Impact of Process Parameters on Vacuum Fluxless Solder Reflow Performance in Backend Applications with Bump Pitch of 15µm and Below

Lei Jing, Vladimir Kudriavtsev, Taylor Nguyen, Jed Hsu, Tapani Laaksonen, Alvin Lin, Xinxuan Tan, Kay Song, Alex Chow, Chris Lane and Zia Karim
Yield Engineering Systems (YES)
3178 Laurelview Court, Fremont, CA 94538, USA
zkarim@yieldengineering.com

***Abstract*— In this paper, an experimental methodology was developed to characterize reflow performance for next-generation micro-bumps using bump shapes distinguished by the ratio of bump height to diameter. Existing technology deals with 50µm bumps; this study analyzed micro-bumps with a pitch of 15µm and below, correlating reflow performance to reflow process parameters.**

Keywords—Formic Acid Reflow, FOWLP, fine pitch micro-bumping process characterization.

I. INTRODUCTION

To date, two major approaches have been employed to conduct bump wafer reflow at 15µm pitch and below. While some industry leaders are working on hybrid bonding, others are pushing the state-of-the-art with ever-finer pitches. Extending micro-bumps to finer pitches leverages the existing solder/copper pillar infrastructure, and several foundries, OSATs and IDMs are currently working on fine-pitch bump technologies [1].

The effort was made to understand the requirements and development challenges for fine-pitch micro-bumps and to explore interactions through system parameters, multi-physics simulations, and digital processing methods. Extensive bump shape data at various combinations of soak and reflow time, process parameters, and temperatures were collected using x2500 microscope vertical cross-section photographs and were investigated to define the reflow regime.

II. EXPERIMENTAL

A. *Background*

The challenge faced in the equipment industry is the specific limitations of demo processes. In order to conduct demos, equipment manufacturers receive process wafers with devices on them (which we refer to as reference wafers) from prospective customers. These reference wafers are only available in limited numbers and cannot be sent out for advanced processing, characterization, or thermal analysis. They must be studied in-house and it is crucial to accurately predict their thermal performance and optical characteristics during subsequent processing. If the process window is missed by just a few degrees, the results will be unfavorable, particularly in fine-pitch applications. This becomes even more challenging when the wafer contains various patterns, each requiring its own optimal conditions. With limited wafers available, process tuning R&D becomes difficult. Our approach helps to overcome this challenge.

We have developed a new dual chamber processing tool designed to resolve the challenges faced by process engineers in the equipment industry. The chamber (Figs. 1(A,B)) can simultaneously heat and cool the wafer using a heating module that operates at atmospheric pressure to heat the wafer and a cold N_2 showerhead located in the lower chamber that cools the wafer quickly to provide a quenching effect and provide a specific amount of N_2 dilution for acid isolation. An exhaust ring surrounds the main chamber and pumps potential particles away from the wafer. The chamber operates at low pressure, between 0.1 torr and 300 torr, and approaches atmospheric pressure during the cooling phase.

The heating of the wafer (Fig.1 C) is accomplished with several short-wavelength infrared lamps above the wafer. The temperature of the filament at the operating power is about 2370 Kelvin (K). Normally the lamp filament temperature is 1200 – 1800 K during the temperature ramp phase. From Wien's displacement law, one can calculate the corresponding peak temperatures for blackbody radiation to be about 2.4 and 1.6 µm, respectively. The wafer temperature is about 423 and 523 K during the process. The peak wavelengths at these temperatures are about 6.8 and 5.5 µm respectively.

The description of the heating process is complicated by the wavelength and the temperature dependence of the emissivity of the wafer. The heating of the wafer as a function of time t can be described by the equation

$$mc_p \frac{dT}{dt} = \epsilon_s AQ - 2\epsilon_l A\sigma T^4 + 2\epsilon_l A\sigma T_a^4$$

where m is the mass of the wafer, c_p the specific heat of the wafer material, T the temperature, ϵ_s the wafer emissivity at the

incoming radiation wavelength, ϵ_l the emissivity at the wavelength corresponding to the temperature of the wafer, A the area of one side of the wafer, Q the irradiance of the electromagnetic radiation coming from the lamps, σ the Stefan-Boltzmann constant, and T_a the ambient temperature.

Several physical and mechanical principles were combined to achieve the desired effect, including low thermal inertia short wavelength infrared heating, precise thermal and temporal control of power for each heating component/lamp, optimized spatial distribution of heaters, wafer rotation to achieve circumferential averaging, optimized shapes of lamp reflectors to accentuate heating where needed, wafer rotation to create a tornado-like mixing vortex during the soak step, upper chamber isolation using a circumferential accessory through N_2 purge flow, and an exhaust manifold around the chamber to facilitate uniform departure of formate salts byproduct, chamber wall heating to control formic acid condensation and salt formation, linear radial injection of diluted formic acid mixture to sustain and enhance the mixing vortex, low pressure operation during formic acid (FA) soak to maximize diffusion, vent/purge pumping cycles to enhance transport in the interpillar space through pressure-driven convective action, and low-pressure reflow with convective action completely shut.

