AIDC

INTEL AI DEVCON 2018

AIDC
INTEL AI DEVCON 2018

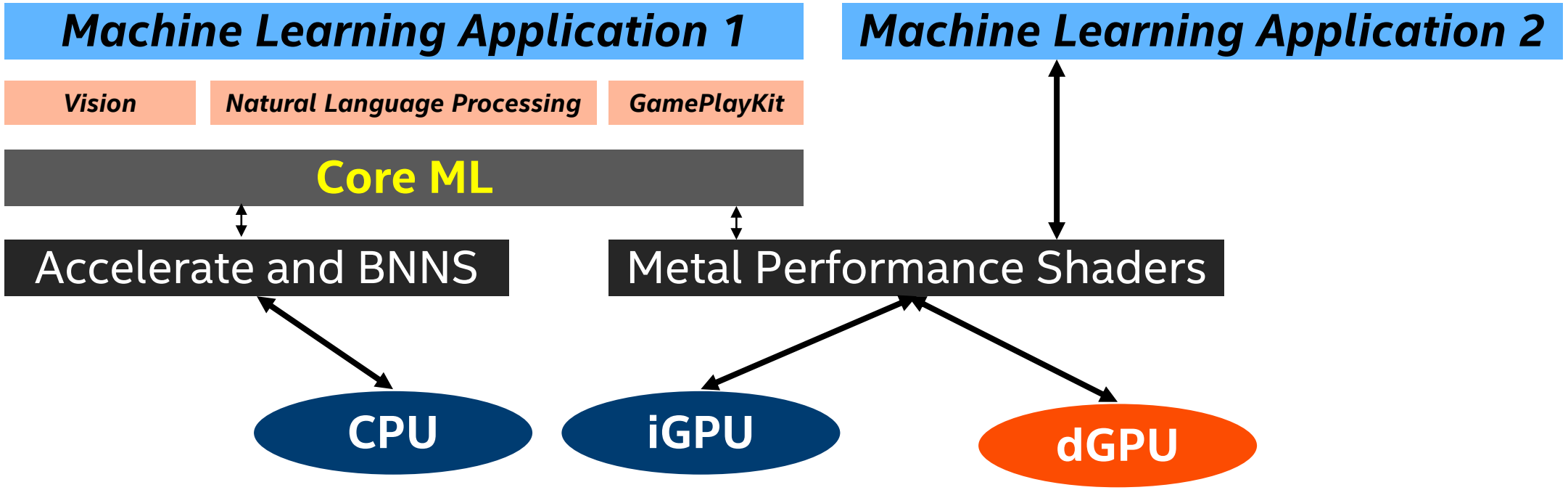
ACCELERATE MACHINE LEARNING ON MACOS WITH INTEL® INTEGRATED GRAPHICS

HISHAM CHOWDHURY

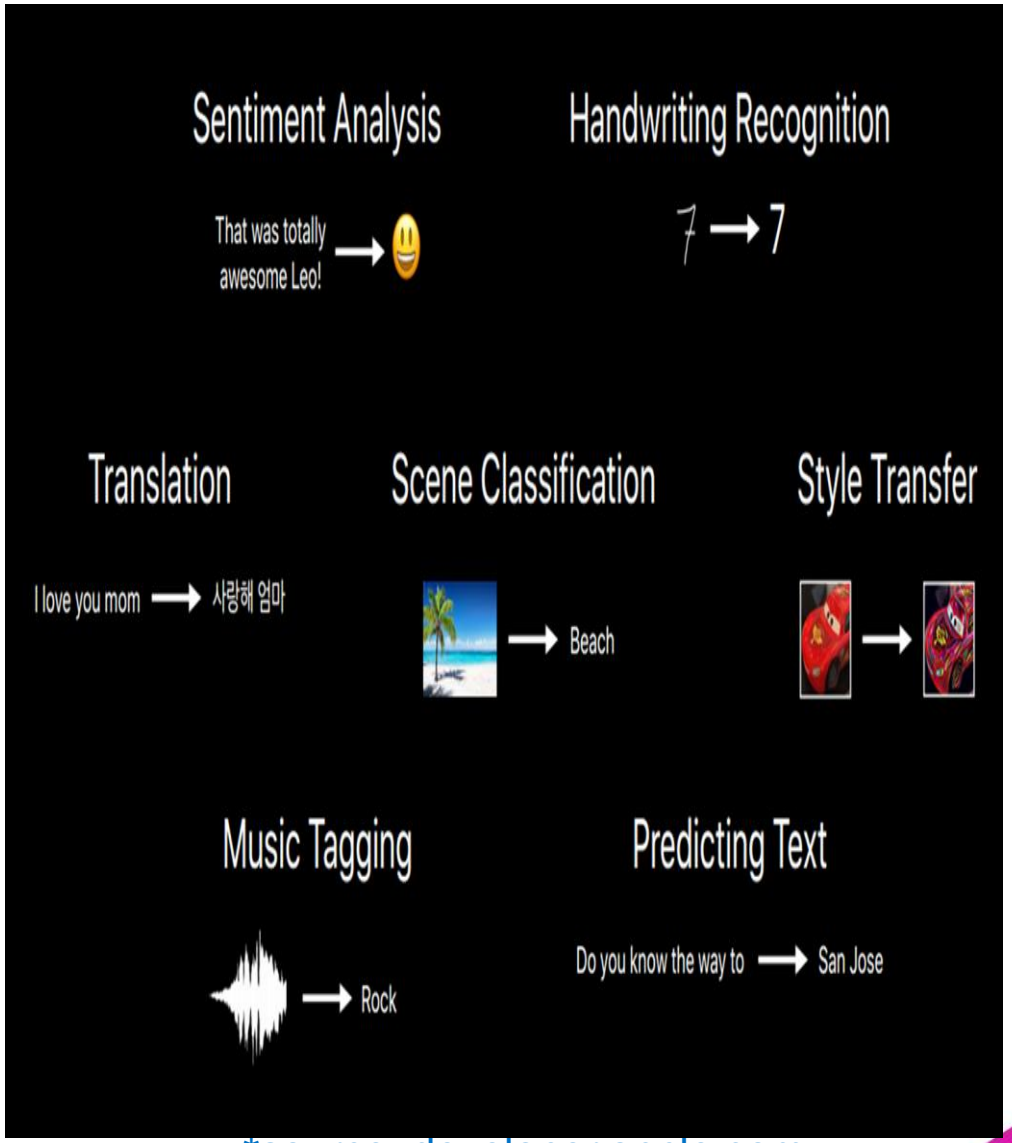
MAY 23, 2018

intel AI

APPLE MACHINE LEARNING STACK

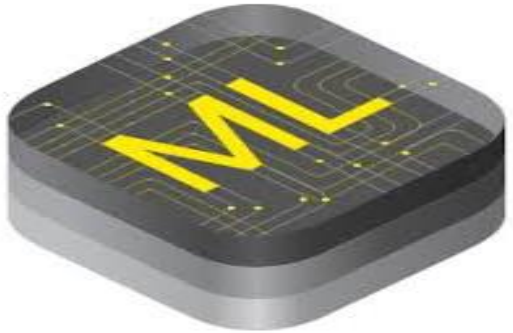


ML USAGE ON MACOS

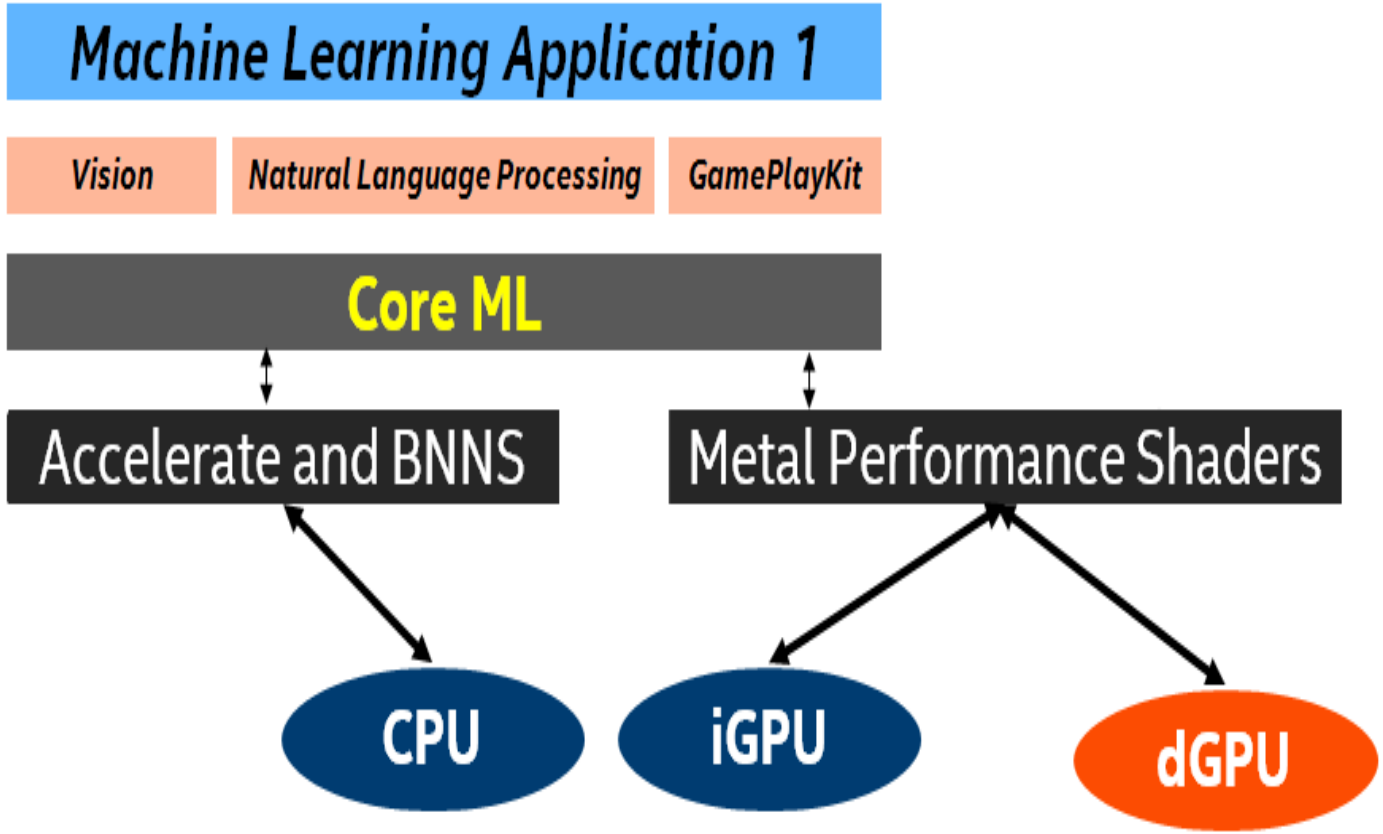


*source: developer.apple.com

USING CORE ML

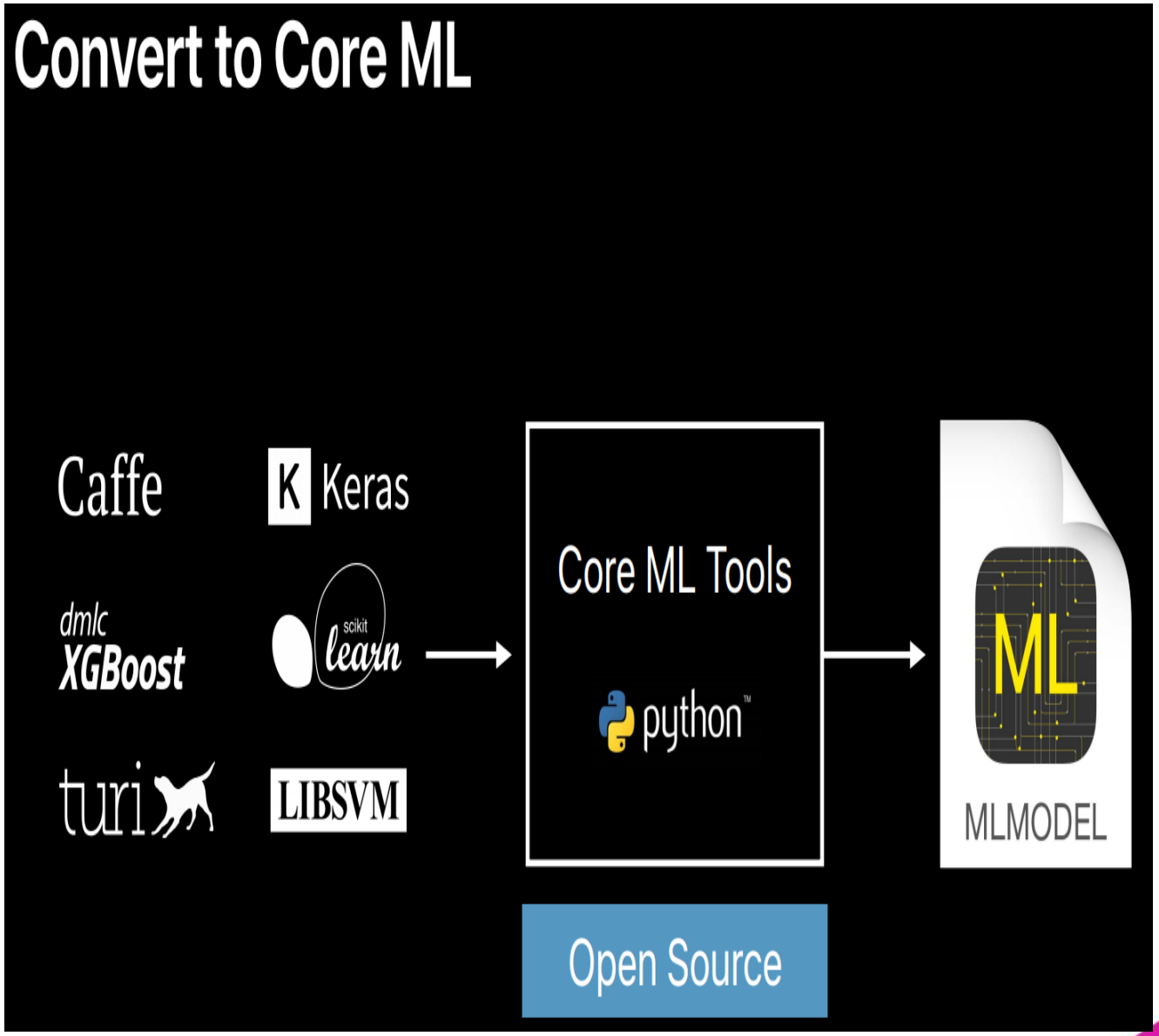


