Direct-On-Chip Hotspot Targeted Microjet Cooling for Ultra-fast Inference at Scale Running on Groq Language Processing Unit (LPU™)

Feifan Xie[1], Shuhang Lyu[1], Zhi Yang[2], Tiwei Wei[1,*]
[1]School of Mechanical Engineering, Purdue University, West Lafayette, IN 47907
[2]Groq Inc, 400 Castro St #600, Mountain View, CA 94041
*Corresponding email: tiwei@purdue.edu

***Abstract*—This paper investigates the thermal management strategies targeted at hotspot mitigation based on the Groq Language Processing Unit (LPU™) architecture with functionally sliced power map. To address this challenge, this study explores the concept of direct-on-chip hotspots targeted microjet cooling corresponding to the power map through the entire LPU™ chip. The customized heat transfer coefficient map can be achieved by tailoring the inlet nozzle diameter and pitch, as well as the flow conditions to match the hotspots within the chip, or high-power densities. Based on the modeling results, this direct-on-chip microjet cooling can successfully address the potential temperature non-uniformity from different functional blocks within the logic die. It achieves a temperature nonuniformity within 4℃ across the logic die and significantly reduces the thermal coupling between LPU™ and the nearby chiplet in an exploratory use case scenario setting.**

Keywords— Groq LPU™, Chiplet, Inference, Hotspot-targeted cooling, Microjet cooling, Direct-on-chip cooling, Thermal decoupling.

I. INTRODUCTION

Over the past decade, data centers have integrated a myriad of processing cores such as GPUs (Graphics Processing Units), TPUs (Tensor Processing Units), and FPGA (Field-Programmable Gate Array) into warehouse-scale computers (WSCs), thereby creating diverse many-core systems. These systems are equipped with sophisticated IO controllers that facilitate efficient remote direct memory access. In contrast, Groq Language Processing Unit (LPU)™ introduces a streamlined approach to data flow through Stream Programming. It incorporates a large, single-level scratchpad static random-access memory (static RAM or SRAM) that ensures a fixed and deterministic latency. The LPU™ explicitly allocates tensors in both space and time domains, unlocking substantial memory concurrency and offering enhanced computational flexibility across multiple dimensions, encompassing device, hemisphere, memory slice, bank, and address offset [1-3].

The unique, highly structured LPU™ architecture shown in Fig.1, offers great advantage in performance. First, it's fully deterministic in nature, which indicates every data flow direction and progress are predetermined at software level, therefore, spatial power distribution and duration are known at any given time even before the program runs. Secondly, on Groq LPU™ silicon, compute/ memory / IO regions are distinctively positioned, leading to a relative stable/ consistent power distribution trend for different use cases. Compute heavy regions are normally high power density region, which presents great opportunity to integrate hotspot targeted cooling strategy. Each functional block operates at different power levels, which may lead to temperature non-uniformity distribution across the large silicon area of the LPU™, inducing to unbalanced thermomechanical stress and overall out-of-plane deformation. These factors pose significant reliability concerns for the package, necessitating careful management of hotspots across different functional blocks to mitigate thermomechanical stress within the die.

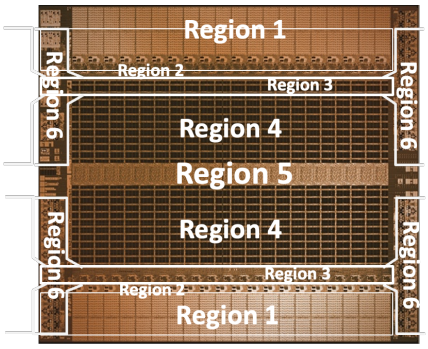

Fig.1: Groq LPU™ structured architecture [1-3].

In literature, various cooling methods have been suggested to address hotspots with high heat fluxes. Sharma, C. S., et., [4-5] has introduced a passive energy-efficient, hotspot-targeted embedded liquid cooling system. The design incorporates fine channels positioned over the hotspots, coarse channels covering the background, and introduces a flow throttling zone to regulate flow in various regions. The optimized designs achieve effective coolant flow distribution without the need for external flow control devices, and performance is only minimally impacted by the manifold geometry used for supplying coolant to the microchannel heat transfer structure. Snyder, G. J., et al., [6] presented an active hot spot cooling solution, employing embedded thermoelectric cooling (eTEC) and capitalizing on wafer fabrication process technology. Radmard, V., et al., [7] introduced an impinging chip-attached micro pin fin direct liquid cooling solution for hotspot targeted applications. They reported the lowest thermal resistance (0.015 K/W) in the field of single-phase cooling research.

The thermal performance is further enhanced by combining microchannels with an impinging microjet array, forming a hybrid Si heat sink for package-level hotspot cooling. A hybrid microchannel heat sink with ultra-low power consumption is introduced, featuring a central hotspot [8]. This design aims to

minimize temperature gradients, enhance hotspot cooling, and prevent overcooling at the inlet region. Bar-Cohen, A., & Wang, P. [9] examined thermoelectric micro-coolers and two-phase microgap coolers for effectively handling high heat flux hot spots. The underlying physical phenomena along with fundamental modeling equations and typical results, are thoroughly described. However, existing microchannel cooling solutions face challenges like nonuniform flow distribution, leading to additional temperature irregularities. In our previous study Wei, T., [10-12] demonstrated a low-cost energy-efficient hotspot targeted cooling solution with polymer-based impingement jet cooling is introduced, where the location of the impinging jet nozzles that eject the coolant onto the chip can be aligned to the location of the hotspots. The primary emphasis was directed only towards mitigating the heat generated specifically at the hotspot location, without accounting for the background power dissipation. The core strategy aimed to increase the local flow rate exclusively within the hot spot region. However, the total background power dissipation might be even higher than the hotspot power despite the lower power density. Therefore, the development of effective cooling strategies that address both hot spot and background power dissipation are needed.

