UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI
INGEGNERIA INFORMATICA,
MODELLISTICA, ELETTRONICA
E SISTEMISTICA
DIMES

Report Tecnico n.3

Tecniche di progettazione low power RTL-based

Corso di Progettazione Low Power

Corso di Laurea Magistrale in Ingegneria Elettronica

Unical, aa 2020/2021

Andrea Alecce Matricola 214611

Prof. F. Frustaci

Sommario

1. **Intro** 3
2. **Non ottimizzato** 3
3. **Pipeline** 6
4. **Reordering** 7
5. **Clock Gating** 9

1. Intro

In questa relazione si illustrano e simulano in ambiente Vivado alcune tecniche low power applicate direttamente a livello di RTL. In particolare, si utilizzeranno le tecniche di :

- Gate Reordering
- Pipeline
- Clock Gating

Ci si concentra dunque su tecniche che fanno uso di codice VHDL, sfruttando alcuni tool messi a disposizione da Vivado, al fine di ottimizzare dal punto di vista energetico il circuito. In questa fase di design, non si ha accesso al singolo transistor, dunque si è fortemente vincolati. In particolare, su FPGA bisogna seguire un flusso di progettazione ben specifico e di seguito illustrato:

- RTL Design: stesura del codice che descrive il circuito;
- RTL Simulation: simulare il codice ottenendo le forme d'onda digitali e un file (.saif) in cui è memorizzata l'attività di ogni nodo;
- RTL Power Analysis: le informazioni sulle attività di switching vengono fornite al tool Power Analyzer di Vivado, oltre che quelle sul modello di potenza dissipata dal device. Il tool fornirà il valore stimato della potenza del circuito.

Si evidenzia che l'attività di switching corretta si ottiene solamente dopo che il progetto è stato sintetizzato e implementato (Post-Implementation).

La **potenza statica** (leakage) del chip FPGA non può essere modificata. In FPGA, la potenza statica è principalmente dovuta alle risorse che seppur non utilizzate sono comunque presenti ed alimentate, non controllabili a livello di RTL.

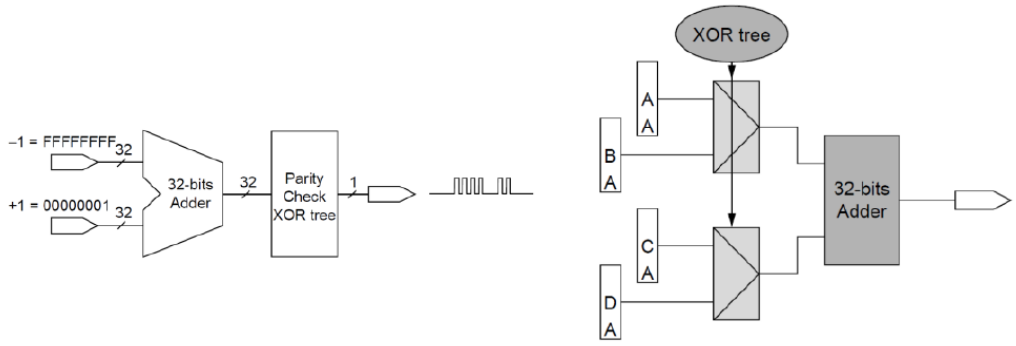
Anche la **potenza di cortocircuito** del chip FPGA non può essere modificata, in quanto dipende dal tempo di salita/discesa dei segnali.

La componente di potenza su cui si può agire a questo livello è quella **dinamica**. In genere, non si può agire sulla tensione di alimentazione ne sulla frequenza, poiché legata alla specifica applicazione. Il parametro che permette di ottenere i maggiori benefici è il **fattore di switching**.

Con la riduzione del fattore di switching si intende ridurre il numero di transizioni dei segnali (Transition Count, TC) non volute, dovute ai glitch.

2. Design non ottimizzato

Ai fini dell'analisi sull'ottimizzazione energetica, si considera il seguente circuito:



Il circuito di destra è quello principale, mentre a sinistra quello di controllo. Il circuito principale presenta:

- 4 registri
- 2 MUX;
- Un sommatore a 32 bit.