- Easy to use high level framework for ML needs
- Easy to integrate ML models in your code
- Provides tools to convert already trained model to core ML models
- Core ML can use CPU or GPU path depending on the application profile
- Built on top of highly optimized CPU and GPU primitives
- Complexities are abstracted from the application



CORE ML WORKFLOW

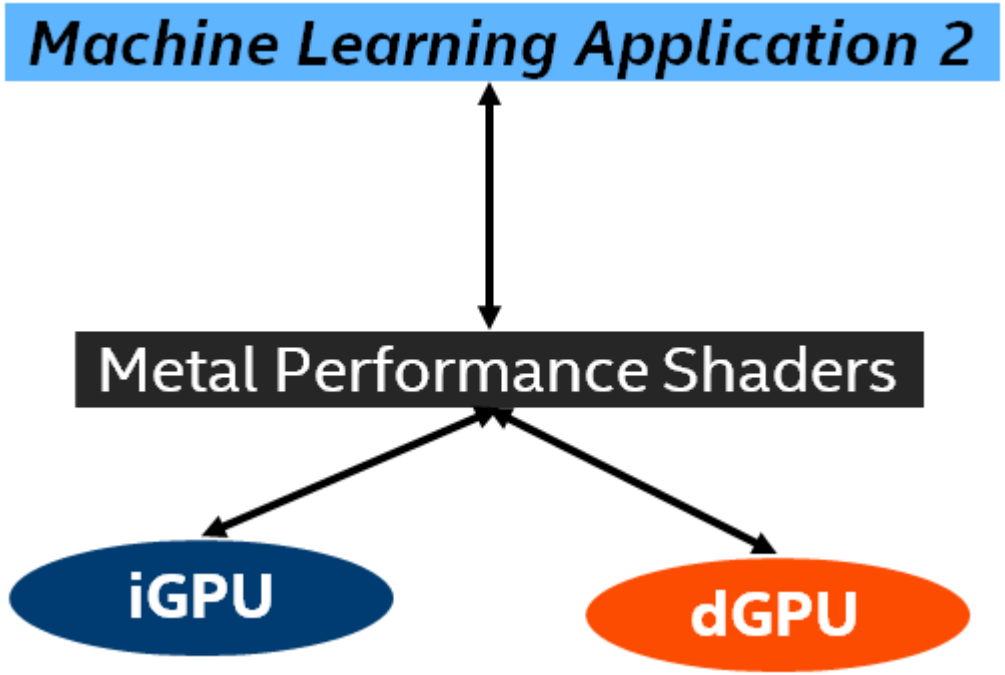
- Integrate model in xcode
  - Load existing .mlmodel provided by Apple or
  - Convert already trained model(caffe, keras etc) to core ML model (.mlmodel)
- do predict:
  - all the magic will happen underneath



https://pypi.org/project/coremltools/

USING METAL PERFORMANCE SHADERS (MPS)

- Can build your ML application directly using Metal Performance Shaders (MPS)
- Low level primitive based framework: MPSCNN, MPSMatrix etc
- More control and low overhead: create the topology graph
- Executes on the GPU
- Optimized for the underlying GPU



