Scheduling and Dispatching in Semiconductor Fabs

The Context

A semiconductor factory (fab) produces chips by processing silicon wafers through hundreds of precise manufacturing steps.

- Wafers move in batches, called lots, with each lot containing multiple wafers.
- After each processing step, lots wait in queues before moving to the next step. This waiting inventory is called Work-In-Process (WIP).
- Lots must share expensive, specialized machines (such as lithography and etching tools), which typically handle only one lot at a time.
- Some lots, known as hot lots, are high-priority due to customer deadlines or critical orders.
- The fab environment is dynamic: machines can fail, hot lots can arrive unexpectedly, and queues can grow unpredictably.

The Challenge

Deciding which lot to send to which machine and when is known as scheduling and dispatching. Poor decisions can lead to:

- Long waiting times for lots, increasing cycle time.
- Idle machines, wasting costly resources.
- Hot lots missing their deadlines.
- Bottlenecks and congestion at certain machines.

Objectives

The main goals of effective scheduling and dispatching are:

- Minimize total waiting and processing time for lots.
- Maximize machine utilization and productivity.
- Ensure high-priority (hot) lots are completed on time.
- Adapt schedules quickly to real-time changes, such as equipment failures or rush orders.

Benefits

Achieving these objectives provides significant benefits:

- Faster delivery of critical orders.
- Higher factory throughput and efficiency.
- Greater flexibility and resilience in operations.

Solution Approach

Use reinforcement learning, machine learning, and optimization algorithms to build an adaptive scheduling and dispatching system that prioritizes hot lots, balances machine utilization, and dynamically responds to real-time fab conditions to minimize cycle time and maximize efficiency.

Automated Material Handling Systems (AMHS): Routing and Collision Avoidance for OHTs

Semiconductor factories (*fabs*) rely heavily on **Overhead Hoist Transport (OHT)** systems to automatically move silicon wafers between processing tools. OHTs operate on an overhead rail network and carry wafer containers (*Front Opening Unified Pods — FOUPs*), enabling fully automated and continuous material handling.

However, the highly complex and dynamic environment of a fab creates significant operational challenges for OHT systems, which can reduce throughput, increase wafer cycle time, and risk wafer damage.

Key challenges specific to OHTs include:

- Traffic congestion at intersections and shared track segments, creating bottlenecks and delays.
- Rigid, static routing, which fails to adapt to changing fab conditions such as congestion or tool availability.
- Inefficient FOUP dispatching and hot-lot prioritization, causing delays in urgent orders and starving critical tools.
- Lack of dynamic rerouting when a track segment becomes blocked or a tool goes down.
- Collisions or near-misses at junctions or merge points, increasing risk to wafer quality and equipment.

These challenges highlight the need for intelligent, adaptive, and coordinated control of OHT systems, improving fab efficiency, responsiveness to dynamic conditions, and operational safety.

Benefits

- Dynamic Routing: AI adapts paths in real-time to avoid congestion and blocked tracks.
- Hot-Lot Prioritization: Ensures urgent wafers move faster, reducing cycle time.
- Collision Prevention: RL minimizes near-misses at junctions for safer operations.
- Tool Starvation Avoidance: Smart dispatching keeps critical tools supplied.
- Scalability: Handles growing OHT fleets without manual tweaks.
- Cost Savings: Fewer delays, less wafer damage, and lower downtime.

Solution Approach

Develop an **intelligent, adaptive control framework** using reinforcement learning and predictive modeling to optimize routing, dispatching, and collision avoidance of OHT systems in dynamic fab environments.

Defect Source Detection in Semiconductor Manufacturing Using GenAI-Based Image and Video Analytics

The Context

Semiconductor manufacturing (fabs) involves highly precise and defect-sensitive processes such as wafer fabrication, lithography, etching, and packaging.

- Each wafer undergoes hundreds of **inspection** steps, generating high-resolution images and videos.
  - **Inspection:** "Is there anything wrong here?" — detecting visible defects or anomalies.
- These data are vital for identifying defect sources, monitoring process quality, and ensuring alignment.

