Windows® 10 May 2019 Update for Machine Learning Acceleration on Intel® Integrated Graphics

Introduction

Microsoft's Game Developers Conference (GDC) announcement in March 2019 emphasized the rising adoption of the Windows* machine learning (Windows ML) API and Direct Machine Learning (DirectML) across a wide range of applications, especially gaming engines. The Windows ML API handles the hardware abstraction for graphics processing unit (GPU) acceleration by interfacing with DirectML for per operator/layer execution. DirectML is a high-performance, low level API for running machine learning operations. DirectML API is part of the DirectX* family. The GDC announcement also highlights how independent hardware vendors (IHVs) such as Intel are collaborating with Microsoft to improve DirectML operator performance by providing architecture-specific optimization, called MetaCommands.

Intel's earlier post in May 2018 introduced the Windows ML API and the DirectML API implementation on Intel® hardware via the DirectX 12 DirectCompute interface. In October 2018, Intel announced the first version of DirectX 3D* 12 MetaCommand support in the Intel® Graphics Driver that shipped in Microsoft's Windows® 10 October 2018 Update. That announcement also highlighted the significant performance boost seen when the DirectML convolution operator is executed as a DirectX 3D* 12 MetaCommand instead of the default High-Level Shader Language (HLSL) shader path.

With the upcoming release of the Windows 10 May 2019 Update, we at Intel are excited to give our customers a peek into the latest MetaCommand feature updates and Windows ML performance improvements to expect in the latest Intel® Graphics - Windows® 10 Drivers

Windows* ML Performance Improvement

In the Windows 10 May 2019 Update, Windows ML API performance on Intel® Integrated Graphics is improved through a combination of graph-level optimizations in the Windows ML Runtime and DirectX 3D* 12 MetaCommand performance enhancements in the Intel Graphics Driver. This combination enables Intel to fully engage the Intel GPU's compute hardware engine capabilities to provide significant improvement in performance compared to the Windows® 10 October 2018 Update.

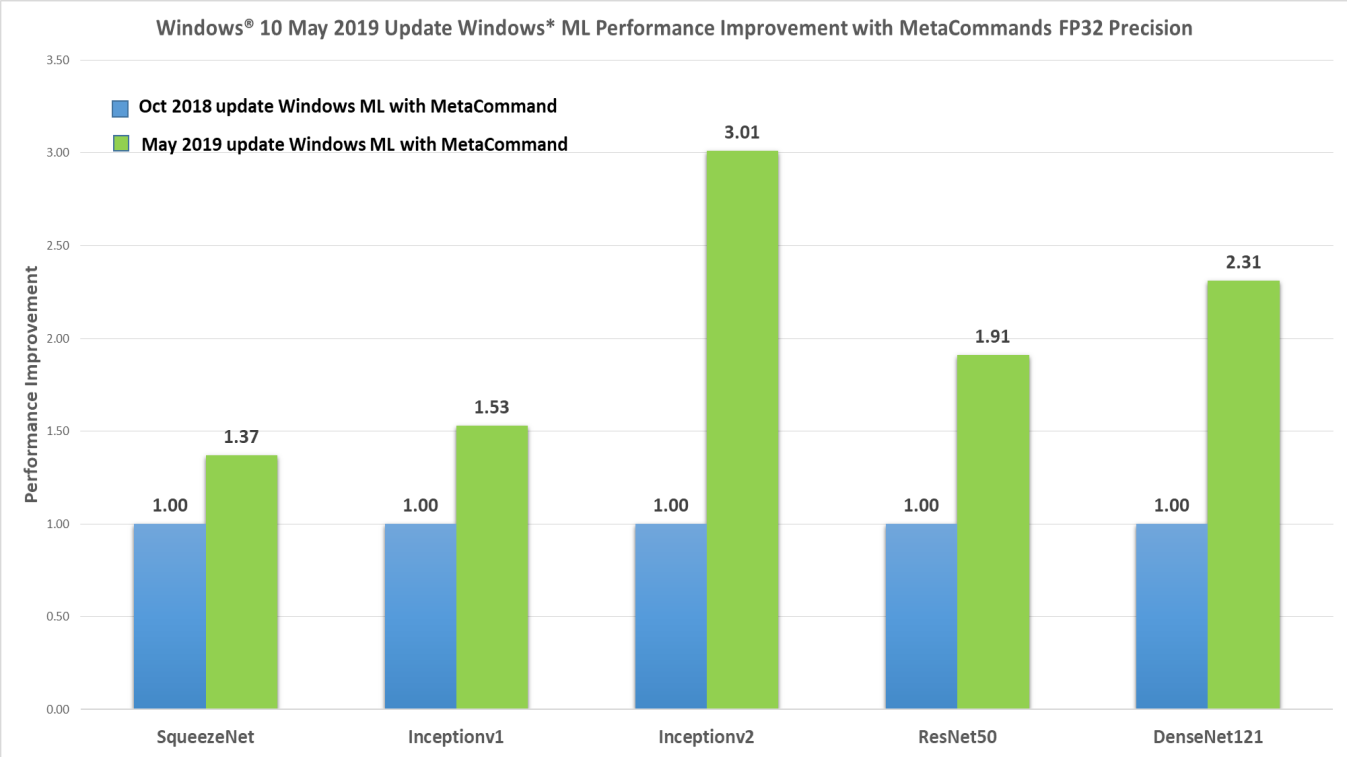
Windows* ML Runtime Graph-Level Optimizations

