Volume 4, Issue 1, February 01, 2019

TC-CPS Newsletter

Technical Articles

- Junlong Zhou, Yue Ma, Jin Sun, Tongquan Wei, *"Resource Management for Improving Reliability of Heterogeneous MPSoC Systems"*
- Yukun Ding, Yiyu Shi, *"Real-Time Boiler Control Optimization with Machine Learning"*.
- Longfei Wang, S. Karen Khatamifard, Ulya R. Karpuzcu, Selçuk Köse, *"Exploring On-Chip Power Delivery Network Induced Analog Covert Channels"*.
- Meng Li, Kaveh Shamsi, Yier Jin, David Z. Pan, *"IP Protection and Supply Chain Security through Logic Obfuscation"*.
- Zhaoyan Shen, Zili Shao, *"Nonvoltile Memory and Storage for CPS"*.
- Yachen Zhang, Long Chen, *"3D Cooperative Mapping for Connected Ground and Aerial-Based Robots"*.

Summary of Activities

Call for Contributions

IEEE TC-CPS Cyber-Physical Systems

Resource Management for Improving Reliability of Heterogeneous MPSoC Systems

Junlong Zhou[†], Yue Ma[‡], Jin Sun[†], and Tongquan Wei[§]
[†]School of CSE, Nanjing University of Science and Technology, Nanjing 210094, China
[‡]Department of CSE, University of Notre Dame, Notre Dame, IN 46556, USA
[§]Department of CST, East China Normal University, Shanghai 200062, China

1 Background

To help meet high performance and low power demands in many applications, heterogeneous multiprocessor systems-on-a-chip (MPSoCs) that integrate I/O components, processing units as well as dedicated hardware and memory on a single silicon die have been introduced [1]. The heterogeneity of an MPSoC system is in the sense that the functionality and computing capability of processing units such as CPU and GPU are distinctively different. In general, heterogeneous MPSoCs can be classified into performance-heterogeneous MPSoC (PH-MPSoC) and function-heterogeneous MPSoC (FH-MPSoC) [2]. In a PH-MPSoC, cores having the same functionality but different power-performance characteristics are interspersed together, whereas in an FH-MPSoC, cores having very different functionality are integrated on the same die. For example, Nvidia's Jetson TX2 board is the so-called PH-MPSoC as it adopts ARM big.LITTLE architecture which integrates high-performance cores with low-power cores.

Heterogeneous MPSoCs are widely deployed in embedded systems. In such systems, soft-error reliability (SER) due to transient faults and lifetime reliability (LTR) due to permanent faults both are imperative design concerns. Transient faults (leading to soft errors) appear for a short time and then disappear without damage to the hardware, and are caused by cosmic radiation or electromagnetic interference (see Fig. 1(a)). Permanent faults (leading to hard errors) continue to exist until the faulty hardware is repaired or replaced, and are caused by circuit wear-out or manufacturing defects (see Fig. 1(b)) .

2 Soft-Error Reliability Model

Soft errors are modeled by the exponential distribution with an average rate r. The fault rate r indicates the expected number of failures occurring per second and it exponentially increases with reduction of core frequency [3]. Suppose $r(f)$ is the core's fault rate running at frequency f. It is expressed as

$$r(f) = r_0 \cdot 10^{\frac{\omega(1-f)}{1-f_{\min}}}, \tag{1}$$

where r_0 is the core's fault rate at the maximal frequency $f_{\max}$, ω is a hardware-dependent constant that indicates the sensitivity of fault rate to frequency scaling, and $f_{\min}$ is the core's minimum frequency. Clearly, reducing frequency leads to exponential increase in fault rate. The SER of a task is defined as the probability of being successfully executed without suffering any transient faults. Using the exponential failure model, the task SER during the execution time et is formulated as

$$SER = e^{-r(f)\cdot et}. \tag{2}$$

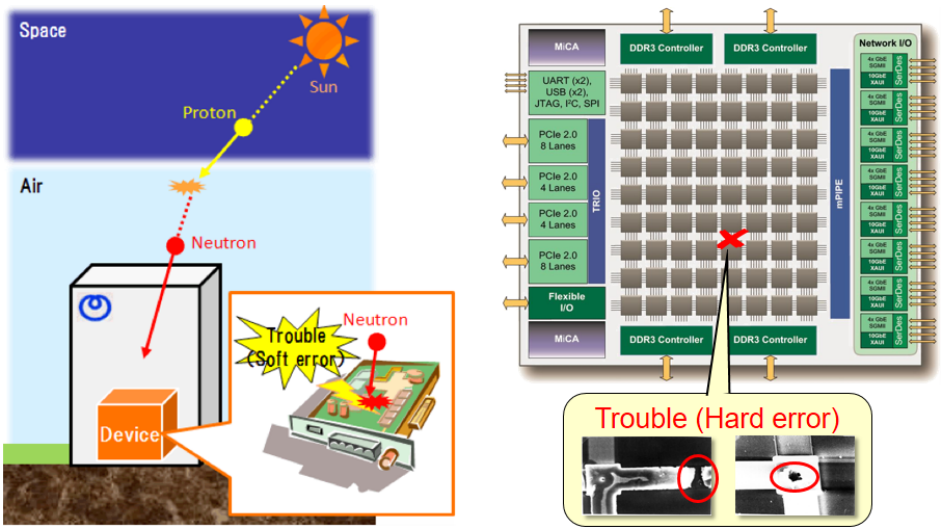


Figure 1: Soft and hard errors.

3 Lifetime Reliability Model

Multiple wear-out effects can lead to permanent faults. We consider wear due to electromigration (EM), stress migration (SM), time-dependent dielectric breakdown (TDDB), and thermal cycling (TC). Wear due to EM, SM, or TDDB exponentially depends on the chip temperature while wear due to TC relies on not only the temperature amplitude but also the cycle maximum temperature [4]. For improving LTR, the chip temperature and the thermal cycling effects both should be mitigated. Mean time to failure (MTTF) is commonly utilized to quantify LTR. The formulas to derive the MTTFs due to EM, SM, and TDDB as well as the number of cycles to failure due to TC are described below.

- $$MTTF_{\mathrm{EM}} = \frac{A_{\mathrm{EM}}}{D^{\lambda}} e^{\frac{E_{\mathrm{act_{EM}}}}{\mathscr{B}T}}, \tag{3}$$

  where A_{EM} denotes a hardware specific constant, D denotes the current density, $E_{\mathrm{act_{EM}}}$ denotes the active energy for EM, λ denotes an empirical constant, $\mathscr{B}$ denotes the Boltzmann constant, and T denotes the runtime temperature [5].

- $$MTTF_{\mathrm{TDDB}} = A_{\mathrm{TDDB}} (\frac{1}{V})^{(\vartheta_1 - \vartheta_2 T)} e^{\frac{\vartheta_3 + \vartheta_4/T + \vartheta_5 T}{\mathscr{B}T}}, \tag{4}$$

  where A_{TDDB} denotes a fitting constant, V denotes the supply voltage, and $\vartheta_1 - \vartheta_5$ denote empirical fitting parameters [6].

- $$MTTF_{\mathrm{SM}} = A_{\mathrm{SM}} |T_0 - T|^{-\lambda} e^{\frac{E_{\mathrm{act_{SM}}}}{\mathscr{B}T}}, \tag{5}$$

  where A_{SM} denotes a fitting constant, T_0 denotes the mental deposition temperature, and $E_{\mathrm{act_{SM}}}$ denotes the activation energy for SM [7].

- $$N_{\mathrm{TC}} = A_{\mathrm{TC}} (\Delta T - T_{\mathrm{th}})^{q} e^{\frac{E_{\mathrm{act_{TC}}}}{\mathscr{B}T_{\max}}}, \tag{6}$$

  where A_{TC} denotes an empirical constant, ΔT denotes the amplitude of thermal cycling, T_{th} denotes the temperature where inelastic deformation begins, q denotes the Coffin-Manson exponent constant, $E_{\mathrm{act_{TC}}}$ denotes the activation energy for TC, and $T_{\max}$ denotes the maximal temperature during the cycle [8].

4 Resource Management for Improving Reliability

Numerous investigations have been made into resource management for improving reliability. Chou et al. [9] developed an efficient fault-aware resource management scheme for network-on-chip (NoC) systems. In this scheme, spare cores on the chip are carefully placed to improve system SER and LTR. Das et al. [10] proposed a genetic heuristic that increases system reliability by determining task-to-core mapping and task operating frequency. Dynamic voltage and frequency scaling aware and Q-learning based optimization techniques [11] are designed for multicore systems in the presence of both permanent and transient faults. Zhou et al. [12] proposed to use MTTF to evaluate SER and LTR, and derived a novel analytical formula for calculating the MTTF of a core due to transient faults. Although these methods are effective in improving SER and LTR, they do not consider the heterogeneity of MPSoCs.

Recently, several efforts are made into increasing SER and LTR for heterogeneous MPSoCs. Yue et al. [13] presented two dynamic recovery based scheduling algorithms for improving system SER under the deadline as well as LTR constraints. The SER is increased by recovering failed tasks and the LTR is ensured by reducing core's frequencies. Considering the effects of hardware- and task-level variations on reliability, Zhou et al. [14] proposed a resource management scheme to maximize SER for real-time applications running on heterogeneous MPSoCs without violating the deadline, peak temperature, and LTR constraints. Especially for the big.LITTLE type heterogeneous MPSoCs, Yue et al. [15] developed an on-line framework for increasing SER while meeting the LTR constraint. Unlike the approaches [13, 14, 15], an evolutionary-based task scheduling algorithm that determines the execution order, allocation, replication and frequency of tasks, is proposed by Zhou et al. [16] to address the problem of maximizing SER and LTR under the requirements of satisfying energy budget, deadline, and task precedence.

References

[1] J. Zhou, T. Wei, M. Chen, J. Yan, X. S. Hu, and Y. Ma, "Thermal-aware task scheduling for energy minimization in heterogeneous real-time MPSoC systems," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 35, no. 8, pp. 1269-1282, 2016.

[2] A. Prakash, S. Wang, A. E. Irimiea, and T. Mitra, "Energy-efficient execution of data-parallel applications on heterogeneous mobile platforms," in *ICCD*, pp. 208-215, 2015.

[3] D. Zhu, R. Melhem, and D. Mosse, "The effects of energy management on reliability in real-time embedded systems," in *ICCAD*, pp. 35-40, 2004.

[4] Y. Xiang, T Chantem, R. P. Dick, X. S. Hu, S. Li. "System-level reliability modeling for MPSoCs," in *CODES+ISSS*, pp. 297-306, 2010.

[5] J. Srinivasan, et al. "The impact of technology scaling on lifetime reliability," in *DSN*, pp.177-186, 2004.

[6] J. Srinivasan, et al. "Exploiting structural duplication for lifetime reliability enhancement," in *ISCA*, pp. 520-531, 2005.

[7] "Failure mechanisms and models for semiconductor devices," Joint Electron Device Engineering Council, Tech. Rep., JEP 122-B, 2003.

[8] M. Ciappa, et al. "Lifetime prediction and design of reliability tests for high-power devices in automotive applications," *IEEE Transactions on Device and Materials Reliability*, vol. 3, no. 4, pp. 191-196, 2003.

[9] C. Chou and R. Marculescu. "FARM fault-aware resource management in NoC," in *DATE*, pp. 1-6, 2011.

[10] A. Das, A. Kumar, B. Veeravalli, C. Bolchini, and A. Miele, "Combined DVFS and mapping exploration for lifetime and soft-error susceptibility improvement in MPSoCs," in *DATE*, pp. 1-6, 2014.

[11] T. Kim, Z. Sun, H. Chen, H. Wang, and S. X. D. Tan, “Energy and lifetime optimizations for dark silicon many core microprocessor considering both hard and soft errors," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 9, pp. 2561-2574, 2017.

[12] J. Zhou, X. S. Hu, Y. Ma, and T. Wei, “Balancing lifetime and soft-error reliability to improve system availability," in *ASPDAC*, pp. 685-690, 2016.

[13] Y. Ma, T. Chantem, R. P. Dick, and X. S. Hu, “Improving reliability for real-time systems through dynamic recovery," in *DATE*, pp. 515-520, 2018.

[14] J. Zhou, T. Wei, M. Chen, X. S. Hu, Y. Ma, G. Zhang, and J. Yan, “Variation-aware task allocation and scheduling for improving reliability of real-time MPSoCs," in *DATE*, pp. 171-176, 2018.

[15] Y. Ma, J. Zhou, T. Chantem, R. P. Dick, S. Wang, and X. S. Hu, “On-line resource management for improving reliability of real-time systems on Big-Little type MPSoCs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, in press, 2019. DOI: 10.1109/TCAD.2018.2883990.

[16] J. Zhou, J. Sun, X. Zhou, T. Wei, M. Chen, S. Hu, and X. S. Hu, “Resource management for improving soft-error and lifetime reliability of real-time MPSoCs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, in press, 2019. DOI: 10.1109/TCAD.2018.2883993.

