



# 5 Steps to an AI Proof of Concept

## A five-step approach to success with proof of concepts (PoC) for image recognition, natural language processing and predictive maintenance

### Table of Contents

Introduction .....	1
Step 1. Confirm the Opportunity.....	2
Step 2. Characterize the Problem and Profile the Data .....	3
Step 3. Architect and Deploy the Solution ...	4
Step 4. Evaluate for Business Value .....	5
Step 5. Scale Up the PoC .....	6
Start small, stay manageable.....	7
References and Resources .....	8

### Introduction – the Role of the PoC in AI

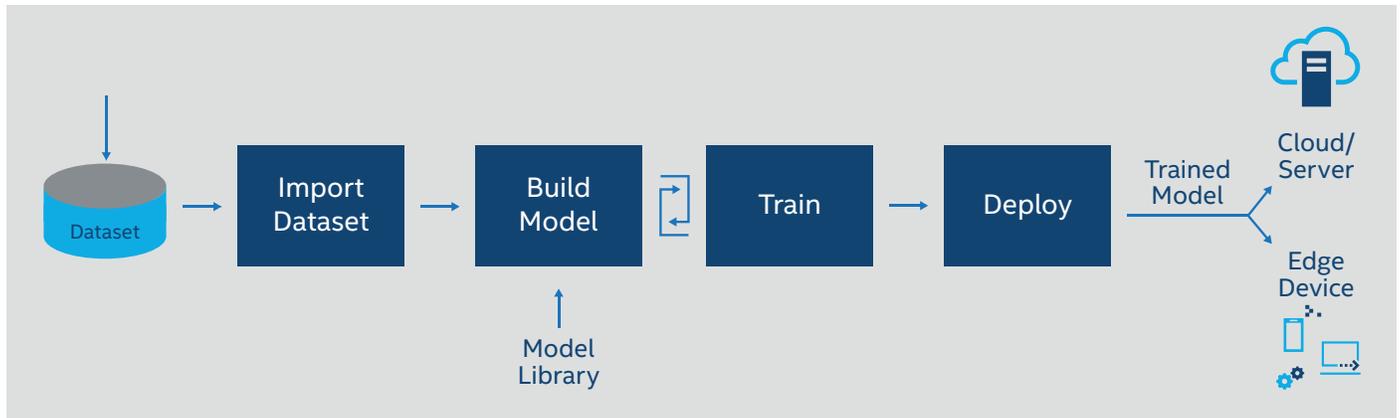
An artificial intelligence (AI) software program is one that can sense, reason, act and adapt. It does so by first 'learning' from a large and diverse data set, which it uses to train models about the data. Once trained, the model is then deployed to infer results from similar, new or unseen data, for example turning verbal speech into text, identifying anomalies in a series of images, or calculating when a piece of machinery is about to fail. We show this sequence in Figure 1.

While AI algorithms have existed for many years, we have recently seen a rapid expansion in AI-based capabilities across the enterprise. This is due to a number of factors, first that processing and data storage costs have fallen at similarly dramatic rates. In parallel, computer scientists have advanced AI algorithm design, including neural networks, leading to greater accuracy in training models.

With AI becoming more prevalent, it has also boosted infrastructure innovation. At Intel, we have been working to embed AI-related features directly into our hardware: [The 2nd Generation Intel® Xeon® Scalable processors offer scalable performance](#) for the widest variety of AI workloads, breakthrough performance in deep-learning model training and inference, and the [Intel® Nervana™ Neural Network Processor](#), incorporates a new architecture built from the ground up for neural networks.

Such advances further accelerate adoption of AI, creating huge opportunities for organizations looking to make smarter decisions and create smarter processes, delivering tangible business benefit. A 2017 study from Accenture, across multiple industries and geographies, found that Artificial Intelligence can increase profitability by 38 percent, generating over \$14 trillion dollars of economic impact in the coming decades<sup>1</sup>.

