由于CPU架构设计原因，目前主流CPU（Intel Xeon E5-2600系列）单颗最大支持40条PCIe通道，双路服务器最多支持80条PCIe通道，所以支持两个GPU以上的服务器，GPU与GPU之间是有多种不同的连接形式，一般来讲最常见的几种连接形式是以下几种：

PLX— 指GPU与GPU之间采用同一个PLX Switch芯片相连，Switch芯片由CPU控制，同一个Switch之下的GPU互传数据，速度较快。

PHB— 指GPU与GPU之间通过同一个CPU相连，同一个CPU之下的GPU互传数据，速度较快。

SOC— 指GPU与GPU之间跨CPU相连，采用这种链接形式，GPU之间互传数据，速度最慢。

PXB— 指GPU与GPU之间采用多个PLX switch芯片，目前市面上主流产品几乎不采用这种设计。

NV#— 指GPU与GPU之间通过NVlink相连，在NVlink直连情况下GPU与GPU之间互传数据，速度非常快，在NVlink间接连接情况下，由于NVlink设计原因，使得程序有很大的优化空间，GPU与GPU互传数据速度也要明显高于其他链接形式。

下面是我司几款主流产品GPU连接的架构图

超微SYS-7048GR-TR

```
user@user-ubuntu:~$ nvidia-smi topo  -m
        GPU0    GPU1    GPU2    GPU3    CPU Affinity
GPU0     X      PHB     SOC     SOC     0-7,16-23
GPU1    PHB      X      SOC     SOC     0-7,16-23
GPU2    SOC     SOC      X      PHB     8-15,24-31
GPU3    SOC     SOC     PHB      X      8-15,24-31

Legend:

  X    = Self
  SOC  = Connection traversing PCIe as well as the SMP link between CPU sockets(e.g. QPI)
  PHB  = Connection traversing PCIe as well as a PCIe Host Bridge (typically the CPU)
  PXB  = Connection traversing multiple PCIe switches (without traversing the PCIe Host Bridge)
  PIX  = Connection traversing a single PCIe switch
  NV#  = Connection traversing a bonded set of # NVLinks
user@user-ubuntu:~$
```

通过上图可以看到，GPU0 和 GPU1 是一个 CPU 控制，GPU2 和 GPU3 是另外一个 CPU 控制，缺点比较明显（GPU0 或 GPU1 传数据到 GPU2 或 GPU3 的时候，速度很慢），客户编写 GPU 通讯函数的时候可以针对这种情况做优化。

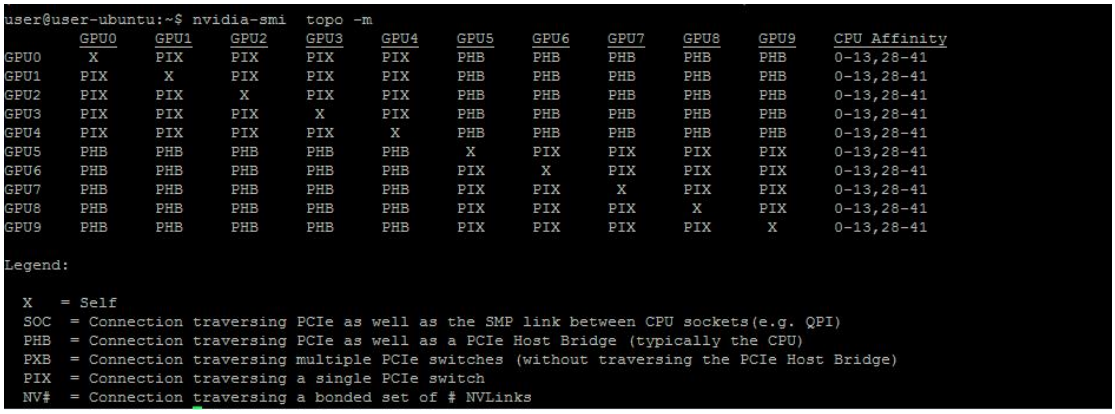
超微 SYS-4028GR-TR

```
[root@gpu01 ~]# nvidia-smi topo -m
        GPU0    GPU1    GPU2    GPU3    GPU4    GPU5    GPU6    GPU7    mlx4_0  CPU Affinity
GPU0     X      PIX     PHB     PHB     SOC     SOC     SOC     SOC     PHB     0-9,20-29
GPU1    PIX      X      PHB     PHB     SOC     SOC     SOC     SOC     PHB     0-9,20-29
GPU2    PHB     PHB      X      PIX     SOC     SOC     SOC     SOC     PHB     0-9,20-29
GPU3    PHB     PHB     PIX      X      SOC     SOC     SOC     SOC     PHB     0-9,20-29
GPU4    SOC     SOC     SOC     SOC      X      PIX     PHB     PHB     SOC     10-19,30-39
GPU5    SOC     SOC     SOC     SOC     PIX      X      PHB     PHB     SOC     10-19,30-39
GPU6    SOC     SOC     SOC     SOC     PHB     PHB      X      PIX     SOC     10-19,30-39
GPU7    SOC     SOC     SOC     SOC     PHB     PHB     PIX      X      SOC     10-19,30-39
mlx4_0  PHB     PHB     PHB     PHB     SOC     SOC     SOC     SOC      X

Legend:

  X    = Self
  SOC  = Connection traversing PCIe as well as the SMP link between CPU sockets(e.g. QPI)
  PHB  = Connection traversing PCIe as well as a PCIe Host Bridge (typically the CPU)
  PXB  = Connection traversing multiple PCIe switches (without traversing the PCIe Host Bridge)
  PIX  = Connection traversing a single PCIe switch
  NV#  = Connection traversing a bonded set of # NVLinks
```

