Overcoming Cleaning Challenges for AI Devices: Optimizing a Cleaning Process for High Reliability

Haley Reid, Miller Armistead
KYZEN Corporation
Nashville, Tennessee, United States of America

Abstract

Every technological revolution has profoundly impacted people's lifestyles and the knowledge acquired in each era. The advent of search engines in the 1990s emphasized the importance of utilizing available information on the Internet and transforming it into valuable outputs rather than attempting to memorize extensive and intricate information. The advancements in artificial intelligence (AI) applications are expected to bring substantial changes and opportunities across various industries, including manufacturing, healthcare, finance, and transportation. With the ability to analyze and process large amounts of data, AI applications enable people to work more efficiently and increase productivity.

Companies like Nvidia, AMD, and Google utilize TSMC's CoWoS packaging technology to incorporate multiple processors and High Bandwidth Memory (HBMs) within a single package. From a packaging perspective, this can be considered a form of large flip-chip packaging, with package sizes up to at least 60x60mm. Additionally, the use of µ-bumps or Cu pillar bumps creates small standoff gap heights typically ranging from 25-40um. This poses a significant challenge in achieving thorough cleanliness of AI devices, as existing wafer form cleaning tools require extensive time and effort. If any residue from the ionic flux used in the process is left behind, the reliability of the AI devices may be compromised.

In this presentation, KYZEN will explore cleaning strategies tailored to overcome the challenges of CoWoS packaging technologies. The discussion will cover the cleaning obstacles unique to these packages, along with steps to optimize the cleaning process and achieve the desired level of cleanliness for AI devices. This will include cleaning tool selection, cleaning chemical selection, wash and rinse parameters, and inspection methods. The goal is to outline a successful cleaning process that guarantees the intricate devices' high reliability and extended device life.

Introduction

Artificial intelligence (AI) innovations enable computer systems to perform tasks previously only achievable through human intervention. AI-powered devices can now engage in reasoning, decision-making, and problem-solving using various technologies such as machine learning, deep learning, and natural language processing [1]. The advantages of AI-powered processing differ across industries and applications. Still, a few notable examples are 24-hour availability, increased efficiency, enhanced safety and risk reduction, and consistent performance throughout the device's lifespan. This technology has been implemented in healthcare, cyber security, military systems, agriculture, and other industries requiring high reliability and repeatability.

The high level of processing required for AI applications has led to increasingly complex electronic packages. Some commonly used packaging types are fan-out wafer level packaging (FOWLP), fan-out panel level packaging (FOPLP), and chip-on wafer on substrate (CoWoS) packaging. All AI packaging types are characterized by intricate geometries, low gap heights, large die sizes, and sensitive metallurgy that support advanced computing but are challenging to clean off residues such as flux and other manufacturing contaminants. A single GPU training cluster for AI applications costs over $4 billion for the server investment alone, ensuring the most reliable electronics possible [2]. It has been seen that in high-reliability applications, uncleaned or partially cleaned flux residues can cause premature device failure [3].

One of the most widely used packaging types is CoWoS, which represents 2.5D IC packaging technology. Compared to other technologies, this packaging is very common in AI applications due to space, power, and cost reduction. This is achieved by the integration of Systems on Chip (SoC) and high-bandwidth memory (HBM) bonded to a silicon interposer with through-silicon vias (TSV), which is then bonded to a PCB substrate. However, a few notable cleaning challenges for CoWoS packages limit the final devices' throughput and reliability.

The cleaning process for CoWoS is divided into two steps. The first step is to chip on the wafer (CoW) cleaning. This is traditionally done using a singular spray nozzle and centrifugal force to allow penetration under the 25-40µm gap heights. This method tends not to thoroughly clean the central die; a single wafer can take over 90 minutes to clean. After the CoW cleaning, the wafers are soldered to the substrate. The fully assembled package is then cleaned using an inline spray-in-air machine, but incomplete cleaning generally occurs due to the large die size, requiring multiple wash cycles. A partially cleaned package will have a short use life, poor thermal management, and low output during production.

This paper will focus on designing a cleaning process for CoWoS packages that results in clean and reliable components. The four cleaning process parameters that will be tested are time, energy, solvency, and temperature. These are the most critical parameters when investigating any cleaning application. The overall aim is to understand how to step through the cleaning process development to allow for high throughput of clean packages in the fastest time.