This study investigates the thermal management strategies targeting hotspot mitigation on Groq LPU™. The paper addresses the potential temperature non-uniformity balancing from individual functional blocks and explores thermal coupling mitigation solutions between the LPU™ and neighboring chiplet as exploratory use cases. **Section II** introduces detailed geometry dimensions and power information for a representative use case. A comprehensive modeling of the LPU™ package is performed to demonstrate the temperature distributions within the LPU™. The modeling methodology, utilizing the FEM (Finite Element Method) model and applied thermal boundary conditions are also introduced in this section. **Section III** explores the concept of microjet cooling directly targeted at on-chip hotspots by power map and cooling map alignment. Detailed design guidelines are proposed based on the modeling comparison and analysis. **Section IV** focuses on investigating the temperature distributions within the LPU™ die, as well as the thermal coupling effects between the Groq LPU™ and neighboring chiplets. Moreover, hotspots thermal management solutions are also proposed and explored in this section to minimize the temperature gradient across the Groq LPU™ chip.

II. Model Introduction and Methodology

A. *Groq LPU™ chip power map*

In the Groq LPU™ design, thermally significant functional blocks are visualized in Fig.2. The power virus use case for these blocks differs: region 1, region 2, region 3, region 4 and region 5 consumes 285.7 W, 51 W, 2.5 W, 29.6 W and 1 W, respectively as modeling conditions. The region 6 on the left edge of the LPU™ is assumed to 0 W. As a result, the power virus use case consumption of a LPU™ reaches 369.9 W. Additionally, the dimensions of the LPU™ die are 25.3 mm × 28.58 mm. The power distribution details of a dummy representative workload are listed in Table I.

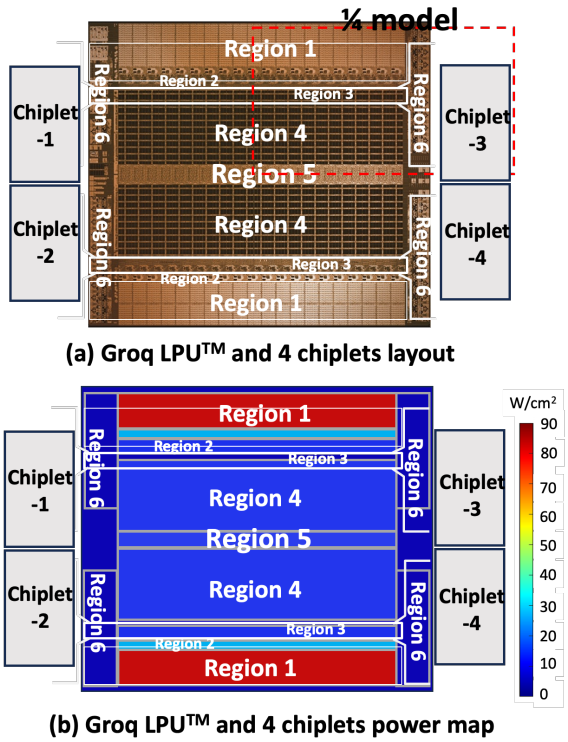

Fig.2: Exploratory use case for Groq LPU™ chip package layout: (a) integration of the Groq LPU™ and 4 chiplets on the silicon substrate; (b) power virus use case of the Groq LPU™ and chiplets.

TABLE I: POWER VIRUS USE CASE WORKLOADS OF DIFFERENT FUNCTIONAL BLOCKS

(note: power virus use case for thermal modeling purposes only)

Functional block region	Heat flux (W/cm^2)	Local power (W)
Region 1	83.3	35.6
Region 2	28.98	1.97
Region 3	10.41	0.989
Region 4	12.73	9.39
Region 5	15.15	1.726
Region 6	0	0

Table II lists the critical dimensions and thickness of the LPU™ die and the neighboring chiplets in the package platform. Both the LPU™ die and chiplet stacks are assumed to have an equivalent height, resulting in a planarized package after assembly. In our FEM model, TIM-1 is excluded, leading to bare die cooling where the liquid directly contacts the silicon die backside surface. The gap between the central large LPU™ die and the surrounding chiplets is assumed to be filled with mold material with a thermal conductivity of 1.2 W/(m·K). Uniform power density assumptions are applied for each LPU™ die functional module and individual chiplet dies.

TABLE II: PACKAGES PARAMETERS FOR LARGE LPU™ DIE FUNCTIONAL MODULE AND INDIVIDUAL CHIPLET DIES.

Component /material	X(mm)	Y(mm)	Z(micrometer)
Groq LPU die	28.58	25.3	700
Groq LPU BEOL	28.58	25.3	10
Chiplet (overall size)	12	8	720
Chiplet ctrl Si	12	8	50
Chiplet ctrl BEOL	12	8	3
Microbump layer			15
Si substrate	44.66	25.3	100
Si substrate BEOL	44.66	25.3	10
Cu pillar	44.66	25.3	50
Package substrate	65	65	1400

B. Finite Element Model (FEM) for Groq LPU™

A Thermal Finite Element Model (FEM) with a simplified 1⁄4 symmetry thermal model is utilized to evaluate the thermal performance of the Groq LPU™ chip package, as illustrated in Fig.3. The heat is generated in the thin active region of the chip, marked as front-end-of-line (FEOL). The FEOL is modelled as a surface heat flux at interface between BEOL and silicon substrate. These heat fluxes are utilized to define the heat generation for the LPU™ die and chiplet modules. The total power is around 20 W for each chiplet module. Considering the PCB side, an equivalent thermal resistance of 20 K/W is assumed. Moreover, effective thermal conductivity properties are used for all the BEOL and solder bump layers [14-16].

