

[ NCCL ]([index.html](#))

[2.25](<https://docs.nvidia.com/deeplearning/sdk/nccl-archived/index.html>)

- \* [Overview of NCCL]([overview.html](#))

- \* [Setup]([setup.html](#))

- \* [Using NCCL]([usage.html](#))

- \* [Creating a Communicator]([usage/communicators.html](#))

- \* [Creating a communicator with options]([usage/communicators.html#creating-a-communicator-with-options](#))

- \* [Creating a communicator using multiple ncclUniqueIds]([usage/communicators.html#creating-a-communicator-using-multiple-nccluniqueids](#))

- \* [Creating more communicators]([usage/communicators.html#creating-more-communicators](#))

- \* [Using multiple NCCL communicators concurrently]([usage/communicators.html#using-multiple-nccl-communicators-concurrently](#))

- \* [Finalizing a communicator]([usage/communicators.html#finalizing-a-communicator](#))

- \* [Destroying a communicator]([usage/communicators.html#destroying-a-communicator](#))

- \* [Error handling and communicator abort]([usage/communicators.html#error-handling-and-communicator-abort](#))

- \* [Asynchronous errors and error handling]([usage/communicators.html#asynchronous-errors-and-error-handling](#))

- \* [Fault Tolerance]([usage/communicators.html#fault-tolerance](#))

- \* [Collective Operations]([usage/collectives.html](#))

- \* [AllReduce]([usage/collectives.html#allreduce](#))

- \* [Broadcast]([usage/collectives.html#broadcast](#))

- \* [Reduce]([usage/collectives.html#reduce](#))

- \* [AllGather]([usage/collectives.html#allgather](#))

- \* [\[ReduceScatter\]\(usage/collectives.html#reducescatter\)](#)
- \* [\[Data Pointers\]\(usage/data.html\)](#)
- \* [\[CUDA Stream Semantics\]\(usage/streams.html\)](#)
  - \* [\[Mixing Multiple Streams within the same ncclGroupStart/End\(\) group\]\(usage/streams.html#mixing-multiple-streams-within-the-same-ncclgroupstart-end-group\)](#)
- \* [\[Group Calls\]\(usage/groups.html\)](#)
  - \* [\[Management Of Multiple GPUs From One Thread\]\(usage/groups.html#management-of-multiple-gpus-from-one-thread\)](#)
  - \* [\[Aggregated Operations \(2.2 and later\)\]\(usage/groups.html#aggregated-operations-2-2-and-later\)](#)
- \* [\[Nonblocking Group Operation\]\(usage/groups.html#nonblocking-group-operation\)](#)
- \* [\[Point-to-point communication\]\(usage/p2p.html\)](#)
- \* [\[Sendrecv\]\(usage/p2p.html#sendrecv\)](#)
- \* [\[One-to-all \(scatter\)\]\(usage/p2p.html#one-to-all-scatter\)](#)
- \* [\[All-to-one \(gather\)\]\(usage/p2p.html#all-to-one-gather\)](#)
- \* [\[All-to-all\]\(usage/p2p.html#all-to-all\)](#)
- \* [\[Neighbor exchange\]\(usage/p2p.html#neighbor-exchange\)](#)
- \* [\[Thread Safety\]\(usage/threadsafety.html\)](#)
- \* [\[In-place Operations\]\(usage/inplace.html\)](#)
- \* [\[Using NCCL with CUDA Graphs\]\(usage/cudagraph.html\)](#)
- \* [\[User Buffer Registration\]\(usage/bufferreg.html\)](#)
  - \* [\[NVLink Sharp Buffer Registration\]\(usage/bufferreg.html#nvlink-sharp-buffer-registration\)](#)
  - \* [\[IB Sharp Buffer Registration\]\(usage/bufferreg.html#ib-sharp-buffer-registration\)](#)
  - \* [\[General Buffer Registration\]\(usage/bufferreg.html#general-buffer-registration\)](#)
  - \* [\[Memory Allocator\]\(usage/bufferreg.html#memory-allocator\)](#)
- \* [\[NCCL API\]\(api.html\)](#)
  - \* [\[Communicator Creation and Management Functions\]\(api/comms.html\)](#)

