SMCGen: Generating Reconfigurable Design for Sequential Monte Carlo Applications

Thomas C.P. Chau*, Maciej Kurek*,
James S. Targett*, Jake Humphrey†, George Skouroupathis*, Alison Eeλ†, Jan Maciejowski‡,
Benjamin Cope*, Kathryn Cobden*, Philip Leong*, Peter Y.K. Cheung†, Wayne Luk*

*Department of Computing, †Department of Electrical and Electronic Engineering, Imperial College London, UK
‡Department of Engineering, University of Cambridge, UK
§School of Electrical and Information Engineering, University of Sydney, Australia
¶Altera Europe Limited

{c.chau10,mk306,james.targett10,jake.humphrey11,georgios.skouroupathis11,p.cheung,w.luk}@imperial.ac.uk
{aje46,jmm1}@cam.ac.uk philip.leong@sydney.edu.au {bcope,kcobden}@altera.com

Abstract—The Sequential Monte Carlo (SMC) method is a simulation-based approach to compute posterior distributions. SMC methods often work well on applications considered intractable by other methods due to high dimensionality, but they are computationally demanding. While SMC has been implemented efficiently on FPGAs, design productivity remains a challenge. This paper introduces a design flow for generating efficient implementation of reconfigurable SMC designs. Using it, users can develop efficient multiple-FPGA SMC applications without any knowledge of reconfigurable computing. The proposed design flow consists of a parametrisable SMC computation engine, and an open-source software template which enables efficient mapping of a variety of SMC designs to reconfigurable hardware. Design parameters that are critical to the performance and to the solution quality are tuned using a machine learning algorithm based on surrogate modelling. Experimental results on 3 case studies show that design performance is substantially improved after parameter optimisation. The proposed design flow demonstrates its capability of producing reconfigurable implementations for a range of SMC applications that have significant improvement in speed and in energy efficiency over optimised CPU and GPU implementations.

Keywords—Sequential Monte Carlo; Machine Learning; FPGA

I. INTRODUCTION

Sequential Monte Carlo (SMC) methods are a set of online posterior density estimation algorithms that perform inference of unknown quantities from observations. The observations arrive sequentially in time and the inference is performed on-line. A common application is in the guidance, navigation and control of vehicles, particularly mobile robots [1] and aircraft [2]. For these applications, it is critical that high sampling rates can be handled in real-time. SMC methods also have applications in economics and finance [3] where minimising latency is crucial.

SMC methods are often preferable over Kalman filters and hidden Markov models, as they do not require exact analytical expressions to compute the evolving sequence of posterior distributions. Moreover, they can model high-dimensional data using non-linear dynamics and constraints, are parallelisable, and can greatly benefit from hardware acceleration. Acceleration of SMC methods has been studied in applications such as air traffic management [4, 5], robot localisation [6], object tracking [7] and signal processing [8].

Although earlier publications in [4, 6, 7, 8] attempted to implement SMC efficiently on FPGAs, design productivity remains a challenge. Firstly, while different sets of SMC parameters produce the same accuracy, they have very different computational complexity. For example, the performance of SMC relies on the a set of random samples, which are called particles in the following. The more complex the problem, the larger the number of particles needed. Using excessive numbers of particles unfortunately causes prohibitive run-time without increasing solution accuracy. The parameter space spans multiple dimensions and the objective function can be non-convex, making exhaustive optimisation impractical. Secondly, customising designs for different SMC applications requires tremendous effort.

In this paper, we propose a SMC design flow for reconfigurable hardware. A computation engine captures the generic control structure shared among all SMC applications. A framework for mapping software to hardware is derived, so users can specify application-specific features which are automatically converted to efficient hardware. Timing model relates design parameters to performance constraints. To enable rapid learning of a large design space, a machine learning algorithm is used to automatically deduce characteristics of the design space.

