Network Flow-based Simultaneous Retiming and Slack Budgeting for Low Power Design

Speaker:Yuchun Ma

Authors: Bei Yu[1] Sheqin Dong[1] Yuchun Ma[1] Tao Lin[1] Yu Wang[1] Song Chen[2] Satoshi GOTO[2]

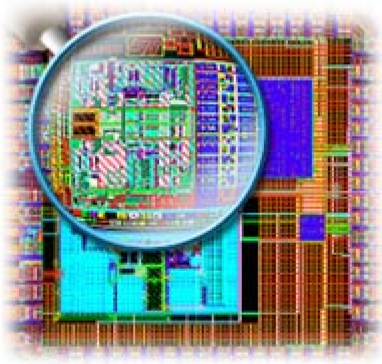
[1]Department of Computer Science & Technology
Tsinghua University, Beijing, China

[2]Graduate School of IPS, Waseda University, Kitakyushu, Japan

Outline

- Introduction
  - Previous Works
  - Problem Formulation
- Methodology
  - MILP Formulation
  - Remove Redundant Constraint
  - Convex Cost Dual Flow Algorithm
- Experimental Results

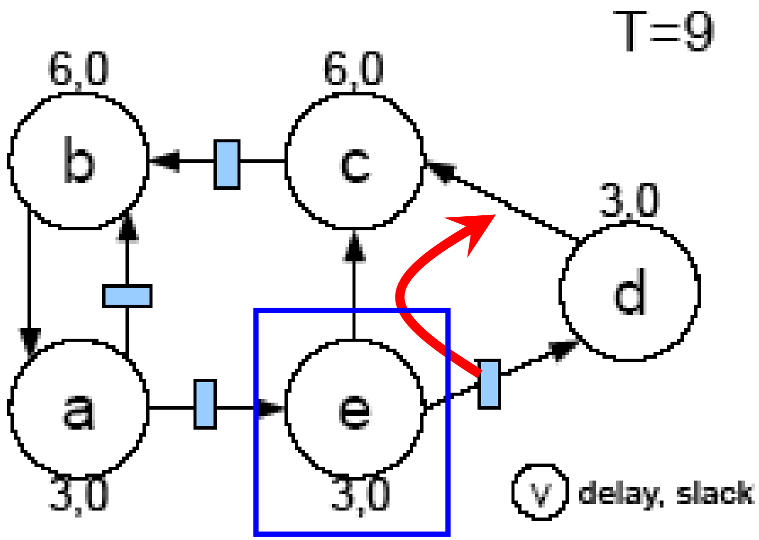
Background



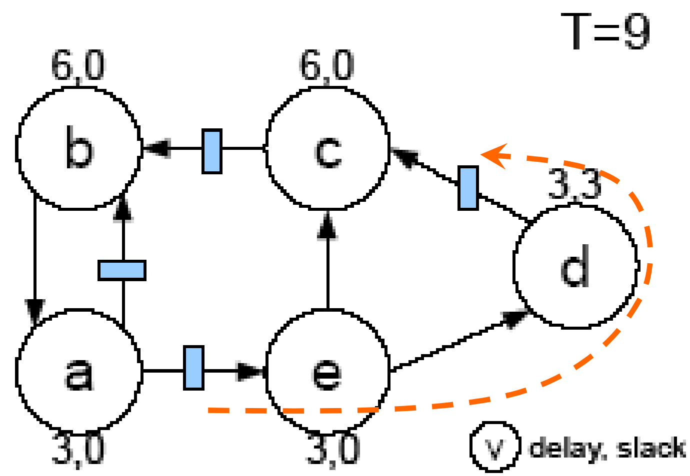
- More devices trend to be put in the small silicon area.
- The clock frequency is pushed even higher.
- Timing constraint and Low Power become significant requirement.
- Slack budget aims at
  - slowing down as many components as possible without violating the system's timing constraints.
  - The slowed-down components can be further optimized to improve system's area, power dissipation, or other design quality metrics.
- Retiming
  - is one of the most powerful sequential optimization techniques
  - relocates the flip-flops (FFs) in a circuit while preserving its functionality

Motivation

- Relocate FFs to increase potential slack without violating timing constraint.