- Manual and rule-based systems often fail to capture rare, subtle, or evolving defect modes and require constant reprogramming.
- Video feeds also hold insights on operational inefficiencies, equipment anomalies, and safety violations — but remain underutilized.

The Challenge

- Rare and novel defect sources often go undetected or misclassified, while false positives/negatives increase scrap and rework.
- Traditional models degrade over time due to process drifts and evolving defect modes.
- Identifying root causes of observed defects across steps is complex and time-consuming.
- High-resolution image/video streams are difficult to analyze consistently and at scale.
- Video monitoring for operational anomalies remains largely manual and reactive.

Objectives

- Detect and trace defect sources, including novel and rare patterns, with high accuracy.
- Minimize false alarms, scrap, and rework while improving traceability.
- Continuously adapt to process changes and emerging failure modes.
- Automate real-time monitoring of equipment and operations for quality assurance.
- Scale analytics effectively across large, high-volume image and video datasets.

Solution Approach

Deploy a GenAI-powered image and video analytics platform to detect, trace, and explain defect sources and operational anomalies in real time, while adapting to evolving processes and minimizing false alarms.

Virtual Metrology for Quality Control in Semiconductor Manufacturing Using AI

The Context

Virtual Metrology (VM) is an AI-driven technique that predicts wafer measurements without physical testing. It uses real-time process data (e.g., sensor readings, equipment parameters) and machine learning to estimate key quality metrics (e.g., film thickness, etch depth, critical dimensions) that traditionally require slow, costly metrology tools.

In semiconductor fabs, physical metrology ensures wafers meet specifications but is expensive, time-consuming, and samples only a fraction of wafers. At the same time, fabs generate abundant real-time process data that correlates with wafer quality but remains underutilized.

The Challenge

- Physical metrology introduces delays, high costs, and limited sampling.
- Undetected process drifts can lead to scrap and yield loss.
- High-dimensional, noisy process data and variability make modeling challenging, and models degrade over time.

Objectives

- Predict wafer quality metrics directly from real-time process data, enabling 100% virtual coverage.
- Provide fast, robust, and adaptive feedback for process control and yield improvement.
- Reduce metrology costs and improve throughput without compromising quality.

Solution Approach

Deploy an AI-powered Virtual Metrology system that predicts wafer quality outcomes from process and equipment data using advanced DL techniques — to model complex dependencies, adapt to drift, and deliver fast, accurate predictions.

Benefits

- Faster feedback and timely corrective actions.
- 100% wafer monitoring, improved yield, and process stability.
- Lower metrology costs with higher confidence in quality assurance.

AI-Driven Predictive Process Control and Maintenance in Semiconductor Manufacturing

The Context

Semiconductor manufacturing requires tightly controlled processes and high equipment uptime to ensure wafer quality, yield, and productivity.

- Equipment parameters like gas flow, temperature, and pressure must be optimized in real time to maintain process stability and avoid deviations.
- Equipment failures (e.g., in pumps, valves, robots) can cause unexpected downtime, scrap, and yield loss.
- Modern fabs collect rich sensor and operational data that can enable AI-driven predictions and proactive interventions.

The Challenge

- Process parameters can drift due to wear, contamination, or environmental changes, degrading wafer quality and yield.
- Traditional process control is reactive and struggles to adjust to dynamic variations quickly.
- Equipment failures are often detected too late, leading to unplanned downtime, costly repairs, and production losses.
- Manual monitoring of large volumes of sensor data is infeasible and error-prone.

Objectives

- Predict wafer measurements and adjust process parameters dynamically to maintain stability and maximize yield.
- Detect early signs of equipment degradation and schedule maintenance proactively before failures occur.
- Improve process consistency, reduce scrap, and enhance equipment availability.
- Leverage existing sensor and operational data streams effectively.

Solution Approach

Deploy AI-driven predictive process control and predictive maintenance systems that use machine learning models (e.g., regression, deep learning, anomaly detection) to analyze real-time sensor and equipment data, optimize process parameters dynamically, and identify early signs of equipment failure for planned maintenance.

Benefits

- Improved wafer quality and maximized yield through dynamic process optimization.
- Reduced process deviation and scrap rates.
- Increased equipment uptime and reduced unplanned maintenance costs.
- Higher productivity and more stable operations.