ACCELERATE USING INTEL® INTEGRATED GPU

END-TO-END AI COMPUTE

DATACENTER

Many-to-many hyperscale for stream and massive batch data processing

Ethernet & Wireless

GATEWAY

1-to-many with majority streaming data from devices

Wireless and non-IP wired protocols

- ✓ *Secure*
- ✓ *High throughput*
- ✓ *Real-time*

EDGE

1-to-1 devices with lower power and often UX requirements

- Intel® Xeon® Processors
- Intel® Xeon Phi™ Processors*
- Intel® Core™ & Atom™ Processors
- Intel® Integrated GPU
- Intel® FPGA
- Crest Family (Nervana ASIC)*
- Movidius Myriad (VPU)
- Intel® GNA (IP)*

Inference on the Edge

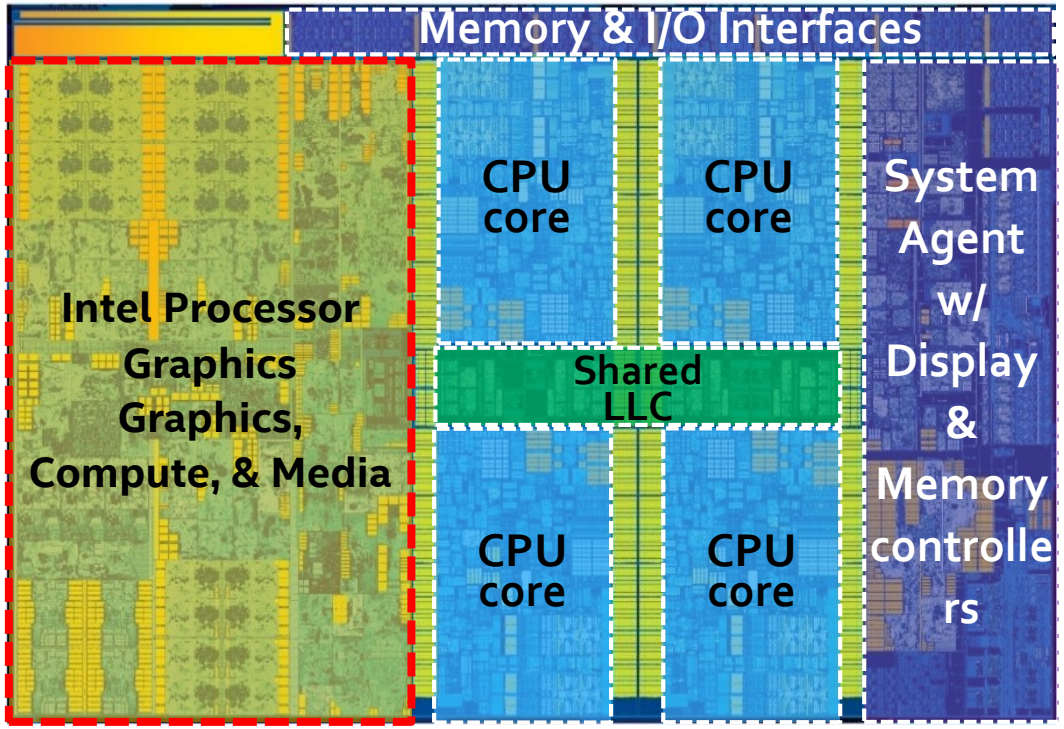
Intel® Integrated GPU

*source: apple.com

- All Macbook lines have Intel® Integrated GPU = take advantage of already built-in engine for ML needs
- MPS Intel layer highly optimized for Intel® Integrated GPU = High HW efficiency
- No need to write your own algorithms for ML primitives = use MPS or core ML
- Offload the CPU

INTEL® INTEGRATED GPU FOR AI

- Performs 3D/media/display and GPGPU (compute)
- All MacBook and MacBook pro lines comes with Intel Integrated Graphics
- EUs (execution units): executes GPGPU kernels
- Raw compute capability measured in FLOPS or OPS; but memory bandwidth and efficiency key
- Run AI Inference on GPU using set of ML primitives → MPS