<sup>1</sup> [https://www.accenture.com/t20170620T055506\\_\\_w\\_\\_/us-en/\\_acnmedia/Accenture/next-gen-5/insight-ai-industry-growth/pdf/Accenture-AI-Industry-Growth-Full-Report.pdf?la=en](https://www.accenture.com/t20170620T055506__w__/us-en/_acnmedia/Accenture/next-gen-5/insight-ai-industry-growth/pdf/Accenture-AI-Industry-Growth-Full-Report.pdf?la=en)



**Figure 1.** AI systems learn, and then infer results, from data

Despite this clear potential, many organizations are yet to get started with AI and adoption is often not necessarily happening as fast as many reports from the media and academia might suggest<sup>2</sup>. As enterprises look to begin their AI journeys, the more common use cases exist around natural language processing (NLP), computer vision and predictive maintenance. See customer examples in table on the next page.

Example opportunities for deep and machine learning with these use cases include:

- **Vertical:** Many organizations are looking to solve challenges specific to their industries, for example manufacturing process and spares management, retail inventory management and patient outcomes in healthcare.
- **Line of Business:** Across industries, corporations will have similar needs depending on individual lines of business. For example, natural language processing has applications in customer service departments, and image recognition and predictive maintenance have relevance to supply chain applications.
- **Technology Architecture:** Many examples of AI that we come across have similar architectures, even if they use different data pools and deliver different results. For example, the image processing and anomaly detection used by one customer to detect solar panel defects, can be based on a similar platform to that which conservationists might use to 'listen' for behavioral changes in bats.
- **IT-related:** Some applications of AI can exist across applications and services, because they are about managing data flows, pre-empting bottlenecks, predicting faults and responding quickly to failures and breaches.

The wide variety of potential opportunities creates a number of challenges – which opportunities will yield the best results, and how to ensure a successful outcome? The role of the Proof of Concept (PoC) is to enable decision makers to answer these questions while maximizing value and minimizing risk.

<sup>2</sup> <https://www.gartner.com/newsroom/id/3856163>

### What is a Proof of Concept?

A proof of concept (POC) is a 'closed' but working solution which can be evaluated and tested subject to clear criteria, from understanding requirements to delivering success. For any AI project or program, PoCs enable decision makers to:

- Deliver more immediate value
- Gain skills and experience
- Test hardware, software and service options
- Identify and resolve potential data bottlenecks
- Highlight impacts on IT infrastructure and the wider business
- Raise the positive profile of AI and grow user trust

### Step 1. Confirm the Opportunities

It is vital to be clear from the outset on what you are looking to achieve with AI, why it matters to your enterprise, and how you can be sure it will deliver. If you have not yet identified the primary opportunities to benefit from AI, then you should assess where AI can make the most immediate difference:

- Consider what others in your industry are doing with AI.
- Look for areas of your business that have a clear problem to be solved or value to be gained from AI.
- Work with existing pools of expertise, using the skills and experience you already have in-house.

Having identified a shortlist of areas where AI might benefit your organization, you can test each opportunity against several criteria. This review does not need to take long, but the following questions can identify gaps in planning and guard against the temptation to rush into an AI project:

Natural Language Processing	Computer Vision	Predictive Maintenance
<p>When FinTech start-up Clinc decided to build Finie, an AI Personal Assistant App designed to help people interact with their personal finances using natural language, it realized that existing natural language algorithms weren't sufficient to deliver the customer experiences it was looking for. <a href="#">In collaboration with Intel</a>, Clinc used the latest in machine learning and deep learning technologies to create a customer-facing AI solution.</p>	<p>Gourmet candy retailer Lolli &amp; Pops uses computer vision and AI to provide personalized customer experiences. Through computer vision, Lolli &amp; Pops "Magic Makers**" recognizes loyalty scheme members as they enter the store. <a href="#">Using AI-enhanced analytics</a>, the retailer accesses members' preferences and makes personalized recommendations – giving shoppers VIP treatment, ensuring they keep coming back.</p>	<p>Deutsche Telekom uses SAP solutions running on Intel® Xeon® processor E7 Family processor-based cloud servers to collect performance, temperature, vibration, or rotation sensor data and perform predictive analytics. <a href="#">This drives the organization's predictive maintenance</a>, proactively reducing machine downtime and maintenance costs by identifying fatigued or worn parts before major damage occurs.</p>

