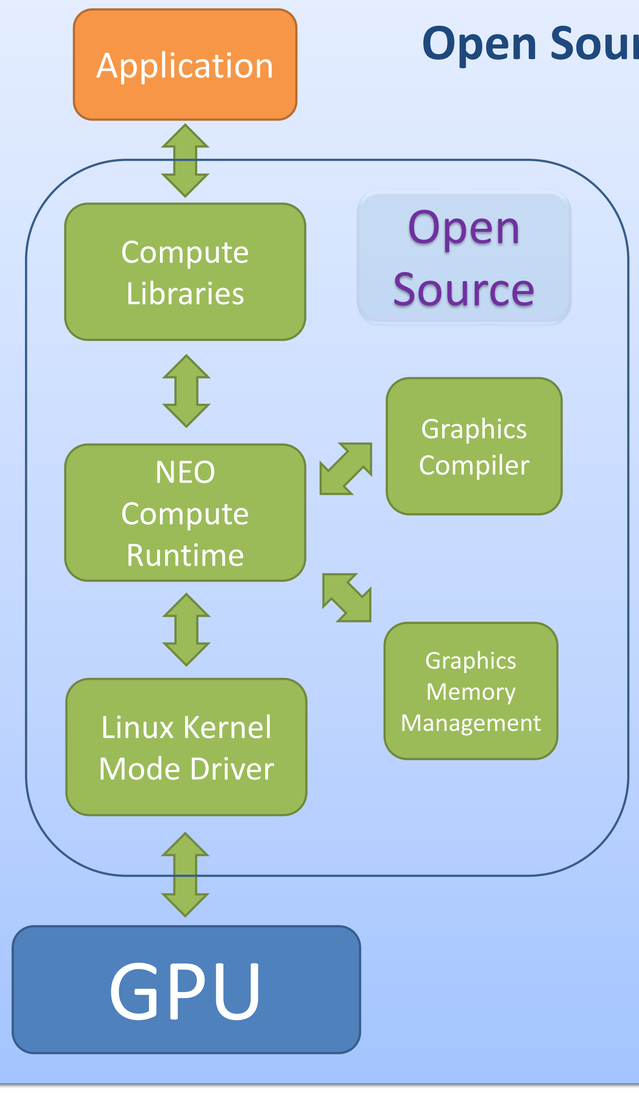
SHAPING OPEN SOURCE COMPUTE ECOSYSTEM WITH NEO AND COMPUTE LIBRARIES

Michal Mrozek, Intel® Corp.

Open Source Software Stack

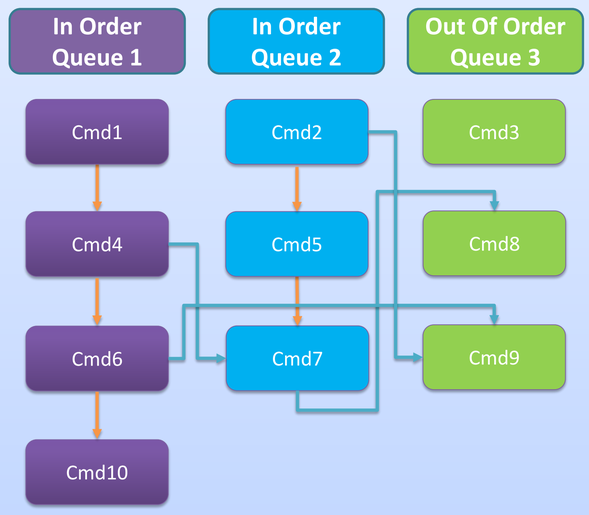


- All software components are available in open source.
- Driver releases every week with validated package.
- Everything works on open source kernels without any custom patches.
- Single unified driver for all BDW+ Intel® GPU devices.
- **N**ew **E**nhanced **O**penCL™ driver.

We are open for feedback and contributions!