Temperature monitoring is a crucial component of the system. Several wafers with varying optical characteristics/ emissivity were designed and built, each with nine embedded K-type thermocouples (Fig1 C). To tune wafer uniformity, we introduced the concept of heater power ratios, in which the individual heater powers are scaled relative to a reference. Optimal heater power ratios were experimentally developed for each class of wafers. The wafer with thermocouples and lead wires is manually placed inside the process chamber, followed by a chamber pump down, allowing the monitoring wafer to match the thermal response of the reference wafer being processed. This monitoring wafer can be used for both thermal tuning and characterization and can serve as a carrier wafer for processing multiple coupons at various locations. Considerable effort was put into characterizing the measurement system calibration and accuracy, understanding that thermocouple measurements can inherently operate with ±1°C uncertainty. Despite this uncertainty, we were able to consistently differentiate wafer performance with at least 0.2°C resolution and tune heating power to improve performance with matching resolution.

By selecting the silicon wafer doping level we can match the test wafer and the customer wafer temperature behavior. The doping level of the silicon controls both the emissivity and the electrical resistivity of the wafer. Typical p-type silicon wafers sold for wafer handling purposes have a resistivity of 1 – 50 Ohm-cm. The resistivity of 1 ohm-cm corresponds to a p-type doping level of $1.5x10^{16}$ atoms/cm^3 and 50 ohm-cm to a p-type doping level of $2.7x10^{14}$ atoms/cm^3 [2]. A p-type doping level of 10^{19} atoms/cm^3 corresponds to a resistivity of 0.009 ohm-cm. The emissivity of a silicon wafer depends also on the wafer thickness. For a 350 µm thick silicon wafer with a p-type doping level of 10^{16} atoms/cm^3 the emissivity is less than 0.02 at 300K and less than 0.1 at 800 K between the wavelengths 1.6 5 µm [3]. The emissivity of 700 µm-thick p-type silicon wafers with a doping level of 10^{19} atoms/cm^3 has emissivity of 0.68 - 0.7 between the wavelengths 1.6 – 5 µm at 473 K [4]. Figures 2A and 2B, which are drawn from the data in Table 1 of reference [4], show the dependence of emissivity on the wavelength, temperature, and dopant level.

To achieve effective averaging of properties, wafer rotation was employed as one of the techniques. The criticality of thermal measurement metrology for repeatability was also recognized. The control of temperature in real-time, with minimal inertial effects, was accomplished by utilizing short wavelength lamps with tunable control. The thermal losses for various wafers were analyzed through simulations and non-uniform power compensation was implemented to maintain uniform temperature across the wafer.

The system can operate thermally in various control modes, including standard PID control, with the PID coefficients separately tuned for each wafer type. This mode uses a contact-type pin with an embedded fast-response thermocouple, which has a predetermined thermal lag. Therefore, if a wafer needs to be set to a specific temperature Tr, the control can be precisely adjusted for that. To ensure reliable and fast-response contact-type measurement, considerable effort went into its development. Several experimental measurements were conducted using different matching materials and interfaces. Additionally, computer simulations using ANSYS were performed to characterize thermal resistance at the interface, compute transient responses, and compare them with the experimental data. The second mode is to use one of the nine-point wafer thermocouples for control. This mode cannot be used in production but is helpful in process development when a precise wafer temperature is required. The control system then ensures that central or edge wafer temperatures are precisely met as specified. The third mode of operation involves specifying fixed heater powers or custom loading different temporal power profiles. This mode is especially useful if an optimal time-dependent power process of record is already present and low deviation is required [6].

B. Experimental Reflow Process

In a micro-bump state or in 3D stacking of semiconductor devices, diffusion of reactants such as formic acid is critical. By controlling the diffusion coefficient or diffusivity of the formic acid vapor, the reaction mechanism can be improved for better results. In this invention, diffusivity is improved by controlling the formic acid concentration, partial pressure of formic acid and temperature dependent gas mobility to enhance the soak process.

