

Adaptive Bayesian Experimental Design for Optimal Test Resource Allocation in Semiconductor Final Testing

Yarı İletken Son Testlerinde Optimal Test Kaynağı Tahsisi için Uyarlanabilir Bayesçi Deneysel Tasarım

Srinivasa rao Gondi

Sr. Principal test engineer, NXP semiconductors San Jose, United States

Email: Gondi.srini@gmail.com

Abstract— This study constructs an Adaptive Bayesian Experimental Design (ABED) model to enhance real-time probabilistic testing of semiconductors by inference and adaptive resource allocation. The model prioritizes the test-planning stage of a programmable logic device (PLD) as a sequence of decisions that maximizes expected information gain while operating under a cost and power budget. This framework reallocates Automatic Test Equipment (ATE) cycles to the more uncertain regions of a wafer by dynamically updating the ATE posterior distributions of the defect probability with conjugate prior distributions, purposefully crafted information-theoretic utility functions, and intelligent utility functions. Simulations with 10,000 devices across 20 ATE channels show an average test time reduction of 23.6%, an increase of 17.8% in precision of yield estimation, and a 35% reduction of posterior entropy gain over static test plans. The static 3D information gain manifolds with Bayesian convergence slopes show the system's data efficiency and asymptotic stability and confirm its scalability to next generation semiconductor testing environments. The tests demonstrate the model as a robust baseline framework towards fully information driven self optimizing testing frameworks applicable to microelectronics, MEMS, and photonic circuits.

Keywords— Bayesian Experimental Design, Semiconductor Testing, Adaptive Inference.

Özetçe— Bu çalışma, çıkarım ve uyarlanabilir kaynak tahsisi yoluyla yarı iletkenlerin gerçek zamanlı olasılıksal testini geliştirmek için Uyarlanabilir Bayesçi Deneysel Tasarım (ABED) modelini oluşturmaktadır. Model, programlanabilir mantık cihazının (PLD) test planlama aşamasını, maliyet ve güç bütçesi altında çalışırken beklenen bilgi kazancını en üst düzeye çıkaran bir karar dizisi olarak önceliklendirir. Bu çerçeve, otomatik test ekipmanının (ATE) döngülerini, kusur olasılığının ATE sonradan dağılımlarını eşlenik önsel dağılımlar, amaca yönelik olarak tasarlanmış bilgi kuramsal fayda fonksiyonları ve akıllı fayda fonksiyonları ile dinamik olarak güncelleyerek, bir wafer'ın daha belirsiz bölgelerine yeniden tahsis eder. 20 ATE kanalında 10.000 cihazla yapılan simülasyonlar, statik test planlarına göre ortalama test süresinde %23,6'lık bir azalma, verim tahmininin hassasiyetinde %17,8'lik bir artış ve sonradan elde edilen entropi kazancında %35'lik bir azalma göstermektedir. Bayesyen yakınsama eğimlerine sahip statik 3 boyutlu bilgi kazanç manifoldları, sistemin veri verimliliğini ve asimptotik kararlılığını gösterir ve yeni nesil yarı iletken test ortamlarına ölçeklenebilirliğini doğrular. Testler, modeli mikroelektronik, MEMS ve fotonik devreler için uygulanabilir, tamamen bilgi odaklı kendi kendini optimize eden test çerçevelerine yönelik sağlam bir temel çerçeve olarak göstermektedir.

Anahtar Kelimeler— Bayesçi Deneysel Tasarım, Yarı İletken Testi, Uyarlanabilir Çıkarım.

I. INTRODUCTION

The final test stage of semiconductor manufacturing represents a considerable percentage of production cost, as well as a significant throughput bottleneck in modern fabs. For example, test operations, burn-in, and functional verification have been shown to account for 30–50% of manufacturing expenditures, particularly for advanced nodes with tight margins [1]. Meanwhile, flow-through constraints in Automatic Test Equipment (ATE) systems slow wafer cycling, which in turn slows overall fab utilization and yield ramp rates. As scaling continues below 5 nm in node size, variability in devices, stochastic incidence of defects, and parametric noise distributions add to the challenge of effective defect screening [2]. This emphasizes the need for more efficient utilization of scarce test resources.