Windows ML runtime evaluates the trained model using the Open Neural Network Exchange (ONNX) Model Inference Engine. Such evaluation consists of a graph compilation process, which determines variables such as GPU submissions count and memory usage that heavily influence the overall topology performance. With data-driven analysis, Intel and Microsoft successfully co-engineered the fusion of certain nodes in the Windows ML API runtime graph processor, thereby reducing the number of layers required to perform the desired inference evaluation. As a result, the GPU execution time and submissions reduced drastically, thereby significantly improving performance[Φ]

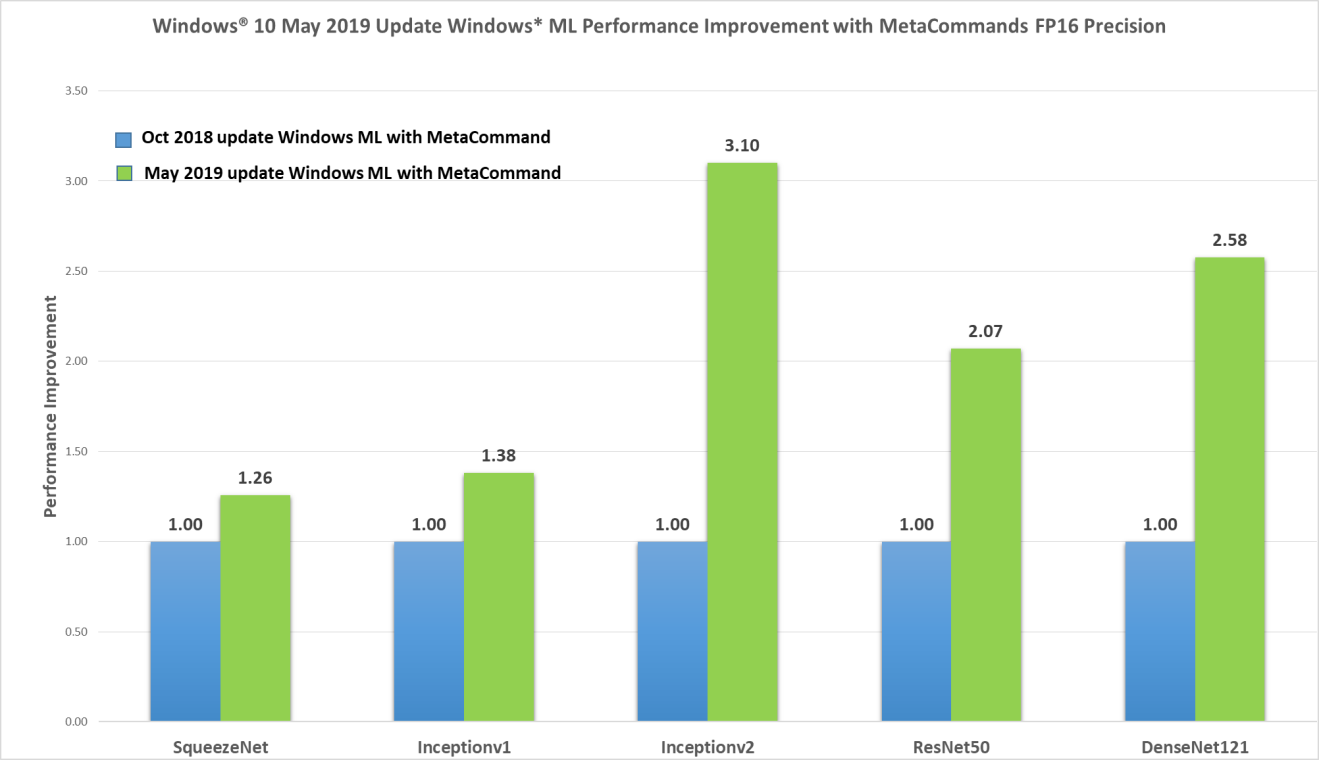
[Φ] Performance results are based on testing as of May 3rd 2019 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark* and MobileMark*, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Windows® 10 May 2019 Update Windows ML API Performance Improvement with MetaCommand