INTEL® INTEGRATED GPU FOR AI ON MACOS

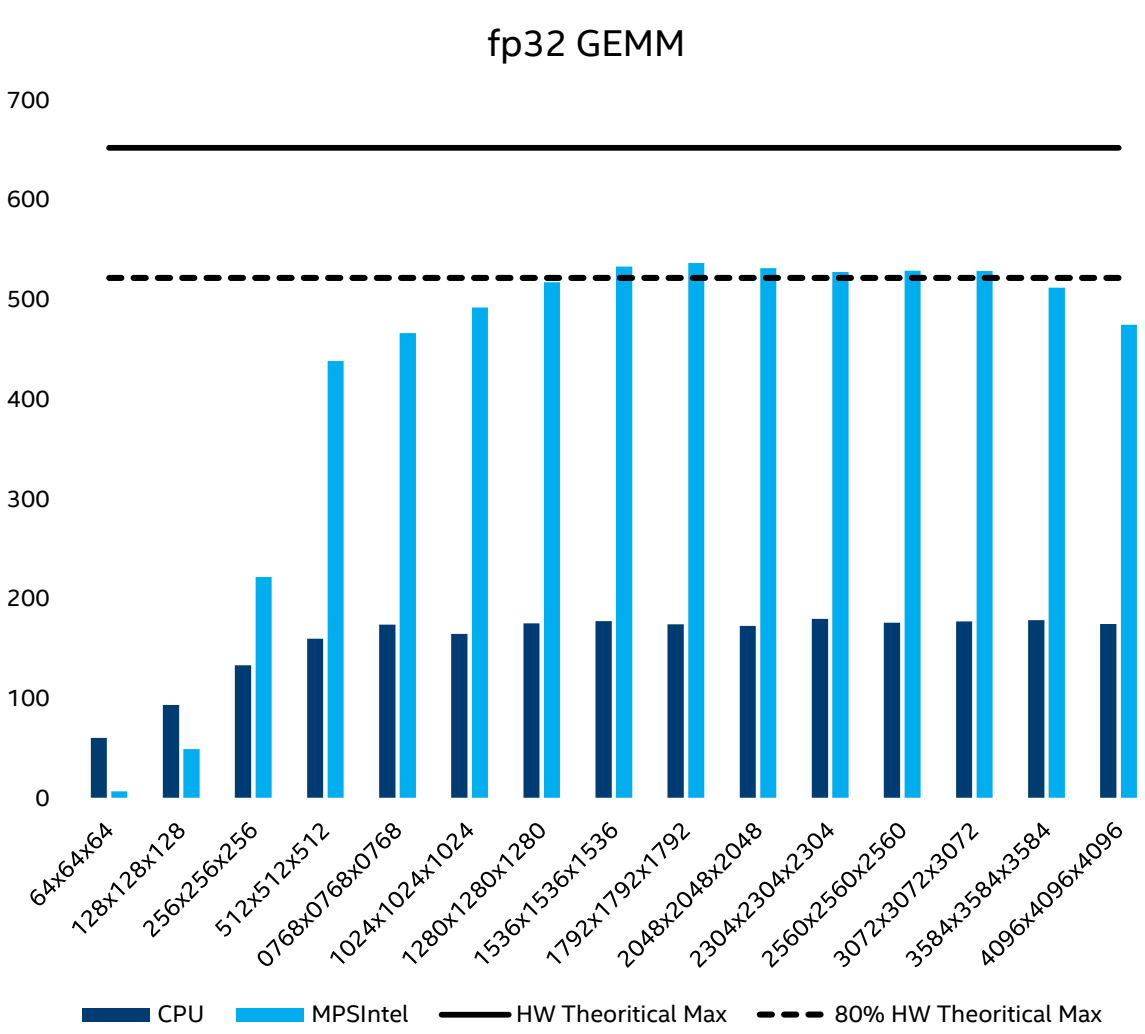
INTEL MPS LAYER

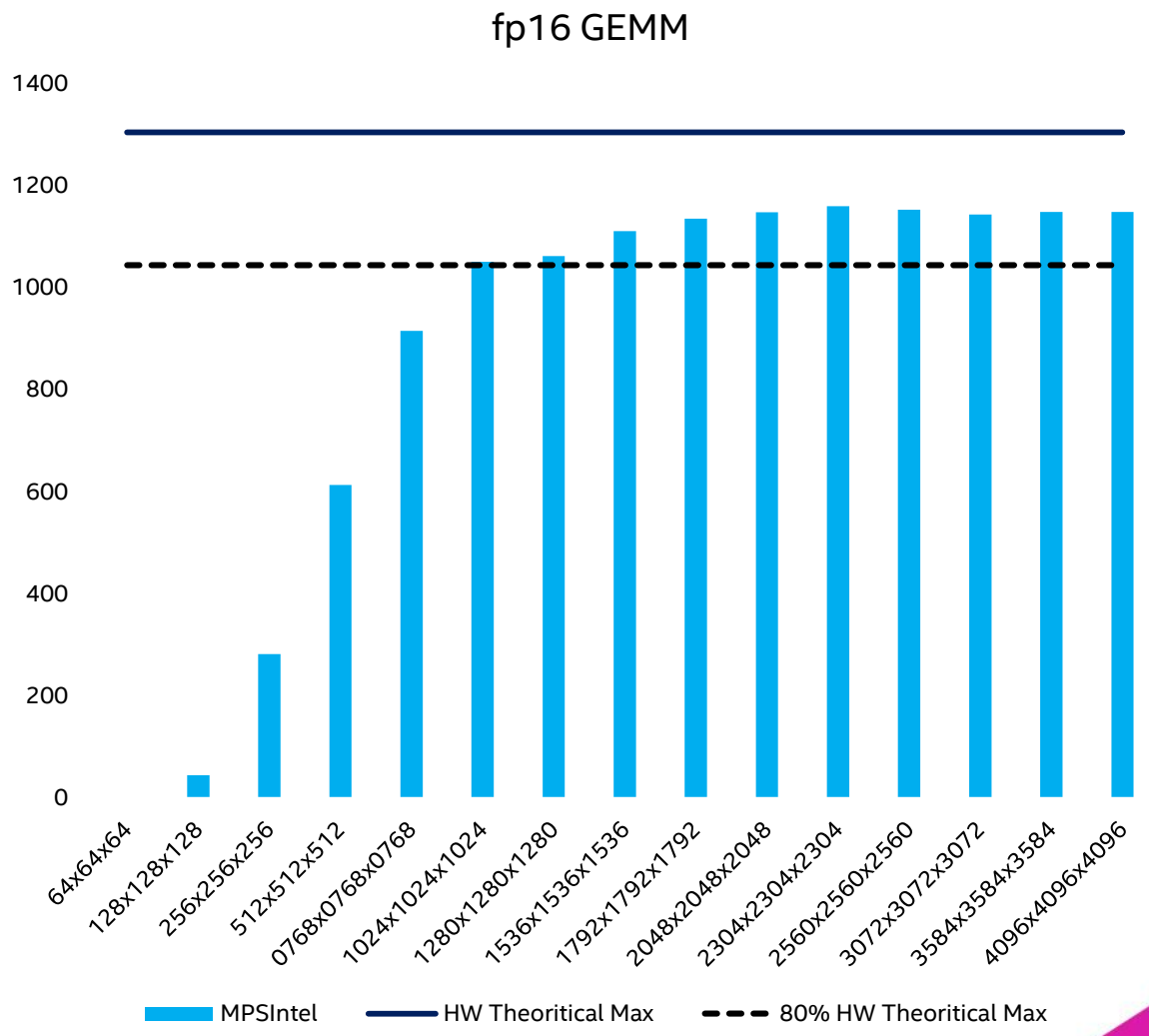
- Highly tuned and optimized algorithms are used for Intel® Integrated GPU
  - 80% HW efficiency
  - Based on the use cases and data-set, highly tuned kernels are deployed to maximize performance
  - Common use cases heavily optimized to fit the HW resources
  - Different algorithm deployed to maximize performance (ex. various conv)
- Use of Custom Hardware Features
  - Uses HW features to accelerate MPS primitives to fit the hardware best
- Kernel Selection Framework
  - Based static profiling and input combination picks the best implementation for the primitimives
  - Optimized for key topologies such as : Inception, VGG, Resnet, Alexnet, Mobilenet etc