- \* [\[ncclGetLastError\]\(api/comms.html#ncclgetlasterror\)](#)
- \* [\[ncclGetErrorString\]\(api/comms.html#ncclgeterrorstring\)](#)
- \* [\[ncclGetVersion\]\(api/comms.html#ncclgetversion\)](#)
- \* [\[ncclGetUniqueId\]\(api/comms.html#ncclgetuniqueid\)](#)
- \* [\[ncclCommInitRank\]\(api/comms.html#ncclcomminitrank\)](#)
- \* [\[ncclCommInitAll\]\(api/comms.html#ncclcomminitall\)](#)
- \* [\[ncclCommInitRankConfig\]\(api/comms.html#ncclcomminitrankconfig\)](#)
- \* [\[ncclCommInitRankScalable\]\(api/comms.html#ncclcomminitrankscalable\)](#)
- \* [\[ncclCommSplit\]\(api/comms.html#ncclcommsplit\)](#)
- \* [\[ncclCommFinalize\]\(api/comms.html#ncclcommfinalize\)](#)
- \* [\[ncclCommDestroy\]\(api/comms.html#ncclcommdestroy\)](#)
- \* [\[ncclCommAbort\]\(api/comms.html#ncclcommabort\)](#)
- \* [\[ncclCommGetAsyncError\]\(api/comms.html#ncclcommgetasynccerror\)](#)
- \* [\[ncclCommCount\]\(api/comms.html#ncclcommcount\)](#)
- \* [\[ncclCommCuDevice\]\(api/comms.html#ncclcommcudevice\)](#)
- \* [\[ncclCommUserRank\]\(api/comms.html#ncclcommuserrank\)](#)
- \* [\[ncclCommRegister\]\(api/comms.html#ncclcommregister\)](#)
- \* [\[ncclCommDeregister\]\(api/comms.html#ncclcommderegister\)](#)
- \* [\[ncclMemAlloc\]\(api/comms.html#ncclmemalloc\)](#)
- \* [\[ncclMemFree\]\(api/comms.html#ncclmemfree\)](#)
- \* [\[Collective Communication Functions\]\(api/colls.html\)](#)
  - \* [\[ncclAllReduce\]\(api/colls.html#ncclallreduce\)](#)
  - \* [\[ncclBroadcast\]\(api/colls.html#ncclbroadcast\)](#)
  - \* [\[ncclReduce\]\(api/colls.html#ncclreduce\)](#)
  - \* [\[ncclAllGather\]\(api/colls.html#ncclallgather\)](#)
  - \* [\[ncclReduceScatter\]\(api/colls.html#ncclreducescatter\)](#)
- \* [\[Group Calls\]\(api/group.html\)](#)

- \* [ncclGroupStart](api/group.html#ncclgroupstart)
- \* [ncclGroupEnd](api/group.html#ncclgroupend)
- \* [ncclGroupSimulateEnd](api/group.html#ncclgroupsimulateend)
- \* [Point To Point Communication Functions](api/p2p.html)
  - \* [ncclSend](api/p2p.html#ncclsend)
  - \* [ncclRecv](api/p2p.html#ncclrecv)
- \* [Types](api/types.html)
  - \* [ncclComm\_t](api/types.html#ncclcomm-t)
  - \* [ncclResult\_t](api/types.html#ncclresult-t)
  - \* [ncclDataType\_t](api/types.html#nccldatatype-t)
  - \* [ncclRedOp\_t](api/types.html#ncclredop-t)
  - \* [ncclScalarResidence\_t](api/types.html#ncclscalarresidence-t)
  - \* [ncclConfig\_t](api/types.html#ncclconfig-t)
  - \* [ncclSimInfo\_t](api/types.html#ncclsiminfo-t)
- \* [User Defined Reduction Operators](api/ops.html)
  - \* [ncclRedOpCreatePreMulSum](api/ops.html#ncclredopcreatepremulsum)
  - \* [ncclRedOpDestroy](api/ops.html#ncclredopdestroy)
- \* Migrating from NCCL 1 to NCCL 2
  - \* Initialization
  - \* Communication
  - \* Counts
  - \* In-place usage for AllGather and ReduceScatter
  - \* AllGather arguments order
  - \* Datatypes
  - \* Error codes
- \* [Examples](examples.html)