No potential slack in this circuit.



The potential slack can be increased from 0 to 3

Simultaneous retiming and slack budget process

Previous Works

- **Retiming**:
  - [C.E.Leiserson et al. Algorithmica 1991]: first work
  - [N. Maheshwari et al. TCAD 1998]: flow based Min-area retiming
  - [H. Zhou, ASPDAC 2005]: incremental Min-period retiming
  - [J. Wang & H. Zhou DAC 2008]: incremental Min-period retiming
- **Slack Budgeting**:
  - [R. Nair et al. TCAD 1989]: ZSA, suboptimal heuristic
  - [C. Chen et al. TCAD 2002]: Maximum-Independent-Set, NP-complete
  - [S.Ghiasi et al. ICCAD 2004]: Flow based algorithm

Existing Work for slack budgeting (2)

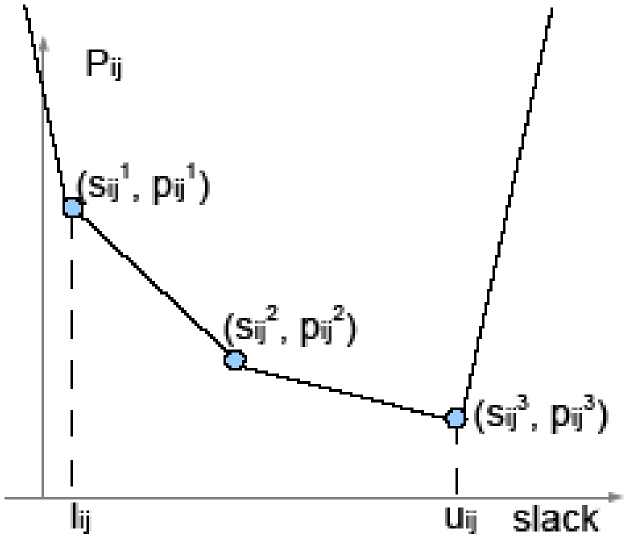
- **Retiming + Slack Budgeting**:
  - [Y. Hu et al. DAC 2006]: dual-Vdd, MILP
  - [S. Liu et al. ASPDAC 2010]: heuristic; MIS based
  - NP-hard
  - Easily trapped in local optimum
- **In previous works**:
  - A few works consider simultaneous Retiming and Slack Budgeting
  - MILP method or heuristic method
- **In our works**:
  - Network-Based Algorithm
  - Speedup

Contribution of Our Algorithm

- Formulate retiming and slack budgeting problem as an Integer Linear Programming (ILP) problem.
- Transform the problem to the convex cost dual network flow problem with just a little loss of optimality.
- Experimental results show the efficiency of our approach which can not only achieve more reduction of power consumption but also get speedup over previous works.

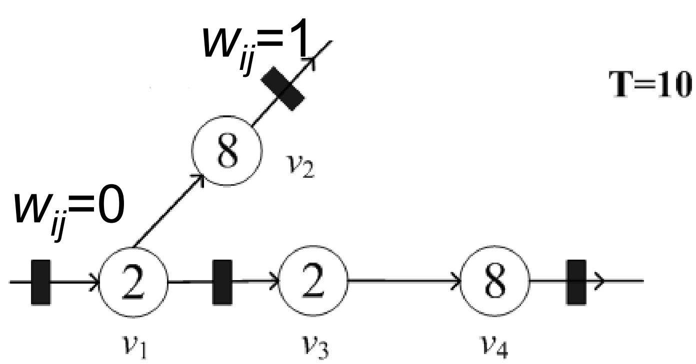
Power Slack Curve

- Each gate i is given k discrete slack levels
  - Each point is connected to its neighboring point(s) by a linear segment.
  - Power Slack Curve is a convex decreasing function.