With the combination of Windows ML runtime Graph-level optimizations and additional operators supported as MetaCommands on Windows® 10 May 2019 Update, we can see approximately 3x speedup[Φ] for FP32 and approximately 2x speedup[Φ] for FP16, compared to the Windows 10 October 2018 Update.



[Φ] Performance results are based on testing as of May 3rd 2019 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark* and MobileMark*, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.



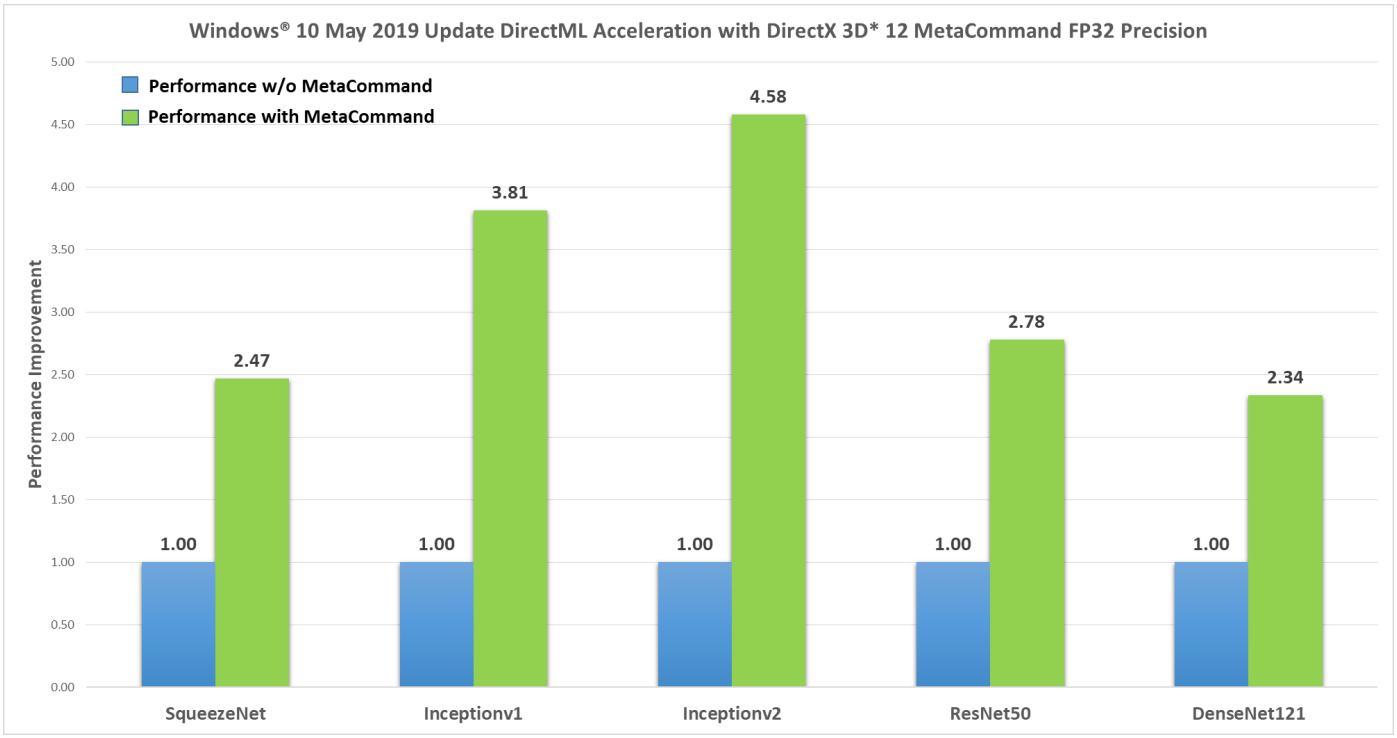
Windows® 10 May 2019 Update DirectML Acceleration with MetaCommand

