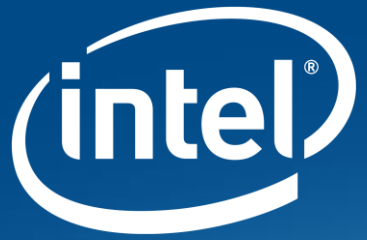


ACCELERATE DEEP LEARNING INFERENCE USING INTEL TECHNOLOGIES

HEALTH AND LIFE SCIENCES APPLICATIONS

Health and Life Sciences

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 or 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. Refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice Revision #20110804.

Legal Notices and Disclaimers (1 of 2)

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 www.intel.com.

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

This document contains information on products, services, and 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.

Any forecasts of goods and services needed for Intel's operations are provided for discussion purposes only. Intel will have no liability to make any purchase in connection with forecasts published in this document.

Arduino* 101 and the Arduino infinity logo are trademarks or registered trademarks of Arduino, LLC.

Altera, Arria, the Arria logo, Intel, the Intel logo, Intel Atom, Intel Core, Intel Nervana, Intel Xeon Phi, Movidius, Saffron, and Xeon are trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

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

Legal Notices and Disclaimers (2 of 2)

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. Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software, or service activation. Learn more at intel.com, or from the OEM or retailer.

No computer system can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit www.intel.com/performance.

Cost-reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Statements in this document that refer to Intel's plans and expectations for the quarter, the year, and the future are forward-looking statements that involve a number of risks and uncertainties.

A detailed discussion of the factors that could affect Intel's results and plans is included in Intel's SEC filings, including the annual report on Form 10-K.

The products described may contain design defects or errors, known as *errata*, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Performance estimates were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown." Implementation of these updates may make these results inapplicable to your device or system.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

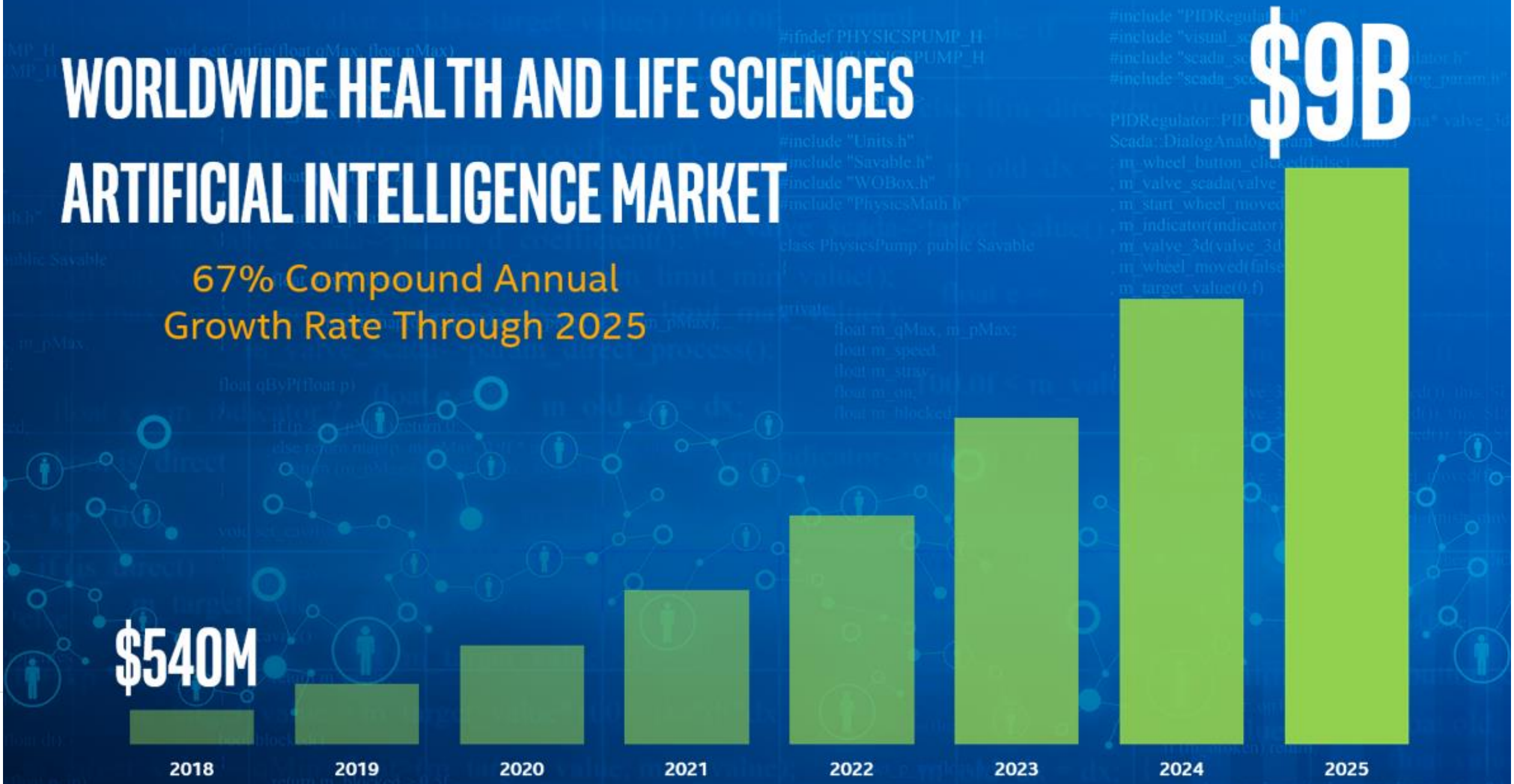
Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