The contributions of this paper are as follows:
• A design flow to reduce the development effort of SMC applications on reconfigurable systems (Section III). Users can design efficient, multiple-FPGA SMC applications for arbitrary problems without any knowledge of reconfigurable computing, and the software template is open-source.

1Available online: http://thomascep.github.io/smcgen
A machine learning approach that explores the SMC design space automatically and tunes design parameters to improve performance and accuracy (Section IV). The resulting parameters can be applied to the hardware design at run-time without the need for resynthesis. It is demonstrated that parameter optimisation enables the design space to be explored an order of magnitude faster without sacrificing quality. Compared with previous work [4, 6], we have achieved better quality of solutions and faster designs.

The utility of this approach in terms of design productivity and performance is quantified over a diverse set of SMC problems. Three applications are implemented on Altera and Xilinx-based reconfigurable platforms, with varying numbers of FPGAs. For these problems, the number of lines of code for the FPGA implementation is reduced by approximately 76%, and significant speedup and energy improvement over CPU and GPU implementations (Section V) are demonstrated.

II. BACKGROUND AND RELATED WORK

A. SMC Methods

SMC methods estimate the unobserved states of interest based on the given observations [9]. The target posterior density \( p(s_t|m_t) \) is represented by a set of particles, where \( s_t \) is the state and \( m_t \) is the observation at time step \( t \). A sequential importance resampling (SIR) algorithm [10] is used to obtain a weighted set of \( N_p \) particles \( \{s^{(i)}_t\}_{i=1}^{N_p} \). The importance weights \( \{w_t^{(i)}\}_{i=1}^{N_p} \) are approximations to the relative posterior probabilities of the particles such that \( \sum_{i=1}^{N_p} w_t^{(i)} = 1 \). This process is described in Algorithm 1 and involves 5 stages of computation:

Algorithm 1 SMC methods

1: for each time step \( t \) do
2: \( \text{idx}1 \leftarrow 0 \)
3: Initialisation
4: while \( \text{idx}1 < \text{itl} \_\text{outer} \) do
5: \( \text{idx}2 \leftarrow 0 \)
6: \( \text{itl} \_\text{inner} \leftarrow 3 + 5 \exp(\frac{5 \times \text{idx}1}{\text{itl} \_\text{outer}}) \)
7: for each particle \( p \) do
8: while \( \text{idx}2 < \text{itl} \_\text{inner} \) do
9: Sampling
10: Importance weighting
11: \( \text{idx}2 \leftarrow \text{idx}2 + 1 \)
12: end while
13: end for
14: \( \text{idx}1 \leftarrow \text{idx}1 + 1 \)
15: if \( \text{idx}1 \leq \text{itl} \_\text{inner} \) then
16: Resampling
17: end if
18: end while
19: Update
20: end for

1) Initialisation: Weights \( \{w_t^{(i)}\}_{i=1}^{N_p} \) are set to \( \frac{1}{N_p} \).
2) Sampling: Next states \( \{s_t^{(i)}\}_{i=1}^{N_p} \) are computed based on the current state \( \{s_{t-1}^{(i)}\}_{i=1}^{N_p} \).
3) Importance weighting: Weight \( \{w_t^{(i)}\}_{i=1}^{N_p} \) is updated based on a score function which accounts for the likelihood of particles fitting the observation. Within each iteration \( \text{idx}1 \), the sampling and importance weighting stages are iterated \( \text{itl} \_\text{inner} \) times so that those particles with sustained benefits are assigned higher weights. As \( \text{idx}1 \) increases, the set of particles reflects a more accurate approximation, so \( \text{itl} \_\text{inner} \) is increased exponentially.
4) Resampling: By removing the particles with small weights and replicating those with large weights \( \text{itl} \_\text{outer} \) times in a time step, the problem of degeneracy is addressed [11]. Without this step, only a small number of particles will have substantial weights for inference.
5) Update: State \( s_t \) is obtained from the resampled particle set \( \{s_t^{(i)}\}_{i=1}^{N_p} \) via weighted average or more complicated functions that will be shown below.