- Is it clear what problem you are looking to solve, its specific requirements and how you can measure success? Have you already considered or deployed other solutions to deal with this problem, and ruled them out in favor of AI?
- Is the scope of the opportunity well-bounded? For example, can you draw a simple picture covering the data set it uses, key components, the people it will affect and other dependencies? Will it be part of a larger solution?
- Do you have the technology resources and funding you need to make this happen? Can you access the data sources you need, without technological, contractual or other impediments?
- Is the business impact significant enough to merit the effort? High visibility wins are important to grow user trust in AI and for broader stakeholder engagement.
- Is impetus and buy-in sufficient, for example in the form of executive sponsorship? Is the affected line of business fully invested in solving this problem?
- Are timescales appropriate? Is the delivery team clearly defined, with sufficient time, skills and motivation to make it happen?
- Does the organization have a wider data science and/or AI strategy, and does this align with its goals? What data science infrastructure and expertise does the organization already have?
- What is the plan following a successful PoC – does funding exist to maintain or scale the solution? Is your operational IT department briefed and ready to participate?

Ultimately these questions test for solution value, cost and risk, which can be characterized as a business case – though a formal document may be too much for a simpler PoC.

For a broader view of readiness for AI, consider reading [The AI Readiness Model white paper](#).

## Step 2. Characterize the Problem and Profile the Data

After you have identified and tested your opportunity, you can turn your attention to understanding and articulating

the problem to be solved in more detail, mapping it to broad categories such as reasoning, perception or computer vision.

Part of the challenge, particularly for those at earlier stages of their AI journey, is having sufficient skills in-house. Intel helps companies through its technical experts and consulting partners, and it also provides training courses. These include [12-week self-directed primers on Machine Learning and Deep Learning](#), designed to help build developers' understanding of how to map the business problems to relevant Intel AI Technologies.

In your AI workflow, this is also a good time to ask a number of more technical questions which may influence the solution. For example:

- Do you favor any hardware/software and why (benchmark data, TCO, preferred supplier)?
- Do security/regulatory/data/other needs favor on-premise systems versus cloud?
- Will your solution be self-serviced locally or provisioned in the data center?
- What is the current data center percentage percentage-of-utilization, and how important is per-watt performance?
- What cadence and quantity of new data will you get for training/inference?
- How will both raw data and resulting insights be kept secure, at rest and in motion?

## Step 3: Architect and Deploy the Solution

The next question is how to design and deploy the solution being tested in the PoC. As shown in Figure 3, this will consist of a stack of technologies, including:

- Underlying products and systems infrastructure
- AI-specific software to drive the infrastructure
- Enabling AI frameworks to support the planned solution
- Visualization and front-end software and/or hardware

At this stage, you may be wondering whether to build, buy or reuse hardware and software, and/or whether to make use of cloud services. We set out the options for build versus

buy and so on, in our companion white paper [Select the Best Infrastructure Strategy to Support Your AI Solution](#). Customer investments in market-leading Intel® Xeon® processors mean that, in many cases, initial exploratory testing can take place using existing infrastructure.

Infrastructure and software that is built and tested in line with best practice still requires taking the requirements of AI into account. In particular, this includes the need for a constant feed of high-quality data. Data scientists can work in partnership with IT systems architects to design the deployment architecture from the data center to the edge, taking into account software integration, network connectivity, physical issues and other aspects. Multiple options may need to be tested: this should be encouraged using a test-and-learn approach such so maximum experience is gained.

When completed, you can work through other AI-specific elements of the solution – constructing the models, training and tuning.

### Constructing the models

Model construction is the core AI task. It involves data scientists using training data and managing parameters to conduct iterative test runs. In this way, they can check models for initial convergence accuracy before sending them for broader training and tuning.

