Course: IST-652
Final Project

Instructor: Professor D. Landowski
Student: Jeff Victor

Introduction

Computer manufacturers compete for business; each wants to offer the fastest computer. In order to design the fastest computer, they must know which type of component will provide the largest performance boost.

This project explored the factors that contribute to the increase in performance of business workloads due to improvements in computer technology. It determined the factors that contribute the most to performance.

Objective

Determine whether computer performance data can justify a conclusion about the computer component choices made by computer manufacturers.

Data and Software

The data used for this project included submissions for the SAP SD 2-Tier benchmark will be used for this project.

Each record includes 26 fields. Of those, seven are performance results, another seven are types or quantities of physical resources that contribute to performance (e.g. amount of RAM), and three are software types and versions.

There are over 1,000 records in the dataset.

The data is freely available at https://www.sap.com/dmc/exp/2018-benchmark-directory/assets/export-sd.csv .

Through experimentation, the following fields were identified as the most important contributors to performance increases:

- Date of submission
- Performance (the "SAPS" metric)
- Quantity of CPU cores in the computer
- Performance of each CPU core
- CPU clock rate

After that was determined, software to clean the data was improved. Preprocessing steps included:

- Remove unnecessary fields
- Fix a small number of records that had an obvious mistake and enough information in the other fields to fix the mistake
- Remove records that used a performance measurement system that was sufficiently different that it invalidated a comparison of data
- Remove records that were missing one of the key data fields

- Fill in empty fields that could be derived from other fields
- Simplify categorical fields to ease potential opportunities for aggregation
- Convert date information to a standard format
- Write the preprocessed data to a new data file.

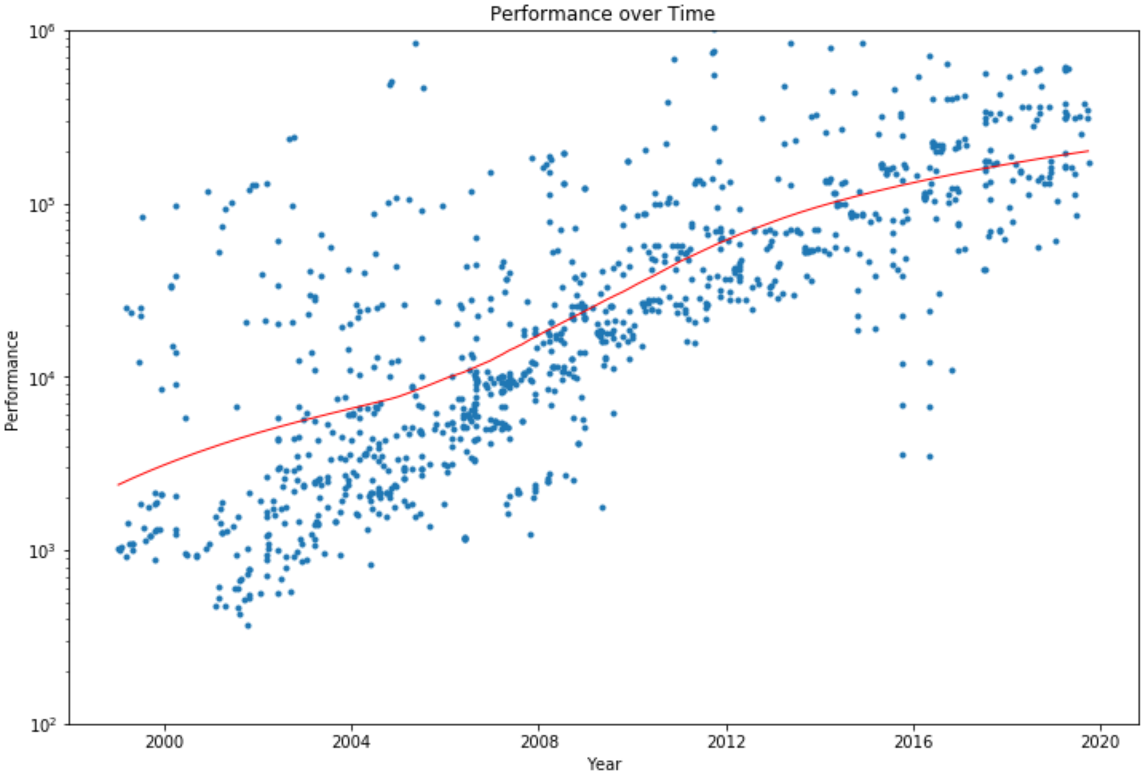
Software Data Analysis

The preprocessed data was read into a "pandas" dataframe to enable use of the many tools included in that package. For example, the submissions were aggregated by determining the median value for each component across each year. Other fields were derived from the fields in the cleaned data, including:

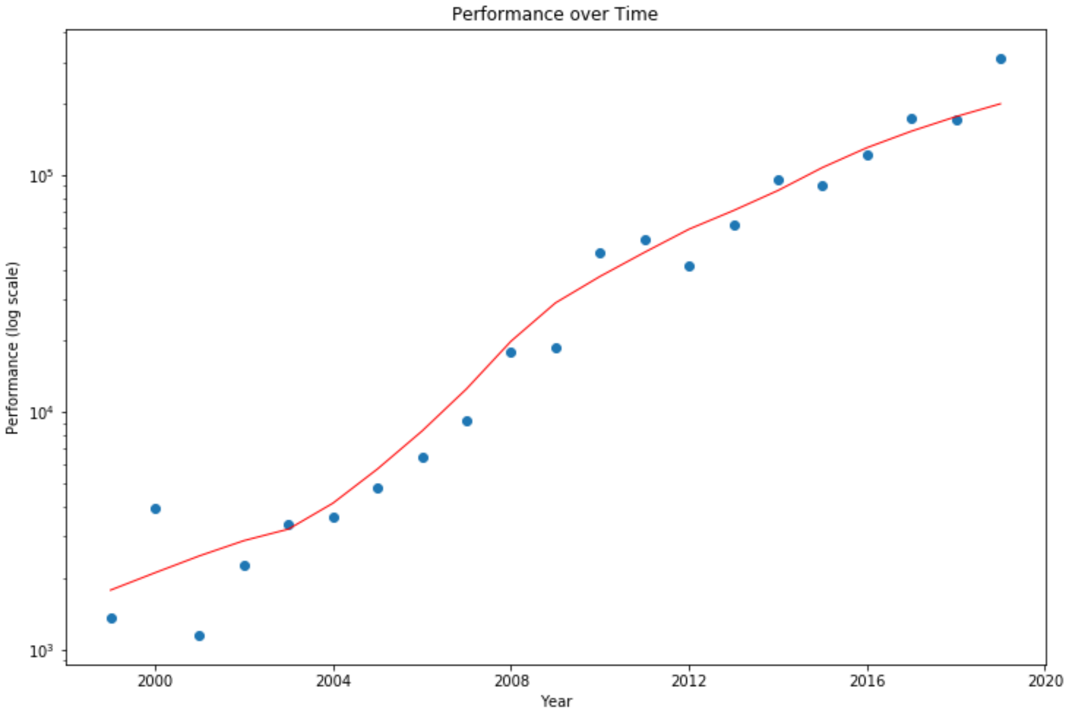
- Performance per CPU chip
- Performance per CPU core
- Quantity of CPU cores per CPU chip

In order to better understand the changes over time, additional fields were added that store the percentage change in all of those values, per year. This completed the process of deriving additional data fields.

