Best Practices in Energy-Efficient High Performance Computing

Michael Ott[1], Dieter Kranzlmüller[2]

Abstract: High Performance Computing is a key technology for Environmental Computing and will continue to facilitate future research and breakthrough discoveries in the field for years to come. However, on the doorstep to Exascale computing, it becomes ever more clear that HPC by itself is major consumer of energy with a significant environmental footprint and keeping HPC sustainable also in future will require considerable effort. With first Exascale systems being projected at a power envelope of 30–40 MW, every percentage point in increased energy efficiency will count and every promising technology to increase it should be explored. Among the interesting technologies for energy efficient HPC are direct-liquid hot-water cooling and adsorption chilling that both allow for reducing the energy required to cool computer systems, as well as energy-aware scheduling that can help in reducing the power consumption of supercomputers. This paper will give an overview of these technologies and report on the operational experience at the Leibniz Supercomputing Centre that has been exploring and driving these technologies for years.

Keywords: Energy-efficiency; HPC; Direct-Liquid Cooling; Adsorption Refrigeration; Energy-Aware Scheduling

1. Introduction

Many applications in Environmental Computing rely on numerical simulation or other computational intensive methods for which High Performance Computing (HPC) is an indispensable tool. As the performance of supercomputers keeps to grow, ever larger datasets can be crunched and more detailed simulations performed on such systems. While these increasing capabilities facilitate deeper insight and novel scientific results, they also come at a cost in terms of environmental impact itself as the most capable supercomputers (as measured by the Top500 list of the fastest HPC systems in the world [Me14]) are not only becoming faster but also more energy hungry: in the November 2007 list, the #1 system in the Top500 consumed 2.3 MW (BlueGene/L, Lawrence Livermore National Laboratory); in November 2012, the power consumption of the fastest system in the world was 8.2 MW (Titan, Oak Ridge National Laboratory); the #1 system in the latest list from November 2017 consumes 15.4 MW (Sunway TaihuLight, National Supercomputing Center in Wuxi); and the first Exascale systems to be deployed

[1] Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities, 85748 Garching, Germany,michael.ott@lrz.de
[2] Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities, 85748 Garching, dieter.kranzlmueller@lrz.de

in the 2021-2022 timeframe are even projected to consume more than 30 MW of electrical power. Although this number will be lower with later deployed Exascale systems as technology advances, there is a clear trend in continuously increasing power consumption of HPC systems. Given that the majority of the world's electrical energy is generated from fossil fuels, this kind of power consumption results in a significant carbon footprint that renders a technology that is supposed to help with environmental problems into an active contributor towards the very same problems. It is therefore important that such systems are operated as energy efficient as possible to do as less harm to the environment as possible.

Besides the environmental implications of this continuous increase in power consumption, there are also very tangible economic and scientific implications: an ever larger fraction of the total budget for HPC systems is being spent on electricity bills instead of hardware (and hence on computational performance). Ultimately this means that less science can be performed within a given budget, particularly in Europe where electricity prices are generally higher than in other parts of the world such as the US or Asia. Energy efficient operation of such systems is therefore vital to allow for more science per Euro or Dollar spent.

The remainder of this paper is organized as follows: Section 2 will give an overview on energy-efficient HPC operations with a particular focus on the "4 Pillar Framework for energy efficient HPC data centers". Sections 3 to 5 provide an overview over key technologies for energy efficient HPC operations and Section 6 will give a summary.

2. Energy Efficiency in HPC

Energy efficiency in HPC must be tackled from multiple angles, from the hardware and infrastructure side as well as from the software side. In the "4 Pillar Framework for energy efficient HPC data centers" [WAS13] the authors list four key aspects of HPC operation that need to be optimized: applications, system software, system hardware, and building infrastructures:

- *Applications* need to be optimized for efficient execution on a given system hardware. Although some optimization techniques exist that aim particularly at improving energy efficiency (such as choosing algorithms that yield the same results with less energy, e.g. for sorting [Bu09]), any optimization that reduces the runtime of an algorithm (time to solution) will ultimately result in lower energy consumption (energy to solution). Therefore, classical and well-known performance optimization techniques can be applied to achieve this goal.
- *System Software* encompasses all software components that are required to execute applications on a particular system hardware, such as the operating system, runtime libraries, and workload schedulers. They facilitate the use of power saving features in the hardware, provide for efficient execution of applications, and

enforce energy saving policies. The optimization goal here is to reduce the energy consumption of the system without impacting application performance.