根据上图能分析出该机型一共有 4 个 PLX switch 芯片， 一个 CPU 控制两个 PLX switch，GPU0-GPU3 互传数据速度较快，GPU4-GPU7 互传数据速度较快， 但是 0-3 任意一个 GPU 传数据到 4-7 任意一个 GPU 时，需要跨 CPU，速度较慢，客户编写 GPU 通讯函数的时候可以针对

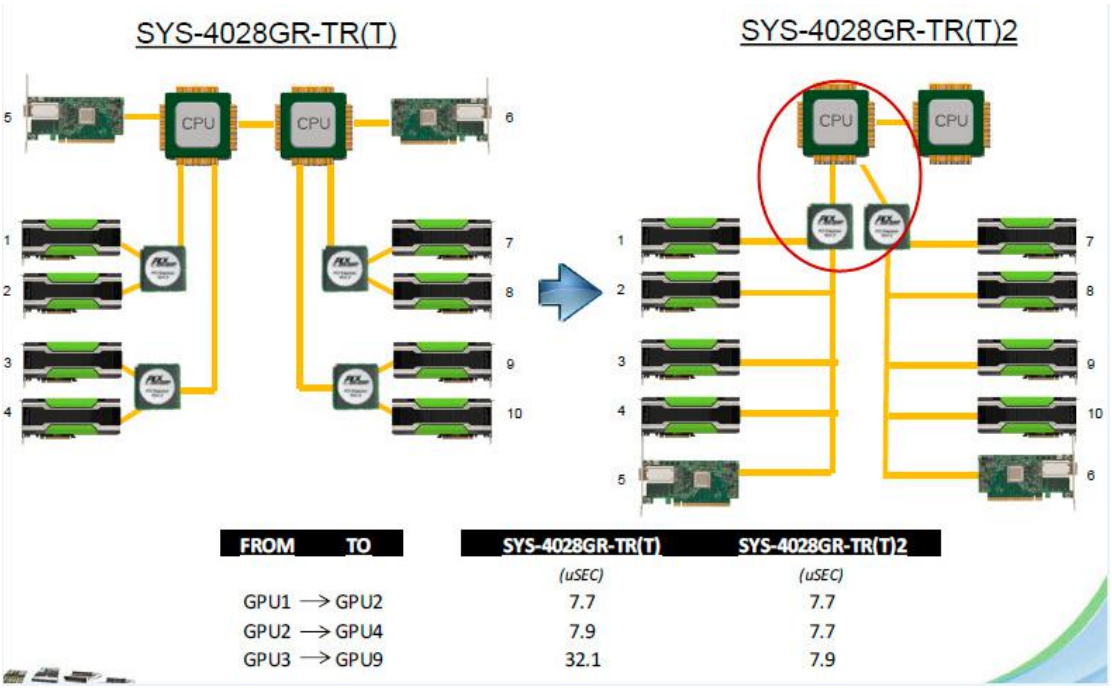
这种情况做优化。

超微 SYS-4028GR-TR2



```
user@user-ubuntu:~$ nvidia-smi  topo -m
        GPU0    GPU1    GPU2    GPU3    GPU4    GPU5    GPU6    GPU7    GPU8    GPU9    CPU Affinity
GPU0     X      PIX     PIX     PIX     PIX     PHB     PHB     PHB     PHB     PHB     0-13,28-41
GPU1    PIX      X      PIX     PIX     PIX     PHB     PHB     PHB     PHB     PHB     0-13,28-41
GPU2    PIX     PIX      X      PIX     PIX     PHB     PHB     PHB     PHB     PHB     0-13,28-41
GPU3    PIX     PIX     PIX      X      PIX     PHB     PHB     PHB     PHB     PHB     0-13,28-41
GPU4    PIX     PIX     PIX     PIX      X      PHB     PHB     PHB     PHB     PHB     0-13,28-41
GPU5    PHB     PHB     PHB     PHB     PHB      X      PIX     PIX     PIX     PIX     0-13,28-41
GPU6    PHB     PHB     PHB     PHB     PHB     PIX      X      PIX     PIX     PIX     0-13,28-41
GPU7    PHB     PHB     PHB     PHB     PHB     PIX     PIX      X      PIX     PIX     0-13,28-41
GPU8    PHB     PHB     PHB     PHB     PHB     PIX     PIX     PIX      X      PIX     0-13,28-41
GPU9    PHB     PHB     PHB     PHB     PHB     PIX     PIX     PIX     PIX      X      0-13,28-41

Legend:

  X    = Self
  SOC  = Connection traversing PCIe as well as the SMP link between CPU sockets(e.g. QPI)
  PHB  = Connection traversing PCIe as well as a PCIe Host Bridge (typically the CPU)
  PXB  = Connection traversing multiple PCIe switches (without traversing the PCIe Host Bridge)
  PIX  = Connection traversing a single PCIe switch
  NV#  = Connection traversing a bonded set of # NVLinks
```

根据上图可以分析出该机型一共有两个 PLX switch 芯片，并且通过一颗 CPU 控制，也就是说所有 GPU 之间互传数据的话，速度都比较理想，且该机型支持高达 10 个 GPU，该机型是 4028GR-TR 的升级版，是目前最适合做深度学习的 GPU 服务器（目前未发现其他厂商有类似产品）。