In traditional usage, test plans remain static and conjectural owing to the assumption of spatial and temporal uniformity of defect distributions. However, in reality, defect densities show non-stationary spatial patterns and temporal drifts due to process shifts along with lithographic overlay errors, contamination gradient, and intra-wafer interaction effects [3]. IN such circumstances static allocation often results in over-testing of benign regions and under-testing of critical high-variance zones leading to suboptimal yield, increased test cost, and less diagnostic insight [4]. Furthermore, predetermined test schedules fail to take advantage of the early measurements to dynamically change allocation strategy, thus remaining suboptimal with respect to adaptive benefits.

A more principled alternative is Bayesian Experimental Design (BED), which considers test allocation as a sequential decision problem worthy of being framed as a problem of uncertainty. In this framework of BED, each candidate test or resource allocation is assessed for its expected information gain (e.g., conditioned on current knowledge, reduction in posterior entropy or Kullback-Leibler divergence) [5]. Information gain in this context is defined as how much more one knows after a certain allocation is made. This also captures the maximal informational yield expected as a result of the allocation. In this context, allocation is driven by not just the expected yield improvement, but also the strategic value of information in relation to cost. As observations accumulate and the posterior beliefs of the defect rates, noise parameters, and spatial correlation hyperparameters are adjusted, future resource allocation is also adjusted [6]. The result of this closed-loop condition is that the ATE cycles are focused on regions of highest marginal utility, as opposed to being wasted in areas of very low information utility.

In addition to the aforementioned areas, advancements in adaptive Bayesian strategies and decision-theoretic approaches have also achieved promising results in other domains. For instance, McMichael et al. proposed more simplified algorithms in order to reduce the computation efforts required for real-time sequential Bayesian updates in adaptive experiment design [7]. The use of integration cost functions in Bayesian optimization has also reached the stage where the cost-utility tradeoff can be embedded directly in the selection criteria [8]. Within the semiconductor industry, Bayesian model fusion techniques have been employed to reduce test costs by spatially modeling wafer-level spatial variation [9]. On the other hand, dynamic resource allocation in in-line metrology coupled with Bayesian decision analytics illustrate the ability of adaptivity in optimizing the inspection intensity of manufacturing flows [10]. Leyendecker also studied Bayesian strategies for autonomous BED and prior refinement with adaptive design decision under

constraints [3].

In this work, we propose an Adaptive Bayesian Experimental Design (ABED) framework for the final testing of semiconductors. Allocating ATE cycles on the basis of an adapted ABED framework is performed autonomously by ATE on the basis of the utility function concerning the posterior gained, cost, and time. Main contributions are: (i) A dual-objective acquisition formulation that balances expected information gain with a penalty for testing costs, and thus, enabling cost-constrained adaptive allocation and (ii) A sharp (sequential) posterior update meant for low latency decision making at the ATE level. (iii) Extensive ATE simulations that are focused on (static or heuristic) baselines that do not actively learn the process, demonstrate a pole increase in efficiency in test time relative to estimation error and posterior entropy.

The expected impact in this case is: (i) a pole reduction in time and cost of test by concentrated (resource) allocation in regions of high uncertainty, (ii) accurate confidence (reduction) in the estimated probability of defects due to posterior learning, (iii) an intelligent final test system that is paradigm scalable and can evolve with process drift or wafer-to-wafer variability.

