Knowing The Spec to Explore The Design via Transformed Bayesian Optimization

Donger Luo*1, Qi Sun*2, Xinheng Li1, Chen Bai3, Bei Yu3, Hao Geng1,4

1ShanghaiTech University  2Zhejiang University
3Chinese University of Hong Kong  4Shanghai Engineering Research Center of Energy Efficient and Custom AI IC
Outline

1 Introduction
2 Algorithms
3 Experiment
4 Conclusion
Digital IC design back end is automated using powerful EDA tools.

While the front end of digital IC design still requires lots of manpower.
Agile Design Based on Chisel

<table>
<thead>
<tr>
<th>Cores &amp; Accelerators</th>
<th>BOOM</th>
<th>Rocket</th>
<th>Gemmini</th>
<th>Hwacha</th>
</tr>
</thead>
<tbody>
<tr>
<td>SoC Libraries</td>
<td>rocketchip</td>
<td>rocketchip &amp; chisel-utils</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chisel-based Parameterized SoC Impl.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compiler</td>
<td>Chisel Frontend</td>
<td>FIRRTL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RISC-V SoC RTL description</td>
<td></td>
<td>Verilog files</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Agile development is gradually being adopted to reduce chip design costs and accelerate design cycles.
- Chisel designs, like Rocket Core, Berkeley Out-of-Order Machine, and Gemmini, are configurable and parameterizable RISC-V processors.
Gemmini SoC
### Gemmini SoC Parameters Examples

Parameters (19 in the paper) include:

- selection of CPU
- cache sizes
- accelerator configurations

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Stage</th>
<th>Candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>cpu_type</td>
<td>CPU core</td>
<td>RocketBig/BoomMed/BoomLarge</td>
</tr>
<tr>
<td>L2TLBs</td>
<td></td>
<td>512, 1024</td>
</tr>
<tr>
<td>nWays</td>
<td>L2 Cache</td>
<td>4, 8, 16</td>
</tr>
<tr>
<td>capacityKB</td>
<td></td>
<td>512, 1024</td>
</tr>
<tr>
<td>tile_Rows/Columns</td>
<td></td>
<td>1, 4, 8</td>
</tr>
<tr>
<td>mesh_Rows/Columns</td>
<td></td>
<td>8, 16, 32, 64</td>
</tr>
<tr>
<td>sp_capacity</td>
<td>Accelerator</td>
<td>256, 512, 1024, 2048, 4096</td>
</tr>
<tr>
<td>sp_banks</td>
<td></td>
<td>4, 16</td>
</tr>
<tr>
<td>dma_buswidth</td>
<td></td>
<td>128, 256</td>
</tr>
</tbody>
</table>
 Vivo 5G chip with EHRM and transformed Gaussian process

Overall Flow

VLSI Evaluation Flow

SoC Parameters
- e.g. tile=1, mesh=16
- sp_cap=256, acc_cap=64
- BOOM Core, Rocket Core

Next Suggestion

EH RM Acquisition

Predicted Mean & Variance

Transformed Gaussian Process

Verilog

Chisel files

Power Analysis

LLMs’ RoCC tests
- Transformer-small
- Bert-base
- Transformer-large
- Bert-large
- GPT1-small

Transformed Bayesian Optimization

Dataset
- Design
- Power
- Area
- Cycles

X Y1 Y2 Y3

SoC Parameters

Designer

LLMs

Synthesis

Power

Transformed Bayesian Optimization

VLSI Evaluation Flow

Overall Flow
A GP model is defined as a collection of random variables, any finite number of which have a joint Gaussian distribution.

A GP is completely specified by its mean function $m(x)$ and covariance function $k(x, x')$:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')),$$

(1)

where $x \in \mathbb{R}^d$ represents the input variable vector.
The Gaussian process is transformed as:

\[
f(x) = f^* - \frac{1}{2}g^2(x) \quad g(x) \sim GP (m_0, K),
\]

so \( f(x) \) will not beyond the given spec’s value \( f^* \).
The expected hypervolume improvement EHVI is defined as the expectation of hypervolume’s improvement with respect to the posterior predictive distribution of the GP:

$$EHVI(y) = \mathbb{E}_{p(y|D)}[I(y)],$$

(3)
EHRM aims to find a configuration of SoC architecture parameters with an expectation closest to the given spec, in other words, with the smallest expected hyper-regret:

\[ x_{t+1} = \arg \min_{x \in X} \mathbb{E}[\text{EHRM}(x)] = \arg \min_{x \in X} \mathbb{E}[Hr(x)]. \]
Experimental Setting

- Gemmini-based RISC-V SoC
- 5 Popular LLMs as performance evaluation
- Cadence Joules and Genus for power and area evaluation
- ASAP7 7nm PDK
### Experiment

<table>
<thead>
<tr>
<th>Metric \ Methods</th>
<th>GLSVLSI’07</th>
<th>HPCA’07</th>
<th>DAC’16</th>
<th>ASPDAC’20</th>
<th>ICCAD’21</th>
<th>ISCA’23</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>HV&lt;sub&gt;0,1&lt;/sub&gt;</td>
<td>0.6320</td>
<td>0.6491</td>
<td>0.6789</td>
<td>0.6610</td>
<td>0.6716</td>
<td>0.6398</td>
<td>0.7063</td>
</tr>
<tr>
<td>HV&lt;sub&gt;0,2&lt;/sub&gt;</td>
<td>0.7129</td>
<td>0.7255</td>
<td>0.7231</td>
<td>0.7144</td>
<td>0.7251</td>
<td>0.6929</td>
<td>0.7472</td>
</tr>
<tr>
<td>HV</td>
<td>0.5577</td>
<td>0.5636</td>
<td>0.5891</td>
<td>0.5758</td>
<td>0.5975</td>
<td>0.5609</td>
<td>0.6208</td>
</tr>
<tr>
<td>Average</td>
<td>0.6342</td>
<td>0.6460</td>
<td>0.6637</td>
<td>0.6504</td>
<td>0.6647</td>
<td>0.6312</td>
<td>0.6914</td>
</tr>
<tr>
<td>Ratio(%)</td>
<td>91.72</td>
<td>93.43</td>
<td>95.99</td>
<td>94.07</td>
<td>96.13</td>
<td>91.29</td>
<td>100</td>
</tr>
</tbody>
</table>

![Graphs](image)

- (a) Hypervolume vs Runtime (hours)
- (b) Norm. ADRS vs Runtime (hours)
The suggestion keeps getting close to the given spec QoR values and finally reaches the targets.
An architecture design space exploration method based on the transformed Bayesian optimization approach.

The constructed model utilizes the given spec QoR metric values as additional information to learn.

A tailored acquisition function is developed for optimization in multiple metrics (e.g., cycles, power, and area).
Thanks!