Real-Time Boiler Control Optimization with Machine Learning

Yukun Ding, Yiyu Shi
University of Notre Dame

1 Introduction

As coal-fired power plants currently produce 41% of global electricity [31], proper control of coal-fired boilers in producing electricity is not only essential to the safety of power plant operation, but also directly affects boilers stability, energy efficiency, and sustainability, thus having huge socioeconomic and environmental impacts [11]. How to optimally control boilers' operating condition in real-time is, however, difficult. The combustion process inside a boiler is highly complex and nonlinear with strong-coupling and time-delayed influences. It is well understood in literature that it is not easy to achieve high efficiency in operating large utility boilers and most existing practices in the industry are highly sub-optimal [5].

Nonuniform temperature distribution inside a boiler is known to cause tube rupture, a frequent failure mechanism for boiler operations. But to maintain a uniform temperature distribution inside a boiler is difficult even for domain expert in practice due to the dynamic air flow inside the boiler. One of the most frequently used practices to deal with nonuniform temperature distribution is spraying water inside a boiler, which introduces unnecessary efficiency loss and additional operating cost [11, 22]. Another practice is to remold a boiler by re-arranging super-heater panels to alleviate the uneven temperature distribution [37], which requires to shut down the boiler and cannot be done in real-time. Temperature distribution inside a boiler has been studied using various computational fluid dynamics (CFD) methods under steady-state conditions [11, 22]. However, employing the CFD methods for real-time boiler control is not feasible because of the extremely high computational cost of solving the CFD models [2].

To avoid solving the physical-based CFD models, researchers have proposed to use machine learning-based methods to predict boiler temperature or other related parameters in order to control boiler combustion process [17, 16, 42]. However, such formulations have mainly focused on reduction of pollutant emission, not for the uniform distribution of temperature inside boilers. There are some works focusing on the boiler efficiency optimization [9, 15] where external measurements (e.g. exhaust gas temperature) are used to estimate the boiler efficiency through some models due to the difficulties in obtaining temperatures within the boiler. Instead, in this work we have collected the temperature distribution data within the boiler from our industry partner, which allows precise and accurate observation of the combustion efficiency and stability. Moreover, due to the strong-coupling, nonlinear and large time delay characteristics of boilers, existing solutions using neural networks often result in a complicated black-box optimization problem and thus can hardly ensure good real-time performance [16, 32, 42, 5].

In this paper, we use a new formulation to the boiler control optimization problem based on inputs from industry, i.e., maintaining a uniform distribution of temperature in different zones and a balanced oxygen (O_2) content from the flue in a coal-fired power plant. We develop a new but practical solution framework to solve the proposed real-time control optimization problem by combining machine learning and optimization techniques. We validate the formulation and show high solution quality using a real industry boiler dataset. Our results suggest that, in specific scenarios, a dedicated system with simpler models can be more desirable than using more powerful models in terms of both performance and computational efficiency.

The rest of the paper is organized as follows. We first give a background about the boiler control problem. Then we present our formulation of the problem and the solution. We then report the experimental results and make conclusions.

2 Background

We give a brief introduction of the operation of a power plant boiler. Pulverized coal is fed into the furnace from different coal feeders with a proper volume of airflow, both of which are controlled by their respective throttle openings to maintain a desired air-to-fuel ratio for combustion. The water circulates in a water-steam system and

absorbs the radiation energy from the furnace continuously until it becomes high-pressure superheated steam in the superheater. Through the steam turbines, the thermal energy is transferred to mechanical work and finally becomes electricity through generators. In a power plant, a central controller determines the desired setpoints for various subsystem controllers, a critical one of which is the combustion controller that determines the feed rates of coal and airflows [19].

Since combustion quality ultimately determines the production efficiency, we focus on combustion control in this paper. In general, higher temperature inside the furnace and lower O_2 content in the flue indicate higher efficiency. But to ensure sustainably high combustion efficiency and stability, it is desirable to also maintain a balanced high temperature distribution and low O_2 content in the flue, as a balanced distribution of both temperature and O_2 content indicates that both flames and pressures are uniformly distributed, and thus promising the stability and safety of the boiler. However, existing formulations [16, 32, 14, 42] have not considered the temperature distribution, and not mentioned the distribution of O_2 content.

In current industry practice, the combustion controller consists of a set of PI/PID controllers and pre-computed set-pints (computed theoretically and fine-tuned empirically). The well-established control system based on PI/PID was improved by advanced PI/PID controller such as auto-tuning PID [39]. In recent years, machine learning-based prediction control has been studied widely and included in commercial boiler control solutions [13, 15]. The prediction model was used for steady state optimization initially and then gradually for real-time optimization [23, 27, 19, 6]. From the algorithm perspective, the modeling approaches are dominated by neural networks and their variants, e.g. vanilla feed-forward network, radial basis function (RBF) network, double linear fast learning network [17, 16, 5]. The optimization problems are solved by various heuristic search algorithms, e.g. genetic algorithm (GA), differential evolution (DE), particle swarm optimization (PSO), ant colony optimization (ACO) [44, 24, 41]. Even though the recent advancement on more computationally efficient neural network [3, 33, 35, 18, 25, 34, 10, 36], due to the considerable number of variables, the computational requirement remains a challenge which degrades the real-time performance [43].

3 Problem Formulation and Solution Framework

We formulate a new boiler control problem in this section by not only maintaining a high temperature and low O_2 content, but also maintaining a uniform distribution of temperature in different zones and O_2 content inside the flue in a coal-fired power plant. The goal of achieving a balanced distribution of temperatures and O_2 content can be captured by a quadratic penalty function of the deviation of temperatures from the average value and the difference between O_2 content from two sides in the flue. Certainly there are other options but quadratic penalty function is employed, because it is differentiable, suitable for capturing the deviation from the desired value, and relatively simple for optimization. For effective combustion, we also want to maintain a high temperature and low O_2 content inside a boiler. Together, we can use a weighted sum of these components as our objective with the constraints under real operation such as the given range of controllable variables and their sum. The polynomial objective function has four terms, indicating the variance of temperature in different zones, the difference of O_2 content in two sides of flue, the average temperature, and the average O_2 content, respectively. The problem needs to be solved continuously for every time stamp t based on data including operations from $t-1$ and prior in order to achieve the goal of real-time control for the boiler. f_i^T and f_j^O define prediction models for temperature and O_2 respectively where i and j denote the index of models for temperature and O_2 content at different zones.

The structure of proposed real-time boiler control framework is shown in Figure 1. The prediction models, f_i^T and f_j^O trained on historical data, provide the symbolic expression of temperature and O_2 content based on control variables and other measured uncontrollable variables, which are denoted as x_{t-1} and M_{t-1} respectively. In every time step, M_{t-1} in the symbolic expression will be replaced by the latest observed values and only the controllable variables x_{t-1} and the optimization objective V_t remain. Then the optimization model takes the resulted expression and solves the nonlinear programming problem to give the optimal combination of controllable variables, which is the control input to the boiler. An error compensation module, which will be discussed later, is employed to further improve the prediction accuracy. The time cost for solving the optimization model at every time step depends on the choice of optimization algorithm and the complexity of the problem determined by the prediction model. Since the

control loop needs to be continuously solved for every time step as soon as possible, the runtime performance of the optimization model is the critical consideration.

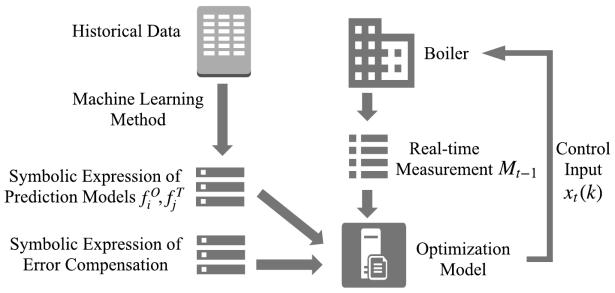


Figure 1: Structure of the solution framework

We employ machine learning-based approaches for predicting both temperature and O_2 content. We notice that there is a special mathematical structure of the given problem that the constraints are linear with respect to controllable variables x and the objective is quadratic with respect to the predicted values $T(i)$ and $O(i)$. Therefore, among many possible choices of machine learning techniques, we use the epsilon-support vector regression (ε-SVR) with linear kernel [29] as the prediction model. Such a choice will render a nice mathematical structure for the optimization model, which in turn enables us to choose an effective optimization technique to solve the problem efficiently. We obtain the linear prediction models through the ε-SVR linear kernel method. Plugging the function of the prediction models back into the objective function with some rearrangement, we obtain a quadratic programming model as follows (by dropping the subscript t for simplicity):

$$\begin{aligned} \min_{x} \quad & V(x) = \frac{1}{2} x^{\mathrm{T}} H x + f^{\mathrm{T}} x + c \\ s.t \quad & A_q \cdot x \leq b_q; \quad A_e \cdot x = b_e; \quad B_l \leq x \leq B_u. \end{aligned} \tag{7}$$

where H is a real symmetric matrix of coefficients, f is a vector of coefficients, and c is a constant term, all of which can be easily constructed based on values from the prediction model. A_q, A_e, b_q, b_e, B_l, and B_u are compact representation of known constraints.

Therefore, at each time step, we end up with a quadratic programming problem for the optimization model. We adopt an efficient algorithm for quadratic programming, the interior point convex (IPC) method [8, 20, 7]. It uses a presolve procedure to remove redundancies and simplify constraints. It then tries to find a point where the Karush-Kuhn-Tucker (KKT) conditions hold, and use multiple corrections to improve the centrality of the current iteration.

As discussed before, the working mechanism of a coal-fired boiler is extremely complex and time-varying, and not all factors are observable through measurements. Therefore, a machine learning-based prediction model of this kind may produce time-related local bias because of the change of underlying hidden factors, such as the fluctuation of coal quality and the restart of boilers. By borrowing an idea from the field of control, we add an error compensation part to further improve the prediction accuracy by compensating the local bias, which is estimated by computing an average difference between the actual output and predicted output for a prior window size of time steps, and adding this value to the future predicted output to decrease residuals [38, 4]. At every time step, the latest prediction error is obtained and the new compensation value is added to $T_t(i)$ and $O_t(j)$ as constants. The window size S is another algorithmic tuning parameter of this method. Note that the error compensation can be effective because the input and output are sampled from a physically continuous system, thus the adjacent prediction errors may implicitly stores contextual information and can be used for better prediction. This work is also an example of how a well-trained machine learning model can be improved by leveraging its physical meaning. This approach can be extended to other applications, where prediction is made about a continuous system and prediction error is available in online operation.

The prediction model is trained offline and does not need to be re-trained any more. The time to solve the optimization problem dominates the latency in the control loop which is the lag from the observation to the correspond-

ing control operations. Because the lag is not taken into consideration when building the prediction model according the dataset, the closer the lag is to zero the better. It is also the reason to simplify the complex highly nonlinear optimization problem to a quadratic programming problem, which is very important for a real-time solution.

4 Experimental Results

4.1 Experiment Setup

We conduct experiments using the real dataset collected by our industry partner from a production power plant boiler as discussed before. It contains more than 13,000 samples collected in a span of more than two months at a sample rate of 432 seconds. Each sample corresponds to a time stamp with 49 features including temperatures in six zones, O_2 content in two sides of flue, generation load, Nitric oxide in two sides of flue, twelve coal feed rates, and sixteen throttle openings, etc.

For comparison purpose, we also implement different algorithms to show the effectiveness of our proposed algorithm. The alternative options used for the prediction model include the ε-SVR with a RBF kernel, the classic three layer feed forward neural network (NN) with tangent-sigmoid activation function for the hidden layer, the vanilla recurrent neural network (RNN) [40], and the LSTM model [12]. The alternative options for the optimization model include some popular heuristic search algorithms, including GA, DE, PSO, and Sequential Quadratic Programming (SQP) [21]. All tuning parameters are selected by Bayesian optimization [26] or grid search on a validation dataset.

4.2 Comparison of Prediction Models

We first compare the prediction accuracy among the five prediction models. For each model, there are also different ways of organizing the input data (or feature selection for ε-SVR based methods). Three variants are considered: (A) non predicting data from the current time stamp, (B) all data from the current time stamp, and (C) all data from both the current time stamp and a varying number of past time steps. Most existing work on boiler optimization uses the type (A) data [42, 32, 5] as they assume a steady state model. Type (B) data is a special case of type (C) data with zero previous time step data.

To show the importance of organizing input data properly, we apply all the three types of data to the first three methods, and only type (B) data to RNN and LSTM models. The reason for the latter is that RNN and LSTM needs time-dependent data and the models themselves can be trained to capture the time-delayed effect through internal memories. The accuracy metrics used are averaged Mean Squared Error (MSE) and mean absolute percentage error (MAPE) of the six zones for temperature and two side for O_2. The prediction accuracy for temperature is reported in Table 1. As it can be seen, for the first three methods, results from type (B) and (C) data are significantly better than those from type (A) data and results from type (C) data are the best.