To improve process control and performance to meet urgent demand, on micro-sized solder bumps less than 15µm we introduce chemical vapor in low pressure chamber conditions from 0.1 to 300 Torr. Figure 3 describes process flow and Figures 4 (A, B) demonstrate temperature variation vs. time. The x-axis shows time in seconds, the left y-axis shows temperature in degrees Celsius, and the right Y axis shows chemical vapor flow timing and amount. Time intervals used here are for reference purposes only. The process starts with pumping the system to low vacuum (0.1Torr~300Torr) for an oxygen-free process environment. Chemical vapor with controlled concentration is applied to the wafer for solder surface oxide cleaning, and nitrogen is input into the chamber from the

dissipator. The chamber gas is controlled between 150ºC -180ºC in a formic acid environment. Chemical vapor is injected into the chamber through nozzles controlled by either a bubbler or vaporizer. The chemical mixture is disseminated over the substrate by diffusion and forced convection (Figure 5 A). In this transport, both convection (rotation) and mass diffusion play an important role. In Figures 5 (B, C), we illustrate the effect of low pressure on improving mass diffusion in a simulated region with bumps.

According to the Chapman-Enskog equation, the binary diffusion coefficient D_{ij} in the multicomponent mixture can be calculated as

$$D_{ij} = \frac{1.86\ x\ 10^{-3} T^{3/2} \sqrt{\frac{1}{M_1} + \frac{1}{M_2}}}{p\sigma_{12}^2 \Omega}$$

where p is pressure, σ is the Lennard-Jones collision diameter, M is molecular mass, T is temperature, and omega is the collision integral. So, with a decrease in chamber pressure diffusion increases proportionally. Diffusion also increases with an increase in temperature. For example, the diffusion coefficient of formic acid HCOOH at atmospheric pressure is 0.2 cm^2/s, at 2 torr is 77 cm^2/s, and at 0.5 torr is 311 cm^2/s, while the 300mm wafer area is 706 cm^2. Thus, the superior species transport performance of the low-pressure process is quite evident.

After the wafer is coated with the chemical vapor, the excess vapor is removed from the chamber with vacuum. The wafer support adjusts the wafer to the heating position, with the target temperature 220°C~250°C for the reflow process. After the reflow/soldering is completed, the lamps are turned off (or reduced to a safe idle power).

C. *Metrology Methodology*

While measurement companies are working to address the challenges of full wafer coplanarity measurements, we developed a practical and robust method for conducting our experiments and characterizing reflow performance. Our approach utilized the x2500 microscopy tool (Keyence VHX-7000), which has the limitation of not being able to fit a full wafer under the microscope and is largely limited to coupons or smaller pieces cut from a full wafer.

Our first method involved the manual digitization of photographed images using the digitization software WebplotDigitizer 4.6 (Figures 6 (A, B)). Each photograph, containing several solder bumps and pillars, produced digital (x,y) tabulated traces with approximately 25-35 data points defining the shape of the bump. These shapes represented vertical cross-sections, and the maximum width (Wm) and maximum height (Hm) were determined. Positioning offset corrections were performed, allowing all profiles to be plotted superimposed with matching statistical deviation and uniformity estimates (Fig. 6 B).

Our digitization analysis of vertical cross-sections proved extremely useful in providing accurate reflow shape measurements with a precision down to 0.25 μm. We were able to see a practical correlation between temperature differences at key locations, the H/D ratio and its deviations, as well as reflow temperature and duration. We also found that for fine pitches near 10 μm, optimal temperature needs to be selected with a precision close to 2 °C, and reflow time with a precision close to 5 sec.

Bump diameter variation at different focal planes is a measure of co-planarity [1,5]. Manual digitization is time-consuming and difficult for all possible process development coupons, so a faster technique was needed for quick overall assessments. Our second method utilized bump top view microscopic photography, which provides a co-planarity visual metric to the operator. The microscope was focused on a specific vertical depth within the bump height, measuring bump diameter variation, but not height variation. The results were immediately apparent, allowing us to easily differentiate between reflows that were within specification and those that were not. To speed up the process of vertical bump shape digitization, a future method could utilize a digital device such as an iPad or an Android tablet with a pen. This method involves manually drawing the edge shape of each bump on the image, creating a fixed-color edge, such as black or red. The images can then be batch-processed in software such as Matlab or Labview, where precise edge detection can be achieved. Unlike gray-colored bump photos generated by microscopy, the process of manual drawing allows for 100% reliable edge detection. This method is significantly faster than the traditional point-by-point placement on the visible edge and has the potential to significantly speed up the digitization process.

