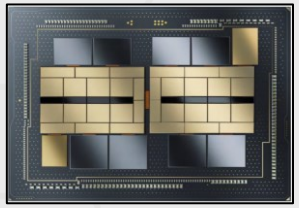
Silicon Integrated Microfluidic Cooling for High Power 2.5D FPGA

Sreejith Kochupurackal Rajan, Ankit Kaul, Muhannad S. Bakir

School of Electrical and Computer Engineering, Georgia Institute of Technology

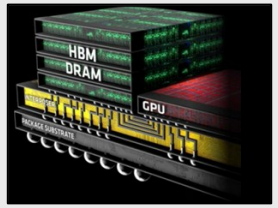
Georgia Tech | Center for Research into Novel Computing Hierarchies

Heterogeneous Chiplet-based Future of Compute



Intel Ponte Vecchio GPU

Source: Intel

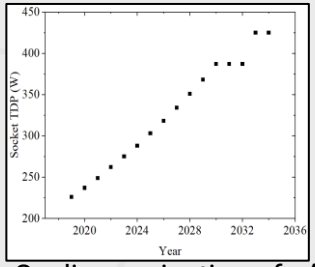


HBM DRAM

Source: AMD

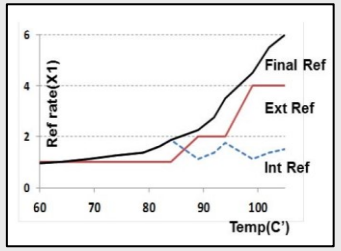
- Slowing of Moore's law is causing a shift to chiplet-based systems
- 2.5D packages are a promising option due to high inter-chiplet bandwidth

Dense Integration: Thermal Challenges



Scaling projection of CPU socket powers

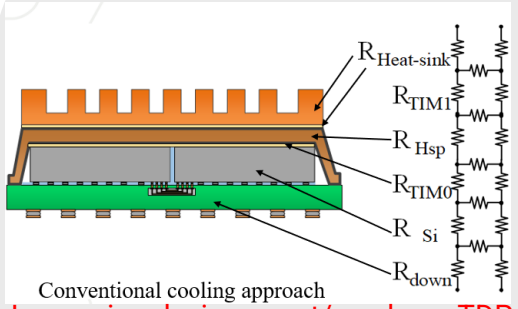
Source: IRDS 2020



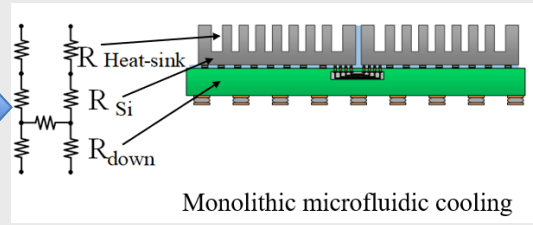
Adaptive Refresh considering Temperature (ART) for HBM

- Aggregate package powers increasing
- Dense integration can also lead to thermal coupling
- Left unchecked, this can lower overall system performance

Microfluidic cooling for 2.5D: Concept



Conventional cooling approach



Monolithic microfluidic cooling

Increasing device count/ package TDP (Higher absolute temperatures)

Die-to-die thermal coupling

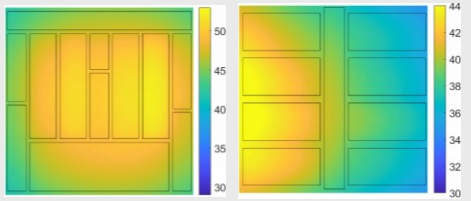
Heterogeneity in die profiles (mismatch representing more TIM etc.)

Low R_{th} to deal with increasing package TDPs

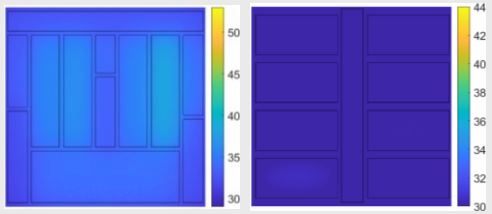
Reduced die-to-die thermal coupling

Ultra-small form-factor heat sinks

Microfluidic cooling for 2.5D: Finite Volume Modeling



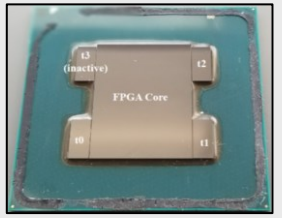
Transient temperature contours with air-cooling



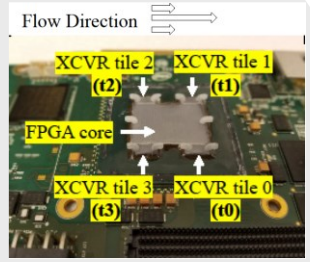
Transient temperature contours with microfluidic cooling

- Reduction in aggregate device temperatures
- Reduction in transient and steady state thermal coupling

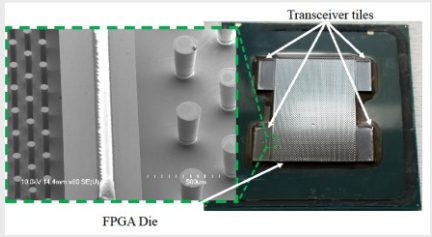
Experimental Demonstration



De-lidded Stratix 10 GX FPGA



Post heatsink assembly



Close-up of the heatsink

	Stock	Monolithic with inlet at 52.5°C
Total power (W)	114.32	172.06 (5% increase in number of compute cores and a further 5% increase in core clock frequency)
FPGA core Temp.	59.98	57.37
XCVR 0 Temp.	63.02	63.68
XCVR 1 Temp.	71.09	58.26

Performance Comparison

- Increased clock-speeds and resource utilization with better cooling

Georgia Tech