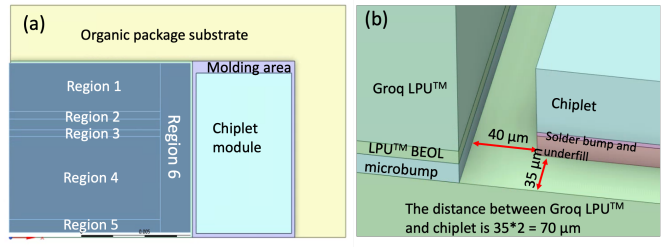

Fig.3: Groq LPU™ chip package: (a) top view of the FEM quarter model; (b) side view of multi-chip module package.

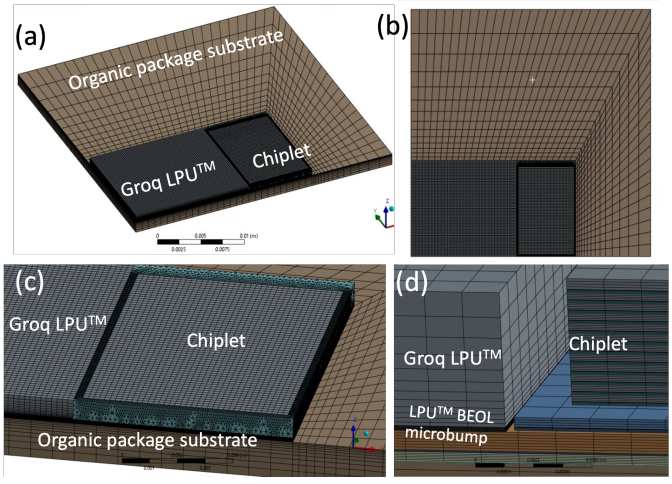

Fig.4: Meshing details for the FEM model including the Groq LPU™ die and chiplet die package: (a-b) over view of the package model meshing; (b-c) side view of multi-chip module meshing details.

Through detailed mesh sensitivity study with elements from 511,211 to 1,044,963. The element number 971,275 are employed to construct a precise thermal model capable of encompassing the necessary details, including the LPU™ power map. The details of the FEM modeling meshing are illustrated in Fig.4. The heat is generated in the thin active region of the chip, marked as front-end-of-line (FEOL). The specified maximum allowable LPU™ temperature is 105 ℃. The inlet temperature of the liquid (DI-water) is kept as 10ºC by using a heat exchanger in the closed-loop.

III. CONCEPT OF HOTSPOT TARGETED COOLING

In this section, effective cooling strategies that address both hotspot and background power dissipation will be investigated. This concept is shown in Fig.5. To explore this scenario, we conducted a Computational Fluid Dynamics (CFD) and thermal analysis within the context of a large die thermal test vehicle containing two distinct heat sources. To manage computational expenses, we employed a scaled-down cooler model with a simplified representation of the hotspot region. This approach allowed us to investigate alterations in localized nozzle diameters and pitch, aiding in understanding the thermal and hydraulic behaviors within the jet cooling system.

As illustrated in Table III, four test cases have been established based on the nozzle diameter and nozzle pitch in the hotspot (HS) and background (BG) zones. Design 1 (D1), used as a baseline test case, is designed with uniform nozzle jet cooling with a nozzle diameter of 0.3 mm and a pitch of 2 mm in both the HS and BG areas. For design 2 (D2), the inlet nozzle pitch within the HS region is reduced from 2 mm to 1 mm, resulting in a denser nozzle array there. In Design 3 (D3), the nozzle diameter in the BG region is increased from 0.3 mm to 0.6 mm, while maintaining a nozzle diameter of 0.3 mm and a pitch of 1 mm in the HS region. As a benchmark, for design 4 (D4), the nozzle diameter within the HS region is increased from 0.3 mm to 0.5 mm while maintaining a 1 mm nozzle pitch, in comparison with Design 2. In this section, the temperature, fluid distribution and pressure drop of the four different cases will be evaluated to better understand the thermal and flow behaviors within the hotspot targeted cooling. This study will help to provide design guidelines regarding the optimal cooling strategies for both hot spot and background power dissipation. The CFD and thermal analysis is performed at 3 LPM, with HS at 200 W/cm^2 and BG at 50 W/cm^2.

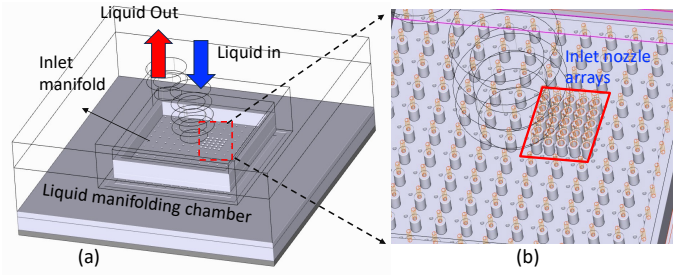

Fig.5: Hotspot targeted cooler design based on configuration with wider and more nozzles in the hot spot region: (a) overall cooler design with multiple nozzles; (b) enlarge view of the local refined nozzles for hotspot cooling.

TABLE III: VARIATIONS OF THE NOZZLE DIAMETER AND NOZZLE PITCH FOR HIGH POWER HOTSPOT REGION AND LOW POWER BACKGROUND REGION.

