White Paper

Original Equipment Manufacturers (OEMs),
Original Design Manufacturers (ODM),
Independent Software Vendors (ISVs)

AI-based Inline Inspection Improves Semiconductor Wafer Quality

Semiconductor wafer thinning can be a source of defects. An inline vision inspection system, running on Intel® architecture CPU- and GPU-based servers, uses AI and provides 100% die inspections in real time

Using thinner silicon substrate is a growing trend in semiconductor packaging where wafer stacking or high density packaging of semiconductors benefits from a thin (tens of micrometers) silicon wafer. In its most recent report on wafer thinning, Yole Group predicts the thinned wafer market will grow significantly to 135 million thinned wafers by 2025[1].

However, the process can lead to expensive defects on the wafer surface. Even minor defects such as die cracks on the thinned wafers can result in a huge impact on downstream packaging, and lead to thermal issues, or other reliability issues.

Historically, wafer die samples were inspected at each step in the assembly process. If a quality or reliability issue on one or more chip package samples was found, then production was halted while teams of experts search for the root cause. Once a defect was found, the entire batch of semiconductor devices that shared silicon dies from the same wafer must be reinspected. After re-inspection, any affected die was discarded wasting time and resources.

Now, artificial intelligence (AI) inference can classify the types of defects early in the process improving the quality of the final chips.

Wafer Defects Have Big Impact

Defects in wafers during the die preparation stage of the silicon assembly process can lead to a variety of quality and reliability issues in the final packaged semiconductor devices. Here are some of the key issues that can arise:

Thermal Issues: Defects like cracks and scratches can affect the thermal conductivity and thermal dissipation properties of a semiconductor device. This can result in uneven heat distribution, hot spots, and reduced thermal performance. In some cases, it can lead to thermal runaway, where excessive heat damages the semiconductor or surrounding components.

Thermo-Compression Bonding (TCB) Issues: TCB is a common process used in semiconductor packaging. Defects in the wafer can lead to improper bonding during TCB, causing weak or unreliable connections between the die and the package. This can result in intermittent electrical connections that can lead to device failures over time.

Mechanical Integrity: Cracks and scratches can compromise the structural integrity of the die. During handling, assembly, or in-field use, these defects can propagate, leading to die delamination or even complete package failure. Mechanical stress from thermal cycling can exacerbate these issues.

[1] https://www.yolegroup.com/press-release/memory-cis-and-power-electronics-are-driving-the-wafer-thinning-equipment-market-before-a-new-wave-of-innovation-by-2025/

Table of Contents

Electrical Issues: Foreign materials on the wafer surface, such as particles or contaminants, can interfere with the electrical functionality of the semiconductor. They can create short circuits, alter the electrical characteristics of the device, or cause electrostatic discharge (ESD) events, all of which can result in device malfunction or failure.

Reliability Degradation: Defects in the die can weaken the overall reliability of the semiconductor device. They can lead to early-life failures, reliability issues in the field, and reduced mean time between failures (MTBF), which is critical in applications where long-term reliability is essential.

Yield Reduction: Defects during die prep can lead to a decrease in manufacturing yield, increasing production costs and potentially reducing the number of functioning semiconductor devices. Lower yields can also impact the competitiveness of semiconductor manufacturers.

DGB Inline Inspection Solution

Inline AI inference provides the ability to inspect 100 percent of the wafers just after the backgrinding process to identify any defects on wafer surface. Inline vision inspection systems use a light source and three 16,000-pixel resolution line scanning industrial cameras to capture wafer images. The overall image size is 15360 x 8640 pixels which provides the granularity to see defects that are only microns in size.

The real-time detection results are displayed on a monitor connected to an industrial PC (IPC), and the algorithm's inference time is approximately 1 second. The entire detection process can be completed within 3-to-5 seconds. In case defects are detected, the device triggers an alert, meanwhile this information is synchronized with a wafer grinding/polishing equipment to stop its operation and minimize further losses.

Powered by Intel® Core™ i7

This inline vision inspection system is powered by the Intel® Core™ i7 processor, which is part of the 13th Gen Intel® Core™ product family.

These CPUs provide I/O of up to 16 PCIe 5.0 lanes and DDR5 memory up to 5600 MT/s. The Intel Core i7 product family offers SKUs with up to 24 cores comprised of eight performance-cores and 16 efficient-cores and up to 32 threads. The performance cores can reach 5.8 GHz with Intel® Thermal Velocity Boost to elevate performance.