Intel, the Intel logo, Pentium, Celeron, Atom, Core, Xeon, Movidius, Saffron, and others are trademarks of Intel Corporation in the United States and other countries.

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

WORLDWIDE HEALTH AND LIFE SCIENCES ARTIFICIAL INTELLIGENCE MARKET

67% Compound Annual Growth Rate Through 2025

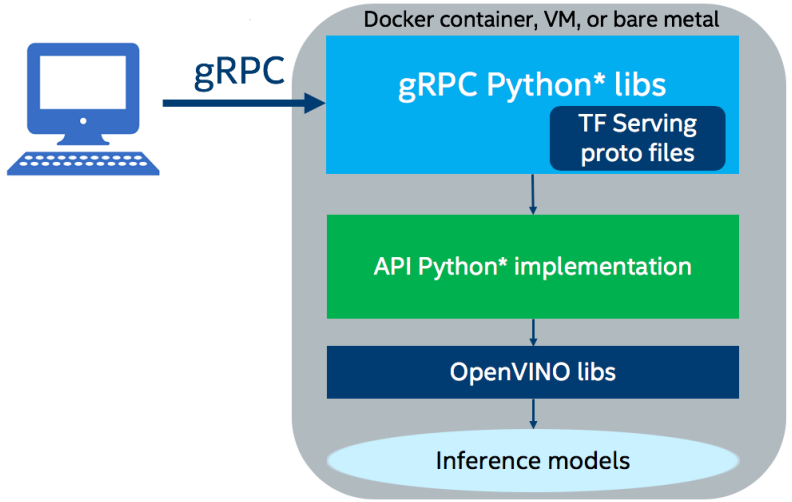


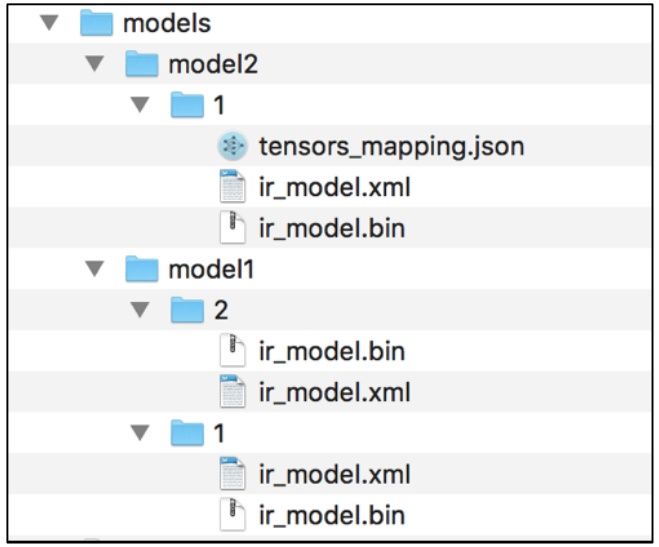
Artificial Intelligence for Enterprise Applications, Tractica, 1Q 2018