Table 1: Temperature prediction models

Model	Data Type	MSE	MAPE
SVR (linear)	(A)	975.3	2.06%
	(B)	289.2	1.12%
	(C)	**164.8**	**0.82%**
SVR (RBF)	(A)	1860.1	2.88%
	(B)	1199.8	2.24%
	(C)	246.6	1.01%
NN	(A)	1268.8	2.55%
	(B)	344.4	1.23%
	(C)	181.0	0.87%
RNN	(B)	635.6	1.89%
LSTM	(B)	841.7	1.98%

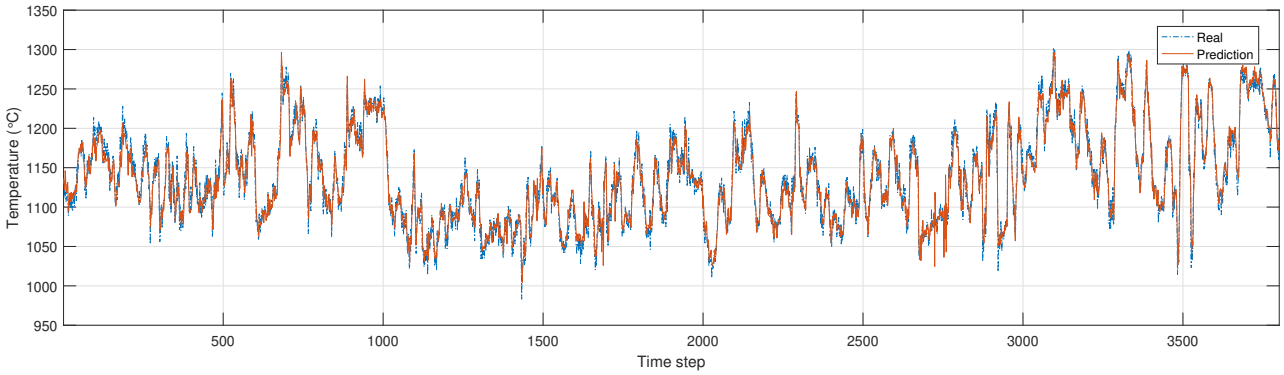


Figure 2: Comparison of real temperature and predicted temperature in zone 1

In terms of methods, although RNN and LSTM seem to be the most suitable models for time series data, we suspect the limited data size (albeit one of the largest in the literature) and the peculiarity of the system dynamics have prevented RNN and LSTM from achieving a stable solution such that the hidden states memorizing past information are weaker than raw data from past steps when supporting the subsequent predictions. The proposed ε-SVR linear model performs the best with the least average MSE value. Even when comparing results from type (B) data for all five methods, ε-SVR linear is still better than RNN and LSTM. The temperature prediction result on test dataset by ε-SVR linear model is illustrated in Figure 2.

We also offer the following reasons to explain why our proposed ε-SVR with linear kernel performs the best. First, the measurements of temperature and O_2 content contain some outliers because of the unstable airflow inside the boiler. The ε-SVR with ℓ_1 loss is less sensitive to such outliers when compared to the ℓ_2 loss in other methods. Second, ε-SVR treats errors less than ε as zero, and is thus less sensitive to sensor noises than others, which further helps to reduce unnecessary updates in the training process. Third, the simple linear regression is less likely to be overfitting compared to those more complex nonlinear models. Moreover, even though NN can provide better prediction performance, it cannot be used in the control system as it lead to a highly nonlinear optimization problem which is too complex to be solved in real-time. The results of O_2 prediction is quite similar to that in the temperature prediction.

4.3 Impact of Error Compensation

In this section, we report the impact of error compensation on the solution quality. Different window sizes can be tried on the historical data and the window size S can be selected with a trade-off between complexity and performance. Figure 3 shows the impact of different window sizes of error compensation on temperature prediction in different zones. The window size in x-axis indicates how many previous samples are used to calculate the error compensation, which is the average error for previous predictions. The y-axis stands for how much MSE are reduced by using error compensation with a given window size and thus the higher the better.

A rough rise and fall trend can be observed as expected in Figure 3. When the window size is small, which means only a few latest errors are used to calculate the compensation, the error compensation makes prediction worse (negative values in Figure 3) because high randomness dominates the compensation. When window size increases, the compensation helps to reduce prediction errors and these curves reach a peak since a proper window size enables us to discover a local bias covered by randomness. As the window size keeps increasing, these curves fall down as too many prediction errors from long ago are used. If the window size reaches a very large value, all curves will finally converge to a narrow range around zero because it leads compensation to a near-zero value, which is not shown within this figure. It is worth noting that predicted temperatures with lower accuracy tend to get more improvement from error compensation. We suspect those zones are more sensitive to some hidden factors and thus have more apparent local bias. Similar observations also hold for O_2 content prediction. We finally select 50 as the window size for error compensation calculation to strike a right balance. With error compensation, we further reduced the average MSE of temperatures and O_2 content by 7.4% and 3.4% respectively.

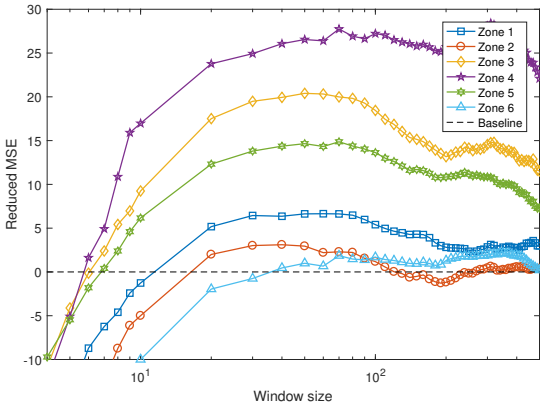


Figure 3: The reduced MSE versus window size of error compensation

More complicated approaches such as SVR and NN also have been tested, but surprisingly, even though they get better result under some settings, the simplest average strategy gets the best and most stable performance overall under this practical circumstance. This is probably caused by the high randomness of the boiler system and its variance along with time.

4.4 Comparison of Optimization Algorithms

We compare various optimization algorithms on the test dataset using the best prediction model obtained in last section. At each time step, the optimization algorithm will produce one objective value and for all test samples, the objective values are collected for each model. We report the comparison results in Table 2, where the solution quality is measured by the objectives collected for all test samples, and we report their mean, minimum, maximum and standard deviation value for simplicity. The smaller the objective value, the better the solution quality. The computation time is measured by the time to converge in seconds on a desktop with an Intel i5-4590 3.3GHz CPU.

Table 2: Comparison of different optimization algorithms

Solving	Time	Objectives			
Algorithm	(sec)	Mean	Min	Max	Std
IPC	**0.16**	**0.085**	**-0.207**	**0.419**	**0.140**
DE	81.5	0.127	-0.168	0.497	0.145
SQP	159	0.117	-0.189	0.434	0.138
PSO	N/C	0.235	-0.121	0.599	0.151
GA	N/C	0.586	0.158	1.093	0.234

We see from Table 2 that only IPC, DE, and SQP can provide a converged solution within the given time interval, while PSO and GA cannot. IPC outperforms DE and SQP on both runtime and result quality significantly. This is expected, as IPC is a most suited optimization algorithm for the special mathematical structure as derived in this work, while other algorithms are generic optimization techniques.

Based on the same prediction model, we observe that solutions from IPC based control are able to reduce the temperature standard deviation by 42.5%, and O_2 content difference by 61.5% when compared to the the original test data without optimization. At the same time, we see 32°C higher average temperature and 38.6% lower average O_2 content, indicating that the proposed model can also improve combustion efficiency simultaneously.

5 Conclusions

Equipped with the unique dataset collected from a real power plant, we introduce a new formulation for boiler control problem that focuses on maintaining not only high temperature and low O_2 content, but also a balanced distribution of temperature and O_2 content. To overcome the foremost challenge of developing a real-time solution, we propose a new algorithmic framework that incorporates a machine learning-based prediction model, an optimization model, and an error compensation model. Experimental results validate the effectiveness and efficiency of the solution. The solution framework can be extended to other Cyber-Physical Systems where the online control or optimization is constrained by the complexity of prediction and its formulation.

References

[1] A. Conn, N. Gould, and P. Toint. A globally convergent lagrangian barrier algorithm for optimization with general inequality constraints and simple bounds. *Mathematics of Computation of the American Mathematical Society*, 66(217):261–288, 1997.

[2] L. I. Díez, C. Cortés, and A. Campo. Modelling of pulverized coal boilers: review and validation of on-line simulation techniques. *Applied Thermal Engineering*, 25(10):1516–1533, 2005.

[3] Y. Ding, J. Liu, J. Xiong, and Y. Shi. On the universal approximability and complexity bounds of quantized relu neural networks. *arXiv preprint arXiv:1802.03646*, 2018.

[4] N. R. Draper, H. Smith, and E. Pownell. *Applied regression analysis*, volume 3. Wiley New York, 1966.

[5] W. D. Feng, M. Li, M. Li, and H. Pu. Combustion optimization based on rbf neural network and multi-objective genetic algorithms. In *Genetic and Evolutionary Computing, 2009. WGEC'09. 3rd International Conference on*, pages 496–501. IEEE, 2009.

[6] J. C. Foreman. *Architecture for intelligent power systems management, optimization, and storage*. University of Louisville, 2008.

[7] J. Gondzio. Multiple centrality corrections in a primal-dual method for linear programming. *Computational Optimization and Applications*, 6(2):137–156, 1996.

[8] N. Gould and P. L. Toint. Preprocessing for quadratic programming. *Mathematical Programming*, 100(1):95–132, 2004.

[9] Y. Gu, W. Zhao, and Z. Wu. Online adaptive least squares support vector machine and its application in utility boiler combustion optimization systems. *Journal of Process Control*, 21(7):1040–1048, 2011.

[10] S. Han, H. Mao, and W. J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.

[11] H. Hasini, M. Z. Yusoff, N. H. Shuaib, M. H. Boosroh, and M. A. Haniff. Analysis of flow and temperature distribution in a full scale utility boiler using cfd. In *Energy and Environment, 2009. ICEE 2009. 3rd International Conference on*, pages 208–214. IEEE, 2009.

[12] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[13] N. Inc. Boiler optimization software solution, 2017.

[14] A. Kusiak, A. Burns, and F. Milster. Optimizing combustion efficiency of a circulating fluidized boiler: A data mining approach. *International Journal of Knowledge-based and Intelligent Engineering Systems*, 9(4):263–274, 2005.

[15] A. Kusiak and Z. Song. Combustion efficiency optimization and virtual testing: A data-mining approach. *IEEE Transactions on Industrial Informatics*, 2(3):176–184, 2006.

[16] G. Li and P. Niu. Combustion optimization of a coal-fired boiler with double linear fast learning network. *Soft Computing*, 20(1):149–156, 2016.

[17] G.-Q. Li, X.-B. Qi, K. C. Chan, and B. Chen. Deep bidirectional learning machine for predicting no x emissions and boiler efficiency from a coal-fired boiler. *Energy & Fuels*, 31(10):11471–11480, 2017.

[18] J. Liu, J. Zhang, Y. Ding, X. Xu, M. Jiang, and Y. Shi. Pbgan: Partial binarization of deconvolution based generators. *arXiv preprint arXiv:1802.09153*, 2018.

[19] X. Liu, P. Guan, and C. Chan. Nonlinear multivariable power plant coordinate control by constrained predictive scheme. *IEEE transactions on control systems technology*, 18(5):1116–1125, 2010.

[20] E. Mezura-Montes, J. Velázquez-Reyes, and C. A. Coello Coello. A comparative study of differential evolution variants for global optimization. In *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, pages 485–492. ACM, 2006.

[21] J. Nocedal and S. J. Wright. *Sequential quadratic programming*. Springer, 2006.

[22] H. Y. Park, S. H. Baek, Y. J. Kim, T. H. Kim, D. S. Kang, and D. W. Kim. Numerical and experimental investigations on the gas temperature deviation in a large scale, advanced low nox, tangentially fired pulverized coal boiler. *Fuel*, 104:641–646, 2013.

[23] M. Pechenizkiy, J. Bakker, I. Žliobaitė, A. Ivannikov, and T. Kärkkäinen. Online mass flow prediction in cfb boilers with explicit detection of sudden concept drift. *ACM SIGKDD Explorations Newsletter*, 11(2):109–116, 2010.

[24] X. Peng and P. Wang. An improved multiobjective genetic algorithm in optimization and its application to high efficiency and low nox emissions combustion. In *Power and Energy Engineering Conference, 2009. APPEEC 2009. Asia-Pacific*, pages 1–4. IEEE, 2009.

[25] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European Conference on Computer Vision*, pages 525–542. Springer, 2016.

[26] J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[27] M. Szega and G. T. Nowak. An optimization of redundant measurements location for thermal capacity of power unit steam boiler calculations using data reconciliation method. *Energy*, 92:135–141, 2015.