Intel Arc GPU and OpenVINO Accelerate AI

The Intel Arc A770 Graphics is a dedicated discrete graphics card (dGPU) that serves as the powerhouse behind AI inference workloads. Its robust performance is harnessed to process both large image sizes and vast quantities of images in a very short time. The GPU's high-performance Intel Xe HPG microarchitecture enables high-performance AI inference-based applications.

In the consumer PC market, Intel offers ODM-branded consumer graphics cards and Intel-branded workstation graphics cards. In the embedded PC market, Intel has partnered with ODMs who offer their branded graphics cards powered by the Intel® Arc™ GPU. These ODMs include Matrox, ADLINK, Advantech and Asus, with several to come . As designed for embedded applications, these graphics cards come with a variety of form factors and longer life than consumer graphics cards.

For AI performance, the inline vision inspection system makes use of the Intel Distribution of OpenVINO, an Intel-developed, open-source solution for optimizing and deploying AI inference, in domains such as media analytics, machine vision and visual inference, natural language processing, and more.

The Intel Arc A770 Graphics is optimized for use with the OpenVINO inferencing engine. In combination with OpenVINO, the Xe GPUs achieve high AI performance levels that reduces production costs, improves factory output and efficiency.

The Intel® Distribution of OpenVINO™ toolkit converts and optimizes AI models, allowing them to be deployed across a mix of Intel hardware and environments (see Fig. 1). For this system, the use of OpenVINO allowed the application to optimize the GPU for inference.

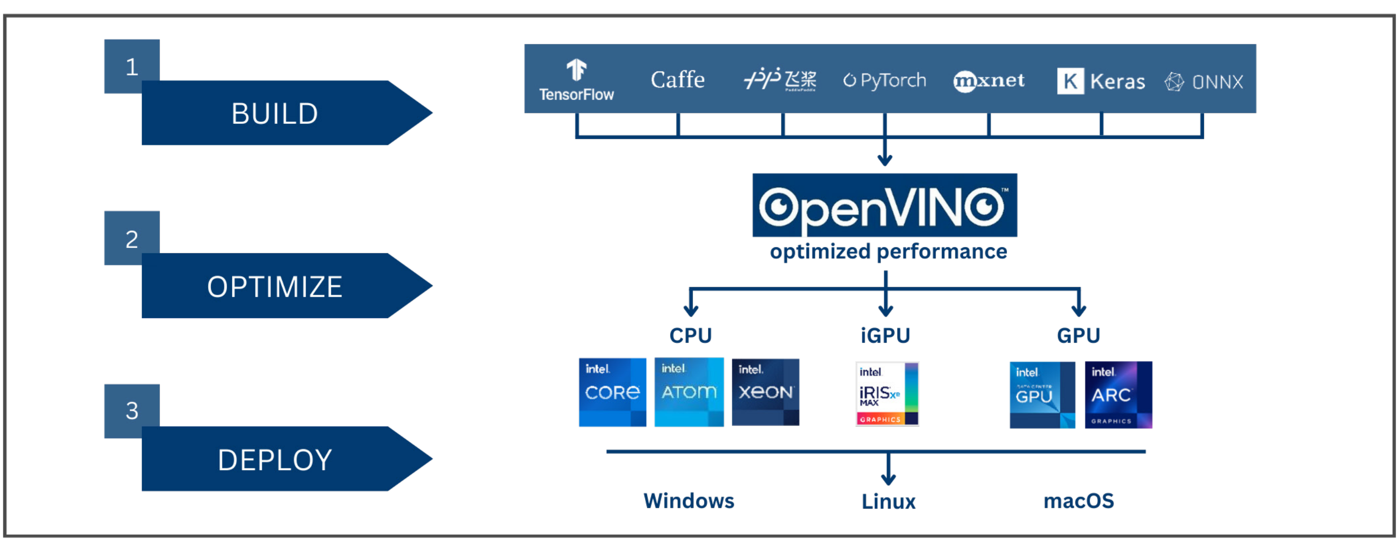

Figure 1. Three phases of Intel Distribution of OpenVINO AI interference software.