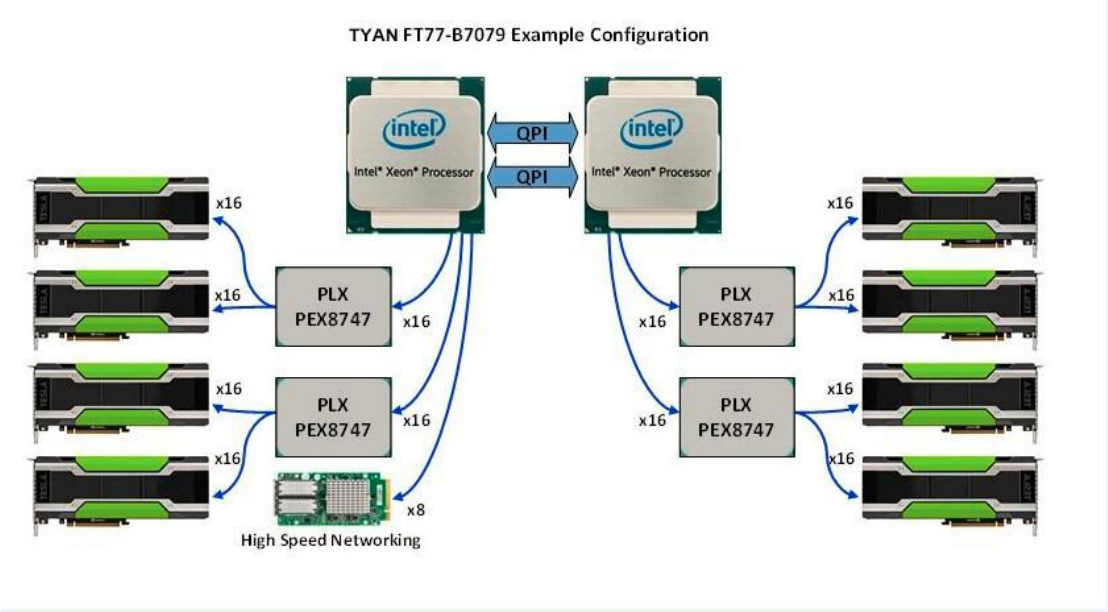


上图为 4028GR-TR&4028GR-TR2 架构参考图。

泰安 7079

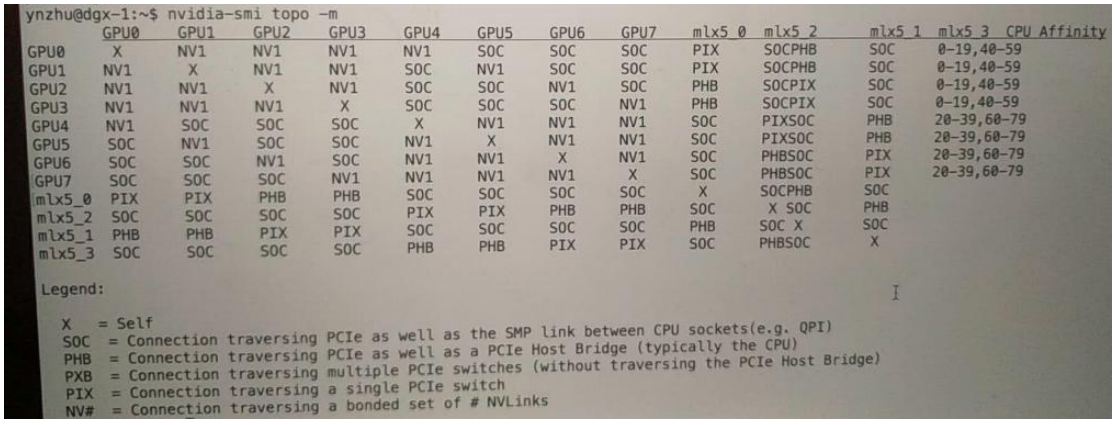
```
user@user-ubuntu:~$ nvidia-smi topo -m
        GPU0    GPU1    GPU2    GPU3    GPU4    GPU5    GPU6    GPU7    CPU Affinity
GPU0     X      PIX     PHB     PHB     SOC     SOC     SOC     SOC     0-7,16-23
GPU1    PIX      X      PHB     PHB     SOC     SOC     SOC     SOC     0-7,16-23
GPU2    PHB     PHB      X      PIX     SOC     SOC     SOC     SOC     0-7,16-23
GPU3    PHB     PHB     PIX      X      SOC     SOC     SOC     SOC     0-7,16-23
GPU4    SOC     SOC     SOC     SOC      X      PIX     PHB     PHB     8-15,24-31
GPU5    SOC     SOC     SOC     SOC     PIX      X      PHB     PHB     8-15,24-31
GPU6    SOC     SOC     SOC     SOC     PHB     PHB      X      PIX     8-15,24-31
GPU7    SOC     SOC     SOC     SOC     PHB     PHB     PIX      X      8-15,24-31

Legend:

  X    = Self
  SOC  = Connection traversing PCIe as well as the SMP link between CPU sockets(e.g. QPI)
  PHB  = Connection traversing PCIe as well as a PCIe Host Bridge (typically the CPU)
  PXB  = Connection traversing multiple PCIe switches (without traversing the PCIe Host Bridge)
  PIX  = Connection traversing a single PCIe switch
  NV#  = Connection traversing a bonded set of # NVLinks
user@user-ubuntu:~$
```

