ThinkSystem Intel Flex 140 12GB Gen4 Passive GPU
Product Guide

The ThinkSystem Intel Flex 140 12GB Gen4 Passive GPU is a general-purpose GPU optimized for media stream density and quality. It is designed for data center and edge applications, with high levels of reliability, availability, and scalability. The GPU is suitable for media processing and delivery, Windows and Android cloud gaming, virtualized desktop infrastructure (VDI), and AI visual inference applications.

The Intel Data Center GPU Flex 140 suitable for workloads that involve lighter AI models, such as simple object detection. Due to the higher number of video engines, the Flex Series 140 GPU can support a high number of streams.

Figure 1. ThinkSystem Intel Flex 140 12GB Gen4 Passive GPU

Did you know?

The Intel Flex Series GPUs support an open, flexible, standards-based software stack together with oneAPI so developers can build high-performance, cross-architecture applications and solutions. This helps organizations reduce the complexity, cost, and time requirements to bring new solutions to market, enabling engineers and programmers to innovate instead of maintaining code.

Part number information

The following table shows the ordering information for the Flex 140 GPU.

Table 1. Ordering information

Part number	Feature code	Description
4X67A86130	BU00	ThinkSystem Intel Flex 140 12GB Gen4 Passive GPU

The option part number includes the following:

- One Flex 140 GPU
- Low profile (2U) and full-height (3U) adapter brackets
- Documentation

Features

Intel Data Center GPU Flex Series is a flexible, robust, and the industry's most open GPU solution for the intelligent visual cloud. The GPUs support a diverse range of workloads in the industry starting with media streaming and cloud gaming, followed by support for AI visual inference and virtual desktop Infrastructure workloads. It supports an open, standards-based software stack optimized for density and quality with critical server capabilities for high reliability, availability, and scalability. This helps reduce the need for data centers to use disparate solutions and manage heterogenous or proprietary environments.

The ThinkSystem Intel Flex 140 12GB Gen4 Passive GPU offers the following features:

- Open Architecture

  The Intel Flex Series GPU supports an open, flexible, standards-based software stack together with oneAPI so developers can build high-performance, cross-architecture applications and solutions. This helps organizations reduce the complexity, cost, and time requirements to bring new solutions to market, enabling engineers and programmers to innovate instead of maintaining code.

- Built-In AV1 Encode

  Services built on the royalty-free open-source AV1 codec mean lowering operational expenses while providing higher video quality. Advanced video coding (AVC), High Efficiency Video Coding (HEVC), and VP9 support also comes standard with the Intel Data Center GPU.

- No Licensing Fees

  Intel provides free virtual GPU software license to customers for lowering their total cost of ownership in VDI deployments. Furthermore, this can act as a catalyst to accelerate the GPU adoption rate in VDI deployments where graphics & encode accelerations are desired/preferred but can often be cost prohibitive.

- Flexible vGPU Management

  The Intel Data Center GPU Flex Series supports multiple vGPU configuration & scheduling options for VDI solutions to meet different customer workload and QoS requirements—such as Linux KVM open-source virtualization technology and VMware ESXi. It also supports both VMware Horizon and Citrix DaaS† on top. The Flex Series accelerators have strong ecosystem support. The top two industry leading desktop & application virtualization solutions, VMware Horizon and Citrix DaaS, will be supported.

- Flexible Performance

  The Flex 140 accelerator has two GPUs on a single card, supporting heterogeneous vGPU profiles. Having fewer virtual machines per GPU contributes to more predictable quality of service, shorter GPU scheduling queues, and strong performance isolation across machines. Smaller physical cards and fewer virtual machines means you can cover a variety of user needs more economically.

- High Density Virtual Desktop

  The Intel Data Center GPU Flex 140 is ideal for VDI deployments targeting knowledge workers persona - office productivity, video, and browser workloads whose graphics and compute performance requirements are low.

Technical specifications

The following table lists the specifications of the Flex 140 GPU.

Table 2. Specifications

Feature	**Flex 140**
GPU Architecture	X^e HPG
GPUs per card	2
Execution units	256 (128 per GPU)
Intel X^e Cores	16 cores (8 per GPU)
Media Engine	4 (2 per GPU)
Ray Tracing	16 RT units
FP32 peak performance	8 TFLOPS
FP16 peak performance	52 TFLOPS
INT8 peak performance	105 TOPS
INT4 peak performance	210 TOPS
GPU Memory	12 GB ECC GDDR6 (6 GB per GPU)
Memory Bandwidth	336 GB/s (168 GB/s per GPU)
Virtualization (SR-IOV)	62 Virtual Functions (VFs)
System Interface	PCIe Gen 4, x8 lanes (x16 mechanical)
Form Factor	PCIe low profile, single width
Max Power Consumption	75 W
Thermal Solution	Passive
Encoding and decoding	• H.264 Hardware Encode/Decode • H.265 (HEVC) Hardware Encode/Decode • AV1 Encode/Decode • VP9 Bitstream & Decoding

Server support

The following tables list the ThinkSystem servers that are compatible.