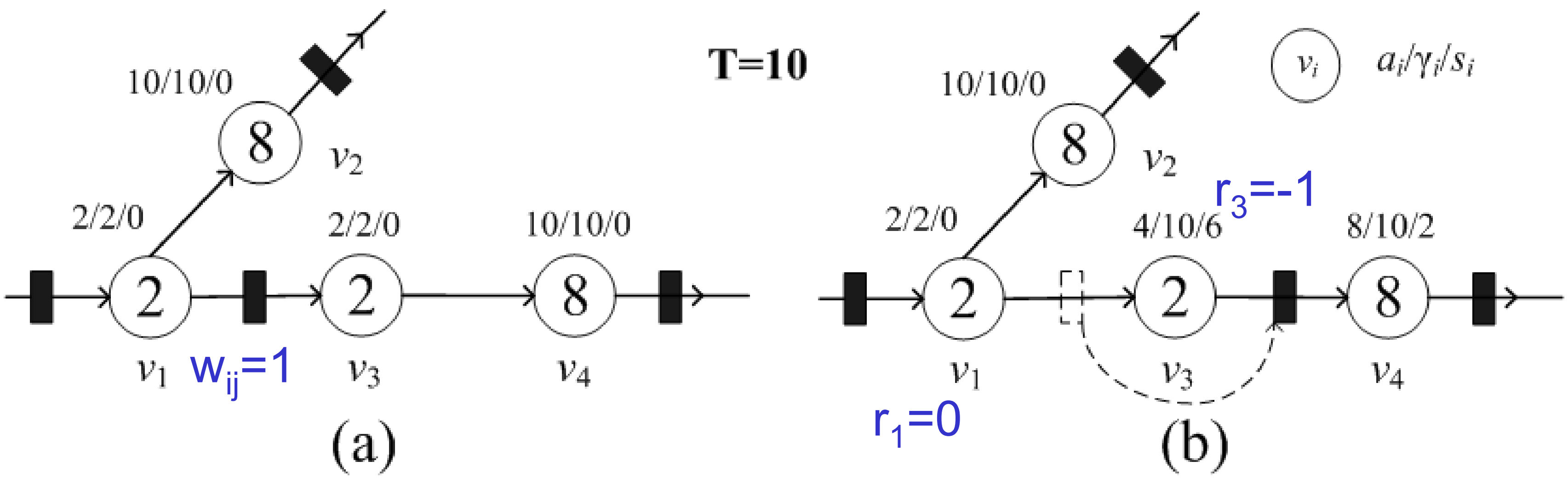
Problem Definition

- **Input**:
  - Directed graph G = (V; E; d; w) as synchronous sequential circuit.
    - $i \in V$: combinational gate
    - $e_{ij} \in E$: signal passing from gate i to j
    - d_i : delay of gate i
    - w_{ij} : number of FF on edge e_{ij}
  - Period constraint T
  - Power-slack tradeoff for each slack level
- **Output**: reallocation represented by r , so
  - Minimize power consumption
  - Under the period constraint



Labels Definition

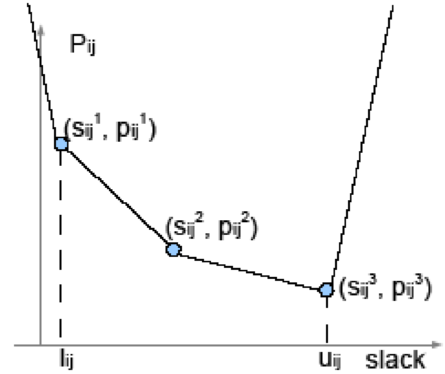
- Three non-negative labels, $a_i/\gamma_i/s_i$: $V \rightarrow R^*$
  - the latest arrival time, require time, and slack of gate i.
- An integer label $r : V \rightarrow Z^*$
  - the number of FFs are moved from the outgoing edges to the incoming edges of each node
  - when r is negative value, it means FFs are moved back from the incoming edges to outgoing edges
  - the number of FFs on edge (v_i, v_j) is $w_{ij} + r_j - r_i$