Table I summarises the parameters of the SMC methods described in Section II-A.

B. SMC Applications

1) Stochastic Volatility: These models are used extensively in mathematical finance [12, 13], and describe volatility as a stochastic process which better reflects the behaviour of many financial instruments but are computationally expensive. In this work, the sampling function shown in Equation 1 is employed, where \( y_t \) is the observable time varying volatility and \( s_t \) represents the stochastic log-volatility process. \( \beta \) and \( \phi \) are empirical constants.

\[
y_t = \beta \exp(s_t/2)\epsilon_t, \epsilon_t \sim \mathcal{N}(0,1) \\
s_t = \phi s_{t-1} + \mathcal{N}(0,1)
\]

(1)

The sampling function in Equation 2 is implied by Equation 1. The state transition from \( s_{t-1} \) to \( s_t \) is used to draw random samples \( s_t \) from the existing pool of particles.

\[
s_t \sim \mathcal{N}(\phi s_{t-1}, 1)
\]

(2)

2) Robot Localisation: SMC methods are applied to mobile robot localisation [1], and this application is used as an example throughout the paper. At regular time intervals, a robot obtains sensor values, identifies its location and commits a move. The robot needs to be aware of the locations of other moving objects in the environment.

The sampling stage is described by Equations 3 and 4. The robot estimates its updated state \( s_t \) based on the current
known location \((x, y)\) and heading \(h\). State is affected by external reference status \(r_t\) which contains displacement \(\delta\) and rotation \(\gamma\). Importance weighting is used to calculate the likelihood of a location based on the observation, i.e. the sensor values.

\[
\begin{align*}
(s_t^i) &= \begin{pmatrix} x_t^i \\ y_t^i \\ h_t^i \end{pmatrix} = \begin{pmatrix} x_{t-1}^i + \delta_t^i \cos(h_{t-1}^i) \\ y_{t-1}^i + \delta_t^i \sin(h_{t-1}^i) \\ h_{t-1}^i + \gamma_t^i \end{pmatrix} \\
(r_t^i) &= \begin{pmatrix} \delta_t^i \\ \gamma_t^i \end{pmatrix} = \begin{pmatrix} N(\delta_t, \sigma_\delta^2) \\ N(\gamma_t, \sigma_\gamma^2) \end{pmatrix}
\end{align*}
\]

(3) Air Traffic Management: SMC methods are applied to model predictive control (MPC) optimisation where control actions at discrete time intervals are determined to minimise an error criteria [2]. An example is air traffic management which avoids dangerous encounters by maintaining safe separation distances between aircraft.

At each sampling instant, the control sequence over a number of future time steps, called the prediction horizon \(H\), is estimated. A state is a set of control sequences \(\{s_t^i, \delta, ..., H-1\}_{t=1}^{N_p}\) being picked within a permitted range and applied to the current reference status \(r_{t-1}\) to compute the future set of reference statuses \(\{s_t^i, \delta, ..., H-1\}_{t=1}^{N_p}\). During importance weighting, a score function evaluates the quality of estimation for each particle, and weights the product of scores over the horizon. If any particle violates any constraint, its weight is set to zero. The first control \(s_t^0\) in the sequence, is obtained by selecting the best one among \(\{s_t^0, \delta, ..., H-1\}_{t=1}^{N_p}\). Then the selected control is committed to form reference \(r_t\).