Experimental Methodology

Test Equipment

The cleaning machine used for this testing was a single nozzle spray box designed and manufactured at KYZEN Corporation to allow filming during cleaning trials, as shown in Figure 1. The image on the left shows a front view of the machine, while the picture on the right shows a right profile view. The front view of the machine shows the stainless-steel construction, a shelf for most sized cameras, and the front viewing window that can be used for additional camera angles. The pump at the bottom of the left image is a 1/3 horsepower Dayton Carbonator Pump Motor. The heater and pump are controlled using the grey control box attached to the left side of the cleaning chamber. The right view of the machine highlights the additional viewing windows on both sides of the chamber, the arm used to adjust the angle of the spray bar, and the 5-gallon holding tank.

Figure 1. Spray-in-air cleaning machine used during testing.

The inside of the cleaning machine and the slide carrier plate are shown in Figure 2. The image on the left is a top-down view of the machine with the lid removed. The center of the picture shows the spray nozzle positioned over the left side of the glass slide. Underneath the glass slide is the slide carrier to ensure the slide is positioned repeatably during testing. The image on the right is taken from the shelf, seen in Figure 1, into the chamber. This view was used during this test to record cleaning.

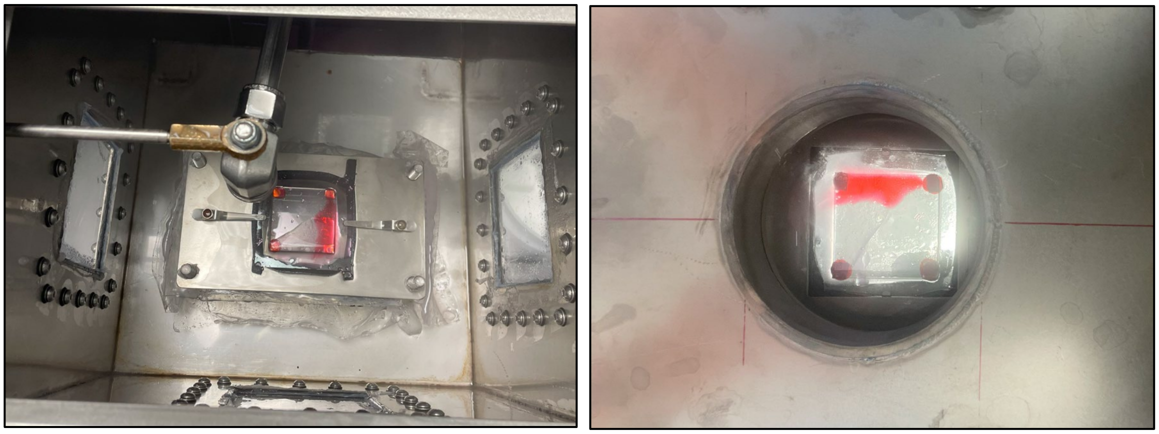

Figure 2. A view of the inside of the cleaning chamber.

The cleaning chemical used was a water-soluble flux cleaning agent, CA-W, diluted to 3% by volume. The holding tank was filled with 7.5 liters.

Materials

Glass slide test vehicles were used in this testing to simulate low standoff heights and large die sizes. KYZEN, Amp Think, LLC., BlueRing Stencils, and ITW EAE designed and manufactured the slides used during this testing. The test vehicle is shown in Figure 3. The image on the left shows the 70x70mm base glass slide underneath the 50x50mm glass die with red epoxy in the corners of the top die. The picture on the right illustrates the 25µm standoff height of the top die from the bottom. Interconnects were placed on the test vehicles but were quickly removed during reflow due to poor adhesion; future work will have interconnect morphology attached. Figure 4 shows a slide reflowed with Water Soluble Flux #1(WS-1). The slides were reflowed at a 30° angle using a typical lead-free reflow profile; the angle allowed the flux to wick under the slide, but no complete contamination was achieved, no matter the angle or amount of flux used. All testing will be completed using the water-soluble flux vehicles, WS-1 and WS-2, due to the availability of the flux-only material. The WS-2 was dyed red with a small amount of liquid food coloring for easy video processing.

