Guidelines on Thermal Management Solutions for Modern Packaging Technologies – A Review

A. Fodor, G. Chindris, D. Pitica and R. Jánó
Applied Electronics Department
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
alexandra.fodor@ael.utcluj.ro

***Abstract*— Thermal management for modern electronic systems can follow two fundamental paths: one concentrates design efforts towards low power/high efficiency electronic circuits/components, the other implies, where the previous one reaches its limits, optimizing thermal transfer in the entire system at chip level, package level, and PCB/assembly level. Thus, exploring all methods of improving heat transfer at chip/package/assembly levels can generate a guideline for common/specific approach that can be used at the earliest stage of a design.**

The current research gives an overview of latest packaging technology along with required thermal management measures and proposes a consistent methodology of Design for Thermal Management.

Keywords—Thermal, management, packaging, cooling solutions

I. Introduction

Semiconductor technology, coupled with modern, miniaturized packaging form the backbone of today's high performance electronic systems [1]. New trends in device packaging are emerging, bringing more challenges with them especially regarding their thermal management. Semiconductor packaging is moving into the 3rd dimension - driven by the need for higher levels of integration, improved electrical performance, or reduction of timing delays, the need for shorter vertical interconnects is forcing a shift from 2D to 2.5D and 3D package designs[1], as seen in figure 1.

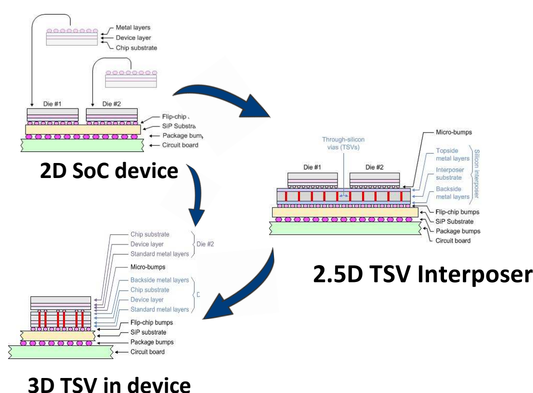

Fig. 1. Device packaging migration, from 2D to 3D [2]

The purpose of this article is to deliver a review of the existent cooling techniques, starting with the most established ones, for 2D packaging, and after that, continuing with the new trends in electronics cooling for 2.5D and 3D packaging.

II. 2D Packaging and Cooling Solution

First, let's see which packages fall into 2D packages category, and what are the cooling techniques specific to each package.

A. MCMs

Multi-Chip Modules were first developed in the beginning of the '90s, and their accepted definition is that an MCM is a collection of more than one bare die on a common substrate. Generally, the MCMs functions are implemented to perform operations done by various electronic packages: propagation of power and signals and, ensuring mechanical support, encapsulation, (to provide protection against dust and chemicals), and a capability of removing the dissipated heat generated by the chips.

Thermal management of Multi-chip Modules can be achieved using the following three techniques: a thermal conduction module with Direct Solder Attach Cooling (DiSAC), a dual layer thermal interface (TIM) thermal design, and a MCM design with Small Gap Technology (SGT) and a hermetic seal, these techniques being thoroughly described in [7].

B. SoCs

A System-on-Chip is an integrated circuit that comprises all components of an electronic system on a single chip. It may contain digital, analog, mixed-signal, radio-frequency functions and hardware accelerators, all on a single chip substrate. Because an SoC includes both the hardware and the software, it uses less power, has better performance, requires less space and is more reliable than multi-chip systems [2].

A cooling solution that can successfully be applied to SoC devices is the usage of Thermal Copper Pillar Bumps (CPB) which involves the integration of a thin-film thermoelectric material into solder bumps for flip chip packaging. This enables active thermal management – and power generation – right at the chip's surface using an industry-accepted manufacturing approach to ensure seamless implementation. Unlike conventional solder bumps, which provide an electrical

path and a mechanical structure, each thermal CPB acts as a solid-state heat pump on a micro-scale [11].

C. SiPs

SiPs (System-in-Package) can be described as multiple bare dice and/or chip-scale package devices mounted on a common substrate. The substrate and its components are built into a single package. The SiP performs all or most of the functions of an electronic system. Dies containing integrated circuits may be stacked vertically on a substrate. They are internally connected by fine wires that are bonded to the package. Alternatively, with flip chip technology, solder bumps are used to join stacked chips together [2].

The main advantage that this IC brings over SoCs, is that SiPs can contain analog, digital and radio frequency dice in the same package, and each die is implemented using the most suitable technology process.

