Monolithic Integration of a Micropin-Fin Heat Sink in a 28-nm FPGA

Thomas E. Sarvey, Yang Zhang, *Student Member, IEEE*, Colman Cheung, Ravi Gutala, Arifur Rahman, *Senior Member, IEEE*, Aravind Dasu, and Muhannad S. Bakir, *Senior Member, IEEE*

***Abstract*— Microfluidic cooling has been demonstrated as an effective means of cooling microelectronic circuits with a very low convective thermal resistance and potential for integration in close proximity to the area of heat generation. However, microfluidic cooling experiments to date have been limited to silicon with resistive heaters representing the heat generating circuitry. In this paper, a micropin-fin heat sink is etched into the back side of an Altera Stratix V field-programmable gate array (FPGA), built in a 28-nm CMOS process. Thermal and electrical measurements are made running a benchmark pulse compression algorithm on the FPGA. Deionized water is used as a coolant with flow rates ranging from 0.15 to 3.0 mL/s and inlet temperature ranging from 21 °C to 50 °C. An average junction-to-inlet thermal resistance of 0.07 °C/W is achieved.**

***Index Terms*— Deep reactive ion etching (DRIE), field-programmable gate arrays (FPGAs), high-performance computing (HPC), microfluidic cooling, thermal management.**

I. Introduction

As MORE functionality and higher density logic continue to be packed into increasingly dense systems, traditional cooling systems are being pushed to their limit, leading to the problem of dark silicon and throttled performance. Microfluidic cooling, first demonstrated by Tuckerman and Pease [1], has the potential to solve this cooling challenge for high-power and high-performance integrated circuits. Microfluidic cooling has the potential for very low junction-to-fluid thermal resistance in a very small form factor. Low thermal resistance opens the possibility of cooling very high heat flux integrated circuits with a moderate inlet temperature, or moderate heat fluxes with an elevated inlet temperature. Cooling with elevated inlet temperatures can reduce or eliminate the need for chilling of the coolant below maximum outside ambient temperatures and open the possibility of waste heat reuse, increasing data center energy efficiency [2].

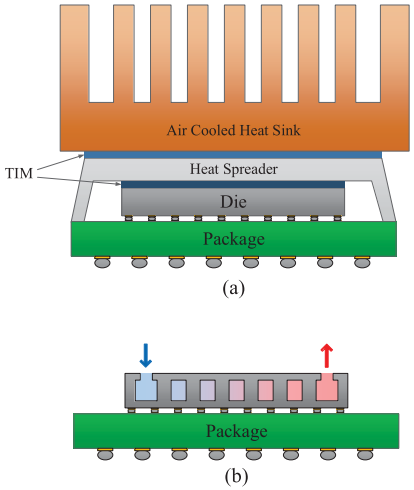

Fig. 1. (a) Traditional microelectronic system. (b) Microelectronic system with monolithically integrated MFHS.

The most common method of cooling microelectronics has long been an air-cooled heat sink (ACHS) mounted on top of the packaged integrated circuit, as shown in Fig. 1(a). Efficacy is limited with this solution and can be improved in several ways through the use of microfluidic cooling. First, due to the small heat sink dimensions and properties of the liquid coolants, much lower convective thermal resistances can be achieved with microfluidic cooling when compared with direct air cooling. In addition, in the traditional configuration shown in Fig. 1(a), there is a large conductive thermal resistance due to the large distance and, more importantly, several material interfaces through which heat must conduct in order to reach the heat sink. In order to improve thermal resistance at these interfaces, two levels of thermal interface material (TIM) are used, but these interfaces still remain a major bottleneck in total junction-to-ambient thermal resistance. By etching the heat sink directly into the silicon die, conductive thermal resistance between the heat source and heat sink is minimized.

The very low profile achievable with microfluidic heat sinks (MFHSs) also makes them compatible with many dense 2.5-D and 3-D systems, as shown in Fig. 2. The examples shown in Fig. 2 use a silicon interposer for signal and fluid routing, but a traditional organic package could also be used.

Integrating the MFHS in an interposer, as shown in Fig. 2(a), can offer thermal resistances superior to typical package-level

Manuscript received June 17, 2016; revised September 13, 2016 and April 19, 2017; accepted May 31, 2017. Date of publication September 15, 2017; date of current version October 5, 2017. This work was supported in part by DARPA under ICECool Applications under Grant HR0011-14-1-0002 and in part by the DARPA Young Faculty Award under Grant N66001-12-1-4240. Recommended for publication by Associate Editor K. Ramakrishna upon evaluation of reviewers' comments. *(Corresponding author: Thomas E. Sarvey.)*

T. E. Sarvey, Y. Zhang, and M. S. Bakir are with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: tsarvey@gatech.edu; steven.zhang@gatech.edu; muhannad.bakir@mirc.gatech.edu).

C. Cheung, R. Gutala, A. Rahman, and A. Dasu are with the Programmable Solutions Group, Intel, San Jose, CA 95134 USA (e-mail: ccheung@altera.com; rgutala@atlera.com; arahman@altera.com; adasu@altera.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCPMT.2017.2740721

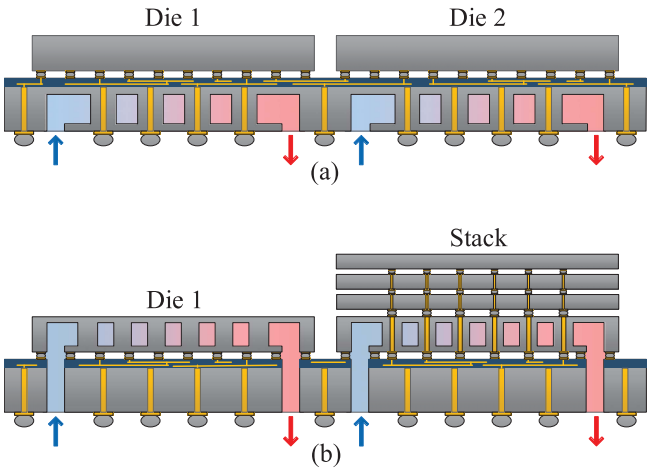

Fig. 2. Microfluidic cooling integrated in the (a) interposer and (b) back side of the die.