MILP formulation

- Power consumption
  - Power Slack Curve

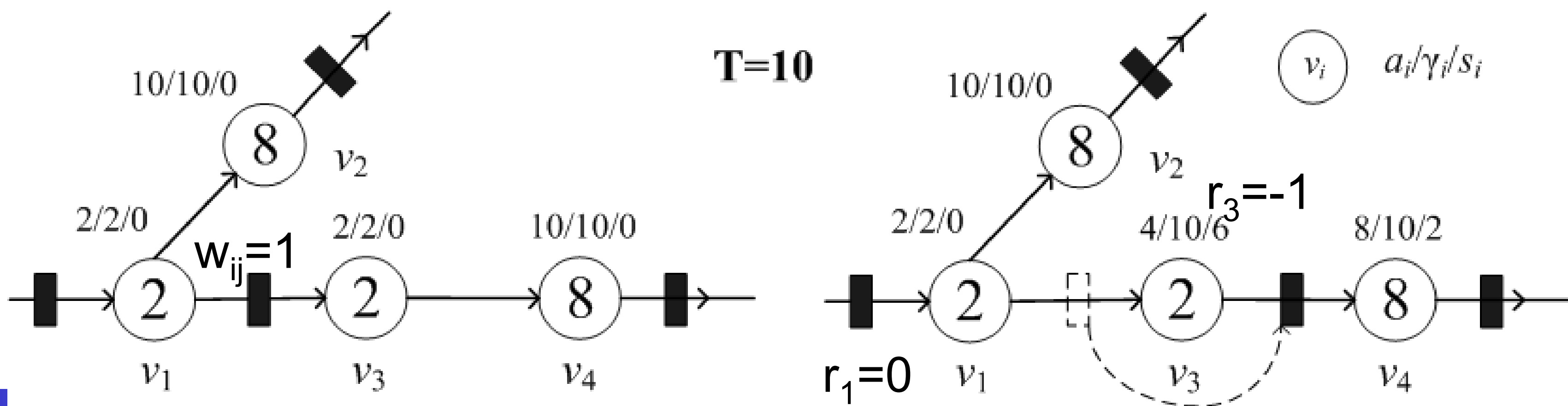
$$\min \sum_{i \in V} P(s_i)$$



- Retiming constraints
  - The number of FFs on edge (v_i, v_j): $w_{ij} + r_j - r_i \geq 0$

- Period constraint T

$$a_i \geq d_i + s_i \qquad \forall i \in V$$
$$a_i \leq T \qquad \forall i \in V$$
$$a_j \geq a_i + d_i + s_i \quad if\ r_i - r_j = w_{ij}$$



MILP formulation

Condition for $\Phi(G) \le T$

$$\begin{array}{rl} a_i \ge d_i + s_i & \forall i \in V \\ a_i \le T & \forall i \in V \\ r_i - r_j \le w_{ij} & \forall (i,j) \in E \\ a_j \ge a_i + d_i + s_i & \text{if } r_i - r_j = w_{ij} \end{array}$$

Suppose $R_i = r_i + a_i / T$
$\Rightarrow a_i = T \cdot R_i - T \cdot r_i.$

$$\begin{array}{llll}
\min & \sum_{i \in V} P(\bar{s}_i) & & (II) \\
\text{s.t.} & \bar{R}_i - \bar{r}_i \ge \bar{s}_i & \forall i \in V & (IIa) \\
 & \bar{R}_i - \bar{r}_i \le T & \forall i \in V & (IIb) \\
 & \bar{r}_j - \bar{r}_i \ge -T \cdot w_{ij} & \forall (i,j) \in E & (IIc) \\
 & 0 \le \bar{R}_i, \bar{r}_i \le \bar{N}_{ff} & \forall i \in V & (IId) \\
 & \bar{s}_i = \{\bar{s}_i^1, \cdots, \bar{s}_i^k\} & \forall i \in V & (IIe) \\
 & 0 \le \bar{s}_i \le T & \forall i \in V & (IIf) \\
 & \bar{R}_j - \bar{R}_i \ge t_{ij} & \forall (i,j) \in E & (IIg) \\
 & t_{ij} \ge \bar{s}_j - T \cdot w_{ij} & \forall (i,j) \in E & (IIh)
\end{array}$$

t_{ij} is an introduced intermediate variable which connects (IIg) and (IIh)