[28] I. C. Trelea. The particle swarm optimization algorithm: convergence analysis and parameter selection. *Information processing letters*, 85(6):317–325, 2003.

[29] V. Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2013.

[30] C. Wang, Y. Liu, S. Zheng, and A. Jiang. Optimizing combustion of coal fired boilers for reducing nox emission using gaussian process. *Energy*, 2018.

[31] World Coal Association. Coal and Electricity. Technical report, World Coal Association, 2016.

[32] W. Xu and C. Taihua. The balanced model and optimization of nox emission and boiler efficiency at a coal-fired utility boiler. In *Conference Anthology, IEEE*, pages 1–4. IEEE, 2013.

[33] X. Xu, Y. Ding, S. X. Hu, M. Niemier, J. Cong, Y. Hu, and Y. Shi. Scaling for edge inference of deep neural networks. *Nature Electronics*, 1(4):216, 2018.

[34] X. Xu, Q. Lu, T. Wang, Y. Hu, C. Zhuo, J. Liu, and Y. Shi. Efficient hardware implementation of cellular neural networks with incremental quantization and early exit. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, 14(4):48, 2018.

[35] X. Xu, Q. Lu, L. Yang, S. Hu, D. Chen, Y. Hu, and Y. Shi. Quantization of fully convolutional networks for accurate biomedical image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8300–8308, 2018.

[36] X. Xu, T. Wang, Q. Lu, and Y. Shi. Resource constrained cellular neural networks for real-time obstacle detection using fpgas. In *2018 19th International Symposium on Quality Electronic Design (ISQED)*, pages 437–440. IEEE, 2018.

[37] C. Yin, L. Rosendahl, and T. J. Condra. Further study of the gas temperature deviation in large-scale tangentially coal-fired boilers. *Fuel*, 82(9):1127–1137, 2003.

[38] G. Zhang, R. Veale, T. Charlton, B. Borchardt, and R. Hocken. Error compensation of coordinate measuring machines. *CIRP Annals-Manufacturing Technology*, 34(1):445–448, 1985.

[39] S. Zhang, C. W. Taft, J. Bentsman, A. Hussey, and B. Petrus. Simultaneous gains tuning in boiler/turbine pid-based controller clusters using iterative feedback tuning methodology. *ISA transactions*, 51(5):609–621, 2012.

[40] Y. Zhang, H. Dai, C. Xu, J. Feng, T. Wang, J. Bian, B. Wang, and T.-Y. Liu. Sequential click prediction for sponsored search with recurrent neural networks. *arXiv preprint arXiv:1404.5772*, 2014.

[41] Y. Zhang, Y. Ding, Z. Wu, L. Kong, and T. Chou. Modeling and coordinative optimization of no x emission and efficiency of utility boilers with neural network. *Korean Journal of Chemical Engineering*, 24(6):1118–1123, 2007.

[42] H. Zhao and P.-h. Wang. Modeling and optimization of efficiency and nox emission at a coal-fired utility boiler. In *Power and Energy Engineering Conference, 2009. APPEEC 2009. Asia-Pacific*, pages 1–4. IEEE, 2009.

[43] W. Zhao, G. Zhao, M. Lv, and J. Zhao. Fuzzy optimization control for nox emissions from power plant boilers based on nonlinear optimization 1. *Journal of Intelligent & Fuzzy Systems*, 29(6):2475–2481, 2015.

[44] L. Zheng, H. Zhou, C. Wang, and K. Cen. Combining support vector regression and ant colony optimization to reduce nox emissions in coal-fired utility boilers. *Energy & Fuels*, 22(2):1034–1040, 2008.

Exploring On-Chip Power Delivery Network Induced Analog Covert Channels

Longfei Wang[1], S. Karen Khatamifard[2], Ulya R. Karpuzcu[2], Selçuk Köse[1]
[1]Department of Electrical and Computer Engineering, University of Rochester
[2]Department of Electrical and Computer Engineering, University of Minnesota

1 Introduction

On-chip power delivery network is an essential part of modern integrated circuits. With a sophisticated control by the power management unit, an off-chip voltage level is converted and regulated to a dedicated voltage applicable to the on-chip load circuits. Meanwhile, high power conversion efficiency is maintained as load current changes. Components of a representative on-chip power delivery network [1, 2, 3, 4, 5, 6, 7, 8, 9] are shown in Figure 1. Within this network, output voltage of an off-chip voltage converter is supplied to the global power grid through VDD C4 pads. The inputs of on-chip voltage regulators (VRs) are connected to the global power grid and the outputs of on-chip VRs are connected to the local power grids. Global ground distribution supplies the ground plane and is connected to the package through GND C4 pads. Multiple voltage domains can be enabled by the distributed VRs providing disparate local power grids. Significant amount of work has been performed to demonstrate potential security vulnerabilities of shared resources within multi/many-core processors. One of these inevitably shared resources is the constituent of the power management and delivery subsystem such as the global power grid shown in Figure 1. Our recent works [10, 11] reveal a new covert channel due to shared power budget enforced by hierarchical on-chip power management. In this article, a previously unexplored, novel class of analog covert channel leveraging switching noise modulation is uncovered. The threat model and related mechanisms to form the covert communication are detailed and proof of concept results are provided.

2 Threat Model

Covert channels make leakage of sensitive information possible even when there is no designated media for transmission of such information [12]. The transmitting and receiving end can be, respectively, referred to as the source

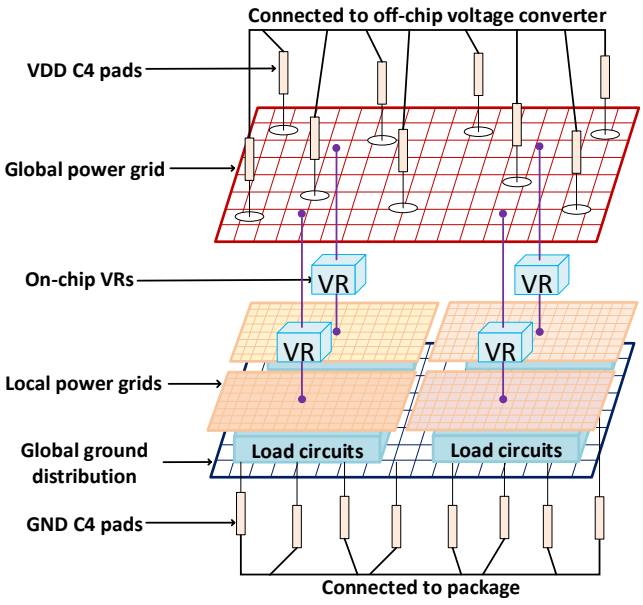


Figure 1: On-chip power delivery network.

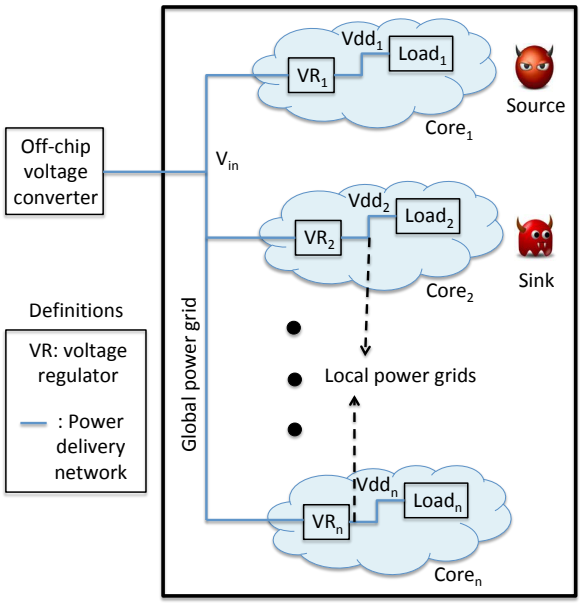


Figure 2: Power delivery network induced covert channel.

and sink. Both the source and sink can be either hardware components or software components sharing hardware resources [13]. The source is assumed to have access to the sensitive information but not to the network, which can be accessed by third parties. On the other hand, the sink is assumed to be capable of transmitting information through the network but have no access to the sensitive information. Due to the existence of covert channels, the sensitive information can be transmitted from the source to the sink and further to the third parties through the network. The covert communication, however, is not apparent to the hardware and software layers residing in the same system. In this article, on-chip VRs are considered as the source and sink. Covert communication is established through the shared global power grid.

3 Proof of Concept Results

As a proof of concept example, a power delivery network consisting of an off-chip VR and multiple on-chip VRs is considered, as shown in Figure 2. The output voltage of the off-chip VR is connected to the global power grid, which supplies the inputs of the on-chip VRs. There are n cores with each powered by an on-chip VR. The output of each on-chip VR is connected to a local power grid providing the supply voltage of the load circuit within that specific core. It is assumed that the sink core is idle while the source core is active when covert channel needs to be established. Without loss of generality, low-dropout (LDO) regulators are implemented as the on-chip VRs for this example. LDOs similar to [5] are adopted and an RC chain model is utilized for the power grid. Power grid parameters from [14] are applied. The activity of the source is modeled as a transient current at the output of VR_1 with the pattern decided by a random bit stream. When there is no covert communication, the load current of VR_1 consists of some leakage current and this transient current. When sensitive information needs to be transmitted from the source to the sink, a periodic current encoding the information will be added to $Load_1$. Meanwhile, as the sink is idle, the load current of VR_2, $Load_2$, only carries a small transient current and some leakage current. Due to the added periodic current to $Load_1$, fluctuations at the input of VR_1 are introduced as the control loop of VR_1 begins to respond. Such fluctuations also occur at the input of VR_2 due to the shared global power grid. The control loop of VR_2 also responds to maintain a constant supply voltage Vdd_2.

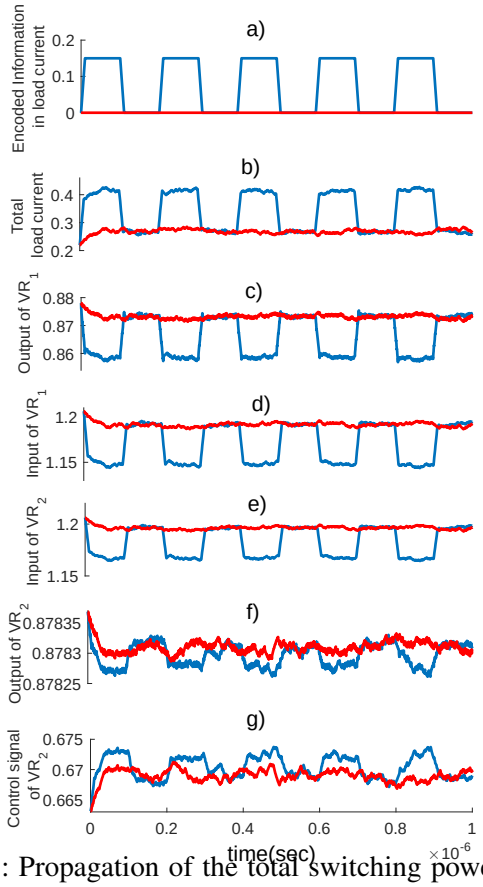


Figure 3: Propagation of the total switching power noise from source to sink over the global power delivery network with intentional noise (i.e., encoded information leak) shown in blue and without intentional noise shown in red.

The transmission of the voltage fluctuations from the source to the sink core is simulated in Cadence according to the above discussion utilizing a 45nm CMOS process. The simulation results with and without intentional noise are demonstrated in Figure 3 with, respectively, blue and red lines. The encoded information reflected in the load current of the source is shown in Figure 3a. The total load current of the source is shown in Figure 3b. The output of VR_1 experiences fluctuations due to the periodically switching load current as featured in Figure 3c which in turn lead to fluctuations at the input of VR_1 as shown in Figure 3d. Such fluctuations can further propagate to the input of VR_2 through the shared global power grid as can be seen from Figure 3e. This intentional noise generated at the output of VR_1 results in some fluctuations at the output of VR_2 seen from Figure 3f. The control signal of

VR_2, which is visible to both the local and global controllers in a hierarchical power management system, is also highly affected by the encoded information due to the tight integration of the power delivery network that consists of the distributed VRs as demonstrated in Figure 3g. As supported by the preliminary data, the sensitive information encoded in the form of a switching load current at the output of an on-chip VR can propagate through the global power grid and be sensed by the local power controller of the other cores. Considering the digital control of on-chip VRs that is implemented in state-of-the-art integrated systems, sensitive information carried by the control signal of the on-chip VR at the sink side can be processed digitally without dedicated hardware.

4 Discussions

This proof-of-concept demonstrates the crucial need for security-aware design of on-chip power delivery network. As the number of voltage regulators that co-reside on a single die increases, the distributed power delivery networks require tighter integration which leads to the increased number of shared resources such as capacitors (be it flying or decoupling), inductors, and most importantly the local and global power/ground interconnection network. Additionally, the design of each individual VR becomes more complicated due to the challenges such as reliability, stability, power efficiency, response time, area, and workload-awareness. Each additional feature to tackle any of these challenges would potentially make the power delivery network more vulnerable against covert communication attacks similar to those explored in this paper. We claim that security should be included within these challenges early in the design process not only at the system or architectural level but also at the low level (analog/mixed signal/digital) circuit design.

