Iterative Frequency Tuning Targeting Energy Efficiency Ratio for FPGA-based Post-Quantum Cryptographic Cores

Srijeet Guha[1], and Andrea Guerrieri[2]

[1]NVIDIA Graphics, India

[2]HES-SO Valais-Wallis, School of Engineering, Sion, Switzerland

contact: andrea.guerrieri@ieee.org

Motivation

- Electronic Design Automation tools such as high level synthesis (HLS) have raised the level of abstraction, allowing the use of FPGAs in multiple domains, including post-quantum cryptography (PQC).
- Among power performance area (PPA) trade offs, the energy efficiency ratio is of increasing interest, especially with the need to integrate PQC cores into battery-powered portable appliances.
- To target energy efficiency at the HLS level, post-place, and route metrics should be taken into account, which makes an exhaustive design space exploration (DSE) extremely time-consuming.
- In this paper, we have developed an **iterative frequency tuning framework** capable of extracting the best quality, energy-efficient design in logarithmic time complexity.
- With this framework, we are able to converge to the best design frequency with a speedup of **2.89×** with respect to the classical approach.
- The framework also allows the selection of the criticality of each metric (ie. power consumed, wall clock time and area consumed on the FPGA fabric), thus altering the design cost function to suit the requirements of the HLS designers.

Iterative Frequency Tuning

Observations on Classical Frequency Tuning

- Most of the exploration time is consumed in placement, and routing of the design. HLS consumes comparatively very little time.
- Some metrics like latency, wall clock time (which are indicative of performance), and LUTs, FFs consumed (which are indicative of area) can be approximated just from the high-level synthesis.
- Dynamic power consumption is directly proportional to the operating frequency and capacitance of the circuit. Capacitance is closely related to the area consumed. Thus, the **minima of the power-frequency curve occur in a region near the minima of the area-frequency curve.**

Number Theoretic Transform NTT, original code of Kyber-768

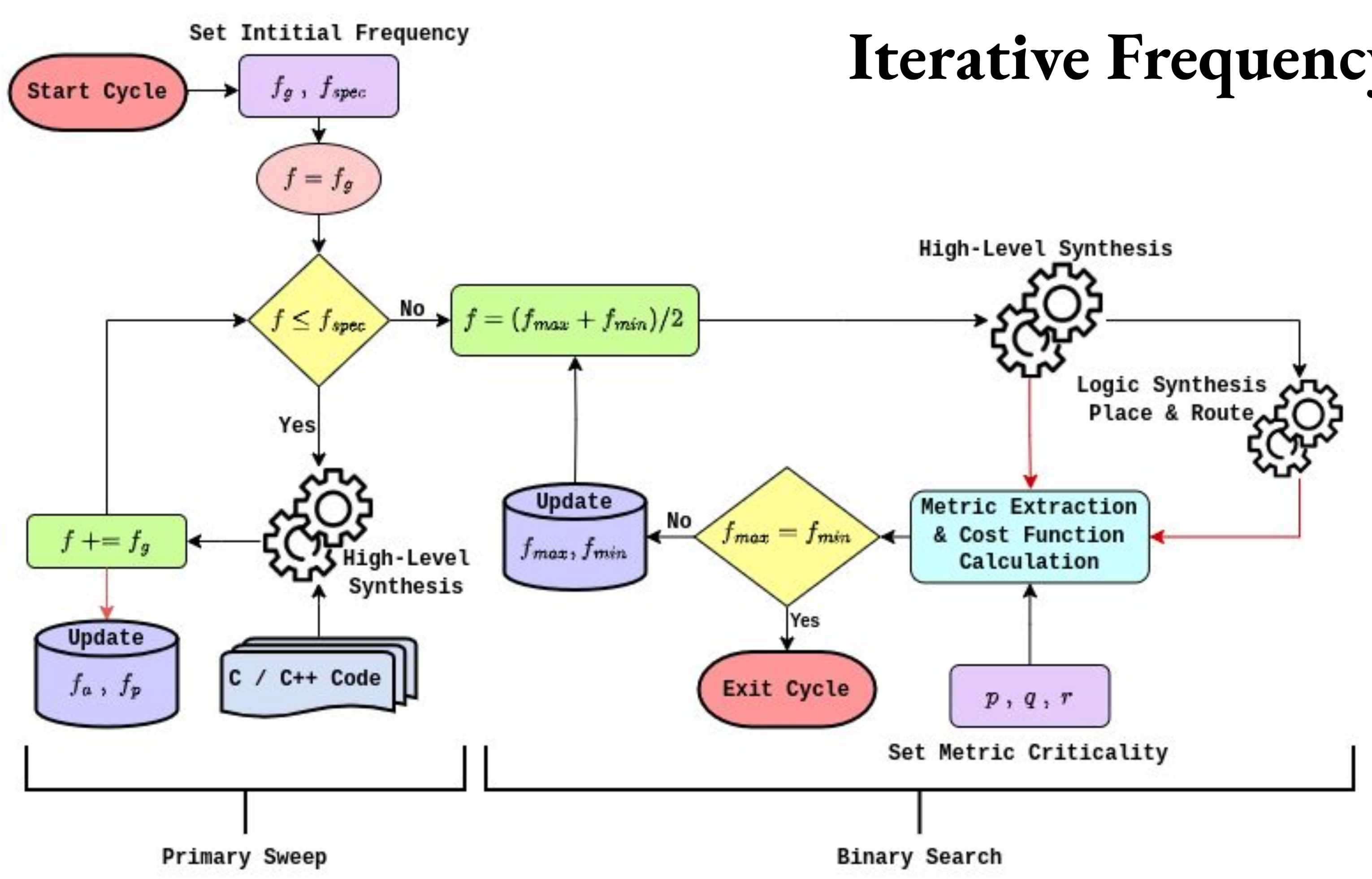
```
k = 1;
for(len = 128; len >= 2; len >>= 1){
    for(start = 0; start < 256; start = j + len){
    zeta = zetas[k++];
        for(j = start; j < start + len; j++){
        #pragma pipeline

            t = fqmul(zeta, r[j + len]);
            r[j + len] = r[j] - t;
            r[j] = r[j] + t;
} } }
```