In thermal management of a SiP device, the things that need to be considered are the basic elements of package assembly, and also design aspects related to the overall operational requirements of the system and manufacturing process, as well as supply chain management and test. Classical cooling techniques do apply, but a more efficient one can be the usage of AlSiC Flip-chip lids. Due to thermally induced stress induced by attaching a microprocessor lid to the printed circuit board, choosing materials that have compatibility in regards to their thermal expansion coefficients is required. This is not invariably possible, as there are various composing materials in these assemblies with differences in CTE (Coefficient of Thermal Expansion) values, meaning that a "one consistent CTE solution" would be difficult to manage. An alternate approach is to increase the stiffness of the system. This can be done by choosing stiffer material or by increasing the thickness of the materials in the system. Increasing thickness for increased stiffness may not meet the critical height requirements of the system. Increased material stiffness without an increase in material density (decreased material density is more desirable) would be the solution. AlSiC, which has a density of 3 g/cm3 – 1/3 the density of copper, has a stiffness value that is greater than the copper equivalent. And since AlSiC has significantly lower density increasing thickness to achieve greater assembly stiffness is possible without significant weight penalty, being 1/3 lighter than copper [4]. AlSiC Flip-chip lids for SiP based applications bring higher thermal and a general reliability, along with a higher packaging performance and flexibility.

Next, 2.5D and 3D packaging will be presented, followed by specific cooling techniques, applicable for both technologies.

III. 2.5D PACKAGING

2.5D packaging and interconnection technology is a rising semiconductor packaging technology that grants reduction of costs and a general reliability increase over 3D packaging technology. 2.5D packaging technology is a fast growing one which permits the combination of homogenous and non-homogenous chips on a single interposer bringing an addition in performance and miniaturization. In some implementations, 2.5D semiconductor packaging mounts one or more semiconductor dies on the undercarriage of the interposer, thereby positioning those semiconductor dies in a relatively tight cavity between the interposer and substrate. While this achieves a more compact semiconductor package configuration, it also raises challenges related to thermal management, as it may be quite difficult to control the operating temperature of semiconductor dies mounted beneath the interposer due to space and/or airflow limitations. Accordingly, mechanisms for controlling the operating temperature of chips mounted beneath an interposer in a 2.5D packaging configuration are desired.

The major difference between a conventional 2D IC and a 2.5D, is that 2.5D IC has a silicon interposer located between the substrate and the dice, where the silicon interposer has through-silicon vias (TSVs) which connect the metallization layers on its upper and lower surfaces. In this case, the dice are attached to the silicon interposer using micro-bumps, which are approximately 10um in diameter. The silicon interposer is attached to the substrate through typical flip-chip bumps, which will be approximately 100um in diameter. The tracks on the silicon interposer's topside and backside metal layers are created using the same processes as the tracks on the silicon chips [2].

The advantage 2.5D IC packaging technology brings is that it's an additional step from the long-established 2D IC technology, and brings a massive development in capacity and performance. There are yield advantages as well, as it is easier to manufacture multiple small dice than a single large one.

Lithography is a key component of 2.5D integration; nonetheless the requirements for interposer exposure systems differ significantly from the requirements of front-end lithography tools. In particular, 2.5D lithography faces specific challenges in regards to resolution, overlay, sidewall angle, exposure field size, depth of focus, warped substrate handling, and backside alignment. As with all middle- and back-end processes, interposer manufacturing must be extremely cost efficient and high yielding [8].

IV. 3D PACKAGING

A "three dimensional integrated circuit" (3D IC) is defined as an integrated circuit assembled by stacking silicon wafers and/or dies. The interconnections are vertical using TSVs, obtaining the behavior of a single device, for performance improvements at reduced power and reduced footprint size. A 3D IC is just one of the integration schemes that uses the z-direction to accomplish electrical performance benefits.

The technique can be used with a stack of similar die (the homogenous approach), for example to build 'memory cubes', stacks of pure memory die with the controller logic on a separate die at the bottom. It is also being developed for use with a mix of different die (the heterogeneous approach), although this is more complex and challenging [6].

3D integration is a newly proposed design approach for reducing the constraints regarding the delay and power consumption of interconnects. Still, this increased level of integration can bring new limitations and layout challenges, including the challenges associated with higher temperatures. The power dissipation leads to critical temperature increase on

solder joints; it diminishes the previously reduced thermo-mechanical reliability of the system. Stacked IC technologies, more than Moore, developed to manage recent limitation of IC density reduction, lack heat dissipation schemes. Important progress in thermal management is greatly needed for these packages; it becomes even more demanding as miniaturization in IC packaging continues. Electronic packages include several different materials with dissimilar thermal and mechanical properties. Thermal stresses appear due to the discrepancies between coefficients of thermal expansion at the junction of separate materials. Accelerated testing is introduced and its purpose is to foresee the reliability of a 3D package in known conditions. The chance of a certain breakdown depends on the operation and the functioning environment. An acceptable level of reliability is necessary in every case, in order to attain customer satisfaction. The material compatibility is fundamental for a high-reliability stacked design.

