

# Intel<sup>®</sup> Machine Learning Scaling Library

## **Developer Guide**

---



## Legal Information

---

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

The products and services described may contain defects or errors known as errata which may cause deviations from published specifications. Current characterized errata are available on request.

Intel, the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.

\* Other names and brands may be claimed as the property of others

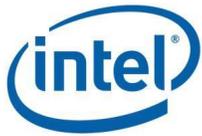
© Intel Corporation.

Optimization Notice
Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.  Notice revision #20110804

# Contents

---

- Legal Information .....2**
- 1. Introduction.....4**
- 2. Using Intel® Machine Learning Scaling Library .....5**
  - 2.1. Prerequisites..... 5
  - 2.2. Generic Workflow..... 5
  - 2.3. Launching Sample Application..... 6
    - 2.3.1. Launching the Sample ..... 6
    - 2.3.2. Sample Description ..... 7
  - 2.4. Statistics Collection..... 7
- 3. Environment Variables.....9**



# 1. Introduction

---

Intel® Machine Learning Scaling Library is a library providing an efficient implementation of communication patterns used in deep learning. Some of the library features include:

- Built on top of MPI, allows for use of other communication libraries
- Optimized to drive scalability of communication patterns
- Works on various interconnects: Intel® Omni-Path Architecture, InfiniBand\*, and Ethernet
- Common API to support Deep Learning frameworks (Caffe\*, Theano\*, Torch\*, etc.)

The Intel® MLSL package comprises the Intel MLSL Software Development Kit (SDK) and the Intel® MPI Library Runtime components.

This document provides usage instructions for Intel MLSL, and its configuration reference. For installation instructions, system requirements, and other information, refer to the README file supplied with the library.



## 2. Using Intel® Machine Learning Scaling Library

---

### 2.1. Prerequisites

Before you start using the Intel® ML SL, make sure to set up the library environment. Use the command:

```
$ source <install_dir>/intel64/bin/mlslvars.sh
```

This sets up the environment both for the C/C++ and Python\* bindings.

Here and below, <install\_dir> is the Intel ML SL installation directory, which is /opt/intel/mlsl by default.

### 2.2. Generic Workflow

Below is a generic flow of using the Intel® ML SL in C++:

1. Initialize the library:

```
Environment::GetEnv().Init(&argc, &argv);
```

2. Create a Session and a Distribution objects:

```
Session *s = Environment::GetEnv().CreateSession();
Distribution *d = Environment::GetEnv().CreateDistribution(<args>);
```

3. Set the global mini-batch size (equal to the sum of local batch sizes):

```
s->SetGlobalMinibatchSize(<args>);
```

4. For each layer, create an Operation object, as follows:

- a. Create an OperationRegInfo object:

```
OperationRegInfo *ori = s->CreateOperationRegInfo(<args>);
```

- b. Depending on the type of parallelism, add input/output activation shapes or shapes of parameters (weights or biases) to the OperationRegInfo object:

```
ori->AddInput(<args>); // to add input activation shape
ori->AddOutput(<args>); // to add output activation shape
ori->AddParameterSet(<args>); // to add weight shape
ori->AddParameterSet(<args>); // to add bias shape
```

- c. Create an Operation object using the OperationRegInfo and Distribution objects:

```
size_t opIdx = s->AddOperation(ori, d);
Operation *op = s->GetOperation(opIdx);
```

5. Invoke the Commit() method of the Session object:



```
s->Commit();
```

6. In each forward propagation iteration for each layer, use `Activation` objects of the `Operation` object to exchange data for model parallelism:

```
op->GetInput(<activation_index>)->WaitComm();  
op->GetOutput(<activation_index>)->StartComm();
```

7. In each backward propagation iteration for each layer:
  - a. Use `Activation` objects of the `Operation` object to exchange gradients with respect to data for model parallelism:

```
op->GetOutput(<activation_index>)->WaitComm();  
op->GetInput(<activation_index>)->StartComm();
```

- b. Use `ParameterSet` objects of the `Operation` object to exchange gradients with respect to parameters (weights or biases) for data parallelism:

```
op->GetParameterSet(<parameter_index>)->StartGradientComm();  
op->GetParameterSet(<parameter_index>)->WaitGradientComm();
```

- c. In the case of distributed parameter update, use `ParameterSet` objects of the `Operation` object to exchange parameter increments for data parallelism:

```
op->GetParameterSet(<parameter_index>)->StartIncrementComm();  
op->GetParameterSet(<parameter_index>)->WaitIncrementComm();
```

8. Delete the `Session` and `Distribution` objects:

```
Environment::GetEnv().DeleteSession(s);  
Environment::GetEnv().DeleteDistribution(d);
```

9. Finalize the library:

```
Environment::GetEnv().Finalize();
```

The workflow described above is implemented in a sample application `m1sl_test.cpp` distributed with Intel® ML SL. There is also a similar sample Python script `m1sl_test.py`, implementing the same workflow.