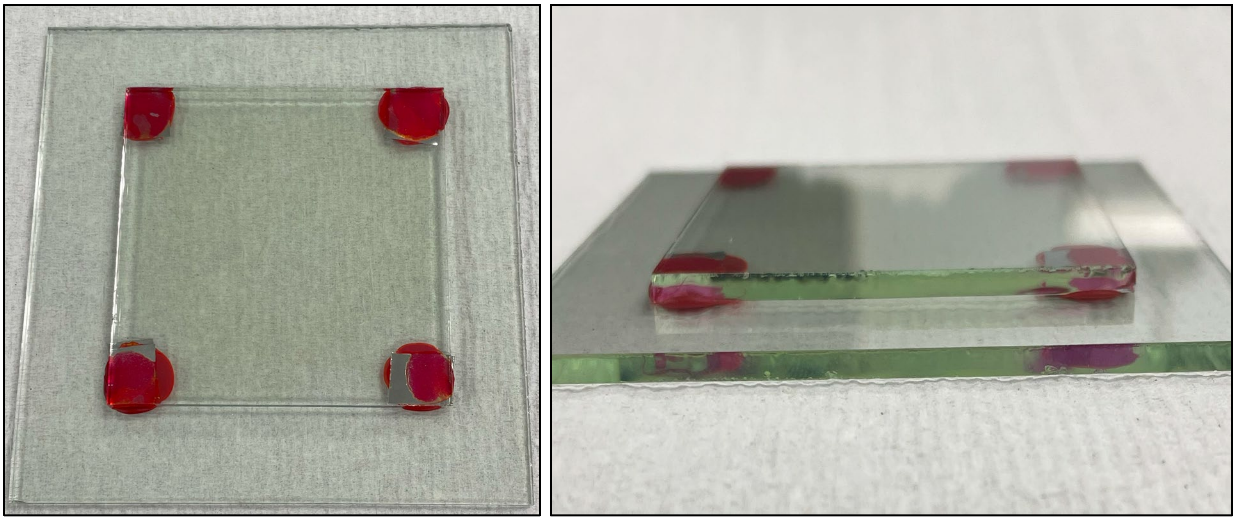

Figure 3. An image of the glass slide vehicle used.

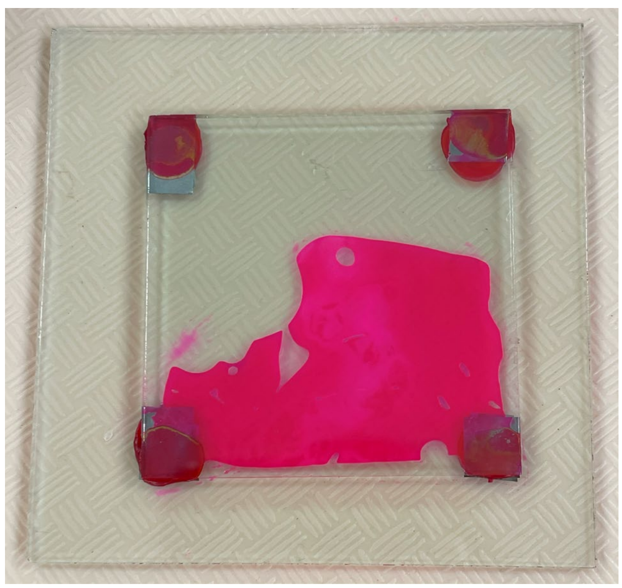

Figure 4. Glass slide test vehicle after reflowing with WS1.

Results

Each test was recorded for the entire duration using an Osmo Action 4 camera. The percent of relative flux removal was then calculated using a Python video processing script developed at KYZEN for this application. The four parameters mentioned in the introduction, time, energy, solvency, and temperature, will be varied to create different processing conditions. The tests performed to vary the four parameters are listed in Table 1. The grey-shaded cells indicate the parameter of interest for each run.

Table 1. Test Parameters

Run	Flux Vehicle	Concentration	Time (min)	Pressure (bar)	Temperature °C
1	WS-1	3%	1	5.5	50
2	WS-1	3%	5	5.5	50
3	WS-1	3%	30	5.5	50
4	WS-1	3%	5	7	50
5	WS-1	3%	5	8.3	50
6	WS-1	3%	5	7	60
7	WS-1	3%	5	7	70
8	WS-1	3%	5	7	80
9	WS-2	3%	1	5.5	50
10	WS-2	3%	5	5.5	50
11	WS-2	3%	30	5.5	50
12	WS-2	3%	5	7	50
13	WS-2	3%	5	8.3	50
14	WS-2	3%	5	7	60
15	WS-2	3%	5	7	70
16	WS-2	3%	5	7	80

After the tests, the videos were processed using the Python script to determine each run's percent change in flux. The graphs in Figures 5 through 8 show the relative change of the flux throughout the videos; each graph represents 3 to 4 runs based on the grey groupings in Table 1.

