SWITCHengines for processing big datasets in research

SEPBDR **E** 4.2017

An application in informatics

Andri Lareida, Institute for Informatics, University of Zurich

My Research

My research interest mainly lies with decentralized peer-to-peer (P2P) systems, focusing on using information present in networks to predict user behavior and on exploiting locality in overlay networks. Although, video streaming is causing the main part of Internet traffic, peer-to-peer (P2P) file sharing applications still account for a large portion of the total Internet traffic. The most prominent among these applications is BitTorrent. Internet Service Providers (ISPs) face the challenge of network congestion during peak traffic hours, which is made more difficult by this P2P traffic.

My Challenges

As part of my PhD thesis, I collected a data set of peers being active in the BitTorrent network. Over the course of 3 months, traffic data of more than 70,000 shared files were collected, of which at least 10,000 were active on a single day. The collected raw data is 5 TBytes big and was reduced to 1 TByte for further analysis. Processing that data with traditional methods would have been extremely complex. Therefore I started to look at big data analysis tools like Hadoop and Spark. With Spark, the processing was very efficient, but still even processing data of one day takes 20 minutes. For the infrastructure to run my Spark jobs, I used existing hardware at our lab. With this cluster, processing the whole dataset takes over 30 hours.

My Solution

Out of curiosity, I wanted to compare the performance of our local hardware with a virtual environment like SWITCHengines. I expected SWITCHengines to be less efficient due to its virtualization of storage and compute power. I transferred my data and compute jobs to a Spark cluster on SWITCHengines. After tuning the installation, I got a surprising result: SWITCHengines was as fast as our hardware in the lab. Therefore, SWITCHengines is a very good alternative to our cluster. I could imagine going for SWITCHengines when the replacement is due. So we would not have to care about hardware maintenance and still get the compute power when we need it. Furthermore, I could profit from the scalability of SWITCHengines: With a higher number of machines I get the results for my research much faster!

Further information:
swit.ch/engines