I multiplexer fanno passare determinati segnali verso il sommatore:

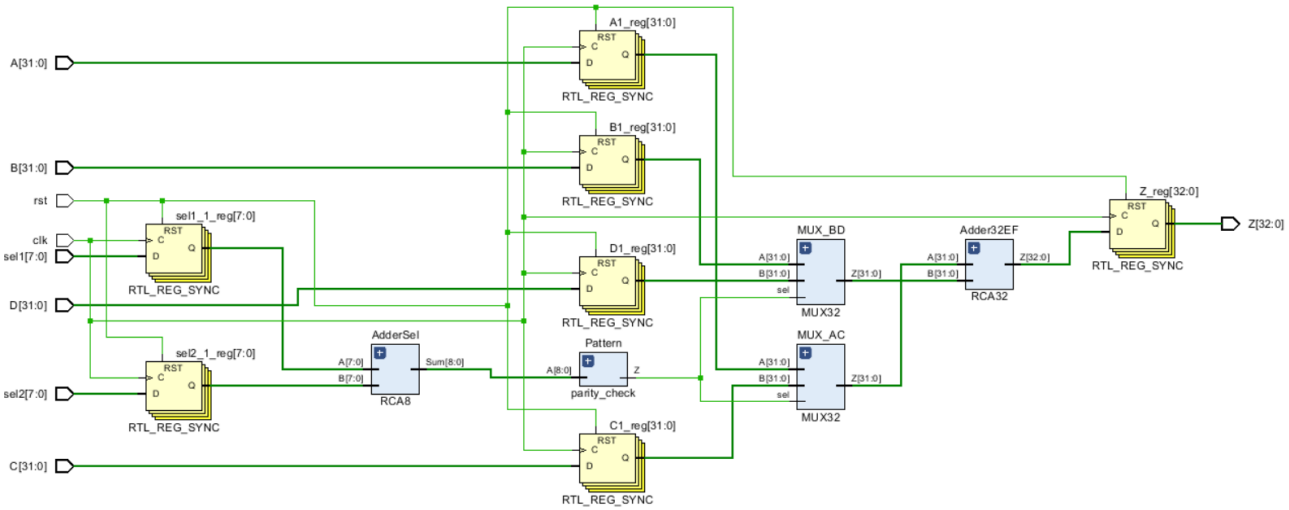
- Sel0: somma tra A e C;
- Sel1: somma tra B e D.

I Mux sono pilotati dal segnale generato dalla sezione di controllo, che per sua natura sarà affetto da glitch. Questo circuito di controllo, nominato **Parity Check XOR Tree**, è composto proprio da una catena di XOR restituisce un unico bit, 0 se il numero di bit in ingresso è pari, 1 altrimenti. Questo circuito presenta:

- Un sommatore
- Una catena di xor;

L'input dello xor tree viene da un sommatore. Gli output bit del sommatore non si assestano nello stesso momento, a causa della propagazione del riporto (MSB più lenti), causando glitch. A causa della presenza di questi glitch, il sommatore del circuito di calcolo svolgerà tante somme dovute alle transizioni non volute del segnale di selezione, sprecando potenza.

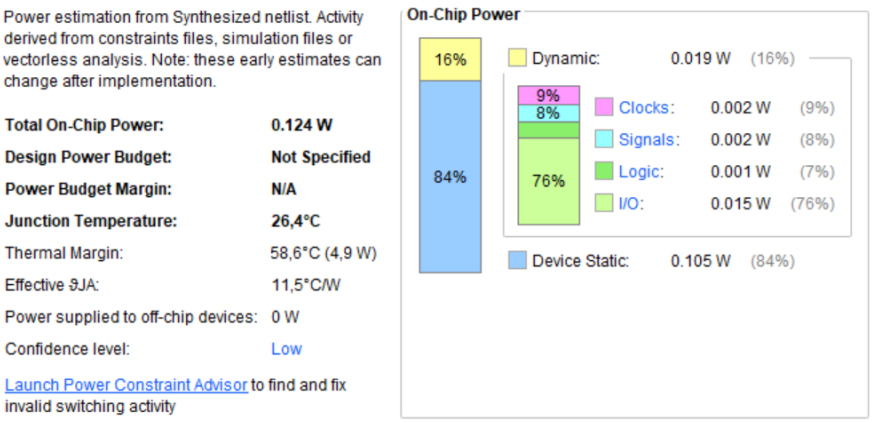
Lo schematic RTL del circuito realizzato in Vivado è il seguente:



