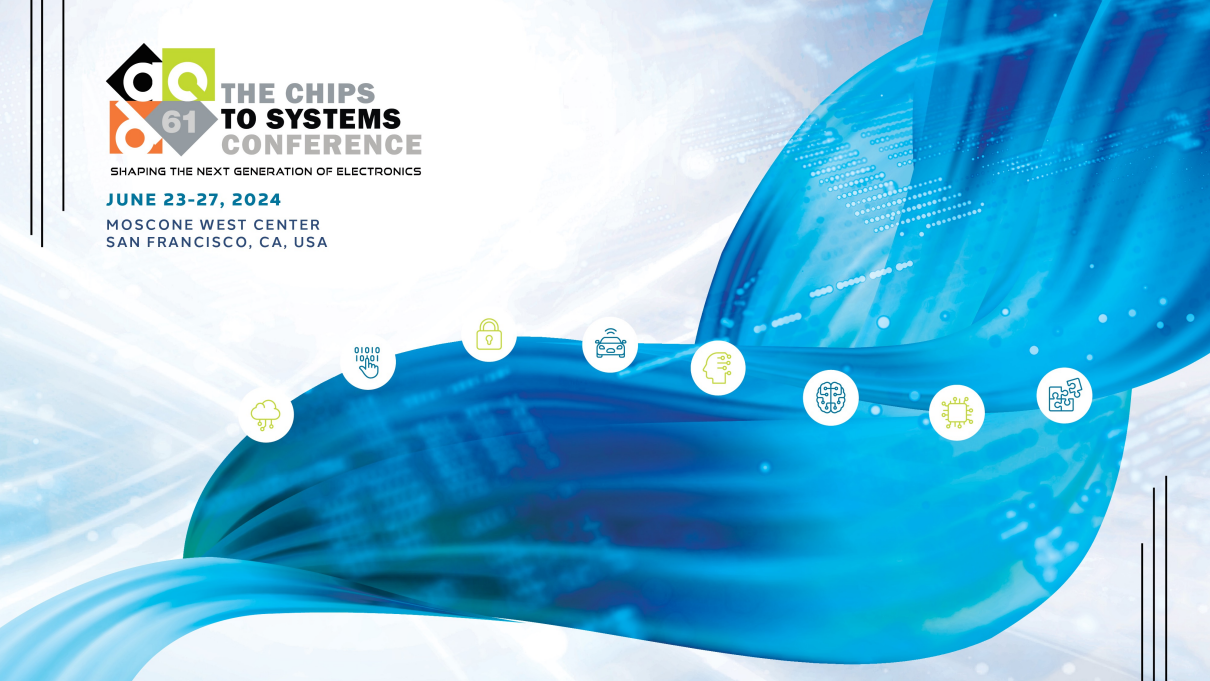

61
THE CHIPS
TO SYSTEMS
CONFERENCE
SHAPING THE NEXT GENERATION OF ELECTRONICS
JUNE 23-27, 2024
MOSCONE WEST CENTER
SAN FRANCISCO, CA, USA

61 THE CHIPS TO SYSTEMS CONFERENCE
SHAPING THE NEXT GENERATION OF ELECTRONICS
JUNE 23-27, 2024
MOSCONE WEST CENTER
SAN FRANCISCO, CA, USA

Knowing The Spec to Explore The Design via Transformed Bayesian Optimization

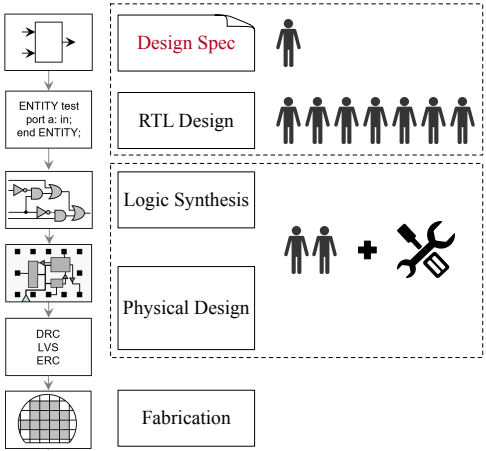
Donger Luo[*1], Qi Sun[*2], Xinheng Li[1], Chen Bai[3], Bei Yu[3], Hao Geng[1,4]