### Training and tuning

Training and tuning is the most computationally intensive part of the AI workflow. Here, data scientists determine under what parameters their models converge most efficiently given the available training data, while dealing with traditional IT concerns of job scheduling and infrastructure management.

This is highly labor intensive, with data scientists spending their time manually wrangling data and executing hundreds of experiments. This can also be made easier by the [Intel® Nervana™ Deep Learning Studio](#) – a comprehensive software suite, the solution enables groups of data scientists to reduce these testing cycles, develop and deploy custom, enterprise-grade deep learning solutions in record time.

## Step 4. Evaluate for Business Value

You will have defined evaluation criteria for the PoC as part of solution design: to engineers, these can be translated into evaluation criteria that can be designed, measured, and continuously tested, preferably in an automated manner.

The following evaluation criteria may be applied around business value:

## Can I use a standard CPU for AI?

While Graphics Processing Units have played a role in advancing the kind of algorithmic processing involved in AI, deep learning (DL) is now practical on general purpose, CPU-based architectures.

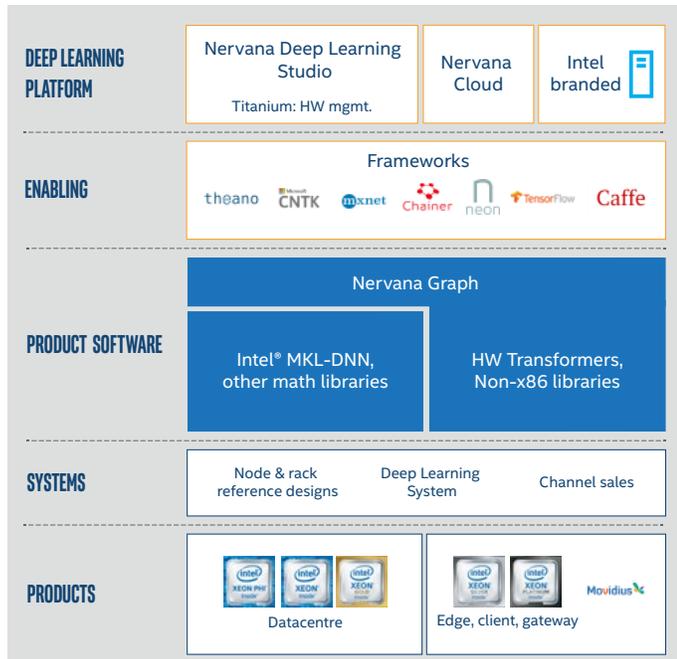
In the past, deep learning training on CPU took an unreasonably long time because processors lacked hardware and especially software optimizations. That is no longer the case. The 2nd Generation of Intel® Xeon® Scalable processors have dramatically narrowed the performance gap. Intel® Xeon® Scalable processor performance executing deep learning tasks has increased by up to 127-fold for training throughput compared to the previous generation without optimized software<sup>3</sup>.

In addition, Intel® Xeon® Scalable processors can scale out very efficiently to achieve almost any deep-learning throughput profile. The ability to use CPUs addresses a number of challenges organizations face with GPU-only based AI:

- GPU architecture requires the data pipeline to be temporarily copied to a GPU data store then back, which breaks the typical data flow and processing tool chain.
- Compared to CPU-based nodes, it can be difficult to scale and manage computation on a large number of GPU-based nodes in non-cluster mode, reducing potential time-to-train savings.
- Memory constraints can exist, particularly for organizations wanting to process very large images within the small memory footprint of a GPU (16 or 32GB), for example in healthcare and geospatial applications.
- Underutilization can occur with any domain-specific architecture. With a general-purpose CPU, idle nodes can be used for other workloads and/or rented as IaaS.

An increasing number of organizations are recognizing the benefits of CPUs for deep learning. Intel is working with customers including Facebook, deepsense.ai, OpenAI, AWS, EMR, Databricks, Alibaba, Microsoft and Cloudera. This list will grow further as the gap between CPU and GPU AI performance closes.