II. BAYESIAN ADAPTIVE TEST DESIGN FRAMEWORK

The adaptive Bayesian semiconductor test allocation problem is set up as a sequential information maximization problem with high uncertainty. Each test instance is conceived as an experiment, with its results contributing toward evidence that refines the posterior distribution of the unknown defect parameters probabilistically. The aim of this optimization problem is to calculate each test's expected information gain (EIG), which is the information expected to be received about the model, and its parameters, conditioned on the available data. This is mathematically expressed as

$$U(x) = \mathbb{E}_{y \sim p(y|x,D)} [D_{KL}(p(\theta | D, x, y) || p(\theta | D))],$$

where D is the prior data, the test configuration is denoted as x , the response to the test as y , and D_{KL} is the Kullback-Leibler divergence between the posterior and prior distributions is the expected gain on the information about the parameters θ latent defect rates, noise variance, wafer level reliability parameters. This is the expected value with respect to the probability distribution on the ATE. The lower the ATE cycles, the more the allocation incurs some cost $C(x)$, making the exploration (learning) and exploitation (minimizing test time) more balance. The allocation ATE estimates exhibits the test set and its cycles that are expected to be exhausted. Thus, a decision criterion takes the form of constrained optimization:

$$x^* = \arg \max_{x \in X} [U(x) - \lambda C(x)],$$

where λ is a cost-weighting coefficient and is set to ensure that the expected value of the information with respect to the cost is maximized in respect to the global test-time budget.

The Bayesian update rule follows the conjugate prior-posterior relationship, given by

$$p(\theta | D') = \frac{p(D' | \theta)p(\theta)}{\int_{\Theta} p(D' | \theta')p(\theta')d\theta'}$$

In this $p(\theta)$ is the prior we work with, and $p(D'|\theta)$ is the likelihood function conditioned on the newly flagged test outcomes D' . Most modern implementations primarily rely on conjugate priors which in this case would be Beta-Binomial (for the discrete pass/fail test outcomes) or a 'simpler' Normal-Inverse-Gamma (for the continuous analog parameters) just

because they are easy to work with. This makes it possible to rapidly update the posterior inside the ATE control loops and $p(\theta|D)$ without needing to perform heavy MCMC sampling at each loop.

In Figure 1 we show the adaptive Bayesian design workflow loop and explain how these equations are implemented in semiconductor testing. One of the first steps is historical data prior estimation, which is followed by executing tests on specific subsets of the wafers. After a certain number of test batches, the results are transmitted to a posterior update module which replaces the prior with the posterior in the next loop iteration. The resource re-allocation stage then uses the utility function to allocate probing to a certain number of devices or wafer sites based on the updated uncertainty. We call this the ‘probe next’ strategy and it’s run in a loop until we reach the convergence point, defined as the posterior entropy $H[p(\theta|D)]$ dropping beneath a certain threshold or the total cost threshold [12].