D. *Process Integration*

To begin the process, we cut coupons using a reference wafer and create a carrier wafer with embedded thermocouples using the thermal matching procedure described in [6]. This carrier wafer matches the thermal responses of the reference wafers. We then map the coupons onto the carrier wafer and process them while analyzing using the metrology methods discussed in the previous section. This approach allows us to combine thermal mapping with shape mapping, with the latter becoming essentially a measure of temperature uniformity, as under-reflow correlates with lower temperature and over-reflow correlates with higher temperature. Reducing the temperature envelope results in reduced shape variation. Once all coupons/locations meet the required specifications, a transition needs to be made from the carrier wafer to full-size reference wafer processing. However, this transition may introduce a fixed thermal offset. Since the reference wafer cannot have any thermocouples, we characterize this wafer using top-view co-planarity analysis to validate the full-wafer performance and final thermal setpoints.

III. RESULTS AND DISCUSSIONS

Several experiments were conducted with coupons placed at various locations on the wafer and bump characteristic dimensions were measured. Figure 7 displays a portion of the data, where each dot represents a single measurement with a matching pair of bump height and diameter. It is evident that there is a bias towards 8 and 12 microns, where we anticipate favorable results. However, we conducted experiments under a

wide range of conditions, resulting in a greater spread of H/D ratios.

Our analysis determined that the center of the optimized process window is at a temperature of X°C with a reflow time of A seconds. The shapes of the bumps were digitized and analyzed based on height (H), diameter (D), and the height-to-diameter ratio (H/D). We then mapped the results to demonstrate the consistency of the post-reflow bump shape. The mechanisms of solder melting and reflow were also studied based on the energy input level, which is defined by the product of time above liquid (TAL) and reflow temperature. As shown in Figure 8A, there is always a tradeoff between temperature and TAL. Figures 8 (B, C and D) provide clear patterns in the reflow results and describe the optimal process tuning range, which must be carefully maintained to avoid under- and over-reflow results. With a higher energy input, there is a higher chance that micro-bumps will collapse and over-reflow. Conversely, at lower energy input levels, solders may not receive enough heat to complete the reflow process, which is known as under-reflow.

Meanwhile, we found that formic acid vapor can remove a thin layer of metal oxide from the micro-bump surface. By combining low pressure processing with the high diffusion coefficient of formic acid gas and low acid number density, we were able to achieve excellent formic acid spreading for dense pillar arrays, as shown in Figure 8B. One critical parameter for full wafer inspection is co-planarity. The vertical cross-sections of micro-bumps and their digitization analysis provided precise reflow shape measurements with accuracy down to 0.25µm. However, it is important to also consider the initial distribution data prior to reflow, as it can affect these measurements. Bump diameters can be measured from the top using imaging techniques, and variations in diameter can be used to assess co-planarity. Figures 9 illustrates the full wafer co-planarity performance. If the bump diameters, as measured on microscope images, exhibit consistent uniformity, then reflow is considered successful. However, diameter measurements alone cannot determine if the result is due to over-reflow or under-reflow. To accurately assess the outcome, multiple measurements must be performed that cover a range of conditions, including under-temperature/under-reflow, normal conditions, and over-temperature/over-reflow. By doing so, the proper diameter can be determined and utilized with top-view images. Co-planarity measurements are done using top view, where bumps are visible as round white circles. The microscope focuses on a particular vertical plane, and depending on the plane's height, the diameter will either be very small, like a dot (bump top tip), or the maximum diameter of a pillar when the plane no longer crosses the bumps. By interactively zooming in on the plane to the maximum diameter of several individual bumps, diameter variations can be measured. Additionally, as we measure the carrier wafer temperature, we can identify the conditions and diameters at which temperatures are lower or higher.

IV. SUMMARY

Using the ratio of bump height to diameter, we have developed an experimental methodology to characterize reflow performance for next-generation micro-bumps with a pitch of 15µm and below, correlating reflow performance to reflow process parameters.

We have developed a processing chamber, process flow, and integration that can achieve near-10 micron bump pitch reflow results. In this study, we have explored the physics and chemistry involved in making this possible, as well as the experimental methodology and preparations required, particularly with respect to thermal tuning. Additional details regarding thermal tuning can be found in our reference [6]. We have also demonstrated a methodology for measuring reflow using coupons and carrier wafers with matched properties, as well as microscope measurements and digital image processing for vertical (side) cross-sections. Further, we have shown the tradeoff between process setpoints and the time-temperature domain and confirmed good co-planarity results using top view diameter measurements.

ACKNOWLEDGMENT

The authors acknowledge the contributions of colleagues from Engineering and Application Labs at YES and especially of Astor Huang, Robin Abuluyan, and Sean Higgins.

REFERENCES

[1] "Scaling Bump Pitches in Advanced Packaging", https://semiengineering.com/scaling-bump-pitches-in-advanced-packaging/.

[2] https://www.siegertwafer.com/calculator_ok.html

[3] B.Sopori, W. Chen, J. Madjdpour, and N. M. Ravindra "Calculation of Emissivity of Si Wafers" Journal of ELECTRONIC MATERIALS, Vol. 28, No. 12, 1385 1999.

