FULL-CHIP INTERCONNECT POWER ESTIMATION AND SIMULATION CONSIDERING CONCURRENT REPEATER AND FLIP-FLOP INSERTION

Weiping Liao and Lei He

Electrical Engineering Department
University of California, Los Angeles, CA 90095
{wliao, lhe}@ee.ucla.edu

Abstract

In this paper, we study the full-chip interconnect power modeling. We show that repeater insertion is no longer sufficient to achieve the target frequencies specified by ITRS, and develop concurrent repeater and FF insertion schemes. Considering structural interconnects, layer assignment and concurrent repeater and FF insertion for delay specification, we develop a cycle-accurate microarchitecture-level interconnect power simulation. The simulation reduces the over-estimation by up to 2.46X compared to power estimation based on purely stochastic interconnects and fixed switching factor. Furthermore, we show that interconnect pipelining has a lower IPC but can improve throughput by up to 2.03X. This indicates that the traditional design flow optimizing IPC and clock frequency *separately* may no longer be valid.

1. INTRODUCTION

Integrated circuit designers always have the desire to increase circuit speed for better performance. The leading high-performance processors have reached clock frequencies above 3GHz. As the system delay is dominated by the interconnect delay, an increasing number of repeaters and flip-flops (FFs) are used to reduce the interconnect delay [1, 2, 8]. Consequently, the power consumed by interconnects including repeaters and FFs gains a growing significance in the total system power. Because the power dissipation has become a primary design constraint, a number of microarchitecture level power simulators [3, 4, 6] are developed to estimate power and verify power-reduction innovations at the early design stage. All these simulators, however, do not explicitly characterize the power consumed by interconnects including repeaters and FFs.

In this paper, we study full-chip level interconnect power modeling and reduction. We consider structural interconnects, layer assignment, and concurrent repeater and FF insertion for minimizing number of FFs (*min-FF* solution) or minimizing power (*min-power* solution). Related work in the literature include: [2] estimates the power for global interconnect repeater insertion based on the stochastic wire length distribution [7], and studies delay-power trade-off for minimizing repeater power. Over-simplified repeater model (i.e., single-model to be defined in Section 2) is used and no FF insertion is considered. [8] estimates the number of repeaters and FFs with consideration of routing tree topology. Its repeater model is the same as the model in [2], and its FF insertion is in essence equivalent to our min-FF solution. Different from our paper, none of the above two papers considers a min-power solution for concurrent repeater and FF insertion, and neither do they study interconnect layer assignment or cycle-accurate microarchitecture-level interconnect power simulation. Furthermore, no leakage power is considered in [2, 8].

The rest of this paper is organized as follows. In Section 2, we study repeater and FF insertion for individual wires. In Section 3, we apply concurrent repeater and FF insertion to the full-chip level considering random interconnects. In Section 4, we study micro-architectural level interconnect power estimation and cycle-accurate power simulation. We conclude in Section 5.

2. REPEATER AND FLIP-FLOP INSERTION

2.1. Interconnect and Device Models

In this paper, we model interconnects by the Π-type distributed RC circuit, and consider multiple interconnect layers. Top layers are used for wide and long global interconnects, and bottom layers are used for short local interconnects. Between them are the layers for intermediate interconnects. For the simplicity of presentation, we assume all wires are global wires in this section, and define the distinction of global and non-global wires in Sections 3 and 4. We assume that a unit length interconnect has resistance R_w and capacitance C_w, and model an inverter by its gate capacitance, drain capacitance and its effective resistance. We represent the gate, drain capacitances and effective output resistance for a minimum size inverter as C_0, C_p and R_0, respectively. A repeater can be a single inverter, or a cascaded inverters chain.

We use Elmore delay to calculate interconnect delay, i.e.

$$T_d = \sum_i R_i \cdot C_{down} \quad (1)$$

where T_d is the total delay, R_i is the resistance of a wire segment and C_{down} is the sum of downstream capacitance of R_i. We consider interconnect power including dynamic power and leakage power given by Equation (2) and (3), respectively:

$$P_{dynamic} = \frac{1}{2}\alpha V_{DD}^2 f_{clk} \cdot (S(C_0 + C_p) + lC_w + N_F \cdot C_F) \quad (2)$$

$$P_{leakage} = V_{DD} I_{off} (S + N_F \cdot S_F) \quad (3)$$

where f_{clk} is the clock frequency, l is the wire length, α is the switching factor, I_{off} is the unit leakage current, and S is the total inverter size. Furthermore, N_F is the total number of FFs, C_F is

This research is partially supported by NSF CAREER Award CCR-0093273, SRC contract 1008, and Hewlett-Packard. We used computers donated by Intel. Address comments to lhe@ee.ucla.edu.