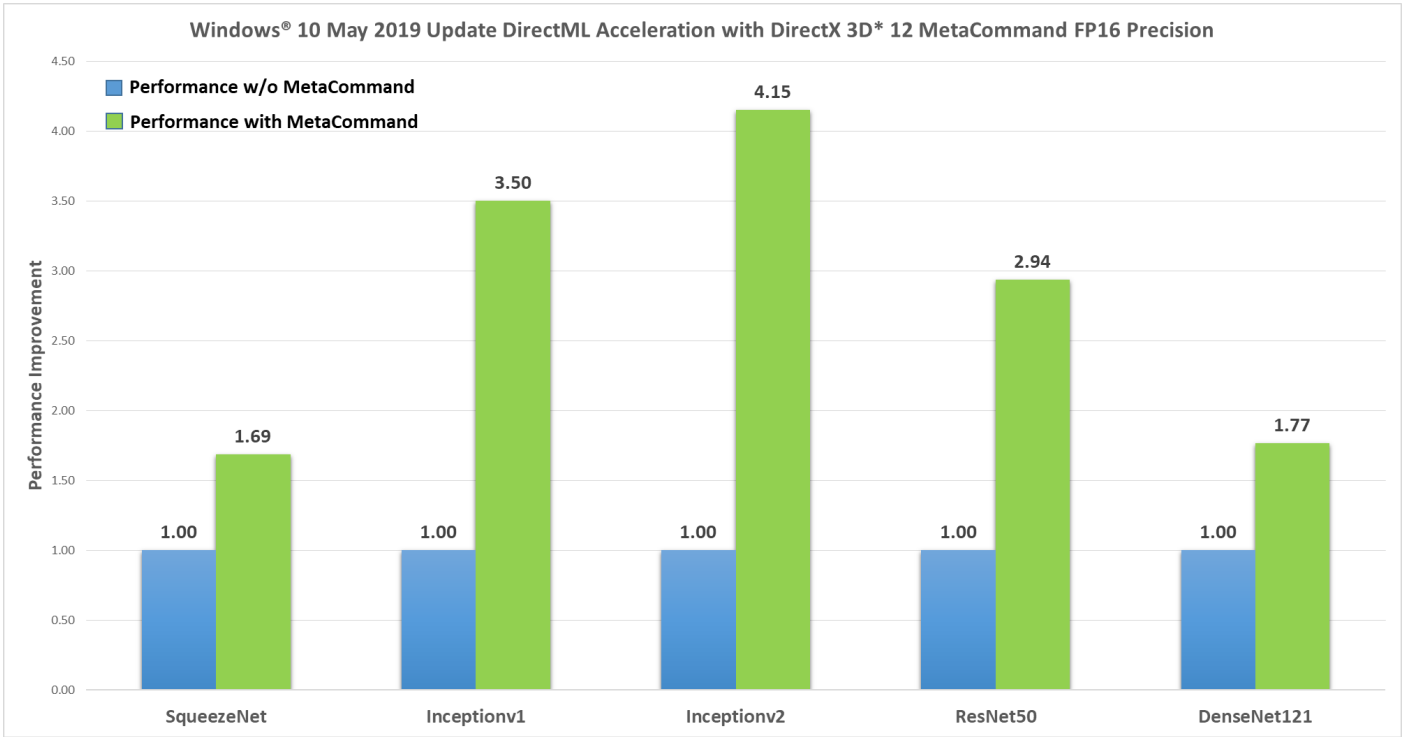
DirectML operators are hardware accelerated by MetaCommands that use architecture-specific optimizations. In addition to Convolution (both FP16 and FP32) and General Matrix Multiplication (GEMM) (FP32), the Intel® Graphics Windows® 10 Driver now adds support for the following operators as MetaCommands.