通过上图可以看到泰安 7079 采用和超微 4028GR-TR 类似的设计，这种设计也是目前市面上绝大多数厂商八卡机主流设计。



上图为泰安 7079 架构参考图。

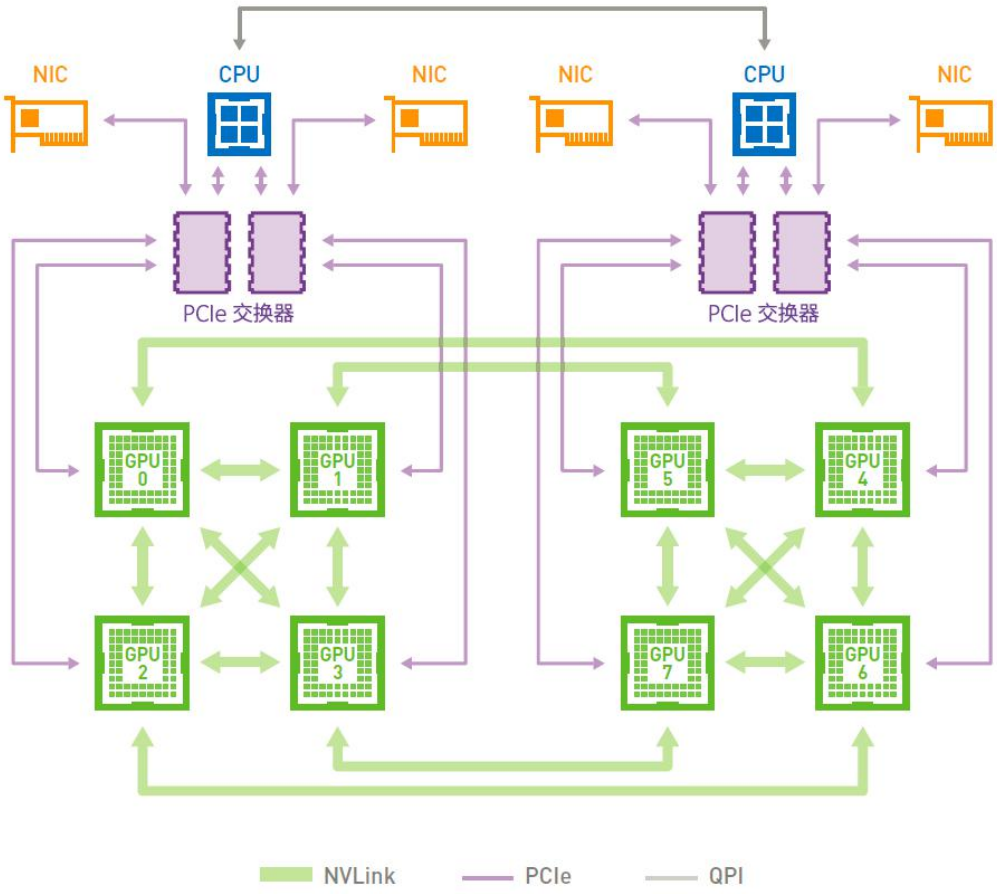
Nvidia DGX-1



```
ynzhu@dgx-1:~$ nvidia-smi topo -m
        GPU0    GPU1    GPU2    GPU3    GPU4    GPU5    GPU6    GPU7    mlx5_0  mlx5_2  mlx5_1  mlx5_3  CPU Affinity
GPU0     X      NV1     NV1     NV1     NV1     SOC     SOC     SOC     PIX     SOCPHB  SOC     0-19,40-59
GPU1    NV1      X      NV1     NV1     SOC     NV1     SOC     SOC     PIX     SOCPHB  SOC     0-19,40-59
GPU2    NV1     NV1      X      NV1     SOC     SOC     NV1     SOC     PHB     SOCPIX  SOC     0-19,40-59
GPU3    NV1     NV1     NV1      X      SOC     SOC     SOC     NV1     PHB     SOCPIX  SOC     0-19,40-59
GPU4    NV1     SOC     SOC     SOC      X      NV1     NV1     NV1     SOC     PIXSOC  PHB     20-39,60-79
GPU5    SOC     NV1     SOC     SOC     NV1      X      NV1     NV1     SOC     PIXSOC  PHB     20-39,60-79
GPU6    SOC     SOC     NV1     SOC     NV1     NV1      X      NV1     SOC     PHBSOC  PIX     20-39,60-79