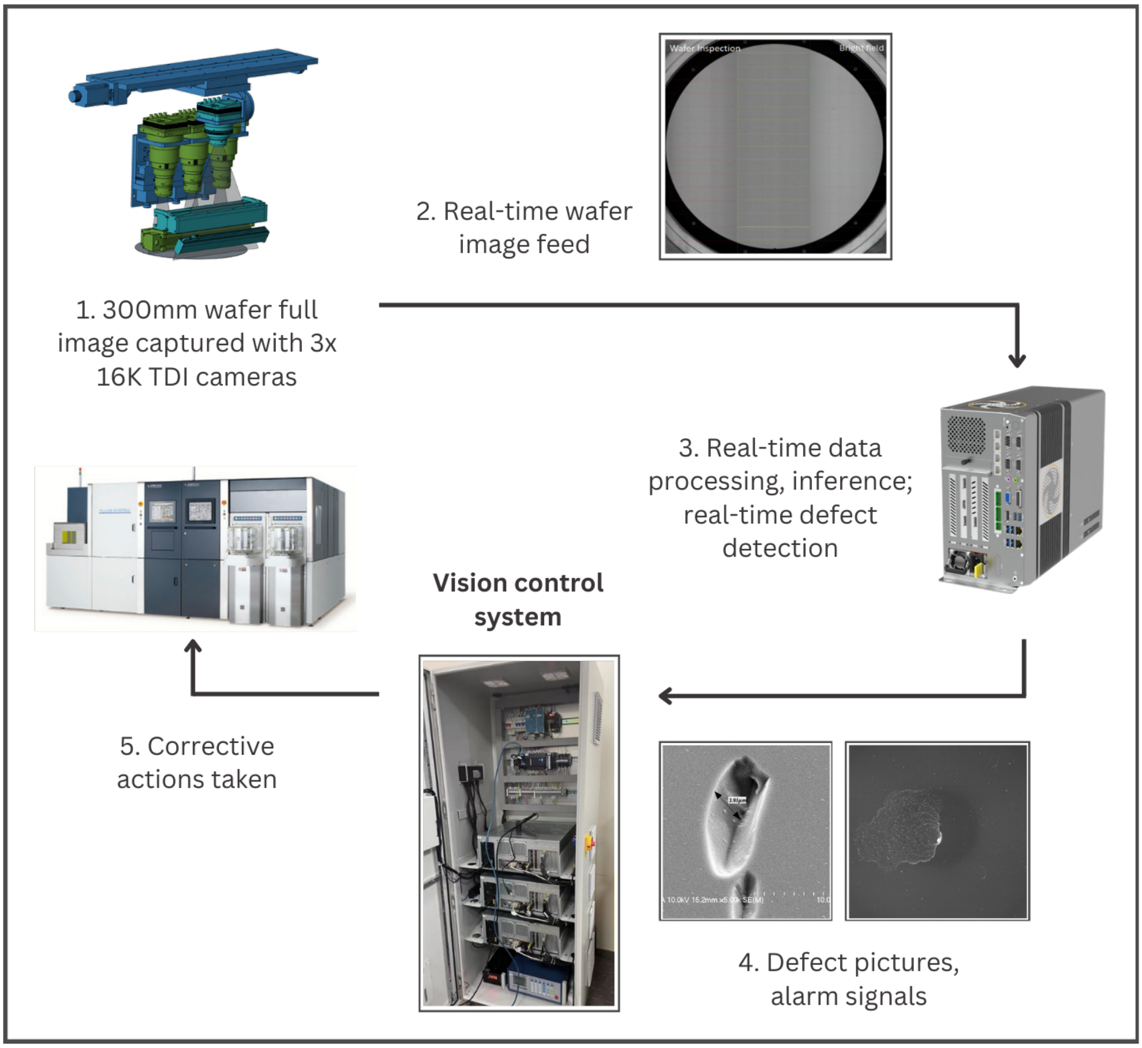

Figure 2. This image depicts the process in which the inline vision inspection system inspects and detects defects in a semiconductor wafer.

Building the Wafer Thinning Quality System

For this application, see Figure 2, three cameras and inspection lights are installed inside the wafer grinder. The cameras take images of wafers and create a real time wafer image feed that sends the images to the data labeling and training systems.

Real-time defect detection happens at the deployment stage with AI inference taking place on the vision controller. When defective wafers are spotted, an alarm signal is generated and sent, along with wafer images, to vision control for corrective action.

Figure 2 shows the inline vision inspection system used in the data processing and inference part of the system: This includes Industrial PC, a lighting controller and the Z-axis motor controller. The three IPCs connect with the cameras installed in the wafer grinder and an Intel Arc A770 GPU processes the AI inference.

Once the defect information has been captured, the Intel Arc A770 GPU's AI inference can be used to classify the defect type distinguishing between cracks, grinding marks, and others (see Figure 3).