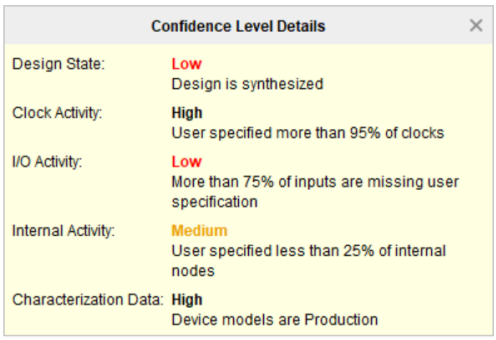
Il design utilizza 116 LUT e 177 FF per realizzare il circuito. Si valutano i TC di un generico segnale (G1), ovvero il segnale in uscita all' ultimo sommatore:

```
(G1\[2\] (T0 614429) (T1 469763) (TX 15808) (TZ 0) (TB 0) (TC 425))
(G1\[3\] (T0 600331) (T1 492501) (TX 7168) (TZ 0) (TB 0) (TC 514))
(G1\[4\] (T0 624548) (T1 467357) (TX 8095) (TZ 0) (TB 0) (TC 551))
(G1\[5\] (T0 601721) (T1 491007) (TX 7272) (TZ 0) (TB 0) (TC 542))
(G1\[6\] (T0 688755) (T1 405779) (TX 5466) (TZ 0) (TB 0) (TC 474))
(G1\[7\] (T0 609893) (T1 482875) (TX 7232) (TZ 0) (TB 0) (TC 557))
(G1\[8\] (T0 571551) (T1 521218) (TX 7231) (TZ 0) (TB 0) (TC 561))
(G1\[9\] (T0 577675) (T1 513365) (TX 8960) (TZ 0) (TB 0) (TC 599))
```

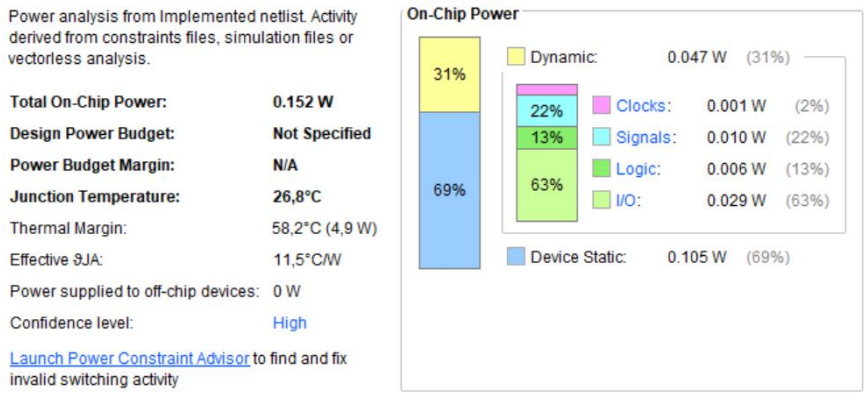
Dal report power si nota che la potenza totale dissipata è pari a 0.124W di cui 0,105W sono relativi alla potenza statica. I restanti 0.019W sono relativi alla potenza dinamica.



Si evidenzia che il livello di confidenza è basso, come descritto in maniera dettagliata nella seguente figura:



Per migliorare il livello di confidenza, si effettua una simulazione post-implementation fornendo il file .saif precedentemente generato alla simulazione, ottenendo i seguenti valori di potenza.



Si riportano i valori di potenza dissipata da ogni singolo componente in maniera più precisa.

Pno_opt [mW]	Clock [mW]	Signal [mW]	Logic [mW]
18.428	1.774	10.456	6.197

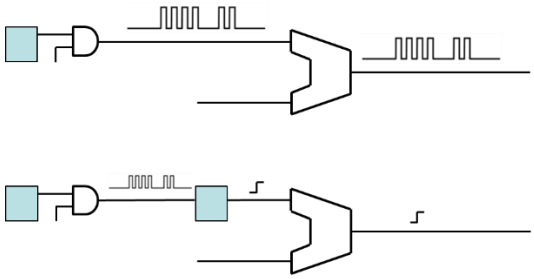
La **logic** è la dissipazione delle risorse, quindi delle LUT e dei Flip Flop. **Signals** invece è la componente di energia che si dissipa sulle interconnessioni. Siccome il numero di Flip Flop nel nuovo circuito non è variato, non si dovrebbe trovare una variazione sulla dissipazione del **clock** in quanto la rete di clock andrà a pilotare lo stesso numero di componenti sequenziali. Ci si aspetta però che la dissipazione di potenza sui segnali e/0 sulla logica sia inferiore. Quello sugli **I/O**, che è la componente maggiore, rimarrà uguale in quanto non sono state fatte modifiche sul file TOP, quindi sulla frequenza di ingressi e uscite.