GPU7    SOC     SOC     SOC     NV1     NV1     NV1     NV1      X      SOC     PHBSOC  PIX     20-39,60-79
mlx5_0  PIX     PIX     PHB     PHB     SOC     SOC     SOC     SOC      X      SOCPHB  SOC
mlx5_2  SOC     SOC     SOC     SOC     PIX     PIX     PHB     PHB     SOC      X SOC  PHB
mlx5_1  PHB     PHB     PIX     PIX     SOC     SOC     SOC     SOC     PHB     SOC X   SOC
mlx5_3  SOC     SOC     SOC     SOC     PHB     PHB     PIX     PIX     SOC     PHBSOC  X

Legend:

  X    = Self
  SOC  = Connection traversing PCIe as well as the SMP link between CPU sockets(e.g. QPI)
  PHB  = Connection traversing PCIe as well as a PCIe Host Bridge (typically the CPU)
  PXB  = Connection traversing multiple PCIe switches (without traversing the PCIe Host Bridge)
  PIX  = Connection traversing a single PCIe switch
  NV#  = Connection traversing a bonded set of # NVLinks
```

由上图可以看到 DGX-1 所有的 GPU 都是通过 NVlink 形式直接或者间接相连，若要完全避免 GPU 通讯延迟问题，程序通讯函数要根据下图架构来做优化，由于接口速度问题（NVlink 单向传输速度比 PCI-E 3.0x16 单向传输速度快 5 倍），采用 NVlink 形式连接，远比市面上

采用 PCI-E 形式连接互传数据要快的多。



上图为 DGX-1 架构参考图。