ICCAD'03, November 11-13, 2003, San Jose, California, USA.

the total capacitance of one FF, and S_F is the total gate size of one FF. We assume 100nm technology in this paper, with parameters in Table 1, where the wire widths and heights are obtained from ITRS roadmap[1], C_w and R_w are calculated by Berkeley Predictive Technology Model [9], the I_{off} is from [10], the α is 0.15 [12] and is fixed for logic and interconnects except the structure interconnects with cycle-accurate power simulation in Section 4.2, and the other values are obtained from SPICE simulations.

Technology		100nm	
I_{off}(uA/u)		6.33	
α		0.15	
Minimum size inverter	R_0(KΩ)	12.0	
	C_0(fF)	0.91	
	C_p(fF)	0.75	
FF	C_F (fF)	16.6	
	S_F	10	
Interconnects			
	Global	Intermediate	Local
Width (nm)	335	160	122.5
Height (nm)	670	272	196
R_W (KΩ/m)	89.106	459.559	832.986
C_W (pF/m)	204.802	180.068	176.188

Table 1: Technology parameters.

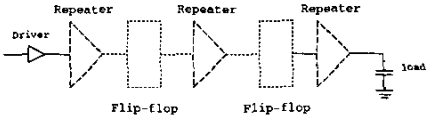

Figure 1: The repeater and FF insertion problem in two-pin nets.

In this paper, we assume all interconnects are two-pin nets. This assumption has been used widely in the literature for high-level estimation [2, 7]. Specifically, as shown in Figure 1, we assume every interconnect has one driver and one load. Both driver and load are inverters with the $4X$ minimum inverter size. We study the repeater and FF insertion for two objective functions: one is to meet the delay target with minimum number of FFs, or *min-FF*; and the other is to meet the delay target with minimum total interconnect power consumption, or *min-power*.

2.2. Min-FF Solution

It has been assumed in [2, 10] that for repeater insertion, the input capacitance C_{in} and effective resistance for each repeater are equal to $S * C_0$ and $\frac{R_0}{S}$ respectively, where S is the size of the repeater. Under this assumption, each repeater is a single inverter, named as *single* model. To drive a large load, a repeater may contain a chain of cascaded inverters, where C_{in} of a repeater is equal to C_0 times the size of first inverter in the inverter chain. We call this type of repeater as *cascaded* repeater. An inverter in a cascaded repeat is a *stage*, and the size ratio between two consecutive inverters is the *stage ratio*. In addition, we also consider a *hybrid* model where the first stage is a chain of cascaded inverters, but the rest stages are single inverters. The hybrid model may lead to a good solution when the inverter in the last stage of first repeater is large enough to drive the rest single repeaters. We illustrate the three repeater insertion models in Figure 2.

[1]Note the width and height of global wires are from 130nm technology as we assume the global interconnects do not have to scaled [1].

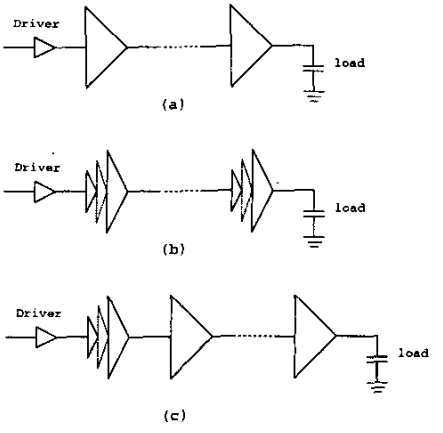

Figure 2: The three models for repeater insertion: (a) single model; (b) cascaded model; and (c) hybrid model.