You can use this sample as a reference when applying Intel ML SL for your framework (for C++ or Python, respectively). See the section below for description and instructions on using the sample application.

For detailed API description, refer to the *Intel ML SL API Reference* at `<install_dir>/doc/API_Reference.htm`.

## 2.3. Launching Sample Application

Intel® ML SL supplies a sample application, `m1sl_test.cpp` or `m1sl_test.py`, which demonstrates the usage of Intel ML SL API.

### 2.3.1. Launching the Sample

1. For the C++ sample, build `m1sl_test.cpp`:



```
$ cd <install_dir>/test
$ icpc -O2 -I<install_dir>/intel64/include/ -L<install_dir>/intel64/lib
-lmlsl -lmpi -ldl -lrt -lpthread -o msl_test msl_test.cpp
```

2. Launch the `msl_test` binary or the `msl_test.py` script with `mpirun` on the desired number of nodes ( $N$ ).

The application takes one argument `num_groups`, which will define the type of parallelism, based on the following logic:

- `num_groups = 1` – data parallelism, for example:

```
$ mpirun -n 8 -ppn 1 ./msl_test[.py] 1
```

- `num_groups = N` – model parallelism, for example:

```
$ mpirun -n 8 -ppn 1 ./msl_test[.py] 8
```

- `num_groups > 1` and `num_groups < N` – hybrid parallelism, for example:

```
$ mpirun -n 8 -ppn 1 ./msl_test[.py] 2
```

## 2.3.2. Sample Description

The application is set up to run a test for two layers. It sets output on the 1<sup>st</sup> layer and checks input for the 2<sup>nd</sup> layer in an `fprop()` call. Similarly, for the `bprop()` call, it sets a gradient with respect to input for the 2<sup>nd</sup> layer and checks the gradient with respect to output for the 1<sup>st</sup> layer. For weights, it sets gradients with respect to weights, checks the gradients accumulation, modifies weights in a `wtinc()` call, and then verifies the expected values in an `fprop()` call for both layers.

The application prints parameters for input and output feature maps and weights, whether the communication is required, and what type of communication is required. The test status is printed as PASSED or FAILED. You can `grep` for FAILED to see if a test failed.

## 2.4. Statistics Collection

Intel® MSL statistics allow you to monitor the time spent by operations in Intel MSL during the computation and communication phases.

You can start and stop monitoring operations using the `MSL::Statistics` API.

You can also fetch, print and reset Intel MSL statistics data to understand whether your neural network is computation or communication bound.

Intel MSL provides the following statistics:

- The total time spent by all operations in the computation phase
- The total time spent by all operations in the communication phase
- The total time expected to spend in the communication phase in an isolated environment (that is, time spent in blocking communication calls with no computation involved between communication calls)
- The total communication size of all operations
- The time spent per operation in the computation phase
- The time spent per operation in the communication phase



- The expected time spent per operation in the communication phase in an isolated environment
- The communication size for each operation

By analyzing communication time collected in an isolated environment you can understand the impact of the computation phase on the communication phase.

To enable or disable statistics collection, use the `MLSL_STATS` environment variable.



## 3. Environment Variables

### MLSL\_ROOT

#### Syntax

```
MLSL_ROOT=<path>
```

#### Arguments

<code>&lt;path&gt;</code>	Installation directory of the Intel® MLSL.
---------------------------	--

#### Description

Set this environment variable to specify the installation directory of the Intel® MLSL.

### MLSL\_NUM\_SERVERS

#### Syntax

```
MLSL_NUM_SERVERS=<nserver>
```

#### Arguments

<code>&lt;nserver&gt;</code>	The number of servers per node.
<code>&gt;= 0</code>	The default value is 4. The maximum value is 16.

#### Description

Set this environment variable to define the number of endpoint servers per node.

**NOTE:** Each server is a separate process, which uses CPU resources. Take that into account when setting the `OMP_NUM_THREADS` variable for OpenMP\*. The recommended value for `OMP_NUM_THREADS` is `max_cores - num_servers`, where `max_cores` is the maximum number of cores, and `num_servers` is the number of endpoint servers per node.