PERFORMANCE CASE STUDY

MPS ON INTEL

PERFORMANCE: VS HW

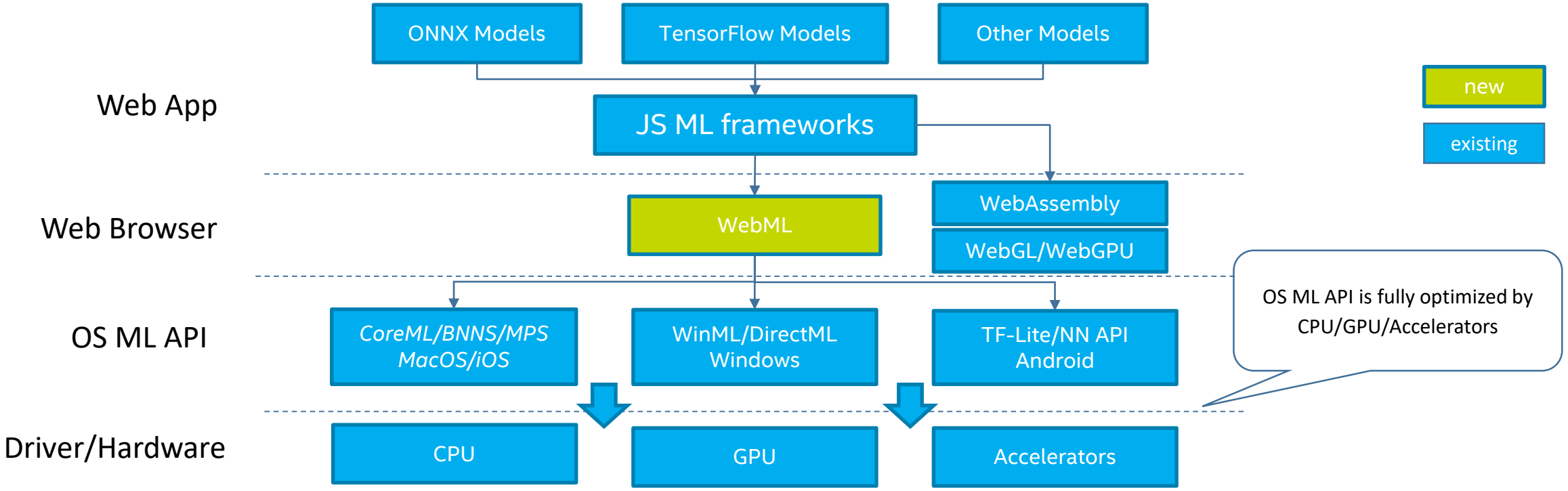




**See disclaimer section

WebML POC using optimized MPS

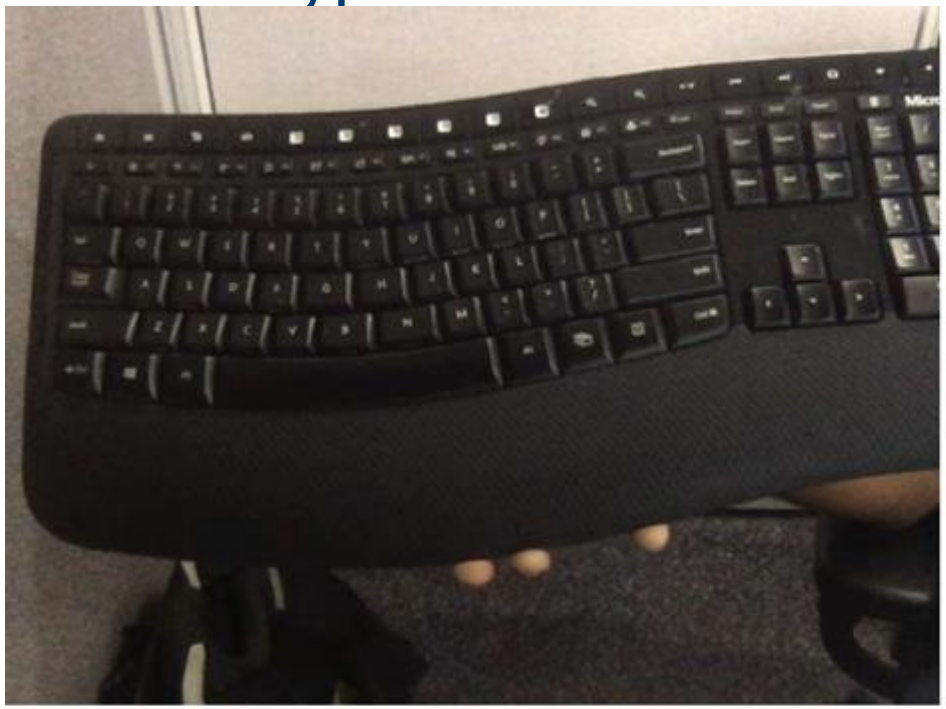
PROPOSED *WEBML*: ACCELERATED WEB MACHINE LEARNING API



- Standard-based ML Web API focus on pre-trained model inferencing
- Integrate with other Web APIs, e.g. text, multimedia, sensors and VR/AR, for real-time AI-based apps on client devices
- Web ML workloads run on top of OS ML API and fully exploit the CPU/GPU/Accelerator performance on client devices

