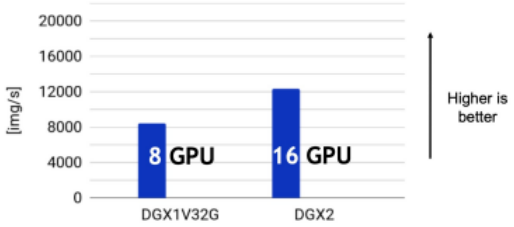
GPU选型

> 如何排查性能瓶颈，参考。此外需注意3060、3090、3080Ti、A4000、A40、A100、A5000等安培架构的卡需要**cuda11.x**才能使用，请使用较高版本的框架。

AutoDL平台分配GPU、CPU、内存的机制为：按租用的GPU数量成比例分配CPU和内存，算力市场显示的CPU和内存均为每GPU分配的CPU和内存，如果租用两块GPU，那么CPU和内存就x2。此外GPU非共享，每个实例对GPU是独占的。

一. 选择CPU

CPU非常重要！尽管CPU并不直接参与深度学习模型计算，但CPU需要提供大于模型训练吞吐的数据处理能力。比如，一台8卡NVIDIA V100的DGX服务器，训练ResNet-50 ImageNet图像分类的吞吐就达到8000张图像/秒，而扩展到16卡V100的DGX2服务器却没达到2倍的吞吐，说明这台DGX2服务器的CPU已经成为性能瓶颈了。



我们通常为每块GPU分配固定数量的CPU逻辑核心。理想情况下，模型计算吞吐随GPU数量线性增长，单GPU的合理CPU逻辑核心数分配可以直接线性扩展到多GPU上。AutoDL平台的算力实例提供了多种CPU分配规格。每块GPU应配备至少4~8核心的CPU，以满足多线程的异步数据读取。分配更多的核心通常不会再有很大的收益，此时的数据读取瓶颈通常源于Python的多进程切换与数据通信开销（如使用PyTorch DataLoader）。那么怎么省钱克服数据读取瓶颈呢，不妨在AutoDL平台试试C++和CUDA编写的NVIDIA DALI数据读取加速库吧。在我们的测试中，单核CPU实例的数据读取能力就超过了基于Python的八核心实例，真正做到了为模型训练保驾护航。

AutoDL中高性能CPU的选择有：

1. 内蒙A区 A5000 / 3090 / A40用到的AMD EPYC 7543 CPU
2. 内蒙A区 A100用到的AMD EPYC 7763 CPU
3. 北京A区 3090用到的Intel(R) Xeon(R) Gold 6330 或 AMD EPYC 7642 CPU
4. 深圳A区 3090用到的Intel(R) Xeon(R) Gold 6330

> 服务器的CPU一般不如桌面CPU的主频高，但是核心数量多。因此您从以前使用桌面CPU切换到服务器CPU上后，需要充分利用多核心的性能，否则无法发挥服务器CPU的性能。如何利用请戳

二. 选择GPU

AutoDL平台上提供的GPU型号很多。我们按照GPU架构大致分为五类：

1. NVIDIA Pascal架构的GPU，如TitanXp，GTX 10系列等。 这类GPU缺乏低精度的硬件加速能力，但却具备中等的单精度算力。由于价格便宜，适合用来练习训练小模型(如Cifar10)或调试模型代码。
2. NVIDIA Volta/Turing架构的GPU，如GTX 20系列, Tesla V100等。 这类GPU搭载专为低精度(int8/float16)计算加速的TensorCore, 但单精度算力相较于上代提升不大。我们建议在实例上启用深度学习框架的混合精度训练来加速模型计算。 相较于单精度训练，混合精度训练通常能够提供2倍以上的训练加速。
3. NVIDIA Ampere架构的GPU，如GTX 30系列，Tesla A40/A100等。 这类GPU搭载第三代TensorCore。相较于前一代，支持了TensorFloat32格式，可直接加速单精度训练 (PyTorch已默认开启)。但我们仍建议使用超高算力的

float16半精度训练模型，可获得比上一代GPU更显著的性能提升。

4. 寒武纪 MLU 200系列加速卡。 暂不支持模型训练。使用该系列加速卡进行模型推理需要量化为int8进行计算。 并且需要安装适配寒武纪MLU的深度学习框架。
5. 华为 Ascend 系列加速卡。 支持模型训练及推理。但需安装MindSpore框架进行计算。