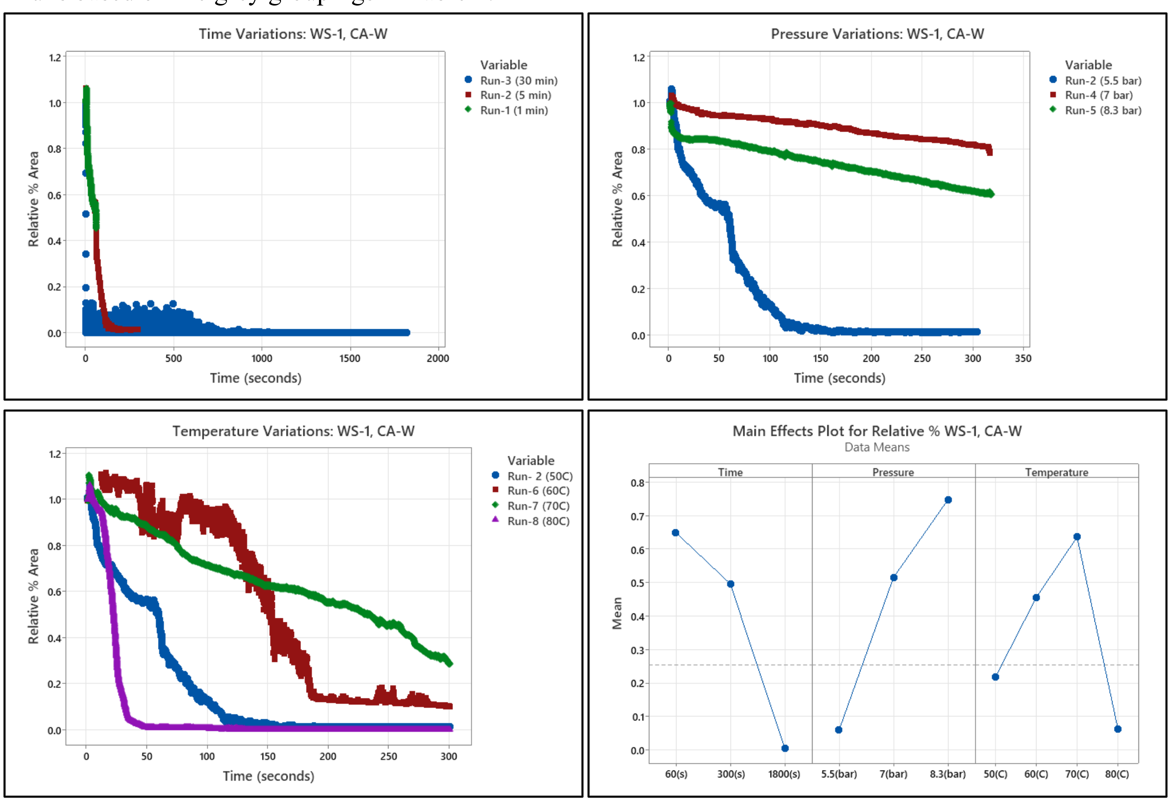

Figure 5. Runs 1 through 8 results.

