White Paper
December 2021

IT@Intel: Transforming Manufacturing Yield Analysis with AI

Intel IT is using artificial intelligence (AI) to accelerate yield ramp by clustering and classifying manufacturing failure patterns

Authors

Artificial Intelligence Group

Nitzan Kalvari
AI Product Manager, Intel IT

Nir Lotan
AI Products Team Manager, Intel IT

Merav Zidon
Business Line Manager, Intel IT

Intel Fab Sort Manufacturing

Robert E. Lofgren
Yield Analysis Engineer

Table of Contents

Executive Summary

Expert yield analysis engineers have always performed end-of-line yield analysis at Intel's silicon wafer factories (fabs). But as the number of products and volume grow in Intel's manufacturing environment, a manual detection approach to yield analysis poses several challenges:

- Limited human-hour resources prevent engineers from reviewing and documenting every issue in every wafer in every lot.
- Detection accuracy depends on an engineer's experience level.
- Knowledge sharing between fabrication sites is cumbersome and slow.

Intel is changing the paradigm of yield analysis from this manual, reactive "pull" approach to a proactive "push" approach, which is helping to find problems such as failing tools, fleet mismatches and process parameter shifts, quickly and accurately. The more quickly such issues are identified, the sooner they get fixed and overall yield is improved.

The solution is characterized by the following:

- **Advanced machine-learning and deep-learning algorithms** for issue detection, clustering, classification, data gathering and in the future, root cause analysis.
- **Autonomous end-to-end detection**, where the above tasks are performed automatically and then the results are pushed to the yield analysis engineers for further investigation.
- **Tight integration** with existing yield analysis tools and systems.

Our unique solution enables end-of-line issue detection to identify multiple issues on the same wafer, and to examine 100 percent of wafers in every lot. The combination of artificial intelligence (AI) and yield analysis engineers' knowledge enables them to support more products, use knowledge captured collectively across fabs and shorten time to resolution. Overall, this solution is propelling us along our Industry 4.0 journey toward complete automation of the root cause analysis process and better yield.

Business Challenge

Intel is one of the world's leading high-volume manufacturers, with 15 wafer fabrication plants (fabs) in production worldwide at 10 locations. Like all manufacturers, Intel strives to improve manufacturing yield without driving up costs. Artificial intelligence (AI) has enormous potential to help achieve this goal, moving us closer to the Industry 4.0 vision of complete automation of manufacturing processes.

In semiconductor manufacturing, a single silicon wafer is composed of tens to hundreds of individual microelectronic integrated circuit units called dies. Wafers are produced in "lots," meaning product that is manufactured during a specific time period. Each wafer undergoes many manufacturing line (also known as "inline") steps as it moves through the fab, each of which involves a complex interplay between state-of-the-art manufacturing tools and highly advanced electro-chemical-mechanical processes. Various problems can occur, such as a tool beginning to fail, the fleet of tools running mismatched or a change in one processing step inadvertently impacting another processing step. All of these issues—and many others—can introduce manufacturing line variability, negatively impacting the end-of-line yield.

Yield analysis engineers inspect end-of-line wafer for die-level functional health indicators. One thing they look for are gross failure areas (GFAs),[1] which appear as patterns that indicate a problem has occurred in the fab. Different problems cause different-looking patterns (see Figure 1). Until recently, the yield analysis engineers used a manual technique to deduce what went wrong inline from the end-of-line perspective. This exercise in pattern recognition serves as input for root cause analysis of an issue. Over multiple years of experience with wafer analyses, yield analysis engineers have cataloged dozens of baseline patterns that relate to specific inline problems.

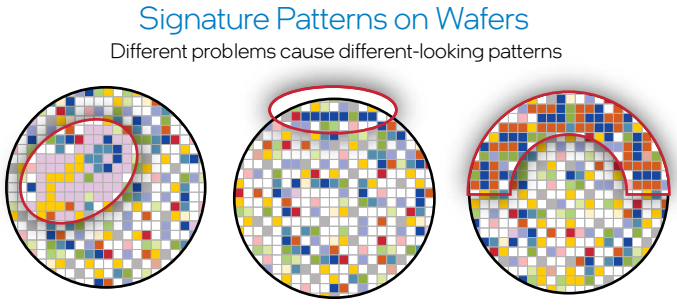

Figure 1. Different inline problems with manufacturing tools and process manifest in different patterns on wafers.