[4] N.M.Ravindra, B.Sopori, O.H.Gokce, X.Cheng, A.Shenoy, L.Jin, S.Abedrabbo, W.Chen, and Y.Zhang " Emissivity Measurements and Modeling of Silicon-Related Materials: An Overview" International Journal of Thermophysics, Vol. 22, No. 5, September 2001, 1593.

[5] "Copper Pillar & Micro Bump Inspection Requirements and Challenges", RudolphTechnologies,https://www.circuitnet.com/news/uploads/1/Copper_Pillar_App_rev2.pdf.

[6] V.V. Kudriavtsev, Lei Jing, T. Laaksonen, Z. Karim, and C. Lane "Pillars of Wafer Temperature Uniformity and Tuning for sub-10µ Reflow Applications, IMAPS DTC Conference, March 15th, 2023

[7] Su-juan Zhong, L. Zhang, Mu-lan Li, Wei-min Long, Feng-jiang Wang "Development of lead free interconnection materials in electronic industry during the past decades: Structure and properties", Materials and Design 215 (2022) 110439, pp. 1-59

[8] Siliang He, Yu-An Shen, B. Xiong, F. Huo, J. Li, and Hiroshi Nishikawa "Behavior of Sn-3Ag-0.5Cu solder/Cu fluxless soldering via Sn steaming under formic acid atmosphere", J. of Materials Research and Technology, 2022:21, p. 2352-231

[9] Siliang He, "Fluxless soldering under a formic acid atmosphere using Sn-3.0Ag-0.5Cu solder", Ph.D. Dissertation, University of Osaka, School of Engineering, 2022, pp. 1-139