- Pooling MetaCommand operator
- FP16 GEMM MetaCommand operator

These new MetaCommands improve the performance[Φ] of several prominent image classification topologies such Resnet50 and InceptionV1. The plots below show the relative inference performance gains[Φ] across topologies when the inference workload is executed both with and without MetaCommands.

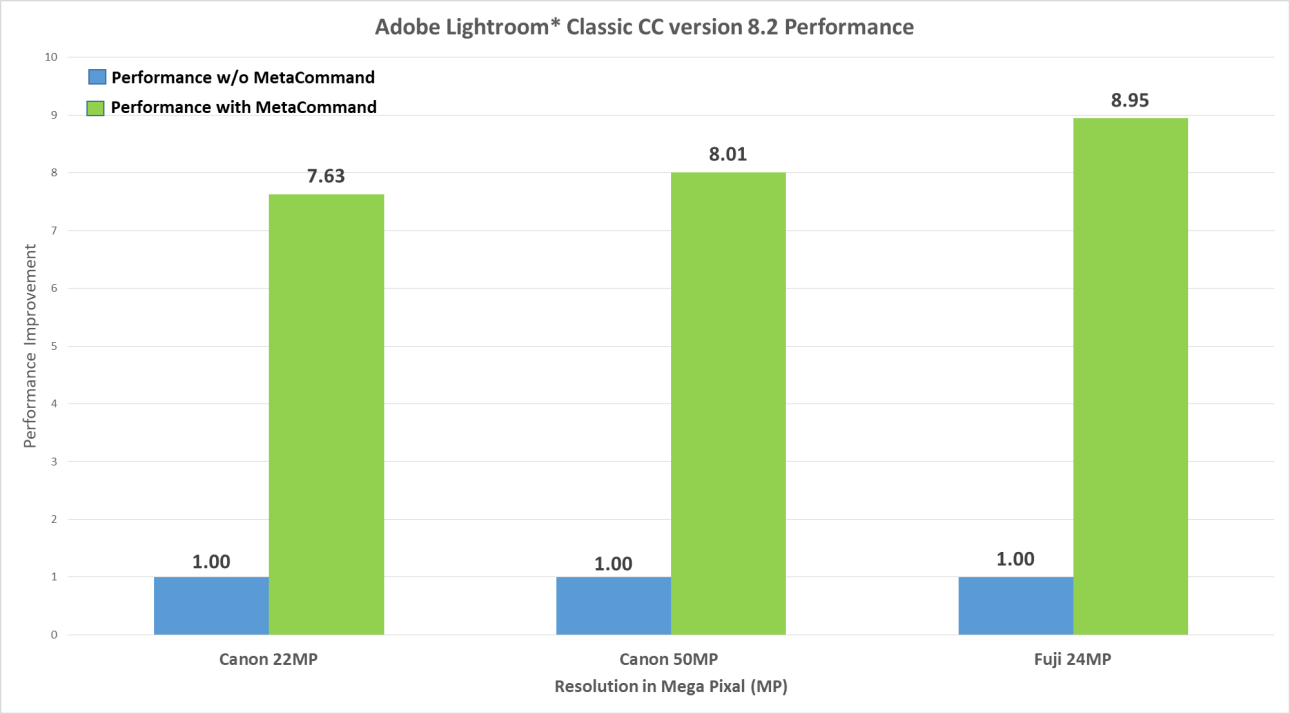
[Φ] Performance results are based on testing as of May 3rd 2019 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark* and MobileMark*, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.