OPTIMIZING HLS APPLICATIONS WITH OPENVINO

Source: https://www.intel.ai/openvino-model-server-boosts-ai-inference-operations/#gs.m6x98m

OPENVINO MODEL SERVER





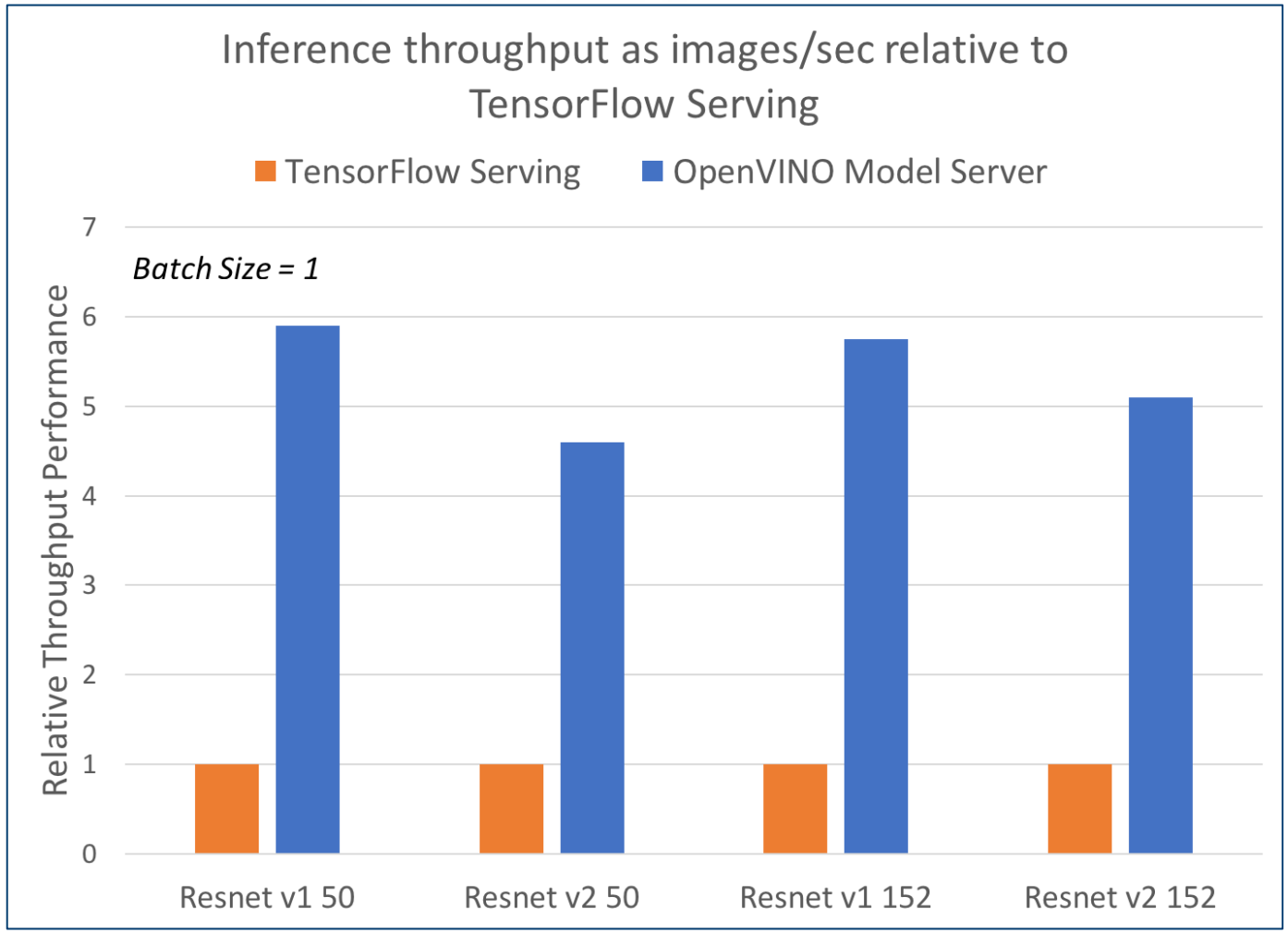
- Same gRPC API as TensorFlow Serving
- Implemented as a Python* service
- Fully compatible with same clients
- Optimized for Intel® CPU, FPGA, VPA
- Suited for Docker containers