For more information on how Intel technology can provide the foundation for your AI PoC, see [The Anatomy of an AI PoC infographic](#).



**Figure 3.** The AI solution architecture can be represented as a stack

- **Accuracy:** Is the solution delivering results and insights correctly and are they repeatable?
- **Completeness:** Is the solution making correct use of all data sources?
- **Timeliness:** Are insights delivered to the point of need, at the time of need?

Additional criteria apply to the solution, and whether it works as expected:

- **Scale:** Will the solution continue to function if data volumes or user numbers grow over time, or in bursts?
- **Compatibility:** Is the solution open to integration with third-party data sources and services, using standard protocols?
- **Flexibility:** Can the solution adapt to changing circumstances, should the data needs or models change?
- **Engineering:** How straightforward is it to debug incorrect outputs from a trained model?

Finally, the solution needs to be evaluated based on what is known in AI circles as ‘explainability’, that is, decision making quality. Criteria around explainability include:

- **Bias:** How to ensure the AI system does not have a biased view of the world (or perhaps an unbiased view of a biased world) based on shortcomings of the training data, model, or objective function? What if its human creators harbor a conscious or unconscious bias?

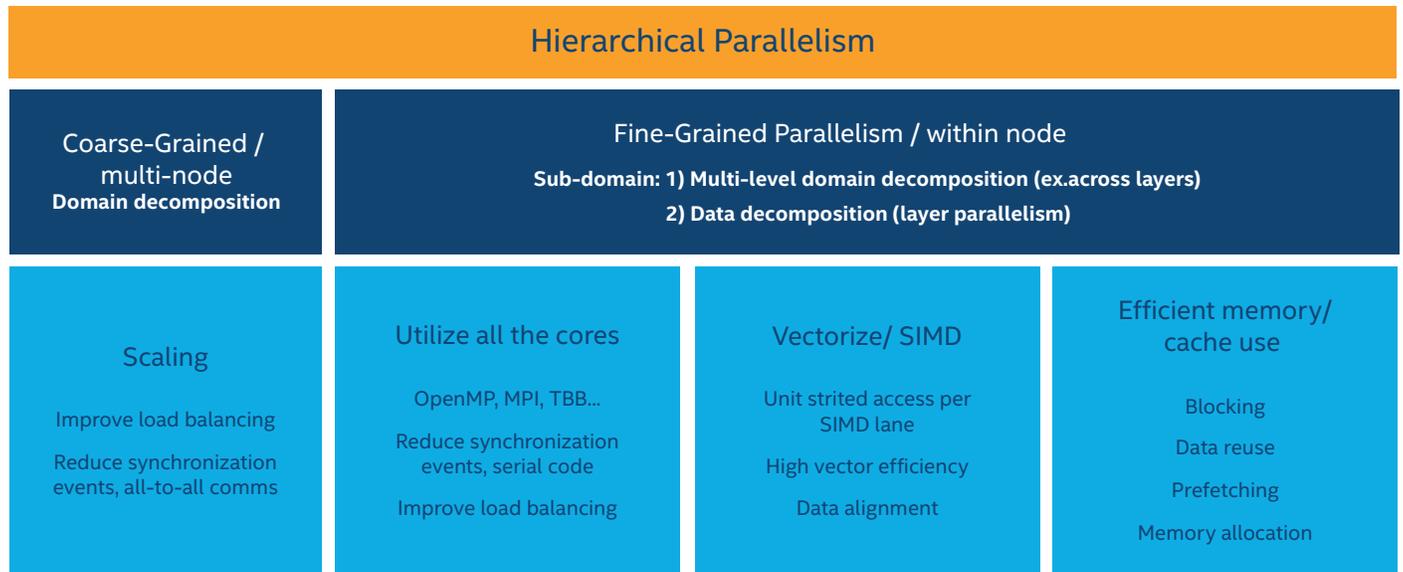
- **Fairness:** If decisions are made based on an AI system, how can it be verified that they were made fairly? And what does fair mean in this context – fair for who?
- **Causality:** Can the model provide not only correct inferences, but also some explanation for the underlying phenomena?
- **Transparency:** Are AI-based insights explained in terms the user can understand? And on what basis can a finding be questioned?
- **Safety:** How will users gain confidence in the reliability of the AI system, with or without transparency on how it reaches conclusions?