$\bar{N}_{ff} = N_{ff} \cdot T$ where N_{ff} is the constant of upper limit for both R_i and r_i

$\bar{R}_i = R_i \cdot T$ $\quad \bar{s}_i = d_i + s_i$ $\quad \bar{r}_i = r_i \cdot T$ $\quad \bar{s}_i^j = s_i^j + d_i$

MILP formulation

- Solved by ILP Solver, but unacceptable runtime
- Need more effective method
- Without two constraints, convex cost dual network algorithm [R. K.Ahuja et al. 2003]
- Removes constraint (IIh), add penalty function P(tij):
- Generate new problem (III)
- Given solutions of (III), heuristic generate solution of (II):

$$
\begin{array}{llll}
\min & \sum_{i \in V} P(\bar{s}_i) & & (II) \\
\text{s.t.} & \bar{R}_i - \bar{r}_i \geq \bar{s}_i & \forall i \in V & (IIa) \\
& \bar{R}_i - \bar{r}_i \leq T & \forall i \in V & (IIb) \\
& \bar{r}_j - \bar{r}_i \geq -T \cdot w_{ij} & \forall (i,j) \in E & (IIc) \\
& 0 \leq \bar{R}_i, \bar{r}_i \leq \bar{N}_{ff} & \forall i \in V & (IId) \\
& \bar{s}_i = \{\bar{s}_i^1, \cdots, \bar{s}_i^k\} & \forall i \in V & (IIe) \\
& 0 \leq \bar{s}_i \leq T & \forall i \in V & (IIf) \\
& \bar{R}_j - \bar{R}_i \geq t_{ij} & \forall (i,j) \in E & (IIg) \\
& t_{ij} \geq \bar{s}_j - T \cdot w_{ij} & \forall (i,j) \in E & (IIh)
\end{array}
$$

$$
\begin{array}{lll}
\min & \sum_{i \in V} P(\bar{s}_i) + \sum_{(i,j) \in E} P(t_{ij}) & (III) \\
\text{s.t.} & (IIa) - (IIg) & \\
& t_{ij} \geq -T \cdot w_{ij}, \quad \forall (i,j) \in E &
\end{array}
$$

convex cost dual network flow problem?

$$\bar{s}_j = \min(t_{ij} + T \cdot w_{ij}, \bar{s}_j), \forall i \in FI(j)$$

Remove Redundant Constraint

- Remove constraint (IIb)
- Consider new problem (III'), which replaces (IIa) and (IIb) by $\bar{R}_i - \bar{r}_i = \bar{s}_i$

$$
\begin{aligned}
\min \quad & \sum_{i \in V} P(\bar{s}_i) && (II) \\
\text{s.t.} \quad & \boxed{\begin{aligned} & \bar{R}_i - \bar{r}_i \geq \bar{s}_i && \forall i \in V && (IIa) \\ & \bar{R}_i - \bar{r}_i \leq T && \forall i \in V && (IIb) \end{aligned}} \\
& \bar{r}_j - \bar{r}_i \geq -T \cdot w_{ij} && \forall (i,j) \in E && (IIc) \\
& 0 \leq \bar{R}_i, \bar{r}_i \leq \bar{N}_{ff} && \forall i \in V && (IId) \\
& \bar{s}_i = \{\bar{s}_i^1, \cdots, \bar{s}_i^k\} && \forall i \in V && (IIe) \\
& 0 \leq \bar{s}_i \leq T && \forall i \in V && (IIf) \\
& \bar{R}_j - \bar{R}_i \geq t_{ij} && \forall (i,j) \in E && (IIg) \\
& t_{ij} \geq \bar{s}_j - T \cdot w_{ij} && \forall (i,j) \in E && (IIh)
\end{aligned}
$$