- *System Hardware* comprises not only the actual computational hardware but also networks and auxiliary components such as storage systems or backup archives. The hardware choice will have a significant impact on the energy consumption of the overall system and therefore needs to be performed wisely during the procurement process as later optimizations are only hard to achieve. In particular, the computer architecture needs to allow for efficient execution of the application mix that will run on the system.
- *Building Infrastructures* summarize all non-IT infrastructures that are required to operate an HPC system. Optimizations in this pillar have to cover power delivery, cooling, and re-use of generated waste heat.

While each of these four pillars may be approached separately, it will in general be more efficient to cover as many of them simultaneously in a holistic approach. For example, energy efficiency features of the system hardware can only be properly exploited with optimized system software. Additionally, operational parameters of the building infrastructure could be fed to the workload scheduler of the system software stack, e.g. to execute energy hungry applications only when there is sufficient headroom in cooling capacity. Likewise, information on future workloads could be communicated from the workload scheduler to the building infrastructure in order to operate the cooling and power delivery systems at their energy efficiency sweet spot. The possibilities are endless and only limited by complexity and available work force to implement the connections between different components.

For a pure HPC data centre such as the Leibniz Supercomputing Centre (LRZ) that has only limited control over the applications that are being run on its HPC systems by its clients, it may be hard or impossible to tackle the applications pillar. Nevertheless, even in such an environment the users can be encouraged to save energy where possible if they are given access to the necessary tools. For example, the energy consumption of an application run could be made available to the user to give them a feeling of the energy that is consumed by their jobs. In fact, most batch scheduling systems allow for reporting consumed energy as part of the standard job report. On top of that, access to more fine-grained energy measurements would also allow the users to optimize specific regions of their code.

Besides supporting its users in energy efficiency optimizations of their applications, efforts at LRZ concentrated mainly on the remaining three pillars in its long history on energy efficient HPC operations: In 2010, it pioneered direct-liquid hot-water cooling (HT-DLC) and adsorption chillers with the CoolMUC-1 system in order to evaluate technologies for reducing the electrical power consumption of the cooling infrastructure. In 2012, LRZ deployed its HPC flagship SuperMUC, the first Top10 system to use HT-DLC and energy-aware scheduling mechanisms for increased energy efficiency. In 2016, it installed second generation adsorption chillers for production use in its CoolMUC-2 compute cluster. In 2017, CoolMUC-3 has been deployed at LRZ, the first HPC system that employs HT-DLC for all components and hence operates completely without energy hungry air-cooling. Later in 2018, LRZ will put its latest flagship SuperMUC-NG into operation that will combine advanced HT-DLC with next-generation adsorption refrigeration.

The following sections will describe each of these technologies and their contributions towards energy efficient HPC operations in more detail.

3. Direct Liquid Hot Water Cooling

As the first law of thermodynamics also holds for HPC, all electrical energy consumed by an HPC system is ultimately transformed into (waste-)heat and, depending on the point of view, computational results are a mere by-product. Since removing the heat energy from the compute system and disposing of it requires additional electrical energy, it is important to do so in an energy-efficient manner in order to reduce the overall energy consumption of the system.

Most data centers use air cooling according to ASHRAE class A1 specifications [AS11a]. In such an environment, cold air of 18°C–27°C is intaken by fans from the front of a compute node, cools the hot components, and is being blown out at higher temperatures at the back of the node. The warm air then needs to be cooled down by computer room air handlers (CRAH) before it is being returned to the computer room. The CRAH units themselves use chilled water (<14°C) to reject the heat. Unless operated in climate zones that can sustain such low temperatures with ambient air, energy hungry mechanical chillers are usually required to generate the necessary cold. Depending on the climate conditions of the operating site, the typical energy consumption of the cooling infrastructure for such air cooled installations can be as high as 30% of the IT energy consumption.