Figure 3. Types of wafer defects detected by the inline vision inspection system.

This defect information (area, width, height, edge to edge) is then sent to the system for analysis and is used to determine if these anomalies are within alarm limits (i.e., they meet the quantization rule). There are alarm limits and stop limits for the process that either stop the entire process, warn the operator or ignore the defect (see Figure 4).

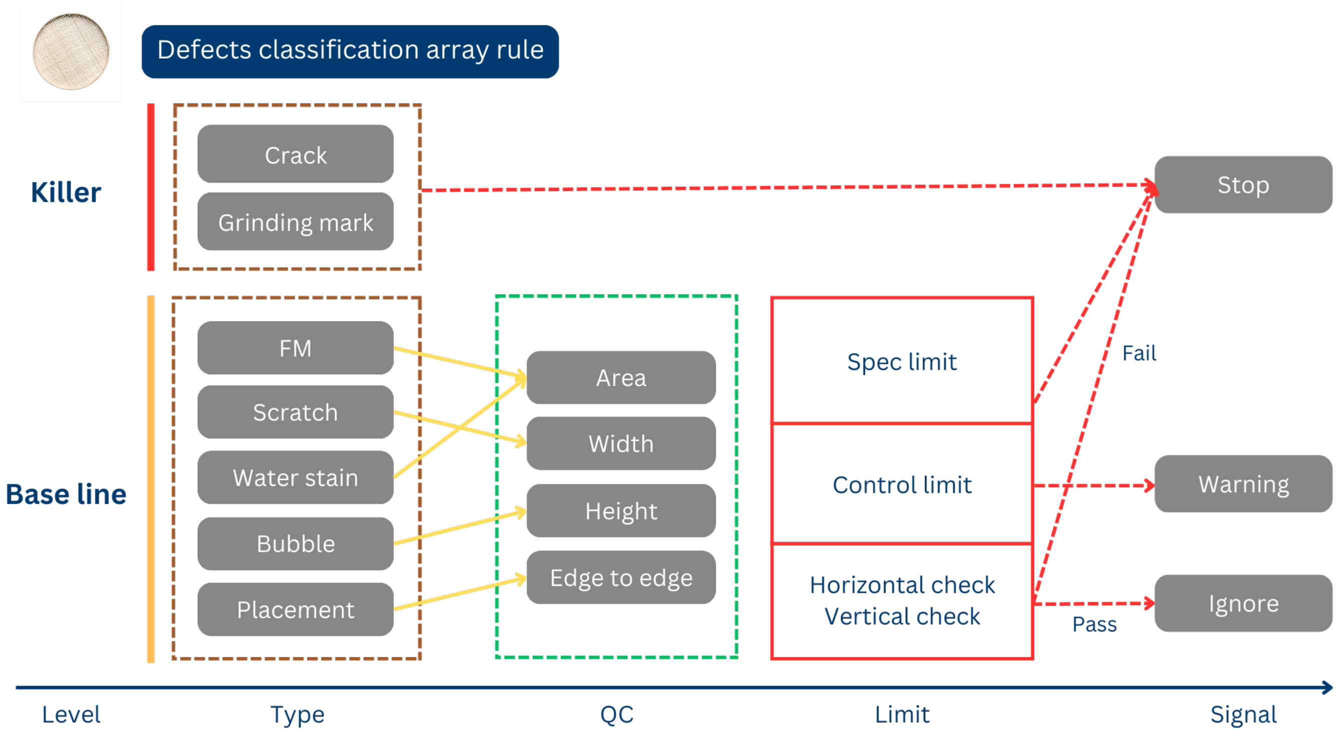

Figure 4. Defect analysis process.

Conclusion

Wafer thinning is one of the many complex processes involved in creating a semiconductor device. Maintaining wafer quality is critical. Today, these devices are sampled for quality at various stages of fabrication, which means a flaw created at the start of the fabrication process might not be identified until much later in the process. Examining wafers at the beginning of the process can save significant amount of money by identifying defects before too much time and money has been spent on fabrication.

AI is a technology that can be used for sampling 100% of the wafers in real time to improve quality significantly. Intel has worked to provide the compute power built into the data processing and inference part of the inline vision inspection system, running on the edge based on Intel® architecture CPUs and GPUs. This system shows how AI inference can be inserted into the fabrication process in real time and can use the power of GPUs to inspect for flaws and defects.

Learn More

Intel Core i7 processor

Intel Arc GPU

Intel Distribution of OpenVINO

Notices & Disclaimers