Equation 5 illustrates a control tuple that consists of roll angle \(\phi\); pitch angle \(\tau\); and thrust \(T\). Equation 6 shows a reference that consists of the current position in 3 dimensional space \((x, y, a)\), heading angle \(\chi\), air speed \(V\) and mass \(M\). For more details of the model, refer to [5].

\[
\begin{align*}
(s_t^i) &= \begin{pmatrix} x_t^i \\ y_t^i \\ h_t^i \end{pmatrix} = \begin{pmatrix} \chi_t - 1 + V_{t-1} \cos(\chi_{t-1}) \cos(\tau_t^i) \\ y_{t-1} - V_{t-1} \sin(\chi_{t-1}) \cos(\tau_t^i) \\ \alpha_{t-1} + V_{t-1} \sin(\tau_t^i) \end{pmatrix} \\
(r_t^i) &= \begin{pmatrix} \delta_t^i \\ \gamma_t^i \end{pmatrix} = \begin{pmatrix} -\sin(\tau_t^i) \\ \sin(\tau_t^i) \end{pmatrix}
\end{align*}
\]

(5) (6)

III. SMC DESIGN FLOW

This section introduces a design flow for generating reconfigurable SMC designs. The design flow has 2 novel features to minimise hardware redesign efforts: (1) A generic high-level mapping where application-specific features are specified in a software template and automatically converted to hardware. The template supports the parameter optimisation described in Section IV. (2) A parameterisable SMC computation engine which is made up of customisable building blocks and generic control structure that maximises design reuse.

![Figure 1. Design flow for SMC applications. Users only customise the application-specific descriptions inside the dotted box.](image)

Figure 1 shows the proposed design flow. Starting with a functional specification such as software codes or mathematical descriptions, the users identify and code application-specific descriptions (Section III-A). The design flow automatically weaves these descriptions with the computation engine (Section III-B) to form a complete multiple-FPGA system. In this work the synthesis tool employed was Maxeler’s MaxCompiler, which uses Java as the underlying language. MaxCompiler also supports a list of FPGAs such that low level configurations, such as I/O binding, are performed automatically.

A. Specifying Application Features

Users create a new SMC design by customising the application-specific Java descriptions inside the dotted box of Figure 1. These descriptions correspond to Def (Code 1), FPGA Func (Code 2) and CPU Func.

Code 1 illustrates the class Def where number representation (floating-point, fixed-point with different bit-width), structs (state, reference), static parameters (Table I) and system parameters are defined. Users are allowed to customise number representation to benefit from the flexibility of FPGA and make trade-off between accuracy and design complexity. State and reference structs determine the I/O interface. Static parameters are defined in this class, while dynamic parameters are provided at run-time. System parameters define device-specific properties such as clock speed and parallelism.

**FPGA Func:** Sampling and importance weighting (line 9 and 10 of Algorithm 1) are the most computation intensive functions, and accelerated by FPGAs. Code 2 illustrates how these two FPGA functions are defined. Given current state \(s_{\text{in}}\), reference \(r_{\text{in}}\) and observation \(m_{\text{in}}\) (sensor value in
this example), an estimation state \( s_{out} \) is computed. Weight \( w \) accounts for the probability of an observation from the estimated state. The weight is calculated from the product of scores over the horizon. In this example, the weight is equal to the score as the horizon length is only 1.

**CPU Func:** *Initialisation and update* are functions running on the CPU. They are responsible for obtaining and formatting data and displaying results. *Resampling* is independent of applications so users need not to customise it.

\[
\text{public class Def} \\
\text{static final DFETYPE float_t = KernelLib.dfeFloat(8,24);} \\
\text{static final DFETYPE fixed_t = KernelLib.dfeFixOffset(26,-20,SignMode.TWOSCOMPLEMENT);} \\
// State Struct \\
\text{public static final DFETYPE state_t = new} \\
\text{DFETYPE{}} \\
\text{new StructFieldType(’x’, compType);} \\
\text{new StructFieldType(’y’, compType);} \\
\text{new StructFieldType(’h’, compType);}} \\
// Reference Struct \\
\text{public static final DFETYPE ref_t = new} \\
\text{DFETYPE{}} \\
\text{new StructFieldType(’d’, compType);} \\
\text{new StructFieldType(’r’, compType);}} \\
// Static Design parameters (Table I) \\
\text{public static int NWall = 8, NSensor = 20;} \\
\text{public static int FPGA_resampling = 0, Use_DRAM = 0;} \\
\text{public static int H=1 ,N A=1 ;} \\
// System Parameters \\
\text{public static int NC_inner = 1, NC_P = 2;} \\
\text{public static int CLK_core = 120, CLK_mem = 350;} \\
\text{public static int FPGA_resampling = 0, Use_DRAM = 0;} \\
\text{public static int NWall = 8, NSensor = 20;} \\
\}