The typical "pull" approach to identifying a GFA involves many challenges:

- **Time-consuming.** GFA detection is a repetitive and labor-intensive job. And, as the number of products and volume grow in Intel's manufacturing environment, it is impractical to hire enough yield analysis engineers to review and document 100 percent of the end-of-line material. Due to human resource constraints, this manual approach is time-consuming and non-scalable.
- **Limited experience.** It takes many years for yield analysis engineers to gain experience in manufacturing process technologies to perform GFA detection accurately. This task tends to be a balance between art and science; engineers need many years of experience to learn how to distinguish accurately between random statistical "noise" and real GFAs/patterns/signatures. Thus, results of analysis and consistency depend heavily on the engineer's skill.
- **Convoluted by multiple failures.** Two or more inline problems may have affected a wafer, potentially leading to multiple patterns on one wafer. Due to resource and experience constraints, yield analysis engineers may identify and characterize only one inline problem, while the other problem(s) go undetected and unresolved, which may hinder fixing the root cause of both issues.
- **Siloed information.** Because Intel manufactures wafers in several sites in parallel, knowledge transition between yield analysis engineers requires meetings, which can slow knowledge sharing.
- **Delayed issue detection.** New issues can appear that are not on the list of baseline patterns. These unknown issues may go undetected until their repeat occurrence is captured by an experienced human eye. Delays in new issue detection due to limited visual sampling or experience may come at a significant cost to manufacturing health and yields.

  Speedy detection and quantification of material at risk due to a fab event or excursion is highly critical. Intel's fabs run 24/7 and process thousands of wafers every hour. The longer a failed tool, mismatched fleet or an unintentional process shift runs uncontained, the more the material is at risk of degraded yields. An automated GFA classification solution can help improve yield by alerting the yield analysts of inline problems that can then be quickly addressed—preventing even more wafers from being affected.

As Intel's product portfolio expands and becomes more complex, the business risk of undetected issues, incorrect signature attribution and the time it takes to identify even known signatures continues to grow. Intel IT is committed to helping Intel Manufacturing accelerate issue detection, improve accuracy and provide multi-product coverage for the issue detection cycle through an automated AI-based solution (see Figure 2).

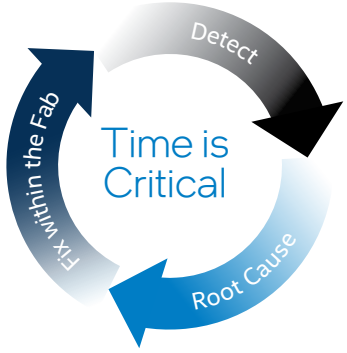

Figure 2. To minimize inline problems' impact on yield and cost, we engage in an ongoing—and time-critical—cycle of issue detection, root cause analysis and in-fab correction of problems.

[1] Although not an industry-standard term, we commonly refer to GFAs as "crashes" and hence to our AI engine that automates GFA detection as "Auto Crash."

Solution Overview: Automated, Integrated GFA Detection

Overall, integrating AI into all of Intel's manufacturing processes to solve a variety of manufacturing challenges is part of our vision for the Factory of the Future. As part of that effort, we have developed a transformational approach to improving and accelerating yield analysis. AI performs the repetitive, labor-intensive detection, then pushes the results to yield analysis engineers for root cause analysis.

The solution encompasses three key elements that make our solution unique in the industry:

- **AI model.** We developed a dedicated AI workflow that uses machine-learning, deep-learning and image-processing techniques to perform automated pattern recognition. AI can identify and document multiple GFAs per wafer, and learn to capture patterns that affect yield.
- **Autonomous end-to-end detection.** While the algorithm is important to the success of the overall solution, automation is the real game-changer. The legacy GFA tools were limited and required manual intervention and manual queries. Now, the automated push approach produces data for root cause analysis and calculates yield impact trends.
- **Holistic integration.** The algorithms' results are seamlessly integrated into the existing manufacturing workflow methods and tools, which improves ease of use and our ability to extend the algorithms' business value to additional use cases.

Other important aspects of the solution include structured data collection without manual documentation, a shift from monotonous pattern classification to focus more on root cause analysis and replacing local execution with a central system that can process and store far more data than is possible on a local client.

One important note: Our solution is not intended to replace yield analysis engineers. Instead, machines execute what they do best, and the engineers perform more complex intellectual tasks, such as applying business knowledge and finding the root cause of detected issues.