ACHSs, without modifying the logic dice [3]. In order to bring the heat sink as (thermally) close to the area of heat generation as possible, the MFHS can be etched into the back side of the active silicon dice, as shown in Fig. 2(b). These microfluidic-cooled dice can then be stacked to form a 3-D stack. Signaling and power delivery are achieved with through silicon vias (TSVs) passing through the MFHSs. Although microfluidic cooling limits how far silicon dice can be thinned, high-aspect-ratio TSVs can be fabricated in micropin-fins to limit TSV capacitance [4]. Unlike traditional cooling methods implemented on the top of the active silicon, MFHSs can be integrated into multiple tiers in a 3-D stack, allowing cooling to scale with the number of high-power tiers [5], [6]. This could potentially enable the stacking of multiple high-power tiers that could not be cooled with a single heat sink.

Since Tuckerman and Pease [1] achieved a thermal resistance of 0.09 °C cm^2/W using microchannels etched into silicon, a great deal of effort has focused on achieving improved thermal resistance and characterizing microchannel or micropin-fin heat sinks with correlations to predict performance [7]–[9]. However, research to date has focused on passive silicon dice with resistive heaters representing the heat-producing circuitry. In this paper, we present a functional microfluidic-cooled CMOS circuit.

An Altera Stratix V field-programmable gate array (FPGA), built in a 28-nm process, was postprocessed to integrate a micropin-fin heat sink directly into the back of the flip-chip bonded silicon die, a few hundred micrometers from the active circuitry. This microfluidic-cooled FPGA was then tested with deionized water as a coolant at several flow rates and inlet temperatures. All testing was performed with Altera Stratix V DSP development boards. A comparison was made with a stock board with the default ACHS. Testing was performed at flow rates ranging from 0.15 to 3.0 mL/s and with inlet temperatures ranging from 21 °C to 50 °C.

II. Fabrication

The Stratix V FPGA consists of a silicon die that is flip-chip bonded onto an organic substrate. The back side of the

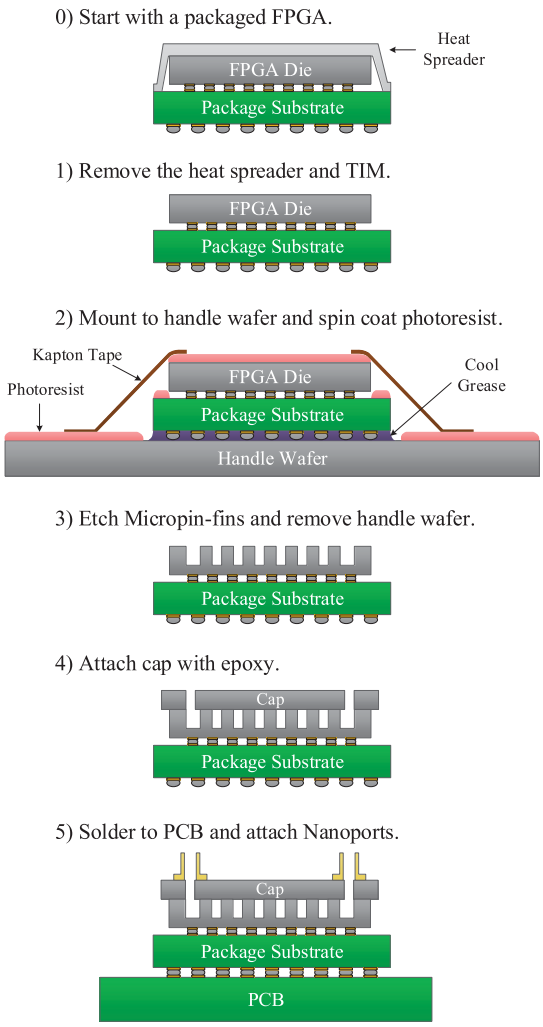

Fig. 3. Fabrication process for etching micropin-fins into the back side of a packaged FPGA die.

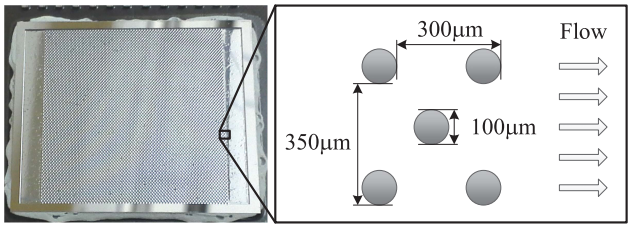

Fig. 4. Image of the etched back side of the silicon FPGA die along with the micropin-fin dimensions.

die was used to etch a micropin-fin heat sink in the same bulk silicon as the active circuitry. The Bosch process was used in order to etch silicon with vertical micropin-fin sidewalls. This batch process could be completed at the wafer level, but in this case was applied to a single chip at the die level. This fabrication flow was used for this proof of concept due to the relative ease of acquiring packaged parts, but may be different when optimized for scalability and manufacturing throughput.

The process used to add microfluidic cooling to a packaged Stratix V FPGA is shown in Fig. 3. First, the metal lid was removed along with TIM 1 on the back side of the die.