Examples](examples.html#communicator-creation-and-destruction-examples)

- \* [Example 1: Single Process, Single Thread, Multiple Devices](examples.html#example-1-single-process-single-thread-multiple-devices)

- \* [Example 2: One Device per Process or Thread](examples.html#example-2-one-device-per-process-or-thread)

- \* [Example 3: Multiple Devices per Thread](examples.html#example-3-multiple-devices-per-thread)

- \* [Example 4: Multiple communicators per device](examples.html#example-4-multiple-communicators-per-device)

- \* [Communication Examples](examples.html#communication-examples)

- \* [Example 1: One Device per Process or Thread](examples.html#example-1-one-device-per-process-or-thread)

- \* [Example 2: Multiple Devices per Thread](examples.html#example-2-multiple-devices-per-thread)

- \* [NCCL and MPI](mpi.html)

- \* [API](mpi.html#api)

- \* [Using multiple devices per process](mpi.html#using-multiple-devices-per-process)

- \* [ReduceScatter operation](mpi.html#reducescatter-operation)

- \* [Send and Receive counts](mpi.html#send-and-receive-counts)

- \* [Other collectives and point-to-point operations](mpi.html#other-collectives-and-point-to-point-operations)

- \* [In-place operations](mpi.html#in-place-operations)

- \* [Using NCCL within an MPI Program](mpi.html#using-nccl-within-an-mpi-program)

- \* [MPI Progress](mpi.html#mpi-progress)

- \* [Inter-GPU Communication with CUDA-aware MPI](mpi.html#inter-gpu-communication-with-cuda-aware-mpi)

- \* [Environment Variables](env.html)

- \* [System configuration](env.html#system-configuration)
- \* [NCCL\_SOCKET\_IFNAME](env.html#nccl-socket-iframe)
  - \* [Values accepted](env.html#values-accepted)
- \* [NCCL\_SOCKET\_FAMILY](env.html#nccl-socket-family)
  - \* [Values accepted](env.html#id2)
- \* [NCCL\_SOCKET\_RETRY\_CNT](env.html#nccl-socket-retry-cnt)
  - \* [Values accepted](env.html#id3)
- \* [NCCL\_SOCKET\_RETRY\_SLEEP\_MSEC](env.html#nccl-socket-retry-sleep-msec)
  - \* [Values accepted](env.html#id4)
- \* [NCCL\_SOCKET\_NTHREADS](env.html#nccl-socket-nthreads)
  - \* [Values accepted](env.html#id5)
- \* [NCCL\_NSOCKS\_PERTHREAD](env.html#nccl-nsocks-perthread)
  - \* [Values accepted](env.html#id6)
- \* [NCCL\_CROSS\_NIC](env.html#nccl-cross-nic)
  - \* [Values accepted](env.html#id7)
- \* [NCCL\_IB\_HCA](env.html#nccl-ib-hca)
  - \* [Values accepted](env.html#id8)
- \* [NCCL\_IB\_TIMEOUT](env.html#nccl-ib-timeout)
  - \* [Values accepted](env.html#id9)
- \* [NCCL\_IB\_RETRY\_CNT](env.html#nccl-ib-retry-cnt)
  - \* [Values accepted](env.html#id10)
- \* [NCCL\_IB\_GID\_INDEX](env.html#nccl-ib-gid-index)
  - \* [Values accepted](env.html#id11)
- \* [NCCL\_IB\_ADDR\_FAMILY](env.html#nccl-ib-addr-family)
  - \* [Values accepted](env.html#id12)
- \* [NCCL\_IB\_ADDR\_RANGE](env.html#nccl-ib-addr-range)
  - \* [Values accepted](env.html#id13)

\* [NCCL\_IB\_ROCE\_VERSION\_NUM](env.html#nccl-ib-roce-version-num)  
\* [Values accepted](env.html#id14)

\* [NCCL\_IB\_SL](env.html#nccl-ib-sl)  
\* [Values accepted](env.html#id15)