**Code 1:** State, control and parameters for the robot localisation example.

\[
\text{public class Func} \\
\text{public static DFETYPE sampling{} \\
\text{DFETYPE s_in, DFETYPE s_out{}} \\
\text{s_out = state_t.newInstance(this);} \\
\text{s_out.x = s_in.x + nrand(c_in.d,S0.5) \times \cos(s_in.h);} \\
\text{s_out.y = s_in.y + nrand(c_in.d,S0.5) \times \sin(s_in.h);} \\
\text{s_out.h = s_in.h + nrand(c_in.r,S0.1);} \\
\text{}} \\
\text{return s_out;}} \\
\text{public static DFETYPE weighting{}} \\
\text{DFETYPE s_in, DFETYPE sensor{}} \\
\text{DFETYPE score = exp(-1^powl(est(s_in)-sensor,2)/S0.5);} \\
\text{// Constraint handling} \\
\text{bool succeed = est(s_in)>0 ? true : false;} \\
\text{// Weight accumulation} \\
\text{DFETYPE w = succeed ? score : 0; //weight} \\
\text{return w;} \\
\}
\]

**Code 2:** FPGA functions (Sampling and importance weighting) for the robot localisation example.

### B. Computation Engine Design

To allow customisation of the computation engine, the engine and data structure are designed as shown in Figure 2(a) and 2(b) respectively. The computation engine employs a heterogeneous structure that consists of multiple FPGAs and CPUs. FPGAs are responsible for sampling, importance weighting and optionally resampling index generation, and fully pipelined to maximise throughput. To exploit parallelism, particle simulations (sampling and importance weighting) are computed simultaneously by every processing core on each FPGA. Processing cores can be replicated as many times as FPGA resources allow. In situations where the computed results have to be grouped together, data are transferred among FPGAs via the inter-FPGA connection. To maximise the system throughput, remaining non-compute-intensive tasks that involve random and non-sequential data accesses are performed on the CPUs. FPGAs and CPUs communicate through high bandwidth connections such as PCI Express or InfiniBand.

From the control paths (dotted lines) of Figure 2(a), we see that there are 3 loops matching Algorithm 1: (1) inner, (2) outer, and (3) time step. First, the inner loop iterates \( \text{itl}_{\text{inner}} \) number of times for sampling and importance weighting, \( \text{itl}_{\text{inner}} \) increases with the iteration count of the outer loop. Second, the outer loop iterates \( \text{itl}_{\text{outer}} \) times to do resampling. The resampling process is performed \( \text{itl}_{\text{outer}} \) times to refine the pool of particles. The particle
The data will be used by the kernel the next state on the prediction horizon. After getting a state as a reference one each cycle and calculates the next state on the prediction horizon. The reference and weight streams have information of \( N_A \) agents in \( N_P \) particles.

Changing the values of \( \text{itl}_\text{outer}, \text{itl}_\text{inner} \) and \( N_P \) at run-time is allowed since they only affect the length of the state streams, and not the hardware data path. The computation engine is fully pipelined and outputs 1 result per clock cycle.