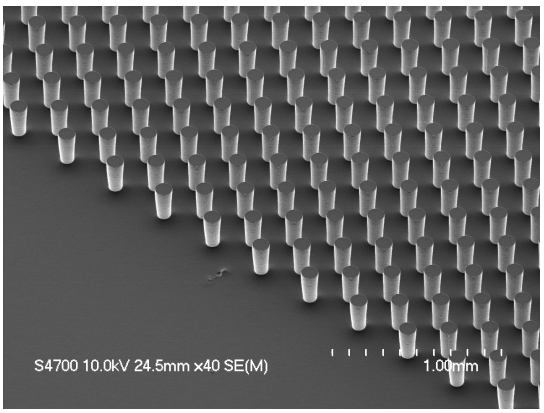

Fig. 5. SEM image of micropin-fins etched using the same process in a silicon wafer.

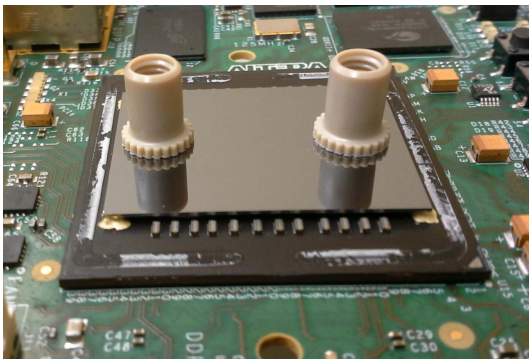

Fig. 6. Processed FPGA soldered to development board with silicon cap and nanoports for fluid delivery.

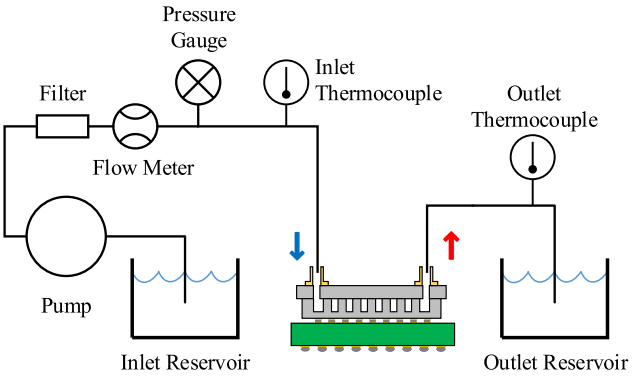

Fig. 7. Diagram of the open-loop system used to test the FPGA.

The flip-chip bonded die, along with the package substrate, was then attached to a carrier wafer with cool grease and Kapton tape to protect the package substrate and sides of the die. Photoresist was then spin-coated on the exposed back side of the silicon die, and the Bosch process was used to etch micropin-fins to a depth of approximately 240 μm. Inlet and outlet plena were formed in the same etching step by etching regions on either side of the micropin-fin array without any micropin-fins. A photograph of the etched die along with the micropin-fin dimensions can be seen in Fig. 4.

An SEM image of identical micropin-fins fabricated with the same process in a silicon wafer can be seen in Fig. 5. In general, aspect ratio and surrounding features are known to affect the profile and depth of etched cavities [10], [11]. Tapering of the micropin-fin sidewalls is visible on the micropin-fins closest to the inlet and outlet regions, but is minor in the rest of the array. Significant tapering, where the micropin-fin base is narrower than the top, could reduce fin efficiency and thermal performance by limiting the cross-sectional area through which heat can conduct up the micropin-fins.

A separate silicon lid was fabricated with an inlet and an outlet port. The lid was first tacked on to the top of the etched FPGA with high-temperature epoxy in order to provide a smooth surface for resoldering to the development board. After soldering, the lid was permanently secured with epoxy and nanoports were attached to deliver coolant. A photograph of the resoldered FPGA with nanoports can be seen in Fig. 6.

III. Testing

The FPGA was loaded with a custom pulse compression algorithm designed to mimic common DSP-style use cases of FPGAs and also to utilize a large amount of the FPGA resources. The algorithm consisted of nine soft computing cores that could be toggled ON and OFF during run time. The FPGA was tested in an open-loop system, shown in Fig. 7, with deionized water as a coolant. Testing was conducted with the Altera Stratix V DSP development board. The voltage regulator module (VRM) on the board was run at a current higher than its datasheet rating, so air was blown over it in order to prevent overheating.

Flow rate was measured with a rotameter, which was calibrated by repeatedly filling a known volume of deionized water at the experiment temperature. Variation between these repeated measurements was found to be less than 0.03 mL/s for the flow rates used in this paper. The pressure gauge used to measure pressure drop across the heat sink was calibrated using an Omega DPI610 calibrator to within 0.1 kPa. K-type thermocouples were used to make fluid temperature measurements at the inlet and outlet of the micropin-fin heat sink. The relative uncertainty between temperature measurements was found to be 0.1 °C in the temperature ranges used.

Die temperature measurements were taken using the Altera Power Monitor tool, which retrieves measurements from an on-die temperature diode with a resolution of 1 °C. At 20 °C, the temperature sensor on the FPGA die was found to have an offset from the thermocouples that was smaller than this resolution.

The FPGA was first tested with a flow rate of 2.4 mL/s, running zero to nine cores in order to vary the FPGA power. The inlet water temperature was 20.5 °C and the ambient air temperature was 19.3 °C. Temperature measurements can be seen in Table I. A stock Stratix V DSP development board was also tested for comparison, using the stock ACHS with which it was bundled.

