NSIGHT COMPUTE

v2019.5.0 | November 2019

Customization Guide
TABLE OF CONTENTS

Chapter 1. Introduction........................................................................................................1
Chapter 2. Sections.............................................................................................................2
  2.1. Section Files........................................................................................................2
  2.2. Section Definition...............................................................................................5
  2.3. Metric Options...................................................................................................5
  2.4. Missing Sections...............................................................................................5
Chapter 3. Rule System....................................................................................................7
  3.1. Writing Rules.....................................................................................................7
  3.2. Integration.........................................................................................................7
  3.3. Rule System Architecture................................................................................8
  3.4. NvRules API...................................................................................................9
  3.5. Rule File API...................................................................................................9
  3.6. Rule Examples................................................................................................10
Chapter 4. Source Counters...........................................................................................11
Chapter 5. Report File Format........................................................................................13
  5.1. Version 7 Format............................................................................................13
LIST OF TABLES

Table 1  Top-level report file format ................................................................. 13
Table 2  Per-Block report file format ................................................................. 13
Table 3  Block payload report file format .......................................................... 14
Chapter 1. INTRODUCTION

The goal of NVIDIA Nsight Compute is to design a profiling tool that can be easily extended and customized by expert users. While we provide useful defaults, this allows adapting the reports to a specific use case or to design new ways to investigate collected data. All the following is data driven and does not require the tools to be recompiled.

While working with section files or rules files it is recommended to open the Sections/Rules tool window from the Profile menu item. This tool window lists all sections and rules that were loaded. Rules are grouped as children of their associated section or grouped in the [Independent Rules] entry. For files that failed to load, the table shows the error message. Use the Reload button to reload rule files from disk.
Chapter 2. SECTIONS

The Details page consists of sections that focus on a specific part of the kernel analysis each. Every section is defined by a corresponding section file that specifies the data to be collected as well as the visualization used in the UI to present this data. Simply modify a section file to add or modify what is collected.

2.1. Section Files

By default, the section files are stored in the sections sub-folder of the NVIDIA Nsight Compute install directory. Each section is defined in a separate file with the .section file extension. Section files are loaded automatically at the time the UI connects to a target application or the command line profiler is launched. That way, any changes to section files become immediately available in the next profile run.

A section file is a text representation of a Google Protocol Buffer message. The full definition of all available fields of a section message is given in Section Definition. In short, each section consists of a unique Identifier (no spaces allowed), a Display Name, an optional Order value (for sorting the sections in the Details page), an optional Description providing guidance to the user, an optional header table, and an optional body with additional UI elements. A small example of a very simple section is:

```
Identifier: "SampleSection"
DisplayName: "Sample Section"
Description: "This sample section shows information on active warps and cycles."
Header {
    Metrics {
        Label: "Active Warps"
        Name: "smsp__active_warps_avg"
    }
    Metrics {
        Label: "Active Cycles"
        Name: "smsp__active_cycles_avg"
    }
}
```

On data collection, this section will cause the two PerfWorks metrics `smsp__active_warps_avg` and `smsp__active_cycles_avg` to be collected.
More advanced elements can be used in the body of a section. Currently, NVIDIA Nsight Compute supports tables and various bar charts. The following example shows how to use these in a slightly more complex example. The usage of regexes is allowed in tables and charts in the section Body only and follows the format regex: followed by the actual regex to match PerfWorks metric names.

The supported list of metrics that can be used in sections can be queried using NVIDIA Nsight Compute CLI with option --query-metrics. Each of these metrics can be used in any section and will be automatically be collected if they appear in any enabled section. Look at all the shipping sections to see how they are implemented.

<table>
<thead>
<tr>
<th>Sample Section</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Active Warps</td>
<td>15,990,870.78</td>
<td>Active Cycles</td>
</tr>
</tbody>
</table>
Identifier: "SampleSection"
DisplayName: "Sample Section"
Description: "This sample section shows various metrics."
Header {
  Metrics {
    Label: "Active Warps"
    Name: "smsp__active_warps_avg"
  }
  Metrics {
    Label: "Active Cycles"
    Name: "smsp__active_cycles_avg"
  }
}
Body {
  Items {
    Table {
      Label: "Example Table"
      Rows: 2
      Columns: 1
      Metrics {
        Label: "Avg. Issued Instructions Per Scheduler"
        Name: "smsp__inst_issued_avg"
      }
      Metrics {
        Label: "Avg. Executed Instructions Per Scheduler"
        Name: "smsp__inst_executed_avg"
      }
    }
  }
  Items {
    Table {
      Label: "Metrics Table"
      Columns: 2
      Order: ColumnMajor
      Metrics {
        Name: "regex:.\+__elapsed_cycles_sum"
      }
    }
  }
  Items {
    BarChart {
      Label: "Metrics Chart"
      CategoryAxis {
        Label: "Units"
      }
      ValueAxis {
        Label: "Cycles"
      }
      Metrics {
        Name: "regex:.\+__elapsed_cycles_sum"
      }
    }
  }
}
2.2. Section Definition

Protocol buffer definitions are in the NVIDIA Nsight Compute installation directory under `extras/FileFormat`.