Case No.	Hotspot region		Background region	
	Nozzle diameter	Nozzle pitch	Nozzle diameter	Nozzle pitch
D1	0.3 mm	2 mm	0.3 mm	2 mm
D2	0.3 mm	1 mm	0.3 mm	2 mm
D3	0.3 mm	1 mm	0.6 mm	2 mm
D4	0.5 mm	1 mm	0.3 mm	2 mm

The temperature distributions for the four different cases are illustrated in Fig.6. Fig. 6(a) presents the reference design under uniform nozzle diameter and pitch cooling in HS and BG zones. In Fig. 6(b), representing D2 test case, the reduction of nozzle pitch from 2 mm to 1 mm in the HS region causes an increase in the inlet nozzle array from 3×3 to 5×5, thereby resulting in a higher inlet nozzle number. Due to the constant total flow rate at 3 LPM for all the test cases, the single inlet velocity decreases as the total inlet number increases, leading to higher HS temperatures compared to D1 in Fig. 6(a). Notably, the total pressure drop decreases due to the increased total inlet nozzle numbers within the HS region, as shown in Fig. 7 and Fig. 8. Furthermore, the HS temperature becomes even worse when the inlet nozzle diameter in the BG region increases from 0.3 mm to 0.6 mm, as seen in comparison to D2 in Fig. 6(b). This occurs because the pressure within the BG nozzle region becomes lower than in the HS region, causing an increased bypass flow within the nozzle channels, depicted in Fig. 7(c). Meanwhile, the pressure drop of D3 reduces significantly due to the overall increase in the inlet nozzle area within the BG region, as indicated in Fig. 8.

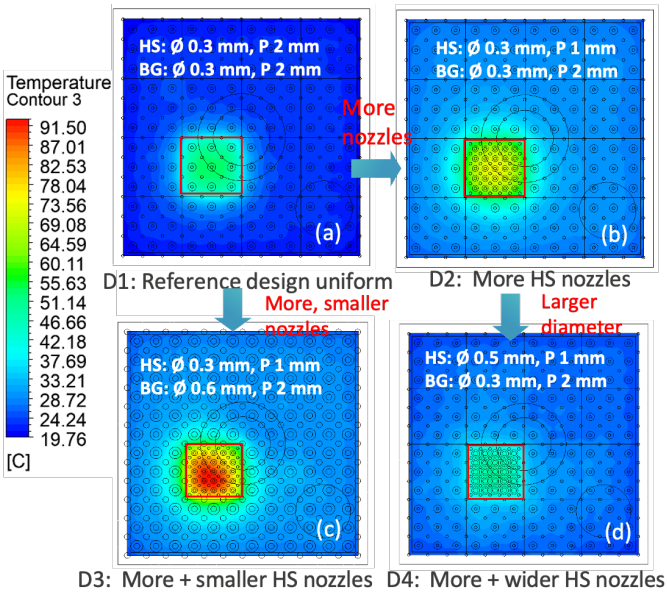

Fig.6: Temperature distribution map across the chip heat surface under four different test cases: (a) design 1; (b) design 2; (c) design 3; (d) design 4 (Thermal analysis at 3 LPM, HS = 200 W/cm^2, BG = 50 W/cm^2).

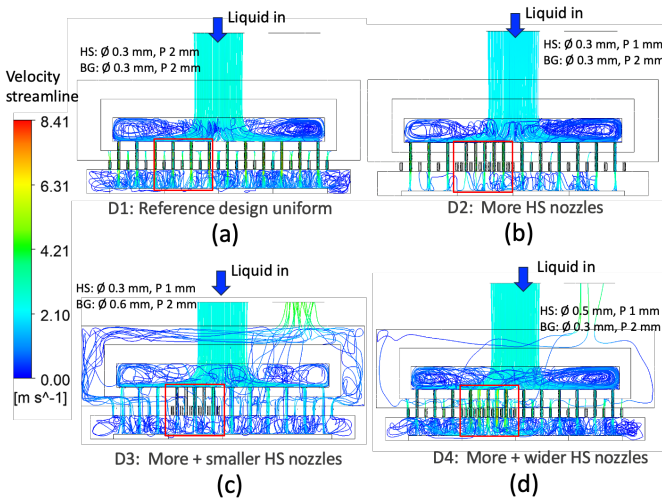

Fig.7: Flow distribution and velocity streamline within the HS-targeted jet cooler using CFD simulation for the four tests cases: (a) design 1; (b) design 2; (c) design 3 and (d) design 4 (Thermal analysis at 3LPM, HS = 200 W/cm^2, BG = 50 W/cm^2).

In the case of D4 depicted in Fig. 6(d), designed with a larger nozzle diameter (0.5 mm) and finer pitch (1 mm) within the HS region, alongside a smaller nozzle diameter (0.3 mm) and larger nozzle pitch (2 mm) within BG region, results in the lowest temperature compared to the other cases. This lower HS temperature primarily arises from the compromise in the pressure drop within the HS and BG regions. As depicted in Fig. 7(d), the larger nozzle and finer pitch within the HS region facilitate increased liquid flow through this area, consequently leading to a lower temperature in the HS. The pressure drop for D4 in Fig.8 shows lower value compared to that of D1 and D2.

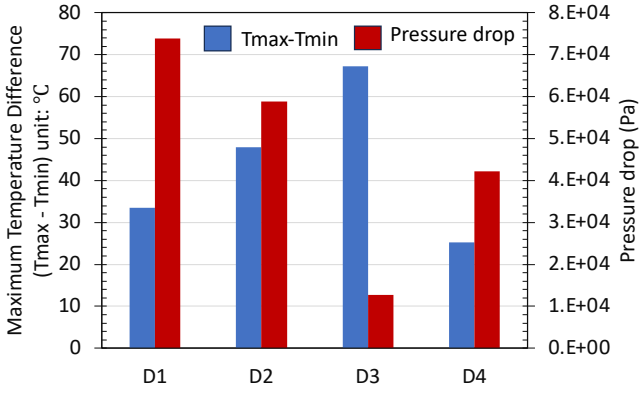

Fig.8: Maximum temperature difference and pressure drop comparison for the four test cases (Thermal analysis at 3 LPM, HS = 200 W/cm^2, BG = 50 W/cm^2).

To further understand the trade-off between thermal and pressure considerations, the temperature profile across the HS and BG regions is depicted in Fig.9. Through the implementation of wider or finer nozzles pitch in the hotspot area, the flow rate at this specific location increases, while at the cost of reduced flow rates in the background region. Thus, among the investigated configurations across the four different cases, the most optimal configuration is observed in the D4 scenario, incorporating a larger nozzle diameter and fine nozzle

pitch within the HS region, while maintaining a balance between the nozzle diameter and pitch within the BG region.