Table 3. Server support (Part 1 of 4)

Part Number	Description	2S AMD V3				2S Intel V3			4S 8S Intel V3			Multi Node			GPU Rich				1S V3	
		SR635 V3 (7D9H / 7D9G)	SR655 V3 (7D9F / 7D9E)	SR645 V3 (7D9D / 7D9C)	SR665 V3 (7D9B / 7D9A)	ST650 V3 (7D7B / 7D7A)	SR630 V3 (7D72 / 7D73)	SR650 V3 (7D75 / 7D76)	SR850 V3 (7D97 / 7D96)	SR860 V3 (7D94 / 7D93)	SR950 V3 (7DC5 / 7DC4)	SD535 V3 (7DD8 / 7DD1)	SD530 V3 (7DDA / 7DD3)	SD550 V3 (7DD9 / 7DD2)	SR670 V2 (7Z22 / 7Z23)	SR675 V3 (7D9Q / 7D9R)	SR680a V3 (7DHE)	SR685a V3 (7DHC)	ST250 V3 (7DCF / 7DCE)	SR250 V3 (7DCM / 7DCL)
4X67A86130	ThinkSystem Intel Flex 140 12GB Gen4 Passive GPU	N	N	N	N	N	N	N	N	N	N	N	1	2	N	N	N	N	N	N

Table 4. Server support (Part 1 of 4)

Part Number	Description	Edge					Super Computing					1S Intel V2			2S Intel V2		
		SE350 (7Z46 / 7D1X)	SE350 V2 (7DA9)	SE360 V2 (7DAM)	SE450 (7D8T)	SE455 V3 (7DBY)	SD665 V3 (7D9P)	SD665-N V3 (7DAZ)	SD650 V3 (7D7M)	SD650-I V3 (7D7L)	SD650-N V3 (7D7N)	ST50 V2 (7D8K / 7D8J)	ST250 V2 (7D8G / 7D8F)	SR250 V2 (7D7R / 7D7Q)	ST650 V2 (7Z75 / 7Z74)	SR630 V2 (7Z70 / 7Z71)	SR650 V2 (7Z72 / 7Z73)
4X67A86130	ThinkSystem Intel Flex 140 12GB Gen4 Passive GPU	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N

Table 5. Server support (Part 1 of 4)

Part Number	Description	AMD V1					Dense V2				4S V2		8S	4S V1			1S Intel V1			
		SR635 (7Y98 / 7Y99)	SR655 (7Y00 / 7Z01)	SR655 Client OS	SR645 (7D2Y / 7D2X)	SR665 (7D2W / 7D2V)	SD630 V2 (7D1K)	SD650 V2 (7D1M)	SD650-N V2 (7D1N)	SN550 V2 (7Z69)	SR850 V2 (7D31 / 7D32)	SR860 V2 (7Z59 / 7Z60)	SR950 (7X11 / 7X12)	SR850 (7X18 / 7X19)	SR850P (7D2F / 2D2G)	SR860 (7X69 / 7X70)	ST50 (7Y48 / 7Y50)	ST250 (7Y45 / 7Y46)	SR150 (7Y54)	SR250 (7Y52 / 7Y51)
4X67A86130	ThinkSystem Intel Flex 140 12GB Gen4 Passive GPU	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N

Table 6. Server support (Part 4 of 4)

Part Number	Description	2S Intel V1								Dense V1			
		ST550 (7X09 / 7X10)	SR530 (7X07 / 7X08)	SR550 (7X03 / 7X04)	SR570 (7Y02 / 7Y03)	SR590 (7X98 / 7X99)	SR630 (7X01 / 7X02)	SR650 (7X05 / 7X06)	SR670 (7Y36 / 7Y37)	SD530 (7X21)	SD650 (7X58)	SN550 (7X16)	SN850 (7X15)
4X67A86130	ThinkSystem Intel Flex 140 12GB Gen4 Passive GPU	N	N	N	N	N	N	N	N	N	N	N	N

Software stack

The Flex Series GPU supports an open, flexible, standards-based software stack with oneAPI cross-architecture programming. The stack includes open source components and libraries, tools and frameworks so developers can create high-performance, cross-architecture media applications and solutions to meet a wide range of use cases. This open approach removes the barriers to proprietary models where code portability and the ability to adopt new architectures across multiple vendors is limited.

