CUDA accelerated Racon User Guide

One of the goals of the NVIDIA Genomics team is to accelerate *de novo* assembly with long read data. Over the past few months we have implemented a CUDA accelerated version of consensus generation using the Partial Order Alignment algorithm. To make it easily available for the community, we have augmented the open source polishing tool *Racon* to use it. This user guide goes over how to make use of the newly added CUDA-POA support in *Racon*.

In order to run the CUDA accelerated setup, the following conditions need to be met on the host machine -

1. CUDA version – 10.1
2. GPU Architecture - Turing or Volta (compiled for Pascal as well, but untested)
3. Minimum driver version - 418.40
4. Operating system – Ubuntu 16.04 and onwards
5. Compiled with GCC version 5.4.0

All of the functionality of default *Racon* has been maintained. Our work has simply extended the application to support CUDA accelerated polishing, and in the future will also support CUDA accelerated alignment.

To enable CUDA acceleration -
1. Use `-c N` to run with GPU where N is number of GPU batches. Each batch takes around 2.5GB of GPU memory. It's best to get the maximum number of batches which your GPU will tolerate.
2. If you run with '-c' you can add an additional `-b` flag - this runs with banded alignment and gives another performance boost (You will also be able to run with more batches because banding requires more memory). The `-b` option is still a bit experimental and uses a fixed band, we will add Suzuki-style adaptive banding soon.
3. By default racon support multi GPU mode. If not specified, it uses all available GPUs, assumed them to be the same, and distributes the specified batches (N) equally among them. If you would like to limit which GPU(s) to run on, please use the CUDA_VISIBLE_DEVICES=<> environment variable.

Disclaimer - This release represents an early access to the alpha version of our work. The effort is still under active development, and much remains to be tested and validated. Please expect frequent updates and bug fixes. We would also appreciate

your candid feedback on the performance, accuracy and overall user experience as it would help us improve future iterations and serve the community better.

If you have any questions, comments or concerns, feel free to reach out to us -
Mike Vella - mvella@nvidia.com (Timezone - GMT)
Joyjit Daw - jdaw@nvidia.com (Timezone - US EST)