In teoria, la potenza di I/O non si dovrebbe neanche considerare, nel calcolo della possibile energia da ridurre. Si sta considerando un sotto circuito, che sarà realisticamente all'interno di un'architettura molto più complessa, dunque non avrà contributi in termini di I/O, motivo per il quale può non essere considerata.

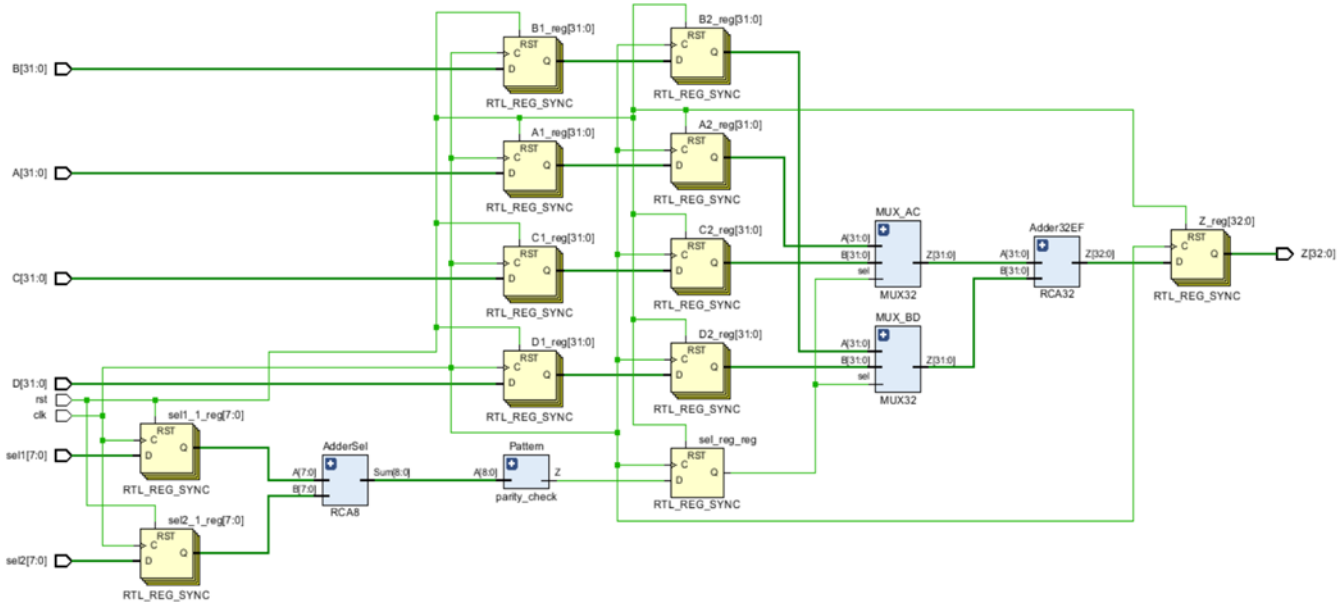
3. Design con Pipeline

Una prima idea per abbattere i consumi è quello di inserire, all'interno dell'architettura non ottimizzata, uno stadio di pipeline. Ciò contribuisce alla riduzione dei glitch, in quanto permette di sincronizzare i dati che viaggiano sui segnali in ingresso ed evitando transizioni non volute a valle. Questo metodo però implica un contributo in dissipazione dovuto alla presenza dei nuovi registri, oltre che una maggiore latenza. In questo caso, bisogna raggiungere un compromesso.