https://github.com/IntelAI/OpenVINO-model-server

Advantages for AI Applications

- Support for Multiple Frameworks
- Support for gRPC interface
- Ease of Transition from existing API
- Improved performance
- Support for Intel FPGAs and Intel Movidius VPUs
- Ease of Python service implementation
- Ease of installation and integration

OPENVINO MODEL SERVER

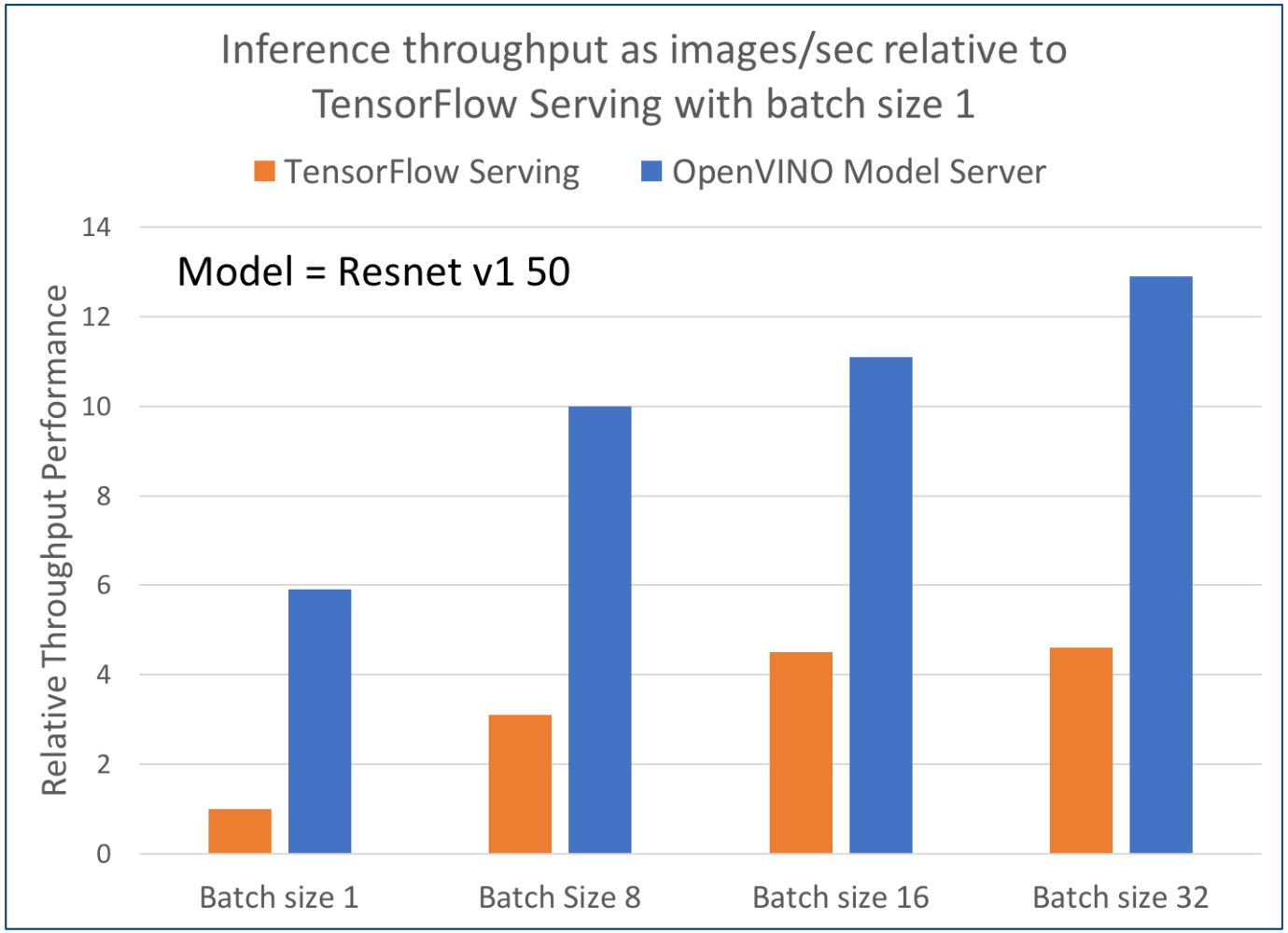


Up to 5x improvement over TensorFlow Serving

Performance results are based on internal testing done on 27th September 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. Test configuration: Dual Intel® Xeon® Platinum 8180 processor @ 2.50GHz, 376.28GB total system memory, Ubuntu-16.04-xenial operating system.

Refer: https://www.intel.ai/openvino-model-server-boosts-ai-inference-operations/#gs.m6x98m

OPENVINO MODEL SERVER



Higher is better.