The combined solution—algorithm, automation and integration—currently provides two services:

- **Baseline pattern detection.** For 100 percent of end-of-line wafers, the solution can use the pattern examples provided to it to automatically determine if a wafer has a known (baseline) inline issue, without manual intervention. This aspect of the solution looks for issues we know exist in the manufacturing environment to some extent.
- **Unknown pattern detection.** The solution can also report information about all the patterns that are currently impacting yield and the level of impact. Yield analysis engineers can use the report to identify new insights, such as a new pattern, a known pattern that has a changed definition or a change in level of yield impact. This latter information can help engineers set their investigation priorities. Once the engineers complete the root cause analysis for a previously unknown pattern, the new pattern is added to the baseline pattern repository, and the AI model is retrained to recognize it.

This solution accelerates the speed at which inline problems can be identified, tagged and subsequently resolved for increasing total yields. The solution can help yield analysis managers increase yield in a talent-constrained environment. In other words, the solution helps maintain a high-quality product without additional yield analysis engineering head count.

While the manufacturing industry has made previous steps toward automating GFA detection, our use of machine-learning operations (MLOps) to drive acceleration and scalability is unique. The solution is providing transformational business value (see Table 1).

Table 1. Transformational Business Value Accruing from AI-based GFA Pattern Detection

	Manual Analysis	AI-based Analysis
Benefits of AI	Limited scalability: Only a subset of end-of-line volume on one product or across multiple products is analyzed.	Highly scalable—*every wafer of every lot* is analyzed, which catches more issues, and can be used as an accurate dataset for root cause investigation. The solution can quickly expand to multiple products.
	Limited to a single gross failure area (GFA) identification and documentation per lot.	Multiple GFAs can be found and documented per wafer.
	Quality of issue identification is based on a yield engineer's experience—it may be biased, and differs from person to person.	Consistent, objective and reproducible tagging of known patterns, plus the ability to find new GFAs for investigation. As the solution improves over time, it can potentially achieve experienced human-level accuracy for all baseline patterns for a variety of products.
Benefits of end-to-end automation	"Pull" reactive approach for GFA search.	"Push" approach to help automatically detect active GFAs.
	Requires about two days to update the aggregated yield impact summary report.	Yield impact summary is automated.
	Dataset creation for root cause analysis, trend and yield impact calculation is labor-intensive.	Enables easy dataset generation for root case analysis, including trends over extended periods of time and yield impact calculation.
	Knowledge sharing (see the sidebar, "A Closer Look at the Virtual Factory") is based on meetings and presentation materials, which can slow down inline fixes.	Virtual Factory integration is automatic and fast, enabling easy knowledge-capturing and sharing, which in turn speeds inline fixes.
Benefits of holistic integration with manufacturing environment	Analysis results are isolated from other analysis and data exploration tools.	Results are fully integrated with existing processes and tools, making the solution easy to adopt and use.

A Closer Look at the Virtual Factory

Sharing solutions across factories leads to increased manufacturing efficiency and quality

Intel implemented a "Virtual Factory" concept nearly 20 years ago. The foundational assumption is that Intel's factories have many commonalities, so sharing solutions and information across all sites helps eliminate unnecessary effort and allows every factory to benefit from a breakthrough solution or idea. Whether it is an ergonomic solution, a new Manufacturing Execution System (MES) or an upgrade to a fab tool, once validated, the change is "Copy Exactly!" to all the factories. Our AI-based solution that automates GFAs on end-of-line wafers is no exception. We have integrated the solution into the Virtual Factory network. When the solution finds a new gross failure area (GFA) pattern and the yield analysis engineers complete their root cause analysis, the new pattern can be added to the list of known patterns at all fabs—improving yield not only at the fab at which the issue was found, but at all of Intel's fabs around the world.

Solution Architecture

Our AI-based GFA detection solution is part of our overall effort to use AI throughout Intel's business processes to improve business outcomes (see the sidebar, "Using MLOps to Accelerate AI Model Productization"). The solution consumes data from the existing fab data lake, such as a list of wafers that previously suffered from a particular pattern signature and information about the overall wafer population. The AI-based models are trained using the baseline patterns. Once training is complete, the models are inferenced on all streaming material and provide classification results, along with yield impact measurements, to existing yield analysis tools. In parallel, AI models are running to identify new patterns that exist on wafers.

Tight integration with existing downstream visualization and analysis tools enables engineers to perform deep-dives into the data when warranted. They can also perform additional custom analysis if necessary. For example, the yield analysis operational management tool allows engineers to customize their analysis views of data and also allows fab users to have simplified web-based assessments for data viewing and data entry. Other tools integrated with the AI-based GFA detection solution include a root cause analysis tool and a dashboard and reporting tool.