The pulse compression algorithm uses 80% of the logic, 93% of memory blocks, and 98% of the DSP blocks on the FPGA. In addition to many subtraction, addition, multiplexing,

TABLE I
FPGA THERMAL AND POWER MEASUREMENTS WITH MFHS AND ACHS

Cores	MFHS FPGA Power (W)	ACHS FPGA Power (W)	MFHS FPGA Temp (°C)	ACHS FPGA Temp (°C)
0	13.2	13.7	21–22	43
1	15.4	16.0	21–23	46
2	17.6	18.3	22–23	49
3	19.8	20.5	22–23	51
4	21.9	22.8	22–23	53
5	24.0	25.1	22–23	56
6	26.2	27.5	22–23	59
7	28.3	29.8	22–24	61[1]
8	30.4	—	22–24	—
9	32.4	—	22–24	—

[1]Temperature warning on board illuminated.

look-up-table, and memory operations, 346 18-bit multiplications are done every clock cycle. Counting only these multiplications, 934 GOPS are performed when operating all nine cores at 300 MHz.

In order to capture the thermal gradient produced by heating of the fluid, measurements were taken with fluid flowing in both directions, as the temperature diode is located at the edge of the chip. Therefore, the temperatures of the microfluidic cooled FPGA are all reported as a range of two values representing flow in both directions.

A maximum die temperature of 60 °C was set to match the default on-board temperature warning indicator of the Stratix V DSP development kit. Although better thermal results could undoubtedly be achieved with a larger more powerful ACHS, it should be noted that the stock ACHS ran six computing cores before reaching this maximum temperature, while the microfluidic cooled FPGA ran all nine cores (a 1.5× improvement in throughput) while maintaining a die temperature below 24 °C (with additional power as per Table I). Although the air-cooled solution results in a higher junction temperature, it meets market requirements for Stratix V target applications.

The FPGA heat flux is lower than many high-power processors, and the low-profile ACHS with which it came is significantly less effective than the best available ACHSs. Since the temperature has a linear relationship with thermal resistance and power ($T_j = T_{\text{in}} + R_{\text{th}} P$), the temperature can be predicted for higher power and higher performance air cooling, assuming a constant thermal resistance. This is demonstrated in Fig. 8, where the average temperatures and powers from Table I are plotted. Lines are fit to the temperature versus power data points of the liquid-cooled FPGA at 2.4 mL/s and the stock air-cooled FPGA. An additional line shows the projected temperature with a powerful hypothetical ACHS with a junction-to-ambient thermal resistance of 0.25 °C/W.

As can be seen, at an FPGA power of 160 W, the microfluidic-cooled FPGA in this paper would have a die temperature of 31.5 °C, while the die cooled with the hypothetical high-performance ACHS would have a temperature of 61.3 °C. At 300 W, these temperatures would be 40.6 °C and 96.3 °C, respectively.

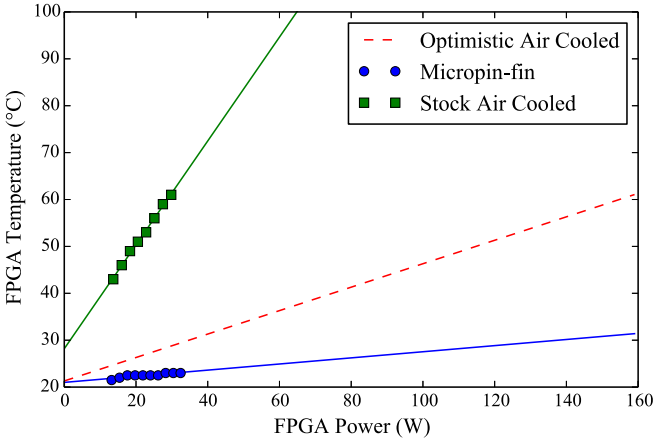

Fig. 8. Die temperature versus die power.

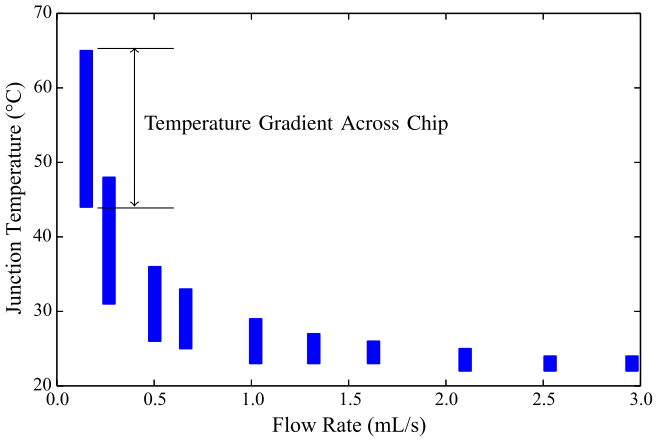

Fig. 9. Die temperature versus flow rate. Die temperature is a range representing the measurements with the temperature diode near the inlet and near the outlet.

A. Variable Flow Rate Testing

As flow rate through a micropin-fin heat sink increases, the convective thermal resistance decreases. This relationship between Nusselt number, which is proportional to heat transfer coefficient, and Reynolds number, which is proportional to flow rate, has been measured for a variety of micropin-fin geometries [7]–[9], [12].

The microfluidic-cooled FPGA was tested with several different flow rates, running the same pulse compression algorithm with all nine cores and an inlet temperature between 20.3 °C and 20.9 °C. The results can be seen in Fig. 9. As flow rate increases, the FPGA temperature decreases due to decreasing heating of the fluid as well as decreasing convective thermal resistance. At a maximum flow rate of 3.0 mL/s, a minimum average thermal resistance of 0.07 °C/W was achieved.

As flow rate increases, the temperature gradient from inlet to outlet due to heating of the fluid also decreases. For a given power, the temperature gradient across the chip is approximately equal to the temperature rise of the fluid, which is related to flow rate as $\Delta T \propto 1/\dot{m}$, where ΔT is the temperature rise of the fluid and $\dot{m}$ is the mass flow rate.