https://github.com/intel/compute-runtime

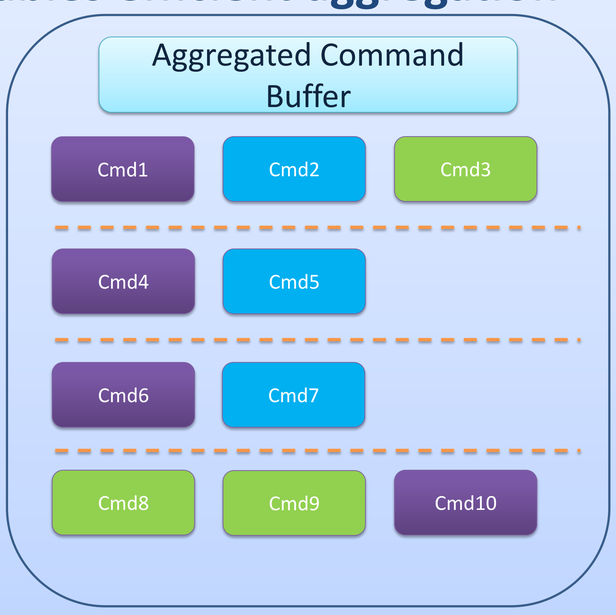
Driver N:1 architecture enables efficient aggregation




```
//Suggested enqueue sequence
enqueue(ioq1, cmd1);
enqueue(ioq2, cmd2, event1); //ret ev1
enqueue(ooq3, cmd3);
enqueue(ioq1, cmd4, event2); //ret ev2
enqueue(ioq2, cmd5);
enqueue(ioq1, cmd6, event3); //ret ev3
enqueue(ioq2, cmd7, event2, event4); //wait for ev2,ret ev4
enqueue(ooq3, cmd8, event4); //wait for ev4
enqueue(ooq3, cmd9, event1&event3); //wait for ev1 & ev3
enqueue(ioq1, cmd10);
flush() // flush all queues
```

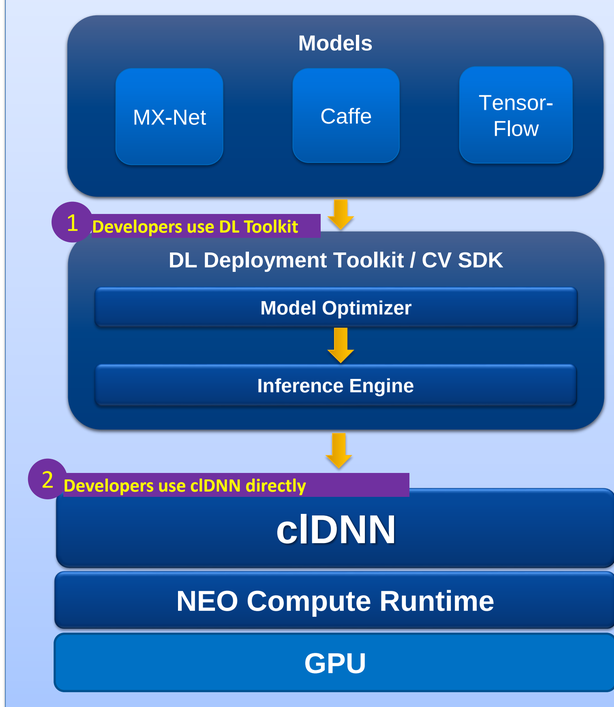
Check out our samples:

https://github.com/intel/compute-samples



- Create at least one out of order queue to enable aggregation mechanism, utilize it to submit tasks without dependencies.
- Prefer round-robin enqueue sequence.
- Avoid user events.

Compute Library for Deep Neural Networks (clDNN)

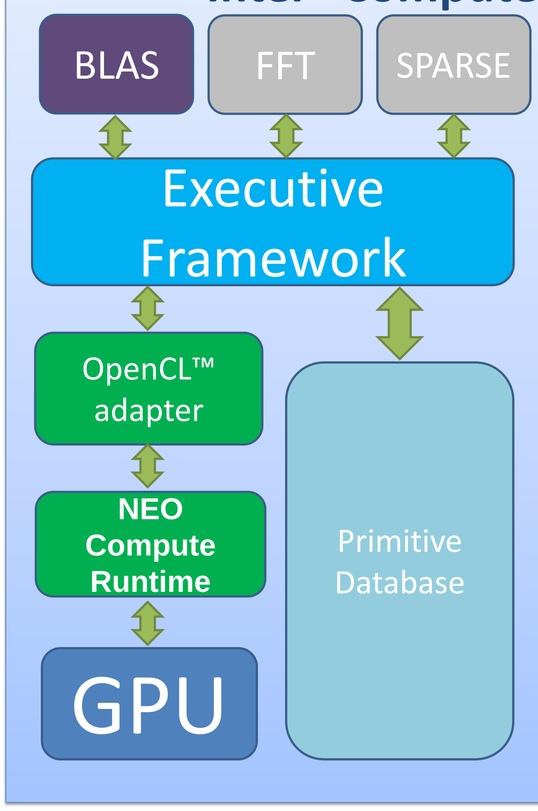


- clDNN contains kernels-primitives optimized for DNN inference acceleration on Intel® SKL+ GPU devices.
- It supports most of commonly known latest neural network topologies.
- Delivered with Intel® Deep Learning (DL) Deployment Toolkit (DT) which supports Caffe, Tensor-Flow and MX-Net models.
- Intel® DL DT contains Model Optimizer, which converts and optimizes trained models. Inference Engine executes those models on GPU using clDNN.