To see the list of available PerfWorks metrics for any device or chip, use the `--query-metrics` option of the NVIDIA Nsight Compute CLI.

2.3. Metric Options

Sections allow the user to specify alternative options for metrics that have a different metric name on different GPU architectures. Metric options use a min-arch/max-arch range filter, replacing the base metric with the first metric option for which the current GPU architecture matches the filter. While not strictly enforced, options for a base metric are expected to share the same meaning and subsequently unit, etc., with the base metric. In addition to its alternatives, the base metric can be filtered by the same criteria (currently min/max architecture). This is useful for metrics that are only available for certain architectures.

2.4. Missing Sections

If new or updated section files are not used by NVIDIA Nsight Compute, it is most commonly one of two reasons:

**The file is not found:** Section files must have the `.section` extension. They must also be on the section search path. The default search path is the `sections` directory within the installation directory. In NVIDIA Nsight Compute CLI, the search paths can be overwritten using the `--section-folder` and `--section-folder-recursive`
options. In NVIDIA Nsight Compute, the search path can be configured in the Profile options.

**Syntax errors:** If the file is found but has syntax errors, it will not be available for metric collection. However, error messages are reported for easier debugging. In NVIDIA Nsight Compute CLI, use the **--list-sections** option to get a list of error messages, if any. In NVIDIA Nsight Compute, error messages are reported in the Sections/Rules Info tool window.
Chapter 3.
RULE SYSTEM

NVIDIA Nsight Compute features a new Python-based rule system. It is designed as the successor to the *Expert System* (un)guided analysis in NVIDIA Visual Profiler, but meant to be more flexible and more easily extensible to different use cases and APIs.

### 3.1. Writing Rules

To create a new rule, you need to create a new text file with the extension `.py` and place it at some location that is detectable by the tool (see Nsight Compute Integration on how to specify the search path for rules). At a minimum, the rule file must implement two functions, `get_identifier` and `apply`. See Rule File API for a description of all functions supported in rule files. See NvRules for details on the interface available in the rule's `apply` function.

### 3.2. Integration

The rule system is integrated into NVIDIA Nsight Compute as part of the profile report view. When you profile a kernel, available rules will be shown in the report's Details page. You can either select to apply all available rules at once by clicking *Apply Rules* at the top of the page, or apply rules individually. Once applied, the rule results will be added to the current report. By default, all rules are applied automatically.
3.3. Rule System Architecture

The rule system consists of the Python interpreter, the *NvRules C++ interface*, the *NvRules Python interface* (NvRules.py) and a set of rule files. Each rule file is valid Python code that imports the NvRules.py module, adheres to certain standards defined by the Rule File API and is called from the tool.

When applying a rule, a handle to the rule *Context* is provided to its apply function. This context captures most of the functionality that is available to rules as part of the *NvRules*.
API. In addition, some functionality is provided directly by the NvRules module, e.g. for global error reporting. Finally, since rules are valid Python code, they can use regular libraries and language functionality that ship with Python as well.

From the rule Context, multiple further objects can be accessed, e.g. the Frontend, Ranges and Actions. It should be noted that those are only interfaces, i.e. the actual implementation can vary from tool to tool that decides to implement this functionality.

Naming of these interfaces is chosen to be as API-independent as possible, i.e. not to imply CUDA-specific semantics. However, since many compute and graphics APIs map to similar concepts, it can easily be mapped to CUDA terminology, too. A Range refers to a CUDA stream, an Action refers to a single CUDA kernel instance. Each action references several Metrics that have been collected during profiling (e.g. instructions executed) or are statically available (e.g. the launch configuration). Metrics are accessed via their names from the Action.

Each CUDA stream can contain any number of kernel (or other device activity) instances and so each Range can reference one or more Actions. However, currently only a single Action per Range will be available, as only a single CUDA kernel can be profiled at once.