Uno schema esemplificativo è riportato in figura:



Lo schematic RTL del circuito realizzato in Vivado è il seguente:

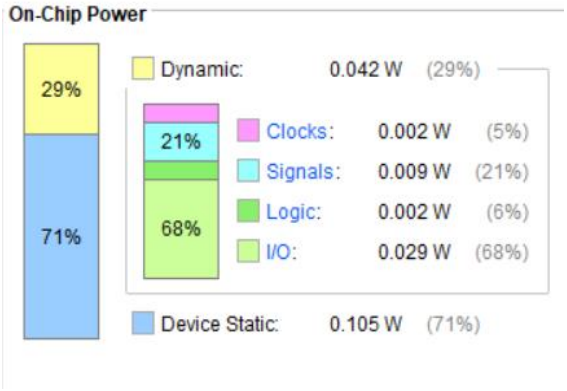


È stato inserito quindi uno stadio di pipeline sulla linea degli ingressi e sul segnale di selezione in uscita dall'albero di xor, quest'ultimo critico dal punto di vista dei glitch. In questa configurazione, ci si assicura che i segnali siano sincronizzati.

Il design utilizza 306 FF rispetto ai 177 precedenti per realizzare il circuito: ci si aspetta dunque un aumento di potenza dinamica sul clock. Si valutano i TC del segnale G1:

```
[G1\[1\] (T0 561132) (T1 527168) (TX 11700) (TZ 0) (TB 0) (TC 152))
[G1\[2\] (T0 650035) (T1 446825) (TX 3140) (TZ 0) (TB 0) (TC 176))
[G1\[3\] (T0 637424) (T1 459677) (TX 2899) (TZ 0) (TB 0) (TC 180))
[G1\[4\] (T0 493006) (T1 594984) (TX 12010) (TZ 0) (TB 0) (TC 197))
[G1\[5\] (T0 619169) (T1 469021) (TX 11810) (TZ 0) (TB 0) (TC 204))
[G1\[6\] (T0 636083) (T1 451203) (TX 12714) (TZ 0) (TB 0) (TC 224))
```

Che risultano diminuiti. Dal report power si nota che la potenza totale dissipata è pari a 0.147W di cui 0.105W sono relativi alla potenza statica. I restanti 0.042W sono relativi alla potenza dinamica.



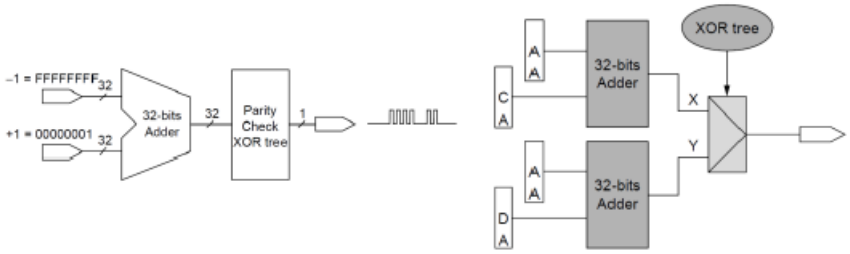
Si riportano i valori di potenza dissipata da ogni singolo componente in maniera più precisa.

Pno_opt [mW]	Clock [mW]	Signal [mW]	Logic [mW]
13.654	2.315	8.756	2.465

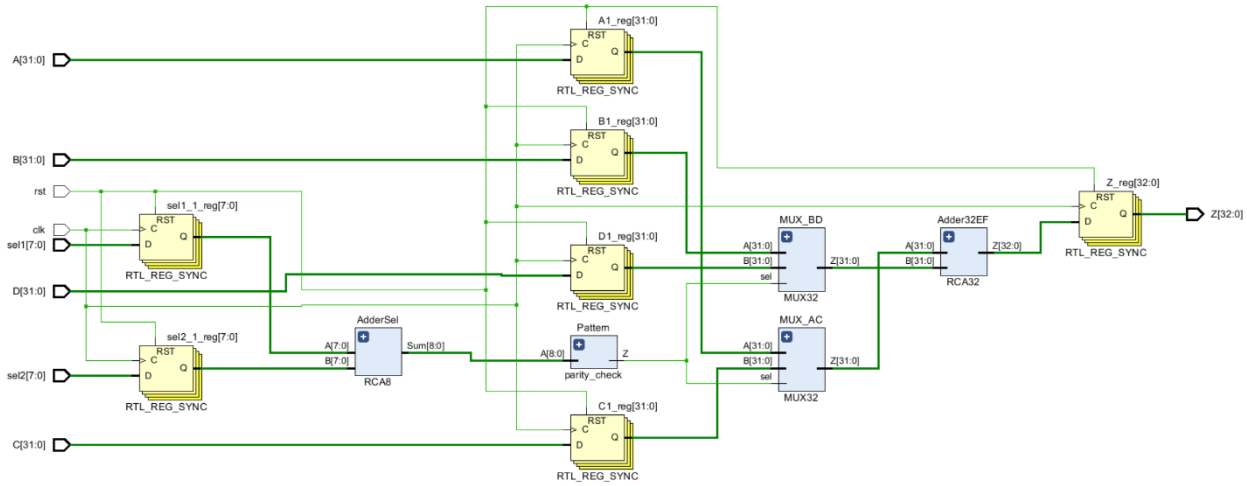
Come previsto, la potenza sulla linea di clock è aumentata, ottenendo però una diminuzione della potenza dinamica. Infatti, si è ottenuto una diminuzione del 25.9%.