Would you like to know MORE?

https://github.com/intel/clDNN

Intel® Compute Libraries for GPU

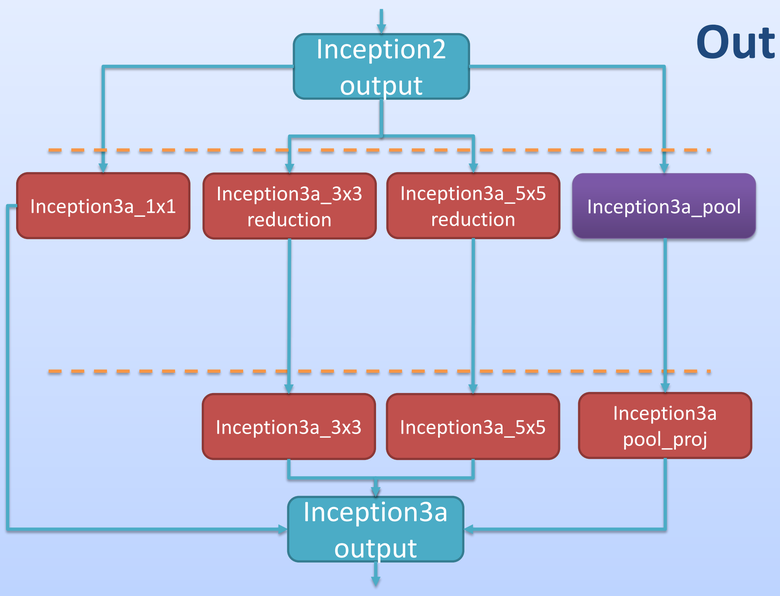


- Extensible repository-framework for compute functions.
- First delivered part is Basic Linear Algebra Subprograms (BLAS) Level 1, 2, 3 single and complex FP32.
- This is a prototype with ongoing development.
- Work on other libraries is currently in progress (i.e. FFT, SPARSE).

Feel free to communicate with us:

https://github.com/intel/clGPU

Out of Order Queue accelerates Neural Networks



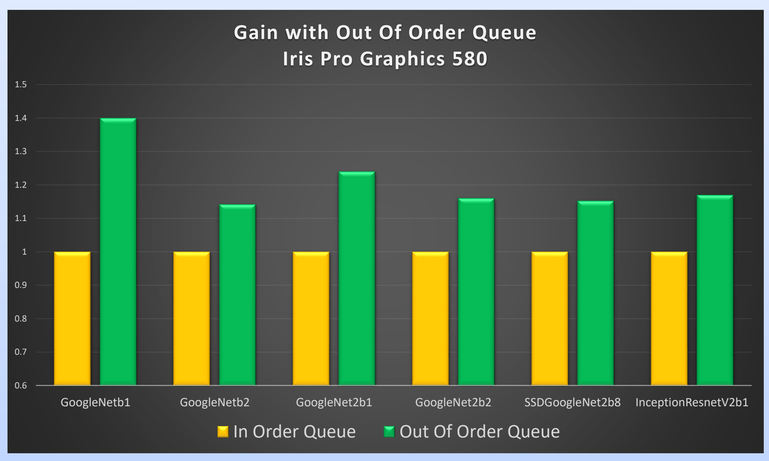

```
//no events are used in all calls below
clEnqueueBarrierWithWaitList( queue )

clEnqueueNDRangeKernel(queue, inception3a_pool )
clEnqueueNDRangeKernel(queue, inception3a_1x1 )
clEnqueueNDRangeKernel(queue, inception3a_3x3_reduce )
clEnqueueNDRangeKernel(queue, inception3a_5x5_reduce )

clEnqueueBarrierWithWaitList( queue )

clEnqueueNDRangeKernel( queue, inception3a_pool_proj )
clEnqueueNDRangeKernel( queue, inception3a_3x3 )
clEnqueueNDRangeKernel( queue, inception3a_5x5 )

clEnqueueBarrierWithWaitList( queue )
```



- Some topologies contain steps consisting of multiple independent compute pieces.
- For small batches those kernels may underutilize GPU if they are executed sequentially.
- clDNN utilizes one Out Of Order Queue and groups independent kernels together.
- clEnqueueBarrierWithWaitlist with no events ensures that all previous commands are completed.
- With Out Of Order Queue driver can remove synchronization points between kernels and GPU can execute them together in one shot.
- That gives performance boosts **up to 40%.**

Unleash the power & performance of Intel® Graphics with Compute Libraries and NEO driver!

International Workshop on OpenCL
14-16 May 2018

IWOCL '18, May 14–16, 2018, Oxford, United Kingdom

ACM ISBN 978-1-4503-6439-3/18/05.
https://doi.org/10.1145/3204919.3207895