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

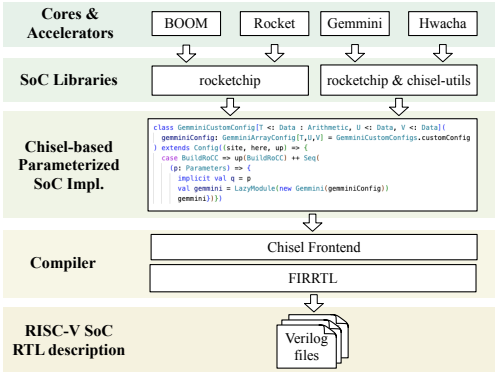
Outline

Digital IC Design Flow

- Digital IC design back end is automated using powerful EDA tools.
- While the front end of digital IC design still requires lots of manpower.

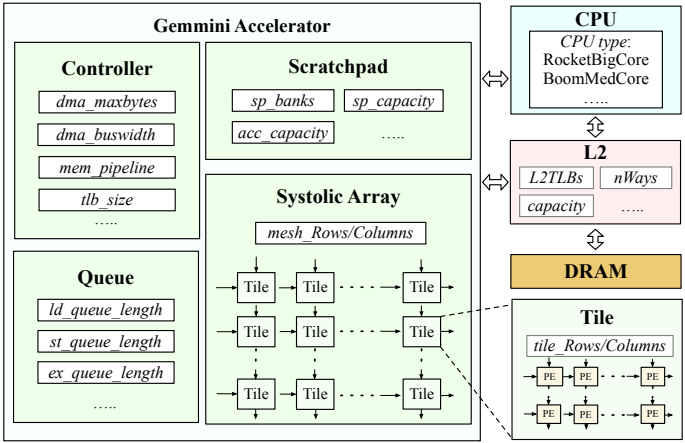


Agile Design Based ob Chisel



- 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 Accelerator
Controller
dma_maxbytes
dma_buswidth
mem_pipeline
tlb_size
.....
Scratchpad
sp_banks
sp_capacity
acc_capacity
.....
Systolic Array
mesh_Rows/Columns
Tile
Queue
ld_queue_length
st_queue_length
ex_queue_length
.....
CPU
CPU type:
RocketBigCore
BoomMedCore
.....
L2
L2TLBs
nWays
capacity
.....
DRAM
Tile
tile_Rows/Columns
PE

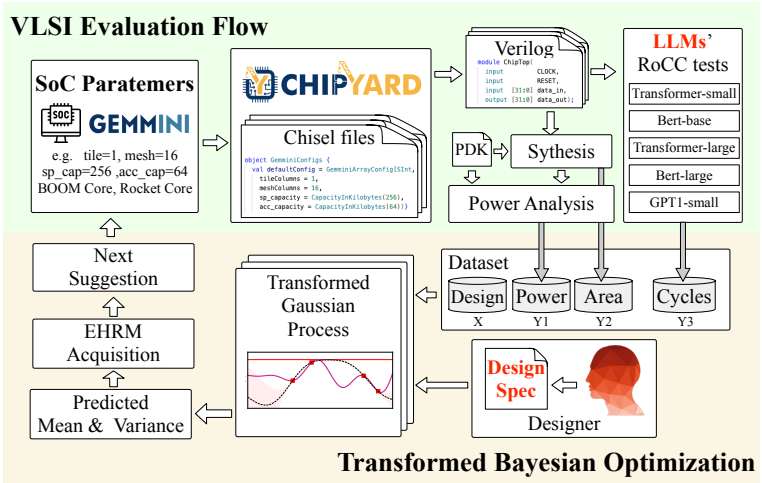
Gemmini SoC Parameters Examples

Paramaters (19 in the paper) include:

- selection of CPU
- cache sizes
- accelerator configurations

Parameters	Stage	Candidates
cpu_type	CPU core	RocketBig/BoomMed/BoomLarge
L2TLBs	L2 Cache	512, 1024
nWays		4, 8, 16
capacityKB		512, 1024
tile_Rows/Columns	Accelerator	1, 4, 8
mesh_Rows/Columns		8, 16, 32, 64
sp_capacity		256, 512, 1024, 2048, 4096
sp_banks		4, 16
dma_buswidth		128, 256

Overall Flow



Vanilla Gaussian Process

- 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 $m(\boldsymbol{x})$ and covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$:

$$f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')), \tag{1}$$

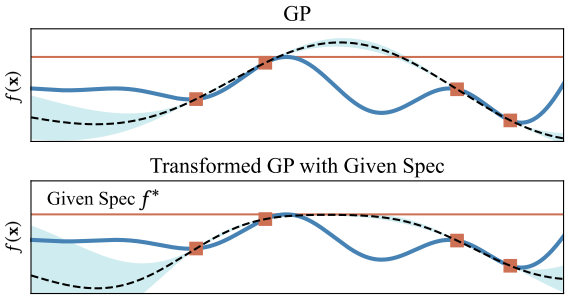
where $\boldsymbol{x} \in \mathbb{R}^d$ represents the input variable vector.