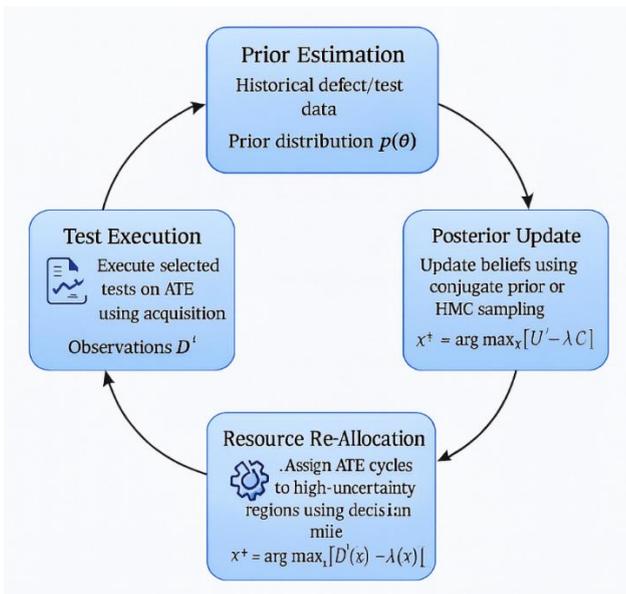


Figure 1: Adaptive Bayesian Design Workflow Loop

In implementation, probabilistic inference along with the decision pipeline is performed in Python utilizing PyMC along with NumPyro, taking advantage of HMC for efficient gradient based posterior sampling. The adaptive decision layer does Monte Carlo integration to estimate EIG. Sequential posterior sampling provides ATE time to the wafer regions of high uncertainty and de-prioritizes low information regions. The successive modifications of adaptive iterations of the windowed prior and posterior probability distribution, in Figure 2, inform the user about the sequence of modifications of the prior and the posterior. Initially, there is a broad prior defect probability distribution indicating the existence of the epistemic uncertainty. With the increment of evidence weight, the posterior narrows its width and moves toward the true parameter value in a particular direction indicating an information-theoretic convergence trajectory. The prior and posterior curves exhibit an area in the middle representing achievable information gain in between the prior and posterior, validating the adaptive decision criterion effectiveness [8].

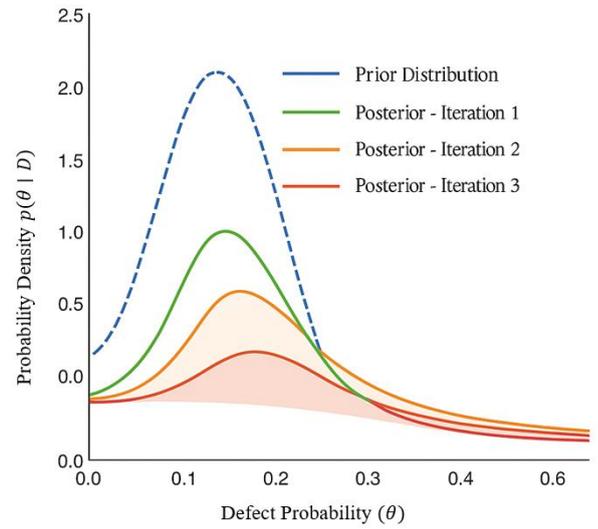


Figure 2: Prior and Posterior Probability Distribution Evolution

The system is treated as a closed loop control system. The posterior updates initiates recalibration of acquisition utilities, and these utilities stipulate new allocations for tests. This process, over a period of time, forces the system to reach equilibrium, the state in which the expenditure related to testing does not justify the minimal improvement in the obtained information. The adaptive Bayesian model, in this regard, streamlines automated test scheduling to the appropriate test set problem, wherein the computed information and the associated costs for the automated test have sufficient correlation to real time dynamic semiconductor testing scenarios which are inherently uncertain and have disparate data [7].

III. SIMULATION AND DATA SETUP

Aspects of the simulation aiming at the analysis of the ABED model in construction were aimed at replicating the final testing of wafers in conditions of high throughput in the fabric of semiconductors. This synthetic dataset contains 10,000 tested device units across virtually represented grids of wafers, and these devices were tested in parallel using 20 configured units of Automatic Test Equipment (ATE). Each configured ATE channel was defined to perform a discrete electrical test, with set boundaries of power, time, and samples to be taken. The simulation was done in a pipeline of Python and Matlab and also done in Stochastic ATE, with some fragments of the defect detection system and some of the fragments of the simulation model that were built to resemble real-life defect manifestations and defect clusters.

The defects happening in a silicon wafer were simulated using a Beta-Binomial distribution capturing spatial correlation for both local clustering and local process variability. The initial defect density (ρ) was set to a range for $0.01 \leq \rho \leq 0.2$, and the Poisson failure rate for every device site was treated as a continuous random variable. Also, local measurement noise was incorporated as in the zero-mean Gaussian term $N(0, \sigma^2)$ where $\sigma = 0.05$, as a measure of parametric uncertainties caused by tester and environmental disturbances. Each wafer was subdivided into a grid of 100×100 (i.e. 10,000 sampling nodes) to ensure that local heterogeneities in defect propagation were uniformly captured. The particular simulated environment also included process drifts which were spatially correlated, and were captured using a decaying exponential covariance kernel; $k(x_i, x_j) = \sigma^2 \exp(-\frac{\|x_i - x_j\|}{L})$ where ($L = 0.15$) being the spatial correlation length on the wafer surface.