\* [NCCL\_IB\_TC](env.html#nccl-ib-tc)  
\* [Values accepted](env.html#id16)

\* [NCCL\_IB\_FIFO\_TC](env.html#nccl-ib-fifo-tc)  
\* [Values accepted](env.html#id17)

\* [NCCL\_IB\_RETURN\_ASYNC\_EVENTS](env.html#nccl-ib-return-async-events)  
\* [Values accepted](env.html#id18)

\* [NCCL\_OOB\_NET\_ENABLE](env.html#nccl-oob-net-enable)  
\* [Values accepted](env.html#id19)

\* [NCCL\_OOB\_NET\_IFNAME](env.html#nccl-oob-net-ifname)  
\* [Values accepted](env.html#id20)

\* [NCCL\_UID\_STAGGER\_THRESHOLD](env.html#nccl-uid-stagger-threshold)  
\* [Values accepted](env.html#id21)

\* [NCCL\_UID\_STAGGER\_RATE](env.html#nccl-uid-stagger-rate)  
\* [Values accepted](env.html#id22)

\* [NCCL\_NET](env.html#nccl-net)  
\* [Values accepted](env.html#id23)

\* [NCCL\_NET\_PLUGIN](env.html#nccl-net-plugin)  
\* [Values accepted](env.html#id24)

\* [NCCL\_TUNER\_PLUGIN](env.html#nccl-tuner-plugin)  
\* [Values accepted](env.html#id25)

\* [NCCL\_PROFILER\_PLUGIN](env.html#nccl-profiler-plugin)  
\* [Values accepted](env.html#id26)

\* [NCCL\_IGNORE\_CPU\_AFFINITY](env.html#nccl-ignore-cpu-affinity)

- \* [Values accepted](env.html#id27)
- \* [NCCL\_CONF\_FILE](env.html#nccl-conf-file)
  - \* [Values accepted](env.html#id28)
- \* [NCCL\_DEBUG](env.html#nccl-debug)
  - \* [Values accepted](env.html#id30)
- \* [NCCL\_DEBUG\_FILE](env.html#nccl-debug-file)
  - \* [Values accepted](env.html#id31)
- \* [NCCL\_DEBUG\_SUBSYS](env.html#nccl-debug-subsys)
  - \* [Values accepted](env.html#id32)
- \* [NCCL\_COLLNET\_ENABLE](env.html#nccl-collnet-enable)
  - \* [Value accepted](env.html#value-accepted)
- \* [NCCL\_COLLNET\_NODE\_THRESHOLD](env.html#nccl-collnet-node-threshold)
  - \* [Value accepted](env.html#id33)
- \* [NCCL\_TOPO\_FILE](env.html#nccl-topo-file)
  - \* [Value accepted](env.html#id34)
- \* [NCCL\_TOPO\_DUMP\_FILE](env.html#nccl-topo-dump-file)
  - \* [Value accepted](env.html#id35)
- \* [NCCL\_SET\_THREAD\_NAME](env.html#nccl-set-thread-name)
  - \* [Value accepted](env.html#id36)
- \* [Debugging](env.html#debugging)
- \* [NCCL\_P2P\_DISABLE](env.html#nccl-p2p-disable)
  - \* [Values accepted](env.html#id37)
- \* [NCCL\_P2P\_LEVEL](env.html#nccl-p2p-level)
  - \* [Values accepted](env.html#id38)
  - \* [Integer Values (Legacy)](env.html#integer-values-legacy)
- \* [NCCL\_P2P\_DIRECT\_DISABLE](env.html#nccl-p2p-direct-disable)
  - \* [Values accepted](env.html#id39)



\* [NCCL\_SHM\_DISABLE](env.html#nccl-shm-disable)

\* [Values accepted](env.html#id40)

\* [NCCL\_BUFFSIZE](env.html#nccl-buffersize)

\* [Values accepted](env.html#id41)

\* [NCCL\_NTHREADS](env.html#nccl-nthreads)

\* [Values accepted](env.html#id42)

\* [NCCL\_MAX\_NCHANNELS](env.html#nccl-max-nchannels)

\* [Values accepted](env.html#id43)

\* [NCCL\_MIN\_NCHANNELS](env.html#nccl-min-nchannels)