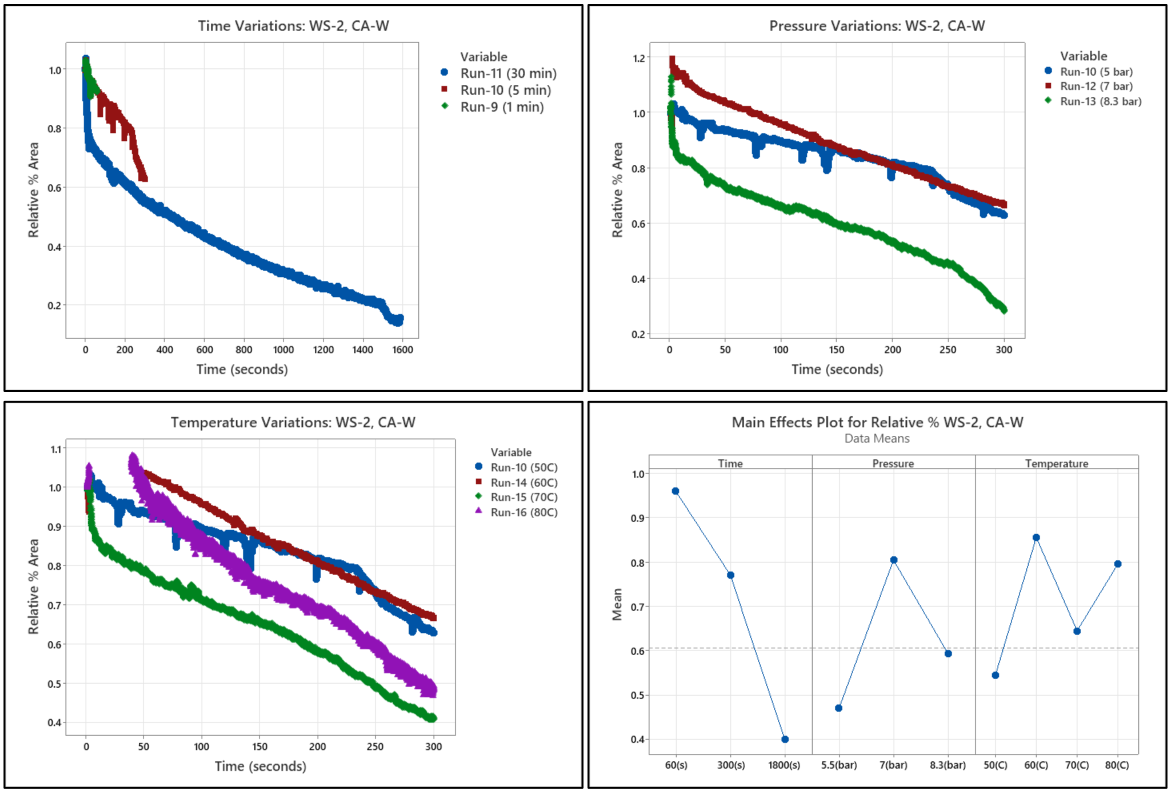

Figure 6. Runs 9 through 16 results.

The flux was selected by color for the area change, as seen in the bottom row in Figure 9. When choosing the flux in the Python script, the red adhesive and red dyed flux had some interactions.