The 3D/SiP approach raises the need for thermal simulations, as a stacked structure has multiple extensive heat sources in a reduced volume that has to be considered. The underfill material between interposer layers has low thermal conductivity and therefore, the removal of heat from the inner parts of a module may require the creation of thermal paths or changing the layout such that the dice producing most of the heat are located on the outmost interposers. Interposers should have a great thermal conductivity. Thinner interposers and dice reduce thermal resistance and enhance of a package. Thermal vias can be used also to reduce junction-to-air thermal resistance [3].

A complete thermal analysis implies the application of basic heat transfer theory. The differences are situated in covering issues that are unique to the on-chip circumstances: for instance, on-chip geometries are greatly rectilinear in essence and contain rectangular geometric symmetries; the major heat sources, the devices, situated in 3D tier, and the points that a user is usually interested in examining temperature are inside the device layers.

Technology scaling has caused the feature sizes to shrink continuously, whereas interconnects, unlike transistors, have not followed the same trend. In the nanometer era, a larger portion of the total chip capacitance comes from interconnects. With the introduction of vias and repeaters to compensate for the performance loss of the long wires, the interconnect power consumption rises dramatically. Designing 3D integrated circuits is one of the recently proposed approaches to overcome the problems associated with interconnects. When components are placed on a 3D architecture, the length of interconnects and the large power consumption associated with them can be reduced. However, 3D integration introduces challenges due to the high power density resulting from the placement of computational units on top of each other. High power densities are already a major concern in 2D circuits, and in 3D systems the problem is even more severe.

Thermal hot spots cause a rise in cooling costs, negatively impact reliability and reduce performance. The significant increase in cooling costs necessitates designing for temperature margins that are lower than the worst-case. Hot spots accelerate failure mechanisms such as stress migration, dielectric breakdown and electro migration, which cause permanent device failures. Leakage is exponentially linked to temperature, and an incremental feedback loop exists between leakage and temperature, which can cause high increases in temperature and damage the circuit if it remains out of control. High temperatures also harmfully affect performance, as the actual operating speed of devices decreases as temperature increases.

Addressing thermal hot spots solely is not enough to achieve better reliability, as temperature gradients in time and space define device reliability at moderate temperatures. The failure rate due to thermal cycling rises with the magnitude and frequency of temperature cycles [9].

V. Cooling solutions for 2.5D and 3D packaging

In [30] a thermal management solution for 2.5D semiconductor packaging is presented. The method states that lower semiconductor dies can be cooled by thermally coupling the lower semiconductor dies to a heat sink positioned above the interposer, to an upper semiconductor die, to a heat sink affixed beneath a substrate or to free-flowing air circulating above the interposer or beneath the substrate. The thermal coupling can be achieved using heat pipes, thermal vias, or other conductive passage ways.

In classical cooling methods, a heat sink or a microchannel cooler is placed at the top of the chip to dissipate the generated heat. In designing a 3D chip stack with an image sensor at the top and underlying A/D converters and signal processing circuits, the cooling methods cannot be located at the top.

Chip level spot cooling is presented as another approach, in [5]. The solution for the spot cooling has two broad forms: at the silicon level and at the packaging level.

A. Silicon Level

Lots of solutions have been explored at the silicon level. Among these is the usage of micro-channel heat exchangers (MCHEs), thin-film thermoelectric coolers (TECs), micro heat pipes, and localized copper bumps to increase heat spreading on a large surface area, or in the case of an MCHE, conveying it to a different location. Figure 2 shows a stacked MCHE that is used for cooling a silicon substrate.

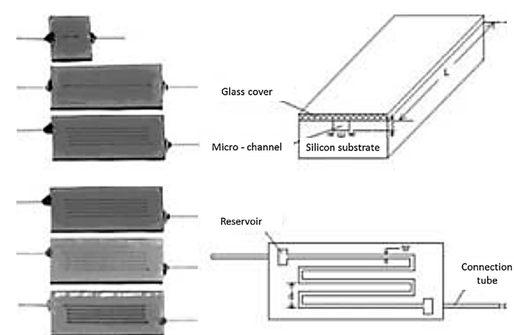

Fig. 2. Stacked micro-channel heat exchanger for substrate cooling [10]

The accomplishment of these techniques (except for a few of highly customized applications, such as larger scale computers) has been unsatisfactory and weak. TECs, bumps, micro heat pipes and such, have had little success in reducing

the peaks witnessed as a result of the power density variation. Often, because of the aspect ratio or the area, the heat flux exceeds 400-500 W/cm^2. In some high-power applications, it exceeds 1-2 kW/cm^2. With such great heat fluxes, the limits of this technology can be foreseen. A noteworthy approach to this problem has been the electrical control of the chip. Dynamic thermal management (DTM) has been studied as a technique for controlling CPU power dissipation. DTM refers to a range of possible hardware and software tactics that work dynamically, at run-time, to control a chip's working temperature.