In high performance computing, some sites have switched to liquid cooling according to ASHRAE W1-W5 standards [AS11b]. For many of those sites, the main incentive to switch to liquid cooling is the fact that it allows for higher compute densities and more powerful processors due to the superior thermal properties of water over air. Yet, if operated at sufficiently high water temperatures, (>30°C, ASHRAE W3-W5), water cooling also allows for heat rejection without energy hungry chillers as such high temperatures can be sustained in most climate zones year round with cooling towers only. This so-called free-cooling can reduce the energy spent on cooling by as much as 90%. However, such high temperatures require the cooling liquid to be delivered as close to the hot components (processor, graphics card, memory, voltage regulators) of the compute node as possible to operate them within their thermal specifications. In order to achieve this, most vendors use water-cooled heat sinks that attach directly to the hot components and allow for sufficient heat transfer from them directly to the water (see Fig. 1 for an example of a water-cooled compute node).

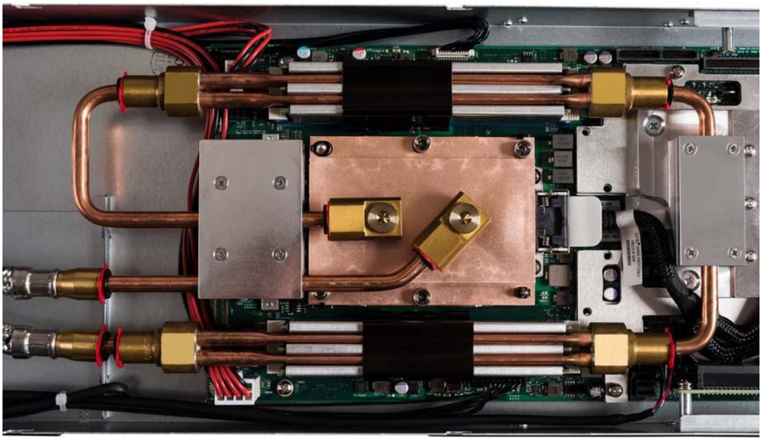

Fig. 1: Direct-liquid water-cooled compute node (Megware SlideSX®-LC, photo courtesy of Megware Computer Vertrieb und Service GmbH)

Due to water's superior thermal properties, even 45°C water keeps the hot components in such a liquid cooled compute system cooler than in a 20°C air cooled system [AH12]. This technology is called Direct Liquid Hot Water Cooling (HT-DLC).

LRZ started deploying HT-DLC systems as early as 2010 with the CoolMUC-1 Linux cluster. Ever since, all HPC systems at LRZ have been based on this technology, including the flagship system SuperMUC (phase 1 and phase 2). While CoolMUC-1 was a first-of-a-kind system with many custom-built parts, the technology has significantly matured over time and is now readily available from multiple vendors off the shelf. However, even the first generation systems did not show any higher failure rates than standard air cooled systems and there was no single instance of water leakage into the computer room many data centre operators might be afraid of.

The latest development in HT-DLC is to cover as many components of the IT system possible with direct liquid cooling. While the first systems only covered CPUs and processors that account for 70% of a compute nodes power consumption, 100% HT-DLC systems are now available. They also cool voltage regulators, network components, and power supplies directly with hot water and hence avoid the need for any air flow through the compute racks. Consequently, the racks can be thermally insulated which reduces heat radiation into the computer room that would have to be dealt with by CRAH units.

The energy savings that can be achieved with HT-DLC are three-fold: (i) the removal of the fans from the compute nodes saves up to 5% of the total energy consumption of the node itself; (ii) the lower operating temperature of the CPUs due to the superior thermal properties of water over air lead to lower leakage currents in the CPU which saves another 4% of the compute nodes energy consumption; (iii) the ability to use free cooling year-round instead of mechanical chillers typically saves another 15% or more depending on climate conditions (all numbers based on operational experience at LRZ).