AI-Driven Recipe Optimization in Semiconductor Manufacturing

The Context

In semiconductor manufacturing, a **recipe** defines the precise process settings (e.g., gas flow, RF power, pressure, temperature) and sequences executed by tools during wafer processing steps like etching, deposition, or annealing.

- Recipes directly impact wafer quality, yield, and tool productivity.
- Traditionally, recipes are developed and tuned through manual, trial-and-error experimentation, which is time-consuming and suboptimal.
- As product complexity and variability increase, efficient and adaptive recipe optimization becomes critical for maintaining productivity and quality.

The Challenge

- Manual recipe development is slow, costly, and prone to suboptimal settings.
- Process interactions are highly complex and nonlinear, making it difficult to predict outcomes of parameter changes.
- Finding optimal trade-offs between competing objectives (e.g., uniformity vs. throughput) is challenging.
- Variations in tools, products, and environments require frequent retuning.

Objectives

- Automatically generate optimal process recipes that achieve target wafer specifications and maximize yield.
- Reduce recipe development time and engineering effort.
- Adapt recipes to new products, tools, and conditions efficiently.
- Balance competing objectives like wafer quality, cycle time, and throughput.

Solution Approach

Deploy an AI-driven recipe optimization platform that leverages process simulations, predictive models, and optimization algorithms (e.g., Bayesian optimization, reinforcement learning) to explore the process parameter space, predict outcomes, and generate optimal recipes that meet desired specifications and improve productivity.

Work Context Analysis

Type of work: Industry-specific, purpose-driven

Work Description:

- **Industry:** Semiconductor manufacturing
- **Function / Business Process:** Manufacturing operations optimization, quality control, and process automation
- **Relevance:** Directly influences both the top line (by improving yield, accelerating time-to-market, and enhancing product quality) and the bottom line (through operational efficiency, reduced waste, lower maintenance costs, and optimized resource utilization)
- **Significance:** Makes a substantial contribution to revenue growth and cost reduction by improving yields, reducing scrap, optimizing processes, and enabling predictive, adaptive manufacturing operations

High-level description of work:

- **purpose:** Transform semiconductor manufacturing through AI-driven optimization of critical fab operations to achieve higher yields, faster production cycles, improved quality, and lower costs
- **what it involves:** Implementing machine learning and AI systems across six key areas: scheduling and dispatching, automated material handling, defect detection, virtual metrology, predictive maintenance, and recipe optimization
- **what it produces:** Intelligent manufacturing systems that autonomously optimize production parameters, predict equipment failures, detect quality issues in real time, and adapt dynamically to changing conditions — leading to significant improvements in operational efficiency, product quality, and manufacturing competitiveness

This work represents a strategic digital transformation initiative that directly enhances the competitiveness and profitability of semiconductor manufacturing operations.

Work Stakeholders & Actors

Buyer Personas

Primary Buyer Personas:

- VP of Manufacturing/Operations — Responsible for production efficiency, operational excellence, and fab performance
- Plant/Fab General Manager — P&L owner of the manufacturing facility
- CTO (Chief Technology Officer) — Shapes technology strategy and enables innovation initiatives
- Chief Digital Officer (CDO) — Champions digital transformation initiatives

Secondary Influencers:

- CEO — Provides strategic approval for major capital and transformation initiatives
- CFO — Approves budgets and ensures financial viability

Actors and Users Involved in the Work

Direct Users:

- Manufacturing Engineers — Configure and optimize AI algorithms for specific processes
- Process Control Engineers — Monitor virtual metrology predictions and predictive maintenance alerts
- Production Planners/Schedulers — Operate AI-driven scheduling and dispatching systems
- Quality Engineers — Analyze defect detection outputs and investigate root causes
- Fab Operators — Respond to AI system alerts and act on dashboard notifications
- Maintenance Engineers/Technicians — Act on predictive maintenance recommendations
- Operations Analysts/Industrial Engineers — Review KPIs, trends, and optimization reports

Supporting Actors:

- Data Scientists/AI Engineers — Develop, train, validate, and maintain AI models
- IT/OT Teams — Integrate AI solutions into MES/ERP and shop floor systems
- Equipment Vendors — Provide data interfaces, integration support, and service alignment

Current Best-Practices

How is the work done today (what is the best-practice)?