The solution is designed for scalability. We use industry-standard software throughout (see Table 2), and we also use a modular approach so that we can add more models as the solution grows, to support additional product types. The algorithms combine several models that include both machine learning and deep learning. In late 2021, we had 16 models in production tagging about 2,500 wafers per day—and these numbers continue to grow as we add new patterns to the baseline pattern repository and add support for additional products.

Table 2. Automated GFA Detection Solution Software

Component	Technologies
Programming Language	Python
Machine- and Deep-learning Frameworks	Python, TensorFlow, Seldon
Analytics Orchestration	Argo, Apache Kafka based on the Confluent Platform
Storage	MinIO database, Network File System (NFS), ElasticSearch
Containers	Docker, Kubernetes
OS	Linux, Ubuntu

The solution is built using cloud-native microservices running on a central server. Both model training/retraining and inference run in a private cloud, using a 20-node Kubernetes cluster equipped with Intel® Xeon® Silver 4215R processors (see Figure 3). The cluster, which also runs several other manufacturing AI solutions, serves all of Intel's fabs around the globe.

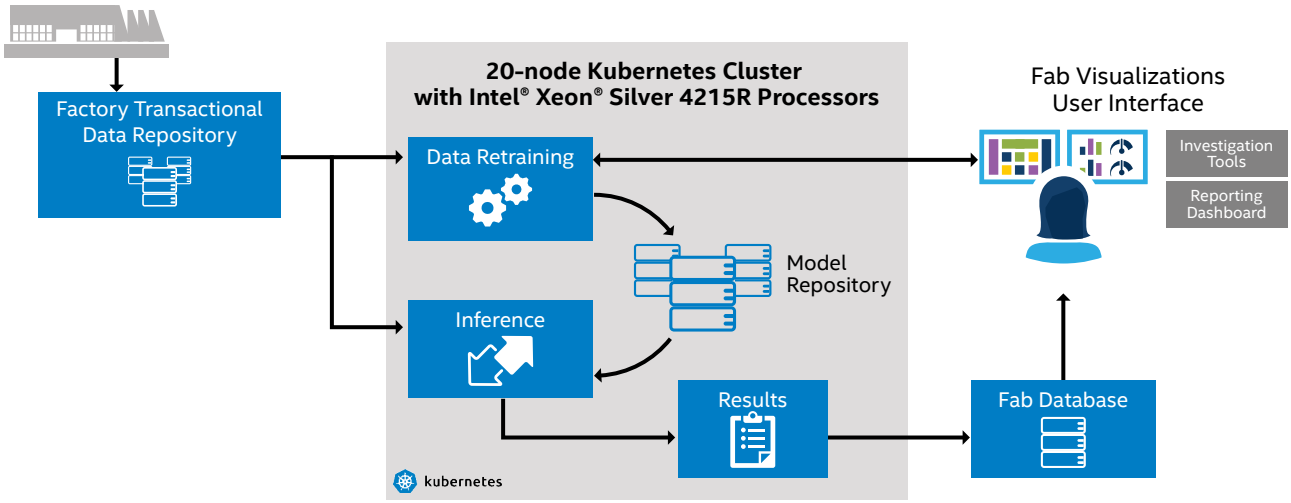

Figure 3. Our AI-based solution for detecting gross failure areas (GFAs) includes input data, the ability to retrain the models and full integration of results with existing analysis and visualization tools and processes.

Results

The success of our use of AI for end-of-line quality control and root cause analysis can be measured in several ways—decrease in defect density, GFA detection accuracy and coverage are several metrics that we track. Our results include the following achievements:

- Early detection of GFAs, including unknown issues that could not have been identified by humans.
- Detection of multiple GFAs on a single wafer, enabling multiple root causes to be fixed simultaneously.
- Solution deployment to all factories producing Intel's leading products, enabling emerging issues in one factory to all be communicated to all the factories.
- 100 percent coverage of the wafers and lots.
- >90 percent accuracy in detecting baseline patterns.

By adding AI to the yield analysis process and integrating the overall solution into the manufacturing environment using automation, we have transitioned from a manual "pull" approach to a far more efficient "push" approach. This helps us maintain the high yield we had in previous process generations while keeping the same level of yield analysis staff. Plus, the yield improvement we are gaining can enable earlier release of products to market.