Intel enables the software ecosystem through industry collaborations, initiatives and standards bodies. It also provides ongoing leadership, investment and technical contributions to the open source community.

The common set of software capabilities integrates into popular middleware and frameworks, and the stack is delivered in validated productized containers or reference stacks. The containers can be orchestrated with Kubernetes on bare metal or in VMs using SR-IOV virtualization with tools to assign and manage workloads. The toolset is designed to speed time-to-market and enable flexible deployment of multiple workloads on the same GPU.

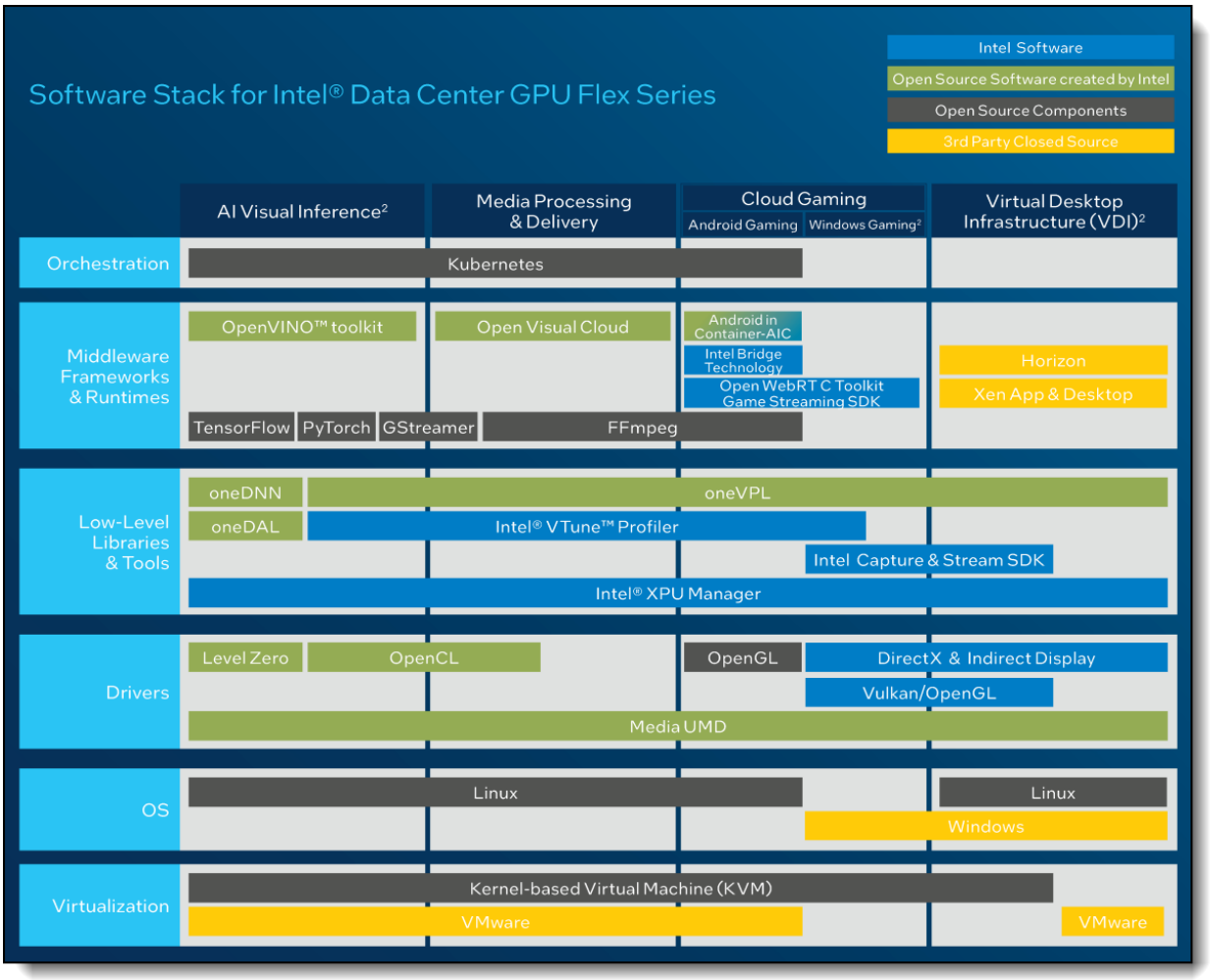

Figure 2. Software Stack for Intel Data Center GPU Flex Series (source: intel.com)