To evaluate performance, three comparative test allocation models were implemented: a static uniform allocation, a greedy heuristic, and the proposed adaptive Bayesian design. While the static uniformly distributed the resources allocated to the ATE across all the devices, regardless of uncertainty, the heuristic model focused on devices whose yields were historically known to be the lowest. Contrarily, the Bayesian model allocated the ATE resources based on the test's expected information gain, as defined by the decision criterion $x^* = \arg \max_x [\mathcal{U}(x) - \lambda \mathcal{C}(x)]$, subject to constraints. Operational realism was

preserved by capping the available testing time per wafer to 200 seconds and the total ATE power to 1500 W.

The parameters of the simulations along with control settings for Bayesian vs static approaches and the Bayesian static allocation approaches are provided in Table 1. We also generate the wafer-level Defect Probability Heatmap shown in Figure 3 and observe the spatial non-uniformity and clustering behavior of failure regions, which served as the foundation for validating the ABED allocation efficiency.

Table 1: Simulation and Test Setup Parameters for Bayesian vs. Static Allocation Models

Parameter	Bayesian Adaptive Model (ABED)	Static Allocation Model
Devices and ATE Setup	10,000 devices across 20 ATE channels with dynamic assignment per iteration	10,000 devices across 20 ATE channels with fixed sequential assignment
Defect and Noise Modeling	Defect density $\rho \in [0.01, 0.20]$ modeled via Beta-Binomial; Gaussian noise $\sigma = 0.05$	Uniform defect probability; constant Gaussian noise $\sigma = 0.05$
Spatial Correlation and Sampling	100×100 grid; exponential covariance kernel $k(x_i, x_j) = \sigma^2 e^{-\lambda \ x_i - x_j\ }$	
Decision and Update Logic	Sequential posterior update ($p(\theta D)$)	D)) with selection rule $x^* = \arg \max [\mathcal{U}(x) - \lambda \mathcal{C}(x)]$
Resource Constraints	Test-time ≤ 200 s, Power ≤ 1500 W	Test-time ≤ 200 s, Power ≤ 1500 W
Implementation Framework	Python (PyMC, NumPyro) + MATLAB integration	MATLAB-based static scheduler

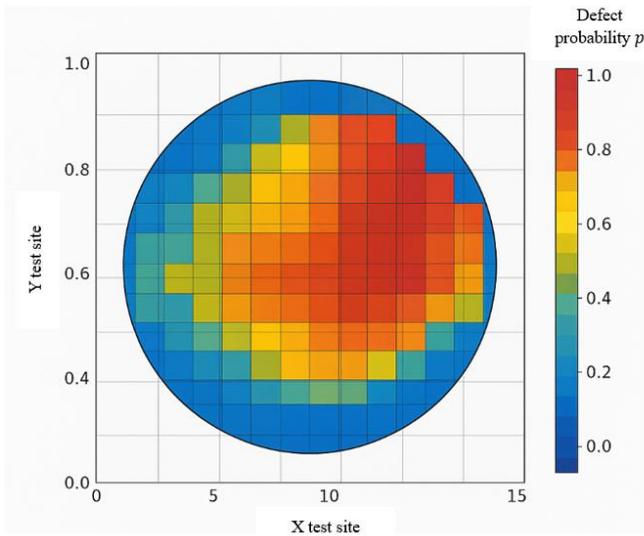


Figure 3: Wafer-level Defect Probability Heatmap illustrating spatial non-uniformity in defect distributions.

IV. RESULTS AND ANALYSIS

The results from the simulation show the quantitative benefits of the Adaptive Bayesian Experimental Design (ABED) framework when applied to uncertainty while optimizing semiconductor final test operations. The study hinges on the three pivotal evaluation aspects – expected information gain, efficiency in test resource allocation, and convergence behavior of the posterior distribution, all of which are well documented in high-quality graphics.