Performance results are based on internal testing done on 27th September 2018 and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure. Test configuration: Dual Intel® Xeon® Platinum 8180 processor @ 2.50GHz, 376.28GB total system memory, Ubuntu-16.04-xenial operating system.

Refer: https://www.intel.ai/openvino-model-server-boosts-ai-inference-operations/#gs.m6x98m

NETWORK LOADING OPTIMIZATIONS

Reduced model load times for faster performance

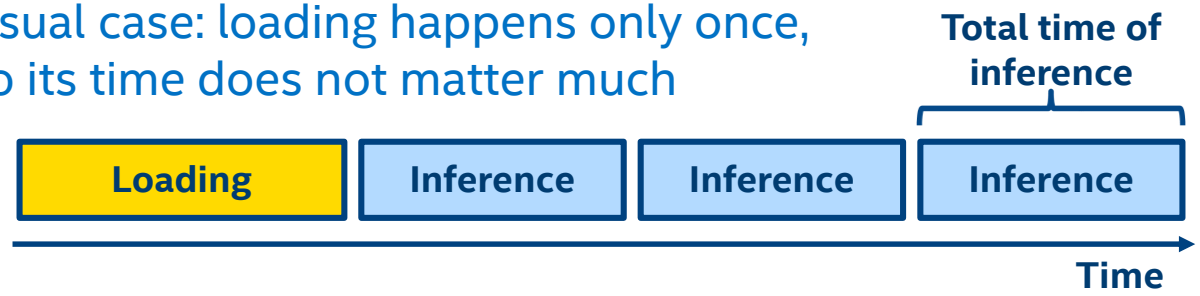
Audience: *Helpful when shape size changes from inference to inference, and resizing is undesirable (e.g., leads to accuracy degradation)*

Problem: *Shape change requires reloading of the model which can be slow*

UseCase: *Input shape is defined by a previous network in the pipeline (i.e., in the case of object detection/classification), or ROI is defined by operator (common case in medical applications)*

