The Blamange design unlocks the full parallel compute power of the Jetson by using inexpensive Quad Ethernet cards in a Flat Neighborhood Network design. This allows us to get close to full utilization (4GBps) from the PCI-e bus and overlap our communication with our computation. Below is a simple example of FNN. It allow each node to see each other node through an unblocked switching system with switch backplane speeds not being a bottleneck. You can get 2GBps per node with 12 GigE cards and three 4-port switches.



Our early experiments with 16, 32, and 48 nodes have scaled linearly on our Tensor tests. We now want to build out the 128 node model to see if it scales to a point that we can estimate the size to recreate the performance of the Google TPUs. <https://en.wikipedia.org/wiki/Tensor_processing_unit>

Below is the wiring diagram for our 128 node cluster. Note that this is before we have access to the Tensor processing available in the Volta architecture. That will be another game-changer.