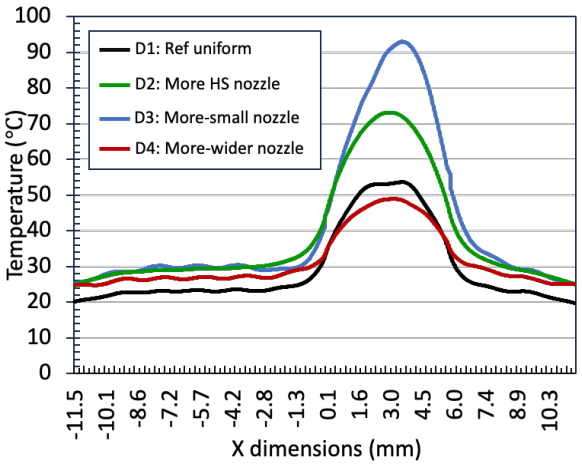

Fig.9: Temperature profile across the HS region and BG region (Thermal analysis at 3LPM, HS = 200 W/cm^2, BG = 50 W/cm^2).

IV. HOTSPOTS THERMAL MANAGEMENT FOR LPU^{TM}

Based on the studies from Section III, we investigate a customizing impingement jet cooling solutions that can regulate both the inlet nozzle diameter and nozzle pitch to align with the power map throughout the entire LPU^{TM} chip. Furthermore, we can adjust the nozzle array dimensions for various chiplet modules like LPU^{TM}, TPU, and other chiplet. Our newly developed predictive model enables the extraction of an equivalent average heat transfer coefficient value based on the inlet nozzle diameter and nozzle pitch [13].

In this section, we aim to demonstrate our concept by implementing targeted cooling directly on the hotspots of the Groq LPU^{TM} package. As a baseline, we investigate the temperature distribution when implementing uniform nozzle jet cooling on the Groq power virus use case power map, demonstrating significant temperature difference across the entire LPU^{TM} chip surface. Two solutions have been proposed to address these temperature non-uniformity concerns. The first solution involves utilizing a dummy heat spreader to guide the heat through the bottom silicon substrate package to the dummy spreader. The second solution focuses on hotspot-targeted cooling, by aligning the heat transfer coefficient map with the power distribution of hotspots on the chip. Within our modeling, we've outlined two primary objectives as guiding principles: first, maintaining a specified maximum allowable LPU^{TM} temperature of 105 ℃, and restricting the maximum chiplet temperature to 95 ℃; second, minimizing the temperature difference between the maximum temperature and the lowest temperature possible, ensuring a more uniform temperature distribution for enhanced package thermal mechanical reliability.

A. Uniform nozzle array jet cooling on LPU^{TM}

As depicted in Fig.10, we've outlined the temperature distribution within the silicon substrate package, illustrating the temperature maps for the Groq LPU^{TM} die package. The varied power levels applied within the Groq LPU^{TM} chip result in a large temperature difference across the LPU^{TM} chip. The highest temperature region is observed within the high-power density zone at edges, under a heat flux level of 83.3 W/cm^2, while the lowest temperature region is found in the center of the silicon, applied with a heat flux of 15.5 W/cm^2. The temperature contour shows that heat spreading occurs through the neighboring chiplet to the LPU^{TM} due to the proximity of the chiplet to the low power region in the LPU^{TM} region.

For a more comprehensive understanding of the thermal interaction between the LPU™ and neighboring chiplet, we've plotted temperature profiles along both the LPU^{TM} and neighboring chiplet in five distinct horizontal directions in Fig.11. The presence of the low power region between the LPU^{TM} and neighboring chiplet allows heat to also flow through the neighboring chiplet toward the colder LPU^{TM} region. Consequently, the maximum temperature difference observed within the LPU^{TM} die stands at 30 °C.

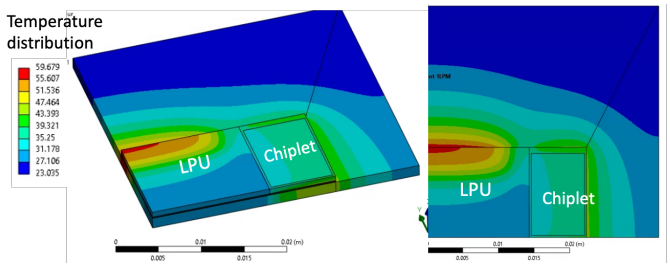

Fig.10: Temperature distribution between the Groq LPU^{TM} chip as well as the neighboring chiplet.

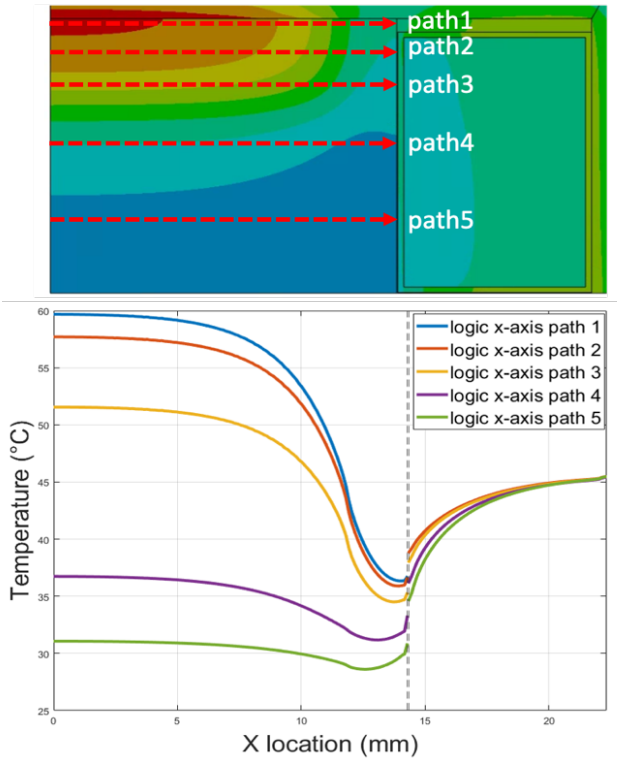

Fig.11: Temperature distribution and profile along the horizontal direction between the Groq LPU^{TM} chip as well as the neighboring chiplet.

