Product Brief

Optimized AI Cloud Services

Build & Deploy AI at Scale in Intel® Tiber™ Developer Cloud

Accelerate AI with Managed, High-performance & Cost-efficient Infrastructure

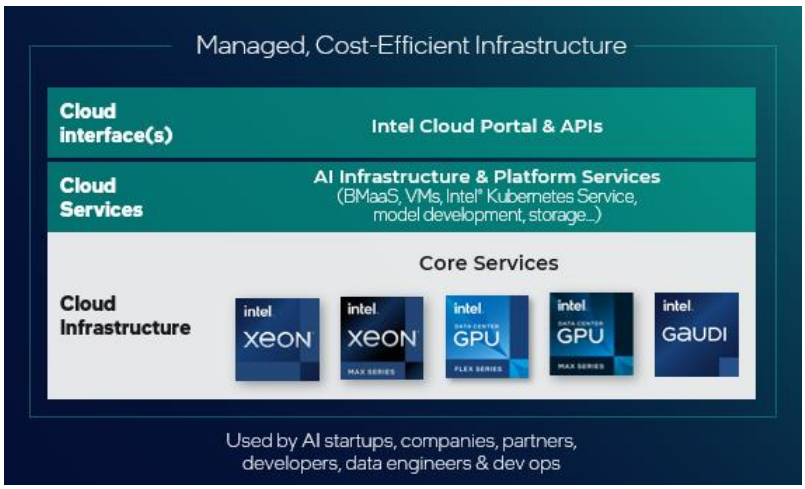
Getting value from AI, along with enough compute resources to scale your solution and optimize for efficiency is a challenge. Overcome the barriers of expensive or inaccessible cloud-compute resources. Intel® Tiber™ Developer Cloud, a managed, high-performance and cost-efficient AI cloud service, helps AI startups, companies and developers build AI solutions by providing a platform to develop and deploy AI models, applications and services at scale with **best price-performance**.

With designed-for-AI architecture, open platforms, and open software, users can innovate quickly with flexibility to advance their AI solutions on Intel's cloud.

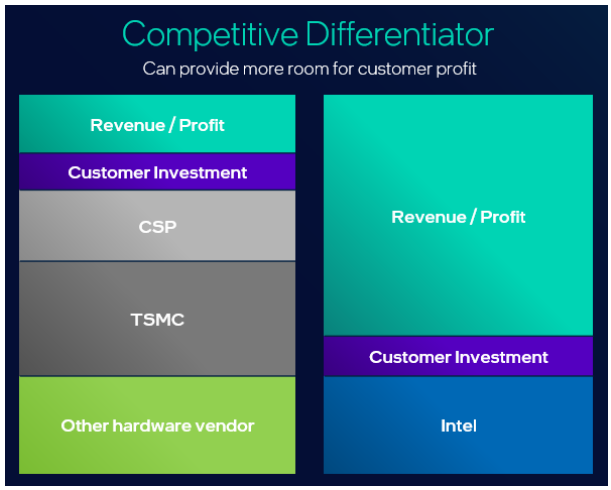
intel. tiber. Developer Cloud

More Value for AI Investments

Maximize value for AI compute using current and next gen hardware systems. Gain performance and productivity from software optimizations. With Intel Tiber Developer Cloud:



- **Build & Deploy AI at Scale**—Develop AI models, applications and solutions. Deploy training and inference production workloads at scale. Deliver up to 10-100X more performance using common tools.[1]
- **Maximize AI Compute Resources**—Choose the best accelerator for every use case for optimal price-performance. Systems include Intel CPUs with built-in AI acceleration, GPUs, and Intel® Gaudi® AI accelerators. Utilize virtual machines (VMs), full systems or clusters and take advantage of Intel® Kubernetes Service and storage. Intel's full, vertically-integrated cloud solution can provide more room for customer profit with:
  - Better price-performance for bare-metal AI accelerator/GPU SKUs compared to other CSPs
  - ELA subscriptions for 3, 6, and 12-month bare-metal SKUs offering favorable discounts
  - Customers gained up to 400% cost savings for select AI workloads vs. on-prem or another CSP.[2]
- **Open Software, Open Ecosystem Advantage**—To help the industry and developer communities innovate and accelerate AI, Intel's open platforms and a comprehensive, open software stack provides flexibility, ease, and choice in hardware. Intel's cloud is built on an open software foundation with oneAPI standards-based heterogeneous programming delivering code reuse and portability across multiple different vendors' hardware.



Example at right is illustrative in nature and does not reflect actual percentages in a fluctuating market, Your results and costs may vary.

Build & Deploy AI Easier—Key Usages

AI startups, companies, data scientists and developers can take advantage of Intel's latest platforms in Intel Tiber Developer Cloud. More than 32,500 users spanning AI companies, independent software vendors (ISVs), academics, and more are already using this platform.