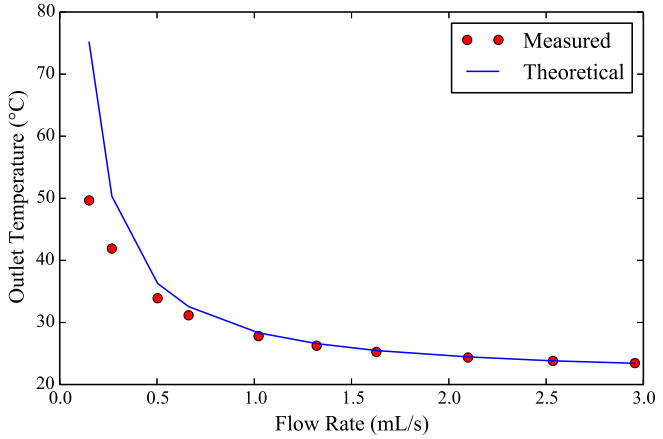

Fig. 10. Measured outlet temperature and theoretical outlet temperature (no heat loss) versus flow rate.

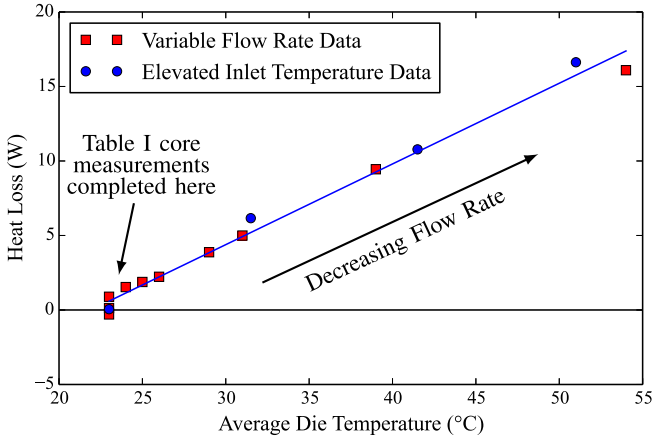

Fig. 11. Heat loss versus average die temperature.

The measured outlet water temperature and the predicted outlet water temperature are plotted in Fig. 10. The difference in the measured and calculated outlet water temperatures is due to heat loss through alternate heat paths to ambient air, such as the board and tubes.

In order to quantify heat loss to the surrounding ambient air, heat loss was quantified as

$$Q_{\text{loss}} = Q_{\text{in}} - \dot{m} C_p \left(T_{\text{out}} - T_{\text{in}}\right) \tag{1}$$

where Q_{in} is the measured power of the FPGA chip, $\dot{m}$ is the water mass flow rate, C_p is the specific heat of water, and T_{out} and T_{in} are the measured outlet and inlet water temperatures, respectively. C_p is a relatively weak function of temperature and was taken to be 4.18 J/(°C g). The density of the water was taken to be 1 g/mL for the purposes of converting measured volumetric flow rate to mass flow rate. The heat loss is plotted versus average die temperature in Fig. 11. Data points were used from this variable flow rate experiment as well as the elevated inlet temperature experiment presented in Section III-B.

Heat produced on the FPGA die has many thermal paths: through the MFHS and into the liquid, or through the package, board, etc., to the surrounding ambient air. When the MFHS is operated with a high flow rate and the coolant is near the temperature of ambient air, as is the case in the left side of Fig. 11, the majority of the die heat is captured in the fluid. If the efficacy of the heat sink is limited through a restricted flow rate, or an elevated inlet temperature, the die temperature rises relative to the surrounding ambient air and more heat is lost through these alternate heat paths to the ambient air. A higher ambient temperature, reduced airflow around the board, and insulation would all increase the fraction of heat captured by the coolant (and increase die temperature).

After fitting a line to the points in Fig. 11, the slope can be used to calculate the thermal resistance from the FPGA die to the ambient air through the board, tubes, etc. It was calculated to be 1.8 °C/W for this test setup.

Pressure drop was also measured as a function of flow rate while running all nine computing cores, which can be seen in Fig. 12. These pressure measurements were made outside

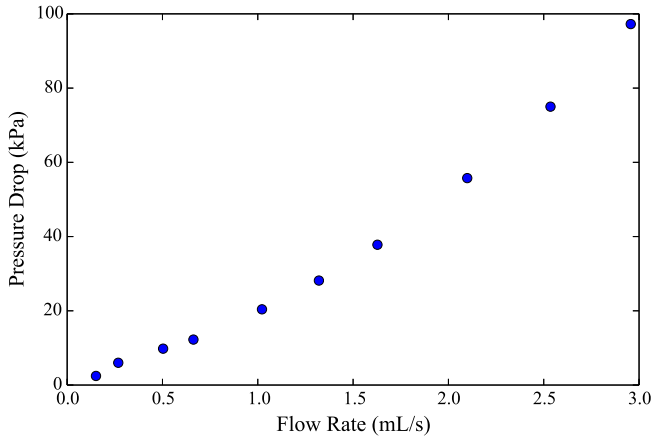

Fig. 12. Pressure drop versus flow rate.

of the chip and therefore include pressure drop across the inlet and outlet ports. A maximum pressure drop of 100 kPa was set as a conservative limit in order to prevent fluid leakage. A much higher pressure drop could be sustained with improved cap bonding [13].

