Machine Learning Based Control of small-scale Autonomous Data Centers

Rickard Brännvall

Dept. of Computer Science, Space and Electrical Engineering
Luleå University of Technology
Luleå, Sweden

Supervisors:

Fredrik Sandin and Jonas Gustafsson

To Grethel.

Abstract

The low-latency requirements of 5G are expected to increase the demand for distributed data storage and computing capabilities in the form of small-scale data centers (DC) located at the edge, near the interface between mobile and wired networks. These edge DC will likely be of modular and standardized designs, although configurations, local resource constraints, environments and load profiles will vary and thereby increase the DC infrastructure diversity. Autonomy and energy efficiency are key objectives for the design, configuration and control of such data centers. Edge DCs are (by definition) decentralized and should continue operating without human intervention in the presence of disturbances, such as intermittent power failures, failing components and overheating. Automatic control is also required for efficient use of renewable energy, batteries and the available communication, computing and data storage capacity.

These objectives demand data-driven models of the internal thermal and electric processes of an autonomous edge DC, since the resources required to manually define and optimize the models for each DC would be prohibitive. One aim of this thesis is to evaluate empirically *how machine learning methods can be employed* for internal control of small modular DCs. Experiments with small server clusters are presented, which were performed in order to investigate *what operational parameters should be considered* in the design of such advanced control strategies for *energy efficiency and autonomy*. Empirical work on grey box models is carried out to explore *how transfer learning can facilitate the deployment* of machine learning models for modular DC in varying configurations and environmental contexts without training from scratch.

The first study calibrates a data driven thermal model for a small cluster of servers to sensor data and subsequently uses it for constructing a model predictive controller for the server cooling fan. The experimental investigations of cooling fan control continues in the next study, which explores operational sweet-spots and energy efficient holistic control strategies. The ML-based controller from the first study is then re-purposed to maintain environmental conditions in an exhaust chamber favourable for drying apples, as part of a practical study of how excess heat produced by computation can be used in the food processing industry. A fourth study describes the RISE EDGE lab – a test bed for small data centers – built with the intention to explore and evaluate related technologies for micro-grids with renewable energy and batteries, 5G connectivity and coolant storage. Finally, the last study presented builds on the first paper and develops a thermal node recurrent neural network model constrained by a hierarchical modular prior. The model is trained on one small server cluster and is subsequently deployed to another server environment requiring only minimal fine-tuning with a small set of observation data.

Contents

Acknowledgments

The research work for this thesis has been carried out at the RISE Research Institutes of Sweden ICE Data Center research facilities in Luleå, Sweden, with academic supervision from the Embedded Intelligent Systems Lab (EISLAB), Department of Computer Science, Electrical and Space Engineering, at Luleå University of Technology.

Therefore I express my gratitude to Professor Jerker Delsing, the head of Electronic Systems, to Professor Marcus Liwicki, the head of Machine Learning, and to Professor Jonas Ekman, the head of department, for providing me with the opportunity of continuing my postgraduate studies and conducting this research. Additionally, I am thankful to the entire staff and personnel at RISE and LTU, for the pleasant research- and social environment at both institutes.

This thesis could not have been written without the guidance and contribution of ideas from my supervisor Associate Professor Fredrik Sandin (EISLAB), and co-supervisor Dr. Jonas Gustafsson (RISE) to both of whom I owe special gratitude for their kind support, challenging discussions and scientific guidance.

I am grateful for the possibility to conduct my postgraduate studies as an industrial doctorate student at RISE AB in Luleå. For that, and for his guidance and kind advice, I express my gratitude here to Tor-Björn Minde, head of the lab. I also thank Professor Jon Summers, the scientific leader of RISE ICE Data Center, for his inspirational leadership and comments.

Special respects I owe to Professor Gareth W. Peters at Heriot-Watt University for supervising me during my studies at UCL, and to Professor Tomoko Matsui for welcoming me during my stays at the Institute of Statistical Mathematics in Tokyo.

I acknowledge all my friends and colleagues for their valuable comments, advice and encouragement. The list of names include Jeffrey, Erik, Sebastian, Daniel, Magnus, Filip, Louise, Jacob, Gustav, György, Anki, Johan, Mattias, Johannes, Jeanette, Mikko, Emilia, Yohannes, Rickard, Joar, Jay, Lee, Masato, Kumiko, Alex, Amir, and many more.

I thank my family for all the love, encouragement, and understanding given, during the entire period of my education. Therefore, to my father Allan, my mother Agneta, my grandmother Grethel, and my sister Mari and brother Henrik, I am eternally grateful.

Finally I thank Mitsuo for always being there to support me, and to my son Loki for being the light in my day and reminding me of what is important.

Luleå, July 2020
Rickard Brännvall

Part I

Chapter 1

Introduction

"A wise man can learn more from a foolish question than a fool can learn from a wise answer."

Bruce Lee

1.1 Data Centers in Society

From the latter half of last century, data centers has come to fill an important function in society by processing, storing, and relaying ever increasing amounts of data. They support a multitude of digital services for the general public, private business, as well as governmental, educational, welfare, and financial institutions. One can really say that data centers form "the backbone of the global digital infrastructure" [18]. In addition to the hardware and software making up the computing infrastructure, data centers also consist of facilities and equipment for maintaining the components through their life-cycle, providing them with power as well as controlling the environment such that it remains favourable for long term operation.

Brown, Richard et al. [12] categorizes the data center into primary computing infrastructure and two supporting facility systems for environmental control (cooling) and power distribution. The processing and transmission of data consumes electrical power and transforms it (almost completely) into low-grade heat - heat which must be removed from the computing equipment. Due to the relatively low temperature this heat is often exhausted into the environment.

The required power and cooling fluctuate with time and are often difficult to predict, as it depends on the demand of the data centers computing resources, which can vary by the time of day, the day of the week, and season [48]. In some applications it can also depend on unexpected end-user behaviour, or unpredictable external events. The cost and capacity to provision cooling within target envelopes for energetic performance is also influenced by the weather, which is troubled by forecast uncertainties familiar to everyone.

Electricity prices also vary as market supply and demand conditions change on daily,

weekly and seasonal cycles – fluctuations that are partly random and unpredictable. Data center operators cannot assume 100% power grid availability and will almost uniformly include battery (UPS) and diesel generator capacity to their facilities. Further complexity is added to the data center modelling and control landscape by the recently popularised micro-grid connected data centers, which have renewable power sources as well as substantial battery capacity.

The industry has also experienced momentous and consistent growth, with capacity, measured in computing floor area, up 10 percent in terms of Compound Annual Growth Rate (CAGR) over the last decade [18]. At the same time technological advances in design of the electronics, its packaging and thermal management increase computing density and capacity (see e.g. Glinkowski, Mietek [27]). Analysis of historical trends shows a consistent pattern with the energy efficiency of computing units doubling every 18 months [40].

Data centers are highly technologically complex entities and provide critical services to modern society. They are thus required to be highly dependable and must demonstrate actionable plans for business-continuity and recovery in times of crisis or disaster [28], with transparent availability agreements about yearly downtime, reliability and contingency.