Choice of Metrics

The framework allows us to adjust the criticality of each parameter in the PPA product as per design requirements. The resulting cost function (ϕ) that achieve is:

$$\text{Cost Function } (\phi) = \text{Power}^p \times \text{Performance}^q \times \text{Area}^r : p, q, r \in W$$

Iterative Frequency Tuning Framework



Iterative Frequency Tuning Framework

Primary Sweep

- The primary sweep involves using a HLS compiler for an approximate measure of performance and area. The sweep starts at the granular frequency and at each iteration, the synthesis frequency is increased by a factor of frequency of granularity
- The sweep is used to record the frequencies of maximum performance and minimum area to reduce the frequency range for binary search.

Binary Search

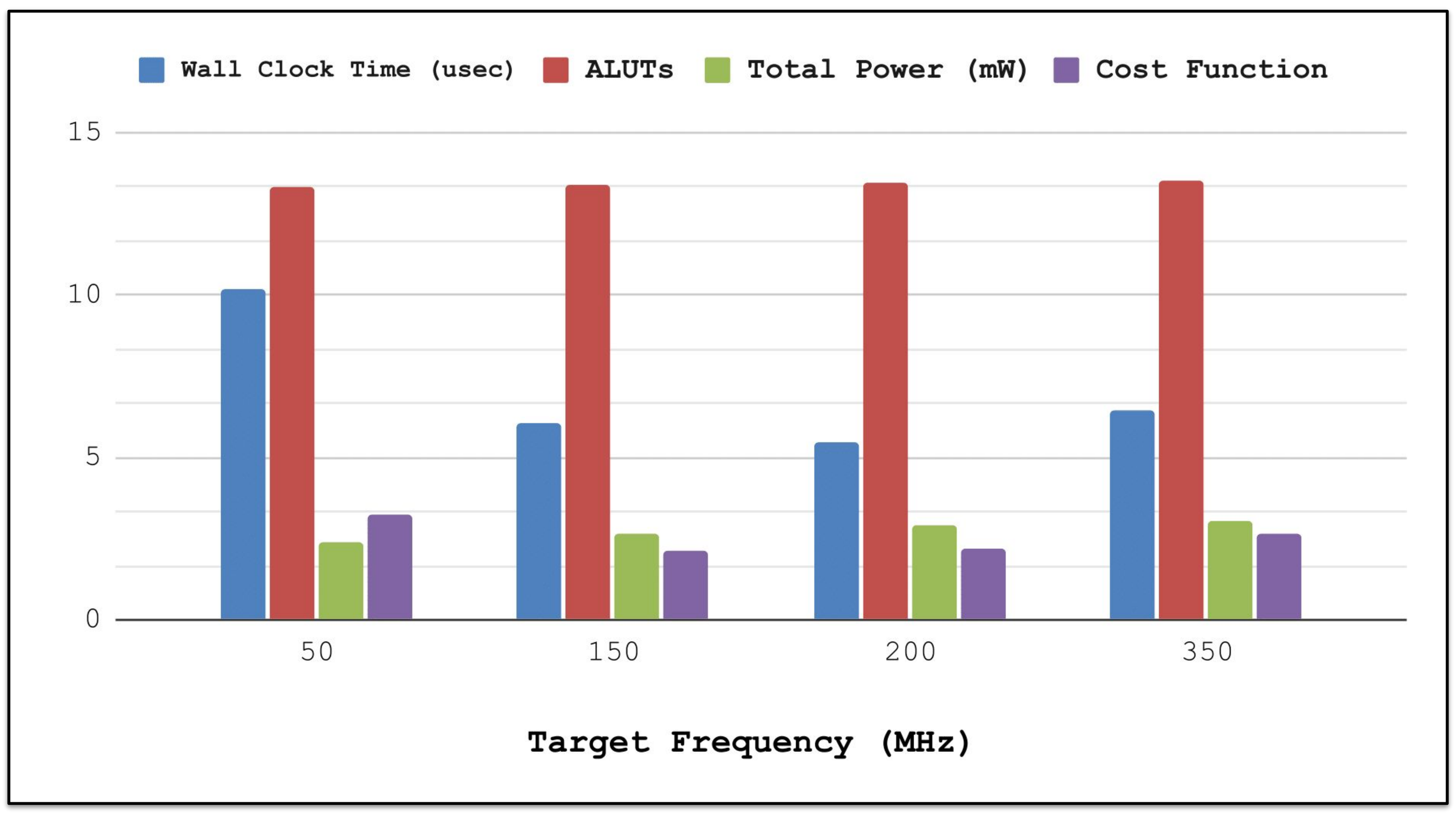
- Binary search is applied over the calculated frequency range, using a full cycle of HLS, placement, and routing to find the value of ϕ at f_{min}, f_{max} and $(f_{min} + f_{max})/2$.
- Based on the calculated values, we update f_{min} and f_{max} and repeat the above step until we converge at a common frequency