We study the power optimization problem under given delay target for interconnects. The existing analytical repeater insertion methods [2, 10] can only be used for single model. We find the solution by the following enumeration. For cascaded model, we enumerate the number of repeaters, the first inverter size, the uniform stage ratio and the stage number for each repeater. Again, we assume that all repeaters are identical. For hybrid model, we enumerate the number of repeaters, the design of first cascaded repeater, and the uniform design of the rest of repeaters using the single model. For each combination, we calculate the delay and power. If the delay is smaller than our delay target, we call this combination as a *valid solution*. We choose the valid solution with the smallest number of FFs. If there are more than one valid solution, we choose the one with lowest power consumption. We also do pruning during enumeration. If we have obtained a valid solution with repeater size S, all solutions with repeater size greater than S should be skipped because they definitely consume more power. If a wire is too long to meet the delay target, we insert FFs to break the wires into shorter wires. If we have obtained valid solutions for one wire, we can reuse the solutions for wires of a same length.

Clock	Longest wire to meet the delay target (mm)		
GHz	Single model	Cascaded model	Hybrid model
1	9.24	10.33	13.12
2	3.88	4.54	5.95
3	1.78	2.84	3.56

Table 3: The longest wire that repeater insertion alone is able to meet the delay target without FF insertion. The delay target is 80% of the clock period.

Table 2 shows our experiment results from all three models

(GHz)	Wire length (mm)	Min-FF power (mW)									Min-power power (mW)			Min-power vs. Min-FF Redu %
		Single model			Cascaded Model			Hybrid Model						
		Power	FF	R#	Power	FF	R#	Power	FF	R#	Power	FF	R#	
1	**4**	**0.1082**	**0**	**1**	**0.1082**	**0**	**1**	**0.1082**	**0**	**1**	0.1082	0	1	0
	8	**0.3347**	**0**	**3**	**0.3309**	**0**	**2**	**0.2868**	**0**	**2**	0.2186	1	2	34.69
	10	0.2797	1	2	0.6972	0	3	0.4578	0	3	0.2735	2	3	2.22
2	4	0.2283	1	1	0.2877	0	1	0.2877	0	1	0.2283	1	2	0
	8	0.4701	2	3	0.5797	1	2	0.5797	1	2	0.4611	3	4	1.91
	10	0.7867	2	6	0.6458	2	3	0.8628	1	4	0.5775	4	5	26.59
3	4	0.3682	2	3	0.4070	1	2	0.4070	1	2	0.3682	2	3	0
	8	0.7520	4	5	1.1163	2	3	1.0579	2	6	0.7431	5	6	1.18
	10	0.9537	5	6	1.1402	3	4	1.6888	2	6	0.9269	8	9	2.81

Table 2: The power consumption for different wire lengths and different clock frequencies, under three models for repeater insertion. The symbol "R#" means the number of repeaters. For 1GHz and 4mm-long wire, the three models achieve a same solution.

we discuss above. We use the wire lengths 4mm, 8mm, and 1cm, and clock frequencies 1GHz, 2GHz and 3GHz. We assume that the delay target is 80% clock period. No FF insertion is needed for wires up to 8mm and 1GHz clock (see **highlights** in the table). In this case, the hybrid model achieves 14.31% power reduction compared with the single model for the 8mm wire under 1GHz system clock. The hybrid model also has the smallest number of FFs for the same wire and that delay target. This can be further illustrated in Table 3. For target delay, the longest wire without FF insertion in the hybrid model can be $1.5X$ of that in the single model.

2.3. Min-power Solution

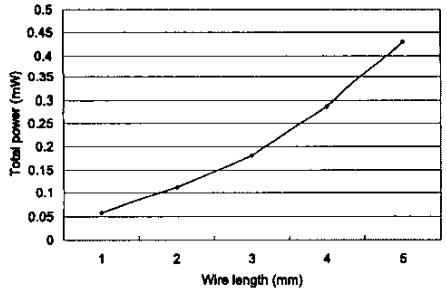

Figure 3: The power consumption with repeater insertion vs. different wire lengths. We choose 100nm technology and the delay target as 400 ps. Only the results from the hybrid model are shown.