Usually, the packaging and fans for a CPU or computer are designed to keep a safe operating temperature even when the chip is dissipating the maximum amount of power possible for a persistent period of time, and therefore generating the very high amounts of thermal energy. DTM allows packaging engineers to design systems for a targeted continuous thermal value that is much closer to for-real average-case benchmarks. If a particular workload operates above this point for continued periods, a DTM response will work to decrease chip temperature.

In essence, DTM allows designers to concentrate on average thermal conditions in their designs, rather than worst-case conditions. The main goals of DTM can be stated as follows:

- to provide cheap hardware or software responses;
- to reliably reduce power;
- to impact performance as little as possible.

To better understand the process, the maximum amount of power is dissipated when all of the structures within the processor are active at maximum switching activity. In reality, the maximum power dissipation is limited by the software code that can maximize the usage and switching activity of the hardware.

Special maximum power benchmarks can be written to exploit the switching activity of the processor. These benchmarks are often relatively esoteric, perform no meaningful calculation, and dissipate higher power than 'real' programs.

Therefore, TM techniques could be used exclusively to target power levels seen in maximum power benchmarks and would hardly be invoked during the course of typical applications.

B. Package Level

Cooling solutions can be applied, at the package level, to the top of the silicon, or instead the component can be immersed in a heat transfer fluid. The goal of these methods is to rapidly remove the heat from the device or to provide a medium that can efficiently absorb the thermal peaks by improved thermal dispersion.

A number of techniques have been studied by many researchers in the field with varied success. Table 1 presents such techniques and their respective characteristics [10].

TABLE I. PACKAGE LEVEL COOLING TECHNIQUES – ADVANTAGES AND DISADVANTAGES.

Technology	Advantages	Disadvantages
Fan sinks with heat pipe (hybrid)	Compact, versatile	Reliability, space, limited to ambient temperature
Thermoelectric	Spot cooling	Reliability, low capacity
Liquid cooling	High surface heat transfer	Sealing, cost, maintenance, packaging
Direct immersion	High capacity	Cost, sealing, packaging
Refrigeration	Sub-ambient	Cost, power, packaging, space
Cryogenics	Super cooling	Cost, power, packaging

VI. CONCLUSIONS

In order to develop a thermally optimal electronic system, there are several key factors that need to be considered, even before starting the development process. Firstly, the architecture needs to be established, meaning that the design engineer should know what packaging technology will be used. Secondly, the appropriate cooling technique must be chosen, depending on the heat dissipation levels that the system could produce, and moreover, a combination of cooling techniques must be applied. Considering these factors, prior to the actual development of the electronic system, will result in creating a reliable, and optimal electronic system, in terms of thermal management.

ACKNOWLEDGMENT

This paper is supported by the Sectorial Operational Programme Human Resources Development POSDRU/159/1.5/S/137516 financed from the European Social Fund and by the Romanian Government.

REFERENCES

[1] H.K. Charles Jr., "Technology Advances and Globalization in Electronic and Electro-Optical Packaging", John Hopkins APL Technical Digest, Volume 28, Number 1, 2008.

[2] Maxwell C., „2D vs. 2.5D vs. 3D ICs 101" , EE Times Mag. April 2012

[3] Khan, Navas et al., "Development of 3-D Stack Package Using Silicon Interposer for High-Power Application", IEEE Transactions on Advanced Packaging, vol. 31, No. 1, 2008, pp. 44-50.

[4] M. Occhionero, „Microprocessor RoHS Requirements: Demands a Need for Lightweight High Stiffness AlSiC Microprocessor Lids", CPS Technologies, May 2011.

[5] Chip Scale Review, The International Magazine for the Semiconductor Packaging Industry, Volume 15, Number 3, May-June 2011.

[6] Maxfield, C. „Bebop to the Boolean Boogie: An Unconventional Guide to Electronics", Newnes, Elsevier Inc., Oxford 2009, ISBN 0-7506-7543-8.

[7] Azar K., „Multi-Chip Modulde Thermal Management", Qpedia Thermal management eMagazine, Volume 4, January 2011.

[8] Ruhmer K., „Lithography Challenges for 2.5D Interposer Manufacturing", ECTC, September 2014.

[9] Coskun A.K., Rosing T.S., Whisnant K., Gross K., „Temperature-aware MPSoC scheduling for reducing hot spots and gradients", Proceeding of.Asia South Pacific Des. Autom. Conf., January 2008, pp. 49-54.

[10] Kaveh Azar, Bahman Tavassoli, Qpedia Thermal Management – Electronics Cooling Book,Advanced Thermal Solutions, Volume 3, January 2009

[11] Riley G., „Bump Cooling", Nextreme Thermal Solutions, Dec. 2007