As stated in the beginning of this section, the first law of thermodynamics also holds for HPC. Consequently, there is no such thing as waste heat, but only heat energy. Capturing as much of this heat energy directly in hot water is important to re-use it for other purposes such as heating office buildings in winter or as process heat in industrial applications and thereby reducing the carbon footprint of the HPC operations even further.

4. Adsorption Refrigeration

While HT-DLC is a prerequisite for re-using the excess heat of supercomputers, making reasonable use of this heat in summer has remained a challenge. Conventional district heating networks usually operate at temperatures above 70°C and according to the International Energy Agency, more than 90% of industrial process heat is required at temperatures above 60°C in the European Union. Unfortunately, such temperature levels are too high to cool IT equipment safely within specifications.

However, adsorption refrigeration can use warm water starting from 55°C as the driver for a thermally driven refrigeration process to produce chilled water. While it is border line in terms of operational safety, such temperature levels can be delivered by a direct liquid cooled supercomputer. The attractiveness of using adsorption refrigeration in a data centre to produce cold water is that it requires almost no electricity. Instead, the energy required to produce cold can be taken from the excess heat of other systems. And although there is a clear trend towards HT-DLC in high performance computing, many IT systems will remain in a typical HPC data centre that require air cooling or humidity control and therefore need chilled water for their operation.

In adsorption chillers, the production of cold and the usage of driving heat do not occur continuously. However, by combining two adsorption processing chambers a de-facto continuous mode of operation can be achieved. Within each of the chambers, the following alternating processes occur:

1. Evaporation and adsorption, and
2. Desorption and condensation.

A schematic overview of the adsorption process is shown in Fig. 2. The right hand graphic shows the schematic overview of the adsorption process. The left side shows the semi-continuous adsorption process for an adsorption chiller with two chambers.

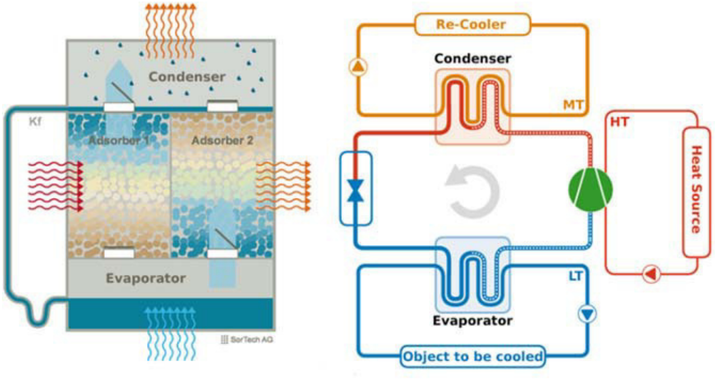

Fig. 2: High level overview of adsorption process

Three closed hydraulic circuits are essential to operate an adsorption chiller:

- HT circuit: High Temperature (driving) circuit
- MT circuit: Medium Temperature (re-cooling) circuit
- LT circuit: Low Temperature (cold water) circuit

During the evaporation and adsorption phase, water evaporates from the reservoir at the Evaporator (right picture bottom) and is adsorbed by the adsorbent (silica gel or zeolite). The evaporation process generates cold which is transported via the LT circuit to the consumer. During the desorption and condensation phase, heat from the HT circuit (the excess heat from the supercomputer) is used to expel the water from the adsorbent. The expelled water is liquified in the condenser and returned to the water reservoir. The heat from the complete process is removed via the MT circuit and typically rejected via cooling towers.

LRZ has been using adsorption refrigeration ever since it has been using HT-DLC: The excess heat of CoolMUC-1 has been used to drive a Sortech ACS08 adsorption chiller that cooled a rack of an air cooled GPU cluster in the same computer room [AH12]. CoolMUC-2's excess heat drives six SorTech eCoo 2.0 adsorption chillers to produce 60kW of cold that are required to cool the storage racks of the SuperMUC Phase 2 flagship system [Wi17]. After the de-comissioning of CoolMUC-1, the remaining adsorption chiller will be connected to the newly installed CoolMUC-3 compute cluster and continue to provide chilled water for other IT systems. The next flagship system SuperMUC-NG will use next generation zeolite adsorption chillers from Fahrenheit to generate all cold required for its remaining air cooled components and therefore will be computer room neutral in terms of heat dissipation into ambient air.