Next Steps

We continually improve the AI-based GFA detection solution by retraining models and adding new models and algorithms. Additional improvements include the following:

- Correlate inline data to patterns detected at end-of-line to provide more precise inputs for root case analysis.
- When applicable, perform automated root case analysis and provide details about the tool, process or parameters that caused the issue.
- Link patterns between different fabs.
- Broaden the solution's use of historical information to provide additional insights. For example, the solution could point out that a similar pattern was handled by a particular yield analysis engineer. Also, expanded historical data may uncover commonalities between patterns.
- Incorporate employee feedback with regard to new issues and the system's performance in the case of false detection.
- Scale to more products and new silicon process technologies.

Using MLOps to Accelerate AI Model Productization

Deploying AI faster and enabling self-maintaining, cost-effective AI services in production at scale

Intel IT works with Intel Manufacturing to apply artificial intelligence (AI) across Intel to transform critical work, optimize processes, eliminate scalability bottlenecks and generate significant business value (more than USD 1.5B return on investment in 2020). Our efforts unlock the power of data to make Intel's business processes smarter, faster and more innovative, from product design to manufacturing to sales and pricing.

To enable this operation at scale, we developed Microraptor, which is a set of machine-learning operations (MLOps) capabilities that are reused in all of our AI platforms. Microraptor enables world-class MLOps to accelerate and automate the development, deployment and maintenance of machine-learning models. Our approach to model productization helps avoid the typical logistical hurdles that often prevent other companies' AI projects from reaching production. Our MLOps methodology enables us to deploy AI models to production at scale through continuous integration/continuous delivery, automation, reuse of building blocks and business process integration.

Our MLOps methodology provides many advantages:

- The AI platforms abstract deployment details and business process integration so that data scientists can concentrate on model development.
- We can deploy a new model in less than half an hour, compared to days or weeks without MLOps.
- Our systematic quality metrics minimize the cost and effort required to maintain the hundreds of models we have in production.

For more information, read the IT@Intel white paper, "Push-button Productization of AI Models."

Conclusion

We are committed to providing continuous innovation that will improve the quality and velocity of Intel's manufacturing environment. We are taking a unique approach to GFA detection, using AI and automation to transform end-of-line yield analysis. The solution we developed for autonomous end-to-end issue detection achieves greater than 90 percent accuracy in baseline pattern recognition and can now identify multiple GFAs per wafer, enabling us to perform root cause analysis on several issues at once to improve wafer quality. Our solution is tightly integrated into the existing manufacturing tools, such as those used for data visualization, making it easy for Intel Manufacturing staff to adopt and use the solution.

Although our current solution is specific to GFA detection on silicon wafers, our overall push approach to yield analysis can be applied to other AI-based product types:

- Use AI to mimic experts work, where machines execute better than humans. Refocus experts in more complex intellectual tasks, such as applying business knowledge. In our case, conducting root case analysis.
- Autonomous end-to-end process provides users with the output they can evolve to future needs. In our case, all active GFAs, known and new.
- Adopt the solution easily and quickly, and integrate it with existing processes and tools.
- Plan for scale in order to increase business impact and value.

Related Content

If you liked this paper, you may also be interested in these related stories:

- Push-Button Productization of AI Models white paper
- Building Intel's AI Center of Excellence blog
- Improving Sales Account Coverage with Artificial Intelligence white paper
- Revolutionizing Product Validation Using AI white paper
- Developing a Scalable Predictive-Maintenance Architecture white paper

For more information on Intel IT best practices, visit **intel.com/IT**.

IT@Intel

We connect IT professionals with their IT peers inside Intel. Our IT department solves some of today's most demanding and complex technology issues, and we want to share these lessons directly with our fellow IT professionals in an open peer-to-peer forum.

Our goal is simple: improve efficiency throughout the organization and enhance the business value of IT investments.

Follow us and join the conversation on Twitter or at #IntelIT.

Visit us today at intel.com/IT or contact your local Intel representative if you would like to learn more.

Contributors

Alex Freilikhman, Yield Analysis Engineer, Intel Fab Sort Manufacturing

Andrew Marin, Yield Analysis Engineer, Intel Fab Sort Manufacturing

Amrish Menjoge, Yield Engineering Manager, Intel Logic Technology Development

Shai Monzon, Industry 4.0 Engagement Manager, Intel IT

Yossi Revah, ISR Analytics Manager, Intel Manufacturing and Operation Automation

Acronyms

AI	artificial intelligence
fab	wafer fabrication plant
GFA	gross failure area
MLOps	machine-learning operations

1221/WWES/KC/PDF