algorithm.

group **rccl_group_api**

When managing multiple GPUs from a single thread, and since RCCL collective calls may perform inter-CPU synchronization, we need to "group" calls for different ranks/devices into a single call.

Grouping RCCL calls as being part of the same collective operation is done using ncclGroupStart and ncclGroupEnd. ncclGroupStart will enqueue all collective calls until the ncclGroupEnd call, which will wait for all calls to be complete. Note that for collective communication, ncclGroupEnd only guarantees that the operations are enqueued on the streams, not that the operation is effectively done.

Both collective communication and ncclCommInitRank can be used in conjunction of ncclGroupStart/ncclGroupEnd, but not together.

Group semantics also allow to fuse multiple operations on the same device to improve performance (for aggregated collective calls), or to permit concurrent progress of multiple send/receive operations.

page **deprecated**

Global *mscclUnloadAlgo* (mscclAlgoHandle_t mscclAlgoHandle)

This function has been removed from the public API.

dir **src**

page **index**

9.1 Introduction

RCCL (pronounced "Rickle") is a stand-alone library of standard collective communication routines for GPUs, implementing all-reduce, all-gather, reduce, broadcast, reduce-scatter, gather, scatter, and all-to-all. There is also initial support for direct GPU-to-GPU send and receive operations. It has been optimized to achieve high bandwidth on platforms using PCIe, xGMI as well as networking using InfiniBand Verbs or TCP/IP sockets. RCCL supports an arbitrary number of GPUs installed in a single node or multiple nodes, and can be used in either single- or multi-process (e.g., MPI) applications.

The collective operations are implemented using ring and tree algorithms and have been optimized for throughput and latency. For best performance, small operations can be either batched into larger operations or aggregated through the API.

9.2 RCCL API Contents

- *Collective Communication Operations*
- *Group semantics*
- *MSCCL Algorithm*

9.3 RCCL API File

- *nccl.h.in*

CHAPTER

TEN

LICENSE

Attributions

Contains contributions from NVIDIA.

Copyright (c) 2015-2020, NVIDIA CORPORATION. All rights reserved. Modifications Copyright (c) 2019-2024 Advanced Micro Devices, Inc. All rights reserved. Modifications Copyright (c) Microsoft Corporation. Licensed under the MIT License.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

- Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
- Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
- Neither the name of NVIDIA CORPORATION, Lawrence Berkeley National Laboratory, the U.S. Department of Energy, nor the names of their contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS ``AS IS'' AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

The U.S. Department of Energy funded the development of this software under subcontract 7078610 with Lawrence Berkeley National Laboratory.

This code also includes files from the NVIDIA Tools Extension SDK project.

See:

https://github.com/NVIDIA/NVTX

for more information and license details.

CHAPTER

ELEVEN

ATTRIBUTIONS

Contains contributions from NVIDIA.

Copyright (c) 2015-2020, NVIDIA CORPORATION. All rights reserved.

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

- Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
- Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
- Neither the name of NVIDIA CORPORATION, Lawrence Berkeley National Laboratory, the U.S. Department of Energy, nor the names of their contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS ``AS IS'' AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

The U.S. Department of Energy funded the development of this software under subcontract 7078610 with Lawrence Berkeley National Laboratory.

This code also includes files from the NVIDIA Tools Extension SDK project.

For more information and license details, see https://github.com/NVIDIA/NVTX

INDEX

M

N

P

R