### MLSL\_SERVER\_AFFINITY

#### Syntax

```
MLSL_SERVER_AFFINITY=<proclist>
```

#### Arguments

<code>&lt;proclist&gt;</code>	A comma-separated list of logical core numbers. The server with the <i>i</i> -th index is pinned to the <i>i</i> -th core in the list. The number should not exceed the number of cores on the node.
-------------------------------	--



n-1, n-2, n-3, n-4, ...	This is the default value – servers are pinned to cores in the reversed order. n – the number of available cores
-------------------------	---

### Description

Set this environment variable to define the processor core affinity for endpoint servers for best performance.

**NOTE:** The recommended values can be retrieved by querying `/proc/interrupts`:

- For the Intel® Omni-Path Fabric (Intel® OP Fabric) you are recommended to map servers to the cores handling the Intel OP Fabric send direct memory access (SDMA) interrupts.
- For Ethernet you are recommended to map servers to the cores handling the Tx/Rx interrupts.

## MLSL\_STATS

### Syntax

MLSL\_STATS=<arg>

### Arguments

<arg>	Binary indicator
1	Enable statistics collection.
0	Disable statistics collection. This is the default value.

### Description

Set this environment variable to enable statistics collection. See [Statistics Collection](#) for details.

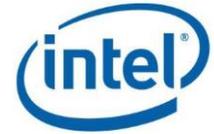
## MLSL\_LOG\_LEVEL

### Syntax

MLSL\_LOG\_LEVEL=<level>

### Arguments

<level>	Logging information level
0	The <i>error</i> level. Prints out critical errors that lead to application termination. This is the default value.



1	The <i>informational</i> level. Prints out informational messages about the application progress.
2	The <i>debug</i> level. Prints out detailed informational messages that are most useful to debug an application.
3	The <i>trace</i> level. Prints out more detailed informational messages than in the <i>debug</i> level.

**Description**

Set this environment variable to print logging information about the application. Higher levels include information logged in the lower levels.

**MLSL\_SERVER\_CREATION\_TYPE**

**Syntax**

MLSL\_SERVER\_CREATION\_TYPE=<arg>

**Arguments**

<arg>	Launch method index
0	Use MPI spawn API
1	Use <code>mpirun</code>

**Description**

Set this environment variable to specify a method for launching endpoint servers.

**MLSL\_HOSTNAME\_TYPE**

**Syntax**

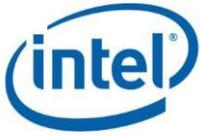
MLSL\_HOSTNAME\_TYPE=<arg>

**Arguments**

<arg>	Index of host name retrieval method
0	Use <code>MPI_Get_processor_name</code>
1	Use <code>gethostname</code> and <code>getaddrinfo</code>
2	Use the interface IP address

**Description**

Set this environment variable to specify a method of retrieving host names of endpoint servers.



If using the value 2, make sure to also specify the interface using one of the variables below:

- `MLSL_IFACE_NAME` – select an interface by prefix name
- `MLSL_IFACE_IDX` – select an interface by index

## MLSL\_HOSTNAME

### Syntax

`MLSL_HOSTNAME=<name>`

### Arguments

<code>&lt;name&gt;</code>	The hostname.
---------------------------	---------------

### Description

Set this environment variable to specify the hostname to be used for launching endpoint servers. The variable must be set for each Intel MLSL process.

For example, to set the variable through Intel® MPI Library `-gtool` option:

```
$ mpirun ... -gtool "env MLSL_HOSTNAME=name1:0; env MLSL_HOSTNAME=name2:1"
```

## MLSL\_MSG\_PRIORITY

### Syntax

`MLSL_MSG_PRIORITY=<arg>`

### Arguments

<code>&lt;arg&gt;</code>	Binary indicator
1	Enable reversed order of communication calls execution.
0	Disable reversed order of communication calls execution. This is the default value.

### Description

Set this environment variable to enable executing communication calls in the order opposite to the order they are started.

When performing backward propagation, the gradient reduction communication generally starts from the last layers of the network topology finishing at the first ones. With this experimental feature enabled, the communication for first layers has a higher priority, which allows layer one to minimize the time spent in SGD and start a forward pass for the next iteration as soon as possible.

**NOTE:** This feature is experimental and has the following limitations: it can be enabled only for a specific implementation of the `Allreduce` algorithm and only when the number of communicating processes is a power of two.



## MLSL\_MSG\_PRIORITY\_THRESHOLD

### Syntax

MLSL\_MSG\_PRIORITY\_THRESHOLD=<arg>

### Arguments

<arg>	Message size in bytes
> 0	The default value is 10000

### Description

Set this environment variable to define the lower threshold of message sizes where the experimental priority feature is enabled. The variable only takes effect when `MLSL_MSG_PRIORITY` is set.