5 Conclusions

On-chip power delivery network provides regulated voltage levels to the load circuits while at the same time is vulnerable to information leakage through shared resources. A power delivery network induced analog covert channel enabled by shared global power grid and switching noise modulation is investigated in this article. Due to the strong correlation between the input and output of on-chip VRs, fluctuations can be introduced at the input of VR at the source side due to added switching load current. Such fluctuations propagate through the shared global power grid and are finally sensed by the local power control circuitry of the other cores. Proof of concept results for the on-chip power delivery network induced analog covert channel are demonstrated through Cadence simulations. Increased design complexity and shared resources necessitate inclusion of security features at the early design stage.

6 Acknowledgement

This work is supported in part by the NSF CAREER Award under Grant CCF-1350451, in part by the NSF Award under Grant CNS-1715286, in part by SRC Contract NO: 2017-TS-2773, and in part by the Cisco Systems Research Award.

References

[1] I. Vaisband, R. Jakushokas, M. Popovich, A. V. Mezhiba, S. Köse, and E. G. Friedman, *On-Chip Power Delivery and Management, Fourth Edition.* Springer, 2016.

[2] S. B. Nasir, Y. Lee, and A. Raychowdhury, "Modeling and Analysis of System Stability in a Distributed Power Delivery Network with Embedded Digital Linear Regulators," *Proceedings of the IEEE International Symposium on Quality Electronic Design (ISQED)*, pp. 68-75, February 2019.

[3] J. F. Bulzacchelli, Z. Toprak-Deniz, T. M. Rasmus, J. A. Iadanza, W. L. Bucossi, S. Kim, R. Blanco, C. E. Cox, M. Chhabra, C. D. LeBlanc, C. L. Trudeau, and D. J. Friedman, "Dual-Loop System of Distributed

Microregulators with High DC Accuracy, Load Response Time Below 500 ps, and 85-mV Dropout Voltage," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 4, pp. 863-874, April 2012.

[4] P. Zhou, D. Jiao, C. H. Kim, and S. S. Sapatnekar, "Exploration of On-Chip Switched-Capacitor DC-DC Converter for Multicore Processors using a Distributed Power Delivery Network," *Proceedings of the IEEE Custom Integrated Circuits Conference (CICC)*, pp. 1-4, September 2011.

[5] S. Lai, B. Yan, and P. Li, "Localized Stability Checking and Design of IC Power Delivery with Distributed Voltage Regulators," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 32, no. 9, pp. 1321-1334, September 2013.

[6] W. Yu and S. Köse, "Exploiting Voltage Regulators to Enhance Various Power Attack Countermeasures," *IEEE Transactions on Emerging Topics in Computing*, vol. 6, no. 2, pp. 244-257, April-June 2018.

[7] W. Yu, O. A. Uzun, and S. Köse, "Leveraging On-Chip Voltage Regulators as a Countermeasure Against Side-Channel Attacks," *Proceedings of the 52nd Annual Design Automation Conference*, pp. 1-6, June 2015.

[8] O. A. Uzun and S. Köse, "Converter-Gating: A Power Efficient and Secure On-Chip Power Delivery System," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 4, no. 2, pp. 169-179, June 2014.

[9] L. Wang, S. K. Khatamifard, O. A. Uzun, U. R. Karpuzcu, and S. Köse, "Efficiency, Stability, and Reliability Implications of Unbalanced Current Sharing among Distributed On-Chip Voltage Regulators," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 11, pp. 3019-3032, November 2017.

[10] S. K. Khatamifard, L. Wang, S. Köse, and U. R. Karpuzcu, "POWERT Channels: A Novel Class of Covert Communication Exploiting Power Management Vulnerabilities," *Proceedings of the IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, February 2019.

[11] S. K. Khatamifard, L. Wang, S. Köse, and U. R. Karpuzcu, "A New Class of Covert Channels Exploiting Power Management Vulnerabilities," *IEEE Computer Architecture Letters (CAL)*, vol. 17, no. 2, pp. 201 - 204, July - December 1 2018.

[12] Department of Defense Trusted Computer System Evaluation Criteria (Orange Book), December 26, 1985.

[13] H. Ritzdorf, "Analyzing Covert Channels on Mobile Devices," M.S. Thesis, ETH, 2012.

[14] R. Zhang, K. Wang, B. H. Meyer, M. R. Stan, and K. Skadron, "Architecture Implications of Pads as a Scarce Resource," *Proceedings of the 41st Annual International Symposium on Computer Architecture (ISCA)*, pp. 373-384, June 2014.

IP Protection and Supply Chain Security through Logic Obfuscation

Meng Li[1], Kaveh Shamsi[2], Yier Jin[2], David Z. Pan[1]
[1]Department of Electrical and Computer Engineering, University of Texas at Austin
[2]Department of Electrical and Computer Engineering, University of Florida

1 Introduction

Recent decades have witnessed the globalization of the integrated circuit (IC) supply chain propelled by the ever-increasing design complexity and cost. However, such globalization comes at a cost. Although it has helped to reduce the overall cost by the worldwide distribution of IC design, fabrication, assembly and deployment, it also introduces serious intellectual property (IP) privacy violations due to reverse engineering [1, 2]. As shown in Figure 1, given a packaged IC, after de-packaging, de-layering, imaging and post-processing, the original circuit netlist can be reconstructed. Over the last decade, such reverse engineering techniques have developed rapidly, demonstrating successful reconstruction of products of leading semiconductor companies in advanced technology nodes [3].

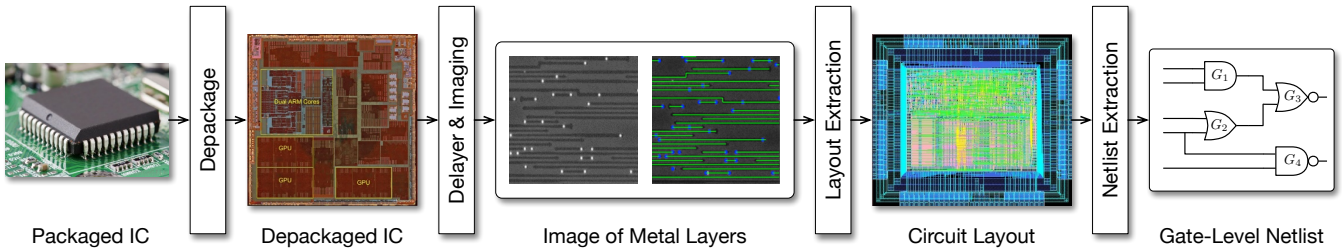


Figure 1: Physical reverse engineering flow.

To thwart reverse engineering, logic obfuscation techniques, including logic locking [4, 5, 6] and IC camouflaging [7, 8], are proposed to hide the critical circuit components against the attackers. IC camouflaging leverages fabrication-level techniques to build circuits whose functionality cannot be easily deduced using known physical reverse engineering techniques [7]. Camouflaging units and cells are first designed and then inserted throughout the netlist with different insertion strategies in the IC camouflaging process. Logic locking arguments the original design with additional key-controlled logic gates (i.e., key-gates), key-bits and an on-chip memory to enhance the circuit programmability and hide the circuit functionality [9]. The correct circuit functionality only manifests itself upon the programming of the correct key-bit.

The insertion of the key-gates or camouflaging cells for circuit obfuscation does not imply security against the reverse engineering. Over the past decade, different attack strategies have been proposed to infer the functionality of the obfuscated netlist. The arms race between the logic obfuscation and reverse engineering inspires stronger and more rigid obfuscation algorithms and drives the evolution of the entire area. In the article, we will first review the evolution of the obfuscation and attack techniques, based on which, we will sketch the future directions in the area.

2 Overview of the Logic Obfuscation Research

The attack model for a given security problem defines the capabilities and intentions of the attacker. While different attack models have been proposed [7, 10, 9], the most widely used attack model, referred to as the oracle-guided attack, grants the attackers with the access to the following two components:

- The obfuscated netlist, which can be acquired from the physical reverse engineering process. In the netlist, the attackers cannot determine the functionality of the camouflaging cells or the logic values of the key bits.
- A functional circuit, which can be acquired from the open market and is treated as a black-box circuit. The attackers cannot observe or probe the internal signals of the functional circuit directly.

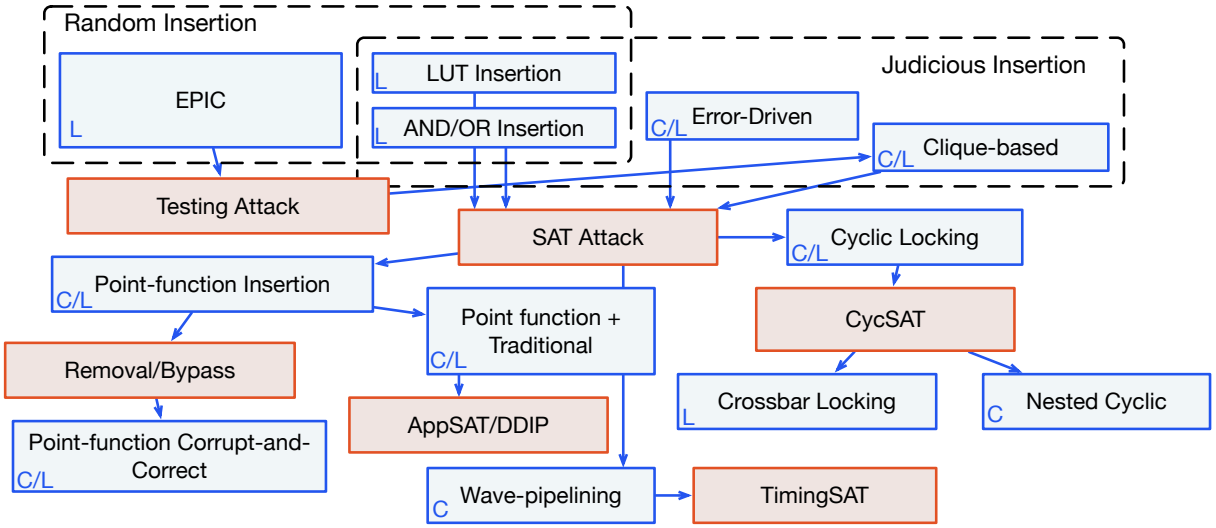


Figure 2: Overview of the evolution of hardware obfuscation. The red boxes denote attacks and blue boxes denote defenses. The C(camouflaging) and L (locking) tags describe the category of defenses.

Given the obfuscated netlist, the attackers can select a sequence of input vectors, import them into the functional circuit through the scan chain, query the functional circuit and observe the corresponding outputs. The correct circuit functionality can be inferred given the collected input-output pairs.

Over the past decade, extensive research has been conducted for both logic obfuscation and reverse engineering. Most of the researches focus on combinational logic with a few studies on sequential logic [11]. The arms-race between the attack and protection is summarized in Figure 2.

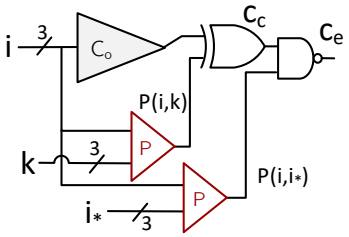
The first logic locking strategy, EPIC, is proposed in [4]. EPIC randomly inserts key-controlled XOR/XNOR gates into the netlist and proposes a key distribution framework using the public-key cryptography. EPIC only considers the brute force attack, which exhaustively enumerate the key space. Following EPIC, [12] proposes a new locking strategy that leverages key-controlled lookup tables (LUTs), denoted as barriers. Compared with EPIC, [12] leverage heuristic methods to judiciously insert the LUTs into the circuit and maximize the output error probability for incorrect keys.

To deobfuscate the obfuscated circuit, the testing-based attack is proposed in [7] and gets further enhanced in [13]. The attack leverages hardware testing principles, i.e., sensitization and justification, to select the input vectors to query the functional IC and demonstrates a strong capability to attack the obfuscated circuits. However, it is observed that when different key-bits can interfere with each other, the testing-based attack can be hindered. Therefore, [7] proposes a clique-based obfuscation strategy to insert key-gates or camouflaging gates that form a clique in the circuit so that the number of interfering key pairs are maximized.

In response to the clique-based obfuscation, the SAT-based attack is proposed [14]. The deobfuscation problem is formulated into an SAT problem, based on which the input vectors that are guaranteed to prune the incorrect circuit functionalities, denoted as discriminating inputs, are acquired. The SAT-based attack demonstrates a strong capability to deobfuscate all the existing protection strategies within minutes even for a large key size.