Figure 3 shows the design of the FPGA kernel. Blocks that require customisation are darkened. The sampling and importance weighting process can be accelerated using multiple cores, in which each of them is responsible for a sub-part of the inner loop iterations or particles. \( N_C \) represents the number of processing cores being used on one FPGA, and \( N_{\text{Board}} \) is the number of FPGA boards being used. \( \min(1, \frac{\text{bandwidth}}{\text{size}(\text{state}) \cdot \text{freq}}) \) accounts for the limitation of bandwidth between FPGAs and CPUs.

\[
T_{\text{step}} = \text{itl}_\text{outer} \cdot (T_{\text{skei}} + T_{\text{resample}} + T_{\text{cpu}} + T_{\text{transfer}})
\]

(7)

\( T_{\text{skei}} \) is the time spent on sampling and importance weighting in the FPGA kernels. Since the data is organised as a stream as described in Section III-B, the time spent on sampling and importance weighting is linear with \( N_P, N_A \), and \( H \). It is iterated \( \text{itl}_\text{inner} \) times in the inner loop. The sampling and importance weighting process can be accelerated using multiple cores, in which each of them is responsible for a sub-part of the inner loop iterations or particles. \( N_C \) represents the number of processing cores being used on one FPGA, and \( N_{\text{Board}} \) is the number of FPGA boards being used. \( \min(1, \frac{\text{bandwidth}}{\text{size}(\text{state}) \cdot \text{freq}}) \) accounts for the limitation of bandwidth between FPGAs and CPUs.

\[
T_{\text{skei}} = \frac{\text{itl}_\text{inner} \cdot N_P \cdot N_A \cdot H}{N_C \cdot N_{\text{Board}} \cdot \text{freq}} \cdot \min\left(1, \frac{\text{bandwidth}}{\text{size}(\text{state}) \cdot \text{freq}}\right)
\]

(8)

\( T_{\text{resample}} \) is the time spent on generating the resampling indices. It takes \( N_P \cdot PW + N_P \cdot N_A \) cycles to generate the cumulative probability distribution function, and a further \( 3 \cdot PL \cdot N_P \) cycles to generate particle indices. \( PW \) and \( PL \) are the length of the pipelines. \( T_{\text{resample}} \) can be omitted if resampling is processed by the CPUs.
\[ T_{\text{resample}} = \frac{N_P \cdot PW + N_P \cdot N_A + 3 \cdot PL \cdot N_P}{\text{freq}} \]

\[ T_{\text{cpu}} = \alpha_1 \cdot H \cdot N_P \cdot N_A \]

\[ T_{\text{transfer}} = \frac{N_P \cdot N_A \cdot (H \cdot \text{sizeof(state)})}{\text{bandwidth}} \]

IV. Optimising SMC Computation Engine

The design parameters in Table I have great impact on the performance. 3 questions manifest when finding optimised customisation of the engine: (1) Which sets of parameters have the best accuracy? (2) For the same accuracy, which sets of parameters meet the timing requirement? (3) How can we reduce the design parameter exploration time?

A. Effect of the Design Parameters

Referring to Table I, the SMC computation engine has up to 6 design parameters, each of which adds a dimension to the design space. It is ineffective to exhaustively search for the best set of parameters. Furthermore, the performance curve of each dimension can be non-linear and constrained by the real-time requirement and FPGA resources.

To answer questions 1 and 2, consider the robot localisation application. Its solution quality is measured by the root mean square error (RMSE) in localisation. We study the effect of changing design parameters using the functional specification in Figure 1, e.g. a C program. Its fast build time helps us to perform analysis effectively but its performance is too slow for real-time operation. The timing model described in Section III-C estimates the run-time of the FPGA implementation.

When \( N_P \) and \( \text{itl}_{\text{outer}} \) are explored together as shown in Figure 4, we see an uneven surface. Although non-linear, the trend of RMSE decreasing as \( N_P \) and \( \text{itl}_{\text{outer}} \) are increased is evident. The valid parameter space is constrained by the real-time requirement. The parameter space is darkened for those parameters leading to an RMSE greater than 1 m (Question 1). Moreover, the dark region with a run-time longer than the 5 seconds real-time requirement is marked as invalid (Question 2).