For AI Compute & Deployment

Develop and optimize models, run small- and large-scale AI training (LLMs or genAI) and inference production workloads. Utilize small to large VMs, full systems or clusters with Intel CPUs, GPUs, and Gaudi accelerator systems. Scale from 7 to hundreds of billions of parameters.

For Companies & Enterprise Use

System integrators (SIs), ISVs, and third-party SaaS organizations use Intel's cloud to run AI training and inference production workloads at scale, for certification and benchmarking, and for third-party AI SaaS compute services.

For Developers

Intel's cloud provides an easy path to access and use Intel-optimized AI software accelerated 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 - obtain Intel® MLOps Certified Developer accreditation for AI development and design
- Try out LLM workload code samples to see how they perform on Intel architecture

GPU-based and Gaudi accelerator-based Jupyter notebooks are available in the cloud pre-loaded with GenAI models, training essentials and software tools for easy development. They also support Visual Studio Code*.

Proven Customer Value & Benefits

seekr

LLM & AI solution production deployment

- Moved from on-prem with A100s & other CSP to Intel's cloud using Intel® CPUs, GPUs & Intel® Gaudi® accelerators
- **2X** inference volume, **50%** faster inference
- **20%** faster AI training
- Up to **400%** cost savings for select AI workloads

prediction Guard

API service production deployment

- Moved from A100s to Intel's cloud using Gaudi
- **2X** throughput increase + decrease in costs for some models, **50%** latency reduction
- **2X** performance increase on 7B parameter models
- Significant cost savings

SELECTON technologies

Built applications using Low-Rank Adaptation of LLMs (LoRA) on virtual hardware

- Moved from Nvidia GPUs to Intel Max Series GPU & 4th Gen Intel® Xeon®
- Achieved exceptional results quoted its co-founder/CEO
- Reached new levels of performance & efficiency

Get Started Today

Setting up your Intel cloud account is easy. The cloud has an easy to use UI, a modern interface, and streamlined workflows help optimize end-to-end AI pipelines. It's simple to get started with quick onboarding and education modules for training for AI LLM/MLOps and oneAPI. No hardware installations or acquisition, software downloads and configuration setup are required.

Different service and support tiers are offered to meet customers' varying flexibility and compute needs. Cloud credits may also be available to get started, contact your Intel representative for details.

Join now at cloud.intel.com.

Simplify Your AI & Cloud Journey

Intel Tiber Developer Cloud is also available through Intel® Tiber™ AI Studio.

Intel Tiber AI Studio's MLOps platform streamlines the whole model lifecycle, so you can focus on deploying better models for the business. Support for Intel Tiber Developer Cloud is built-in, while re-usable software templates make it easy to create different cluster sizes and start developing. Get started now.

intel.tiber™ AI Studio

Intel Tiber Developer Cloud—Services & Support Options

1. Preview environment

Access pre-production systems for evaluation and optimizing applications and solutions for next gen architectures with advanced features. Includes Intel® Xeon® 6 e-core & p-core preview systems—bare metal.
Coming soon: Intel® Gaudi® 3 AI Accelerator preview systems.

	2. Standard (initial starting account)	3. Premium (paid account)	4. Enterprise (paid account)
Users	AI startups and developers, data scientists, performance engineers, researchers, academia Single-user access	AI and enterprise companies/developers Multi-user access	AI and enterprise companies Multi-user access
Usages	▪ Build AI applications, optimize for new features and best performance ▪ Schedule GPU access ▪ Education and development ▪ Hardware evaluation—Run applications, workloads, and LLM workload code samples on different architectures	**Standard usages +** ▪ Build and deploy AI training and inference production workloads ▪ Develop, optimize and deploy AI models, applications and solutions ▪ AI compute ▪ Certification, software validation and benchmark testing ▪ Create file storage volume with access to storage as a service (StaaS), maximum 50 TB. Quota is adjustable per account on request. Coming soon: Object storage.	**Premium usages +** ▪ High-performance, cost-optimized Intel compute and deployment services for third-party AI SaaS providers ▪ File storage maximum 500 TB. Quota is adjustable per account on request.
Hardware Access	▪ 4th Gen Intel® Xeon® processors -VMs and bare metal ▪ Intel® Xeon® CPU Max Series processors—bare metal ▪ Intel® Data Center GPU Max Series and Intel® Data Center GPU Flex Series—bare metal	**Standard hardware access +** ▪ 5th Gen Intel® Xeon® processors, plus all Xeon processors in the cloud ▪ Intel Data Center GPU Max and Flex Series—bare metal access to single node systems and clusters ▪ Intel® Gaudi®2 processors—bare metal access for pre-qualified, select customers ▪ Access to k8s clusters	**Same as Premium hardware access** Access to supercomputing Gaudi 2 clusters (64+ nodes)
AI Infrastructure Services	▪ LLM model training and optimization ▪ AI model deployment via CLI/SSH automation ▪ Hosting platform for deploying AIaaS ▪ Coming soon: VMs on Max Series GPUs	**Standard services +** ▪ Intel® Kubernetes Service for AI and general purpose, container deployment via K8s APIs ▪ Bare metal as a service (BMaaS) on large-scale Gaudi 2 clusters (up to 32 nodes) ▪ BMaaS on Max Series GPU clusters ▪ VMs on Gaudi accelerators	**Same as Premium services access** One click access to super-computing Gaudi cluster (32+ nodes) with fast storage for checkpoints and object storage
	Run open source AI foundational models—examples include: ▪ Technology Innovation Institute* (TII) Falcon LLM ▪ MosaicML* MPT ▪ Meta AI* Llama 2 ▪ Hugging Face* BigScience Large Open-science Open-access Multilingual Language Model (BLOOM) ▪ Databricks* Dolly ▪ Stability.AI* Stable Diffusion		
Technical Support	Community forum support (no SLA)	Support through Intel technical engineers Monday-Friday 8 a.m.-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 Tiber Developer Cloud portal—cloud.intel.com Discounts for long-term contracts/reserve pricing are available, contact your Intel representative	Available with cloud credits + Monthly subscription rate noted in Intel Tiber Developer Cloud portal—cloud.intel.com Discounted founders rate and long-term contracts/reserve pricing are available, contact your Intel representative

