Infiniband 调研报告

W Huang

日期：2023 年 10 月 17 日

1 需求描述

我们当前并行计算的主要问题有两个：

1. 单机计算内存带宽不够，导致数据阻塞，只有一两个核能充分工作。要直接解决这个问题，只能更换主板。
2. 多机通讯采用了以太网，延迟太高，导致 mpi 函数很慢。

本文关注第二个问题。我们现在的以太网设备速率是 10Gbps，对当前项目来说已经足够，当然不排除未来需要更高速率的可能性。目前以太网不能满足延迟要求，我们考虑将网卡、网线、交换机全套设备更换成 infiniBand。

我们采用 osu benchmark 进行测试，我们的以太网的通讯延迟大约是 infiniBand 的 32 倍。其中 infiniBand 的测试在学校超算平台进行。

2 采购方案

当前市场上 infiniBand 的品牌只有 Mellanox 一家，主流配置有 100Gbps/200Gbps/400Gbps 三种。考虑到我们的计算并不是运行在 GPU 上的数据密集型计算，采用 100Gbps 组网方案即可。采购单如下（南京星涌网络科技有限公司报价）

产品类型	型号	关键参数	单价（估）
交换机	MSB7800-ES2F	36 口，单口 100Gbps，吞吐量 7Tbps，延迟 90ns，136W	89000
网卡	MCX653105A-ECAT	6 代产品，双口连接，100Gbps，延迟 600ns	5300
	MCX653106A-ECAT	6 代产品，单口连接，100Gbps，延迟 600ns	4300
	MCX555A-ECAT	5 代产品，单口连接，100Gbps，延迟 600ns	3900
网线	MFA1A00-E005	100Gbps，5 米	3800

选择最便宜的网卡，总估价为 $88000 + 16 \times 3900 + 16 \times 3800 = 211200$ 元。

双口的作用仅仅是防止意外损坏，不能让速度翻倍。另外 Mellanox 的产品从 5 代到 6 代只是把上限从 100Gbps 提高到了 200Gbps，我们 100Gbps 的配置没必要买 6 代产品。

采购可以联系星涌网络销售人员（康亮，微信号 luyeesk），或者咨询其它国内代理商。