Department of Physics and Astronomy
University of Heidelberg

Bachelor Thesis in Physics
submitted by

Nico Schwersenz

born in Mosbach (Germany)

2019

Optimizing Cooling Performance and Efficiency of the BrainScaleS System

This Bachelor Thesis has been carried out by Nico Schwersenz at the
Kirchhoff Institute for Physics in Heidelberg
under the supervision of
Dr. Johannes Schemmel

Optimizing Cooling Performance and Efficiency of the BrainScaleS System

The server room of the Electronic Vision(s) group contains 20 BrainScaleS systems. Every system is built around a wafer, that contains neuromorphic hardware. Each of the systems consumes 1.2 kW of electric power in standby, that have to be cooled. An air conditioning capable of 20 kW cooling power is therefore installed to be able to use most of the systems. The air conditioning produces air as low as 16 °C. This air is just cold enough to keep the BrainScalesS systems at a temperature of 50 °C. As the neuromorphic circuits on the systems are analog, this temperature of 50 °C is desired to be able to run stable experiments. The aim of this thesis was to increase cooling performance of the system while keeping the cooling as efficient as possible. In order to minimize the temperature difference between air inlet and wafer, two different fans were installed in the system. These fans were then tested at different speeds and in different numbers to decrease the power consumption of the cooling. In the course of this thesis also a ducts was installed into the system to increase the air flow around the wafer and increase cooling performance. The goal of $\Delta T = 25\,\mathrm{K}$ was reached with a power usage effectiveness below 1.11 by a cardboard duct and the current heat sinks. For a 3D printed duct the power usage effectiveness is at 1.15, which could be improved further.

Optimieren der Kühlleistung und Kühleffzienz des wafer-scale Systems

Der Serverraum der Electronic Vision(s) Gruppe beinhaltet 20 BrainScaleS Systeme. Jedes System ist um einen Wafer aufgebaut, der neuromorphe Schaltkreise enhält. Jedes dieser Systeme hat eine Leistungsaufnahme von 1.2 kW im Standby-Betrieb, die zu kühlen sind. Eine Klimaanlage, die über eine Kühlleistung von 20 kW verfügt, ist deshalb installiert, um viele System gleichzeitig nutzen zu können. Die Klimaanlage erzeugt Kaltluft mit einer Minimaltemperatur von 16 °C. Diese Luft is gerade kalt genug, um die BrainScaleS Systeme bei einer Temperatur von 50 °C zu halten. Die neuromorphen Schaltkreise auf den Systemen sind analog, weshalb es für stabile Experimente notwendig ist, diese Temperature von 50 °Czu halten. Das Ziel dieser Arbeit war, die Kühlleistung zu erhöhen. Gleichzeitig sollte auch die Kühleffizienz verbessert werden. Um die Temperaturdifferenz zwischen Lufteinlass und Wafer zu minimieren, wurden zwei verschiedene Lüfter im System installiert. Diese Lüfter wurden bei unterschiedlicher Leistung und in unterschiedlicher Anzahl getestet, um die Leistungsaufnahme der Kühlung zu reduzieren. Im Rahmen dieser Thesis wurden ebenso Luftleitbleche im System installiert, die den Luftstrom um den Wafer und damit die Kühlleistung erhöhen sollten. Das Ziel von $\Delta T = 25\,\mathrm{K}$ wurde mit einer 'power usage effectiveness' unter 1.11 durch ein Kartonluftleitblech erzielt. Ein 3D gedrucktes Luftleitblech, welches noch verbessert werden kann, erreichte eine 'power usage effectiveness' von 1.15.

Contents

1 Introduction

There are two approaches to simulate neurons and therefore the human brain's behaviour. The first one is the digital approach, where individual neurons are simulated on traditional computers. For more complex neural networks this method is becoming inefficient and consumes large computational power. A higher efficiency is desired, not only because the brain is very efficient, but also to save resources (time, energy,...).

The approach for simulating neurons, which the Electronic Vision(s) group at the Kirchhoff-Institute for Physics in Heidelberg pursues, is the hardware approach. Neurons are implemented as analog circuits on the High Input Count Analog Neural Network (HICANN) [1] [2] chip on a wafer in the BrainScaleS [3] system. HICANN chips are partitioned in groups of 8, called reticles, of which 48 exist on each wafer.

The server room of the Electronic Vision(s) group contains 20 BrainScaleS systems. Each of the BrainScaleS systems consumes about 1.2 kW of electrical power in standby, that have to be cooled. Therefore, an air conditioning, capable of 20 kW cooling power, is installed to be able to use most of the systems. The air conditioning produces air as low as 16 °C. This air is just cold enough to keep the BrainScalesS systems at a temperature of 50 °C. As the neuromorphic circuits on the systems are analog, this temperature of 50 °C is desired to be able to run stable experiments.

The server-room will be moved to the European Institute for Neuromorphic Computing (EINC) [4] as soon at it will be finished. The new server room contains a heat exchanger, which is able to produce air down to a temperature of 25 °C. As this temperature is 9 °C higher than currently, it would not possible to run the BrainScaleS systems in the new building at a wafer temperature of 50 °C with the current cooling. The goal was therefore to investigate, if it is possible to reduce the temperature difference between air inlet and wafer of the BrainScaleS system to 25 K. And if so, can power consumption of the cooling be reduced. This means reducing the air flow on colder locations and increasing it on hotter locations, i.e. no cold air is at the outlet.

To maximize the cooling performance, more powerful fans were built into the systems. Furthermore, a duct was built out of cardboard first, then was 3D printed, to increase the air flow at the wafer's heat sinks and therefore cooling performance. For increasing the efficiency of the cooling, the speed of certain fans was lowered. Fans have also been omitted. The duct, that should increase performance, is also expected to increase efficiency, as it should prevent cold air to pass the wafer unheated.

After a short theoretical overview the BrainScaleS system will be described in more detail. It is followed by the main part of this thesis: the measurements. The measurements will be evaluated and summarized. The thesis ends with an outlook.

2 Theory

2.1 Thermal Resistance

For the description of heat sinks the most important parameter is the thermal resistance R defined by equation 2.1. It describes the temperature increase of an object due to an heat flow through it. The highest recommended power for a heat sink is called Thermal Design Power (TDP, given in W) . Which temperature increase ΔT manufacturers take for calculating the TDP is unclear, as the thermal resistance depends on the air flow (see section 2.2) and therefore the used fans. Most server's cpus run at 65 °C case temperature. With an average room temperature of 20 °C, ΔT can be assumed to be 45 K. In general the following is the case: the higher the TDP value of the heat sink, the lower the temperature of the system at full power.

$$R = \frac{\Delta T}{P} \tag{2.1}$$

2.2 Energy Transport

The air flow is usually given in Cubic Feet per Minute (CFM) or in $\mathrm{m^3\,min^{-1}}$. Without added fans it is called natural convection, forced convection is the name for the air flow with added fans or some other force. While it is clear that fans reduce the temperature of a heat sink and therefore reduce thermal resistance, it is complex how thermal resistance depends on air flow in theory. This behaviour is covered by the field of fluid dynamics and can be described by the Navier-Stokes-Equation [5]. It is highly non-linear, because of turbulence and its behaviour on complex geometries can only be simulated. Therefore measurements of the thermal resistance vs. air flow are given by the manufacturers, see figure 3.4.

It is still possible to describe the air flow by the means of energy transport. Equation 2.2 - commonly called heat transfer equation in its integrated form - shows how much power is dissipated by transporting some material of mass m with heat capacity c, heated by a temperature difference ΔT, away from the heat source. Air has a heat capacity of $1\,\mathrm{kJ\,kg^{-1}\,K^{-1}}$ [6], for water it is $4.2\,\mathrm{kJ\,kg^{-1}\,K^{-1}}$ [7]. The same mass of water can transport 4.2 times the amount of energy. While being a lot denser a lower volumetric flow is needed for the same energy transport. This is the reason why water cooling is more effective and more efficient.

$$P = \frac{\partial m}{\partial t} c \Delta T \tag{2.2}$$

2.3 Cooling Efficiency: PUE

The cooling efficiency of a data center is usually compared via the Power Usage Effectiveness (PUE) . The PUE describes the relation between the power spent including all facility factors, like lighting - which can not be included in this thesis (P_{total}) - and the power driving the computation (the 48 V power $P_{48\,\mathrm{V}}$), see equation 2.3. The PUE does not include factors of climate e.g. the same data center placed at the north pole and at the equator would have the same PUE. The power for an air conditioning is therefore neglected. The PUE is 1 for a perfect system, where no energy is needed for running the facility itself, but only the servers.

$$PUE = \frac{P_{total}}{P_{48\,\mathrm{V}}} = 1 + \frac{P_{Cooling}}{P_{48\,\mathrm{V}}} \tag{2.3}$$

3 System Overview and Techniques

First, the system itself will be described containing information about the construction, that will be needed for understanding the following chapter.

Second, used heat sinks and fans will be introduced and possible improvements to the cooling of the system will be explained.

Third, the techniques used for improving the cooling performance and efficiency will be presented.

3.1 System Overview

The wafer-scale system consists of the following components (from bottom to top in the picture,): Wafer I/0 PCB , bracket, Field Programmable Gate Arrays (FPGAs) , insertion frame, wafer, positioning mask, main PCB , top cover, AnaB (containing only plugs) and power supply (PwrAux). They can be seen in figure 3.1. For a list of abbreviations see 7.2.

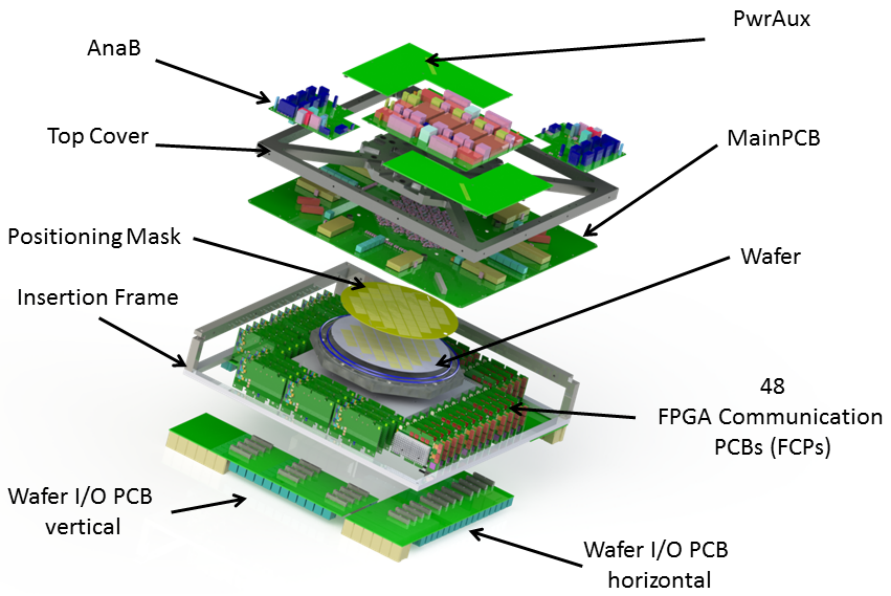

Figure 3.1: Explosion view of the brainScaleS system

The main heat sources of a BrainScaleS system are the wafer in the center with its 48 reticles and the 48 FPGA boards surrounding it, see figure 3.2. A single FPGA board consumes 10 to 12 W and one reticle consumes about the same power, which will be converted into heat in the end. The total power of one system is therefore about 1.2 kW.

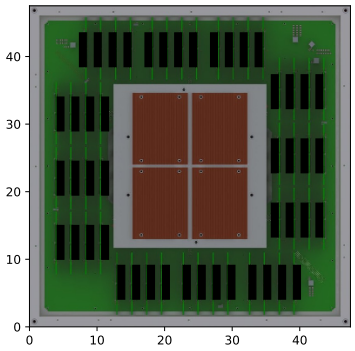

Figure 3.2: Bottom view of the BrainScaleS system: The wafer in the center is covered by an aluminium base plate and four copper heat sinks, which are surrounded by 48 FPGA boards. The FPGAs heat sinks are marked by black rectangles

An wafer, especially designed for power tests, called power wafer, was used during all experiments. This wafer differs in the functionality from the wafer used for neuromorphic experiments, which is called usual wafer from now on: the reticles are not working as neuromorphic circuits but resistors and their power consumption is twice as high. For a better orientation the wafer directions have been introduced: north, south, west, east. In figure 3.2 this corresponds to top, bottom (not the vertical positions), right and left, because the system is pictured from the bottom view and is therefore mirrored.

On this wafer a reticle distribution has to be found so that the dissipated power is maximized. The limit for the power dissipation is the 50 °C limit of the wafer. In order to maximize the dissipated power, the heat distribution has to be as even as possible across the wafer. The Reticles have to be turned on accordingly.

This distribution is difficult, because the reticles affect the temperatures differently according to their position, as the air flow is not symmetrical in every direction, but air flows from the south to the north. Reticles closer to the air inlet have an effect on all the other wafer temperatures, because heated air from the south passes all other directions, too. Reticles closer to the air outlet on the other hand just have an effect to their close surrounding. As a consequence the temperature distribution on the wafer is not even, but has differences of at least 2 K (see e.g. fig. 4.9).

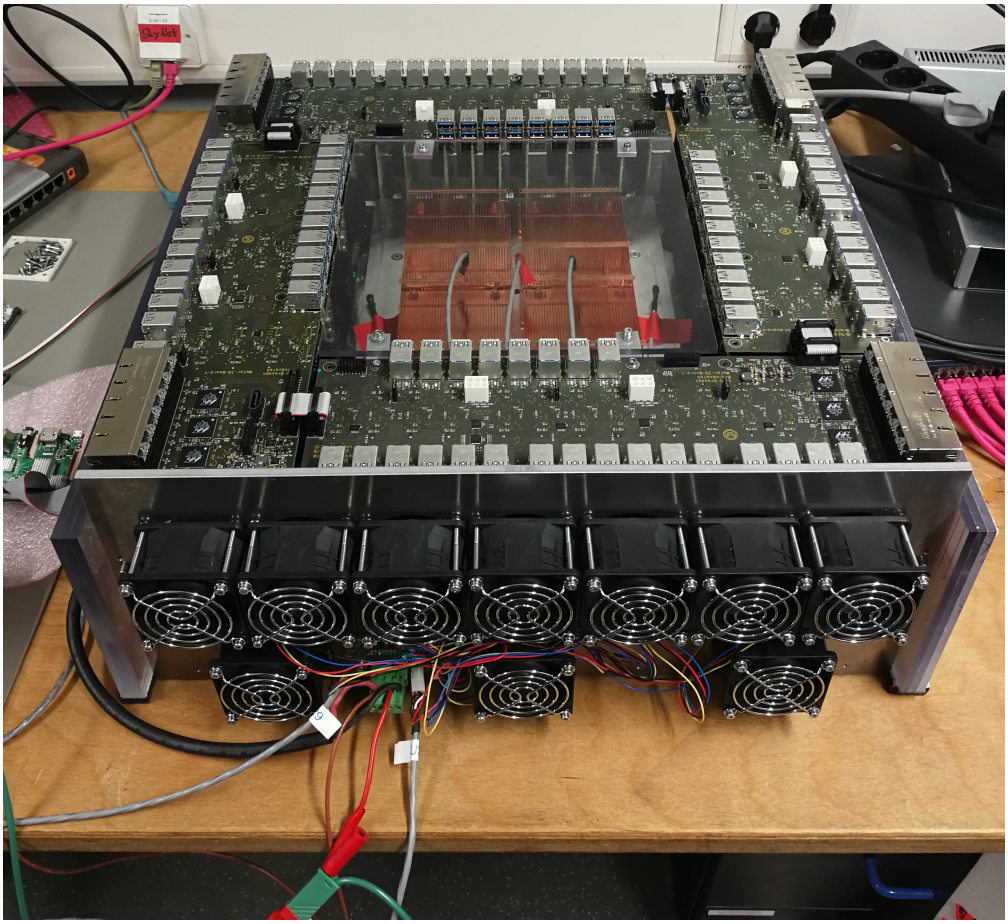

Figure 3.3: Picture of the test setup

The test setup can be seen in figure 3.3. The BrainScaleS system can be seen from the bottom view and the south, where the 10 fans are mounted on an aluminium plate. The sides are covered with three plexiglass plates, on which the system stands. One can see the quadratic plexiglass plate with a distance to the four copper heat sinks, where 5 added temperature sensors are placed with cables, mounted on an aluminium base plate. The four wios surrounding the plexiglass plate are connected to the FPGAs beneath them. In the south, cables of a voltmeter can be seen. The fan board is where these cables are connected to.

3.2 Fans and Heat sinks

The wafer is currently cooled by 4 copper heat sinks [8]. Every FPGA board has its own heat sink and 10 fans [9] are installed in total. On top there are three of these fans cooling the power supply. The other seven fans are pushing air through the heat sinks of FPGAs and wafer. Different heat sinks [10] were selected, but not tested yet and different fans [11] were used in later test. Heat sink 2 was chosen for its low thermal resistance compared to others and for its different working principle: heat pipes instead of vapor chamber base for heat sink 1. Also because of its greater build height of 2 U $\approx$ 89 mm it was expected to perform better at lower air flows, which could increase efficiency. All of the compared heat sinks are from Dynatron, because they don't have a minimum order quantity and still are mass-produced and therefore cheap, while custom made heat sinks have a minimum order quantity of a few hundreds and are more expensive.

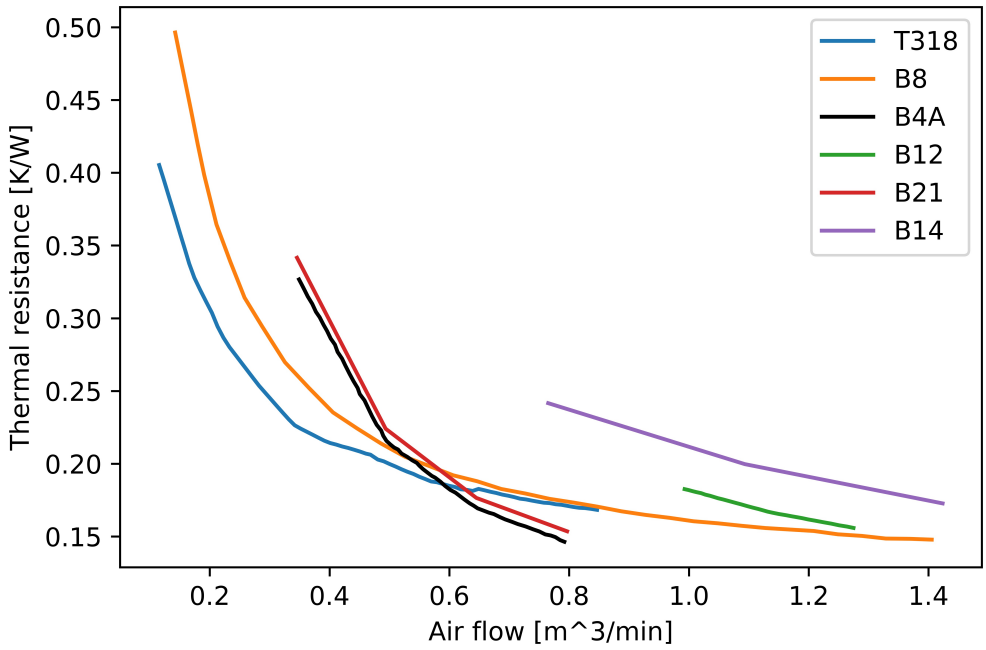

Figure 3.4: Thermal resistance vs. air flow of a few heat sink models from Dynatron including heat sink 1 (T318) and heat sink 2 (B8)

In figure 3.4 the thermal resistance vs. air flow is shown for different heat sinks. The plots were extracted from the specification sheets which just covered a small air flow spectrum. Therefore, fits were done to extrapolate the thermal resistance. The fit function was taken from the specification sheets and is an exponential function. The fit does not work perfectly for higher air flows, but gives a first impression of the behaviour of the heat sinks. Heat sink 1 reaches saturation faster than heat sink 2 and at a higher

thermal resistance, which can be seen in the shallow slope of the graph. Heat sink B4A was also considered, but has the same build height as heat sink T318. Because of a higher expected efficiency heat sink B8 was chosen.

The static pressure vs. air flow of fan A and fan B is shown in figure 3.5 shows . Data were again extracted from the specifications sheets. While both fans reach almost the same air flow at zero pressure, fan b delivers a static pressure more than twice as high at zero air flow compared to fan a. This pressure keeps a high airflow in small cross sections, which will be needed for tests including a duct. A higher pressure could lead to higher efficiency. A fact, that's influence is not further discussed, is the more focused air flow of fan B - compared to fan A - due to its two counter rotating fans. Two motors and two blades are built into one fan, which has two power plugs.

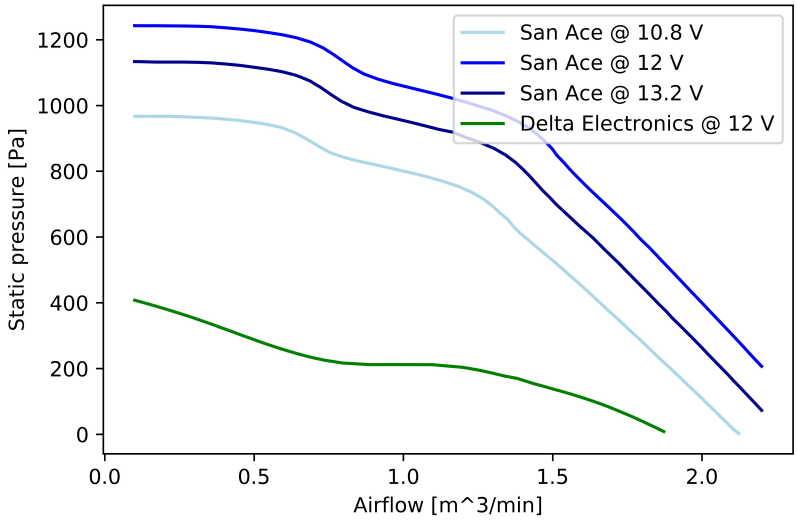

Figure 3.5: Air flow vs. static pressure of Delta Electonics (fan A) and San Ace (fan B). While both have approximately the same air flow at zero pressure, the static pressure at zero air flow is more than twice as high for fan B.

3.3 Techniques

To improve the cooling performance of the system, following steps are possible:

- using different fans
- using different heat sinks
- installing a duct

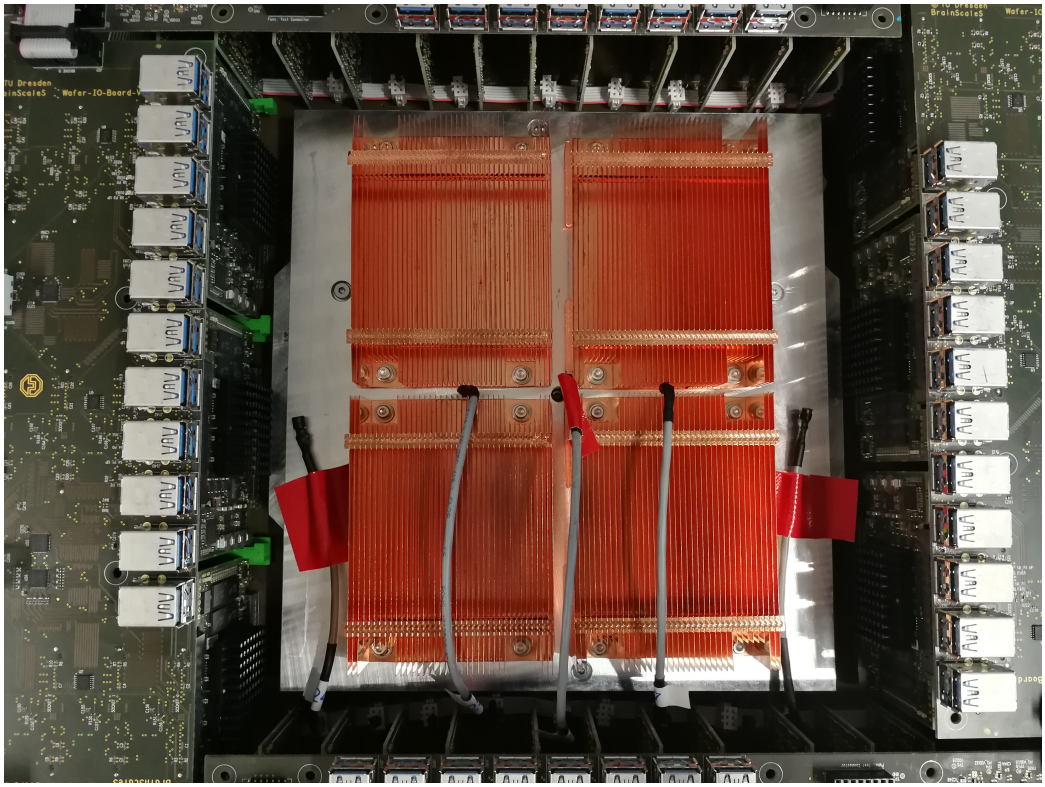

Figure 3.6: Zoomed in bottom view of the test setup. The plexiglass plate is removed so that the installed sensors can be seen better. One sensor row can be seen in the north, one is almost not visible in the south. 5 sensors are placed in the middle of the center via cables connected to the row in the south

To investigate these steps, 28 temperature sensors [12] were installed into the system. Ribbon cables were taped flat to the main pcb to minimize the influence on the air flow: one in front of the wafer, one behind and one in a row of FPGAs at the side. 23 sensors were directly plugged to the ribbon cables. Because the space inbetween wafer heat

sinks is not wide enough to tape a ribbon cable to the aluminium plate, 5 sensors are connected with normal cables to one ribbon cable in the south and placed in the middle of the wafer. In this configuration determining the temperature distribution of the air perpendicular to the air flow around the wafer is possible, where about half of heat is produced. The heat contribution of the FPGAs is assumed to be symmetrical, because the FPGAs are arranged almost symmetrically to the air flow direction. Therefore one sensor row at the side is enough to estimate the influence of the FPGAs' heat. The currents of the 48 V input of the system and the fan power supply was measured with a current clamp [13]

The experiment was set up in a room of about 50 m^3 with an air conditioning [14] set to 19 °C. The procedure of the measurements was turning on different voltages on the wafer-scale system, then turning on the fans and finally turning on the FPGAs and reticles. Which FPGAs and reticles were switched on specifically is discussed in chapter 4. Waiting for the temperature equilibrium takes about half an hour, which is why one measurement about an hour. Sometimes an undesired influence by the air conditioning or by people opening the experiment room's door, see figure 4.13, occurred. Measurement had to be extended due to these factors so that there was only time for one measurement for most of the configurations of fans, heat sinks and ducts.

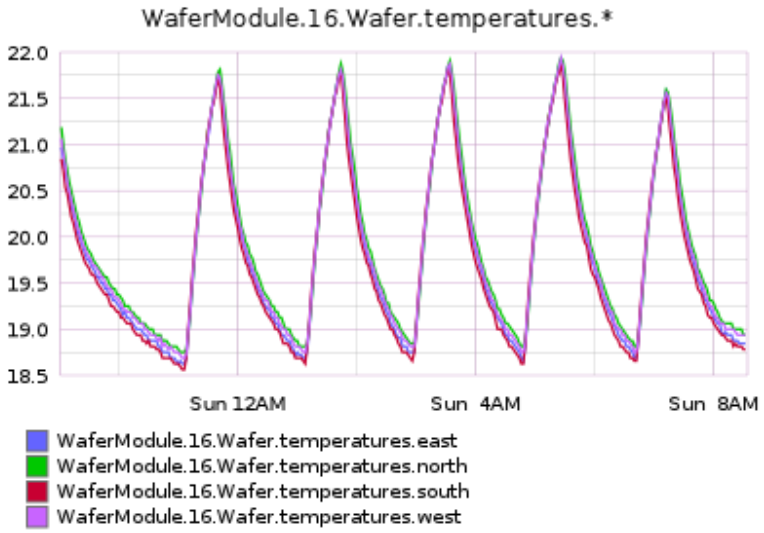

Figure 3.7: Temperature oscillation of 3 K and 2 h caused by the switching of the air conditioning.

These configurations sometimes only differ by a few degrees in thermal equilibrium, while the air conditioning already oscillates with an amplitude of 3 K, see 3.7, and a period of 2 h. The switching of the air conditioning is unpredictable, therefore an absolute temperature error of 3 K must be assumed for all experiments.

4 Measurements

Two plots were created for a measurement: one showing how the temperatures of the wafer and the sensors evolve over time and one showing the temperature distribution of the system in equilibrium. Improvements and further measurements were derived from these plots.

4.1 Current State

The current state includes 10 fans of type a in total and four heat sinks of type 1. Only 22 of the 48 reticles and their corresponding FPGAs were switched on to get the desired 50 °C wafer tempertature. The exact reticle distribution can be seen in table 7.4. The first distribution is valid until section 4.7.

In figure 4.1 the temporal temperature profile of all sensors is shown. The left plot shows absolute temperatures, the right plot shows the same temperatures, but relative to a reference temperature. The reference temperature is given by a temperature sensor close to the air inlet near a fan. This sensor is expected to have the air inlet temperature. Still there are some influences on the sensor's temperature, because the air gets heated by the fans itself, possibly some turbulences of air between FPGAs occur. All sensors reach thermal equilibrium after about half an hour.

Figure 4.2 shows the system from the bottom in equilibrium. The wafer temperatures are at 53.1 °C on average and are expected to be higher than the air temperatures, as they are near the heat source. The influence of the wafer heat sinks on the air temperature can be seen in the temperature of the northern row of air sensors.

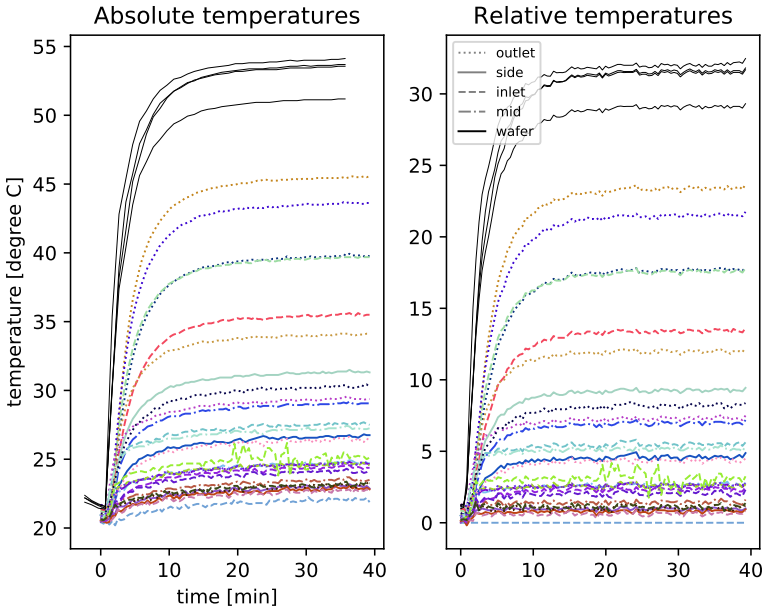

Figure 4.1: Temporal temperature profile of the system in the current state. The absolute temperatures are on the left, the temperatures relative to the reference sensor on the right. The 28 mounted sensors are in color, the wafer temperatures in black. The line style depends on the position: sensors on the air inlet are dashed, on the outlet dotted, in the mid dash-dotted and on the side lined.

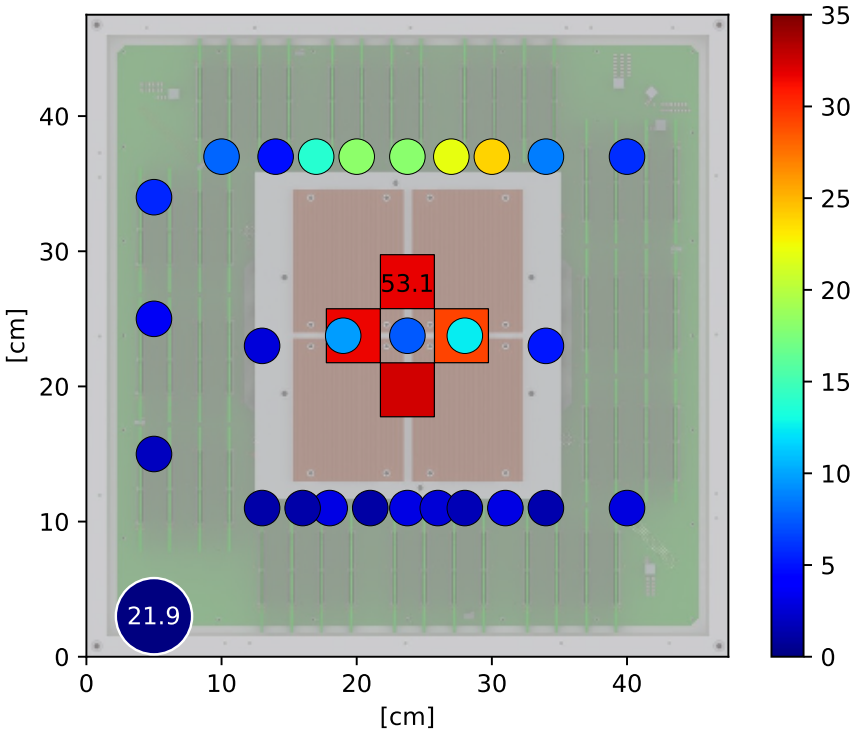

Figure 4.2: Heat map of the system in the current state in thermal equilibrium. Plotted are relative temperatures of all sensors color coded by the colorbar on the right side. The air sensors are marked by a circle at their actual position in the system. The reference sensor on the bottom left is marked by a larger circle. The reference temperature is written into the circle. The wafer temperatures are indicated by a square and the average is written into the northern square.

4.2 Cardboard Inlet

As systems in use for neuromorphic tests are currently mounted vertically, the air flow is not optimal as in the previous section described. The air has to flow through a duct to get to the fans. This is simulated by a cardboard air guide mounted on the air inlet side. This duct has the same height as the system height (about 20 cm), a distance from the fans of 6 cm and extends all along the fan side of the system. The resulting inlet would be on the bottom right in figure 4.3.

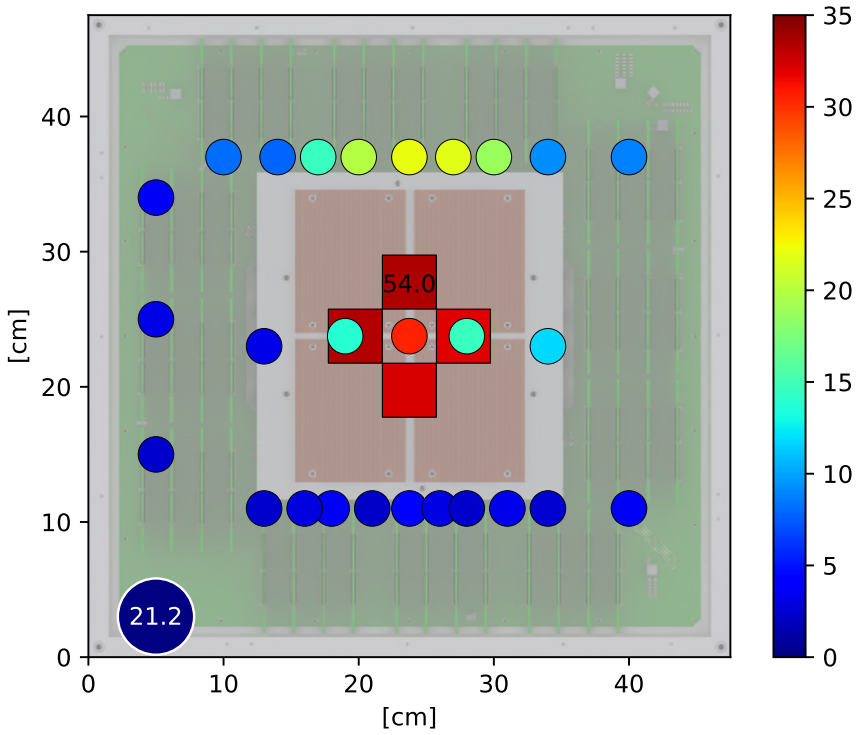

Figure 4.3: Heat map of the system in equilibrium, see fig. 4.2. A piece of cardboard was mounted in front of the inlet to simulate the effect of the vertical system installation.

For the air guide on the inlet the temperature of the wafer increases by $2K$ compared to figure 4.2. The difference of the hottest air sensor is $5K$. The duct on the inlet therefore reduces efficiency of the cooling. Future systems will be build horizontally into a rack system, where there won't be a duct and this effect does not occur.

There is an air guide not only on the inlet, but also on the outlet. The influence of this duct on the outlet on the air flow is unpredictable, because it has fans built in.

It is unexpected that the sensor right at the center shows a temperature really close

to the wafer temperature. This is due to the mounting of this sensor. It was placed in a hole, that is thermally connected to the wafer via heat sink compound. This sensor has to be neglected until section 4.4. The experiment in section 4.1 was carried out as one of the last experiments, which is why it does not show the higher temperature of the centered sensor.

4.3 Top sealed

There are many gaps between the wios, air can escape without moving through the whole system. It was of particular interest to find out how closing these gaps would affect the cooling performance. All gaps were closed with tape.

Compared to figure 4.1, figure 4.4 shows only minor changes. One may conclude that the air pressure inside the system produced by the fans is not high enough to push a significant amount of air through the gaps. It is also possible that closing the gaps does not affect the system, because there are already many ways for the air to avoid the heat sinks. For the next experiments the top will remain sealed, as there is a noticeable amount of air moving out of the system.

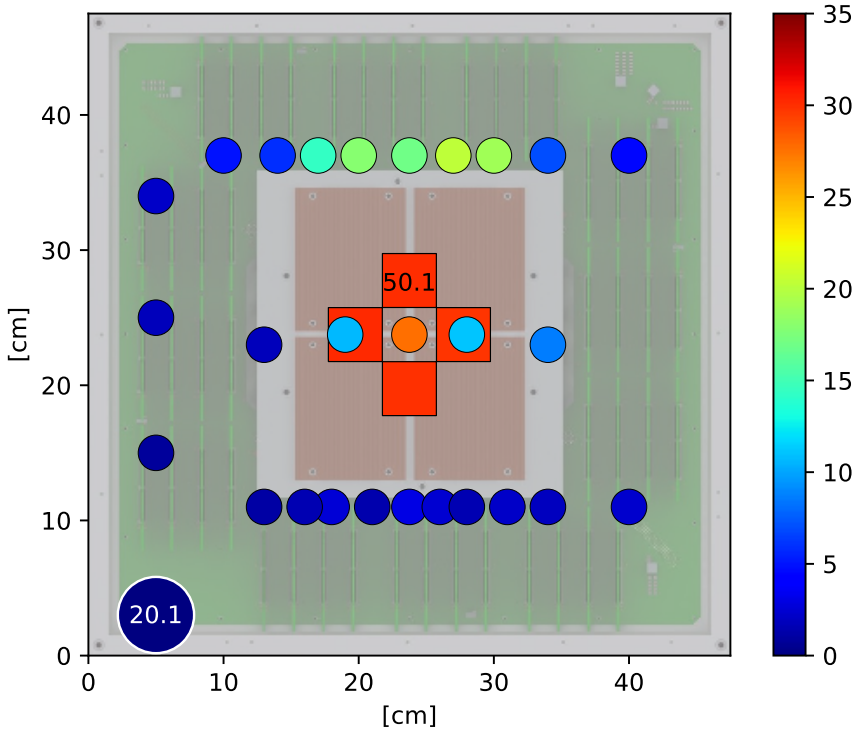

Figure 4.4: Heat map of the system in equilibrium, see fig. 4.2, with sealed gaps between wios.

4.4 All FPGAs switched on

A reticle on a usual wafer uses roughly half the power as the reticle on the power wafer. This was the reason to switch on just half of the reticles and their accompanying FPGAs for the previous measurements. This leads to a distortion of the results: On a usual wafer most of the reticles are switched on and so are their corresponding FPGAs.

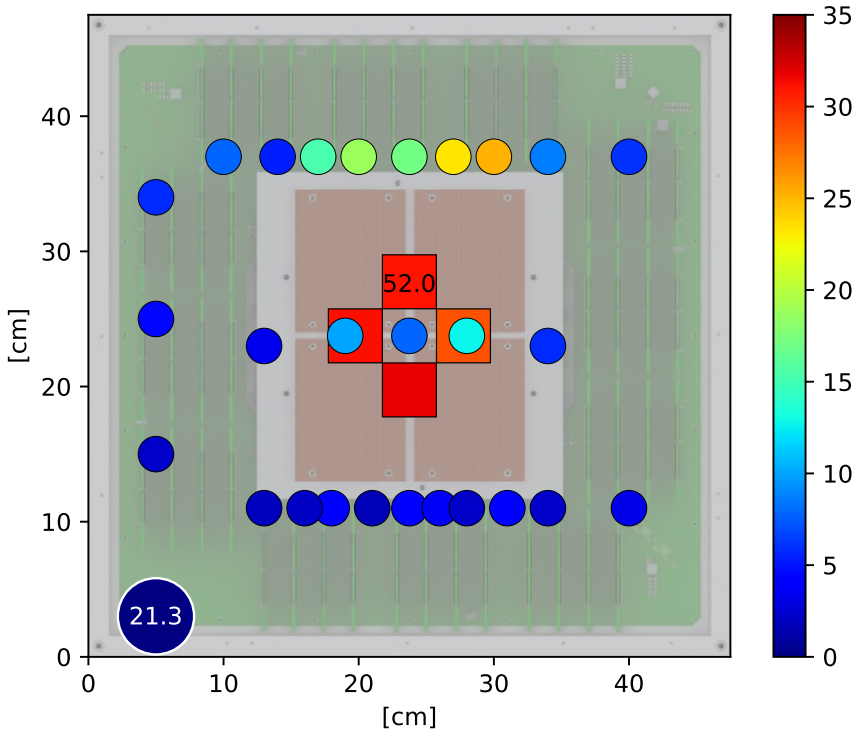

Figure 4.5: Heat map of the system in the current state in thermal equilibrium,see fig. 4.2. All FPGAs are switched on.

An uneven heat distribution of heat amongst the wafer's temperature sensors can be seen in figure 4.5. The wafer temperature in the north stays the same, while the other three temperatures increase by 1 K, 2 K and 3 K.

The reason for that is the direct influence of the additionally switched on FPGAs. The air gets more heated by these FPGAs and reaches heat sink 1 already with a higher temperature, the sensor in the south gets hotter. With increasing distance from the inlet the air moving through heat sink 1 gets more mixed with the colder air from the sides, which is not noticeably warmer than in figure 4.4. The sensor in the north is therefore not affected by the change of switched on FPGAs.

The position of the centered sensor was corrected in this measurement, see figure 4.5.

The temperature of this sensor can be taken into account from now on. It is about 3 K lower than the neighboring temperatures. Air in the center can move more freely, because of a small distance between adjacent heat sinks and is therefore colder. The air is warmer close to the neighboring sensors, because the fin density is high there.

4.5 Fans sealed

All fans are currently screwed to a aluminium plate by four M3 screws and have a grille on both sides for finger protection. To decouple the vibrations of the fans from the system, every fan is mounted with eight rubber washers. This results in a small distance between fan and aluminium plate, where air can escape. A 3D printed part (Squishy Washer-Alike Grille, SWAG) made of TPU plastic [15] with a shore D hardness of 45 [16] was introduced to prevent air from escaping and to decouple the fans' vibrations from the system due to the materials flexibility. A 3D model of the printed part made in OpenSCAD can be seen in figure 4.7. A small drawback of this part, that should be fixed in the design, is that the grilles facing into the system can't be mounted anymore.

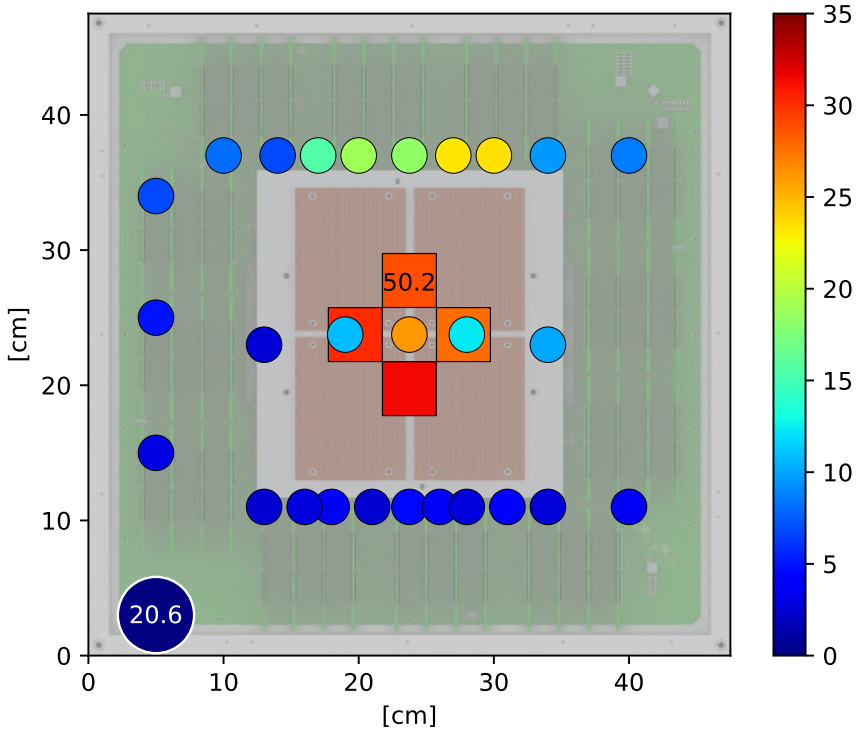

Figure 4.6: Heat map of the system in the current state in thermal equilibrium, see fig. 4.2, with sealed fans.

Compared to figure 4.5, figure 4.6 shows a reduction of the wafer temperature by 1 K. One of the air sensors in the north also shows a difference of 1 K. As the used part increases the cooling performance, it will be used in the next experiments.

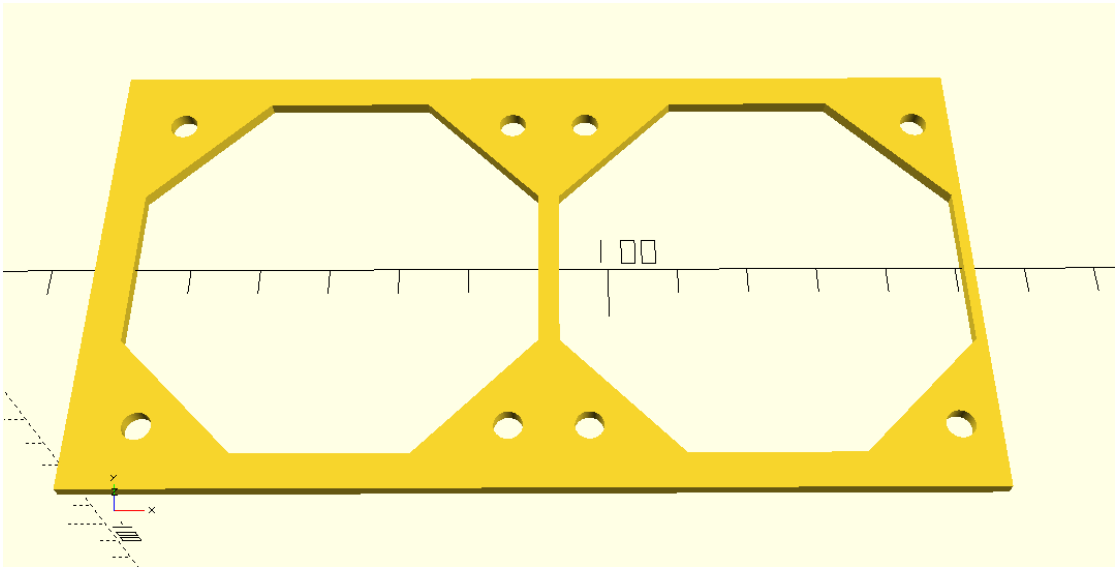

Figure 4.7: 3D model of the printed part made with OpenSCAD.

4.6 Cardboard duct

The next step was to include a duct for heat sink 1. The questions was: what should this duct look like? It was reasonable to minimize the air flow between heat sink 1 and the plexiglass. The heat sinks fins were covered with a piece of cardboard, while the sides were closed with tape. Now a part had to be constructed, that widens up toward the air inlet in order to collect air from a larger surface and guide it into the first part and through the heat sinks.

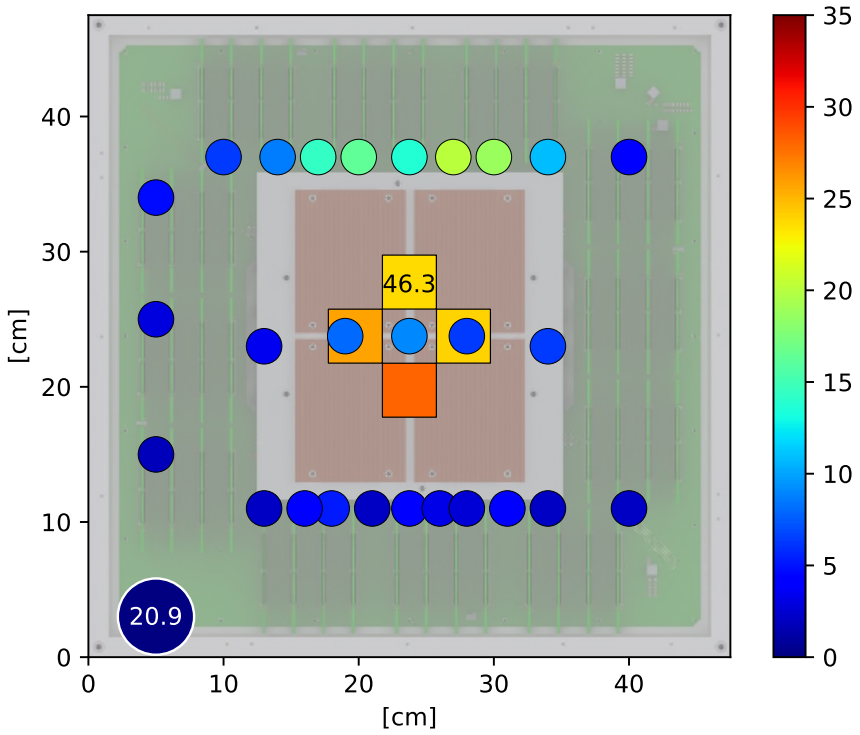

Figure 4.8: Heat map of the system in equilibrium, see fig. 4.2 with a cardboard duct.

At first the plexiglass was removed, because then it was easier to mount the duct. The outer most column of FPGAs has a distance of about 5 cm to the side wall. To reduce air flow in this area, a piece of foam was built in on the left and the right.

The average wafer temperature decreases by by 4 K compared to figure 4.6. The assumption was that colder air from the sides mixes with the hotter air from the wafer heat sinks. The duct prevents this mixing and should lead to higher temperatures in the north. Instead the opposite effect is observed, see figure 4.8: with the duct installed, the temperature in the north decreases more than other wafer temperatures. The explanation may be that more air passes the second row of wafer heat sinks and

air can't move through the system in the gap between wafer heat sink and plexiglass anymore.

4.7 New Reticle Distribution

As the wafer temperatures differed by about 5 K in the previous section, a new distribution of reticles was found, see table 7.4, which reduces the temperature difference to 1.5 K. As before 22 reticles are switched on so that the power consumption stays the same compared to previous tests. Figure 4.10 is intended to be the new reference to following measurements.

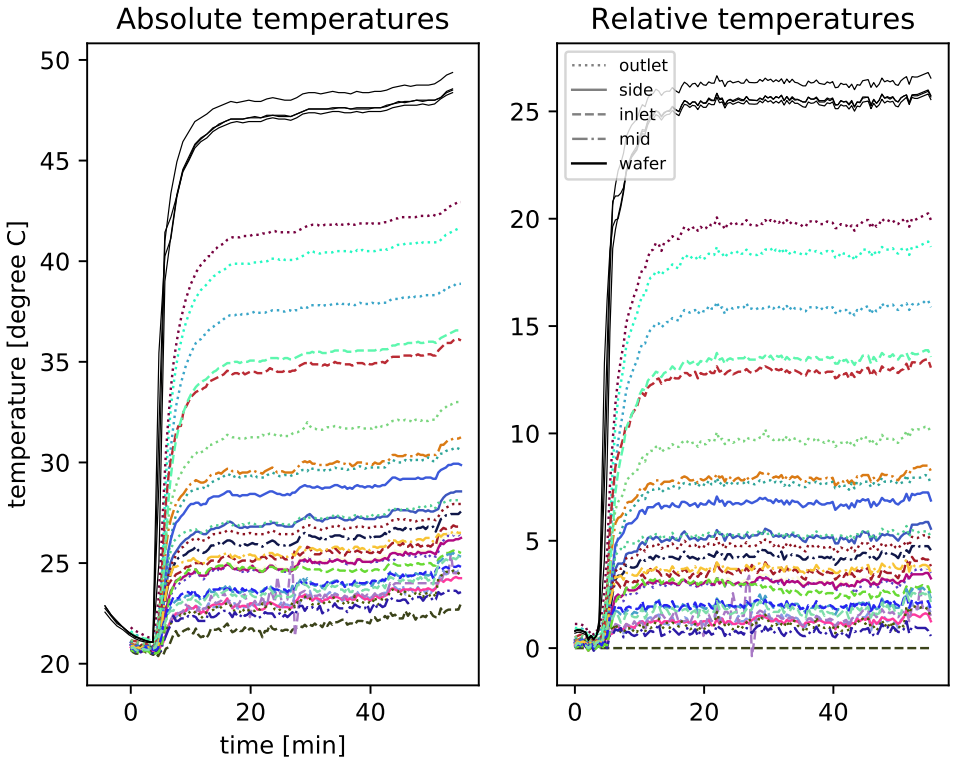

Figure 4.9: This temporal temperature profile with a different reticle distribution will be reference from now on.

This state was improved by mounting the plexiglass on top again, which did not make a difference to the temperatures (because the duct already keeps the air inside). The next measurement included extending the duct right to the fans again with foam. That means closing the small gap between fan and beginning of the FPGA so that 3 of the 7 fans push air through the duct only. This change also did not reduce temperatures further. Because of that, the state shown in figure 4.9 and figure 4.10 represents the best cooling performance possible with fan A and heat sink 1.

One can now conclude that the saturation of either heat sink 1 or of fan A is reached.

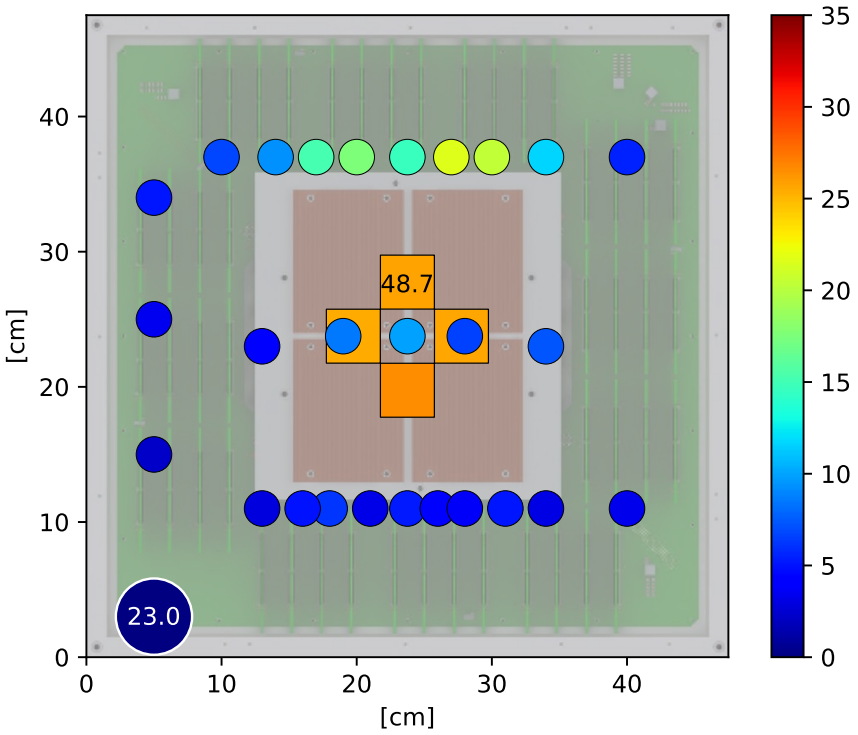

Figure 4.10: Heat map of the system with duct and different reticle distibution.

4.8 Reducing Power Consumption of the Cooling with Fan A

Starting from section 4.7 the power of the fans was reduced in order to increase cooling efficiency. The question was: how much can the power of the fans be reduced while staying within the given temperature limits? The temperature limits are 50 °C for the wafer and 60 °C for the FPGAs. As the three centered fans already ran at full speed and the temperature difference between inlet and wafer is already about 25 K, they were not slowed down at all.

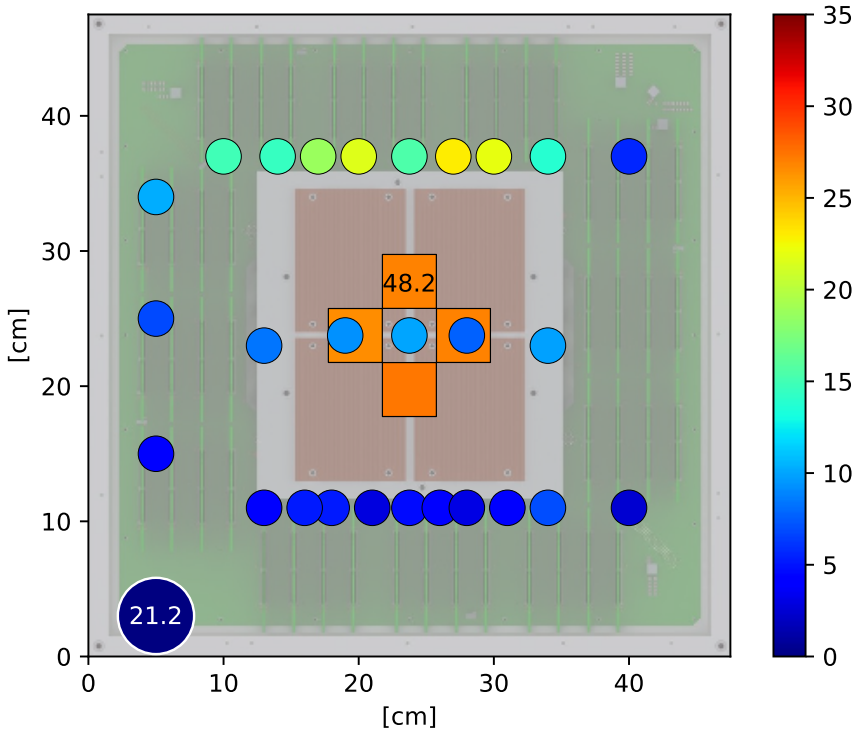

Figure 4.11: Heat map of the system with duct and four outer fans turned down to 42 % of the maximum fan RPM. Plotted are temperatures relative to the temperature of the reference sensor on the bottom left marked by a larger circle. The reference temperature is written into the circle.

First the speed of the three fans cooling the power supply was not changed. The two outer most fans were reduced step by step until the FPGAs were about to reach 60 °C. This was the case when the fan speed was about 50 % of the maximum RPM of 13000. The power consumption for this fan configuration is 151 W ± 3 W, which is a power reduction of $33 \pm 5W$ compared to section 4.7.

Then the four outer most fans were turned down to 42 %, where all FPGAs remain

below 55 °C. The power consumption is 127 W ± 3 W, which is a reduction of 58 W ± 5 W. Any further slowing down of these fans would lead to an increase in FPGA temperatures beyond 60 °C, but almost no reduction of power: At a total power consumption of 122 W ± 3 W the FPGAs are above 65 °C, the power consumption is 5 W ± 3 W lower than before.

As the power consumption for four slowed down fans is lower than that of two slowed down fans, while maintaining a lower FPGA temperature, one can conclude: Two fans of type A are more efficient than one.

4.9 Testing Fan b

For the purpose of testing two fans of type B, two fans on the top had to be removed, as fan B has two power connectors due to its two built-in motors. They were mounted besides the centered fan. They deliver a higher static pressure and are able to reduce the wafer temperatures by another 3 K compared to 4.7. Results can be seen in 4.12. The temperature difference between inlet and wafer is about $23K$.

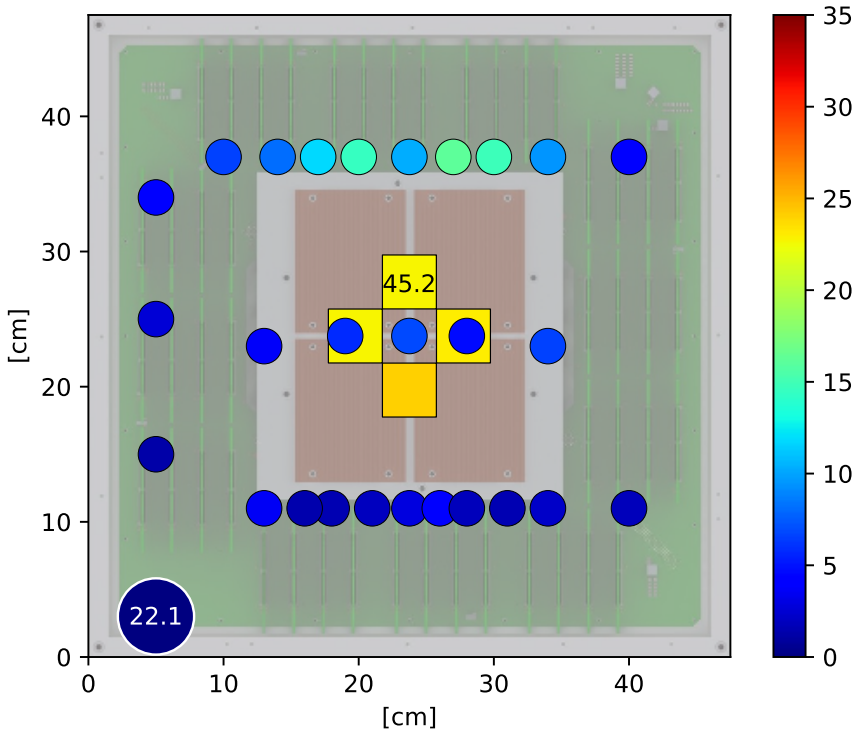

Figure 4.12: Heat map of the system in thermal equilibrium, see fig. 4.2, with duct and two fans of type B besides the centered fan.

These fans consume noticeably more power: 10 pieces of fan a needed 184 W ± 4 W,or 18.4 W ± 0.4 W per fan. 6 pieces of fan A and 2 pieces of fan B need 205 W ± 4 W, or 47 W ± 1 W for one fan of type B and not 37 W as given in the data sheet. The higher power consumption of fan B might be result of the increased voltage of 12.6 V ± 0.1 V of the power supply versus the rated 12 V.

4.10 Comparison of Fan A and Fan B

As fan B needs more than twice the power than fan A, the goal was to speed it down to the same power as fan B and compare the wafer temperatures. If the temperatures are the same, then the efficiency of both fans is equal.

As there are only 6 fans of type A and 2 fans of type B, a reference measurement of the power for 8 fans of type A has to be done. Two of them will then be replaced by two fans of type A.

The reference power with 8 fans of type a is $146\,\mathrm{W} \pm 3\,\mathrm{W}$. The two fans of type B had to be set to about 50 % RPM of the maximum RPM of 17000 to reach this power. The wafer temperatures are 2 K lower for fan B compared to fan A, which is therefore more efficient. This behaviour is expected, because these counter rotating fans can deliver a higher static pressure needed for the reduced cross section of duct and high fin density.

As fan B is more efficient than fan A, installing 3 fans of type B would be reasonable. But at least one more fan a has to be omitted to be able to connect it to the fan board.

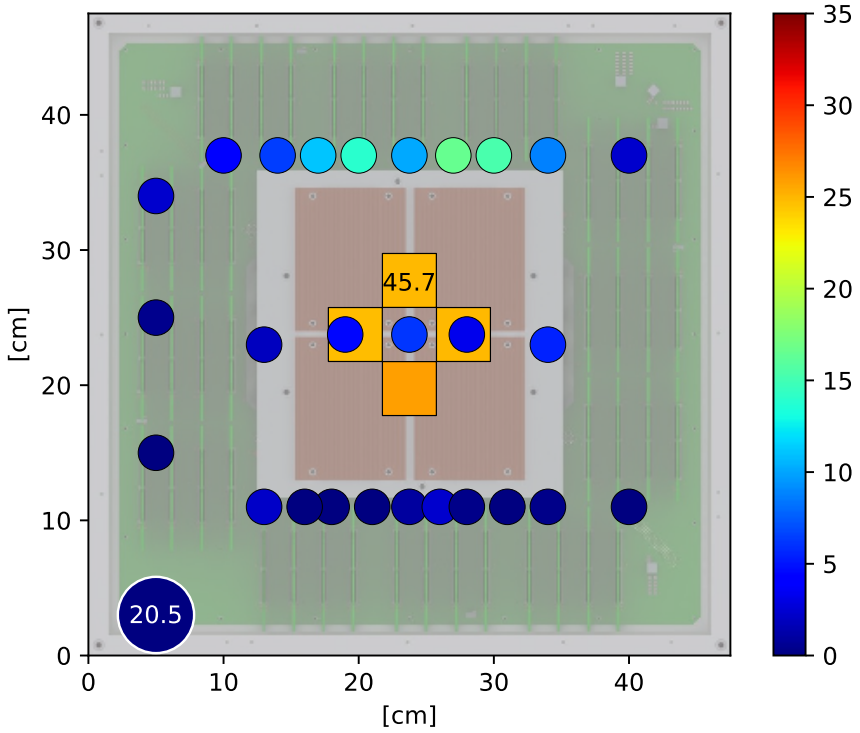

Figure 4.13: Heatmap of the system with two slowed down fans of type B besides the centered fan.

Heat sink 1 should now be at least close to saturation. To proof it, a test with three

fans of type B was done. This test led to another decrease of the wafer temperatures by 1 K compared to two fans of type B. Because increasing the static pressure and therefore increasing the air flow through heat sink 1 only leads to minor changes in temperature, saturation of heat sink 1 is reached.

The thermal resistance of the heat sinks can be calculated with a few assumptions. If all four heat sinks 1 contributes in the same way to the total thermal resistance, the temperature difference between inlet and wafer is 25 K and the power of the wafer is half the power consumed by the system, then the thermal resistance for one heat sink 1 is:

$$R = \frac{25\,\mathrm{K}}{624\,\mathrm{W}} \times 4 = 0.16\frac{\mathrm{K}}{\mathrm{W}} \tag{4.1}$$

Errors are not calculated as this calculation is meant to be an estimation. While the error of the temperature is known, it is unclear, if exactly half of the total energy for the system (without fans) is spent on the reticles. It is also not certain that all four heat sinks contribute to the thermal resistance in the same way. This estimation shows that heat sink 1 is in saturation.

4.11 3D printed Duct

For easier installation two parts were 3D printed to replace the duct made of cardboard, see figure 4.14. It is currently made from pla [17], which seems to withstand the temperature, although it has a glass transition temperature of 60 °C. It might have to be replaced, if it wears to fast. One part is the main duct which covers the fins of heat sink 1, one part is the transition to the first part and widens to the fans. Both parts are currently taped to the aluminium plate. The transition part should be modified to be able to screw it to a Wio.

Figure 4.14: 3D printed duct installed into the system. The southern sensor row can be seen in front of it.

The results of replacing the cardboard duct with the 3D printed duct can be seen in fig 4.15. The change between the cardboard duct and the printed part regarding temperatures is about 1 K, see fig 4.12, even though the printed duct is not extended to the fans, but to the row of FPGAs in the south only.

With fans B slowed down, the temperature difference between cardboard duct, see figure 4.10, and printed duct is neglectable with 1 K.

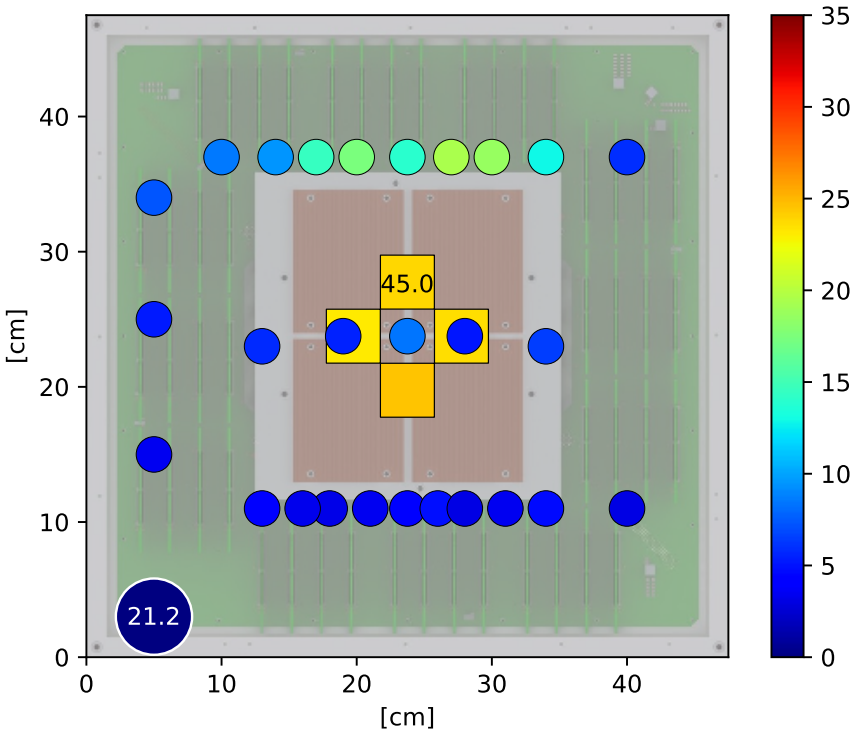

Figure 4.15: Heat map of the system with two fans of type B and a 3D printed duct.

5 Evaluation

The results of chapter 4 can be summarized in one bar graph, see figure 5.1). The bars are labelled like the sections in chapter 4. Additionally, an optimization was included, where the four outer fans were slowed down to 42 % and two fans of type B were installed. This experiment is shown as Fan B optimized. The cardboard duct was included from the measurement 'Cardboard duct on', or was replaced by the 3D printed duct, which is labelled explicitly. To compare the efficiency of fan A and fan B, see section 4.10 the reference 'Fan A comparison' with 8 fans of type A was included. It has to be compared to 'Fan B comparison'. Two measurements with 3 fans of type B are included as well, the first measurement with the cardboard duct, the second one with the 3D printed duct. The last measurement ' " optimized' is the same setup as in the previous measurement '2 fans B 3D printed duct', but all fans were slowed down so that the total power for the cooling is 70 W. This is less than half the power of the current state. The total errors mainly consist of the temperature error of 3 K caused by the air conditioning. The error of the voltage and current read out makes up a small part of the total error and can be neglected for this comparison.

The grey bar shows the temperature difference between the wafer temperatures and the temperature of the inlet. The sensor close to the inlet was taken to measure the reference temperature, the temperature of the four wafer sensors was averaged. The temperature difference between wafer and inlet could be reduced to 22.6 K with the cardboard duct and 3 fans of type B. The 3D printed duct is with 23.0 K only 0.4 K higher. This is a reduction of 8.3 K compared to the current state and a reduction of 9.8 K if the effect of the vertical mounting is included.

$$R = \frac{\Delta T_{Wafer,Inlet}}{P_{48\mathrm{V}}} \tag{5.1}$$

$$P(\Delta T = 25\,\mathrm{K}) = \frac{25\,\mathrm{K}}{R} \tag{5.2}$$

$$PUE = 1 + \frac{P_{cooling}}{P(\Delta T = 25\,\mathrm{K})} \tag{5.3}$$

The green bar contains information about cooling efficiency at the desired temperatures of 50 °C and 60 °C for wafer and FPGAs respectively. Different improvements were done to increase efficiency. But they do not result in the same temperature difference between wafer and inlet. They have to be normalized by means of equation 2.1: For a given configuration of heat sinks, fans and ducts it is possible to calculate the thermal

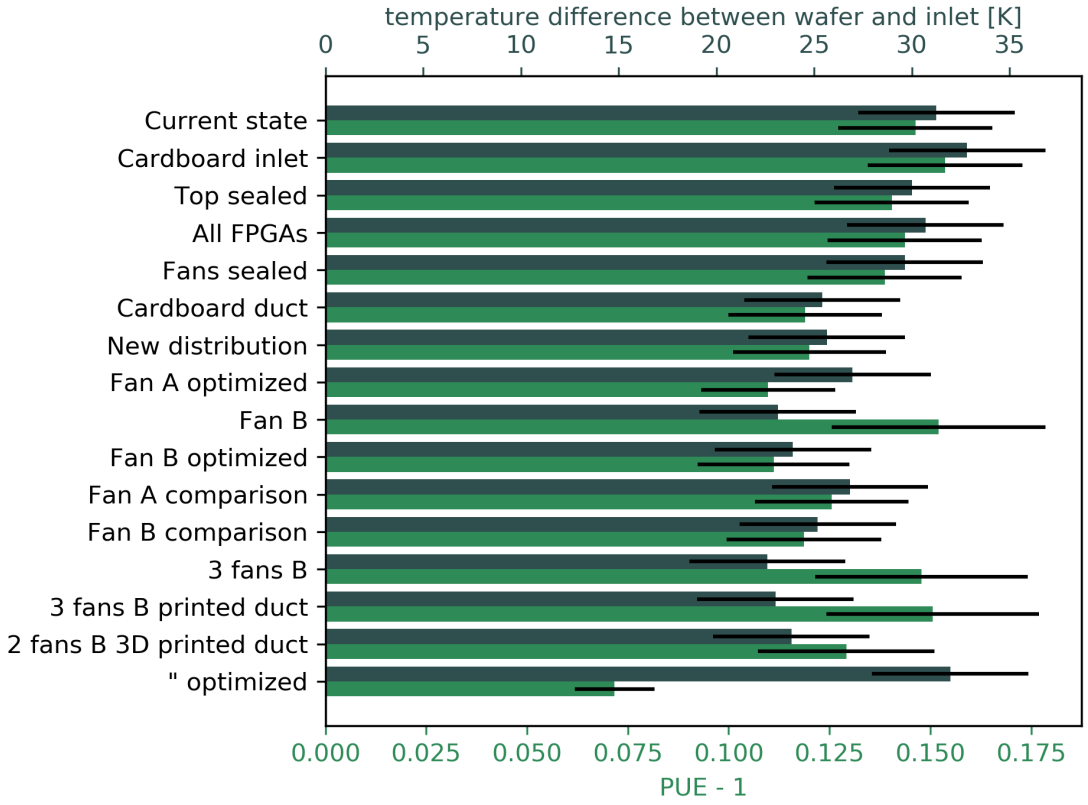

Figure 5.1: PUE and temperature difference between inlet and wafer of all experiments.

resistance with the measured power of the 48 V input. One can now calculate the theoretical power consumption of the wafer for a desired temperature difference of 25 K with this thermal resistance. This power consumption has to be taken for $P_{48\mathrm{V}}$ to calculate the PUE, see the following equations 5.1. These PUEs can now be compared. A lower PUE means higher efficiency.

The $PUE = 1.07 \pm 0.02$ for the optimization with fan B is the lowest of all configurations. All fans were turned down to a speed of about 50 %.

A good compromise between PUE and ΔT is the setting with a duct, 2 fans of type B and four slowed down outer fans. There are two versions, labelled 'Fan B optimized' and '2 fans B 3D printed duct'. The temperature differences between wafer and inlet are below 24 K and a PUE below 1.13. Also if the wafer runs too cold, the fan speed can be reduced, resulting in a temperature difference of 32 K and a PUE of 1.07.

6 Summary and Outlook

Different configurations of fans were tested on the BrainScaleS system, a duct was installed into the system and parts of the system were sealed in order to minimize the temperature difference between the air inlet and the wafer. Sealing the space between Wios led to an increase of efficiency and a decrease of wafer temperature by 1 K and sealing the fans also reduced the wafer temperature by 1 K. The four outer fans can be reduced to below 50 % of the maximum fan speed without heating the FPGAs above 60 °C to decrease the cooling power. A duct made of cardboard reduced the wafer temperature by 4 K, a 3D printed duct by 3 K. Fan B was able to reduce the wafer temperature by another 3 K. Also, it is more efficient than fan A, but the optimum between temperature and power consumption with these fans was not figured out yet. Heat sink 1 is already in the saturation regime with 2 fans of type B, which is why a third fan B reduces the wafer temperature only by 1 K more.

With all improvements included, the temperature difference between wafer and inlet could be reduced by 8.3 K compared to the current state by 9.8 K if the effect of the vertical mounting is included.

The 3D printed duct can still be improved to get closer to the cooling performance of the cardboard duct. It also has to be modified to be able to install it easily. The printed fan sealing also needs to be improved so that the inner fan grilles can be mounted.

Fan B should be used for future experiments with the selected heat sink 2 for increased efficiency. Heat sink 2 is expected to perform better than heat sink 1 and reduce power consumption of the cooling further.

The installation of a liquid cooling should also be considered. Due to the higher thermal capacity of this liquid and its higher density, a higher heat exchange could be reached. Lowering the thermal resistance by a factor of 5 - depending on the liquid cold plate - are reachable. The used liquid can be coupled to a heat exchanger with larger heat sinks, that would not fit into the BrainScaleS system, without the need of any fans. This would be the state of the art concept for reducing power consumption of the cooling and increasing its efficiency.

Although the concept of liquid cooling has advantages, its realization is more complex. Preparations have to be made to keep the liquid in the cycle and to prevent liquid reaching the electronics. That could be topic of another bachelor thesis.

For the current setup I recommend the installation of the 3D printed duct, as well as the 3D printed fan sealing, as soon as they are improved. Parts could also be 3D printed to seal the gaps between Wios. Two fans of type B should be used for higher efficiency and a good performance.

7 Appendix

7.1 Specifications

Fan	Static pressure	Max. air flow	Power
Delta Electronics AFC0612DE-AF00	433 Pa	$1.87\,\mathrm{m^3\,min^{-1}}$	18 W
San Ace 9CRA0612P6K001	1100 Pa	$2.3\,\mathrm{m^3\,min^{-1}}$	37 W

Table 7.1: Specifications of the used fans.

Heat sink	TDP	Principle	Weight	Height
Dyantron T318	165 W	Vapor chamber base	435 g	1 U
Dynatron B8	205 W	Heat pipes	600 g	2 U

Table 7.2: Specifications of the used heat sinks.

Sensor	Accuracy
Maxim ds18b20	0.5 K

Table 7.3: Specifications of the used sensors.

7.2 Reticle Distributions

Configuration	Switched on reticles
Initial distribution	0, 2, 4, 5, 8, 14, 16, 21, 23, 24, 25, 26, 27, 29, 31, 33, 34, 35, 36, 41, 43, 44
New distribution	0, 2, 4, 5, 6, 8, 14, 16, 21, 23, 25, 28, 28, 29, 31, 33, 34, 35, 36, 38, 41, 44

Table 7.4: Two different configurations of switched on reticles.

List of Abbreviations

CFM	Cubic Feet per Minute
FPGA	Field Programmable Gate Array
Main PCB	Main Printed Circuit Board
PUE	Power Usage Effectiveness
RPM	Rotations Per Minute
SWAG	Squishy Washer-Alike Grille
TDP	Thermal Design Power
Wio	Wafer I/O PCB

Bibliography

[1] Johannes Schemmel et al. An accelerated analog neuromorphic hardwaresystem emulating nmda- and calcium-basednon-linear dendrites. *https://arxiv.org/pdf/1703.07286.pdf*, 2017.

[2] Hicann. https://www.kip.uni-heidelberg.de/vision/previous-projects/facets/neuromorphic-hardware/waferscale-integration-system/hicann/.

[3] Johannes Schemmel et al. A wafer-scale neuromorphic hardware system for large-scale neural modeling. 2010.

[4] European institue for neuromorphic computing. http://www.vba-mannheim-und-heidelberg.de/pb/,Lde/Startseite/Ueber+uns/Neubau+European+Institute+for+Neuromorphic+Computing+_EINC_+INF245.

[5] Navier-stokes equation. https://www.math.hu-berlin.de/~jwolf/web/NSE-ss2014.pdf.

[6] Specific heat capacities of air. https://www.ohio.edu/mechanical/thermo/property_tables/air/air_cp_cv.html.

[7] Specific heat capacity of water. https://en.wikipedia.org/wiki/Specific_heat_capacity.

[8] heat sink 1. Dynatron t318. https://www.dynatron.co/product-page/t318.

[9] fan a. Delta electronics afc0612de-af00. https://www.delta-fan.com/Download/Spec/AFC0612DE-AF00.pdf.

[10] heat sink 2. Dynatron b8. https://www.dynatron.co/product-page/b8.

[11] fan b. San ace 9cra0612p6k001. https://sda.sanyodenki.us/data/cooling/catalog/Counter_Rotating_Fan.pdf.

[12] Maxim. ds18b20. https://datasheets.maximintegrated.com/en/ds/DS18B20.pdf.

[13] Current clamp. http://storage-download.googleapis.com/spec/MS2102.pdf.

[14] Air conditioning. https://www.breeze24.com/media/pdf/c2/0e/b8/BRC073_Daikin-IR-FB-databook.pdf.

[15] Innofil. Innoflex. http://blog.innofil3d.com/wp-content/uploads/2015/05/Innofil3D-Technical-Data-Sheet-Innoflex-45-131201.pdf.

[16] Shore scale. https://wiki.polymerservice-merseburg.de/index.php/SHORE-H%C3%A4rte.

[17] Innofil. Pla. https://www.innofil3d.com/wp-content/uploads/2016/05/TDS-Innofil3D-PLA-160608.pdf.

Statement of Originality (Erklärung):

I certify that this thesis, and the research to which it refers, are the product of my own work. Any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Ich versichere, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg, June 27, 2019

.....
(signature)