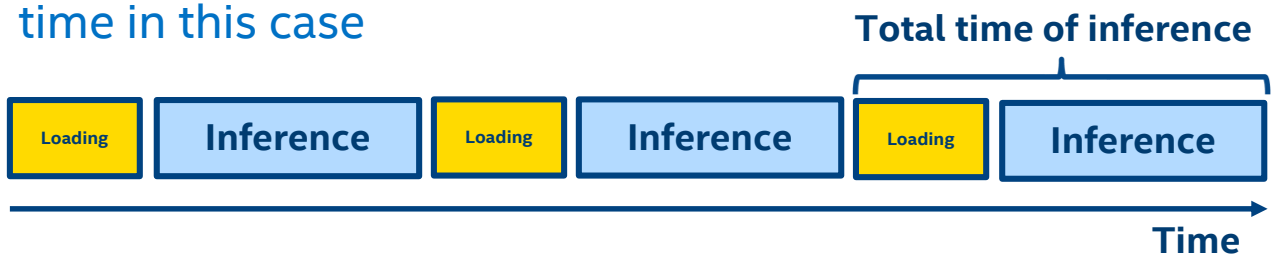
Performance

Usual case: loading happens only once, so its time does not matter much



Changing shape case: loading happens before each inference, so its time matters

Loading time optimization decreases total inference time in this case



CPU
iGPU
VPU
FPGA
GNA

Win 10
Win Serv
Linux
Mac

COMMAND LINE DEPLOYMENT MANAGER

Introducing a new tool to help generate deployment package with customer application and Inference Engine runtime on target device

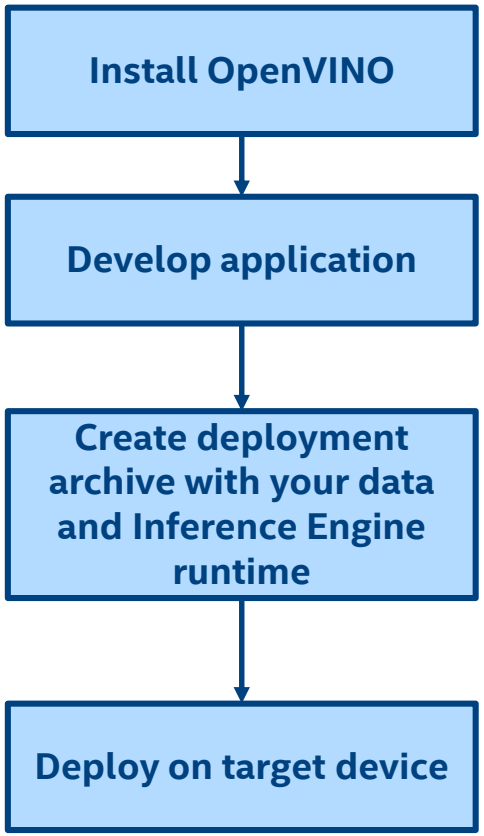
Audience: Developers that want to deploy inference-engine parts with their data (pre-compiled application with required data: models, configs, etc)

Problem: OpenVINO entire package size is too big and usually contains components that are not suitable for particular system. Not all systems have Internet access to mitigate this by online installer usage

UseCase: Generate deployment archives (tarball)