Transformed Gaussian Process

The Gaussian process is transformed as:

$$f(\mathbf{x}) = f^* - \frac{1}{2}g^2(\mathbf{x}) \quad g(\mathbf{x}) \sim GP\left(m_0, K\right), \tag{2}$$

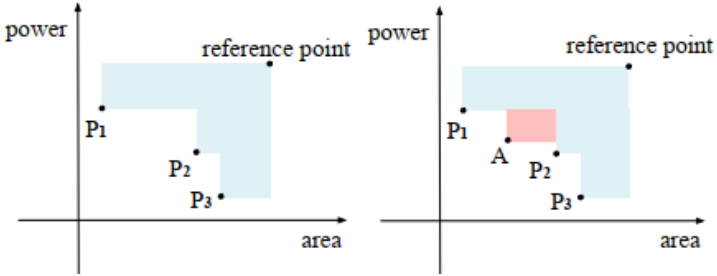
so $f(x)$ will not beyond the given spec's value f^*.



Acquisition Function

The expected hypervolume improvement EHVI is defined as the expectation of hypervolume's improvement with respect to the posterior predictive distribution of the GP:

$$\mathrm{EHVI}(\boldsymbol{y}) = \mathbb{E}_{p(\boldsymbol{y}|\boldsymbol{D})}[I(\boldsymbol{y})], \tag{3}$$



EHRM Acquisition

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

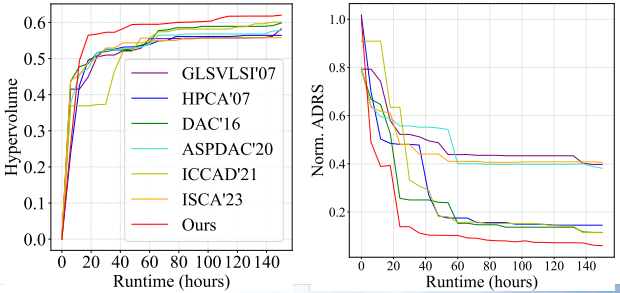
$$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in X} \mathbb{EHRM}(\mathbf{x}) = \arg\min_{\mathbf{x}\in X} \mathbb{E}[Hr(\mathbf{x})]. \tag{4}$$

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

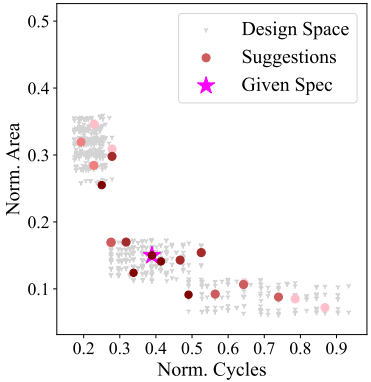
Metric \ Methods	GLSVLSI'07	HPCA'07	DAC'16	ASPDAC'20	ICCAD'21	ISCA'23	**Ours**
$HV_{0,1}$	0.6320	0.6491	0.6789	0.6610	0.6716	0.6398	**0.7063**
$HV_{0,2}$	0.7129	0.7255	0.7231	0.7144	0.7251	0.6929	**0.7472**
HV	0.5577	0.5636	0.5891	0.5758	0.5975	0.5609	**0.6208**
Average	0.6342	0.6460	0.6637	0.6504	0.6647	0.6312	**0.6914**
Ratio(%)	91.72	93.43	95.99	94.07	96.13	91.29	**100**



(a) (b)

Experiment

- The suggestion keeps getting close to the given spec QoR values and finally reaches the targets.



Conclusion

- 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).

61 THE CHIPS TO SYSTEMS CONFERENCE
SHAPING THE NEXT GENERATION OF ELECTRONICS
JUNE 23-27, 2024
MOSCONE WEST CENTER
SAN FRANCISCO, CA, USA

Thanks!