WEBML POC

- Prototype WebML API in Chromium M65 on MacOS



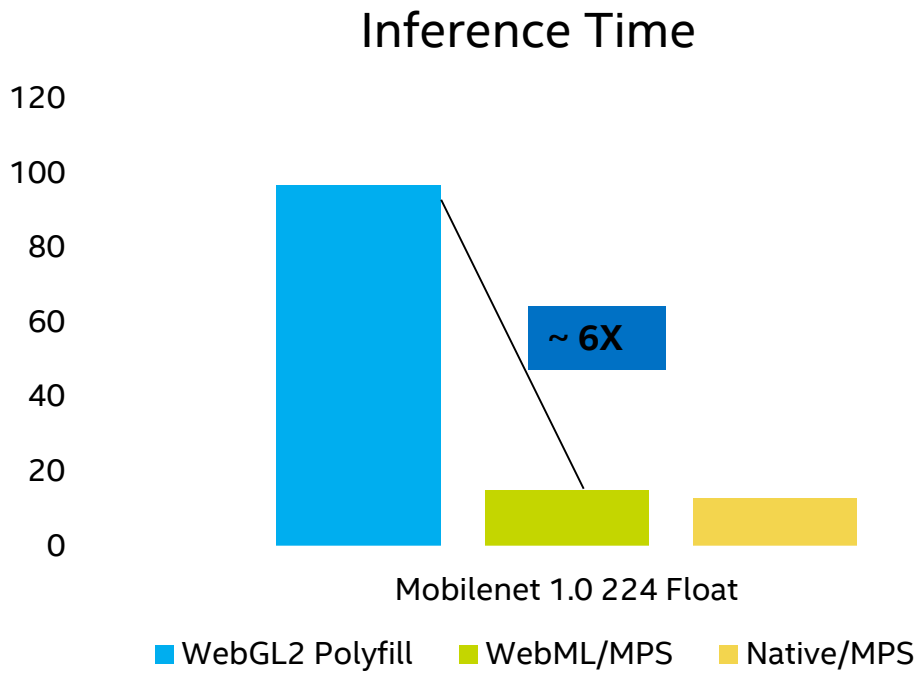
inference time: 14.60 ms

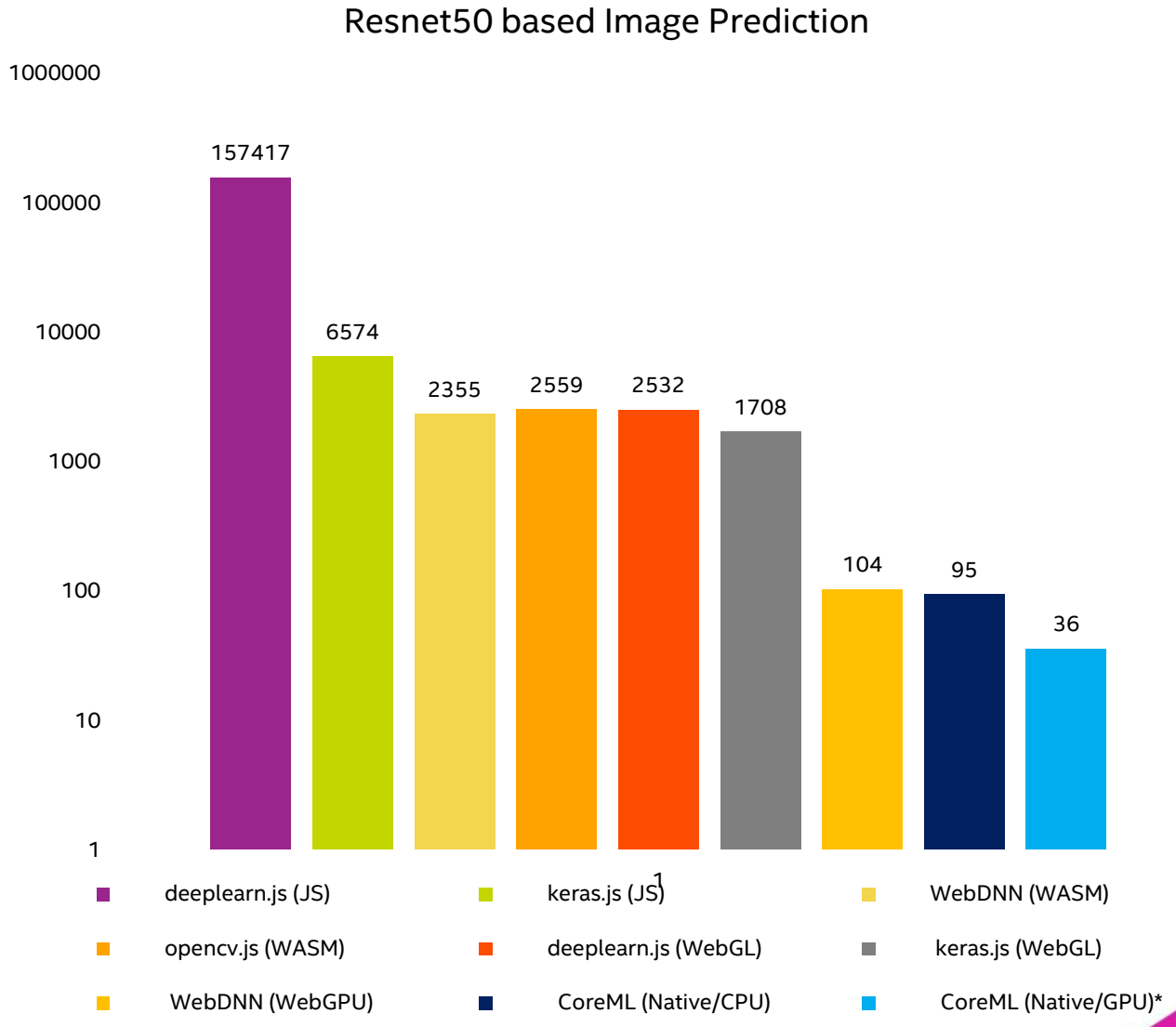
#	Label	Probability
1	computer keyboard, keypad	90.82%
2	space bar	4.89%
3	typewriter keyboard	3.69%

inference time: 14.90 ms

#	Label	Probability
1	toilet tissue, toilet paper, bathroom tissue	54.74%
2	studio couch, day bed	13.84%
3	paper towel	7.07%

PERFORMANCE SUMMARY





- Observed significant speedup on CPU/GPU comparing to existing Web APIs
- Can bring close-to-native performance to Web apps
- Will scale with new dedicated ML hardware accelerators

**See disclaimer section

DEMO

TAKE ADVANTAGE OF INTEL® INTEGRATED GPU COMPUTE POWER TO DO INFERENCE ON MACOS!!!