1.1.1 Data Centers at the Center

What we call a data center is in itself not precisely defined, but instead covers everything from small computer rooms of a few square meters up to hyperscalers a thousand times larger - the appropriate scale being determined by the particular business needs as well as constraints such as availability of grid power and physical space (see for example Brown, Richard et al. [12], Glinkowski, Mietek [27]). A slow shift to ever larger data centers have however been observed, fueled by the growth of Internet traffic, an increased demand for compute and the rapid increase of end-user subscriptions to digital services such as social media, streaming services and Internet of Things (IoT) applications. The data center services are now considered as utilities rather than assets to manage [16], and benefit from economies of scale, for example in amortising cost of investment in security measures over a larger server population – a larger population that by itself increases flexibility for managing aggregate workload.

This development has made it possible to move computation and storage from local data centers to highly efficient central hyperscalers providing the majority of cloud services used today - it is estimated [69] that by 2020 about 70 percent of US data center overall electricity usage will be consumed by large- and hyper-scale facilities (with floor capacity over $500m^2$). These large data centers are often located where there are good conditions for operating data centers, for example stable power supply, good network infrastructure, affordable land and climates that favour lower cooling costs and provide access to renewable energy sources, which can strengthen the operators environmental profile. However, these conditions are often not present in major cities, where people tend to live and work.

1.1.2 Data Centers at the Edge

Latency, or the response time induced by the physical distance that the digital information has to travel does matter for many applications, where for example algorithmic trading ventures pay big dollars renting server space right next-door to the securities exchanges. For technologies that rely on earlier mobile communications systems, using centralized data centers was considered acceptable since the 4G (and earlier) latency dominated that of the wired communication (˜20-30ms). This will no longer be the case, since with 5G this drops to 1-10 ms, which means that the larger part of the system latency will now be in the communication between the base station and the data center.

Latency-sensitive applications and devices will demand that the compute resources move closer to the end-user – or even collocate with the access point such that, for example an IoT control loop or real time analytics of streaming data can benefit from the latency of down to 2ms in the 5G radio link over shorter distances. The Telecommunications Industry Association [71] predicts that many existing applications compute resources traditionally delivered by centralized data centers will instead be required at the edge of the network. Cloud technologies and streaming data services will require even faster caches at the very edge of the 5G networks.

From being a niche technology, edge data centers are thus expected to be placed everywhere where human or IoT presence demands it, which is likely to drive development towards mass production and mass distribution of more or less standardized units, e.g. containerized solutions. Although each edge data centers would have a lower computing capacity, they would likely have comparable heat load density as larger scale data centers (considering this invariant for energy efficient facilities as in Patterson, Michael K. D. G. Costello, et al. [61]). This points to two major challenges for the edge data center concept: 1) How to provide sufficient power, and 2) How to design efficient heat rejection (cooling) systems. Compared to centralized data centers that can be optimally placed for access to power grid and cooling resource, many edge data centers will likely be located in cities where competition for electric power is high (and costly) and heat rejection is complicated as they are co-located inside existing buildings.

1.2 Sustainability

It has been widely reported in science and media that data centers are consuming large and growing amounts of electrical power. The impact of the industry on global emission of green house gasses is therefore significant [46, 10]. Reports from the U.S.' Environmental Protection Agency show a tremendous growth in the electric energy use of the country´s data centers over the last two decades, estimating the total 2014 annual consumption at 70 billion kWh. Globally, the industry represents 2% of the world's total electricity demand and substantial growth is expected over the next decade for the power used, not only by data centers themselves, but also by the networks than connect them with the end users as larger amounts of data is communicated annually [34]

Up to 40% of the power consumed by a typical data center is accounted for by cooling

systems, which is comparable to the around 40% of the total energy consumption that the servers themselves use for computation [14, 64]. Large variations in the overhead incurred by heat removal can however be observed, depending on the type of cooling technology, the scale of the deployment, the climate conditions at the geographic location, and the production year of the IT equipment used (as older-generation servers and servers near the end of life are less efficient).

Thermal management of data centers is therefore an important area of research and an obvious target for optimization. The objectives are for example to find more efficient cooling strategies that reduce the demand for chiller work and facility electricity [39, 29, 38, 17, 35] and to find productive reuses for the excess heat generated by the data center through its normal operations, for example recent work on absorption refrigeration, organic Rankine cycles, desalination, clean water production, piezoelectric, thermoelectric, heating buildings, swimming pools, biomass- or power plant co-location [21, 78, 56, 23], where high potential is found for heat pumps to upgrade the excess heat as a source for district heating [19, 15, 75, 3, 2, 63].

1.3 Data Center Automation

An important objective for this licentiate thesis is the investigation of machine learning applications that support the development of autonomous data centers[1]. Much of the work presented here has been carried out and funded through the ITEA3/Vinnova project: Autonomous data centers for long term deployment, or in short AutoDC.

The objective statement of the project application states:

> The aim of AutoDC is to provide an innovative design framework for autonomous data centers. An autonomous data center should be able to, without any human intervention, from a best effort perspective continue its operation independent of contextual interference, such as intermittent power failure, failing components, overheating, etc.

The project AutoDC description continues:

> With an expected continued growth in the data center market, the cost of operating and maintaining the data center footprint will increase. The administration and maintenance cost in large and mega-scale data centers is one third of the OPEX and challenges in this market segment are due to the vast amount of equipment and personnel.
>
> Part of the growth, except in the usage scenario of mega-scale data centers, is predicted to be in the market segment of edge computing, where infrastructure is close to application usage such as urban areas, intra-urban transportation routes and areas of dense congregation of devices.

[1] this work focus on small scale deployments, like edge data centers

> These future requirements will need distributed data centers and equipment leading to an increase in the cost of operation and maintenance.

The report develops the relevance and motivation of the project:

> Beyond the obvious need for adding more redundancy in the form of back-up components, this level of autonomy will require a detailed knowledge of the ecosystem's condition and context.
>
> It also requires beyond state-of-the- art ability to autonomously detect failures and maintenance activities as well as controlling the environment employing AI trained software, without human intervention.

The report further identifies key technologies:

> Data center complexity is increasing and there is a need for holistic and integrated approaches. Developments in modular data center components lend themselves to applications, and will facilitate this project by enabling experimentation to take place in laboratory environments.
>
> However, the concepts developed will be clearly scalable to larger applications. Sensors measuring the physical properties of components and their environment in addition to equipment that monitors power and cooling must be well integrated and improved from today's state-of-the-art offerings. Innovative and sophisticated control peripherals will also be required to replace human intervention.
>
> A powerful data analytics engine is required to achieve data collection from the various monitoring systems, which is then consolidated with external data sources and periodically stored as appropriate records to allow for both real-time and off-line ecosystem modelling and machine learning data analysis.
>
> The analytics results will ensure proper actions are applied to the control systems for optimised power, cooling, network and server operation, which is essential to maintain the data center "health" within desired parameters to reach identified target KPI values.

1.4 Problem Formulation

Given the problems outlined in the sections above, and particularly with respect to the AutoDC project specification, the following research questions have been formulated:

Q1: *How can machine learning methods be employed for control over internal thermal and electric processes of small modular data centers?*

Q2: *What operational parameters should be considered (especially with respect to energy efficiency and autonomy) in the design of advanced machine learning based control strategies?*

Q3: *How can transfer learning facilitate the deployment of machine learning models for modular data centers (in varying configurations and environmental contexts) without training from scratch?*

1.4.1 Delimitations

The licentiate research focus on the thermal management of small autonomous edge data centers, leaving the equally important power distribution and load balancing problems for the remaining post-graduate research.