GPU型号的选择并不困难。对于常用的深度学习模型，根据GPU对应精度的算力可大致推算GPU训练模型的性能。AutoDL平台标注并排名了每种型号GPU的算力，方便大家选择适合自己的GPU。

GPU的数量选择与训练任务有关。一般我们认为模型的一次训练应当在24小时内完成，这样隔天就能训练改进之后的模型。以下是选择多GPU的一些建议：

- 1块GPU。适合一些数据集较小的训练任务，如Pascal VOC等。
- 2块GPU。同单块GPU，但是你可以一次跑两组参数或者把Batchsize扩大。
- 4块GPU。适合一些中等数据集的训练任务，如MS COCO等。
- 8块GPU。经典永流传的配置！适合各种训练任务，也非常方便复现论文结果。
- 我要更多！用于训练大参数模型、大规模调参或超快地完成模型训练。

最后需注意：3060、3090、3080Ti、A4000、A40、A100、A5000等安培架构的卡需要cuda11.x才能使用，请使用较高版本的框架。

三. 选择内存

内存在充足的情况下一般不影响性能，但是由于AutoDL的实例相比本地电脑对内存的使用有更严格的上限限制（本地电脑内存不足会使用硬盘虚拟内存，影响是速度下降），比如租用的实例分配的内存是64GB，程序在训练时最后将要使用64.1GB，此时超过限制的这一时刻进程会被系统Kill导致程序中断，因此如果对内存的容量要求大，请选择分配内存更多的主机或者租用多GPU实例。如果不确定内存的使用，那么可以在实例监控中观察内存使用情况。

附GPU型号简介

型号	显存	单精(FP32)	半精(FP32)	说明
TITAN Xp	12GB	12.15 T	12.15 T	比较老的Pascal架构的GPU，用作入门比较合适
1080 Ti	11GB	11.34 T	11.34 T	和TITANXp同时代的卡，同样适合入门，但是11GB的显存偶尔会比较尴尬

型号	显存	单精(FP32)	半精(FP32)	说明
2080Ti	11GB	13.45 T	53.8 T	图灵架构GPU，性能还不错，老一代型号中比较适合做混合精度计算的GPU。性价比高
V100	16/32GB	15.7 T	125 T	老一代专业计算卡皇，半精性能高适合做混合精度计算
3060	12GB	12.74 T	约24T	如果1080Ti的显存正好尴尬了，3060是不错的选择，适合新手。需要使用cuda11.x
A4000	16GB	19.17 T	约76T	显存和算力都比较均衡，适合进阶过程使用。需要使用cuda11.x
3080Ti	12GB	34.10 T	约70T	性能钢炮，如果对显存要求不高则是非常合适的选择。需要使用cuda11.x
A5000	24GB	27.77T	约117T	性能钢炮，如果觉得3080Ti的显存不够用A5000是合适的选择，并且半精算力高适合混合精度。需要使用cuda11.x
3090	24GB	35.58 T	约71T	可以看做3080Ti的扩显存版。性能和显存大小都非常够用，适用性非常强，性价比首选。需要使用cuda11.x
A40	48GB	37.42 T	149.7 T	可以看做是3090的扩显存版。算力和3090基本持平，因此根据显存大小进行选择。需要使用cuda11.x
A100 SMX4	40/80GB	19.5 T	312 T	新一代专业计算卡皇，除了贵没缺点。显存大，非常适合做半精计算，因为有NVLink加持，多卡并行加速比非常高。需要使用cuda11.x