### Step 5. Scale Up the PoC

The problem has been defined, the solution designed, the data profiled and modelled. The PoC has been built, tested and deployed. So, what next?

Positive experiences among users can lead to greater demand and therefore higher levels of success. But they also risk the PoC becoming a victim of too much interest. You can do a number of things to ensure that your PoC remains a success story, enabling it to be built upon in support of a broader AI strategy:

- **Scale up inference capabilities.** AI does not scale in a linear fashion – when moving from a single-node configuration for example, 50 processors will not necessarily result in 50 times the performance. You will need to test and optimize a multi-node configuration in much the same way as you have done in your single-node configuration.
- **Scale up broader infrastructure.** AI success requires you to examine every link in the chain of inference. Review existing technology platforms, networks and storage with an aim to increase the amount of data available, its timeliness and latency. This will minimize the potential for future bottlenecks while maximizing the value you can derive from your data sources.
- **Tune and optimize the PoC Solution.** As time goes on, you will develop more skills for improving and enhancing the AI solution you have deployed. You can optimize software around areas such as data curation and labelling, and can experiment with, train and deploy new models that may give better results.
- **Scale out to other business scenarios.** Your PoC may have applications in other parts of your business, for example a predictive maintenance solution may have been deployed for one area of your manufacturing and you can now be broadened. You can adopt a portfolio approach to manage how you extend the PoC across a wider user base.
- **Plan for management and operations.** By their nature, many AI use cases require the systems to perform inference in real-time, rather than offline or batch mode. In addition, models may need to be re-trained and updated over time.



**Figure 4.** Performance optimization on modern platforms

These factors will put additional requirements on service delivery. Ensure sufficient time and skilled resource is pre-allocated, so that the PoC can continue to deliver.

Figure 4 shows a number of areas you can look at to further optimize your AI solution. At Intel, we've been directly [optimizing the most popular AI frameworks](#) for Intel® architecture and producing significant performance increases. These include Theano\* and TensorFlow\*, and we intend to enable even more frameworks in the future through the [Intel® nGraph™ Compiler](#)<sup>4</sup>.

Additionally, [BigDL](#) was created by Intel to bring deep learning to big data. It is a distributed deep learning library for Apache Spark\* that can run directly on top of existing Spark or Apache Hadoop\* clusters, and allows your development teams to write deep learning applications as Scala or Python programs.

### Start small, stay manageable

To maximize the chances of success and deliver value quickly, we recommend starting small and manageable, making sure objectives are clear and business-focused from the outset.

At Intel, we are fully committed to helping our customers deliver on the potential of AI with:

- **Solutions** – Intel's data scientists, technical services and reference solutions teams develop, apply and share AI solutions to expedite your journey from data to insight.
- **Platforms** – Intel offers several turnkey, full stack and user-friendly systems that can be quickly deployed to accelerate the AI innovation cycle.

- **Tools** – Intel's AI software suite features productivity tools for data scientists and developers that compress the deep learning innovation cycle.
- **Frameworks** – Intel is optimizing the most popular open-source community frameworks for deep learning to deliver peak performance across a range of processor platforms.
- **Libraries** – Intel is accelerating AI applications by optimizing primitives and creating the Intel® nGraph™ Compiler, to enable frameworks to use any target hardware with peak performance.
- **Hardware** – Intel's comprehensive product portfolio spans the data center to the edge and addresses all current AI approaches.

Intel is enabling its strong ecosystem and partner network to accelerate AI progress through wide industry collaboration. Like the organizations we serve, we are on a journey, pushing the forefront of AI computing through cutting-edge R&D into areas such as neuromorphic and quantum computing.

This is only the beginning.