Control objectives (explicit and implicit) are mainly formulated with respect to energy efficiency and autonomy, and do not directly consider impact on OPEX/CAPEX, equipment life-cycle or SLA compliance.

The focus is on practical applications of machine learning for control, rather than investigating conventional automatic control theory. Similarly, only limited comparison is made with basic control methods, although we recognize that comparative analysis is important as these methods (like PIDs) can be cost effective in comparison to model predictive control for some applications.

Fault detection and predictive maintenance are important areas to consider for the design of autonomous data centers, but these topics are out of scope for this licentiate work. We hope to address these topics in future work.

1.5 Thesis Outline

This thesis is made up of two parts. In part I, Chapter 1 presents the background of this work, and Chapter 2 introduces the methods that are the building blocks for the scientific papers presented in the second part. The contributions of each paper is summarized in Chapter 3. In closing part I, Chapter 4 presents the conclusion of the Thesis and suggest future work within the area.

Part II contains the scientific papers that contribute to this work. Paper A, published in the proceedings of the 2019 IEEE 17th International Conference on Industrial Informatics (INDIN 2019), demonstrates how a data driven thermal model for a small cluster of servers can be calibrated to sensor data and used for constructing a model predictive controller (MPC) for the server cooling fan. The experimental investigations of cooling fan control is continued in Paper B which explores operational sweet-spots and energy efficient holistic control strategies (presented at the Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems, ITherm 2020). Paper C (accepted for the 33rd International Conference on Efficiency, Cost, Optimization, Simulation and Environmental Impact of Energy Systems, ECOS 2020) re-purposes the MPC from the previous paper on server fan control to produce environmental conditions in an exhaust chamber favourable for drying apples, in other words a practical study how excess heat produced by computations can be used in the food processing industry. Experimental investigations into autonomous data centers commence with Paper D (presented at the E2DC 2020 workshop at The Eleventh ACM International Conference on

Future Energy Systems) that describes the RISE EDGE lab – a test bed for small data centers – built with the intention to explore and evaluate related technologies for micro-grids with renewable energy and batteries, 5G connectivity and coolant storage. Paper E presents work to be submitted to a journal that builds on the work in paper A and develops a thermal node recurrent neural network model to investigate an application of transfer learning.

Chapter 2

Methods

"Learn to be indifferent to what makes no difference."

Marcus Aurelius

2.1 Energy Efficiency and Performance Metrics

A number of performance and sustainability indices has been defined to evaluate and compare data centers. Of these, Power Usage Effectiveness (PUE) is most widely recognized and used in practise. It was proposed in 2007 by a consortium of data center operators, who defined it as the ratio between the average facility power consumption and the average power consumption of the computing infrastructure, that is

$$\text{PUE} = \frac{\text{[Total facility energy]}}{\text{[Computing equipment energy]}}, \tag{2.1}$$

which uses quantities aggregated over one year and where the denominator include computing nodes, network and storage. The PUE index has been examined in great detail in many studies (see for example the Green Grid white paper by Avelar et al. [7]), but in summary one can say that PUE is reduced when the overhead of supporting infrastructure is reduced by making cooling and power distribution more efficient. A PUE of 1.4 or above are not uncommon, while it has been demonstrated that best practise can push PUE closer to the ideal score of 1.

At the time of writing this report, the world record was reported by the BTDC research data center deployment in Boden [26] with a PUE of less than 1.04. Another example from the region is Facebook's hyperscale data center in Luleå, which reports a yearly average PUE of 1.05. A lower PUE not only strengthens the sustainability profile of a data center operator, but can also translate into dollars saved. In the Facebook example with annual cooling costs of 9.7 million SEK, each hundredth shaved off the PUE saves about 2 million SEK [18].

Energy Reuse Effectiveness (ERE) is a metric defined as a correction to PUE for the amount of energy captured and re-used for productive purposes outside the data center [60].

$$\text{ERE} = \text{PUE} - \frac{\text{[Energy reuse]}}{\text{[Computing equipment energy]}} \tag{2.2}$$

Better performance as measured by ERE can thus come from either the same causes that improve PUE, or by increasing the quality or quantity of the heat harvested.

The above performance indices capture two major avenues for improving the sustainability of data centers, which have also previously been investigated most in research. Other important objectives relate to data center water usage, carbon emissions and equipment life cycle. Performance indices associated to these areas have been proposed, such as Carbon Usage Effectiveness (CUE) [9] and Water Usage Effectiveness (WUE) [62] that each attempt to quantify the efficiency of the use of natural resources important for data center operation.

A complete sustainability analysis should take into account the whole chain of processes, from material acquisition, over manufacturing, distribution, operation and maintenance of equipment in the data center, to the final disposal stage. Estimates of life cycle contributions for the ICT and media sectors, however, indicate that the overwhelming share, at least pertaining to aspects related to greenhouse gas emissions and electricity use, originates in the operation phase [47].

Efficient cooling and thermal management of data centers is a topic that has received considerable attention in recent years, see for example the review article by Ebrahimi et al. [22] and the references therein. Although liquid cooling shows advantages compared to air cooling, it is only the latter mode that is addressed in this work as liquid cooling is still largely limited to exascale and high-performance computing (HPC) environments.

2.2 Monitoring and Control

As discussed in the Introduction, the efficient operation of a data center requires planning and coordinating the use of resources across a variety technologies and infrastructures: Software, hardware, power distribution and facility thermal management [27, 18]. A recent review of control strategies in data centers can be found in the PhD thesis of Lucchese [42], which also compiles several papers on different control strategies that target energy efficiency and heat recovery.

There are complex interactions between the different systems. For example when considering the options for workload allocation, the optimal physical placement of computing on servers inside the data center depends on the provisioning of cooling [51, 11], flexible workload scheduling that best matches a fluctuating electricity price [58], or load-balancing across data centers in different locations to reflect geographical demands and availability constraints [41, 72], shutting down servers or keeping them idle to permit quick deployment of new unexpected workloads, etc.

The specific control objectives should support high-level management goals, such as 1) operational cost savings, 2) delivering against service level agreements (SLAs) and

3) reducing the environmental footprint. Ideally these goals are aligned, but one can easily find examples when they are opposed, e.g. cost constraints limiting investments in sustainability; servers are kept running idle to have spare capacity to quickly meet volatile demand, but this practise can compromise energy efficiency.

Formulating a comprehensive optimal control policy is by itself a formidable task, as the data centers are considered among the most complex cyber-physical systems employed in our society. Furthermore, there is no general agreement on what is the appropriate set of performance metrics to include in a holistic control objective, or how to give weight among them, which can result in a mixed basked of conflicting objectives [12]. Practical Data Center Infrastructure Management (DCIM) systems instead compartmentalizes the control systems for the different resources, each focusing on managing only a subset with a local objective function [27]. It can then be expected that holistic control strategies that optimize towards global objectives should perform better than control systems based on local information only [59] provided that the challenges of system complexity can be alleviated, for example by better coordination among subsystems or more powerful modeling abstractions. There are also opportunities to simplify data centers systems by new technological developments, such as efficient abstractions for software orchestration and load balancing, or investing in liquid cooling systems that exhibit simpler thermal dynamics.

Overheating the equipment affect the computing capacity as servers start throttling, the temperature-dependent leakage increases (which add to power consumption), and risk damaging the equipment and decrease the mean time to failure. It is important also to control humidity in conjunction with temperature, as too dry conditions increase the risk of static discharge, and over-provisioning cooling under too humid conditions can promote moisture condensation. All these harmful consequences of control failures increase operating costs through breached up-time agreements, maintenance work and equipment replacement.