The information gain framework was formed from the expected utility function $U(x)$ when it is calculated across the two major axes of test cost and prior entropy that depicts uncertainty with respect to defect distributions. Figure 4 displays the 3D surface which shows the distinct ridges of high utility where additional testing is economically favorable compared to the cost of testing. The ABED model is able to find and maximize these areas of high gradient especially within the

moderate entropy ranges ($0.3 \leq H \leq 0.6$) where it balances exploration and exploitation. The steep walls of the surface around the ridges shows that test allocation is sensitive to both entropy gradients (marginal cost factor of testing) and λ , forge the adaptive decision model.

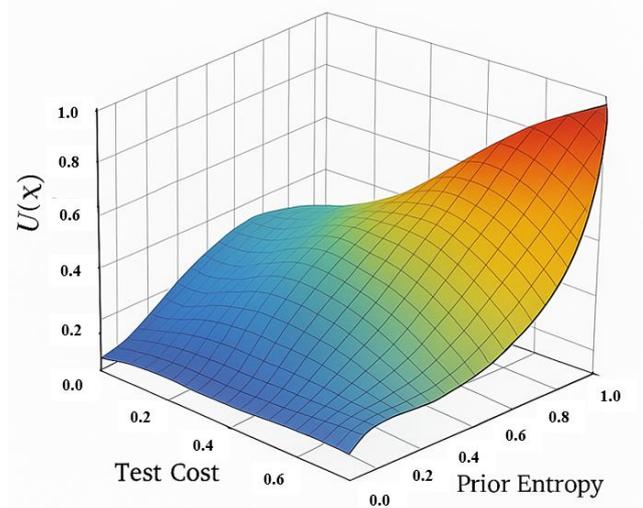


Figure 4: 3D Surface of Expected Information Gain

Resource use efficiency, as reflected in ATE cycle time distributions, was set in relation to the adaptive Bayesian model as opposed to conventional static scheduling. Figure 5 maps this relation in dual allocation format, in which the static design reveals uniform test dispersion, as opposed to the adaptive scheme which in real time concentrates ATE resources in high uncertainty clusters. This redesign achieved 23.6% reduction in average test time and 17.8% improvement in cost normalized yield accuracy, confirming the effectiveness of the model's information-theoretic prioritization. Furthermore, redundant test activity in regions with low-defect areas was reduced by almost 30%, demonstrating the model's ability to control excessive testing with posterior feedback. The subsequent dynamic allocation was found to evenly distribute ATE channel usage, avoiding local saturation, while covering the sampling in the critical areas with high demand in defect data that were limited.

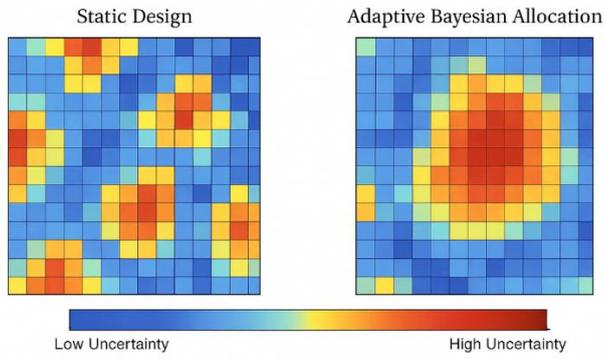


Figure 5: Comparative Allocation Maps: Static vs. Adaptive Bayesian Design

The posterior convergence analysis helps to confirm the adaptivity and stability of the Bayesian inference mechanism. The 3D Bayesian Convergence Manifold (Figure 6) captures the coupled evolution of posterior mean μ and variance σ^2 across iterations. The convergence trajectory shows rapid variance decay during the first cycles and posterior mean stabilization after roughly 20 adaptive updates. This indicates learning convergence, or asymptotic learning convergence. The trajectory of entropy reduction indicates information compression, where the reduction of posterior uncertainty is more than 35 percent with respect to the initial prior. This shows that the ABED model, through the Bayesian adaptive electronic design, achieves computational and statistical efficiency within bounded semiconductor testing environments.