$$
\begin{aligned}
\min \quad & \sum_{i \in V} P(\bar{s}_i) + \sum_{(i,j) \in E} P(t_{ij}) && (III) \\
\text{s.t.} \quad & (IIa) - (IIg) \\
& t_{ij} \geq -T \cdot w_{ij}, \quad \forall (i,j) \in E
\end{aligned}
$$

$$
\begin{aligned}
\min \quad & \sum_{i \in V} Q(\bar{s}_i) + \sum_{(i,j) \in E} P(t_{ij}) && (III') \\
\text{s.t.} \quad & (IIc) - (IIg) \\
& \bar{R}_i - \bar{r}_i = \bar{s}_i && \forall i \in V \\
& t_{ij} \geq -T \cdot w_{ij} && \forall (i,j) \in E
\end{aligned}
$$

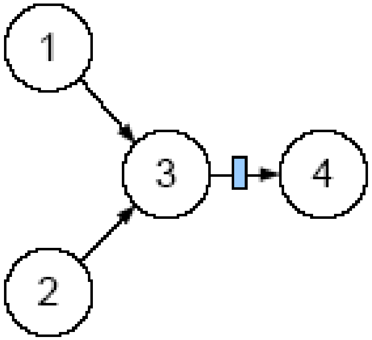
$$
Q(\bar{s}_i) = \begin{cases} P(\bar{s}_i^*) & if \quad \bar{s}_i \leq s_i^* \\ P(\bar{s}_i) & if \quad \bar{s}_i > s_i^* \end{cases}
$$

Theorem 1

For every optimal solution $(\bar{R}, \bar{r}, \bar{s})$ of problem (III), there is an optimal solution $(\bar{R}, \bar{r}, \hat{s})$ of problem (III'), and the converse also holds.

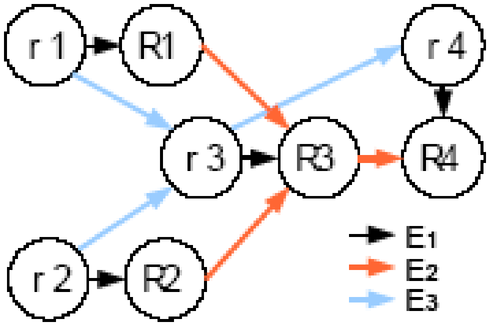
Primal Network Flow Problem

- Solve problem (III) by Convex Cost Dual Flow:
  - Transformation to Primal Network Flow Problem
    - Split vertex i into two vertex $\bar{r}_i$ and $\bar{R}_i$
    - $(\bar{r}_i, \bar{R}_i) \in \bar{E}_1$, $(\bar{R}_i, \bar{R}_j) \in \bar{E}_2$, $(\bar{r}_i, \bar{r}_j) \in \bar{E}_3$
  - Further simplify problem as follow:



$$\min \sum_{(i,j)\in\bar{E}} P(s_{ij}) \qquad (IV)$$

$$\begin{aligned} \text{s.t.} \quad & \mu_j - \mu_i \geq s_{ij} && \forall (i,j) \in \bar{E} && (IVa) \\ & 0 \leq \mu_i \leq \bar{N}_{ff} && \forall i \in \bar{V} && (IVb) \\ & l_{ij} \leq s_{ij} \leq u_{ij} && \forall (i,j) \in \bar{E} && (IVc) \end{aligned}$$



$$\min \sum_{(i,j)\in \bar{E}} P(s_{ij}) \qquad (IV)$$

$$\text{s.t.} \quad \mu_j - \mu_i \geq s_{ij} \qquad \forall (i,j) \in \bar{E} \qquad (IVa)$$

$$\boxed{\begin{array}{lll} 0 \leq \mu_i \leq \bar{N}_{ff} & \forall i \in \bar{V} & (IVb) \\ l_{ij} \leq s_{ij} \leq u_{ij} & \forall (i,j) \in \bar{E} & (IVc) \end{array}}$$