To address safety and energetic concerns, organisations like the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) publish guidelines for both air and liquid cooled data center equipment [4, 5, 6]. For example, the recommended (level A1) dry-bulb temperature envelope stretches between 15 to 32 °C, for relative humidity between 20 and 80 %, and dew point 17 °C. These are quite wide upper and lower limits, which may allow operation far from conditions that are optimal from a cost and environmental perspective.

At the individual server level, important strategies for reducing energy usage while maintaining quality of service can be identified by the following three categories:

1. Optimal scheduling of the use of digital components to reduce power consumption, for example by turning on equipment only when it is required [79, 33]

2. Dynamic regulation of the central processing unit (CPU) clock-frequency so that it is at minimum when operating low digital workload (low utilization) and increasing the frequency when the digital demand is increased [50, 49]

3. Optimizing the power consumption of the server fans while staying within the thermal constraints for the CPU, memory and other components [77, 30].

An efficient server fan controller aims to hold component temperature at a fixed set point, while trying to save cooling power, as for example studied for closed loop settings in Wang et al. [77]. Controls based on physical models of the system face considerable challenges that spring from non-trivial long range spatial and temporal dependencies. Computational Fluid Dynamics can model both static thermal properties and dynamic flows in detail, such as demonstrated in data center settings in Iyengar et al. [32], however only at computational costs that are prohibiting for use in online closed loop controls.

The Proportional-Integral-Derivative (PID) controller is popular in real industry applications because of its simplicity and relative robustness (see for example review in Åström and Hägglund [80]). It is model-free and transparent, with clear physical interpretations for each of its three parameters that can be tuned by straightforward, albeit time-consuming, standard procedures , see e.g. [54].

2.3 Models of the Server, Rack and Data Center

One important means of reducing inefficiencies in individual servers as well as whole data center rooms is to appropriately provision the cooling air such that hot spots, re-circulation and bypass air flows are avoided. Re-circulation is when hot air from the IT equipment mixes with the cold air provided by the cooling system. This is a sub-optimal mode of operation and is due to insufficient flow of cold air being supplied to the cold aisle. Bypass is the phenomenon when part of the cold air flow is not used to cool the equipment, but instead passes directly back to the computer room air conditioning (CRAC). Both these inefficiencies unfortunately occur quite often in data centers and are important targets for improvements, either by devising advanced control strategies for the cooling system, or by design solutions such as containment which is a common practise where a barrier is raised between the hot aisle and cold aisle [53].

To understand bypass, re-circulation, hot spots and other phenomena relating to the temperature and air distribution within the server room, Computational fluid dynamics (CFD) provides a powerful modelling tool with higher resolution compared to other experimental approaches [1]. The traditional methods of Finite Elements and Finite Volumes are however computationally demanding, although recent work on lattice Boltzmann demonstrate performance of real-time simulation for some systems (see e.g. de Boer et al. [20] for a comparison of the different methods).

Proper Orthogonal Decomposition reduced order models [65, 66, 67] and Resistor-Capacitor Networks [70, 8] provide more favourable computation times for modelling data centres compared to CFD however at the trade-off of reduced accuracy.

2.3.1 Compact models

At the individual server level, compact models for the thermal inertia of their internal components have demonstrated good performance, see for example the review by Pardey

et al. [57] for a description of these models and methods for their calibration. These predict exhaust temperature in response to internal heating and time-varying ambient temperatures, such as those developed independently by VanGilder et al. [73] and Erden et al. [24] in the context of using as simple, compact components for heat producing servers and racks for larger CFD based models.

Lucchese et al. [45] propose a control-oriented, non-linear, thermal model of computer servers, and identify model parameters by CFD simulations of an idealized server set-up. Further simulations are then used to (in silico) both validate the model and test a Receding Horizon Control (RHC) strategy aiming to keep IT component temperatures below thermal bounds while minimizing fan energy use. Building on this work, Eriksson et al. [25] first describe means for collecting detailed data from Open Compute Servers used by Facebook [55] and develop a control-oriented model for the thermal dynamics of the CPU and RAM as functions of the computational load and mass flow produced by server fans.

VanGilder et al. [74] propose a compact model for data center applications with various chilled-water cooling systems. It includes a heat exchanger term in series with an "additional mass" term, which captures the thermal inertia of the cooling system. The first term is either a discretized numerical model of a simple counterflow heat exchanger, or by a quasi-steady-state model for applications when accuracy can be traded for model simplicity. This compact model i developed further in Healey et al. [31] that adds further "additional mass" components for room, plenum, walls, floor, ceiling and a water storage tank in order to better represent the complete thermal mass of a data center. The model is used to simulate a chiller failure and compared with a real-world incident in a 300kW size data center.

A similar line of modelling by thermally connected nodes in a network is followed by Lucchese and Johansson [43], but with greater detail in the internal components of the server, such that the thermal mass of each CPU and its associated aluminium fin heat exchanger are included. This allows finer details of the components transient temperatures to be explained. Combined with a simple model of the thermal leakage for the hot chip, the thermal network is then used in a RHC strategy to control the cooling fans under an objective function that penalizes energy consumption (from both sources) while keeping the CPUs within temperature constraints. The thermal network based model of this work is employed in Lucchese and Johansson [44] to examine heat recovery from Open Compute Windmill (V2) servers. There the model based controller is given the objective to maximizes exhaust air temperature and encourage fan control signal smoothness, all while keeping CPU temperatures within envelopes for safe operation.

2.3.2 Physically guided ML models

Karpatne et al. [37] discuss *Physically Guided Neural Networks* (PGNN) - a framework for incorporating guidance from physics into training neural network by two means: 1) letting a (crude) physical model provide features for the neural network, and 2) explicitly including a term in the training cost function that penalizes solutions inconsistent with

known laws of physics. The framework is applied to modelling lake temperature gradients. The framework is extended in a vision paper on *Theory Guided Data Science* (TGDS) by Karpatne and a wider group of collaborators in machine learning and the physical sciences [36], which presents a taxonomy of ways in which data science algorithms can be anchored with scientific knowledge - in summary: 1) restrict model space to physically consistent solutions, 2) guide a model towards physically consistent solutions by e.g. by initialization, priors or regularization, 3) refine output of data science algorithms by scientific knowledge, 4) build hybrid models that mix theory-based components and (pure) data-driven components, and 5) augment theory based models to make better use of observational data. The aim of the endeavour is three-fold: i) build models that generalise better (also from sparse data), ii) improve physical interpretability, and iii) reduce computational cost. Many examples of TGDS are reviewed including from the fields of hydrology, turbulence modelling, bio-medicine, climate science and material discovery. Continuing the work, Muralidhar et al. [52] employees a neural network that predicts intermediate physical quantities in addition to the main target which is the drag force on particles suspended in moving fluids. They demonstrate how the sequential multi-task formulation yields a model that not only performs better on the main task, but is also interpretable and more consistent with the laws of fluid dynamics.

2.4 RISE ICE Experimental Data Center Facility

The subarctic part of Europe has become a favorable location for data center establishments like Facebook, Google, Microsoft, due to the beneficial climate that permits free air-cooling, a stable power grid with a large share of renewables to a low cost and acceptable latency time to the large internet exchanges hubs [76]. Currently the installed data center power in the north of Sweden constitutes 175 MW, and there are several ongoing establishments in the process of strengthening the region as data center industrial area.