If the value of \( S \) is considered, the parameter optimisation problem expands to 3 dimensions as shown in Equation 12.

\[ \text{minimise } \text{RMSE} = f(N_P, \text{itl}_{\text{outer}}, S) \]

\[ \text{subject to } \text{RMSE} \leq 1 \text{ m, } T_{\text{step}} \leq 5s, \]

B. Parameter Optimisation

Now we come to question 3, the parameter optimisation problem, which is difficult as construction of an analytical model combining timing and quality of solution is either impossible or very time consuming. Furthermore the design space is constrained by multiple accuracy and real-time requirements. We cannot use a design unless the results are within certain error bound. The problem is further aggravated by the curse of dimensionality. We use an automated design exploration approach which is facilitated by a machine learning algorithm developed in [16]. The approach allows the performance impact of different parameters to be determined for any design based on our SMC computation engine.

A surrogate model is employed to enable rapid learning of the valid design space and deal with a large number of parameters. The idea is illustrated in Figure 5. Firstly, a number of randomly sampled designs is evaluated (Figure 5(a)). Secondly, the results obtained during evaluations are used to build a surrogate model. The model provides a regression of a fitness function and identifies regions of the parameter space which fail any of the constraints (Figure 5(b)). Thirdly, the surrogate model output is used to calculate the expected improvement (Figure 5(c)). Finally, the exploration converges to the parameter set that is expected to offer the highest improvement. Parameter sets in the invalid region are disqualified (Figure 5(d)).

Our SMC computation engine is made customisable so the design productivity of FPGA is increased from the optimisation approach which is already applicable to CPUs and GPUs.

V. Evaluation

A. Design Productivity

We first analyse how the proposed design flow can reduce design effort. In Table II, user-customisable code is classified into three parts: (a) Def is the definition of state, reference and parameters. (b) FPGA Func is the description of sampling and importance weighting functions. (c) CPU Func is the initiation, resampling and update part running on CPU. On average, users only need to customise 24% of the source code. Moreover, automatic design space optimisation greatly saves overall design time. As we will see in the applications.
below, we are able to choose the optimal set of parameters without doing exhaustive search.

![Diagram](image)

**Figure 5.** Illustration of automatic parameter optimisation: (a) Sampling parameter sets; (b) Building surrogate model; (c) Calculating expected improvement; (d) Moving to the point offering the highest improvement.

### B. Application 1: Stochastic Volatility

The design flow is first evaluated using a small-sized application on a microATX computer. The stochastic volatility model was described using the flow and targeted to a Xilinx Virtex-6 XC6VSX475T FPGA at 150 MHz. Parallel single precision floating-point data paths are used to maximise resource utilisation and hence performance. Limited by I/O constraints, 16 processing cores are chosen. The resulting design uses 70,674 LUTs (24%), 448 DSPs (22%) and 394 block RAMs (19%). The CPU is an Intel Core i7 870 quad-core processor clocked at 2.93GHz.

The design space has 2 dimensions, \( N_P \) and \( S \). Out of 420 sets of design parameters, the machine learning approach evaluates 20 of the candidates, and obtains an optimal set of parameters \( N_P = 708, S = 1.5 \) which minimises the estimation error.

Table III summarises the performance of CPU and reconfigurable systems using the same set of tuned parameters. Both systems have the same microATX form factor for fair comparison. Since the data size being processed is very small, the processing time of reconfigurable system is dominated by the overhead of invoking the FPGA kernel.

### C. Application 2: Mobile Robot Localisation

Now we look at an application with larger data set. For this example the same reconfigurable system as application 1 is used. Two processing cores are instantiated in an FPGA. Core computation in the sampling and importance weighting process is implemented using fixed-point arithmetic to optimise resource usage. The result utilises 148,431 LUTs (50%), 1,278 DSPs (63%) and 549 block RAMs (26%).

The design space has 3 dimensions: \( t_{il\_outer} \), \( N_P \) and \( S \). Out of 945 sets of parameters, 52 sets are evaluated to minimise the localisation error within the 5 seconds real-time constraint.