After over 8 years of operational experience, adsorption refrigeration at LRZ has proven to be a useful technology to drive down the energy consumption and cooling costs in a data centre. While typical mechanical chiller setups yield an energy efficiency ration

(EER) of less than 1:5 (i.e., for every 5kW of heat energy removed, 1kW of electrical energy needs to be spent), experience at LRZ shows an EER of more than 1:17 for the adsorption chiller setup of CoolMUC-2. Energy savings with adsorption chilling can therefore easily amount to 15% (of the energy spent on IT and cooling) compared to traditional chillers. However, as outlined in [OWH17], adsorption cooling makes most sense in environments where as much of the excess heat as technically feasible can be recovered with HT-DLC as any remaining air cooled component significantly harms the overall energy efficiency.

5. Energy Aware Scheduling

The technologies described in the previous two sections covered the pillars System Hardware and Building Infrastructure of the 4 Pillar Framework. Both aimed at the cooling aspect of HPC, i.e. at how to dispose of the excess heat of a supercomputer efficiently. Yet, in terms of energy efficiency it is certainly more efficient to spent less electricity in the first place. However, for a data centre like LRZ that provides HPC as a service and that has no control over the application that its clients run on its systems, this is not quite straight forward to do. It needs to be tackled in a generic way that works on all applications and without any contribution from the users.

As the biggest contributor to the energy consumption of a supercomputer are its main processors, one angle to approach this is to reduce their consumption. Modern processors that are used in HPC provide different technologies - many of them derived from mobile processors - to save energy. One of them is called dynamic voltage and frequency scaling (DVFS) and allows for lowering the clock speed and the supply voltage of the processor. As the power dissipation of a processor is mainly determined by the supply voltage, the goal is to lower it as much as possible in order to drive down its power consumption. However, the required supply voltage for stable operation of the processor in turn is dependent on its clock frequency - the higher the frequency, the higher the voltage needs to be. So in order to reduce the power consumption of a processor, the voltage and the frequency have to be lowered simultaneously. Yet, this has implications on the processor's computational performance that is largely determined by its clock frequency. Since the main purpose of a supercomputer is to perform computations and saving energy is only a secondary target, the goal is to lower the processor frequency only if it does not hurt application performance.

Many scientific applications that run on HPC systems are not compute but memory bound. That is, their performance is determined by the performance of the memory subsystem and not by the actual computational performance of the main processors. During the execution of such applications the processor spends many clock cycles waiting for data from the slow memory subsystem without doing any productive work. This fact can be leveraged for saving energy by lowering the clock frequency (and supply voltage) of the processor which will reduce its computational performance but not the memory performance. Consequently, this intervention will have no impact on the actual execution time of a memory bound application (time-to-solution) but can reduce its energy consumption (energy-to-solution) significantly. Whether an application is memory or compute bound can be determined during its execution by leveraging the processor's performance counters that allow for measuring a multitude of performance characteristics of an application.

With the deployment of SuperMUC, LRZ also introduced automatic DVFS for memory bound applications under the umbrella of Energy Aware Scheduling (EAS) in SuperMUC's work load manager IBM Load Leveller [Au14]. The EAS features have been developed jointly by IBM and LRZ and allow for predicting an application's runtime and energy consumption at different processor frequencies. The predictions are based on performance counter measurements of previous executions of the same application and a energy model that is specific to the microprocessor architecture of SuperMUC. Under EAS, memory bound applications run at a default frequency of 2.3 GHz. If the application is not memory bound (according to the performance counter measurements of the previous execution) and a runtime improvement of more than 5% at a frequency of 2.5 GHz is predicted by EAS, the application may run at the higher frequency. Strictly compute bound applications may even run at 2.7 GHz if EAS predicts a runtime improvement of more than 12%. This approach efficiently allows for saving energy on applications that would not profit from higher processor frequencies and at the same time ensures that applications that are compute bound are executed at the highest frequency possible.