#### Learn More

- **More Information at:** [ai.intel.com](https://ai.intel.com)
- **Explore - Use Intel's performance:** [optimized libraries & frameworks](#)
- **Engage - Contact your Intel representative for help and POC opportunities**

<sup>4</sup> Please note that each framework has a varying degree of optimization and configuration protocols, so visit [ai.intel.com/framework-optimizations/](https://ai.intel.com/framework-optimizations/) for full details.

## References and Resources

Intel publishes case studies, reference solutions and reference architectures on [ai.intel.com](https://ai.intel.com) for customers to use in scoping and building their own similar AI solutions.

The Challenges and Opportunities of Explainable AI

<https://ai.intel.com/the-challenges-and-opportunities-of-explainable-ai/>

The Future of Retail is All About Artificial Intelligence <https://ai.intel.com/future-retail-artificial-intelligence>

Intel AI academy – learn the basics <https://software.intel.com/en-us/ai-academy/basics>

Loihi – Intel's New Self-Learning Chip Promises to Accelerate Artificial Intelligence <https://newsroom.intel.com/editorials/intels-new-self-learning-chip-promises-accelerate-artificial-intelligence/>

How to Get Started as a Developer in AI, Singh, Niven, <https://software.intel.com/en-us/articles/how-to-get-started-as-a-developer-in-ai>

Your AI Personal Assistant for Finance and Banking, <https://www.intel.com/content/www/us/en/analytics/artificial-intelligence/ai-personal-assistant.html>

Predictive Analytics Helps Reduce Machine Maintenance Costs, <https://www.intel.co.uk/content/www/uk/en/big-data/intel-sap-telekom-predictive-analytics-paper.html>



<sup>3</sup> INFERENCE using FP32 Batch Size Caffe GoogleNet v1 256 AlexNet 256.

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. 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. For more complete information visit <http://www.intel.com/performance> Source: Intel measured as of June 2017 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 for Inference throughput:

Processor :2 socket Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores HT ON , Turbo ON Total Memory 376.46GB (12slots / 32 GB / 2666 MHz).CentOS Linux-7.3.1611-Core , SSD sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB , Deep Learning Framework caffe version: f6d01efbe93f70726ea3796a4b89c612365a6341 Topology :googlenet\_v1 BI OS:SE5C620.86B.00.01.0004.071220170215 MKLDNN: version: ae00102be506ed0fe2099c6557df2aa88ad57ec1 NoDataLayer. Measured: 1190 imgs/sec vs Platform: 2S Intel® Xeon® CPU E5-2699 v3 @ 2.30GHz (18 cores), HT enabled, turbo disabled, scaling governor set to "performance" via intel\_pstate driver, 256GB DDR4-2133 ECC RAM. CentOS Linux release 7.3.1611 (Core), Linux kernel 3.10.0-514.el7.x86\_64. OS drive: Seagate\* Enterprise ST2000NX0253 2 TB 2.5" Internal Hard Drive. Performance measured with: Environment variables: KMP\_AFFINITY='granularity=fine,compact,1,0', OMP\_NUM\_THREADS=36, CPU Freq set with cpupower frequency-set -d 2.3G -u 2.3G -g performance. Deep Learning Frameworks: Intel Caffe: (<http://github.com/intel/caffe/>), revision b0ef3236528a2c7d2988f249d347d5fdae831236. Inference measured with "caffe time --forward\_only" command, training measured with "caffe time" command. For "ConvNet" topologies, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs from [https://github.com/intel/caffe/tree/master/models/intel\\_optimized\\_models](https://github.com/intel/caffe/tree/master/models/intel_optimized_models) (GoogLeNet, AlexNet, and ResNet-50), [https://github.com/intel/caffe/tree/master/models/default\\_vgg\\_19](https://github.com/intel/caffe/tree/master/models/default_vgg_19) (VGG-19), and [https://github.com/soumith/convnet-benchmarks/tree/master/caffe/imagenet\\_winners](https://github.com/soumith/convnet-benchmarks/tree/master/caffe/imagenet_winners) (ConvNet benchmarks; files were updated to use newer Caffe prototxt format but are functionally equivalent). GCC 4.8.5, MKLML version 2017.0.2.20170110. BVLC-Caffe: <https://github.com/BVLC/caffe>, Inference & Training measured with "caffe time" command. For "ConvNet" topologies, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training. BVLC Caffe (<http://github.com/BVLC/caffe>), revision 91b09280f5233cafc62954c98ce8bc4c204e7475 (commit date 5/14/2017). BLAS: atlas ver. 3.10.1.