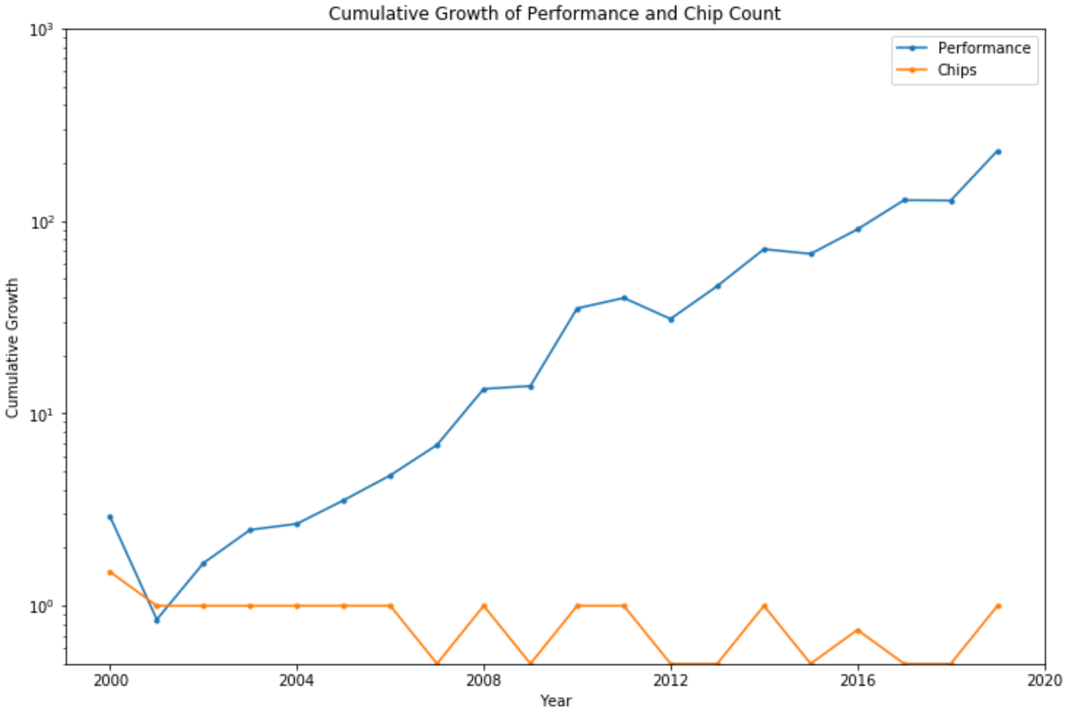
The data was explored through plots. The first plot was used to determine if the data had so much noise that trends would be difficult to identify. Fortunately, the performance data made a clear trend. A Lowess curve was added to the scatter plot to highlight the trend. Note that this plot has a log scale for the y-axis. The plot shows exponential growth over 20 years, slowing only slightly for 2016-2019 compared to the period 2004-2016.



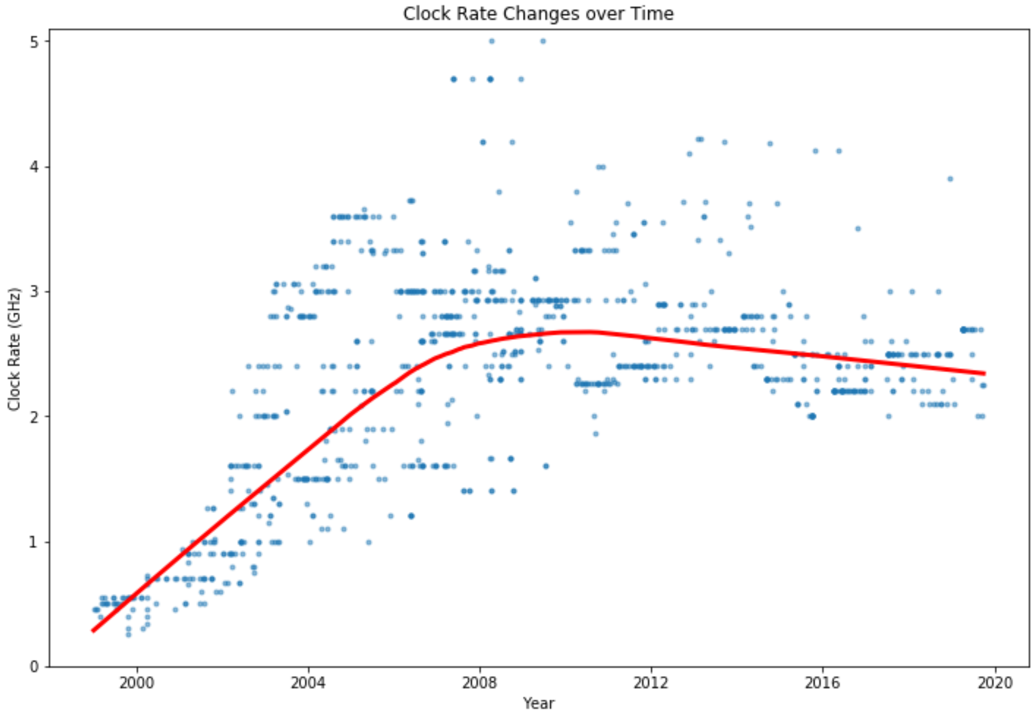
However, there is enough noise that the median annual values were plotted to improve trend visiability. This removed the outliers and the trend is even clearer.