The Frontend provides an interface to manipulate the tool UI by adding messages or graphical elements such as line and bar charts or tables. The most common use case is for a rule to show at least one message, stating the result to the user. This could be as simple as "No issues have been detected," or contain direct hints as to how the user could improve the code, e.g. "Memory is more heavily utilized than Compute. Consider whether it is possible for the kernel to do more compute work."

3.4. NvRules API

The NvRules API is defined as a C/C++ style interface, which is converted to the NvRules.py Python module to be consumable by the rules. As such, C++ class interfaces are directly converted to Python classes and functions. See the NvRules API HTML documentation for the classes and functions available in this interface.

3.5. Rule File API

The Rule File API is the implicit contract between the rule Python file and the tool. It defines which functions (syntactically and semantically) the Python file must provide to properly work as a rule.

Mandatory Functions

- get_identifier(): Return the unique rule identifier string.
- apply(handle): Apply this rule to the rule context provided by handle. Use NvRules.get_context(handle) to obtain the Context interface from handle.
- get_name(): Return the user-consumable display name of this rule.
- get_description(): Return the user-consumable description of this rule.
- get_section_identifier(): Return the unique section identifier that maps this rule to a section. Section-mapped rules will only be available if the corresponding
They implicitly assume that the metrics requested by the
section are collected when the rule is applied.

- **evaluate(handle):**
  
  Declare required metrics and rules that are necessary for this rule to be applied. Use `NvRules.require_metrics(handle, [...]` to declare the list of metrics that
  must be collected prior to applying this rule.

  Use e.g. `NvRules.require_rules(handle, [...])` to declare the list of other
  rules that must be available before applying this rule. Those are the only rules that
  can be safely proposed by the `Controller` interface.

### 3.6. Rule Examples

The following example rule determines on which GPU architecture a kernel was
running.

```python
import NvRules

def get_identifier():
    return "GpuArch"

def apply(handle):
    ctx = NvRules.get_context(handle)
    action = ctx.range_by_idx(0).action_by_idx(0)
    ccMajor =
        action.metric_by_name("device_attribute_compute_capability_major").as_uint64()
    if ccMajor >= 7:
        ctx.frontend().message("Running on Volta or newer")
    else:
        ctx.frontend().message("Running or Pascal or older")
```
Chapter 4.
SOURCE COUNTERS

The Source page provides correlation of various metrics with CUDA-C, PTX and SASS source of the application, depending on availability. The followings columns are shown:

- Address (SASS only)
- Source
- Live Registers
- Sampling Data
- (Source Counters)

Which Source Counter metrics are collected and the order in which they are displayed in this page is controlled using section files, specifically using the ProfilerSectionMetrics message type. Each ProfilerSectionMetrics defines one ordered group of metrics, and can be assigned an optional Order value. This value defines the ordering among those groups in the Source page. This allows, for example, you to define a group of memory-related source counters in one and a group of instruction-related counters in another section file.

```
Identifier: "SourceMetrics"
DisplayName: "Custom Source Metrics"
Metrics {
   Order: 2
   Metrics {
      Label: "Instructions Executed"
      Name: "inst_executed"
   }
   Metrics {
      Label: ""
      Name: "collected_but_not_shown"
   }
}
```

If a Source Counter metric is given an empty label attribute in the section file, it will be collected but not shown on the page.
<table>
<thead>
<tr>
<th>Address</th>
<th>Source</th>
<th>Sampling Data (All)</th>
<th>Sampling Data (No Issue)</th>
<th>Instructions Executed</th>
<th>Predicated-On Thread Instructions Executed</th>
</tr>
</thead>
<tbody>
<tr>
<td>000000</td>
<td>ADDIB, RX</td>
<td>1</td>
<td>1 (1)</td>
<td>89,000</td>
<td>2,000,000</td>
</tr>
<tr>
<td>000000</td>
<td>ADDIB, RX</td>
<td>1</td>
<td>1 (1)</td>
<td>89,000</td>
<td>2,000,000</td>
</tr>
<tr>
<td>000000</td>
<td>ADDIB, RX</td>
<td>1</td>
<td>1 (1)</td>
<td>89,000</td>
<td>2,000,000</td>
</tr>
<tr>
<td>000000</td>
<td>ADDIB, RX</td>
<td>1</td>
<td>1 (1)</td>
<td>89,000</td>
<td>2,000,000</td>
</tr>
<tr>
<td>000000</td>
<td>ADDIB, RX</td>
<td>1</td>
<td>1 (1)</td>
<td>89,000</td>
<td>2,000,000</td>
</tr>
<tr>
<td>000000</td>
<td>ADDIB, RX</td>
<td>1</td>
<td>1 (1)</td>
<td>89,000</td>
<td>2,000,000</td>
</tr>
<tr>
<td>000000</td>
<td>ADDIB, RX</td>
<td>1</td>
<td>1 (1)</td>
<td>89,000</td>
<td>2,000,000</td>
</tr>
<tr>
<td>000000</td>
<td>ADDIB, RX</td>
<td>1</td>
<td>1 (1)</td>
<td>89,000</td>
<td>2,000,000</td>
</tr>
<tr>
<td>000000</td>
<td>ADDIB, RX</td>
<td>1</td>
<td>1 (1)</td>
<td>89,000</td>
<td>2,000,000</td>
</tr>
</tbody>
</table>