Primal Network Flow Problem

$$\min \sum_{(i,j)\in \bar{E}} \bar{P}(s_{ij}) + \sum_{i\in \bar{V}} B(\mu_i)$$

$$\text{s.t.} \quad \mu_j - \mu_i \geq s_{ij} \qquad \forall (i,j) \in \bar{E}$$

- Remove constraints
- P(sij) can be modified to eliminate the bounds on s_{ij}

$$\bar{P}(s_{ij}) = \begin{cases} P(u_{ij}) + M(s_{ij} - u_{ij}) & \bar{s}_{ij} > u_{ij} \\ P(s_{ij}) & 0 \leq \bar{s}_i \leq T \\ P(l_{ij}) - M(s_{ij} - l_{ij}) & \bar{s}_{ij} < l_{ij} \end{cases}$$

- The bounds on μ_i can also be eliminated by adding into objective a convex cost function $B(\mu_i)$

$$B(\mu_i) = \begin{cases} M \cdot (\mu_i - \bar{N}_{ff}) & if \quad \mu_i > \bar{N}_{ff} \\ 0 & if \quad 0 \leq \bar{\mu}_i \leq \bar{N}_{ff} \\ -M \cdot \mu_i & if \quad \mu_i < 0 \end{cases}$$

Lagrangian Relaxation

- Step 2: Lagrangian Relaxation
  - Lagrangian relaxation to eliminate constraints
  - Lagrangian sub-problem:

$$L(\vec{x}) = \min \sum_{e(i,j)\in \bar{E}} [P(s_{ij}) + x_{ij}s_{ij}] + \sum_{i\in \bar{V}} [B_i(\mu_i) + x_{0i}\mu_i]$$

  - Introduce start node v0
  - Final Lagrangian subproblem:

$$L(\vec{x}) = \min \sum_{e(i,j)\in E} [P_{ij}(s_{ij}) + x_{ij}s_{ij}]$$

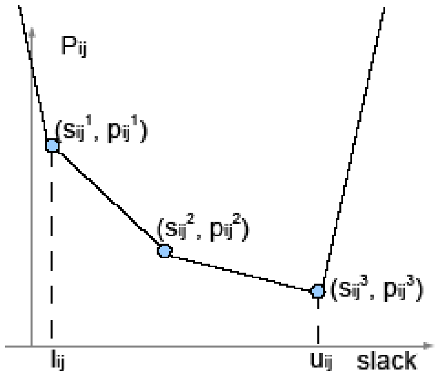
$$\text{s.t.} \sum_{j:e(i,j)\in E} x_{ij} - \sum_{j:e(j,i)\in E} x_{ji} = 0 \quad \forall i \in V$$

$$x_{ij} \geq 0 \qquad \forall (i,j) \in E_1 \cup E_2 \cup E_3$$

Convex Cost-scaling Approach

- Step 3: Convex Cost-scaling Approach
  - Define function $H_{ij}(x_{ij}) = \min_{s_{ij}}\{P_{ij}(s_{ij}) + x_{ij}s_{ij}\}$
  - Hij (xij) is concave, define convex function

$$C_{ij}(x_{ij}) = -H_{ij}(x_{ij})$$



  - Final problem is a min-cost flow problem

$$L(\vec{x}) = \min \sum_{e(i,j)\in E} C_{ij}(x_{ij})$$

$$\text{s.t.} \sum_{j:e(i,j)\in E} x_{ij} - \sum_{j:e(j,i)\in E} x_{ji} = 0 \quad \forall i \in V$$

$$0 \le x_{ij} \le M \qquad \forall (i,j) \in E_1 \cup E_2 \cup E_3$$

$$-M \le x_{ij} \le M \qquad \forall (i,j) \in E_4$$

  - For optimal flow x, construct residual network G(x)
  - In G(x), solve shortest path distance d(i)
  - Apply (i) = d(i) and sij = (i)
  - Final solve the problem!!