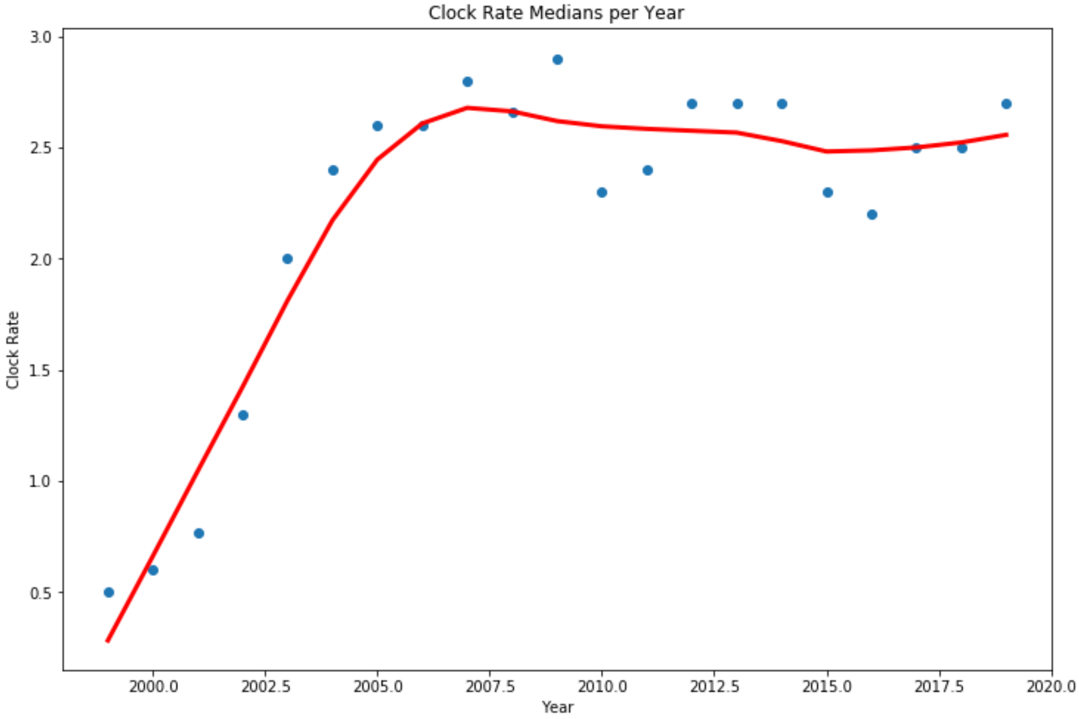


A simple plot of two values showed that, in general, computer manufacturers did not increase the number of CPU chips in order to increase performance. For this reason, this field was ignored for the rest of the analysis.