Although the hybrid model provides the better power consumption for the same wire length, FF number and clock frequency, we also observe from the Table 2 that the single model with more FFs actually has lower power consumption than the hybrid model with fewer FFs. The reason is that for all repeater insertion models, the resulting power consumption is super-linear w.r.t the wire length as shown in Figure 3, where the wire length increases by $4X$ from 1mm to 4mm, the power consumption increases by more than $10X$. It is easy to see that instead of inserting FF merely to meet the delay target, we can reduce power by aggressively inserting more FFs. Figure 4 shows the power for different wire lengths

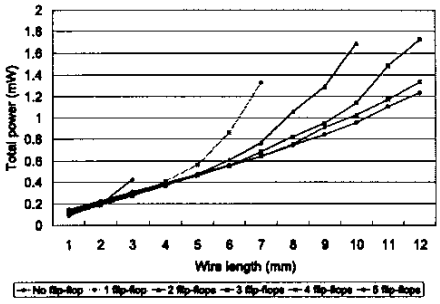

Figure 4: The power for different wire lengths, under different FF insertion. The delay target is 240ps corresponding to 80% of clock period for a 3GHz system clock. Other settings are the same as those in Figure 3.

for same target delay but different numbers of FFs. According to the Figure 4, when enough FFs are inserted, the power curve becomes nearly linear with respect to wire length. On the other hand, FF insertion is not always beneficial. The more FFs inserted, the more power consumed by the FFs. There exists a point where the extra power consumed by FFs outweighs the power saving by FF insertion, i.e. there is an optimal number of FF to be inserted for minimal power consumption.

The min-power solution finds the concurrent repeater and FF insertion with the minimum power and with delay less than the delay target. Again we use enumeration to find the min-power solution. We enumerate a range of reasonable FF numbers. For each number, we find the repeater insertion solution as discussed before. Finally, we choose the solution with the minimum total power. We present the results under min-power FF insertion and hybrid repeater model in Table 2. The min-power method can reduce the interconnect power by up to 34.69% compared with the min-FF method.

2.4. Runtime Reduction

In our implementations, we use table lookup for concurrent repeater and FF insertion solution since there is no closed-form solution. Tables are built for each interconnect length and clock fre-

quency. Each table entry contains the concurrent repeater and FF insertion solution and the optimal power. With table lookup we can greatly reduce runtime and speed up our calculation for full-chip interconnect power in Section 3 and 4.

3. FULL-CHIP POWER ESTIMATION

In this section we study the full-chip interconnect power estimation. As we have already seen in the last section, the hybrid model achieves the largest power reduction and least number of FFs compared with the other two models. In the rest of this paper, we only use hybrid model for interconnect power estimation unless specified otherwise.

3.1. Layer Assignment

Because wires in top layers have a smaller RC constant that helps to reduce delay, it is beneficial to assign long interconnects to top layers. We have developed the following *layer assignment*: we assume that all global interconnects are accommodated by 50% of the total area of the top two layers (and the rest area in the top layers is used by power/ground and clock routing), and use Equation (4) to solve the minimum length of global interconnects l_{gmin}:

$$2 \times 50\% \times Chip\_size = \int_{l_{gmin}}^{l_{max}} Global\_pitch\_width \cdot l \cdot i(l) dl \quad (4)$$

where l_{max} is the maximum length of interconnects and it is $2\sqrt{N}$ with N being the total number of gates on the chip, $i(l)$ is the length density function. l_{gmin} can be used as the length boundary between global and intermediate interconnects.

We find the length boundary between intermediate and local interconnects l_{mmin} by the following Equation (5):

$$2 \times Layer\_Pairs \times Chip\_size = \int_{l_{mmin}}^{l_{gmin}} intermediate\_pitch\_width \cdot l \cdot i(l) dl \quad (5)$$

where $Layer\_Pairs$ is the number of pairs of intermediate layers, and the area utilization rate is 100% for the intermediate layer. We keep increasing $Layer\_Pairs$ until the interconnects with the length of l_{mmin} can meet the delay target without repeater insertion. Interconnects with length less than l_{mmin} are local interconnects and are assigned to local layers.

3.2. Power Estimation

In this section we study full-chip interconnect power estimation. We obtain the chip size from ITRS and assume the chip area for random logic by subtracting cache area from the total chip area. We use the length density function $i(l)$ from the stochastic length distribution methodology[7] to calculate the boundaries between local, intermediate and global interconnects in layer assignment. We set the length of one gate pitch as the square root of the logic gate area obtained from ITRS. The typical rent's exponent of 0.55 is used. The gate count, gate area, and gate pitch are shown in Table 4.