B. Elevated Inlet Temperature Testing

The FPGA was also tested with an elevated inlet temperature, varying from 21 °C to 50 °C at a flow rate of 3.1 mL/s. These temperatures can be seen in Fig. 13. As expected, the FPGA die temperature tracks the water inlet temperature very closely, with an average junction temperature rise above inlet of 2.1 °C and 0.8 °C at average inlet temperatures of 20.9 °C and 50.1 °C, respectively. Temperature rise above inlet (and hence apparent junction-to-inlet thermal resistance) decreases with increasing temperature due to increased heat loss to the surrounding ambient air, which was 19.7 °C.

Pressure drop is also plotted as a function of inlet temperature in Fig. 14. As water temperature increases, its viscosity decreases, leading to the downward trend in pressure drop versus water temperature shown in Fig. 14.

C. Clock Speed

In addition to increasing performance through increased silicon utilization, improved cooling can also benefit performance

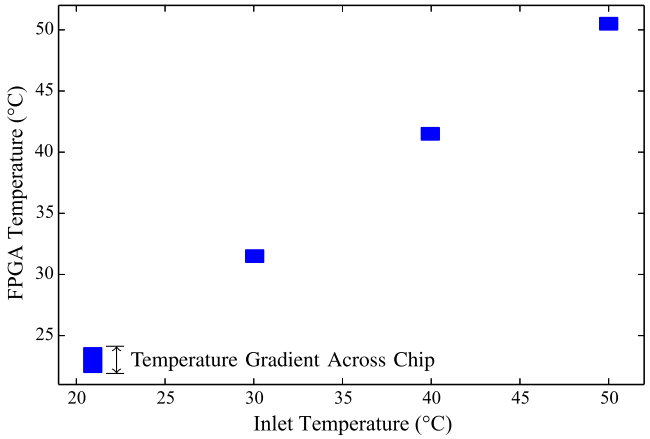

Fig. 13. FPGA temperature range versus inlet temperature.

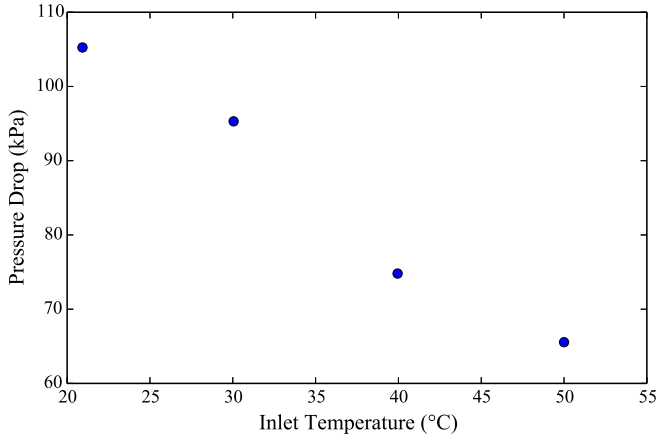

Fig. 14. Pressure drop versus inlet temperature.

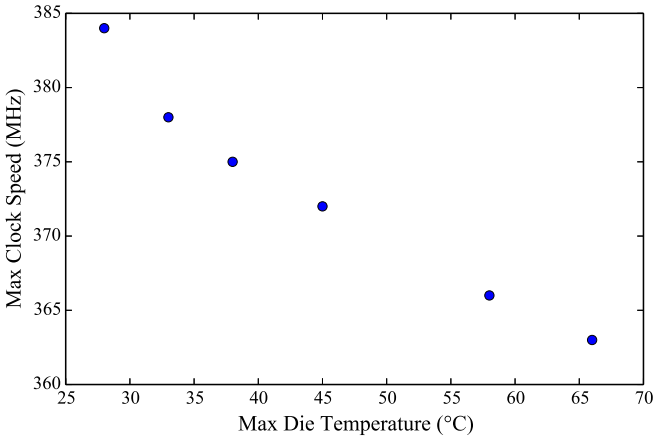

Fig. 15. Maximum FPGA clock speed without glitches versus maximum die temperature.

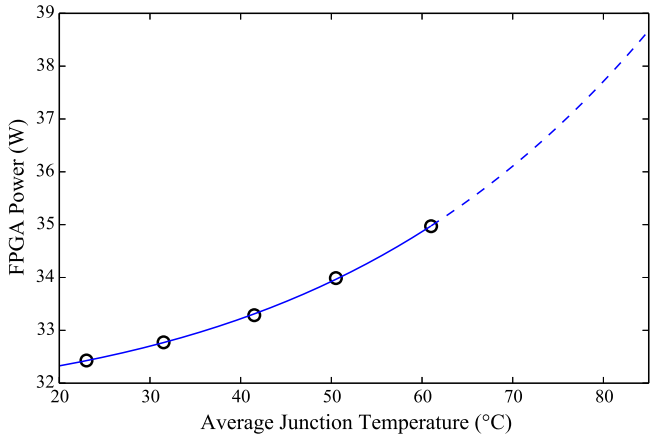

Fig. 16. FPGA power versus average die temperature.

in terms of clock speed. Both the transistors and interconnects experience enhanced performance with decreased temperature in planar bulk CMOS. Transistor threshold voltage and mobility tend to decrease as temperature increases [14]. Although decreased threshold voltage partially counteracts decreased mobility, a net decrease of drive current is observed by Lin *et al*. [14] in simulations of a 45-nm technology. Lin *et al*. [14] observed a 9% increase in delay time between simulations of a nine-stage inverter chain at 25 °C and 125 °C. The dependence of total critical path delay can, however, vary widely depending on chip design and process technology.