Table IV compares the performance of our reconfigurable system with CPU, GPU and a previous system in [6] which has not been optimised by our proposed approach. The reconfigurable system is 8.9 times and 1.2 times faster than the CPU and GPU, respectively. With parameter optimisation that maximise accuracy, our work achieves a better RMSE than the previous work (0.15m vs. 0.52m).

### D. Application 3: Air Traffic Management

The air traffic management system is able to control 20 aircraft simultaneously. The FPGA part runs on a 1U machine hosting 6 Altera Stratix V GS 5SGSD8 FPGAs clocked at 220 MHz, each of which has a single precision floating-point data path that consumes 166,008 LUTs (63%), 337 multipliers (9%) and 1,528 block RAMs (60%). The CPU part runs on 2 Intel Xeon E5-2640 CPUs clocked at 2.53GHz. Both parts are connected via InfiniBand.

This application has 4 design parameters leading to a space with 4000 sets of parameters. The optimisation target is to minimise the time of aircraft spending in the air traffic control region. Machine learning reduces the number of evaluations to 1% as indicated in Table V. Hence, the parameter optimisation time is reduced from days to hours.
Table V

<table>
<thead>
<tr>
<th>Parameter set obtained</th>
<th>Parameter sets evaluated / total</th>
</tr>
</thead>
<tbody>
<tr>
<td>all_gates</td>
<td>41 / 4000</td>
</tr>
<tr>
<td></td>
<td>20 / 5000</td>
</tr>
</tbody>
</table>

Table VI

<table>
<thead>
<tr>
<th>CPU</th>
<th>GPU</th>
<th>This work</th>
<th>Ref. FPGA [4]</th>
</tr>
</thead>
<tbody>
<tr>
<td>opt.</td>
<td>opt.</td>
<td>w/o opt.</td>
<td>w/o opt.</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Clik. freg. (MHz)</td>
<td>2.660</td>
<td>1.50</td>
<td>220</td>
</tr>
<tr>
<td>Number of cores</td>
<td>550</td>
<td>1100</td>
<td>600</td>
</tr>
<tr>
<td>Power (W)</td>
<td>198</td>
<td>28.25</td>
<td>11.6</td>
</tr>
</tbody>
</table>

Table VI summarises the performance of the CPU, GPU, and reconfigurable system. To ensure fair comparisons, we scale the CPU and GPU systems to similar form factors with the reconfigurable system. The scaling is based on the fact that the sampling and importance weighting process is evenly distributed to every GPU and computed independently, while the resampling process is computed on the CPU no matter how many GPUs are used. The reconfigurable platform is faster and more energy efficient than the other systems.

We also compare the performance of our work with a reference implementation that uses an Altera Stratix IV FPGA [4]. That implementation is only large enough to support 4 aircraft and it does not have the flexibility to tune parameters without re-compilation. Our design exploration approach is able to select the set of parameters that produces the same quality of results and is up to 73 times faster.

VI. CONCLUSION

This paper demonstrated the feasibility of generating highly-optimised reconfigurable designs for SMC applications, while hiding all aspects of reconfigurable computing from the user. A software template makes the computation engine portable and facilitates code reuse, the number of lines of user-written code being decreased by approximately 76% for an application. We further established that a surrogate software model combined with machine learning can be used to rapidly optimise designs, reducing optimisation time from days to hours; and that the resulting parameters can be utilised without resynthesis.

Ongoing and future work is focused on incorporating device-specific parameters, such as the level of parallelism and clock speed, into the machine learning approach [16]. We are currently investigating run-time optimisation of parameters based on our initial work in [6]. We will also automate the design flow to allow translation of popular software programming languages (e.g. R, MATLAB) directly to reconfigurable designs.

ACKNOWLEDGMENTS

This work is supported in part by the European Union Seventh Framework Programme under grant agreement number 257906, 287804 and 318521, by UK EPSRC, by the Australian Research Council under Linkage Project LP110200413, by Maxeler University Programme, by HiPEAC NoE, by Altera, by Xilinx, and by the Croucher Foundation.

REFERENCES