More details about Advanced Technologies in Intel® Tiber Developer Cloud

CPUs	4th Gen Intel® Xeon® Scalable processors with built-in AI acceleration—VM and bare metal access 2 sockets, 256 GB memory, 2 TB disk—see site for advanced AI capabilities 5th Gen Intel® Xeon® Scalable processors—Bare metal access Intel® Xeon® CPU Max Series processors with high-bandwidth memory—Bare metal access Intel® Xeon® 6 e-core and p-core preview systems
GPUs	Intel® Data Center GPU Max Series 1100 and 1550 large-scale GPU clusters—Bare metal access using batch service for AI and ML training. Includes innovative features with Xe-Core, supports SIMT and SIMD models, Intel® Xe Link, data type flexibility, ray-traced hardware acceleration, and more. Intel® Data Center GPU Flex Series—Bare metal access Supports media streaming, AI visual inference, cloud gaming, virtual desktop infrastructure (VDI), virtualization and digital content creation. Accelerates a variety of ray tracing, simulation, and image-enhancement workloads.
AI Accelerators	Intel® Gaudi®2 AI Accelerators for Deep Learning large-scale (128 node) clusters—Bare metal access for select premium and enterprise customers. Gaudi processors are the best AI accelerators for deep learning training and inference of LLMs and genAI with performance and cost efficiency.[3] Gaudi optimized software provides easy access to state-of-the-art models ranging from small-scale computer vision and NLP models to efficient handling of multi-billion parameter models. Coming soon: Intel Gaudi 3 AI Accelerator preview systems
Optimized Software & Tools	Multiple Intel AI tools and optimized frameworks for PyTorch* and TensorFlow* and Hugging Face Optimum Habana SynapseAI 1.15, 1.16 preview for Intel Gaudi processors Intel® oneAPI Base Toolkit—Intel® oneAPI DPC++/C++ Compiler, performance libraries, and advanced analysis, debug and code migration tools Intel® HPC Toolkit—Intel® Compilers (oneAPI DPC++/C++, Fortran), Intel® MPI Library Intel® Rendering Toolkit Intel® Quantum SDK **Jupyter notebooks**: Begin your development journey with a familiar Jupyter notebook, where you can run GenAI models, write and run your Python code inline, and learn oneAPI on Intel's newest CPUs and GPUs.

Tiber Cloud References & Resources

- Intel Tiber Developer Cloud (cloud.intel.com)
- Intel AI Tools, Libraries & Frameworks Optimizations
- Prediction Guard case study: Derisking LLMs for Enterprise
- Seekr: Building Trustworthy LLMs for Evaluating & Generating Content at Scale
- Intel® Tiber™ AI Solutions
- Intel® Liftoff for Startups

Simplify & Secure Your AI & Cloud Journey

The world is evolving fast. Advances in technology are redefining computing, AI, and cloud usages. While it is now possible to create a transformative GenAI or a cutting-edge viral application, these breakthroughs are unprecedented in their complexity, vulnerability, and costs. Intel provides a portfolio of software solutions for the next generation of innovators: built for speed, ready for growth, trusted for security, and optimized for cost. Get more value from your AI and cloud investments with solutions that deliver strong price performance, automated workflows and trusted secure environments.
Learn more: Intel® Tiber™ Enterprise Solutions

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 intel.com/content/www/us/en/developer/articles/technical/software-ai-accelerators-ai-performance-boost-for-free.html.
2. Prediction Guard case study, CIO.com Seekr article. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.
3. Gaudi & Xeon Advance Inference Performance for GenAI, Intel Gaudi Enables a Lower Cost Alternative for AI Compute and GenAI