Energy Efficiency in Large Scale Information Retrieval Systems

Matteo Catena
ISTI-CNR
National Research Council of Italy
Pisa, Italy
matteo.catena@isti.cnr.it

Abstract

Web search engines are large scale information retrieval systems, which provide easy access to information on the Web. High performance query processing is fundamental for the success of such systems. In fact, web search engines can receive billions of queries per day. Additionally, the issuing users are often impatient and expect sub-second response times to their queries (e.g., 500 ms). For such reasons, search companies adopt distributed query processing strategies to cope with huge volumes of incoming queries and to provide sub-second response times.

Web search engines perform distributed query processing on computer clusters composed by thousands of computers and hosted in large data centers. While data center facilities enable large-scale online services, they also raise economical and environmental concerns. Therefore, an important problem to address is how to reduce the energy expenditure of data centers. Moreover, another problem to tackle is how to reduce carbon dioxide emissions and the negative impact of the data centers on the environment.

A large part of the energy consumption of a data center could be accounted to inefficiencies in its cooling and power supply systems. However, search companies already adopt state-of-the art techniques to reduce the energy wastage of such systems, leaving little room for more improvements in those areas. Therefore, new approaches are necessary to mitigate the environmental impact and the energy expenditure of web search engines.

One option is to reduce the energy consumption of computing resources to mitigate the energy expenditure and carbon footprint of a search company. In particular, reducing the energy consumption of CPUs represents an attractive venue for web search engines. Currently, CPU cores frequencies are typically managed by operating system components, called frequency governors. We propose to delegate the CPU power management from the OS frequency governors to the query processing application [2]. Such search engine-specific governors can reduce up to 24% a server power consumption, with only limited (but uncontrollable) drawbacks in the quality of search results with respect to a system running at maximum CPU frequency.

Since users can hardly notice response times that are faster than their expectations we advise that web search engine should not process queries faster than user expectations and,

consequently, we propose the Predictive Energy Saving Online Scheduling (PESOS) algorithm, to select the most appropriate CPU frequency to process a query by its deadline, on a per-core basis [3]. PESOS can reduce the CPU energy consumption of a query processing server from 24% up to 48% when compared to a high performance system running at maximum CPU core frequency.

To reduce the carbon footprint of web search engines, another option consists in using green energy to partially power their data centers. Stemming from these observations, we propose a new query forwarding algorithm that exploits both the green energy sources available at different data centers and the differences in market energy prices [1]. The proposed solution maintains a high query throughput, while reducing by up to 25% the energy operational costs of multi-center search engines.

Supervisor: Dr.-Ing. Nicola Tonellotto (ISTI-CNR)
Available from: `http://hpc.isti.cnr.it/~catena/docs/thesis.pdf`

References

[1] Roi Blanco, Matteo Catena, and Nicola Tonellotto. Exploiting Green Energy to Reduce the Operational Costs of Multi-Center Web Search Engines. In *Proc. WWW*, pages 1237–1247, Montreal, Canada, 2016. IW3C2.

[2] Matteo Catena, Craig Macdonald, and Nicola Tonellotto. Load-sensitive CPU Power Management for Web Search Engines. In *Proc. SIGIR*, pages 751–754, Santiago, Chile, 2015. ACM.

[3] Matteo Catena and Nicola Tonellotto. Energy-efficient query processing in web search engines. *IEEE Transactions on Knowledge and Data Engineering*, 29(7):1412–1425, 2017.