Product Brief

Optimized AI Cloud Services

Accelerate & Scale AI in Intel® Developer Cloud

Accelerate AI with Access to the Latest Intel Hardware & Software Innovations

When implementing AI-driven solutions, optimizing the performance and ROI of your hardware and software can be challenging. The ideal solution should offer all that plus flexibility, choice, and end-to-end workflow optimization. You also want assurance that your investments will offer continued value during future transitions to new architectures. Intel® Developer Cloud is designed to help developers and companies accelerate, optimize, and scale AI.

To help developers and companies accelerate AI, Intel® Developer Cloud provides a development environment with the latest Intel hardware and software innovations to build and test AI, machine learning, HPC and security applications for cloud, enterprise, client, and edge deployments. There are three service tiers: a standard free, a premium, and an enterprise paid service tier.

More Value for Your AI

Maximize value for AI compute from current and next gen hardware systems. Gain performance and productivity from software optimizations.

- **Choose the best accelerator for every use case:** Build and test applications and workloads on the most advanced CPUs designed for machine learning, GPUs, and Intel® Gaudi®2 AI accelerators.
- **Scale with ease and flexibility:** Run small and large-scale LLM and generative AI training, model optimization, and inference workloads. Utilize virtual machines, full systems or clusters.
- **Stay one step ahead with pre-release hardware:** Start using Intel processors and platforms up to a year before they're generally available.
- **Get the freedom to develop and deploy anywhere:** With oneAPI's open, standards-based programming model, you get the flexibility to choose the best hardware for your workloads, from any vendor, using the same code base.
- **Boost performance with AI software accelerators:** Intel Developer Cloud includes built-in access to Intel-optimized versions of popular frameworks and tools for every stage of the AI workflow to deliver 10-100X performance gains using the tools you already know.[1]
- **Get gen AI solutions to market faster:** Intel Developer Cloud supports a wide range of open source AI foundational models, so you can build LLMs and other generative AI products using HuggingFace, Stable Diffusion, Meta AI Llama 2, and more.

AI, Optimized

Build and deploy AI at any scale on managed, cost-effective infrastructure with Intel® Developer Cloud and cnvrg.io. cnvrg.io's MLOps platform streamlines the whole model lifecycle, so you can focus on deploying better models for the business. Support for Intel Developer Cloud is built-in, while re-usable software templates make it easy to create different cluster sizes and start developing.

intel Developer Cloud

cnvrg.io

Test, Build & Deploy AI Easier–Key Usages

Customers and developers can try many of Intel's latest platforms to determine which accelerator is best for their specific needs. More than 2,000 users spanning developers, ISVs, academics, and enterprise companies are already using Intel Developer Cloud.

For developers, Intel Developer Cloud provides an easy path to access and use Intel-optimized AI software on Intel hardware. Common usages include architecture evaluation, application development and optimization, model and workload optimization, research and academia learning, education/ training for oneAPI and LLM/MLOps, plus AI development and design certification. You can try out LLM workload code samples to see how they perform on Intel architecture.

The cloud includes Jupyter notebooks for easy development and supports Visual Studio code.

For companies and enterprise, system integrators (SIs), independent software vendors (ISVs), and 3rd party SaaS organizations, usages include running and testing AI training and inference production workloads at scale, certification and benchmarking, and Intel compute services for third-party AI SaaS.

For AI compute, run small and large-scale AI training (LLMs or generative AI), model optimization and inference workloads. Utilize small to large virtual machines (VMs), full systems or clusters with Intel GPUs, CPUs, and Habana Gaudi2 accelerator systems. Scale from 7 to hundreds of billions of parameters.

The table below outlines the different service and support tiers offered for Intel Developer Cloud.