4. Design con Gate Level Reordering

Una ulteriore tecnica per ridurre i glitch e quindi la potenza dinamica dissipata è quella del Gate Level Reordering, scrivendo il codice riarrangiando topologicamente il sistema. Si considera il seguente circuito:

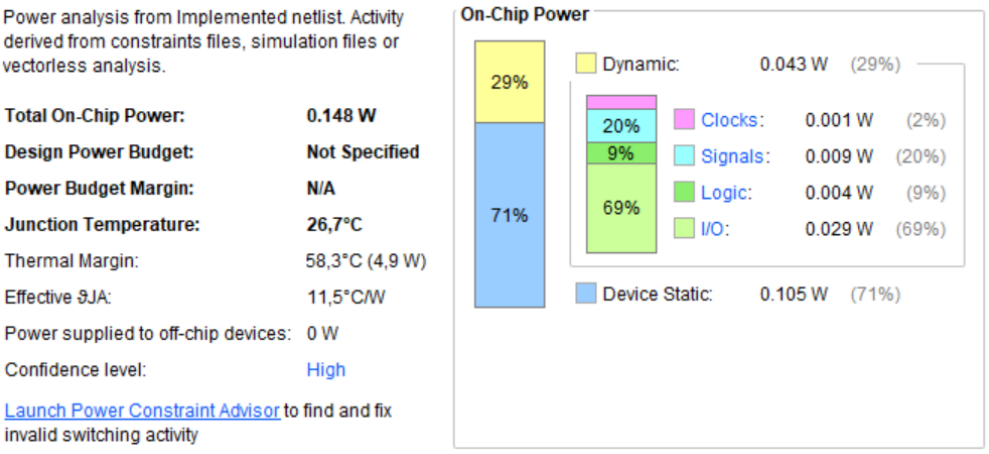


La differenza dall'architettura precedente è che gli ingressi dei sommatori provengono direttamente dai registri, che generano meno glitch. Si procede, come per i casi precedenti, alla simulazione su Vivado. Lo schematic RTL del circuito realizzato è il seguente:



Il design utilizza 177 FF, mentre il numero di LUT aumenta da 116 a 123, per la presenza del sommatore in più. Si valutano i TC del segnale G1:

```
(G1\[0\] (T0 527821) (T1 567782) (TX 4397) (TZ 0) (TB 0) (TC 273))
(G1\[1\] (T0 674941) (T1 419840) (TX 5219) (TZ 0) (TB 0) (TC 309))
(G1\[2\] (T0 543677) (T1 550512) (TX 5811) (TZ 0) (TB 0) (TC 365))
(G1\[3\] (T0 645962) (T1 450872) (TX 3166) (TZ 0) (TB 0) (TC 373))
(G1\[4\] (T0 607807) (T1 489446) (TX 2747) (TZ 0) (TB 0) (TC 383))
(G1\[5\] (T0 570485) (T1 524432) (TX 5083) (TZ 0) (TB 0) (TC 348))
(G1\[6\] (T0 581468) (T1 513643) (TX 4889) (TZ 0) (TB 0) (TC 421))
(G1\[7\] (T0 646787) (T1 450920) (TX 2293) (TZ 0) (TB 0) (TC 341))
(G1\[8\] (T0 578561) (T1 516585) (TX 4854) (TZ 0) (TB 0) (TC 501))
(G1\[9\] (T0 539867) (T1 556429) (TX 3704) (TZ 0) (TB 0) (TC 390))
```



La potenza statica è la stessa di prima. Quella dinamica è diminuita invece da 47mW a 43mW. Non considerando quella associata agli I/O si ha una diminuizione da 17mW a 14mW. In termini percentuali, si ha una riduzione del 17.64%.