To enhance the resilience against the SAT-based attack, how to increase the number of discriminating inputs becomes an important question. Point function-based obfuscation strategies are thus developed [8, 5]. As shown in Figure **??**, a point function realized by an AND-tree with camouflaged tree inputs can be inserted into the netlist, which can be proved to require an exponential number of discriminating inputs to deobfuscate. The point function-based obfuscation can be further combined with traditional random obfuscation strategy or re-synthesis [8, 5] to defend against the removal attacks, which try to leverage the structural footprint of the point functions to detect and remove them [9].

Although the AND-tree based strategies achieve an exponential increase of resilience against the SAT-based attacks with respect to the AND-tree size, [16, 17] observes that the output error probability for an incorrect key also reduces exponentially. Hence, an approximate SAT-based attack, denoted as the AppSAT attack, is proposed



(a)

input	output	key							
$I_2I_1I_0$	O	K_0^*	K_1	K_2	K_3	K_4	K_5	K_6	K_7
000	0	0	0	0	0	0	0	0	0
001	1	1	0	1	1	1	1	1	1
010	1	1	1	0	1	1	1	1	1
011	0	0	0	0	1	0	0	0	0
100	0	0	0	0	0	1	0	0	0
101	1	1	1	1	1	1	0	1	1
110	0	0	0	0	0	0	0	1	0
111	0	0	0	0	0	0	0	0	1

(b)

Figure 3: A point-function scheme example. The truth table on the right demonstrates that a SAT attack will have to query all input patterns to disqualify all possible keys (red cells). Figure and table are adapted from [15].

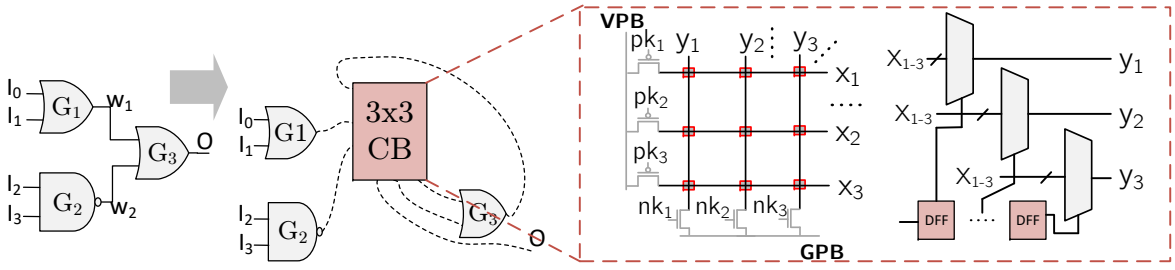


Figure 4: Cyclic wire obfuscation leveraging a crossbar structure. The crossbar can be implemented with multiplexers and a scan-chain as well however with significantly more area overhead. Figures are adapted from [22].

in [16, 17], which can efficiently obtain a netlist that functions correctly with very high probability. The AppSAT attack forces the protection to consider both the output error probability and the number of discriminating inputs.

To defend against the AppSAT attack, [18, 19] propose to introduce unconventional structures for circuit obfuscation. In [19], circuit wire interconnections are camouflaged by introducing dense, nested, "fake" cyclic structures into the netlist, which does not impact the circuit functionality but significantly complicates the AppSAT attack process. [18] proposes to deliberately remove the flip-flops in the circuit to create a mixture of single-cycle and multiple-cycle paths in the circuit, which cannot be directly resolved by the AppSAT attacks either. Cyclic obfuscated circuits were first broken with the CycSAT algorithm proposed in [20]. It was later shown that CycSAT may run into problems when attacking dense cyclic circuits as cycle enumeration of such circuit may be necessary. As for timing-based camouflaging schemes, TimingSAT [21] was proposed and tested on benchmark circuit as well.

3 Future Directions

While the arms race between the logic obfuscation and reverse engineering leads to better obfuscation strategies and significantly boosts the whole area, there are still critical challenges to be addressed, including formal proof of security, more generic and quantitative security analysis, full system-level obfuscation flow, etc. These challenges motivate the following research directions:

- Cryptographic criteria over Boolean functions: state-of-the-art cryptoanalysis usually enjoys high formalism. The cryptographic primitives, including the encryption and decryption engines, have been widely used with good security properties under the basic assumptions on the hardness of mathematical problems. Hence, if we can bridge the cryptography area with the logic obfuscation, a better security metric can be expected. In fact, Boolean functions have been of great importance in cryptography. Boolean values of 0 and 1 are

treated as elements of the finite field GF (2). There are also several characteristics of Boolean functions that are significant in cryptography as certain functions will create more confusion when used in cryptographic primitives [23], including non- linearity, correlation-immunity, strict avalanche criterion, etc. How to bridge these cryptographic criteria with the logic obfuscation strategies can be a promising direction.

- Obfuscation protection and attack with new attack models: develop more diversified attack models to capture the attackers' capability on reverse engineering the hardware IP. To formally evaluate the impact of the fault attack and the side channel attack, the attack models need to be defined, including the attackers' knowledge, the accessibility to the circuit, the fault injection techniques and so on. Based on the attack models, protection and attack strategies can be proposed and formalized to further enhance the practicality and applicability of the logic obfuscation strategies. Meanwhile, the recently proposed performance locking [24] can be another promising direction beyond the function locking that is intensively studied.

- Full chip logic obfuscation and EDA flow evaluation: instead of simply focusing on the combinational logic, full chip logic obfuscation needs to be considered and demonstrated with prototype chips. Full chip obfuscation requires formal security analysis for the sequential logic and rigorous evaluation of the impact of circuit partitioning. Meanwhile, the impact of EDA tools and current EDA flow needs to be evaluated, especially considering the efficiency of the algorithm and the security implication of different design stages, including placement, routing, and so on.

References

[1] R. Torrance and D. James, "The state-of-the-art in IC reverse engineering," in *Proc. Int. Conf. on Cryptographic Hardware and Embedded Systems*, pp. 363–381, Springer, 2009.

[2] S. E. Quadir, J. Chen, D. Forte, N. Asadizanjani, S. Shahbazmohamadi, L. Wang, J. Chandy, and M. Tehranipoor, "A survey on chip to system reverse engineering," *ACM J. on Emerging Technologies in Computing Systems*, vol. 13, no. 1, pp. 6:1–6:34, 2016.

[3] Chipworks, "Intel's 22-nm Tri-gate Transistors Exposed." http://www.eet.bme.hu/~mizsei/Montech/intel-s-22-nm-trigate-transistors-exposed.html, 2012.

[4] J. A. Roy, F. Koushanfar, and I. L. Markov, "EPIC: Ending piracy of integrated circuits," in *Proc. Design, Automation and Test in Europe*, pp. 1069–1074, 2008.

[5] Y. Xie and A. Srivastava, "Mitigating SAT attack on logic locking," in *Proc. Int. Conf. on Cryptographic Hardware and Embedded Systems*, pp. 127–146, 2016.

[6] M. Yasin, A. Sengupta, M. T. Nabeel, M. Ashraf, J. J. Rajendran, and O. Sinanoglu, "Provably-secure logic locking: From theory to practice," in *Proc. ACM Conf. on Computer & Communications Security*, pp. 1601–1618, ACM, 2017.

[7] J. Rajendran, M. Sam, O. Sinanoglu, and R. Karri, "Security analysis of integrated circuit camouflaging," in *Proc. ACM Conf. on Computer & Communications Security*, pp. 709–720, 2013.

[8] M. Li, K. Shamsi, T. Meade, Z. Zhao, B. Yu, Y. Jin, and D. Z. Pan, "Provably secure camouflaging strategy for IC protection," in *Proc. Int. Conf. on Computer Aided Design*, pp. 28:1–28:8, 2016.

[9] M. Yasin, B. Mazumdar, O. Sinanoglu, and J. Rajendran, "Security analysis of Anti-SAT," in *Proc. Asia and South Pacific Design Automation Conf.*, 2017.

[10] M. Yasin, O. Sinanoglu, and J. Rajendran, "Testing the trustworthiness of ic testing: An oracle-less attack on ic camouflaging," *IEEE Trans. on Information Forensics and Security*, vol. 12, no. 11, pp. 2668–2682, 2017.

[11] K. Shamsi, M. Li, D. Z. Pan, and Y. Jin, "KC2: Key-condition crunching for fast sequential circuit obfuscation," in *Proc. Design, Automation and Test in Europe*, 2019.

[12] A. C. Baumgarten, "Preventing integrated circuit piracy using reconfigurable logic barriers," 2009.

[13] Y. W. Lee and N. A. Touba, "Improving logic obfuscation via logic cone analysis," in *Proc. IEEE Latin-American Test Symp.*, pp. 1–6, 2015.

[14] M. El Massad, S. Garg, and M. V. Tripunitara, "Integrated circuit (IC) decamouflaging: Reverse engineering camouflaged ICs within minutes.," in *Proc. Network and Distributed System Security Symp.*, 2015.

[15] K. Shamsi, T. Meade, M. Li, D. Z. Pan, and Y. Jin, "On the approximation resiliency of logic locking and ic camouflaging schemes," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 2, pp. 347–359, 2019.

[16] K. Shamsi, M. Li, T. Meade, Z. Zhao, D. Z. Pan, and Y. Jin, "AppSAT: Approximately deobfuscating integrated circuits," in *Proc. IEEE Int. Symp. on Hardware Oriented Security and Trust*, 2017.

[17] Y. Shen and H. Zhou, "Double DIP: Re-evaluating security of logic encryption algorithms," in *Proceedings of the on Great Lakes Symposium on VLSI 2017*, pp. 179–184, ACM, 2017.

[18] L. Zhang, B. Li, B. Yu, D. Z. Pan, and U. Schlichtmann, "TimingCamouflage: Improving circuit security against counterfeiting by unconventional timing," in *Proc. Design, Automation and Test in Europe*, 2018.

[19] K. Shamsi, M. Li, T. Meade, Z. Zhao, D. Z. Pan, and Y. Jin, "Cyclic obfuscation for creating sat-unresolvable circuits," in *Proc. IEEE Great Lakes Symp. on VLSI*, 2017.

[20] H. Zhou, R. Jiang, and S. Kong, "CycSAT: SAT-based attack on cyclic logic encryptions," in *Proc. Int. Conf. on Computer Aided Design*, pp. 49–56, IEEE, 2017.

[21] M. Li, K. Shamsi, Y. Jin, and D. Z. Pan, "TimingSAT: Decamouflaging Timing-based Logic Obfuscation," *Proc. IEEE Int. Test Conf.*, 2018. submitted.

[22] K. Shamsi, M. Li, D. Z. Pan, and Y. Jin, "Cross-lock: Dense layout-level interconnect locking using cross-bar architectures," in *Proc. IEEE Great Lakes Symp. on VLSI*, 2018.

[23] C. Carlet, "Boolean functions for cryptography and error correcting codes," *Boolean models and methods in mathematics, computer science, and engineering*, vol. 2, pp. 257–397, 2010.

[24] M. Zaman, A. Sengupta, D. Liu, O. Sinanoglu, Y. Makris, and J. Rajendran, "Towards provably-secure performance locking," in *Proc. Design, Automation and Test in Europe*, pp. 1592–1597, 2018.

Nonvoltile Memory and Storage for CPS

Zhaoyan Shen[1], Zili Shao[2]
[1]Shandong University
[2]Chinese University of Hong Kong

1 Introduction

Nonvoltile Memory (NVM) is gaining great attention in both academia and industrial. Many NVMs have emerged, such as NAND Flash memory, Phase Change random access memory (PCM), Resistive random access memory (ReRAM), Magnetoresistive random access memory (MRAM), and Ferroelectric random access memory (FeRAM), each with its own peculiar properties and specific challenges. With advantages such as low latency, low power consumption, high density, and high resistance, many research papers have been proposed to adopt NVMs to cyper-physical systems to provide reliable storage system.

The most well studied NVM is NAND flash memory. An important goal of NAND flash development has been to reduce the cost per bit and to increase maximum chip capacity so that flash memory can compete with magnetic storage devices, such as hard disks. NAND flash has been widely adopted in embedded applications such as MMC or CF card flash memory, mobile devices including cellular phones and mp3 players, and many others. Compared with It has become an indispensable technology to partly bridge the gap between DRAM and storage performance.

As with flash memory a decade ago, NVMs are attracting a great deal of interest and much work is being conducted on the issue of how different technologies can be integrated in the memory hierarchy. The numerous announcements from different companies seeking to mass produce NVMs justifies the need to take a step back to discuss and classify the options for integration that have been investigated in state-of-the-art work. This paper surveys state-of-the-art work on improving the utilization of NVMs. Specially, we introduce the three types of NVM, namely NAND Flash, PCM, and ReRAM.