The predictions of EAS have proven to be quite accurate and have allowed for energy savings in the order of 5% since the deployment of SuperMUC. For the user EAS is transparent and they do not have to modify their applications in any way to improve their energy efficiency. With SuperMUC-NG, LRZ will deploy similar techniques based on the SLURM workload scheduler that will allow for more fine-grained control over frequency adjustments and different program phases with different performance properties in a single application.

6. Summary

While High Performance Computing remains a key technology for multiple applications in environmental computing, it has its own challenges to cope with in terms of environmental impact and sustainability. As the power consumption of HPC systems continues to grow and will exceed 30 MW for upcoming Exascale systems, energy efficiency is one of the main challenges that need to be addressed. Multiple approaches and technologies exist to boost the energy efficiency in day-to-day HPC operations, three of which have been presented in more detail in this paper: direct-liquid hot-water cooling, adsorption refrigeration, and energy aware scheduling. From the operational experience of LRZ in almost a decade, all of them have proven to make valuable contributions towards energy efficient HPC operations and could easily be deployed in other (potentially smaller) data centres as well. While the previous implementation of EAS on SuperMUC was based on IBM's proprietary Load Leveler batch scheduler, the next generation of EAS to be deployed with SuperMUC-NG will be open-source and made available to others.

Further efficiency potential may be tapped by tighter integration of all four pillars of the framework for energy efficient HPC data centers, particularly by connecting the system software to the building infrastructure.

References

[AH12] Auweter, Axel; Huber, Herbert: Direct Warm Water Cooled Linux Cluster Munich: Cool- MUC. inSiDE, 10:26, 2012.

[AS11a] ASHRAE TC 9.9: Thermal guidelines for data processing environments-expanded data center classes and usage guidance. American Society of Heating, Refrigerating and Air Conditioning Engineers (ASHRAE), White Paper, 2011.

[AS11b] ASHRAE TC 9.9: Thermal Guidelines for Liquid Cooled Data Processing Environments. American Society of Heating, Refrigerating and Air Conditioning Engineers (ASHRAE), White Paper, 2011.

[Au14] Auweter, Axel; Bode, Arndt; Brehm, Matthias; Brochard, Luigi; Hammer, Nicolay; Huber, Herbert; Panda, Raj; Thomas, Francois; Wilde, Torsten: A Case Study of Energy Aware Scheduling on SuperMUC. In: Proceedings of the 29th International Conference on Supercomputing - Volume 8488. ISC 2014, Springer-Verlag New York, Inc., New York, NY, USA, pp. 394–409, 2014.

[Bu09] Bunse, Christian; Höpfner, Hagen; Roychoudhury, Suman; Mansour, Essam: Choosing the "Best" Sorting Algorithm for Optimal Energy Consumption. In: ICSOFT. 2009.

[Me14] Meuer, Hans Werner; Strohmaier, Erich; Dongarra, Jack; Simon, Horst D.: The TOP500: History, Trends, and Future Directions in High Performance Computing. Chapman & Hall/CRC, 1st edition, 2014.

[OWH17] Ott, M.; Wilde, T.; Huber, H.: ROI and TCO analysis of the first production level installation of adsorption chillers in a data center. In: 2017 16th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm). pp. 981–986, May 2017.

[WAS13] Wilde, Torsten; Auweter, Axel; Shoukourian, Hayk: The 4 Pillar Framework for energy efficient HPC data centers. Computer Science - Research and Development, pp. 1–11, 2013.

[Wi17] Wilde, Torsten; Ott, Michael; Auweter, Axel; Meijer, Ingmar; Ruch, Patrick; Hilger, Markus; Kühnert, Steffen; Huber, Herbert: CooLMUC-2: A Supercomputing Cluster with Heat Recovery for Adsorption Cooling. SEMI-THERM conference proceedings, 2017.