More info, reference, resources…

Apple Machine Learning:

https:// www.developer.apple.com/machine-learning

Intel Graphics:

https://www.intel.com/content/www/us/en/architecture-and-technology/visual-technology/graphics-overview.html

https://en.wikipedia.org/wiki/Intel_HD,_UHD_and_Iris_Graphics

MPS:

https://developer.apple.com/documentation/metalperformanceshaders

WebML:

https://discourse.wicg.io/t/api-set-for-machine-learning-on-the-web/2491/9

ACKNOWLEDGEMENTS

- Sudhir Tonse (Intel)
- Arzhange Safdarzadeh (Intel)
- Aaftab Munshi (Apple)
- Richard T Trinh (Intel)
- Joseph Van De Water (Intel)
- Sachin Sane (Intel)
- Ningxin Hu (Intel)
- Sriram Murali (Intel)
- many others

LEGAL DISCLAIMER

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at [intel.com].

Results have been estimated or simulated using internal Intel analysis or architectural simulation or modeling and provided to you for information purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

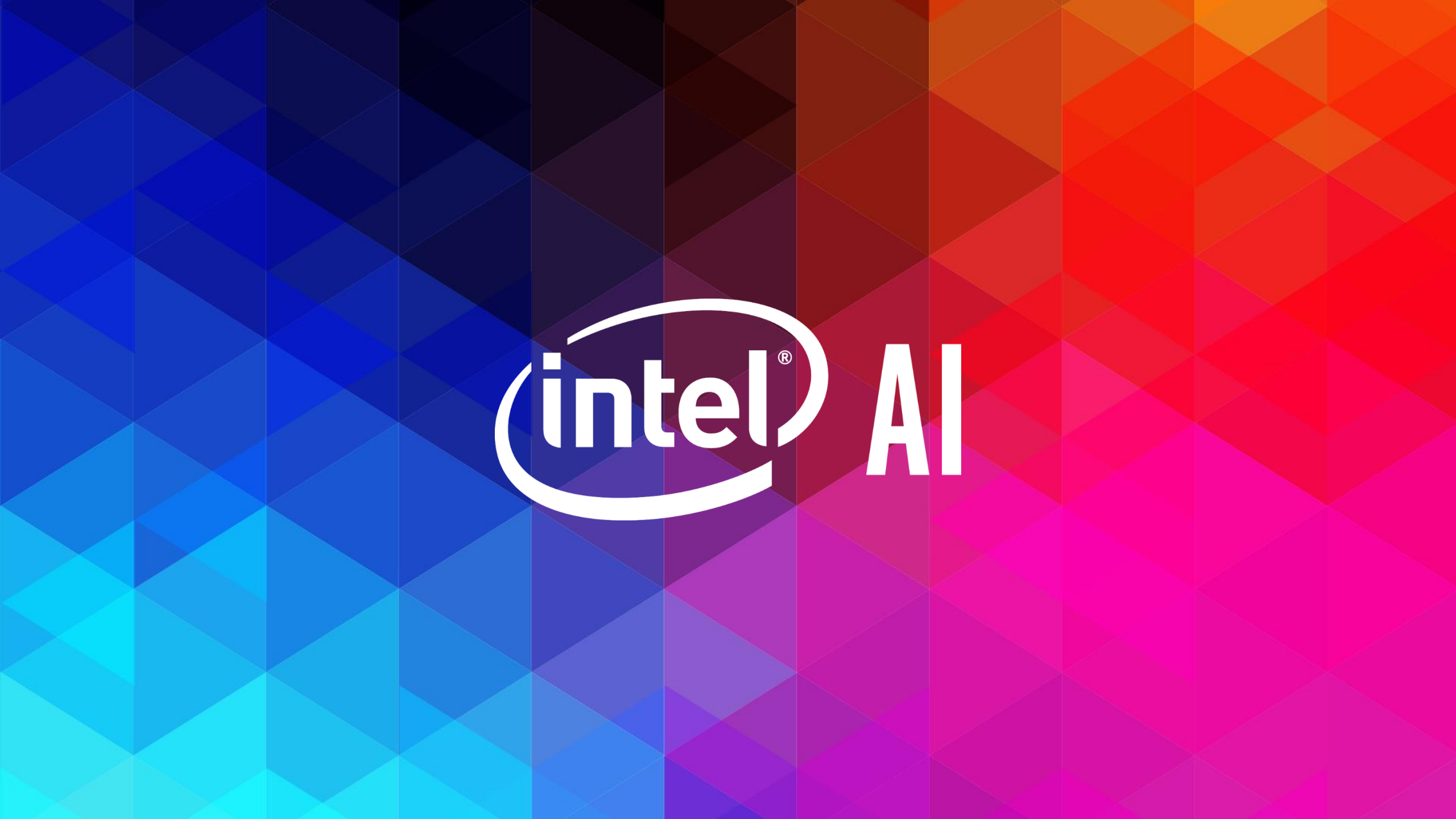
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors and Intel Integrated GPU. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change 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. For more information go to www.intel.com/benchmarks.

Optimization Notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Configurations used for test and perf data: Macbook Pro 13" with Intel Iris Graphics 550, 530 some with fixed 850Mhz frequesncy and some with dynamic frequency. All testing was performed at Intel

Intel, the Intel logo, are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries.
*Other names and brands may be claimed as the property of others.

© Intel Corporation.


intel®
AI