登临 Hamming™ V1

NPI 工具集使用说明

DL-DG/SW-067A-01

2025-03-03

商标和许可

和其它苏州登临科技有限公司的其它登临科技的图标为苏州登临科技有限公司的商标。本手册中提及的所有其他商标均为其各自所有者的财产。

通知

所购买的产品、服务和特性由苏州登临科技有限公司与客户签订的合同规定。本文件中描述的所有或部分产品、服务和特性可能不在采购范围或使用范围内。除非合同中另有规定，本文件中的所有声明、信息和建议均按“原样”提供，无任何明示或暗示的保证或陈述。
本手册中的信息如有更改，恕不另行通知。本文件在编制过程中已尽一切努力确保内容的准确性，本文件中的所有声明、信息和建议不构成任何明示或暗示的保证。

苏州登临科技有限公司

苏州工业园区扬富路11号南岸新地一期商务楼栋5号1101室，江苏，中国

http://www.denglin.ai

Email：support@denglin.ai

更新历史

版本	更新描述
01	第一次发布。

目录

1 简介

NPI工具集是登临研发的、在新产品导入过程中用来评估硬件环境正确性及硬件关键指标的一组测试工具。NPI工具由以下工具组成：

- **bandwidthTest**：该工具是本地（Local）带宽测试程序。能够测量设备到设备的复制带宽、主机到设备的可分页和页锁定内存的复制带宽，以及设备到主机的可分页和页锁定内存的复制带宽。
- **tuFLOPS**：该工具可以用来评估登临GPU TU的峰值算力。

> **说明**：
>
> 每个NPI工具的使用方法请参考下文章节 2 NPI使用说明，或参考NPI工具集目录下各工具的README.md文件。

2 NPI 使用说明

2.1 tuFLOPS

2.1.1 简介

登临tuFLOPS工具用来计算登临GPU TU的峰值算力。

2.1.2 使用方法

首先激活DLI V1 SDK环境：`source <SDK_PATH>/env.sh`，然后在命令行中输入带参数的./tuFLOPS 命令。

.tuFLOPS [OPTION]

参数含义如下：

- -d 指定数据类型
- -n 指定循环次数
- -D 用于指定设备ID （多卡机器上选择设备，默认值为0）

以上参数可以使用缺省值。

命令执行如下图所示：

```
tuFLOPS Usage:
./tuFLOPS -d fp16 -n 10000
./tuFLOPS -d fp16 -n 10000 -D 0
dtype: int4/uint4/int8/uint8/fp16/fp32
Usage: ./tuFLOPS [OPTIONS]

Options:
  -h,--help                   Print this help message and exit
  -D INT                      device_id
  -n INT                      num_runs
  -d TEXT                     data_type
```

2.1.3 输出结果

输出结果如下图，会打印数据类型、conv shape、循环次数、运行时最大功率、运行时间和TU算力等。

```
Device: Goldwasser L256, id: 0, clusters: 4
    Data   Type    : int8
    Input  Shape   : 1, 64, 64, 512
    Filter Shape   : 512, 1, 1, 512
    Repeat Times   : 10000
    Rated  Power   : 55W
    Exec   Power   : 29W
    Total  Time    : 0.0834743 s
    TU     Perf    : 257.26 T
done.
```

2.2 bandwidthTest

2.2.1 简介

登临bandwidthTest工具是基于CUDA的带宽测试程序。用来测量以下内容：

- 设备到设备（Device to Device）的数据传输带宽。
- 主机到设备（Host to Device）的可分页和页锁定内存的数据传输带宽。
- 设备到主机（Device to Host）的可分页和页锁定内存的数据传输带宽。

其中主机到设备和设备到主机的数据传输带宽主要用来衡量PCIE的实测带宽，设备到设备的数据传输带宽用来衡量DDR的实测带宽。

2.2.2 使用方法

./bandwidthTest [OPTION]

参数含义如下：

```
Options:
--help  Display this help menu
--csv   Print results as a CSV
--device=[deviceno]Specify the device device to be used
       all - compute cumulative bandwidth on all the devices
       0,1,2,...,n - Specify any particular device to be used
--memory=[MEMMODE] Specify which memory mode to use
       pageable - pageable memory
       pinned  - non-pageable system memory
--mode=[MODE] Specify the mode to use
       quick - performs a quick measurement
       range - measures a user-specified range of values
       shmoo - performs an intense shmoo of a large range of values
--htod    Measure host to device transfers
--dtoh    Measure device to host transfers
--dtod    Measure device to device transfers
--dtod_sm Measure device to device transfers using kernel mode for testing
--wc      Allocate pinned memory as write-combined
--cputiming      Force CPU-based timing always Range mode options
--start=[SIZE]     Starting transfer size in bytes
--end=[SIZE]      Ending transfer size in bytes
--increment=[SIZE] Increment size in bytes
```

```
[CUDA Bandwidth Test] - Starting...
Running on...

 Device 0: Goldwasser L256
 Quick Mode

 Host to Device Bandwidth, 1 Device(s)
 PINNED Memory Transfers
   Transfer Size (Bytes)        Bandwidth(GB/s)
   32000000                     10.3

 Device to Host Bandwidth, 1 Device(s)
 PINNED Memory Transfers
   Transfer Size (Bytes)        Bandwidth(GB/s)
   32000000                     11.6

 Device to Device Bandwidth, 1 Device(s)
 PINNED Memory Transfers
   Transfer Size (Bytes)        Bandwidth(GB/s)
   32000000                     72.0

Result = PASS
```