\* [Values accepted](env.html#id44)

\* [NCCL\_CHECKS\_DISABLE](env.html#nccl-checks-disable)

\* [Values accepted](env.html#id45)

\* [NCCL\_CHECK\_POINTERS](env.html#nccl-check-pointers)

\* [Values accepted](env.html#id46)

\* [NCCL\_LAUNCH\_MODE](env.html#nccl-launch-mode)

\* [Values accepted](env.html#id47)

\* [NCCL\_IB\_DISABLE](env.html#nccl-ib-disable)

\* [Values accepted](env.html#id48)

\* [NCCL\_IB\_AR\_THRESHOLD](env.html#nccl-ib-ar-threshold)

\* [Values accepted](env.html#id49)

\* [NCCL\_IB\_QPS\_PER\_CONNECTION](env.html#nccl-ib-qps-per-connection)

\* [Values accepted](env.html#id50)

\* [NCCL\_IB\_SPLIT\_DATA\_ON\_QPS](env.html#nccl-ib-split-data-on-qps)

\* [Values accepted](env.html#id51)

\* [NCCL\_IB\_CUDA\_SUPPORT](env.html#nccl-ib-cuda-support)

\* [Values accepted](env.html#id52)

\* [NCCL\_IB\_PCI\_RELAXED\_ORDERING](env.html#nccl-ib-pci-relaxed-ordering)

\* [Values accepted](env.html#id53)

\* [NCCL\_IB\_ADAPTIVE\_ROUTING](env.html#nccl-ib-adaptive-routing)

\* [Values accepted](env.html#id54)

\* [NCCL\_IB\_ECE\_ENABLE](env.html#nccl-ib-ece-enable)

\* [Values accepted](env.html#id55)

\* [NCCL\_MEM\_SYNC\_DOMAIN](env.html#nccl-mem-sync-domain)

\* [Values accepted](env.html#id56)

\* [NCCL\_CUMEM\_ENABLE](env.html#nccl-cumem-enable)

\* [Values accepted](env.html#id57)

\* [NCCL\_CUMEM\_HOST\_ENABLE](env.html#nccl-cumem-host-enable)

\* [Values accepted](env.html#id58)

\* [NCCL\_NET\_GDR\_LEVEL] (formerly

NCCL\_IB\_GDR\_LEVEL))(env.html#nccl-net-gdr-level-formerly-nccl-ib-gdr-level)

\* [Values accepted](env.html#id59)

\* [Integer Values (Legacy)](env.html#id60)

\* [NCCL\_NET\_GDR\_READ](env.html#nccl-net-gdr-read)

\* [Values accepted](env.html#id61)

\* [NCCL\_NET\_SHARED\_BUFFERS](env.html#nccl-net-shared-buffers)

\* [Value accepted](env.html#id62)

\* [NCCL\_NET\_SHARED\_COMMS](env.html#nccl-net-shared-comms)

\* [Value accepted](env.html#id63)

\* [NCCL\_SINGLE\_RING\_THRESHOLD](env.html#nccl-single-ring-threshold)

\* [Values accepted](env.html#id64)

\* [NCCL\_LL\_THRESHOLD](env.html#nccl-ll-threshold)

\* [Values accepted](env.html#id65)

\* [NCCL\_TREE\_THRESHOLD](env.html#nccl-tree-threshold)

\* [Values accepted](env.html#id66)

\* [NCCL\_ALGO](env.html#nccl-algo)

\* [Values accepted](env.html#id67)

\* [NCCL\_PROTO](env.html#nccl-proto)

\* [Values accepted](env.html#id68)

\* [NCCL\_NVX\_DISABLE](env.html#nccl-nvx-disable)

\* [Value accepted](env.html#id69)

\* [NCCL\_P2P\_DISABLE](env.html#nccl-p2p-disable)

\* [Value accepted](env.html#id70)

\* [NCCL\_P2P\_P2P\_LEVEL](env.html#nccl-p2p-p2p-level)

\* [Value accepted](env.html#id71)

\* [NCCL\_RUNTIME\_CONNECT](env.html#nccl-runtime-connect)

