

# Towards energy-efficient workload placement in Data Centers

Rania Elnaggar (student & presenter) , Raj Yavatkar

Intel Corporation  
{rania.elnaggar, raj.yavatkar}@intel.com

Jonathan Walpole

Portland State University  
walople@cs.pdx.edu

A new era of computing is characterized by a shift to aggregate computing resources in large-scale *Data Centers* (DCs) with the advent of new technologies such as Cloud Computing. Energy efficiency becomes critical to DCs as they continue to grow in size and capacity. In fact, energy costs in DCs maintain an upward trend that is poised to surpass the cost of the equipment and infrastructure in a typical lifetime usage model. We distinguish between two types of DC energy-related costs; first, the cost of power, cooling and facilities provisioning in *capital expenditure* (CapEx); and second, the operational energy cost of powering servers, cooling and facilities equipment in *operational expenditure* (OpEx). The CapEx cost component is proportional to peak power and set power budget, while the OpEx component is proportional to average power. In our research we focus on improving operational energy expenditure and hence do not confront efficiencies related to under/over capacity provisioning.

We contend that achieving improved *energy-efficiency* for a DC should be driven by a holistic approach that takes into consideration all system components to achieve maximum synergetic energy savings. This strategy governs a set of local policies and protocols that effectively use existing power-saving features within each system component in a way that is proportional to the workload of the component, and to that of the overall system. In our work we present an example of such a holistic strategy that distributes workloads to servers within the DC such that the overall power consumption of the DC is minimized, for a given total workload. In defining such a strategy, we initially examine the compute-related part of the DC, thus excluding networking, cooling and power delivery overheads. We consider a DC workload that is a mix of online and offline workloads that are mostly characterized as massively parallel, or throughput-oriented.

Our strategy treats a DC's power expenditure as a collective cost function based on the aggregate power costs obtained from individual load-based *power profiles* of all the servers within the DC. The purpose of a power profile is to chart a server's power usage over a normalized workload range from 0% to 100% of the server's maximum capacity, represented using a throughput metric of jobs/sec. Such a power profile treats the server as a *black box* and does not interfere with the server's own fine-grained power management features implemented either through Dynamic Voltage and Frequency Scaling (DVFS), or through putting components to sleep. An example of such a standard power profile is the one specified in the SPEC Power benchmark. We consider such a power profile to be a valid representation of a server's energy proportionality.

In our methodology, for a given DC throughput level, workloads are distributed to servers such that the aggregate cost of the overall DC compute power is minimized. In our initial investigation we only consider a DC system comprised of homogeneous servers that share the same power profile. We have obtained preliminary results that show the validity of our approach and how it exploits power saving opportunities. We also show that operating servers at their maximum capacity is not always an energy-efficient choice because of increasing non-linearity in the power profiles of modern servers. Our cost function can be extended to encompass additional metrics such as electricity prices, cooling and networking costs.

We also show how the same approach is valid for heterogeneous DCs and workloads, and can be used in multi-DC workload placement scenarios. It is worth noting that our approach avoids the need to coordinate power management policies of the numerous servers and components within the DC; it also eliminates requiring control interfaces (such as ACPI) to be defined at various levels within a DC's hierarchy, and thus greatly simplifies DC-level power management. Our strategy can hence be classified as *passive (or uncoordinated) power management*.

Addressing DC-level power management through workload distribution was previously attempted by simply powering-off some servers while operating the remaining servers at full capacity, according to the total DC workload; however, the resulting power-on latency makes this approach inefficient in term of responsiveness. We propose keeping an active set of servers that are not powered-off when idle but are rather put into a low-power active state (active-idle). The number of servers in the active set is determined by taking into consideration the latency, workload patterns and Service Level Agreements (SLAs).

**Demo:** Yes