The following components are part of the Intel oneAPI Base Toolkit (individual tools can be downloaded separately):

- oneAPI Deep Neural Network Library (oneDNN)
- oneAPI Data Analytics Library (oneDAL)
- oneAPI Video Processing Library (oneVPL)
- Intel VTune Profiler

The following Intel-optimized tools are part of the Intel AI Analytics Toolkit:

- TensorFlow
- PyTorch

For more information about Intel oneAPI, see https://intel.com/oneapi

Operating system support

The following table lists the supported operating systems:

Tip: These tables are automatically generated based on data from Lenovo ServerProven.

Table 7. Operating system support for ThinkSystem Intel Flex 140 12GB Gen4 Passive GPU, 4X67A86130

Operating systems	SD530 V3	SD550 V3
Microsoft Windows Server 2022	Y	Y
VMware vSphere Hypervisor (ESXi) 7.0 U3	Y	Y

Auxiliary power cable

The Flex 140 GPU does not require an auxiliary power cable.

Operating environment

The Flex 140 GPU has the following operating characteristics:

- Ambient temperature
  - Operational: 0°C to 85°C
  - Storage: -40°C to 70°C
- Relative humidity:
  - Operational: 5 to 85%
  - Storage: 5 to 90%

Physical specifications

The Flex 140 GPU has the following physical specifications:

- Height: 69 mm (including edge connector)
- Length: 168 mm
- Width: 19 mm
- Weight: 400 g

Warranty

One year limited warranty. When installed in a Lenovo server, the GPU assumes the server's base warranty and any warranty upgrades.

Related publications

For more information, refer to these documents:

- ThinkSystem and ThinkAgile GPU Summary:
  https://lenovopress.lenovo.com/lp0768-thinksystem-thinkagile-gpu-summary
- ServerProven compatibility:
  https://serverproven.lenovo.com/
- Intel Data Center GPU Flex Series product page:
  https://www.intel.com/content/www/us/en/products/details/discrete-gpus/data-center-gpu/flex-series.html
- Intel oneAPI:
  https://intel.com/oneapi

Related product families

Product families related to this document are the following:

- GPU adapters

Notices

Lenovo may not offer the products, services, or features discussed in this document in all countries. Consult your local Lenovo representative for information on the products and services currently available in your area. Any reference to a Lenovo product, program, or service is not intended to state or imply that only that Lenovo product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any Lenovo intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any other product, program, or service. Lenovo may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

Lenovo (United States), Inc.
8001 Development Drive
Morrisville, NC 27560
U.S.A.
Attention: Lenovo Director of Licensing

LENOVO PROVIDES THIS PUBLICATION ”AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. Lenovo may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

The products described in this document are not intended for use in implantation or other life support applications where malfunction may result in injury or death to persons. The information contained in this document does not affect or change Lenovo product specifications or warranties. Nothing in this document shall operate as an express or implied license or indemnity under the intellectual property rights of Lenovo or third parties. All information contained in this document was obtained in specific environments and is presented as an illustration. The result obtained in other operating environments may vary. Lenovo may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any references in this publication to non-Lenovo Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this Lenovo product, and use of those Web sites is at your own risk. Any performance data contained herein was determined in a controlled environment. Therefore, the result obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

This document, LP1830, was created or updated on March 6, 2024.

Send us your comments in one of the following ways:

- Use the online Contact us review form found at:
  https://lenovopress.lenovo.com/LP1830
- Send your comments in an e-mail to:
  comments@lenovopress.com

This document is available online at https://lenovopress.lenovo.com/LP1830.

Trademarks

Lenovo and the Lenovo logo are trademarks or registered trademarks of Lenovo in the United States, other countries, or both. A current list of Lenovo trademarks is available on the Web at https://www.lenovo.com/us/en/legal/copytrade/.

The following terms are trademarks of Lenovo in the United States, other countries, or both:
Lenovo®
ServerProven®
ThinkAgile®
ThinkSystem®

The following terms are trademarks of other companies:

Intel® and VTune™ are trademarks of Intel Corporation or its subsidiaries.

Linux® is the trademark of Linus Torvalds in the U.S. and other countries.

Microsoft®, Windows Server®, and Windows® are trademarks of Microsoft Corporation in the United States, other countries, or both.

Other company, product, or service names may be trademarks or service marks of others.