RISE ICE data center is a research facility in Luleå, in the very north of Sweden, which conducts a variety of research in the data center field e.g. operation optimization, evaluation of cooling techniques, facility measurement and machine learning applications for data center control. Several ongoing projects focus on heat reuse and how a data center can function as a heating source for greenhouses, fish farms and biomass dryers, but also how to incorporate them in an urban and industrial symbiosis.

The facility houses several experimental test beds for data center research, including those that allow detailed control over the thermal environment. The work presented in this thesis makes use of the following: 1) the "chimney" set-up for experiments on natural draft and free cooling of multiple servers, 2) the 10 meter wind tunnel set-up, and 3) the EDGE lab set-up.

The 10m generic server wind tunnel mimics data center environments for cooling experiments and is described in detail in Sarkinen et al. [68], Paper B of this compilation thesis. It is daisy chaining a liquid cooled heat exchanger, a heater, a humidifier, a droplet collection box and a fan connected by a long 200mm tube to the working section. The wind tunnel fan, humidifier and heater are controlled via PID controllers to maintain a

prescribed pressure drop, temperature and humidity to the server.

The RISE EDGE lab is described in detail in Brännvall et al. [13], Paper D in this compilation thesis. It is a test bed for small data centers - built with the intention to explore and evaluate related technologies for micro-grids with renewable energy and batteries, 5G connectivity and coolant storage.

Chapter 3

Contributions

"This report, by its very length, defends itself against the risk of being read."

Winston Churchill

3.1 Paper A

Title Digital Twin for Tuning of Server Fan Controllers
Authors Rickard Brännvall, Jeffrey Sarkinen, Joar Svartholm, Jonas Gustafsson, Jon Summers
Abstract Cooling of IT equipment consumes a large proportion of a modern data centre's energy budget and is therefore an important target for optimal control. This study analyses a scaled down system of six servers with cooling fans by implementing a minimal data driven time-series model in TensorFlow/Keras, a modern software package popular for deep learning. The model is inspired by the physical laws of heat exchange, but with all parameters obtained by optimisation. It is encoded as a customised Recurrent Neural Network and exposed to the time-series data via n-step Prediction Error Minimisation (PEM). The thus obtained Digital Twin of the physical system is then used directly to construct a Model Predictive Control (MPC) type regulator that executes in real time. The MPC is then compared in simulation with a self-tuning PID controller that adjust its parameters on-line by gradient descent.
Personal contribution General idea developed by me in collaboration with Joar Svartholm and Jonas Gustafsson. Method development, data collection, implementation and evaluation of results by the main author. Manuscript preparation and revision by main author in collaboration with Jonas Gustafsson and Jon Summers.

3.2 Paper B

Title Experimental Analysis of Server Fan Control Strategies for Improved Data Center Air-based Thermal Management

Authors Jeffrey Sarkinen, Rickard Brännvall, Jonas Gustafsson, and Jon Summers
Abstract This paper analyzes the prospects of a holistic air-cooling strategy that enables synchronisation of data center facility fans and server fans to minimize data center energy use. Each server is equipped with a custom circuit board which controls the fans using a proportional, integral and derivative (PID) controller running on the servers operating system to maintain constant operating temperatures, irrespective of environmental conditions or workload. Experiments are carried out in a server wind tunnel which is controlled to mimic data center environmental conditions. The wind tunnel fan, humidifier and heater are controlled via separate PID controllers to maintain a prescribed pressure drop across the server with air entering at a defined temperature and humidity. The experiments demonstrate server operating temperatures which optimally trade off power losses versus server fan power, while examining the effect on the temperature difference, ΔT. Furthermore the results are theoretically applied to a direct fresh air cooled data center to obtain holistic sweet spots for the servers, revealing that the minimum energy use is already attained by factory control. Power consumption and Power Usage Effectiveness (PUE) are also compared, confirming that decreasing the PUE can increase the overall data center power consumption. Lastly the effect of decreased server inlet temperatures is examined showing that lower inlet temperatures can reduce both energy consumption and PUE.
Personal contribution The main work for data collection, implementation and analysis was done by the main author Jeffrey Sarkinen. I contributed to developing the general idea, method discussions, analysis, manuscript preparation and background research together with Jeffrey Sarkinen, Jonas Gustafsson and Jon Summers.

3.3 Paper C

Title Data Center Excess Heat Recovery:A Case Study of Apple Drying
Authors Rickard Brännvall, Louise Mattsson, Erik Lundmark and Mattias Vesterlund
Abstract Finding synergies between heat producing and heat consuming actors in an economy provides opportunity for more efficient energy utilization and reduction of overall power consumption. We propose to use low-grade heat recovered from data centers directly in food processing industries, for example for the drying of fruit and berries. This study analyses how the heat output of industrial IT-load on servers can dry apples in a small-scale experimental set up. To keep the temperatures of the server exhaust airflow near a desired set-point we use a model predictive controller (MPC) re-purposed to the drying experiment set-up from a previous work that used machine learning models for cluster thermal management. Thus, conditions with for example 37 °C for 8 hours drying can be obtained with results very similar to conventional drying of apples. The proposed solution increases the value output of the electricity used in a data center by capturing and using the excess heat that would otherwise be exhausted. The results from our experiments show that drying foods with excess heat from data center is possible with potential of strengthening the food processing industry and contribute to food self-sufficiency in northern Sweden.

Personal contribution General idea and method development by the main author in collaboration with Louise Mattsson and Mattias Vesterlund. Data collection, implementation and analysis by the main author together together with Erik Lundmark and Louise Mattsson. Manuscript preparation and responding to reviewer feedback by main author in collaboration with Louise Mattsson, Erik Lundmark and Mattias Vesterlund.

3.4 Paper D

Title EDGE: Microgrid Data Center with Mixed Energy Storage
Authors Rickard Brännvall, Mikko Siltala, Jeffrey Sarkinen, Jonas Gustafsson, Mattias Vesterlund, Jon Summers

Low latency requirements are expected to increase with 5G telecommunications driving data and compute to EDGE data centers located in cities near to end users. This article presents a testbed for such data centers that has been built at RISE ICE Datacenter in northern Sweden in order to perform full stack experiments on load balancing, cooling, microgrid interactions and the use of renewable energy sources. This system is described with details on both hardware components and software implementations used for data collection and control. A use case for off-grid operation is presented to demonstrate how the test lab can be used for experiments on edge data center design, control and autonomous operation.
Personal contribution General idea and method development by the main author in collaboration with Mikko Siltala, Jonas Gustafsson and Mattias Vesterlund. Data collection, implementation and analysis by the author together together with Mikko Siltala, Jonas Gustafsson and Jeffrey Sarkinen. Manuscript preparation by main author in collaboration with Mikko Siltala, Jeffrey Sarkinen, Jonas Gustafsson, Mattias Vesterlund, Jon Summers.

The main work for building the test bed, installing the sensors, controls and data collection software components was done by the other authors together with other staff at RISE ICE Data Center.

3.5 Paper E

Title Machine Learning based Thermal Control of Edge Server Cluster: Towards Transfer Learning
Authors Rickard Brännvall, Jeffrey Sarkinen, Jon Summers, Jonas Gustafsson, Fredrik Sandin

We investigate a machine-learning based approach to efficient thermal control of a small cluster of six servers that are heating a co-located space. A time-series model is calibrated by prediction error minimization (PEM) and is subsequently used to control the cooling fans so that a chamber receiving the exhaust air is held at a selected temperature set-point, while the CPUs are kept within safe operation thermal limits. The model features a hierarchical regularising prior that connect related modules in the system to

reduce the number of degrees of freedom of the model, and it is found to outperform an ordinary recurrent neural network (RNN) both for CPU and exhaust chamber temperature predictions. We investigate if this modular design facilitates transfer learning by fine tuning the model calibrated on the server cluster on another experimental set-up, where a server is located in a wind tunnel. The transferred model is finetuned with only a few epochs on a small data-set and is found to outperform a model trained from scratch over hundreds of epochs. The proposed model is implemented as a custom RNN in a open-source software package for machine learning.

Personal contribution General idea developed by the main author in collaboration with Fredrik Sandin and Jonas Gustafsson. Method development, data collection, implementation and evaluation of results by the main author. Drafting of manuscript and further improvements in collaboration with Fredrik Sandin.

Chapter 4

Conclusions and future work

"The whole future lies in uncertainty:
live immediately."

Seneca

4.1 Conclusions

This work investigates *how machine learning based methods can be exploited for control of small server clusters in edge data centers*. A data-driven grey-box model is trained to predict temperature measurement time-series from components of the system, and used implement Model Predictive Control (MPC) for thermal management.

In Paper E, a Thermal Node Recurrent Neural Network with a hierarchical prior was employed to *investigate the feasibility of transfer learning* by pre-training a model on one server cluster and fine-tuning it to another set-up with only a few epochs of parameter updates on a small data set.

The grey-box models are based on thermodynamic equations, with machine learning approximations in place of physical relations that are hard to capture in simple models. An open source machine learning library [1] is used for the analysis and control, demonstrating the ease at which complicated objective functions can be expressed using modern software that supports automatic differentiation and arbitrary loss norms. The models are formulated as custom Recurrent Neural Networks (RNNs), which allows the use of the software library's machine learning functionality, thereby avoiding much for the boilerplate code required to adapt it to the problem.

To *further the investigation of energy-efficient data centers*, the MPC is deployed in an experimental study in Paper C of how to reuse the waste heat in a food industrial process. A small-scale experiment shows that results similar to conventional hot-air drying can be achieved using the waste heat, without compromising food quality criteria for reliable shelf life.

[1] TensorFlow/Keras for Python

In a related line of investigations, the delicate interplay between CPU operating temperature and the overall server power demand was studied in Paper B with the aid of a server wind tunnel, demonstrating that a server CPU temperature sweet spot can be found that optimally trades off microelectronics rate of energy losses versus fan power consumption. A theoretical analysis built on the experimental analysis is applied to a real data center with direct fresh air coolers in Paper B to examine the effect of server operating temperatures and data center energy use, in which holistic sweet spots were obtained. The optimal range of CPU operating temperatures was higher than the server sweet spot, and somewhat surprisingly placing it close to the factory operating conditions of the server.

While the CPU temperature sweet spots are relevant for energy efficient data centres in general, the work presented in this thesis also *investigates what operational parameters should be considered for small autonomous EDGE data centers* with solar panels, batteries and coolant thermal energy storage (TES) tanks. Paper D explores different strategies for off-grid operation by scheduling the use of battery and the TES tanks. Variables important for the dimensioning of batteries and TES tank are are also analysed using experimental data, with the conclusion that including one or both can be appropriate for designs required to continue operation during off-grid/black-out scenarios, albeit with slight adverse effect on energy performance.

4.2 Future Work

For future work, I propose to extend the thermal node RNN model to larger server setups, such that more complex interactions between components can be investigated, and test whether a similar modular grey-box approach using hierarchical priors can alleviate the problem of rapid growth of parameters also for models of even larger systems.

A first step investigating transfer learning for thermal node networks was taken in Paper E, but we believe that the formulation with hierarchical prior could also be explored for larger modular data center designs, and compared with other methods for transfer learning as for example employed recently for neural networks.

The grey-box models developed in this work still require manual work to implement the thermodynamic equations that they are based on as RNNs. An avenue for future investigation would be to use the *Theory Guided Data Science* framework [36], briefly discussed in Section 2.3.2. Such "hybrid" models could benefit from the flexibility and ability for generalization of deep learning methods, while incorporating constraints imposed by physics.

This work focused on the thermal processes, but future work should also investigate modelling approaches for power distribution and load scheduling for microgrid connected edge data centers, with the end goal of formulating and testing holistic control and scheduling strategies for matching demand for computation with the available resources, e.g. renewable power, battery and coolant stored.

References

[1] Alissa, H. A., Nemati, K., Sammakia, B., Ghose, K., Seymour, M., and Schmidt, R. (2015). Innovative approaches of experimentally guided CFD modeling for data centers. In *2015 31st Thermal Measurement, Modeling & Management Symposium (SEMI-THERM)*. IEEE.

[2] Antal, M., Cioara, T., Anghel, I., Gorzenski, R., Januszewski, R., Oleksiak, A., Piatek, W., Pop, C., Salomie, I., and Szeliga, W. (2019). Reuse of data center waste heat in nearby neighborhoods: A neural networks-based prediction model. *Energies*, 12(5):814.

[3] Antal, M., Cioara, T., Anghel, I., Pop, C., and Salomie, I. (2018). Transforming data centers in active thermal energy players in nearby neighborhoods. *Sustainability*, 10(4):939.

[4] ASHRAE (2011). Thermal Guidelines for Data Processing Environments. . Technical report, The American Society of Heating, Refrigerating and Air-Conditioning Engineers.

[5] ASHRAE (2012). Datacom Equipment Power Trends and Cooling Applications. . Technical report, The American Society of Heating, Refrigerating and Air-Conditioning Engineers.

[6] ASHRAE (2014). Liquid Cooling Guidelines for Datacom Equipment Centers. . Technical report, The American Society of Heating, Refrigerating and Air-Conditioning Engineers.

[7] Avelar, V., Azevedo, D., and French, A. (2012). PUE: A Comprehensive Examination of the Metric. Technical report, The Green Grid.

[8] Ayoub, R., Indukuri, K., and Rosing, T. S. (2011). Temperature aware dynamic workload scheduling in multisocket cpu servers. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 30(9):1359–1372.

[9] Belady, Christian et al. (2010). Carbon Usage Effectiveness (CUE): A green grid data center sustainability metric. Technical report, The Green Grid.

[10] Belkhir, L. and Elmeligi, A. (2018). Assessing ict global emissions footprint: Trends to 2040 & recommendations. *Journal of Cleaner Production*, 177:448–463.

[11] Beloglazov, A., Abawajy, J., and Buyya, R. (2012). Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. *Future Generation Computer Systems*, 28(5):755–768.

[12] Brown, Richard et al. (2007). Report to Congress on Server and Data Center Energy Efficiency: Public Law 109-431. Technical report.

[13] Brännvall, R., Siltala, M., Gustafsson, J., Sarkinen, J., Vesterlund, M., and Summers, J. (2020). Edge: Microgrid data center with mixed energy storage. Forthcoming in the proceedings of the Eleventh ACM International Conference on Future Energy Systems (e-Energy '20).