B. *Hotspots targeted jet cooling schemes on LPU™*

As demonstrated in the above thermal analysis, uniform cooling strategy at packaging level is not ideal to mitigate the temperature gradient issue within logic silicon. Therefore, here we introduced another hotspot targeted package level cooling strategy. Impingement cooling nozzles were strategically placed based on the power (power density) level for logic functional blocks. To be consistent with our previous hotspots targeted jet cooling experimental measurement data, we assume that the inlet temperature of the liquid (DI-water) is kept as 10ºC by using a heat exchanger in the closed-loop.

From previous literature review [12-13], we have demonstrated and validated an analytical predict model for the heat transfer coefficient and pressure drop analysis for hotspots jet cooling. Our experimental data show that the Nu shows as a function of the Re with a power-law trend with an exponent of 0.67, listed as below:

$$Nu = 1.24 \cdot Re^{0.67} \tag{1}$$

where Nu is based on the measured averaged chip temperature. The hydraulic characteristics lengths of the dimensionless number Nu and Re are both based on the inlet nozzle diameter d_i. The local heat transfer coefficient *htc* also shows as a function of the local flow rate per nozzle $\dot{V}$ with a power-law trend with an exponent of 0.67.

For the local flow rate analysis, a unit cell approach is used as a first estimation, to assess the improvement in cooling at the targeted chip areas. Based on the $Nu - Re$ correlation, the relation between the local heat transfer rate htc and local inlet nozzle flow rate $\dot{V}$ is shown with a power-law trend with an exponent of 0.67, derived from equation (1). Therefore, the expected heat transfer coefficient htc^* for the hotspot cooler can be extracted as below:

$$htc^* = m^{0.67} \cdot htc \tag{2}$$

$$\dot{V}^* = m \cdot \dot{V} \tag{3}$$

$$m = \frac{N^2}{M} \tag{4}$$

$$\dot{V} = \frac{\dot{V}_{tot}}{N \times N} \tag{5}$$

Where $\dot{V}_{tot}$ is the total flow rate for the cooler, and $\dot{V}$ is the local flow rate per nozzle. Also, N × N is the inlet nozzle array. Similarly, the expected pressure drop for the targeted cooler is shown as below:

$$\Delta p \sim m^2 \cdot \dot{V}^2 \tag{6}$$

where $\dot{V}$ is the averaged local flow rate per nozzle, $\dot{V}^*$ is the averaged local flow rate for the hotspot targeted cooler. N^2 is the total inlet nozzle number with the array cooler. M is the total inlet nozzle number for the hotspot targeted cooler. And m is defined as the ratio between N^2 and M.

Based on the predictive model, Table IV below lists 5 different cases with corresponding HTC value at each main power block, assuming liquid (DI-water) cooling with inlet temperature kept as 10ºC. Combined with Groq LPU™ deterministic power profile, it exhibits great potential to achieve uniform temperature distribution in the silicon.

TABLE IV: HEAT TRANSFER COEFFICIENT APPLIED FOR DIFFERENT FUNCTIONAL REGION

(Note: using water as the working fluid, with an inlet temperature of 10 ℃ the Heat Transfer Coefficient Unit: W/m²/K)

Name	Heat flux (W/cm^2)	HTC-Case 1	HTC-Case 2	HTC-Case 3	HTC-Case 4	HTC-Case 5
Region 1	83.3	39200	26133	19600	15680	13067
Region 2	28.98	13635	9090	6817	5454	4545
Region 3	10.41	4900	3267	2450	1960	1633
Region 4	12.73	5992	3995	2996	2397	1997
Region 5	15.15	7127	4752	3564	2851	2376
Region 6	0	2450	1633	1225	980	817

The equation (1) can be further developed to equation (7), where a=1.24, b=0.67 are constant for Nu-Re correlation. d_i is the local inlet diameter. K_f, ρ and μ are water properties of thermal conductivity, density and viscosity. Q_l and N_l are local volumetric flow rate and number of local inlet nozzles (for certain region), respectively. Given a total inlet flow rate, which is then redirected to each heat region with certain Q_l, this equation can help with designing a hot-spot targeted cooling system with fixed dimensionless number $d_i/L=0.3$ and $H/L=0.2$ where H is the distance from nozzle to hot surface, and L is the pitch of nozzle system. A detailed parameter design for testing conditions is shown in Table.V.

$$\bar{h} = a d_i^{-b-1} k_f \left(\frac{4\rho Q_l}{\mu \pi N_l}\right)^b \tag{7}$$

TABLE V: NOZZLE NUMBER CORRESPONDING TO THE HEAT TRANSFER COEFFICIENT APPLIED FOR DIFFERENT FUNCTIONAL REGION

(Note: using water as the working fluid, with an inlet temperature of 10 ℃ the Heat Transfer Coefficient Unit: W/m²/K)

Name	Heat flux (W/cm^2)	Nozzle diameter (mm)	Nozzle pitch (mm)	Nozzle number	Total local flow rate mL/min	HTC-Case 1
Region 1	83.3	0.42	1.41	30 = (2×15)	54	39200
Region 2	28.98	0.91	3.05	8 = (1×8)	20.4	13635
Region 3	10.41	1.36	4.55	6 = (1×6)	9	4900
Region 4	12.73	0.68	2.27	30 = (3×10)	10.8	5992
Region 5	15.15	1.02	3.40	8 = (1×8)	10.20	7127
Region 6	0	0.78	2.61	9 = (1×9) each	4.8	2450