2 NAND Flash Memory

NAND flash memory is a type of EEPROM devices. A flash memory chip consists of multiple planes, each of which consists of thousands of blocks (a.k.a. erase blocks). A block is further divided into hundreds of pages. Each page has a data part (e.g., 4-16KB) and a spare area part (e.g., 128 bytes). Flash memory supports three main operations, namely read, write, and erase. Reads and writes are normally performed in units of pages. A read is typically fast (e.g., 50μs), while a write is relatively slow (e.g., 600μs). A unique constraint of NAND flash is that pages in a block must be written sequentially, and pages cannot be overwritten in place, meaning that once a page is programmed (written), it cannot be written again until the entire block is erased. An erase is typically slow (e.g., 5ms) and must be done in block granularity. A Flash Translation Layer (FTL) is used to emulate the Flash memory as a block device. It has three main components: address translator, garbage collector, and wear leveler.

FTL plays an important role in NAND flash management, and many studies for FTL have been conducted. FTL designs can be mainly categorized into three types [5]: page-level mapping [1], block-level mapping [7, 2, 12], and hybrid-level mapping [24, 20, 15, 23]. The page-level FTL records the mapping between logical page number and physical page in NAND Flash. It provides efficient address translation time, low garbage collection overhead, and high space utilization. However, it suffer from significant memory space requirement. Block-level FTL maps logical block number to a physical block number, which requires much less mapping information. However, in block-level FTL, a logical page can only be written to a physical page with the designated page offset within a physical block. Thus, block-level FTL is not as good as page-level FTL in terms of the flexibility and space utilization. To overcome the shortcomings of the above two mapping schemes, hybrid-level FTL is proposed to balance the space overheads and flexibility. The technique proposed in this paper is based on hybrid-level FTL.

With the development flash memory, the density of flash device keeps increasing. However, this also make the lifetime of flash memory much shorter. To prolong the flash lifespan, several techniques have been presented to

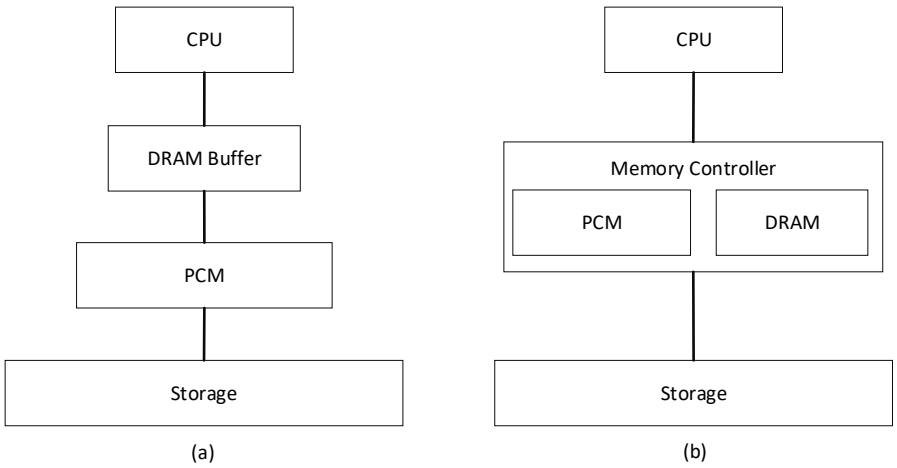


Figure 1: PCM/DRAM Memory Integration.

enhance the reliability [10, 22, 21, 9]. These approaches can provide good solutions to enhance the reliability of 2-D NAND flash, which could not be directly applied to 3-D flash. There have been several approaches to address the reliability issues of 3-D flash memory [11, 4]. Nevertheless, they focused primarily on the system structure and hardware implementation of 3-D flash. [25] is a good supplement for these approaches by helping them effectively reduce bit errors caused by program disturb and read disturb to further improve the reliability of 3-D flash memory.

Recently, there is a new trend of flash design, called open-channel SSD, which directly exposes the internal channels and its low-level flash details to the host [18, 14]. With open-channel SSD, the responsibility of flash management is shared between the host software and hardware device. Compared with conventional SSD design, open-channel SSD exposes its internal geometry details (e.g., the layout of channels, LUNs, and flash blocks) to software applications. Applications have the flexibility of scheduling I/O tasks among different channels to fully utilize the raw flash performance. Applications can directly operate the device hardware through the ioctl interface, which allows them to bypass many intermediate OS components, such as file system and the block I/O layer.

3 Phase Change Memory

Phase-change memory (PCM) has been intensively studied as a promising candidate main memory. Comparing with dynamic random access memory (DRAM), which has been widely used as the main memory for decades, PCM needs no refresh energy and consumes much lower leakage energy. In addition, PCM provides DRAM-like byte-addressable access and has the characteristics of non-volatility, low power, better scalability and higher density. As such, it has been regarded as potential alternative to replace DRAM to build main memory for energy optimization. Unfortunately, directly replace DRAM with PCM as main memory encounters the challenge of limited lifetime of PCM. For example, state-of-the-art process technology has demonstrated that the maximum writes number of PCM is around 10^8 to 10^9. In this section, we summarize the research works that extend the lifetime of PCM and the works that integrate PCM in the main memory subsystem

Extending the lifetime of PCM-enhanced memory systems has been one of the major focuses in recent years. PTL [17] proposes to add a translation layer so the constraints of PCM can be concealed for embedded systems to use PCM in a transparent manner. It designs an effective wear leveling technique that exploit application-specific features in embedded systems and periodically move hot areas of an application across the whole area in a PCM chip. Curling-PCM [13] proposes an effective application-specific wear leveling technique to evenly distribute write activities across the PCM chip so that the endurance of PCM-based embedded systems is enhanced. To further reduce the write traffic, Zhou et al. studied the memory lines written to PCM and observed that a significant portion of bits stay unchanged. They then proposed differential write that compares each bit to be written with the one in the PCM, and writes the device cell only if necessary [27].

As shown in Figure, there are two possible organizations when integrating PCM in the main memory subsystem.

The first approach is to utilize PCM as the main memory and make a small DRAM buffer between CPU and PCM to reduce the write traffic to PCM and improves the access latency of off chip misses [16]. Another alternative is to combine both DRAM and PCM as the hybrid memory, and the memory controller is in charge of allocating and reclaiming DRAM or PCM space. To improve the performance and lifetime, Dhiman et al. proposed a PCM-aware policy that dynamically monitors writes into each memory page and migrate write-intensive PCM pages to DRAM [6]. Zhang et al. reduced the frequent migration by adopting a modified Multi-Queue Algorithm [26] to categorize pages according to their hotness (i.e., the number of writes). Pages in highest ranked queues are placed in DRAM.

4 Resistive Random Access Memory

Metal-oxide resistive random access memory (ReRAM) is a kind of emerging non-volatile memory that can perform matrix-vector multiplication and sum operation efficiently in a crossbar structure. ReRAM has been widely studied to perform processing-in-memory (PIM) for several applications. With ReRAM-based PIM, data movement between memory and CPU can be eliminated, thus, releasing computational resource, saving energy, and reducing latency.

PRIME [3] proposes a novel PIM architecture to accelerate NN applications in ReRAM based main memory. In PRIME, a portion of ReRAM crossbar arrays can be configured as accelerators for NN applications or as normal memory for a larger memory space. RPBFS [8] utilizes ReRAM to accelerate graph traversal. In RPBFS, the ReRAM-based memory banks are separated into graph banks and master banks. The compressed adjacency lists are persistently mapped and scattered over multiple graph banks by an efficient mapping scheme. The master bank is selected for a graph to perform graph traversal through collaboration with graph banks. GRAPHR [19] follows the principle of near-data processing and explores the opportunity of performing massive parallel operations with low hardware and energy cost. GRAPHR divides ReRAM into memory ReRAM and graph engine (GE). The memory ReRAM stores the graph data in compressed sparse representation. GEs (ReRAM crossbars) perform the efficient matrix-vector multiplications on the sparse matrix representation. Re-Mining [] adopts ReRAM to accelerate the blockchain mining process. By eliminating data movement and providing high parallelism, ReRAM has significantly improved the performance of these applications. In the future, ReRAM would be applied to more other applications.

5 Conclusion

In this paper, we have introduced trhee very promising NVM technologies: NAND flash, PCM, and ReRAM, each of which has its pros and cons. Those NVMs are in different stages of development, with some are already being manufactured and others still in the prototype phase. NVM brings new research opportunities for optimizing computing systems. It should be an interesting direction to continuously exploit specific characteristics of NVM.

References

[1] Amir Ban. Flash file system, 1995. US Patent 5,404,485.

[2] Renhai Chen, Zhiwei Qin, Yi Wang, Duo Liu, Zili Shao, and Yong Guan. On-demand block-level address mapping in large-scale nand flash storage systems. *IEEE Transactions on Computers*, 64(6):1729–1741, 2015.

[3] Ping Chi, Shuangchen Li, Cong Xu, Tao Zhang, Jishen Zhao, Yongpan Liu, Yu Wang, and Yuan Xie. Prime: A novel processing-in-memory architecture for neural network computation in reram-based main memory. In *ACM SIGARCH Computer Architecture News*, volume 44, pages 27–39. IEEE Press, 2016.

[4] Won-seok Cho, Sun Il Shim, Jaehoon Jang, Hoo-sung Cho, Byoung-Koan You, Byoung-Keun Son, Ki-hyun Kim, Jae-Joo Shim, Choul-min Park, Jin-soo Lim, et al. Highly reliable vertical nand technology with biconcave shaped storage layer and leakage controllable offset structure. In *2010 Symposium on VLSI Technology*, pages 173–174. IEEE, 2010.

[5] Tae-Sun Chung, Dong-Joo Park, Sangwon Park, Dong-Ho Lee, Sang-Won Lee, and Ha-Joo Song. A survey of flash translation layer. *Journal of Systems Architecture*, 55(5-6):332–343, 2009.

[6] Gaurav Dhiman, Raid Ayoub, and Tajana Rosing. Pdram: A hybrid pram and dram main memory system. In *2009 46th ACM/IEEE Design Automation Conference*, pages 664–669. IEEE, 2009.

[7] Yong Guan, Guohui Wang, Chenlin Ma, Renhai Chen, Yi Wang, and Zili Shao. A block-level log-block management scheme for mlc nand flash memory storage systems. *IEEE Transactions on Computers*, 66(9):1464–1477, 2017.

[8] Lei Han, Zhaoyan Shen, Duo Liu, Zili Shao, H Howie Huang, and Tao Li. A novel reram-based processing-in-memory architecture for graph traversal. *ACM Transactions on Storage (TOS)*, 14(1):9, 2018.

[9] Pei-Han Hsu, Yuan-Hao Chang, Po-Chun Huang, Tei-Wei Kuo, and David Hung-Chang Du. A version-based strategy for reliability enhancement of flash file systems. In *Proceedings of the 48th Design Automation Conference*, pages 29–34. ACM, 2011.

[10] Min Huang, Zhaoqing Liu, Liyan Qiao, Yi Wang, and Zili Shao. An endurance-aware metadata allocation strategy for mlc nand flash memory storage systems. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 35(4):691–694, 2016.

[11] Jiyoung Kim, Augustin J Hong, Sung Min Kim, Kyeong-Sik Shin, Emil B Song, Yongha Hwang, Faxian Xiu, Kosmas Galatsis, Chi On Chui, Rob N Candler, et al. A stacked memory device on logic 3d technology for ultra-high-density data storage. *Nanotechnology*, 22(25):254006, 2011.

[12] Duo Liu, Tianzheng Wang, Yi Wang, Zhiwei Qin, and Zili Shao. A block-level flash memory management scheme for reducing write activities in pcm-based embedded systems. In *Proceedings of the Conference on Design, Automation and Test in Europe*, pages 1447–1450. EDA Consortium, 2012.

[13] Duo Liu, Tianzheng Wang, Yi Wang, Zili Shao, Qingfeng Zhuge, and Edwin Sha. Curling-pcm: Application-specific wear leveling for phase change memory based embedded systems. In *2013 18th Asia and South Pacific Design Automation Conference (ASP-DAC)*, pages 279–284. IEEE, 2013.

[14] Jian Ouyang, Shiding Lin, Song Jiang, Zhenyu Hou, Yong Wang, and Yuanzheng Wang. Sdf: software-defined flash for web-scale internet storage systems. In *ACM SIGPLAN Notices*, volume 49, pages 471–484. ACM, 2014.

[15] Zhiwei Qin, Yi Wang, Duo Liu, Zili Shao, and Yong Guan. Mnftl: An efficient flash translation layer for mlc nand flash memory storage systems. In *Proceedings of the 48th Design Automation Conference*, pages 17–22. ACM, 2011.

[16] Moinuddin K Qureshi, Vijayalakshmi Srinivasan, and Jude A Rivers. Scalable high performance main memory system using phase-change memory technology. In *ACM SIGARCH Computer Architecture News*, volume 37, pages 24–33. ACM, 2009.

[17] Zili Shao, Naehyuck Chang, and Nikil Dutt. Ptl: Pcm translation layer. In *2012 IEEE Computer Society Annual Symposium on VLSI*, pages 380–385. IEEE, 2012.

[18] Zhaoyan Shen, Feng Chen, Yichen Jia, and Zili Shao. Didacache: A deep integration of device and application for flash based key-value caching. In *15th {USENIX} Conference on File and Storage Technologies ({FAST} 17)*, pages 391–405, 2017.