Figure 1 A. Processing tool front end

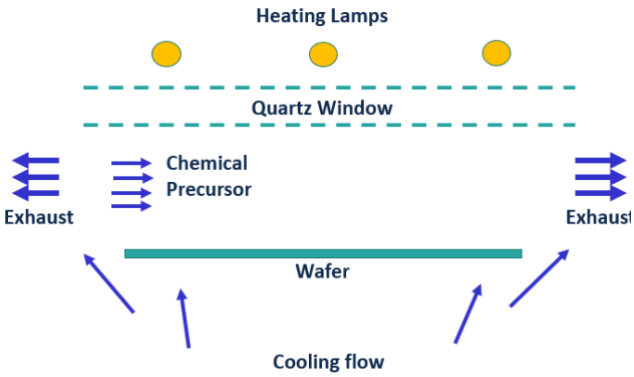

Figure 1 B. Functional chamber schematic

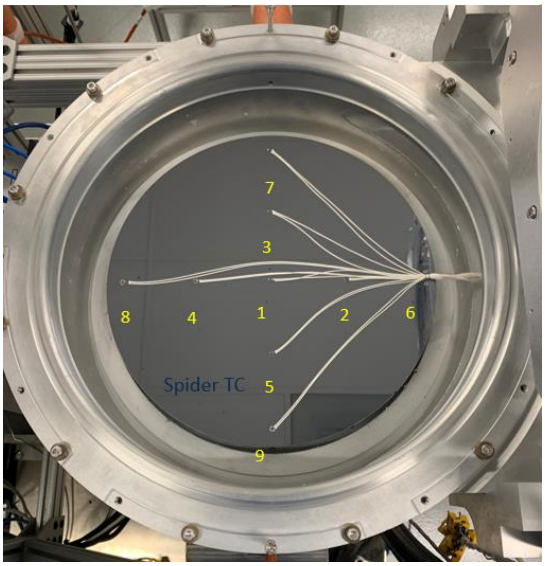

Figure 1 C Wafer with thermocouples

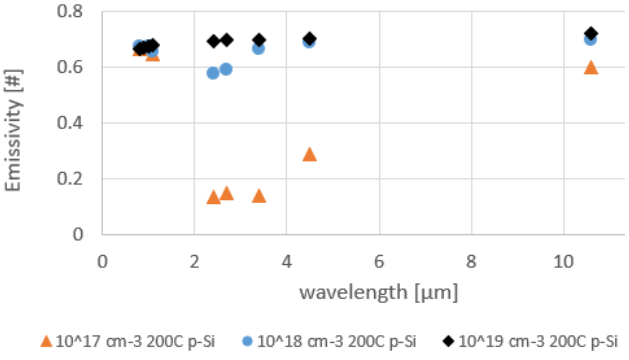

Figure 2A. Silicon emissivity: wavelength and dopant dependence

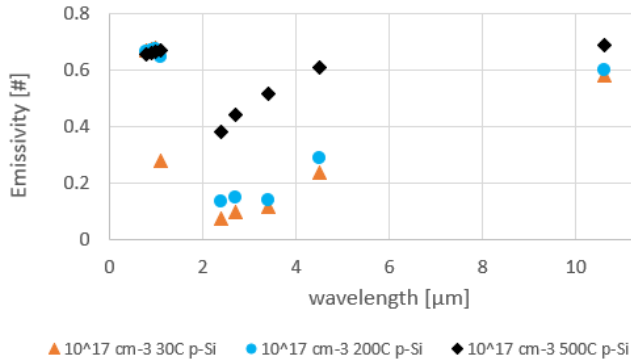

Figure 2B. Silicon emissivity: wavelength and temperature dependence

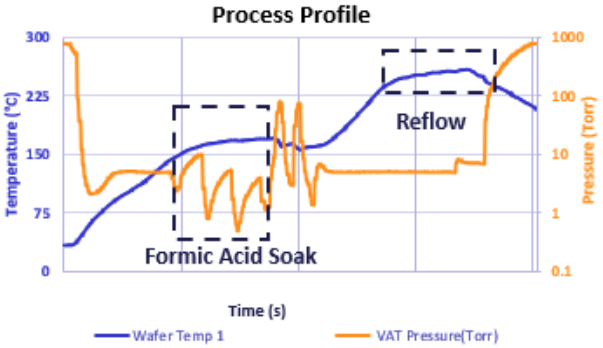

Figure 3. Process Flow

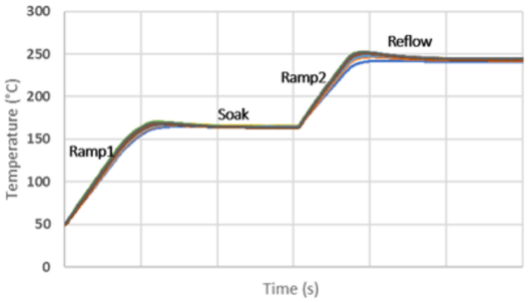

Figure 4A. Temperature history loaded from nine locations on the wafer

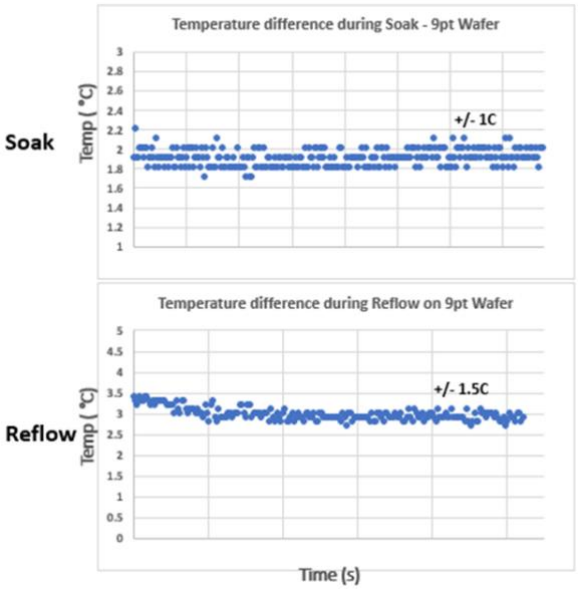

Figure 4 B. Temperature Uniformity vs Time at FA soak and reflow

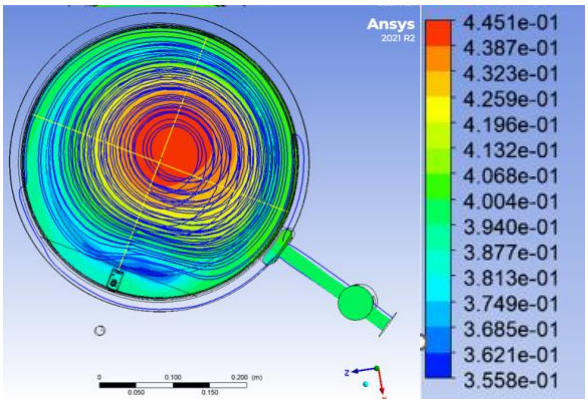

Figure 5 A. Chamber Chemical Mixture Distribution showing FA mass fraction and streamlines showing velocity distribution with wafer rotation