In order to test the dependence of maximum clock frequency on die temperature with the microfluidic-cooled FPGA, temperature was varied by varying flow rate with an inlet temperature between 24.3 °C and 24.7 °C. Due to current limitations from the on-board VRM, seven of the nine cores were run. The outputs from all cores were monitored through the Altera Signaltap tool in order to detect glitches that occurred in the output waveforms. The maximum clock speeds at which all seven cores operated with no glitches can be seen in Fig. 15 as a function of the die temperature measured on the side of the chip closest to the outlet. Decreasing the maximum die temperature from to 66 °C to 28 °C yielded an improvement of 21 MHz, a 6% improvement in clock speed (with an accompanying increase in power).

D. Die Power

Chip power consists of dynamic power, which has little dependence on temperature, and static/leakage power, which comes from several components, such as subthreshold leakage, gate leakage, and reverse bias junction current. Subthreshold leakage current tends to be the most significant temperature dependent component of the power [15], [16] and is given by [17]

$$I_{\rm ds} = \mu_0\ C_{\rm ox}\frac{W}{L}(m-1)\upsilon_T^2 {\rm e}^{(V_{gs}-V_{\rm th})/m\upsilon_T} \times (1-{\rm e}^{-V_{\rm ds}/\upsilon_T}) \quad (2)$$

where μ_0 is the zero bias mobility, $C_{\rm ox}$ is the gate oxide capacitance, W/L is the channel width to length ratio, m is the subthreshold swing coefficient, $V_{\rm th}$ is the threshold voltage, and υ_T is the thermal voltage, given by $\upsilon_T = k_b T/q$.

The total FPGA power versus average die temperature is plotted in Fig. 16. Die temperature was varied by varying inlet temperature at a constant flow rate of 3.1 mL/s since this provided nearly uniform die temperatures (Fig. 13). Due to the temperature-dependent leakage power, the measured total FPGA power increases by 2.6%, 4.8%, and 7.8% at 41.5 °C, 50.5 °C, and 61 °C relative to power dissipation at 23 °C. A trend curve using a first-order approximation of (2) is also

shown in Fig. 16. From an efficiency standpoint, the measured increase in FPGA power at elevated temperatures provides another strong motivation for effective cooling.

IV. Conclusion

In this paper, a micropin-fin heat sink was etched into the back side of an Altera Stratix V FGPA. The FPGA was tested with a pulse compression algorithm to demonstrate functionality and perform thermal benchmarks. Die temperature and power were measured as a function of flow rate and inlet temperature. An average junction-to-inlet thermal resistance of 0.07 °C/W was achieved at a flow rate of 3.0 mL/s and pressure drop of 97 kPa. This thermal resistance is sufficiently low to cool future generations of FPGAs and other high heat flux processors. The FPGA was also cooled with inlet water temperatures up to 50 °C, enabling high efficiency through heat exchange directly to ambient air, or waste heat reuse.

Future work may focus on both enhancement of the MFHS and the benchmarking algorithm loaded on to the FPGA. Pressure drop and thermal resistance could be improved through optimization of the micropin-fin heat sink and ports. The FPGA offers an opportunity to benchmark the performance gained through increased clock speed and chip utilization with many algorithms and architectures.

V. Acknowledgment

The authors would like to thank D. Ibbotson, K. Chandrasekar, and V. Mahadev of Altera Corporation for their help with this paper.

References

[1] D. B. Tuckerman and R. F. W. Pease, "High-performance heat sinking for VLSI," *IEEE Electron Device Lett.*, vol. EDL-2, no. 5, pp. 126–129, May 1981.

[2] S. Zimmermann, M. K. Tiwari, I. Meijer, S. Paredes, B. Michel, and D. Poulikakos, "Hot water cooled electronics: Exergy analysis and waste heat reuse feasibility," *Int. J. Heat Mass Transf.*, vol. 55, pp. 6384–6390, Nov. 2012.

[3] T. E. Sarvey *et al.*, "Embedded cooling technologies for densely integrated electronic systems," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Sep. 2015, pp. 1–8.

[4] H. Oh, J. M. Gu, S. J. Hong, G. S. May, and M. S. Bakir, "High-aspect ratio through-silicon vias for the integration of microfluidic cooling with 3D microsystems," *Microelectron. Eng.*, vol. 142, pp. 30–35, Jul. 2015.

[5] Y. Zhang, L. Zheng, and M. S. Bakir, "3-D stacked tier-specific microfluidic cooling for heterogeneous 3-D ICs," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 3, no. 11, pp. 1811–1819, Nov. 2013.

[6] T. Brunschwiler *et al.*, "Heat-removal performance scaling of interlayer cooled chip stacks," in *Proc. 12th IEEE Thermal Thermomech. Phenomena Electron. Syst. (ITherm)*, Jun. 2010, pp. 1–12.

[7] R. Prasher *et al.*, "Nusselt number and friction factor of staggered arrays of low aspect ratio micropin-fins under cross flow for water as fluid," *ASME Trans. J. Heat Transfer*, vol. 129, no. 2, pp. 141–153, 2007.

[8] T. Brunschwiler *et al.*, "Interlayer cooling potential in vertically integrated packages," *Microsyst. Technol.*, vol. 15, no. 1, pp. 57–74, Aug. 2008.

[9] A. Kosar and Y. Peles, "TCPT-2006-096.R2: Micro scale pin fin heat sinks—Parametric Performance evaluation Study," *IEEE Trans. Compon. Packag. Technol.*, vol. 30, no. 4, pp. 855–865, Dec. 2007.

[10] J. Karttunen, J. Kiihamaki, and S. Franssila, "Loading effects in deep silicon etching," *Proc. SPIE*, vol. 4174, pp. 90–97, Sep. 2000.

[11] I. W. Rangelow, "Critical tasks in high aspect ratio silicon dry etching for microelectromechanical systems," *J. Vac. Sci. Technol. A, Vac. Surf. Films*, vol. 21, no. 4, pp. 1550–1562, 2003.