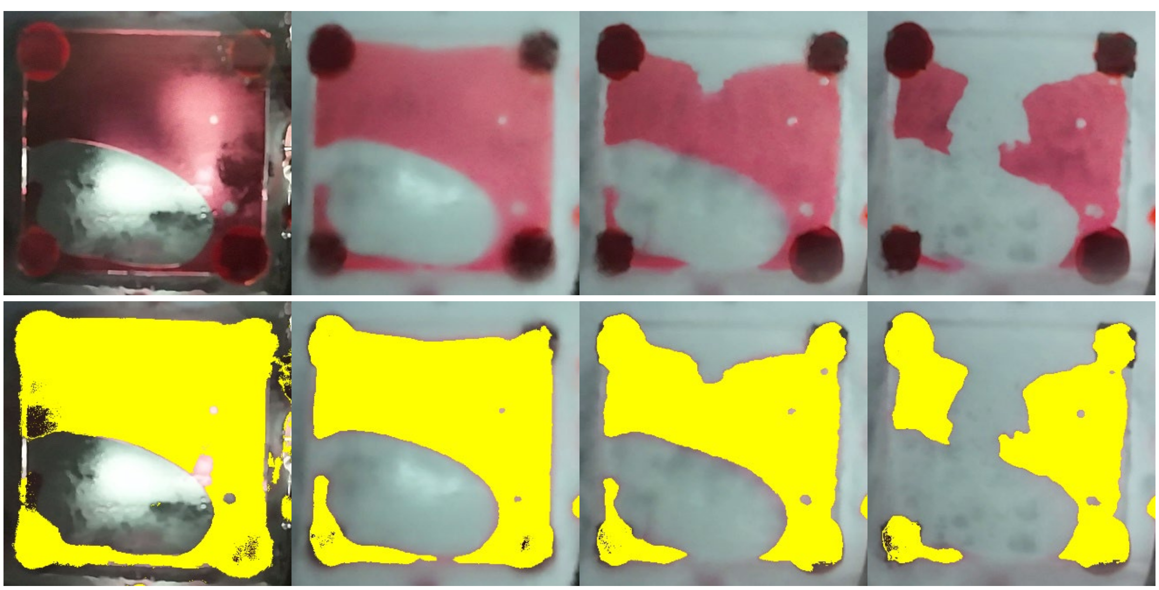

Figure 7. Images from run 16. The top row represents the images taken from the video. The bottom row represents the flux selection for the area change over time.

Discussion

The testing conducted focuses on CoW cleaning with a large dummy die. The results will be analyzed to provide recommendations for effectively cleaning challenging components and assemblies. The critical parameters tested include time, pressure, temperature, and solvency selection, which are essential to successful cleaning process implementation.

Time Effects

AI cleaning process are a tradeoff between the reliability of the final product and the output of the product manufacturing operations. The length of the cleaning process is a dependable method for complete cleaning, but it can significantly inhibit the production deadlines crucial to a fast-paced industry. To ensure the most profitable and reliable assemblies, the cleaning time needs to be optimized to the lowest possible time that still produces the highly reliable components required for AI applications.

This study examined the effects of different cleaning cycle times, 1, 5, and 30 minutes while maintaining all other factors consistently. Results indicated that extending the cleaning time beyond 300 seconds improved cleanliness levels if any flux remained at that time. This is demonstrated in Figures 5 and 6. However, it is crucial to consider that time is a limited resource in high-volume manufacturing settings such as for AI applications.

The selection of cleaning time should be based on carefully evaluating various cleaning process metrics rather than choosing the most extended time available. Additionally, the wash cycle duration will directly influence the rinsing time in a spray-in-air inline cleaning machine. Incomplete rinsing can lead to electrical failures, underscoring the importance of time in the wash and rinse.

Pressure Effects

In a spray-in-air AI assembly cleaning application, spray pressure plays a significant role, and many additional factors must be considered. The spray pressure will allow the fluid to be pushed underneath a low-gap component, while the spray nozzle type, position, and angle are also important considerations. Due to time limitations, this evaluation focused solely on the spray pressure aspect. Further testing would be required to assess the full impact of the spray nozzle types and positions.

During this testing, 5.5, 7, and 8.3 bar pressures were applied using an industry-standard fan nozzle. The spray nozzle was positioned adjacent to the top die to enable deflection underneath the component, as shown in Figure 2. A spray pressure that is too high can cause the fluid not to make sufficient contact to remove the flux, while a low spray pressure can mimic a flooding action under the die. The WS-1 and CA-W showed that the lowest pressure, 5.5 bar, was the most influential factor after time for cleaning of the flux. However, WS-1 also showed that 7 bar outperformed 5 bar marginally. The WS-2 flux showed that 8.3 bar was the most effective cleaning pressure, while 5 and 7 bar were relatively equal. There was not a definitive result for the pressure testing. However, it was seen that the highest pressure did not correlate to the best cleaning between the two flux types.

The two fluxes showed different responses to spray pressure variations. Further investigations will need to be performed to determine the root cause. In general electronics cleaning, a spray pressure that is too high can cause the fluid not to make sufficient contact to remove the flux, while a low spray pressure can mimic a flooding action under the die.

Temperature Effects

The temperatures tested for this paper ranged from 50°C to 80°C, encompassing the normal range of cleaning temperatures for semiconductor cleaning. The lowest effective wash temperature is generally the best choice in any cleaning application, especially AI applications with high production volumes. This is due to energy costs to keep the machine heated and material compatibility concerns of sensitive metals and polymers.

The best temperature for AI cleaning applications will depend on the flux and cleaning agents chosen for manufacturing assemblies. Elevated cleaning temperatures aid in the solubility of fluxes during any assembly cleaning and lower the surface tension of the wash bath to allow cleaning under low-gap components. However, there is a tradeoff between aiding solubility and material compatibility. During the testing performed for this paper, it was observed that the epoxy holding the two glass slides together was no longer adequate for additional testing. This highlights the compatibility concerns that could be observed during production.