Source Counters
This section documents the internals of the profiler report files (reports in the following) as created by NVIDIA Nsight Compute. The file format is subject to change in future releases without prior notice.

5.1. Version 7 Format

Reports of version 7 are a combination of raw binary data and serialized Google Protocol Buffer version 2 messages (proto). All binary entries are stored as little endian. Protocol buffer definitions are in the NVIDIA Nsight Compute installation directory under `extras/FileFormat`.

Table 1  Top-level report file format

<table>
<thead>
<tr>
<th>Offset [bytes]</th>
<th>Entry</th>
<th>Type</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Magic Number</td>
<td>Binary</td>
<td>NVP:0</td>
</tr>
<tr>
<td>4</td>
<td>Integer</td>
<td>Binary</td>
<td>sizeof(File Header)</td>
</tr>
<tr>
<td>8</td>
<td>File Header</td>
<td>Proto</td>
<td>Report version</td>
</tr>
<tr>
<td>8 + sizeof(File Header)</td>
<td>Block 0</td>
<td>Mixed</td>
<td>CUDA CUBIN source, profile results, session information</td>
</tr>
<tr>
<td>8 + sizeof(File Header) + sizeof(Block 0)</td>
<td>Block 1</td>
<td>Mixed</td>
<td>CUDA CUBIN source, profile results, session information</td>
</tr>
</tbody>
</table>

Table 2  Per-Block report file format

<table>
<thead>
<tr>
<th>Offset [bytes]</th>
<th>Entry</th>
<th>Type</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Integer</td>
<td>Binary</td>
<td>sizeof(Block Header)</td>
</tr>
</tbody>
</table>
### Table 3  Block payload report file format

<table>
<thead>
<tr>
<th>Offset [bytes]</th>
<th>Entry</th>
<th>Type</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Block Header</td>
<td>Proto</td>
<td>Number of entries per payload type, payload size</td>
</tr>
<tr>
<td>4 + sizeof(Block Header)</td>
<td>Block Payload</td>
<td>Mixed</td>
<td>Payload (CUDA CUBIN sources, profile results, session information, string table)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Offset [bytes]</th>
<th>Entry</th>
<th>Type</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Integer</td>
<td>Binary</td>
<td>sizeof(Payload type 1, entry 1)</td>
</tr>
<tr>
<td>4</td>
<td>Payload type 1, entry 1</td>
<td>Proto</td>
<td></td>
</tr>
<tr>
<td>4 + sizeof(Payload type 1, entry 1)</td>
<td>Integer</td>
<td>Binary</td>
<td>sizeof(Payload type 1, entry 2)</td>
</tr>
<tr>
<td>8 + sizeof(Payload type 1, entry 1)</td>
<td>Payload type 1, entry 2</td>
<td>Proto</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>Integer</td>
<td>Binary</td>
<td>sizeof(Payload type 2, entry 1)</td>
</tr>
<tr>
<td>...</td>
<td>Payload type 2, entry 1</td>
<td>Proto</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Notice

ALL NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO THE MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE.

Information furnished is believed to be accurate and reliable. However, NVIDIA Corporation assumes no responsibility for the consequences of use of such information or for any infringement of patents or other rights of third parties that may result from its use. No license is granted by implication of otherwise under any patent rights of NVIDIA Corporation. Specifications mentioned in this publication are subject to change without notice. This publication supersedes and replaces all other information previously supplied. NVIDIA Corporation products are not authorized as critical components in life support devices or systems without express written approval of NVIDIA Corporation.

Trademarks

NVIDIA and the NVIDIA logo are trademarks or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright

© 2018-2019 NVIDIA Corporation. All rights reserved.

This product includes software developed by the Syncro Soft SRL (http://www.sync.ro/).