\* [Value accepted](env.html#id72)

\* [NCCL\_GRAPH\_REGISTER](env.html#nccl-graph-register)

\* [Value accepted](env.html#id74)

\* [NCCL\_LOCAL\_REGISTER](env.html#nccl-local-register)

\* [Value accepted](env.html#id75)

\* [NCCL\_LEGACY\_CUDA\_REGISTER](env.html#nccl-legacy-cuda-register)

\* [Value accepted](env.html#id76)

\* [NCCL\_SET\_STACK\_SIZE](env.html#nccl-set-stack-size)

\* [Value accepted](env.html#id77)

\* [NCCL\_GRAPH\_MIXING\_SUPPORT](env.html#nccl-graph-mixing-support)

\* [Value accepted](env.html#id79)

\* [NCCL\_DMABUF\_ENABLE](env.html#nccl-dmabuf-enable)

\* [Value accepted](env.html#id80)

\* [NCCL\_P2P\_NET\_CHUNKSIZE](env.html#nccl-p2p-net-chunksize)

\* [Values accepted](env.html#id81)

\* [NCCL\_P2P\_LL\_THRESHOLD](env.html#nccl-p2p-ll-threshold)

- \* [Values accepted](env.html#id82)
- \* [NCCL\_ALLOC\_P2P\_NET\_LL\_BUFFERS](env.html#nccl-alloc-p2p-net-ll-buffers)
  - \* [Values accepted](env.html#id83)
- \* [NCCL\_COMM\_BLOCKING](env.html#nccl-comm-blocking)
  - \* [Values accepted](env.html#id84)
- \* [NCCL\_CGA\_CLUSTER\_SIZE](env.html#nccl-cga-cluster-size)
  - \* [Values accepted](env.html#id85)
- \* [NCCL\_MAX\_CTAS](env.html#nccl-max-ctas)
  - \* [Values accepted](env.html#id86)
- \* [NCCL\_MIN\_CTAS](env.html#nccl-min-ctas)
  - \* [Values accepted](env.html#id87)
- \* [NCCL\_NVLS\_ENABLE](env.html#nccl-nvls-enable)
  - \* [Values accepted](env.html#id88)
- \* [NCCL\_IB\_MERGE\_NICS](env.html#nccl-ib-merge-nics)
  - \* [Values accepted](env.html#id89)
- \* [NCCL\_MNNVL\_ENABLE](env.html#nccl-mnnvl-enable)
  - \* [Values accepted](env.html#id90)
- \* [NCCL\_RAS\_ENABLE](env.html#nccl-ras-enable)
  - \* [Values accepted](env.html#id91)
- \* [NCCL\_RAS\_ADDR](env.html#nccl-ras-addr)
  - \* [Values accepted](env.html#id92)
- \* [NCCL\_RAS\_TIMEOUT\_FACTOR](env.html#nccl-ras-timeout-factor)
  - \* [Values accepted](env.html#id93)
- \* [Troubleshooting](troubleshooting.html)
- \* [Errors](troubleshooting.html#errors)
- \* [RAS](troubleshooting.html#ras)
- \* [RAS](troubleshooting/ras.html)

- \* [\[Principle of Operation\]\(troubleshooting/ras.html#principle-of-operation\)](#)
- \* [\[RAS Queries\]\(troubleshooting/ras.html#ras-queries\)](#)
- \* [\[Sample Output\]\(troubleshooting/ras.html#sample-output\)](#)
- \* [\[GPU Direct\]\(troubleshooting.html#gpu-direct\)](#)
- \* [\[GPU-to-GPU communication\]\(troubleshooting.html#gpu-to-gpu-communication\)](#)
- \* [\[GPU-to-NIC communication\]\(troubleshooting.html#gpu-to-nic-communication\)](#)
- \* [\[PCI Access Control Services \(ACS\)\]\(troubleshooting.html#pci-access-control-services-ac\)](#)
- \* [\[Topology detection\]\(troubleshooting.html#topology-detection\)](#)
- \* [\[Shared memory\]\(troubleshooting.html#shared-memory\)](#)
- \* [\[Docker\]\(troubleshooting.html#docker\)](#)
- \* [\[Systemd\]\(troubleshooting.html#systemd\)](#)
- \* [\[Networking issues\]\(troubleshooting.html#networking-issues\)](#)
- \* [\[IP Network Interfaces\]\(troubleshooting.html#ip-network-interfaces\)](#)
- \* [\[IP Ports\]\(troubleshooting.html#ip-ports\)](#)
- \* [\[InfiniBand\]\(troubleshooting.html#infiniband\)](#)
- \* [\[RDMA over Converged Ethernet \(RoCE\)\]\(troubleshooting.html#rdma-over-converged-ethernet-roce\)](#)

