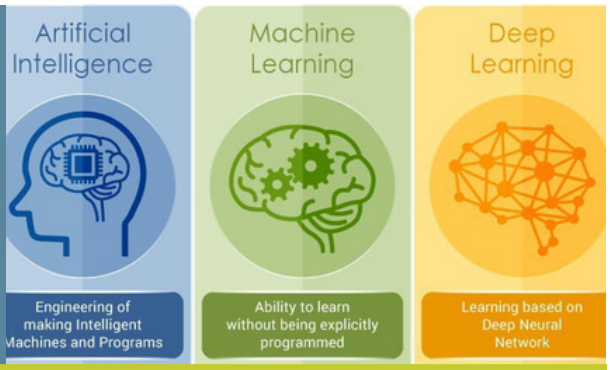
BEIJING POSTS & TELE. UNIVERSITY

How Beijing Posts & Telecommunications University accommodates its students and faculties with sufficient GPU resources for everyone.



OBJECTIVES

- Everyone can use GPUs when they need them.
- Multiple AI jobs can run on a single GPU at the same time.
- GPUs can "smartly transfer" computing power between idle jobs and busy jobs.
- GPUs can return resources automatically.
- Compatible with different types of NVIDIA cards.

CHALLENGES

The University of Beijing Posts & Telecommunications - BUPT is the country's leading Artificial Intelligence and Machine Learning institute.

As the need for AI teaching exploded in recent years, BUPT faced big challenges in figuring out how to utilize the GPUs fully - simply too many people and not enough GPUs.

At first, the school thought buying more GPUs would solve the problem, but it didn't. As the wait time to use the GPUs got longer and longer, everyone was frustrated.

Therefore, as Professor Xiao puts it - "There's gotta be a better way to solve this problem."

"XPU is a singularly transformative solution for us.

We were able to harness the power of our GPUs like never before thanks to XPU. Our overall productivity has skyrocketed; the difference is night and day!"

Professor Bo Xiao
School of Artificial Intelligence
Beijing Posts & Tele. University

SOLUTIONS

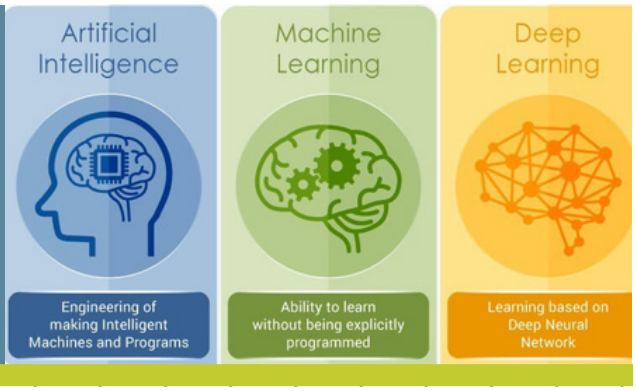
XPU stepped in with a virtualization solution. It allowed everyone to "chop" a piece of the physical GPU to suit their needs on demand.

First, a single GPU can now be split into 16 virtual GPUs, each with self-defined computing power and memory size; it is then mapped to the AI workload from the GPU's kernel layer, entirely isolated as if everyone owns their dedicated GPU.

Please continue reading on the next page.

BEIJING POSTS & TELE. UNIVERSITY

How Beijing Posts & Telecommunications University accommodates its students and faculties with sufficient GPU resources for everyone.



The following figure shows a virtualized GPU share of a NVIDIA A10 GPU was created with memory size of 4096 MB.

```
[root@ubuntu198 ~]# docker exec -it tensorflow /bin/bash
root@217c4bb0bde1:/notebooks# nvidia-smi
Wed Mar 15 03:44:08 2023
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 515.65.01    Driver Version: 515.65.01    CUDA Version: 11.7     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                               |                      |               MIG M. |
|===============================+======================+======================|
|   0  XPU-NVIDIA A10      On   | 00000000:00:08.0 Off |                    0 |
|  0%   29C    P8    18W / 150W |      0MiB /  4096MiB |      0%      Default |
|                               |                      |             Disabled |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes:                                                                  |
|  GPU   GI   CI        PID   Type   Process name                  GPU Memory |
|        ID   ID                                                   Usage      |
|=============================================================================|
|  No running processes found                                                 |
+-----------------------------------------------------------------------------+
root@217c4bb0bde1:/notebooks#
```

You can then assign your AI workload to utilize this vGPU; it is entirely isolated as if installed on your machine.

Next, XPU took care of resource recycling and Auto Power Transfer on virtual GPUs automatically, freeing up resources after jobs were finished and load balancing between virtual GPUs to achieve maximum usage.

Now, BUPT can support massive AI training generated by students with only a limited number of GPUs simultaneously. XPU also provides unmatched compatibility on nearly all types of NVIDIA GPUs, whether legacy or newer.

Tesla	H100 A800/A100/A10/A16/A30/A40 T4 V100 P100/P40/P6/P4
RTX	A6000/A5000/A4000
Quadro	RTX8000/RTX6000/RTX5000/RTX4000 P6000/P5000/P4000
GeForce	All 30XX, for example 3090/3080Ti All 20XX, for example 2080/2080Ti All 10XX, for example 1080