Intel® Developer Cloud Services & Support Options

	Standard	Premium	Enterprise
Users	AI & HPC developers, data scientists, researchers, academia Single-user access	Enterprise customers/developers Single-user access	Enterprise customers Multi-user access
Usages	▪ Evaluation—test applications, workloads, & LLM workload code samples on different architectures ▪ Build AI & HPC applications, optimize for new features & best performance ▪ Schedule GPU access ▪ Training & development, obtain Intel® Certified Developer accreditation for AI development & design	▪ Standard services usages + ▪ AI training & inference production workloads ▪ Model optimization & deployment ▪ Certification, software validation & benchmark testing	▪ Premium services + ▪ High-performance, cost-optimized Intel compute services for third-party AI SaaS providers
Hardware access	▪ 4th & 5th Gen Intel® Xeon® Scalable processors—bare metal & virtual machine (VM) ▪ Intel® Xeon® CPU Max Series processors—bare metal ▪ Intel® Data Center GPU Max Series & Intel® Data Center GPU Flex Series—bare metal ▪ Habana® Gaudi®2 processors for Deep Learning—bare metal access for pre-qualified, select customers	▪ Standard services access + ▪ Intel® Xeon® Scalable processors & Intel Data Center GPU Max & Flex Series—bare metal access to single node systems & clusters ▪ Access to k8s clusters	Same as Premium services
Technical support	Community forum support (no SLA)	Support through Intel technical engineers Monday-Friday, 8 a.m. to 5 p.m. (per user's local region) 1 business day SLA	Premium Support through Intel technical engineers (phone, chat, help request tickets) 1 hour to 1 business day SLA, 24x7
Cost	**Free** + available with cloud credits for certain instance types such as bare metal services Optional upgrade option for extended use, pay-as-you-go	Available with cloud credits + based on an hourly rate noted in Intel Developer Cloud portal Discounts for long-term contracts/ reserve pricing are available, contact your Intel representative	Available with cloud credits + monthly subscription rate noted in Intel Developer Cloud portal Discounted founders rate & long-term contracts/ reserve pricing are available, contact your Intel representative

Proven Customer Value & Benefits

Fine-Tuning LLMs with LoRA on Intel Max Series GPU for Predictions, E-commerce & Personal Assistants: Moonshot AI is breaking new ground in leveraging LLMs to make predictions. SiteMana is harnessing LLMs to automate e-commerce marketing. Selecton Technologies is carving a niche by developing an AI personal assistant for gamers utilizing LLMs.

Learn more

Revolutionizing Email Generation: Through advanced machine learning, SiteMana is providing businesses with a method to engage with anonymous traffic by identifying visitors with high buyer intent and retargeting them with personalized messaging. The approach is secure and legal, all the while respecting individual privacy and without uncovering personal identities.

Learn more

Derisking LLMs for Enterprise: Prediction Guard's platform enables companies to adopt the latest wave of AI models, like those used in ChatGPT, without compromising on privacy or security.

Learn more

Intel® Technologies in Intel® Developer Cloud

Hardware Catalog: CPUs	▪ **4th & 5th Gen Intel® Xeon® Scalable processors** with advanced AI accelerators and capabilities including: - Intel® Advanced Matrix Extensions (Intel® AMX) - Intel® Advanced Vector Extensions 512 (Intel® AVX-512) - Intel® QuickAssist Technology (Intel® QAT) - Intel® Data Streaming Accelerator (Intel® DSA) - Intel® In-Memory Analytics Accelerator (Intel® IAA) - VNNI/bfloat16	▪ Virtual machine (VM) & bare metal access (2 sockets, 256 GB memory, 2 TB disk)
	▪ **Intel® Xeon® CPU Max Series processors**, the only x86-based processor with high-bandwidth memory (HBM) & includes Intel® AMX, Intel® AVX-512, Intel® Deep Learning Boost (VNNI, bfloat16), and Intel® DSA. ▪ Coming soon: **Pre-production 5th Gen Intel® Xeon® Scalable processors** (codenamed Emerald Rapids)	▪ Bare metal access: DDR5 (8 channels 4,800 MT/S (1DPC), PCIe 5 (80 lanes), 64GB HBM2e, up to 56 cores /1 TB/S memory bandwidth. HBM-only mode, flat mode or cache mode.
GPUs	▪ **Intel® Data Center GPU Max Series** 1100 & 1550 includes innovative features with Xe-Core driving compelling Op/CLK for critical data formats for AI & HPC with vector & matriX engines (Intel® Xe Matrix Extensions (Intel® XMX)) & supports both SIMT and SIMD models, Intel® Xe Link, data type flexibility, ray-traced hardware acceleration, & more. ▪ **Intel® Data Center GPU Flex Series** supports media streaming, AI visual inference, cloud gaming, virtual desktop infrastructure (VDI), virtualization & digital content creation. Accelerates a variety of ray tracing, simulation, and image-enhancement workloads with built-in Intel® XMX AI acceleration, AV1 hardware encode & decode, and ray-traced hardware acceleration.	▪ Bare metal access using batch service for AI and ML training. Up to 128 ray tracing units, 128X(raised e) cores, up to 64MB L1 cache, 408MB L2 cache, up to 128GB HBM2e ▪ Bare metal access: Up to 32 Xe (raised e) cores & ray tracing units, up to 4 Xe media engines
AI Accelerators	▪ **Intel® Gaudi®2 Processors for Deep Learning** are the best AI accelerators for deep learning training and inference of LLMs and generative AI with performance and cost-efficiency. Gaudi2 optimized software provides easy access to state-of-the-art models ranging from small-scale computer vision & NLP models to efficient handling of multi-billion parameter models. Designed for efficient scalability, Gaudi2 accelerators are ideal for customers who need competitive performance on the most challenging workloads at a highly-efficient cost.	▪ Bare metal access for pre-qualified select customers