Current Approaches:

- Manual scheduling and dispatching using MES systems with rule-based algorithms
- Static material handling routing with predetermined paths for OHT systems
- Sampling-based quality control using physical metrology tools on limited wafer quantities
- Reactive maintenance based on scheduled intervals or equipment failures
- Rule-based defect detection using traditional image processing and statistical process control
- Manual recipe optimization through design of experiments and expert knowledge

What are the work effectiveness, performance, and quality measures?

Key Performance Indicators:

- Cycle Time — Time from wafer start to finish
- Overall Equipment Effectiveness (OEE) — Combination of equipment availability, performance, and quality
- Yield and First Pass Yield — Percentage of wafers meeting specifications without rework
- Throughput and Equipment Utilization — Production rate and proportion of productive equipment time
- Mean Time Between Failures (MTBF) — Measure of equipment reliability
- Defect Density — Number of defects per unit area of wafer
- Work-in-Process (WIP) — Inventory levels of wafers in the fab

What are representative values and trends of these KPIs?

Current Industry Benchmarks:

- Exact KPI values vary widely depending on technology node, fab maturity, and product type, but current practices aim to optimize all the above measures to improve competitiveness.

Historical Trend: Traditional approaches have delivered incremental improvements in these KPIs over time, but gains have diminished, creating an opportunity for AI-driven transformation to achieve breakthrough performance.

Elevator Pitch Narrative for the Future Proposition

What is the aspirational target state for work performance and value? What will be considered significant improvement in the domain?

Aspirational Target State:

- Shortened cycle time through optimized scheduling and material flow
- Improved overall equipment effectiveness (OEE) with predictive maintenance and intelligent dispatching
- Higher yield through real-time defect detection and process optimization
- Increased equipment utilization through AI-driven scheduling and analytics
- Comprehensive quality coverage enabled by virtual metrology and predictive quality control
- Proactive, predictive maintenance that minimizes unplanned downtime
- Faster recipe development through automated, AI-driven parameter tuning

Significant Improvement Threshold: Achieving meaningful gains across these core KPIs is viewed as transformational in semiconductor manufacturing, enabling better efficiency, quality, and responsiveness.

What will be the propensity to buy a solution that achieves the target state?

High Propensity to Buy Drivers:

- Competitive pressure to adopt advanced manufacturing capabilities
- Need to maximize return on high-capital equipment investments
- Imperative to ensure zero-defect quality through predictive controls
- Demand for faster time-to-market and improved profitability
- Industry-wide trend toward AI adoption as a differentiator
- Clear and measurable impact on operational and financial performance

Expected Purchase Readiness: Strong among leading semiconductor manufacturers and progressive fabs aiming for operational excellence and competitive advantage.

Key Challenges & Ideas

What are the challenges in achieving the aspirational target? What makes achieving the target state difficult?

Primary Challenges:

- Data and system integration — Connecting heterogeneous data from hundreds of tools with real-time processing across legacy MES and equipment systems
- AI reliability and trust — Ensuring AI predictions are accurate and robust enough to support critical production decisions in regulated environments
- Organizational transformation — Managing the shift from manual, expertise-based operations to AI-driven automation while maintaining compliance and change management

What are the unique ideas that we believe will allow us to overcome the challenges?

Unique Solution Approaches:

- Leveraging fab-scale digital twins to validate and deploy AI models in realistic virtual environments before production rollout
- Implementing hybrid human-AI intelligence systems where AI augments human expertise with continuous learning and adaptation
- Building domain-specific AI frameworks tailored to semiconductor manufacturing physics, constraints, and operational requirements

How differentiated will the ideas be? Will the differentiation be sustainable? What will we need to do to sustain differentiation?

Differentiation Strategy:

- High differentiation — Semiconductor-specific AI models combined with deep fab domain expertise create significant competitive barriers
- Sustainability approach — Continuous R&D investment, proprietary data advantages, comprehensive customer ecosystems, and ongoing development of specialized AI talent
- Platform lock-in — Delivering integrated solutions that become indispensable infrastructure for fab operations

Market Opportunity

What is the market size (potential) for this solution?

