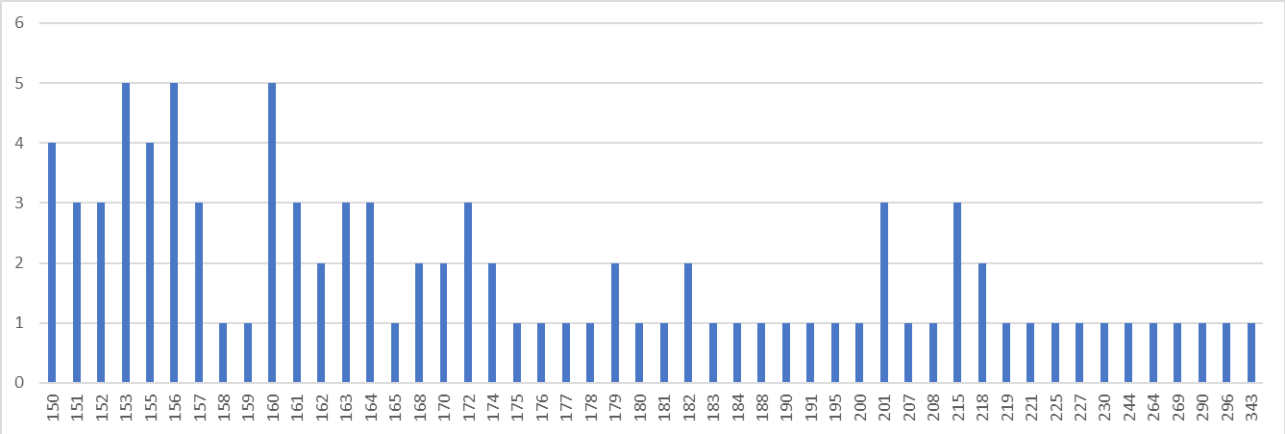
EECE 5640
HOMEWORK 5

Shreyash Jadhav

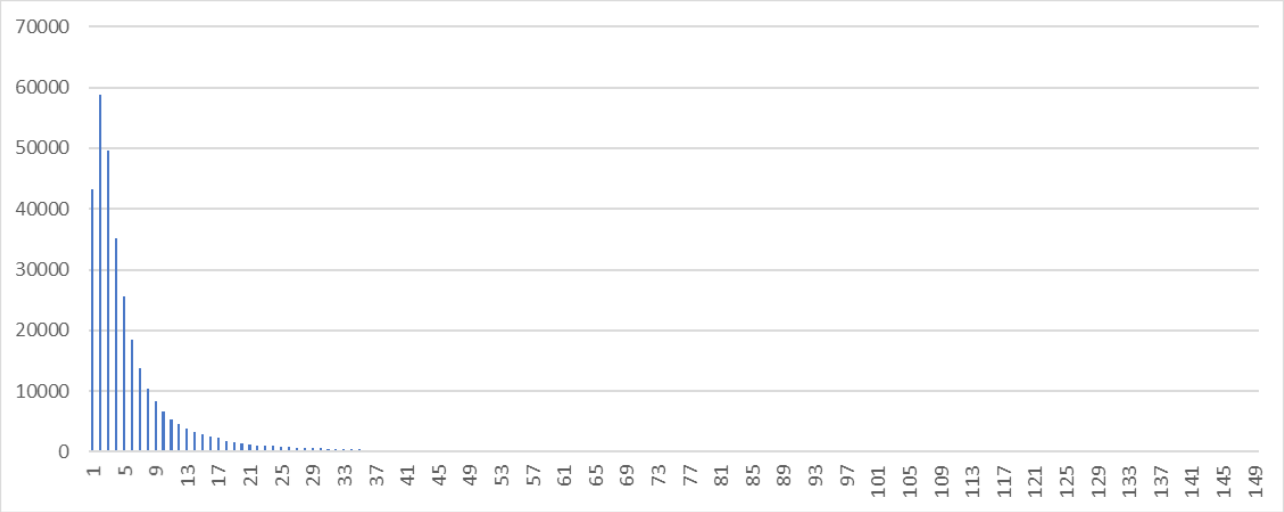
Q1.

Author with most number of co-author is Author no. 3336 with 343 co-authors.
Below is the distribution of the co-authors and authors.
The X-axis indicates the number of co-authors of a certain author and Y-axis indicates the number of authors that correspond with number of author on X-axis.
We can see that the most authors have 2 co-authors seen from the Lower 50% graph.

Top 50%:



Lower 50%:



Q2.

Performance wise CUDA is perform better compared to MPI because it will not have communication overhead that MPI program have if they are allocated different nodes.
CUDA cores are on the single chip hence memory transfers will be faster in case of CUDA. But when there are any conditional loops MPI will perform better as it using CPU for computation which have branch predictor and other mechanism to improve the performance and which GPU lacks.

Q3. Read the Pascal whitepaper provided, and then identify the key features that were introduced in the Pascal P100 architecture, as compared against the Volta V100 architecture (make sure to identify the source for the information you obtained on the V100). Please do not just repeat what you read in the Pascal whitepaper, go into more detail on each of the features you identify.

Pascal P100 vs Volta V100

Pascal:

- **NVLink**: 160 GB/s bidirectional GPU to GPU communication With the increase in the use of the GP-GPU and the ratio of the GPU to CPU decreasing, PCIe have become the bottleneck in the communication where CPU and GPU are connected. Alone PCIe cannot handle the communication between the GPU and GPU in multi GPU system. So the new NVLink provides with the high speed bidirectional communication between the GPU's in the system.
- **HBM2**: Stacked High Bandwidth Memory. This memory is provided on the same physical package as GPU and thus helps in reducing the power consumption. It has also provided a 3X increased memory bandwidth compared to its predecessor.
- **Unified Memory**: In order make programmers life easier this feature was introduced in the architecture. This provides with unified virtual address space for CPU and GPU so that the programmer does not have to worry about underlying hardware. This also supports hardware pagging faults which is combined with the 49-bit virtual address space (512 TB) which cobined helps in transparent migration of data between virtual address space of CPU and GPU.
- **Preemptive compute**: This feature allows compute tasks to be pre-empted at an instruction level rather than the thread level which was the case in the previous generations. This feature prevents long running programs in monopolizing the system resources or timing out.
- **Support for FP16** : This feature was added for the deep learning application. For deep learning we require layers to train our models and depending upon the weights and biases. These are mostly small number and does not require FP32 or FP64. So using FP32 we can save the memory usage and thus speed up the training peroid.
- Used **16nm FinFET** to decrease the power consumption.

V100:

- **Tensor cores** where introduced which had major performance impact for deep learning applications.
- V100 had low cache hit latency due to the unified memory for L1 cache which yielded boost in the performance of the system.
- There was **improved** in the **HBM2** and **Nvlink 2.0** which helped r**educe the latency**.
- Improved in FinFET technology to **12nm FinFET** provided better power efficiency.

References:
1. https://en.wikipedia.org/wiki/Volta_(microarchitecture)
2. https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf