Direct Power Semiconductor Cooling – Potentials, Challenges and Development Approaches

Benjamin Pessl [1)] Gregor Gaßner [2)] Hannes Hofstetter [2)] Michael Haider-Peterseil[2)]

1) Magna Powertrain / Engineering Center Steyr, St. Valentin, Austria

E-Mail: benjamin.pessl@magna.com

2) Magna Powertrain / Engineering Center Steyr, St. Valentin, Austria

E-Mails: gregor.gassner@magna.com, hannes.hofstetter@magna.com, michael.haider-peterseil@magna.com

ABSTRACT: Automotive traction inverters contain power semiconductors with high electrical power capability. Even though the overall efficiency is being improved continuously, the absolute level of waste heat is increasing, due to concurrently rising power levels. This circumstance results in demanding cooling needs and approaches the limits of typical indirectly cooled solutions using thermal interface materials (TIM) or thermal pads. Direct cooling – whereby the coolant is in direct contact with a low-thermal-resistance power-device-mounted heat sink – extends the solution space and enables to effectively address the described limitations. This paper provides a comprehensive overview of challenges to be dealt with during the development of direct cooling solutions for power semiconductors. Detailed insights are presented for thermal and hydraulic optimization, posing the foundation for electrically and energetically performant solutions. It is shown how the utilization of advanced modelling and simulation methods can enhance development speed and helps exploiting technologies to their limits. Complemented by insights into integration-related aspects like geometrical tolerances, corrosion protection and reliability, the paper examines some of the main subjects relevant during development. It concludes with a summary of interactions between the presented influences.

KEY WORDS: Semiconductor cooling, wavy fin, pin fin, direct cooling, thermal management, heat sink optimization, CFD-simulation

1. INTRODUCTION

As vehicle electrification moves forward, the challenges faced in electric drivetrains are steadily increasing. By moving from silicon-based to silicon carbide-based semiconductors for power conversion, the inverter efficiency can be increased (1). However, due to simultaneously soaring power levels, the cooling demand for operating power semiconductors are rising accordingly.

In this paper, we investigate the potentials of direct power semiconductor cooling, using advanced simulation methods. With this approach we can drastically reduce the required time for assessing a multitude of design variants and prevent several expensive hardware loops.

2. MOTIVATION FOR DIRECT COOLING

State-of-the-art cooling solutions comprise a cooling channel, encasing a cooling medium, the power semiconductor devices, and a thermal interface material (TIM) in between (1). By permanently exerting homogeneous pressure on the assembly, the TIM ensures a thermally acceptable interface from power source to the cooling channel. With rising waste heat, such solutions approach their thermal limits, requiring more performant cooling architectures.

One of the most promising approaches addressing this phenomenon is *direct cooling* (2). There is no uniform definition available for this term. In this paper, the term direct cooling is used to describe a cooling architecture, wherein:

- a heat sink is permanently attached to the surface of the heat source (e.g., by soldering or sintering), and
- the coolant is in direct contact with this heat sink.

In Fig 1, the difference between indirect and direct cooling is described.

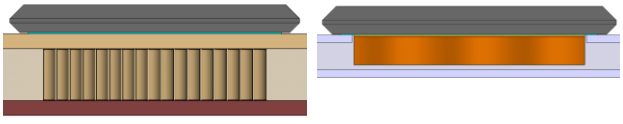


Fig. 1 Indirect cooling using pin fin heat sink (left) vs. direct cooling using a fin structure attached to the power module (right).

Compared to indirect cooling, the thermal path is improved with direct cooling. This is mainly achieved by:

- reducing the contact resistances on the two sides of the TIM

- increasing the joining material heat conductivity (solder / sinter layer instead of TIM)
- decreasing the involved wall thicknesses through which the heat needs to be conducted.

The combined effects result in a decreased thermal resistance from junction to fluid ($R_{thJ\text{-}F}$), yielding lower chip temperatures at constant power losses. In the next chapters, these effects are investigated.

3. HEAT SINK GEOMETRIES AND PLACEMENT

A variety of heat sink architectures (geometries) is available in industry (3). Three of the most relevant ones are pin fins (cylindrical solids protruding from a base plate, towering into the coolant flow), lanced fins (thin sheet-metal-based structures incorporating intermittent stampings) and the folded wavy fins (thin sheet-metal-based structures with defined geometrical properties). These architectures are represented in Fig. 2 and Fig. 3, respectively, whereby they are integrated into generic cooling channel geometries.

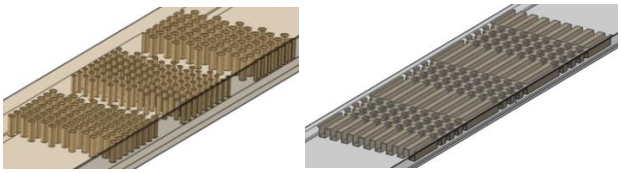


Fig. 2 Pin fin assembly (left) vs. lanced fin assembly (right)

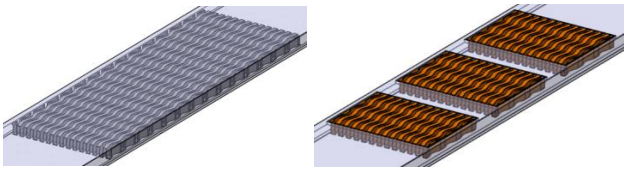


Fig. 3 Wavy fin assembly (left) vs. Direct wavy fin assembly (right)

All heat sink architectures have their specific properties making them suitable for different purposes. However, besides others, their common aim is providing sufficient surface area for the heat transfer into the coolant. In this context, the placement over the heat sources plays a crucial role. The more cooling surface can be placed directly below the heat source (semiconductor chips), the more effectively heat is removed from there.
In Fig. 4, the following aspects are illustrated:

- Pin fins typically come with larger dimensions (diameter) that folded structures. At a given chip size, this results in lower active cooling surface, taking the same height as a basis.
- By shifting the heat sink in the horizontal plane, the active cooling surface is directly affected. Pin fins are particularly sensitive towards this effect, due to their large diameters.

a *b* *c*

Fig. 4 Schematic illustration depicting a power module containing semiconductor chips, a pin fin structure (a), a wavy fin structure (b) and a lanced fin structure (c).

Compared to the most common heat sink architecture used for indirect cooling – the pin fin, the wavy fin architecture offers more surface area below the heat source, enabling a better cooling performance.
From cooling performance perspective, the lanced fins lie in between. Like wavy fins, they offer a large cooling surface area. In contrast, however, the coolant intermittently collides perpendicularly with the structure's front sides, leading to higher pressure drops.

4. SIMULATION MODEL DESCRIPTION

Traditionally, simulation-based workflows comprise an initial setup that is defined. After a first simulation loop, the results are manually analyzed and the set of parameters for the next iteration is derived. The described process repeats itself until the target corridor is reached. This time-consuming process can be drastically sped up by implementing automatic operating optimizer tools. By doing so, a large set of results can be generated in a short amount of time, enabling the engineers to focus on the analysis and interpretation without the need to regularly define new parameter sets and start simulation loops manually.
In this paper, we make use of the above-described approach. The 3D-CAD geometry of the cooling channel is used to perform a numerical simulation based on the continuity equation (1), in which ρ represents the coolants density, ϕ the specific term of the fluid flow, $\underline{u}$ the velocity and T the temperature (8).

$$\frac{\partial}{\partial t}(\rho\phi) + \nabla \cdot \left(\rho \underline{u} \phi\right) = \nabla \cdot (\Gamma \nabla \phi) + Q_\phi \qquad (1)$$

The 3D-model is automatically generated by the CAD tool. This is where the so-called “SHERPA” algorithm comes into action. By looped parameter definition, calculation within the CFD simulation tool (4), and comparison with set boundary conditions, it generates multi-objective Pareto optimization studies. In this process, global boundary conditions (coolant flow rate, ambient temperature etc.) are considered.

During each cycle, SHERPA refines the trade-off front by adding new data points. Each point represents one fully evaluated design.

5. HEAT SINK OPTIMIZATION

For efficiently and effectively analyzing the potential of a particular cooling architecture it is vital to follow a purposive workflow.

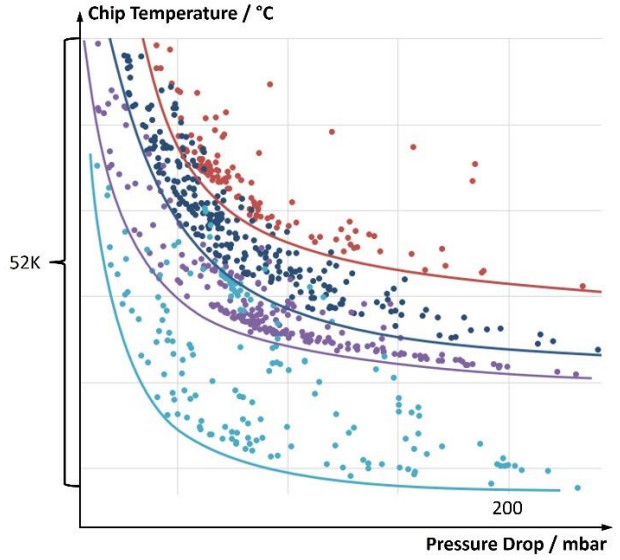


Fig. 5 Representation of calculated design variants and resulting Pareto fronts for pin fin (red), lanced fin (dark blue), wavy fin (purple) and wavy fin direct cooling (light blue)

The selected approach comprises an interaction of 3D-CAD software with CFD-simulation tools and an optimizer. First, a fully parameterized 3D-model of the entire heat sink and cooling channel is created. Second, the optimizer accesses this model, creating numerous design variants by parameter variation. This is done within pre-defined parameter ranges, considering boundary conditions like manufacturability or physical availability (e.g., discretization of the sheet metal thickness). Third, these design variants are processed within a computational fluid dynamics (CFD) simulation tool, yielding temperature distribution and pressure drop. Fourth, the gained results are plotted in a diagram, whereby the abscissa represents the pressure drop and the ordinate the maximum temperature across all the heat sources (semiconductor chips). Each design variant is represented by one data point in this graph. With increasing number of data points, an enveloping curve emerges along the left-hand side of the point cloud (5). This so-called *Pareto front* represents the achievable limits of the investigated technology – seen in Fig 5. Thereby, we were able not only to compare several singular designs, but also to identify each investigated technologies’ thermal and hydraulic characteristics. The following deductions could be made from the studies’ results exemplarily depicted in Fig. 5:

- With pin fin architecture (indirect cooling), the maximum chip temperature is the highest of the investigated technologies. It was found, however, that this architecture shows the highest gradient in the Pareto fronts high pressure drop regions. Thereby, the achievable thermal benefit with increasing pressure drop is the highest of the investigated technologies.
- Lanced fins (indirect cooling) show an increased thermal performance at constant pressure drop. This behavior is achieved by an increased active cooling area and advantageous flow conditions.
- With wavy fins (indirect cooling), the thermal performance can be improved once again. Due to the architecture’s hydrodynamic and thermal properties, this can be already achieved at significantly lower pressure drop levels, compared to pin fins, for example.
- Direct cooled wavy fins show a remarkable gain in thermal performance. This is true especially at low pressure drop regions. By minimizing the overall thermal resistance from junction to fluid and providing a high active cooling surface area, this architecture holds significant potential for chip temperature decrease and, equally, for semiconductor chip downsizing.

5.2. Variables

The parameterized 3D-models include several variables accessible to the optimizer for modification (6). In Fig. 6, Fig. 7, and Fig. 8, some of them are depicted for the respective technologies.

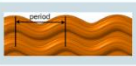
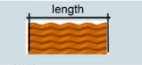
Nb	Parameter	Overview	Nb	Parameter	Overview
1	Sheet metal thickness		5	Period of sinus curve	
2	Height of folded fins		6	Cooling channel gap	
3	Width of folded fins		7	Total length of folded fins	
4	Amplitude of sinus curve		8	Gap to Cooling Channel Wall	

Fig. 6 Excerpt of parameters used for optimization of wavy fins.

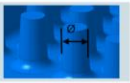
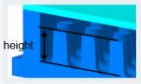
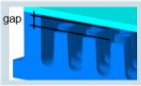
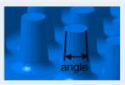
Nb	Parameter	Overview	Nb	Parameter	Overview
1	Diameter of pins		5	x-distance	
2	Height of pins		6	y-distance	
3	Pin gap		7	Min. gap between pins	
4	Draft angle		8		

Fig. 7 Excerpt of parameters used for optimization of lanced fins.

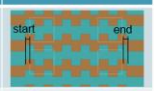
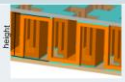
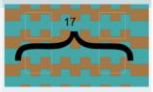
Nb	Parameter	Overview	Nb	Parameter	Overview
1	Sheet metal thickness		5	start/end Folded fins	
2	Height of folded fins		6	Offset value	
3	Width of folded fins				
4	Amount of lanced fins				

Fig. 8 Excerpt of parameters used for optimization of pin fins.

5.1. Boundary Conditions

Each parameter in the 3D-model can automatically be modified by the optimizer within a pre-defined range. This range takes different aspects into account, narrowing down the investigation space to practically feasible designs. As an example, the wall distance between wavy fins must not be lower than 1 mm, allowing particle contaminants to pass through without clogging the heat sink. Moreover, there are constraints like permitted maximum pressure drop or a permitted maximum temperature level, towards which optimization is to be carried out.

For a representative comparison, a set of boundary conditions for the coolant volume flow and the thermal loss of the power module was defined and used in all the executed simulations. These conditions are 10 l/min at 65°C coolant temperature and 1200 W of power loss per module. These values have been defined in accordance with typical ranges within automotive traction applications.

Typically, the permissible pressure drop within an automotive inverter is in the range between 90 to 250 mbar, with some exceptions. Maximum permissible temperature levels are higher in case of SiC MOSFETs, compared to Si IGBTs. Including margins, the level is in the range between 150 and 175 °C, in some cases slightly higher.

6. SENSITIVITY ANALYSIS

During a sensitivity analysis, we fundamentally examined wavy-fin-based direct cooling architectures for the influence of various manufacturing and assembly tolerances on the resulting change in performance (7). Not only did we consider each tolerance separately. Instead, we investigated the interactions of different tolerances, using the parameterized models. Fig. 8 schematically shows the superimposed effects different parameters have on the results.

6.1. Tolerances

During assembly of such a direct cooling solution, mechanical tolerances need to be considered. Both attaching the heat sink to the power module, and attaching the power module to the cooling channel, are characterized by such. Similarly, the involved parts themselves show a manufacturing-prone variation in their dimensions.

For this reason, the tolerance chain was investigated within a sensitivity analysis, yielding insights on the expected influence on thermal and hydraulic performance. Fig. 9 depicts an example of a variation of the lanced fins focusing on three different variables: sheet-metal thickness, width, and height:

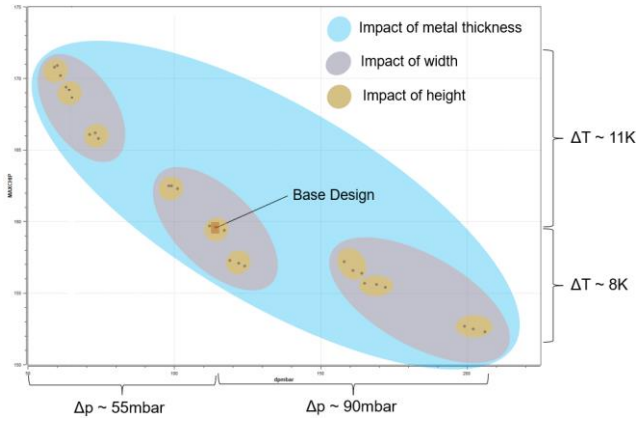


Fig. 9 Superimposed effects of different parameters, with Δp (pressure drop) and ΔT (temperature difference).

In practice, two of the three variables are fixed – in this case the height and the width – and the thickness is varied to the minimum and maximum tolerance limits.

At the example of the lanced fin architecture, it becomes obvious, that the sheet-metal thickness has a high influence on the thermal behavior – see Fig. 10. Consequently, special attention needs to be paid to this dimension during the design process, restricting its tolerances to a sufficiently narrow band, preventing undesired thermal behavior from occurring.

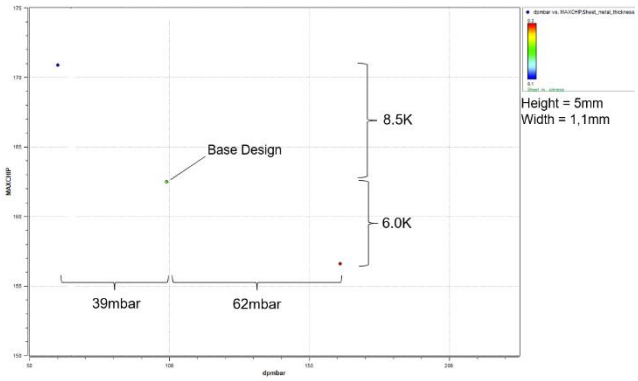


Fig. 10 Influence of the sheet-metal thickness on pressure drop and max. chip temperature (lanced fin architecture).

The second priority in terms of tolerances in this example is the width between the lanced fins. It still exhibits a noticeable impact on the hydraulic and thermal behavior, but already to a lower extent compared to the sheet-metal thickness, as seen in Fig. 11:

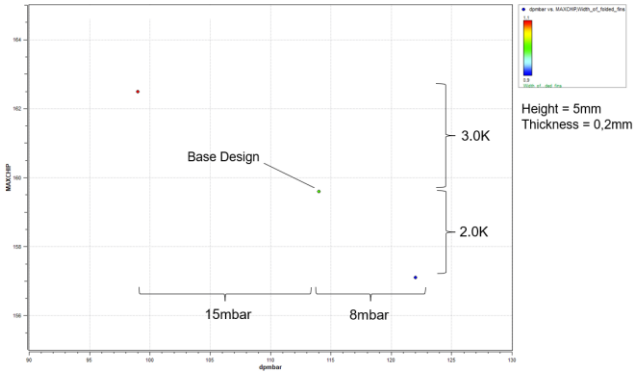


Fig. 11 Influence of the width between the lanced fins on pressure drop and max. chip temperature.

The least sensitive parameter is the height, which only slightly influences the behavior compared to the other two parameters and their tolerances – as seen in Fig. 12.

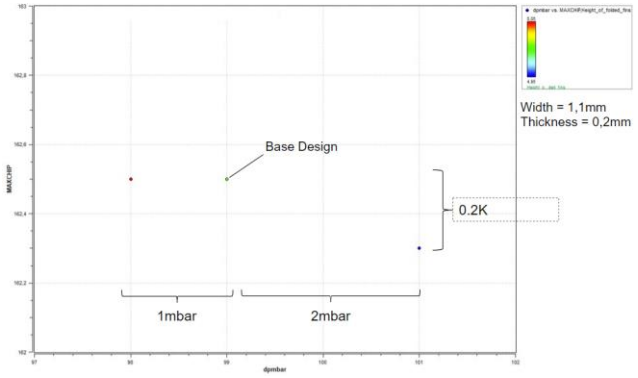


Fig. 12 Influence of the lanced fin height on pressure drop and max. chip temperature.

6.2. Pressure Drop

There is an interdependence of pressure drop and maximum chip temperature. The higher the coolant flow velocity, the better the thermal performance and the worse the hydraulic performance will be. Moreover, the coolant guiding effects of the cooler assembly play a crucial role. It needs to be made sure that most of the coolant passes through the heat sink, with a flow velocity in the favorable range of 1 to 2 m/s to obtain an adequate balance between pressure drop and chip temperature. Leakage flow rates between the heat sink and the cooling channel wall need to be minimized, since their contribution to heat removal is marginal.

As can be seen in the Pareto diagram (Fig. 5), the wavy fin heat sink architecture for direct cooling is particularly effective in the lower pressure drop range. With a moderate increase in pressure drop, a significant decrease in chip temperature can be obtained. This effect saturates with increasing pressure drop, such that thermally, there is only a minor benefit expectable.

For higher pressure drops, heat sink architectures like the pin fin, offer a higher optimization potential. However, pressure drop comes at the cost of energetical effort (operation of the coolant pump).

6.3. MATERIAL INFLUENCE

One of the main tasks in engineering is finding the right balance between costs and benefit. Using different materials can be an appropriate approach to that. Therefore, two heat sink materials were investigated and compared in their performance. It was shown that using aluminum instead of copper increased the maximum chip temperature by 9 K, as seen in Fig. 13.

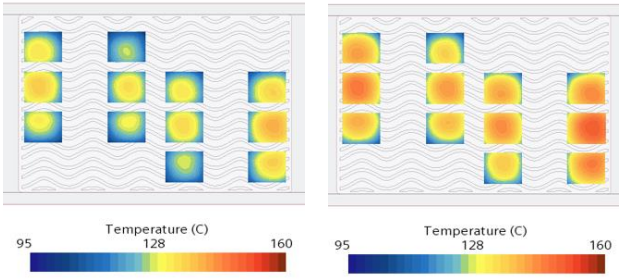


Fig. 13 Temperature distribution for copper (left) and aluminum (right) wavy fin heat sinks in direct cooling.

7. VALIDATION / MEASUREMENTS

The simulation models used in this study have been created following the example of existing, previously validated models. Due to their broad similarity in relevant aspects, the models are well-suited for the presented study using a generic power module setup.

Validation is always done for concrete hardware to calibrate the remaining uncertainties. This process fundamentally involves the following steps:

- Hardware procurement and assembly

- Test bench operation / measurement
- CFD-simulation of the concrete assembly
- Comparison of measurement and simulation data
- Loops of adjusting the heat transfer coefficient in the model until measured and simulated temperatures match.

In Fig. 14 and Fig. 15, respectively, exemplary validation cases are depicted. By continuously enhancing the pool of available, validated simulation models, the overall confidence level can be enhanced. This allows to perform meaningful simulation studies even before concrete hardware is available.

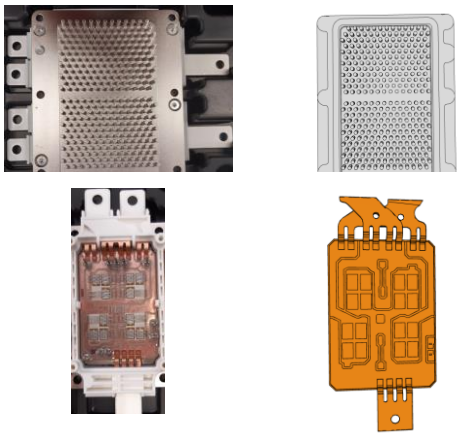


Fig. 14 Hardware and CAD geometry for a validation case

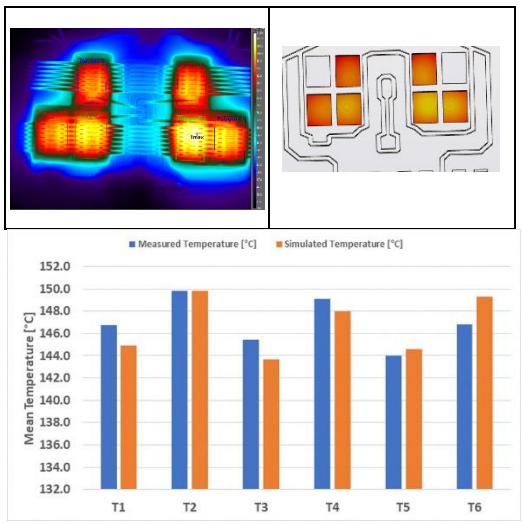


Fig. 15 Exemplary validation of measured and simulated temperature.

8. CHALLENGES AND INTERACTIONS

Besides the thermal and hydraulic performance, there is a multitude of additional challenges to be dealt with in case of direct cooling. All the following aspects need to be considered during development:

- Reliability of the heat sink attach layer
- Reduced temperature spread from chips to heat sink
- Coefficient of thermal expansion (CTE) mismatch
- Material compatibility with the coolant
- Contamination of the heat sink / clogging
- Corrosion / corrosion protection / surface coating
- Abrasion through particles in the cooling loop
- Local surface temperature / boiling effects
- Cooling channel integration

Many of the abovementioned aspects have strong interdependencies, requiring a systematic approach not only on component-, but on subsystem-, system- and in some cases even vehicle-level. Some of the aspects are represented in Fig. 16, showing exemplary domains of interaction.

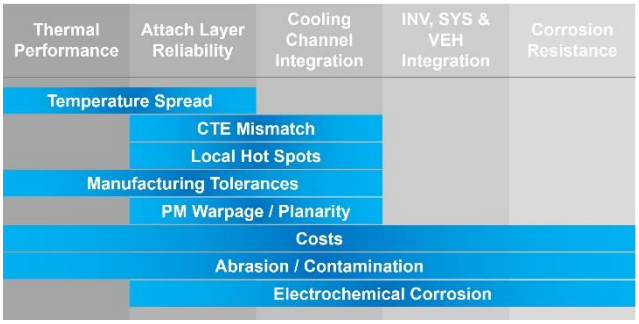


Fig. 16 Selected aspects relevant for direct cooling applications, including their domains of interactions.

We know how to address them in concrete projects, working with partners from industry and academia.

9. CONCLUSION

Direct cooling using wavy fin heat sinks proves to offer thermal benefits – already at low pressure drops, compared with indirect cooling using pin fins, or the slightly better-performing lanced fins (indirect cooling). In some cases, the maximum chip temperature can be decreased by 20 K or more, resulting in the potential to either downsize the semiconductor, or to increase the current output at constant chip size.

Pin fin-based solutions for indirect cooling show a potential at higher pressure drops beyond 200 mbar. However, it does not attain the cooling performance of the direct cooled wavy fins.

From a thermal point of view, the lanced fins sit well in between the two architectures, forming a hybrid between them. Even though not reaching the thermal performance of direct cooled wavy fins, this architecture offers the advantage of a closed cooling channel without the need for the integration using sealings.

The presented simulation-based approach enabled a considerable reduction in investigation time and associated costs, including

hardware. By utilizing an optimization algorithm in conjunction with parameterized 3D-models, a multitude of design variations could be processed in CFD-simulation software. Making use of already-validated simulation models of similar scope, a high confidence in the results could be achieved, such that no additional hardware loops were required during this investigation.

The obtained thermal and hydraulic potentials justify activities aiming for the implementation of (wavy fin-based) direct cooling for power modules. Considering the additional challenges this brings, the overall cost-benefit ratio needs to be considered during dedicated development activities.

REFERENCES

(1) Alexander Stippich, "Exploiting the Full Potential of Silicon Carbide Devices via Optimized Highly-Integrated Power Modules", Institute for Power Electronics and Electrical Drives (ISEA) RWTH Aachen University.

(2) Ekaterina Abramushkina, Assel Zhaksylyk, Thomas Geury, Mohamed El Baghdadi and Omar Hegazy, “A Thorough Review of Cooling Concepts and Thermal Management Techniques for Automotive WBG Inverters: Topology, Technology and Integration Level”, *Energies*, vol. 14, no. 16, 2021.

(3) Naser Sahiti, “Thermal and Fluid Dynamic Performance of Pin Fin Heat Transfer Surfaces”, Faculty of Engineering University of Erlangen-Nürnberg, 2006.

(4) Georgios Mademlis, Raik Orbay, Yujing Liu, Nimananda Sharma, Rickard Arvidsson, Torbjörn Thiringer, “Multidisciplinary cooling design tool for electric vehicle SiC inverters utilizing transient 3D-CFD computations”, eTransportation, vol. 7, 2021.

(5) Moritz Hildemann, Judith Verstegen “Quantifying uncertainty in Pareto fronts arising from spatial data”, *Environmental Modelling & Software*, vol. 141, 2021.

(6) Feng Han, Hong Guo, Xiaofeng Ding, “Design and optimization of a liquid cooled heat sink for a motor inverter in electric vehicles”, *Applied Energy*, vol. 291, 2021.

(7) Joshua E. Aviles, Luis E. Paniagua-Guerra, Bladimir Ramos-Alvaradoa, “Liquid-cooled heat sink design for a multilevel inverter switch with considerations for heat spreading and manufacturability”, *Applied Thermal Engineering*, 2022.

(8) Schwarze R., “CFD-Modellierung”, *Berlin Heidelberg: Springer*, 2013.