Technology		100nm
System clock		2GHz
Total gate count		14267075
Gate area		6.5 um^2
Gate pitch		2.55 um
Global and intermediate interconnect boundary	in gate pitch	1389
	in mm	3.542
Intermediate and local interconnect boundary	in gate pitch	85
	in mm	0.217

Table 4: The length boundaries decided by layer assignment with the gate pitch and system clock.

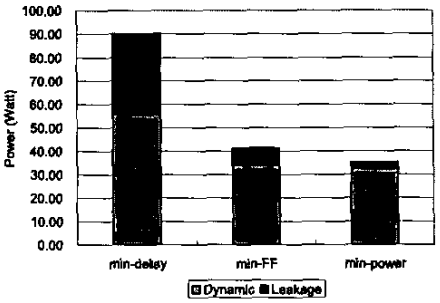

Figure 5: The full chip interconnect power for three repeater and FF insertion solutions. Note the result for *min-delay* is larger than results in [2] because we consider the power from drain capacitances, leakage power and intermediate interconnects that are all ignored by [2].

Solution	Number of equivalent repeaters	number of FFs
min-delay	45664369	N/A
min-FF	10022679	15908
min-power	3912611	158163

Table 5: The number of repeaters and FFs insertion in all three solutions in Figure 5. The min-delay solution does not use any FF insertion.

Figure 5 shows the full chip interconnect power calculated by the three different repeater and FF insertion solutions. In the first solution, repeaters are inserted for *minimum delay*, or *min-delay*, i.e., we insert repeaters as long as it can reduce delay and we do not insert any FF. The power reduction from the min-power method mainly comes from the reduced repeater area. We define one *equivalent repeater* as one minimum size inverter. A repeater with total size S can be mapped to S equivalent repeaters. For any repeater and FF insertion solution, the total power is decided by total wire capacitances, the number of equivalent repeaters and FFs. Table 5 shows the total number of equivalent repeaters and FFs for all three solutions. From Table 5 we can see that the min-FF and min-power solutions reduce the number of equivalent repeaters by $4.56X$ and $11.67X$, respectively. Although the number of FFs in min-power solution is almost $10X$ of that in min-FF solution, min-power solution still save 14.99% power as it reduce the number of

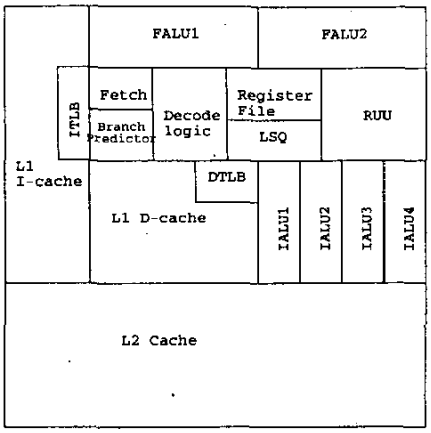

Figure 6: The floorplanning of the SuperScalar processors we study.

equivalent repeaters by 60.96%.

4. MICROARCHITECTURE-LEVEL POWER ESTIMATION AND CYCLE-ACCURATE SIMULATION

In this section we first develop the microarchitecture level interconnect power estimation that improves the power estimation in Section 3. We then develop the cycle-accurate interconnect power simulation, and show that there is a significant accuracy gap between power estimation and the accurate power simulation. We conclude that power simulation should be used for accurate power modeling.

4.1. Microarchitecture-level Power Estimation

Stochastic interconnect distribution is assumed in [2, 10, 8] and our Section 3. However, major components in a system-on-a-chip are often connected by varieties of busses that can be modeled accurately. We define two types of interconnects: *random* interconnects and *structural* interconnects. The random interconnects are interconnects inside each module and can be calculated by the same stochastic model as in Section 3.2. The structural interconnects are address and data busses between related modules, and their lengths are decided by the floorplan of the layout.

We consider high-performance SuperScalar processors, and summarize the configuration of the processor under study in Table 6. Considering gate counts for modules in Table 7 and wire length minimization between modules, we design the floorplan shown in Figure 6. We measure the lengths of busses according to the Manhattan distances between the centers of modules connected by the busses.The details about the bit-width and lengths for all busses can be found in the technical report [11].