[19] Linghao Song, Youwei Zhuo, Xuehai Qian, Hai Li, and Yiran Chen. Graphr: Accelerating graph processing using reram. In *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 531–543. IEEE, 2018.

[20] Tianzheng Wang, Duo Liu, Yi Wang, and Zili Shao. Ftl 2: a hybrid f lash t ranslation l ayer with logging for write reduction in flash memory. In *ACM SIGPLAN Notices*, volume 48, pages 91–100. ACM, 2013.

[21] Yi Wang, Luis Angel D Bathen, Nikil D Dutt, and Zili Shao. Meta-cure: A reliability enhancement strategy for metadata in nand flash memory storage systems. In *Proceedings of the 49th Annual Design Automation Conference*, pages 214–219. ACM, 2012.

[22] Yi Wang, Min Huang, Zili Shao, Henry CB Chan, Luis Angel D Bathen, and Nikil D Dutt. A reliability-aware address mapping strategy for nand flash memory storage systems. *IEEE Transactions on computer-aided design of integrated circuits and systems*, 33(11):1623–1631, 2014.

[23] Yi Wang, Duo Liu, Meng Wang, Zhiwei Qin, Zili Shao, and Yong Guan. Rnftl: A reuse-aware nand flash translation layer for flash memory. *ACM Sigplan Notices*, 45(4):163–172, 2010.

[24] Yi Wang, Zhiwei Qin, Renhai Chen, Zili Shao, Qixin Wang, Shuai Li, and Laurence T Yang. A real-time flash translation layer for nand flash memory storage systems. *IEEE Transactions on Multi-Scale Computing Systems*, 2(1):17–29, 2016.

[25] Yi Wang, Zili Shao, Henry CB Chan, Luis Angel D Bathen, and Nikil D Dutt. A reliability enhanced address mapping strategy for three-dimensional (3-d) nand flash memory. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 22(11):2402–2410, 2014.

[26] Wangyuan Zhang and Tao Li. Exploring phase change memory and 3d die-stacking for power/thermal friendly, fast and durable memory architectures. In *2009 18th International Conference on Parallel Architectures and Compilation Techniques*, pages 101–112. IEEE, 2009.

[27] Ping Zhou, Bo Zhao, Jun Yang, and Youtao Zhang. A durable and energy efficient main memory using phase change memory technology. In *ACM SIGARCH computer architecture news*, volume 37, pages 14–23. ACM, 2009.

3D Cooperative Mapping for Connected Ground and Aerial-Based Robots

Yachen Zhang, Long Chen
School of Data and Computer Science, Sun Yat-sen University

1 Introduction

Nowadays, the simultaneous localization and mapping (SLAM) approach [1] has been widely used for robot applications. In such a approach, a moving robot is expected to estimate its position and pose through repeatedly observing environmental features (e.g., corners, pillars, etc.) and to incrementally build the map according to the information of observed environments [2]. Due to the wide view field in horizontal directions (e.g., 360° for the full view) and the high precision in distance measurements, lidar sensors already become indispensable parts for both robot and autonomous driving systems. Normally, the performance of lidar-based SLAM approach depends on the quality of collected data and the efficiency of utilized algorithms. The former however further depends on employed lidar sensors. The high laser channel number not only provides huge amount of data, which can guarantee a high reconstruction degree of the environment, but also leads to increased computational burden on processing resources [3].

To improve the efficiency of 3D lidar construction, multiple robots for data collection have been considered in mapping of large scaled fields [4]. However, the problem is how to efficiently merge the maps built by each individual robot. In most cases, it is impossible to plan the robots' paths in advance and their initial locations are usually different [5]. In these cases, we need to search for similar route sections between each two robots to calculate their pose transform [6]. The similar route sections are related to the posture relationship for two robots. Therefore, the two local map for each single robot could be merged to a global map for whole environment [7]. In most instances, the multiple robots system is built on the ground to achieve cooperative mapping. However, ground-based multi-robot system is limited by the sensor¡¯s altitude, difficult to realize a top-down looking configuration [8]. The aerial-based robot could capture the top information of the buildings with 3D lidar. In this paper, we propose 3D cooperative mapping for connected ground and aerial-based robots. A complete and detailed map could be constructed by combining the data collected by ground and aerial-based robots.

2 Method Overview

In order to obtain a complete and detailed 3D map, we propose a 3D cooperative mapping for connected ground and aerial-based robots. For clarity yet without loss of generality, here the theory part is introduced with a simple use case of two robots. However, our method is not limited to such case and can be easily applied to scenarios with more robots. In the mentioned use case, we made following assumptions:

- Each robot is installed with a full view 3D lidar of the same sensor setup.
- The 3D lidar on the ground-based robot should be placed horizontally.
- The 3D lidar on the aerial-based robot should be placed vertically.
- The ground and aerial-based robots explore same environment.

The fourth assumption is necessary for merging the individual maps generated by each robot and can be guaranteed by pre-defining an overlap between the explored areas of both robots. Please note that such definition is very coarse, because the exact location and size of their common route segments are unknown. The ground and aerial-based robots will share the common segments to construct a complete and detailed 3D map for the same environment, which is illustrated in Fig. 1.

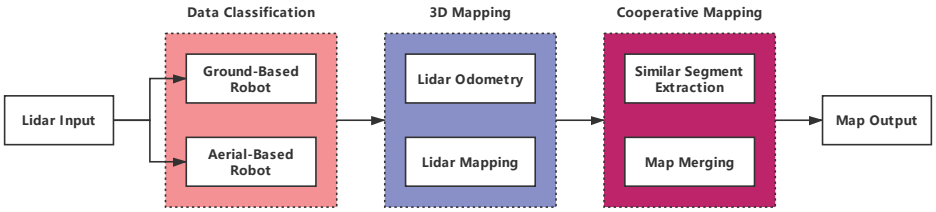


Figure 1: Method Overview

There are three parts in this method to transform the lidar data to a global map. For the same environment, the data will be classified to ground-based and aerial-based robots. Then, the 3D Mapping method will be used to get local map for each single robot. The local observation frames and global observation frames will be divided in this part. Next, the local and global observation frames will be used to search the similar segment to get the posture transformation. Finally, two single local maps will be merged to a global map through the posture transformation mentioned above.

3 3D Cooperative Mapping Model

In the proposed method, we utilize three parts to acquire final global map. In the same environment, the ground-based robot and aerial-based robot will collect the data from each own perspective. Due to the placement of each robot, the coordinate system of the data we acquired is also different. The lidar on the ground-based robot is placed horizontally and on the aerial-based robot is placed vertically. Therefore, the coordinate system for two robots is different. What we need to do is to unified there two coordinate systems by:

$$X_{k+1} = R_k X_k + T_k \tag{8}$$

Where X_{k+1} and X_k mean the coordinates for frame k and $k+1$. R_k and T_k represent the corresponding rotation and translation matrix for the posture transformation between frame k and $k+1$. When the two coordinate systems are unified, the 3D mapping section could be processed.

Since the sensors are the same, the same SLAM algorithm can be used for both ground and aerial-based robots. To get accurate 3D map for the environment, the method will be found on the KITTI vision benchmark suite [9]. After time and complexity consideration, V-LOAM was chosen as our 3D mapping algorithm [10]. Then, the local and global observation frames will be divided through **Lidar Odometry** and **Lidar Mapping**. These two types of observation frames will be used in next section to achieve cooperative mapping. After processed, there will be two types of observation frames and two single map for each robot.

The similar route segment is an unique contact between two independent robots during exploration of the environment. Thus, how to extract the similar route segment is the main mission in cooperative mapping tasks. The local observation frames mentioned before will be used to search the similar segment for ground and aerial-based robots. In this part, the point cloud matching algorithm is crucial for similar segment extraction to describe whether two frames are similar or not. Due to processing speed and accuracy, the normal distribution transform(NDT) algorithm is chosen [11]. Through iterative matching, the similar frame pairs will be achieved through selecting the most similar frames. Then a local map for each robot will be divided to two parts: similar segment and dissimilar segment. The global observation frames of similar segment from two single robot will construct a submap, which is part of the local map for ground and aerial-based robots. By matching two submaps, a posture transformation matrix will be achieved. Finally, the two single map for each robot will be merged to a complete and detailed map through the posture transformation mentioned.

4 Conclusions

A ground-based multi-robot system is limited by the sensor's altitude, difficult to realize a top-down looking configuration. Therefore, in this paper, we propose 3D cooperative mapping for connected ground and aerial-based robots. The ground-based robot will get the information from down looking configuration, where the aerial-based robot will get the information from top looking configuration. By merging maps constructed from each own perspective, a complete and detailed map will be achieved. There are still many problems that we need to explore for 3D cooperative mapping with ground and aerial-based robots in the future.

References

[1] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: part i," *IEEE robotics & automation magazine*, vol. 13, no. 2, pp. 99–110, 2006.

[2] J. Zhang and S. Singh, "Loam: Lidar odometry and mapping in real-time." in *Robotics: Science and Systems*, vol. 2, 2014, p. 9.

[3] S. Kohlbrecher, O. Von Stryk, J. Meyer, and U. Klingauf, "A flexible and scalable slam system with full 3d motion estimation," in *Safety, Security, and Rescue Robotics (SSRR), 2011 IEEE International Symposium on*. IEEE, 2011, pp. 155–160.

[4] S. Saeedi, M. Trentini, M. Seto, and H. Li, "Multiple-robot simultaneous localization and mapping: A review," *Journal of Field Robotics*, vol. 33, no. 1, pp. 3–46, 2016.

[5] B. Kim, M. Kaess, L. Fletcher, J. Leonard, A. Bachrach, N. Roy, and S. Teller, "Multiple relative pose graphs for robust cooperative mapping," 2010.

[6] J. W. Fenwick, P. M. Newman, and J. J. Leonard, "Cooperative concurrent mapping and localization," in *ICRA*. Citeseer, 2002, pp. 1810–1817.

[7] F. Amigoni, S. Gasparini, and M. Gini, "Merging partial maps without using odometry," in *Multi-Robot Systems. From Swarms to Intelligent Automata Volume III*. Springer, 2005, pp. 133–144.

[8] J. Zhang and S. Singh, "Aerial and ground-based collaborative mapping: an experimental study," in *Field and Service Robotics*. Springer, 2018, pp. 397–412.

[9] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[10] J. Zhang and S. Singh, "Visual-lidar odometry and mapping: Low drift, robust, and fast," in *IEEE International Conference on Robotics and Automation(ICRA)*, Seattle, WA, May 2015.

[11] P. Biber and W. Straßer, "The normal distributions transform: A new approach to laser scan matching," in *IROS*, vol. 3, 2003, pp. 2743–2748.

1 Conferences and Workshops

- IEEE International Conference on Embedded Software and Systems (ICESS) 2019
- Asian Hardware Oriented Security and Trust Symposium (AsianHOST) 2018
- IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA) 2018
- IEEE Non-Volatile Memory Systems and Applications Symposium (NVMSA) 2018

2 Special Issues in Academic Journals

- ACM Transactions on Cyber-Physical Systems (TCPS) special issue on Human-Interaction-Aware Data Analytics for Cyber-Physical Systems

Newsletter of Technical Committee on Cyber-Physical Systems (IEEE Systems Council)

The newsletter of Technical Committee on Cyber-Physical Systems (TC-CPS) aims to provide timely updates on technologies, educations and opportunities in the field of cyber-physical systems (CPS). The letter will be published twice a year: one issue in February and the other issue in October. We are soliciting contributions to the newsletter. Topics of interest include (but are not limited to):

- Embedded system design for CPS
- Real-time system design and scheduling for CPS
- Distributed computing and control for CPS
- Resilient and robust system design for CPS
- Security issues for CPS
- Formal methods for modeling and verification of CPS
- Emerging applications such as automotive system, smart energy system, internet of things, biomedical device, etc.

Please directly contact the editors and/or associate editors by email to submit your contributions.

Submission Deadline:

All contributions must be submitted by **Jul. 1st, 2019** in order to be included in the February issue of the newsletter.

Editors:

- Bei Yu, Chinese University of Hong Kong, Hong Kong, byu@cse.cuhk.edu.hk

Associate Editors:

- Long Chen, Sun Yat-Sen University, China, chenl46@mail.sysu.edu.cn
- Wuling Huang, Chinese Academy of Science, wuling.huang@ia.ac.cn
- Yier Jin, University of Florida, USA, yier.jin@ece.ufl.edu
- Abhishek Murthy, Philips Lighting Research, USA abhishek.murthy@philips.com
- Rajiv Ranjan, Newcastle University, United Kingdom, raj.ranjan@ncl.ac.uk
- Muhammad Shafique, Vienna University of Technology, Austria, mshafique@ecs.tuwien.ac.at
- Yiyu Shi, University of Notre Dame, USA, yshi4@nd.edu
- Ming-Chang Yang, Chinese University of Hong Kong, Hong Kong, mcyang@cse.cuhk.edu.hk