#### Configuration for training throughput:

Processor :2 socket Intel(R) Xeon(R) Platinum 8180 CPU @ 2.50GHz / 28 cores HT ON , Turbo ON Total Memory 376.28GB (12slots / 32 GB / 2666 MHz).CentOS Linux-7.3.1611-Core , SSD sda RS3WC080 HDD 744.1GB,sdb RS3WC080 HDD 1.5TB,sdc RS3WC080 HDD 5.5TB , Deep Learning Framework caffe version: f6d01efbe93f70726ea3796a4b89c612365a6341 Topology :alexnet\_BIOS:SE5C620.86B.00.01.0009.101920170742 MKLDNN: version: ae00102be506ed0fe2099c6557df2aa88ad57ec1 NoDataLayer. Measured: 1023 imgs/sec vs Platform: 2S Intel® Xeon® CPU E5-2699 v3 @ 2.30GHz (18 cores), HT enabled, turbo disabled, scaling governor set to "performance" via intel\_pstate driver, 256GB DDR4-2133 ECC RAM. CentOS Linux release 7.3.1611 (Core), Linux kernel 3.10.0-514.el7.x86\_64. OS drive: Seagate\* Enterprise ST2000NX0253 2 TB 2.5" Internal Hard Drive. Performance measured with: Environment variables: KMP\_AFFINITY='granularity=fine,compact,1,0', OMP\_NUM\_THREADS=36, CPU Freq set with cpupower frequency-set -d 2.3G -u 2.3G -g performance. Deep Learning Frameworks: Intel Caffe: (<http://github.com/intel/caffe/>), revision b0ef3236528a2c7d2988f249d347d5fdae831236. Inference measured with "caffe time --forward\_only" command, training measured with "caffe time" command. For "ConvNet" topologies, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs from [https://github.com/intel/caffe/tree/master/models/intel\\_optimized\\_models](https://github.com/intel/caffe/tree/master/models/intel_optimized_models) (GoogLeNet, AlexNet, and ResNet-50), [https://github.com/intel/caffe/tree/master/models/default\\_vgg\\_19](https://github.com/intel/caffe/tree/master/models/default_vgg_19) (VGG-19), and [https://github.com/soumith/convnet-benchmarks/tree/master/caffe/imagenet\\_winners](https://github.com/soumith/convnet-benchmarks/tree/master/caffe/imagenet_winners) (ConvNet benchmarks; files were updated to use newer Caffe prototxt format but are functionally equivalent). GCC 4.8.5, MKLML version 2017.0.2.20170110. BVLC-Caffe: <https://github.com/BVLC/caffe>, Inference & Training measured with "caffe time" command. For "ConvNet" topologies, dummy dataset was used. For other topologies, data was stored on local storage and cached in memory before training. BVLC Caffe (<http://github.com/BVLC/caffe>), revision 91b09280f5233cafc62954c98ce8bc4c204e7475 (commit date 5/14/2017). BLAS: atlas ver. 3.10.1.

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](http://intel.com)

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. For more complete information visit [www.intel.com/benchmarks](http://www.intel.com/benchmarks)

Estimated results 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. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

Intel, Xeon, Nervana, nGraph, and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.

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

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](http://intel.com)

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. For more complete information visit [www.intel.com/benchmarks](http://www.intel.com/benchmarks)

Estimated results 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.

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

Intel, Xeon, Nervana, nGraph, and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.

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

© Intel Corporation  
0318/RD/CAT/PDF  
337357-001EN