Fig.12 shows the temperature versus distance plot for various HTC cases. Starting from the center of the Groq LPU™ to outer edge of adjacent chiplet module, temperature slightly decreases due to thermal spreading effects till the edges of LPU™ die, then the relative temperature difference between logic dies and chiplet module is the result of different HTC

boundary conditions on top of silicon. When higher HTC is applied to logic die, one can expect it to have a lower temperature compared to chiplet module. This also gives us an opportunity to not only mitigate temperature variation reliability concern of logic die, but also strategically manage thermal budget within a package. HTC case 5 shows higher logic die temperature compared to chiplet module for example, whereas when HTC cases 1 is applied, logic and chiplet module are within a narrow temperature fluctuation range. Because of thermal coupling between Groq LPU™ and chiplet module, the selection of HTC on LPU™ will also have non-negligible impact on neighboring chiplet die temperatures. Utilizing proposed hotspot targeted cooling strategies, the temperature gradient across the LPU™ can be limited to below 4 ℃, using optimal case 1.

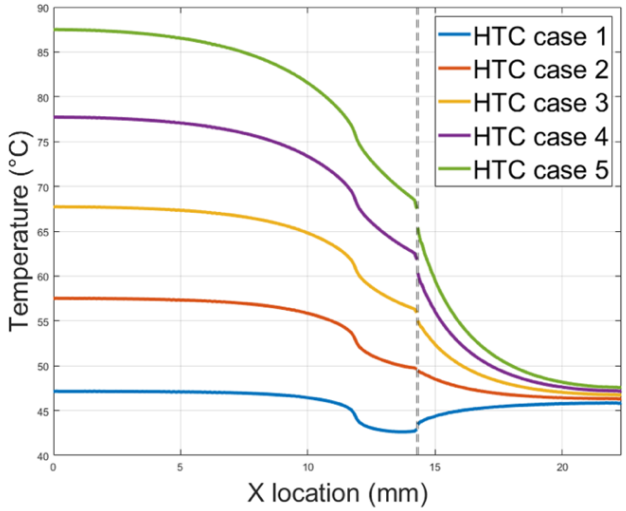

Fig.12: Temperature distribution along the horizontal direction of the LPU™ and neighboring chiplet, under hotspots targeted cooling.

C. *Dummy silicon heat spreader*

Cooling on the lidded package could potentially improve the heat distribution of the high-power density function block. However, this solution would also cause the heat to spread to the chiplet modules, resulting in an increased temperature there. To address this, bare die jet cooling with dummy heat spreader is proposed to mitigate the thermal nonuniform distribution across the LPU™ chip, as well as maintain a lower temperature of chiplet module shown in Fig.13. Utilizing our hotspot targeted package level cooling strategy, a tailored heat transfer coefficient cooling will be applied on the heat spreader by customizing the nozzle diameters and nozzle pitch. Fig.13 illustrates how the peak temperature of the LPU™ decreases due to the heat spreading effects generated by the dummy silicon heat spreader. The heat transfer pathway passes through the silicon substrate layer, the microbump bonding layer, and exits through the dummy silicon heat spreader, benefiting from an efficient tailored impingement jet cooling method. Fig.14 illustrates the impact of different heat transfer coefficient values applied on the heat spreader, demonstrating an efficient cooling strategy to minimize the temperature different across the Groq LPU™ die.

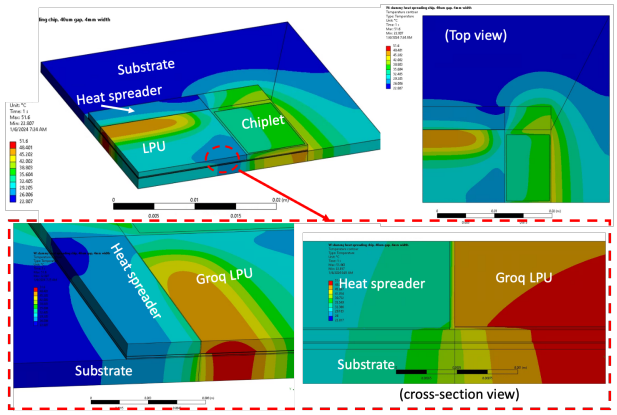

Fig.13: Temperature profile along the horizontal direction between the Groq LPU™ chip as well as the neighboring chiplet.

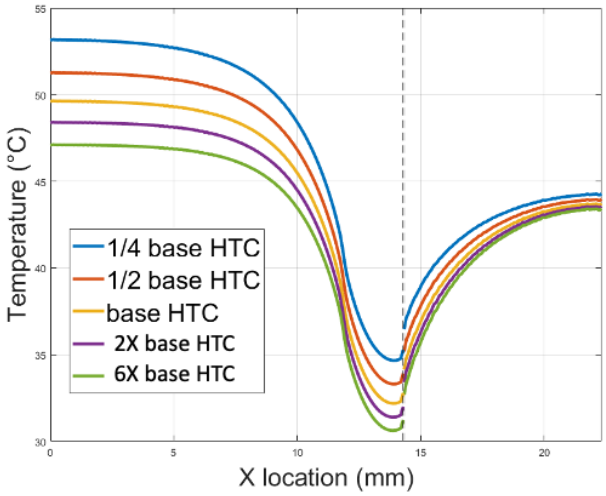

Fig.14: Temperature distribution between the LPU™ chip as well as the neighboring chiplet.

V. Conclusion

In our study, we've successfully developed efficient cooling strategies addressing the hotspot with different power functional block within LPU™ chip. This research explores the concept of microjet cooling directly targeted at on-chip hotspots by power virus use case power map / cooling map alignment throughout the Groq LPU™ chip. The customized cooling map can be achieved by tailoring the inlet nozzle diameter and pitch, as well as the flow conditions to align to hotspots location within the chip, or power densities for various workloads. Our modeling analysis identified the optimal configuration among the investigated setups, emphasizing a larger nozzle diameter and fine nozzle pitch in the hotspot region while maintaining a balanced nozzle diameter and pitch in the background region across four different cases.