[14] Capozzoli, A. and Primiceri, G. (2015). Cooling systems in data centers: State of art and emerging technologies. volume 83, pages 484–493.

[15] Carbó, A., Oró, E., Salom, J., Canuto, M., Macías, M., and Guitart, J. (2016). Experimental and numerical analysis for potential heat reuse in liquid cooled data centres. *Energy Conversion and Management*, 112:135–145.

[16] Carr, Nicholas G. (2005). The End of Corporate Computing. Technical report. In: MITSloan Management Review.

[17] Cho, J. and Kim, Y. (2016). Improving energy efficiency of dedicated cooling system and its contribution towards meeting an energy-optimized data center. *Applied Energy*, 165:967–982.

[18] Clipp, Celeste et al. (2014). Digital Infrastructure and Economic Development: An impact assessment of Facebook's data center in Northern Sweden. Technical report, The Boston Consulting Group.

[19] Davies, G., Maidment, G., and Tozer, R. (2016). Using data centres for combined heating and cooling: An investigation for london. *Applied Thermal Engineering*, 94:296–304.

[20] de Boer, G. N., Johns, A., Delbosc, N., Burdett, D., Tatchell-Evans, M., Summers, J., and Baudot, R. (2018). Three computational methods for analysing thermal airflow distributions in the cooling of data centres. *International Journal of Numerical Methods for Heat & Fluid Flow*, 28(2):271–288.

[21] Ebrahimi, K., Jones, G. F., and Fleischer, A. S. (2014a). A review of data center cooling technology, operating conditions and the corresponding low-grade waste heat recovery opportunities. *Renewable and Sustainable Energy Reviews*, 31:622–638.

[22] Ebrahimi, K., Jones, G. F., and Fleischer, A. S. (2014b). A review of data center cooling technology, operating conditions and the corresponding low-grade waste heat recovery opportunities. *Renewable and Sustainable Energy Reviews*, 31:622–638.

[23] Ebrahimi, K., Jones, G. F., and Fleischer, A. S. (2015). Thermo-economic analysis of steady state waste heat recovery in data centers using absorption refrigeration. *Applied Energy*, 139:384–397.

[24] Erden, H., Ezzat Khalifa, H., and Schmidt, R. (2013). Transient thermal response of servers through air temperature measurements. volume 2.

[25] Eriksson, M., Lucchese, R., Gustafsson, J., Ljung, A.-L., Mousavi, A., and Varagnolo, D. (2017). Monitoring and modelling open compute servers. volume 2017-January, pages 7177–7184.

[26] Fredriksson, S., Gustafsson, J., Olsson, D., Sarkinen, J., Beresford, A., Käufeler, M., Minde, T. B., and Summers, J. (2019). Integrated thermal management of a 150kw pilot open compute project style data center. In *2019 IEEE 17th International Conference on Industrial Informatics (INDIN)*, volume 1, pages 1443–1450.

[27] Glinkowski, Mietek (2014a). Data center defined. Technical report. In: ABB review: Data centers.

[28] Glinkowski, Mietek (2014b). Designed for uptime. Technical report. In: ABB review: Data centers.

[29] Ham, S.-W., Kim, M.-H., Choi, B.-N., and Jeong, J.-W. (2015). Energy saving potential of various air-side economizers in a modular data center. *Applied Energy*, 138:258–275.

[30] Han, X. and Joshi, Y. (2012). Energy reduction in server cooling via real time thermal control. In *Annual IEEE Semiconductor Thermal Measurement and Management Symposium*, pages 74–81.

[31] Healey, C., VanGilder, J., Condor, M., and Tian, W. (2018). Transient data center temperatures after a primary power outage. pages 865–870.

[32] Iyengar, M., Hamann, H., Schmidt, R. R., and Vangilder, J. (2007). Comparison between numerical and experimental temperature distributions in a small data center test cell. In *2007 Proceedings of the ASME InterPack Conference, IPACK 2007*, volume 1, pages 819–826.

[33] Jiang, H.-P., Chuck, D., and Chen, W.-M. (2016). Energy-aware data center networks. *Journal of Network and Computer Applications*, 68:80–89.

[34] Jones, N. (2018). The information factories, data centers are chewing up vast amount of energy - so researchers are trying to make them more efficient. *Nature*.

[35] Jouhara, H. and Meskimmon, R. (2014). Heat pipe based thermal management systems for energy-efficient data centres. *Energy*, 77:265–270.

[36] Karpatne, A., Atluri, G., Faghmous, J., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., and Kumar, V. (2017a). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2318–2331.

[37] Karpatne, A., Watkins, W., Read, J., and Kumar, V. (2017b). Physics-guided neural networks (pgnn): An application in lake temperature modeling.

[38] Khalaj, A. H. and Halgamuge, S. K. (2017). A review on efficient thermal management of air- and liquid-cooled data centers: From chip to the cooling system. *Applied Energy*, 205:1165–1188.

[39] Khalaj, A. H., Scherer, T., and Halgamuge, S. K. (2016). Energy, environmental and economical saving potential of data centers with various economizers across australia. *Applied Energy*, 183.

[40] Koomey, J., Berard, S., Sanchez, M., and Wong, H. (2011). Implications of historical trends in the electrical efficiency of computing. *IEEE Annals of the History of Computing*, 33(3):46–54.

[41] Liu, Z., Lin, M., Wierman, A., Low, S., and Andrew, L. L. H. (2015). Greening geographical load balancing. *IEEE/ACM Transactions on Networking*, 23(2):657–671.

[42] Lucchese, R. (2019). *Cooling Control Strategies in Data Centers for Energy Efficiency and Heat Recovery*. PhD thesis, Luleå University of Technology.

[43] Lucchese, R. and Johansson, A. (2019a). On energy efficient flow provisioning in air-cooled data servers. *Control Engineering Practice*, 89:103–112.

[44] Lucchese, R. and Johansson, A. (2019b). On server cooling policies for heat recovery: Exhaust air properties of an open compute windmill v2 platform. pages 1049–1055.

[45] Lucchese, R., Olsson, J., Ljung, A.-L., Garcia-Gabin, W., and Varagnolo, D. (2017). Energy savings in data centers: A framework for modelling and control of servers' cooling. *IFAC-PapersOnLine*, 50(1):9050–9057.

[46] Malmodin, J. and Lundén, D. (2016). The energy and carbon footprint of the ict and e&m sector in sweden 1990-2015 and beyond. In *ICT for Sustainability 2016*. Atlantis Press.

[47] Malmodin, J., Moberg, Å., Lundén, D., Finnveden, G., and Lövehagen, N. (2010). Greenhouse gas emissions and operational electricity use in the ICT and entertainment & media sectors. *Journal of Industrial Ecology*, 14(5):770–790.

[48] Minet, P., Renault, E., Khoufi, I., and Boumerdassi, S. (2018). Analyzing traces from a google data center. In *2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC)*. IEEE.

[49] Mittal, S. (2014). Power management techniques for data centers: A survey. *arXiv preprint arXiv:1404.6681*.

[50] Mohan Raj, V. and Shriram, R. (2016). Power management in virtualized datacenter – a survey. *Journal of Network and Computer Applications*, 69:117–133.

[51] Mukherjee, T., Banerjee, A., Varsamopoulos, G., Gupta, S. K., and Rungta, S. (2009). Spatio-temporal thermal-aware job scheduling to minimize energy consumption in virtualized heterogeneous data centers. *Computer Networks*, 53(17):2888–2904.