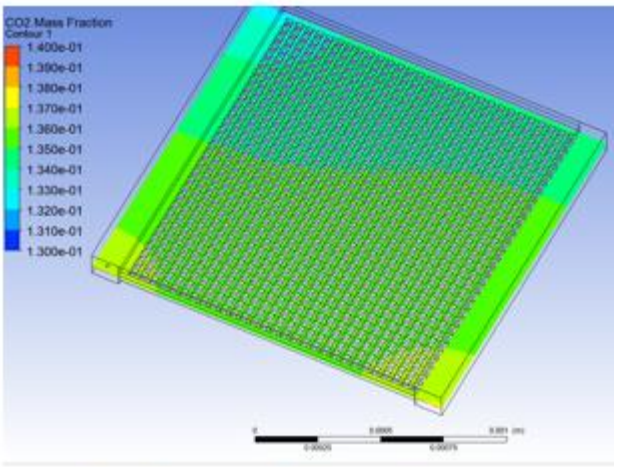

Figure 5 B. Diffusion at Low pressure

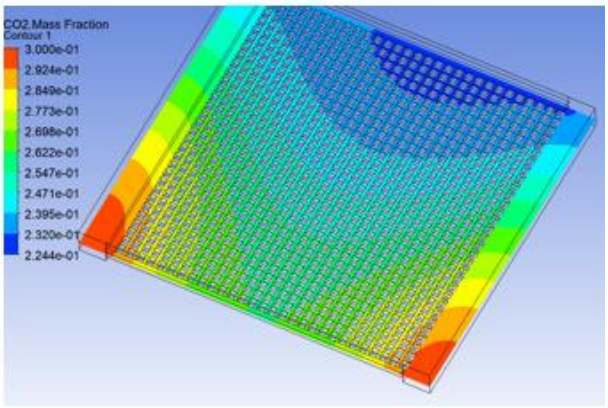

Figure 5 C. Diffusion at atmospheric pressure

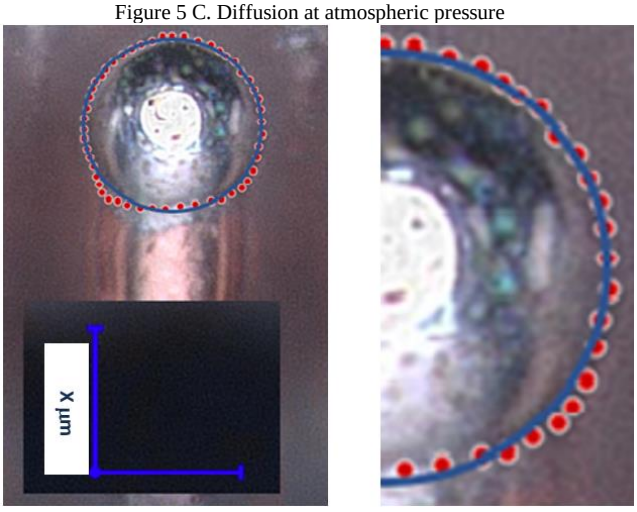

Figure 6A. Digitization Process using WebplotDigitizer software. Dimensions are marked on the image with its zoom in and edge points (red dots) are manually placed at the edge of the bump. Data exported into Plotly for post-processing.

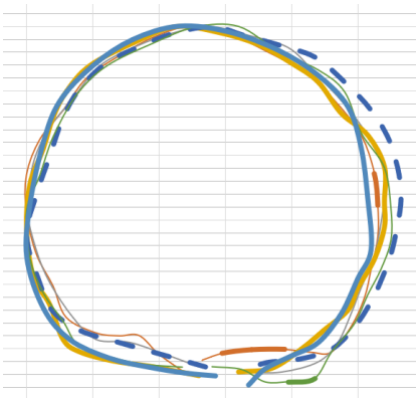

Figure 6B. Example of digitized bump shape analysis (bump diameter 11 micron). Six neighboring bumps processed and superimposed in same coordinate system.
Average Rt =H/D= 1.005769014
H/D uniformity is 5.9%

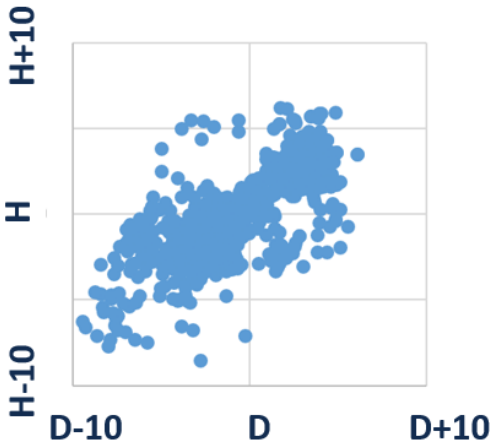

Figure 7. Experimental data showing (H,D) pairs. X-axis is diameter and Y-axis is height in micron.

