XPU FAQ

1、 What is XPU?

XPU is a container GPU virtualization product introduced by YOYOWORKS LLC ("YOYO"). The core is to split the GPU at the kernel layer into many resource shares, and then simulate them into XPU devices (vGPU) for containers. XPU framework unbinds AI applications with physical GPUs, and binds them to virtual GPUs. With isolating fault, video memory, and computing in the scope of virtual GPU, XPU greatly improves business applications concurrency and achieves better GPU hardware utilization.

2、 What does XPU support?

- Hardware

  XPU is designed to support mainstream accelerators, including GPU, FPGA and ASIC, but currently it only support NVIDIA GPUS, including:

  - Pascal and later NVIDIA GPUs
  - Tesla/RTX/Quadro/Titan/GeForce GPUs

- Software

  XPU is designed to support all mainstream Linux distributions, and the following OS version are well-tested:

  RHEL 7.x/8.x/9.x, CentOS 7.x/8.x, AlmaLinux 8.x/9.x, Rocky 8.x/9.x
  Ubuntu 18.04/20.04/22.04 LTS

- CUDA version

  XPU supports CUDA 8.0 and later versions

3、 Does XPU support vGPU for VM?

No yet. XPU can be used for containers for now. However, XPU is compatible with NVIDIA GRID vGPU. That is, XPU can be installed in VM with GRID vGPUs, and then can help containers share one GRID vGPU.

4、 Can my legacy Docker containers run on XPU?

Yes. All containers, including NGC images, are designed to be run on XPU without any issues.

5、 What API XPU support?

XPU support NVIDIA CUDA API, OpenGL/OpenEGL, and Vulkan API for Linux 64 bit.

6、 How many GPUs can XPU support on one machine?

XPU can support up to 8 GPUs on one machine. If you need more GPUs, please contact XPU support team at support@yoyoworks.com for help.

7、 What is the maximum vGPU number can XPU virtualizes a physical GPU to?

16 for now. If you need more vGPUs, please contact XPU support team at support@yoyoworks.com for help.

8、 How many vGPUs XPU support for one container?

One backend physical GPU is needed for each vGPU assigned to the container. That is, the maximum vGPU for the container is the number of physical GPUs.

9、 Does XPU support ARM architecture?

Currently XPU is well tested on x86_64 architecture only. However, XPU is designed to be architecture neutral. Since NVIDIA GPU and Docker are supported on ARM architecture (XPU requires Dockers and NVIDIA GPU), XPU is supposed to be compatible with ARM. If you need XPU on ARM, please contact XPU support team at support@yoyoworks.com for help.

10、 Can XPU virtualize more than one vGPUs based on one physical GPU and assign them to one container?

No. If you need more GPU resource, you can increase the memory and computing shares to one XPU, and assign it to the container.

11、 Can XPU virtualize two physical GPUs to one vGPU and assign it to one container?

No yet.

12、 How XPU is licensed? And what about the technical support?

On-premises:

It is provided to the user in the form of a software installation package. When the user installs for the first time, he/she comes with a license of the Express version, which supports at most two vGPUs on each physical NVIDIA GPU. If the standalone node needs to support more vGPUs on a physical GPU, obtain relevant GPU information through the GPU information collection tool xpu-gpu-collect in the XPU software package and contact YOYO at support@yoyoworks.com for support.

On Clouds:

The current version XPU is available via some AWS instances types freely. If you need run XPU on other AWS instance type, on your local machines, or you need more technical support and service, please contact business@yoyoworks.com for help.