Windows® 10 May 2019 Update DirectML Acceleration with DirectX 3D* 12 MetaCommand FP32 Precision
Performance w/o MetaCommand
Performance with MetaCommand
Performance Improvement
5.00
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
1.00
2.47
1.00
3.81
1.00
4.58
1.00
2.78
1.00
2.34
SqueezeNet
Inceptionv1
Inceptionv2
ResNet50
DenseNet121


Windows® 10 May 2019 Update DirectML Acceleration with DirectX 3D* 12 MetaCommand FP16 Precision
Performance w/o MetaCommand
Performance with MetaCommand
Performance Improvement
4.50
4.00
3.50
3.00
2.50
2.00
1.50
1.00
0.50
0.00
1.00
1.69
1.00
3.50
1.00
4.15
1.00
2.94
1.00
1.77
SqueezeNet
Inceptionv1
Inceptionv2
ResNet50
DenseNet121

Adobe Lightroom* MetaCommand Case Study

While the examples above reflect the performance in some typical convolutional neural network (CNN) topologies, the benefits of Windows ML's GPU acceleration with MetaCommands can be seen in a variety of applications too. The Adobe "Enhance Details" feature demonstrated in the Lightroom* Classic CC version 8.2 application uses Windows ML API to improve image quality by demosaicing RAW images. This application shows up to approximately 9x speedup[Φ] with MetaCommands enabled for FP16 precision on Intel Integrated Graphics.



[Φ] Performance results are based on testing as of May 3rd 2019 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark* and MobileMark*, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Conclusion

More DirectX 3D 12 MetaCommand functionality will be supported to accelerate DirectML operators in upcoming Intel Graphics Windows 10 Drivers. Be sure to check for the latest Intel Graphics Windows 10 Driver to obtain the most current benefits of MetaCommand improvements for Windows ML and DirectML applications.

For more information on running Windows ML on the PC, please visit *AI on the PC: Create New Usages and Deploy them to Your PC*.

Configuration Disclosure

Test Environment and system configuration under which the performance claim or benchmark data was obtained.

- Platform: Intel® Core™ i7-7567U processor with Iris® Plus graphics 650
- Graphics Driver: Intel® Graphics driver 26.20.100.6813
- Operating System: Windows® 10 May 2019 Update version 18362
- Operating System Power plan : High Performance
- Benchmark application: Tests performed using Microsoft's WinMLRunner tool in the Windows ML sample application. The inferences/sec is calculated from the application's reported "Evaluate" time over 1000 iterations.
- Pretrained models were obtained from the ONNX documentation.
- Company that performed the testing: Intel, 1900 Prairie City Rd, Folsom, CA 95630
- Date of testing: May 3rd 2019

Notices

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.

Copies of documents which have an order number and are referenced in this document may be obtained by calling 1-800-548-4725 or by visiting www.intel.com/design/literature.htm.

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

Microsoft, Windows, and the Windows logo are trademarks, or registered trademarks of Microsoft Corporation in the United States and/or other countries.

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

© 2019 Intel Corporation