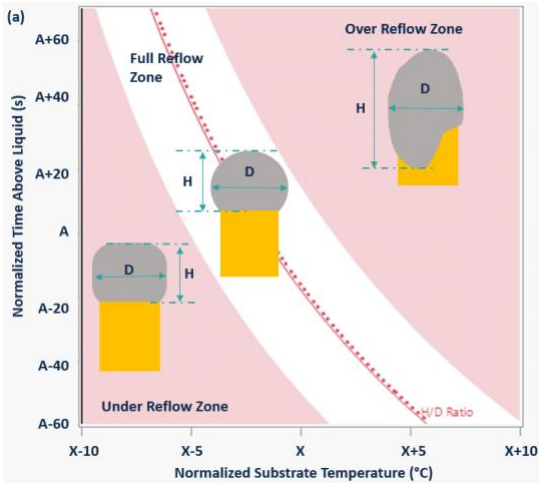

Figure 8A. Processing Map Summary

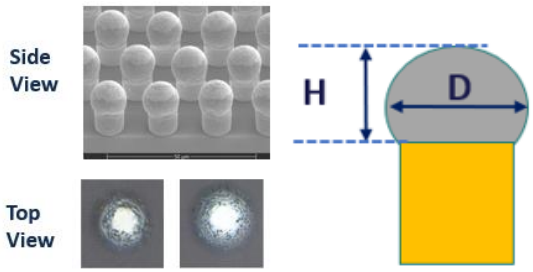

Figure 8B. Full reflow

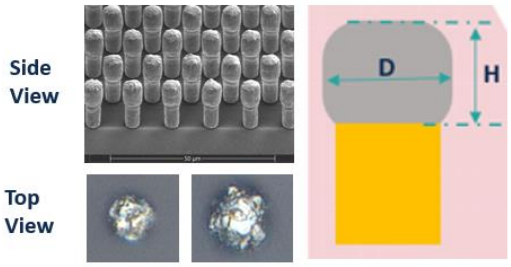

Figure 8C. Under Reflow

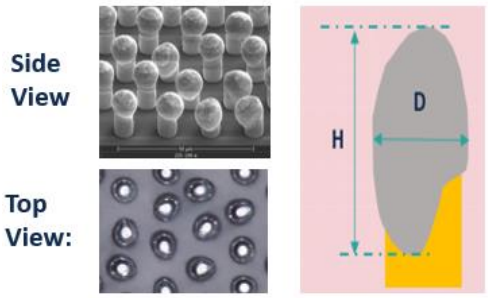

Figure 8D. Over reflow

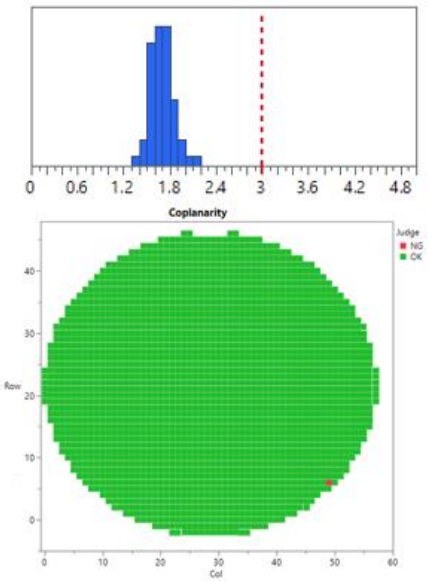

Figure 9. Full Wafer Co-planarity Analysis with Microscope top inspection, full reflow

Impact of Process Parameters on Vacuum Fluxless Solder Reflow Performance in Backend Applications with Bump Pitch of 15µm and Below

Lei Jing
YES (Yield Engineering Systems)
Fremont, California, USA
ljing@yieldengineering.com

Vladimir Kudriavtsev
YES (Yield Engineering Systems)
Fremont, California, USA
VKudriavtsev@yieldengineering.com

Taylor Nguyen
YES (Yield Engineering Systems)
Fremont, California, USA
tnguyen@yieldengineering.com

Jed Hsu
YES (Yield Engineering Systems)
Fremont, California, USA
jhsu@yieldengineering.com

Tapani Laaksonen
YES (Yield Engineering Systems)
Fremont, California, USA
tlaaksonen@yieldengineering.com

Xinxuan Tan
YES (Yield Engineering Systems)
Fremont, California, USA
xtan@yieldengineering.com

Alvin Lin
YES (Yield Engineering Systems)
Fremont, California, USA
alin@yieldengineering.com

Kay Song
YES (Yield Engineering Systems)
Fremont, California, USA
ksong@yieldengineering.com

Alex Chow
YES (Yield Engineering Systems)
Fremont, California, USA
achow@yieldengineering.com

Chris Lane
YES (Yield Engineering Systems)
Fremont, California, USA
clane@yieldengineering.com

Zia Karim
YES (Yield Engineering Systems)
Fremont, California, USA
zkarim@yieldengineering.com