[\\_\\_\[NCCL\]\(index.html\)](#)

\* [\[Docs\]\(index.html\)](#) »

\* [Migrating from NCCL 1 to NCCL 2](#)

\* [\[ View page source\]\(\\_sources/nccl1.rst.txt\)](#)

\* \* \*

# Migrating from NCCL 1 to NCCL 2

If you are using NCCL 1.x and want to move to NCCL 2.x, be aware that the APIs have changed slightly. NCCL 2.x supports all of the collectives that NCCL 1.x supports, but with slight modifications to the API.

In addition, NCCL 2.x also requires the usage of the "Group API" when a single thread manages NCCL calls for multiple GPUs.

The following list summarizes the changes that may be required in usage of NCCL API when using an application that has a single thread that manages NCCL calls for multiple GPUs, and is ported from NCCL 1.x to 2.x:

## ## Initialization¶

In versions 1.x, NCCL had to be initialized using `ncclCommInitAll` at a single thread or having one thread per GPU concurrently call `ncclCommInitRank`. NCCL 2.x retains these two modes of initialization. It adds a new mode with the Group API where `ncclCommInitRank` can be called in a loop, like a communication call, as shown below. The loop has to be guarded by the Group start and end API.

```
ncclGroupStart();  
for (int i=0; i<ngpus; i++) {  
    cudaSetDevice(i);  
    ncclCommInitRank(comms+i, ngpus, id, i);  
}
```

```
}  
  
ncclGroupEnd();
```

## ## Communication ¶

In NCCL 2.x, the collective operation can be initiated for different devices by making calls in a loop, on a single thread. This is similar to the usage in NCCL 1.x. However, this loop has to be guarded by the Group API in 2.x. Unlike in 1.x, the application does not have to select the relevant CUDA device before making the communication API call. NCCL runtime internally selects the device associated with the NCCL communicator handle. For example:

```
ncclGroupStart();  
  
for (int i=0; i<nLocalDevs; i++) {  
    ncclAllReduce(..., comm[i], stream[i]);  
}  
  
ncclGroupEnd();
```

When using only one device per thread or one device per process, the general usage of the API remains unchanged from NCCL 1.x to 2.x. The usage of the group API is not required in this case.

## ## Counts ¶

Counts provided as arguments are now of type `size_t` instead of integer.

## In-place usage for AllGather and ReduceScatter¶

For more information, see [“In-place Operations”](#).

## AllGather arguments order¶

The AllGather function had its arguments reordered. The prototype changed from:

```
ncclResult_t ncclAllGather(const void* sendbuff, int count, ncclDataType_t datatype,  
    void* recvbuff, ncclComm_t comm, cudaStream_t stream);
```

to:

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

The `recvbuff` argument has been moved after `sendbuff` to be consistent with all



the other operations.

## ## Datatypes¶

New datatypes have been added in NCCL 2.x. The ones present in NCCL 1.x did not change and are still usable in NCCL 2.x.

## ## Error codes¶

Error codes have been merged into the `ncclInvalidArgument` category and have been simplified. A new `ncclInvalidUsage` code has been created to cover new programming errors.

[Next ]([examples.html](#) "Examples") [ Previous]([api/ops.html](#) "User Defined Reduction Operators")

\* \* \*

(C) Copyright 2020, NVIDIA Corporation

Built with [Sphinx](<http://sphinx-doc.org/>) using a [theme]([https://github.com/rtfd/sphinx\\_rtd\\_theme](https://github.com/rtfd/sphinx_rtd_theme)) provided by [Read the Docs](<https://readthedocs.org>).