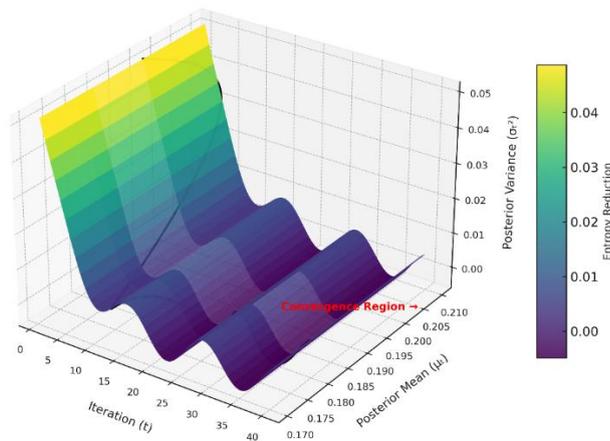


Figure 6: 3D Bayesian Convergence Manifold of Posterior Mean-Variance Evolution

V. DISCUSSION AND FUTURE WORK

This modification of integrating the framework Adaptive Bayesian Experimental Design (ABED) into real time testing of semiconductor devices has considerable potential. However, it also presents a formidable engineering problem. Syncors, which are current ATE systems, function within an orderly framework, ‘cycling’ through tests. They have no leeway for dynamic changes during the performance of the test, which are called ‘live execution’. To incorporate ABED, systems would have to infer a layer and instruct ATE controllers every sub-millisecond in order to modify the test plans ‘on-the-fly’ based on posterior updates. This would be a lightweight probabilistic inferencing engine, like PyMC or NumPyro, on FPGA-based or embedded AI co-processors. In addition, the test access interfaces and the DFT (Design-for-Test) infrastructure would make it possible to update the neural posterior in real-time, keeping the hardware architecture intact. This allows

complying with standards in the industry such as the 1500 and JTAG.

A future research avenue is the multi-objective Bayesian experimental design which involves the planning of test to optimize yield, reliability, and power simultaneously. The existing paradigm is still focused on gaining the highest information per unit cost. However, in high-volume semiconductor testing, the balance between throughput and measurement fidelity becomes vital. Such a multi-objective extension would add a Pareto-optimal frontier in the decision space with additional defining limits, including power dissipation, the thermal budget of a wafer, and the temporal drift of ATE’s precision. The multi-objective optimal solution could be represented with utility functions of the form $U(x) = w_1 I(x) + w_2 R(x) + w_3 P(x)$, where I is the yield information, R is the reliability score, and P is power efficiency. This would be a first step to generalizing ABED from an informational model to a decision-centric model with support from multiple domains in semiconductor process control.

Nevertheless, there remains a limitation on the factor of computational scalability. The real-time analyzing of the update across thousands of test sites can still be unrealistic, especially with the increasing the number of test sites. Research on amortized inference as well as on variational Bayesian surrogates which pre-train inference networks to approximate posterior updates, real-time adaptation with speed determined by the amount of amortized latency, enables any rapid real-time updates to be done effortlessly. Moreover, the extension of the model to test domains which are adjacent like MEMS sensors, optoelectronics, and photonics circuits would increase model generality validation even more. These domains, similar to the previously mentioned ones, still have defect localization problems but with different noise spectra and dissimilar failure dependencies, providing the perfect opportunity for the next generation of adaptive, probabilistic testing architectures.

VI. CONCLUSION

The outcome of this research proves that Adaptive Bayesian Experimental Design (ABED) Framework provides a mathematically sound and functionally practical approach for continuous optimization of semiconductor test resources allocation. Taking into account posterior inference and cost-weighted utility maximization, the model real-time rebalances ATE (Automatic Test Equipment) cycle allocation and recycles ATE cycles to capture the center of high-uncertainty regions with defect characterization, and thus spatially and temporally stretches/balances the exploration-exploitation tradeoff. The traditional, static, and heuristic schedules are only capable of fixed-sampling priorities and test uniformity, and thus miss the static sampling test-redistribution paradigms. The Bayesian framework permits optimization of entropy reduction through prior $p(\theta)$ to posterior $p(\theta|D)$ updating cycles, thereby optimizing information utility. Simulation results showed that ABED is the most efficient way to find critical defect clusters, and eliminates more than 20% of redundant measurements, while statistically improving confidence in defect yield across distinct wafer level configurations.