Valutando il Transition Count (TC) sullo stesso segnale G1 rispetto a prima, si nota una riduzione di questo valore da 425 a 273, pari al 35.74%.

Pno_opt [mW]	Clock [mW]	Signal [mW]	Logic [mW]
13.126	1.088	8.418	3.520

5. Clock Gating

Un'altra tecnica utilizzabile per la riduzione della potenza è quella del **Clock Gating**, mirata alla dissipazione di potenza delle linee di clock. In particolare, serve a ridurre la potenza dinamica di un circuito in fase di standby: tutti gli elementi collegati al clock vengono disabilitati/congelati, evitando transizioni inutili.

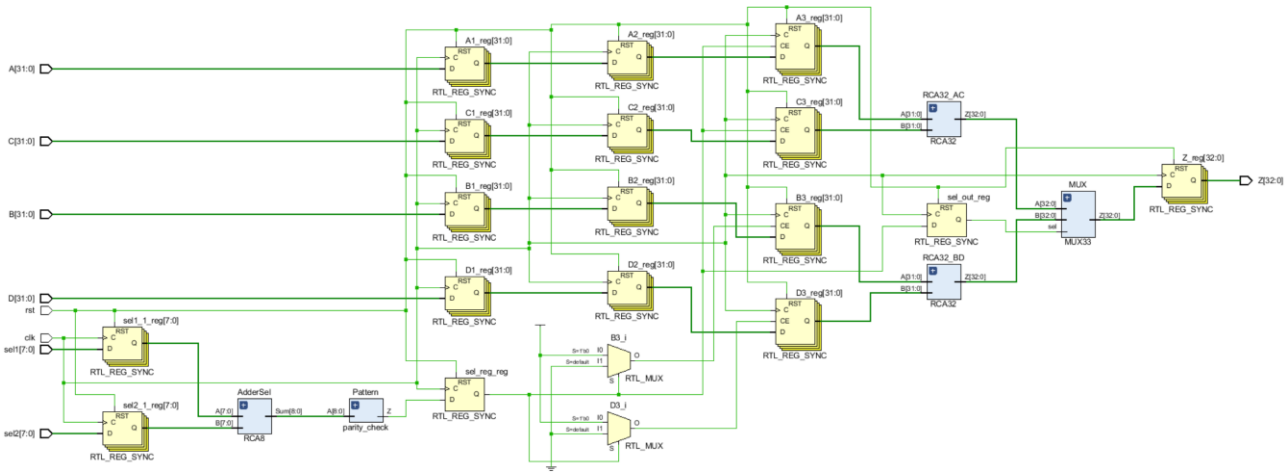
Teoricamente, è possibile applicare il clock gating a **livello globale** e a **livello locale**. Agire a livello globale permette di disabilitare dalla fonte il clock, permettendo il massimo guadagno in termini di energia, ma a discapito di una maggiore latenza quando il sistema riparte, oltre che un overhead maggiore. A livello locale invece, si ha un guadagno minore ma i problemi riscontrati si abbattono.

L'abilitazione o meno di una linea di clock può essere fatta tramite una AND.

In FPGA questo approccio non è però consigliabile, in quanto il clock è presente su linee dedicate. Inserire elementi come porte AND su queste linee va contro il principio di ottimizzazione della tecnologia riprogrammabile. Esistono comunque alcune *primitive*, situate su queste linee dedicate, utili per ridurre ulteriormente la dissipazione di potenza. Quello che si fa invece è mantenere costante gli ingressi degli elementi sequenziali in modo da congelare l'uscita: in questo caso si parla di **data gating**.

Il clock gating in FPGA avviene sfruttando il **Clock Enable** (**CE**) dei vari registri, in quanto tutte le risorse ne sono già predisposte.

Si combinano tutte le tecniche viste ed adottate finora per apprezzarne i vantaggi in termini energetici. . Lo schematic del circuito realizzato è il seguente:



Si notano la presenza di due registri, in cascata, sull'uscita del circuito di controllo. Al fine di mantenere la sincronia tra i segnali in ingresso e quello di selezione, sono presenti un totale di 3 registri in cascata. Il segnale di CE dei registri dei segnali di ingresso, prima dei due RCA32, viene gestito dal segnale di selezione in uscita dal circuito di controllo. Infatti, il valore di selezione è ben definito un periodo di clock prima che i segnali di ingresso entrino nel circuito combinatorio di somma; in funzione del valore di selezione, verranno congelate le uscite della coppia di ingressi non interessati.