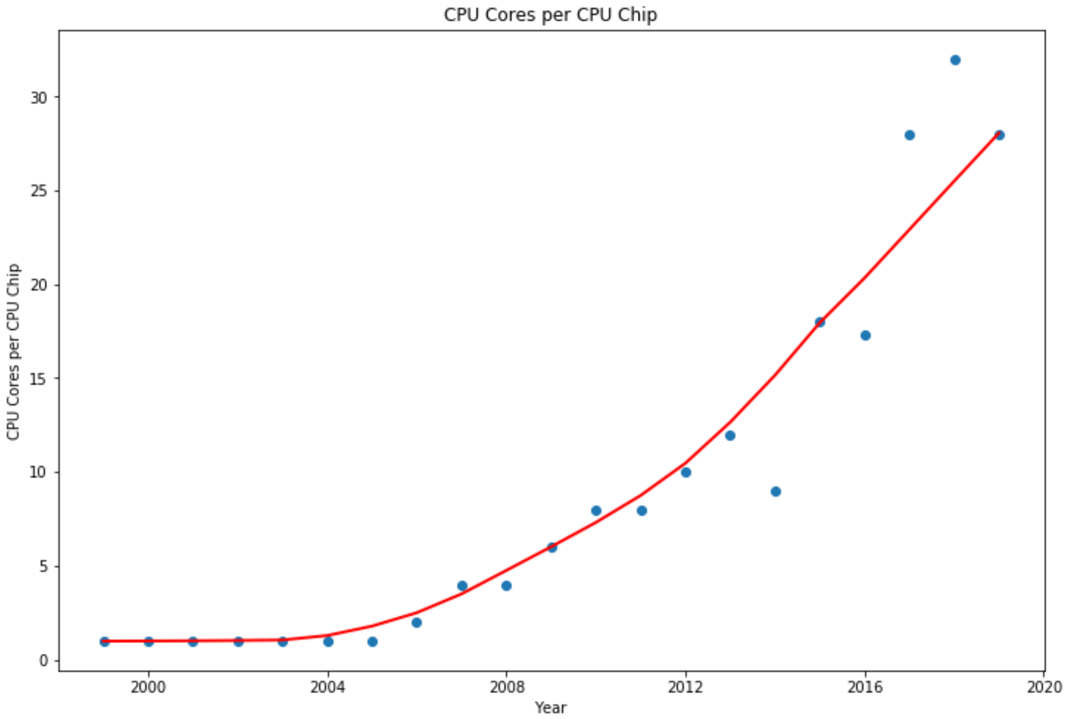


A pair of plots similar to the performance trend plots display the CPU clock rate for each benchmark submission. They showed significant growth in clock rate from 2002 to 2006, followed by a slow **decrease** in clock rate. If clock rate contributed to the overall exponential performance increase, it did so only until 2006.

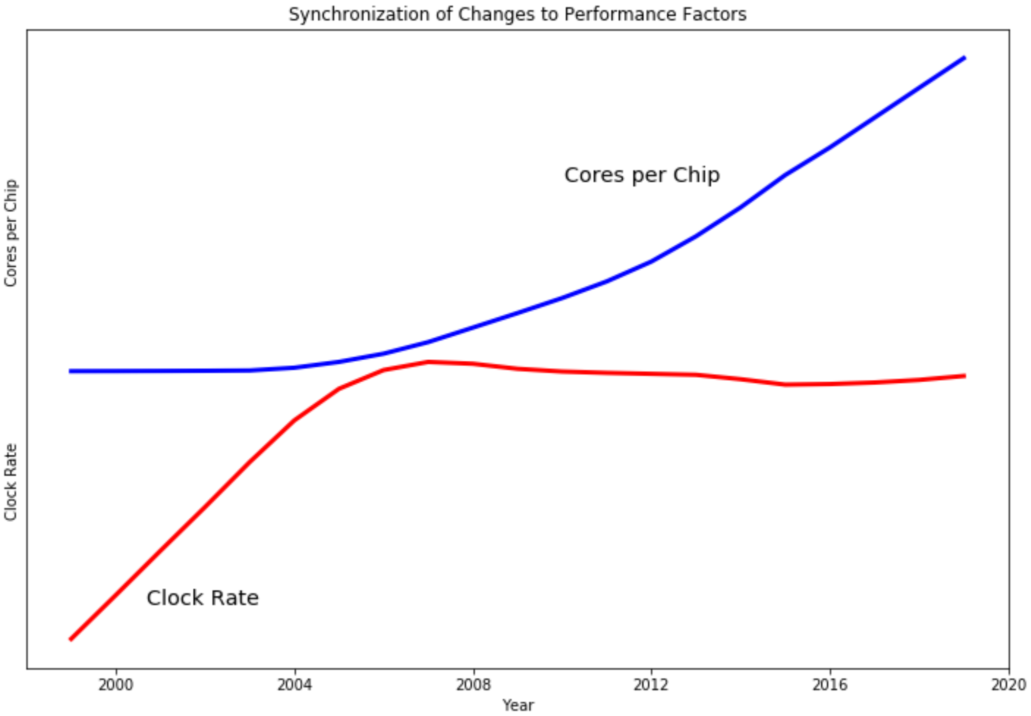




A plot of CPU cores per CPU chip showed that originally, CPU chips only had one core each. This changed starting in 2006, and the growth was close to exponential until about 2013, then showed a linear increase.



Noticing that clock rate stopped increasing at about the same time as core count began increasing, both were plotted together, normalized and scaled.



Conclusion

Computer performance, as measured by commonly used benchmark software, provided important insight into choices made by computer manufacturers. Before 2006, they relied heavily on increases in CPU clock rate to deliver regular increases in computer performance.

By 2006, CPU manufacturers had discovered that further clock rate increases would be limited by the laws of physics. After 2006, they relied largely on increasing the number of CPU cores that were built into each CPU chip.

The data shown above clearly illustrate this shift.