Experiments Setup

- Implemented in C++
- 3.0GHz CPU and 6GB Memory
- 19 cases from the ISCAS89

TABLE I
The Characteristics of Test cases

Test case name	*Gates count*	*Edges count*	*Max outputs*	*Max inputs*	T_{min}
s27.test	11	19	4	2	20
s208.1.test	105	182	19	4	28
s298.test	120	250	13	6	24
s382.test	159	312	21	6	44
s386.test	160	354	36	7	64
s344.test	161	280	12	11	46
s349.test	162	284	12	11	46
s444.test	182	358	22	6	46
s526.test	194	451	13	6	42
s526n.test	195	451	13	6	42
s510.test	212	431	28	7	50
s420.1.test	219	384	31	4	40
s832.test	288	788	107	19	98
s820.test	290	776	106	19	92
s641.test	380	563	35	24	238
s713.test	394	614	35	23	262
s838.1.test	447	788	55	4	80
s1238.test	509	1055	192	14	110
s1488.test	654	1406	56	19	166

Experimental results

TABLE II: Comparisons with Optimal ILP and Previous Work [18]

Benchmark	T	Power Consumption			Total Slacks			Runtime(s)		
		Optimal ILP	[18]	ours	Optimal ILP	[18]	ours	Optimal ILP	[18]	ours
s27.test	20	800	824	850	40	40	30	0.02	0.0	0.0
s208.1.test	28	3542	9118	4772	1770	290	1988	0.39	0.44	0.06
s298.test	24	6498	8888	8010	1330	660	1240	0.78	0.69	0.07
s382.test	44	6456	9038	9958	3011	2071	1895	>1000	10.56	0.12
s386.test	64	8836	12870	9564	2484	807	2324	4.58	1.03	0.1
s344.test	46	9876	11848	9894	1855	1064	1760	0.82	2.53	0.09
s349.test	46	9938	12472	9894	1852	912	1780	0.79	4.49	0.11
s444.test	46	8938	14032	11884	2962	1025	1939	>1000	12.04	0.12
s526.test	42	7602	14106	11498	3626	1307	2356	42.57	1.67	0.17
s526n.test	42	7752	11734	11548	3616	2089	2366	30.32	4.72	0.17
s510.test	42	13976	17492	14846	2237	937	2040	>1000	1.62	0.17
s420.1.test	50	4574	17920	9224	5906	1050	4466	1.29	16.91	0.14
s832.test	98	13652	14518	16274	5175	4525	4171	71.96	151.26	0.24
s820.test	92	13552	17694	16448	5261	3493	4103	68.98	13.18	0.25
s641.test	238	13334	20408	14424	7925	6067	7604	2.24	92.97	0.26
s713.test	262	13018	21228	14322	8522	6363	8112	2.27	121.1	0.27
s838.1.test	80	6004	18898	17556	14048	9016	9912	1.48	256.9	0.4
s1238.test	110	6096	10444	8208	16764	14635	15792	0.23	448.6	0.34
s1488.test	166	21292	23799	27836	15313	14791	13024	>1000	670.7	0.53
Avg	-	9249.3	14070	11947.9	5457.7	3744.3	4573.8	-	95.3	0.19
Diff	-	1	+52%	+29%	1	-31%	-16%	-	1	0.002

[18] S.Liu et al., "Simultaneous slack budgeting and retiming for synchronous circuits optimization", ASPDAC 2010

Conclusion

- The retiming and slack budgeting problem can be simultaneously solved by formulating the problem to a convex cost dual network flow problem.
- With the power slack curve, the discrete slack budgeting can be solved by convex cost dual network flow formulation.
- Both the theoretical analysis and experimental results show the efficiency of our approach
  - Reduce power consumption by 15%
  - 500 times speedup.

Thank You