Implementation and Results

TABLE I

RESULTS: FREQUENCY SWEEPING. TABLE SHOWS: COST FUNCTION (ϕ), WALL CLOCK TIME (WCT), INITIATION INTERVAL (II) AND TOTAL POWER (TP). THE HIGHLIGHTED VALUES REPRESENT THE MINIMUM VALUES FOR EACH PARAMETER, WHICH IS NOT THE SAME AS THE FINAL PPA (150MHZ).

Target Freq. (MHz)	Fmax (MHz)	II	Latency (cc)	WCT (us)	LUTs ($\times 10^3$)	FFs ($\times 10^3$)	TP (mW)	Energy (uJ)	ϕ ($\times 10^8$)
50	153.30	1	**1556**	10.15	**13.33**	13.42	**2389**	24.25	3.23
100	213.72	1	1567	7.33	13.36	13.56	2584	18.95	2.53
150	261.30	1	1574	6.02	13.37	13.77	2670	16.08	**2.15**
200	289.60	1	1583	**5.47**	13.49	14.00	2928	**16.00**	2.16
250	336.36	2	2479	7.37	13.45	14.16	2973	21.91	2.95
300	379.08	2	2500	6.59	13.51	14.20	3046	20.09	2.71
350	388.35	2	2500	6.44	13.52	14.25	3038	19.56	2.64
400	376.36	3	3906	10.38	13.55	14.76	3034	31.49	4.27
450	390.01	3	3914	9.93	13.57	14.81	3071	30.51	4.14
500	375.38	3	3914	10.43	13.57	14.80	3055	31.85	4.32
550	397.62	5	6227	15.66	14.03	20.17	3261	51.07	7.16
600	400.48	11	13734	34.29	16.07	24.18	3715	127.40	20.47
650	402.34	12	15137	37.62	16.07	24.05	3701	139.23	22.37
700	359.97	13	16024	44.51	16.01	23.99	3660	162.92	26.08
750	375.92	14	17449	46.42	16.04	24.03	3823	177.46	28.46
800	382.70	15	18326	47.89	16.11	32.36	3895	186.52	30.06
850	413.24	16	19759	47.81	16.13	32.40	3996	191.05	30.82
900	458.72	17	20631	44.98	16.09	32.37	3890	174.95	28.15
950	411.23	18	22069	53.67	16.10	32.38	3887	208.62	33.59
1000	370.78	19	23447	63.24	16.10	32.42	3892	246.12	39.61



32nd IEEE International Conference on Circuits and Systems

Nancy, France

November 18 - 20, 2024