The number of long interconnects are reduced with the introduction of structural interconnects. Therefore, we need to recalculate the overall wire length distribution and layer assignment. The interconnect density function $i(l)$ for a system is now the sum of all interconnect density functions among all modules and busses, given by Equation (6)

Parameter	Value
Processor Core	
RUU size	64 instructions
LSQ size	32 instructions
Fetch Queue size	8 instructions
Fetch width	4 instructions/cycle
Decode width	4 instructions/cycle
Issue width	4 instructions/cycle
Commit width	4 instructions/cycle
Functional Units	3 integer ALUs, 1 integer multiply/divide, 1 FP add, 1 FP multiply/divide
Branch Predictor	Combined, Bimodal 4K table 2-Level 1K table, 10-bit history 4K chooser
Memory Hierarchy	
L1 instruction-cache	64K, 4-way (LRU) 32B blocks, 1-cycle latency
L1 data-cache	64K, 4-way (LRU) 32B blocks, 1-cycle latency
L2	Unified, 2M, 8-way (LRU) 64B blocks, 12-cycle latency
TLB	128 entry, fully associative 30-cycle miss latency

Table 6: The configuration of the SuperScalar processors we simulate.

		Min-FF power		Min-power power	
		Total	Dynamic	Total	Dynamic
Power	*go*	22.11	19.78	21.35	19.72
	li	17.28	14.95	16.53	14.91
	perl	13.65	11.32	12.91	11.28
	compress	19.18	16.85	18.43	16.80
	swim	18.80	16.46	18.05	16.42
	mgrid	16.99	14.66	16.25	14.62
	fpppp	13.17	10.84	12.40	10.80
	applu	15.16	12.83	14.42	12.79
Geometry mean		16.92	14.55	16.15	14.50
Power Estimation		33.56	31.22	31.7	30.07
Difference		1.98X	2.15X	1.96X	2.07X

Table 9: The interconnect power with both random and structural interconnects. Leakage power is omitted because it is not affected by clock gating. The unit of power is Watt.

$$i(l) = \sum_k i_k(l) \quad (6)$$

where subscript k iterates over all modules and busses. Using the same number of layers as in Table 4, the new length boundaries with consideration of structural interconnects are shown in table 10. Compared to Table 4, the boundaries for both global/intermediate and intermediate/local are reduced due to the reduced number of long interconnects. In other words, a higher portion of random interconnects can be assigned to the global and intermediate layers

Module	Microarchitecture	gate count	Min-FF (W)			Min-power (W)		
			Total	Dynamic	Leakage	Total	Dynamic	Leakage
Fetch	Fetch queue	241231	0.3025	0.2970	0.0055	0.3025	0.2970	0.0055
Decode	Decode logic	1227692	2.3127	2.2112	0.1015	2.3068	2.2114	0.0954
Branch	Branch Predictor	861538	5.4764	5.1480	0.3284	5.4034	5.1356	0.2678
Ruu	Register Update Unit	3446154	7.5120	7.0054	0.5066	7.3599	6.9754	0.3846
Lsq	Load/Store Queue	1598154	3.2002	3.0427	0.1574	3.1834	3.0414	0.1420
IALU[1-4]	One integer unit	861538	1.4934	1.4371	0.0563	1.4922	1.4374	0.0548
FALU[1-2]	One floating-point unit	1723077	3.5108	3.3322	0.1786	3.4887	3.3300	0.1587
	Sum:	14267075	31.7990	30.1172	1.6821	31.5022	30.0703	1.4319

Table 7: Modules and their correspondent microarchitecture, gate count, and power under different repeater and flip-flop insertion models. The caches and register files are not considered because they are purely memory array.

	Min-FF					Min-power				
	Total power	Dynamic	Leakage	Rep	FF	Total power	Dynamic	Leakage	Rep	FF
Only random	41.49	33.87	7.61	10022679	15908	35.27	32.31	2.97	3912611	158163
w/ structural	33.56	31.22	2.34	2219416	69	31.7	30.07	1.63	1969583	16179
Reduction	$1.24X$	$1.08X$	$3.25X$	$4.52X$	$230X$	$1.11X$	$1.08X$	$1.82X$	$1.99X$	$9.76X$

Table 8: Total interconnect power without gating. The "Rep" and "FF" represent the number of equivalent repeaters and FF, respectively. The power is in the unit of Watt.

for reduced delay and in turn reduced buffer numbers. This may help to reduce interconnect power.

Global and intermediate interconnect boundary	in gate pitch	499
	in mm	1.272
Intermediate and local interconnect boundary	in gate pitch	24
	in mm	0.061

Table 10: New interconnect length boundaries between local, intermediate and global wires, after we distinguish the structural interconnects and random interconnects.