[52] Muralidhar, N., Islam, M., Marwah, M., Karpatne, A., and Ramakrishnan, N. (2019). Incorporating prior domain knowledge into deep neural networks. pages 36–45.

[53] Niemann, J., Brown, K., Avelar, V. (2011). Impact of hot and cold aisle containment on data center temperature and efficiency. . Technical report, Schneider Electric Data Center Science Center.

[54] Ntogramatzidis, L. and Ferrante, A. (2011). Exact tuning of pid controllers in control feedback design. *IET Control Theory and Applications*, 5(4):565–578.

[55] Open Compute Project (2019). Open Compute Project. http://opencompute.org/.

[56] Oró, E., Allepuz, R., Martorell, I., and Salom, J. (2018). Design and economic analysis of liquid cooled data centres for waste heat recovery: A case study for an indoor swimming pool. *Sustainable Cities and Society*, 36:185–203.

[57] Pardey, Z., Demetriou, D., Erden, H., VanGilder, J., Khalifa, H., and Schmidt, R. (2015). Proposal for standard compact server model for transient data center simulations. volume 121, pages 413–421.

[58] Parolini, L., Sinopoli, B., and Krogh, B. H. (2011). Model predictive control of data centers in the smart grid scenario. *IFAC Proceedings Volumes*, 44(1):10505–10510.

[59] Parolini, L., Sinopoli, B., Krogh, B. H., and Wang, Z. (2012). A cyber–physical systems approach to data center modeling and control for energy efficiency. *Proceedings of the IEEE*, 100(1):254–268.

[60] Patterson, Michael K., Bill Tschudi, et al. (2010). Ere: A Metric for Measuring the Benefit of Reuse Energy From a Data Center . Technical report, The Green Grid.

[61] Patterson, Michael K. D. G. Costello, et al. (2007). Data center TCO: a comparison of high- density and low-density spaces. Technical report.

[62] Patterson, Michael K., Dan Azevedo, et al. (2011). Water Usage Effectiveness (WUE): A green grid data center sustainability metric. Technical report, The Green Grid.

[63] Pärssinen, M., Wahlroos, M., Manner, J., and Syri, S. (2019). Waste heat from data centers: An investment analysis. *Sustainable Cities and Society*, 44:428–444.

[64] Rong, H., Zhang, H., Xiao, S., Li, C., and Hu, C. (2016). Optimizing energy consumption for data centers. *Renewable and Sustainable Energy Reviews*, 58:674–691.

[65] Samadiani, E. and Joshi, Y. (2010a). Multi-parameter model reduction in multi-scale convective systems. *International Journal of Heat and Mass Transfer*, 53(9-10):2193–2205.

[66] Samadiani, E. and Joshi, Y. (2010b). Proper orthogonal decomposition for reduced order thermal modeling of air cooled data centers. *Journal of Heat Transfer*, 132(7):1–14.

[67] Samadiani, E. and Joshi, Y. (2010c). Reduced order thermal modeling of data centers via proper orthogonal decomposition: A review. *International Journal of Numerical Methods for Heat and Fluid Flow*, 20(5):529–550.

[68] Sarkinen, J., Brannvall, R., Gustafsson, J., and Summers, J. (2020). Experimental analysis of server fan control strategies for improved data center air-based thermal management. In *2020 19th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*. IEEE (Forthcoming).

[69] Shehabi, Arman et al. (2016). United States Data Center Energy Usage Report. Technical report.

[70] Skadron, K., Stan, M. R., Huang, W., Velusamy, S., Sankaranarayanan, K., and Tarjan, D. (2003). Temperature-aware microarchitecture. In *Conference Proceedings - Annual International Symposium on Computer Architecture, ISCA*, pages 2–13.

[71] Telecommunications Industry Association (TIA) (2018). Edge Data Centers. Technical report.

[72] Toosi, A. N., Qu, C., de Assunção, M. D., and Buyya, R. (2017). Renewable-aware geographical load balancing of web applications for sustainable data centers. *Journal of Network and Computer Applications*, 83:155–168.

[73] VanGilder, J., Healey, C., Pardey, Z., and Zhang, X. (2013). A compact server model for transient data center simulations. volume 119, pages 358–370.

[74] VanGilder, J. W., Healey, C. M., Condor, M., Tian, W., and Menusier, Q. (2018). A compact cooling-system model for transient data center simulations. In *2018 17th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*. IEEE.

[75] Wahlroos, M., Pärssinen, M., Manner, J., and Syri, S. (2017). Utilizing data center waste heat in district heating – impacts on energy efficiency and prospects for low-temperature district heating networks. *Energy*, 140:1228–1238.

[76] Wahlroos, M., Pärssinen, M., Rinne, S., Syri, S., and Manner, J. (2018). Future views on waste heat utilization – case of data centers in northern europe. *Renewable and Sustainable Energy Reviews*, 82:1749–1764.

[77] Wang, Z., Bash, C., Tolia, N., Marwah, M., Zhu, X., and Ranganathan, P. (2010). Optimal fan speed control for thermal management of servers. In *Proceedings of the ASME InterPack Conference 2009, IPACK2009*, volume 2, pages 709–719.

[78] Zhang, P., Wang, B., Wu, W., Shi, W., and Li, X. (2015). Heat recovery from internet data centers for space heating based on an integrated air conditioner with thermosyphon. *Renewable Energy*, 80:396–406.

[79] Zhu, H., Liao, X., Laat, C., and Grosso, P. (2016). Joint flow routing-scheduling for energy efficient software defined data center networks: A prototype of energy-aware network management platform. *Journal of Network and Computer Applications*, 63:110–124.

[80] Åström, K. J. and Hägglund, T. (2001). The future of pid control. *Control Engineering Practice*, 9(11):1163–1175.

Part II

Paper A

Digital Twin for Tuning of Server Fan Controllers

Authors:
Rickard Brännvall, Jeffrey Sarkinen, Joar Svartholm, Jonas Gustafsson and Jon Summers

Reformatted version of paper originally published in:
2019 IEEE 17th International Conference on Industrial Informatics (INDIN 2019).

Paper B

Experimental Analysis of Server Fan Control Strategies for Improved Data Center Air-based Thermal Management

Authors:
Jeffrey Sarkinen, Rickard Brännvall, Jonas Gustafsson and Jon Summers

Reformatted version of paper accepted for publication in:
The Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm 2020).

Paper C

Data Center Excess Heat Recovery: A Case Study of Apple Drying

Authors:
Rickard Brännvall, Louise Mattsson, Erik Lundmark and Mattias Vesterlund

Reformatted version of paper accepted for publication in:
33rd International Conference on Efficiency, Cost, Optimization, Simulation and Environmental Impact of Energy Systems.

Paper D

EDGE: Microgrid Data Center with Mixed Energy Storage

Authors:
Rickard Brännvall, Mikko Siltala, Jonas Gustafsson, Jeffrey Sarkinen, Mattias Vesterlund and Jon Summers

Reformatted version of paper accepted for publication in:
8th International Workshop on Energy-Efficient Data Centres at e-Energy '20: The Eleventh ACM International Conference on Future Energy Systems

Paper E

Machine Learning based Thermal Control of Edge Server Cluster: Towards Transfer Learning

Authors:
Rickard Brännvall, Jeffrey Sarkinen, Jon Summers, Jonas Gustafsson, Fredrik Sandin

To be submitted.