FPGA 集群上的分布式计算

Why FPGA

知乎问题：**如何评价微软在数据中心使用 FPGA 代替传统 CPU 的做法?**

我对 FPGA 业界主要的遗憾是，FPGA 在数据中心的主流用法，从除微软外的互联网巨头，到两大 FPGA 厂商，再到学术界，大多是把 FPGA 当作跟 GPU 一样的计算密集型任务的加速卡。然而 FPGA 真的很适合做 GPU 的事情吗？前面讲过，**FPGA 和 GPU 最大的区别在于体系结构，FPGA 更适合做需要低延迟的流式处理，GPU 更适合做大批量同构数据的处理。**

因此我们提出了 ClickNP 网络编程框架 [5]，**使用管道（channel）而非共享内存来在执行单元（element/kernel）间、执行单元和主机软件间进行通信**。需要共享内存的应用，也可以在管道的基础上实现，毕竟 CSP（Communicating Sequential Process）和共享内存理论上是等价的嘛。ClickNP 目前还是在 OpenCL 基础上的一个框架，受到 C 语言描述硬件的局限性（当然 HLS 比 Verilog 的开发效率确实高多了）。理想的硬件描述语言，大概不会是 C 语言吧。

低延迟的流式处理，需要最多的地方就是通信。然而 **CPU 由于并行性的限制和操作系统的调度，做通信效率不高，延迟也不稳定**。此外，**通信就必然涉及到调度和仲裁**，CPU 由于单核性能的局限和核间通信的低效，调度、仲裁性能受限，硬件则很适合做这种重复工作。因此我的博士研究把 FPGA 定义为通信的「大管家」，不管是服务器跟服务器之间的通信，虚拟机跟虚拟机之间的通信，进程跟进程之间的通信，CPU 跟存储设备之间的通信，都可以用 FPGA 来加速。

成也萧何，败也萧何。缺少指令同时是 FPGA 的优势和软肋。每做一点不同的事情，就要占用一定的 FPGA 逻辑资源。如果要做的事情复杂、重复性不强，就会占用大量的逻辑资源，其中的大部分处于闲置状态。这时就不如用冯·诺依曼结构的处理器。数据中心里的很多任务有很强的局部性和重复性：**一部分是虚拟化平台需要做的网络和存储，这些都属于通信；另一部分是客户计算任务里的，比如机器学习、加密解密**。我们首先把 FPGA 用于它最擅长的通信，日后也许也会像 AWS 那样把 FPGA 作为计算加速卡租给客户。

另：**深度学习训练和推理有何不同** FPGA 适合对流式数据做推理，利用流水线并行。

Mipsology

https://www.eet-china.com/news/20190117094301.html

我最近参加了在硅谷举行的2018年Xilinx开发者论坛(XDF)。在这个论坛上，我了解到一家名为Mipsology的AI领域初创公司，声称已经解决了采用现场可编程门阵列(FPGA)的AI相关问题。 Mipsology的宏伟愿景是利用FPGA可实现的最高性能来加速神经网络(NN)计算，而不受其部署中固有的限制。

Mipsology展示了每秒可执行超过2万张图像的能力，基于Xilinx新发布的Alveo板，处理一系列NN，包括ResNet50、InceptionV3、VGG19及其它深度学习模型等。

这公司似乎要钱，感觉不太可用。

HLS4ML

https://fastmachinelearning.org/hls4ml/

HLS4ML可以加载Keras代码到HLS。

```
KerasJson: keras/KERAS_3layer.json
KerasH5:   keras/KERAS_3layer_weights.h5 #You can also use h5 file from Keras's model.save() without supplying json file.
InputData: keras/KERAS_3layer_input_features.dat
OutputPredictions: keras/KERAS_3layer_predictions.dat
OutputDir: my-hls-test
ProjectName: myproject
XilinxPart: xcku115-flvb2104-2-i
ClockPeriod: 5

IOType: io_parallel # options: io_serial/io_parallel
HLSConfig:
  Model:
    Precision: ap_fixed<16,6>
    ReuseFactor: 1
    Strategy: Latency
  LayerType:
    Dense:
      ReuseFactor: 2
      Strategy: Resource
      Compression: True

```

`KerasJson, KerasH5` 这些是Keras的模型的文件格式。pytorch对应的是 `pt` 文件。

`weights` 指神经网络的权重，指 hls4ml 可以加载Keras训练好的数据做推理

`XilinxPart` 这个part与我们Nexys4版的part差距很大，不知道我们的板子能不能做。

问题：

- HLS不会用，不知道输出接口是什么样的，能不能和Nexys4DDR适配。

总体考虑