Considering the new layer assignment, we apply the power estimation method based on the stochastic length distribution to each module independently and obtain the interconnect power for each module (see Table 7). We also apply concurrent repeater and FF insertion to obtain the interconnect power for busses. In Table 8, adding power for all modules and busses, we obtain the total interconnect power at the microarchitecture level and compare it with the full-chip interconnect power estimation from Section 3. Based on this table, not considering structural interconnects (as in [2, 10] and Section 3) over-estimates the interconnect power by $1.24X$ and $1.11X$ for the min-FF and min-power solutions, respectively. Part of the power reduction is due to reduced number of long interconnects, which in turn reduces the number of equivalent repeaters and FFs. The equivalent repeaters are reduced by $4.52X$ and $1.99X$ for min-FF and min-power solutions, respectively. Compared to the min-FF solution, the min-power solution uses slightly fewer repeaters. Only 69 FFs are needed to meet the delay constraint in the min-FF solution,[2] but 16179 FFs are used by the min-power solution for power reduction. With consideration of power used by FFs, the min-power solution reduces the full-chip interconnect power by 5.5% compared to the min-FF solution. Min-power solution actually provides us the lower bound of full-chip interconnect power, as min-power solution may greatly reduce IPC (instruction per cycle) and therefore is not necessarily used in practice.

[2]We assume 2GHz clock here. More FFs are needed by the min-FF solution under higher clock rates.

4.2. Cycle-accurate Power Simulation

To develop cycle-accurate interconnect power simulation, we further incorporate our interconnect power models with concurrent repeater and FF insertion in the *sim-outorder* simulator of SimpleScalar toolset [14]. We can perform the following cycle by cycle simulation: if a module is accessed, we count its active (dynamic + leakage) interconnect power, otherwise we only count its leakage power. On the other hand, for each bus, we count the number of bit-line transitions in every cycle. The dynamic power in that cycle equals to the number of transitions times the dynamic switching power per bus bit-line. Note the dynamic switching power is the full switching power ($\frac{1}{2}CV^2$) without the empirical fixed switching factor. The leakage power for each bus is always equal to the total number of bit lines times the leakage power per bus bit-line. By counting only leakage power for idle modules we *implicitly* consider clock gating.

We ran simulations for a variety of SPEC 95 benchmarks. During each simulation, the benchmark is first fast forwarded by 10 million instructions to avoid the startup effect, and is then simulated for 10 million instructions. Table 9 reports the total interconnect power obtained by cycle-accurate simulation. By applying clock gating, the interconnect power based on geometric mean of all benchmarks can be reduced by $1.98X$ and $1.96X$ for min-FF and min-power solution, respectively. Combining the over-estimation factors in Section 4.1, the overall reduction of over-estimation is $2.46X$ and $2.18X$ for min-FF and min-power solution, respectively. Given such big differences in power, the cycle-accurate interconnect power simulation is needed to obtain accu-

rate interconnect power and validate power reduction innovations.

4.3. Performance Impact

In this section, we study this performance impact from FF insertion with min-FF solution. In microprocessor designs, FF insertion affects the IPC because it requires addition pipeline stages at microarchitecture level to amortize the interconnect delay. Therefore, we can not improve processor performance by simply increasing clock. For a given program, we choose BIPS (Billion Instruction Per Second) to represent the performance as shown in (7):

$$BIPS = \frac{IPC \times cycle\_frequency}{10^9} \quad (7)$$

where $cycle\_frequency$ is the processor clock frequency.

Table 11 compares the IPC and the BIPS for the cases without and with FF insertion. The maximum clock we can achieve without FF insertion is 880MHz. With FF insertion, although the IPC always decreases, BIPS increases by $1.32X$ to $2.03X$. In conventional design flow, the system design optimizes IPC and an *independent* VLSI design optimizes clock frequency. This design flow may be no longer valid as indicated by Table 11. Therefore, iterations between system design and VLSI design or simultaneous optimization [13] [3] is necessary for better performance and power efficiency. Our concurrent repeater and FF insertion in this paper can be viewed as an initial effort on simultaneous optimization.