Get Started Today

Setting up your Intel Developer Cloud account is easy. Intel offers quick onboarding and training for AI LLM/MLOps and oneAPI. Join now at cloud.intel.com.

> We maximized our solver performance in a cost-efficient way on Intel® Xeon® processors using components from the Intel oneAPI HPC Toolkit. We're now getting ready to enable the workloads to run across CPU and GPU architectures.
>
> **Johannes Gutekunst**
> Chief Technology Officer (CTO), Dive Solutions GmbH

Intel® Technologies in Intel® Developer Cloud

Software Catalog: Optimized software & tools	▪ **Intel® oneAPI Base Toolkit** - Intel® oneAPI DP++/C++ Compiler - Libraries for deep neural network, math, data analytics, DPC++, threading, collective communications, video processing, and imaging, signal processing, data compression and cryptography - Analysis, debug & code migration tools–Intel® VTune™ Profiler, Intel® Advisor, Intel® Distribution for GDB*, Intel® DPC++ Compatibility Tool ▪ **Intel AI tools & optimized frameworks** - Intel-optimized deep learning frameworks for TensorFlow and PyTorch - Intel® OpenVINO™ toolkit - Intel® Neural Compressor - Intel® Extension for Scikit-learn*, Intel® Optimization for XGBoost*, Intel® Distribution of Modin, Intel® Distribution for Python* - Model Zoo for Intel® Architecture ▪ **Intel® HPC Toolkit**: Intel® Compilers (DPC++/C++, C++ and Fortran), Intel® Inspector, Intel® MPI Library, Intel® Trace Analyzer & Collector ▪ **Intel® Rendering Toolkit**: Intel® Embree, Intel® Open Image Denoise, Intel® Open Volume Kernel Library, Intel® Open Path Guiding Library, Intel® Implicit SPMD Program Compiler, Intel® OSPRay, Intel® OSPRay Studio, Intel® OSPRay Studio for Hydra, Intel® OpenSWR ▪ **Intel® Quantum SDK** ▪ **Jupyter notebooks:** Begin your development journey with a familiar Jupyter notebook, where you can write and run your Python code inline on Intel's newest CPUs and GPUs.
AI Services	▪ **Run open source AI foundational models – some examples include:** - Technology Innovation Institute* (TII) Falcon LLM - MosaicML* MPT - Hugging Face* BigScience Large Open-science Open-access Multilingual Language Model (BLOOM) - Stability.AI* Stable Diffusion - Meta AI* Llama 2 - Databricks* Dolly

References & Resources

- Intel Developer Cloud (cloud.intel.com)
- Intel AI Tools, Libraries & Frameworks Optimizations
- Intel® Liftoff for Startups
- Intel Enterprise AI Solutions
- Enterprise Software Solutions by Intel

Notices and Disclaimers

1. Performance varies by use, configuration, and other factors. Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. Learn more at www.Intel.com/PerformanceIndex and https://www.intel.com/content/www/us/en/developer/articles/technical/software-ai-accelerators-ai-performance-boost-for-free.html.

Roadmap Notice: All information provided in this question is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.
Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex. Results may vary.
Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. No product or component can be absolutely secure.
Your costs and results may vary. Intel technologies may require enabled hardware, software or service activation.
Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Printed in USA 0324/CCZ/TAN/PDF Please Recycle 359594-001US