Across all tests for temperature variations, it was observed that an elevated temperature correlated with faster flux removal. The water-soluble cleaning agent CA-W showed that both flux types cleaned faster with elevations in temperature. However, WS-2 was not completely removed in 300 seconds, even at 80°C.

Solvency Effects

In all high-reliability cleaning processes, a match between the contamination and the cleaning agent chosen will determine the ease with which the process can be optimized. A mismatch between the flux and cleaning agent will make optimizing all other process parameters much more tedious and, overall, less effective. This was shown during the testing with WS-1 and WS-2.

The two water-soluble fluxes chosen for this testing exhibited different solubility with the CA-W material. WS-1 was thoroughly cleaned under several of the testing conditions using CA-W. The flux vehicle WS-2 was significantly more challenging to clean with CA-W and was not fully cleaned under any tested parameters. When choosing a chemistry to clean WS-1, CA-W would be a good starting point for an overall process. However, it is unlikely that CA-W would be selected in a process cleaning WS-2.

Suggested Process Optimization Methods

Tests 1 through 16 highlight many methods to optimize a cleaning process and several ways to inspect test or production assemblies for cleanliness. Based on the findings above, a suggested starting point for a successful cleaning process is to determine the inspection methods and test vehicles that are representative of the manufacturing process under investigation.

Once the inspection methods are chosen, a baseline solubility testing between the fluxes in use and the cleaning agents of interest, whether engineered or water only, needs to be performed. The cleaning agent that showed the quickest dissolution of the flux can then be immersion tested with all sensitive parts, such as any aluminum pads present, to determine if the product could be a good match and test any temperature sensitivities. Then, the chemical with the best-starting solubility and metal compatibility should undergo testing by changing one parameter at a time and inspecting after each change. There will often be process limitations that will hold one parameter at a particular level. This is often the case with time variations due to throughput demand in AI manufacturing.

After static testing is complete, the testing performed for this paper indicated that temperature is a factor that has a minimum value needed for complete cleaning under a large die. The data from immersion solubility testing must be considered to test temperature variations successfully. For example, if an aluminum pad starts to discolor after 5 minutes at 80°C, a lower temperature would be suggested if a second wash could be needed during manufacturing. The test vehicles will need to be placed in the cleaning machine used for production while keeping all factors but temperature identical. After stepping through all temperatures for the tests, the test vehicles will need to be inspected. The lowest temperature to produce the cleanest test vehicles should be used to move on to pressure testing. Pressure variations should be tested as systematically as the temperature variation testing.

Conclusion

A systematic approach to cleaning parameter adjustment is the most important for reliable AI application packages. This will include stepping through individual cleaning parameters and having well-designed test vehicles that represent the final product of the process. In this work, the four cleaning process variations of time, energy, temperature, and solvency relate to more machine settings than were investigated. However, once the cleaning agent, time, pressure, and temperature are chosen, making additional changes, such as spray nozzle type or rinse temperature, will be easier to narrow down.

AI applications require packages to be entirely contamination-free to avoid the high cost of failure. Each manufacturing process will have unique challenges, but having a non-biased method to step through the cleaning process will ensure that the process is optimized and will allow for the highest throughput available with minimal losses.

Acknowledgments

The authors would like to thank Alan McCready, Jake McMasters, and Aaron McWright of KYZEN Corporation for designing and manufacturing the spray testing machine; Wayne Raney (KYZEN) and Mike Butler at ITW for designing and manufacturing the test vehicles; Adam Klett, Ph.D. (KYZEN) for designing the Python video processing script; and Jason Chan, Scott Cain, Chelsea Jewell, and Ram Wissel of KYZEN for consulting during the project.

References

[1] Soori, M. (2023) Artificial intelligence, machine learning and deep learning in advanced robotics, a review, *Cognitive Robotics,* Volume 3 Pages 54-70.

[2] Patel, D. and Nishball, D. (2024) 100,000 H100 Clusters: Power, Network Topology, Ethernet vs InfiniBand, Reliability, Failures, Checkpointing, *Semianalysis,* https://www.semianalysis.com/p/100000-h100-clusters-power-network

[3] IPC-CH-65B (2011) Guidelines for Cleaning Printed Boards and Assemblies, *IPC Standards,* Page 13.