Market Opportunity:

- The global semiconductor manufacturing industry is a large, established market with hundreds of advanced fabs operating worldwide
- AI and software solutions for smart manufacturing are a rapidly growing segment within the broader industrial automation market
- Target customers include major semiconductor manufacturers, such as leading foundries and integrated device manufacturers
- Geographic focus on regions with high concentrations of advanced fabs, including Asia (Taiwan, South Korea, China), North America, and Europe

What will it replace?

Displacement Targets:

- Manual scheduling and process control systems that depend on expert intervention and slow optimization cycles
- Legacy MES systems lacking real-time optimization and predictive analytics
- Reactive maintenance and quality control based on fixed schedules and limited sampling
- Rule-based defect detection and material handling systems with static routing and high false-positive rates

How difficult will it be to sell? What will be the most significant hurdles in selling this solution?

Key Sales Challenges:

- Risk aversion and stringent validation requirements, as fabs demand extensive proof of reliability and phased implementation to avoid production disruptions
- Technical and integration complexity due to deep interfaces with existing equipment, control systems, and compliance requirements
- Significant investment and organizational change management required to transition from expert-driven to AI-driven operations

Timeline & Budget Estimations

How long will it take to build the solution?

Development Timeline:

- Phase 1 — Foundation: Core AI platform development, data integration framework, and initial algorithm prototypes
- Phase 2 — Pilot Implementation: Single-use case deployment, customer validation, and iterative refinement
- Phase 3 — Full Platform: Complete multi-module solution with all six AI applications integrated
- Overall development is expected to progress across these phases before reaching a comprehensive market-ready solution

How long will the opportunity last? Is it time-bound?

Market Window:

- Current opportunity exists to establish an early-mover advantage, as the semiconductor industry is still in the early stages of AI adoption
- The market is expected to evolve over the coming years as adoption increases and competition grows
- Technology advancement in semiconductor nodes ensures continuous demand for optimization and new AI capabilities
- The opportunity is not strictly time-bound, as fabs consistently seek operational improvements

What resources will be needed to become market ready?

Resource Requirements:

- Technical team: AI/ML engineers, semiconductor domain experts, software architects, and systems integration specialists
- Partnership network: Equipment vendors, system integrators, and pilot customer collaborations
- Infrastructure: Cloud computing resources, development environments, and testing facilities
- Regulatory and compliance: Quality assurance expertise aligned with semiconductor industry standards
- Sales and marketing: Industry-focused business development and technical sales capabilities
- Capital investment: Funding for R&D, talent acquisition, infrastructure, and pilot programs

StructiDoc AI – Document Ingestion as a Service: Unlocking Unstructured Data in Semiconductor Operations

The Context

Semiconductor fabs generate and rely on massive volumes of **unstructured and semi-structured operational data and documentation**, including:

- **Process Operations Manuals & Work Instructions:** SOPs, emergency response procedures, recipe tuning guides, process flow diagrams.
- **Equipment Logs & Maintenance Records:** Handwritten or scanned logs, calibration sheets, failure reports, PM checklists.
- **Statistical Process Control (SPC) Charts:** Printed control charts, exception reports, Cpk trend analyses, and process capability studies.
- **Wafer Processing Documentation:** Wafer travelers, lot cards, recipe files, process specifications, and run sheets.
- **Inspection & Quality Reports:** Annotated wafer maps, die-level defect patterns, audit checklists, metrology certificates, yield reports, and pareto analyses.
- **Engineering Drawings & Layouts:** Cleanroom layout drawings, tool placement maps, utility distribution schematics, contamination control zone diagrams.
- **Material & Supplier Documentation:** Certificates of Analysis (CoA), incoming inspection reports, chemical inventory logs, supplier qualification records.
- **Regulatory & Safety Documents:** MSDS sheets, environmental impact forms, OSHA compliance reports, ITAR documentation, export control records.