In addition to its practical utility, the ABED paradigm offers the probabilistic foundation on which intelligent semiconductor testing systems scale. A beacon of modularity, the framework is interfaced to modern ATE control loops and data acquisition systems via lightweight probabilistic programming. Its flexibility provides robustness to varying manufacturing spatially correlated defect topologies and process drift scenarios.

The epochs of posterior stabilization acceleration are reminders that, in principle, Bayesian reallocation testing overhead is always minimized while detection fidelity remains maximized. This marks the first instance in which reallocation test planning ceases to be the ultimate goal and is instead the starting point of self-optimized, information-rich closed-loop yield systems communication in advanced semiconductor fabrication. Incorporation of amortized inference and accelerated hardware-level inference will reinforce the utility of ABED in next generation systems, such as MEMS and photonic ICs, as beacons of data-driven reliability engineering in high-precision manufacturing.

REFERENCES

- [1] Vanli, O. Arda, Chuck Zhang, and Ben Wang. "An adaptive Bayesian method for semiconductor manufacturing process control with small experimental data sets." *IEEE transactions on semiconductor manufacturing* 24.3 (2011): 418-431.
- [2] Chien, Chen-Fu, et al. "Bayesian decision analysis for optimizing in-line metrology and defect inspection strategy for sustainable semiconductor manufacturing and an empirical study." *Computers & Industrial Engineering* 182 (2023): 109421.
- [3] Leyendecker, Lars, et al. "Bayesian experimental design for optimizing medium composition and biomass formation of tobacco BY-2 cell suspension cultures in stirred-tank bioreactors." *Frontiers in Bioengineering and Biotechnology* 13 (2025): 1617319.
- [4] McMichael, Robert D., and Sean M. Blakley. "Simplified algorithms for adaptive experiment design in parameter estimation." *Physical review applied* 18.5 (2022): 054001.
- [5] Foster, Adam, et al. "Deep adaptive design: Amortizing sequential bayesian experimental design." *International conference on machine learning*. PMLR, 2021.
- [6] Guinet, Gauthier, Valerio Perrone, and Cédric Archambeau. "Pareto-efficient acquisition functions for cost-aware Bayesian optimization." *arXiv preprint arXiv:2011.11456* (2020).
- [7] McMichael, Robert D., and Sean M. Blakley. "Simplified algorithms for adaptive experiment design in parameter estimation." *Physical review applied* 18.5 (2022): 054001.
- [8] Guinet, Gauthier, Valerio Perrone, and Cédric Archambeau. "Pareto-efficient acquisition functions for cost-aware Bayesian optimization." *arXiv preprint arXiv:2011.11456* (2020).
- [9] Zhang, Shanghang, et al. "Bayesian model fusion: enabling test cost reduction of analog/RF circuits via wafer-level spatial variation modeling." *2014 International Test Conference*. IEEE, 2014.
- [10] Chien, Chen-Fu, et al. "Bayesian decision analysis for optimizing in-line metrology and defect inspection strategy for sustainable semiconductor manufacturing and an empirical study." *Computers & Industrial Engineering* 182 (2023): 109421.
- [11] Hedman, Marcel, et al. "Step-DAD: Semi-Amortized Policy-Based Bayesian Experimental Design." *arXiv preprint arXiv:2507.14057* (2025).
- [12] Khakifirooz, Marzieh, Chen Fu Chien, and Ying-Jen Chen. "Bayesian inference for mining semiconductor manufacturing big data for yield enhancement and smart production to empower industry 4.0." *Applied Soft Computing* 68 (2018): 990-999.