Benchmark	Without FF		With FF		With FF BIPS
	IPC	BIPS	IPC	BIPS	improve factor
go	0.893	0.714	0.618	1.236	$1.73X$
li	1.455	1.28	0.954	1.909	$1.49X$
perl	1.045	0.92	0.784	1.568	$1.704X$
compress	1.727	1.52	1.050	2.010	$1.32X$
swim	1.231	1.08	0.949	1.898	$1.76X$
mgrid	1.665	1.465	1.036	2.072	$1.41X$
fpppp	0.825	0.726	0.739	1.477	$2.03X$
applu	1.418	1.25	0.864	1.729	$1.38X$

Table 11: IPC and BIPS for cases without and with FF insertion. The clocks for the two cases are 880MHz and 2GHz, respectively. Min-FF solution is applied.

5. CONCLUSIONS AND DISCUSSIONS

We have shown that repeater insertion is no longer sufficient to achieve the target frequencies specified by ITRS, and have developed concurrent repeater and FF insertion schemes minimizing the number of FFs (min-FF) or minimizing the power (min-power), both subject to the delay specification. We have observed that the power in the min-FF solution has a super-linear growth with respect to the wire length, but the power has a nearly linear growth in the min-power solution.

Considering structural interconnects, layer assignment and concurrent repeater and FF insertion for delay specification, we have developed a cycle-accurate microarchitecture-level interconnect power simulation. The simulation reduces the over-estimation by up to $2.46X$ compared to power estimation based on purely stochastic interconnects and fixed switching factor.

Furthermore, we show that although interconnect pipelining has a lower IPC but it can improve throughput by up to $2.03X$. In conventional design flow, the system design optimizes IPC and an *independent* VLSI design optimizes clock frequency. This design flow may be no longer valid as indicated by the above example. We plan to simultaneously optimize microarchitecture and floorplanning with consideration of FF insertion explicitly.

[3] No FF insertion is considered in [13].

6. REFERENCES

[1] D. Sylvester and K. Keutzer, "Imapct of small process geometries on microarchitectures in systems on a chip," *Proceedings of the IEEE*, vol. 89, no. 4, pp. 467–489, 2001.

[2] P. Kapur, G. Chandra and K. C. Saraswat, "Power estimation in global interconnects and its reduction using a novel repeater optimization methodology," in *DAC 02*, June 2002.

[3] W. Ye, N. Vijaykrishnan, M. Kandemir, and M. J. Irwin, "The design and use of simplepower: a cycle-accurate energy estimation tool," in *DAC*, 2000.

[4] D. Brooks, V. Tiwari, and M. Martonosi, "Wattch: A framework for architectural-level power analysis optimization," in *ISCA*, 2000.

[5] *The International Technology Roadmap for Semiconductors*. http://public.itrs.net/, 2001.

[6] W. Liao, J. M. Basile, and L. He, "Leakage power modeling and reduction with data retention," in *ICCAD 02*, Nov 2002.

[7] J. A. Davis, V. K. De and J. Meindl, "A stochastic wire-length distribution for gigascale integration (gsi)-part 1:derivation and validation," *IEEE Transaction on Electron Devices*, vol45, March 1998.

[8] P. Cocchini, "Concurrent flip-flop and repeater insertion for high performance integrated circuits," in *ICCAD*, Nov 2002.

[9] *Berkeley Predictive Technology Model (BPTM) 0.10μm SPICE Model Cards*, July 2000.

[10] K. Banerjee and A. Mehrotra, "Power dissipation issues in interconnect performance optimization for sub-180 nm designs," in *Proceedings of 2002 Symposium on VLSI Circitus*, 2002.

[11] W. Liao and L. He, *Full-Chip Interconect Power Estimation and Simulation Considering Concurrent Repeater and Flip-Flop Insertion.* Technical Report "UCLA ENG 03-238" (available at http://eda.ee.ucla.edu/), University of California at Los Angeles, 2003.

[12] D. Sylvester and K. Keutzer, "Getting to the bottom of deep submicron," in *Proc. Int. Conf. on Computer Aided Design*, pp203–211, 1998.

[13] J. Cong, A. Jagannathan, G. Reinman, and M. Romesis, "Microarchitecture evaluation with physical planning," in *DAC*, June 2003.

[14] D. Burger and T. Austin, *The simplescalar tool set version 2.0.* Technical Report CS-TR-97-1343, University of Wisconsin-Madison, 1997.