These documents are central to maintaining yield, uptime, compliance, and continuous improvement — but their **unstructured nature and inconsistent formats make them hard to integrate into AI-driven analytics, MES systems, and decision support platforms.**

The Challenge

Traditional document processing approaches (manual data entry, legacy OCR, or rule-based RPA) fall short in the fab environment because:

- **Complex, dynamic formats:** Multi-column engineering drawings, nested SPC tables, wafer maps with defect annotations, and process flow diagrams lose critical meaning when treated as plain text.

- **Volume & variability:** Tens of thousands of documents generated per week across multiple process areas, each with slight variations in format, structure, and equipment-specific layouts.
- **Context loss:** Legacy tools extract raw text without preserving the relationships among process parameters, equipment states, yield impacts, and corrective actions.
- **Error-prone & time-consuming:** Manual transcription and validation increase cycle times, delay excursion response, and introduce errors that impact yield analysis.
- **Compliance & security risks:** Cloud OCR services often violate data residency, ITAR, export control, and intellectual property protection requirements in sensitive manufacturing environments.
- **Knowledge retention challenges:** Critical process knowledge trapped in documents becomes inaccessible when experienced engineers leave or during shift transitions.

Objectives

- Transform all plant-generated unstructured and semi-structured documentation into **structured, machine-interpretable knowledge** compatible with MES, FDC, and YMS systems.
- Enable fab data to feed **predictive maintenance, yield optimization, excursion detection, RAG-powered reasoning agents, and digital twin systems.**
- Preserve the **hierarchy, traceability, and process context** of each extracted data element while maintaining lot genealogy and equipment history.
- Comply with stringent fab security, privacy, ITAR, export control, and audit standards while scaling seamlessly to meet 24/7 operational volumes.
- **Accelerate cycle time** for root cause analysis, excursion response, and process optimization decisions.

Benefits

- **Operational Intelligence:** Makes critical process knowledge easily searchable, analyzable, and actionable across all shifts and engineering teams.
- **Faster Excursion Response:** Structured insights from equipment logs, SPC reports, and wafer maps enable rapid root cause identification and corrective actions.
- **Proactive Maintenance & Yield:** Historical equipment data and process trends improve preventive maintenance scheduling and reduce unplanned downtime.
- **Compliance & Audit Readiness:** Secure, structured archiving of critical records for internal audits, customer qualifications, and regulatory inspections.
- **Knowledge Preservation:** Captures and digitizes tribal knowledge from experienced engineers, ensuring continuity during workforce transitions.
- **Efficiency & Cost Reduction:** Reduces manual work, eliminates transcription errors, and accelerates time-to-insight for yield and quality investigations.
- **Foundation for AI Transformation:** Enables advanced AI workflows, predictive analytics, and comprehensive fab digitalization initiatives.

Solution Approach

StructiDoc AI applies a **vision-aware, multimodal AI platform** specifically tuned for semiconductor fab operations:

- **Layout-Aware Parsing:** Accurately extracts information from complex formats like cleanroom layouts, SPC charts, wafer maps, process flow diagrams, and annotated defect patterns.
- **Contextual Reasoning:** Uses Vision-Language Models (VLMs) to infer relationships between process parameters, equipment conditions, yield impacts, and quality metrics (e.g., recipe steps $\rightarrow$ expected SPC limits $\rightarrow$ defect signatures $\rightarrow$ corrective actions).
- **Structured Outputs:** Converts raw documents into JSON/XML/graph representations optimized for integration into fab MES, SECS/GEM protocols, FDC systems, and analytics platforms.
- **Semiconductor-Specific Intelligence:** Pre-trained on fab terminology, process flows, equipment types, and industry-standard formats to minimize customization requirements.
- **Secure Deployment:** Offers cloud-based or on-premises options, with HIPAA, SOC 2, ITAR, export control, and data sovereignty compliance for sensitive IP protection.
- **Flexible Integration:** REST APIs and SDKs support seamless deployment into existing fab IT/OT ecosystems, including MES, YMS, and LIMS platforms.

By unlocking the full breadth of fab-generated unstructured data — from **process operations manuals and wafer travelers to SPC charts and equipment maintenance logs** — StructiDoc AI provides the foundation for **intelligent, compliant, and AI-optimized fab operations** that accelerate yield improvements and reduce cycle times.