[12] T. E. Sarvey, Y. Zhang, Y. Zhang, H. Oh, and M. S. Bakir, "Thermal and electrical effects of staggered micropin-fin dimensions for cooling of 3D microsystems," in *Proc. IEEE Intersoc. Conf. Thermal Thermomech. Phenomena Electron. Syst. (ITherm)*, May 2014, pp. 205–212.

[13] D. C. Woodrum, T. Sarvey, M. S. Bakir, and S. K. Sitaraman, "Reliability study of micro-pin fin array for on-chip cooling," in *Proc. 65th IEEE Electron. Compon. Technol. Conf. (ECTC)*, May 2015, pp. 2283–2287.

[14] S. C. Lin and K. Banerjee, "Cool chips: Opportunities and implications for power and thermal management," *IEEE Trans. Electron Devices*, vol. 55, no. 1, pp. 245–255, Jan. 2008.

[15] A. Agarwal, S. Mukhopadhyay, A. Raychowdhury, K. Roy, and C. H. Kim, "Leakage power analysis and reduction for nanoscale circuits," *IEEE Micro*, vol. 26, no. 2, pp. 68–80, Mar. 2006.

[16] Y. Liu, R. P. Dick, L. Shang, and H. Yang, "Accurate temperature-dependent integrated circuit leakage power estimation is easy," in *Proc. Design, Autom. Test Eur. Conf. Exhibit.*, Apr. 2007, pp. 1–6.

[17] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits," *Proc. IEEE*, vol. 91, no. 2, pp. 305–327, Feb. 2003.

Thomas E. Sarvey received the B.S. degrees in physics and computer engineering from the University of Maryland, College Park, MD, USA, in 2012. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the Georgia Institute of Technology, Atlanta, GA, USA.

His current research interests include densely integrated 2.5-D and 3-D electronic systems and their enabling thermal technologies.

Yang Zhang (S'13) received the B.S. degree in microelectronics and math (double major) from Peking University, Beijing, China, in 2012. He is currently pursuing the Ph.D. degree in electrical engineering with the Georgia Institute of Technology, Atlanta, GA, USA.

Colman Cheung received the B.S. degree in electrical engineering from the University of Illinois at Carbondale, Carbondale, IL, USA, in 1985, and the M.S. degree in electrical engineering from San Diego State University, San Diego, CA, USA, in 2008.

He started his career as a Digital Design Engineer in computing and networking. He is currently with Altera (part of Intel), San Jose, CA, USA, as a System and Design Engineer in signal processing and communications.

Ravi Gutala is a Senior Member of Technical Staff with the Programmable Solutions Group at Intel, San Jose, CA, USA. He has over 25 years of experience in the semiconductor industry and performed pioneering work in high-endurance flash memory design. He currently has 18 granted patents in the area of circuit/memory design.

Arifur Rahman (SM'12) received the B.S. degree in electrical engineering from the Massachusetts Institute of Technology, Cambridge, MA, USA, the M.S. and Ph.D. degrees in electrical engineering from the New York University (NYU) Polytechnic School of Engineering, Brooklyn, NY, USA, and the M.B.A. degree from Santa Clara University, Santa Clara, CA, USA.

He was at Xilinx, San Jose, CA, USA, Agere Systems, Allentown, PA, USA, Lattice Semiconductor, Allentown, and the NYU Polytechnic School of Engineering. During his tenure at Altera, San Jose, CA, USA, and Xilinx, he contributed to the deployment of 2.5-D/3-D integrated circuit technologies in field programmable gate arrays. He is the Director of System Planning with the Intel Programmable Solution Group, San Jose, CA, USA. He has authored more than 35 papers and has been granted 75 patents.

Aravind Dasu received the Ph.D. degree in electrical engineering from Arizona State University, Tempe, AZ, USA.

He is a Principal Investigator with the CTO Office, Intel's Programmable Solutions Group, San Jose, CA, USA. His current research interests include reconfigurable computing.

Muhannad S. Bakir (SM'12) received the B.E.E. degree in electrical engineering from Auburn University, Auburn, AL, USA, in 1999, and the M.S. and Ph.D. degrees in electrical and computer engineering from the Georgia Institute of Technology (Georgia Tech), Atlanta, GA, USA, in 2000 and 2003, respectively.

He is currently a Professor with the School of Electrical and Computer Engineering at Georgia Tech. His current research interests include 3-D electronic system integration, advanced cooling and power delivery for 3-D systems, biosensors and their integration with CMOS circuitry, and nanofabrication technology.

Dr. Bakir was a recipient of the 2013 Intel Early Career Faculty Honor Award, the 2012 DARPA Young Faculty Award, and the 2011 IEEE CPMT Society Outstanding Young Engineer Award. He was an Invited Participant in the 2012 National Academy of Engineering Frontiers of Engineering Symposium. In 2015, he was elected by the IEEE CPMT Society to serve as a Distinguished Lecturer and was an Invited Speaker at the U.S. National Academies Frontiers of Sensor Science Symposium. He and his research group have received more than 20 conference and student paper awards including six from the IEEE Electronic Components and Technology Conference, four from the IEEE International Interconnect Technology Conference, and one from the IEEE Custom Integrated Circuits Conference. His group was awarded the 2014 Best Paper of the IEEE TRANSACTIONS ON COMPONENTS, PACKAGING, AND MANUFACTURING TECHNOLOGY in the area of advanced packaging. He is an Editor of the IEEE TRANSACTIONS ON ELECTRON DEVICES and an Associate Editor of the IEEE TRANSACTIONS ON COMPONENTS, PACKAGING, AND MANUFACTURING TECHNOLOGY.