Furthermore, we've applied this concept practically by cooling the hotspots on the Groq LPU™ chip package platform. Initially, we analyzed the temperature distribution by implementing uniform nozzle jet cooling on the Groq LPU™ non uniform power map of power virus use case, showing noticeable temperature variations of about 30 ℃ across the entirety of the LPU™ chip surface. To address these temperature

nonuniformity issues from the uniform nozzle jet cooling, we proposed two solutions: one is focusing on targeted hotspot cooling by aligning the cooling map with hotspot power distribution and the other solution is involving a dummy heat spreader to enhance the heat flow through the bottom silicon substrate package. This targeted cooling approach demonstrated achieving a uniform temperature within 4 ℃ difference across the chip.

In summary, direct on-chip impingement jet cooling presents distinct advantages over other cooling technologies. Customizing nozzle parameters to different power maps (workloads) enables achieving uniform temperatures across different chip module inside the package.

REFERENCES

[1] Abts, D., Ross, J., Sparling, J., Wong-VanHaren, M., Baker, M., Hawkins, T., ... & Kurtz, B. (2020, May). Think fast: A tensor streaming processor (TSP) for accelerating deep learning workloads. In *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)* (pp. 145-158). IEEE.

[2] Abts, D., Kim, J., Kimmell, G., Boyd, M., Kang, K., Parmar, S., ... & Ross, J. (2022, August). The Groq Software-defined Scale-out Tensor Streaming Multiprocessor: From chips-to-systems architectural overview. In *2022 IEEE Hot Chips 34 Symposium (HCS)* (pp. 1-69). IEEE Computer Society.

[3] Abts, D., Kimmell, G., Ling, A., Kim, J., Boyd, M., Bitar, A., ... & Ross, J. (2022, June). A software-defined tensor streaming multiprocessor for large-scale machine learning. In *Proceedings of the 49th Annual International Symposium on Computer Architecture* (pp. 567-580).

[4] Sharma, C. S., Tiwari, M. K., Zimmermann, S., Brunschwiler, T., Schlottig, G., Michel, B., & Poulikakos, D. (2015). Energy efficient hotspot-targeted embedded liquid cooling of electronics. *Applied Energy*, *138*, 414-422.

[5] Sharma, C. S., Schlottig, G., Brunschwiler, T., Tiwari, M. K., Michel, B., & Poulikakos, D. (2015). A novel method of energy efficient hotspot-targeted embedded liquid cooling for electronics: An experimental study. *International Journal of Heat and Mass Transfer*, *88*, 684-694.

[6] Snyder, G. J., Soto, M., Alley, R., Koester, D., & Conner, B. (2006, March). Hot spot cooling using embedded thermoelectric coolers. In *Twenty-Second Annual IEEE Semiconductor Thermal Measurement And Management Symposium* (pp. 135-143). IEEE.

[7] Radmard, V., Azizi, A., Rangarajan, S., Fallahtafti, N., Hoang, C. H., Mohsenian, G., ... & Sammakia, B. (2021, June). Performance Analysis of Impinging Chip-Attached Micro Pin Fin Direct Liquid Cooling Package for Hotspot Targeted Applications. In 2021 20th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (iTherm) (pp. 220-228).

[8] Rao, X., Jin, C., Zhang, H., Song, J., & Xiao, C. (2023). A hybrid microchannel heat sink with ultra-low pressure drop for hotspot thermal management. International Journal of Heat and Mass Transfer, 211, 124201.

[9] Bar-Cohen, Avram, and Peng Wang. "Thermal management of on-chip hot spot." (2012): 051017.

[10] Wei, T., Oprins, H., Cherman, V., Beyne, E., & Baelmans, M. (2019). Low-cost energy-efficient on-chip hotspot targeted microjet cooling for high-power electronics. *IEEE Transactions on Components, Packaging and Manufacturing Technology*, *10*(4), 577-589.

[11] Lyu, S., Wu, Q., & Wei, T. (2023, May). Hotspot-targeted Cooling Scheme with Hybrid Jet Impingement/Thermal Through Silicon Via (TSV). In 2023 22nd IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm) (pp. 1-9).

[12] Wei, T., Oprins, H., Cherman, V., De Wolf, I., Beyne, E., & Baelmans, M. (2019, May). First demonstration of a low cost/customizable chip level 3D printed microjet hotspot-targeted cooler for high power applications. In *2019 IEEE 69th Electronic Components and Technology Conference (ECTC)* (pp. 126-134). IEEE.

[13] Wei, T., Oprins, H., Fang, L., Cherman, V., Beyne, E., & Baelmans, M. (2022). Heat transfer and pressure drop correlations for direct on-chip microscale jet impingement cooling with alternating feeding and draining jets. International Journal of Heat and Mass Transfer, 182, 121865.

[14] Oprins, H., Wei, T., Cherman, V., & Beyne, E. (2023, May). Liquid jet impingement cooling of high-performance interposer packages: a hybrid CFD–FEM modeling study. In *2023 22nd IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)* (pp. 1-10). IEEE.

[15] Oprins, H., Vandevelde, B., Badaroglu, M., Gonzalez, M., Van der Plas, G., & Beyne, E. (2013, May). Numerical comparison of the thermal performance of 3D stacking and Si interposer based packaging concepts. In *2013 IEEE 63rd Electronic Components and Technology Conference* (pp. 2183-2188). IEEE.

[16] Oprins, H., & Beyne, E. (2014, May). Generic thermal modeling study of the impact of 3D-interposer material and thickness options on the thermal performance and die-to-die thermal coupling. In *Fourteenth Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)* (pp. 72-78). IEEE.