Inoltre, è stato implementato un clock gating di tipo **esplicito**, come riportato da questo estratto del codice VHDL del TOP:

```
if (sel_reg='1')then
    A3<=A2;
    C3<=C2;
end if;
if (sel_reg='0') then
    B3<=B2;
    D3<=D2;
end if;
```

Si valutano i TC del segnale G1, nettamente diminuiti:

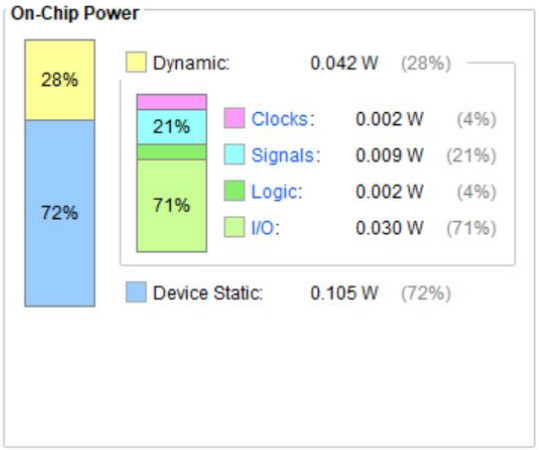
```
(G1\[0\] (T0 618176) (T1 480308) (TX 1516) (TZ 0) (TB 0) (TC 64))
(G1\[1\] (T0 640066) (T1 457680) (TX 2254) (TZ 0) (TB 0) (TC 77))
(G1\[2\] (T0 561804) (T1 536392) (TX 1804) (TZ 0) (TB 0) (TC 71))
(G1\[3\] (T0 712707) (T1 385288) (TX 2005) (TZ 0) (TB 0) (TC 125))
(G1\[4\] (T0 592152) (T1 505223) (TX 2625) (TZ 0) (TB 0) (TC 119))
(G1\[5\] (T0 574491) (T1 523354) (TX 2155) (TZ 0) (TB 0) (TC 139))
(G1\[6\] (T0 694868) (T1 403102) (TX 2030) (TZ 0) (TB 0) (TC 150))
(G1\[7\] (T0 618546) (T1 479179) (TX 2275) (TZ 0) (TB 0) (TC 164))
(G1\[8\] (T0 613593) (T1 484181) (TX 2226) (TZ 0) (TB 0) (TC 136))
(G1\[9\] (T0 571495) (T1 526408) (TX 2097) (TZ 0) (TB 0) (TC 177))
```

Dal report power si nota che la potenza totale dissipata è pari a 0.147W di cui 0.105W sono relativi alla potenza statica, come nel caso precedente. I restanti 0.042W sono relativi alla potenza dinamica.

Power analysis from Implemented netlist. Activity derived from constraints files, simulation files or vectorless analysis.

Total On-Chip Power:	**0.147 W**
Design Power Budget:	**Not Specified**
Power Budget Margin:	**N/A**
Junction Temperature:	**26,7°C**
Thermal Margin:	58,3°C (4,9 W)
Effective ϑJA:	11,5°C/W
Power supplied to off-chip devices:	0 W
Confidence level:	High

Launch Power Constraint Advisor to find and fix invalid switching activity



Pno_opt [mW]	Clock [mW]	Signal [mW]	Logic [mW]
11.931	1.670	8.518	1.742

Con questa architettura, la componente che ha subito la maggior attenuazione è quella della logic, dovuta principalmente all'uso del clock gating e alla relativa riduzione di glitch. In particolare, si apprezza una riduzione della potenza pari al a riduzione della potenza pari al 35.25%. Si riporta una tabella riepilogativa delle tecniche utilizzate.

	Pno_opt [mW]	Clock [mW]	Signal [mW]	Logic [mW]
Non opt	18.428	1.774	10.456	6.197
Pipeline	13.654	2.315	8.756	2.465
Reordering	13.126	1.088	8.418	3.520
Clock Gating/reord/pipe	11.931	1.670	8.518	1.742