Ease of Use

Smaller deployments



CPU

iGPU

VPU

FPGA

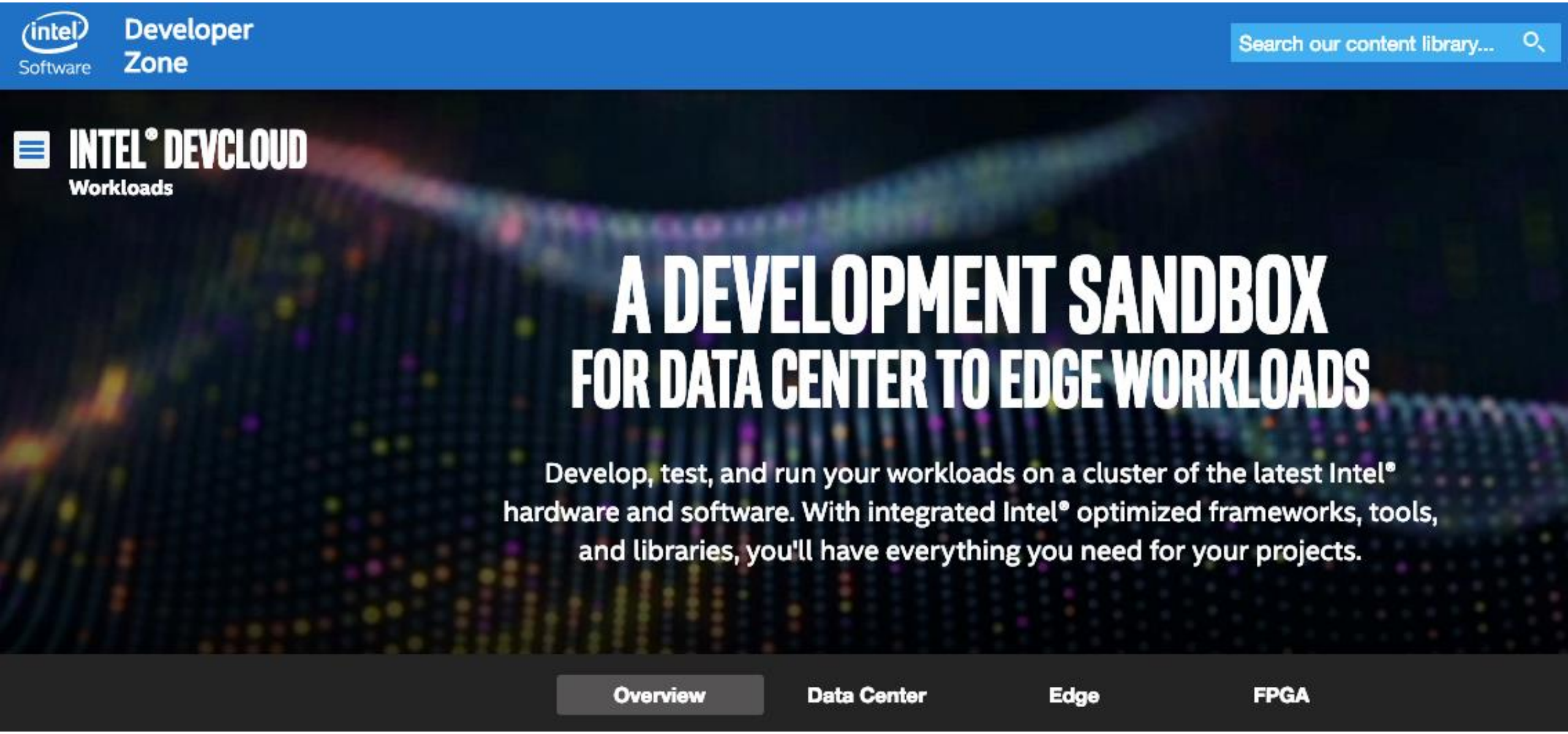
GNA

Win 10

Win Serv

Linux

Mac



software.intel.com/en-us/devcloud

intel

CASE STUDIES

GE CT AXIAL CLASSIFICATION

Source: https://www.intel.ai/wp-content/uploads/sites/69/OpenVINO_GEHealthcare_DLPerf_SolutionBrief.pdf

GE PNEUMOTHORAX DETECTION

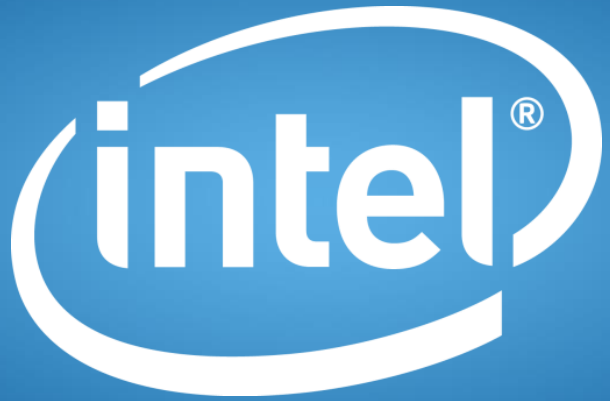
source: https://www.intel.ai/wp-content/uploads/sites/69/Pneumothorax_Whitepaper_GEHC_FINAL-1.pdf

SIEMENS CARDIAC MRI SEGMENTATION

Source: https://www.intel.ai/wp-content/uploads/sites/69/SiemensHealthineers2ndGenXeonScalable.pdf

MAXQ – ACCELERATING STRKE DETECTION

Source: https://builders.intel.com/docs/aibuilders/maxq-ai-automates-and-accelerates-suspected-intracranial-hemorrhage-ich-detection-with-ai-driven-imaging-insight.pdf


intel